

Applied Data Analysis (CS401)



Lecture 1
Intro to ADA
20 Sep 2023

EPFL

Robert West



Important websites



<http://ada.epfl.ch>

Your main entry point. All materials linked from there.



<https://go.epfl.ch/ada2023-ed>

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



<https://github.com/epfl-ada/2023>

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

dlab
0110110



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

dlab
0110110

dlab

ada

⇒ Call me Bob

bob
10010

My path toward data science

- Born in Ingolstadt, Bavaria, Germany



MY RESEARCH

Develop
neat
algorithms

Give back to
the real world
(\Rightarrow applications)

Address
real-
world
issues



Distill raw data
into insights into
people & the world

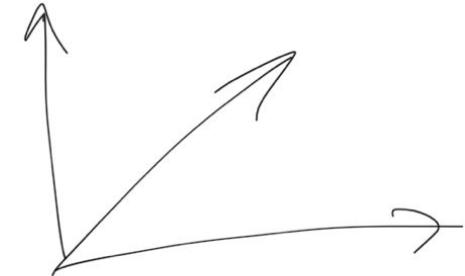
Leverage
large datasets



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

When necessary, generate
my own data (human comput-
ation, 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 abbreviate this course as **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 for disentangling data
 - How to disentangle datasets with correlated variables
- Causal analysis of observational data
 - How to deal with “found data”
 - Correlation != causation

Syllabus (cont'd)

- Learning from data
 - 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 5 students (may switch groups after Homework 1)
 - Homework of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 30% **Final exam** (date TBD)
 - Mini data analysis project
 - Done on laptop, individually, on campus
 - Final exams of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 25% **Project** (more details soon)
 - Your own freestyle data analysis
 - Done in groups of 5 students (same as for homework)
 - Milestones spread throughout the semester
 - Projects of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 15% **Quizzes**
 - Weekly, online (on Moodle), 5 questions in 10 minutes



Grading

- 30% **Homework assignments (2)**
 - Involving skills required from data scientists
 - Groups of 5 students (may switch groups after Homework 1)
 - Homework of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 30% **Projects (2)**
 - Milestones spread throughout the semester
 - Projects of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 25% **Final Project (1)**
 - Done in groups of 5 students (same as for homework)
 - Milestones spread throughout the semester
 - Projects of [2017](#), [2018](#), [2019](#), [2020](#), [2021](#), [2022](#)
- 15% **Quizzes**
 - Weekly, online (on Moodle), 5 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 13:15–14:45**
- In person only:
 - BCH 2201 (next to UNIL)
 - CE 1 106
- Labs are complementary to lectures, not simply more detail on same
- You solve exercises that we make available the day before, can ask questions and get help from assistants
- In certain weeks: homework/project office hours (probably on Zoom, in parallel to exercises)
- Weekly online quizzes on Moodle (see next slide)

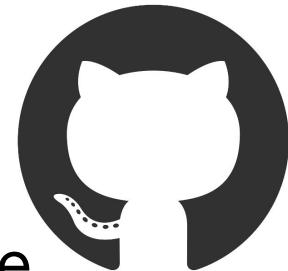
Weekly quizzes

- Held online on Moodle
- 5 questions, to be answered within 10 minutes of starting
- 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 i (held in week $i + 1$) is about lecture material of week i
- Goal:
 - Engage continuously with course material
 - Think (not just find right slide)
- No one is expected to get everything right
- 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](#), [2022](#)
 - 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/ada2023-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)

Watch-at-home videos

- Throughout the semester, we'll release videos with supplemental information; e.g.,
 - Intro to lab sessions ([already available!](#))
 - Project instructions
 - Homework 1 postmortem
 - Homework 2 postmortem

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 4 teammates
- By Fri 6 Oct 23:59, complete the registration form (to be done by each team member individually):
<https://forms.gle/kEX3asGKPKRFBhN2A>
- 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

