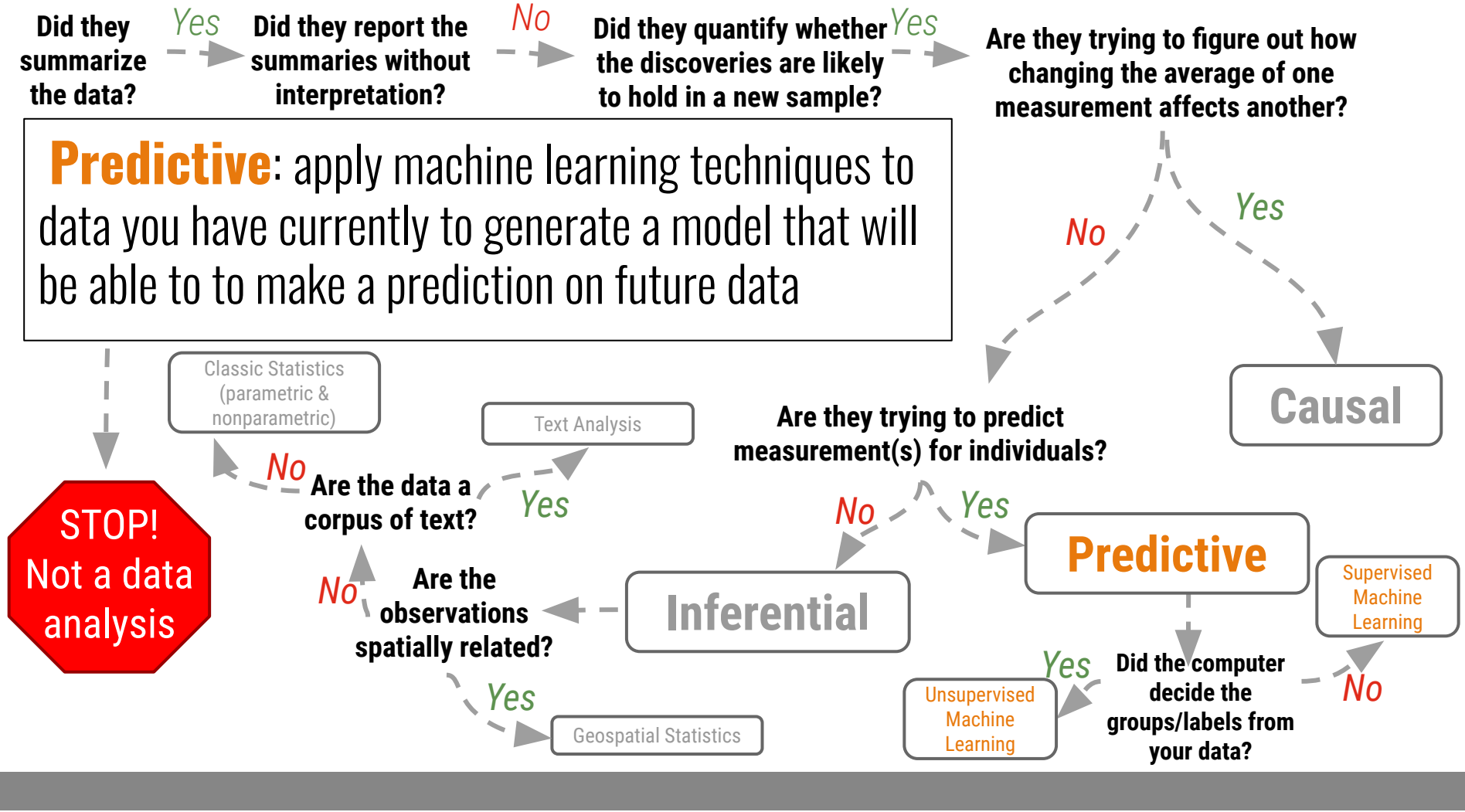
Plan:

1. Define predictive analysis
2. Explain the relationship between ML and AI
3. Walk through the four general steps to predictive analysis

Machine Learning: Basics

Shannon E. Ellis, Ph.D
UC San Diego

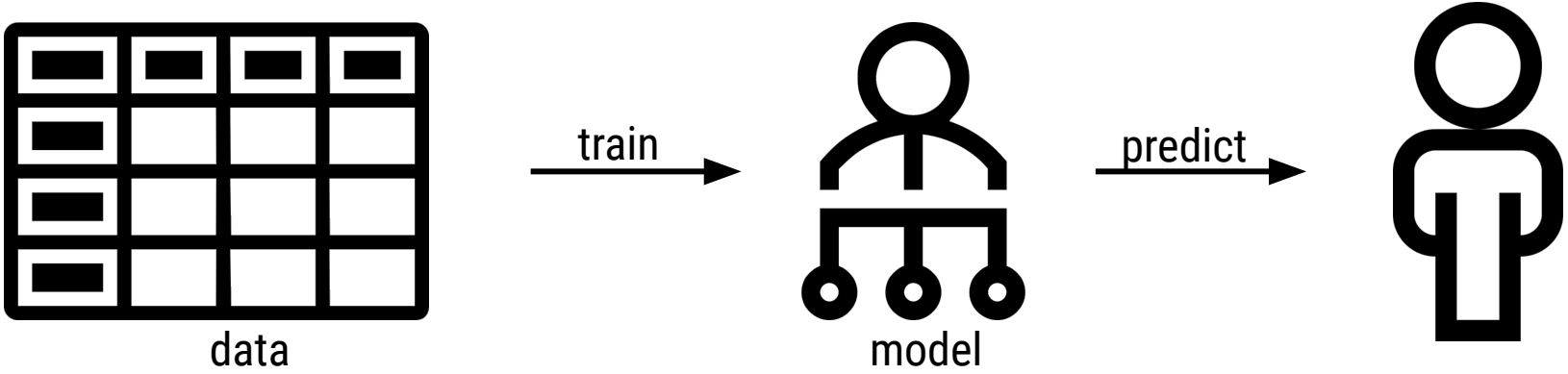
Department of Cognitive Science
sellis@ucsd.edu


Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
Predictive: apply machine learning techniques to data you have currently to generate a model that will be able to to make a prediction on future data
No
Yes
Causal
Are they trying to predict measurement(s) for individuals?
No
Yes
Predictive
Supervised Machine Learning
Did the computer decide the groups/labels from your data?
Yes
Unsupervised Machine Learning
No
Inferential
Are the observations spatially related?
No
Yes
Geospatial Statistics
Are the data a corpus of text?
No
Yes
Text Analysis
Classic Statistics (parametric & nonparametric)
STOP! Not a data analysis

- **Problem:** Detecting whether credit card charges are fraudulent.
- **Data science question:** Can we use the time of the charge, the location of the charge, and the price of the charge to predict whether that charge is fraudulent or not?
- **Type of analysis:** Predictive analysis

predictive analysis uses data you have now to make predictions in the future

machine learning approaches are used for predictive analysis!



