

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

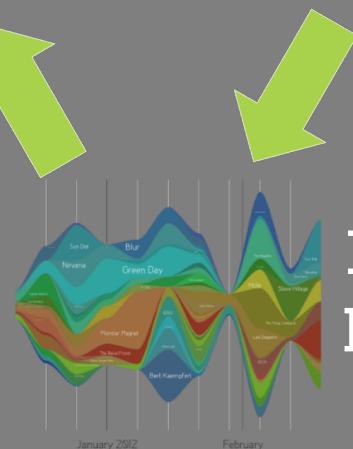


Digging Around in Data

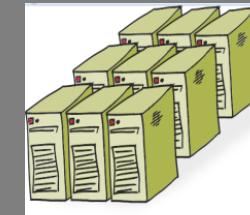
Clean, prep

$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \underline{\underline{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}}$$

Hypothesize Model



Evaluate Interpret



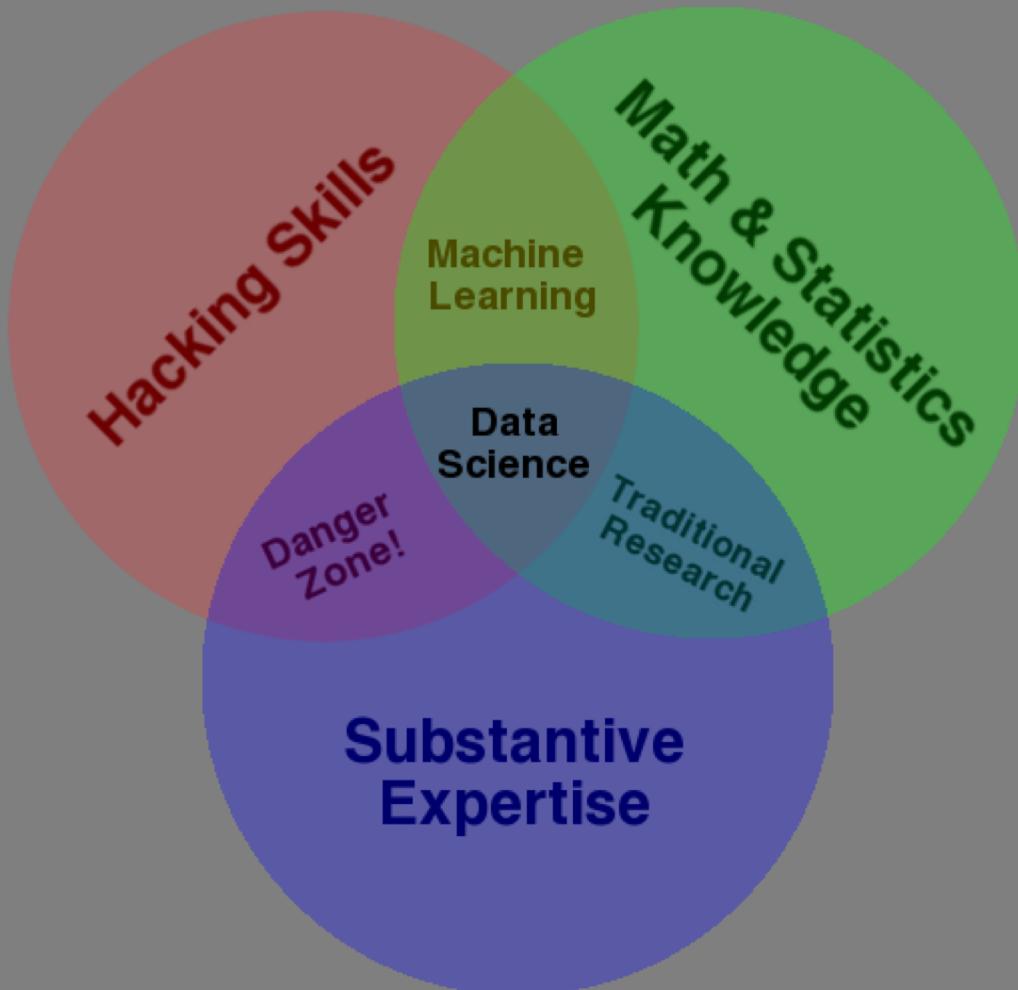
DATA SCIENCE APPLICATIONS

- 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 Cannarella¹, Joshua A. Spechler^{1,*}