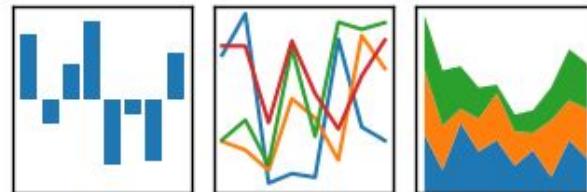
- Homeworks and exams to be done as [Jupyter Notebooks](#)
- You will submit a pre-executed .ipynb file
 - We don't care how you produce it
 - Option 1: local Python installation (e.g., [Anaconda](#) + [JupyterLab](#))
 - Option 2: [Google Colab](#) = notebook hosted by Google
 - Option 3: [noto](#) = notebook hosted by EPFL
- To get started: come to Friday's lab session ("[Exercise 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++



pandas

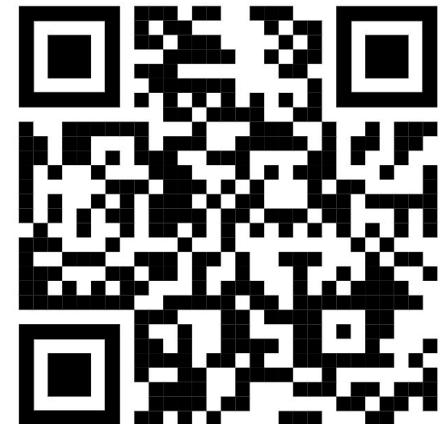
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$





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



Student assistants (SAs) = MS students





WE WANT YOU!

- 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/ada2023-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?

Commercial break

ADA students:
sharp like
teeth!



≡ MENU

Harvard
Business
Review

What is data science?



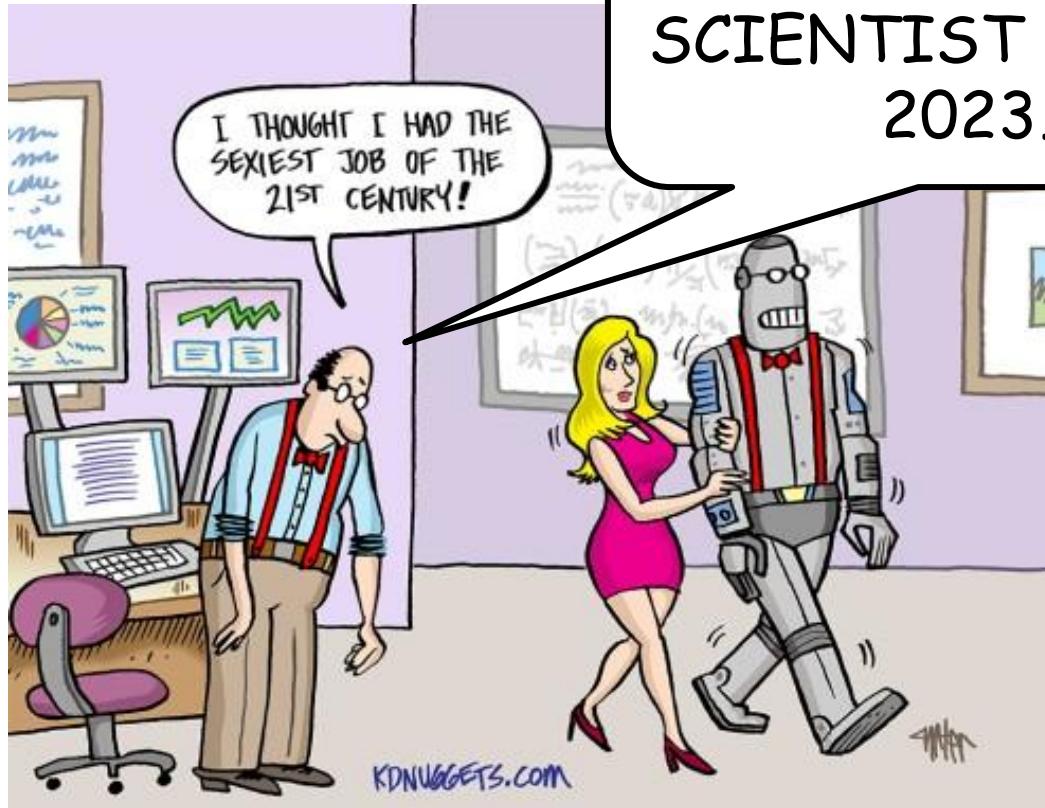
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Why now?



