06.08.2024

Statistical Methods in AI (CS7.403)

Lecture-3: Basic Data Transformations, Data Visualization, Intro to Performance Measures, Benchmarking

Ravi Kiran (ravi.kiran@iiit.ac.in)

https://ravika.github.io

@vikataravi

Center for Visual Information Technology (CVIT)

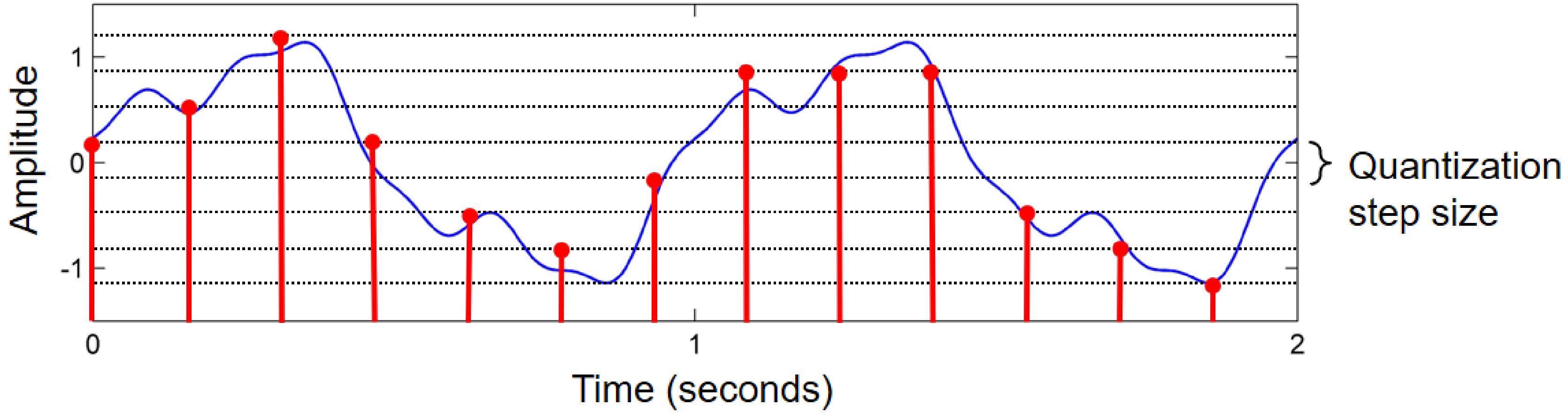
IIIT Hyderabad

Lecture Outline

- *ML Workflow (Previous Lecture)*
- *Data Representations (Previous Lecture)*
- Basic Data Transformations
- Data Visualization

Quantization

1. Continuous → Discrete ('Rounding off')

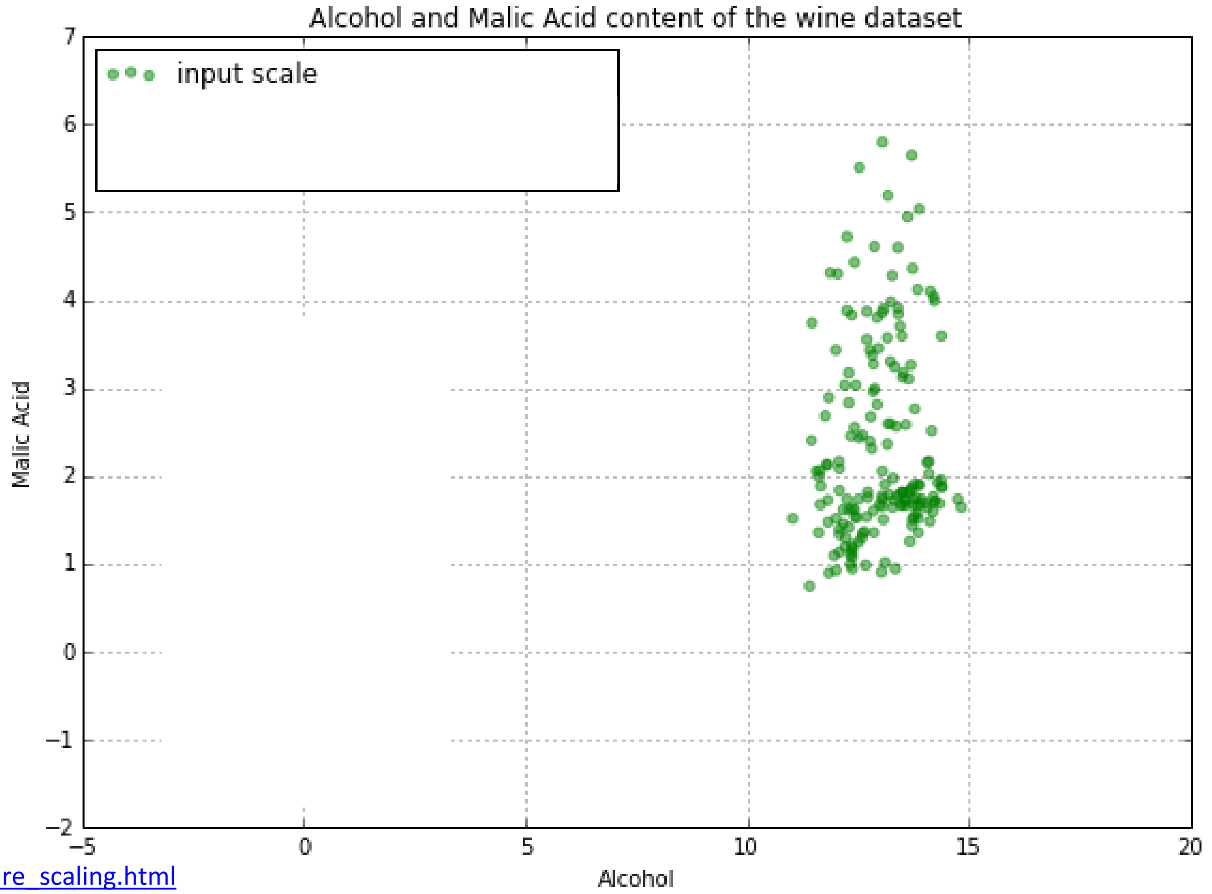


2. Binary Quantization ('Thresholding')

Data Normalization

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

-
-
-
-



https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Popular normalization approaches

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

MinMax Scaling

Standardization
(Unit Normal Scaling)

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

Data Normalization (applied to each feature)

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

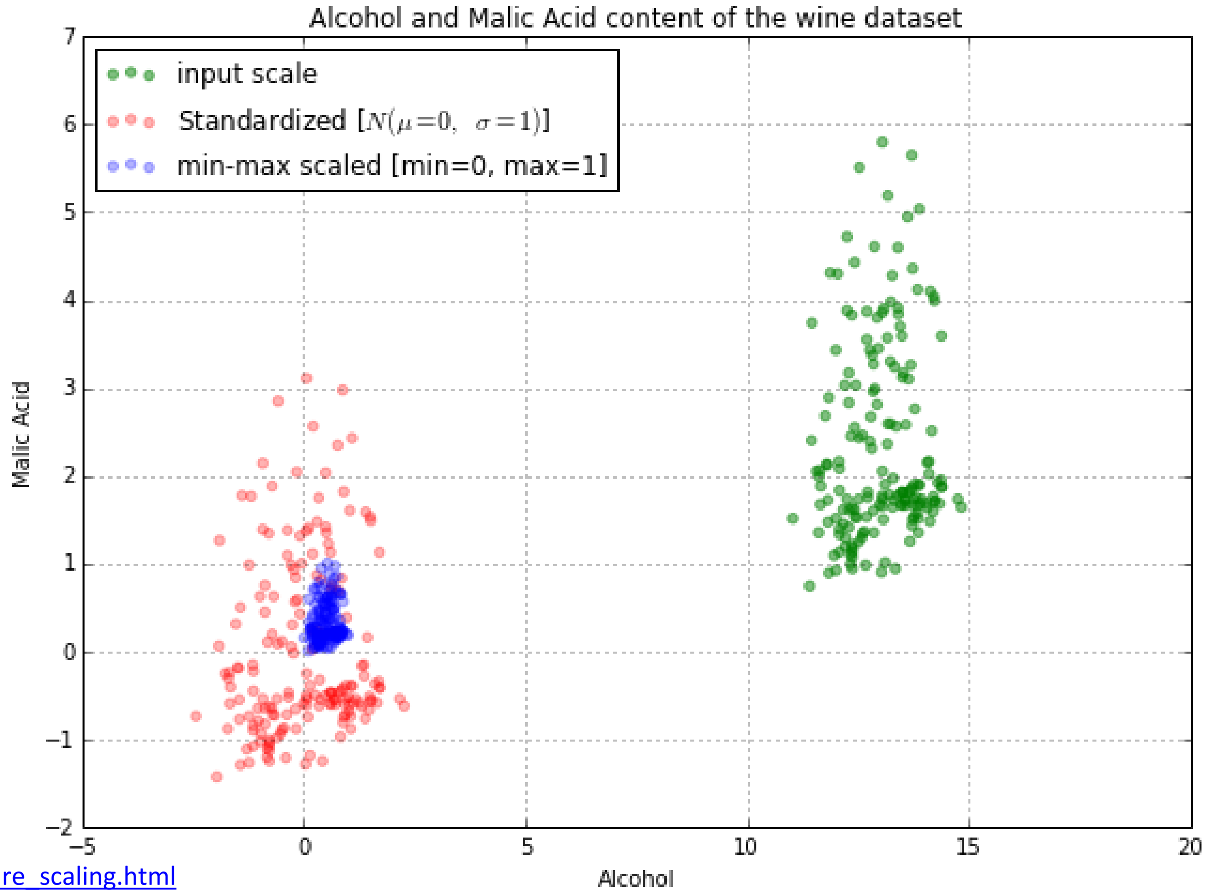
-
-
-
-

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

MinMax Scaling

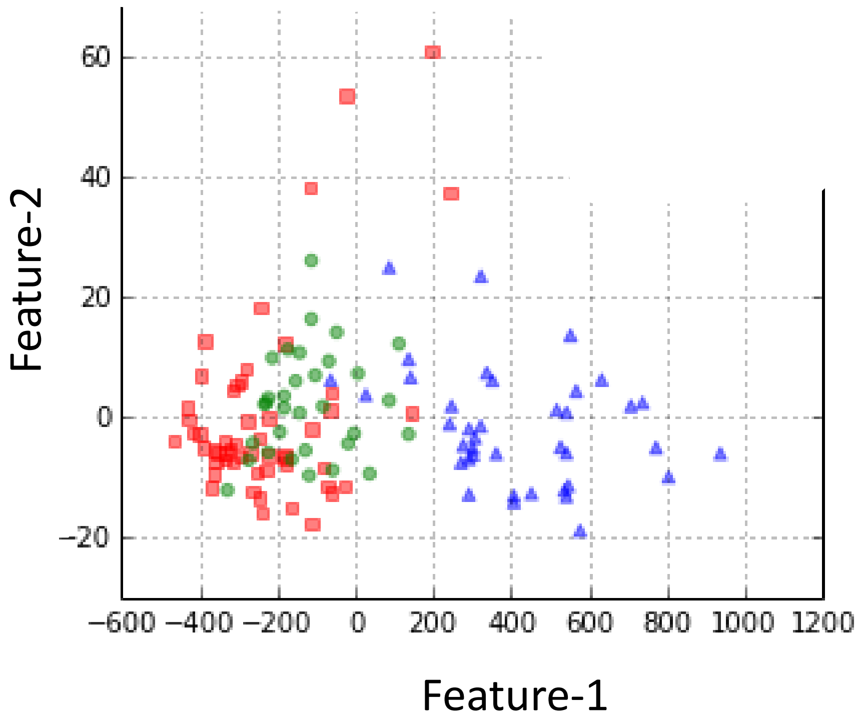
Standardization
(Unit Normal Scaling)

$$z = \frac{x - \mu}{\sigma}$$



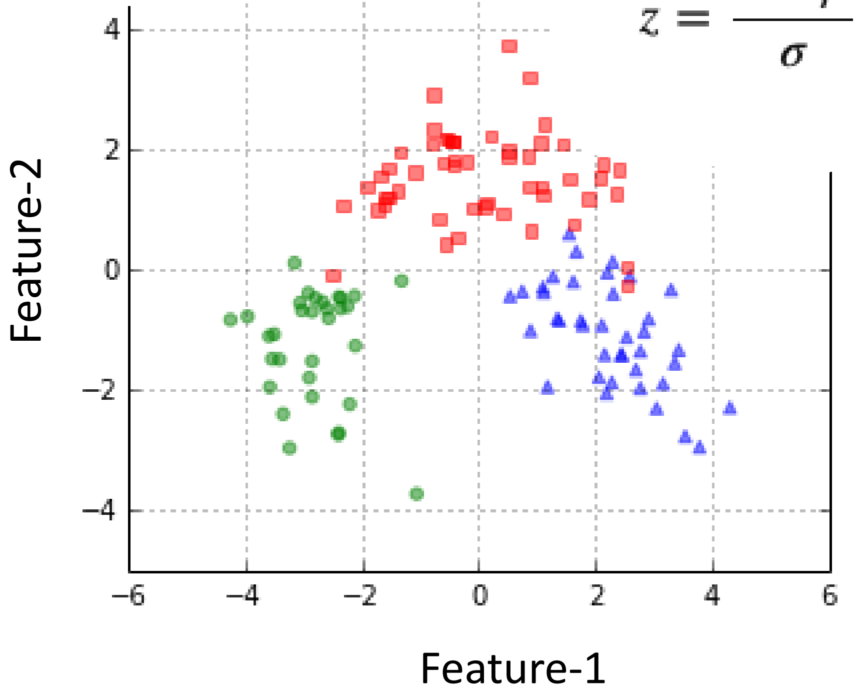
https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Before standardization



After standardization

$$z = \frac{x - \mu}{\sigma}$$



https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

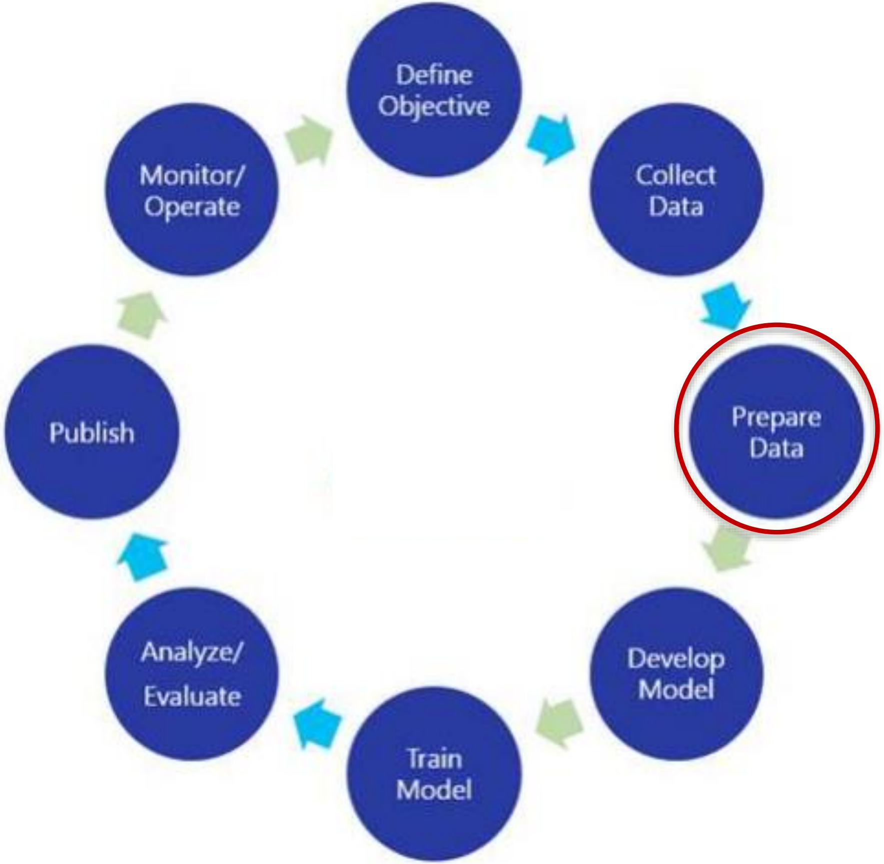
Why normalize data ?

- Uniform treatment of all features
- (Empirically) Helps stabilize optimization and lead to faster convergence
- Disadvantages?

