# What is data science?

## Know it when you see it

Prof. Bisbee

Vanderbilt University

Lecture Date: 2022/08/24

Slides Updated: 2022-08-24

# Agenda

1. Meet the instructor
    - Prof. Bisbee: *james.h.bisbee@vanderbilt.edu*
2. Course Motivation
    - What is data science (DS) & why should we care?
3. Course Objectives
    - **Content:** Critical thinking, analysis, presentation
    - **Skills:** Computing and analysis in R
4. Course Expectations & Syllabus review

# Meet the instructor

- PhD from NYU Politics in 2019
- Postdocs at Princeton Niehaus & NYU CSMaP
- Published some things
  - Methods-ey: external validity 1, 2; measurement 3, 4
  - Substantive: economics & populism 1; Covid-19 & U.S. politics 2, 3; IPE 4; academic naval-gazing 5
- Popular press
  - Monkey Cage articles 1, 2
  - Podcast / Radio interviews

# Meet the instructor

- Current research
  - YouTube + polarization
  - Twitter + misinformation
  - Telegram + white supremacists
- Is my current research agenda data science?

# What is "data science"?

- What is data?
- What is science?

# What is data?

- "It is a capital mistake to theorize before one has data." Sherlock Holmes
  - Data **informs**
- "Torture the data, and it will confess to anything." Ronald Coase, Nobel Prize Laureate in Economics
  - Data **lies**
- “Here’s an open secret of the big data world: all data is dirty. All of it.” Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*
  - Data is **invalid**

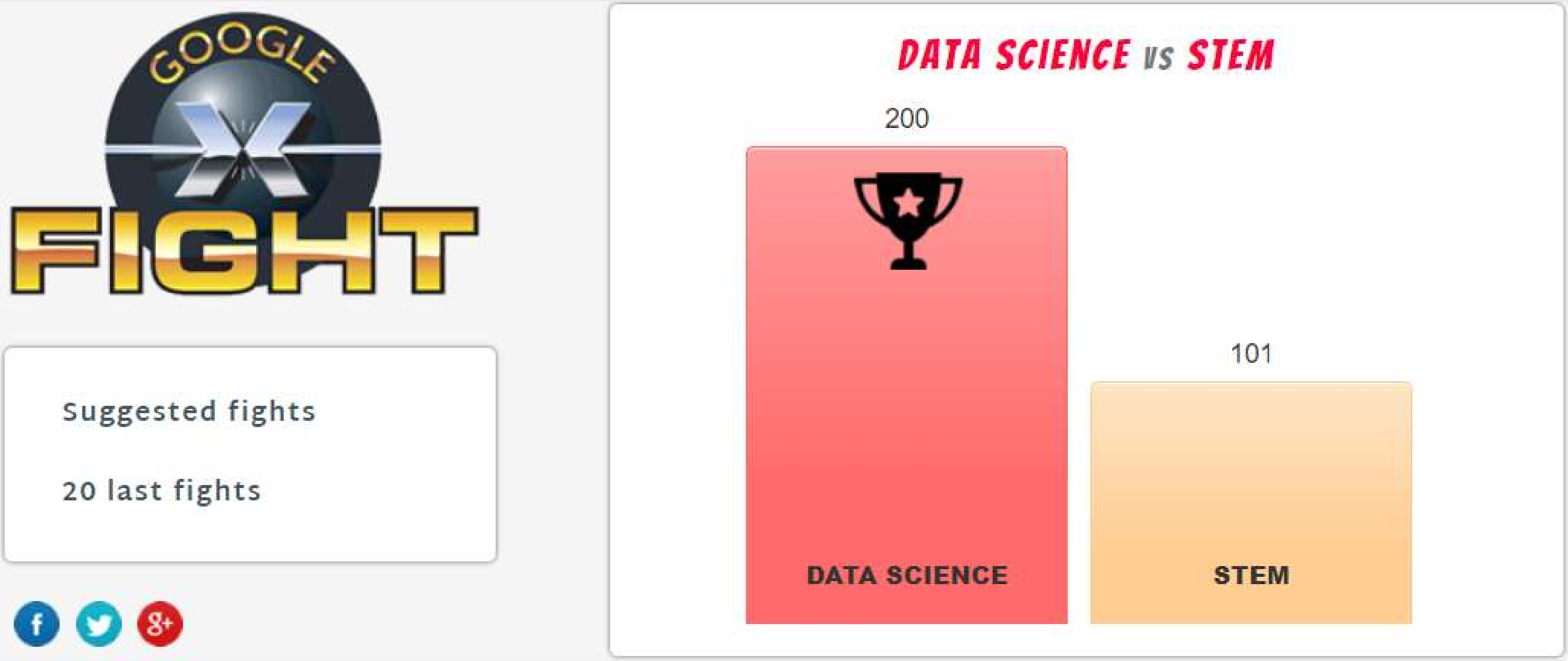
# What is science?

- Simplification, codification, abstraction
  - Science identifies patterns in data...
  - ...to make predictions about the future
- As such, it is inherently:
  - Causal
  - Empirical
  - Theoretical

# What is data science?

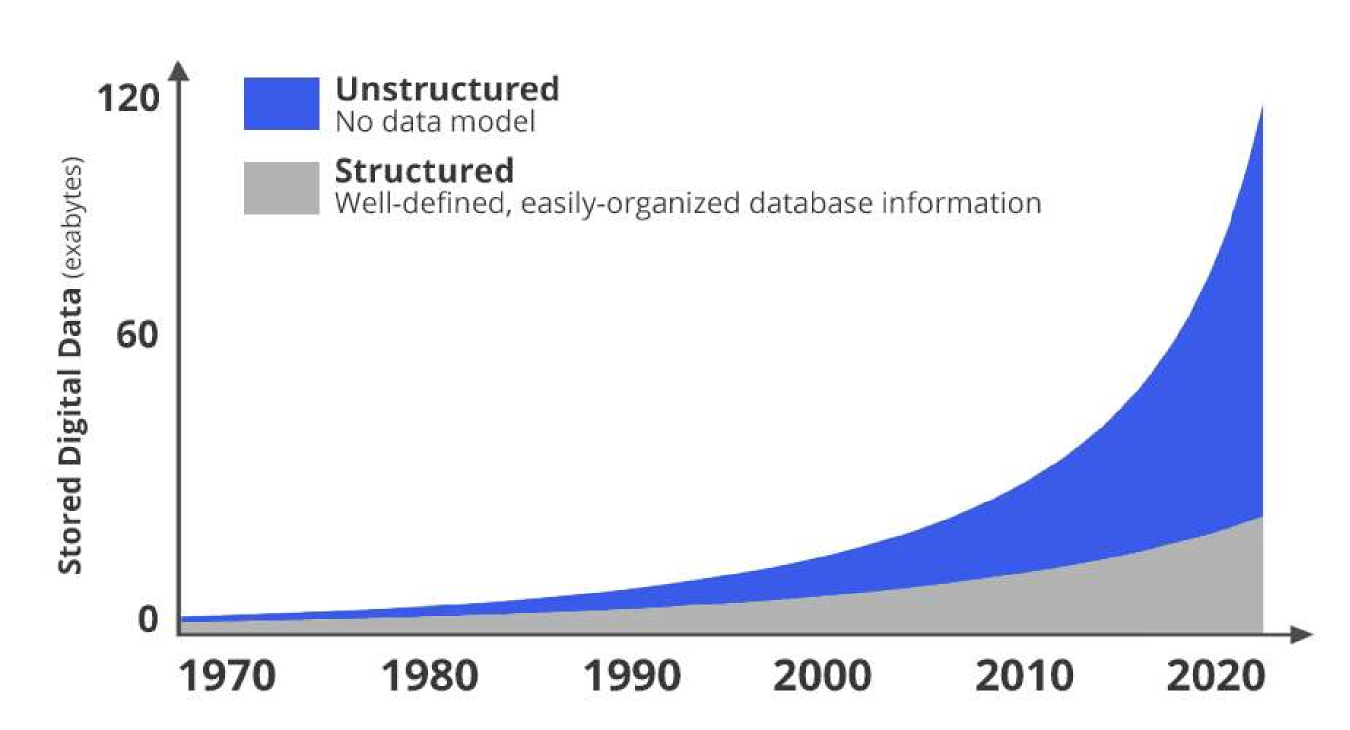
- Data: informs / lies / invalid
- Science: simplification / codification / abstraction
- Data + science = ?

