Applied Data Analysis (CS401)

Lecture 1
Intro to ADA
10 Sep 2025

EPFL

Maria Brbić

Important websites

http://ada.epfl.ch

Your main entry point. All materials linked from there.

https://edstem.org/eu/courses/2502/discussion

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

https://github.com/epfl-ada/2025

Used for exercises, homework, project, and final exam.

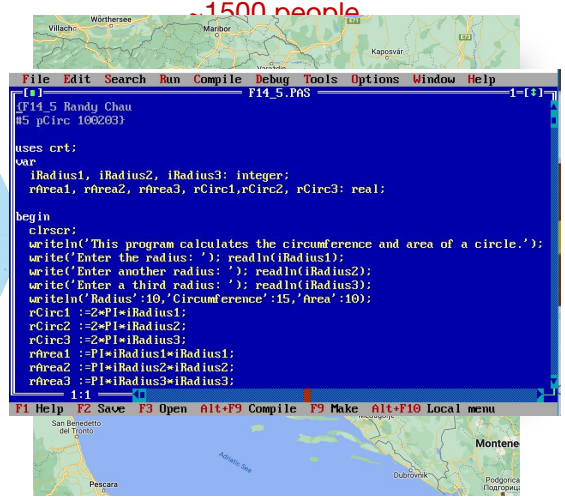
Credits

- Robert West
- Now your ML instructor

About your instructor

~ 3.850.000 people

~1500 people

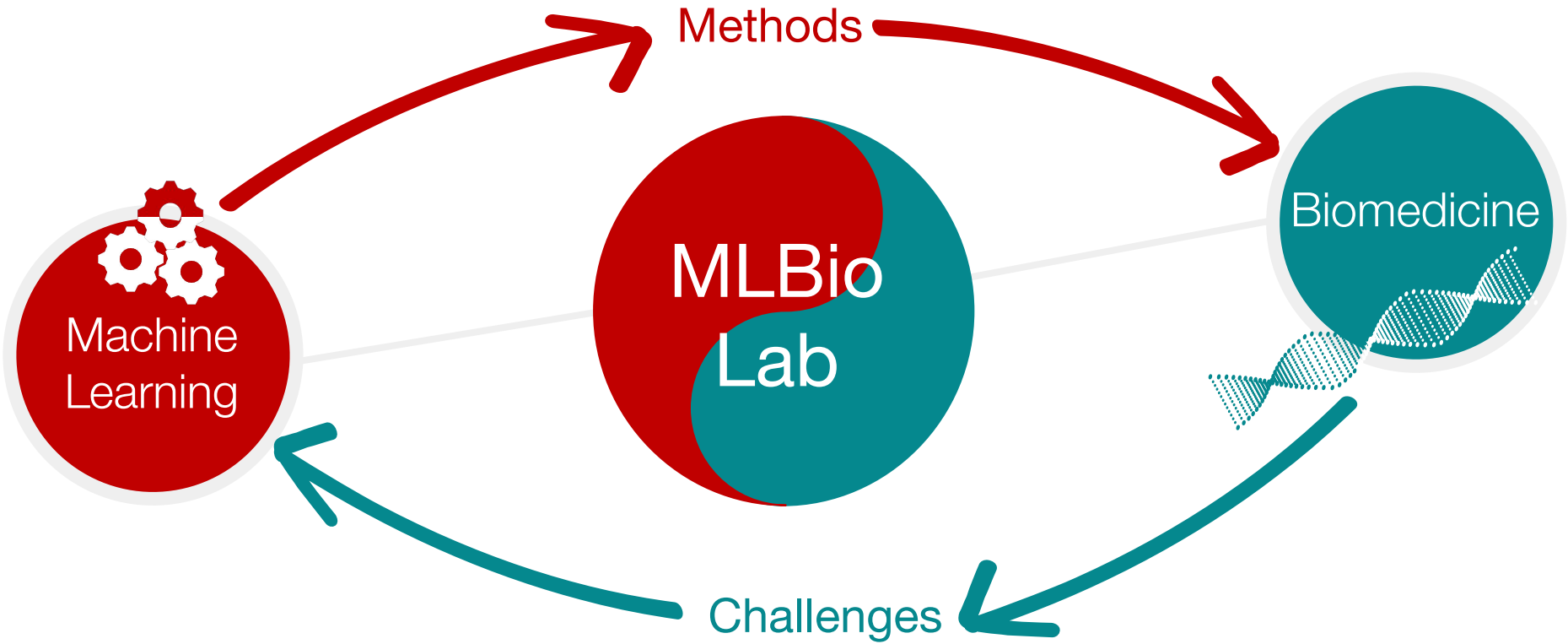


- Born in Tučepi, Croatia

- Education:
  University of Zagreb, Croatia
  Stanford University, USA