Many more approaches exist …

- https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#plot-all-scaling-normalizer-section
- Non-linear: log(), exp(), sin(), pow(,)

Workflow of a Machine Learning Problem



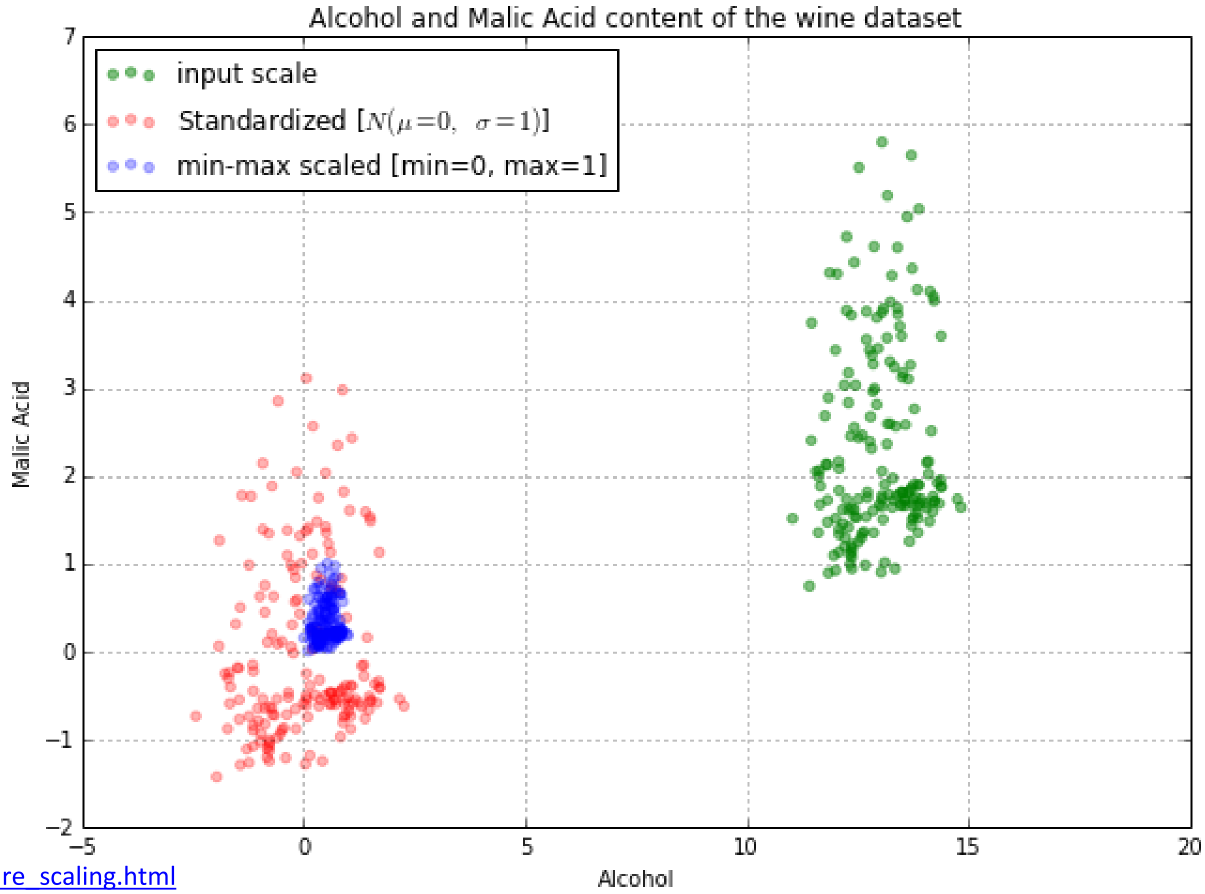
Lecture Outline

- *ML Workflow*
- *Data Representations*
- *Basic Data Transformations*
- Data Visualization

Data Normalization (applied to each feature)

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

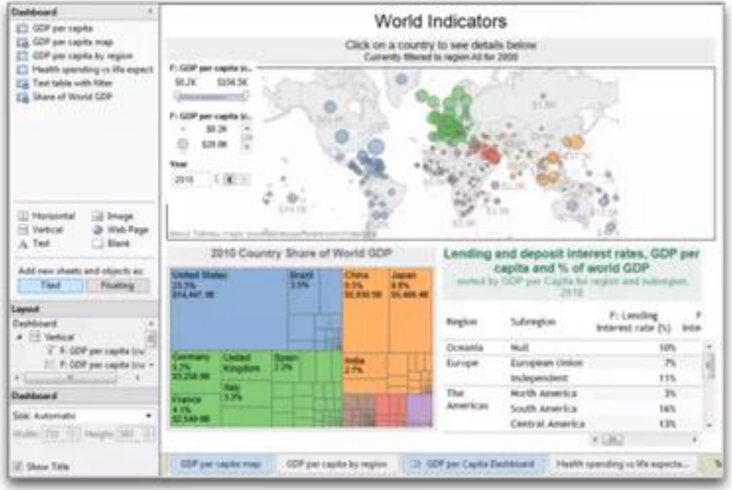
- $$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$
- MinMax Scaling
- Standardization (Unit Normal Scaling)
- $$z = \frac{x - \mu}{\sigma}$$



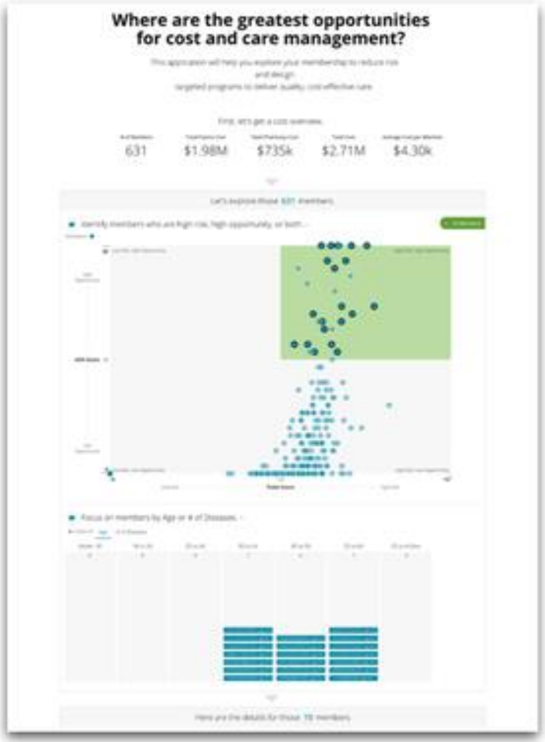
https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

Gazing at Data: Data visualization

data exploration



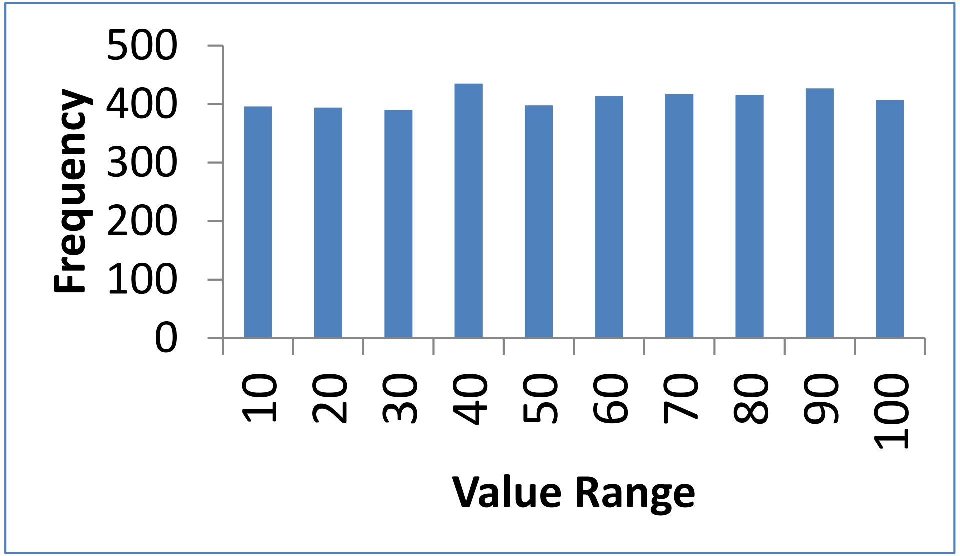
data presentation



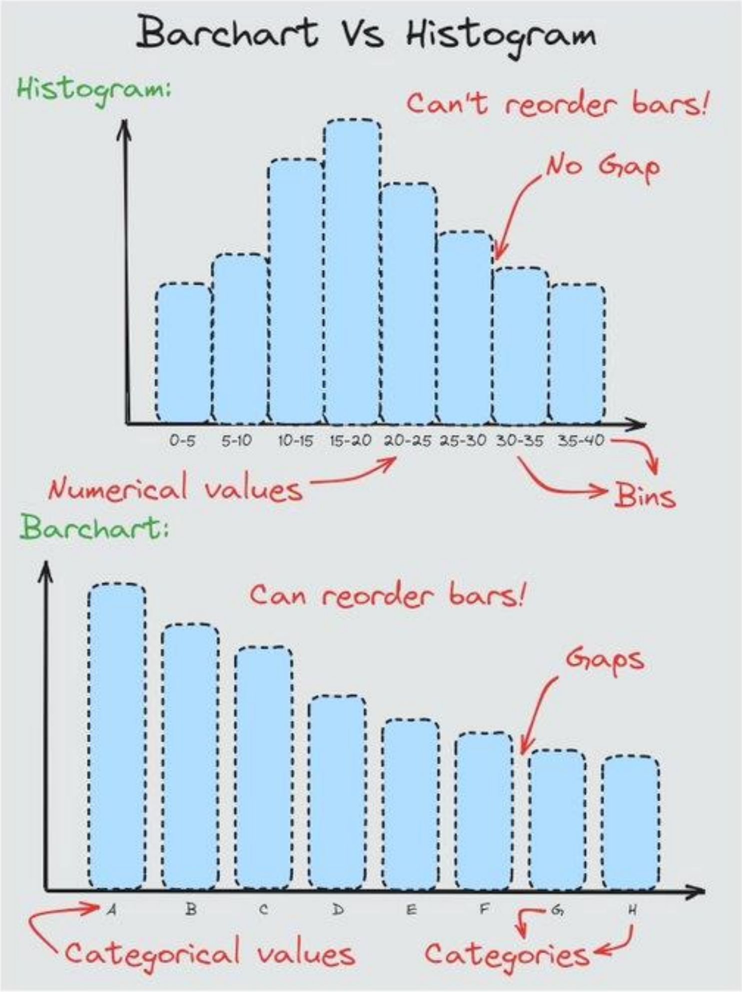
Why Plot Data?

22.65
23.03
12.07
47.68
86.20
20.02
48.38
55.23
65.30
75.71
69.73
77.95
58.90
6.45
32.39
74.35
82.87
47.05
34.26
12.54

42.12	67.24	59.13		81.49
53.42	40.54	89.97		21.85
93.43	51.93	49.30		43.76
51.91	13.12	73.88		60.29
41.28	66.24	62.15		46.87
92.09	26.50	83.53		70.99
46.21	10.85	29.61		62.15
84.90	15.37	35.00		83.23
26.56	5.78	72.59		12.47
93.15	3.67	49.80	…	43.05
53.77	82.80	43.59		32.35
14.94	63.71	9.30		1.31
42.53	62.74	99.91		53.17
46.29	67.34	32.65		23.94
57.39	10.61	54.07		53.28
60.10	2.25	77.55		12.05
17.02	80.73	29.60		9.96
97.01	19.84	76.59		45.90
86.80	19.11	4.80		1.24
30.40	67.94	55.53		58.25
		73.13		0.23



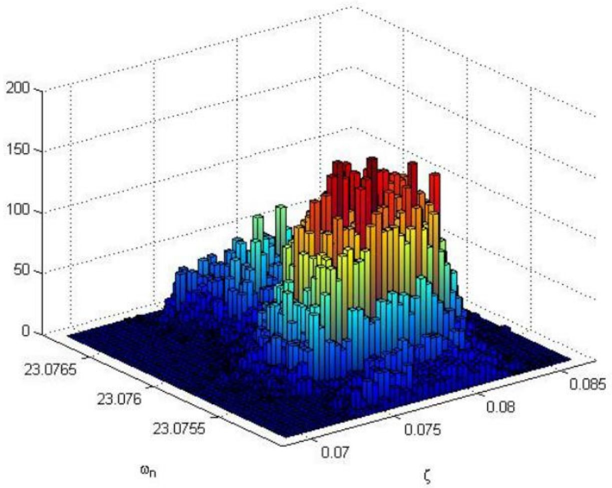
- Visualization of data provides specific insights into the nature of the data.
- Depending on the plot, we gain different insights

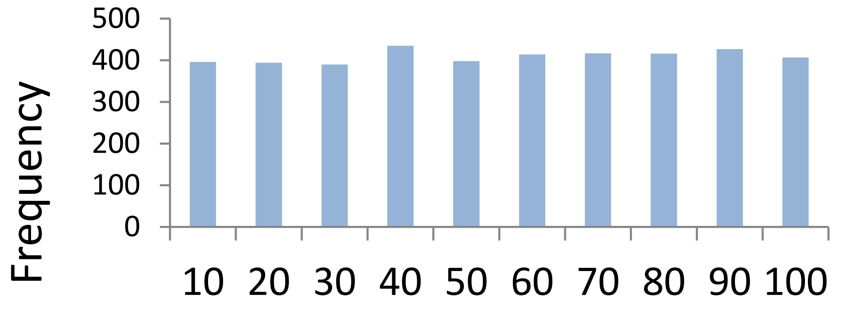


https://pbs.twimg.com/media/F1ej_z8X0AEiTvf?format=jpg&name=small

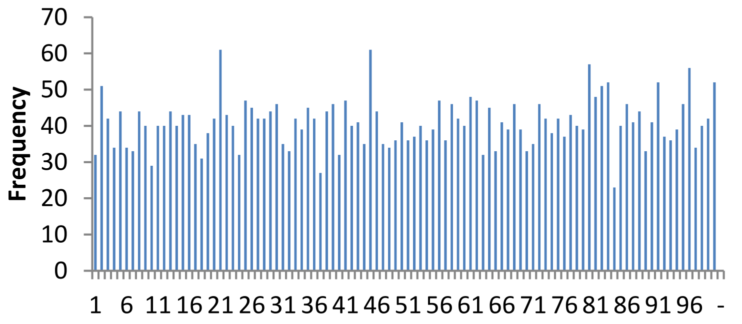
Histogram

- Count of items in each bin
  - Not a bar chart of Data
  - Approximation of Distribution
- Visualize one feature at a time
- Possible to extend to two
- Dependency on bins ($\sqrt{n}$; $2\sqrt[3]{n}$)





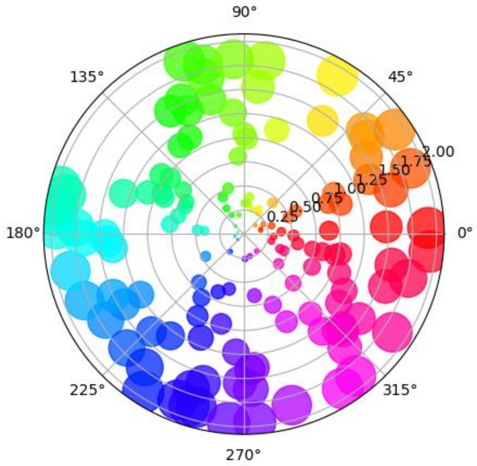
Bin Size = 10

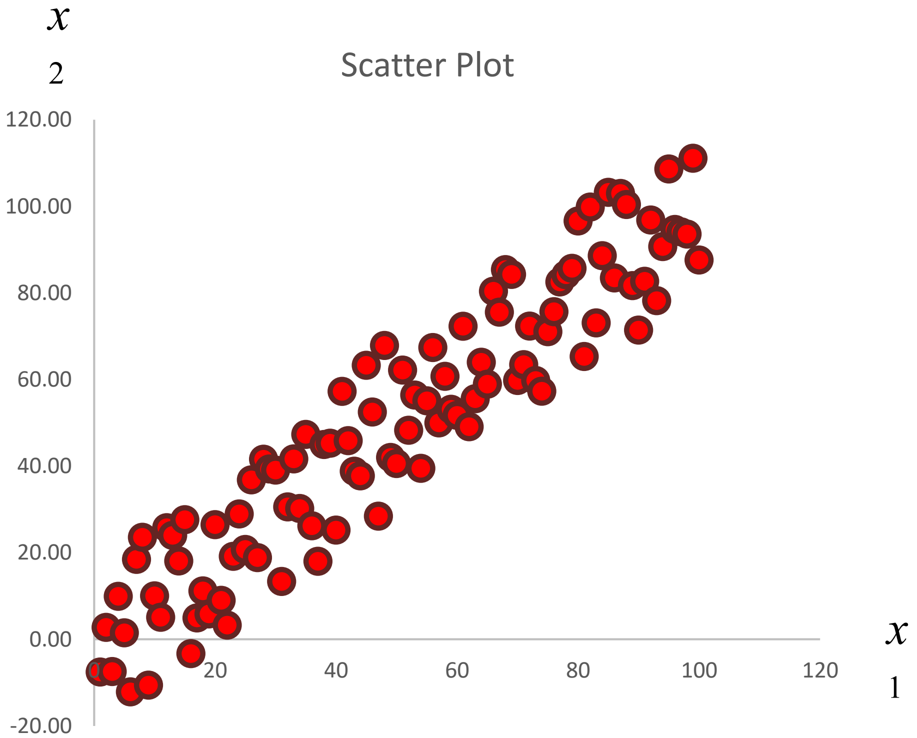


Bin Size = 1

Scatter Diagram

- Plots two features at a time
- Captures the correlation between the two
- Other formats possible

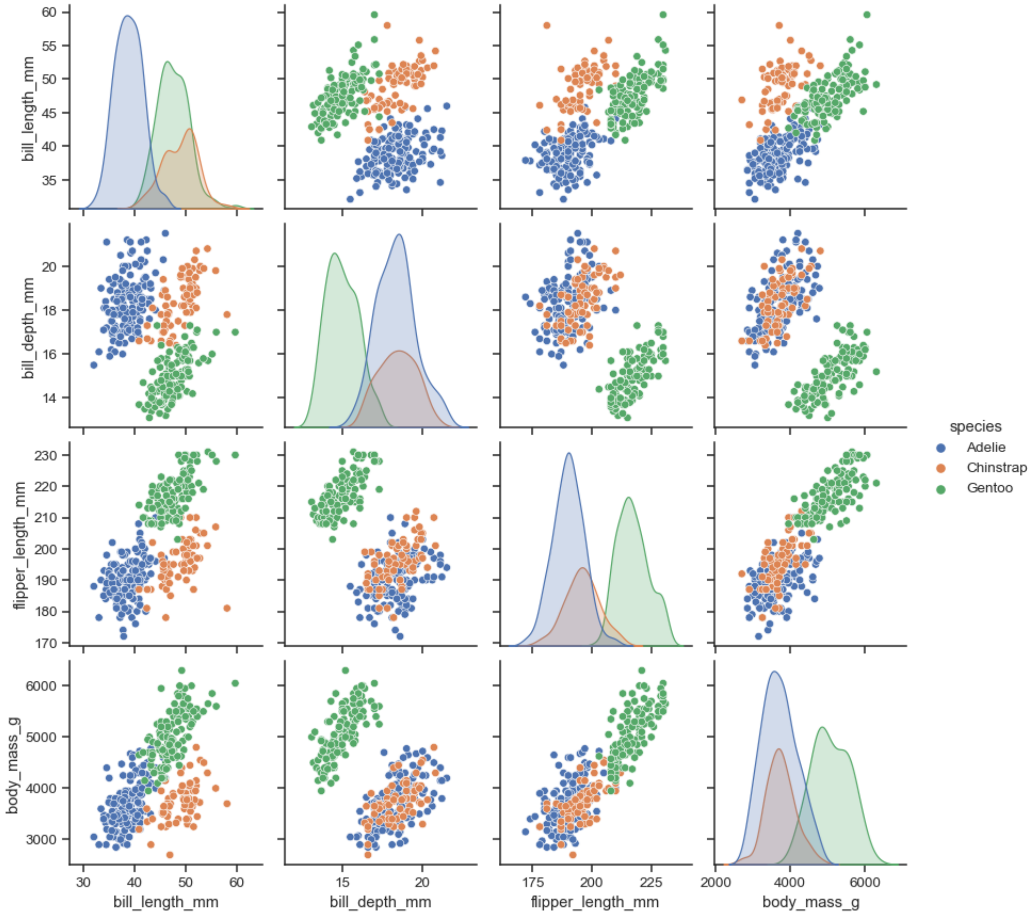




Polar Plot Courtesy: Scatterplot Documentation [matplotlib.org]

