NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 1: Lecture 2
Overview – Solving Problems with Data

NYU Center for Data Science

What is data science?
What are steps for applying data science?

DS-UA 112
Introduction to Data Science

Week 1: Lecture 2
Overview – Solving Problems with Data

Adapted from Nolan, Steinhardt, Lau, and Salganik

Announcements

- Please access Syllabus on NYU Classes
- Please check Week 1 agenda on NYU Classes
  - Homework 0
  - Lab 1
  - Survey 1

Remember to post on Piazza and participate in Instructor, TA or Grader Office Hours

Announcements

- NYU Classes
  - Weekly Agenda, Zoom Conference, Syllabus
- JupyterHub
  - Class Materials, Submission Labs/Homework/Projects
- Piazza
  - Announcements, Discussion
- Gradescope
  - Submission Homework/Projects, Retrieve Exams

Please complete Survey 1 by Monday February 10

Review: What characterizes data?

Inside the Fight to Save Alaska's 20 Native Languages from Dying Out

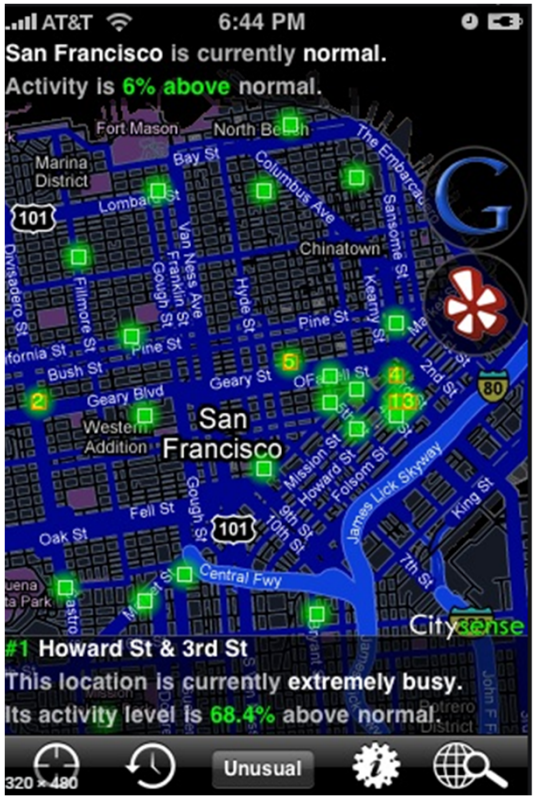
Alaskans are racing to collect their elders' knowledge of 20 native languages and design a pedagogy that breaches the generational divide.

By Agnes Walton

Volume

"[Our] corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion. The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over."

Review: What characterizes data?



Volume

WIRED

NYC Now Knows More Than Ever About Your Uber and Lyft Trips

In 2007, New York City's Taxi and Limousine Commission, in a belated ... because along with those readers came GPS trackers that became a ... The ride-hail companies, though, remain wary of such data operations.

Jan 31, 2019

Review: What characterizes data?

Velocity

NYU

Memorandum

Date: January 24, 2020

To: THE NYU COMMUNITY

From: Carlo Ciotoli, MD, Associate Vice President for Student Health and Executive Director of the Student Health Center

Re: The Emergence of the Novel Coronavirus

An AI Epidemiologist Sent the First Warnings of the Wuhan Virus

WIRED · 2 days ago

Review: What characterizes data?

