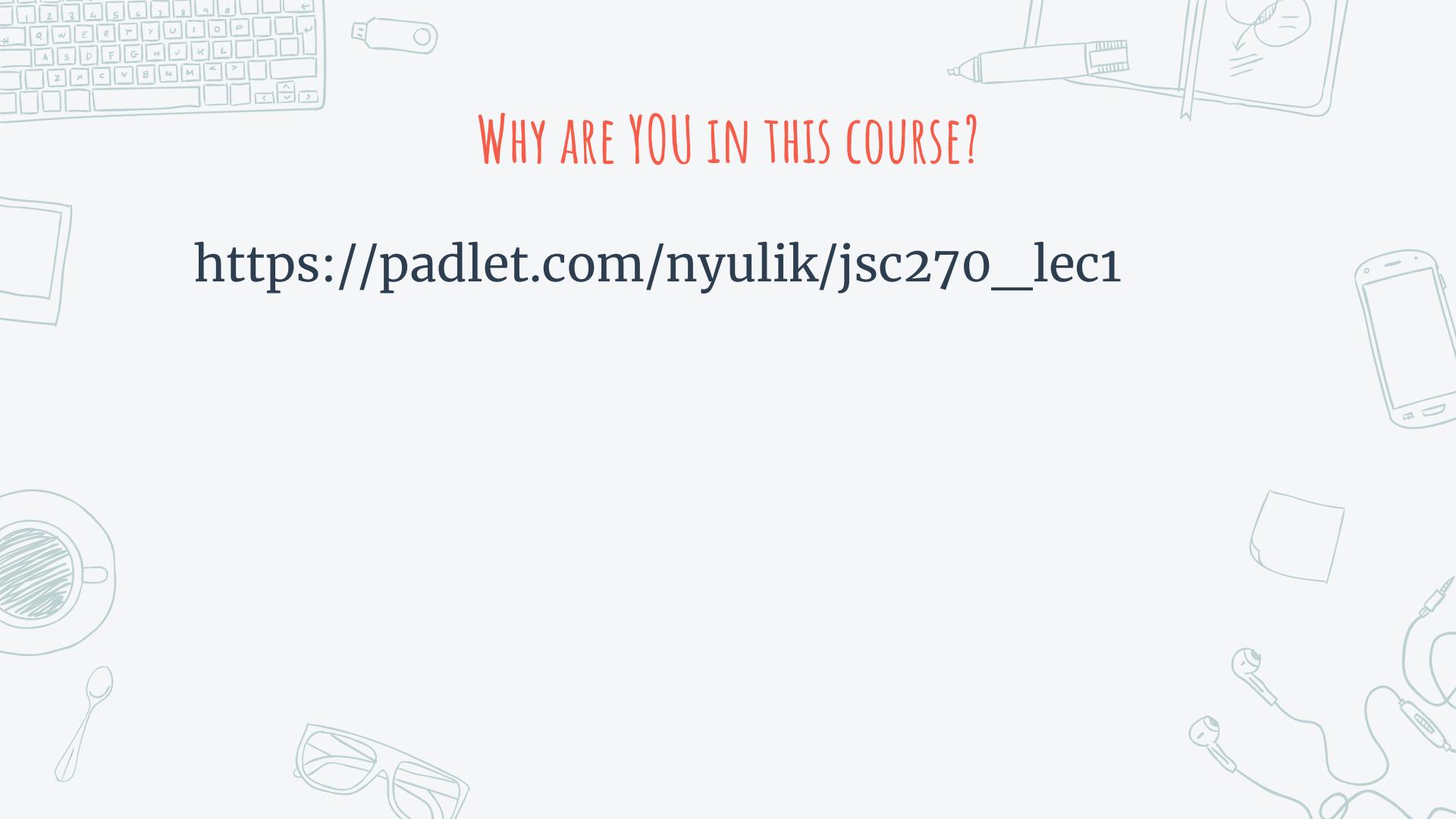


JSC 270 - LECTURE 1

<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

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/>

This screenshot shows a browser window for the Perusall web application. The URL is https://app.perusall.com/courses/jsc270/assignments/5166297c9f0e. The page title is "Perusall" and the course name is "JSC270". The main content area displays an assignment titled "DataScience vs BigData 2013.1508" due on Jan 16, 2021 at 8:59 am PST. The assignment details state: "Data Science vs Big Data" and "Due Mon Jan 16, 2021 8:59 am PST". Below the assignment title, there is a note: "Please read and annotate this paper with your comments. Don't hesitate commenting on each other's comments. Discussions and questions are very welcome!". A progress bar indicates "Assignment is in progress. You have started 0 conversations and have posted 0 follow-up answers or comments. 0 people responded to or uploaded your annotations." At the bottom of the assignment view is a large green button labeled "Work on assignment". On the left side of the screen, there is a sidebar with various course navigation links: "My Courses", "Course home", "My scores", "Notifications", "About", "Add to my calendar", and "Unread from course". Below this, under "Readings", there is a "Library" section with "DataScience vs BigData 2013.1508" and "Data Science vs Big Data". Under "Assignments", it says "Jan 16 Data Science vs BigData 2013.1508". Under "Chats", "Groups", and "One-on-One", there are no items listed. Under "Hashtags", there are "#grades", "#homework", "#logistics", and "#election". The status bar at the bottom shows "9:11 AM" and "2021-01-16".



This screenshot shows a browser window for the Perusall web application. The URL is https://app.perusall.com/courses/jsc270/assignments/5166297c9f0e#comment-115507. The page title is "Perusall" and the course name is "JSC270". The main content area displays a discussion thread. The first comment is by "DataScience vs BigData 2013.1508" on Jan 16, 2021 at 8:59 am PST, with the text: "Companies have realized they need to hire data scientists, academic institutions are scrambling to put together data science programs, and publications are touting data science as a hot—yet “sexy”—career choice. However, there is confusion about what exactly data science is, and this confusion could lead to disillusionment as the concept diffuses into meaningless buzz." Below this, there is a section titled "In this article, we argue that there are good reasons why it has been hard to pin down exactly what is data science. One reason is that data science is intricately intertwined with other important concepts also of growing importance, such as big data and data-driven decision making. Another reason is the natural tendency to associate what a practitioner does with the definition of the practitioner's field; this can result in overlooking the fundamental nature of data. We believe that trying to pin down the boundaries of data science precisely is not of the utmost importance. What is more important is to understand the field in an academic sense, for in order to better serve business effectively, it is important (i) to understand its relationships to other important related concepts, and (ii) to begin to identify the fundamental principles underlying data science. Once we embrace (ii), we can much better understand and explain exactly what data science has to offer. Furthermore, only once we embrace (ii) should we be comfortable calling it data science. In this article, we present a perspective that addresses all these concepts. We close by offering, as examples, a partial list of fundamental principles underlying data science." At the bottom of the comment, there is a link to "Download PDF" and a DOI number: "DOI: 10.1089/big.2013.1508 • MARY ANN LIBERT, INC. • VOL. 1 NO. 1 • MARCH 2013 BIG DATA BD51". On the right side of the screen, there is a sidebar with "Current conversation" and a message input field: "Enter your comment or question and press Enter. Mention a friend by typing @. Add highlighting by typing #". The status bar at the bottom shows "9:11 AM" and "2021-01-16".