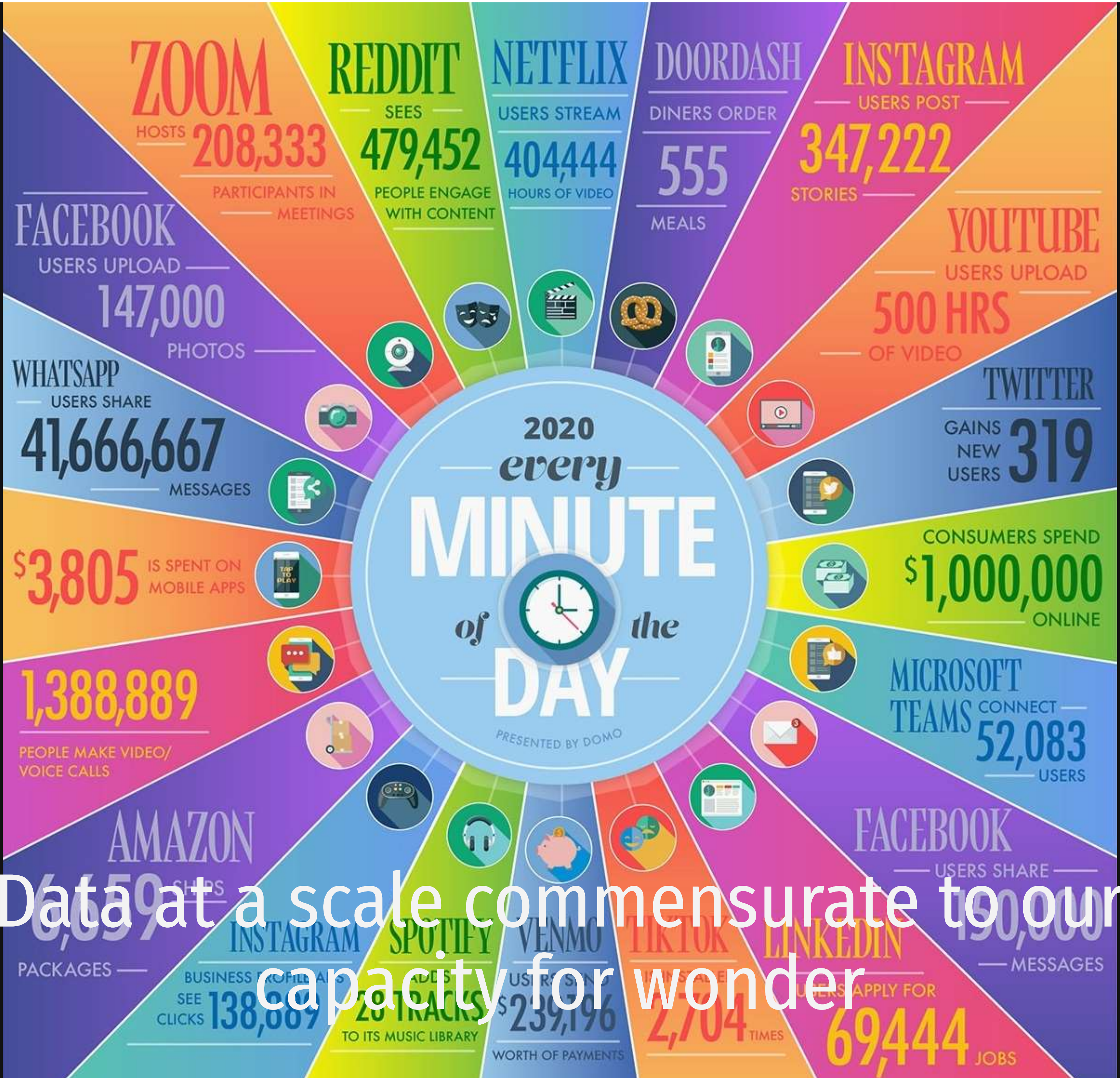
# Why are you here?



# Is this all just a fad?

- No

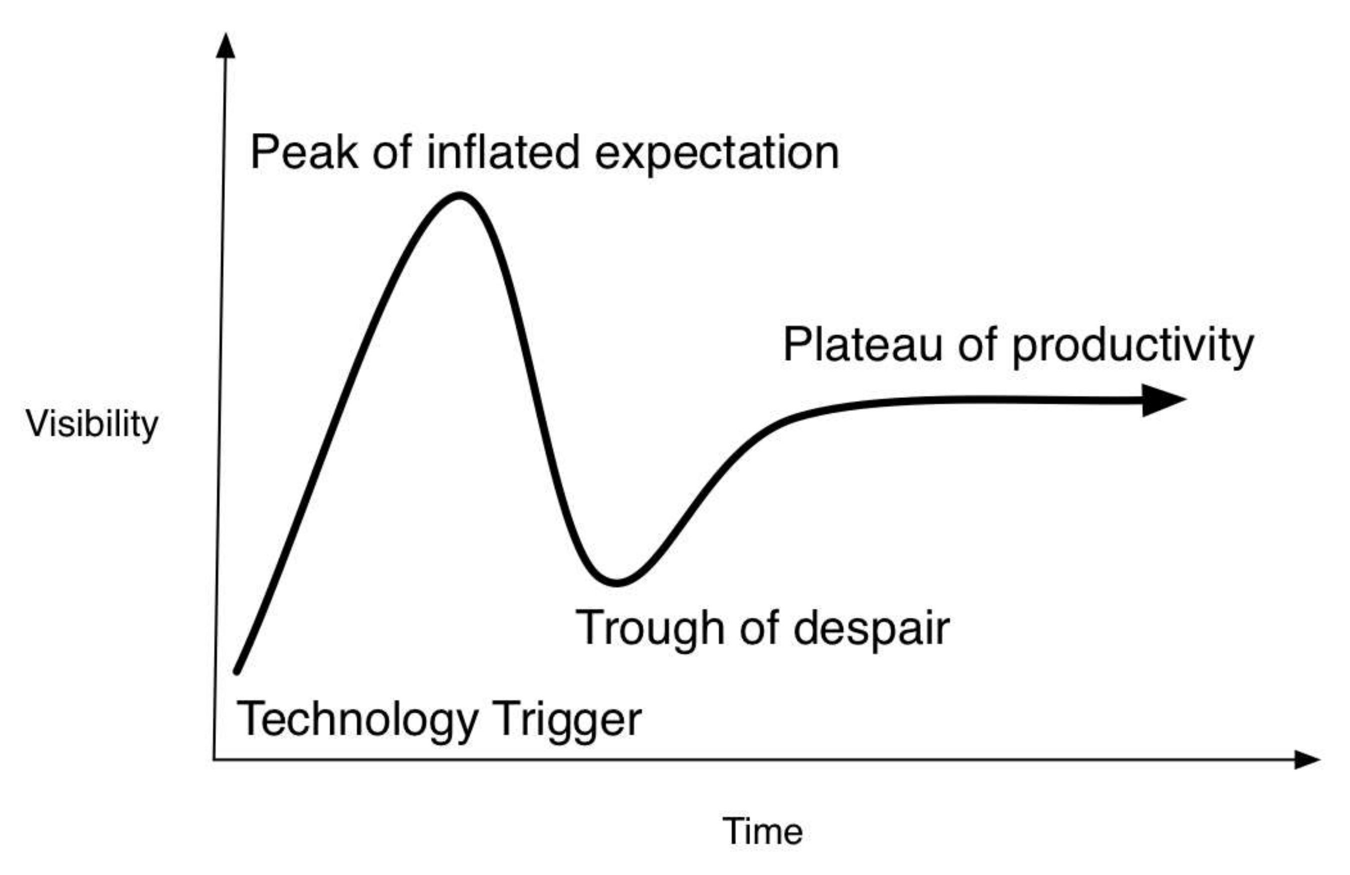




Data at a scale commensurate to our capacity for wonder

# Is this all just a fad?

- But there are faddish qualities



# Wait so WHAT is data science?

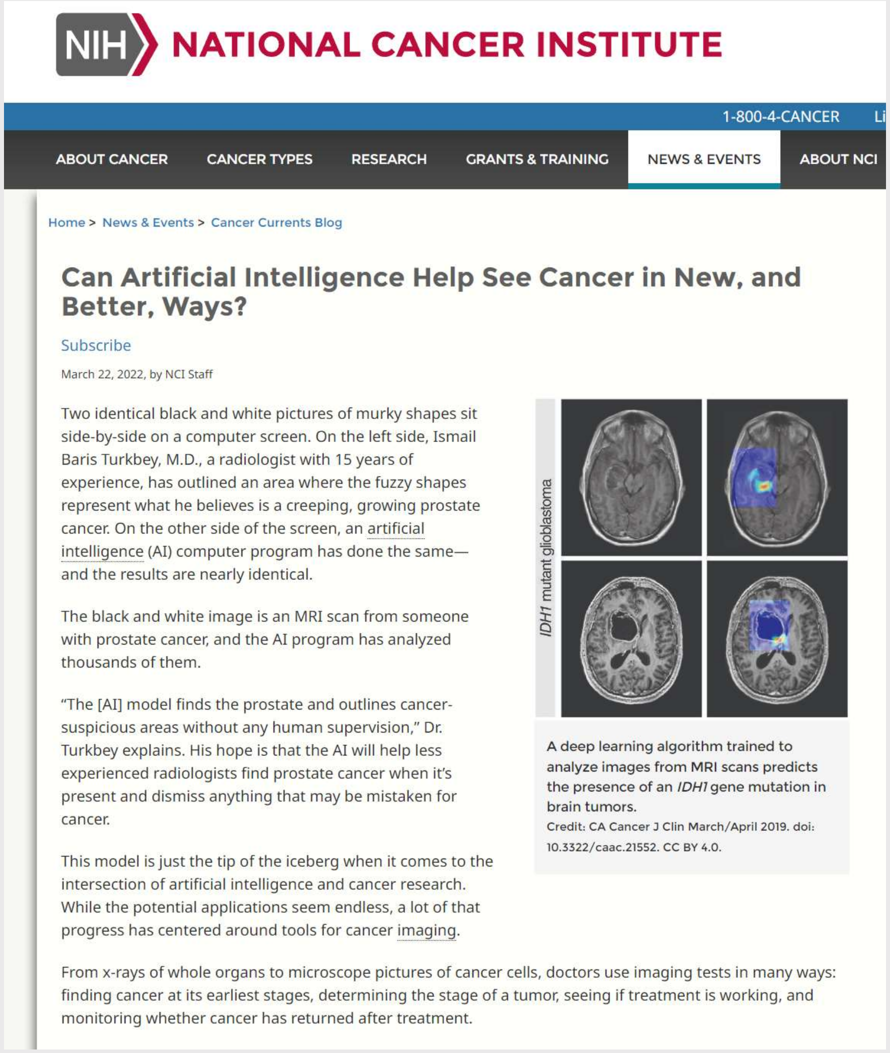
- A series of examples
- Data science is for **everybody**

