

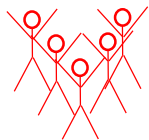
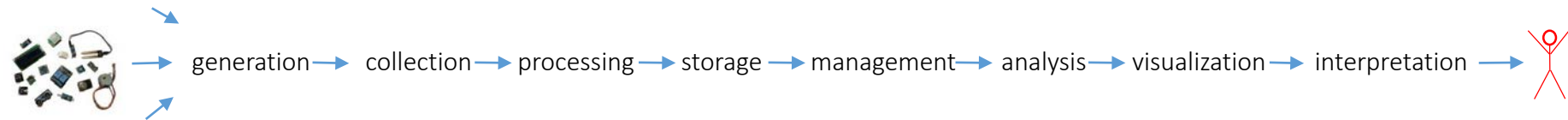
Data for Good: Data Science at Columbia

Jeannette M. Wing

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

Federal University of Rio de Janeiro
Brazil
May 9, 2018

Data life cycle



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

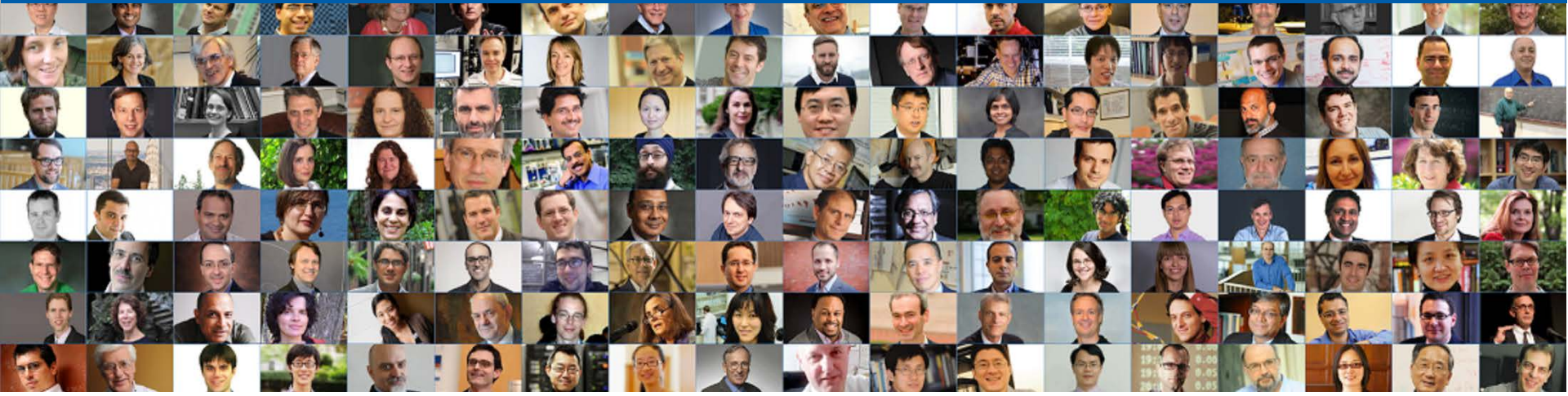


12 Schools
300+ Faculty

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



Foundations



Cybersecurity



Data, Media and Society



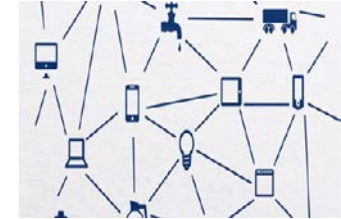
Financial Analytics



Health Analytics



Smart Cities



Sense, Collect, and Move



Computational Social Science



Computing Systems



Materials Discovery Analytics

Education

Degree:

Master of Science

2014



Non-Degree:

Certification

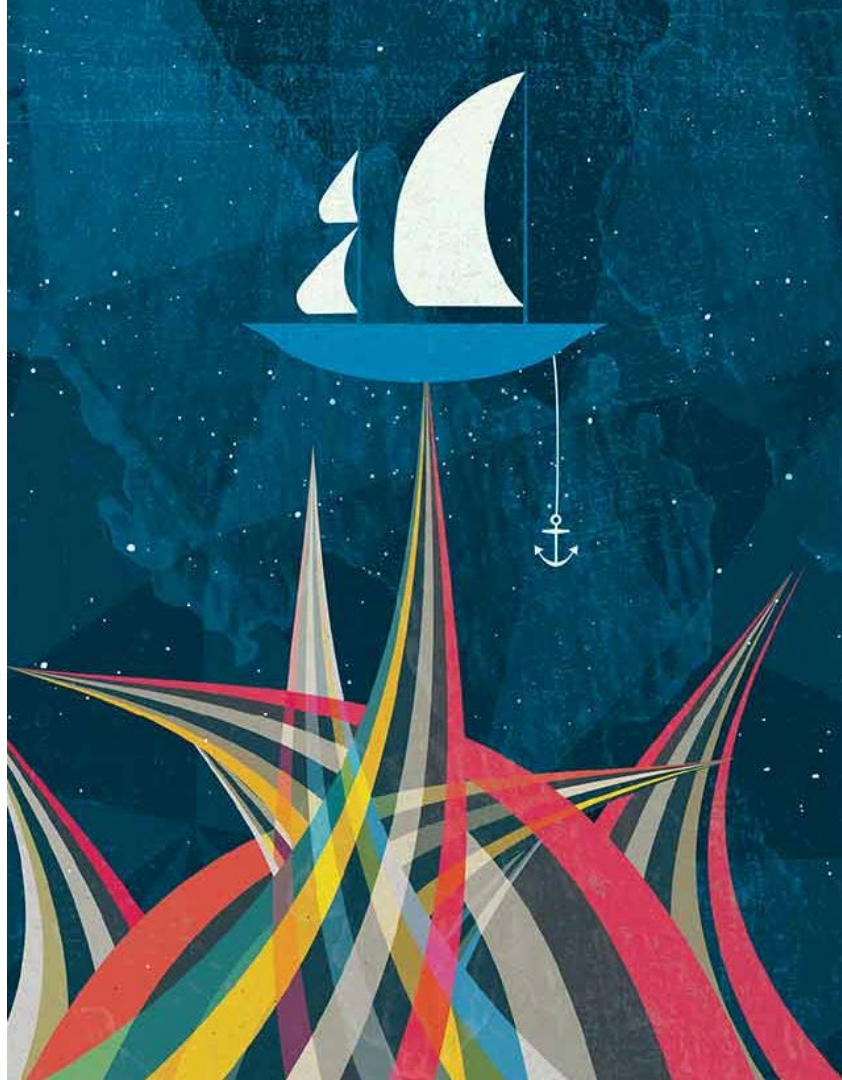
2013



ColumbiaX:

Online Courses

2016



Industry Affiliates Program



Bloomberg®



citi VENTURES



Goldman
Sachs



Google Cloud

J.P.Morgan

Johnson & Johnson



Microsoft Research

MOODY'S

synergic
partners

2σ TWO SIGMA



Vanguard®

industry.datascience.columbia.edu

Capstone Projects

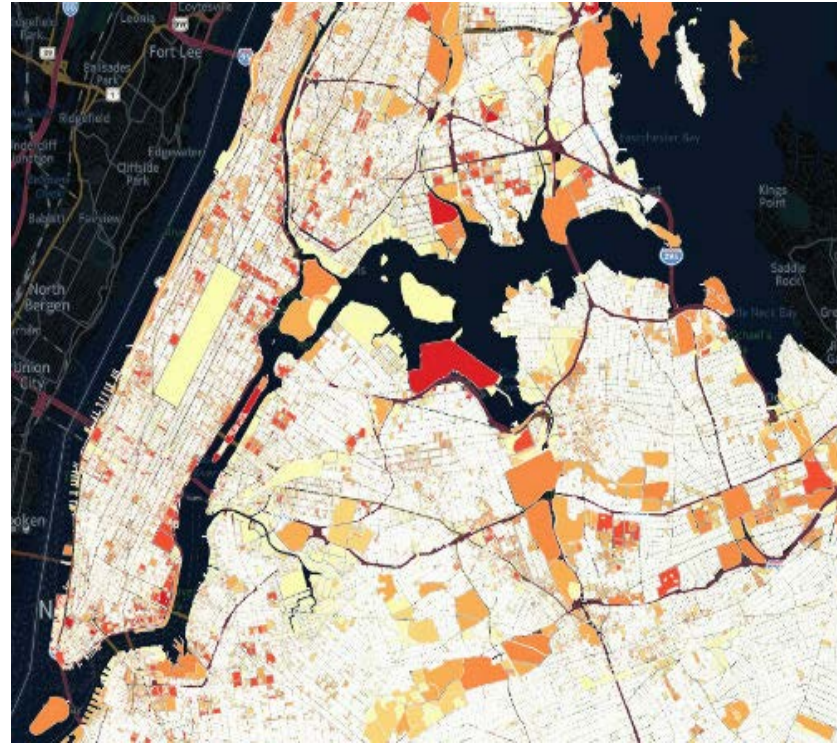
Students working with Industry Affiliates