<https://jsc270.github.io/>

SYLLABUS - REFLECTION QUIZZES

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

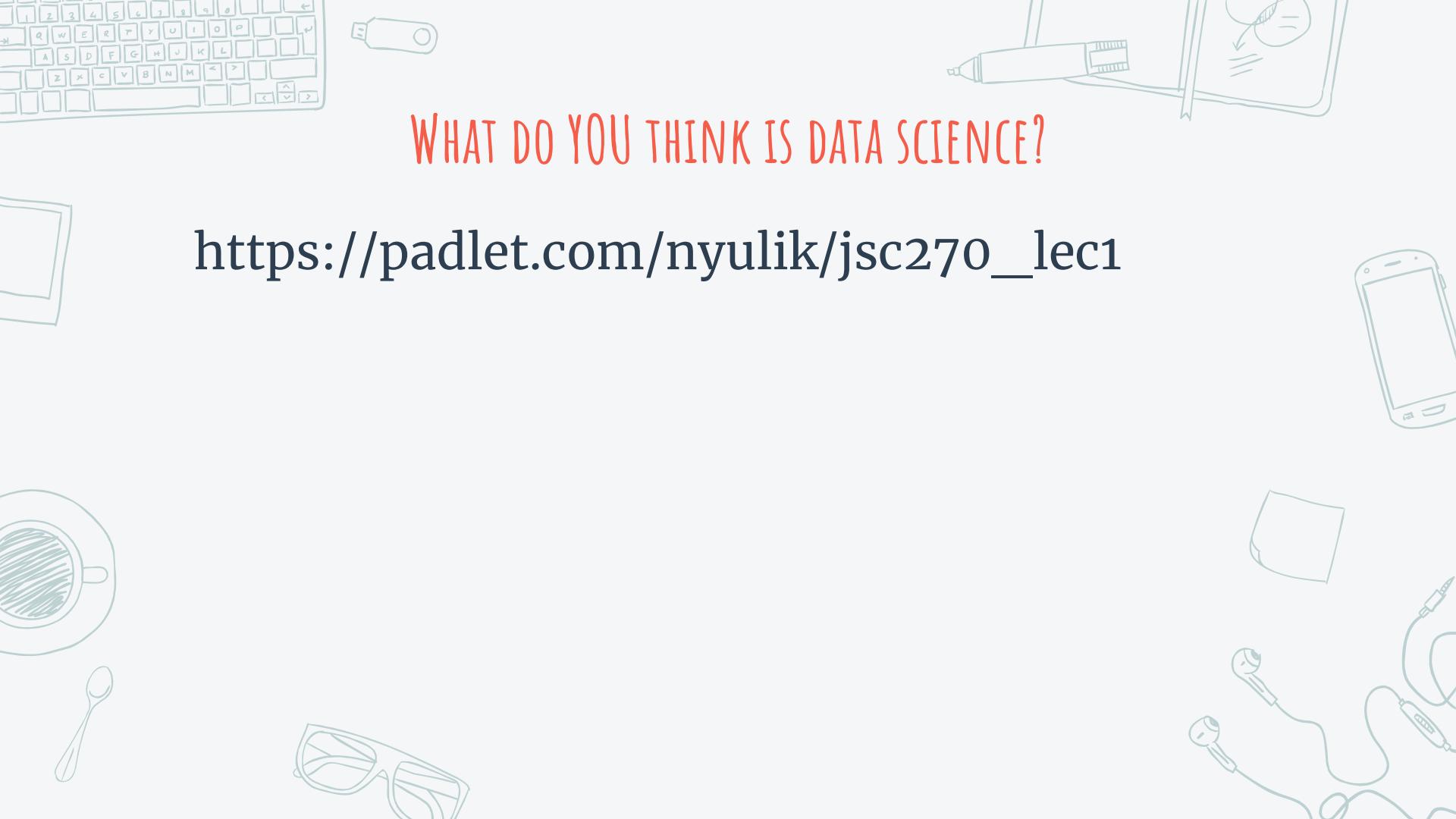
<https://jsc270.github.io/>

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?

