National Institutes of Health (NIH) Workshop: Harnessing Artificial Intelligence and Machine Learning to Advance Biomedical Research

Porter Neuroscience Building, Rooms 610/640
NIH Campus

Monday, July 23, 2018

Contents

Workshop Description

Artificial intelligence (AI) and machine learning (ML) are advancing rapidly and in use across many industries, including biomedical research and healthcare delivery. For this full-day public workshop, NIH brought together leaders in innovation and science to explore the opportunities for AI and ML to accelerate medical advances from biomedical research. Workshop participants heard from leading industry experts and scientists who are employing AI/ML in biomedical research settings. Speakers covered a range of issues, including the promise of integrating AI technology into healthcare, how it is being used in biomedical research, and its potential for enhancing clinical care and scientific discovery. Craig Mundie, who served on the President's Council of Advisors on Science and Technology (PCAST) and was formerly Microsoft's Chief Research Strategy Officer, delivered the keynote address.

Access to the videocast of this meeting can be found at the following link: https://videocast.nih.gov/summary.asp?live=28053&bhcp=1. Discussions on Twitter can be found using the hashtag #2018biomedAI.

The content of the presentations summarized in this report represents the views of the presenters and are based on their own work, experiences, and opinions.

Executive Summary

Convened to discuss AI and ML in biomedical research, the attendees of this workshop explored how these technologies currently are used in biomedical research, ways to expand their use, and how to leverage existing resources to improve collaborations and networks. Keynote speaker Mr. Craig Mundie explored the ways in which mankind and machines can work together to advance biomedical research. He questioned whether the amount of biomedical data should be limited to only critical information, not expanded to collect all the data possibly available. He proposed the use of gaming methodologies to interrogate discrete biomedical data and competitions to provide incentive for people to participate in biomedical ML endeavors. With the use of AI, the paradigm of studying a population to understand what is occurring in an individual may be upended, shifting to the collection of individual data to create a composite for population data.

Other speakers addressed data sharing, which is a limitation for use of ML in biomedical research. A significant amount of data is needed for ML, but data sharing is currently not properly incentivized. Efforts to address improved data sharing were discussed.

ML has been used to make significant advancements in radiology and imaging. These advancements have reduced the time needed to annotate data and allowed expanded use; thus, more and richer data are available. Future advancements in use of AI in radiology include expanded personalized medicine, increased integration of radiological data with other data, and expanded use of radiology to improve global health.

AI is expanding into the realm of in-home patient monitoring through wireless technology. The advancements in this area were discussed, as well as the challenges with integrating this new type of data with existing data. AI has also been used to connect patients with clinical trials more efficiently using technologies like natural language processing (NLP). ML can be used to assist

in protocol trial design and has been used to perform predictive modeling of human disease states. ML can be used to review genetic variants in individuals in a specific cell type to understand a particular disease, such as Alzheimer's disease, and to understand the regulatory code of genomic control elements with a level of complexity not possible by human analysis.

Lastly, attendees discussed challenges and opportunities in this field. Challenges include recruiting the correct expertise to participate in biomedical research and fostering ways to improve collaboration and networking among scientists from the different fields. Training materials to facilitate collaborators' basic understandings of each other's fields—for example, clinical research, genetics, or computer science—are also lacking. Significant challenges were noted with respect to data, including data sharing, data harmonization, and issues related to data in electronic medical records. Challenges related to current methods were also discussed. Opportunities were noted within these same areas, including the expanding access to training materials, existing ML framework models that can be leveraged for future research, and opportunities that NIH might consider to foster AI and ML in biomedical research.

Promising future aspects of AI include the interpretation of information with complex structure, improvement of images, and assemblage of new data types. Combining medical imaging with genomic data and clinical test results will provide insights not allowed by analysis of data in isolation, and using AI in this space could ultimately lead to new types of therapy.

Introductory Remarks by the NIH Director

Francis Collins, M.D., Ph.D.

The advent of AI and ML, big data, cloud computing, and robotics may represent the Fourth Industrial Revolution. One of the initial hallmarks in this area was the Human Genome Project, which spanned 1990 to 2003 and resulted in one of the first large data sets intended to study the genome. The outcome of this project spawned several other NIH initiatives that combined biology and information science to study and interpret the data: Examples include The Cancer Genome Atlas (TCGA), the Human Epigenome Atlas, and the International HapMap Project.

One example of a large biomedical data set is the NIH Human Microbiome Project, which generates resources for the comprehensive characterization of the microbiome and the analysis of its role in health in disease. Over the last 10 years, this project has resulted in one of the major advances of our time by providing new insights into health and disease. Phase 1 was 5 years long, ended in 2013, and characterized microbial communities from 300 individuals across several body sites. It generated large data sets, with more than 14 TB of publicly available information. Phase 2, which began in 2014 and is ongoing, is integrative and focuses on specific health and disease areas, including pregnancy and preterm birth, irritable bowel disease, and prediabetes.

The *All of Us* Research Program, through the Precision Medicine Initiative, is another example of a large data set and aims to enroll 1 million participants; it officially launched on May 6, 2018. Participants will vary in lifestyle, socioeconomic status, environment, and biology. The resulting data set will provide access to an unprecedented number of variables, such as environmental,

social, behavioral, and biological/clinical information, enabled through the informed consent of participants. There is opportunity for AI and ML to foster discovery in this program.

