CSE303

Lecture 1: Introduction to Data Science

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.

The Nobel Prize in Physics 2024

John Hopfield created an associative memory that can store and reconstruct images and other types of patterns in data. Geoffrey Hinton invented a method that can autonomously find properties in data

HOW TO USE DATA?

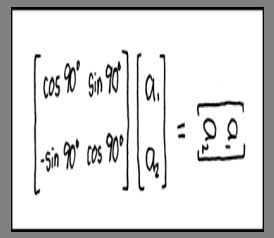
- Data => exploratory analysis => knowledge models => product / decision making
- Data => predictive models => evaluate / interpret => product / decision making

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

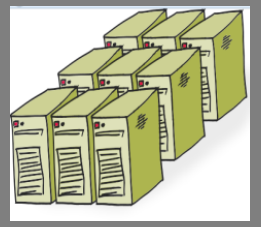
DATA SCIENTIST'S PRACTICE

Digging Around in Data

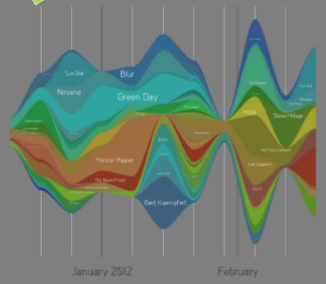
Clean, prep



Hypothesize Model



Large Scale Exploitation



Evaluate Interpret

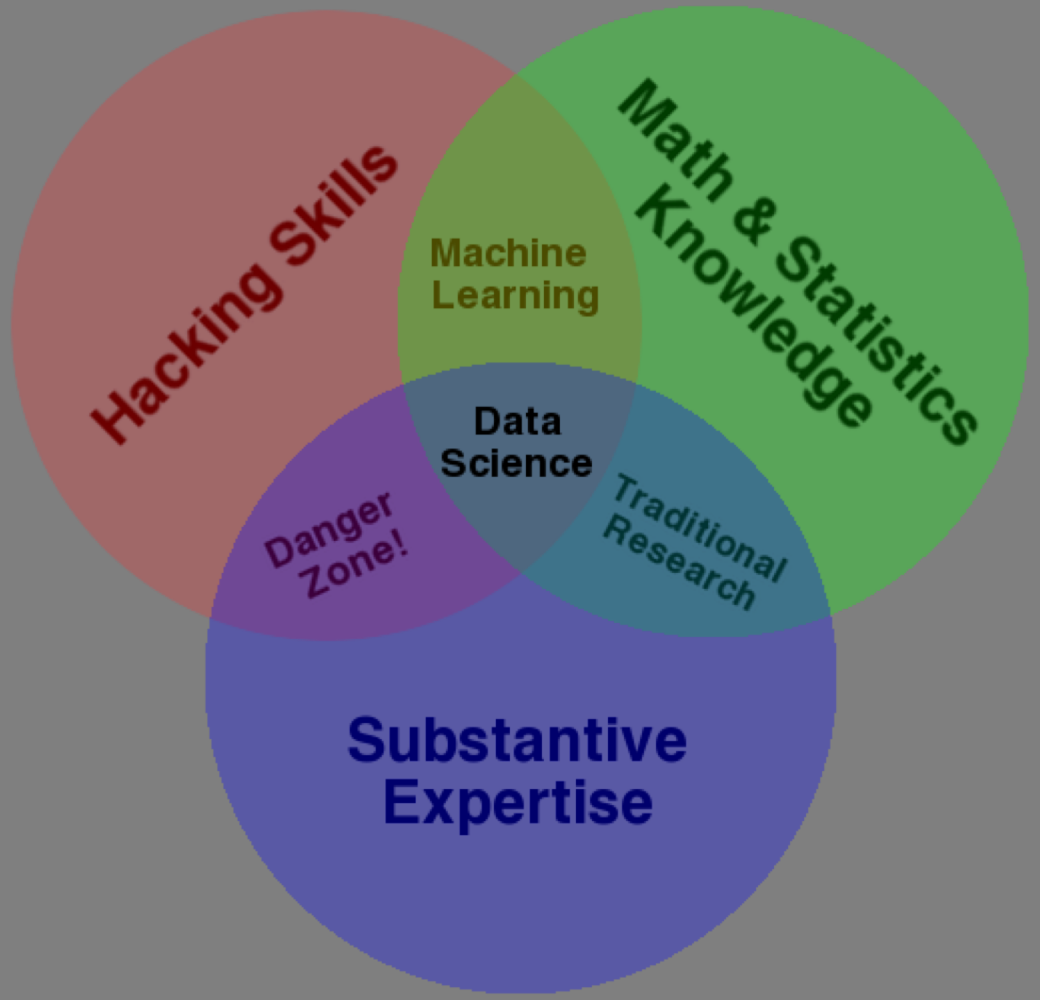
DATA SCIENCE APPLICATIONS

- 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
- Genotype and Phenotype Data → 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"

