Introduction to Data Science and Analytics (DSC510)

University of Cyprus

Introduction

George Pallis

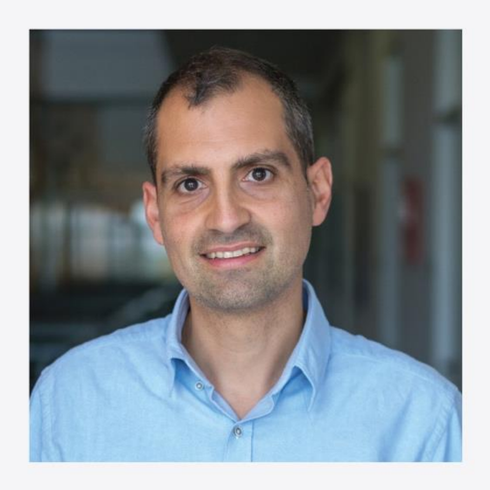
Course Website

https://piazza.com/

Your main entry point. All materials linked from there.

Main communication channel.

About me



George Pallis, PhD

Professor at Computer Science Department

Research interests:
Edge Computing, Big Data Analytics

Google Scholar:
https://scholar.google.com/citations?user=kNkLOHcAAAAJ&hl=el

Co-director of the MSc Data Science Programme

Introduction to Data Science and Analytics

- 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…

Syllabus

- Data wrangling
  - Obtaining, preparing, handling data, data sampling
- Data interpretation
  - Exploration of data, communication of results (data visualization), PCA
- Observational studies
  - How to deal with “found data”
- Applied machine learning
  - Supervised learning, Unsupervised learning, Practical issues of machine learning
- Handling specific types of data
  - Text, graphs, networks
- Fairness, privacy data and ethics, applications, use cases

Logistics

- **Lectures**
  - Tuesdays 9:00 - 11:30 (Room 146 – FST01)
  - Fridays 8:30 – 10:00 (Room 146 – FST01)
- **Labs:**
  - Wednesdays 11:30 – 13:00 (Room 201 – FST01)
  - Instructor: Dr. Pavlos Antoniou

Grading

- 10% **Homework assignments**
  - continuous assessment during the semester (homework and in-class quizzes)
- 30% **Project**
  - Groups of 3 students
  - More details later on
- 60% **Final exam**

Students who accumulate more than four (4) unexcused absences from lectures will be ineligible to take the course examinations.

Similarly, students who accumulate more than three (3) unexcused absences from lab sessions will be ineligible to submit the final project.

Learning Outcomes

- Construct a coherent understanding of the techniques and software tools required to perform the fundamental steps of the Data Science pipeline.
- Perform data acquisition (data formats, dataset fusion, Web scrapers, REST APIs, open data, big data platforms, etc.)
- Perform data wrangling (fixing missing and incorrect data, data reconciliation, data quality assessments, etc.)
- Perform data interpretation (knowledge extraction, critical thinking, team discussions, ad-hoc visualizations, etc.)
- Perform result dissemination (reporting, visualizations, publishing reproducible results, ethical concerns, etc.)

Teaching methods

- Physical in-class recitations and lab sessions
- In-class quizzes/exams
- Course project

Expected student activities

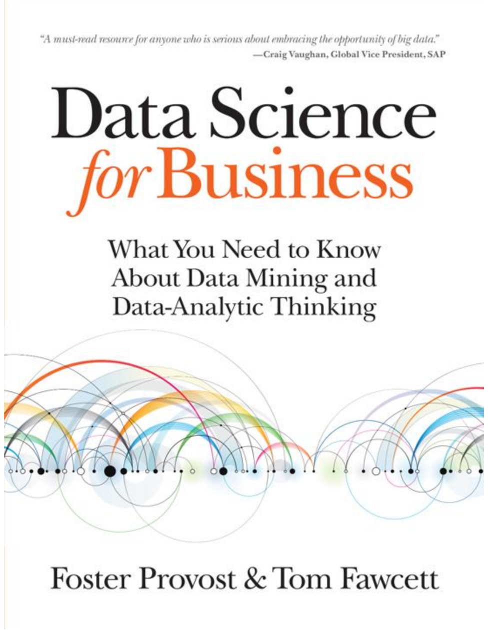
- Attend the lectures and lab sessions
- Complete 2-3 in-class quizzes (held during lab sessions)
- Conduct the class project
- Read/watch the pertinent material before a lecture
- Engage during the class, and present their results in front of the other colleagues

Prerequisites

Basics of

- probabilities and stats
- programming python™

Course Book



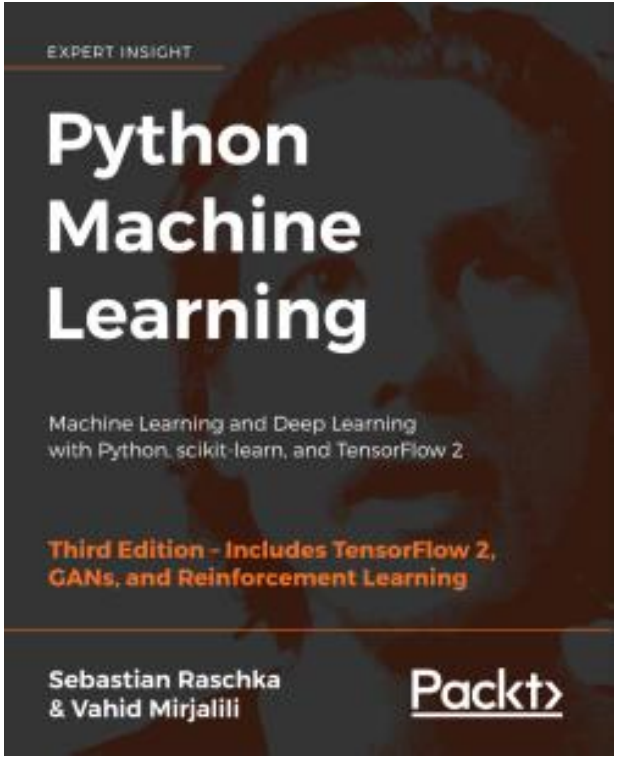
"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know
About Data Mining and
Data-Analytic Thinking

Foster Provost & Tom Fawcett

https://data-science-for-biz.com



EXPERT INSIGHT

Python
Machine
Learning

Machine Learning and Deep Learning
with Python, scikit-learn, and TensorFlow 2

Third Edition - Includes TensorFlow 2,
GANs, and Reinforcement Learning

Sebastian Raschka
& Vahid Mirjalili

Packt>

https://github.com/rasbt/python-machine-learning-book-3rd-edition

- Participate actively in classes and labs
- Give us **feedback**

Why Data Science Matters

Data science enhances critical thinking

The world is complicated! Decisions are hard.

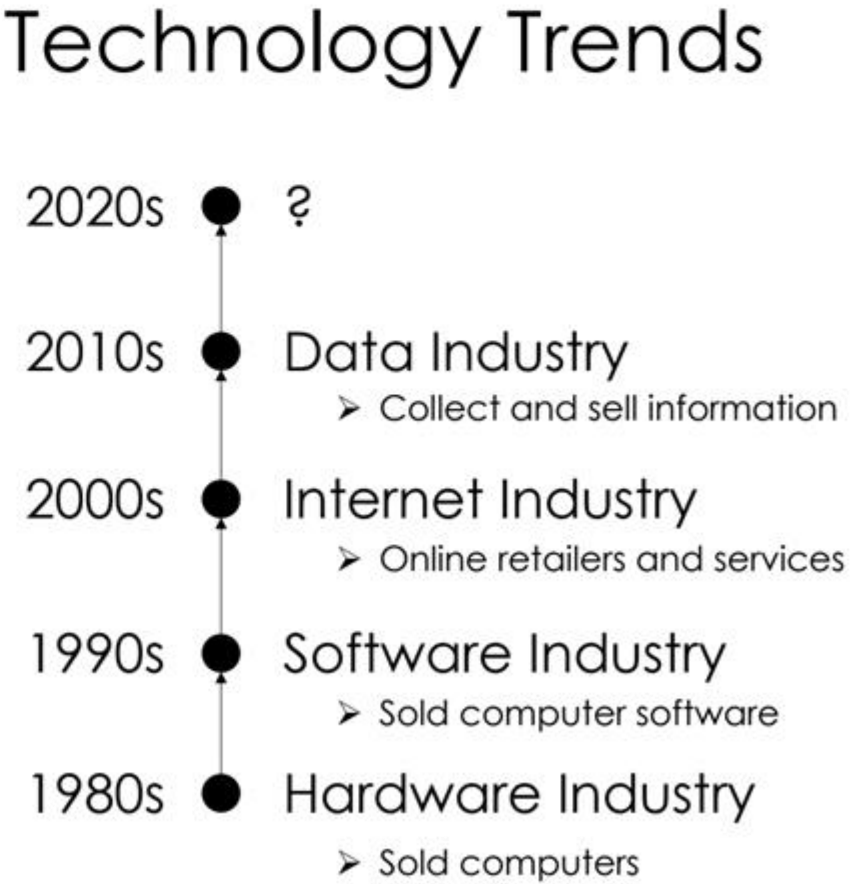
Data is used everywhere to answer hard questions and make tough decisions:

- Science
- Medicine
- Social science
- Engineering
- Sports

Claims about data come up in discussing almost any important issue:

- Instead of "Aquinas says," now it's "the data says."
- It is usually not easy to tell what the data "says"
- **Empower yourself** to participate in the arguments that shape your life and your society

Data is changing the world



Technology Trends

2020s ● ?

2010s ● Data Industry

- ➢ Collect and sell information

2000s ● Internet Industry

- ➢ Online retailers and services

1990s ● Software Industry

- ➢ Sold computer software

1980s ● Hardware Industry

- ➢ Sold computers

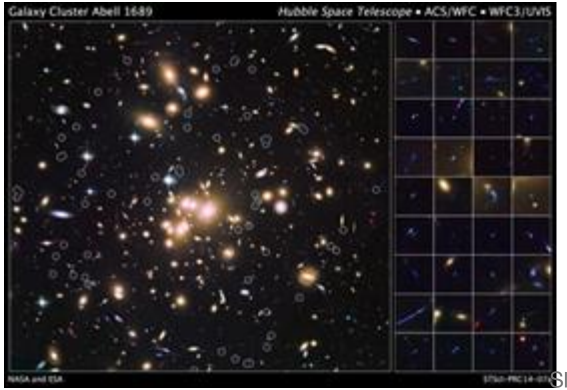
From Joey Gonzalez.

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 (2010)

Data variety explodes


Galaxy Cluster Abell 1689
Hubble Space Telescope • ACS/WFC • WFC3/UVIS
NASA and ESA
STScI-PRC14-07

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

The darker side of data science?

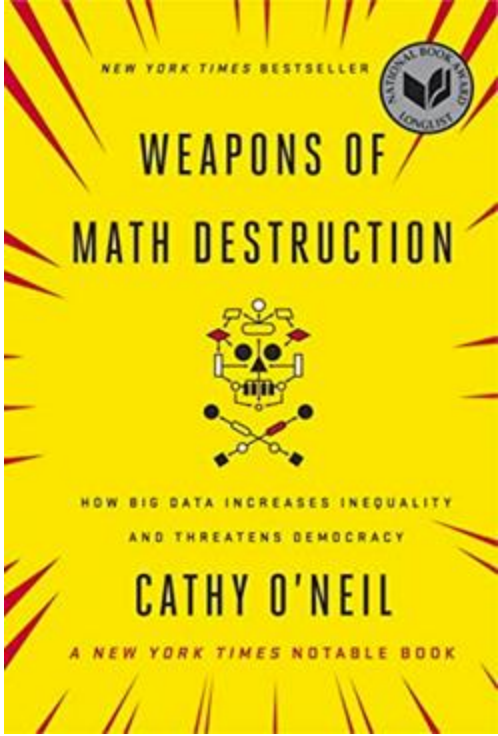
Obscuring complex decisions:

- Mortgage-backed securities → market crash
- Teaching scores & job advancement

Reinforcing historical trends and biases:

- Hiring based on previous hiring data
- Recidivism and racially biased sentencing
- Social media, news, and politics

We will discuss the ethics of data science throughout the class!


NEW YORK TIMES BESTSELLER
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK

NPR author interview
with Cathy O'Neil

But…we are optimistic!

Knowledge is empowering.

Data science offers **immense potential** to address challenging problems facing society.

The future is in your hands, and I believe:

You will use your knowledge for good.

Data science enhances critical thinking

The world is complicated! Decisions are hard.

Data science is a fundamentally human-centered field that facilitates decision-making by quantitatively balancing tradeoffs.

- To quantify things **reliably** we must:
  - Find relevant data;
  - Recognize its limitations;
  - Ask the right questions;
  - Make reasonable assumptions;
  - Conduct an appropriate analysis; and
  - Synthesize and explain our insights.

- Apply **critical thinking and skepticism** at every step; and

- Consider how our decisions **affect others**.

Our primary goal for you in this course

The world is complicated! Decisions are hard.

Data science is a fundamentally human-centered field that facilitates decision-making by quantitatively balancing tradeoffs.

- To quantify things **reliably** we must:
  - Find relevant data;
  - Recognize its limitations;
  - Ask the right questions;
  - Make reasonable assumptions;
  - Conduct an appropriate analysis; and
  - Synthesize and explain our insights.
- Apply **critical thinking and skepticism** at every step; and
- Consider how our decisions **affect others.**

After this course, you should be able to take data and produce useful insights on the world's most challenging and ambiguous problems.

What is DSC510?

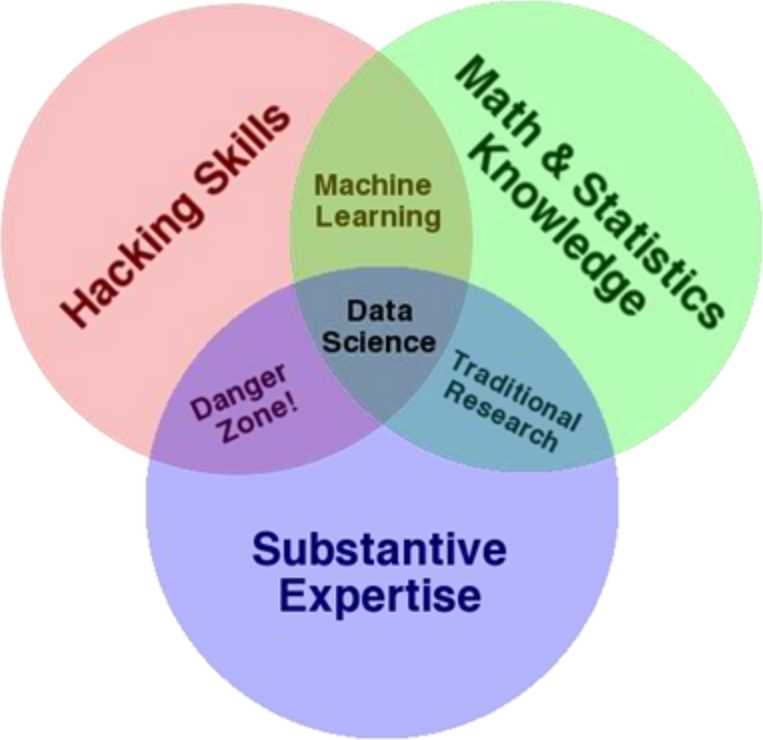
PRINCIPLES AND TECHNIQUES OF DATA SCIENCE

Data science is a fundamentally interdisciplinary field

Data Science is the application of data centric, computational, and inferential thinking to:

