Applied Data Analysis (CS401)

Robert West

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

dlab

Important websites

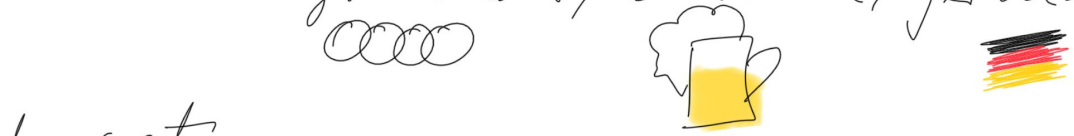
http://ada.epfl.ch
(Slides linked from there)

https://github.com/epfl-ada-2018

https://icmattermost.epfl.ch/cs-401
Sign up at above link (use your EPFL email address)

ABOUT ME

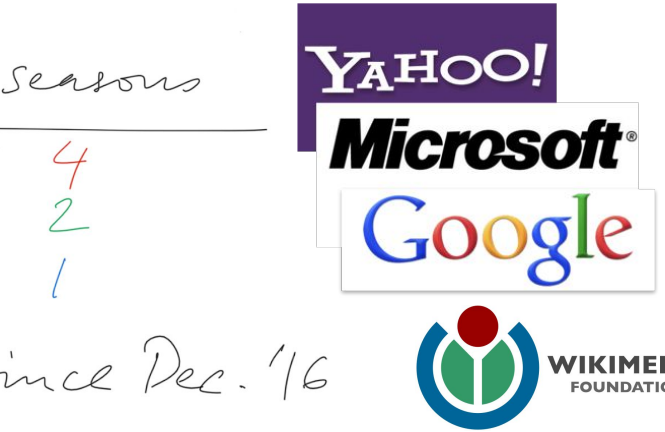
- Born in Ingolstadt, Bavaria, Germany



- Education:

School	Location	Degree	# seasons
TUM	Germany	Diplom	4
Mcgill	Canada	MS	2
Stanford	USA	PhD	1

- Assistant professor at EPFL since Dec. '16
- Heading Data Science Lab



ABOUT ME

- Born in Ingolstadt, Bavaria, Germany

- Education:

School	Location	Degree	# seasons
TUM	Germany	Diplom	4
McGill	Canada	MS	2
Stanford	USA	PhD	1

- Assistant professor at EPFL since Dec. '16
- Heading Data Science Lab
- Recent daddy

6.10.16
4444 g
55.5 mm

dlab
011011

dlab

- Born in
- Educatio

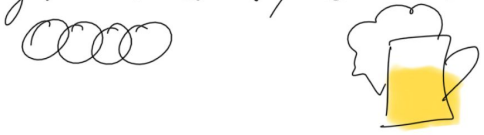
School
TUM
Mcgill
Stanford

- Assistan
- Heading
- Recent da

Dec. '16

ABOUT ME

- Born in Ingolstadt, Bavaria, Germany



- Education:

School	Location	Degree	# seasons
TUM	Germany	Diplom	4
McGill	Canada	MS	2
Stanford	USA	PhD	1

- Assistant professor at EPFL since Dec. '16
- Heading Data Science lab
- Recent daddy

6.10.16
4444 g
555 mm

dlab
ollolo

dlab

ada

⇒ Call me Bob

bob
loolo

My path toward data science

– Born in Ingolstadt, Bavaria, Germany

MY RESEARCH

Develop neat algorithms

Distill raw data into insights into people & the world

Give back to the real world (⇒ applications)

DATA SCIENCE

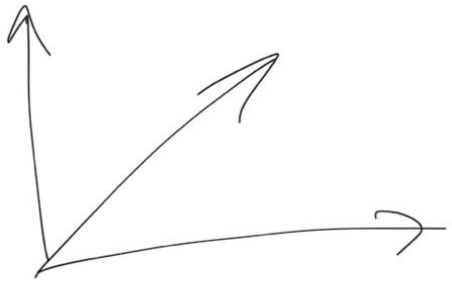
Leverage large datasets

Address real-world issues

Draw meaningful conclusions from "found data" (a.k.a. observational studies)

When necessary, generate my own data (human computation, crowdsourcing)

