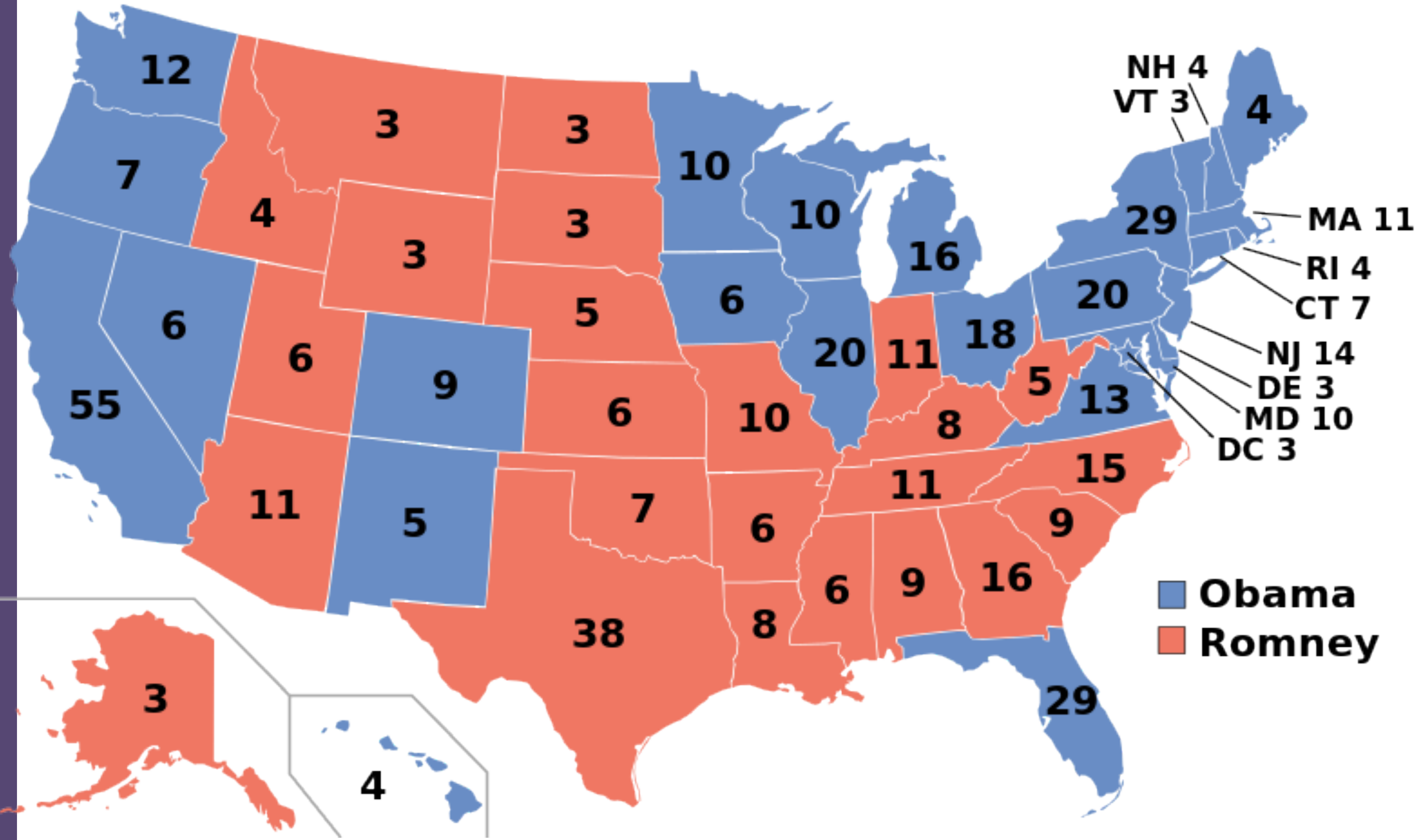
UNIVERSITY *of* WASHINGTON

Data Science III: Scaling, Applications, Ethics (Data Science in the Wild)

INFX 575

Today…

- A “spanning basis” of data science examples
- Perspectives on data science as a discipline
- Course logistics



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Nate Silver

source: randy stewart

“The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes.”

Nate Silver, Oct. 26, 2012

fivethirtyeight.com

“…the argument we’re making is exceedingly simple. Here it is: Obama’s ahead in Ohio.”

Nate Silver, Nov. 2, 2012

fivethirtyeight.com

“The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign.”

Nate Silver, Nov. 10, 2012

DailyBeast

Related: Obama campaign's data-driven ground game

"In the 21st century, the candidate with [the] best data, merged with the best messages dictated by that data, wins."

Andrew Rasiej, Personal Democracy Forum

"…the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

My view:

Data science is about answering questions using large, noisy, and heterogeneous datasets

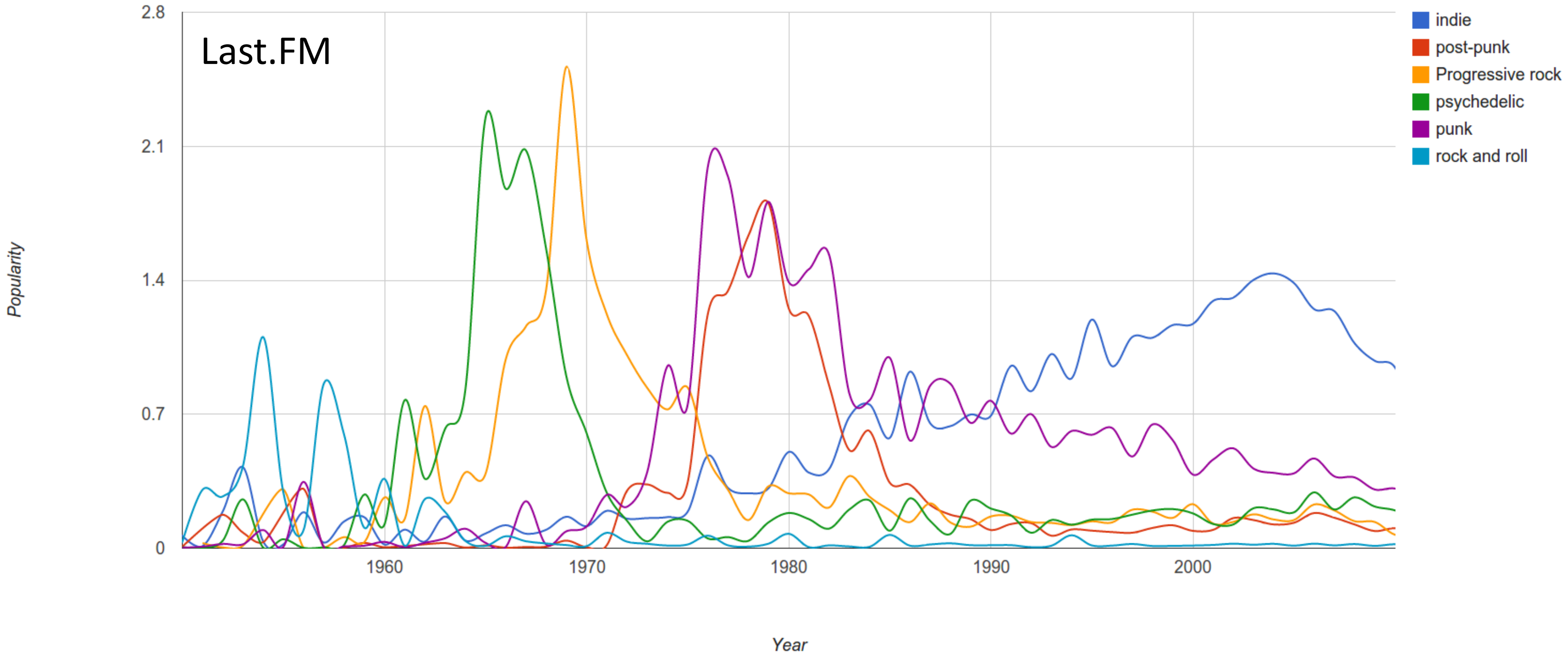
The technology is as important as the methods

But the results are more important than both

last.fm

Question:
How has the popularity of genres of music changed over time?