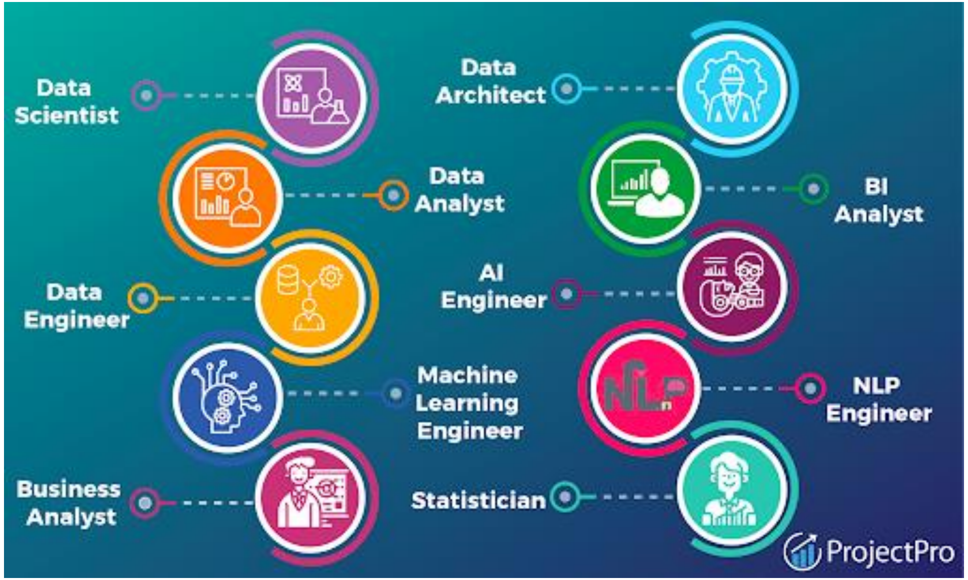
- Understand the world (science).
- Solve problems (engineering).

Data Science Venn Diagram



by Drew Conway in 2010 (link)

Data science in industry



The major tasks that data scientists say they work on regularly.

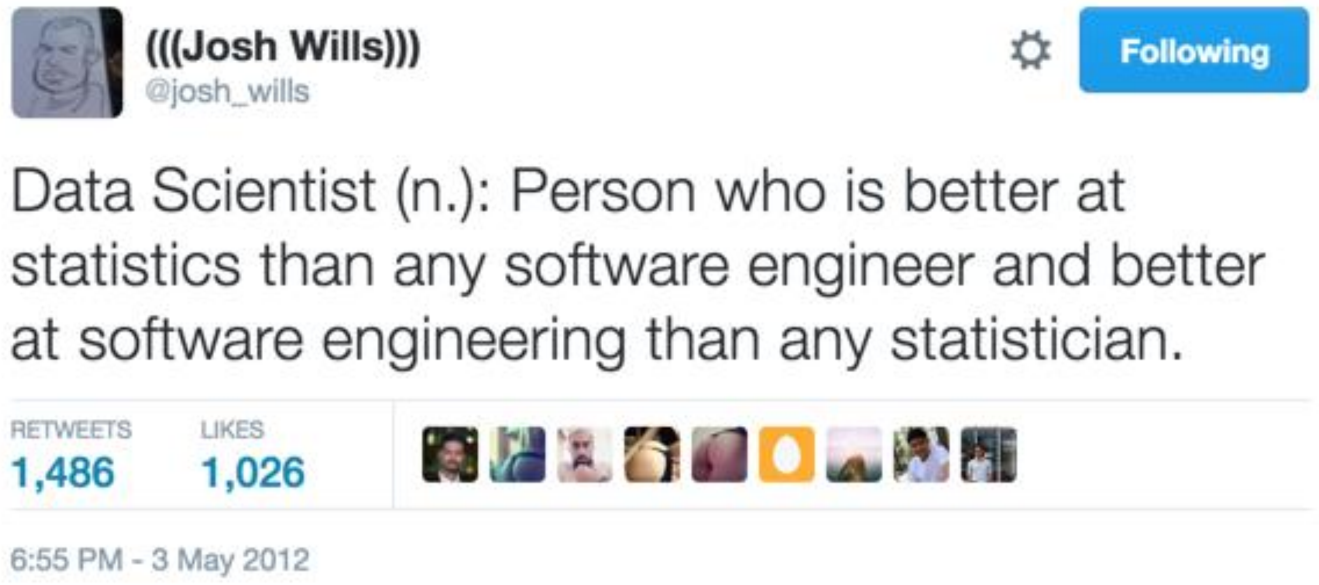
Data analysis

“... 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.**”