BUZZWORDS



Machine learning

Data mining

Social & information network analysis

Computational social science

Natural language processing

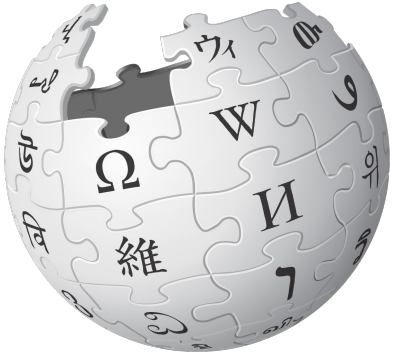
Human computation, crowdsourcing

Information retrieval

Data analysis

"... the process of **inspecting, cleaning, transforming, and modeling data** with the goal of **discovering useful information**, suggesting conclusions, and supporting decision-making."

"Data analysis has multiple facets and approaches, encompassing **diverse techniques** under a variety of names, **in different** business, science, and social science **domains.**"



Applied data analysis

- This course is about **breadth,** not depth
- *"What methods, principles, and tools are out there?"*, rather than *"How can I become an expert in deep learning for computer vision?"*
- Data science is a fast-paced, shifting field
- Obsessing on one tool or technique won't pay off in a few years
- Be ready to explore and keep learning on your own
- Goal of this class: Enable you to conduct a full-fledged data science project from start to finish
- That said, depth matters, too…

Complementary courses:
Machine learning
NLP
DIS
Data viz

Let’s call this course **Ada,** not A-D-A, in honor of Ada Lovelace, “the world’s first computer programmer.”

https://en.wikipedia.org/wiki/Ada_Lovelace

Topics

- Obtaining, preparing, managing, handling data
  - “Data wrangling”, “slicing and dicing”, etc.
- Stats
  - How to support (and be suspicious of) claims about data
- Observational studies
  - How to deal with “found data”
  - Correlation != causation
- Data visualization
  - Exploration of data, communication of results

Topics (cont'd)

- Applied machine learning
  - Supervised learning
  - Unsupervised learning
- Handling specific types of data
  - Text
  - Graphs, networks
- Scaling up to terabytes and beyond
  - MapReduce, Spark

Meeting times

- **Thursdays** 8:15 - 10:00 in SG 1
  - Lectures
- **Fridays** 13:15 - 15:00 in BCH 2201
  - “Lab sessions”
  - Hands-on tutorials
  - Homework post-mortem
  - Invited speakers from industry and academia
  - Second half of each session: “office hours”
  - Bring your laptops!

Grading

- 30% **Homework** (4 during the semester)
  - Involving skills required from data scientists
  - Groups of 3 students (may switch groups after HW1)
  - Peer reviewing (more details later)
  - Last year's homework
- 30% **Final exam** (in January, TBD)
  - Mini data analysis project
  - Done on laptop, individually
  - Last year's final
- 40% **Project** (more details in October)
  - Proposal (15%), milestone (15%), report (50%), poster presentation (20%)
  - Groups of 3 students (same as for homework)
  - Last year's projects
- Average of {homework, final, project} to be computed as **harmonic mean**
  - arithmetic_mean([1,6]) = 3.5; harmonic_mean([1,6]) = 1.7

Project

- We'll provide datasets
- You need to form and pitch a crisp project idea
- Free to propose own datasets (more risky)
- Powerful compute cluster, running Apache Spark
- Goal: not a loose collection of results -- tell a story with the data!
  - Last year's data stories
  - Nice example data story

Main communication channel: Mattermost

- Open-source Slack alternative
- Channel: https://icmattermost.epfl.ch/cs-401
- Sign up at above link (use your EPFL email address!)
- Install desktop and mobile clients
- Mandatory! We'll send all important announcements only here
- Feel free to use for private team communication

Homework and projects: GitHub

- De-facto standard for managing and sharing code
- All students in this class need a GitHub account
- Register it here: https://goo.gl/EfVKhJ (complete form once per team; by next Friday, 9/28)
- Homework and project submissions done via GitHub

Prerequisites

Basics of

- Databases
- Probabilities and stats
- Programming
  - You won’t survive if you can’t program
  - Homework: Python required
  - Project: up to you, but we support only Python
  - Brush up your Python skills (many great online courses out there)

Anaconda: our Python environment

- We only support Anaconda (“Adaconda”...)
- Introduction: tomorrow’s lab session
- Before lab session:
  - Install Anaconda on your laptop
  - https://www.anaconda.com/download
  - Version: Python 3.6!