What is machine learning?

"Machine learning is the science of getting computers to act without being explicitly programmed"

- Andrew Ng, Stanford, ex-Google, chief scientist at Baidu, Coursera founder, Stanford Adjunct Faculty

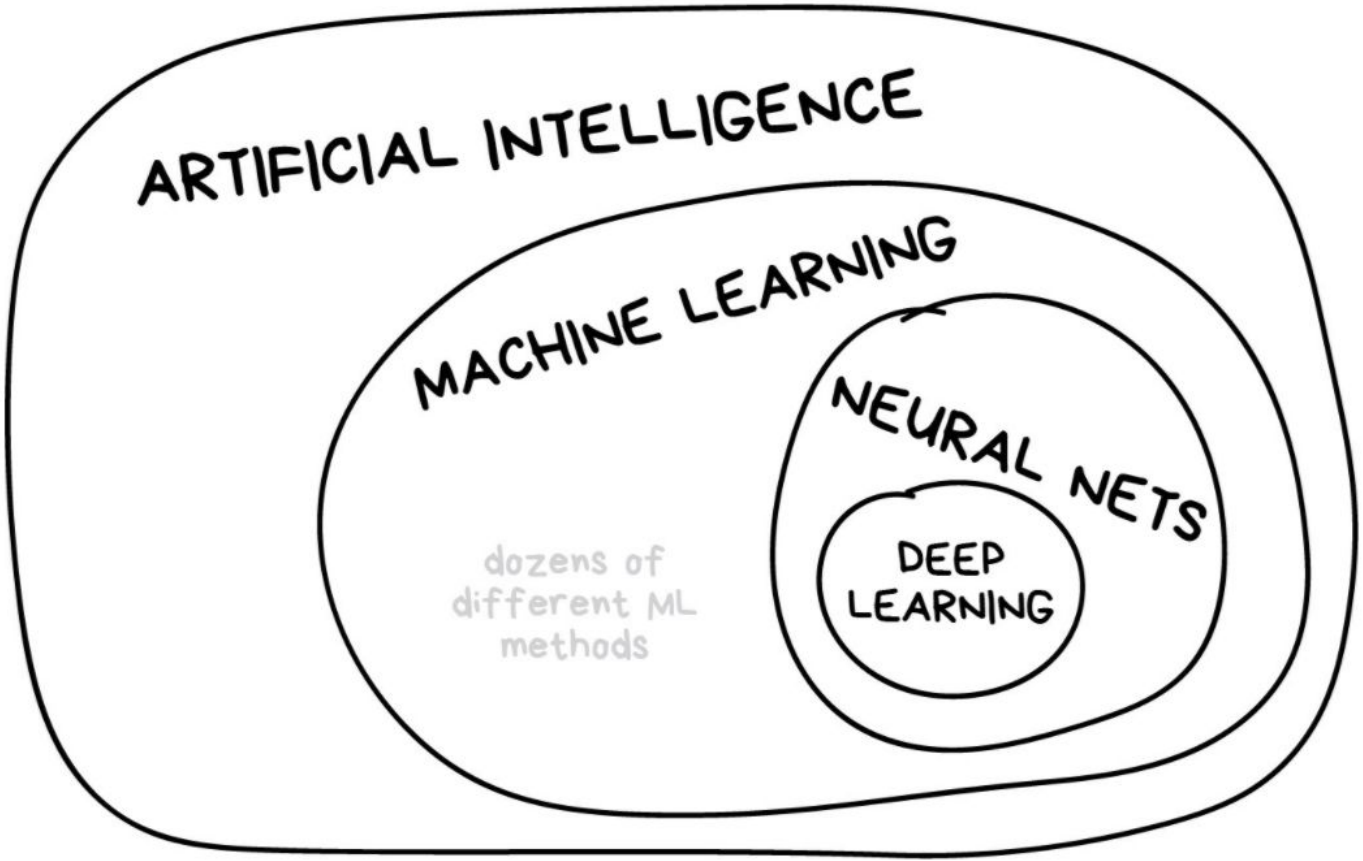


Image source: https://vas3k.com/blog/machine_learning/

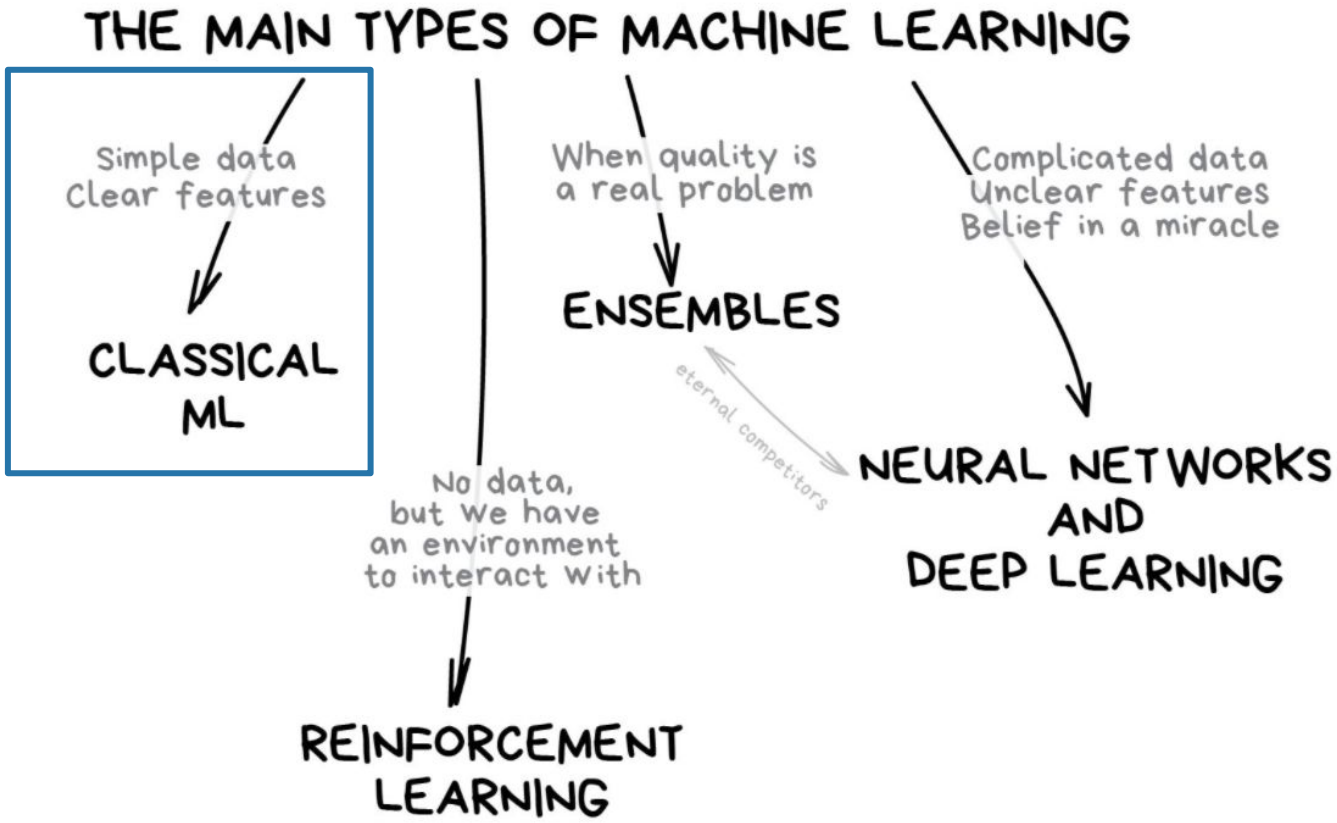
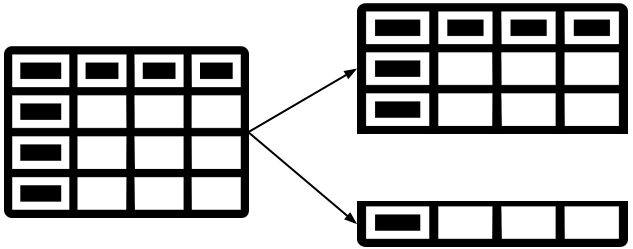


