Welcome to the Datazine, in which Stanford Data Science Scholars present key highlights of their data science-related research and a glimpse of themselves as interesting individuals.

The research presented shows data science at a very high level of relevance and innovation, communicated clearly and vividly. As someone who has spent more than 50 years in related research, I would unhesitatingly direct someone here who asks “What is real data science all about?”

The Scholars’ research, in a wide range of knowledge-based fields, applies modern techniques to obtain scientifically valid insights from relevant data. Other research in computing, statistics and mathematics advances the techniques and their effective use.

The diversity of fields is striking, and a part of our goal for the Scholars. But even more striking and encouraging has been to observe how this diverse group works together. Their complementary insights and team spirit enrich their work, with insights communicated back to colleagues in their individual fields.

Stanford Data Science supports the scholars in their research, along with their participation in projects that are helping to build a better data science community for Stanford and beyond. The Datazine itself is an example.

Data science has a crucial role in meeting current and future challenges to our planet. Its success in that role in turn depends crucially on talented and motivated young researchers such as those presenting themselves here.

John Chambers

The Stanford Data Science Scholars program is part of Stanford Data Science, one of the Long Range Vision initiatives of Stanford University. Its mission is to weave Data Science into the fabric of the university, and at its core, support data science research through early career scholars. Stanford Data Science is led by Faculty Director Emmanuel Candès and staff Executive Director Chris Mentzel, under the organization of the Vice Provost and Dean of Research office and in partnership with all seven schools at Stanford University.

For more information, please visit: https://datascience.stanford.edu.

Stanford Data Science Scholars are supported in part through the generosity of the Ram and Vijay Shriram family and Stanford’s Bing Presidential Fund.

The 2020 Datazine was created and edited by current scholars Fátima Pardo-Avila, Anna N. Rasmussen, Mae Velloso-Lyons and Tongtong Zhang. Design by Mae Velloso-Lyons.
How does climate drive dengue incidence at the population level? How is the climate–dengue relationship influenced by susceptible availability? Previous experiments and models suggest that climate affects mosquito-borne disease transmission such as dengue. However, it is difficult to infer climate–dengue mechanisms empirically from observational data at the population level due to complex nonlinear interactions between climate and other factors.

In this study I used long time series (19 years) of dengue incidence data, estimated susceptible population size, and climate data with a data-driven approach for nonlinear systems called empirical dynamic modeling to infer drivers and their interactive, state-dependent effects on dengue dynamics in Puerto Rico. My colleagues and I found evidence of climate effects on dengue dynamics at the population level, but only when susceptible availability was high. Knowing when climate plays an important role in dengue transmission is important for epidemic forecasting and timing of public health interventions.

This figure shows how temperature and rainfall effects on dengue vary depending on susceptible availability (i.e., how many people could be infected with dengue), which was estimated using a proxy for the susceptible population size denoted by λ. Plotted here are the rates of change of dengue incidence against temperature and rainfall. Neither climate driver has an effect on dengue incidence when susceptible availability is small. However, above a certain threshold climate effects are observed: temperature has mostly a positive effect, and rainfall has a negative effect. The red and blue solid straight lines represent median regressions (dashed red and blue lines represent 0.05 and 0.95 quantile regressions) of temperature and rainfall effects, respectively.

I received a degree in dental surgery from Karolinska Institutet and an MS in statistics from Stanford University. I am currently a PhD candidate in the Department of Biology at Stanford University. I study the ecology and evolution of infectious diseases using mathematical and statistical modeling and various data mining approaches. I try to understand how diseases are transmitted between different species and what environmental mechanisms drive disease dynamics using epidemiological and genomic data. My current work focuses on dengue, yellow fever, West Nile and COVID-19 in the Americas, and canine distemper in wildlife in Alaska and Yellowstone. My goal is to promote One Health by finding sustainable win–win solutions for public health and wildlife conservation.

Fun facts about me:

I grew up speaking four languages at home. • I play the acoustic/electric guitar and the flute. • I love animals and swimming.
The smartphone has radically changed how political information arises in Americans’ daily lives. In contrast to older media platforms like print and television, bits of news can be accessed throughout the day, from any manner of sources. Politics can even be encountered incidentally -- e.g., when scrolling through social media, or when entertainment content references the state of the world. Part of my work is to unravel exactly what ‘news’ means in this new media environment, and how to measure human interaction with it, so that the educational and psychological effects of modern news consumption can be better understood.

As part of the Screenomics Lab, I’ve helped to collect and analyze moment-by-moment smartphone usage. Participants around the world can install our custom app that uploads an exact screenshot from their smartphone every few seconds, for weeks. This lets us computationally dissect exactly what appeared before the eyes of smartphone users, even capturing elusive ‘incidental’ news encounters.

In this figure*, I zoom in on a 30-hour slice of life from an American Screenomics participant, collected in mid-2017. The goal is to simply illustrate the complexity of news consumption via the smartphone.

In gray, I highlight each moment that the participant’s smartphone screen was active over this time period, as measured by the Screenomics Lab’s custom screen-capture app. Then, for either of the selected news topics, I used word-matching to detect moments throughout the day in which a key news topic arose. With further scrutiny, we can detect which occurred incidentally, e.g., by scrolling by a headline on Reddit, and which occurred intentionally, e.g., by reading a news article.

