Machine learning

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego
jfleischer@ucsd.edu

@jasongfleischer

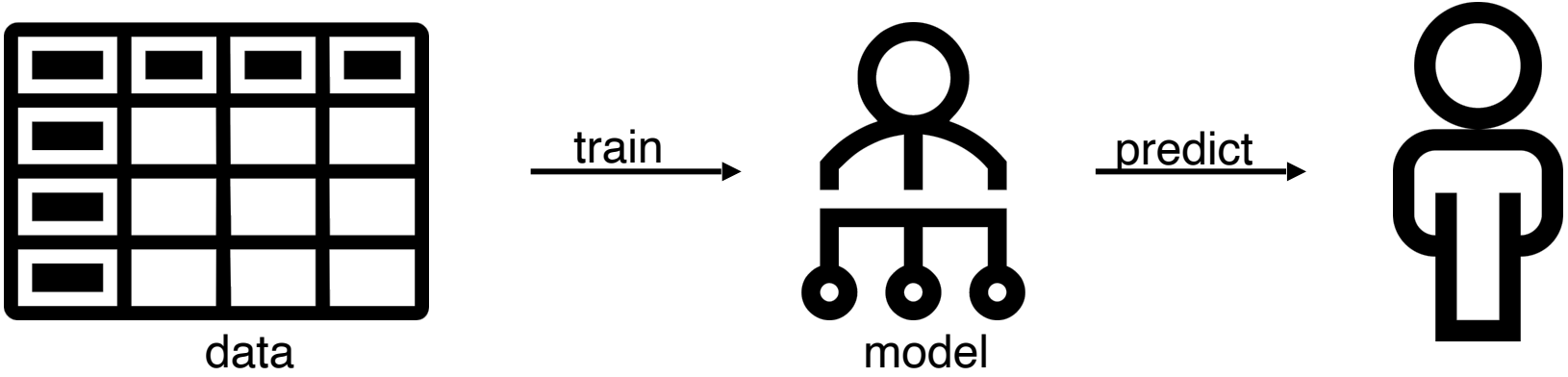
https://jgfleischer.com

- **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!