Data:
List of artists with user-supplied genre tags; playlists from users



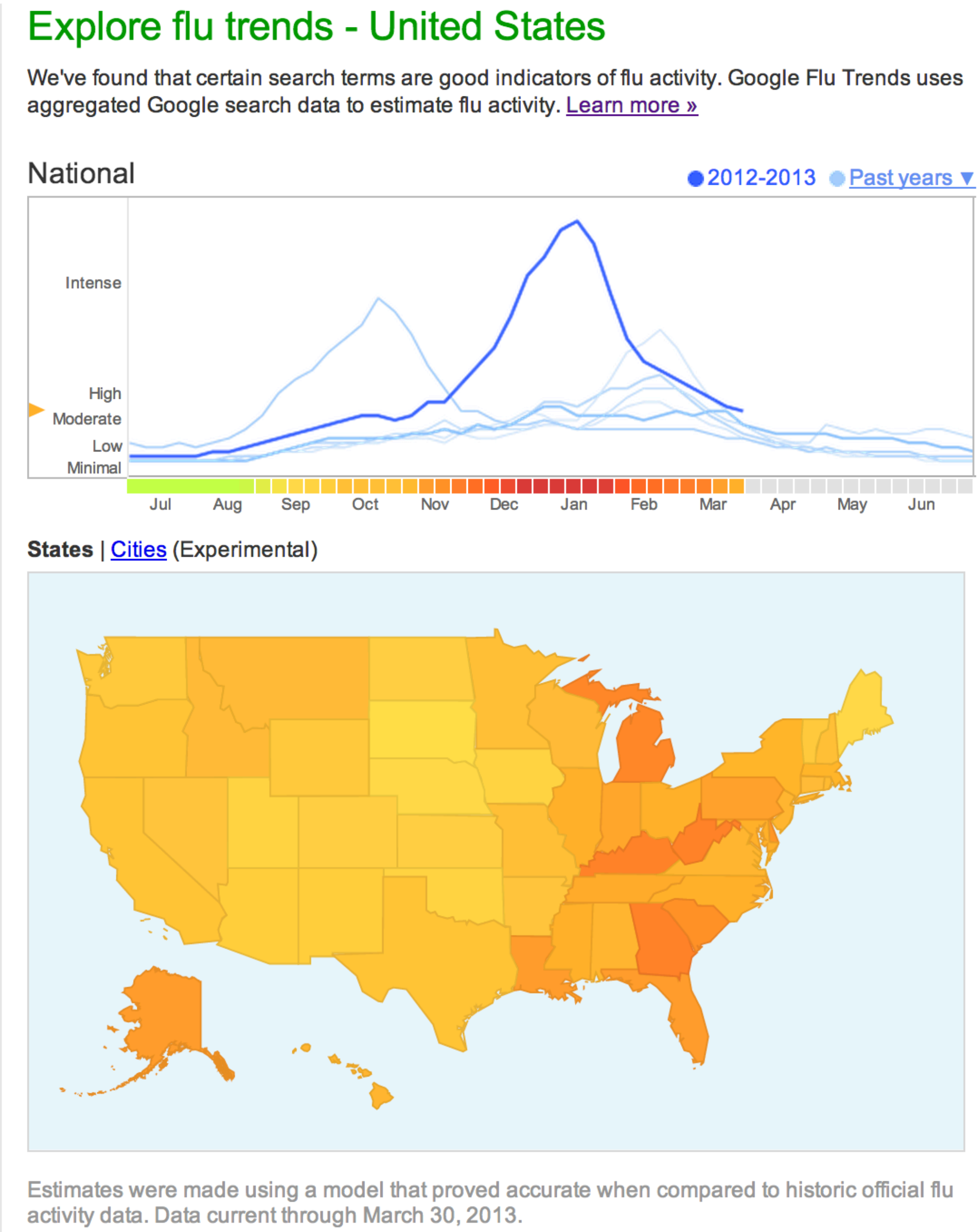
“Since we have a massive amount of user tag data available we can easily correlate tags and years and measure “popularity” of a genre by counting the number of artists formed in a specific year.”

Janni Kovacs, Last.FM

Google

Question:
How early and accurately can we predict flu outbreaks, so we can plan production levels of flu vaccine?

Dataset:
Search histories of users



Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)

Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.

source: http://www.google.org/flutrends/us/#US

flu risk

“Scientific hindsight shows that Google Flu Trends far overstated this year's flu season….”

“Lots of media attention to this year's flu season skewed Google's search engine traffic.”

David Wagner, Atlantic Wire, Feb 13 2013

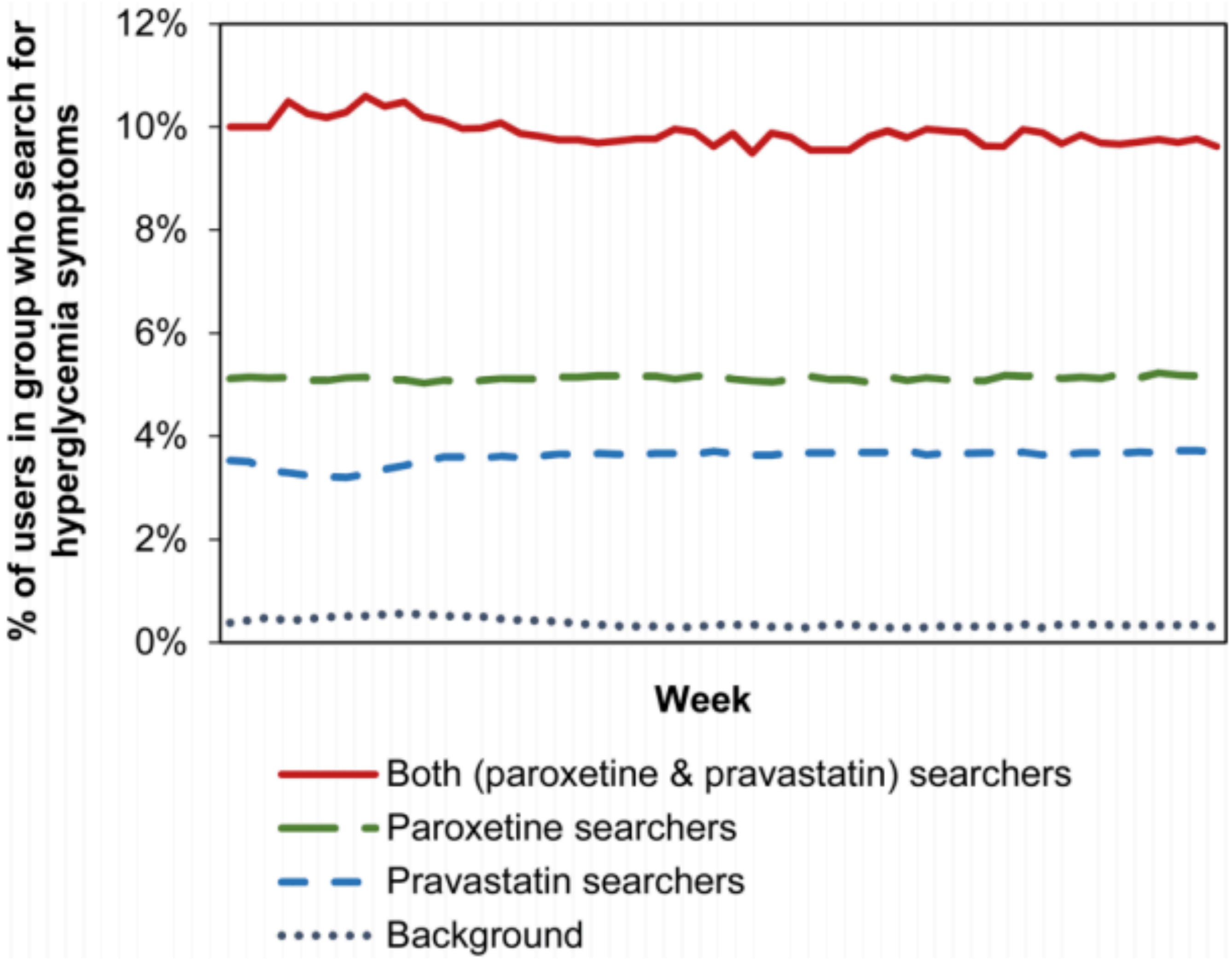
http://www.google.com/permissions/using-product-graphics.html

Question:

Do people that take paroxetine and pravastatin together exhibit hypoglycemia symptoms?