Pair Plot

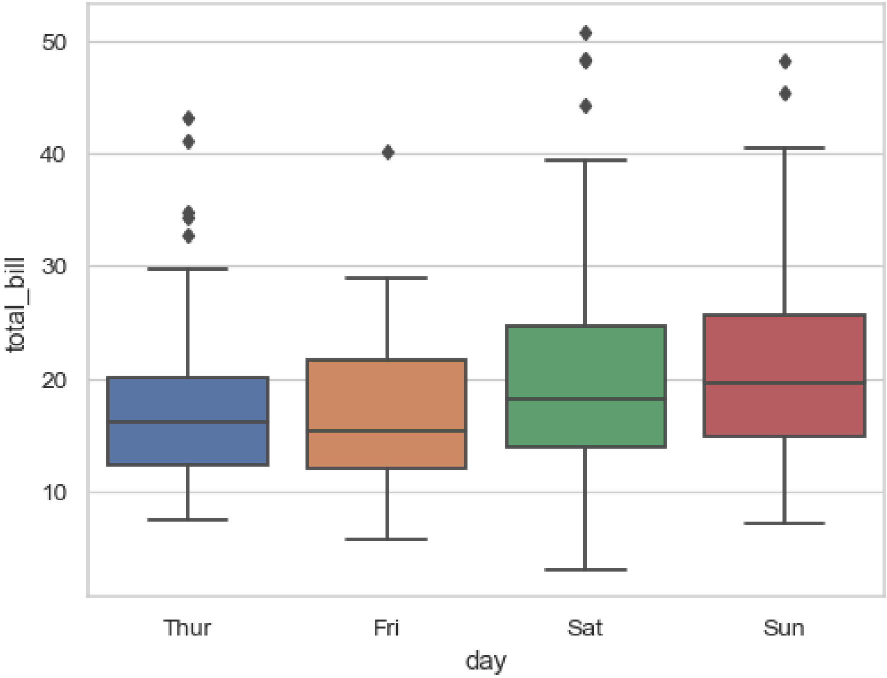
- Plot each pair of features as a matrix
  - Diagonal entries are histograms (densities)
  - Off-diagonal entries are scatter plots
- Can use other plots at each cell.



Plot Courtesy: Seaborn Documentation [seaborn.pydata.org]

Box Plot

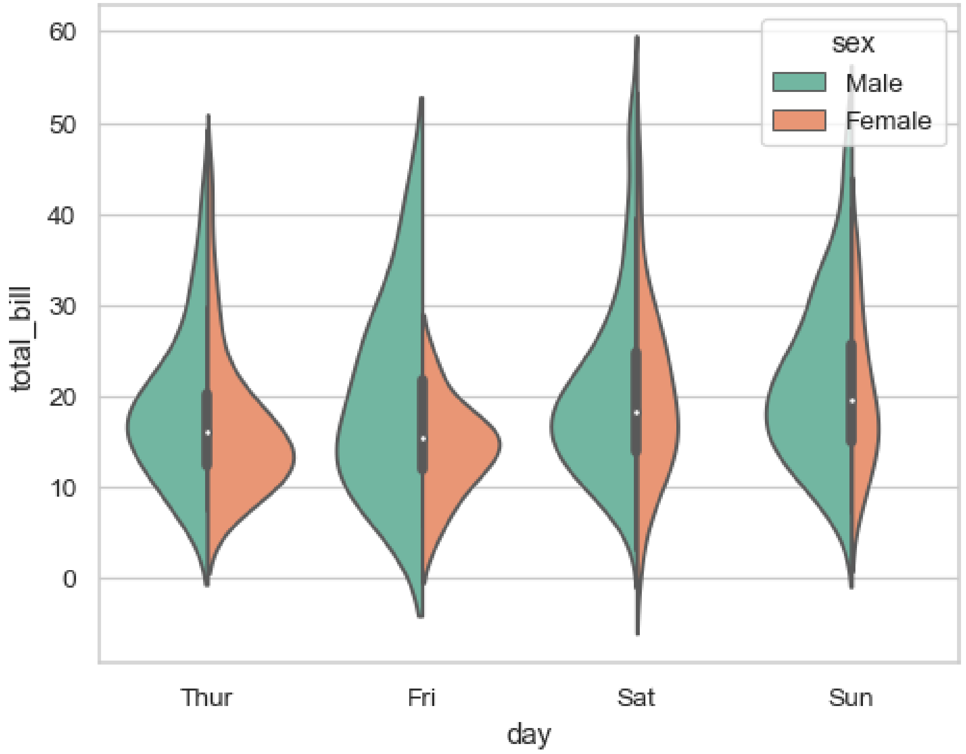
- Show median and quartiles of each feature
  - Outliers are removed
  - Box-and-whisker plot
- Whiskers can represent other percentiles/data
- Simpler than histograms of each feature



Plot Courtesy: Seaborn Documentation [seaborn.pydata.org]

Violin Plot

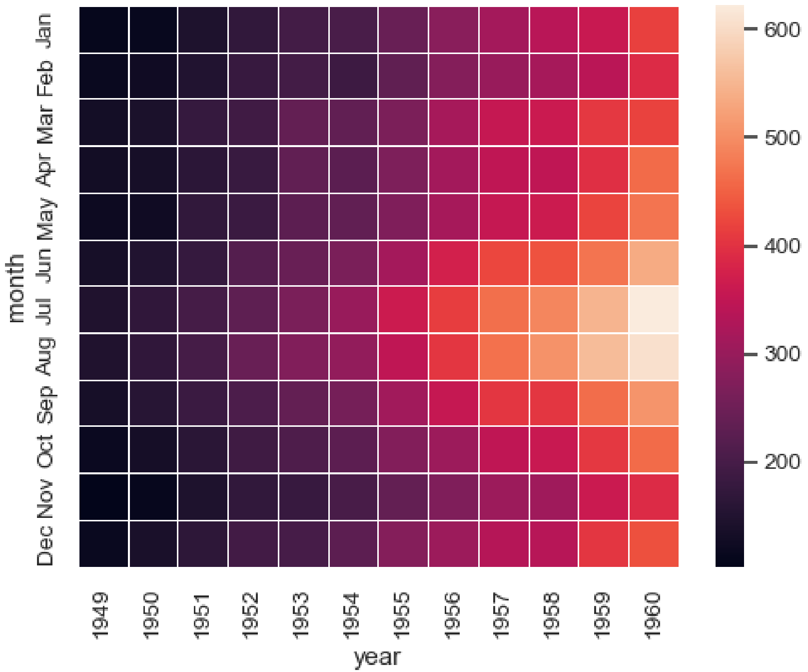
- Shows the density plot of each feature
  - Similar to Box Plot
- Either side can represent different densities
- Densities are smoothened estimates from data



Plot Courtesy: Seaborn Documentation [seaborn.pydata.org]

Heat Map

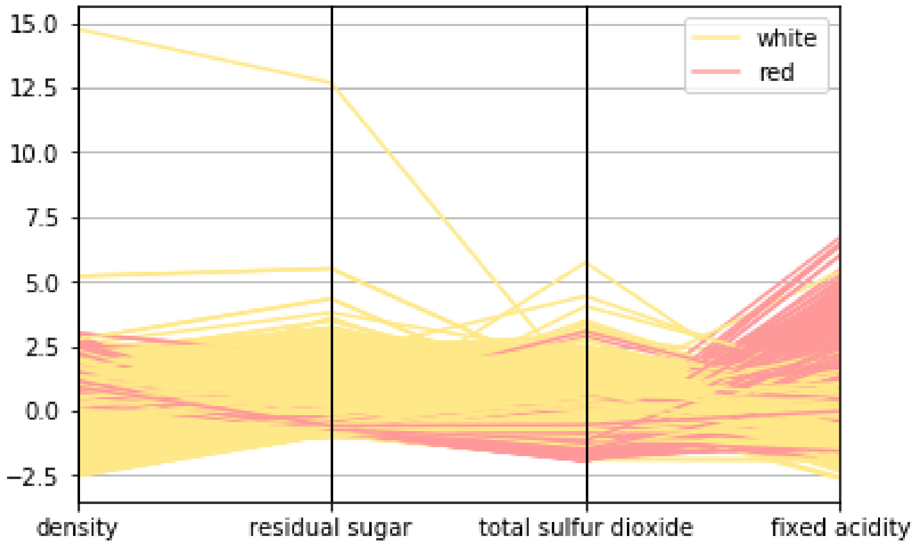
- A color-coded representation of 2D data
- Can be raw data, 2D histogram or any other function of 2 variables
- A color map accompanies the heat map
- We will learn other metrics in future that may be visualized as a heat map



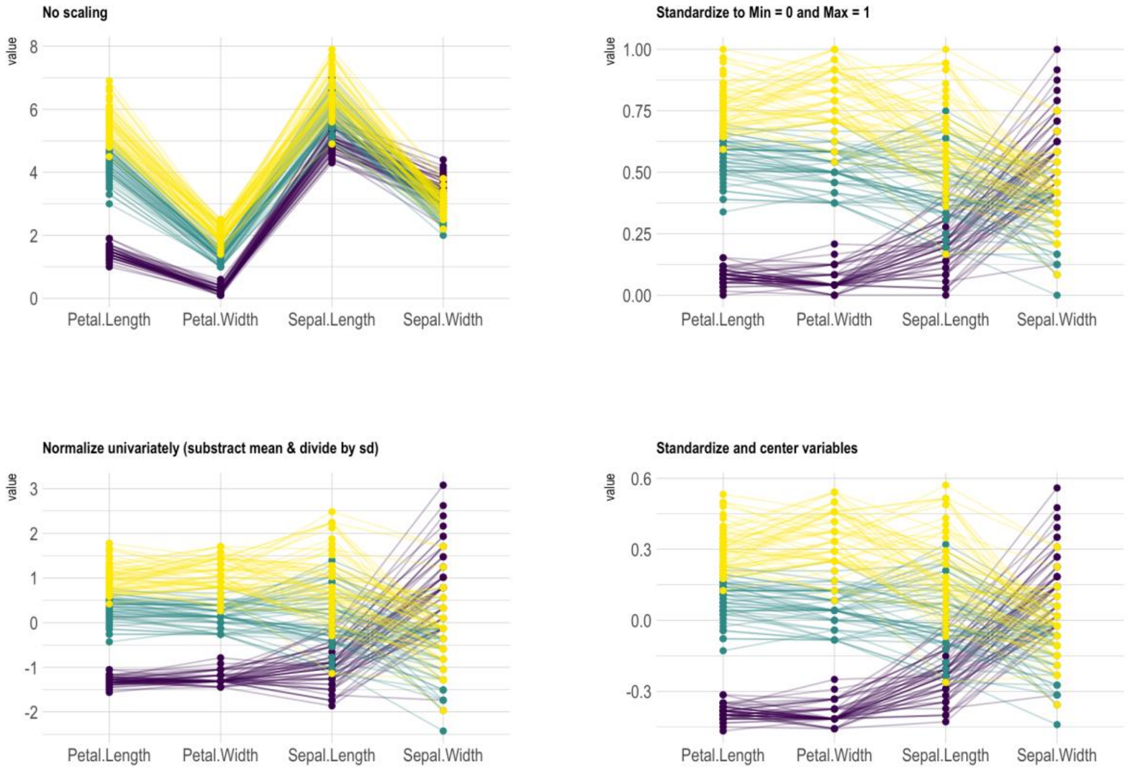
Plot Courtesy: Seaborn Documentation [seaborn.pydata.org]

Direct Visualization

- Parallel Co-ordinates
  - Each vertical line is a dimension
  - A data item is connected by line segments
  - Large number of samples clutters the visualization



Parallel Coordinates



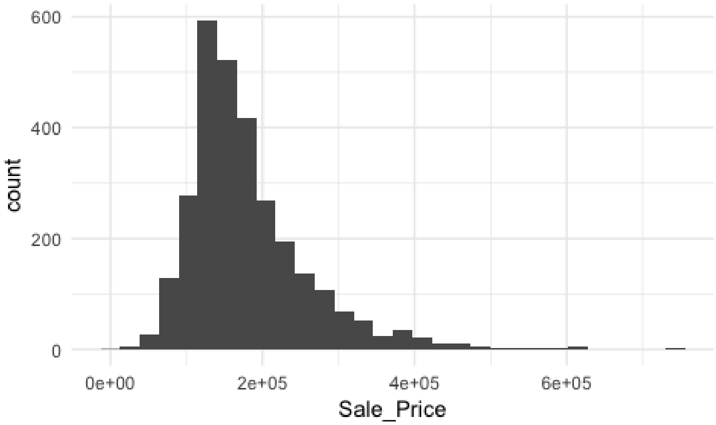
https://www.data-to-viz.com/graph/parallel.html

Leaderboard

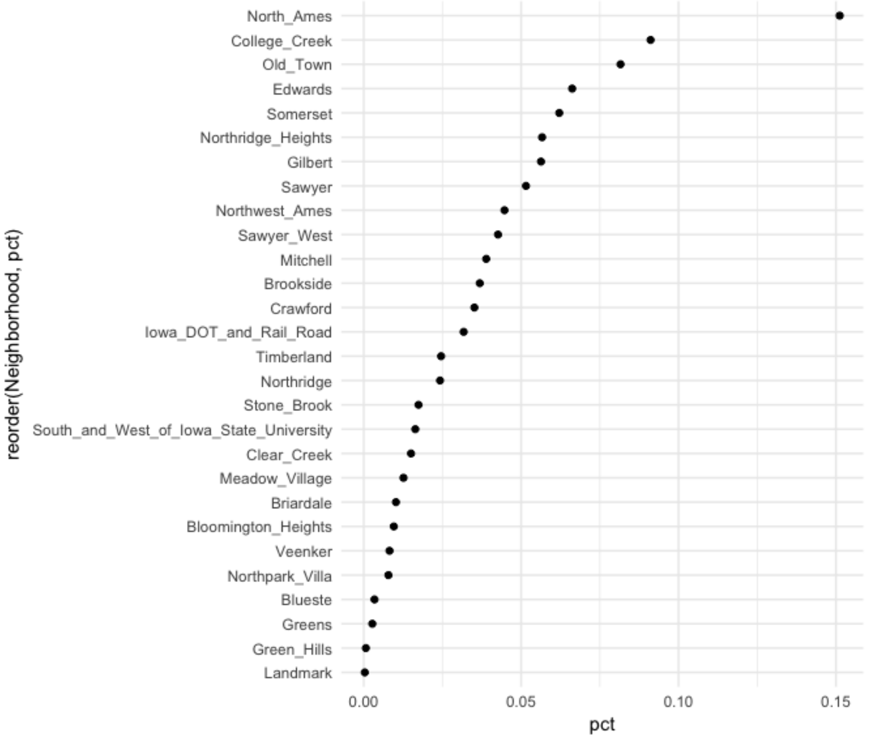
Proposals		Proposal Dollars		Awards		Award Dollars	
1st Computer Science	10	Recreation & Tourism	$9.18M	1st Biology	20	1st Computer Scien	$8.92M
2nd Biology	9	Health Sciences	$8.95M	2nd Computer Science	17	2nd Biology	$8.84M
2nd Electrical & Comp Engr	9	3rd Computer Scien	$8.92M	Health Sciences	12	3rd Electrical & Com	$4.56M
4th Chemistry and Biochemi	8	4th Biology	$8.84M	3rd Electrical & Comp Eng	12	4th Chemistry and B	$3.67M
Journalism	6	Psychology	$6.53M	5th Chemistry and Biochem	10	Health Sciences	$3.55M
Mathematics	4	Undergraduate Studi	$5.18M	Secondary Education	7	Recreation & Tourism	$3.06M
Deaf Studies	4	7th Chemistry and B	$3.67M	Elementary Education	7	Secondary Education	$2.85M
Elementary Education	4	8th Electrical & Com	$3.65M	Journalism	5	Elementary Educatio	$2.67M
Art	4	Secondary Education	$2.85M	Recreation & Tourism M	5	Communication Stud	$2.39M
Health Sciences	3	Elementary Educatio	$2.67M	Mechanical Engineering	5	Psychology	$2.17M

Top 10

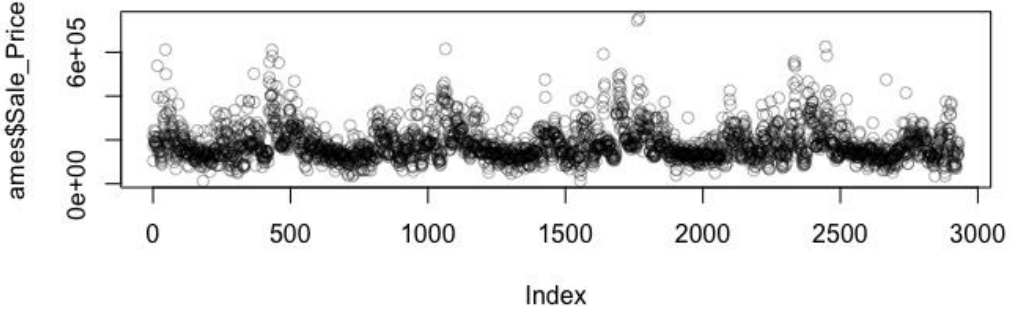
https://www.juiceanalytics.com/writing/writing/parallel-coordinates