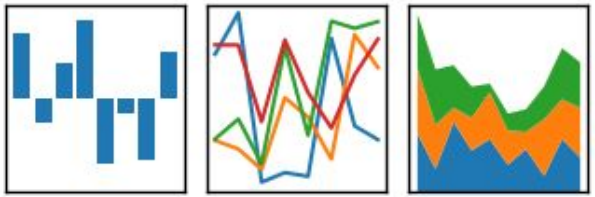
Python++

machine learning in Python

Lightning-Fast Cluster Computing

pandas

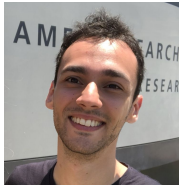
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Instructor

Bob West

Teaching assistants (TAs)



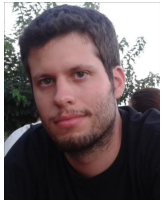
Tiziano Piccardi

Jérémie Rappaz

Kristina Gligoric

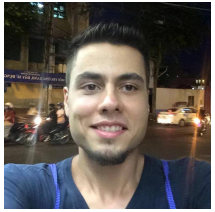
Ramtin Yazdanian

Panagiotis Sioulas



Panayiotis Smeros

Periklis Chrysogelos



Tugrulcan Elmas

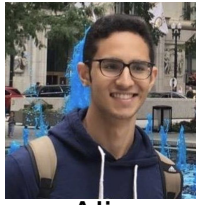
Student assistants (AEs)

Téo Stocco

Nuno Gonçalves

Léonore Guillain

Armand Boschin



Ali Benlalah

Samuel Beuret

Quentin Bacuet

Quentin Rebjock

- Help each other on Mattermost
- Participate actively in classes and labs
- Give us **feedback**
- **Peer reviewing** (next slide)

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2018-lec1-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- What would you like the instructor to (not) wear next time?
- …

Peer reviewing

- You will read and provide feedback on 2 other groups' homework submissions
- **Give:** 10 (out of 110) points for giving us feedback, per homework
- **Take:** you get to see 6 reviews of your work
- At end of semester: Best Reviewer Awards

Questions?

“Data science”

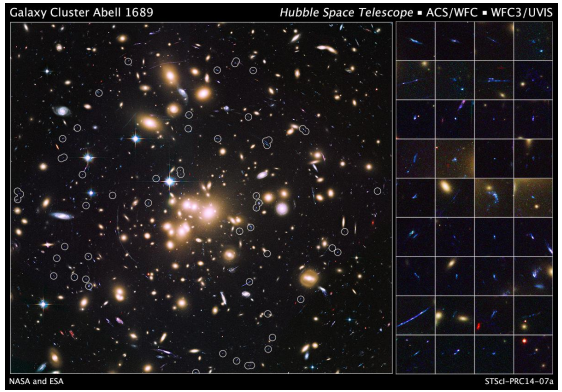
- All science is (or should be) based on data, per definitionem
- So how is “data science” different from plain old “science”?

Data volume explodes

"Between the dawn of civilization and 2003, we only created **five exabytes** of information; now we're creating that amount **every two days.**"

