

Machine learning

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

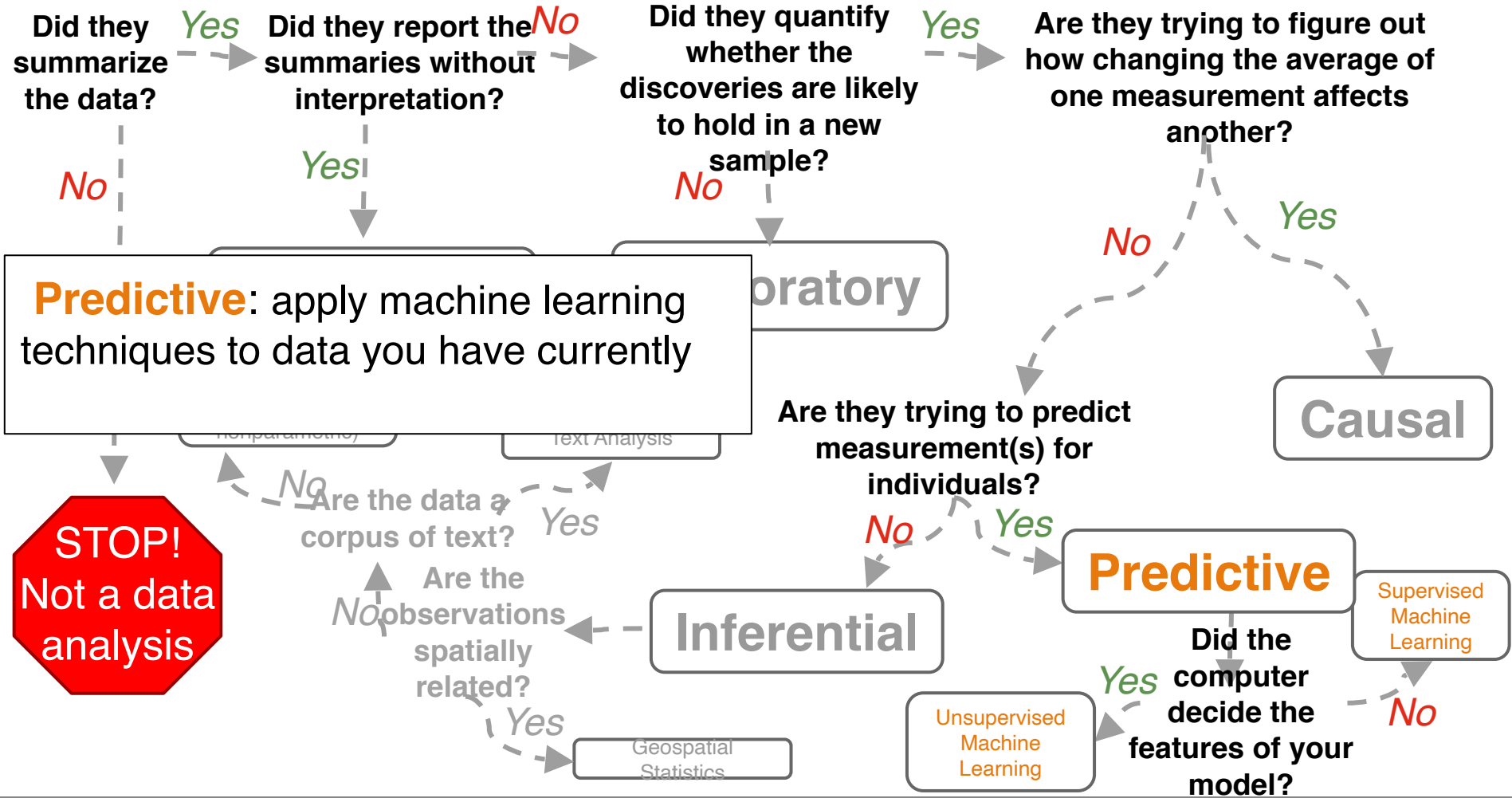
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>

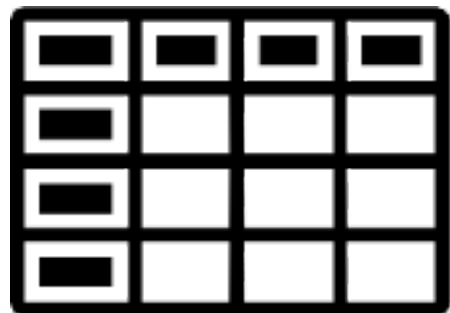


- **Problem:** Detecting whether credit card charges are fraudulent.
- **Data science question:** Can we use the time of the charge, the location of the charge, and the price of the charge to predict whether that charge is fraudulent or not?
- **Type of analysis:** Predictive analysis



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

machine learning
approaches are used for
predictive analysis!



data

train →



model

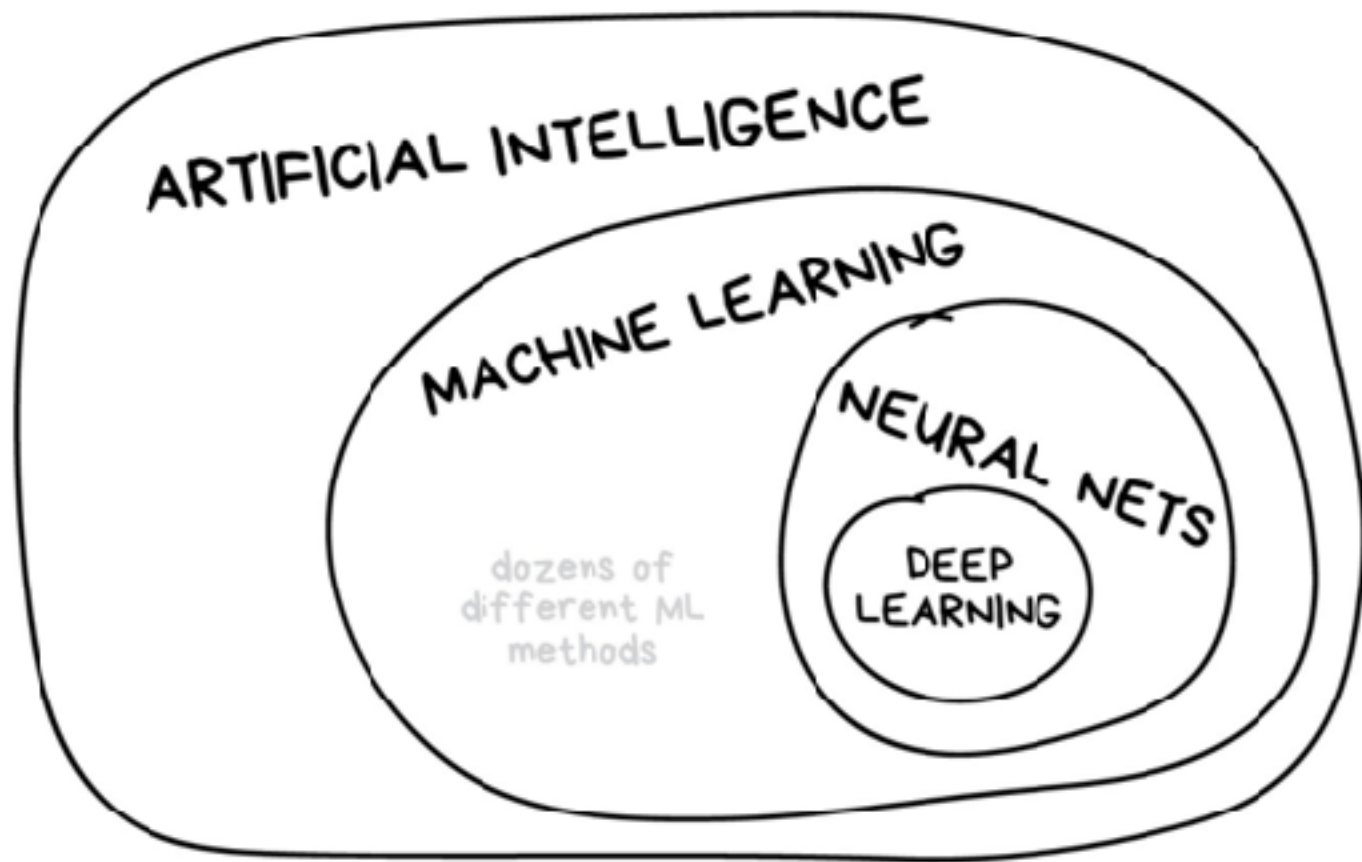
predict →



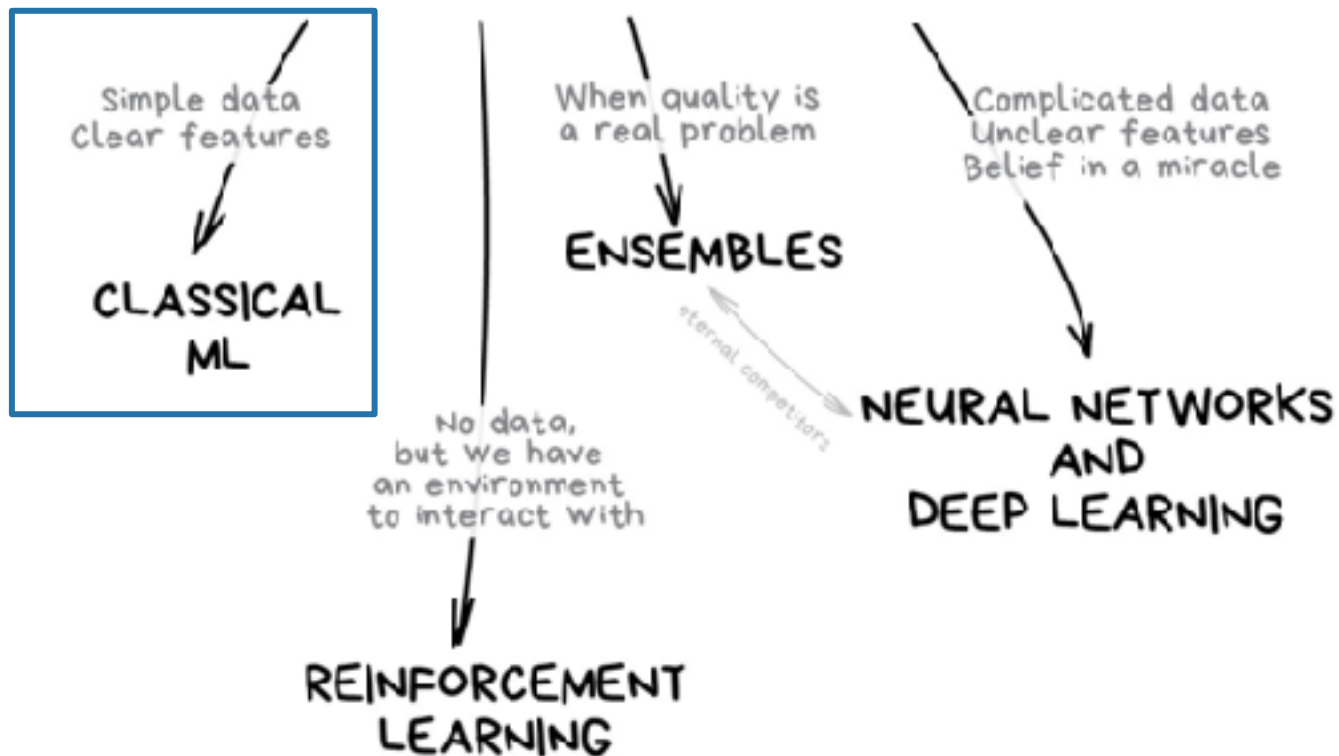
What is machine learning?

“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



THE MAIN TYPES OF MACHINE LEARNING





Prediction Questions

Which of these
questions is most
appropriate for
machine learning?

A How common is watching Sesame Street in the US?

B What is the effect of watching Sesame Street on children's brains?

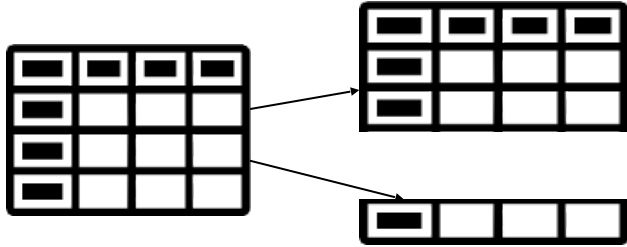
C What is the relationship between early childhood educational programming and success in elementary school?

D Can we use information about one's early childhood to predict their success in elementary school?

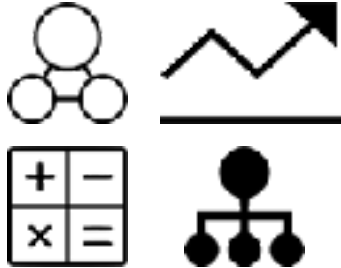
E How does Sesame Street cause an increase in educational attainment?

