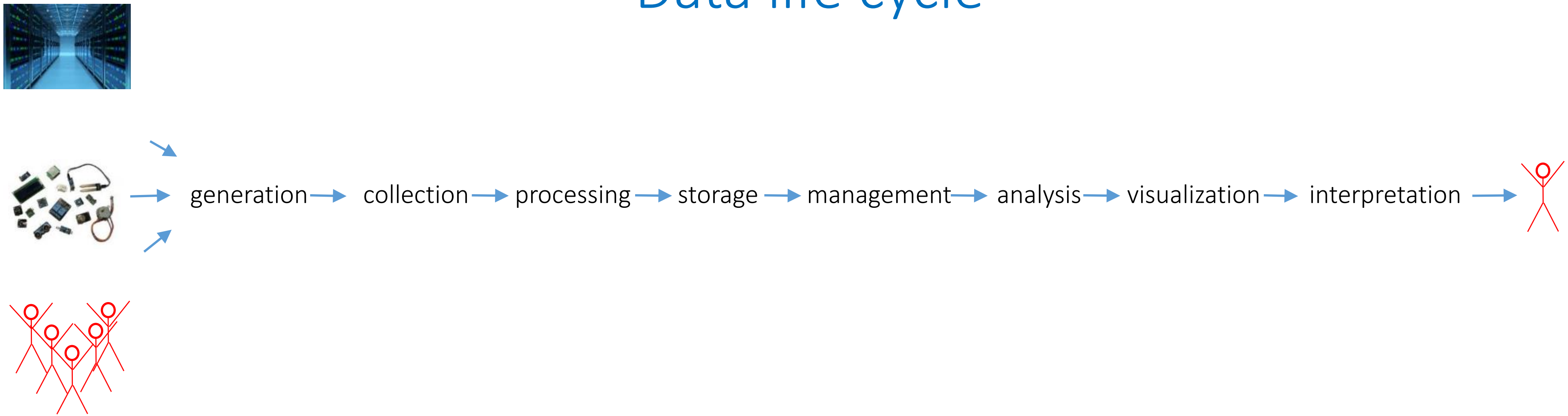
COLUMBIA UNIVERSITY
Data Science Institute

Data for Good:
Data Science at Columbia

Jeannette M. Wing
Avanessians Director of the Data Science Institute
Professor of Computer Science
Columbia University

Brazilian National Academy of Sciences
Rio de Janeiro, Brazil
May 9, 2018

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

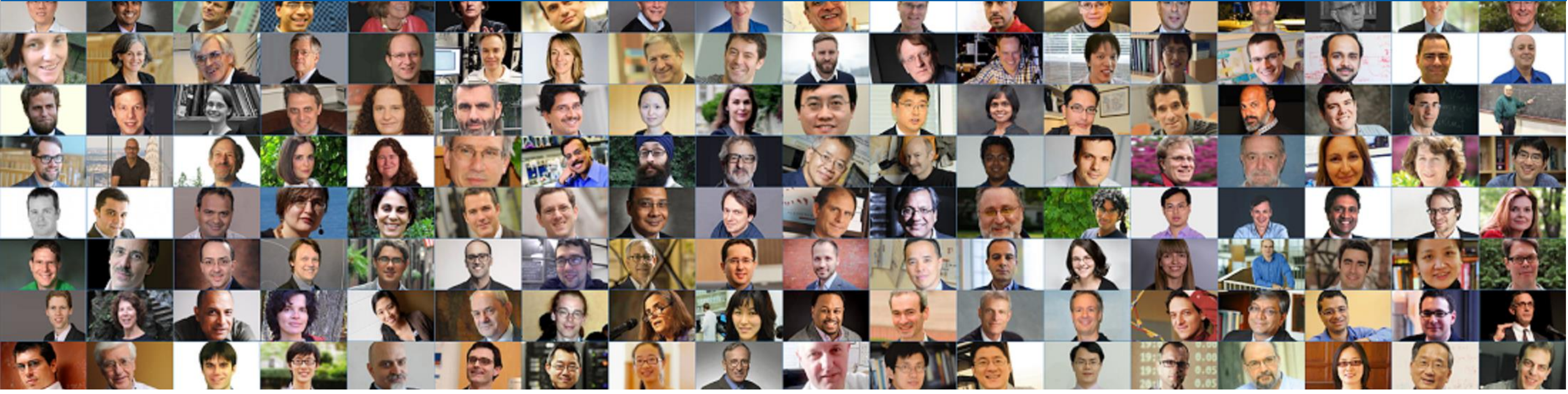
Data for Good

COLUMBIA UNIVERSITY
Data Science Institute

12 Schools

300+ Faculty

- Arts and Sciences
- Architecture, Planning, and Preservation
- Business
- Dental Medicine
- Engineering and Applied Sciences
- International and Public Affairs
- Journalism
- Law
- Medicine
- Nursing
- Public Health
- Social Work



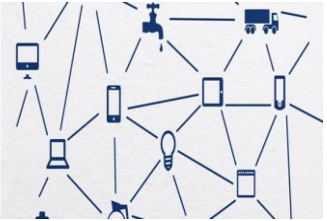
Centers and Frontiers

datascience.columbia.edu/data-science-centers

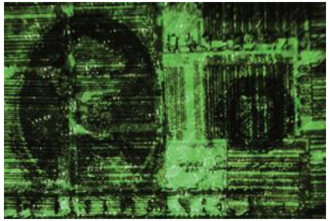
Foundations

Cybersecurity

Data, Media and Society



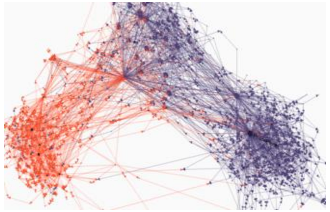
Sense, Collect, and Move



Financial Analytics

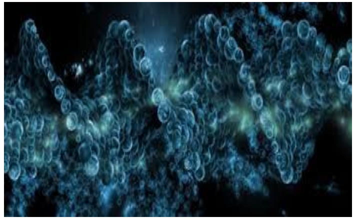
Health Analytics

Smart Cities



Computational Social Science

Computing Systems



Materials Discovery Analytics

Education

Degree:

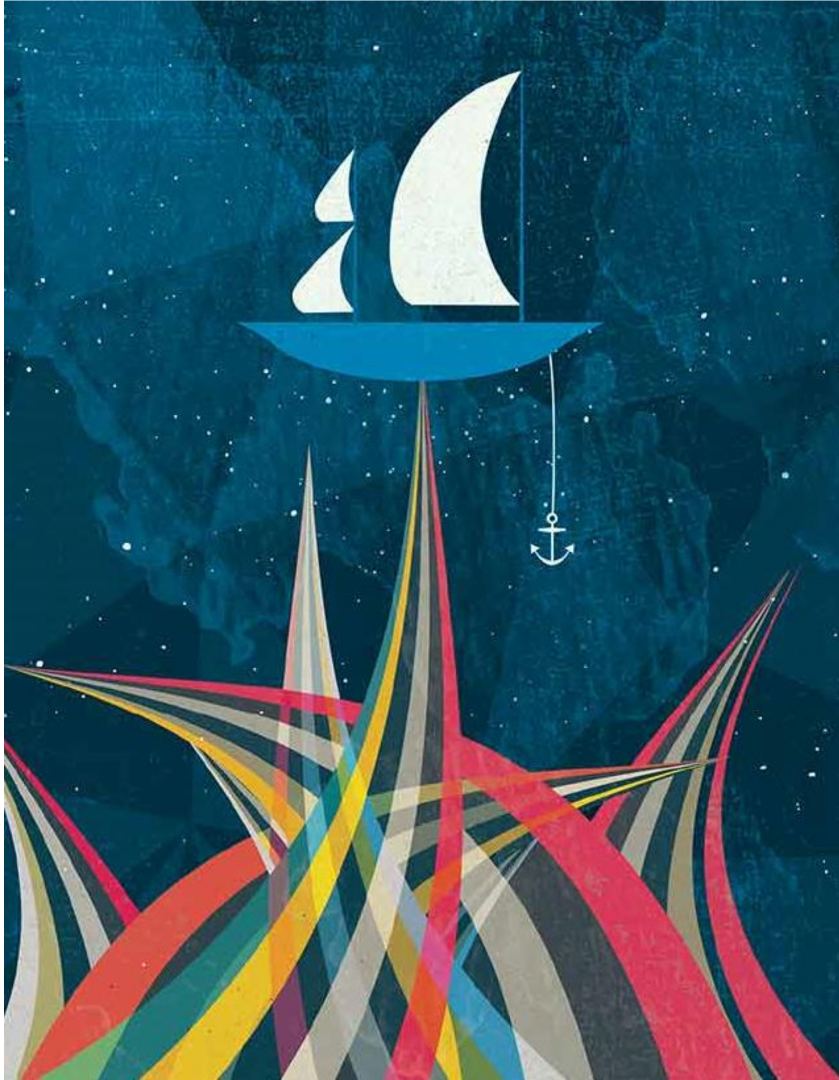
Master of Science 2014

Non-Degree:

Certification 2013

ColumbiaX:

Online Courses 2016



Data Science Student Employers

AIG, Adhoc, Alvarez & Marsal, Amazon, American Express, Amgen, Amper Music, Amphora, Audible, Barclays, BCT Partners, Blackrock, Capital One, Capital One Labs, CartoDB, CKM Advisors, Clipper Data, Collibra, Comcast, Creative Chaos, Crisis Text Line, Deloitte, Digital Reasoning, Droice, Early Signal, eBay, EMC Corporation, eScience Institute at U of Washington, Facebook, Factset Research Systems, FarePortal, FDNY, GE Research, Glassdoor, Goldman Sachs, Google, Guy Carpenter, Handy, Hover, IBM, IBM Social Good Fellowship, Intersection, Intuit, Jet.com, Jobdiva, Kora Capital Management, Manhattan DA's Office, McKinsey, MediaMath, Merck, Microsoft, Milliman Max, MoneyLion, Morgan Stanley, Mount Sinai, MSCI, Mylan Pharma, NBCUniversal, Nestle Waters, Netflix, NYC Department of Buildings, OnDeck, Palantir, Paypal, Pfizer, Pfizer, Pixel Place, Point72, Primus, Quaera, Quarterspot, RBC, Red Ventures, Resolvity, SAP, Satmap, Singapore Bank, Spotify, Spreemo, Springbot, Swiss Re, Synergic Partners, TAPAD, The Hartford., Tomorrow Networks, Trans Org Analytics, Tremor Video, Trifecta, Trinnacle Capital, TuneIn, Twitter, Uber, Uncommon Schools, United Nations, Venus Tech Ventures, Verisk Analytics, Viacom, Vulcan, Walmart, Yelp.

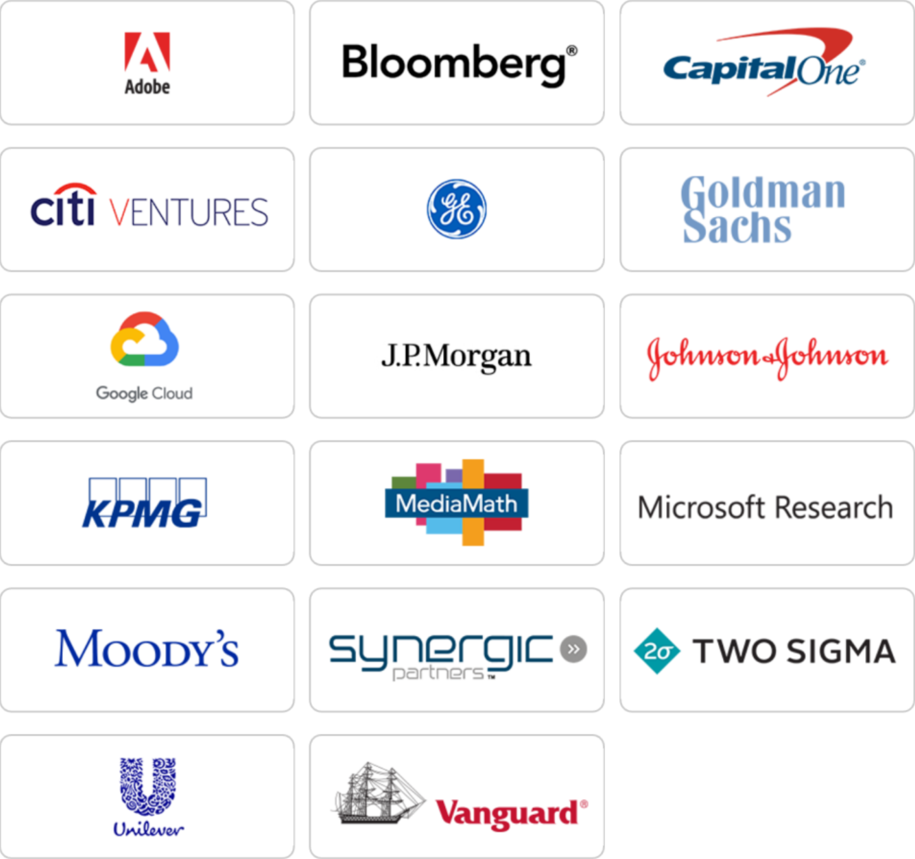
datascience.columbia.edu/data-science-careers

Top Data Scientist Jobs:

Data Scientist
Business Analyst
Data Analyst
Statistician
Senior Data Scientist
Chief Scientist
Research Scientist
Analytics Manager
Senior Business Analyst
Analytics Consultant
Business Intelligence Consultant
Data Architect
Research Analyst
Chief Data Scientist
Director of Analytics
Quantitative Analyst
Senior Web Analyst
Lead Analyst
Entrepreneur

100% Placement of Inaugural Class

Industry Affiliates Program



industry.datascience.columbia.edu

Capstone Projects

Students working with Industry Affiliates

Predicting Trash Hot Spots in New York City

This team analyzed 6 years of 311 data, tax records and pollution data for NYC's Dept. of Environmental Protection.

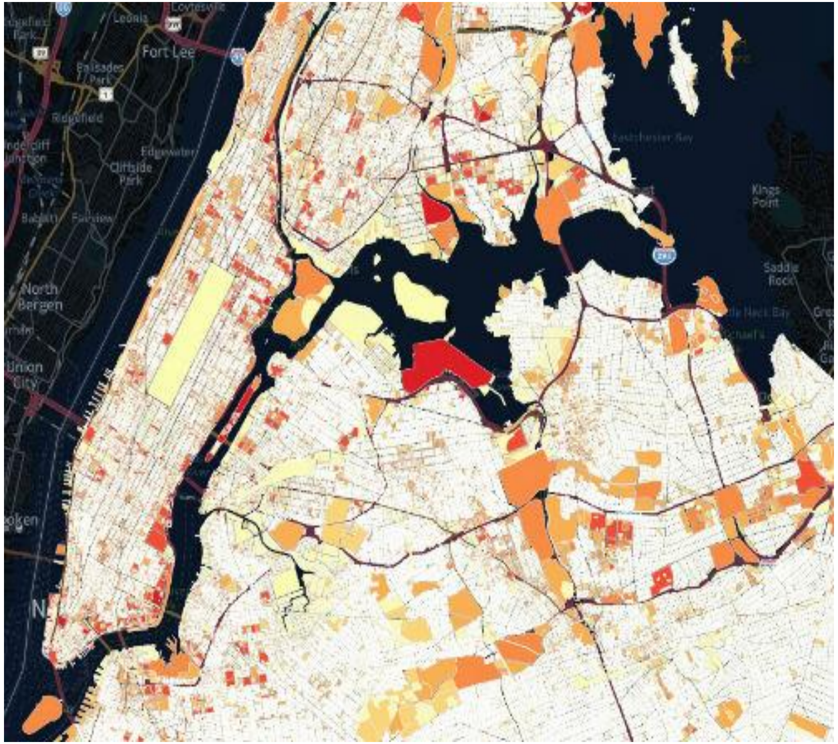
Created data map of neighborhoods that city could use to target cleanups, especially after rains and snow storms.

Found that population density is the single best predictor of trash complaints.