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.



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.



DSI Feeds New York City's Start Up Community



[Agolo](#) algorithmically curates your twitter feed



[Wordseye](#) lets anyone create 3D images



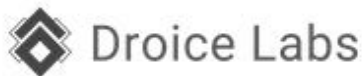
[eBrevia](#) uses machine learning to summarize legal documents



[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

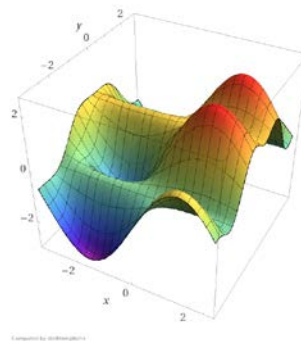
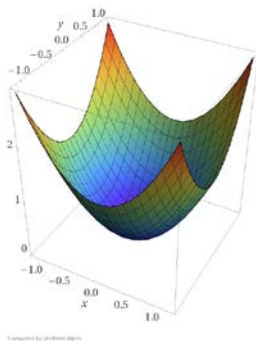
Mission

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

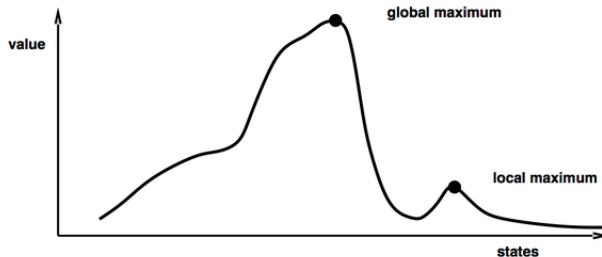


Ji Xu, Daniel Hsu, Arian Maleki, “Global analysis of Expectation-Maximization for mixtures of two Gaussians,” NIPS 2016

