

Machine learning

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>

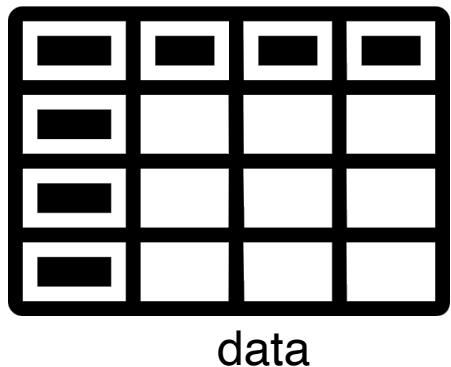
- **Problem:** Detecting whether credit card charges are fraudulent.
- **Data science question:** Can we use the time of the charge, the location of the charge, and the price of the charge to predict whether that charge is fraudulent or not?
- **Type of analysis:** Predictive analysis



Robert Hecht-Neilsen and Zeus (and others) sold HNC for \$810M in 2002
Around here lots of people see him as a major contributor to the development of neural networks and data science as we know them today

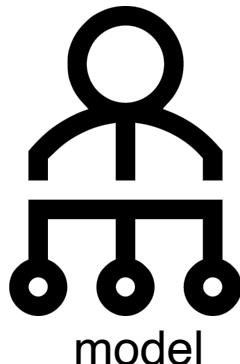
predictive analysis
uses data you have now
to make predictions in
the future

machine learning
approaches are used for
predictive analysis!



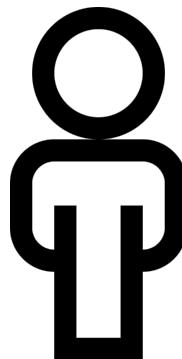
data

train →



model

predict →

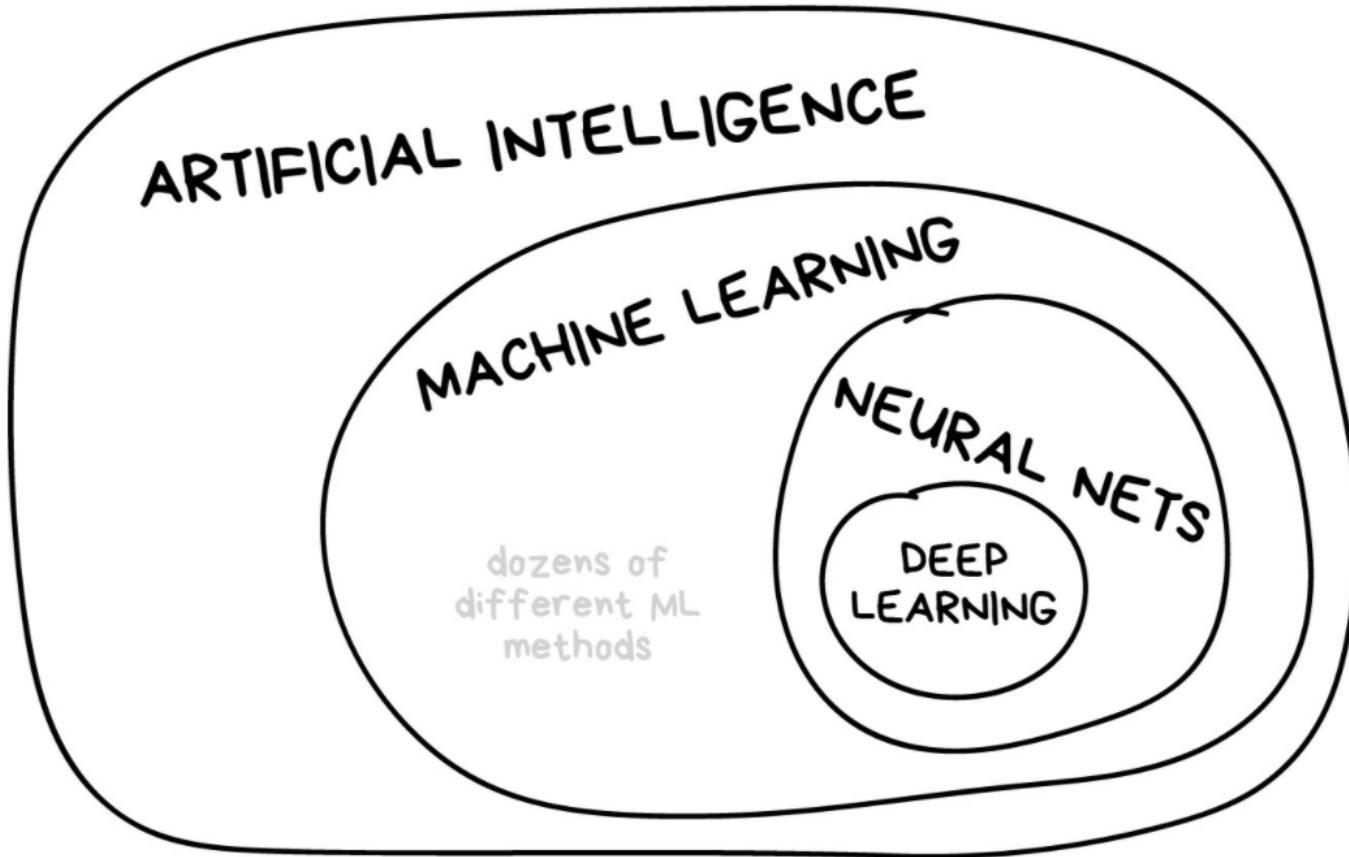


In contrast to statistical approaches which care more about the model
accurately reflecting the process than nailing the predictions

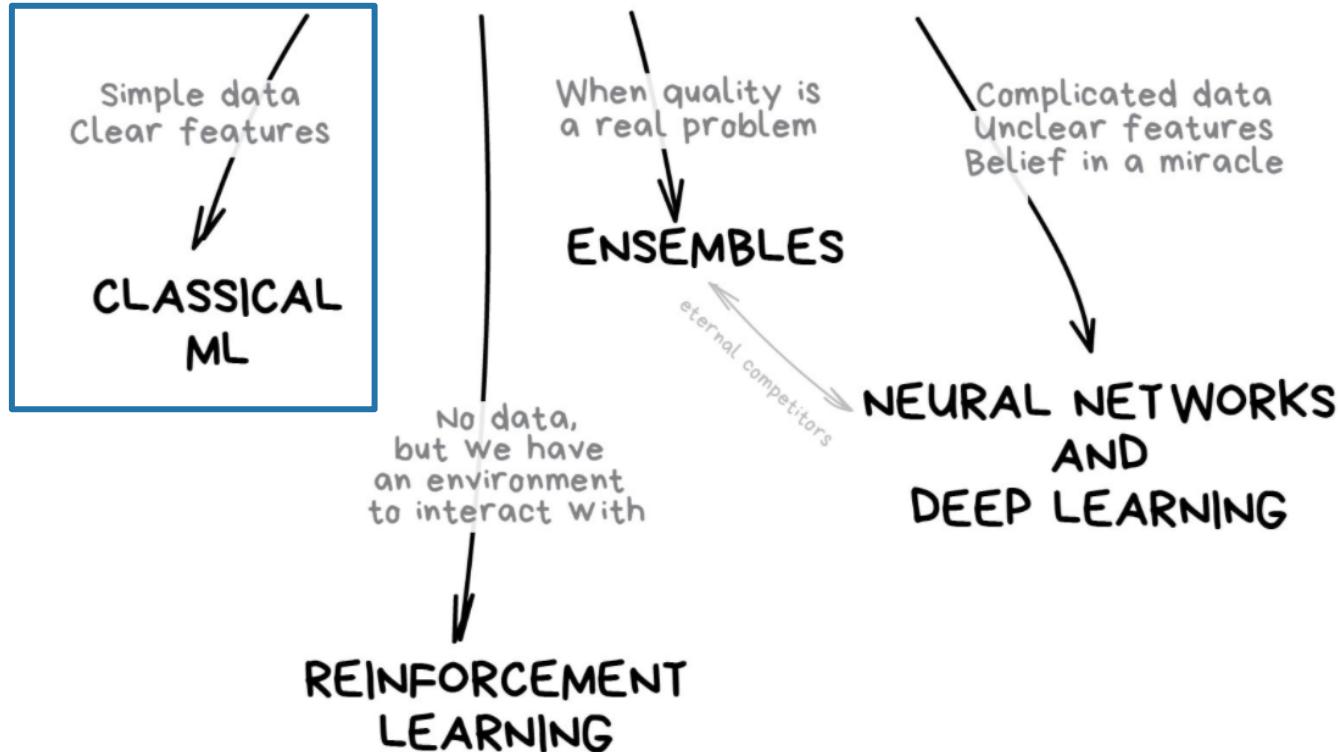
What is machine learning?

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

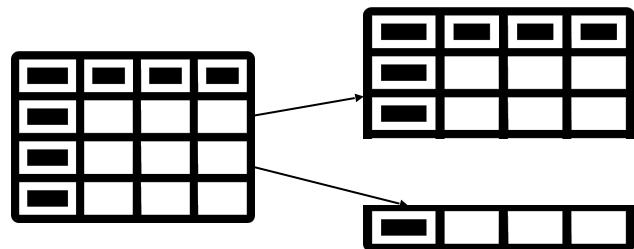
- Andrew Ng, Stanford, ex-Google, chief scientist at Baidu, Coursera founder, Stanford Adjunct Faculty



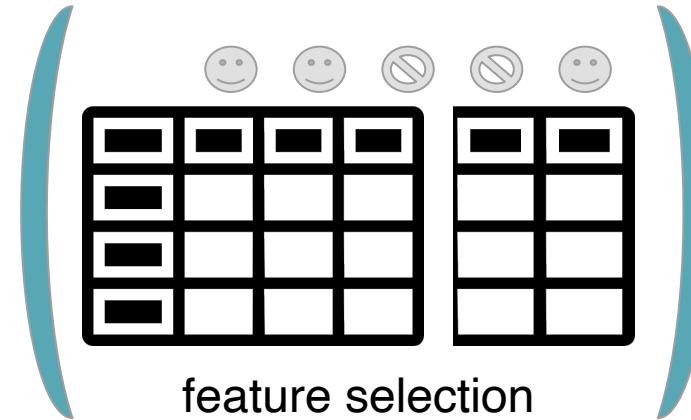
THE MAIN TYPES OF MACHINE LEARNING



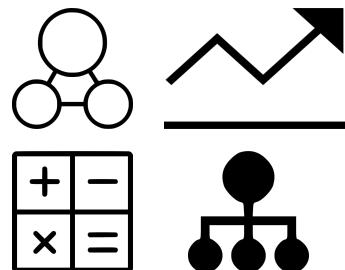
Machine Learning - The usual ingredients in the recipe



data
partitioning



feature selection



model selection



model assessment

Drowning in data?



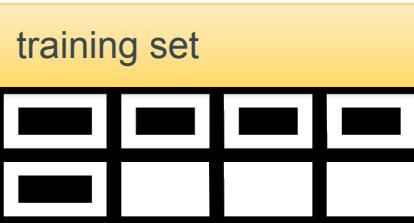
Image borrowed from
<https://medium.com/@tmaoconnor/marketing-is-drowning-in-data-yet-its-starving-for-answers-f15ee890a732>

Limited data?



Image ID: 237163860 | www.depositphotos.com

Data partitioning:
Evaluating how well you will
generalize to new data with huge
amounts of data



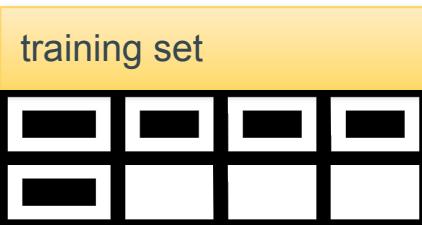
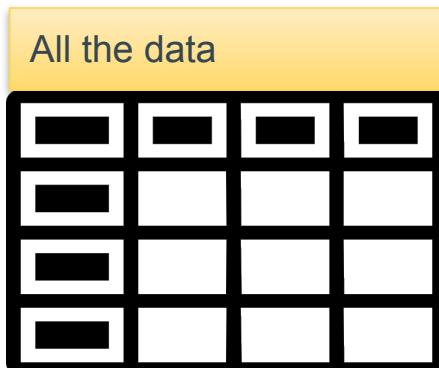
Data used to build
your model



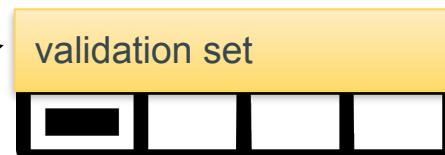
Used to assess how good
the final model is
(Generalization)



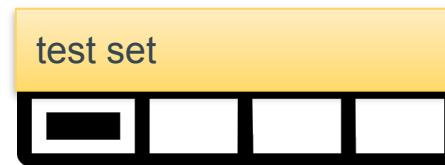
Data partitioning:
Finding the best version of your
model through validation



Data used to build
your model



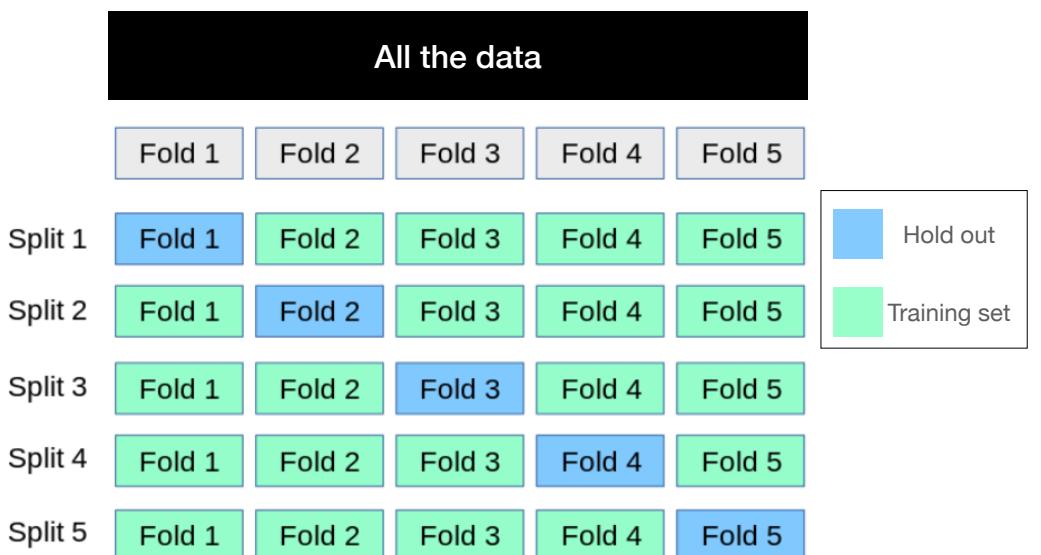
Fine-tune the model to
increase prediction
accuracy
(Hyper-parameter tuning)



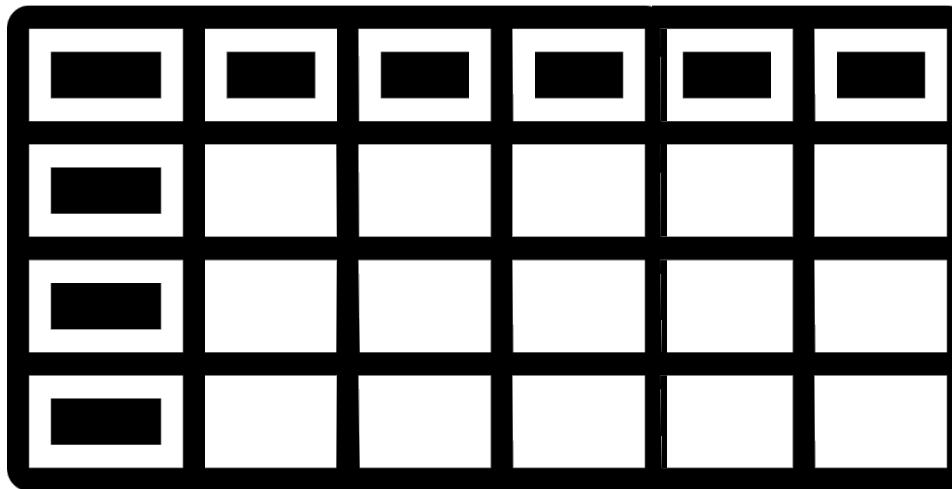
Used to assess how good
the final model is
(Generalization)