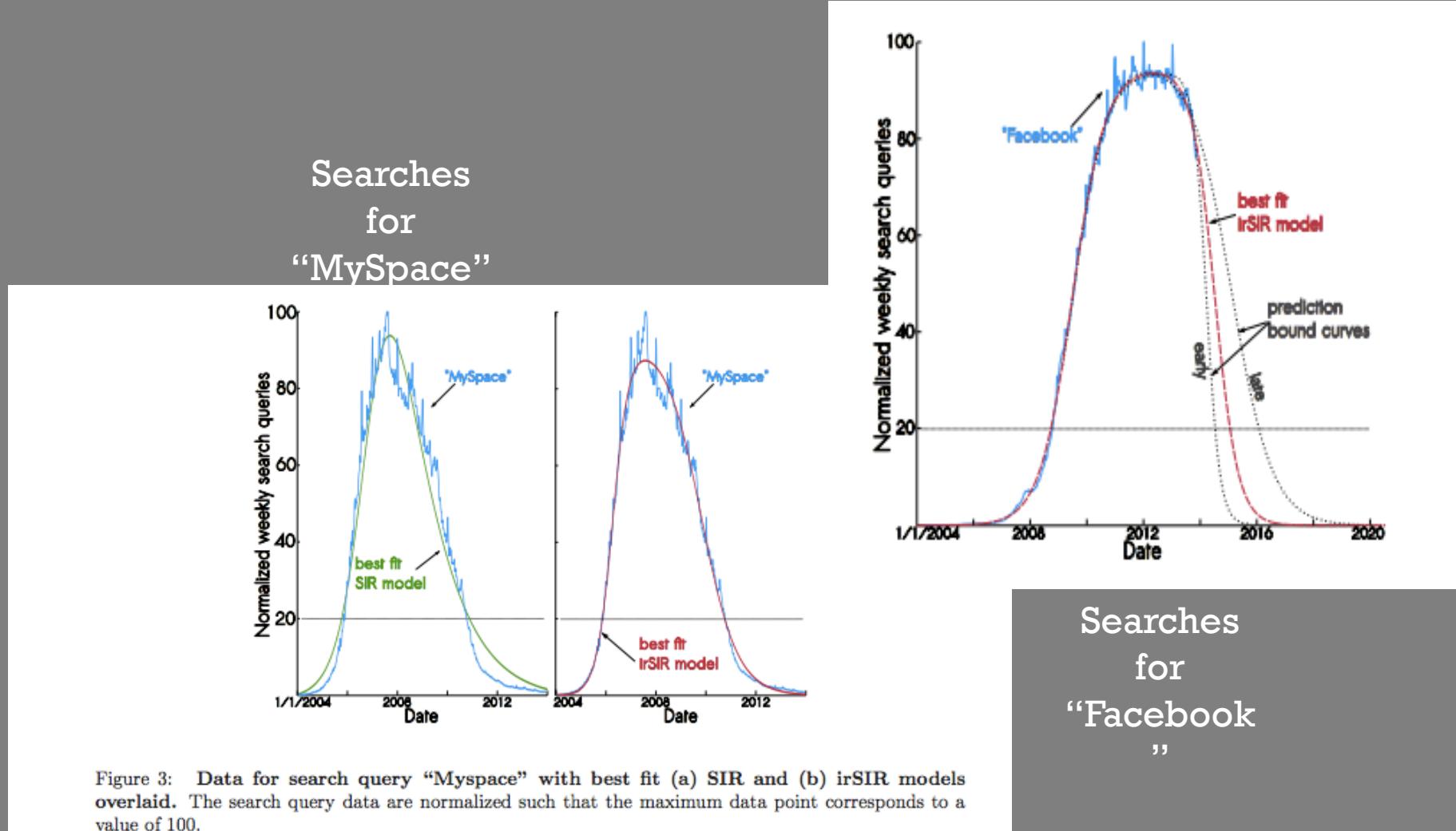
¹ Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