# Historians: Identify Shakespear

- Use original texts written by Shakespeare and Marlowe (among others)
- Apply natural language processing (NLP) to characterize styles of writing
- Demonstrate that Shakespeare was at least heavily influenced by collaborators

# Biologists: Identify Cancer

- Use x-rays of patients
- Apply image analysis to identify cancerous areas
- Reproduce expert analysis, facilitating early detection



## Can Artificial Intelligence Help See Cancer in New, and Better, Ways?

March 22, 2022, by NCI Staff

Two identical black and white pictures of murky shapes sit side-by-side on a computer screen. On the left side, Ismail Baris Turkbey, M.D., a radiologist with 15 years of experience, has outlined an area where the fuzzy shapes represent what he believes is a creeping, growing prostate cancer. On the other side of the screen, an artificial intelligence (AI) computer program has done the same—and the results are nearly identical.

The black and white image is an MRI scan from someone with prostate cancer, and the AI program has analyzed thousands of them.

"The [AI] model finds the prostate and outlines cancer-suspicious areas without any human supervision," Dr. Turkbey explains. His hope is that the AI will help less experienced radiologists find prostate cancer when it's present and dismiss anything that may be mistaken for cancer.

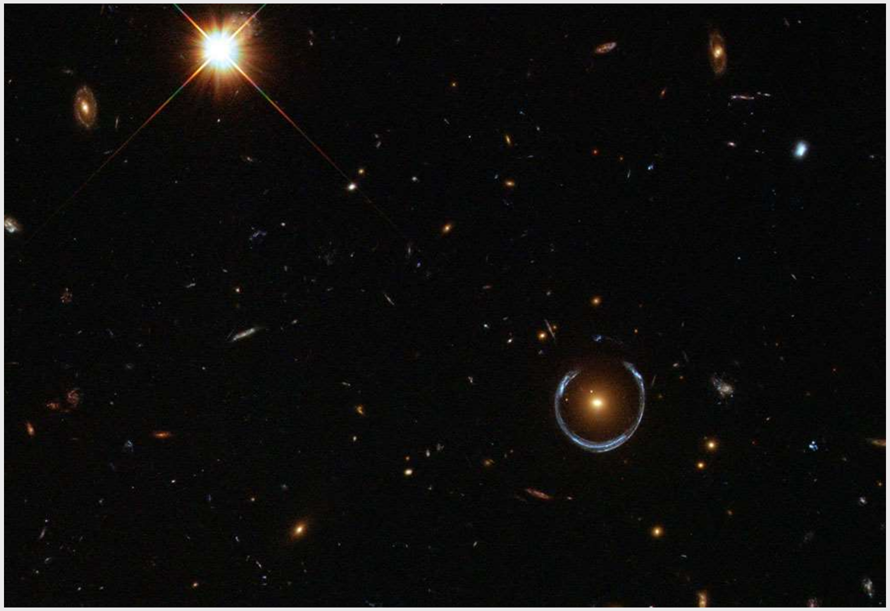
A deep learning algorithm trained to analyze images from MRI scans predicts the presence of an *IDH1* gene mutation in brain tumors.
Credit: CA Cancer J Clin March/April 2019. doi: 10.3322/caac.21552. CC BY 4.0.

This model is just the tip of the iceberg when it comes to the intersection of artificial intelligence and cancer research. While the potential applications seem endless, a lot of that progress has centered around tools for cancer imaging.

From x-rays of whole organs to microscope pictures of cancer cells, doctors use imaging tests in many ways: finding cancer at its earliest stages, determining the stage of a tumor, seeing if treatment is working, and monitoring whether cancer has returned after treatment.

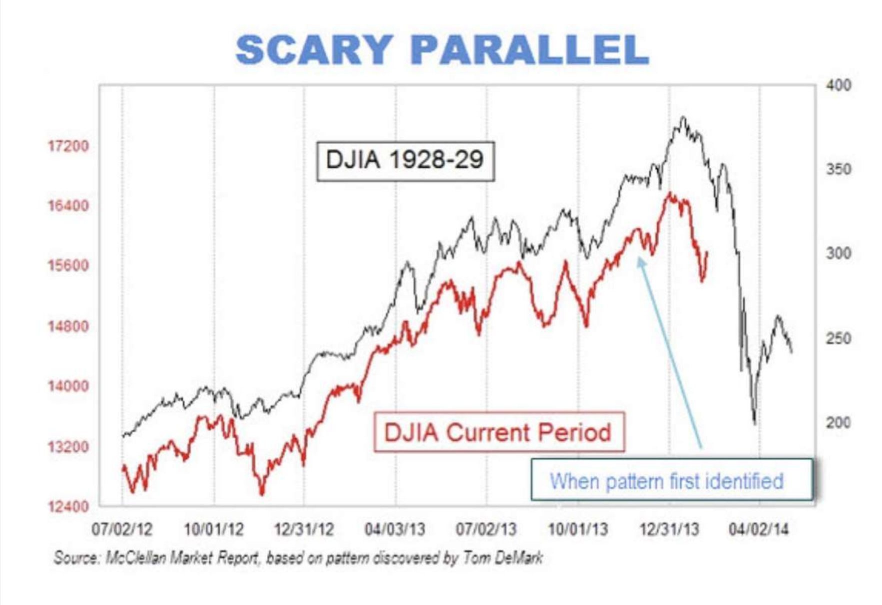
# Astronomers: Detect Dark Matter

- Use satellite photos of deep space
- Apply machine learning to detect gravitational lensing
- Streamline analysis



# Economists: Predict stock prices

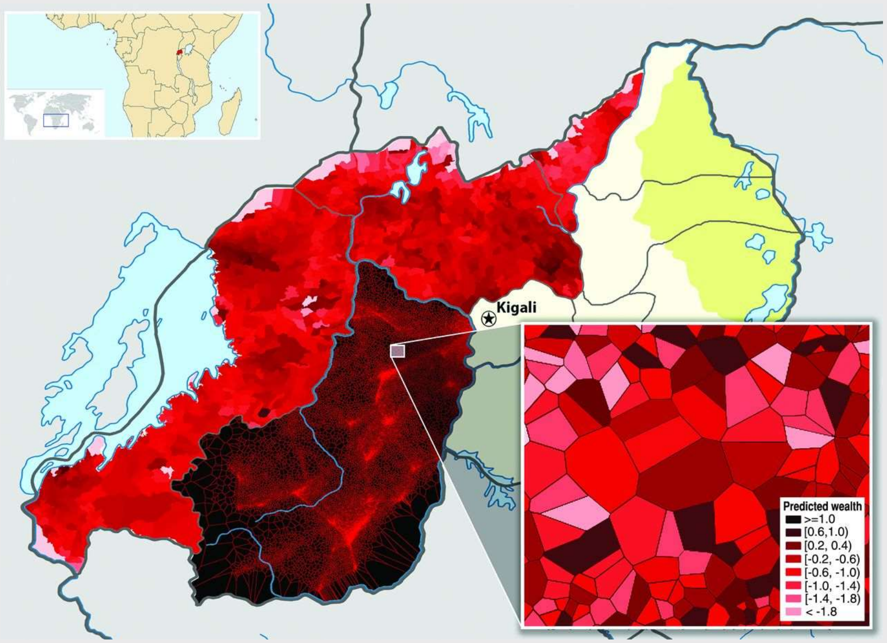
- Use time series data of stock prices
- Apply Long Short Term Memory Networks (LSTM) to predict future prices
- Make KEE$SH!!