Dataset:

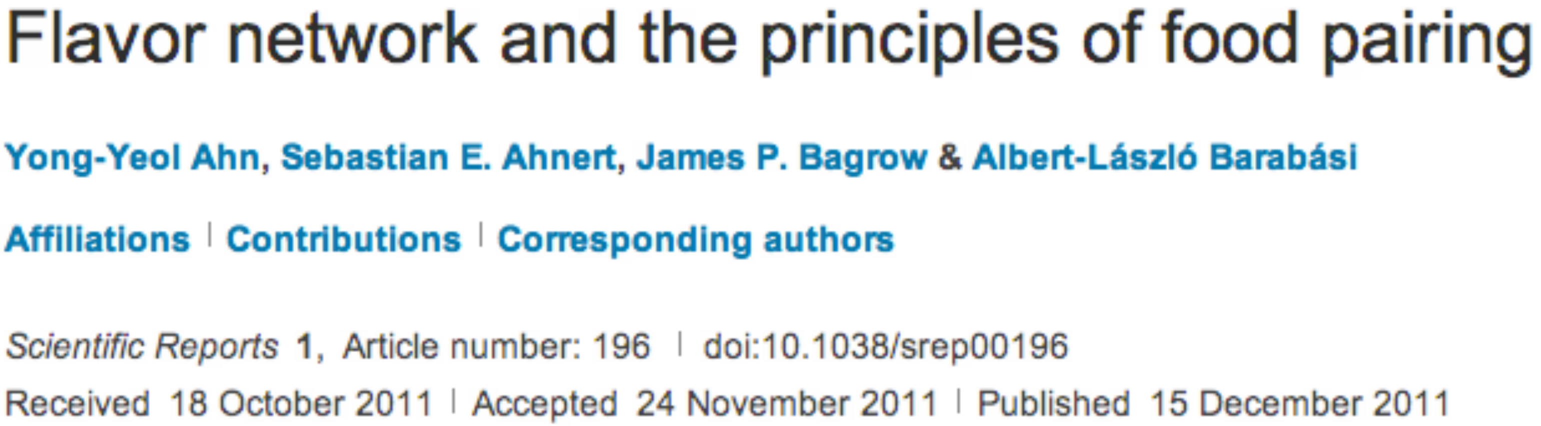
Search engine histories



Ryen W White,Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, Eric Horvitz, ***Web-scale pharmacovigilance: listening to signals from the crowd****, J Am Med Inform Assoc, March 2013, doi:10.1136/amiajnl-2012-001482*

Question:
What pairs of foods go well together?

Data:
Large repositories of recipes



Flavor network and the principles of food pairing

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási

Affiliations | Contributions | Corresponding authors

Scientific Reports **1**, Article number: 196 | doi:10.1038/srep00196

Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.

Question:

Has the expression of emotion in the literature changed over time?

Dataset:

Literature over the last 100 years

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

1) Convert all the digitized books in the 20th century into n-grams (Thanks, Google!) (http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

2) Label each 1-gram (word) with a mood score. (Thanks, WordNet!)

3) Count the occurences of each mood word

sum of all the occurrences of words with a particular mood

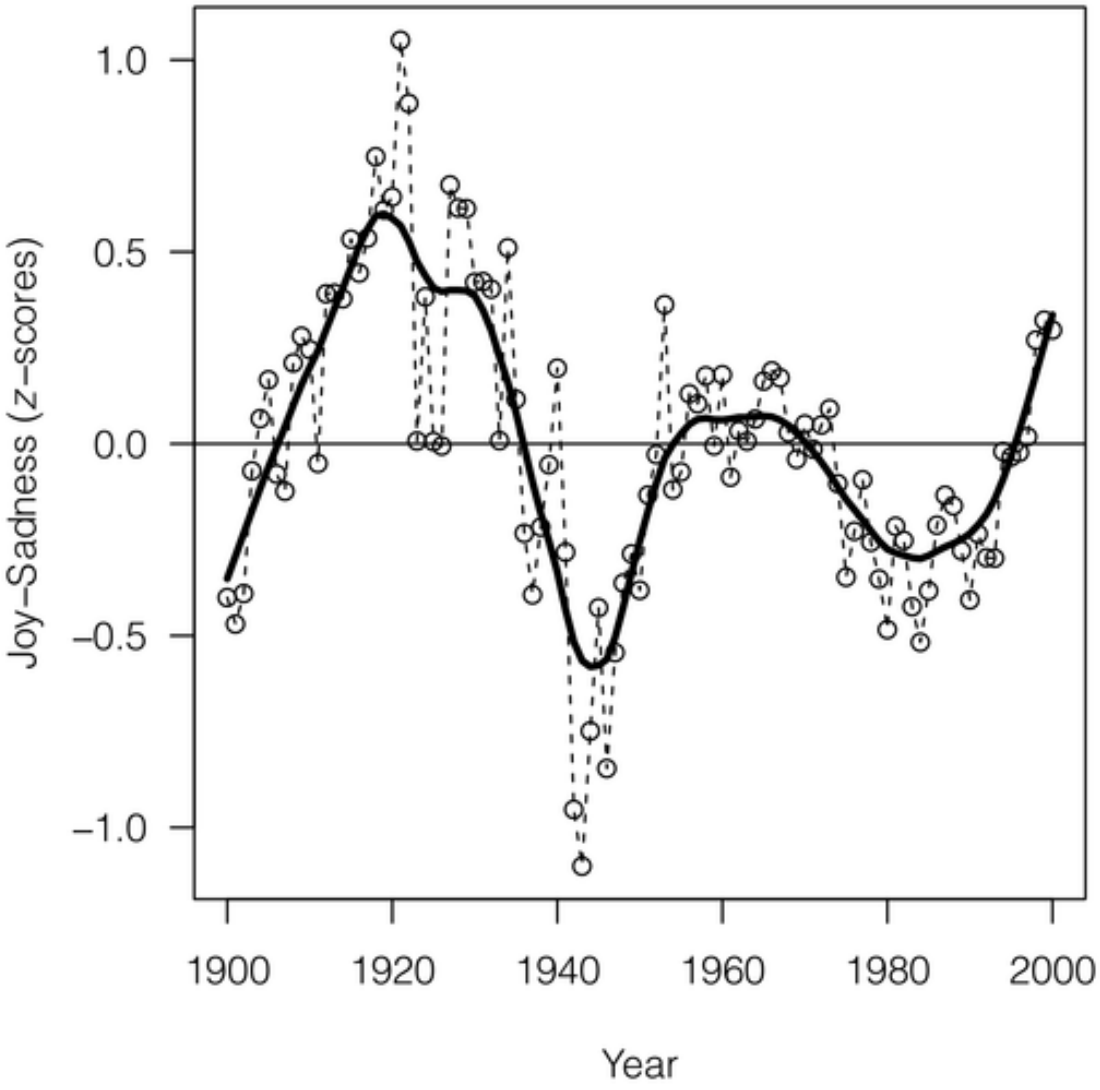
number of occurrences of word i

$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

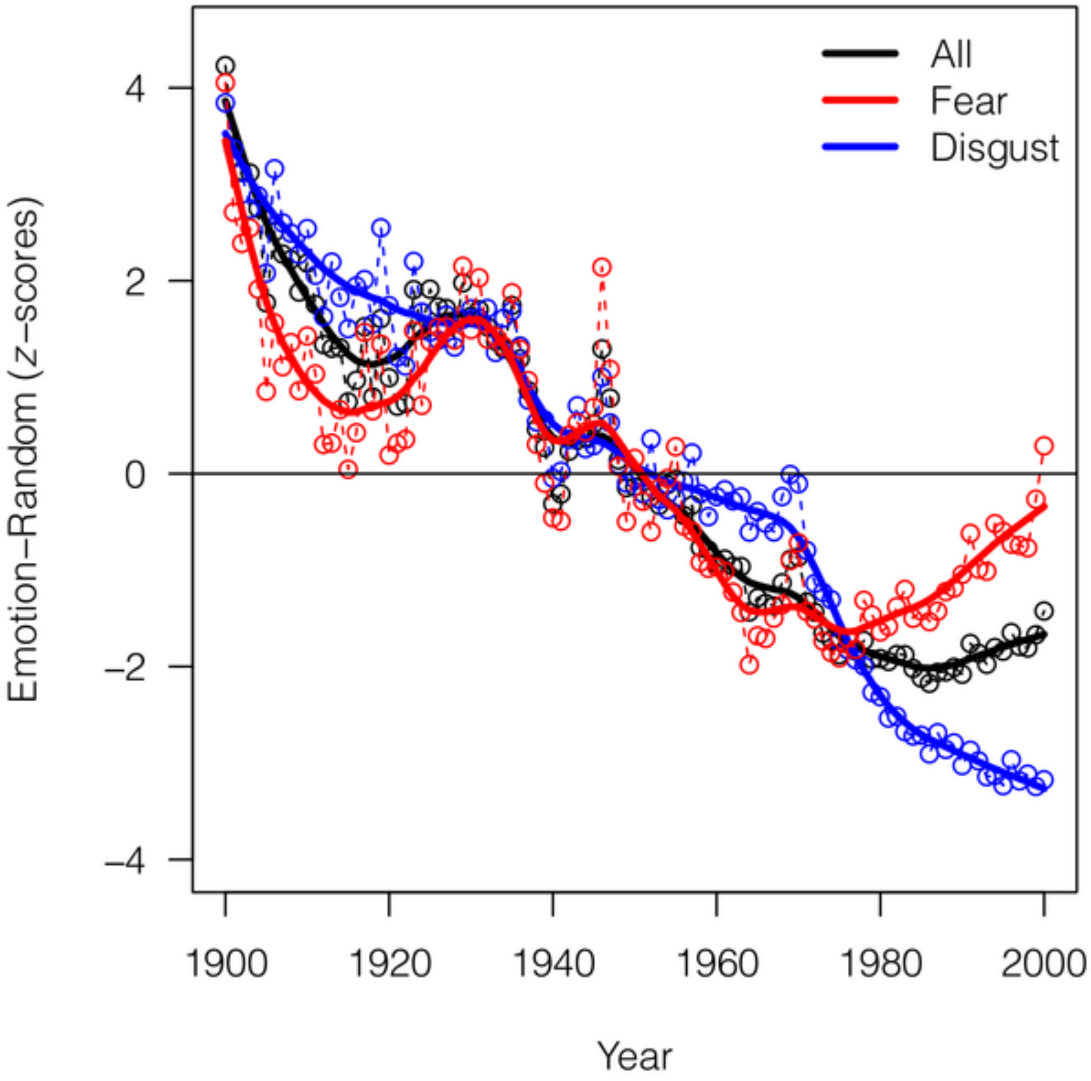
number of occurrences of the word “the”

