Data Science and Applied AI Postdoctoral Scholars Program

Candidates:	Applicants must hold a PhD in computer science, statistics, or a related field by the start date of the program
Application Window:	November 1, 2019 - January 31, 2020
Applicants Notified:	Applications will remain open until the positions are filled
Earliest Start Date:	July 1, 2020
Appointments:	For up to three years, renewed annually
Apply Online:	https://uchicago.inforeadv4.com

The Center for Data and Computing and the Center for Applied AI at the University of Chicago seek applications for Postdoctoral Scholars who wish to deepen their knowledge of cutting-edge data science and computing research while developing additional expertise in a specific, applied problem domain. Through an innovative partnership, postdoctoral scholars will have access to resources at two data science and artificial intelligence research centers:

- The Center for Data and Computing (CDAC) is the intellectual hub and incubator for data science research at the University of Chicago. Co-located with the University of Chicago Computer Science Department, we catalyze discoveries by exploring new data and computing methods, foundations, and platforms in the context of real-world applications.
- The Center for Applied AI at the Chicago Booth School of Business is committed to creating revolutionary advances in applications of AI through groundbreaking, interdisciplinary research. We bring together researchers, professionals, and industry leaders to propel AI's evolution into new frontiers by creating tools that respond directly to real-world problems across a diverse array of fields.

This unique program provides postdocs with the opportunity to pursue original research on significant questions in data science, while also developing specialized domain expertise in one or more complementary areas such as behavioral science, healthcare, and public policy. Drawing on the University of Chicago's top-ranked programs, world-renowned faculty, as well as a vibrant and quickly expanding data science ecosystem, this program will allow postdoctoral scholars to engage in field-defining data science and artificial intelligence research. Our positions carry a competitive salary, generous research funding stipends, and benefits.

As part of our mission to catalyze a dynamic, multidisciplinary data science community, we actively recruit, support, and mentor scholars from all genders, ethnicities, and backgrounds. Diversity of thought, experience, and background are essential for the generation of new ideas, and we aim to build an inclusive environment where all voices are heard, respected, and considered.

Program Benefits:

- **Joint Mentorship:** Scholars will receive joint mentorship from both a data science researcher and a domain expert. This mentorship will provide postdoctoral scholars with multiple perspectives on their research and career guidance. Mentors will provide

ongoing evaluation and advice through regular meetings, as well as opportunities to promote the scholar's accomplishments in public forums.

- **Interdisciplinary Training & Collaboration:** To extend and deepen scientific, technical, and communication skills, scholars will gain broad exposure to diverse research fields with up to 50% collaborative work on cutting-edge research projects.
- **Independent Research:** Scholars will have the freedom to pursue their own research interests with up to 50% work on independent projects and no teaching responsibilities.
- **Unique Datasets:** Scholars will have privileged, unique access to large-scale datasets from a variety of sectors including urban studies, marketing, medicine, science, computing and communications, and economics.
- **Cohort Program:** The program will host weekly seminars where scholars can connect with members of their cohort, share knowledge, and gain insight through guest lectures, industry speakers, and other activities. Scholars will have autonomy and budget to select, host, and invite speakers, with support from CDAC administrative staff.
- **Outreach and Impact:** Scholars will have considerable opportunities to establish new relationships and translate their research into real world impact by leveraging our network of academic, civic, government, and industry connections.
- **Academia/Industry Ready:** Experience gained during the program will help scholars prepare for diverse career paths from tenure-track academic positions to leadership opportunities within innovative companies.

Successful Applicants Will:

- Hold a PhD in computer science, statistics or a related field by the start date of the scholarship.

Applications Should Include:

- Curriculum vitae;
- A one-paragraph summary of the candidate's current research;
- Research statement that outlines research goals, potential projects of interest, and motivation for seeking a postdoctoral appointment at UChicago (maximum of 3 pages);
- 1-2 representative publications
- Three letters of reference;
- Names of potential UChicago faculty mentors;
- (Optional) Applicants are encouraged to include a letter of collaboration from a UChicago faculty mentor who has agreed to mentor the applicant if the scholar is accepted into the program. Please use the following template for the letter:
  - "If Dr. [insert full name of applicant] is accepted as a **Data Science and Applied AI Postdoctoral Scholar** at the University of Chicago, it is my intent to act as a mentor on a project of mutual interest."

