DATA SCIENCE INSTITUTE
COLUMBIA UNIVERSITY

Data For Good:

Ensuring the Responsible Use of Data to Benefit Society

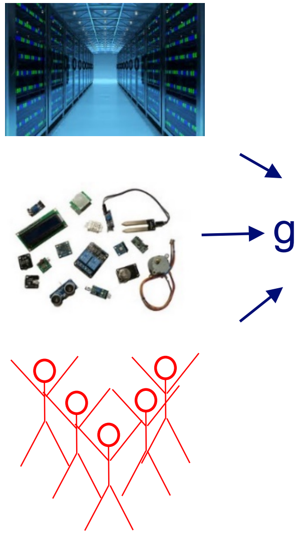
Jeannette M. Wing

Avanessians Director of the Data Science Institute and Professor of Computer Science
Columbia University

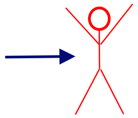
Adjunct Professor of Computer Science
Carnegie Mellon University

Symposium on Data Science and Statistics
Virtual Pittsburgh, PA
June 5, 2020

Data Life Cycle



→ generation → collection → processing → storage → management → analysis → visualization → interpretation →



privacy and ethical concerns throughout

What is Data Science?

Definition:

Data science is the study of extracting value from data.

Mission

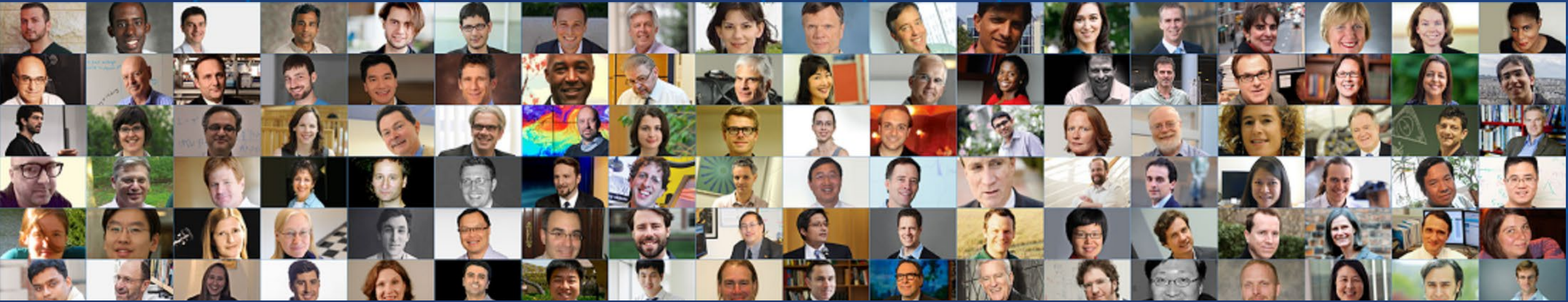
Advance the state of the art in data science

Transform all fields, professions, and sectors through the application of data science

Ensure the responsible use of data to benefit society

Tagline

Data for Good



DATA SCIENCE INSTITUTE
COLUMBIA UNIVERSITY

17 Schools, Colleges, and Institutes

Graduate School of Architecture, Planning and Preservation
School of the Arts
Graduate School of Arts and Sciences
Barnard College
Columbia Business School

College of Dental Medicine
The Earth Institute
Columbia Engineering
School of International and Public Affairs
Columbia Journalism School
Columbia Law School

School of Nursing
Vagelos College of Physicians and Surgeons
Mailman School of Public Health
School of Social Work
Teachers College
Zuckerman Institute

Cross-Cutting Centers

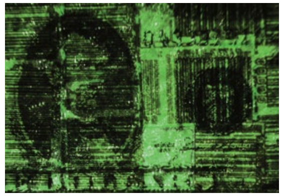
datascience.columbia.edu/data-science-centers

Foundations

Computing Systems

Cybersecurity

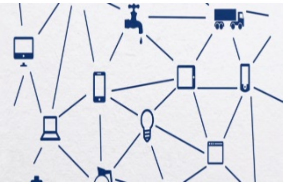
Data, Media, and Society



Financial Analytics

Health Analytics

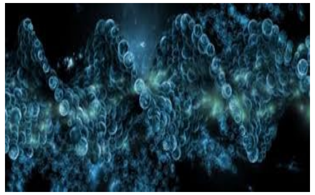
Smart Cities



Sense, Collect, and Move

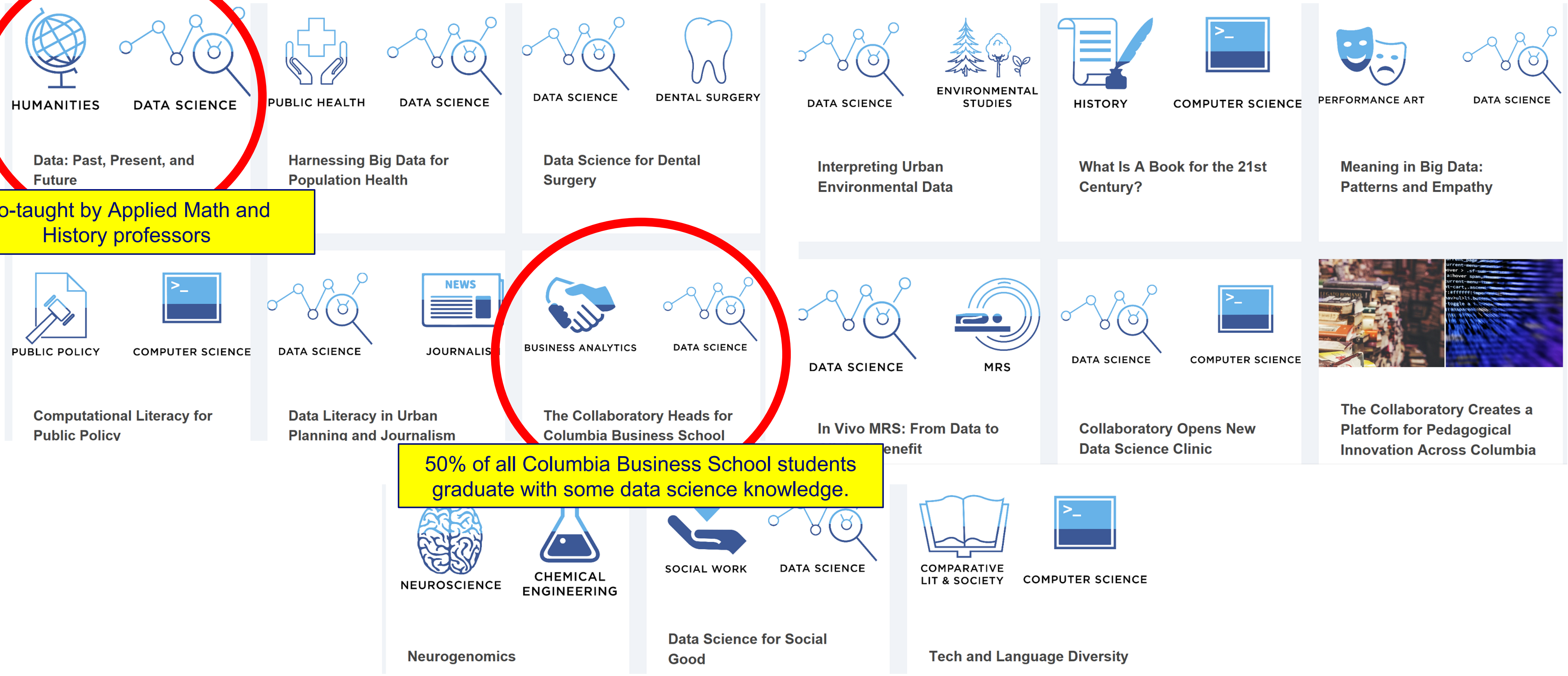
Computational Social Science

Education



Materials Discovery Analytics

Collaboratory
(Columbia Entrepreneurship + DSI)



HUMANITIES · DATA SCIENCE

Data: Past, Present, and Future

Co-taught by Applied Math and History professors

PUBLIC HEALTH · DATA SCIENCE

Harnessing Big Data for Population Health

DATA SCIENCE · DENTAL SURGERY

Data Science for Dental Surgery

DATA SCIENCE · ENVIRONMENTAL STUDIES

Interpreting Urban Environmental Data

HISTORY · COMPUTER SCIENCE

What Is A Book for the 21st Century?

