Data, Modeling, and Computation for Human Benefit

May 2022

Objective of this Document

Yale scholars across disciplines are using and advancing data science and computation to enable cutting-edge research, teaching, and applications that benefits human knowledge and society. This document surveys many of the efforts Yale is pursuing to build on its current and anticipated strengths in this broad field.

Context

The world is undergoing a data revolution comparable to the industrial revolution in its potential transformative impact. The confluence of pervasive data and computation promise to shape virtually all fields of endeavor. This benefit could be new perspectives on the fundamental knowledge that is core to the academic enterprise, for example: knowledge of science (determine how the brain works; search for exoplanets; simulate new materials), society (understand at a much higher degree of resolution the causes of economic growth), and the arts (uncover connections across disparate objects collections). It can result in tools that solve societal challenges, for example: problems of health (healthcare tailored to each individual through the combination of health record and genetic data), climate (harness the power of sensors to target carbon emissions), or public services (designing interventions in the education system that improve outcomes).

Academia and large swathes of the public sector recognize these opportunities. Every university has a major initiative in this area. Therefore, rather than attempting to survey the entire field, this document is selective; it identifies selected focus areas in which Yale feels it can deeply and uniquely contribute. The intellectual proximity and rich opportunities for knowledge transfer between efforts that develop foundational data science methods and tools with a wide range of disciplinary domains in which they can be utilized and extended is a unique feature of Yale. We structure these as three sets of related ideas, each rooted in an area of distinctive strength at Yale: (1) mathematics and the mathematical sciences, (2) social sciences, and (3) biology and the health sciences.

"Rooted in" does not imply that the efforts are limited to those disciplines, but rather identifies the source of Yale's unique potential to contribute. Each focus area is likely to connect with many, often unanticipated, areas of scholarship and discovery. In fact, this integration across disciplines is a key feature of scholarship in this realm, which benefits from and reinforces the connections across Yale's schools.

Further, each area includes *both* methods that can be called "fundamental" that arise from that area and "applications" of those fundamentals. As methods are often applied to fields other than those that inspire their development, our categorization here is not meant to be exhaustive.

This document introduces each of the three areas, and profiles focus topics in each.

Rooted in Mathematics and the Mathematical Sciences

The University Science Strategy Report clearly identified the need for deep mathematical understanding of data: methods of its analysis; mechanisms of its generation; and tools of its curation. Yale's deep strengths in the

mathematical and computational sciences have been of singular importance in creating an environment where new fast algorithms and tools for understanding signals and data can be developed for the broadest and deepest impact. These strengths have led to the development of recognized world class research in statistical foundations, computational modeling, and the computational systems that make modern computing on these data possible in ways that preserve privacy and security. We describe three critical areas anchored in these mathematical and computational strengths where Yale can lead at the forefront of transformational research.

Mathematical, Algorithmic and Statistical Foundations: *Kline Tower Institute for the Foundations of Data Science*

The avalanche of research in Data Science is the result of three key developments: the increased availability of data provided by the Internet and advances in sensor technology, advances in computer systems that make it easy to access and analyze vast amounts of data and to perform large computations in a reasonable amount of time, and the development of new methods of analyzing and detecting patterns in data. The development and analysis of these methods are central activities in the Mathematical, Algorithmic and Statistical foundations of Data Science. We strive to improve the classical methods of statistical inference, to better understand recent methods, and to develop new methods of analysis that satisfy unmet needs.

The fields of Machine Learning and Artificial Intelligence have made amazing advances through the development of Deep Learning and its offshoots. These are algorithmic advances in need of statistical and mathematical foundations. We seek a firm statistical understanding of these methods so that we know how much confidence to put in their results. We need to understand why these methods work as well as they do and when we should expect them to fail. Such an understanding of the nature of inference will be mathematical.

The methods of Data Science are typically abstract and may be implemented by many different types of algorithms. They may be improved by developing new algorithms that are faster or more accurate than those that came before. Some of these improvements will be general, while others will be just for the sorts of data we find in new applications or for newly developed hardware.

The Data Science revolution is just beginning: as new disciplines examine new data sources and consider new problems, they will discover a need for new methods and new domains of expertise. Researchers in the Foundations of Data Science will regularly interact with these disciplines so they can help provide methods that meet these needs.