The CDAC team will also be available to help identify potential faculty mentors if you move forward in the application process. As examples of faculty who may be available as mentors, please see the list of projects and mentors on the following pages and the CDAC website.

Contact:
For questions about this application, please contact cdac@uchicago.edu.

Examples of Projects & Mentors

Sections:

Scholars will have the freedom to pursue their own research interests, as well as working on collaborative projects in cutting-edge research areas. Below we list examples of projects and faculty mentors (non-exhaustive). If you are interested in working on a particular project or mentor, please indicate the area(s) and mentor name within your application.

Business, Behavioral Science & Economics

Creating a Recommender System for Investor Portfolios

"Traditional methods to understand financial markets use data on firms' fundamentals and asset prices. But the asset prices we observe reflect the aggregate behavior of individual investors, both institutional and retail investors. In this project, you would be helping to create models of investors' behavior to explain current, and predict future, asset prices. This framework can be used to evaluate financial market regulations, to estimate the impact of fiscal and monetary policies, and to improve portfolio management decisions."
Mentor: Ralph Koijen, AQR Capital Management Professor of Finance and Fama Faculty Fellow, Booth School of Business

Understanding Signatures of Interactive Human Communication

Our lab is working on several projects using NLP and textual analysis to answer questions concerning fundamental human behavior in communication and learning. For example, we are attempting to define a signature that could identify an example of dialogue from that of debate. Similarly, we are interested in understanding whether the language used for giving reasons versus rationalizations could be identified. Lastly, we are investigating the differences in learning when done for oneself versus someone else. An example of this is looking at knowledge acquisition followed by explanation to another versus acquisition followed by regurgitation. You would have the opportunity to develop skills related to conducting experiments, to help develop a project from concept through to analysis, and to share expertise in NLP with the lab team.
Mentor: Jane Risen, Professor of Behavioral Science and John E. Jeuck Faculty Fellow, Booth School of Business

Energy & Environment

Data-driven Environmental Enforcement

The Energy and Environment Lab invites a postdoc to collaborate on a suite of projects that leverage advances in monitoring technology and machine learning approaches to inform environmental policy, under the mentorship of *Michael Greenstone*, the Milton Friedman Distinguished Service Professor in Economics, the College, and the Harris School; Director of the Energy and Environment Lab, the Becker Friedman Institute, and the Energy Policy Institute at Chicago.

Congestion and Traffic Safety

Many cities across the United States have adopted Vision Zero, the policy goal of eliminating all traffic-related deaths and serious injuries. But what are the costs of achieving Vision Zero and what are the most efficient policy instruments to get there? Monitoring technologies offer the potential to revolutionize urban policy by providing governments with big data to inform policymaking. As part of NYC's Vision Zero, the Department of Transportation is more than quintupling the number of speed cameras in the city. Leveraging our access to unique data of taxi, for-hire vehicle, and city fleet trips to model the impacts of new traffic cameras on vehicle crashes, slowdowns, and congestion spillovers. The post-doc would utilize large administrative datasets on camera enforcement, vehicle crashes, segment-level traffic speeds, and high-resolution driver behavior, to help measure the costs and benefits of enforcement strategies for fatality/injury reduction to inform optimal policy for urban traffic safety.