NEW YORK CITY DEPARTMENT OF ENVIRONMENTAL PROTECTION

DEP

www.nyc.gov/dep

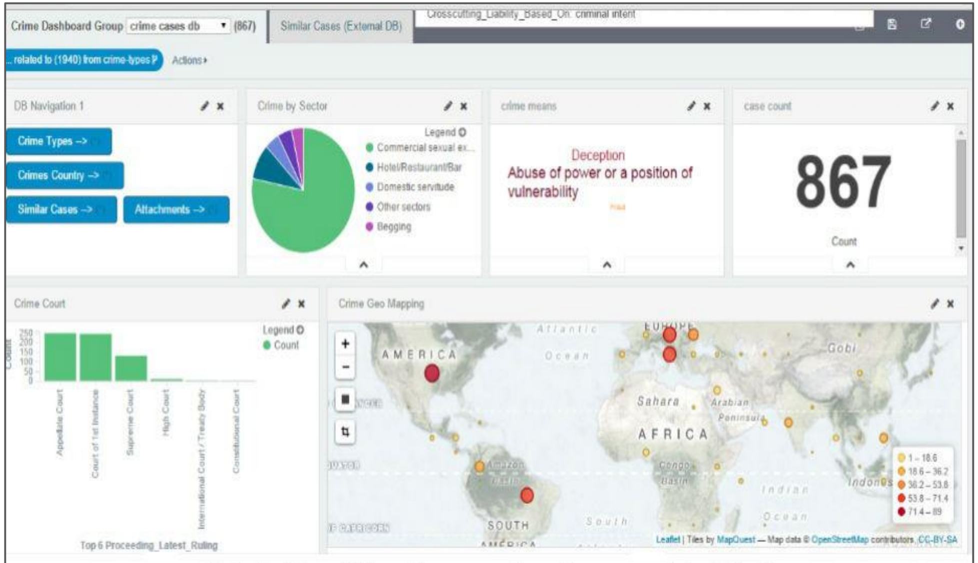


Automating Case Law Analysis

Lawyers must understand legal precedents
for plotting trial strategy and predicting trial outcome.

This team used case law from the U.N. Office on Drugs and Crimes.

Developed an interactive system for analyzing data on legal precedents.



Main dashboard filtered for cases where there was criminal intent

synergic partners™

DSI Feeds New York City’s Start Up Community

Agolo algorithmically curates your twitter feed

Wordseye lets anyone create 3D images

eBrevia uses machine learning to summarize legal documents

Fero Labs brings machine learning to factories to optimize production

TextIQ uses natural language processing to change the consumption of big data

Droice Labs uses artificial intelligence to predict how drug treatments affect patients

Chipscan, a cybersecurity firm, identifies malice in chips

Vidrovr analyzes video collections to make them searchable

Research Highlights

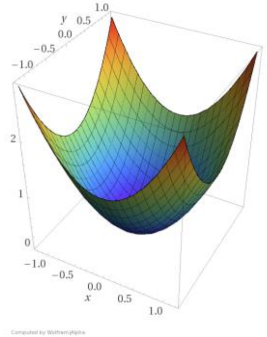
Mission

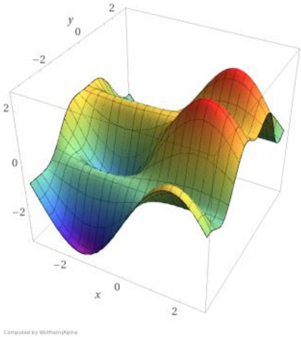
Advance the state of the art in data science

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

Ensure the responsible use of data to benefit society

Understanding Expectation Maximization



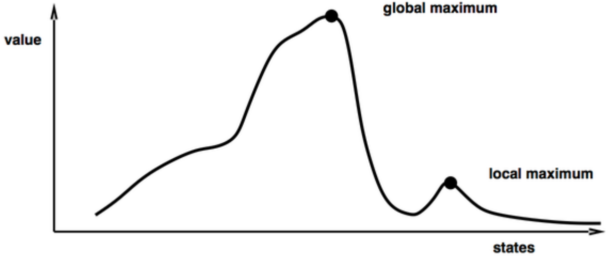


Ji Xu, Daniel Hsu, Arian Maleki, “Global analysis of Expectation-Maximization for mixtures of two Gaussians,” NIPS 2016

Expectation-Maximization (E-M)

E-M: local optimization procedure for Maximum Likelihood Estimation (MLE) in statistical models [Dempster, Laird, & Rubin, 1977].

- [DLR'77] cited 50,000+ times; algorithm ubiquitous in statistical applications.
- Finds stationary point of likelihood objective (e.g., local maximizers).
- Does **not** necessarily find MLE.
- Statistical theory about MLE does not generally apply to local maximizers, and hence does not generally apply to E-M.



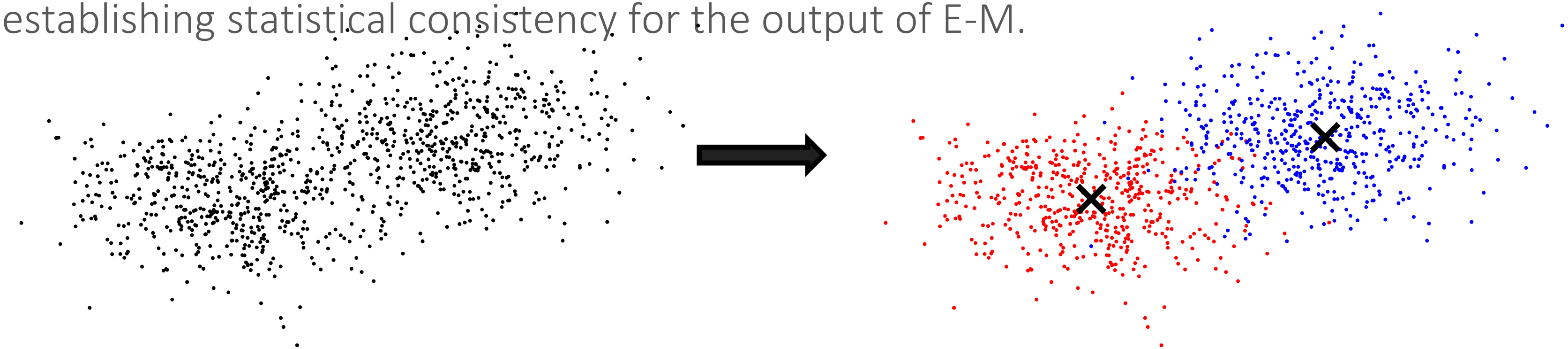
Our New Result

The first non-trivial global convergence analysis of E-M.

For uniform mixtures of two multivariate Gaussians with a (known) common covariance but different (unknown) means, we prove

- E-M iterates to converge to one of two equivalent global maximizers at linear rate, for all possible initializations, except in a particular measure-zero set.

- E-M gives the right answer for the data for which the model was designed, i.e., establishing statistical consistency for the output of E-M.



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

Using Big Data 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).

The Tumor Microbiome

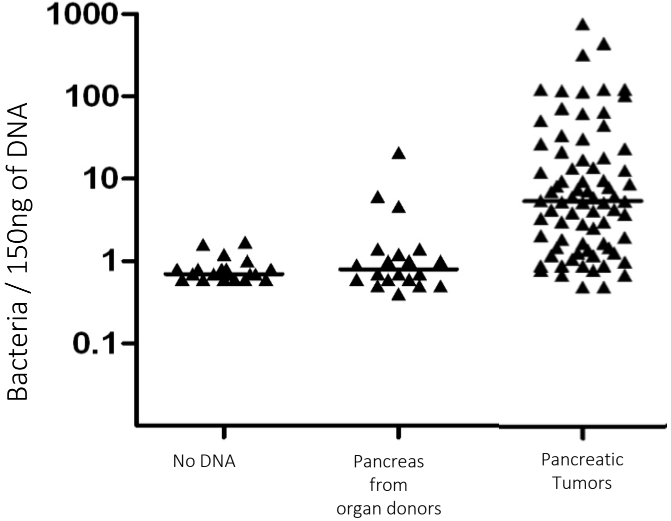
Tumors generally thought to be sterile environments

Microbiota survey finds *Lactobacillus* & *Bifidobacteria* on healthy breasts, while *Escherichia* & *Bacillus* predominate on cancerous breasts

Urbaniak, Camilla, et al. "Microbiota of human breast tissue." Applied and environmental microbiology 80.10 (2014): 3007-3014.
APA

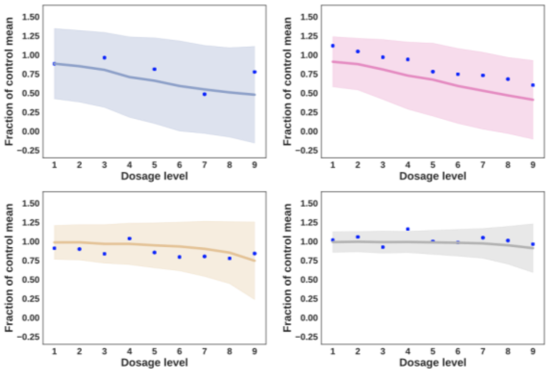
Xuan, Caiyun, et al. "Microbial dysbiosis is associated with human breast cancer." PloS one 9.1 (2014): e83744.