The disciplines at Yale that use Data Science require more from methods than the industrial applications that gave birth to so much of the field. They need their analyses to be right. They often need to do more than detect and classify data: they require understanding of data and models of the processes that generate it. They need to dynamically predict what will happen if something is changed or an intervention is made. They require integration with causal inference and scientific modeling. Some will require entirely new ways of describing their data. The language of these is again mathematics.

In fact, most foundational research requires advances in mathematical understanding. The analysis of what happens under a model, or the proof that an algorithm is correct, or that a method works under reasonable assumptions all depend on mathematics, and much of that mathematics has yet to be developed. Research in fields including probability, analysis, combinatorics, and topology have already played an important role.

The departments of Statistics & Data Science, Computer Science, Electrical Engineering, and Mathematics all have faculty and students who are making important contributions to the foundations of Data Science.

The Yale Kline Tower Institute's mission is to support and foster the mathematical, algorithmic, and statistical foundations of this emergent field and applications in multiple domains. It will facilitate cross-disciplinary collaboration in a variety of ways: through organizational support and funding for faculty research, training to students and postdoctoral fellows, and by aiding partnerships with leading researchers and practitioners in the field. It will facilitate collaboration between researchers developing methods with those who employ those methods, and it will host workshops and programs that will make Yale a focus of the worldwide Data Science community.

Computational, Model-Driven Science and Engineering: *Proposed Applied Computational Science and Engineering effort*

An interrelated set of questions emerge as one considers scientific problems arising from systems with well-defined mathematical models. The capabilities of modern computer systems have enabled high-fidelity numerical simulations of complex models that complement traditional experiments, drive experimental design, and even serve as a new paradigm for data-driven discovery. This interplay has transformed many fields of science and engineering, driven by an integrative approach that combines computer simulations, experiments, and advanced instrumentation to drive the discovery process. There is a need for the creation of techniques for the co-development of computationally efficient algorithms, physically grounded models, and computer systems. With computation supplementing experimentation, trustworthy techniques and tools for modeling and interpreting data in a principled and reproducible manner are critical—there is a need for a new research framework for computation-driven science and engineering. With our historic strength in fast algorithms for numerical analysis and scientific computing, as well as wavelets, harmonic analysis, and diffusion geometry, Yale's Applied Mathematics and Physics programs provide world renowned strength on which to build such an effort.

While machine learning is currently very effectively deployed for science with the availability of copious data and computation, how these complex correlations are uncovered by algorithms can remain mysterious when the mechanisms of data generation are not completely understood. In many fields such as physics, for instance, the existence of known physical laws and symmetries in nature may hold the clue to illuminating machine learning methods. We are now in a unique situation wherein the well understood order in nature can be utilized to inform and help refine methods for inference. Machine learning algorithms can now be provided priors and models from rich domain expertise and knowledge across disciplines from astronomy, physics to chemistry to biology.

Incorporating well-understood laws and analytic models from science can help us make foundational advances in machine learning and AI which will in turn open up a new discovery space in science. A new symbiosis that builds on the interrelation between machine learning for science and uses of science to improve machine learning represents a new intellectual frontier. It is already clear that machine learning and AI will fundamentally transform how we do science itself, and Yale must embrace its unique position as an institution that fosters discovery across computational and scientific domains to open new pathways to foundational discoveries.

This new 'data-enabled' science and engineering, where data are constantly being analyzed, modeled, and used to revise experimental approaches, and where likewise the modeling and data analysis methods themselves are informed by experimental results, will drive scientific discovery in a manner that exhibits commonality between data and phenomena across domains. This kind of dynamic feedback loop is unprecedented in science and augurs new discovery pathways for breakthroughs.

To address these opportunities, efforts in computational science and engineering should accompany and complement more data-centric methods of AI and machine learning, leading to a virtuous cycle between data generating models and cross-validation with data-driven analyses with experimental data. This leads to a framework to advance an initiative in "modeling and AI for science and engineering."

