W4995 Applied Machine Learning
Fall 2021

Lecture 1

Dr. Vijay Pappu

A little about me...

B.Tech

Ph.D.

Data Scientist

Engineering Manager

Logistics

- Course website:
  - https://courseworks2.columbia.edu/courses/141188
  - Assignments schedule is also posted (in Syllabus section)
  - Please post any questions in the discussions section of the webpage

- Course grading:
  - 5 programming assignments - **50%**
  - 1 in-class midterm - **20%**
  - 1 project - **30%**

- Class attendance is optional

Logistics

- My information
  - **Email:** vp2472@columbia.edu
  - **Office hours:** Wednesdays (4-6PM @DSI Mudd conference room)

- Course Assistants (CAs) & Office hours (all times in ET):
  - Nandini Agarwal - Mon 4-6 PM
  - Keshaw Singh - Tue 12-2 PM
  - Shouvik Mani - Tue 4-6 PM
  - Vaibhav Bagri - Thu 5-7 PM
  - Yue Kuang - Thu 3-5 PM
  - Jaidev Shah - Fri 7-9 PM

Course materials

- The slides will be made available on course website post the in-class lecture
- The classes are recorded (everything)
  - I might repeat your questions so that the mic can pick up.
- The class recordings will be available to students attending in-person as well in the “Video Library” section of the course webpage.

Assignments

- Programming assignments should be submitted in Python
- We will use Github classroom for assignment submissions
- 5 assignments in total (3 before midterm)
- Each assignment contributes 10% towards your final grade
- Late submissions are not allowed and will result in no points

Project

- You can work in teams of 4-6 (ideal)
  - This will result in ~20-25 teams
- Project deliverables include functional code, report and a final presentation (to be scheduled in the study week)
- Project description and key milestones will be shared soon on courseworks

Plagiarism

The use of words, phrases, or ideas that do not belong to the student, without properly citing or acknowledging the source, is prohibited. This may include, but is not limited to, copying ***computer code*** *for the purposes of completing assignments for submission.*

Columbia University Plagiarism policy

- CAs check the homeworks for plagiarism
- Copied code could result in no points for all involved
- Leveraging code snippets from other sources (Stack overflow, open source libraries) is allowed.
    - It is important to mention the source if it is substantial amount of code.

Useful resources

- The course does not have one recommended book, but would leverage material from the following resources:
  - Introduction to Machine Learning with Python
  - Learning from Data
  - Applied Predictive Modeling
  - Deep Learning

Before we begin...

This course is **not** about...

Before we begin...

This course is **not** about...

Theoretical underpinnings of Machine Learning

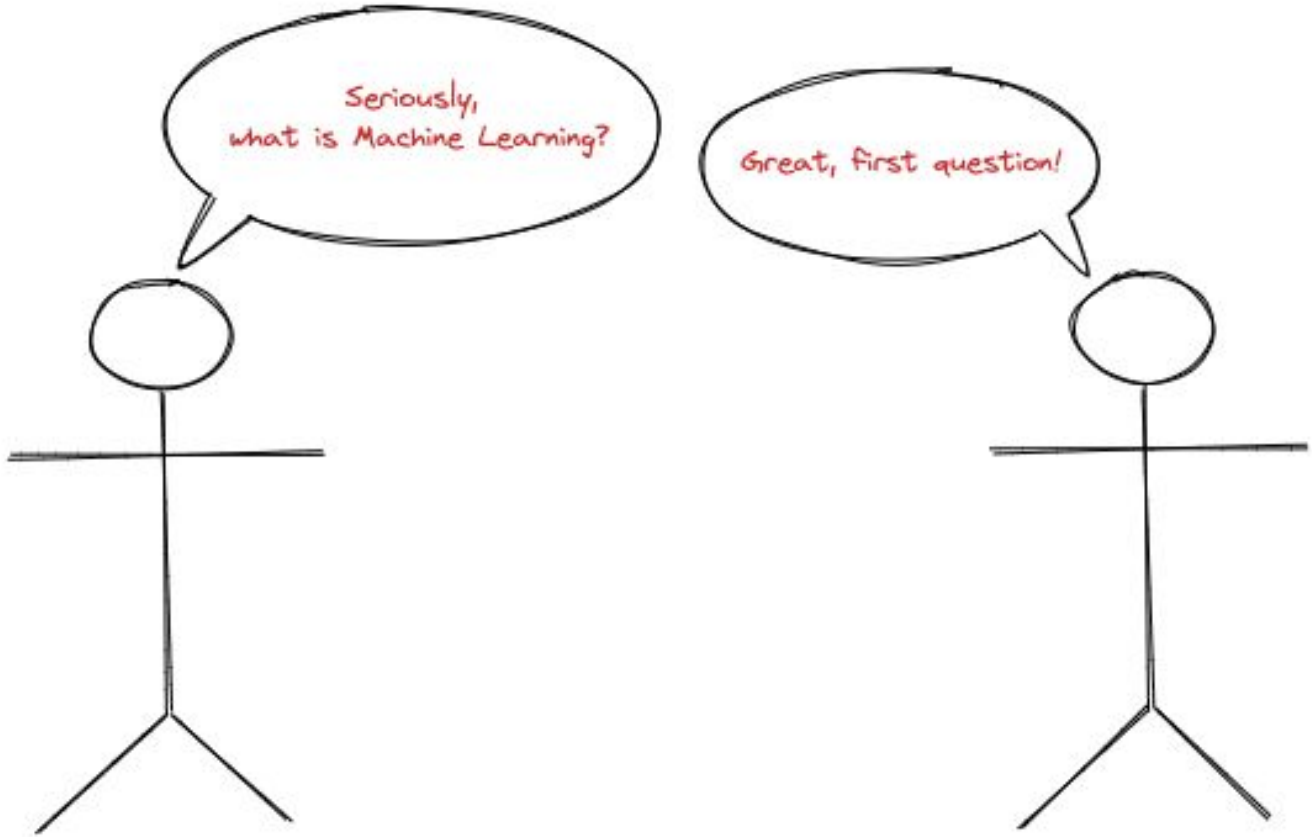
Before we begin...

However, this course is about...

Applying Machine Learning to real-world applications

In today’s lecture, we will cover...

- Introduction
- Exploratory Data Analysis & Visualizations


Seriously,
what is Machine Learning?
Great, first question!

Machine learning (**ML**) is the study of computer algorithms that can improve automatically through experience and by the use of data.

https://en.wikipedia.org/wiki/Machine_learning

I like this definition better...

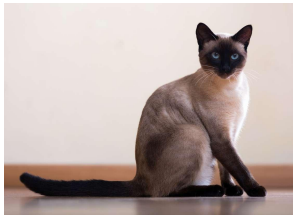
Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so.

https://en.wikipedia.org/wiki/Machine_learning

Heuristics v.s. ML system

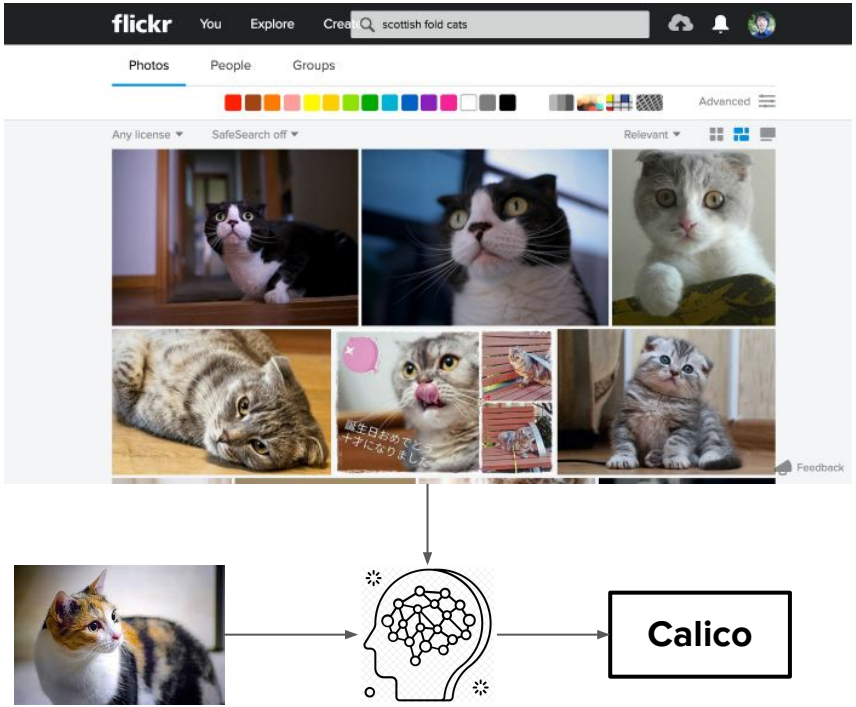
Calico - orange, black and/or grey markings over a white coat.

Tortoiseshell - white, orange and/or beige over a black coat.

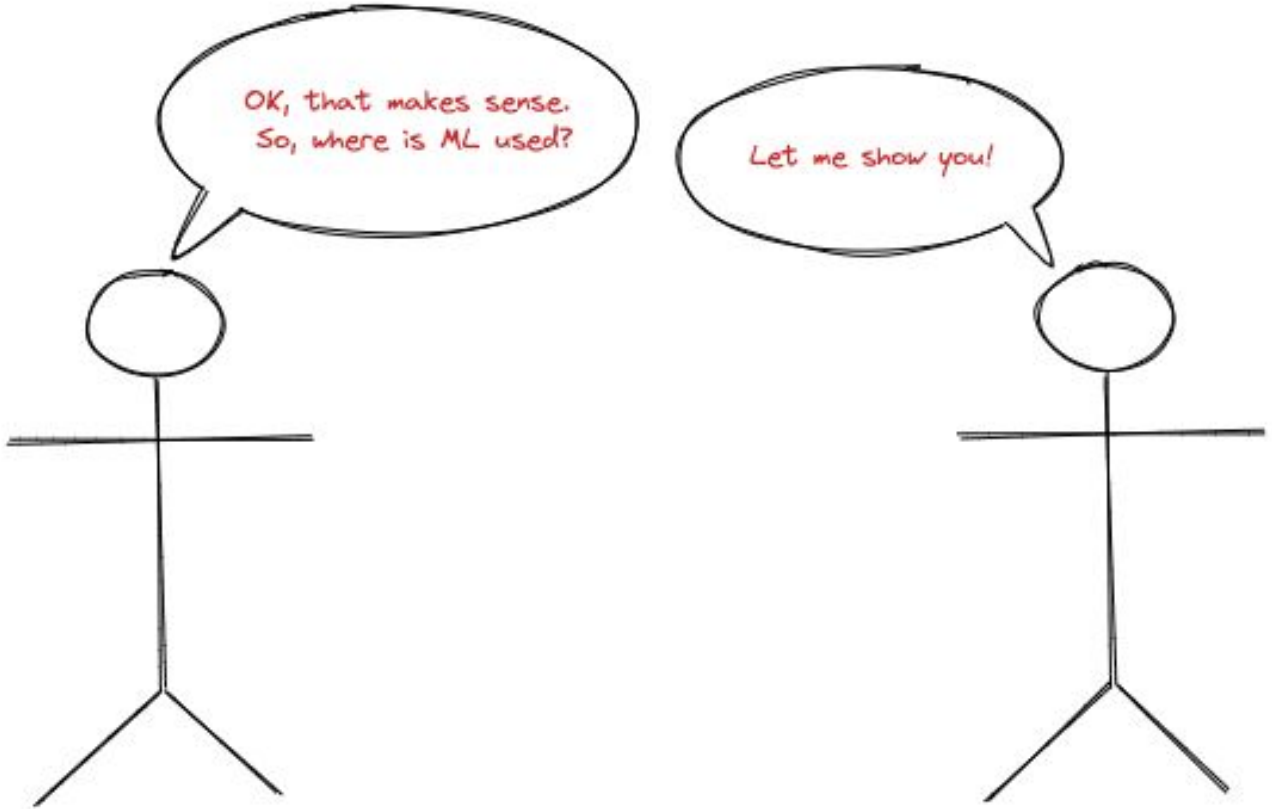


Siamese - angular-looking with black and tan coloring.

Heuristics



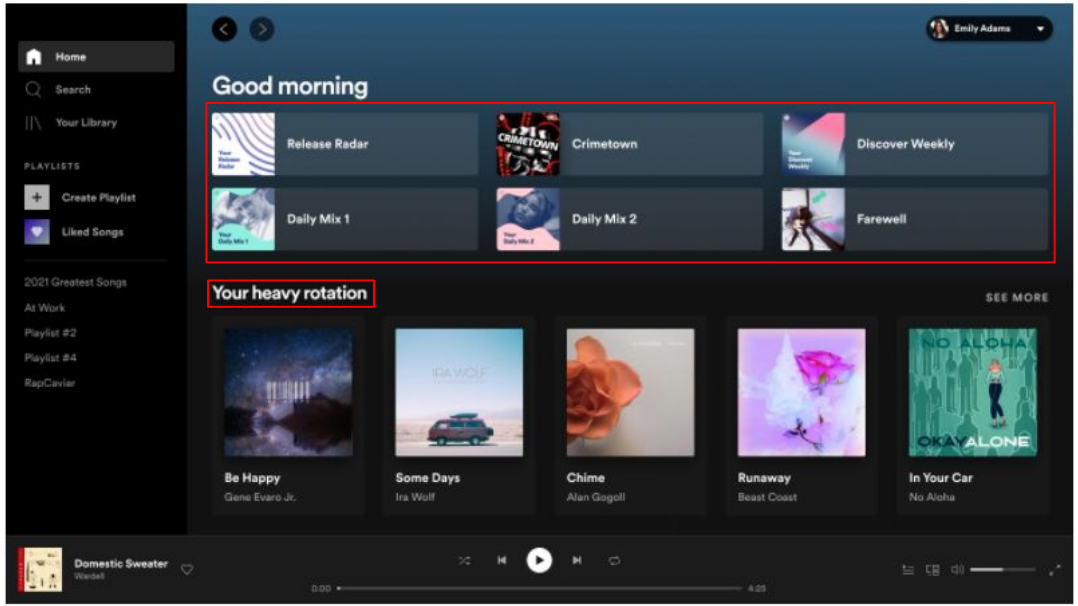
ML system

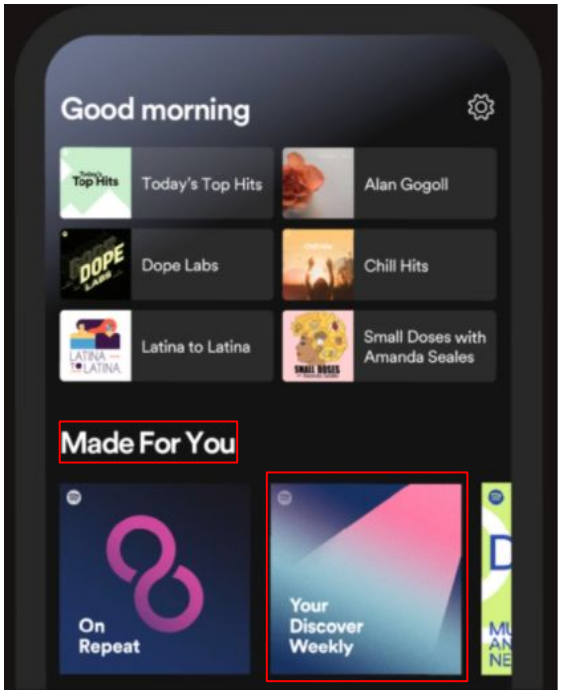

OK, that makes sense.
So, where is ML used?
Let me show you!

The short answer is...

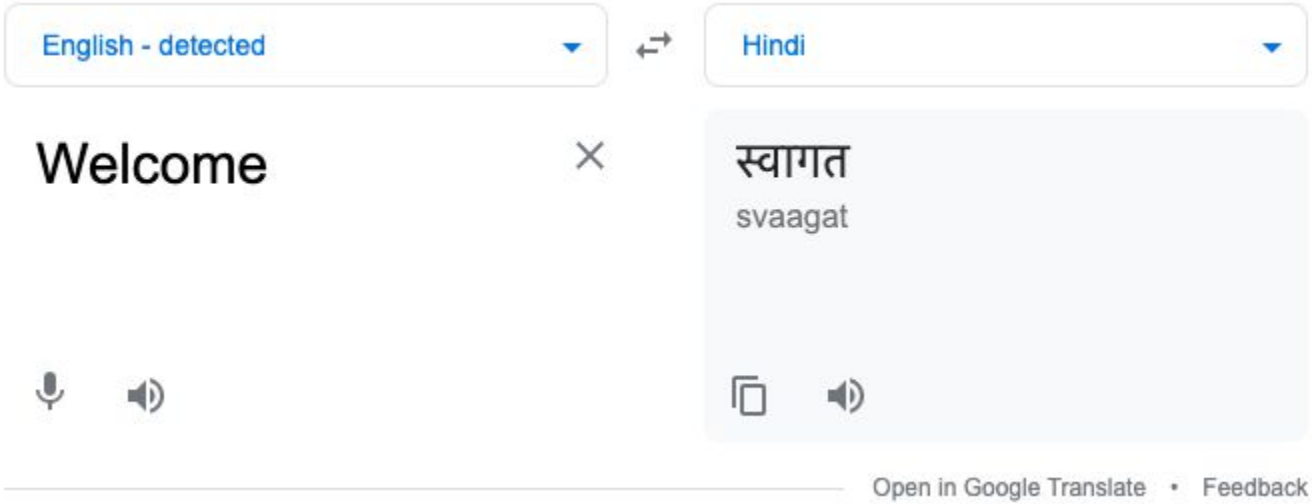
EVERYWHERE

(Obvious) Examples of Machine Learning

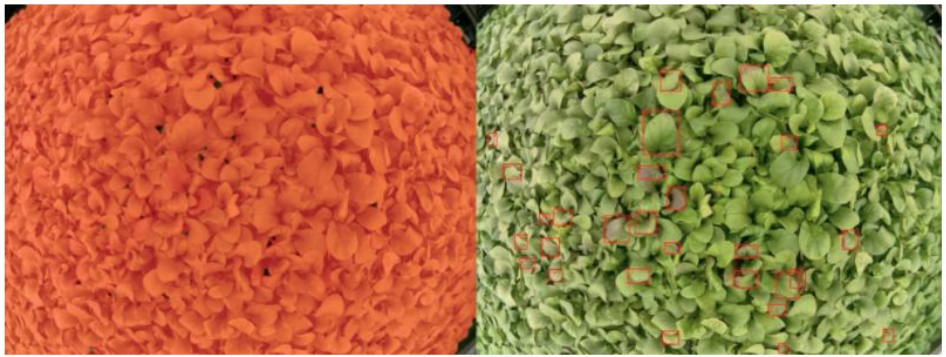




(Obvious) Examples of Machine Learning



(Not so obvious) Examples of Machine Learning



Computer vision identifies crop health issues on a tray of arugula

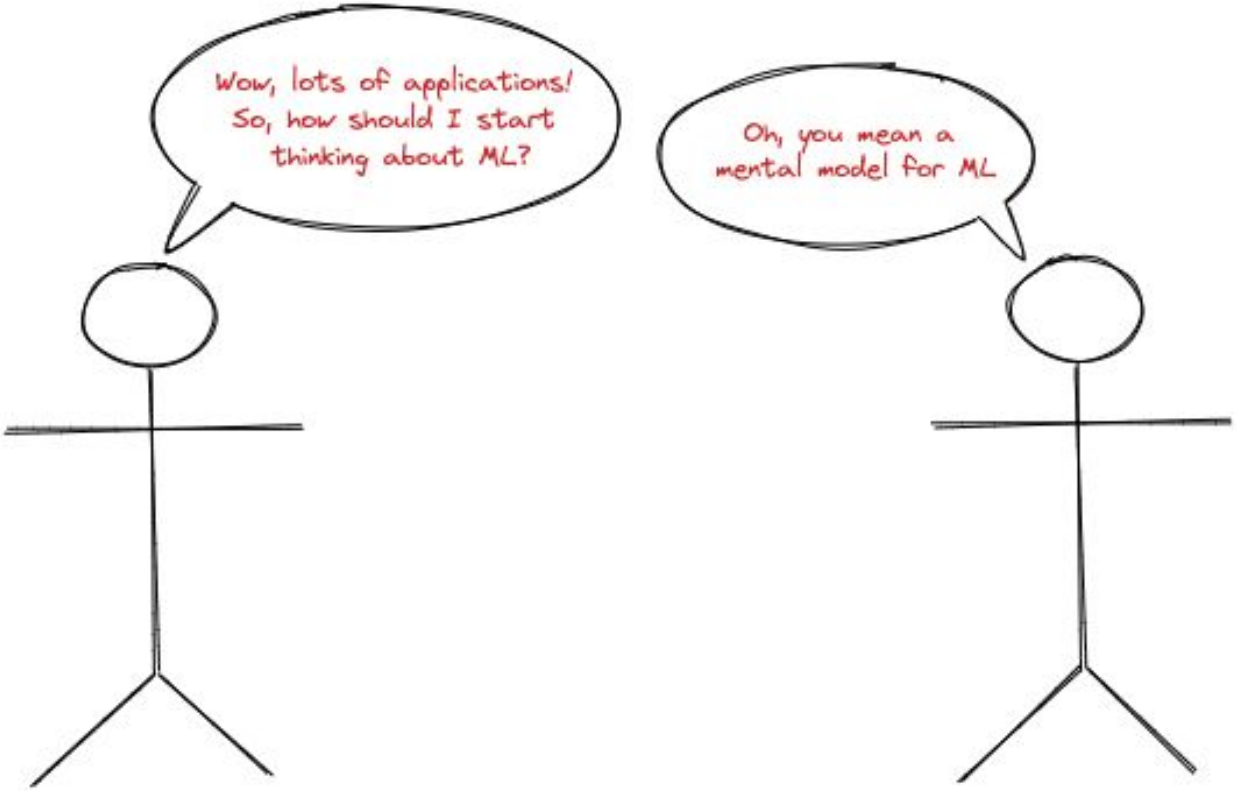
Example of the various recipes being tested for one crop over time.

