ISSN2641-9084(print)
ISSN2641-9092(online)

Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era

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 2019

Mabel L. Rice, Editor
Technical editor: Suzanne Scales

MASC Report No. 123
The University of Kansas

TABLE OF CONTENTS

MASC Report No. 123

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-third annual research policy retreat hosted by the Merrill Center: Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era. 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 continues to be on opportunities and challenges of big data for research in public universities.

Our keynote speaker for the event was Dr. Daniel Reed, Senior Vice President for Academic Affairs, the University of Utah. He called for an integrative, holistic perspective on the challenges and opportunities of the explosive growth of big data in our lives. Science, social structures, and commercial enterprise merge in unprecedented ways that demand our attention.

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.

Nineteen administrators, faculty, and students from four institutions in Kansas, Arkansas and Nebraska attended in 2019, which marked our twenty third retreat. Additionally, a librarian from the National Library of Medicine/National Institutes of Health 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. The 2017 retreat topic and discussions were on university research planning in the era of big data. The 2018 retreat topic and discussions were on rethinking and re-engineering incentives for scholarly activities across the research enterprise in an open access environment.

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 2019 Merrill policy retreat on: *Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era.*

Executive Summary

Big Data: Big Challenges, Big Opportunities
The Reification of Consilience

Daniel Reed, Senior Vice President for Academic Affairs
University of Utah

- Each individual's Weltanschauung is shaped by the totality of their life experiences and it defines their perspectives, philosophy and understanding of the cultural, economic, and scientific milieu. An explosive growth of knowledge has had a negative effect on one's ability to see the integrative whole. The original seven liberal arts have given way to the speciation of disciplines. As a result, academics feel the deep loss to satisfy their needs for convergent conversation and reflection. Disciplinary, interdisciplinary, and transdisciplinary collaboration and shared insights are needed for a multitude of technological revolutions and socioeconomic disruptions. The emergence of big data and machine learning are potential opportunities to holistically reunify divergent domains.

- Three socioeconomic and technical developments have led to the explosive growth of data. First, interconnected mobile devices and the associated growth of social media have created volumes of consumer data of economic value. Second, new scientific instruments are changing the nature of academic research. With large scientific data readily available, hypothesis-driven experimentation is now being complemented by exploring what might the existing data reveal. Third, small sensors such as consumer health devices, environmental monitors and connected household objects are providing a rich source of data for understanding human behavior and interactions.

- The recent rise of machine learning depends on the confluence of rich sources of data, low-cost high-performance computing and deep learning. Deep learning's recent success depends on large volumes of training data and powerful computing systems. These tasks, once solely in the human cognitive domain, raise social, economic and ethical questions.

- The explosive growth of data has brought about many challenges such as issues with data retention. Security, privacy, and bias loom large in big data and machine learning discussions, particularly individuals' data. Big data and machine learning are accelerating, and we must recognize there are benefits as well as unexpected consequences. It will take an engaged and thoughtful debate to define both a social consensus and legal and ethical framework.

Growing Diversity in Data Science: Shared Lessons from Clinical Trials

Robert D. Simari, MD, Department of Cardiovascular Medicine
University of Kansas School of Medicine
University of Kansas Medical Center

- Data science must evolve with the social changes that are underway. Individuals who make decisions on how data sets are generated and analyzed make decisions based on their life experiences. In this paper, the importance and challenges of diversity in clinical trials is applied to the emerging field of data science.

- The lack of diversity in clinical trials can be attributable to factors such as lack of trust, and social and financial barriers of diverse populations. To overcome this, an area of important focus is the diversity of the investigative team, which may also be paramount in data science. Diversity in data science is important because it may limit bias in data sets and the analysis of the data. The likelihood that data science focuses on the social and medical issues affecting diverse groups is enhanced with a diverse investigative team. The data science workforce will suffer if diverse groups are not considered for advanced training in the field.

- There are similar challenges to developing diverse data science teams as there are to diverse clinical trial teams. Prepared doctoral graduates are needed to enter both fields and there remains a lack of diversity in the disciplines. Efforts need to be focused when students begin to develop aptitude for the STEM fields. The University of Kansas Medical Center has engaged with the Kansas City, Kansas community with a series of programs that attempt to meet the cornerstones of what such programs should include. These programs include a university run Head Start program, faculty involvement in multiple public K-12 programs, and faculty and staff advocacy that enabled sidewalks and grocery stores to be built in food deserts in the community.

Quantifying Biomedical Data Reuse in an Open Science Ecosystem

Lisa Federer, PhD, MLIS, Data Science and Open Science Librarian
Office of Strategic Initiatives, National Library of Medicine,
National Institutes of Health

- The amount of data that is available today has exploded due to advances in a range of research disciplines. Not only do we have more data today, it is freely available. More data sharing is the result of the adoption of policies requiring researchers to share their data. Though not all researchers share their data because they are required to, some share as a move toward open science. Part of the trend of open science includes open access publications, but it also encompasses digital objects across the research lifecycle including data and code.

- With advances in technology, researchers have a wealth of public data available to them. The universe of publicly available research is vast with the National Library of Medicine (NLM) playing a significant role providing access to biomedical data. NLM is just part of the data sharing picture with the National Institutes of Health (NIH), institutional repositories and generalist repositories also contributing.

- Time, effort and funding has made research data publicly available, though what happens with the datasets remains unknown. Understanding data reuse can pave the way for rewarding researchers. Article citations are a measure to quantify scientific impacts, yet data and code are research outputs also meriting reward. Major research funders have formally recognized datasets as research products to demonstrate researchers' impact.

- Data are an important research output that merits a reward system to incentivize sharing. There are technological challenges that hinder rewarding data sharing. Though attempts have been made to standardize data citation, adherence remains low and there is still debate if data citations are the appropriate acknowledgement. Despite not having a reliable way to quantify data reuse, the future reward of data sharing is something to be considered. Not too far in the future, data reuse will feasibly be tracked and quantified and it is important to think about metrics to use for giving credit and rewards for the reuse of data.

Journal Programs and Cross-Disciplinary Research

Marianne Reed, Digital Initiatives Manager
University of Kansas Libraries

- In order for innovative cross-disciplinary research to find its audience, it must be easily discovered by scholars, professional practitioners, and the public. Journal publishing programs in libraries operate under the principle that investment in open access publishing of quality peer-reviewed research is the best way to make that research visible to a global audience and to shift control of publishing from commercial entities to the academy. Library publishers are therefore not constrained, as commercial publishers are, by the need to publish only research that will ensure a profit. This means that library publishing programs can provide a home for cross-disciplinary journals that break new ground and that may take time to find an audience.

- The lack of a profit imperative for library publishing programs also means that the platform for hosting journals is provided to journals at little or no cost, which makes library publishing very attractive to editors looking for a place to publish a new journal. Once the infrastructure is operational, the cost to add a new journal to the system is negligible because the costs of maintaining the technology are already covered. This lowers the financial barriers to starting new journals, allowing editors to focus on the task of finding and publishing excellent peer-reviewed research instead of fundraising.

- Journal platforms used by library publishers are designed so that journals published on those systems automatically follow best practices and standards, such as those outlined by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-MPH) that make the content readily discoverable by internet search engines. These platforms also integrate the use of machine-readable licenses that clearly indicate how the content can be used. In addition to infrastructure that ensures visibility, library publishing programs benefit from existing library expertise in collaboration, technology, copyright, data management, scholarly publishing, information literacy, digital preservation, and the effective promotion of online research.

Convergence Research in the Age of Big Data: Team Science, Institutional Strategies, and Beyond

Daniel Sui, Vice Chancellor for Research and Innovation
Jim Coleman, Provost and Executive Vice Chancellor for Academic Affairs
University of Arkansas

- With the explosion of big data in the last ten years, discussions on interdisciplinary research now emphasize convergence research through a team science approach. The authors of this paper present how to facilitate convergence research in the age of big data by exploring the concept of convergence research, outlining key elements of a team science approach, and discussing institutional strategies, opportunities and challenges.

- More than ever we need interdisciplinary collaboration and team work to address the multiple challenges to society which cannot be resolved by any individual discipline. Big data and data science are emerging as the fourth paradigm, following the previous three paradigms in empirical, theoretical, and computational approaches to science. Now is the time for higher education to conduct convergence research through big data and team science to address grand societal challenges that reach beyond traditional boundaries.

- By emphasizing the need for convergence research, the authors are not abandoning the need for traditional-based research and individual-based inquiries. We need more cutting-edge discipline-based work to enhance our convergence efforts. All research must be conducted individually at some point, even in large team projects. Through the dialectal process of convergent/divergent, disciplinary/interdisciplinary, individual/team-based approach, our research enterprise has been propelled to a level of excellence to make the world a better place for all.

Making Mountains out of Molehills: Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era

Daniel Andresen, Director, Institute for Computational Research in Engineering and Science. Professor, Department of Computer Science
Eugene Vasserman, Department of Computer Science
Kansas State University

- In this paper, the authors present a "Researcher's Hierarchy of Needs", based loosely on Maslow's "Hierarchy of Needs" in the context of interdisciplinary research in a "big data" era. As in Maslow's model of needs, those needs at higher levels can only be expressed if lower levels needs are met. In the research environment, these levels are shared vision, social capital/relationships, domain expertise, technical expertise, and data and software. In this researcher's model of needs, two researchers who have a shared vision but lack the data for their research will be unsuccessful. Considering the hierarchy model for researchers, they suggest that researchers and institutions recognize that interdisciplinary research is both difficult and rewarding.

- There are several overarching issues when considering interdisciplinary research centered around big data. Interdisciplinary research is extraordinarily challenging in universities where the environments are typically siloed by departments in which individuals within the unit communicate at a much higher level. Cybersecurity and privacy present more challenges as data sharing occurs between research groups. So, too, research environments are lacking in the support needed for today's interdisciplinary research. Those institutions who can best support researchers will have a strong competitive advantage. The molehills need to be converted to mountains at every level of the hierarchy: infrastructure should be well resourced, professionals trained in big data across disciplines, and institutions should be committed to a planned and systemic infrastructure.

Training for Cross-Disciplinary Research and Science as a Team Sport

Jennifer L. Clarke, PhD, Professor, Food Science and Technology, Statistics
Bob Wilhelm, Ph.D., Vice Chancellor for Research and Economic Development
University of Nebraska-Lincoln

- Members of the University of Nebraska, a land grant highly research active university, recognize the increasing significance of data and computing across disciplines. With faculty, postdoctoral scholars, and students working together in a cross-disciplinary environment and leveraging advances in data and computing, we can further institutions and make discoveries that benefit humankind.

- This way of thinking—scientific research as a team sport for communal benefit—represents challenges to faculty such as lack of knowledge and resources to plan for use of data beyond their own projects. Another challenge includes the faculty time and effort to prepare data or code to meet the guidelines for proper data sharing. A more sustainable model for cross-disciplinary data services and management is needed, as it is difficult for researchers to secure all the financial support needed. Proper documentation and sharing of code are a requirement for some publications, and institutions are challenged with how to support researchers with meeting this requirement. The University of Nebraska-Lincoln is addressing this campus need with the training of individuals with advanced training in data science who manage multiple projects across disciplines. These application specialists will be tasked with facilitating transdisciplinary research through data knowledge and advanced cyberinfrastructure. As repositories of institutional memory, they will enable the use and reuse of data and code from university projects.

- Research has evolved into a team sport with members from various disciplines working toward a shared goal, enabled by advances in data science and cyberinfrastructure. Institutions of higher education must enable convergent research through avenues of support such as: reproducibility of data, University Libraries, application specialists, and strategic investments in transdisciplinary teams.

Protecting the Value of Interdisciplinary Collaborations in the Development of a New Budget Model

Carl Lejuez, Interim Provost and Executive Vice Chancellor,
University of Kansas

- If you want to know an administrator's priorities, you need look no further than the budget. In its efforts to build a more stable and fiscally healthy institution, the University of Kansas developed a new budget model. A Responsibility-Centered Management (RCM) model, or a hybrid of it, has been adopted by many U.S. higher education institutions. This model offers a decentralized budget with a percentage of revenue controlled by the unit that generated the revenue. KU refers to its hybrid RCM, where less than 100% of the funds are returned to the unit, as a Priorities Centered Management (PCM) model. Our budget is aligned with our priorities including research, student success, career development, outreach, and diversity, equity and inclusion.

- KU had a $20 million budget reduction in Fiscal Year 2019. A series of townhall meetings were held to educate the Lawrence campus community on how the new model will align resource allocations with strategic priorities. A working group, in consultation with campus leadership, developed and shared guiding principles for the new budget. The overall PCM model was enhanced with meetings with the provost's direct reports, town hall presentations, and meeting with faculty, staff and students.

- The new budget allocation model will take effect in Fiscal Year 2021. The first structural feature of the model is the creation of three broad categories in which budgetary resources can be allocated: 1) foundational priorities, 2) institutional strategic priorities and 3) units allocations, including academic and support units. The funding for academic units will be based on performance in a set of priority areas. Additional budget strategies include subsidies outside of unit allocation and implementation of guardrails to reduce the impact of potential budget fluctuations to units. The PCM Model provides support for interdisciplinary collaborations at a time when there is great awareness of the benefits of interdisciplinary initiatives.

Cross-Disciplinary Research: From Nuclear Physics to Cosmic Ray Detection and Medical Applications

Christophe Royon, Foundation Distinguished Professor
Tommaso Isidori
Nicola Minafra
Department of Physics and Astronomy
University of Kansas

- The Large Hadron Collider (LHC) at Cern, Switzerland is the highest energetic collider in the world. The collider provides a better understanding of proton structure and reproduces conditions as close to possible to the Big Bang, where new particles might be produced General purpose detectors including ATLAS and CMS are large detectors built to identify all kinds of particles produced after the interactions. Recently, strange events were observed at the LHC where protons are found to be intact after interacting, though they lost part of their energy. This means that it is possible to detect intact protons after the interaction with detectors.

- At the University of Kansas (KU), fast silicon detectors together with their readout electronics have been developed to achieve this goal. At KU, multipurpose electronics boards were designed to measure precisely the time when particles cross the detector. A test-stand was built to test the full chain from the detector to the read-out electronics. The amplifier that was designed at KU can be used for a full range of detectors and applications. The performance of the amplifier designed at KU is better than commercial ones and the cost is much lower.

- Three possible applications using Ultra Fast Silicon detectors and electronics that were developed at KU are discussed. The first is a project in collaboration with NASA to measure within one single detector the nature and the energy of cosmic ray particles originating from the sun. This will eventually help with the precise measurement of radiation between Earth and Mars, needed to send astronauts to Mars. The second application will measure radiation in cancer treatment with millimeter squared precision and, if successful, will allow a more optimized dose during treatments. An additional medical application deals with PET imaging. The third application is a better understanding of catalysis in chemistry. This could have implications for the way medicine is absorbed and improve the interface of human cells and medicine. It will also improve the methods to desalinize sea water.

Complexities of Conducting Cross-Disciplinary Biomedical Research

Jennifer Larsen, MD, Vice Chancellor for Research
W. Scott Campbell, PhD, Senior Director of Research and IT
University of Nebraska Medical Center

- Solving complex health related problems requires large teams with a broad range of skills. There are many "complexities" that must be considered when building an effective team for cross disciplinary biomedical research. An effective team must define the rules of engagement, which takes time and effort. Teams should form an environment where all members are understood which includes not only the vocabulary in conversations, but also using a common format for capturing and storing data. An effective multidisciplinary team includes team members with terminology expertise and the ability to translate between disciplines.

- Another complexity in cross disciplinary biomedical research is data transfer and storage. More teams are working with large research files which need to be stored, and moving the data is time consuming. Data sharing can create new risks if those researchers who are sharing data are not as knowledgeable about privacy and security issues. Protected health information, protected individual information, as well as other sensitive data might require special controls for the access of data, as well as the ability to audit who has accessed data.

- There are special considerations with global sites, teams or focus. There are increasing and changing rules and regulations on moving data, samples, equipment or team members between countries. Lastly, the public needs to be part of the communication before and after data is shared, to understand the value and make use of the of the results that are found.

Big Data: Big Challenges, Big Opportunities
The Reification of Consilience

Daniel A. Reed, Senior Vice President of Academic Affairs
University of Utah

Each individual's *Weltanschauung* is shaped by the totality of their life experiences and it defines their perspectives, philosophy, and understanding of the cultural, economic, and scientific milieu. (The phrase "world view," the nearest English equivalent, seems rather prosaic by comparison.) Each perspective is necessarily constrained, bringing biases, both explicit and implicit. If in doubt, spend a bit of time looking at a simple doorway—any example will do—and think about what you truly see.

Beyond the superficial, a humble doorway, like many human objects, embodies large fractions of our culture, history, and innovation: protection and the rule of law; security, privacy and mathematics (locks); metallurgy and materials science; environmental systems and fluid dynamics; trade and economic specialization; manufacturing and replication; microbiology and cellular structure (wood); human social dynamics and structures; art, design and esthetics; paint, chemistry and polymers; and mechanical advantage and physics, to name just a view. On any cursory examination, each of us typically sees but a few of these things, lost in the minutiae of daily life, but they remain there to see despite our obliviousness.

The explosive growth of knowledge has had similar, deleterious effects on our ability to see the integrative whole. Intellectual consilience is increasingly obscured by increasing specialization and the seeming triumph of reductionism over holistic perspective. Is there any academic anywhere who has not heard or repeated the old joke, "You learn more and more about less and less, until you know everything about nothing, then they give you a Ph.D.?" (See *Simplifying Communication* and *Shaping the Message, Using the Medium.*)

Humor aside, the original seven liberal arts, the *Trivium* (grammar, logic, and rhetoric) and *Quadrivium* (arithmetic, geometry, music, and astronomy), have given way to the repeated speciation of disciplines, each with their own arcane argot, incomprehensible to all but the speciated initiate. Yet the three big and enduring questions about matter and the universe, life and its processes, and the human condition are deeply intertwined. How did it all begin? How does it work? How will it end? What does it mean? Philosophy, ethics, mathematics, the physical and biological sciences are all elements of our doorway, Plato's Cave manifest in new ways. (See *Eudora, You Got the Love?)*

As academics, we ardently seek to be the embodiment of Raphael's *Causarum Cognitio* (*The School of Athens*) when disciplinary isolation means Pieter Bruegel the Elder's *The Tower of Babel* may often be more apt. The concomitant loss of a *lingua franca*, an ontology of shared discourse, and a deep and binding epistemology of knowledge endanger our ability and our deep need for convergent conversation and reflection.

CRISPR and gene editing, climate change and the Anthropocene, technological revolutions, and socioeconomic disruption all cry out for disciplinary,

interdisciplinary, and transdisciplinary collaboration and shared insights. Across this cacophony, the emergence of big data and machine learning is a potential Diogenean lantern, illuminating a mechanism to reunify divergent domains in holistic ways, an enabler for collaborative Renaissance teams. (See *Renaissance Teams: Reifying the School at Athens*.)

As with any new approach, the combination of big data and machine learning brings both great opportunities and equally grave risks. Data from disparate academic and social sources can be used to predict when students may struggle, but must be used wisely lest privacy be compromised or bias be introduced. Similarly, social and e-commerce data can be used for targeted advertising and product marketing, but must never be used to discriminate against certain groups. Finally, data from multiple scientific domains can be used to glean insights into complex interdependent phenomena such as the effects of human behavior on climate change.

The Rise of Big Data

One of the enduring lessons of computing is that quantitative change begets qualitative change, with viability determined by the ratios of speeds, capacities, costs, and market scale. The smartphone of today embodies the same principles as the mainframe computers of the 1960s. Dramatic shifts in both component capacity and performance made what were then room-sized, multimillion dollar systems available now for hundreds of dollars to billions of people, with data volumes dwarfing those heretofore available. (See *The Zeros Matter: Bigger Is Different*.)

Put another way, today's smartphone is more powerful and more interconnected than the supercomputers of yesteryear. The iconic Cray X-MP supercomputer of 1985 cost roughly $18M in 2018 U.S. dollars, with a peak performance of 800 million floating point operations per second (800 megaflops) and a 56 kilobit/second network connection. By comparison, a 2017-era Apple iPhone costs roughly $700 (U.S.), has a performance of roughly 3000 million floating point operations per second (3000 megaflops), and a broadband network speed of roughly 5-10 megabits/second that can access the vast network that is the Internet. (See *HPC, Big Data and the Peloponnesian Wars*.)

Similarly, big data denotes data of a volume and scale that dwarfs government and enterprise data scales of prior years, made possible by the same quantitative technological changes. Commercial terabyte data stores were once nation-scale resources; today, they are consumer storage devices. This explosive data growth rests on three socioeconomic and technical developments. First, ubiquitous, interconnected, mobile computing devices, and the associated growth of social media and e-commerce, have created enormous volumes of consumer behavioral data, whose large economic value have been unlocked by predictive machine learning. Every major corporation and many governments and universities now leverage this data to tailor marketing messages and to shape products and services.

Second, new scientific instruments, themselves enabled by quantitative computing changes, are transforming the nature of academic research. As an example, the Large Synoptic Survey Telescope (LSST),[1] is designed to survey the southern sky and help understand dark matter and dark energy and the formation and structure of the Milky Way and will produce tens of terabytes of sky survey data each night and petabytes per year. Analogously, National Ecological Observing

Network (NEON)[2] is a continental-scale observation facility designed to collect long-term open access ecological data to better understand how U.S. ecosystems are changing.

In science, the rise of big data has profound social and potentially democratizing implications. For most of scientific history, scientific data has been both difficult and challenging to obtain. Indeed, the experimental method—hypothesis, experiment, theory—is rooted in the capture of new data to validate ideas. As Richard Feynman once described science, a researcher guesses at a law that would explain the currently inexplicable, they then derive the consequences of the putative law, then they make further observations to see if the consequences predicted match the reality now found. (See *The Epistemology of Science*.) With large volumes of scientific data now readily available, hypothesis-driven experimentation is now being complemented by an abduction inversion—what interesting things might the existing data reveal?