Leveraging Satellite Data to Reduce Oil & Gas Methane Emissions

The meteoritic rise of shale oil and gas (O&G) drilling in the United States poses significant challenges for reducing greenhouse gas emissions. The methane emitted has around 30 times greater short-term global warming potential than CO2, contributing aggressively to climate change. Reliable estimates of emitted methane are essential to fully understand and mitigate the environmental threat presented by shale drilling. While some estimates suggest that approximately 2.3% of gross natural gas production is leaked per year, accurate monitoring of emissions remains extremely challenging. Currently, regulators visit individual facilities to measure emissions; but due to budgetary constraints and a fast-growing industry inspector can visit only a fraction of the facilities each year. This project will leverage a wealth of administrative data and novel remote sensing data from recently-launched satellites to estimate facility-level methane emissions. Leveraging these unique data and state-of-the-art machine learning techniques, the project will help regulators re-design their monitoring and enforcement strategy to realize improvements in regulatory efficiency and reductions in greenhouse gases.

Beyond Inspection Targeting: Deterrence through Machine Learning

Building on a three-year partnership with the Environmental Protection Agency (EPA), this project aims to scale a machine learning-driven framework across inspection targeting programs at EPA. The Clean Water Act (CWA) is one program where data-driven inspection targeting can directly influence environmental policy. Using state-of-the-art machine learning models, we can generate risk scores for the likelihood individual firms will violate CWA standards, and use these model-generated risk scores to study facility compliance behavior and identify the most effective approaches to deterrence in a randomized field trial.

Mentor: Michael Greenstone, Milton Friedman Distinguished Service Professor in Economics, the College, and the Harris School, University of Chicago; Director, Becker Friedman Institute for Research in Economics; Director, Energy Policy Institute at the University of Chicago (EPIC); Director, Tata Center for Development at the University of Chicago

Foundations of Data Science

Data Markets and the Economics of Data

Data has been called the new oil, and it is certainly true that the sheer volume, variety, and velocity of data being produced and stored is generating enormous value for the individuals and organizations that know how to tap into and refine it. But data is not just another commodity inhabiting an economic and social system; it has given rise to an entirely new economy and

society. This research project will investigate theoretical, empirical, and technological foundations of the new data and artificial intelligence economy comprising emerging data markets, data integration and the transformational services that fund these changes. Sub-projects include research into: pricing models for data, architecture of data sharing platforms, data discovery, and data provenance and integration.

Mentor: Michael Franklin, Liew Family Chairman, Department of Computer Science
Sr. Advisor to the Provost for Computation and Data Science

Machine Learning for Physical Systems

Much of machine learning is focused on recognizing patterns and making predictions based on training data. However, in many physical science settings, the complexity of the task is too high for effective learning giving the amount of available data. In these settings, it is essential to incorporate knowledge of the underlying physical system to mitigate the effect of limited data. Examples include using a combination of training data and models of a CT scanners operation to develop better medical image reconstruction methods, leveraging both observational and simulated data to develop better climate predictions, and building deep learning-based surrogate models for computationally demanding PDE-based simulators of physical systems. While there are isolated examples of successes in these regimes, little is known on a fundamental level. What are optimal machine learning methods that leverage both training data and physical models? How does sample complexity scale with the type of physical system and the accuracy of our models? Which kinds of PDE models are most amenable to deep surrogate models? This project will focus on developing new methodology and theory for machine learning for physical system that will address these and other open problems.

Mentor: Rebecca Willett, Professor, Statistics, Computer Science, and the College

Operational Analytics for Communications Systems

Many applications of machine learning to computer communications systems such as the Internet rely on models that are trained offline, on snapshots of data. Yet, in many operational systems, data arrives as a continuous stream—often as a timeseries, and decisions must be made on short timescales (e.g., milliseconds). In operational systems, designers must face difficult challenges and design tradeoffs concerning encoding of timeseries data, efficiently labeling large quantities of data, distinguishing anomalous activity from model drift, and trading off model accuracy versus model or feature complexity. In this research project, we will explore these challenges in the context of networked systems. Possible avenues include device identification and anomaly detection in industrial control systems and consumer “IoT” smart homes; streaming video quality estimation; content moderation on social media platforms (e.g., Facebook); and security vulnerability detection in networked systems.
Mentor: Nick Feamster, Neubauer Professor of Computer Science; Director, Center for Data and Computing

