Data science: What is it and why does it matter?

Jesse Gronsbell

Department of Statistical Sciences
University of Toronto

Boston High School Data Science Initiative
October 5, 2022

Today's roadmap

- Who I am
- How I got here
- What data science is
- Why data science matters

Who I am

Assistant Professor in Statistics, University of Toronto

Who I am

Assistant Professor in Statistics, University of Toronto

I develop statistical learning methods for
high volume, high noise health data

Who I am

Assistant Professor in Statistics, University of Toronto

I develop **statistical learning methods** for
high volume, high noise health data

Methods to analyze data so you can learn something from it

Who I am

Assistant Professor in Statistics, University of Toronto

I develop statistical learning methods for
high volume, high noise health data

Big and messy data sets from electronic health records,
smartphones, biobanks, etc.

What that actually means

Chatting with physicians about the data I am analyzing

My favorite physician, Dr. Paul Varghese

What that actually means

Writing and debugging code

main Infairness-Project / PheCap_analysis.R Go to file ···

jlgrons Update PheCap_analysis.R Latest commit f18e3c6 on Sep 1 History

1 contributor

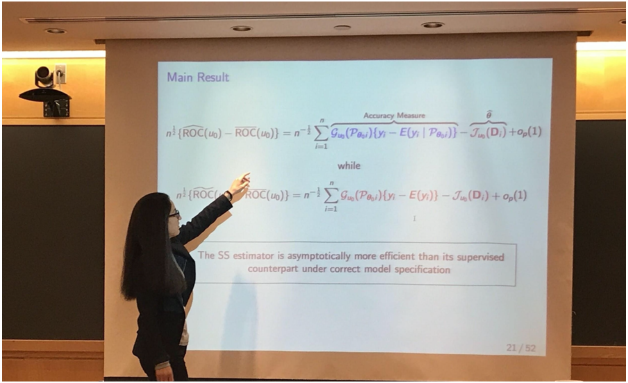
116 lines (89 sloc) | 4.04 KB Raw Blame

```r
library(PheCAP)

data(ehr_data)

data <- PhecapData(ehr_data, "healthcare_utilization", "label", 0.4, patient_id = "patient_id")


surrogates <- list(
       PhecapSurrogate(
             variable_names = "main_ICD",
             lower_cutoff = 1, upper_cutoff = 10),
       PhecapSurrogate(
             variable_names = "main_NLP",
             lower_cutoff = 1, upper_cutoff = 10),
       PhecapSurrogate(
            variable_names = c("main_ICD", "main_NLP"),
             lower_cutoff = 1, upper_cutoff = 10))


system.time(feature_selected <- phecap_run_feature_extraction(data, surrogates))

Y_train <- c(data$frame[data$training_set, data$label])
X_train <- as.matrix(data$frame[data$training_set,
             feature_selected$selected])
```

An example from my GitHub

What that actually means

Doing math so I understand why certain methods work



Defending my PhD thesis a long time ago

How I got here

A very long time ago

I didn't think I would go to college

How I got here

Fired from my job as a line chef at El Azteco

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

Harvard, PhD in Biostatistics with Tianxi Cai

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

Harvard, PhD in Biostatistics with Tianxi Cai

Stanford, Postdoc in Biomedical Data Science with Lu Tian

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

Harvard, PhD in Biostatistics with Tianxi Cai

Stanford, Postdoc in Biomedical Data Science with Lu Tian

Alphabet's Verily Life Sciences, Data Scientist

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

Harvard, PhD in Biostatistics with Tianxi Cai

Stanford, Postdoc in Biomedical Data Science with Lu Tian

Alphabet's Verily Life Sciences, Data Scientist

University of Toronto, Assistant Professor in Statistics

Now

How I figured it out

I took every opportunity in statistics that was given to me

How I figured it out

I took every opportunity in statistics that was given to me

- Baseball pitch classification
- Ballistic missile defense
- Electronic health records
- Mobile health
- COVID-19 testing strategies
- . . .

What I learned

> The best thing about being a statistician is that you get to play in everyone's backyard.

