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