https://boweryfarming.com/artificial-intelligence/

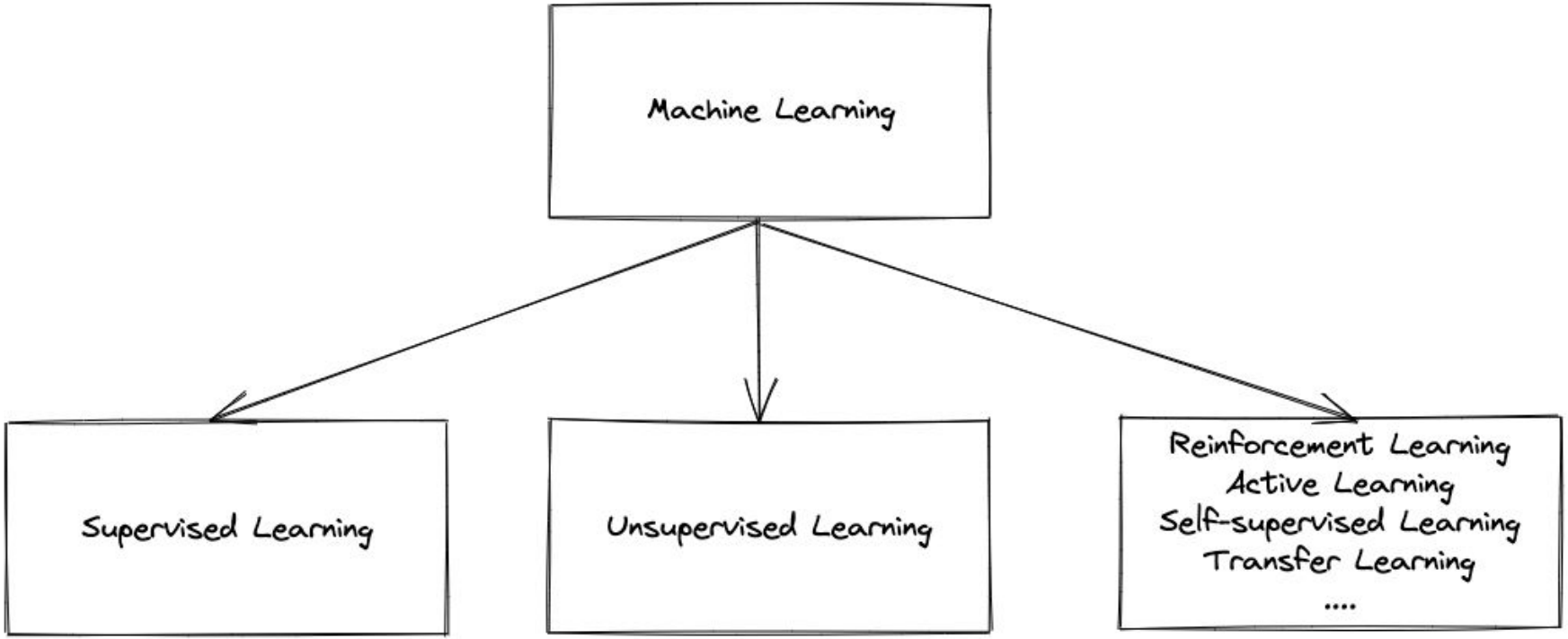
(Not so obvious) Examples of Machine Learning

A protein’s function is determined by its 3D shape.

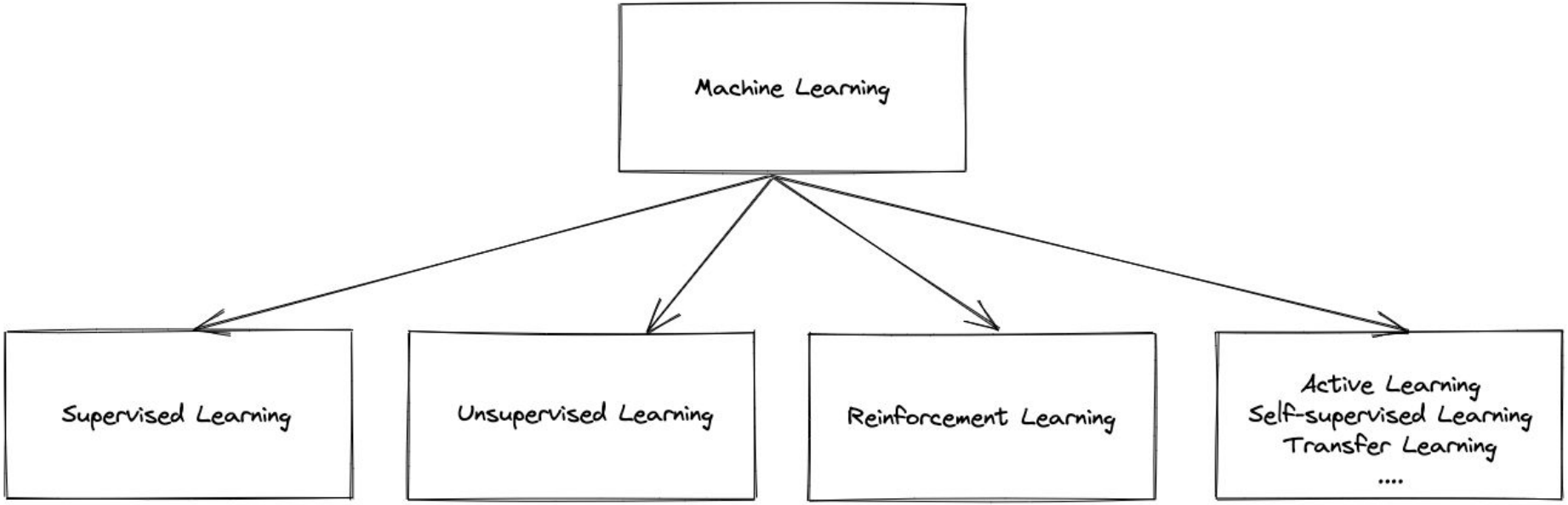
https://www.nature.com/articles/d41586-020-03348-4


Wow, lots of applications! So, how should I start thinking about ML?
Oh, you mean a mental model for ML

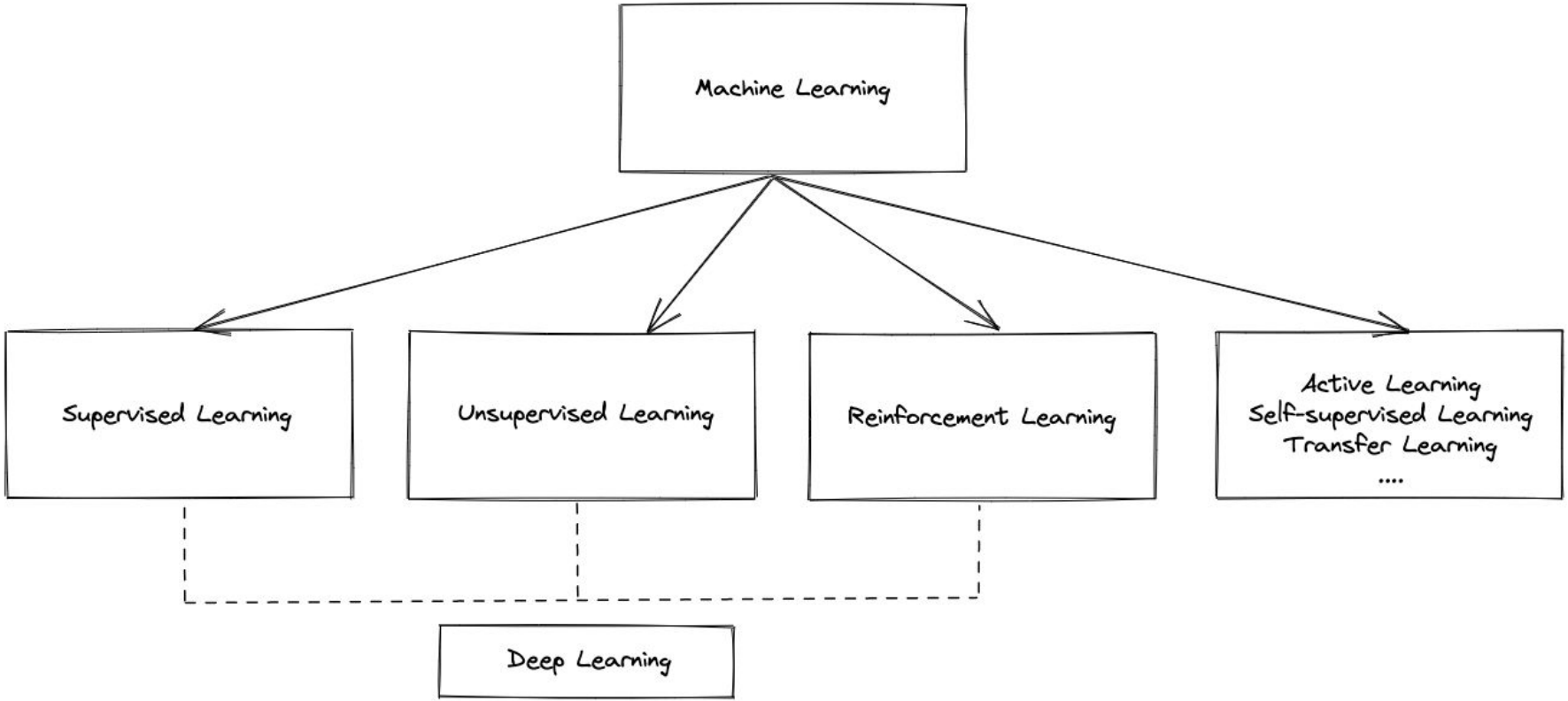
Until 1990's...


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Active Learning
Self-supervised Learning
Transfer Learning
....

Recently...

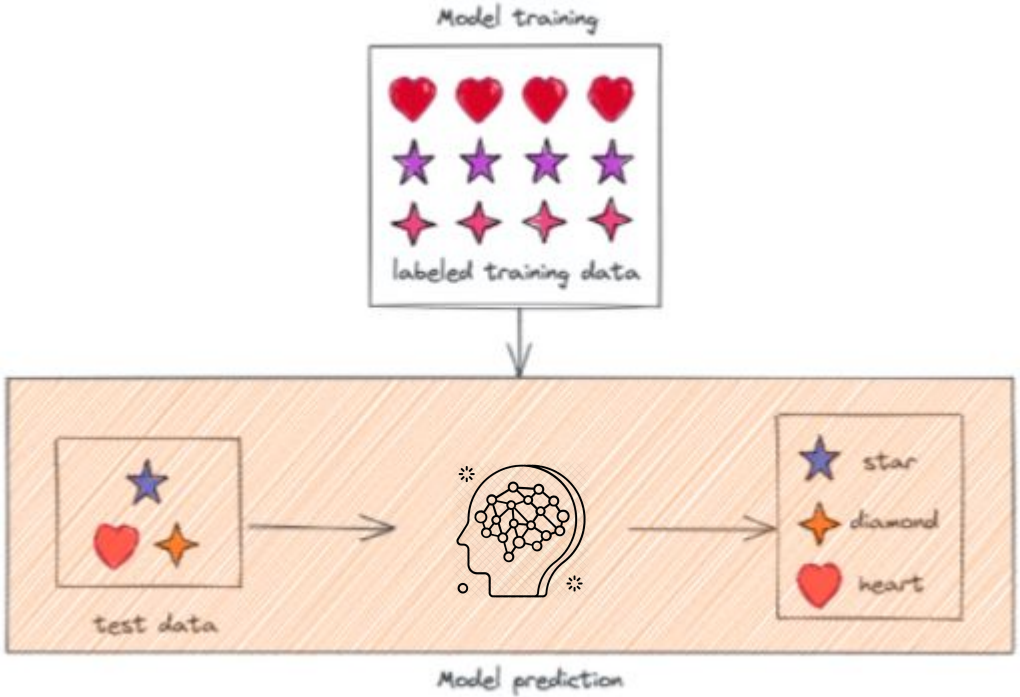

Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Active Learning
Self-supervised Learning
Transfer Learning
....

One of the reasons...



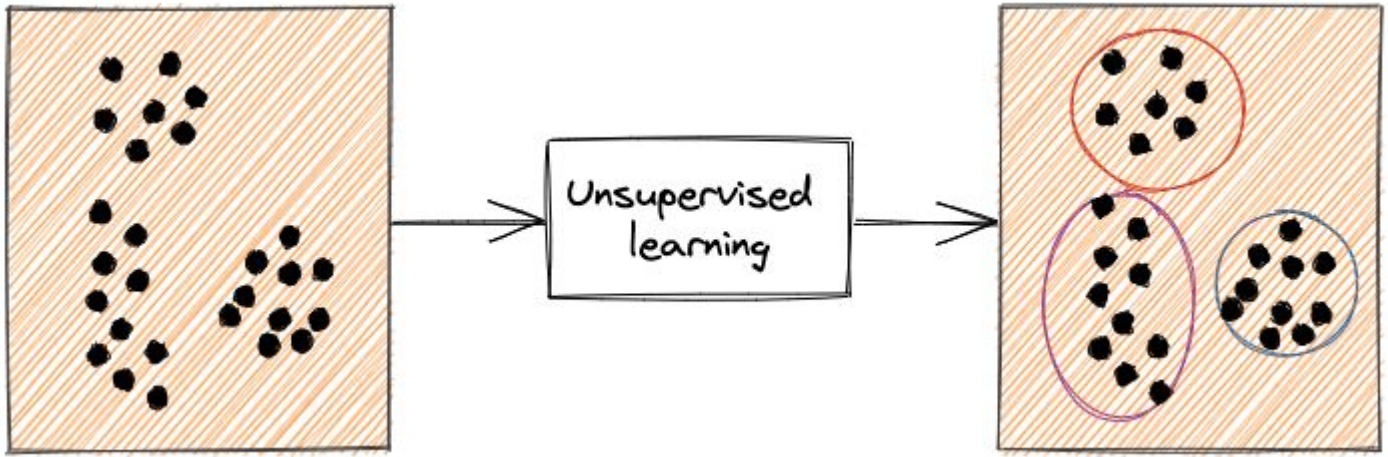
Supervised Learning

- Supervised learning algorithms learn a function that maps inputs to an output from a set of **labeled** training data.

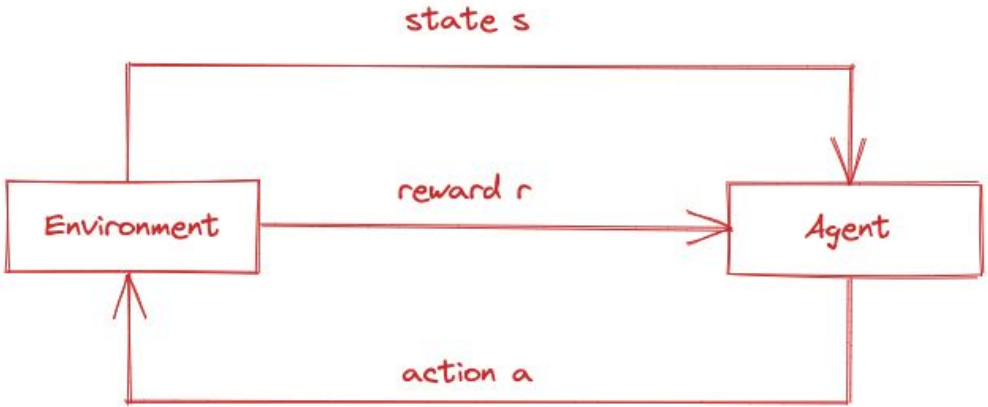


Unsupervised Learning

- Unsupervised learning algorithms learn patterns from **unlabeled** data samples.

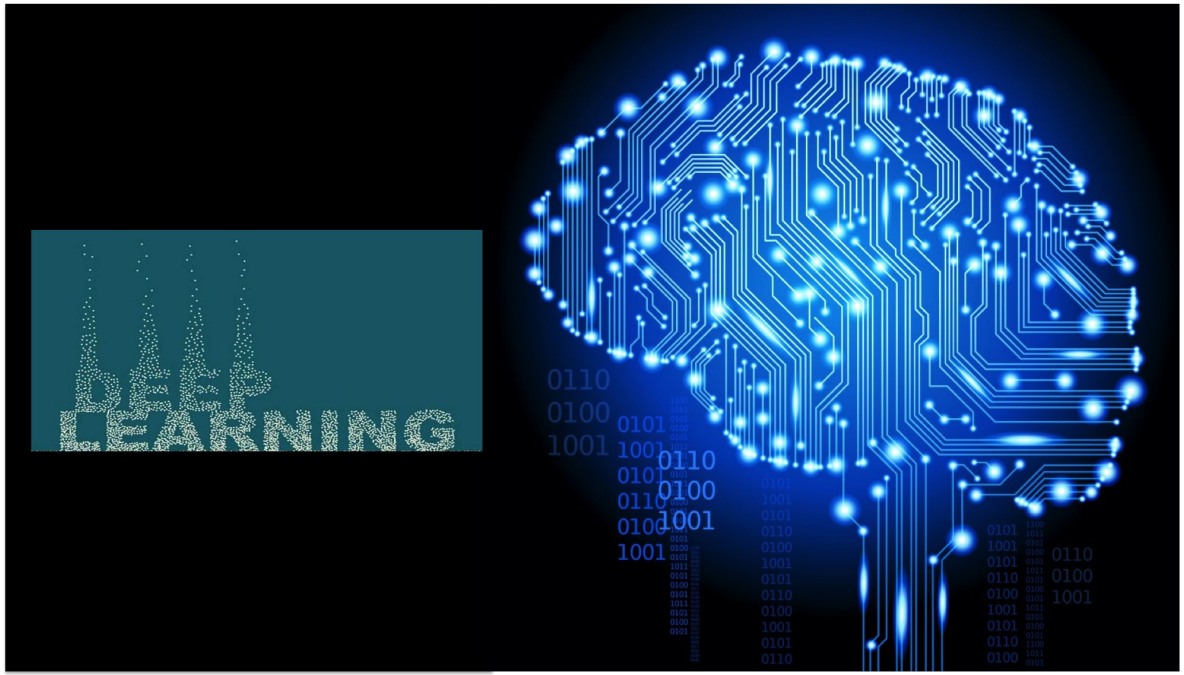


Reinforcement Learning



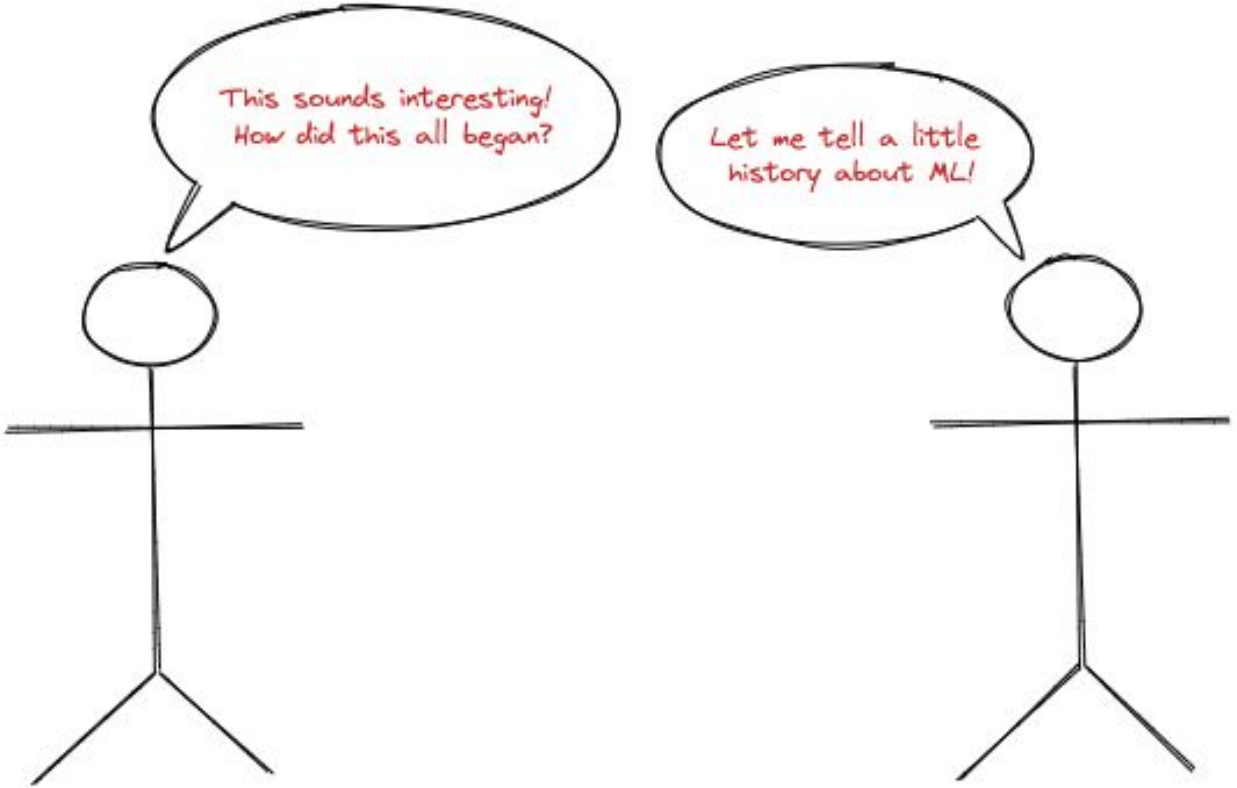
Deep Learning

- Deep learning is a class of ML algorithms that uses multiple layers to progressively extract higher-level features/abstractions from raw inputs.