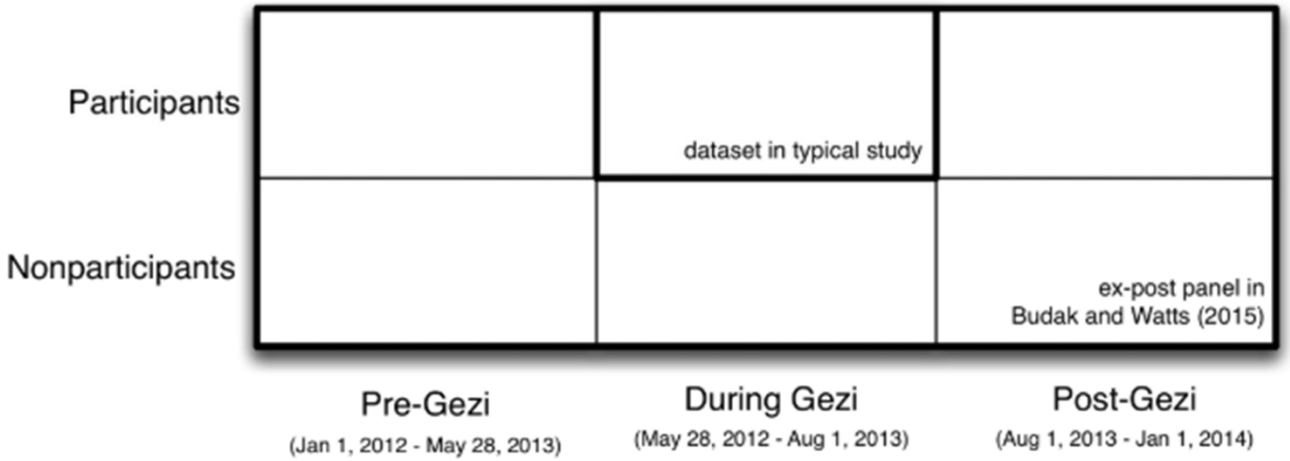
Reuters UK

Turkey escalates crackdown on dissent six years after Gezi protests

Turkey escalates crackdown on dissent six years after Gezi protests ... were inspired by the worldwide “Occupy” protests and Arab uprisings ...

Mar 18, 2019

Velocity



Review: What characterizes data?

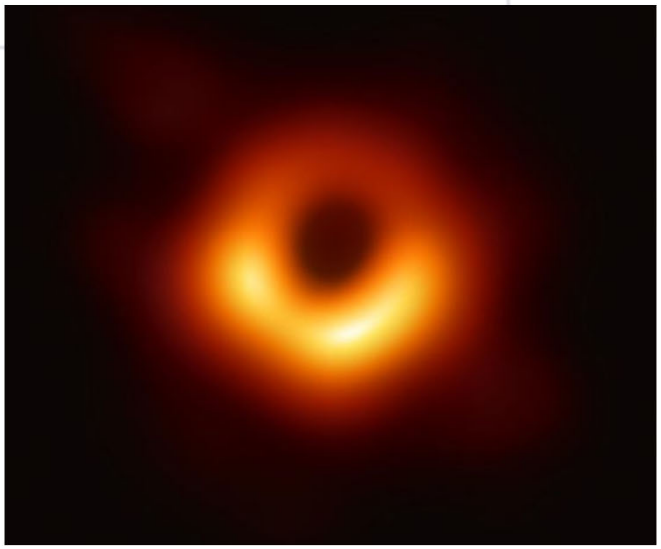
Smithsonian

The Top Ten Scientific Discoveries of the Decade

The Top Ten Scientific Discoveries of the Decade ... fossils were originally identified as Homo sapiens, but a 2019 analysis determined that the ... An image of the environment around the black hole at the center of Messier 87,

1 month ago

Variety



Review: What characterizes data?

Not Reactive

The cost of racial animus on a black candidate: Evidence using Google search data ☆

Seth Stephens-Davidowitz ✉

Highlights

- Google search data is used as a new measure of racial animus in the United States.
- In places where racial animus is highest, Obama did worse than comparable Democratic candidates.
- Racial animus cost Obama about 4 percentage points, giving his opponent the equivalent of a home state advantage.
- Analysis with Google searches suggest racism cost Obama substantially more votes than survey-based analysis.

What characterizes data?

"[T]echnologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable."

Ursula Franklin, 1989

Confounded

"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

Charles Goodhart, 1975

What characterizes data?