When you don't have huge data

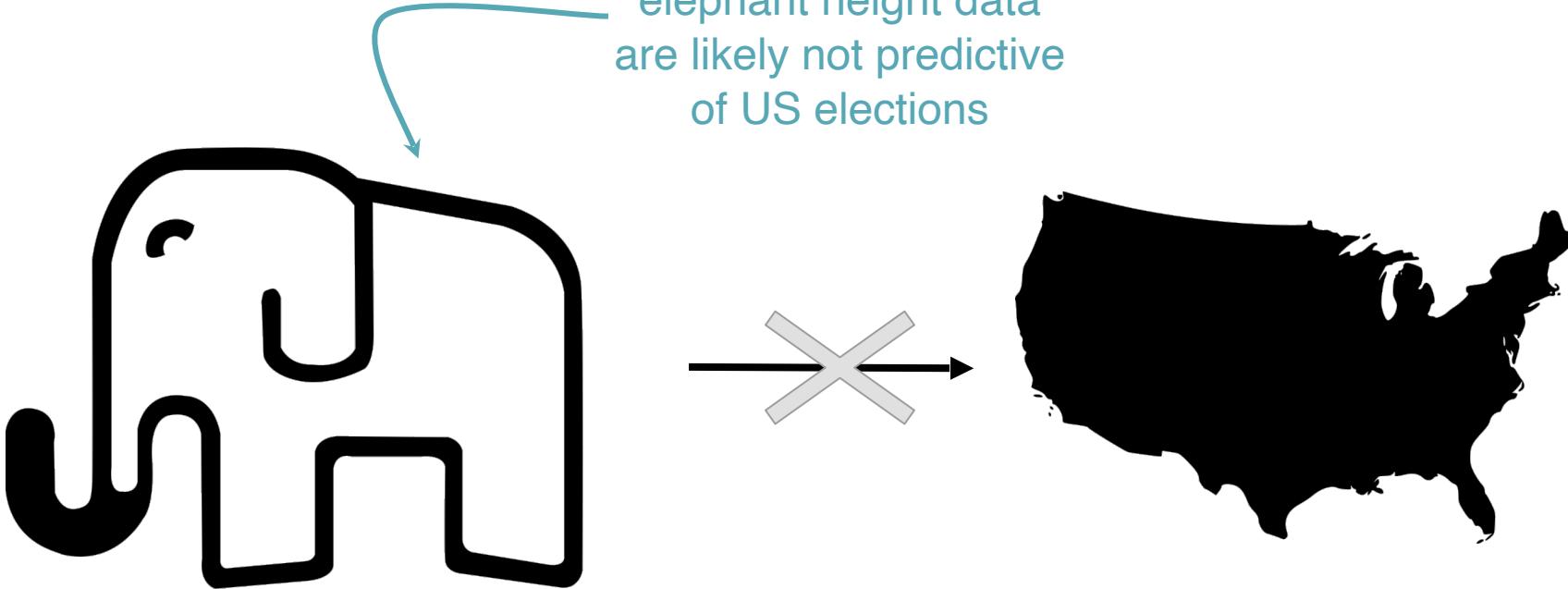
Cross validation



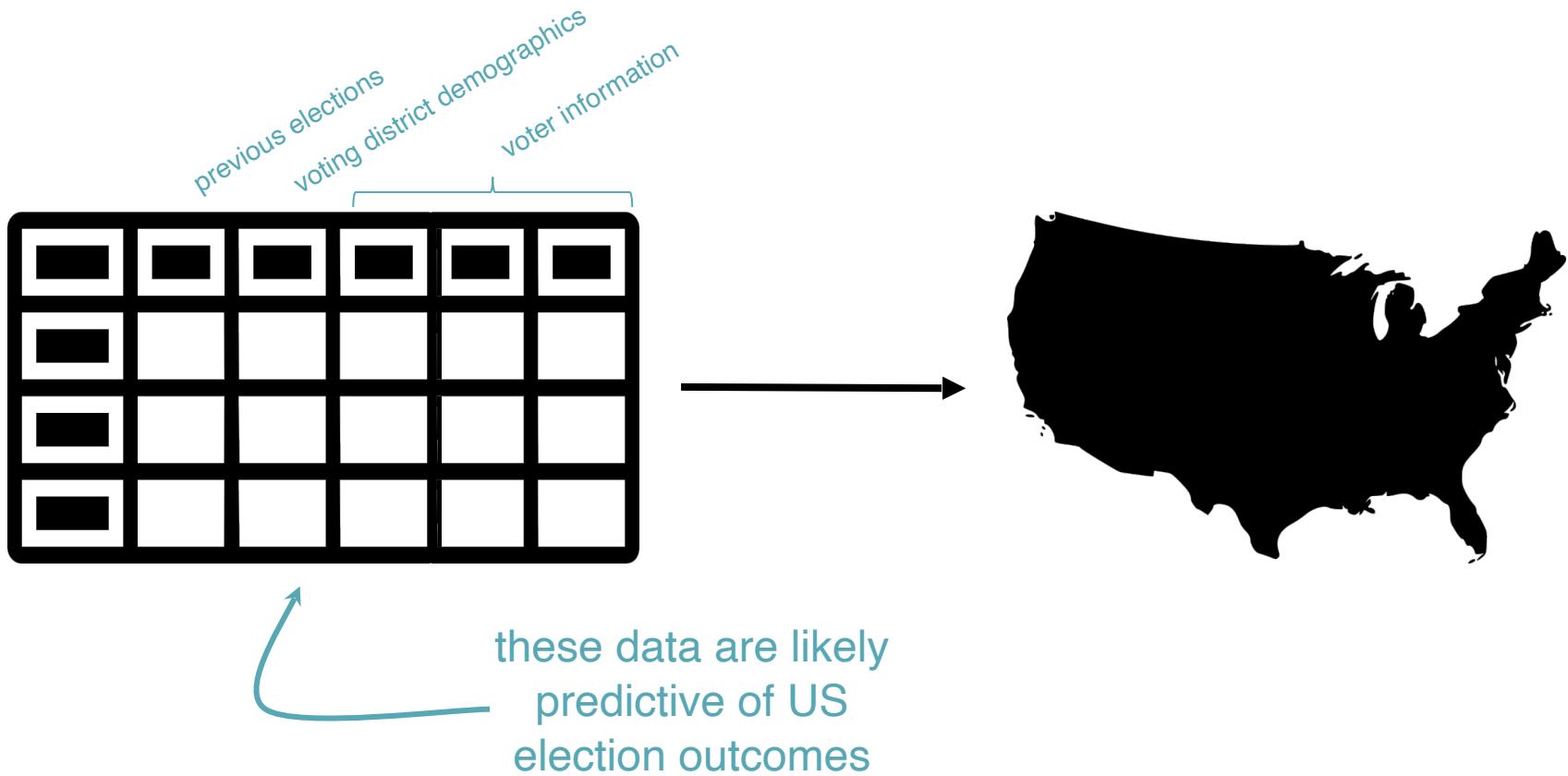
Use the mean of the hold out set performances to estimate either test or validation error

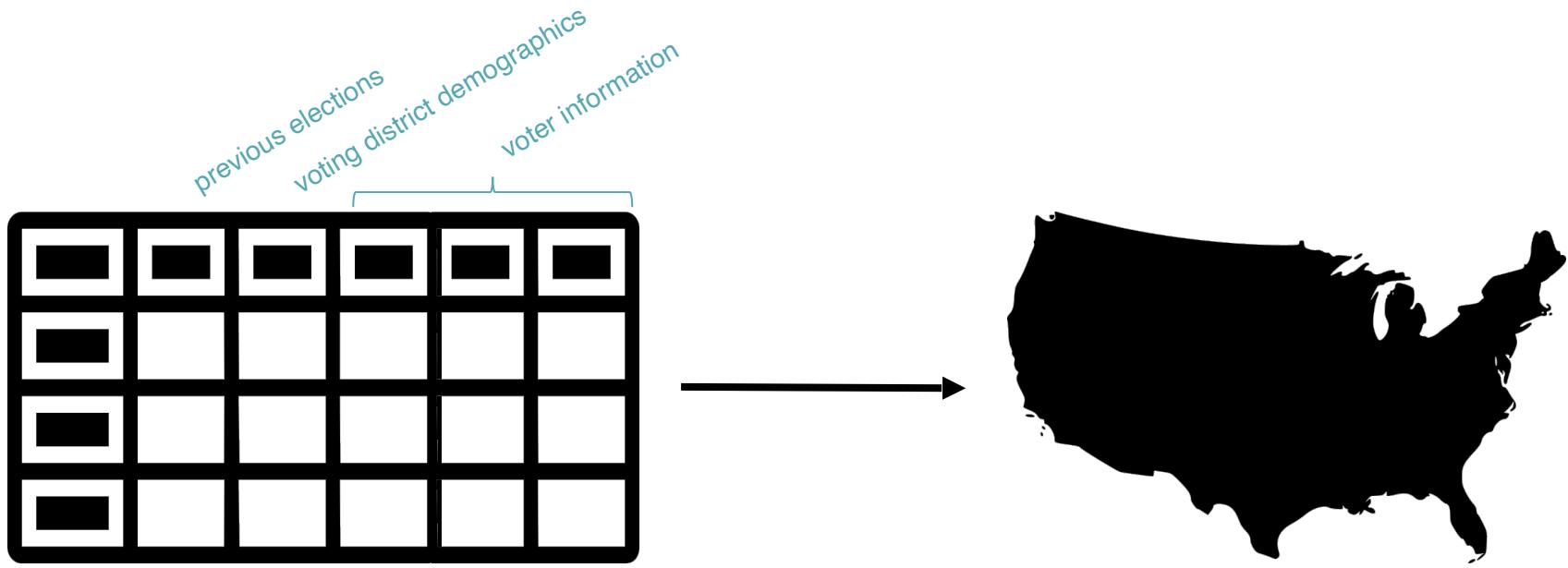


**feature
selection**

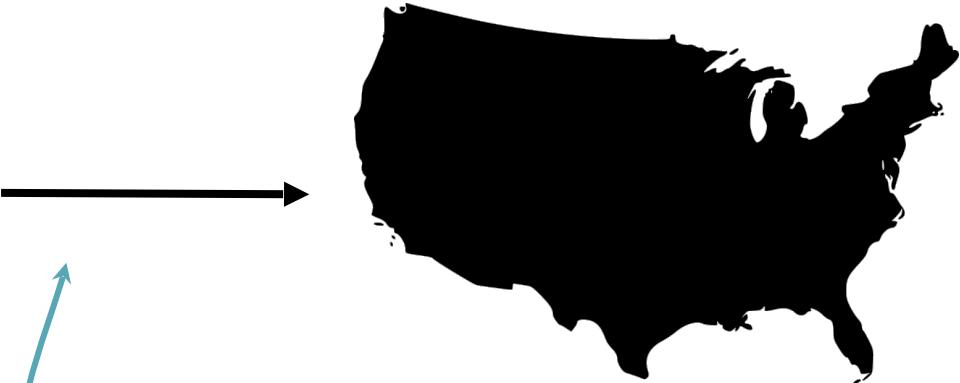
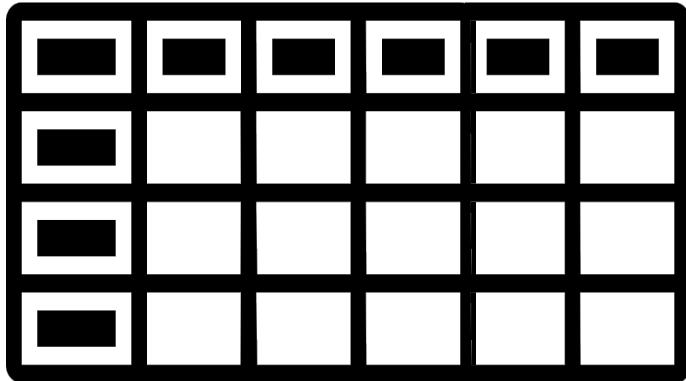


elephant height data
are likely not predictive
of US elections



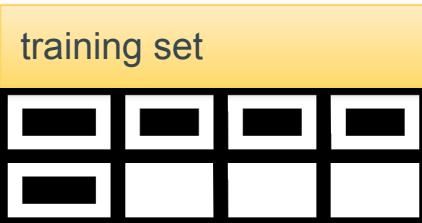
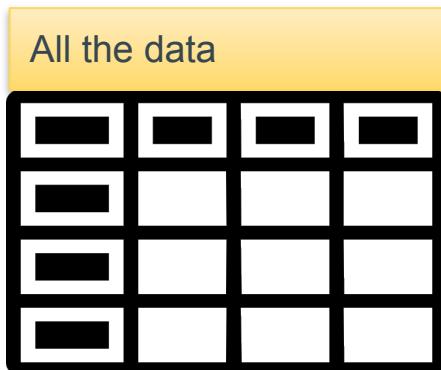


feature selection determines which variables are most predictive and includes them in the model



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

Data partitioning:
Finding the best version of your
model through validation



Data used to build
your model



Fine-tune the model to
increase prediction
accuracy
(Hyper-parameter tuning)

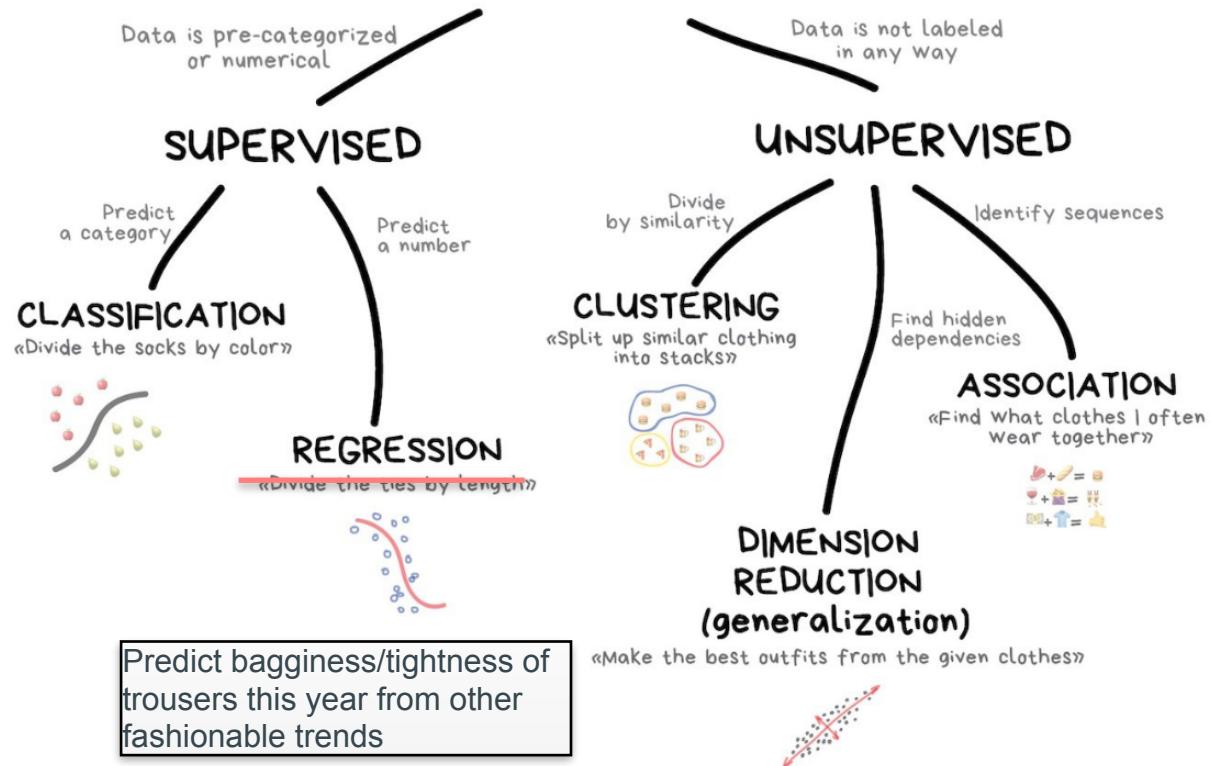


Used to assess how good
the final model is
(Generalization)

Use validation set to select the features!