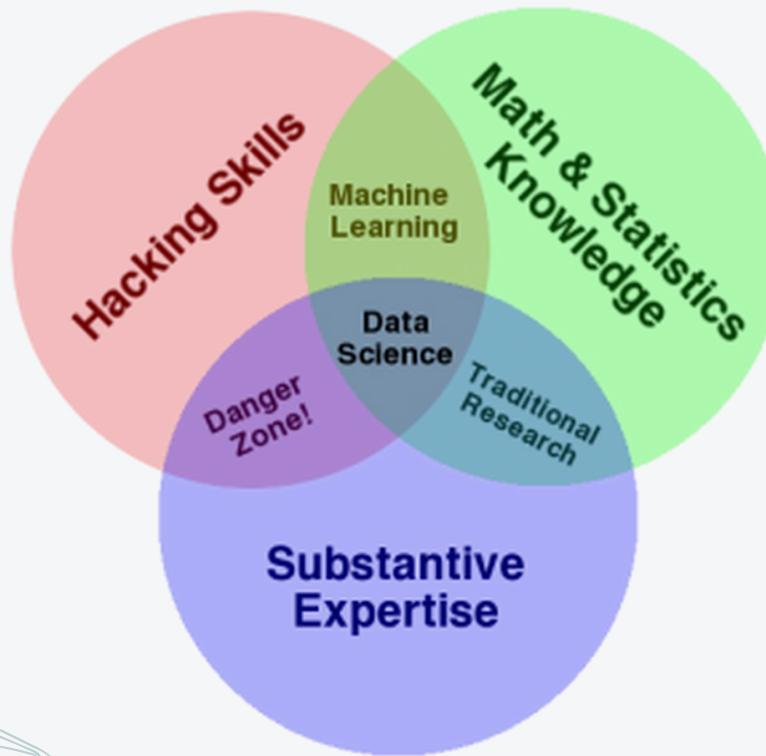
https://padlet.com/nyulik/jsc270_lec1

WHAT OTHERS THINK IS DATA SCIENCE

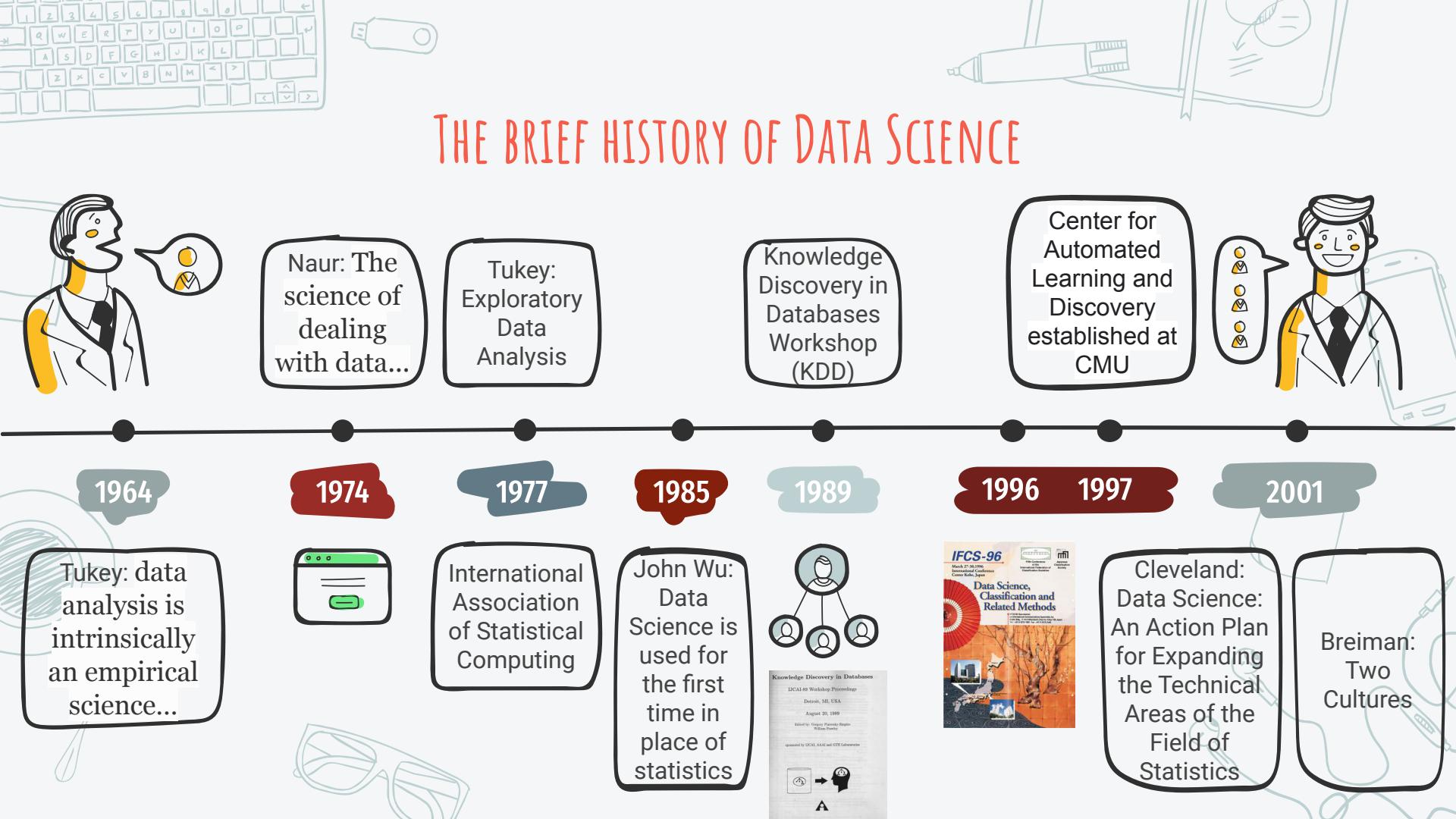
By ‘Data Science’ we mean almost everything that has something to do with data: Collecting, analyzing, modeling..... yet the most important part is its applications--all sorts of applications. This journal is devoted to applications of statistical methods at large....

Journal of Data Science, 2003

FROM THE TAXONOMY OF DATA SCIENCE (2010)



THE BRIEF HISTORY OF DATA SCIENCE



THE BRIEF HISTORY OF DATA SCIENCE (CONT'D)

2002 - Launch of Data Science Journal

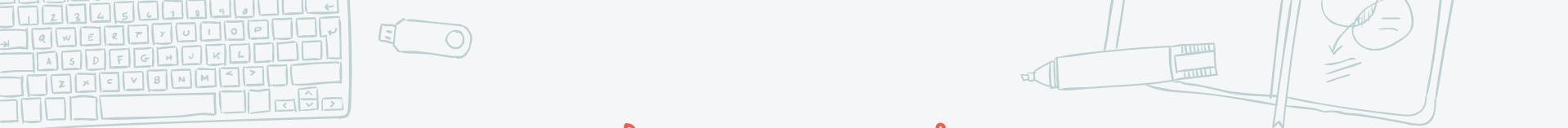
2003 - Launch of Journal of Data Science

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!)

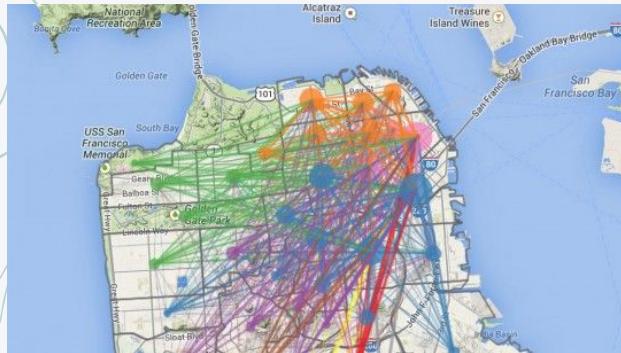


BUT WHAT IS DATA?

genetic data

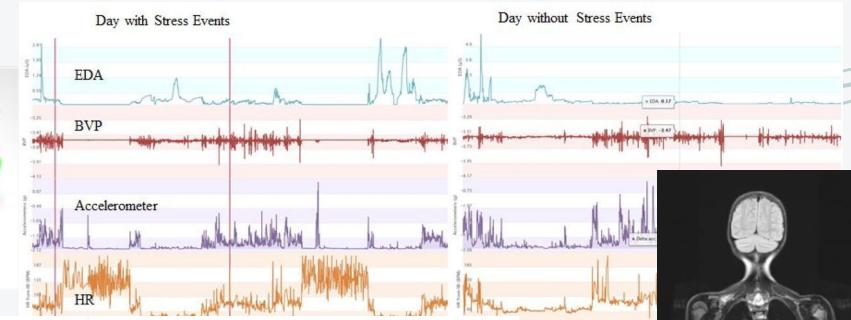
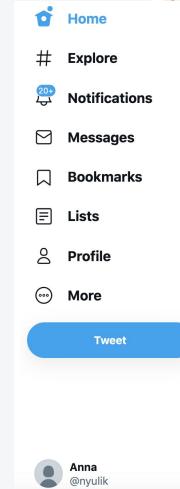
170 180 190

```
ATCTCTTGGCTCCAGCATCGATGAAGAACGCA
TCATTTAGGAAAGTAAAAGTCGTAACAAAGGT
GAACTGTCAAACCTTTAACAAACGGATCTCTT
TGTTGCTTCGGCGGGCGCAAGGGTGCCTC
GCCCTGCCGTGGCAGATCCCCAACGCCGGCC
TCTCTTGGCTCCAGCATCGATGAAGAACGAG
CAGCATCGATGAAGAACGAGCGAAACGCGAT
CGATACTTCAGTGGTCTTAGCGAAACTGTCA
CGGATCTTGGCTCCAGCATCGATGAAGAAC
AACACGGATCTTGGCTCCAGCATCGATGAA
CGGATCTTGGCTCCAGCATCGATGAAGAAC
GATGAAGAACGAGCGAAACGCGATATGTAAT
```



questionnaire

strongly agree
Agree ✓
Disagree
v disagree



-  Bernie Sanders  See new Tweets
Wealth of Elon Musk on March 18, 2020: \$24.5 billion
Wealth of Elon Musk on January 9, 2021: \$209 billion