Confounded

Facebook Creates Pages for Businesses Whether They Like It or Not

And they won't take them down unless you send them a bunch of documents

Another way Facebook auto-generates pages is by using the information available on Wikipedia. Such a page was created for Justin Cappos, an associate professor of computer science and engineering at New York University Tandon School of Engineering. Before Facebook imported the content from his Wikipedia page, Cappos says he had never had a social media account because he thinks they don't offer a positive usability-privacy tradeoff.

What characterizes data?

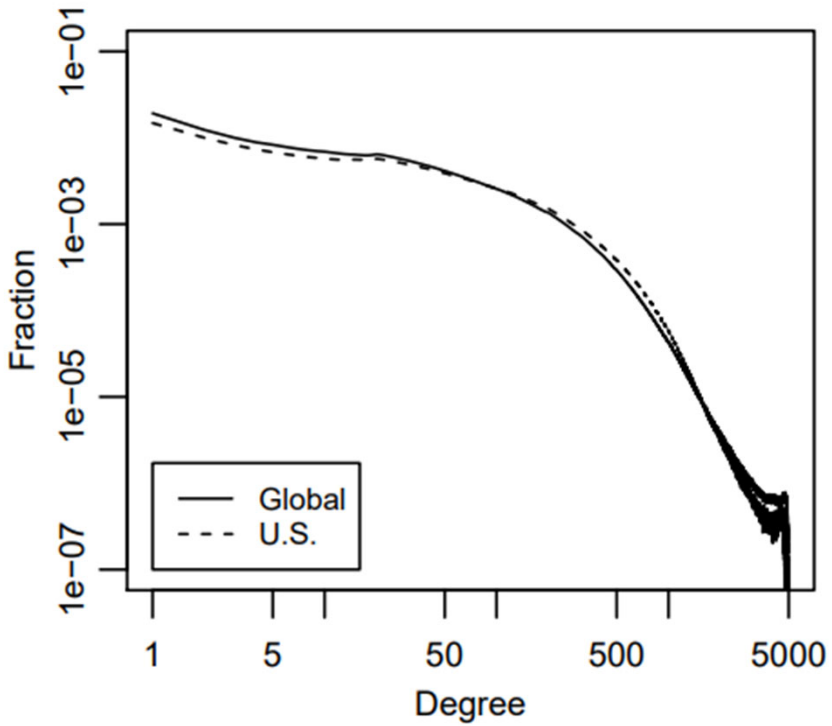
Confounded

Facebook Creates Pages for Businesses Whether They Like It or Not

And they won't take them down unless you send them a bunch of documents

"I frankly do not want to have a Facebook presence of any kind," Cappos told me in a recent interview. "The company has proven itself a poor steward of the information it has collected (both ethically and unethically). While the data collected by Facebook is completely legal for them to take off of Wikipedia, it is certainly not a desirable outcome from my standpoint."

What characterizes data?

