Applications of Data Science, Session 1: 1-4 PM PT

While we wait to get started:

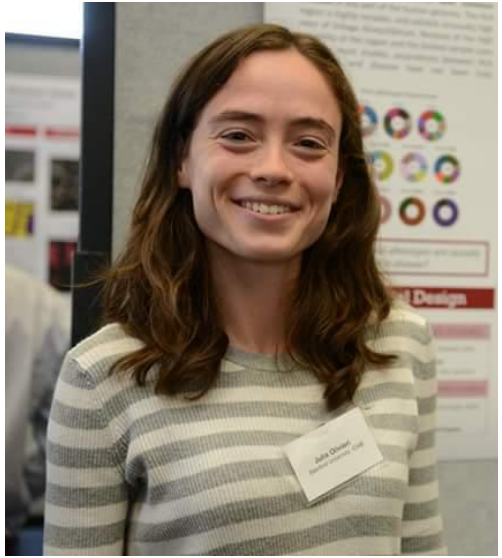
1. Log into Piazza
2. Locate the "Introductions" thread
3. Introduce yourself and share 1) the city you're Zooming in from, and 2) what you hope to get out of this workshop.

Session 1:

Applications of Data Science

Instructor: Julia Olivieri
TA: Julie Wang

Julia Olivieri



University of the Pacific

Welcome to Applications of Data Science

- TA: Julie Wang
- Don't message questions in Zoom: Use the 7/28 Workshop Questions thread in Piazza
- There will be two five-minute breaks, one after each of the first two modules
- Meeting time: today 1-4 PM, Friday August 4th 1-4 PM

For questions during class: Post in Piazza

Cameras on or off?

- I'll let you know when I think it would be helpful to turn cameras on
- Feel free to keep your cameras off for now
- I encourage you to add a profile picture to Zoom if you haven't already

Active learning through in-class Exercises

- Exercises provide opportunities to actively engage with the material
- We will devote time to working on these exercises throughout the workshop
- You will be requested to share your work for each exercise on Piazza
- I recommend you open a document to work on the exercises now
- You are encouraged to post with your name, though anonymous posting is still allowed
- You are encouraged to interact with each others' answers: upvote answers you agree with, comment if you have something else to say

Schedule for today

Module 1: What is Data Science?

Questions

Break

Module 2: What makes an application suitable for data science?

Questions

Break

Module 3: Prediction: The art of forecasting

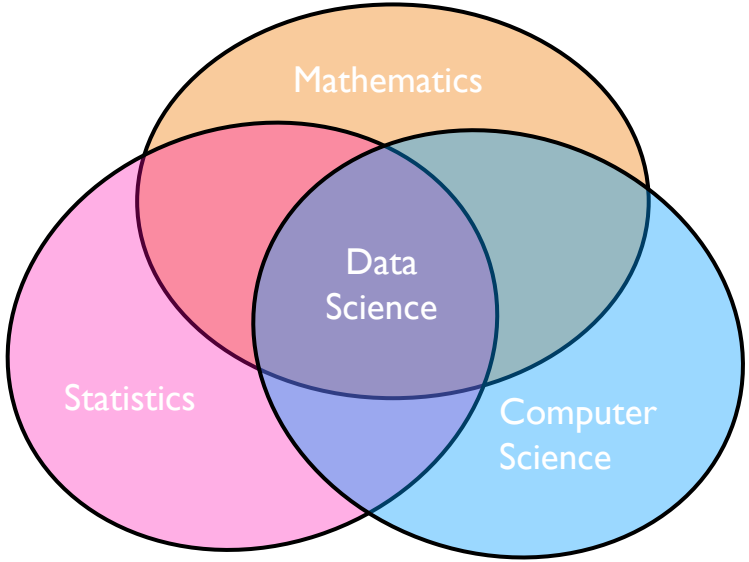
Questions

Module 1: What is Data Science?

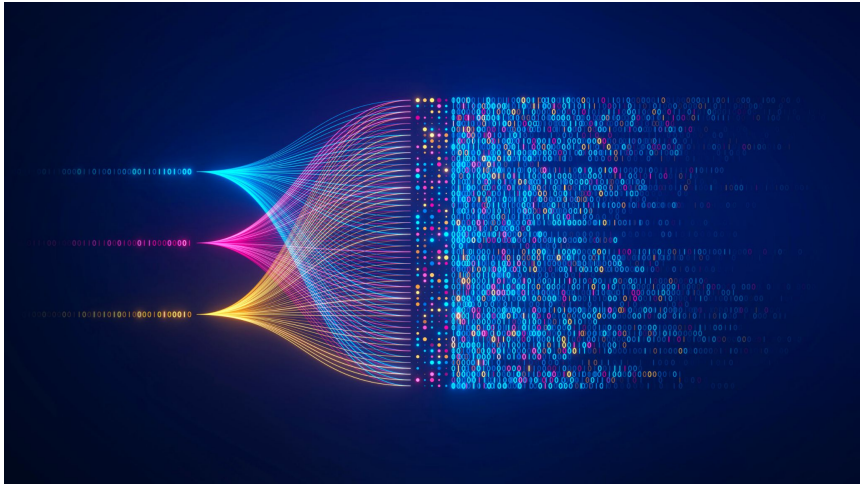
★ Exercise 1: How would you define "data science"?

How would you define "data science"?

- Data science involves using real-world data to answer questions and solve problems
- Involves methods to collect, analyze, and interpret data
- Often involves making predictions or decisions about the future

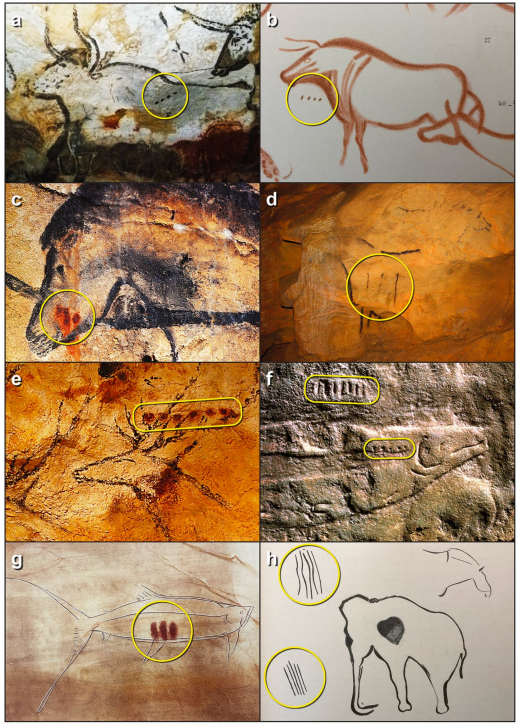


Data collection has ancient roots



Data collection has ancient roots

37,000 years ago



Bacon, B., Khatiri, A., Palmer, J., Freeth, T., Pettitt, P., & Kentridge, R. (2023). An Upper Palaeolithic Proto-writing System and Phenological Calendar. Cambridge Archaeological Journal, 33(3), 371-389. doi:10.1017/S095977432200041 5

Data-informed prediction of the future is not new

- Edmond Halley used Newton's laws of gravity and motion to calculate the effects of Jupiter and Saturn on the orbits of comets

- He collected 24 comet observations and realized that the orbital elements of a comet that had appeared in 1682 were the same as comets that had appeared in 1607 and 1531

- He concluded (correctly) that these were all the same comet, returning every 76 years

Takeaway: Data science isn't limited to the popular tools of today

We want to take an application-focused rather than method-focused perspective today