divide by the number of words with that mood label

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030



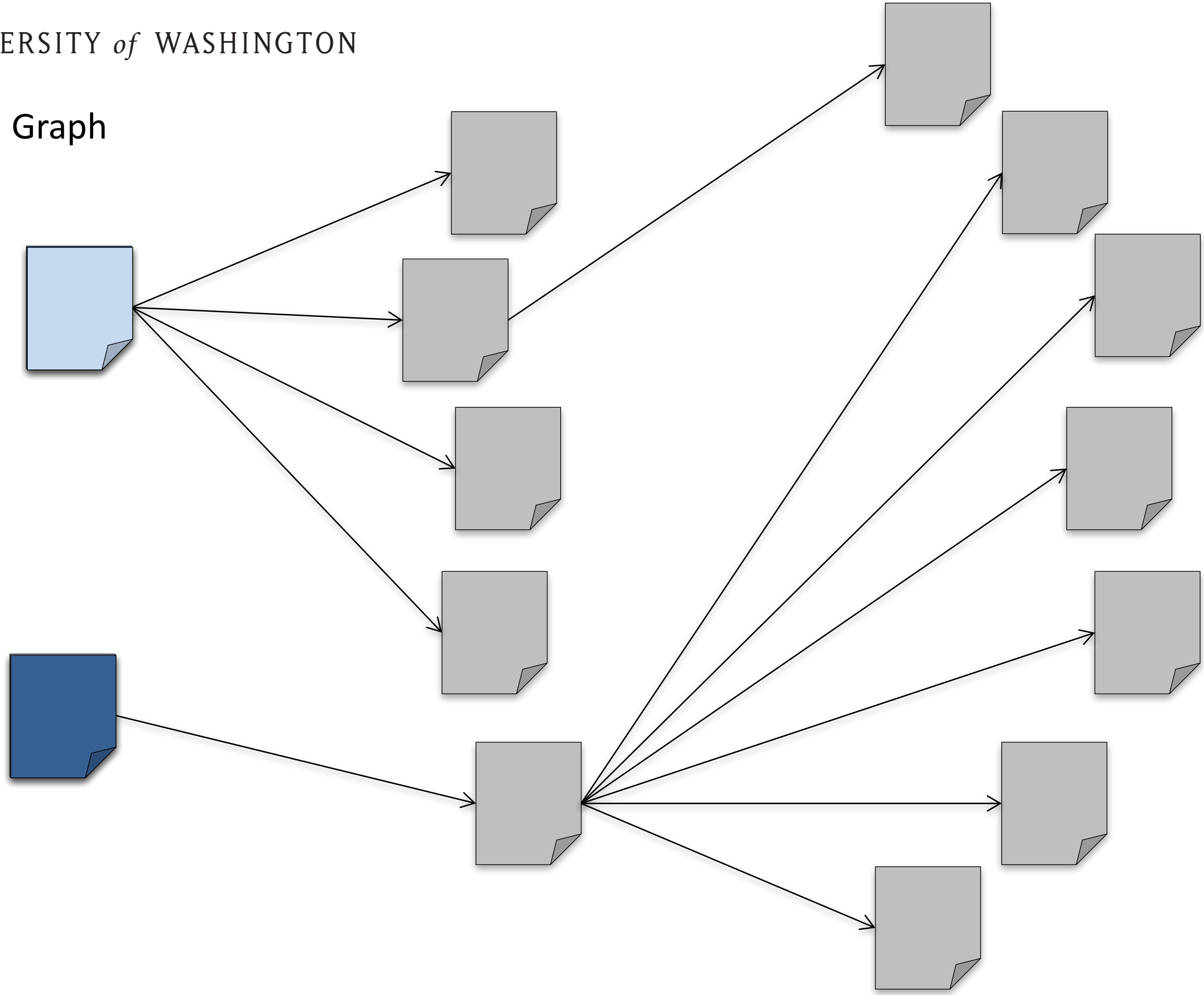
Question:

How do we measure the influence of a scientific paper?

Dataset:

Citation network between papers

Citation Graph



Key Takeaways:

Data science is pervasive

It’s about repurposing data collected in some other context

It’s about extracting signal from large, noisy, heterogeneous sources

It’s about insight, not always sophisticated mathematics

October 22, 2012

Six Italian seismologists convicted of manslaughter for failing to predict magnitude 6.3 earthquake in April 2009.

Locals were concerned about seismic activity; researchers deemed "too reassuring" in the verdict.

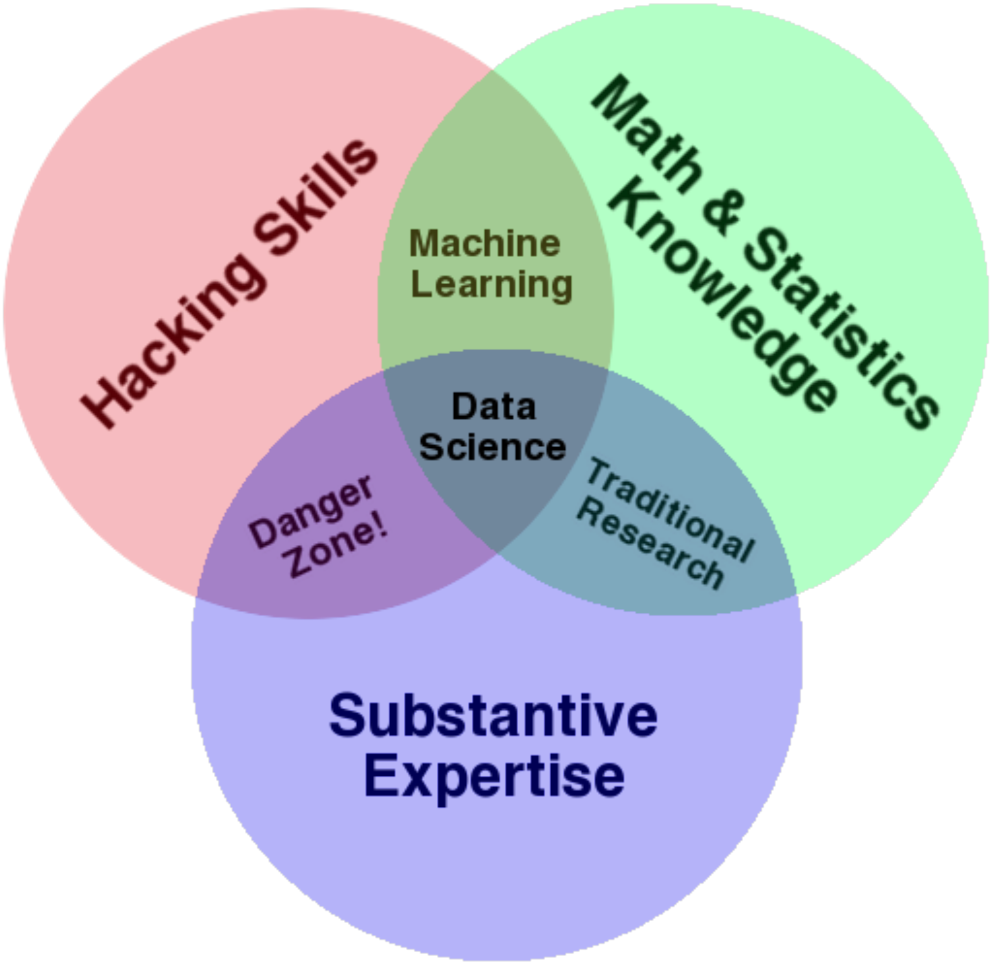
Credit: TheWiz83, Creative Commons Attribution-ShareAlike 3.0 Unported

PERSPECTIVES ON DATA SCIENCE

Three views I'd like to share

- Skills Perspective
  - Drew Conway's Venn Diagram
- Task Perspective
  - Data Science "Workflow"
- Output Perspective
  - "Data Products"

Skills Perspective:
Drew Conway's Data Science Venn Diagram



"I worry that the Data Scientist role is like the mythical "webmaster" of the 90s: master of all trades."

