University Research Planning in the Data Era: Working with the Levers and Pulleys that Tie Together Research Information, from Big Data to Local Details

Merrill Series on
The Research Mission of Public Universities

A compilation of papers originally presented at a retreat
sponsored by The Merrill Advanced Studies Center
July 2017

Mabel L. Rice, Editor
Technical editor: Suzanne Scales

MASC Report No. 121
The University of Kansas

TABLE OF CONTENTS

MASC Report No. 121

Panel 3: Researchers

Panel 4: Researchers

Panel 5: Research Administrator

Panel 6: Researcher and Administrator Panel

Introduction

Mabel Rice

The Fred and Virginia Merrill Distinguished Professor of Advanced Studies and Director, Merrill Advanced Studies Center, University of Kansas

The following papers each address an aspect of the subject of the twenty-first annual research policy retreat hosted by the Merrill Center: *University research planning in the data era: Working with the levers and pulleys that tie together research information, from big data to local details.* We are pleased to continue this program that brings together University administrators and researcher-scientists for informal discussions that lead to the identification of pressing issues, understanding of different perspectives, and the creation of plans of action to enhance research productivity within our institutions. This year the focus was on opportunities and challenges of big data for research in public universities.

Our keynote speaker for the event was Dr. Michael Huerta of the National Library of Medicine, National Institutes of Health. He is helping to lead the institute's Big Data to Knowledge (BD2K) initiative which will support research and development in the area of data science and associated technologies. Importantly, BD2K will also work to change policies and practices at NIH to raise the prominence of data in the biomedical research enterprise by increasing data sharing, supporting community-based standards efforts, and making data sets discoverable, citable, and linked to the scientific literature.

Benefactors Virginia and Fred Merrill make possible this series of retreats: The Research Mission of Public Universities. On behalf of the many participants over two decades, I express deep gratitude to the Merrills for their enlightened support. On behalf of the Merrill Advanced Studies Center, I extend my appreciation for the contribution of effort and time of the participants and in particular, to the authors of this collection of papers who found time in their busy schedules for the preparation of the materials that follow.

Twenty administrators, faculty, and students from five institutions in Kansas, Iowa and Nebraska attended in 2017, which marked our twenty first retreat. Additionally, two executives from the American Speech-Language-Hearing Association attended this year. Though not all discussants' remarks are individually documented, their participation was an essential ingredient in the general discussions that ensued and the preparation of the final papers. The list of all conference attendees is at the end of the publication.

The inaugural event in this series of conferences, in 1997, focused on pressures that hinder the research mission of higher education. In 1998, we turned our attention to competing for new resources and to ways to enhance individual and collective productivity. In 1999, we examined in more depth cross-university alliances. The focus of the 2000 retreat was on making research a part of the public agenda and championing the cause of research as a valuable state resource. In 2001, the topic was evaluating research productivity, with a focus on the very important National Research Council (NRC) study from 1995. In the wake of 9/11, the topic for 2002 was “Science at a Time of National Emergency”; participants discussed scientists coming to the aid of the country, such as in joint research on preventing and mitigating bioterrorism, while also recognizing the difficulties our universities face because of increased security measures. In 2003 we focused on graduate education and two keynote speakers addressed key issues about retention of students in the doctoral track, efficiency in time to degree, and making the rules of the game transparent. In 2004 we looked at the leadership challenge of a comprehensive public university to accommodate the fluid nature of scientific initiatives to the world of long-term planning for the teaching and service missions of the universities. In 2005 we discussed the interface of science and public policy with an eye toward how to move forward in a way that honors both public trust and scientific integrity. Our retreat in 2006 considered the privatization of public universities and the corresponding shift in research funding and infrastructure. The 2007 retreat focused on the changing climate of research funding, the development of University research resources, and how to calibrate those resources with likely sources of funding, while the 2008 retreat dealt with the many benefits and specific issues of international research collaboration. The 2009 retreat highlighted regional research collaborations, with discussion of the many advantages and concerns associated with regional alliances. The 2010 retreat focused on the challenges regional Universities face in the effort to sustain and enhance their research missions, while the 2011 retreat outlined the role of Behavioral and Social sciences in national research initiatives. Our 2012 retreat discussed the present and future information infrastructure required for research success in universities, and the economic implications of that infrastructure, and the 2013 retreat discussed the increasing use of data analysis in University planning processes, and the impact it has on higher education and research. The 2014 retreat looked at the current funding environment and approaches which could be used to improve future funding prospects. The 2015 retreat addressed the opportunities and challenges inherent in innovation and translational initiatives in the time of economic uncertainty that have an impact on goals to enhance research productivity. The 2016 retreat focused on the building of infrastructure to meet the changing needs in research.

Once again, the texts of this year’s Merrill white paper reveal various perspectives on only one of the many complex issues faced by research administrators and scientists every day. It is with pleasure that I encourage you to read the papers from the 2017 Merrill policy retreat on: *University research planning in the data era: Working with the levers and pulleys that tie together research information, from big data to local details.*

Executive Summary

Realizing the Promise of a Digital Ecosystem for Science and Scholarship

Michael F. Huerta, PhD, Associate Director for Program Development and NLM Coordinator of Data Science and Open Science, National Library of Medicine, National Institutes of Health

- The National Library of Medicine (NLM) joined the National Institutes of Health (NIH) in 1968. NLM conducts and supports research and training in information science, informatics, and data science. It is also the world's largest biomedical and medical library. In addition to its vast collection of book, journal, manuscripts and other items, the NLM is home to hundreds of digital data and information resources. It receives and delivers a vast amount of digital content for user including researchers, healthcare providers, and the general public.

- Medicine and biomedicine are a substantive scope of the NLM. As biomedical research becomes increasingly digital, the NLM will likely pay attention to digital research objects (DROs), which might include software used to generate or analyze research data, as well as models and workflows used in research. After the NLM applies information science, informatics and data science to the digital research objects, they are findable, accessible, interoperable, and re-usable (FAIR). The processes of NLM applied to DROs make those objects compliant with FAIR principles.

- When DROs are findable, accessible, interoperable, re-usable, and attributable, they make possible a more data-centric and open paradigm of science and scholarship. To bring DROs into an open ecosystem, first the data must be shared. The benefits and objections to data sharing are discussed. Most biomedical research does not use a data-centric and open approach, but rather a concept-centric approach. This is about to change with both society expectations to data from funded research to be available and directives from federal government to encourage data sharing. The author discusses what the NIH is doing to make digital research objects findable, accessible, interoperable and re-usable.

- Key issues have been identified for NLM to assume the leadership role for data science and open science at NIH. One is to engage with others across the NIH as economies of scale and experience can be realized with a strategic enterprise approach. Another is the use of evidence based value assessment of data to provide guidance about future investments in data, infrastructure and policy. Other priorities include strategic engagement beyond NIH; development of a data-savvy workforce; promotion of open science; and research and innovation in data science and open science.

- The cumulative biomedical knowledgebase and breathtaking powerful scientific technologies available today present significant opportunities to understand health and mitigate illness. A digital ecosystem wrought of data science and open science promises to multiply these opportunities many-fold.

From Hospital Informatics Laboratories to National Data Networks: Positioning Academic Medical Centers to Advance Clinical Research

Lemuel R. Waitman, Professor, Department of Internal Medicine, Associate Vice Chancellor for Enterprise Analytics
University of Kansas Medical Center

- Pioneering academic medical centers (AMC) have been leaders in developing medical informatics systems to improve patient care and aggregate biomedical data to advance research. The potential to aggregate biomedical data now extends to all healthsystems. Led by the National Institutes of Health and the Patient Centered Outcomes Research Institute's creation of PCORnet, federal, nonprofit, and industry sponsors along with clinicians, patients, and investigators are seeking to capitalize on these new clinical data. Institutions are creating local, regional and national data networks that can support research and realize the vision of a learning heath system.

- The 2010 proposal for the University of Kansas Medical Center's Clinical and Translational Science Awards (CTSA) program, Frontiers, provides an example of a regional vision for biomedical informatics. The program's central aim was creating HERON Clinical integrated data repository. The open source i2b2 for data integration and warehousing was implemented. In addition to i2b2, Frontiers biomedical informatics adopted and promoted REDCap as a common tool for research data capture across the enterprise and partner institutions.

- Frontiers biomedical informatics' choice of i2b2 and REDCap was fortuitous for supporting broader collaboration nationally. Frontiers biomedical informatics saw high alignment with its work for integrating data in support of the CTSA program and the PCORI funding opportunity to create a Clinical Data Research Network (CDRN). Frontiers worked with other institutions in the Midwest to organize a response and create the Greater Plains Collaborative (GPC) and successfully compete for an initial Phase 1, and subsequent Phase 2 CDRN contract.

- In addition to becoming a viable Clinical Data Research Network, network partners' efforts were shifted to support a new data infrastructure, the PCORnet Common Data Model (CDM). As they worked to develop the network, the difference in perspectives between epidemiology focused coordinating center data modeling team and those embedded in health systems with rich clinical research goals was apparent.

- While much of PCORnet's activity was establishing governance and data infrastructure, the network also collaboratively prioritized and devised three national demonstration projects: the prospective ADAPTABLE pragmatic trial and two observational studies regarding obesity.

- As the University of Kansas Medical Center and peers in the Greater Plains Collaborative complete four years of building PCORnet, they reflect that this participation has impacted the campuses. The majority of the campuses are involved in all three demonstration projects. Their network leads the national collaborate research group for advancing PCORnet's cancer research; and Dr. Russ Waitman has served as the national chair for the PCORnet data committee. PCORI announced in 2017 that it will transition infrastructure support to a newly created nonprofit which will in turn contract with

Clinical Data Research Networks instead of networks contracting with PCORI. Though this will provide flexibility, questions arise as to how structure informs network design and collaboration.

Cross-disciplinary Activities in Big Data for Agricultural Innovation

Carolyn J. Lawrence-Dill, Ph.D., Associate Professor, Department of Genetics, Development & Cell Biology and Department of Agronomy
Iowa State University

- Agriculture is broad, involving not only crops and animals, but also the ecosystems that support their growth and development. Pressures on agricultural systems are increasing and there are pressures for improvements in agriculture, which tell us that we need to discover, design, and invent news ways to improve agricultural products.

- Solving agricultural problems involves a multidisciplinary approach. A way to engage a broader group is to make data that describes ecosystems, crops and animals more accessible to researchers. This extreme data sharing is in keeping with long-standing traditions in science. Limiting access to data stands in the way of agricultural innovation, and that position cannot be supported.

- Data standardization seeks to improve both human and machine access to and analysis of data. Phenotype is the primary datatype selected for crop improvement and it includes many different types of data imaginable, making standardization difficult. The development of standards is in it's infancy, with the first standard for data was released only two years ago. In opposition, is the view against the development and use of standards for this emerging field of research. The concerns against standardization make the debate a topic at scientific meetings where phenotyping is a focus.

- There is a need for scientist with broad expertise to work together to address agricultural issues. Through the Iowa State University Plant Sciences Institutes (PSI) Faculty Scholar initiative, researchers working in the areas of plant sciences, data sciences, and engineering are funded to focus on plant phenomics problems. Another Iowa State initiative on this front is a grant from the National Science Foundation in Predictive Plant Phenomics (P3) that supports graduate education and research.

- The general approach to agricultural improvement must evolve to meet anticipated future needs. Researchers are developing the infrastructure and human resources to support the development of a new paradigm for research that results in agriculture innovation.

Developing Data Science at UNL: Progress, Challenges, and Opportunities for Research

Jennifer L. Clarke, PhD, University of Nebraska

- Over the past several years we have seen a groundswell of interest and investment in data science, as the author has come to appreciate data science as more encompassing endeavor that encourages interdisciplinary research. The author was hired by the University of Nebraska-Lincoln in 2013 in the primary role as Director of the Quantitative Life Sciences Initiative (QLSI), whose mission is to develop expertise and resources in data science and 'Big Data' to meet the growing needs for the disciplines in the Life Sciences. Advances in computing has brought us the era of "Big Data", which can be defined as more data than one is accustomed to or more than one can manage. The four common attributes of Big Data are volume, velocity, variety, and veracity.

- One of the challenges of the 21st century science is how to get from data to information to knowledge when data are large, noisy and complex. This process requires a diverse skill set drawn from many disciplines. To meet the national workforce needs in data science, we need to rethink undergraduate, graduate and continuing education. Through a process of identifying opportunities to benefit the campus and stakeholders, UNL decided to develop an interdisciplinary doctoral program in Complex Biosystems. QLSI has active research and partnerships with local, regional, national, and international organization. These partnerships are critical to the success of the initiative because the field is evolving. Partnerships are an effective way to stay informed of developments, and they provide opportunities for graduate training.

- A recent area of emphasis for QLSI is reproducible research and Big Data management and analysis. The collaborate activities help support faculty associated with federally supported research centers comply with standards and expectations. How to finance the maintenance and sharing of data remains a challenge that must be overcome.

Enhancing and Automating University Reporting Of R&D Expenditure Data Using Machine Learning Techniques

Joshua L. Rosenbloom, Iowa State University, National Bureau of Economic Research
Rodolfo Torres, University of Kansas
Joseph St. Amand, University of Kansas
Adrienne Sadovsky, University of Kansas

- Most of what we know about research and development performed at the nation's colleges and universities is derived from data collected by National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES) as part of its Higher Education R&D (HERD) survey. The data collected in the HERD Survey are derived from institutional responses to an annual survey sent by NCSES.

- Responsibility for responding to the HERD survey at research universities is likely delegated to one or more specialist and this method of collecting data results in three distinct problems: responding is costly, there is a lag time in the availability of data, and there are inconsistencies of data collection across institutions, and even variation within an institution. To address these problems, the authors engaged in an experiment to apply techniques of machine learning to automated project classification. They determined these are potentially feasible but require further efforts.

- The goal of their project is to develop a classification algorithm that can be used to either supplement or replace human judgement in classifying sponsored research projects. Working with the University of Kansas Office of Research, they obtained complete data from approximately 1500 historical projects. The process and results of the experiment of applying machine learning to predict project classification are discussed in the paper. Among the different machine learning models, the authors found that the Logistic Regression classifier provides the best overall performance.

- The authors have not yet succeeded in developing a set of classifiers that precisely reproduce the human judgements underlying the University of Kansas' response to the HERD Survey, though it is not clear this should be a measure of the project's success. The project has been successful in showing that developing reasonably accurate machine-learning classifiers is possible. Future goals for the project include assessing the ability of the classifiers to successfully classify projects at other institutions. Classifiers can further be refined through adding additional projects from other institutions to the training data set.

Clinical Research and Data: HIPAA, the Common Rule, the General Data Protection Regulation, and Data Repositories

Amy Jurevic Sokol, Associate General Counsel
The University of Kansas Medical Center

- The way we do clinical research has changed. This article touches on different legal aspects arising at the intersection of technology, data, and clinical research—specifically HIPAA (the Health Insurance Portability and Accountability Act), human subjects research, the European data law (the General Data Protection Regulation), and data repositories. It explains how two different law-making bodies, the US and EU, have tried to balance the needs of the use of data with the privacy and risk issues.

- There is not one overarching law that protects all data. Instead, the US has a patchwork of federal and state laws that protect different types of data. HIPAA applies a different standard than that of the Common Rule and FDA Regulations, which require there are provisions in place to protect privacy of subjects and confidentiality of data. HIPAA applies to "covered entities", and may or may not apply to researchers, depending on the situation.

- Some researchers incorrectly believe removing certain information de-identifies data under HIPAA. To be considered de-identified, it must meet the requirements of safe harbor or expert determination. The safe harbor requires removal of specific identifiers of the patient and the patient's relatives, employers or household members. Under the expert determination method, it must be determined that the risk is very small that the information could be used alone or combined with other available information to reidentify an individual. HIPAA and its regulations do not apply to de-identified data under either method.

- Researchers often need information that is not available in properly de-identified data sets. A limited data set (LDS) is protected health information where some information is permissible to remain, and some information has been removed. HIPAA Privacy Regulation require covered entities must enter a data use agreement with recipients of LDS.

- There are two separate legal analyses that must occur when creating a data repository; does HIPAA apply and is it considered human subject research under the Common Rule. Each time protected health information is accessed for research, then the requirements for access must be met. There is the HIPAA analysis and the Common Rule analysis for accessing data. If a limited data set or fully identifiable protected health information is requested, then certain circumstances and conditions must by met under HIPAA. The Common Rule analysis is equally as complicated.

- Issues arise when US researchers want to use data from other countries for their research. Researchers who use data from multiple countries must navigate not only their own country's laws, but also the international legal waters.

Hitting the Mark– Facilitating Research Administration to Support the Institutional Strategic Plan

Ian Czarnezki, MBA, Director of Operations, Office of the Vice President for Research, Kansas State University

- Kansas State University has a bold vision to be recognized as one of the nation's Top 50 Public Research Universities by 2025. This vision presents a challenge for research leadership on how to monitor the progress and facilitate growth. K-State will need to roughly double its research expenditure to achieve Top 50 public university status.

- Due to the scope of the bold vision, K-State needs the ability to understand how each award impacts the progress toward the overall goal. To accurately assess the progress towards the institutional goals, information needs to be harvested from each of K-State's three disparate systems; human resource information system, financial information system, and research administration system. K-State has undertaken a reporting initiative to provide a cohesive and timely view of the research activities. K-State Consolidated Award Tracking System (K-CATS) is the branded research administration intelligence solution that gives leadership and stakeholders insight into the research activities.

- The HERD Project is a collaborative effort between Kansas State University and Microsoft. This reporting solution will allow for greater insight into K-State's research activities compared to other institutions. K-State is utilizing the wealth of information regarding its research activities to help align research with funding opportunities, highlight interdisciplinary partners, and to move its research efforts forward.

Influencing the Culture of Scholarly and Professional Communities to Advance Clinical Research and Accelerate Knowledge Translation

Margaret A. Rogers*, Chief Staff Officer for Science and Research American Speech-Language-Hearing Association
Michael Cannon. Director, Serial Publications and Editorial Services American Speech-Language-Hearing Association

- Professional and scientific associations for health care disciplines have an opportunity to help shape how evidence-based practice becomes integrated into the fabric of the professions that they support. The efforts of these associations to "bridge the research-to-practice gap" are numerous, with the most promising are efforts that make use of big data, especially when coupled with text and data mining, semantic computing, and artificial intelligence. Three areas have been evolving over the past 75 years that have shaped the priorities of the American Speech-Language Hearing Association: evidence-based practice and implementation science movements; rapid changes in healthcare; and big data and data science, which is redefining scientific publishing.

- In this paper, the authors discuss the historical roots of evidence-based practice and data driven outcomes improvement. Physician Archibald Cochrane's work yielded the terms effective, efficient and equitable. Coupled with three additional domains, safe, patient-centered, and timely, these became the cornerstone for assessing ROI for health care expenditures in the U.S. The Commonwealth Fund supports research that compares health care quality and expenditures across high-income countries. Despite the authoritative data from this report, it is a puzzling, yet predictable phenomenon that it has not had a more influential effect. It has been observed that despite compelling scientific evidence, behavior and attitudes do not necessarily change, and if so, change will be slow. Everett Roger's *diffusion of innovations theory* and Prochaska and DiClemente's subsequent *transtheoretical model of change* are presented. Other theories of change also contribute to our understanding of how behaviors and attitudes might be influenced to promote the adoption of evidence-based practices.

- Dissemination and implementation science is a growing focus of research that seeks to lessen the gap of new knowledge and its application by identifying the factors that influence change. Estimates of the time it takes for research to become translated into evidence-based policies, programs, and practices is about 15 and 20 years. Though it is a challenging process, there is a consensus that evidence-based practices need to be integrated into clinical care at a more rapid pace. Understanding the strategies and factors that can help or hinder new knowledge has become a central focus in health care. Using a combination of dissemination approaches, perhaps the most important of which include social learning opportunities, could help achieve the goal of improved health through better evidence-based decisions.

- Clinical data registries hold much promise to fill in gaps in the investigator-initiated research. Because clinical data registries and electronic health records accumulate large samples of patient populations, there are questions that are best addressed through big data and data science. The vision of a learning health care system is that analyses of large clinical data repositories will provide information about what works best for whom under which circumstances. Decisions can then be made about improving services and outcomes. After these adjustments, new data will provide information on the success and failure of the adjustments. Learning health care systems are expected to accelerate the rate at which evidence-base practices and innovations in health care are adopted; thereby, reducing the research-to-practice gap.

- There are many ways that professional and scientific associations can leverage their innate strengths to increase the implementation strategies. Publishers, such as the American Speech-Language-Hearing Association, have increasingly adopted continuous publishing models so important research can be timely released. The standardization of the data behind the full range of publication steps has also shaved from the time it takes to disseminate knowledge. In this publishing era, granular tagging is applied to articles, extending a user's discovery. All these advances are emblematic of the tide of big data flooding all publishers. The next two decades should lead to measurable improvements in reducing the gap from research to practice.

Towards a Research Profiling Ecosystem
Weaving Scholarly, Linked Open and Big Data

David Eichmann, School of Library and Information Science &
Iowa Graduate Program in Informatics
The University of Iowa

- Research profiling systems provide programmatic support for discover and use of research and scholarly information. Many systems have been developed including open source, commercial, and local institutional systems, such as Loki, the University of Iowa's research profiling system. The work on extending Loki into the Semantic Web serves as a substantial case study in modular architectures extending into Linked Open Data (LOD). Loki is an investigator-rather than institutionally-focused, supporting many phases of the research life cycle.

- Several approaches in the design of Loki proved to be valuable. Work involved definition of a Loki ontology and the mapping of relational database entities into the resulting ontological concepts, including synthesizing the tag library layer of the architecture from an entity-relationship diagram. Furthermore, the clean partitioning of the logical components allowed them to represent those components as discrete triple stores, supporting an overall LOD environment of interlinked triple stores that reflected the modularity of the initial tag library design. Several conscious design decisions were made in the development of Loki. Initially, they opted to develop a Loki ontology that directly represents the semantics of their local environment. Subsequently, they mapped the Loki ontology to the VIVO ontology, demonstrating the value in maintaining separation between the representational and conceptual levels in the overall information architecture.

- CTSAsearch is a federated search engine using VIV-compliant Linked Open Data published by 87 institutions. User feedback on CTSAsearch showed a desire for more sophisticated search than what was provided by a simple 'bag-of-words' relevance list. The default search mode currently used has been successful in pruning low level relevance hits from results. For "reasonable" result scales, approximately 200 queries, useful force graph visualizations are possible. Challenges arise when queries return thousands of results, leading to a network hairball. Two approaches have been taken to address this challenge: one is aggregating results at the institution; and a second is inter-institutional community visualization.

Aligning Data Collection with Multi-Dimensional Construct Definitions: The Example of Behavioral Tasks for Measuring Risk–taking Behavior

Carl W. Lejuez, Dean, College of Liberal Arts & Sciences,
University of Kansas

- Data plays an important role in the understanding of real world risk taking behavior. Ensuring the quality of that data requires an understanding of the rational for the tasks developed and used. It also requires a clear sense of what useful existing data or new behavioral tasks can provide and where they fall short.

- The Iowa Gambling Task (IGT) and the Balloon Analogue Risk Task (BART) assess risk taking in different ways, and together they may provide a strong comprehensive picture. IGT was the original standard for measuring risk taking. Slovic's Devil Task was the first behavioral task designed and used to assess Risk Taking Propensity (RTP), a commonly used behavioral measure currently used. The simplicity of the Devil's Task led to the development of the BART, a computerized measure of RTP, that allows for complex ways to study complex risk behavior seen in the real world.

- In the BART task, the participant is presented with a balloon and asked to pump the balloon by clicking a button on the screen. As the balloon inflates, winnings are added; however, if pumped beyond the explosion point, the balloon explodes, and the participant loses the money earned on the balloon. Existing data shows a correlation between risk taking on the task and current levels of real world risk behavior. However, there appears to be no evidence of risk taking on the BART at one time point that predicts future risk-taking behavior. Several studies have been done to understand how risk taking is impacted by external factors in the real world by manipulating those factors in a controlled laboratory. The studies presented include one that examined the effects of varying cash reward magnitudes on RTP; another that examined the effect of peer influence on BART RTP; and one that examined the impact of anxiety on risk taking.

- Isolating risk taking in a controlled laboratory setting and providing the opportunity to manipulate key variables thought to impact risk behavior in the real world should be the focus of experimental studies. This work has begun with the BART, but work including studies that bring in genetic factors, neural assessment, as well as environmental factors is crucial to further progress.

Aligning Researcher Practice to Support Public Access to Data

Surya K. Mallapragada, Associate Vice President for Research
Iowa State University

- There is a national move towards open science and open enquiry. The resources and systems for openly sharing publications are well developed, though the policies for data sharing are less defined. Open access to data will be effective if there are common standards for communicating data and a cohesive strategy used.

- AAU-APLU Public Access Working Group is working on common goals for data sharing, federal agency recommendations and guidance for research institutions. At Iowa State University, the implementation is being coordinated across the Library, IT services and the Office of the Vice President for Research. A faculty committee is providing perspective for the rigorous process of data sharing. Key questions for consideration to develop polices are: what is the purpose of sharing, what data should be shared, what is the standard for documenting data, what are options for data storage, and how to train researchers to adopt the new mindset.

If a Tree Fell in a 300 Million-year Old Forest, Did it Leave a Data Trail?

Joseph A. Heppert, Ph.D., Vice President for Research
Texas Tech University

- Researchers in many fields at universities are creating masses of data at a record rate. This paper explores the sources of this increase in data, describes the challenges created by the ever-increasing pace of data creation, and looks at the strategies universities are considering in managing the expansion of data creation.

- Because of technological advances, compute capacity per dollar increases, and a decrease in the price of data storage capacity, investigators are analyzing more comprehensive and realistic data sets. These big data applications are being used in the analyses of varying fields. Universities are addressing this challenge of supporting research with big data by investing in high-quality high-performance research computing (HPC). The University of Kansas (KU) has only had a centralized HPC strategy for five years.

- The increase in the size of data sets has offset the cost savings of declining cost of computing, network and storage capacity. Though there are desires to off-load HPC computing and storage capacity to the cloud, unfortunately, most present analyses of cloud computing services do not support moving university enterprise HPC to a cloud platform. Accessing data that is stored in the cloud adds to the cost of the service, yet, glacial cloud storage which is used for long-term archiving is more cost-effective.

- There are several key challenges facing universities and researchers. Investigators have flocked to low-cost and sometimes low-quality technologies for data storage. In response, funding agencies have begun to intervene out of concern for data integrity and accountability. The mandate for Office of Science and Technology (OSTP) to make data collected through federal funding available to the public has created a dilemma for research universities as few have the server capacity for public access and security concerns are an issue. Another challenge is the leakage of academic research and development activities.

- To promote a healthy data culture in higher education, the following is recommended; provide economical access to professionally maintained computing capacity and archival storage; give ownership of computational and storage hardware to commercial vendors; facilitate the transition of research records to electronic records; standardize meta-data to identify data sources and ownership; create internal training and policies to minimize the volume and extended time of retained data; engage disciplinary experts to incorporate data management best practices; develop shared application interfaces to bring computing tasks to large data sets, create institutional capacity to ensure compliance; and continue dialog with funding agencies about sustainable support for research data archives.

Data, Consent, Privacy, and Insight

Daniel A. Reed, University of Iowa

- The changes brought about by technology are deep and profound. Some of the changes include the creation of megacities, concentration of wealth in a small fraction of the population, direct consumer engagement resulting in elimination of some existing companies and creation of new ones, and polarization of social perspectives and political opinions.

- Technological change continues against this backdrop of social issues. Digitization of our world ranks at the top of technological change with examples including big data, deep learning, automation, biomedicine advances, and environmental change and global warming.

- Data is important in both enabling technical changes and remediating the damaging effects. Within this context, the paper discusses the scale and scope of big data, the privacy and legal challenges created by digital data flows, and the emerging issues surrounding sensors and passive data. Thoughts are shared on a new model of digital privacy. Combining three principles creates a more nuanced model for data sharing: one principle that attaches a lifetime to data at the time of its release, a second principle limits sharing of data, and a third principle is claims-based access that would specify the purpose for which the data can be used.

- Within the broader context of social and technological change, we must ask wise and thoughtful questions about how this data is used and by whom. Only by concurrently considering social implications and technological capabilities can we create sustainable approaches.

Research Planning at Nebraska
Research and Economic Development Growth Initiative (REDGI): 2012-2017

Steve Goddard, Vice Chancellor for Research & Economic Development
University of Nebraska-Lincoln

- The University of Nebraska-Lincoln's launching of the Research and Economic Development Growth Initiative (REDGI) is an example of the use of data and analytics in research planning. In 2011, the University of Nebraska Lincoln was ranked as one of the top US universities in research growth over the previous 10 years. In his 2011 State of University Address, Chancellor Harvey Perlman emphasized the need to increase the academic stature and gave these specific goals: increase total research expenditures to $300 million, increase academic stature through increased faculty awards and memberships, increase the number of faculty working with scientists in the private sector, and increase student enrollment by 20%.

- The Office of Research & Economic Development (ORED) was charged with carrying out the research growth goals, a mission that would require buy-in from research-active administrators, faculty and staff. From the fall of 2011 through spring 2012, targeted forums with key audiences were held to discuss issues to accomplish the goals and solicit input on the strategies. Following these forums, the Research and Economic Development Growth Initiative (REDGI) was created with two broad goals: to enhance the quality of research, scholarship, and creative activity at UNL, and to increase the quality and quantity of industry partnerships.

- Metrics would be a driving force for REDGI. The approach was to use a variety of analytical tools to better understand UNL's scholarly strengths and market position. REDGI defined specific metrics and actions to meet each of its two goals. The REDGI roll-out to campus included promotions and education campaign to engage the campus in the effort of the new platform for disseminating data and analytics to track their progress toward the REDGI goals. REDGI dashboards were developed and made available specific to the university, college and department levels.

- Metrics are provided on the success of the project. Total research expenditures for the Fiscal year 2017 nearly meet the fiscal year 2018 goal of $300 million. The growth goals for industry funding were not met, but UNL exceeded the goals for faculty engagement in sponsored programs and exceed by almost double the number of faculty awards and memberships.

- Several lessons were learned from the REDGI experience. Engaging leaders at all levels is critical to success. Incorporating goals into the story of the research institution is critically important. New staffing is necessary with any new, large undertaking. Most important, strong and clear data measurements and analytics is critical.

Realizing the Promise of a Digital Ecosystem for Science and Scholarship[i]

Michael F. Huerta, PhD, Associate Director for Program Development and NLM Coordinator of Data Science and Open Science, National Library of Medicine, National Institutes of Health

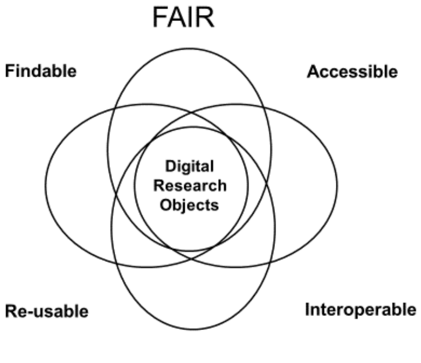
What is the National Library of Medicine and What Does It Do?

The National Library of Medicine (NLM) joined the National Institutes of Health (NIH) in 1968. As an NIH Institute, NLM conducts and supports research and training in information science, informatics, and data science. The NLM is also the world's largest biomedical and medical library, tracing its origins to the library of the Office of the Surgeon General of the Army in 1836[ii]. Today, in addition to its large collections of physical items, including books, journals, manuscripts, photographs, and other items, NLM is also home to hundreds of digital data and information resources. These include major resources, such as ClinicalTrials.gov[iii], which houses information about and from hundreds of thousands of clinical trials, and MedlinePlus[iv], which provides authoritative consumer health information, as well as smaller resources serving important niche purposes, such as TOXNET, which is a collection of databases and information products related to toxicology, environmental health, and hazardous substances[v].

Every day, NLM receives more than 10 terabytes of digital content from more than 3000 users, and delivers more than 100 terabytes to more than 4 million users, often through application programming interfaces. Users of these resources include researchers, healthcare providers, and the general public. The Library supports activities that engage all categories of users to make its resources known, understood and used. For example, to facilitate and enhance health information access to the general public throughout the country, including many rural areas, the NLM supports the National Network of Libraries of Medicine[vi]. Through its eight regional medical libraries, the Network reaches more than 6500 points of presence across the country in academic health science, community college, tribal college and public libraries, as well as other organizations, such as community health centers.

With medicine and biomedicine as its substantive scope, the NLM has been paying attention to that literature for more than 180 years. In 1879, John Shaw Billings, Director of the Library of the Surgeon General of the Army, and Robert Fletcher compiled and had published *Index Medicus*, an index of medical books, journals, and pamphlets[vii]. Stewarded by NLM, *Index Medicus* continued to be the authoritative index of the medical literature until 2004, but by 1964 the Library had started compiling citations and indexing much of the biomedical and medical literature digitally, in a database system called MEDLARS. In 1971, this database became available online (mostly through university libraries) as MEDLINE[viii]. And, in 1997, PubMed, which included MEDLINE bibliographic data and more, was launched for free use by anyone on the World Wide Web. Today, PubMed contains more than 27 million bibliographic citations[ix].

In addition to bibliographic data, NLM has established databases and resources for particular types of research data. These include GenBank, a database containing all publicly available DNA sequence data with annotations[x], and others[xi]. As biomedical research becomes increasingly digital, the NLM will likely pay attention to research objects beyond research data and bibliographic data. Such digital research objects[xii] (DROs) might include software used to generate or analyze research data, as well as models, workflows, etc., used in research. (It is important to note that "pay attention to" covers a broad range of possible activities.) To DROs, such as citations or datasets, NLM applies: information science to curate acquired objects, informatics to compute in context on these objects, and data science to extract insight from these objects. After this, the DROs are findable (e.g., by having had metadata assigned), accessible (e.g., through publicly accessible databases), interoperable (e.g., by having adopted common data-related standards), and re-usable (e.g., by linking one set of objects, like publication citations, to another set of objects, like the datasets reported on in those publications). Thus, the processes of NLM applied to DROs make those objects compliant with the FAIR Principles[xiii]. In addition, NLM is interested in making DROs attributable (e.g., through PubMed Identifiers, PMIDs, or GenBank Accession Numbers) and sustainable (NLM considers this carefully before committing to hosting DROs).

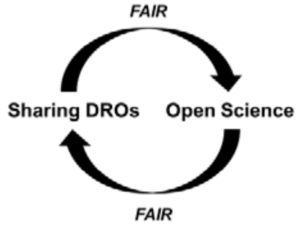


The Importance of Being FAIR

When DROs are findable, accessible, interoperable, re-usable, and attributable, they make possible a more data-centric and open paradigm of science and scholarship, where the products and processes of research can populate an ecosystem that allows for others than those who produced specific DROs to add value to the science and scholarship around them. The starting point for bringing DROs into this more open ecosystem is to share DROs, especially data. Benefits of sharing data and other DROs can include (depending how interoperable they are with other data and tools): providing a deeper understanding of the publications and ideas with which they are associated, gaining additional insight by reanalysis of the data, boosting statistical power to answer particular questions by aggregating multiple datasets, ability to apply big data analytic methods, broadening opportunities for collaboration, enhancing accountability (e.g., assessing reproducibility) and increasing return on research investments for research participants, science, and society.

Of course, there are also objections to such sharing, including: the costs incurred by making data and other DROs FAIR and sharing them, the possibility of others using the shared data to publish before the lab that produced the data does, concerns about intellectual property, patient privacy, and confidentiality, the fact that credit is accrued to investigators for papers published but not for data-related activities (so efforts not directed at publishing represent a net loss), the concern that data will not be understood and that the data will be misused. Most of these objections can be, and in some cases have already been, overcome (e.g., support of funders for data-sharing activities, use of embargo periods so sharing data happens after publication, and a host of policy and practice solutions to address intellectual property and patient privacy concerns). Yet, some, such as the lack of incentives for data-sharing, will require broad changes in the enterprise, while others still, such as the misuse of data, may never be fully resolved.

Conducting biomedical research in a more data-centric and open paradigm has repeatedly been shown to add significant value to science and scholarship. This was perhaps most famously demonstrated by the Human Genome Project, a 13-year project launched in 1990 that fundamentally changed the direction of biomedical research, and transformed our understanding of health and illness[xiv]. The spectacular success of the Human Genome Project was powered by collaborations and interactions of investigators in the ecosystem wrought of its findable, accessible, interoperable, re-usable data, tools, and infrastructure. Since then, many large-scale data-centric and open research initiatives have proven the value of these paradigms, including the Human Connectome Project and its subsequent related initiatives[xv], the NIH Human Microbiome Project[xvi], and the Genotype-Tissue Expression initiative[xvii]. And, the use of data-centric and open paradigms continues today as projects like All of Us[xviii] and the Adolescent Brain Cognitive Development Study[xix] get underway.



Data-centric & open paradigms have proven successful

From Concept-Centric and Closed to Data-Centric and Open

Despite many examples of data-centric and open approaches being used, most biomedical research is not conducted that way. For most research, the major public products are scientific papers reporting conclusions *about* the data, but the data themselves are almost never seen by others, much less shared with others. Thus, the currency of most biomedical research is not the data, but ideas and concepts about the data; it is concept-centric. And, since data are not made available to others, most research remains closed rather than open. This, however, is about to change. As society increasingly expects data from federally funded research to be broadly accessible, as computational and communication technologies become ever more powerful, as the scientific opportunities afforded by open and data-centric paradigms become more obvious, and as bipartisan policy directives from executive and legislative branches of the federal government encourage data sharing, it is likely that data-centric and open paradigms will soon be used beyond the confines of one-off initiatives.

An important policy directive was issued on February 22, 2013 by the Director of the White House Office of Science and Technology Policy (OSTP), Dr. John Holdren, wherein federal agencies with annual research and development budgets exceeding $100 million were directed to increase public access to the results of the research they conduct or support, including both the publications and the data underlying those publications[xx].

The National Institutes of Health issued its plan for meeting the OSTP directive in February 2015[xxi]. Regarding access to research publications, the NIH already had a policy in place, and NLM had already developed PubMed Central as the infrastructure to provide public access to them. Starting in 2008, NIH required "scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to PubMed Central immediately upon acceptance for publication"[xxii]. Since the OSTP directive, several other agencies across the federal government have opted to use PubMed Central as the infrastructure for publications supported by those agencies. The NIH plan for making research data more accessible in response to the OSTP directive and in the interest of better science, is described below.

What is NIH Doing to Make Digital Research Objects FAIR?

As data and other DROs become more broadly accessible, it is important that NIH encourage and facilitate these objects being findable, accessible, interoperable, and re-usable. There are existing and ongoing efforts at NIH that do or could support the FAIR principles; some of these are described, below.

Findable – PubMed is a powerful platform for discovery of the biomedical literature, with coverage from 1946 to the present, and more selectively before 1946. The full contents of some 5600 journals are indexed with a curated, hierarchically organized terminology (MeSH[xxiii]) that allows for sophisticated search and retrieval of citations. This infrastructure could be leveraged to make other DROs, such as datasets, findable, perhaps by building on MeSH in ways that would be well suited to categorize and find datasets, with a pointer to the locations of the datasets.

Some publication citations in PubMed already link to datasets[xxiv]. And, it is expected that within the next year, investigators submitting papers to PubMed Central will be able to also deposit in PubMed Central the datasets associated with those papers. Both mechanisms allow data to be findable via the literature.

Data repositories make their constituent data findable and NLM supports a portal with information about, and links to, some 70 data repositories that are supported by NIH[xxv], and that allow data egress and ingress. This portal can be used to identify data repositories containing data of interest, and the search mechanisms available for the respective repositories can be used to find specific datasets.

Looking forward, NLM is now exploring specifications that could be used to describe datasets with appropriate metadata, ideally in ways that would be widely applicable across much of the diverse data landscape of biomedical research.

Accessible – As was mentioned, the 2013 OSTP directive to increase public access to research results supported with federal funding included increasing access to both research publications and research data. Highlights of the NIH plan to make research data more accessible include that policies would apply to all NIH mechanisms of research support, including grants, contracts, and intramural projects, and would apply at all levels of support, regardless of the amount of budget.