“Data science”

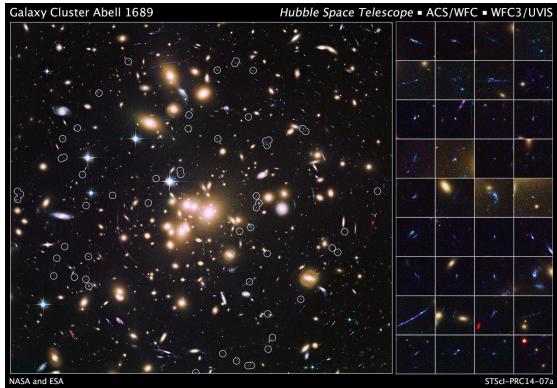
- All science is (or should be) based on data, per definitionem
- So how is “data science” different from plain old “science”?

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.**”

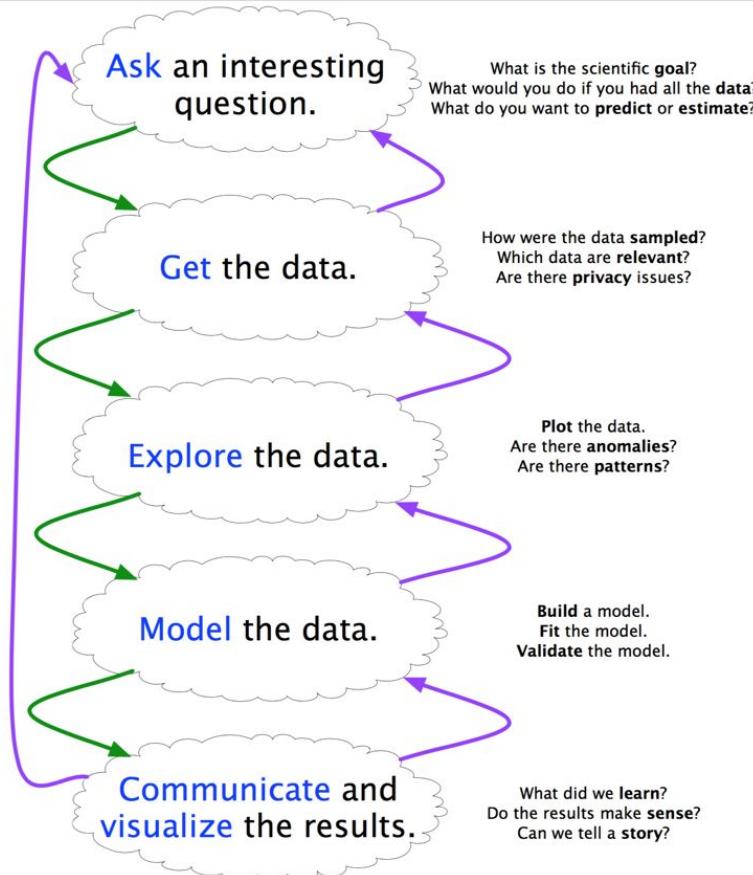
Eric Schmidt, Google (2010)