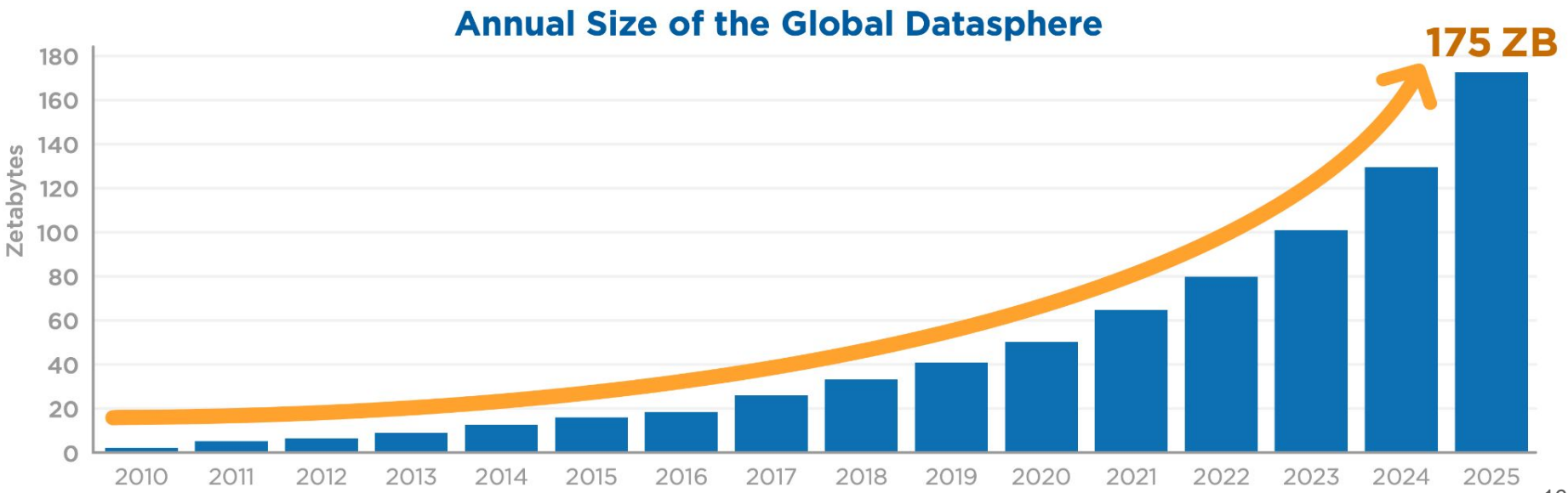
How much raw data is there?

2.5 quintillion bytes of data created each day

How much raw data is there?

90% of the data in the world was generated in the last two years (Forbes, 2018)

Figure 1 – Annual Size of the Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

FIGURE 22

Top 20 job roles in increasing and decreasing demand across industries

Increasing demand

1	Data Analysts and Scientists
2	AI and Machine Learning Specialists
3	Big Data Specialists
4	Digital Marketing and Strategy Specialists
5	Process Automation Specialists
6	Business Development Professionals
7	Digital Transformation Specialists
8	Information Security Analysts
9	Software and Applications Developers
10	Internet of Things Specialists
11	Project Managers
12	Business Services and Administration Managers
13	Database and Network Professionals
14	Robotics Engineers
15	Strategic Advisors
16	Management and Organization Analysts
17	FinTech Engineers
18	Mechanics and Machinery Repairers
19	Organizational Development Specialists
20	Risk Management Specialists

Decreasing demand

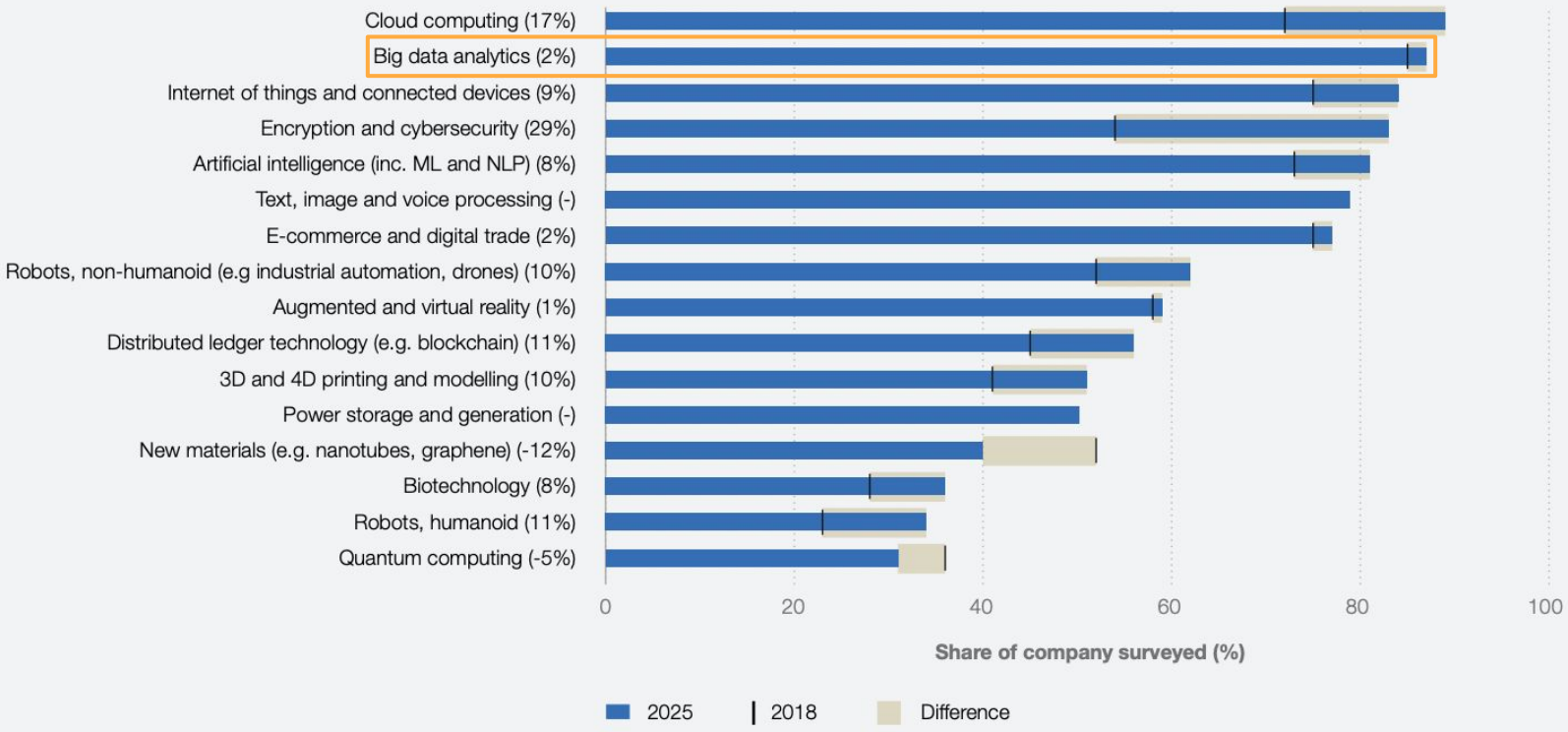
1	Data Entry Clerks
2	Administrative and Executive Secretaries
3	Accounting, Bookkeeping and Payroll Clerks
4	Accountants and Auditors
5	Assembly and Factory Workers
6	Business Services and Administration Managers
7	Client Information and Customer Service Workers
8	General and Operations Managers
9	Mechanics and Machinery Repairers
10	Material-Recording and Stock-Keeping Clerks
11	Financial Analysts
12	Postal Service Clerks
13	Sales Rep., Wholesale and Manuf., Tech. and Sci.Products
14	Relationship Managers
15	Bank Tellers and Related Clerks
16	Door-To-Door Sales, News and Street Vendors
17	Electronics and Telecoms Installers and Repairers
18	Human Resources Specialists
19	Training and Development Specialists
20	Construction Laborers

Source

Future of Jobs Survey 2020, World Economic Forum.

FIGURE 18

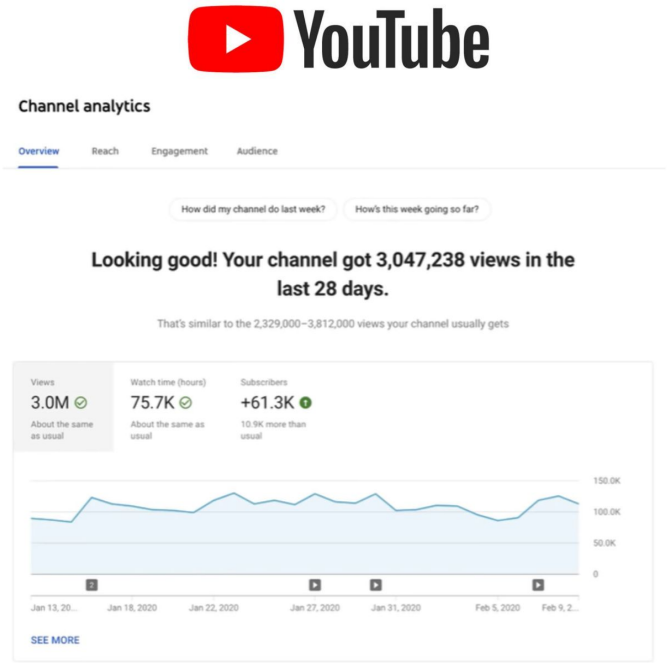
Technologies likely to be adopted by 2025 (by share of companies surveyed)