Data management plans would be expected from all applications and proposals for research support, and would provide information such as the type and amounts of data expected to be generated or collected, the data-related standards to be used, how data might be made available to others, provisions for re-use, etc. The data management plans would be part of the review process, with the review of the plan being able to affect the merit score.

Peer review of the data management plans would allow plans to be reviewed on a project-by-project basis, with the expertise and norms of that particular research community brought to bear on the assessment. Peer review of data management plans, with that review affecting the overall score of the review, will also raise the salience of the plans with applicants and reviewers, encouraging an appropriate level of consideration being paid to them both parties.

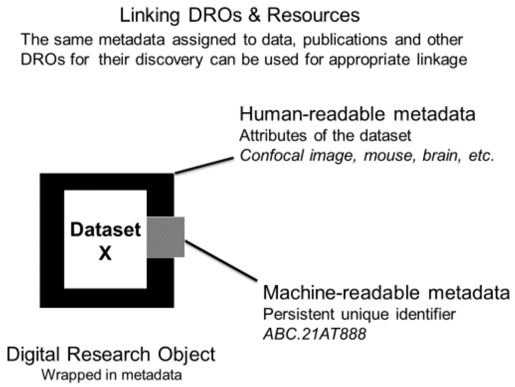
Interoperable – Data, and other DROs, are made interoperable with other DROs, tools and data resources through the use of standards. The NLM develops, supports the development of, and stewards widely-used standards, particularly for biomedical literature, healthcare information technology, and certain types of research data. These standards include[xxvi] terminologies, such as the Unified Medical Language System[xxvii] and SNOMED CT, coding systems like LOINC[xxviii], and metadata tagging specifications like JATS[xxix].

Across NIH, data repositories and major research initiatives supported by NIH, its institutes and centers, also specify data-related standards. Some of these, such as the Human Connectome Project, have incentivized investigators beyond that initiative[xxx] to adopt their data-related standards as their adoption allows investigators not supported by the initiative to rigorously compare their data with the initiative's data.

As the value of the use of common standards becomes more evident, institutes and centers of the NIH are increasingly communicating about and coordinating such efforts. For example, the NIH Clinical Common Data Elements Task Force maintains a conversation among all institutes and centers on this topic and is currently considering how to best harmonize related infrastructure, as well as developing and documenting best practices for standing up common data element initiatives. The NLM has also developed web resources on behalf of the Task Force, including a portal to NIH collections of common data elements and related resources[xxxi], and a repository for common data elements used in research conducted or supported by NIH[xxxii]. The repository allows users to search for specific common data elements, by topic, funding opportunity announcement, etc., as well as serving as a platform to compare and harmonize similar but distinct common data elements. Such harmonization mitigates the unnecessary proliferation of these types of standards.

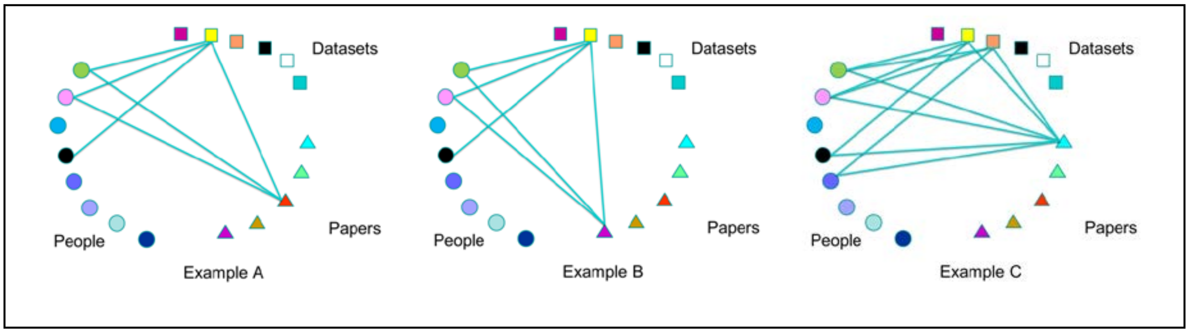
NIH is now launching pilot projects to create cloud instances as virtual spaces for data, analytic tools, repositories, and other DROs. Digital research objects that populate this NIH Data Commons will need to be compliant FAIR principles and certain standards[xxxiii], enhancing their interoperability.

Re-usable – Digital research objects are re-usable and most useful when they are linked to each other. Such linkage depends upon metadata of the DRO (metadata are data or information about DRO). For example, if the DRO is a research dataset, it might have both human-readable and machine-readable types of metadata. The human-readable metadata might be a set of descriptors such as "confocal image, mouse, brain" to reflect the instrument source of the data, and the organism and tissue type from which the data were collected. The machine-readable metadata might be a string of alpha numeric characters. Ideally, the identifier is unique so it resolves to the intended object, and persistent so that version of that object will not be lost over time, and its provenance tracked (i.e., changes to the object can be monitored by relating the identifier to identifiers of subsequent, modified, versions of it).


Linking DROs & Resources
The same metadata assigned to data, publications and other DROs for their discovery can be used for appropriate linkage
Human-readable metadata
Attributes of the dataset
Confocal image, mouse, brain, etc.
Dataset X
Machine-readable metadata
Persistent unique identifier
ABC.21AT888
Digital Research Object
Wrapped in metadata

Linkage of DROs forms the basis of a digital ecosystem because such linkage allows DROs to interact in an automated and dynamic way. A simple example of such an ecosystem is shown below, where each disc represents a particular person, each square a particular dataset, and each triangle a particular scientific publication. And, each of these objects is associated with other specific objects through linkage of their respective persistent unique identifiers, shown as lines connecting them.

In Example A, below, three individuals participated in activities resulting in a dataset (i.e., they are the data authors), and two of them were authors on the paper. In Example B, the same dataset produced by the same three data authors resulted in a second publication by the same two paper authors. In Example C, the dataset and data authors from examples A and B, as well as an additional dataset authored by a subset of data authors in previous example and a



new author, served as the basis of a scientific publication for which all authors of both datasets served as paper authors.

At this time, through some of NIH's research data repositories, bibliographic data platform (i.e., PubMed), administrative systems for grants, investigator's identifiers, and other DRO identifiers, simple linkage like that illustrated here is possible for certain sets of DROs (e.g., as does the National Database for Autism Research[xxxiv]).

Now, imagine an ecosystem where all DROs have persistent unique identifiers and are associated with (linked to) all DROs appropriately. So, in addition to identifiers of particular datasets, publications, investigators and their affiliations, also present and appropriately linked in this ecosystem are identifiers such as those for the specific instrument used to collect the data (e.g., the particular magnetic resonance imaging scanner), the specific data-related standards (e.g., NIfTI-1[xxxv] data format), the software used to statistically analyze the data (e.g., AFNI_17.2.05[xxxvi]), pre-registered experimental protocols[xxxvii], etc. For any given DRO (whether a dataset, paper, data author, paper author, software tool, etc.), such an ecosystem would allow a person - or a computer - to be aware of all of the other DROs directly linked to it. Of, course, this awareness of associations need not be limited to the first degree of association, but higher levels; analysis of such higher dimensional relationships across networks of DROs could provide interesting insights about the nature of the science, itself. Comprehensive awareness of associations in and by the ecosystem could be maintained by something like a blockchain approach[xxxviii],[xxxix], with the DRO links representing the transactions tracked. Such an approach could provide a way to characterize the DROs and maintain provenance at-scale in an open, distributed, and reliable manner, adding significant value to the ecosystem of science and scholarship.

Data Science and Open Science at NIH: Looking Forward

With the retirement of Dr. Donald A. B. Lindberg as the Director of NLM, the Director of NIH asked a working group of the Advisory Committee to the Director to examine the nexus of NLM's purview and expertise and the future of data science and open science in biomedicine[xl]. Later that year, the working group issued a report[xli] with six recommendations, all of which were adopted by the Director of NIH. One of these was that "NLM should be the intellectual and programmatic epicenter for data science at NIH" and another that "NLM should lead efforts to support and catalyze open science, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods." Soon thereafter, Dr. Patricia Flatley Brennan accepted the position as Director of NLM.

Since Dr. Brennan's arrival at NLM, input has been solicited and received from many, including NLM leadership and staff, leadership of NIH institutes and centers, external experts, and the general public through meetings, workshops, committee deliberations, town halls, and published requests for information (some of these activities have been undertaken as part of the decadal NLM strategic planning process). Informed by these ideas from diverse perspectives, a view has emerged of the key issues that need to be addressed as NLM assumes the leadership role for data science and open science at NIH.

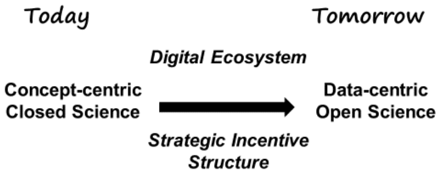
A clear and urgent priority for biomedical data science and open science is to engage with others across NIH to develop solutions for sustainability. As more large scale, large cohort studies, initiatives, and research programs are launched, the valuable data they produce must be housed, curated, and disseminated. Economies of scale and experience can be realized with a strategic enterprise approach, solving the same problem once rather than multiple times, converging on common standards, common architectures, coordination of acquisition activities around compute, and developing best practices for implementing and maintaining data assets and related infrastructure. Of course, it is important that trans-NIH approaches are flexible enough to meet the needs of particular studies, initiatives, and programs.

Another important contributor to sustainability is the use of evidence-based value assessment (e.g., cost/benefit analyses). Decisions such as those about: which data to keep, at what level particular datasets should be curated, how long specific datasets should be kept, which infrastructure should be invested in, which policies should be implemented and at what level of compliance, would all benefit by having an empirically-derived base of evidence for support. Such evidence could be used to develop criteria and heuristics for guidance about future investments in data, infrastructure, and policy.

Other priorities include the: 1) Strategic engagement beyond NIH, as data, science, and scholarship do not respect borders of nations, economic sectors or disciplines. 2) Development of a data-savvy workforce, including not only data scientists, *per se*, but data scientists cross-trained in biomedicine, biomedical scientist cross-trained in data science, both intramurally and extramurally. And, as data science and open science figure more prominently in NIH portfolios of extramurally-supported research, program officers, scientific review officers, and scientific policy staff of NIH must become more familiar with these areas. 3) Promotion of open science through changes in policies, engagement with the public around data and open science issues, and the development of tools designed specifically for use by public to facilitate their participation in research activities. 4) Research and innovation in data science and open science, developing new analytic approaches and tools, solutions to challenges of curation at-scale, and exploration of various flavors of artificial intelligence to harness the dynamic and expanding ecosystem of science and scholarship.

Finally, it is important to note that the behaviors of individuals and the practices of research-related organizations in a closed, concept-centric paradigm of science and scholarship are very different than those required for an open, data-centric paradigm. For example, in the former data are not shared, while the latter depends upon sharing data (and other DROs). The flip from the former paradigm to the latter will require changes in incentives to both people and organizations comprising the biomedical research enterprise (e.g., universities, funders, publishers, professional societies, regulatory agencies, etc.). Ideally, these incentives would be distributed across the entire enterprise and would be strategically aligned with each other to be mutually and maximally reinforcing, and avoiding unintended consequences. Due to wide variations in how poised various biomedical research domains are for adopting a data-centric and open paradigm (e.g., genomics already is largely thus; not so for epidemiology), such strategic incentive structures would likely be best designed and developed domain-by-domain, rather than across all areas of biomedicine at once.

In closing, the cumulative biomedical knowledgebase and breathtakingly powerful scientific technologies available today present significant opportunities to understand health and mitigate illness. A digital ecosystem wrought of data science and open science promises to multiply these opportunities many-fold. With the right incentives in place, this promise could be realized in the foreseeable future.



References

[i] Supported by the National Institutes of Health, National Library of Medicine

[ii] https://www.nlm.nih.gov/about/briefhistory.html

[iii] https://clinicaltrials.gov/

[iv] https://medlineplus.gov/

[v] https://toxnet.nlm.nih.gov/

[vi] https://nnlm.gov/

[vii] https://www.nlm.nih.gov/services/indexmedicus.html

[viii] https://www.nlm.nih.gov/pubs/factsheets/medline.html

[ix] https://www.ncbi.nlm.nih.gov/pubmed/

[x] https://www.ncbi.nlm.nih.gov/genbank/

[xi] https://wwwcf.nlm.nih.gov/nlm_eresources/eresources/search_database.cfm

[xii] Defined here as digital instantiations or representations of products and processes of research

[xiii] https://www.nature.com/articles/sdata201618

[xiv] https://www.nature.com/news/human-genome-project-twenty-five-years-of-big-biology-1.18436

[xv] http://www.humanconnectome.org/about-ccf

[xvi] https://commonfund.nih.gov/hmp

[xvii] https://commonfund.nih.gov/gtex

[xviii] https://allofus.nih.gov/

[xix] https://abcdstudy.org/about.html

[xx] https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

[xxi] https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

[xxii] https://publicaccess.nih.gov/

[xxiii] https://www.nlm.nih.gov/mesh/

[xxiv] https://www.ncbi.nlm.nih.gov/projects/linkout/

[xxv] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

[xxvi] https://www.nlm.nih.gov/healthit/snomedct/

[xxvii] https://www.nlm.nih.gov/research/umls/

[xviii] https://www.nlm.nih.gov/research/umls/loinc_main.html

[xix] https://jats.nlm.nih.gov/index.html

[xxx] http://www.nature.com/neuro/journal/v19/n9/box/nn.4361_BX2.html?foxtrotcallback=true

[xxxi] https://www.nlm.nih.gov/cde/

[xxxii] https://cde.nlm.nih.gov/home

[xxxiii] https://datascience.nih.gov/BlogFAIR

[xxxiv] https://ndar.nih.gov/

[xxv] https://afni.nimh.nih.gov/node/11

[xxxvi] *Ibid.*

[xxxvii] http://www.sciencemag.org/careers/2015/12/register-your-study-new-publication-option

[xxxviii] http://www.tandfonline.com/doi/full/10.1080/02763869.2017.1332261

[xxxix] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477

[xl] https://acd.od.nih.gov/working-groups/nlm.html

[xli] https://acd.od.nih.gov/documents/reports/Report-NLM-06112015-ACD.pdf

From Hospital Informatics Laboratories to National Data Networks: Positioning Academic Medical Centers to Advance Clinical Research

Lemuel R. Waitman, Professor, Department of Internal Medicine,
Associate Vice Chancellor for Enterprise Analytics
University of Kansas Medical Center

Since the 1960s and 1970s, pioneering academic medical centers (AMC) have been leaders in developing medical informatics systems to improve patient care and aggregate biomedical data to advance research. Since the HITECH Act in 2008 and the subsequent adoption of Electronic Health Records (EHR), the potential to aggregate biomedical data now extends beyond pioneering academic medical centers to all healthsystems. Led by the National Institutes of Health and the Patient Centered Outcomes Research Institute's creation of PCORnet, federal, nonprofit, and industry sponsors along with clinicians, patients, and investigators are seeking to capitalize on these new clinical data and link them to traditional billing and claims data sources. These institutions are creating local, regional and national data networks that can support prospective and observational research and realize the vision of a learning health system.

I. Developing National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA) Capacity for Biomedical and Informatics Research

As described previously [Merrill Waitman, Lushington, Warren], academic medical center's pursuit of National Institutes of Health Clinical and Translational Science Awards (CTSA) [Zerhouni] catalyzed the development and integration of informatics capabilities to support clinical and translational research. The 2010 proposal for University of Kansas Medical Center's CTSA program, Frontiers, provides an example of a regional vision for biomedical informatics as illustrated in Figure 1. While the precise steps and integration varied from this plan over the five years since award (2011-2016), the program largely succeeded in achieving these complementary aims. For our central aim, creating the HERON clinical integrated data repository: business agreements, and oversight processes were successfully established between the university and health system leadership, the open source i2b2 software was implemented [Murphy], data was increasingly mapped to national standards aligned with meaningful use standards available from the National Library of Medicine, and over 2 billion facts for 2 million patients were integrated. Over 800 access and data requests from faculty have been approved and investigators have executed over 50,000 queries. Notably, HERON provided a platform for integrating electronic health records with existing national registries (e.g. NAACCR hospital tumor registry, national cardiovascular research database CathPCI, trauma, and cystic fibrosis) and organizational benchmarking activities (e.g. Visient University Health System Consortium).

In addition to i2b2 for data integration and warehousing, Frontiers biomedical informatics adopted and promoted REDCap (https://projectredcap.org/) [Harris] as a common tool for research data capture across the enterprise and our partner institutions: streamlining access to all with a KUMC campus login. Adoption has been dramatic: with over 2,500 data collection projects in production for over 4,000 users. REDCap use has also extended beyond traditional clinical trial electronic case report forms to also support registries and administrative needs across the medical center and our health system partners; increasing awareness of the CTSA capabilities for the campus and adoption of REDCap at other campuses in the Kansas City region.

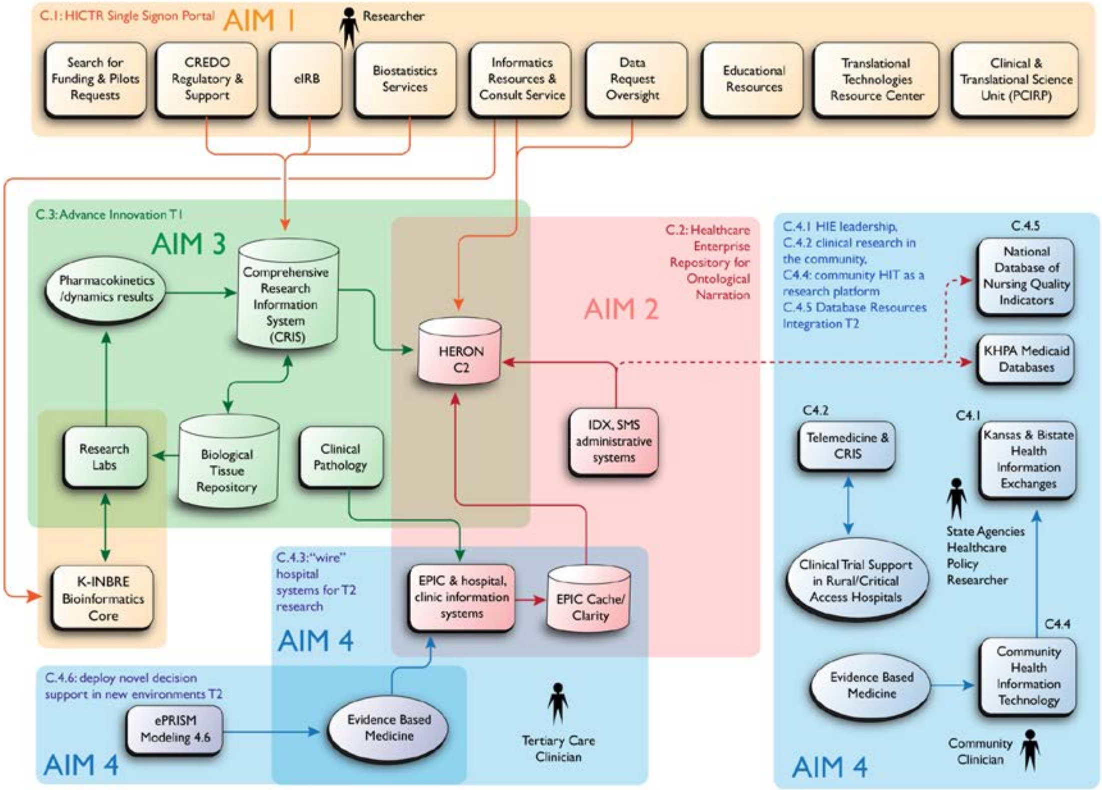

C.1: HICTR Single Signon Portal
AIM 1
Researcher
Search for Funding & Pilots Requests
CREDO Regulatory & Support
eIRB
Biostatistics Services
Informatics Resources & Consult Service
Data Request Oversight
Educational Resources
Translational Technologies Resource Center
Clinical & Translational Science Unit (PCIRP)
C.3: Advance Innovation T1
AIM 3
Pharmacokinetics /dynamics results
Comprehensive Research Information System (CRIS)
HERON C2
C.2: Healthcare Enterprise Repository for Ontological Narration
AIM 2
C.4.1 HIE leadership,
C.4.2 clinical research in the community,
C4.4: community HIT as a research platform
C.4.5 Database Resources Integration T2
C.4.5
National Database of Nursing Quality Indicators
KHPA Medicaid Databases
Research Labs
Biological Tissue Repository
Clinical Pathology
IDX, SMS administrative systems
C4.2
Telemedicine & CRIS
C4.1
Kansas & Bistate Health Information Exchanges
State Agencies Healthcare Policy Researcher
Clinical Trial Support in Rural/Critical Access Hospitals
K-INBRE Bioinformatics Core
C.4.3: "wire" hospital systems for T2 research
AIM 4
EPIC & hospital, clinic information systems
EPIC Cache/ Clarity
C4.4
Community Health Information Technology
Evidence Based Medicine
C.4.6: deploy novel decision support in new environments T2
ePRISM Modeling 4.6
Evidence Based Medicine
Tertiary Care Clinician
AIM 4
Community Clinician
AIM 4

Figure 1. Conceptual Model of Clinical and Translational Science Award Biomedical Informatics and Specific Aims

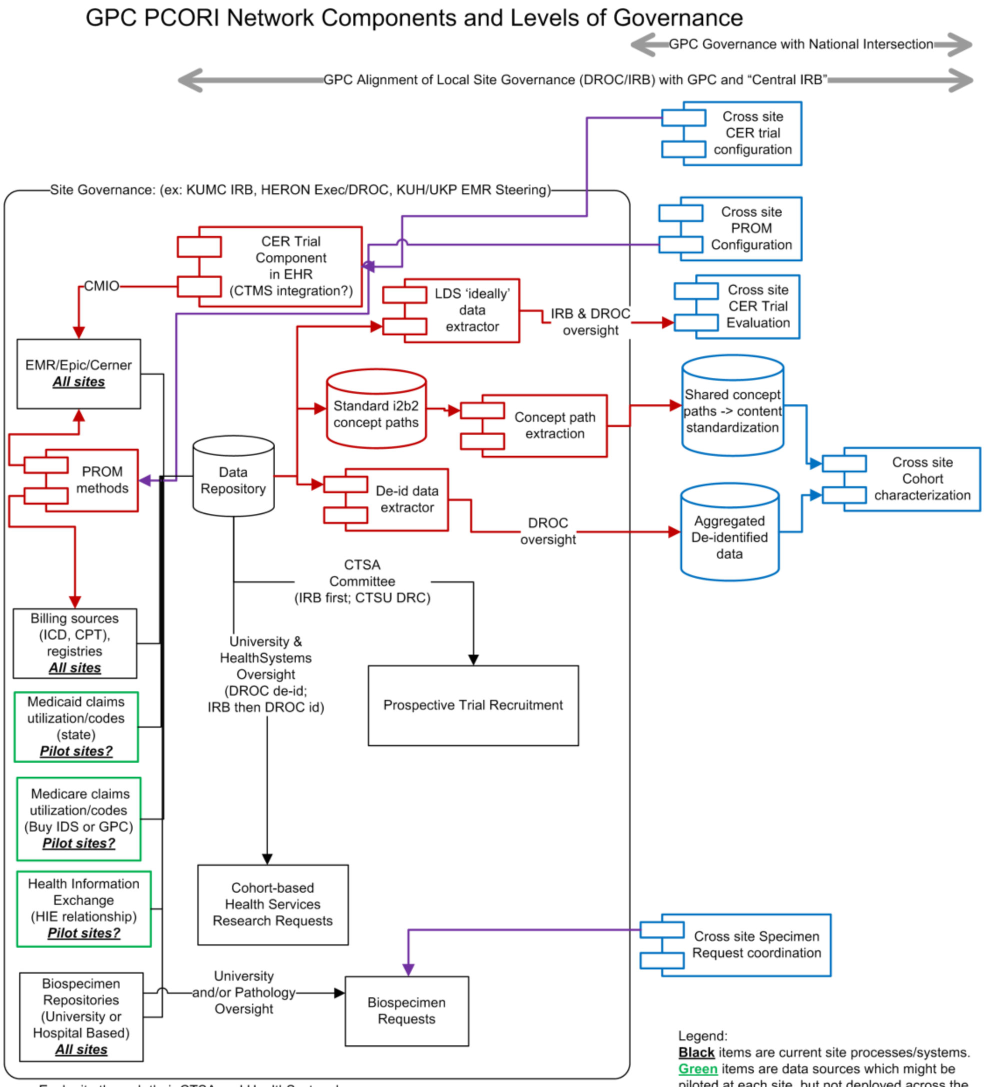
II. CTSA Informatics Infrastructure's Potential for National Interoperable Data Research

Frontiers biomedical informatics' choice of i2b2 and REDCap was fortuitous for supporting broader collaboration nationally. Our campus' efforts to make HERON a strong example of extending i2b2 at enterprise scale for a campus and its heightened integration with REDCap was of interest to other academic medical center CTSA programs. Often i2b2 was implemented for specific informatics projects or to only provide feasibility assessment. These limited scope or reduced functionality implementations of i2b2 would often hinder adoption by the broader research community at a medical center. By 2013, HERON was seen by the broader community as a successful example for data integration which was coincidental with the announcement by the Patient Centered Research Institute (PCORI) that they were creating PCORnet, the National Patient-Centered Clinical Research Network [Fleurence]. This effort was supported by several funding announcements in Spring 2013 that would support the creation of Patient Powered Research Networks, Clinical Data Research Networks, and a Coordinating Center for the national network.

PCORI and its stakeholders from the National Institutes of Health (NIH), the Food and Drug Administration (FDA), patient organizations and insurance plans had a vision to improve our nation's capacity to conduct clinical research. Current state was seen as an environment where a high percentage of decisions made in clinical practice were not supported by the best evidence. Patients health outcomes were not improving and disparities in outcomes were also either stagnant or widening. The machinery of developing prospective or retrospective clinical research studies was slow and expensive or unreliable at creating reproducible research. As a result, clinical research was also unattractive to hospital and health system administrators who often didn't see how this complex additional activity benefited patient care at their institutions. As PCORI's mission is patient centered and more oriented towards pragmatic research (relative to basic science and early translational research support by the NIH) there was a strong emphasis of generating evidence to support daily decision making for doctors, patients, and their families [Tricoci P et al.] PCORI, it's stakeholders of funder and researchers saw that existing networks, recently enabled by the adoption of electronic health records, might provide a platform for conducting research more effectively but also bring together patients, providers and scientists to work as a connected community. PCORnet's goal was "to improve the nation's capacity to conduct clinical research by creating a large, highly representative, national patient-centered network that supports more efficient clinical trials and observational studies."

Frontiers biomedical informatics saw high alignment with its work for integrating data in support of our CTSA program and the PCORI funding opportunity to create a Clinical Data Research Network (CDRN) which required the ability to incorporate at least two health systems with over 1 million patients' data and have rich physician and patient engagement. Networks needed to demonstrate governance and the ability to collect patient reported outcomes as well as embed comparative effectiveness trials in the clinical workflow. Some existing data networks were poised to respond while KUMC and related CTSA programs didn't have existing data networks in place. Frontiers worked with other institutions in the Midwest to organize a response and create the Greater Plains Collaborative (GPC) [Waitman, Aaronson] and successfully competed for an initial

Phase 1 CDRN contract (http://frontiersresearch.org/frontiers/sites/default/files/frontiers/documents/GPC-PCORI-CDRN-Research-Plan-Template-KUMCv44.pdf ; awarded in 2014 for 18 months) and the subsequent Phase 2 contract (http://frontiersresearch.org/frontiers/sites/default/files/Phase%20II%20Proposal.pdf ; awarded in 2015 for 3 years). The Greater Plains Collaborative initially included 10 partner institutions covering an estimated 11.8 million lives, 13 hospitals, 430 clinics, 1800 primary care providers, and 7600 specialists. Each network had to develop their ability to characterize obese patient populations but had flexibility in choosing a common and rare disease focus. GPC worked with community stakeholders to identify breast cancer as its common condition and guided by CTSA leadership choose Amyotrophic Lateral Sclerosis (ALS or Lou Gerhig's disease) as its rare condition. GPC adopted i2b2 and REDCap as common technologies based on their increased adoption across CTSA programs or related programs at other institutions. The Greater Plains Collaborative high level architecture and governance is shown in Figure 2. The initial 10 GPC institutions included: University of Texas Health Sciences Center at San Antonio, University of Texas Southwestern, University of Kansas Medical Center, Children's Mercy Hospital, University of Nebraska Medical Center, University of Iowa, University of Wisconsin, Medical College of Wisconsin, University of Minnesota, and the Marshfield Clinic. In the phase two proposal the Greater Plains Collaborative expanded to include the University of Missouri and Indiana University (http://www.gpcnetwork.org/?q=AboutUs).


GPC PCORI Network Components and Levels of Governance
GPC Governance with National Intersection
GPC Alignment of Local Site Governance (DROC/IRB) with GPC and "Central IRB"
Cross site CER trial configuration
Site Governance: (ex: KUMC IRB, HERON Exec/DROC, KUH/UKP EMR Steering)
Cross site PROM Configuration
CER Trial Component in EHR (CTMS integration?)
CMIO
LDS 'ideally' data extractor
IRB & DROC oversight
Cross site CER Trial Evaluation
EMR/Epic/Cerner
All sites
Standard i2b2 concept paths
Concept path extraction
Shared concept paths -> content standardization
PROM methods
Data Repository
De-id data extractor
Cross site Cohort characterization
DROC oversight
Aggregated De-identified data
CTSA Committee (IRB first; CTSU DRC)
Billing sources (ICD, CPT), registries
All sites
University & HealthSystems Oversight (DROC de-id; IRB then DROC id)
Prospective Trial Recruitment
Medicaid claims utilization/codes (state)
Pilot sites?
Medicare claims utilization/codes (Buy IDS or GPC)
Pilot sites?
Cohort-based Health Services Research Requests
Health Information Exchange (HIE relationship)
Pilot sites?
Cross site Specimen Request coordination
Biospecimen Repositories (University or Hospital Based)
All sites
University and/or Pathology Oversight
Biospecimen Requests

Legend:
Black items are current site processes/systems.
Green items are data sources which might be piloted at each site, but not deployed across the network.
Red items are new components deployed at each site across the network
Blue items are components deployed centrally
Purple lines show the feedback processes to configure sites for PROM, CER, and coordinating amongst biospecimen repositories

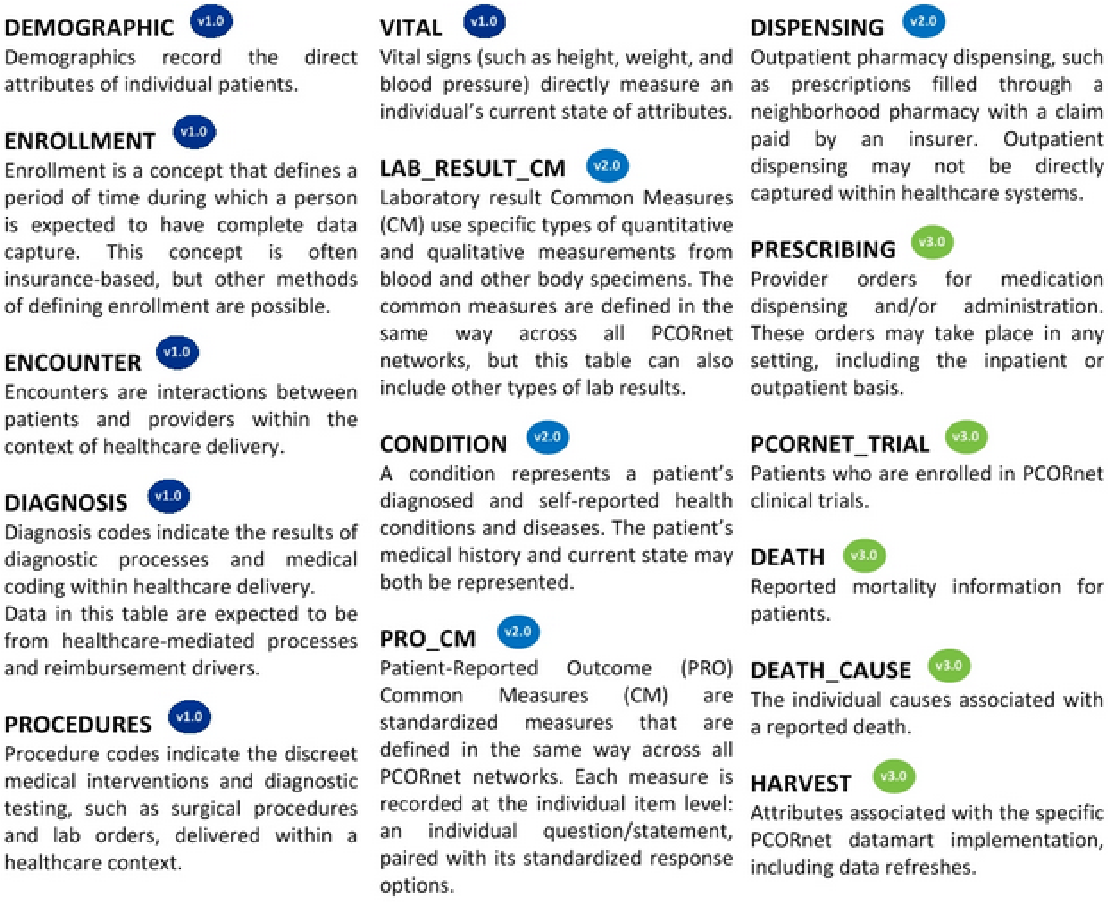
Each site through their CTSA and HealthSystem has current approaches to governance and oversight. This typically involves:
A) IRB protocol and oversight of the honest broker protocol governing the data repository
B) IRB oversight for identified data requests
C) IRB oversight for prospective trials
D) HealthSystem oversight for data requests (de-id and identified)
E) CTSA oversight for contacting prospective trial patients (concerns of overlap)
F) HealthSystem CMIO/CIO oversight of EMR modifications and components

Author: Russ Waitman
August 18, 2013

Figure 2. Greater Plains Collaborative Architecture and Governance

III. Managing data expectations amongst differing data constituents: Clinical Researchers ("their data") and Epidemiologists ("all the data")

In addition to meeting our contractual milestones outlined in our proposals for becoming a viable Clinical Data Research Network, all network partners were instructed to also shift effort to support adoption and implementation of a new data infrastructure: the PCORnet Common Data Model (CDM) that was to be based on a data model created for the FDA by a portion of the PCORnet coordinating center at Harvard Pilgrim's Insurance (the Mini-Sentinel Common Data Model). Mini-Sentinel was created to support adverse drug event surveillance using insurance claims administrative data: classically diagnoses, procedures, and hospitalizations across insured populations who were "covered" by an insurance plan during well defined enrollment periods. The PCORnet Common Data Model version 3.0 is shown in Figure 3.



DEMOGRAPHIC v1.0
Demographics record the direct attributes of individual patients.

ENROLLMENT v1.0
Enrollment is a concept that defines a period of time during which a person is expected to have complete data capture. This concept is often insurance-based, but other methods of defining enrollment are possible.

ENCOUNTER v1.0
Encounters are interactions between patients and providers within the context of healthcare delivery.

DIAGNOSIS v1.0
Diagnosis codes indicate the results of diagnostic processes and medical coding within healthcare delivery. Data in this table are expected to be from healthcare-mediated processes and reimbursement drivers.

PROCEDURES v1.0
Procedure codes indicate the discreet medical interventions and diagnostic testing, such as surgical procedures and lab orders, delivered within a healthcare context.

VITAL v1.0
Vital signs (such as height, weight, and blood pressure) directly measure an individual's current state of attributes.

LAB_RESULT_CM v2.0
Laboratory result Common Measures (CM) use specific types of quantitative and qualitative measurements from blood and other body specimens. The common measures are defined in the same way across all PCORnet networks, but this table can also include other types of lab results.

CONDITION v2.0
A condition represents a patient's diagnosed and self-reported health conditions and diseases. The patient's medical history and current state may both be represented.

PRO_CM v2.0
Patient-Reported Outcome (PRO) Common Measures (CM) are standardized measures that are defined in the same way across all PCORnet networks. Each measure is recorded at the individual item level: an individual question/statement, paired with its standardized response options.

DISPENSING v2.0
Outpatient pharmacy dispensing, such as prescriptions filled through a neighborhood pharmacy with a claim paid by an insurer. Outpatient dispensing may not be directly captured within healthcare systems.

PRESCRIBING v3.0
Provider orders for medication dispensing and/or administration. These orders may take place in any setting, including the inpatient or outpatient basis.

PCORNET_TRIAL v3.0
Patients who are enrolled in PCORnet clinical trials.

DEATH v3.0
Reported mortality information for patients.

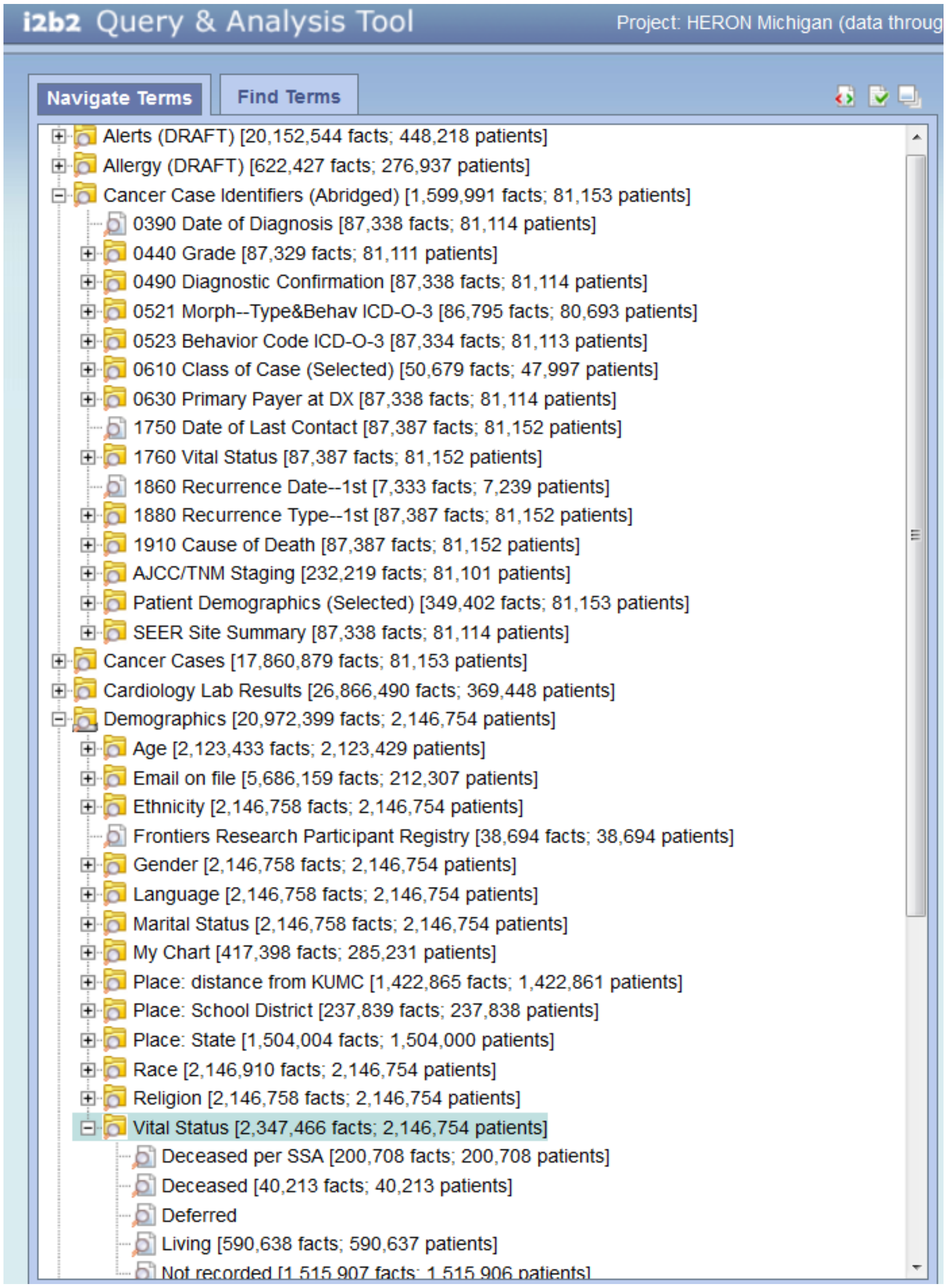
DEATH_CAUSE v3.0
The individual causes associated with a reported death.