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

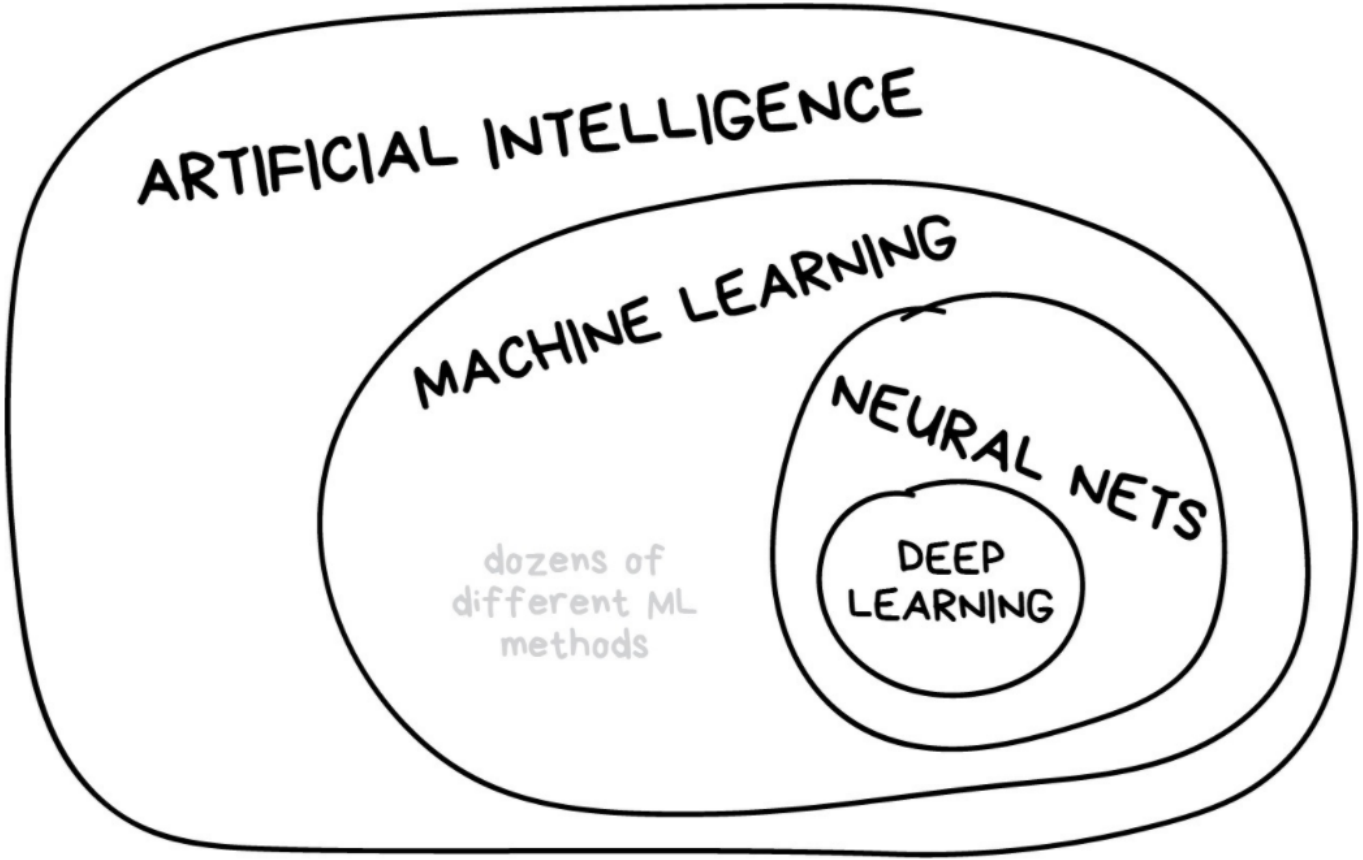


Image source: https://vas3k.com/blog/machine_learning/

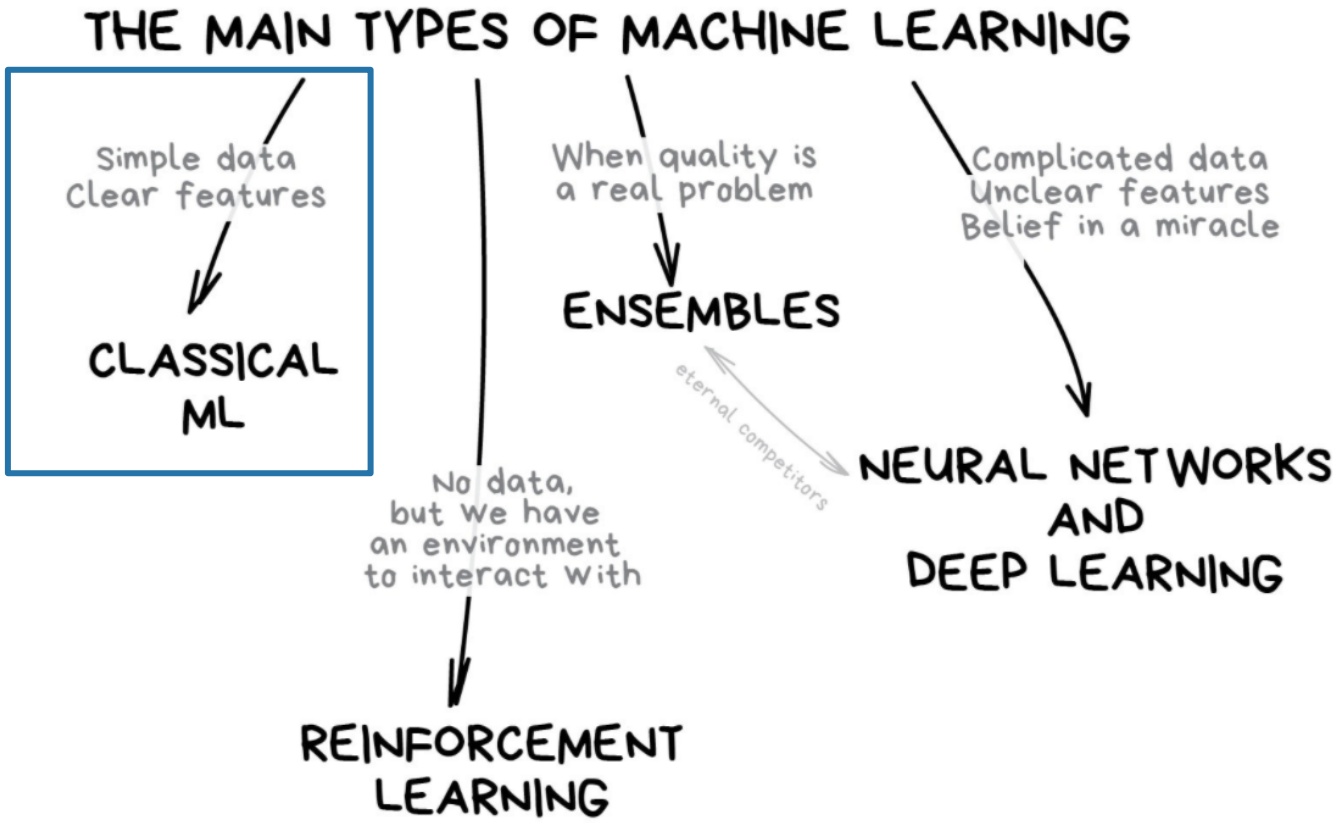
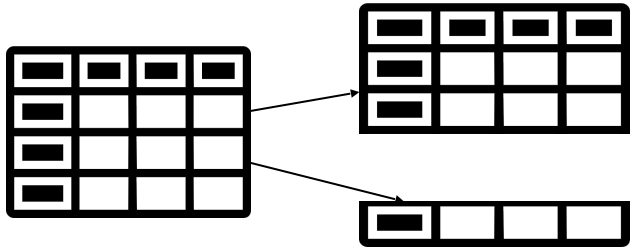


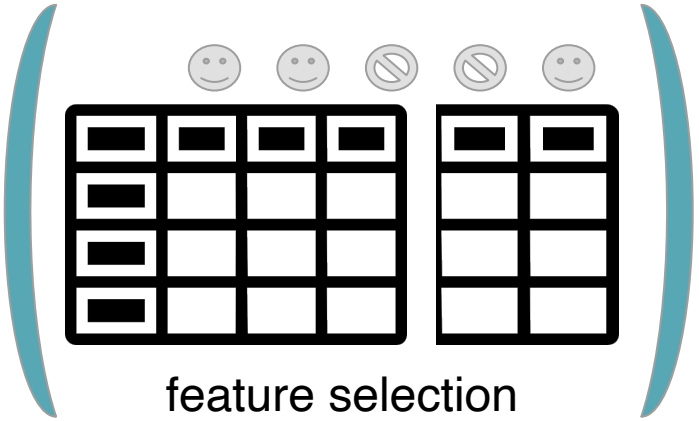
Image source: https://vas3k.com/blog/machine_learning/