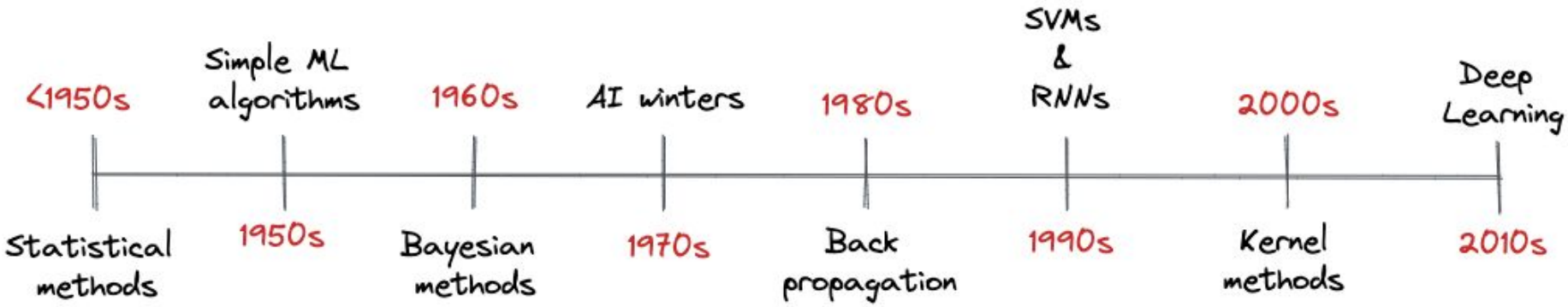


What about others?

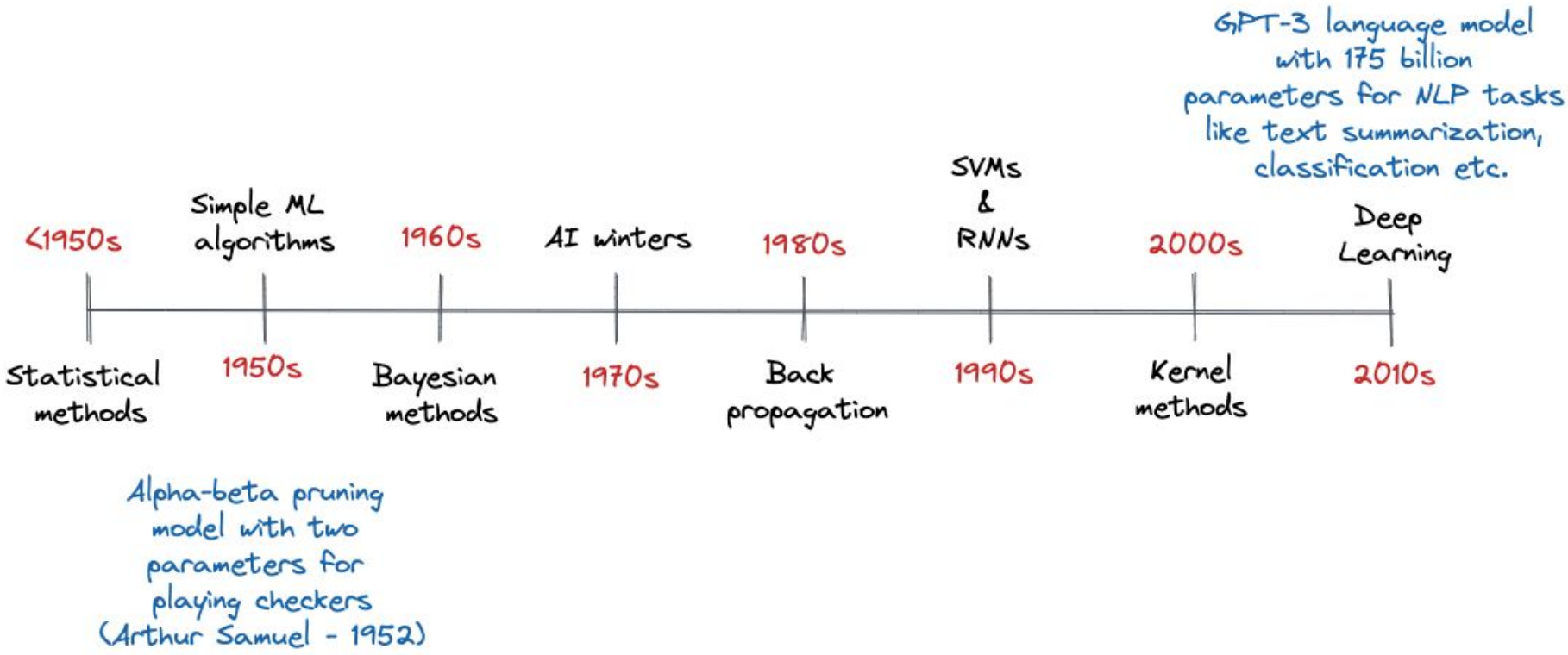
- Active learning
- Self-supervised learning
- Transfer learning

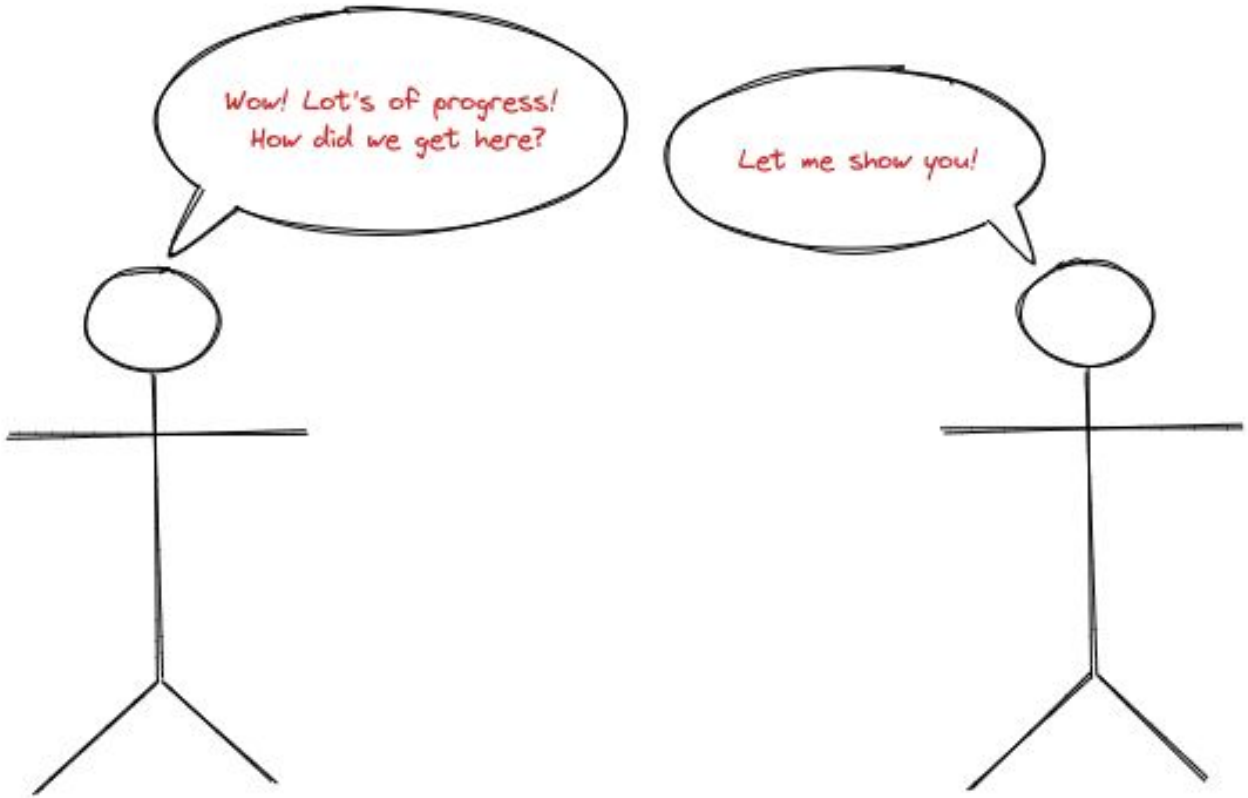

This sounds interesting! How did this all began?
Let me tell a little history about ML!

Timeline

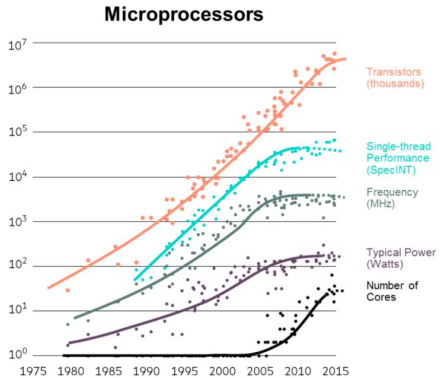

<1950s
Statistical methods
Simple ML algorithms
1950s
1960s
Bayesian methods
AI winters
1970s
1980s
Back propagation
SVMs & RNNs
1990s
2000s
Kernel methods
Deep Learning
2010s

Model complexity




Wow! Lot's of progress!
How did we get here?
Let me show you!

Three reasons...



Computational power[1]

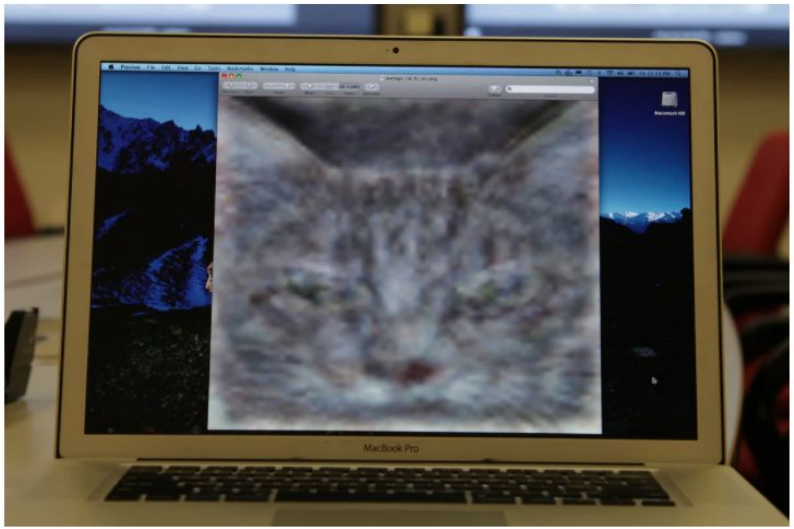
Big data

Breakthrough in Deep Learning

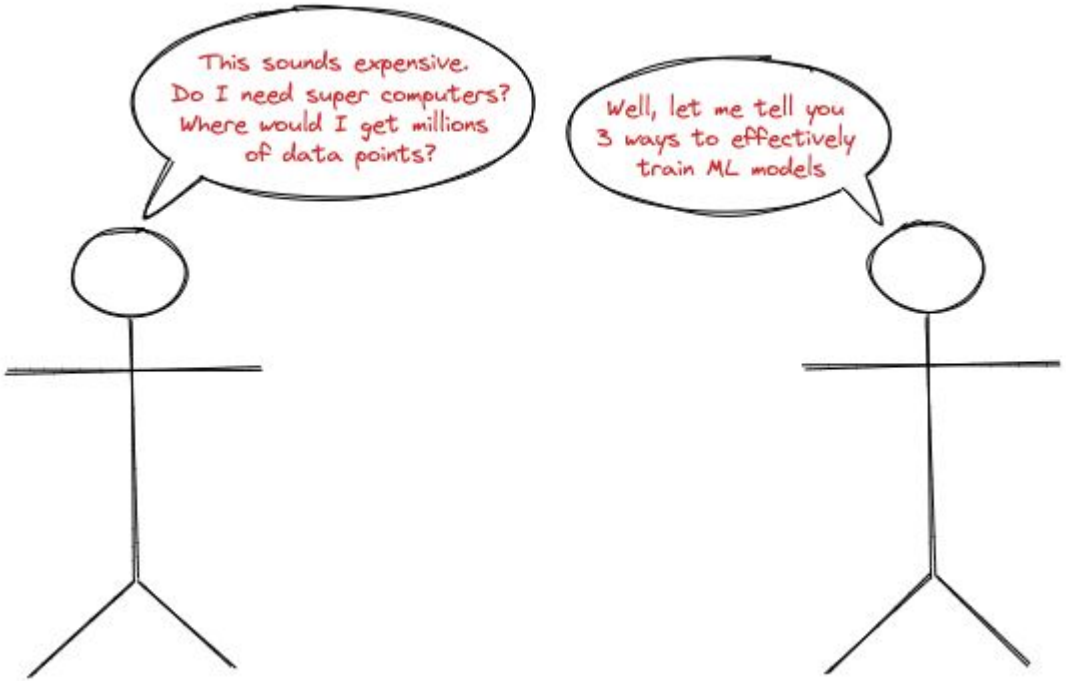
[1] - https://www.offgridweb.com/preparation/infographic-the-growth-of-computer-processing-power/

One example…

How Many Computers to Identify a Cat? 16,000



An image of a cat that a neural network taught itself to recognize. Jim Wilson/The New York Times


This sounds expensive. Do I need super computers? Where would I get millions of data points?
Well, let me tell you 3 ways to effectively train ML models

Computers have become powerful and accessible...

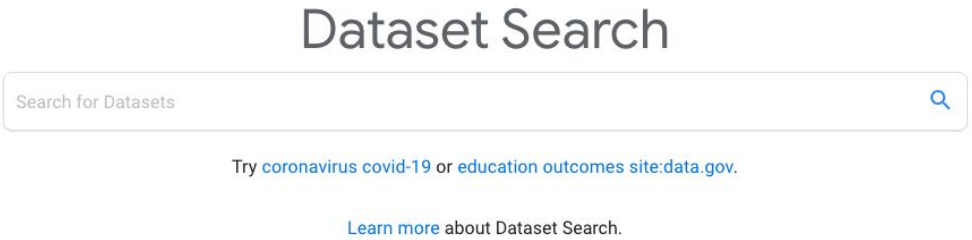
Introducing Amazon EC2 High Memory Instances with up to 12 TB of memory, Purpose-built to Run Large In-memory Databases, like SAP HANA

Posted On: Sep 27, 2018

Starting today, Amazon EC2 High Memory instances with up to 12 TB of memory are generally available.

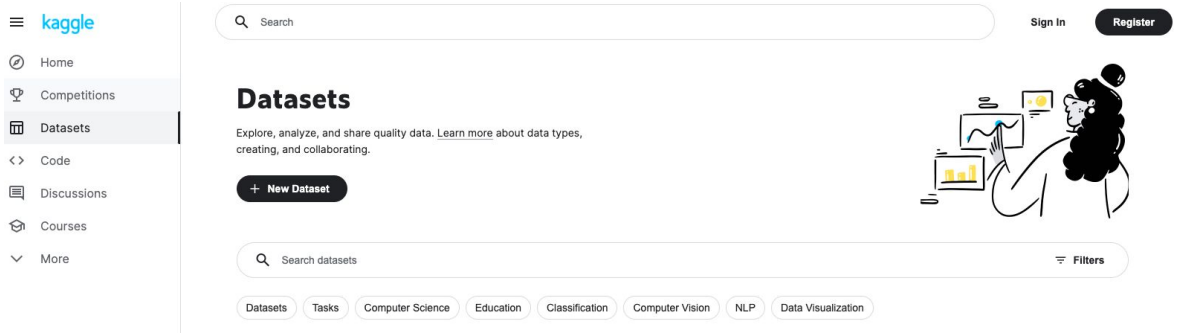
EC2 High Memory instances offer 6 TB, 9 TB, and 12 TB of memory in an instance. These instances are purpose-built to run large in-memory databases, including production deployments of the SAP HANA in-memory database, in the cloud. EC2 High Memory instances allow running large in-memory databases in the same Amazon Virtual Private Cloud (VPC), where connected business applications relying on the database are running, reducing the management overhead associated with complex networking and ensuring predictable performance.