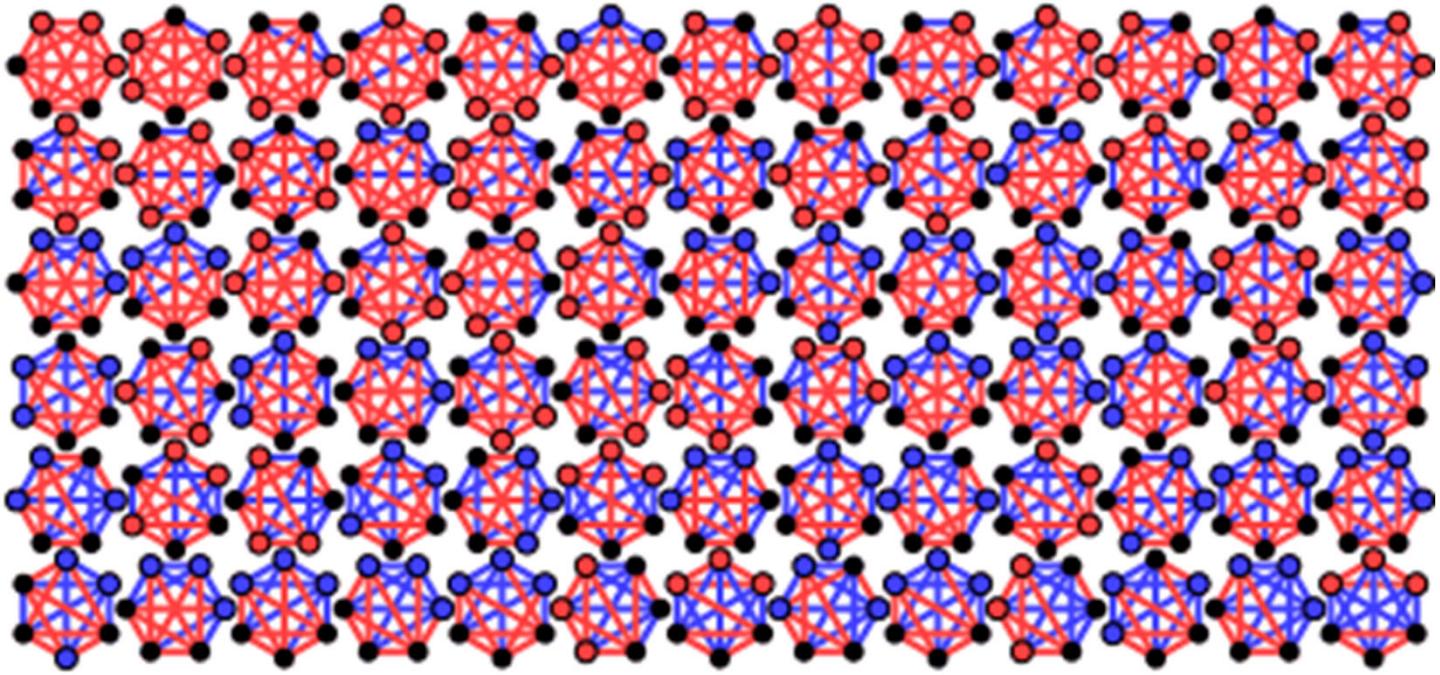


Confounded

for a small anomaly near 20 friends. This kink is due to forces within the Facebook product to encourage low friend count individuals in particular to gain more friends until they reach 20 friends. The distribution shows a clear cutoff at 5000 friends, a limit imposed by Facebook on the number of friends at the time