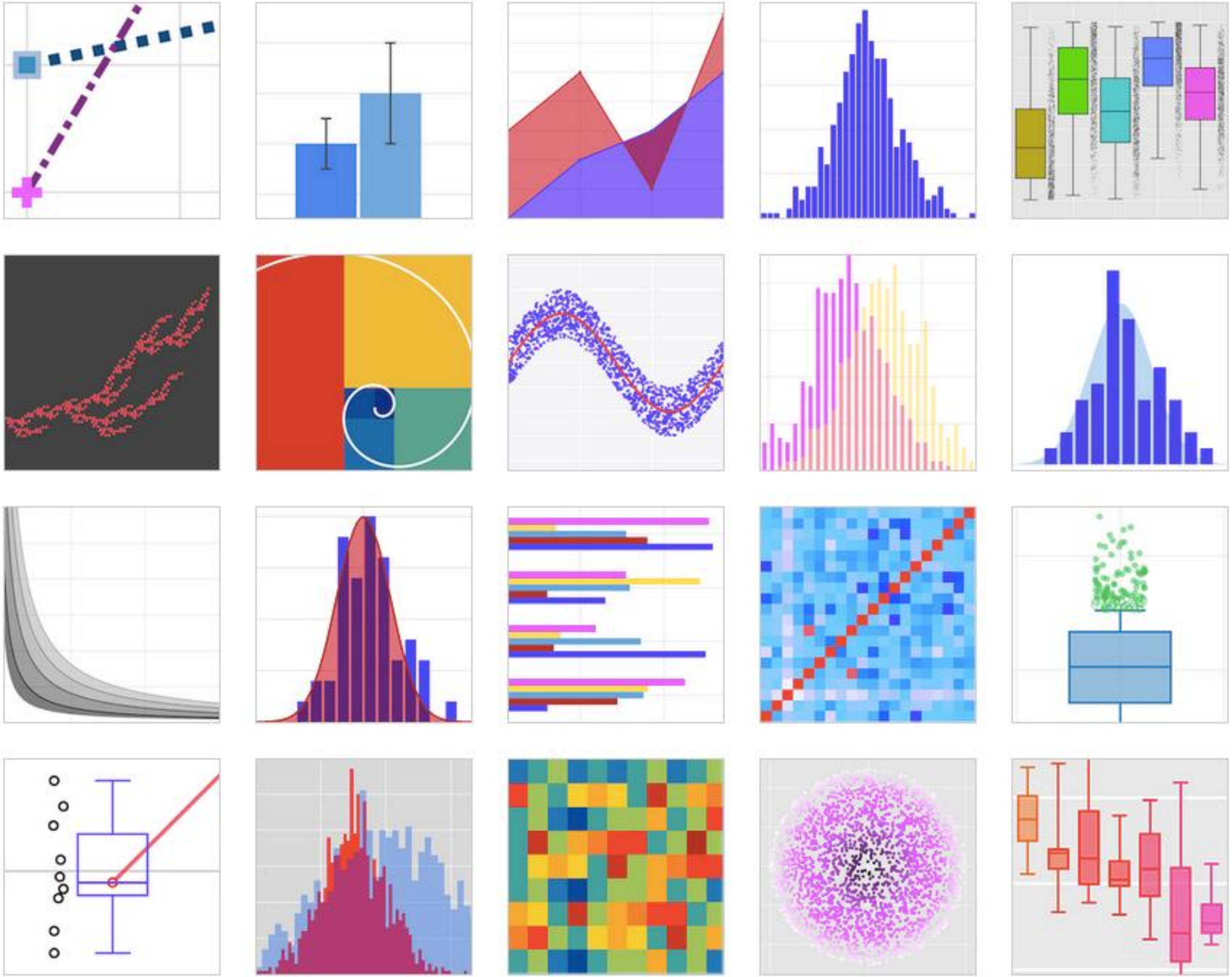
600
400
200
0
count
0e+00
2e+05
4e+05
6e+05
Sale_Price



North_Ames
College_Creek
Old_Town
Edwards
Somerset
Northridge_Heights
Gilbert
Sawyer
Northwest_Ames
Sawyer_West
Mitchell
Brookside
Crawford
Iowa_DOT_and_Rail_Road
Timberland
Northridge
Stone_Brook
South_and_West_of_Iowa_State_University
Clear_Creek
Meadow_Village
Briardale
Bloomington_Heights
Veenker
Northpark_Villa
Blueste
Greens
Green_Hills
Landmark
reorder(Neighborhood, pct)
0.00
0.05
0.10
0.15
pct



ames$Sale_Price
0e+00
6e+05
0
500
1000
1500
2000
2500
3000
Index



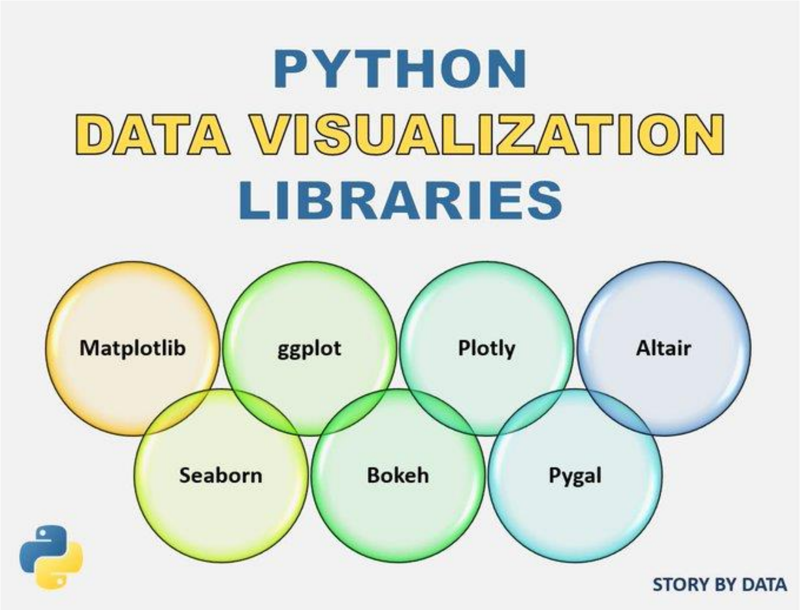
"In good information visualization, there are no rules, no guidelines, no templates, no standard technologies, no stylebooks ... You must simply do whatever it takes."

—Edward Tufte

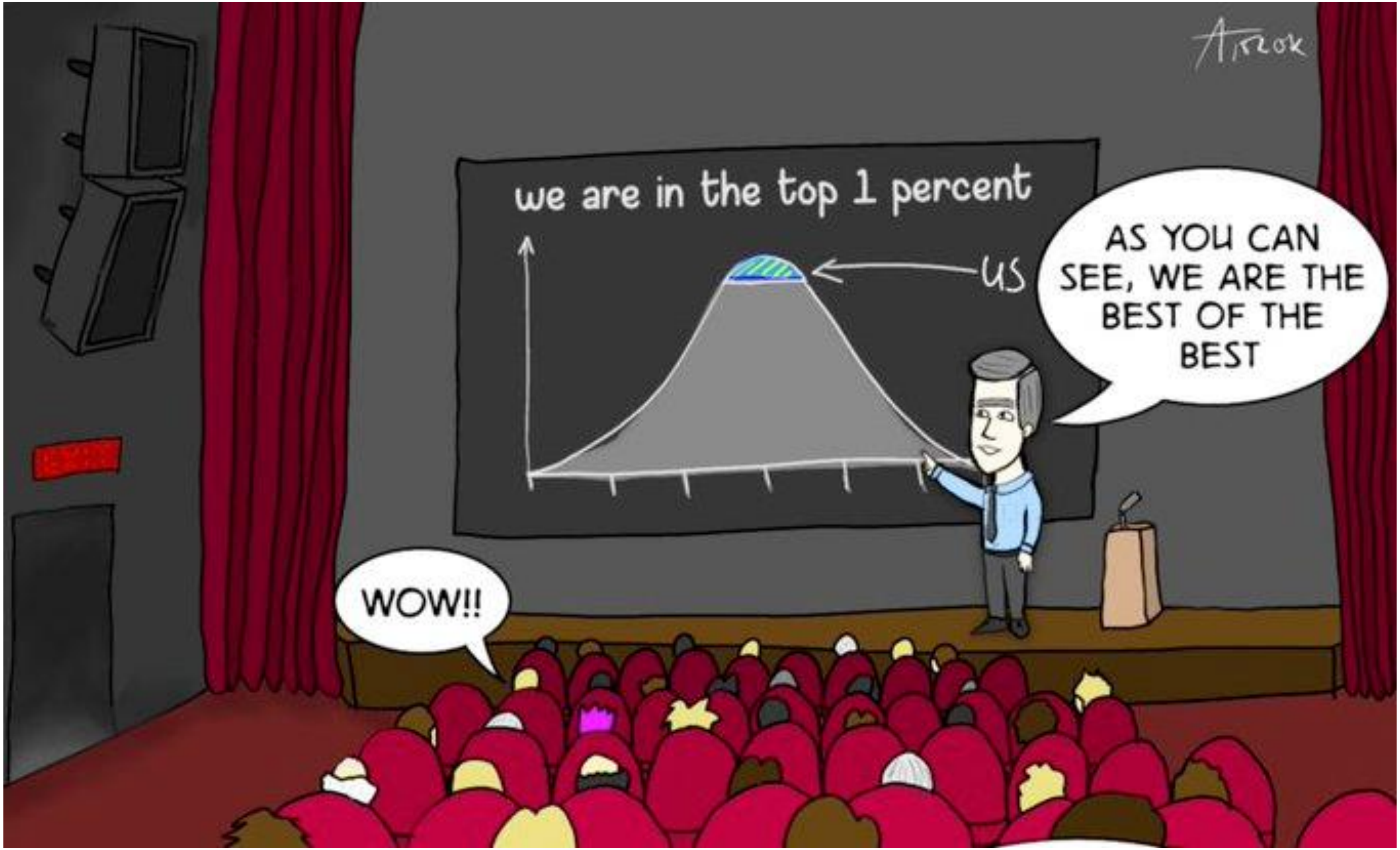
VISAGE.CO

Resources

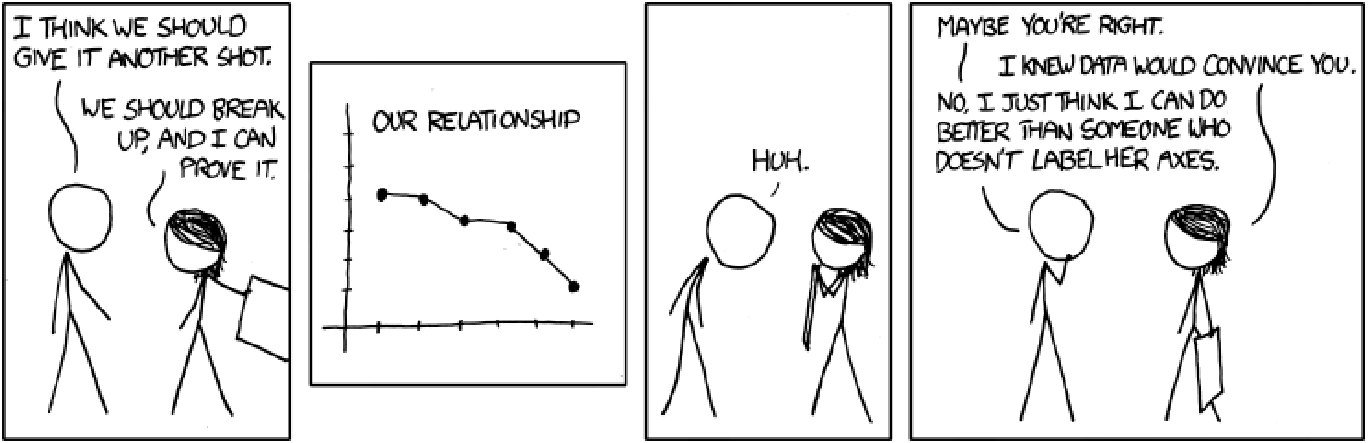
- https://towardsdatascience.com/5-quick-and-easy-data-visualizations-in-python-with-code-a2284bae952f

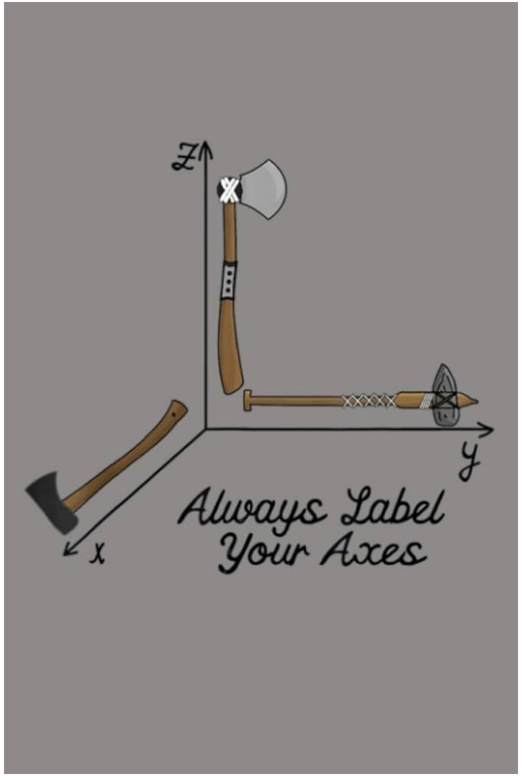


https://twitter.com/storybydata/status/1166337648341991424

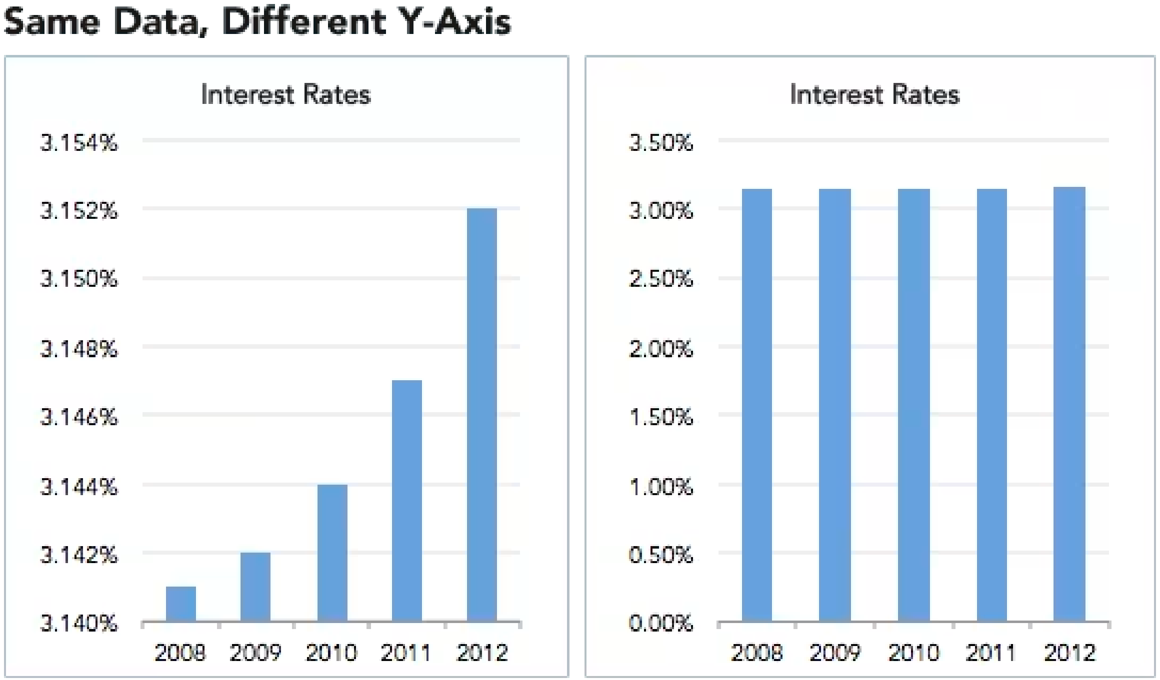


https://pbs.twimg.com/media/F23DL9KWwAAIE0-


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

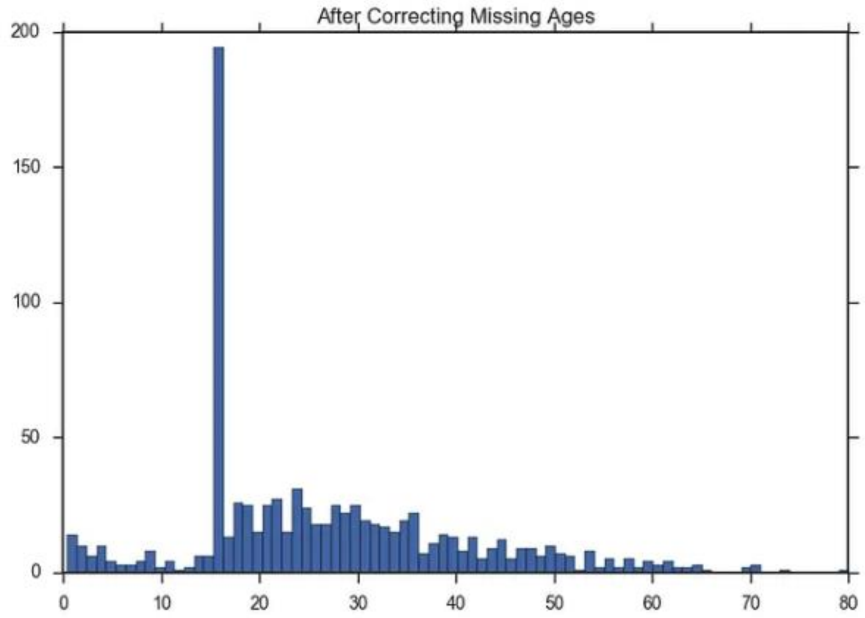

z
y
x
Always Label Your Axes

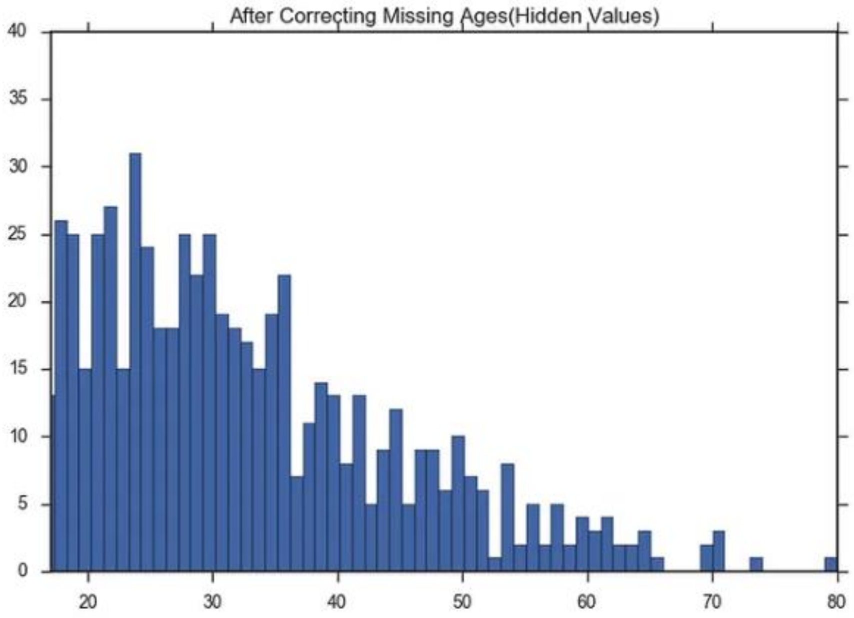
Lies that visualizations tell and how to spot them