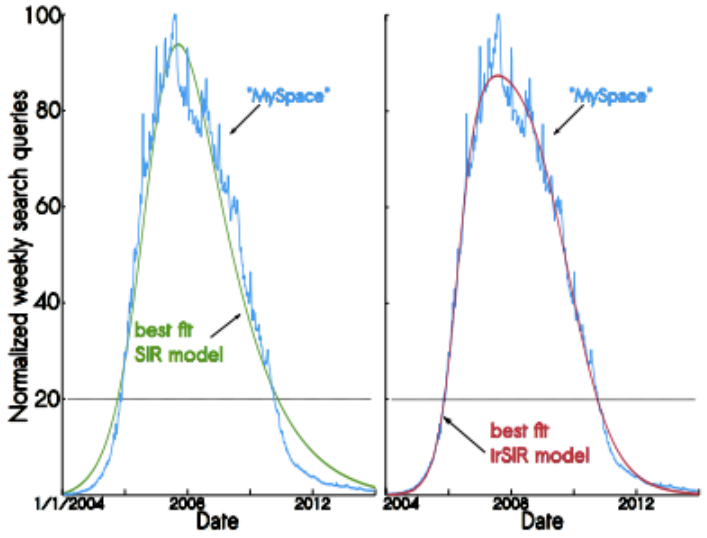
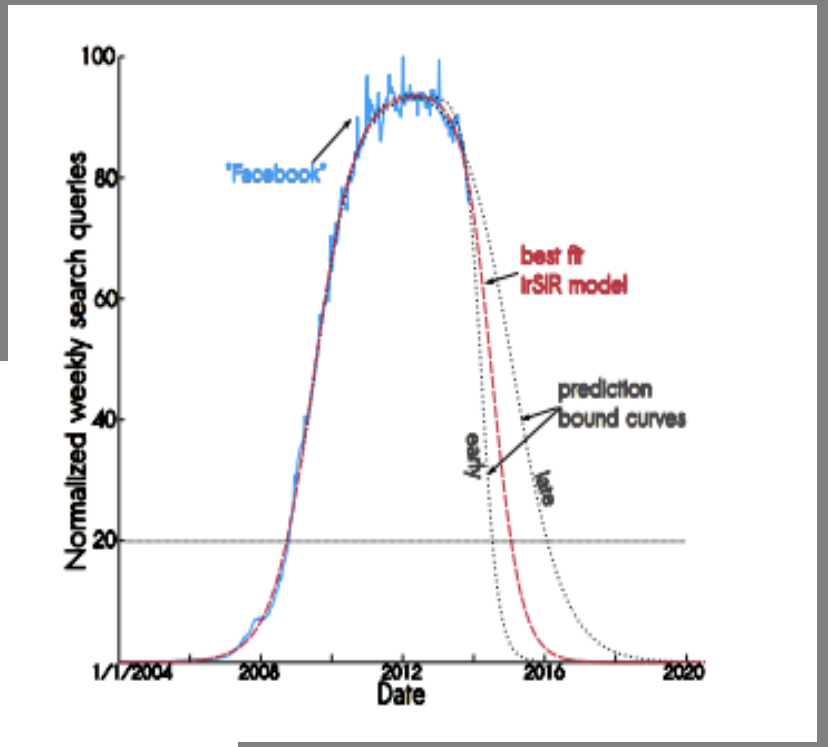


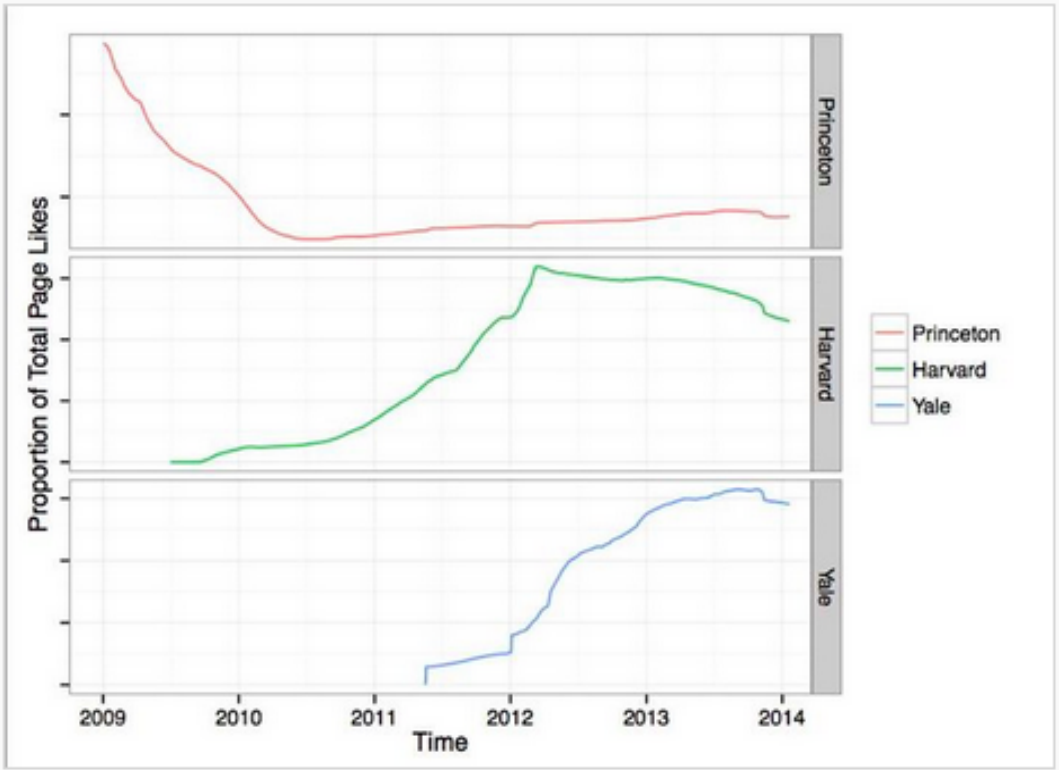
Figure 3: **Data for search query "Myspace" with best fit (a) SIR and (b) irSIR models overlaid.** The search query data are normalized such that the maximum data point corresponds to a value of 100.

