Applied Data Analysis (CS401)

Lecture 1
Intro to ADA
21 Sep 2022

EPFL

Robert West

Important websites

http://ada.epfl.ch

Your main entry point. All materials linked from there.

https://go.epfl.ch/ada2022-ed

Main communication channel. Sign in with your EPFL email address (or simply access via Moodle).

https://github.com/epfl-ada/2022

Used for homework, project, and final exam.

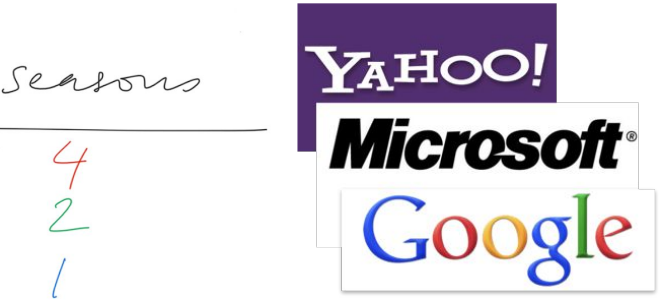
ABOUT ME

- 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

triple

6.10.16
4444 g
555 mm

dlab
011011o

YAHOO!
Microsoft
Google

WIKIMEDIA FOUNDATION

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
555mm

dlab
ol1o1o

triple

⇒ 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

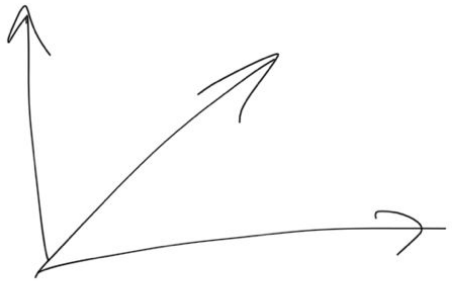
Leverage large datasets

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

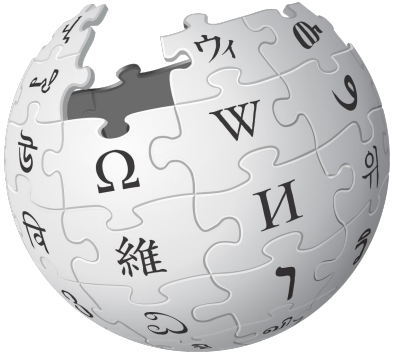
Human computation, crowdsourcing

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.**"



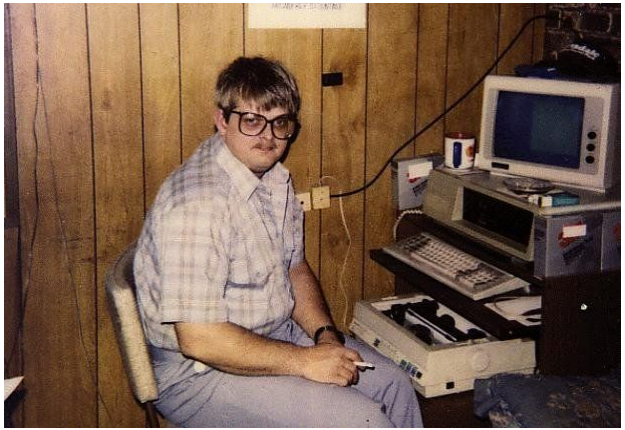
Applied data analysis

- 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 applied to images of cats?"*
- 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




Place
Ada Lovelace

Syllabus

- Handling data
  - “Slicing and dicing”: obtaining, preparing, juggling data
- Visualizing data
  - Exploration of data, communication of results
- Describing data
  - How to support (and be suspicious of) claims about data
- Regression analysis
  - How to disentangle datasets with correlated variables
- Observational studies
  - How to deal with “found data”
  - Correlation != causation

Syllabus (cont'd)

- Machine learning
  - Supervised learning
  - Unsupervised learning
  - Applied aspects of machine learning
- Handling specific types of data
  - Handling text data
  - Handling network data
- Scaling to massive data