U.S. minimum wage in 2009: \$7.25 an hour
U.S. minimum wage in 2021: \$7.25 an hour

Our job: Raise the minimum wage to at least \$15, tax the rich & create an economy for all.

6.2K 24.8K 133.9K
-  Kamala Harris  @KamalaHarris · Jan 7
We have witnessed two systems of justice: one that let extremists storm the US Capitol yesterday, and another that released tear gas on peaceful protesters last summer. It's simply unacceptable.

14.8K 53.3K 357.1K
-  Carlos D. Bustamante liked
 Andrew Yang  @AndrewYang · 9h
If you don't impeach a guy who sent a mob to your house that resulted in multiple deaths there's not much left.

2.9K 21.3K 159.3K
-  Jonathan Dursi liked
 Ilhan Omar  @ilhanMN · 6h
I will officially introduce two articles of impeachment against Donald J. Trump tomorrow.

1) Abuse of power for attempting to overturn the election results in Georgia.
2) Incitement of violence for orchestrating an attempted coup against our country.





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

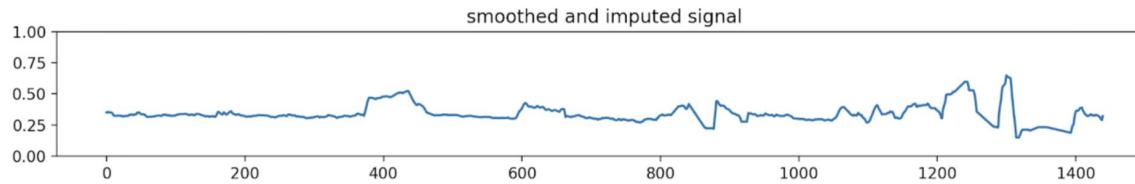
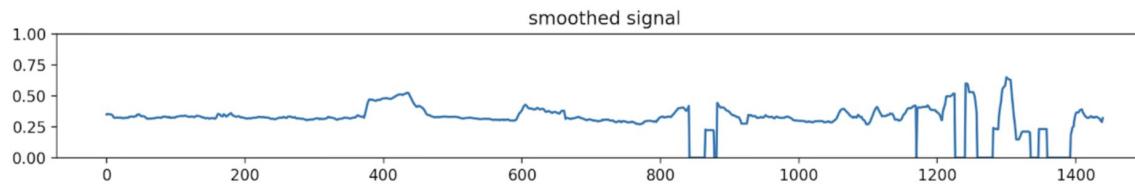
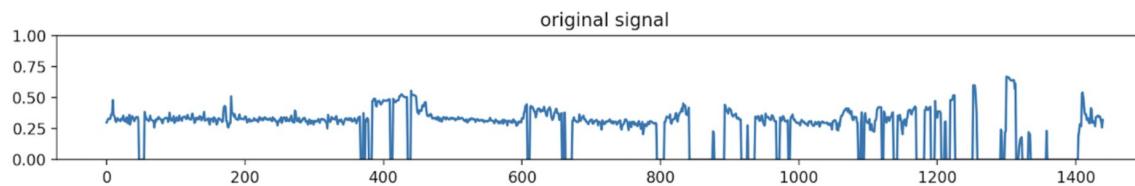
LinkedIn gained 120 new users

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

1.5 billion people ($\frac{1}{3}$ world population) are active on Facebook every day.

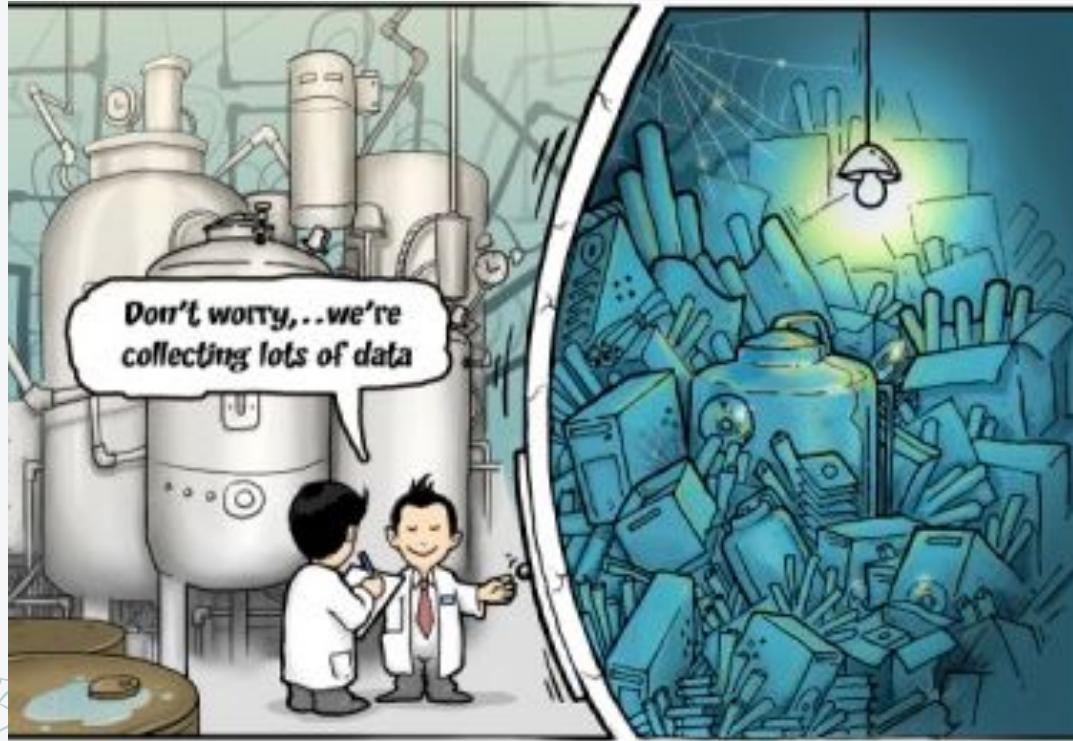
$\frac{2}{3}$ of the world's population now own a mobile phone.