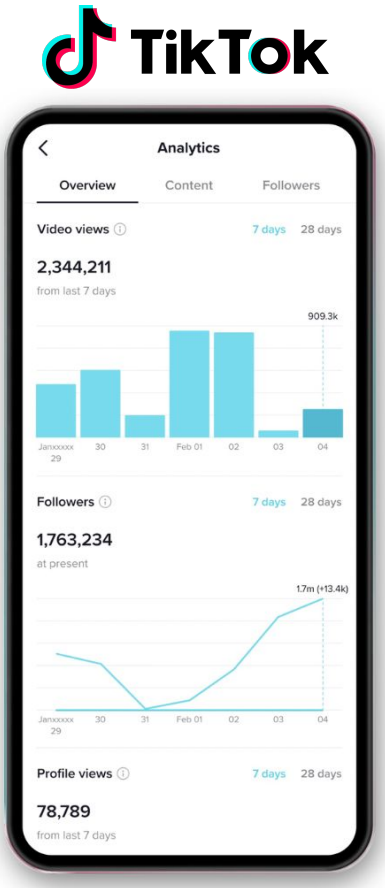


Source

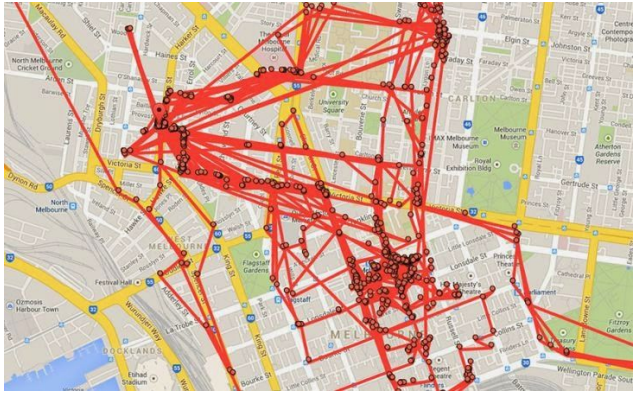
Future of Jobs Survey 2020, World Economic Forum.

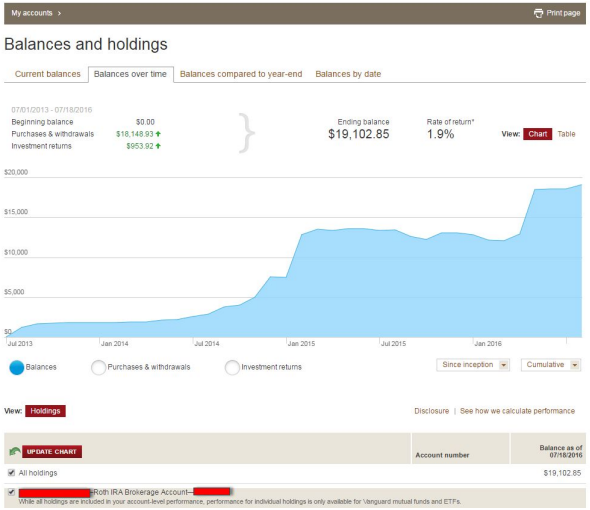
Data is available all around us

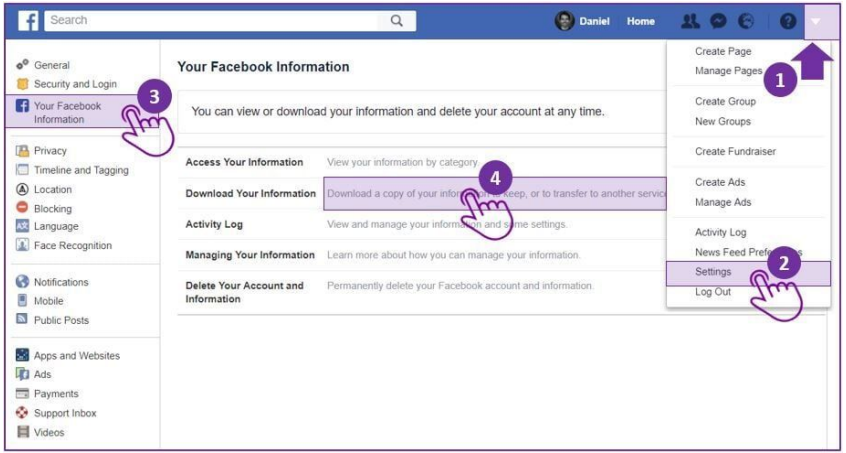




Personal data analytics

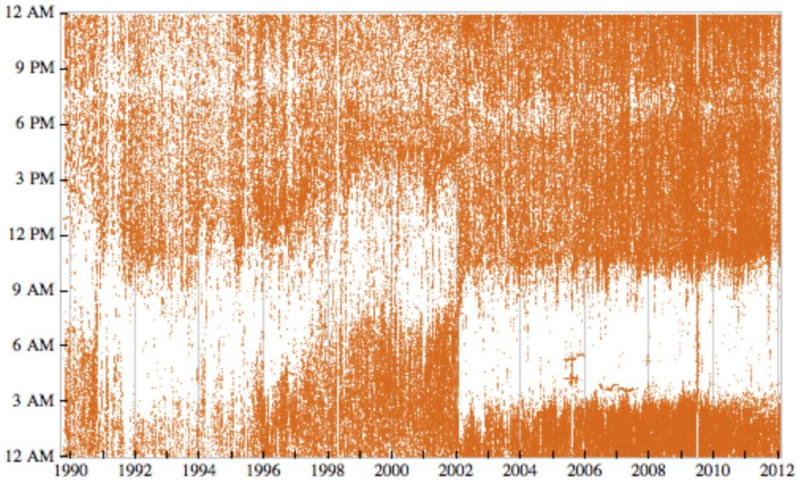






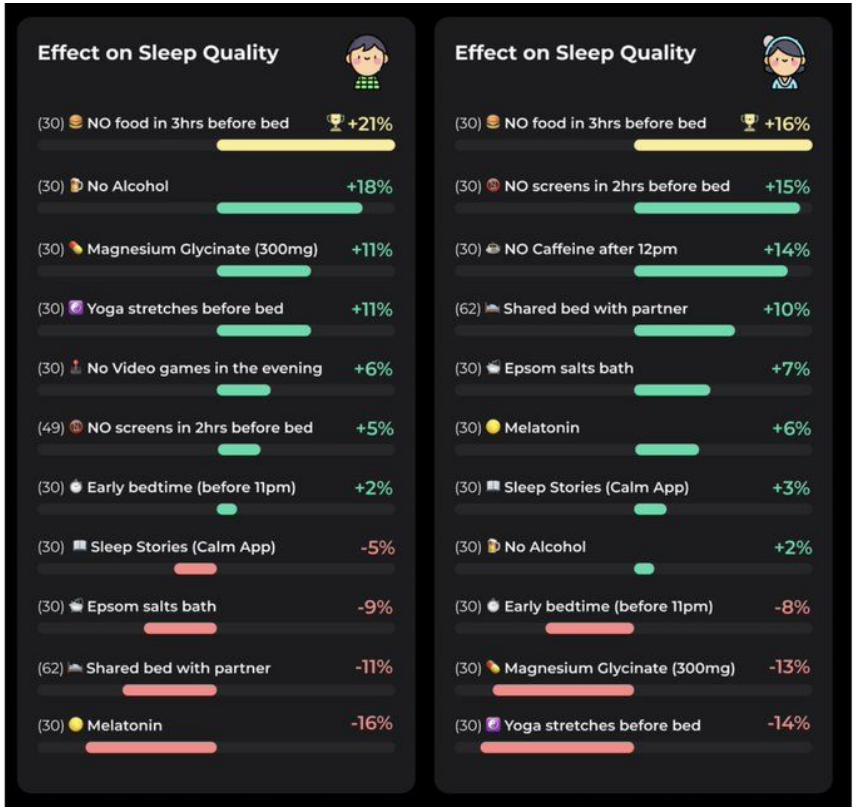
Stephen Wolfram tracked his personal data for 20 years