Third, concurrently with the deployment of a modest number of large-scale scientific instruments, very large numbers of small, inexpensive sensors are being deployed worldwide. This Internet of Things (IoT) now includes billions of consumer health devices, environmental monitors, and smart and connected household objects (doorbells, cameras, and thermostats), each a rich source of data for understanding human behavior and interactions.

The opportunity posed by heterogeneous big data is obvious—statistically rare events are manifest at scale, and the fusion of data from multiple sensors and domains offers opportunity for correlation and holistic understanding. Yet big data's very scale brings challenges, for humans are rarely either accurate or effective in repetitive, manual analysis. Technology for producing and recording data is of little value unless there are effective ways to extract insights from it. As the late Nobel Laureate, Herbert Simon wisely noted,

> *What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.*

The goal of machine learning is to focus limited human attention on salient data attributes, while automating the laborious and error-prone attributes of data processing. From experiment and theory to computational modeling, this new model of data-driven exploratory discovery has been called the fourth paradigm of scientific discovery.[3]

The Machine Learning Revolution

Machine learning—the use of computing technology to glean insights from data, identify patterns, and make decisions with little or no human intervention—is an idea that dates to the very beginning of modern computing. Its recent rise depends on the confluence of rich sources of big data, inexpensive, high-performance computing, and deep learning. Though the latter is but one of a wide range of learning techniques, deep neural networks have transformed many notions of practical machine learning.

Though a detailed description of machine learning techniques is beyond the scope of this brief review, it is instructive to view deep learning as a subset of machine learning, which is itself a subset of artificial intelligence. As the name suggests, deep neural networks (DNNs) were inspired by biological neural networks and consist of many levels of sim-

ple algorithmic neurons that progressively identify and extract higher level features from their inputs. Thus, each level of a network trained to recognize faces might move from pixels to edges, then features, then faces.

There are many variants of DNNs, each with strengths and weaknesses and varying applicability to specific domains (e.g., handwriting, speech recognition, image and feature identification, drug discovery, advertising, fraud detection, or computer vision). As noted above, deep learning's recent success depends on large volumes of training data (big data) and powerful computing systems, particularly GPUs and targeted hardware such as TPUs[4] to support DNN configuration and training. Once trained, DNNs can then be deployed on much more modest hardware with new data, yielding highly accurate predictions and identifications.

More recently, the appearance of Generative Adversarial Networks (GANs)[5] has led to breakthroughs in both competitive games and in automated digital object creation. As the name suggests, GANs consist of two neural networks in competition; a generative network creates candidates, and a discriminator network evaluates them. As an example, one might use a GAN to create artificial images of human faces, where the generator creates the facial images and the discriminator accesses them for validity. Using GANs, Google's AlphaZero system has automatically learned winning strategies for games such as chess and Go.[6]

The automation of many tasks long considered solely in the human cognitive domain raises many important social, economic, and ethical questions. The data used to train DNNs can introduce bias (e.g., by training facial recognition systems with images lacking a wide range of ethnic backgrounds). Likewise, the creation of "deep fakes" (news, images, or videos combining algorithmically generated attributes with actual images or videos) can be used to sway social or political sentiment or to facilitate fraud.

Challenges and Opportunities

As with any change, the explosive growth of big data has brought a new set of challenges. What data should be retained and for how long? Who pays for the data retention and for how long? How does one ensure data accessibility for long periods, particularly as storage and retrieval technologies continue to change rapidly? Industry, government, and academia all struggle with this balance, as do consumers. Magnetic tapes, floppy disks, and tape cartridges all had their day; few working readers remain. Unlike books and papers, digital data must be repeatedly transferred to new media to be preserved.

In reality, there are few economic or social incentives to retain data for long periods, particularly when in many domains, the costs to acquire new data are so low. The exceptions of course, are when the data is the record of a rare or non-reproducible event or when the costs of data reproduction are exorbitant. Once the disciplinary value of data dissipates, creators often have little incentive to retain the data.

More perniciously, the long-term value of data may accrue to those other than the creators or maintainers, particularly when insights are gleaned from transdisciplinary data fusion. Maintaining metadata is equally important, as it defines the provenance and content for data capture. In many cases, the metadata is as valuable, and sometimes more valuable, than the data itself, as it shapes the context for data fusion and integration.

New government policies for data preservation, particularly for experimen-

tal reproducibility and validation, together with rising data volumes are placing unprecedented financial and regulatory pressures on academic institutions for data management. (See *Research Data Sustainability and Access*.) Historically, one of the fundamental functions of libraries has been triage—deciding what materials should be discarded, which should be preserved, and which should be monitored closely for future evaluation.

It is imperative that the analogs of such policies in the digital age be defined based on thoughtful experiment and assessment. Simply put, we need an interoperable research and data marketplace that exposes and sustains true costs and benefits, recognizing that the costs of data preservation are not self-similar across temporal and spatial scales.

Finally, security, privacy, and bias loom large in any discussion of big data and machine learning, particularly data associated with individuals. (See *Information Privacy: Changing Norms and Expectations.*) Who is liable when data breaches inevitably occur? How are disparate international laws reconciled with transnational data flows? When and where do you have the "right to be forgotten?"[7] Who controls use of an individual's data and when is consent required? How can we best determine the bias or lack of bias in machine learning predictions? How are these policies tested and validated?

Final Thoughts

The big data and machine learning revolution is accelerating. Beyond its institutional effects, the consumerization of artificial intelligence, with deep neural networks now embedded in Internet-connected consumer devices—edge AI—is reshaping the nature of computing and society. (See *Come to the Supercomputing Big Data Edge*.)

Targeted face recognition systems such as Amazon's DeepLens are now available for ~$250 (US), and any technically savvy hobbyist can build an equivalent device for ~$100 (US) using a Raspberry Pi computer and open source face recognition software, with concomitant privacy risks. The same technology, however, is enabling improved cancer detection via feature identification and urban environmental monitoring and smart cities.

As with any new technology, we must choose wisely regarding acceptable use, recognizing that there are always expected benefits and unexpected consequences. Only engaged and thoughtful debate, one dependent on a diverse, educated and engaged citizenry, can balance benefits and risks to define both a social consensus and acceptable legal and ethical frameworks. (See *Public Intellectuals: Seeing the Stars.*)

Our future depends on an inclusive *Weltanschauung*, a reunification of speciated perspectives, one where we all see our doorway to the future as a holistic opportunity and cautionary future. There are no easy answers; there never have been.

The future awaits. Come, let us reason together.

References

[1] Large Synoptic Survey Telescope (LSST), Opening a Window of Discovery on the Dynamic Universe, https://www.lsst.org

[2] National Ecological Observatory Network (NEON): Open Data to Understand How Our Aquatic and Terrestrial Ecosystems Are Changing, https://www.neonscience.org

[3] The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data--intensive-scientific-discovery
[4] N. P. Jouppi, C. Young, N. Patil, and D. Patterson, "A Domain-Specific Architecture for Deep Neural Networks," *Communications of the ACM*, September 2018, Vol. 61 No. 9, Pages 50-59
[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozai, A. Courville, and Y.Bengio, "Generative Adversarial Nets," Advances in Neural Information Processing Systems, pp. 2672-2680, 2014
[6] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play," *Science*, Vol. 362, Issue 6419, pp. 1140-1144, December 8, 2018
[7] Data Protection in the EU, https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

Growing Diversity in Data Science: Shared Lessons from Clinical Trials

Robert D. Simari, MD
Department of Cardiovascular Medicine
University of Kansas School of Medicine
University of Kansas Medical Center

The demographic nature of western society is rapidly changing. In the United States the population is aging and becomingly increasingly diverse (1). Next year, there will not be a majority race among those under 18. By 2060 there will be no majority within the entire US population. The implications of these changes are enormous and the academic enterprise will not be spared. The work of academia and the work force of academia will be forever changed within these ongoing social changes. Data science has the potential to alter the fundamental framework of biomedicine. Machine learning and artificial intelligence have the capacity to identify mechanisms and associations that may lead to innovations in disease prevention and therapy. Yet data science must evolve with the social changes underway.

In the field of clinical trials the diversity of the investigators and the subjects have major impact on the conduct and applicability of the trials. Unlike the experimental nature of clinical trials, data science is observational in nature. Yet decisions in how data sets are generated and analyzed are made by individuals and groups of individuals. The thoughts and actions of each of those individuals is based on their life experiences. As such, diversity of the investigative team can impact the conduct and outcome of data science. In this paper, I will extend the importance and challenges of diversity in clinical trials and apply it to the new field of data science.

In 2018, an editorial in Nature was published entitled "When will clinical trials reflect diversity"(2). In this essay the authors demonstrate the lag between social change and inclusiveness in subjects in clinical trials. In spite of federal requirements and expectations, clinical trial subjects are mainly white and male. Yet, it is widely understood that social determinants of health are the major drivers of health in the United States. As such, exclusion of diverse populations in clinical trials may result in underpowered studies whose results may not be generalizable to the broader population. Further, the exclusion of large segments of our population from these therapeutic trials is not just.

The reasons for a lack of diversity in clinical trials find their roots in social injustices throughout our country's history. Historic injustices in medical experimentation have led to significant lack of trust among diverse populations. This lack of trust has led to challenges in the recruitment of underrepresented communities. Furthermore, there are social and financial barriers that are inherent to inclusion in trials. Missing work, lack of transportation and limited exposure to available trials are often cited as barriers. Progress to overcome this lack of diversity has been slow in spite of the multiple drivers. One particular area of focus is to address the primacy of diversity of the investi-

gative team. Diverse investigators will be more suited to design trials and approaches that favor more inclusion and strategies to recruit broader populations. The need for diversity of the investigative team may also be paramount in developing diversity in data science.

Why is it important to have diverse investigative teams in data science? First, the generation of data bases may be biased by the teams that generate them. Having a diversity of thought, opinion and background may limit known and unconscious bias that can affect the data sets. Second, the analysis of data can be similarly biased without a diversity of thought. Third, the critical social and medical issues that impact diverse populations who bear the greatest burden of social determinants can be impacted through data science. The likelihood that data science focuses on these issues will be enhanced if diverse investigative teams are developed. Finally, the population of majority students in the pipeline will certainly decrease. If diverse groups are not considered for advanced training, the overall workforce in data science will suffer.

The rate of societal changes in diversity is noticeably greater than changes in diversity in academia. Data from the National Science Foundation demonstrate that the gap for underrepresented populations is greater for higher academic degrees than entry level degrees, is greater for men than women and is greater for black and African-Americans than for Hispanics and Latinos (3). Furthermore, these gaps are greater in the sciences associated with data science, mathematics and engineering. These gaps and the rates at which they are lessening do not suggest that the current problem in diversity will be self-limited.

So what strategies might be successful in developing diverse investigative teams in data science? Unfortunately, the challenges faced by the field of clinical trials are applicable to data science as well. Prepared doctoral graduates in advanced fields of science and technology are required. For clinical trials the fields are medicine, nursing and biomedical studies. In data science it is mathematics, engineering, medicine, computer science and statistics. As discussed above, major challenges remain with a lack of diversity in these disciplines.

As the trends in diversity demonstrate, passive approaches will not be successful towards the goal of diverse investigative teams. Intentional and continuous efforts will be required. Additionally these efforts must be focused at the time in which students begin to develop aptitudes for STEM fields. In my opinion the cornerstones of such programs include the following:

1. Coordination among centers of higher education and communities of underrepresented minorities to provide valued and culturally competent programs to support attainment of educational milestones.
2. Exposure of diverse students to careers in data science throughout their pregraduate careers preferably with diverse role models, mentors and sponsors.
3. Coordination with governmental and non-governmental organizations to address social determinants of health, many of which act as social determinants of education and achievement.

At the University of Kansas Medical Center, we have engaged our Kansas City, Kansas community through a series of programs that attempt to meet the cornerstones defined above. These programs start at the earliest stage of development

with Project Eagle, a large university run Head Start program supporting some of the neediest and youngest members of our community. Our faculty are engaged in multiple programs throughout the public K-12 school including a successful grant which funds STEM curriculum development for our local high schools and a summer and Saturday science academy. Finally, through the advocacy of our faculty and staff, we have enabled building of sidewalks and grocery stores in food deserts in our community.

For students interested in health based careers, we have a formal shadowing program, a post bac degree for underrepresented minorities that includes a guarantee of medical school admission and transition to a prematriculation program for successful graduates. Taken together, this series of programs creates an important mechanism for students to meet the educational demands of STEM careers.

The field of data science has the potential to revolutionize academia and our society. Yet it is subject to the unconscious biases currently present in both. We must strive to broaden the diversity of investigators who participate in data science in order to avoid the pitfalls currently present in the field of biomedical clinical trials. This will require intentional and consistent efforts throughout the educational hierarchy and interspersed throughout our communities. The future of data science is dependent on those efforts.

References

1. U.S. Census Bureau (2017). 2017 National Projections Tables. https://www.census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html Accessed 9/5/19.
2. Knepper TC, McLeod HL. When will clinical trials reflect diversity. Nature. 2018 May; 557(7704):157-159. doi: 10.1038/d41586-018-05049-5.
3. National Science Foundation. Women, minorities, and persons with disabilities in science and engineering. Arlington, VA: NSF 17-310, Jan 2017. https://www.nsf.gov/statistics/2017/nsf17310/digest/about-this-report/. Accessed 9/5/19.

Quantifying Biomedical Data Reuse in an Open Science Ecosystem

Lisa Federer, PhD, MLIS, Data Science and Open Science Librarian
Office of Strategic Initiatives, National Library of Medicine
National Institutes of Health

The last decade has seen a significant shift in the ways that academic and research communities think about research data. Data can now be generated more quickly and cheaply than ever before, a phenomenon that is clearly evident in the case of genomic data. The process of sequencing the first human genome, under the auspices of the Human Genome Project, took about thirteen years by the time it was complete in 2003 and cost about $2.7 billion, requiring the collaboration of research institutions from around the world (1). Today, a human genome can be sequenced in about 24 hours at a cost of around $1,000. As a result of such advances not only in the field of genomics, but across the range of research disciplines, the amount of data available today has exploded.

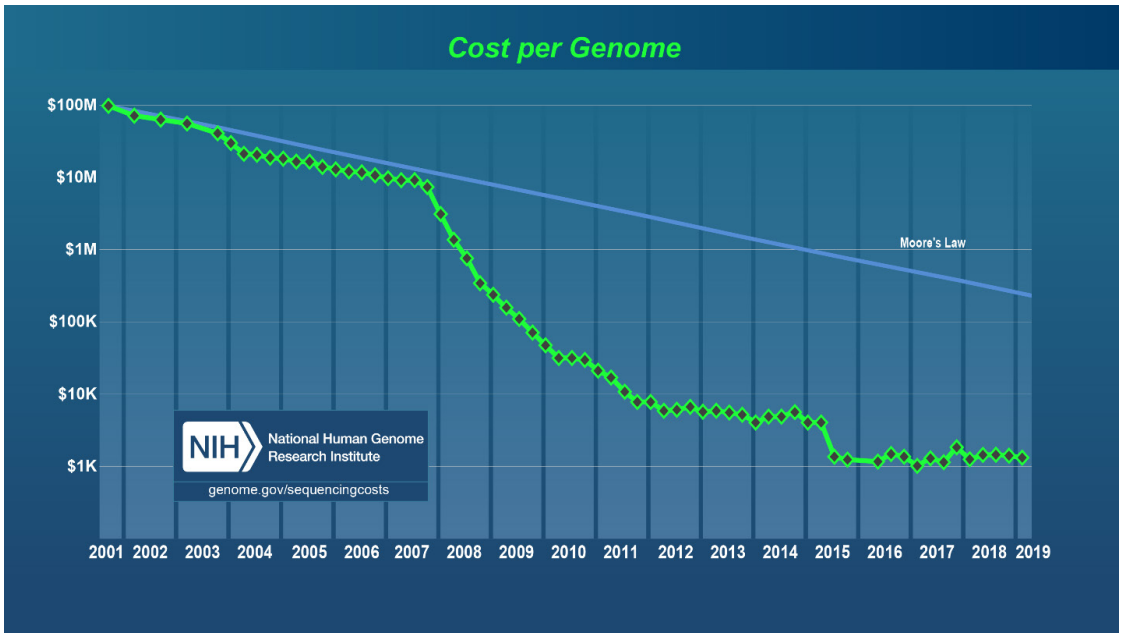


Image source: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Not only do we have more data than ever before, but those data are also increasingly freely available through repositories and other sharing mechanisms. This move toward sharing data has been driven in part by the adoption of policies that require researchers to share their data. PLOS was among the first publishers to adopt such a policy, stating that open access to the literature is only part of making research open, since "without similar access to the data underlying the findings, the article can be of limited use" (2). The International Committee of Medical Journal Editors has instituted a clinical data sharing policy for its

member journals, noting that researchers have "an ethical obligation to responsibly share data generated by interventional clinical trials because trial participants have put themselves at risk" (3). Many funders have also adopted such policies, making data sharing a condition for investigators accepting grant funding. At the National Institutes of Health (NIH), a number of policies govern sharing data of different types; in different domains of research, such as clinical data about mental health (4); specific research initiatives, such as the Human Connectome Project (5); or funding mechanisms (6), with plans underway for an overarching policy on data management and sharing that will apply to all NIH-funded research (7,8).

Not all investigators share their data simply because they are required to do so; a growing number of researchers have adopted sharing practices as part of a cultural shift towards open science, a trend in which research products are made openly available. The move toward open access publications is one part of this trend, but open science encompasses digital objects from across the entire research life cycle, including data and code. Enabling access to research products is seen as a way to "foster equality, widen participation, and increase productivity and innovation in science" (9). In light of recent concerns about irreproducible research, open science practices are beneficial to increasing transparency and thereby enhancing research reproducibility (10).

As a result of these advances in technology and changes in science policy and culture, researchers have a wealth of public data available to them. The National Library of Medicine (NLM) plays a significant role in this data sharing ecosystem. As the world's largest biomedical library, NLM not only houses a comprehensive collection of literature, but also provides access to a wide range of biomedical data through a number of databases administered by NLM's National Center of Biotechnology Information (NCBI). Each day, NLM sends out over 115 terabytes of data to 5 million users, as well as adding to its own data holdings by receiving over 15 terabytes of data from around 3,000 users. While a significant source of biomedical research data, NLM is only one part of the big picture for data sharing. NIH alone hosts or funds over 80 domain-specific repositories, housing data related to a specific disease, of a specific type, or funded by a specific institute of the NIH (11). As of this writing, the Registry of Research Data Repositories (re3data) lists over 1,200 repositories collecting data related to the life sciences (12). Add to this the many institutional repositories that house their investigators' data, as well as generalist repositories, such as Mendeley Data, Zenodo, Dryad, and figshare, that accept a range of data types from various disciplines, and it becomes clear that the universe of publicly available biomedical research data is vast.

Despite all the time, effort, and funding that has been put into making research data publicly available, a fundamental question nonetheless remains relatively unanswered: what happens to all of these datasets? In theory, reusing existing data rather than collecting more yields many benefits to science and society on the whole. Reusing data increases the return on investment of the original funding by yielding additional discoveries and knowledge, as well as saving funds that would have been spent on collecting new, potentially duplicative, data. The time for translation from research findings to life-saving clinical applications may be sped up by reusing existing data rather than taking additional time to collect new. Making data publicly avail-

able can also help democratize the practice of science, enabling researchers who may not have access to large amounts of funding or expensive laboratory technology to nonetheless contribute to knowledge creation.

Understanding data reuse can also pave the way for meaningfully rewarding researchers who share their data. Being able to reward researchers who share might also make data sharing more appealing for researchers who are not necessarily open data enthusiasts. Some researchers consider data sharing a burden or worry that making their data available opens them up to being "scooped," concerns that might be mitigated by providing credit for researchers who share data. Science is, after all, a credit economy; if a research team wants to build on my ideas, they need not pay me to do so, but instead give me credit by citing my article in their publication. While a citation in and of itself has no actual monetary value, it indirectly has very real value as a means for demonstrating a researcher's scientific productivity and impact, which in turn form the basis for career advancement in the form of professional recognition, tenure, promotion, and funding. Citations to articles, though an imperfect measure, are a method to quantify the difficult-to-define concept of scientific impact. However, such measures privilege journal articles as the only research output meriting reward, when in fact other research outputs, such as data or code, can have meaningful impact.

A move toward rewarding data sharing is in large part a culture change that must be driven by stakeholders involved in scientific reward, particularly funders and institutions. Indeed, several major funders, including the National Science Foundation (NSF) and NIH have already formally recognized datasets as research products that can be reported to demonstrate researchers' scientific impact for grant applications as well as progress toward grant aims on progress reports (13,14). Some institutions likewise have begun to consider data and other research products in considerations of researchers' scientific output and impact; the Montreal Neurological Institute (MNI), for example, has adopted an institution-wide open science policy that recognizes shared data as a research output in the tenure and promotion process (15).

However, technological challenges remain that hinder efforts to reward data sharing. Using article citations as a means of quantifying impact works because we have well-established mechanisms for tracking such citations. While the exact citation style may differ from one journal to another, authors generally understand how and where to cite an article, and journals know how to appropriately tag citations to enable them to be tracked by systems that capture citations. The same is not true for data; while groups like FORCE11, CODATA, and the International Council for Scientific and Technical Information (ICSTI) have made efforts to help standardize data citation (16,17), uptake among authors and publishers remains relatively low. In fact, some debate remains about whether data citations are even the most appropriate way to acknowledge the contribution of shared data. Some authors choose to recognize data creators in the article's acknowledgement, and some data creators have argued that they should be co-authors on any papers that arise from secondary analysis of their shared data, although sharing data alone does not satisfy the authorship criteria outlined by the ICMJE (18). In the absence of a widely-adopted standard for citing data reuse, quantifying data reuse is impossible in

practice, so even the adoption of policies that reward data sharing will be difficult to implement.