As shown in the pop-out, incidental encounters with political news can lead to intentional encounters, blurring the distinction between the two. The screenshots here are synthetic recreations of the originals, as a standard privacy preservation measure. *This figure can be found in Reeves et al (2019) in Human–Computer Interaction, 1-52.

I am a fifth-year PhD Candidate in the Department of Communication at Stanford. I research how the structure of media platforms can influence our behaviors and opinions, especially regarding political news consumption. The most unique data I have collected and analyzed is from the Screenomics Lab here at Stanford, and I am also on a team sifting through a massive panel of TV news consumption data from the Nielsen Company. Prior to Stanford, I have earned degrees in political science & economics, and I have been on research or product teams at Harvard SEAS, Microsoft Research, and Instagram.

Fun facts about me:

I’ve done voiceover narration for educational science videos. • I’ve driven coast-to-coast four times. • I’ve performed ukulele on live radio in Yangon, Myanmar.
San Francisco Bay is a highly nitrogen polluted estuary surrounded by approx. 8 million human inhabitants. Despite large inputs of nitrogen from wastewater treatment plants, urban runoff, and agricultural runoff, San Francisco Bay has not yet suffered many classic symptoms of eutrophication such as harmful algae blooms or dead zones. Using decades of water quality data collected by the USGS, a time-series of bacterial and archaeal ‘barcode’ genes, and metagenomics, we are investigating the ecology of microorganisms that eat and breathe nitrogen.

Our sequencing efforts have revealed a previously unknown ammonia-oxidizing archaea bloom (abundance shown in purple bars) in South San Francisco Bay. This large population of ammonia-oxidizing archaea consumes ammonia (blue line) leading to high and persistent nitrite accumulation (pink line), a rare occurrence in oxygenated waters.

By exploring large genomic and water quality datasets we can better understand the relationship between the Bay’s smallest constituents and the fate of excess nitrogen; and how this may change in the future.

What does my data look like and how do I deal with it?

Months of field sampling and lab work eventually culminate in very large text files of DNA sequences (.fasta files). Then I use computationally and memory-intensive bioinformatics pipelines to get biologically meaningful information like species “barcodes” or genomes from the raw sequences. I use several types of databases to try to classify the function and/or taxonomy of sequences. I summarize functional, taxonomic, and sequence count data into large matrices, then I get all that information into R! I use a variety of statistics to understand the data, especially linear algebra to convert large matrices into smaller matrices. Some of my favorite packages include: vegan, phyloseq, and, of course, ggplot2!

I am a 4th year PhD Candidate in the Department of Earth System Science at Stanford University focused on geomicrobiology under the direction of Professor Chris Francis. By examining DNA, RNA, nitrification rates, and historical water quality data, I hope to connect the ecology of microorganisms to biogeochemical cycling in San Francisco Bay and give insight into how the system may change in the future.

While I love to go sampling on the boat, most of my time is spent wrangling data in R.

I am motivated to study microbes because I love life and the Earth we live on. Growing up I thought I would study charismatic mammals or trees, but during my undergraduate thesis microorganisms captured my heart. After exploring a few career avenues, I found a way to study big problems by thinking on a much smaller scale. Microbes have played an important role in shaping the Earth for billions of years and are increasingly important to understand as humans address the climate crisis and other human impacts on the environment.

Fun facts about me:
I enjoy drinking cider. • I love to hike. • My favorite book is Jane Austen’s Persuasion.
In this digital era, how do young people communicate with their parents through digital devices? Beyond directly asking them, we can answer this question by objectively observing their activities and behaviors in the digital world. In particular, I am leading a paper using data from the Human Screenome Project at Stanford University to describe and examine patterns of young adults’ communication with their mothers and fathers. Screenshots were passively and unobtrusively collected from individuals’ smartphones every five seconds over weeks and months, encrypted, transmitted, and stored in our secure server. Using a custom designed optical character recognition (OCR) module, text was extracted from each screenshot and stored as metadata.

Based on these data, I use a combination of text-based search and human judgment to identify communication events between young adults and their parents. Events identified include phone calls, SMS, video calls, WhatsApp messaging, Facebook interactions, and location sharing. In addition, I also identify the communication partner (e.g., whether it is mom or dad), and directionality of the communication (i.e., who initiates the event) for each event. In total, based on 457,905 screenshots from 10 young adults in the U.S., I identified 1,413 communication events between them and their mother and/or father. An especially interesting observation is the idiosyncrasy of the communication patterns for each young adult, which has been dismissed in prior research. The fine granularity of the screenomics approach makes this observation possible.

I am a postdoctoral scholar at Stanford University, affiliated with the Departments of Pediatrics and Communication. My current work is mainly involved with The Screenomics Lab and the Solutions Science Lab. My research applies innovative statistical and computational methods to a range of data sources to investigate family dynamics and youth well-being across multiple time scales.

I am primarily interested in family systems processes and adolescent and young adult educational attainment and career development. My research can be grouped by data source and corresponding methodologies into three projects: (1) Using multi-family member panel data to study socio-cultural context, family dynamics and adolescent and young adult development and achievement; (2) Using high-dimensional, large-scale data to build machine learning models for predicting young adult achievement from early family experiences; (3) Using digital big data to observe and analyze family dynamics on micro time scales and investigate the role of technology in family and adolescent life.

