Data Science in Action

Machine Learning for Self-Driving Cars

CELEHS, Harvard University

Jesse Gronsbell

The teaching team

- ⋆ **Jesse Gronsbell**, Assistant Professor of Statistics (U of T)
- ⋆ **Aaron Sonabend**, Data Scientist (Google)
- ⋆ **Andy Beam**, Assistant Professor of Epidemiology (Harvard)
- ⋆ **Junwei Lu**, Assistant Professor of Biostatistics (Harvard)
- ⋆ **Nitin Sharma**, Technical Director of AI Research (Paypal)
- ⋆ **Bryan Cai**, PhD student in CS (Stanford)

Read more about us here

Let's start!

Question

What do you think data science is?

Other thoughts...

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**

A brief history of data science

1964 Tukey: data analysis is intrinsically an empirical science…

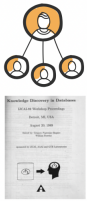
1974 Naur: The science of dealing with data…

1977 Tukey: Exploratory Data Analysis

International Association of Statistical Computing

1985 Jeff Wu: Data Science is used for the first time in place of statistics

1989 Knowledge Discovery in Databases Workshop (KDD)



1996 **1997** Center for Automated Learning and Discovery established at CMU

2001 Cleveland: Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

Breiman: Two Cultures

A brief history of data science

2002

DATA SCIENCE Journal

Launch of *Data Science Journal*

2003

Launch of *Journal of Data Science*

JDS Journal of Data Science

2005

National Science Board Report

"...develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists"

2008

Professional use of data scientist in industry

(Patil & Hammerbacher)

2011

Data are everywhere!

Now

Over 8 million articles referring to data science on Google Scholar

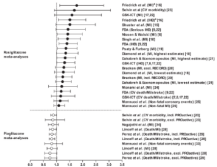
Why is data science what it is today?

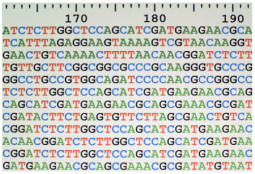
By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling..... yet the most important part is its applications - all sorts of applications. This journal is devoted to applications of statistical methods at large . . .

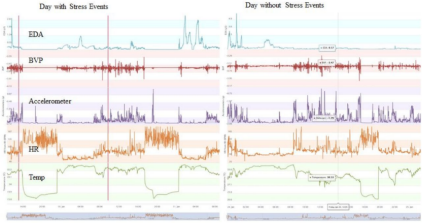
Journal of Data Science, **2003**

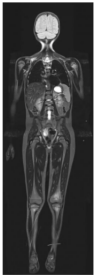
But, what is data?

	No Cold	Cold
Placebo	31	109
Vitamin C	17	122









How much data is there?

- ⋆ IDC: ‘Global Datasphere’ reached 18 zettabytes (2018)
  - ○ zettabyte: 10^{21} bytes, trillion gigabytes

How much data is there?

- IDC: 'Global Datasphere' reached 18 zettabytes (2018)
  - zettabyte: 10^{21} bytes, 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?

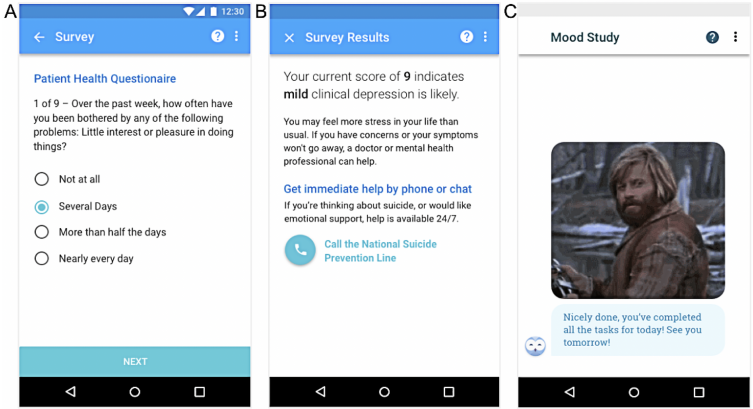
- ⋆ IDC: 'Global Datasphere' reached 18 zettabytes (2018)
  - ○ zettabyte: 10^{21} bytes, trillion gigabytes

