Monday, August 19

8:30 – 9:00 am Breakfast

9:00-9:40 am David Yang, Harvard University
Title: "Data, Autocracies, and the Direction of Innovation"

Abstract: Data is a key input for developing certain modern technologies, particularly in the realm of AI. Many of these technologies aim to predict human behaviors, which could greatly benefit the survival of autocratic regimes. How does modern autocracy shape data-intensive innovation and the direction of technological change? We provide a framework to model: (1) data in the technological production process, where data is a non-rival, excludable input whose mass collection is often considered unethical; and (2) political economic incentives to collect data and develop AI algorithms that sustain autocracy. We support this conceptual framework with consistent empirical evidence from China's world-leading AI facial recognition industry. We examine the facial recognition firms who receive data from the government by providing public security services, as well as the scientists who conduct facial recognition research.
(Joint with Martin Beraja and Noam Yuchtman)

9:40-10:20 am Abeer ElBahrawy, City, University of London
Title: "Coordinating in the Dark: the lives and deaths of Bitcoin Marketplaces"

Abstract: Dark markets are commercial websites which are accessible via darknets (e.g. Tor) and often specialize in selling drugs, weapons, and other illicit goods. Bitcoin is the standard currency for trading on dark markets. Recently, dark markets have seen a dramatic increase in their customer base and transaction volume. Multiple successful police raids and scams have shut down many of the largest dark markets. We analyze Bitcoin transactions network for 52 dark markets. We investigate the dynamics of 18 dark markets before and after dark market shutdowns. First, we show that users migrate quickly to other dark markets following shutdowns. Second, we describe the characteristics of migrant users. Finally, we study how migrant users coordinate on new dark markets following shutdowns. In this presentation, I will discuss these results in the context of our broader analysis of the Bitcoin ecosystem.

10:20 - 10:50 am Break

10:50-11:30 am Rediet Abebe, Cornell University
Title: "Impacting Policy Through Computational Approaches"

Abstract: Access to healthcare and health information is of major global concern. The stark inequality in the availability of health data by country, demographic groups, and socioeconomic status impedes the identification of major public health concerns and implementation of effective health policy. A key challenge is understanding health information needs of under-served and marginalized communities. Without understanding people's everyday needs, concerns, and misconceptions, health organizations lack the ability to effectively target education and programming efforts.

In this presentation, we focus on the lack of comprehensive, high-quality data about information needs of individuals in developing nations. We propose an approach that uses search data to uncover health information needs of individuals in all 54 nations in Africa. We analyze Bing searches related to HIV/AIDS, malaria, and tuberculosis; these searches reveal diverse health information needs that vary by demographic groups and geographic regions. We also shed light on discrepancies in the quality of content returned by search engines and discuss potential for using computationally-informed interventions to improve access to health information.

In the last part of the talk, we explore how to use a mix of techniques from algorithm and mechanism design as well as computational social science, along with insights from other disciplines, to inform policy in health and other domains and the Mechanism Design for Social Good initiative -- a vibrant and growing community of researchers and practitioners working towards improving access to opportunity for historically under-served and marginalized communities.

11:30-12:10 am Tymofiy Mylovanov, University of Pittsburgh
Title: "Detecting political disruption using voting patterns and media coverage: The case of Ukraine"

Abstract: We identify time periods of disruption in the voting patterns of the Ukrainian parliament for the last three convocations. We compare two methods: ideal point estimation (PolSci) and faction detection (CS). Both methods identify the revolution in Ukraine in 2014. The faction detection method also detects structural changes prior to the revolution (election of the president whose tenure was ended early by the revolution), while the ideal points method performs stronger after 2014, identifying a disruption around voting on constitutional changes to implement Minsk II agreements between separatists and Ukraine. The ideal point method is better at detecting position changes of the members of parliament, while the faction method is better at detecting changes in relationships between different MPs. The results suggest that after 2014, the Ukrainian parliament has become more consolidated, but the distribution of its political positions continues to evolve in response to changes in geo-political conditions. We then study TV media coverage of the political events in Ukraine by channels that are owned by different oligarchs. We track polarization of the parliament to polarization in the coverage by different TV channels, establishing a link between polarization in views of oligarchs and political polarization in the parliament.

12:10 – 1:30pm Lunch

1:30-2:10 pm Elaine O. Nsoesie, Boston University
Title: "Non-traditional Approaches to Public Health Surveillance"

Abstract: Data from a variety of sources, including social media, e-commerce websites and remote sensing, offer unique opportunities for studying and addressing problems in public health. In this talk, we will present examples on the use of data from a variety of sources for public health surveillance. We will also discuss the biases inherent in these datasets and potential implications on the public's health.