In the absence of a reliable means to quantify data reuse, it is still worthwhile to consider how we will eventually reward data sharing at some point in the future. Careful consideration of the meaning of data's value and impact may help avoid some of the perverse incentives that have arisen as a result of the ways that bibliometrics are used to measure the impact of articles by citations (19). One issue is that not all citations to an article mean the same thing, yet all are counted equally when it comes to measuring impact by citation count. Eugene Garfield enumerated fifteen reasons for citing articles, not all of them positive, including "criticizing previous work" and "disclaiming work or ideas of others" (20). For example, the paper in which Andrew Wakefield incorrectly connected autism to vaccinations has been relatively highly cited, with 184 citations according to Google Scholar, 76 citations according to Web of Science, and 74 citations according to Scopus. The disparity in citation counts across various platforms presents its own complication, but also problematic is that most of these citations are in the context of articles that discredit his findings, and simple citation counts would not be able to distinguish this article from another that has been cited a similar number of times.

Similarly, not all instances of data reuse are identical. In genomics research, it is a common practice to pool multiple datasets from different studies and different researchers to achieve adequate statistical power, and the standardization of this type of data means it is possible to do so, since data from multiple sources will be largely interoperable (21). Clinical data, on the other hand, is far less standardized, with researchers often recording the same concept using different terminology or phrasing questions to patients in slightly different ways that mean it is often infeasible to combine clinical datasets even if they are on similar topics (22). If a researcher creates a dataset that is reused as one of several hundred combined together in a genomic study, should that reuse be counted the same as a clinical dataset that is used on its own to entirely form the basis for a new study? Datasets themselves may also have different value based on their contents as well as varied potential for reuse. For example, compare a dataset collected from patients with an extremely rare disease and a dataset collected from patients with heart disease, the most common cause of death in the United States. A dataset on a common condition with high disease burden will almost certainly be reused more than one that covers a rare, and therefore likely less-researched disease. However, it could be argued that the rare disease dataset has greater value since it would be more difficult to re-collect such data than it would be to re-collect data on heart disease. Relying simply on counts of data citation makes it difficult to meaningfully reward researchers in ways that recognize the complexity of data collection and research.

As we move toward a future that is not far off when data reuse can be feasibly tracked and quantified, it will be important for institutions, funders, and other stakeholders to think about how to incorporate metrics for reuse into the scientific system of credit and reward. Overlooking data as an important research output that merits its own recognition and reward means we risk disincentivizing sharing. On the other hand, oversimplifying the practice of rewarding data creators for reuse means we risk creating some of the perverse incentives that have

arisen from bibliometrics and led to undesirable research practices like excessive self-citation. It is therefore worth careful consideration now of how we can create a reward system that meaningfully recognizes the place of shared data in the research ecosystem.

References

1. National Human Genome Research Institute. The Human Genome Project FAQ [Internet]. 2018 [cited 2019 Sep 4]. Available from: https://www.genome.gov/human-genome-project/Completion-FAQ
2. Silva L. PLOS' new data policy: Public access to data [Internet]. EveryONE: PLOS ONE Community Blog. 2014 [cited 2017 Apr 2]. Available from: http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/
3. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. N Engl J Med [Internet]. 2017;376(23):2277–9. Available from: http://www.nejm.org/doi/10.1056/NEJMe1705439
4. National Institute of Mental Health. NOT-MH-15-012: Data Sharing Expectations for Clinical Research Funded by NIMH [Internet]. 2015 [cited 2019 Sep 5]. Available from: https://grants.nih.gov/grants/guide/notice-files/not-mh-15-012.html
5. National Institutes of Health. RFA-MH-10-020: The Human Connectome Project (U54) [Internet]. 2009 [cited 2019 Sep 5]. Available from: https://grants.nih.gov/grants/guide/rfa-files/rfa-mh-10-020.html
6. Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). NIH Data Sharing Policies [Internet]. U.S. National Library of Medicine; 2019 [cited 2019 Sep 4]. Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
7. National Institutes of Health. NOT-OD-19-014: Request for Information (RFI) on proposed provisions for a draft data management and sharing policy for NIH funded or supported research [Internet]. 2018 [cited 2018 Nov 11]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html
8. National Institutes of Health. National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research [Internet]. 2015 [cited 2017 Jul 19]. Available from: https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf
9. Levin N, Leonelli S, Weckowska D, Castle D, Dupré J. How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice. Bull Sci Technol Soc [Internet]. 2016;36(2):128–41. Available from: http://journals.sagepub.com/doi/10.1177/0270467616668760
10. Shrout PE, Rodgers JL. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. Annu Rev Psychol [Internet]. 2017; Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-122216-011845
11. Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). NIH Data Sharing Repositories [Internet]. U.S. National Library of Medicine; 2019 [cited 2019 Sep 4]. Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

12. re3data.org. Life Sciences Repositories [Internet]. 2019 [cited 2019 Sep 4]. Available from: https://www.re3data.org/search?subjects[]=2 Life Sciences
13. National Institutes of Health. NIH and Other PHS Agency Research Performance Progress Report (RPPR) Instruction Guide [Internet]. 2017. Available from: https://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf
14. National Science Foundation. Dissemination and sharing of research results [Internet]. Vol. 2012. 2010. Available from: http://www.nsf.gov/bfa/dias/policy/dmp.jsp
15. Ali-Khan SE, Jean A, MacDonald E, Gold ER. Defining success in open science. MNI Open Res [Internet]. 2018 [cited 2018 Nov 11];2:2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29553146
16. Data Citation Synthesis Group. Joint declaration of data citation principles [Internet]. Martone M, editor. FORCE11; 2014 [cited 2017 May 19]. Available from: https://www.force11.org/group/joint-declaration-data-citation-principles-final
17. CODATA. CODATA-ICSTI Data Citation Standards and Practices [Internet]. [cited 2019 Jan 3]. Available from: http://www.codata.org/task-groups/data-citation--standards-and-practices
18. International Committee of Medical Journal Editors. Defining the Role of Authors and Contributors [Internet]. 2017 [cited 2017 May 19]. Available from: http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html
19. Stephan P. Research efficiency: Perverse incentives. Nature. 2012;484(7392):29–30.
20. Garfield E. Can citation indexing be automated? In: Statistical Assocation Methods for Mechanized Documentation. 1964. p. 84–90.
21. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform [Internet]. 2012 Jun [cited 2018 Apr 1];10(2):117–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23105939
22. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: Current status and challenges. J Am Med Informatics Assoc. 2011;18(3):341–6.

Journal Programs and Cross-Disciplinary Research

Marianne A. Reed
Digital Initiatives Manager
University of Kansas Libraries
mreed@ku.edu
https://orcid.org/0000-0001-8452-6103

In the late 1990s, as the cost of commercially-published journal subscriptions increased and library budgets held steady or declined, university libraries began to explore ways to change the scholarly publishing ecosystem through initiatives designed to shift control of academic publishing from commercial entities back to the academic community.[1] One strategy employed was the creation of library publishing programs that supported faculty in the publication of quality open access journals that made peer-reviewed scholarly research freely available to anyone in the world with a computer and an Internet connection. Since they were not driven by profit, library journal programs could publish cross-disciplinary research that had scholarly merit, but was not considered viable by commercial publishers. These programs provided universities with new opportunities to showcase the research of their faculty and to make it available to a worldwide audience.[2]

In 1998, the Public Knowledge Project was founded and began the process of creating the first open-source journal publication system that supported the entire editorial workflow.[3] The Open Journal Systems (OJS) software is available free of charge and can be installed wherever the prerequisite technical requirements are met.[4] This has made it attractive to new publishing programs in libraries all over the world, allowing them to put scarce resources into producing high quality content instead of paying a commercial firm for journal hosting.

Business models of these library journal programs vary, but almost all provide basic journal hosting services free of charge. A few library publishing programs offer additional tiers of fee-based services that further enhance the publication process, such as website customization, graphic design, copyediting, and promotional services. However, many editors do not have outside funding and, instead, collaborate with other scholars who can provide these services free of charge. These collaborations are actually helpful to a new cross-disciplinary journal, since it creates a pool of researchers that are invested in the success of a journal and will be more likely to serve as peer reviewers and to promote the journal throughout their academic networks.

Barriers to Publishing Cross-Disciplinary Research

Financial Considerations

The publication of new cross-disciplinary journals is less appealing to commercial publishers because these journals may initially have a smaller audience and may never be commercially viable. Conversely, some commercial publishers may establish cross-disciplinary journals, only to cease publishing issues when they are no longer profitable. For example, the journals program at the University of Kansas (KU) Libraries was approached by a member of the KU faculty asking the program to take over the publication of an important journal in his field that

was no longer going to be published by a commercial publisher. With the full assistance of the publisher, the journal moved to our OJS system and continues to publish high-quality peer-reviewed research in its new home.

Article publication fees charged by some commercial open access journals to finance journal costs may be a barrier to authors of cross-disciplinary research who do not have outside funding available to them. Lack of outside funding can also be a barrier to research in the humanities or when publishing with international partners that might be unable to obtain funding.

Disciplinary Considerations

New forms of cross-disciplinary research submitted to traditional journals in the field are often rejected because their cross-disciplinary approach does not fit within the stated scope of the journals.

Also, in order for the research in new cross-disciplinary journals to be taken seriously, the editors of new cross-disciplinary journals must be well-respected members of their fields that are willing to risk failure if the journal does not find an audience. Finding scholars that can afford to take that risk may be difficult; untenured faculty who are still establishing their scholarly reputations often prefer to do editorial work for traditional journals in their discipline, because those are the positions that are given the most weight by tenure committees. Senior faculty who serve as editors of traditional journals may be reluctant to take a chance on a new cross-disciplinary journal, because failure of the journal may tarnish their scholarly reputation.

Access and visibility

Commercially published journals are typically kept behind paywalls, where access is by subscription only unless the author has paid the publisher to ensure that the research will be openly available. This limits the audience for that research to those readers, or their institutions, that are able to pay for access.

In addition, as journal prices rise and library budgets decline in purchasing power, less-accessed journals may be dropped from library subscription packages, which limits visibility and access to research in those journals. Interlibrary loan is sometimes used to provide access from the collection of another institution, but those services are not available to all readers.

How Library Journal Programs Support Interdisciplinary Research

Enhanced Visibility

Journal publishing in libraries adheres to standards such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[5] that exposes journal content to the web search engines that make it available to a worldwide audience.

In addition to making content widely discoverable, library journal programs encourage journals to follow best practices in licensing open materials, such as the use of Creative Commons licenses[6] that make it clear to the reader how the research can be re-used without asking for permission from the copyright holders.

The visibility of journals hosted by library publishing programs often results in more submissions to the journal, as scholars interested in cross-disciplinary research discover these outlets for publication. Another benefit of visibility is the possibilities for new research partnerships for authors as other scholars with similar interests find the research.

Lower Cost, Lower Risk

Publishing through library publishing programs is cost effective for the journal, since the lack of a profit motiva-

tion means that the costs of the library publishing infrastructure is subsidized so that journal hosting can be provided to journals free of charge. Traditional library values emphasizing the preservation and dissemination of knowledge extend into their journal programs, which often provide for the long-term preservation of journal content.

Open access publishing eliminates the need for the journal to allocate resources to subscription management, including reminders, renewals, tracking of payments, etc. Some smaller scholarly societies that publish journals find themselves in the predicament of paying more to manage subscriptions than the subscriptions provide in revenue. By moving journals to a library publishing program and making them open access, a scholarly society not only saves those costs, but also provides a wider outlet for the research of its membership.

Once the publishing infrastructure is in place, there is very little cost to the library journals program to publish a new cross-disciplinary journal. This lowers the financial risk to the publisher if a new journal fails to find an audience and ceases publication. Since library publishers have less to lose if they take a chance on a new journal, journals are not pressured to publish a minimum number of articles per issue or to find an audience in a short amount of time. This gives library-published journals the freedom to experiment with new forms and combinations of scholarship that may not be commercially viable.

Existing print-only journals can eliminate substantial printing costs by moving to an online-only model on a library publishing platform. This has the added benefit of making the research in the journal much more visible to an audience beyond its previous subscribers.

Library Expertise

Journals published by libraries benefit from existing library expertise in scholarly publishing, project management, experience collaborating with peers to manage scarce resources, copyright, digital preservation, and adherence to standards for the effective promotion of online research.

Libraries are also experts at building communities of interest. For example, the University of Kansas Libraries' journal publishing program periodically hosts Editor Forums in the Libraries where faculty editors from all disciplines can meet to talk about the challenges of journal editing with other editors. An email discussion list allows KU editors to continue the conversations and ask for advice from other editors if a situation arises between meetings. This community of editors provides support to new editors and allows experienced editors to share their extensive knowledge of scholarly publishing.

Visibility is the Key to Success

The visibility and discoverability of journals in library publishing programs results in a large number of downloads of journal content. For example, articles in the 24 journals hosted by the University of Kansas Libraries on the OJS platform[7] **were downloaded over 2.7 million times in 2019**. See *Figure 1* for a list of these journals, as well as those additional journals that are available through KU ScholarWorks, KU's online institutional repository.

Common Strategies for Visibility

Open Access

- Make as much article content as possible publicly available without a paywall to encourage use and citation.
- Add all older issues and articles whenever possible when publishing an established print journal

Journals and Serials

KU Libraries provides journal editors with the technical infrastructure to publish their journals on either of two platforms: **KU ScholarWorks**, KU's institutional repository, which makes journals visible to a wide audience and assures their long term preservation and **Open Journal Systems (OJS)**, which makes journals visible and assures their preservation, but also supports the entire editorial management workflow, including article submission, multiple rounds of peer-review, and indexing.

- American Studies (OJS)
- Auslegung: A journal of philosophy (OJS)
- Biodiversity Informatics (OJS)
- Center for East Asian Studies Publication Series (KU ScholarWorks)
- Chimères (OJS)
- Digital Treatise (OJS)
- Focus on Exceptional Children (OJS)
- Folklorica: Journal of the Slavic and East European Folklore Association (OJS)
- Human Communication & Technology (OJS)
- IALLT Journal of Language Learning Technologies (OJS)
- Indigenous Nations Journal (KU ScholarWorks)
- Infrastructure Research Institute Reports (KU ScholarWorks)
- Issues in Language Instruction (OJS)
- Journal of Amateur Sport (OJS)
- Journal of Copyright in Education & Librarianship (OJS)
- Journal of Dramatic Theory and Criticism (OJS)
- Journal of Intercollegiate Sport (OJS)
- Journal of Melittology (OJS)
- Journal of Montessori Research (OJS)
- Journal of Russian American Studies (OJS)
- Journal of Undergraduate Research (KU ScholarWorks)
- Kansas Law Review (KU ScholarWorks)
- Kansas Journal of Medicine (OJS)
- KU Field Methods in Linguistic Description (KU ScholarWorks)
- Latin American Theatre Review (OJS)
- Merrill Series on The Research Mission of Public Universities (OJS)
- Midcontinent Geoscience (OJS)
- Novitates Paleoentomologicae (OJS)
- Scientific Papers of the University of Kansas Museum of Natural History (KU ScholarWorks)
- Slavia Centralis (KU ScholarWorks)
- Slovene Linguistic Studies (KU ScholarWorks)
- Social Thought and Research (KU ScholarWorks)
- Special Publication of the University of Kansas Museum of Natural History (KU ScholarWorks)
- Treatise Online (OJS)
- Undergraduate Research Journal for the Humanities (OJS)
- University of Kansas Paleontological Contributions (KU ScholarWorks)

Figure 1: Journals supported by the KU Libraries journal publishing program

online for the first time. The more visible content there is, the more likely that readers will discover the journal.

Make Articles Easier for Search Engines to Find

Another strategy that can used by library journals programs to enhance visibility is to make article information available through other not-for-profit entities so that internet search engines can find journal content more easily. Here are some examples:

- Register DOIs with Crossref[8] or DataCite[9] to provide permanent links to journal articles and, as a byproduct of this process, to make article information more available to search engines.
- Encourage authors to add their Open Researcher and Contributor ID (ORCID) when submitting an article. An ORCID is a unique researcher identification number used to connect research outputs such as articles to a particular researcher. An advantage of using ORCID is not only that researchers with similar names can clearly identify the work that is theirs, but also that granting agencies are starting to use ORCID as an optional way to add faculty publications to grant applications. Rather than typing the information for each publication, integration with ORCID allows the information to be imported directly to the grant application.[10]
- Help journals apply for inclusion in the Directory of Open Access Journals (DOAJ) and, once they are successful, upload article information to DOAJ as it is published.[11]
- Include machine-readable Creative Commons licenses for articles whenever possible to make it clear to readers how the content can be used. Machine-readable Creative Commons licenses also allow search engines to include those articles in results when a user searches for content that is licensed for re-use.[12]
- Follow the best practices in Google Scholar's *Inclusion Guidelines for Webmasters*[13] so that Google Scholar indexes the journal articles.

Editors as Partners

Predatory open access journals that charge authors large fees to publish poorly researched or scholarly research that does not undergo rigorous peer-review have warped many scholars' perceptions of open access publishing as sub-standard. There is a commonly-held belief, possibly nurtured by those that benefit, that only through commercial publishing channels can quality research be published.

However, for all journals, whether published commercially or not, the reputation of the journals are closely tied to the reputation of the editor and editorial team. Indeed, open access journals with respected editorial teams that do not charge author fees have no incentive to publish poor research, as there is no commercial reward for quantity over quality.

Editors can greatly contribute to the success of their cross-disciplinary journals by following a few simple guidelines:

- Journals that are managed by one person trying to do everything are not sustainable and usually cease publication after a few years. Share the burdens of edi-

torship and ensure continuity of the journal by making sure that there is a group of scholars involved in running the journal, soliciting content, and performing reviews. Including other scholars also brings other perspectives to the journal and helps minimize the risks of publishing in a new area of scholarship.

- Ensure that information on the journal website—e.g., focus and scope, author guidelines, copyright information, and publication agreements—is clear and complete.
- Require abstracts to be submitted with articles. More information on the article's page gives readers incentive to download the article.
- Support all readers by following the World Wide Web Consortium's *Web Content Accessibility Guidelines* (WCAG) to ensure that the journal website and journal content are accessible.[14]
- Encourage authors to share their articles in their online institutional repositories with links back to the original published content. Since content in institutional repositories is placed higher in search results by Google's search algorithm, this will increase the visibility of your journal.
- Editors should use their existing scholarly and social media networks to promoting the journal. Sending calls for papers, announcements of special issues, and the publication of new issues to a disciplinary discussion list or forum is an excellent way to call attention to a journal. Announcements at conferences are often very successful, too, especially if connected to a presentation at the conference.
- Solicit articles from well-respected scholars that are acquainted with members of the editorial team. A request from a friend is much more likely to be taken seriously by a reputable scholar.
- Invite selected authors from the journal to be reviewers for future issues.
- The online guide for editors, **Resources for Editors of Scholarly Journals**,[15] is a good starting point for those who are considering starting a journal or who are looking for information about managing an existing journal. For those journals on the OJS platform, the **Using Open Journal Systems** page can be especially helpful.[16]

Summary

Library publishing programs are uniquely positioned to help cross-disciplinary journals prosper. For scholars thinking of starting a new journal or for editors of a journal that is looking for a new home, library publishing programs are the ideal partner to help the journal, and its editors, succeed.

References

1. Chapman, A.L. and David E. Shulenburger. "The State of Research Endeavors: View from the Campus-wide Leadership Level." *Merrill Series on The Research Mission of Public Universities* (1997): 67-68. https://doi.org/10.17161/merrill.1997.8173
2. Steinmetz, Joseph E. "The Role of Universities in Promoting Scholarly Work in the Emerging Open Access World." *Merrill Series on The Research Mission of Public Universities* (2018): 1-9. https://doi.org/10.17161/merrill.2018.9120

3. "History." Public Knowledge Project. Accessed June 16, 2019. https://pkp.sfu.ca/about/history/.
4. "Download Open Journal Systems." Public Knowledge Project. Accessed June 16, 2019. https://pkp.sfu.ca/ojs/ojs_download/
5. "Open Archives Initiative Protocol for Metadata Harvesting." Open Archives Initiative. Accessed June 16, 2019. https://www.openarchives.org/pmh/
6. "CC Licenses and Examples." Creative Commons. Accessed June 16, 2019. https://creativecommons.org/share-your-work/licensing-examples/
7. "Journals@KU." University of Kansas Libraries. Accessed June 16, 2019. https://journals.ku.edu
8. "Home." Crossref. Accessed June 16, 2019. https://www.crossref.org/
9. "Home." DataCite. Accessed June 16, 2019. https://datacite.org/
10. "Home." ORCID. Accessed June 16, 2019. https://orcid.org/
11. "DOAJ (Directory of Open Access Journals.)" DOAJ. Accessed June 16, 2019. https://doaj.org/
12. "Home." Creative Commons. Accessed June 16, 2019. https://creativecommons.org/
13. "Inclusion Guidelines for Webmasters." Google Scholar. Accessed June 16, 2019. https://scholar.google.com/intl/en/scholar/inclusion.html
14. "Web Content Accessibility Guidelines (WCAG) 2.1." World Wide Web Consortium. Accessed June 16, 2019. https://www.w3.org/TR/WCAG21/
15. "Resources for Editors of Scholarly Journals: Getting Started." Digital Publishing Services, University of Kansas Libraries. Accessed June 16, 2019. https://guides.lib.ku.edu/journal_editors
16. "Resources for Editors of Scholarly Journals: Using Open Journal Systems (OJS), version 3.x." Digital Publishing Services, University of Kansas Libraries. Accessed June 16, 2019. https://guides.lib.ku.edu/journal_editors/learn_ojs3

Convergence Research in the Age of Big Data: Team Science, Institutional Strategies, and Beyond

Daniel Sui, Vice Chancellor for Research and Innovation
Jim Coleman, Provost and Executive Vice Chancellor for Academic Affairs
University of Arkansas

While there have been distinct subjects for human intellectual inquires for thousands of years, discipline-based and interdisciplinary research have been with us in the academy for only the last two centuries (Fredeman et al., 2010). Disciplines covering specifically defined subject matters emerged as universities expanded in size since the Industrial Revolution and especially as universities evolved increasingly to stress research alongside teaching (Klein, 1990). Disciplines that could hire their own faculty, design their own curriculum, grant their own Ph.D.s, publish specialized journals, and hold their own annual meetings have been the driving force for the spectacular growth of both the educational and research enterprise of higher education throughout the world, especially among American Universities (Jacobs, 2013; Woeler and Millar, 2013). In the natural sciences and engineering, the hardening of disciplines was aided by industrial demand for specialized researchers. Before that time, scholars were expected to be generalists.