* E-mail: Corresponding spechler@princeton.edu

Abstract

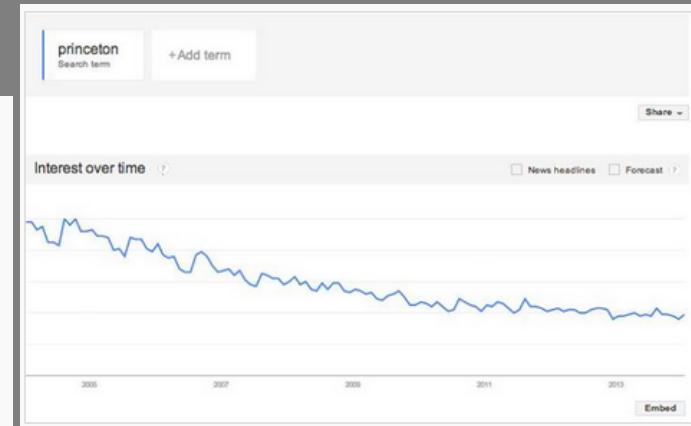
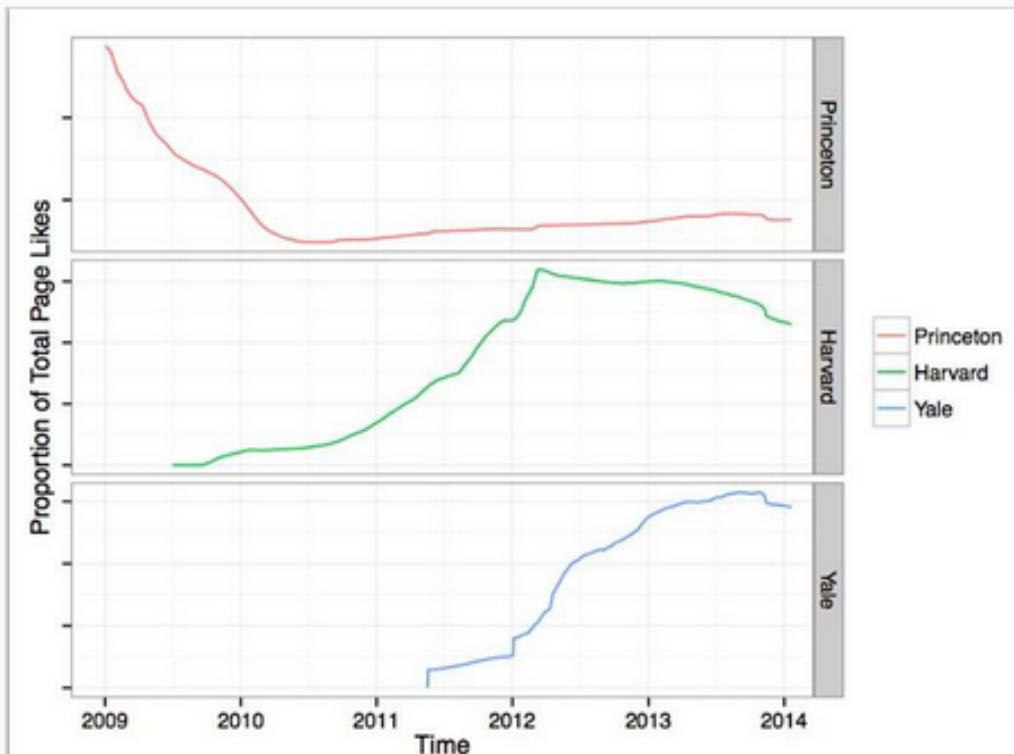
The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional 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?



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,..."

DATA SCIENTISTS ARE IN HIGH DEMAND

The screenshot shows the Harvard Business Review website. At the top, there's a search bar and navigation links for 'THE MAGAZINE', 'BLOGS', 'VIDEO', 'BOOKS', 'CASES', 'WEBINARS', and 'COURSES'. Below that, a guest user is logged in. A banner for 'THE MAGAZINE' October 2012 issue is displayed, featuring an article titled 'Data Scientist: The Sexiest Job of the 21st Century' by Thomas H. Davenport and D.J. Patil. A key icon indicates an 'ARTICLE PREVIEW'. The main content area has a dark background with white text.

The screenshot shows the CNBC website. At the top, there's a search bar and navigation links for 'HOME U.S.', 'NEWS', 'MARKETS', 'INVESTING', 'TECH', 'SMALL BIZ', 'VIDEO', 'SHOWS', and 'PRIMETIME'. A banner for 'NEW SHOW SQUAWKalley' is prominently displayed. Below it, a section titled 'BIG DATA | A CNBC SPECIAL REPORT' is shown. Another section discusses 'Why your kids will want to be data scientists' by John Phillips. The bottom part of the page features a SAP advertisement and another 'BIG DATA' section.