RAW VS PROCESSED DATA





... AND THEN DATA SCIENCE ...



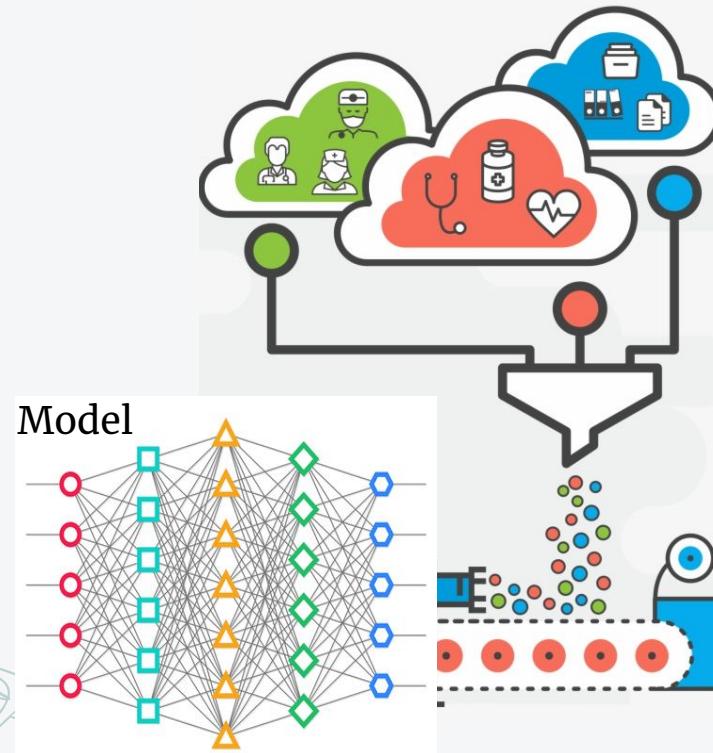
THE DATA SCIENCE PIPELINE



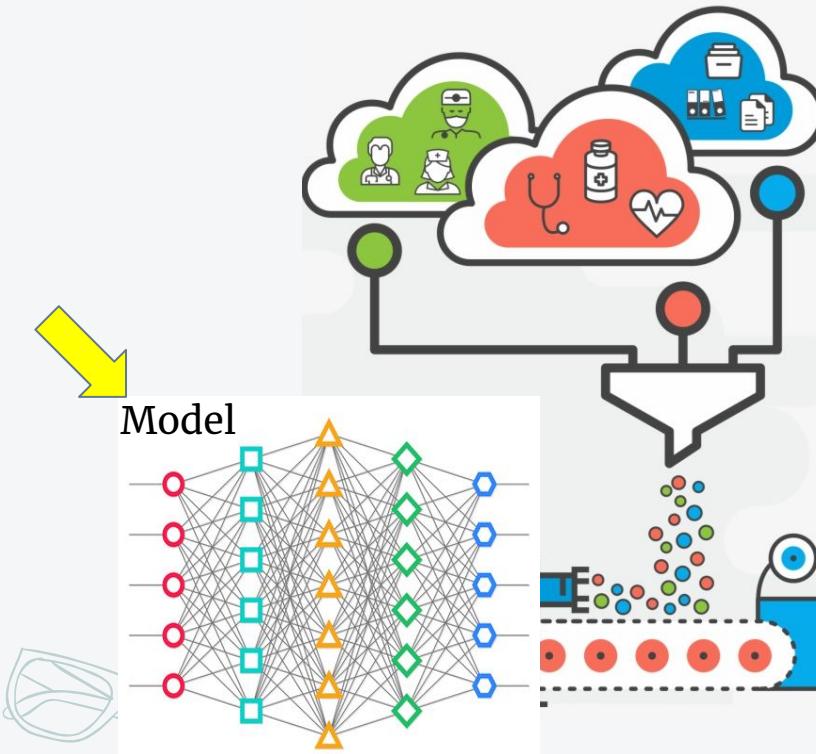
TAKING DATA SERIOUSLY



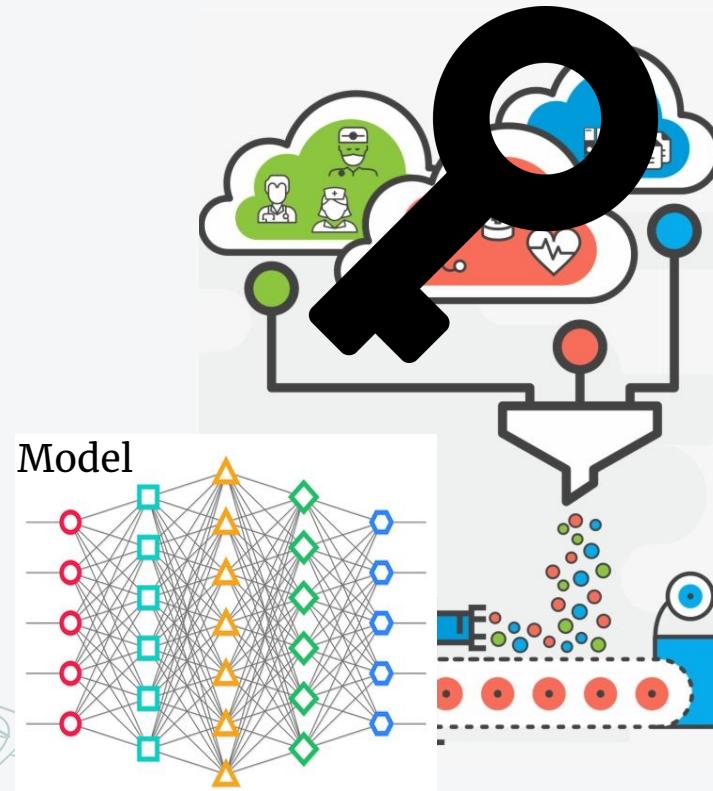
TAKING DATA SERIOUSLY



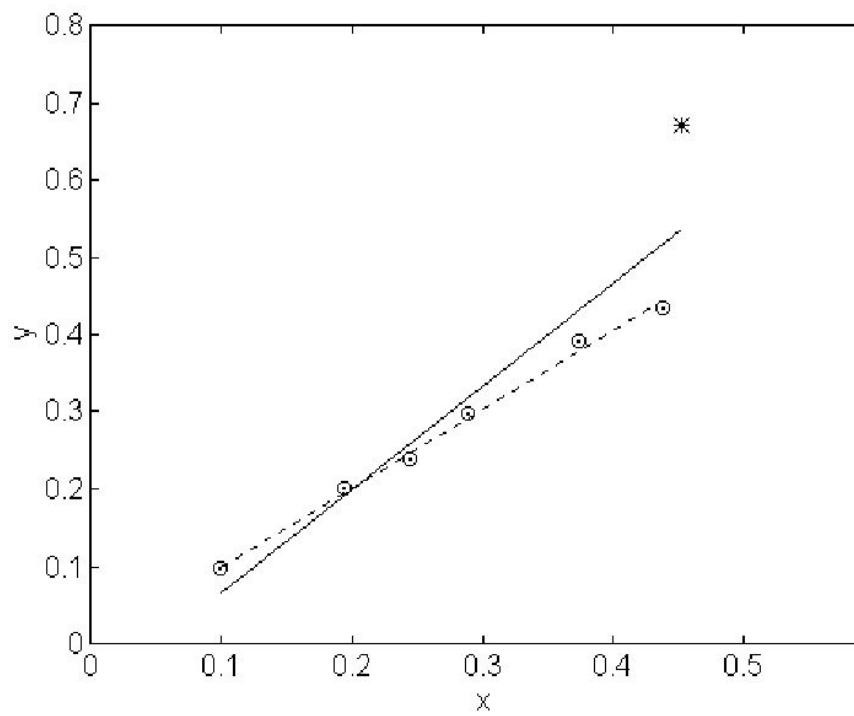
TAKING DATA SERIOUSLY



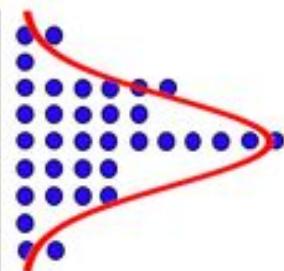
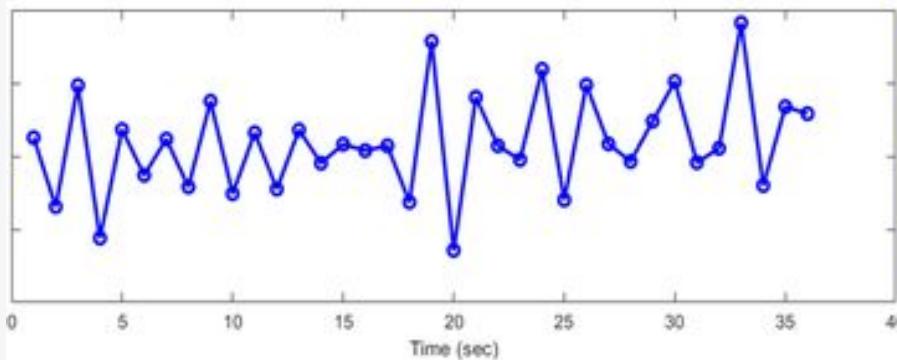
TAKING DATA SERIOUSLY



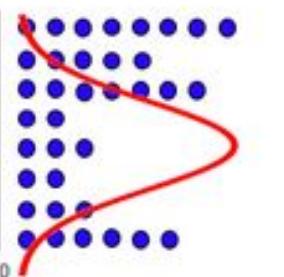
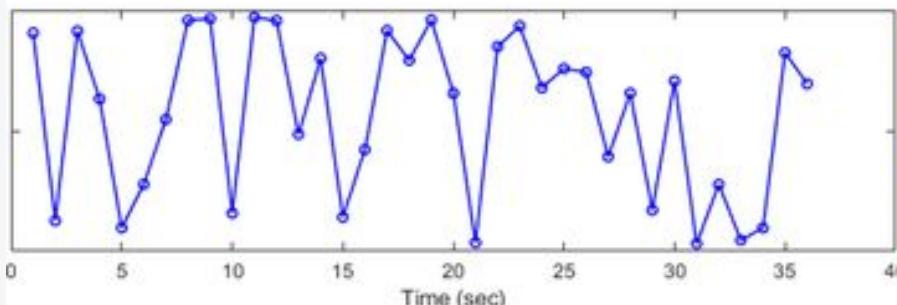
THE IMPORTANCE OF CLEANING



THE IMPORTANCE OF DATA EXPLORATION



The Gaussian distribution (red) fits the data well.



