Welcome

“Data for Good: Ensuring the Responsible Use of Data to Benefit Society”

Jeannette Wing

Twitter Hashtag: #ACMLearning

Tweet questions & comments to: @ACMeducation

Post-Talk Discourse: https://on.acm.org

Additional Info:

- Talk begins at the top of the hour and lasts 60 minutes
- On the bottom panel you’ll find a number of widgets, including Twitter and Sharing apps
- For volume control, use your master volume controls and try headphones if it’s too low
- If you are experiencing any issues, try refreshing your browser or relaunching your session
- At the end of the presentation, you will help us out if you take the experience survey
- This session is being recorded and will be archived for on-demand viewing. You’ll receive an email when it’s available.

Data for Good: Ensuring the Responsible Use of Data to Benefit Society

Speaker: Jeannette Wing

Moderator: Paul Leidig

ACM.org Highlights

For Scientists, Programmers, Designers, and Managers:

- Learning Center - https://learning.acm.org
  - View past TechTalks & Podcasts with top inventors, innovators, entrepreneurs, & award winners
  - Access to O'Reilly Learning Platform – technical books, courses, videos, tutorials & case studies
  - Access to Skillsoft Training & ScienceDirect – vendor certification prep, technical books & courses
- Ethical Responsibility – https://ethics.acm.org

By the Numbers

- 2,200,000+ content readers
- 1,800,000+ DL research citations
- $1,000,000 Turing Award prize
- 100,000+ global members
- 1160+ Fellows
- 700+ chapters globally
- 170+ yearly conferences globally
- 100+ yearly awards
- 70+ Turing Award Laureates

Popular Publications & Research Papers

- Communications of the ACM - http://cacm.acm.org
- Queue Magazine - http://queue.acm.org
- Digital Library - http://dl.acm.org

Major Conferences, Events, & Recognition

- *https://www.acm.org/conferences*
- *https://www.acm.org/chapters*
- *https://awards.acm.org*

Welcome

“Data for Good: Ensuring the Responsible Use of Data to Benefit Society”

Jeannette Wing

Twitter Hashtag: #ACMLearning

Tweet questions & comments to: @ACMeducation

Post-Talk Discourse: https://on.acm.org

Additional Info:

- Talk begins at the top of the hour and lasts 60 minutes
- On the bottom panel you’ll find a number of widgets, including Twitter and Sharing apps
- For volume control, use your master volume controls and try headphones if it’s too low
- If you are experiencing any issues, try refreshing your browser or relaunching your session
- At the end of the presentation, you will help us out if you take the experience survey
- This session is being recorded and will be archived for on-demand viewing. You’ll receive an email when it’s available.


DATA SCIENCE INSTITUTE
COLUMBIA UNIVERSITY

Data For Good:

Ensuring the Responsible Use of Data to Benefit Society

Jeannette M. Wing

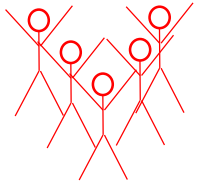
Avanessians Director of the Data Science Institute and Professor of Computer Science
Columbia University

Adjunct Professor of Computer Science
Carnegie Mellon University

ACM Tech Talk
April 30, 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

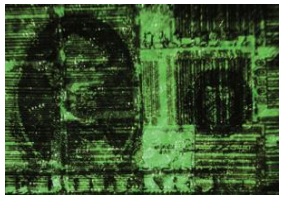
datascience.columbia.edu/data-science-centers

Foundations

Computing Systems

Cybersecurity

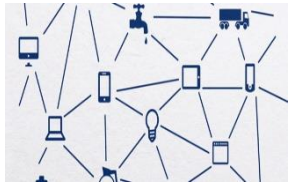
Data, Media, and Society



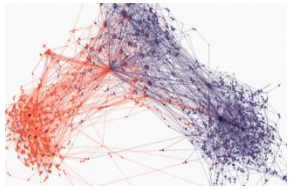
Financial Analytics

Health Analytics

Smart Cities

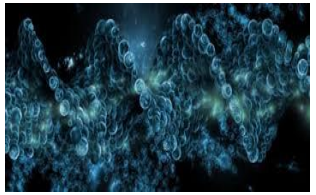


Sense, Collect, and Move



Computational Social Science

Education



Materials Discovery Analytics

Collaboratory (Columbia Entrepreneurship + DSI)