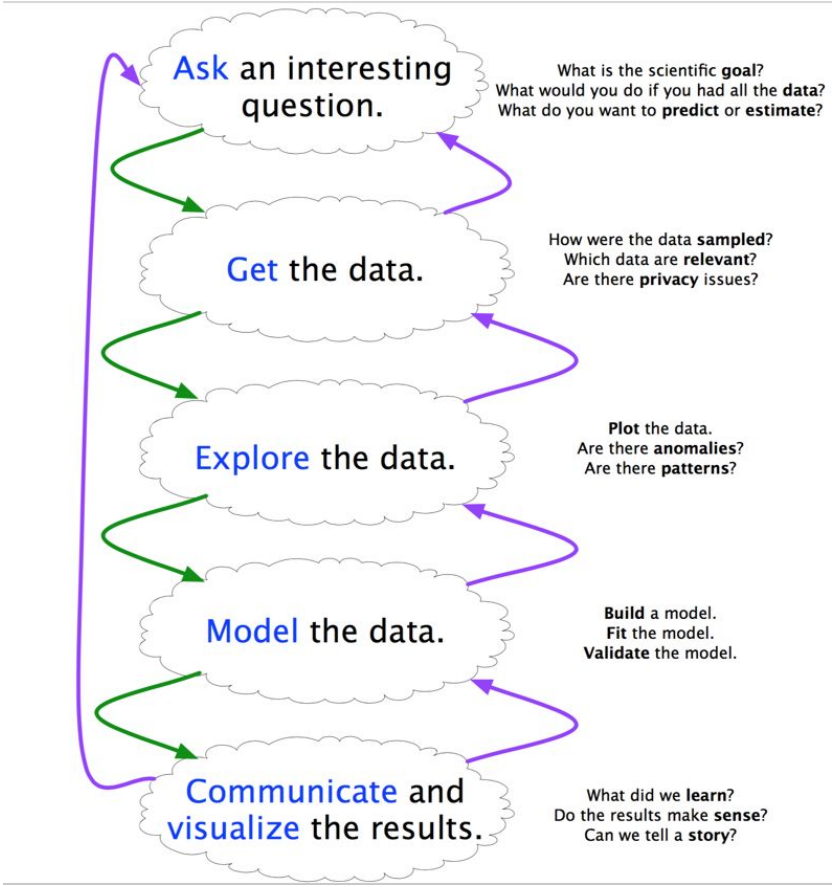
Eric Schmidt, Google (and others)

Data variety explodes



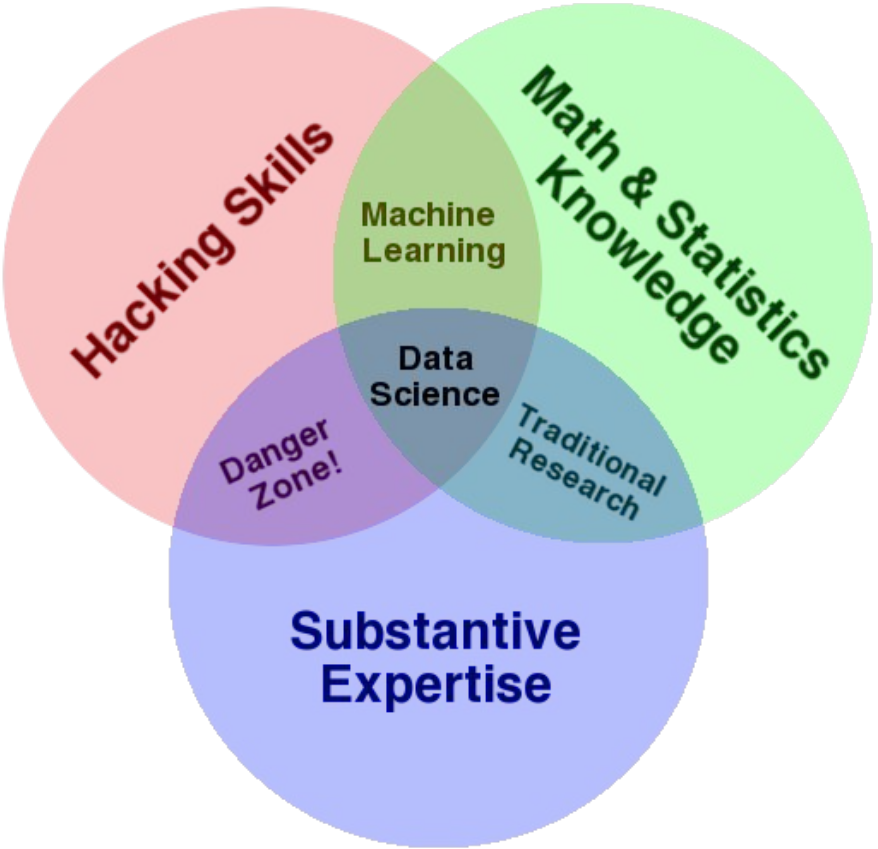
Text (indexed Web pages, email), **networks** (Web graph, Google+, knowledge graph), **images, maps, logs** (search logs, server logs, GPS logs), **speech,** …