Source: McClellan Market Report, based on pattern discovered by Tom DeMark

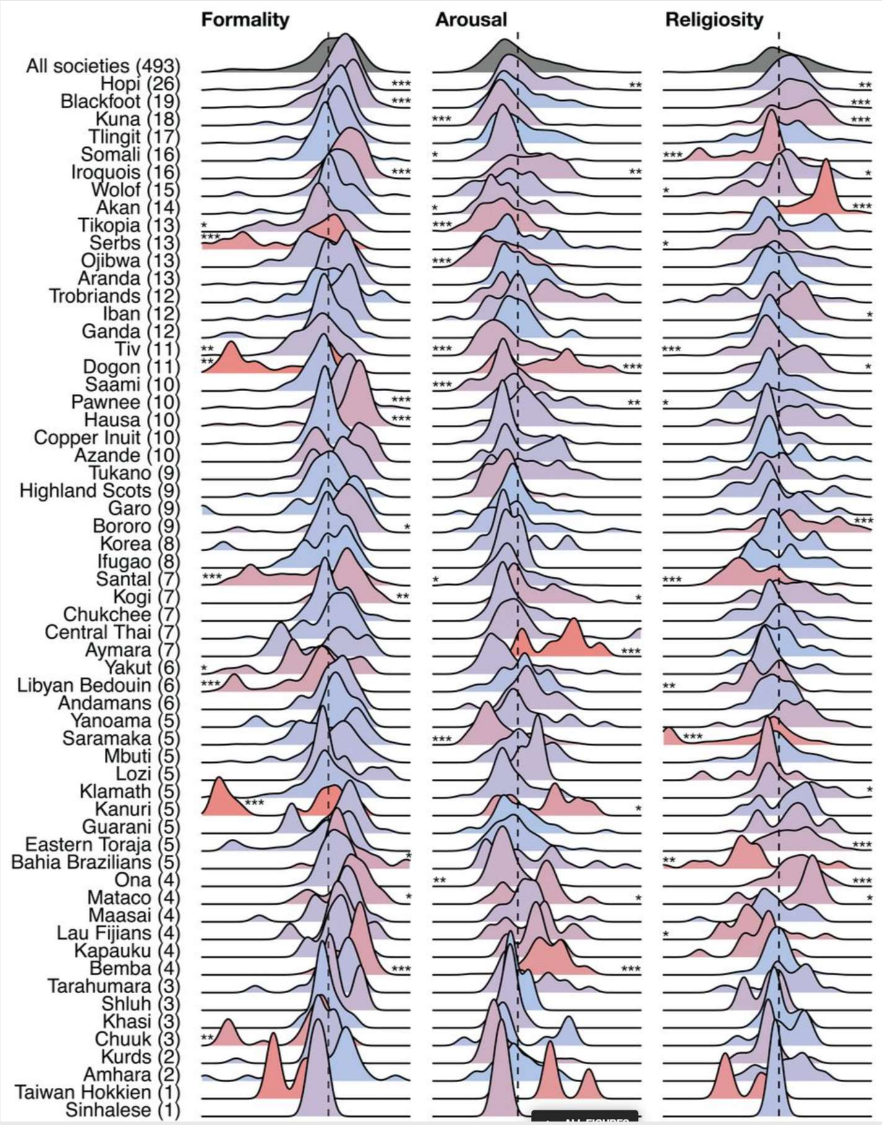
# Social Scientists: Measure Poverty

- Use cell phone data
- Apply machine learning to learn relationships between calling and wealth
- Empower aid agencies around the globe



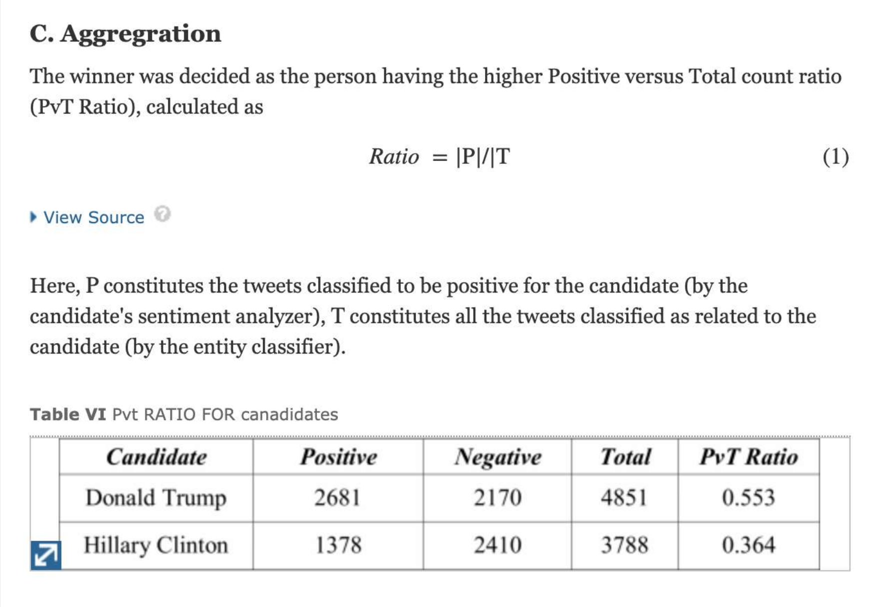
# Musicologists: Describe Music

- Use audio recordings and ethnographic labels
- Apply factor analysis to distill labels to three dimensions
- Bring the world closer together / anger traditional musicologists



# Political Scientists: Predict Polls

- Use tweets written by candidates
- Apply basic algebra to predict winner
- Start a blog



**C. Aggregration**

The winner was decided as the person having the higher Positive versus Total count ratio (PvT Ratio), calculated as

$$Ratio\ = |\mathrm{P}|/|\mathrm{T} \tag{1}$$

Here, P constitutes the tweets classified to be positive for the candidate (by the candidate's sentiment analyzer), T constitutes all the tweets classified as related to the candidate (by the entity classifier).

**Table VI** Pvt RATIO FOR canadidates

| *Candidate* | *Positive* | *Negative* | *Total* | *PvT Ratio* |
|---|---|---|---|---|
| Donald Trump | 2681 | 2170 | 4851 | 0.553 |
| Hillary Clinton | 1378 | 2410 | 3788 | 0.364 |

# WHAT IS DATA SCIENCE?!

- How is data science different from science that uses data?
- Readymade versus custommade