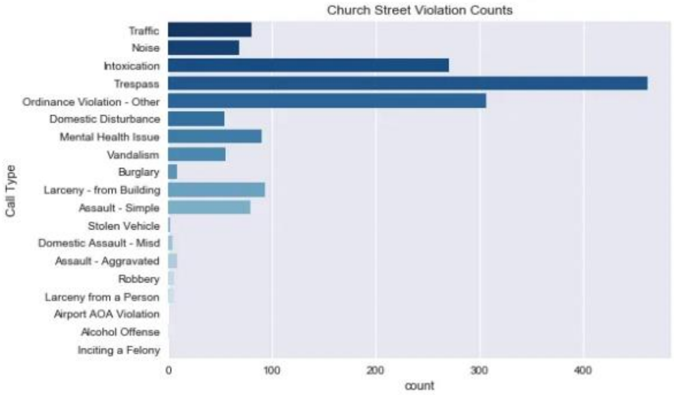
https://www.heap.io/blog/how-to-lie-with-data-visualization

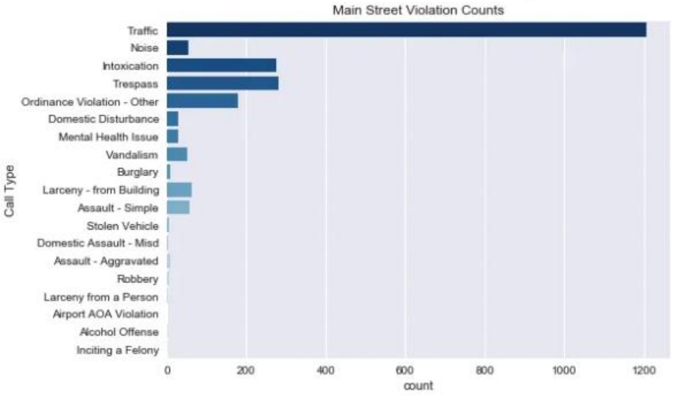
Lies that visualizations tell and how to spot them

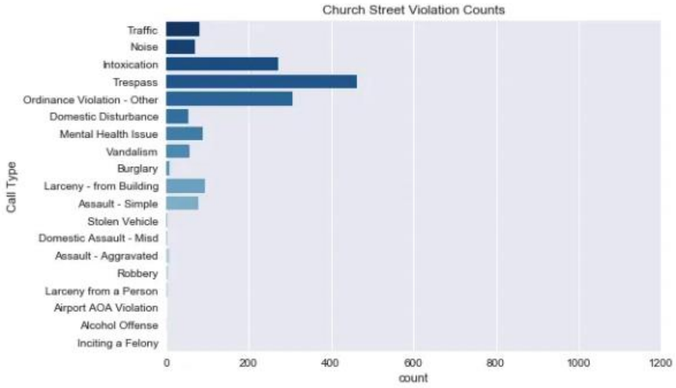


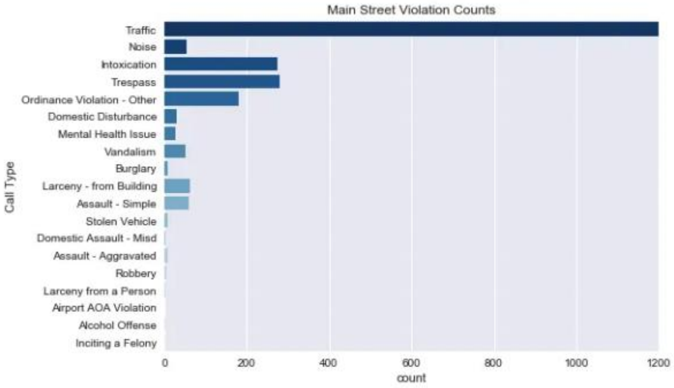


Lies that visualizations tell and how to spot them

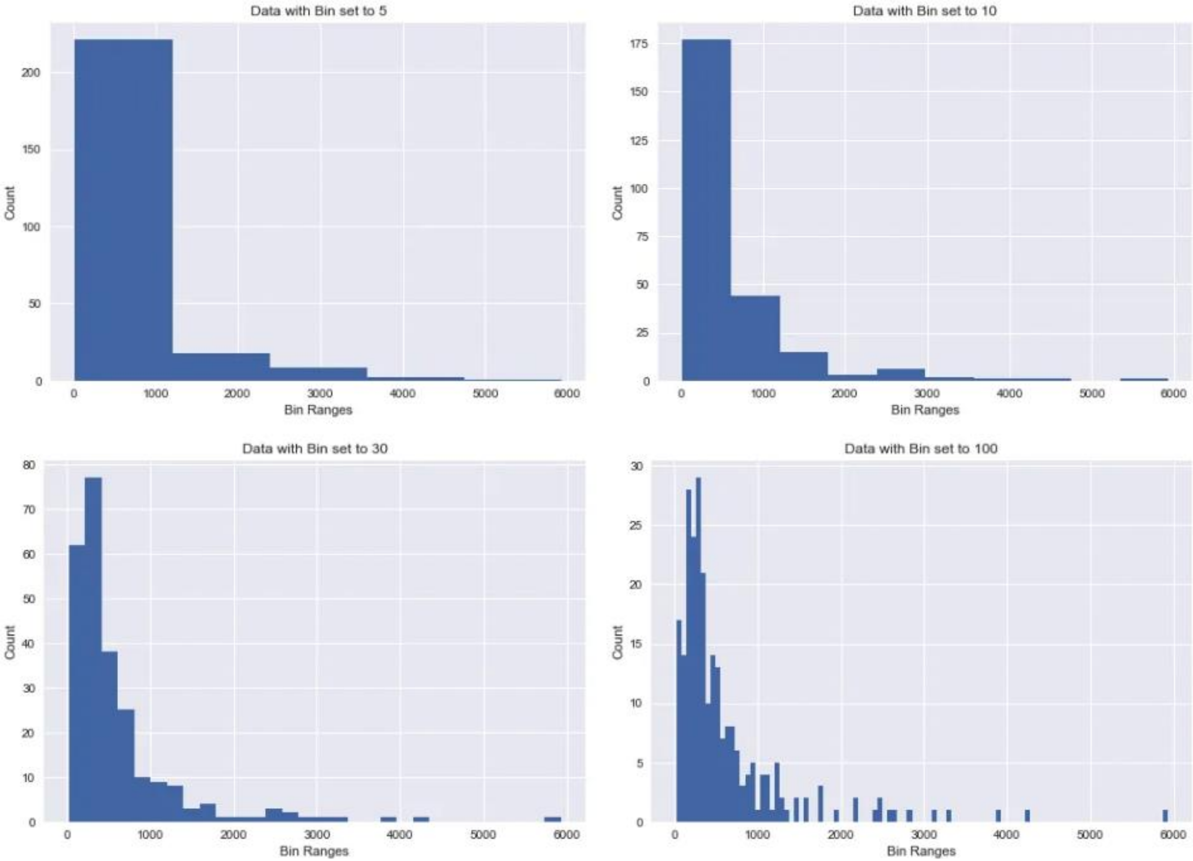




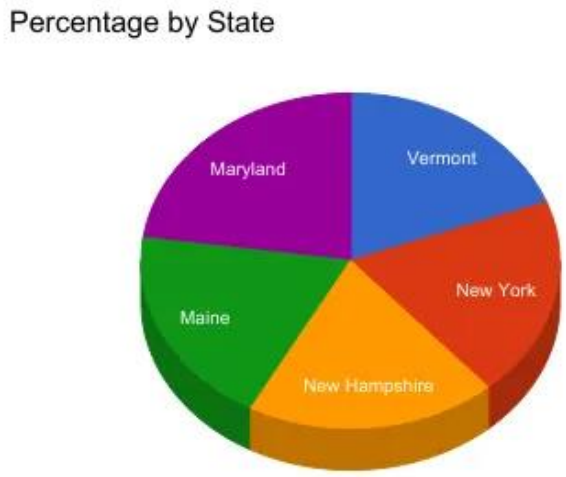




Lies that visualizations tell and how to spot them

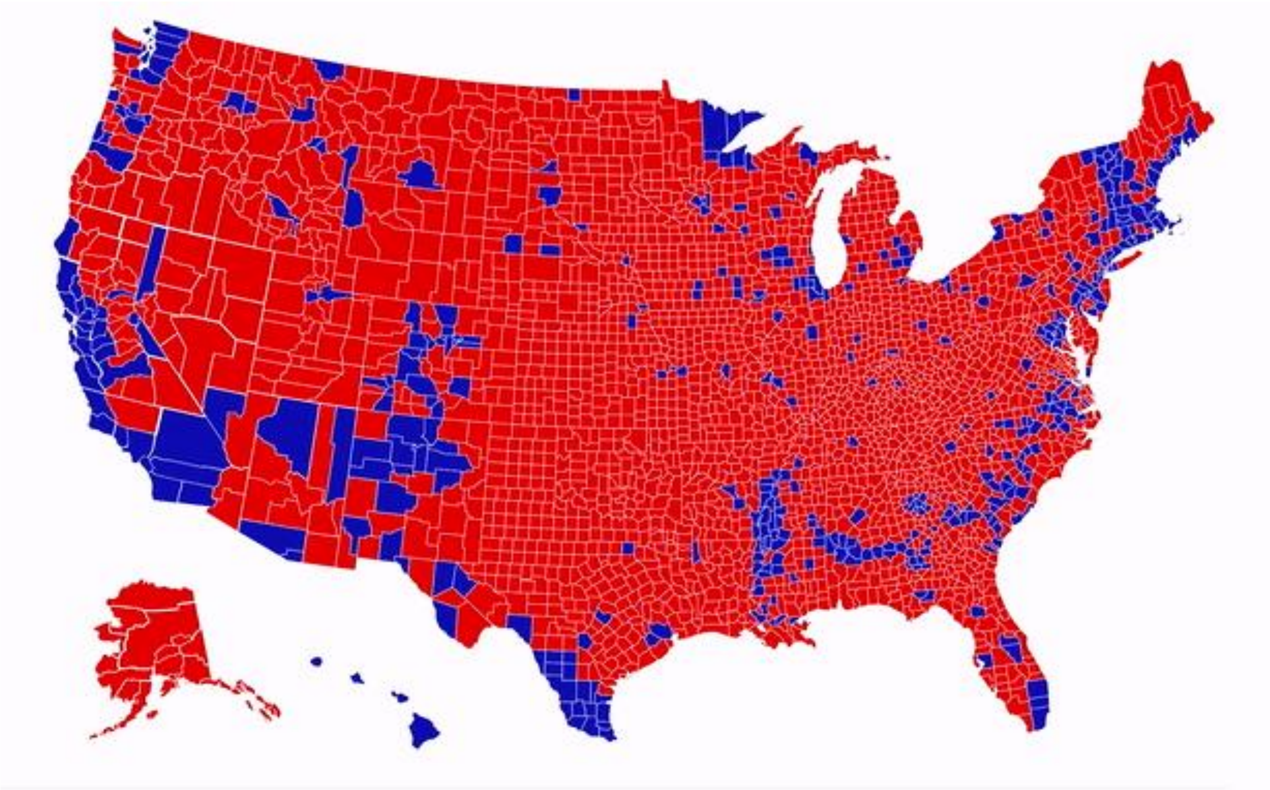


Lies that visualizations tell and how to spot them

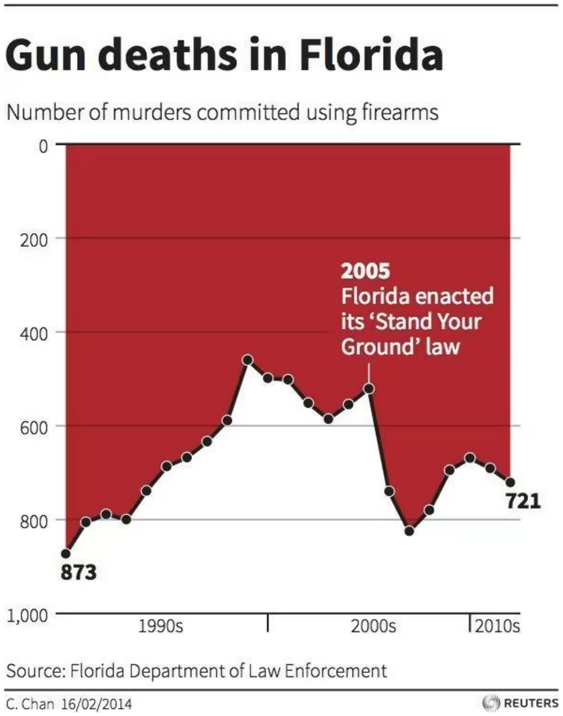


https://www.heap.io/blog/how-to-lie-with-data-visualization

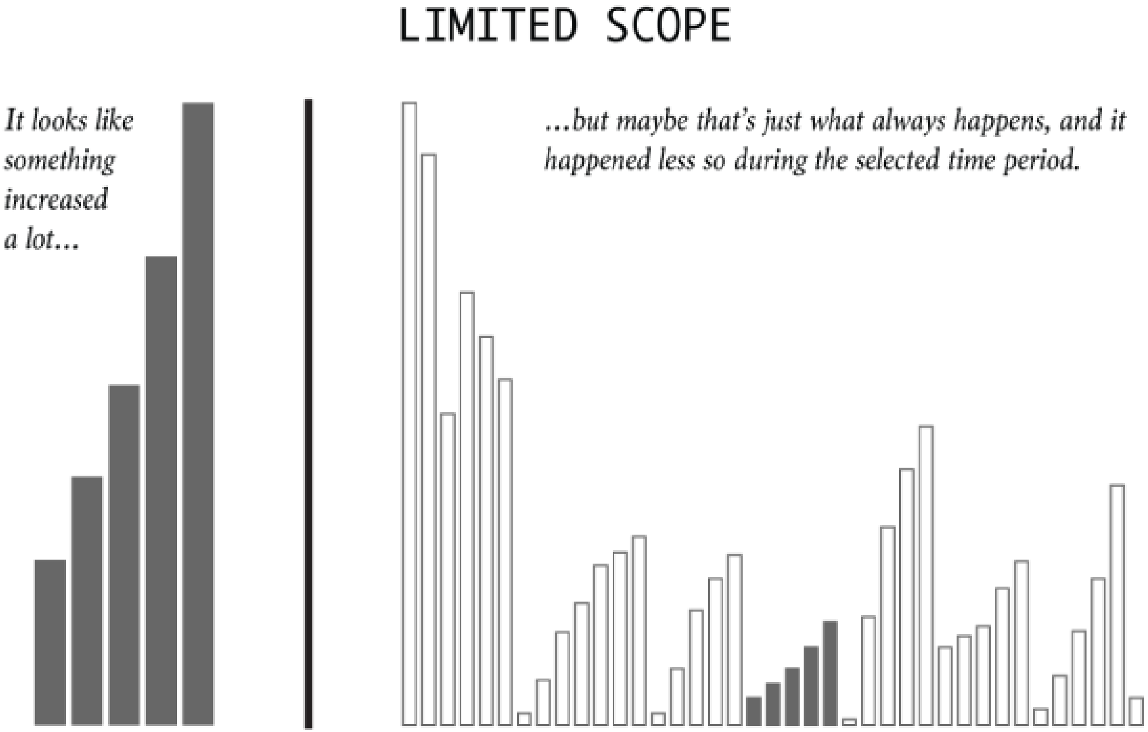
Lies that visualizations tell and how to spot them