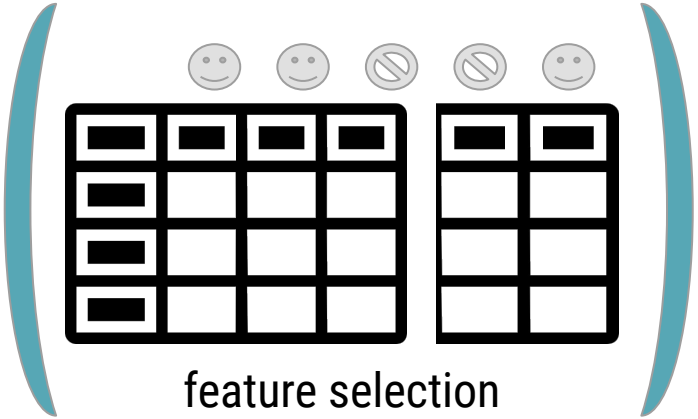
Image source: https://vas3k.com/blog/machine_learning/

Machine Learning Generalizations

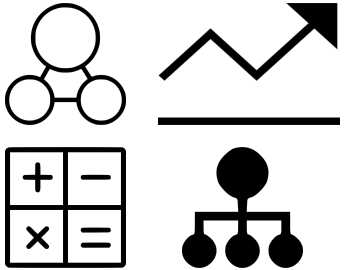
Basic Steps to Prediction



data partitioning

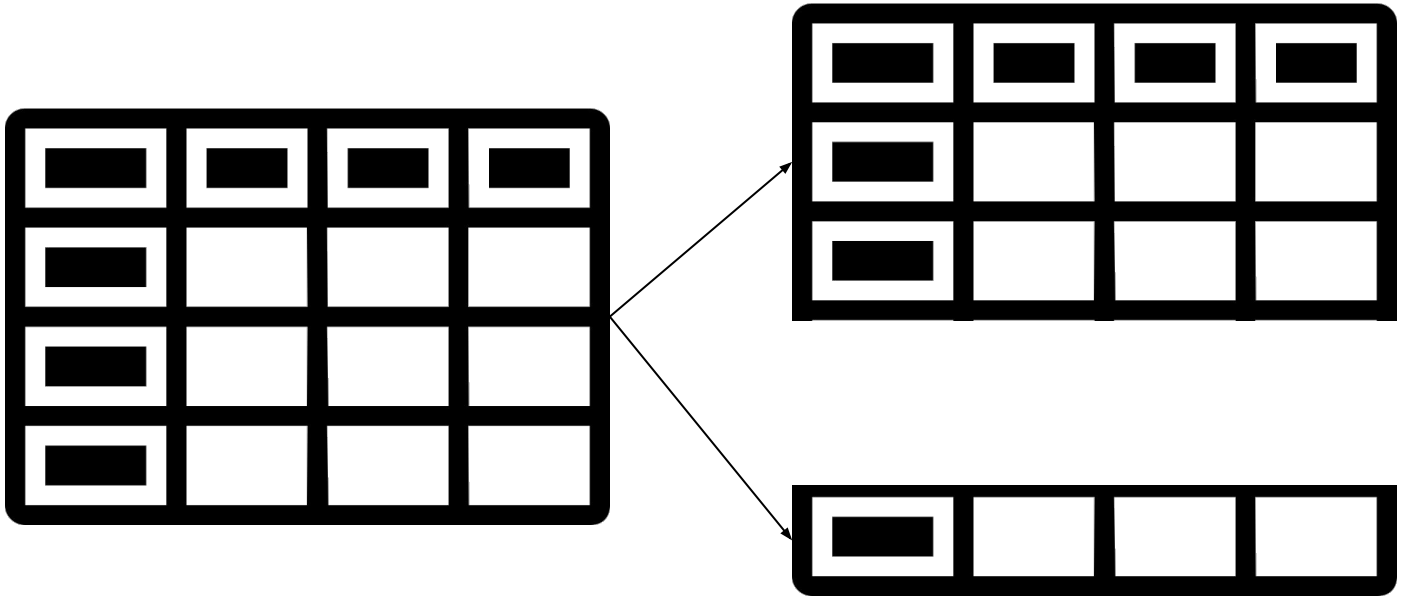


feature selection



model selection

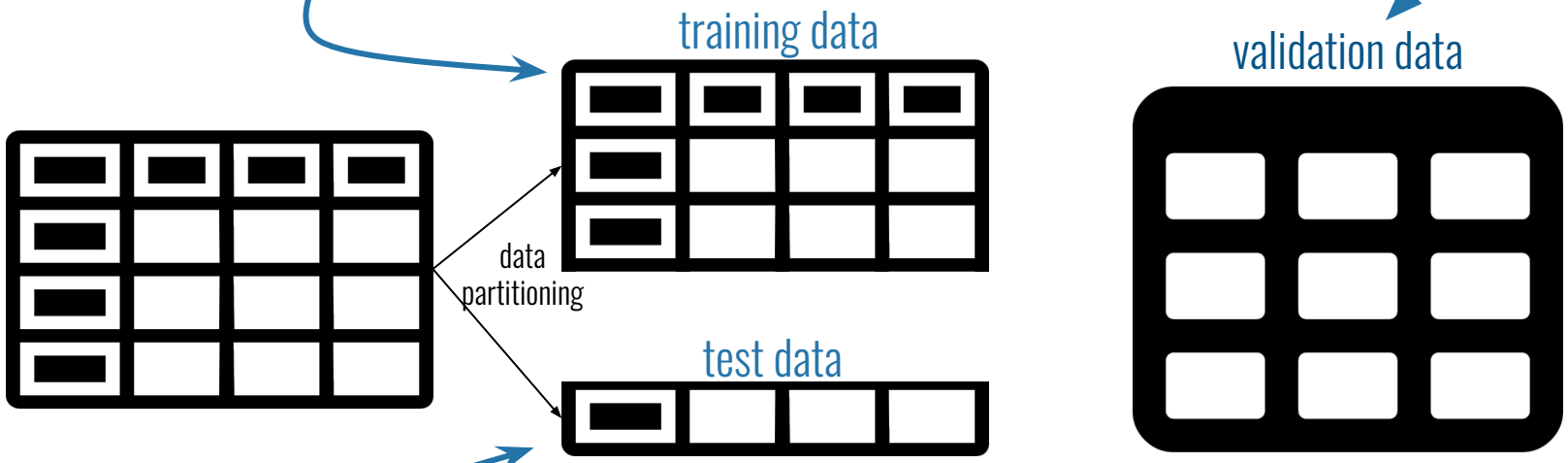
model assessment