-- Aaron Kimball, CTO Wibidata

Task Perspective:
A Typical Data Science Workflow

DB

1) Preparing to run a model

"80% of the work"

-- Aaron Kimball

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

ML/Stats

2) Running the model

Vis

3) Interpreting the results

"The other 80% of the work"

Output perspective:
Data Science is about *Data Products*

- "Data-driven apps" (Mike Loukides)
  - Spellchecker
  - Machine Translator
- Interactive visualizations
  - Google flu application
  - Global Burden of Disease
- Online Databases
  - Enterprise data warehouse
  - Sloan Digital Sky Survey

Data science is about building data products, not just answering questions

Data products empower others to use the data.

May help communicate your results (e.g., Nate Silver's maps)

May empower others to do their own analysis
(e.g., Global Burden of Disease)

The Data Scientific Method

- *Start with a Question*
- *Leverage your current data*
- *Create features and run tests*
- *Analyze the results and draw insights*
- *Let the data frame a conversation*

src: DJ Patil, Josh Elman, LeWeb, London, 2012

- Science
  - Start with a question
- Data Science
  - Start with a question, *and typically some existing data*

Key Takeaways:

Skills perspective: programming, mathematical rigor, domain expertise

Task perspective: wrangling, modeling, communicating

Output perspective: delivering useful data products, not just answers in a vacuum

BIG DATA LIMITATIONS

Big Data vs. Statistics

- Statistics is about drawing inferences about a population based on the properties of a sample
- What if we have access to (almost) the entire population?
  - All the credit card transactions
  - All the papers
  - All the clicks
  - All the photographs

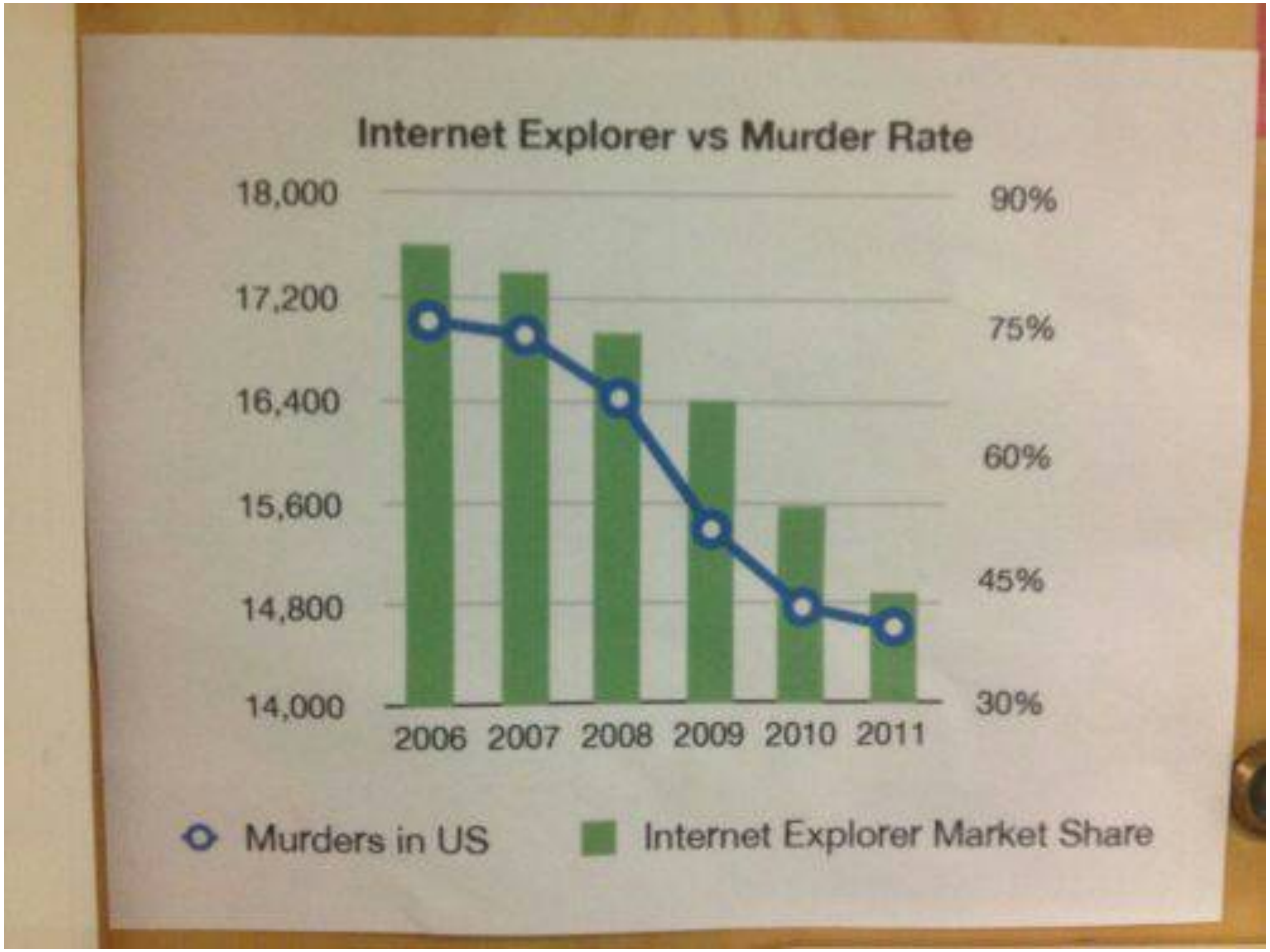
What about Big Data?

"Classical statistics was fashioned for small problems, a few hundred data points at most, a few parameters."

"The bottom line is that we have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of classical statistics."

Bradley Efron,
Bayesians, Frequentists, and Scientists

http://www-stat.stanford.edu/~ckirby/brad/papers/2005BayesFreqSci.pdf


Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006
2007
2008
2009
2010
2011
Murders in US
Internet Explorer Market Share

Positive Correlations

- Number of police officers and number of crimes (Glass & Hopkins, 1996)
- Amount of ice cream sold and deaths by drownings (Moore, 1993)
- Stork sightings and population increase

(Box, Hunter, Hunter, 1978)

The “curse” of Big Data?

“…the curse of big data is the fact that when you search for patterns in very, very large data sets with billions or trillions of data points and thousands of metrics, you are bound to identify coincidences that have no predictive power.”

Vincent Granville

http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data

Vincent Granville's Example

- Consider stock prices for 500 companies over a 1-month period
- Check for correlations in all pairs