Needed: A method to the madness



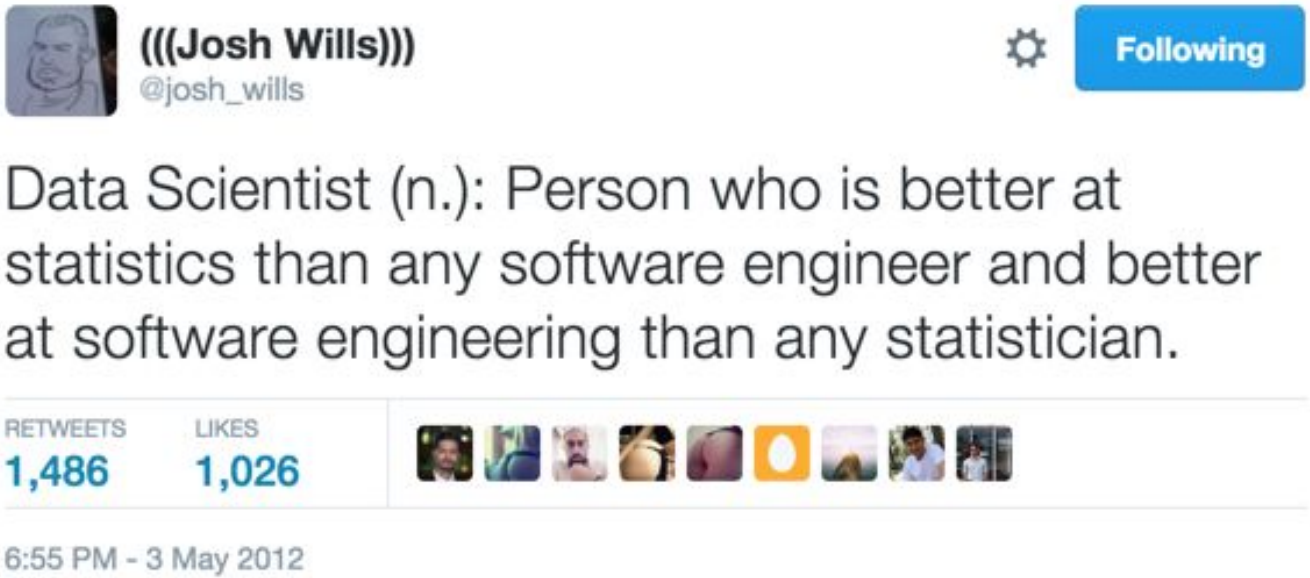
- Scientific method 1.0:
  - Focused on “Model the data”
  - Scientist has hypothesis prior to analyzing the data
- Scientific method 2.0:
  - Systematic cycle (“data science pipeline”)
  - “Explore the data” becomes increasingly important
  - **Data as a first-class citizen**

Scientist 2.0



“A data scientist is someone who can obtain, scrub, explore, model, and interpret data, blending hacking, statistics, and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

Hilary Mason, chief scientist at bit.ly



Josh Wills, Data Scientist at Slack


data
is the new oil
we need to find it,
extract it, refine it,
distribute it and
monetize it.
David Buckingham

More data often beats better algorithms

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Large Language Models in Machine Translation

Thorsten Brants Ashok C. Popat Peng Xu Franz J. Och Jeffrey Dean

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94303, USA
{brants,popat,xp,och,jeff}@google.com