Medicine & Health

Creating Personalized Incentives to Drive Diabetes Patients' Behavior

Physical exercise can product a significant health benefit for diabetics, but not all patients have the same natural inclinations for exercise. We are working to develop a heterogenous treatment model that would create individualized incentives to help diabetes patients succeed at an

exercise regimen. Come have an outsized impact on this project in its early stages as we formulate initial data needs and begin sourcing additional measures with which to create our model.
Mentor: Rebecca Dizon-Ross, Associate Professor of Economics and Charles E. Merrill Faculty Scholar, Booth School of Business

Early Childhood Metric Initiative

Early childhood suffers from a lack of quantifiable data that is easy to collect at scale. Work with a team of engineers, computer scientists, and early childhood experts to build a non-intrusive, wearable technology that leverages machine-learning to collect real-time, real-world data to measure young children's early language environments (i.e. the quantity and quality of language interactions they are exposed to). As part of the project, develop machine learning algorithms and models to analyze the large-scale adult-child interaction data collected from this wearable technology. This large dataset will enable researchers and practitioners to better understand the relationships between family demographics, parental inputs, and child outcomes and identify effective approaches, as well as enable policymakers to hone in on the most effectives programs and policies to enhance children's early language environments. This audio dataset will also allow experts in natural language processing to develop and refine speech processing algorithms.
Mentor: Dana Suskind, MD, Professor of Surgery and Pediatrics, Director, Pediatric Cochlear Implantation Program, Co-Director, Thirty Million Words (TMW) Center for Early Learning + Public Health

Nightingale Project

Machine learning, we are told, will transform medical diagnosis and patient care: by integrating 'big data' on patients' history and physiology, algorithms can dramatically improve the quality of doctors' decisions, with the potential both to reduce waste, avoid misdiagnosis, and produce breakthrough discoveries. For example, if massive datasets of ECG waveforms could be linked to national mortality registries, we could supercharge the current research, and find better, more consistent ways to allocate life-saving defibrillators. But most clinical data like this is siloed by different institutions and unavailable to researchers. Further complicating things, in order to protect patient privacy, public medical datasets are almost universally limited to a single, easily de-identified stream of information, like a set of X-rays.

The goal of the Nightingale Project in the Booth Center for Applied AI is to gather and share just the sort of rich, multidimensional data needed to feed AI-enabled discovery. Work with a team of engineers, data analysts, and medical experts to build a secure platform that can warehouse curated de-identified clinical datasets linked to ground truth outcomes. Using this initiative as a proof of concept to develop other privacy tools, own and work through the de-identification, sharing, and privacy components of large-scale datasets -- work that will include writing models and creating security-related challenges -- with the goal of making the data available to researchers securely.
Mentor: Sendhil Mullainathan, Faculty Director, Center for Applied AI, Roman Family University Professor of Computation and Behavioral Science, Chicago Booth

Pediatric Cancer Data Commons
Collecting, aggregating, harmonizing, and sharing data from children with cancer is essential to making new discoveries and developing new cures. Too often, data are siloed and disconnected, drastically reducing the usefulness of these valuable resources. The Pediatric Cancer Data Commons (PCDC) at UChicago brings together researchers from around the world with the goal of building data dictionaries for all types of pediatric cancer. Consensus data models are balloted with experts from around the world, including clinicians, ontologists/taxonomists, statisticians, and data scientists. The resulting dictionary is used for harmonizing data from completed clinical trials and is subsequently leveraged as a framework for collecting data on new studies. The data are made available to the worldwide research community through a public-facing cohort discovery tool. Data are further connected to other sources through common identifiers, allowing novel new data sets to be developed for research and discovery.