To advance these methods and this broader discovery agenda, Yale should build an initiative bridging out from applied and computational mathematics to connect with this range of data enabled science and engineering domains. It is critical to build support structures (for faculty hiring and development, computing infrastructure, data analytic support, internal and external funding support, curriculum and training, innovation and entrepreneurship, and partnership with industry and national labs) for computational sciences and engineering more broadly. An emphasis on the role of trained staff modelers and programmers to drive the development of scientifically focused software will place Yale in a unique position to advance science and engineering. Strategic partnerships with industry are also vital, as industry is increasingly engaged in strategizing to address many scientific challenges with methods from AI (e.g., AlphaFold by the Google DeepMind team) and developing a variety of domain-agnostic tools for research and education (e.g., Google Collaboratory, TensorFlow) as well as computing infrastructures (e.g., Cloud Storage, CPU, GPU and TPU, Anton). Finally, these efforts can help us to reimagine a liberal arts education for the age of data and computation, in which the computational approach informs core problem-solving.

Trustworthy Computing, Privacy, and Cybersecurity: *proposed Center for Privacy, Accountability, Verification, and Economics of Blockchain Systems*

As the pervasiveness of data tools and methods increases, new systems and frameworks for distributed data and computation have emerged to ensure trustworthy, reliable computational systems, as well as ready access to critical data necessary for point of service or other insights. Systems for maintaining privacy of data as well as computing without revealing privacy have likewise emerged as a critical frontier in the management of cryptographic systems that encode data, whether about healthcare, financial markets, or other sensitive personal information.

Yale's Computer Science Department has historic leadership in consensus-based algorithms and their theoretical limits, as well as active world class research groups in cybersecurity, verifiable computing, and smart contracts. Recent recruiting has broadened this focus to include blockchain and cryptocurrencies, generating compelling new research. These systems and frameworks for secure and trustworthy computation have put society on the cusp of a new digital transformation (sometimes referred to as "Web3") in which vast distributed, consensus-based information networks allow for decentralized systems for sharing information, making inferences, and forming new social structures. The future of AI, machine learning, and large scale big-tech computing will involve and be transformed by these massive distributed systems and datasets, which at scale will be deployed through blockchain and increasingly consensus-based systems.

Yale currently supports core expertise in the mathematics and algorithms of these cryptographically based systems, as well as in the health, policy, and economic applications of such technologies. Whether in the management of secure and privacy-protecting healthcare databases, ensuring reliability and fidelity of voting systems, or development of new crypto-systems that serve identifiable and societally beneficial economic goals, Yale should lead in creating novel and fundamental research and teaching that furthers these applications.

Likewise, with quantum computing and the possibility of encryption breaking technology ever advancing (even on our own campus) new mathematical systems for quantum resistant encryption have become a vital and

active area of research, with various methods (lattice-based cryptosystems, notably) taking on a central position in modern cryptography research.

But more than data science, these ideas and computational systems represent a technological framework that relies on data, verifies its provenance, and extracts benefits based on its intrinsic value. As AI, machine learning, and algorithmically driven insights increasingly guide society, trustworthy computational methodologies and verifiable, consensus-based distributed systems will define a future of computing on which just, peaceful, and equitable societies can be built and advanced.

Rooted in the Social Sciences

Data science has a long history in social sciences: a significant portion of the history of statistics is the history of the systematic collection and analysis of demographic and economic data. Every advance in data collection has transformed the social sciences and related professional fields, such as the post-war explosion of national accounts and randomized surveys. Likewise, the social sciences have a storied tradition and outstanding intellectual community at Yale, with faculty and students at the leading edge of all core social science disciplines poised to push forward the use of data and computation in social science research, as well as to inform the societal aspects of understanding and shaping the impact of technology.

Enabling Research: *Data-Intensive Social Sciences Center*

Social science has long illuminated important issues and is well-positioned to offer new insights into the role of individuals, groups, institutions, and markets in social life. The availability of massive amounts of social and behavioral data, rapidly increasing computational power, and potent new methods for analyzing all sorts of data are transforming how many social scientists do their work. Digitalization of very large databases, the data-streams produced by social media and as a by-product of commercial transactions, and vast archives of text and images give researchers the capacity to achieve remarkable precision and texture in the description and analysis of the patterns of social behavior. Computational power and analytical methods permit the design and launching of research programs that would have seemed like science fiction just a few years ago. Social scientists are tackling large and important questions relevant to social problem-solving and finding new answers that are improving our understanding of the world and informing the design of more effective and equitable solutions to our problems.