Grading

- 30% **Homework assignments (2)**
  - Involving skills required from data scientists
  - Groups of 4 students (may switch groups after Homework 1)
  - Homework of 2017, 2018, 2019, 2020, 2021
- 30% **Final exam** (date TBD)
  - Mini data analysis project
  - Done on laptop, individually, (probably) on campus
  - Final exams of 2017, 2018, 2019, 2020, 2021
- 25% **Project** (more details soon)
  - Your own freestyle data analysis
  - Done in groups of 4 students (same as for homework)
  - Milestones spread throughout the semester
  - Projects of 2017, 2018, 2019, 2020, 2021
- 15% **Quizzes**
  - Weekly, online (on Moodle), 10 questions in 10 minutes

Grading

- 30% **Homework assignments (2)**
  - Involving skills required from data scientists
  - Groups of 4 students (may switch groups after Homework 1)
  - Homework of 2017, 2018, 2019, 2020, 2021
- 30%
  -
  -
  -
- 25%
  -
  - Done in groups of 4 students (same as for homework)
  - Milestones spread throughout the semester
  - Projects of 2017, 2018, 2019, 2020, 2021
- 15% **Quizzes**
  - Weekly, online (on Moodle), 10 questions in 10 minutes

This class will be hard work, but it will get you a job.

Grading (cont'd)

- To obtain a meaningful grade distribution, scaling/shifting will be applied to each of {homework, project, exam, quizzes} before taking weighted average (standard practice at EPFL)
- While intermediate grades are a good indication of where you stand, remember there might be some wiggle
    - → Don't rely on intermediate grades to decide whether you can afford to skip the exam etc.

Meeting logistics: Lectures

- **Wednesdays 8:15 - 10:00**
- If you want to see it live, come to class! (No live streaming)
- Lectures are also recorded and made available after class

Meeting logistics: Lab sessions

- **Fridays 13:15 - 15:00**
- In person only: BCH 2201 (next to UNIL)
- You solve exercises that we make available the day before, can ask questions and get help from assistants
- Can connect with assistants in the spirit of office hours
- Sometimes brief tutorials
- Possibly invited speakers from industry and academia
- Weekly online quizzes on Moodle (see next slide)

Weekly quizzes

- Held online on Moodle
- During first 15 minutes (13:15-13:30) of each Fri lab session
- Week 1 (this week): no quiz
- Week 2 (next week): Quiz 1: no real questions, just to let you get familiar with the setup
- Week 3: Quiz 2: the first quiz that counts
- Quiz in week $i + 1$ is about lecture material of week i
- Goal: to encourage you to continuously engage with the course material
- Your two lowest-scoring quizzes from entire semester won't count

Project

- We'll provide a number of datasets
- You need to form and pitch a crisp project idea
- Free to combine with other datasets (at your own risk)
- Goal: not a loose collection of results -- tell a story with the data!
  - Data stories of 2017, 2018, 2019, 2020, 2021
  - Nice example data story

Homework and projects: GitHub

- De-facto standard for managing and sharing code
- All students in this class need a GitHub account
- Homework and project submissions done via GitHub

Main communication channel:

- Class forum, available via Moodle
- Also accessible directly, outside of Moodle:
  https://go.epfl.ch/ada2022-ed
  (sign in using the same email address as for Moodle)
- Central place to ask all class-related questions
- Mandatory! We'll send important announcements on Ed only
- Help each other (without cheating, of course)

General note on communication

- Multiple platforms used in ADA for various tasks (as in real life): Ed, GitHub, Google docs, ADA website
- To avoid confusion,
  - familiarize yourself with communication guidelines
  - all materials will be linked from the website as a central point of entry: https://ada.epfl.ch
  - all discussions will take place on Ed

Group registration

- Must form teams within 2 weeks, starting now (in time for release of Homework 1)!
- Get started immediately to find 3 teammates
- By Fri 7 Oct 23:59, complete the registration form (to be done by each team member individually): https://forms.gle/tgCU1yPKvVmSWuNE9
- Can change team after Homework 1 (but try to avoid it)