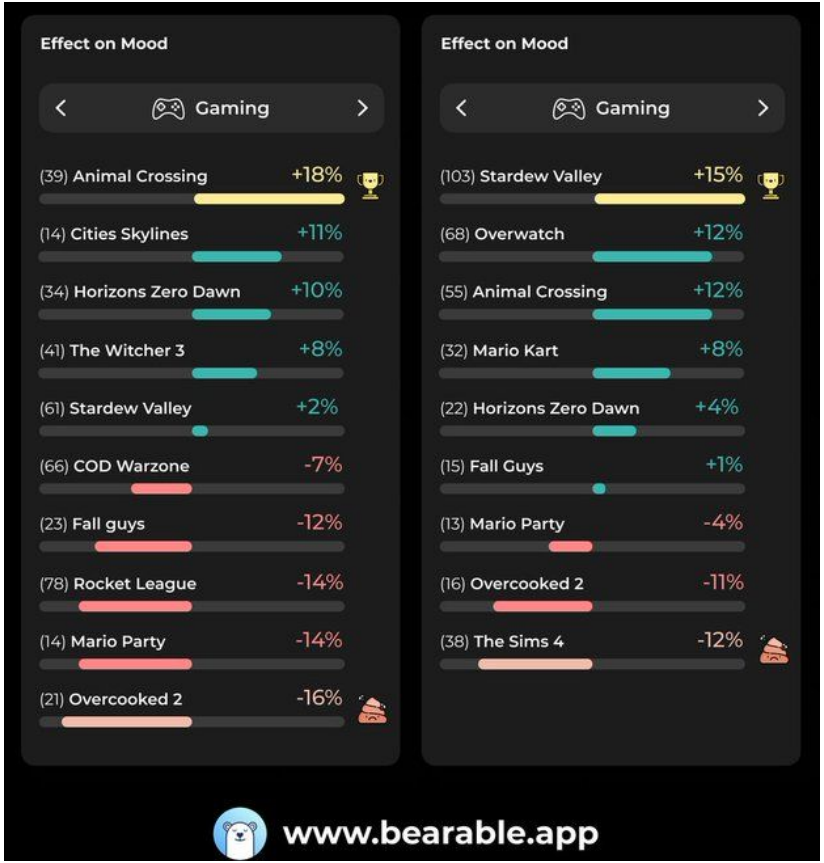
Each orange dot is the time of an email



https://writings.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/

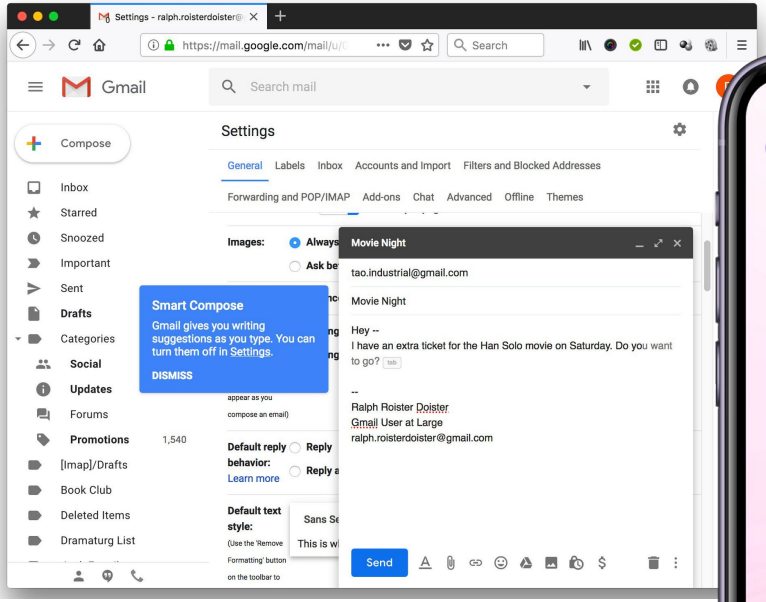
Data science on a personal level

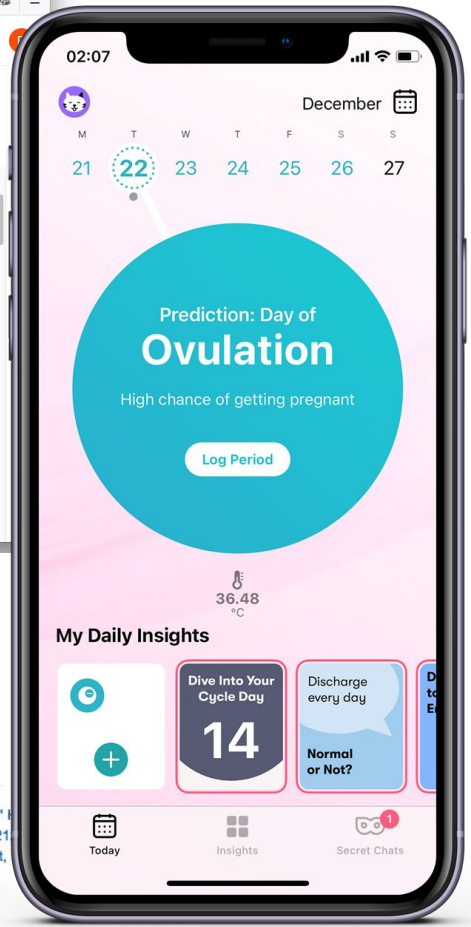




https://twitter.com/gregisenberg/status/1596160063227408384

Data science in daily life





★ Exercise 2: Name other applications of data science that you notice in your daily life

Questions?

Break!

Module 2: What makes an application suitable for data science?

★ Exercise 3: Which problem(s) are not currently well-solved by data science?

Features that make a problem suitable for data science

1. Quantifiability
2. Data availability
3. Data quality
4. Predictability
5. Understanding the impact

Quantifiability

- Desired output can be quantified or measured in some way
- Quantifiable outputs:
  - Temperature
  - Amount of money earned
  - Strength of earthquake
- Unquantifiable outputs:
  - Happiness
  - Animal intelligence
  - Strength of friendship

Quantifiability: Animal intelligence

Different definitions of intelligence:

- Brain size
- Number of neurons
- Ability to manipulate surroundings
- Ability to learn new things
- Ability to plan ahead
- Ability to solve artificial problems like puzzles and mazes

The general term "intelligence" isn't easily quantifiable

Data availability

- Variables necessary for prediction must be available
- Data available:
  - Youtube viewer counts
  - Basketball player stats
  - Crop yield
- Data unavailable:
  - Many health records due to privacy regulations
  - Netflix streaming numbers
  - Historical studies

Data availability: TV viewership

- Previously: Nielsen ratings were used to measure American TV viewership

- In the era of streaming, viewership numbers are often hidden even from the creators themselves

Show	NETWORK	RATING	OVERALL RANK	NETWORK RANK	RENEWED OR CANCELLED	NUMBER OF SEASONS	GENRE
Top 10 (24.8>)							
All in the Family	CBS	30.2	1st	1st	RENEWED	5	Sitcom
Sanford and Son	NBC	29.6	2nd	1st	RENEWED	4	Sitcom
Chico and the Man	NBC	28.5	3rd	2nd	RENEWED	1	Sitcom
The Jeffersons	CBS	27.6	4th	2nd	RENEWED	1	Sitcom
M*A*S*H	CBS	27.4	5th	3rd	RENEWED	3	Sitcom
Rhoda	CBS	26.3	6th	4th	RENEWED	1	Sitcom
Good Times	CBS	25.8	7th	5th	RENEWED	2	Sitcom
The Waltons	CBS	25.5	8th	6th	RENEWED	3	Drama
Maude	CBS	24.9	9th	7th	RENEWED	3	Sitcom
Hawaii Five-O						7	Drama
Mary Tyler Moore						5	Sitcom
The Rockford Files						1	Drama
Little House on a P						2	Drama
Kojak						2	Drama

‘I Don’t Know How My Show Is Doing’ Streamers run on data, but that doesn’t mean they’re sharing it with showrunners.