PERFORMANCE ART · DATA SCIENCE

Meaning in Big Data: Patterns and Empathy

PUBLIC POLICY · COMPUTER SCIENCE

Computational Literacy for Public Policy

DATA SCIENCE · JOURNALIS

Data Literacy in Urban Planning and Journalism

BUSINESS ANALYTICS · DATA SCIENCE

The Collaboratory Heads for Columbia Business School

50% of all Columbia Business School students graduate with some data science knowledge.

DATA SCIENCE · MRS

In Vivo MRS: From Data to enefit

DATA SCIENCE · COMPUTER SCIENCE

Collaboratory Opens New Data Science Clinic

The Collaboratory Creates a Platform for Pedagogical Innovation Across Columbia

NEUROSCIENCE · CHEMICAL ENGINEERING

Neurogenomics

SOCIAL WORK · DATA SCIENCE

Data Science for Social Good

COMPARATIVE LIT & SOCIETY · COMPUTER SCIENCE

Tech and Language Diversity

Industry Affiliates Program

industry.datascience.columbia.edu

Alibaba Group

BOSE

aws

Baidu 百度

Bloomberg

Capital One

DiDi

EY Building a better working world

ELSEVIER

GE

Goldman Sachs

Google Cloud

Johnson & Johnson

J.P.Morgan

KPMG

LUCA Telefónica DATA UNIT

MediaMath

Microsoft Research

MOELIS & COMPANY

MOODY'S

Neoway

RALPH LAUREN

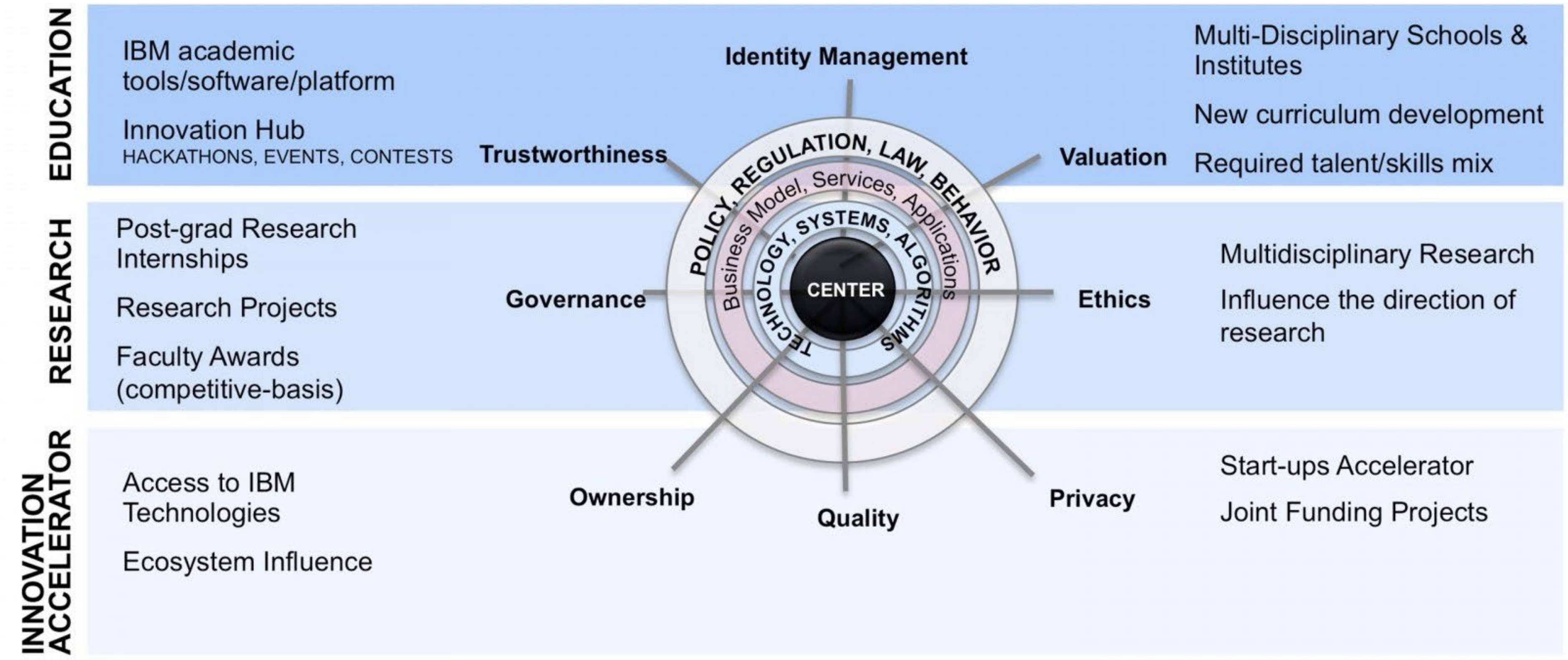
TDF VENTURES

2σ TWO SIGMA

Unilever

Vanguard

Columbia-IBM Center on Blockchain and Data Transparency



Identity Management
Trustworthiness
Valuation
Governance
Ethics
Ownership
Quality
Privacy
POLICY, REGULATION, LAW, BEHAVIOR
Business Model, Services, Applications
TECHNOLOGY, SYSTEMS, ALGORITHMS
CENTER

EDUCATION

IBM academic tools/software/platform

Innovation Hub
HACKATHONS, EVENTS, CONTESTS

Multi-Disciplinary Schools & Institutes

New curriculum development

Required talent/skills mix

RESEARCH

Post-grad Research Internships

Research Projects

Faculty Awards (competitive-basis)

Multidisciplinary Research

Influence the direction of research

INNOVATION ACCELERATOR

Access to IBM Technologies

Ecosystem Influence

Start-ups Accelerator

Joint Funding Projects

Mission

Advance the state of the art in data science

Transform all fields, professions, and sectors through the application of data science