HARVEST v3.0
Attributes associated with the specific PCORnet datamart implementation, including data refreshes.

Figure 3. PCORnet Common Data Model version 3.0

As we worked to develop the network and focus on our common condition, breast cancer, the differences in perspective between the epidemiology focused coordinating center data modeling team and those embedded in health systems with rich clinical research goals became apparent. Clinical researchers often are intimately familiar with unique data available to their profession and disease area. For example, Figure 4 illustrates how incorporating standardized hospital tumor registry tables defined by the North American Association of Central Cancer Registries (NAACCR) can act as a standard against which one can compare clinical and billing information from electronic and administrative systems at a health system. Figure 4 also illustrates HERON's ability to link to EHR patient portal usage (MyChart) and incorporate the social security administration death master file (SSDMF) so clinical researchers may exclude patients who have died outside their health system from being contacted for clinical trials or use SSDMF status for outcomes research. This leads to a gross observation that trialists and clinical researchers want access to "their data". They are familiar with relevant clinical workflows and registries unique to their profession and would like it incorporated in a manner similar to how they are used to seeing this data. Since the majority of their research is at a single site, they are less concerned with aligning data to national standards which may actually hinder their interpretability.

In contrast, PCORnet's data coordination was centered at Harvard Pilgrims which is epidemiology focused. Upon reflection, that was clear from the initial funding announcement which called for participants to: 1) create research-ready datasets that included comprehensive data from EHRs to describe patients' care experience over time and in different care settings. 2) CDRNs were to utilize multiple rich data sources to support research, such as electronic health records, insurance claims data, and data reported directly by patients. This called for CDRNs to establish relationships with external data partners (Centers for Medicare and Medicaid, State, private insurers).


i2b2 Query & Analysis Tool
Project: HERON Michigan (data throug
Navigate Terms
Find Terms

- Alerts (DRAFT) [20,152,544 facts; 448,218 patients]
- Allergy (DRAFT) [622,427 facts; 276,937 patients]
- Cancer Case Identifiers (Abridged) [1,599,991 facts; 81,153 patients]
  - 0390 Date of Diagnosis [87,338 facts; 81,114 patients]
  - 0440 Grade [87,329 facts; 81,111 patients]
  - 0490 Diagnostic Confirmation [87,338 facts; 81,114 patients]
  - 0521 Morph--Type&Behav ICD-O-3 [86,795 facts; 80,693 patients]
  - 0523 Behavior Code ICD-O-3 [87,334 facts; 81,113 patients]
  - 0610 Class of Case (Selected) [50,679 facts; 47,997 patients]
  - 0630 Primary Payer at DX [87,338 facts; 81,114 patients]
  - 1750 Date of Last Contact [87,387 facts; 81,152 patients]
  - 1760 Vital Status [87,387 facts; 81,152 patients]
  - 1860 Recurrence Date--1st [7,333 facts; 7,239 patients]
  - 1880 Recurrence Type--1st [87,387 facts; 81,152 patients]
  - 1910 Cause of Death [87,387 facts; 81,152 patients]
  - AJCC/TNM Staging [232,219 facts; 81,101 patients]
  - Patient Demographics (Selected) [349,402 facts; 81,153 patients]
  - SEER Site Summary [87,338 facts; 81,114 patients]
- Cancer Cases [17,860,879 facts; 81,153 patients]
- Cardiology Lab Results [26,866,490 facts; 369,448 patients]
- Demographics [20,972,399 facts; 2,146,754 patients]
  - Age [2,123,433 facts; 2,123,429 patients]
  - Email on file [5,686,159 facts; 212,307 patients]
  - Ethnicity [2,146,758 facts; 2,146,754 patients]
  - Frontiers Research Participant Registry [38,694 facts; 38,694 patients]
  - Gender [2,146,758 facts; 2,146,754 patients]
  - Language [2,146,758 facts; 2,146,754 patients]
  - Marital Status [2,146,758 facts; 2,146,754 patients]
  - My Chart [417,398 facts; 285,231 patients]
  - Place: distance from KUMC [1,422,865 facts; 1,422,861 patients]
  - Place: School District [237,839 facts; 237,838 patients]
  - Place: State [1,504,004 facts; 1,504,000 patients]
  - Race [2,146,910 facts; 2,146,754 patients]
  - Religion [2,146,758 facts; 2,146,754 patients]
  - Vital Status [2,347,466 facts; 2,146,754 patients]
    - Deceased per SSA [200,708 facts; 200,708 patients]
    - Deceased [40,213 facts; 40,213 patients]
    - Deferred
    - Living [590,638 facts; 590,637 patients]
    - Not recorded [1,515,907 facts; 1,515,906 patients]

Figure 4: North American Association of Central Cancer Registries (NAACCR) hospital tumor registry data incorporated within the KUMC HERON i2b2 data warehouse.

Figure 5 from the Phase 1 GPC proposal provides a visualization of PCORnet's goal of complete and comprehensive data by revealing gaps in data for a typical academic medical center's cancer center. Data after breast cancer diagnosis is reasonably complete during treatment but is often missing for common data elements prior to diagnosis (e.g. vital signs, common labs) since the medical center predominantly provides specialty care.

A year after diagnosis, we may see similar decline as the patient's primary care is provided outside the academic medical center.

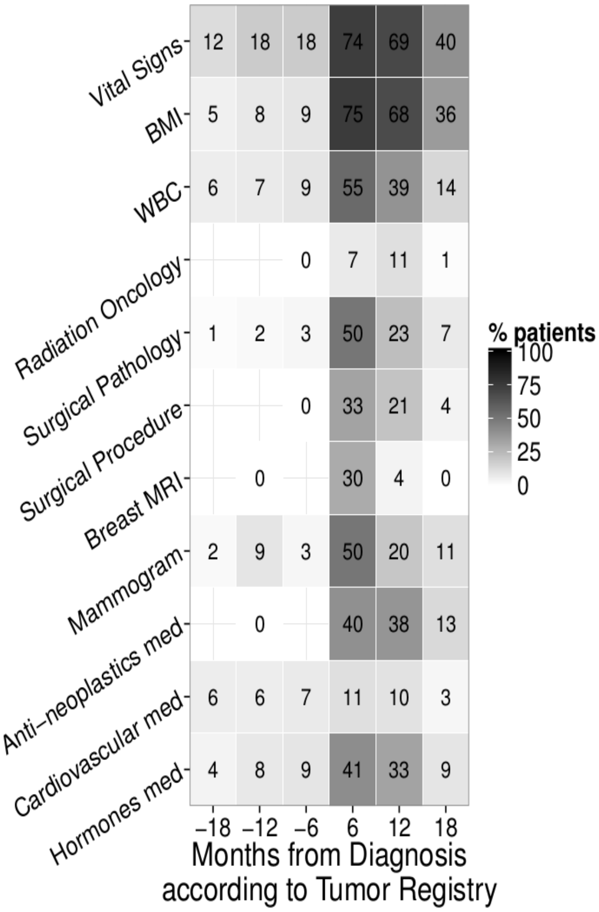
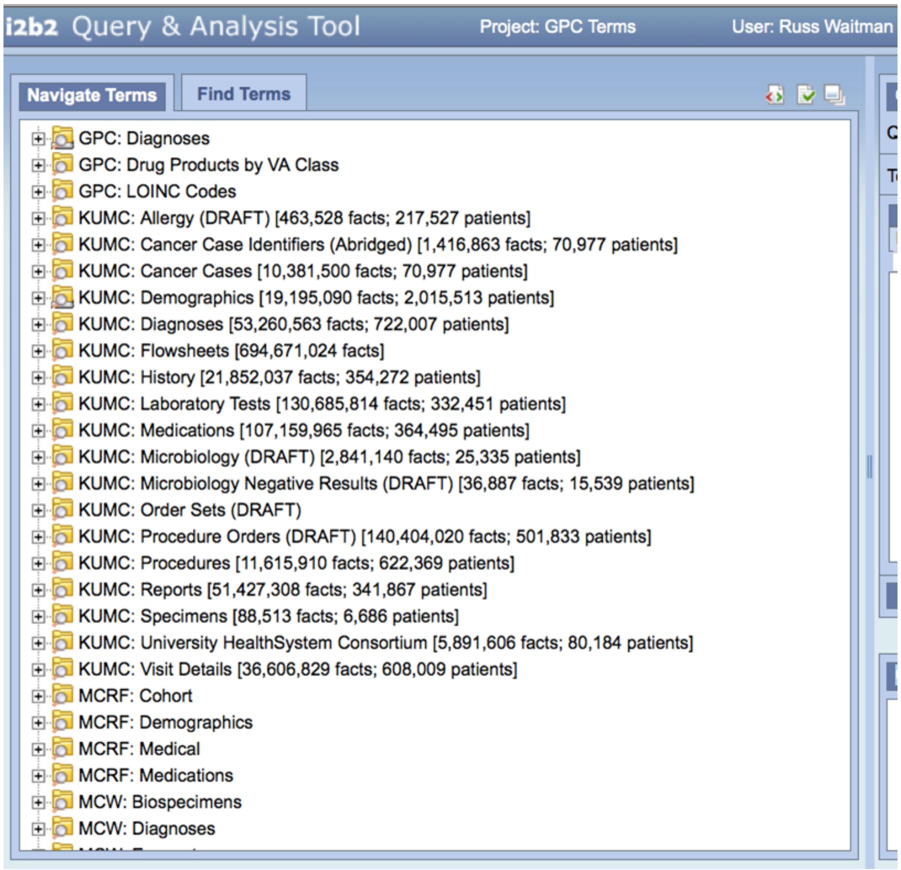


Figure 5. Comprehensive and complete data example from the University of Kansas Cancer Center: heat map of percentage of proposed data elements from the EHR and billing sources recorded in six month intervals surrounding the data of breast cancer diagnosis specified by the hospital tumor registry.

This highlights a challenge for network infrastructure teams: clinical researchers seeking data specific to their professions specific needs and workflow while epidemiologists seeking transforming data out of specific workflows and registries into a unified common data model that harmonizes "all the data". Medical centers struggle as they participate in multiple national initiatives with how to manage both kinds of customers (predominately local clinical researchers versus national network epidemiologic driven) and potentially conflicting national common data models required for participation in different efforts (NIH CTSA, All of Us for the National Cancer Institute, PCORnet, Mini-Sentinel for FDA, etc).

GPC has chosen various strategies to promote data exchange and compare terminology alignment across partners. Sites use i2b2 to incorporate rich clinical data sources such as cancer registries and share their terminologies using a centralized website (https://babel.gpcnetwork.org) shown in Figure 6 but also using software developed in partnership with Harvard University and the SCILHS network (http://scilhs.org/) to transform data from i2b2 into the PCORnet Common Data Model (https://github.com/kumc-bmi/i2p-transform). Sites align their data in i2b2 to use consistent terminologies for common variable such as diagnoses, procedures, demographics and laboratory results. This allows the network to support local and regional investigators who can directly use i2b2 to determine study feasibility and KUMC developed software to extract data from i2b2 (Data Builder; https://informatics.gpcnetwork.org/trac/Project/wiki/DataBuilder) but also have their data quality checked as it's transformed into the PCORnet CDM. The PCORnet CDM in turn allows the medical centers in the GPC to participate in the national research initiatives and also supports GPC level investigators developing studies that will leverage the PCORnet CDM.



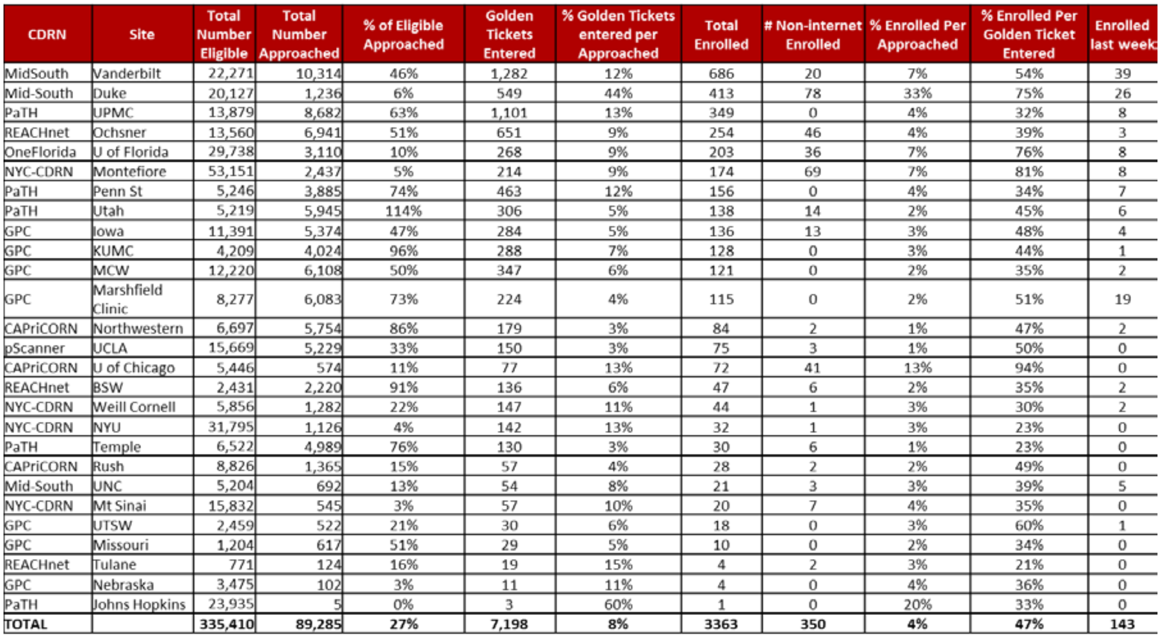
GPC: Diagnoses
GPC: Drug Products by VA Class
GPC: LOINC Codes
KUMC: Allergy (DRAFT) [463,528 facts; 217,527 patients]
KUMC: Cancer Case Identifiers (Abridged) [1,416,863 facts; 70,977 patients]
KUMC: Cancer Cases [10,381,500 facts; 70,977 patients]
KUMC: Demographics [19,195,090 facts; 2,015,513 patients]
KUMC: Diagnoses [53,260,563 facts; 722,007 patients]
KUMC: Flowsheets [694,671,024 facts]
KUMC: History [21,852,037 facts; 354,272 patients]
KUMC: Laboratory Tests [130,685,814 facts; 332,451 patients]
KUMC: Medications [107,159,965 facts; 364,495 patients]
KUMC: Microbiology (DRAFT) [2,841,140 facts; 25,335 patients]
KUMC: Microbiology Negative Results (DRAFT) [36,887 facts; 15,539 patients]
KUMC: Order Sets (DRAFT)
KUMC: Procedure Orders (DRAFT) [140,404,020 facts; 501,833 patients]
KUMC: Procedures [11,615,910 facts; 622,369 patients]
KUMC: Reports [51,427,308 facts; 341,867 patients]
KUMC: Specimens [88,513 facts; 6,686 patients]
KUMC: University HealthSystem Consortium [5,891,606 facts; 80,184 patients]
KUMC: Visit Details [36,606,829 facts; 608,009 patients]
MCRF: Cohort
MCRF: Demographics
MCRF: Medical
MCRF: Medications
MCW: Biospecimens
MCW: Diagnoses

Figure 6. Greater Plains Collaborative Babel terminology service illustrating terminology types from University of Kansas, Marshfield Clinic, and the Medical College of Wisconsin.

IV. From Infrastructure Building to National Studies: prospective interventional and observational PCORnet demonstration projects

While much of PCORnet's activity was establishing governance and data infrastructure, the network also collaboratively prioritized and devised three national demonstration projects: the prospective ADAPTABLE pragmatic trial and two observational studies regarding obesity.

Aspirin Dosing: A Patient-Centric Trial Assessing Benefits and Long-term Effectiveness (ADAPTABLE) (http://theaspirinstudy.org/) is PCORnet's first pragmatic clinical trial. It is not only important clinically but provides the richest test of a medical center's willingness and capability to conduct trials in a novel, more efficient manner. ADAPTABLE seeks to compare the effectiveness and safety of two doses of aspirin (81 mg and 325 mg) in 20,000 high-risk patients with atherosclerotic cardiovascular disease (ASCVD). It's cost per enrolled patient is an order of magnitude lower than traditional trials (ADAPTABLE = $850; 3 Simple, NIH Pragmatic Trials: $2,260 to $13,269; Industry Trials: $8,500). It's key innovation is to determine eligible patients by screening electronic health records for defined by a computable phenotype and then approach them via predominantly high volume, low cost channels (email, EHR patient portals, and physical mailers). Patients then visit the ADAPTABLE website and enter their "Golden Ticket" provided by the approach email/letter. After reviewing the trial and online consent videos, the patients are consented and enrolled, typically at their convenience in their home. ADAPTABLE includes researchers from 8 PCORnet CDRNs and 35 health systems along with an "Adaptor Team composed of 8 patients representing each CDRN supported by the Health eHeart Alliance PPRN. While enrollment has now exceed 5000 patients, Figure 7 provides a snapshot of enrollment yield rates across sites in June 2017.



CDRN	Site	Total Number Eligible	Total Number Approached	% of Eligible Approached	Golden Tickets Entered	% Golden Tickets entered per Approached	Total Enrolled	# Non-internet Enrolled	% Enrolled Per Approached	% Enrolled Per Golden Ticket Entered	Enrolled last week
MidSouth	Vanderbilt	22,271	10,314	46%	1,282	12%	686	20	7%	54%	39
Mid-South	Duke	20,127	1,236	6%	549	44%	413	78	33%	75%	26
PaTH	UPMC	13,879	8,682	63%	1,101	13%	349	0	4%	32%	8
REACHnet	Ochsner	13,560	6,941	51%	651	9%	254	46	4%	39%	3
OneFlorida	U of Florida	29,738	3,110	10%	268	9%	203	36	7%	76%	8
NYC-CDRN	Montefiore	53,151	2,437	5%	214	9%	174	69	7%	81%	8
PaTH	Penn St	5,246	3,885	74%	463	12%	156	0	4%	34%	7
PaTH	Utah	5,219	5,945	114%	306	5%	138	14	2%	45%	6
GPC	Iowa	11,391	5,374	47%	284	5%	136	13	3%	48%	4
GPC	KUMC	4,209	4,024	96%	288	7%	128	0	3%	44%	1
GPC	MCW	12,220	6,108	50%	347	6%	121	0	2%	35%	2
GPC	Marshfield Clinic	8,277	6,083	73%	224	4%	115	0	2%	51%	19
CAPriCORN	Northwestern	6,697	5,754	86%	179	3%	84	2	1%	47%	2
pScanner	UCLA	15,669	5,229	33%	150	3%	75	3	1%	50%	0
CAPriCORN	U of Chicago	5,446	574	11%	77	13%	72	41	13%	94%	0
REACHnet	BSW	2,431	2,220	91%	136	6%	47	6	2%	35%	2
NYC-CDRN	Weill Cornell	5,856	1,282	22%	147	11%	44	1	3%	30%	2
NYC-CDRN	NYU	31,795	1,126	4%	142	13%	32	1	3%	23%	0
PaTH	Temple	6,522	4,989	76%	130	3%	30	6	1%	23%	0
CAPriCORN	Rush	8,826	1,365	15%	57	4%	28	2	2%	49%	0
Mid-South	UNC	5,204	692	13%	54	8%	21	3	3%	39%	5
NYC-CDRN	Mt Sinai	15,832	545	3%	57	10%	20	7	4%	35%	0
GPC	UTSW	2,459	522	21%	30	6%	18	0	3%	60%	1
GPC	Missouri	1,204	617	51%	29	5%	10	0	2%	34%	0
REACHnet	Tulane	771	124	16%	19	15%	4	2	3%	21%	0
GPC	Nebraska	3,475	102	3%	11	11%	4	0	4%	36%	0
PaTH	Johns Hopkins	23,935	5	0%	3	60%	1	0	20%	33%	0
TOTAL		**335,410**	**89,285**	**27%**	**7,198**	**8%**	**3363**	**350**	**4%**	**47%**	**143**

Figure 7. PCORnet's ADAPTABLE trial enrollment and recruitment yield rates across sites circa June 2017.

PCORnet also undertook two observational studies addressing controversial subjects with the largest sample sizes to date. The bariatric surgery study (largely adult patients with seven GPC sites participating) includes 48 institutions, 11 CDRNs, 3 PPRNs, and 65,000 people (1,000 of who are adolescents) and studies which surgical approach is best for treating severe obesity between Roux-en-y gastric bypass, Adjustable gastric banding, or Sleeve gastrectomy shown in Figure 8.

This study focuses on one, three and five year outcome that matter to obese patients: weight loss, improvement in diabetes, and risk of adverse events.

The pediatric obesity survey studies whether antibiotics given to children increase risk for obesity and includes 10 Participating CDRNs, 4 Participating PPRNs, 41 Institutions, and 650,000 children. Its main effect analyses evaluates antibiotics use during the first 24 months and weight outcomes at five and ten years of age.

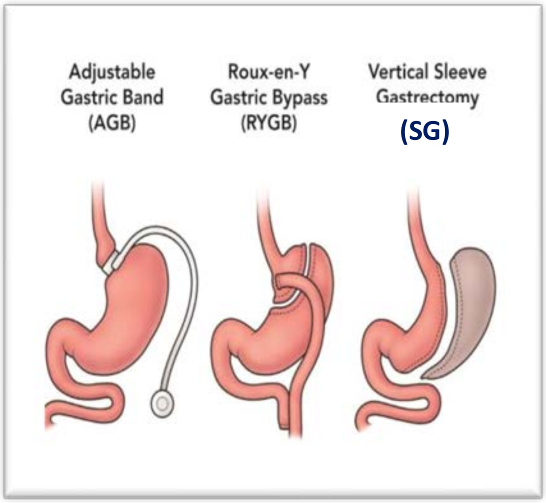

Adjustable Gastric Band (AGB)
Roux-en-Y Gastric Bypass (RYGB)
Vertical Sleeve Gastrectomy (SG)

Figure 8. PCORnet bariatric study surgical approaches (disseminated by the PCORnet coordinating center).

V. How does a medical center (KUMC) and the GPC (peer AMCs) fit into the evolving national landscape?

As the University of Kansas Medical Center and our peers in the Greater Plains Collaborative complete four years of building PCORnet, we reflect upon how our participation has impacted our campuses. The majority of our campuses are involved in all three demonstration projects. Our network, led by Dr. Elizabeth Chrischilles at Iowa, leads the national collaborative research group for advancing PCORnet's cancer research and Dr. Russ Waitman has served as the national chair for the PCORnet data committee. GPC sites have been responsive to national common data model queries issued from the national coordinating center to each participating site and all GPC sites adopted the SmartIRB national reliance model before any other network. The GPC strategy to merge Medicare and Medicaid Claims was successful and has created a centralized claims repository for integrating EHR and CDM data (https://informatics.gpcnetwork.org/trac/Project/wiki/GROUSE). While the national structure of PCORnet is evolving, in many ways the GPC has served as a data fitness camp for academic medical centers to participate in national data intensive research. PCORI announced in 2017 that it would transition infrastructure support for PCORnet to a newly created non-profit: the Patient Centered Research Foundation (http://www.pcrfoundation.org/) which will in turn contract with Clinical Data Research Networks instead of networks contracting with PCORI. This will give the networks and foundation flexibility in seeking varied sponsors and funders who may seek to use the network. Questions arise though as to how structure informs network design and collaboration. Currently, PCORnet is coordinated in largely a traditional model where recruiting sites serve the central coordinating center. But new trends, such as reciprocal IRB and the Greater Plains Collaborative's use of complementary reciprocal data sharing shift the model to allow interoperable data exchange and coordination so that trials made led by each participating site or medical center. GPC currently sees it role as shifting to an intermediary, member governerned collaborative providing the following services and roles: 1) helping member improve their regulatory, patient and clinician engagement to complement data infrastructure for participating in national research, 2) contracting with Patient Center Research Foundation (PCRF) PCORnet 2.0, 3) governing peer to peer data
Sharing to complement SMART IRB (http://www.gpcnetwork.org/sites/default/files/GPC%20Resource%20Guide_Pilot%20Program%20Supplement.pdf), 4) providing a forum for members to share technology and be accountable to one another for data quality and capability, and 5) consolidating data assets as needed such as CMS claims via GROUSE. This structure, shown in Figure 9 would allow grants to be awarded at the member site and contracting for services if needed at higher organizational levels (GPC or PCRF – formerly refered to as NewCo). National collaborative opportunities would flow-

down through GPC who coordinates regional quality and capability as well as supports centralized resources.

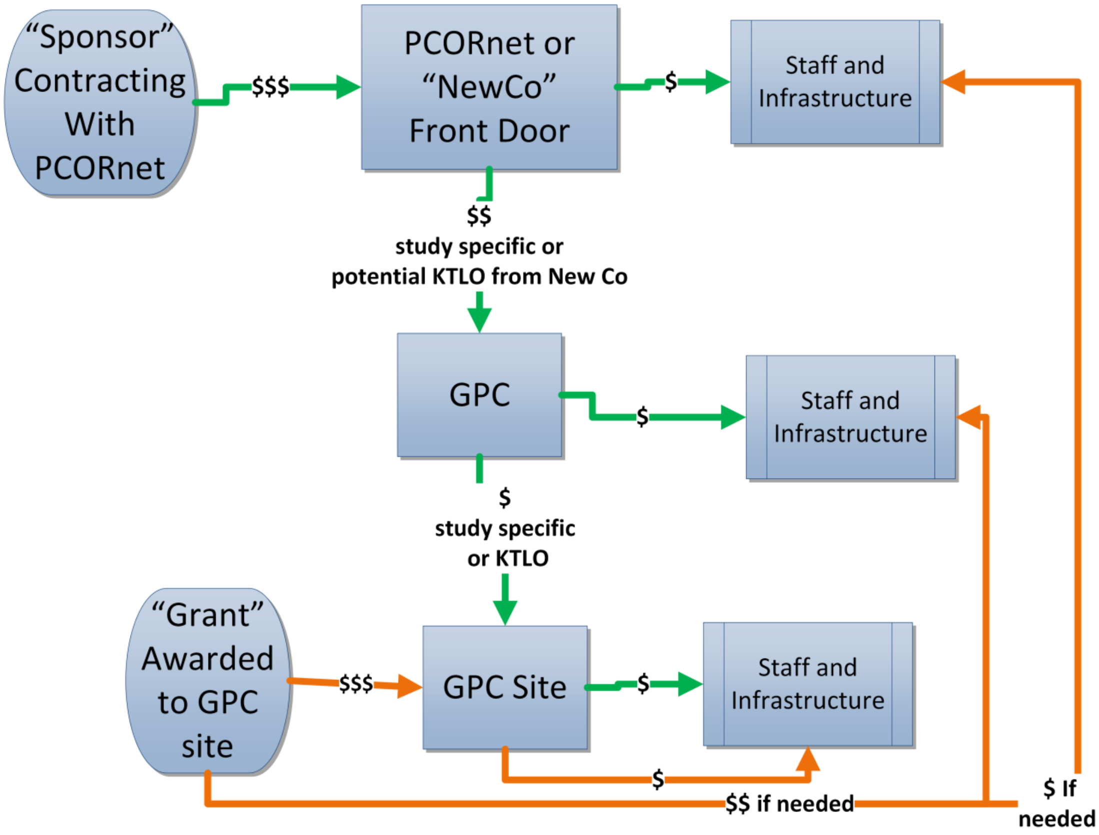


Figure 9. Proposed Greater Plains Collaborative sustainability model post Phase 2 PCORI contract

PCORnet and the Greater Plains Collaborative have catalyzed dramatic changes in regulatory, patient engagement, and data infrastructure to support research at the point of care. But, data, technology, and organizational relationship are very fluid so we are constructing these capabilities in a very dynamic time. Medical centers will continue to demonstrate their desire to lead by providing responsive regulatory and contracting activities, flexible data infrastructure, and the ability to deploy informatics interventions at the point of care with integrated and responsive approaches to patient, clinician and researcher engagement for both research and care delivery.

Acknowledgements

Portions of this text and figures were previously used in the university CTSA application biomedical informatics sections in 2010 (authors: Lemuel Waitman, Judith Warren, and Gerald Lushington) and 2016 (Lemuel Waitman) and in the PCORI PCORnet 2013 phase 1 and 2015 phase 2 applications for the Greater Plains Collaborative (authors: Lemuel

Waitman, Gary Rosenthal, Lauren Aaronson, Prakash Nadkarni, James Campbell, Daniel Connolly). Multiple references to PCORnet and especially the three demonstration projects were summarized from PCORnet slides developed during 2015-2018 by multiple individuals from PCORI and the PCORnet networks and coordinating centers; many are available at the PCORnet website (http://pcornet.org/) and the Commons (http://pcornetcommons.org/). We would like to acknowledge the participation of Frontiers, Greater Plains Collaborative, and PCORnet leadership and numerous investigators who have contributed to the development and ongoing progress of informatics and pragmatic research infrastructure in the Midwest and nationally.

References

Waitman LR, Lushington G, Warren JJ. Advancing Clinical and Transformational Research with Informatics at the University of Kansas Medical Center. Information Systems as Infrastructure for University Research Now and in the Future. Merrill Series on the Research Mission of Public Universities Mabel L. Rice, Editor. MASC Report No. 116 The University of Kansas 2012. 90-106.

Zerhouni EA. Translational and clinical science–time for a new vision. NEJM 2005, 353:1621-23.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30.

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81. doi: 10.1016/j.jbi.2008.08.010. Epub 2008 Sep 30.

Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA. 2009 Feb 25;301(8):831-41. doi: 10.1001/jama.2009.205. Erratum in: JAMA. 2009 Apr 15;301(15):1544. PubMed PMID: 19244190.

Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):578-82. doi: 10.1136/amiajnl-2014-02747. Epub 2014 May 12.

Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORNet Clinical Research Data Network. J Am Med Inform Assoc. 2014 May 12.

Cross-disciplinary Activities in Big Data for Agricultural Innovation

Carolyn J. Lawrence-Dill, Ph.D., Associate Professor, Department of Genetics, Development & Cell Biology and Department of Agronomy
Iowa State University

Agriculture is broad, involving not only crops and animals, but also the ecosystems that support their growth and development. Agricultural products are major economic drivers the world over. As the population increases and climates change, pressures on agricultural systems increase. At the same time, we seek to improve how we produce agricultural products by reducing inputs including pesticides, herbicides, antibiotics, and fertilizers. Taken together, these pressures tell us that we need to discover, design, and invent new ways to improve agricultural products.

Solving complex agricultural problems involves a multidisciplinary approach involving expertise from engineering, data sciences, and plant sciences. One way to engage a broader group in addressing these problems is to make the data that describe ecosystems, crops, and animals more easily accessible, comprehensible, and available to researchers. This extreme data sharing perspective is in keeping with long-standing traditions in science. It is widely recognized that information must be communicated or it is effectively lost (which is a driving force behind why research results are published) and that research results should be reproducible. Contrary to these basic assertions, Longo and Drazen state (in a commentary ironically entitled "Data Sharing") that "someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters" (2016). This implies that those generating data must routinely fail to describe their materials and methods sufficiently to enable true reproducibility. The authors go on to assert that, "There is concern among some front-line researchers that the [research] system will be taken over by what some researchers have characterized as *research parasites*." It is possible that in some areas of research this perspective will prevail, but in the area of agricultural innovation stakes are high. To the degree that limiting access to data stands in the way of innovation, that position cannot be supported.

Data standardization seeks to improve both human and machine access to and analysis of data. The need for standardization was well described by Lohr, who reported that 50-80% of a data scientist's time is spent aggregating and formatting data for analysis (2014). For crop improvement, a primary datatype selected for improvement is 'phenotype.' Phenotypes constitute all observable characteristics of an organism, so phenotypic descriptions include those traits that should be selected for improvement. Unfortunately for standardization, phenotypes include as many different types of data as one can imagine. From leaf angle to root depth, from infrared spectrograms to molecular gel patterns, much can be observed. To add to this issue, metadata about the environment the organisms experience and even the level of observation (e.g., single plant versus average across a field) must be documented. The development of standards that would enable data discovery, simple aggregation, and wholesale analysis are largely lacking and where they exist, use is spotty, which is not surprising given that MIAPPE (Minimal Information About a Plant Phenotyping Experiment), the first well-described standard for collecting and describing plant phenotype data, was only released two years ago (Krajewski et al., 2015).

In argument against the development and use of standards for this emerging field of research, there are many discussions held among those working in the field debating whether the area is ready for standardization. The concern is that if standards are developed and their use is required too soon, novel mechanisms for data representation might be missed. What's more, if standardization constrains how scientists think of these data, some opportunities to develop new ideas for methods of analysis could be missed entirely, making the debate on creation and use of standards a hot topic at many scientific meetings where phenotyping is a focus.

Beyond the topic of how data are formatted and made accessible, the need for scientists with broad expertise to work together to address agricultural issues involving the measurement and analysis of plant phenotypes remains. Unfortunately, cross-disciplinary efforts in the area of crop improvement remain an exception with most researchers working within well-described and narrow disciplinary boundaries. To push scientists to work more broadly, initiatives such as the Iowa State University Plant Sciences Institute (PSI) Faculty Scholars program have been initiated. For PSI Scholars, research funding is provided to faculty members working in the area of predictive plant phenomics (where phenomics is the set of all possible phenotypes a species could produce across all possible environments). PSI Director Patrick Schnable developed a program modeled on the Max Planck Institute in Germany and the HHMI and the HHMI-GBMF fellowship programs, which fund people rather than projects. Researchers working in the areas of plant sciences, data sciences, and engineering are funded to focus on plant phenomics problems and a community atmosphere is created among PSI Scholars by getting the group together weekly during the academic school year. Unlike traditional grant funding, PSI scholars themselves are funded rather than specific projects, giving them freedom to pursue specific projects they find to be of use to develop the discipline. For details see https://plantsciences.iastate.edu/about_us/psi_faculty_scholars/plant-sciences-institute-announces-psi-faculty-scholars/. Another mechanism Iowa State researchers have developed to push on this front involves student training. A grant from the National Science Foundation in Predictive Plant Phenomics (P3) supports novel graduate education and research aimed at creating graduates with expertise in plant sciences, data sciences, and engineering. These local developments are reinforced by research networks like the North American Plant Phenotyping Network, a new organization founded by the broad research community, and by the creation of PHENOME, a new scientific meeting, first convened in 2017, which is organized by the research community and supported by the American Society for Plant Biology.

Because current approaches to agricultural improvement do not show the gains necessary to meet anticipated future needs, it is clear that the general approach to agricultural improvement must evolve. Through the development of shared data access and analysis mechanisms and by supporting cross-disciplinary collaborative activities focused on phenotype measurement and analysis, researchers are actively developing the infrastructure and human resources required to support the development of a new paradigm for research that results in agricultural innovation.

References

Krajewski P, Chen D, Ćwiek H, van Dijk AD, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S. Towards recomme dations for metadata and data handling in plant phenotyping. J Exp Bot. 2015 Sep;66(18):5417-27.

Lohr, S. "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights" *New York Times* 17 August 2014.

Longo, D.L. and Drazen, J.M. Data Sharing. N Engl J Med 2016; 374:276-277 Jan 21, 2016 DOI: 10.1056/NEJMe1516564

Developing Data Science at UNL: Progress, Challenges, and Opportunities for Research

Jennifer L. Clarke, PhD, University of Nebraska

Over the past several years we have seen a groundswell of interest and investment in what is commonly referred to as `data science'. As a statistician, I am biased to consider data science as simply what statisticians have been doing for decades. However, I have come to appreciate that data science can be a broader and more encompassing endeavor, one that encourages interdisciplinary research.

I was hired in 2013 by the University of Nebraska-Lincoln as a faculty member in the Department of Statistics and the Department of Food Science and Technology. My primary role on campus is as Director of the Quantitative Life Sciences Initiative (QLSI), a program of excellence whose mission is to develop expertise and resources in data science and `Big Data' for disciplines in the Life Sciences. I advocate for resources and expertise related to turning data into knowledge, e.g., develop graduate and undergraduate curricula on data topics in the life sciences, serve as a liaison between UNL and stakeholders with interests in Big Data, and enable research in data and the life sciences. I report to the Dean of Agricultural Research within the Institute for Agriculture and Natural Resources, the UNL Vice Chancellor for Research and Economic Development, and to the QLSI faculty advisory committee.

Why QLSI? Over the past 20 years, and certainly over the last few years, the ways in which we analyze, process, store and interact with data have been rapidly changing (and there is no indication that this process is slowing). The pace of change can be quickly exemplified by a quick look at the types of media and communication devices we use today (MP3 players, BlueRay discs, smart phones) compared to a few decades ago (VHS tapes, floppy discs, answering machines)[see Figure 1]. Advances in computing have brought us the era of `Big Data', a term with different meanings to different constituencies. A good working definition, albeit relative to each individual, is more data than one is accustomed to or more than one can manage. Experts continue to discuss what aspects of data define `Big Data' [see Figure 2]; four common attributes of Big Data are

- *Volume* or scale of data, e.g., in petabytes or exabytes
- *Velocity* or speed of data, e.g., streaming data from sensors
- *Variety* or different types of data, e.g., text and images and GPS-tagged locations, and
- *Veracity* or level of uncertainty, e.g., missing or inaccurate data.

The last attribute, veracity, applies to any type of data and hence is not exclusive to Big Data (see [1] for an ongoing discussion of Big Data). However, it is an important attribute to keep in mind as a reminder that the amount of useful information in data may not scale with data volume. The discussions around Big Data are happen-ing now because (1) academic disciplines are becoming more quantitative; (2) data collection is becoming easier and less expensive; and (3) there is enough available computing power to analyze larger amounts of data than has previously been possible [2].

This brings us to one of the challenges of 21st century science: How to get from data to information to knowledge when data are large, noisy, and complex. The data-to-knowledge process requires a diverse skill set that draws upon expertise from multiple disciplines. For example, a key societal challenge is feeding a growing global population in a manner that is resource efficient and environmentally sustainable. The development of such a process will involve improvements in weather prediction, farm management practices, plant and animal breeding, and food storage and transportation, as well as reductions in food waste. Each of these improvements can only be achieved with the collection and analysis of data by scientists with domain knowledge as well as advanced data management, analysis, and communication skills. This combination of skills is relatively unusual and requires considerable education and training to achieve.