John Tukey

What I would've learned if I were younger

The best thing about being a ~~statistician~~ **data scientist** is that you get to play in everyone's backyard.

What is data science

Question

What do you think data science is?

What is data science

Textbook answer

By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling...... yet the most important part is its applications - all sorts of applications. This journal is devoted to applications of statistical methods at large

...

Journal of Data Science, **2003**

What is data science

Textbook answer

By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling...... yet the most important part is its applications - all sorts of applications. This journal is devoted to applications of statistical methods at large

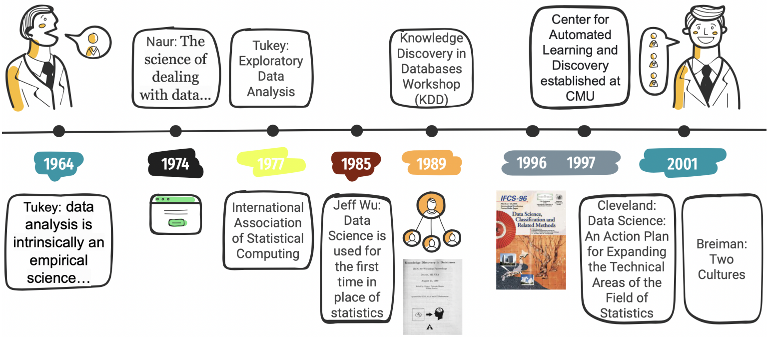
...

Journal of Data Science, **2003**

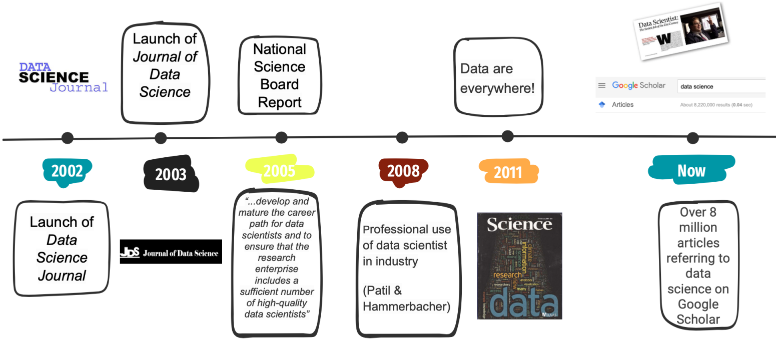
Truthful answer

It depends who you ask

A brief history of data science



A brief history of data science



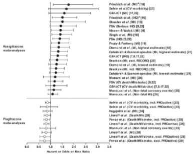
Something we can all agree on

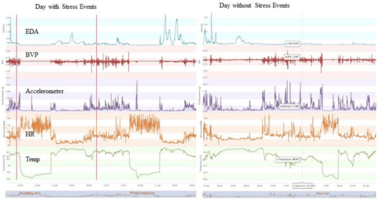
By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling...... yet the most important part is its applications - all sorts of applications. This journal is devoted to applications of statistical methods at large . . .

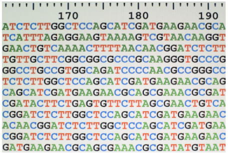
Journal of Data Science, **2003**

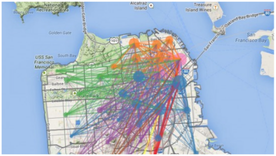
But, what is data?

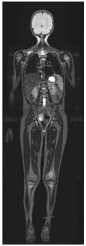
	No Cold	Cold
Placebo	31	109
Vitamin C	17	122











How much data is there

- IDC: ‘Global Datasphere’ reached 33 zettabytes (2020)
  - zettabyte: trillion gigabytes

How much data is there

- IDC: 'Global Datasphere' reached 33 zettabytes (2020)
  - ⋆ zettabyte: trillion gigabytes

This is a lot of data!

Imagine that each bit is a penny, which is around 0.1 inches thick. One ZB made up of a stack of pennies would be 2,550 lightyears. This can get you to the nearest star system, Alpha Centauri, 600 times.

How much data is there

- IDC: 'Global Datasphere' reached 33 zettabytes (2020)
  - ⋆ zettabyte: trillion gigabytes

Some stats...