Ever since the emergence and growth of strong disciplines, there have been calls for interdisciplinary collaboration in the academy for both education and research (Graff, 2015). Calls in the early twentieth century for interdisciplinarity often focused on teaching. This was a reaction to the creation of the disciplinary major. The general education (GenEd) movement in the United States and elsewhere also aimed to make university education more relevant to the needs of modern citizenship. In the 1960s again, interdisciplinarity was often advocated as a means to make university education more relevant. It was widely felt that disciplines were ill-equipped to prepare students to address pressing social problems (Abbott, 2001). By the start of the twenty first century, this interest in interdisciplinary education had been matched by a growing interest in interdisciplinary research as well (NAS, 2005a; Atkinson and Crowe, 2006; Szostak, 2013).

With the explosion of Big Data during the past ten years, discussions on interdisciplinary research has entered a new phase, with a strong emphasis on convergence research through a team science approach. The goal of this paper is to present a synoptic overview on how we may facilitate convergence research in the age of big data. The rest of this paper is organized as the following. After a brief introduction, there is an overview on the general concept of convergence research and how it is different from traditional multi-, inter-, and trans-disciplinary work. The following section outlines key elements of a team science approach for conducting data-driven convergence research following the emerging fourth paradigm. The next session discusses institutional strategies, opportunities, and challenges to promote convergence research, followed by a summary and conclusion in the last section.

Convergence Research: An Overview

Convergence is the new buzz word these days in science, business, public policy and beyond, as evidenced by over 5,000 books published recently with "convergence" as part of the book title according to Amazon.com. Inevitably, convergence means different things to different audiences despite Watson (2016) mounting convincing evidence that convergence is really at the heart of scientific progress throughout history. In the context of research, we draw primarily on the NSF's definition (https://www.nsf.gov/od/oia/convergence/index.jsp) and the National Academies' Report on Convergence (NAS, 2014).

Growing convergence research at the U.S. National Science Foundation (NSF) was identified in 2016 as one of 10 Big Ideas for Future NSF Investments. Convergence research is a means of solving vexing research problems - in particular, complex problems focusing on societal needs (Bainbridge and Roco, 2016). Convergence research typically entails integrating knowledge, methods, and expertise from different disciplines and forming novel frameworks to catalyze scientific discovery and innovation. Convergence research is closely related to other forms of research that span across different disciplines – multi-, inter-, and trans-disciplinarity (Figure 1), but also has its distinctive meaning. It is the closest to trans-disciplinary research which was historically viewed as the pinnacle of evolutionary integration across disciplines (Bergmann et al., 2012).

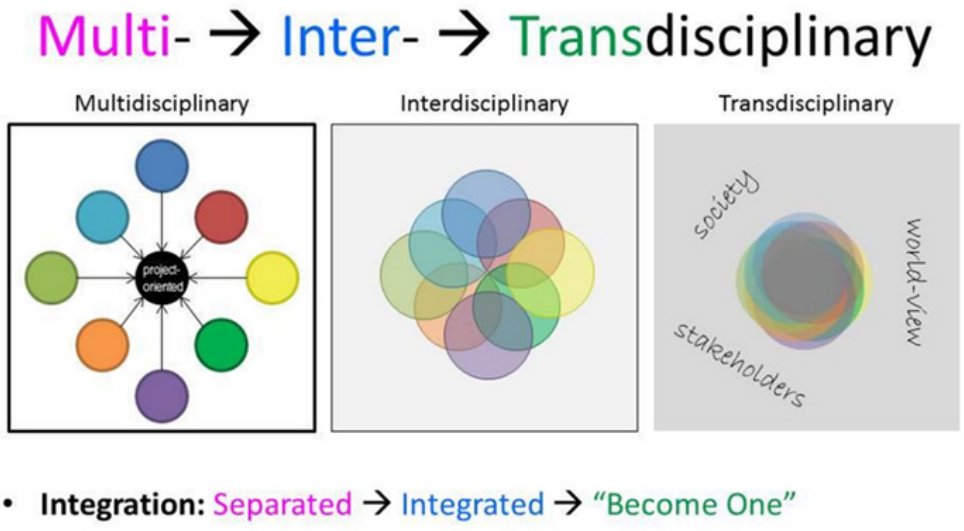


Figure 1. Multi-, Inter-, and Trans-disciplinarity
[Source: https://nanohub.org/groups/howpeoplelearnnano/crossdisciplinary_nature_of_nanotechnology, fair non-commercial use via Creative Commons agreement]

According to NSF and NAS, convergence research must have two primary characteristics (Figure 2):

- Transdisciplinarity (Figure 1) - Deep integration across disciplines. As experts from different disciplines pursue common research challenges, their knowledge, theories, methods, data, research communities and languages become increasingly intermingled or integrated. New

frameworks, paradigms or even disciplines can form sustained interactions across multiple communities.

- Stakeholder synergy: Research driven by a specific and compelling problem that draws together academic researchers, policy makers and industry partners. Convergence research is generally inspired by the need to address a specific challenge or opportunity, whether it arises from deep scientific questions or pressing societal needs.

op holistic and robust theoretical frameworks, problem-solving strategies, and innovative ways of collaboration in new exciting areas of research.

The continuing explosion of big data (both quantitative and qualitative) during the past ten years is transforming how we can conduct research in multiple fields. We strongly believe that data-driven approach will serve as a catalyst to stimulate future convergence research and the emerging team science will play increasingly important roles in convergence research due to its mandates on transdisciplinarity and stakeholder synergy.

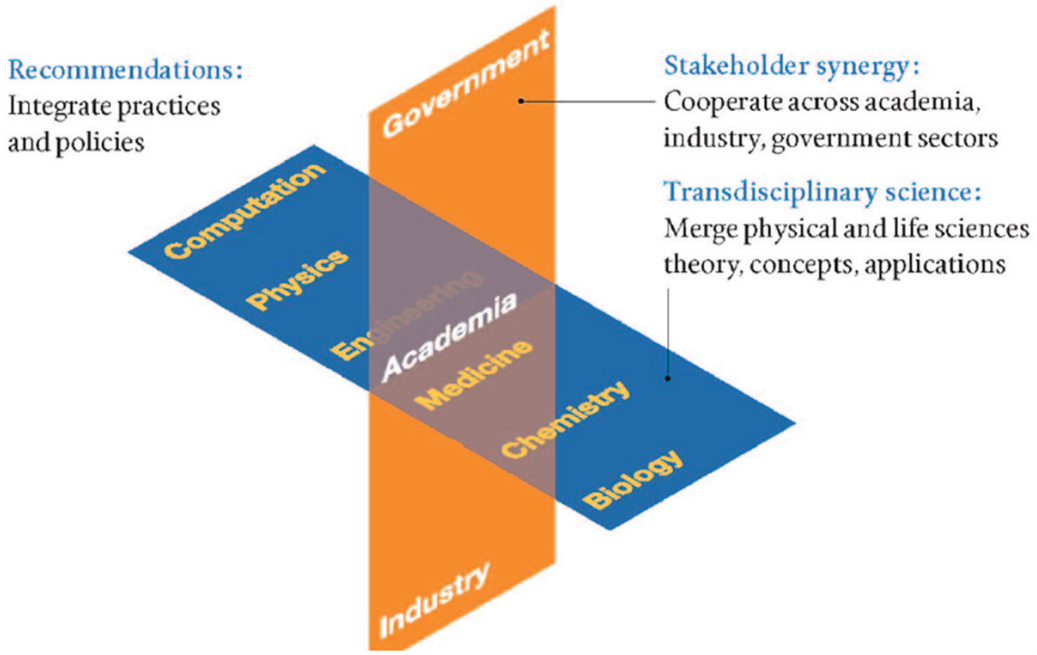


Figure 2. Two-dimensions of Convergence Research
[Source: NAS, 2014]

Since its inception, the convergence paradigm intentionally brings together intellectually diverse researchers to develop effective ways of communicating and synergizing across disciplines by adopting common frameworks and a new scientific language, which may, in turn, empower researchers to devel-

Convergence Research and Big Data: A Team Science Approach

As demonstrated by the convergence research projects recently funded by NSF (https://www.nsf.gov/od/oia/convergence/index.jsp), a variety of diverse approaches have been proposed and used to conduct convergence research, but two

approaches featured prominently in convergence research - data science and team science approaches.

Transdisciplinary research by the 4th paradigm.

According to Jim Gray at IBM (https://jimgray.azurewebsites.net), for over two thousand years, science has been conducted according to three paradigms - empirical science, theoretical science and computational science until big data exploded onto the scene. The availability of big data has transformed multiple fields, including physical sciences, medical/health sciences, engineering, social sciences, and even humanities. The emerging data-driven fourth paradigm to conduct basic research provides new opportunities to grow convergence research to a new level (Hey et al., 2010). Although data science needs new infrastructure, theoretical framework, and domain specific techniques, it is integral part and parcel of the fourth industrial revolution.

Through convergence research, the rapidly emerging field of data-intensive science (aka eScience) will continue to transform the world's scientific and computing research communities and inspire the next generation of scientists. Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets. The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud-computing technologies. The fourth paradigm of discovery based on data-intensive science offers insights into how the potential of convergence research can be fully realized.

Stakeholder synergy by team science.

Stakeholder synergy – the integration of academia, industry, and government – is the second defining characteristics for convergence research. To achieve stakeholder synergy, a team science approach is needed. In general, team science is a collaborative effort to address a scientific challenge that leverages the strengths and expertise of professionals trained in different fields (NAS, 2005b; Wuchty, 2007). Although traditional single investigator driven approaches are ideal for many scientific endeavors, coordinated teams of investigators with diverse skills and knowledge may be especially helpful for studies of complex social problems with multiple causes.

Over the past two decades, there has been an emerging emphasis on scientifically addressing multi-factorial problems, such as climate change, the rise of chronic disease, and the health impacts of social stratification. This has contributed to a surge of interest and investment in team science. Increasingly, scientists across many disciplines and settings are engaging in team-based research initiatives. These include small and large teams, uni- and multi-disciplinary groups, and efforts that engage multiple stakeholders such as scientists, community members, and policy makers (Fiore, 2008; Disis, 2010). Academic institutions, industry, national governments, and other funders are also investing in team science initiatives.

A growing trend within team science is cross-disciplinary science in which team members with training and expertise in different fields work together to combine or integrate their perspectives in a single research endeavor. Cross-disciplinary team science has been identified as a means to engage in expansive studies that address a broad array of complex

and interacting variables. It is seen as a promising approach to accelerate scientific innovation and the translation of scientific findings into effective policies and practices.

In addition, the science of team science (SciTS) is a rapidly emerging field focused on understanding and enhancing the processes and outcomes of team science (Stokols et al., 2008). A key goal of SciTS is to learn more about factors that maximize the efficiency, productivity, and effectiveness of team science initiatives. A diverse group of scholars contributes to SciTS (Falk-Krzesinski et al., in press). They bring conceptual, historical, and methodological approaches from a wide variety of disciplines and fields, including public health, management, communications, and psychology. They have created measures to assess team science processes and outcomes, and to influence contextual and environmental conditions (Table 1). Applying these measures can help researchers evaluate team science, improve the quality of ongoing initiatives, and develop best practices.

Table 1. Major areas of inquiry in SciTS [Source: Stokols, 2008]

- Methods and models for the study of team science
- The structure and organization of team science, particularly the collaborative processes moderated by a variety of contextual environmental factors
- Team characteristics and dynamics, such as the elements of effective leadership, ideal team composition, and communication styles
- Design and outcomes of training programs to support team science
- Translation of team science findings to practice and policy
- Scientific and societal outcomes of team science such as scientific discoveries and innovations, knowledge dissemination, and long-term public health impacts

Among the multiple insights gained from the research in SciTS, we now know that interpersonal dynamics among team members are the key for the success of a team science project. Team members' collaborative skills and experiences can be very useful to guide our future efforts of data science-driven team science convergence research. In addition, the success of team science is influenced by a variety of contextual environmental influences (Börner et al., 2010). These factors influence each stage of a scientific initiative, with implications for efficiency, productivity, and overall effectiveness. For example, funding trends from government, industry, and private foundations can exert a huge influence on how research is being conducted. The recent emphasis by both public and private funding agencies on convergence and interdisciplinary collaborative projects addressing society's grand challenges will surely further stimulate and promote team science approaches in science (NAS, 2018). Institutional infrastructure and resources for communication and data sharing are also very important. Team processes, including the existence of agreements related to proprietary rights to data and discovery (King and Persily, 2019), as well as mechanisms for feedback and reflection, can also shape the outcome of team efforts. Last but not least, organizational policies, such as those relating to promotion and tenure, can also significantly incentivize or discourage team-based endeavors.

Institutional Strategies for Promoting Convergence Research: Opportunities and Challenges

We have seen many strategies put in place with the purpose of facilitating and stimulating convergence research and

team science in our previous and current roles at several institutions. This has provided us with some empirical evidence of how various strategies have fared. But empirical evidence has been difficult to interpret for several reasons. Sometimes it is just simply unclear whether there have been positive results. At other times, increases in team-based research and convergence research have improved, but there is no control group that would enable one to determine whether the strategy that was implemented actually caused the improvement. For example, we have been at institutions that have created research space designed to encourage team-based, interdisciplinary, and/or convergence research. At those institutions, the facilities were populated with some of the institution's most productive and collaborative scientists. The research programs in those facilities showed great success in productivity and collaborative research. Yet, it is impossible to answer the question whether those highly productive scientists became more collaborative, or were more productive, than they would have been had they stayed in their original facilities.

Nonetheless, we believe that there are multiple golden opportunities to conduct team-based convergence research using big data right now, but there exist certain challenges and barriers we need to overcome. We'd like to share some of those challenges, our experience with a few things that we have tried, and the University of Arkansas's institutional strategies to promote convergence research via a team-based data science approach.

Opportunities.

The new digital economy and new business models.

The fourth Industrial Revolution (Schwab, 2015) is unfolding rapidly in front of us, driven by new innovations and advances in AI, block chain, cloud computing, and data analytics. The economy will grow increasingly digital and be built upon digital platforms. Convergence research will be needed to address the pressing issues of this economy head-on. At the University of Arkansas, we have been implementing several efforts to marshal our resources in this direction including the creation of a center of excellence in block chain research, a cross-campus data science degree, and identifying convergence research in data science as an institutional investment.

New funding opportunities from government, industry, and private foundations.

The new digital economy has created new demands for data-driven convergence research. The U.S. federal agencies have all developed a new data strategy (https://strategy.data.gov). NSF has been leading the funding opportunities through its Harnessing the Data Revolution initiative (https://www.nsf.gov/cise/harnessingdata). NIH has developed a similar data science strategic plan (https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for__Data_Science_Final_508.pdf) with focus on infrastructure, analytic tools, data ecosystem, stewardship, and workforce development.

With the growth of new data-driven businesses and the use of data to enhance the traditional industry/business, every company/business is in the process of developing a new data strategy (https://www.forbes.com/sites/bernardmarr/2019/03/13/why-every-company-needs-a-data-strategy-for-2019/#3ff319e64cbb). Once again, these create new opportunities for da-ta-driven convergence research. Private foundations and non-profit organizations have also increased their investments in research related to data science and data analytics (https://www.rockefellerfoun-

dation.org/blog/introducing-data-science-social-impact).

Potentially major scientific breakthroughs that won't be accomplished otherwise.

We can learn a lot from the natural world, or at least find strong metaphors in the processes of organismal evolution. Ecological systems are often characterized by having edge effects. Edge effects are changes in population or community structures that occur at the boundary of two or more habitats.[1] There is often a relative explosion of biodiversity that occurs in these areas where two or more separate eco-systems overlap. Many major advances in science occur in the area where disciplines overlap. For example, these "edge effects" in science have led to new fields such as molecular biology and biomedical engineering. We believe that major advances in the integration of big data into convergence research will also occur on the "edges" where disciplines, approaches, vocabulary, and more overlap. By facilitating interactions at the edges, institutions should be able to also facilitate advances in convergence research.

New opportunities to integrate arts, humanities, and social sciences with STEM fields.

The promotion of convergence research has further raised broader awareness that all human knowledge are branches of the same tree (NAS, 2018). We strongly believe the on-going trend towards convergence research also provides a golden opportunity to further integrate the arts, humanities, and social sciences with the traditional STEM fields, especially for those scholars practicing in the emerging digital humanities. Data-driven team science approach could potentially be added to humanity scholars' methodological repertoire (Dobson, 2019). We are fully aware of the on-going debates about the future of digital humanities (Gold and Klein, 2019) and a major cultural change needed for the practice of humanity scholarship. One thing we are absolutely convinced is that with the increasing automation and adoption of robotics in everything we do, we need to find creative ways (more than ever) to integrate the arts, humanities, and social science with STEM fields (Levit, 2018). To us, this integration will represent convergence research at its highest/deepest levels.

Challenges.

There are, of course, many challenges inherent in our institutions to fully engaging in convergence research. We list some of those below.

Academic culture.

Graduate students are generally trained to have a primary allegiance to a discipline, which carries through to the faculty years. The importance of belonging to a discipline is enforced or promulgated in several ways in universities. For example, academic departments are generally built around disciplines, partly because they may need to be for curricular reasons. Departments are not just an administrative unit, though, but become like a family unit, where members of the department work to garner resources for their unit and to foster the success of the department and hence the discipline. Furthermore, faculty often receive their most important rewards from their discipline (e.g., awards from professional societies, grants from discipline specific panels, and respect that comes from being recognized in a discipline). Also, in general, the academic culture has focused recognition of research on individuals generally for their independent research contributions (e.g., membership in the National Academies, fellows of major societies, and research recognition awards on university campuses), not to teams.

Thus, the incentives driving the participation in convergence research may often need to be self-created through personal excitement about a specific question or collaboration, or because of potential access to research funding.

Also, disciplines can create or enforce their own culture, like many human societies, by creating their own vocabulary and vernacular for communicating their work. The vocabulary can become an intentional or unintentional barrier for members of other disciplines to integrate or collaborate.

Furthermore, disciplines have evolved their own way of collecting and sharing data. For example, Hampton et al (2013) examined how big data may impact the future of the field of ecology. Their meta-analysis indicated that, as a group, ecologists tend to design their own experiments to answer specific questions, and to a large degree do not have a culture of sharing or reusing data. Where data is shared in open databases, the vast majority was shared in genetic databases (e.g., GenBank)- genetics is a science that has been driven in many ways by a culture of data sharing.

Costs of maintaining computational infrastructure and data storage.

The amount of data is increasing at an extraordinary rate. In some fields, we now collect more data in a single year than had been collected in all of human history. This is putting tremendous pressure on the capacity of universities to manage data as well as maintaining the computational capacity to analyze to support research computation. Many universities built their research computing infrastructure with heavy support from external, often federal, grants. The availability of these funds has not kept pace with the dramatically changing needs for storage and computation. Furthermore, it has been difficult for many universities to develop a sustainable business model to support the increasing capacity needed for data storage and computation. Part of the challenge with storage is deciding what data needs to be kept. Physical libraries, at least that we know of, do not save and catalogue every document ever written, at least partly because there is no business or operational model that would allow that approach to be functional. We think that decisions will need to be made regarding benefit and cost to ensure optimal storage and use of data, and the criteria that can be applied with respect to which data should be saved.

Data science vs. statistical support.

We have both been involved in implementing initiatives to develop data science programs. A challenge that we observed is the definition of "data scientists" varies among various fields or individuals. For example, one of us was involved in facilitating the development of an institutional partnership that involved biomedical disciplines from one campus and computational disciplines from another. The biomedical researchers demonstrated great enthusiasm for the bioinformatics expertise of the computational scientists. But, when digging a little deeper, some of the biomedical scientists were excited because they viewed the strength in bioinformacticists as a way to acquire help analyzing their data - essentially looking for statistical support of their research. This was a mismatch for the data scientists who focused their research on the development of new tools and/or the fundamental science of data analysis. This created tension in the partnership. Although this occurred several years ago, the term "data scientist" still means different things to different people, and this can inhibit the formation of teams participating in convergence research.

Correlation/predictive analytics vs cause/mechanism.

Both of us in our current roles as provost (Coleman) and vice chancellor for research (Sui) are using big data and predictive analytics to ask questions and to help drive resource allocation. These tools are powerful. But we also worry that the rapidly growing ability to correlate and tease apart how different variables are related to each other can lead to inappropriate conclusions regarding causality. It can be easy to mistake correlation for causality.

One humorous example of the potential to confuse correlation and causality is the predictive value of ice cream consumption for weight gain over an annual period. A big data analysis would show that ice cream consumption peaks during summer months in the US and declines in winter months. Alternatively, weight gain follows an almost exact opposite pattern. Weight gain in US population reaches a peak in winter months and reaches its low point in summer. Thus, it turns out that ice cream consumption is in fact a great correlative (inverse) predictor of weight gain over the course of a year - the higher the ice cream consumption the lower the weight gain. This is a wonderful conclusion for those of us who love ice cream. But, unfortunately, although ice cream consumption over the course of a year is a great inverse predictor of weight gain over a year, it has no causal relationship. Both curves are driven by seasonal temperatures and culture.

As it becomes easier and easier to construct predictive analysis relating variables, it will also become more and more likely that causation and correlation can become confused. This can easily lead to bad decisions on policy or resource allocation. Furthermore, transdisciplinary research and team science always increase the difficulty and complexity of reproducibility and replicability. Also, managing large teams in research projects entails new human dynamics, and not all teams succeed admirably. In fact, some teams end in catastrophic failures.