For this workshop report, the definitions of AI, ML, and deep learning (DL) are framed as follow:

- AI: A larger umbrella of computer intelligence; a program that can sense, reason, act, and adapt
- ML: A type of AI that uses algorithms whose performance improves as they are exposed to more data over time
- DL: A subset of ML in which multilayered neural networks learn from vast amounts of data

Clinical applications of AI, ML, and DL include imaging (e.g., pathology diagnostics), diagnostics in dermatology or ophthalmology, radiology, cancer treatment, robotic surgery, and NLP of electronic medical record (EMR) data. Basic science applications include interpretation of imaging, neuroscience (e.g., the Brain Research through Advancing Innovative Neurotechnologies® [BRAIN] Initiative), genomic analyses (e.g., variants, risk of disease, gene structure), microbiome/metagenomics (e.g., multiorganism study), and epigenomics (e.g., histone marks, TF binding, enhancers, DNA methylation). A recently published paper, which Dr. Collins co-authored, discussed technology that accurately predicted DNA methylation values in whole-genome sequencing of multiple human tissues.

NIH Institutes and Centers currently invested in AI and ML include the National Institute of Biomedical Imaging and Bioengineering, the National Institute of General Medical Sciences, the National Library of Medicine (NLM), and the National Human Genome Research Institute, among many others. Currently, 667 active NIH research projects (for a total of $377 million) in the NIH RePORTER mention AI, ML, or DL.

The data science landscape at NIH is already rich and continues to expand. Large data sets being produced include the Genotype-Tissue Expression Project (200 TB), the Genomic Data Commons/TCGA (>4 PB), the Trans-Omics for Precision Medicine Project (approaching 15 PB), and the short-read archive and the database of Genotypes and Phenotypes (15 PB). Another NIH effort is the intramural Biowulf NIH super computer, which is number 88 among the 100 most powerful computers in the world and the first devoted strictly to biomedical applications.

Efforts in the data science space are ongoing to address challenges and create new opportunities. Many researchers invest a significant amount of time in accessing the data and making it analyzable; this aspect needs improvement. NIH Data Commons is establishing best practices to work with different data sets across cloud environments. NIH is also working with cloud providers to establish cost-effective deals for both intramural and extramural NIH staff and grantees. NIH leaders and experts must determine where limitations exist, whether in hardware or architecture. Recently, NIH released a Strategic Plan for Data Science to maximize the value of data generated through NIH-funded research.

NIH should find a balance between a bottom-up approach (which is the current predominant approach) and some form of top-down approach to encourage collective collaborative engagement, such as consortia, which has not been applied in this space. Incentives and shared resources could be employed.

Keynote Address: Mankind and Machines—Learning to Understand Human Biology, Together

Craig Mundie, President, Mundie & Associates

Previous societal evolutions were about 100 years apart until the computing revolution. Computing is changing society more rapidly, and it is harder to adapt. The evolution of computing has seen significant changes in architecture, such as tensor flow, and it is important to recognize and anticipate these changes. When considering advancements, we must be mindful of what will provide the most capacity. In the next 10 to 20 years, quantum machines will be part of the landscape, and it will be important to prepare and think about how this technology can be incorporated.

Machine intelligence has seen an evolution similar to that of computing. The advent of big data occurred outside the realm of biomedicine and was incorporated separately into biomedical fields. Big data sets, causal learning, and transfer learning are important for artificial general intelligence (AGI), but in some cases big machines are needed, but not big data. When AGI will be widespread is still up for question, but most prognostications about computer-driven advancement timelines have been inaccurate.

On the biomedical side, electronic data started first with EMRs and then arrived in other areas, such as imaging. People have worked hard to find ways to integrate EMR data, but Mr. Mundie is not convinced those are the correct data. He also acknowledged the importance of genome and proteome data but is less convinced of the importance of other "-omics." Other types of computing technologies that are currently coming from physicists but will be used in biomedical science include the following:

- Bayesian networks and inference
- Pearl-esque probabilistic causal learning
- Monte Carlo simulation and tree search
- Hypothesis-free, unsupervised DL
- High-scale modeling for prediction and forward simulation
- Quantum-inspired optimizations, including sampling, minimization, and training neural networks

Last fall, while performing an evaluation of candidate analytical tool sets, researchers inadvertently left off the metadata and reviewed only the raw proteomic data. The tool found an appropriate explanation that was simpler than what was currently known. Adding the metadata revealed a suitable answer that included the other answer, but the pathway was more convoluted. This suggests that scientists should have tools to allow the molecular data to speak for itself. This happened around the same time that Google's AlphaGo Zero results were published. In AlphaGo Zero, the group used tabular learning with no human data, examples, or interventions. It

recognized and rediscovered patterns. As it advanced, it understood information in current human knowledge and beyond. This approach can be applied in other places to discover new knowledge. In another recent example, a group at the University of California, Irvine enabled a computer to solve a Rubik's cube without human help based on Monte Carlo tree search techniques. Earlier this year, OpenAI was created as a way to develop an AI platform; it was a massive multiplayer online simulation and strategy game that used 256 GPUs and 128,000 CPU cores. In August, another team created a bot to play Dota 2, a multiplayer online game that involves international tournaments with professional players. The bot learned by playing a version of itself, and it has beaten professional players in one-on-one scenarios and will play in a multiplayer tournament. Professional players were interested in playing against the bot to train themselves. A similar scheme could be used in biological research.