Abstract

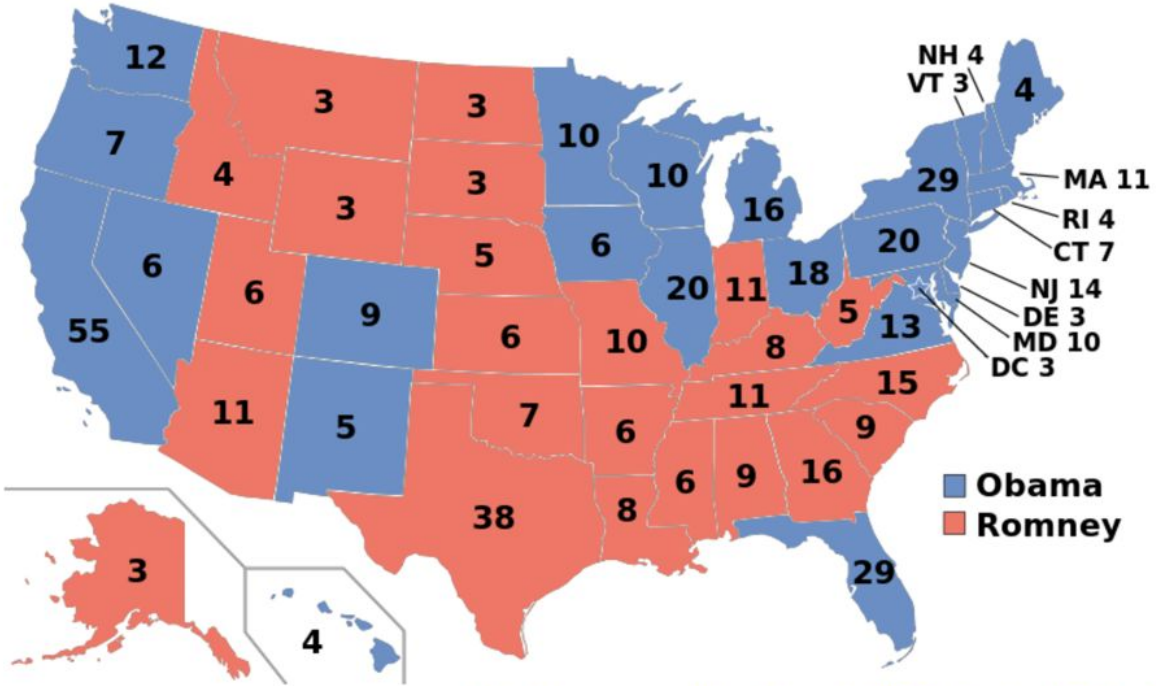
This paper reports on the benefits of large-scale statistical language modeling in machine translation. A distributed infrastructure is proposed which we use to train on up to 2 trillion tokens, resulting in language models having up to 300 billion n-grams. It is capable of providing smoothed probabilities for fast, single-pass decoding. We introduce a new smoothing method, dubbed *Stupid Backoff*, that is inexpensive to train on large data sets and approaches the quality of Kneser-Ney Smoothing as the amount of training data increases.

How might one build a language model that allows scaling to very large amounts of training data? (2) How much does translation performance improve as the size of the language model increases? (3) Is there a point of diminishing returns in performance as a function of language model size?

This paper proposes one possible answer to the first question, explores the second by providing learning curves in the context of a particular statistical machine translation system, and hints that the third may yet be some time in answering. In particular, it proposes a *distributed* language model training and deployment infrastructure, which allows direct and efficient integration into the hypothesis-search algorithm rather than a follow-on re-scoring phase.

2008, 2012: Predicting elections

"Silver, who made his name by using cold hard math to call 49 out of 50 states in the 2008 general election and all 50 in 2012"



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

2016: Predicting elections?

How I Acted Like A Pundit And Screwed Up On Donald Trump

Trump's nomination shows the need for a more rigorous approach.

By Nate Silver
Filed under 2016 Election
Published May 18, 2016

Polls whiz kid Nate Silver and presidential candidate Donald Trump.

Photo illustration by *Slate*. Images by Slaven Vlasic/Getty Images and Ethan Miller/Getty Images.

2008, 2012: Winning elections

"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, CITO Research

2016: Winning elections

2016 CTR CORRECT THE RECORD AMERICAN BRIDGE

Stephen Lam / Reute

A $1 Million Fight Against Hillary Clinton's Online Trolls

A super PAC has a plan to defend the Democratic presidential front-runner and her supporters on social media. Will it work?