Fun facts about me:

I play keyboard for an amateur rock band. • My hometown, Weifang, China, is known as the kite capital of the world. • My favorite trip to date was a cross-country road trip from Pennsylvania to California.
These figures show plots for two young adults’ (YA) communication patterns with their mom and dad. The horizontal axis indicates time of the day, and the vertical axis indicates what day it was in their participation of the study. Thus, each grid represents for each day, when the communication events occurred. Each colored bar indicates both the occurrence and the duration (in cases of phone/video calls) of a communication event. The color indicates whether the communication event was with dad (green) or mom (red), or within a family communication group (purple). Dots that overlay the colored bars indicate that those events were sent or initiated by the young adult. With such high resolution of observing and describing family digital communication, an important finding is both the within- and between-person differences in their communication patterns. The differences include day-to-day variabilities in frequency, duration, timing, communication partner (i.e., mother, father, or the family group), and interactivity within each young adult.

Zooming out, we can also see individual differences in who they communicated with the most, and whether there was a routinized family communication. An important next step is to capture these different patterns and then examine how these differences may influence family and young adult well-being.
There are many evolutionary advantages to large body sizes. And there is a reason lions prefer eating gazelles to giraffes. A less well-known advantage of a big body is: energy efficiency. Pound-for-pound, big plants and animals burn fewer calories than small ones, but the limits to that efficiency are unknown. The largest animal with a known daily energy budget in the wild is a walrus, which weighs about a ton and a half. So what is the energy budget of a blue whale, the largest animal that has ever lived? Are they the most energy efficient animal on earth? Or are there limits to energy efficiency? Unfortunately, traditional energy measurements do not work for blue whales; they are just too big. I estimate their energy budgets indirectly by measuring oxygen consumption. If you have ever seen a whale blow, you know their breaths are both big and distinct. I count individual breaths using sensors attached to a whale with suction cups and measure the volume of air in each breath using drones and infrared cameras.

This is two days in the life of a blue whale. On the left, we have the dive profile of an animal that fed intensively from dawn until dusk (229 times, to be exact). We also know this animal breathed almost one thousand times that day. On the right is the dive profile of the same whale a few days later. On this day, the animal was traveling south looking for food. Although it dived about the same number of times as the feeding day, we know it consumed less oxygen because it took only 70% as many breaths. Fun fact: the spikes in the dive profile on the right, down to about 40m, correspond to when the whale was singing.

In addition to blue whales, I’m also measuring the energy budgets of their smaller relatives. Minke whales are only one-tenth the size of blue whales, with humpback and fin whales filling out the size range. By looking at trends from “small” whales up to the biggest, we will better understand the limits to energy efficiency and body size.

I am a 3rd year PhD candidate at Stanford University’s Hopkins Marine Station in the Goldbogen Lab, where we study the drivers and limits of extreme body size. My research investigates the energy budgets of large whales, asking the question “how many calories does a blue whale burn in a day?” My methods include the use of “bio-loggers”: animal-borne sensors that track movement and the environment. Essentially a FitBit for critters, these devices yield gigabytes of data from each deployment. Data science techniques help me visualize and identify important physiological events, like feeding and breathing.

Fun facts about me:

I have a darling blue nose pitbull named Bowie. • My beef short rib chili once brought a grown man to tears. • I’m an amateur woodworker (emphasis on amateur).
California houses America's largest “lifer” population, with 25% of its 115,000 prisoners serving life sentences. Each year, the Board of Parole Hearings (BPH) conducts thousands of parole hearings to decide whether to grant prisoners early release. While each hearing is transcribed into about 150 pages of dialogue and sent to the BPH and governor’s office for review, capacity constraints mean that, in practice, only grants of parole are reviewed. Broadly, this project aims to assist reviewers in identifying decisions that most need a second look. One of the approaches is a simple unsupervised technique for using language models to identify procedural anomalies in long-form legal text.

Another approach (see diagram, right), is a weak supervision technique for extracting case factors from the text, to then be used as features in a downstream anomaly detection model. The hearings cover a large range of topics, and as a result, so do the features we want to extract from the hearings. These topics include questions about the commitment offense (e.g. Was the parole candidate the sole participant in the crime? Was the parole candidate under the influence of drugs at the time of the crime?), questions about pre-commitment factors (e.g. Was there a history of drug usage prior to the crime? Was there a history of gang involvement prior to the crime? Was there abuse in the family?), questions about behavior in prison (e.g. What kind of programming did the candidate participate in? What was the extent of disciplinary writeups?), and questions about post-release plans (e.g. If the parole candidate were released today, would he or she have a place to live? A job offer?).

A visual example of what we can do with these features is a nearest neighbor visualization. For a hypothetical parole candidate, in blue, we can show historical parole hearings of candidates with similar factors. The closer on the plot, the more similar. The blue person refers to the hypothetical candidate. The green circles refer to historical parole hearings that resulted in parole being granted. The red circles refer to hearings that resulted in a denial, with a larger circle showing a longer amount of time until the next parole hearing.

Hi! I’m Jenny, a PhD student in the Computational Social Science program in the Management Science and Engineering department. Over the last few years, I’ve been working on a natural language processing and causal inference project on legal data, transcripts of hearings from California’s Board of Parole Hearings. I’m excited about bridging the gap between technology and policy and working at every step of that ladder. More recently, I’ve started exploring the topic of differential privacy, a more methodological question about how data science can be done in such a way where we protect the attributes of each underlying data point.