Prerequisites

Basics of

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

Python environments

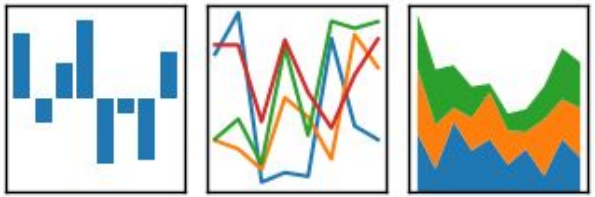
- Homeworks and exams to be done as Jupyter Notebooks
- You will submit a pre-executed .ipynb file
  - We don't care how you produce it
  - Option 1: local Python installation (e.g., Anaconda + JupyterLab)
  - Option 2: Google Colab = notebook hosted by Google
  - Option 3: noto = notebook hosted by EPFL
- To get started: come to Friday's lab session ("Tutorial 0")
- "Homework 0": do it yourself at home after lab session (optional, not graded)
- Doing Homework 0 is the best way of making sure you're set up correctly for later homework, project, exam

Python++

machine learning in Python

pandas

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



POLLING TIME

- **“What is your prior experience with Python?”**
- Scan QR code or go to https://web.speakup.info/room/join/66626

Instructor

Bob West

Teaching assistants (TAs) = PhD students

Akhil Arora

Manoel Horta

Lars Klein

Aryo Lotfi

Silin Gao

Valentin Hartmann

Halima Schede

Chenyang Wang

Marija Sakota

Bhargav Srinivasa Desikan

Martin Josifoski

Student assistants (SAs) = MS students

Roberto Ceraolo

Eugénie Chabenat

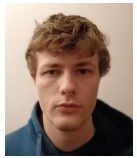
Edvin Maid

Marie Biolková

Jozef Coldenhoff

Francesco Salvi

Giacomo Orsi



Erwan Serandour

Mete Ismayil

Miloš Vujasinović

Hugo Casademont

Ana-Arina Raileanu

Nicolas Baldwin

Margaux L'Eplattenier

- Help each other on Ed
- Participate actively in classes and labs
- Give us **feedback**

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2022-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?
- …

Questions?

What is data science?