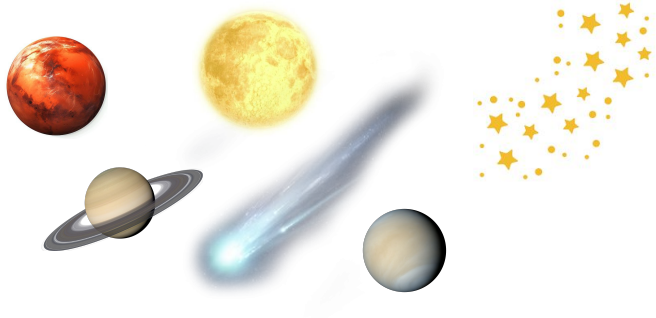
- Assistant Professor at EPFL since Sep ’22
  Machine Learning for Biomedicine (MLBio) lab

Our research @ MLBio

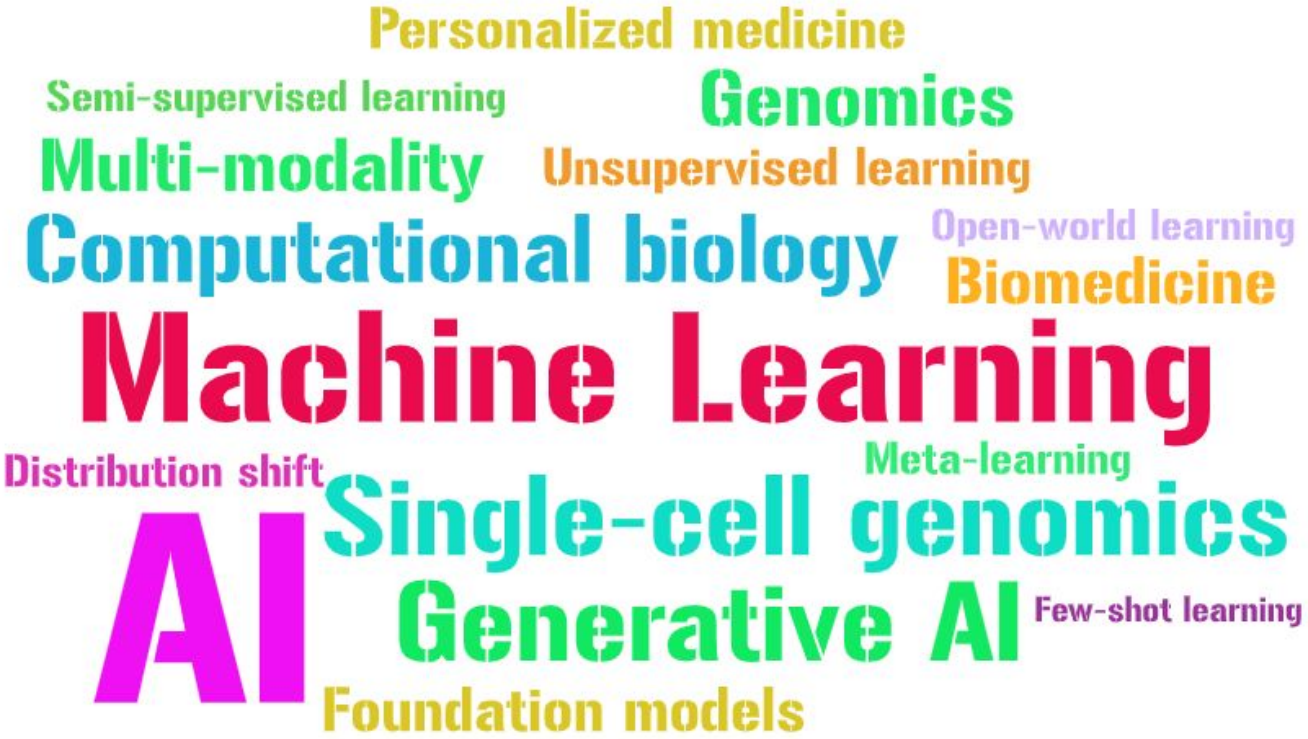


Our research @ MLBio

- Develop new AI methods
  - Generative AI, foundation models, multi-modal models
- Collaborate with biologists and medical researchers and have new “untouched datasets” or collect datasets
- Gain new insights from these datasets what are interesting questions that need new AI algorithms to be answered?
- Apply AI algorithms we develop to advance biomedical research and drive new discoveries in biology and medicine



