**UC San Diego**
**Division of Social Sciences**
**Defining the Interdisciplinary Future of Data Science**

## Defining Data Science

Data Science is an emerging, independent discipline. Consequently, the field is currently wrestling with the questions of what Data Science is, what it can be, and what scientific and social problems are unique to the modern proliferation of massive amounts of contextual and personal data [1]. UC San Diego is well-positioned not only to shape the future of Data Science, but to *define* it.

The Division of Social Sciences views Data Science as a distinct academic discipline that advances mathematical and computational methods applied to data—data that is predominantly about humans, and used by humans to understand and make decisions about humans in a social context.

We believe UC San Diego offers a unique, yet critical, perspective on Data Science that can sculpt the future of the field and cement UC San Diego's dominance in this domain. In particular, our strengths in Engineering and Mathematics as well in the Social Sciences and the Humanities provide opportunities to integrate the technical features of Data Science like machine learning and data analysis with the human, cognitive, and behavioral domains.

## Data Science and the Social Sciences

The social sciences are central to modern Data Science, as most Data Science applications are concerned with predicting, understanding, and emulating human behavior. Additionally, academic Social Science disciplines are critical to advancing state-of-the-art techniques, making sense of data sources, formalizing research objectives, and tackling foundational questions.

Since modern machine learning often aims to emulate human performance in naturalistic tasks such as language processing, image recognition, and semantic analysis, it is no surprise that advances in methods often come from a joint investigation of human and machine performance. For instance, models of semantic relatedness used to characterize human memory and reasoning have served as the basis of modern semantic analysis in natural language processing [2]. Moreover, since humans have vast expertise in real-world reasoning, training algorithms to reproduce human performance in addition to ground-truth improves classification accuracy [3]. Developing and improving data science algorithms by analyzing and incorporating human behavior is an active area of research in areas ranging from the integration of neural and structured representations [4] to developing resource efficient algorithms that fail as gracefully as humans under computational constraints [5-8].

Likewise, methods developed by the Social Sciences for the study of human behavior and society are a staple of Data Science. These include the substantial cross-pollination between computational Linguistics and Natural Language Processing in semantic and grammar models, classic developments in psychometrics like Factor Analysis, Multidimensional Scaling, and Cluster Analysis; time-series models from econometrics; social network analysis from Sociology; and interface and visualization design from HCI. These methods and others were developed in response to the unique challenges inherent to making sense of human, linguistic, and social data, thus placing the Social Sciences in an essential role in advancing Data Science methodology.

Furthermore, research and applications in the emerging field of Computational Social Sciences often drive innovation in Data Science by identifying new connections between classes of questions and sources of data. For example, we can analyze millions of books digitized by Google Books to uncover language evolution, cultural and political structures, and thought suppression [9] or combine huge amounts of data from archeology, criminology, economics, geography, history, political science, and psychology to analyze the effect of climate on human conflict [10].

Outside the academic setting, social scientists use data science methods to tackle concerns in their own industries. These applications include social network analysis to identify "hubs" and trend-setters for purposes of marketing, GIS in urban planning and development, psychometric tools to develop human resources testing materials, semantic analysis of social media to forecast stock prices, econometric models for data-driven development of market mechanisms in the sharing economy.

**Social Science Questions in Data Science**

Social science questions, addressed with suitable methods and expertise, are fundamental to developing a mature field of Data Science. Here are illustrations of such questions:

> *Why are some problems more amenable to purely data-driven approaches using generic learning algorithms than to domain-specific structure, or vice versa? What factors determine data quality with respect to a particular question? Can we ascertain* a priori *whether a particular question can be answered via a particular data source? Can we estimate the data requirements for adequate algorithmic performance in a given domain? When can training on synthetic data substitute for training on real-world data?*
> These questions emphasize both the science and art in Data Science, and indicate a need for Data Science to serve not only as an applied field, but as a distinct scientific discipline, with its own foundational research questions and theory.
>
> *How do we combine unstructured, data-driven machine learning algorithms with human domain expert knowledge [11,12]? More generally, how can we design systems that integrate human intelligence with algorithmic data-science predictions, leveraging humans' rich understanding of the world to improve predictions, and to help human decision-makers with algorithmic forecasts [13]?*
> These questions highlight the critical role that humans and human-level understanding play. For instance, human-machine hybrid teams regularly outperform human- or machine-alone systems at chess tournaments; why might this be? How can humans

understand and explain the reasoning behind particular machine learning predictions [14]? That is, how can an algorithm predicting a particular asset acquisition for an investor explain its prediction when the representations the forecast is based on are non-linear combinations of human-accessible variables?

*How does society balance the demand for data and individuals' rights to privacy? Is it possible to develop provably anonymous data-gathering strategies?*
If individuals give up their own communication data for access to convenient services, they necessarily reveal data about individuals they interact with; thus giving up one's data comes with negative externalities for others. What does anonymity mean in an era of uniquely identifying meta-data? Classically, "anonymized data" removes specific unique identifiers, but in sufficient quantities, even seemingly anonymized data (like browsing/location history from a cellphone) can uniquely identify an individual.