Are machines here to help us, or is it our job to help machines? A recent *Harvard Business Review* article posited two tenets that may not be true: that humans must train machines and that humans must explain the outcomes of the tasks. Computers should be allowed to come up with the answer. Perhaps, instead, humans should be trained by machines, as in the case of Dota 2 players training against the bot.

People are interested in prescience about their own health. When applied, analysis of personalized biological data, including genomic and proteomic data, can help better identify the timeline and characteristics of human cancers. The future of medicine will likely start with individual data and personalized approaches, then synthesize cumulative information to establish trends for the population, not the other way around.

While some believe that it is best to collect as much data and data types as possible, it may be preferable to define and isolate the most useful data. What data are useful may well depend on the question being asked. A game platform could allow AI evolution to determine the relevant data. Some data may also be more accessible due to lack of privacy issues.

Life can be considered a tournament of multiplayer games, where winning is getting the most people to have the longest, highest-quality existence. Each game is unique. Players include Mother Nature, the environment, humans, and the bots helping humans. In this analogy, the genome is the rulebook, and the game board is a fixed 2D array of lots full of proteins. Humans can implement different elements of AI to biomedicine to help play the game.

Mr. Mundie provided answers to the questions that he posited at the beginning of his talk:

- Is human biology too complicated for humans to figure out? *Yes.*
- We are getting more data, but do we know how to map it and understand it? *Not really.*
- If it is too hard for humans, what about for machines? *Looks possible, even likely.*
- If they could figure it out, could they explain it to us in a useful way? *Unlikely.*
- How is that valuable? *In limited ways during the early stages, as a part of "unit testing" and to build confidence.*
- How would you go about it? *Use rapidly evolving computing advances with genomic and longitudinal proteomic assay to build a gaming platform for "the game of life."*

- How would it be different from current efforts? *Discard the existing history of pathways and interventions and focus on personalized data to understand the population.*

Researchers should act like gamers, using machines to improve their understanding and skill. Giving people a common toolkit and allowing competition is how many companies were started, and it might work for AI in biomedicine. Ultimately, humans will trust—rather than understand—what machines do, just as they do with the current computerized infrastructure in their lives.

Using AI will likely permit experimentation to be conducted at a lower cost and higher speed with powerful technology to study biological systems. Planning would require balancing investment and policy tradeoffs at a societal scale. The purity of data is important, which is the primary reason that EMR data are so challenging to use. Mr. Mundie believes that, ultimately, molecular data are more important. With AI, we can emphasize wellness and prevention over intervention and reduce healthcare costs. Machines might be able to decide which -omics provide the best answers to guide health care.

Current challenges include complexity in programs, systems, and networks, as well as physical and virtual interconnections. Emergent behaviors of hyperscale networks cannot yet be predicted. Approaches that help address these challenges include unit testing, formal specifications, and composability to help computers reason through the correctness of designs. Randomized controlled trials (RCTs) are still necessary, because while AI may be able to determine correlations, it cannot yet determine causation. Prospective data will likely be required for many years; a model never contains all the variables.

Building the game-of-life platform may be the next research grand challenge.

Applications of ML, AI, and DL in the Clinic and Community

Data Sharing
David Heckerman, M.D., Amazon (formerly Microsoft)

Unlike algorithms in AI that are an openly shared commodity, some biomedical data that would benefit from analyses with AI/ML are not always readily available. Data is at the heart of AI and ML, so it is important to expand it as a resource to continue innovation. There is not a lack of data; rather, some groups that own large amounts of data choose not to share it. It is important to understand why data are not shared and to determine ways to incentivize data sharing. Although some people say that data are not shared due to issues of legality and privacy, this can be overcome and is not a significant barrier to sharing. For example, 23andMe asked customers to share their data for research, and 80 percent of them agreed.

The motivations for a lack of data sharing affect different groups in different, but significant, ways:

- Doctors in private practice spend their careers collecting patient data and contact lists, which they sell for revenue, and they have come to rely on this money source.

- Researchers are under pressure to be productive to secure tenure and funding. They are likely to keep their collected data for fear that they might lose credit if the data were shared.
- Institutions seek the notoriety and positive public image of exclusive breakthroughs, as opposed to sharing credit with numerous other entities.

A Blue Ribbon Panel on Enhanced Data Sharing from a Cancer Moonshot [SM] Taskforce crafted a report of recommendations to motivate groups to provide early data availability.

Researchers can be encouraged to share data by—

- Giving more credit for data generation, curation, and analysis. This can include positive scoring during grant review, an s-index (for sharing) similar to the h-index as part of the grant evaluation process, and encouraging journals to publish separate author lists (as *Science* does).
- Implementing an NIH approval process for data sharing plans. NIH grant mechanisms currently include a data sharing requirement designed to incentivize sharing.
- Implementing stage gates within funding with a "no sharing, no payment" policy for continued funding.

