JSC 270 - Lecture 1

https://jsc270.github.io/

Why are YOU in this course?

https://padlet.com/nyulik/jsc270_lec1

What are we trying to achieve in this course?

- Critical thinking when handling data
- Use of basic tools when dealing with data
  - Pre-processing
  - Exploration
  - Modeling
  - Assessment
  - Visualization
- Comfortably working with real data
- Have fun!

The Team

Professor

Anna Goldenberg (Anna)
CS

TA

Lauren Erdman
(CS)

TA

Matthew Edwards
(Stats)

Syllabus - Lectures/Labs

https://jsc270.github.io/

Lectures Mondays 1-2pm
Discussion/Office Hours 2-3pm
(Occasional invited talk 2-3pm)

Labs Wednesdays 12-2pm
Prepare you to successfully complete your homeworks in python

Syllabus - Evaluation

https://jsc270.github.io/

Assessment	Weight
Assignment 1	10%
Assignment 2	15%
Assignment 3	20%
Assignment 4	20%
Perusal Papers (6)	12%
Reflection Quizzes (6)	12%
Group Presentations (2)	10%
Universal Income	1%

Assignment 1

Data Science for COVID Due Jan 25 11:59AM

Data exploration and visualization

https://jsc270.github.io/assign_docs.html

Syllabus - Perusall

https://jsc270.github.io/

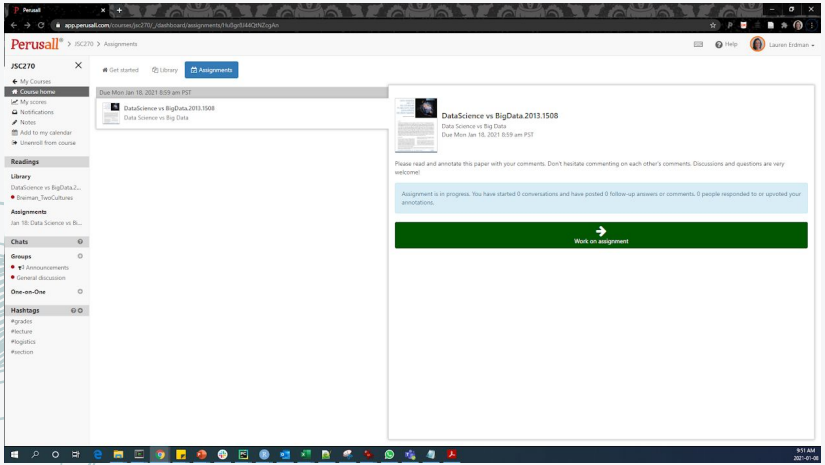
We need to learn to read scientific papers, hence Perusall! https://app.perusall.com/

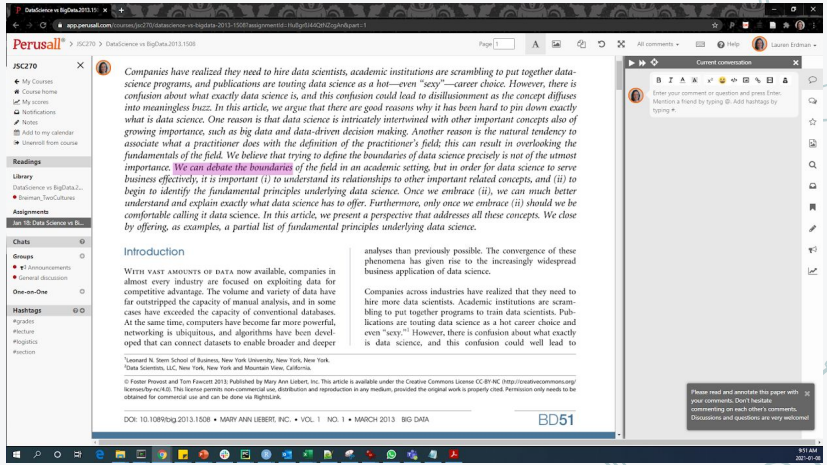
Annotate provided papers and comment on each other's annotations

Syllabus - Perusall

https://jsc270.github.io/

https://app.perusall.com/





Syllabus - Reflection Quizzes

https://jsc270.github.io/

The point is for you to think out of the box in data science terms and for us to ensure that you are all on the same page to move forward.

Should not take too long:

1-3 questions per quiz

Paragraph per question

Syllabus - Forum

https://jsc270.github.io/

A Discourse instance has been set up by the CS department: https://bb-2021-01.teach.cs.toronto.edu/c/jsc270

You should definitely be able to login with your CS instances. You should also be able to login with utoronto.ca ids (if they are the same as your CS ids). Please let me/TAs know if your ids are not the same.

Questions regarding course setup?

What do YOU think is data science?

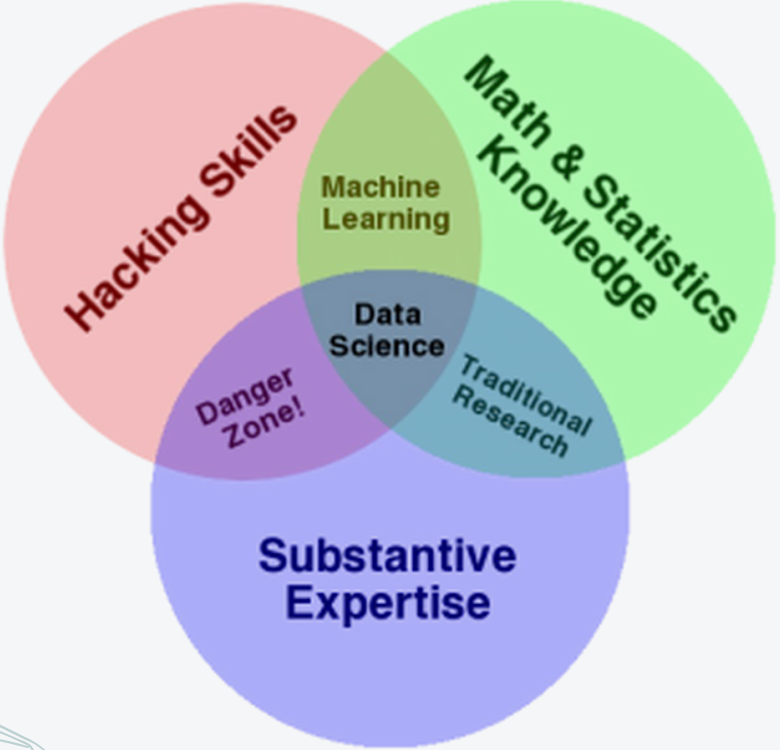
https://padlet.com/nyulik/jsc270_lec1

What others think is data science

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

From the Taxonomy of Data Science (2010)



The brief history of Data Science

1964

Tukey: data analysis is intrinsically an empirical science...