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

Raw vs processed data

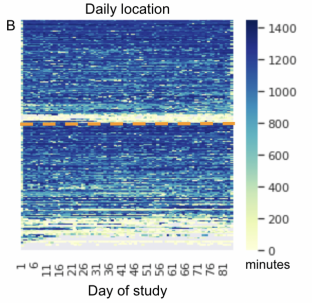
Example: Verily Baseline Mood Study



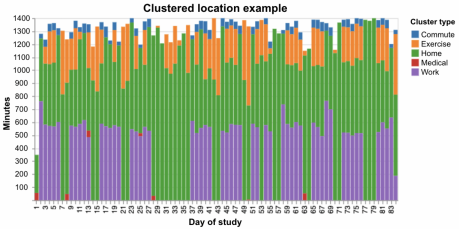
Nickels et al 2021

Raw vs processed data

Raw location data

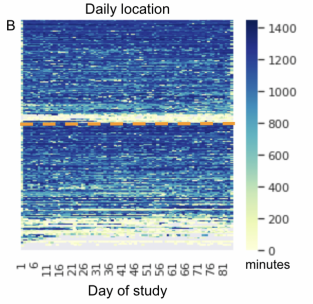


Processed location data

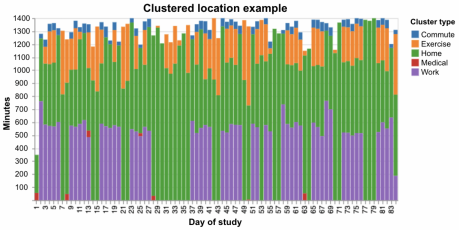


Raw vs processed data

Raw location data



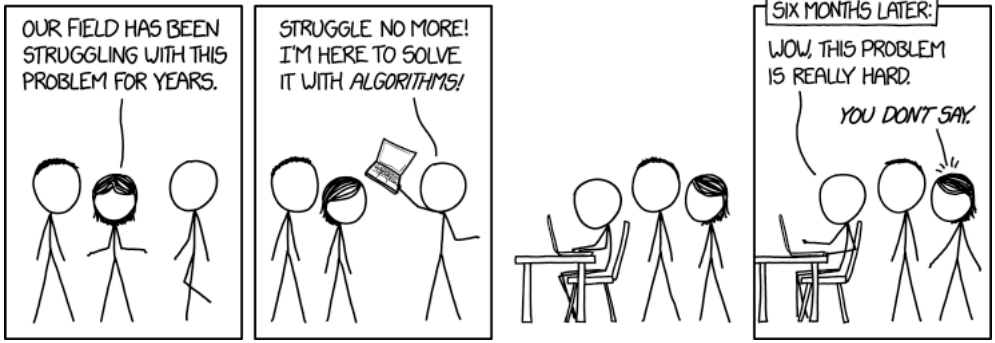
Processed location data



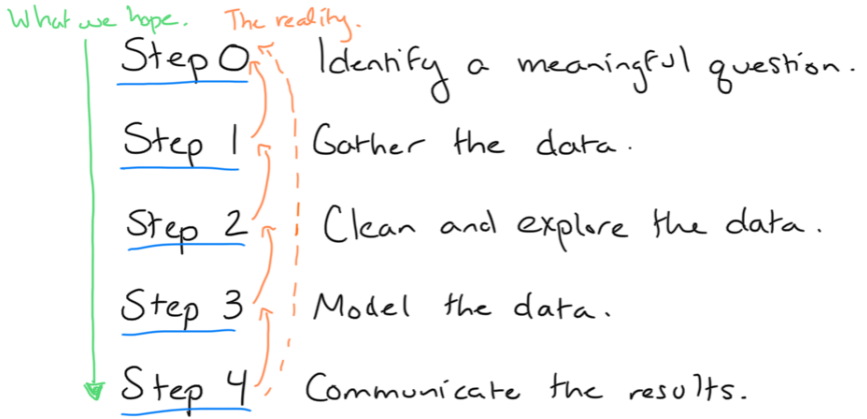
It takes a lot of work to go from raw to processed data