Data is publicly available…

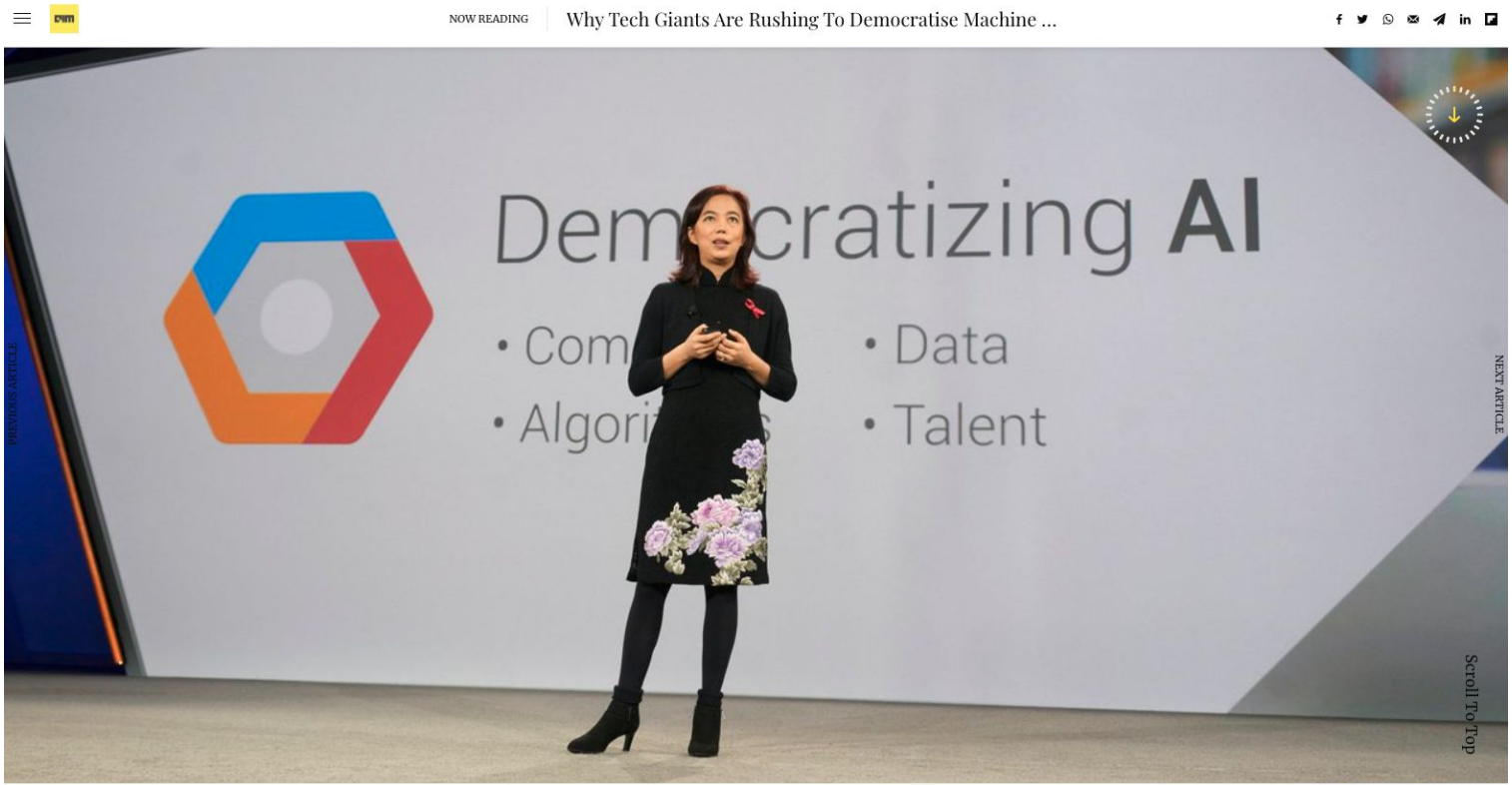


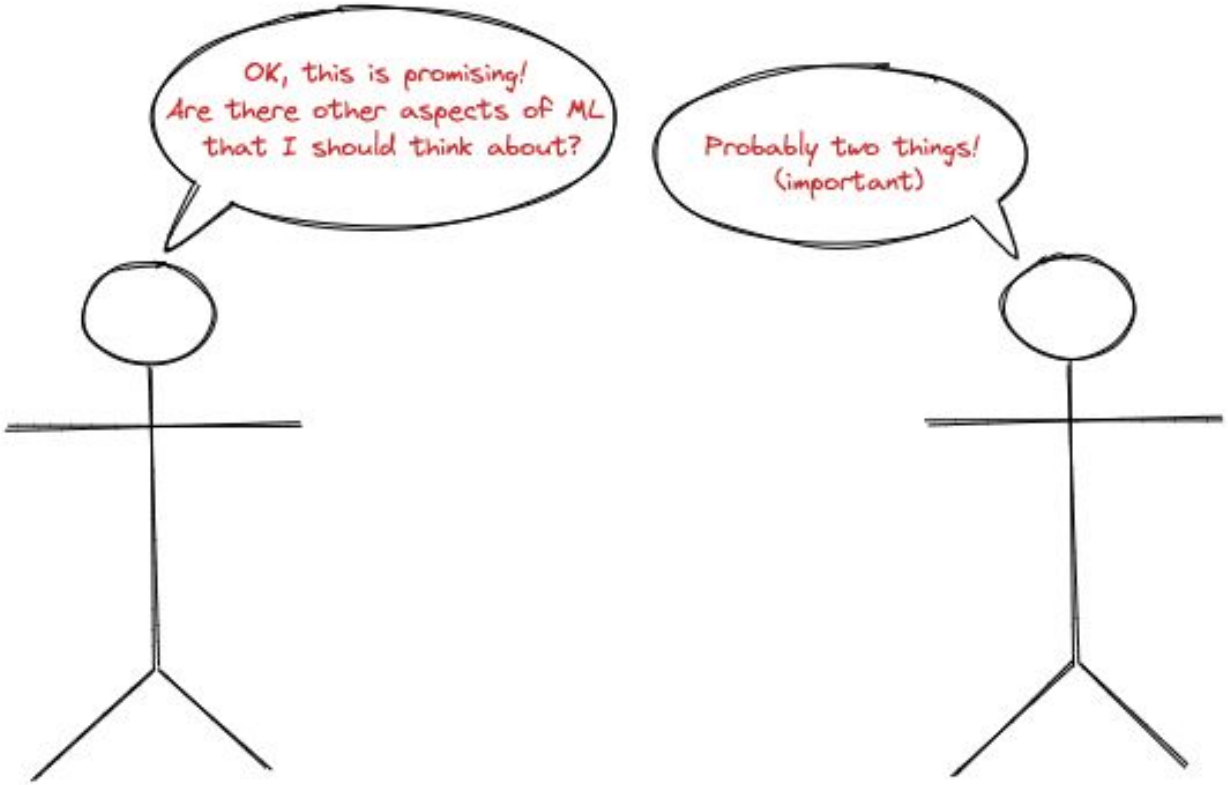
https://datasetsearch.research.google.com/

https://www.kaggle.com/datasets



Access to ML is being democratized...




OK, this is promising!
Are there other aspects of ML that I should think about?
Probably two things!
(important)

Ethics

Zoom's Virtual Background Feature Isn't Built for Black Faces

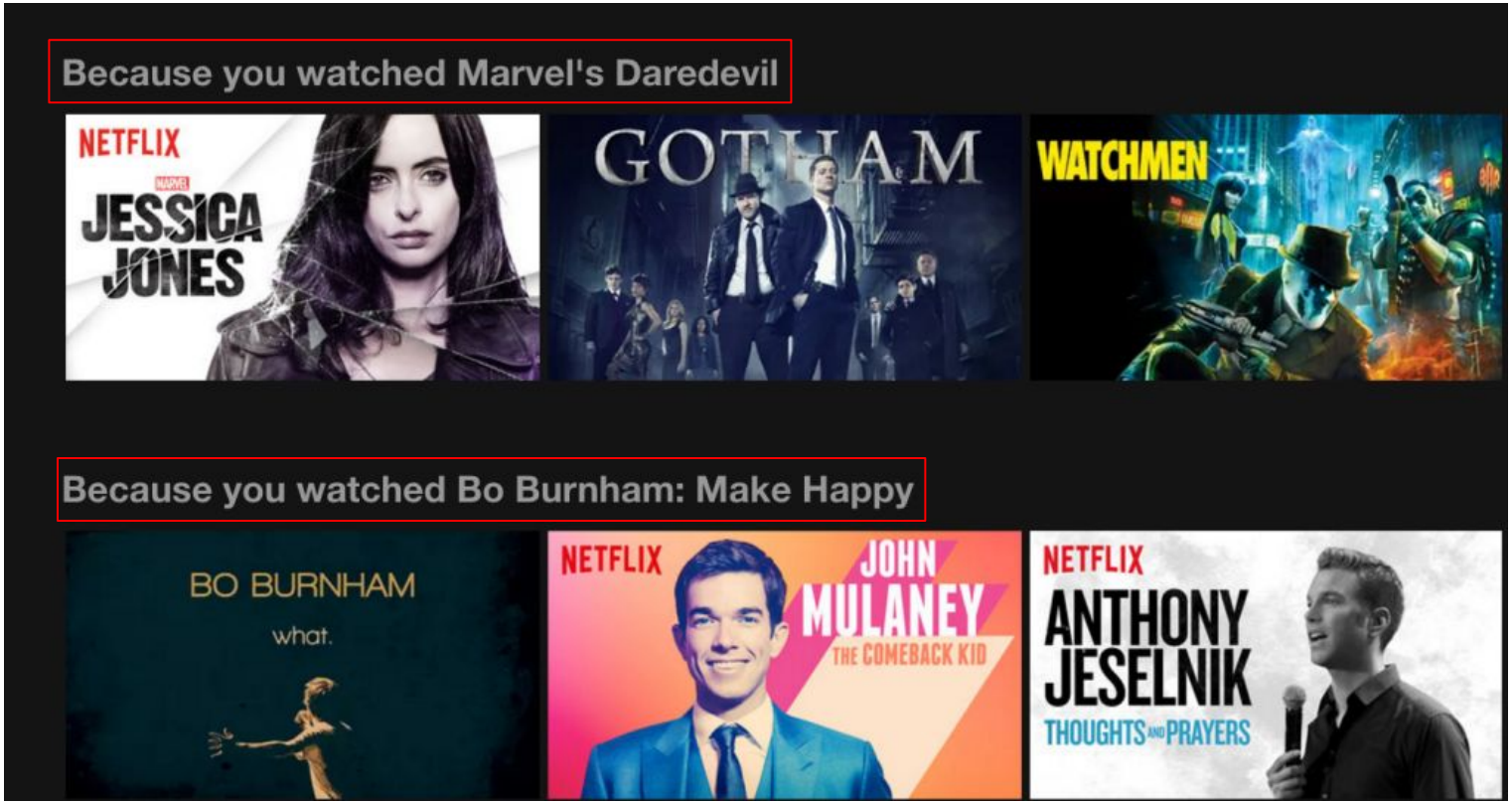
A scientist warns that bias in facial recognition software could lead to false arrests, lost job opportunities

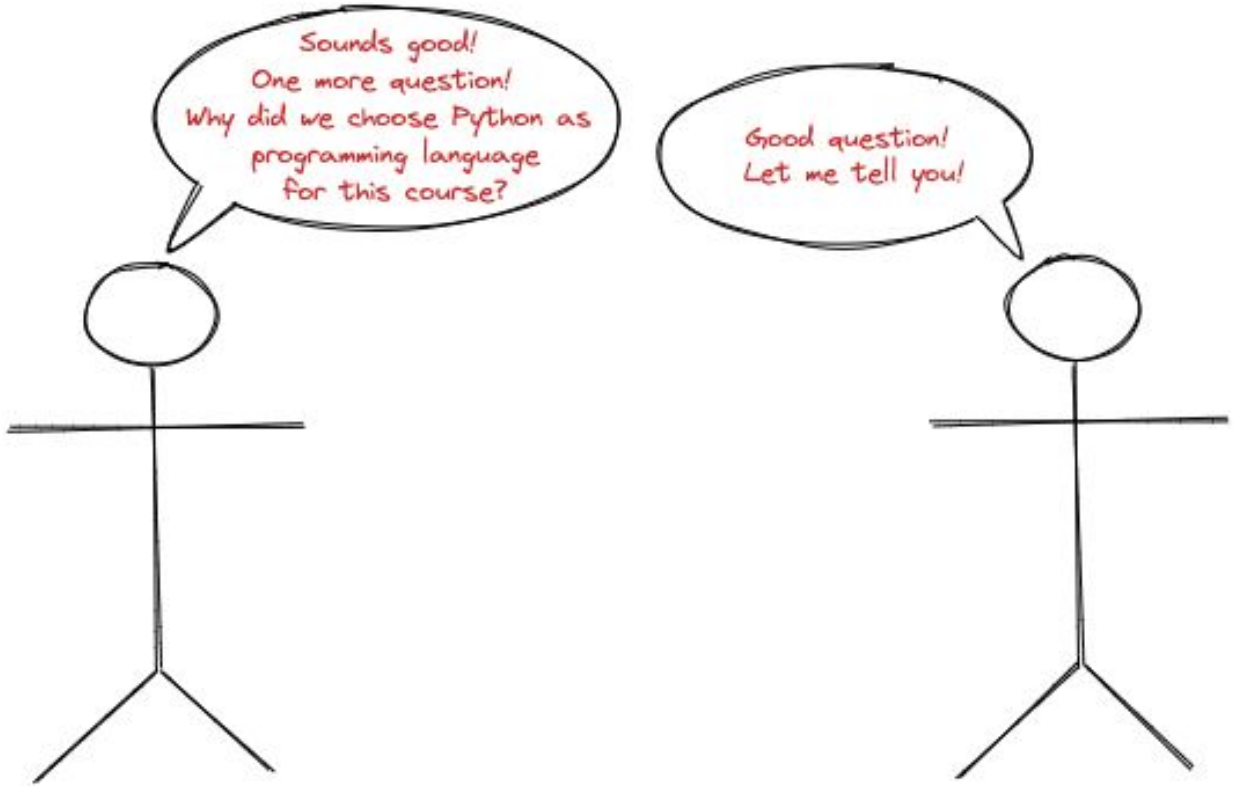
Drew Costley Oct 26, 2020 · 2 min read ★

Ainissa Ramirez says she's seen Black and dark-skinned colleagues disappear into their virtual backgrounds on Zoom calls a few times this year. And she isn't the only one.

"I have heard reports that Black people are fading into their Zoom backgrounds because supposedly the algorithms are not able to detect faces of dark complexions well," Ramirez, PhD, former professor of mechanical engineering at Yale University, tells *OneZero*.

Explainability




Sounds good!
One more question!
Why did we choose Python as programming language for this course?
Good question!
Let me tell you!

Python is the de-facto language for ML

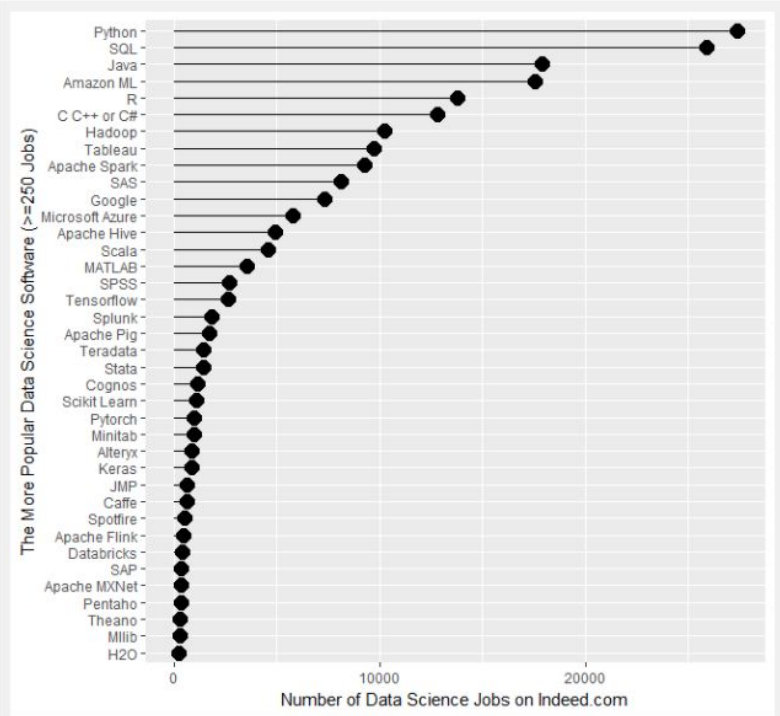
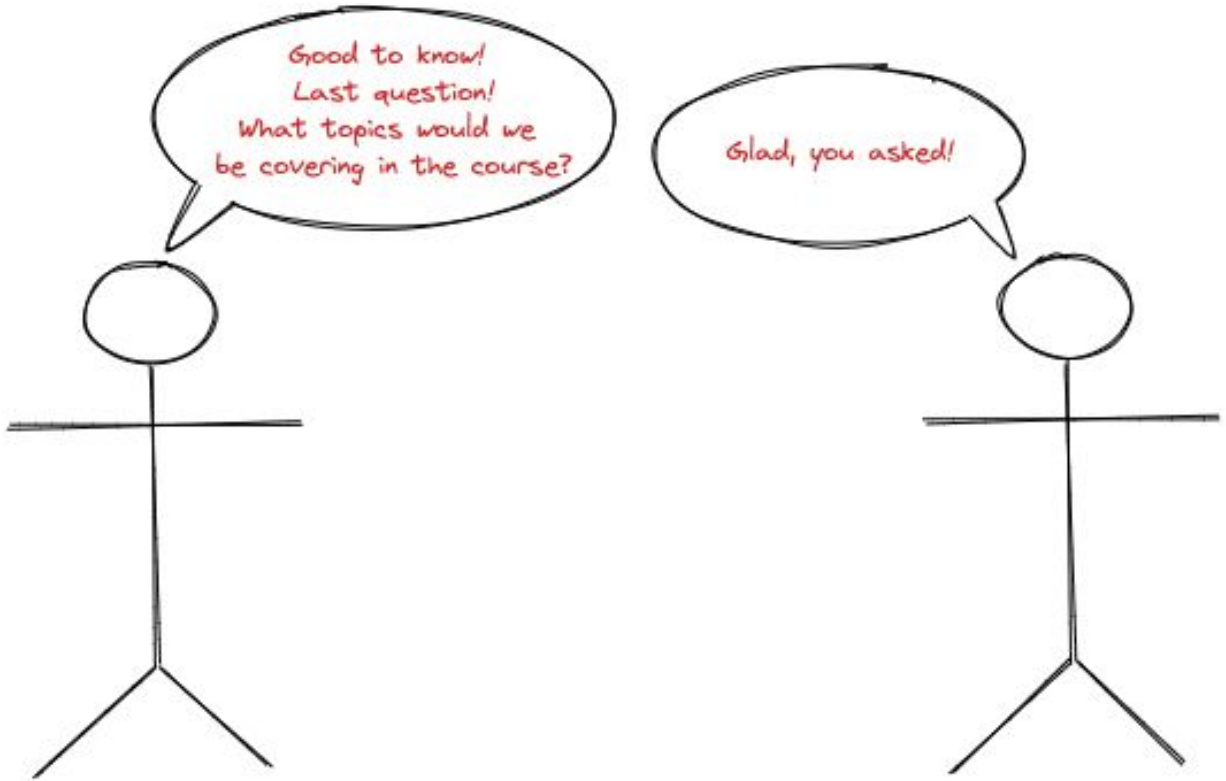


Figure 1a. Number of data science jobs for the more popular software.

http://r4stats.com/articles/popularity/

Great suite of matured libraries for ML tasks


Good to know!
Last question!
What topics would we
be covering in the course?
Glad, you asked!

Course schedule

	Week	Topics	Assignments	By the end of class
1	09/15	Introduction Exploratory Data Analysis & Visualization		Students would be familiar with basic data exploration
2	09/22	Introduction to supervised learning Preprocessing	HW1 posted	
3	09/29	Linear models for regression Linear models for classification, SVMs	HW1 due	
4	10/06	Trees, Forests & Ensembles Gradient Boosting, Calibration	HW2 posted	
5	10/13	Model evaluation Parameter tuning & automatic machine learning	HW2 due	Students would be familiar with training & evaluation of linear and ensemble models

Course schedule

	Week	Topics	Assignments	By the end of class
6	10/20	Model Interpretation & Feature Selection Linear & non-linear dimensionality reduction Clustering & mixture models	HW3 posted	
	10/27	Midterm		
7	11/03	Learning with imbalanced data Learning with sparse data	HW3 due	
8	11/10	Deep Neural Networks (DNN) Convolutional Neural Networks	HW4 posted	
9	11/17	Advanced neural networks	HW4 due	Students would be familiar with applying neural networks to different tasks

Course schedule

	Week	Topics	Assignments	By the end of class
10	12/01	Working with text data Topic models for text data Word & document embeddings	HW5 posted	Students would be familiar working with text data
11	12/08	Content-based recommendations Collaborative filtering & matrix factorization Recommendations using DNNs	HW5 due	Students would be familiar training and evaluating recommender systems
	12/15	Project presentations		

Questions?

Let’s take a 10 min break!

Exploratory Data Analysis
&
Visualization

Exploratory Data Analysis (EDA) is an approach of analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

https://en.wikipedia.org/wiki/Exploratory_data_analysis

Why do we do EDA?