(Past literature)



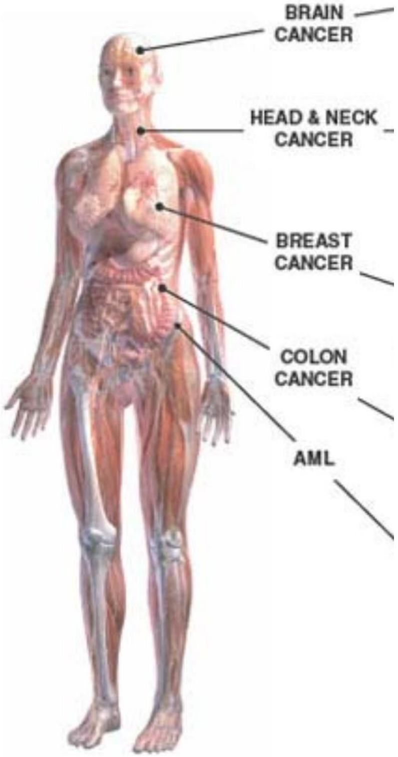
collaboration w/ Ravid Straussman (Weitzmann)
& Todd Golub (Broad Institute)

Predicting Personalized Cancer Therapies



David Blei (Statistics, Computer Science, Data Science Institute); Raul Rabadan (Systems Biology and Biomedical Informatics); Anna Lasorella (Pathology and Cell Biology and Pediatrics), Wesley Tansey (Systems Biology)

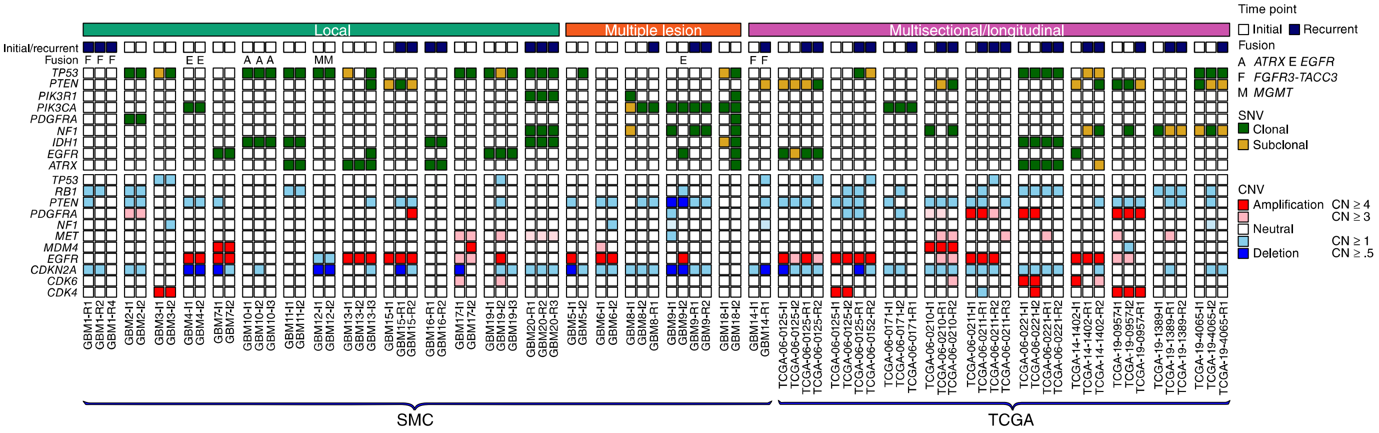
Genetically, No Two Tumors are Alike



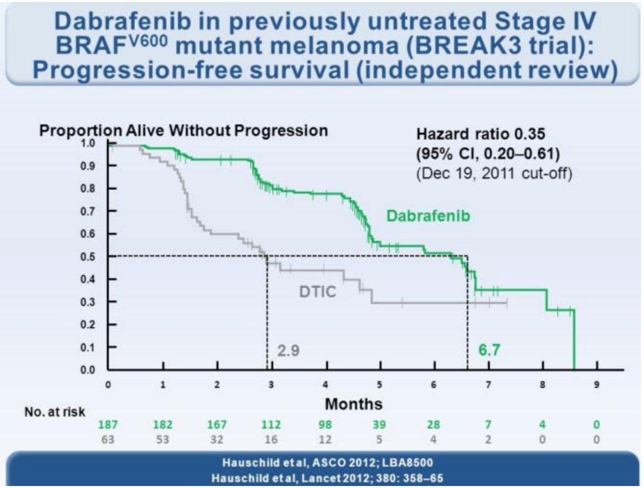
Historically, the location of the tumor determined the treatment

Puzzle: Why do some patients respond very well and others don’t?

Answer: Genomic makeup of tumors is diverse; each tumor is unique even for the same site of origin



Bringing Machine Learning and Cancer Research Together



Dabrafenib in previously untreated Stage IV BRAF^{V600} mutant melanoma (BREAK3 trial): Progression-free survival (independent review)

Hazard ratio 0.35
(95% CI, 0.20–0.61)
(Dec 19, 2011 cut-off)

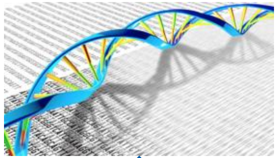
Hauschild et al, ASCO 2012; LBA8500
Hauschild et al, Lancet 2012; 380: 358–65

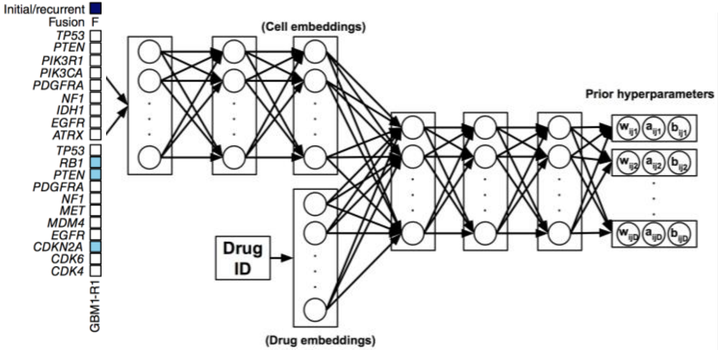
Current methods look at single targetable mutations

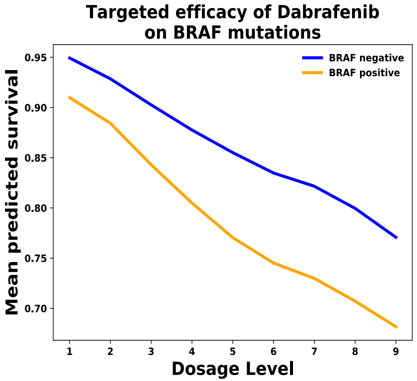
Some great success stories (e.g., BRAF-mutated melanoma)

Key idea: Can we use state-of-the-art machine learning methods to make treatment recommendations personalized to your specific tumor?

Goal of personalized medicine: Choosing the right drug, for the right patient, at the right dosage.