Ensure the responsible use of data to benefit society

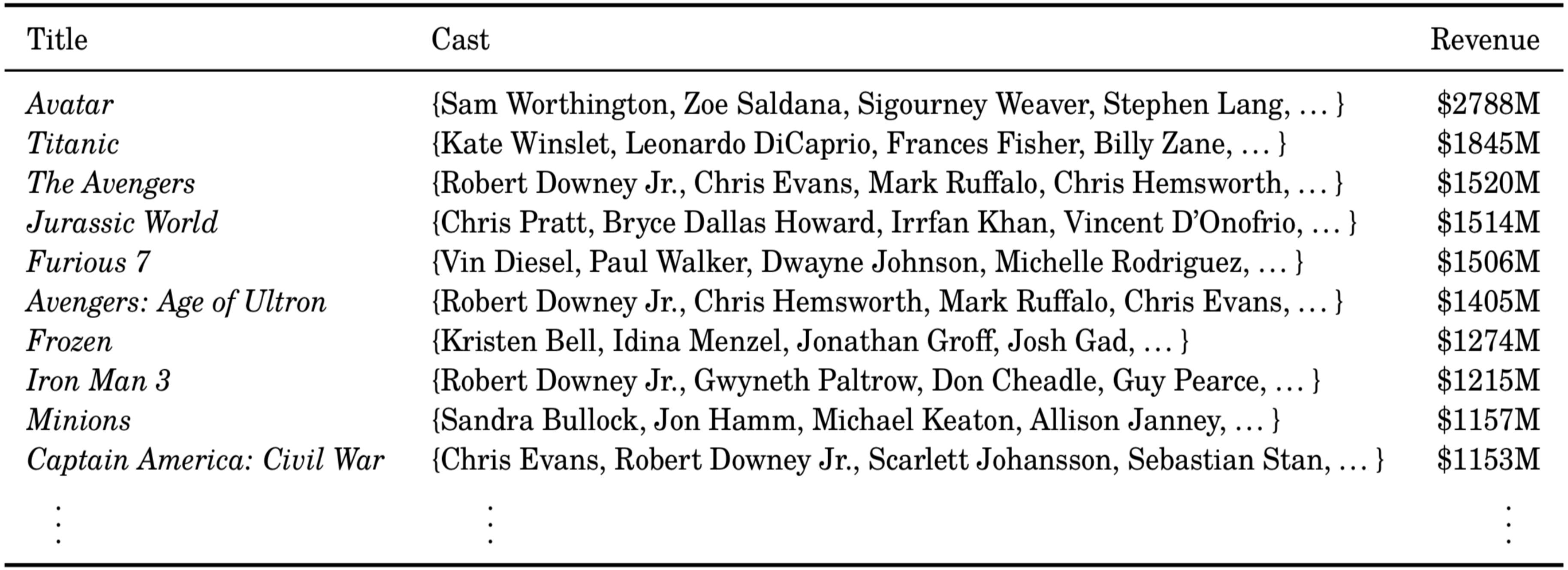
Multiple Causal Inference

Yixin Wang and David M. Blei, “The Blessings of Multiple Causes,” arXiv:1805.06826v2 [stat.ML], June 19, 2018.

Understanding Causal Effect

What happens to movie revenue **if** we place an actor in a movie**?**

$\boxed{\text{Goal: } E[Y_i(\boldsymbol{a})]}$ $E[Y_i \mid do(\boldsymbol{a})]$



Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, … }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, … }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, … }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, … }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, … }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, … }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, … }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, … }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, … }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, … }	$1153M
⋮	⋮	⋮

Many Applications

Classical Causal Inference

THINK ABOUT CONFOUNDERS

MEASURE CONFOUNDERS

$\{w_1, \dots, w_n\}$

ESTIMATE CAUSAL EFFECTS

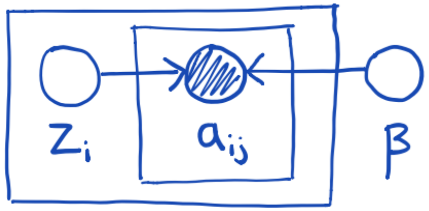
$$\mathbb{E}[Y_i(a)] = \mathbb{E}[\mathbb{E}[Y_i(A_i) | A_i = a, W_i]]$$

Strong ignorability:
No unobserved confounders

- **Confounders** affect both the causes and the outcomes.
- We should correct for all confounders in causal inference, which requires in theory to measure **all confounders**.
- But, whether we have measured all confounders is (famously) **untestable.**

New Idea: The Deconfounder

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \dots, \hat{z}_n\}$$

$$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$$

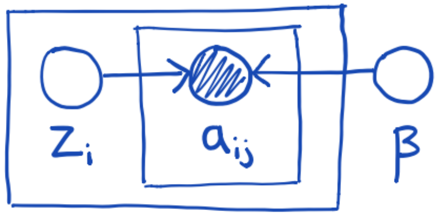
ESTIMATE CAUSAL EFFECTS

$$\mathbb{E}[Y_i(a)] = \mathbb{E}[\mathbb{E}[Y_i(A_i) \mid A_i = a, Z_i]]$$

1. **Fit** a “local latent-variable model” of the assigned causes (e.g., Factor Analysis).
2. **Infer** the latent variable for each data point; it is a substitute confounder.
3. **Correct** for the substitute confounder in a causal inference.

New Idea: The Deconfounder

MODEL ASSIGNED CAUSES



ESTIMATE SUBSTITUTE CONFOUNDERS

$$\{\hat{z}_1, \ldots, \hat{z}_n\}$$

$$\hat{z}_i = \mathbb{E}[Z_i \mid A_i = a_i]$$

ESTIMATE CAUSAL EFFECTS

$$\mathbb{E}[Y_i(a)] = \mathbb{E}[\mathbb{E}[Y_i(A_i) \mid A_i = a, Z_i]]$$

Assumption:
No unobserved single-cause confounder

Weaker assumptions: No unobserved single-cause confounder.
(But no need to measure all confounders.)

Checkable procedure: We can check if the substitute confounder is good.

Unbiased inference: We prove the deconfounder gives unbiased causal inference.

Back to Movies

- With the deconfounder,
  (1) Sean Connery's (James Bond) value goes up.
  (2) Bernard Lee's (M) and Desmond Llewelyn's (Q) values go down.

- We can now answer questions such as: What happens to revenue if we place Desmond Llewelyn in *A Beautiful Mind*? How about Sean Connery?

- The deconfounder **corrects for unobserved confounders**: genre, sequel, etc.

Advance the state of the art in data science

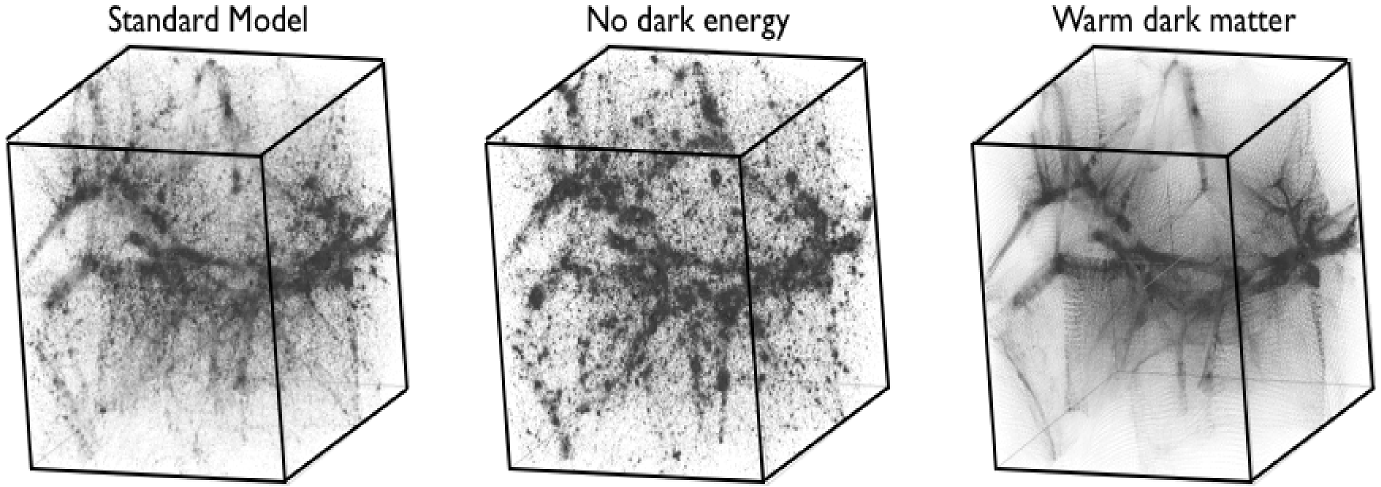
Transform all fields, professions and sectors through the application of data science

Ensure the responsible use of data to benefit society

Biology and Big Data:
Understanding Tumor Microbiome to Combat Cancer

Geller, L.*, Barzily-Rokni, M.*, Danino, T., Shee, K., Thaiss, C., Livny, R., Avraham, R., Barczak, A., Zwang, Y., Mosher, C., Smith, D., Chatman, K., Skalak, M., Bu, J., Cooper, Z., Tompers, F., Ligorio, M., Qian, Z., Muzumdar, M., Michaud, Gurbatri, C., M., Mandinova, A., Garrett, W., Jacks, T., Ogino, S., Ferrone, C., Thayer, S., Warger, J., Trauger, S., Johnston, S., Huttenhower, C., Gevers, D., Bhatia, S., Golub, T. Straussman, R. Tumor-microbiome mediated resistance to gemcitabine. *Science* 357, 1156–1160 (2017).