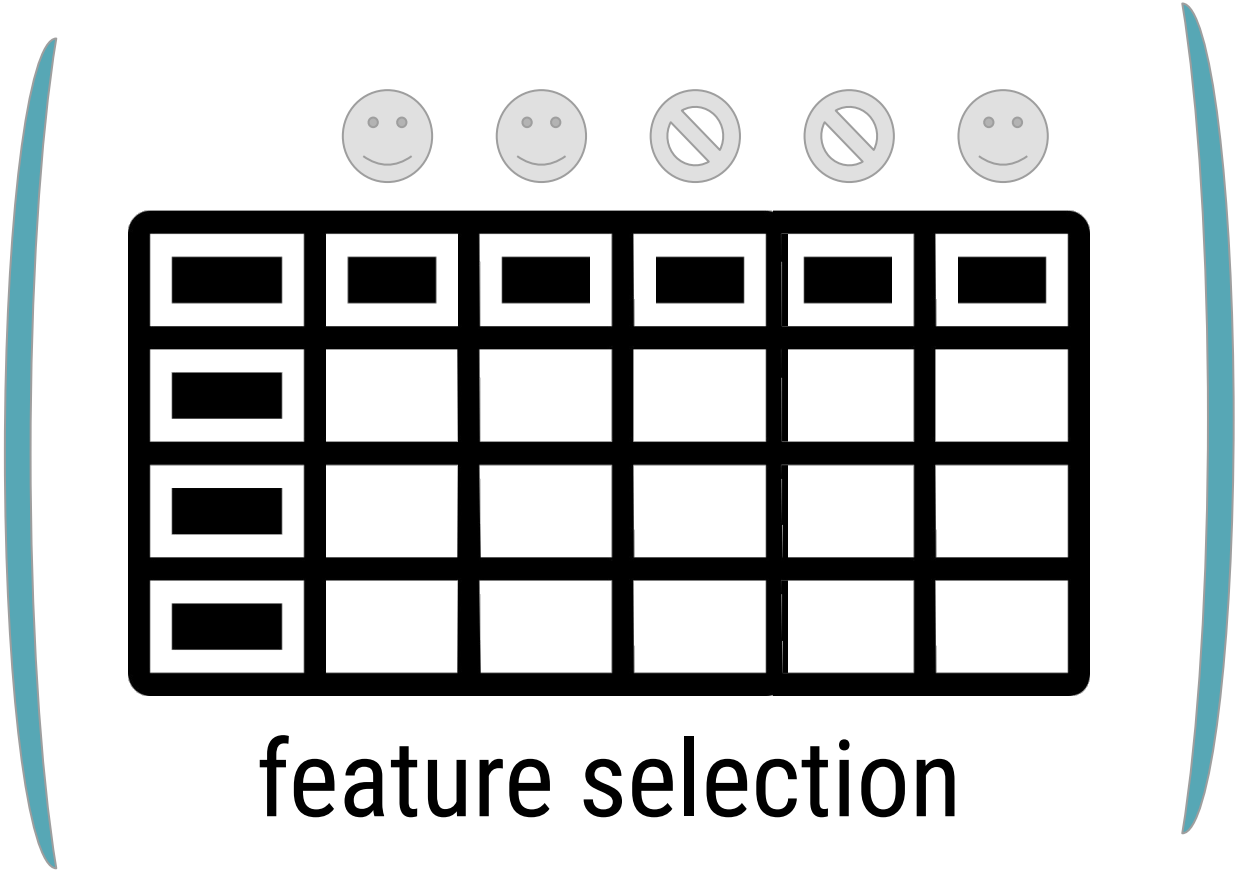
data partitioning

the data used to build your predictive model

new and independent data set used to assess if prediction model is generalizable

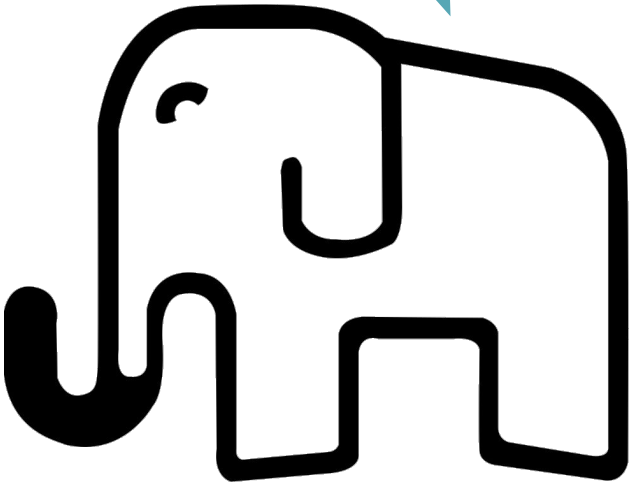


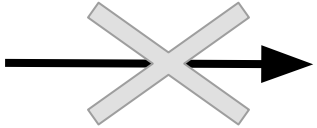
Data from original dataset that was held out and not used in training the model ; helpful in fine-tuning prediction accuracy