Aside: Cross-correlation of timeseries

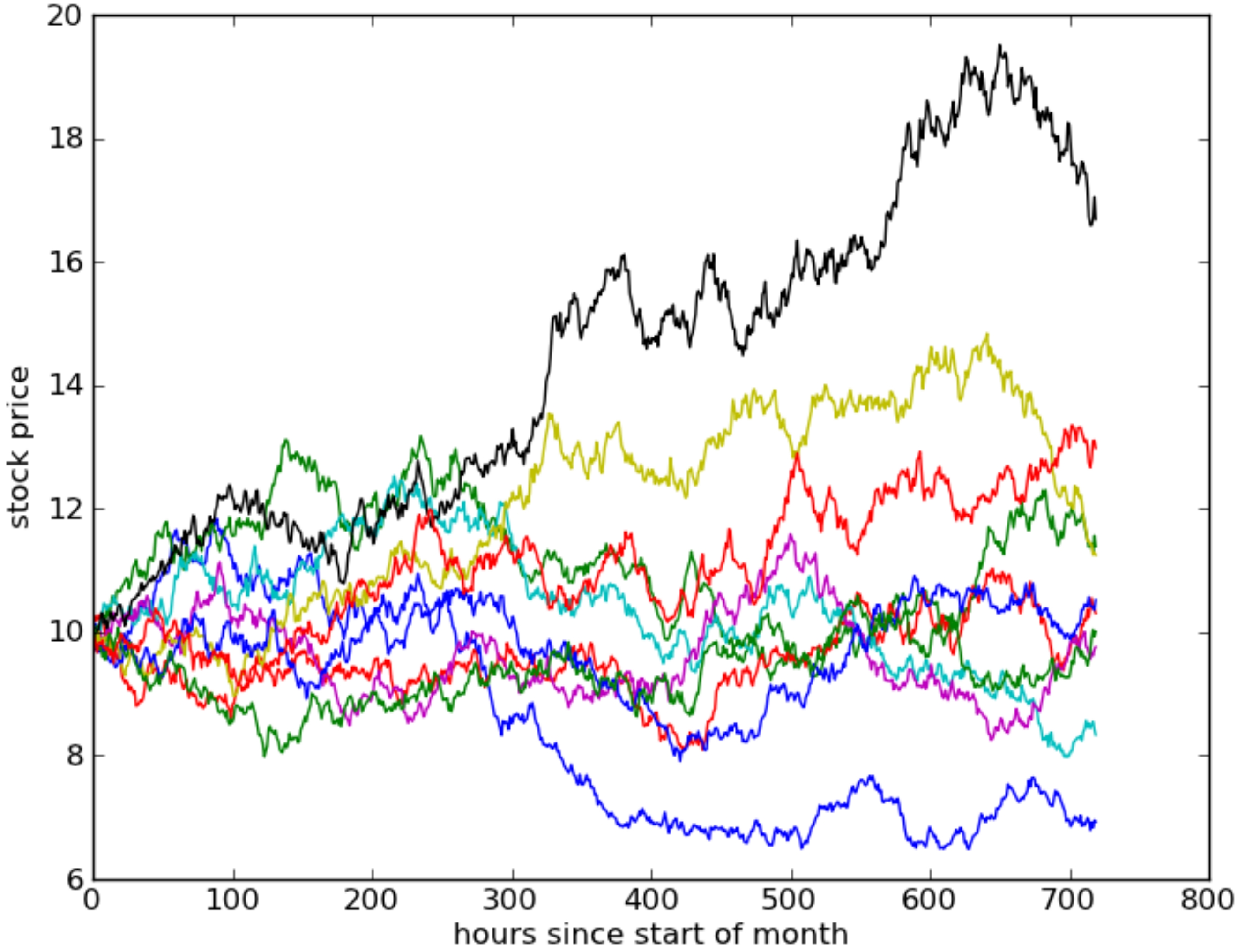
What happens when both are high at time i? Both low? One high, one low?

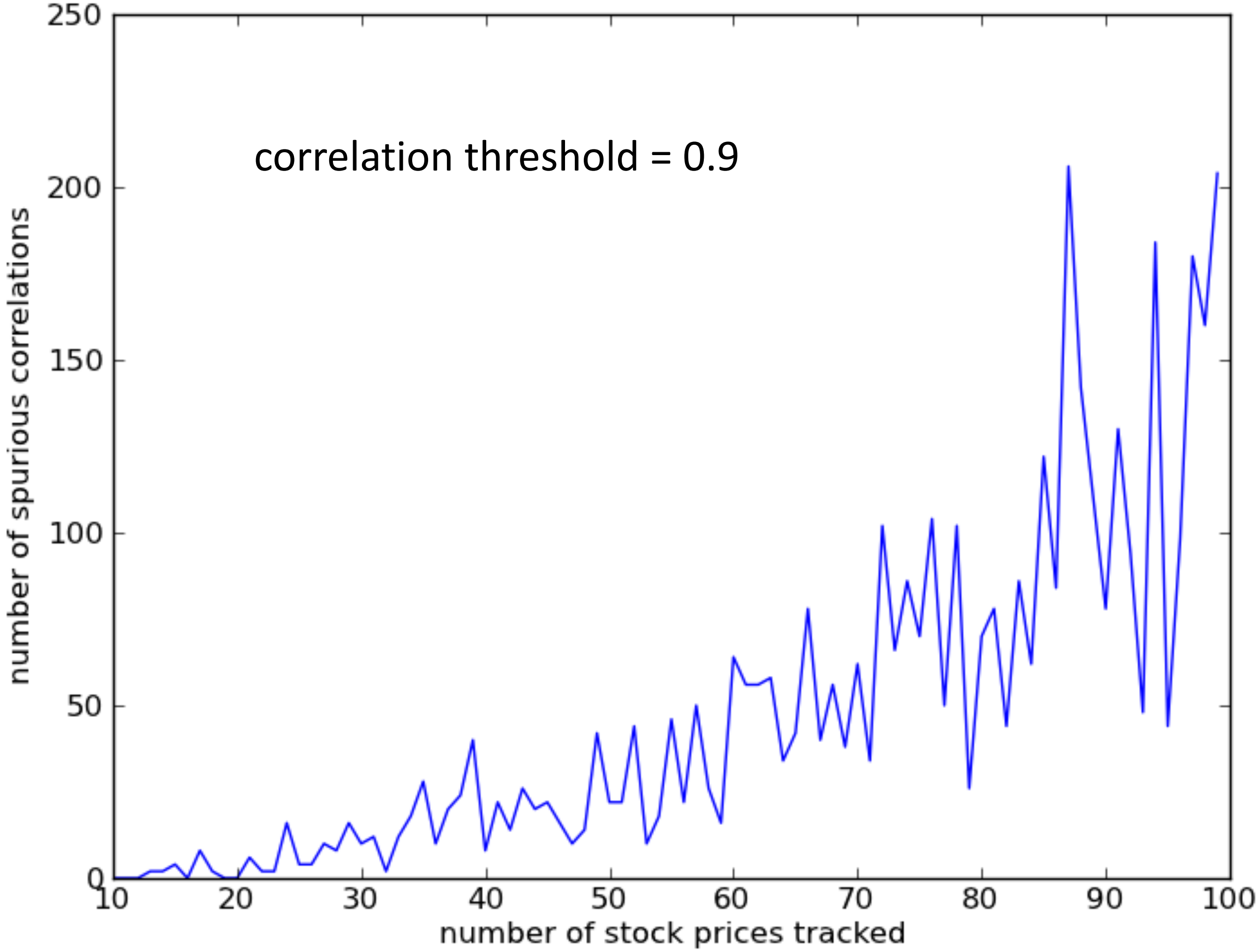
$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_{i}^{N}(x_i - u_x)(y_i - u_y)$$

$$\mathrm{corr}(x, y) = \frac{cov(x, y)}{\sqrt{\sum_i (x_i - u_x)^2}\sqrt{\sum_i (y_i - u_y)^2}}$$

standard deviation

random walk
each step is normally distributed @ 1% of current price




correlation threshold = 0.9
number of spurious correlations
0
50
100
150
200
250
number of stock prices tracked
10
20
30
40
50
60
70
80
90
100

Is Big Data different?

- Big P vs. Big N
  - P = number of variables (columns)
  - N = number of records
- Marginal cost of increasing N is essentially zero!
- But while >N decreases variance, it amplifies bias
  - Ex: You log all clicks to your website to model user behavior, but this only samples current users, not the users you want to attract.
  - Ex: Using mobile data to infer buying behavior
- Beware multiple hypothesis tests
  - "Green jelly beans cause acne"
- Taleb's "Black Swan" events
  - The turkey's model of human behavior

Jonah Lehrer, 2010, The New Yorker
The Truth Wears off

John Davis, University of Illinois
"Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades."

Anders Pape Møller, 1991
"female barn swallows were far more likely to mate with male birds that had long, symmetrical feathers"
"Between 1992 and 1997, the average effect size shrank by eighty per cent."

Jonathan Schooler, 1990
"subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it."
The effect became increasingly difficult to measure.