Machine Learning Generalizations

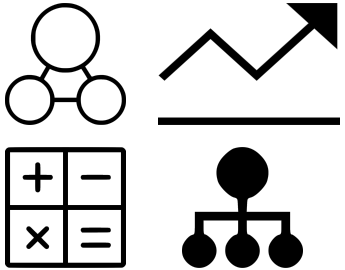
Basic Steps to Prediction



data
partitioning



feature selection

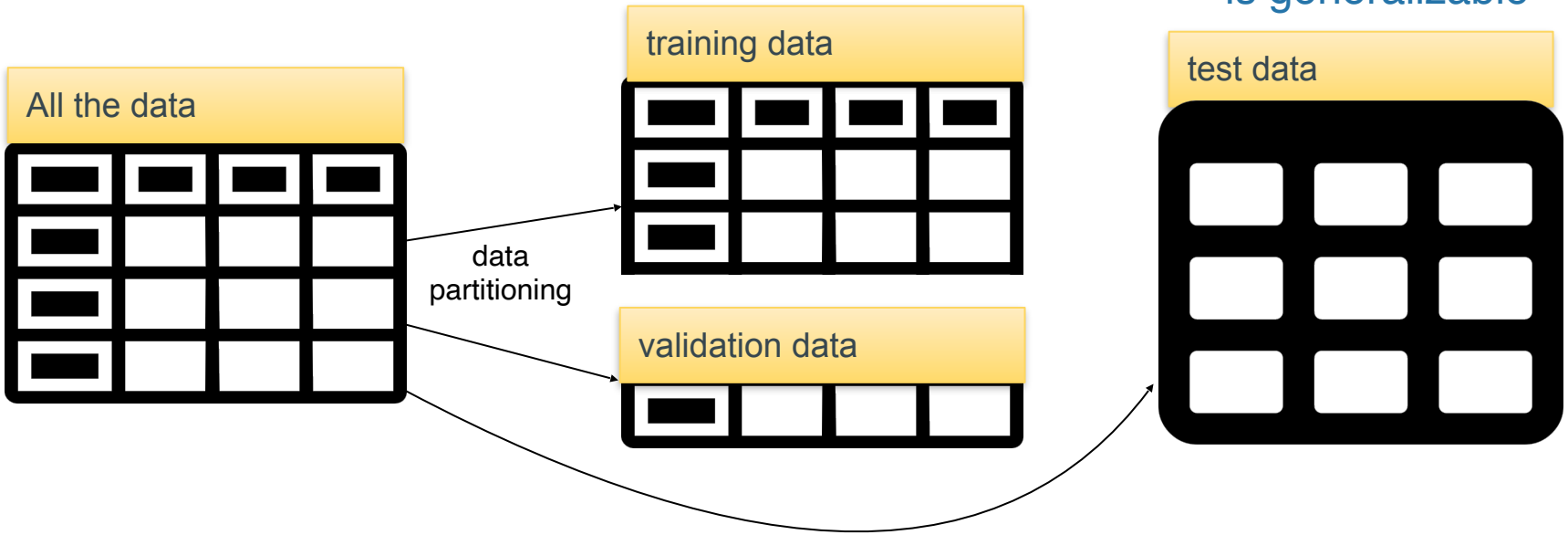


model selection

model assessment

the data used to build your model

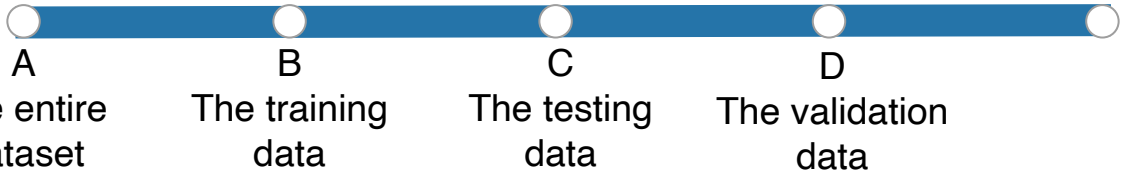
Data not used in training or validating the model; used to assess if model is generalizable

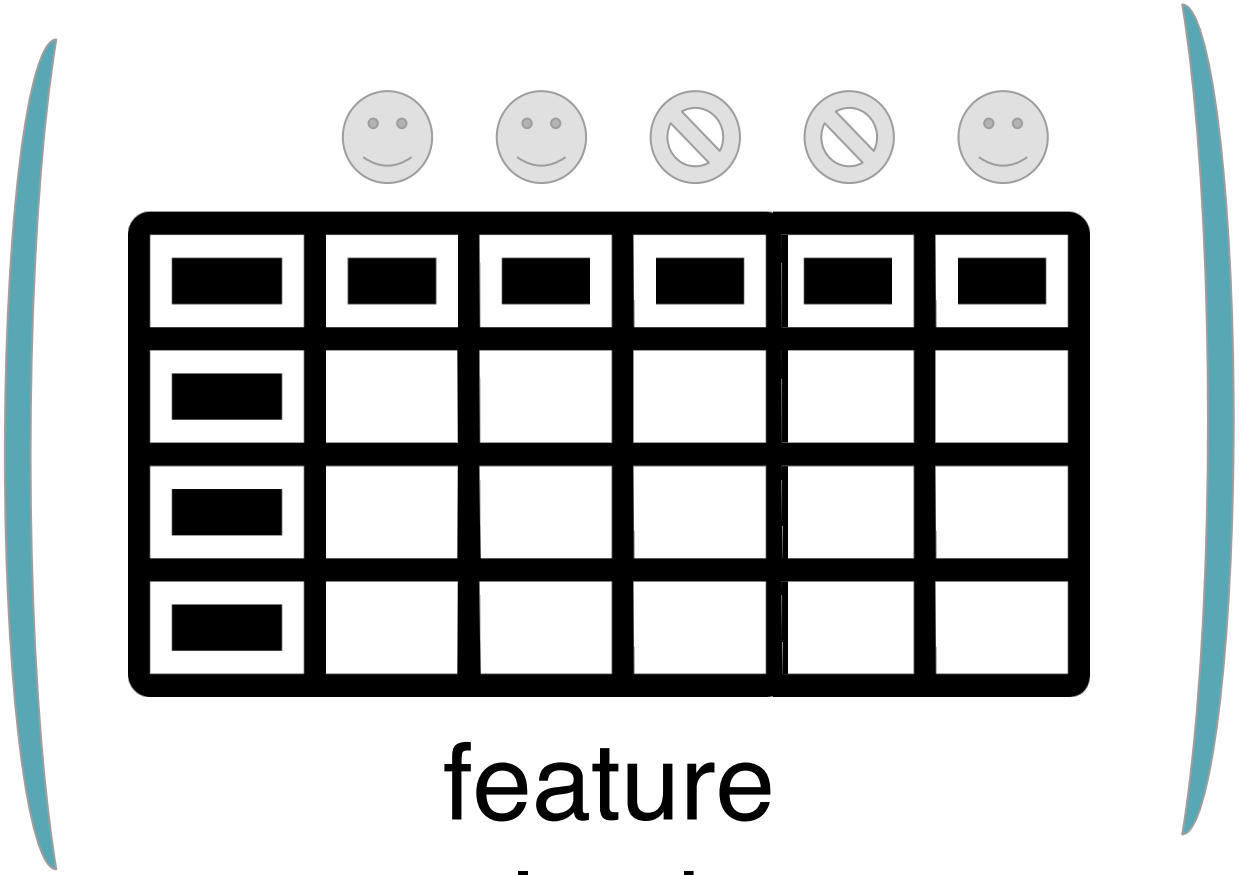


Data not used in training the model; used to fine-tune the model to increase prediction accuracy

Data Partitioning

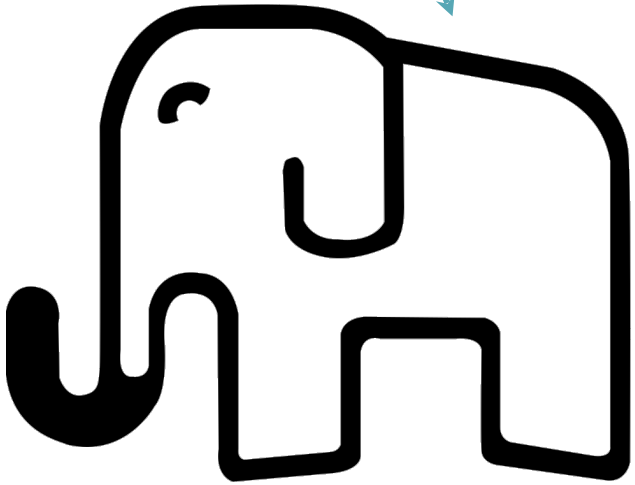
What portion of the data are typically used for building the model?