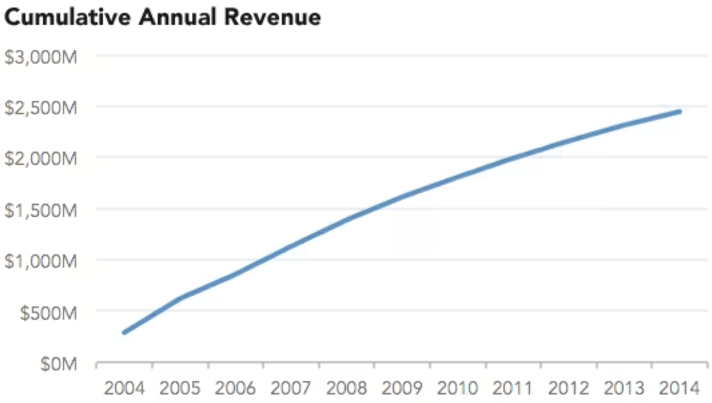
Lies that visualizations tell and how to spot them

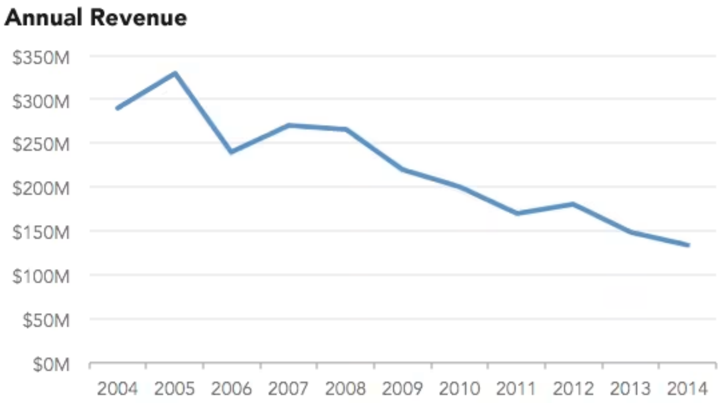


Lies that visualizations tell and how to spot them

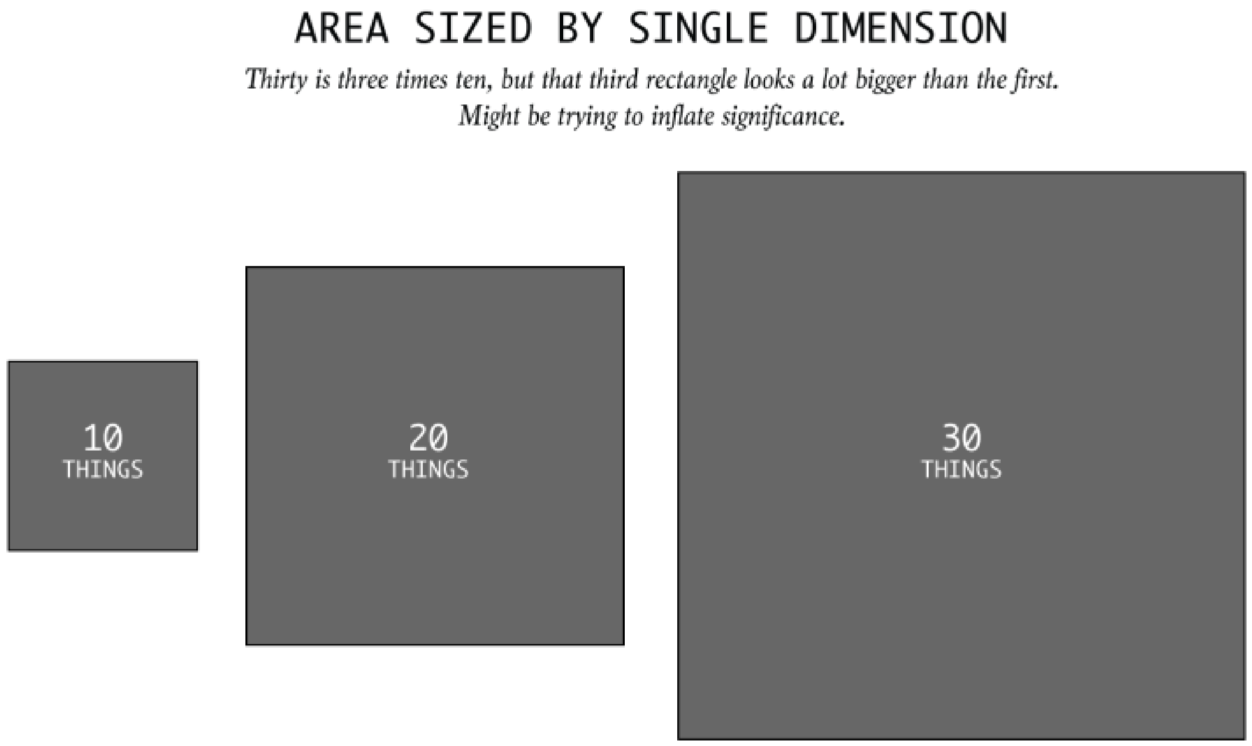


Lies that visualizations tell and how to spot them

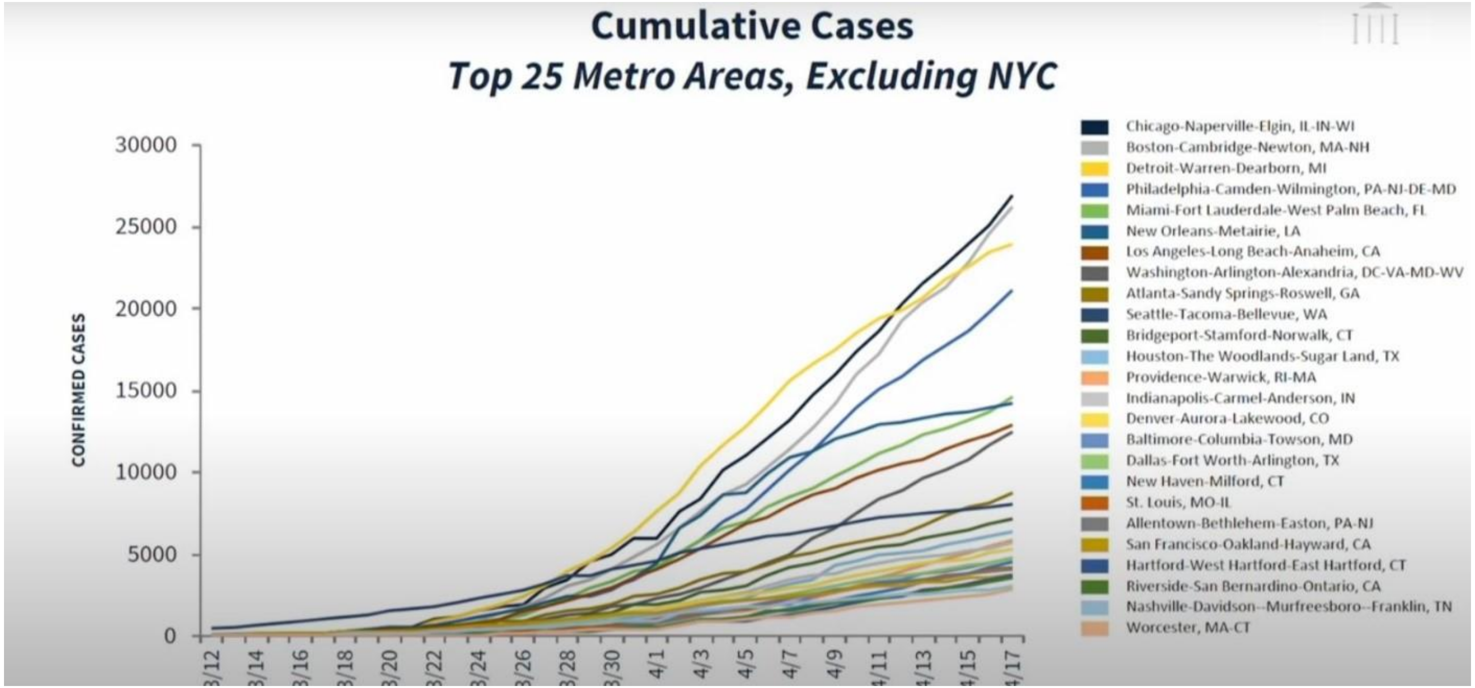




Lies that visualizations tell and how to spot them



Lies that visualizations tell and how to spot them



References

- https://towardsdatascience.com/5-ways-data-visualizations-can-lie-46e54f41de37
- https://www.everviz.com/blog/lies-damned-lies-and-visualizations/
- https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Smell the data

Phil Duan
@philduan

Inspecting large amount of data manually and visualizing the exact data fed into the network (after all the filtering, post-processing, etc.) is one of the best ML practices I've learned from @karpathy.

> **Jason Wei** @_jasonwei · Oct 3, 2023
>
> One pattern I noticed is that great AI researchers are willing to manually inspect lots of data. And more than that, they build infrastructure that allows them to manually inspect data quickly. Though not glamorous, manually examining data gives valuable intuitions about the
>
> Show more

11:06 AM · Oct 3, 2023 · **16.3K** Views

The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

I’ve been at OpenAI for almost a year now. In that time, I’ve trained a **lot** of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

Smell the data

Delip Rao e/σ
@deliprao

Subscribe

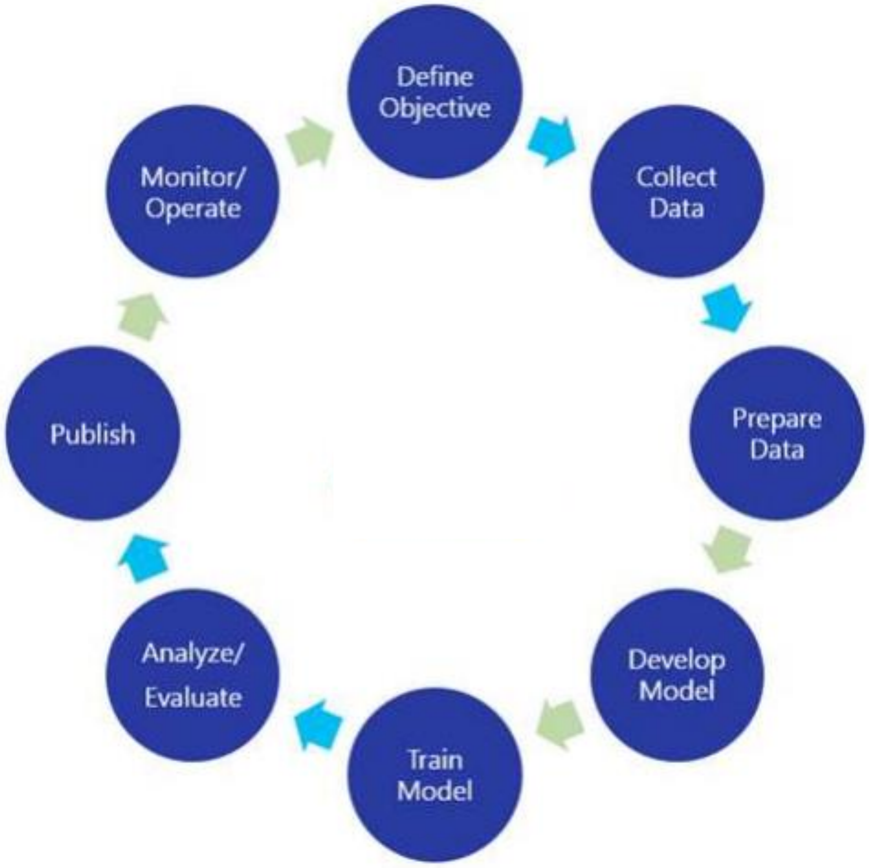
One of the best pieces of advice I embodied from my advisor is “smell the data”. You pay for it in compute and other ways, if you don’t do it, and from my experience working with others, most don’t. That’s one of the reasons why we have overly complicated archs, objectives, and reward functions.

Perversely, we reward these complications as “novelty” in literature, further incentivizing not smelling the data. But real-world deployments will not care about novelty. Here, simplicity and efficiency are rewarded. And guess what? Both come from taking time to smell the data.

I will

SMELL THE DATA

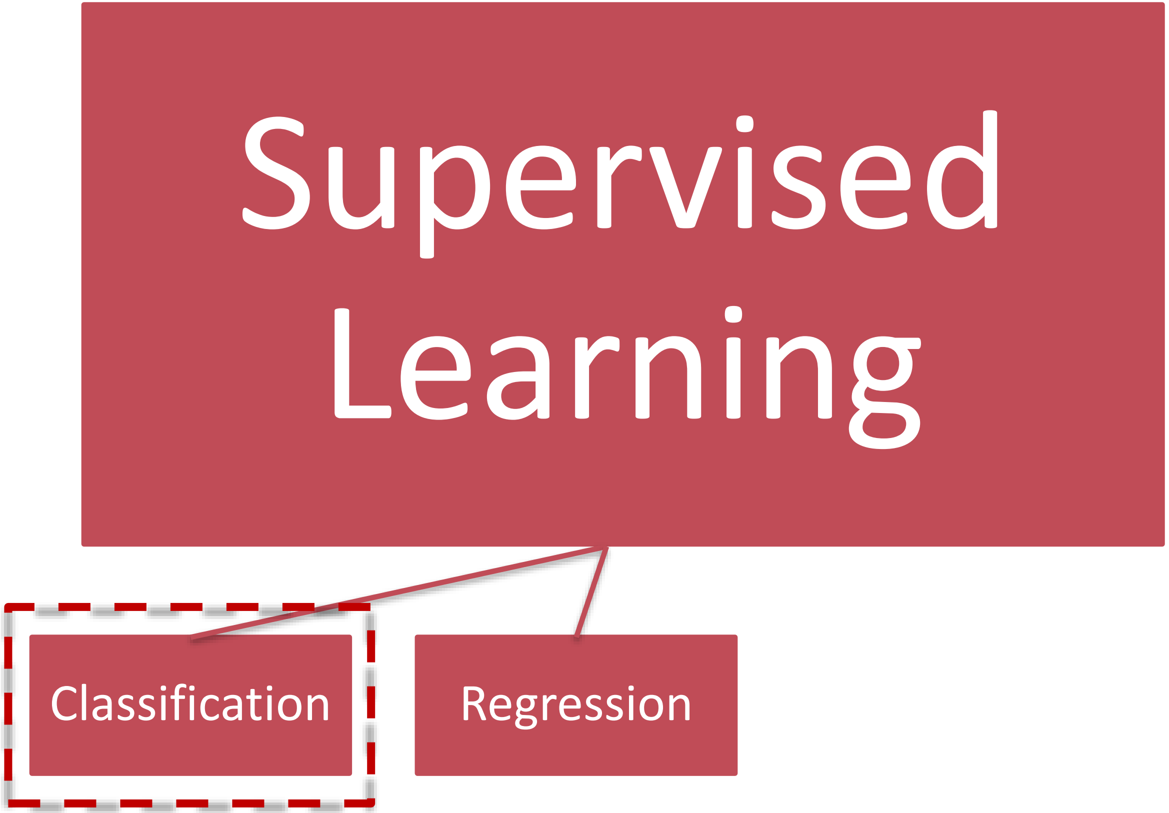
Workflow of a Machine Learning Problem



Lecture Outline

- *ML Workflow*
- *Data Representations*
- *Data Visualization*
- Intro to Supervised Learning
  - Taxonomy
  - Models

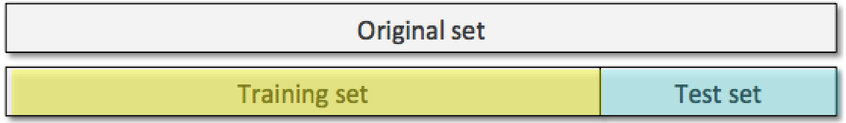
Supervised Learning


Supervised Learning
Classification
Regression

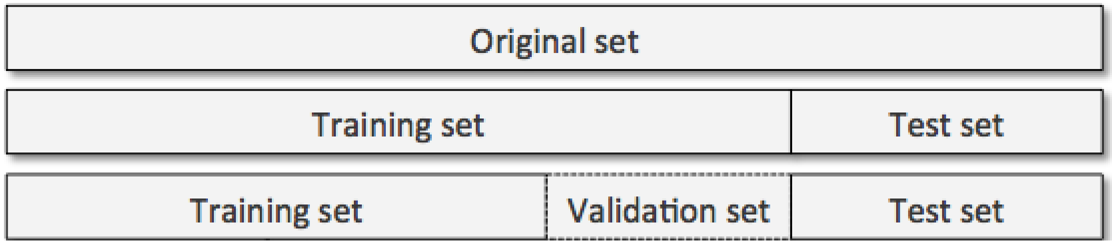
An interview analogy

1. Collect worked out problems (Q, S are both known)
2. Prepare on ALL the available problems.
3. Go for interview.

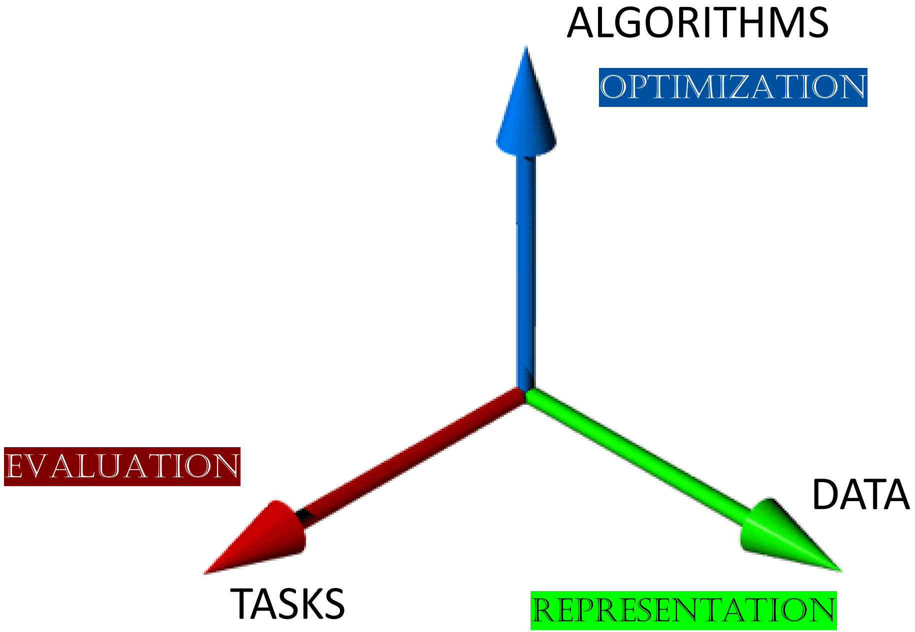
1. Collect worked out problems (Q,S are both known)
2. Randomly set aside a small number of problems.
3. Prepare on rest of the problems.
4. Take a mock interview containing all the 'set aside' problems.
5. Score answers and compare with solution.
6. Use mistakes to decide which topics to prepare better on.
7. Go for interview.