MENU

Harvard Business Review

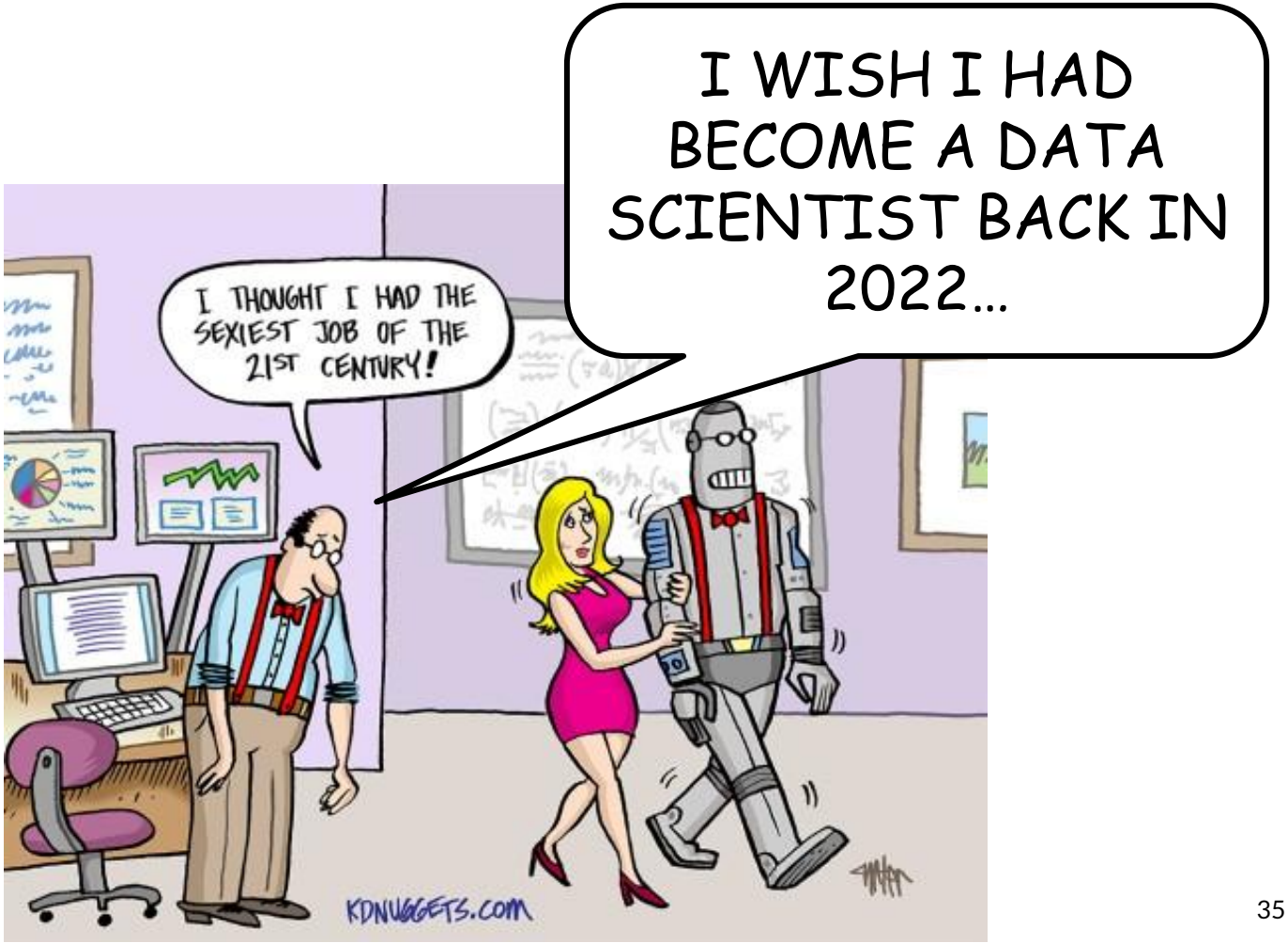
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Why now?


I WISH I HAD BECOME A DATA SCIENTIST BACK IN 2022...
I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!
KDNUGGETS.COM

“Data science”

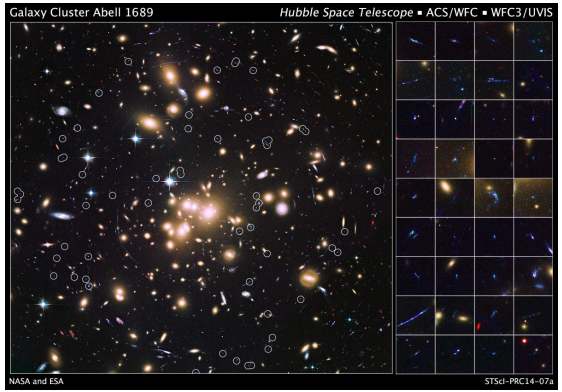
- All science is (or should be) based on data, per definitionem
- So how is “data science” different from plain old “science”?

Data volume explodes

"Between the dawn of civilization and 2003, we only created **five exabytes** of information; now [in 2010] we're creating that amount **every two days.**"