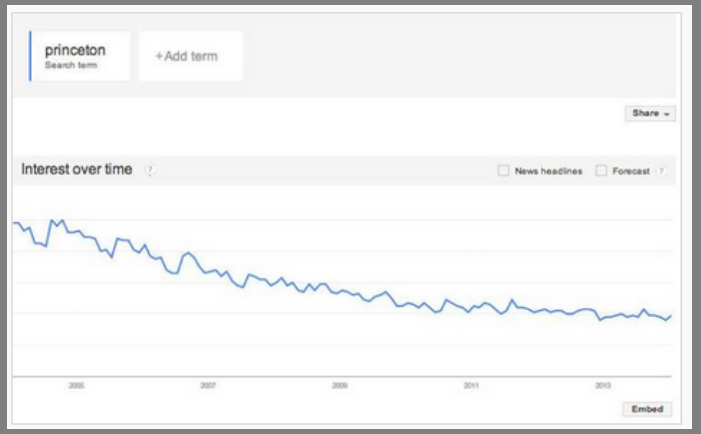


Searches for "Facebook"

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:



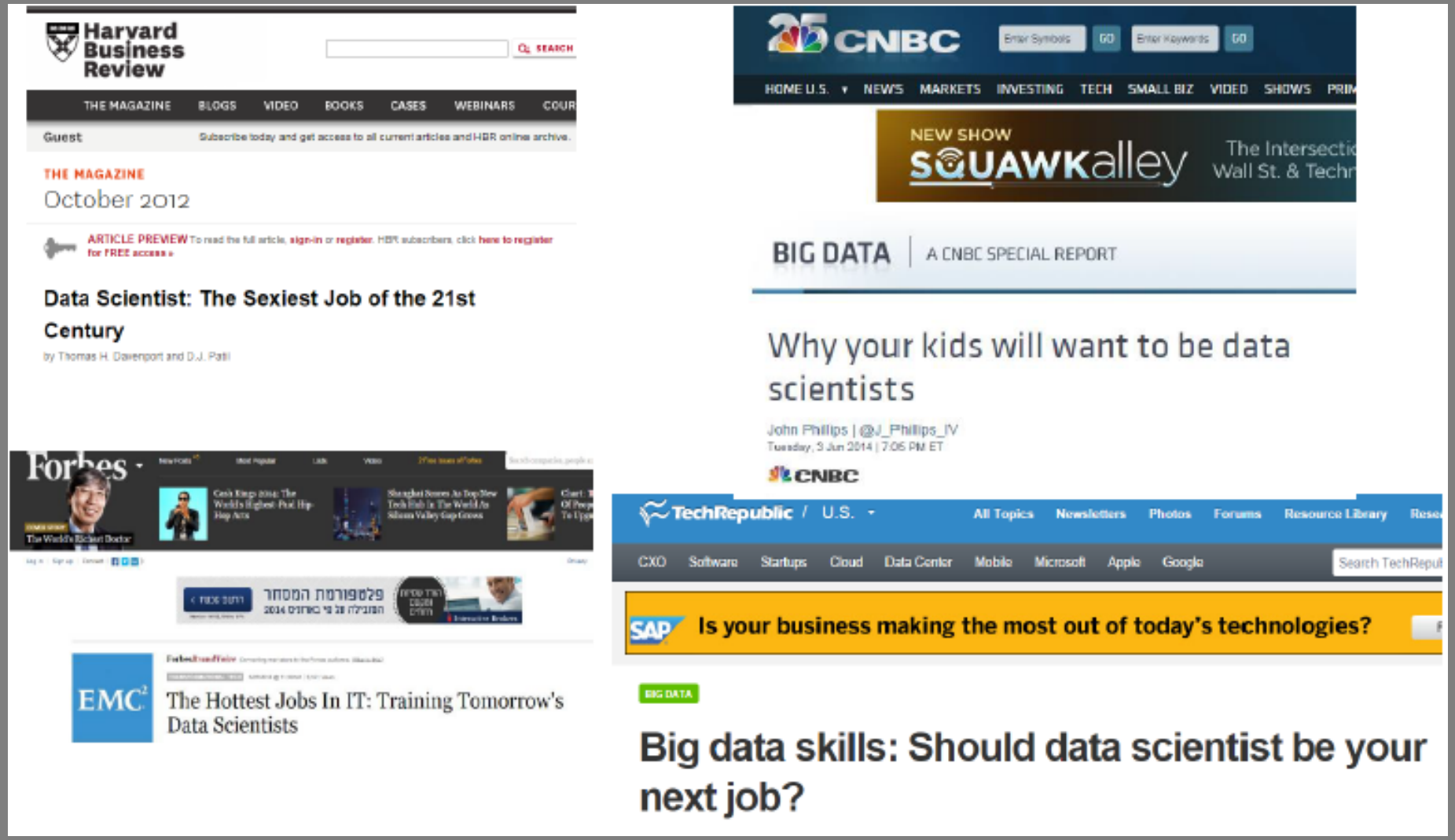