Some strategies we have employed.

Both of us have been involved in implementing several strategies to facilitate interdisciplinary, transdisciplinary, and convergence research in our various roles. The results of these various strategies have been either mixed or hard to decipher. None of these strategies were outright failures. In most every case, the strategies facilitated positive outcomes. The challenge is determining whether the strategy really facilitated changing the culture of research, or whether the strategies produced positive outcomes enhancing research infrastructure or supporting the most highly productive, energetic, and/or ambitious faculty. And, it is also hard to determine whether the resources allocated to these initiatives created the highest ROI with respect to increasing team-based research. We list some strategies that we were involved in implementing along with links to websites describing some of them for those interested in further information:

a. Designing research facilities to facilitate convergence research. Examples that we were involved with: Bond Life Sciences Center, University of Missouri, https://bondlsc.missouri.edu/ and Bioscience Research Collaborative, Rice University, https://brc.rice.edu. By any measure, the researchers housed in these facilities have been successful, and that many of the researchers work in collaborative teams. These state-of-the-art of facilities provided great

environments for productive researchers.

b. Seed grant programs to support interdisciplinary or convergence research: A Chancellor's Research and Innovation Fund, funded partially through athletics revenue, was created at the University of Arkansas, https://chancellorsfund. uark.edu/innovation-and-collaboration. This program has seeded several collaborations in its first three years, and some seed grants have led to successful large collaborative research proposals. We have focused the resources on new collaborations. We should point out, though, the program has not yet run long enough to determine the return on its investment.

c. Creating interdisciplinary structures that don't compete with disciplinary homes. The Bond Life Science Center was designed with a specific model around faculty lines, space, distribution of funds equivalent to indirect cost recovery, and salary savings obtained through support of salary by research grants to minimize competition between departments across campus and the center. This model at least worked at the start of the center in facilitating departments across campus moving faculty into the facility. At the Desert Research Institute (www.dri.edu), five disciplinary units (Biological Sciences, Earth Science, Energy, Atmospheric Science, and Hydrological Sciences) were combined into three larger units (Earth and Ecosystem, Hydrological Sciences, and Atmospheric Science), and the administrative savings were used to create two interdisciplinary centers, selected through a faculty review process, focused on bringing teams from across disciplines in institute together to solve larger issues. The original two centers were focused on alpine watersheds and arid lands environmental restoration. These have morphed into different areas of strength and need.

d. Interdisciplinary graduate and post-graduate programs; The University of Arkansas has six interdisciplinary graduate programs that cross department and college lines (https://graduate-and-international.uark.edu/more-information/our-staff/interdisciplinary.php) reporting to the Graduate School and International Studies that have helped to support interdisciplinary work. At Virginia Commonwealth, several interdisciplinary Ph.D. programs were created - one that has become particularly successful and distinctive is Media, Art and Text (https://matx.vcu.edu/).

We are aware that in recent years, cluster hiring has been a common practice among multiple institutions to promote interdisciplinary collaboration. Although there are positive reports on this new practice, cluster hiring has its own problems (https://www.aplu.org/ members/commissions/urban-serving-universities/student-success/cluster.html). In addition, we believe that the global movement towards an open science paradigm has and will continue to promote convergence research and interdisciplinary collaboration despite renewed emphasis on IP protection in the U.S. and some other countries.

Summary and Conclusion

Multiple grand challenges, ranging from dealing with global climate change and addressing the widening income and health disparity, to fighting terrorism and combating misinformation and fake news, can't be resolved by any individual discipline. More than ever, we need interdisciplinary collaboration and teamwork. Following the previous three paradigms in empirical, theoretical, and computational approaches to science, the growth of big data and data science are emerging as the fourth paradigm in the form of eScience that could potentially further facilitate convergence research, which in most cases also call for a team science approach. New insight from studies of the science of team science provide further guidance related to the composition, size, and leadership of teams. Indeed, there are no better times in the history of higher education than now to conduct convergence research through big data and team science to address grand societal challenges that transcend traditional disciplinary boundaries.

However, we do want to conclude this paper with one caveat - by emphasizing the need of convergence research and interdisciplinary collaboration, we are NOT abandoning/marginalizing the traditional discipline-based research nor individual-based inquiries. In fact, we need more cutting-edge discipline-based work in order to be more productive and effective in our convergence efforts (Jacobs, 2013). Likewise, by arguing for the need of a team science approach and collaboration, we do not want to marginalize individual-based endeavors. To contrary, we believe all research must necessarily be conducted individually at some point, even for projects involving large teams. So, it is not either/or; moving forward, we need both discipline-based, individual research and convergent, team-based transdisciplinary endeavors. It has been (and continues to be) through the dialectal process of convergent/divergent, disciplinary/interdisciplinary, individual/team-based approaches that our research enterprise has been propelled to a new level for excellence to make our world a better place for all.

Acknowledgement:

The authors gratefully acknowledge the assistance John Post and Angela Bolinger provided in the preparation of this paper. DS also acknowledges the fruitful discussions he had with his former NSF colleagues on topics related to harnessing the data revolution, growing convergence research, and team science, which planted the seeds for this paper.

References

Abbott, A. 2001. *Chaos of Disciplines*. Chicago, IL.: University of Chicago Press.

Atkinson, J. and Crowe, M. (eds.), 2006. *Interdisciplinary Research: Diverse Approaches in Science, Technology, Health and Society*. West Sussex, UK.: John Wiley & Sons.

Bainbridge, W.S. and M. C. Roco (Eds.), 2016. *Handbook of Science and Technology Convergence* 1st ed. Berlin, Springer.

Bergmann, M., T. Jahn, T. Knobloch, W. Krohn, C. Pohl, and E. Schramm, 2012. *Methods for Transdisciplinary Research: A Primer for Practice.* Chicago, IL.: University of Chicago Press.

Börner, K., Contractor, N., Falk-Krzesinski, H.J., Fiore, S.M., Hall, K.L., Keyton, J., Spring, B., Stokols, D., Trochim, W., and Uzzi, B., 2010. A Multi-Level Systems Perspective for the Science of Team Science. *Science Translational Medicine* 2, 49cm24.

Disis, M., and Slattery, J., 2010. The Road We Must Take: Multidisciplinary Team Science. Science *Translational Medicine* 2, 22cm29.

Dobson, J.E., 2019. *Critical Digital Humanities: The search for a methodology*. Urbana, IL.: University of Illinois Press.

Falk-Krzesinski, H. J., Contractor, N., Fiore, S. M., Hall, K. L., Kane, C., Keyton, J., Klein, J.T., Spring, B., Stokols, D. & Trochim, W. (in press). Mapping a Research Agenda for the Science of Team Science. *Research Evaluation*.

Fiore, S.M., 2008. Interdisciplinarity as teamwork—How the science of teams can inform team science. *Small Group Research* 39, 251–277.

Frodeman, R., J. T. Klein, and C. Mitcham (eds.), 2010. *The Oxford Handbook of Interdisciplinarity*. New York, N.Y.: Oxford University Press.

Gold, M.K. and L. F. Klein (eds.), 2019. *Debates in the Digital Humanities*. Minneapolis, MN.: Univ Of Minnesota Press.

Graff, H. J. 2015. *Undisciplining Knowledge*. Baltimore, MD.: Johns Hopkins University Press.

Hampton, S.E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter, 2013. Big data and the future of ecology. *Frontiers in Ecology and Environment* 11(3): 156-162.

Hey, T., Tansley, S., & Tolle, K. (Eds.), 2010. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, available on-line at: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery.

Jacobs, J. A., 2013. *In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University*. Chicago, IL.: University of Chicago Press.

King, G. and N. Persily, 2019. A new model for industry-academic partnerships. PS: Political Science and Politics. Publisher's version available on-line at http://j.mp/2q1IQpH (last accessed on July 7, 2019).

Klein, J. T., 1990. *Interdisciplinarity: History, Theory, and Practice*. Detroit, MI.: Wayne State University Press.

Levit, A., 2018. *Humanity Works: Merging technologies and people for the workforce of the future.* New York, NY.: Kogan Page Inspire.

National Academies of Sciences (NAS), 2005a. *Facilitating Interdisciplinary Research*. Washington D.C., National Academies Press.

National Academies of Sciences (NAS), 2005b. *Enhancing the Effectiveness of Team Science*. Washington D.C., National Academies Press.

National Academies of Sciences (NAS), 2014. *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond*. Washington D.C., National Academies Press.

National Academies of Sciences (NAS), 2018. *The Integration of the Humanities and Arts with Sciences, Engineering, and Medicine in Higher Education: Branches from the Same Tree.* Washington, DC: The National Academies Press.

Schwab, K., 2015. *The Fourth Industrial Revolution*. https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab

Stokols, D., Hall, K.L., Taylor, B.K., and Moser, R.P., 2008. The Science of Team Science: Overview of the Field and Introduction to the Supplement. *American Journal of Preventive Medicine* 35, S77–S89.

Szostak, R., 2013. The state of the field: Interdisciplinary research. *Issues in Interdisciplinary Studies* 31: 44-65

Watson, P., 2016. *Convergence: The idea at the heart of science*. New York, NY.: Simon & Schuster.

Woelert, P. and V. Millar, 2013. The 'Paradox of Interdisciplinarity' in Australian Research Governance. *Higher Education* 66 (6): 755–767.

Wuchty, S., Jones, B.F., and Uzzi, B., 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science* 316, 1036–1038.

Making Mountains out of Molehills: Challenges for Implementation of Cross-Disciplinary Research in the Big Data Era

Daniel Andresen and Eugene Vasserman
Department of Computer Science, Kansas State University
{dan, eyv} @ksu.edu

We present a "Researcher's Hierarchy of Needs" (loosely based on Maslow's Hierarchy of Needs) in the context of interdisciplinary research in a "big data" era. We discuss multiple tensions and difficulties that researchers face in today's environment, some current efforts and suggested policy changes to address these shortcomings and present our vision of a future interdisciplinary ecosystem.

Big data, as noted by Dr. Francine Berman of the San Diego Supercomputer Center, is crucial to maintaining competitiveness in today's research environment. She notes, "More scientists will depend on exabyte data than on exaflop machines." Big data is also a new strategic advantage, and the new shared environments for scientists and researchers to explore. Virtually all of the NSF's *10 Big Ideas for 2019* relate to interdisciplinary research on big data.

Several overarching issues come to mind when considering interdisciplinary research centered around big data. For example, interdisciplinary communication is challenging, as similar terms may mean different things, and each party is to develop a sufficiently sophisticated vocabulary to be able to communicate across the research areas. Also, finding the right people with the right skills is particularly challenging. For example, each of the authors is in the department of computer science, yet even with our own discipline we had difficulty finding collaborators with the skills necessary to deal with large quantities of data in an efficient, effective manner. In addition, that funding agency has emphasized the importance of interdisciplinary collaboration, even going so far as insisting on it in any calls for proposals, yet adding significant barriers in the form of legislation such as CUI, ITAR, and other regulations. Cybersecurity and privacy also come heavily into play, as large datasets become increasingly likely to contain data which is plausibly personally identifiable, leading to dangers both in researchers learning information which is supposed to be hidden, and hackers causing potentially embarrassing and financially ruinous data leaks. Finally, comprehensive institutional support is frequently lacking, where the infrastructure is inadequate to support the desired scale and types of research, and the rewards systems for researchers fails to incentivize interdisciplinary research.

In examining interdisciplinary research, particularly in the academic environment, one must take into account the whole environment, and not simply focus on skills, Cyberinfrastructure, or other more easily tackled subproblems.

In this paper we first present a model for interdisciplinary research in a big data environment. We then discuss two overarching issues, interdisciplinary communication (both at the data and interpersonal levels), and the effects of the need for cybersecurity and privacy. We finish by offering several suggestions and observations for changes at the institutional level.

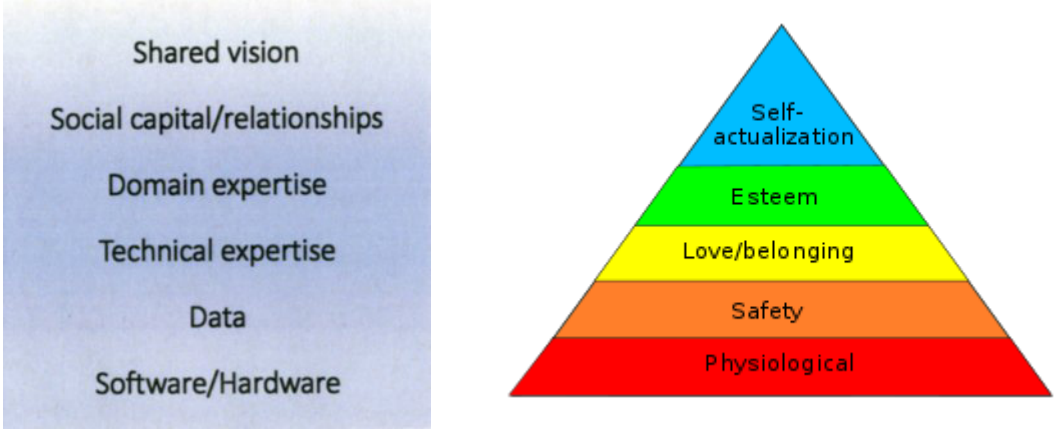


Figure 1. A Research Hierarchy of Needs (left), and Maslow's Hierarchy of Needs (right).

A Research Hierarchy of Needs

That realization drives us to introduce a Research Hierarchy of Needs loosely modeled on Maslow's Hierarchy of Needs[1]. And Maslow's hierarchy of needs, needs which are underserved at the lower levels of the hierarchy prevent the full expression of needs at higher levels in the hierarchy. So for instance if a person is in danger, the need for self esteem is substantially diminished. Similarly, in research, if the two researchers have a strong shared vision but lack the data on which to base their research, they will be unsuccessful.

1. *Shared Vision* The end goal for a relationship among researchers is that moment when they agree on the shared vision containing the research goal, outline the general plan, and realize they have the resources to accomplish it.
2. *Social capital/relationships* The shared vision is built on the ability to work together, to communicate well, and develop sufficient trust in the other to warrant risking valuable resources (e.g., time, effort, and expertise). Such relationships typically evolve over time, motivated by mutual respect, shared interests, and shared goals – frequently developing from loose, unofficial ties. Conversely, they can develop quickly if sufficient incentives (positive or negative) are introduced.
3. *Domain expertise* Frequently, particularly for larger projects, much of the actual work will be accomplished by a pool of experts – typically post-doctoral associates or graduate students – who are well-versed in the theory and experience in the associated research areas. However, in a big-data environment, we have noticed a serious shortage of qualified personnel with sufficient experience and knowledge of big data tools, leaving significant data analysis either unaccomplished or unattempted.
4. *Technical expertise* The domain experts cannot accomplish their goals, however, without a rich cyberinfrastructure ecosystem.[2] It is impossible to accomplish rich, large-scale data analysis on a typical academic IT infrastructure consisting of personal computers, Wi-Fi, and an Internet connection. Specialized hardware and

software environments that can easily handle multi-terabyte- to petabyte-level data analysis (both storage and computation) are required – and these require support staff (frequently referred to as "HPC Facilitators"[3]) who are experts at helping researchers effectively use these tools at scale. These facilitators are frequently domain scientists with doctoral degrees in related fields who have accumulated sufficient training and expertise to be effective in their hybrid role.

5. *Data* Clearly the basis for any research with the foundation in big data is the data itself. There are multiple issues that arise typically when acquiring the data needed to accomplish the research in question. The first issue is, of course, getting permission to acquire the data. As noted below, security, privacy, and tradition can all conspire to make getting permission difficult. Also, as the data is acquired, frequently sufficient metadata is not collected to make the easily searchable and curatable. Finally, after the data is collected, metadata attached (if this vital step is not effectively ignored – many researchers consider naming a directory 'lab3_day4_ir_rat_exp1' to be sufficient labeling), the cost of storing the data and making it available can prove prohibitive.
6. *Software/hardware* In the end, all data and analysis must occur in a sufficiently-scaled software and hardware environment. With local HPC compute resources available at every Top-100 research institution[4], and the benefits in productivity and prestige accrued[5], most institutions have cyberinfrastructure in place for previous-generation science. However, given the massive increase in scale required by today's research, institutions are scrambling to find the right mix of local and cloud resources, while also seeking effective funding models to support this vital expansion. Cyberinfrastructure is the new laboratory environment, and resources allocations (such as F&A) need to reflect this on par with more traditional efforts.

Interdisciplinary Communication

My colleague at Kansas State University, Dr. Doina Caragea, noted when asked about the toughest part of interdisciplinary research, "**I can't understand your problems, and you can't understand my possibilities**." Building the vocabulary and depth of understanding for sound interdisciplinary research is generally extraordinarily challenging, requiring significant investments in time and energy. However, university environments tend to be heavily siloed, typically by department, and as Kleinbaum notes, "…pairs of individuals that are in the same business unit, subfunction, and office location communicate at an estimated rate that is 1,000 times higher."[6] Overcoming a 1000x communication disadvantage is challenging, and workers even 30 meters apart have a perceived 1KM of distance between them.[7] Overcoming the siloes to build the social capital is needed to overcome the remorseless logic that in general, the short-term, annual-report-driven return on incremental efforts invested is likely to be better as extensions on existing domain successes rather than risky large-scale new projects. Building an understanding of each disciplines' vocabulary, workflows, rewards/

priorities, and arranging these into a mutually-beneficial project structure flows significantly more easily when the principals have an established foundation of trust and respect to build upon.

Cybersecurity and Privacy

Data sharing, and even data coexistence is challenging, especially as the data volumes increase and the amount of trust and collaboration between research groups decreases. The data may contain all kinds of sensitive information, from personally identifiable information, to financial records, to intellectual property, sensitive but unclassified, export controlled, etc. Careful design at the domain expertise level, and special security controls at the software level are required. An extreme case is that of complete separation (no group is allowed to interact with another) or even allow data analysis processes to coexist on the same physical hardware, which makes data sharing impossible but provides an excellent level of information leakage protection. The level of special controls and the associated difficulty in navigating them must be agreed upon at the level of shared vision (in terms of benefits and trade-offs).

In general, the more data sharing is allowed, the higher the risk of unforeseen information leakage. Pre-processing data before it is shared is an effective way to preserve privacy, but it is highly work-intensive, and is sometimes difficult to reuse once prepared: different methods of pre-processing are required depending on the types of analyses that collaborators would like to run on the data set. An example of pre-processing is an algorithm that adds noise to the original data while preserving desired statistical properties. Data redaction as pre-processing can also be effective, but requires significant domain knowledge to perform correctly, as the privacy properties of the redacted data set are heavily dependent on the type of data being shared. For instance, simple de-identification (removing names and identifiers) is usually not enough, and additional work is required to provide better anonymization[8] as demonstrated in practice by assigning names to user IDs in the Netflix challenge data set, which only contained film star ratings and randomly assigned user identifiers.[9] This was made possible by comparing the Netflix dataset (public but deidentified) to the IMDb data set (public but **not** deidentified) and inferring the Netflix user identities through similarities in watched films and star rankings.

Shared databases are fundamentally vulnerable to data extraction through the combination of multiple queries.[10] Some modern alternatives allow non-experts to query a database while the framework enforces privacy constraints.[11] This appears to be one of the more usable alternatives as of the time of writing, and allows the database to be fully shared, as long as it can only be queried using the PINQ platform.[4]

Institutional Support

We find the current research environments lacking in providing the type of comprehensive, universal support needed for today's interdisciplinary research. Steve Blank comments, "[I]nnovation is not a point activity, it's an end-to-end process. You need a pipeline."[12] We also see a strong competitive advantage for those institutions who can best enable their researchers (both professionals and students) in discovery, funding, and reputation. As such, we have several specific recommendations:

1. *Continue deliberately building opportunities for bridges across disciplines.* Funding – e.g., new NIH/NSF interdisciplinary CFPs – can certainly be a strong motivator, but build-

ing other opportunities for ties to organically form can pay off. One company deliberately reduced the number of coffee machines by a factor of 20 to force interaction among different groups – and sales increased by 20%![13] While this approach may not work for faculty (we suspect there would be a sudden explosion in in-office coffeemakers), variations like shared coffee machines between departments or university-funded coffeeshop accounts could help relationships develop.

2. *Make big data competence universal.* Given the need for competence (or at least familiarity) with big data tools in virtually every research area and job occupation today, from astronomy to zoology, we suggest that every student and postdoc be trained early in their tenure at the university. For example, at Kansas State University, we have made use of remote workshops offered through XSEDE which last two days, require no prior programming experience, and introduce participants to tools like Hadoop/Spark and TensorFlow.[14] These have been popular across multiple colleges and departments, and we have suggested that they form a basis for requiring every student in the College of Engineering at least be exposed to these tools. Longer-term workshops (typically one week) through Data Carpentry can offer another quality option.[15]
3. *Build the community of experienced interdisciplinary researchers.* We recommend that institutions incentivize interdisciplinary research. At present, there represents an implicit penalty for an interdisciplinary grant – e.g., given that most departments are assessed based on research expenditures, a grant with a colleague in the same department is better (at least in the short term) than a grant with a colleague outside the department. Similarly, adding young researchers to a grant proposal may strategically grow the number of experienced interdisciplinary researchers, in the short term it may decrease the odds of an individual proposal's success. Institutions may find it advantageous to weight interdisciplinary achievements (papers, grants, and other artifacts) more heavily in evaluations for tenure or promotion; they may also want to set an expectation that larger interdisciplinary efforts will have a certain percentage (say, 10%) of faculty who are relatively new to these environments.

Conclusion

In light of our "Research Hierarchy of Needs", we suggest that both researchers and their institutions recognize both the difficulty and rewards in interdisciplinary research, and the need to adapt in the modern research era to the size, speed, and scale of data and its contribution to science. We need to convert the "molehills" to "mountains" of resources at every level of the hierarchy: hardware/storage/cyberinfrastructure should be well resourced with dedicated staff; human capital with training in big data should be ubiquitous across disciplines; and there should be an institutional commitment to replacing a culture of scarcity with planned, systemic infrastructure on par with traditional science environments like laboratories and other physical resources.