Machine Learning Generalizations

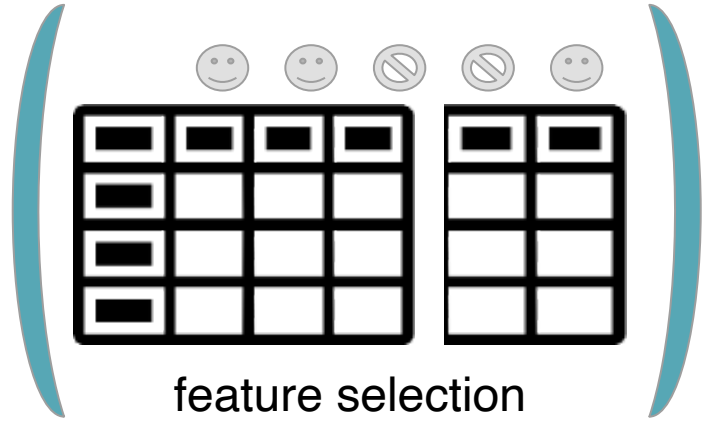
Basic Steps to Prediction



data
partitioning



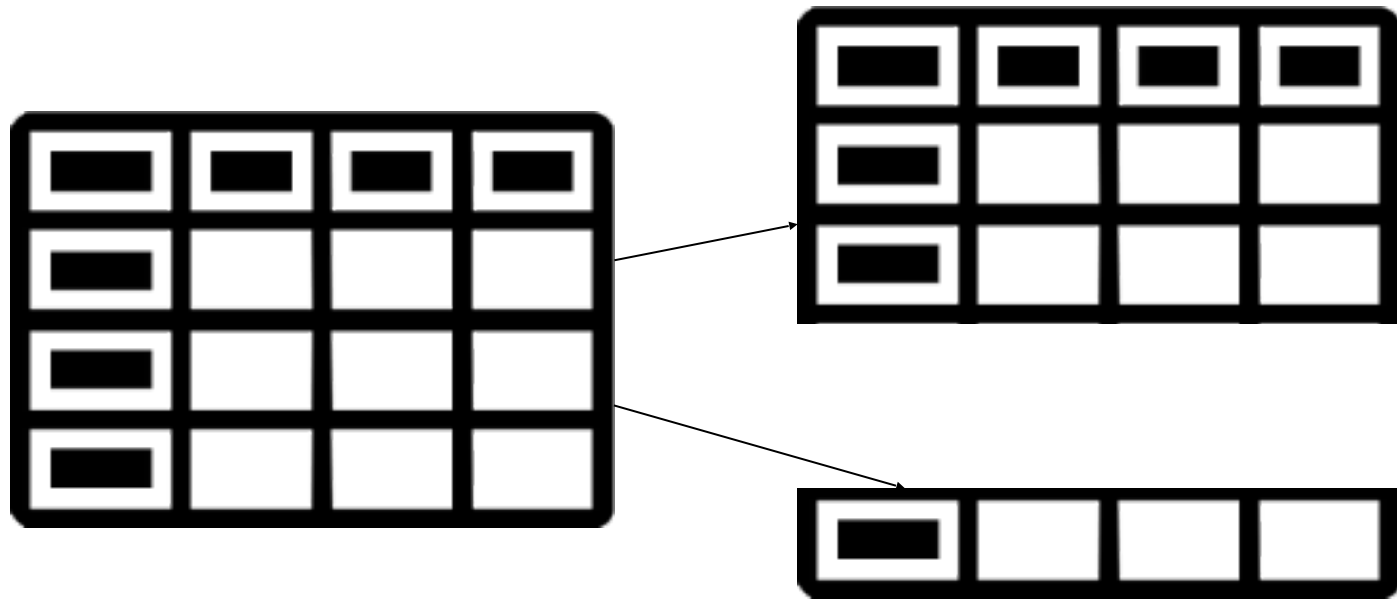
model selection



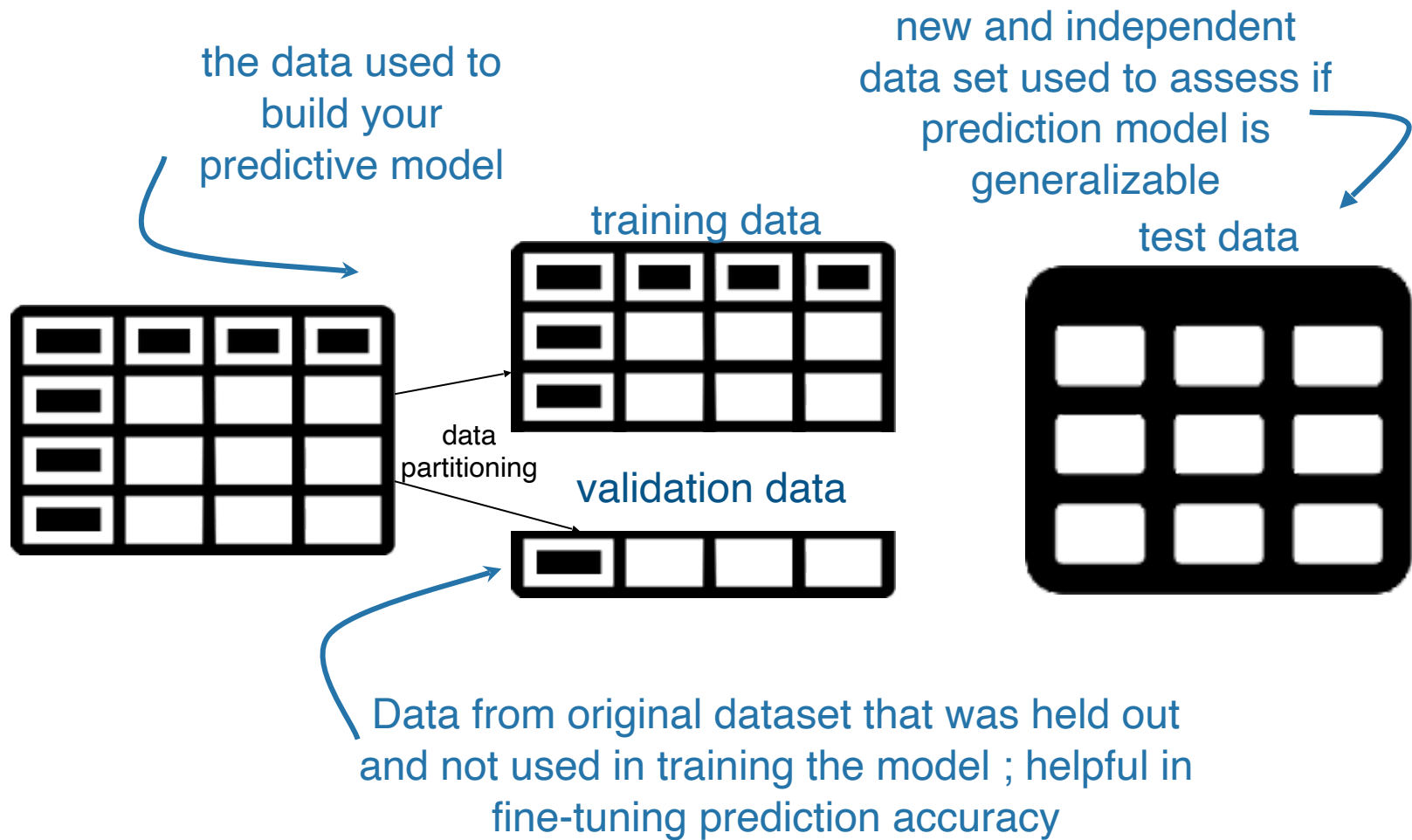
feature selection



model assessment



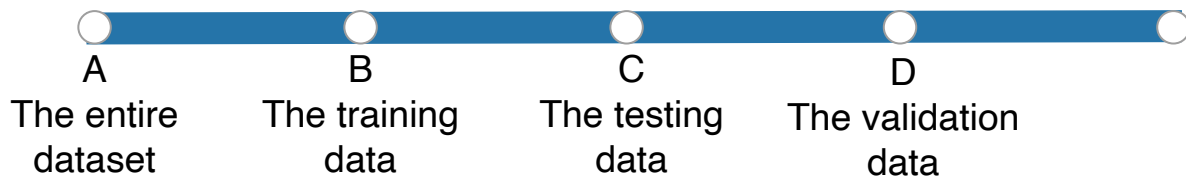
data partitioning

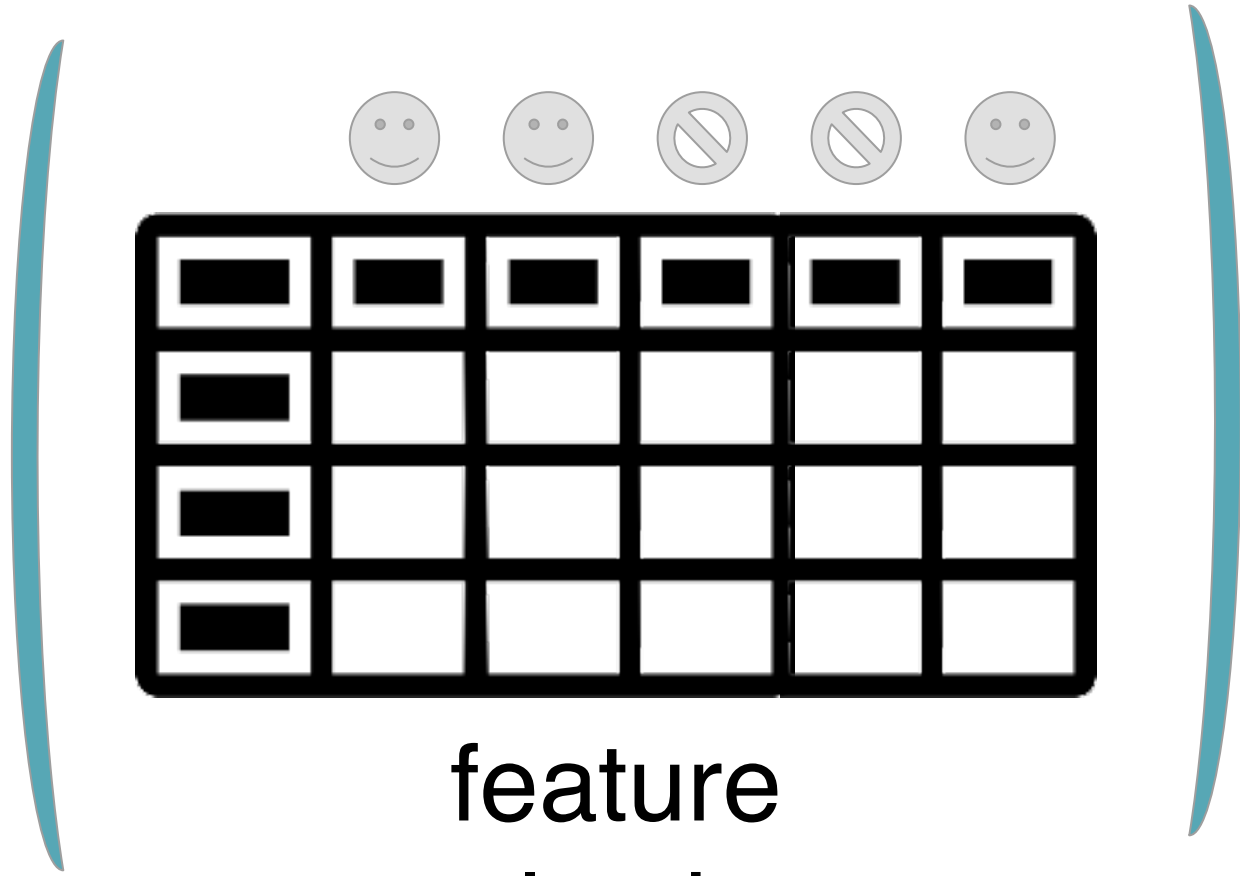


Data Partitioning



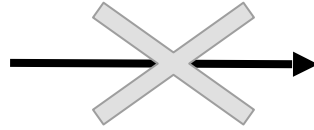
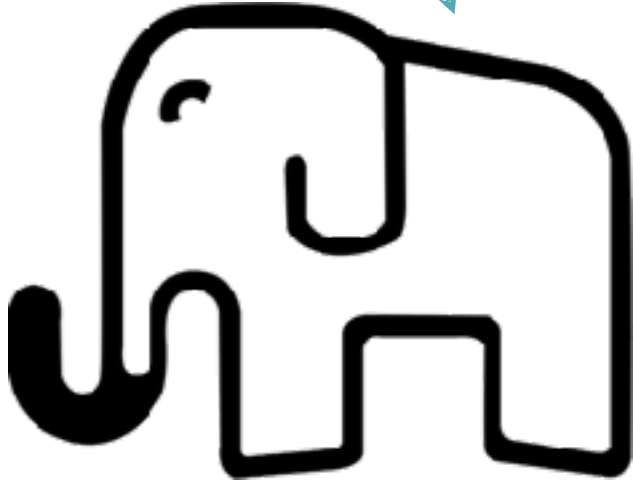
What portion of the data are typically used for generating the model?

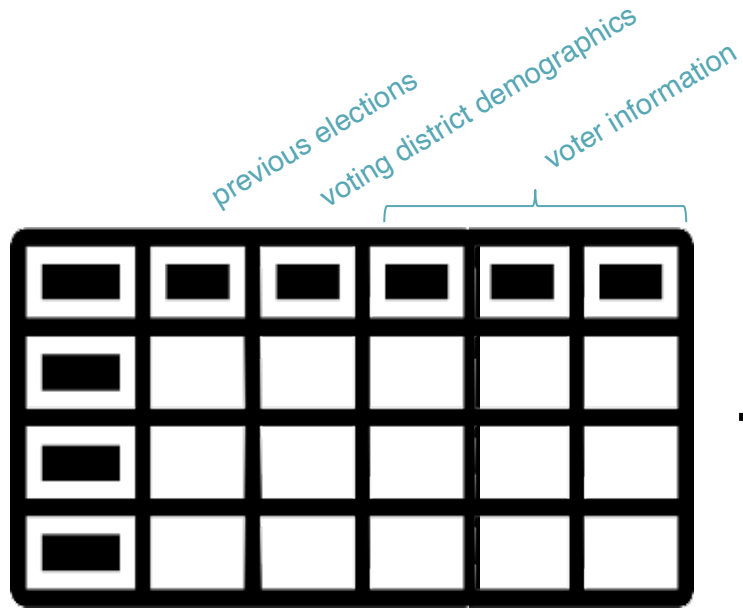




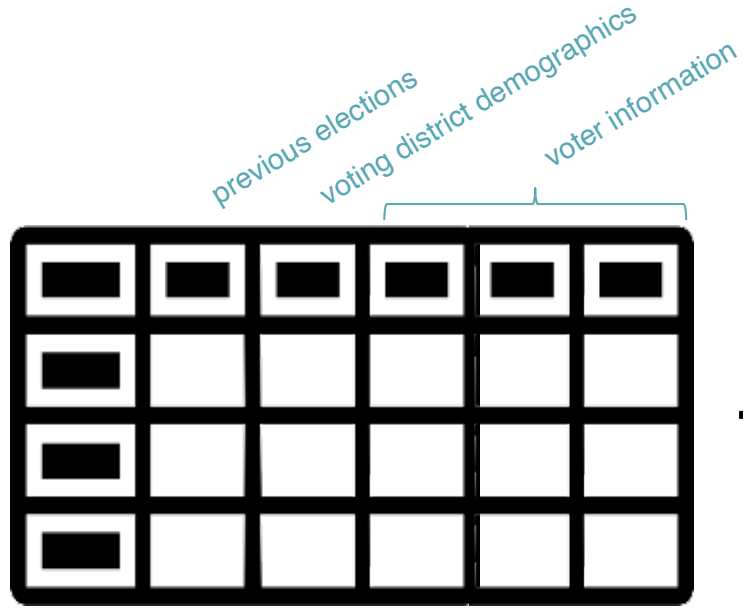
feature
selection

elephant height data
are likely not predictive
of US elections



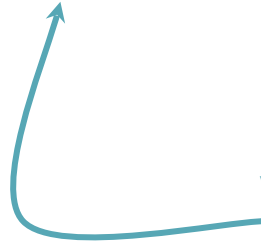


these data are likely
predictive of US
election outcomes



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

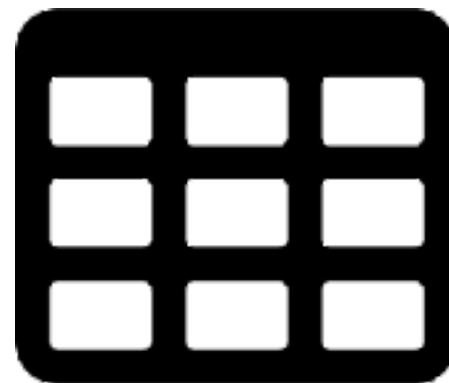
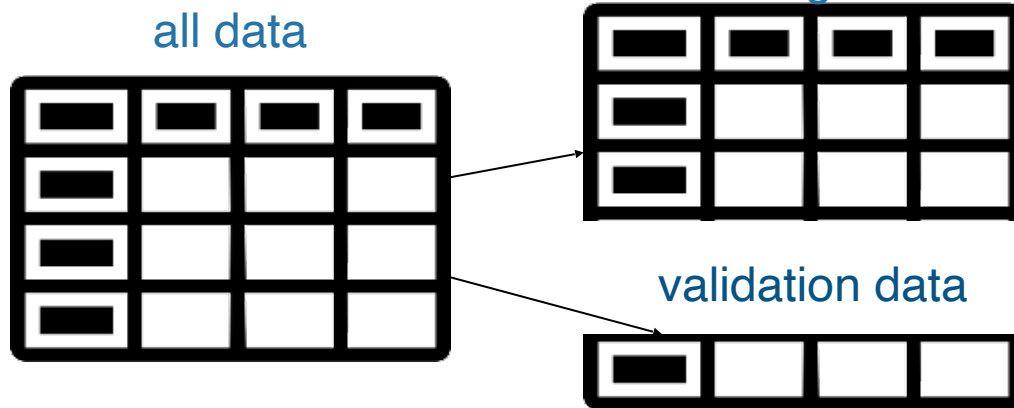
■	■	■	■	■	■
■					
■					
■					



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

Try different
feature sets

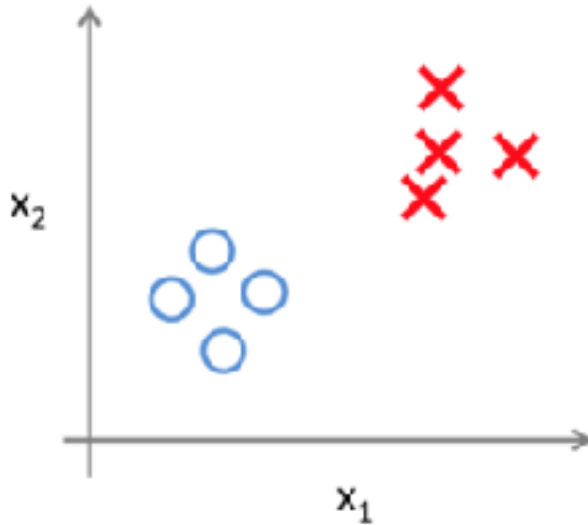
Evaluate best
feature set on
test data



Use validation set to select the features!