The screenshot shows the TechRepublic website. At the top, there's a search bar and navigation links for 'U.S.', 'All Topics', 'Newsletters', 'Photos', 'Forums', 'Resource Library', and 'Research'. Below that, a navigation bar includes links for 'CXO', 'Software', 'Startups', 'Cloud', 'Data Center', 'Mobile', 'Microsoft', 'Apple', and 'Google'. A prominent yellow banner at the bottom features the SAP logo and the text 'Is your business making the most out of today's technologies?'. The overall design is clean with a blue header and white text.

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
UNIVERSITY OF CALIFORNIA

RESEARCH HIGHLIGHTS | NEWS | ABOUT US | RESEARCH UNITS | FACULTY EXPERTISE | RESEARCH POLICIES & ADMINISTRATION | TECH TRANSFER | FUND YOUR RESEARCH

CONTACT US | HOME

HOME > DATA SCIENCE

Data Science

DATA SCIENCE

OVERVIEW

INSTITUTES FOR DATA SCIENCE

Names Reference

Press Releases

PEOPLE

CAREER OPPORTUNITIES

2013-14 LECTURE SERIES

CAMPUS EVENTS

Archives

NEWS

INSTITUTE AND PROGRAMS



SCIENTIFIC AMERICAN™

Sign In / Register

Search: Search ScientificAmerican.com

More Science in Scientific American Volume 309, Issue 4

Subscribe

News & Features

Topics

Blogs

Videos & Podcasts

Education

75%

WHO WE ARE

University of Washington

eScience Institute

Supporting Data-Driven Discovery In All Fields

New Ph.D. Tracks in "Big Data"

NYU

DATA SCIENCE AT NYU

About | What is data science? | Research | Academics | News | Contact Us

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, among researchers and professionals, with tools to harness the power of big data.

LEARN MORE

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 centers are large enough to fit 5.555 football fields. (Source: Gartner)

75%

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

University of Washington

eScience Institute

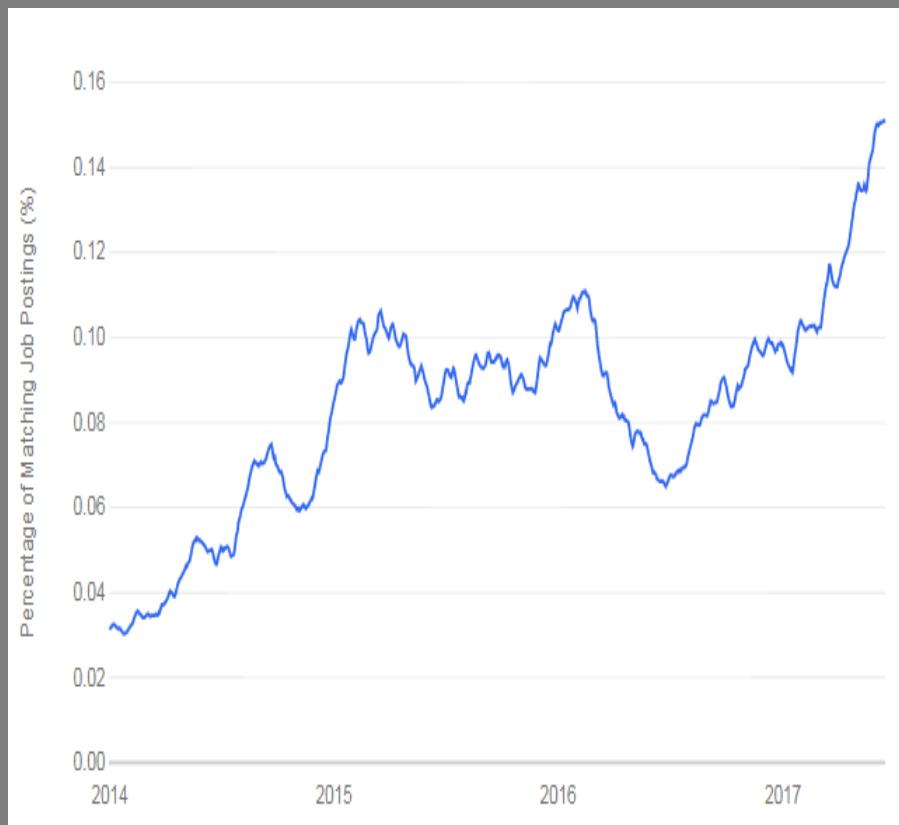
Supporting Data-Driven Discovery In All Fields