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

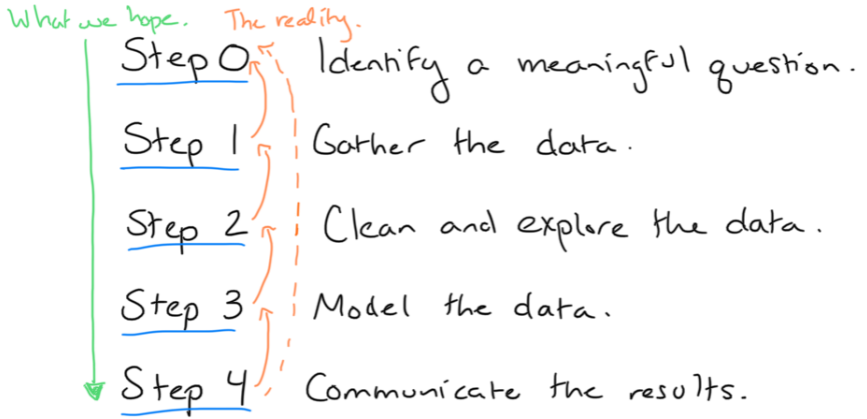
Data science problems are hard



Taking data seriously: The data science process



Taking data seriously: The data science process



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 scientific goal?
- ⋆ What do you want to predict or estimate?
- ⋆ What impact will the conclusion of your analysis have?

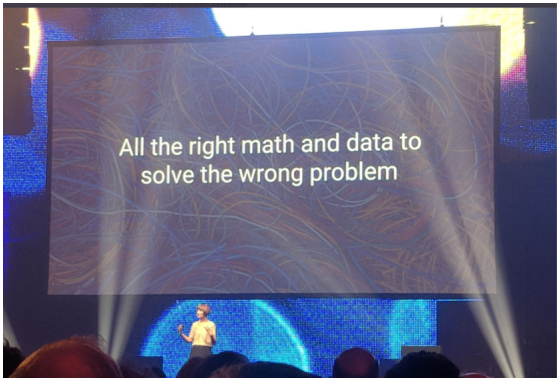
Step 0: Identify a meaningful question

Things to think about when developing your question:

- ⋆ What is the scientific goal?
- ⋆ What do you want to predict or estimate?
- ⋆ 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?*

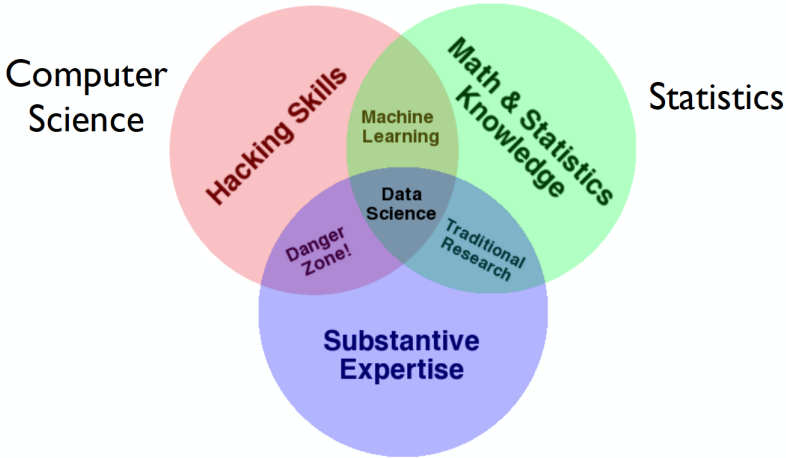
You go off to build the model...