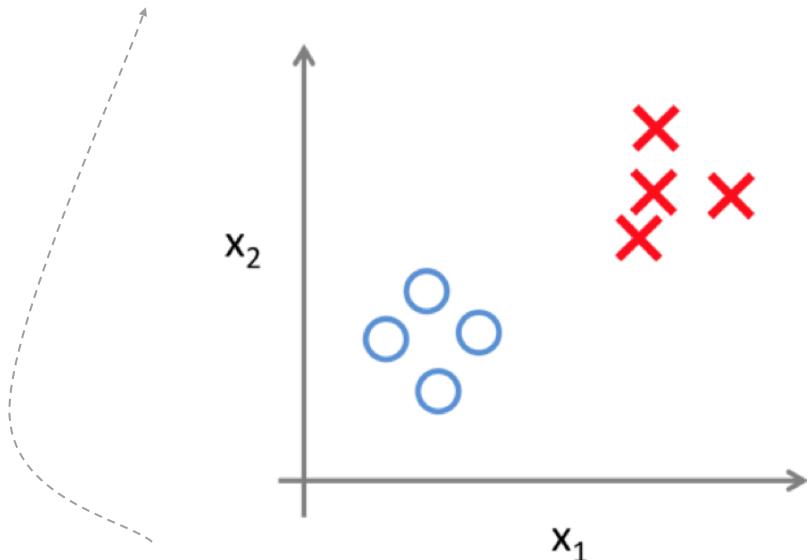
ML Algorithms Overview

CLASSICAL MACHINE LEARNING



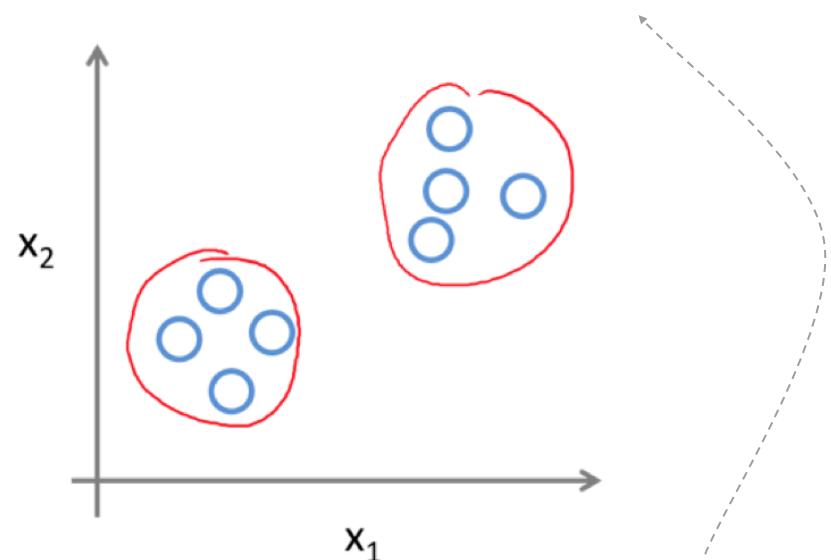
Two modes of machine learning

Supervised Learning



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

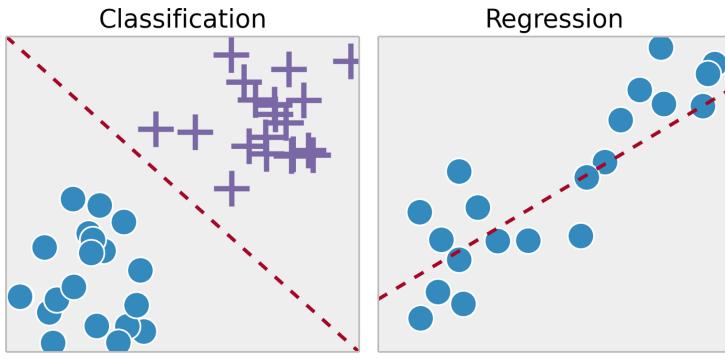
Unsupervised Learning



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

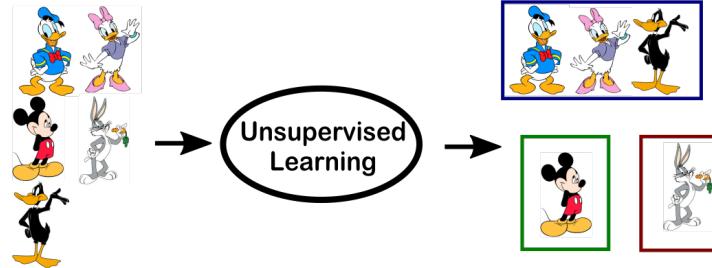
Approaches to machine learning

Supervised Learning



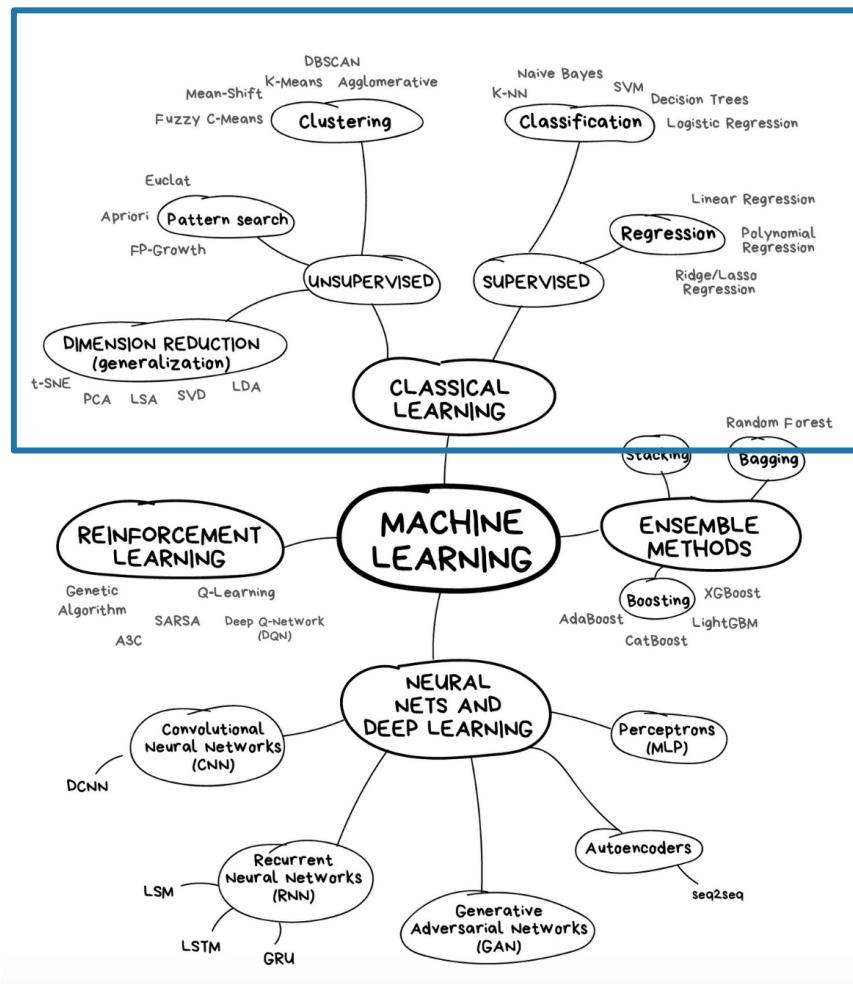
categorical variables

Unsupervised Learning



Clustering (categorical)
& dimensionality reduction (continuous)

can automatically identify
structure in data

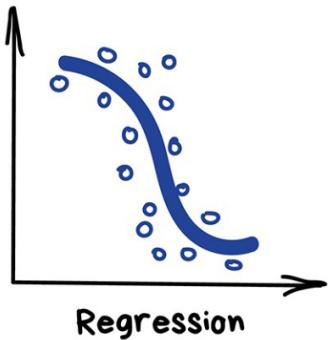


Regression

"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

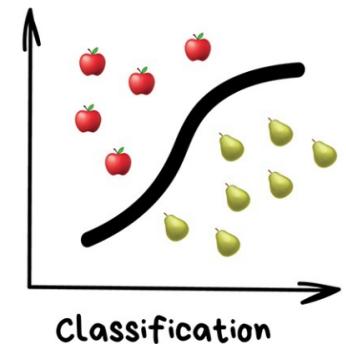


Classification

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

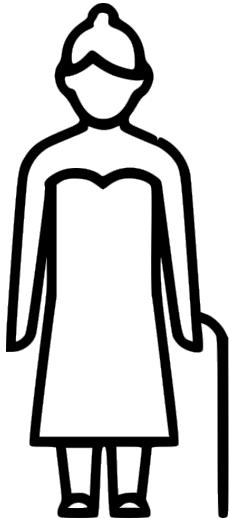
Today used for:

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



Popular algorithms are Linear and Polynomial regressions.

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



Regression:

predicting continuous variables
(i.e. Age)

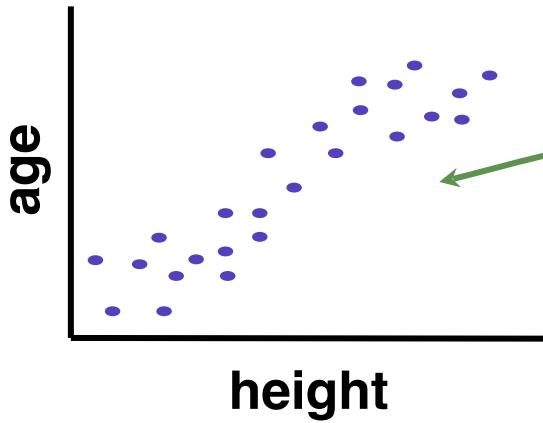
continuous variable prediction



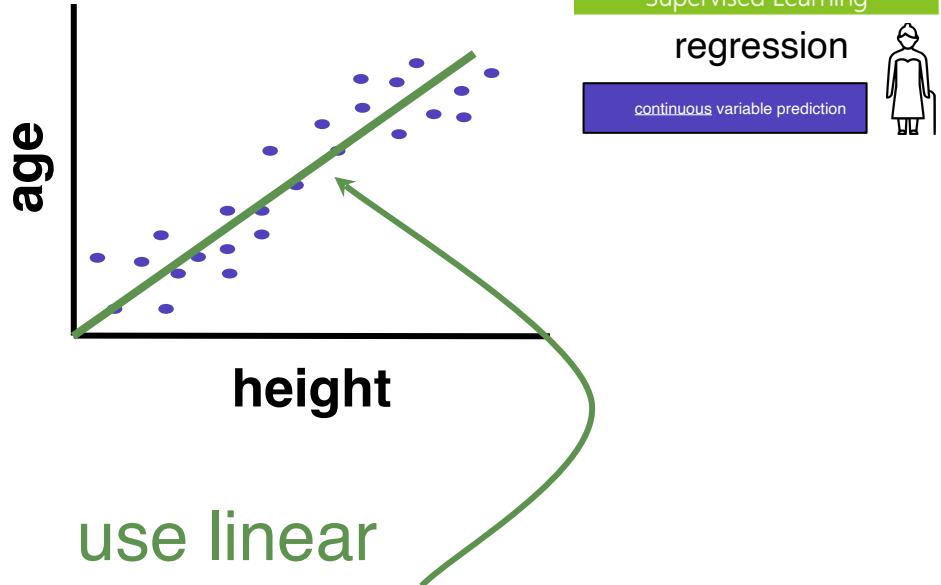
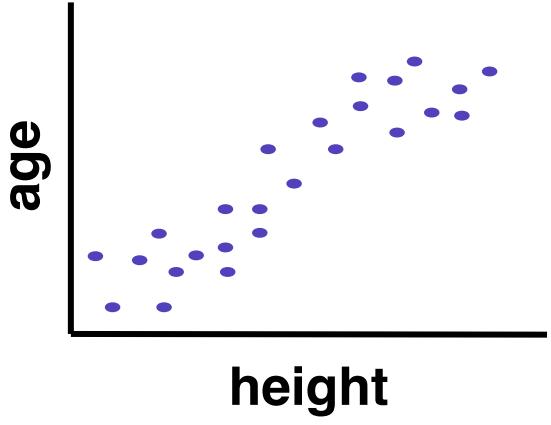
Classification:

predicting categorical variables
(i.e. education level)

categorical variable prediction



the training data
will be used to
build the
predictive
model



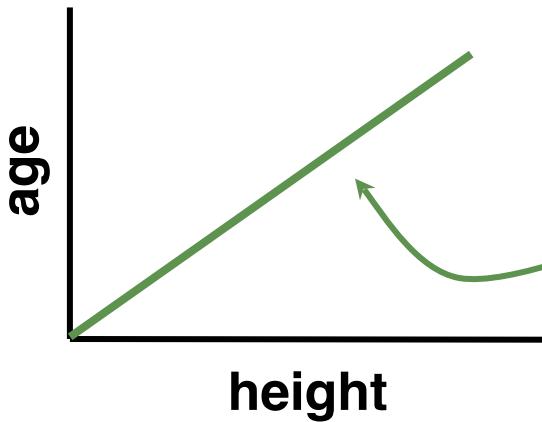
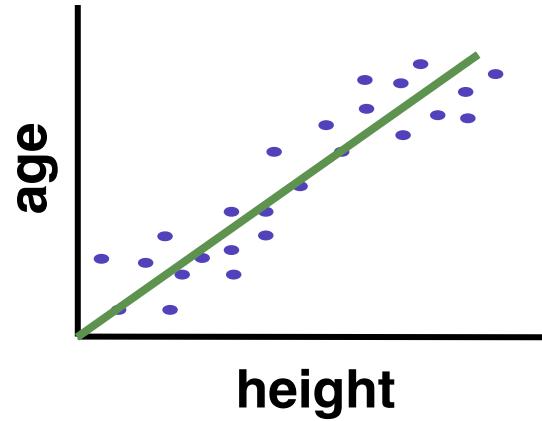
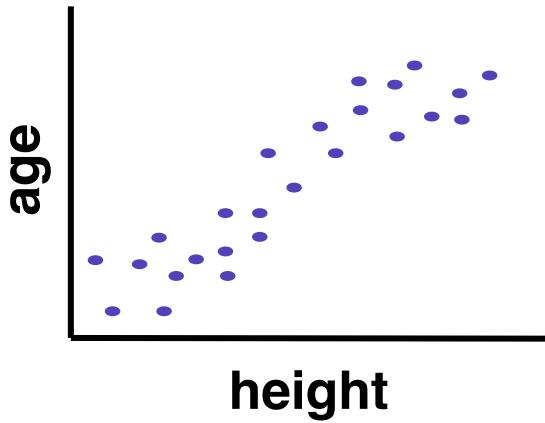
use linear
regression to
model the
relationship

Supervised Learning

regression

continuous variable prediction

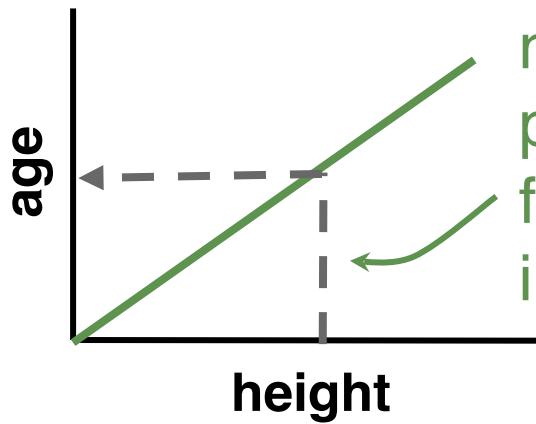
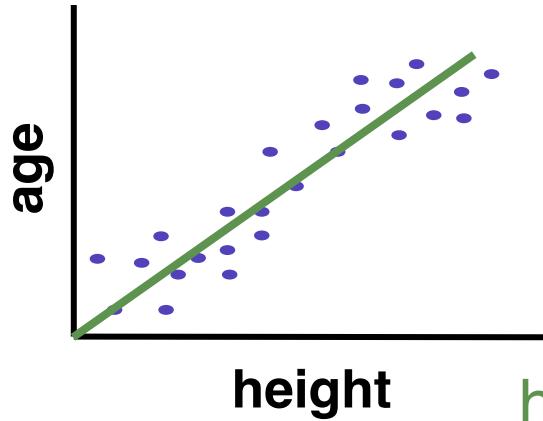
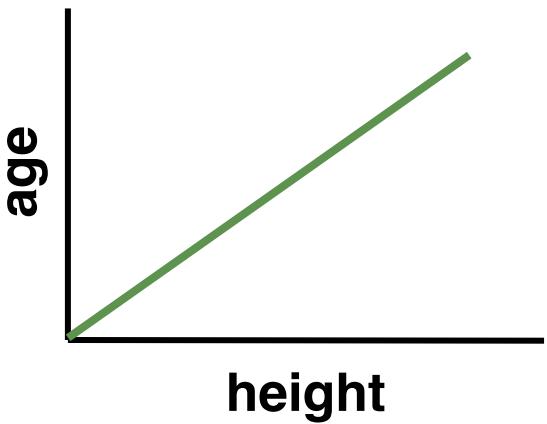
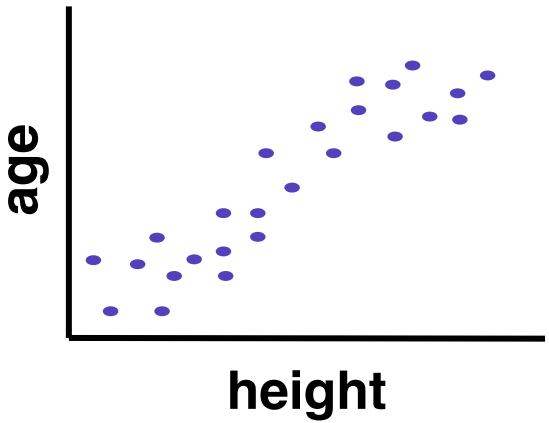




For prediction, the individual values in the training data are *not* important. We only need the model.



continuous variable prediction

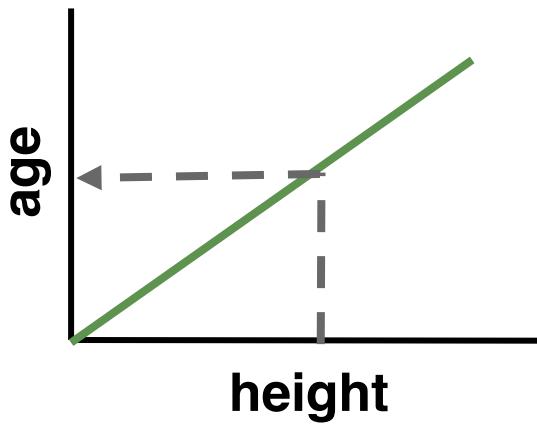
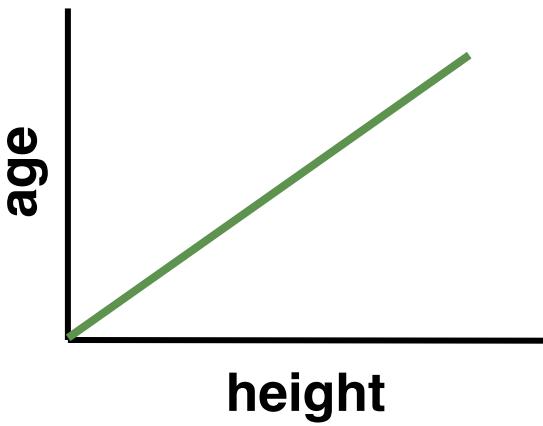
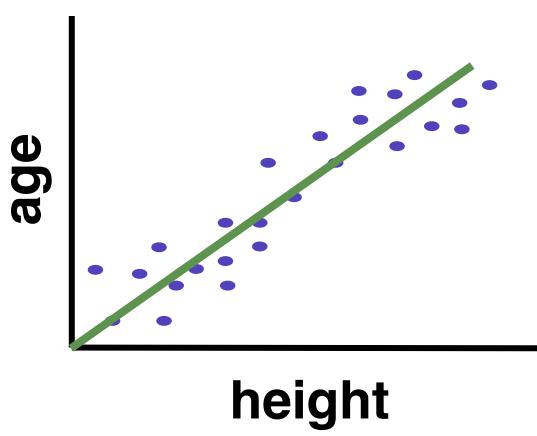
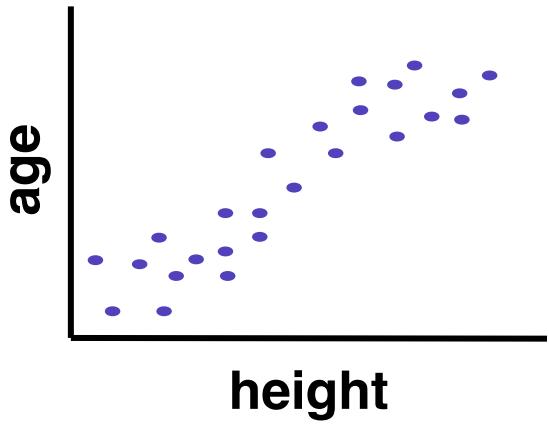