WHO WE ARE

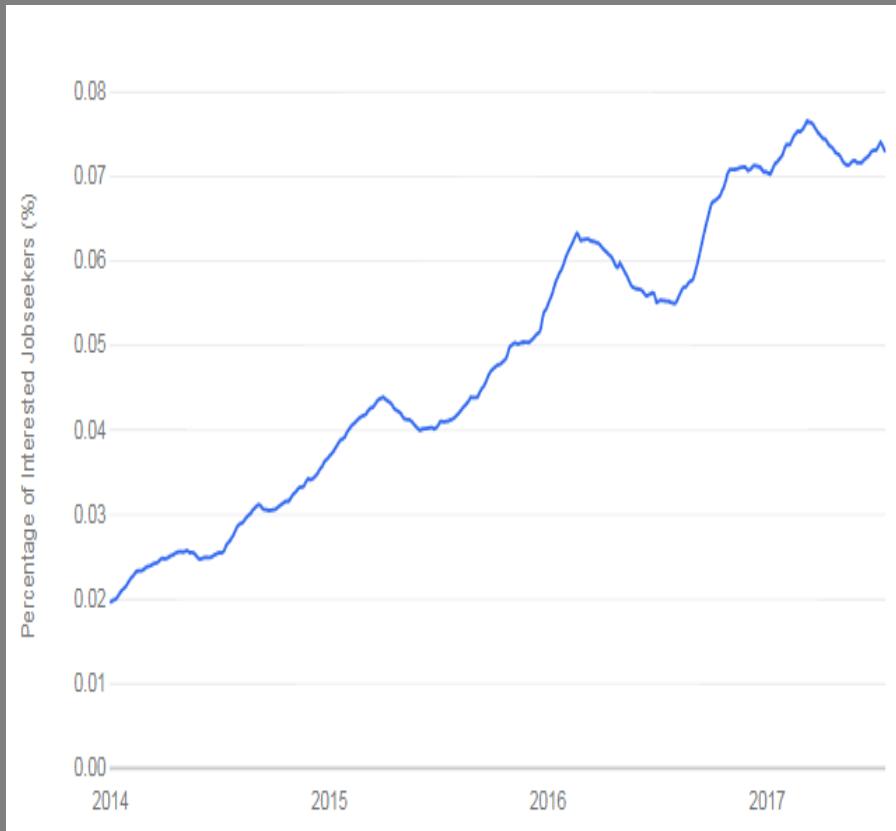
New Ph.D. Tracks in "Big Data"