and based on Princeton search trends:

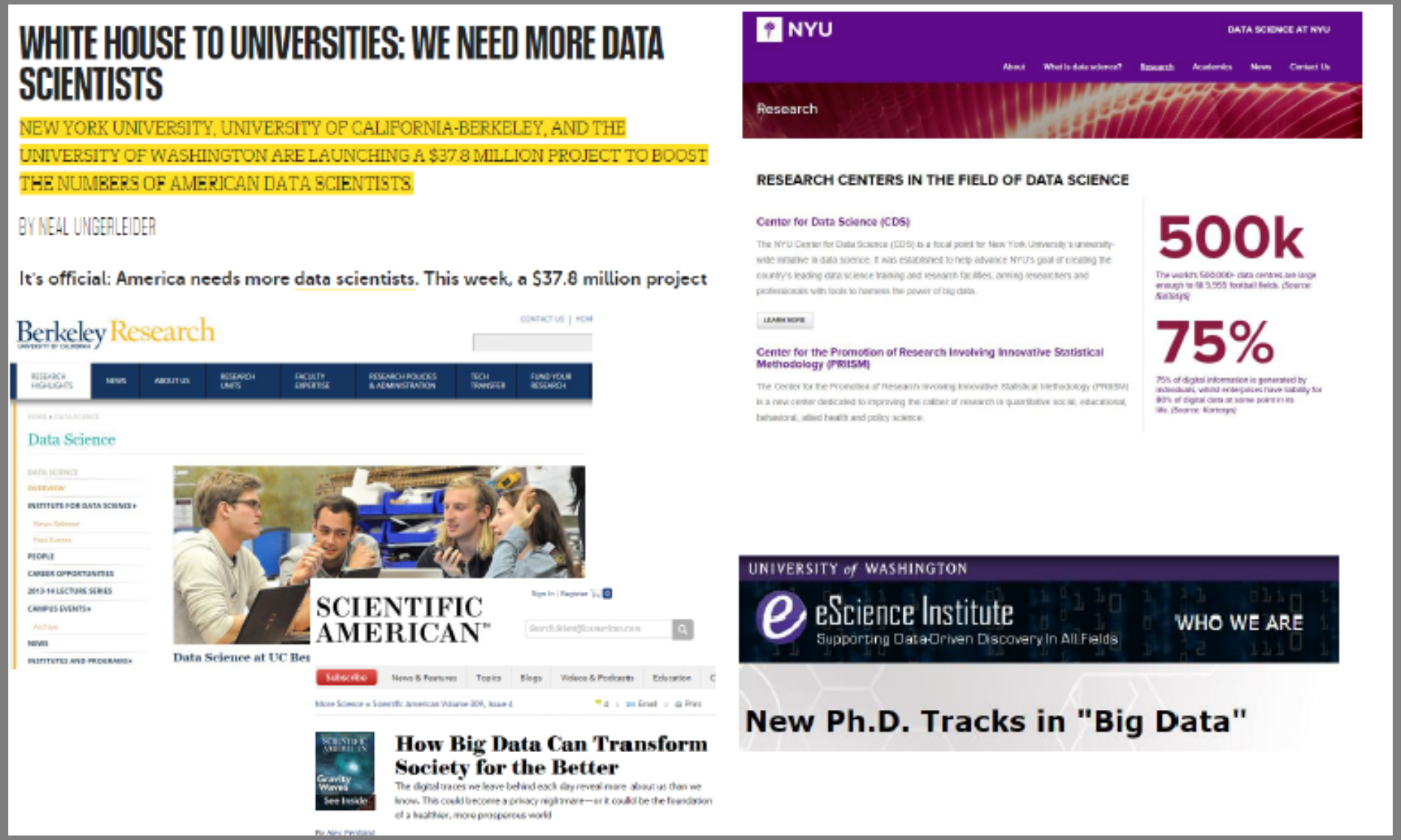
"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…