- Explore
- Inform
- Communicate

Data types

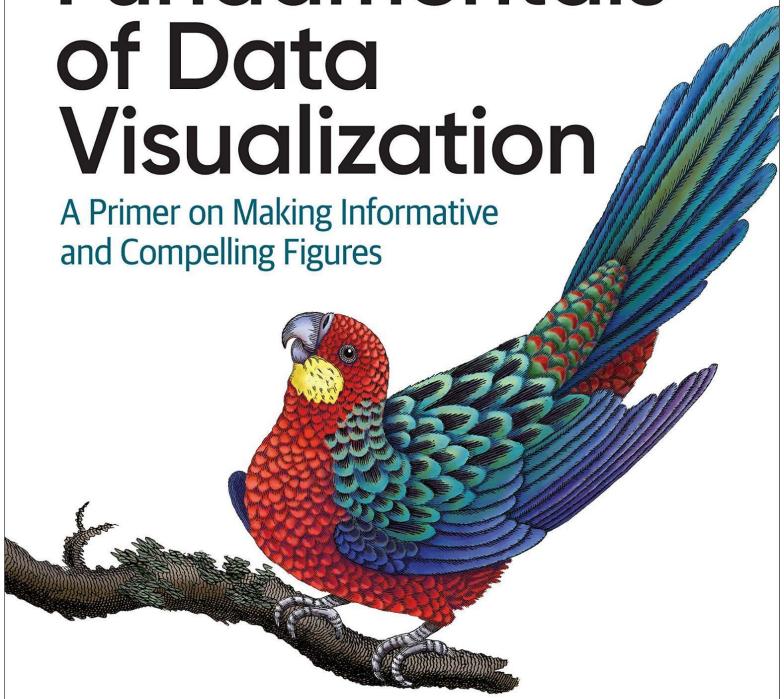
- Quantitative/numerical continuous - 1, 3.5, 100, 10^10, 3.14
- Quantitative/numerical discrete - 1, 2, 3, 4
- Qualitative/categorical unordered - cat, dog, whale
- Qualitative/categorical ordered - good, better, best
- Date or time - 09/15/2021, Jan 8th 2020 15:00:00
- Text - The quick brown fox jumps over the lazy dog

Data Visualization

O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

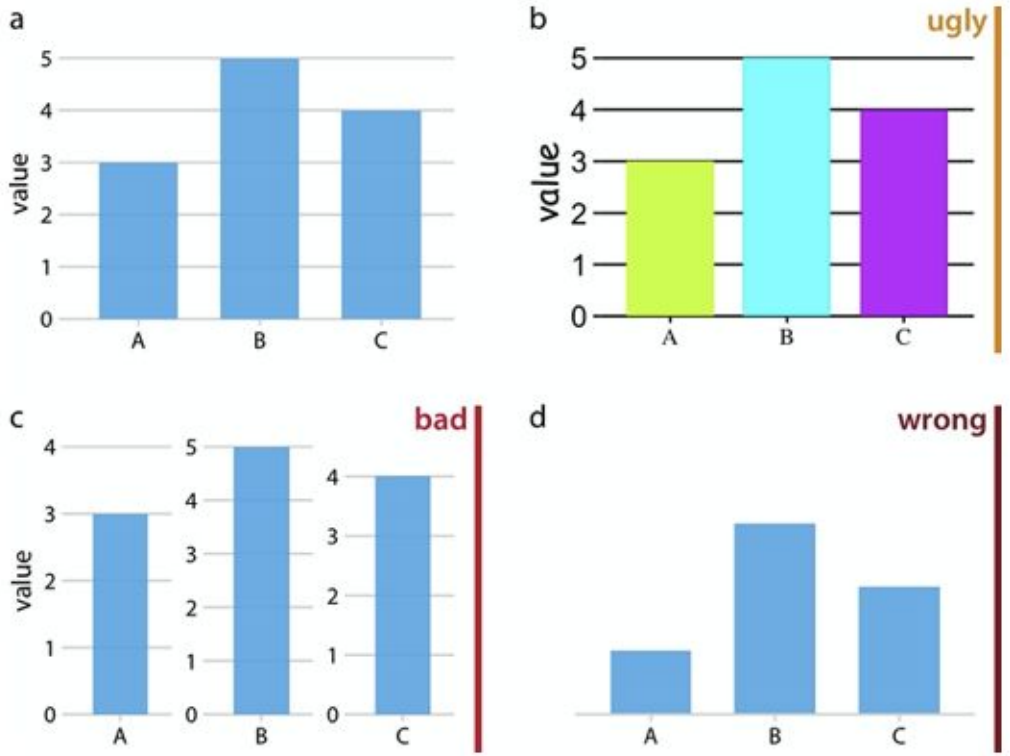


Claus O. Wilke

Ugly, Bad & Wrong figures

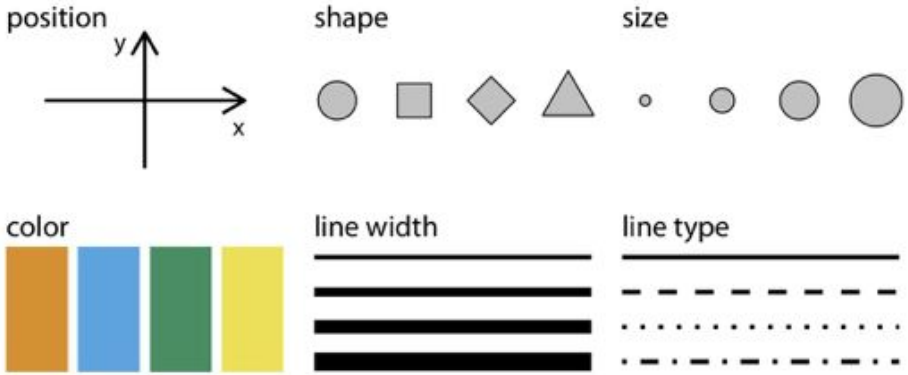
- Ugly
  - A figure that has aesthetic problems but otherwise is clear and informative
- Bad
  - A figure that has problems related to perception; it may be unclear, confusing, overly complicated, or deceiving
- Wrong
  - A figure that has problems related to mathematics; it is objectively incorrect

Ugly, Bad & Wrong figures



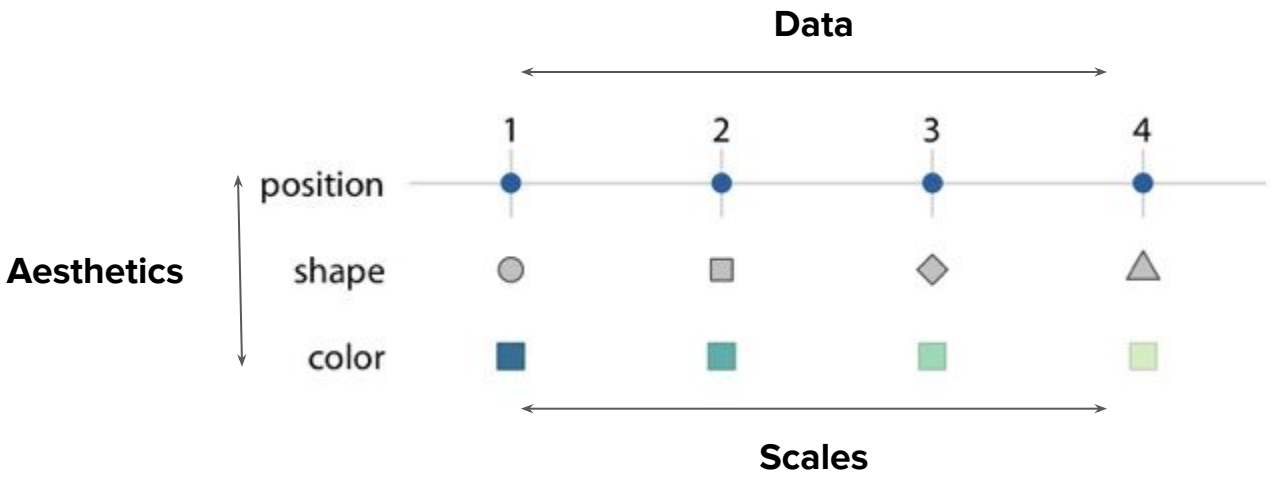
Aesthetics in data visualization

- Aesthetics refer to a quantifiable set of features that are mapped to the data in a graphic.
- Aesthetics describe every aspect of a given graphical element.
- Some aesthetics like position, size, color and line width work for both continuous & discrete data, while others (shape & line type) work for only discrete data



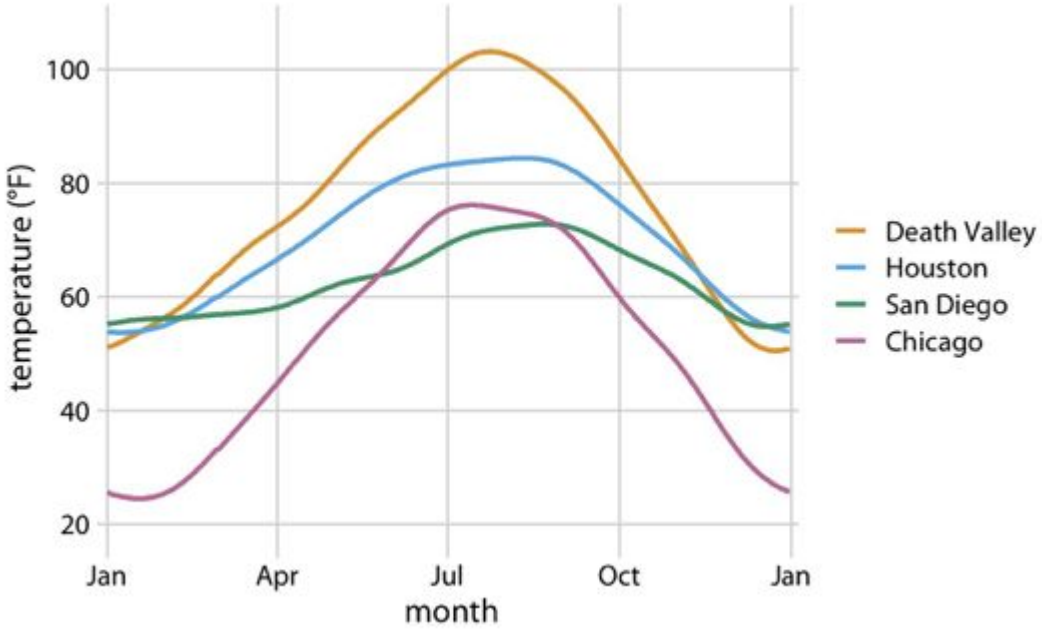
Scales

- Scales are the mapping between data values and aesthetics values.



A typical data visualization chart

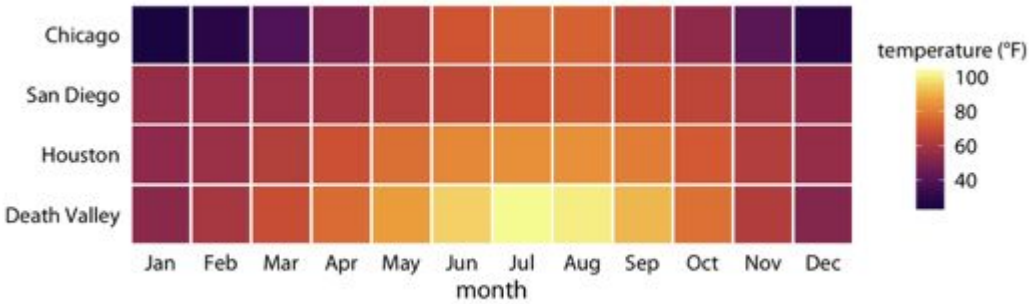
- A typical data visualization chart uses **three** scales.



- **Two position scales:**
  - month (x-axis)
  - temperature(y-axis)
- **One color scale:**
  - location

A typical data visualization chart

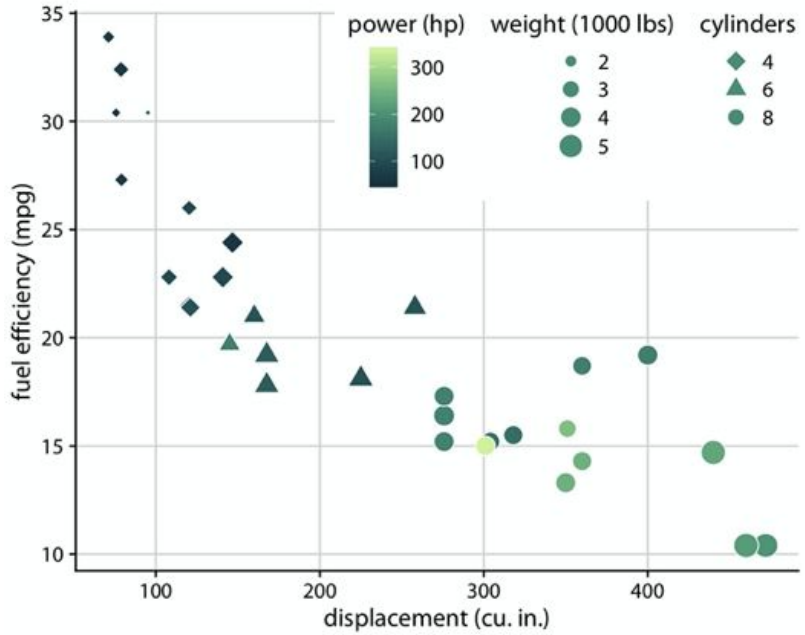
- A typical data visualization chart uses **three** scales.



- **Two position scales:**
  - month (x-axis)
  - location (y-axis)
- **One color scale:**
  - temperature

An (a)typical data visualization chart

- This visualization chart uses **five** scales.

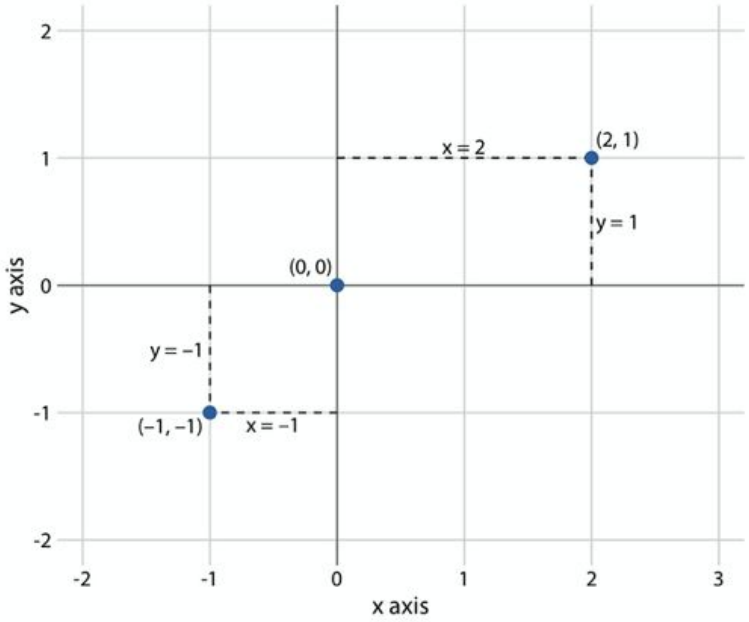


- **Two position scales:**
  - displacement (x-axis)
  - fuel efficiency (y-axis)
- **One color scale:**
  - power
- **One shape scale:**
  - cylinders
- **One size scale:**
  - weight

Position scale

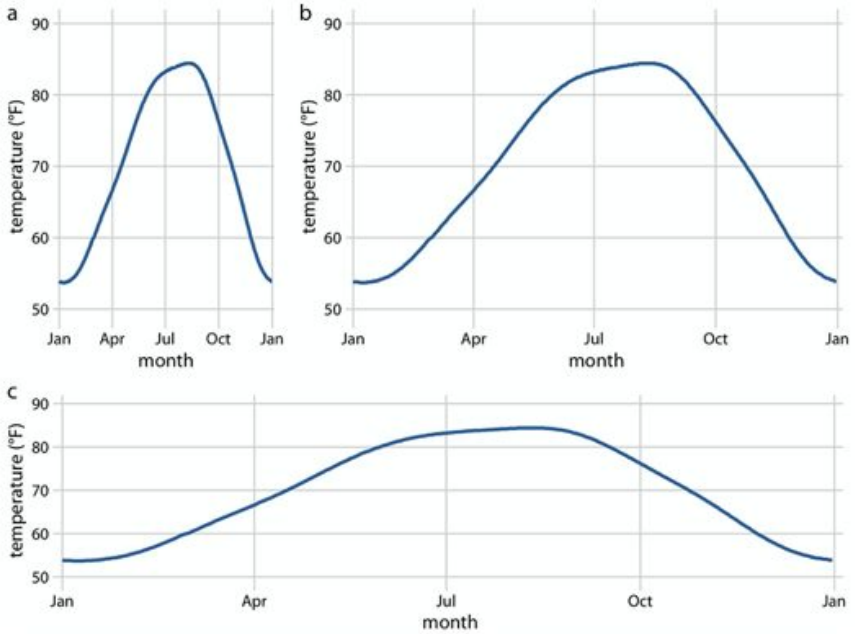
Cartesian coordinate system

- The most widely used coordinate system for data visualization.



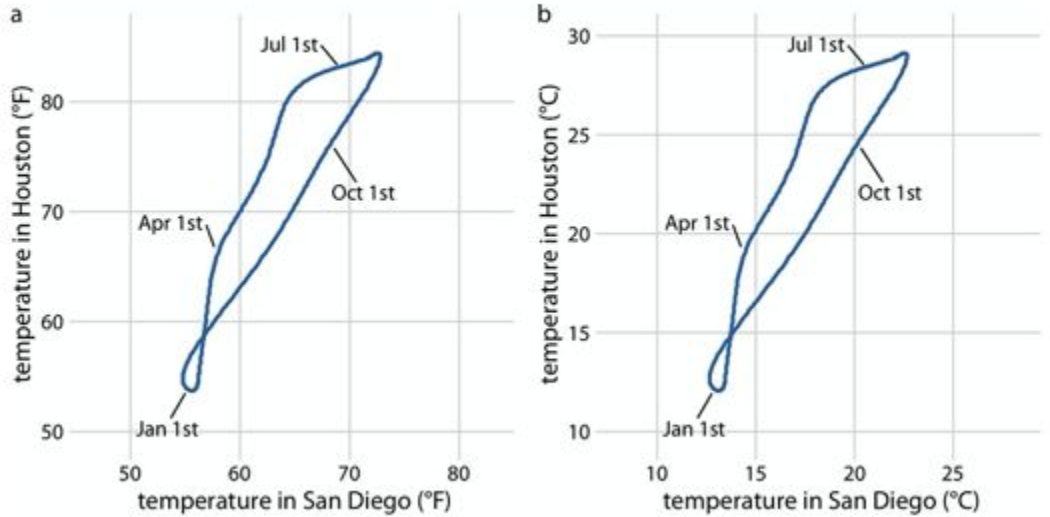
Valid data visualization charts

- All figures show the same information with different aspect ratios



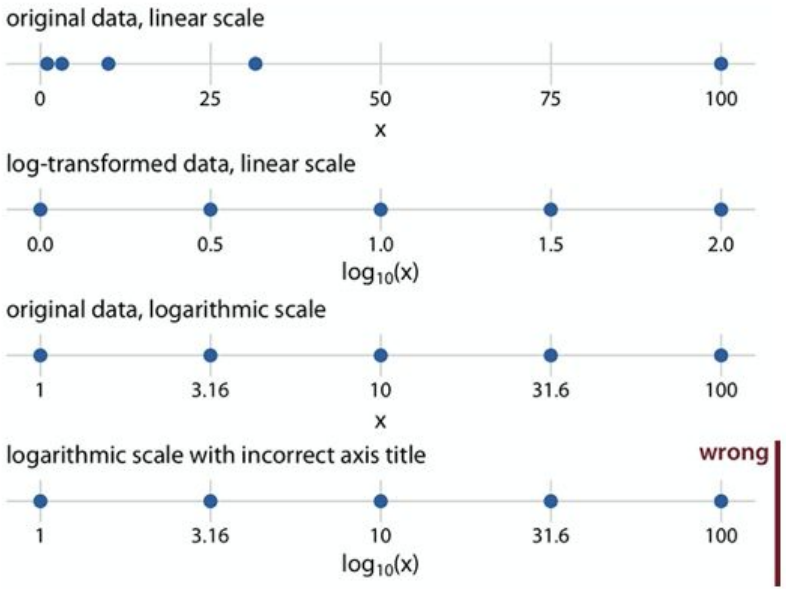
Valid data visualization charts

- Since the same quantity (temperature) is plotted on both axes, the grid spacings should be same.



Nonlinear axes

- Logarithmic scale is the most commonly used nonlinear scale.

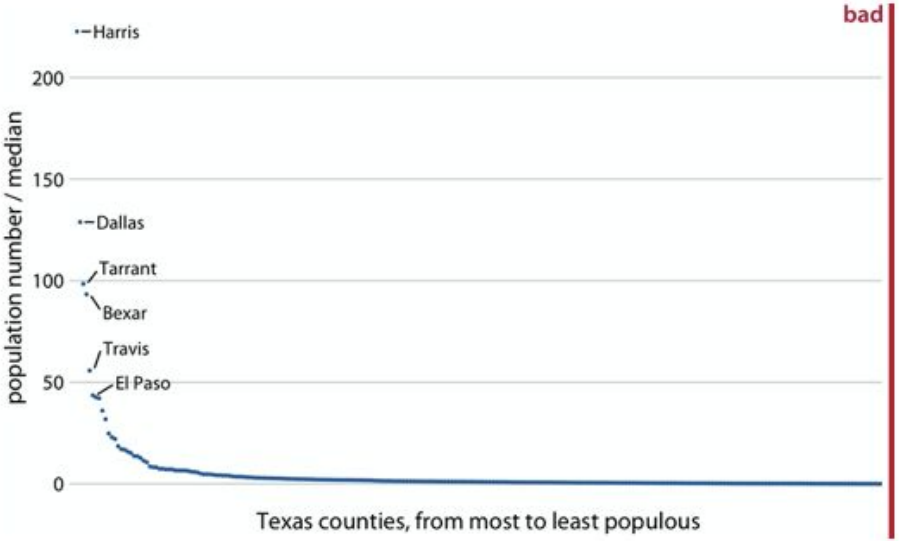


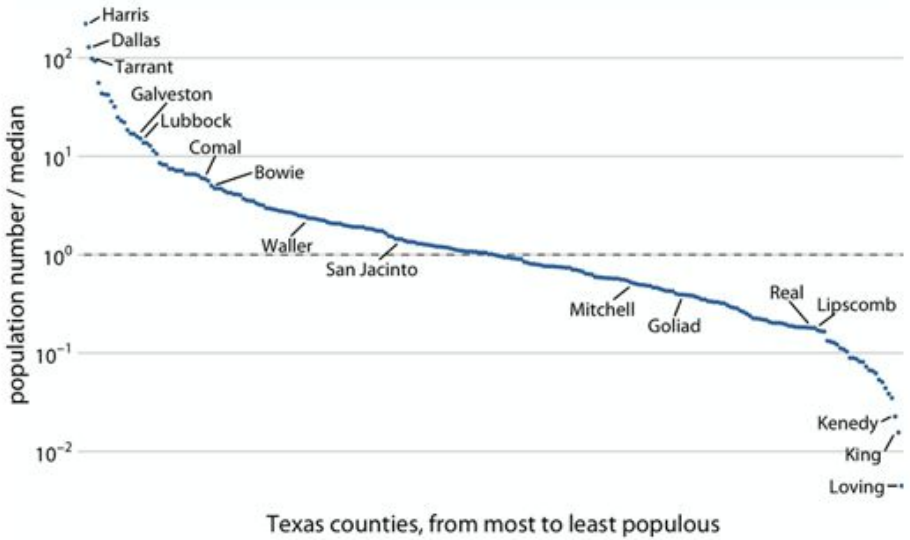
[1, 3.16, 10, 31.6, 100]

[0, 0.5, 1, 1.5, 2]

Logarithmic scale - an example

- Logarithmic scale is typically useful to represent ratios.