Joseph Rhine, 1930s, coiner of the term extrasensory perception
Tested individuals with card-guessing experiments. A few students achieved multiple low-probability streaks. But there was a "decline effect" – their performance became worse over time.

http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Publication Bias

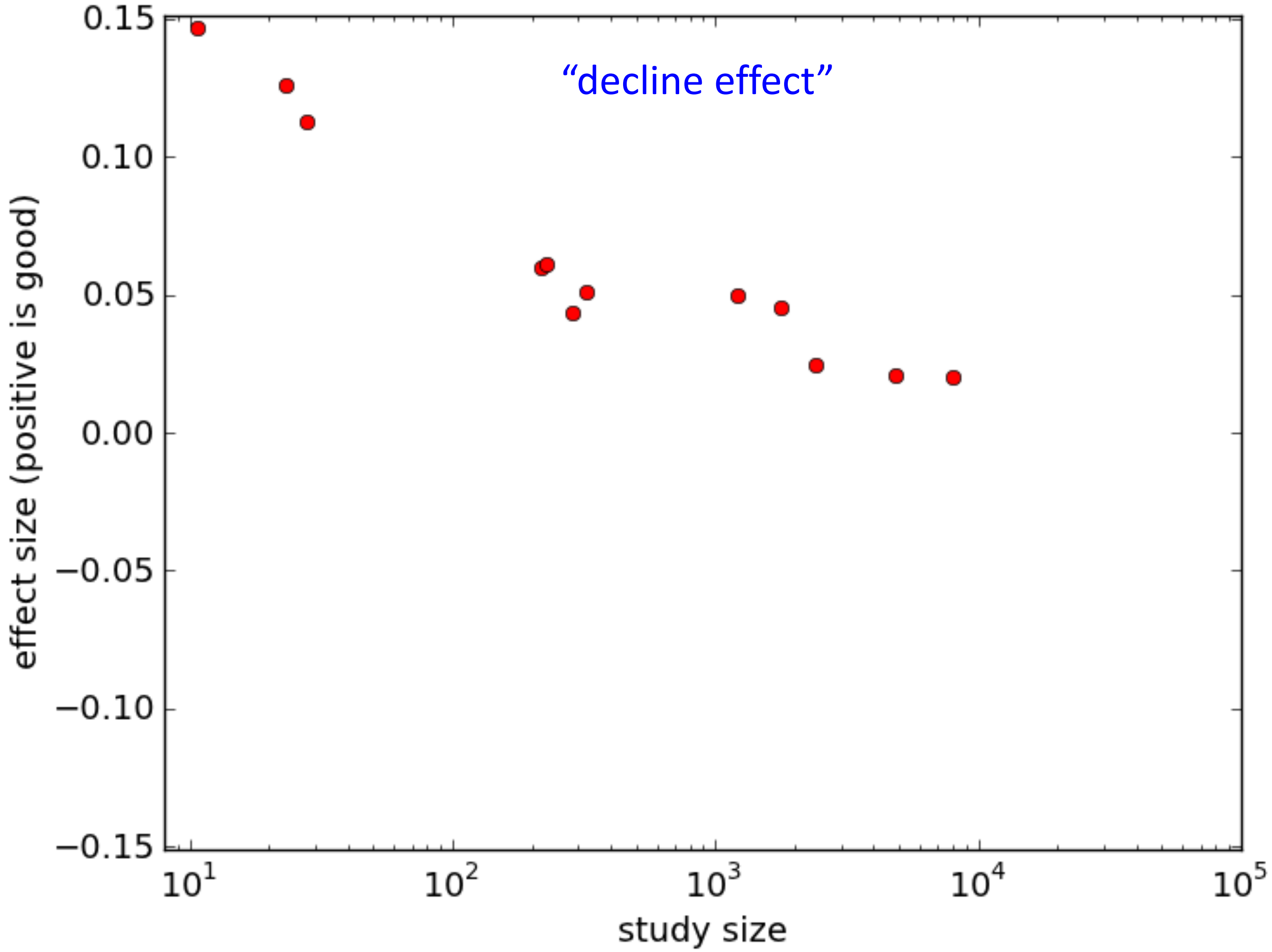
"In the last few years, several meta-analyses have reappraised the efficacy and safety of antidepressants and concluded that the therapeutic value of these drugs may have been significantly overestimated."

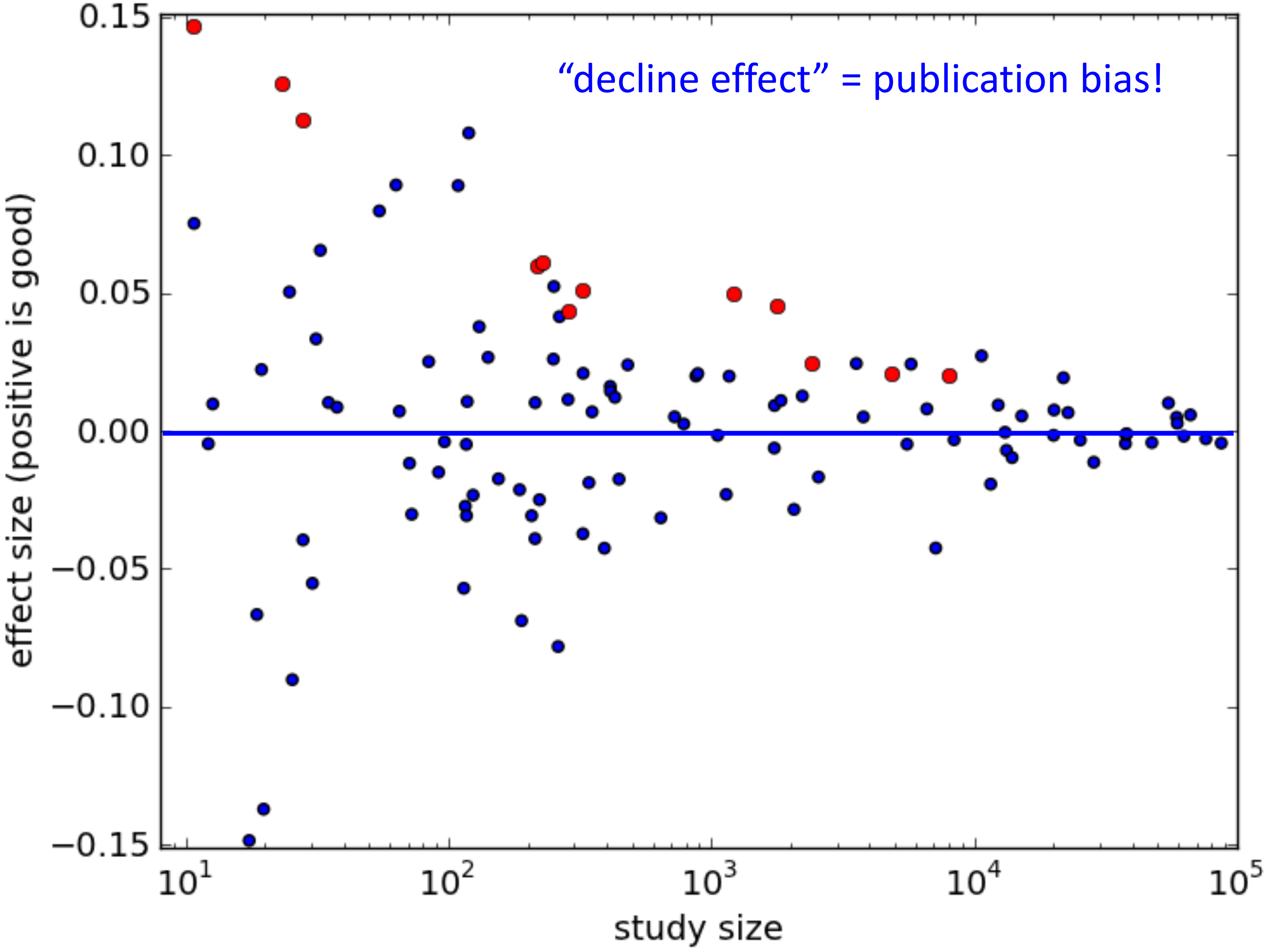
"Although publication bias has been documented in the literature for decades and its origins and consequences debated extensively, there is evidence suggesting that this bias is increasing."