Naur: The science of dealing with data...

1974

Tukey: Exploratory Data Analysis

1977

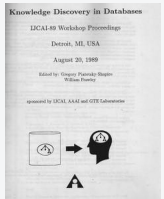
International Association of Statistical Computing

Knowledge Discovery in Databases Workshop (KDD)

1985

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

1989



Center for Automated Learning and Discovery established at CMU

1996 **1997**

2001

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

Breiman: Two Cultures

The brief history of Data Science (cont'd)

2002 - Launch of Data Science Journal

2003 - Launch of Journal of Data Science

2005 - National Science Board report that said that NSF in partnership should *"...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"*

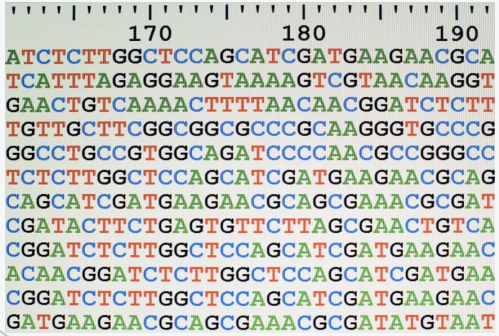
2008 - Professional use/position of data scientist in industry (Patil and Hammerbacher)

2009 - Astro2010 Decadal Survey "The Revolution in Astronomy Education: Data Science for the Masses"

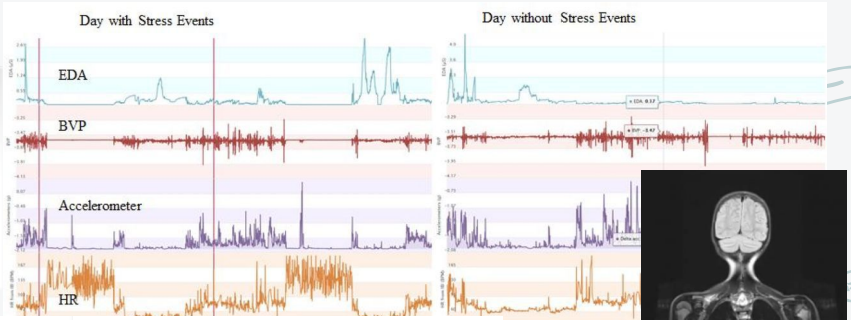
Now - 8,110,000 articles on Google Scholar referring to data science directly (there are more!)

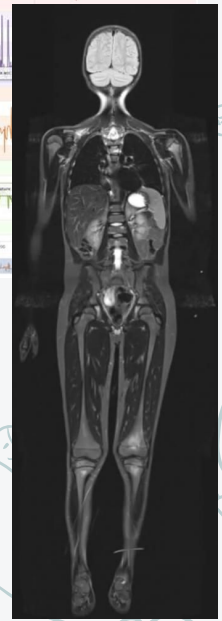
But what is data?

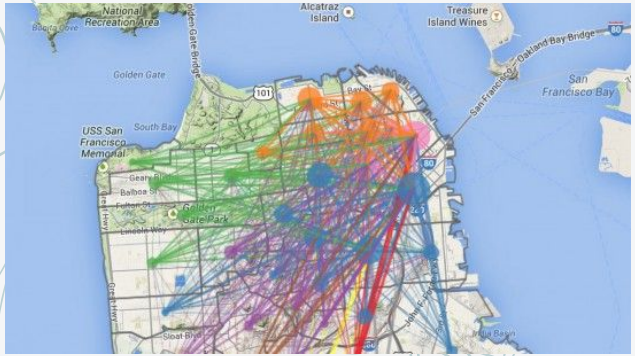
genetic data



questionnaire







How much data is there?

IDC: 'Global Datasphere' reached **18 zettabytes (2018)**

zettabyte: 10^{21} bytes, trillion gigabytes (1,000,000,000,000,000,000,000)

In just one minute:

Twitter users sent 473,400 tweets

Snapchat users shared 2 million photos

Instagram users posted 49,380 pictures

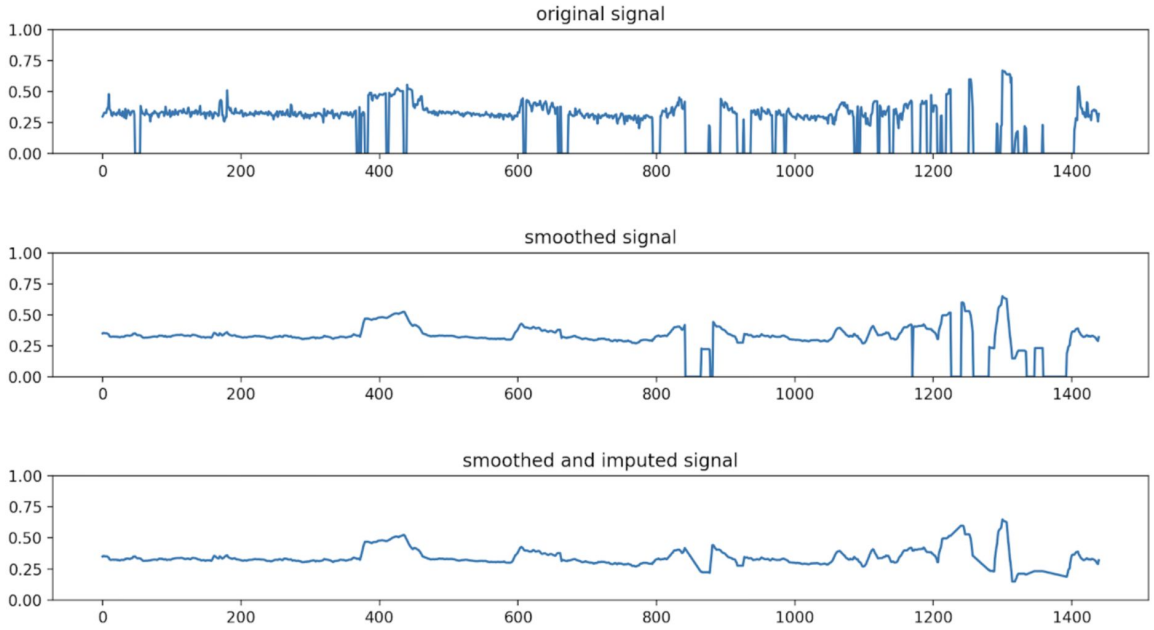
LinkedIn gained 120 new users

Google processes more than 40,000 searches/sec, 3.5 billion searches/day.