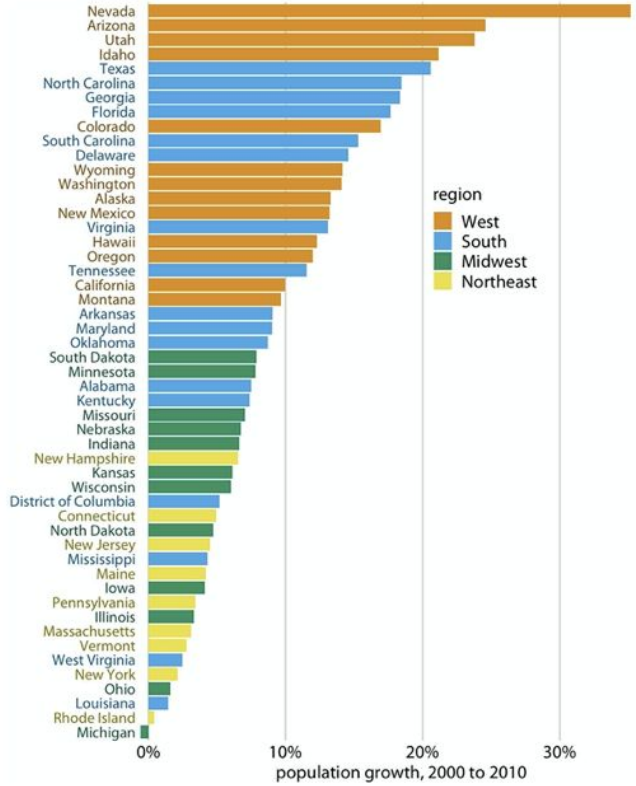
Color scale

Use-cases for color in data visualization

- Three fundamental use-cases for using color in data visualization:
  - distinguish groups of data
  - Represent data values
  - Tool to highlight

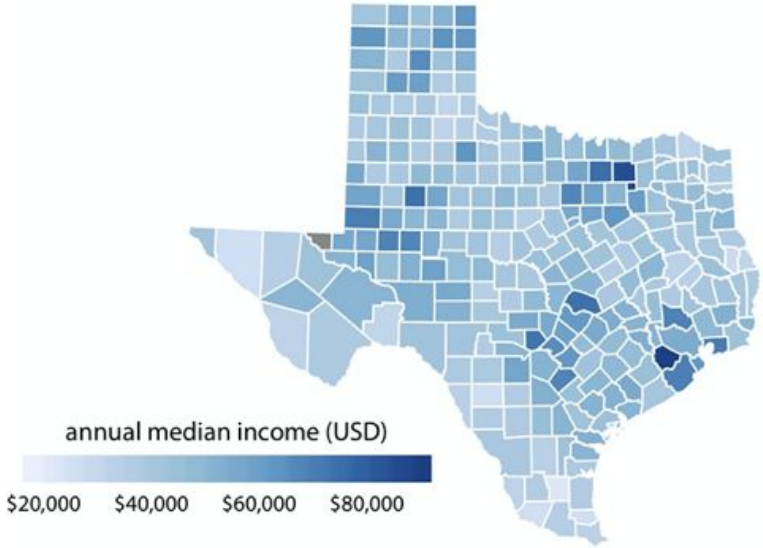
Color as a tool to distinguish

Qualitative color scale

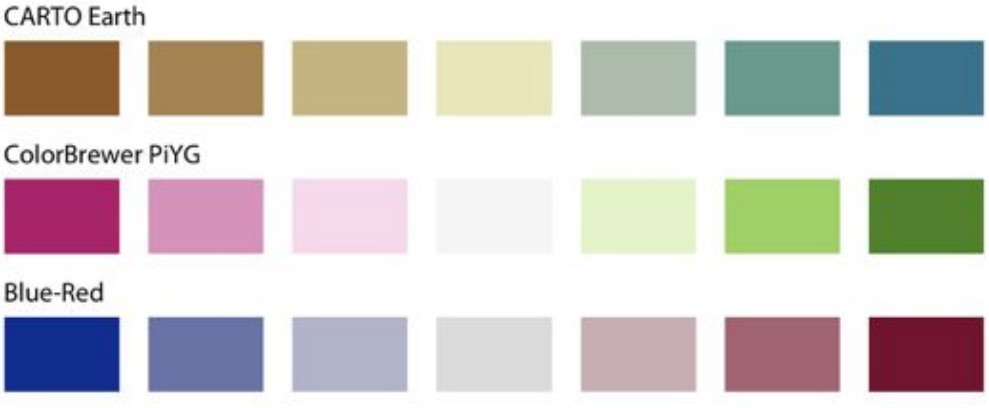


Color to represent data values

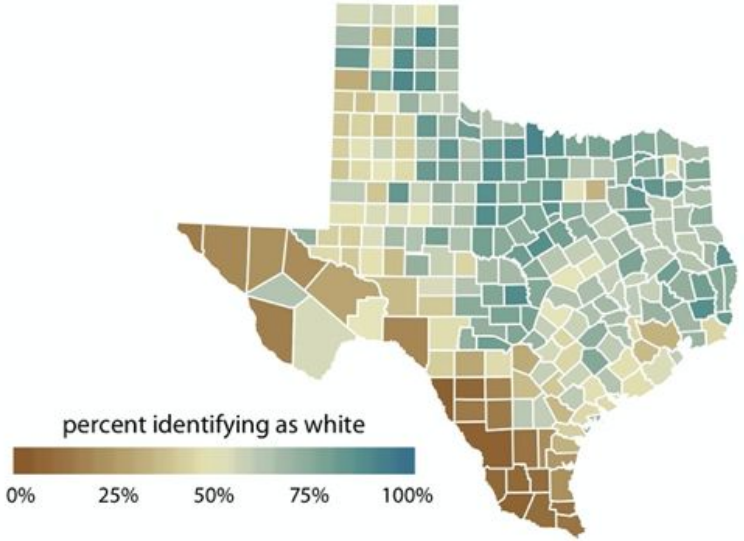
Sequential color scale



Color to represent data values

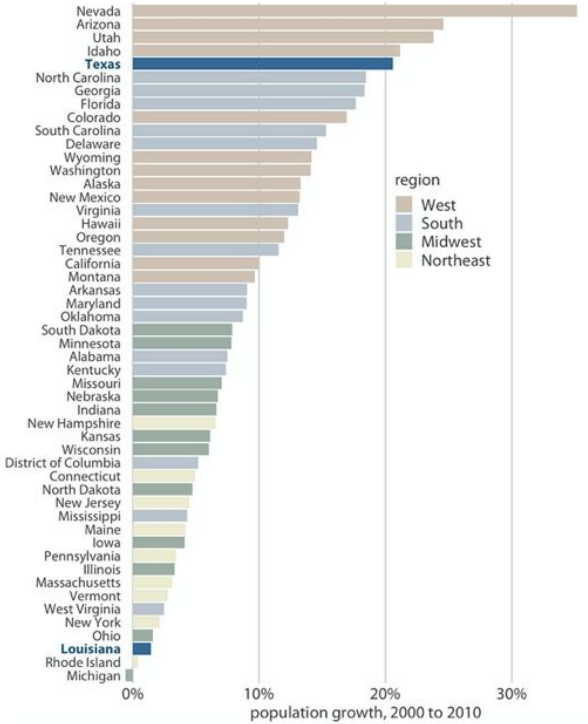


Diverging color scale



Color as a tool to highlight

Accent color scale



The two neighboring states Louisiana and Texas experienced among the highest and lowest population growth from 2000 to 2010.

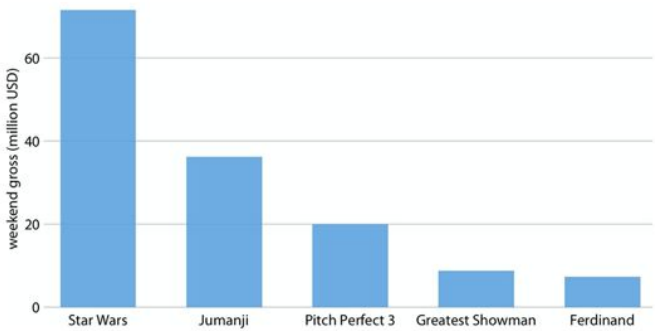
Visualization Collections

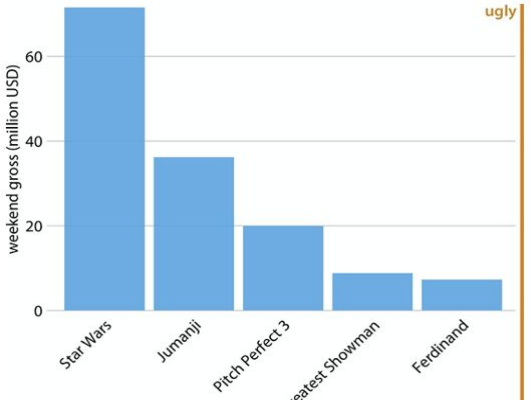
Visualizing data

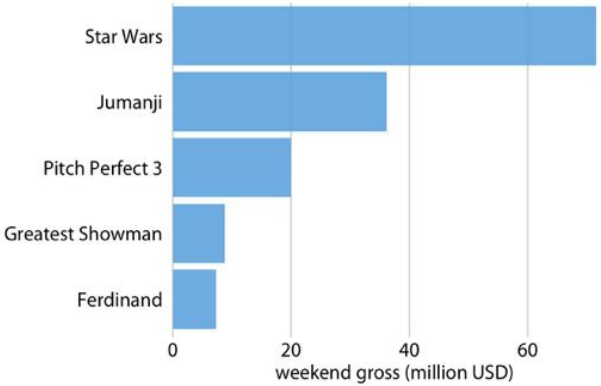
- Typically, we would like to visualize the following kinds of data:
  - Amounts
  - Distributions
  - Proportions
  - X-Y relationships
  - Uncertainty

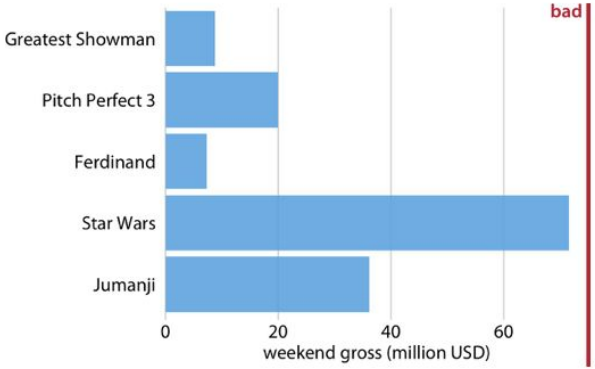
Visualizing amounts

Visualizing amounts - bar plots



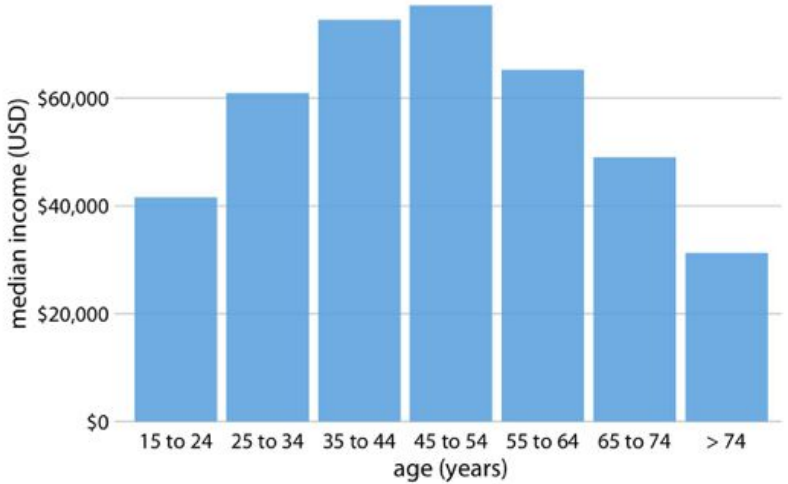


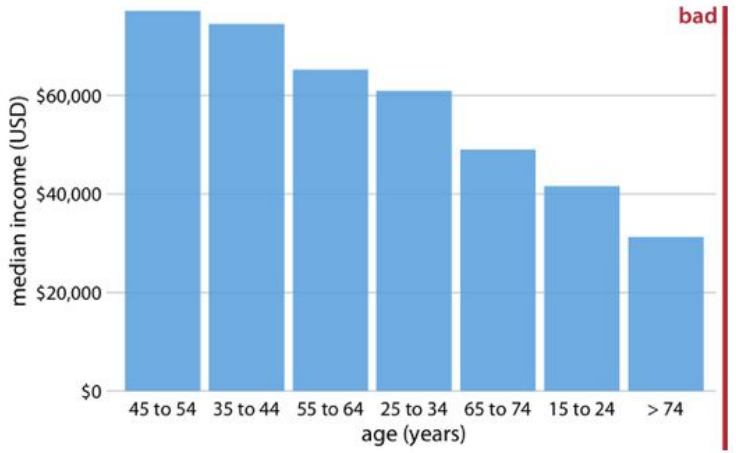




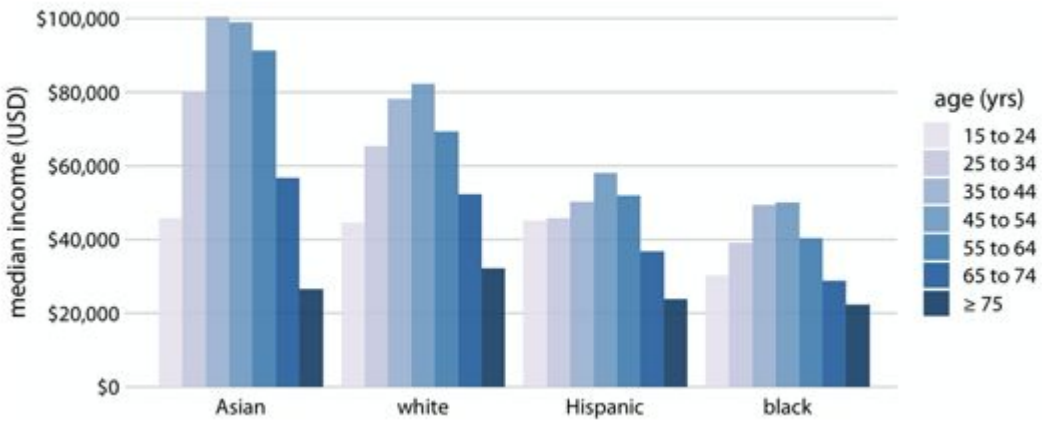
Visualizing amounts - bar plots

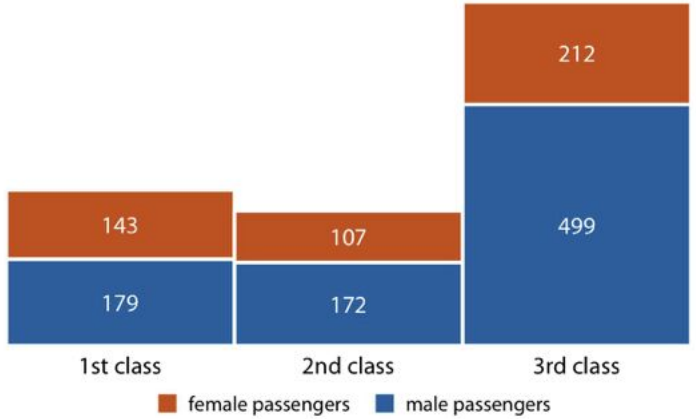
- The bars should not be ordered if they represent ordered categories



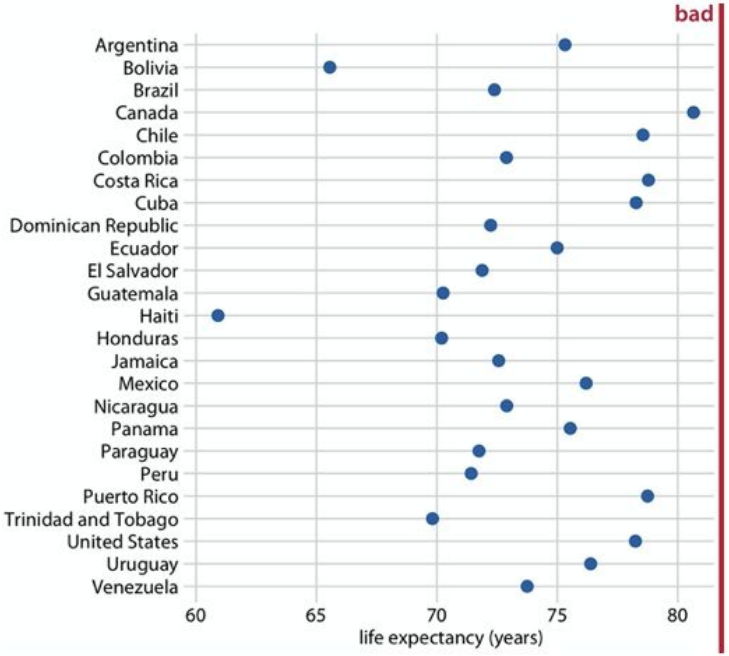


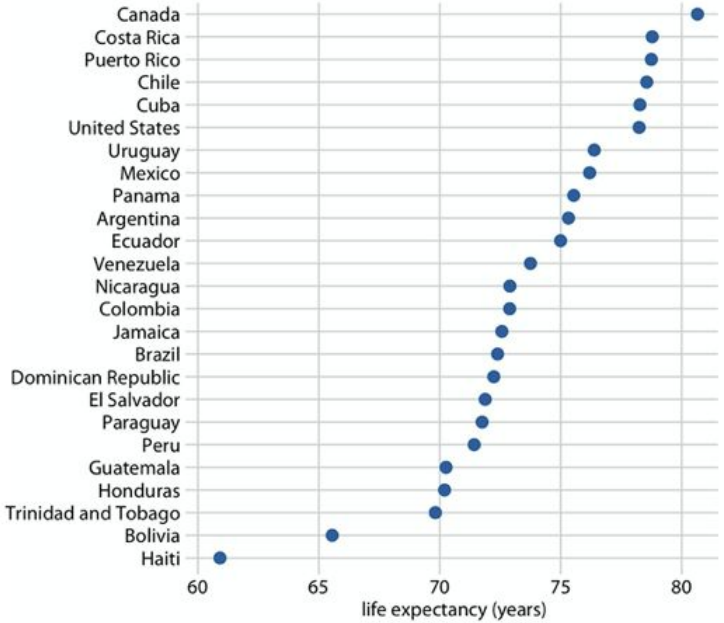
Visualizing amounts - grouped & stacked bars