- ⋆ 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

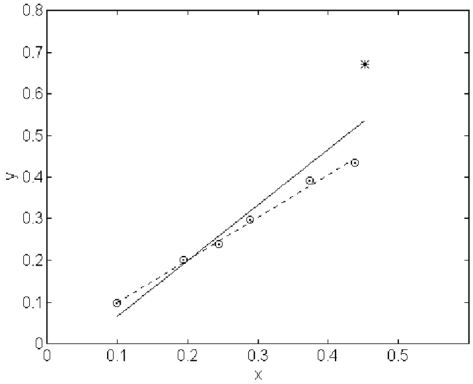
In data science projects, you will be:

- ⋆ 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

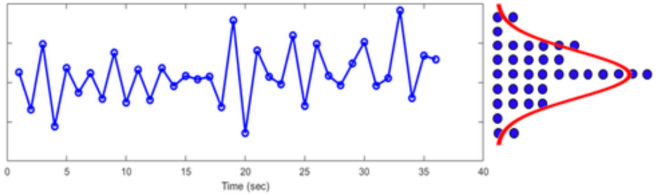
Often the most time consuming aspect of data science!

The importance of data cleaning

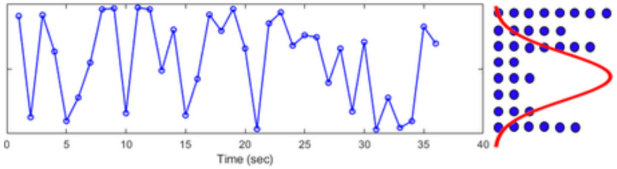


Failure to clean the data can impact your conclusions

The importance of data exploration



The Gaussian distribution (red) fits the data well.



The Gaussian distribution (red) **does not** fit the data well.

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, e.g. descriptive statistics, clustering

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, e.g. descriptive statistics, clustering

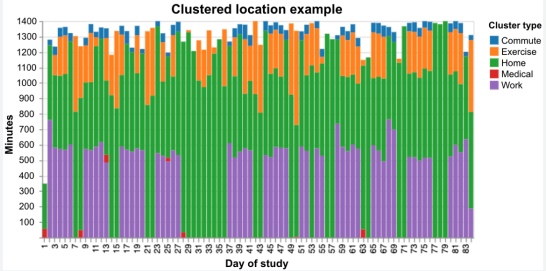


Figure 4. Example of clustered location data for 1 participant for the duration of the study. The total number of minutes (vertical axis) with categorized locations (denoted by various colors in the legend) are plotted as stacked bars for each day of the study on the horizontal axis. Note the week-long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

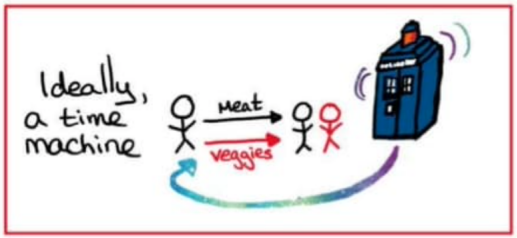
Warning

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

Description vs prediction vs causal inference

Causal inference

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

Table 3

Point and SE estimates based on MAD for the ATE of ADA vs. IFX, with respect to one-year treatment response rate, among IBD patients in EMR data based on various methods, including the naive CC estimator (CC_{Naive}) that completely ignores confounding bias. 95% CIs are percentile-based CIs from resampling and p-values are for testing $H_0 : \Delta = 0$ based on inverting percentile CIs.

Estimator	Estimate	SE	95% CI (Pct)	p-value
CC_{Naive}	0.014	0.099	(-0.201, 0.177)	0.822
CC_{AIPW}	-0.125	0.153	(-0.416, 0.164)	0.592
SS_{AIPW}	0.033	0.109	(-0.265, 0.180)	0.778
SS_{DR}	-0.067	0.036	(-0.164, -0.002)	0.044