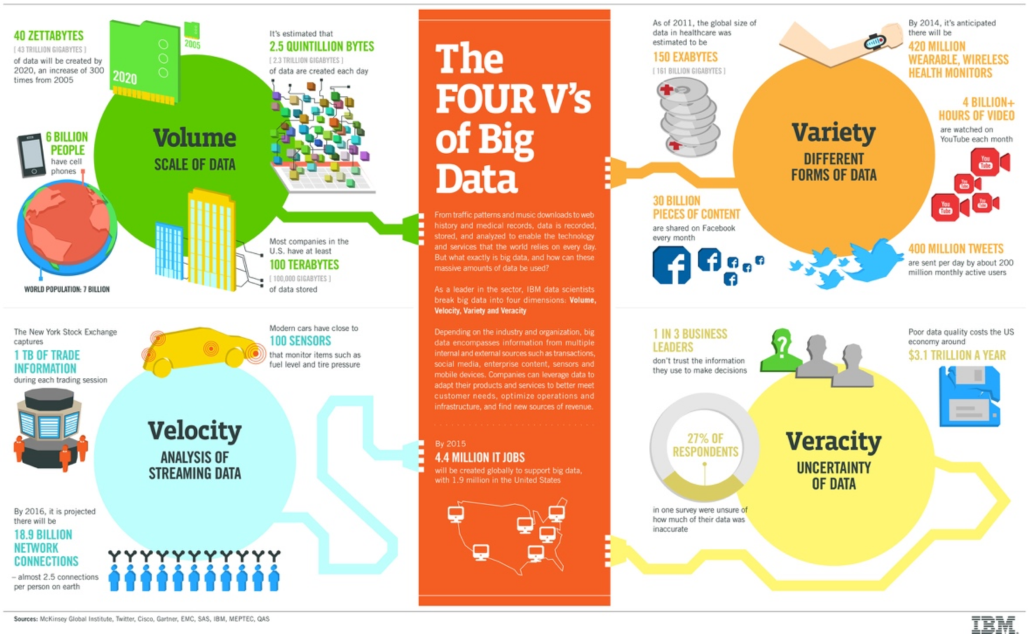
We need to rethink undergraduate, graduate, and continuing education if the academic community is going to fulfill the national workforce needs in data science. When I arrived at UNL I spent a lot of time learning about the campus, identifying a set of initial goals and plans for evaluation for QLSI, building connections with external partners, and developing buy-in among faculty and administrators. This process revealed several opportunities for the development of data science for the life sciences that would benefit both the campus and its stakeholders. One idea that I pursued in Spring of 2014 was an undergraduate major in data science/informatics that would provide students with a coherent set of curricula in the information sciences. Although this idea garnered strong support from several constituencies on campus, the academic administration decided that the university should increase undergraduate enrollment in order to accommodate an additional major.

We decided to develop an interdisciplinary doctoral program in Complex Biosystems [3]. This program was primarily driven by junior faculty from three separate colleges whose research programs required access to graduate students with training in the quantitative life sciences. All students participate in an initial year of core training before selecting advisors and a program specialization; the current specializations are microbial interactions, integrated plant sciences, systems analysis, pathobiology and biomedical sciences, and computational organismal biology, ecology, and evolution (COBEE). Qualified faculty can participate in one or more specializations. We also co-host a graduate student recruitment event each year with the Office of Graduate Studies and existing graduate programs in the life sciences. This event is very popular and has increased both program awareness and recruitment success rates.

QLSI has active research and/or educational partnerships with local, regional, national, and international organizations. These include the Midwest Big Data Hub (midwestbigdatahub.org/), the North American Plant Phenotyping Network (http://nappn.plant-phenotyping.org/), the Nebraska Food for Health Center (http://foodforhealth.unl.edu/), the Fraunhofer Institute for Integrated Circuits (https://www.iis.fraunhofer.de/en.html), the Great Plains Network (https://www.greatplains.net/), and CyVerse (http://www.cyverse.org/). These partnerships are critical to the success of the Initiative for several reasons. First, data science is a rapidly evolving discipline and partnerships are an effective way to become aware of the latest developments. Second, these partnerships provide opportunities for cutting-edge graduate training experiences. Finally, the research reputation of Nebraska and the UNL research funding portfolio both benefit from such collaborations.

A recent area of emphasis for QLSI is reproducible research and Big Data management and analysis. We have partnered with UNL Libraries and Office of Graduate Studies to support and promote the use of ORCiD (https://orcid.org/) and common metadata standards. We also encourage the use of shared research infrastructure such as NSF XSEDE, the Open Science Grid (OSG), Galaxy (https://galaxyproject.org/), and CyVerse [see Figure 3]. These activities are of particular interest to faculty associated with federally supported research centers who are obligated to comply with federal data sharing standards and expectations. Sharing and hosting large amounts of research data can be both time consuming and costly, while universities that receive public research funding have an obligation to conduct `open science' and share their research products. How to finance the maintenance and effective sharing of data in the era of Big Data and the Internet of Things (IoT) remains an open challenge [4], and one we must surmount if we are to remain the stewards of research for public benefit.

Fig. 1. A graphic example of how our relationships with data and modes of communication have changed rapidly over the past few decades with advances in computing and information technology



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States

Volume SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least 100 TERABYTES [100,000 GIGABYTES] of data stored

Velocity ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session

Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth

Variety DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT are shared on Facebook every month

By 2014, it's anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around $3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Fig. 2. The Four V's of Big Data as an infographic from IBM. The debate over the 'V's of Big Data continues, with some favoring only 3 'V's (without veracity) and others advocating for 5 'V's (including value).

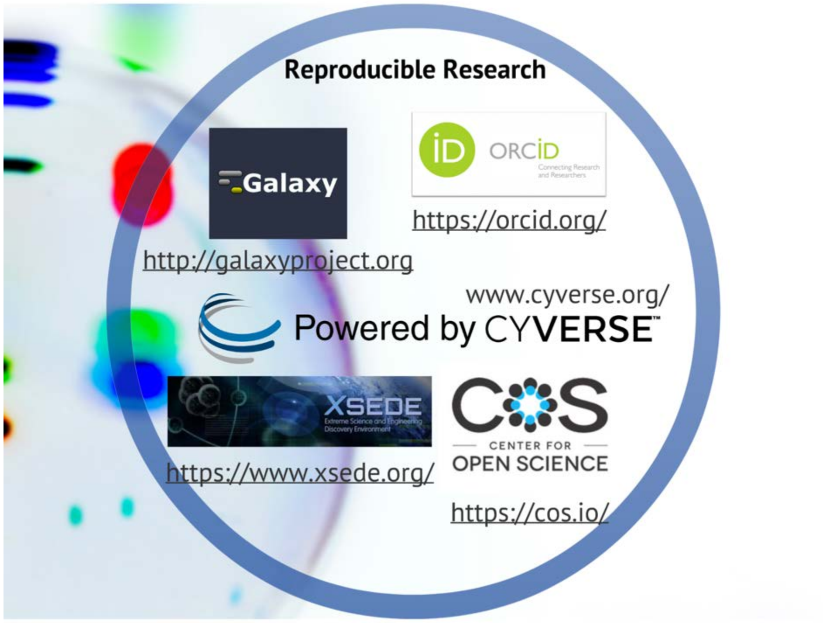


Fig. 3. Several examples of resources that enable reproducible research. These include tools for researcher disambiguation, data analysis, distributed computing, and open science.

Works Cited

1. Laney, Doug. Batman on Big Data. Post to the Gartner Blog Network, November 13, 2013. http://blogs.gartner.com/doug-laney/batman-on-big-data/
2. King, Gary. Big Data is not about the data! Presentation at Shanghai Jiao Tong University, January 4, 2017. https://gking.harvard.edu
3. Schrage, Scott. New doctoral program links life sciences with big data. Nebraska Today, October 28, 2016. http://news.unl.edu/newsrooms/today/article/new-doctoral-program-links-life-sciences-with-big-data/
4. CDWVoice. The future is data-driven, but IoT has its challenges. Forbes BrandVoice, May 2, 2017. https://www.forbes.com/sites/cdw/2017/05/02/the-future-is-data-driven-but-iot-has-its-challenges/#52abf1511459

Enhancing and Automating University Reporting Of R&D Expenditure Data Using Machine Learning Techniques[1]

Joshua L. Rosenbloom, Iowa State University, National Bureau of Economic Research

Rodolfo Torres, University of Kansas

Joseph St. Amand, University of Kansas

Adrienne Sadovsky, University of Kansas

Higher Education Spending on Research and Development

In 2014, U.S. Colleges and Universities reported spending $67.3 billion on Research & Development (R&D). While this figure constitutes only about 15 percent of the nation's total R&D effort, colleges and universities performed more than half of U.S. basic research.[2] Most of what we know about R&D performed at the nation's colleges and universities—not only aggregate totals, but also expenditures by field of study and by individual institutions – is derived from data collected by the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES) as part of its Higher Education R&D (HERD) survey. The HERD survey continued and expanded a data collection effort that was started in 1972 as the Academic R&D Expenditures Survey.

The data collected by the HERD survey are widely used by both university administrators and academic researchers interested in understanding the nation's scientific enterprise. University leadership is interested in tracking total R&D expenditures and rankings of R&D expenditures as an indicator of research prowess. Most universities work to move up in the rankings by increasing their expenditures. Scholars interested in the political economy of federal science funding have used HERD expenditure data to track the expansion of the nation's cadre of research universities and to assess the tendency of the political system to promote more equal distribution of funds across states and regions (Geiger and Feller 1995; Graham and Diamond 1997; Feller 2001). Others have used more disaggregated data on expenditures at the

[1] The material in this article is based in part upon work supported by the National Science Foundation under Grant Numbers SMA-1547513 and SMA-1547464. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2] National Science Board 2016, chapters 4, 5

discipline level to explore either how funding is related to scientific productivity (Adams and Griliches, Rosenbloom et al 2015) or to uncover factors that influence the allocation of federal R&D funding across institutions (Lanahan et al 2016; Rosenbloom and Ginther forthcoming).

The data collected in the HERD and the earlier Academic R&D Survey are derived from institutional responses to an annual survey distributed by NCSES. Colleges and Universities undoubtedly take different approaches to compiling the necessary data, but at research universities with specialized research administration staff, responsibility for responding to the survey is likely delegated to one or more specialists within the office of sponsored research or institutional research.

The aforementioned method of collecting data on college and university R&D expenditures results in three distinct problems. First, responding to the HERD is costly in terms of the time required to accurately report the requested data. Second, because of the nature of the annual survey and the lead time involved in tabulating responses, the data are available only with a long lag. While the data are useful for retrospective analysis, the lags make them far less valuable for setting institutional strategy or performing real-time analysis of R&D activity. Finally, the effort to classify projects by their purpose and field of study is inevitably subjective, making it problematic to make comparisons across institutions and introducing spurious variation within institutions when responsibility for data collection shifts from one person to another.[3]

To address these problems, we have been engaged in an experiment to apply techniques of machine learning to automate project classification. Successful development of classification algorithms would reduce the cost of responding to the HERD survey, allow for essentially real-time tracking of expenditures, and offer the potential to increase the consistency of classification over time and across institutions. As we describe here, our proof of concept investigation suggests that such approaches are potentially feasible, but require further efforts.

Application of Machine Learning to Classify Sponsored Research Projects

With the growth of large data sets and the declining cost of computation, application of machine learning techniques to identify data patterns and make predictions based on these patterns has become increasingly common.[4] The goal of our project is to develop a classification algorithm that can be used to either supplement or replace human judgement in classifying sponsored research projects. To do so, we begin with a set of sponsored projects awards that have already been classified by Research Administration staff at the University of Kansas. In

[3] The survey categorizes projects into 4 different purposes (applied research, basic research, development, and other), and 40 different scientific fields of study (e.g., Bioengineering and Biomedical Engineering, Astronomy and Astrophysics, Political Science and Government, etc.).

[4] For an overview of machine learning and associated terminology see: https://en.wikipedia.org/wiki/Machine_learning

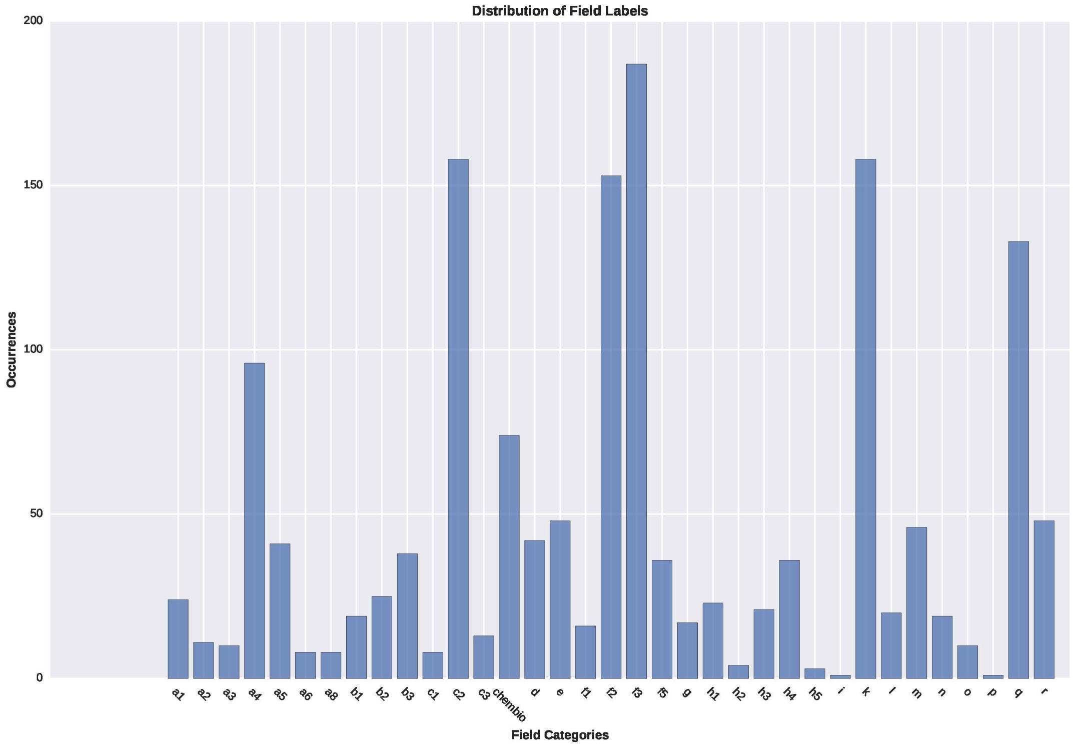
the language of machine learning, this is an example of "supervised learning."

Working with staff in the University of Kansas Office of Research, we obtained a data set of historical sponsored project awards. After dropping awards for which we did not have complete data, we were left with approximately 1,500 projects. For each of these projects, the data included information on the:

- Project sponsor
- Principle Investigator (PI) home unit
- Project abstract describing the project
- Human-assigned classification of the project's purpose and field of study.

Figures 1 and 2 show the distribution of projects across fields of study and purpose based on the human-assigned classification. In addition to the full list of NSF-defined fields of study, our data include KU-specific fields of "Chem-Bio" and "CEBC" (that combines projects across chemistry, chemical engineering and biomedical sciences) that are used for internal institutional purposes.[5] As Figure 1 illustrates, there are some fields for which we do not have a large number of projects. The distribution of projects by purpose is also somewhat uneven, as illustrated in Figure 2.

Figure 1: Distribution of Projects by Field of Study

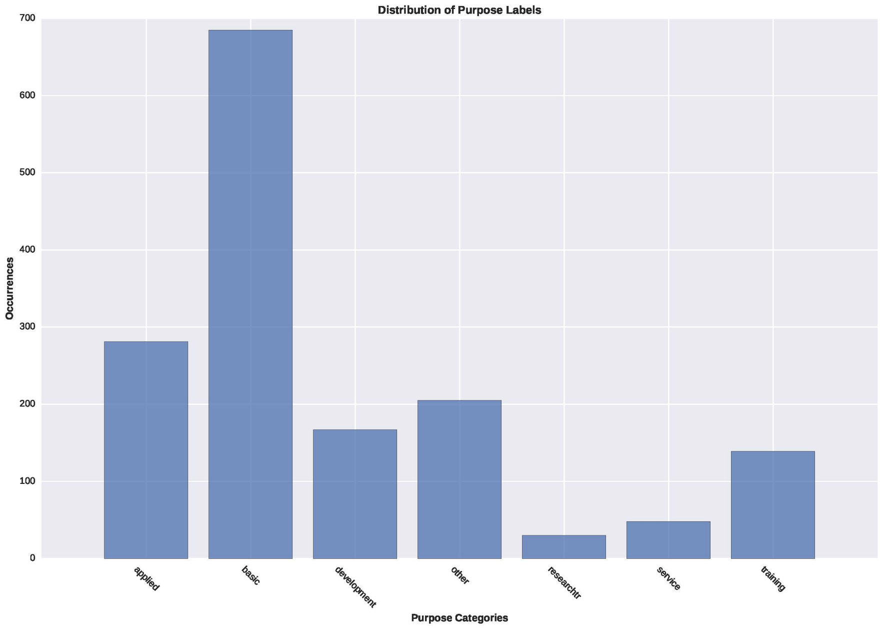


[5] For reporting purposes, expenditures in the Chem-Bio and CEBC categories are split evenly between the Chemistry and Biological and Biomedical Sciences and Chemistry and Chemical Engineering, respectively.

Notes to Figure 1: The University of Kansas Fields of Study are denoted with the following codes.

Field	Code
Computer and Information Sciences	A
Aerospace / Aeronautical / Astronautical Engineering	B1
Bioengineering and Biomedical Engineering	B2
Chemical Engineering	B3
Civil Engineering	B4
Electrical, Electronic, and Communications Engineering	B5
Industrial and Manufacturing Engineering	B6
Mechanical Engineering	B7
Metallurgical & Materials Engineering	B8
Other Engineering	B9
Atmospheric Sciences and Meteorology	C1
Geological and Earth Sciences	C2
Ocean Sciences and Marine Sciences	C3
Other Geosciences, Atmospheric, and Ocean Sciences	C4
Agricultural Sciences	D1
Biological and Biomedical Sciences	D2
Health Sciences	D3
Natural Resources and Conservation	D4
Other Life Sciences	D5
Mathematics and Statistics	E
Astronomy and Astrophysics	F1
Chemistry	F2
Materials Science	F3
Physics	F4
Other Physical Sciences	F5
Psychology	G
Anthropology	H1
Economics	H2
Political Science and Government	H3
Sociology, Demography, and Population Studies	H4
Other Social Sciences	H5
Other Sciences	I
Business Management and Business Administration	K
Communication and Communications Technologies	L
Education	M
Humanities	N
Law	O
Social Work	P
Visual and Performing Arts	Q
Other Non-S&E Fields	R

Figure 2: Distribution of Projects by Purpose



The major source of information about each project comes from the proposed statement of work, which is treated as a "bag of words." As a first step, we pre-process the data by standardizing word forms and eliminating "stop-words" (e.g. the, is, to). Individual (or collections of) words are converted to a numerical form based on their frequency of occurrence within and between project abstracts. In machine learning, these numerical representations are referred to as "features." The goal of the machine learning algorithm is to assess which specific combination of features are useful to discriminate between purpose/field categories.

Once the data are processed, we experimented with a selection of commonly used classifiers to identify features that provide predictive power. Following standard practice, we split the data into training and testing samples. The training sample contains approximately 70% of the observations, while the testing sample contains the remaining 30%. The models are trained on the training sample (which is further split into ÛÙÈÐÕÐÕÎш and validation samples) via a cross-validation procedure, which is necessary to prevent the models from over-fitting (i.e. "memı orizing") the data. We estimate the prediction error of the models using the validation samples, and use the testing sample as an assessment of generalization error.

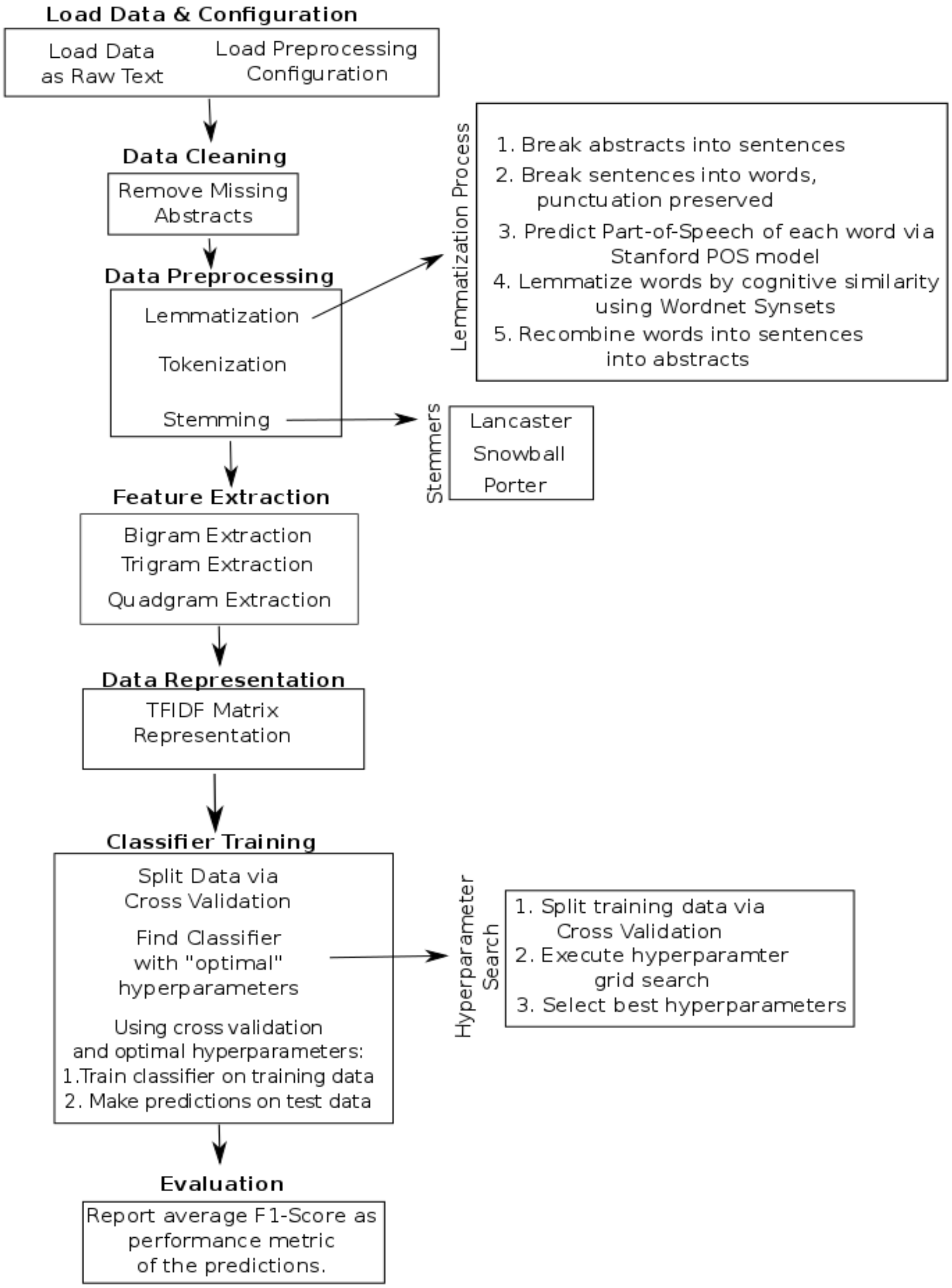

Load Data & Configuration
Load Data as Raw Text
Load Preprocessing Configuration
Data Cleaning
Remove Missing Abstracts
Data Preprocessing
Lemmatization
Tokenization
Stemming
Lemmatization Process
1. Break abstracts into sentences
2. Break sentences into words, punctuation preserved
3. Predict Part-of-Speech of each word via Stanford POS model
4. Lemmatize words by cognitive similarity using Wordnet Synsets
5. Recombine words into sentences into abstracts
Stemmers
Lancaster
Snowball
Porter
Feature Extraction
Bigram Extraction
Trigram Extraction
Quadgram Extraction
Data Representation
TFIDF Matrix Representation
Classifier Training
Split Data via Cross Validation
Find Classifier with "optimal" hyperparameters
Using cross validation and optimal hyperparameters:
1.Train classifier on training data
2. Make predictions on test data
Hyperparameter Search
1. Split training data via Cross Validation
2. Execute hyperparamter grid search
3. Select best hyperparameters
Evaluation
Report average F1-Score as performance metric of the predictions.

Figure 3: Schematic Representation of Data Analysis Steps

We treat the prediction of purpose and field categories as two separate classification tasks. For each classification problem we tried the following classifiers:

- Decision Tree
- Support Vector Machine
- Logistic Regression
- Random Forest
- Naïve Bayes
- Neural Network

All of these classification schemes are binary: reporting a probability that the project belongs to a particular purpose/field. For each purpose/field, the classifier yields a predicted probability (between 0 and 1) or a categorical determination that the project belongs to that purpose/field.

We first find classifiers for each purpose/field by assessing their performance (as described below) and then assign projects to a single purpose/field using the purpose/field with the highest predicted probability across all the classifiers.

The result of our analysis is a prediction of the purpose/field to which each project should be assigned. Comparing the human- and machine-assigned results produces a two-way contingency table depicted in Figure 4. Projects for which the two classifications agree (T1 and T2) are successful predictions, whereas cases where the assignment is different (F1 and F2) are unsuccessful.

Figure 4: Project Classification Outcomes

Actual Outcome	**Predicted Outcome**	
	In Field/Purpose	**Not in Field/Purpose**
In Field	T1	F1
Not in Field	F2	T2

A number of measures of the performance of machine learning algorithm are possible. The "accuracy" of the predictions is simply the number of correct predictions as a share of all predictions:

(T1+T2)/(T1+T2+F1+F2)

Where the distribution of outcomes is uneven, however, this measure may not be very illuminating. For example, in a binary classification problem where 90% of observations are not in a field, simply guessing that no projects belong to that field would yield an accuracy of 0.9, but would be a thoroughly uninformative classifier.

To correct for this, two other measures of prediction success have been proposed and are routinely used in evaluating the effectiveness of machine learning algorithms. They are:

- Precision = T1/(T1+F2), and
- Recall = T1/(T1+F1)

Intuitively, Precision measures the share of all projects belonging to a classification that are correctly identified; Recall measures the share of projects that are predicted to be in the field that are correctly predicted. The F-1 score, which is the harmonic mean of Precision and Recall, is generally viewed as the best single summary of classifier performance.

Results

Among the different machine learning models, we found that the Logistic Regression classifier provides the best overall performance. Figure 5 summarizes the performance of the classifiers for each field of study, and Figure 6 reports performance for the classifiers for project purpose. In each case, we compare F-1 scores from the cross-validation results to those obtained using the testing sample. In cases where the number of projects was too small, we do not have any projects in the testing sample, so we cannot compute an F-1 score.

Figure 5: Comparison of F-1 Scores for Field of Study in Cross-Validation and Testing Samples

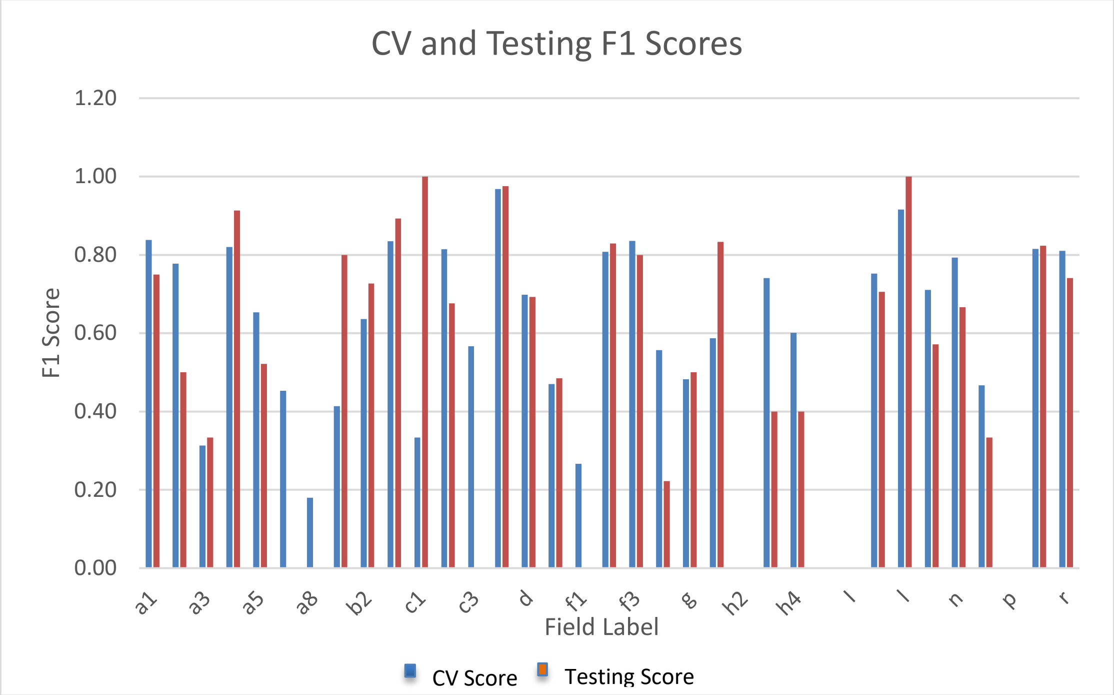
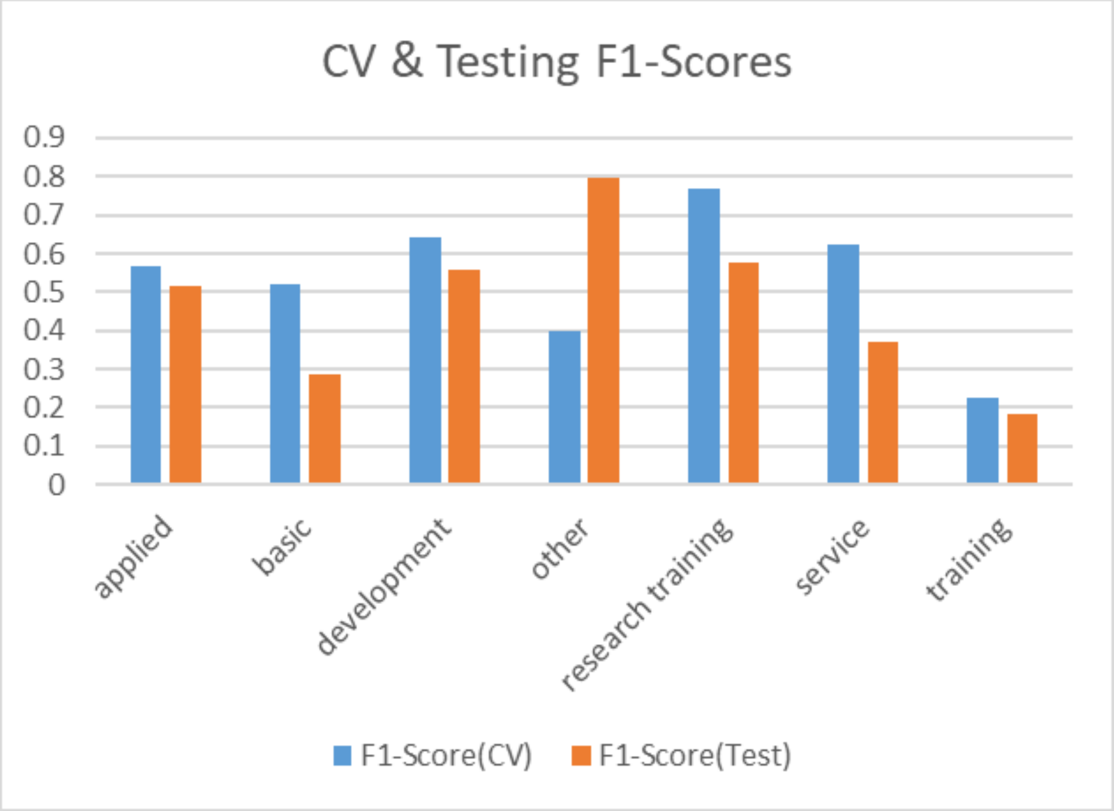


Figure 6: Comparison of F-1 Scores for Purpose in Cross-Validation and Testing Samples



As shown earlier (Figures 1 and 2), the distribution of projects by purpose and field is highly uneven; this difference accounts for much of the variation in classifier performance across the purpose/field categories. Figure 7 plots the relationship between the F-1 score and the number of projects assigned to each field in the training sample. F-1 scores rise sharply as the number of projects increases from 0 to about 40, reaching a range of 0.6-0.9 at this point. For cases with more than 60 projects the F-1 scores are clustered around 0.8.

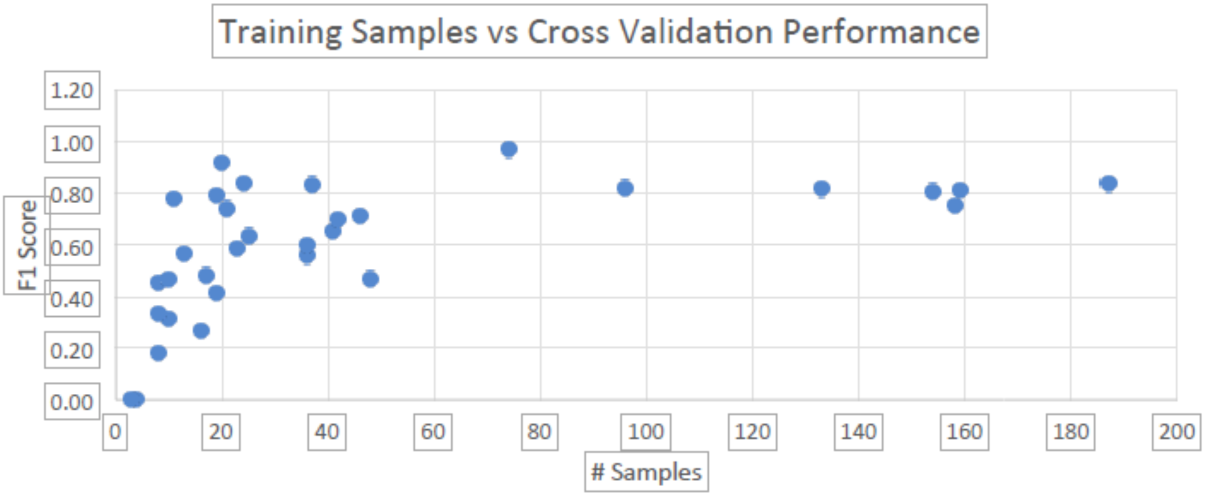


Figure 7: F-1 Scores for Field of Study vs. Number of Projects in Training Sample

Figure 8 shows comparable relationships for project purpose.

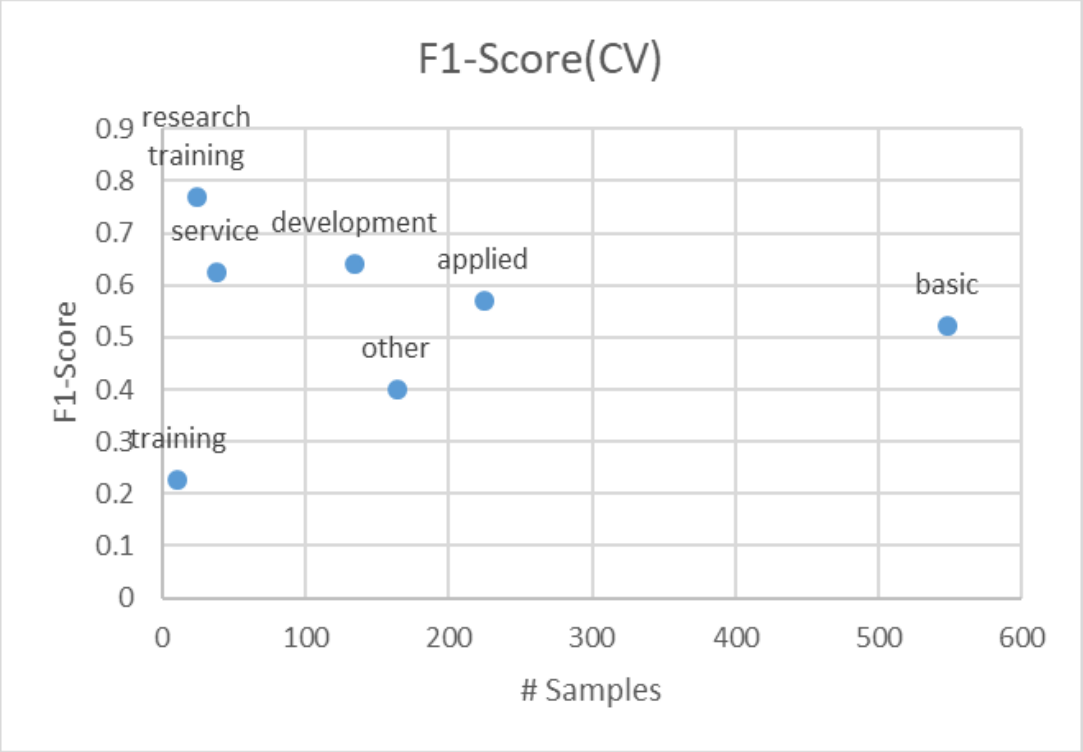


Figure 8: F-Scores for Purpose vs. Number of Projects in Training Sample

Discussion

We have not yet succeeded in developing a set of classifiers that will precisely reproduce the human judgements underlying the University of Kansas's response to the HERD survey. But it is not clear that this is the appropriate measure of the project's success.

First, while we have relied on human judgements to train the classifiers, it is not entirely obvious that we should regard the human assessments as constituting the ground truth in this case. Furthermore, different individuals at KU have classified projects for the HERD survey over the years; this may have added additional subjectivity or inconsistency in the classification of projects. It may be that the classifiers are more consistent in their judgment than humans. Evaluating this possibility requires more careful examination of the cases in which the two approaches produce different results. Careful analysis of these cases may help clarify the root of the disagreement and yield additional insights.

Second, the end product of the classification process is an aggregated report on expenditures broken down by field of study and purpose. Rather than focusing on the accuracy of the individual project classifications, it may prove more valuable to look at the extent to which aggregated results from the machine classifiers approximate the aggregated results from the human classifiers.

Conclusions

As a proof of concept, we believe that the current project has been successful in demonstrating that it is possible to develop reasonably accurate machine-learning classifiers. We believe that many of the problems encountered so far will be reduced by expanding the training data set to include additional examples.

One initial objective of our project was to bring greater uniformity to HERD reporting across institutions. Future goals for this project include assessing the ability of our classifiers to successfully classify projects at other institutions. Adding additional projects from other institutions to the training data set may also offer opportunities to further refine the classifiers we have developed.

References

Adams, James D. and Zvi Griliches. 1998. "Research Productivity in a System of Universities." Annals of Economics and Statistics 49/50: The Economics and Econometrics of Innovation, 127-62.

Feller, Irwin. 2001. "Elite and/or Distributed Science." In Maryann P. Feldman and Albert N. Link, eds, Innovation Policy in the Knowledge Based Economy. Norwell, MA and Dordrecht, Nether-lands.: Kluwer Academic.

Geiger, Roger and Irwin Feller. 1995. "The Dispersion of Academic Research in the 1980s." Journal of Higher Education 66, 336-60.

Graham, H. Davis and Nancy Diamond. 1997. The Rise of American Research Universities: Elites and Challengers in the Postwar Era. Baltimore, MD: Johns Hopkins University Press.

Lanahan, Loren, Alexandra Graddy-Reed and Maryann P. Feldman. 2016. The Domino Effects of Federal Research Funding. PLOS ONE (June 21) https://doi.org/10.1371/journal.pone.0157325

Rosenbloom, Joshua L., Donna K. Ginther, Ted Juhl and Joseph A. Hep-pert. 2015. "The Effects of Research & Development Funding on Scientific Productivity: Academic Chemistry, 1990-2009." PLOS ONE (September 15) https://doi.org/10.1371/journal.pone.0138176

Rosenbloom, Joshua L. and Donna K. Ginther. Forthcoming. "Show me the Money: Federal R&D Support for Academic Chemistry, 1990-2009." Research Policy.

National Science Board (2016) Science and Engineering Indicators 2016. Arlington, VA: National Center for Science and Engineering Statistics (NCSES) https://www.nsf.gov/statistics/2016/nsb20161/#/report

Clinical Research and Data: HIPAA, the Common Rule, the General Data Protection Regulation, and Data Repositories

Amy Jurevic Sokol, Associate General Counsel
The University of Kansas Medical Center

First, some context. My first computer was an Apple IIe. It had a 1.023MHz CPU, 64KB of RAM and booted off of floppy disks with a whopping 360KB capacity. For its time it was amazing; however, there was no such thing as "big data" thirty-four years ago. That computer, which I loved so much, would've failed miserably if given the task of crunching "big data" numbers. Fast forward to today. My iPhone has 128GB and my computer has a terabyte of storage. That terabyte is 8000GB, or to put that into context: more than 8000 times the storage capacity of that Apple IIe. In short, "big data"—and the machines required to process it—are now a reality.