Fun facts about me:

In my spare time I enjoy doing henna, such as at university events! • I love taking trips to stargaze! • My go-to cuisine for takeout is Ethiopian food.
Sex differences have been reported in many traits, such as height. In most cases, males and females have different average levels but overlapping distributions of the trait (e.g., some males are on average taller than females, but some females are taller than some males). However, it is unknown to what extent these differences in these traits are due to differences in the genetics of the trait. I worked on a project to examine sex differences in the genetics of biomarker levels in a population of over 300,000 men and women from the UK Biobank. We built a model to examine the genetics of thirty-three blood and urine lab tests and found that the majority of these biomarkers did not show sex-related effects, with the exception of testosterone. For testosterone, we found a large number of genetic variants that showed effects in males but not females and vice versa. Previous studies of testosterone genetics focused on males; however, testosterone is not just a male hormone, and by considering sex in our analysis we are able to better understand how genetics relates to testosterone levels in both men and women.

I used ggplot in R to make this plot. We highlighted genetic variants that we found to have associations with testosterone in females and not males (orange), males and not females (purple), or both males and females (blue). For example, a genetic variant highlighted in orange will be far from the axis in females but then close to zero in males, indicating that we see a strong association between that variant and testosterone in females but not males. The genes for the top ten most significant genetic variants in males or females are also labeled in the diagram.

I am a sixth year PhD candidate in the Biomedical Informatics program working in Dr. Russ Altman’s lab. I completed my undergraduate degree at Smith College, where I started out as a Biochemistry major, but then discovered programming and data analysis and was hooked. I then went to graduate school seeking to use computer science and statistics to study problems in biology. My research leverages data analysis methods to better understand sex-, hormone-, and drug-related variability in genome-wide association studies and gene expression data. By studying these topics, I hope to contribute to improving health outcomes across the sex-gender spectrum.

Fun facts about me:
I grew up in the Boston area and still miss the seasons despite being in CA for six years! • My fluffy gray cat Winston loves to go for walks. • I like to bake, especially sourdough bread and bagels.
Are neural networks really “black boxes”? My current research studies the dynamics of neural networks during training. A neural network model is well defined in its architecture and the optimization process used to train the model, but how the model changes during training is not well understood or easy to predict.

In my research we view the training dynamics as a physical process and borrow tools from physics to predict the trajectory these models take during training. To test our predictions we train neural network models on commonly used datasets and save statistics of their parameters through training time. Our hope is that through better understanding a neural network's training dynamics we can improve their optimization process and better understand how these models make predictions.

This figure is an example of how we learn about neural network models as they train. In this figure we look at a well studied neural network model (VGG-16) and focus on a specific layer within the model. We look at a specific statistic, the squared Euclidean norm, for each convolutional filter from this specific layer during training on a commonly used dataset of classified images (CIFAR100).

The colored lines are the empirical norms for these filters and the black dotted lines are our theoretical predictions. There is a striking alignment of our theory with the observed output statistics of the model.

I am a PhD Candidate at the Institute for Computational and Mathematical Engineering focused on computational neuroscience and deep learning theory under the direction of Professor Surya Ganguli and Daniel Yamins. Prior to Stanford I received my bachelor's degree in applied math from Brown University where I developed an enthusiasm for statistics and its broad applications in biology. I came to Stanford to further my understanding of statistics and computer science, and soon found myself very interested in their intersection and the growing field of deep learning. My research focuses on the dynamics of artificial neural network models during training and how to leverage this understanding to generate hypotheses for how biological neural networks might learn.

Fun facts about me:
I make really good miso soup. • I spent a summer hiking the PCT. • I currently live in a one-hundred year old cottage.
How do young children grow and change? A large amount of theory has addressed this question, most famously Piaget’s stage theory, but it has received little empirical attention. We use data provided by Kinedu, Inc., a developer of a parenting app, to address this question. While using the app, parents answer dozens of milestone questions such as “Can your child stand on their toes?” or “Does your child babble to imitate conversations?”

I use a novel cross-validation procedure to compare a variety of item response models fit to this data. I find that a three dimensional model fits the data well and generates accurate predictions. These predictions can be used for a variety of purposes including helping parents understand their child’s development and targeting interventions. The first of the three dimensions is highly associated with age, but the second and third dimensions are not, making them particularly useful for understanding how a child is progressing through their early development.

Understanding the nature of variation in milestones could help shed light on children’s development. This figure illustrates the three developmental dimensions (i.e., factors) that we find. The first dimension, which is highly associated with age, focuses on linguistic and cognitive milestones such as “Does your child make animal sounds?” or “Does your child find hidden objects?” The second dimension corresponds highly on physical milestones—for example, “Does your child turn book pages one at a time?” The third dimension, similar to the first dimension, is driven by linguistic milestones, which highlights how critical language is to early child development. Using multi-factor item response theory models and a new cross-validation method for model comparison, we find that a three-dimensional model best describes developmental variation across the first 55 months. This work has significant practical implications, for example: measures of developmental change should not assume that a single score captures all of the variance in developmental change.

In education, it is wild how much data we now have. Just 20-30 years ago, the only data we had was whether a student showed up to school each day, the grade a teacher gave them, and occasionally a standardized test. Now, with much of education happening online, we often have data on every question that a student has answered — did they get it right or wrong? How much time did they spend on that question? I broadly characterize my work as educational data science—the work of extracting value out of this educational data. In particular, how can we give students better questions to help them learn more? How can we make smarter decisions about what material a student has mastered? How can we identify students that are falling behind and the best ways to help them catch up? This work is critical to preventing the massive and devastating difference in performance that we see across students.