Exercise

- Using Graphs with Colored Edges to Detect Patterns in Networks

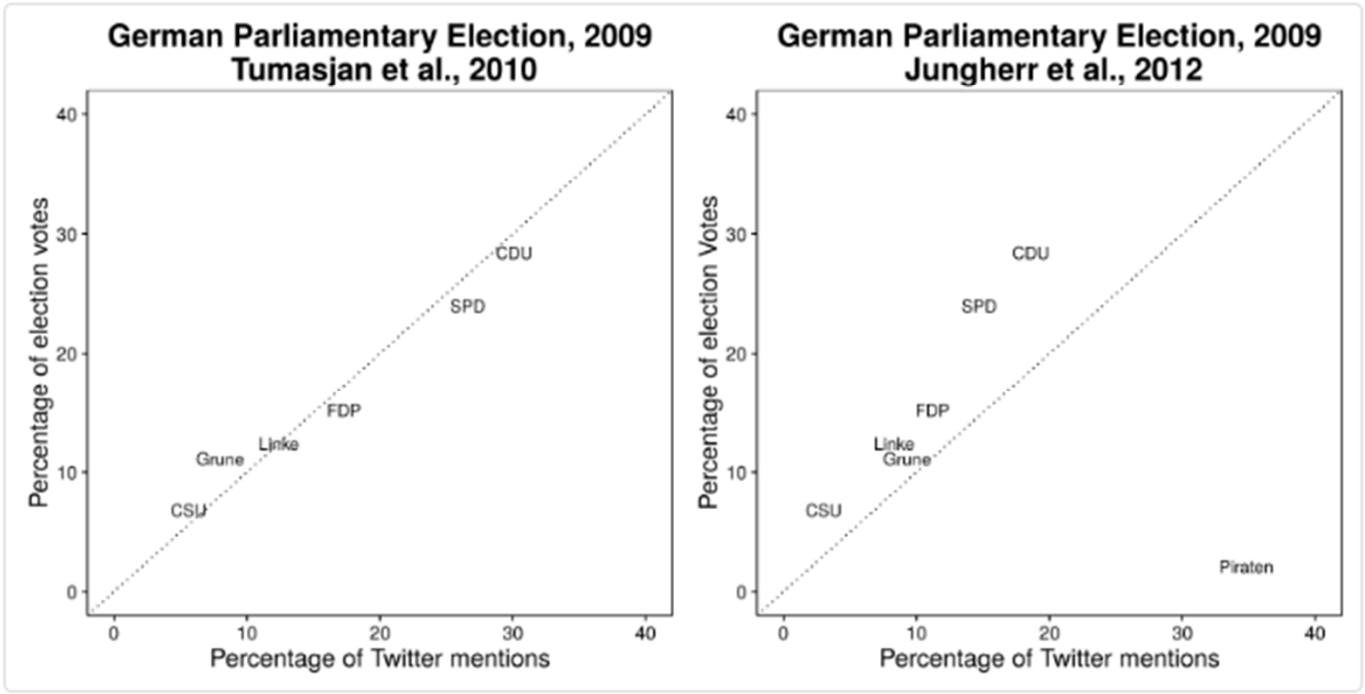


What characterizes data?

"These problems are, and will probably ever remain, among the inscrutable secrets of nature. They belong to a class of questions radically inaccessible to the human intelligence."—The Times of London, September 1849, on how cholera is contracted and spread

Not Representative

- Sample
- Population



What characterizes data?

Messy

Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures

Scott A. Golder*, Michael W. Macy

Abstract

We identified individual-level diurnal and seasonal mood rhythms in cultures across the globe, using data from millions of public Twitter messages. We found that individuals awaken in a good mood that deteriorates as the day progresses—which is consistent with the effects of sleep and circadian rhythm—and that seasonal change in baseline positive affect varies with change in daylength. People are happier on weekends, but the morning peak in positive affect is delayed by 2 hours, which suggests that people awaken later on weekends.

What characterizes data?

Messy

"Reboot NT machine [name] in cabinet [name] at [location]:CRITICAL:[date and time]"

D Land Seasonal Mood Vary with Work, Sleep,
Across Diverse Cultures

Scott A. Golder

Abstract

We identified individual-level diurnal cultures across the globe, using data from millions of public Twitter individuals awaken in a good mood that deteriorates as the day progresses with the effects of sleep and circadian rhythm—and that seasonal changes positive affect varies with change in daylength. People are happier on weekends, but the morning peak in positive affect is delayed by 2 hours, which suggests that people awaken later on weekends.

What characterizes data?

Reuters UK

Turkey escalates crackdown on dissent six years after Gezi protests

Turkey escalates crackdown on dissent six years after Gezi protests ... were inspired by the worldwide “Occupy” protests and Arab uprisings ...