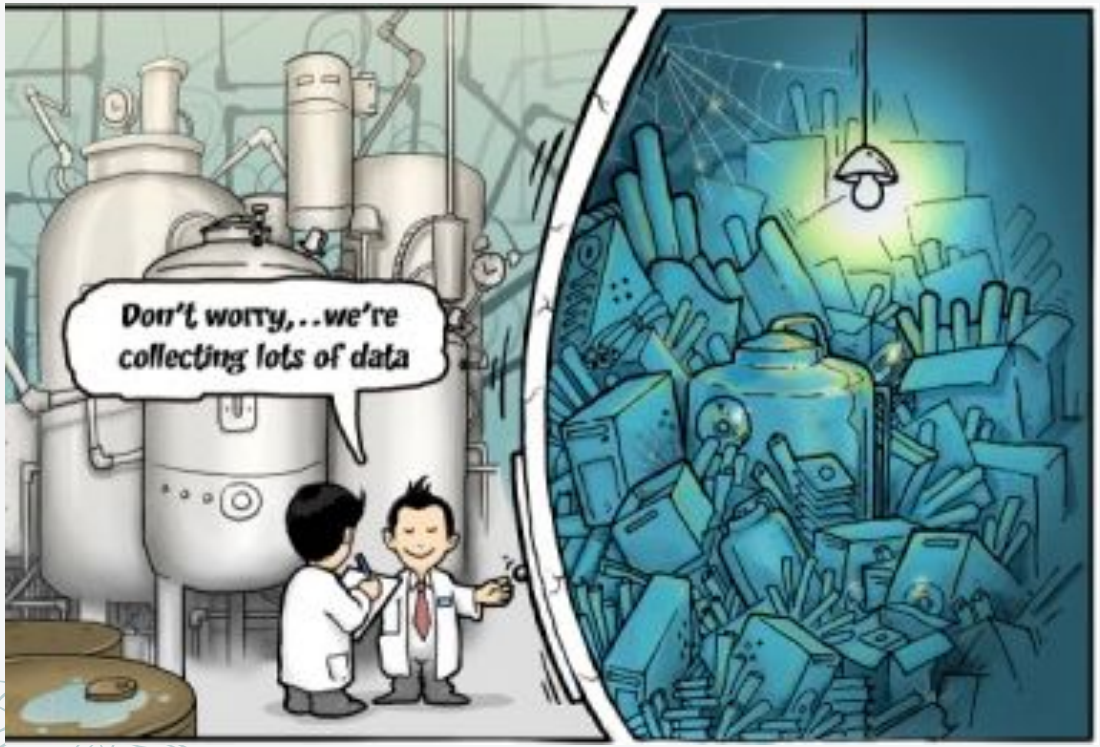
1.5 billion people (⅕ world population) are active on Facebook every day.

⅔ of the world's population now own a mobile phone.

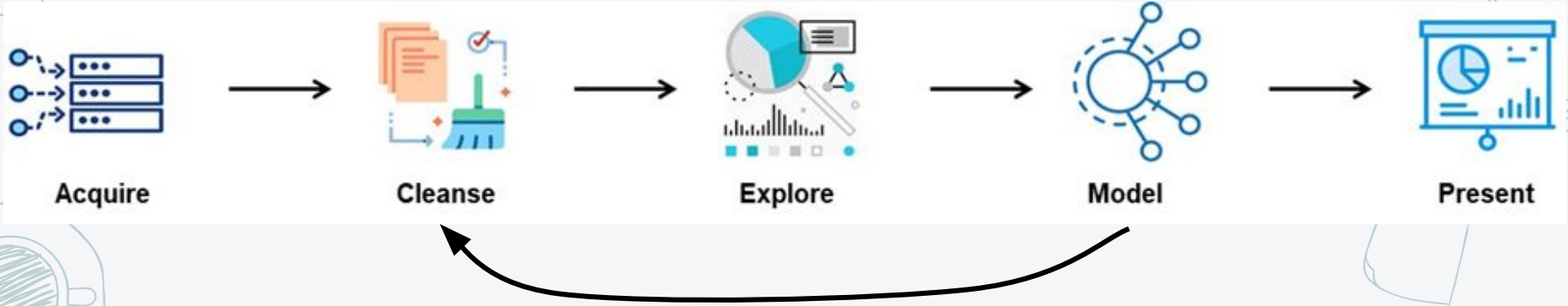
Raw vs Processed Data



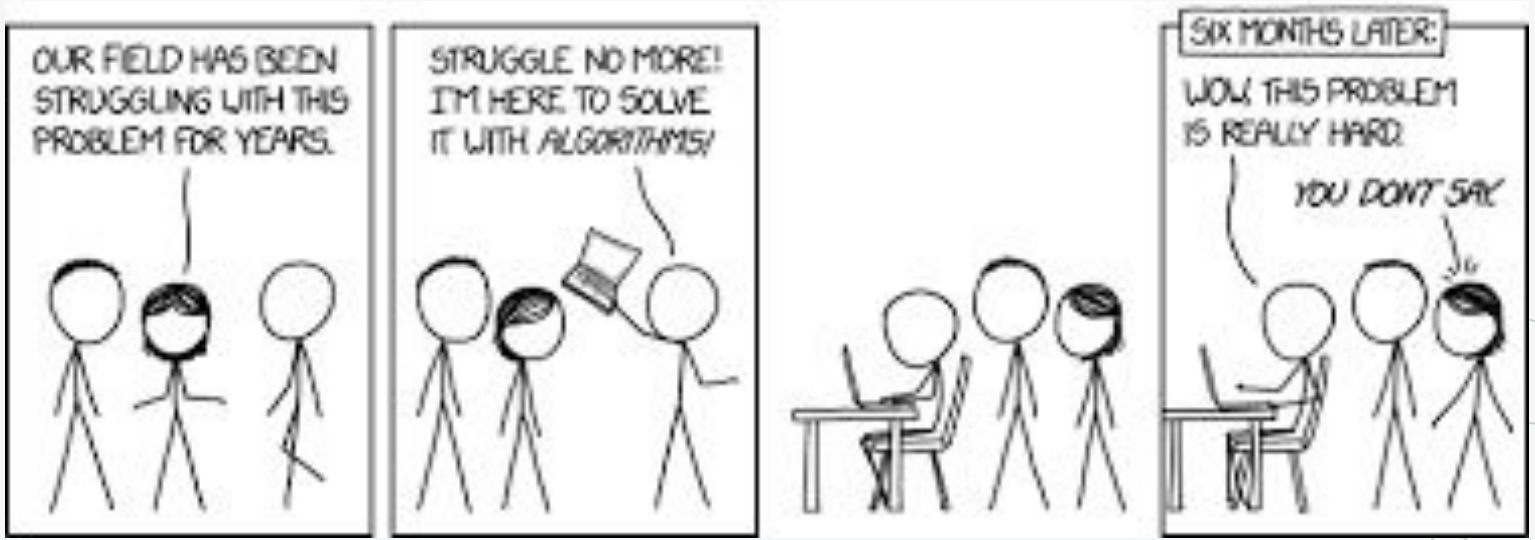
... AND THEN DATA SCIENCE ...