Following college, I taught high school math at Warren Early College High School in Warren- ton, North Carolina as part of Teach for America. I then worked for three years at Khan Academy on the math content team. Motivated by the possibility of leveraging the explosion of educational data gathered by organizations like Khan Academy, I joined the Stanford Graduate School of Education as a PhD student in 2016. I focus both on data science in education—e.g., measuring learning—and the education of data science—e.g., helping someone build the skill of gathering insights from data.

Fun facts about me:

My favorite thing to do is go for a long walk and listen to Minnesota sports talk radio.
• I love pasta. • I grew up with a very protective dog named Snuggles.
Have changes in digital environments, such as the use of search engines, led to a greater concentration or a broader dispersal of scientific citations? Using the Web of Science database, we analyzed the distribution of citation counts made in four broad scientific fields (Health; Humanities; Mathematics and Computer Sciences; and Social Sciences) during 1996–2014. If the inequality of citation distribution has become more unequal in recent decades, it would imply the gaze of scholars has been concentrated on a few highly cited papers. Our research team found that frequently adopted inequality measures, such as the Gini coefficient, were biased when applied to the distribution of count variables with lots of zeros. Once we adjusted inequality measures with our suggested resampling method, we found little evidence that the inequality in citation distributions has changed over time. This work was recently published in Sociological Science.

The above results highlight that we cannot rely on having big data by itself to provide the right answers to our questions. We need to scrutinize and understand the structure of the data to answer our questions in the right way. For example, the Gini coefficient is a simple and widely-used measure of inequality so we often underestimate how biased our results can be. We also often think our results cannot be wrong because the Gini coefficient is calculated across millions of papers and citations. However, our research shows there is always the possibility that the underlying structure of the data can obscure results, even when using simple measures on large datasets.
The ribosome is the molecule in the cell responsible for making proteins. It can be found in the cells of all living organisms, and it is critical for the survival of all cells. For that reason, most antibiotics target and stop the function of bacterial ribosomes. Understanding the regulation of the ribosome is crucial to understanding how cells work.

The ribosome is made of around 200K atoms, and thanks to the field of structural biology, we have been able to obtain “snapshots” of how those atoms are placed and interact with each other. Figure 1 shows one snapshot of the ribosome. In this structure, the protein being synthesized is shown inside the ribosome, (in the exit tunnel), and is still attached to a tRNA (a molecule that serves to translate from messenger RNA (mRNA) to protein). Structures like the one shown in Figure 1 can serve as starting points of molecular dynamics simulations, which can produce movies of the ribosome performing different functions.

The ribosome plays such an important role for the survival of every cell that it is not hard to imagine that its function would be highly regulated. One of the mechanisms that cells use to regulate protein expression is the synthesis of translational arrest peptides (APs). APs are sequences of amino acids that, as they are being produced by the ribosome, they can stall the ribosome, thus stopping protein synthesis.

The SecM AP of Mannheimia succiniiciproducens (SecM(Ms)) is one of the most efficient arrest peptides, and it surprisingly consists of a sequence of only eight amino acids (HAPIRGSP). SecM(Ms) is routinely used as a force sensor to study processes such as protein folding.

While SecM(Ms) has been characterized biochemically, no structure has been described so far, thereby limiting our full understanding of how it functions. Based on available structures of similar systems, we have generated an atomic model of a ribosome with SecM(Ms) and used it as a starting conformation for molecular dynamics simulations. Our model is shown in Figure 2. Our goal is to describe how SecM(Ms) can cause the stalling of the ribosome.
**Structural biology and data science**

All living organisms are made of atoms, interacting with each other in particular ways in order to create life. The goal of Structural Biology is to figure out how atoms in the cells are arranged and how they interact. Biophysical methods have allowed scientists to experimentally figure out the position of atoms in molecules (i.e. determine structures). The collection of known molecular structures is available through public databases, such as the Protein Data Bank (PDB).

The information necessary to create a protein is encoded in its amino acid sequence, which can be seen as a sequence of letters known as amino acids. How to go from a sequence of letters to a three dimensional structure is still not well understood. With a growing number of available structures and advances in Machine Learning, the past couple of years have seen great progress in protein structure prediction.

As mentioned previously, we can use these static structures as starting conformations of molecular dynamics simulations. These simulations allow us to sample structures inaccessible by current biophysical methods. With a growing number of available structures, obtained experimentally and from MD simulations, combined with advances in Machine Learning, we can dream of a near future in which we will be able to take the sequence of a protein, and predict not only its structure, but also its thermodynamics and function. This is an exciting time to be a structural biologist.

**How my figures were created**

Representations of molecular structures of the ribosome were created with VMD 1.9.3. Analysis and plots were generated with python, using matplotlib and msmbuilder libraries.

I was born and raised in Mexico City. I obtained my undergraduate degree in Basic Biomedical Research from the Na
tional Autonomous University of Mexico. During that time, I discovered my love for Molecular and Structural Biology, in particular in understanding how biological processes occur at the atomic level. This motivated me to obtain a PhD in Chemistry at the Hong Kong University of Science and Technology, where I used computational tools to model how the RNA Polymerase (the enzyme that reads and translates DNA) works. My main focus as a postdoctoral researcher at Stanford has been on the ribosome, but I also enjoy collaborating in protein design projects. I am passionate about outreach, inclusion and diversity in Science, and this year I became co-pres-
ident of the Stanford Latinx Postdoc Association.