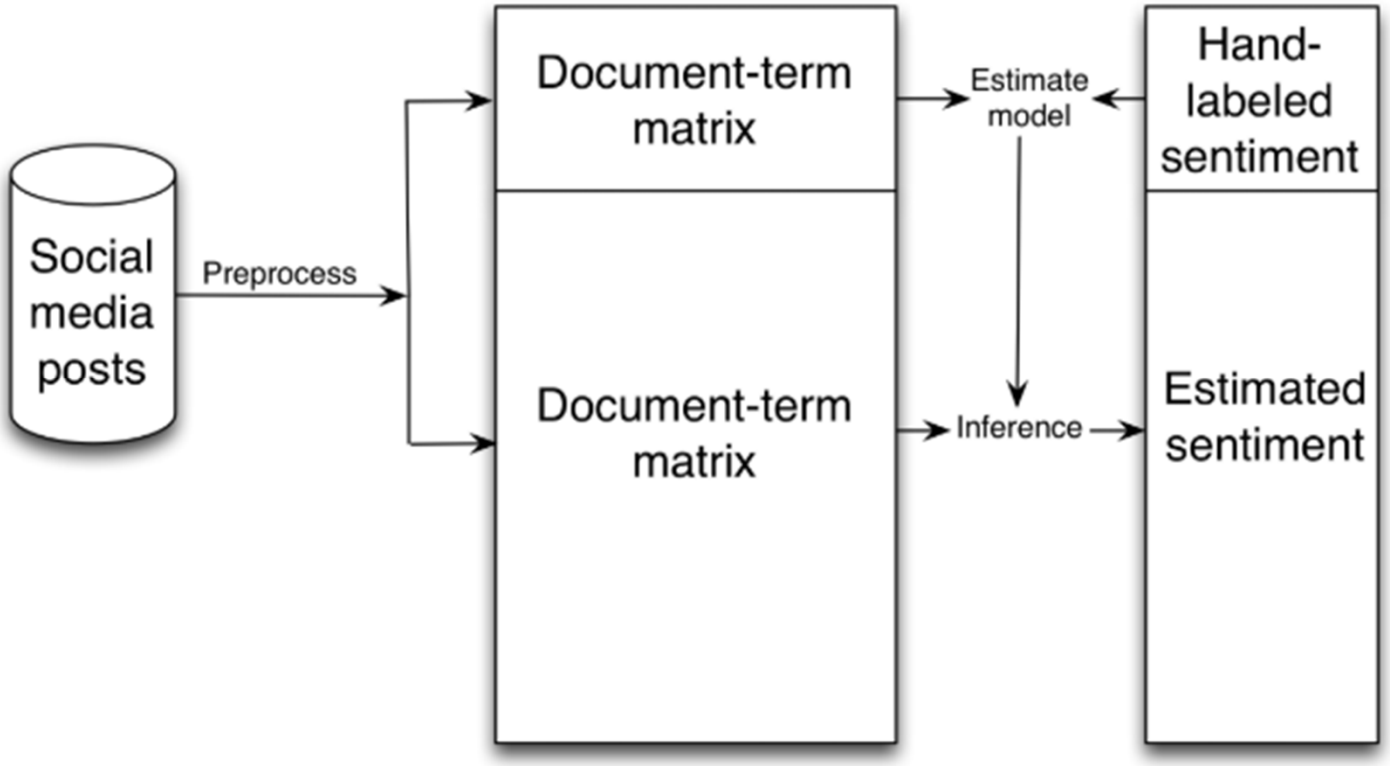
Mar 18, 2019

Drift

“What had happened was that as soon as the protest became the dominant story, large numbers of people … stopped using the hashtags except to draw attention to a new phenomenon … While the protests continued, and even intensified, the hashtags died down. Interviews revealed two reasons for this. First, once everyone knew the topic, the hashtag was at once superfluous and wasteful on the character-limited Twitter platform. Second, hashtags were seen only as useful for attracting attention to a particular topic, not for talking about it.”