“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

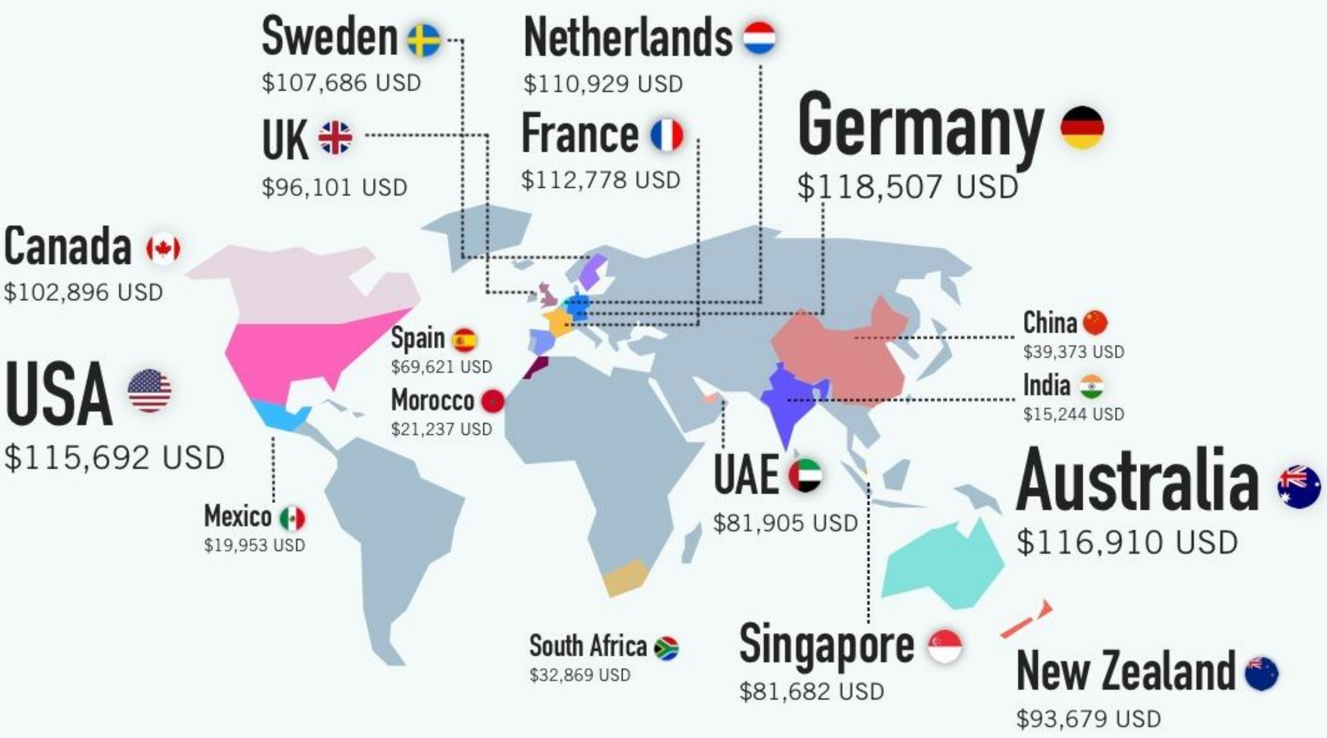


Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

6:55 PM - 3 May 2012

Josh Wills, Data Scientist at Slack

Data scientist salary guide



Sweden
$107,686 USD

UK
$96,101 USD

Netherlands
$110,929 USD

France
$112,778 USD

Germany
$118,507 USD

Canada
$102,896 USD

USA
$115,692 USD

Spain
$69,621 USD

Morocco
$21,237 USD

Mexico
$19,953 USD

China
$39,373 USD

India
$15,244 USD

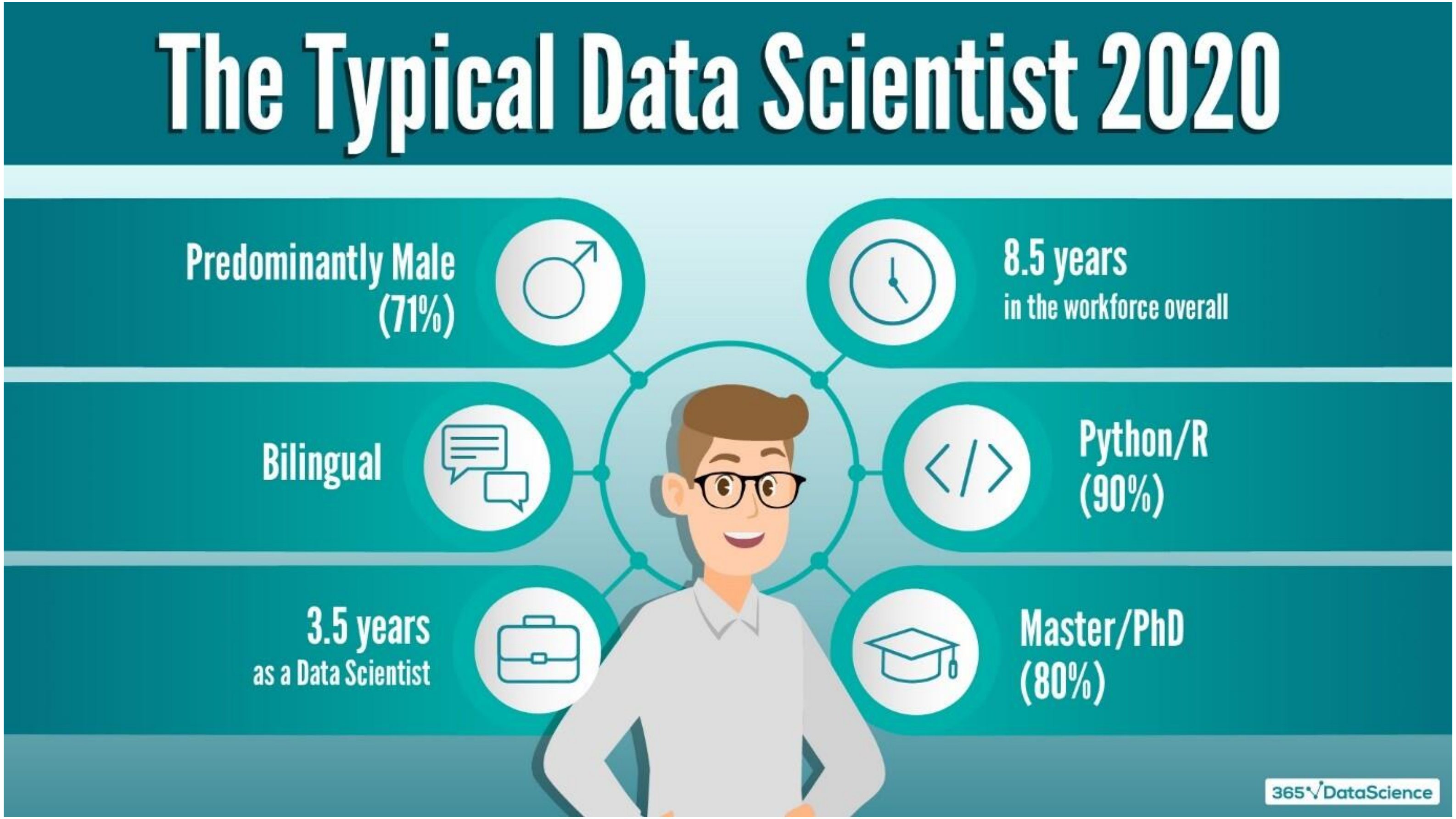
UAE
$81,905 USD

Australia
$116,910 USD

South Africa
$32,869 USD

Singapore
$81,682 USD

New Zealand
$93,679 USD



The Typical Data Scientist 2020

Predominantly Male (71%)

Bilingual

3.5 years as a Data Scientist

8.5 years in the workforce overall

Python/R (90%)

Master/PhD (80%)

Data science requires engineering and scientific insight

Good data analysis is not:

- Simple application of a statistics recipe.
- Simple application of statistical software.

There are many **tools** out there for data science, but they are merely tools.

- **They don't do any of the important thinking!**

"The purpose of computing is insight, not numbers."

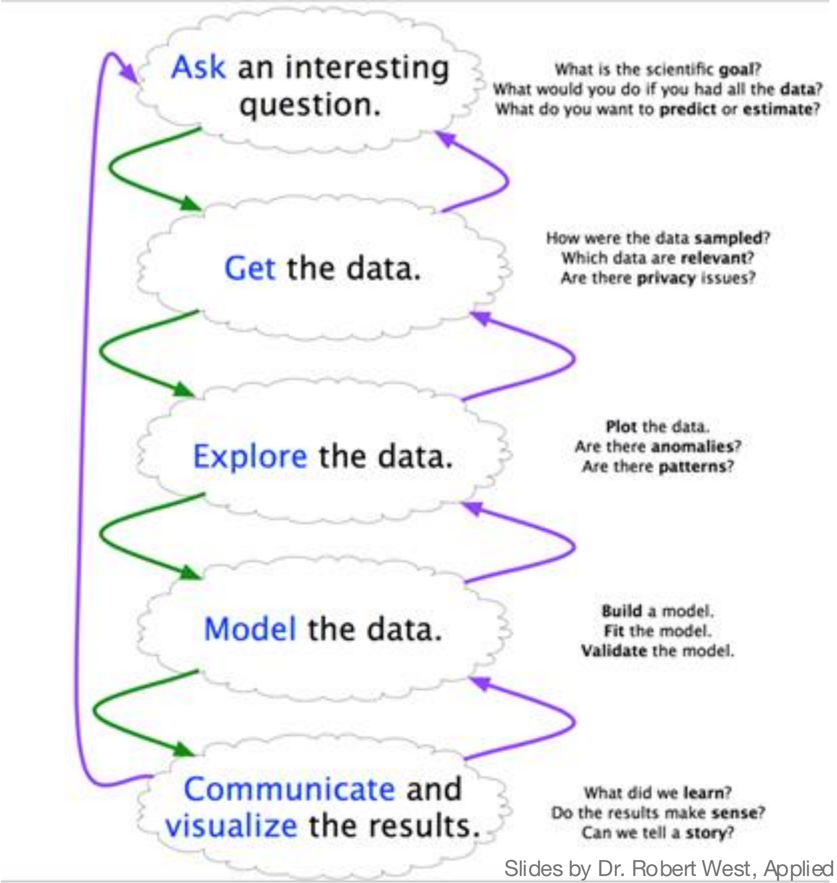
R. Hamming. *Numerical Methods for Scientists and Engineers (1962).*

Example Questions in Data Science

Some (broad) questions we might try to answer with data science:

- What show should we recommend to our user to watch?
- In which markets should we focus our advertising campaign?
- Should I send my kids to daycare?
- What areas of the world are at higher risks for climate change impact in 10 years? 20?
- Where should we put docking ports for our bikes?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?