2:10-2:50 pm Joe Kileel, Princeton University
Title: "Mathematics for cryo-electron microscopy"

Abstract: Single particle cryo-electron microscopy (cryo-EM) is becoming an increasingly popular technique for determining three-dimensional molecular structures at high resolution (Nobel Prize in Chemistry 2017). We will discuss the mathematical principles for reconstruction from cryo-EM data. The talk will focus on computational challenges, in particular, reconstruction of small molecules and heterogeneity analysis.

2:50 – 3:20 pm Break

3:20-4:00 pm Aaron Roth, University of Pennsylvania
Title: "The Ethical Algorithm"

Abstract: Many recent mainstream media articles and popular books have raised alarms over anti-social algorithmic behavior, especially when driven by machine learning. The concerns include leaks of sensitive personal data by predictive models, algorithmic discrimination as a side-effect of machine learning, and inscrutable decisions made by complex models. While standard and legitimate responses to these phenomena include calls for better laws and regulations, researchers in machine learning, statistics and related areas are also working on designing better-behaved algorithms. An explosion of recent research in areas such as differential privacy, algorithmic fairness and algorithmic game theory is forging a new science of socially aware algorithm design. I will survey these developments and attempt to place them in a broader societal context. This talk is based on the forthcoming book "The Ethical Algorithm", co-authored with Michael Kearns and available for pre-order.

4:00-4:40pm Andrew Lo, Massachusetts Institute of Technology
Title: "How Machine-Learning Predictions of Clinical Trial Outcomes Can Help Cure Cancer"

Abstract: Funding for early-stage biomedical innovation has become scarcer even as breakthroughs in our understanding of the biology of human disease are occurring more frequently. The reason for this "Valley of Death" in translational medicine has to do with the increasing risk and uncertainty of drug discovery and development. To counteract this trend, we apply machine-learning techniques to predict drug approvals using drug-development and clinical-trial data from 2003 to 2015 involving several thousand drug-indication pairs with over 140 features across 15 disease groups. Imputation methods are used to deal with missing data, allowing us to fully exploit the entire dataset, the largest of its kind. We show that our approach outperforms complete-case analysis, which typically yields biased inferences. We achieve predictive measures of 0.78, and 0.81 AUC ("area under the receiver operating characteristic curve" is the estimated probability that a classifier will rank a positive outcome higher than a negative outcome) for predicting transitions from phase 2 to approval and phase 3 to approval, respectively. Using five-year rolling windows, we document an increasing trend in the predictive power of these models, a consequence of improving data quality and quantity. The most important features for predicting success are trial outcomes, trial status, trial accrual rates, duration, prior approval for another indication, and sponsor track records. We provide estimates of the probability of success for all drugs in the current pipeline. By providing investors with better tools for assessing the risk and reward of drug development, we hope to attract more capital to this critical sector, reduce the cost of capital, and bring new and better therapeutics to patients faster.

Tuesday, August 20

8:30 - 9:00 am Breakfast

9:00-9:40 am Jörn Boehnke, University of California, Davis
Title: "Identifying substitution patterns using product reviews"

Abstract: Determining which products are close competitors in a differentiated product market is a key question in marketing and industrial organization. The closeness of competition between any two products is determined by the degree of consumer substitutability between them and directly informs firm's optimal price setting. However, estimating demand and thus the substitutability in online marketplaces is becoming computationally complex and not easily scalable due to the large number of products and not easily identifiable relevant product characteristics. We argue that consumer reviews present a promising, free alternative for analyzing product similarity and ultimately predicting substitution patterns. The key idea is that the co-occurrence of words and the sentiment with which the products are described in customer reviews for any two products is a strong predictor of their substitutability. Specifically, we propose an unsupervised machine-learning approach to determine product similarity by analyzing 1.8 million of Amazon product reviews across 14 different product categories. We first generate word embeddings by training Word2vec – a shallow, two-layer neural networks – on the corpus of all consumer reviews in each product category. The resulting vector space represents the distributed numerical representations of the word features in these product reviews. We subsequently analyze the words used in a product's reviews to position the product in the product vector space. The resulting Euclidian distance between any two products approximates the degree of substitutability between them. Finally, using daily price and sales rank data within each category, we estimate product demand and thus own- and cross-price elasticities. We show that our measures of Euclidian distances generated from product reviews predict estimated cross price elasticities. Our proposed method provides an alternative and easily accessible approach to estimating product substitutability and price sensitivity from consumer generated content alone, without requiring any sales data.