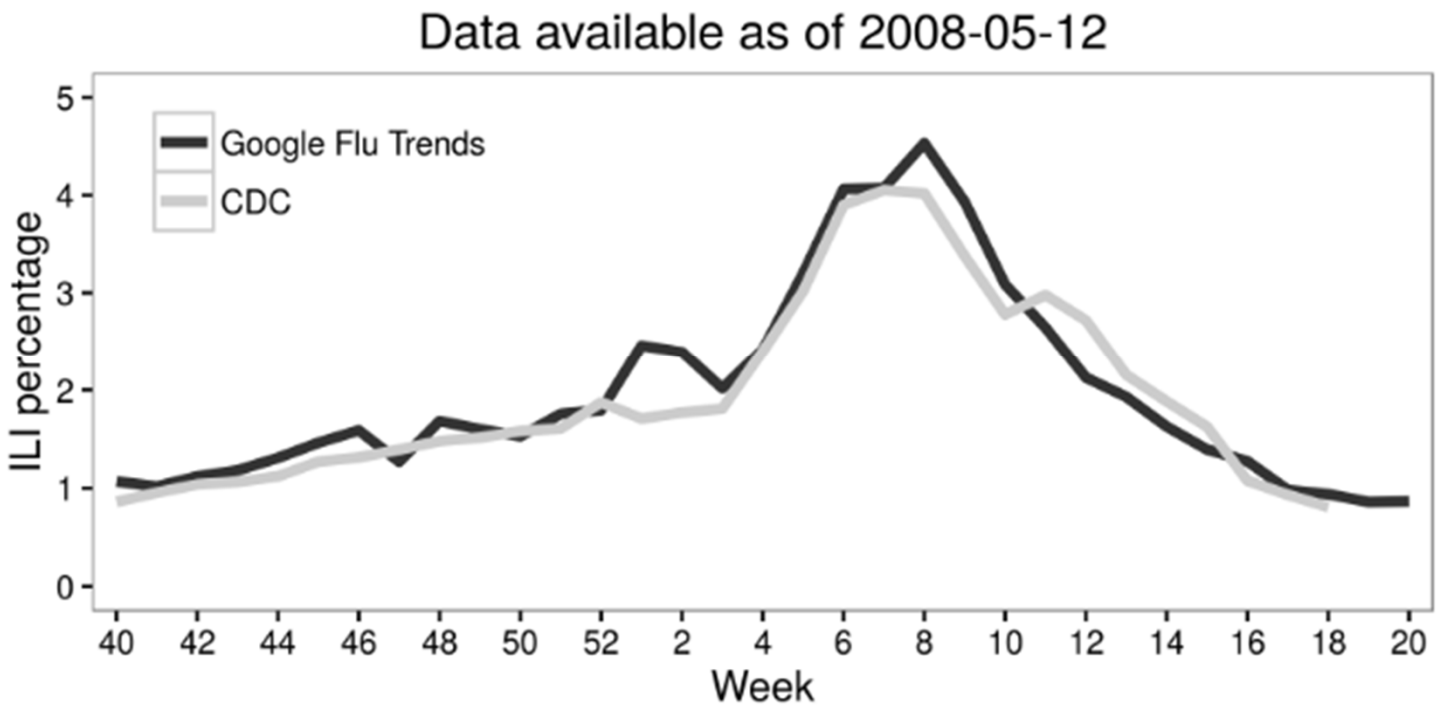
What is Data Science?