Needed: A method to the madness


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

- 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**

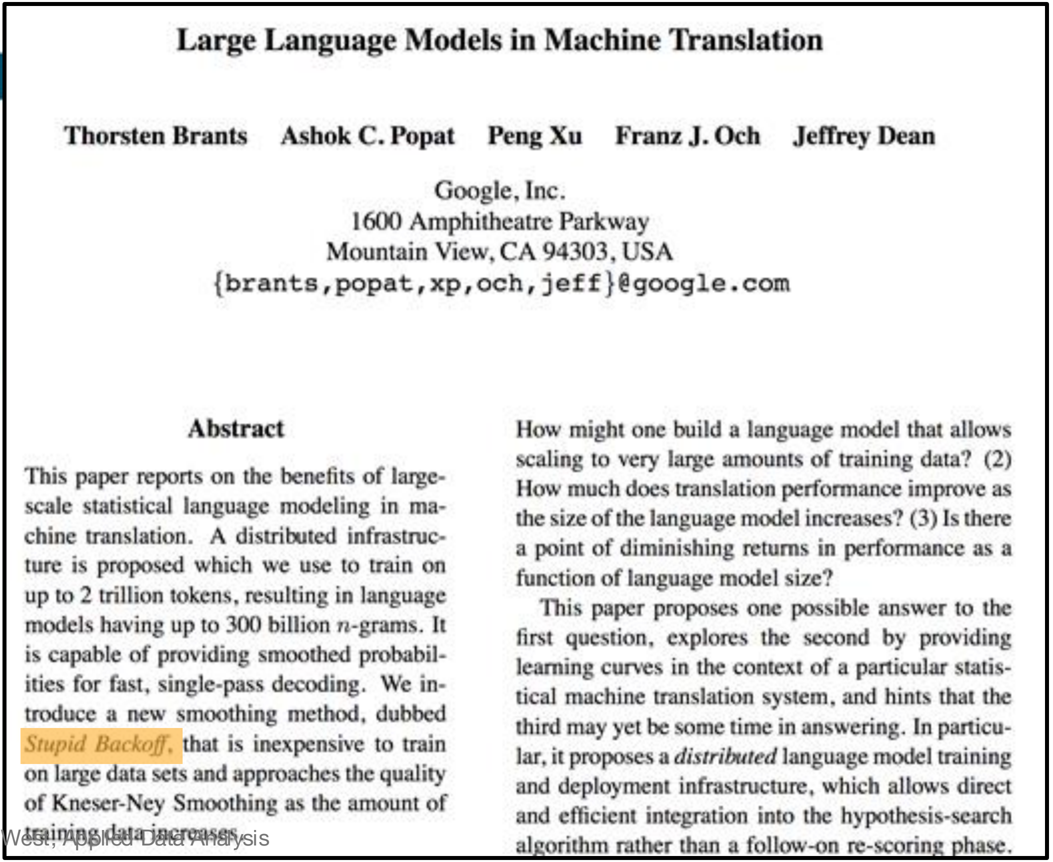
More data often beats better algorithms

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

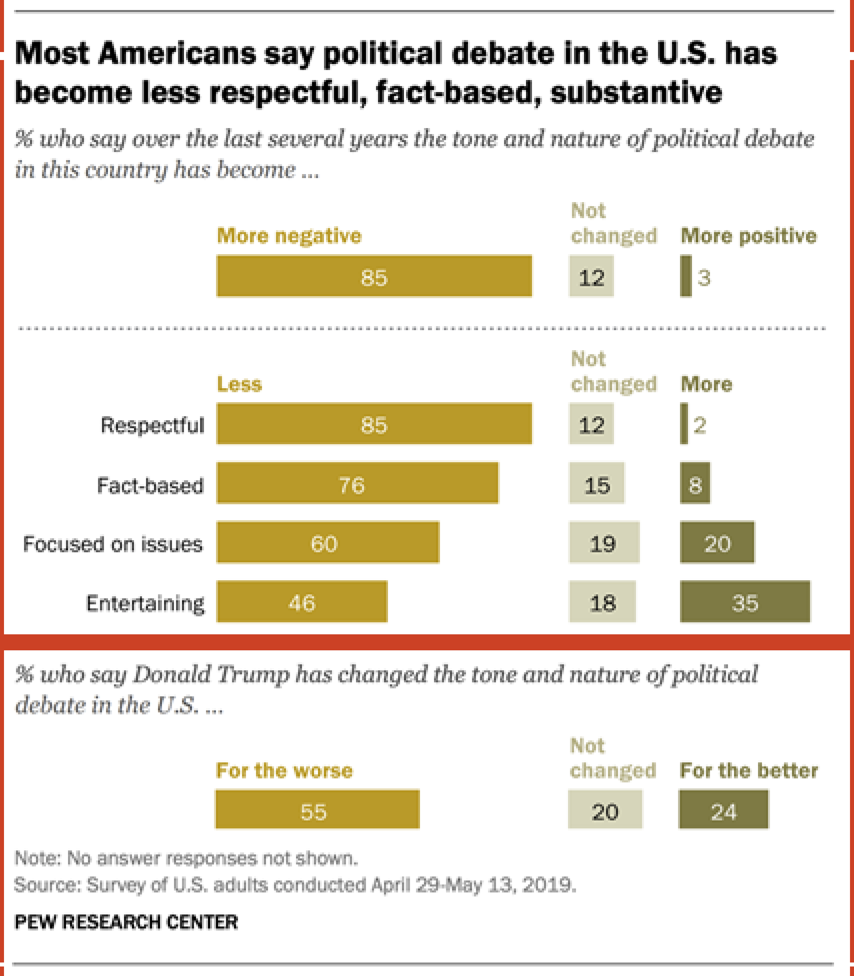
Abstract

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.

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

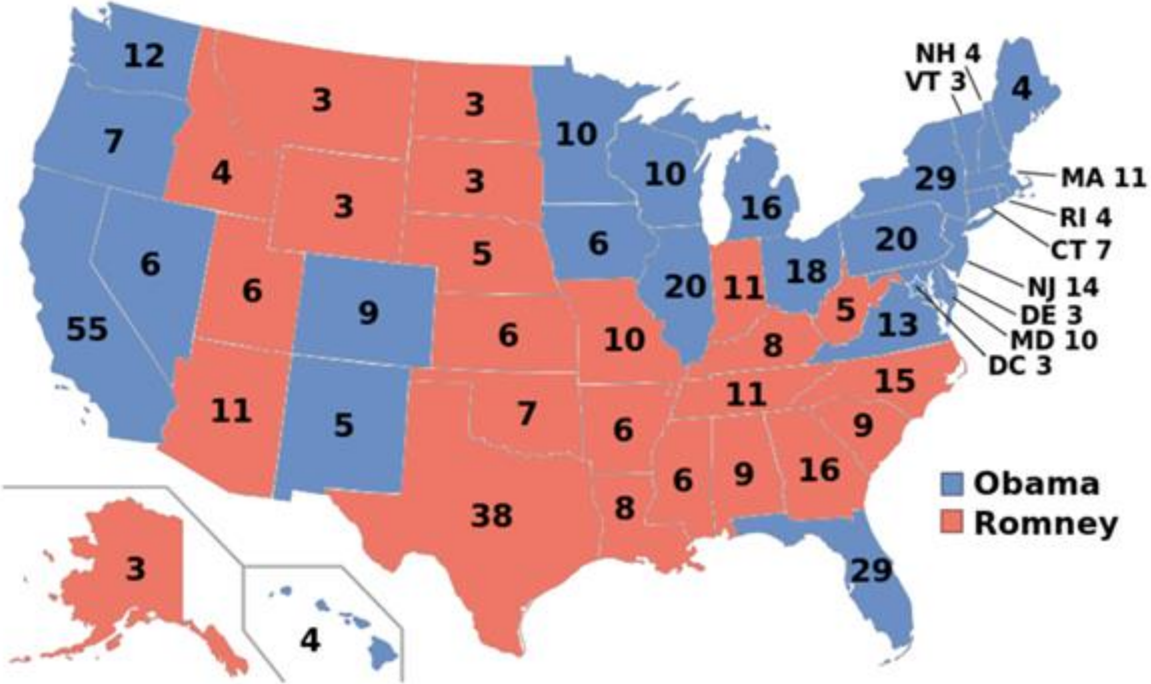
We ask: Do these subjective impressions reflect the true state of US political discourse?