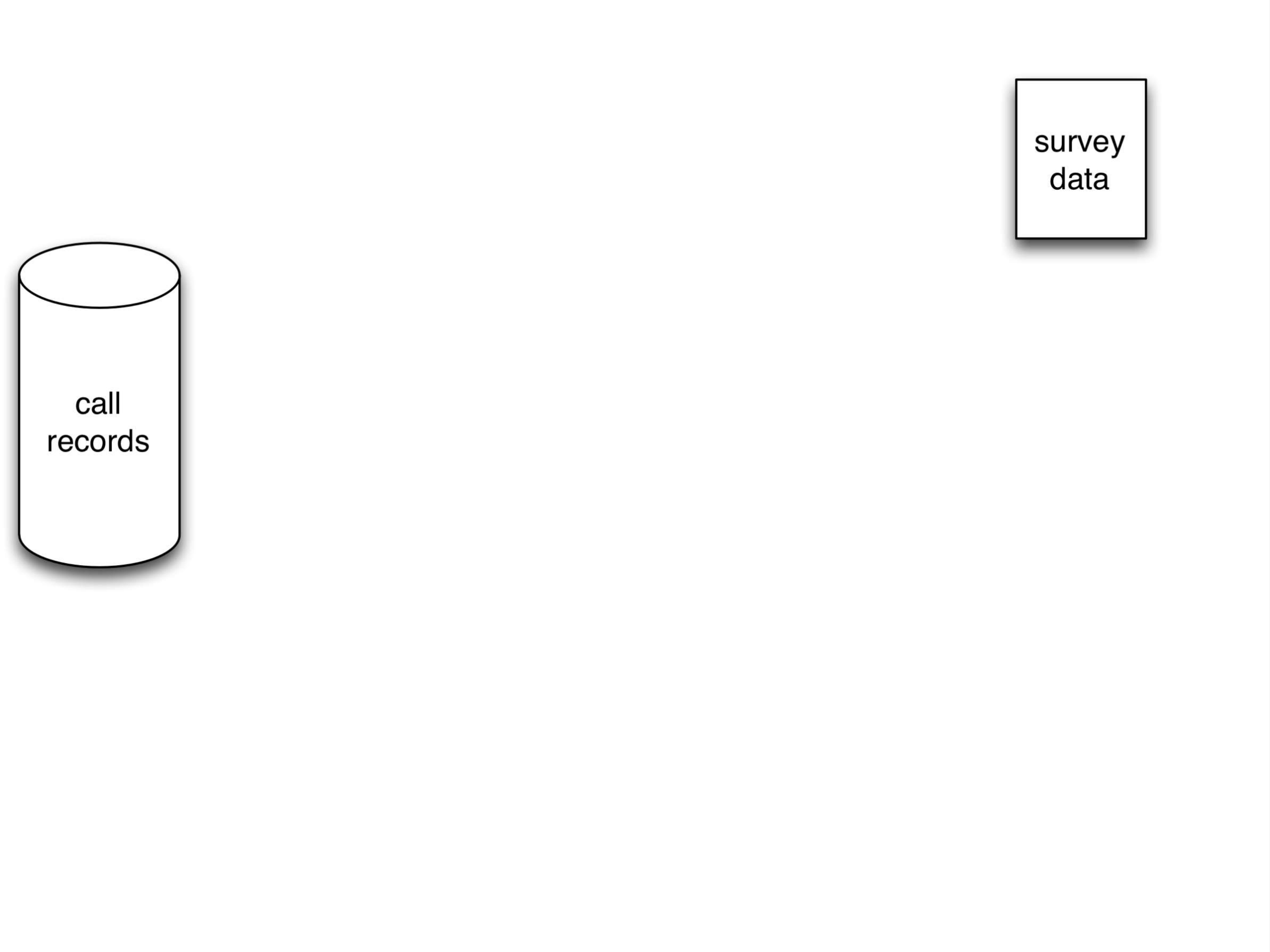
# Poverty Measure Example

Predicting poverty and wealth from mobile phone metadata

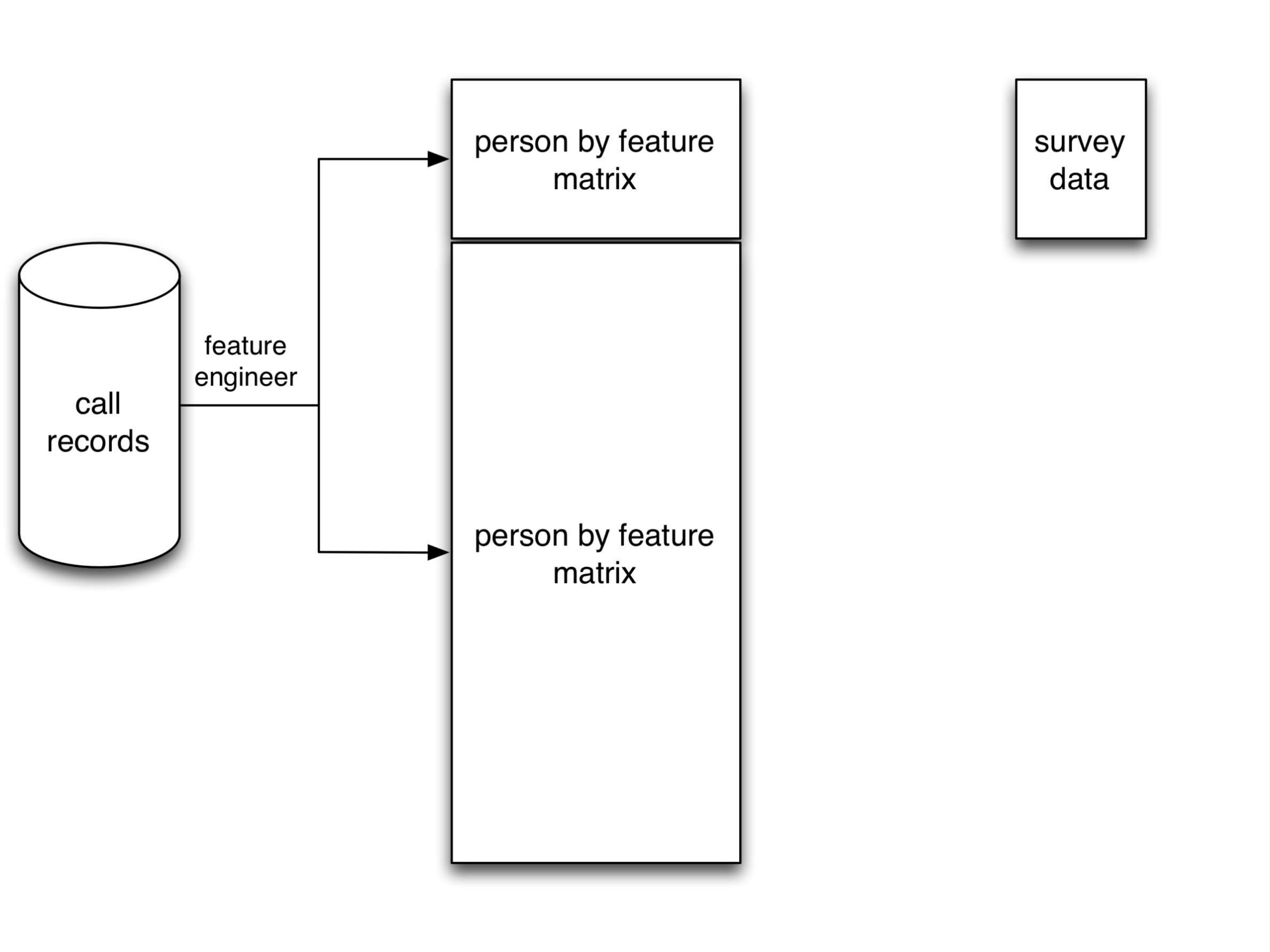
Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