9:40-10:20 am Pablo Azar, Massachusetts Institute of Technology
Title: "Endogenous Production Networks"

Abstract: We develop a tractable model of endogenous production networks. Each one of a number of products can be produced by combining labor and an endogenous subset of the other products as inputs. Different combinations of inputs generate (prespecified) levels of productivity and various distortions may affect costs and prices. We establish the existence and uniqueness of an equilibrium and provide comparative static results on how prices and endogenous technology/input choices (and thus the production network) respond to changes in parameters. These results show that improvements in technology (or reductions in distortions) spread throughout the economy via input-output linkages and reduce all prices, and under reasonable restrictions on the menu of production technologies, also lead to a denser production network. Using a dynamic version of the model we establish that the endogenous evolution of the production network could be a powerful force towards sustained economic growth. At the root of this result is the fact that the arrival of a few new products expands the set of technological possibilities of all existing industries by a large amount — that is, if there are n products, the arrival of one more new product increases the combinations of inputs that each existing product can use from 2^{n−1} to 2^{n} , thus enabling significantly more pronounced cost reductions from choice of input combinations. These cost reductions then spread to other industries via lower input prices and incentivize them to also adopt additional inputs.

10:20 - 10:50 am Break

10:50-11:30 am Sarah Brown, Brown University
Title: "Wiggum: Simpson's Paradox Inspired Fairness Forensics"

Abstract: Both technical limitations and social critiques of current approaches to fair machine learning motivate the need for more collaborative, human driven approaches. We aim to empower human experts to conduct more critical exploratory analyses and support informal audits. I will present Wiggum, a data exploration package with a visual analytics interface for Simpson's paradox inspired fairness forensics. Wiggum detects and ranks multiple forms of Simpson's Paradox and related relaxations.

11:30-12:10 am David Gamarnik, Massachusetts Institute of Technology
Title: "Algorithmic Challenges in High-Dimensional Inference Models. Insights from Statistical Physics"

Abstract: Inference problems arising in modern day statistics, machine learning and artificial intelligence fields often involve models with exploding dimensions, giving rise to a multitude of computational challenges. Many such problems "infamously" resist the construction of tractable inference algorithms, and thus are possibly fundamentally non-solvable by fast computational methods. A particularly intriguing form of such intractability is the so-called computational vs information theoretic gap, where effective inference is achievable by some form of exhaustive search type computational procedure, but fast computational methods are not known and conjectured not to exist. A great deal of insight into the mysterious nature of this gap has emerged from the field of statistical physics, where the computational difficulty is linked to a phase transition phenomena of the solution space topology. We will discuss one such phase transition obstruction, which takes the form of the Overlap Gap Property: the property referring to the topological disconnectivity (gaps) of the set of valid solutions.

12:10 – 1:30pm Lunch

1:30-2:10 pm Charles Epstein, University of Pennsylvania, AMCS
Title: "Geometry of the Phase Retrieval Problem"

Abstract: In several imaging modalities the measured data can be interpreted as the modulus of the Fourier transform of a function describing the unknown object. To reconstruct this object one needs to use some auxiliary information to recover the un-measured phase of the Fourier transform. This is a notoriously difficult problem. I will discuss the underlying geometric reasons for these difficulties, approaches to improving the performance of standard algorithms, as well as entirely new approaches to this problem.

2:10-2:50 pm Anna Gilbert, University of Michigan
Title: "Sparse Metric Repair"

Abstract: Suppose we are given a distance or similarity matrix for a data set that is corrupted in some fashion, find a sparse correction or repair to the distance matrix so as to ensure the corrected distances come from a metric; i.e., repair as few entries as possible in the matrix so that we have a metric. I will discuss generalizations to graph metrics, applications to (and from) metric embeddings, and algorithms for variations of this problem. I will also touch upon applications in machine learning and bio-informatics.

2:50 – 3:20 pm Break

3:20-4:00 pm Zhiping Weng, University of Massachusetts Medical School
Title: "An ENCODE Registry of Candidate cis-Regulatory Elements for Human and Mouse"

Abstract: Comprehensive maps of functional elements are critical for understanding how a genome specifies cell and tissue types and assessing how genomic variants affect development, aging, and disease susceptibility. The goal of the Encyclopedia of DNA Elements (ENCODE) Project is to discover and characterize the full repertoire of functional elements in the human and mouse genomes. An essential component of the repertoire is cis-regulatory elements, which orchestrate transcriptional regulation. Here, we introduce a new registry of candidate cis-Regulatory Elements (cCREs), defined by a biochemical signature that uses chromatin accessibility, histone modification, and transcription factor occupancy data derived from assay-specific and integrative analyses of ENCODE and Roadmap Epigenomics Consortium data. The registry currently contains 1.5 million human and 0.5 million mouse cCREs and their candidate functions in 618 human and 138 mouse biosamples. We tested 151 cCREs with transgenic mouse assays and the validation rates ranged from 75% for highly ranked cCREs to 30% for lower ranked ones. The cCRE landscape recapitulates current understandings of cellular identity, tissue composition, developmental progression, and disease-associated genetic variants. Aided by a dedicated visualization engine called SCREEN, the registry is a resource for exploring noncoding DNA elements and their variants.

4:00-4:40pm X. Shirley Liu, Dana-Farber Cancer Institute & Harvard T.H. Chan School of Public Health
Title: "Hidden immunology signals from tumor RNA-seq data"