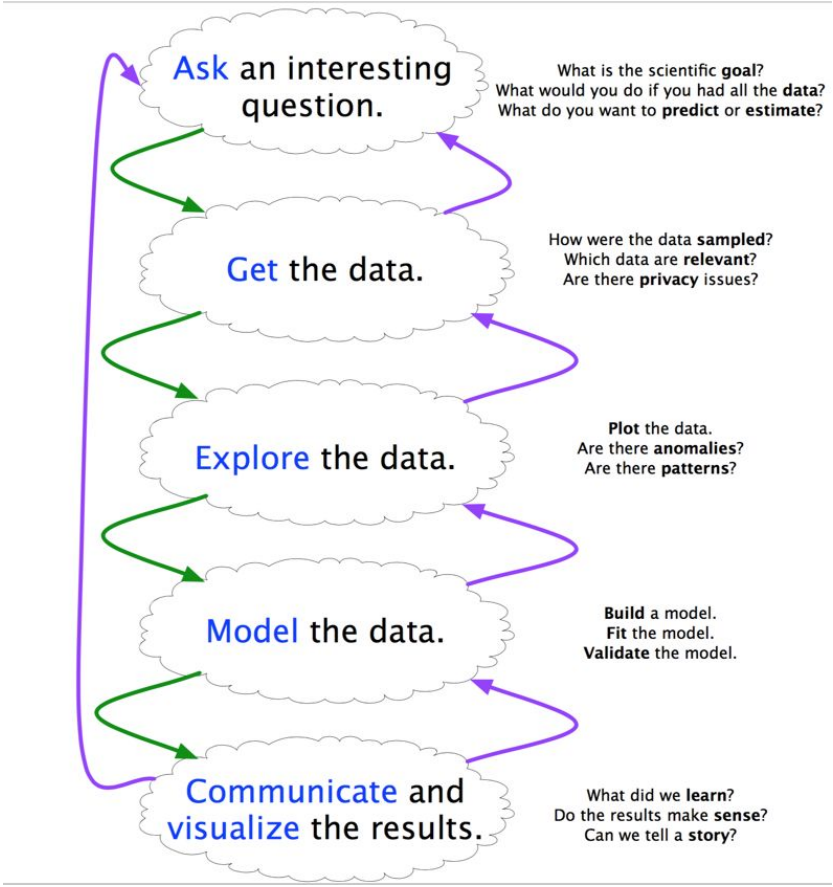
Eric Schmidt, Google (2010)

Data variety explodes



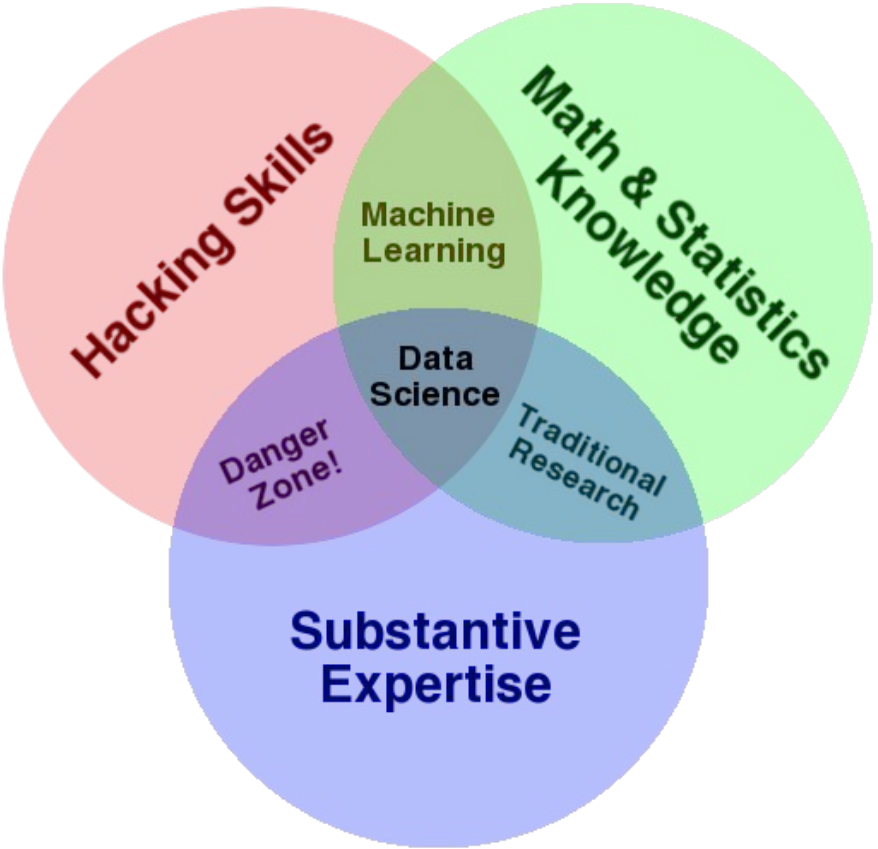
Text (indexed Web pages, email), **networks** (Web graph, Google+, knowledge graph), **images, maps, logs** (search logs, server logs, GPS logs), **speech,** …