# Analysis

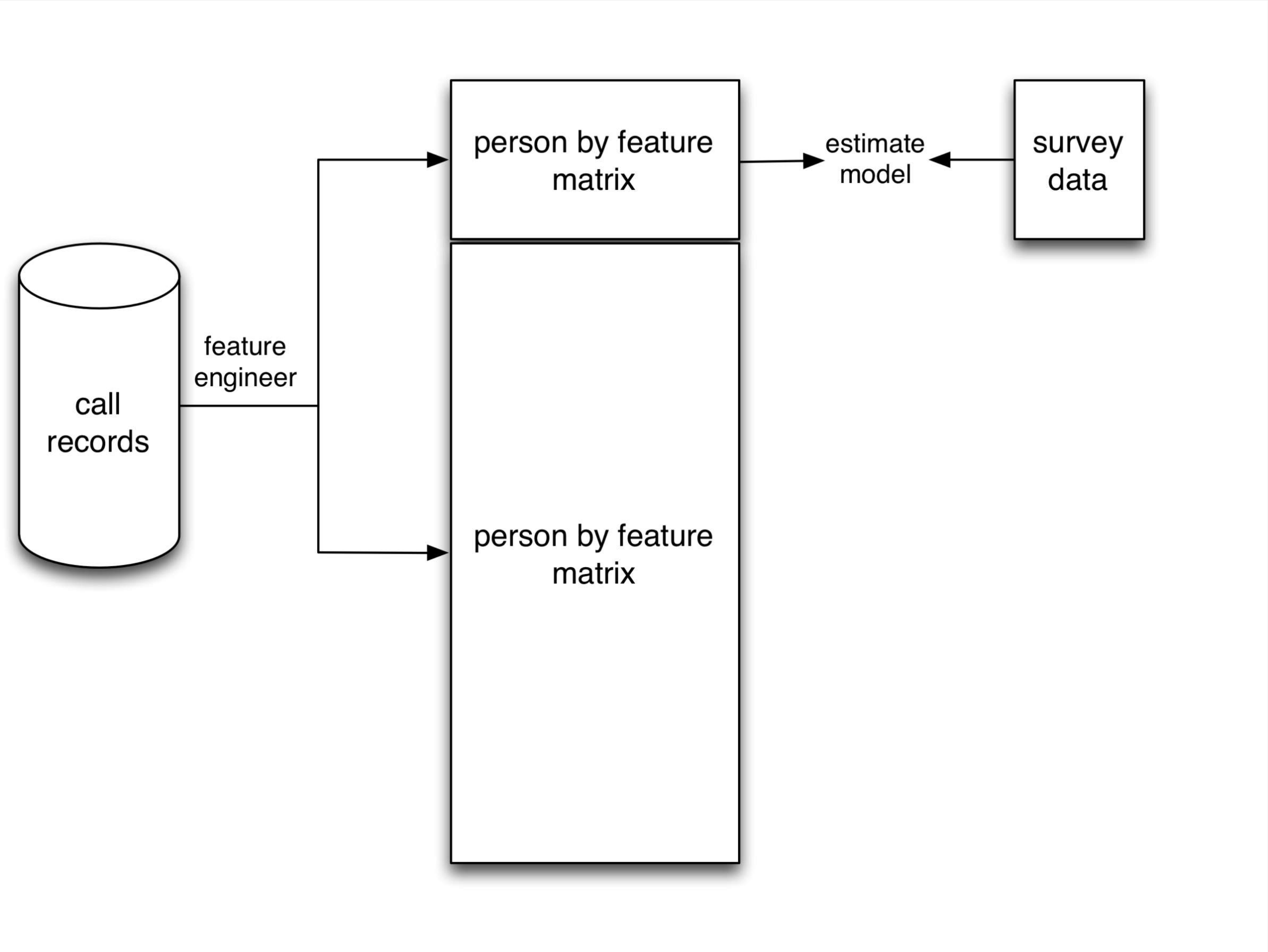
# Analysis



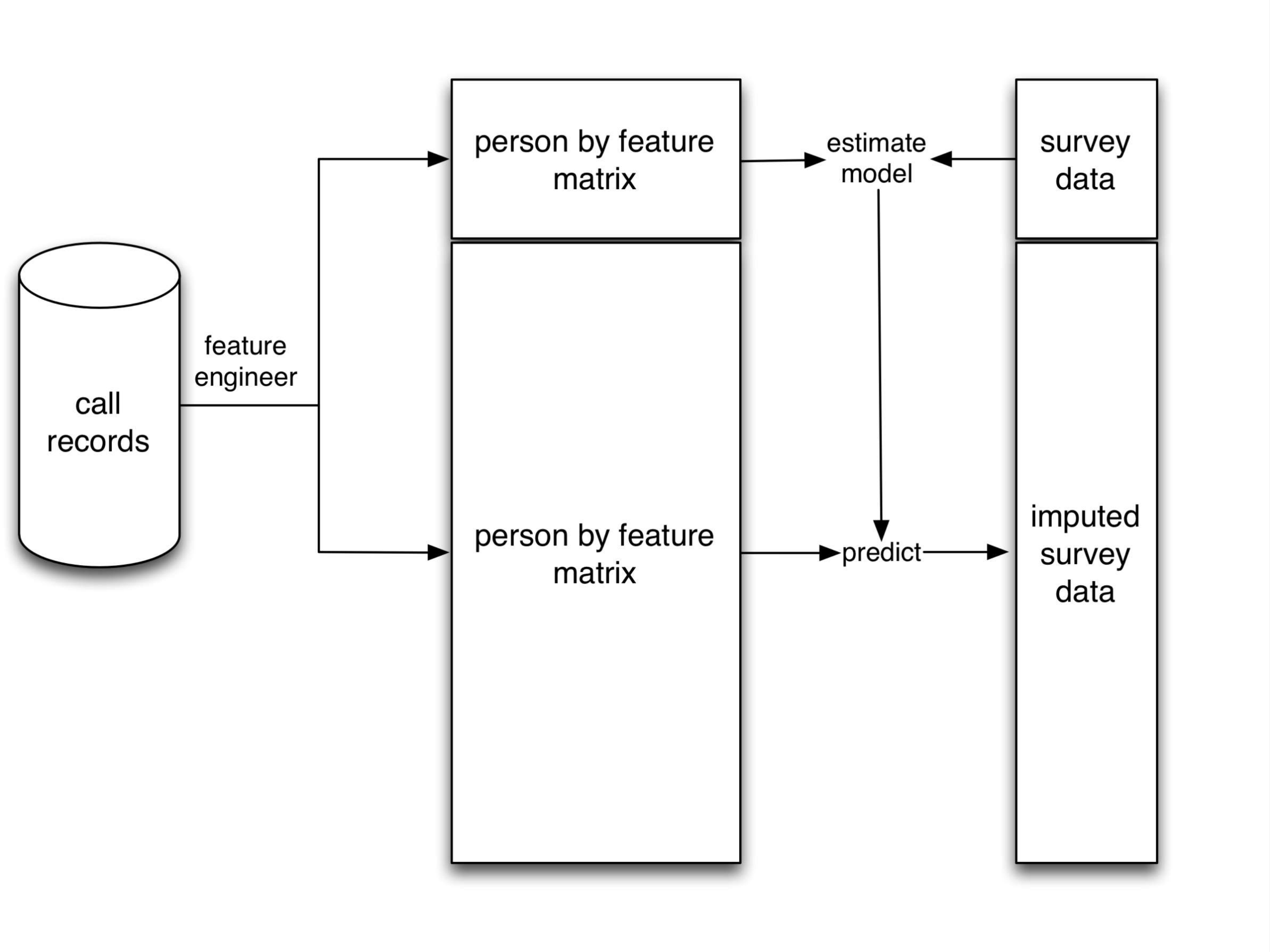
# Analysis


person by feature matrix
survey data
call records
feature engineer
person by feature matrix

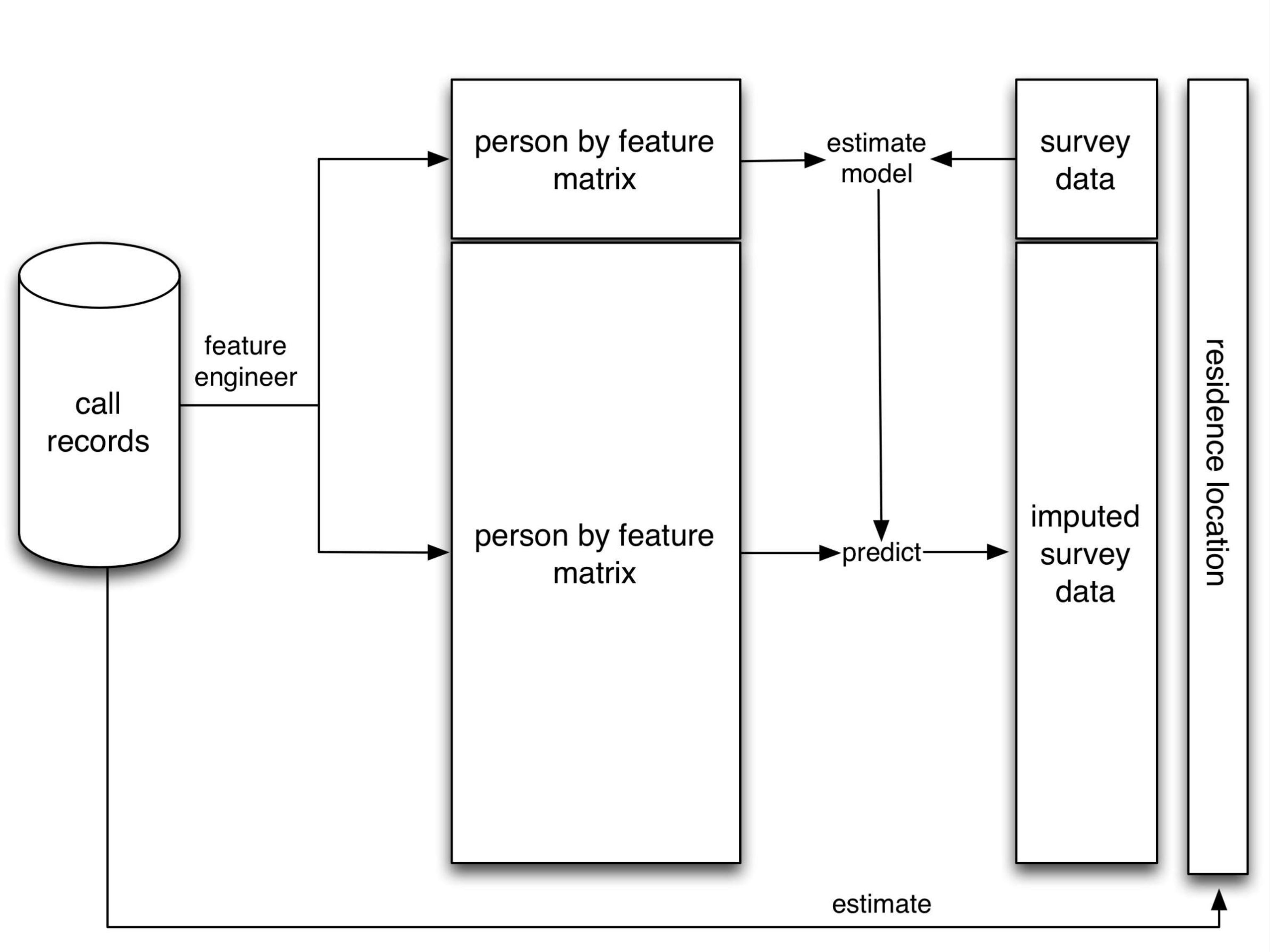
# Analysis


call
records
feature
engineer
person by feature
matrix
estimate
model
survey
data
person by feature
matrix