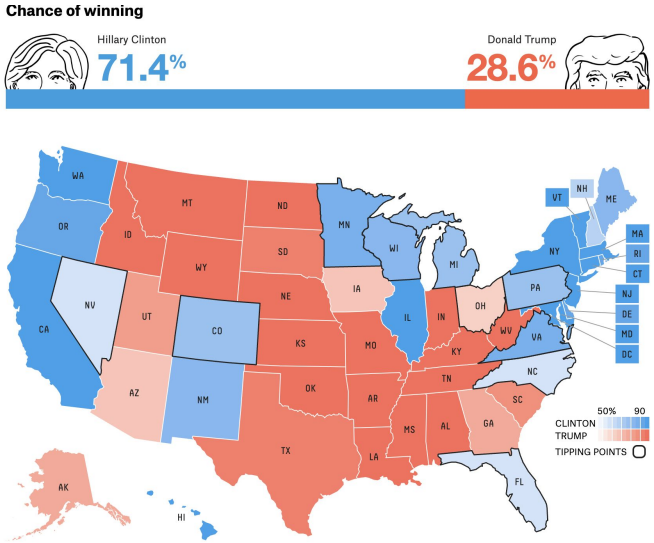
As told to Kathryn VanArendonk and Josef Adalian

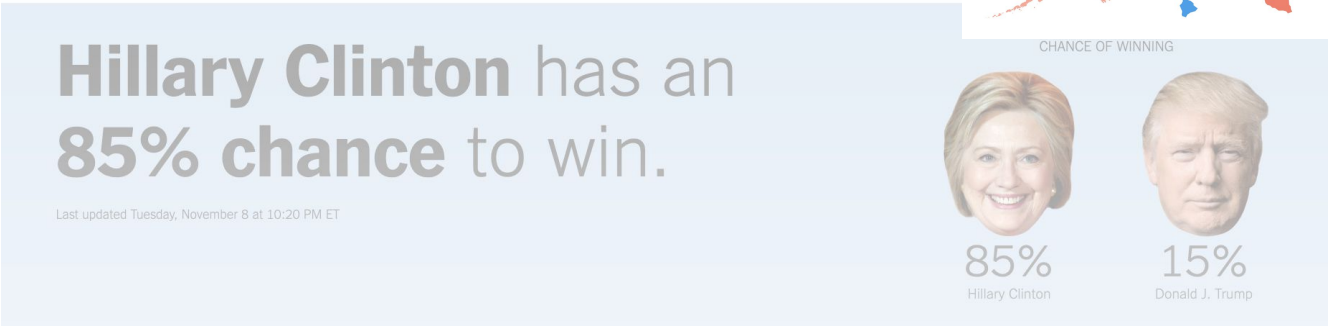
Data Quality

- Data should be of high enough quality to carry meaningful signal (and not be overly biased)
- High-quality data:
  - Contains all relevant information without missing values (complete sales record for a company)
  - Updated frequently
  - High confidence in each measurement
- Low-quality data:
  - Data from developing countries (inconsistent quality and coverage)
  - Data from previous years, e.g. pre- vs post-pandemic (outdated)
  - Data from opt-in surveys (biased)

Data Quality: 2016 US presidential election

- Most polls were not effective at predicting the result of the 2016 US presidential election
- Some of the error was due to "nonresponse bias," where different kinds of people are systematically less likely to respond to a poll





Predictability

- Desired output is at least partially based on patterns in the input data
- Predictable outputs:
  - Credit card fraud detection
  - Ovulation timing
  - Short-term weather
- Unpredictable outputs:
  - Stock market
  - Earthquakes
  - Long-term weather

Predictability: Earthquakes

- No major earthquake has ever been predicted
- The probability that a significant earthquake will occur in a specific area within a certain number of years is the best we can do
- Most earthquakes have no precursor events whatsoever

https://www.usgs.gov/programs/earthquake-hazards/national-

★ Exercise 4: Choose a data science application that hasn't had success yet. Which of the following attributes are missing?

1. Quantifiability
2. Data availability
3. Data quality
4. Predictability

Understanding the Impact

- Just because we can solve a problem, doesn't mean we should
- What are the impacts on society?
- Four principles of ethical research (from the Belmont Report (1979) and the Menlo Report (2011)):
  a. Respect for Persons
  b. Beneficence
  c. Justice
  d. Respect for Law and Public Interest

Respect for Persons

1. Individuals should be treated as autonomous
2. Individuals with diminished autonomy should be entitled to additional protections

Basically: Researchers should not do things to people without their consent - even if the researcher thinks that the thing is harmless, or even beneficial

Beneficence

1. Do no harm
2. Maximize possible benefits and minimize possible harms

In practice, researchers should perform a risk/benefit analysis and then a decision about whether the risks and benefits strike an appropriate ethical balance

Justice

- It should not be the case that one group in society bears the costs of research while another group reaps its benefits
- Researchers shouldn't be allowed to intentionally prey on the powerless
- Minorities should be included in clinical trials so they can benefit from the knowledge gained

Respect for Law and Public Interest

1. Compliance
2. Transparency-based accountability

- Researchers should attempt to identify and obey relevant laws, contracts, and terms of service
- Avoid doing things in secret

Example: Scraping Facebook data and merging with university records

- Beginning in 2006, each year, researchers scraped the Facebook profiles of the Class of 2009 at a "diverse private college in the Northeastern US"
- Data was merged with data from the college
- Rich data resource that helped us understand how social networks form and evolve
- Days after the data was made available, people deduced that the school in question was Harvard
- Students did not give consent for their data to be used
- The dataset was removed from the internet

Example: Emotional contagion among Facebook users

- For a week in January 2012, 700,000 Facebook users were placed in an "emotional contagion" experiment
- Studied the extent to which a person's emotions are influenced by the people they interact with
- Placed in "negativity-reduced," "positivity-reduced," and "control" groups
- Participants did not provide any consent beyond the standard Facebook terms of service
- The study didn't undergo meaningful third-party ethical review

Example: Measuring internet censorship with Encore

- Encore was launched in 2014: System to provide real-time and global measurements of Internet censorship
- Encouraged website owners to add a small code snippet to the source files of their web pages
- Anyone who visited a web page with this code snippet would try to contact a website the researchers were monitoring for possible censorship and report whether it could be accessed
- The user doesn't know this is happening
- Concern that people in certain countries could be exposed to risk if their computers tried to visit certain sensitive webpages
- The study was controversial due to the lack of informed consent of participants

★ Exercise 5: Which principles were violated in each example? How?

1. Scraping Facebook data and merging with Harvard records
2. Emotional contagion among Facebook users
3. Measuring censorship with Encore

Principles:

- Respect for Persons
- Beneficence
- Justice
- Respect for Law and Public Interest

Questions?

Break!

Module 3: Prediction: The art of forecasting

After thinking of a new data science application:
What's next?

- Separate our discussion of applications into three distinct kinds of tasks:
  - Prediction
  - Optimization
  - Generation

After thinking of a new data science application:
What's next?

- Separate our discussion of applications into three distinct kinds of tasks:
  - Prediction
  - Optimization
  - Generation

Prediction of output for new instances based on previous instances

Examples of prediction applications

- Price prediction: What's the optimal price for a good or service?
- Dosage prediction: What dosage of medicine should be prescribed for a given patient?
- Risk assessment: How much risk is predicted to be associated with a given decision?
- Propensity modeling: What actions are customers likely to take based on their past behavior?
- Diagnosis: What disease or condition explains a patient's symptoms the best?
- Document classification: Which category does a document fall in? "Document" can be expanded to include images, sounds, and videos

Questions to answer when brainstorming an application:

1. What output will be predicted?
2. What input will be used to make the prediction?
3. If our model works well, how will it solve the problem?

★ Exercise 6: Car insurance fraud case study

- A car insurance company is losing too much money due to fraudulent claims, despite investigating 30% of claims
- Brainstorm potential predictive modeling solutions to address this problem
  - What would you be trying to predict?
  - What would be the input data?
  - How would this help address the problem?

piazza

Adapted from "Fundamentals of Machine Learning for Predictive Data Analytics" Kelleher, Mac Namee, D'Arcy

Approach 1: Claim prediction

- Build a model to predict the likelihood that a claim is fraudulent
- Claims that are assigned the highest likelihood of fraud could be investigated
- This would help the investigation team focus on the claims most likely to be fraudulent, increasing the number of frauds detected and reducing the amount of money lost to fraud
- Data requirements: A large collection of historical data labeled as "fraudulent" and "non-fraudulent"
- Capacity requirements: A mechanism put into place to inform claims investigators that some claims should be prioritized above others

Approach 2: Member prediction

- Build a model to predict the likelihood that a member will commit fraud in the near future
- The model could be run every quarter, predicting which members were fraud risks
- The insurance company could reach out to these members with a warning about canceling their policies, reducing the amount of fraud they commit
- Data requirements: A large collection of historical claims labeled "fraudulent" and "non-fraudulent" linked to specific members
- Capacity requirements: Requires that the company can contact customers in a way that disincentivizes fraud, but doesn't lose them as a customer
- Could also be ethical implications

Approach 3: Application prediction

- Build a model to predict the likelihood that a new applicant will make a fraudulent claim at some point
- The model could be run every time a new person applied, and applications predicted to result in a fraudulent claim could be rejected
- Data requirements: A large collection of historical claims labeled "fraudulent" and "non-fraudulent" linked to specific applicant details
- Capacity requirements: This model would need to be integrated into the current application approval process