DATA SCIENTIST JOB TREND

Job postings

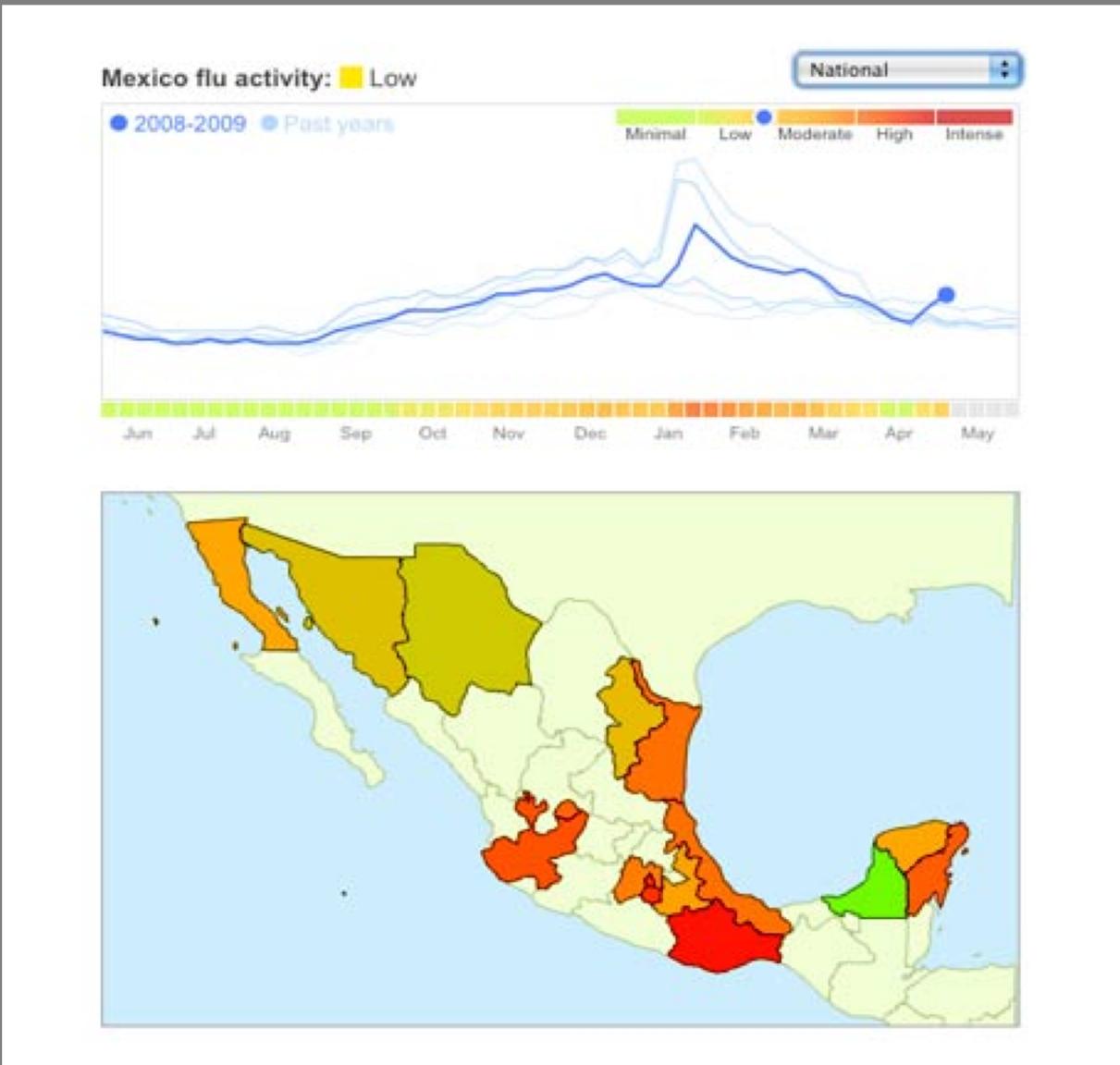


Jobseeker interest



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?

elections2012

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

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

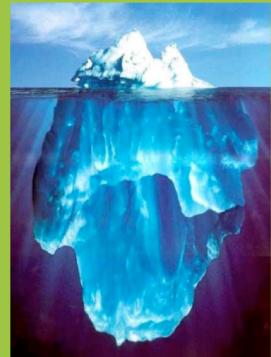
Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST



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

“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)



...

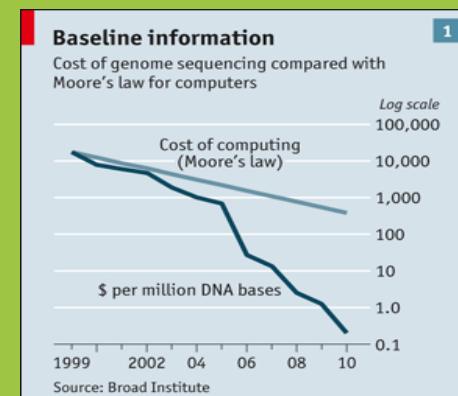


...