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

MODELING

- Descriptive
- Predictive

statistics

- Supervised
- Unsupervised

AI/ML

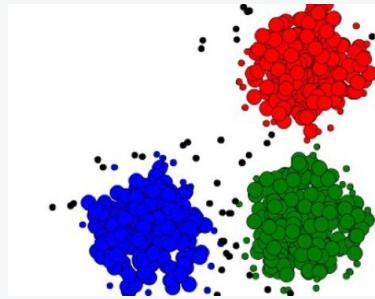
- Generative
- Discriminative

AI/ML/stats

DESCRIPTIVE VS PREDICTIVE

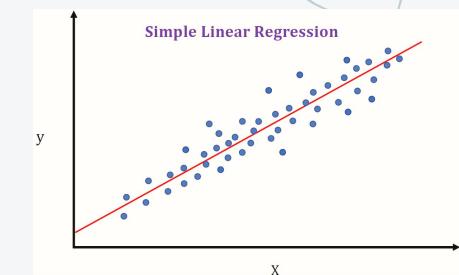
- Descriptive: analysis/model that describe the past

E.g. clustering



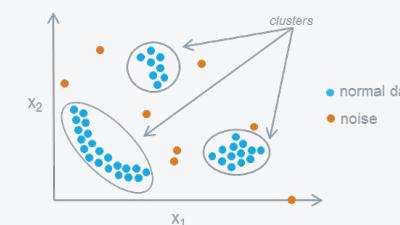
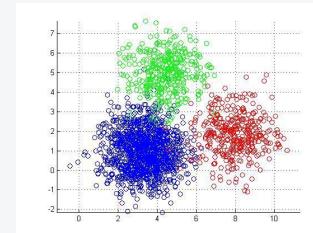
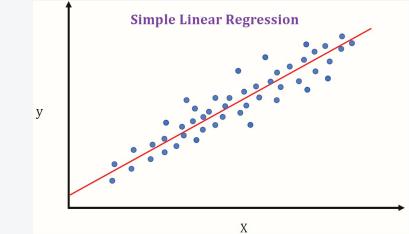
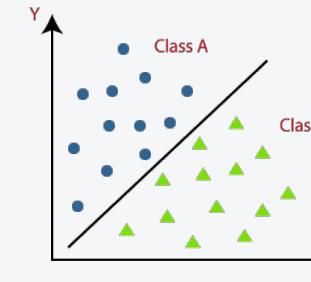
- Predictive: building a model that predicts y for the previously unseen x