how we'll
make
predictions
for a future
individual

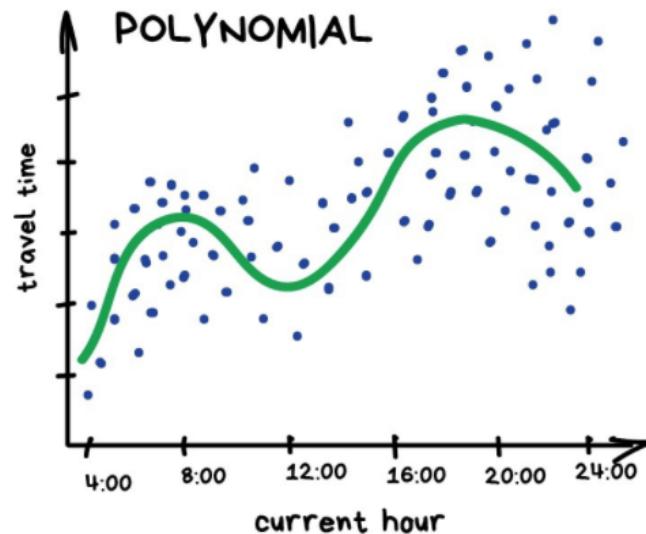
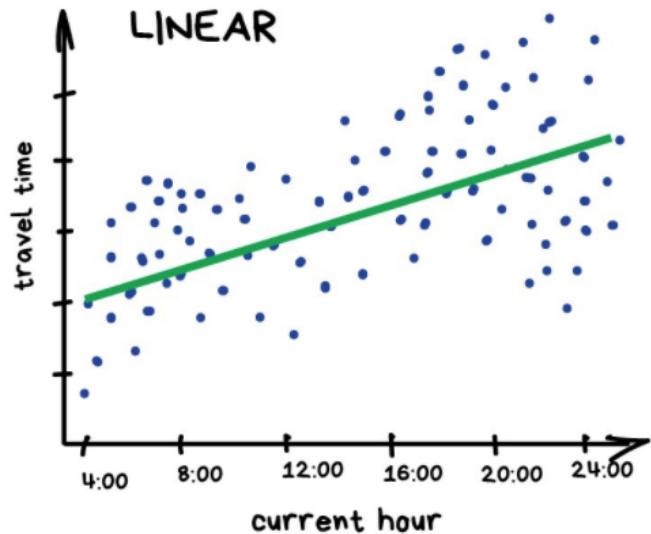
Supervised Learning

regression

continuous variable prediction



PREDICT TRAFFIC JAMS



REGRESSION

<https://forms.gle/HBiDbxR2oRBYxcUh7>

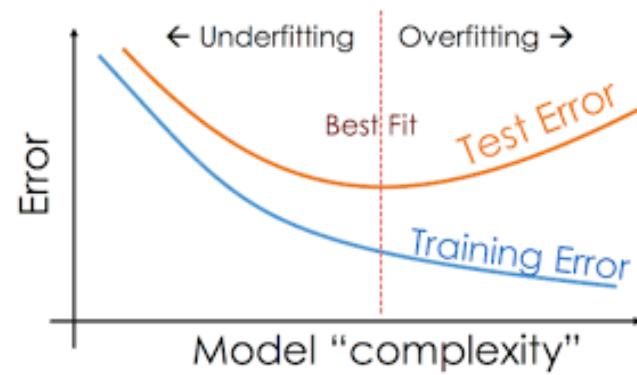
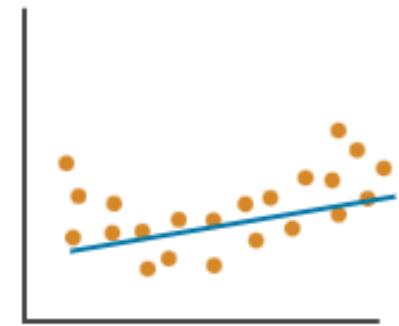
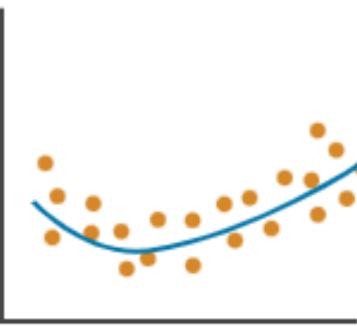
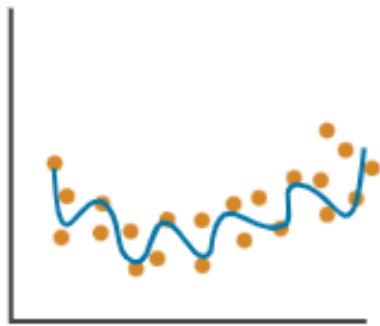


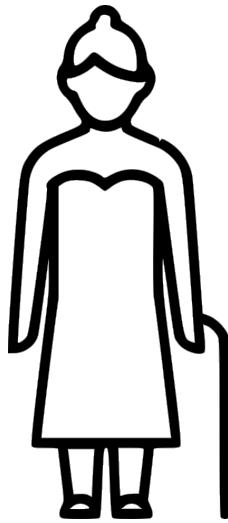
Overfit

Best fit?

Underfit

Regression



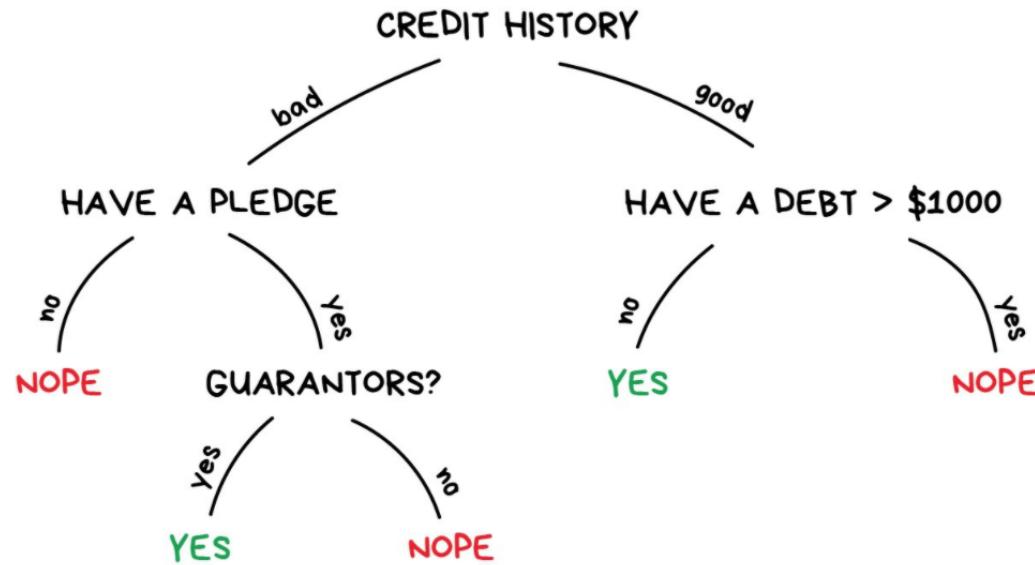


Regression:
predicting continuous
variables
(i.e. Age)



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

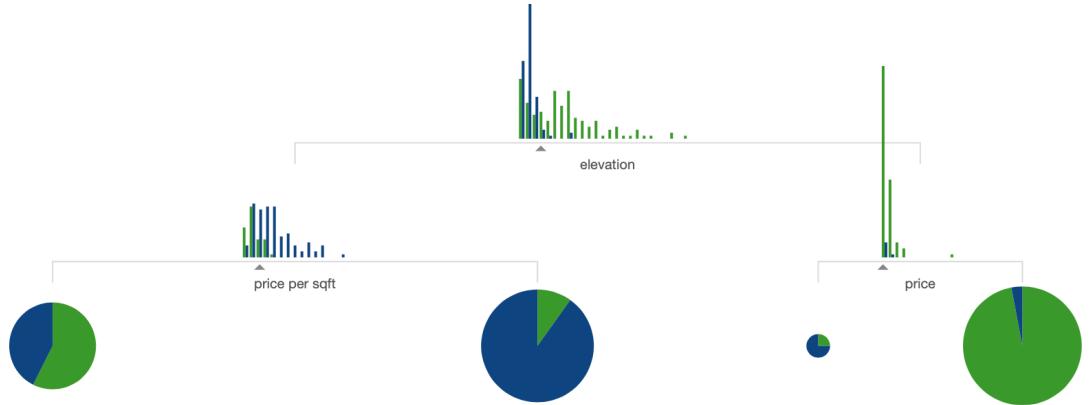
GIVE A LOAN?



DECISION TREE

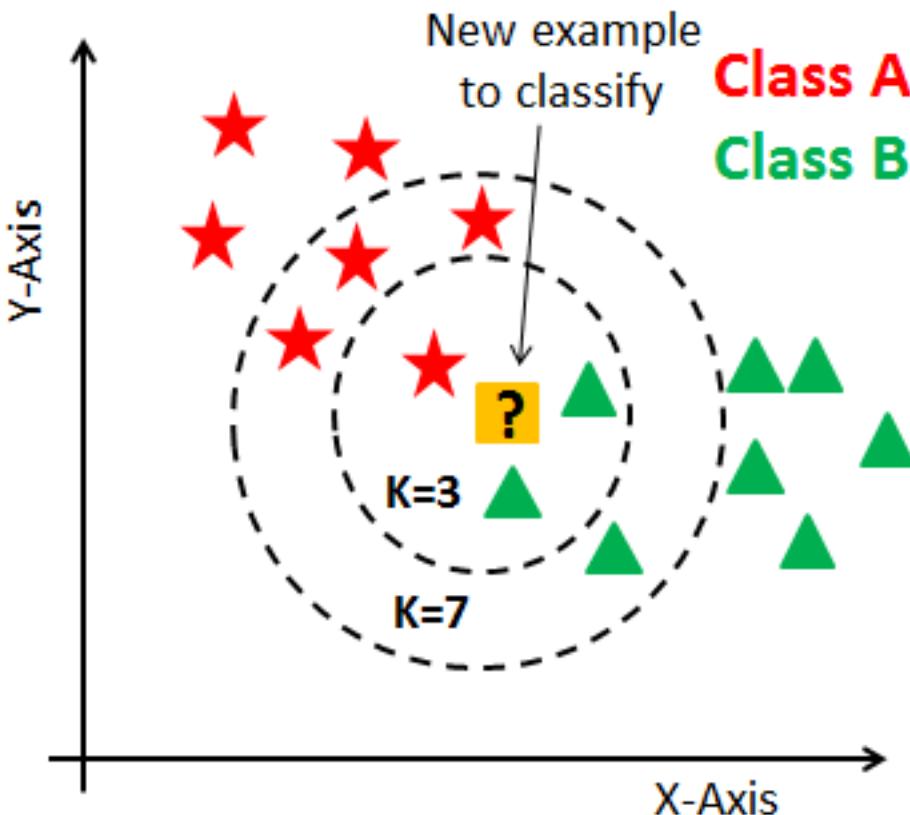
Growing a tree

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

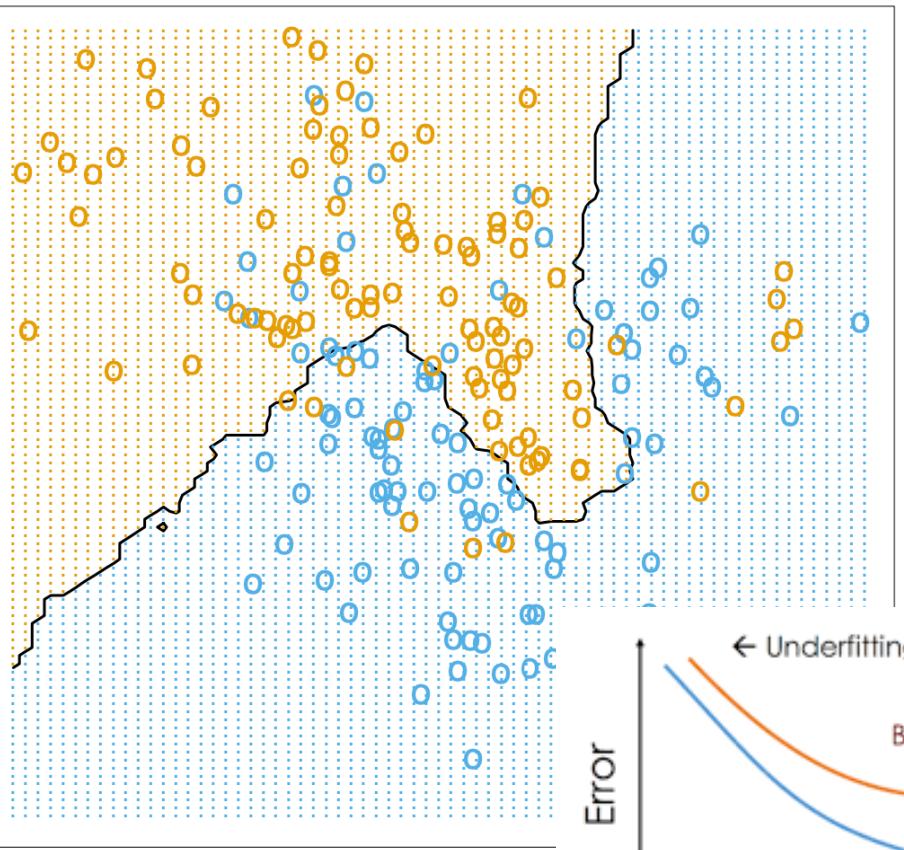


<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

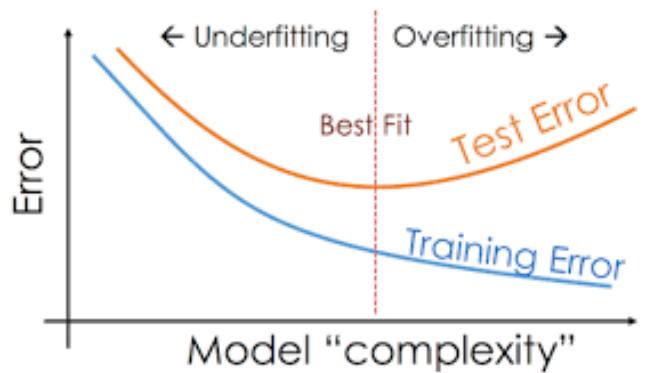
K-nearest neighbors



15-Nearest Neighbor Classifier

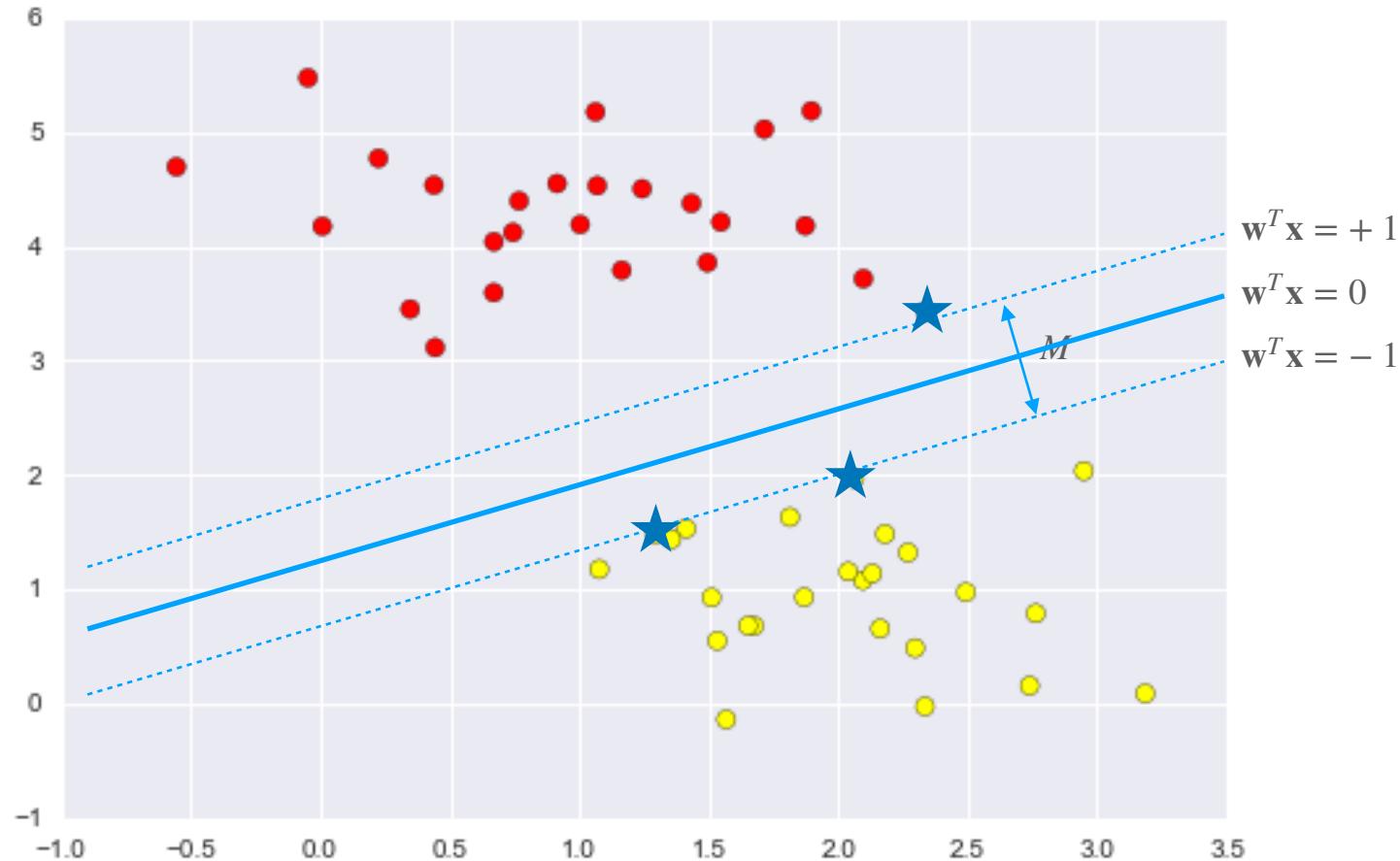


1-Nearest Neighbor Classifier



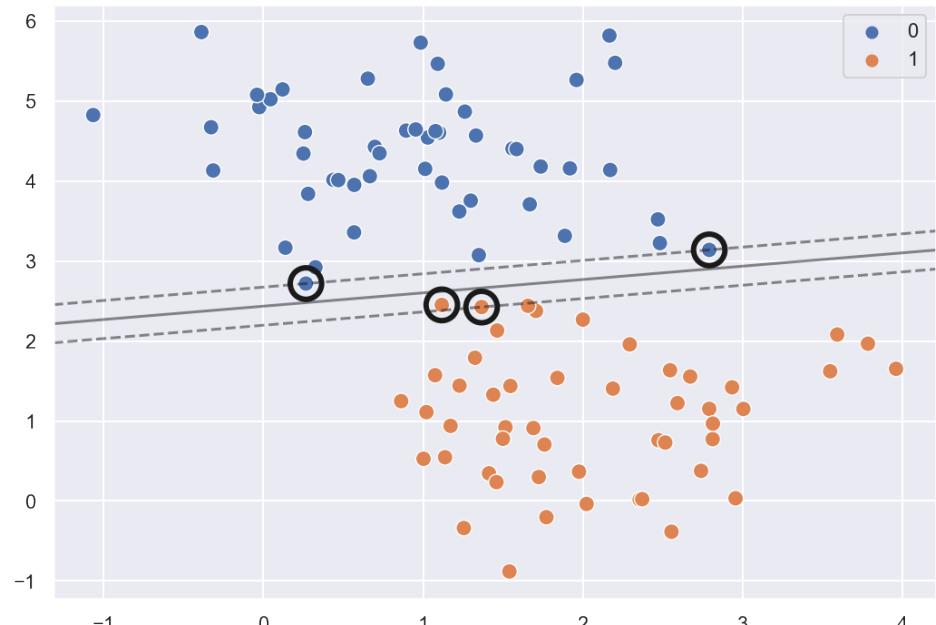
Support Vector Machines:

Find the support
data points that
maximize the
margin M



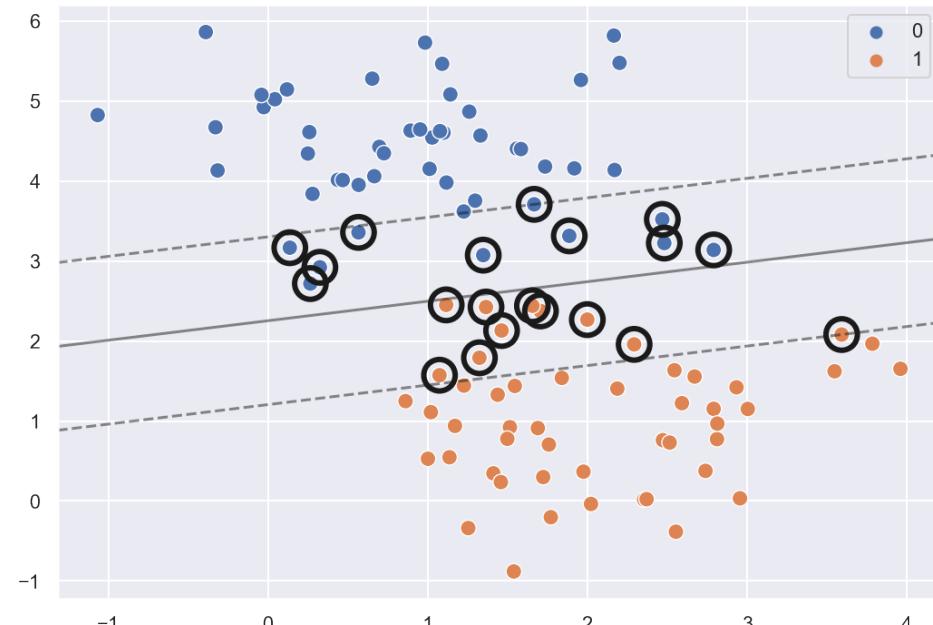
Hard margin

$C = 10.0$



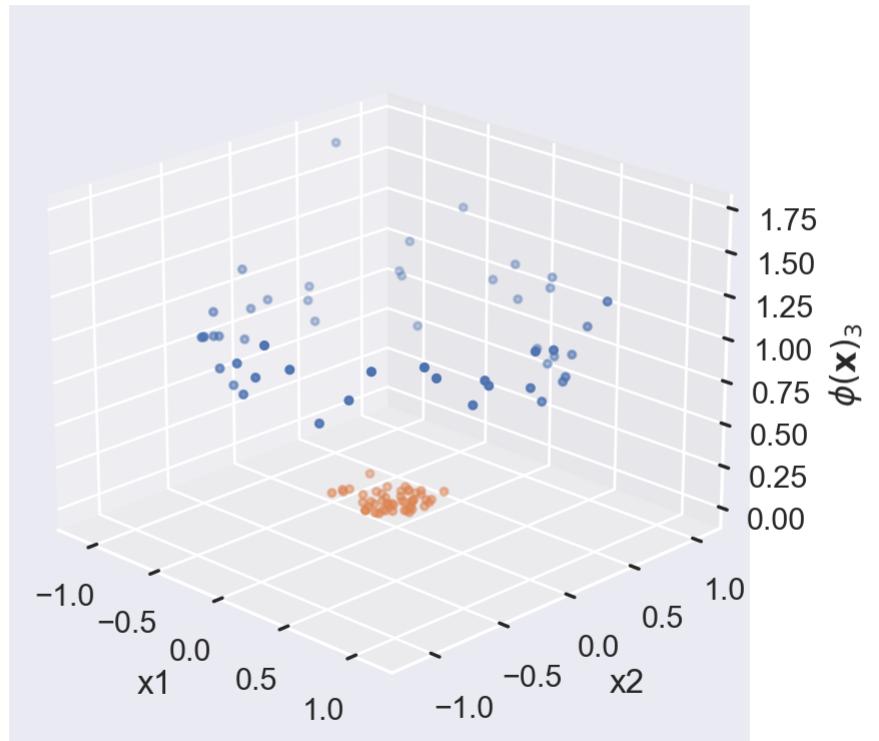
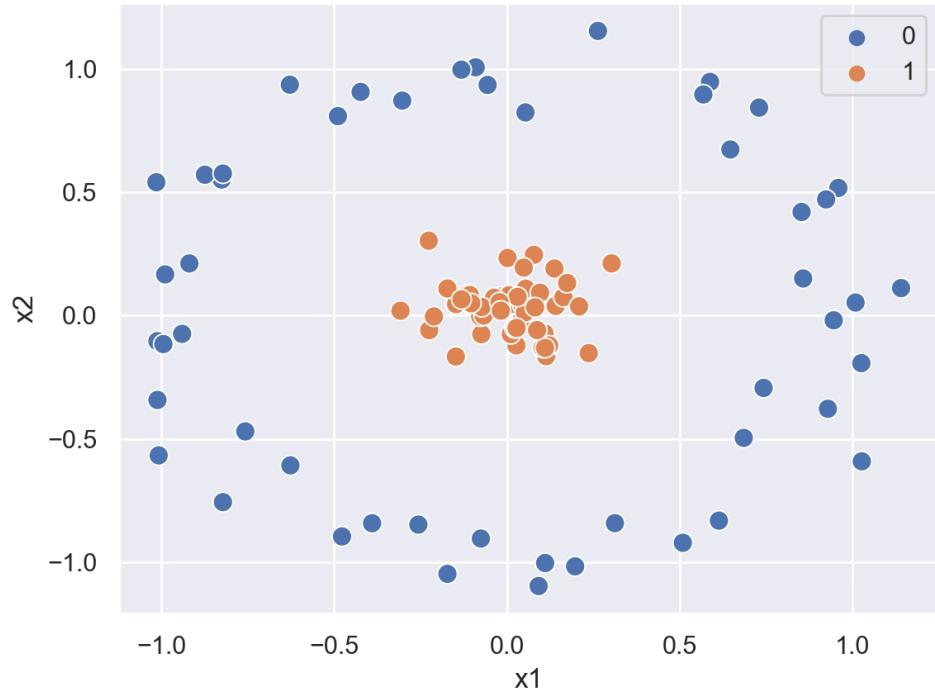
Soft margin

$C = 0.1$

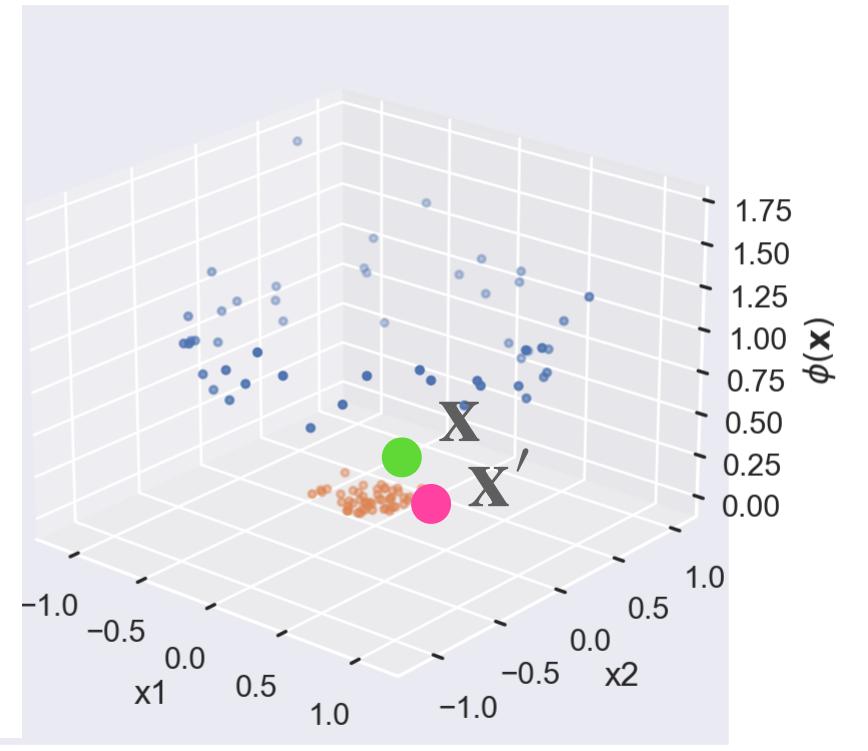
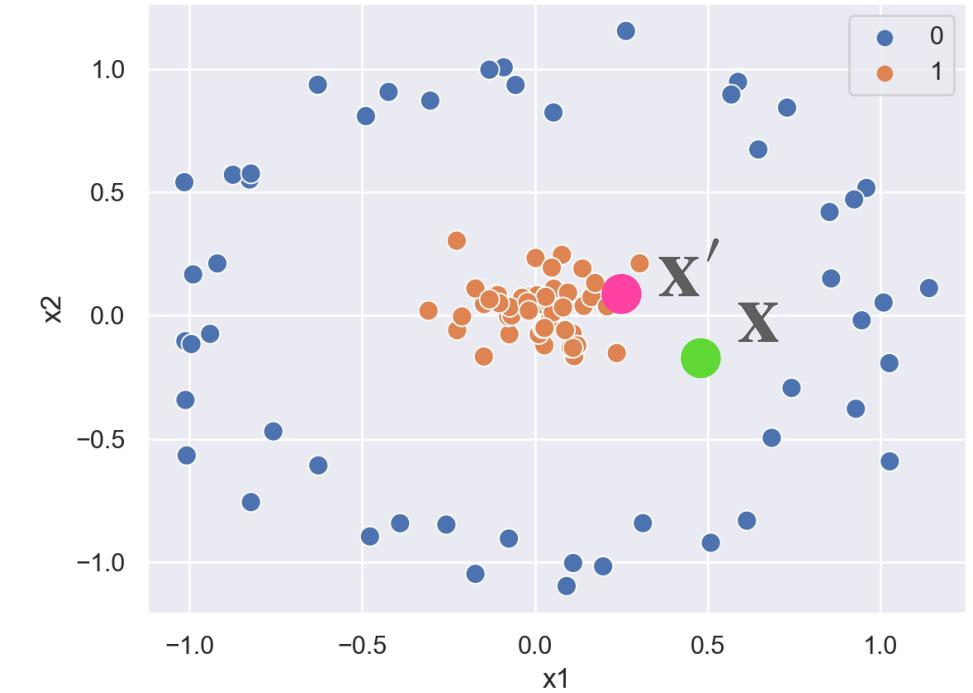


$$\text{Find: } \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \times \sum_{i=1}^n (1 - y_i \times (\mathbf{w}^T \mathbf{x}_i + b))_+$$

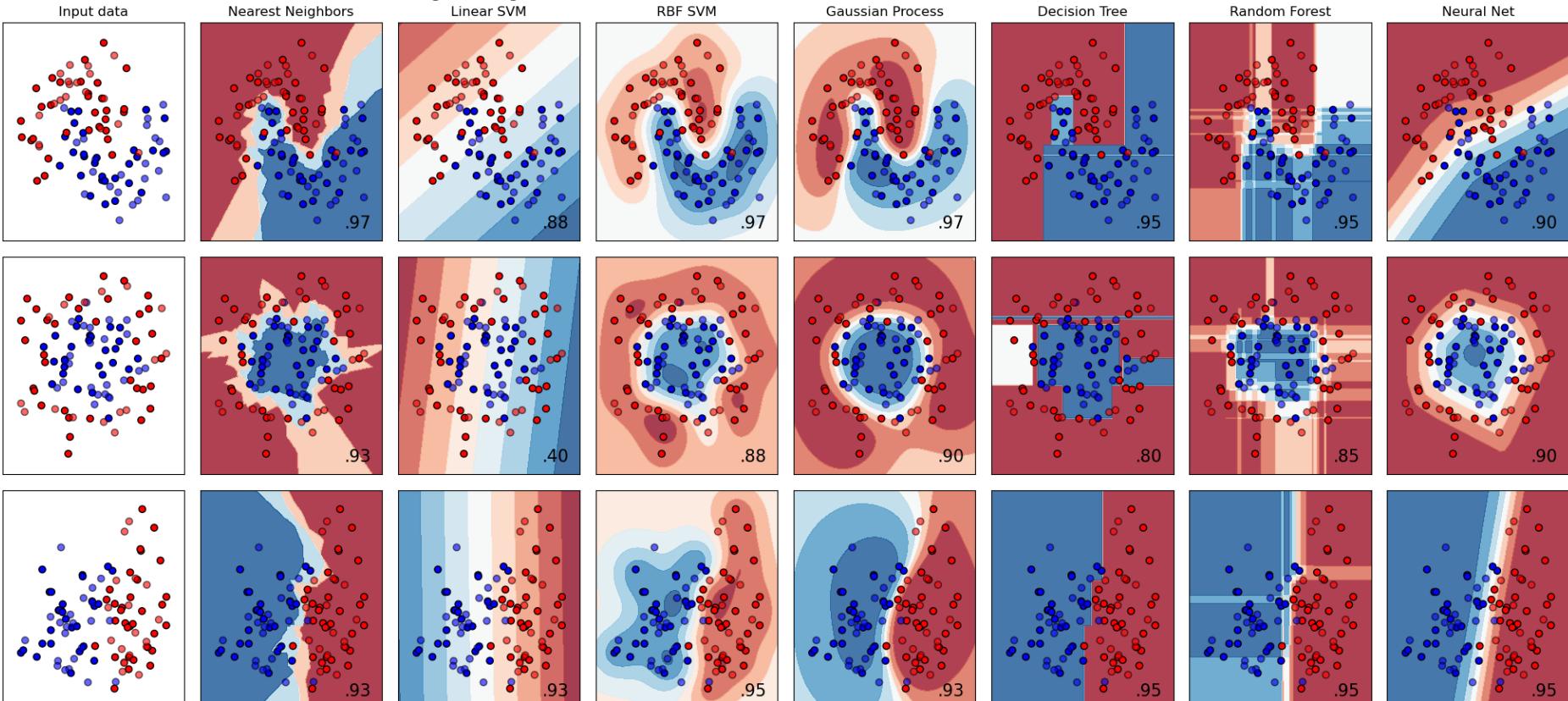
$$\phi(\mathbf{x}) = [x_1, x_2, x_1 \cdot x_2 + x_1^2 \cdot x_2^2]$$



More generally this is the polynomial kernel



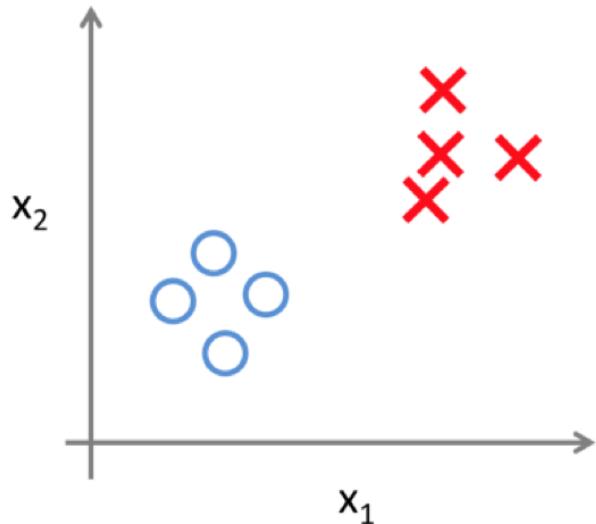
Logistic regression



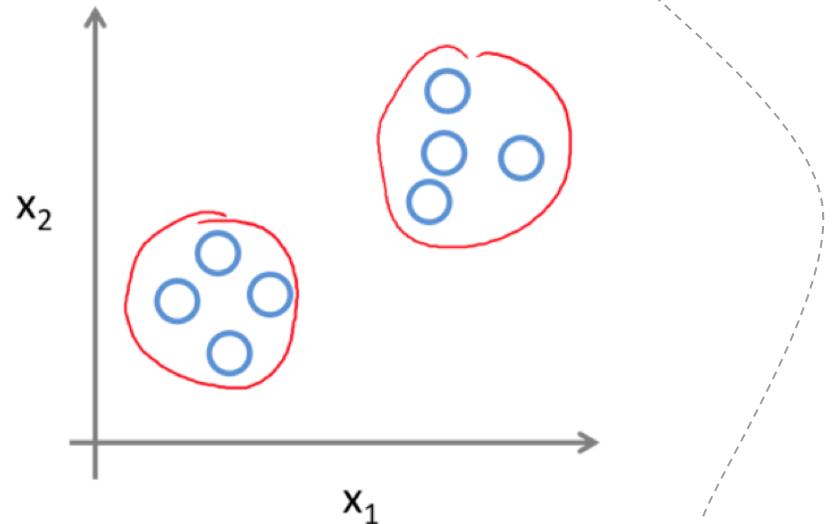
Unsupervised Learning

To modes of machine learning

Supervised Learning



Unsupervised Learning



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

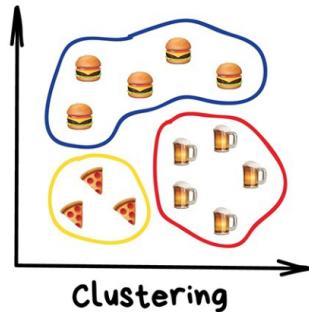
Dimensionality Reduction (Generalization)