**Fun Facts About Me:**

I love cooking recipes I learn from YouTube, I recommend the channels “Cooking with Dog” and Maangchi! • My first programming language was LOGO, but now I prefer programming in Python. • One of my favorite trips was a one-day trip to the Arctic!
How did medieval Europeans envisage the human condition? What strategies did they use to represent it in literature?

My work investigates these questions by applying both quantitative and qualitative methods to the study of medieval fiction. The genre known as “prose romance” was enormously popular across Europe in the high and late Middle Ages (1200-1500) but the unwieldy length of many of the surviving examples and their high degree of textual variance (i.e. unlike for other genres, scribes copying these works didn’t copy the text word-for-word, creating many differences between manuscripts) mean that they are difficult to study using conventional methods of literary analysis. As a result, scholars have tended to focus on shorter works in verse, even though these circulated less widely and may therefore have less claim to represent the era they come from. While researchers applying data science approaches to the study of literature often build corpora of hundreds or thousands of works, this scale is rarely possible for premodern material. My work explores the possibilities of statistical analysis at the “medium scale” (around 1-3 million words).

This figure shows the occurrences of selected “body words” in the *Prose Lancelot*, a long narrative by multiple authors (c. 1220). The correlation between the words for “heart” and “body” in the first two parts suggests that they are interdependent in some way. Further analysis shows that the two terms are often invoked as a semantic binary: the body is presented as unchanging, vulnerable, and exposed, whereas the heart is described as malleable, inexhaustible, and secret. The repetition of this contrast points to a dualistic conception of the human being and to a cultural preoccupation with the tension between “private” and “public” selves.

I received my bachelor’s degree in languages from the University of Oxford, followed by a master’s degree in Medieval and Renaissance Studies from University College London. I came to Stanford in 2015 for a PhD in Comparative Literature, and quickly became enthralled by the quantitative analysis of literature! My research focuses on thirteenth-century literature in Old French and Middle High German, and tries to answer cultural questions through the examination of linguistic phenomena at scales not possible with conventional methods of literary analysis.

**Fun facts about me:**

My home is ruled by a tuxedo rescue cat called Cléa. • I like to crochet hats, scarves, and sweaters for anyone who’ll wear them. • The most beautiful place I’ve ever lived is southern Austria—nothing beats an uninterrupted view of the mountains!
I use computational tools to study the biomedical literature at a large scale. Meta-research, or research of research, is an evolving scientific discipline studying the process of organizing, producing and communicating scientific research. Its overall aim is to calibrate the scientific ecosystem such that new reliable knowledge arises as efficiently as possible, which has been instrumental in identifying the lack of transparency, rigor and reproducibility in current research practice. However, the vast majority of studies in meta-research involve researchers manually extracting information from the published literature, a time-consuming and laborious process prone to errors. With tens of thousands of new articles published per week, manually mapping and monitoring aspects of the scientific literature is unrealistic. My work develops open source computational approaches to studying the scientific literature at a large scale and introduces bridges between metrics and datasets compiled by separate entities in the field. By doing so, I aim to substantially improve our ability to produce meaningful, comprehensive and timely contributions to meta-research.

Recent concerns about the reproducibility of science have led to several calls for more open and transparent research practices. In response, we developed an automated approach to map five indicators of transparency (data sharing, code sharing, conflicts of interest (COI) disclosures, funding disclosures and protocol registration) across the entire open access biomedical literature of 2.75 million articles on PubMed Central (a repository of biomedical articles maintained by the United States National Library of Medicine). We then mapped indicators of transparency across galaxies of science, which divide 18.2 million PubMed articles into around 28,000 clusters of similar articles. The galaxy in grey represents all clusters of articles published between 2015-2019. On top of the grey galaxy, we overlaid colored representations of the proportion of each cluster that is Open access or reports on any of the indicators of transparency. The Open access galaxy is very similar to that of COI and Funding disclosures, suggesting that most of the open literature reports on both. However, with a few notable exceptions, most fields do not share their data or code and do not pre-register their protocols.

After serving as a lieutenant of the Cyprus National Guard, I completed my bachelor’s degree in neuroscience and my medical training at the University of Edinburgh, where I also obtained my licence to practice. I came to Stanford in 2016, where I obtained a master’s degree in statistics and recently graduated with a PhD in epidemiology and clinical research. I am passionate about using, developing and promoting best practice in medical research and clinical practice. I am now an AI Resident at Google, where I am excited to work on the foundations of a more efficient, equitable and evidence-based next generation of clinical medicine.

Fun facts about me:

I love attending new restaurants and writing reviews about them. • I enjoy classical music concerts and still play the violin from time to time. • I can’t resist flights with long layovers in cities I have never been to before.
How does political orientation affect job market outcomes in China? To study this question, we constructed 19,221 unique resumes using R, each resume having a randomly assigned proxy for political orientation. We submitted these resumes to real job vacancies across different industries in China. Contrary to our expectation, we found that a pro-government view doesn’t increase the chance of getting callbacks from employers when compared to not showing political orientation on the resume, while a pro-democracy view decreases the probability of callbacks by 9%.

We first used R Markdown to write three different resume templates, which had the same sections such as education, internships, and extracurriculars. Then, using R we randomized and plugged in three extracurricular activities across the resume templates to proxy for a student’s political orientation: participation in a “Western political philosophy study group” proxied for a pro-democracy view, participation in a “Socialism with Chinese characteristics study group” proxied for a pro-government view, and participation in a “Comic Book Club” indicated an apolitical position (control). The figure on the left shows that the extracurricular activities we used are effective and clean proxies of political orientations -- they change ratings of the job applicant’s political position in the expected directions and have insignificant effects on the rating of other qualities that may affect employment prospects.