The inexorable march of Moore's Law has resulted in changes in all areas of our lives, including how we do clinical research. Researchers and patients are more connected. We store, access, and manipulate data in different ways; we conduct studies in multiple countries sharing data and samples around the world; and cybersecurity and hacking are a reality. This article touches on different legal aspects arising at the intersection of technology, data, and clinical research—specifically HIPAA (the Health Insurance Portability and Accountability Act), human subjects research, the European data law (the General Data Protection Regulation), and data repositories. It attempts to explain how two different law-making bodies, the US and the EU, have tried to balance the necessity of using data for research purposes that benefit society with the privacy issues and risks of that same data.

Health Insurance Portability and Accountability Act (HIPAA) and Human Subjects Research

The US has a patchwork of federal and state laws that protect different types of data. Student records are protected by Family Educational Rights and Privacy Act (FERPA); financial data by the Gramm-Leach-Bliley Act (GLB); health data by the Health Insurance Portability and Accountability Act of 1996 (HIPAA); human subjects research by the Common Rule and FDA Regulations; and personal information, like driver's license number and social security number, by state law. There is no one overarching law that protects all data. Instead, it is a patchwork of laws that sometimes overlap and at other times have large holes of information that is not covered.[1]

The Federal Policy for the Protection of Human Subjects, or the Common Rule, outlines the basic provisions for Institutional Review Boards (IRBs), informed consent, and assurances of compliance.

Generally, it applies to research involving human subjects[2] conducted, supported or otherwise subject to regulation by one of eighteen different federal departments or agencies.[3] A different set of regulations apply to clinical investigations that are regulated by the Food and Drug Administration (FDA) or that support applications for research or marketing permits for products regulated by the FDA.[4] This includes sponsored trials of drugs and devices. Both the Common Rule and FDA Regulations require IRBs to, where appropriate, verify that there are adequate provisions in place to protect the privacy of subjects and to maintain the confidentiality of data.[5] The Common Rule requires that the informed consent form contain a statement describing the extent, if any, to which the confidentiality of records identifying the subject will be maintained.[6]

The Common Rule does not apply to public records or records in which the research subject cannot be identified directly or indirectly linked to the research subject.[7] So, if the information cannot be linked back to the subject, under the Common Rule it does not constitute human subjects research.[8] However, under the FDA regulations it may still be considered a clinical investigation.

HIPAA applies a much different standard than the Common Rule and FDA Regulations. HIPAA applies to "covered entities," which are defined as health care providers that transmit any information in an electronic form in association with standard transactions, health plans, and health care clearing houses.[9] It does not apply to all researchers; it applies to researchers that are covered entities and may apply, depending on the situation, to researchers who work for covered entities or obtain their data from a covered entity.[10] For instance, if a data repository is created by an academic medical center or health system, then HIPAA likely applies. However, if a group of individuals or a disease foundation create a data repository by submitting their own data, HIPAA likely does not apply.

Many researchers believe if they remove the patient's name and social security number they have de-identified data under HIPAA. These researchers would be incorrect. For information to be considered de-identified it has to meet the requirements of either the safe harbor or expert determination. The safe harbor requires removal of the following identifiers of the patient and the patient's relatives, employers, or household members:

- Names;
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
    - The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
    - The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000;

- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
- Telephone numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Fax numbers;
- Device identifiers and serial numbers;
- Email addresses;
- Web Universal Resource Locators (URLs);
- Social security numbers;
- Internet Protocol (IP) addresses;
- Medical record numbers;
- Biometric identifiers, including finger and voice prints;
- Health plan beneficiary numbers;
- Full-face photographs and any comparable images;
- Account numbers;
- Any other unique identifying number, characteristic, or code; and
- Certificate/license numbers.[11]

To fit within the safe harbor method all the identifiers above have to be removed, encoded, or randomized; no exceptions.[12] In addition, the researcher cannot have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.[13] It is important to note that the safe harbor method does not require removal of the physician or other health care provider's information, but only the patient's and family members' information.

Another option is the expert determination method. Under this method, a qualified statistician determines that the risk is very small that the information could be used alone or in combination with other reasonably available information by an anticipated recipient to re-identify an individual. In addition, the expert must document the methods and analysis that support this decision. [14] This method *could* allow certain identifiers to remain that would otherwise have to be removed under the safe harbor method of compliance.

If the data is de-identified under either the safe harbor or expert determination method, then HIPAA and its associated regulations do not apply to that data, and there are no limitations under HIPAA on its use or disclosure. For example, you may sell de-identified data (unless it is subject to a DUA or other agreement that prohibits it).

Researchers often need information that is not available in properly de-identified data sets. The most common request is for dates—often birth, death, admission, and discharge. In this instance, a researcher would use a *limited data set* (LDS), instead of fully identified information. A LDS is information from which the following identifiers of the individual or his or her relatives, employers or household members are removed:

- names;
- street addresses (other than town, city, state and zip code);
- telephone numbers;

- fax numbers;
- e-mail addresses;
- Social Security numbers;
- medical records numbers;
- health plan beneficiary numbers;
- account numbers;
- certificate license numbers;
- vehicle identifiers and serial numbers, including license plates;
- device identifiers and serial numbers;
- URLs;
- IP address numbers;
- biometric identifiers (including finger and voice prints); and
- full face photos (or comparable images).[15]

Examples of information that may remain and still be a LDS include:

- dates such as admission, discharge, service, birth, and death;
- city, state, five digit or more zip code; and
- ages in years, months, days, or hours.

An LDS is still considered protected health information even though there are fewer identifiers and less risk than fully identified protected health information.[16] The HIPPA Privacy Regulations require covered entities enter into a data use agreement with any recipient of a LDS. These agreements must include the following:

- a description of the permitted uses and disclosures of the limited data set;
- a list of who may use or receive the information;
- a requirement that the recipient will not use or further disclose the information, except as permitted by the agreement or as permitted by law;
- a requirement that the recipient use appropriate safeguards to prevent a use or disclosure that is not permitted by the agreement;
- a requirement that the recipient report to the covered entity any unauthorized use or disclosure of which it becomes aware;
- a requirement that the recipient ensure that any agents (including a subcontractor) to whom it provides the information will agree to the same restrictions as provided in the agreement; and
- a statement that the recipient will not re-identify the information or contact the individuals.[17]

A data use agreement has become a standardized agreement which is usually not a difficult agreement to negotiate. Increasingly, data use agreements are also being used for the disclosure of de-identified data, to prohibit the selling, re-disclosure, and noncompetitive use of de-identified data by the recipient.

Data Repositories

Creation of a Data Repository

A data repository is a collection or warehouse of data.[18] It can contain de-identified data, a limited data set, or fully identifiable information. There are two separate legal analyses that must occur when creating a data repository: first, does HIPAA apply; and second, is it considered human subject research under the Common Rule.

The HIPAA analysis starts with a seemingly simple concept: the use or disclosure of protected health information by a covered entity for research purposes requires that certain conditions under the HIPAA Privacy Rule be met. There is a lot of information packed in that one sentence. HIPAA applies to covered entities. So, if the researcher is not a covered entity, is not employed by a covered entity, and does not obtain the information from a covered entity then HIPAA and its associated regulations do not apply. Also, it only applies to protected health information. If the information is de-identified according to the HIPAA Privacy Rule (either by the safe harbor or expert determination method) then HIPAA no longer applies to the de-identified data. Finally, it must be for a research purpose. Research is defined as a "systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge."[19] If the use or disclosure is for health care operations of the covered entity then patient authorization is not required. Health care operations include conducting quality assessment and improvement activities, including outcomes evaluation and development of clinical guidelines, provided that obtaining generalizable knowledge is not the primary purpose; patient safety activities; and population-based activities relating to improving health or reducing health care costs, protocol development, case management and care coordination, and contacting of health care providers and patients with information about treatment alternatives.[20] Therefore, organizations could create a data repository for quality improvement activities without obtaining patient authorization or a waiver of authorization.

If a covered entity uses or discloses protected health information to create a database to conduct research, then the creation of the database itself is a research activity that must meet the requirements of the HIPAA Privacy Rule. If the data repository is created from a limited data set then one option is to use a data use agreement, which would enable the subsequent accessing of the data for research purposes to be achieved through a similar data use agreement.[21] Another option for creating the research repository is to obtain a HIPAA compliant authorization of each person's data that is contained in the repository. This may be a viable option for a small local database that is starting from scratch, but it is unlikely to work for anything but the smallest data repository. The more likely option is a waiver of the authorization requirement.

To create the research repository without obtaining each person's signed authorization, the researcher must get a written waiver of the authorization requirement from either an IRB or a privacy board that meets the Privacy Rule requirements.[22] The covered entity—prior to the use or disclosure—would obtain documentation of the following:

- Identification of the IRB or Privacy Board and the date on which the alteration or waiver of authorization was approved;
- A statement that the IRB or Privacy Board has determined that the alteration or waiver of authorization, in whole or in part, satisfies the three criteria in the Privacy Rule (listed below);

- A brief description of the protected health information for which use or access has been determined to be necessary by the IRB or Privacy Board;
- A statement that the alteration or waiver of authorization has been reviewed and approved under either normal or expedited review procedures; and
- The signature of the chair or other member, as designated by the chair, of the IRB or the Privacy Board, as applicable.[23]

The following three criteria must be satisfied for an IRB or Privacy Board to approve a waiver of authorization under the Privacy Rule:

- The use or disclosure of protected health information involves no more than a minimal risk to the privacy of individuals, based on, at least, the presence of the following elements:
  - an adequate plan to protect the identifiers from improper use and disclosure;
  - an adequate plan to destroy the identifiers at the earliest opportunity consistent with conduct of the research, unless there is a health or research justification for retaining the identifiers or such retention is otherwise required by law; and
  - adequate written assurances that the protected health information will not be reused or disclosed to any other person or entity, except as required by law, for authorized oversight of the research project, or for other research for which the use or disclosure of protected health information would be permitted by this subpart;
- The research could not practicably be conducted without the waiver or alteration; and
- The research could not practicably be conducted without access to and use of the protected health information.[24]

At the same time as the waiver of authorization under HIPAA is being obtained, the IRB is also determining whether or not it considers the data repository to be "human subjects research" covered under the Common Rule.[25] The Common Rule does not apply if both of the following conditions are met: the data was not collected specifically for the currently proposed research project through an interaction or intervention with living individuals, and the investigator(s) cannot readily ascertain the identity of the individual(s) to whom the coded private information pertains. [26] Conversely, obtaining identifiable private information for research purposes constitutes human subjects research. [27]

If the IRB determines that the Common Rule applies, then: (1) either the researcher will get consent from each research subject; or (2) in addition to the HIPAA waiver of authorization, the researcher will also request waiver of informed consent under the Common Rule. To approve a waiver of the requirement to obtain informed consent or approve a consent procedure which does not include, or which alters, some or all of the elements of informed consent requirements the IRB must determine and document the following:

- the research involves no more than minimal risk to the subjects;
- the waiver or alteration will not adversely affect the rights and welfare of the subjects;
- the research could not practicably be carried out without the waiver or alteration; and
- whenever appropriate, the subjects will be provided with additional pertinent information after participation. [28]

Several changes to the Common Rule go into effect July 19, 2018 that impact the informed consent process.[29] These revisions add a provision for secondary research uses of identifiable private information and identifiable biospecimens; [30] add a provision for broad consent for the storage, maintenance and secondary research use of identifiable private information and identifiable biospecimens; [31] modify the list of required elements of informed consent to include an additional statement if private identifiable information or identifiable biospecimens are collected[32] and additional language for the consent form if biospecimens (even if identifiers are removed) will be used for commercial profit[33] or if research of biospecimens will or might include whole genome sequencing; [34] and a provision approving research where a researcher obtains information or biospecimens without consent for the purpose of screening, recruiting, or determining eligibility of a prospective research subject if certain conditions are met. [35]

Once all of the approvals are obtained, the data repository may be populated with data. However, if new data sources are added, data elements collected are changed, or the protocol is revised, then the researcher would have to do the HIPAA analysis again. For instance, if the repository was originally considered to be de-identified or a limited data set, is that still the case after additional data is added? Also the IRB would want to review any modifications to the data repository or associated protocol prior to their implementation.

Accessing Data in a Data Repository

Each time protected health information (even a limited data set) is accessed for a research purpose, then the requirements for access must be met as well. Like the creation of a data repository, this is potentially a two-part analysis. The first part is the analysis under HIPAA; the second, an analysis under the Common Rule.

Under HIPAA, to access de-identified data no additional steps are required unless as part of the protocol that created the data repository the researcher stated that a data access or system access agreement would be signed by researchers accessing the data. In this case, the protocol must be followed. If a limited data set or fully identifiable protected health information is requested, then under the HIPAA Privacy Rule one of the following circumstances and conditions must be met:

- The request is a review preparatory to research and certain representations are obtained from the researcher;
- The research is solely on decedents' information and certain representations are obtained from the researcher;
- A HIPAA-compliant authorization was signed by each subject of the PHI, granting specific written permission for the access and use of the information;
- An IRB or Privacy Board has granted and documented the grant of a waiver or an alteration of the authorization requirement (if an alteration of the authorization is granted, a signed authorization is required from each individual);
- The PHI has been de-identified in accordance with the standards set by the Privacy Rule (in which case, the information is no longer PHI);
- The information is released in the form of a limited data set and a data use agreement between the researcher and the covered entity is signed;
- Informed consent of the individual to participate in the research, an IRB waiver of such informed consent, or other express legal permission to use or disclose the information for the research is grandfathered by the transition provisions. [36]

The Common Rule analysis is just as complicated as the HIPAA analysis, maybe even more so. Under the Common Rule, obtaining identifiable private information for research purposes constitutes human subjects research. This includes information that is already in the possession of the investigator. [37] Conversely, research is not considered human subject research if the research only involves coded private information of human subjects if both of the following conditions are met: (1) the private information was not collected specifically for the proposed research project through an interaction or intervention with living individuals; and (2) the investigator(s) cannot readily ascertain the identity of the individual(s) to whom the coded private information pertains. [38]

Information is considered identifiable when the information can be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems. Additionally, an investigator is broadly defined to include anyone involved in conducting the research. [39] An investigator would not include an honest broker who solely provides de-identified or coded information as long as the honest broker does not collaborate on other activities related to the conduct of this research with the investigator(s) who receive(s) such information.[40] An example of the difference between these two scenarios in plain English is: First, a researcher accesses a database with patient information that identifies the patient, queries the information, and records the results in a way that is coded but maintains the key to decode the data. Or, second, an honest broker, who is not part of the research team, accesses the database, queries the data, codes the results so that the patients are not identifiable, and the honest broker maintains the linking code.

The first example is human subjects research, while the second is not.

Even if the accessing of data is considered human subjects research, it may be exempt under the Common Rule. According to the guidance published by the Office for Human Subjects Protections, the most relevant exemption is if the information obtained by the investigator is recorded in such a manner that the individuals cannot be identified directly or indirectly through identifiers linked to the individuals.[41] Using our previous examples, the researcher would access a database with patient information that identifies the patient, would query the information, and would record the results in a way that is coded with no way to decode the data to identify the individuals.

If the access to the research repository is considered human subjects research that is not exempt, then the investigator must submit a study submission as a new study or a modification to an existing study. In addition, the researcher is likely to request a waiver of the informed consent requirement as well.

The European Union and the General Data Protection Regulation

This final section is provided for illustrative purposes in order to highlight some of the issues that may arise when US researchers want to use data from other countries for their research. I have used the EU as an example. As this article has shown, the US has a complex maze of laws with limitations and exceptions which often makes researchers and their attorneys want to scream in frustration. The EU data protection laws are just as complex, but unlike the US, the EU has a single data protection regime that applies to all data. The previous law was the Data Protection Directive, [42] which is being replaced (effective May 25, 2018) by the General Data Protection Regulation (GDPR).[43] Like HIPAA, the GDPR has sanctions[44] and special rules for personal data breaches. [45] Unlike HIPAA, it also includes a right to be forgotten, or *erasure,* [46] and there are provisions for the portability of data. [47]

The GDPR was designed to harmonize the different data privacy laws across Europe, give all EU citizens control of how their data is used and protected, and to reshape the way organizations across the region approach data privacy. It is intended to reach beyond the territory of the EU to individuals and businesses that offer goods and services to or monitor the behavior of individuals in the EU. [48] This impacts US researchers because US researchers are increasingly wanting to use EU and other countries' data and are thus dragged into the quagmire that is international data protection law. This section provides an overview of the GDPR. What remains to be seen is what impact the GDPR will have on data repositories, big data research, and personal data in "the cloud" as it evolves over time, especially with the large potential sanctions. Since the GDPR is not yet in effect, it has yet to be interpreted by courts, researchers, and lawmakers throughout Europe.

The GDPR applies to "personal data," which is defines as:

> [A]ny information relating to an identified or identifiable natural person ("data subject"); an identifiable natural

> person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. [49]

This broad definition is likely a moving target. An identifiable person defined as someone who can be identified indirectly will change over time as more information becomes publicly available and as technology changes.[50]

Under the GDPR there is not a "de-identified data" safe harbor or expert determination, but instead data can be anonymized and pseudonymised. When data is anonymized, it is no longer personal data because the individual cannot be identified either directly or indirectly.[51] Pseudonymisation is defined by the GDPR as processing personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information. The GDPR not only requires the "additional information" be stored separately, but also requires various technical and organizational measures to ensure that the personal data cannot be attributed to an identified or identifiable natural person.[52] An example of pseudonymisation is encryption, which renders the original data unintelligible and the process cannot be reversed without access to the correct decryption key. The GDPR requires that this additional information (the decryption key) be kept separately from the pseudonymised data. Although the GDPR encourages the use of pseudonymisation to reduce risks to the data subjects, pseudonymised data is still considered personal data and, therefore, remains covered by the GDPR. [53]

The GDPR also recognizes special categories of personal data which are considered to be particularly sensitive. The processing (or use) of data related to these special categories is generally prohibited by the GDPR:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited." [54]

There are two exceptions to this prohibition that are likely to apply to research involving the special categories of data, which includes health data or protected health information under HIPAA. The first exception requires the data subject give *explicit* consent to the processing of the personal data for one or more specified purposes. [55] The GDPR defines consent of the data subject as "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her." [56] If the consent is provided in a document that concerns other matters, the request for consent must be presented in a way that is clearly distinguishable from the other matters, in an intelligible and easily accessible form, and using clear and plain language. [57] In addition, consent should not be considered freely given if the data subject had no genuine or free choice or is unable to refuse or withdraw consent without penalty.[58] However, the GDPR also recognizes that it is often not possible to fully identify the purpose of personal data processing for research purposes at the time of data collection; therefore, data subjects should be allowed to provide their consent to certain areas of research or parts of research projects when in keeping with recognized ethical standards for scientific research.[59] If a researcher wants to reuse the data for research (e.g., secondary research) and does not have explicit consent for the secondary use, then the researcher would decide if the use was comparable to the previously consented use. The "compatibility test" looks at the following factors:

- any link between the purpose(s) for which the personal data was collected and the purpose(s) of the intended use;
- the context in which the personal data was collected, in particular the relationship between data subjects and the controller;
- the nature of the personal data, in particular are there special categories of personal data or personal data related to criminal convictions and offences;
- the possible consequences of the intended further processing for data subjects; and
- whether there are appropriate safeguards. [60]

The second exception is for research purposes that meet the requirements for applicable safeguards outlined in Article 89(1)[61] and based on EU or an EU country's law that is proportionate to the research aim pursued, respect the right to data protection, and provide for suitable and significant measures to safeguard the fundamental rights and interests of the data subject.[62]

For consent to be "waived" the research should have adequate safeguards in place to protect the data subject's information and have a valid research purpose. For secondary research the secondary use must meet the requirements for applicable safeguards outlined in Article 89(1) and be compatible with the initial purpose for collecting the data ("purposes limitation").[63] If personal data has not been provided by the data subject (e.g., secondary research), then unless the exception for research is met, the data subject should be provided with the following: the identity and the contact details of the controller and where applicable the data protection officer and the controller's representative; the purposes for processing the data and the legal basis for the processing; the categories of personal data concerned; the data recipients or categories of data recipients; and where applicable, that the controller intends to transfer personal data to a recipient in a third country or international organization, whether there is an adequacy decision by the Commission, or if the transfer is made subject to appropriate safeguards.[64] The research exception for this requirement is if the provision of the information would be impossible or involve a disproportionate effort and likely render impossible or seriously impair the research. Then, subject to the conditions and safeguards referred to in Article 89(1), the researcher must take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available.[65]

Conclusion

Research and technology are moving forward at an incredible pace. Technology has enabled researchers to store, manipulate and calculate data in new ways, which has created benefits and risks for researchers and data subjects. The US has a patchwork of laws to address the use of data in clinical research and data repositories, but there are some gaps. Further, as the world of research has gotten smaller and data is shared around the globe, "big data" and research has never been more complicated. US laws like HIPAA and the Common Rule complement and contradict other laws like the EU GDPR. So, researchers who use data from multiple countries must navigate not only their own country's laws but also international legal waters often without a clear path.

Endnotes

[1] An example of a data repository falling though a gap in the patchwork of laws is a commercial pharmaceutical company that has a data or biospecimen repository that uses the repository for internal non-funded research, which would not be subject to HIPAA (not a covered entity), the Common Rule (not funded or supported by one of the eighteen Federal agencies or department), or the FDA Regulations (not submitted to the FDA).

[2] A human subject is currently defined as "A living individual about whom a researcher: (1) Obtains data through intervention or interaction with the individual; or (2) Obtains identifiable private information." 45 CFR § 46.102(f). In the revised Common Rule a human subject is defined as "A living individual about whom a researcher: (1) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or (2) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens." Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7260 (Jan. 19, 2017) (to be codified at 45 CFR § 46.102(e)).

[3] 45 CFR § 46.101(a); *see also* Department of Health and Human Services, *Federal Policy for the Protection of Human Subjects ("Common Rule")* https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

[4] 21 CFR § 50.1(a).

[5] 21 CFR § 56.111(a)(7) and 45 CFR § 46.111(a)(7). One of the changes to the Common Rule was the statement that the Secretary of the Department of Health and Human Services after consulting with the Office of Management and Budget's privacy office and other Federal department and agencies that have adopted the Common Rule will issue guidance to assist IRBs in assessing adequate provisions to protect the privacy of research subjects and the confidentiality of data. Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7264 (Jan. 19, 2017) (to be codified at45 CFR § 46.111(a)(7)(i)).

[6] 45 CFR § 46.116(a)(5).

[7] 45 CFR § 46.101(b)(4).

[8] *Id.*

[9] 45 CFR § 160.103.

[10] Department of Health and Human Services, *Research Repositories, Databases, and the HIPAA Privacy Rule*, (January 2004), https://privacyruleandresearch.nih.gov/research_repositories.asp.

[11]45 CFR 164.514(b). *See also* Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, November 26, 2012, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#_edn1.

[12] Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, November 26, 2012, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#_edn1.

[13] 45 CFR § 164.514(b)(2)(ii).

[14] 45 C.F.R. § 164.514. *See also* Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#_edn1.

[15] 45 CFR 164.514 § (e)(2).

[16] National Institutes of Health, *How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?*, February 2, 2007, https://privacyruleandresearch.nih.gov/pr_08.asp.

[17] 45 CFR § 164.514(e)(4).

[18] For a description of types of data repositories, *see generally*, Amy Jurevic Sokol, *Big Issues in Big Data: Considerations for Research within Large Networks*, AHLA CONNECTIONS, August 2017, at 34.

[19] 45 CFR § 164.501.

[20] 45 CFR § 164.501

[21] Department of Health and Human Services, *Research Repositories, Databases, and the HIPAA Privacy Rule*, (January 2004), https://privacyruleandresearch.nih.gov/research_repositories.asp.

[22] 45 CFR § 164.512(i)(1)(i).

[23] 45 CFR § 164.512(i).

[24] 45 CFR § 164.512(i)(2)(ii).

[25] In addition, the IRB may determine in specific situations that the FDA Rules apply. Previously, the FDA did not have the statutory authority to permit an IRB to waive the informed consent requirements; however, an amendment to the Federal Food, Drug and Cosmetic Act has provided FDA with authority to permit an exception from informed consent for minimal risk clinical investigations when specific criteria are met. Food and Drug Administration, *IRB Waiver or Alteration of Informed Consent for Clinical Investigations Involving no More than Minimal Risk to Human Subjects, Guidance for Sponsors, Investigators, and institutional Review Boards*, July 2017, https://www.fda.gov/downloads/Regulatory-Information/Guidances/UCM566948.pdf.

[26] 45 CFR § 46.102; *also see* Department of Health and Human Services, Office for Human Research Protections, *Guidance Coded Private Information or Specimens Use in Research*, October 16, 2008, https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html.

[27] 45 CFR § 46.102(f) (private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects); *see also* Department of Health and Human Services, Office for Human Research Protections, *Guidance Coded Private Information or Specimens Use in Research*, October 16, 2008, https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html.

[28] 45 CFR § 46.116(d).

[29] 45 C.F.R. § 46.101 (l)(4).

[30] Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149; 7266 (Jan. 19, 2017) (to be codified at 45 CFR § 46.104(d)(4)).

[31] Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149; 7266 (Jan. 19, 2017) (to be codified at 45 CFR § 46.116(d)). It is important to note that if an individual has declined to provide broad consent for the storage, maintenance, and secondary research use of identifiable private information or biospecimens, an IRB may not waive consent. *Id.*

[32] *Id.* One of the following must be included: "(i) A statement that identifiers might be removed from the identifiable private information or identifiable biospecimens and that, after such removal, the information or biospecimens could be used for future research studies or distributed to another investigator for future research studies without additional informed consent from the subject or the legally authorized representative, if this might be a possibility; or (ii) A statement that the subject's information or biospecimens collected as part of the research, even if identifiers are removed, will not be used or distributed for future research studies." *Id.*

[33] Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7266 (Jan. 19, 2017) (to be codified at 45 CFR § 46.116(c)(7)). The statement must also include whether the research subject will share in the commercial profit.

[34] Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7266 (Jan. 19, 2017) (to be codified at 45 CFR § 46.116(c)(9)).

[35] Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 7149, 7267 (Jan. 19, 2017) (to be codified at 45 CFR § 46.116(g)). One of the following "(1) The investigator will obtain information through oral or written communication with the prospective subject or legally authorized representative, or (2) The investigator will obtain identifiable private information or identifiable biospecimens by accessing records or stored identifiable biospecimens." *Id.*

[36] National Institutes of Health, *Research Repositories, Databases, and the HIPAA Privacy Rule* (January 12, 2004), https://privacyruleandresearch.nih.gov/research_repositories.asp.

[37] Office for Human Research Protections, *Guidance on Research Involving Coded Private Information or Biological Specimens* (October 16, 2008), https://archive.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm.

[38] *Id.*

[39] *Id.*
[40] *Id.*
[41] 45 CFR § 46.101(b)(4).
[42] Council Regulation (EU) 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995 O.J. (L 281).
[43] Council Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the processing of Personal Data and on the Free Movement of Such Data, and repealing Directive 95/46/EC, (General Data Protection Regulation), 2016 O.J. (L 199) [hereinafter GDPR].
[44] The following sanctions may be imposed: (1) a warning in writing in cases of first and non-intentional non-compliance; (2) regular periodic data protection audits; (3) a fine up to 10,000,000 EUR or up to 2% of the annual worldwide turnover of the preceding financial year in case of an enterprise, whichever is greater; or (4) a fine up to 20,000,000 EUR or up to 4% of the annual worldwide turnover of the preceding financial year in case of an enterprise, whichever is greater. *Id.* at art. 83.
[45] A personal data breach means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed. *Id.* at art. 4(12). Under the GDPR, the Data Controller has a legal obligation to notify the Supervisory Authority without undue delay (within 72 hours) and the reporting requirement not subject to a de minims standard. I*d.* at art. 33. Individuals must be notified if adverse impact is determined, Individuals do not have to be notified if anonymized data is breached or personal data is protected by pseudonymisation techniques like encryption with adequate technical and organizational protection measures. *Id.* at art. 34.
[46] *Id.* at art. 17; there is a research exception to this right as well: "[t]he right to be forgotten will not apply to the extent it is likely to render impossible or seriously impair the achievement of the objectives of the research. The protections articulated in Article 89(1) must be met." *Id.* at art. 17(3(d).
[47] *Id.* at art. 20.
[48] *Id.* at art. 3(2).
[49] *Id.* at art. 4 (1).
[50] John Mark Michael Rumbold & Barbara Pierscionek, *The Effect of the General Data Protection Regulation on Medical Research*, 19 J. MED. INTERNET RES. (February 2017) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5346164/#ref21 *citing* Anonymisation: Managing Data Protection Risk Code of Practice, London: Information Commissioner's Office, (2012), https://ico.org.uk/for-organisations/guide-to-data-protection/anonymisation/webcite. According to various studies the likelihood of re-identification is high if the researcher has the names, zip code and date of birth (Montreal, Canada almost 98% (Khaled El Emam et al., *The Re-identification of Canadians from Longitudinal Demographics*, BMC MED. INFORMATICS & DECISION MAKING (2011), https://bmc-medinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-11-46); the Netherlands more than 99% (*id.*); and United States approximately 87 % (Latanya Sweeney, *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University, Data Privacy Working Paper 3 (2000), https://dataprivacylab.org/projects/identifiability/paper1.pdf)).
[51] GDPR, *supra* note 43, pmbl. 26.
[52] *Id.* at art. 4 (5).
[53] *Id.* at pmbl 26.
[54] *Id.* at art. 9 (1).
[55] *Id.* at art. 4 (2)(a) (unless the EU or the member law provides that the prohibition may not be consent to by the data subject). *Id.*
[56] *Id.* at art. 4(11).
[57] *Id.* at art. 7(2).
[58] *Id.* at pmbl. 42.
[59] *Id.* at pmbl. 33.
[60] *Id.* at pmbl. 50.
[61] "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for

the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in L 119/84 EN Official Journal of the European Union 4.5.2016 order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner." *Id.* at art. 89(1).

[62] "[P]rocessing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject." *Id.* at art. 9(2)(j).

[63] *Id.* at art. 5(b).

[64] *Id.* at art. 14(1).

[65] *Id.* at art. 14 (5)(b).

Hitting the Mark– Facilitating Research Administration to Support the Institutional Strategic Plan

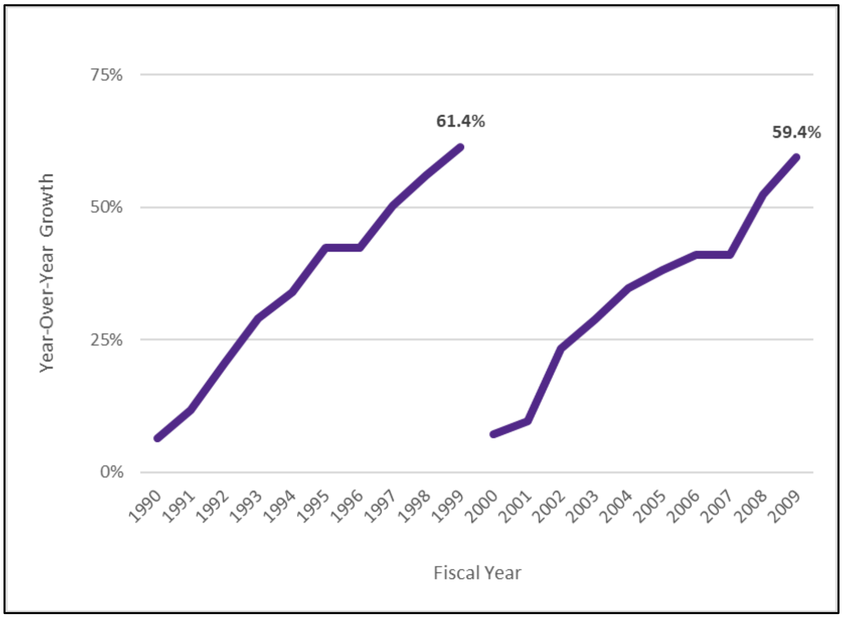
Ian Czarnezki, MBA, Director of Operations, Office of the Vice President for Research, Kansas State University

This paper explores the challenges of facilitating research administration and monitoring progress towards the institutional strategic plan. Kansas State University has a bold vision of being recognized as one the nation's Top 50 Public Research Universities by 2025 (K-State 2025). [1] This bold vision presents a significant challenge for research leadership, how to effectively monitor progress and facilitate growth.

The goal

K-State 2025 identifies two primary sources for assessing our progress towards the research related goals; National Science Foundation Higher Education Research and Development Survey and Arizona State University Center for Measuring University Performance[1].

Figure 1. Kansas State University research expenditures between FY1990 - 2009

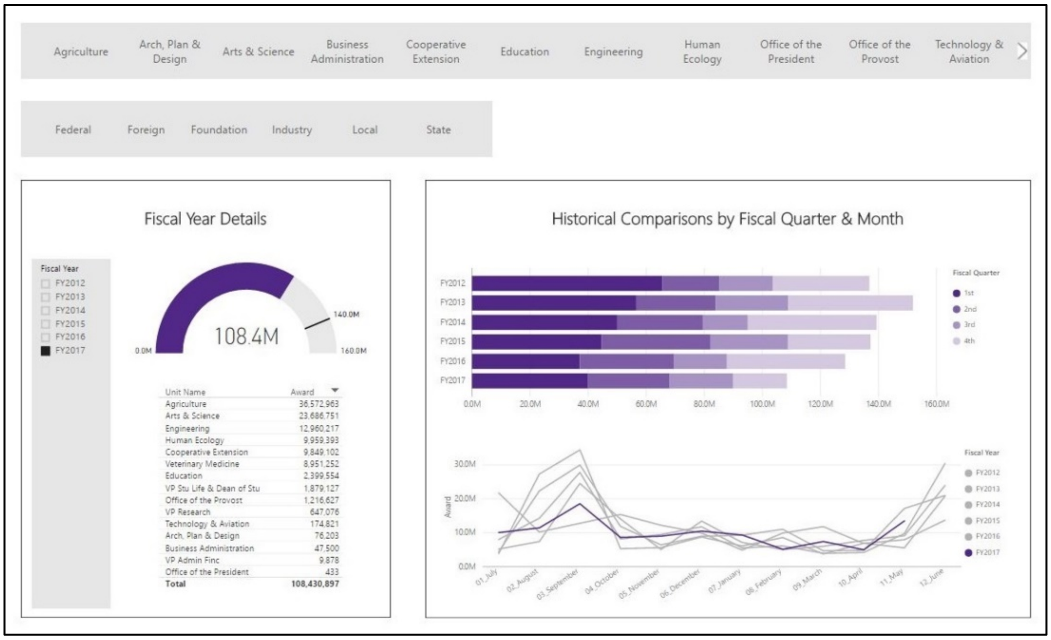


National Science Foundation 1992 Academic S&E R&D Expenditures, Table B-32 [2]
National Science Foundation 1997 Academic R&D Expenditure, Table B-32 [3]
National Science Foundation 2015 Higher Education Research & Development, Table 16 [4]

Our progress

Due to the scope of K-State 2025 we need the ability to understand how each research award impacts our progress. One of the primary challenges to achieving this level of monitoring is integrating the various systems supporting research administration. Kansas State University leverages multiple systems to support our research administration efforts; human resource information system, financial information system, and research administration system. Each of these systems plays an integral role in the overall administration of the research enterprise, unfortunately each system has unique constraints. For example, the human resource information system is bound by the official reporting structure, which presents issues for monitoring multidisciplinary and cluster hire research activities. The financial information system captures transactions associated with research, but is blind to the full research award. In order to accurately assess our progress towards our institutional goals we need to harvest information from each of these disparate systems. Kansas State University has undertaken a reporting initiative to provide a cohesive and timely view of our research activity. Figures 2 and 3 provide an overview of several of the outputs from this initiative.

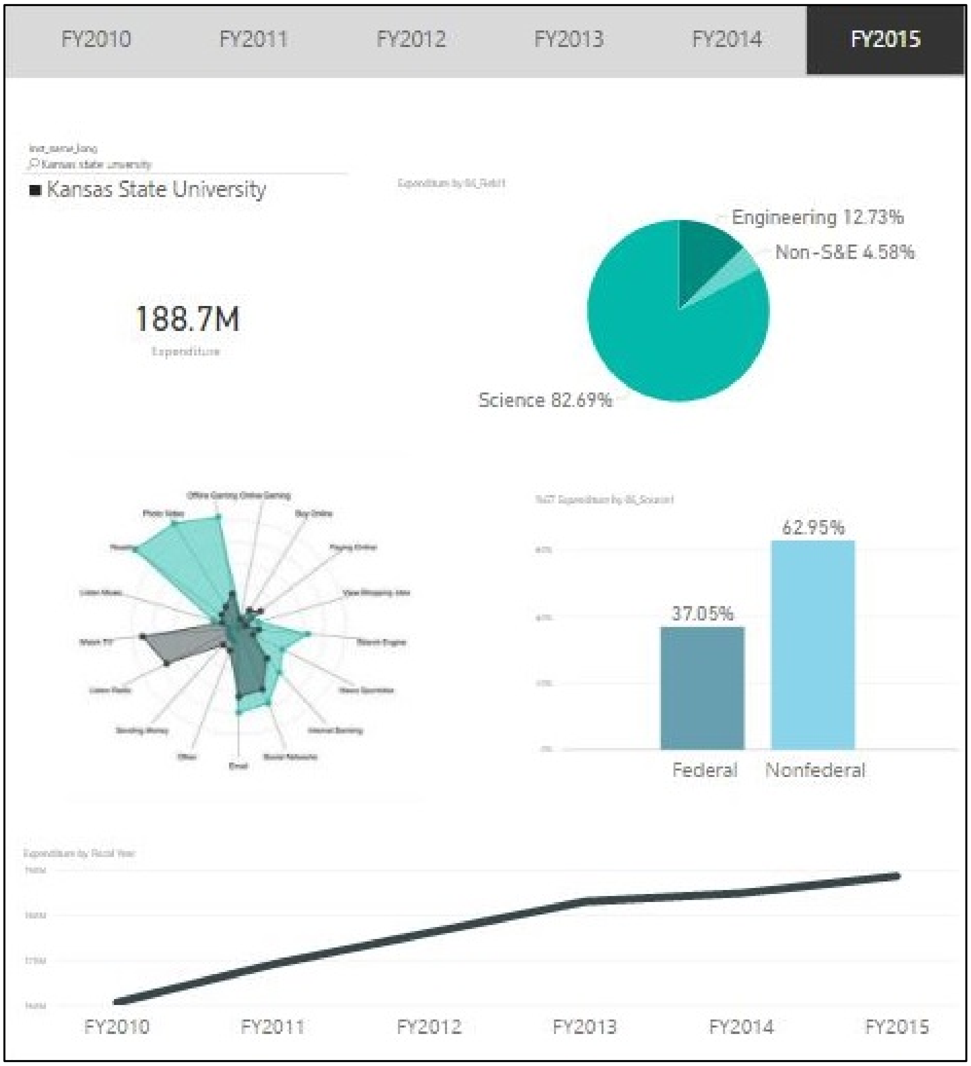
K-State Consolidated Award Tracking System (K-CATS) is our branded research administration business intelligence solution. This solution allows both research leadership and a broad community of stakeholders' insight into our research activities. Figure 2 is screenshot of K-CATS Sponsored Research Award Activity overview, which provides research leadership with near real-time access to award activity and gives greater insight with historical context. In addition the Kansas State Research Awards Dashboards provide a broad audience, including both internal and external stakeholders, with an overview of the current research activities. The dashboards are accessible via the Kansas State University website at http://www.k-state.edu/research/our-research/reports/dashboards.html



Unit Name	Award
Agriculture	36,572,963
Arts & Science	23,686,751
Engineering	12,960,217
Human Ecology	9,959,393
Cooperative Extension	9,849,102
Veterinary Medicine	8,951,252
Education	2,399,554
VP Stu Life & Dean of Stu	1,879,127
Office of the Provost	1,216,627
VP Research	647,076
Technology & Aviation	174,821
Arch, Plan & Design	76,203
Business Administration	47,500
VP Admin Finc	9,878
Office of the President	433
Total	**108,430,897**

Figure 2. K-CATS Sponsored Research Award Activity

Currently underway, the HERD Project is a collaborative effort between Kansas State University and Microsoft. The objective of this collaboration is the development of a reporting solution for the National Science Foundation Higher Education Research Development Survey. The reporting solution will consolidate the survey's data files into a single interactive business intelligence solution thus allowing for great insight into our research activities compared to other institutions. Figure 3 is a draft visualization from the reporting solution which provides an interactive overview of Kansas State University research expenditure profile. Users can interact with the visualization by changing the fiscal year, drill through the disciplines hierarchy (engineering, science, life sciences, etc.) and filter by funding source. It is critical we empower our leadership with as much insight from National Science Foundation HERD Survey as possible, as K-State 2025 identifies it as a primary source of assessing our research activities. Once the project is completed, it is anticipated the reporting solution will be accessible via the Kansas State University website and will be shared with other higher education institutions upon request.