HUMANITIES DATA SCIENCE

Data: Past, Present, and Future

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

Co-taught by Applied Math and History professors

PUBLIC POLICY COMPUTER SCIENCE

Computational Literacy for Public Policy

DATA SCIENCE NEWS

Data Literacy in Urban Planning and Journalism

BUSINESS ANALYTICS DATA SCIENCE

The Collaboratory Heads for Columbia Business School

DATA SCIENCE MRS

In Vivo MRS: From Data to [illegible]enefit

DATA SCIENCE COMPUTER SCIENCE

Collaboratory Opens New Data Science Clinic

The Collaboratory Creates a Platform for Pedagogical Innovation Across Columbia

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

NEUROSCIENCE CHEMICAL ENGINEERING

Neurogenomics

SOCIAL WORK DATA SCIENCE

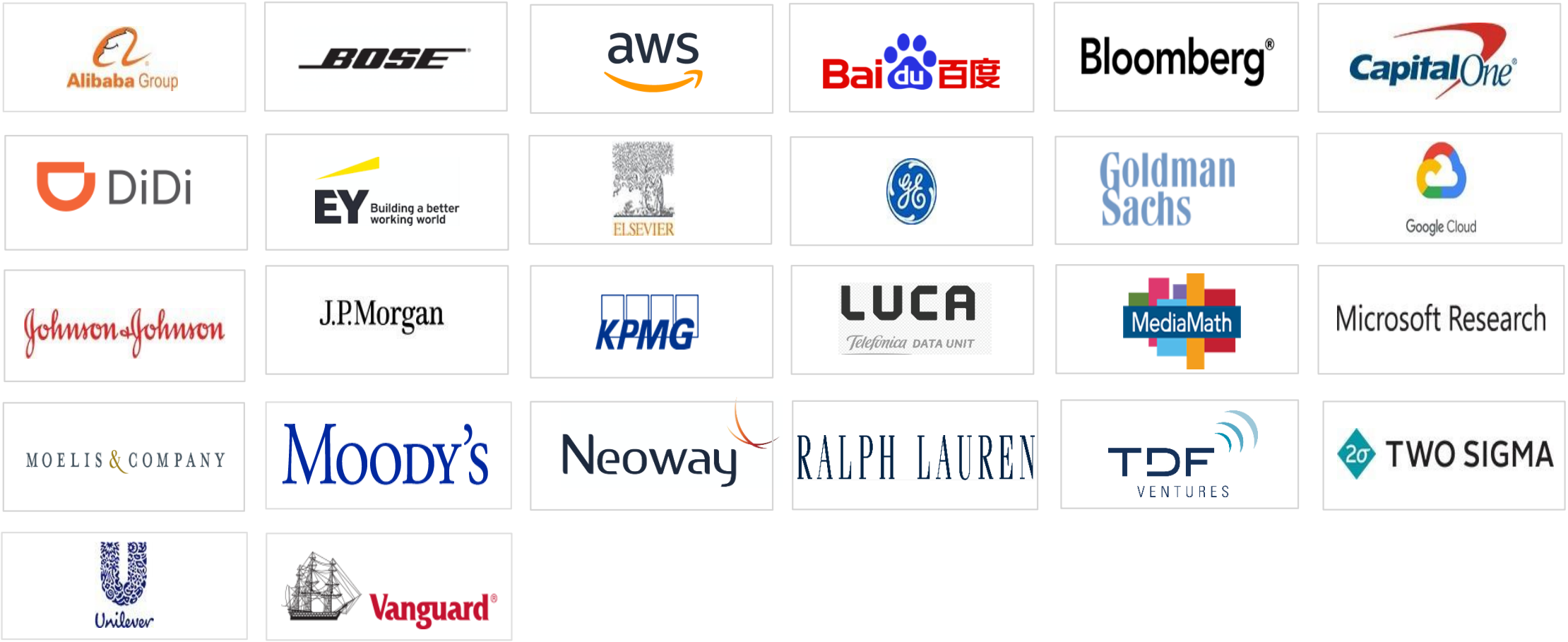
Data Science for Social Good

COMPARATIVE LIT & SOCIETY COMPUTER SCIENCE

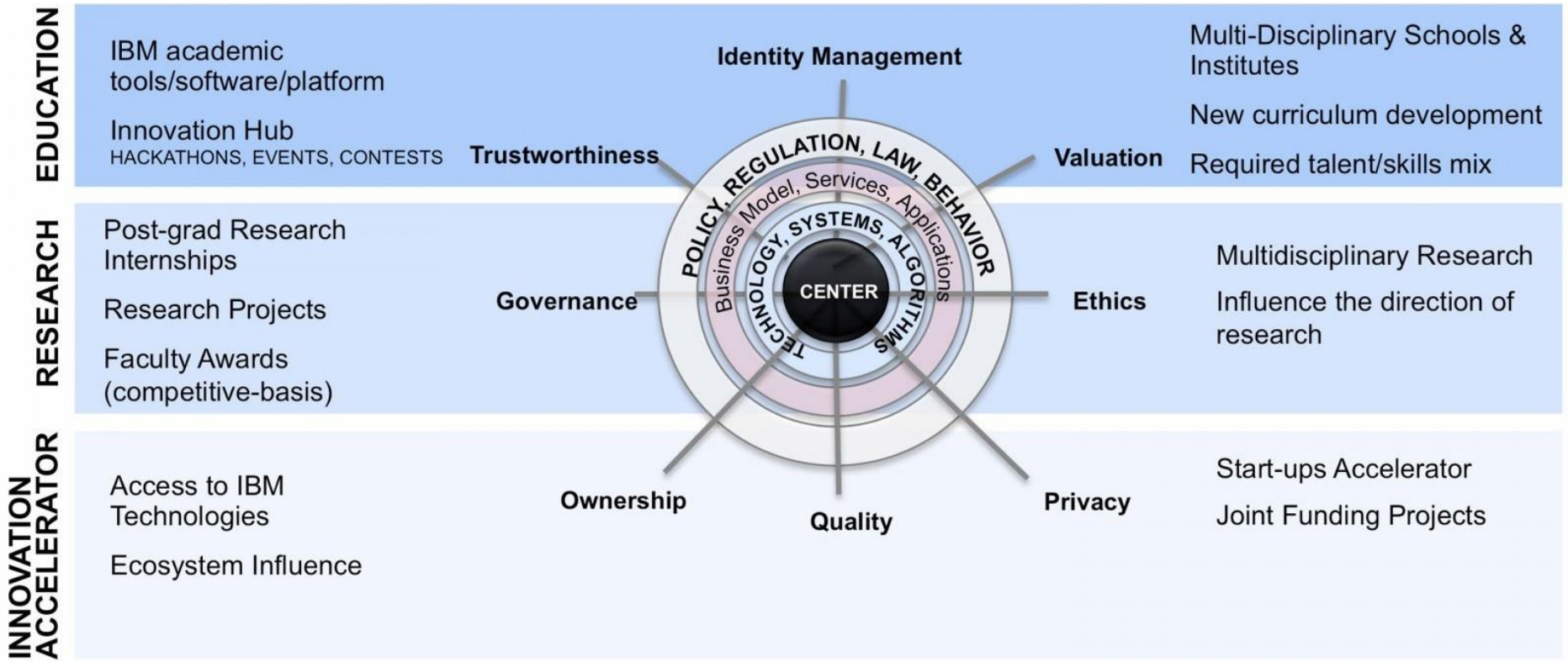
Tech and Language Diversity

Industry Affiliates Program

industry.datascience.columbia.edu



Columbia-IBM Center on Blockchain and Data Transparency



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

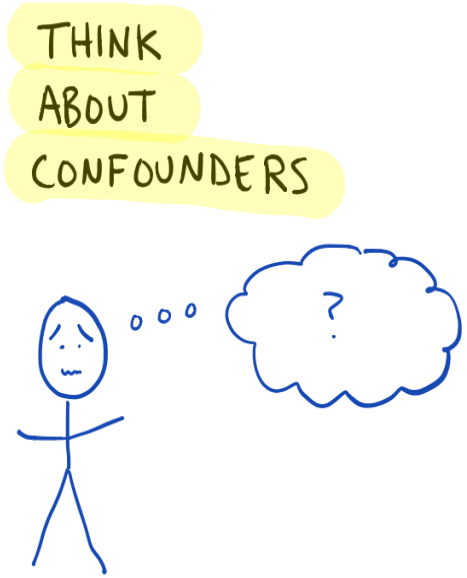