References

1 Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), *370-396*.

2 Lowndes, Julia S. Stewart, et al. "Our path to better science in less time using open data science tools." *Nature ecology & evolution* 1.6 (2017): 0160.

3 Neeman, H.. "A Case Study for HPC Workforce Development and Workforce Meta-Development." (2015).

4 Neeman, H., personal communication, June, 2018.

5 Apon, Amy W., et al. "Assessing the effect of high performance computing capabilities on academic research output." *Empirical Economics* 48.1 (2015): 283-312.

6 Kleinbaum, Adam M., Toby Stuart, and Michael Tushman. *Communication (and coordination?) in a modern, complex organization*. Boston, MA: Harvard Business School, 2008.

7 Kraut, Robert E., Carmen Egido, and Jolene Galegher. "Patterns of contact and communication in scientific research collaborations." *Intellectual teamwork*. Psychology Press, 2014. 163-186.

8 Clete A. Kushida, Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. (2012) "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies." *Medical care* vol. 50 Suppl: S82-101.

9 Arvind Narayanan and Vitaly Shmatikov. (2008) "Robust De-anonymization of Large Sparse Datasets." In *IEEE Symposium on Security and Privacy (S&P)*.

10 Dorothy Elizabeth Robling Denning. (1982) "Cryptography and Data Security." Addison-Wesley.

11 Frank D. McSherry. (2009) "Privacy integrated queries: An extensible platform for privacy-preserving data analysis." In ACM SIGMOD International Conference on Management of data (SIGMOD).

12 Blank, Steve, and Pete Newell. "What your innovation process should look like." *Harvard Business Review* (2017).

13 Waber, Ben, Jennifer Magnolfi, and Greg Lindsay. "Workspaces that move people." *Harvard business review* 92.10 (2014): 68-77.

14 Maiden, T., "XSEDE HPC Workshop: BIG DATA," https://psc.edu/hpc-workshop-series/big-data , accessed 9/9/19.

15 Teal, Tracy K., et al. "Data carpentry: workshops to increase data literacy for researchers." *International Journal of Digital Curation* 10.1 (2015): 135-143. https://datacarpentry.org/

Training for Cross-Disciplinary Research and Science as a Team Sport

Jennifer L. Clarke, PhD, Professor, Food Science and Technology, Statistics
Bob Wilhelm, PhD, Vice Chancellor for Research and Economic Development
University of Nebraska-Lincoln

This article is dedicated to the memory of David R. Swanson, Ph.D. (8/13/65 - 8/12/19), former Director of the Holland Computing Center at the University of Nebraska. David was a wonderful person and a great colleague. He will be missed.

As members of a land-grant highly research active university, we recognize the growing importance of data and computing across all disciplines. We are also aware that addressing most, if not all, societal problems will require knowledge from multiple disciplines. This brings to mind the recent work by Peter Watson in which he makes a compelling argument that many diverse scientific branches are converging on the same truths [1]. Hence training faculty members, postdoctoral scholars, and students to excel in cross-disciplinary environments and leverage advances in data and computing to further research goals is critical to future institutional success. Only by working together and leveraging intellectual resources can we make significant discoveries for the benefit of humankind.

There are multiple, high profile examples in science of large successful collaborative projects. These include The Cancer Genome Atlas [2], the Laser Interferometer Gravitational-Wave Observatory (*LIGO*) Scientific Collaboration [3], ELIXIR (an intergovernmental organization that brings together life science resources from across Europe)[4], and the French Conseil Européen pour la Recherche Nucléaire (CERN) [5]. All of these projects leverage human resources from multiple disciplines as well as the latest in data and computing resources. A primary reason for the ongoing societal impact of these projects is their willingness to share resources with the broader community.

This way of thinking - **scientific research as a team sport for communal benefit** - represents several challenges for most faculty researchers. Foremost, many faculty lack the knowledge and human resources required to plan for the use of their data and/or code outside of their own projects. This means that data and computational pipelines are generated to meet immediate or short-term needs, often associated with a specific project. Once the project is completed, data and code that could benefit other researchers (and even future projects within the same faculty group) languish.

The current guidelines for proper data sharing follow the **F.A.I.R. principles** - Findable, Accessible, Interoperable, and Reusable [4] (see Figure 1). There are existing web-accessible platforms on which data may be shared in a reproducible manner, e.g., Cyverse [6], the Dryad Digital Repository [7], and resources from the National Center for Biotechnology Information (NCBI) [8]. These are complemented by professional organizations that focus on research data management, e.g., the Research Data Alliance [9]. On our campus one of the key resources for information about F.A.I.R. is the **Uni-**

versity Libraries who have embraced the digital age of archiving [10]. Even with these resources, however, faculty time and effort are required to prepare either data or code for further scientific use. The National Institutes of Health and the National Science Foundation both encourage reproducibility and sharing through data sharing policies, yet it is difficult to secure financial support for long-term data storage and maintenance of data repositories. Most educational institutions view federal agency requirements to store and maintain data generated using public funds as an unfunded mandate. At this time the federal agencies have responded with some support via web repositories and databases. Processing and uploading data and associated code, however, remains largely unsupported. This indicates that a more sustainable model for cross-disciplinary data services and management is needed.

Associated with the challenges around proper data sharing and maintenance is computational toolkit development and maturity. As mentioned above, as with data, research groups often develop code for a specific purpose where the priority is one-time use. Once the individual responsible for the code (usually a student or postdoctoral scholar) moves to another project or leaves the university, the code is usually lost. Other research groups who could benefit from the toolkit and associated knowledge are forced to effectively start from scratch.

As with research data, there are web-accessible resources for hosting and sharing code that support efforts toward reproducibility [12,13]; see Figure 2. These include Github, the Science

Figure 1: F.A.I.R. principles (left column) and ways to support implementation [11]

Gateways Community Institute (SGCI) [14], and Cyverse (mentioned previously) where code can be linked with Jupyter notebooks and other tools for documentation and ease of reuse. Publishers are starting to take notice of the need to properly document and test code. For example, De Gruyter (publisher of more than 700 journals in the humanities, social sciences, and law), SPIE (the international society for optics and photonics), and BMC Bioinformatics allow authors to share working code associated with their publications through Code Ocean [15], a platform for code and data sharing to improve research reproducibility. The Nature Publishing Group insisted as a condition of publication in a Nature Research journal that authors make data, code, and associated protocols available in a timely fashion to readers without undue qualifications [16].

or recruitment of **application specialists.** These are individuals with advanced training (i.e., hold or are earning graduate degrees) in a discipline related to data science (e.g., computer science, engineering, physics, statistics, bioinformatics) who serve as catalysts for cross-disciplinary research. They span disciplinary boundaries and can manage multiple projects simultaneously. This is an extension to cross-disciplinary contexts of the basic concept behind the NSF project in Advanced Cyberinfrastructure Research and Education Facilitators [18], the Carpentries [10,19], and the SGCI. As can be seen in Figure 3, effectiveness in data science requires a taxonomy of skills and the idea is to match these with disciplinary knowledge and the ability to communicate within interdisciplinary environments. These individuals would reside within research core facilities or Centers

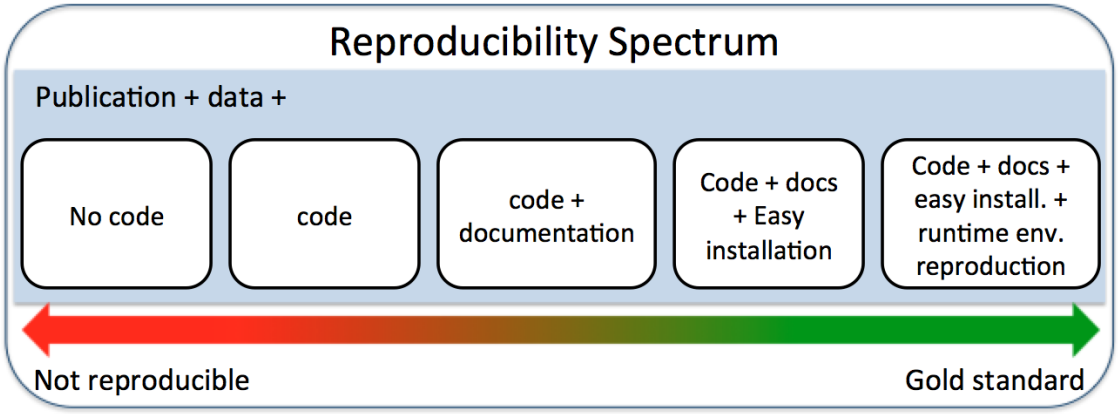


Figure 2. Reproducibility Spectrum for Research and Publication [17].

Hence faculty researchers are discovering that proper documentation and sharing of code are a requirement for publication in many highly respected venues. The challenge for institutions is how to support researchers who should or must meet this requirement.

A proposal on our campus that would provide this support is the development and be tasked with facilitating transdisciplinary research through knowledge of data and advanced cyberinfrastructure. Some are heroically technical while others emphasize the translation of scientific problems to computational solutions. These specialists often serve an `on-boarding' role for new staff/faculty/students to orient them to data and

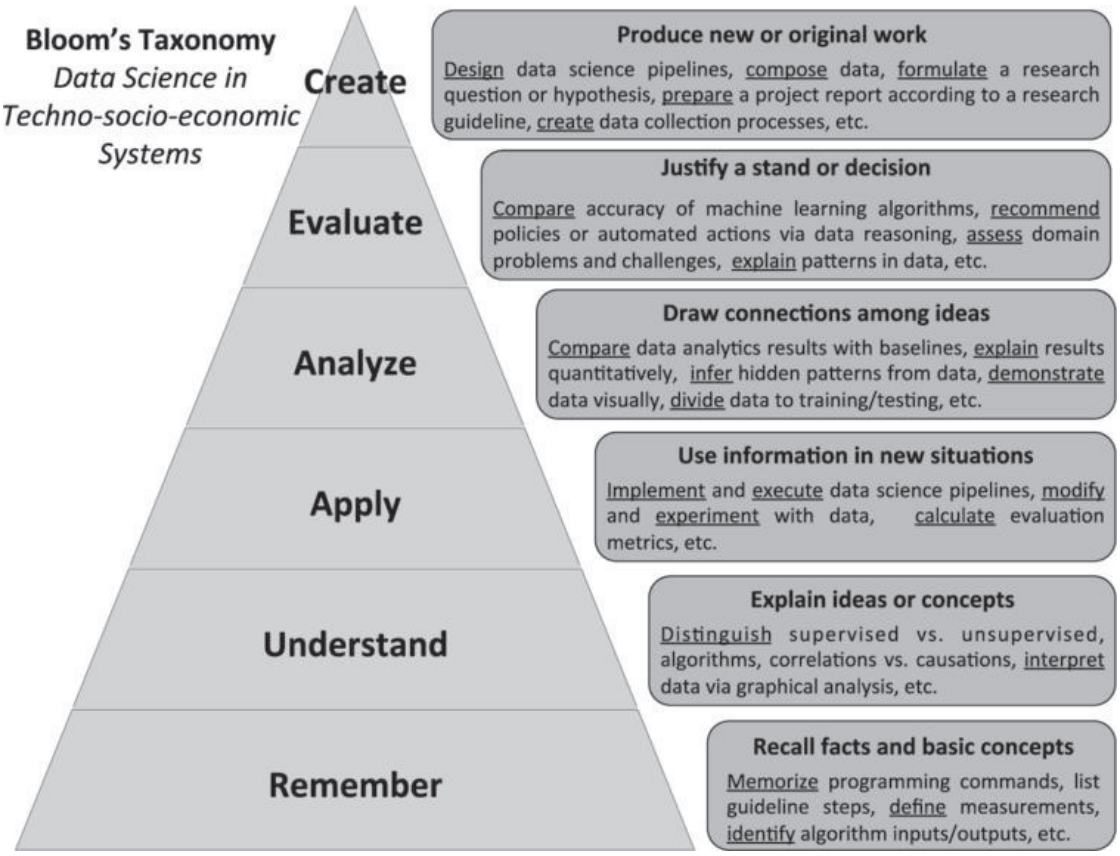


Figure 3. Bloom's Taxonomy applied to Data Science and Computing [20].

computing resources and current interdisciplinary projects. One of their most important roles is as repositories of institutional memory; in other words, they enable the use and reuse of data and code from university research projects. Even if they are graduate students, part of their job as application specialists is to document institutional projects and associated resources as part of building this memory. This leads to efficiencies in research that cannot be gained elsewhere.

It is important to note that even with the resources, tools, and willingness to meet the gold standard for reproducibility, there are additional challenges to conducting research in a transdisciplinary environment. We define **transdisciplinary research** as research that results in new knowledge formed via the integration of those domains that contribute to them; see Figure 4 [21]. Building an effective transdisciplinary team requires strong communication, not only of scientific concepts and ideas but also disciplinary expectations in terms of research output and recognition. The team must have a clearly defined goal and be able to articulate how each contributing discipline is expected to benefit. Faculty members who are willing (and able) to work in such environments need institutional support, as it takes considerable time and effort to build a team and realize the tangible benefits. Fortunately, federal funding agencies, in particular the NSF with their Big 10 Idea of **Growing Convergence Research** [22], are willing to support efforts in this direction. In effect, transdisciplinary research is a ``high risk, high reward'' endeavor. It is the role of the institution to mitigate the risk for fac-

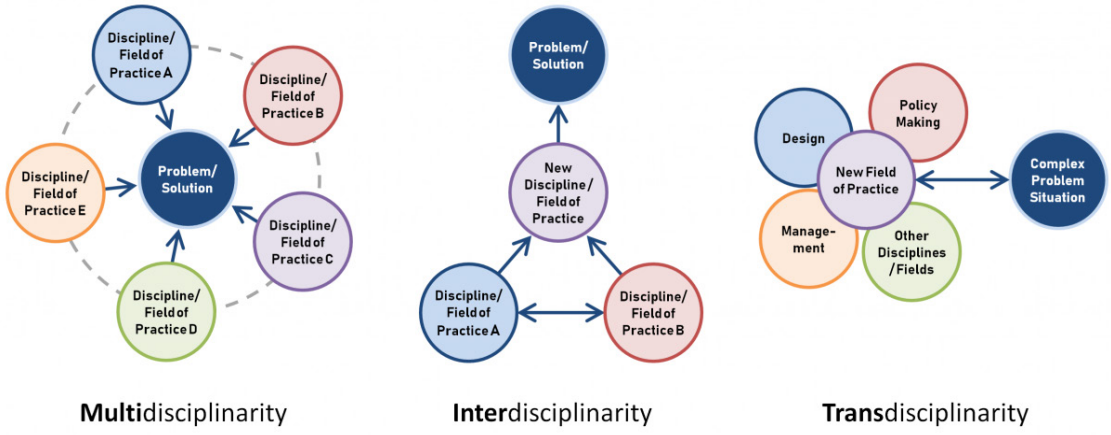


Figure 4. Multidisciplinary, interdisciplinary, and transdisciplinary approaches to research [21].

ulty members so that significant rewards can be realized.

On our campus we provide training opportunities for students and faculty to support acquisition of data and computing skills, reproducibility, and convergent research. These include data and software carpentry workshops, digital commons, and data archiving by the Libraries and the High Performance Computing Center. We also support an interdisciplinary **PhD program in Complex Biosystems** through the Office of Graduate Studies [23]. This program prepares doctoral students to conduct research that requires knowledge of both the data and life sciences. Each student is mentored by a pair of faculty advisors, one from the data and computing disciplines and one from the life sciences. Students in this program have earned predoctoral awards from several agencies including NIH, NSF, and the Foundation for Food and Agricultural Research (FFAR).

In summary, research has evolved into a team sport with members from multiple disciplines, working together toward a shared goal, enabled by continual advances in data science and cyberinfrastructure. As institutions of higher education, our role is to enable convergent research. We have articulated some avenues for support: reproducibility of data and code, University Libraries, application specialists, and strategic investments in transdisciplinary teams research. Convergence research is the future of science as solving some of society's largest challenges, from rural economic vitality to feeding our growing population, requires expertise in data, computing, and multiple scientific disciplines. The institutions represented at this year's Merrill Conference are well placed to play a leading role in the growth of convergence research to address societal challenges.

References

1. Watson, Peter (2017). Convergence: The Idea at the Heart of Science. Simon & Schuster.
2. Hutter, C. and Zenklusen, J. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173(2):283-285. doi:10.1016/j.cell.2018.03.042
3. The Laser Interferometer Gravitational-Wave Observatory (*LIGO*) Scientific Collaboration. Web Page. https://dcc.ligo.org/LIGO-M980279/public Accessed 2019-09-01

4. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18
5. The French Conseil Européen pour la Recherche Nucléaire (CERN). Web page. https://home.cern/ Accessed 2019-09-01
6. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, et al. (2016). The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLOS Biology 14(1): e1002342. https://doi.org/10.1371/journal.pbio.1002342
7. The Dryad Digital Repository. Web Page. https://datadryad.org/ Accessed 2019-09-01
8. The National Center for Biotechnology Information (NCBI). Web Page https://www.ncbi.nlm.nih.gov/ Accessed 2019-09-01
9. Research Data Alliance (2019) "About RDA". Web Page. https://rd-alliance.org/about-rda Accessed 2019-09-01
10. Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo K, Zimmerman N, Hollister J. (2016). Ten simple rules for digital data storage. PeerJ Preprints 4:e1448v2 doi:10.7287/peerj.preprints.1448v2
11. Australian National Data Service. Web Page. https://www.ands.org.au/working-with-data/fairdata/training Accessed 2019-09-01
12. Wilson G. et al. (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. doi: 10.1371/journal.pcbi.1005510
13. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, et al. (2014) Best Practices for Scientific Computing. PLOS Biology 12(1): e1001745. doi: 10.1371/journal.pbio.1001745
14. Science Gateways Community Institute (SGCI). Web Page. https://sciencegateways.org/home Accessed 2019-09-01
15. Code Ocean (2018). De Gruyter Partners with Code Ocean to Improve Research Reproducibility. Press Release of 2018-07-10. Web Page. https://codeocean.com/press-release/de-gruyter-partners-with-code-ocean-to-improve-research-reproducibility Accessed 2019-09-01
16. Nature Publishing Group (2019). Nature Research Editorial Policies. Web Page. https://www.nature.com/nature-research/editorial-policies Accessed 2019-09-01
17. Akalin, A. (2018). Scientific Data Analysis Pipelines and Reproducibility. Towards Data Science. Web Page. https://towardsdatascience.com/scientific-data-analysis-pipelines-and-reproducibility-75ff9df5b4c5 Accessed 2019-09-01
18. Advanced CyberInfrastructure - Research and Education Facilitators (ACI-REF) (2019). Web Page. https://aciref.org/ Accessed 2019-09-01
19. The Carpentries. Web page. https://carpentries.org/ Accessed 2019-09-01.
20. Pournaras, E (2017). Cross-disciplinary higher education of data science – beyond the computer science student. Data Science, vol. 1, no. 1-2, pp. 101-117. doi: 10.3233/DS-170005
21. McPhee, C., Bliemel, M., & van der Bijl-Brouwer, M. (2018). Editorial: Transdisciplinary Innovation. Technology Innovation Management Review, 8(8): 3-6. doi:10.22215/timreview/1173
22. The National Science Foundation (NSF) Big 10 Ideas (2018). Web Page. https://www.nsf.gov/news/special_reports/big_ideas/ Accessed 2019-09-01.
23. PhD Program in Complex Biosystems (2015). University of Nebraska-Lincoln. Web Page. https://bigdata.unl.edu/phd-program Accessed 2019-09-01.

Protecting the Value of Interdisciplinary Collaborations in the Development of a New Budget Model

Carl Lejuez, Interim Provost/Executive Vice Chancellor
University of Kansas

There is a saying that if you want to know what an administrator cares about, do not listen to what they say are priorities. Instead, look at their budget and where they allocate resources.

The new budget model for the Lawrence campus of the University of Kansas (KU) was developed as part of a larger effort to build a more stable and fiscally healthy KU, where priorities and budgetary decisions tell the same story.

For as far back as anyone can remember, KU utilized a historical/incremental budget model that provides the same allocation to units annually. While this model provides relative certainty about available funding each year, it also limits new and higher-risk efforts and relies on the use of special arrangements to provide units trying new things with the funds needed for those efforts. Such arrangements always make sense in the moment but over time can become convoluted and have unpredictable and unintended consequences for the unit in question as well as for the overall budget of the university.

KU's traditional budget model also has limited the ability to prioritize interdisciplinary collaborations because these efforts can often require considerable resources and the traditional model does not allow for incentivizing/rewarding these efforts when they are successful. Such collaborations are important for a number of reasons: they broaden perspectives on research problems, strengthen research capabilities and productivity, allow for cost efficiencies, improve research opportunities for students, expand the ability to establish external partnerships, and create important relationships among faculty. New budget models that are revenue-driven are often thought to undermine collaboration because there is a belief that units would rather go it alone and would not want to work together and share in the benefits of strong collaboration.

Since April 2018, our leadership team, in close collaboration with the KU community, has been working to redesign KU's budget model for revenue allocation. A Responsibility-Centered Management (RCM) model, or a hybrid of it, has been adopted in increasing numbers among higher education institutions around the country. The RCM model offers decentralized budget authority where a percentage of revenue is controlled by the unit that generated that revenue. In a full RCM, this amount is 100% of the revenue generated, but the academic unit also must cover all of its costs, including its facilities and services, from support units such as the libraries. The academic unit would also cover the expenses of strategic priorities and other unit-generated initiatives. Hybrid models return less than 100% and provide additional budget based on historic allocations and/or central priorities aligned with academic units. In some hybrid models, academic units receive fewer funds from central administration but are not required to pay some and/or all of its service costs.