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

DATA SCIENTISTS ARE IN HIGH DEMAND

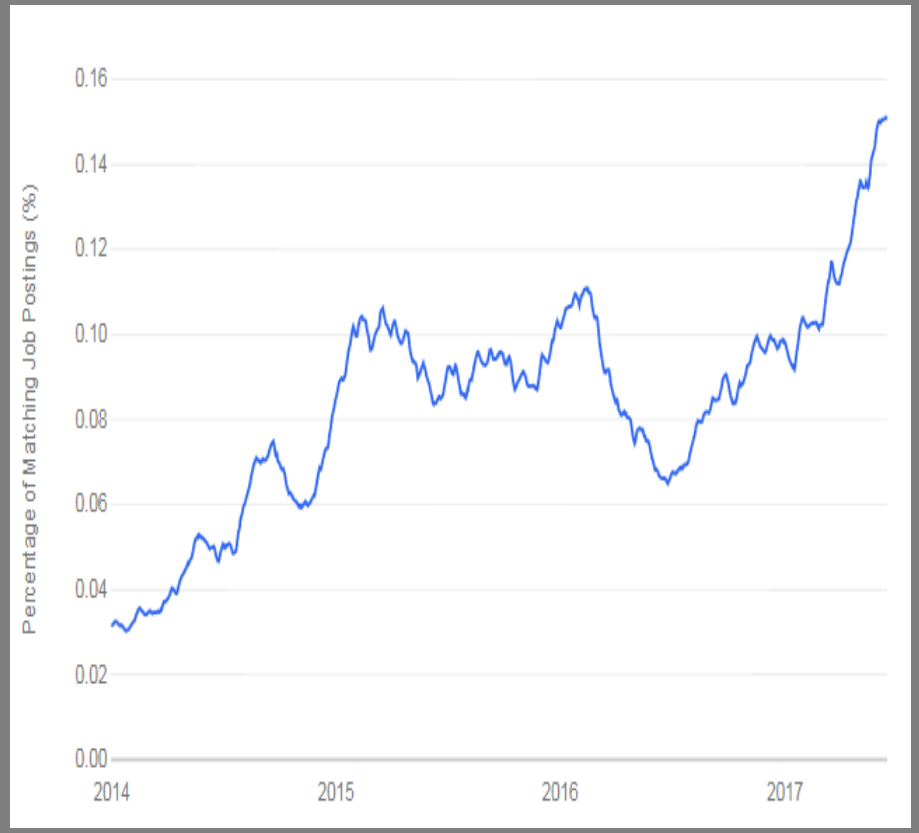


ALSO IN ACADEMIA

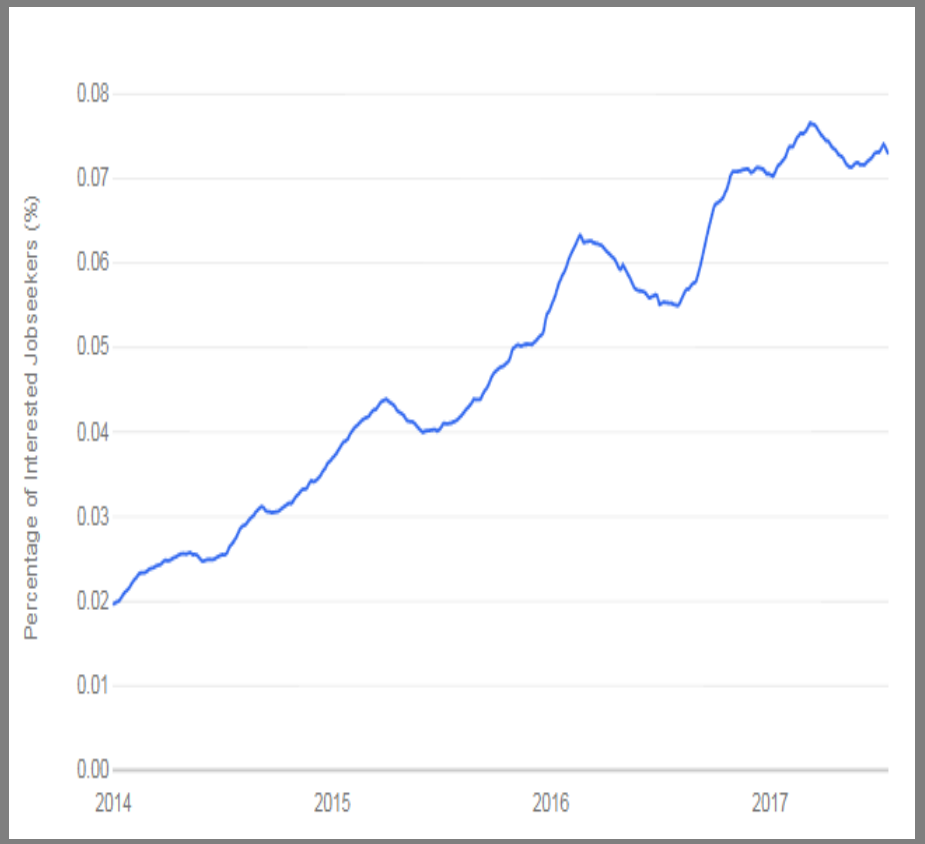


DATA SCIENTIST JOB TREND

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?