DSC510 will teach you the tools to answer such questions using data

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

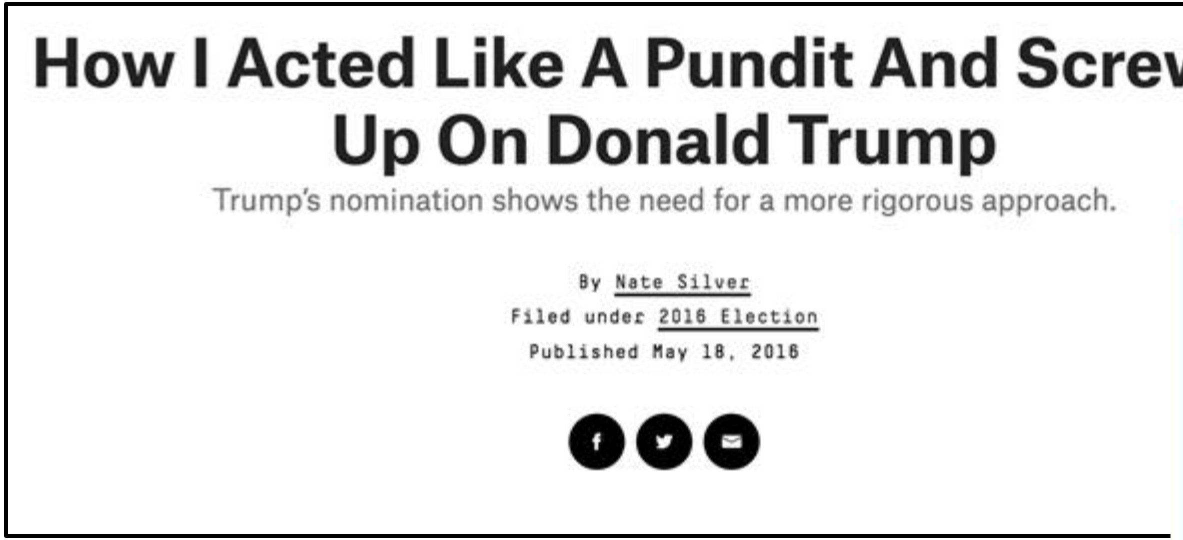
2008, 2012: Predicting elections

> “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.**"

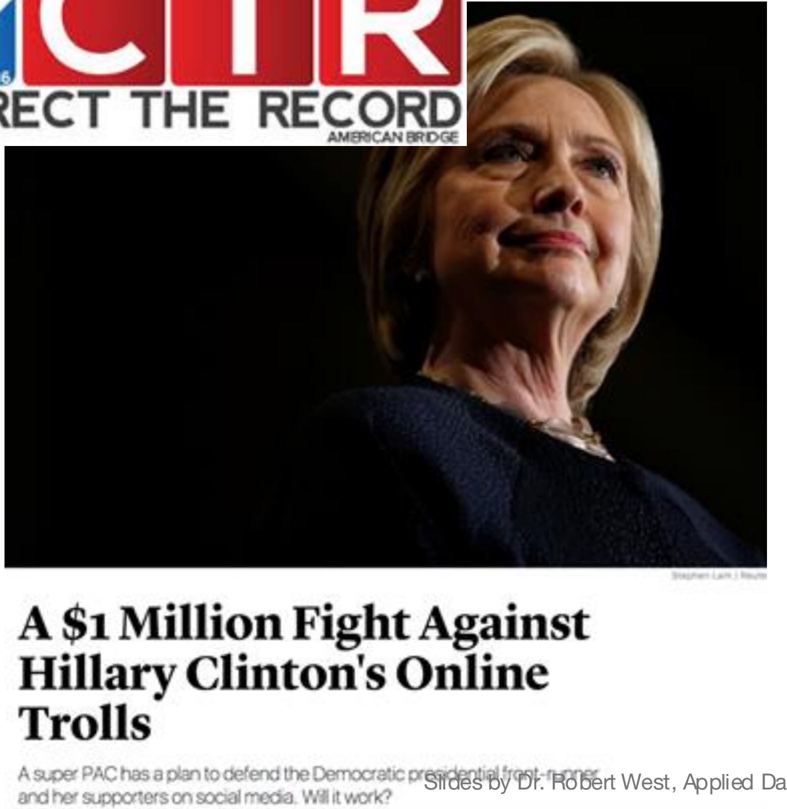
Andrew Rasiej, Personal Democracy Forum

"... 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

CTR
2016
CORRECT THE RECORD
AMERICAN BRIDGE

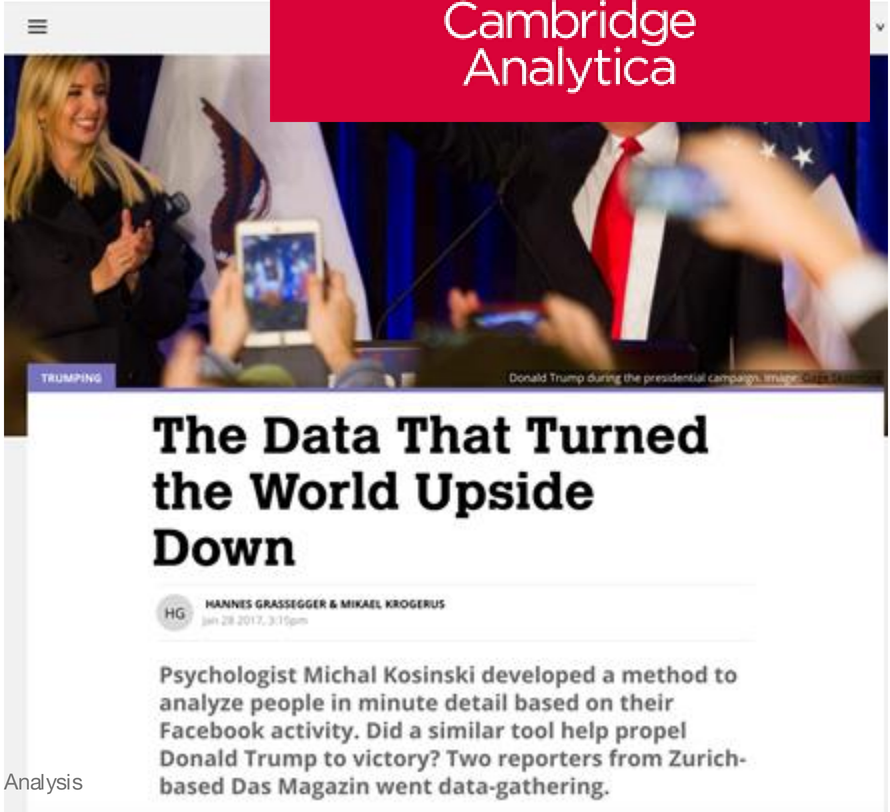


Stephen Lam / Reuters

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:

The Data That Turned the World Upside Down

HG HANNES GRASSEGGER & MIKAEL KROGERUS
Jan 28 2017, 3:10pm

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.