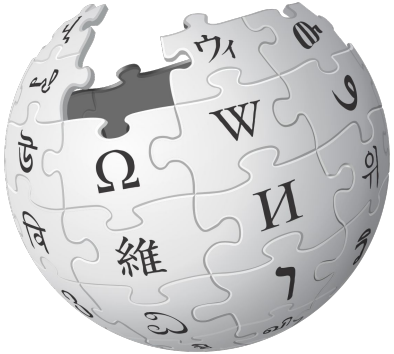
Our research



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

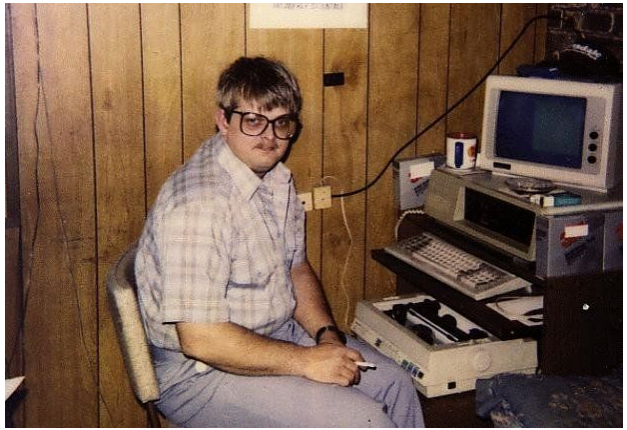
Goal of this class: Enable you to conduct a full-fledged data science project from start to finish

That being 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

- **15% Homework assignment**
  - Involving skills required from data scientists
  - Groups of 5 students
  - Homework of 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024
- **50% Final exam**
  - First part: Multiple-choice questions
  - Second part: Mini data analysis project
    - Done on laptop, individually, on campus
  - Final exams of 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024
- **35% 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, 2023, 2024

Grading

- **15% Homework assignment**
  - Involving skills required from data scientists
  - Groups of 5 students
  - Homework of 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024
- **50%**
  -
  -
  -
  -
- **35% 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, 2023, 2024

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} 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

Deadlines

- **Homeworks**
  - **Homework**
    - Release Nov 5th 2025
    - Due Nov 26th 2025
- **Final exam**
  - Date TBD
- **Project deliverables**
  - **Project milestone P1**
    - Due Oct 1st 2025
  - **Project milestone P2**
    - Due Nov 5th 2025
  - **Project milestone P3**
    - Due Dec 17th 2025

All deadlines are 23:59 CET

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 3:15–4:45**
- In person only:
  - GCC 330
  - CE 14
  - BCH 2201
- 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 quizzes

- Available online on Moodle after every lecture
- 5 questions, to be answered within 10 minutes of starting
- Quiz 2: the first quiz with lecture material questions
- Quiz *i* is about lecture material of week *i*
- Goal:
  - Engage continuously with course material
  - Think (not just find right slide)
- Not graded, for you to recap lecture materials

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, 2023, 2024
  - 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
- ADA Github repository:

https://github.com/epfl-ada/2025

Main communication channel:

- Class forum, available via Moodle
- Also accessible directly, outside of Moodle: https://edstem.org/eu/courses/2502/discussion (sign in using the same email address as for Moodle)
- Central place to ask all class-related questions
- Don’t send us emails
- 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 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

Commercial break

ADA students: sharp like teeth!