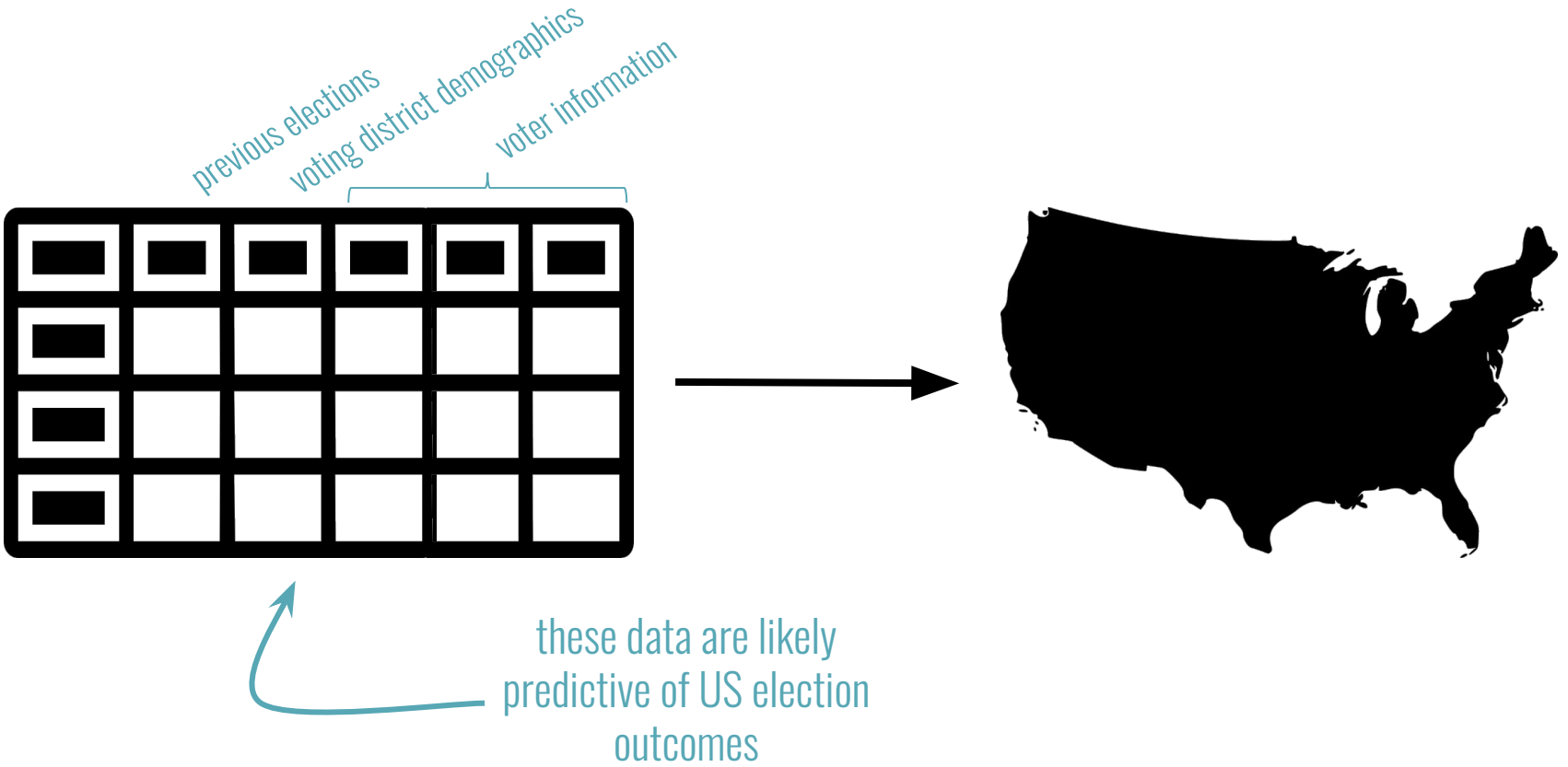


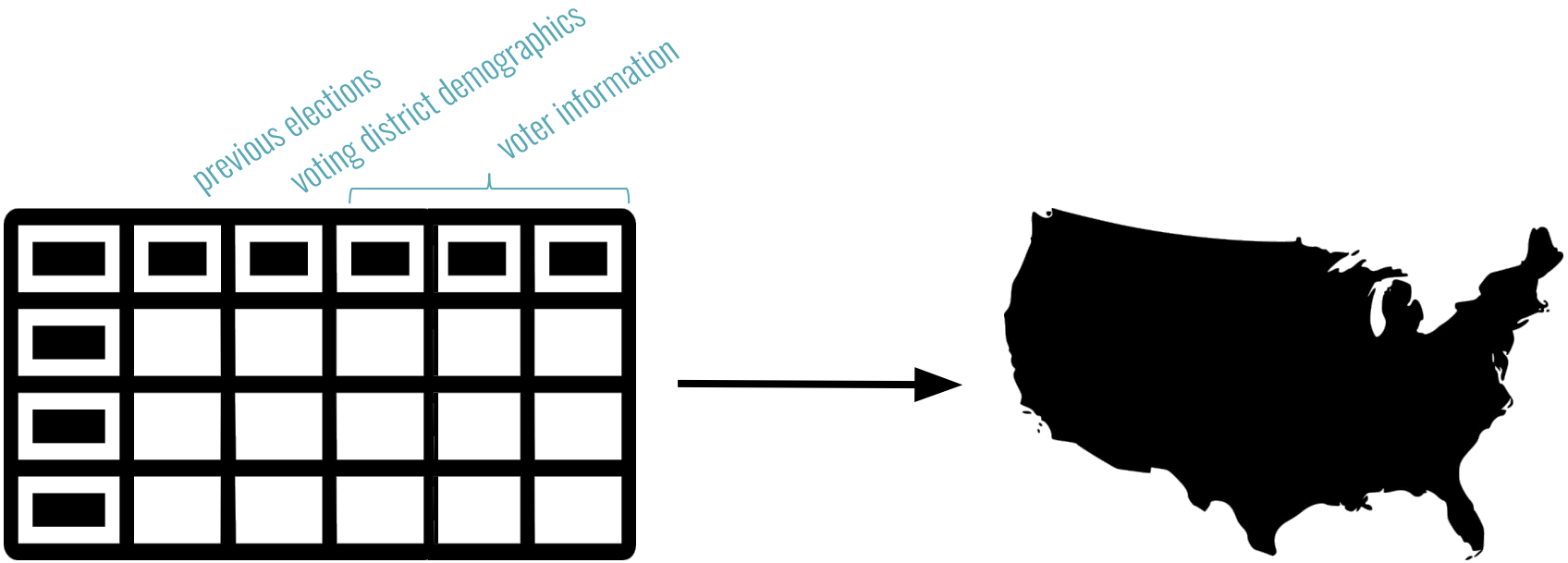
feature selection

elephant height data are likely
not predictive of US elections

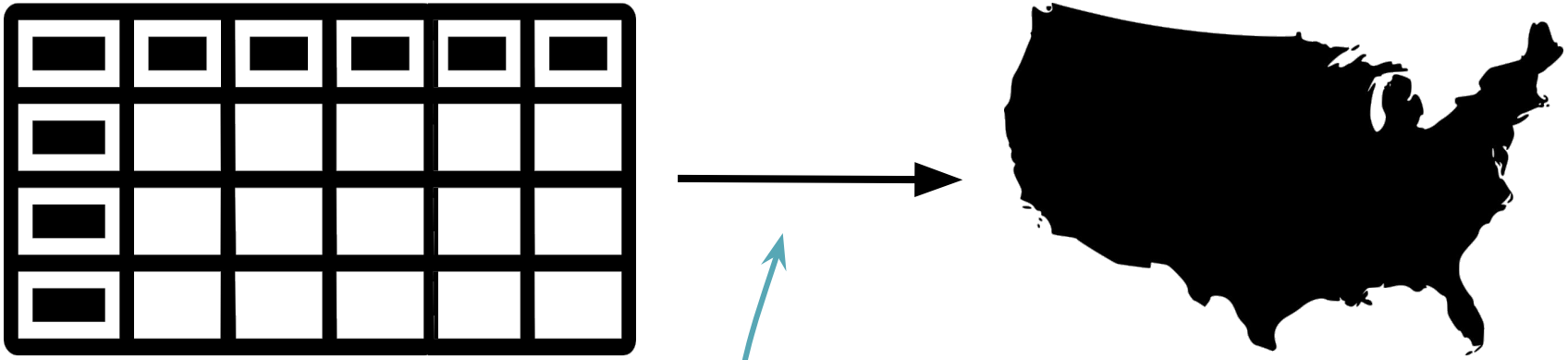




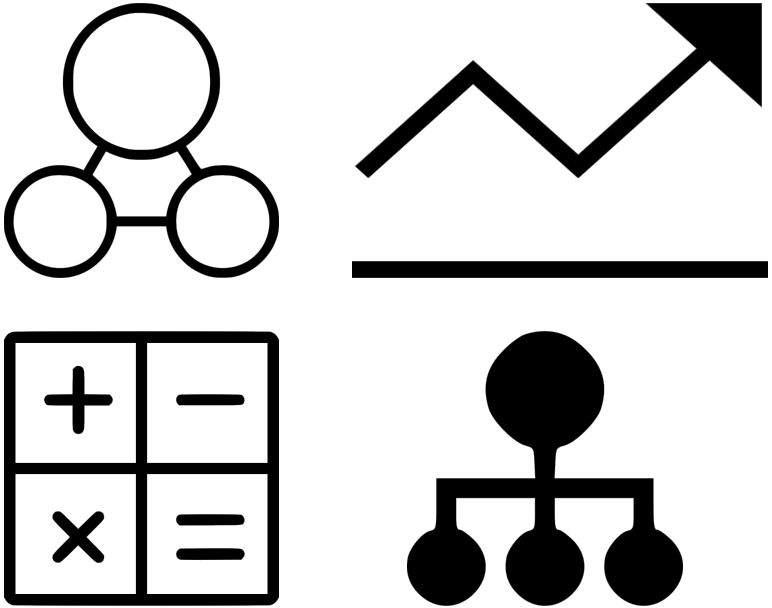

previous elections
voting district demographics
voter information
these data are likely predictive of US election outcomes