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

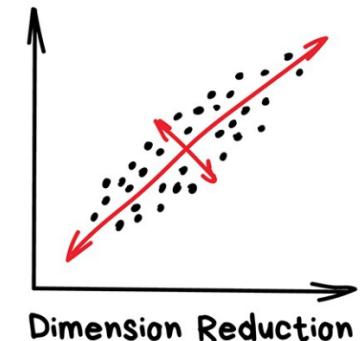


Popular algorithms: [K-means clustering](#), [Mean-Shift](#), [DBSCAN](#)

"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

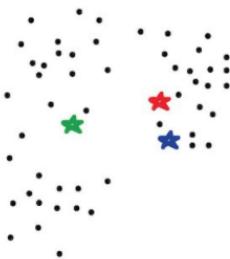


Popular algorithms: [Principal Component Analysis \(PCA\)](#), [Singular Value Decomposition \(SVD\)](#), [Latent Dirichlet allocation \(LDA\)](#), [Latent Semantic Analysis \(LSA, pLSA, GLSA\)](#), [t-SNE](#) (for visualization)

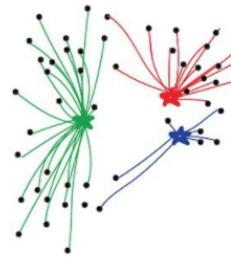
PUT KEBAB KIOSKS IN THE OPTIMAL WAY

(also illustrating the K-means method)

Unsupervised Learning



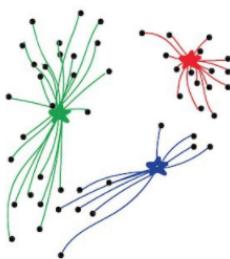
1. Put kebab kiosks in random places in city



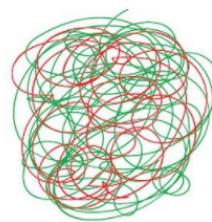
2. Watch how buyers choose the nearest one



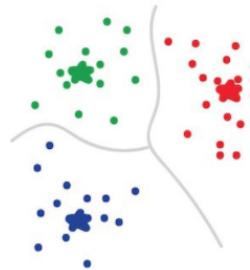
3. Move kiosks closer to the centers of their popularity



4. Watch and move again



5. Repeat a million times



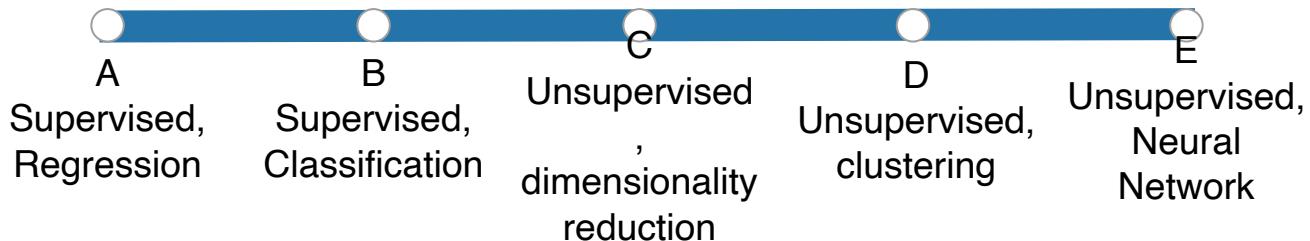
6. Done!
You're god of kebabs!

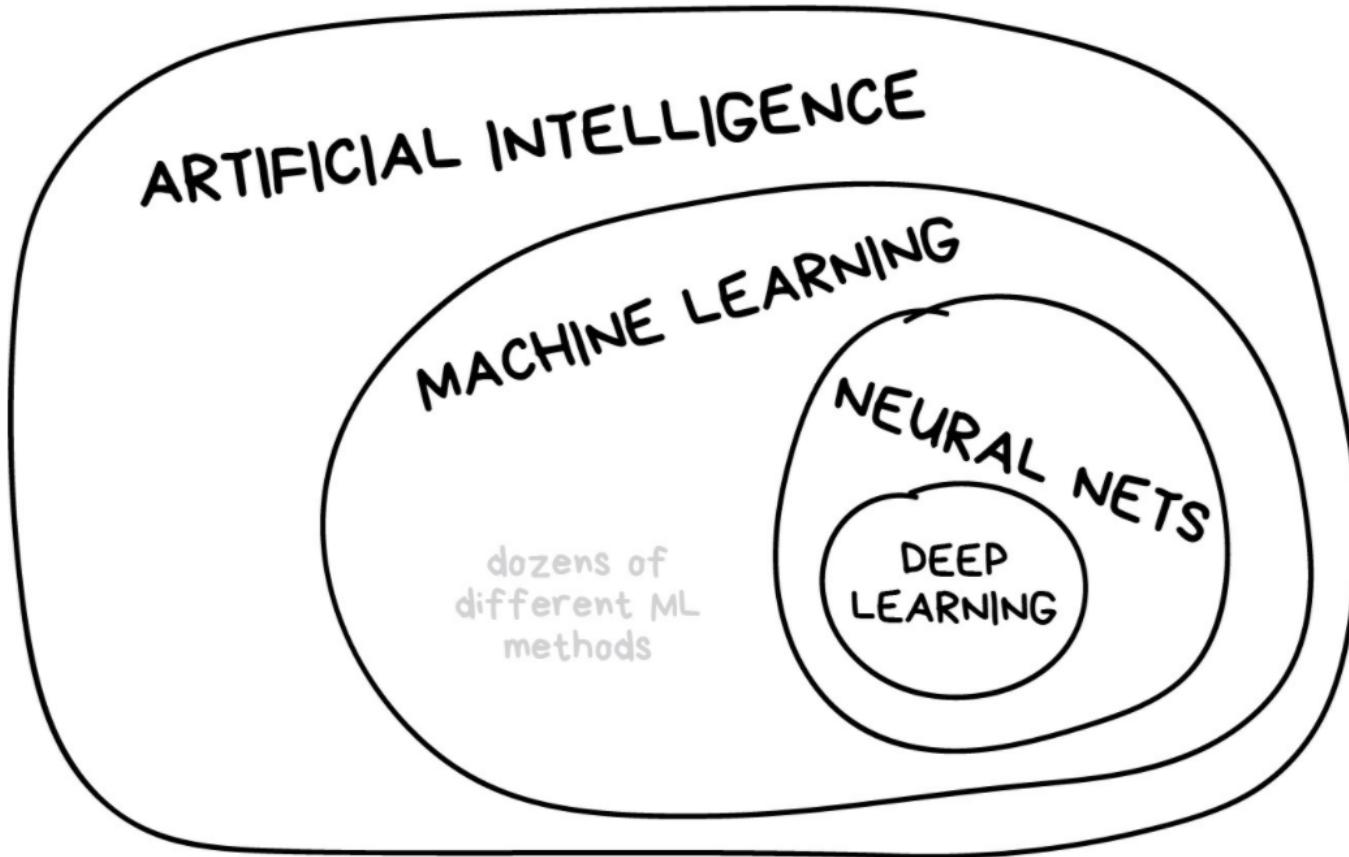


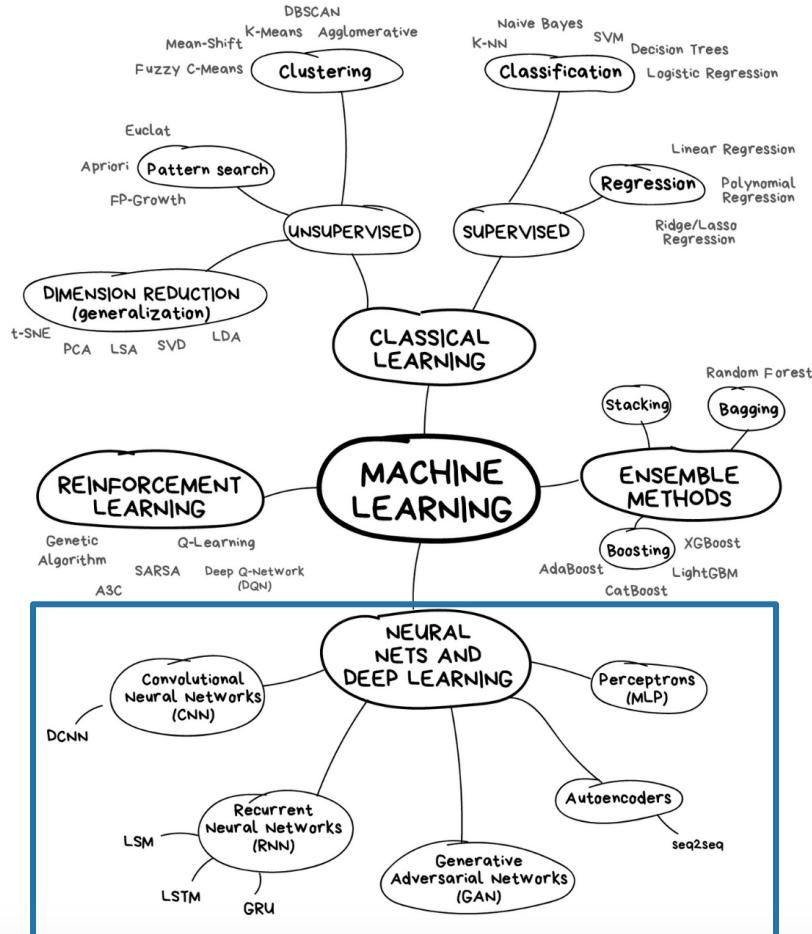
Prediction Approach

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

How would you approach this with machine learning?





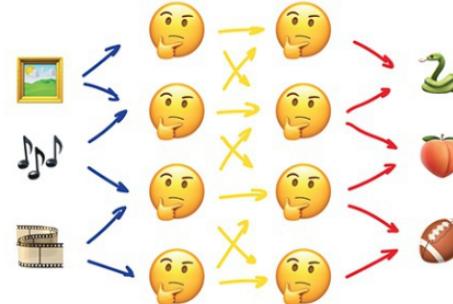


https://vas3k.com/blog/machine_learning/

"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:

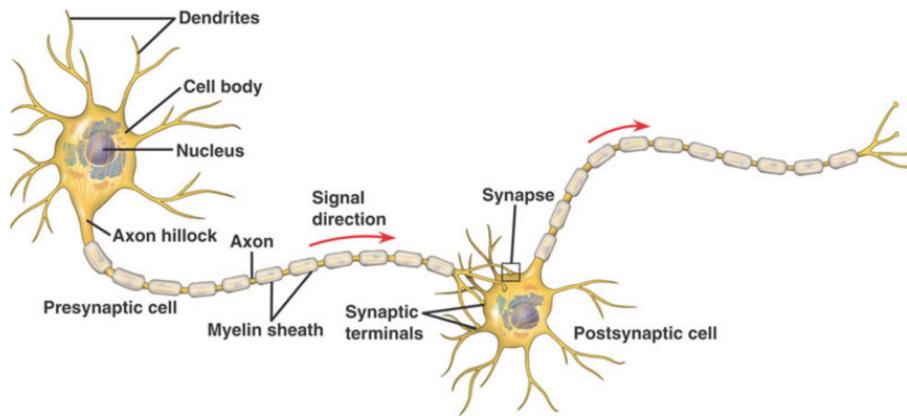
- Replacement of all algorithms above
- Object identification on photos and videos
- Speech recognition and synthesis
- Image processing, style transfer
- Machine translation



Neural Networks

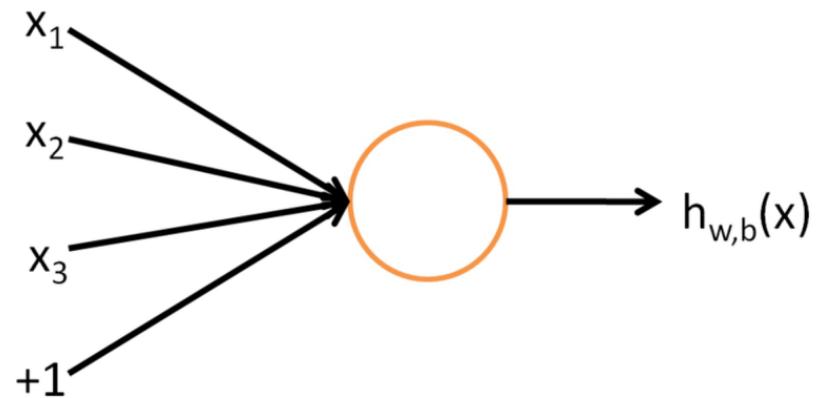
Popular architectures: Perceptron, Convolutional Network (CNN), Recurrent Networks (RNN), Autoencoders

WHAT IS A NEURON?



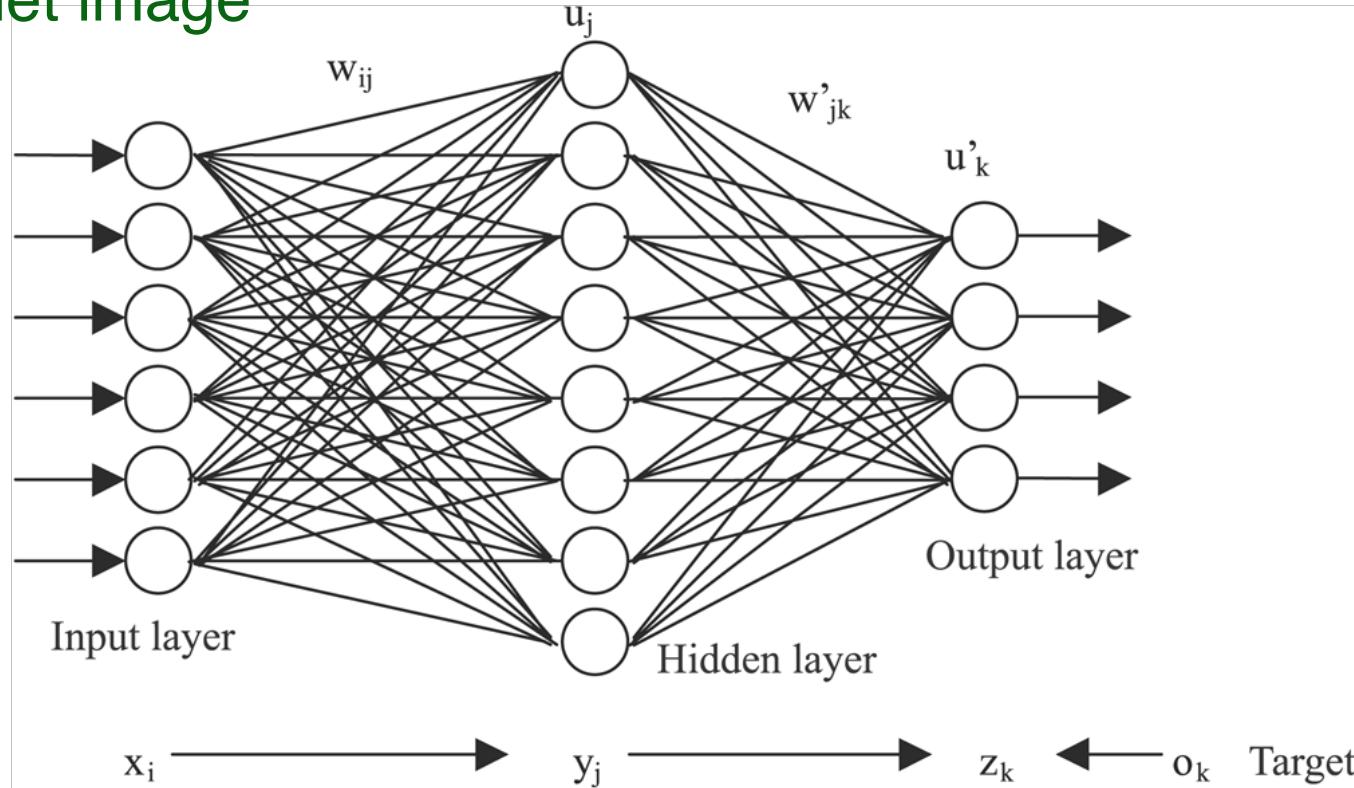
- Receives signal on synapse
- When trigger sends signal on axon

MATHEMATICAL NEURON

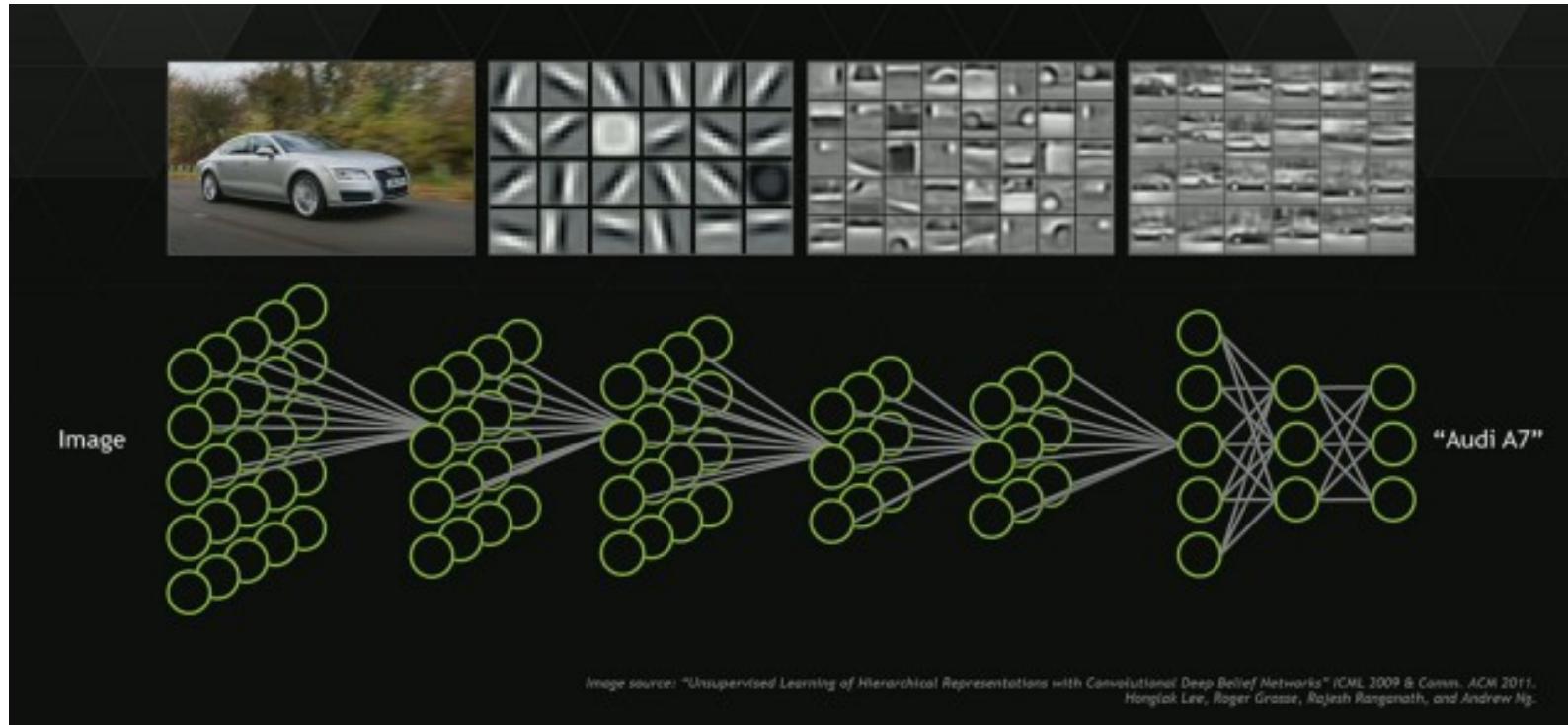


- Mathematical abstraction, inspired by biological neuron
- Either on or off based on sum of input