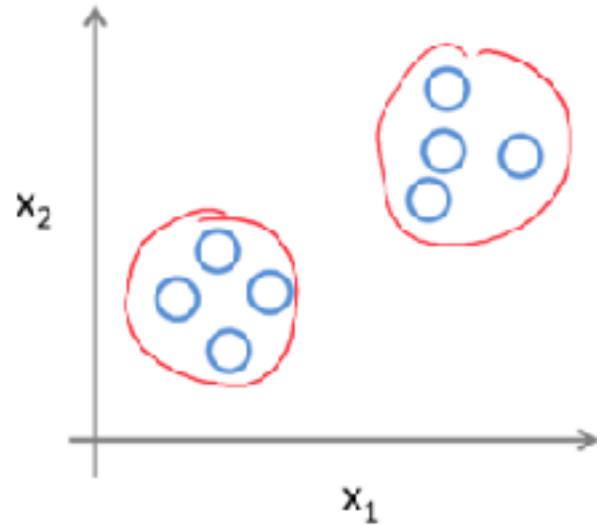
Two modes of machine learning

Supervised Learning



You tell the computer what features to use to classify the observations

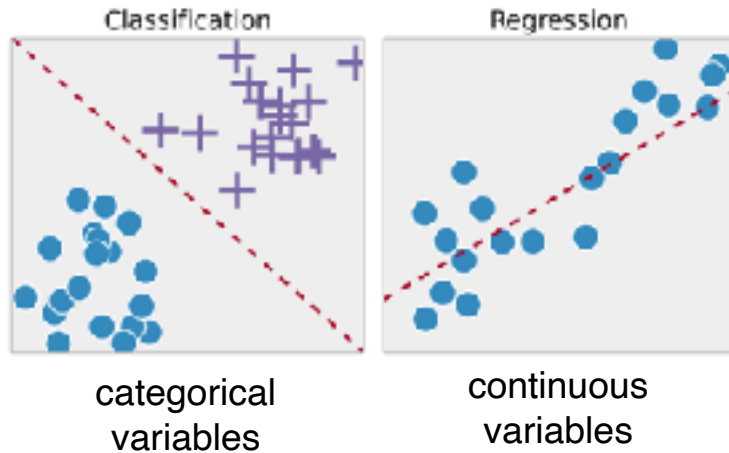
Unsupervised Learning



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

Approaches to machine learning

Supervised Learning



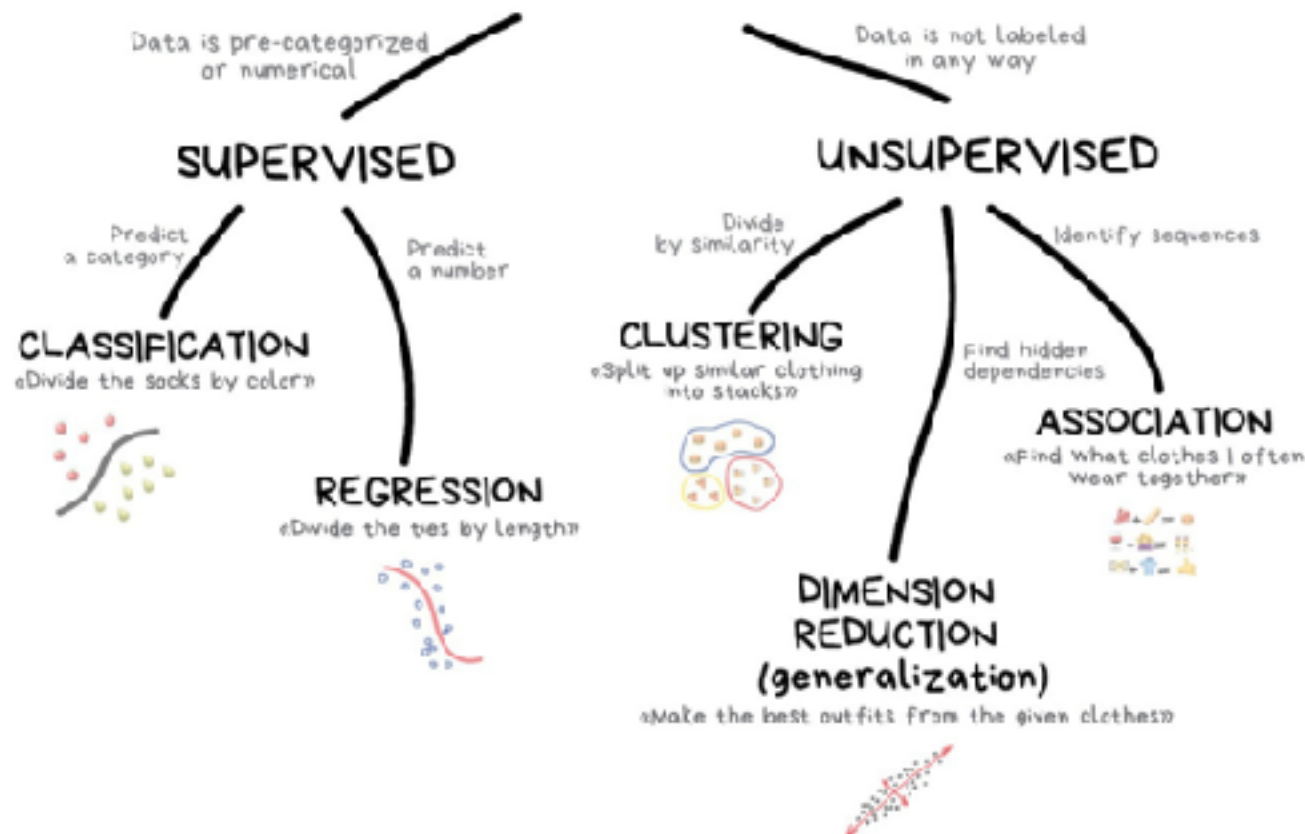
Unsupervised Learning

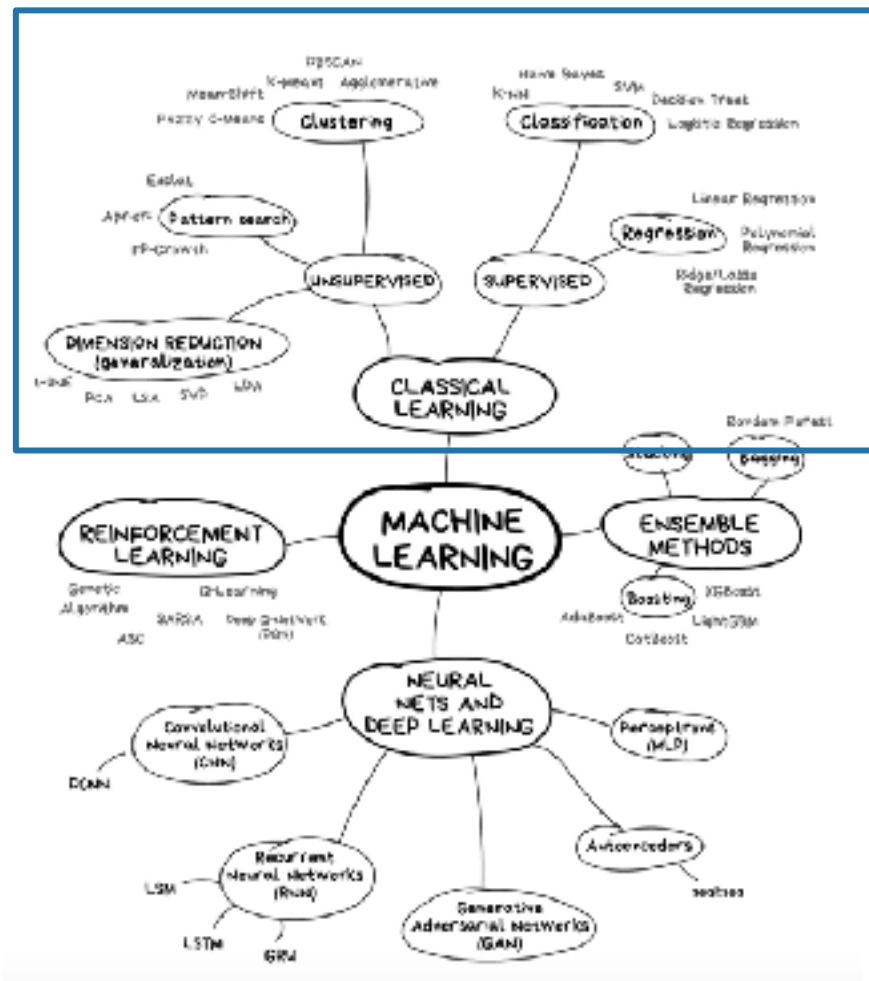


Clustering (categorical)
& dimensionality reduction (continuous)

can automatically identify
structure in data

CLASSICAL MACHINE LEARNING



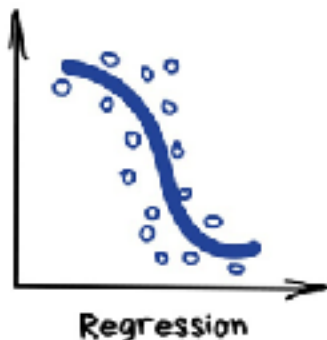


Regression

"Draw a line through these dots. Yep, that's the machine learning"

Today this is used for:

- Stock price forecasts
- Demand and sales volume analysis
- Medical diagnosis
- Any number-time correlations



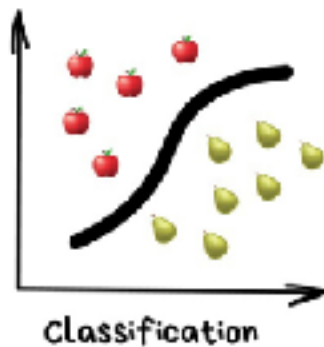
Popular algorithms are Linear and Polynomial regressions.

Classification

"Splits objects based at one of the attributes known beforehand. Separate socks by based on color, documents based on language, music by genre"

Today used for:

- Spam filtering
- Language detection
- A search of similar documents
- Sentiment analysis
- Recognition of handwritten characters and numbers
- Fraud detection



Popular algorithms: Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbours, Support Vector Machine



Regression:

predicting continuous variables
(i.e. Age)

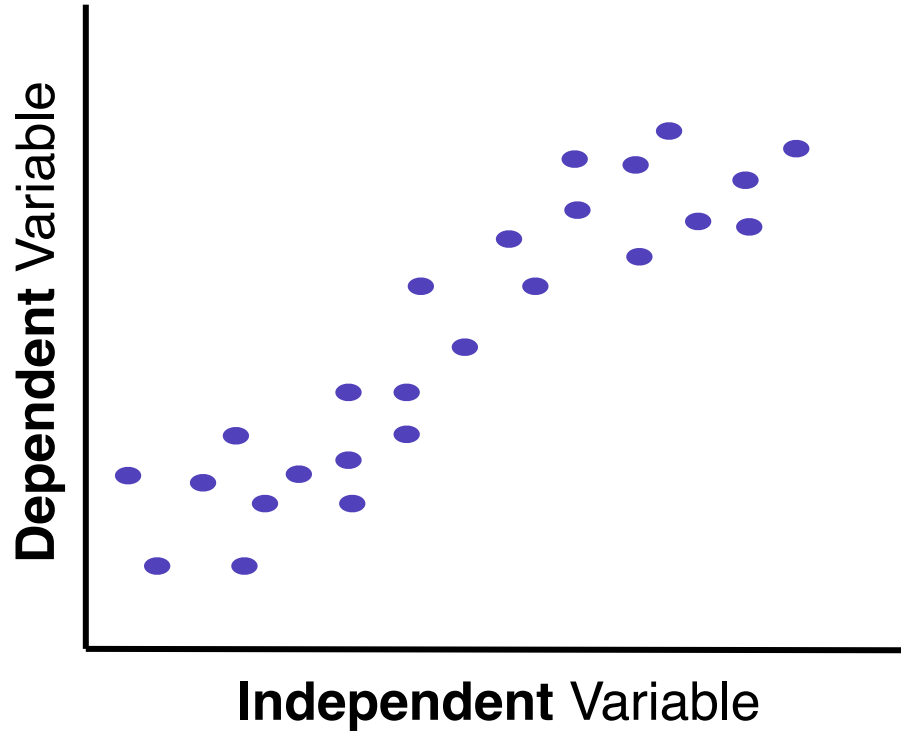
continuous variable prediction

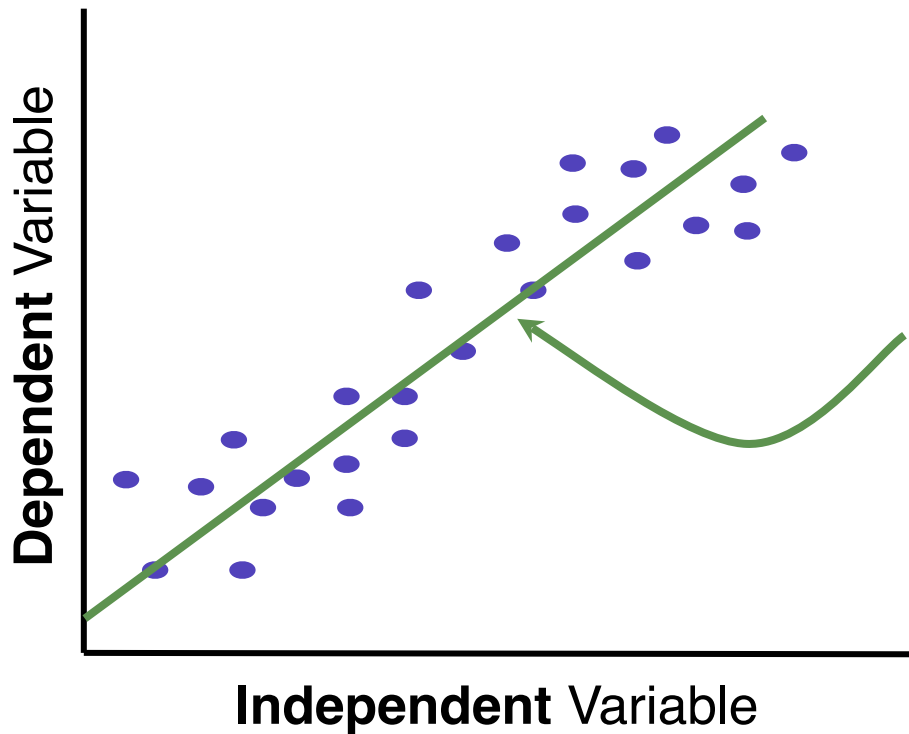


Classification:

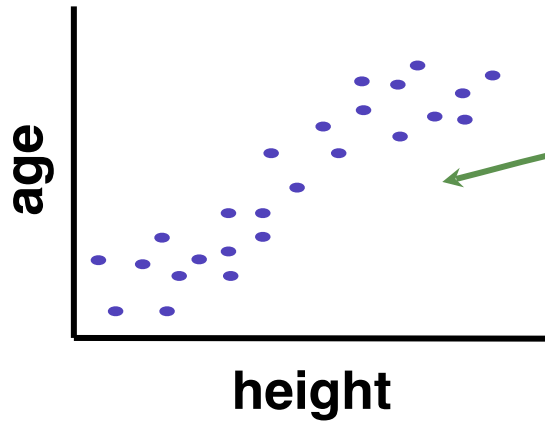
predicting categorical variables
(i.e. education level)

categorical variable prediction





We'll use the
linear relationship
between variables
to generate a
predictive model



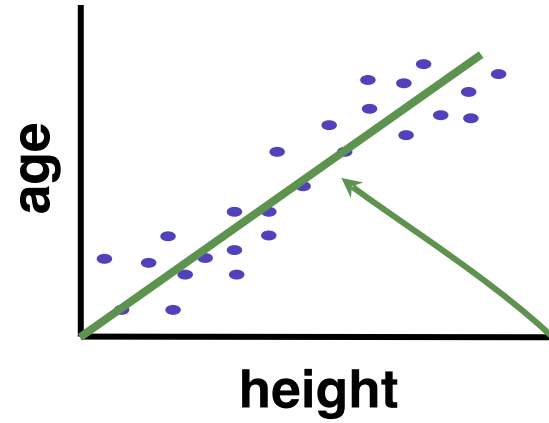
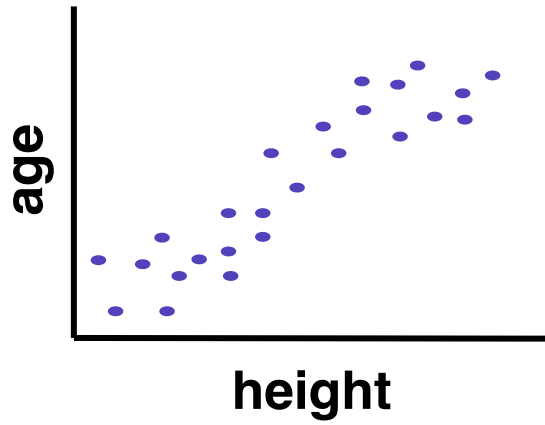
the training data
will be used to
build the
predictive
model

