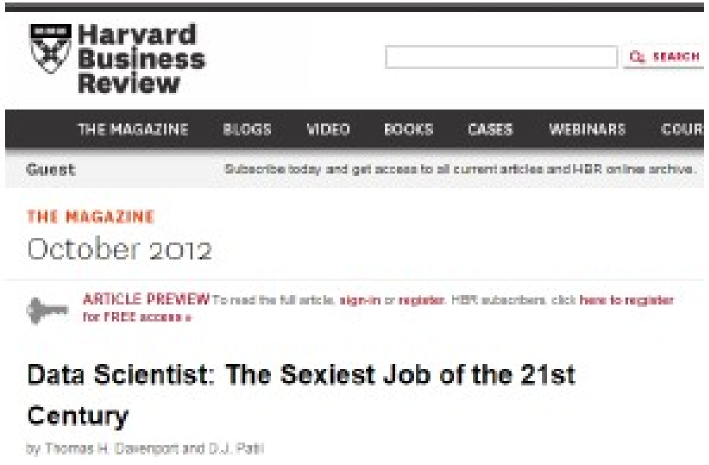
CSE303

Lecture 1: Introduction to Data Science

DATA SCIENTISTS ARE IN HIGH DEMAND



Subscribe today and get access to all current articles and HBR online archive.

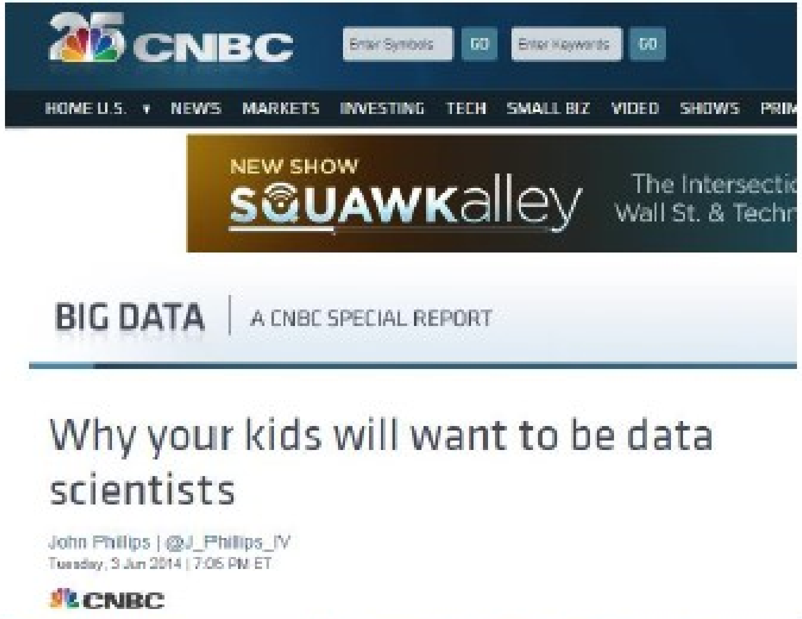
THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »

Data Scientist: The Sexiest Job of the 21st Century

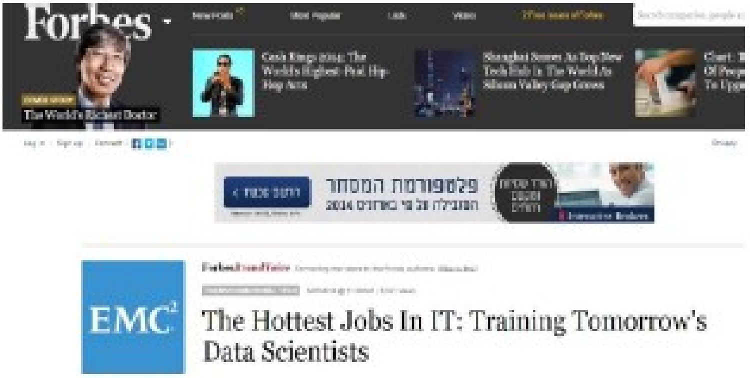
by Thomas H. Davenport and D.J. Patil



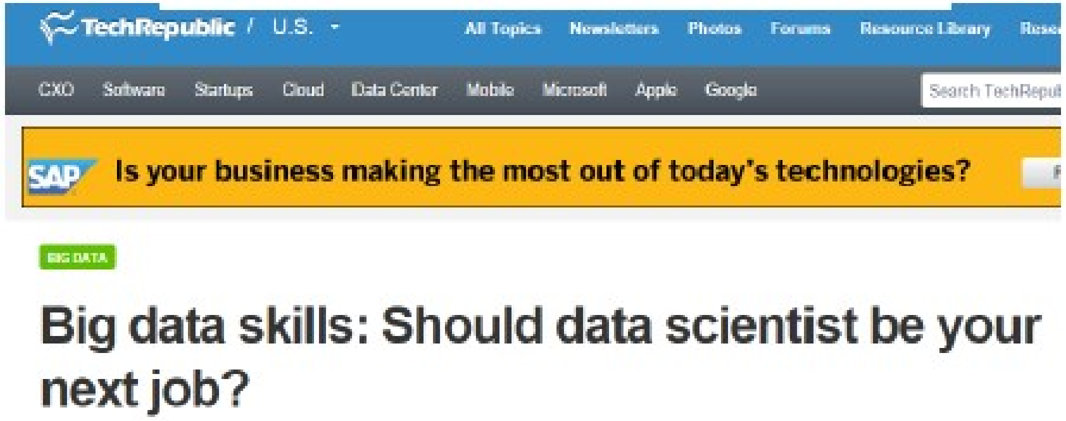
BIG DATA | A CNBC SPECIAL REPORT

Why your kids will want to be data scientists

John Phillips | @J_Phillips_IV
Tuesday, 3 Jun 2014 | 7:06 PM ET


Forbes
The World's Richest Doctor
Cash Kings 2014: The World's Highest-Paid Hip-Hop Acts
Shanghai Soars As Top New Tech Hub In The World As Silicon Valley Gap Grows
EMC²