- In just one minute:
  - ⋆ Twitter users sent 473,400 tweets
  - ⋆ Snapchat users shared 2 million photos
  - ⋆ Instagram users posted 49,380 pictures
  - ⋆ LinkedIn gained 120 new users

How much data is there

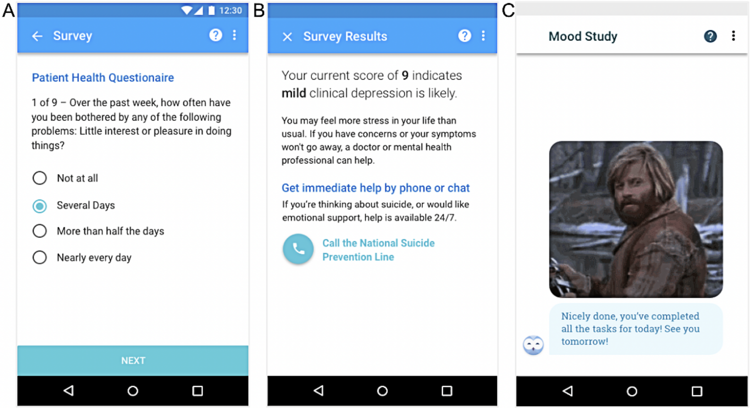
- IDC: ‘Global Datasphere’ reached 33 zettabytes (2020)
  - ⋆ zettabyte: trillion gigabytes

Some stats...

- Google processes more than 40,000 searches/sec and 3.5 billion searches/day
- $\frac{1}{5}$ of the world’s population (1.5 billion people) are active on Facebook every day
- $\frac{2}{3}$ of the world’s population (5 billion people) now own a mobile phone

But it takes a lot of work to understand data

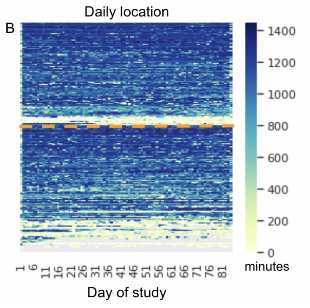
Example: Verily Baseline Mood Study



Nickels et al 2021

Raw vs processed data

Raw location data



Processed location data

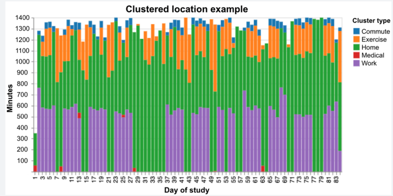
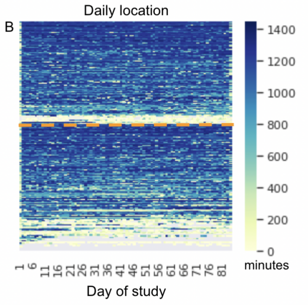


Figure 4. Example of clustered location data for 1 participant for the duration of the study. The total number of minutes (vertical axis) with categorized locations (denoted by various colors in the legend) are plotted as stacked bars for each day of the study on the horizontal axis. Note the week-long increased homestay starting on day 28.

Raw vs processed data

Raw location data



Processed location data

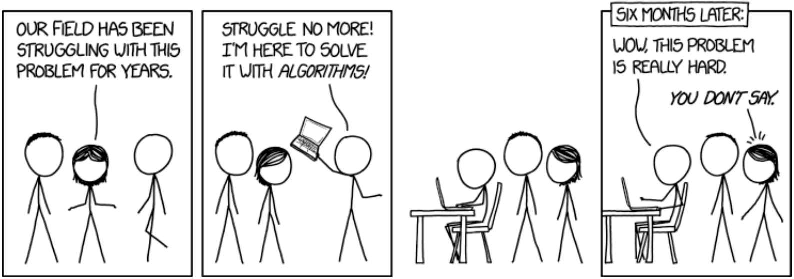
Figure 4. Example of clustered location data for 1 participant for the duration of the study. The total number of minutes (vertical axis) with categorized locations (denoted by various colors in the legend) are plotted as stacked bars for each day of the study on the horizontal axis. Note the week-long increased homestay starting on day 28.