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.**

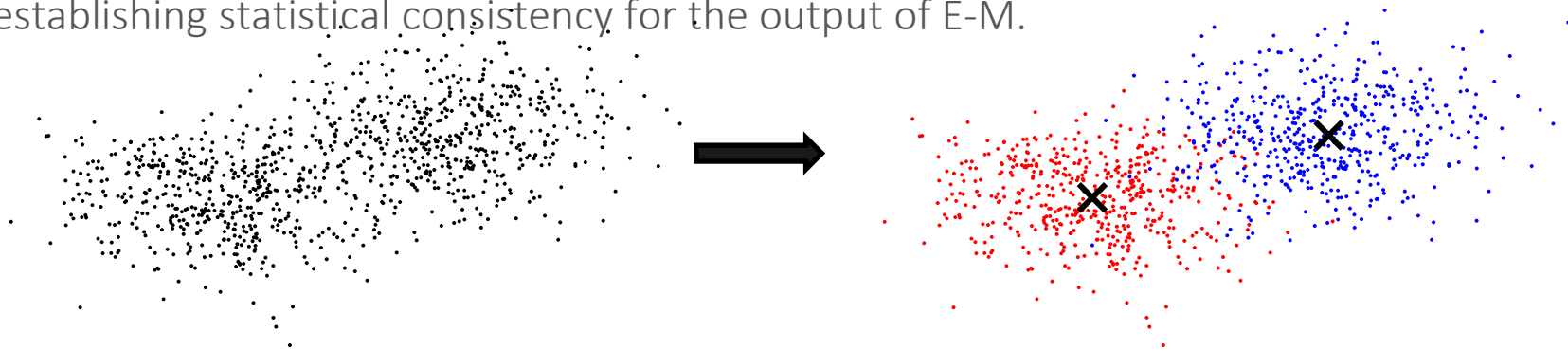


Our New Result

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

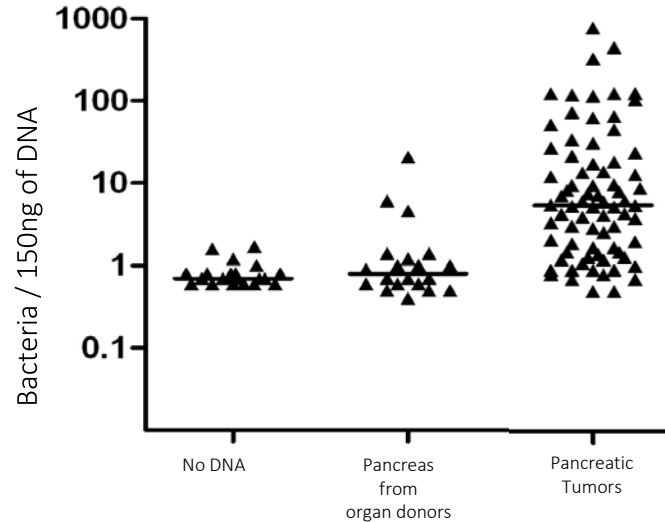


Urbaniak, Camilla, et al. "Microbiota of human breast tissue." *Applied and environmental microbiology* 80.10 (2014): 3007-3014.

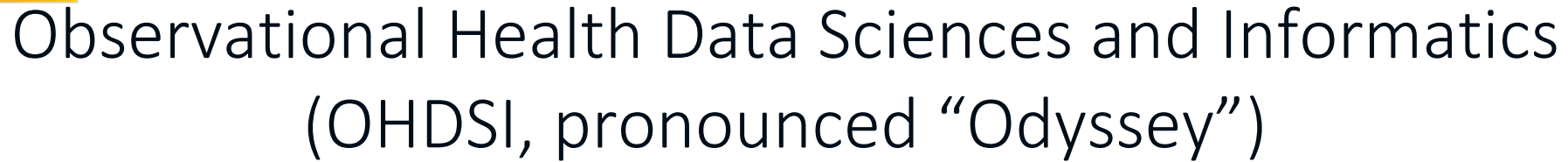
APA

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

(Past literature)

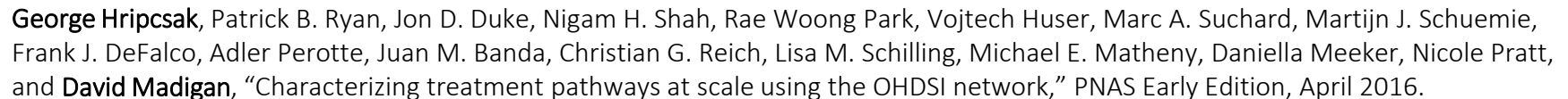


collaboration w/ Ravid Straussman (Weitzmann)
& Todd Golub (Broad Institute)