The Hottest Jobs In IT: Training Tomorrow's Data Scientists

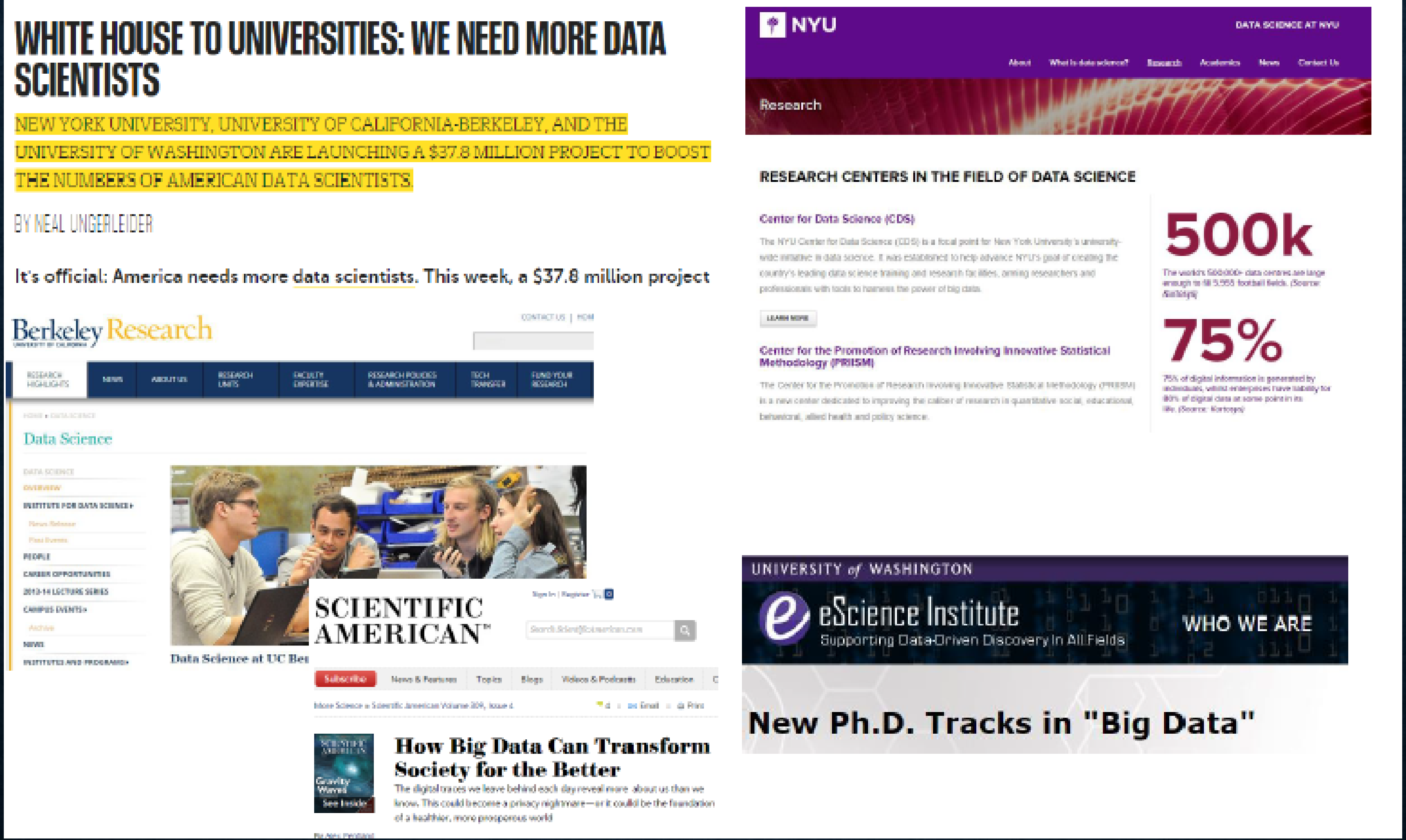


Is your business making the most out of today's technologies?

BIG DATA

Big data skills: Should data scientist be your next job?

ALSO IN ACADEMIA



WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A $37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS.

BY NEAL UNGERLEIDER

It's official: America needs more data scientists. This week, a $37.8 million project

Berkeley Research

Data Science

Data Science at UC Ber

SCIENTIFIC AMERICAN™

How Big Data Can Transform Society for the Better

The digital traces we leave behind each day reveal more about us than we know. This could become a privacy nightmare—or it could be the foundation of a healthier, more prosperous world

By Alex Pentland

NYU

Research

RESEARCH CENTERS IN THE FIELD OF DATA SCIENCE

Center for Data Science (CDS)

The NYU Center for Data Science (CDS) is a focal point for New York University's university-wide initiative in data science. It was established to help advance NYU's goal of creating the country's leading data science training and research facilities, arming researchers and professionals with tools to harness the power of big data.

Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM)

The Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM) is a new center dedicated to improving the caliber of research in quantitative social, educational, behavioral, allied health and policy science.

500k

The world's 500,000+ data centres are large enough to fill 5,955 football fields. (Source: Softlayer)

75%

75% of digital information is generated by individuals, whilst enterprises have liability for 80% of digital data at some point in its life. (Source: Kontoopo)

UNIVERSITY of WASHINGTON

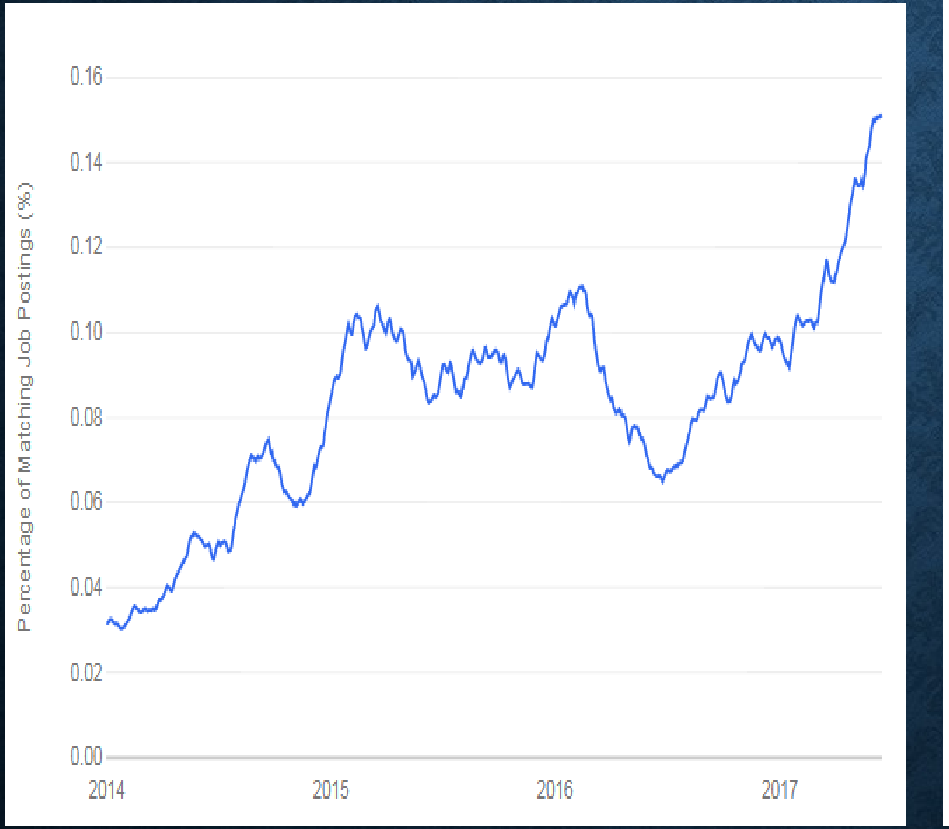
eScience Institute

Supporting Data-Driven Discovery In All Fields

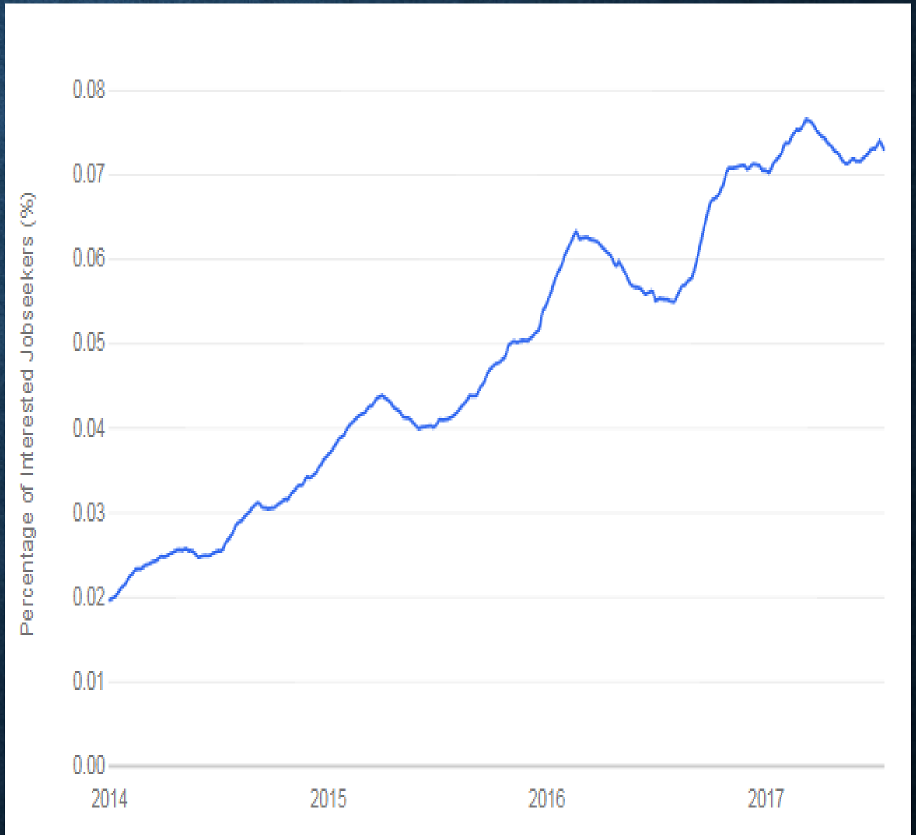
New Ph.D. Tracks in "Big Data"