Making the investments to allow faculty to take full intellectual advantage of rapidly developing analytical methods and providing the infrastructure for innovative data use will ensure that our social scientists remain at—and push forward—the research frontier. Therefore, Yale is establishing the Data-Intensive Social Science Center (DISSC) to support a wide range of research topics, including high-level statistical and computational consulting, assistance with data use agreements, and setting up secure computing environments pursuant to privacy and related concerns, among many others.

AI, Humans, and Society: *Computation and Society Initiative*

The University Science Strategy Committee notes that “not a single aspect of society today will be left untouched by the data revolution.” The SEAS Strategy proposes an even broader, campus-wide effort to address two questions: “How will AI systems change our lives, and how do we make sure that this change is for the better?” and “How do we ensure that computing and AI are optimally applied to solve real-world problems?” Here we take “AI systems” to include the myriad tools of the data and computation revolution, and we endorse the active role of humans in shaping, rather than merely reacting to, these systems for human good. At Yale this effort will proceed along three directions: (1) understanding and managing the societal impact of

technology, (2) reimagining society using the tools of the technological revolution, and (3) designing technologies that better achieve desired outcomes. These directions span the natural and social sciences, united by a desire to put human aims at the center of technological development.

The first direction is understanding and managing the societal impact of technology. Data science and related technological developments in computing and artificial intelligence are transforming the human environment and society. We are experiencing rapid change in communications, politics, work, and markets, with enormous consequences for psychological well-being, economic performance, social equality, identity, and governance. These are among the most pressing issues of the day, and engagement with these issues is both an intellectual challenge and, at the university level, a social responsibility.

The second direction is reimagining society using the tools of the technological revolution. The disjunction between the gradual evolution of institutions and practices, and the rapid change in technical feasibility, opens a space for radical reinvention of key social activities. Rethinking of status quo institutions and practices is possible in the wake of the technological revolution. Consider the enduring features of human society, such as educating the next generation, caring for the sick and the poor, organizing work and exchange, forming teams, choosing leaders, making decisions about the community, and sharing ideas. The ways that we perform these core activities have been built up slowly, over many generations, in response to changing values, local experimentation and accumulating experience, and technical constraints. We are now experiencing rapid, perhaps unprecedented, technological advances which dramatically relax the technical constraints on what is possible. The dimensions of this epochal change in technical possibilities are broad-reaching.

The third direction is a design aim, seeking to engineer the technologies so that they better achieve desired outcomes. “Outcomes” are defined as “use in the world,” rather than simply a set of technical specifications. This could relate to direction two, but it need not be limited to that. It can instead include how to make DMC technologies achieve better outcomes, in the real world of humans. Consider this another input into the design of those technologies. DMC can only benefit humanity if its products and discoveries are effectively implemented. While the work of implementation has often been considered the domain of business or industry, the end application must be carefully considered throughout the scientific process. Without this focus, critical discoveries may take years to be implemented or may never achieve their full potential. It is also essential that consumers (users) be well informed concerning strengths and limitations of these scientific products. Academia has a responsibility to carefully consider the end use and end user at all stages of investigation. Several basic questions pertain:

- What is the human, social, economic and political context in which a scientific product will be used and how do these factors influence the problem to be solved?
- What are available data inputs and resources to support the work? What is the cost/benefit of creating additional data resources?
- How will usefulness be assessed before implementation? What is the level and type of evidence required to justify implementation? What is the degree of certainty to the output? What factors most influence the output?
- Should randomized trials be required to determine effectiveness? If so, what are their basic requirements?
- What are the barriers to widespread adoption? How can we most effectively overcome individual and system-level economic, political, legal and other barriers to implementation?

- How will the psychological, sociological, political, and other impacts be assessed after implementation?