Needed: A method to the madness



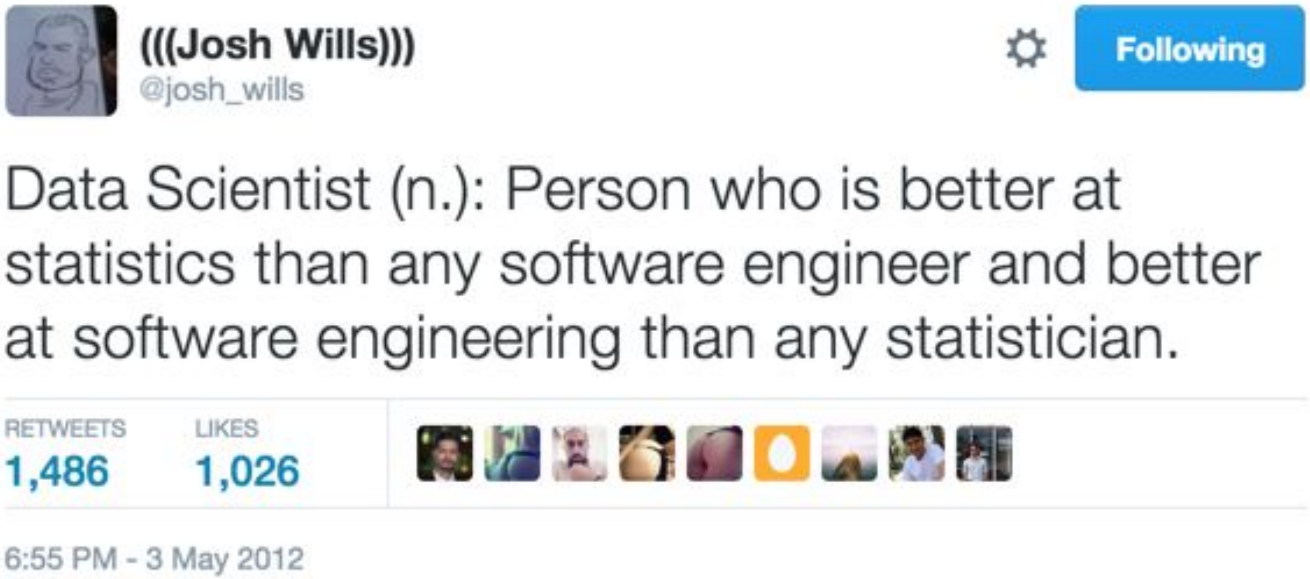
- Scientific method 1.0:
  - Focused on “Model the data”
  - Scientist has hypothesis prior to analyzing the data
- Scientific method 2.0:
  - Systematic cycle (see diagram)
  - “Explore the data” becomes increasingly important
  - **Data as a first-class citizen**

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

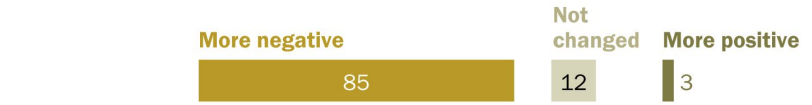
The Unreasonable Effectiveness of Data

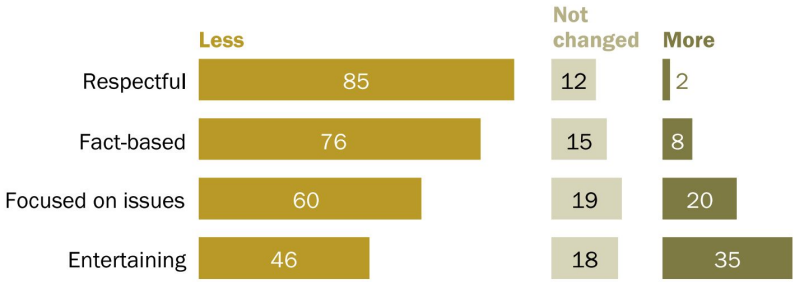
Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

21st-century politics

Most Americans say political debate in the U.S. has become less respectful, fact-based, substantive

% who say over the last several years the tone and nature of political debate in this country has become ...