What happens to movie revenue **if** we place an actor in a movie**?**

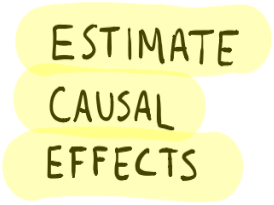
Goal: $E[Y_i(\boldsymbol{a})]$ $E[Y_i \mid do(\boldsymbol{a})]$

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





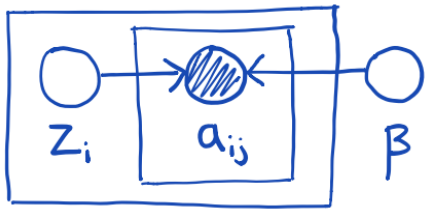
$$\mathbb{E}[Y_i(a)] = \mathbb{E}[\mathbb{E}[Y_i(A_i) \mid A_i = a, W_i]]$$

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

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

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

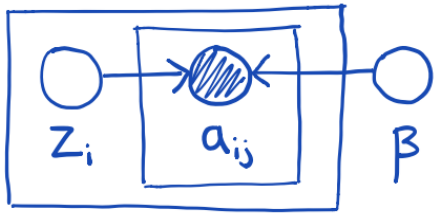
ESTIMATE
CAUSAL
EFFECTS

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

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

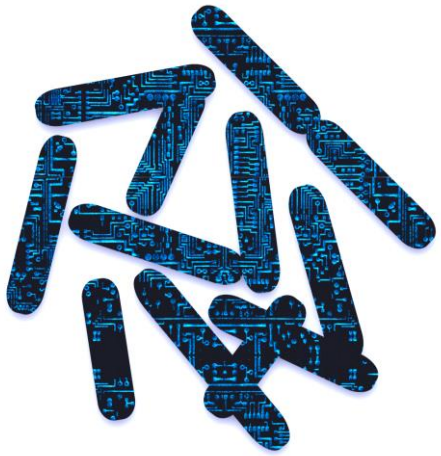
- With the deconfounder,
  (1) Sean Connery's (James Bond) value goes up.
  (2) Bernard Lee's (M) and Desmond Llewelyn's (Q) values go down.
- We can now answer questions such as: What happens to revenue if we place Desmond Llewelyn in *A Beautiful Mind*? How about Sean Connery?
- The deconfounder **corrects 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

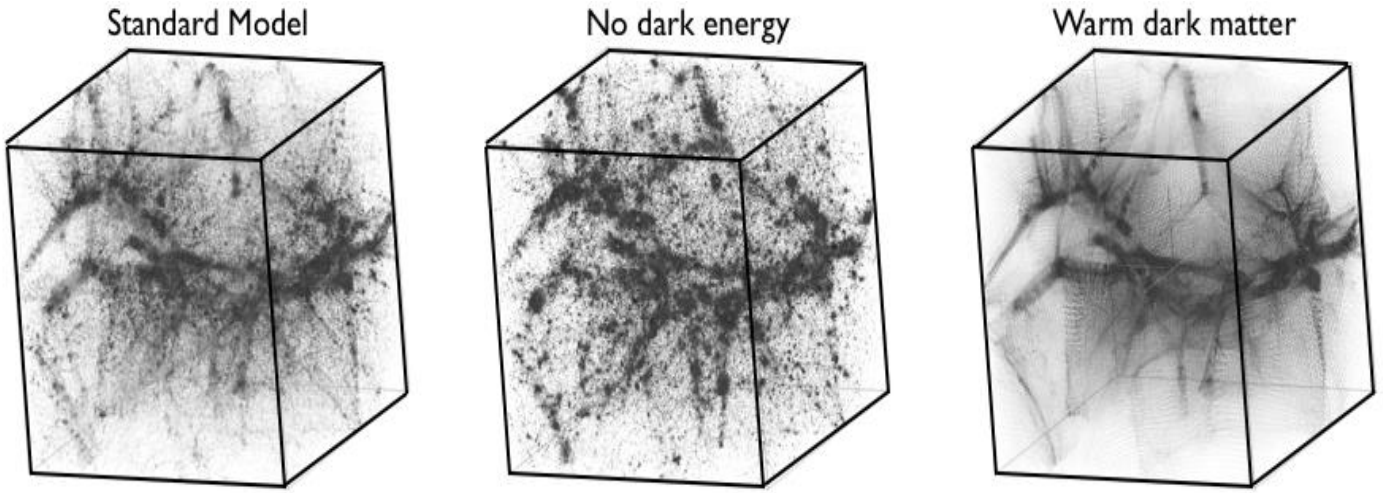
Ensure the responsible use of data to benefit society