Supervised Learning

regression

continuous variable prediction





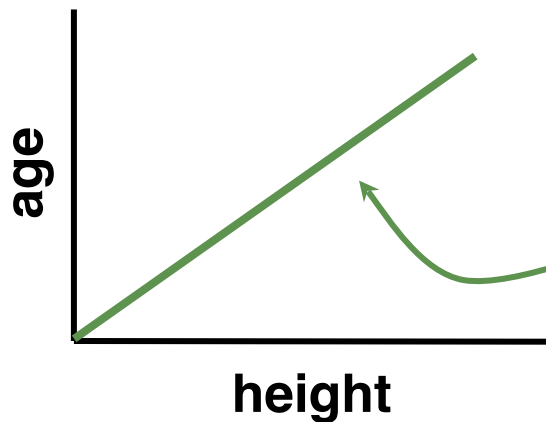
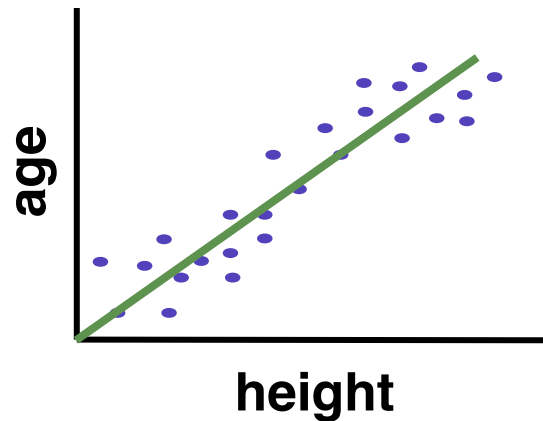
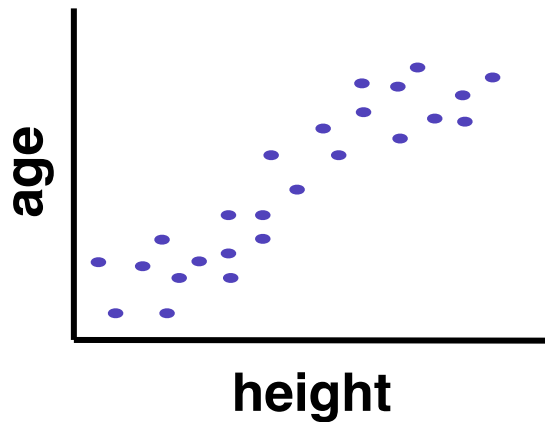
use linear
regression to
model the
relationship

Supervised Learning

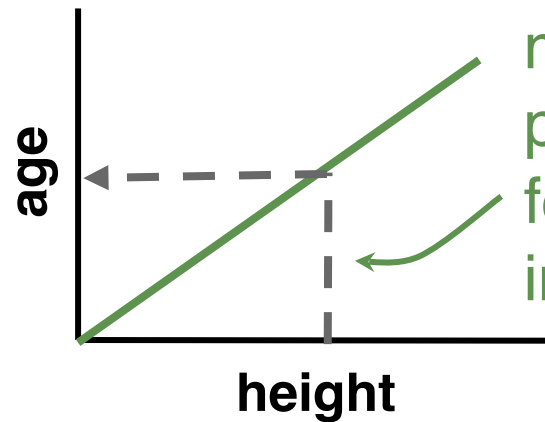
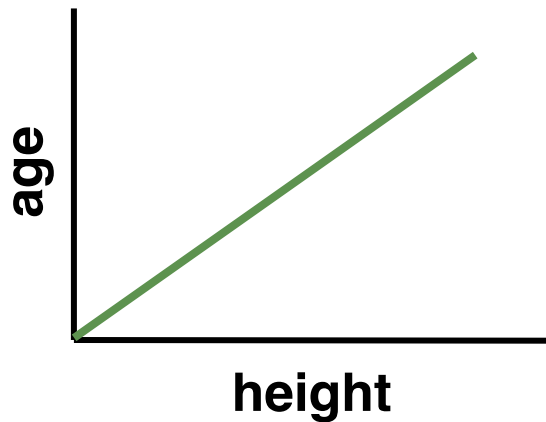
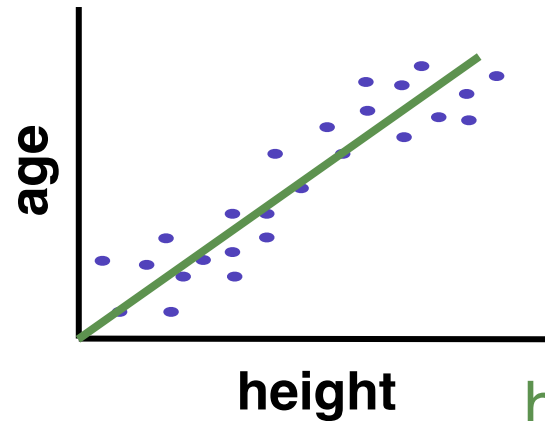
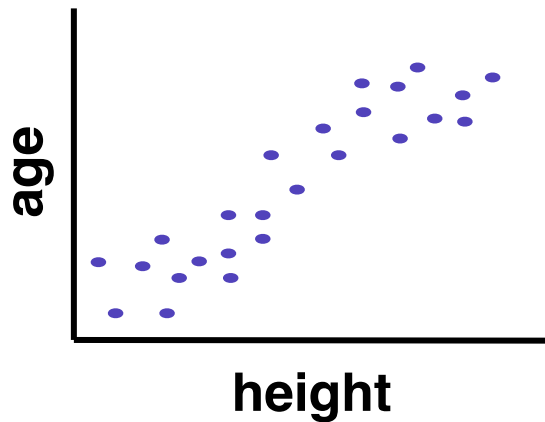
regression

continuous variable prediction

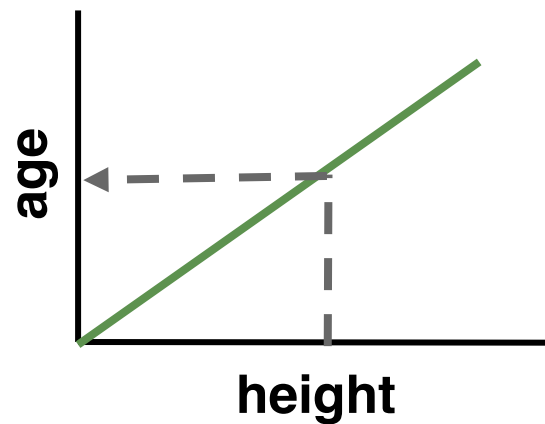
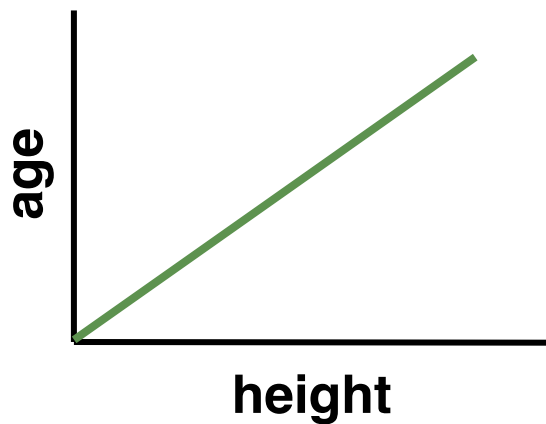
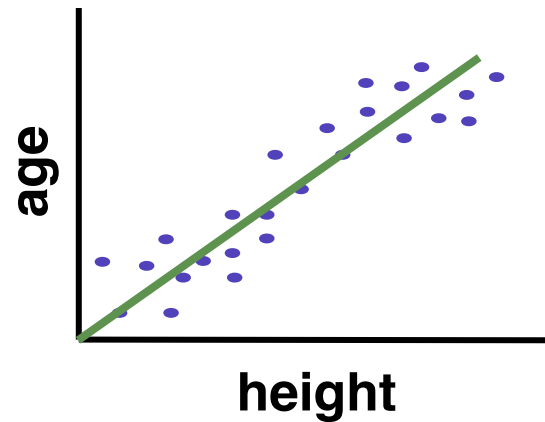
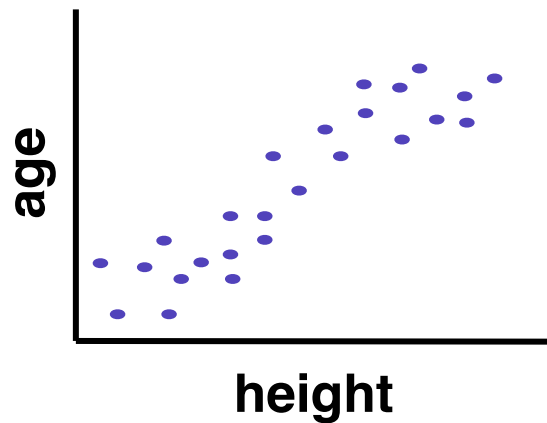




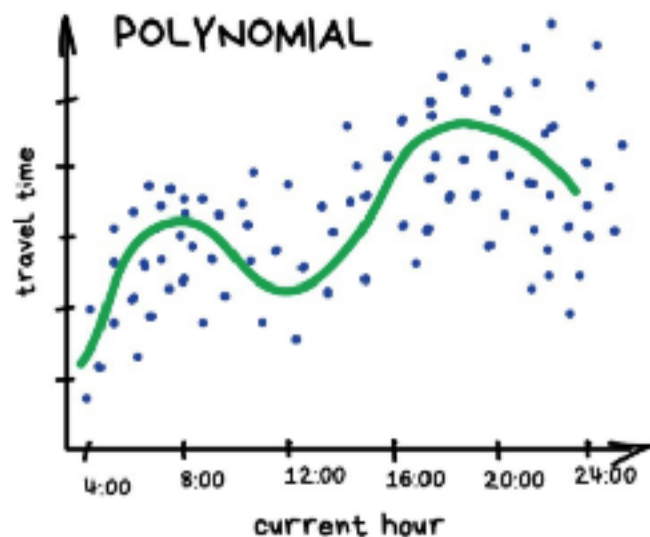
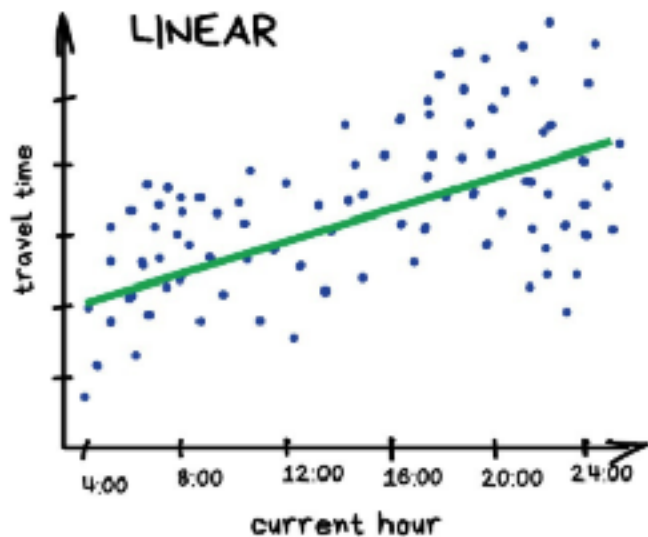
For prediction, the individual values in the training data are *not* important. We only need the model.



how we'll
make
predictions
for a future
individual



PREDICT TRAFFIC JAMS



REGRESSION

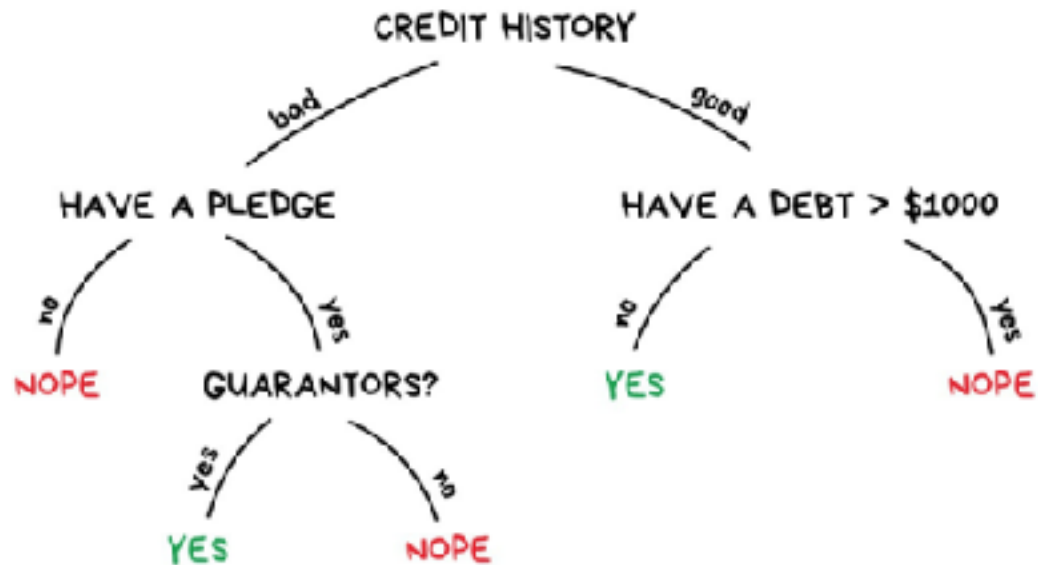


Regression:
predicting continuous
variables
(i.e. Age)



Classification:
predicting categorical
variables
(i.e. give a loan?)

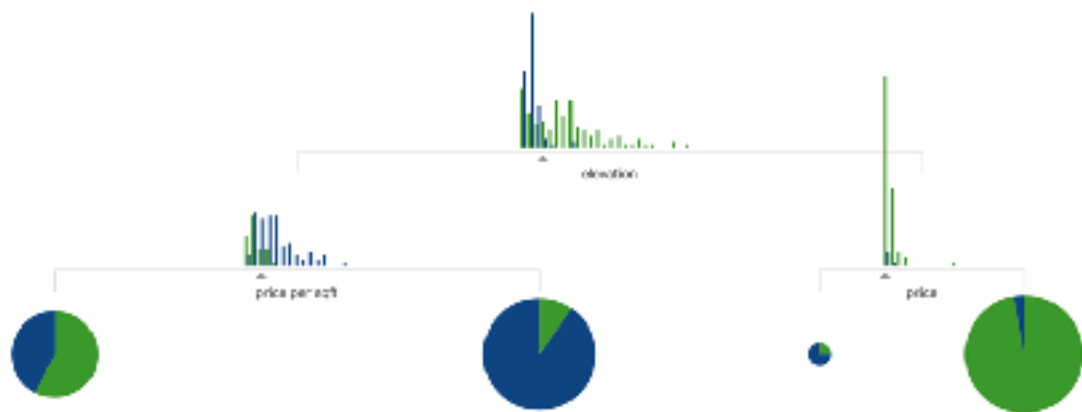
GIVE A LOAN?