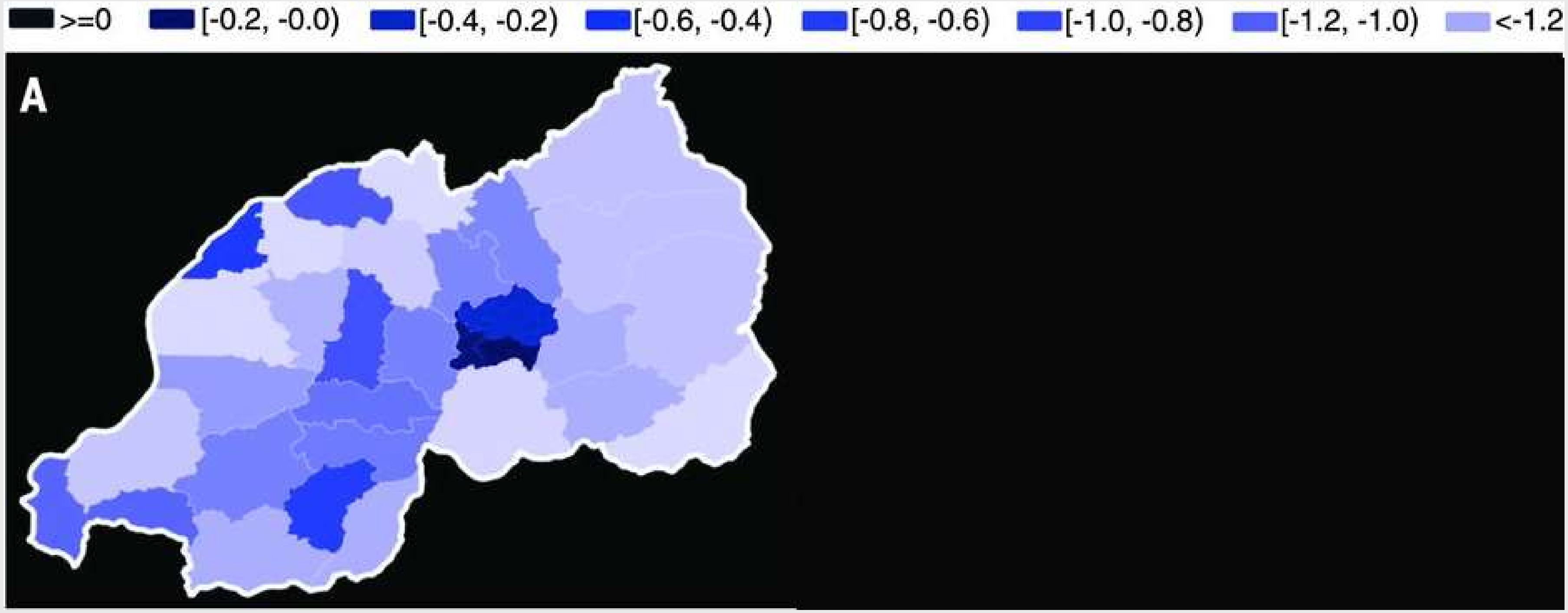
# Analysis


call records
feature engineer
person by feature matrix
person by feature matrix
estimate model
predict
survey data
imputed survey data

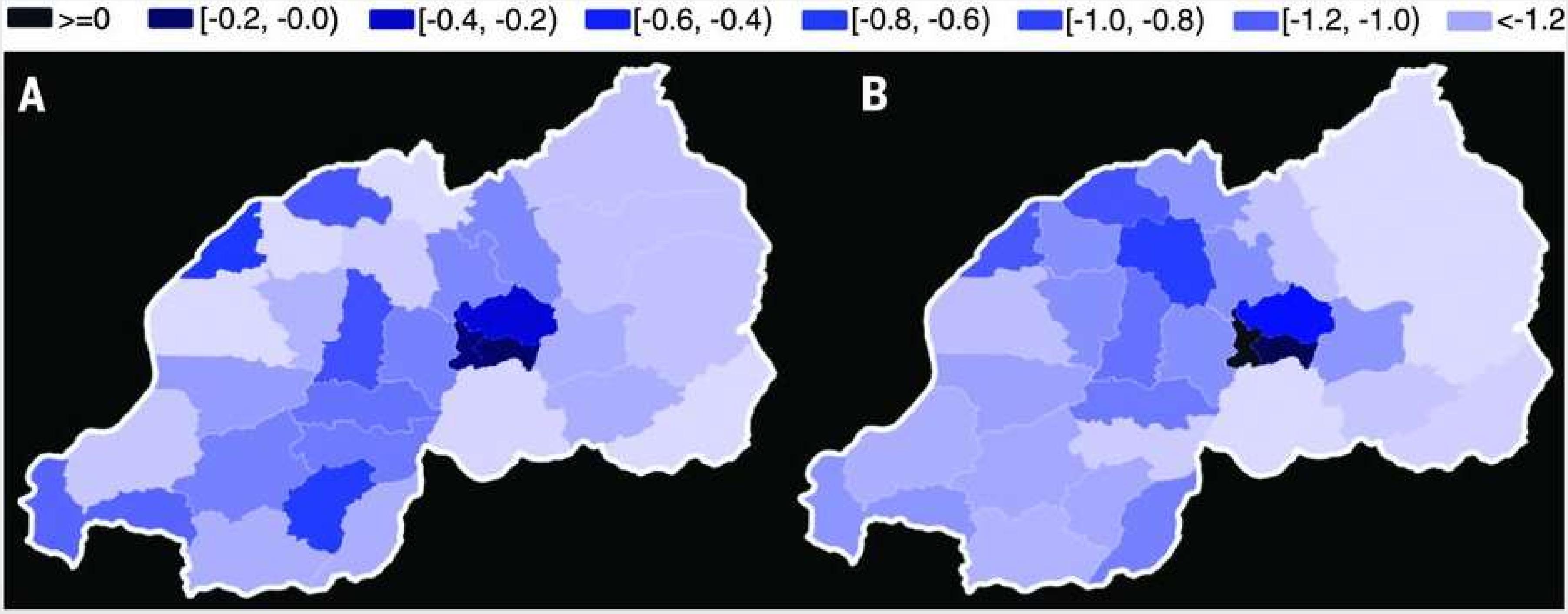
# Analysis


call records
feature engineer
person by feature matrix
person by feature matrix
estimate model
predict
survey data
imputed survey data
residence location
estimate