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

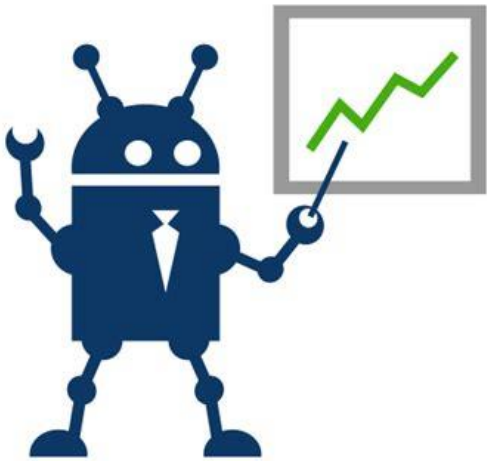


Arushi Gupta, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, "*Non-Gaussian information from weak lensing data via deep learning,*" *Physical Review D*, in press (accepted April 30, 2018), E-print available at https://arxiv.org/abs/1802.01212

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

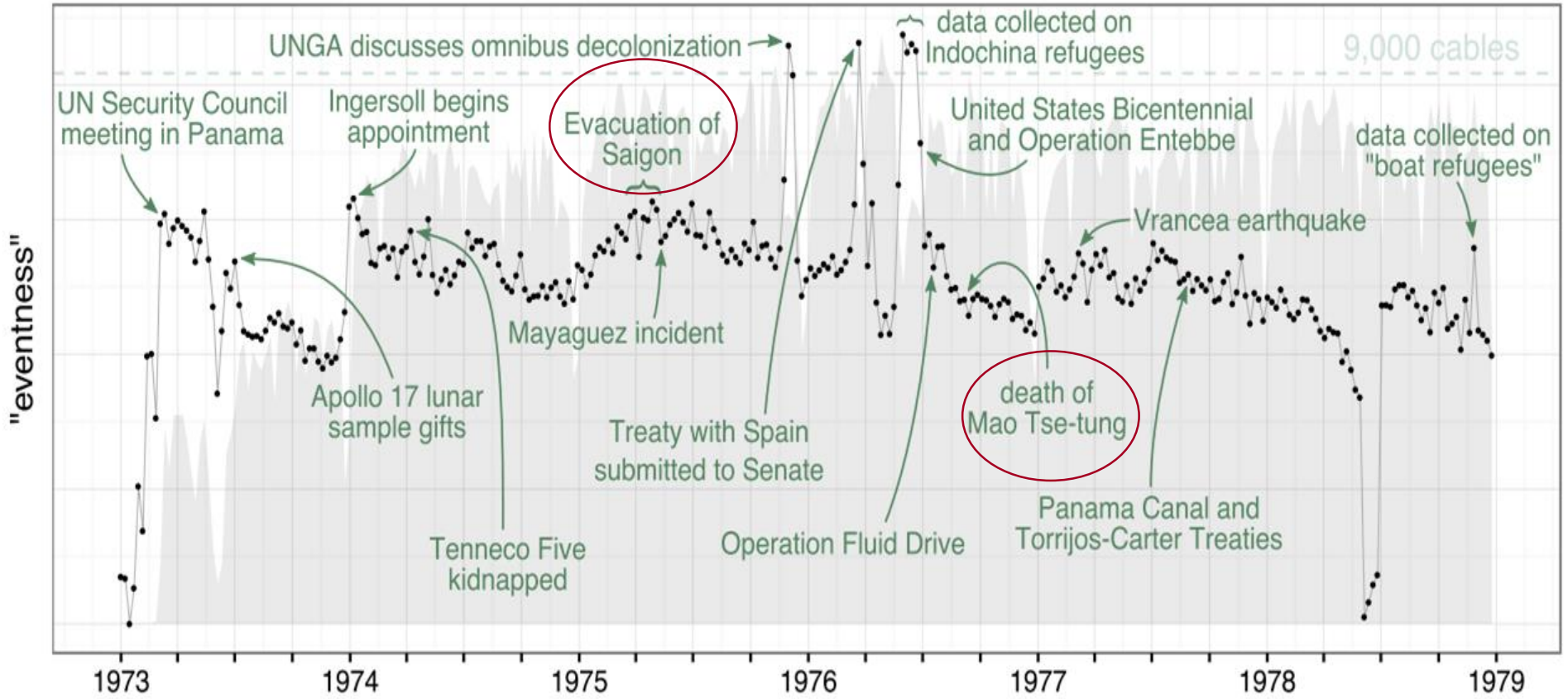


Agostino Capponi, Octavio Ruiz Lacedelli, and Matt Stern, “Robo-Advising as a Human-Machine Interaction System”, August 2018, preprint.