Approach 4: Payment prediction

- Build a model to predict the amount most likely to be paid out by an insurance company after investigating the claim
- This model could be run whenever a new claim is made, and the policy holder could be offered the predicted amount as an alternative to going through the claims process
- Data requirements: Data about all claims, the amount originally requested by each claim, and the amount paid out for each claim
- Capacity requirements: The company would need to be able to run this model whenever a claim was made, and be able to contact customers to give them the offer

What do we do after defining our input and output data:

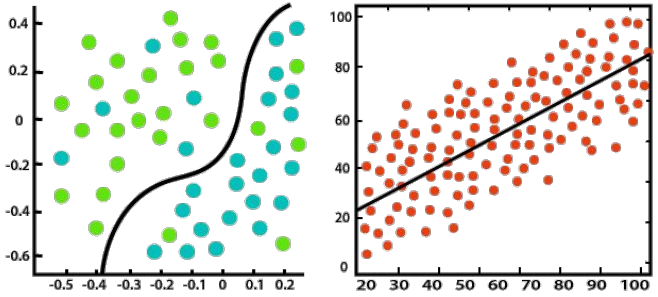
Choosing a model

Choosing which model to use: No free lunch

Wolpert and Macready (2005): Any two optimization algorithms are equivalent when their performance is averaged across all possible problems

- The correct solution will depend on the problem
- It is often best to try multiple algorithmic solutions
- When choosing a model, it is important to consider:
  - Prediction speed
  - Capacity for re-training
  - Interpretability

Regression vs Classification Problem



Classification Regression

- Regression:
  - The goal is to predict a continuous number
  - The outcome can be any value in a given range
  - Example: Predicting income from education, age, and where they live
- Classification:
  - The output we're trying to predict is one of several options
  - The possible outputs are called classes
  - For example: Predicting the species of a flower based on observed characteristics
- To determine whether it's a classification or regression task, ask: Is there continuity in the output?

Overfitting vs Underfitting

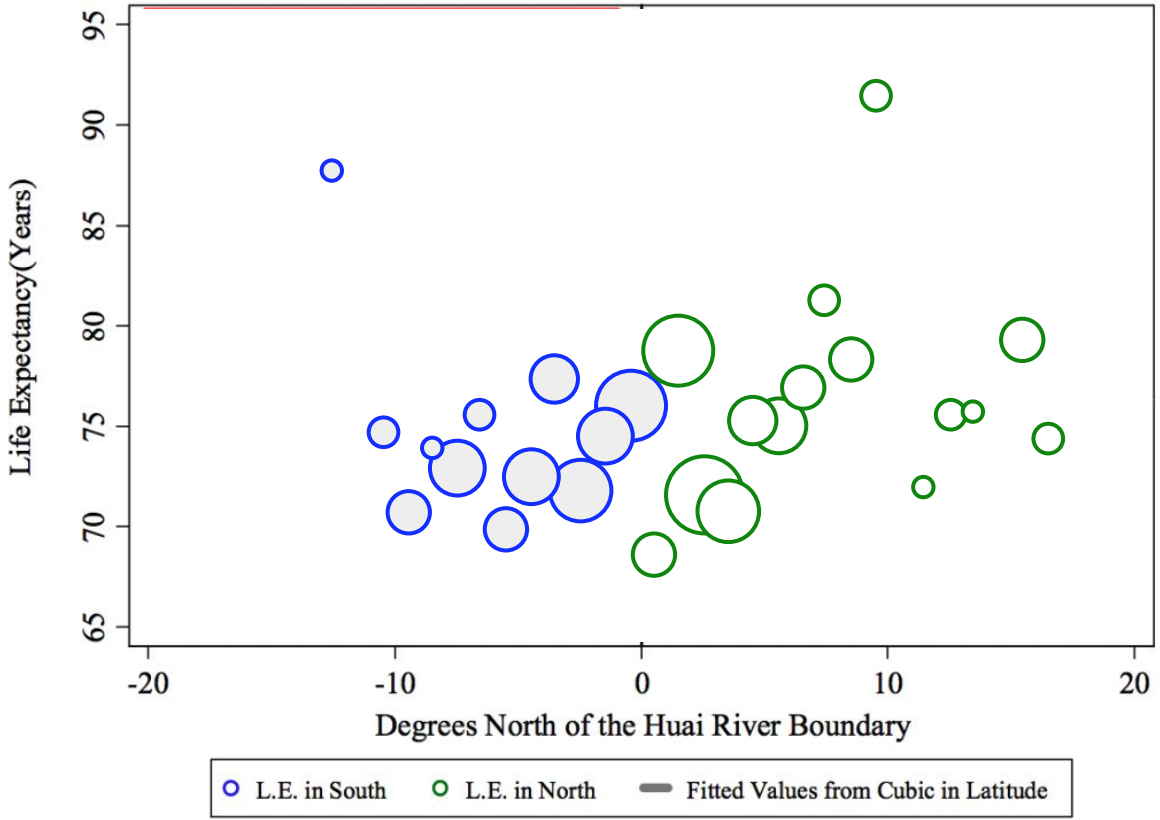
Overfitting: when you fit a model too closely to the training set and so that it works well on the training set but doesn't generalize to new data; too complicated

Underfitting: when you fit a model that doesn't capture all aspects of variability in the data, and does badly even on the training set; too simple

Overfitting in the real world

Plot from a study of life expectancy based on the distance from China's Huai River

Based on this plot, would you say the life expectancy was higher in the north or the south?