DATA SCIENTIST JOB TREND

Job postings

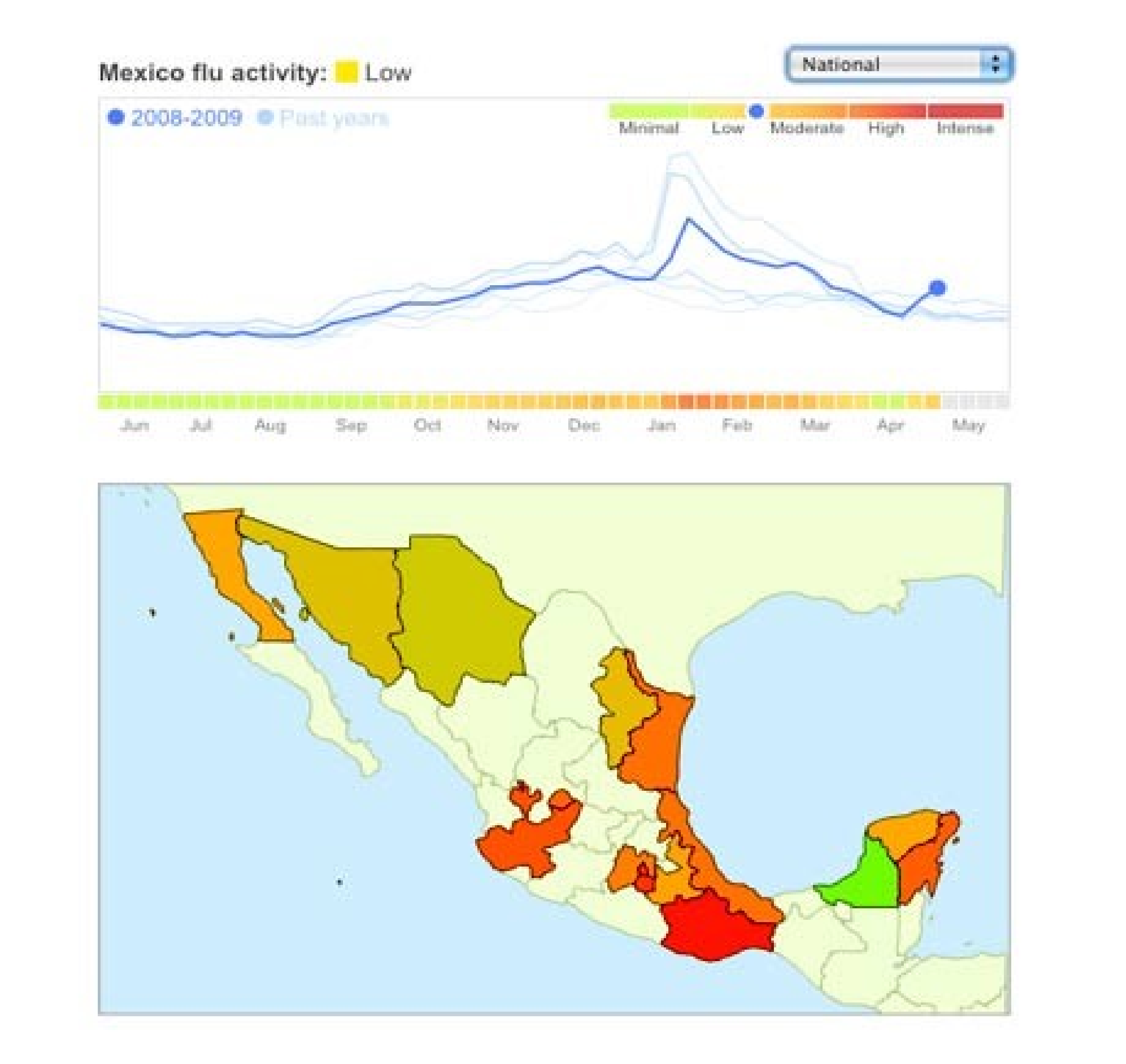


Jobseeker interest



Source: indeed.com

DATA SCIENCE: WHY ALL THE EXCITEMENT?