E.g. regression



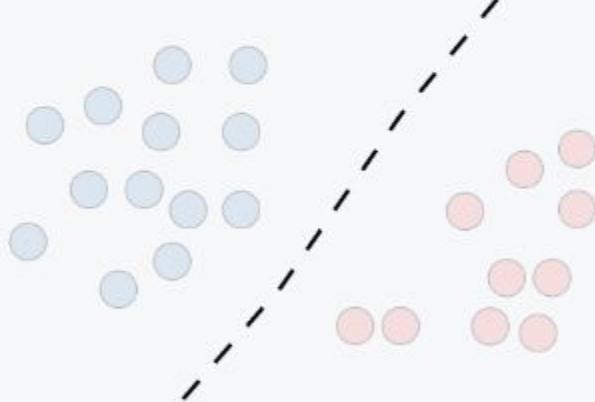
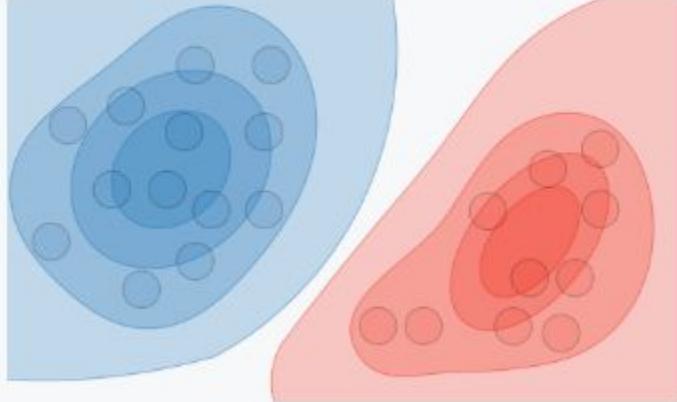
SUPERVISED VS UNSUPERVISED

- 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

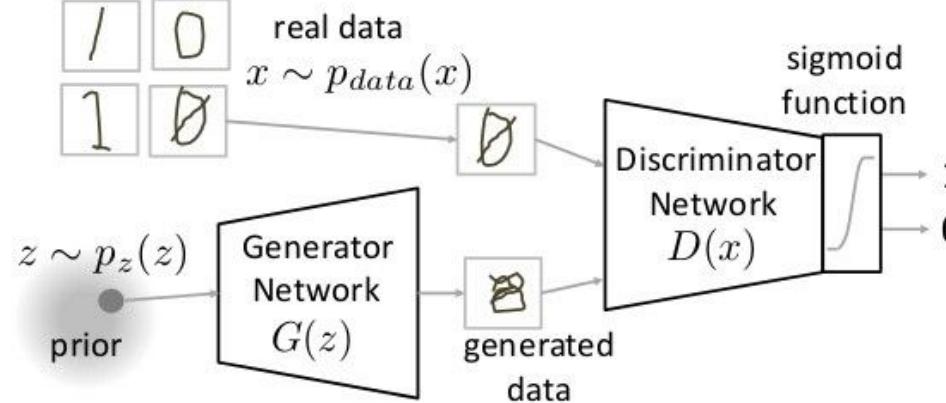
	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

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:

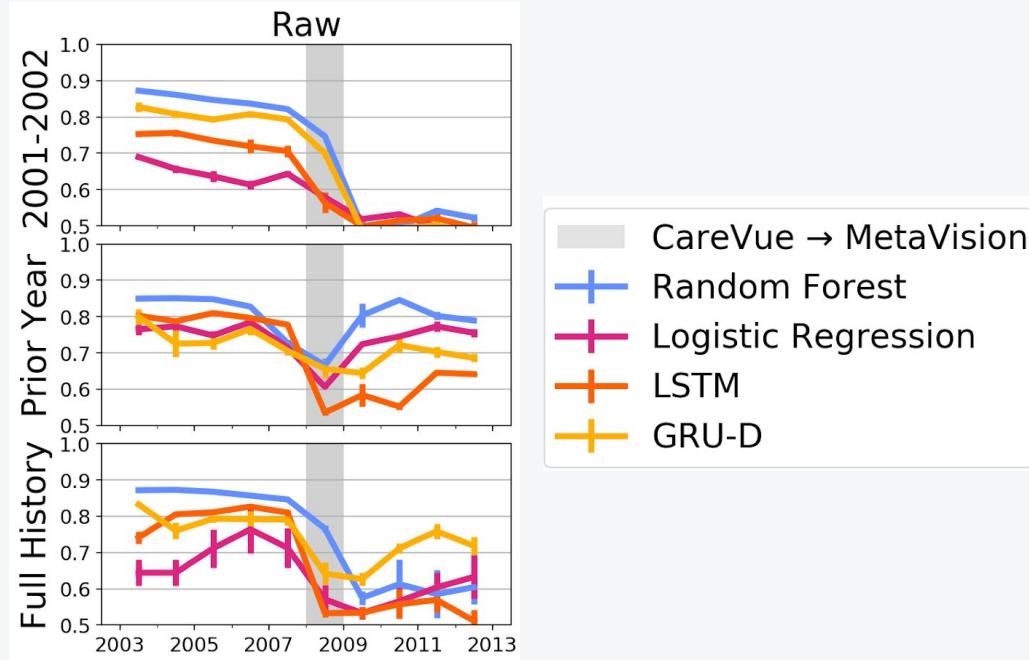
Patient with asthma has pneumonia and is treated more aggressively

Fewer patients with asthma die of pneumonia

Learned:

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

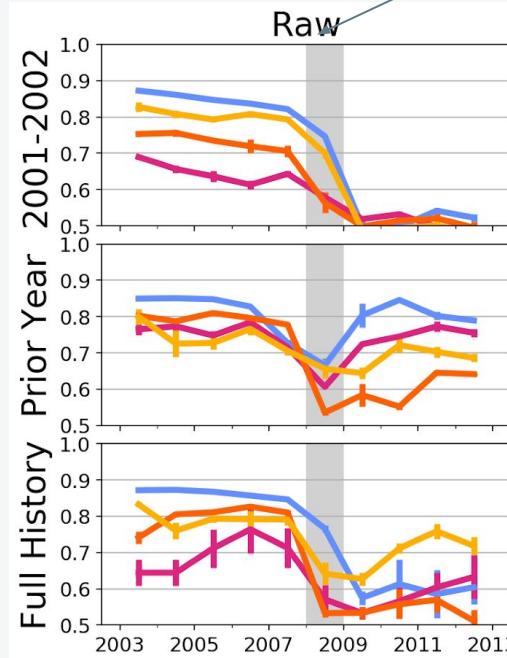
DRAMATIC DATA SHIFTS



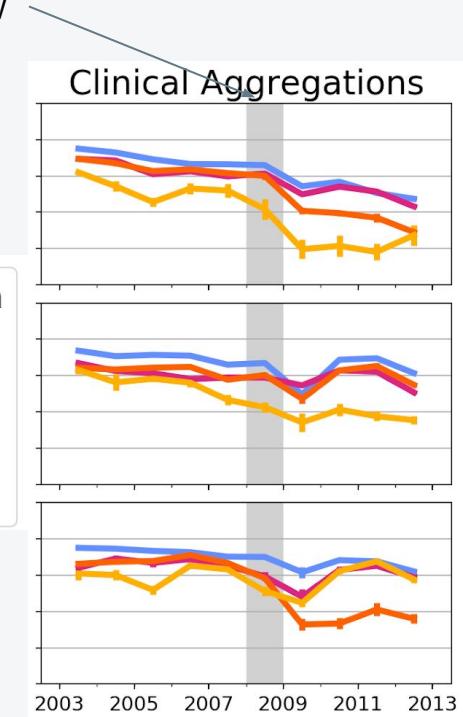
DRAMATIC DATA CHANGES

Change of EMR

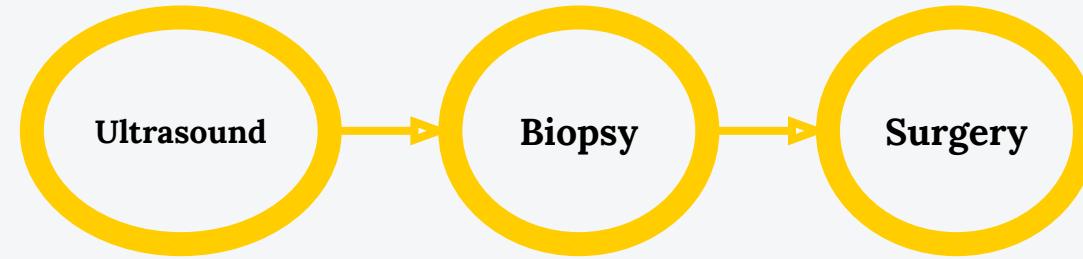
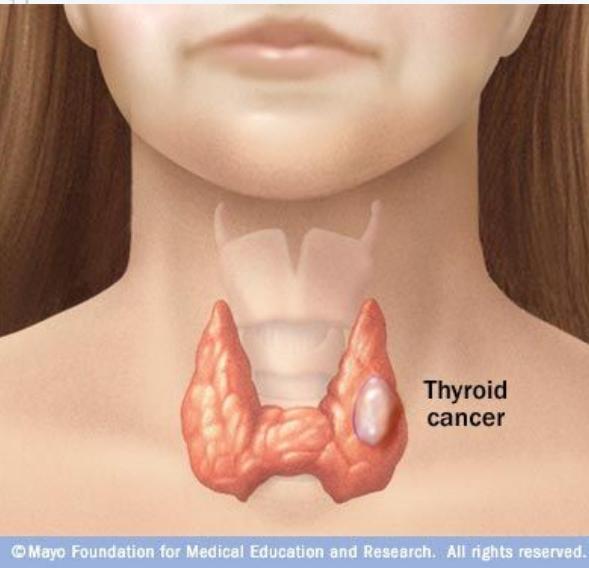
All data looks different now



- CareVue → MetaVision
- Random Forest
 - Logistic Regression
 - LSTM
 - GRU-D



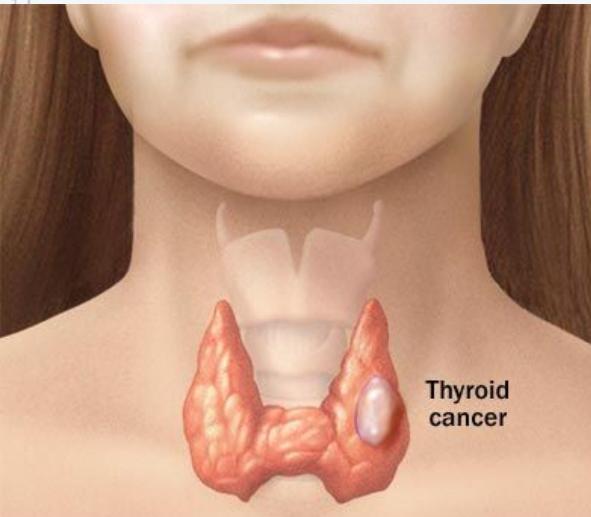
LACK OF CONTEXT



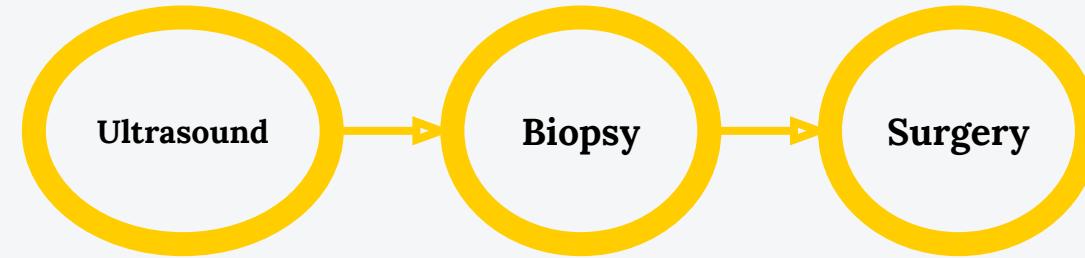
Unnecessary: 67% of surgeries \Rightarrow ~30%

Need classifier whether to have surgery

LACK OF CONTEXT



© Mayo Foundation for Medical Education and Research. All rights reserved.



Unnecessary: 67% of surgeries \Rightarrow ~30%

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

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

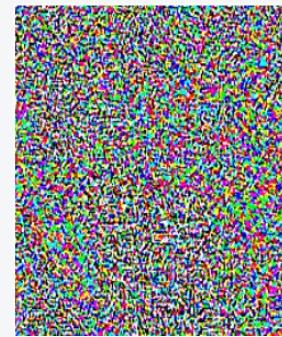
Result: overbooked appointments only for
African American and poor people

SENSITIVITY TO NOISE



Turtle

+



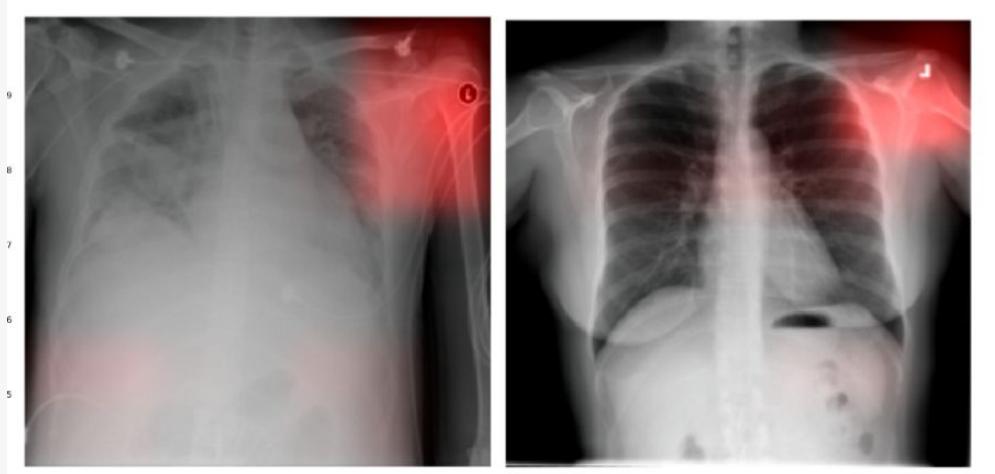
Noise

=



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!