Group registration

- Must form teams within 2 weeks, starting now
- Get started immediately to find 4 other teammates
- By **Fri Sep 26th 23:59**, complete the registration form (to be done by each team member individually): https://go.epfl.ch/ada2025-team-registration

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++

jupyter

scikits learn

machine learning in Python

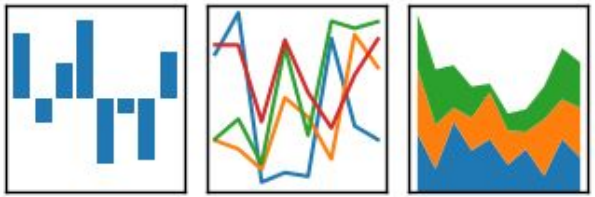
Spark

Lightning-Fast Cluster Computing

seaborn

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://app.sli.do/event/f5usXPBvsT6GLWi5yLuAd6

Instructor

Maria Brbić

Head TAs

Shuo Wen

Siba Panigrahi

TAs: Teaching assistants

Vinko Sabolcec

Mete Ismayilzada

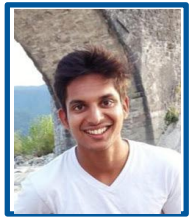
Tim Davidson

Aoxiang Fan

Shuqi Wang

Sevda Ogut

Yulun Jiang



Savyaraj Deshmukh

Alba Carballo Castro

Artyom Gadetsky

Yist Yu

Schuyler James Stoller

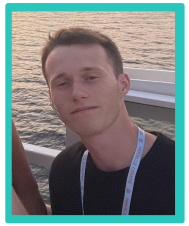
Pierre Beck

Paula Sanchez Lopez

SAs: Student assistants (Master students)

Abdul Karim Mouakeh

Zahra Taghizadeh



Lysandre Costes

Marija Zelic

Alexander Procelewski

Jean Siffert

Alessandro Di Maria

Nastaran Hashemisanjani

Sara Zatezalo

Kyuhee Kim

Amene Gafsi

William Jallot

Yassine Mustapha Wahidy

- 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/ada2025-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?
- …

Questions? 

What is data science?