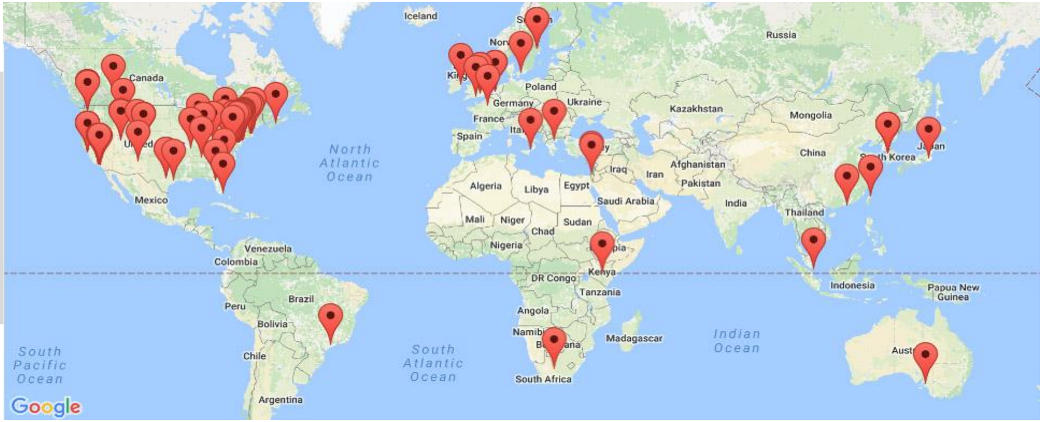
Targeted efficacy of Dabrafenib on BRAF mutations

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

Columbia University is the coordinating center

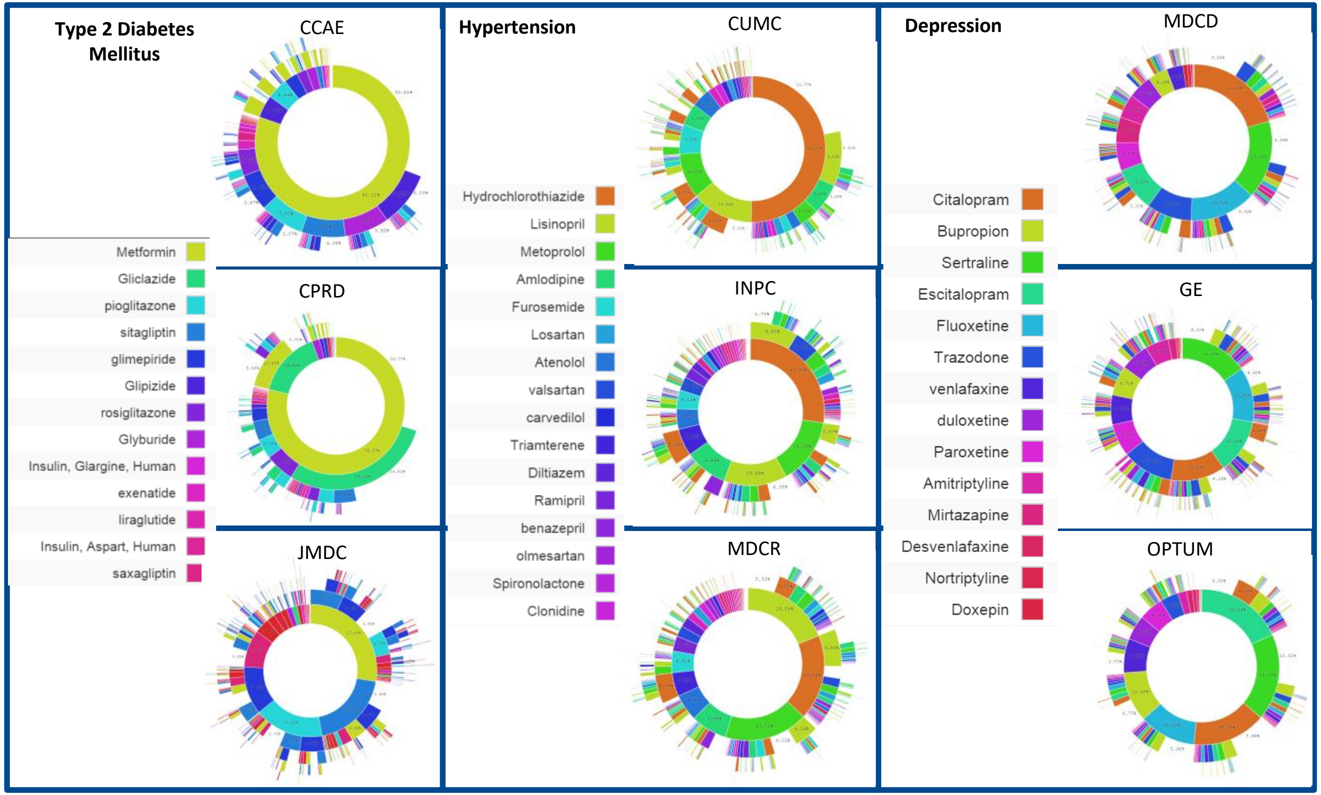
Goal: 1 billion patient records for observational research

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



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



Data Science Institute
COLUMBIA UNIVERSITY

COLUMBIA UNIVERSITY
Computational Social Science

ISERP
INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH AND POLICY

History Lab

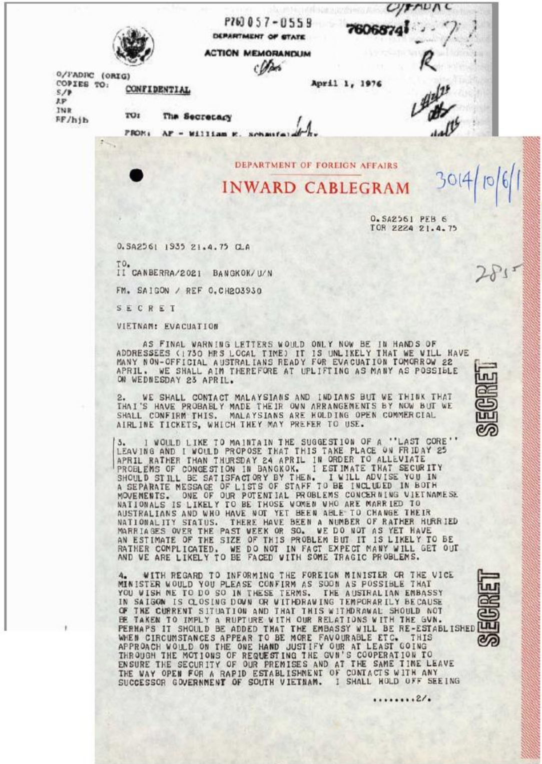
Team of data scientists, social scientists, and domain experts across Columbia and partners at MIT and Microsoft Research

NSF

MacArthur
Foundation

Assembling (What is Now) the World's Biggest Database of Declassified Documents

- The Foreign Relations of the United States (1932-1984). The ~200,000 most important declassified documents selected by State Department historians with access to every government department and agency.
- The State Department Central Foreign Policy Files (1973-1979). 3.2 million State Department records.
- Henry Kissinger Telephone Conversations (1973-1976). 4.5 thousand transcripts of Kissinger Telephone Conversations during his tenure as Secretary of State.
- The Hillary Clinton Emails (2009-2012). 51K individual messages from 32K email threads.
- British Cabinet Papers (1907-1990) 43K documents from the UK's most important decision-making body provided by the British National Archives.
- Azeredo da Silveira Papers (1974-1979) 10K personal papers of Brazil's foreign minister provided by The Center for Research and Documentation of the Contemporary History of Brazil
- In Process: 10 million pages of records declassified by the CIA, including the President's Daily Intelligence Briefs from 1961-1977



O/FADRC

P760057-0559

7606874

DEPARTMENT OF STATE

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. Schaufe

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

SECRET

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.