DECISION TREE

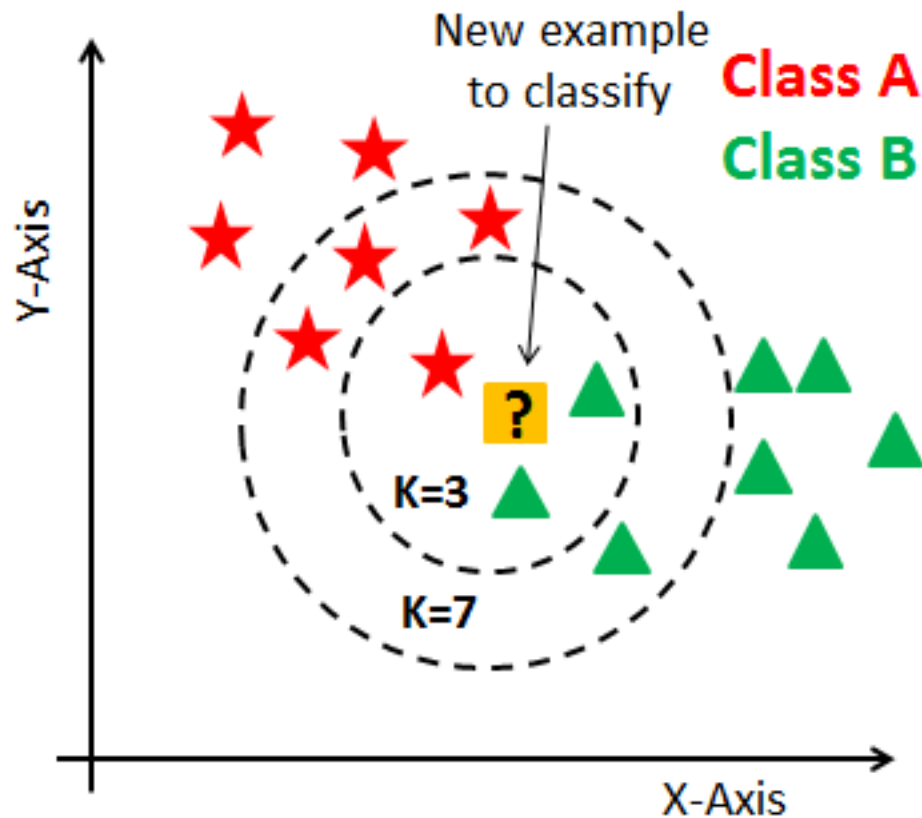
Growing a tree

Additional forks will add new information that can increase a tree's **prediction accuracy**.

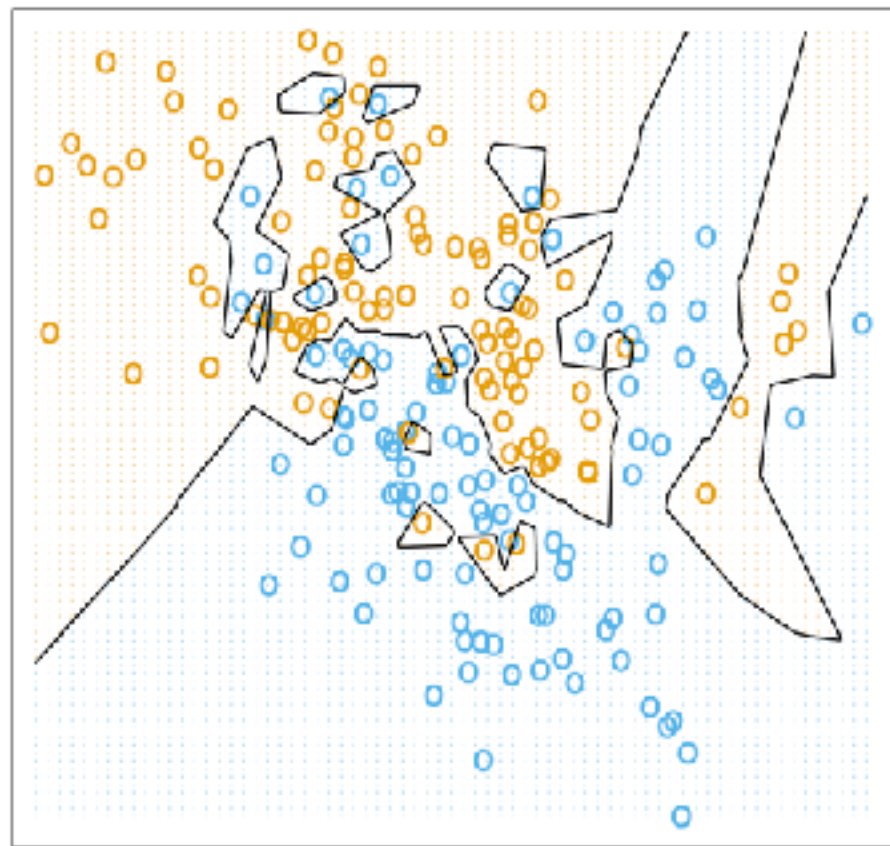


<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

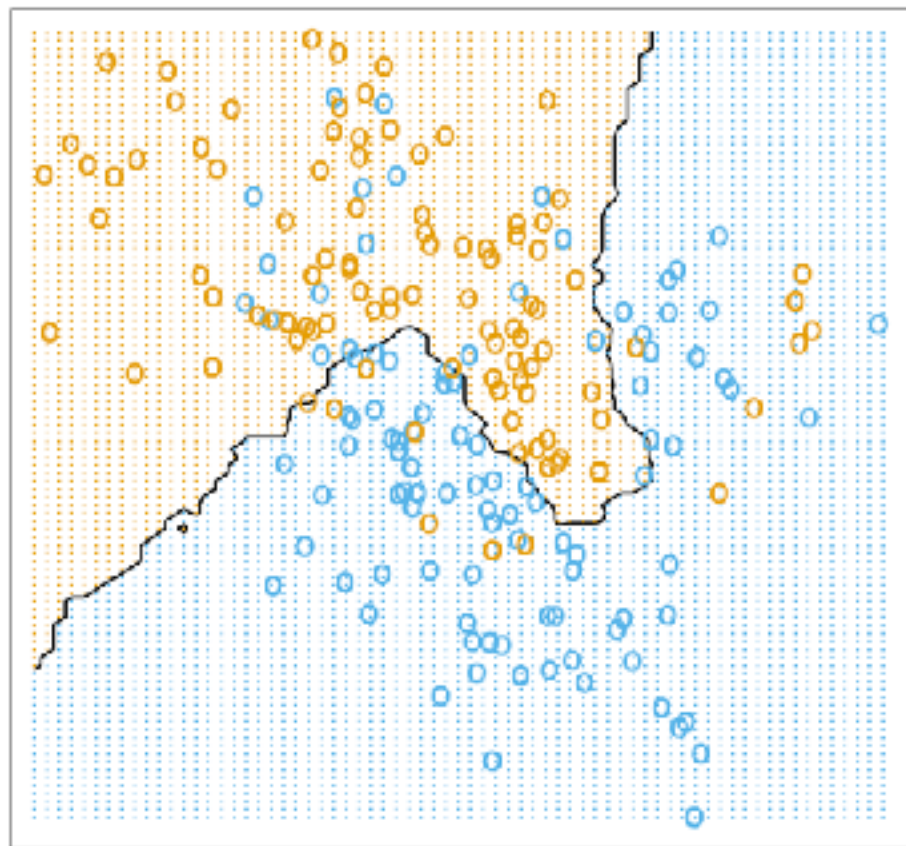
K-nearest neighbors



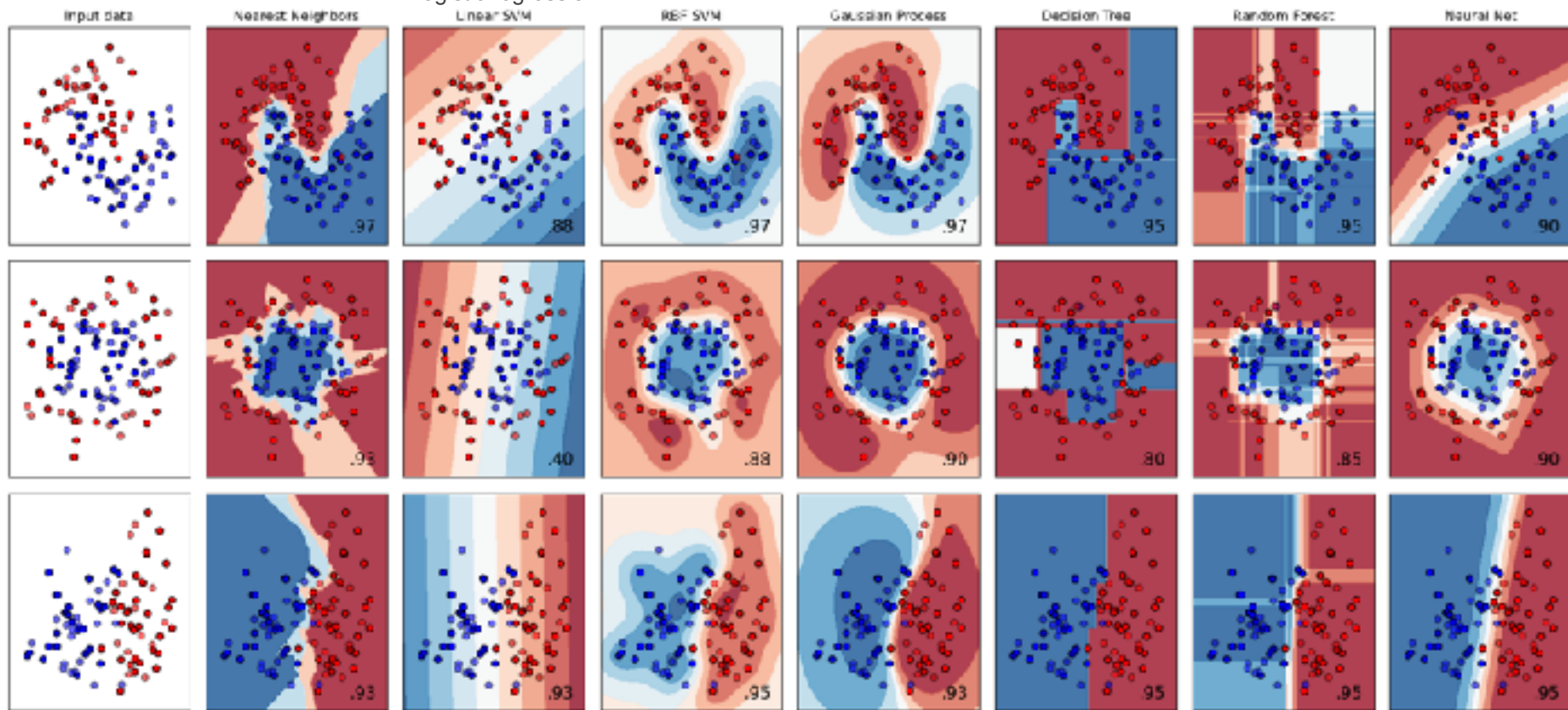
1-Nearest Neighbor Classifier



15-Nearest Neighbor Classifier



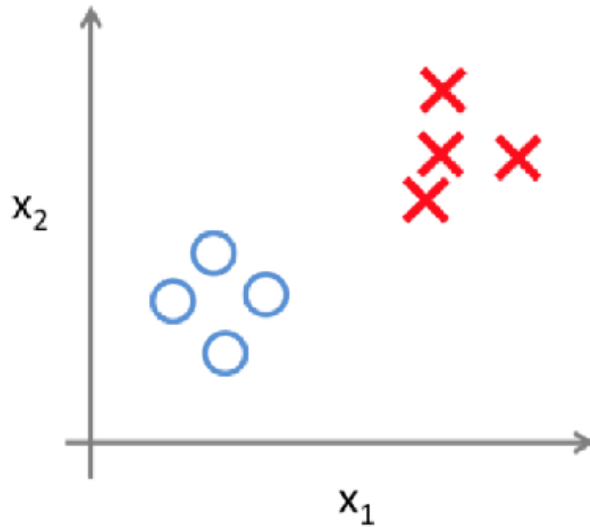
Logistic regression



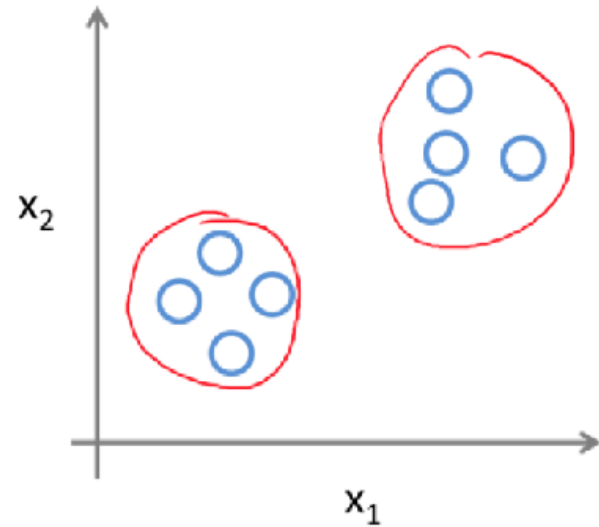
Unsupervised Learning

To modes of machine learning

Supervised Learning



Unsupervised Learning



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

Dimensionality Reduction (Generalization)

Clustering

"Divides objects based on unknown features.
Machine chooses the best way"

Nowadays used:

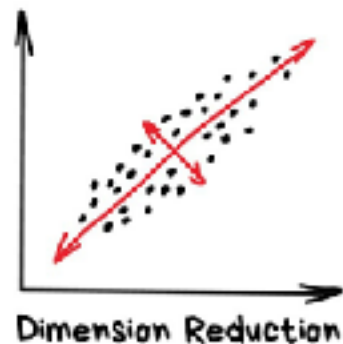
- For market segmentation (types of customers, loyalty)
- To merge close points on a map
- For image compression
- To analyze and label new data
- To detect abnormal behavior



"Assembles specific features into more high-level ones"

Nowadays is used for:

- Recommender systems (★)
- Beautiful visualizations
- Topic modeling and similar document search
- Fake image analysis
- Risk management



Popular algorithms: K-means, clustering, Mean-Shift, DBSCAN

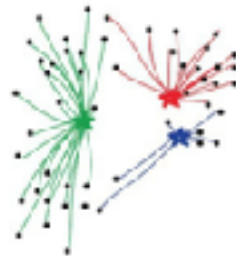
Popular algorithms: Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA, pLSA, cLSA), t-SNE (for visualization)

PUT KEBAB KIOSKS IN THE OPTIMAL WAY

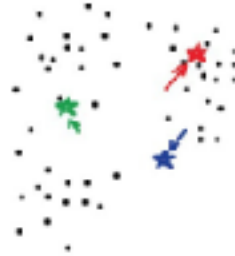
(also illustrating the K-means method)



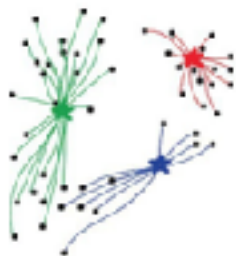
1. Put kebab kiosks in random places in city



2. Watch how buyers choose the nearest one



3. Move Kiosks closer to the centers of their popularity



4. Watch and move again



5. Repeat a million times



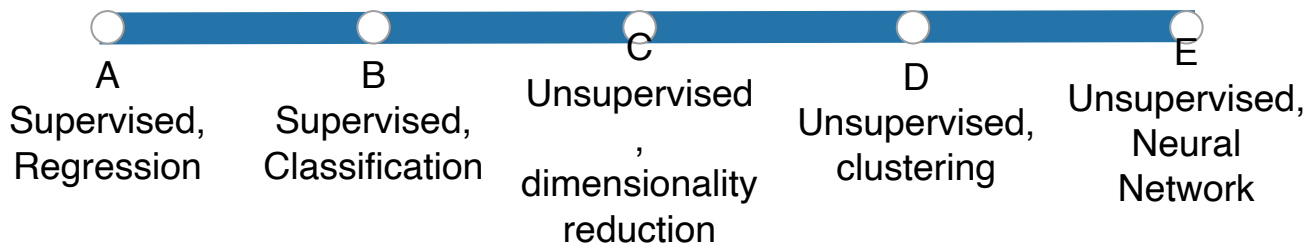
6. Done!
You're god of kebabs!