It took us years to process raw location data

A lot of data $\neq$ a lot of answers

"After analyzing all your data, I think we can safely say that none of it is useful."

Data science problems are hard



Taking data seriously: The data science process

What we hope. The reality.

Step 0 Identify a meaningful question.

Step 1 Gather the data.

Step 2 Clean and explore the data.

Step 3 Model the data.

Step 4 Communicate the results.

Taking data seriously: The data science process

What we hope. The reality.

Step 0 Identify a meaningful question.

Step 1 Gather the data.

Step 2 Clean and explore the data.

Step 3 Model the data.

Step 4 Communicate the results.

All of these steps require a lot of thought!

Step 0: Identify a meaningful question

Things to think about when developing your question:

- What is the goal of your analysis?
- What impact will the conclusion of your analysis have?

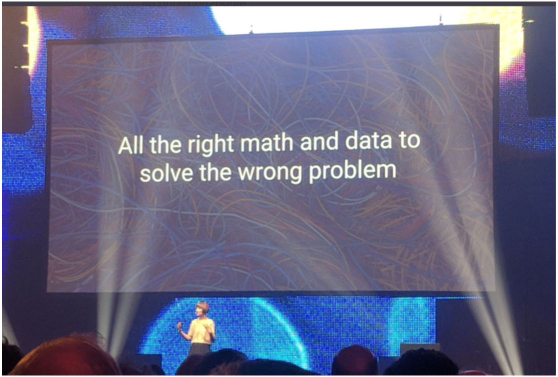
Step 0: Identify a meaningful question

Things to think about when developing your question:

- What is the goal of your analysis?
- What impact will the conclusion of your analysis have?

Think "why am I doing what I am doing?"

Yes, step 0 can go wrong!



Twitter

Step 0... gone wrong

- You: *Can we predict in-hospital mortality within 48 hours of ICU admission using MIMIC data?*

You go off to build the model...

Step 0... gone wrong

- You: *Can we predict in-hospital mortality within 48 hours of ICU admission using MIMIC data?*

You go off to build the model...

- You: *The model has a great AUC!*
- Clinician: *All of the patients that have a high probability of mortality are on end of life care. The model doesn't convey anything new to me.*

Step 0... gone wrong

- You: *Can we predict in-hospital mortality within 48 hours of ICU admission using MIMIC data?*

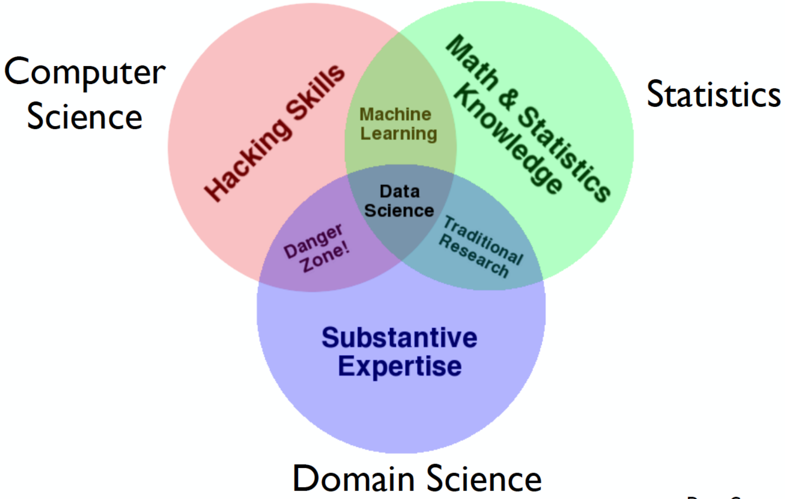
You go off to build the model...