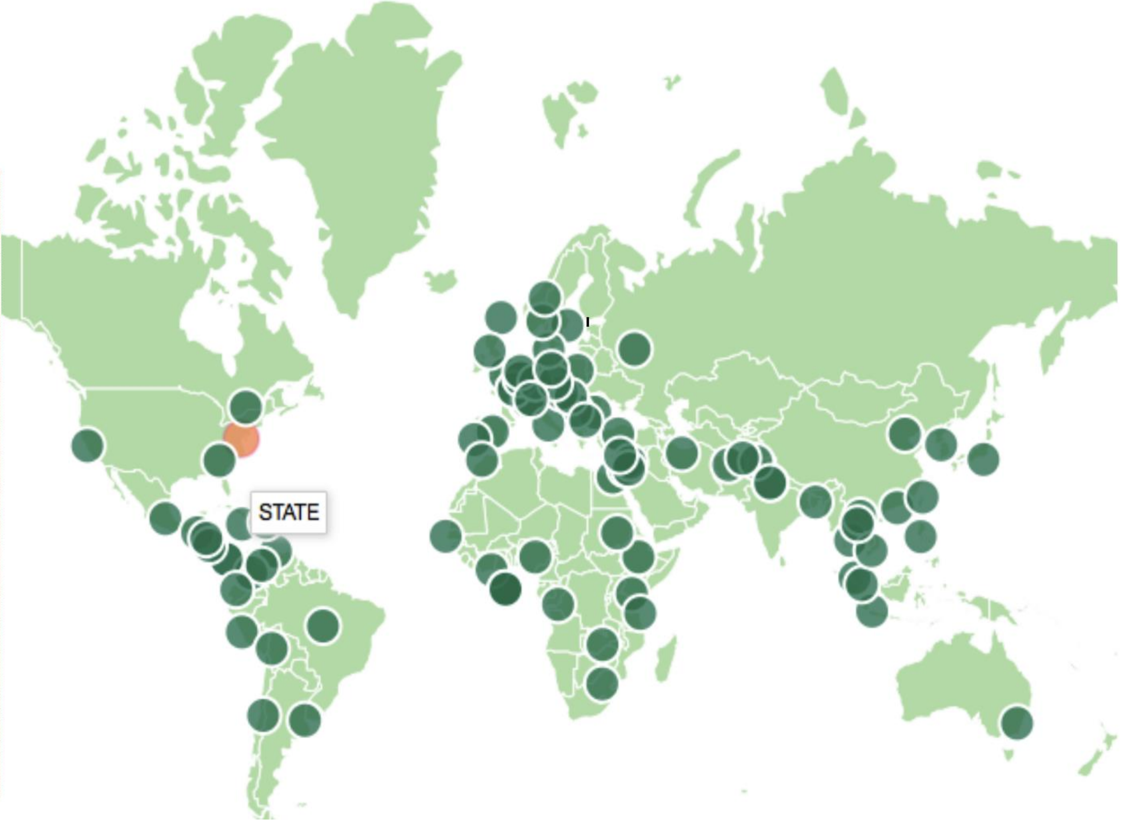
3. I WOULD LIKE TO MAINTAIN THE SUGGESTION OF A ''LAST CORE'' LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLEVIATE PROBLEMS OF CONGESTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN 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.

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

........2/.

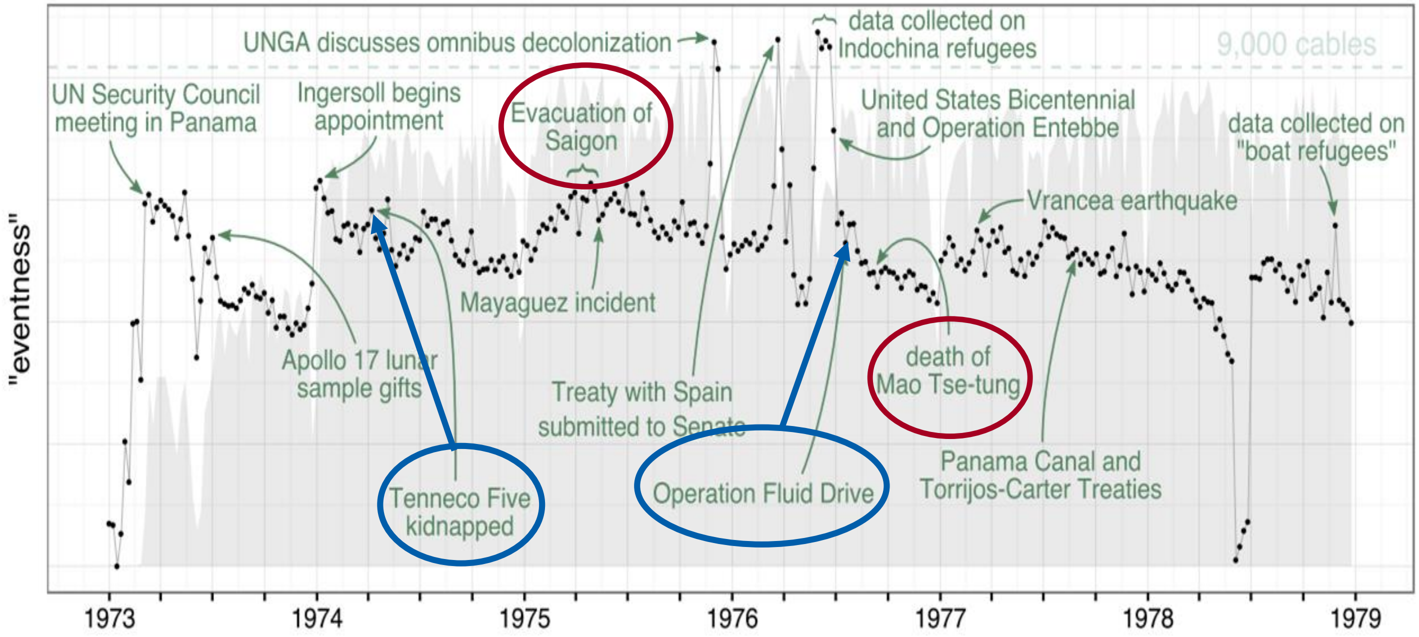
SECRET

SECRET



Allison J. B. Chaney[1], Hanna Wallach[2], **Matthew Connelly**[3], and **David M. Blei**[3]
[1]Princeton University [2]Microsoft Research [3]Columbia University

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



News + Context Drives Risk and Returns

Calomiris, Charles W. and Mamaysky, Harry, How News and Its Context Drive Risk and Returns around the World (April 1, 2017). Columbia Business School Research Paper No. 17-40. Available at SSRN: https://ssrn.com/abstract=2944826 or http://dx.doi.org/10.2139/ssrn.2944826

Novel Contributions

- Analyzed 51 developed and emerging markets

- Main result: The effect of news measures on market outcomes differs by country type and over time.
    - Topic-specific sentiment, frequency and ***unusualness*** of word flow suffice to predict future country-level returns, volatilities, and drawdowns.
    - New events cause more market reaction in ***developed*** than in emerging markets.
    - Economic and statistical significance are high and larger for ***year-ahead*** than monthly predictions.

- Context matters
    - Positive sentiment in Government, Corporate topics -> bad news
    - Positive sentiment in Market topics -> good news

Data for Good:

responsible use of data

J.M. Wing, “Data for Good: Scary AI and the Dangers of Big Data,’’ Institute for Social and Economic Research and Policy, Columbia University, New York, NY, September 20, 2017.

FATES

Fairness

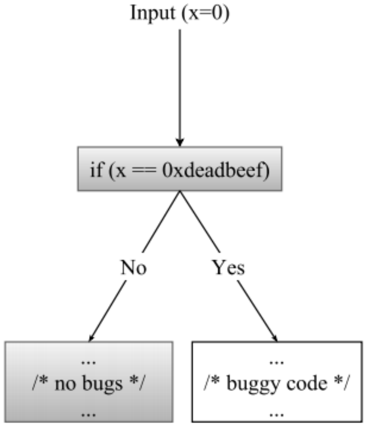
Accountability

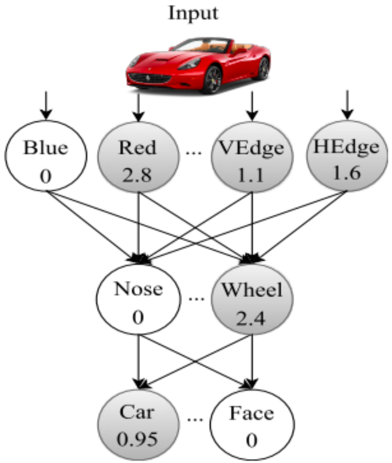
Transparency

Ethics

→ Safety and Security

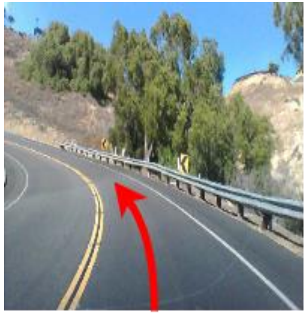
DeepXplore: Testing Deep Learning Systems



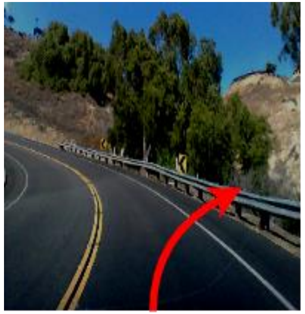


Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana, "Deep Xplore: Automated Whitebox Testing of Deep Learning Systems, *Proceedings of the 26th ACM Symposium on Operating Systems Principles*, October 2017, Best Paper Award.