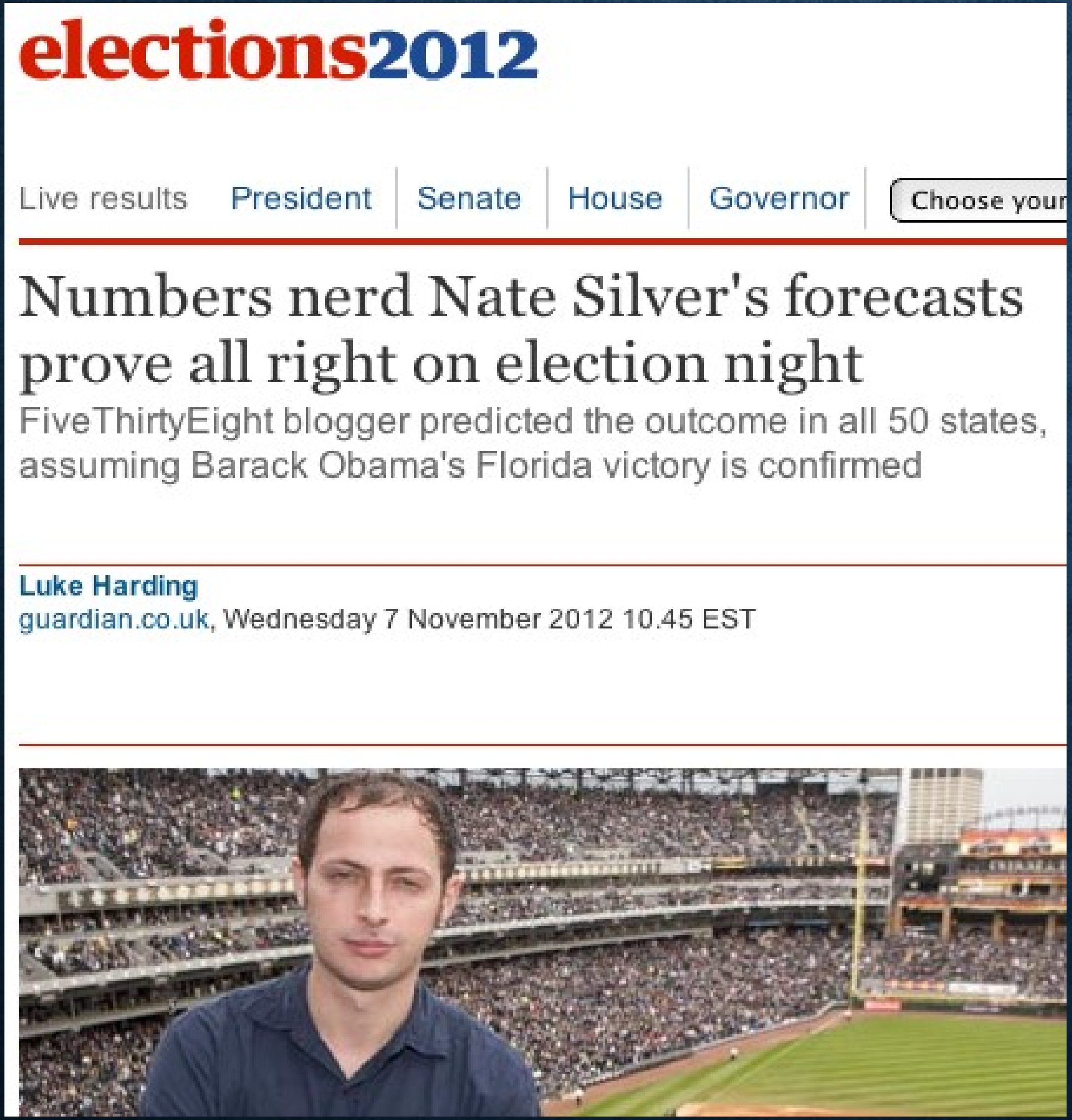


e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

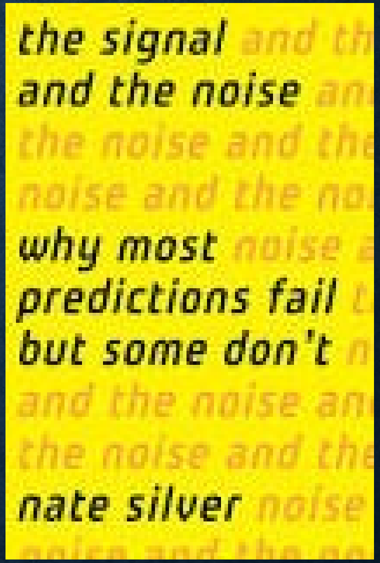
DATA SCIENCE: WHY ALL THE EXCITEMENT?



Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

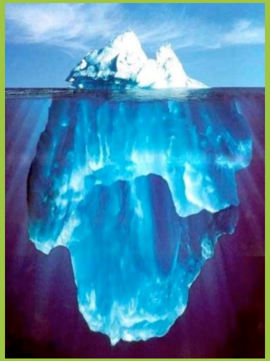
Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST



the signal
and the noise
why most
predictions fail
but some don't
nate silver

“BIG DATA” SOURCES

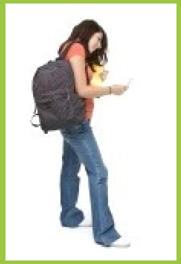
It’s All Happening On-line



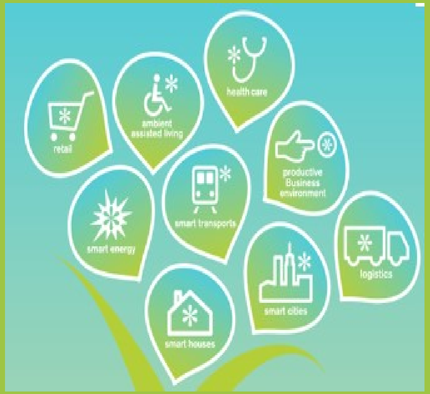
Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

User Generated (Web & Mobile)

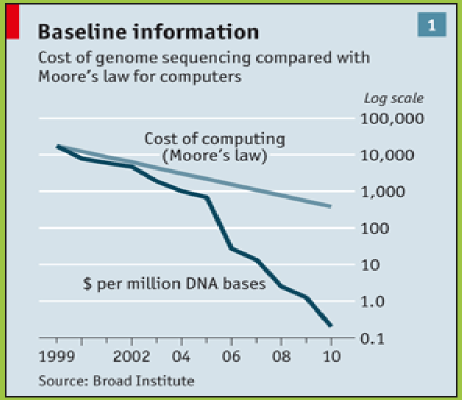
…
..



Internet of Things / M2M



Health/Scientific Computing



Baseline information

Cost of genome sequencing compared with Moore’s law for computers

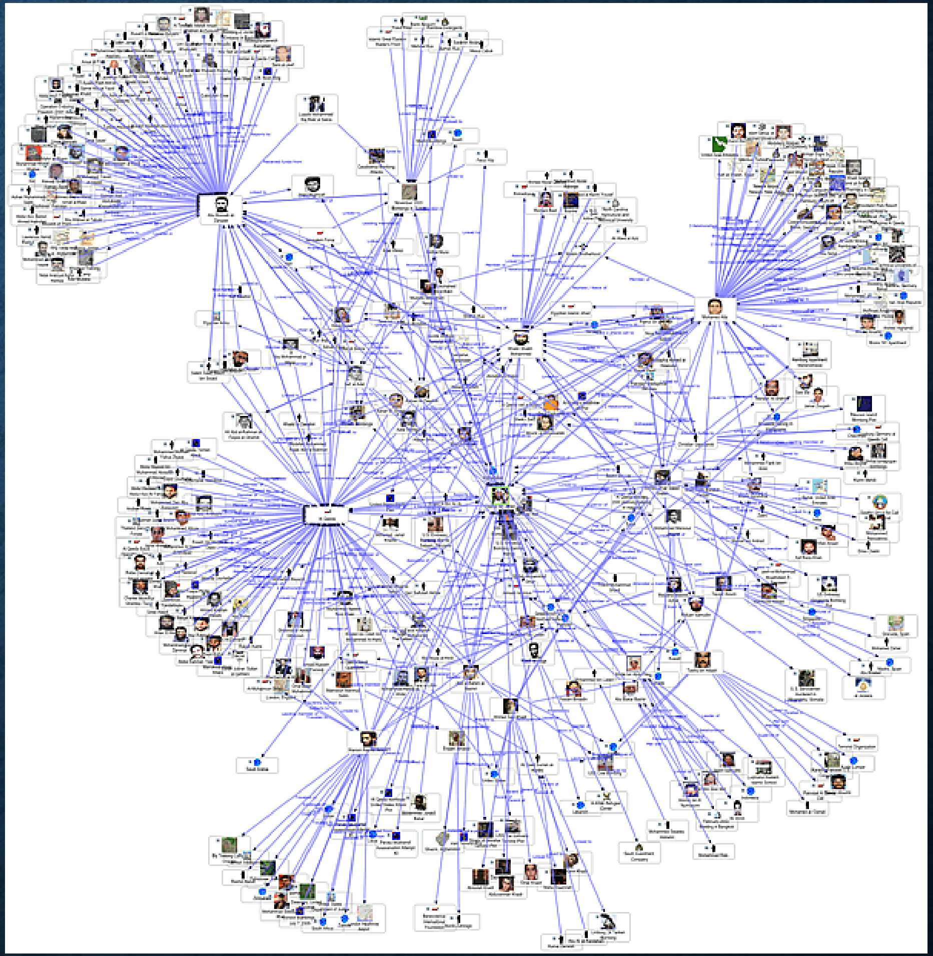
Source: Broad Institute

GRAPH DATA

Lots of interesting data has a graph structure:

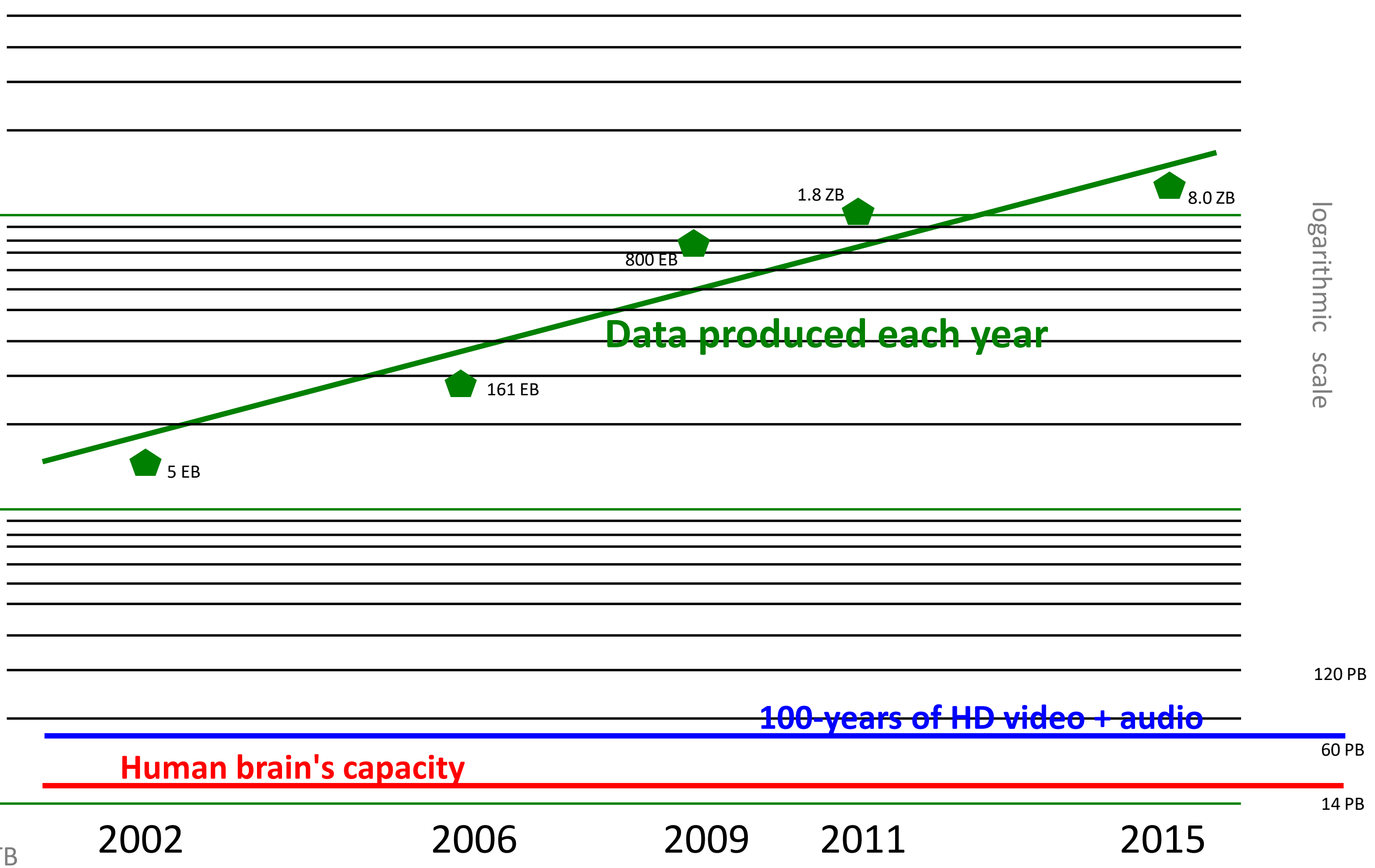
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

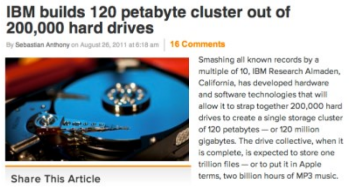
Some of these graphs can get quite large (e.g., Facebook* user graph)



Data, data everywhere...

There's certainly a lot of it!





IBM builds 120 petabyte cluster out of 200,000 hard drives

By Sebastian Anthony on August 26, 2011 at 6:18 am | 16 Comments

Smashing all known records by a multiple of 10, IBM Research Almaden, California, has developed hardware and software technologies that will allow it to strap together 200,000 hard drives to create a single storage cluster of 120 petabytes — or 120 million gigabytes. The drive collective, when it is complete, is expected to store one trillion files — or to put it in Apple terms, two billion hours of MP3 music.

Share This Article

1 Petabyte == 1000 TB
1 TB = 1000 GB

References

(2015) 8 ZB: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

(2011) 1.8 ZB: http://www.emc.com/leadership/programs/digital-universe.htm

(2009) 800 EB: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf

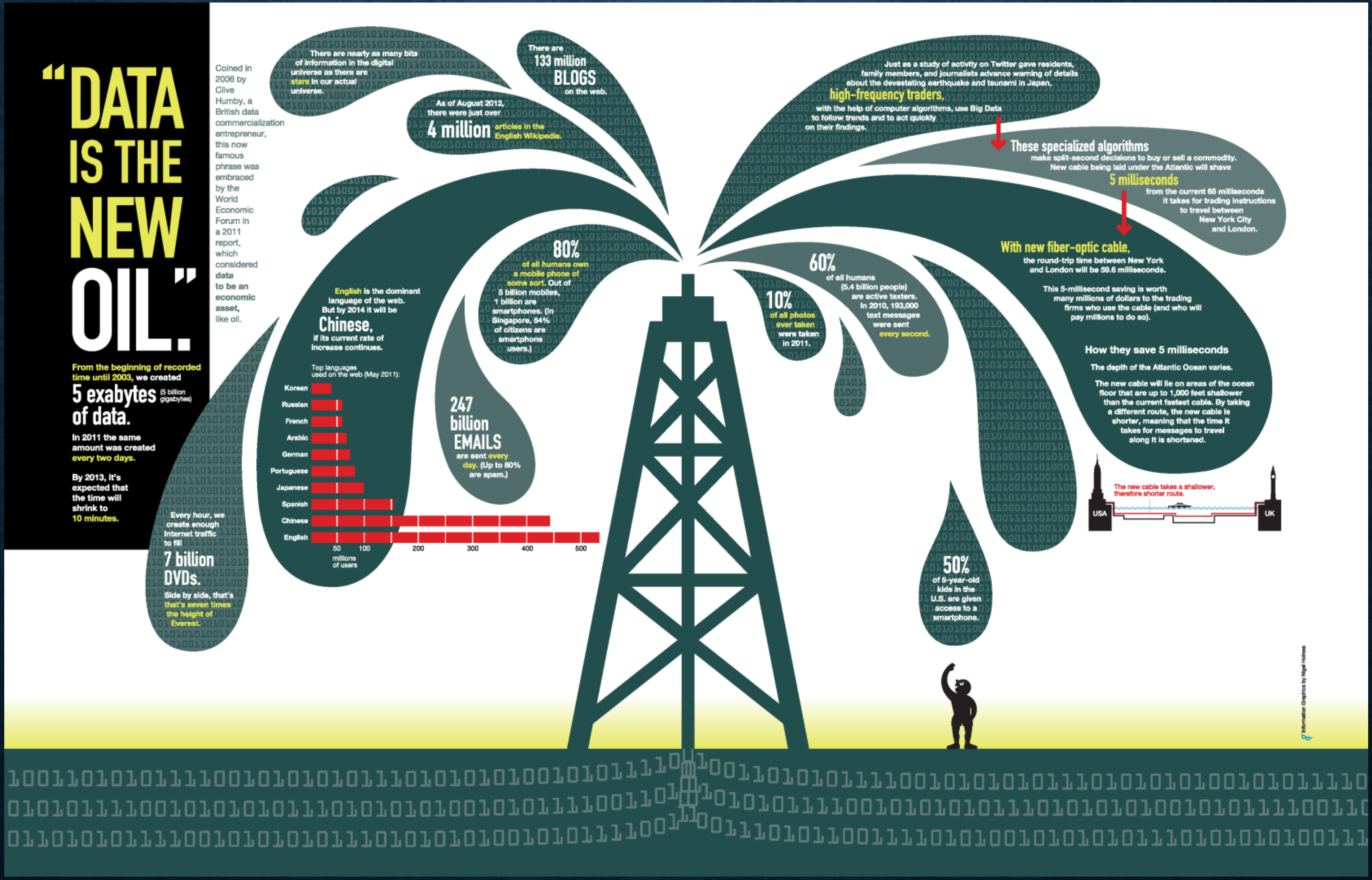
(2006) 161 EB: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