The Data Science Pipeline

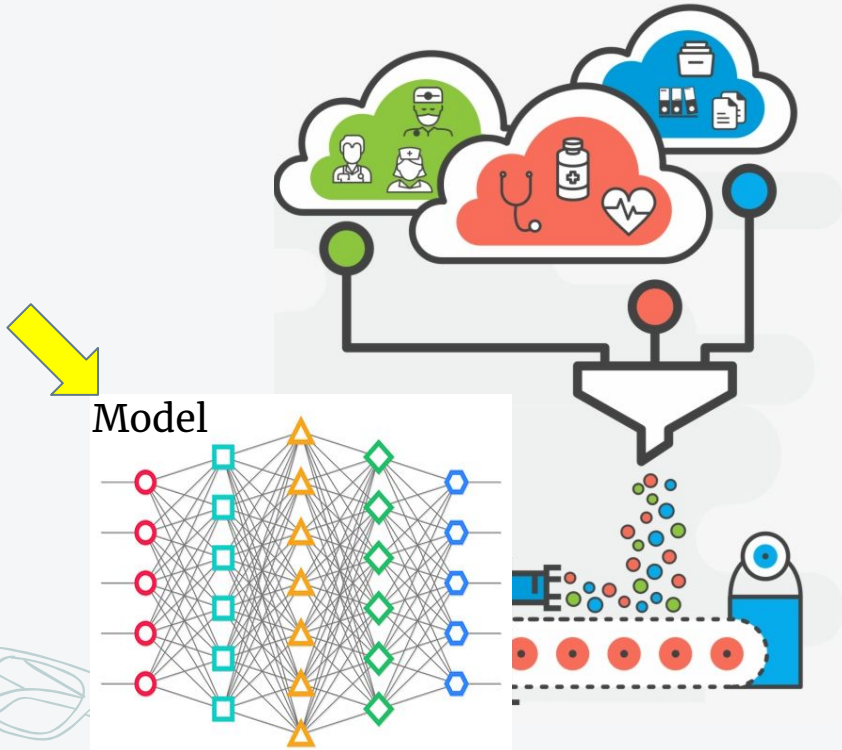


Taking data seriously

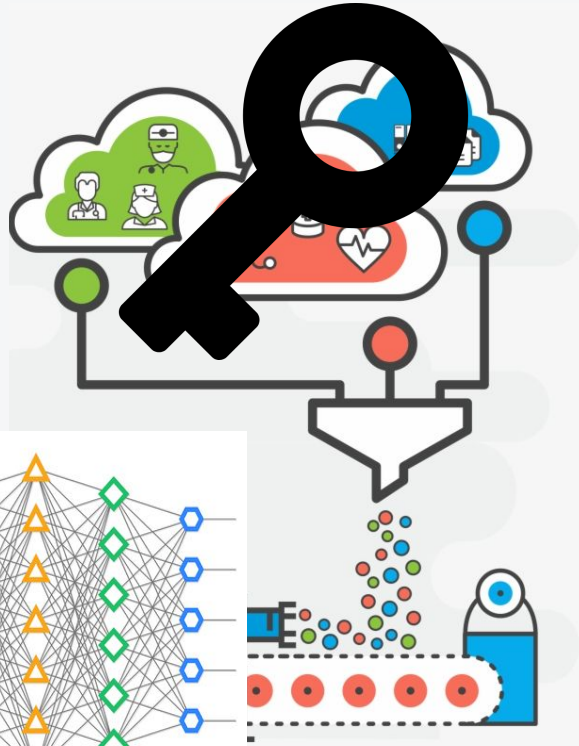


Taking Data Seriously

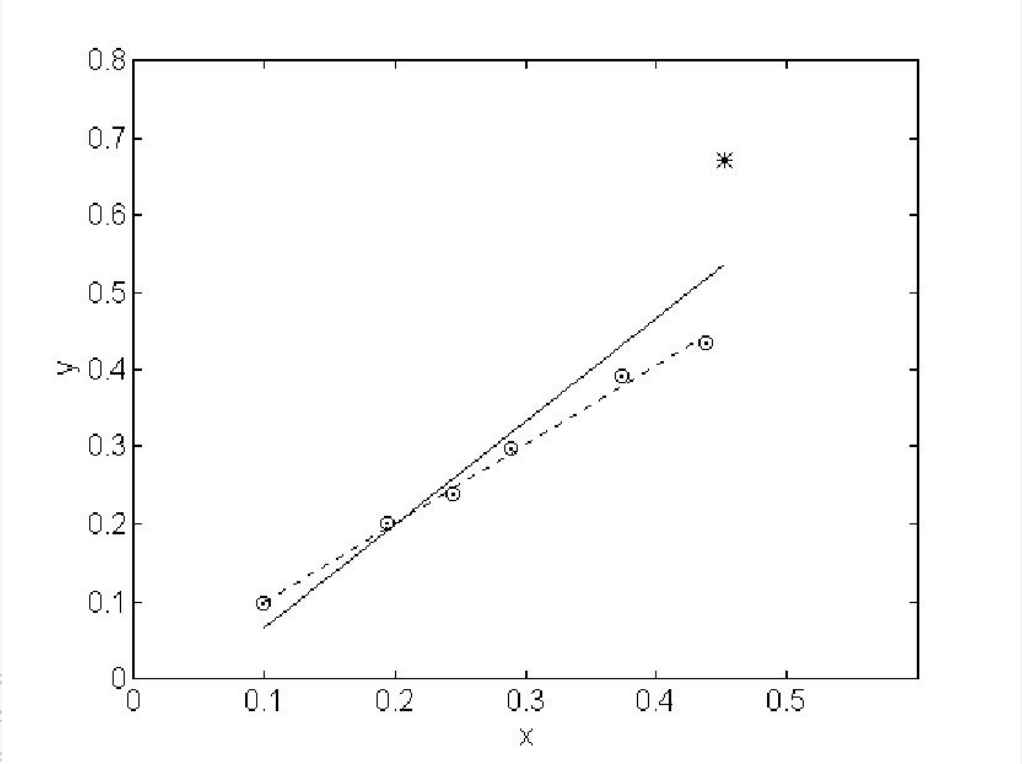
Taking Data Seriously



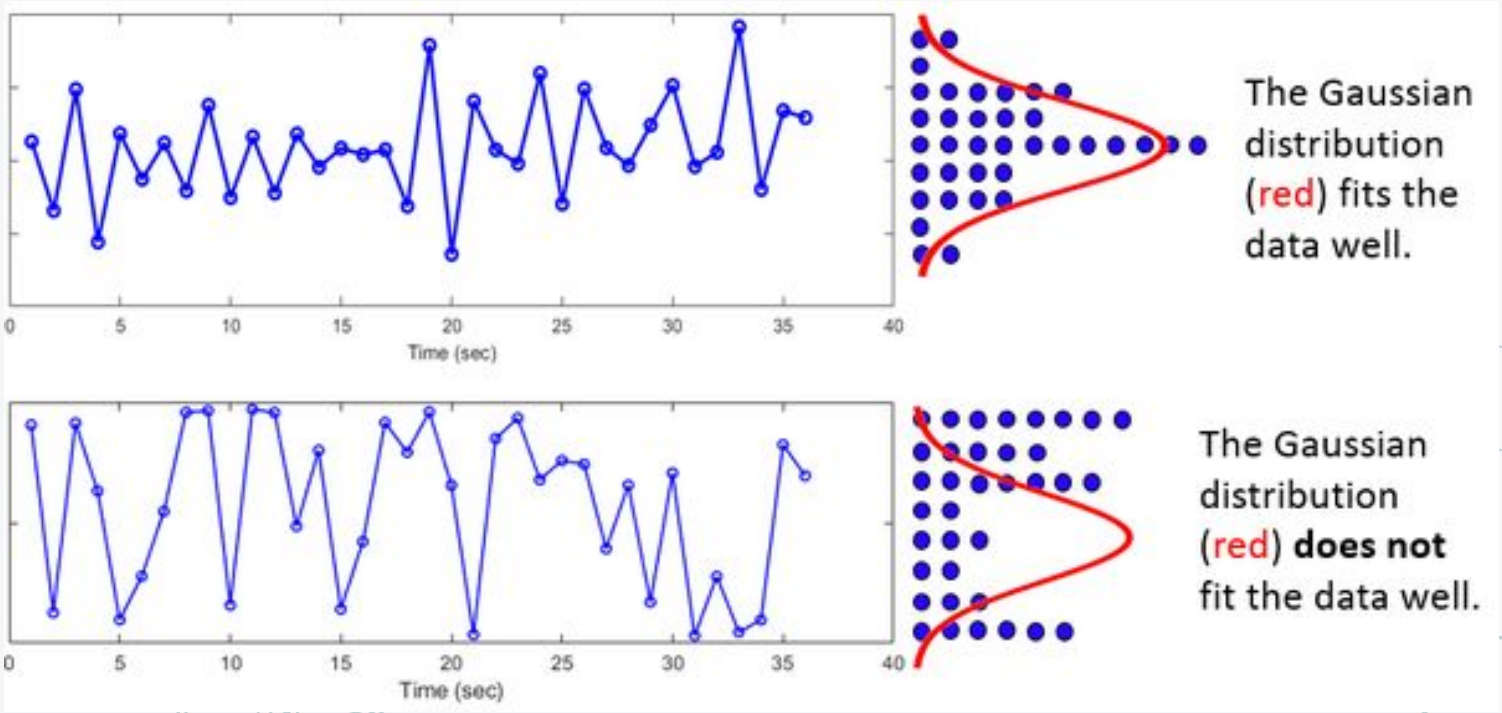
Taking Data Seriously



The importance of Cleaning



The importance of Data Exploration



Modeling

- Descriptive
- Predictive

statistics

- Supervised
- Unsupervised

AI/ML

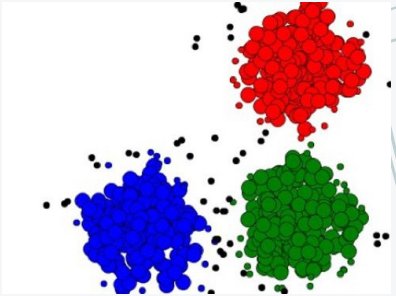
- Generative