Data variety explodes



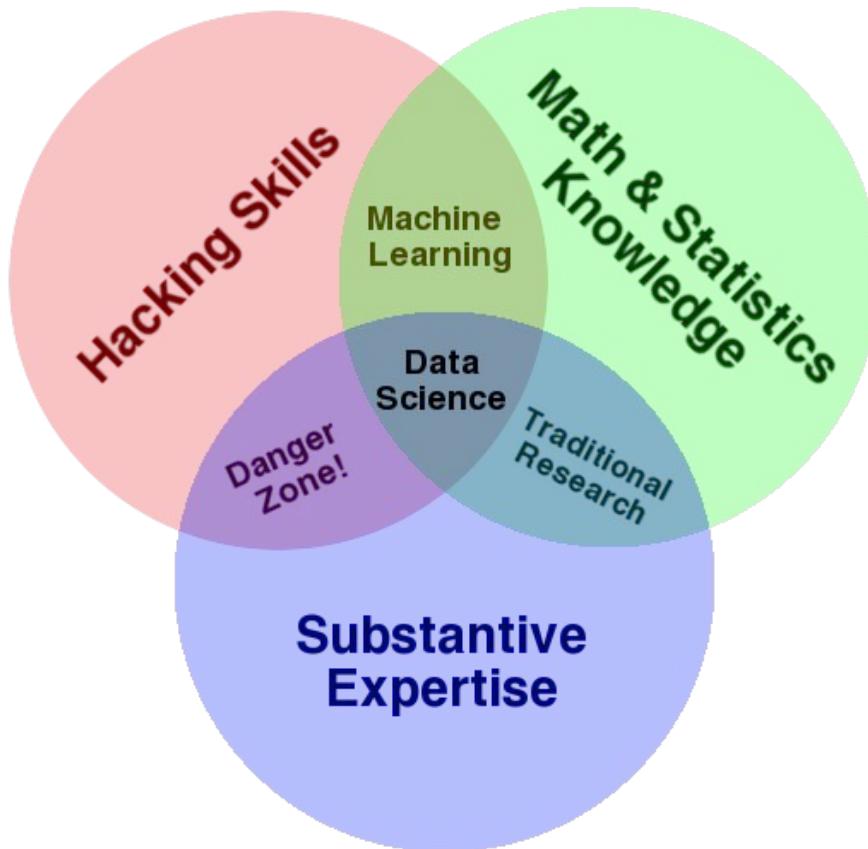
Text (indexed Web pages, email),
networks (Web graph, Google+, knowledge graph), **images, maps, logs** (search logs, server logs, GPS logs), **speech**, ...

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))

@josh_wills



Following

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS

1,486

LIKES

1,026



6:55 PM - 3 May 2012

Josh Wills, Data Scientist at Slack

A dark silhouette of an oil pump jack against a blue sky.

data oil
is the new

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

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

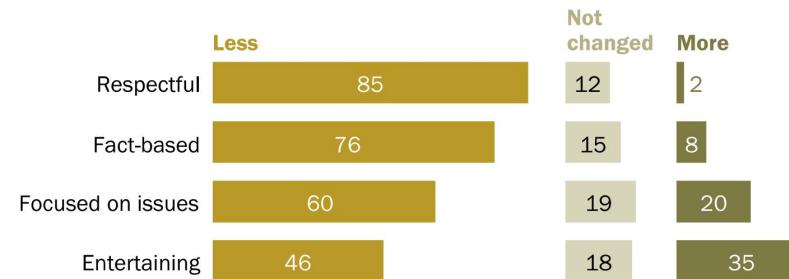
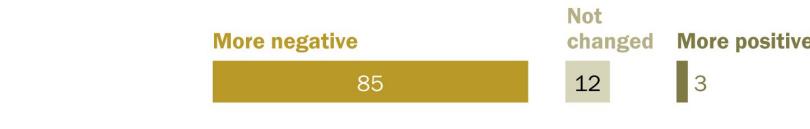
<http://www.incompleteideas.net/Incldeas/BitterLesson.html>

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)