Visualizing amounts - dot plots



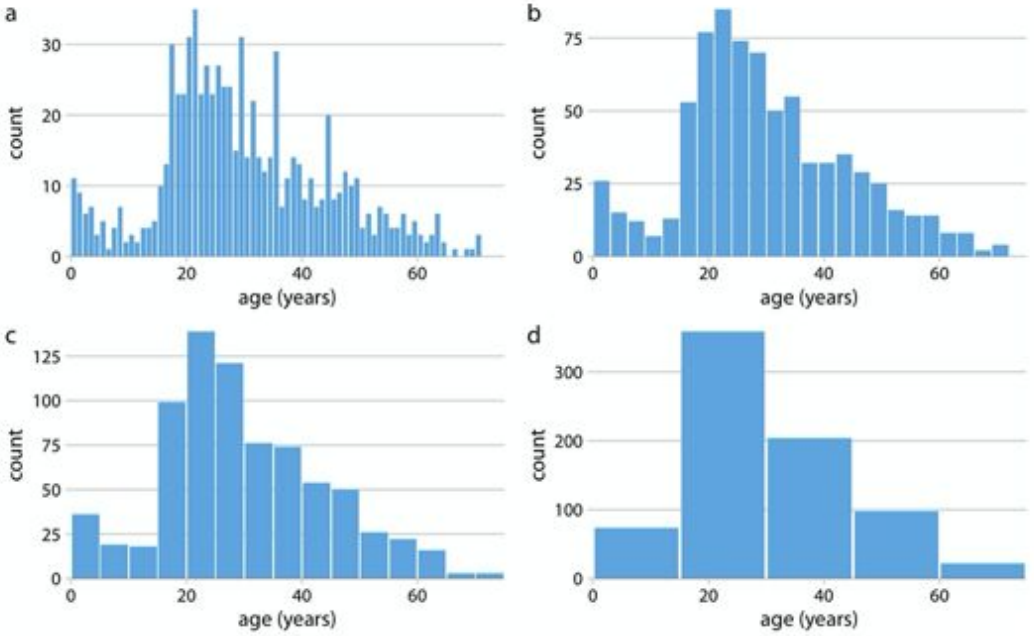


Visualizing distributions

Visualizing distributions - histograms

- When making histograms, always explore multiple bin widths

bin width = 1 year



bin width = 3 years

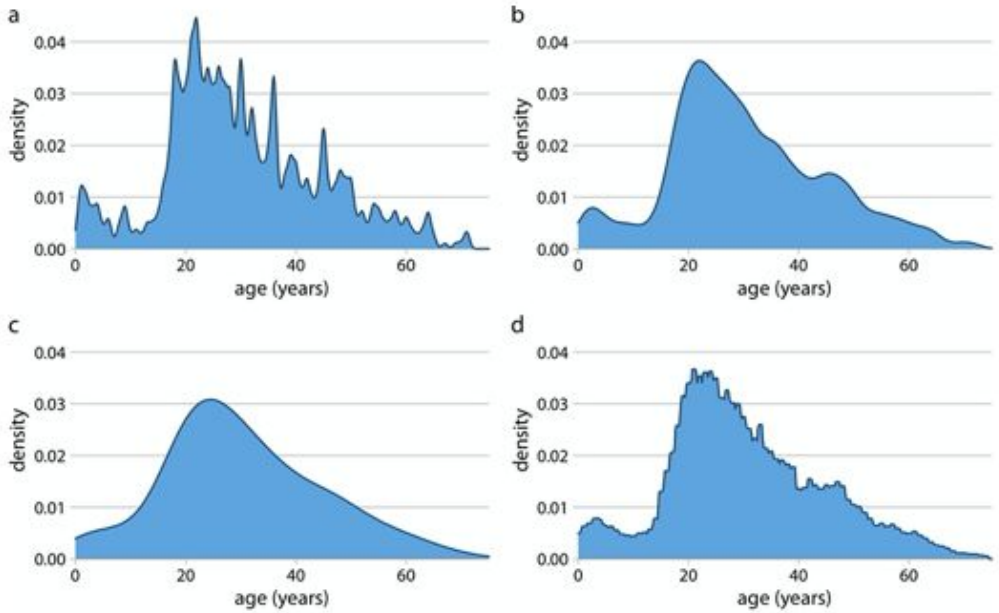
bin width = 5 years

bin width = 15 years

Visualizing distributions - kernel density plots

- Different kernels include Gaussian, Rectangular etc.
- Each kernel is parameterized by bandwidth

Gaussian kernel bandwidth = 0.5



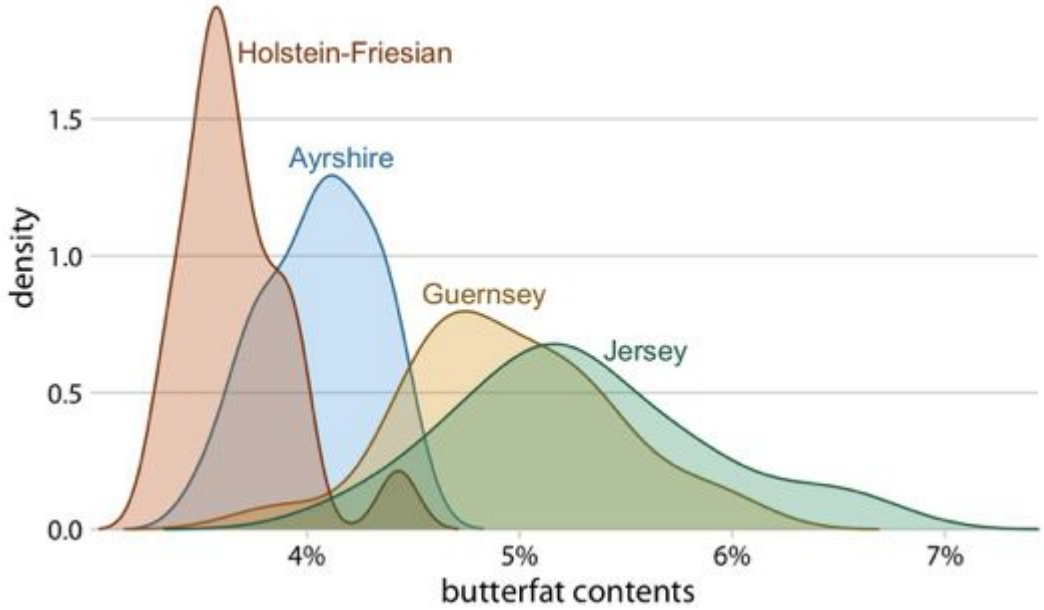
Gaussian kernel bandwidth = 2

Gaussian kernel bandwidth = 5

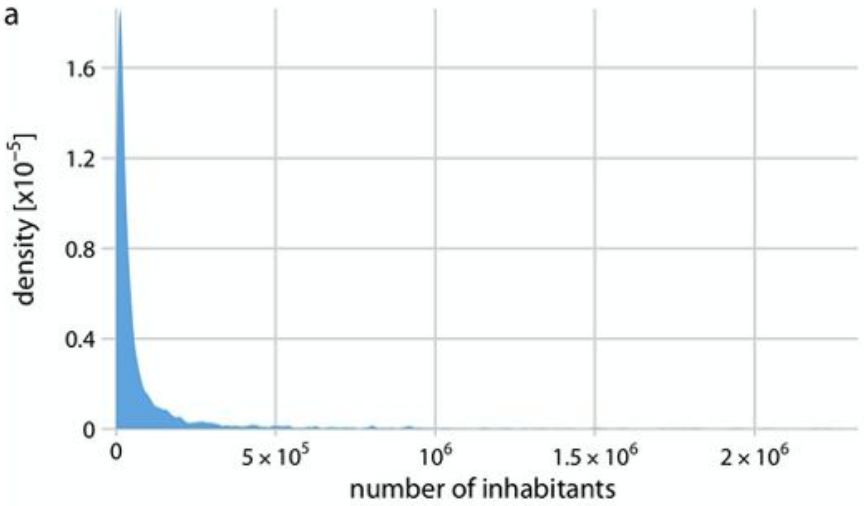
Rectangular kernel bandwidth = 2

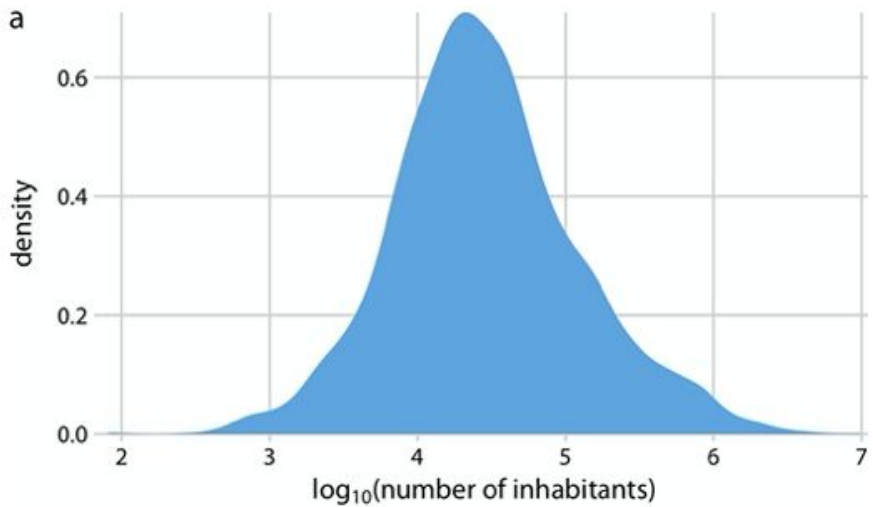
Visualizing distributions - kernel density plots

- To visualize several distributions at once, kernel density plots work better than histograms



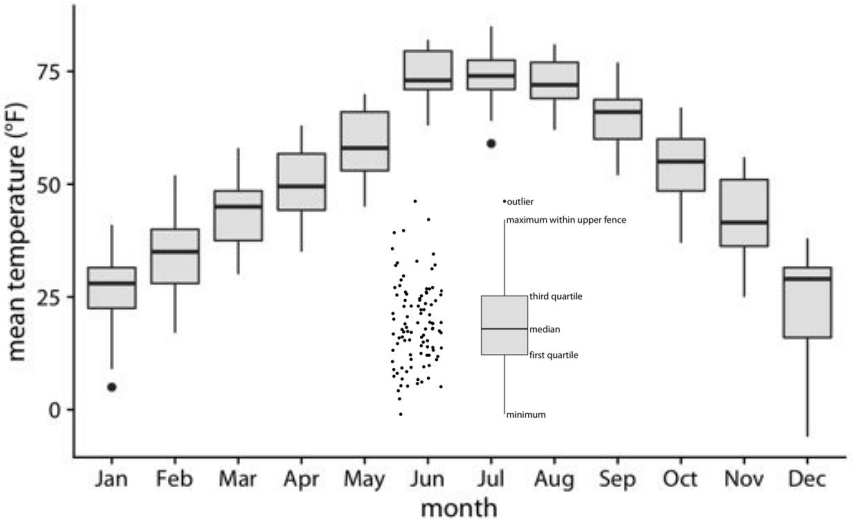
Visualizing distributions - highly skewed distribution



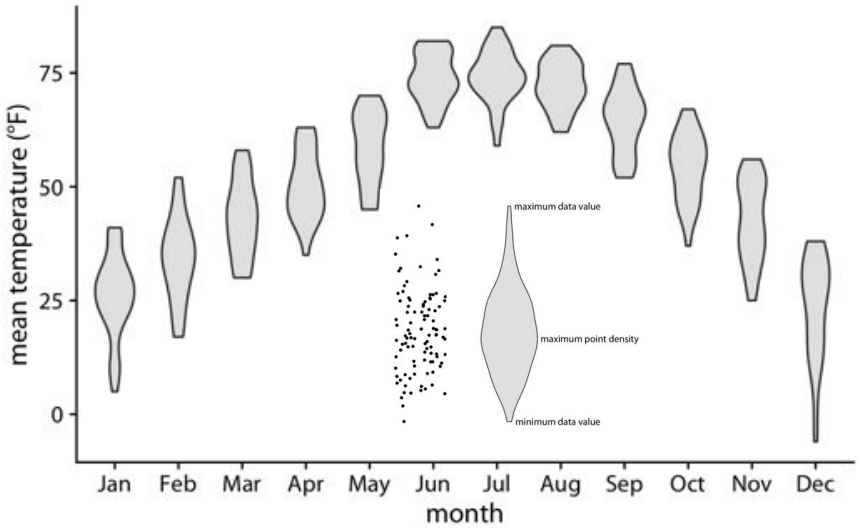


Visualizing distributions - multiple distributions

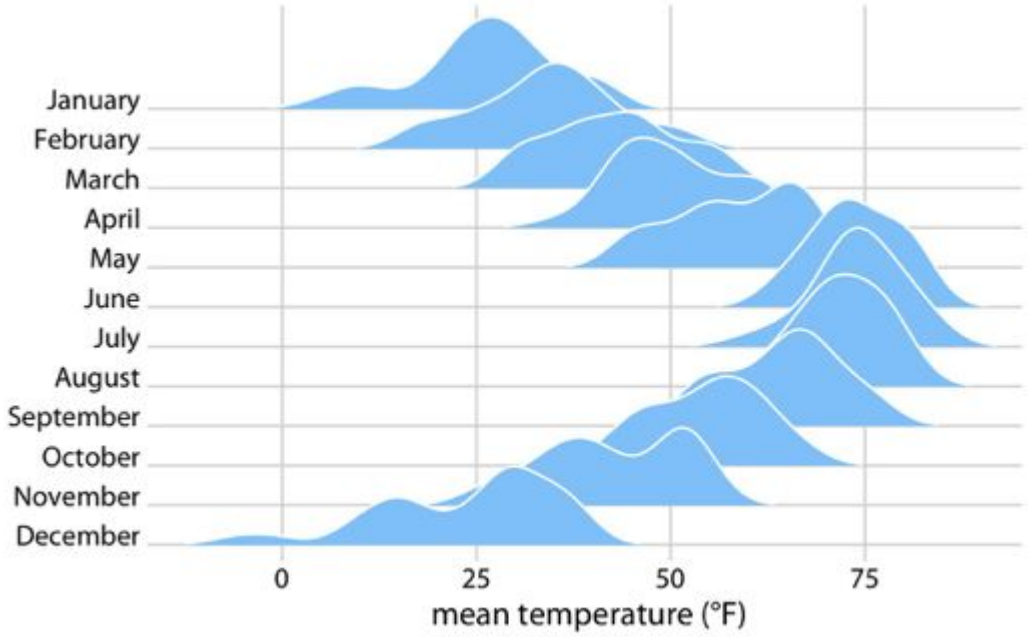
Boxplot



Violin plot



Visualizing distributions - multiple distributions

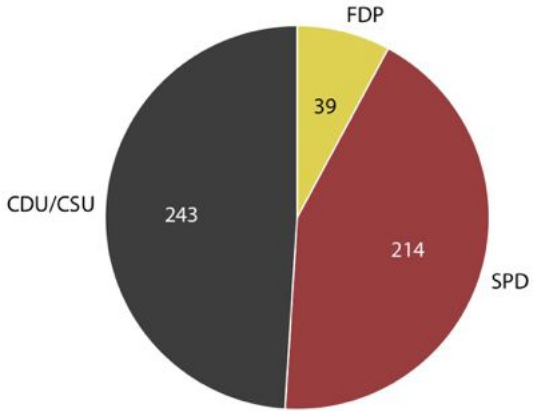


Ridgeline plot

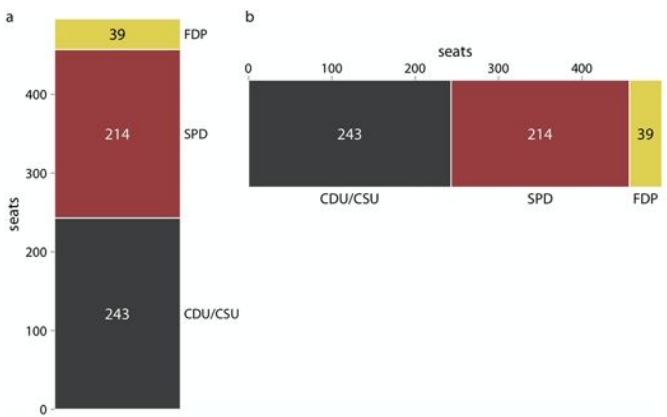
Visualizing proportions

Visualizing proportions - pie charts, stacked & side-by-side bars

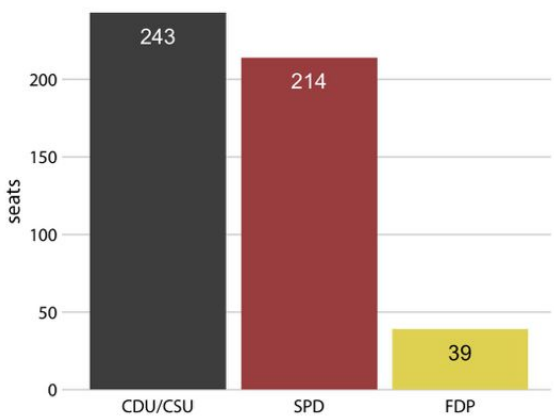
- Pie charts help visually emphasize simple fractions, such as ½, 1/3 , ¼ etc.



Pie chart



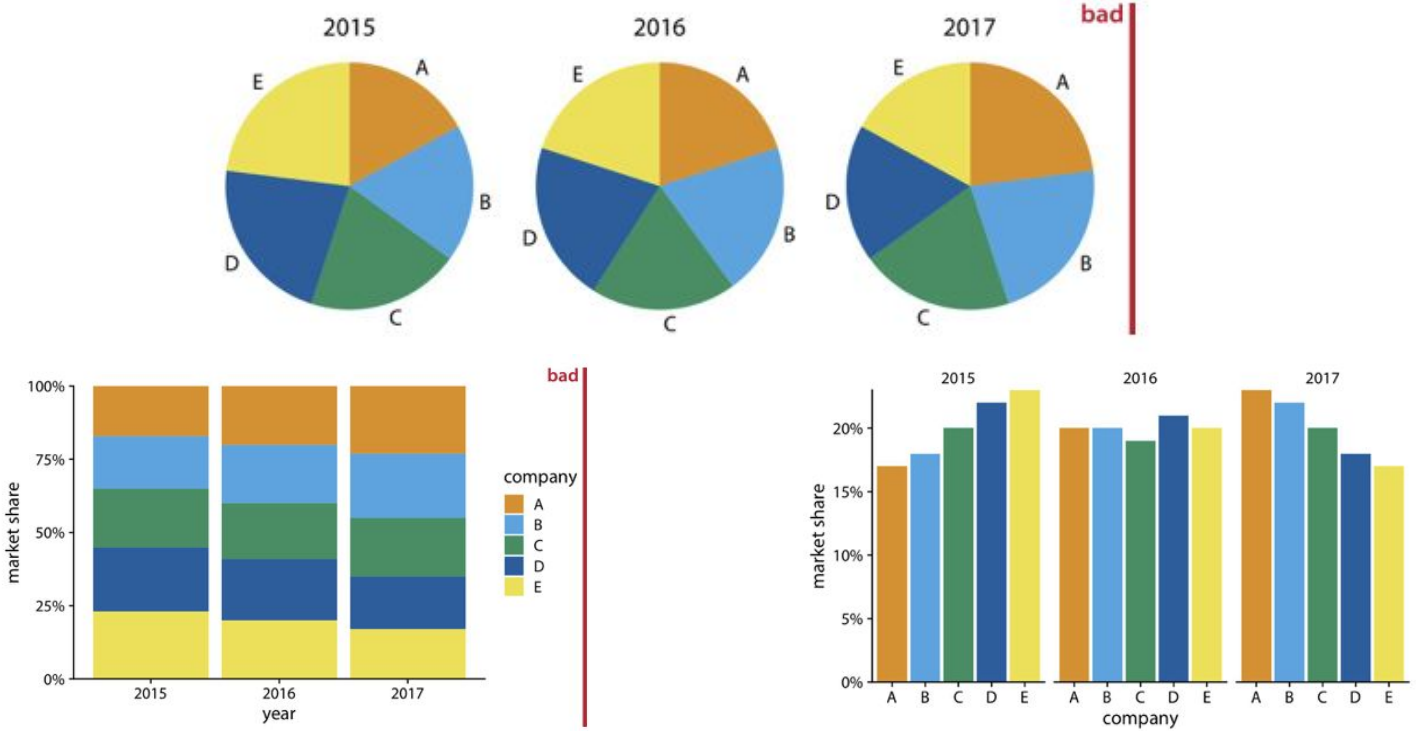
Stacked bar



Side-by-side bar

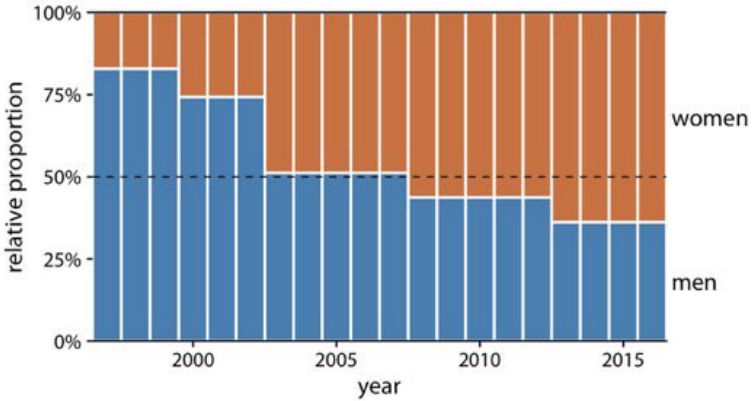
Visualizing proportions - pie charts, stacked & side-by-side bars

- Side-by-side bars help visualize easily changing proportions over time.

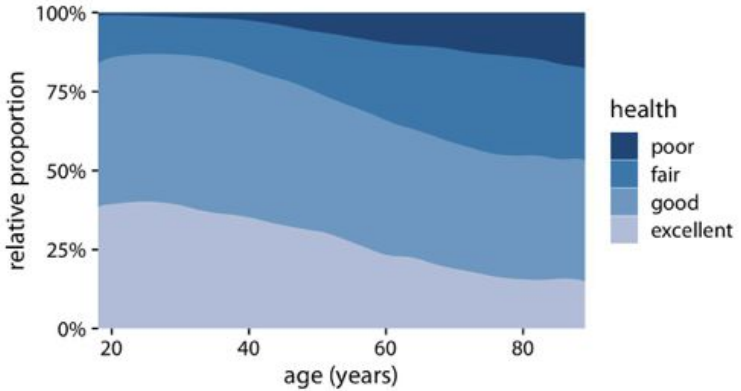


Visualizing proportions - pie charts, stacked & side-by-side bars

- Stacked bars are preferred when there are only two quantities to compare over time.
- Stacked density plots can be used to visualize how proportions change in response to a continuous variable.