Cheng et al 2021

Example: Description vs prediction vs causal inference

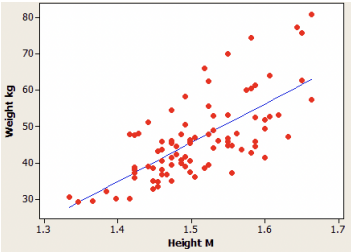
	Description	Prediction	Causal inference
Example of scientific question	How can women aged 60-80 years with stroke be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features, e.g., symptoms, clinical parameters…	• Eligibility criteria • Output, e.g., diagnosis of stroke over the next year • Inputs, e.g., age, blood pressure, history of stroke, diabetes at baseline	• Eligibility criteria • Outcome, e.g., diagnosis of stroke over the next year • Treatment, e.g., initiation of statins at baseline • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

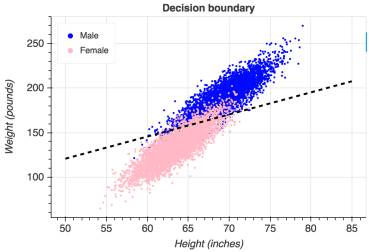
Hernan et al 2019

Supervised vs unsupervised

Supervised

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

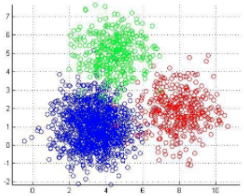


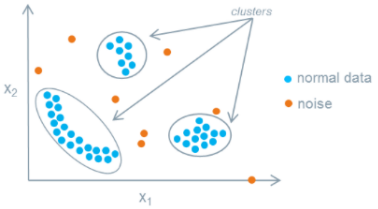


Supervised vs unsupervised

Unsupervised

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

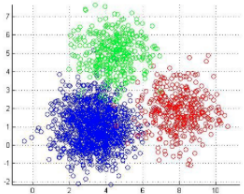


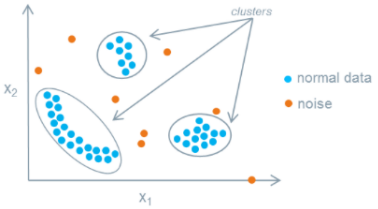


Supervised vs unsupervised

Unsupervised

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





P.S.

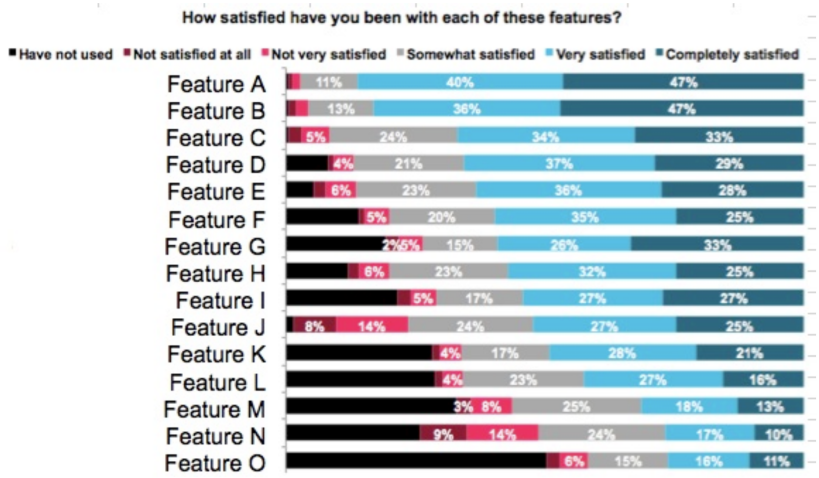
Other approaches exist (e.g. semi-supervised)

Step 4: Communicate the results

Tell a logical story!