Insurers' data and clinical laboratory tests can be made widely available by—

- Making data sharing a requirement for reimbursement, laboratory accreditation, or the streamlined review of laboratory test approvals
- Discouraging institutional review boards from introducing unintentional constraints for participants
- Encouraging payers to share data, which laboratory companies currently are more likely to do

Patient data sharing can be improved by—

- Requiring researchers to provide patients with their results in plain language
- Expanding the Genetic Information Nondiscrimination Act of 2008 to life insurance
- Providing standard consents for multiple situations to allow familiarization
- Giving patients ownership of their own data

Data owners could consider creating a marketplace in which metadata is available and can be searched by interested parties. A party interested in the actual data could purchase it through the market. In this scenario, all participants sign the same legal agreement, and the participating groups provide governance. In this scenario, if a group is afraid of missing opportunities from their data, they can increase its price.

In-Home Noninvasive Patient Monitoring
Dina Katabi, Ph.D., MIT Center for Wireless Networks and Mobile Computing

In-home wireless health monitoring technology that does not require people to wear any sensors or change their lifestyle is being created and can alert a caregivers if there is an emergency. Currently, home monitoring is difficult and, for some conditions, reliant on patient diaries. The new Wi-Fi device monitors breathing, sleep, heart rate, and gait speed through DL and the detection of electromagnetic fields.

A person's movement can be tracked in three dimensions to monitor falls. Gait speed is a health indicator for a variety of conditions: It is an endpoint for Parkinson's disease and multiple sclerosis, a predictor for exacerbations of congestive heart failure and chronic obstructive pulmonary disease, and a surrogate marker for cognitive impairment. Use of this device would allow full-time monitoring. Other advantages to monitoring movement include tracking toilet use, eating, and socialization behaviors. The range for detecting motion is approximately 40 feet.

Monitoring sleep is important for detecting sleep disorders and a variety of diseases. The device monitors brain wave changes similar to the monitoring done currently, but without the need for excessive nodes and wires that make sleep studies uncomfortable and may affect the results. Through machine learning, the device monitors sleep phases with an accuracy rate of 80 percent (similar to the 83 percent accuracy of sleep technicians reading patients' results).

The technology can also monitor breathing with 97 percent accuracy, compared with a chest band. The device is so sensitive that within the breathing pattern data, it can detect heartbeats. The range for detecting breathing and heart rate is around 13 feet. The device can also monitor two people in the same vicinity separately, and it can distinguish humans from pets. Although the technology can distinguish between people, it is not currently designed to work for a large number of people in a small space. Currently, most of the focus of this technology is supervised learning, but it does perform transfer learning to train the device in the laboratory and test it in the home.

With regard to concerns about security and privacy, participants consent to the use of the device, and identifiable information is stored separately from monitoring data; both are encrypted. The device has been designed to ensure that it cannot be used on someone without his or her consent, using instructional parameters at the time of setup.

Training for wireless in-home technology occurs in the laboratory, using small amounts of home data to facilitate transfer learning. People have not collected a lot of data from the home, so interpretation is still new, as is comparing that data to traditional clinical assessments. Laboratory data and data from homes are distinct. Wearables and invisibles can be used together to provide stronger types of data and information for improvement in care and to expand research, as the wireless device can only be used in the home, and the data from wearables tends to lack context.

Use of DL and AI in Radiology
Ronald Summers, M.D., Ph.D., NIH Clinical Center

Before 2012, there were a plethora of automated systems to detect abnormalities that relied on humans to annotate and measure the data. DL has helped solve intractable problems in the

classifier stage of image processing in radiology. Researchers have learned that handcrafted features (shape and thickness) are less important and that large data sets are more informative. As a result, more people can contribute to the field in a variety of ways, allowing a data democratization that has increased progress. Since 2017, there has been a profound increase in DL studies, of which 9 percent are in radiology. DL and ML have improved prediction of disease, although there are few examples of this in radiology compared with prediction by a trained physician.

DL has enabled accurate detection of pancreas segmentation in a way that was not possible before. There have been competitive challenges by professional societies to determine whether a cancer is malignant or benign; the winners used DL. Prediction of colon inflammation through ML has been successful.

Body-part recognition is also possible with ML. Segmentation labels can annotate representative areas that propagate through the entire data set. Diseases and conditions studied with this technology include muscular dystrophy, sarcopenia, obesity, and tumor growth, among others.

DL systems in the imaging and computer vision community are more tolerant of human error in labeling, or "noisy" labels. It is difficult to get a solid reference standard, so consensus points are used. Feedback loops of correction and refinement are used in some projects, such as a polyp detection project, to improve system performance.

Combining data mining reports and images can teach computers to read images to identify body parts and metastases. The identification improves with more descriptors. One complaint about ML is that the rationale for the decision that the computer makes is not clear, but people are creating saliency maps to provide context. ML for chest x-rays used eight terms with accuracy as high as 99 percent for large organs, but it did not work as well for identification of small nodules. Attempts to generate automated radiology reports from data are sometimes successful. Other data sets, such as a collection of measured and classified CT scans, allow sophisticated queries and attempts to detect lesions. Another system, Auto RECIST, is an automated lesion-measuring system with accuracy similar to that of expert radiologists.