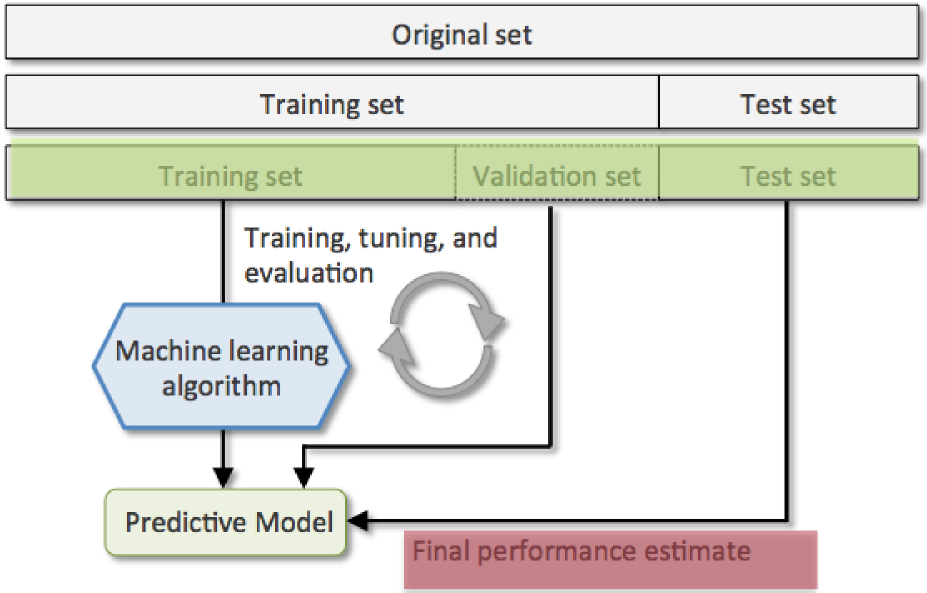


The Train-Validation-Test paradigm

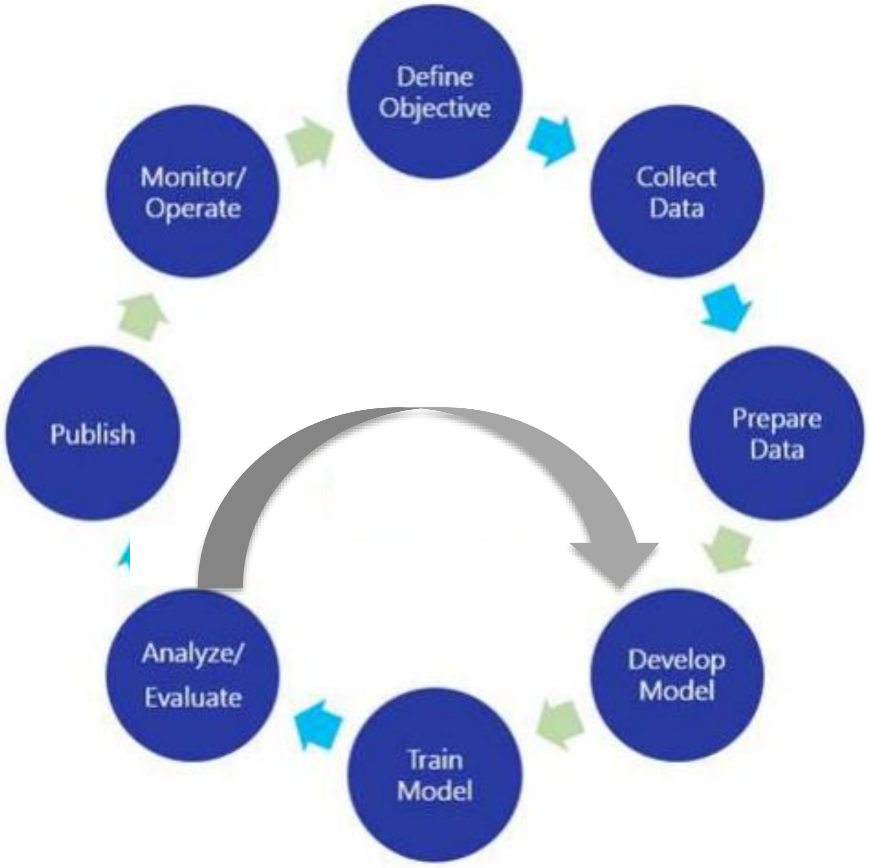


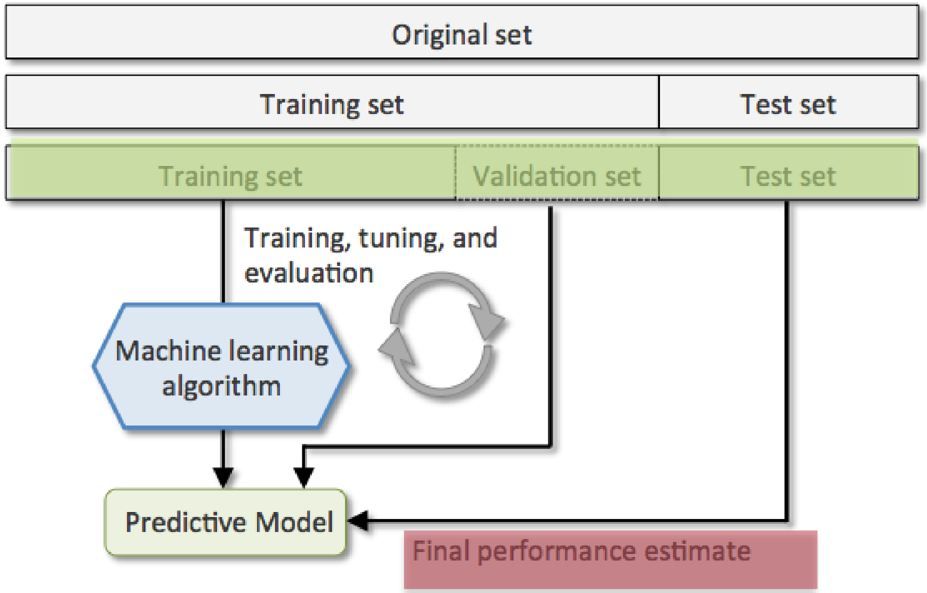
The Train-Validation-Test paradigm

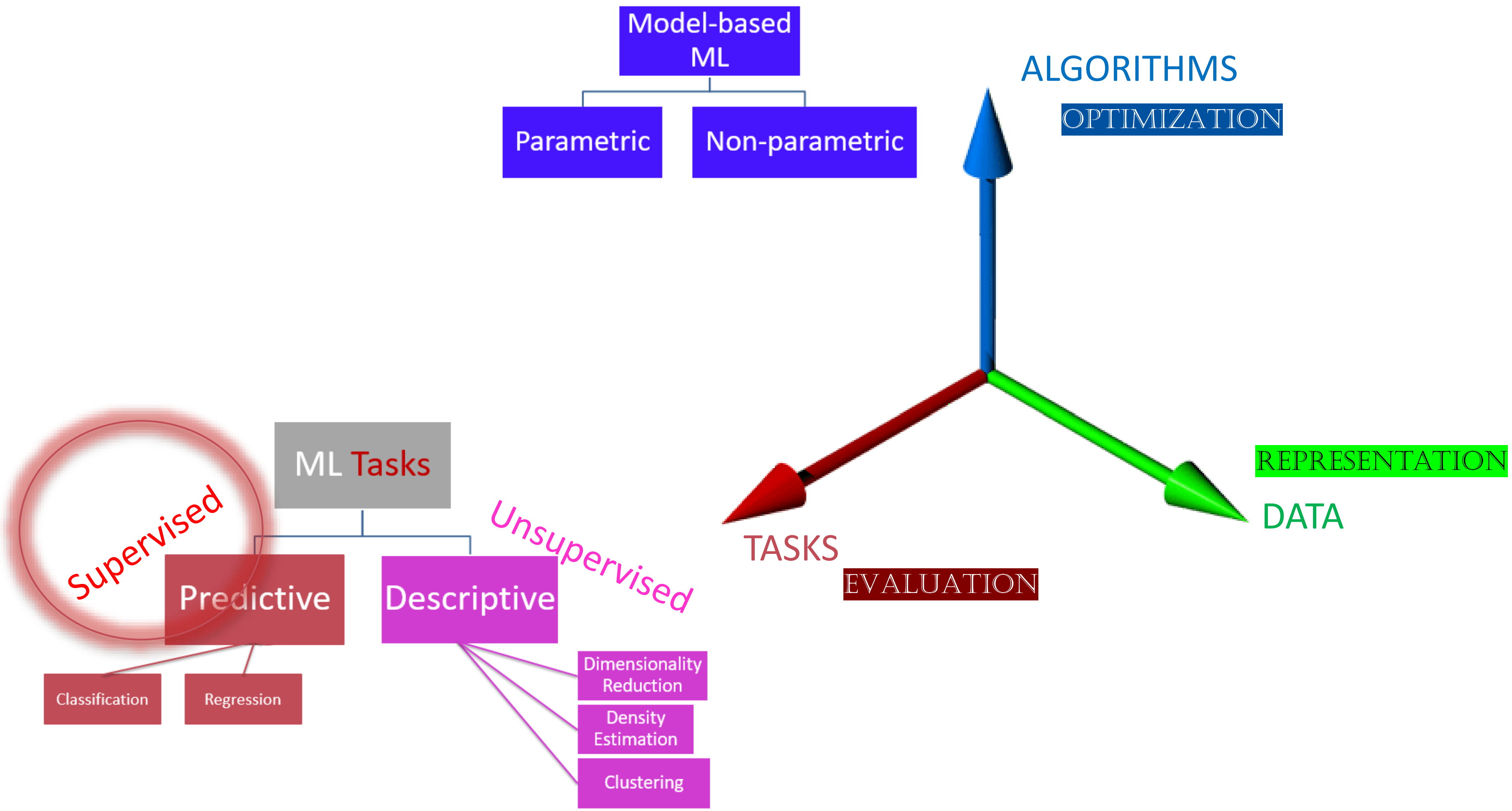


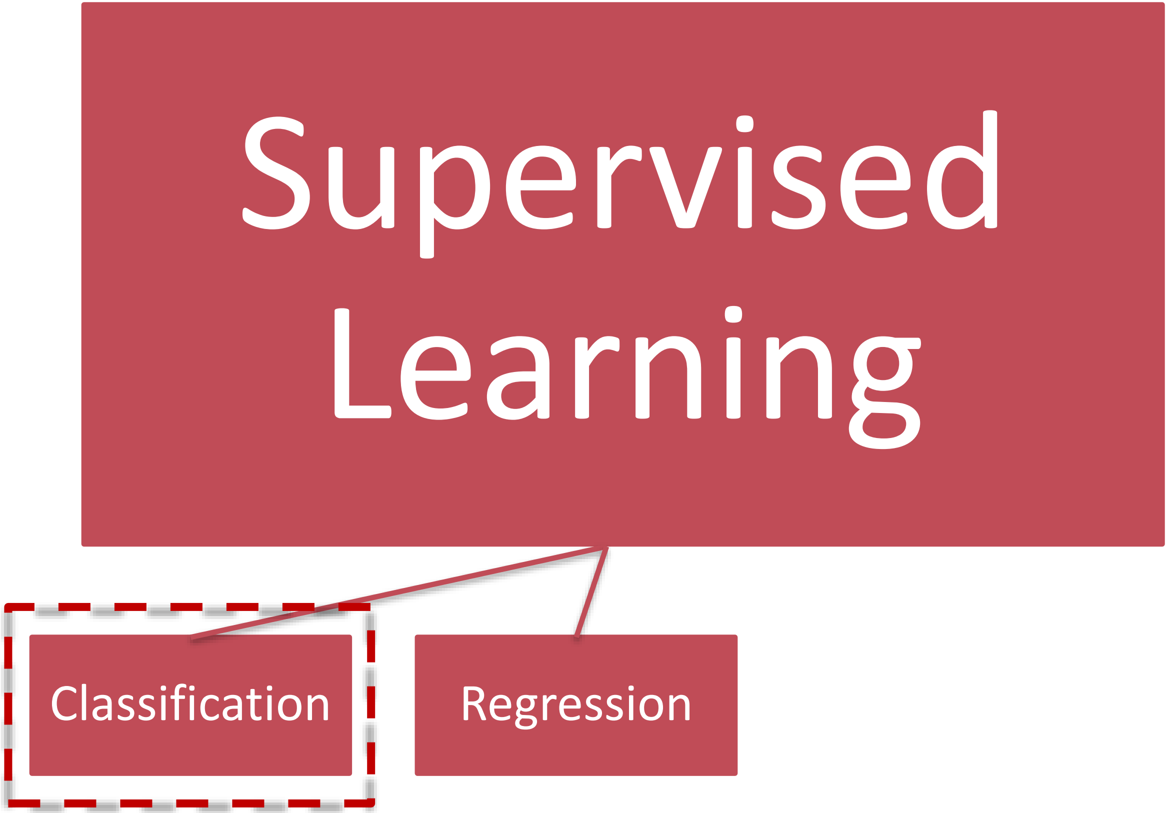


The Train-Validation-Test paradigm






Model-based ML
Parametric
Non-parametric
ALGORITHMS
OPTIMIZATION
REPRESENTATION
DATA
TASKS
EVALUATION
ML Tasks
Supervised
Unsupervised
Predictive
Descriptive
Classification
Regression
Dimensionality Reduction
Density Estimation
Clustering

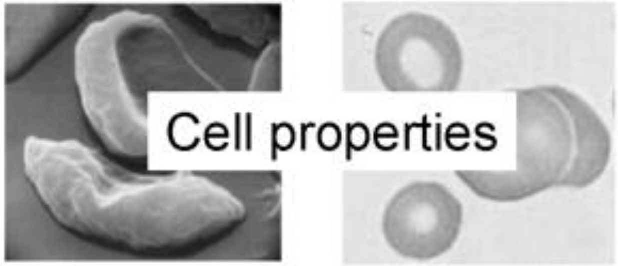

Supervised Learning
Classification
Regression

ML::Tasks → Predictive → Classification

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
...

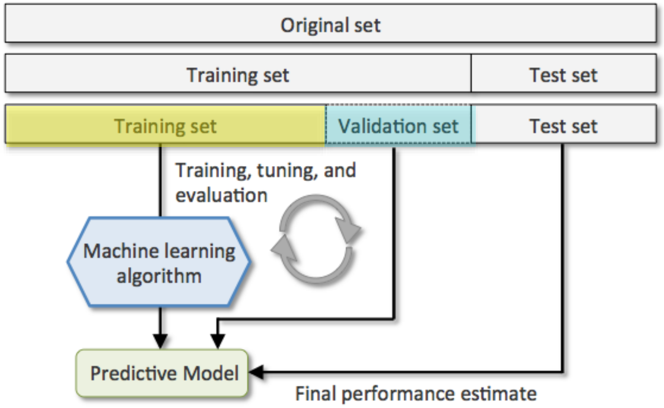


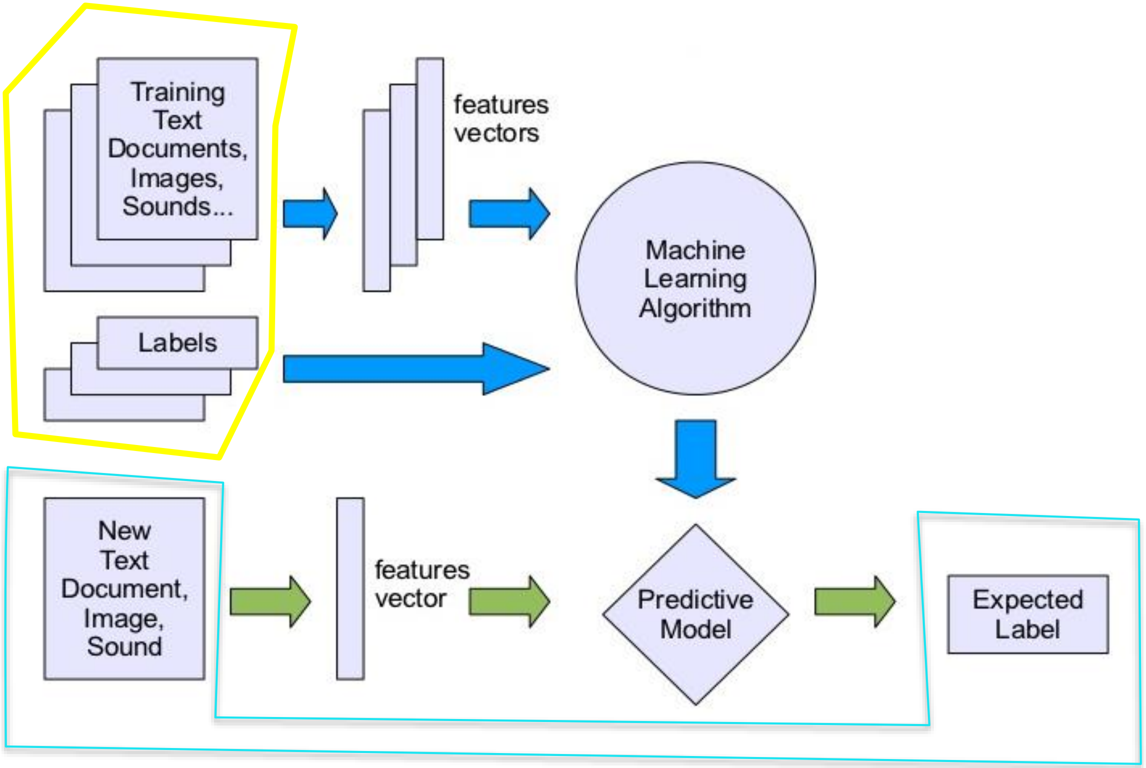
"Anemic cell"
"Healthy cell"