- ⋆ 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

Make your story easy to follow

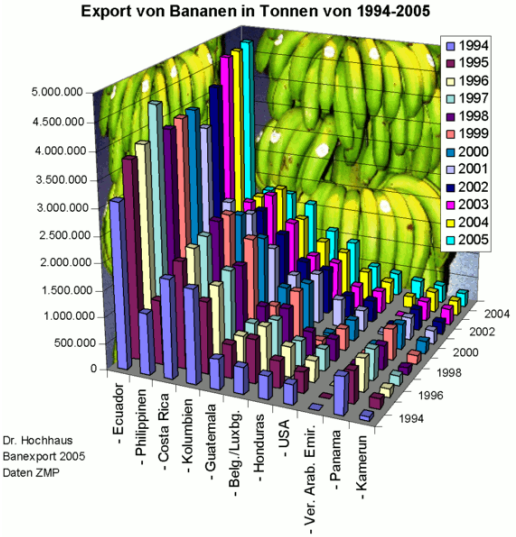


Make your story easy to follow

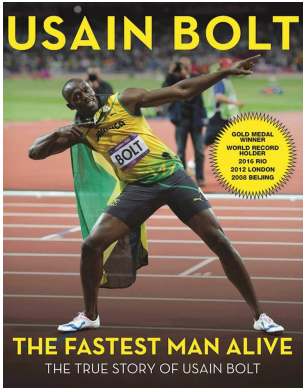
User satisfaction varies greatly by feature

Feature O is least-used feature; what steps can we proactively take with existing users to increase use?

Keep your story simple

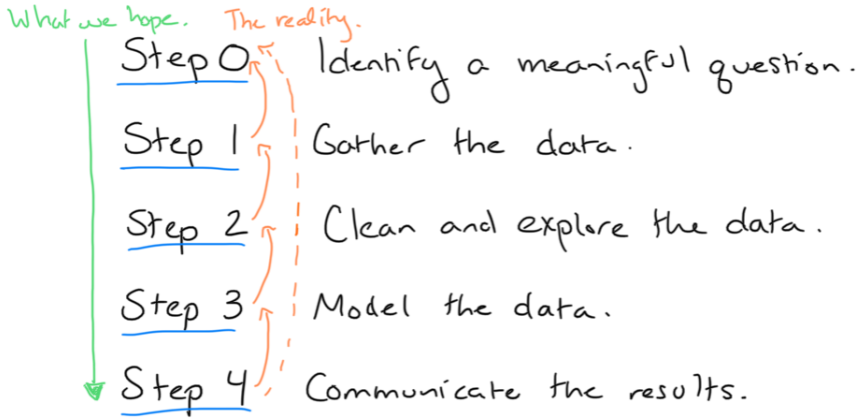


Make your story interesting



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

Our overarching goal



Better understand the data science process

Examples of data science pitfalls

If you aren't careful!

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

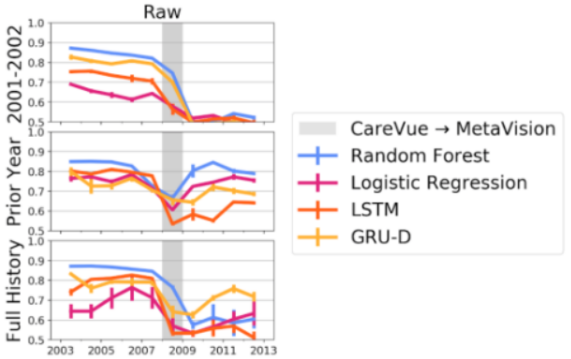
Learned

- ⋆ 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 bias

RESEARCH

RESEARCH ARTICLE

ECONOMICS

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

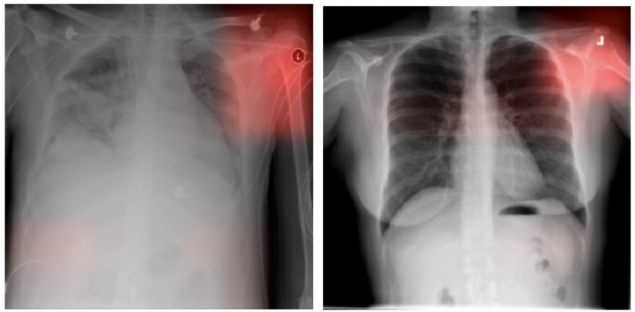
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

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..."

Tukey

Summary

- ⋆ Data can be tricky and has to be handled with care
- ⋆ Critical thinking is a must for each project/problem you face
- ⋆ Not all problems have been solved, but many have
- ⋆ Carefully check your modelling assumptions
- ⋆ Carefully check the performance of your model