- Discriminative

AI/ML/stats

Descriptive vs Predictive

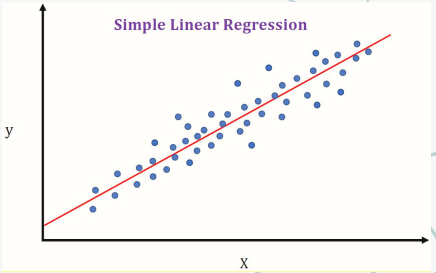
- Descriptive: analysis/model that describe the past

  E.g. clustering



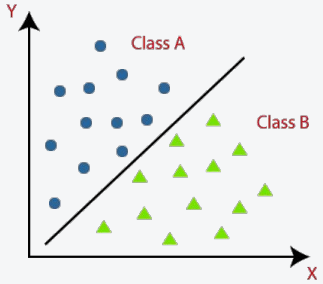
- Predictive: building a model that predicts y for the previously unseen x

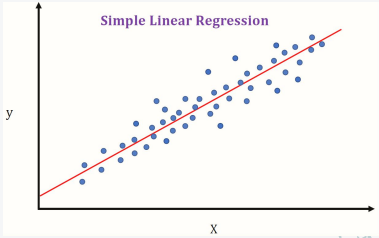
  E.g. regression



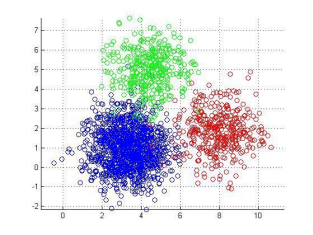
Supervised vs Unsupervised

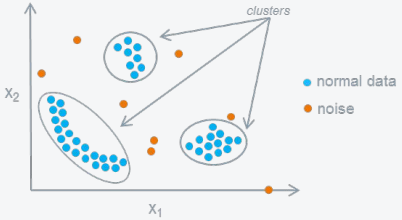
- Supervised: label driven (x,y)