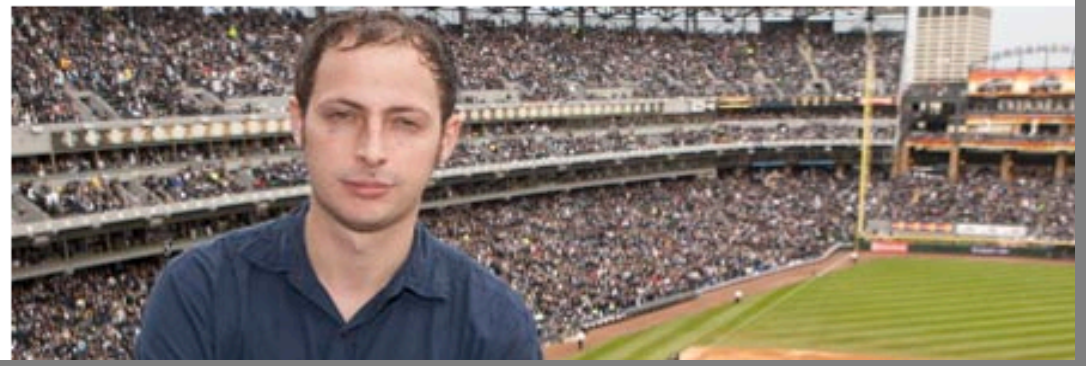
elections2012

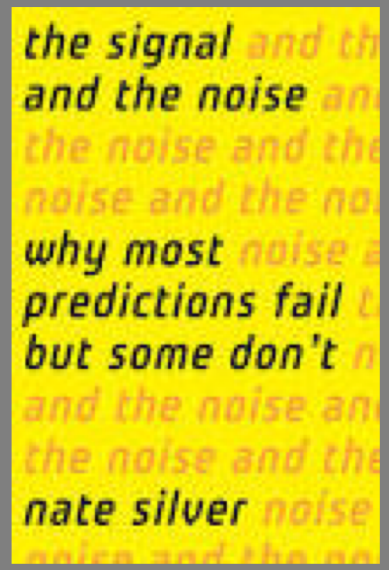
Live results | President | Senate | House | Governor | Choose your

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

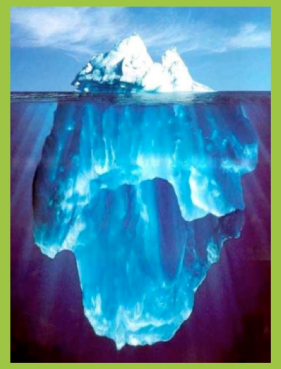
Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST





“BIG DATA” SOURCES

It’s All Happening On-line



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

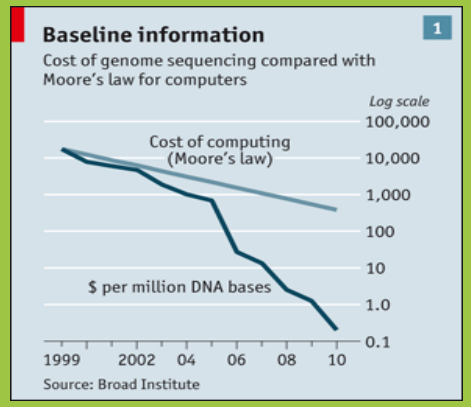
User Generated (Web & Mobile)




…
..