(2002) 5 EB: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

(life in video) 60 PB: in 4320p resolution, extrapolated from 16MB for 1:21 of 640x480 video (w/sound) – almost certainly a gross overestimate, as sleep can be compressed significantly!

(brain) 14 PB: http://www.quora.com/Neuroscience-1/How-much-data-can-the-human-brain-store

"DATA IS THE NEW OIL" – WORLD ECONOMIC FORUM 2011



"DATA IS THE NEW OIL."

From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2012, there were just over 4 million articles in the English Wikipedia.

There are 133 million BLOGS on the web.

English is the dominant language of the web. But by 2014 it will be Chinese, if its current rate of increase continues.

Top languages used on the web (May 2011):

Every hour, we create enough Internet traffic to fill 7 billion DVDs. Side by side, that's that's seven times the height of Everest.

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

10% of all photos ever taken were taken in 2011.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave 5 milliseconds from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Information Graphics by Nigel Holmes

DATA SCIENCE – A DEFINITION

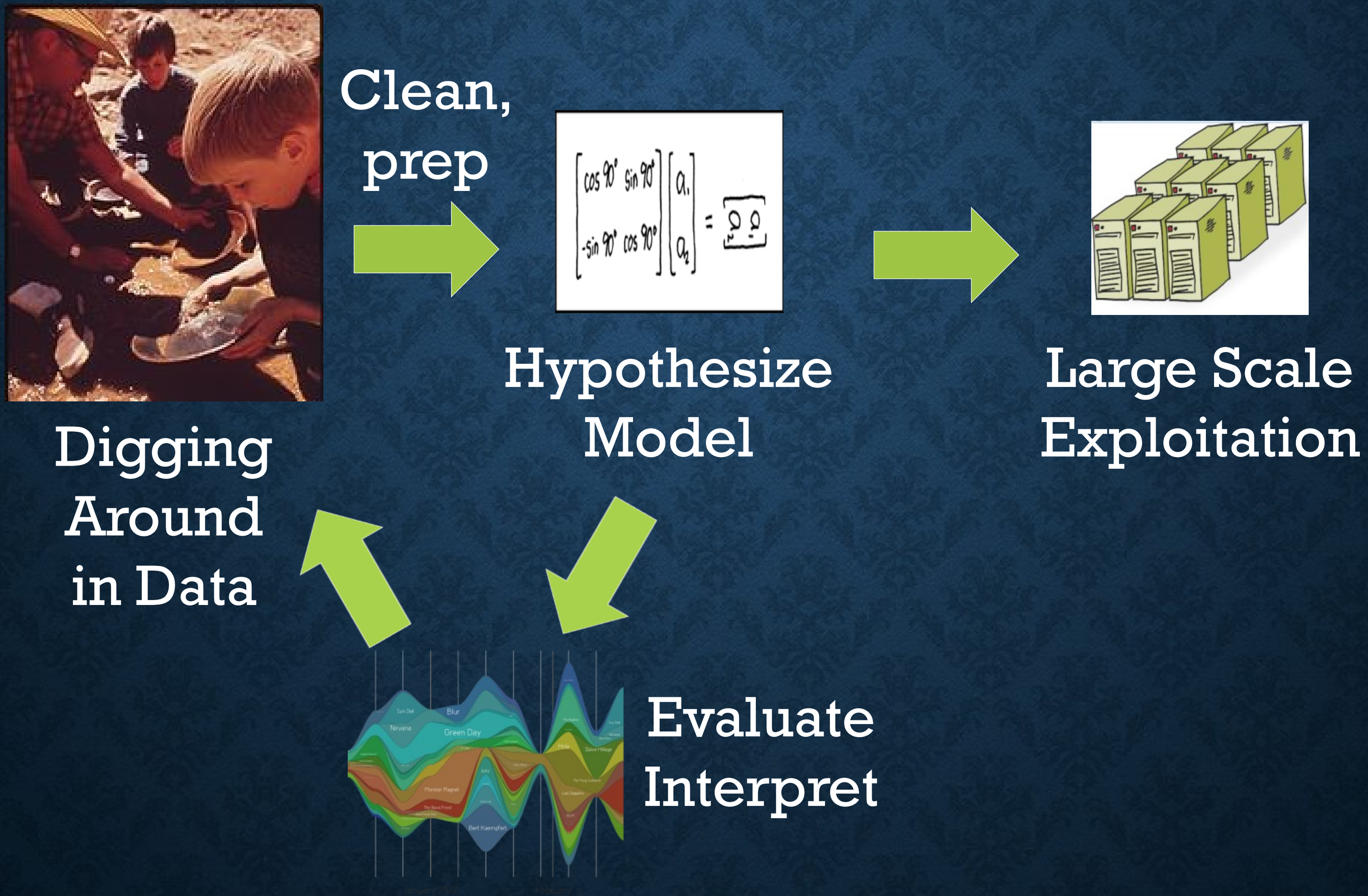
- **Data is a collection of facts.**
- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.
- Information is processed data.

HOW TO USE DATA?

- Data => exploratory analysis => knowledge models => product / decision making
- Data => predictive models => evaluate / interpret => product / decision making

- Exploratory analysis tells us what happened.
- Predictive analysis tells us what could happen next!