Syllabus, revisited

- **Handling data**
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

Quotebank

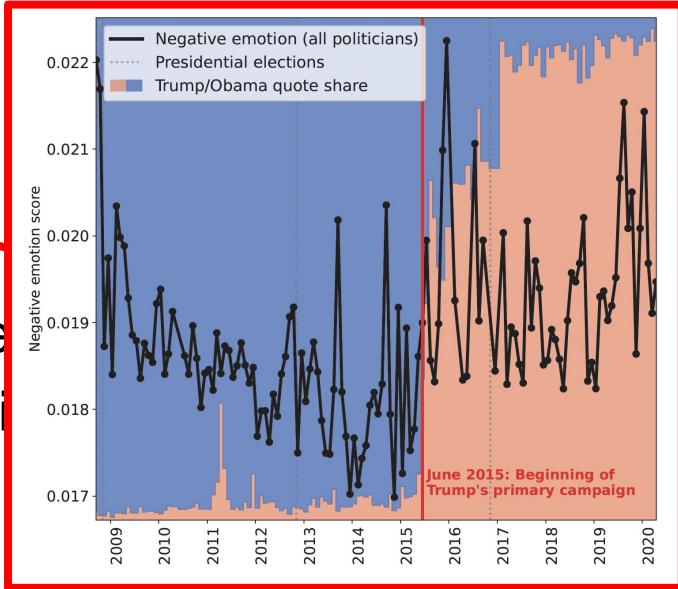


Data: <https://github.com/epfl-dlab/Quotebank>

Web interface: <https://quotebank.dlab.tools/>

Syllabus, revisited

- Handling data
- **Visualizing data**
- Describing data
- Regression analysis for disease
- Causal analysis of observational studies
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data



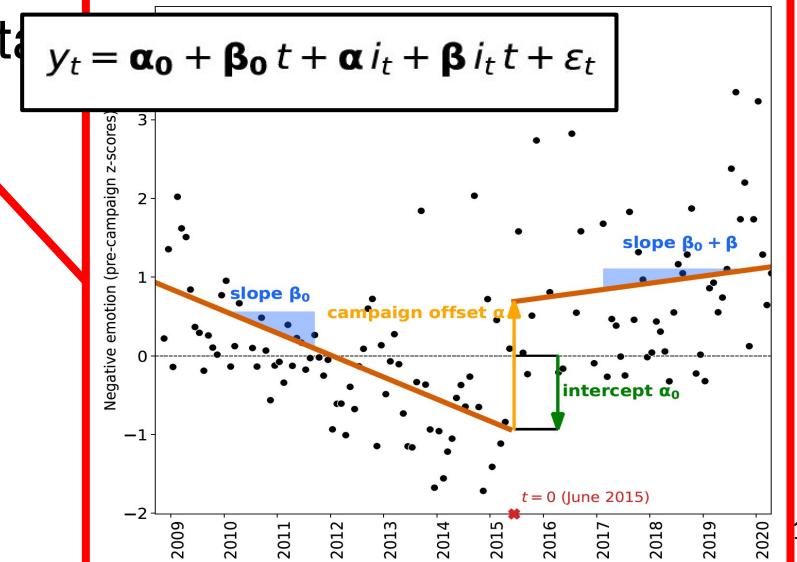
Syllabus, revisited

- Handling data
- Visualizing data
- **Describing data**
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- 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 for disentangling data**
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data



Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- **Causal analysis of observational data**
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data

“What caused the observed increase in negativity?”

Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- **Learning from data**
- Handling text data
- Handling network data
- Scaling to massive data



Syllabus, revisited

- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- **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

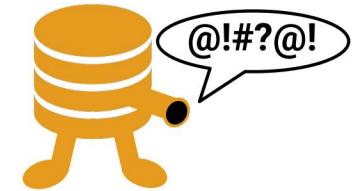
- Handling data
- Visualizing data
- Describing data
- Regression analysis for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- **Handling network data**
- Scaling to massive data

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 for disentangling data
- Causal analysis of observational data
- Learning from data
- Handling text data
- Handling network data
- **Scaling to massive data**