MENU

Harvard Business Review

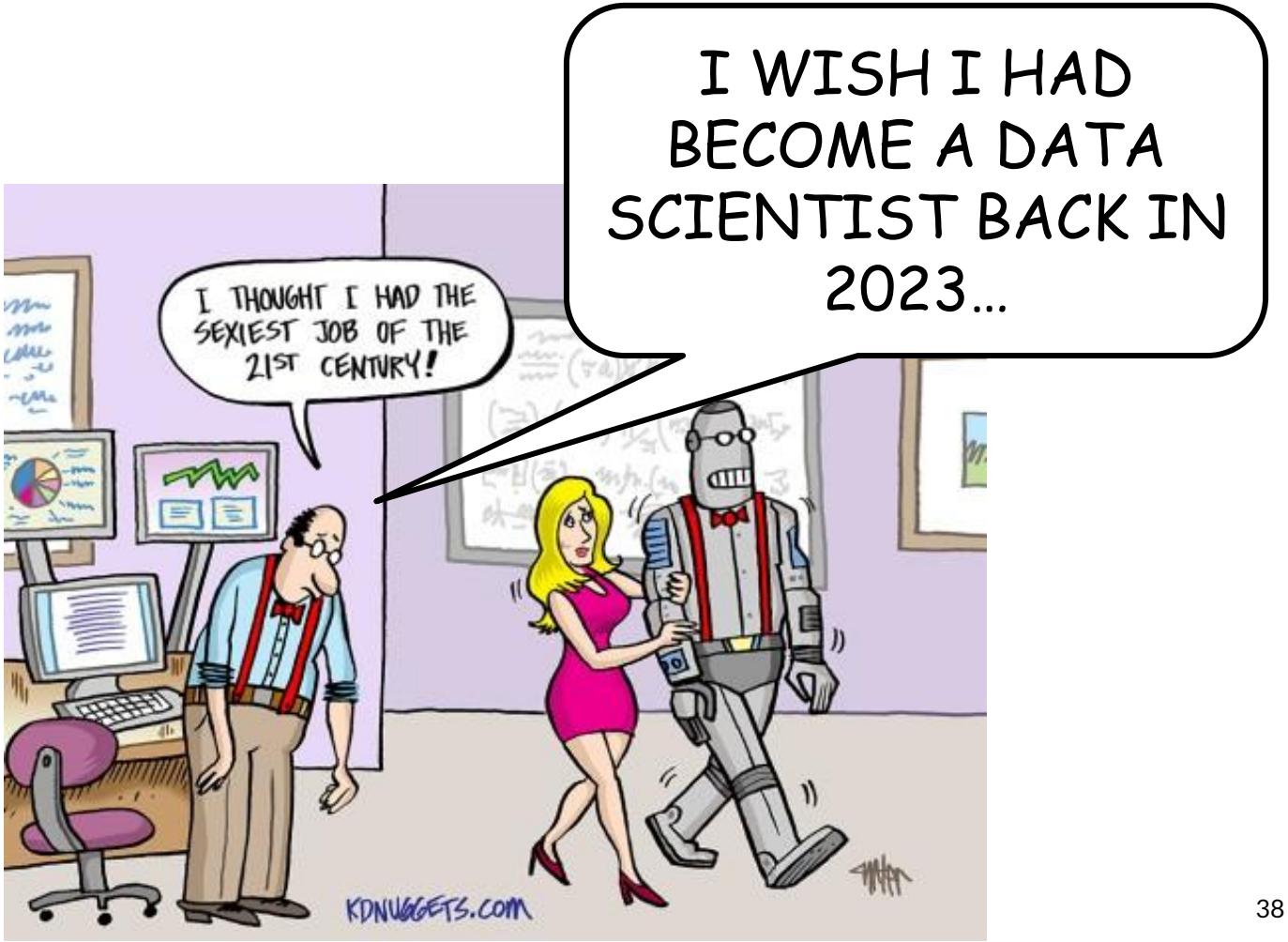
DATA

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 2023...
I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!
KDNUGGETS.COM

“Data science”

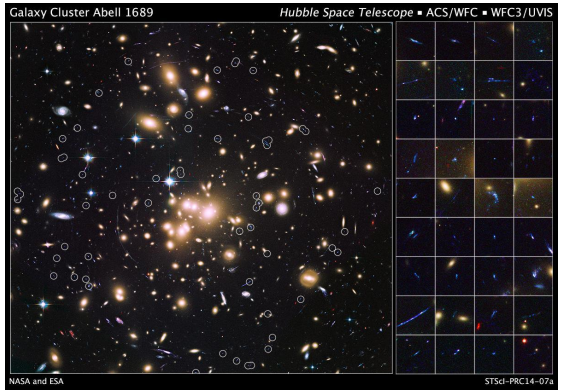
- Most science is (or should be) based on data, per definitionem
- So how is “data science” different from plain old “science”?

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.**"

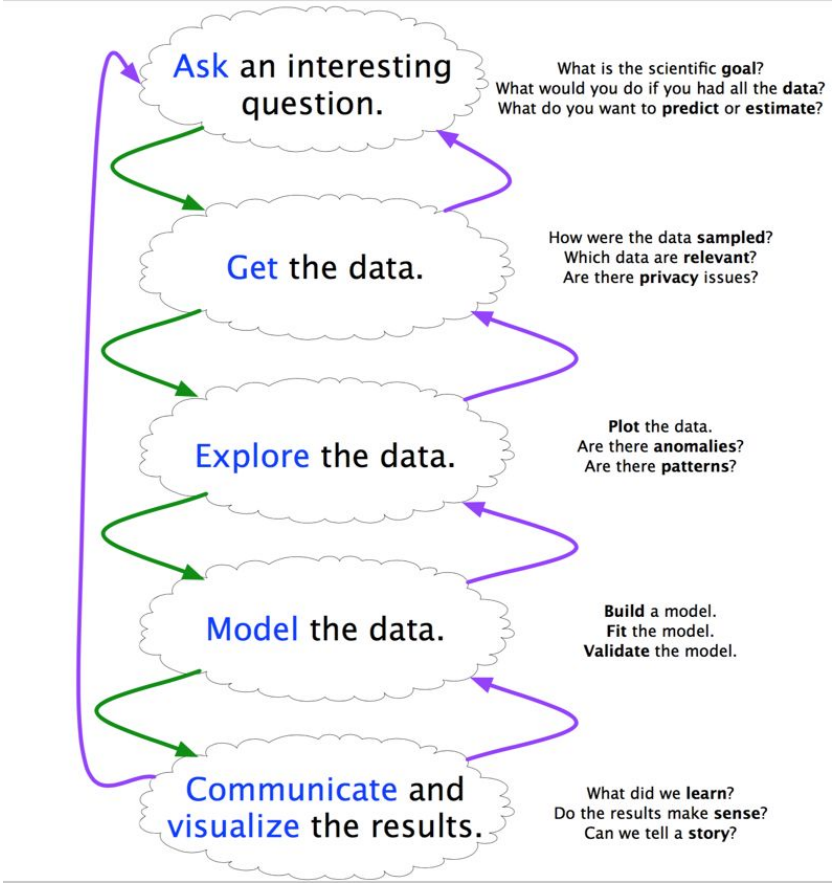
Eric Schmidt, Google (2010)