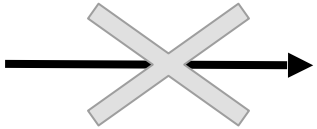


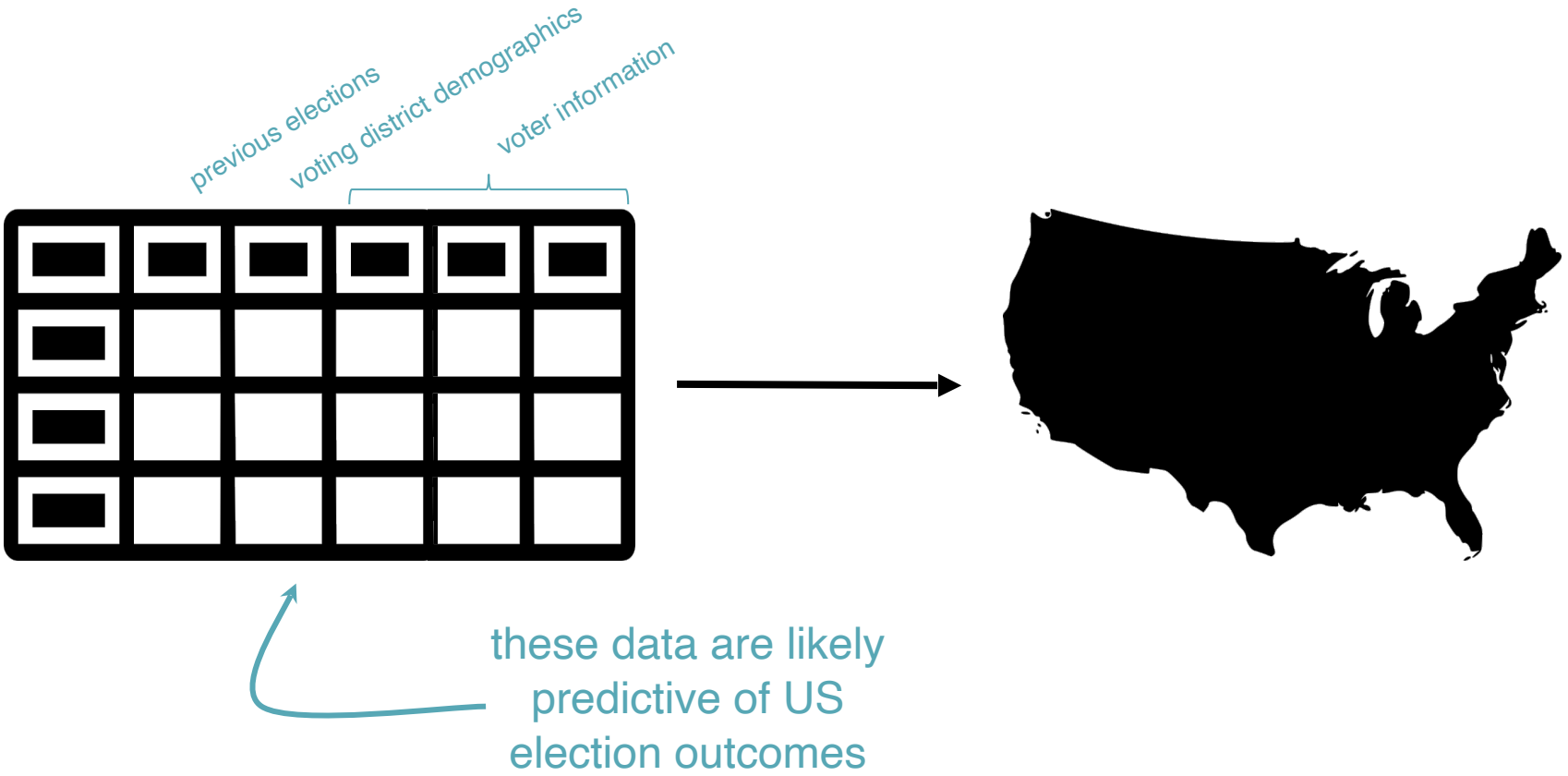


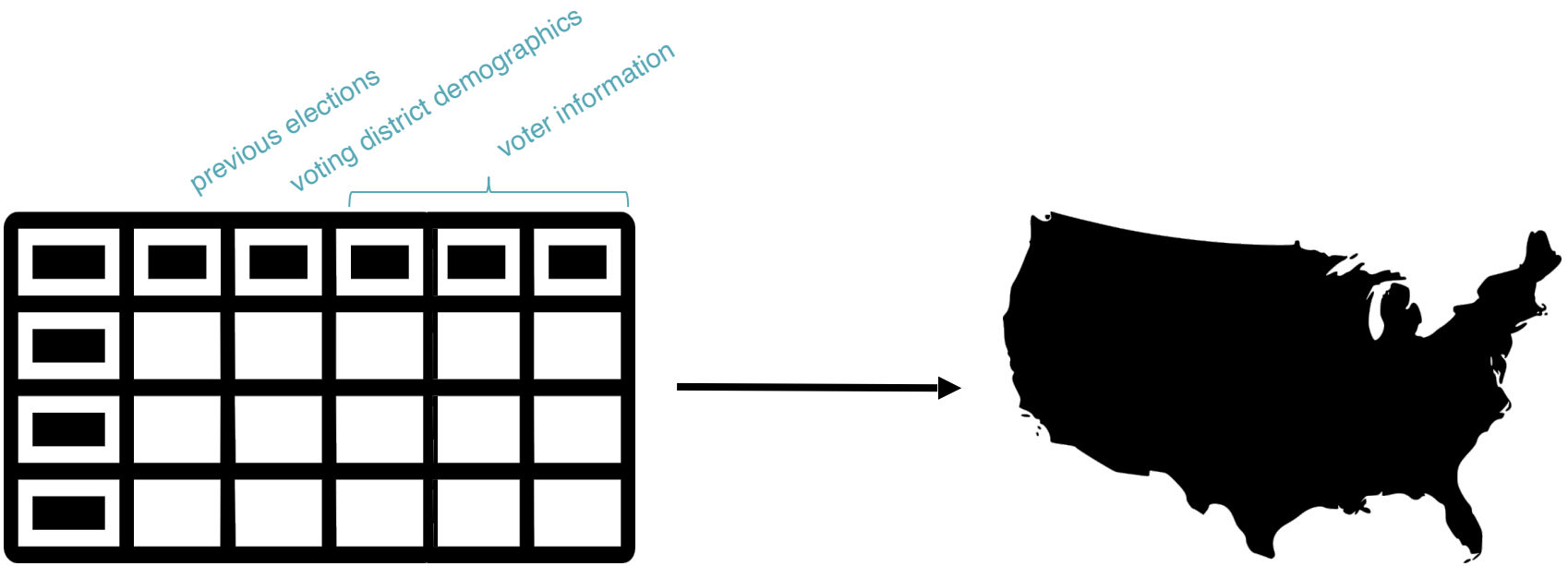
feature
selection

elephant height data
are likely not predictive
of US elections