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

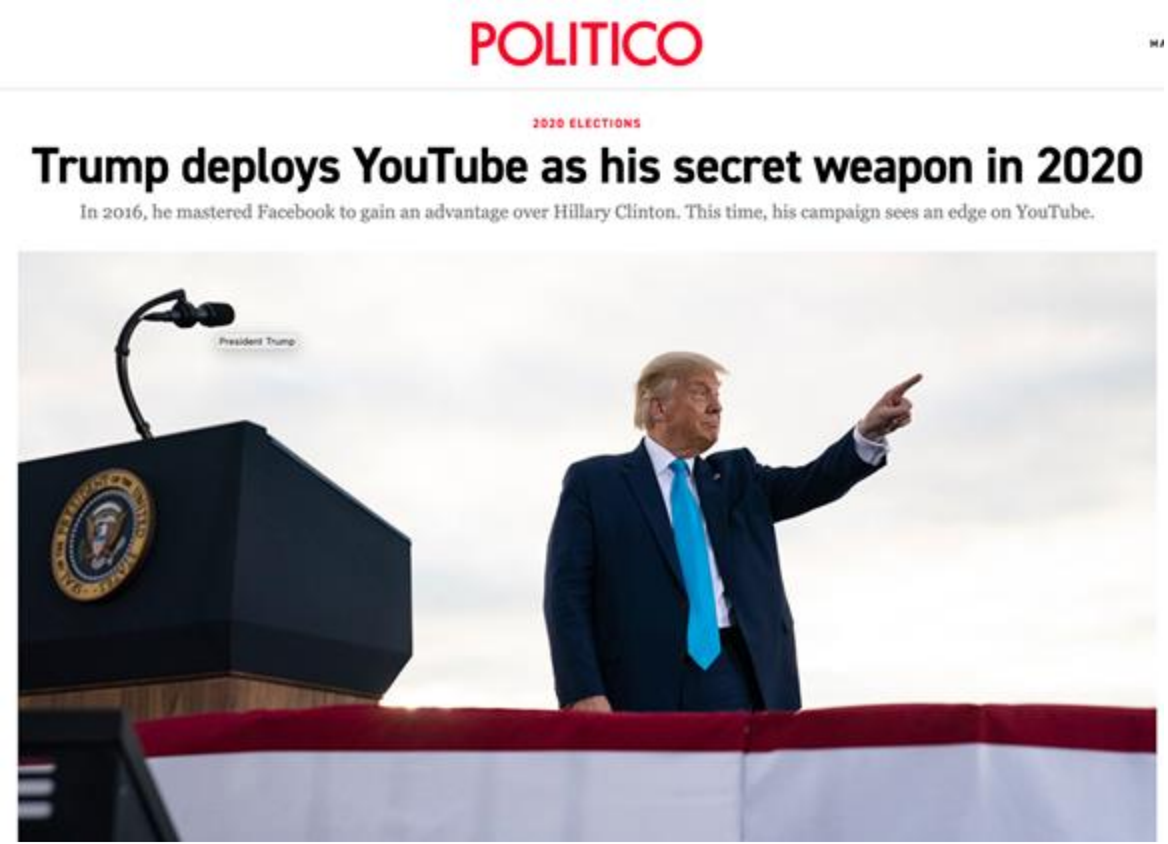
United States politicians' tone became more negative with 2016 primary campaigns

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

2020:



POLITICO

2020 ELECTIONS

Trump deploys YouTube as his secret weapon in 2020

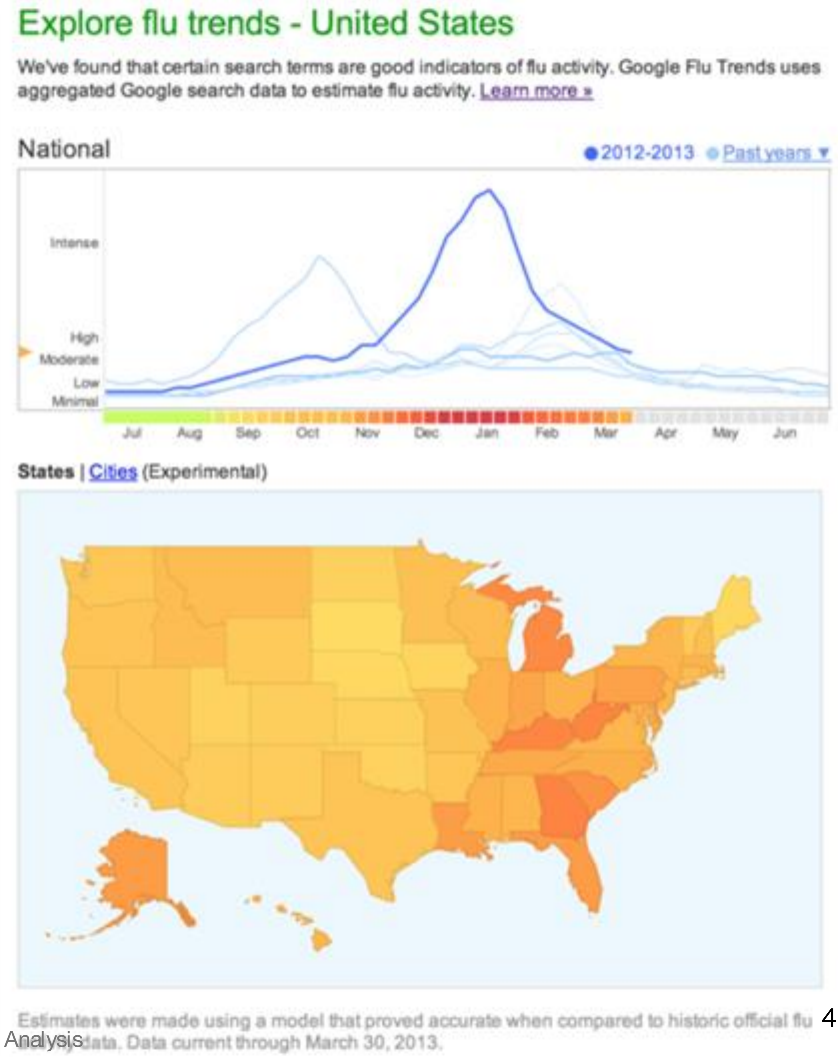
In 2016, he mastered Facebook to gain an advantage over Hillary Clinton. This time, his campaign sees an edge on YouTube.

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.”



Explore flu trends - United States

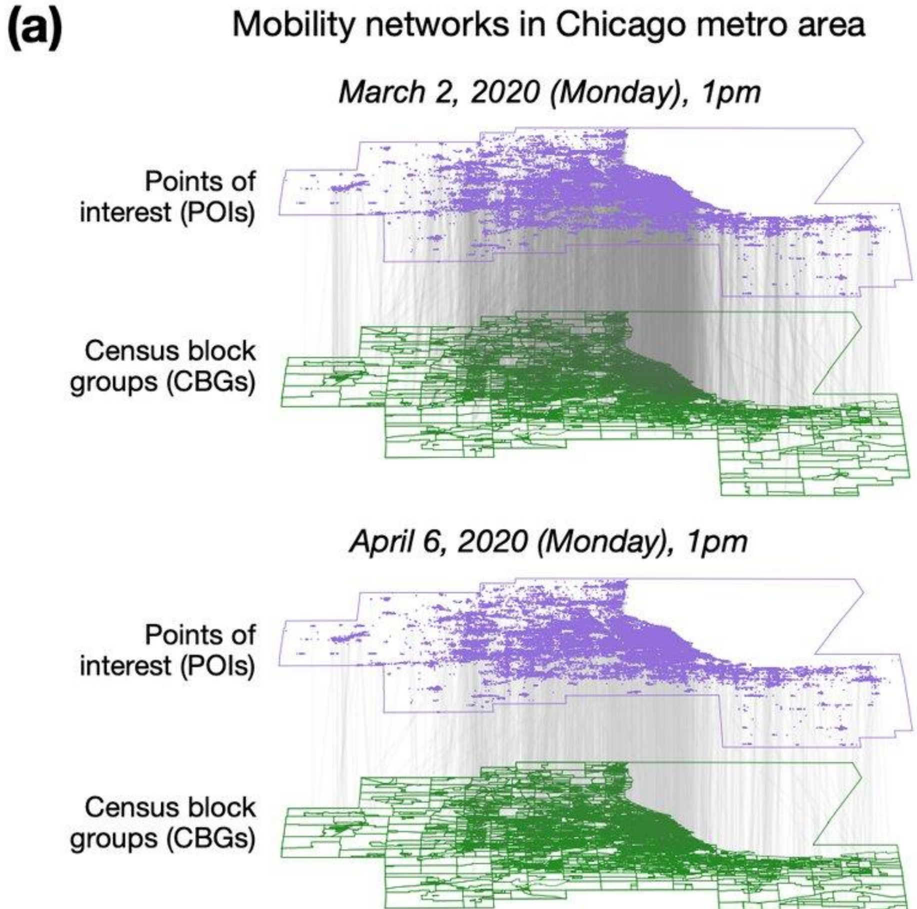
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)

Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 30, 2013.