Cosmology and Neural Networks



Standard Model

No dark energy

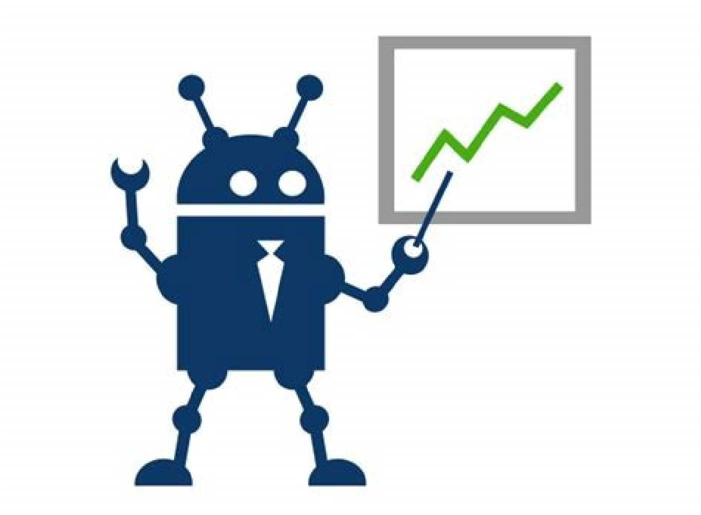
Warm dark matter

Dezso Ribli, Balint Armin Pataki, Jose Manuel Zorrilla Matilla, Daniel Hsu, Zoltan Haiman, Istvan Csabai, “Weak lensing cosmology with convolutional neural networks on noisy data,” Monthly Notices of the Royal Astronomical Society, Volume 490, Issue 2, December 2019, pp. 1843-1860.

Monopsony:
Economics and Machine Learning

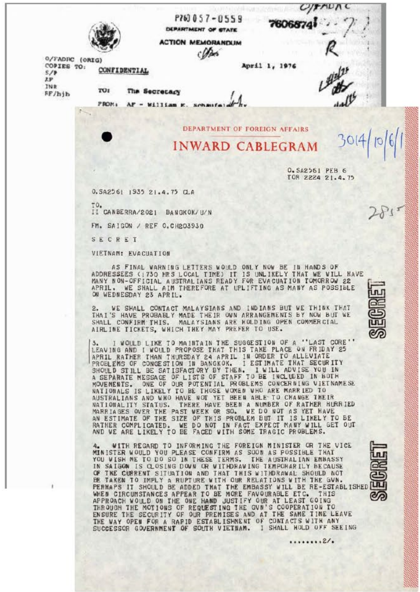
Arindrajit Dube, Jeff Jacobs, Suresh Naidu, and Siddharth Suri, “Monopsony in Online Labor Markets,” forthcoming, *American Economic Review: Insights,* August 2018.

Robo-Advising:
Finance and Reinforcement Learning



Agostino Capponi, Octavio Ruiz Lacedelli, and Matt Stern, “Robo-Advising as a Human-Machine Interaction System”, August 2018, preprint.

Event Discovery:
History and Topic Modeling



O/FADRC

P760057-0559

DEPARTMENT OF STATE

7606874

ACTION MEMORANDUM

R

O/FADRC (ORIG)
COPIES TO:
S/P
AF
INR
RF/hjb

CONFIDENTIAL

April 1, 1976

TO: The Secretary

FROM: AF - William E. Schaufele

DEPARTMENT OF FOREIGN AFFAIRS

INWARD CABLEGRAM

3014/10/6/1

O.SA2561 PEB 6
TOR 2224 21.4.75

O.SA2561 1935 21.4.75 CLA

TO.
II CANBERRA/2021 BANGKOK/U/N

2815

FM. SAIGON / REF O.CH203930

S E C R E T

VIETNAM: EVACUATION

AS FINAL WARNING LETTERS WOULD ONLY NOW BE IN HANDS OF ADDRESSEES (1730 HRS LOCAL TIME) IT IS UNLIKELY THAT WE WILL HAVE MANY NON-OFFICIAL AUSTRALIANS READY FOR EVACUATION TOMORROW 22 APRIL. WE SHALL AIM THEREFORE AT UPLIFTING AS MANY AS POSSIBLE ON WEDNESDAY 23 APRIL.

2. WE SHALL CONTACT MALAYSIANS AND INDIANS BUT WE THINK THAT THAI'S HAVE PROBABLY MADE THEIR OWN ARRANGEMENTS BY NOW BUT WE SHALL CONFIRM THIS. MALAYSIANS ARE HOLDING OPEN COMMERCIAL AIRLINE TICKETS, WHICH THEY MAY PREFER TO USE.

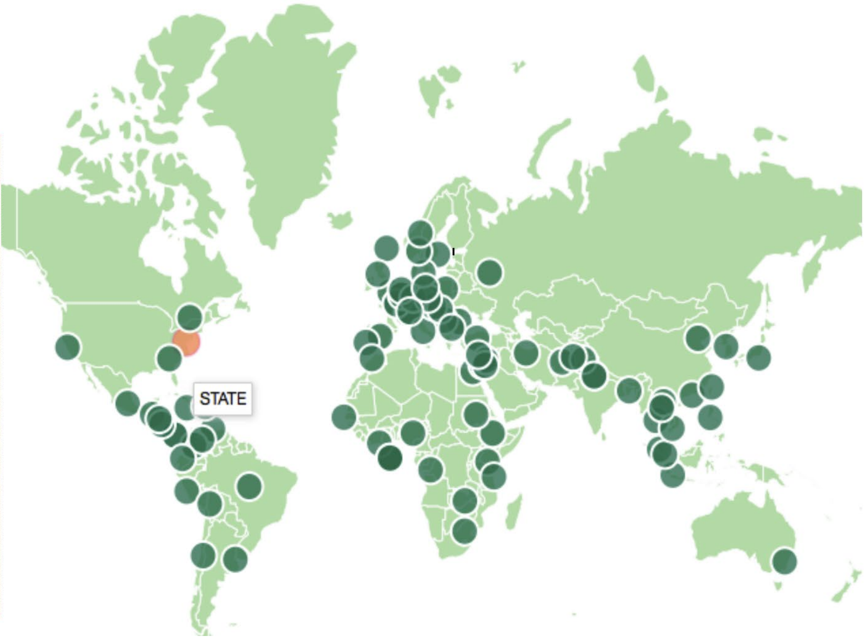
3. I WOULD LIKE TO MAINTAIN THE SUGGESTION OF A ''LAST CORE'' LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLEVIATE PROBLEMS OF CONGESTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN A SEPARATE MESSAGE OF LISTS OF STAFF TO BE INCLUDED IN BOTH MOVEMENTS. ONE OF OUR POTENTIAL PROBLEMS CONCERNING VIETNAMESE NATIONALS IS LIKELY TO BE THOSE WOMEN WHO ARE MARRIED TO AUSTRALIANS AND WHO HAVE NOT YET BEEN ABLE TO CHANGE THEIR NATIONALITY STATUS. THERE HAVE BEEN A NUMBER OF RATHER HURRIED MARRIAGES OVER THE PAST WEEK OR SO. WE DO NOT AS YET HAVE AN ESTIMATE OF THE SIZE OF THIS PROBLEM BUT IT IS LIKELY TO BE RATHER COMPLICATED. WE DO NOT IN FACT EXPECT MANY WILL GET OUT AND WE ARE LIKELY TO BE FACED WITH SOME TRAGIC PROBLEMS.