Internet of Things / M2M



Health/Scientific Computing

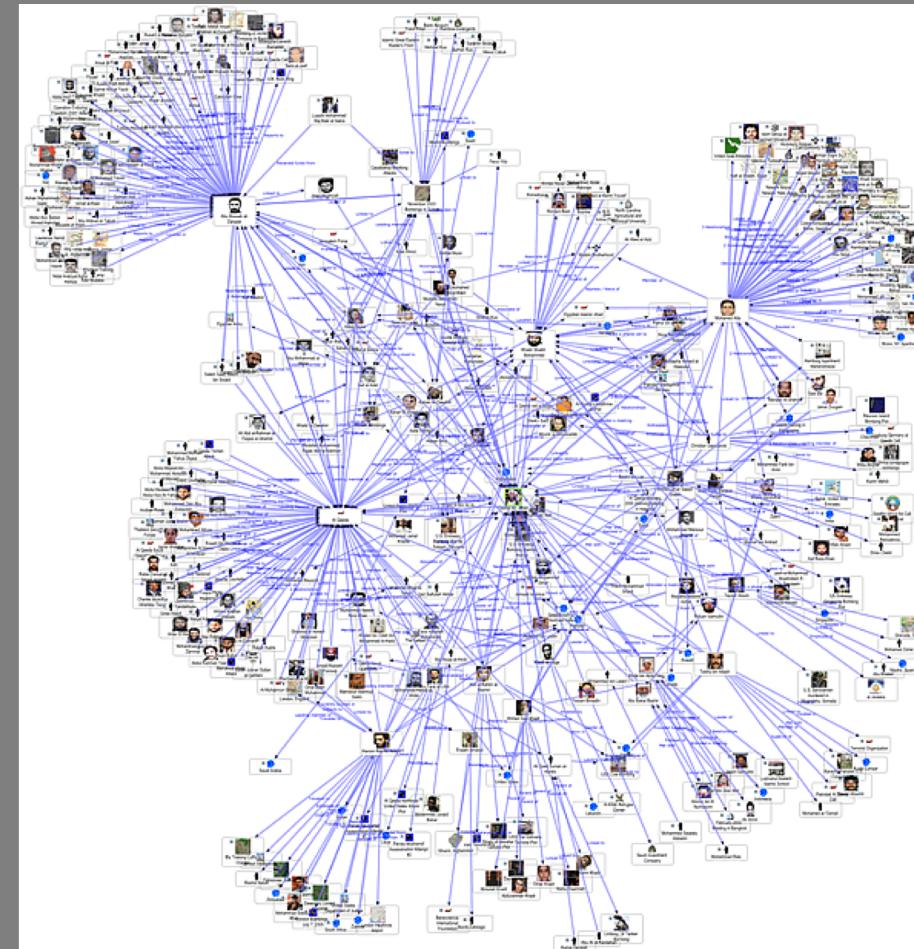


GRAPH DATA

Lots of interesting data
has a graph structure:

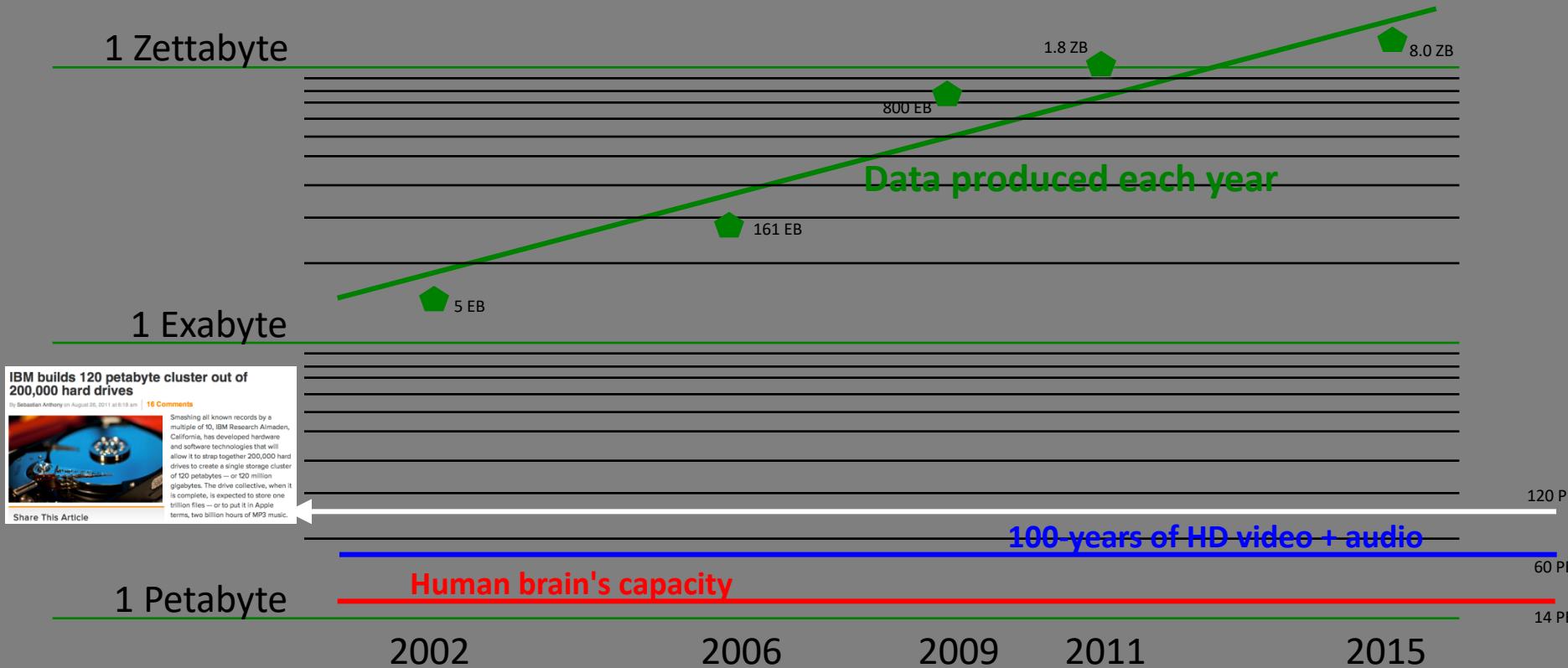
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

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



Data, data everywhere...