Mobility network models of COVID-19 explain inequities and inform reopening


(a)
Mobility networks in Chicago metro area
March 2, 2020 (Monday), 1pm
Points of interest (POIs)
Census block groups (CBGs)
April 6, 2020 (Monday), 1pm
Points of interest (POIs)
Census block groups (CBGs)

Supporting COVID-19 policy response with large-scale mobility-based modeling

Serina Chang[1], Mandy L. Wilson[2], Bryan Lewis[2], Zakaria Mehrab[2], Komal K. Dudakiya[3], Emma Pierson[4], Pang Wei Koh[1], Jaline Gerardin[5], Beth Redbird[6], David Grusky[7], Madhav Marathe[2], Jure Leskovec[1]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA
[2] Biocomplexity Institute & Initiative, University of Virginia, Charlottesville, VA, USA
[3] Persistent Systems, Pune, India
[4] Microsoft Research, Cambridge, MA, USA
[5] Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
[6] Department of Sociology, Northwestern University, Evanston, IL, USA
[7] Department of Sociology, Stanford University, Stanford, CA, USA

ABSTRACT

Social distancing measures, such as restricting occupancy at venues, have been a primary intervention for controlling the spread of COVID-19. However, these mobility restrictions place a significant economic burden on individuals and businesses. To balance these competing demands, policymakers need analytical tools to assess the costs and benefits of different mobility reduction measures. In this paper, we present our work motivated by our interactions with the Virginia Department of Health on a decision-support tool that utilizes large-scale data and epidemiological modeling to quantify the impact of changes in mobility on infection rates. Our model captures the spread of COVID-19 by using a fine-grained, dynamic mobility network that encodes the hourly movements of people from neighborhoods to individual places, with over 3 billion hourly edges. By perturbing the mobility network, we can simulate a wide variety of reopening plans and forecast their impact in terms of new infections and the loss in visits per sector. To deploy this model in practice, we built a robust computational infrastructure to support running millions of model realizations, and we worked with policymakers to develop an intuitive dashboard interface that communicates our model's predictions for thousands of potential policies, tailored to their jurisdiction. The resulting decision-support environment provides policymakers with much-needed analytical machinery to assess the tradeoffs between future infections and mobility restrictions.

1 INTRODUCTION

The COVID-19 pandemic has wreaked havoc on lives and livelihoods across the globe. In an effort to contain the virus, policymakers have turned to non-pharmaceutical interventions, such as restricting mobility, in order to limit contact and reduce disease transmission between individuals. To this end, many US states and local governments have closed or required reduced occupancy at places such as restaurants and gyms [7]. However, these measures come at a heavy cost to individuals and businesses: for example, over 160,000 US businesses closed due to COVID-19 between March and August 2020 [36].

The next few months will continue to pose challenges to public health and economic activity. It is imperative during this time to provide policymakers with analytical tools that can help them develop optimal solutions to minimize new COVID-19 infections while also minimizing damage to businesses. They need a tool that can quantitatively assess, in near real-time, the tradeoffs between mobility and new infections. Furthermore, this tool should be fine-grained, able to test out heterogeneous plans—for example, allowing one level of mobility at essential retail, another level at retail, and yet another at restaurants—so that policymakers can tailor restrictions to the specific risks and needs of each sector. Despite this granularity, the tool also needs to be scalable, supporting analyses for an exponential number of potential policies so that policymakers can select the best option among them for their jurisdiction.

To fulfill these needs, we present a decision-support tool, which we built based on interactions with the Virginia Department of Health (VDH) to support their decision-making on mobility reduction policies. Our approach begins with our state-of-the-art epidemiological model [8], which integrates large-scale mobility and mask-wearing data to accurately capture the spread of COVID-19. Our model overlays transmission dynamics on a time-varying mobility network which encodes the hourly movements of individuals from neighborhoods to specific points of interest (POIs), such as restaurants or grocery stores. Since we model infections in tandem with mobility, our model can provide the multifaceted analyses necessary to understand the costs and benefits of a policy; for example, by quantifying predicted infections and the number of POI visits lost per sector, which can serve as a proxy for economic impact. By design, our model is fine-grained, as it simulates who is getting infected where and when down to the individual POI and hour. Our model is also flexible, since we can modify any one of its inputs—for example, modifying mobility for a subset of POIs to reflect a change in policy, or modifying transmission rates per neighborhood to indicate vaccination effects—and straightforwardly run the model with the new inputs to observe the effects of the hypothetical change.

Finally, to scale our model, we build a robust infrastructure to handle computational challenges. The mobility networks that we model are large, with billions of hourly edges between POIs and neighborhoods. Furthermore, the flexibility of our approach—for

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

CHAT!

Think for 1 minute and write down

2-3 things you are excited to do with data science!

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

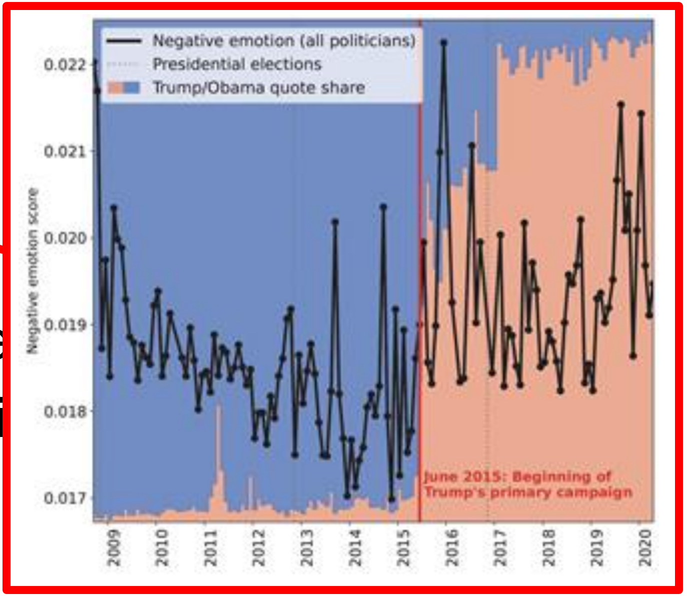
Syllabus, revisited

- **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

Data: https://github.com/epfl-dlab/Quotebank
Web interface: https://quotebank.dlab.tools/

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



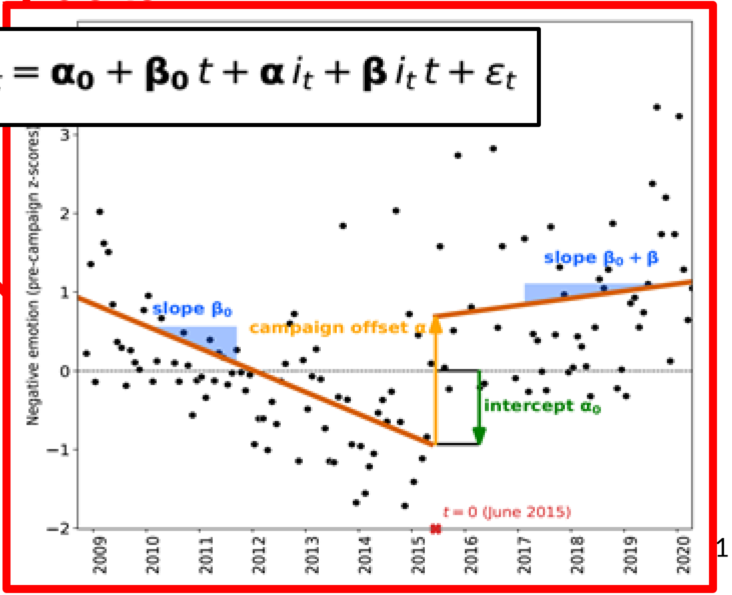
Syllabus, revisited

- 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

“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



$$y_t = \boldsymbol{\alpha}_0 + \boldsymbol{\beta}_0\, t + \boldsymbol{\alpha}\, i_t + \boldsymbol{\beta}\, i_t\, t + \varepsilon_t$$

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

“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

Quotebank

@!#?@!

Quotes were attributed to speakers by a machine learning algorithm

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

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**

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

… and key principles

- Use many data sources
- Be critical
- 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