While there are many excellent examples of the importance of focusing on implementation from the beginning, prognostic modeling in health care illustrates some essential points. The advent of longitudinal electronic health records and images linked with large-scale genetic data combined with access to high performance computing promises to offer dramatic new insights for clinical decision making (i.e., choosing who is most likely to benefit from existing therapies) and for therapeutic discovery (i.e., identifying new therapeutic targets). However, most of the predictive models that have emerged from this work have not, and likely will not, be implemented because they have failed to adequately consider the questions listed above.

Information, Algorithms, and Society

With the arrival of the internet, and more broadly with rapid increases in the capacity to transmit, communicate and process decentralized information among a large population, information has become a central object of interest in computer science, data science and economics. Digital information in particular is central for the allocation and distribution of services and commodities society wide. The question of how to collect, aggregate, and disseminate information among many decentralized and heterogeneous individuals is critical for the functioning of democracy, as the discussion surrounding social networks showed recently, and for the functioning of economies, as the continued search for fair and efficient financial markets shows.

A complete understanding of information clearly requires computer science and information theory; is supported by central statistical insight; and requires the analytic tools of economics to understand the incentives to share and produce information. The key question that underpins the use of information is its value and that of the databases in which it is stored. These questions have found a remarkable convergence in these three fields, yet we are just beginning to grapple with these foundational questions.

Yale builds on a storied history in Economics (in particular, market design) and intersections with growing faculty strength in algorithms, machine learning/optimization and causal inference and ML in Computer Science, Statistics and Data Science and Political Science. The initiative also draws on policy and law related efforts at the Law School, as well as the Operations Research and Economics groups at the School of Management.

Causal Inference

From the explosion of data in the digital economy and social media, to the digitalization and merging of massive datasets from traditional data sources, to the new forms of data such as text, image, and sound, there is a new sea of data available to describe the world in previously unimaginable detail. But beyond advances in description, how are these quantities related to each other? What are the causal linkages? The technology and cost of gathering data has changed dramatically. What new research designs are now possible and cost effective? Businesses conduct thousands of A/B experiments every hour and algorithms create discontinuous breaks in the environment experienced by customers and employees. How can these experiments and natural experiments be identified and analyzed to advance our theoretical and empirical knowledge of human behavior?

Likewise, some hard sciences such as astrophysics and biochemistry also strive to extract scientific understanding, that is inherently causal in nature, from applications of machine learning techniques to scientific datasets. These efforts can be accelerated through interface and intellectual exchange with those in the social sciences pursuing similar core concepts.

Measuring causal effects and establishing causation is a central challenge for the social and behavioral sciences. The modeling techniques and statistical methods for extracting causal relationships from non-experimental data are among the most important tools of modern social science inquiry. The design of randomized experiments is a foundational method across the social sciences, medicine, and policy design in government, business, and the non-profit sector. A university-wide effort to advance the study of causality would engage faculty across the university.

The study of causality, both basic research and applications, is a vibrant area of research. Recent decades have seen the dramatic increase of experimental methods and non-experimental research designs that focused on the assumptions necessary for producing credible causal estimates. These efforts are seen in the publication patterns in leading academic journals and recognized by awards including Nobel prizes.

A Yale Causality Initiative would promote research on causality. Topics for research might include:

- The use of machine learning methods to measure causal effects, including incorporation of context-specific theoretical constraints on model selection and the detection of different treatment effects across demographic and other subject groups.
- Theoretical analysis of experimental design to improve efficiency and reduce sources of bias, including analysis and discovery of methods for creating balanced experimental groups, use of repeated measurements, adaptive designs, and the analysis of treatment effects in complex network structures.
- New technology for measuring of interventions and outcomes, including use of wearables for health and other measurements, geo-location and surveying from mobile devices, remote sensing, internet-based surveys.
- The study of Human-Evidence Interaction (HEI), including how to create experimental designs that are not only scientifically rigorous but persuasive and useful to practitioners and policy makers and to understand the cognitive biases that affect how people evaluate the evidentiary value of alternative research designs.