Our new model has many aspects of an RCM but funds service units separately and provides resources to the academic units based on a proportion of revenue generated as well as outcomes in centrally determined priority areas. Given the latter, we refer to our hybrid RCM as a Priorities-Centered Management (PCM) model. This approach intentionally and explicitly aligns our budget with our priorities, including research; undergraduate and graduate student success; the career development of our people; outreach across our state and beyond; and diversity, equity, and inclusion. It orders allocation in a meaningful manner across (1) foundational priorities, (2) institutional strategic priorities, and (3) unit allocations.

In May 2018, the Lawrence campus of KU (referred to throughout simply as "KU") underwent a $20 million base budget reduction for FY2019, which coincided with the start of efforts to create a new budget allocation model for our campus. To ensure the Lawrence campus community was educated about the rationale for a new budget model and the specifics of the model proposed, we held a series of seven town halls that were livestreamed and provided details on how the new model will better align resource allocations with strategic priorities. We presented information about university-level strategic investments, the way funding is distributed to academic units and academic service units, and funding foundational priorities such as merit-pay increases, building maintenance, and financial reserves. Participants had opportunities to ask questions and voice any concerns, thus contributing to the ultimate design of the model.

Significant budget review and revision work took place during the 2019 fiscal year by a working group in consultation with campus leadership. The working group developed and shared guiding principles that would help shape the development of the new budget model:

1. **Common Good:** We have a responsibility to focus on the greater good of the entire university rather than our individual units. The common good includes what is best for our students, academic excellence, and the overall health and sustainability of the university.
2. **Transparency:** As we go through the process of developing a new budget model, there will be the broadest possible participation, sharing of decisions, and next steps. Diversity of viewpoints is encouraged and accepted. Constant and direct communication occurs so all stakeholders are thoroughly informed about the new budget model processes and issues.
3. **Clarity:** Simplicity is preferred over complexity.
4. **Innovation:** While appreciating what is positive about prior budget models and approaches, we will encourage innovative planning to adapt to the current environment.
5. **Responsiveness:** We remain cognizant of the changing environment and design toward nimbleness to respond to future campus changes.
6. **Respect (for each other and the product):** The budget model will be a result of the investment and input from the entire leadership team and their consideration of feedback solicited from constituencies across campus.

The working group met roughly every week during the 2019 fiscal year with the aim to identify:

- a high-level structure of allocation
- formula for determining fund allocations for academic units and academic service units, and an order of operations that ensures the ability of leadership to plan
- general criteria of incentivized activities
- means of evaluating progress toward unit and institutional goals (incentivized activities)
- mechanisms for reducing extreme swings in fund allocation
- a calendar for implementing the new budget model, evaluating unit performance, and determining future allocations

Throughout the year, direct reports to the provost, including deans, vice provosts, and other high-level directors, met on a weekly basis to discuss the development of the model. The provost also held regular town hall presentations of model progress and held hundreds of individual and small-group meetings of faculty, staff, and students from across campus. These meetings included participant feedback and insight that led to many important changes and a stronger overall model. The resulting PCM model will allow the university to save and invest in priorities and better support program innovation geared toward advances in strategic priorities.

Institutional Overview: The 1-2-3 of the New Budget Allocation Model

The first structural feature of the model is the creation of three broad categories in which budgetary resources can be allocated. The source of these funds is state appropriations, tuition revenue, and certain other operating revenues. Other funding streams, such as course fees, differential tuition, and fee charges for services, will be allocated directly to the appropriate unit and are not part of this budget allocation model or the pool of funds being distributed. KU Endowment funds available to various campus offices are also separate from the budget model.

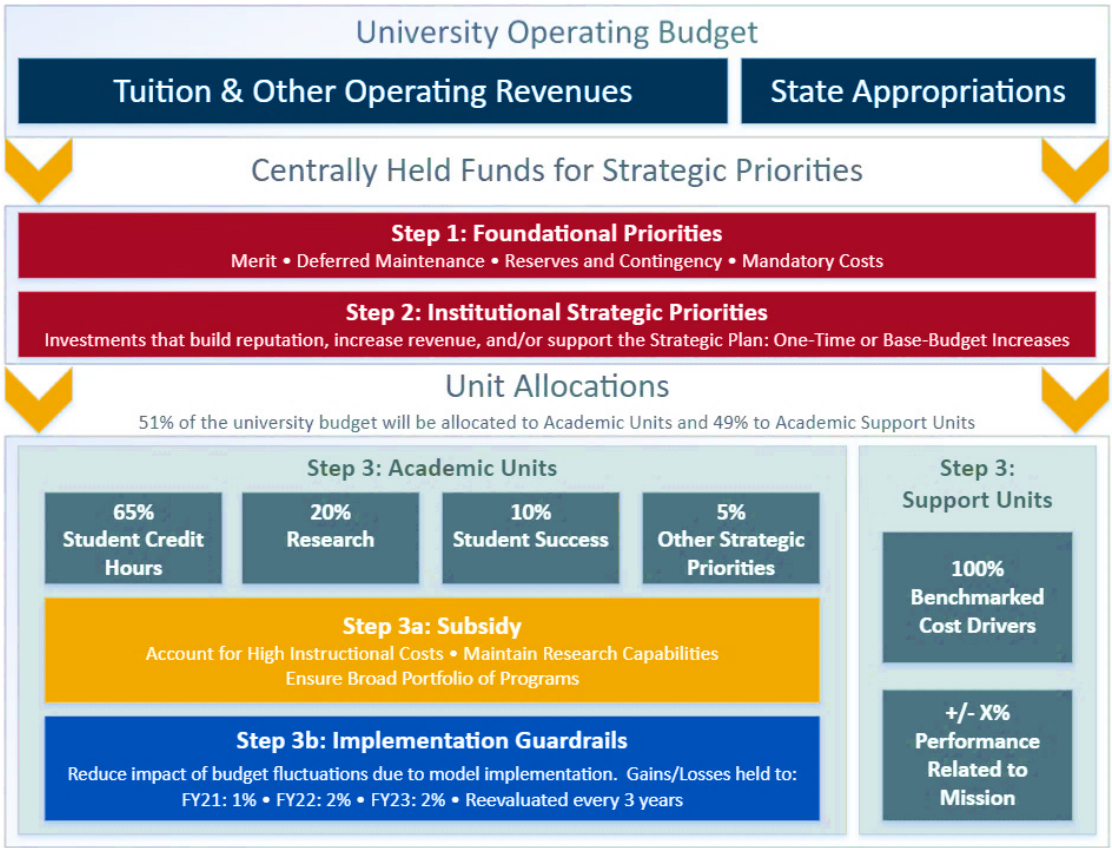
- **Step 1: Funding for Foundational Priorities**. The chancellor and provost will capture a limited pool of funds from state appropriations and tuition revenue. Funds will be used to support central—foundational—priorities, such as restoration of savings/reserves and contingency funds, regular merit raises, deferred maintenance needs, and increased year-to-year mandatory costs (external licensing, subscriptions, etc.).
- **Step 2: Funding for Institutional Strategic Priorities**. Strategic priorities are areas that are part of KU's vision, future goals, and growth. The funds are held by the chancellor and the provost and are expected to be used for one-time or limited-term investments that can build reputation or revenue opportunities. These may also be used to support initiatives determined through the strategic planning process and will largely take the place of "Provost Commitments" currently made throughout the school year on a case-by-case basis. Some Provost Commitments will still be made as needed but shifting funds to a more strategic approach helps ensure more vision-aligned investments.
- **Step 3: Determination of Allocations to Units**. This distribution

will provide resources needed by each academic unit (the schools and the College) and academic support units that provide crucial services to students, faculty, and staff. The pool for allocation is approximately $420 million. In Year One, roughly 51 percent of remaining funds will be allocated to academic units and 49 percent will be allocated to academic support units.

The new budget model, represented in the graph below, will take effect in Fiscal Year 2021 and will be based on performance in Calendar Year 2019. This will allow us time to obtain metrics beyond student credit hours. Once we have all of the necessary data, we will seek to balance investment in foundational priorities and allocation to units with strategic investments where possible. Evaluating the success of our new budget model requires that its structure and its results are clear and easy to understand.

Academic Unit Allocations

As noted above, remaining budget is largely split evenly across the academic (the College of Liberal Arts & Sciences and each of the professional schools) and support units. For academic units, funding will be distributed based on performance across a set of key priority areas. Where possible, performance will be centrally determined, but in some cases a centrally generated score is less possible/meaningful (e.g., research products and faculty development). In those cases, performance will be peer-reviewed by other deans the provost's leadership team with the provost assigning a final score. Percentages of budget allocation assigned for each key priority area are provided below.



- Student Credit Hours – 65%
  - For undergraduate units, 75% of SCH will be assigned to the instruction unit and 25% to the major. Students with more than one major count for both units.
  - For graduate units, SCH are weighted at 2x to support the higher cost of instruction. Graduate SCH is defined by the level of the student, not the course number. As with the undergraduate allocation, 75% is assigned to the instruction unit and 25% to the major.
- Research – 20%
  - Research Grants/Contracts (external direct/indirect costs), 10%
  - Grant Efforts/Success (faculty submitting external proposals and/or faculty with active external awards), 2%
  - Research Products (journal articles, books, chapters, creative/scholarly/legal works, invention disclosures/patents issued), 4%
  - Research Impact (awards/ recognition, citations, invited presentations, editorships, national/disciplinary ranking, societal and economic impact), 4%
- Student Success – 10%
  - Graduation Rates (degrees awarded), 2.5%
  - Time to Degree (time from junior level), 2.5%
  - Student Experience at KU (participation in high-impact experience) , 1.66%
  - Placement Post-KU (placement rates, career counseling), 1.66%
  - Teaching and Mentoring Quality (pedagogical advances, awards to faculty-staff, assessment of teaching, advising, classroom focus, bottleneck courses), 1.66%
- Other Strategic Initiatives – 5%
  - Climate and Support (diversity, equity, and inclusion; faculty/staff development), 2.5%
  - Collaboration and Outreach (internal, local, and global; fundraising and alumni engagement; efficiency), 2.5%

In addition to the budgeting approach for the academic units provided above, several other budgeting strategies are notable:

Subsidy. In some cases, certain academic units may receive a subsidy outside of the SCH unit allocation process. Such subsidies recognize a variation in instructional costs or may, at the provost's discretion, ensure a broad portfolio of disciplines, preserve foundational areas that aren't well suited to the budget models, and/or maintain research capabilities in high-potential areas. The funding methodology for subsidies is to set aside an initial $20 million from unit allocation to the academic units. Units cannot receive a subsidy greater than 100% of its prior-year budget, and units with growth greater than 10% are capped and the difference is added back to the subsidy pool.

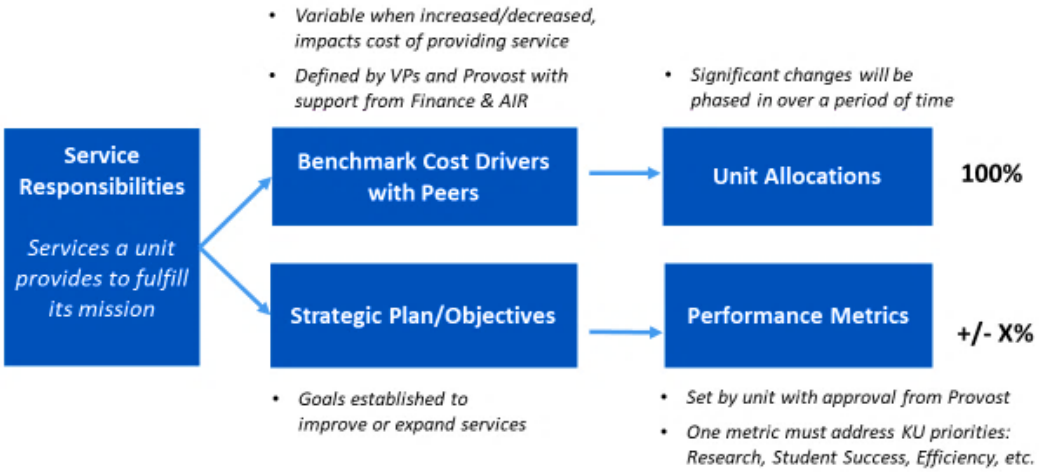
Implementation Guardrails. The budget model includes guardrails, a strategy to reduce the impact of budget fluctuations that could occur with implementation of the model. The strategy states that no school or the College will experience gains or losses greater than the percentages outlined below each year (as compared to the prior year):

- FY20: Historical Base Budget (0% guardrails)
- FY21: 1% guardrails
- FY22: 2% guardrails
- FY23: 2% guardrails
- FY24 & beyond: Continuation and/or % of guardrails reevaluated every 3 years

Academic Service Unit Allocations

Service unit allocations will be funded 100% based on service responsibilities, benchmarked to peers on cost drivers. They will reflect three to ten responsibilities that support the mission of KU and convey the goal and activities of each unit. Service units can be allocated additional funding based on their performance in meeting service responsibilities. Similar to the academic units, service units also will have a set of guardrails in place.

Graphic of Service Unit Allocation



PCM Budget Model Support for Interdisciplinary Collaborations

Flexibility in funding strategic priorities can make it possible to provide increased support for interdisciplinary collaborations. This begins with Step 2 of the model that supports strategic initiatives at a time when there is great awareness that interdisciplinary initiatives are most likely to bring teaching innovation, research impact, and the development of new external resources. These collaborations also support key components of the model that directly reward interdisciplinary work including collaboration, outreach, and efficient use of funds, as well as research impact, expenditures, and pedagogical advancements, all of which benefit from interdisciplinary efforts.

The University of Kansas recognizes that successful grant awards are essential to a robust research enterprise, and interdisciplinary relationships are key to that success. Not only are grant dollars greater and increasingly more available at the intersection of disciplines, there are benefits that extend beyond dollars. In an era of heightened competition for scarce resources, higher education must continually seek ways to collaborate, share data where possible, and bring needed perspective to the problems that research seeks to address.

For more information, please visit http://provost.ku.edu/budget-model-redesign.

Cross-Disciplinary Research: From Nuclear Physics to Cosmic Ray Detection and Medical Applications

Christophe Royon, Tommaso Isidori and Nicola Minafra
Department of Physics and Astronomy
The University of Kansas, Lawrence, USA
christophe.royon@ku.edu

After a short introduction about the Large Hadron Collider at CERN, Switzerland, we will discuss briefly the fast timing detectors built to measure intact protons. The applications of these detectors concerning cosmic-ray detection and medical applications will be described.

The Large Hadron Collider and Timing Measurements

The Large Hadron Collider

The Large Hadron Collider (LHC) located close to Geneva at the border between Switzerland and France collides protons with a centre-of-mass energy of 13 TeV, making it the highest energetic collider in the world. The idea is twofold: a better understanding of the proton structure in terms of quarks and gluons and reproducing conditions as close as possible to the big bang where new particles might be produced. The LHC energy allows getting similar conditions at the particle level at about 10^{-13} seconds after the big-bang. In most cases, the interacting protons are completely destroyed after interactions and general purpose detectors such as ATLAS[1] and CMS[2] have been built to identify and measure all kinds of particles that are produced after the interaction. Since particles need more or less material to be absorbed in a material according to their type and energy, the structure of such detectors is always made of different layers dedicated to measure successively photons, electrons and positrons, pions, protons, neutrons and finally muons that need a lot of material to be absorbed. Only neutrinos cannot be directly measured and appear as missing energy in the detector. Detectors such as ATLAS and CMS are large and heavy; for example, the site of the ATLAS detector being of the same magnitude as Mount Rushmore in the USA and the weight of the CMS detector being larger than the Eiffel Tower in Paris. In addition to these two main experiments, smaller, more dedicated experiments exist such as LHCb, ALICE, TOTEM, MOEDAL...

Recently, some "strange" events were observed at the LHC where protons are found to be intact after interacting. An everyday analogy would be one gets an accident between two trucks (the protons) and both trucks are intact after the accident and, in addition, some small cars (additional particles) are produced during the collision. The two trucks will however be slower: the protons "donate" part of their energy to create the additional particles. The LHC magnets are used as a spectrometer to measure the intact protons in the final state. Namely, the radius of curvature of the intact protons in the final state is smaller than for the beam protons since they lost part of their energy. This clearly means that it is possible to detect these intact protons after interaction by installing detectors very close to the beam. This is why both ATLAS and CMS-TOTEM[3] Collaborations installed detector in so-called roman pots, at few mm from the beams, about 220 m down-

stream the interaction point, to measure the intact protons scattered at very small angles. A scheme of the proton detectors in the case of the CMS-TOTEM collaboration is shown in Figure 1 as an example.

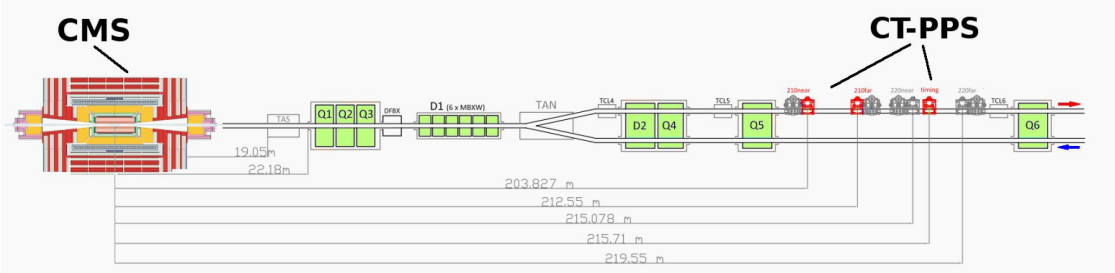


Figure 1. Schematic view of the CMS detector and the roman pot detectors from TOTEM. Only one side is shown.

Measuring proton time-of-flight at the LHC

The LHC collides clouds of hundreds of billions of protons together; this means that there are multiple proton collisions occurring within the same bunch crossing. What we are interested in, as an example, is the production of two photons or two *W*/*Z* bosons together with two intact protons that could be a sign of extra-dimensions in the universe, composite Higgs bosons or axion-like particles[4, 8].The issue is that the two photons or the two *W*/*Z* bosons can originate from a different interaction than the two protons as shown in Figure 2. In order to reject these uninteresting events (called background), it is possible to measure precisely the time of the protons interaction. Namely, we can constrain the protons to originate from the same interaction point as the two photons or *W*/*Z* bosons. Since particles at the LHC travel at the speed of light, time needs to be measured with high precision, of the order of ten picoseconds (1 ps=10^{-12} s). Fast silicon detectors together with their readout electronics have been developed in order to achieve this goal.

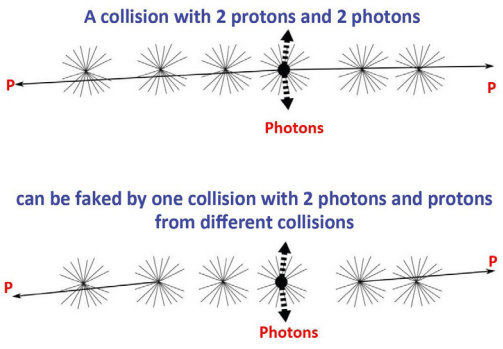


Figure 2. Pile up processes at the LHC.

Performance of Timing Detectors at the University of Kansas

At the University of Kansas (KU), we designed a multi-purpose electronics board to read out silicon or diamond detectors to measure precisely the time at which particles cross the detector, as well as a test-stand in order to test the full chain from the detector to the read-out electronics. The test-stand is equipped with a laser or a radioactive source in front of the silicon detectors (see Figure 3). The system is highly adaptable to different kinds of sensors (diamond or Silicon) and only requires a power supply to operate. The read-out electronics produces a signal that can be analysed using a digital scope or some waveform signal analyser, such as SAMPIC[5, 6, 7]. The amplifier was designed at the University of Kansas and can be used for a full range of detectors and applications.

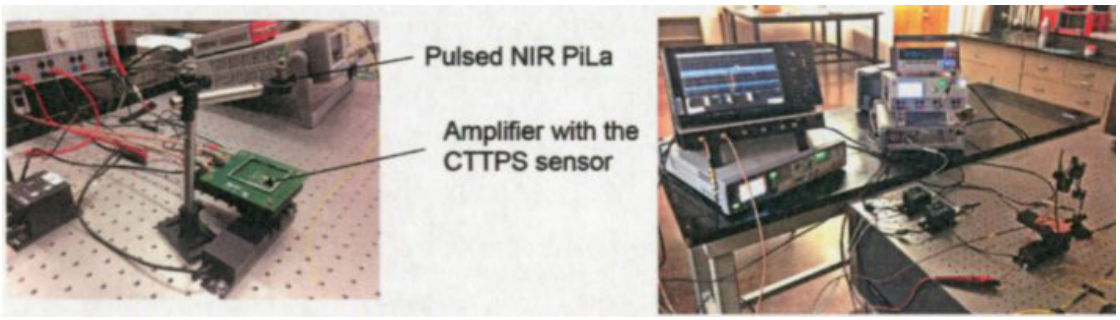


Figure 3. Timing detector test stand at the University of Kansas.

The performance of the timing detector and its amplifier is shown in Figure 4. In order to test the full system in real conditions for nuclear and particle physics, we used a test using a particle beam at Fermilab, Batavia, USA. Using a single layer of silicon sensor, we obtained a resolution of about 39 picoseconds, which means that a resolution better then 15 picoseconds can be achieved with 8 layers of these detectors. In particular, the sensor technology that was used is often referred as Low Gain Avalanche Detectors (LGAD), or Ultra Fast Silicon Detectors. On Figure 4, we can see a photo of the board designed and built at the University of Kansas. The idea was to build a "plug-and-play" amplifier that can be used to test different kinds of sensors for different applications. The performance of the amplifier is similar or even better than commercial ones with a cost about two orders of magnitude lower.

Possible Applications of Timing Detectors and Analysis Techniques

In this section, we will discuss three possible applications using Ultra Fast Silicon detectors and the electronics that was developed at KU, namely the measurement of cosmic rays in collaboration with NASA, of doses applied for cancer treatment in collaboration with KU Medical Center and a better understanding of catalysis in chemistry.