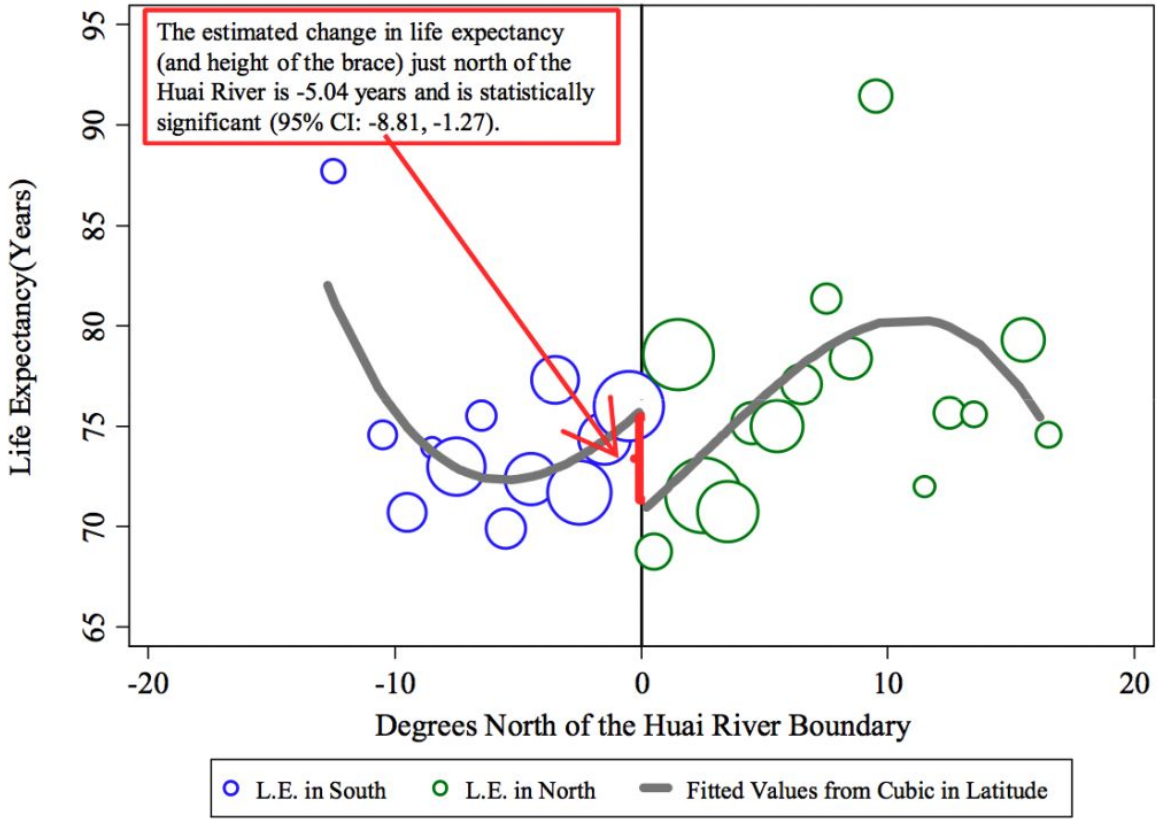


https://statmodeling.stat.columbia.edu/2013/08/05/evidence-on-the-impact/

Overfitting in the real world

The authors fit separate cubic functions to each side of the data

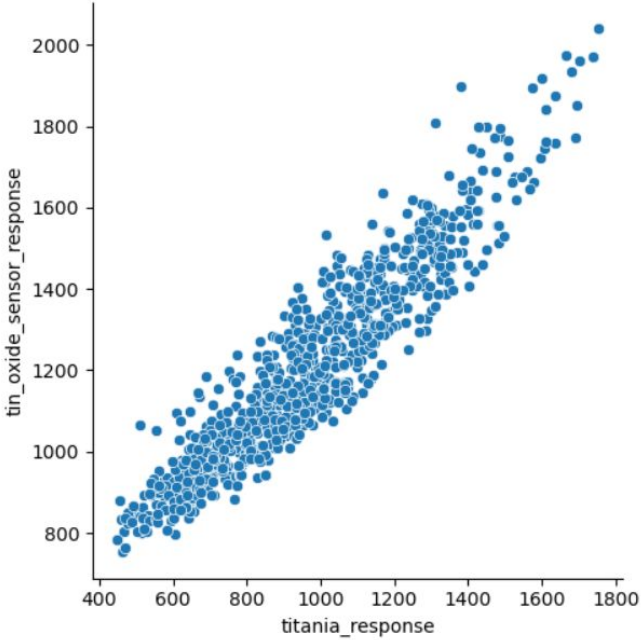
Based on these fitted functions, they claimed that life expectancies in the North were about 5.5 years lower

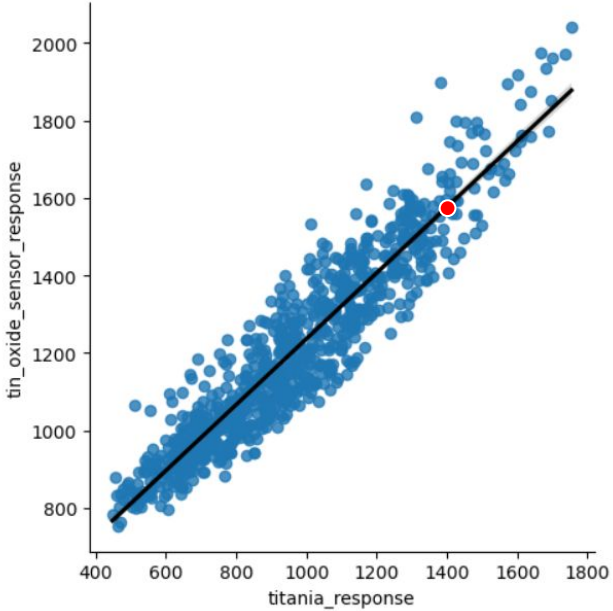


https://statmodeling.stat.columbia.edu/2013/08/05/evidence-on-the-impact/

Most straightforward regression algorithm: Linear regression

- After fitting our model, we can use the regression line to predict new values



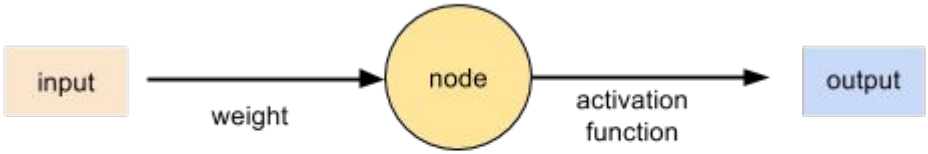


$y = 0.85x + 386$

$x = 1400$

$y = 0.85 \times 1400 + 386$
$= 1{,}576$

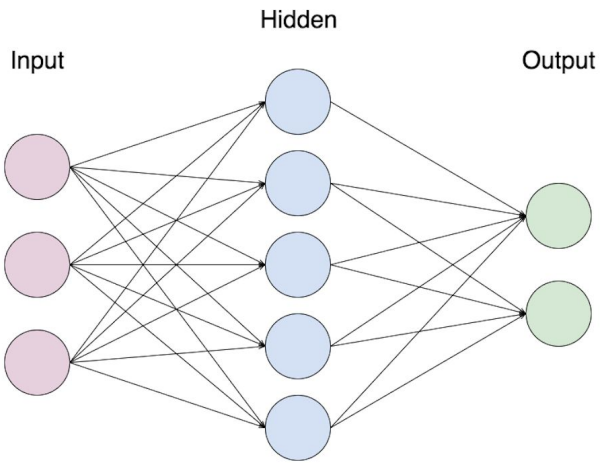
Adding more complexity: Neural networks



output = activation_function (input x weight)

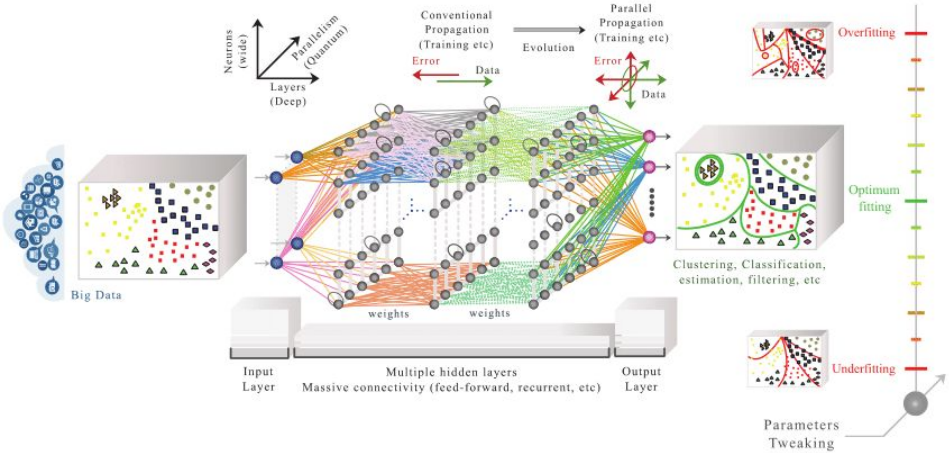
- A node applies a weight to each input, feeds the inputs into an activation function, and passes on the output
- During training the weights are modified until they provide the most accurate predictions on the training data

Adding more complexity: Neural networks

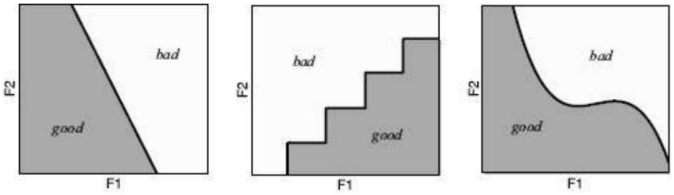


- A node applies a weight to each input, feeds the inputs into an activation function, and passes on the output
- During training the weights are modified until they provide the most accurate predictions on the training data