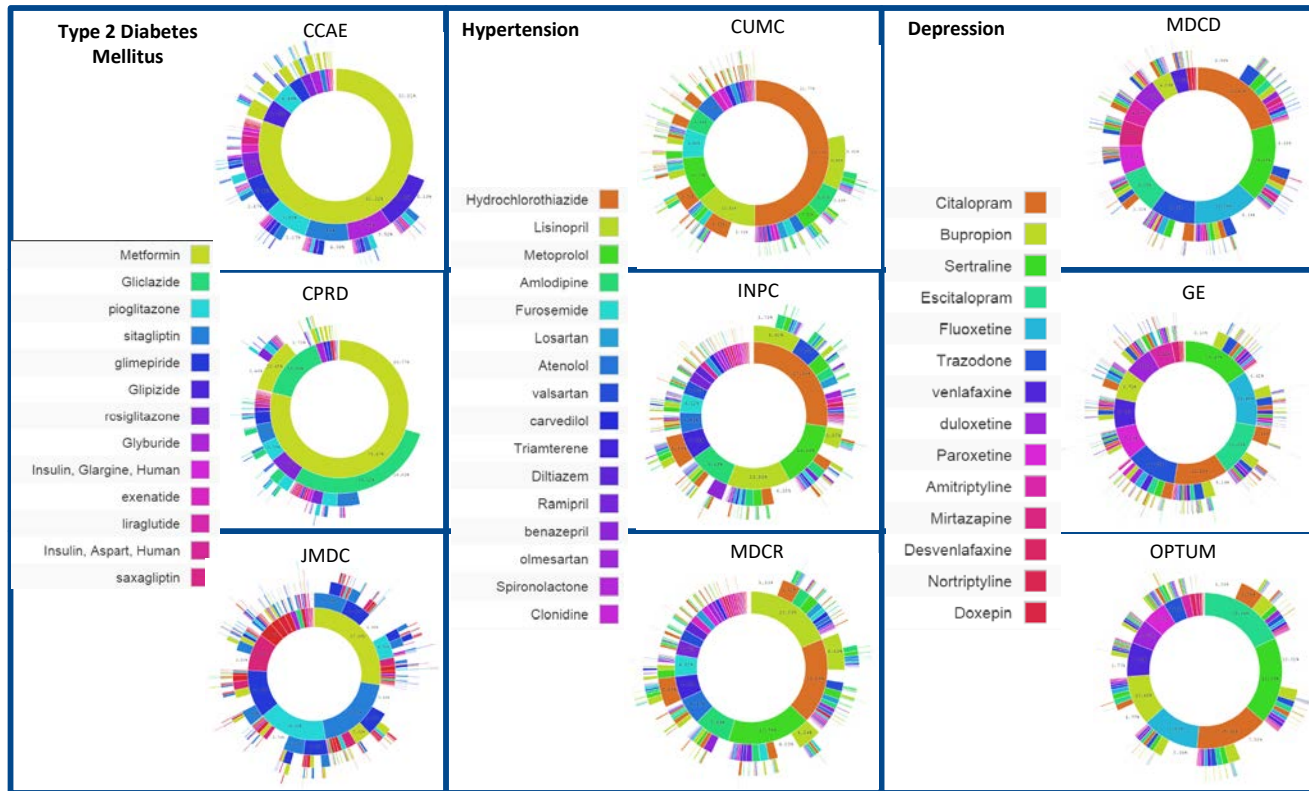


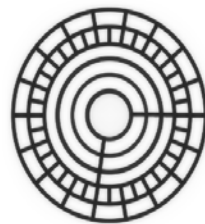
Goal: 1 billion patient records
for observational research