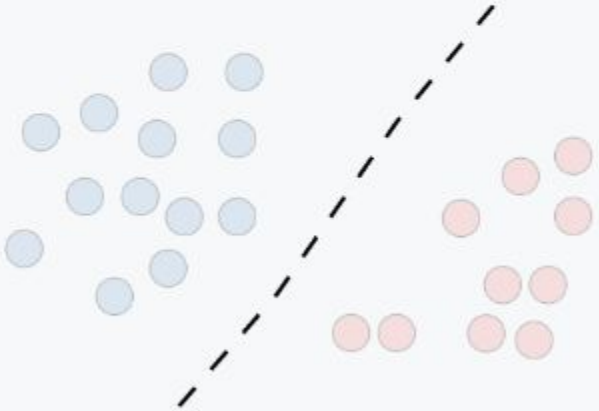
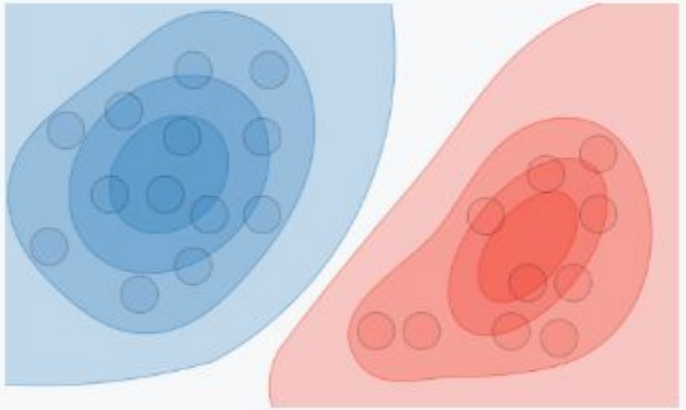
- Unsupervised: no labels (class) assoc w/ input





P.S. There are semi-supervised, self-supervised, etc approaches as well...

Generative vs Discriminative

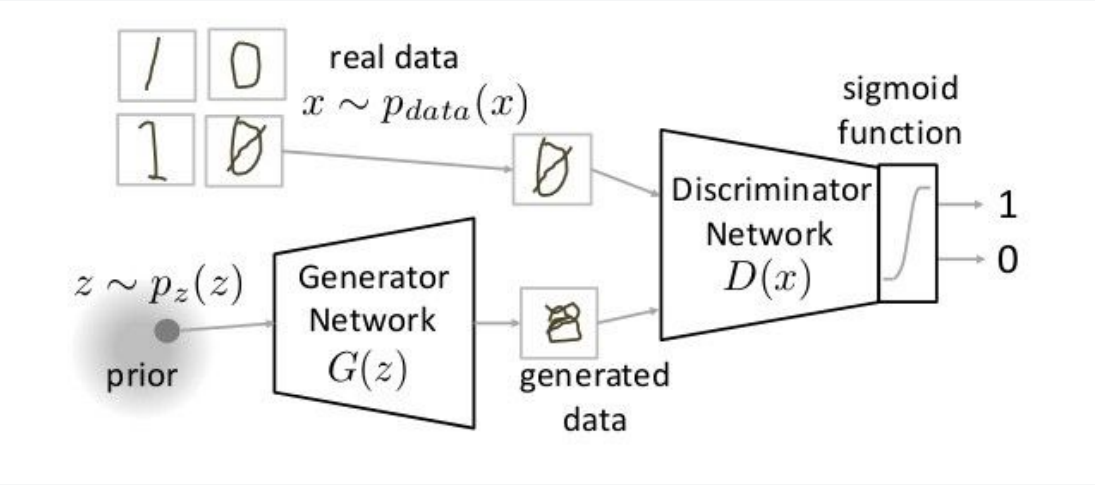
	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Stackoverflow:https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-a-discriminative-algorithm

Generative Supervised?

Generative Supervised?

Generative Adversarial Networks



Predictive Unsupervised?

A bit of an oxymoron:
Predictive implies having labels
Unsupervised means labels are not present

Examples of Pitfalls (if not careful)

Policy Creep

Reality:

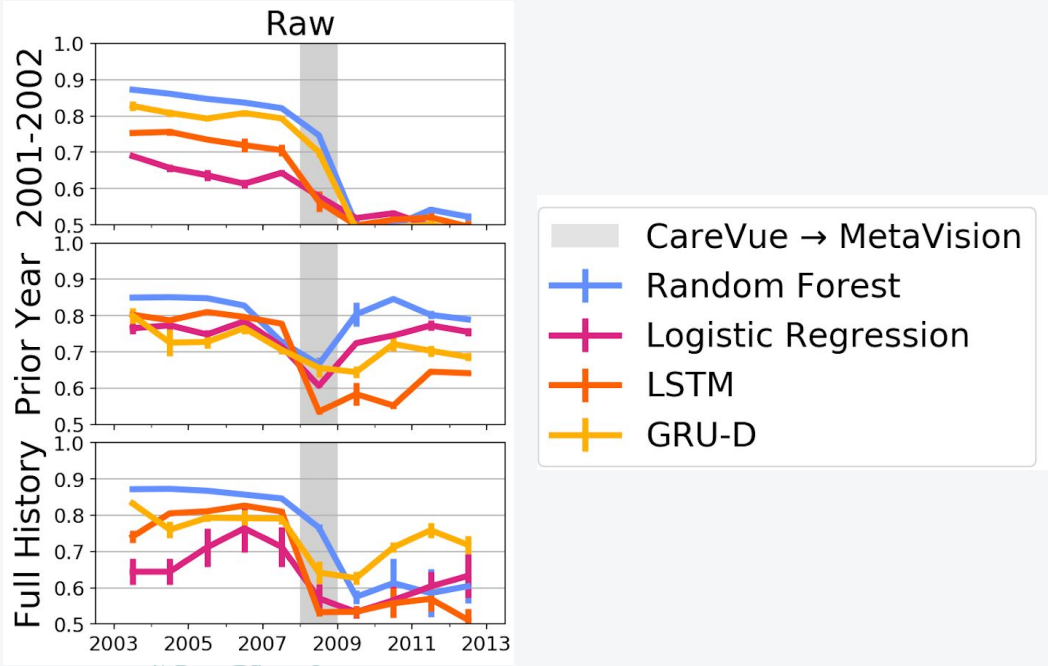
Patient with asthma has pneumonia and is treated more aggressively

Fewer patients with asthma die of pneumonia

Learned:

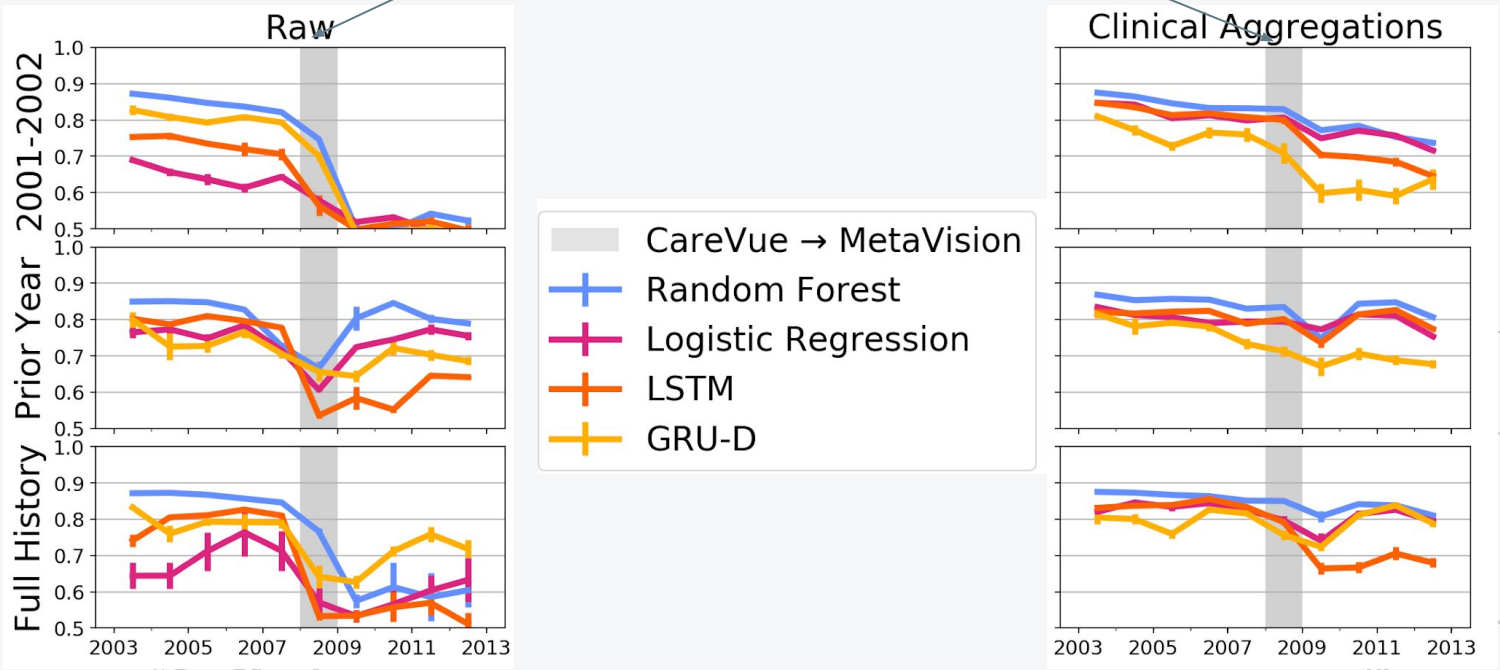
If you get pneumonia, it's better if you already have asthma too!

Dramatic Data Shifts



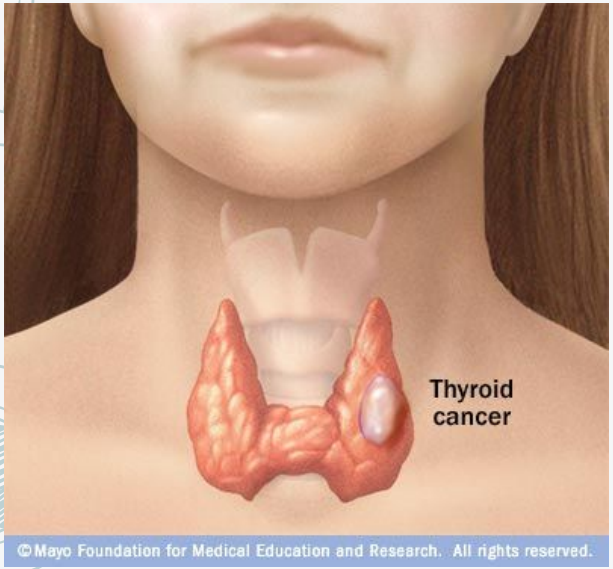
Dramatic Data Changes

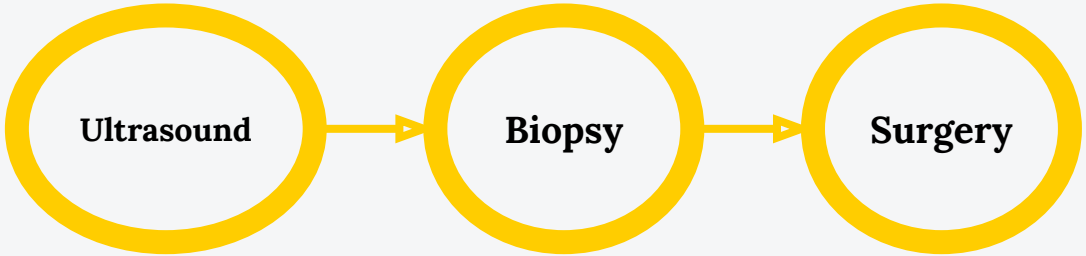
Change of EMR
All data looks different now