DATA SCIENTIST'S PRACTICE



Digging Around in Data

Clean, prep

Hypothesize Model

Large Scale Exploitation

Evaluate Interpret

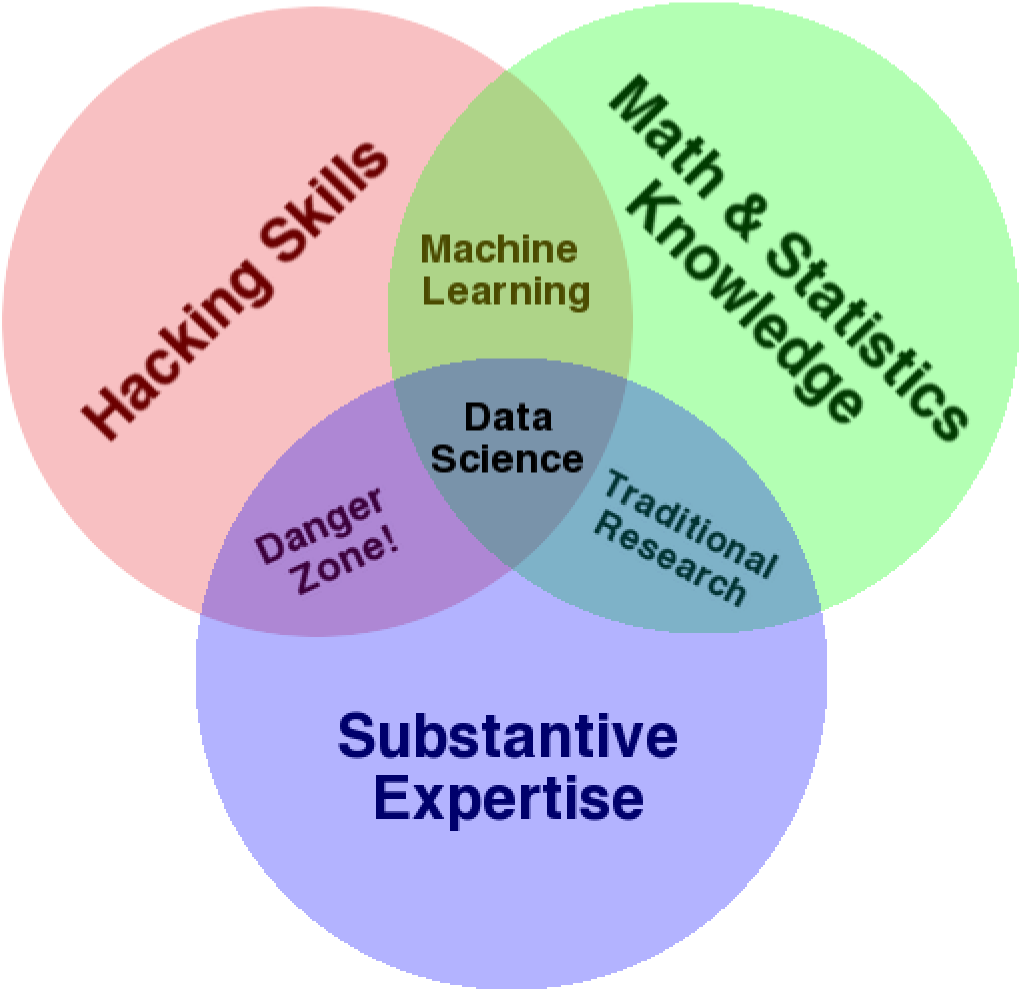
DATA SCIENCE APPLICATIONS

- Marketing: predict the characteristics of high life time value (LTV) customers, which can be used to support customer segmentation, identify upsell opportunities, and support other marketing initiatives
- Logistics: forecast how many of which things you need and where will we need them, which enables learn inventory and prevents out of stock situations
- Healthcare: analyze survival statistics for different patient attributes (age, blood type, gender, etc.) and treatments; predict risk of re-admittance based on patient attributes, medical history, etc.

MORE EXAMPLES

- Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)
- Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things
- Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery
- Software Log Data → Automatic Trouble Shooting (Splunk)
- Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine

DATA SCIENCE – ONE DEFINITION



Drew Conway

WHY “DANGER ZONE?”

Ronny Kohavi* keynote at KDD 2015

- People are incredibly clever at explaining “very surprising results”. Unfortunately most very surprising results are caused by data pipeline errors.

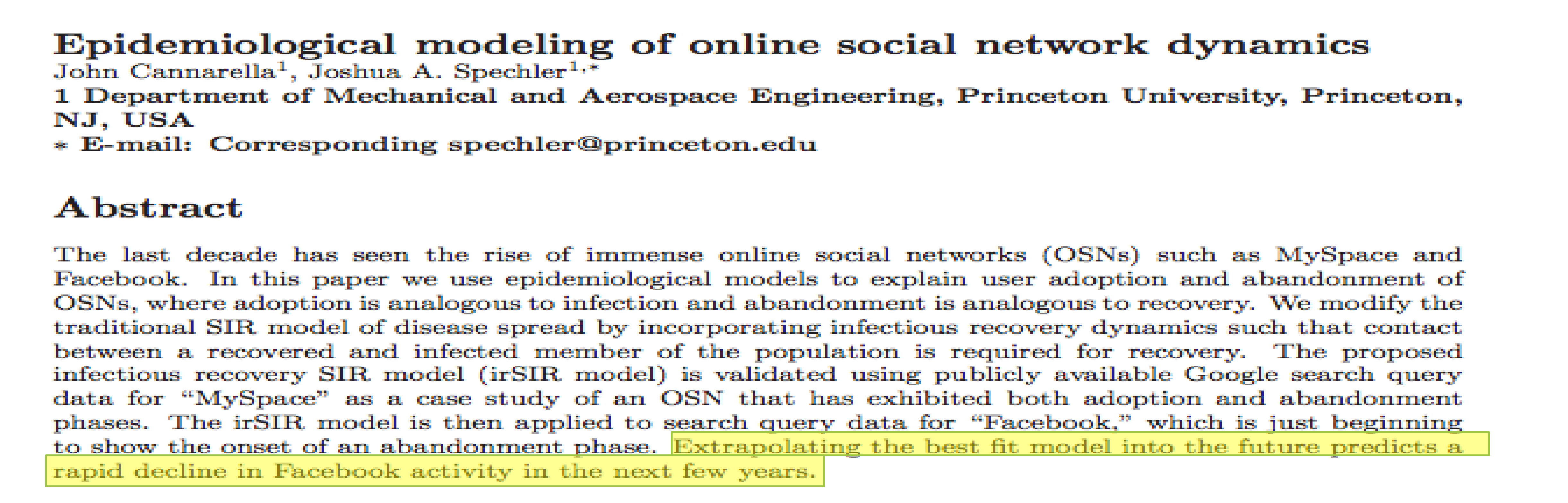
- Beware “HiPPOs” (Highest Paid-Person’s Opinion)

* General Manager for Microsoft’s Analysis and Experimentation Team

WHAT'S HARD ABOUT DATA SCIENCE

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

DATA SCIENCE CONCERNS



Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]
1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
* E-mail: Corresponding spechler@princeton.edu

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

DATA MAKES EVERYTHING CLEARER?

Searches for “MySpace”