SECRET

4. WITH REGARD TO INFORMING THE FOREIGN MINISTER OR THE VICE MINISTER WOULD YOU PLEASE CONFIRM AS SOON AS POSSIBLE THAT YOU WISH ME TO DO SO IN THESE TERMS. THE AUSTRALIAN EMBASSY IN SAIGON IS CLOSING DOWN OR WITHDRAWING TEMPORARILY BECAUSE OF THE CURRENT SITUATION AND THAT THIS WITHDRAWAL SHOULD NOT BE TAKEN TO IMPLY A RUPTURE WITH OUR RELATIONS WITH THE GVN. PERHAPS IT SHOULD BE ADDED THAT THE EMBASSY WILL BE RE-ESTABLISHED WHEN CIRCUMSTANCES APPEAR TO BE MORE FAVOURABLE ETC. THIS APPROACH WOULD ON THE ONE HAND JUSTIFY OUR AT LEAST GOING THROUGH THE MOTIONS OF REQUESTING THE GVN'S COOPERATION TO ENSURE THE SECURITY OF OUR PREMISES AND AT THE SAME TIME LEAVE THE WAY OPEN FOR A RAPID ESTABLISHMENT OF CONTACTS WITH ANY SUCCESSOR GOVERNMENT OF SOUTH VIETNAM. I SHALL HOLD OFF SEEING

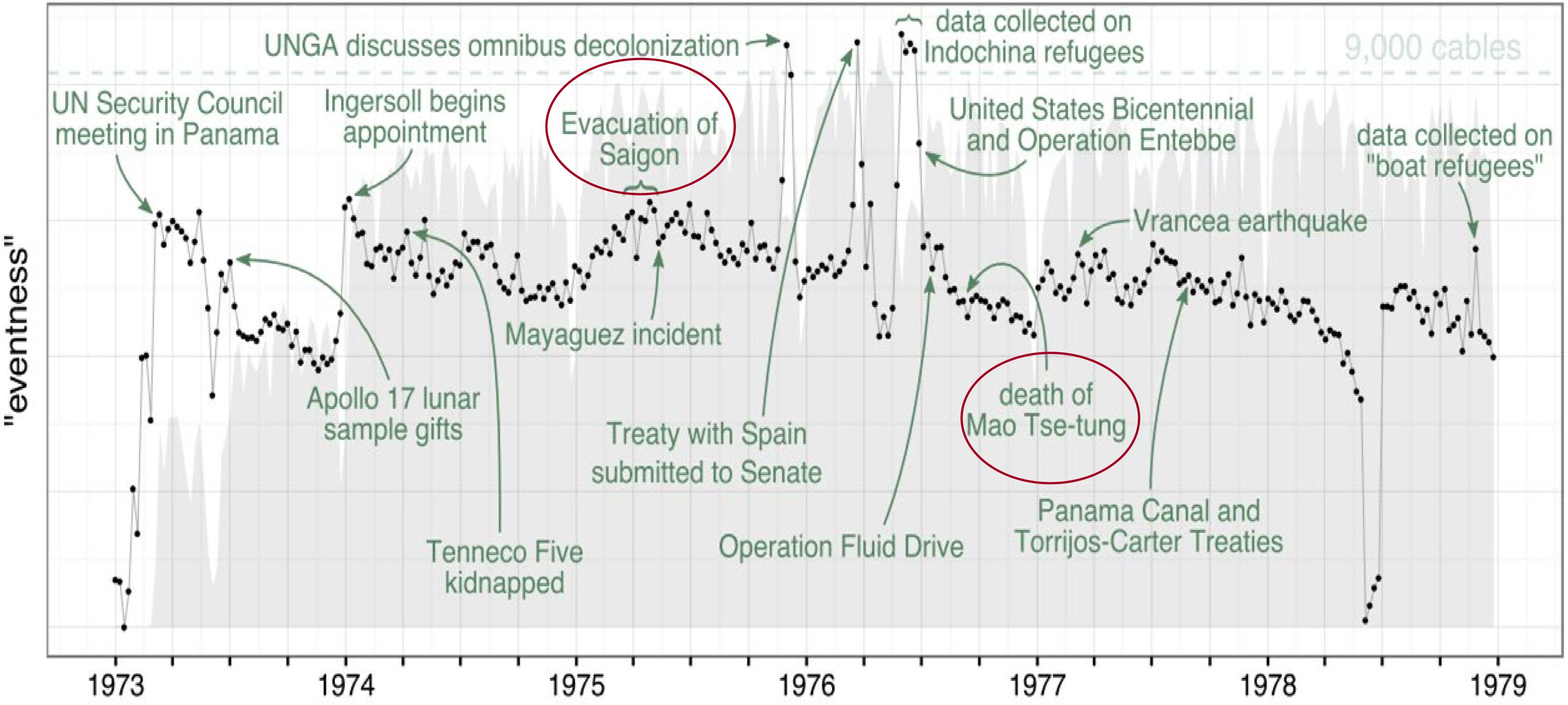
SECRET

........2/.



Allison J. B. Chaney, Hanna Wallach, **Matthew Connelly**, and **David M. Blei,** Detecting and characterizing Events, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, November 2016.

Distinguish between topics describing "business as usual" and those that deviate from such patterns.



Data for Good:

responsible use of data

FAT* → Trustworthy AI

Fairness

Accountability

Transparency

Ethics

Robustness

Interpretability/Explainability

➡ Safety

Security

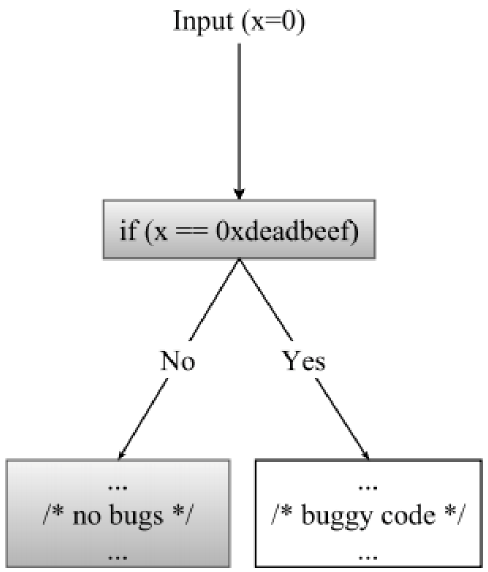
Privacy

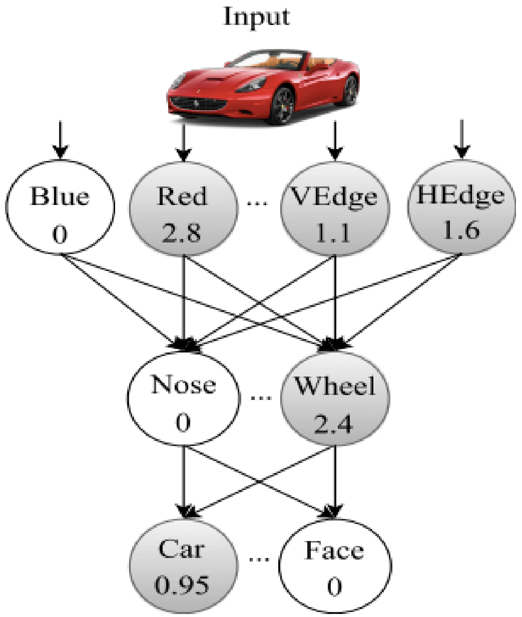
Reliability

Availability

Usability

DeepXplore: Testing Deep Learning Systems

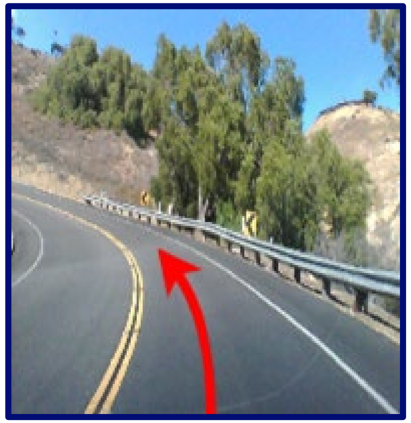




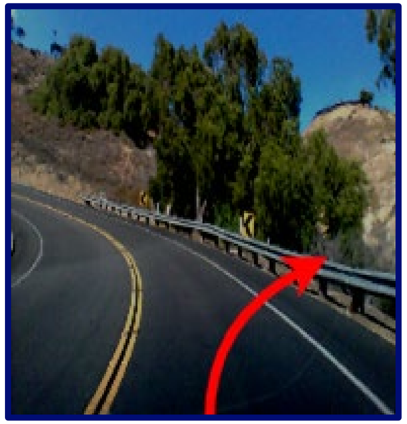
Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana, “Deep Xplore: Automated Whitebox Testing of Deep Learning Systems, *Proceedings of the 26th ACM Symposium on Operating Systems Principles*, October 2017, Best Paper Award.