*How do we identify and anticipate challenges that some applications of data science might pose to the functioning of our democratic political, economic, and cultural institutions?*
It's now widely recognized that data science (from the use of psychometric data scraped from Facebook and other sources) played a pivotal role in the Brexit campaign and the 2016 US presidential election. However, precisely how data analytics were applied, and how central a role these techniques played, are not widely understood, largely because of the lack of transparency surrounding the operations of private companies hired by the respective political campaigns. These recent politically game-changing elections represent just two examples of the profound impact data science is having on political, economic, and social life, that offer clear evidence that we need greater transparency and understanding of data science and its applications, as well as wider and more varied input into debates and decisions about its appropriate methods and applications.

*How can data-intensive organizations avoid perpetuating and reinforcing the inequalities inherent in data as algorithms and automation gain prominence in ever-more important aspects of modern life [15,16]?*
A Google Image search for "babies" at one point returned images of almost exclusively Caucasian babies in the top results [17]. Algorithms aimed to predict recidivism rates weighted heavily on race [18]. Google ads have discriminated by race and gender [19]. What role do these algorithms play in shaping social inequalities, and how do those inequalities influence the algorithms that increasingly play largely hidden, but important, roles in our daily lives?

*How do we develop forms of data-literacy that foster collaboration across disciplines and generate awareness of the social, political and historical embeddedness of data and data infrastructures? As an educational institution, how do we promote data-literacy at all levels, from undergraduate education, broadly conceived, to even the data collection activities of the university itself?*

Decisions in organizations inform how data are gathered, used, and analyzed. Like other human artefacts, data and the information platforms that it is associated to encode specific cultural evaluations of what matters and what doesn't. Producing robust claims about society requires not only competencies in working with and visualizing data, but also in evaluating its validity as a representation of underlying practices, behaviors and attitudes. Creating these forms of data-literacy and data-awareness could also provide incentives for communication, collaboration and sharing across disciplines that, traditionally, have different concepts of data quality.

**Developing the Data Science Institute**

Since the field of Data Science and its central questions span multiple divisions, we anticipate that this multi- and inter-disciplinarity will be reflected in the structure of the UCSD Data Science Institute, its research aims, as well as its educational agenda.

As a starting point, we propose the following concentrations:

- **Data Engineering:** The development of data architectures, algorithms, systems, etc. for capturing, storing and processing an exponentially increasing torrent of data. In industry, holders of jobs with this title are often tasked with establishing the data infrastructure necessary to make analytics and machine learning feasible.

- **Machine Learning and its Foundations:** The mathematical and algorithmic tools for learning from data and their theoretical foundations, including questions such as which problems are more or less amenable to general purpose data-driven approaches, what are the sample, memory, time complexities of specific problems, inter alia.

- **Social Science Oriented Analytics[1]:** Although some AI/ML applications can be entirely machine-centered (a control system for a quadrocopter), most of Data Science aims to generate usable insights for humans about human behavior. Here the focus is on understanding how data science tools can be adapted to answer social science questions, how social science processes generate data and what this means for their analysis, how the results of large-scale machine learning can be made useful/understandable to people, how to seamlessly integrate human intelligence with machine learning in expert systems and for crowd-sourcing applications, and how expert knowledge interacts with machine learning approaches.

- **Data and Society:** Beyond being a tool, Data Science itself should be an object of social science investigation. There is a need to bring cutting-edge research and applications in

[1] Other options: "Human Data Interaction" "Human Algorithm Interaction" "Human Centered Analytics" "Human Targeted Analytics". The goal is to convey that both the subjects and the consumers of data science are usually human, so we can't leave them out of the loop.

data science in conversation with democratic legal frameworks as well as forms of social analysis that examine how values get designed into technical systems. How can we develop provably anonymous data gathering and reporting strategies, balance the need for privacy with the demand for data, incorporate fairness and accountability into algorithms, and counter statistical/algorithmic discrimination to prevent data-driven approaches from perpetuating and reinforcing inequities?  More broadly, we need to understand the social ethics required for a society where data science applications are pervasive.

### The Data Science Initiative and Undergraduate Education

Big data is transforming Social Science as the microscope transformed biology. It transforms not only our ability to study human subjects, but also the disciplines themselves: human communication, markets, sociology, race relations, and politics are changing at an accelerated rate in response to new data-driven media and products. These changes also affect our undergraduate students as they enter a society and labor market shaped by data technology.

As technological change accelerates, inclusion in training and access to researchers become vital to our students' careers. All social science students should be included in training of the basic skills of Data Science (programming and statistics). Our students are already responding to the changing shape of society, with many enthusiastically seeking courses in coding, statistics, and quantitative methods.  Indeed, when our best students complement the standard curriculum of Social Science with some statistics (econometrics) and computer programming skills, they are positioned to be hired into data analytic fields such as financial and research firms.  However, expanding data science education among social science undergraduates requires that technical classes in the social sciences have the same laboratory training as those in engineering.