Lymph node CT data sets, pancreas segmentation data, chest x-ray data, and a deep-lesion data set with more than 30,000 diverse CT scan images that are measured and classified are publicly available. Data sets like the CT scan set include a band of normal tissue around the lesion. Release of the entire CT scan was limited by data storage restrictions.

Images required in routine clinical practice are used to develop new labels with ML. A large number of clinical data scans that are not analyzed for aspects like body composition analysis are available. There is an opportunity to extract more information from existing data.

The future of ML/DL and radiology might include the following:

- New discoveries benefiting individual patients and populations of patients
- Routine integration of radiology with other clinical data
- Improvements in triaging and critical result monitoring

- Expansion of radiology imaging for global health
- More automation and quantitation

Artificial Intelligence and Machine Learning in Biomedical Research

Machine Learning Applications in Pediatrics
Judith Dexheimer, Ph.D., University of Cincinnati

There are more than 300 EMR vendors in the United States that hold patient population data, and most hospitals use EMRs. There is a tremendous amount of data in EMRs, and it is increasing with the inclusion of data from wearables, extra laboratory visit data, and patient comments. It is a potentially rich source for research and data mining, but up to 80 percent of it is unstructured. Quantitative or structured data include such information as vital signs, whereas qualitative or unstructured data include such documents as scanned clinical notes.

In the field of informatics, the person is augmented, not replaced, by the computer. Data are complex, and professional judgment will continue to be vital to shaping care. Capturing the data to enable patient care is important, but the systems are considered user friendly compared with other interfaces. In contrast to Instagram, which has one user type, EMRs must account for more than 100 types of users. There are also more numerous and complex types of data in EMRs. The features must be different, and there are more significant requirements for privacy, safety, and compliance.

NLP, a type of ML that attempts to train computers on natural languages, is used as a primer to create a data set and train a model (with supervision) to make a prediction. Challenges include syntactic ambiguity, speech patterns, negation, and data cleanliness. Some successes in NLP include NLM's MetaMap and the Clinical Text Analysis Knowledge Extraction System. Remaining challenges include lack of access to shared data, lack of annotated data and standards for annotations, lack of real-time implementation, and issues with data quality.

Pediatrics has distinct challenges due to variability based on development that must be taken into account and to differences in delivery of care. Caregivers also frequently perform treatments, and there are multiple historians for a medical history. Use of ML in monitoring sepsis during a pediatric intensive care unit transfer and monitoring appendicitis has increased, but ML is used less in pediatrics than in adult medicine. Main areas to apply ML include healthcare decision support (e.g., real-time patient identification for research studies in emergency department settings, patient surgery cancellation detection to help providers and hospitals), provider decision support, and integration of NLP with patient encounters.

NLP has been used for provider decision support to differentiate between intractable and tractable disease in epilepsy. Those patients who do not appear to respond to medications are eligible for a surgical consult. Using NLP to extract information from doctors' notes, researchers were able to develop a classifier that identified patients for a consult earlier than without ML.

NLP has also been used to monitor patient interactions and make predictions about a risk of suicide through verbal and nonverbal markers. This could be a way to identify people who might

be susceptible and who are not displaying obvious risk characteristics. In this case, audio data and video data of a patient–participant interaction were transcribed and annotated to train a classifier on text and linguistic features. Researchers are currently implementing this effort in the Cincinnati public schools through use of an iPhone app and counselor interviews of students to create a risk score. This study is still ongoing.

The future of this area of research likely includes NLP systems that are fully integrated with EMRs, real-time identification of patients with less processing time, smarter search algorithms, integration of information from multiple sources, and synthetic data. Voice-to-text software currently lacks accuracy, likely due to ambient background noise and because people do not speak the way they take notes.

Many ML models allow researchers to design trials more efficiently. RCT inclusion and exclusion criteria are intended to minimize variability, which may be good to avoid confounders but may limit the trials' relevance to the general population. Adjunct technologies help determine a balance. Simulation of clinical trials is meant to determine maximum efficiency and not replace RCTs.

AI for Healthcare Research: Next-Generation Healthcare
Eileen Koski, IBM Research

To some, it is not a question of whether data are good or bad, but rather what the data are good *for*. EMR data answer different questions to help understand health behaviors. AI in healthcare research is the convergence of three major drivers: need, data, and innovation. Healthcare costs are unsustainable, yet outcomes need to improve. EMRs contain a critical mass of rich health data, and major advances in data science allow handling of complex, diverse data at scale with progress toward multitask, multidomain, continuously adapting intelligence.

Opportunities with data include the ability to accrue real-world, population-level data in the form of images; -omics; and voice, test, sensor/device, video, patient-generated, social, environmental, and behavioral information. New types of data are consistently emerging.

There are many challenges for data, including access, privacy, security, scale, heterogeneity, semantic interoperability, harmonization, and data literacy. The concept of perfect data is a myth; instead, the focus is on understanding the optimal use of different data sets. The intent of the data and how it was created matter. In order to make a clinical decision based on data, one must be able to understand it and its intent.

IBM Research is involved in technology advancement in the fields of computational health, computational biology, devices, and technology at the intersection between life sciences and health care. Patient compliance is another frontier to empower patients toward positive behavioral changes. The four pillars of research at the MIT-IBM Watson AI Lab include AI algorithms (learning and reasoning), the physics of AI (analog AI and quantum computing), applications of AI to industries (health care and security), and advancing shared prosperity through AI (ethics and broad economic prosperity).