DeepXplore

https://github.com/peikexin9/deepxplore



Seed,
No accident

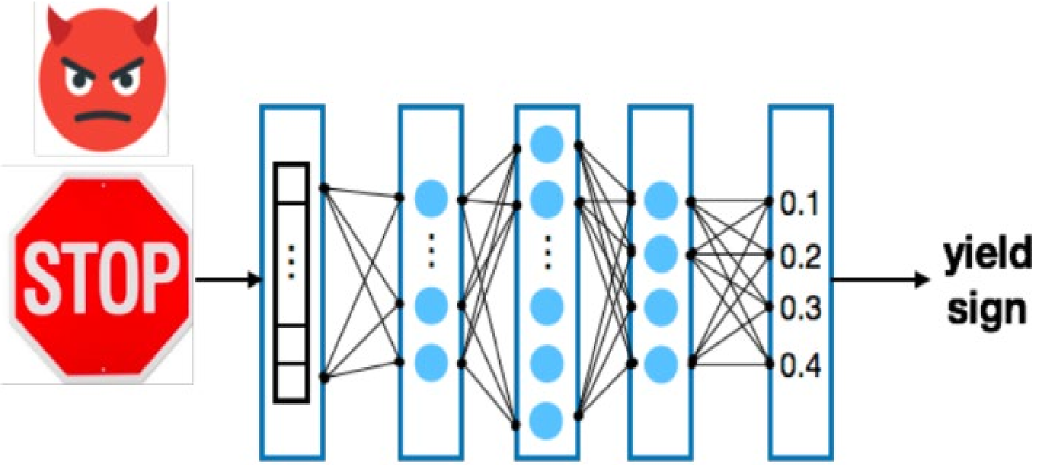


Darker,
Accident

- Efficiently and systematically tests DNNS of hundreds of thousands of neurons without labeled data (only needs unlabeled seeds)
- Key ideas: neuron coverage (akin to code coverage), differential testing, and domain-specific constraints for focusing on realistic inputs
- Testing as a joint optimization problem (maximize both number of differences and neuron coverage)
- Found 1000s of fatal errors in 15 state-of-the-art DNNs for ImageNet, self-driving cars, and PDF/Android malware

DP and Machine Learning: PixelDP

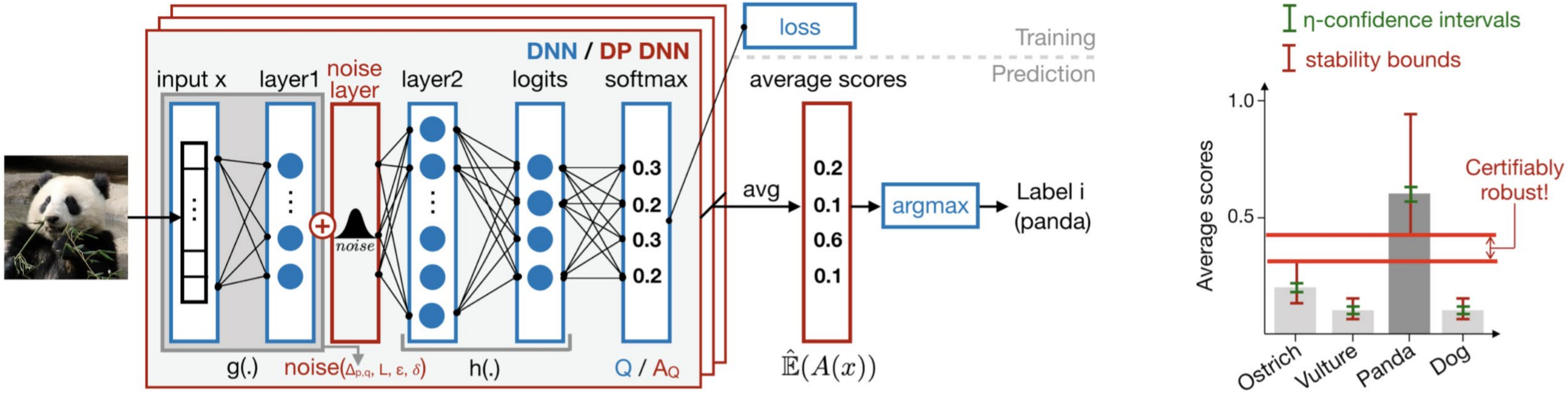
Problem



Mathias Lecuyer, Baggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana, “Certified Robustness to Adversarial Examples with Differential Privacy, arXiv:1802.03471v2 , June 26, 2018, to appear IEEE Security and Privacy (‘’Oakland’’) 2019.

Solution

1. Add a noise layer a la Differential Privacy



(a) PixelDP DNN Architecture

(b) Robustness Test Example

2. Provable guarantee from DP says classifier is robust to some degree of input perturbations.

Data for Good:
tackling societal grand challenges

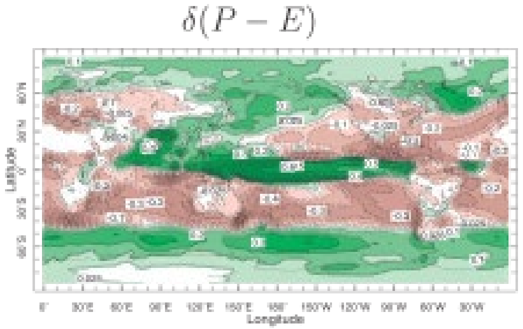
PANGEO: Climate Science and Big Data

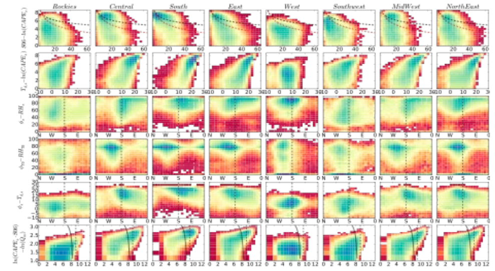
https://pangeo-data.github.io/

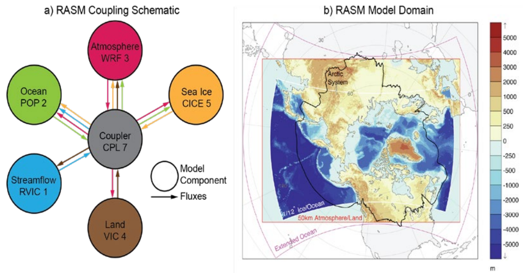
PI: Ryan Abernathey
(Dept. of Earth & Env. Sci., LDEO, Columbia University)

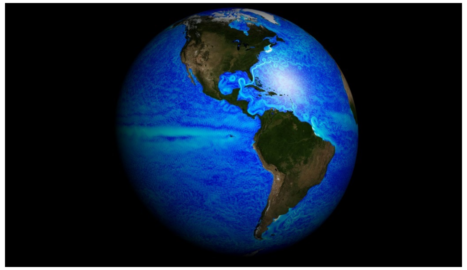
Co-PIs: Chiara Lepore, Michael Tippett, Naomi Henderson, Richard Seager (LDEO)
Kevin Paul, Joe Hamman, Ryan May, Davide Del Vento (National Center for Atmospheric Research)
Matthew Rocklin (Anaconda; formerly Continuum Analytics)

Collaborators: Gavin Schmidt (APAM, Frontiers in Computing Systems (DSI), NASA Goddard Institute for Space Studies (director), V. Balaji (National Oceanographic and Atmospheric Administration Geophysical Fluid Dynamics Lab)