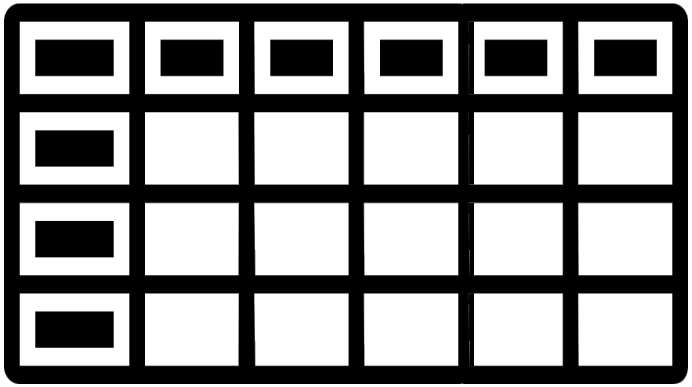




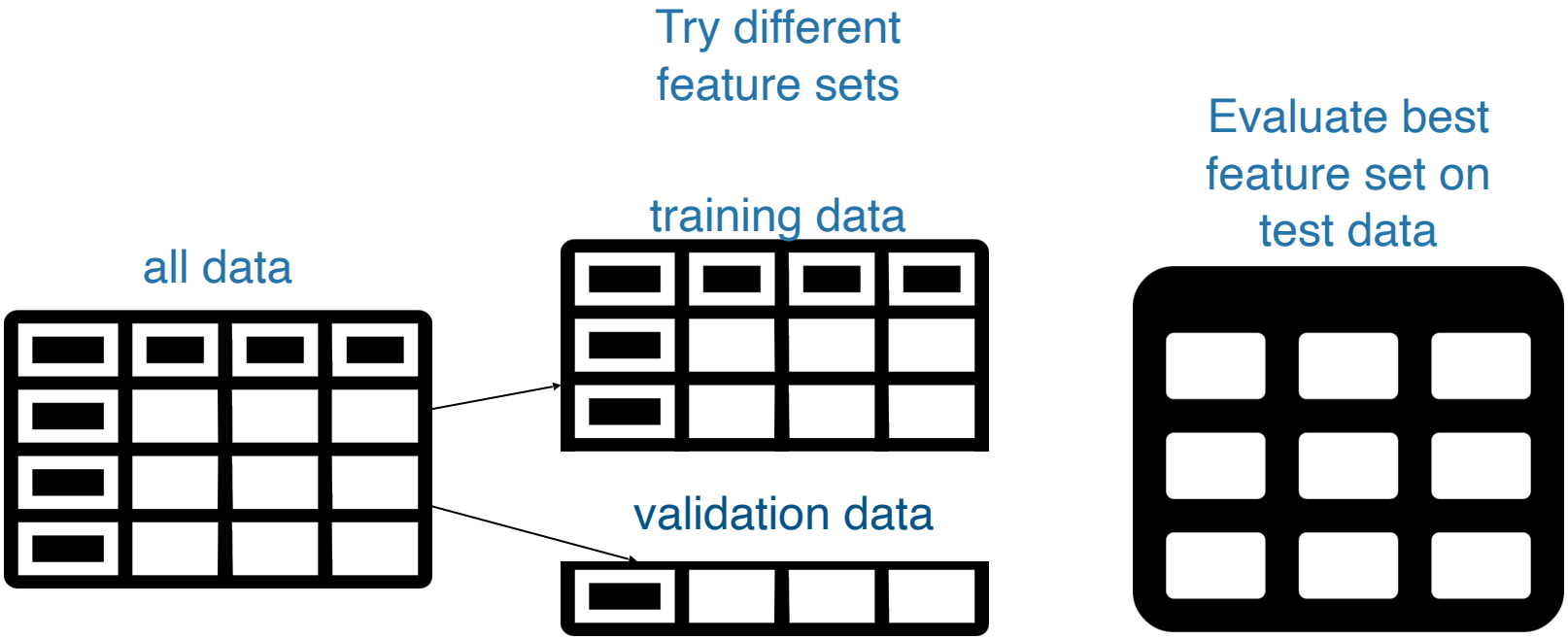

previous elections
voting district demographics
voter information
these data are likely predictive of US election outcomes