National Science Foundation 2010 – 2015 Higher Education Research and Development [5]

Future efforts

To date, our efforts have been focused on monitoring Kansas State University's performance towards our institutional goals, though our efforts could be leveraged to go beyond just monitoring. Under the leadership of the Vice President for Research, we are utilizing this wealth of information regarding our research activities to align researchers with applicable funding opportunities, highlighting interdisciplinary partners, and move Kansas State University's research efforts forward. Though Kansas State University has set a lofty goal, being recognized as one the nation's Top 50 Public Research Universities by 2025 [1], we have the research talent, leadership, and the insight into our data to meet the bold vision of K-State 2025.

References

[1] http://www.k-state.edu/2025/

[2] https://wayback.archive-it.org/5902/20160210143434/http://www.nsf.gov/statistics/rdexpenditures/dst/

[3] https://wayback.archive-it.org/5902/20160210143410/http://www.nsf.gov/statistics/nsf99336/

[4] https://ncsesdata.nsf.gov/herd/2015/

[5] https://www.nsf.gov/statistics/srvyherd/

Influencing the Culture of Scholarly and Professional Communities to Advance Clinical Research and Accelerate Knowledge Translation

Margaret A. Rogers*, Chief Staff Officer for Science and Research
American Speech-Language-Hearing Association

Michael Cannon, Director, Serial Publications and Editorial Services
American Speech-Language-Hearing Association

*Corresponding author: Margaret Rogers – mrogers@asha.org

Professional and scientific associations that represent health care disciplines have a tremendous opportunity to help shape how evidence-based practice becomes integrated into the fabric of the professions that they support. Associations are membership organizations that are typically *not-for-profit* entities. Although their functions vary, common ways in which associations can bring value to a discipline include (a) advocating; (b) setting and enforcing ethical standards; (c) credentialing individuals; (d) accrediting academic training programs; (e) providing continuing education opportunities; (f) communicating news and other information through websites, magazines, blogs, and other social media; (g) publishing scientific journals; and (h) providing other forms of professional and scientific support to their membership. For associations whose members work in the health care sector, decisions made now about how best to plan and prioritize resources for the changing landscape in health care will likely play a critical role in determining how well these disciplines fare over the next decade. Actions taken by associations can affect the rate at which evidence becomes adopted into practice and, likely, the extent to which these practice changes result in improved patient outcomes. The nature of the efforts that are being marshaled by professional and scientific associations to "bridge the research-to-practice gap" are numerous, but perhaps the most promising among these are efforts that make use of big data, especially when coupled with text and data mining (TDM), semantic computing, and artificial intelligence. In this article, the need for and potential application of these technologies to advance evidence-based practice and health care quality will be described—as well as some of the more persistent and prevalent associated challenges in the adoption and implementation of evidence-based clinical practices.

The *data-driven health care*, *value-based purchasing*, and *pay-for-performance* trends in health care, as well as similar trends in other sectors such as K–12 education, have given rise to many challenges that require strategically focused plans of action. Actions taken or not taken by professional and scientific associations over the next few years may have far-reaching effects on the state and health of the disciplines that they support—and, taken together, the future quality of health care. Although there are undoubtedly many paths that could help health care disciplines and others to transition successfully into becoming more data driven and evidence based, inaction is not likely to be one of those paths. Arguably, a "we'll do what we've always done in the ways we've always done it" response is not a wise approach in positioning a discipline to survive, and potentially thrive, in the context of radically new economic models of health care and education right at our doorsteps. In health care, the daunting complexity of the transition to value-based purchasing coupled with the abounding political uncertainty surrounding the Affordable Care Act (ACA), as well as numerous other policy and regulatory uncertainties, has made it very difficult to plan and act. However, the rapid development and adoption of web and computing technologies have made the range of possibilities for action broader than ever before—and the potential for success greater than ever before.

Drivers of Change for Associations Focused on Health Care

Over approximately the past 75 years, three areas have been evolving that have had formative effects on the priorities of the American Speech-Language-Hearing Association. The first is the evidence-based practice and implementation science movements, which encompass a host of efforts aimed at improving the quality of the scientific evidence, the rate of knowledge translation, and the effectiveness of clinical practice. Under the umbrella of evidence-based practice/medicine, many standards and guidelines have been established to improve the design, conduct, and reporting of clinical trials and to reduce bias at every stage of the scientific process, including in the development of clinical practice guidelines. With the emergence of dissemination and implementation science, a new generation of research is being conducted to bridge the research-to-practice gap.

The second area of rapid change is health care—with specific attention given to the central role that outcomes measurement is playing in shifting to a value-based purchasing economy. According to Porter and colleagues (2012), "if nations place outcome measurement at the top of their reform agendas, they can begin to unlock forces of innovation and improvement that may yet help us realize the dream of financially sustainable, high-quality health care." Their argument, from a business perspective, is that health care in the United States has never benefited from the forces of market competition because data has not been sufficiently available to allow comparisons of providers, facilities, or systems. To shift successfully to a value-based purchasing health care economy, outcome measurement needs to be mandated and the data made publically available. This vision has prompted many efforts to develop outcome measures and other quality indicators. It has also prompted recognition of the need to develop clinical data registries and a host of associated supporting efforts—such as determining which patient characteristics are most important to capture—so that data can be adjusted to enable valid comparisons across providers, facilities, and systems, despite the heterogeneity of their patient populations and facility types.

The third evolution is big data and data science, an area that is redefining scientific publishing, boosting the scientific contribution of clinical data registries and electronic health records, and fueling the data-sharing revolution. A detailed account of the sources of these trends in evidence-based practice, health care, and big data is beyond the scope of this article. However, a discussion of some of the key thought leaders and drivers of change is included, as their contribution helps tell the story of how one scientific and professional association, the American Speech-Language-Hearing Association, is readying itself for the future. From across a broad mix of disciplines—such as epidemiology, business, computer science, medicine, social science, statistics, and more—many individuals and ideas have contributed to the aforementioned drivers of change. However, in our quest to improve the human condition through informed decision making based on trustworthy evidence, perhaps the largest game-changer will be the data itself as a contributing form of intelligence.

Historical Roots of Evidence-Based Practice and Data-Driven Outcomes Improvement

In the Introduction to his seminal monograph, *Effectiveness and Efficiency: Random Reflections on Health Service* (1972), Archibald Cochrane confessed that he long held the belief that "all effective treatments must be free" (p. 3). In fact, he stated that he originally wrote that slogan to be displayed on a banner at a rally that he attended when the idea of a National Health Service (NHS) in Great Britain was being debated. During his years as a physician and a prisoner of war (POW), he observed that a large number of POWs with serious, life-threatening conditions survived incarceration without having received medical care. This observation and his experiences in epidemiology led him to the conviction that we need to be able to measure, without bias, the effects of any particular medical action on altering the natural course of health conditions. He termed this type of knowledge "effectiveness," and his work led to the preeminence of the randomized controlled trial for establishing the efficacy of a clinical intervention. He also saw the necessity of being able to account for the costs of management, personnel and materials, and length of stay, among other things, to conduct realistic cost/benefit analyses of care. He termed this dimension "efficiency." His experiences with the inequities of health care associated with poverty led him to herald another "E"—"equality"—which he declared as the third "yardstick" by which the success of the NHS should be measured. These terms have served as a compass ever since, guiding the evolution of health care policy, health care systems, analytic frameworks, and measure development in many countries—and they continue to guide discussions about how to improve health care quality in the United States. As one prominent example of this legacy, the terms *effective*, *efficient*, and *equitable* are now three of the six dimensions that the Institute of Medicine (2011a) recommends as principles to guide quality improvement efforts in health care and that the Agency for

Healthcare Research and Quality (AHRQ, n.d.) endorses as the analytic framework for quality assessment in health care. (The other three domains are *safe*, *patient-centered*, and *timely*.)

As these six quality domains became a cornerstone for assessing the return on investment for health care expenditures in the United States, organizations such as The Commonwealth Fund began to monitor health care quality across economically comparable nations. The Commonwealth Fund supports independent research that compares health care quality and expenditures across high-income countries. In the most recent report, *Mirror, Mirror 2017: International Comparison Reflects Flaws and Opportunities for Better U.S. Health Care*, 72 indicators designed to measure performance across five domains—Care Process, Access, Administrative Efficiency, Equity, and Health Care Outcomes—were compared across 11 high-income nations. As has been consistently the case, the United States ranked lowest with respect to overall quality and ranked poorest in three of the five domains—namely, Access, Equity, and Health Care Outcomes. Furthermore, the report stated that "the United States has the highest rate of mortality amenable to health care and has experienced the smallest reduction in that measure during the past decade" (p. 3). Of the 11 high-income nations included in the study, the United States is the only country lacking universal health insurance coverage. Including both private and public expenditures on health care, the United States spends nearly 17% of its gross domestic product (GDP) on health care. The country with the next highest percent of GDP spent on health care was Switzerland (11.4%), and Australia spent the least (8.8%) of the 11 countries included in the comparison. According to another 2017 report from The Commonwealth Fund, *Aiming Higher: Results from a Scorecard on State Health System Performance*, "nearly all state health systems improved on a broad array of health indicators between 2013 and 2015. During this period, which coincides with the implementation of the ACA's major coverage expansions, uninsured rates dropped and more people were able to access needed care, particularly those in states that expanded their Medicaid programs" (p. 3). Thus, despite the current political uncertainties regarding the fate of the ACA, as of October 2017 its implementation has been associated with significant improvements in the quality of and access to health care in the United States. That these authoritative data are not currently having a more influential effect on the policy makers who aim to dismantle the ACA—let alone on the millions of voters who stand to gain (or lose) so much pending the outcome of this debate—is a puzzling, yet somewhat predictable, phenomenon. It has become a common observation across many sectors that, even in situations where there is a strong body of compelling scientific evidence, behavior and attitudes do not necessarily change accordingly—and if they do, it is quite likely that the rate of such change will be slow and will spread only incrementally throughout a population.

In 1962, Everett Rogers published the first edition of his first book, *Diffusion of Innovations,* and coined this term to describe similar phenomena wherein new technologies, products, scientific findings, or other innovations are gradually adopted by a population. As a doctoral student studying rural agricultural society at Iowa State University in the 1950s, Rogers became interested in modeling the usage patterns and rate at which Iowan farmers were adopting a new weed spray. The *diffusion of innovations theory* is broad enough to model how new ideas, products, scientific findings, and other innovations gain momentum and spread through a population or social system—or fail to catch on at all. In the theory, Rogers describes five different categories of adopters—innovators, early adopters, early majority, late majority, and laggards—and hypothesizes that the strategies most likely to promote adoption will vary based on the adopter category. In the fifth edition of his book in 2003, he put forth the innovation-decision process model and proposed five factors that shape the rate and likelihood of adoption—relative advantage, compatibility, complexity, trialability, and observability. His social science theory has been applied in many other fields—including communications, criminal justice, marketing, public health, and social work—and it has been particularly influential in dissemination and implementation science.

In 1984, Prochaska and DiClemente proposed the *transtheoretical model of change,* which expands on Everett's model of adoption to account for why people cease behaviors (or not). The slow rate at which Americans were ceasing to smoke cigarettes by the 1980s was unanticipated given the large body of convincing medical research linking smoking behavior to cancer and the omnipresence of antismoking campaigns. Prochaska and DiClemente were motivated to understand what led to the success of some people and yet the failure of others in smoking cessation. Their research led them to develop a six-stage model of change through which individuals, or social groups, might progress: precontemplation, contemplation, preparation, action, maintenance, and termination. The notion of readiness for change is inherent in this model in that the actions associated with changing one's behavior are unlikely to occur if the person has not yet contemplated and prepared for the change.

Other theories of change that model the effects of the social environment and interpersonal influences also contribute to our understanding of how provider behaviors and attitudes might be influenced to promote the adoption of evidence-based practices. For example, consider two prominent theories highlighting the role that peer influence plays on individual decision making and behavioral change—the *social norms theory* developed by Perkins and Berkowitz (1986), which was developed to understand and reduce alcohol consumption and alcohol-related injury on college campuses, and the *social network theory* developed by Wasserman and Faust (1994). Social network approaches to changing behavior are showing promise as an effective way to help clinical providers consume clinical research and adopt evidence-based practices. Models of change may be helpful to advance our understanding of why the mere publication of scientific knowledge does not automatically or quickly result in behavioral and attitudinal change, regardless of whether the targeted change entails adopting, stopping, or preventing behavior. These models may also be helpful in identifying successful strategies for bridging the research-to-practice gap.

The Emerging Science of Dissemination and Implementation

Identifying the factors that influence the diffusion of innovations in science and medicine to practice and policy is a growing focus of research often referred to as *dissemination and implementation science*. This research seeks to narrow the gap between the discovery of new knowledge and its application by identifying the factors that influence the pattern and rate of change—and, ultimately, the maintenance of that change in people or systems. Estimates of the length of time that it takes for research to become translated into evidence-based policies, programs, and practices hovers between 15 and 20 years. For example, Morris, Wooding, and Grant (2011) reviewed the literature that attempts to quantify the time lag in the health research translation process. The title of their article sums up their findings: "The Answer Is 17 Years, What Is the Question: Understanding Time Lags in Translational Research." They modeled the stages from innovation at the basic science stage, to testing with human/clinical trials, to guidelines development, to adoption in clinical practice—a process that took an average of 17 years based on the literature that they reviewed. Their findings are similar to the lag time reported in other dissemination and implementation research (see, e.g., Wolff, 2008). For both humane and economic reasons, Morris and colleagues (2011) argue the importance of speeding up this process, and there is now an emerging consensus that evidence-based practices need to be more rapidly integrated into clinical care. However, translating science into clinical application is proving to be a very challenging process; the reality that failure is perhaps more common than success has led to comparing the bench-to-bedside translation process to crossing the "valley of death" (Meslin, Blasimme, & Cambon-Thomsen, 2013, p. 1).

From the inception of a research question to its dissemination and implementation stages, understanding the strategies and factors that can promote or hinder the uptake of new knowledge has become a central focus in health research. For example, in 1995, Andrew Oxman and colleagues published an article in which, once again, the title conveys the findings—*No Magic Bullets: A Systematic Review of 102 Trials of Interventions to Improve Professional Practice*. They searched the literature for studies that examined the effectiveness of dissemination and implementation strategies on changing behavior among health care providers. To be included in their analysis, the study had to report the outcomes of the implementation "intervention" using objective assessments of provider performance in the health care setting. The intervention types included educational materials, conferences, outreach visits, presentations by local opinion leaders, audit and feedback, reminders, marketing materials, local consensus processes, and more. They observed that "interventions to improve professional performance are complex" (p. 1427) and that complex interactions likely occur among "the characteristics of the targeted professionals, the interventions studied, the targeted behavior, and the study design" (p. 1427). Although no one approach stood out as a clear front-runner (i.e., there was "no magic bullet"), they concluded that using a combination of dissemination approaches—perhaps the most important of which include social learning opportunities (e.g., journal groups)—"could lead to substantial improvements in clinical care derived from the best available evidence" (p. 1427). To achieve the goal that Cochrane (1972) originally inspired of better health through informed decisions based on trusted evidence, we will need to understand why, as Colditz (2012) put it, "discovery alone does not lead to use of knowledge; evidence of impact does not lead to uptake of new strategies; organizations often do not support the culture of evidence-based practice; and maintenance of change is often overlooked, leading to regression of system-level changes back to prior state" (p. 7).

Over the past three decades, quality metrics concerning the design, conduct, and reporting of clinical trials, evidence syntheses, and clinical practice guidelines have been developed. The unwavering aim of these metrics has been (a) to decrease the risk of bias and (b) to better ensure that research is reported and, ideally, conducted to maximize its usefulness for replication efforts and evidence syntheses. For the most part, the scientific community is welcoming the emergence of standardized reporting guidelines because stakeholders are recognizing that many of the key details needed for developing systematic reviews and conducting replication studies are not routinely reported in the clinical research literature (see, e.g., Hoffman, Chrissy, & Glaziou, 2013). As can be seen on the EQUATOR Network[1] website (http://www.equator-network.org/), more than 400 reporting guidelines have been developed for many types of clinical research, including randomized controlled trials (e.g., *Consolidated Standards of Reporting Trials* (CONSORT) – http://www.consort-statement.org/) as well as for the development of evidence-based systematic reviews (e.g., *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) – http://www.prisma-statement.org/); and evidence-based clinical practice guidelines (e.g., *Appraisal of Guidelines for Research and Evaluation* (AGREE) – http://www.agreetrust.org/). With these developments, a level of consistency and rigor has been introduced that furthers the objectivity and value of the scientific process—an outcome that benefits researchers, students, journal publishers, editors, reviewers, health care providers, and, most important, patients. To fulfill the promise of evidence-based practice, however, there continues to be a dire need to increase the relevance of clinical research, the quality of scientific reporting, the readiness of providers, and the capacity of health care systems to support the implementation of evidence-based practices.

[1] The UK EQUATOR Centre is hosted by the Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford.

A key component of translating research to practice has been the development of evidence-based clinical practice guidelines. There are numerous definitions and uses of the term *guidelines*, including consensus guidelines, clinical guidance, quality statements and standards, clinical pathways, and evidence-based clinical practice guidelines. For most of the 20th century, practice guidelines were developed by professional societies, including the American Speech-Language-Hearing Association, based on the expert opinions of a selected subset of members. As payers became involved, they began to look to guidelines as a basis for making coverage decisions. Once payers themselves got into the guidelines business, those that they developed had unrealistically high evidence standards that generally could not be met. Many feared that payer-developed guidelines would be used mainly to absolve payers of the responsibility to cover all interventions except those that were most thoroughly researched and validated. The disconnect between guidelines developed by professional societies versus those developed by payers made it clear that a set of standards for guidelines development and quality appraisal was needed so that the same methods and quality standards could be applied regardless of the sponsoring entity. According to the Institute of Medicine (1990, p. 1), in 1989, Congress amended the Public Health Service Act to create the Agency for Healthcare Research and Quality (AHRQ; originally called the Agency for Health Care Policy and Research), largely to address this disconnect. Under the terms of Public Law 101-239, AHRQ "has broad responsibilities for supporting research, data development, and other activities that will enhance the quality, appropriateness, and effectiveness of health care services. The creation of a practice guidelines function within [AHRQ] can be seen as part of a significant cultural shift, a move away from unexamined reliance on professional judgment toward more structured support and accountability for such judgment. Reflecting the first element of this shift, guidelines are intended to assist practitioners and patients in making health care decisions; reflecting the second aspect, they are to serve as a foundation for instruments to evaluate practitioner and health system performance" (p. 2-3).

In 2003, the era of consensus-based guidelines began a steady decline as the AGREE instrument for evaluating the process of practice guideline development and the quality of reporting was born. The Institute of Medicine (2011a) defined *clinical practice guidelines* as "statements that include recommendations, intended to optimize patient care, that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options" (p. 1). A variety of stakeholders have used different methods to develop guidelines across different organizations and countries, but the Scottish Intercollegiate Guidelines Network (SIGN), the National Institute of Clinical Excellence (NICE), and Guidelines International (GIN) are the primary entities setting standards and providing tools for developing evidence-based clinical practice guidelines. As mentioned above, AGREE is the most widely accepted set of standards by which the quality of clinical practice guidelines is evaluated. In addition, developers of clinical practice guidelines often use the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to develop recommendations based primarily on findings from evidence-based systematic reviews (http://www.gradeworkinggroup.org/). The purpose of GRADE is to evaluate the quality of evidence and the strength of the proposed recommendations in a clinical practice guideline. GRADE uses five criteria to evaluate the quality of evidence: risk of bias, consistency of evidence, directness of evidence, precision in results, and publication bias. Based on these criteria, the overall quality of evidence is ranked and categorized into four levels. According to Langhoff-Roos and Shah (2016), the key concept behind the strength of recommendation in the GRADE approach is "the extent to which one can be confident that the desirable consequences of an intervention outweigh its undesirable consequences" (p. 844). All of the efforts that have been devoted to establishing quality standards for clinical research methodology, for reporting clinical research, for developing evidence-based systematic reviews, and for creating evidence-based clinical practice guidelines are moving us closer to bridging the research-to-practice gap. Unfortunately, the relevance of much of the biomedical and behavioral sciences research with respect to its applicability to clinical practice continues to be a very challenging shortcoming (Colditz, 2012).

The Promise of Clinical Data Registries as Learning Health Care Systems

Clinical data registries hold much promise to fill in some of the gaps left unaddressed in the extant body of clinical research that has been initiated by individual scientists or teams of investigators. That *investigator-initiated* research may leave knowledge gaps is not surprising. Researchers necessarily focus deeply on specific topics and, understandably, may give less attention to how their research may be implemented in clinical practice or used to influence policy. Even for interventions whose effectiveness has been established with randomized controlled trials, typically little will be known about how well the intervention will work when applied to a wider range of patients and different settings than those included in controlled studies. Clinical data registries may help to fill in gaps in the body of investigator-initiated research. Moreover, because clinical data registries and electronic health records accumulate large samples of heterogeneous patient populations, there are questions that are arguably best addressed through big data and data science. Information about patient characteristics, the interventions applied, the outcomes of care, and the resources utilized can be amassed in clinical registries and analyzed to address the multifaceted question of "What works best for whom under which circumstances?" (see, e.g., Paul, 1967). This concept is described by the Institute of Medicine's *Learning Health System Series* (2011b). The report states that "health and health care are going digital. As multiple intersecting platforms evolve to form a novel operational foundation for health and health care—the nation's digital health utility—the stage is set for fundamental and unprecedented transformation. Most changes will occur virtually out of sight, and the pace and profile of the transformation will be determined by the stewardship that fosters alignment of technology, science, and culture in support of a continuously learning health system. In the context of growing concerns about the quality and costs of care, the nation's health and economic security are interdependently linked to the success of that stewardship. Progress in computational science, information technology, and biomedical and health research methods have made it possible to foresee the emergence of a learning health system that enables seamless and efficient health delivery of best care practices and the real-time generation and application of new knowledge. Increases in the cost and complexity of care compel such a system" (p. 1).

Central to the goals of cost containment and budgeting health care expenditures is the ability to predict that, based on the incidence of a given condition, a given number of individuals will need a given regiment of treatments at an estimated annual cost. At a minimum, four data domains are needed to inform predictive models of health care expenditures and to serve as a learning health care system, including information about patients (case-mix data), the services and interventions provided, the outcomes of care, and an accounting of the resources utilized to achieve these outcomes. Payment systems based on predictive modeling need first to be able to predict, based on historical data, the rehabilitation potential of individual patients and then, based on evidence about which approach to intervention is likely to be most effective for a given type of patient in a given circumstance, predict the resources that will be required to achieve pre-specified rehabilitation goals. According to Massof (2010), the primary goal for such modeling is to predict the probability of obtaining a specified outcome given a mix of patient-specific factors and the choice of interventions. As patient factors mediate the effects of services on outcomes, case-mix–adjusted data are necessary to compare the effectiveness of different interventions and to model the costs and benefits of providing services for specific patient groupings.

With clinical data registries and data science added to the arsenal of clinical research methods, we can move beyond asking, "What works?" to addressing, "What works best for whom under which circumstances?" As put forth by Rogers and Mullen (2012), it is unlikely that we will be able to address this multifaceted question solely through investigator-initiated research. They assert that large-scale data collection instruments that amass information across each of the four data domains mentioned above are necessary so that adequately powered analyses can be conducted to identify the effects that case-mix and service delivery factors have on patient outcomes. The vision of a learning health care system is that analyses of large clinical data repositories will provide information about what works best for whom under which circumstances. Based on these analyses, decisions will be made about how services and outcomes might be improved.

After providers and systems make targeted adjustments, the newly collected data will provide information about the relative success and failure of the adjustments. Within the context of a learning health care system, data collection, analysis, change, and learning are cyclical, iterative stages. Learning health care systems are anticipated to accelerate the rate at which evidence-based practices and innovations in health care become adopted; thereby, reducing the research-to-practice gap.

Data-Driven Publication Strategies for Straightening the Path to Implementation

As noted previously, the gap between discovery of new knowledge and its implementation can be quite large, and the potential barriers to implementation can be so varied that there is no magic bullet for how to infuse evidence-supported changes into practice across the board. Although the optimal path to any given implementation may be unclear, there are a number of ways that professional and scientific associations can leverage their innate strengths to increase the precision and efficiency of implementation strategies. This is particularly true now because of the shift to a much more data-based and data-driven journal publishing enterprise that has unfolded over the past two to three decades.

Whereas the initial three centuries of journal publishing involved some structured data in the sense of the articles themselves having a defined, sectioned format and both the articles and the journals having bibliographic elements sufficient for tracking and organization of the collection of outputs over time, it was not until the online publishing era that true structuring began to be put in place. Covering that evolution could itself fill may pages—and has (see, e.g., Ware & Mabe, 2015)—but a key milestone in that evolution was the relatively recent shift to the article, rather than the journal, being the "unit."

In an article-based economy, the entire publishing workflow changes. Rather than hold articles for publication until, say, the next quarterly issue of a clinical practice journal, publishers (including the American Speech-Language-Hearing Association) have increasingly adopted continuous publishing models so that important research can be released as soon as it is ready to be read and used. In practical terms, this shaves off a period of time, often up to several months, from the overall amount of time that it takes to get new knowledge into implementation. With a timeline of 17 years associated with implementation, every such reduction counts.

Because of the focus on article-based publication, publishers have sought competitive advantage in time-to-market, and that has led to more radical reductions in the amount of time from the report of findings to the availability of that information in the environment where application of it can be made. Standardization at the XML markup level is now more typically relied on from submission all the way through to the production and publication of research. And it is increasingly driving a shift in the authoring process, as well. In aggregate, the standardization of the data behind the full range of the publication steps has certainly shaved off another several months to a year or more from the time it takes for new knowledge to get into discovery and use.

Likewise, concomitant with this now accelerated flow of new knowledge into the discovery environment, the same data standards underpinning its processing and production are also driving the means of discovery itself. In the era of the issue-based publication model, massive "stacks" of information developed and were archived in libraries and, eventually, stored in online repositories and aggregations, both accessible through a search approach where success was as variable as the searcher's facility with query construction. In the semantic publishing era, very granular tagging is applied to articles so that a publisher's platform can display articles in automatically curated collections by topic and can dynamically link related articles, thereby extending a user's discovery. This, then, shaves additional time from the discovery phase of acquiring and applying new knowledge, although it would be hard to quantify the level of reduction in any meaningful way, given the breadth of disciplines and pace of development of research lines. In aggregate, savings here may more realistically be considered a factor mitigating against the sheer difficulty of coping with the greatly increased flow of new knowledge into the system, as made possible by the previously covered advances.

More important, the opportunities afforded by semantic tagging allow the publisher to offer a well-organized, highly accessible library of information versus a merely browsable online collection of articles that had been produced in print. In the article-based economy, with semantic data better characterizing the articles and being used to drive additional pathways into the stacks and away from arrived-at articles to other articles that otherwise might not have been found, the deeper opportunity now exists for such publishers to better assess the actual patterns of knowledge acquisition or attempted knowledge acquisition—and to marry up those patterns with future curation and promotion activities. The opportunity is thus greater than ever for the society publisher to be a valued connector in the process of bridging research to practice, by virtue of that central, high-value information resource.

In addition to semantic data and the opportunities presented by it, the online article-based economy and its prevailing standards related to the objects themselves, such as the digital object identifier (DOI), have driven significant advances in other forms of data, such as attention-level metrics. A user can now rely on altmetrics, for example, to track how research has been used so far and by whom. For a publisher, imagine the value of that type of multifaceted data layered onto semantic usage data. Just over 10 years ago at the American Speech-Language-Hearing Association, for comparison purposes, virtually none of that data existed. Issues were produced and shelved at libraries or physically placed in mailboxes, and citation data years down the line were the only approximate gauge of how published research was being used.

This is truly the revolutionary point at which the gap between output (research) and use (practice) can begin to be more effectively observed and appraised. On the near-term horizon (and, in some cases, already in place) are advances such as text and data mining layers in publishing platforms, so that automated analyses of the nature of patterns of research (including populations addressed, techniques used, and types of data derived) can be performed vastly more quickly and precisely. In addition, systems providing predictive analytics on citation likelihood are now being put into place. At the user level, the now more commonly in place HTML5 web standards are allowing for greater annotation of content and incorporation of it into shared learning activities.

All of these advances are emblematic of the tide of big data flooding all publishers. For the society publisher, deriving intelligence from that flow and leveraging it alongside programmatic efforts aimed at bridging the research-to-practice gap amounts to a fundamentally different business than what scholarly publishing formerly offered an association. The publishing enterprise within associations must be retooled accordingly, with a greater emphasis placed on strategizing with editors, working more directly with authors, and focusing on different domains—away from day-to-day production and toward a fuller embrace of science writing, content curation, and well-developed curation and promotion skills, especially those focused on support of learning-based usage contexts. If the past two decades have been any guide, the next two should lead to measurable improvements in reducing the gap from research to practice.

References

Agency for Healthcare Research and Quality. (n.d.). The six domains of health care quality. Rockville, MD: Author. Retrieved from http://www.ahrq.gov/professionals/quality-patient-safety/talkingquality/create/sixdomains.html

Colditz, G. A. (2012). The promise and challenges of dissemination and implementation research. In R. C. Brownson, G. A. Colditz, & E. K. Proctor (Eds.), *Dissemination and implementation research in health: Translating science to practice* (pp. 3–22). New York, NY: Oxford University Press.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*. London, England: Nuffield Trust.

Hoffman, T. C., Chrissy, E., & Glaziou, P. P. (2013). Poor description of non-pharmacological interventions: Analysis of consecutive sample of randomised trials. *The BMJ*, *347*, f3755.

Institute of Medicine. (1990). *Clinical Practice Guidelines: Directions for a New Program*. Marilyn J. Field and Kathleen N. Lohr, Editors; Committee to Advise the Public Health Service on Clinical Practice Guidelines. Washington, DC: The National Academies Press. [Available from: http://www.guidelines-registry.cn/uploadfile/2016/0914/20160914112506974.pdf

Institute of Medicine of the National Academies. (2011a). *Clinical practice guidelines we can trust*. Washington, DC: The National Academies Press.

Institute of Medicine of the National Academies. (2011b). *Digital infrastructure for the learning health system: The foundation for continuous improvement in health and health care* [Workshop series summary]. Washington, DC: The National Academies Press.

Langhoff-Roos, J., & Shah, P. S. (2016). Evidence-based guidelines and consensus statements. *Acta Obstetricia et Gynecologica Scandinavica, 95,* 843–844.

Massof, R. W. (2010). A clinically meaningful theory of outcome measures in rehabilitation medicine. *Journal of Applied Measurement*, *11*(3), 253-270.

McCarthy, D., Radley, D. C., & Hayes, S. L. (2015, December). *Aiming higher: Results from a scorecard on state health system performance*. New York, NY: Author. Retrieved from http://www.commonwealthfund.org/publications

Meslin, E. M., Blasimme, A., & Cambon-Thomsen, A. (2013). Mapping the translational science policy 'valley of death'. *Clinical and Translational Medicine*, *2*(1), 1–8.

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*, 510–520.

Oxman, A. D., Thomson, M. A., Davis, D. A., & Haynes, R. B. (1995). No magic bullets: A systematic review of 102 trials of interventions to improve professional practice. *Canadian Medical Association Journal*, *153,* 1423–1431.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*, 109–118.

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International Journal of the Addictions, 21,* 961–976.

Porter, M. E., Larsson, S., & Ingvar, M. (2012). A new initiative to put outcomes measurement at the center of health reform. Retrieved from http://healthaffairs.org/blog/2012/10/31/a-new-initiative-to-put-outcomes-measurement-at-the-center-of-health-reform

Prochaska, J. O., & DiClemente, C. C. (1984). *Transtheoretical therapy: Crossing traditional boundaries of therapy.* Homewood, IL: Dow Jones-Irwin.

Public Health Service Act, 42 U.S.C. § 101-239 (1989). Retrieved from https://www.gpo.gov/fdsys/granule/CRI-1989/CRI-1989-PUBLIC-HEALTH-SERVICE-ACT-SEE-ALSO-H-17257C/content-detail.html

Radley, D. C., McCarthy, D., & Hayes, S. L. (2017). *Aiming Higher: Results from the Commonwealth Fund scorecard on state health system performance.* Retrieved from http://www.commonwealthfund.org/interactives/2017/mar/state-scorecard/assets/1933_Radley_aiming_higher_2017_state_scorecard_FINAL.pdf

Rogers, E. M. (1962). *Diffusion of innovations*. New York, NY: Free Press of Glencoe.

Rogers, M. A., & Mullen, R. (2012). Outcomes measurement in healthcare. In L.A. Golper (Ed.), *Measuring Outcomes in Speech-Language Pathology*. New York, NY: Thieme Medical Publishers, 91-115.

Schneider, E. C., Sarnak, D. O., Squires, D., Shah, A., & Doty, M. M. (2017). *Mirror, mirror 2017: International comparison reflects flaws and opportunities for better U.S. health care.* New York, NY: Author. Retrieved from http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing.* The Hague, the Netherlands: International Association of Scientific, Technical and Medical Publishers. Retrieved from http://www.stmassoc.org/2015_02_20_STM_Report_2015.pdf

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, MA: Cambridge University Press.

Wolff, S. H. (2008). The meaning of translational research and why it matters. *JAMA, 299,* 211–213.

Towards a Research Profiling Ecosystem Weaving Scholarly, Linked Open and Big Data

David Eichmann
School of Library and Information Science &
Iowa Graduate Program in Informatics
The University of Iowa

Research Profiling / Networking

Research profiling systems provide programmatic support for discovery and use of research and scholarly information regarding people and resources – essentially serving as special purpose institutional knowledge management systems. They have also achieved notable adoption by research institutions.[1] A number of systems have been developed, including open source (e.g., VIVO and Harvard Profiles), commercial (e.g., Elsevier Pure) and local institutional systems (e.g., Iowa's Loki and Stanford's CAP).

Multi-site search of research profiling systems has substantially evolved since the first deployment of systems such as DIRECT2Experts.[2] CTSAsearch is a federated search engine using VIVO-compliant Linked Open Data (LOD) published by members of the NIH-funded Clinical and Translational Science (CTSA) consortium and other interested parties. Eighty-seven institutions are currently included, spanning eight distinct platforms and three continents (North America, Europe and Australia). CTSAsearch has data on 174-421 thousand unique researchers (depending upon how you count) and their 10 million publications. The public interface is available at http://research.icts.uiowa.edu/polyglot.

Linked Open Data (LOD) holds substantial promise for tools supporting collaborative and translational science. The NIH-funded Clinical and Translational Science (CTSA) program has already proven to be a significant catalyst for tools supporting research discovery. Our work on extending Loki, the University of Iowa research profiling system, into the Semantic Web serves as a substantial case study in modular architectures extending into LOD.

The Loki Research Profiling System

Loki was developed as a component of the University of Iowa CTSA to support researcher discovery and collaboration. Comprised of investigator-authored research narratives coupled with publication data from MEDLINE and the Web of Knowledge, Loki's functionality expanded to include NIH funding opportunity awareness, demographics data from Human Resources and grant data from the Division of Sponsored Programs. Loki is investigator- rather than institutionally-focused, supporting multiple phases of the research life cycle, from funding opportunity identification (through NIH announcements), to team formation (through expertise search), to proposal creation (through biosketch management) to outcome dissemination (through automated inclusion of investigator publications). The modular nature of Loki's architecture has been a key element of this approach, consisting of

- A database layer, where each source is managed by a separate connector;
- A tag library layer, where each source is mapped into a suite of semantics-based tags; and
- A Java Server Page (JSP) presentation layer, where the semantic tags are woven into HTML and CSS elements to comprise a browser page.

From Tags to Triples

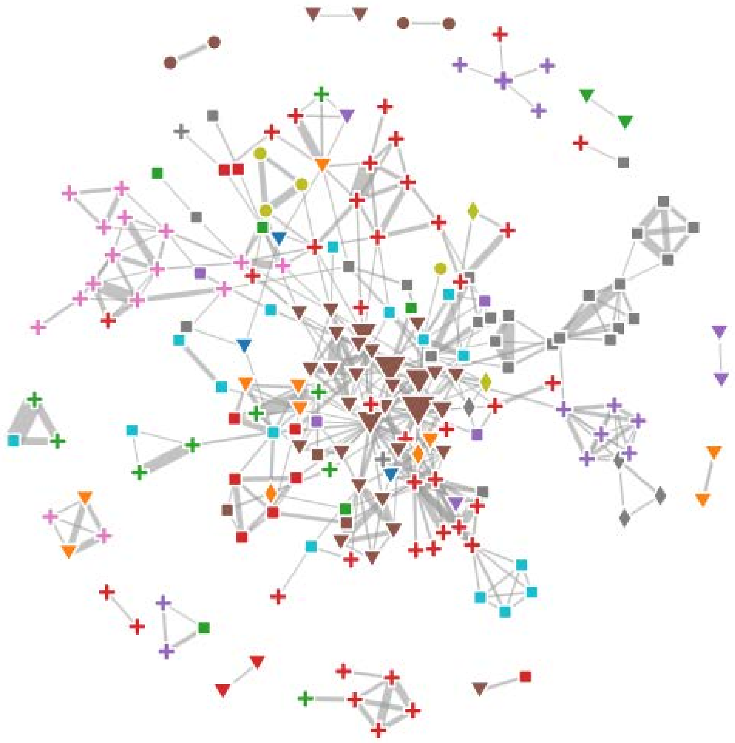
Subsequent work involved definition of a Loki ontology and the mapping of relational database entities into the resulting ontological concepts. Our approach of synthesizing the tag library layer of the architecture from an entity-relationship diagram proved substantially valuable in this work, as much of this mapping proved to be fairly formulaic through the use of the D2R relation to triple mapping tool. Furthermore, the clean partitioning of the logical components (e.g., demographics and publications) of the database layer allowed us to independently represent those components as discrete ontologies, and hence, discrete triple stores – supporting an overall LOD environment of interlinked triple stores that reflected the modularity of our initial tag library design.

Ontological Mapping and Equivalences

The use of ontologies to model complex semantic relationships has become well-established, particularly in certain disciplines, such as biomedicine. Standardization on languages such as OWL have further demonstrated the utility and reusability of such formalisms. VIVO (the ontology) is an excellent example of community adoption of a shared semantic model, and projects such as CTSAsearch have demonstrated the potential for use of these models and the related data beyond that of the original context (i.e. VIVO the application).
As noted above, we opted initially to develop a Loki ontology that directly represents the semantics of our local environment. This was a conscious design decision, as we wished to demonstrate in a practical fashion that the LOD goal of concept mapping and equivalence was possible, and indeed desirable in this domain. We subsequently mapped the Loki ontology to the VIVO ontology within the D2R specification file purely at the ontological level, demonstrating the value in maintaining separation between the representational and conceptual levels in our overall information architecture. At the level of SPARQL query, Loki now is indistinguishable from a native VIVO instance.