Internet of Things / M2M

Health/Scientific Computing

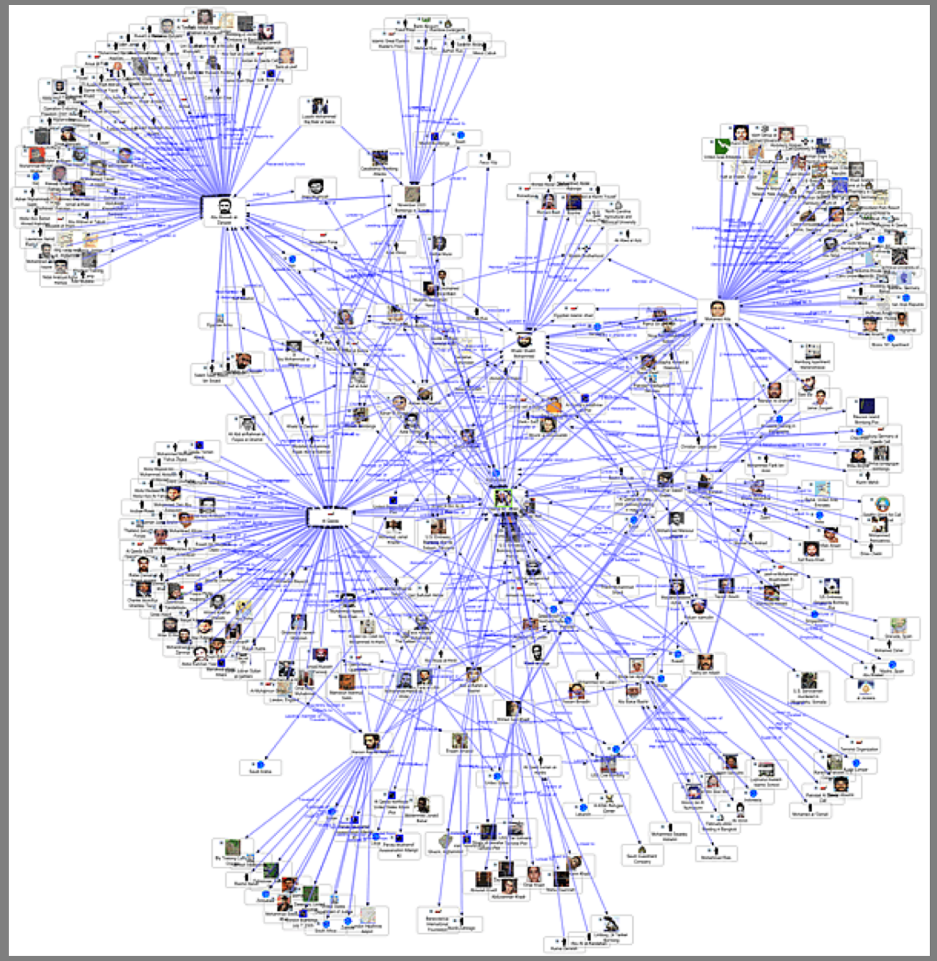


GRAPH DATA

Lots of interesting data
has a graph structure:

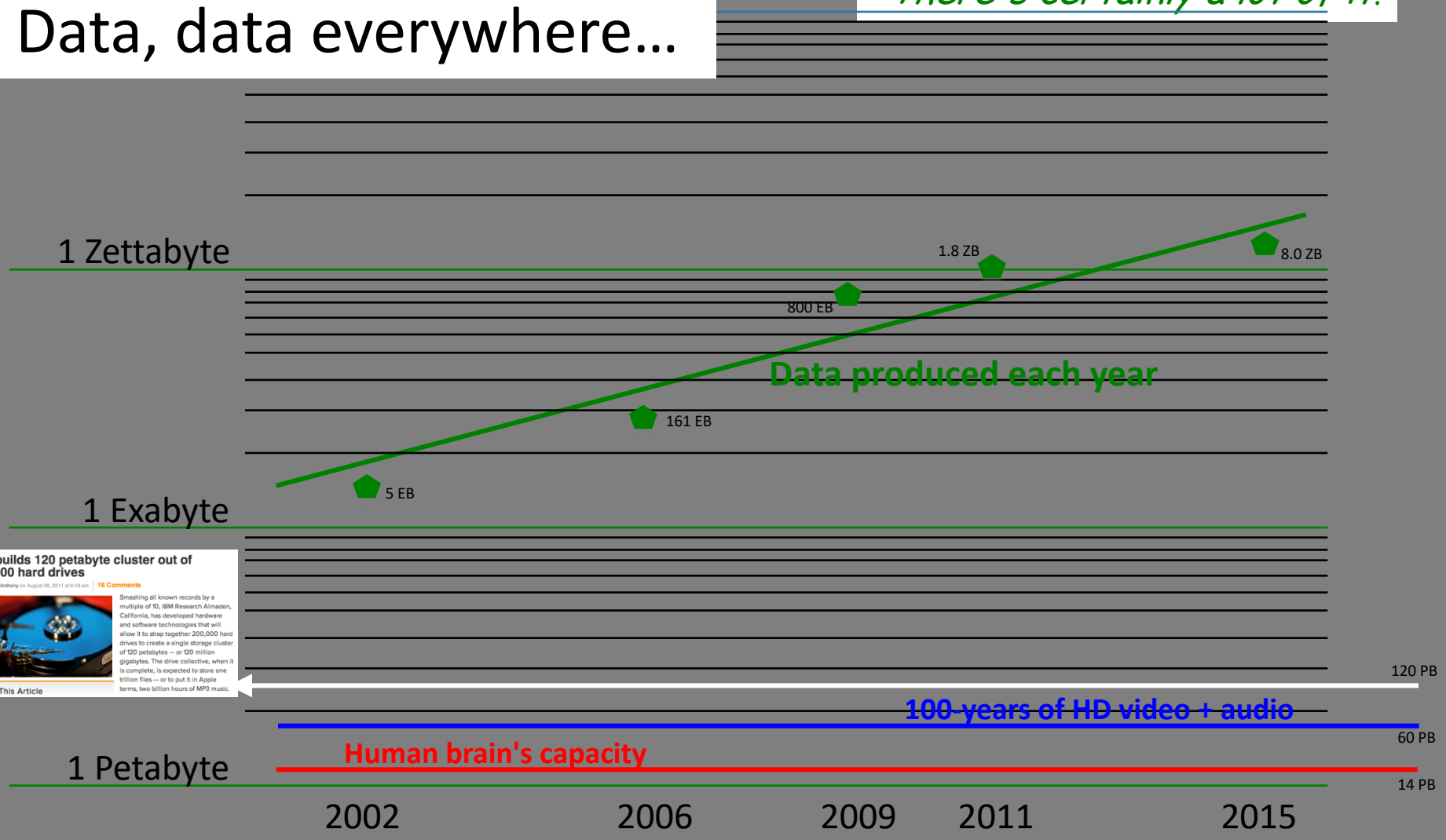
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