IBM has performed analytics on patient similarities to design precision cohorts. Patients who are similar in a clinically meaningful way can be used as a metric for ML and to guide treatment. It will be important to accurately predict who will get a disease, whether current prophylactic interventions are helpful or necessary, and the timing of disease development.

IBM has done predictive modeling in heart failure by working with healthcare centers to validate the data. Sometimes data are sparse, because a condition is rare or a procedure is unusual; it is important to apply the appropriate techniques and validate findings. Research with DL to predict occurrences of epileptic seizures is also ongoing.

Disease progression modeling has been done for Huntington's disease by using clinical fMRI and SNP data. IBM is exploring how to interpret the data to potentially predict the onset of symptoms. IBM has equipped a home with sensors to monitor the health and activity of patients with Parkinson's disease.

Another project, Psych-E, posits that speech is a virtual window to the mind. Asking even a benign question and analyzing the response can identify distinct speech patterns that may be diagnostic. This has shown to be promising in multiple languages and is not dependent on the content. This type of approach could be used in places like college health clinics, the college years being a common time for many affected people to have their first schizophrenic episode. Other programs are addressing healthcare access and helping doctors to practice better medicine.

Other projects are looking at behavior in terms of wellness, including managing stress through the use of wearable devices and offering suggestions for behavior modification. Not everyone is ready to make behavioral changes at a specific time, so behavioral phenotyping is also under study. This will assist care providers with messaging to patients and improved support for both patients and physicians.

Lastly, IBM is developing tools through its IBM Watson Health effort, which is developing products for people to purchase to assist with drug discovery, clinical trial matching, and many other matters.

Deciphering Genome Function with Functional Genomics and Machine Learning
Anshul Kundaje, Ph.D., Stanford University

Since 2003, when the first draft sequence of the human genome was established, there has been a revolution in technology for sequencing and interpretation to enable scaling to population-level sequencing, enabling identification of disease-associated genetic variants.

Most genetic variants are not harmful. Techniques such as genome-wide association studies (GWAS) allowed analysis of case-controlled studies to probe for variants that appear to associate strongly with a disease. One example is the identification of genetic variants in Alzheimer's disease. While helpful, these data are not sufficient because they do not elucidate mechanisms. That is the purpose of functional genomics.

Humans have one genome, but many cell types with differing phenotypes. One approach to mapping includes the biochemical profiling of cell type–specific elements. Only 1.5 percent of the genome encodes for genes; the rest is composed of control elements that fine-tune gene expression. Researchers can use functional genomics to create a digital map of precise locations in the genome that are modified or affected by biochemical modification. Many biochemical markers can be studied in one experiment. Unlike images that can require 200 layers of convolutional neural networks, regulator sequence predictions require three to five layers. The number of layers needed is done by trying a specific range, measuring performance, and optimizing the model. The model can then be used on an independent test set.

Two large projects, ENCODE and the NIH Roadmap Epigenomics Mapping Consortium, analyzed epigenomic data from hundreds of tissues and cell types, including 3 billion genetic coordinates and hundreds of biochemical measurements (e.g., chromatin modification). A 3D "data cube" is created that can be studied with ML probabilistic models and DL to identify tissue-specific control elements, learn the DNA sequence code of control elements, and interpret genetic variation.

ML can be used to "walk" across the genome to look at patterns and combinations for tissue-specific control elements. The combinations can be identified in latent states to infer how many control elements there are. Analysis predicts that there are 2 million control elements for 20,000 genes. It not clear, however, how many are functional. The control elements can be mapped, but the dynamics for how they are regulated in different tissue types must be inferred. Expression data can start to fill in these gaps. For example, expression of the *PAX5* gene in embryonic stem cells can be compared with other cell types. The data can be used to understand disease-associated genetic variation. This is particularly important, because 95 percent of complex disease-associated variants are in disrupted gene control elements.

ML can be used to review genetic variants in individuals in a specific cell type to understand a disease, such as Alzheimer's disease. Researchers initially reviewed the neuronal cells for genetic variants, but the genetic variants were enriched in microglial cells instead. Similar results suggest that there is a potential causal role for these cells to play in disease.

ML can also be used to understand the regulatory code of genomic control elements. If DNA is considered as a language, the control elements can be words that dictate a part of the control, and together they create a grammar that is distinct in different cells types to enable differential gene control. Researchers are trying to understand how the combinations of words give rise to meaning.

Patterns can be detected using the entire genetic sequence as input and defining the output as a biochemical marker. This creates a huge training set to use deep convolution neural networks to perform classification. In this model, there are no human-based assumptions about sequences' properties. The model scans the sequence, establishes a pattern, and builds layers to learn more complexity. Sufficiently complex patterns can be established to predict activity for a given sequence. The model can predict the nucleotide residue biochemical profiles with high accuracy. It can also predict binding profiles with strong correlations. Cumulatively, this can be done for every biochemical marker in every cell type through multitask learning.

The patterns learned by ML are as interesting as the predictions. The model learned thousands of known and novel patterns defining the regulatory code of tissue-specific control elements. Rather than summarizing patterns, a user can study a specific control element and use the model to interpret the sequence elements that are driving the prediction. It is also possible to create a dynamic profile to assess each nucleotide in the control element for its relative contribution. For example, in two active cell types, it is possible to interpret the drivers in the sequence and compare differences. This can be done across hundreds of cell types and validated by experimental data.