CTSAsearch

CTSAsearch(http://research.icts.uiowa.edu/polyglot) is a federated search engine using VIVO-compliant Linked Open Data published by 87 institutions using eight distinct platforms. Since its introduction in 2013, the query and visualization mechanisms in



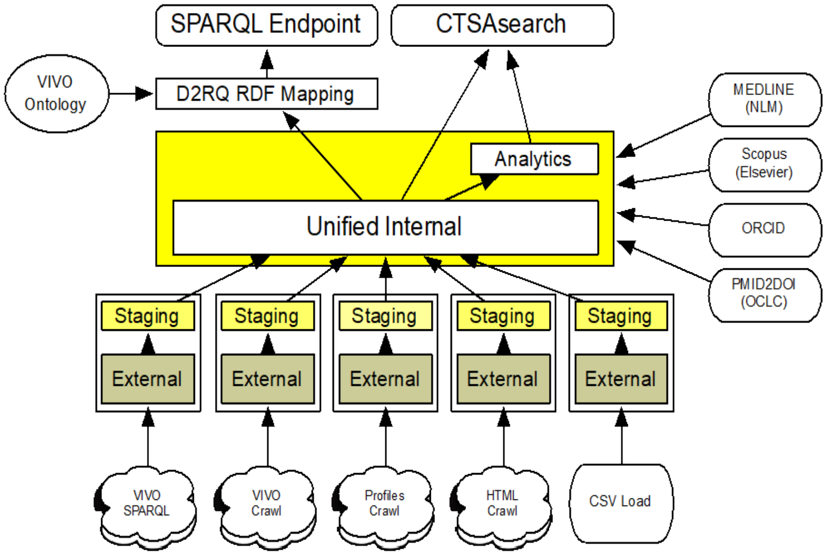
CTSAsearch have proven to be the primary elements of user interest. In particular,the coauthorship relationships between investigators at various institutions forms the principal visualization mechanism, as show in this figure. Each symbol indicates a particular investigator where the specific symbol indicates their home institution and the size of the symbol the relevance to the user's query. Edges between symbols indicate coauthorship with the thickness of the edge indicating the level of joint authorship.

Architecture

CTSAsearch draws both from research networking systems and from multiple data authorities. Given the diversity of information sources, modularity is critical to a robust, adaptable software architecture, as illustrated following.

CTSAsearch is currently comprised of the following:

- 1 VIVO-based SPARQL harvester
- 2 VIVO-based crawlers (due to differences in the respective ontologies
- 1 Profiles-based crawler
- 2 Platform-specific HTML crawlers
- 1 Proprietary API harvester (for Elsevier's Pure)
- 1 CSV-based loader



The resulting information space is comprised of 14.3 million VIVO v. 1.4 – derived triples, 129.3 million VIVO v. 1.6 – derived triples and 74.2 million Profiles-derived triples. The unified internal model aligns these representation variants into a single model which is used for indexing, retrieval and visualization.

Query Formulation using concept recognition

One of the early aspects of user feedback on CTSAsearch was a desire for more sophisticated search than that provided by a simple 'bag-of-words' relevance list. While this google-style search mode is available as an option, the default currently is one supporting full Boolean logic with a greedy concept recognizer processing the Boolean operands. For example, the query "stem cell & ferret" results in two operands, the first bound to UMLS concepts C0018956 and C0038250 (stem cell) and the second C0015859 (ferret). The recognizer aggregates the longest strings of tokens possible in each operand and the resulting concepts and any unrecognized strings are grouped as a Lucene query node. This approach has been very successful in pruning low relevance hits from results (e.g., matches on "cell" above). Finally, each Boolean operand is expanded with the subconcepts for each of its recognized concepts using the UMLS semantic network. This supports retrieval of profiles mentioning more specific descendant concepts by a more generic ancestor query concept.

Author-level co-authorship visualization

The co-authorship connections between the matched profiles are visualized using a force graph implemented in D3. Connections are pre-computed at profile harvesting time using multiple alternative identifiers (DOI, PMID, and PMCID) present in the profile data. OCLC pmid2doi crosswalk data is used to span the identifier spaces. As seen in the figure, useful force graph visualizations are possible for 'reasonable' result scales (n ~ 200). Challenges arise when results are larger – a query term such as "diabetes" returns thousands of results, leading to a network hairball.

Institution-level visualization

I have taken two different approaches to untangling the hairball. The first, and simplest, is aggregating results at the institution (i.e., VIVO instance) level. This clearly limits the number of nodes in the result to the number of VIVO instances for which I have data. However, for our diabetes query, there is little information discernable other than the degree of inter-institutional collaboration present for the topic. I am currently exploring the value of aggregation at smaller granularities (e.g., departments, institutes, etc.).

Inter-institutional community visualization

Focusing on community detection in the network structure is proving to be a far more robust approach to untangling large networks. I use a user-selectable set of community detection algorithms to aggregate community members into a single initial node, and then support zooming into an author-level visualization for a given community. I anticipate that this multi-scalar approach to visualization will accommodate scaling to entire research disciplines.

References

1. Obeid JS, Johnson LM, Stallings S, Eichmann D (2014) Research networking systems: the state of adoption at institutions aiming to augment translational research infrastructure. J Transl Med Epidemiol 2(2): 1026.
2. Weber GM, et al. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. J Am Med Inform Assoc. 2011 Dec; 18 Suppl 1:i157-60. doi: 10.1136/amiajnl-2011-000200. PubMed PMID: 220378.

Aligning Data Collection with Multi-dimensional Construct Definitions: The Example of Behavioral Tasks for Measuring Risk–taking Behavior

**Carl W. Lejuez, Dean, College of Liberal Arts & Sciences,
University of Kansas**

In their seminal work, Jessor and Jessor (1977) defined risk-taking as "behavior that is socially defined as a problem, a source of concern, or as undesirable by the norms of conventional society and the institutions of adult authority, and its occurrence usually elicits some kind of social control response" (p. 33). Focusing more explicitly on the potential consequences or outcomes of such behavior, risk-taking has also been conceptualized as behavior that involves some potential for harm or negative consequence to the individual, but that may also result in a positive outcome or reward (Byrnes et al. 1999; Leigh 1999). Further, the propensity to take risks exists on a continuum, with some risk-taking being adaptive, and only more extreme levels being maladaptive (Bornovalova et al. 2009). The availability of well-developed behavioral methodologies used to study risk-taking behavior and to ensure quality data that are reliable, valid, and meaningful are crucial to the advancement of our understanding of the processes underlying the development and maintenance of risk behaviors (Koffamanus & Kaplan, in press).

Decision-making Risk-taking Tasks

The Iowa Gambling Task

The original gold standard measure of risk taking is the Iowa Gambling Task (IGT; Bechara et al. 1994). The IGT is a decision making task originally developed to examine decisional processes associated with neuropsychological impairment (e.g., Rogers et al. 1999a). At the beginning of the task, the participant is given $2,000, is instructed to maximize earnings over the course of 100 decision-making trials, and is provided with four decks of cards on the computer screen. As described by Bechara et al. (2001), the decks are labelled A, B, C, and D at the top end of each deck. All cards are identical, and each card is associated with hypothetical payoffs or losses (although versions with real financial contingencies are available). Cards from decks A and B pay an average of $100, but also contain cards with higher losses, while cards from decks C and D pay an average of $50, but losses are smaller. Accordingly, 10 draws from decks A and B (the "disadvantageous" decks) lead to a net loss of $250, while 10 draws from decks C and D (the "advantageous" decks) lead to a net gain of $250 (Bechara et al. 1994; Buelow and Suhr 2009).

During the task, the participant clicks on a card from any of the four decks. Once the card is selected, the computer makes a sound similar to that of a slot machine. The selected card appears as either red or black, indicating whether money was lost or gained, and the value of the reward or loss appears at the top of the screen. Following this feedback, the card disappears and the participant selects another card. Each deck of cards is programmed to have 60 cards (30 red and 30 black), although the participant is unaware of how many cards of each type are in each deck. Losses are equally frequent in each deck. Several dependent variables from the IGT indicate RDM, but the most widely reported indices are the number or percentage of disadvantageous choices over 100 trials, with larger values representing greater riskiness.

While there is value in considering risk taking this way, there are concerns about the challenge in being able to repeat the task given the participants goal is to "figure out" the risky options and avoid them. Thus, once this is figured out, regardless of how long that has taken, there is no way to return that same participant to the naïve state in relation to the goals of the task. Researchers have attempted to address this issue, but it remains a challenge for studies that require repeated measures (Almy, Kuskowski, Myers, & Luciana, in press). The nature of the task also presents a conceptual challenge as the task's focus on risky decision making means that it only provides one narrow way to explore risk taking. Indeed, understanding the maladaptive aspects of risk taking from a decision making perspective is at the same time incredibly valuable for understanding one aspect of risk taking but also too narrow to be considered a comprehensive assessment. Given that risk taking is so multidimensional, the argument is not that the IGT should be discarded given its narrowness, but instead that it should be categorized clearly for the aspects of risk taking it does represent well and there should be simultaneous efforts to develop tasks that are targeted at the other dimensions of the construct. The remainder of this paper seeks to discuss this very goal.

Risk-Taking Propensity Tasks

In considering other key dimensions of risk taking, an obvious starting point is the movement away from examining risk taking as solely maladaptive and instead as a continuum, ranging from maladaptive in different ways at its extremes and adaptive in moderation. Such an approach makes intuitive sense with one receiving few rewards when taking no risks whatsoever and too many negative consequences when extreme risks are taken, thereby leaving risk taking in moderation as an ideal goal in most situations, and the outcome measure most relevant being risk taking propensity (RTP). As our research group began to develop a task to do this very thing, we found a simple approach that had been around for some time that could serve as a starting point.

Slovic's Devil Task

The first behavioral task designed and used to assess RTP was Slovic's Devil Task (1966). Although it was not used to assess risk-taking related to substance use in the extant literature, review of the Devil Task is important for its historical significance and relation to current, commonly used RTP behavioral measures.

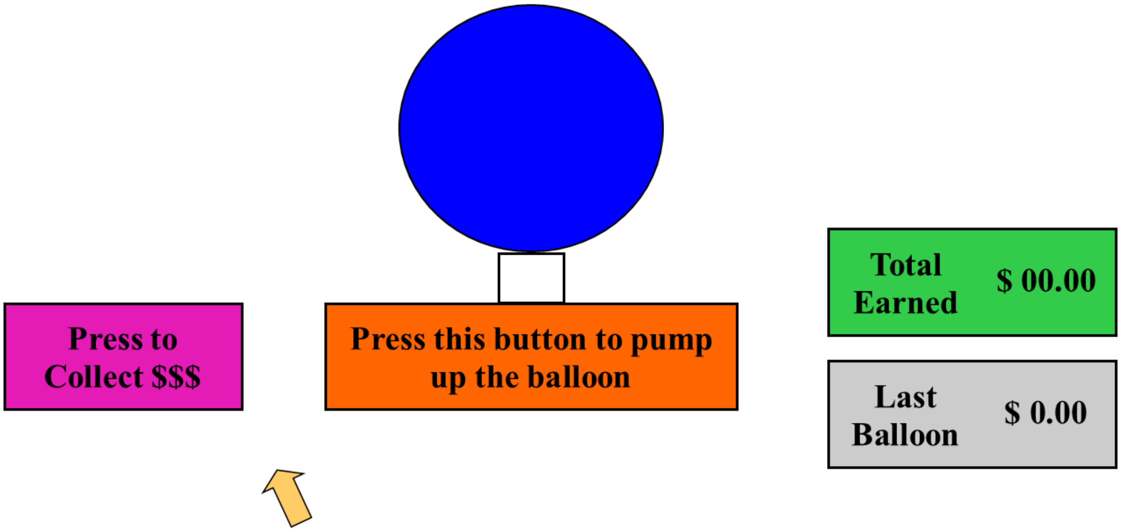
In the Devil Task originally implemented by Slovic (1966), participants were seated before a panel of ten small knife switches and told that nine of the switches were "safe" and that the tenth was a "disaster" switch, with it being impossible to distinguish which was the disaster switch. Each participant was informed that a switch could only be pulled once (i.e., sampled without replacement), thus the likelihood of pulling the disaster switch increased with each subsequent trial. The participant was then asked to pull one of the switches and, if the participant chose a safe switch, he/she was allowed to place one spoonful of candies into a glass bowl. The participant then had to decide whether to pull another switch or to stop and keep the candy he/she had already won. If the participant decided to continue but subsequently pulled the disaster switch, a buzzer would sound and the participant would lose everything he/she had already earned. The task ended when the participant either chose to stop and collect his/her winnings or pulled the disaster switch and lost everything. In the event that the participant pulled nine safe switches in a row, he/she was automatically forced to stop and take his/her nine spoonfuls of candy, as the only remaining switch necessarily was the "disaster" switch. Slovic (1966) argued that because both the probability and magnitude of one's potential loss increases with the number of switches pulled, stopping performance on the task can be considered an index of risk-taking tendencies. A slightly altered version of the Devil Task includes a presentation with ten wooden boxes, where one box contains a devil. The potential reward for this version of the task are stickers (Hoffrage et al. 2003). A computerized version of the Devil Task using boxes also exists (Eisenegger et al. 2010).

The Devil Task is most notable for its historical significance. Despite its limited use, the task possesses multiple features that make it a useful measure for studying risk-taking. For example, the Devil Task discriminates between risk takers without involving learning (Hoffrage et al. 2003). Although few variants of the task have been tested, the task could be a useful vehicle for examining contextual influences on risk-taking such as the impact of varying either the number of switches and/or the magnitude of stakes. Because the task has clear face validity, it is likely useful without modification across developmental stages, and the short administration time makes it ideal for examining the impact of experimental manipulations such as stress or drug/alcohol administration.

Balloon Analogue Risk Task (BART)

The Devil Task provided a simple model of RTP. This simplicity gave rise to our development of a task that took the basic idea of measuring RTP but was done in a way that would allow for increasingly complex ways to study the type of complex risk behaviour seen in the real world. This work led to the development of the Balloon Analogue Risk Task (BART; Lejuez et al. 2002).

The BART is a computerized measure of RTP that models real-world risk behavior through the conceptual framework of risky behaviors in having both the potential for reward as well as the risk of harm (Leigh 1999; Lejuez et al. 2002). In the task, the participant is presented with a balloon and asked to pump the balloon by clicking a button on the screen. With each click, the balloon inflates .3 cm and money is added to the participant's temporary winnings; however, if the participant pumps up the balloon beyond its explosion point, the balloon explodes and he/she loses the money earned on that balloon. The task is depicted in the Figure below.



The explosion point can be varied both across and within studies. Lejuez et al. (2002) used three different balloon colors, with each color associated with a range of 1-8, 1-32, or 1-128 pumps, respectively. The probability that a balloon would explode was arranged by constructing an array of *N* numbers. The number 1 was designated as indicating a balloon explosion. On each pump of the balloon, a number was selected without replacement from the array. The balloon exploded if the number 1 was selected. For example, for the blue balloon ranging from 1-128, the probability that a balloon would explode on the first pump was 1/128. If the balloon did not explode after the first pump, the probability that the balloon would explode was 1/127 on the second pump, 1/126 on the third pump, and so on up until the 128th pump, at which the probability of an explosion was 1/1 (i.e., 100%). According to this algorithm, the average break point would be the midpoint of the range, in this case 64 pumps. Before the balloon pops, the participant can press "Collect $$$" which saves his or her earnings to a permanent bank. If the balloon pops before the participant collects the money, all earnings for that balloon are lost, and the next balloon is presented.

RTP on the BART is defined as the adjusted average number of pumps on unpopped balloons (Bornovalova et al. 2005; Lejuez et al. 2002), with higher scores indicative of greater RTP. In the original version of the task, each pump was worth $.05 and there were 30 total balloons. Participants were provided this information, but were not given information about the breakpoints. This instructional set allowed for the examination of participants' initial responses to the task and to changes as they experienced the contingencies related to payout collections and balloon explosions. Results of the original study also established that the relationships between key outcome variables and BART scores were most evident at the largest balloon range of 1-128, which has largely been used in subsequent studies.

The BART has well-established reliability across a range of samples. Split-third reliability has been examined by comparing scores across the first block of 10 balloons, middle block of 10 balloons, and final block of 10 balloons on the task. The reliability estimates typically indicate strong correlations (>0.7) among the blocks (Lejuez et al. 2010; Lejuez et al. 2002). Extended test-retest reliability has been indicated with the task presented twice across a two-week period, with a nonsignificant increase across administrations (T2–T1Δ = 1.2 adjusted average pumps) and a reasonably robust test-retest correlation (T1/T2 r = .77; White et al. 2008).

It is notable that relationships of performance on the BART to self-report measures of disinhibition are inconsistent. Findings indicate modest, though significant, relationships with sensation seeking (r = ~.20), but typically nonsignificant relationships with self-report and other behavioral measures of impulsivity (Bornovalova et al. 2005; Lejuez et al. 2007; Meda et al. 2009).

The BART is considered to be one of a small number of gold-standard measures of risk-taking (Harrison et al. 2005). Within adolescent studies, RTP as measured by the BART is related to adolescent self-reported engagement in real-world risk-taking behaviors including substance use behaviors and delinquency/safety behaviors (Aklin et al. 2005; Fernie, Goudie, & field, 2010; Lejuez et al. 2003b). Greater levels of RTP on the BART have been found in adolescent ever-smokers as compared to never-smokers as well as in adolescents with conduct and substance problems when compared to matched controls (Crowley et al. 2006; Lejuez et al. 2005). Performance on the BART-Y (a youth version; Lejuez et al. 2007) correlates significantly with a variety of real-world risky behaviors. Specifically, greater RTP on the BART-Y is associated with increased frequency of substance use, gambling, delinquent behaviors, and risky sexual behavior (Aklin et al. 2005; Hamilton, Felton, Risco, Lejuez, & MacPherson, 2014; Lejuez et al. 2007; Lejuez et al. 2005; Macpherson 2010).

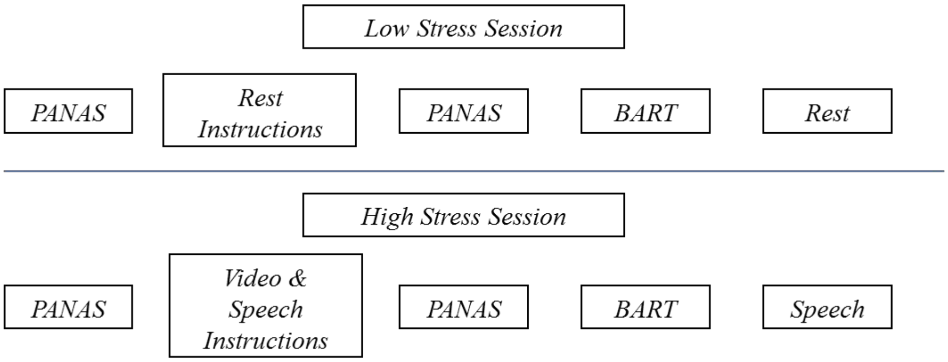
Prediction of Future Risk Behavior

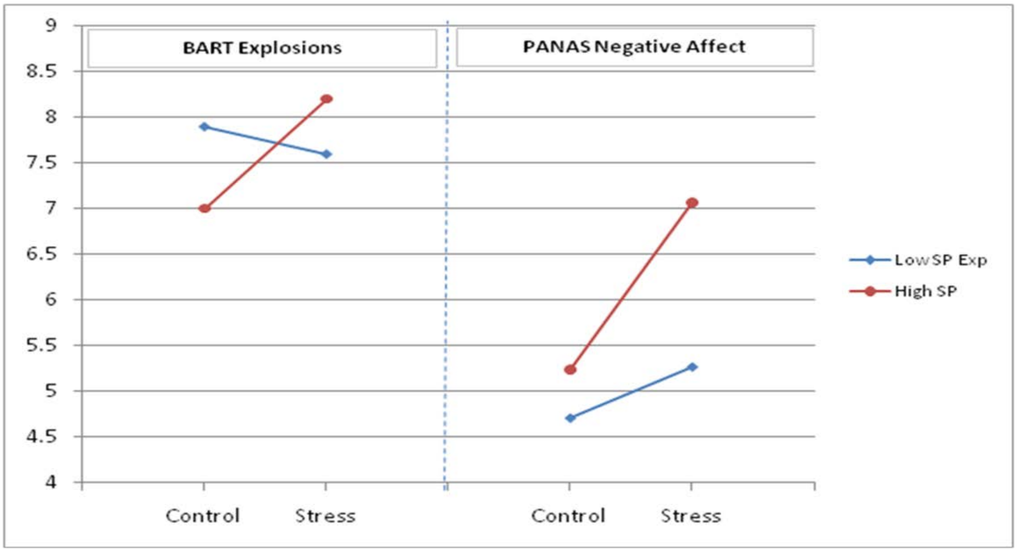
Of particular importance is the extent to which the BART can be used to predict future risk behavior. As noted above, existing data often shows a correlation between risk taking on the task and current levels of real world risk behavior, and a handful of studies have further shown that changes in risk-taking behavior in the real world over a period of time correlate with changes in risk taking on the BART over the same period of time (MacPherson et al. 2010). However, there appears to be no evidence of risk taking on the BART at one time point predicting future risk taking behavior. The absence of prediction data considered together with reasonably solid evidence of cross-sectional relationships between BART suggest that while the BART may not be a task to predict who will be risky in the future, its does a competent job of serving as a proxy of existing risk behavior in the real world. Towards this end, we have completed several studies that leverage this opportunity to understand how risk taking is impacted by a range of external factors in the real world by manipulating those factors in a controlled laboratory environment.

Reward density and category. The BART has one of the few studies examining the effects of varying cash reward magnitudes on RTP, as a function of individual differences in impulsivity (Bornovalova et al. 2009). Specifically, Bornovalova and colleagues manipulated the magnitude of reward/loss value, examining differences in BART score at 1, 5, and 25 cents per pump, as a function of trait impulsivity and sensation seeking. As the magnitude of monetary reward/loss value increased, risk-taking on the task decreased. However, when examined separately among individuals with high impulsivity/sensation seeking as compared to individuals low in these traits, the negative relationship between reward/loss magnitude and riskiness appeared most prominent among individuals low in impulsivity/sensation seeking. Conversely, individuals high in impulsivity/sensation seeking showed little change in riskiness as reward/loss magnitude increased. These findings illustrate that higher reward/loss magnitudes convey less risk-taking on the task, particularly for individuals low in impulsivity-related traits, suggesting that researchers should consider the value of reinforcers that are employed in behavioral risk-taking paradigms when interpreting results.

Peer Influence. Reynolds (2014) examined the effect of peer influence on BART RTP. Results of this study indicated that while no differences existed between the groups at baseline, RTP on the BART was significantly greater at a second experimental session for individuals who had peers encouraging their risk-taking as compared to individuals who completed the second session alone or with peers solely present, but not encouraging. In a second study, Cavalca et al. (2013) found that youth who were smokers (a proxy for higher levels of real world risk taking) were more influenced by peer pressure to be risky than youth who were not smokers. These data suggest that risk taking and the influence of peer pressure on risk taking can be assessed using the BART.

Anxiety. One particularly challenging question in the risk taking literature is the impact of anxiety. While anxiety is more generally thought to be negatively related to risk taking (e.g., Butler & Mathews, 1987), other evidence suggests that risk behaviors increase among those with anxiety in situations in which an anxiety-inducing stimulus is present (Kashdan & Hofmann, 2008). While this dichotomy makes intuitive sense it is difficult to test tease apart the disposition vs situation impact on risk taking of those with elevated anxiety. To address this issue, Reynolds et al. (2013) utilized an anxiety induction with a community sample of adolescents and compared performance on the BART in a high and low stress state (see the top of the Figure below for a diagram of the study). The high stress state included informing participants that they would be giving a speech at the end of the study and in the interim, they watched study confederates giving the speech and getting a difficult response from the audience; the low stress state included watching pleasant videos. Participants completed the BART before the low or high stress induction and again after the induction to examine the impact of the induction on risk taking as a function of anxiety level. Participants were partitioned based on their level of anxiety specific to social phobia (resulting in a high and low social phobia group). As shown in the bottom half of the Figure below, an interaction of group (low social phobia versus high social phobia) by condition was observed. Those in the high social phobia group had significantly increased risk-taking behavior from the control to the experimental session. Moreover, while the report of negative affect was similar for both groups in the low stress condition, the high social phobia group reported significantly higher self-reported distress during the high stress condition. This suggests that a task such as the BART can be useful in isolating the impact of anxiety inducing stimuli, and that for those showing greater impact of that stimuli a corresponding increase in risk taking behavior is observed.





Conclusion

In conclusion, data plays an important role in our understanding of real world risk taking behavior. Ensuring the quality of that data requires an understanding of the rationale for the tasks developed and used. It also requires a clear sense of what useful existing data or new behavioral tasks can provide and where they fall short. The IGT and the BART assess risk taking in different ways, and together they may provide a strong comprehensive picture. Nonetheless, there is little data from any task suggesting strong longitudinal predictions. Thus, at least for now it seems clear that isolating risk taking in a controlled laboratory setting and providing the opportunity to manipulate key variables thought to impact risk behavior in the real world should be the focus on experimental studies. We have begun to conduct this work with the BART and have reviewed some early studies above, but considerable work including studies that bring in genetic factors and neural assessment (Bjork & Pardini, 2015; Gu, Zhang, Luo, Wang, & Broster, in press) as well as a wider range of environmental factors will be crucial to any further progress.

References

Aklin, W. M., Lejuez, C. W., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. *Behavior Research and Therapy, 43,* 215-228.

Amy, B., Kieslowski, M., Malone, S. M., Myers, E., & Luciana, M. (in press). A longitudinal analysis of adolescent decision-making with the Iowa Gambling Task. *Developmental Psychology.*

Bechara, A., Damasio, A. R., Damasio H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7-15.

Bechara, A., Dolan, S., Denburg, N., Hindes, A., Anderson, S. W., & Nathan P . E. (2001). Decision making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. *Neuropsychologia, 39,* 376-389.

Bjork, J. M., & Padroni, D. A. (2015). Who are those risk-taking adolescents?: Individual differences in developmental neuroimaging research. *Developmental Cognitive Neuroscience, 11*, 56-64.

Bornovalova, M. A., Cashman-Rolls, A., O'Donnell, J. M., Ettinger, K., Richards, J. B., Dewit, H., & Lejuez, C. W. (2009). Risk taking differences on a behavioral task as a function of potential reward/loss magnitude and individual differences in impulsivity and sensation seeking. *Pharmacology, Biochemistry, and Behavior, 93,* 258-262.

Bornovalova, M. A., Daughters, S. B., Hernandez, G. D., Richards, J. B., & Lejuez, C. W. (2005). Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a residential substance-use program. *Experimental and Clinical Psychopharmacology, 13,* 311-318.

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa gambling task. *Neuropsychology Review, 19,* 102-114.

Butler, G., & Mathews A. (1987). Anticipatory anxiety and risk perception. *Cognitive Therapy and Research, 11,* 551-565.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychol Bulletin, 125,* 367-383.

Cavalca, E., Liss, T., Lejuez, C. W., Reynolds, E. K., Schepis, T. S., Kong, G., & Krishnan-Sarin, S. (2013). A preliminary experimental investigation of peer influence on risk taking among adolescent smokers and non-smokers. *Drug and Alcohol Dependence, 129,* 163-166.

Crowley, T. J., Raymond, K. M., Mikulich-Gilbertson, S. K., Thompson, L. L., & Lejuez, C. W. (2006). A Risk-Taking "Set" in a Novel Task Among Adolescents With Serious Conduct and Substance Problems. *Journal of the American Academy of Child and Adolescent Psychiatry, 45,* 175-183.

Eisenegger, C., Knoch, D., Ebstein, R. P., Gianotti, L. R R., Sándor, P. S., & Fehr, E. (2010). Dopamine receptor D4 polymorphism predicts the effect of L-DOPA on gambling behavior. *Biological Psychiatry, 67,* 702-706.

Fernie, G., Cole, J. C., Goudie, A. J., & Field, M. (2010). Risk-taking but not response inhibition or delay discounting predict alcohol consumption in social drinkers. *Drug and Alcohol Dependence, 112*, 54-61.

Gu, R., Zhang, D., Luo, Y., Wang, H., & Broster, L. S. (in press). Predicting risk decisions in a modified balloon analogue risk task: Conventional and single-trial erp analyses. *Cognitive, Affective & Behavioral Neuroscience*.

Hamilton, K. R., Felton, J. W., Risco, C. M., Lejuez, C. W., MacPherson, L. (2014). Brief report: The interaction of impulsivity with risk-taking is associated with early alcohol use initiation. *Journal of Adolescence, 37*, 1253-1256.

Harrison, J. D., Young, J. M., Butow, P., Salkeld, G., Solomon, M. J. (2005). Is it worth the risk? A systematic review of instruments that measure risk propensity for use in the health setting. *Social Science & Medicine, 60,* 1385-1396.

Hoffrage, U., Weber, A., Hertwig, R., & Chase, V. M. (2003). How to Keep Children Safe in Traffic: Find the Daredevils Early. *Journal of Experimental Psychology: Applied, 9,* 249-260.

Jessor, R., & Jessor, S. L. (1977). Problem behavior and psychosocial development: A longitudinal study of youth. Academic Press New York

Kashdan, T. B., & Hofmann S. G. (2008). The high-novelty-seeking, impulsive subtype of generalized social anxiety disorder. *Depression and Anxiety, 25,* 535-541.

Koffarnus, M. N., & Kaplan, B. A. (in press). Clinical models of decision making in addiction. *Pharmacology, Biochemistry and Behavior,* https://www.ncbi.nlm.nih.gov/pubmed/28851586

Leigh, B. C. (1999). Peril, chance, adventure: concepts of risk, alcohol use and risky behavior in young adults. *Addiction, 94,* 371-383.

Lejuez, C. W., Aklin, W. M., Daughters, S. B., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2007). Reliability and validity of the youth version of the Balloon Analogue Risk Task (BART- Y) in the assessment of risk-taking behavior among inner-city adolescents. *Journal of Clinical Child and Adolescent Psychology, 36,* 106-111.

Lejuez, C. W., Aklin, W. M., Bornovalova, M. A., & Moolchan, E. (2005). Differences in risk taking propensity across inner-city adolescent ever- and never-smokers. *Nicotine and Tobacco Research, 7,* 71-79.

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. R., Strong, D. R., Kahler, C. W., & Read, J. P. (2003). The Balloon Analogue Risk Task Differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology, 11,* 26-33.

Lejuez C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla C. M. (2003b). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescne, 26,* 475-479.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimntal Psychology: Applied, 8,* 75-84.

Lighthall, N. R., Mather, M., & Gorlick, M. A. (2009). Acute Stress Increases Sex Differences in Risk Seeking in the Balloon Analogue Risk Task. PLoS One, 4, e6002.

MacPherson, L., Magidson, J. F., Reynolds, E. K., Kahler, C. W., & Lejuez, C. W. (2010). Changes in Sensation Seeking and Risk Taking Propensity Predict Increases in Alcohol Use Among Early Adolescents. *Alcohol: Clinical and Experimental Research, 34,* 1400-1408.

Meda, S. A., Stevens, M. C., Potenza, M. N., Pittman, B., Gueorguieva, R., Andrews, M. M., Thomas, A. D., Muska, C., Hylton, J. L., & Pearlson, G. D. (2009). Investigating the behavioral and self-report constructs of impulsivity domains using principal component analysis. *Behavioral Pharmacology, 20,* 390-399.

Reynolds, E. K., Schreiber, W. M., Geisel, K., MacPherson, L., Ernst, M., & Lejuez, C. W. (2013). Influence of social stress on risk-taking behavior in adolescents. *Journal of Anxiety Disorders, 27*, 272-277.

Reynolds, E. K., MacPherson, L., Schwartz, S., Fox, N. A., & Lejuez, C. W. (2014). Analogue study of peer influence on risk-taking behavior in older adolescents. *Prevention Science, 15,* 842-849.

Rogers, R., Everitt, B., Baldacchino, A., Blackshaw, A., Swainson, R., Wynne, K., Baker, N., Hunter, J., Carthy, T., & Booker, E. (1999a). Dissociable deficits in the decision making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms. *Neuropsychopharmacology, 20,* 322-339.

Schepis, T. S., McFetridge, A., Chaplin, T. M., Sinha, R., & Krishnan-Sarin, S. (2011). A Pilot Examination of Stress-Related Changes in Impulsivity and Risk Taking as Related to Smoking Status and Cessation Outcome in Adolescents. *Nicotine and Tobacco Research, 13,* 611-615.

Slovic, P. (1966). Risk-taking in children: Age and sex differences. *Child Development, 37*, 169-176.

White, T. L., Lejuez, C. W., & de Wit, H. (2008) Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Experimental and Clinical Psychopharmacology, 16,* 565-570.

Aligning Researcher Practice to Support Public Access to Data

Surya K. Mallapragada, Associate Vice President for Research,
Iowa State University

There is a national move towards open science and open enquiry. The Royal Society summary report (2012), places “open enquiry at the heart of the scientific enterprise”. The report outlines the different advantages that open access brings, including enabling change, facilitating new ways of doing science, allowing new collaborations to flourish, enabling better communication of science with the public, enhancing reproducibility and quality and impacting transparency and accountability.

Studies have shown that open science can facilitate research and can help researchers succeed by improving citation rates. The extent of benefit was found to vary across disciplines, as seen in Figure 1 (from McKiernan et al., 2016).

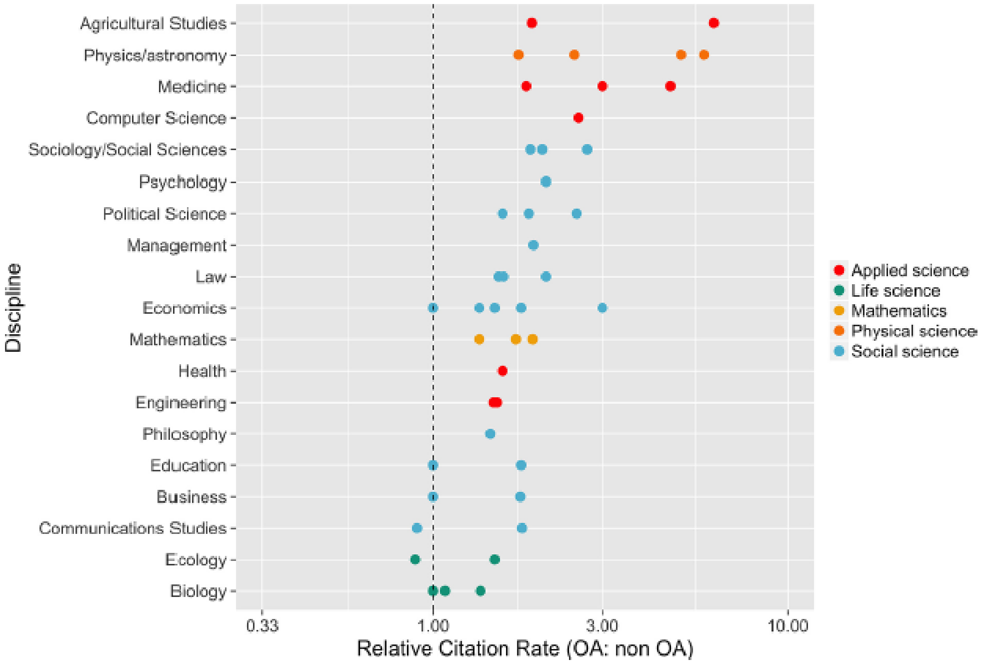


Figure 1. Open access articles get more citations. The relative citation rate is defined as the mean citation rate of open access articles divided by the mean citation rate of non-open access articles. (Reproduced from McKiernan et al, eLIFE, 2016; 5:e16800. License: https://creativecommons.org/licenses/by/4.0/legalcode).

There has been a recent increase in open access policies in the US (Figure 2), and the Holdren memo (2013) mandates public access to research products developed with federal funding.

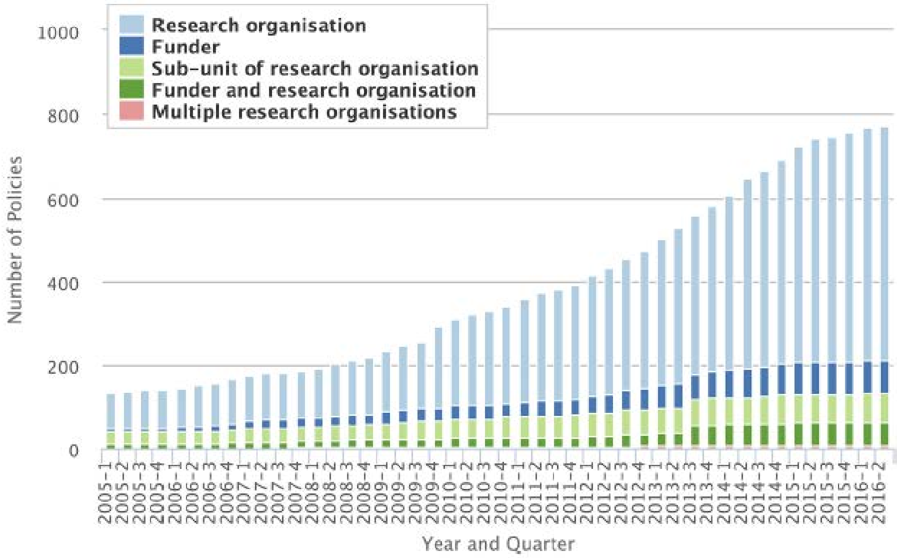
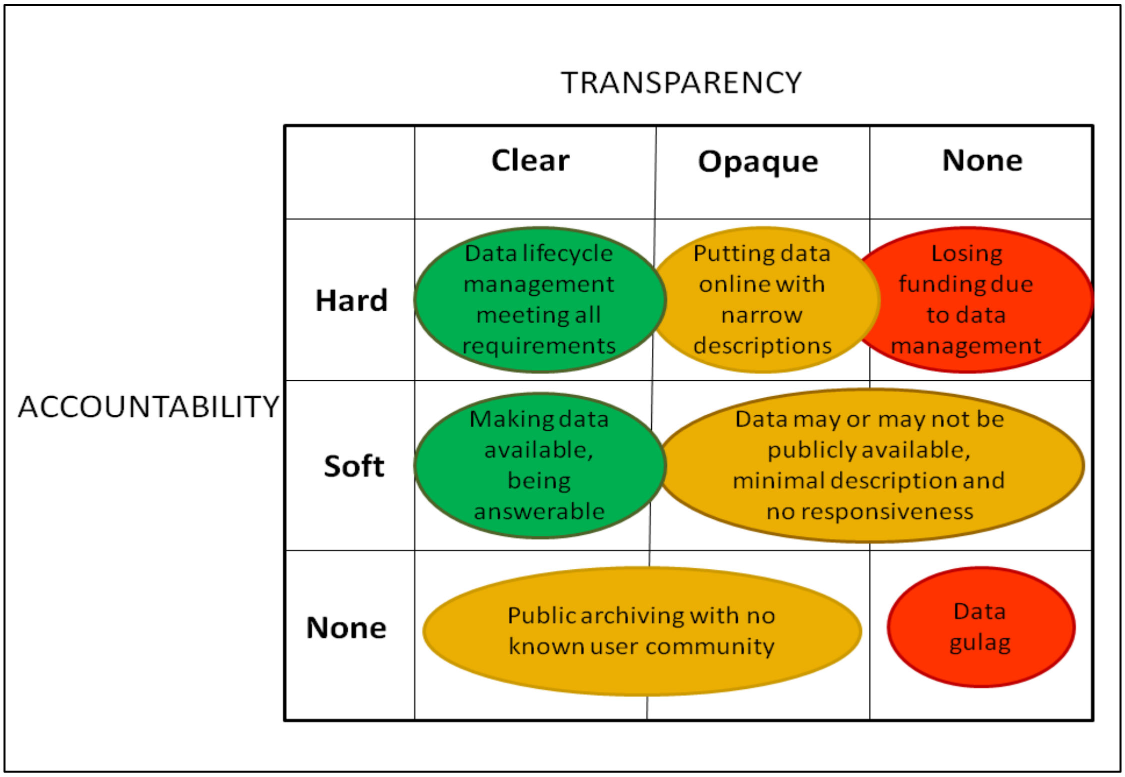


Figure 2. Recent increase in open access policies (Reproduced from McKiernan et al., eLIFE, 2016; 5:e16800. License: https://creativecommons.org/licenses/by/4.0/legalcode)

While the resources and systems for openly sharing publications are well developed, the policies associated with data sharing are less well defined, and sometimes contradictory. Open access to data will be effective only if there are common standards for communicating data and there is a cohesive and uniform strategy used, so that data is usable by others and sharing data is a criterion for career progression. Mayernik (2017) has conducted an interesting analysis of the likelihood of research practices related to open data policies initiatives achieving the goals of open data policy initiatives. Figure 3 presents a model of open data distinguishing between hard and soft accountability as well as different levels of transparency, to arrive at the relative likelihood in each case that research practices will achieve the goals of broad accessibility and usability of data.