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

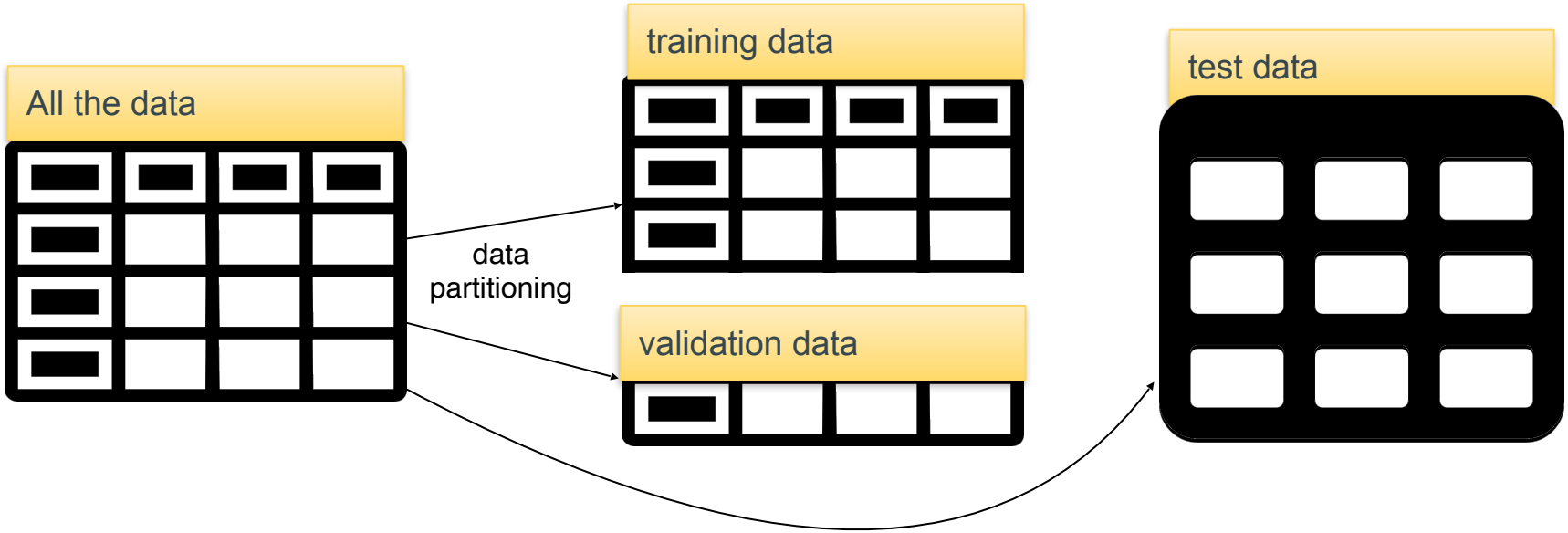


variables that can be used for accurate prediction exploit the relationship between the variables but do NOT mean that one causes the other



Use validation set to select the features!

Try different feature sets



Use validation set to select the features!

ML Algorithms Overview

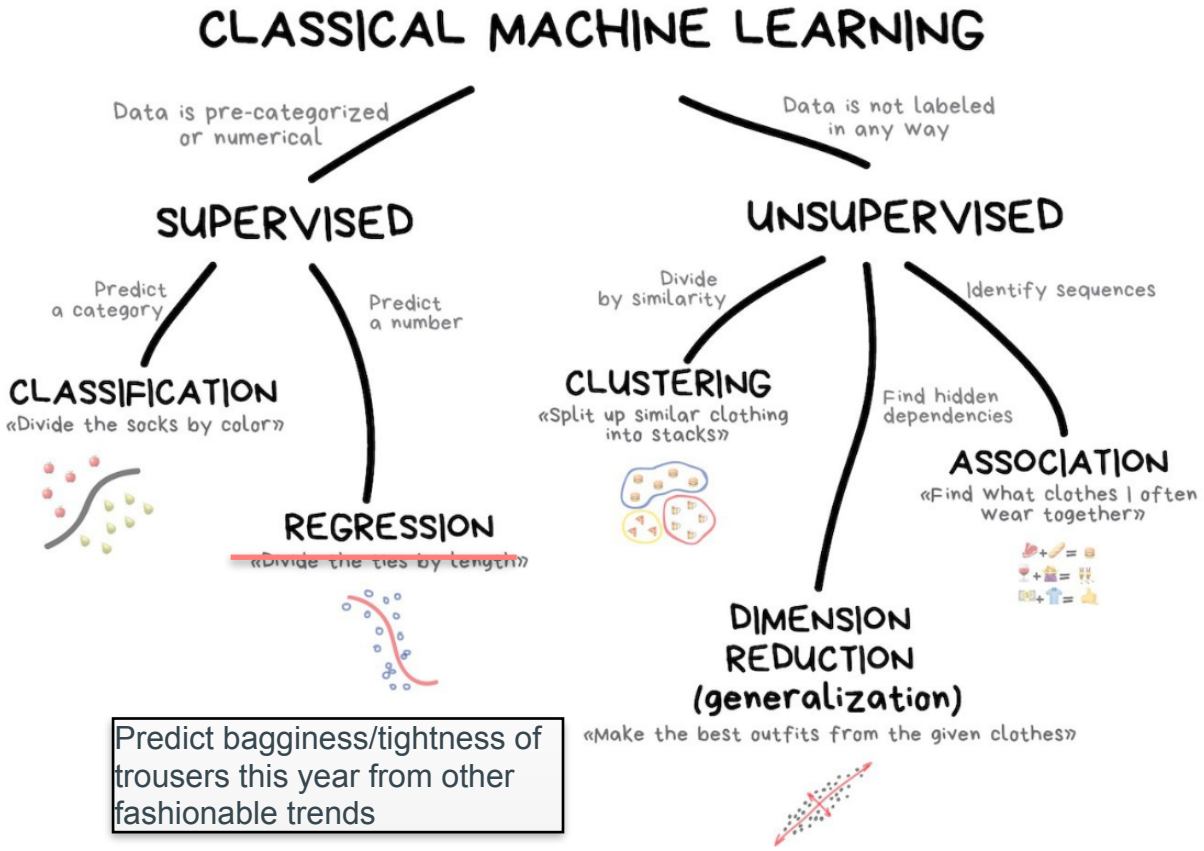
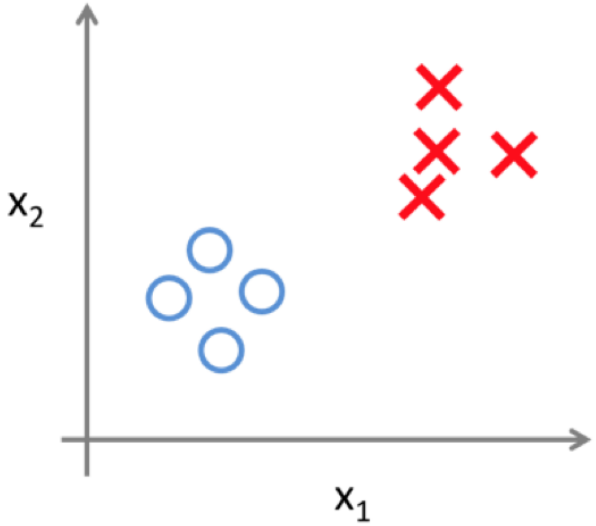