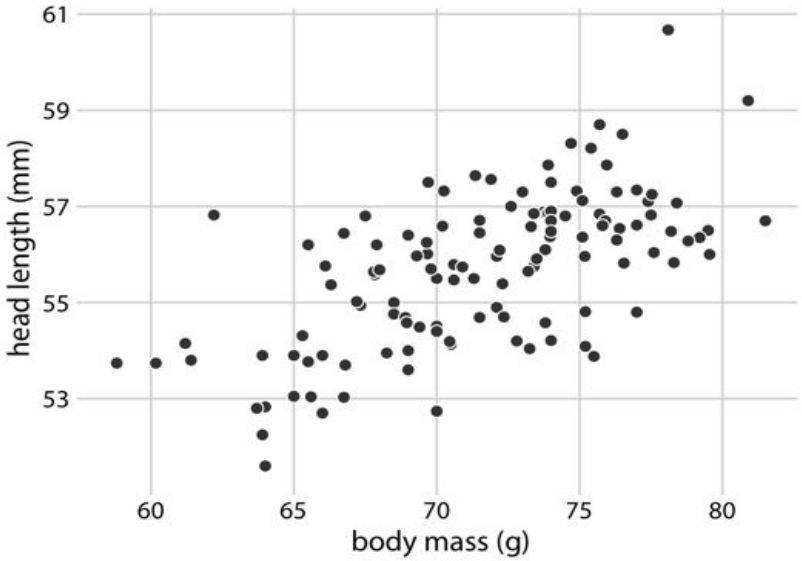
Stacked bar plot

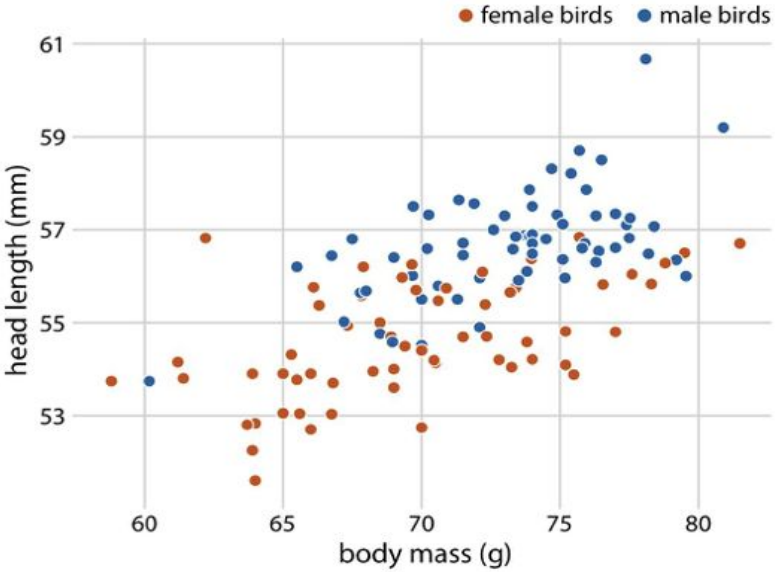


Stacked density plot

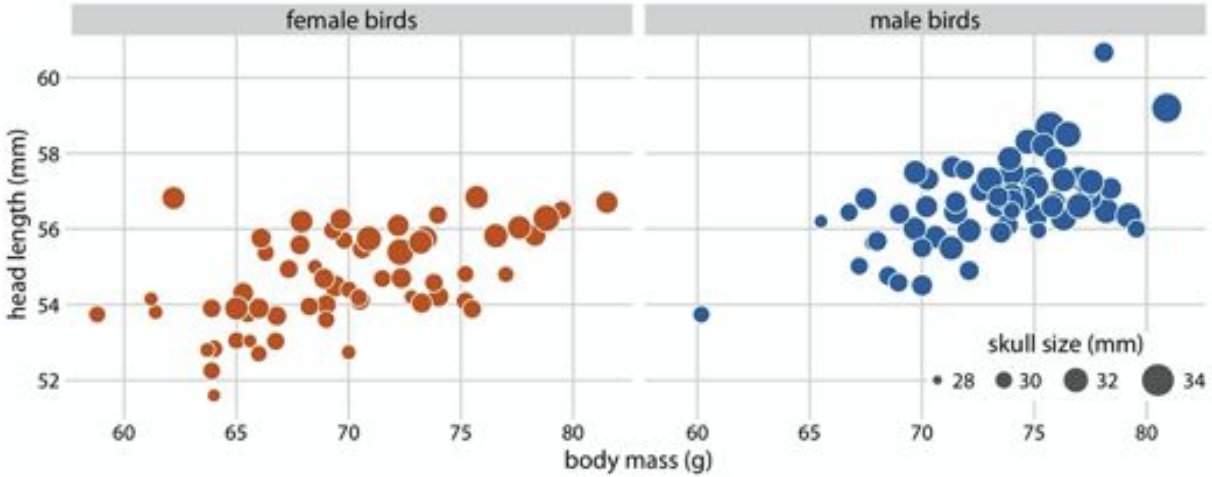
Visualizing X-Y relationships

Visualizing X-Y relationships - scatterplots

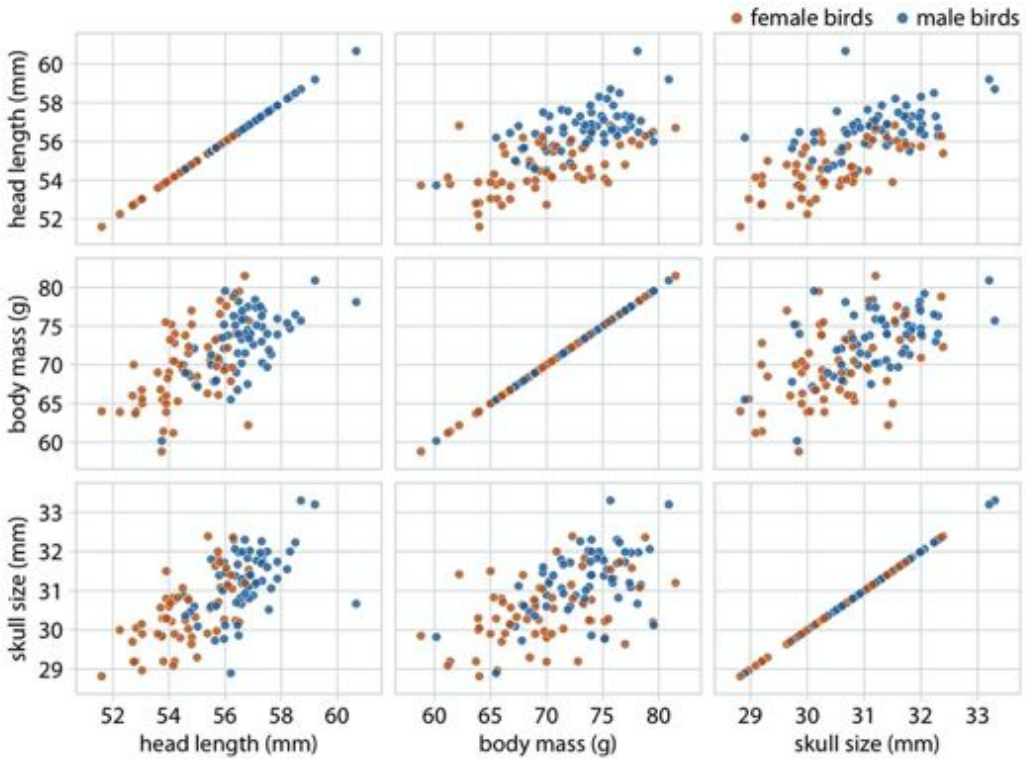




Visualizing X-Y relationships - bubble plots



Visualizing X-Y relationships - scatterplot matrix



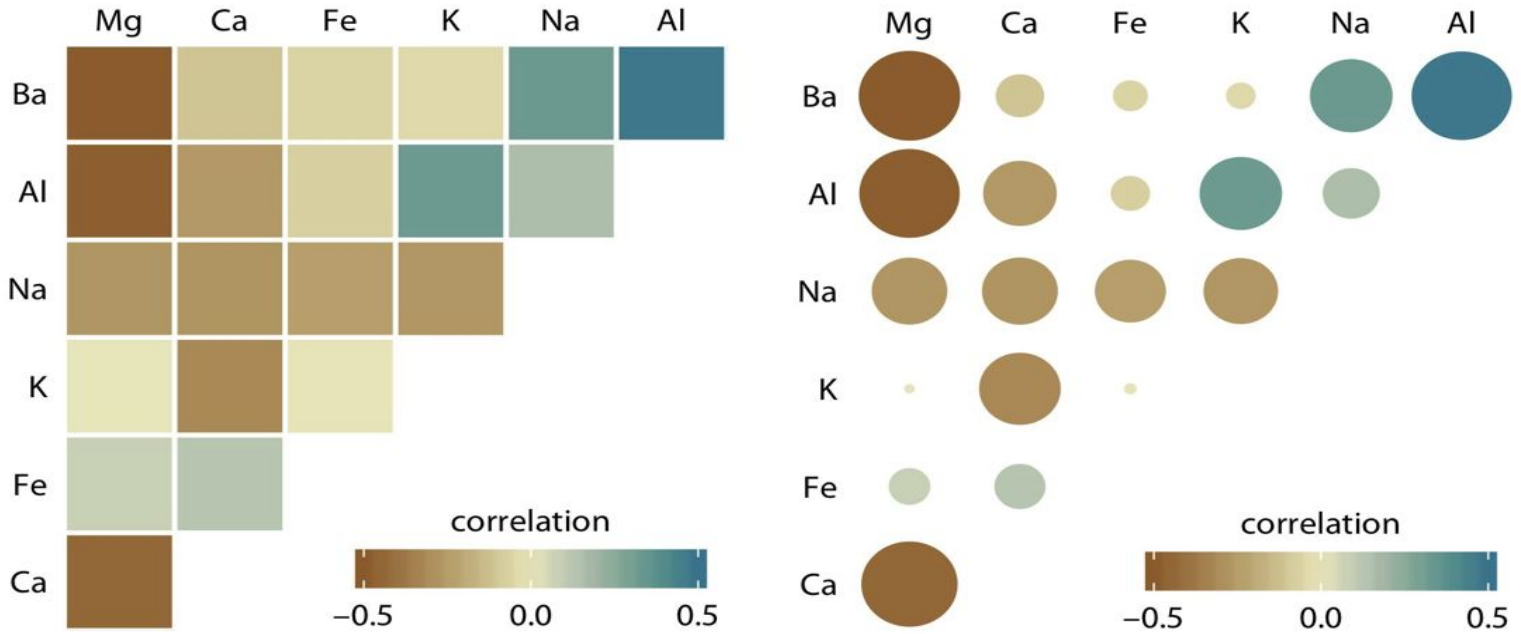
Visualizing X-Y relationships - correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Correlation coefficient

sample means

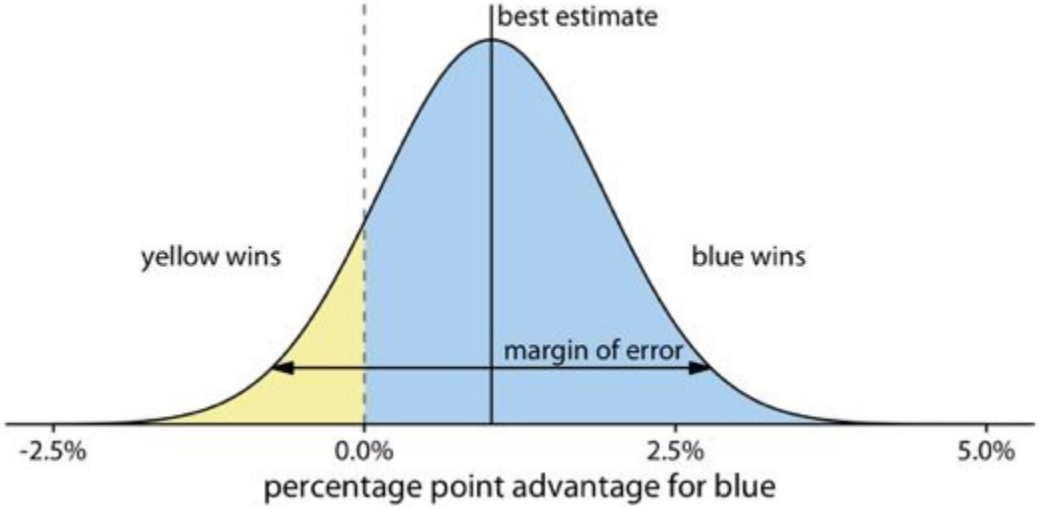
Visualizing X-Y relationships - correlograms

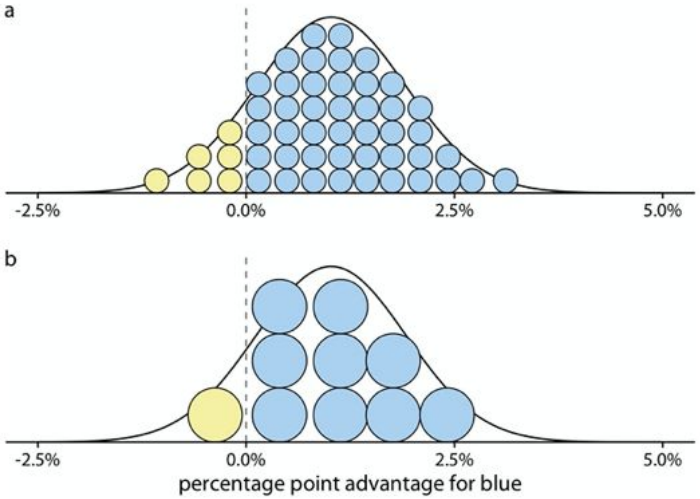


Correlations between mineral content obtained from 214 glass samples during forensic work

Visualizing uncertainty

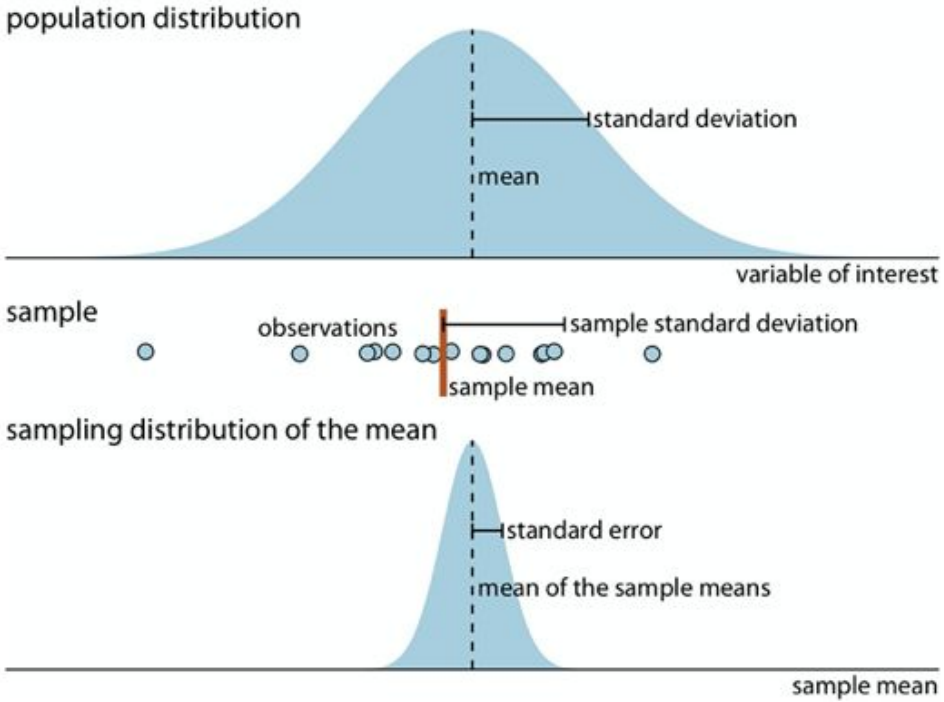
Visualizing uncertainty - probability distribution



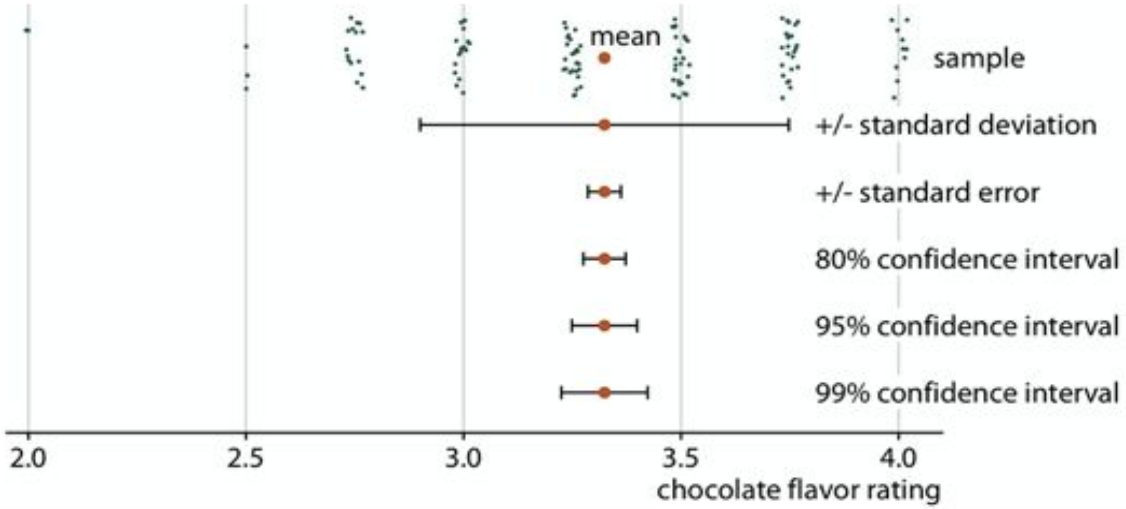


The blue party is predicted to win over the yellow party by ~1 percentage point with a margin of error of 1.76 percentage points.

Visualizing uncertainty - population & sample

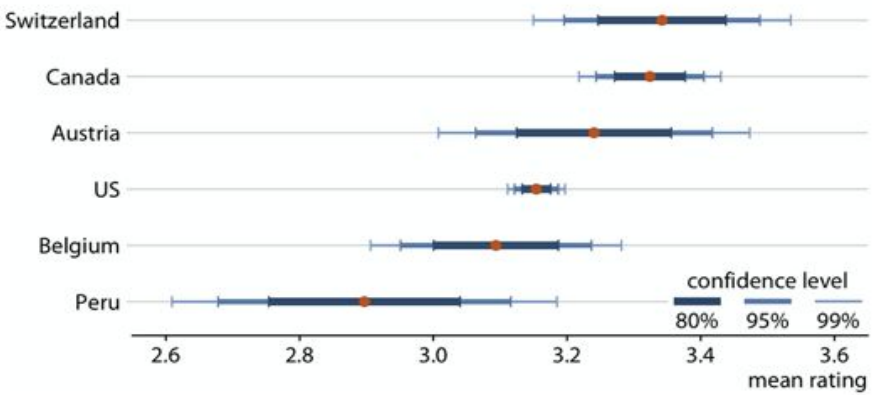


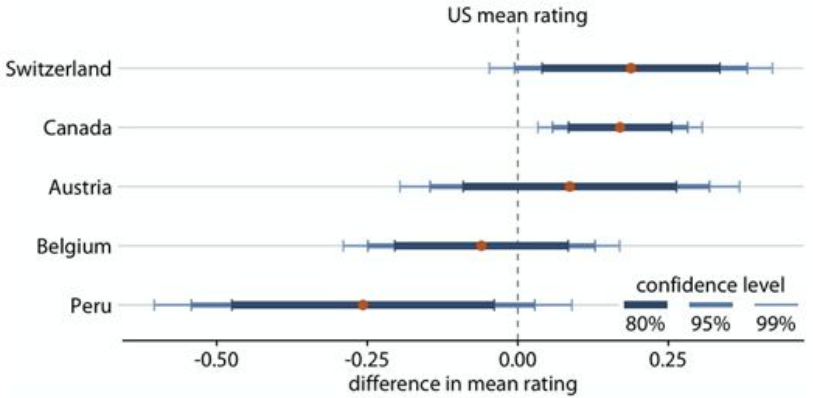
Visualizing uncertainty - confidence intervals



Ratings of Chocolate bars manufactured in Canada

Visualizing uncertainty - comparing parameter estimates





Questions?

In next lecture, we will cover...

- Supervised learning
- Data preprocessing