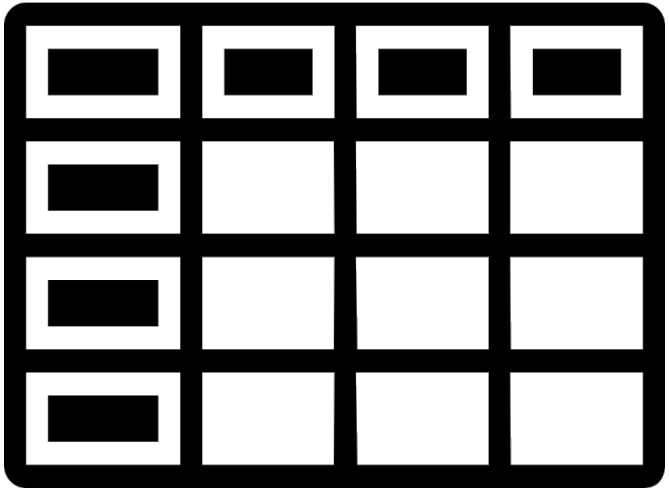
feature selection determines which variables are most predictive and includes them in the model



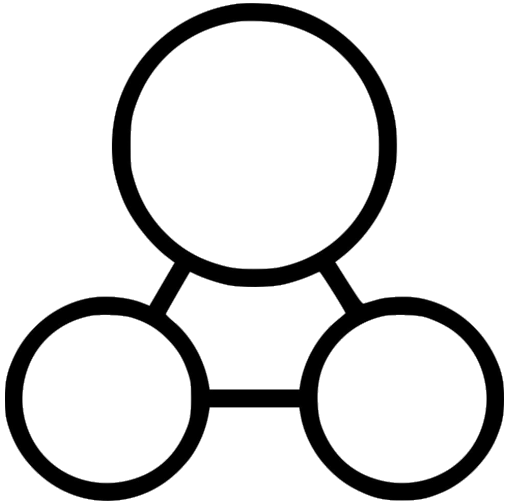
variables that can be used for accurate prediction exploit the relationship between the variables but do NOT mean that one causes the other