Cambridge Analytica

TRUMPING

Donald Trump during the presidential campaign. Image: Gage Skidmore

The Data That Turned the World Upside Down

HG HANNES GRASSEGGER & MIKAEL KROGERUS
Jan 28 2017, 3:15pm

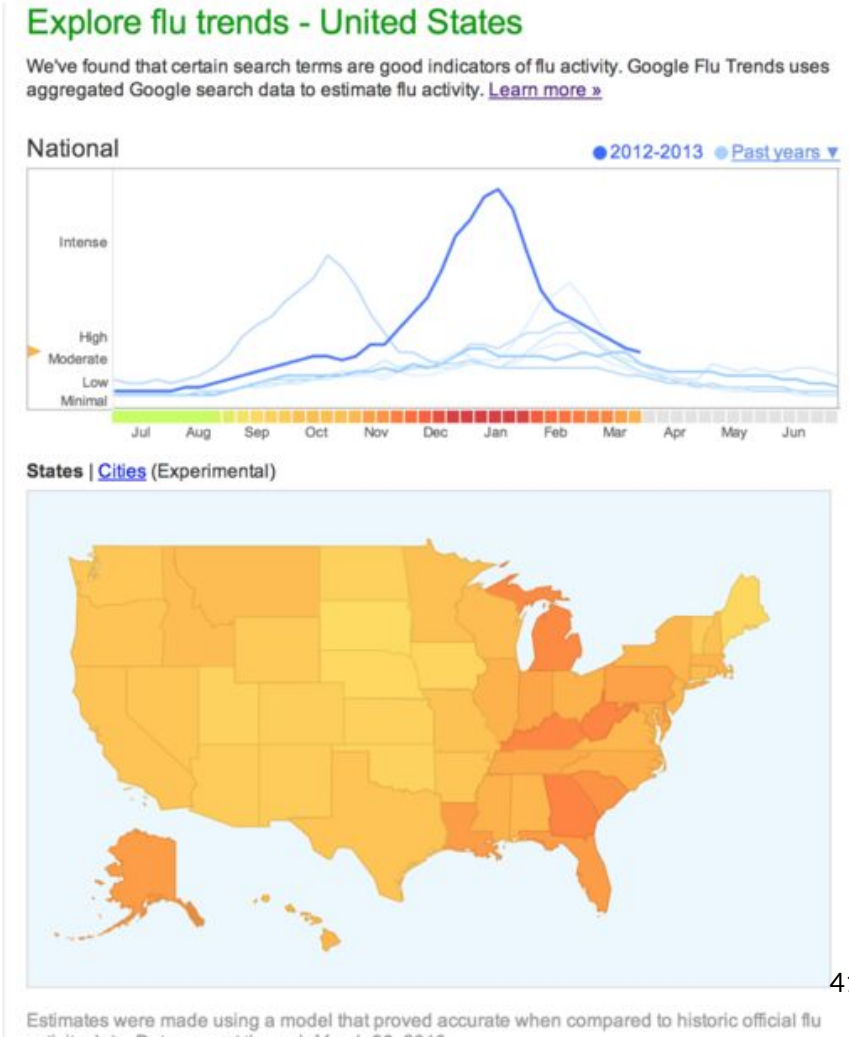
Psychologist Michal Kosinski developed a method to analyze people in minute detail based on their Facebook activity. Did a similar tool help propel Donald Trump to victory? Two reporters from Zurich-based Das Magazin went data-gathering.

Google Flu Trends

Now discontinued…

“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.”



Take care when using oil or data…

Skills we are going to develop...

data munging/scraping/sampling/cleaning in order to get an informative, manageable data set
data storage and management in order to be able to access data quickly and reliably during subsequent analysis
exploratory data analysis to generate hypotheses and intuition about the data
prediction based on statistical tools such as regression, classification, and clustering
communication of results through visualization, stories, and interpretable summaries

… and key principles

- Use many data sources
- Be critical (data is like your friend Steve*)
- Be paranoid (data can be your friend-turned-enemy)
- Understand how the data was collected (recognize biases)
- Use data wisely to remedy biases
- Use statistical models (not just hacking around in Excel)
- Understand correlations (!= causations)
- Communicate your results clearly and carefully

* This statement won’t make sense if you didn’t attend class.

TODO before tomorrow's lab session

- Sign up for Mattermost and install clients (desktop and mobile): https://icmattermost.epfl.ch/cs-401
- Sign up for GitHub
- Register your Mattermost and GitHub accounts with us: https://goo.gl/EfVKhJ (complete form once per team; by next Friday, 9/28)
- Install Anaconda (version: Python 3.6) on your laptop: https://www.anaconda.com/download
- 

Any feedback? -- Let us know!

Give us feedback on this lecture here:
https://go.epfl.ch/ada2018-lec1-feedback

Feedback form available for each lecture and lab session

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more details?
- What would you like the instructor to wear next time?
- …