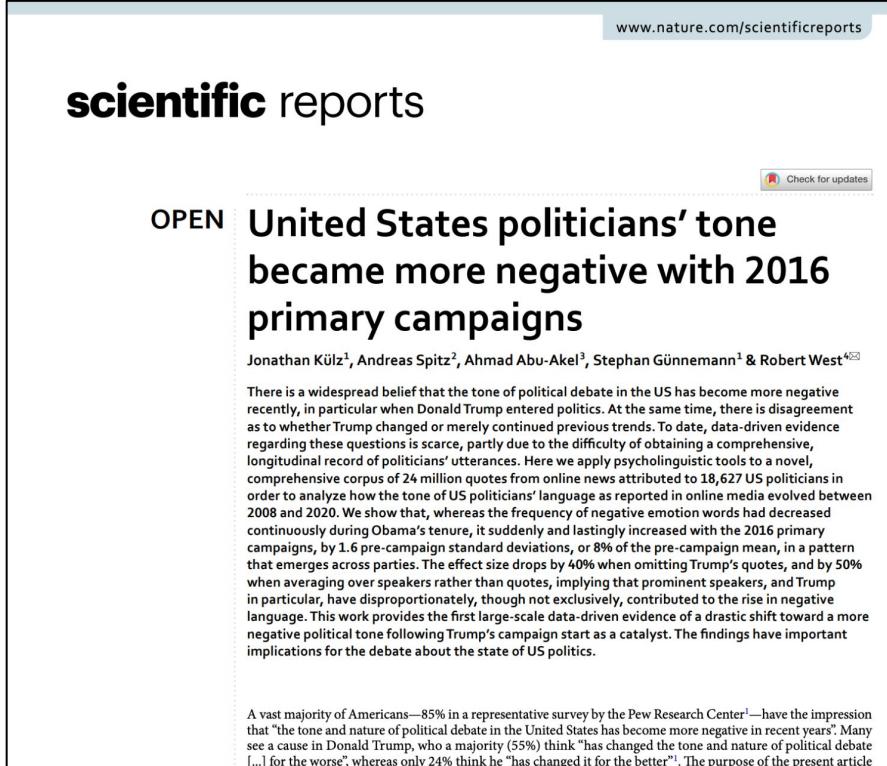
Quotebank



Extracting 235 million quotes
from 127 million news articles
(about 1 terabyte of raw text)
spanning 12 years requires
big-data tools (e.g., Spark)

Curious to learn more?

Full paper available at <https://www.nature.com/articles/s41598-023-36839-1>



The image shows a screenshot of a scientific article from the journal "scientific reports". The article is titled "United States politicians' tone became more negative with 2016 primary campaigns". It is marked as "OPEN" and includes author names: Jonathan Külz¹, Andreas Spitz², Ahmad Abu-Akel³, Stephan Günemann¹ & Robert West^{4,5}. A brief abstract discusses the shift in political tone from 2008 to 2020, noting a significant increase in negative language during the 2016 primaries. The URL www.nature.com/scientificreports/ is visible at the top right of the page.

www.nature.com/scientificreports/

scientific reports

Check for updates

OPEN

United States politicians' tone became more negative with 2016 primary campaigns

Jonathan Külz¹, Andreas Spitz², Ahmad Abu-Akel³, Stephan Günemann¹ & Robert West^{4,5}

There is a widespread belief that the tone of political debate in the US 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 politicians' utterances. Here we apply psycholinguistic tools to a novel, comprehensive corpus of 24 million quotes from online news attributed to 18,627 US politicians in order to analyze how the tone of US politicians' language as reported in online media evolved between 2008 and 2020. We show that, whereas the frequency of negative emotion words had decreased continuously during Obama's tenure, it suddenly and lastingly increased with the 2016 primary campaigns, by 1.6 pre-campaign standard deviations, or 8% of the pre-campaign mean, in a pattern that emerges across parties. The effect size drops by 40% when omitting Trump's quotes, and by 50% when averaging over speakers rather than quotes, implying that prominent speakers, and Trump in particular, have disproportionately, though not exclusively, contributed to the rise in negative language. This work provides the first large-scale data-driven evidence of a drastic shift toward a more negative political tone following Trump's campaign start as a catalyst. The findings have important implications for the debate about the state of US politics.

A vast majority of Americans—85% in a representative survey by the Pew Research Center¹—have the impression that “the tone and nature of political debate in the United States has become more negative in recent years”. Many see a cause in Donald Trump, who a majority (55%) think “has changed the tone and nature of political debate [...] for the worse”, whereas only 24% think he “has changed it for the better”¹. The purpose of the present article

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 4 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 Exercise 0 [here](#) (in prep for Fri lab session)
- 

Any feedback? -- Let us know!

Give us feedback on this lecture here:

<https://go.epfl.ch/ada2023-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?
- ...