model selection



big
datasets



simple
models

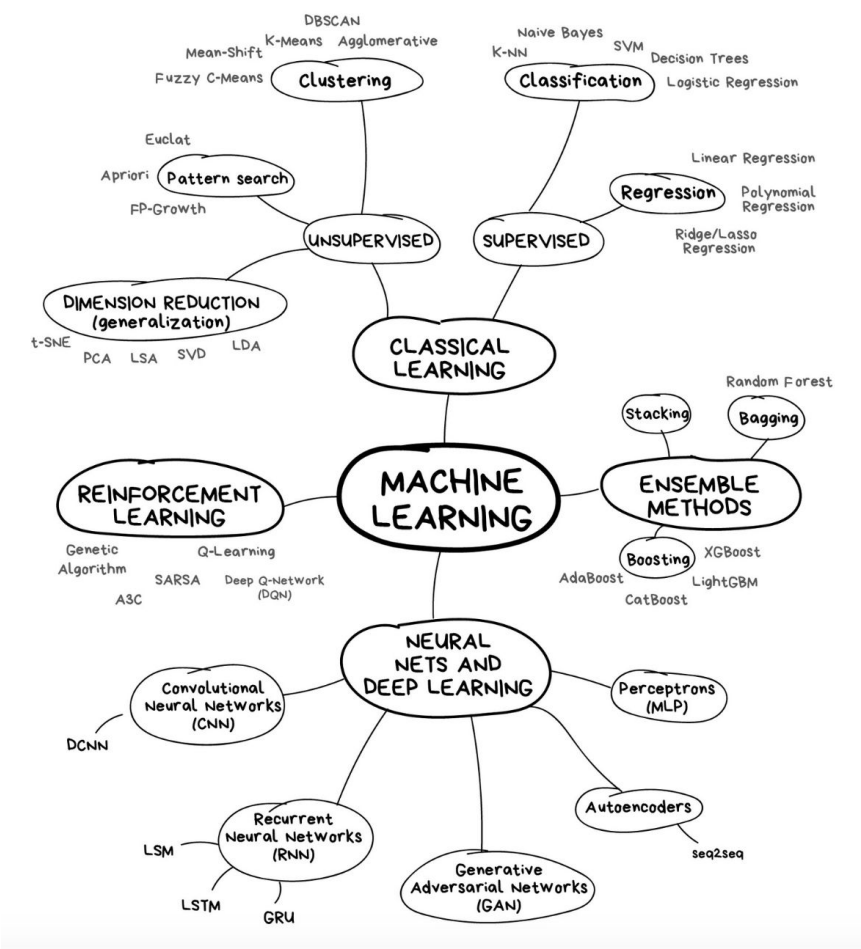
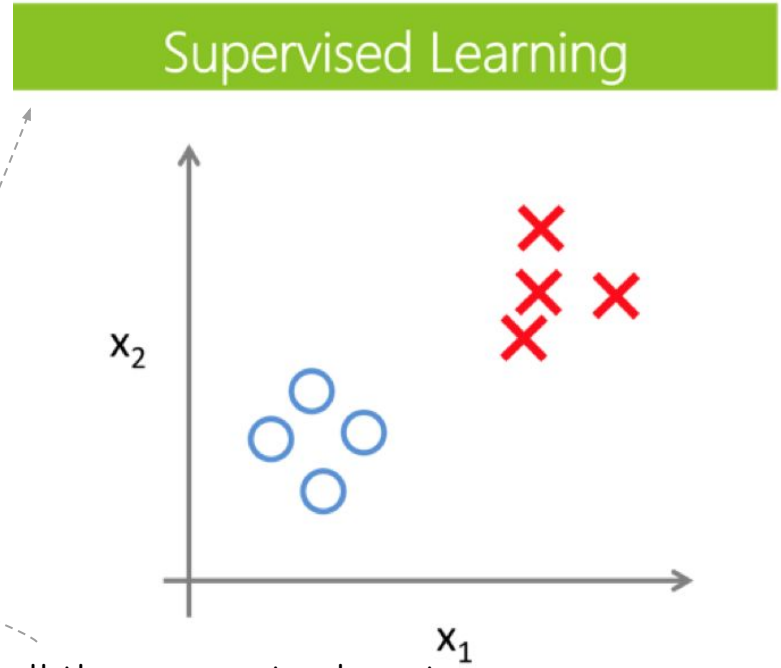
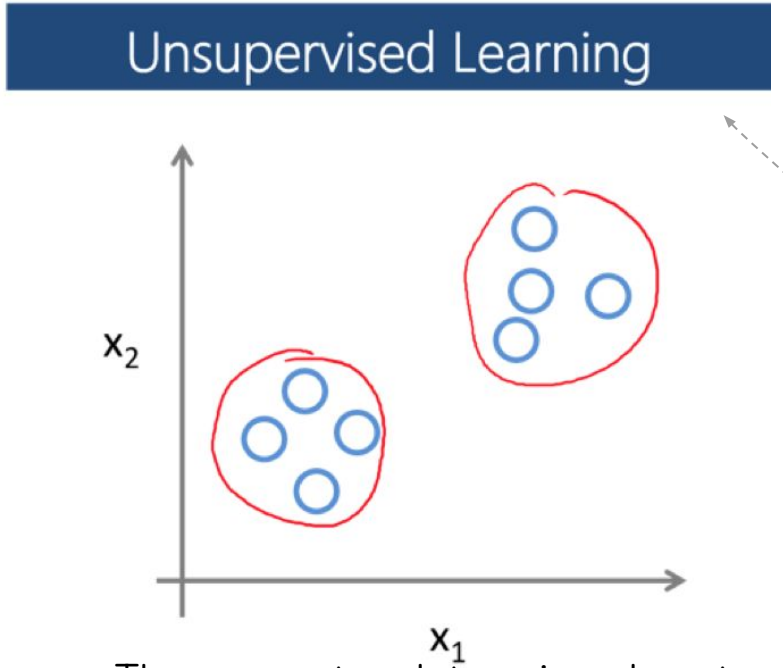


Image source: https://vas3k.com/blog/machine_learning/

To modes of machine learning



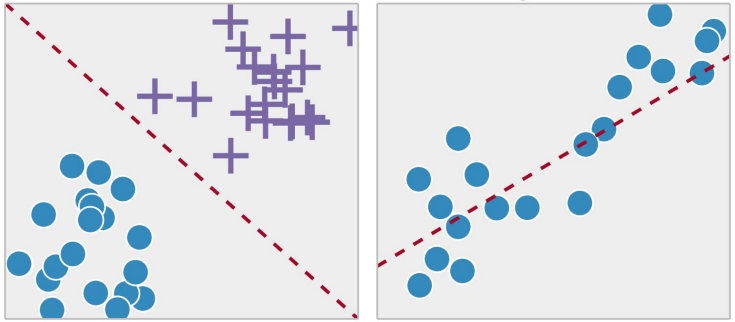


You tell the computer how to classify the observations

The computer determines how to classify based on properties within the data

Approaches to machine learning

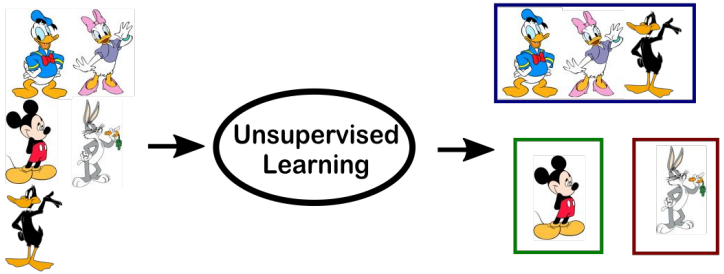
Supervised Learning



categorical variables

continuous variables

Unsupervised Learning



Clustering (categorical)
& dimensionality reduction (continuous)