% who say Donald Trump has changed the tone and nature of political debate in the U.S. ...

Note: No answer responses not shown.
Source: Survey of U.S. adults conducted April 29-May 13, 2019.

PEW RESEARCH CENTER

We ask: Do these subjective impressions reflect the true state of US political discourse?

ADA will teach you the tools to answer such questions using data (see next slides)

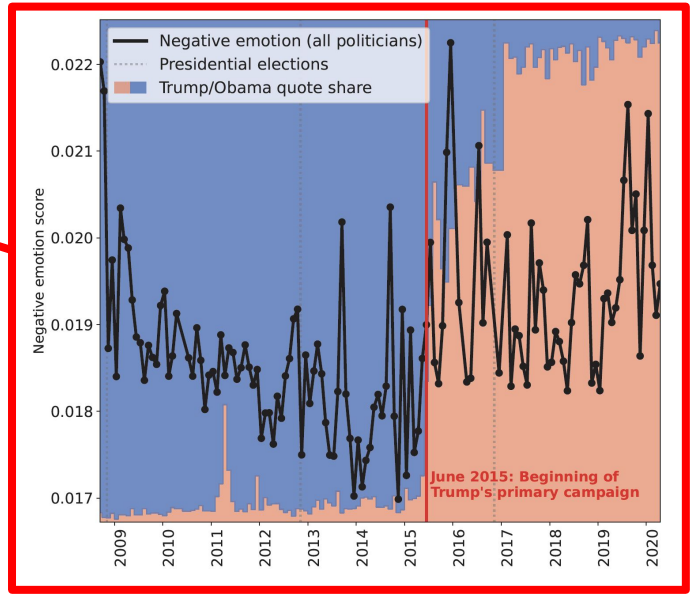
https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/06/PP_2019.06.19_Political-Discourse_FINAL.pdf

Syllabus, revisited

- **Handling data**
- Visualizing data
- Describing data
- Regression analysis
- Observational studies
- Machine learning
- Handling text data
- Handling network data
- Scaling to massive data

Syllabus, revisited

- Handling data
- **Visualizing data**
- Describing data
- Regression analysis
- Observational studies
- Machine learning
- Handling text data
- Handling network data
- Scaling to massive data



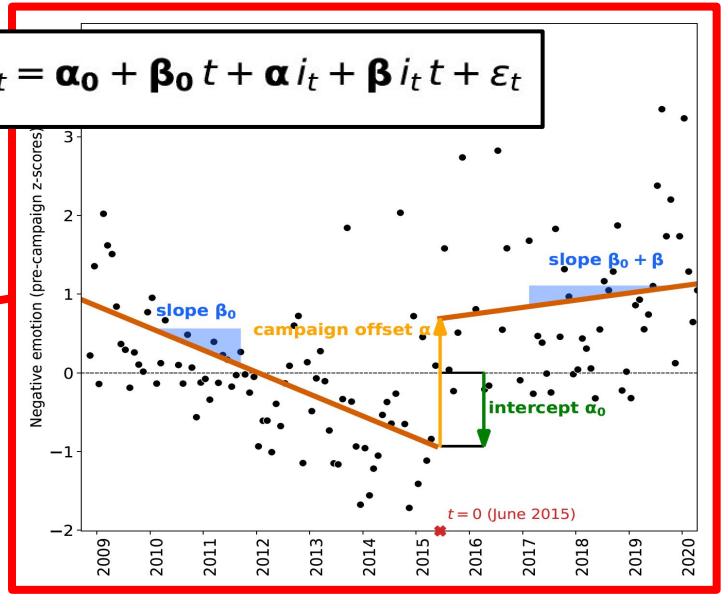
Syllabus, revisited

- Handling data
- Visualizing data
- **Describing data**
- Regression analysis
- Observational studies
- Machine learning
- Handling text data
- Handling network data
- Scaling to massive data

"Is the effect real, or could it have been produced by chance?"