To allow our students to anticipate, and adapt to, changes in society, they must have access to research faculty and graduate students at the cutting edge of Data Science, in classes small enough to allow meaningful interaction.  Such access has two benefits: first, it allows students to see what research and Data Science are like, perhaps even trying them on to see if they fit; second, it gives students a peek at what the next wave might look like, so that they can prepare for a working life in which they can master accelerating technological changes, rather than be displaced by them.

To be specific, inclusion and access have two critical components:

1. Numerous classes at both the lower- and upper-division levels in coding and statistics taught by ladder rank faculty, augmented by labs staffed by graduate students pursuing research using quantitative methods, at lab TA-to-student ratios. These classes should not be concentrated in a few departments, but broadly available across the division, offering access to the wide spectrum of issues in Social Science Analytics.
2. For students interested in pursuing Data Science as a minor, a full curriculum of classes, also

taught by ladder rank faculty and augmented by qualified graduate students, at lab TA-to-student ratios.

With foresight and thoughtful allocation of resources, the Division of Social Science can feasibly achieve inclusion and access to Data Science in undergraduate education for the largest student body on campus. We have the expertise and are eager to make this transformational commitment to our students.

**Recommendations**

We have opened a dialogue on data science across the Division of Social Sciences. There is faculty enthusiasm for a variety of Data Science activities. Here are some recommendations that this interdisciplinary conversation has generated. They fall into three categories that can define the core mission of the Data Sciences Institute:

*1. Advancing Data Science as a distinct field of scientific inquiry and engineering, unifying faculty across many departments and divisions. Examples:*

- Have the data science institute span multiple divisions to reflect the interdisciplinarity of the field
- Organize the data science institute to reflect the variety of concentrations needed in the scientific study of data science as well as the expertise of its practitioners: e.g., Data Engineering, Machine Learning, Human Oriented Analytics, and Data and Society.
- Faculty lines with joint appointments at the DSI, and other departments.
- Temporary faculty fellow appointments like Berkeley BIDS [20] to bring together faculty across departments and divisions.
- Advanced, independent postdoctoral fellowships like those at the NYU CDS [21], wherein fellows are free to collaborate with multiple members of the UCSD Data Science community, thereby helping to bring together the many disciplines and divisions working on data science.
- Cross-department, or rotating department speaker series featuring Data Scientists from industry, non-profit agencies and universities

*2. Improving data science education at the undergraduate, graduate, and faculty levels. Examples:*

- Take advantage of the general and domain specific Data Science courses taught throughout the Division of Social Sciences, by recognizing that they too require heavier investment in lab and TA time.
- Internships and portfolio building projects during education, including capstone projects. In particular, extramural projects wherein students work with outside companies, agencies, or labs.
- An interdisciplinary graduate program in Computational Social Science, at both the masters' and doctoral levels
- Additionally, UC San Diego Data Science should place a special emphasis on Data Ethics,

Privacy, Algorithmic Biases, and, broadly, how Data Science can predict—and influence—human behavior.

*3. Pursuing civic and social good through outreach and community involvement. Examples:*

- A civic good oriented summer program similar to the University of Chicago's Data Science for Social Good Summer Fellowship program [22]
- University and community partnerships creating easy-to-use software tools for search, visualization, and analysis of big data for the lay members of the community
- Public speaker series, and
- Workshops that emphasize the inevitably interdisciplinary future of Data Science.

## References

[1] http://datascience.nyu.edu/what-is-data-science/
[2] https://en.wikipedia.org/wiki/Topic_model
[3] https://www.google.com/patents/US20160148077
[4] https://nips.cc/Conferences/2016/Schedule?showEvent=6208
[5] https://nips.cc/Conferences/2016/Schedule?showEvent=6243
[6] https://sites.google.com/site/icml2015budgetedml/
[7] http://www.darpa.mil/program/xdata
[8] https://nips.cc/Conferences/2015/Schedule?showEvent=4910
[9] http://science.sciencemag.org/content/331/6014/176
[10] http://science.sciencemag.org/content/341/6151/1235367
[11] http://www.darpa.mil/program/data-driven-discovery-of-models
[12] https://nips.cc/Conferences/2016/Schedule?showEvent=6208
[13] https://recsys.acm.org/recsys17/intrs/
[14] http://www.darpa.mil/program/explainable-artificial-intelligence
[15] http://www.fatml.org/
[16] https://recsys.acm.org/recsys17/fatrec/
[17] http://www.bbc.com/news/technology-21322183
[18] https://www.bloomberg.com/features/2016-richard-berk-future-crime/
[19] http://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you
[20] https://bids.berkeley.edu/call-data-science-fellow-applications
[21] http://cds.nyu.edu/nyu-moore-sloan-data-science-fellows/
[22] https://dssg.uchicago.edu/