Yale is well-positioned to contribute to the development of both new methods and innovative applications. The new Data Intensive Social Science Center is designed to provide research frontier infrastructure for data-intensive social science research across the university and to generate awareness of new data and new applications. The Cowles Foundation is a pioneer in the application of mathematics to economics and supports one of the world's premier econometrics groups. The Economic Growth Center is a leading center for the study of developing economies. Affiliated faculty conduct field experimental research and conduct household surveys around the world, including in environments with complex logistical challenges to data gathering. The Institution for Social and Policy Studies pioneered the modern use of randomized experimental trials in political science and ISPS affiliated faculty include leading scholars in the application of experimental methods to the study of political behavior and to other core topics in political science. Paired with Yale's outstanding schools and their many centers, Yale has both research excellence and extraordinary opportunities to produce research frontier applications that combine domain specific knowledge with advanced research designs.

Rooted in Biology and Health Sciences

Yale has long been at the forefront of biological, biomedical, and public health research. Much of this work is based on the collection, analysis, and interpretation of complex data in various domains (genomics and epigenomics, single cell, exposome, biomarker, imaging, mobile device, health statistics, health system, etc.) that can be sparse, incomplete, noisy, and/or multi-dimensional and whose interpretation requires development and application of novel data science methods and tools. At Yale, the large volume of clinical and research data, strong collaboration between biology and medicine, as well as close connections between scholars developing foundational techniques and those generating data in their research provide fertile ground for advances in both the technology and applications of computation and data science in both the lab and in the field.

Invest in Biomedical Data Science Faculty: *Section of Biomedical Informatics and Data Science*

Today the study of biology, medicine, and public health are increasingly driven by large datasets across scales, from single cells and molecules, to tissues, individual patients, and population studies with millions of participants. The data collected are diverse, including those from different omics technologies (genetics, genomics, epigenetics, proteomics, metabolomics, and others), imaging instruments, health records from health care providers and insurance companies, wearable devices, sensors, the exposures of an individual, social media and network information. These data offer great opportunities and challenges to studying basic biological processes, disease etiologies, drug development, patient treatment, public health, and health care deliveries and policies.

In response to the need of biomedical data science for both basic and clinical research, the Yale Center for Biomedical Data Science (YCBDS) (https://medicine.yale.edu/cbds/) was established in 2018 as a focus for research and education in biomedical data science at Yale. Since its inception, YCBDS has offered training workshops on different data science topics, developed seminar series, connected researchers from different disciplines, and established alliances with industry partners, such as the Yale-Boehringer Ingelheim Biomedical Data Science Fellowship Program. Currently, the biomedical data science community at Yale includes researchers from different schools (such as School of Medicine, School of Public Health, Faculty of Arts and Science, School of Engineering and Applied Science) located at the Medical School Campus, Main Campus and West Campus. With the ongoing recruitment of the Senior Associate Dean for Biomedical for Biomedical Informatics and the establishment of the section of Biomedical Informatics and Data Science at the School of Medicine, the biomedical data science community will likely see a rapid growth in the coming years, and this presents an excellent opportunity to coordinate different research and teaching programs and connect with researchers focusing on foundational aspects of data science research.

Human and Machine Intelligence: *Wu Tsai Institute for Neuroscience - Center for Neurocomputation and Machine Intelligence*

The goal of Yale's Wu Tsai Institute (WTI) is to understand human cognition. WTI explores fundamental processes of human cognition by bringing together scientists who study the psychological properties of the mind with those who study the biological properties of the brain. It bridges these often-separated branches of neuroscience through cutting-edge technologies and the common language of data science, providing a computational foundation for building new models and theories of cognition.

The Center Neurocomputation and Machine Intelligence is one of three interdisciplinary centers at WTI. The Center's vision is that fundamental advances in the understanding of human cognition will both require and enable new computational frameworks for machine intelligence. Its focus, therefore, will be to discover next-generation algorithms and models that reflect the mechanisms underlying intelligent behavior in humans and other animals. While the goal is to better understand human cognition, we expect that insights from neuroscience will ultimately lead to advances in artificial intelligence that will benefit a range of applications. The Center will also develop resources to help Yale scientists make sense of the large datasets that are being generated in neuroscience research across a spectrum of instruments, spatial-temporal scales, and species.

The Center will host seminars and events, manage computational coworking space and interface with departments across Yale's campus. With its integrative nature, the Center will bring scientists trained in different areas together to communicate and problem solve together. It will also build a core facility for advanced computation, data visualization, and data management to support the WTI's mission.