It is also possible to predict the molecular impact of genetic variants by simulating what would happen if part of the sequence were changed and seeing how that compares with other cell types. This had been done on a relatively large scale; scores can be created for variants based on their predicted significance for protein–DNA binding, allowing inference as to what the mutation is disrupting. Sometimes the variation or mutation can appear to affect binding in the same location as the variant, but sometimes it affects regions downstream of the mutation. In another example, one nucleotide mutation affects binding in the mutation region and binding downstream. In this way, researchers can analyze disease-associated variants and interpret their effects (e.g., for rheumatoid arthritis and multiple sclerosis). While GWAS cannot predict complex causal variants, this model can precisely predict a causal variant and how it is acting. Not all molecular effects are translated into disease effects. Large amounts of training data are needed, making it difficult to look for disease effects.

Variation is both the power and the challenge of genomics. The background structure of the genome has vast effects on the penetrance of a variant and can result in significantly different phenotypes. It can be so powerful, in fact, that genomic structure can mask otherwise fatal phenotypes. Although the idea of game design is abstract, it would be implemented at the level of the individual and the data would be aggregated. The knowledge base would include the phenotypic expression of everyone as it played the game. The ML framework exists; it simply has not been applied in the biomedical realm.

Fostering AI/ML in Biomedical Research at NIH

Group Discussions

NIH can and will play an important role in fostering AI/ML in biomedical research. Throughout the course of the workshop, speakers and attendees identified several barriers or challenges, as well as opportunities and existing resources. These are summarized here.

Challenges

Recruitment and training. Computer science has been fostered in factories and manufacturing. To accomplish a similar outcome in the biomedical sciences, there is a need for recruitment of promising computer scientists. NIH financial incentives are currently not competitive with other avenues for computer scientists.

Online ML training materials are inaccessible for biologists, and the reverse is also true.

Communication and collaboration. Some networks in this field exist, although perhaps not to the extent that is needed. If new networks are built, it is important that they can grow to include diverse microcommunities with multidisciplinary participants. To expand use of AI and ML, it is important that NIH identify opportunities for public-private partnerships. Some specific areas of communication warrant attention. For example, there is not enough communication between statisticians and ML experts. Peer-to-peer collaboration between computer scientists and biomedical researchers is also needed. Improved communications across disciplines will advance AI/ML in biomedical science.

Data. There is not enough biomedical data available for ML. *Data sharing* is a real and continuous problem that will affect enablement of AI and ML in biomedical research. We must understand the reasons data are not shared and incentivize sharing.

Incentives for *data harmonization* are lacking. Data are not always standardized, because the way a variable is defined (e.g., the normal range for an assay result) can differ. In the case of molecular assays, they are changing rapidly, and it is complicated to harmonize data from different platforms. Researchers must upload and prepare data and manage those costs, which can render the work unaffordable. Getting data in and out is a significant barrier. Most large DL endeavors so far have used data from large consortia to avoid spending months cleaning data. ML and functional genomics data need to be harmonized, prioritized, imputed, and integrated. We can create data cubes but cannot yet scale to create one per individual. Gaining meaning is a challenge. Harmonization will allow a centralized version control that can trigger better ML. Researchers need a better experimental gold standard or better catalogs for causal elements and variants. NIH should define who devotes resources to finalizing data sets.

Weaknesses related to *EMR data* exist. Some would like to see EMR data replaced with objective data; others note that EMR are a rich data source for specific information. Two weaknesses to EMRs were also mentioned: upcoding and treatment policies. In the former, a doctor will put in a code for a disease as a primary diagnosis over why the patient is in the office, because the doctor will make more money that way. This convolutes the data. For the latter, hospitalized patients with pre-existing conditions may get specialized treatment that improves outcomes, because the care regimen is more cautious. These are difficult challenges to overcome.

There is a lack of failure data, particularly in drug development, which could be a drawback for training the models.

NIH will need to bridge the gap between new technology and data and traditional data metrics and not overwhelm healthcare providers with excessive information.

Methods. ML and functional genomics methods are currently modular and isolated. Researchers are just starting to build a model that translates control elements to gene expressions and then to disease phenotypes to create a link of models from variants to phenotypes, end to end.

There is room for improvement in interpretation methods. New methods are needed for quantifying the uncertainty of predictions and interpretation.

AI and biases and inferences. Disparities, such as which patients show up for patient care and when, are inherent and exist before data are collected. In addition, some diseases disproportionally affect some populations, which can affect ML. In genomics, predictive models do not translate across different ancestry groups, and that diversification of data is important.

Biases can be reinforced with ML. One ML algorithm to assess which criminals might reoffend once released took into account factors that created bias, such as race.

ML approaches can make inferences with observational data but are sometimes wrong. There must be prospective analysis and randomization before a technology is embraced. Using ML as a complementary approach to human analyses and observations is still required. For example, a system for identifying breast cancer was used only as an initial screening.