Potential areas of research include: ontology development and data dictionary creation, data harmonization, automated methods of metadata extraction and data ingestion, development and deployment of novel data visualization tools and analytics, data governance and provenance methods and tools, developing novel methods of combining disparate data sets, and developing analytic methods for new modes of risk stratification. Experience with clinical data is preferred but not required.
Mentor: Samuel L. Volchenboum, Associate Professor of Pediatrics & Associate Chief Research Informatics Officer, UChicago Medicine

Public Policy and Society
Combining human and machine intelligence for policy impact
The success of artificial intelligence (AI) for engineering and commercial applications has led to growing interest in using these tools to help solve important social problems like inequality in income, education, health, or criminal justice system involvement. But any realistic assessment of how AI will be used in these areas suggests it will be a complement to, not substitute for, human judgment. That is, AI will be used as decision aids, not decision makers. In previous work (Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan, 2018 Quarterly Journal of Economics) we have found in the context of criminal justice system decision-making that humans on net add negative value to the machine's predictions of defendant risk, although in principle the private information humans can access that algorithms cannot (such as courtroom discussion about the details of the case) could help the human add positive value in at least some cases. Similar issues arise in numerous other policy domains such as medical diagnosis, hiring, credit, and education admissions. The goal of this project is to better understand the potential sources of human and machine comparative advantage by measuring the private information humans have access to in different decision-making domains, trying to understand what are useful sources of signal versus sources of noise for human decisions about when to follow versus over-ride the algorithm's recommendations, and then try to build decision-making systems that lead to the human plus machine together to outperform the decisions implied by the machine's predictions alone.
Mentor: Jens Ludwig, Edwin A. and Betty L. Bergman Distinguished Service Professor, Harris School of Public Policy, Director of University of Chicago Crime Lab, Co-director Education Lab

Corporate influence on the rulemaking process within the United States

The rulemaking process in the United States includes an opportunity for public comment in between a new regulation and implementation of a rule change. During this comment period, not only the public at large, but corporations as well are able to exert influence over rule changes. Using textual analysis techniques, we aim to understand how this process of influence works – to what extent corporate influence affects rule changes and what kinds of changes ultimately result.
Mentor: Marianne Bertrand, Chris P. Dialynas Distinguished Service Professor of Economics and Willard Graham Faculty Scholar, Booth School of Business

Preventing violent encounters with first responders

People who live with serious mental illness or related challenges face heightened risks of violent encounters with first responders. Administrative and qualitative data from police, fire, and other first responders allow us to identify individuals, places, and events associated with such violent encounters. This project will use predictive analytics to improve preventive services and emergency responses for individuals and families who face these risks.
Mentor: Harold Pollack, Helen Ross Professor of Social Service Administration, Co-director, University of Chicago Health Lab

Predicting mortality among high-users of safety-net services in Illinois

Individuals who pass through jails, homeless services, and other safety-net institutions face severe risks of premature mortality from opioid overdose, homicide, and other causes. This machine learning project uses integrated administrative data from diverse city, county, and state data sources in Illinois to identify key risk-factors for premature mortality.
Mentor: Harold Pollack**,** Helen Ross Professor of Social Service Administration**,** Co-director, University of Chicago Health Lab

News-based Sentiment Analysis to Understand Market Movement

Most financial analysis is quantitative, but there is a wealth of data contained in textual artifacts as well. In this project, we are using natural language processing methodologies to conduct sentiment analysis on global news reports in order to better understand how news-based sentiment affects the movement of markets. Other possible avenues of investigation using new NLP methodologies include understanding the macro effects of global economic sentiment and attempting to detect the existence of fake news.
Mentor: Dacheng Xiu, Professor of Econometrics and Statistics, Booth School of Business

Understanding the Effects of Gender on Policy Through Textual Analysis of Congressional Data

Congress generates a substantial amount of textual data from its hearings, meetings, speeches, etc. By conducting sentiment analysis on this data, we are trying to understand how gender influences the likelihood of member participation and the resulting policy decisions. Currently, this project has preliminary results and is entering a second phase of analysis, which will involve additional scraping of data and the generation of new methodologies for textual analysis.
Mentor: Heather Sarsons, Assistant Professor of Economics and Diane Swonk Faculty Fellow, Booth School of Business