"A case in point is the field of biomedical research in autism spectrum disorder (ASD), which suggests that **in some areas negative results are completely absent**" *(emphasis mine)*

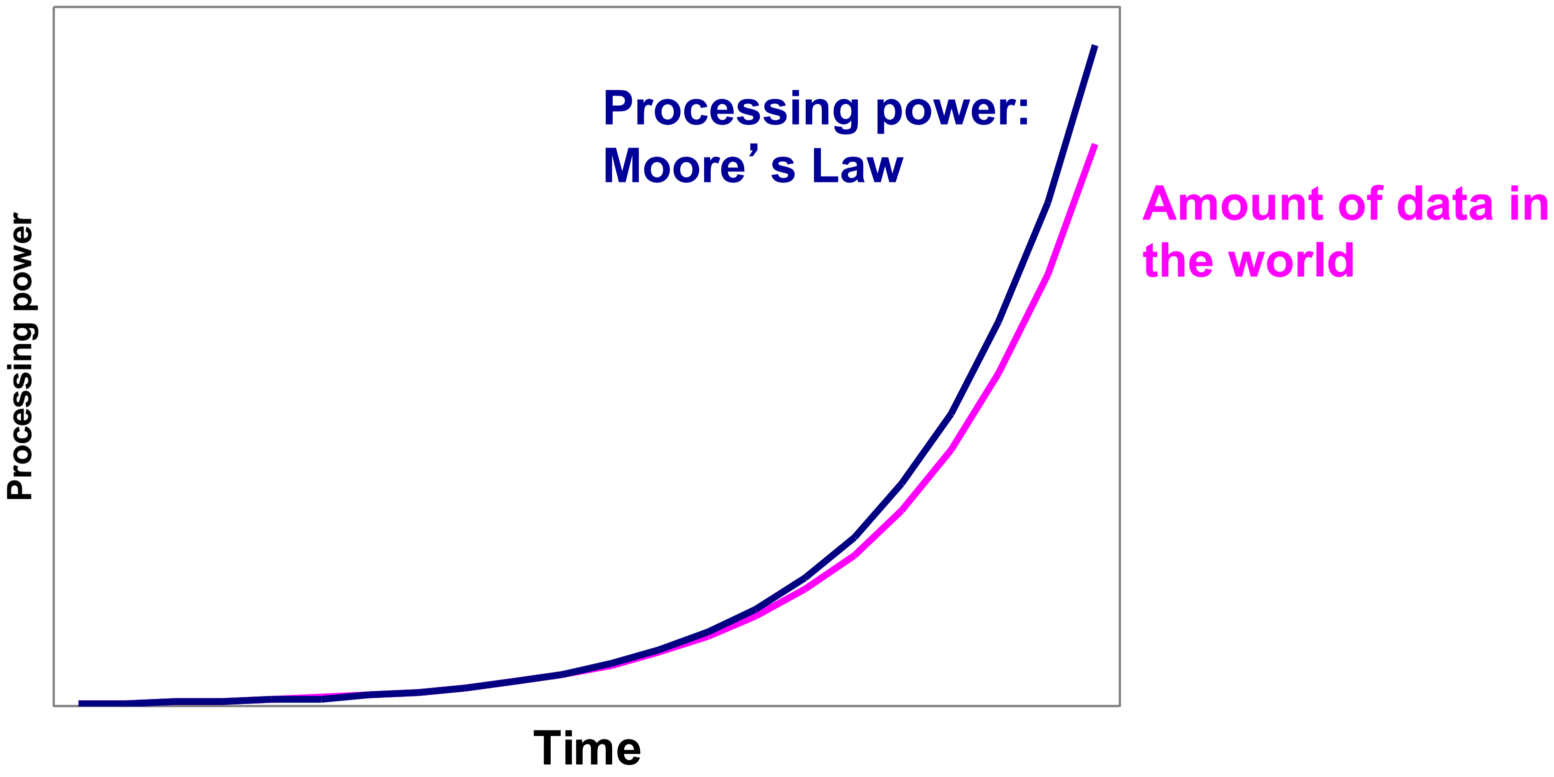
"… a highly significant correlation (R^2= 0.13, p < 0.001) between impact factor and overestimation of effect sizes has been reported."

Publication bias: What are the challenges and can they be overcome?
Ridha Joober, Norbert Schmitz, Lawrence Annable, and Patricia Boksa
J Psychiatry Neurosci. 2012 May; 37(3): 149–152. doi: 10.1503/jpn.120065

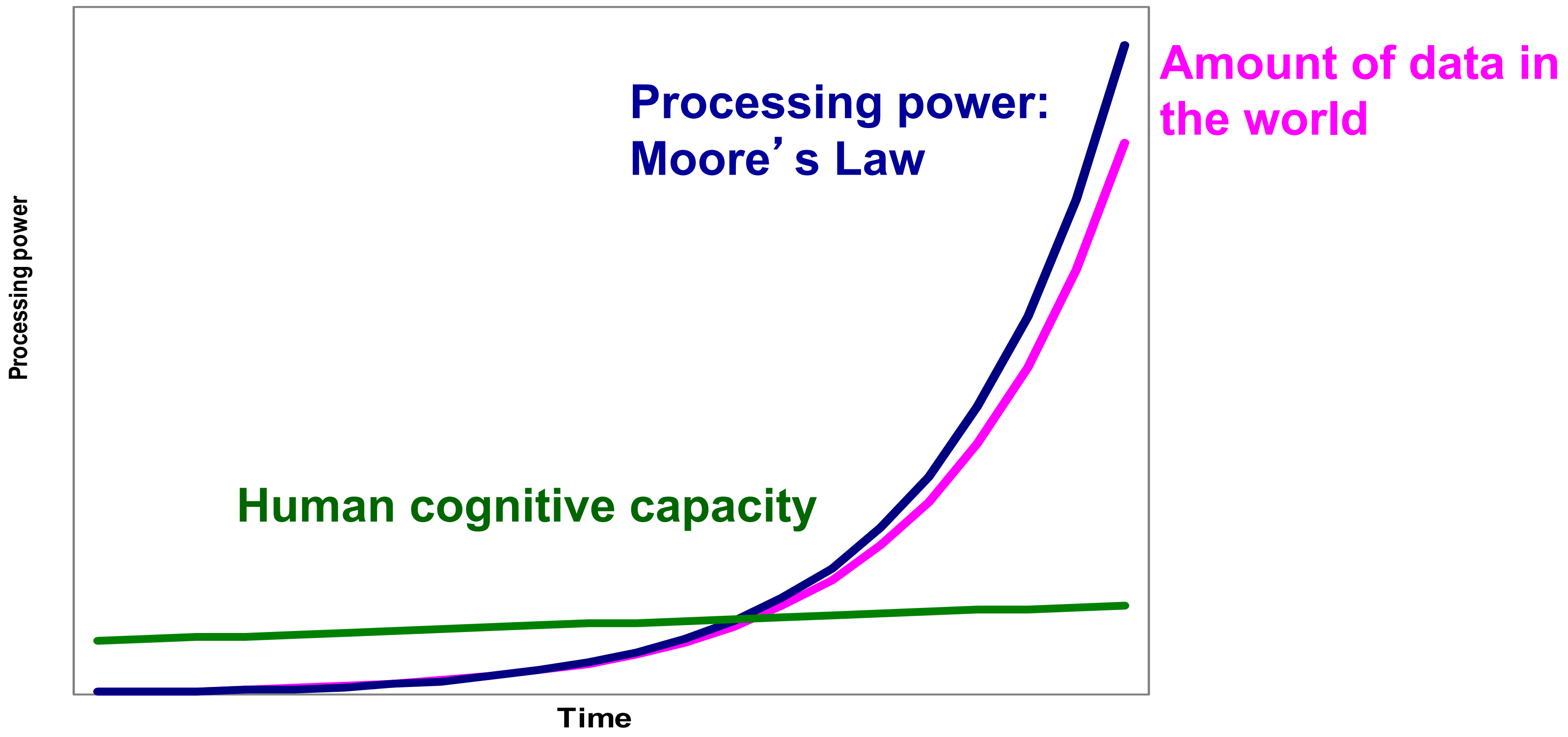

"decline effect"
effect size (positive is good)
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
study size
10^1
10^2
10^3
10^4
10^5


“decline effect” = publication bias!
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
effect size (positive is good)
10^1
10^2
10^3
10^4
10^5
study size

What is the rate-limiting step in data understanding?



What is the rate-limiting step in data understanding?



Idea adapted from "Less is More" by Bill Buxton (2001)

slide src: Cecilia Aragon, UW HCDE

Key Takeaways:

Data science is different than statistics: large, noisy, heterogeneous datasets instead of small, carefully acquired datasets

With large N and large P, there's a risk of spurious correlations that appear convincing

COURSE LOGISTICS

Topics by Week

- Data Science in the Wild; Cloud Computing
- Principles of Big Data Processing
- Abstractions for Big Data Programming
- Systems for Big Data Management
- Algorithms for Big Data Analytics
- Graph and Network Analysis
- Text-as-Data
- Deep Learning (maybe)
- Privacy and Ethics
- Algorithmic Bias

Grades

Group Project	35%
Assignments	30%
Quizzes	20%
In-class	15%

Assignments

- Twitter (Python, data wrangling)
- MapReduce and SQL algorithms
- (likely) Machine Learning in Spark
- NLP & Neural networks
- Algorithmic Bias (some SQL)

Readings

- To be read by the week in which they are listed (first week is an exception)
- Some readings will emphasize research – the field is moving fast!

Quizzes

- Approximately one per week
- Short, in class questions via canvas
- Based on readings and prior lectures

Course Project

- Four milestones
  - – Project Proposal
  - – Project Plan
  - – Project Update
  - – Final Presentation
- Milestones interleaved with other assignments