Estimating

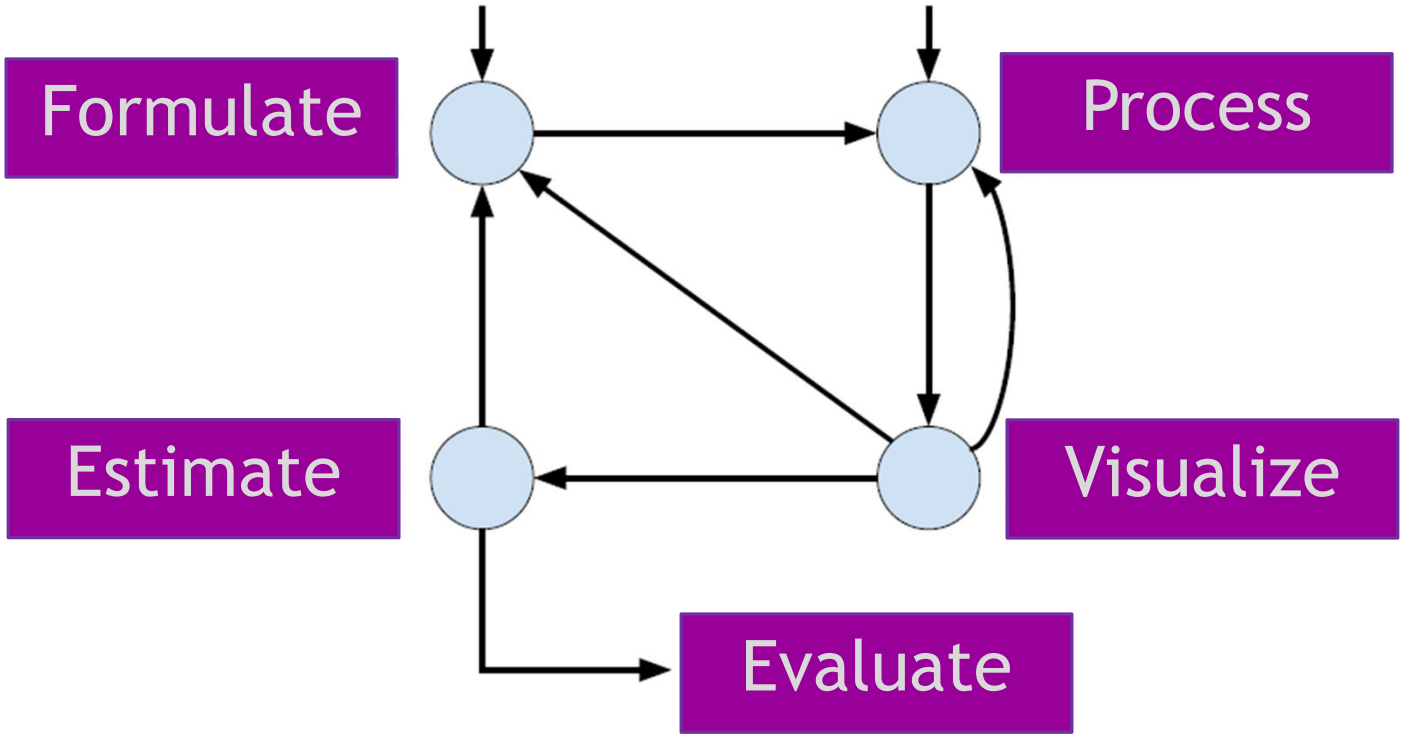


What is Data Science?

Predicting



What is data science?



Exercise

Data Source

All names are from Social Security card applications for births that occurred in the United States after 1879. Note that many people born before 1937 never applied for a Social Security card, so their names are not included in our data. For others who did apply, our records may not show the place of birth, and again their names are not included in our data.

All data are from a 100% sample of our records on Social Security card applications as of March 2019.

Exercise

Data qualifications

People using our data on popular names are urged to explicitly acknowledge the following qualifications.

1. Names are restricted to cases where the year of birth, sex, State of birth (50 States and District of Columbia) are on record, and where the given name is at least 2 characters long.
2. Name data are tabulated from the "First Name" field of the Social Security Card Application. Hyphens and spaces are removed, thus Julie-Anne, Julie Anne, and Julieanne will be counted as a single entry.
3. Name data are not edited. For example, the sex associated with a name may be incorrect. Entries such as "Unknown" and "Baby" are not removed from the lists.
4. Different spellings of similar names are not combined. For example, the names Caitlin, Caitlyn, Kaitlin, Kaitlyn, Kaitlynn, Katelyn, and Katelynn are considered separate names and each has its own rank.
5. When two different names are tied with the same frequency for a given year of birth, we break the tie by assigning rank in alphabetical order.
6. Some names are applied to both males and females (for example, Micah). Our rankings are done by sex, so that a name such as Micah will have a different rank for males as compared to females. When you seek the popularity of a specific name (see "Popularity of a Name"), you can specify the sex. If you do not specify the sex, we provide rankings for the more popular name-sex combination.
7. To safeguard privacy, we exclude from our tabulated lists of names those that would indicate, or would allow the ability to determine, names with fewer than 5 occurrences in any geographic area. If a name has less than 5 occurrences for a year of birth in any state, the sum of the state counts for that year will be less than the national count.

Questions

- Questions on Piazza?
- Question for You!

  Should students have access to the class roster?