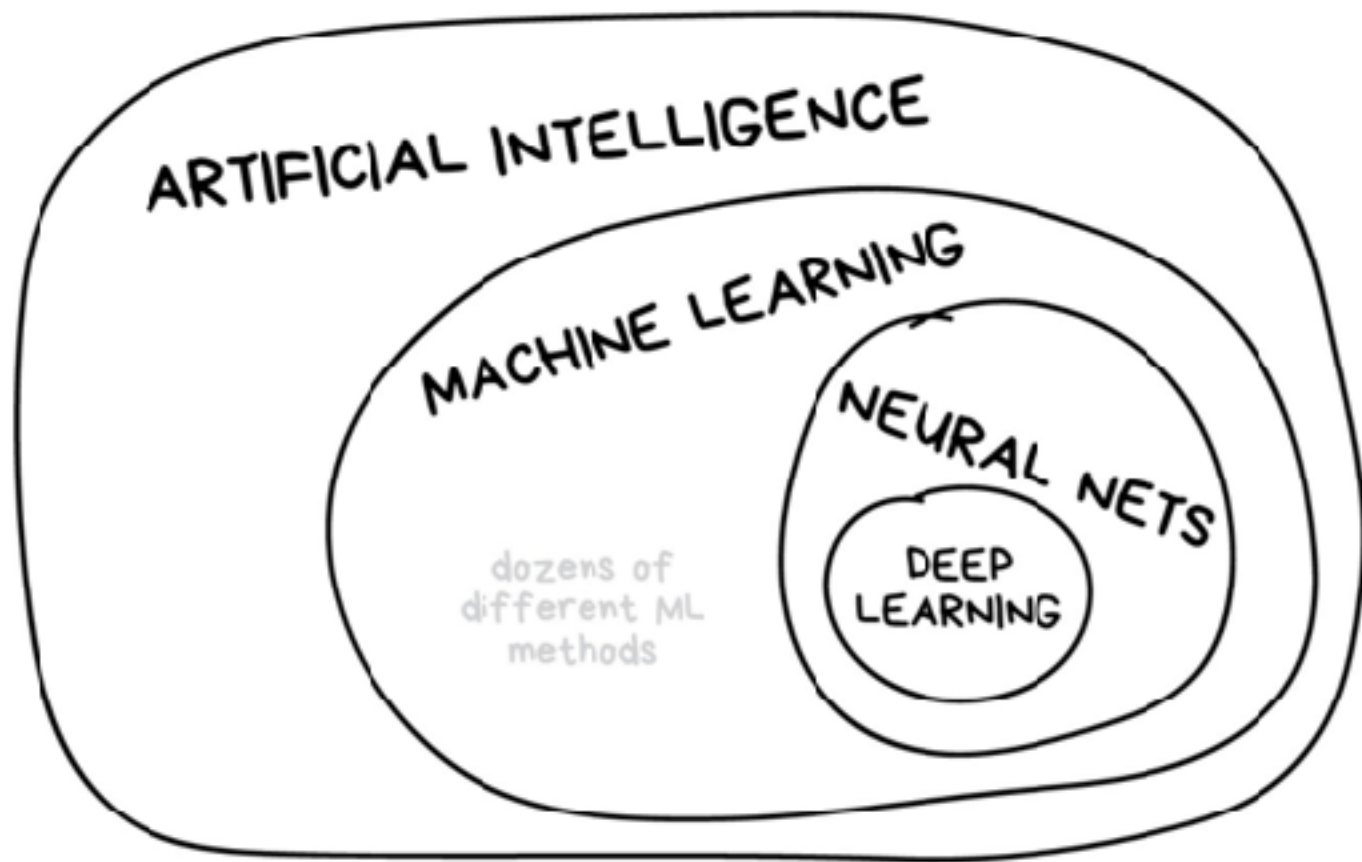


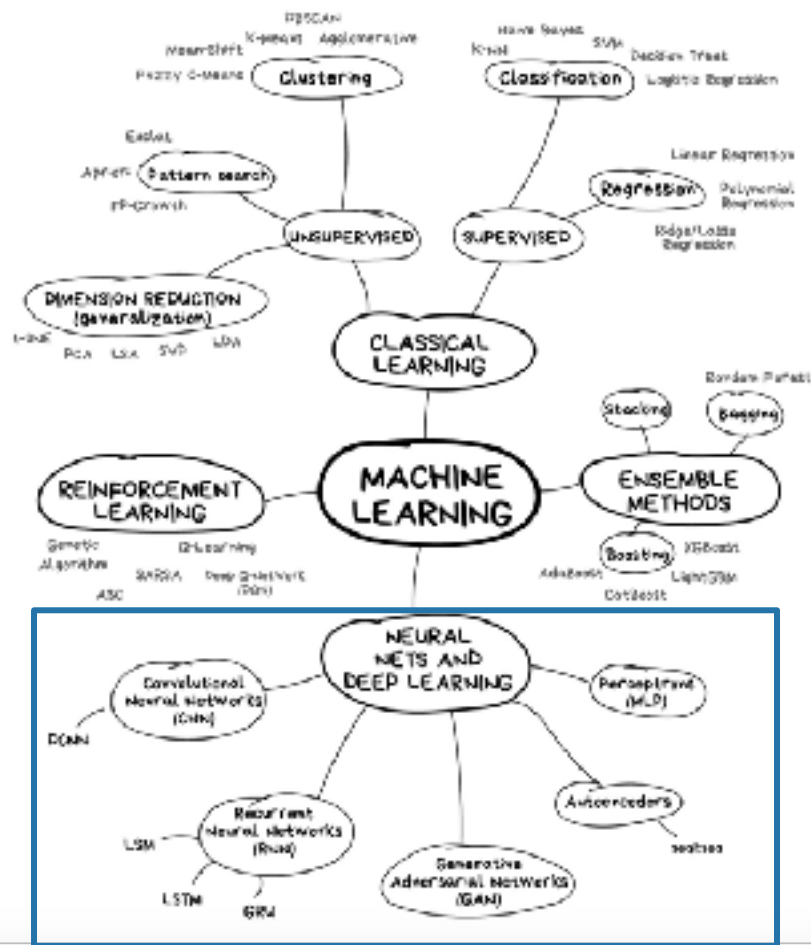
Prediction Approach

You want to predict someone's emotion based on an image.

How would you approach this with machine learning?



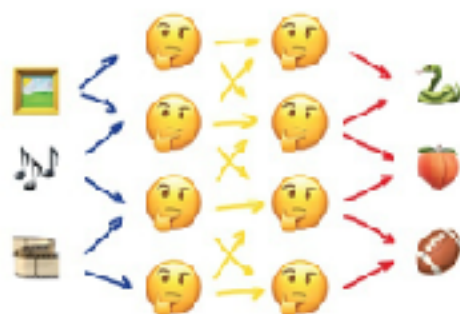




"We have a thousand-layer network, dozens of video cards, but still no idea where to use it. Let's generate cat pics!"

Used today for:

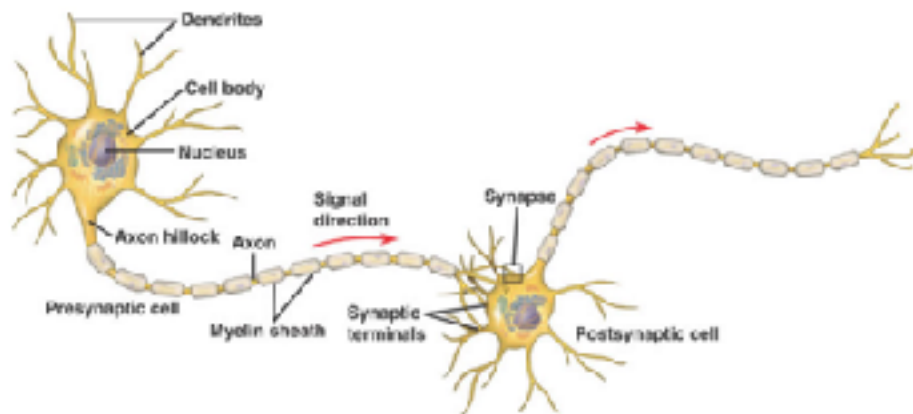
- Replacement of all algorithms above
- Object identification on photos and videos
- Speech recognition and synthesis
- Image processing, style transfer
- Machine translation



Neural Networks

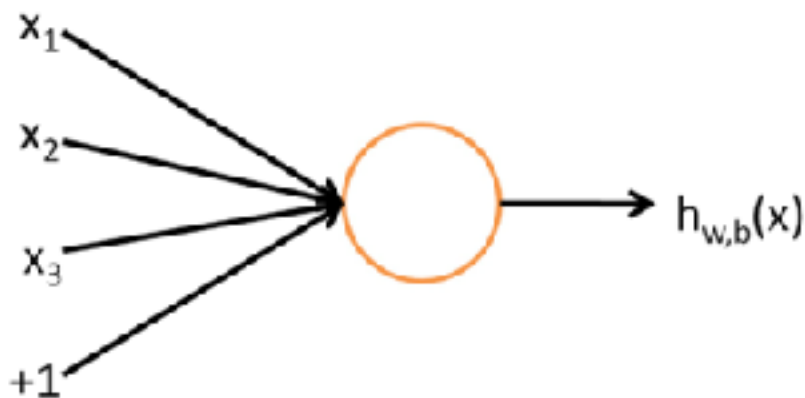
Popular architectures: Perceptron, Convolutional Network (CNN), Recurrent Networks (RNN), Autoencoders

WHAT IS A NEURON?



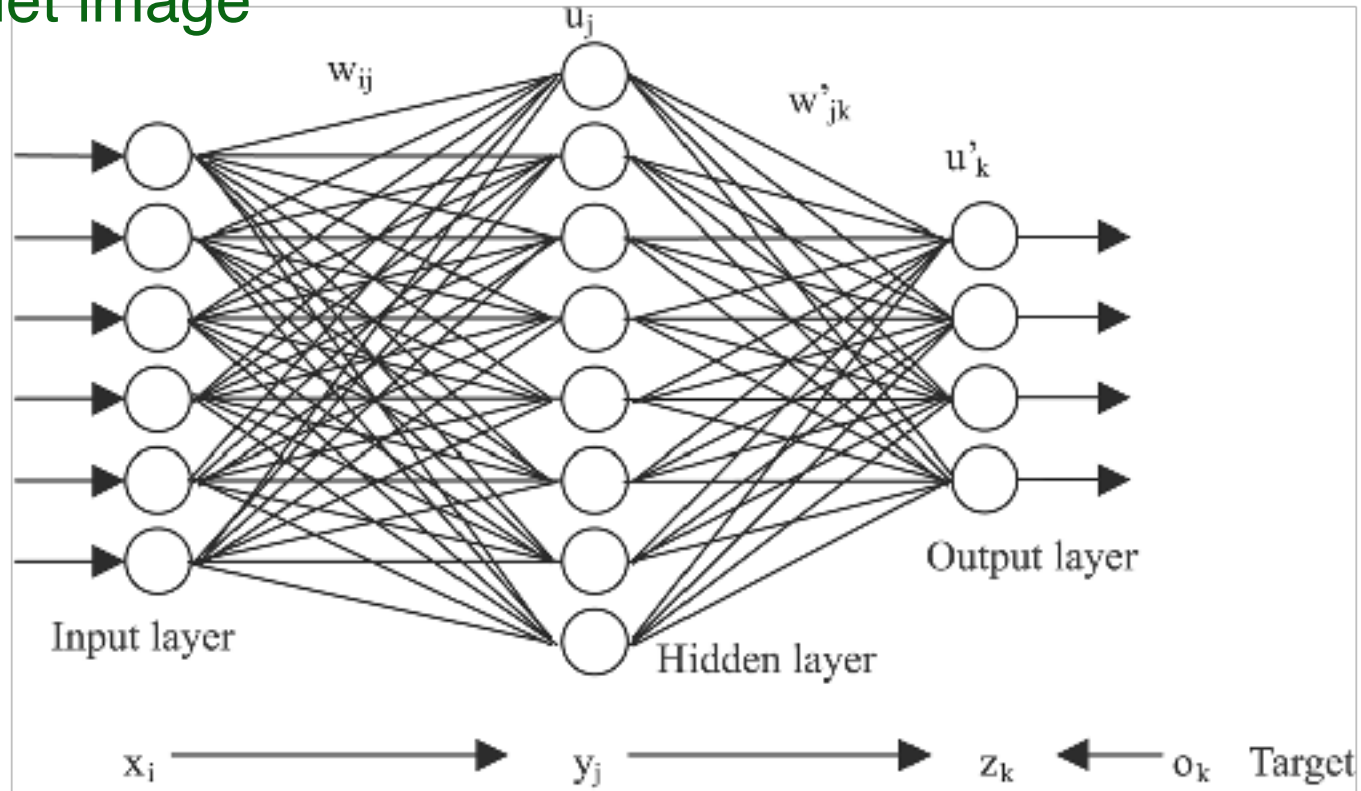
- Receives signal on synapse
- When trigger sends signal on axon

MATHEMATICAL NEURON

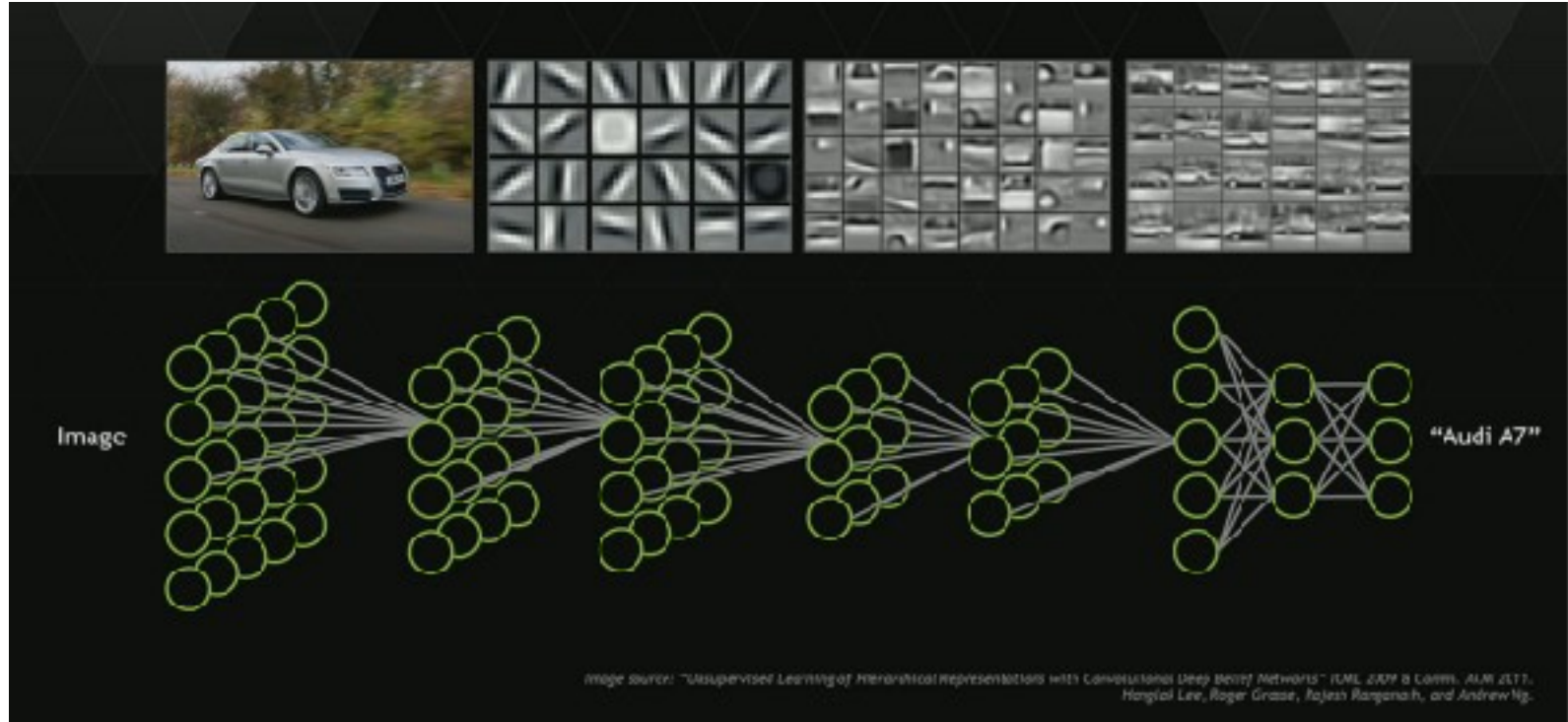


- Mathematical abstraction, inspired by biological neuron
- Either on or off based on sum of input