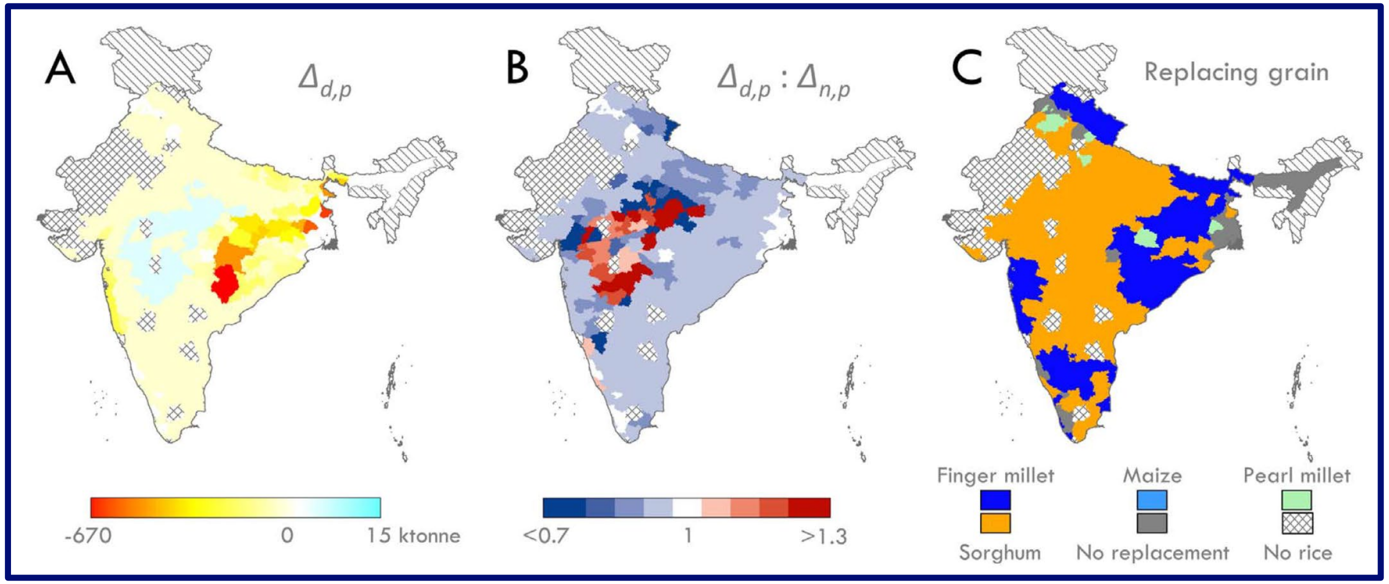






EarthCube
Anaconda Powered by Continuum Analytics
Data Science Institute Columbia University
Lamont-Doherty Earth Observatory
Columbia University | Earth Institute
NCAR National Center for Atmospheric Research
NSF

Data Science and Agriculture



Kyle F. Davis, Ashwini Chhtre, Narasimha D. Rao, Deepti Singh, Ruth DeFries, *Environmental Research Letters,* Volume: 14, Article number: 064013, https://doi.org/10.1088/1748-9326/ab22db

Main Results

Picture from The Economic Times, June 18, 2019

- If India’s crop production continues to homogenize towards rice, food supply in the country may be more vulnerable to increasingly frequent climate shocks (e.g., droughts, extreme heat).

- Increasing the share of production contributed by coarse cereals (such as millets and sorghum) could improve the resilience of India’s food production against climatic changes, especially in the places where coarse cereal yields are already comparable to rice yields.

- More broadly, diversifying crop mixes in agriculturally important areas can help buffer against some aspects of climate change such as droughts and extreme heat.

Healthcare: Observational Health Data Sciences and Informatics (OHDSI, pronounced “Odyssey”)

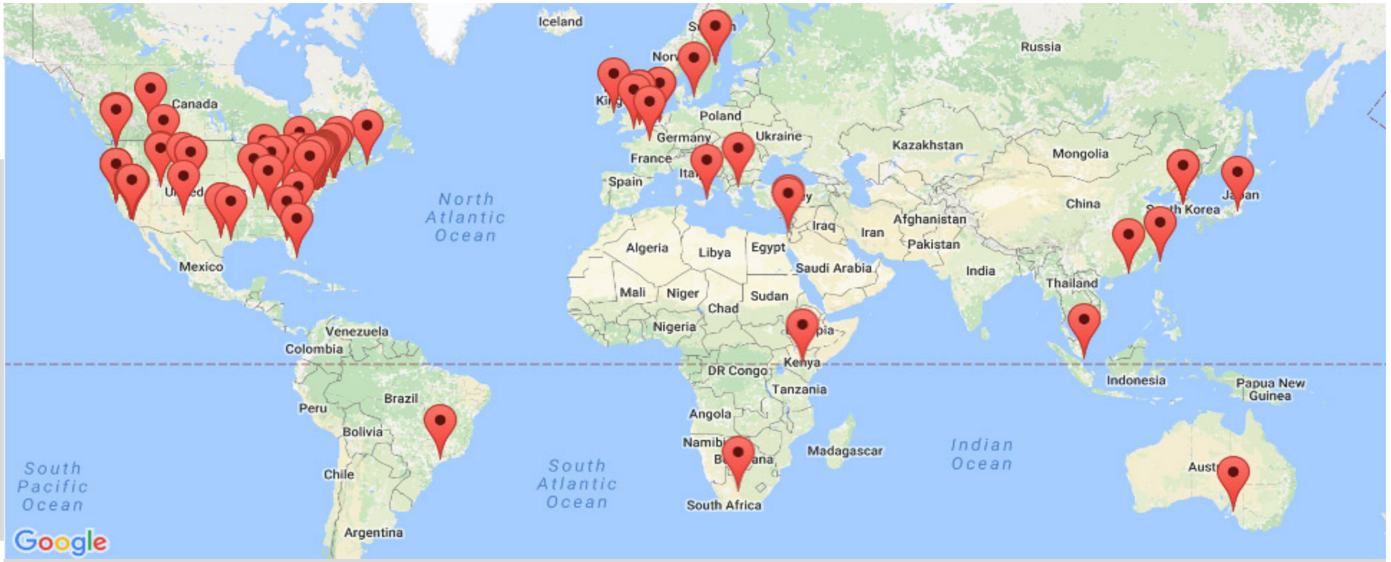
Goal: 1 billion patient records for observational research

25 countries

200 researchers

80 databases

600 million patient records



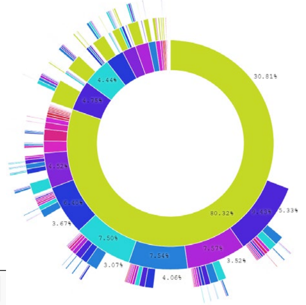
Columbia University is the coordinating center

George Hripcsak, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, Martijn J. Schuemie, Frank J. DeFalco, Adler Perotte, Juan M. Banda, Christian G. Reich, Lisa M. Schilling, Michael E. Matheny, Daniella Meeker, Nicole Pratt, and **David Madigan**, “Characterizing treatment pathways at scale using the OHDSI network,” PNAS Early Edition, April 2016.

Heterogeneity of Observational Research Results

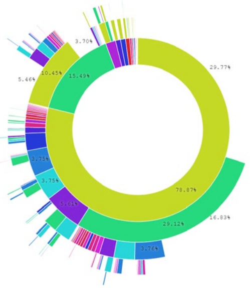
Type 2 Diabetes Mellitus

CCAE

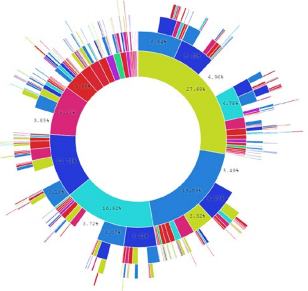


- Metformin 🟨
- Gliclazide 🟩
- pioglitazone 🟦
- sitagliptin 🟦
- glimepiride 🟦
- Glipizide 🟪
- rosiglitazone 🟪
- Glyburide 🟪
- Insulin, Glargine, Human 🟪
- exenatide 🟪
- liraglutide 🟥
- Insulin, Aspart, Human 🟥
- saxagliptin 🟥