I am a PhD candidate in the Department of Political Science. My research focuses on digital participatory institutions (e.g. online forums, social media sites) run by authoritarian governments. Why do authoritarian regimes create internet participation channels that allow the general public to express their grievances? Why do some dictators provide more of these channels for participation than others? How do these e-participatory institutions influence public opinion and the stability of the authoritarian regimes? I use computational and experimental methods with large-scale datasets on China and other authoritarian countries to attempt to answer these questions.

Fun facts about me:

I love dogs, especially golden retrievers! • I really like designing creative field or survey experiments. • In my spare time, I enjoy hiking, movies, and audio books.
Natural language processing is an area of computer science that has seen enormous progress over the past 10 years. Despite being a highly quantitative field of research, however, there is often a lack of rigor when it comes to evaluation. In particular, the choice of hyperparameters (or model selection) plays a major role in determining how well a model performs, such that it is often unclear whether a new model is better because of new ideas, or simply because more effort was put into optimizing the model. In this work, we sought to establish a new framework for evaluation that would clarify the impact of hyperparameter tuning, by encouraging researchers to report expected validation performance —that is, how well would you expect a model to perform, given a fixed computational budget for hyperparameter tuning. In many cases, this requires no additional experimentation, and facilitates a more fair comparison of models.

This figure shows a simple application of our framework, comparing two classifiers (LR and CNN). Standard practice when comparing models is to train multiple instantiations of each, choose the best of each using validation data, and then evaluate them on a held-out test set. Unfortunately, this elides the effort that goes into finding the best instantiation of each type of model.

Here, we show that if researchers provide the same computational budget to each model, and make use of random search for hyperparameter tuning, then these validation results can be used to estimate the expected validation accuracy. Importantly, this shows that there is no such thing, in general, as “the best model”. Rather, there are models which have better or worse expected accuracy as a function of the budget, and this may change depending on whether the budget is large or small. Here, CNN has better maximum accuracy, but worse expected accuracy for a small budget. Full paper at arxiv.org/abs/1909.03004

Having pursued a somewhat idiosyncratic educational trajectory (from engineering to international development studies), I eventually returned to graduate school to earn a Ph.D. in Machine Learning. After graduating from Carnegie Mellon University in 2019, I came to Stanford as a postdoctoral scholar in computer science, specializing in computational social science. Here, I am primarily focused on two broad questions: first, what can we learn about people and society from the text that they produce; and second, how can we improve scientific practice, especially in fields such as data science. At present, I am working on projects related to the historical discourse around U.S. immigration, how issues such as climate change are framed in the media, and changing norms in scientific research and publication.

Fun facts about me:
I have lived in 9 different cities on 3 continents and traveled around the world (going West). • I am a big fan of bicycle touring, with a personal highlight being the trip from Pittsburgh to Washington, DC. • I love coffee so much that I made a short film about it.
In the field of medicine there are lots of questions that can be addressed using causal inference. In a previous research project, I sought to identify the sex-specific (i.e., male versus female) microbial signatures for fat distribution in humans. Although the gut microbiome has been linked to host obesity, the sex-specific association between microbiome and fat distribution are not well understood. After initial findings supported that the gut microbiome signature for fat distribution differs at both the species and genus level in men and women, a follow up study attempted to explain the potential mechanism by introducing metabolomic dynamics into the discovered association. The results are shown in Figure 1. The Association of Microbiome Abundance and Fat Ratio and Figure 2. Correlations and Gradient Effects Between Microbiome and Metabolites in Women and Men.

My current project is to examine: do surgeon and hospital volumes affect cardiothoracic surgery outcomes? To evaluate operative outcomes here, traditional randomized controlled trials are typically not applicable due to ethical and logistical reasons. So, we end up with a messy bulk of observational data from the Medicare program. Since I am attempting to compare the operative outcomes between high and low volume surgeons and high and low volume hospitals, I will need to carve out a clean two-by-two factorial design out of the highly confounded observational dataset. The studied surgical procedures and general study design are shown in Figure 3. This is still a study in progress. More to come!

This figure shows the global relationship between microbiome abundance and fat distribution in men and women. The top bar indicates the distribution of android and gynoid fat ratios within each microbiome abundance tertile. The blue gradient from light to dark encode the quartiles of the fat ratio from the lowest to the highest. The second bar represents the microbiome abundance tertiles. From light to dark purple, it indicates the lowest to the highest microbiome abundance tertiles. The heatmap contains the top 30 most abundant microbiome taxa in men and women ranked from the most abundance to the least bottom up; they compose the rows of the heatmap. Each column represents one sample. The color gradient indicates the abundance of a microbiome taxa in a particular sample. Panels (a) and (b) show the relationships between microbiome abundance and android fat ratio in women and men respectively.

What does my data look like?

The microbiome sequencing data is a string of text composed of A, U, G, C. Then after I processed the raw text data into the operational taxonomic units, I obtained a large contingency table with each cell of the table as the raw counts of the microbial abundance for the corresponding specific bacterial species and study subject. The metabolomic data is a large table composed of metabolomic concentrations. Thus, all the data is on a continuous scale. Medicare data is traditionally stored in a relational database. It is composed of several data tables, each table includes different variables on patient characteristics, insurance claims, provider’s information, etc. It is mostly categorical variables with very few exceptions, such as patient age, and hospitalization days.