Nestor et al, MLHC, 2019

Lack of Context



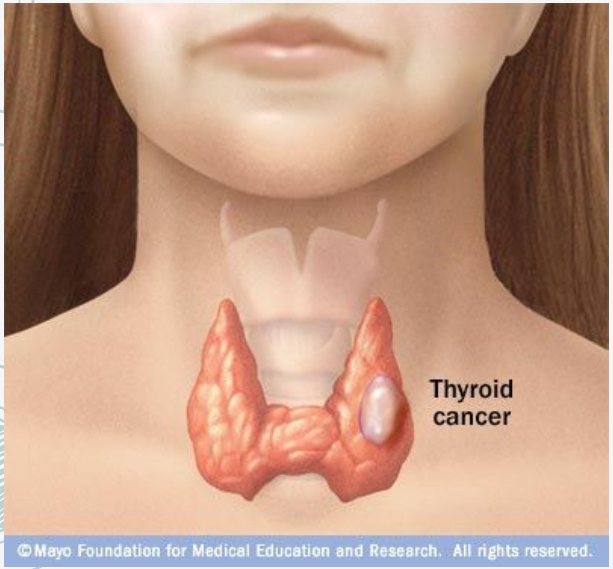


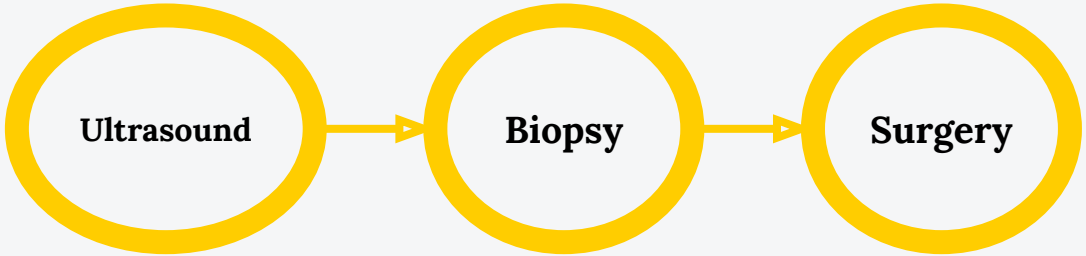
Unnecessary: 67% of surgeries ⇨ ~30%

Need classifier whether to have surgery

Radebe et al, in preparation

Lack of Context





Unnecessary: 67% of surgeries ⇨ ~30%

Small validation set of 10 patients:
7 patients didn't need surgery (but had) -> predict 2/7
3 patients needed surgery -> predict ⅔

Lacking family history data

Radebe et al, in preparation

Encoding bias in the data

Classifier of no-show for appointments
Learned to discriminate based on race and SES

Result: overbooked appointments only for African American and poor people

From personal communication

Sensitivity to noise

Turtle

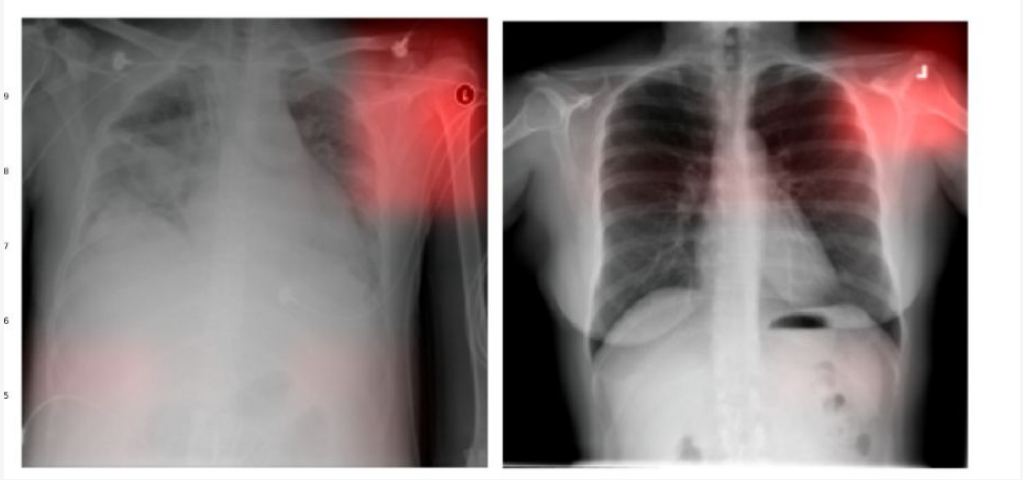
+

Noise

=

Rifle

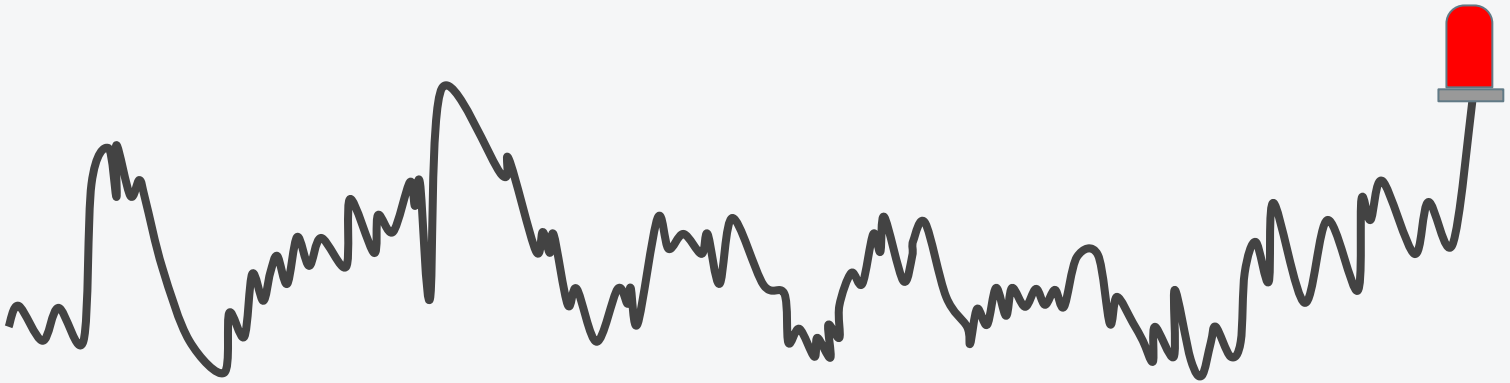
Are we learning about disease or artifacts?



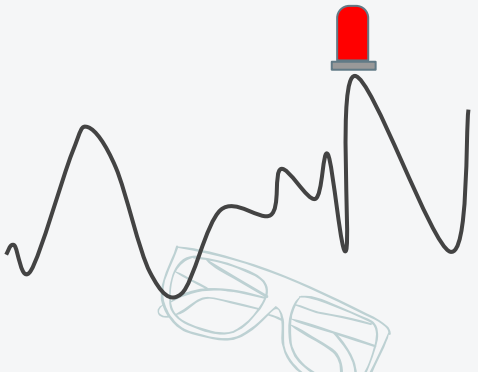
Pneumonia==metal token??

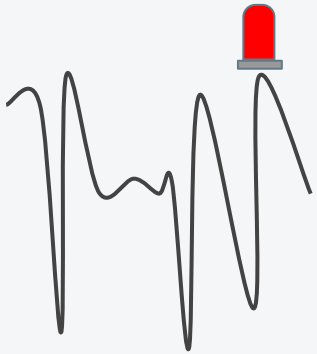
Source: Zech et al (2018)

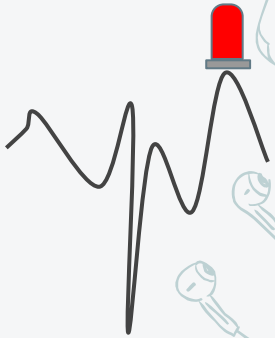
High False Positive Rate



1% error = 30,000+ false positives







Big or small you need to have the right data

Tukey:

> “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…”

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
- Check your modelling assumptions
- Check the performance of your model (does it make sense?)
- Off we go!