Measuring signals of a diamond or Ultra Fast Silicon detector

All applications that we are going to discuss rely on the same principle: we

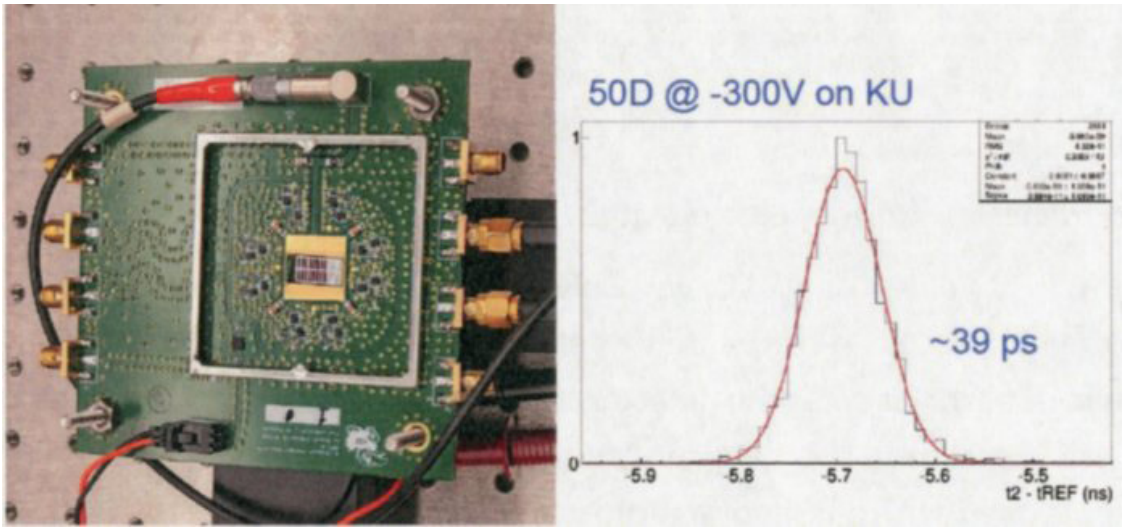


Figure 4.Timing board made at the University of Kansas.

need to analyse the full signal produced by a sensor at the passage of a particle. The applications do not rely too much on timing measurements (as in high energy particle or nuclear physics) but in the measurement of the number of particles or charge that crosses the detector. The idea is similar and it is illustrated in Figure 5. When a particle (for instance a proton) crosses a detector, some pairs of electrons and "lack of electrons", called "holes", are formed and drift slowly for the ions or fast from the electrons towards the electrodes because of the electric field. Detecting the proton passing through the detector is thus possible measuring the signal induced on the electrodes using dedicated electronics.

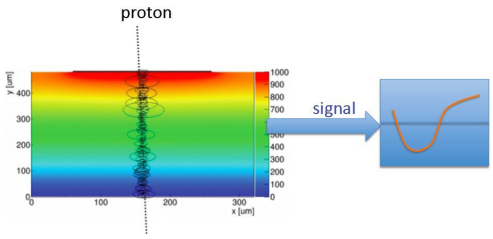


Figure 5. Scheme of signal induced in a silicon detector at the passage of a particle.

Since we want some automatic method to detect particles or to measure deposit charges, a dedicated electronics system was delayed at KU as we mentioned already. The next steps are illustrated in Figure 6. Signals directly read out from a detector need to be amplified. The first step of the KU circuit is then to amplify these signals without affecting too much the properties of the signals, like shape and amplitude with respect to the noise. In order to measure the signal, a very fast digitization is performed (taking as an example 64 measured points in a few nanoseconds). A mathematical interpolation between the different measured points allow then a smooth reconstruction of the full signal. The method allows the precise measurement of the time of crossing of the particle and, at the same time, the signal amplitude and other signal characteristics, like the rise time and the duration.

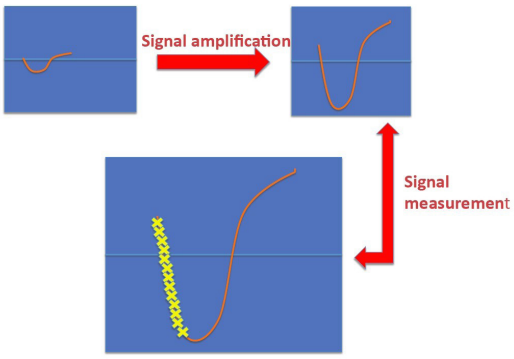


Figure 6. Scheme of amplification of a signal coming out of a Silicon detector.

Application 1: Measurement of cosmic rays with NASA[10]

The idea for this project in collaboration with NASA is to measure the type and energy of cosmic ray particles originating from the sun for a range between keV to GeV as illustrated in Figure 7.

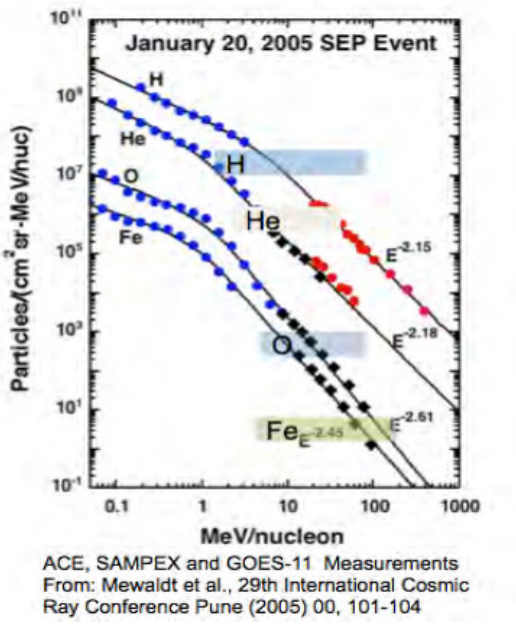


Figure 7. Spectrum of solar cosmic rays.

In order to do so, build a detector made of sandwiches between active layers of Si detectors and absorbers that allows the measurement of different kinds of particles with different energies. Using the very fast digitization described above, it is possible to reconstruct fully the signal in the different Si layers. Since different kinds of particles deposit their energies in the different layers differently, also depending on their energies, it will be possible to reconstruct the full properties of the comic ray radiation analysing the digitized signals. This project aims at preparing a prototype of a cube sat in collaboration with NASA (the AGILE project, shown in Figure 8) in the next

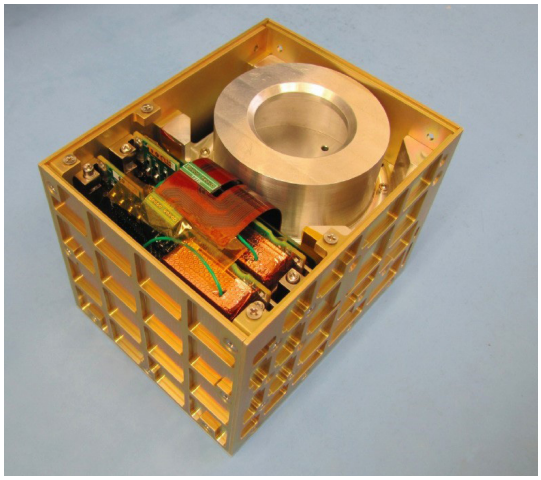


Figure 8. The AGILE project with NASA.

three years and will eventually help with the precise measurement of radiation between the Earth and Mars needed in order to send astronauts to Mars.

The inconvenience of digitizing the signal in each Si layer is that the amount of data originating from the detector will be quite high. This is needed to pre-process data before sending them back to Earth. Advanced analysis techniques have been developed successful in high energy physics to discover the Higgs boson, as an example, or to look for physics not explained by the standard model. The amount of data accumulated by the LHC experiments is very large and requires neural networks or other advanced techniques in order to analyse them. For this application, advanced techniques will be needed to filter and optimize the relevant important data that will be sent back to Earth.

Application 2: Measuring radiation in cancer treatment[9]

The second application deals with the precise measurement of radiation accumulated by the human body during cancer treatment using photon (radiotherapy) or proton (hadrotherapy) beams. The idea is to measure the amount of radiation delivered by the medical particle accelerators with millimetre precision. Furthermore, present techniques are not able to count exactly the number of particles produced, but they measure the average charge deposited inside the sensor. A more precise method consists in counting the number of photons or protons that pass through the sensor. This project is being developed in collaboration with KU Medical Center and, if successful, it will allow a more optimized dose absorbed by the patient during cancer treatments.

Another possible medical application deals with PET imaging. Usually, patients absorb radioactive material that interact with electrons inside the tumour, emitting photons that can be measured and that can be used to create an image of the tumour. The problem is that the human body emits naturally lots of photons and in average, 1 pair on photon originates from the tumour out of 10,000. An advanced analysis is then needed in order to isolate the interesting photons. In order to preselect the photons originating from the tumour, it is possible to measure the time of detection of the photons and

require them to originate from the tumour itself. This would allow to produce a more effective and faster image of the tumour. This development would be a fundamental application as well and has been the interest of many private companies in the world.

Application 3: Understand better catalysis in chemistry

The third application deals with a better understanding of catalysis in chemistry with the application of reaching better methods to desalinize sea water, as illustrated in Figure 9. The idea is to understand better how an interface between two liquids, a solid and a liquid, or a gas and a liquid vary as a function of time when catalysis occurs. Using interferometry technics, we can measure how the interface varies as a function of time by measuring a snapshot every 20 or 30 picoseconds. This will lead to new insights in the mechanism of catalysis and thus in a better understanding of applications where catalysis is needed. This could also have implications on the way medicine is absorbed by human body by improving the interface between human cells and medicine.

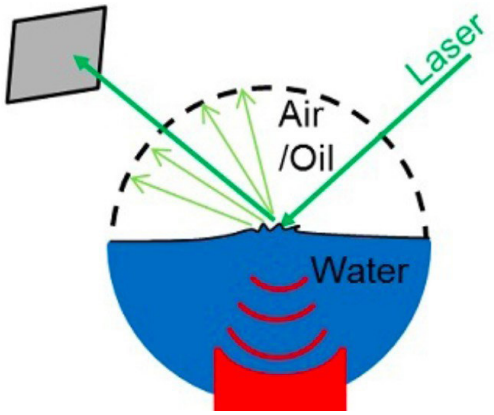


Figure 9. Catalysis measurements in chemistry.

Conclusion

In this short report we describe the fast timing detectors and their electronics developed originally for high energy and nuclear physics as well as the potential applications in cosmic ray measurements, medical analysis and chemistry performed at the University of Kansas.

References

1. ATLAS Coll., CERN-LHCC-99-14, CERN-LHCC-00-15.
2. CMS Coll., CERN-LHCC-94-38.
3. TOTEM Coll., JINST **3** (2008) S08007.
4. S. Fichet, G. von Gersdorff, B. Lenzi, C. Royon, M. Saimpert, JHEP **1502** (2015) 165; S. Fichet, G. von Gersdorff, O. Kepka, B. Lenzi, C. Royon, M. Saimpert, Phys.Rev. D89 (2014) 114004; S. Fichet, G. von Gersdorff, C. Royon, Phys. Rev. Lett. **116** (2016) no.23, 231801; S. Fichet, G. von Gersdorff, C. Royon, Phys.Rev. D**93** (2016) no.7, 075031; C. Baldenegro, S. Fichet, G. von Gersdorff, C. Royon, JHEP **1806** (2018) 131; C. Baldenegro, S. Hassani, C. Royon, L. Schoeffel, Phys. Lett. B**795** (2019) 339; C. Baldenegro, S. Fichet, G. von Gersdorff, C. Royon, JHEP **1706** (2017) 142; E. Chapon, C. Royon, O. Kepka, Phys.Rev. D**81** (2010) 074003; O. Kepka , C. Royon, Phys.Rev. D**78** (2008) 073005.
5. E. Delagnes, D. Breton, H. Grabas, J. Maalmi, P. Rusquart. Nucl. Instrum. Meth. A**787** (2015) 245.
6. A. Apresyan *et al.*. Nucl. Instrum. Meth. A**895** (2018) 158.; N. Minafra *et al.*, Nucl. Instrum. Meth. A**867** (2017) 88; D. Breton, V. De Cacqueray, E. Delagnes, H. Grabas, J. Maalmi, N. Minafra, C. Royon, M. Saimpert, Nucl. Instrum. Meth. A**835** (2016) 51.

7. N. Minafra, PhD thesis, http://cds.cern.ch/record/2139815/files/CERN-THESIS-2016-016.pdf; H. Grabas, PhD thesis, https://cds.cern.ch/record/1700497/files/VD2 GRABAS HERVE 03122013.pdf.
8. KU News, https://news.ku.edu/2019/04/25/research-explores-behavior-quarks-and-gluons-large-hadron-collider;
9. Physics World, 04 June 2019, https://physicsworld.com/a/particle-telescope-technology-could-help-improve-radiotherapy/
10. KU News, https://news.ku.edu/2019/05/03/particle-telescope-will-probe-subatomic-makeup-suns-cosmic-rays-and-could-lead-more

Complexities of Conducting Cross-Disciplinary Biomedical Research

Jennifer Larsen, MD, Vice Chancellor for Research
W. Scott Campbell, PhD, Senior Director of Research and IT
University of Nebraska Medical Center

Cross disciplinary research has become routine at academic health centers because larger teams with a broader range of skills are needed to solve complex health related problems. Researchers routinely reach out to colleagues to understand how their work "at the bench" is, or could be, relevant to future clinical care, as well as, how to better incorporate what has been learned in a clinical trial into the community. To build effective teams, there are many "complexities" that must be anticipated and/or addressed.

The Complexities

Some of the common complexities involved in conducting cross disciplinary biomedical research are outlined below.

Defining the Rules of Engagement

To form an effective team requires time and discussion. Even as the team is assembled and before the data is gathered, teams should discuss rules on how the data will or will not be shared with others, could or could not be moved, and how each member will be acknowledged for their role in any published results. Who will write or lead each manuscript or grant requires frank discussions long before they are written or submitted.

Vocabulary

As simple as it sounds, to form a functional team, an environment must be created where all members can be understood, and their ideas are welcomed. This starts with encouraging everyone to speak in words most can understand by avoiding terminology and acronyms specific to one discipline that might not be understood by others. This is also important if teams want to ever attract new members, including students. Vocabulary goes beyond conversations. It also addresses how the data is captured and stored so that the data can be more easily shared. Using a common format, preferably an established vocabulary standard (e.g., SNOMED, LOINC) that includes meta-data to allow members to understand how the fields are defined, for more consistent and reproducible data collection, as well as queries and analyses, or to combine with other data sets. Team members with terminology expertise are very valuable, and team members who can translate between disciplines are essential to an effective multidisciplinary team.

Data Transfer and Storage

Many teams require data to move from one place to another, like from a research instrument or electronic health record to a data storage space or a research database where the analysis will be performed. More teams are working with large research files, terabytes or more, such as DNA sequencing data or image files (e.g., MRI or other anatomic imaging files) or data from large populations. These data sets often have to be stored in the cloud or in large data centers able to accommodate such large data, but moving files can be time and resource consuming. Large data sets often must be stored in their entirety at the point of creation until the full copy of the dataset is transported, stored and validated for completeness in its new lo-

cation. Often discussions arise regarding "whose data is it". Differences of opinions need to be ironed out, including what federal or commercial entity rights or patient/research subject perceptions that might be involved. The cost of data storage is often underrecognized as well. How that cost will/will not be subsidized by the team members or grants must be determined and use of data steward(s) (e.g., personnel) to maintain and distribute data sets may be necessary to include in the cost structure.

Privacy and Security

The data storage vehicle depends in part on what data is being stored. Protected health information (PHI), Protected individual information (PII), as well as other sensitive data (e.g., student data, high security data) may require special controls for who can access the data and the ability to audit who has accessed the data. Data associated with an FDA application or trial needs to meet FDA's Title 21 Code of Federal Regulations (CFR), Part 11 requirements. Many researchers are not as knowledgeable as they should be of the eighteen PHI identifiers defined by the HIPAA legislation (https://www.hhs.gov/hipaa/index.html and shown in Table 1). As a result, researchers incorrectly believe their data is deidentified and attest as such, when the data is, in fact, still considered identifiable.

Special Considerations with Global Sites, Teams or Focus

There are increasing, and often changing rules, when data, samples, equipment, or team members move between or live in other countries. Countries have varying "export control" regulations concerning what is or is not allowed to cross into or out of their country. For the US, this includes interactions with specific individuals whether located in another country or in the US and specific types of equipment. For many countries, this involves export of biologic samples or data. In particular, the new European Union Data Protection Regulation (EUDPR) introduced in 2016 requires anyone acquiring data in an EU country, even if acquired on the property of and from citizens of another country to meet specific standards and receive EU approval. These standards further apply to data moved from an EU country.

Table 1: 18 PHI identifiers

- Names
- Dates, unless year alone
- Telephone numbers
- Geographic data (address, full zip)
- FAX numbers
- Social Security numbers
- Email addresses
- Medical record number
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers including license plates
- Web URLs
- Device identifiers and serial numbers
- Internet protocol addresses
- Full face photos or comparable images
- Biometric identifier (i.e. fingerprint)
- Any unique identifying number or code

Problem Solving

Teams will always encounter problems including personality disputes, intellectual property disputes, 'I contributed more than you did' disputes, and 'but you promised me' disputes, among others. Ideally, the team would have discussed potential conflicts as the team is developed, including how the team would anticipate solving conflicts and identify a structure, process, or person(s) within or outside the team to resolve disputes if

additional help is needed. Such a resolution strategy is often required in multi-PI grants by funding sponsors. These kinds of agreements are best documented and discussed for new teams so there is no misunderstanding later. If the team has not had such discussions, team leaders or members may need to reach out and find the best mediator after the fact, such as the research integrity officer or another senior leader that all parties agree to listen to for dispute resolution. Team leaders should be proactive—watching for signs of frustration or conflict and address issues before they become impossible to resolve. Michelle Bennett, PhD, of the National Cancer Institute (NCI) assembled a Field Guide for Collaboration and Team Science available on line that provides many practical approaches to common problems (https://www.cancer.gov/about-nci/organization/crs/research-initiatives/team-science-field-guide/collaboration-team-science-guide.pdf).

Discussing Data Sharing with the Public

As time has passed, the public has seen more and more examples of times their data has been shared or "leaked" that they were not aware could occur. Many investigators believe sharing of de-identified data is acceptable and may even assume that no one would care. In fact, many individuals are comfortable with researchers sharing their personal data, even identified data, if they are informed in advance and have given their permission, such as through informed consent or broad consent. However, others may feel differently, even if their data is deidentified, hence the recent class action lawsuit or patients who objected to the University of Chicago Hospital who gave deidentified health data to Google as part of an artificial intelligence project. These concerns can be proactively addressed through the informed consent document, town hall meetings, or other public discussions about the importance of the study and what the study is supposed to accomplish, or having a community advisory board of community leaders to be a sounding board about the methods to be used and help the investigators disseminate the results when they are found. As researchers, if we do not have the public's trust, we may not have funding long-term. We can all do a better job of discussing the value of the data and the project with the public and working with community leaders to implement the data into day-to-day healthcare or other outcomes.

Summary

Team science is here to stay, and curating the datasets assembled by those teams needs to be discussed in advance. This is just one aspect of the complexities of conducting cross disciplinary biomedical research. Data sharing is usually good, often mandated by some funding mechanisms, is essential to multidisciplinary collaborations, and may result in even bigger datasets which can make moving and storing the data more challenging. The nuts and bolts of achieving data sharing, which may include data deidentification, moving large data sets, or loading data into a specific website can be confusing at best, and often difficult, as well. Data sharing may require new tools that are often developed with biomedical informatics experts, who are in too short of supply. Data sharing can create new risks if those researchers who are sharing data are not aware of the pitfalls, particularly when PHI is involved. But lastly, we cannot forget the public, who needs to be part of the communication before data is shared and after, to bring them along, to understand the value, and to fully understand and make use of the results that are found.

RETREAT PARTICIPANTS 2019

Keynote Speaker
Daniel A. Reed, Senior Vice President for Academic Affairs, University of Utah

National Library of Medicine, National Institutes of Health
Lisa Federer, Data Science and Open Science Librarian

University of Arkansas
Jim Coleman, Provost and Executive Vice Chancellor for Academic Affairs
Daniel Sui, Vice Chancellor for Research and Innovation

Kansas State University
Daniel Andresen, Director, Institute for Computational Research in Engineering and Science
Professor, Dept. of Computer Science

The University of Kansas
Mabel L. Rice, Fred and Virginia Merrill Distinguished Professor of Advanced Studies
Director, Merrill Advanced Studies Center
Simon Atkinson, Vice Chancellor for Research
Richard Barohn, Vice Chancellor for Research, KUMC
John Colombo, Acting Dean, College of Liberal Arts & Sciences Director, Life Span Institute
Samantha Ghali, Child Language Doctoral Program Doctoral Candidate
Teresa Girolamo, Child Language Doctoral Program Doctoral Candidate
Carl Lejuez, Interim Provost and Executive Vice Chancellor
Marianne Reed, Digital Initiatives Manager, The University of Kansas Libraries
Christophe Royon, Foundation Distinguished Professor, Department of Physics and Astronomy
Robert Simari, Executive Vice Chancellor, KUMC, Executive Dean & Franklin E. Murphy Professor of Cardiology

University of Nebraska-Lincoln
Jennifer Clarke, Professor, Food Science and Technology, Statistics
Deb Hamernik, Associate Vice Chancellor for Research
Bob Wilhelm, Vice Chancellor for Research and Economic Development

University of Nebraska Medical Center
W. Scott Campbell, Associate Professor, Sr. Director Research Technologies - UNMC
Director of Public Health Laboratory Informatics and Pathology Laboratory Informatics
Department of Pathology/Microbiology
Jennifer Larsen, Vice Chancellor for Research, Louise and Morton Degen Professor of Internal Medicine

Notes