DeepXplore



Seed,
No accident



Darker,
Accident

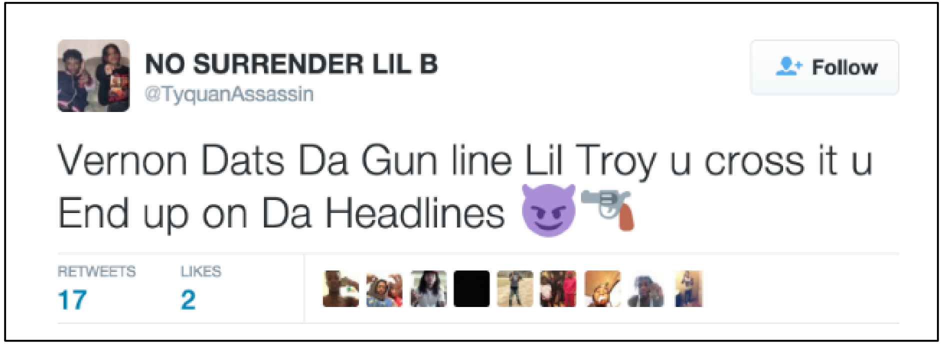
- Efficiently and systematically tests DNNS of hundreds of thousands of neurons without labeled data (only needs unlabeled seeds)
- Key ideas: neuron coverage (akin to code coverage), differential testing, and domain-specific constraints for focusing on realistic inputs
- Testing as a joint optimization 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

https://github.com/peikexin9/deepxplore

Data for Good:

tackling societal grand challenges

Intervening in Gang Violence



NO SURRENDER LIL B
@TyquanAssassin

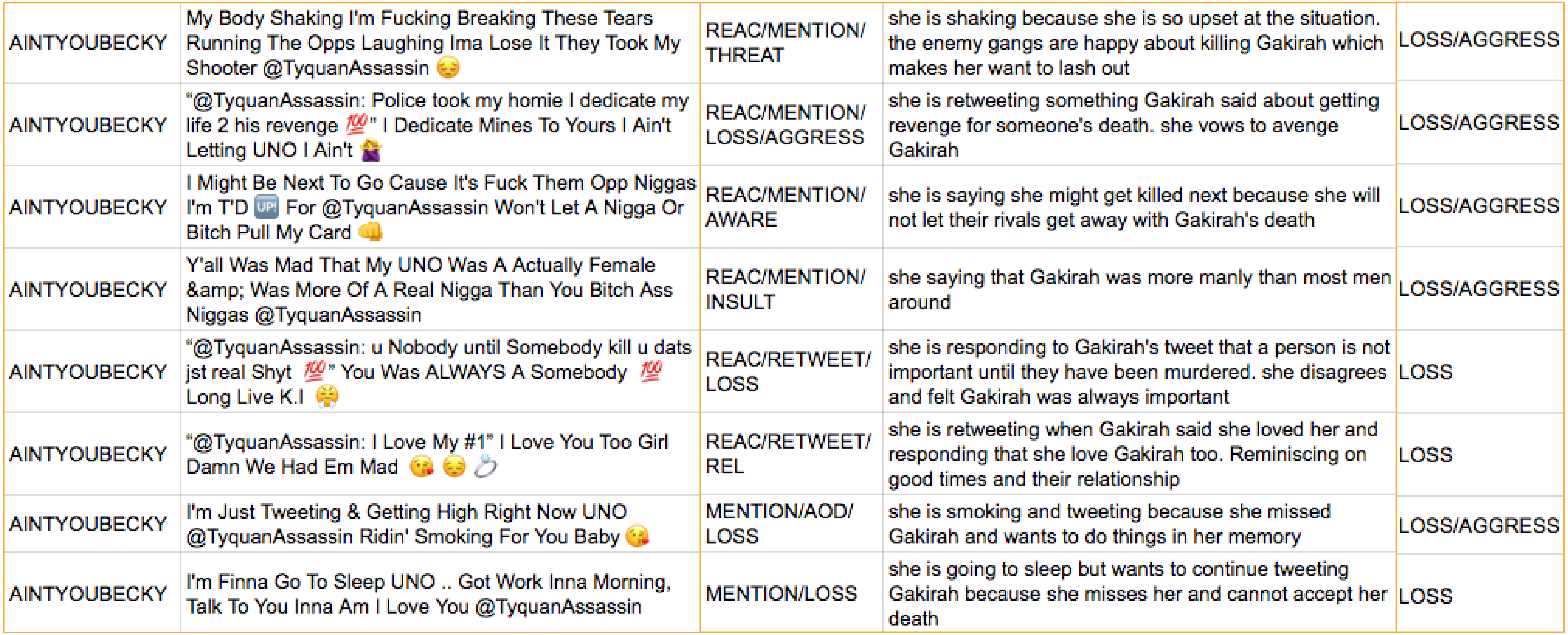
Vernon Dats Da Gun line Lil Troy u cross it u End up on Da Headlines 😈🔫

RETWEETS 17

LIKES 2

Terra Blevins, Robert Kwiatkowski, Jamie Macbeth, **Kathleen McKeown, Desmond Patton, and Owen Rambow,** "Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression," Int'l Conference for Computational Linguistics, October 2016.

Qualitative Analysis



Author	Content	Initial Code	Description	Final Code
AINTYOUBECKY	My Body Shaking I'm Fucking Breaking These Tears Running The Opps Laughing Ima Lose It They Took My Shooter @TyquanAssassin 😔	REAC/MENTION/ THREAT	she is shaking because she is so upset at the situation. the enemy gangs are happy about killing Gakirah which makes her want to lash out	LOSS/AGGRESS
AINTYOUBECKY	"@TyquanAssassin: Police took my homie I dedicate my life 2 his revenge 💯" I Dedicate Mines To Yours I Ain't Letting UNO I Ain't 🙅‍♀️	REAC/MENTION/ LOSS/AGGRESS	she is retweeting something Gakirah said about getting revenge for someone's death. she vows to avenge Gakirah	LOSS/AGGRESS
AINTYOUBECKY	I Might Be Next To Go Cause It's Fuck Them Opp Niggas I'm T'D 🆙 For @TyquanAssassin Won't Let A Nigga Or Bitch Pull My Card 👊	REAC/MENTION/ AWARE	she is saying she might get killed next because she will not let their rivals get away with Gakirah's death	LOSS/AGGRESS
AINTYOUBECKY	Y'all Was Mad That My UNO Was A Actually Female &amp; Was More Of A Real Nigga Than You Bitch Ass Niggas @TyquanAssassin	REAC/MENTION/ INSULT	she saying that Gakirah was more manly than most men around	LOSS/AGGRESS
AINTYOUBECKY	"@TyquanAssassin: u Nobody until Somebody kill u dats jst real Shyt 💯" You Was ALWAYS A Somebody 💯 Long Live K.I 😤	REAC/RETWEET/ LOSS	she is responding to Gakirah's tweet that a person is not important until they have been murdered. she disagrees and felt Gakirah was always important	LOSS
AINTYOUBECKY	"@TyquanAssassin: I Love My #1" I Love You Too Girl Damn We Had Em Mad 😘 😔 💍	REAC/RETWEET/ REL	she is retweeting when Gakirah said she loved her and responding that she love Gakirah too. Reminiscing on good times and their relationship	LOSS
AINTYOUBECKY	I'm Just Tweeting & Getting High Right Now UNO @TyquanAssassin Ridin' Smoking For You Baby 😘	MENTION/AOD/ LOSS	she is smoking and tweeting because she missed Gakirah and wants to do things in her memory	LOSS/AGGRESS
AINTYOUBECKY	I'm Finna Go To Sleep UNO .. Got Work Inna Morning, Talk To You Inna Am I Love You @TyquanAssassin	MENTION/LOSS	she is going to sleep but wants to continue tweeting Gakirah because she misses her and cannot accept her death	LOSS