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

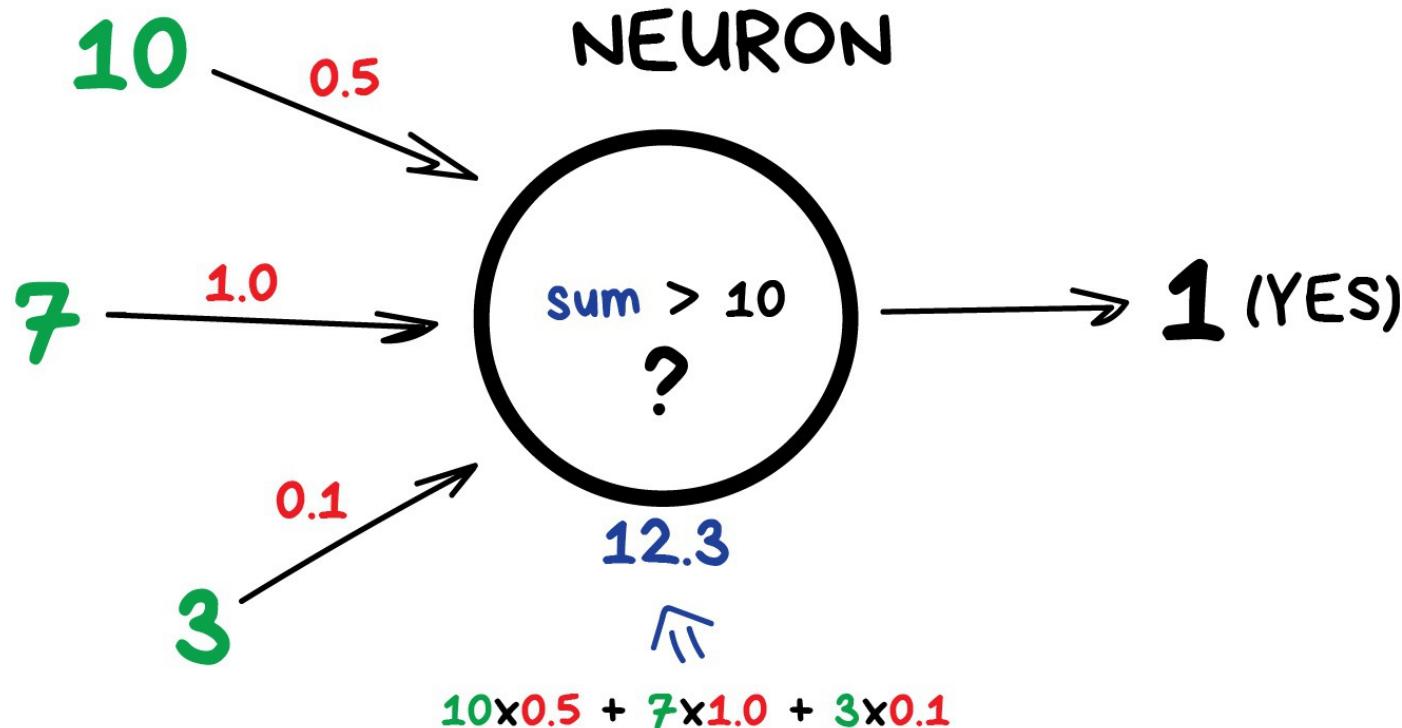


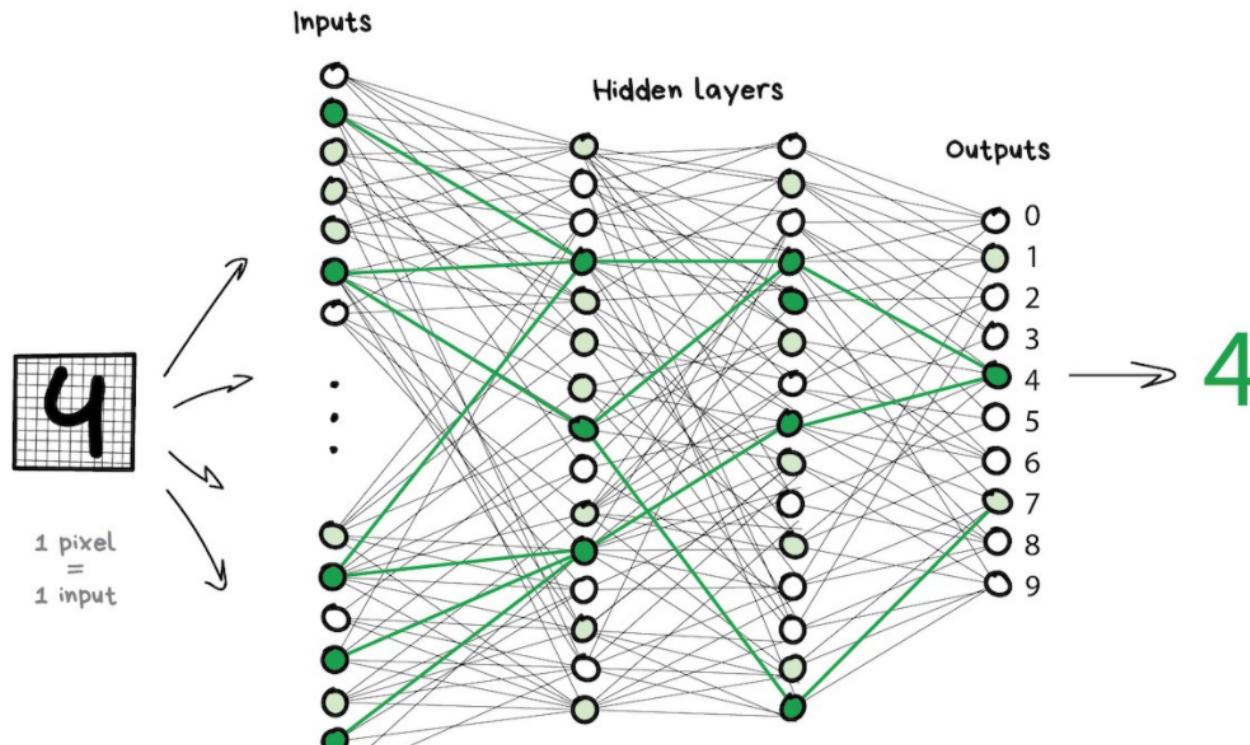
Convolutional neural networks



<https://towardsdatascience.com/understanding-residual-networks-9add4b664b03>

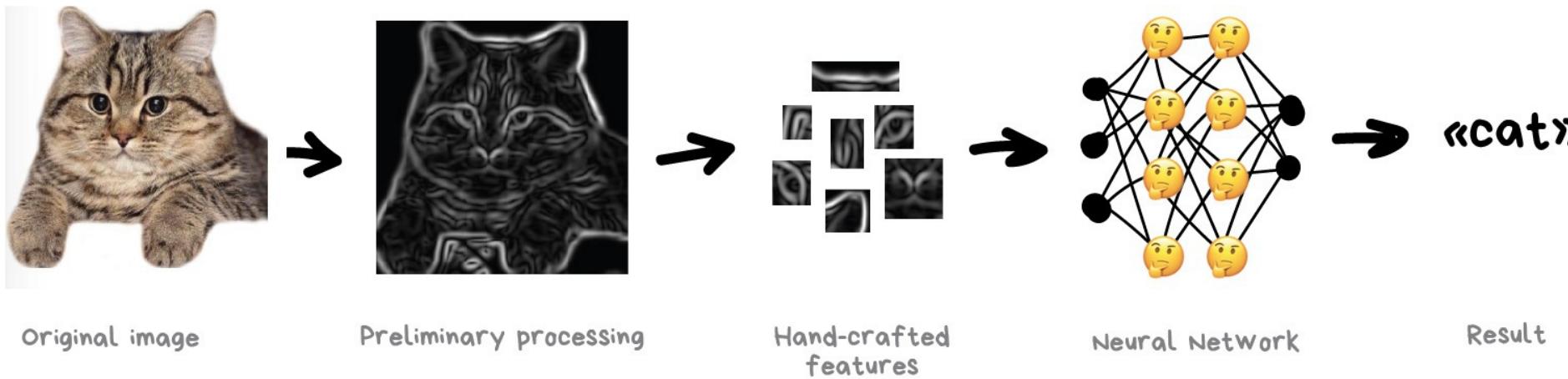
These weights tell the neuron to respond more to one input and less to another. Weights are adjusted when training — that's how the network learns. Basically, that's all there is to it.



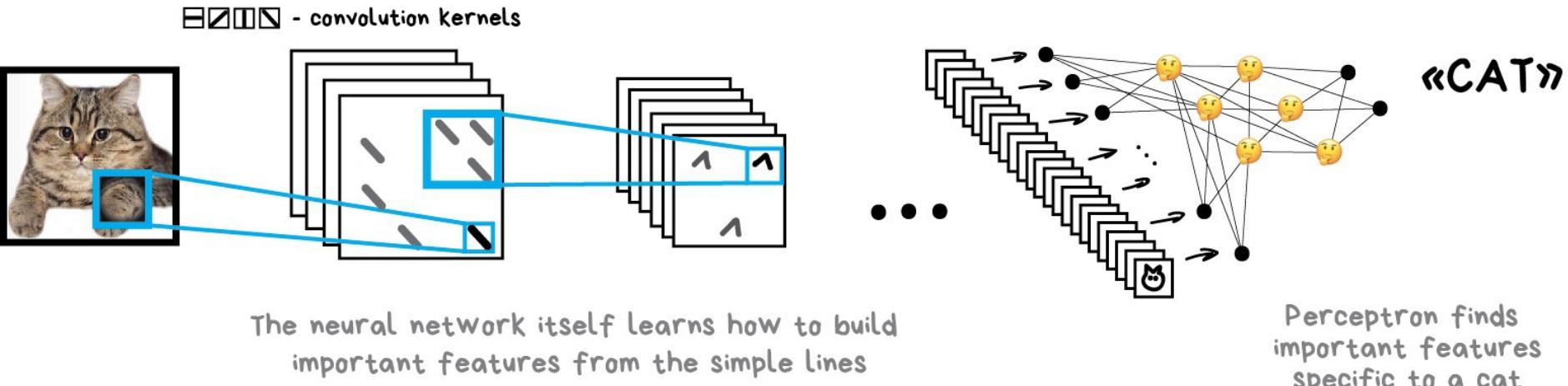


MULTILAYER PERCEPTRON (MLP)

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

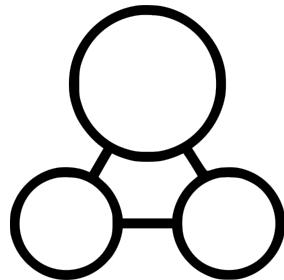
↑ ↑
store gallon

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

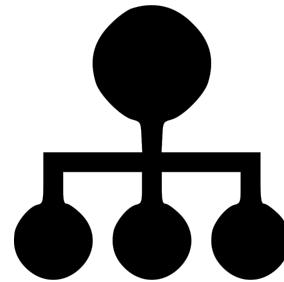
ML jargon

The model is the algorithm + hyperparameters; parameters are the result of training the model

- Algorithm:
 - some math to optimally solve a problem given some training data
- Solver:
 - The software implementation/approximation of the math; many different ways to get the job done
- Hyperparameters:
 - Knobs/Dials that affect how the algorithm solves the problem, the solver is one such
- Parameters:
 - the solution to the optimization problem being solved on the training data



+	-
x	=



model selection

OUR GOAL IS TO PREVENT OVERFITTING!!!!

<https://arstechnica.com/health/2023/11/ai-with-90-error-rate-forces-elderly-out-of-rehab-nursing-homes-suit-claims/>

Preventing overfitting is important!

When AI models make errors it is typically because the dataset is a bad sample, leading to increased likelihood of overfitting patterns that are particular to this bad sample.

Generically there are two issues at play here.

Bad procedures (or carelessness) led to a bad model

Then ethical failings led to this bad model being put into service and kept in service

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

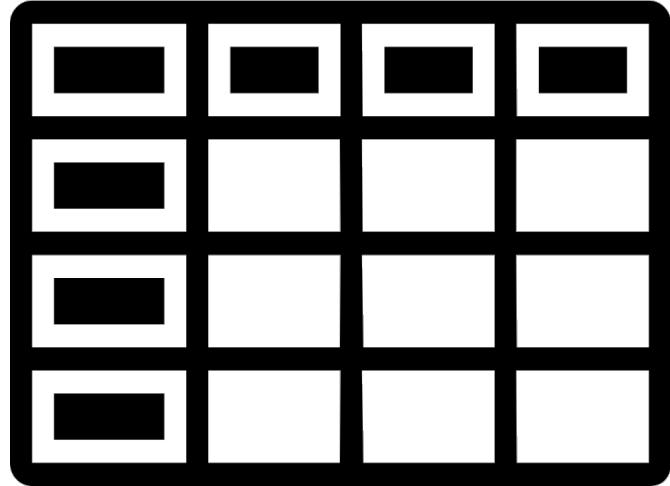
DESPICABLE —

UnitedHealth uses AI model with 90% error rate to deny care, lawsuit alleges

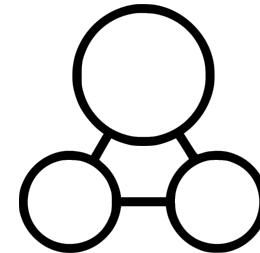
For the largest health insurer in the US, AI's error rate is like a feature, not a bug.

BETH MOLE - 11/16/2023, 3:37 PM



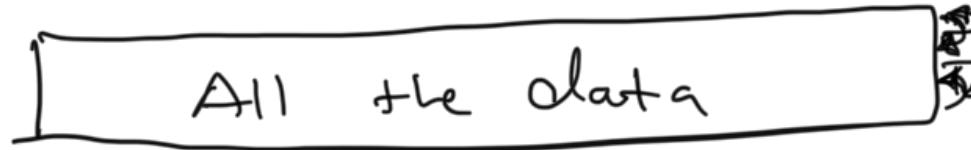


BIGGER
datasets



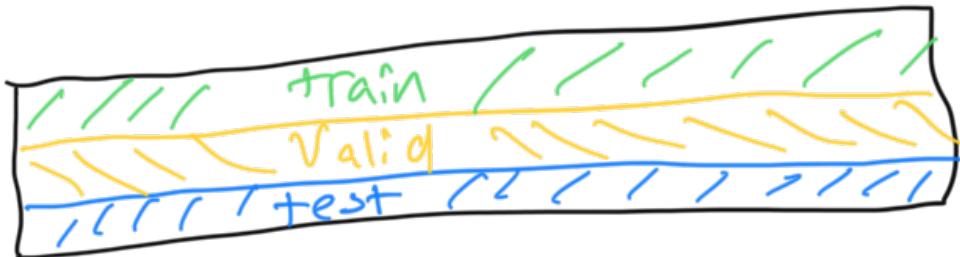
SIMPLER
models

①



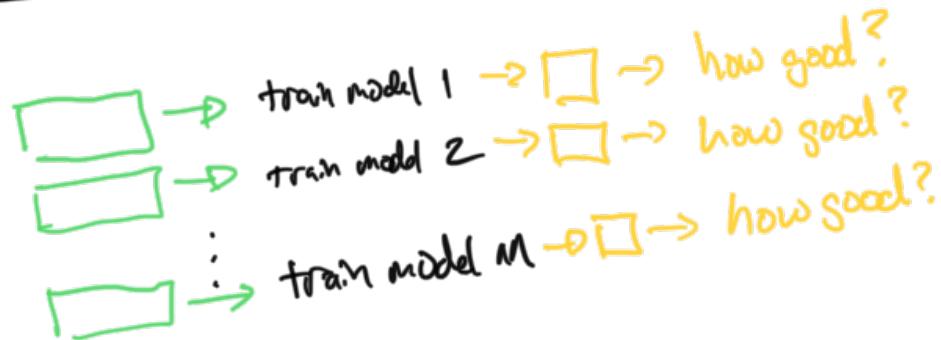
randomly shuffle
rows of dataset

②



Split into
training, validation
and test sets

③



If you have M
models to try do this
once for each

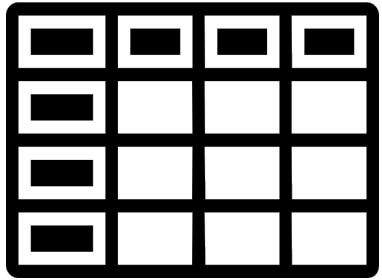
④



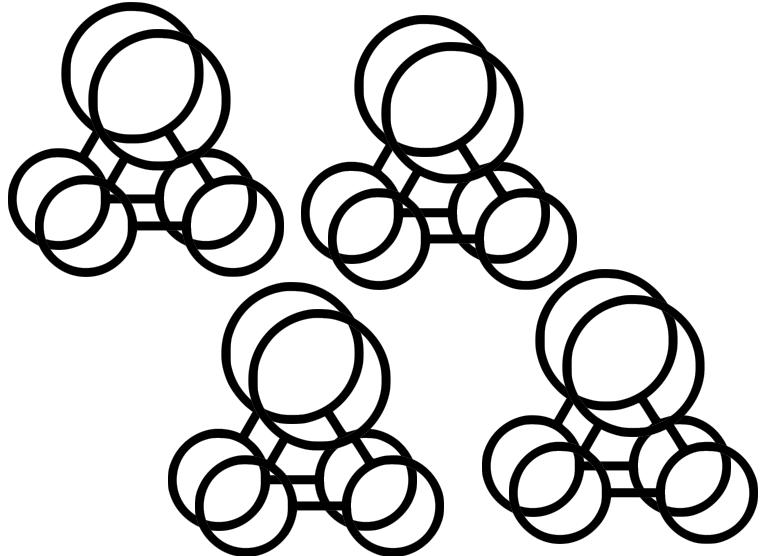
Pick best model from
validation set performance
train it on $T+V$ then test
it on Test set