# Results



# Results

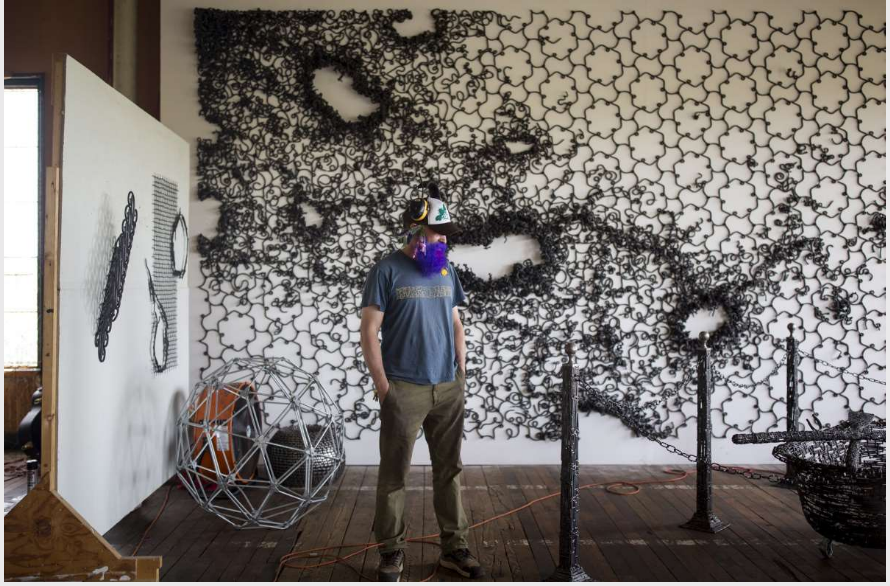


# Results

- 10 times faster
- 50 times cheaper

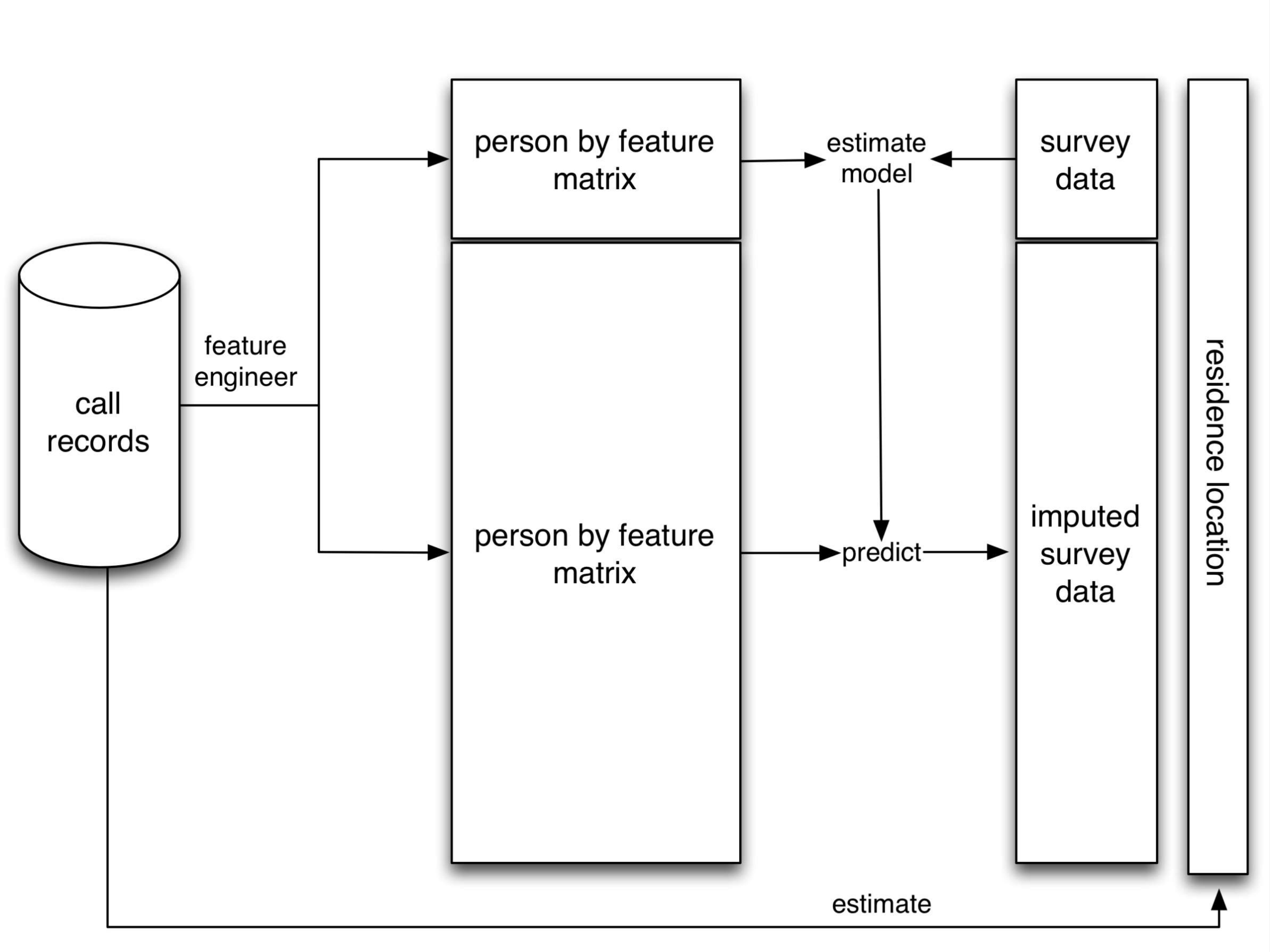
# DS Vs. Science with Data

Readymade



Custommade

# DS Vs. Science with Data



# Course Objectives

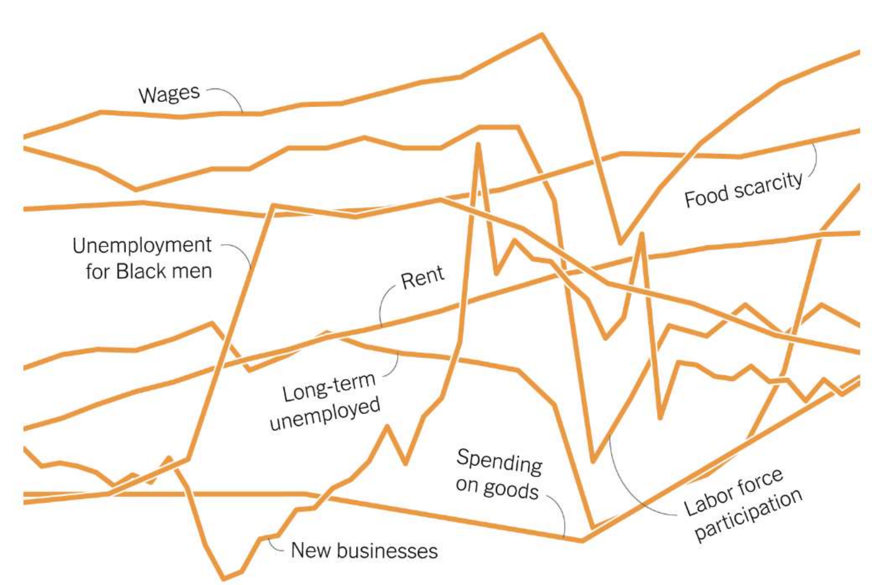
- **Content**: Critical thinking, analysis, presentation
    - How to think about data
        - Data --> theory; Theory --> data
    - How to use data
        - Structured vs. unstructured
    - How to analyze data
        - The basics (but the basics are EVERYTHING)

# Course Objectives

- **Skills**: Computing and analysis in `R`
    - Introduction: no prior experience necessary
    - Opening tabular data
    - Plotting with base `R` and `ggplot`
    - Writing functions

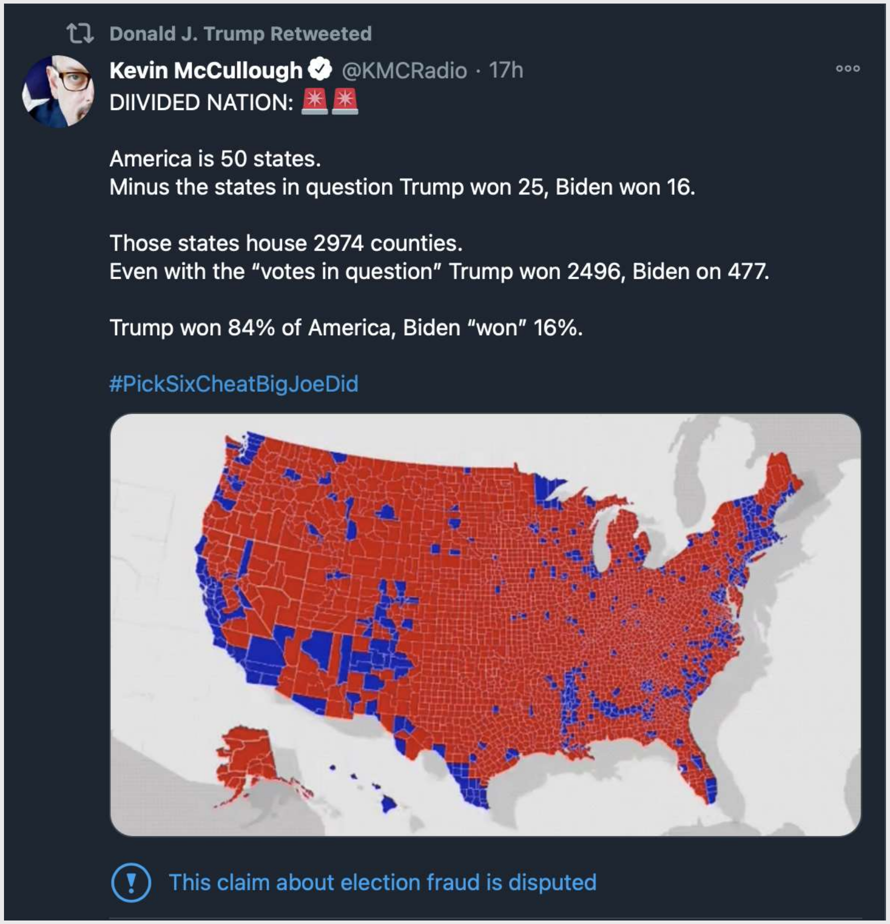
# Course Objectives

- **Perspective**: How to read empirical research



**How the U.S. Economy Is Actually Doing, in 9 Charts**

The unemployment rate doesn't tell the whole story, so we talked to a panel of economists to find out what other measures can shed light.


Donald J. Trump Retweeted

**Kevin McCullough** @KMCRadio · 17h

DIIVIDED NATION: 🚨🚨

America is 50 states.
Minus the states in question Trump won 25, Biden won 16.

Those states house 2974 counties.
Even with the "votes in question" Trump won 2496, Biden on 477.

Trump won 84% of America, Biden "won" 16%.

#PickSixCheatBigJoeDid

This claim about election fraud is disputed

# What does "introduction" mean?

- This is not a "foundations" course
- Will give you experience running code (copy, paste, & **tweak**)
- Will not go through every function in detail
- Will not go through the math behind analysis choices
- Focus on intuition and motivation

# A preview of the substantive stuff

- Predict U.S. elections using survey data (`linear regression`)
- Understand why some movies make more money (`linear regression`)
- Predict college admissions and enrollment (`linear regression`)
- Identify "clusters" of voters (`unsupervised learning`)
- Analyze Twitter data (`sentiment analysis`)
- Predict who wrote contested documents (`natural language processing`)

# How to succeed

- Before class
  - Download lecture notes & data
  - Try to `knit` the code
  - Review lecture notes
- During class
  - Be a prediction algorithm! (i.e., try to predict what code will do)
  - Ask questions...if you have a question, everyone does (5 hours at home vs 5 minutes in class)
- After class
  - Tweak code
  - Be patient with yourself

# The Internet Over Time

- Web 1.0 (1990-2000)
    - Static websites
    - Read-only interaction
    - Company-oriented
    - Owning content
- Web 2.0 (2000-2010)
    - Interactive websites
    - User-generated content
    - Individual-oriented
    - Sharing content
- Web 3.0 (2010-today)

# Parting Question & HW

- Is the internet better today than in 2009? Worse?
  - Why?
  - Post **ungraded** paragraph response to Brightspace
- TA assignments:
  - Each of you is assigned to a specific TA. They will be your primary point of contact.

> - `Aaliyah-Caroline:` Mubarak Ganiyu
> - `Catherine-Jansen:` Sriram Kannan
> - `Jayna-Lucas:` Enya Tan
> - `Luke-Ryan D. Lee:` Amogh Vig
> - `Ryan M. Schaufele-Zongwei:` Quishi Yan

- "Better" & "Worse" suggest ethics / morals / normative thinking
  - Never **EVER** *EVER* lose this lens