Novel Contributions

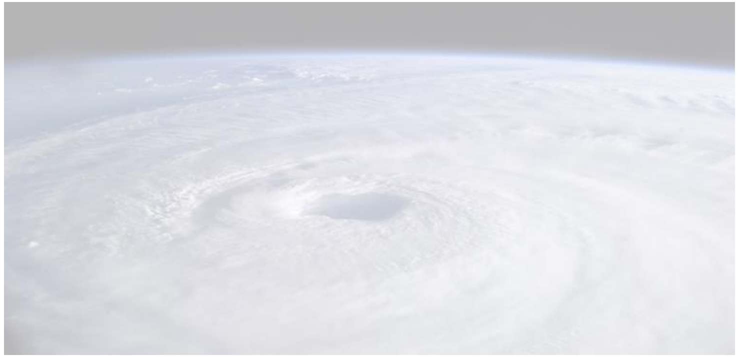
New corpus annotated with discourse intention based on deep read of text and part-of-speech (POS) tags: http://dx.doi.org/10.7916/D84F1R07

NLP resources for the sub-language used by Chicago gang members, specifically POS tagger and glossary

System to identify the emotion conveyed by tweets, using the Dictionary of Affect in Language, specifically ***loss*** or ***aggression***.

Future: Study how close the relationship is between expressions of aggression on Twitter and real world aggression

Pangeo: Big Data and Climate Science

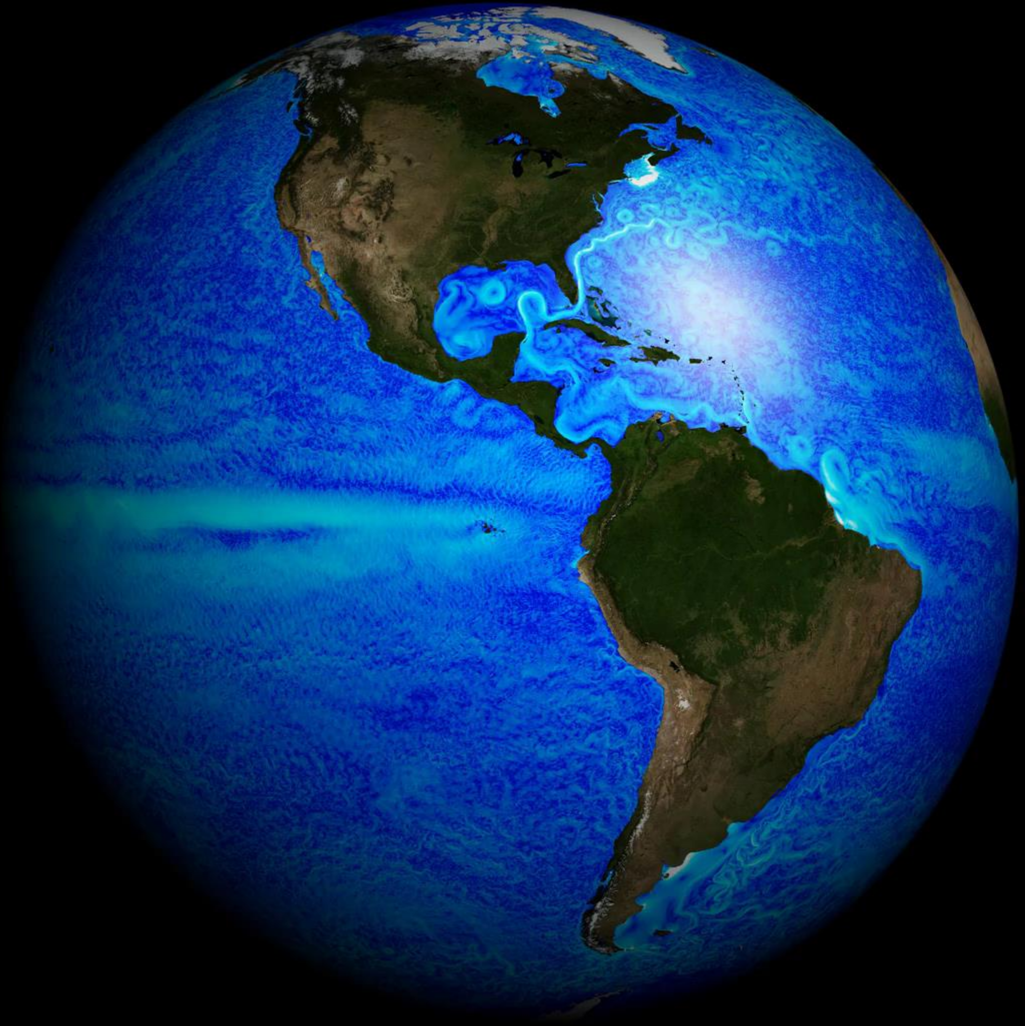


PI: Prof. Ryan Abernathey (Dept. of Earth & Env. Sci., LDEO, Columbia University

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)

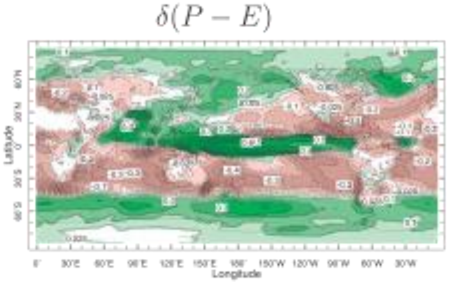
Collaborators: Gavin Schmidt (APAM, Frontiers in Cptg. Systems (DSI), NASA Goddard Institute for Space Studies (director)),
V. Balaji (National Oceanographic and Atmospheric Administration Geophysical Fluid Dynamics Lab)

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



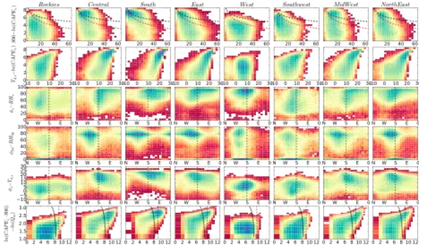
Applications of Pangeo

The Water Cycle of The Global Atmosphere (Henderson, Seager)



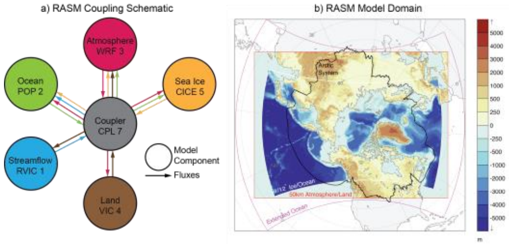
https://doi.org/10.1175/JCLI-D-13-00018.1

Understanding Severe Thunderstorms (Lepore, Tippett)



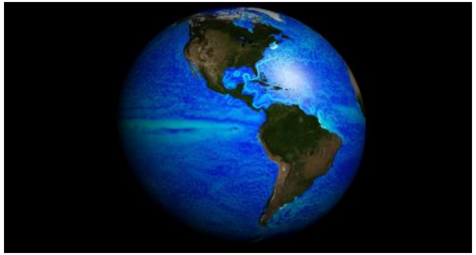
https://doi.org/10.1175/BAMS-D-16-0208.1

Improving Regional Hydrologic Modeling (Hamman)



https://doi.org/10.1175/JCLI-D-15-0415.1

Energetics of Ocean Turbulence (Abernathey)



https://doi.org/10.1175/JPO-D-14-0160.1

Data for Good

Thank You