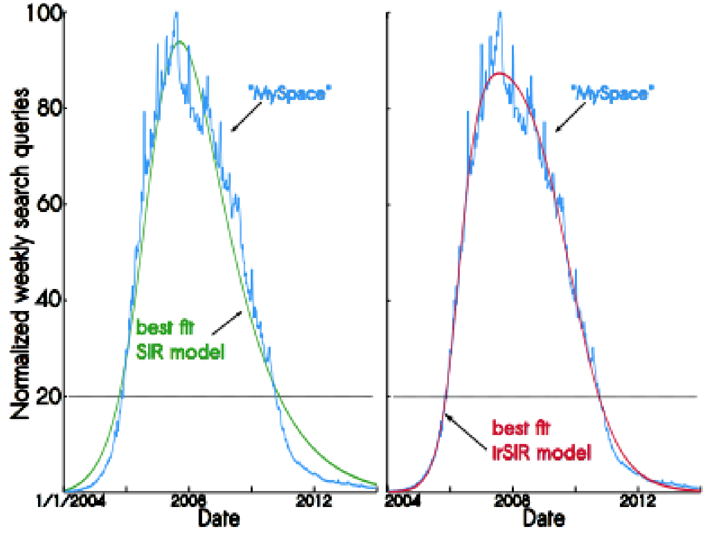
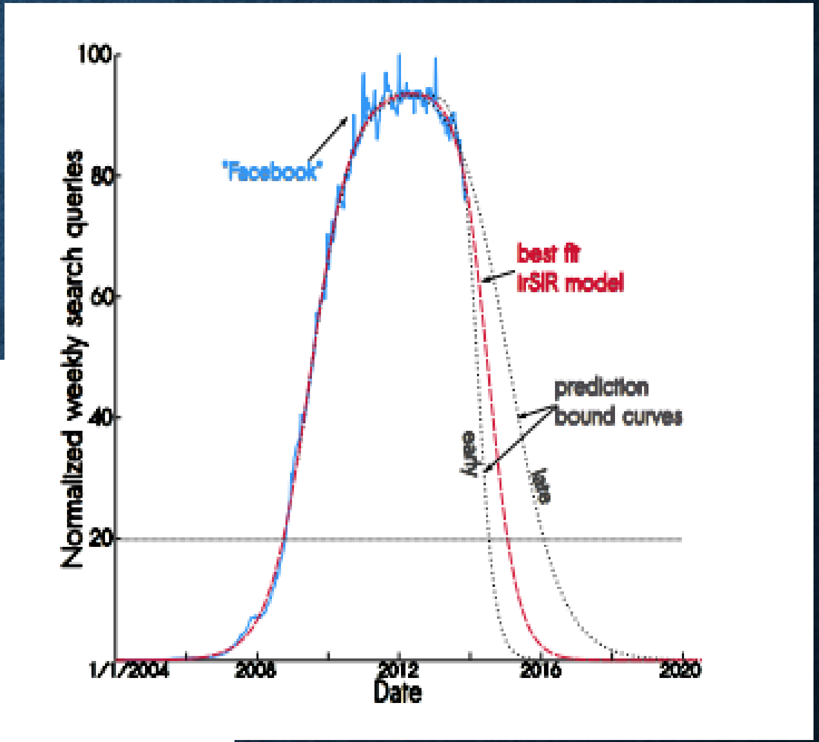
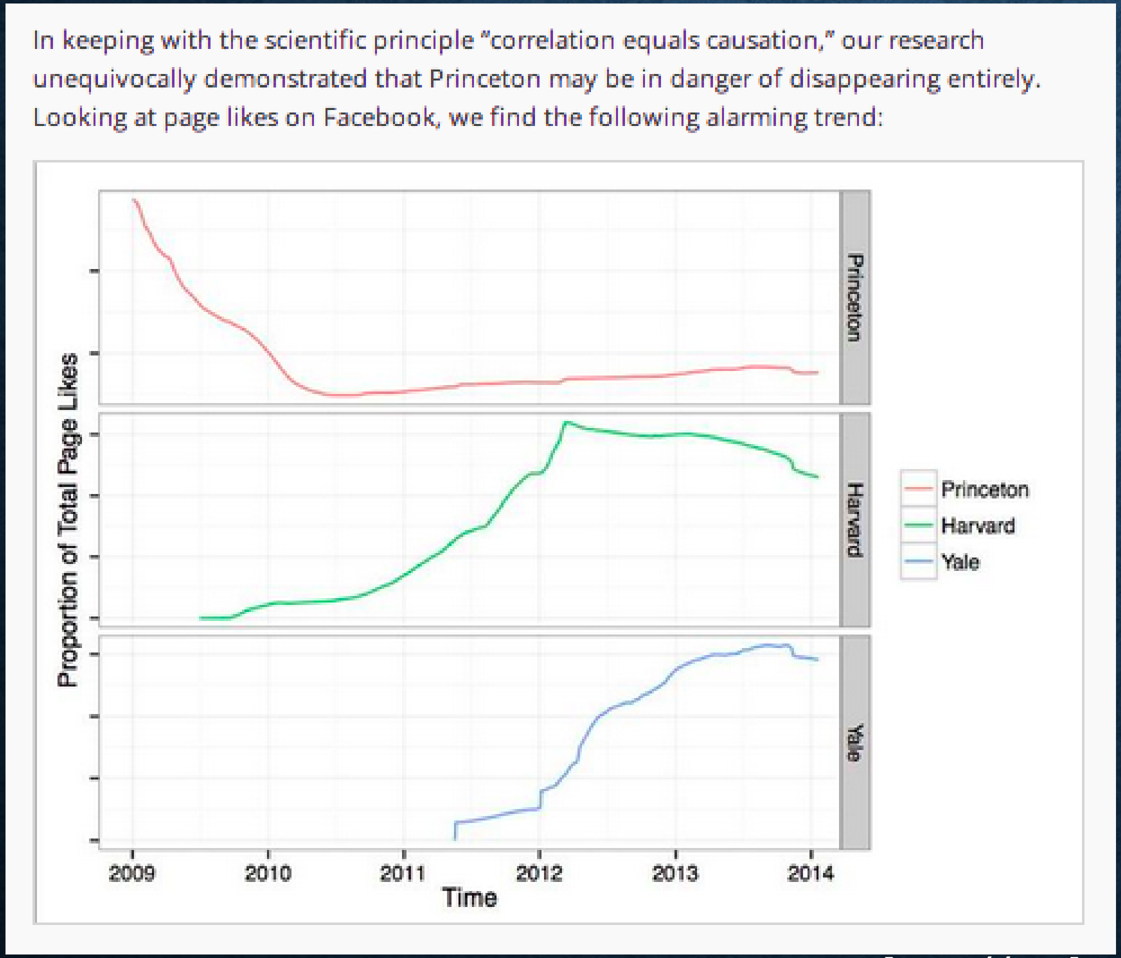


Figure 3: **Data for search query “Myspace” with best fit (a) SIR and (b) irSIR models overlaid.** The search query data are normalized such that the maximum data point corresponds to a value of 100.

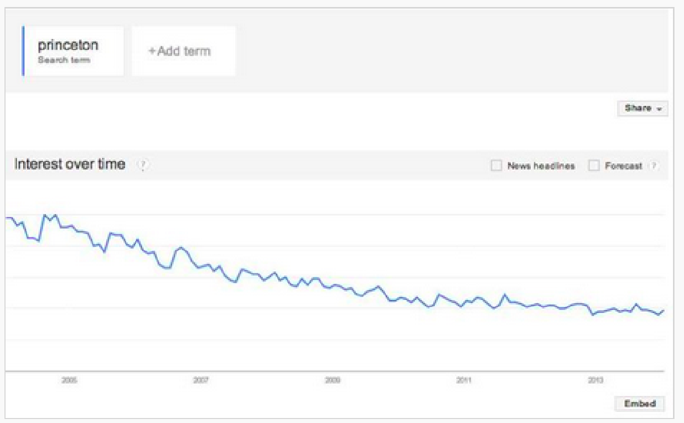


Searches for “Facebook”

DATA MAKES EVERYTHING CLEARER?



In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:



and based on Princeton search trends:

"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…

http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/

CSE303: STATISTICS FOR DATA SCIENCE

ABOUT THE COURSE

- A mixture of theory and practice
- Introductory, broad overview of subjects including
  - Statistics
  - Probability
  - Linear Algebra
  - Predictive models (Regression, Classification, Clustering)
  - Data Visualization
- Relevant Coding Skills
- Language choice: Python
  - Relatively easy to learn (for computer scientist) compared to R (more popular among statisticians)
  - Open source means easy access (as opposed to SAS or MATLAB)
  - https://www.upgrad.com/blog/data-science-programming-languages/
  - https://towardsdatascience.com/top-programming-languages-for-data-science-in-2020-3425d756e2a7

THANK YOU