Background image by Jeremy Bishop.
Hi, there! I am a 3rd year Ph.D. candidate in the Department of Epidemiology and Population Health focusing on causal inference methodology. I have completed my medical training and surgical residency at Peking University, and received a master’s degree in health economics from Stanford. After a gap year working as a policy analyst at the World Health Organization in Geneva, I discovered my interest in epidemiology and bioinformatics. Those brilliant study designs and translational methodologies have never failed to fascinate me! I have conducted several studies on human gut microbiome and body composition. My current research focuses on combining traditional causal inference theory with modern machine learning tools to answer large-scale causal questions in observational settings.

**Fun Facts About Me:**
- I am a mutant; part of my energy supply comes from photosynthesis.
- I love love LOVE adventures! Exploring the mountains, diving into the ocean, gazing at the stars—that sort of thing.
- In my spare time, I cook like a geek, dance like a cook, karate like a dancer.

The inner cords of the figures on the left indicate the correlations between the microbial taxa and metabolites that are both associated with fat distribution.

The outer spherical curves illustrate how much each microbial taxa and metabolite correspond to the fat distribution.

This figure presents the general study design of the causal inference project in the field of cardiothoracic surgery. The four surgical procedures included are aortic valve replacement, mitral valve replacement, coronary artery bypass graphing, and thoracoabdominal aorta aneurysm.

**Hospitals v.s Surgeons**

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital Volume</td>
<td>Surgeon Volume</td>
<td>Random</td>
</tr>
</tbody>
</table>
| High | Low | A
| Low | High | B
| High | Low | C
| Low | High | D

A × 2 factorial

Data: Medicare

AHAS

Physician Compare
I use satellite remote sensing to study agricultural water use. Water rights and agricultural water use is an important environmental and political issue in the Western US. I use physical simulations and statistical modeling to try to infer irrigation patterns from measurements of variables like land surface temperature and vegetation density. Direct measurements of irrigation schedules and volumes are generally not available, so it is common to use evapotranspiration (ET) as a proxy for the amount of water that was applied to a field. We can estimate ET using satellite measurements of variables like vegetation density and land surface temperature and combining that information with other weather and radiative forcing data. One of the main goals of my research is to more explicitly quantify and, where possible, reduce uncertainty in ET estimates from remote sensing models. I plan to do this using methods developed for “inverse problems,” where you develop a physical process-based forward model and then use statistical techniques to “invert” the model and identify the factors that could have produced your set of observations. In my case, the forward model tracks water and energy fluxes at the land surface. The inverse model then tries to identify the irrigation schedules and volumes that are consistent with a series of measurements taken by satellites.

The figure shows the estimated total ET for a region in Northern California over the month of July 2018. Each pixel represents an area of 30 meters by 30 meters. ET is expressed in millimeters (like precipitation is often expressed as a “depth”), but can be converted to a volume by multiplying the value at each pixel by its area. This model produces point estimates of ET over a given time scale, which is currently the norm among standard models. I’m working to develop methods for estimating a credible interval, rather than a single point estimate, that accounts for uncertainty in model inputs and the inherently limited information we have about land surface processes that develop over time.

I am a PhD student in Environmental Engineering. I studied math as an undergraduate, then worked at a database software company for a few years after graduating. Before beginning my PhD at Stanford, I completed a masters program in Public Policy and Management as Carnegie Mellon. I’m interested in the effects of climate change and water management policies (like surface water allocations and curtailments) on agricultural water use.

Fun facts about me:

• I’m a huge NBA fan.
• I enjoy running for exercise.
• My dog is the most popular pet in my apartment building.

Background image by Ivan Bandura.
One of my primary research interests is in the problem of property estimation, which has applications in many fields such as neuroscience, physics, and biology. In many of these applications, it is often very costly to collect data (e.g., observing all butterflies in a forest) and the goal is to estimate properties (functions) of an unknown probability distribution given a limited number of samples. Over a series of recent works, we constructed an efficient universal estimator (one estimator for all properties) that performs as well as any other property specific estimators. En route, we resolved an open problem in the field about efficiently computing the profile maximum likelihood distribution. We also showed many interesting results related to the Bethe/Sinkhorn approximations to the permanent, initially studied in statistical physics and with close connections to belief propagation algorithms in machine learning.

I am also interested in understanding nature from a computational lens. More specifically, I have been working on algorithmically modeling the behaviour of arboreal ants as limited computational agents performing reinforced random walks on graphs. I also conduct research in optimization (solving structured linear systems) and economics (modelling two sided matching markets).

Limited samples

Infer properties of whole population

One of my main research interests is in inferring properties of a probability distribution, given only a limited amount of data. For instance, consider the problem of estimating the number of distinct butterfly species in a forest. Suppose you observe only a few butterflies, many fewer than the total number of butterfly species, could you still estimate the total number of butterfly species? This corresponds to estimating the support size of a distribution given many fewer samples as compared to the support size.

In our work, we provide an efficient universal algorithm to estimate symmetric properties of a distribution that includes support size when given access only to a limited number of samples. Our algorithm is universal in the sense that it is independent of the property of interest and performs as well as any other property specific estimators.

Fun facts about me:

I play a variety of sports such as badminton, tennis, table tennis and frisbee. • In my spare time, I enjoy social dancing. • My favorite travel destination to date is Japan.