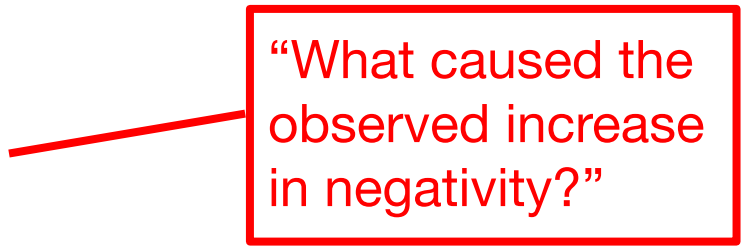
Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- **Regression analysis**
- Observational studies
- Machine learning
- Handling text data
- Handling network data
- Scaling to massive data



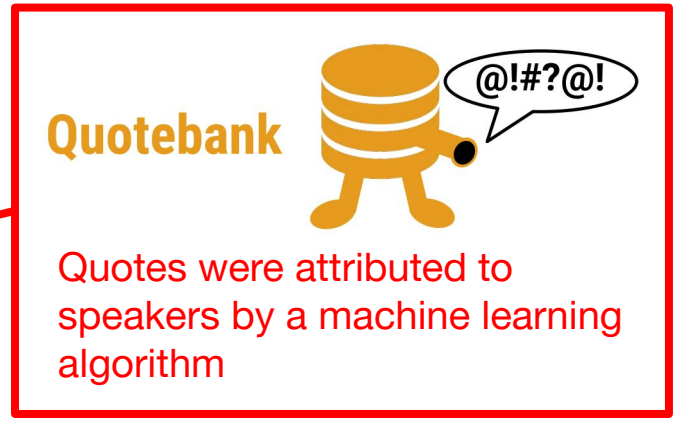
Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis
- **Observational studies**
- Machine learning
- Handling text data
- Handling network data
- Scaling to massive data



Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis
- Observational studies
- **Machine learning**
- Handling text data
- Handling network data
- Scaling to massive data



Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis
- Observational studies
- Machine learning
- **Handling text data**
- Handling network data
- Scaling to massive data

Research question (“Did political discourse become more negative?”) is a question about language == text

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis
- Observational studies
- Machine learning
- Handling text data
- **Handling network data**
- Scaling to massive data

In follow-up work we ask: “Who speaks about whom in what way?” → Construct “who-mentions-whom” network

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis
- Observational studies
- Machine learning
- Handling text data
- Handling network data
- **Scaling to massive data**

Curious to learn more?

Full paper available at https://arxiv.org/abs/2207.08112

United States Politicians' Tone Became More Negative with 2016 Primary Campaigns

Jonathan Külz[1], Andreas Spitz[2], Ahmad Abu-Akel[3], Stephan Günnemann[1], Robert West[4*]

[1]Department of Informatics, Technical University of Munich, Germany

[2]Department of Computer and Information Science, University of Konstanz, Germany

[3]School of Psychological Sciences, University of Haifa, Israel

[4]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Abstract

There is a widespread belief that the tone of US political language has become more negative recently, in particular when Donald Trump entered politics. At the same time, there is disagreement as to whether Trump changed or merely continued previous trends. To date, data-driven evidence regarding these questions is scarce, partly due to the difficulty of obtaining a comprehensive, longitudinal record of

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 Friday's lab session

- Sign up for Ed here and familiarize yourself with it
- If you're not on GitHub yet, sign up for GitHub
- Start looking for 3 teammates
  - You may use "Group formation" category on Ed
- Check out Google Colab and noto (to see if you want to use either of them)
- Check out Tutorial 0 here (in prep for Fri lab session)
- 

Any feedback? -- Let us know!

Give us feedback on this lecture here:
https://go.epfl.ch/ada2022-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?
- …