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



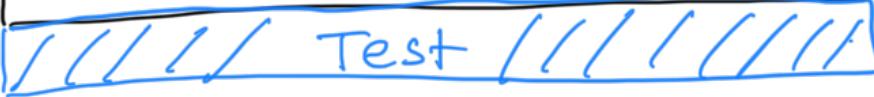
SMALLER
datasets



COMPLEX
models

- ①

All the data

randomly shuffle rows of dataset
- ② split off a test set

- ③

Fold 1
Fold 2
Fold 3
Fold K

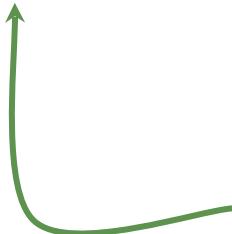
split remaining data into K folds with even(ish) amounts of data w/ same(ish) distributions
- ④  Every split gets its turn as the validation set, use all other splits as training. Validation performance is the mean across splits
- ⑤ Do steps ③ + ④ M times to test M models, pick best validation performance model 



model assessment

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

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

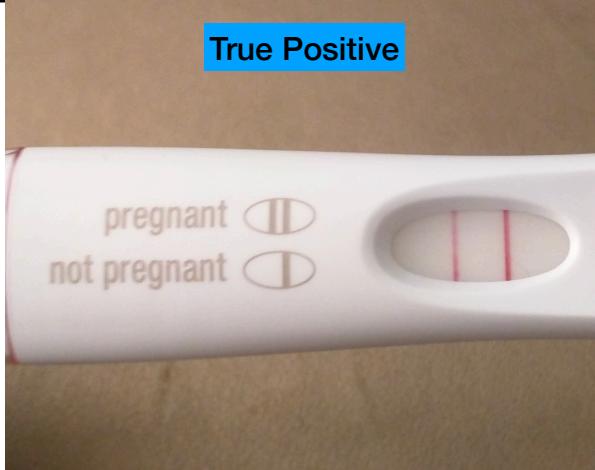
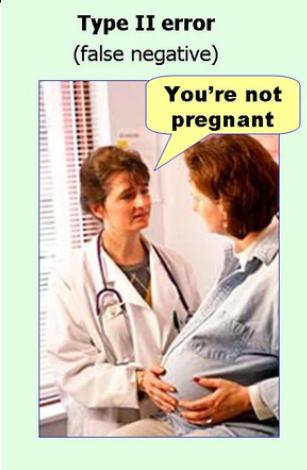
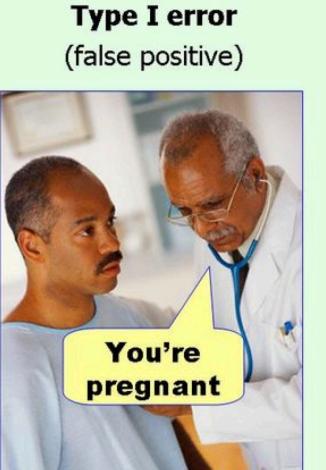
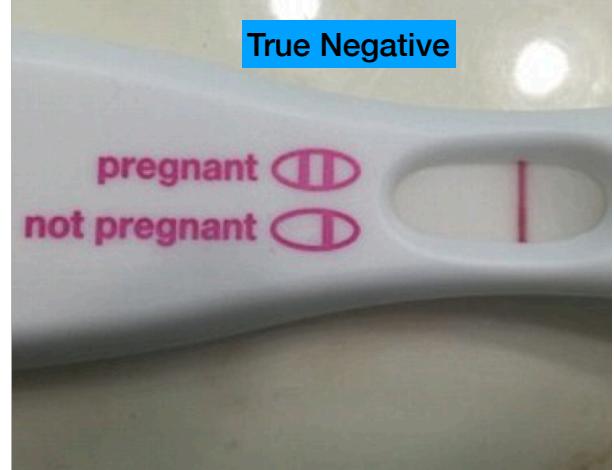
Accuracy can mislead

- 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	
Predicted	Positive	True Positive (TP)	False Positive
	Negative	False Negative	True Negative (TN)

The image shows a 2x2 matrix diagram illustrating medical test results. The columns represent 'Actual' status (Positive or Negative) and the rows represent 'Predicted' status (Positive or Negative). The four quadrants are:

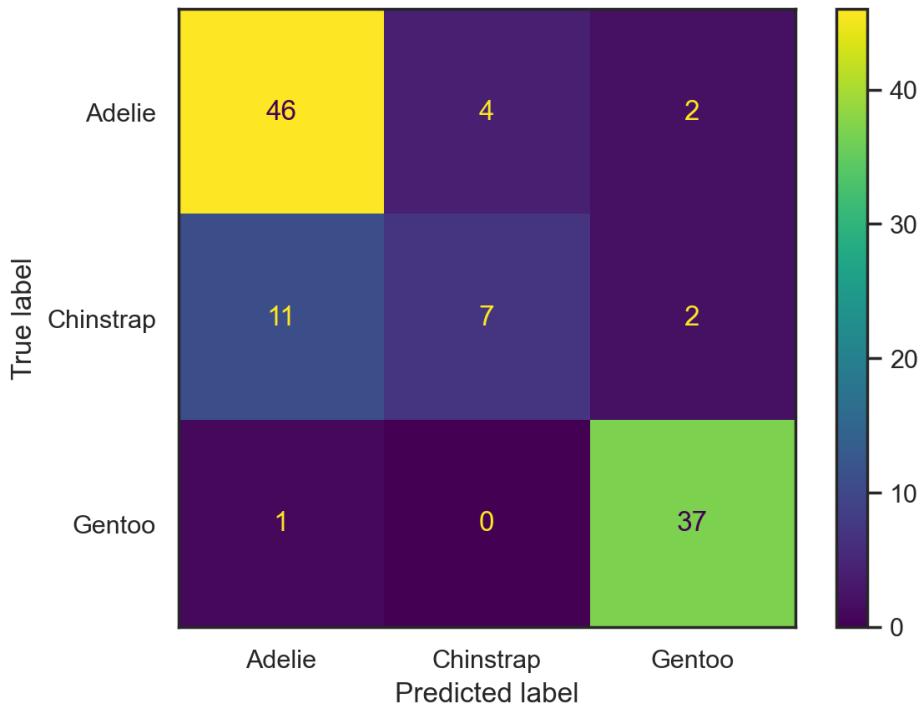
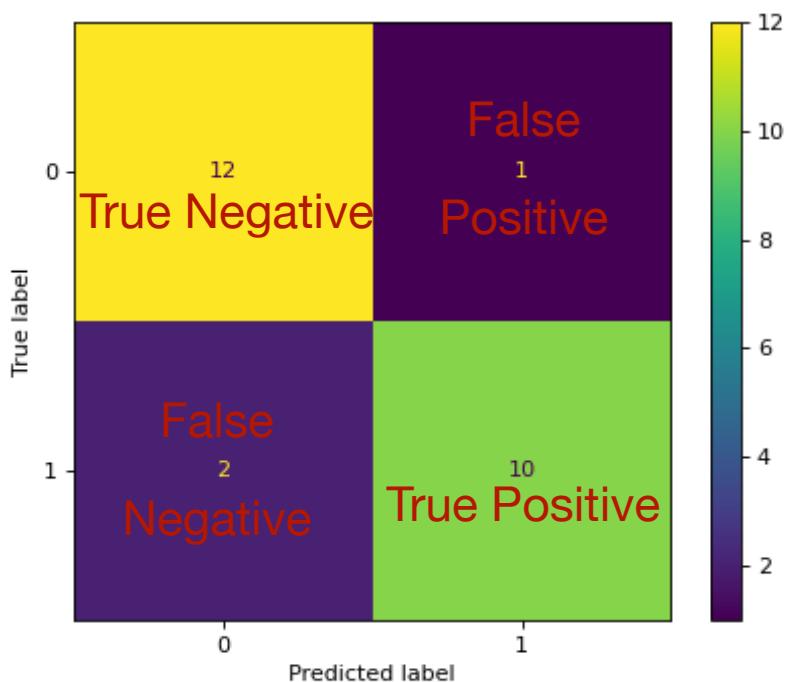
- True Positive (TP):** Actual Positive, Predicted Positive. Image: A digital pregnancy test strip showing a single line for 'not pregnant'.
- False Positive:** Actual Positive, Predicted Negative. Image: A doctor examining a pregnant woman's belly.
- False Negative:** Actual Negative, Predicted Negative. Image: A doctor telling a patient they are pregnant.
- True Negative (TN):** Actual Negative, Predicted Positive. Image: A digital pregnancy test strip showing two lines for 'pregnant'.

	Predicted +		Predicted -
Truly +	<p>True Positive</p>  A photograph of a digital pregnancy test device. The screen displays two horizontal red lines, indicating a positive result. The words "pregnant" and "not pregnant" are visible above and below the test area respectively.		<p>Type II error (false negative)</p>  A photograph of a doctor in a white coat and stethoscope around their neck, talking to a pregnant woman. A yellow speech bubble from the doctor contains the text "You're not pregnant".
Truly -	<p>Type I error (false positive)</p>  A photograph of a doctor in a white coat and stethoscope around their neck, talking to a man. A yellow speech bubble from the doctor contains the text "You're pregnant".		<p>True Negative</p>  A photograph of a digital pregnancy test device. The screen displays one horizontal red line, indicating a negative result. The words "pregnant" and "not pregnant" are visible above and below the test area respectively.

confusion matrix

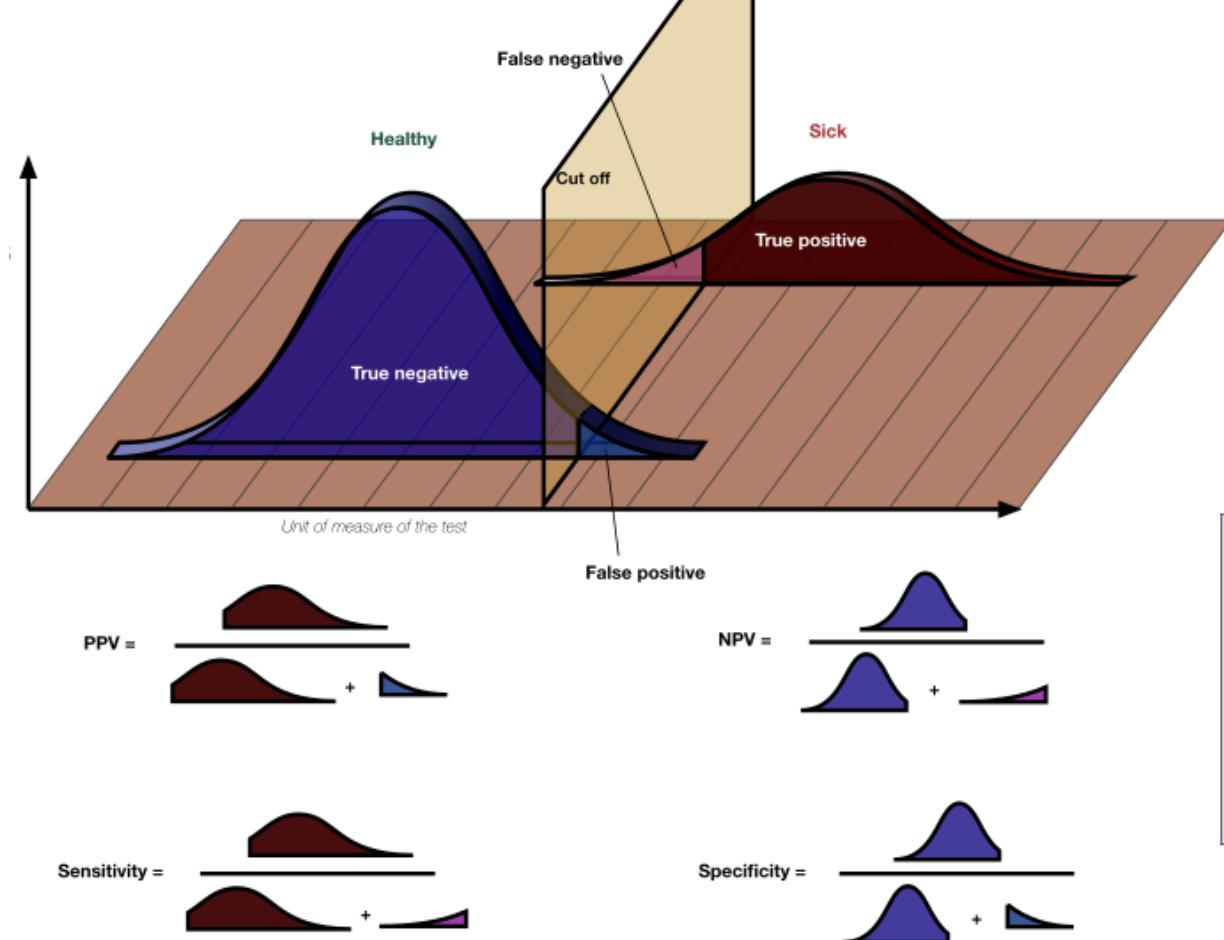
		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Confusion matrix



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

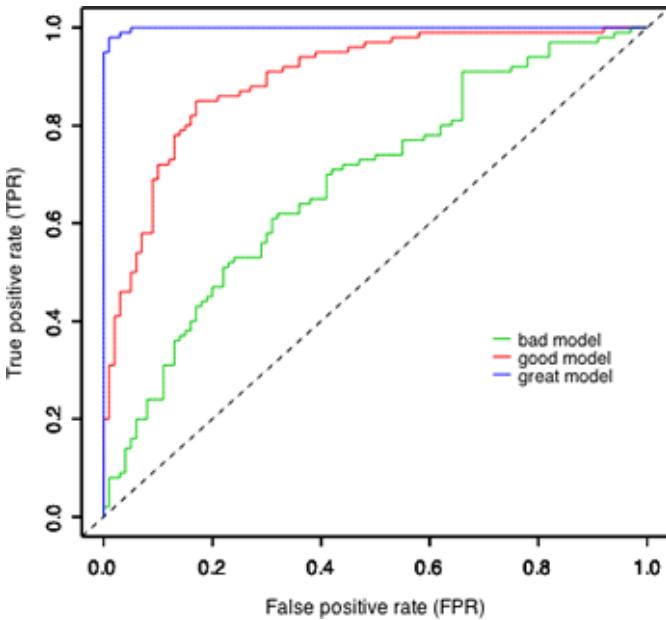
categorical variable prediction



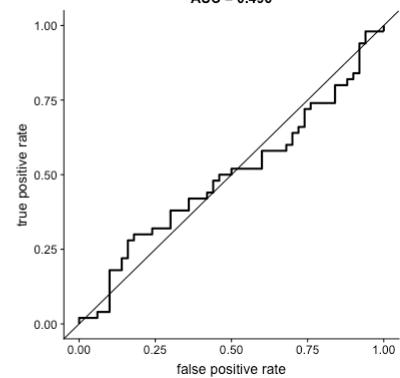
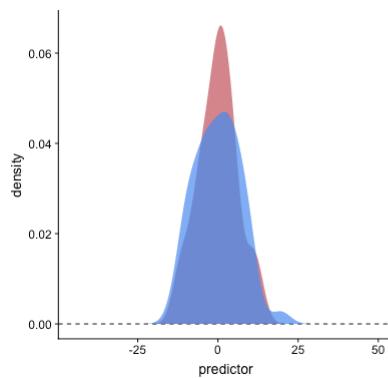
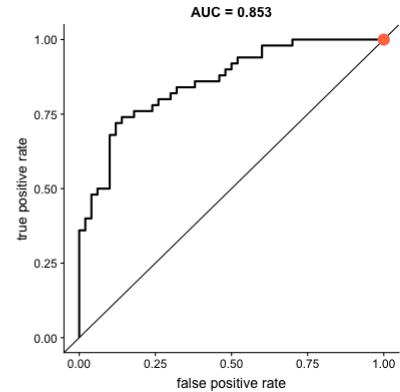
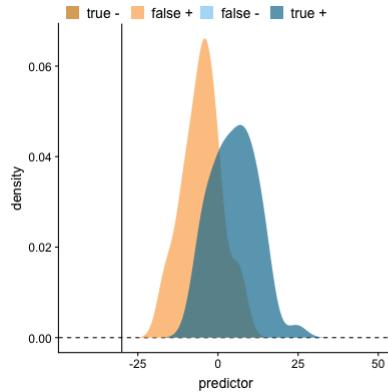
		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Receiver Operating Characteristic

ROC-AUC



The value can range from 0 to 1. However AUC score of a random classifier for balanced data is 0.5





Prediction Approach



[https://forms.gle/
ubN7LFy3FdbSmcEi6](https://forms.gle/ubN7LFy3FdbSmcEi6)