Data variety explodes



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

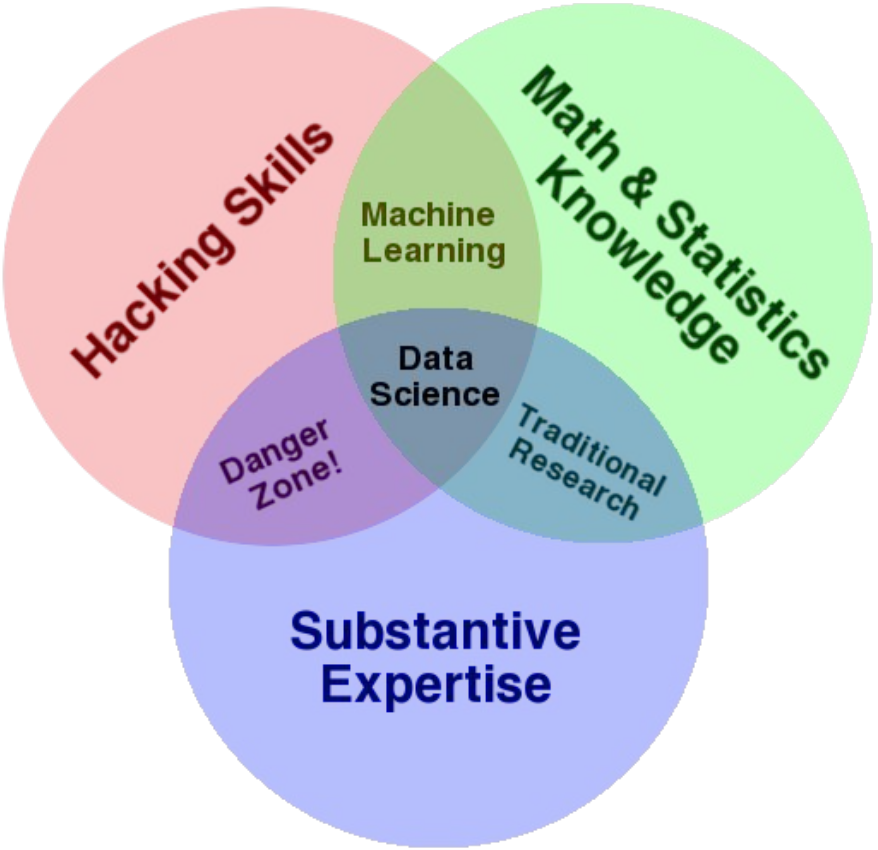
Needed: A method to the madness



- Scientific method 1.0:
  - Focused on “Model the data”
  - Scientist has hypothesis prior to analyzing the data
- Scientific method 2.0:
  - Data-driven science
  - Systematic cycle (see diagram)
  - “Explore the data” becomes increasingly important