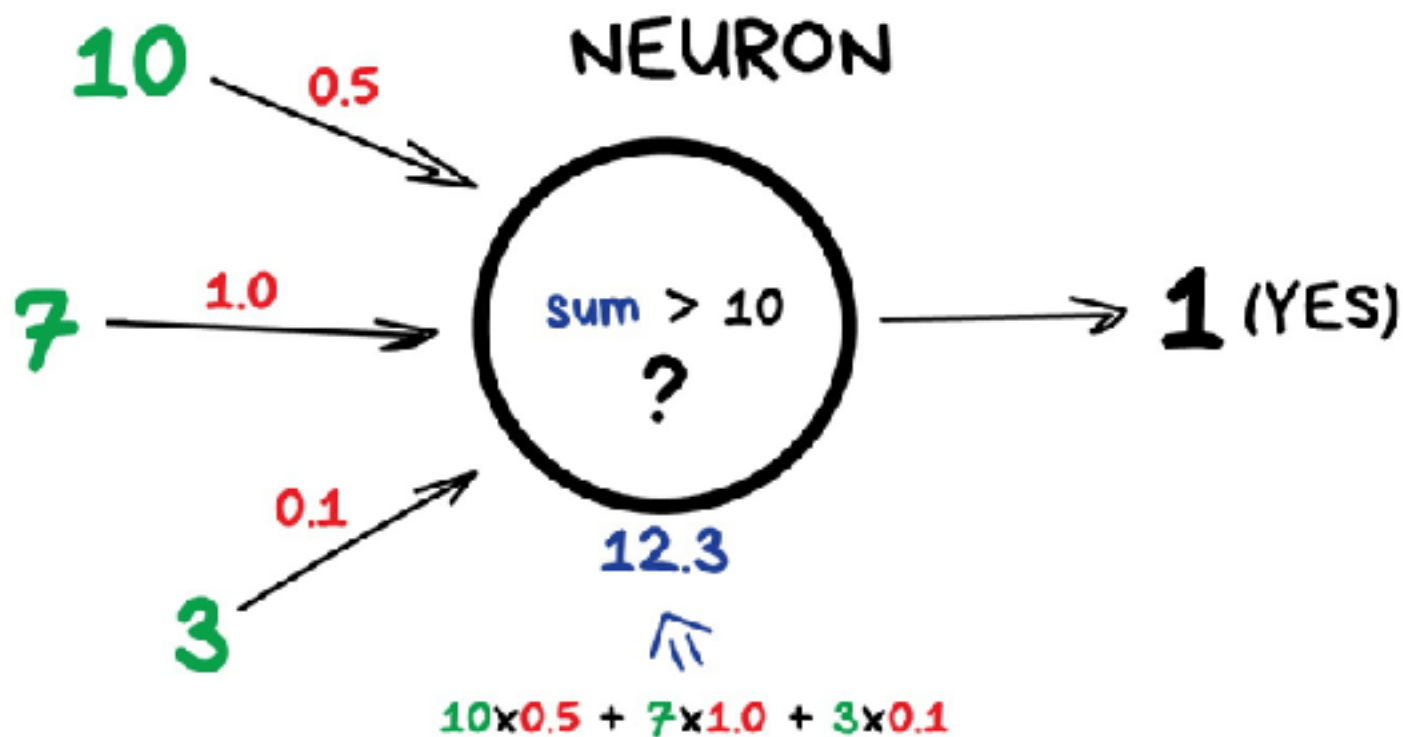
This will likely not be the last time you see this (mostly unhelpful) neural net image

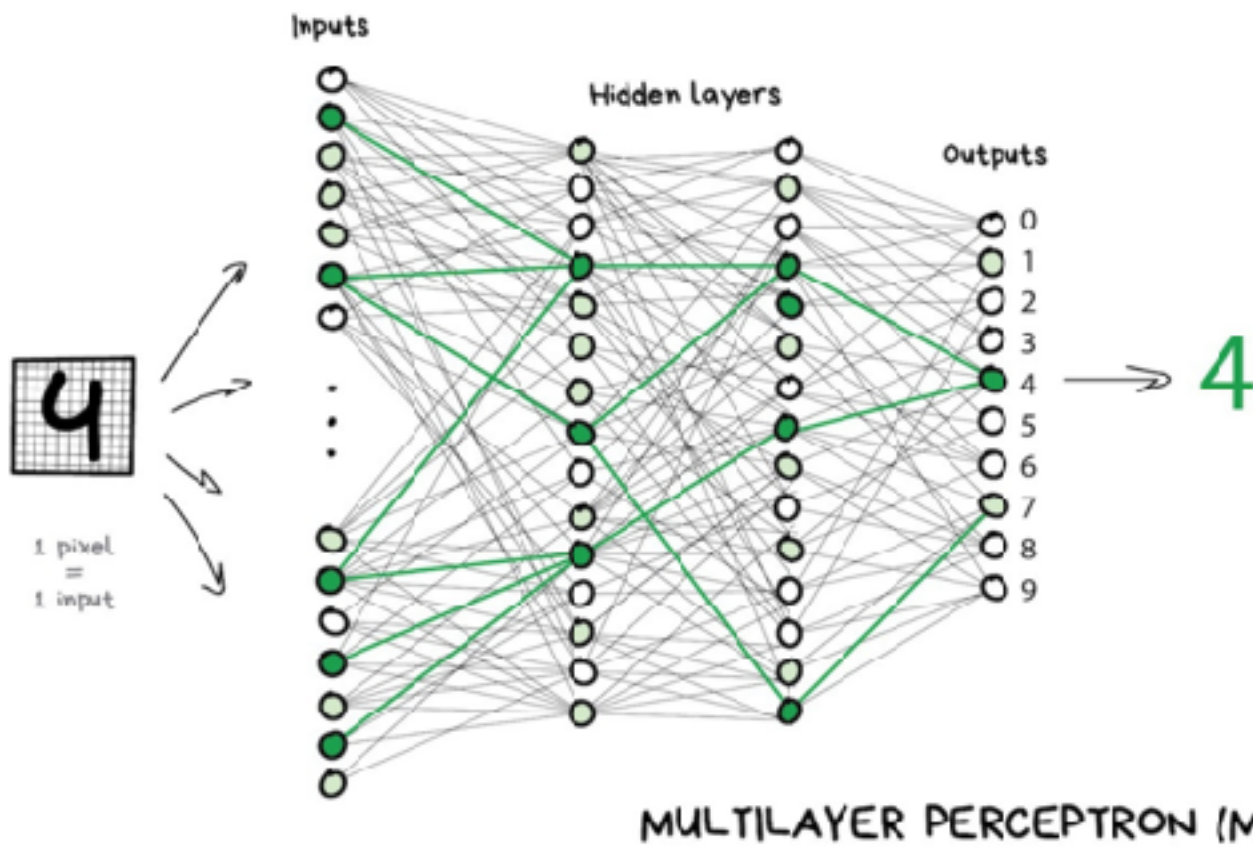


Convolutional neural networks

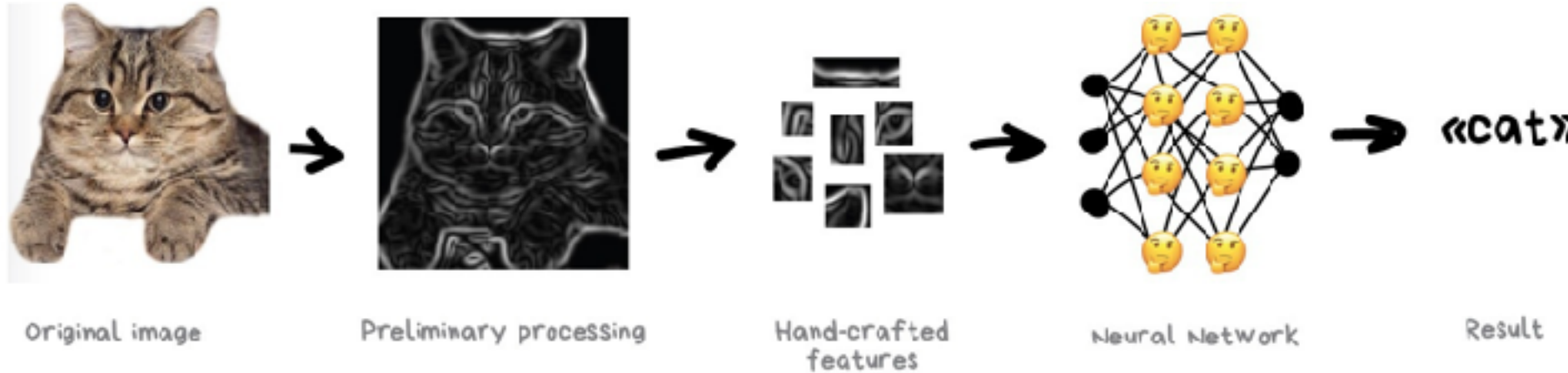


These weights tell the neuron to respond more to one input and less to another. Weights are adjusted when training — that's how the network learns. Basically, that's all there is to it.

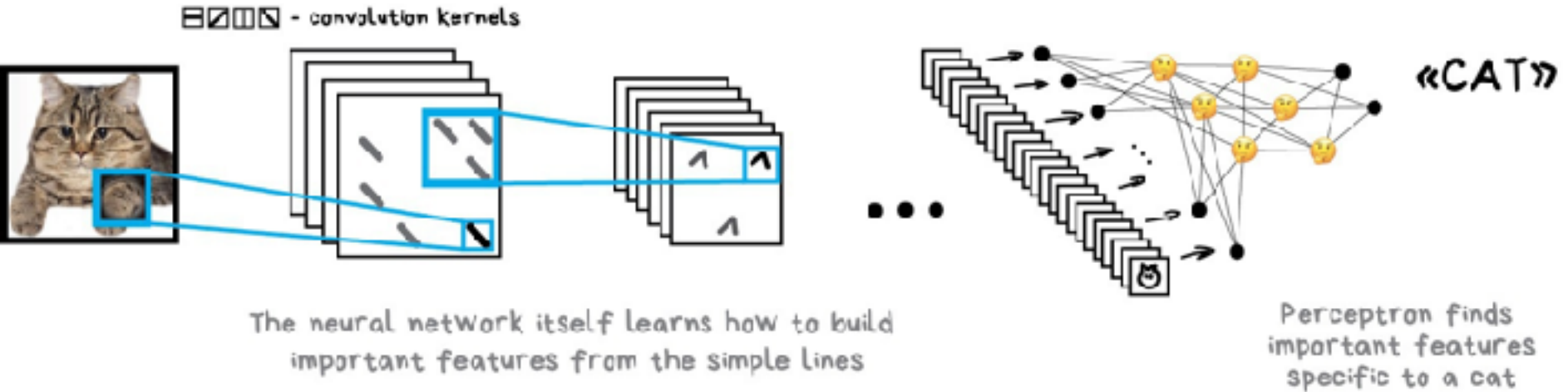




Manually labeling used to be the way...



CNNs avoid manual features



“CNNs are all the rage right now. They are used to search for objects on photos and in videos, face recognition, style transfer, generating and enhancing images, creating effects like slow-mo and improving image quality. Nowadays CNNs are used in all the cases that involve pictures and videos.”

CONVOLUTIONAL NEURAL NETWORK (CNN)



Much of DL success comes from semi-supervised tricks to avoid large hand labelled datasets

Masked LM

- **Solution:** Mask out $k\%$ of the input words, and then predict the masked words
 - We always use $k = 15\%$

the man went to the `[MASK]` to buy a `[MASK]` of milk

store gallon

↑ ↑

- Too little masking: Too expensive to train
- Too much masking: Not enough context



all your base are belong to

@jasongfleischer

...

I used @OpenAI GPT-3 to make some #programming themed candy hearts for you this Valentine's Day. Hope you feel the 100% computer generated ❤️! Here's a selection of the one's I liked best (thread 1/3)





all your base are belong to @jasongfleischer · 1h

...

It's too long to fit on the candy heart generator but 🤖🤖🧠 (bonus post, now the thread is over really)

candy theme: normal

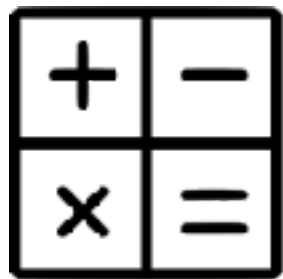
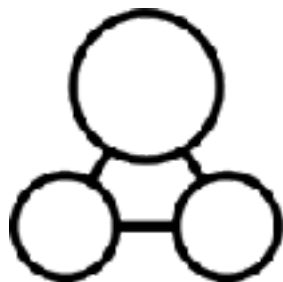
messages: BESTIE, CUTIE PIE, SOUL MATE, SWEET PEA,
UR CUTE, YOU + ME, BE MINE, PICK ME, KISS ME, LOVE
BIRDS, MARRY ME, OOO LA LA, TRUE LOVE, WINK WINK,
XOXO

candy theme: programming

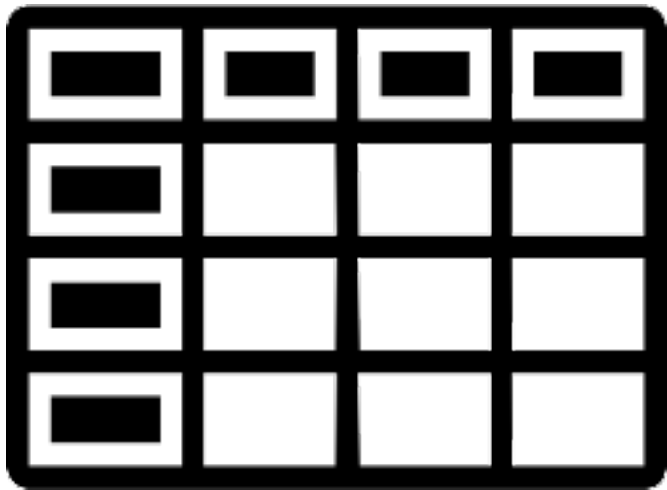
messages:



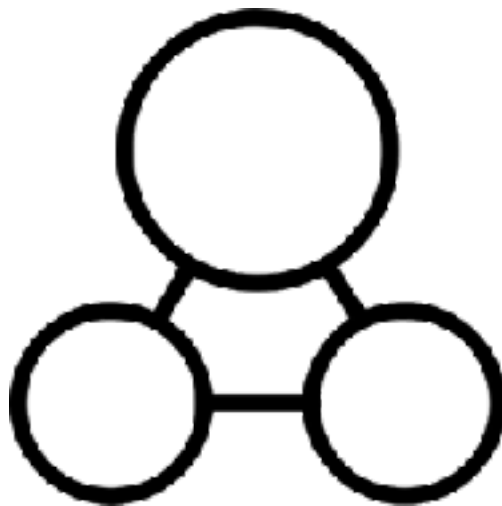
(1) I <3 programmers, (2) Programmers do it better, (3) If you can't code it, you can't date it, (4) Code is poetry, (5) Programmers make the world go round, (6) If you can't code, you can't love, (7) Love is the language of the future, (8) I heart code, (9) Code is life, (10) If you don't code, you



model selection



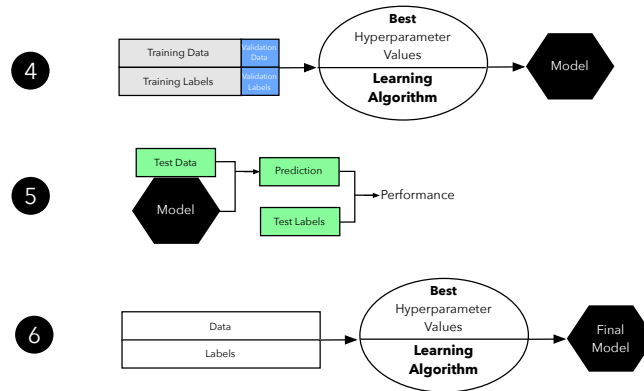
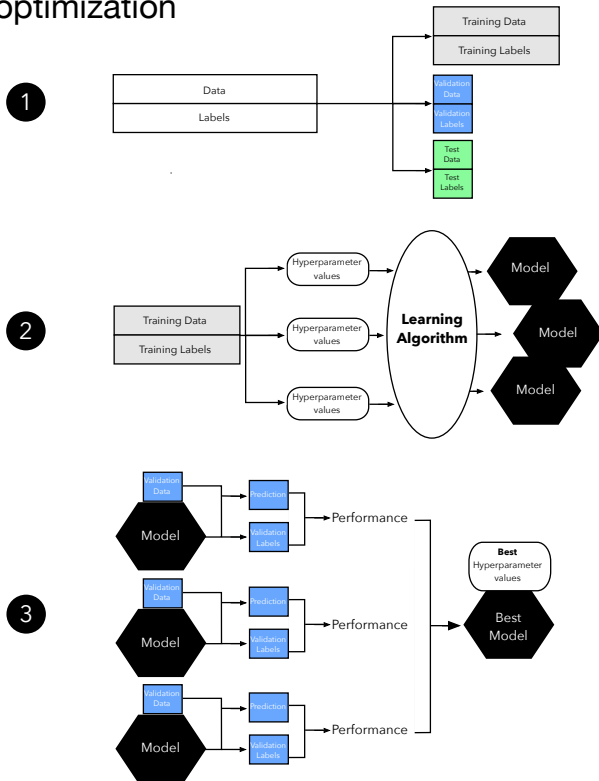
BIG
datasets