	Clear	Opaque	None
Hard	🟩 Data lifecycle management meeting all requirements	🟨 Putting data online with narrow descriptions	🟥 Losing funding due to data management
Soft	🟩 Making data available, being answerable	🟨 Data may or may not be publicly available, minimal description and no responsiveness	
None	🟨 Public archiving with no known user community		🟥 Data gulag

Figure 3. Color scheme shows likelihood (green to red) of research practices related to open data policy initiatives achieving the goals of broad accessibility and usability of data (Reproduced with permission from Mayernik, 2017).

The Open Science Framework is an example of a workflow management system designed to help researchers in this regard (http://osf.io). Data sharing is the focus of an AAU-APLU Public Access Working Group co-chaired by Dr. Sarah Nusser, Vice President for Research at Iowa State University, that is working on common goals for data sharing, federal agency recommendations and guidance for research institutions. At Iowa State University, the implementation is being coordinated across the Library, IT services and the Office of the Vice President for Research, aided by a faculty committee to provide faculty perspective on building a rigorous process for data sharing. Some key questions to consider as we develop institutional policies and practices are to see what the purpose of sharing is, what data should be shared, what the standard should be for documenting data, different options for storing the data, and finally how we can train researchers to adopt this new mindset.

References

1. “Science as an open enterprise”, The Royal Society, London. 2012
2. “How open science helps researchers succeed”, McKiernan, E.C., Bourne, P. E., Brown, C.T. et al., eLIFE, 2016; 5:e16800
3. “Open data: Accountability and transparency”, Mayernik, M.S., Big Data and Society, 1-5 (2017)

If a Tree Fell in a 300 Million-year Old Forest, Did it Leave a Data Trail? [1]

Joseph A. Heppert, Ph.D., Vice President for Research
Texas Tech University

You have just checked into a deluxe ski resort. You grab your skis, take several lifts to get to the top of the mountain, and arrive at an area with a breathtaking view, but startlingly few staff to direct you to the runs. In fact, you are concerned that none of the slopes seem to be formally marked. Nevertheless, deducing that there must be some way off the mountain, you find a slope that shows signs of use, seems free of debris, and is about your skill level. You start down the trail. Switching to another trail, you encounter other skiers and, feeling more at ease, make it to the bottom of the second run. There, you happen across an individual who, based on their clothing, seems to be part of the ski patrol. You politely inquire about the state of the signage on the slopes. "Oh," the individual remarks in an off-hand manner, "We've never bothered to mark or groom the slopes at this resort." As this response is sinking in, you hear and feel a low frequency rumbling. You look up and see a 20-foot wall of snow hurtling down the mountain in your direction…

Of course, this is an allegory, not something that would actually occur in today's ski industry. But it is a fair description of a situation unfolding today in aspects of data management within numerous U.S. research universities.

Researchers in many different fields—genomics, bioinformatics, climate science, fluid dynamics, economics and marketing, etc.-- are creating masses of data at a rate unprecedented in the history of the age of scientific discovery. It is estimated that the entirety of the dataverse reached 1.8 zettabytes (1 zettabyte = 1 trillion gigabytes) in 2011[2]. That nearly doubled by 2012, and is expected to reach 160 zettabytes by 2025 [3]. The developing ability of artists, scientists, and scholars to create massive datasets that are integral to their scholarly work has far outstripped the rate at which university data curation systems and policies adapt to the new condition. This paper outlines some of the origins of this explosion in data volume, describes some of the specific challenges posed by the accelerating pace of data creation, and examines strategies that university systems are considering for managing the unfolding Data Age.

The proliferation of big data in research

In part, the explosion of research problems employing big data sets has been driven by the remarkable technological advancements in high performance computing and networking. Many of today's researchers have never experienced the challenges associated with the early days of computing, when computer code and, often, data sets had to be entered by hand onto punch cards or paper tape, and when CPU register space was severely limited. Recent estimates suggest that compute capacity per dollar has increased by a factor as large as 10^{15} over the past 60 years. [4] At the same time, the decreasing price of data storage capacity, the invention of the Internet in the late 1960's, and the recent move toward 100 Gbps Ethernet capacity has made it more feasible to bring large data sets to the compute capacity together.

These dramatic technological enhancements have allowed investigators to analyze more comprehensive and, presumably, more realistic data sets in a range of computational and modeling applications. Arguably, the signature big data project of the last century was the mapping of the human genome. But today, evolutionary biologists regularly use datasets ranging up to 10 Tb to map previously unknown genomes using de novo assembly strategies. Similar types of big data applications are playing out in analyses of climate data, educational testing and evaluation, modeling of turbulence around aircraft and wind turbines, business analytics, and countless other fields.

How universities are addressing the data challenge

One clear trend among institutions supporting research using big data sets is investment in high-quality high-performance research computing (HPC). The University of Kansas (KU) has only had a centralized HPC strategy for five years, and is therefore a late entrant into this arena. The factors that motivated us to provide centralized support for HPC probably mirror those of many public research universities:

- Saving resources by reducing the proliferation of cold rooms for housing servers.
- Minimizing duplication and underutilization of IT staff.
- Allowing researchers and students to focus on research, not on attempting to maintain clusters and servers.
- Creating an efficient HPC computing and networking environment.
- Providing data management capabilities in response to federal sponsors.
- Minimizing threats to data security.

KU chose to initially undertake a subsidized "condominium" model for support of HPC, with investigators storing hardware purchased for their programs in a common cold room environment. This program was modeled on the "community cluster" model Purdue University uses to support HPC. Many other institutions purchase or lease large amounts of compute and storage capacity and charge investigator grants for use of the resource, while some provide free access to institutional computing resources to all investigators. There is no single model for how major universities offer similar resources to support HPC-based research with large datasets.

I have often told my colleagues and staff, '…the minute it becomes economical, secure, and efficient to off-load HPC computing and storage capacity to the cloud, we want to be out of the business of owning and leasing "enterprise" HPC hardware.' By "enterprise" HPC hardware, I mean multiprocessor cluster units (containing either standard processors or GPUs) and standard spinning storage media that can support 85% or more of common needs for HPC research computing applications. (In contrast, I note that there will always be a need for universities to host hardware for researchers investigating new technologies and configurations of computing, networking and storage hardware.) Universities generally are not adapted to the mission of optimizing technology business processes, nor do our individual operations create the financial efficiencies of operations run by Amazon©, Google©, or Microsoft©. Owned or leased computer hardware also does not support the degree of scalability offered by major cloud computing providers.

Unfortunately, most contemporary analyses of cloud computing services still do not support fully moving university enterprise HPC to a cloud platform. In part, this is because the cost of using dynamic storage in a cloud computing environment is still prohibitively expensive. But the landscape is constantly changing. Most larger university compute customers have transitioned from outright hardware ownership to designed multiple year leases of hardware. This provides a reliable method for maintenance and upgrade of computational hardware at a predictable annual cost that is a fraction of the lump sum investment in a hardware purchase. Many universities are using cloud HPC services when their compute demand bursts over on-campus capacity. Several have experimented with the services of a computing broker to advise them on diversifying their monetary investment in HPC to maximize their computing power.

The current economics of storage technology represents more of a mixed picture. Dynamic cloud storage is currently far more expensive than on-campus spinning media. This is driven by data access costs companies add to the fees for storage capacity. However, glacial cloud storage, which is used for long-term archiving of infrequently accessed data, does provide a cost advantage over even low-tier on-site storage. Many big data research file storage systems are now featuring a seamless interface between dynamic on-site spinning media and glacial cloud storage. Part of the trade-off for employing cloud storage is the advantage of having scalable, immediately accessible storage resources at the expense of giving up some control over institutional data integrity.

Key challenges facing universities and researchers

Though the cost of computing, networking and storage capacity have all exponentially declined over our lifetime, one of the challenges facing universities is that the increase in size of many research data sets has balanced and possibly outstripped the rate of these cost savings. Investigators have noticed this, and, in the absence of long-term institutional strategies for high quality data curation, have flocked to low-cost and sometimes low-quality technologies for data storage. Funding agencies, in turn, have noticed this trend, and have begun to intervene out of concern for data integrity and accountability to taxpayers. Federal agencies funding research have instituted requirements in research proposals for data management plans. NIH has notably created publication and data repositories for investigators funded by the Department of Health and Human Services. In 2013, the Office of Science and Technology

(OSTP) instituted a policy aimed at making the data collected through federal research funding available to the public. [5] This policy provided no hint of how universities were supposed to (a) curate such enormous volumes of research data or (b) fund this policy mandate.

The OSTP mandate has created a dilemma for research universities. Few university IT systems have been engineered with the bandwidth or server capacity to provide public access to research data. In fact, one could argue that security concerns are driving most universities to isolate their research data platforms from public access. Numerous research universities have moved toward creating research DMZ's (sequestered research computing, storage and networking zones) to separate research data flow from business enterprise, education, entertainment and social media traffic, which often dominate day-to-day activity on university networks. But the question facing universities is: Should each research university create its own public research data portal in order to provide access to data from federally funded projects? Six hundred forty universities are listed in the most recent NSF Higher Education Research and Development (HERD) survey. [6] It seems apparent that scattering unrelated disciplinary data among 640 separate data archives, with the accompanying potential for devolution of technology and interface compatibility, might satisfy the letter of the OSTP directive; but would not provide the public with transparent accessibility and use of the archived data.

Another challenge facing research universities is policy makers' concerns about the leakage of technologies (including physical artifacts, information, and tacit knowledge and skills) that contribute to U.S. competitiveness to foreign governments and companies. Though some of this leakage is unintentional, cases of espionage aimed at industrial and academic research and development activities are well documented. According to most security experts, such incidents are increasing in frequency and intensity. This activity is not solely aimed at classified and dual-use technologies, but also at technologies that fall into the category of controlled unclassified information. Controlled unclassified information (CUI) is defined and categorized by the National Archives, which is also responsible for creating standards for its protection. [7] CUI can include an astounding range of data types, including:

- Patent data/proprietary process information.
- Confidential technical information.
- Export controlled information and technology.
- Personal health and financial information.
- Law enforcement data.
- Critical infrastructure specifications.

This area of regulatory controls represents a new world and new challenges for research universities. Not only will CUI regulation affect the nature and scope of some technological information we are can openly publish, but deemed export controls are likely to affect which subjects can be studied by certain groups of foreign nationals.

Key principles for a saner, more useful, and secure data landscape

Resolutions to the issues outlined above will require continued dialog among the academy, and the federal agencies that fund and regulate research. A diverse group of stakeholders in research universities must be engaged in developing proposals to address these concerns. University leaders must engage faculty in developing funding plans and policies to support the efficient management of research data. Representatives of scholarly disciplines must define best practices for data utilization in a way that serves the needs of modern research in their fields. Librarians must examine methods for the efficient curation of scholarly data, and collaborate with disciplinary representatives to create interfaces that support the research culture of the discipline. Information Technology specialists must ensure that systems are resilient, affordable and broadly compatible with the needs of the vast array of disciplines represented in the academy.

At a recent symposium examining the challenges of supporting research using big data sets, there was a discussion of the mandate for public access to data from federally sponsored research. Several thought leaders proposed an alternative to the creation of individual institutional data archives. Following on the pattern of NIH's creation of PubMed [8], and the Census Bureau's development of Federal Statistical Research Data Centers [9], a suggestion was made to gather related data into disciplinary data archives. This solution would create centralized repositories of research data commonly used in similar disciplines. In spite of the potential advantages of this approach, several concerns were discussed. Some of the most interesting research questions derived from big data may stem from finding correlations between datasets not commonly associated with a single discipline. By creating silos composed of commonly associated data sets, we might inadvertently impede interdisciplinary inquiry. But perhaps the larger conundrum with this proposal is that disciplinary societies and associations seem unlikely to volunteer to host and fund continuously growing data archives, and federal agencies have not stepped forward to offer ongoing funding to support their formation. Continued examination of these issues must continue to engage all stakeholders.

In order to promote a healthy data culture in higher education, the following principles seem to form a reasonable framework for future action by research universities:

- Provide economical access to high quality, professionally maintained computing capacity and archival storage.
- Wherever possible give up ownership of computational and storage hardware to commercial vendors who maximize value for compute and storage spending.
- Facilitate, as appropriate, the transition from paper-based research records to electronic records; and streamline the association of various records, research products, and data sets associated with a particular project.
- Standardize meta-data to identify data sources and ownership, associate data sets with original funding sources and publications, and define keywords and data

tags to provide consistency in data curation and searching.

- Create internal data management training and policies that minimize the volume of data retained for long periods of time.
- Engage disciplinary experts to incorporate data management best practices into curricula, create norms for data lifecycles, and develop strategies for centrally storing and curating related data resources.
- Develop shared application interfaces to bring computing tasks to large data sets.
- Create institutional capacity to ensure compliance with CUI controls.
- Continue the dialog with funding agencies about sustainable support for research data archives.

Notes and references

[1] The answer, of course, is: "Yes, assuming someone chooses to include it in a research study."

[2] Webopedia Staff, "How much data is out there?" Webopedia, March 2014, http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html

[3] Andrew Cave, "What will we do when the world's data hits 163 zettabytes in 2025?" Forbes, April 2017, https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#f7788a3349ab

[4] Hardware and AI Timelines, "Trends in the cost of computing." March 2015, https://aiimpacts.org/trends-in-the-cost-of-computing/

[5} OSTP, "Increasing access to the results of federally funded research." February 2013, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] National Science Foundation, "Higher Education Research and Development Survey: Fiscal Year 2015." March 2017, https://ncsesdata.nsf.gov/herd/2015/.

[7] National Archives, "Controlled Unclassified Information (CUI)." October 2017, https://www.archives.gov/cui.

[8] NIH, "PubMed."
https://www.ncbi.nlm.nih.gov/pubmed/.

[9] U.S. Census Bureau, "Federal Statistical Research Data Centers."
https://www.census.gov/about/adrm/fsrdc/locations.html.

Data, Consent, Privacy, and Insight

Daniel A. Reed, University of Iowa

From Toffler's iconic 1970 book, *Future Shock* and its meditations on "much change in too short a period of time" through Friedman's 2005 book, *The World is Flat,* and its commentary on globalization and its far reaching effects, to contemporary assessments of global information flows, much has been written about the accelerating pace of technical change and the associated socioeconomic consequences. Some assessments have been simplistic and motivated by specific social agendas; others have been nuanced and deep. All agree the changes are deep, profound, and substantive. Consider just a few, illustrative examples:

- **Urbanization** is creating new megacities while, even as the world's population grows, rural areas are being depopulated. United Nations predictions suggest that over 66 percent of the population will live in megacities by the year 2050 [1], up from 30 percent in 1950.
- **Stratification** is concentrating a larger fraction of the world's wealth in an increasingly small fraction of the population. Today, the top one percent control the overwhelming majority of global wealth, and socioeconomic mobility (i.e., the ability to change the economic stratum of one's birth) continues to decline.
- **Disintermediation** of existing supply and distribution models and chains is creating new businesses while eliminating others. From e-commerce (e.g., Amazon) through professional services (e.g., E-Trade and Zillow) to consumer services (e.g., Uber and AirBnB), all focus on direct consumer engagement.
- **Polarization** of social perspectives and political opinions is one consequence of such rapid shifts, as people respond to change with anger and fear, fed by a continuous stream of targeted news and information, based on data analytics and machine learning.

Against this backdrop of social issues, technological change continues apace, contributing to and accelerating the social change. As the 21st century industrial revolution, the digitization of our world surely ranks at the top of that technological change. Consider just a few examples:

- **Big data** arises from a combination of e-commerce transactions, the explosive growth of smartphones, the nascent, but rapidly growing Internet of Things (IoT), and the automation of government and business services.
- **Deep learning,** enabled by both big data and massive computing capabilities, is increasingly enabling computing systems to equal or exceed human capabilities on a wide range of speech and vision tasks, as well as routine and (often) non-routine cognitive tasks.
- **Automation** is a direct consequence of deep learning, with machines now managing many manufacturing tasks

and increasingly supplanting humans in areas such as medicine and finance.

- **Biomedicine** advances, triggered by inexpensive DNA sequencing and new insights, have created tailored treatments for heretofore untreatable illnesses, albeit sometimes at exorbitant costs. It has also brought a new world of consumer health sensors and the quantified self movement.
- **Environmental** change and global warming create severe weather, adversely affect agriculture, and may render some portions of the planet uninhabitable. Concurrently, smart sensors and data analytics have enabled precision agriculture and detailed environmental monitoring.

The importance of data in both enabling these technical changes and in potentially remediating the more pernicious effects of others cannot be overemphasized. With that backdrop, the remainder of this paper discusses the scale and scope of big data, the privacy and legal challenges created by digital data flows, and the emerging issues surrounding sensors and passive data. It concludes with some thoughts on a new model of digital privacy, one that combines bounded lifetimes, limited sharing transitivity, and claims-based access.

Data Gets Big

The phrase "big data" has been widely adopted to describe the explosive growth of digital data from a wide variety of sources. Big is, of course, a relative term, depending on both context and use. Just as New Orleans, LA is a "big city" if one lives in rural Iowa, New Orleans is a small city by comparison with Chicago or Beijing. More practically, "big data" means the data volume exceeds the utility and efficacy of the traditional tools used in the relevant context. For a small office, that might mean exceeding the expertise of local users with standard desktop tools. For an academic, government, or business, it might mean exceeding the capabilities of the organization's enterprise storage systems.

The growth of "big data" has been both driven by and aided by the rise of e-commerce, smartphones, and commercial cloud services. Anyone who has browsed the web, updated their social network, purchased an item online, or used a business service, has relied on a rich array of cloud services. In turn, these cloud services operate atop a network of massive data centers. Built by Amazon, Google, Microsoft, Facebook, and others, each data center exceeds the scale of the entire Internet just a few years ago. As such, each costs hundreds of millions of dollars to construct and contains tens of thousands of servers. Each of the major cloud vendors operates a worldwide network of such data centers, serving both consumers and enterprise customers.

Privacy, Ethics, and Law in the Digital Age

As data has become digital and migrated from local devices to the global cloud of data centers, law and policy have struggled to keep pace. Our legal notions of privacy and security are all rooted in the concepts of person and place. In the United States, these derive from the castle doctrine and English common law. We expect that our homes are legally secure, protected from search and seizure without due process. Likewise, we expect our physical selves to be also be legally inviolate.

In contrast to these physical concepts of person and place, where location defines the governing laws and policies, the data associated with our digital personas can be and often are geographically distributed across a worldwide set cloud data centers. Equally importantly, the decisions about where that data resides rest with the cloud service operators, not the consumers with whom that data is associated nor is it dependent on where the consumer may physically reside.

Not only can the data be in multiple jurisdictions, sometimes those jurisdictions can be in legal conflict. This can be true not only across state boundaries but also across international ones. As a hypothetical example, consider a Kenyan national whom a German company employs. As he or she travels from Nairobi to Berlin, then on to the United States and then China, each time he or she uses her smartphone, he or she leaves a global trail of digital data and an equally complex set of conflicting legal jurisdictions.

As this example illustrates, there deep, profound, and unresolved issues about global jurisdiction and legal applicability. For instance, can a country insist that a cloud service provider produce data regarding one of its citizens, even if that data is stored outside the borders of the country? What about citizens of other countries? Alternatively, must the entity seek legal access in the country where the data is stored? What rights do citizens have to protect their digital personas in each jurisdiction? How are conflicting legal expectations managed?

These are not merely hypothetical questions; they are issues currently being litigated. The U.S. Supreme Court recently agreed to hear arguments in a case involving Microsoft [2]. The U.S. sought access to data on a U.S. citizen who was a suspected drug dealer. Microsoft turned over data stored domestically, but it refused to supply the data stored in a data center in Dublin, Ireland, citing the potential precedent that would require it to turn over similar data to other countries.

The high court will also consider whether a warrant is required to access smartphone location data (i.e., the location history of the smartphone based on cellular tower connections). This case illustrates the power of not just data, but metadata. It is not what was discussed on the smartphone that is of value, but where the calls were made.

Although a judicial resolution of such issues would benefit all, the ultimate and appropriate disposition of such questions should be via an update to the *Stored Communications Act of 1986.* Thirty years of technological change have made many of its provisions no longer relevant. To put this in perspective, remember that the first popular web browsers did not appear until the 1990s, smartphones where unknown, and floppy disks were the common medium of data exchange.

Passive and Active Data

The use of email or a smartphone is an active event, requiring the user to engage in an explicit action (i.e., sending an email or making a call). The same is true when making an electronic purchase or using a social network service. In each case, the user has agreed to the terms of service as a condition of use. To be sure, these end user license agreements (EULAs) are often arcane and difficult for a layperson to understand, but they do enumerate the rights of use by both the provider and the consumer. Thus, one may well question whether consent was truly informed, but there is an explicit consent and any use requires explicit action by the user.

The rapid growth of wireless sensors and the Internet of Things (IoT) often removes the element of explicit consent, as data capture is an implicit artifact of other activities. Interconnected smart appliances, networked security systems and cameras, and smart cars all raise questions regarding appropriate and acceptable use of captured data. Who owns the data? Who controls it use?

Wearable health or exercise monitors make these questions intensely personal. Although a user may have accepted the terms of use when the device was purchased, no explicit action is required to generate data; the device captures data continuously during wear and often stores it in a cloud service. More perniciously, this data is highly personal and in a medical context would be protected by the Health Insurance Portability and Accountability Act (HIPAA).

In addition to personal health monitoring, intelligent home assistants can and do capture behavior and context from daily life. Amazon's Echo and Alexa assistant, Google Home and Google Assistant, and Apple's Home Pod and Siri have already raised privacy concerns, with cloud-based voice recognition systems listening for user commands. How the balance of privacy, ease of use, and consumer value will be resolved is yet determined.

Toward a New Model of Data Sharing

As noted earlier, our models of privacy are all rooted in concepts of person and place. Moreover, data sharing is largely a binary choice – to share or not to share. Even when more nuanced policies exist, managing the configuration details is often confusing and difficult.

As we adapt to a brave new world of cyberespionage, corporate and government data breaches, and global data flows, perhaps it is time to reconsider some of our approaches from first principles. These might include more nuanced notions of data ownership, privacy, and security, resting on three principles:

- **Bounded lifetimes.** Today, there are no constraints on how long digital data may persist across the Internet. Once released, it remains as long as any search engine or cloud service algorithm asserts that its retention may have economic value. Employment recruiters and government agencies both exploit this to conduct background checks. Instead, one might attach a lifetime to data at the time of its release, requiring the data's destruction at the end of that period.
- **Controlled Transitivity.** Similarly, data released today is most often available to all entities. Instead, consider a model where data released to an individual or organization has limited transitivity (i.e., it cannot be passed on to others without explicit consent). Thus, one might share a photograph with one person but that individual would be unable to share the photograph with anyone else.
- **Claims-based Access**. Finally, once data is released, there are few limitations on how that data is used. Claims-based access would specify the purpose for which data could be used. Hence, one might make data

available for personal, non-commercial use, but forbid any other uses.

Combining these three ideas creates a more nuanced model for data sharing. As an integrated example, one might share a digital document only with four team members for one week, allowing them to read it but not create copies or share it with others.

Concluding Thoughts

In the early days of the Internet, the *New Yorker* published a cartoon [3] that sparked an Internet meme: "On the Internet, nobody knows you are a dog." The notion of an uncharted frontier, where anonymity ruled, disappeared long ago. Today, ubiquitous sensors and consumer-connected devices; big data from e-commerce, social networks, intelligent assistants, and smart devices; and deep learning, mean the Internet not only knows you are a dog, your smart dog collar shares where you walk and your dogfood preferences are cross-referenced and matched with purchasable chew toys.

Within the broader context of social and technological change, we must ask wise and thoughtful questions about how this data is used and by whom. Only by concurrently considering social implications and technological capabilities can be create sustainable approaches.

References

1. United Nations, *World Urbanization Prospects*, https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf, 2014
2. United States v. Microsoft, http://www.scotusblog.com/case-files/cases/united-states-v-microsoft-corp, 2017
3. Wikipedia, "On the Internet, Nobody Knows You're a Dog," https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog, 2017

Research Planning at Nebraska: Research and Economic Development Growth Initiative (REDGI) 2012-2017

Steve Goddard, Vice Chancellor for Research & Economic Development
University of Nebraska-Lincoln

Research planning is being transformed by the advent of big data and analytics. Robust mechanisms are available to determine whether our research activities and programs are meeting our goals and where to invest our resources, and can provide new ways to evaluate areas of innovation and excellence. The University of Nebraska-Lincoln's launching of the Research and Economic Development Growth Initiative is an example of the use of data and analytics in research planning.

The Context

The University of Nebraska-Lincoln was ranked in 2011 as one of the top U.S. universities in research growth over the previous 10 years. Faculty were publishing in top journals, pursuing large multidisciplinary center grants with greater success and making discoveries that received national and international attention. The University leadership viewed this success as a foundation, and an impetus, for continuing growth in the years ahead.

Chancellor Harvey Perlman in his 2011 State of the University Address emphasized the need for increasing our academic stature and challenged the campus to continue its growth, setting ambitious goals for increasing our growth in research and economic development. The specific challenges presented by Chancellor Perlman included:

- To enhance the quality of research at UNL and increase total research expenditures to $300 million, with at least half of this coming from federal agencies.
- To increase the academic stature of the university and double the number of faculty receiving prestigious national awards and memberships in honorary societies.
- With the establishment of the Nebraska Innovation Campus, to increase the number of faculty working with scientists in the private sector and translating basic and applied research into innovations and job creation.
- To increase enrollment by 20 percent, to 30,000 students.

In addition to the Chancellor's goals, UNL also remained committed to the research growth metrics set by the University of Nebraska Board of Regents, which call for UNL to increase research awards from federal agencies at a rate 20% higher per year than the national average, on a three-year rolling average.

Developing an Action Plan

The Office of Research & Economic Development (ORED) was charged with carrying out the research growth goals, a mission that would require buy-in from research-active administrators, faculty and staff. From the fall of 2011 through spring, 2012, targeted forums with key audiences were held to discuss the key issues related to accomplishing the growth goals. In the 2012 spring semester, more than 300 faculty, administrators and staff participated in focused discussions and open forums, including:

- Deans
- Research Advisory Board
- New faculty hires within the previous 3 years
- Two meetings with department chairs/heads, center directors and deans
- Professorships (named chairs, university professors)
- A meeting open to all faculty meeting, with a webinar available

The forums provided information about the growth goals and solicited input on the strategies and the challenges and/or barriers to achieving the goals. Discussion included how best to increase research quality and academic stature, translating research to innovation, identifying key areas of research growth and new infrastructure needed, and identifying challenges. Following the forums the Research and Economic Development Growth Initiative was drafted by the UNL Office of Research and Economic Development and shared with the Vice Chancellors, Deans, Research Council, Research Advisory Board and the Associate Deans for Research, for their input.

The Research and Economic Development Growth Initiative (REDGI)

The Research and Economic Development Growth Initiative (REDGI) was launched in the summer of 2012, with two broad goals: to enhance the quality and stature of research, scholarship and creative activity at UNL, and to increase the quality and quantity of industry-academia partnerships.

REDGI was a campus-wide initiative that included identifying expected outcomes and defining metrics and mechanisms for reaching them. ORED, in partnership with department chairs, center directors, deans and vice chancellors, would assist faculty in meeting the goals and provide leadership, services and infrastructure to support growth in faculty research and creative activity. ORED made clear that its primary mission is to drive excellence in research, scholarship and creative activity, and that this excellence can be demonstrated through many activities, including research funding, book and journal publications, citations, national honors, patents, business startups and more. But REDGI also would emphasize external funding as key to research excellence, and the primary factor enabling faculty to support more postdoctoral scholars, to hire graduate students and staff and to access other resources to support their scholarship.

A critical component of REDGI would be the use of analytical tools enabling tracking and quantifying of research and scholarly productivity, an increasingly difficult task. ORED had already developed NUgrant, a robust system for tracking research funding, and UNL had acquired licenses for using software systems (e.g., Academic Analytics, Elsevier products, Web of Science) to analyze research productivity using various measures. REDGI would take an inclusive approach, working with the Office of

Academic Affairs, the Institute of Agriculture and Natural Resources and the Office of Institutional Research and Planning to use these analytical tools to better understand UNL's scholarly strengths and relative market position among other research universities. Metrics would be a driving force for REDGI.

REDGI Goals, Metrics and Action Steps

Goal 1: Enhance the quality and stature of research, scholarship and creative activity at UNL

Metrics:

- Increase total research expenditures to $300 million by 2018, with half coming from federal sources
- Increase the number of faculty engaged in extramurally-funded sponsored programs by 10% annually on a three-year rolling average
- Double the number of UNL faculty winning prestigious national awards and memberships in honorary societies

Meeting these metrics would require actions engaging faculty and administrators at all levels, creating faculty development programs, and investing in equipment, infrastructure and technology. A key action was to develop growth plans at the university, college and department (where appropriate) levels aimed at increasing total and federal research expenditures through 2018. These plans would include identifying strategic faculty hires – faculty to provide leadership in pursuing major centers or large-scale funded projects, to fill crucial gaps in research expertise, and to build critical mass for stronger interdisciplinary teams. Enabling pre-tenure faculty to establish high-quality research programs early in their careers and re-engaging associate professors and professors in funded research would also be addressed.

As in any research enterprise, instrumentation and infrastructure are constant needs. Goal 1 emphasized investment in state-of-the-art research instrumentation and infrastructure to increase competitiveness. Tied to this was a campus-wide assessment of technology infrastructure needed to support the growth in research programs. Two areas identified as having high potential for growth were increased sponsored funding in the social sciences, arts and humanities, and developing a process for winning more prestigious national faculty awards.

Goal 2: Increase the quality and quantity of industry-academia partnerships

Relationships with industry are often framed by the potential for academic research to solve a particular challenge or problem. Developing strategic partnerships with industry creates opportunities for basic and applied research collaboration, technology commercialization and new venture formation, resulting in job creation and increased economic development in Nebraska and beyond.

In 2011 UNL had launched development of Nebraska Innovation Campus (NIC), a private-public research campus and designed as a driver of these critical industry-academia partnerships. Fueled by $80 million in public and private investments, it is designed as an innovation hub where companies, entrepreneurs and UNL faculty and students work in a collaborative environment that helps fuel Nebraska's economy.

Metric:

- Increase private-sector funded sponsored program expenditures by 15% annually

Actions to meet this metric required development of growth plans to increase research expenditures from private sector sources, increase proposal submission and subsequent funding from federal programs designed for university-industry partnerships, increase technology commercialization, and university-centric economic development. Nebraska Innovation Campus was critical to this goal, offering attractive opportunities for private sector companies to build synergistic research teams with University faculty and to participate in the education of the next generation of graduates.

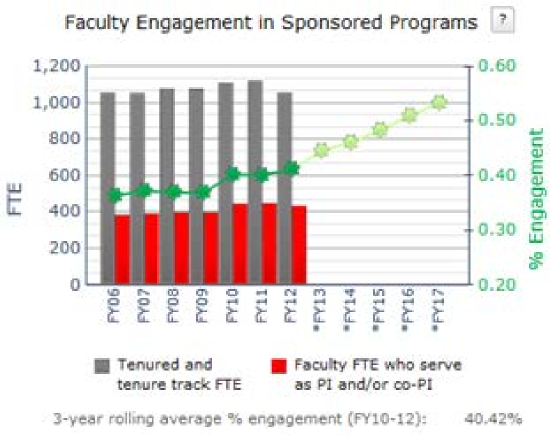
2012: The REDGI Roll-out

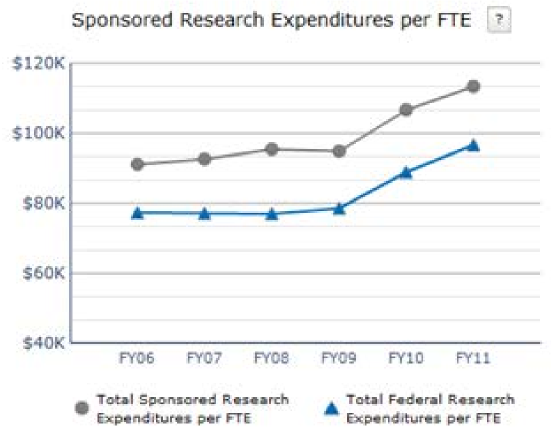
Introducing the campus to REDGI was simultaneously a promotion and education campaign to engage the campus in this new effort and the roll-out of a new platform for disseminating data and analytics that would enable faculty and administrators to measure research outcomes and track their process toward the REDGI goals.

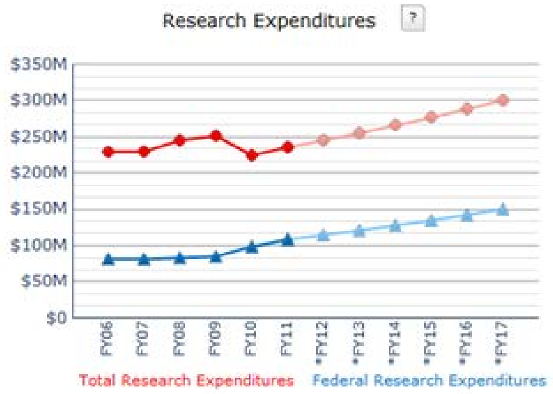
The roll-out included promoting REDGI to all faculty through the UNL and ORED websites, UNL Today news website, ORED's web-based Research News, and other venues. ORED met with research leaders to discuss and present the measurement process and the REDGI "dashboards," the key tools for evaluating progress toward the goals, and offered training in the use of the dashboards.

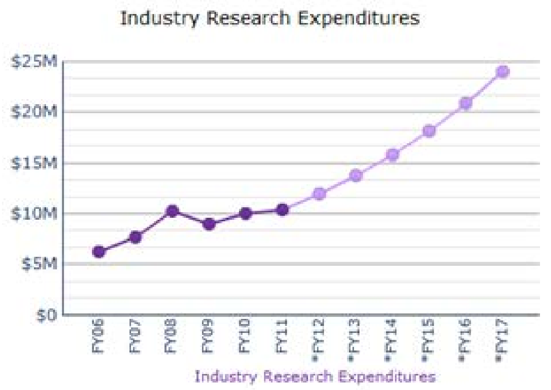
REDGI Dashboards

Dynamic monthly REDGI dashboards were developed and incorporated into the reporting module in NUgrant, UNL's electronic research administration system. Dashboards are available specific to the university, college and department levels. They are easy to access and use, giving a quick snapshot of engagement in sponsored projects, research expenditures, research awards by size and source and other key metrics. Executive leaders receive an updated published binder annually. On the following page is a screen shot of four engagement and expenditure dashboards.









Other Dashboards for other key metrics include:

- Total research expenditures by source: federal, industry, state etc.
- Federal research expenditures by agency
- Federal proposals success rate
- Sponsored proposals and awards per FTE
- Total amount of research award funding by size
- Total number of research awards by size

2017: How did we do?

UNL total research expenditures for just-ended FY2017 reached $295 million, nearly meeting our FY2018 $300M goal. Although we will meet our total growth target we won’t meet our federal goal, largely due to the end of ARRA funding – which occurred during our base year – and the current federal funding environment. The growth goals for industry funding were not met, but we exceeded our goals for faculty engagement in sponsored programs and exceed by almost double the number of prestigious national awards and memberships in honorary societies won by our faculty. Below are some key numbers.

Metrics

- Increase total research expenditures to $300 million by 2017

  FY2018 Goal: $300M total

  FY2016 Actual: $295M total

- Increase private-sector funded sponsored programs expenditures by 15% annually
  FY2018 Goal: $24M
  FY2016 Actual: $13M
- Increase number of faculty engaged in sponsored programs by 10% annually on a 3-yr. rolling avg.
  FY2018 Goal: 53%
  FY2016 Actual: 58%
- Double the number of UNL faculty winning prestigious national awards, memberships in honorary societies
  FY2018 Goal: 14
  FY2016 Actual: 23

Successes Contributing to Future Growth

Nebraska Innovation Campus (NIC), conceived to boost growth of private-public partnerships and university engagement with industry, has become a destination for innovation and is meeting its growth goals. NIC has leased 94 percent of its current facilities and construction has begun on an 80,000-square-foot multi-tenant building to be completed in 2018. A recent Bureau of Business Research report showed that the annual economic impact from NIC business development and operations was $139.9 million in fiscal year 2016. NIC now is home to 20 companies and more than 400 employees.

New research infrastructure on campus includes the Prem S. Paul Research Center at Whittier School, housing interdisciplinary research centers; the Voelte-Keegan Nanoscience Research Center; and Behlen Research Lab, home to defense research.

Lessons Learned from the REDGI Experience

Engaging leaders at all levels is critical to success. A gap in understanding between the Vice Chancellor for Research and some college Deans was problematic. The lack of a regular communication loop resulted in some colleges never really "owning" their metrics, and failing to develop individual strategic growth plans. The college that was most engaged ultimately saw the most success.

Our federal funding goals were set using the last year of the American Recovery and Reinvestment Act as a benchmark. In hindsight, this was unlikely to be an obtainable goal, especially given the current federal funding environment. Ensuring that UNL policies support and encourage an increase in industry-academia partnerships is an important action step.

Incorporating growth goals into the 'story' of our research institution is critically important. The messaging should come from all levels, and many avenues exist for communicating the goals, including mentions and updates in newsletters, annual reports, presentations to faculty, staff, Regents, and other venues. And, as with any new, large undertaking, new staffing is needed to achieve goals. New staff in industry relations is planned and the hire of a national awards coordinator in the past year has already significantly grown our number of awards.

Most important, developing strong, clear data measurements and analytics is critical. The monthly dashboards were particularly effective, especially in units whose leaders "trained" on them, and they are invaluable tools to the staff in ORED.

RETREAT PARTICIPANTS 2017

Keynote Speaker

Michael Huerta, Associate Director for Program Development and NLM Coordinator of Data Science and Open Science, National Library of Medicine, National Institutes of Health

American Speech-Language-Hearing Association

Michael Cannon, Director of Serial Publications and Editorial Services
Margaret Rogers, Chief Staff Officer for Science & Research

Iowa State University

Carolyn J. Lawrence-Dill, Associate Professor, Department of Genetics, Development and Cell Biology and Dept. of Agronomy
Surya K. Mallapragada, Associate Vice President for Research
Joshua L. Rosenbloom, Professor and Chair, Department of Economics

Kansas State University

Ian Czarnezki, Director of Operations, Office of the Vice President for Research

The University of Iowa

Daniel A. Reed, Vice President for Research and Economic Development
David Eichmann, Associate Professor, Director, School of Library and Information Science, Chair, Information Science Subprogram, Iowa Graduate Program in Informatics

The University of Kansas

Mabel Rice, Fred and Virginia Merrill Distinguished Professor of Advanced Studies; Director, Merrill Advanced Studies Center
Richard Barohn, Vice Chancellor for Research, KUMC
John Colombo, Director, Life Span Institute
Teresa Girolamo, Doctoral Candidate, Child Language Doctoral Program
Joseph Heppert, Associate Vice Chancellor for Research
Carl Lejuez, Dean, College of Liberal Arts & Sciences
Kathleen Kelsey Earnest, Data Analyst, Postdoctoral Fellow
Amanda Kulp, Principal Analyst
Claire Selin, Doctoral Candidate, Child Language Doctoral Program
Amy Jurevic Sokol, Associate General Counsel and Risk Manager, KUMC
Russ Waitman, Professor, Department of Internal Medicine, Associate Vice Chancellor for Enterprise Analytics, KUMC

University of Nebraska-Lincoln

Steve Goddard, Interim Vice Chancellor for Research and Economic Development
Jennifer L. Clarke, Director, Quantitative Life Sciences Initiative
Professor, Department of Statistics, Department of Food Science and Technology

Notes