Event Discovery: History and Topic Modeling

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

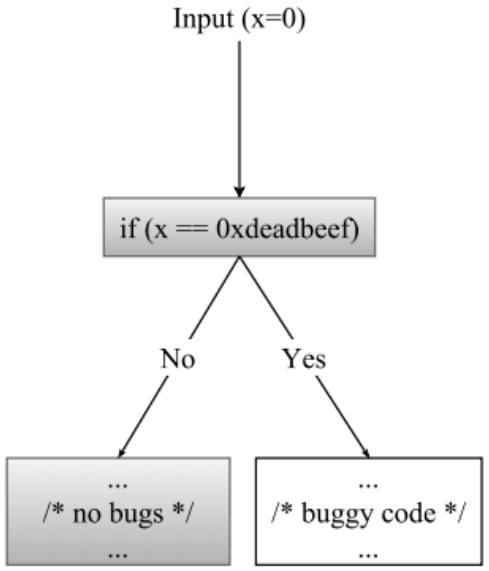
Privacy

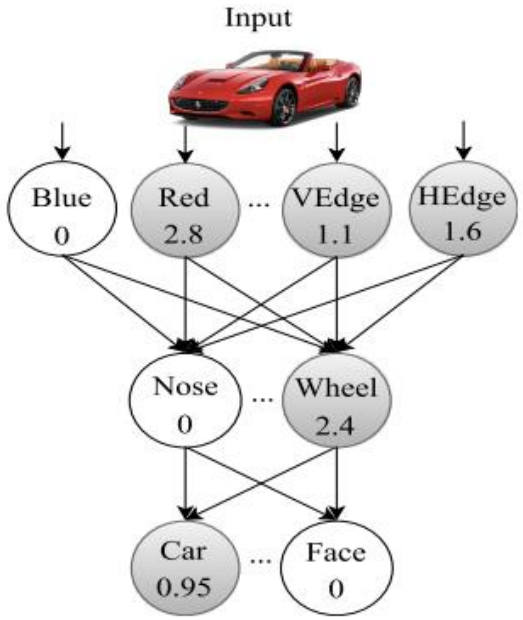
Reliability

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

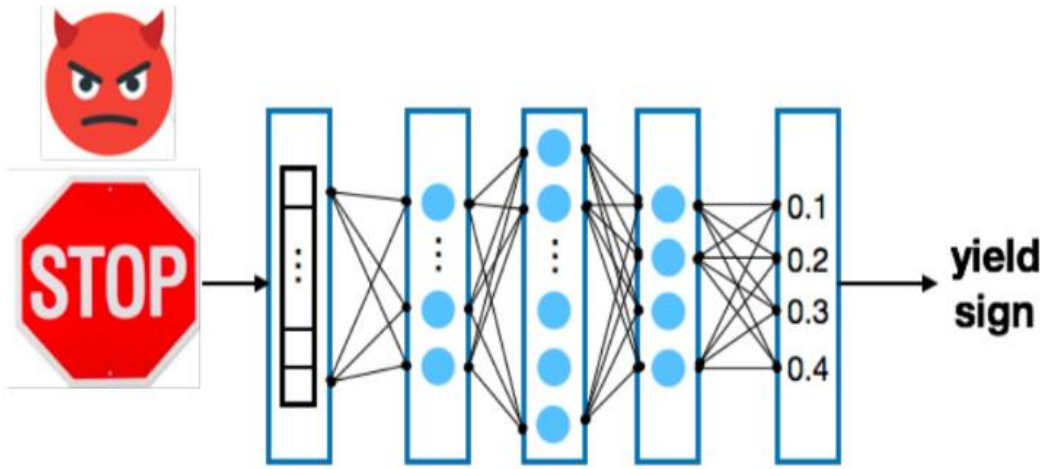
Seed,
No accident

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

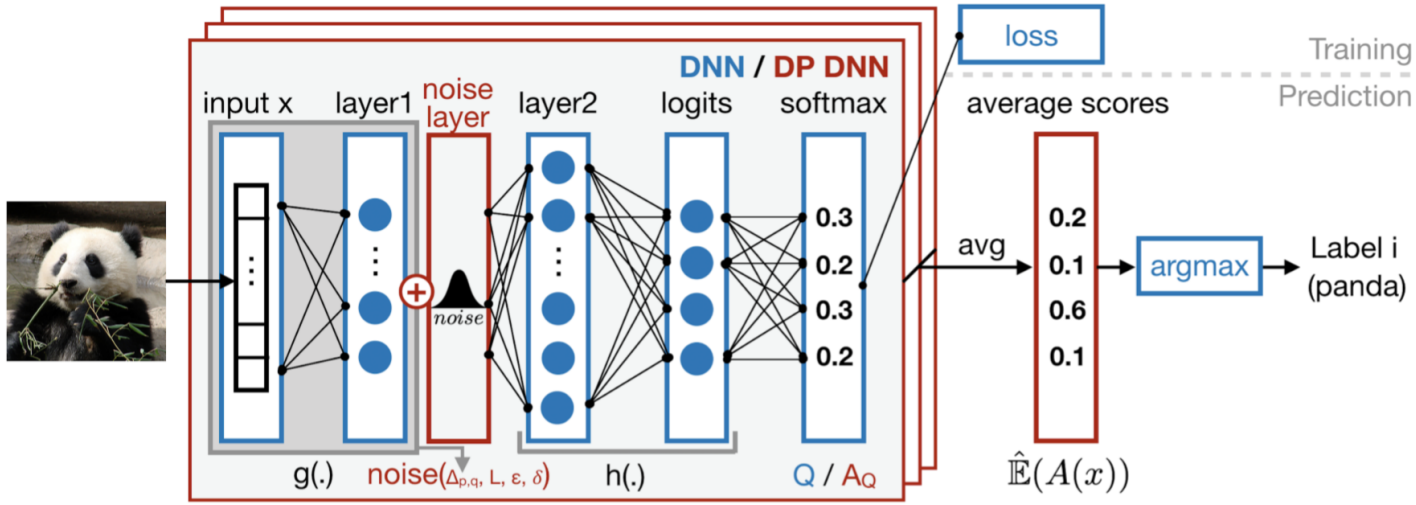
Problem



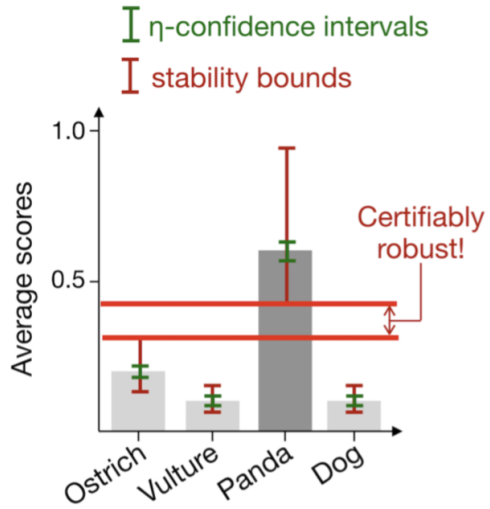
Mathias Lecuyer, Baggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana, “Certified Robustness to Adversarial Examples with Differential Privacy, arXiv:1802.03471v2 , June 26, 2018, to appear IEEE Security and Privacy (‘’Oakland’’) 2019.