There's certainly a lot of it!



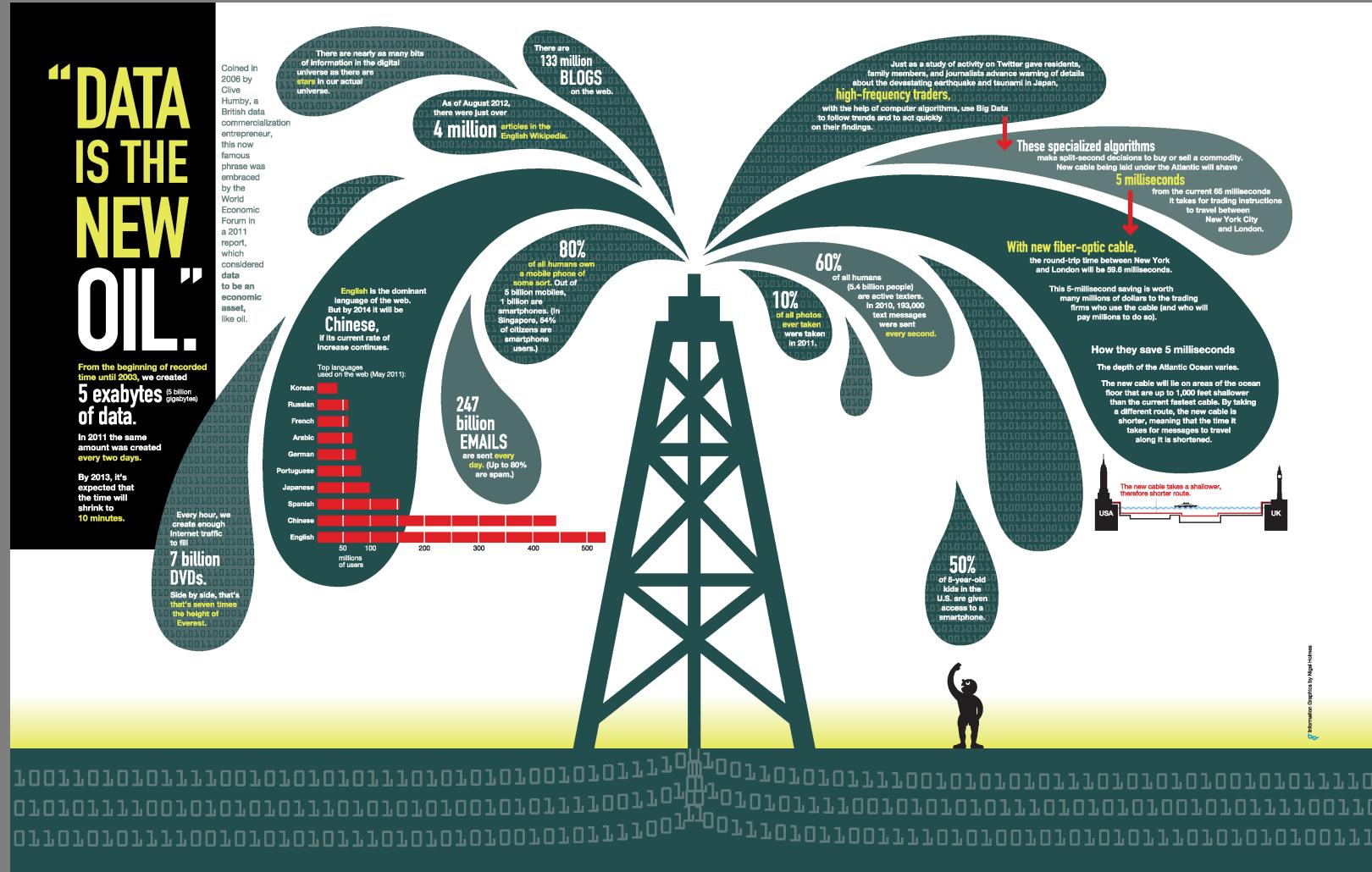
References

- (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