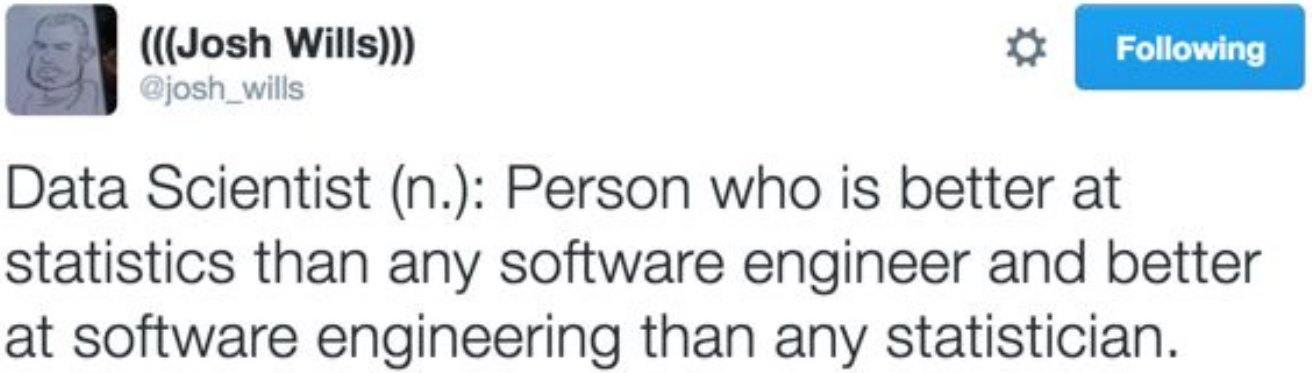
Data as a first-class citizen

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

Programming Life like a software: How Digital Biology Will Disrupt Everything?

X (formerly Twitter)

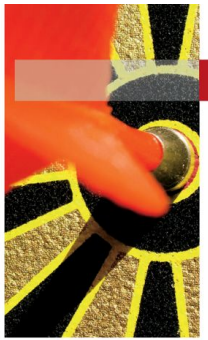
Seth Bannon 🌻 (@sethbannon) on X

"Where do I think the next amazing revolution is going to come? And this is going to be flat out one of the biggest ones ever.

There's no question that digital biology is going to be it."

Jensen Huang, founder & CEO of NVIDIA. (72 kB)

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

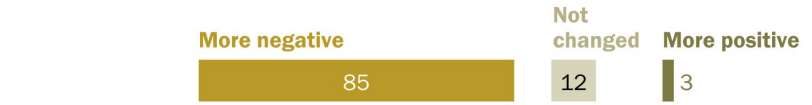
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. 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.

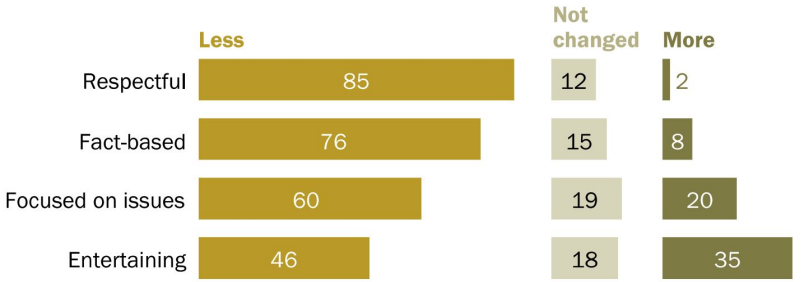
http://www.incompleteideas.net/IncIdeas/BitterLesson.html

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)

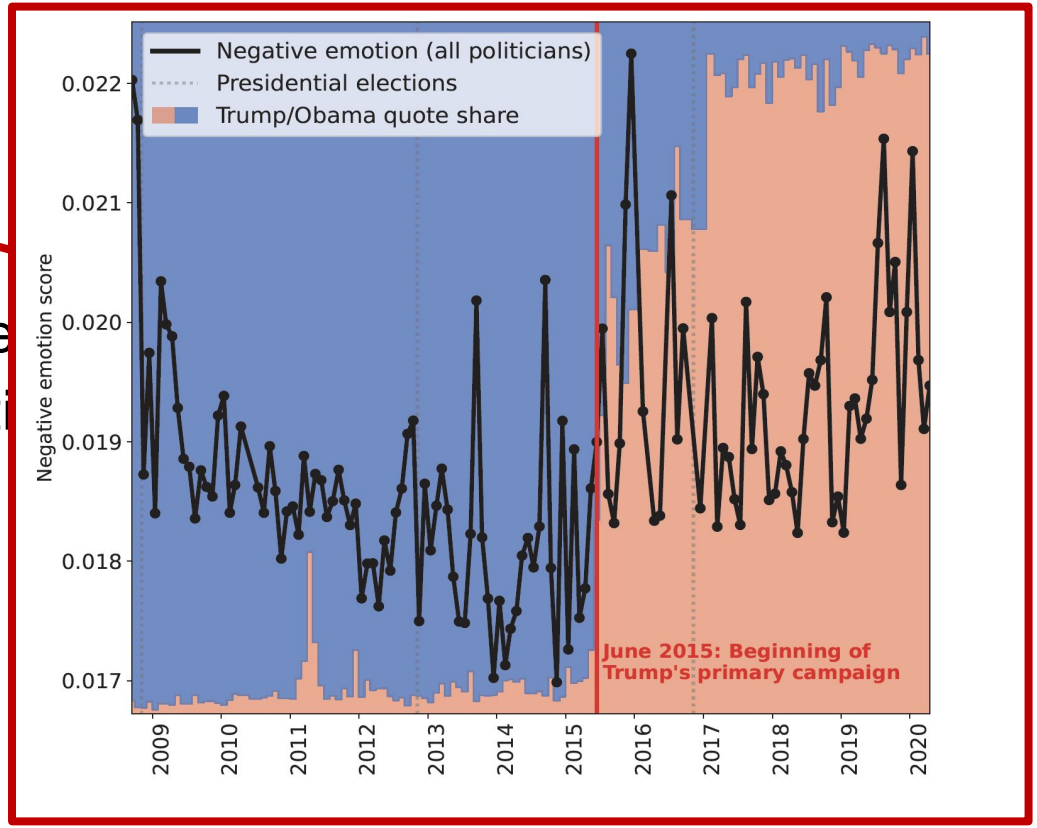
https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/06/PP_2019.06.19_Political-Discourse_FINAL.pdf