CPRD

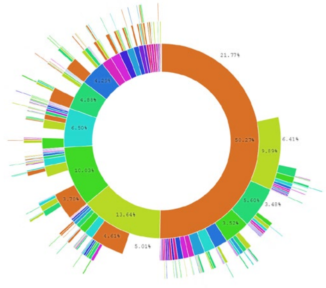


JMDC



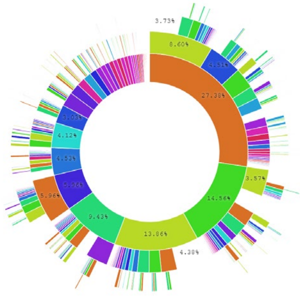
Hypertension

CUMC

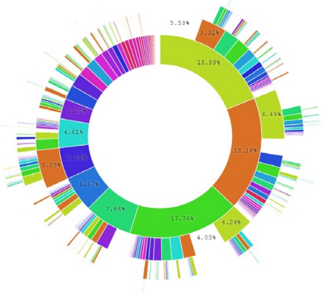


- Hydrochlorothiazide 🟧
- Lisinopril 🟨
- Metoprolol 🟩
- Amlodipine 🟩
- Furosemide 🟦
- Losartan 🟦
- Atenolol 🟦
- valsartan 🟦
- carvedilol 🟦
- Triamterene 🟪
- Diltiazem 🟪
- Ramipril 🟪
- benazepril 🟪
- olmesartan 🟪
- Spironolactone 🟪
- Clonidine 🟪

INPC

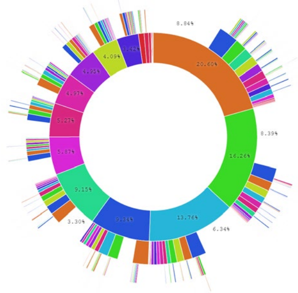


MDCR



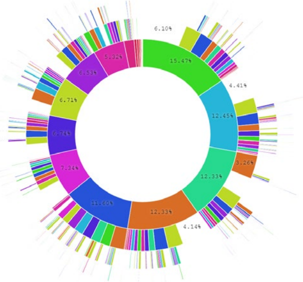
Depression

MDCD

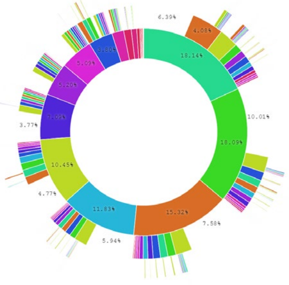


- Citalopram 🟧
- Bupropion 🟨
- Sertraline 🟩
- Escitalopram 🟩
- Fluoxetine 🟦
- Trazodone 🟦
- venlafaxine 🟪
- duloxetine 🟪
- Paroxetine 🟪
- Amitriptyline 🟥
- Mirtazapine 🟥
- Desvenlafaxine 🟥
- Nortriptyline 🟥
- Doxepin 🟥

GE

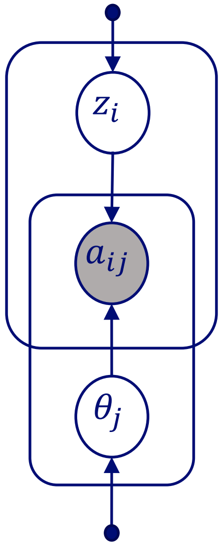


OPTUM



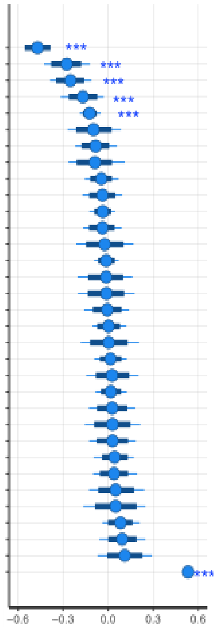
The Medical Deconfounder

Extract EHRs from the OHDSI database



$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2)$$

Fit the medical deconfounder

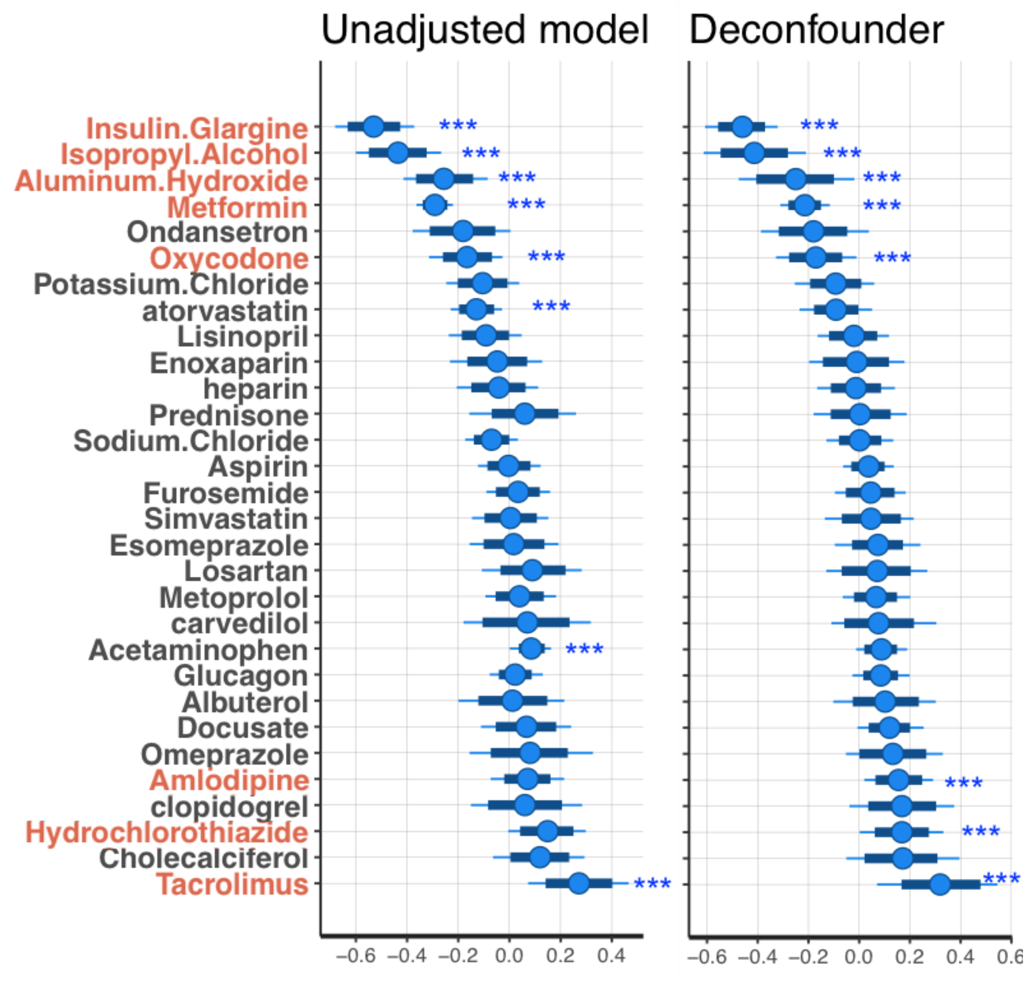


Analyze the causal effects of medications

Evaluate the results by medical literature review

Linying Zhang, Yixin Wang, Anna Ostropolets, Jami J. Mulgrave, David M. Blei, George Hripcsak, “The Medical Deconfounder: Assessing Treatment Effect with Electronic Health Records (EHRs),” arXiv:1904.02098v1, April 2019.

Treatment Effects on Hemoglobin A1c (Type 2 Diabetes)



- The unadjusted model

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij}, \sigma^2)$$

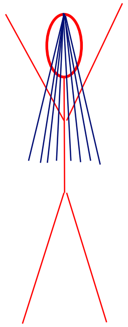
- The medical deconfounder

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2)$$

- The deconfounder reduces both false positive and false negative rates: acetaminophen (c2nc); amolodipine and hydrochorothiazide (nc2c).
- It identifies effective (causal) drugs that are more consistent with the medical literature.

Data for Good

Thank you!



DATA SCIENCE INSTITUTE
COLUMBIA UNIVERSITY