Prediction accuracy dependent on training data

can automatically identify structure in data

Image source: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

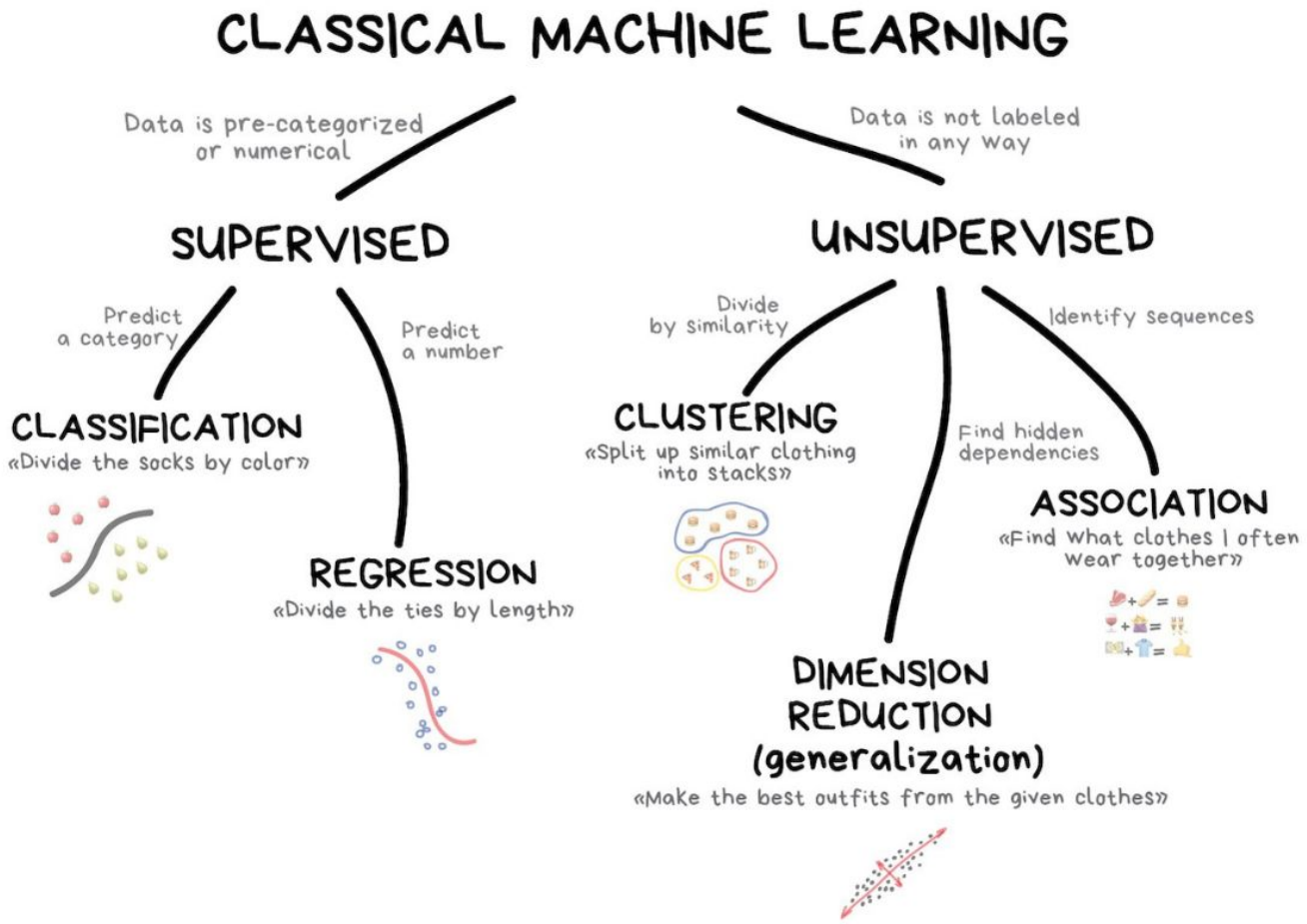
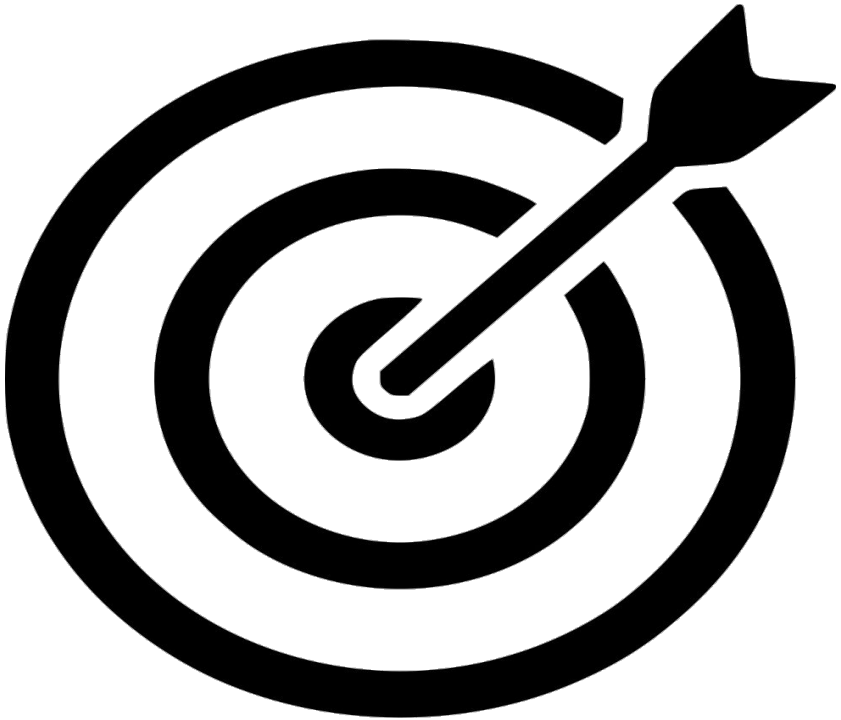


Image source: https://vas3k.com/blog/machine_learning/



model assessment

continuous variable prediction

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

A few outliers can lead to a big increase in RMSE, even if all the other predictions are pretty good

categorical variable prediction

continuous variable prediction

$$\text{Accuracy} = \frac{\#\text{ of samples predicted correctly}}{\#\text{ of samples predicted}} * 100$$

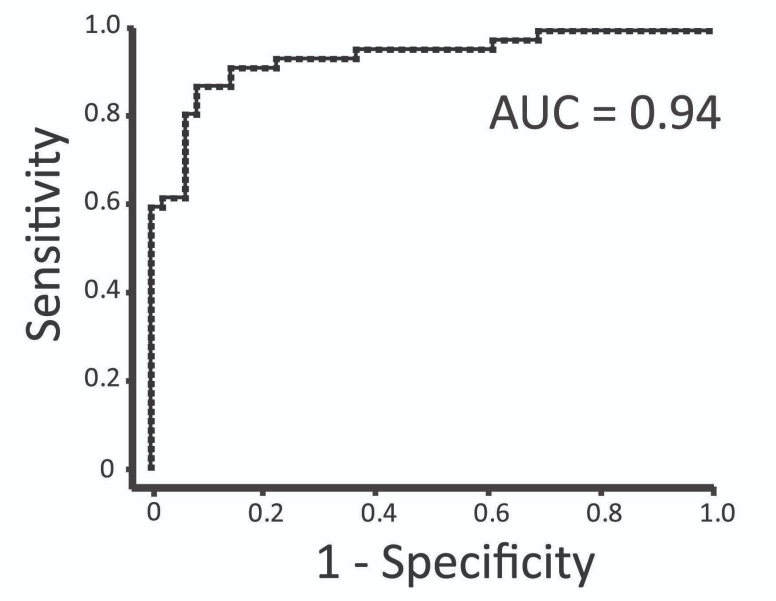
categorical variable prediction

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

A 2x2 table is a type of confusion matrix

Sensitivity

$$\frac{TP}{TP + FN}$$



Specificity

$$\frac{TN}{TN + FP}$$

Accuracy	What % were predicted correctly?
Sensitivity	Of those that *were* positives, what % were predicted to be positive?
Specificity	Of those that were *negatives*, what % were predicted to be negative?