SIMPLE
models

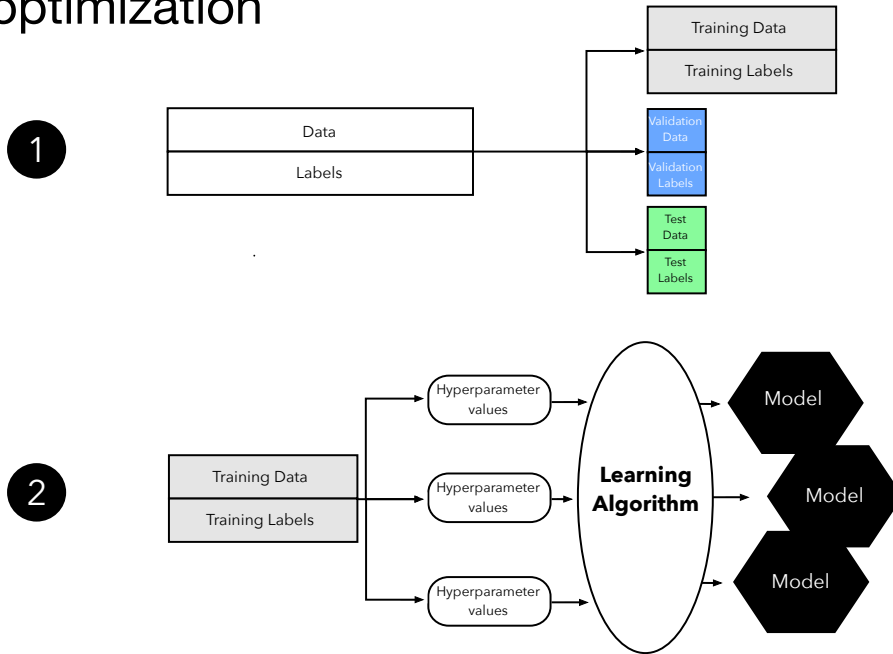
3-Way Holdout

instead of "regular" holdout to avoid "data leakage" during hyperparameter optimization

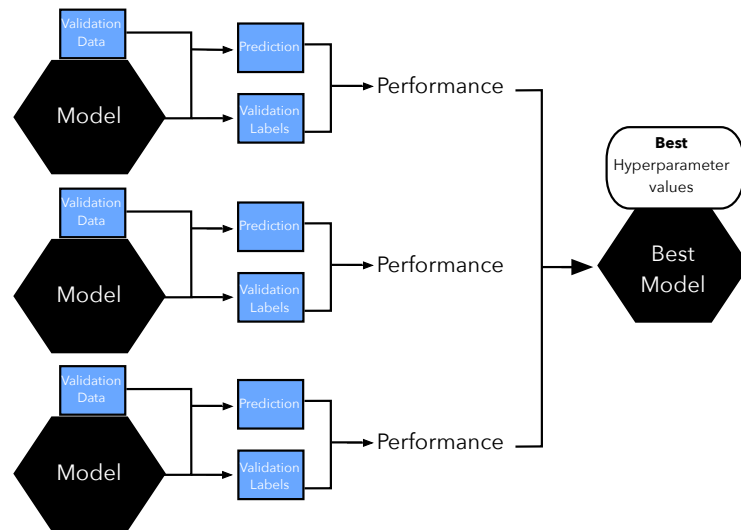


3-Way Holdout

instead of "regular" holdout to avoid "data leakage" during hyperparameter optimization



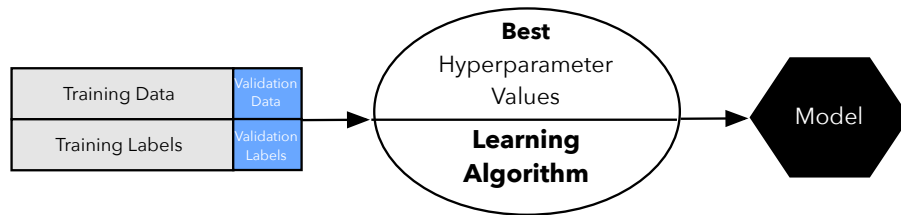
3



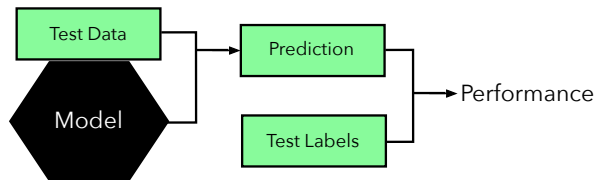
Sebastian Raschka

STAT 451: Intro to ML

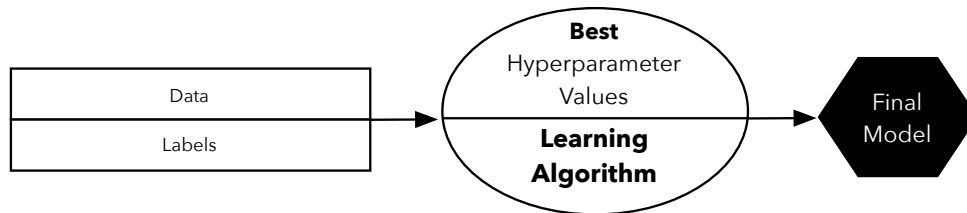
4



5

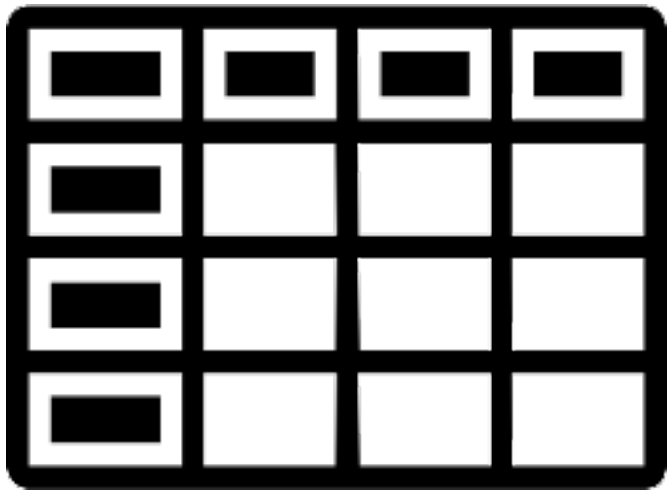


6

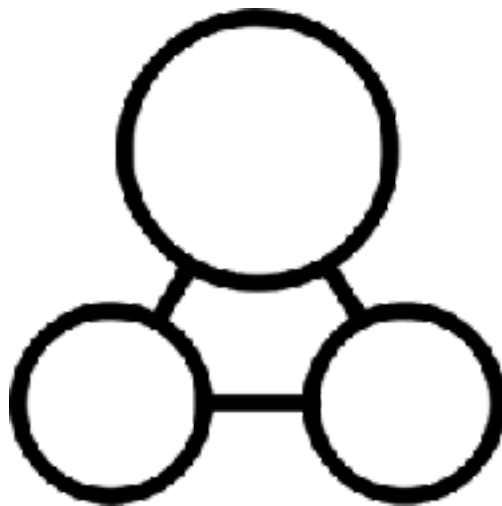


Model, feature, and hyperparameter selection

- Needs: Don't leak information from training/selection process into the test set!
- Trade-offs: Usually not enough data to have completely separate train, validation, test sets. Which one do we prioritize?
 - Low training data -> bad fit
 - Low validation data -> bad selection of model/feature/hparam
 - Low test data -> poor estimate of generalization performance



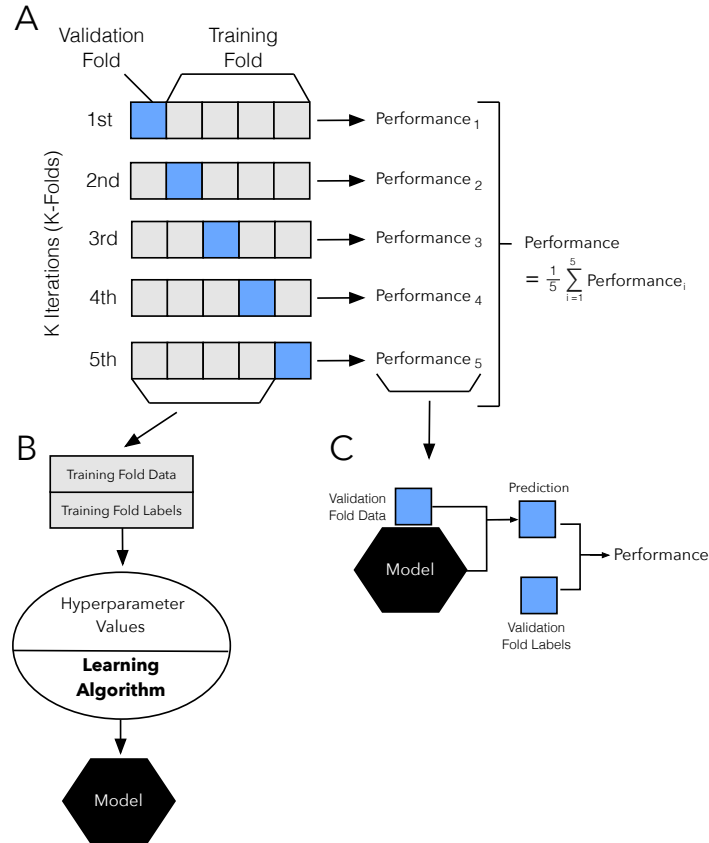
SMALLER
datasets



COMPLEX
models

k-Fold Cross-Validation

- Non-overlapping test data, overlapping training data among folds
- Small k -> biased pessimistic from small training data
- Variance increases with k; LOOCV is unbiased but high variability

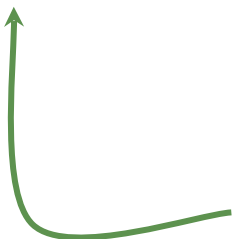




model assessment

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{Predicted}_i - \text{Actual}_i)^2}{N}}$$

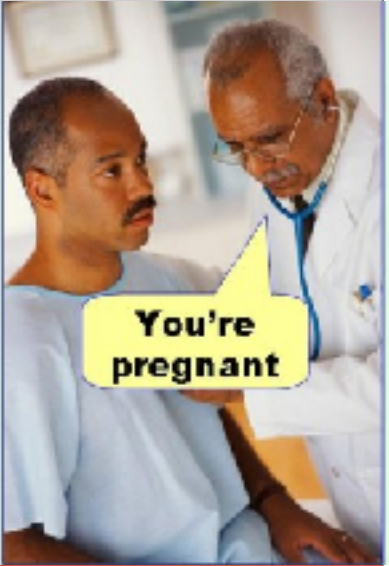

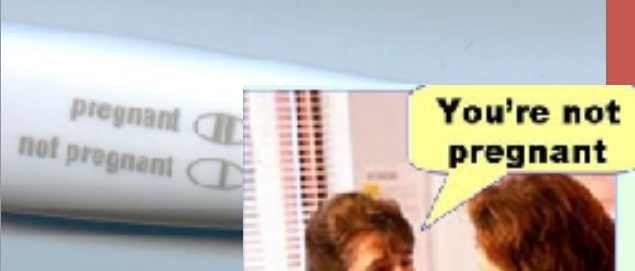
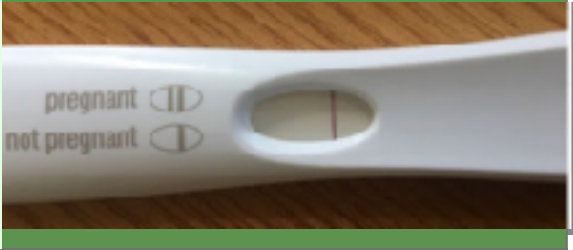


A few outliers can lead to a big increase in RMSE, even if all the other predictions are pretty good

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

Accuracy can mislead

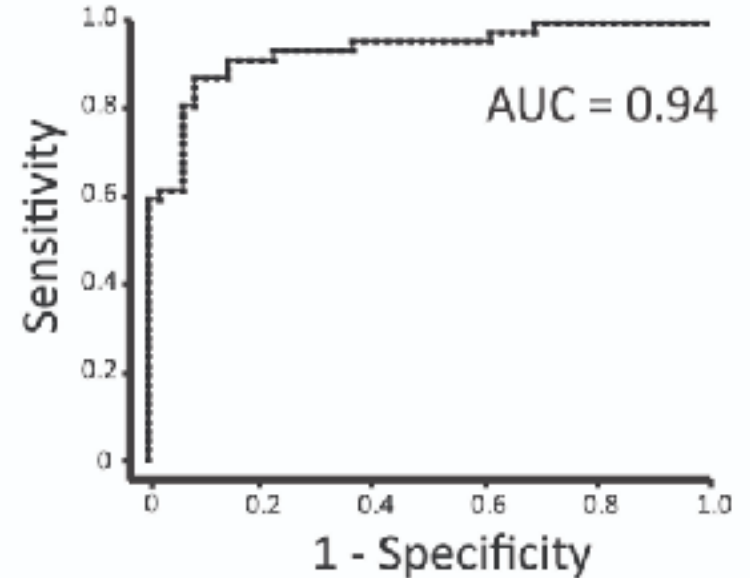
- If classes are imbalanced, what would “chance performance” be?
- The classic example: detect cancer
 - 1/1000 actually have cancer
 - your prediction algorithm misdiagnoses 1% of healthy people
 - Do the math... your algorithm tells 10 people who are healthy they are sick for every 1 person who is actually sick

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

		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
TP
TP + FN



Specificity
TN
TN + FP

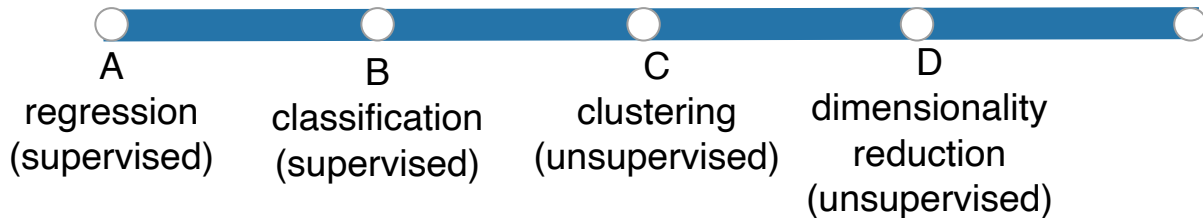
Accuracy	What % were predicted correctly?
Sensitivity	Of those that <i>were</i> positives , what % were predicted to be positive?
Specificity	Of those that were <i>negatives</i> , what % were predicted to be negative?

Prediction Approach



You've been given a dataset with a number of features and have been asked to predict each individual's age.

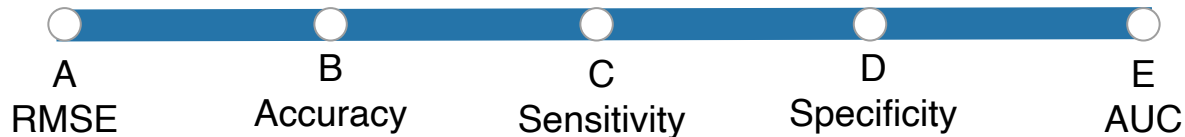
What prediction approach would you use?



Prediction Approach



After predicting each person's age, how would you assess your model?



Prediction Approach



Which would be the error value you'd want from your model?