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

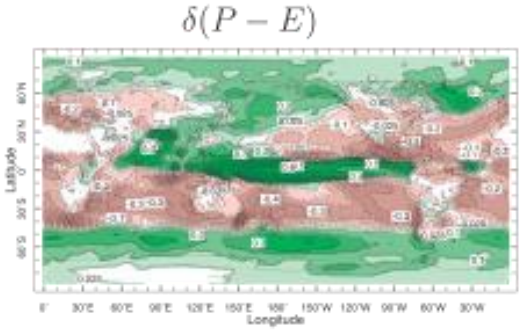
PANGEO: Climate Science and Big Data

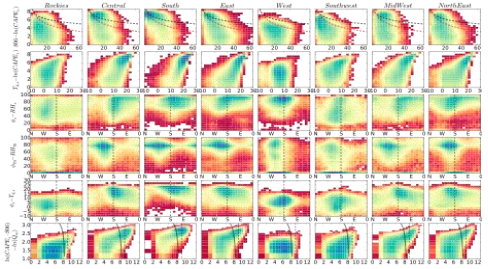
https://pangeo-data.github.io/

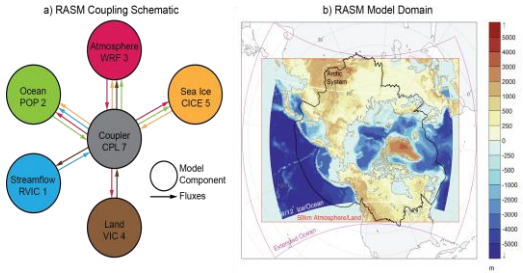
PI: Ryan Abernathey
(Dept. of Earth & Env. Sci., LDEO, Columbia University)

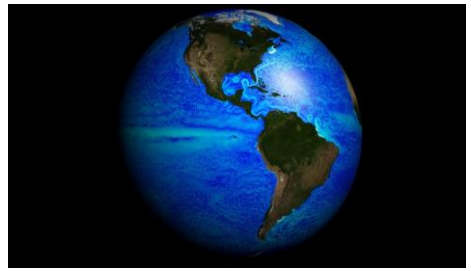
Co-PIs: Chiara Lepore, Michael Tippett, Naomi Henderson, Richard Seager (LDEO)
Kevin Paul, Joe Hamman, Ryan May, Davide Del Vento (National Center for Atmospheric Research)
Matthew Rocklin (Anaconda; formerly Continuum Analytics)

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)