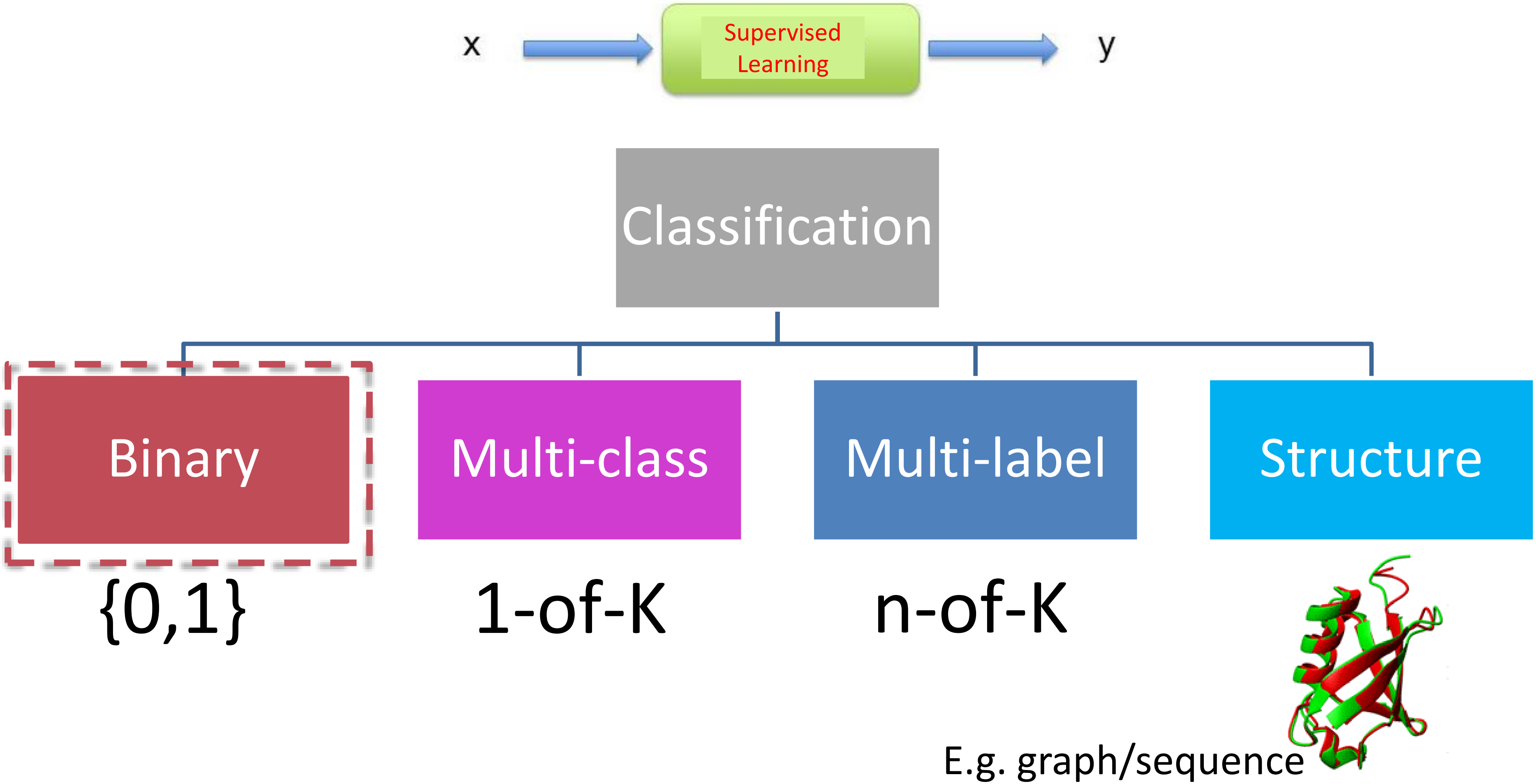
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Discrete Labels

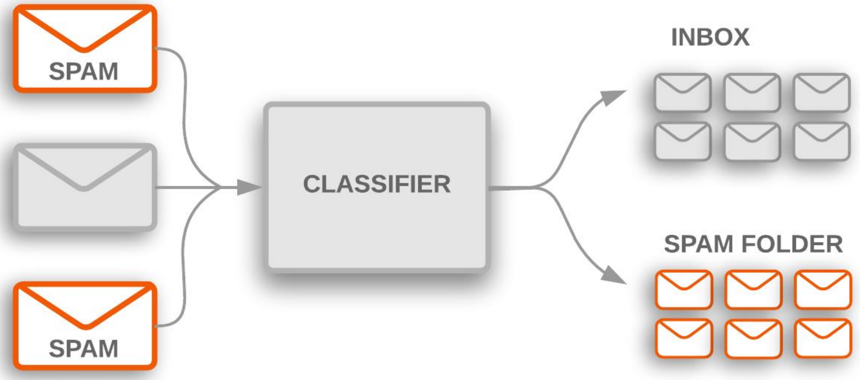
The Train-Validation-Test paradigm






x
Supervised Learning
y
Classification
Binary
Multi-class
Multi-label
Structure
{0,1}
1-of-K
n-of-K
E.g. graph/sequence

Binary Classification



Performance Measures - Accuracy

$$Accuracy = \frac{(100 + 50)}{165} = 0.91$$

$$Misclassification = \frac{(10 + 5)}{165} = 0.09$$

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

- **Pool of 100 patients' data used for validation of a cancer prediction ML model**
- Prediction:
  - 3 have cancer
  - Rest (100-3=97) are healthy.
- Reality:
  - 1 of the 3 did not actually have cancer !
  - 3 from 97 predicted healthy actually have cancer
- Accuracy =

n=	Predicted: NO	Predicted: YES
Actual: NO		
Actual: YES		

- **Pool of 100 patients' data used for validation of a cancer prediction ML model**
- Prediction:
  - 3 have cancer
  - Rest (100-3=97) are healthy.
- Reality:
  - 1 of the 3 did not actually have cancer !
  - 3 from 97 predicted healthy actually have cancer
- Accuracy = (100 - 4) / 100 = 96% !

n=	Predicted: NO	Predicted: YES
Actual: NO		
Actual: YES		

- **Pool of 100 patients' data used for validation of a cancer prediction ML model**
- Prediction:
  - 3 have cancer → selected for chemotherapy
  - Rest (100-3=97) are healthy.
- Reality:
  - 1 of the 3 did not actually have cancer !
  - 3 from 97 predicted healthy actually have cancer → should have been selected for chemotherapy
- Accuracy = (100 - 4) / 100 = 96% !

n=	Predicted: NO	Predicted: YES
Actual: NO		
Actual: YES		

Performance Measures - Accuracy

$$Accuracy = \frac{(100 + 50)}{165} = 0.91$$

$$Misclassification = \frac{(10 + 5)}{165} = 0.09$$

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Performance Measures – Accuracy, TPR, FPR

$$Accuracy = \frac{(100 + 50)}{165} = 0.91$$

$$Misclassification = \frac{(10 + 5)}{165} = 0.09$$

$$FalsePositiveRate(FP) = \frac{(10)}{60} = 0.17$$

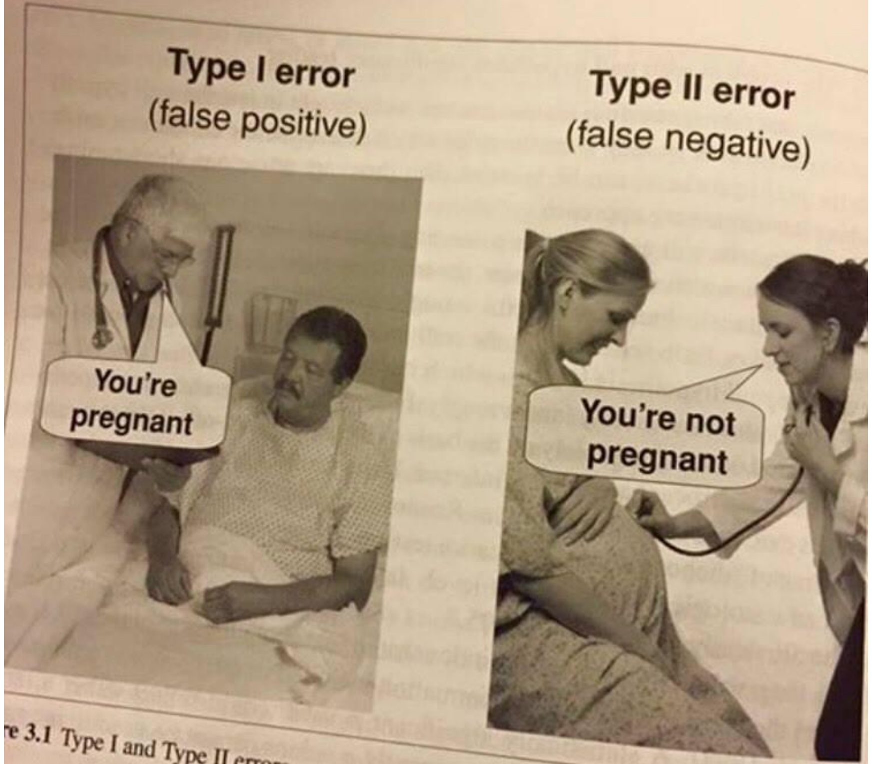
$$FalseNegativeRate(FN) = \frac{(5)}{105} = 0.048$$

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

$$TrueNegativeRate(TN) = \frac{(50)}{60} = 0.833$$

$$TruePositiveRate(TP) = \frac{(100)}{105} = 0.95$$

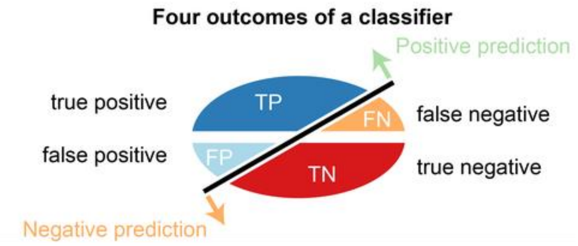
n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



re 3.1 Type I and Type II errors

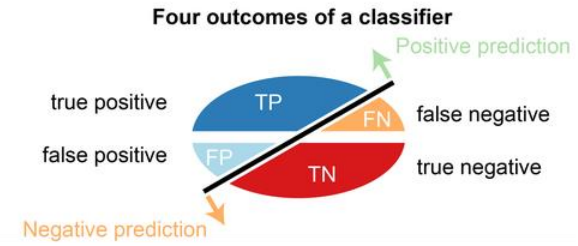
levels to .01 or even .001

Summary of Measures

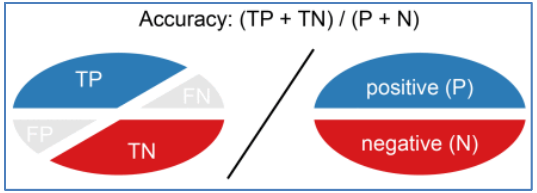


n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Summary of Measures

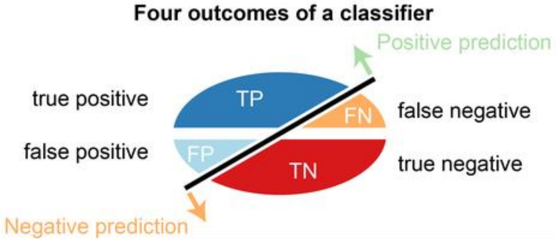


n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

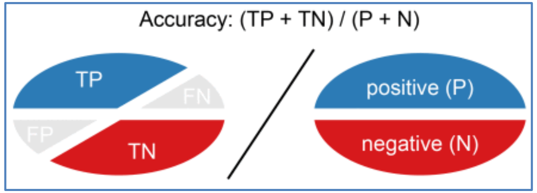


% of correct predictions

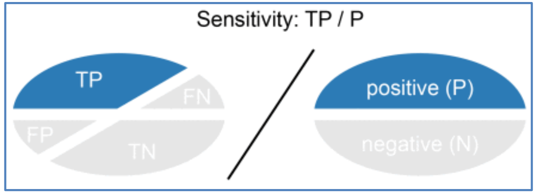
Summary of Measures



n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

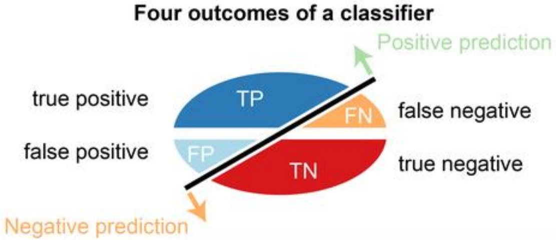


% of correct predictions

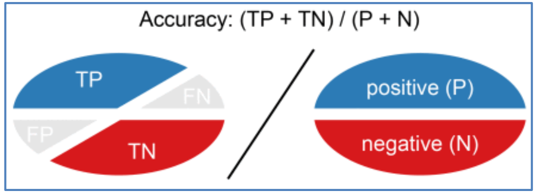


% of + class correctly predicted
[aka Recall / TPR]

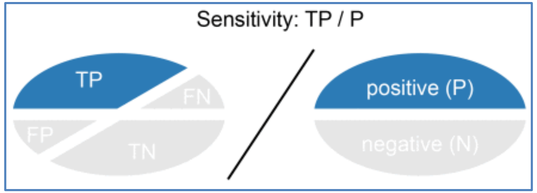
Summary of Measures



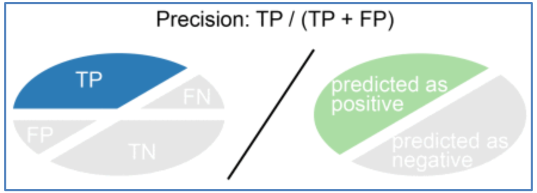
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



% of correct predictions

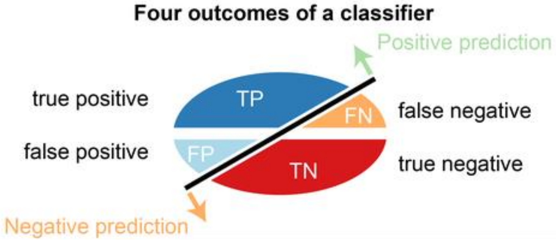


% of + class correctly predicted
[aka Recall / TPR]

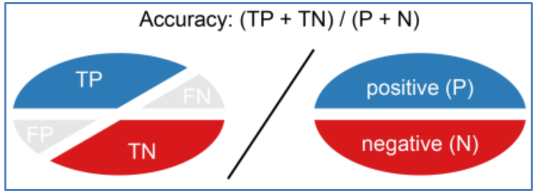


correct prediction of + class
[aka Precision]

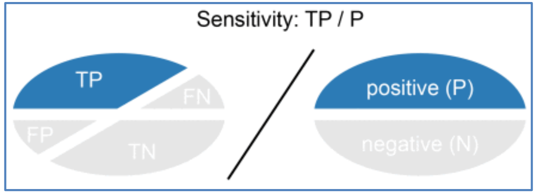
Summary of Measures



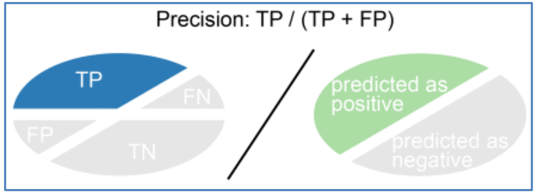
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



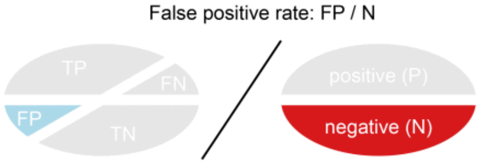
% of correct predictions



% of + class correctly predicted
[aka Recall / TPR]



correct prediction of + class
[aka Precision]



% of – class incorrectly predicted

- **Cancer-Prediction System**
- Precision =
- Recall =
- Accuracy =

- **Cancer-Prediction System**
- Precision = 2/(2+1) = 67%
- Recall = 2/(2+3) = 40%
- Accuracy = (94+2)/100 = 96%

What measure are we optimizing for ?

- Screening for a terminal disease
- Do not want to miss anyone: Maximize Recall
- Classification between apples and oranges
  - Both types of errors are equally imp.: Maximize Accuracy
- Automatic bombing on detecting a target from a drone
  - Should not hurt civilians: Zero False Positives
- Giving access to a secure installation
  - No access to unauthorized personnel: Low False Positives

Cost

- Sometimes, there is a cost for each error
  - E.g. Earthquake prediction
    - False positive: Cost of preventive measures
    - False negative: Cost of recovery

- Detection Cost (Event detection)
  - Cost = C_{FP} * FP + C_{FN} * FN

Farmer Shri MoneyBags and ML-FruitPicker

- MB : I want an automated fruit picker. I will pay a large amount of money for it.
- You (having just finished this course) : Sure
- *You (Thinking): I love large amounts of money* 😎

Farmer Shri MoneyBags and ML-FruitPicker

After 6 months …

- MB : Well ?
- You : I have a High Precision ML-FruitPicker. But its Recall is 20% !
- MB : (confused) Precision ? Recall ?
- *You : (thinking) Should I go over first 3 lectures of SMAI with MB ? He'll probably run away !*
- You : It rejects 80% of good, pickable fruit, but whatever it picks, those fruits are good !
- MB : I'll take your system. How do I transfer large amount of money to you ?
- You : 😲
- MB (seeing your shocked face) : See, in a batch of 100 fruits, 10 fruits are usually bad. Among the 90 good ones, your system will select 18 of them on average. But from any given selection, I pack only 8.