- You: *The model has a great AUC!*
- Clinician: *All of the patients that have a high probability of mortality are on end of life care. The model doesn't convey anything new to me.*

Lesson learned

Data science is a team sport

Taxonomy of data science



Drew Conway

Step 1: Gather the data

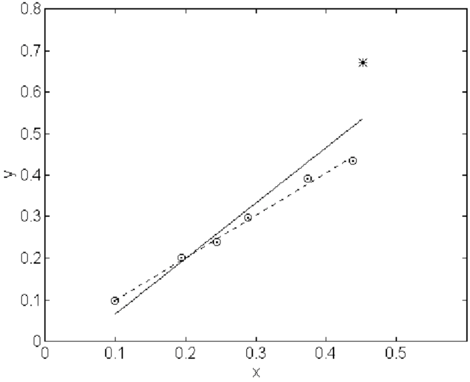
In data science projects, you will be:

- Given data
- Required to download data
- Required to scrape data off of the web
- Responsible for collecting the data
- Some combination of these

Step 2: Clean and explore the data

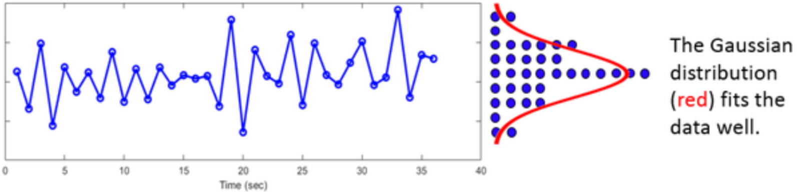
Often the most time consuming part of data science

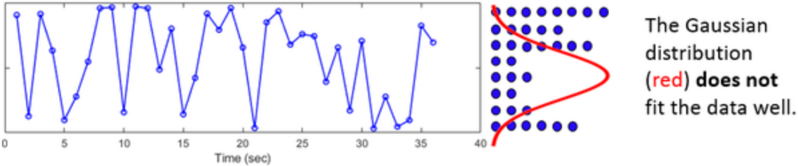
The importance of data cleaning



Failure to clean the data can impact your conclusions

The importance of data exploration





Exploring the data allows you to verify modeling assumptions

Step 3: Model the data

The first step in modeling is to understand the task.

- A **statistician** classifies a modeling task as either:
  - Description
  - Prediction
  - Causal inference
- A **computer scientist** classifies a modeling task as either:
  - Supervised
  - Unsupervised

Step 3: Model the data

The first step in modeling is to understand the task.

- A **statistician** classifies a modeling task as either:
  - Description
  - Prediction
  - Causal inference
- A **computer scientist** classifies a modeling task as either:
  - Supervised
  - Unsupervised

These terms do not necessarily map to each other

Description vs prediction vs causal inference

Description Using data to provide a quantitative summary of certain features of the world

eg. descriptive statistics

Table 1. Key demographics of the 384 participants with minimally sufficient data.

Characteristic	Participants with minimally sufficient data (n=384)	
	Depressed (n=313)	Not depressed (n=71)
Age (years), n (%)		
18-29	123 (39.3)	26 (36.6)
30-39	90 (28.8)	22 (30.9)
40-49	56 (17.9)	14 (19.7)
50-59	36 (11.5)	8 (11.2)
60-69	7 (2.2)	0 (0)
70-79	1 (0.3)	1 (1.4)

Nickels et al 2021

Description vs prediction vs causal inference

Description Using data to provide a quantitative summary of certain features of the world

eg. clustering

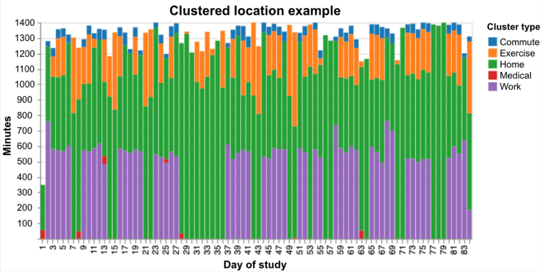


Figure 4. Example of clustered location data for 1 participant for the duration of the study. The total number of minutes (vertical axis) with categorized locations (denoted by various colors in the legend) are plotted as stacked bars for each day of the study on the horizontal axis. Note the week-long increased homestay starting on day 28.

Nickels et al 2021

Description vs prediction vs causal inference

Prediction Using data to map some features of the world ("features") to other features of the world ("outcome")

Table 4

Model performance

Eligible patients captured represents the number of patients identified by the EHR computable phenotype algorithm that could have qualified for the registry. Registry patients missed shows the number of patients enrolled in the registry (N = 179) that the computable phenotype did not identify.