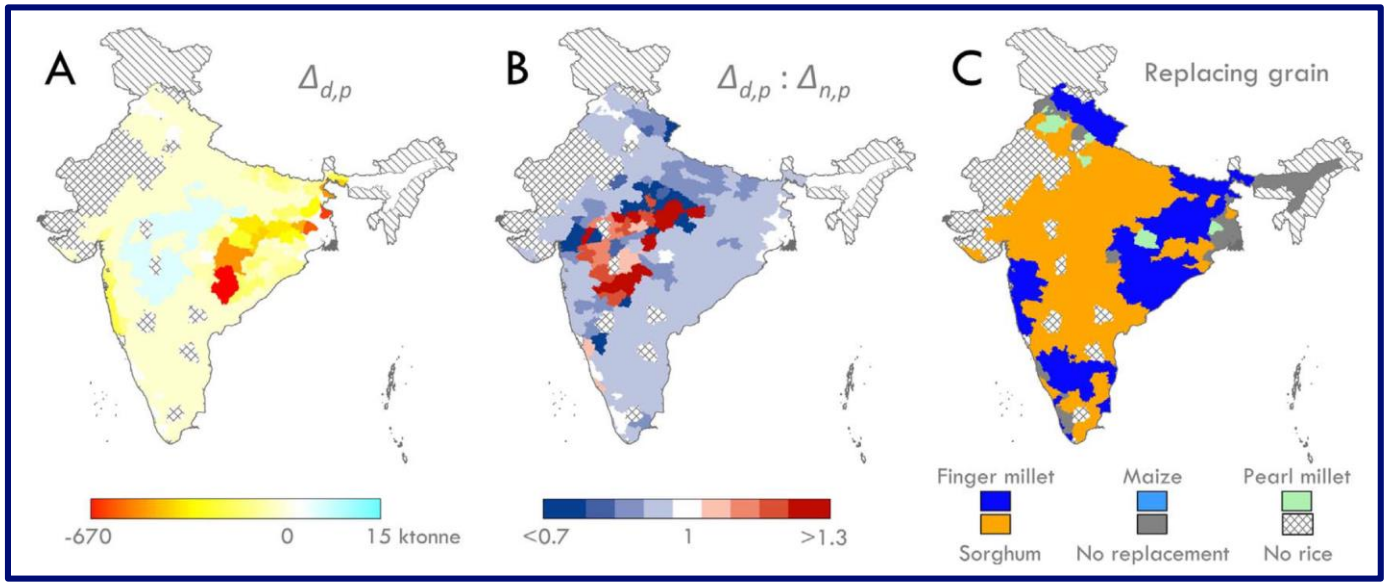








Data Science and Agriculture



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”)

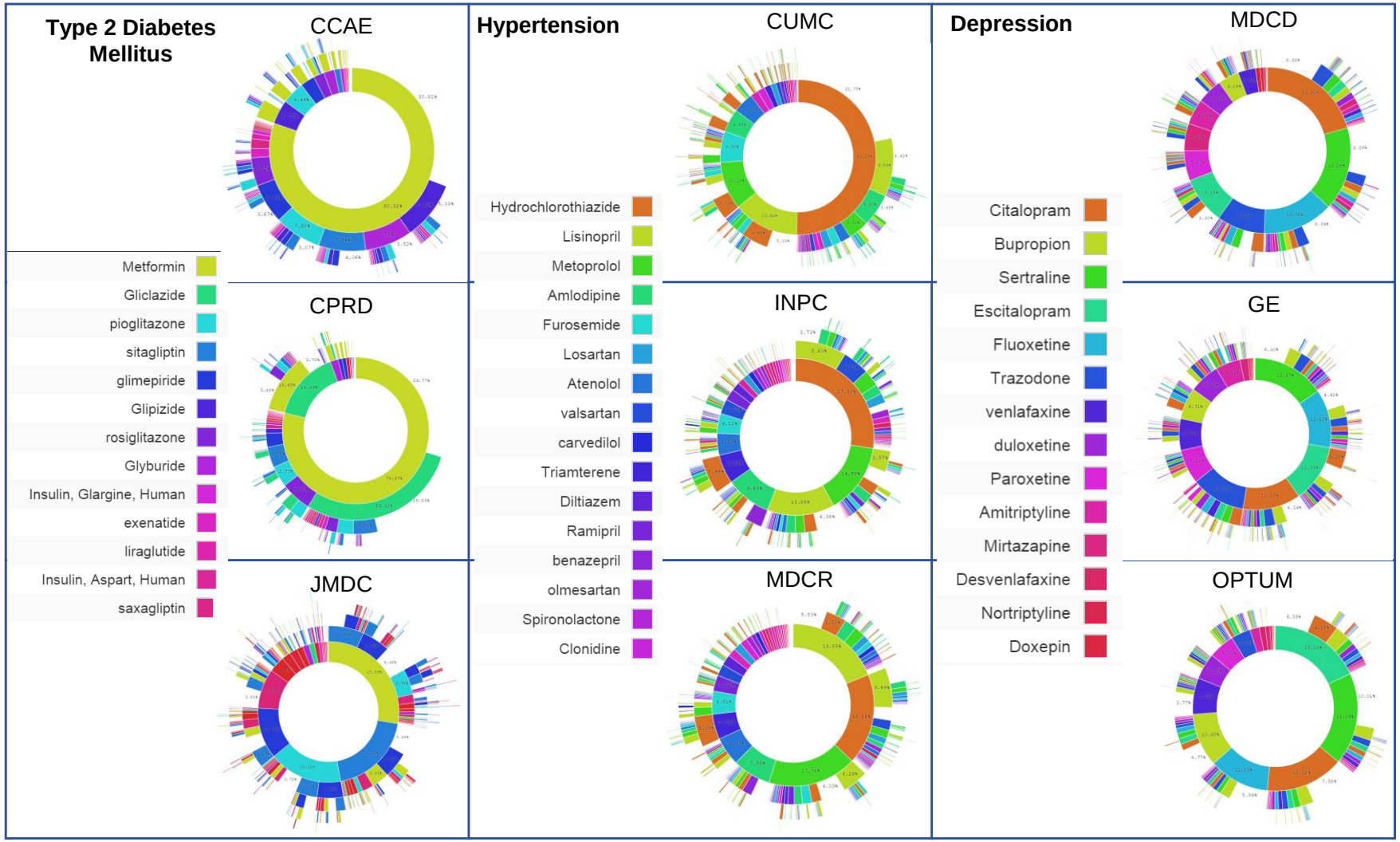
Goal: 1 billion patient records for observational research