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



Heterogeneity of Observational Research Results





History Lab

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



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

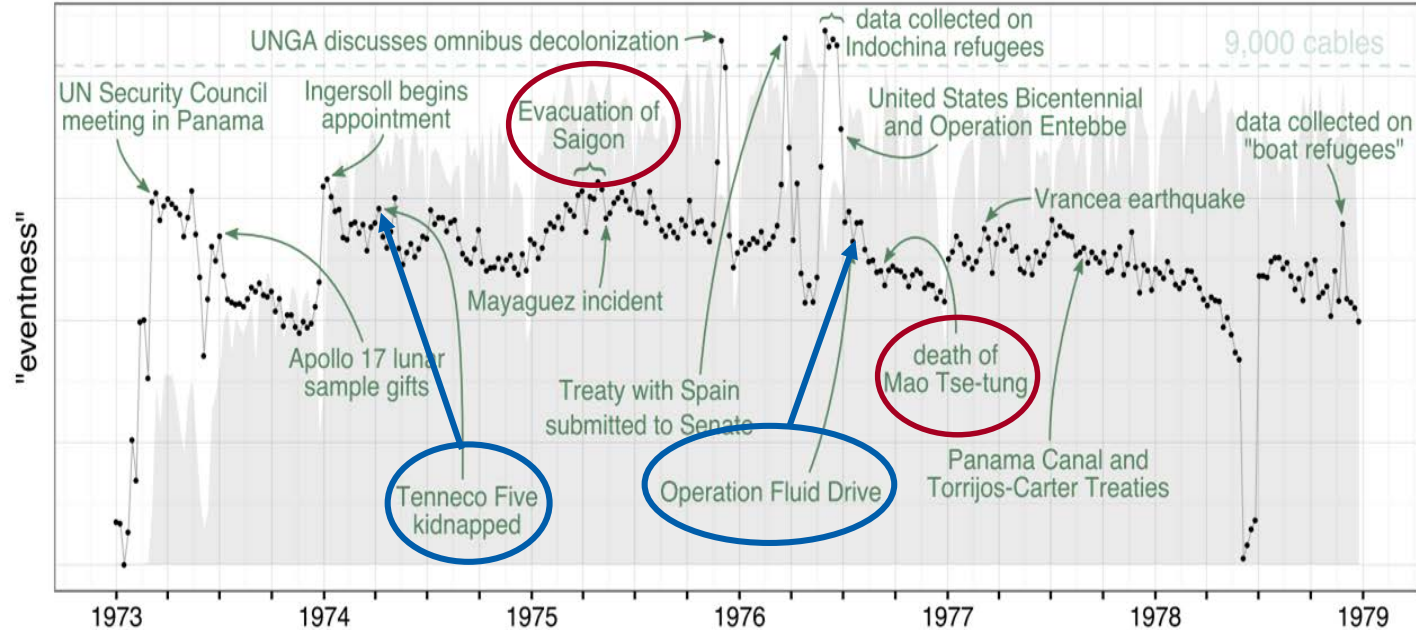
PNO 857-0559
 DEPARTMENT OF STATE
 ACTION MEMORANDUM
 April 1, 1976
 CONFIDENTIAL
 TO: The Secretary
 FROM: AF - William W. Schmeidler
 304/10/6/1
 281
 DEPARTMENT OF FOREIGN AFFAIRS
 INWARD CABLEGRAM
 0.343561 1305 21.4.75 CLS
 TO: CANBERRA/2001 BAWKOR/018
 FM: SAIGON / REF 0.04023530
 3 K C N K 1
 VIETNAM EVACUATION
 AS FINAL WARNING LETTERS WOULD ONLY NOW BE IN HANDS OF ADDRESSEES (120 REF LOCAL TIME) IT IS UNLIKELY THAT WE WILL HAVE MANY NON-OFFICIAL AUSTRALIAN READY FOR EVACUATION TOMORROW 22 APRIL. WE SHALL AIR THEREFORE AT UNLITTING AS EARLY AS POSSIBLE ON WEDNESDAY 25 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 COME" LEAVING AND I WOULD PROPOSE THAT THIS TAKE PLACE ON FRIDAY 25 APRIL RATHER THAN THURSDAY 24 APRIL IN ORDER TO ALLOCATE PROBLEMS OF CONNECTION IN BANGKOK. I ESTIMATE THAT SECURITY SHOULD STILL BE SATISFACTORY BY THEN. I WILL ADVISE YOU IN A SEPARATE MESSAGE OF A LOTS OF DEATH TO BE INFLICTED IN SUCH MOVEMENTS. ONE OF OUR POTENTIAL PROBLEMS CONCERNING VIETNAMESE NATIONALS IS LIKELY TO BE THOSE WOMEN WHO ARE MARRIED TO AUSTRALIANS AND WHO MORE OR LESS HAVE BEEN 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 FACE 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 THOSE TERMS. THE AUSTRALIAN EMBASSY IN SAIGON IS CLOSING DOWN OR WITHDRAWING TEMPORARILY BECAUSE OF THE CURRENT SITUATION AND THIS THIS WITHDRAWAL SHOULD NOT BE TAKEN TO IMPLY A RUPTURE WITH OUR RELATIONS WITH THE DRV. PERHAPS IT SHOULD BE ADDED THAT THE PROBABLY WILL BE RE-EVALUATED WHEN CIRCUMSTANCES APPEAR TO BE MORE FAVOURABLE ETC. THIS APPROACH WOULD ON THE ONE HAND JUSTIFY OUR AT LEAST COMING THROUGH THE NOTIONS OF RESPECTING THE DRV'S COOPERATION TO ENSURE THE SECURITY OF OUR FRIENDS AND AT THE SAME TIME LEAVE THE WAY OPEN FOR A RAPID ESTABLISHMENT OF CONTACT WITH ANY SUCCESSOR GOVERNMENT OF SOUTH VIETNAM. I SHALL HOLD OFF SEEING



Allison J. B. Chaney¹, Hanna Wallach², **Matthew Connelly**³, and David M. Blei³

¹Princeton University ²Microsoft Research ³Columbia 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