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

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

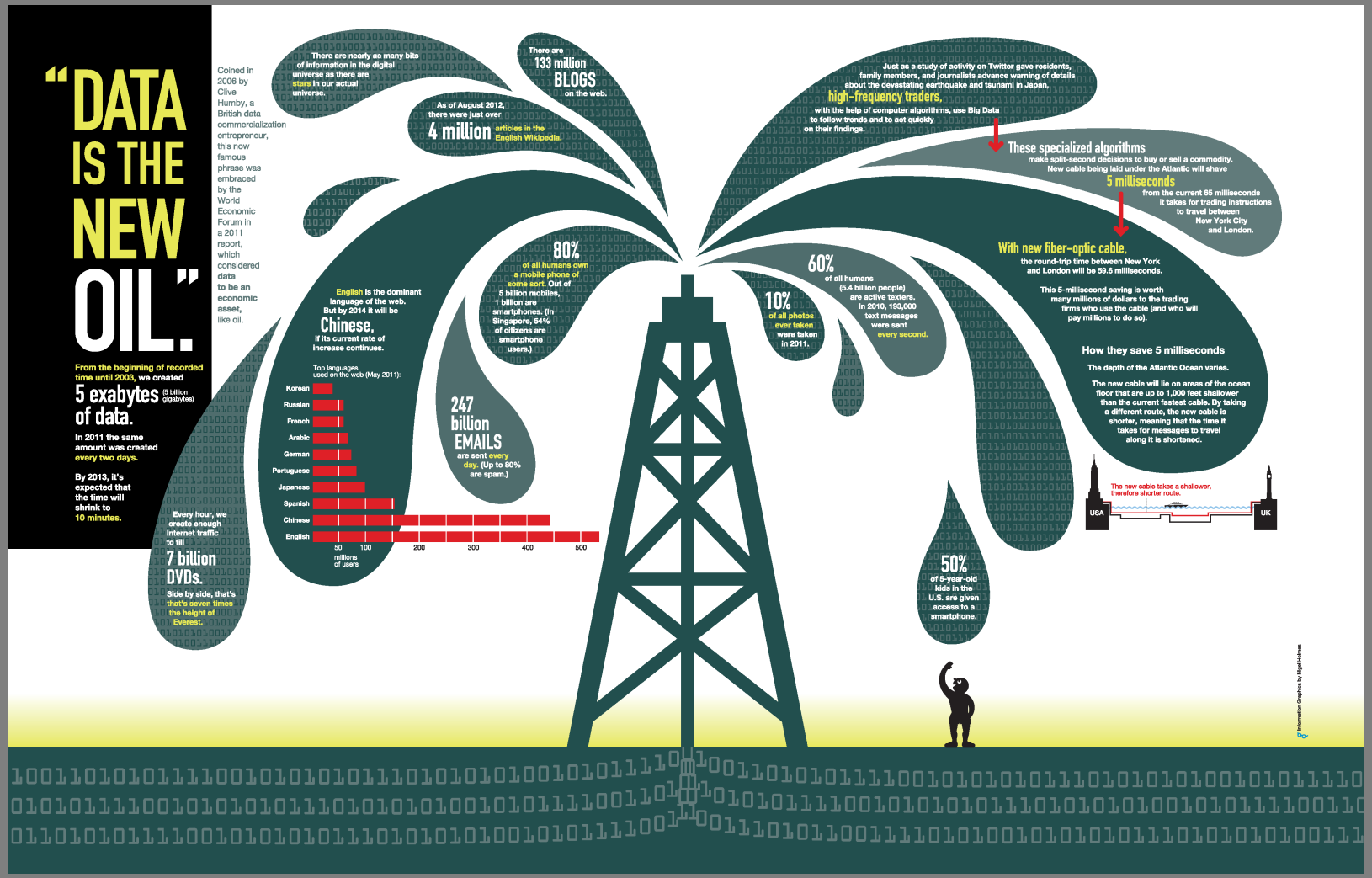
(2006) 161 EB: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

(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



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