Syllabus, revisited

- **Handling data**
- Visualizing data
- Describing data
- Regression analysis 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 dise
- Causal analysis of observati
- Learning from data
- Handling text data
- Handling network data
- Scaling to massive data



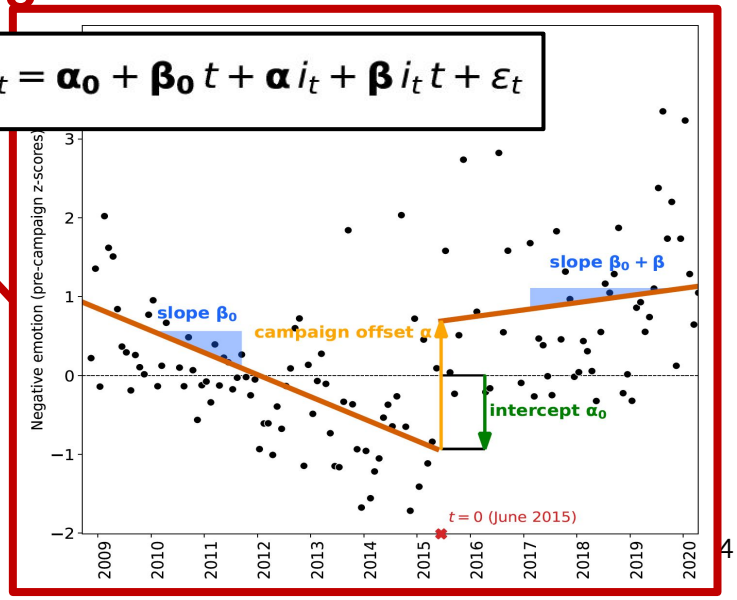
Syllabus, revisited

- 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



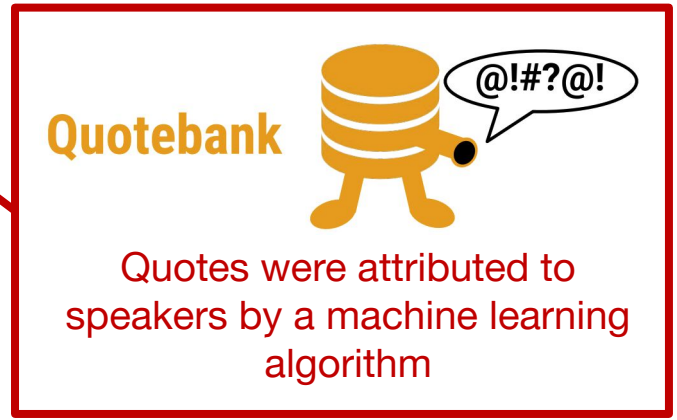
Syllabus, revisited

- 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

"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**

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

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

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]✉

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 Center[1]—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 is to investigate whether these subjective impressions reflect the true state of US political discourse.

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/ada2025-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?
- …