FATES

Fairness

Accountability

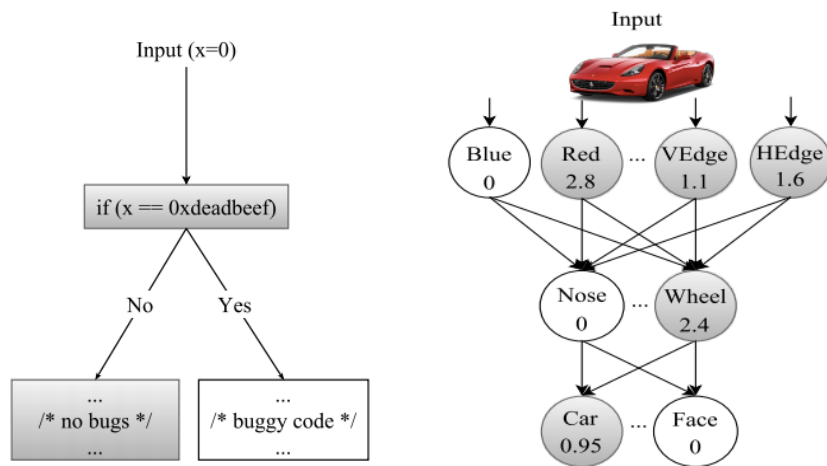
Transparency

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



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 & 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



EARTH CUBE

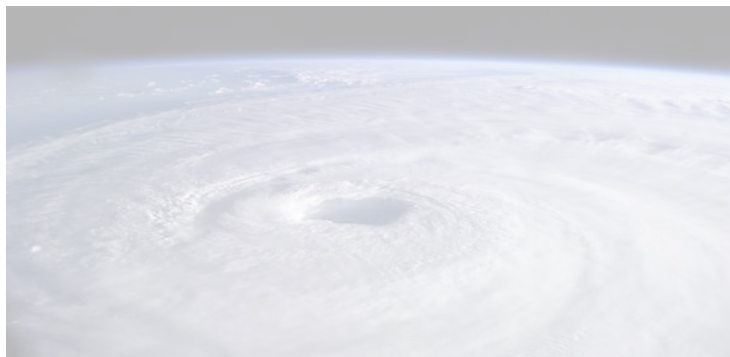


Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

COLUMBIA UNIVERSITY
Data Science Institute



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)),

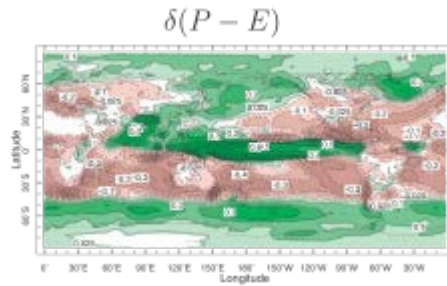
V. Balaji (National Oceanographic and Atmospheric Administration Geophysical Fluid Dynamics Lab)

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



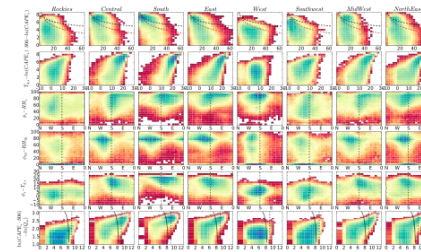
Applications of Pangeo

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



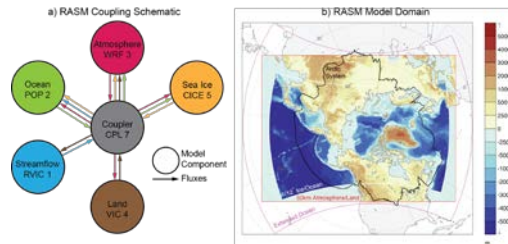
<https://doi.org/10.1175/JCLI-D-13-00018.1>

Understanding Severe Thunderstorms
(Lepore, Tippett)



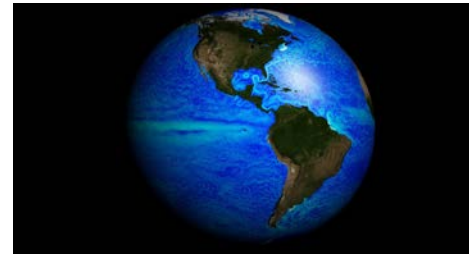
<https://doi.org/10.1175/BAMS-D-16-0208.1>

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