- 25 countries
- 200 researchers
- 80 databases
- 600 million patient records

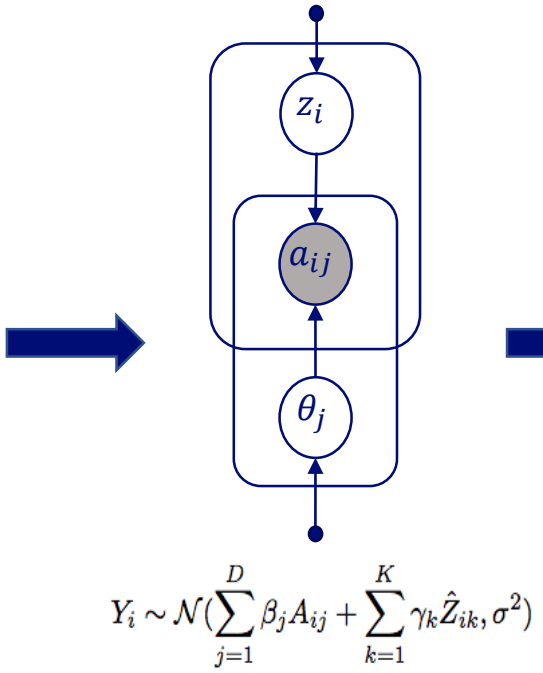
Columbia University is the coordinating center

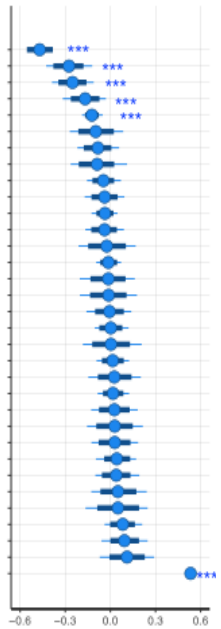
George Hripcsak, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, Martijn J. Schuemie, Frank J. DeFalco, Adler Perotte, Juan M. Banda, Christian G. Reich, Lisa M. Schilling, Michael E. Matheny, Daniella Meeker, Nicole Pratt, and **David Madigan**, “Characterizing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

Heterogeneity of Observational Research Results



The Medical Deconfounder





Extract EHRs from the OHDSI database

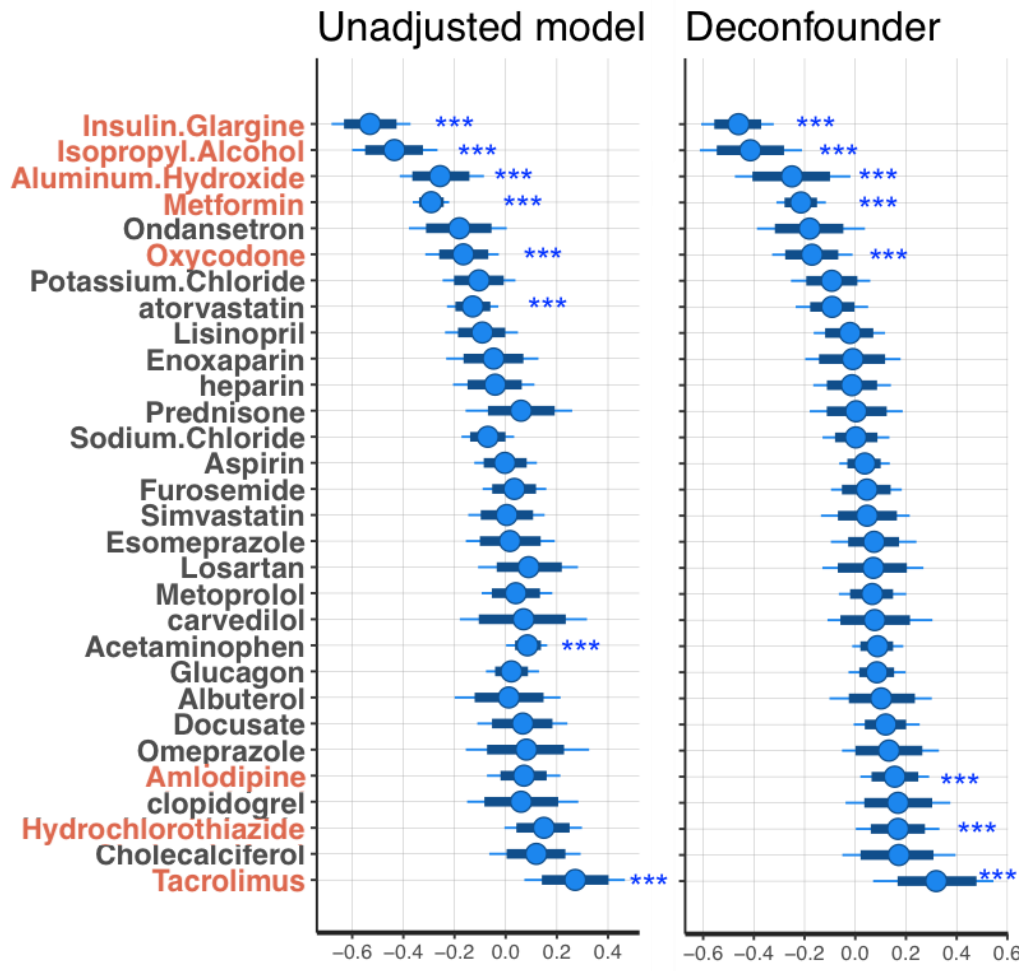
Fit the medical deconfounder

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

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

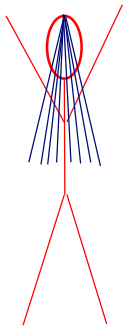
- The medical deconfounder

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

- 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

The Learning Continues…

TechTalk Discourse: https://on.acm.org

TechTalk Inquiries: learning@acm.org

TechTalk Archives: https://learning.acm.org/techtalks

Learning Center: https://learning.acm.org

Professional Ethics: https://ethics.acm.org

Queue Magazine: https://queue.acm.org

Data Science Task Force: http://dstf.acm.org/