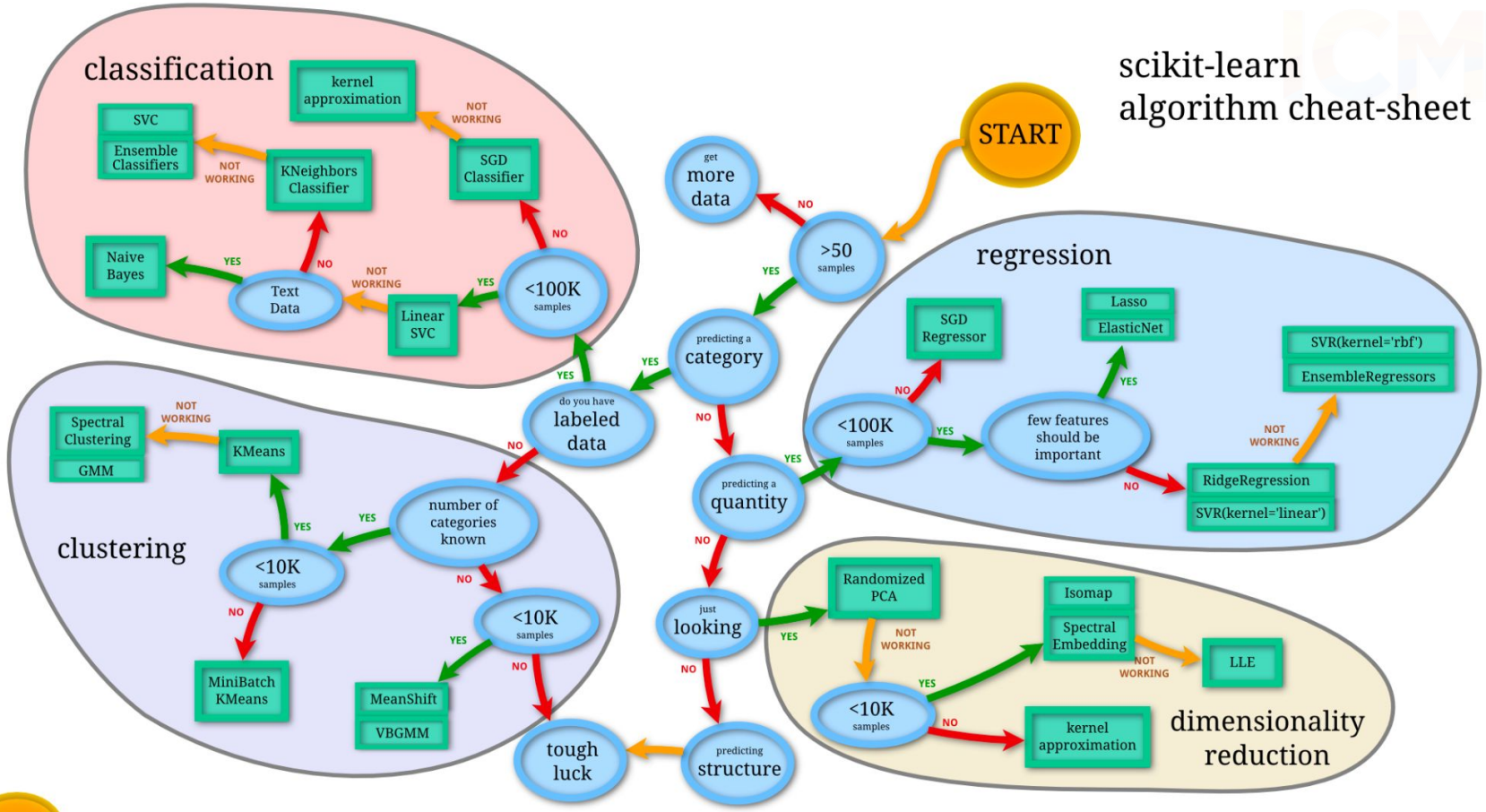
Adding more complexity: Neural networks



- A node applies a weight to each input, feeds the inputs into an activation function, and passes on the output
- During training the weights are modified until they provide the most accurate predictions on the training data


bad
good
F2
F1
bad
good
F2
F1
bad
good
F2
F1

scikit-learn algorithm cheat-sheet



Back

scikit learn

★ Exercise 7: If you have built predictive models in the past: How did you decide which complexity of algorithm to use?

If you haven't, how would you decide?

Application example: Disease susceptibility prediction

Problem: Many people experience diseases or illnesses later in life that could be protected against or mitigated if they were known earlier

Is this problem suitable to approach with data science?

Is disease susceptibility prediction suitable for a data science solution?

- Quantifiability: We can measure whether a person has a particular disease
- Data availability: Companies like 23andMe collect genetic data, as well as information on diseases of participants
- Data quality: Data is collected in-house, so quality can be assured (and uncertainty can be tested and corrected for)
- Predictability: It is known that some people are genetically predisposed to different diseases (though the effect of genetics varies based on disease)
- Understanding the impact: This could result in people living longer lives, because they are able to protect themselves more from diseases they are susceptible to
- Could this have a negative effect? Increase medical insurance costs for some?

How to approach the problem of disease susceptibility prediction?

- Input: Genetic data from many individuals
- Output: likelihood that a given participant will acquire a specific disease over their lifetime
- How this will help solve the problem: Disease risk will be delivered to participants with their ancestry results, along with suggested actions to mitigate risk from any diseases found

Share Print

Bitter Taste

A bitter warning

Scientists believe the bitter taste sense developed to help animals detect toxins or poisons in food. But not everyone can taste the same things.

Your Traits Result

Jamie, your genetics make you **unlikely to detect certain bitter tastes.**

Of 23andMe research participants with genetics like yours:

66% are unlikely to taste.

34% are likely to taste.

Health > Health Predisposition

Print

Celiac Disease

Celiac disease is an autoimmune condition in which the consumption of gluten (found in wheat, barley, and rye) can result in damage to the small intestine. Celiac disease can lead to both digestive and non-digestive problems. This test includes two common variants associated with an increased risk of developing this condition.

Overview | Scientific Details | Frequently Asked Questions

Jamie, you **do not have** the two genetic variants we tested.

You are not likely at risk of developing celiac disease based on your genetic result.

Application example: Product recommendation

Problem: How can consumers be encouraged to buy more products when on the site?

Is this problem suitable for a data science solution?

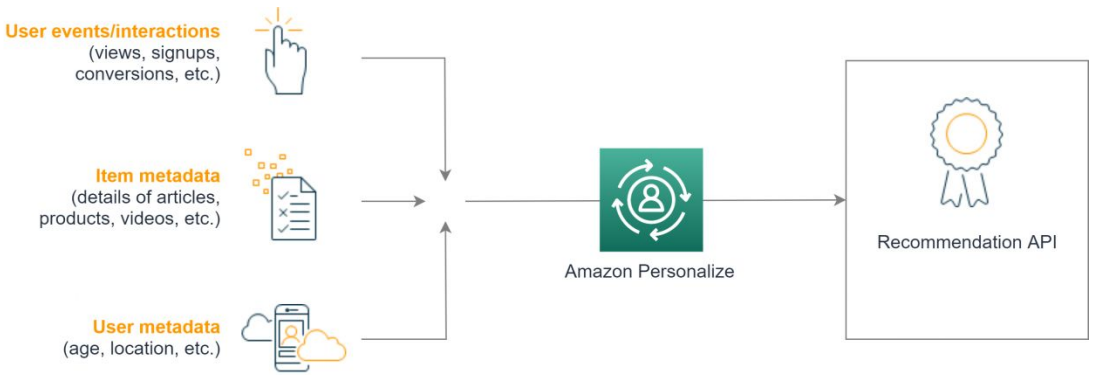
Is product recommendation suitable for a data science solution?

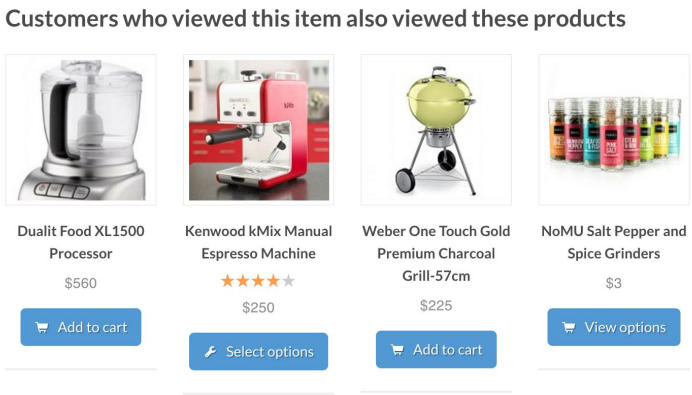
- Quantifiability: Whether or not someone buys a product is measurable
- Data availability: Amazon has lots of data on which products were purchased by which customers
- Data quality: Amazon is able to control what data is saved to avoid missing values
- Predictability: There is often a logical link between a cluster of purchases
- Understanding the impact: Could help consumers find new products more efficiently
- Could also influence customers to buy more than they were planning on

How to approach the problem of product recommendation?

- Input: ordered list of recent purchases for many customers
- Output: The next product a customer will buy
- How this will help solve the problem: products that are likely to be of interest to a customer can be recommended. This is expected to lead to more purchases being made

How to approach the problem of product recommendation?





Review

Framework for determining application suitability:

1. Quantifiability
2. Data availability
3. Data quality
4. Predictability
5. Understanding the impact

Framework for deciding how to approach the problem:

1. What output will be predicted?
2. What input will be used to make the prediction?
3. If our model works well, how will it solve the problem?

Questions?

★ Exercise 8 (Homework): Brainstorm ideas of applications that could be addressed with data science

Ideas could come from:

- Work
- Personal life
- Outside inspiration

Consider whether the application is suitable for data science based on the frameworks we discussed.

What would you consider predicting? How would this model "solve" the problem?