	SC only	SC/Fyler	SC/NLP	SC/NLP/Fyler
PPV, % (95% CI)	82 (72 – 91)	83 (75 – 92)	84 (75 – 93)	85 (77 – 93)
Sensitivity, % (95% CI)	48 (42 – 54)	59 (53 – 65)	64 (57 – 71)	66 (60 – 73)
AUC, % (95% CI)	85 (78 – 92)	89 (83 – 95)	89 (83 – 95)	90 (85 – 95)
Total eligible patients captured, No.	470	518	575	575
New eligible patients captured, No.	323	364	414	413
Registry patients missed, No.	32	25	18	17

AUC = area under the receiver operating characteristic curve; NLP = natural language processing; PPV = positive predictive value; SC = standard codified

Geva et al 2017

Aside: Data terminology

Data elements are often classified as either:

- **Outcome/Response/Output/Target/Label(s)**: The variable(s) that you want to understand better, often denoted as y
- **Covariate/Input/Feature(s)**: The variable(s) that can potentially tell you something about your outcome(s), often denoted as x

Aside: Data terminology

Data elements are often classified as either:

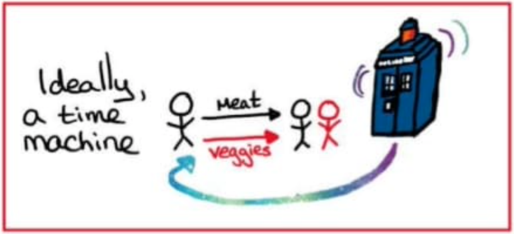
- **Outcome/Response/Output/Target/Label(s)**: The variable(s) that you want to understand better, often denoted as y
- **Covariate/Input/Feature(s)**: The variable(s) that can potentially tell you something about your outcome(s), often denoted as x

Warning

There are often several terms and definitions for a concept in data science!

Description vs prediction vs causal inference

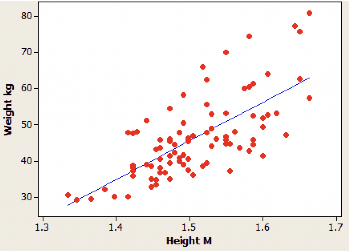
Causal inference Using data to predict certain features of the world if the world had been different

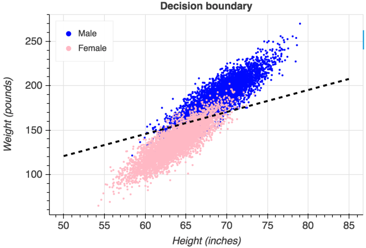


@EpiEllie – A cartoon guide to causal inference

Supervised vs unsupervised

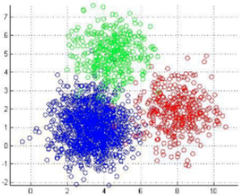
Supervised The data is made of observations with information on both the features and the outcome ("label driven")

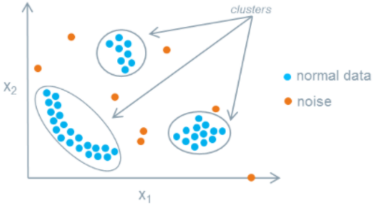




Supervised vs unsupervised

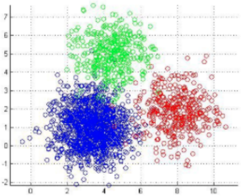
Unsupervised The data is made up of observations with information only on the features

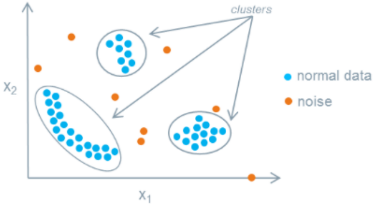




Supervised vs unsupervised

Unsupervised The data is made up of observations with information only on the features





P.S.

Other approaches exist (e.g. semi-supervised)

Step 4: Communicate the results

Tell a logical story

Telling a logical story

A story has a beginning, middle, and end!