Opportunities

Communication, collaboration, and engagement. Improving the understanding of other fields will increase appreciation for each expert's contribution to AI in biomedical research. We need training resources for biological experts to understand practical statistics, ML, and data science, as well as similarly complementary resources for computational scientists to understand abstracted aspects of modeling for biological data. Collaborators of complementary expertise could be paired as part of a positive career trajectory. More flexible collaboration and career transitions between academia and industry would facilitate more participation. NIH should foster more active collaborations among key players, such as between statisticians, geneticists and genomicists, biophysicists, and computational scientists within biomedical science, or between practicing physicians, patients, statisticians, ML researchers, and experimental biologists.

Gaming and prize competitions are good ways to engage AI and ML experts.

Incentivizing high-quality software and models as high-value commodities in academia, beyond merely publications, would foster research. One example of ready-to-use ML models for genomics can be found at http://kipoi.org. This allows users to replicate procedures from the models with less code. Researchers should be encouraged to use models for predictions, retaining, fine tuning, combining, and ultimately contributing them to the research field.

Training. Training opportunities and education foundations for this field are growing. There are many colleges working on ML in medicine. The annual American Medical Informatics Association (AMIA) Clinical Informatics Conference highlights advances. A basic understanding of writing code is needed, but DL does not require prior knowledge to do something useful.

Suggestions for acquiring the requisite background include graduate-level summer school courses, Coursera courses, familiarity with AMIA, and biomedicine-based computational

prediction challenges with the data already harmonized. Most teams are multidisciplinary, so it is not necessary to be an expert biologist to participate.

Data. NIH should encourage multi-institutional data collection to foster ML. Collaborations with groups like Kaiser Permanente would enable access to large amounts of data. Starting with emerging data is likely ideal, such as the *All of Us* Research Program. Incentives are needed to encourage data sharing. It would have to be clear to companies how their data sharing improves their business.

NIH's facilitation of computing and software in one location would accelerate research and help to avoid some of the lack of data harmonization that slows down research. Centralized locations and linking for data sets, software, and computation, like in the cloud, would allow conversion from "dry catalogs" to genomic knowledge bases that could allow smart searches, recommendations, and reasoning engines to empower discovery of relevant data sets.

Promising future aspects of AI include interpretation of information with complex structure, improvement of images, and assemblage of new data types. Combining medical imaging with genomic data and clinical test results will provide insights not allowed by using the data in isolation.

Methods. Frameworks and infrastructure for genomic ML already exist. Researchers build specific layers, or plug-ins, to adapt the framework to their needs. Efforts to build ways to rapidly transfer data into GPUs are ongoing. Modeling functional, large-scale interactions will provide more powerful models. Efforts to learn with less data should be considered. Public competition might establish more rigorous benchmarks and evaluation of models.

Newer models are providing explanations for the AI determination, not just the answers.

Soon models will become more precious than raw data. Creating model repositories will foster research and could include ready-to-use, trained models for regulatory genomics and other biomedical research efforts.

Advisory support. NIH should consider convening an advisory committee with AI/ML experts to provide recommendations on how to proceed with encouraging AI and ML in biomedical research. The NIH Chief Data Strategist will be a leader within the NIH Office of the Director.

Summary Remarks by the NIH Director

Francis Collins, M.D., Ph.D.

With regard to data sets, NIH should—

- Prioritize data sets, including their harmonization and cleanup, to allow ML.
- Enforce access to data from work that was supported by NIH. This is already NIH policy and is working well for genomic data. NIH is working on standards for other data types.

EMRs are an area of frustration, but they remain crucial for data related to medications, laboratories, diagnoses, and unstructured text.

Current NIH programs and areas where NIH can integrate AI include the following:

- BRAIN 2.0
- The *All of Us* Research Program
- Cancer genomics and therapeutics
- Environmental Influences on Childhood Health and Outcomes Program
- The Adolescent Brain Cognitive Development Study
- Model organism databases

Mechanisms that NIH might use to support AI include traditional R01s and other grant types, consortia and cooperative agreements, prizes, competitions, and games.

With regard to hardware, architecture matters more than FP64s, exascale computers may not be the best answer for biomedical AI, and forward planning should be put in place for the next big advance: quantum computing.

Needs related to training future researchers include the following:

- A workforce for biomedical applications of this technology.
- Expertise in AI/ML and biology; NIH should create environments for this training. Review of T32 structure may help determine whether NIH is providing the right training for AI/ML.
- Defining and nurturing career paths. Traditional academic tenure track may not be the right approach.

AMIA covers a segment of this growing community, but much of the biomedical AI/ML/DL community appears to be scattered. NIH should consider ways to foster and empower these researchers. Ideas to improve their support include—

- Better reward systems, including giving credit for software beyond publications and adjusting grant and tenure/hiring review policies to account for their contributions
- Nurturing more effective interactions and collaborations with statisticians, physicians, and basic scientists
- Formation of a new professional society

There is a continuing need to build the NIH "brain trust." Recruiting talented people from Silicon Valley would provide more depth for AI expertise. Messaging is important to convey the opportunities to work on interesting and important health problems. There may be ways to encourage select individuals to spend time with NIH researchers to understand our work. NIH is searching for a Chief Data Strategist to lead this type of effort. Lastly, it will be important to convene a group of visionaries to continue today's discussion; Dr. Collins is considering establishing a working group for the Advisory Council to the Director.