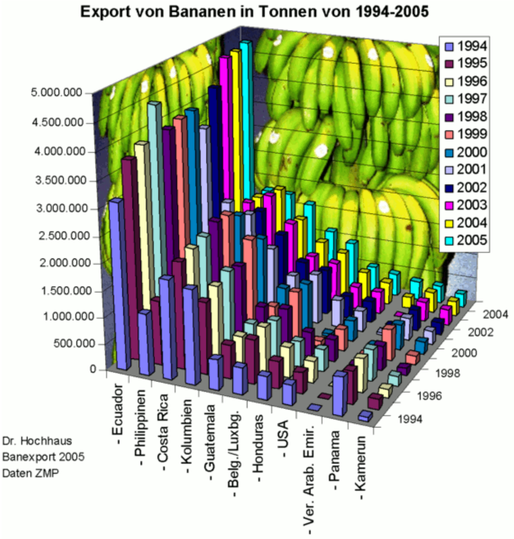
Telling a logical story

A story has a beginning, middle, and end!

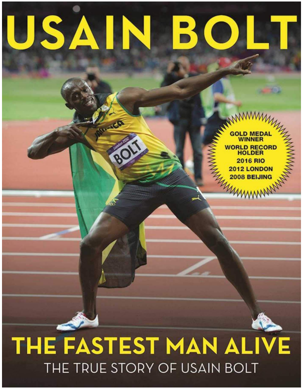
- Introduce interesting characters
  - Explain and create some excitement for your problem
- Put them in a predicament
  - Explain why your problem is hard
- Resolve the predicament
  - Explain your solution
- Leave room for sequels!
  - Discuss future work and improvements

Harvard CSC109

Keep your story simple



Make your story interesting



How fast is Usain Bolt compared to 116 years of sprinting?

Why data science matters

We use data science to learn something about the world and to help make decisions

Why data science matters

We use data science to learn something about the world and to help make decisions

We have to be **extremely careful** of potential pitfalls!

Policy creep

Reality

- Patient with asthma has pneumonia and is treated more aggressively
- Fewer patients with asthma die of pneumonia

Policy creep

Reality

- Patient with asthma has pneumonia and is treated more aggressively
- Fewer patients with asthma die of pneumonia

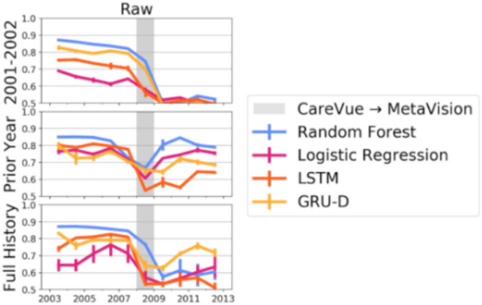
Learned

- If you get pneumonia, it's better if you already have asthma too!

Cabitza et al 2017

Dataset shift

Mortality AUC vs. Time, by model and history used



Nestor et al 2019

Algorithmic (un)fairness

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

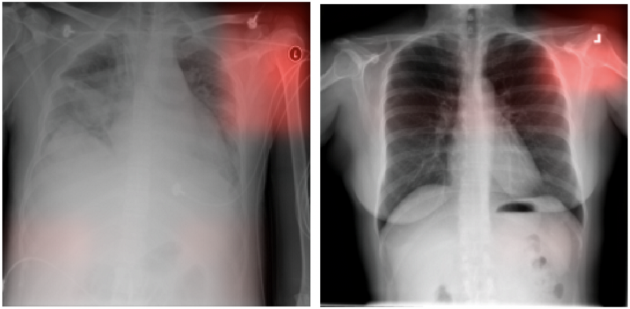
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Obermeyer at al 2019

Sensitivity to noise

Pneumonia or artifact?



(b) **(c)**

Zech et al 2018

Value of data science IRL

- Consider predicting cardiac arrest in the pediatric ICU
- Cardiac arrest occurs in 100 patients out of 3 million per year
- Suppose we can build a highly accurate algorithm to predict cardiac arrest with a true positive rate of 100% and a false positive rate of 1%

Question

Should we deploy such a model in practice?

Take-away: Big or small, you need the right data

> *"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data..."*

John Tukey

Thank you!