Image source: https://vas3k.com/blog/machine_learning/

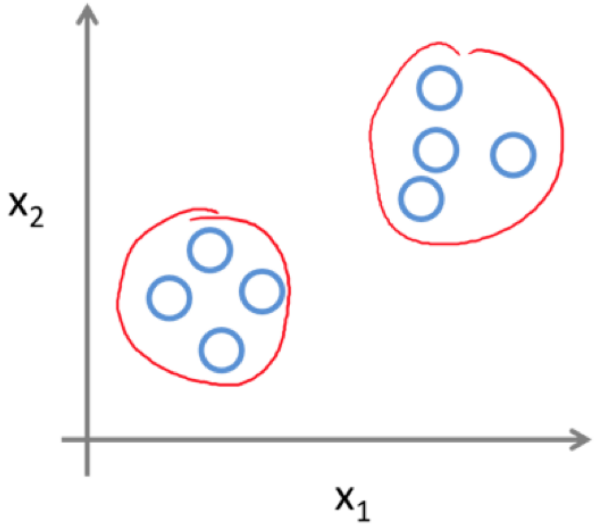
Two modes of machine learning

Supervised Learning



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

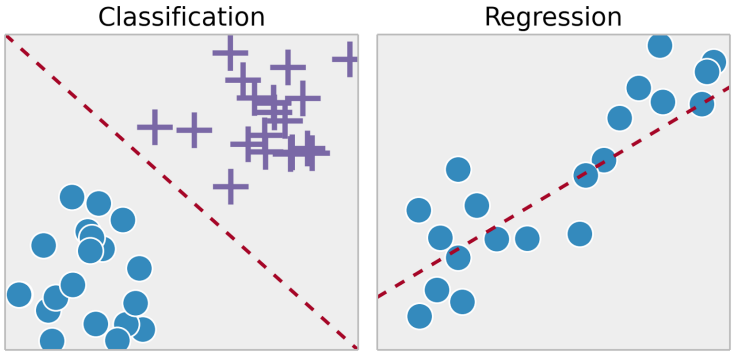
Unsupervised Learning



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

Approaches to machine learning

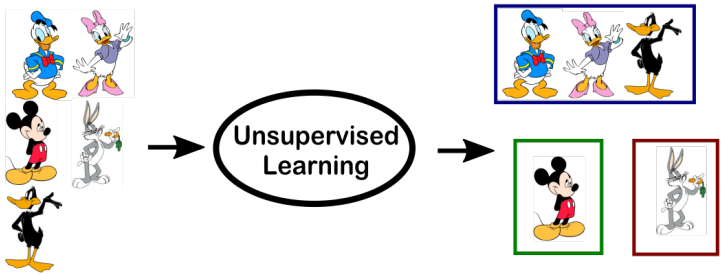
Supervised Learning



categorical variables

continuous variables

Unsupervised Learning



Clustering (categorical)
& dimensionality reduction (continuous)

can automatically identify structure in data

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

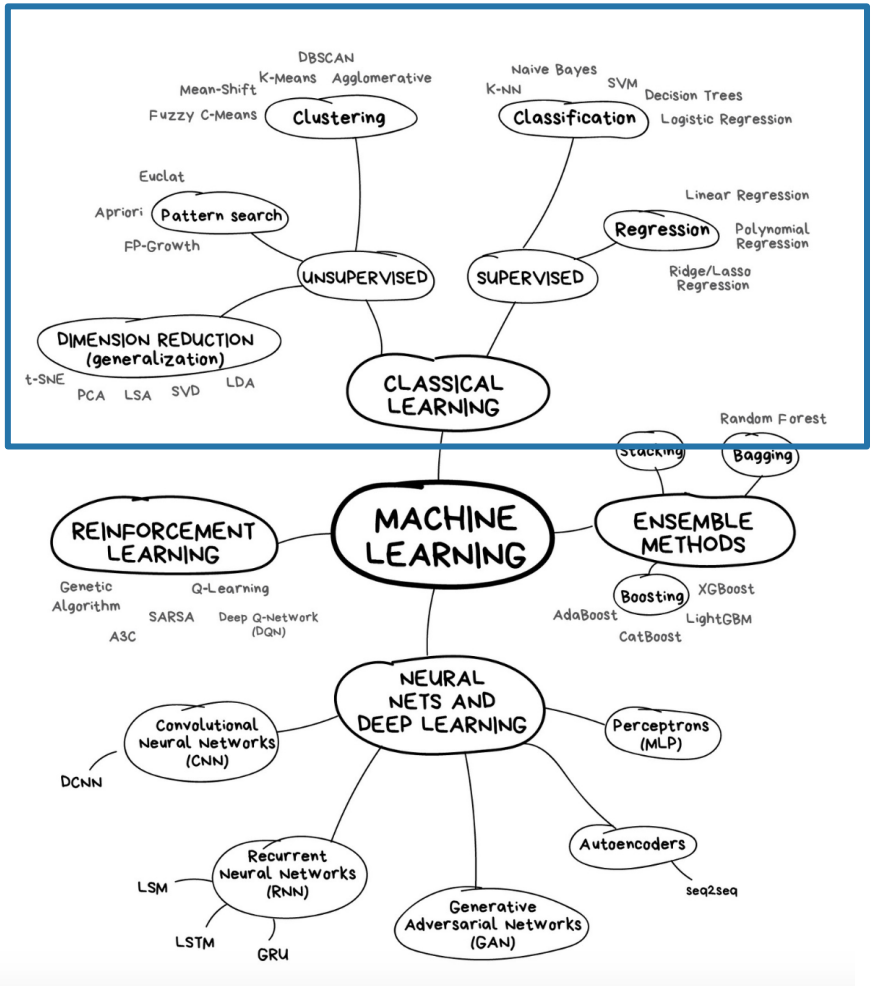
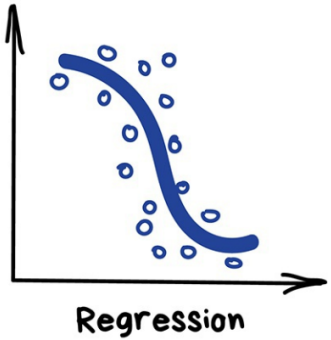


Image source: https://vas3k.com/blog/machine_learning/

Regression

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



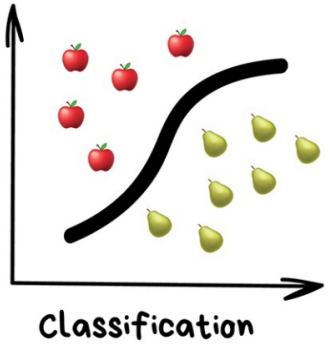
Today this is used for:

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

Popular algorithms are Linear and Polynomial regressions.

Classification

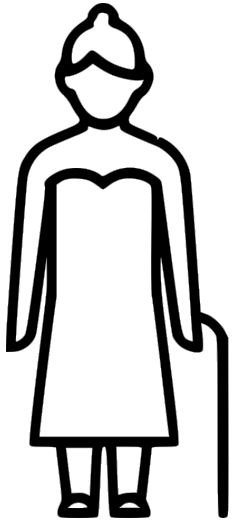
"Splits objects based at one of the attributes known beforehand. Separate socks by based on color, documents based on language, music by genre"



Today used for:
– Spam filtering
– Language detection
– A search of similar documents
– Sentiment analysis
– Recognition of handwritten characters and numbers
– Fraud detection

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

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



Regression:

predicting continuous variables
(i.e. Age)

continuous variable prediction

Classification:

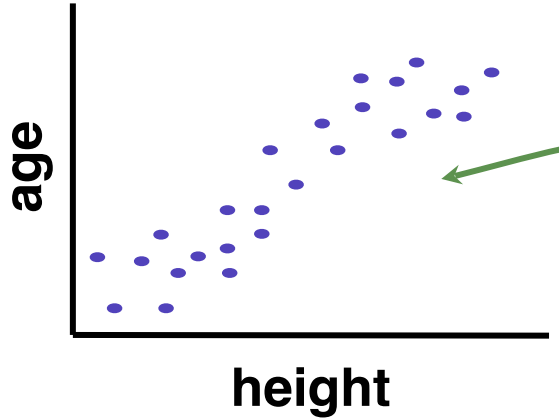
predicting categorical variables
(i.e. education level)

categorical variable prediction

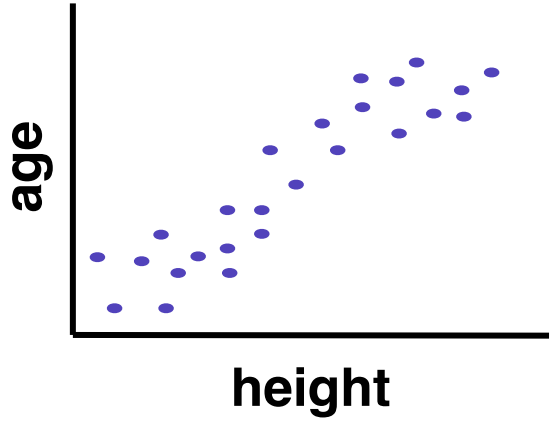
Supervised Learning

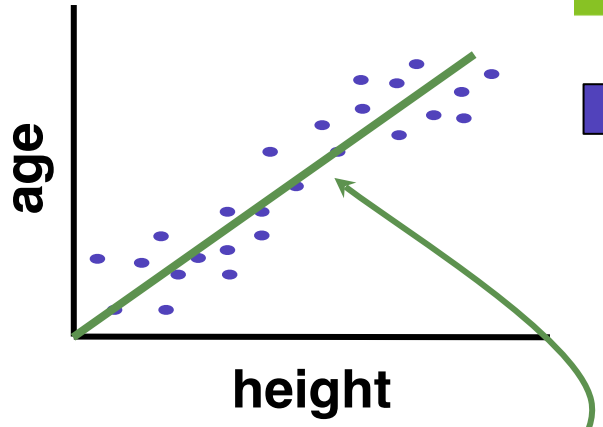
regression

continuous variable prediction



the training data will be used to build the predictive model



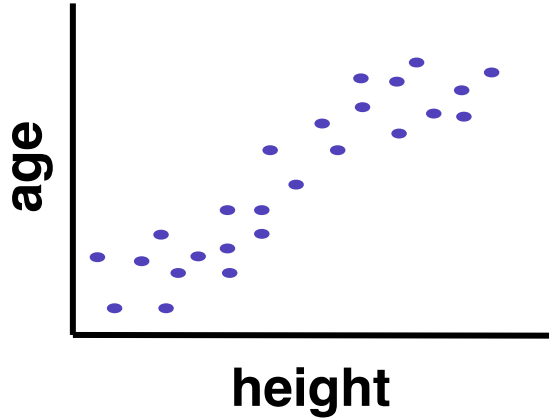


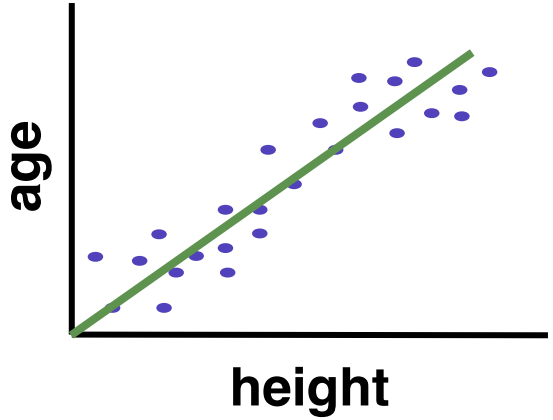
Supervised Learning

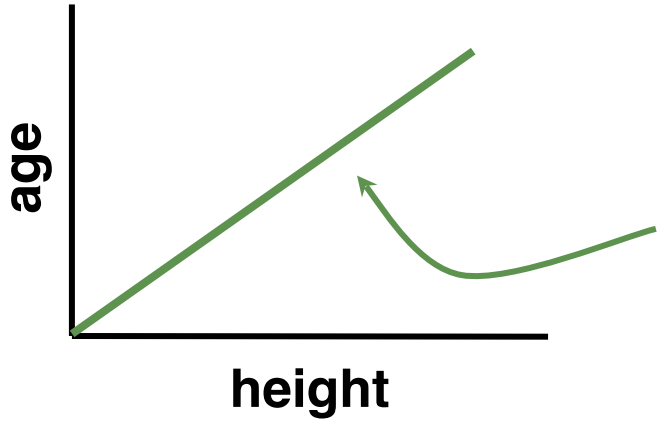
regression

continuous variable prediction

use linear regression to model the relationship





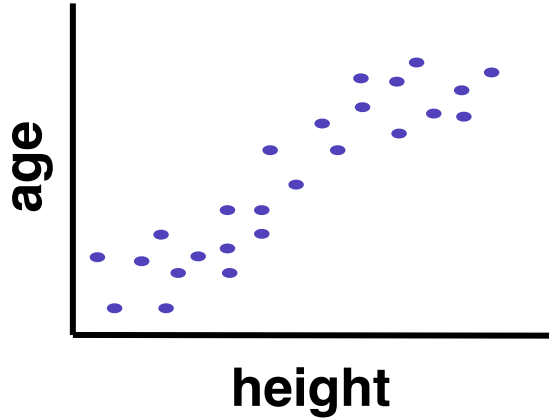


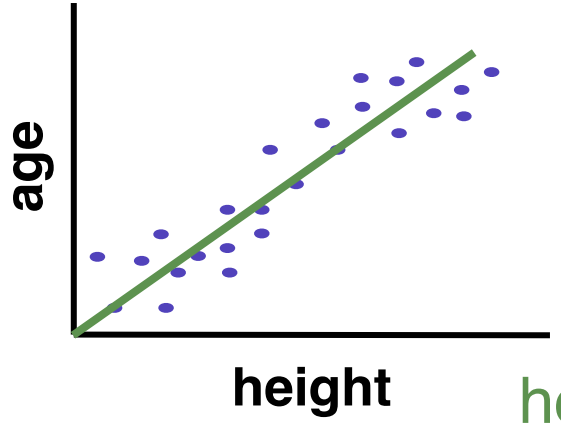
For prediction, the individual values in the training data are *not* important. We only need the model.

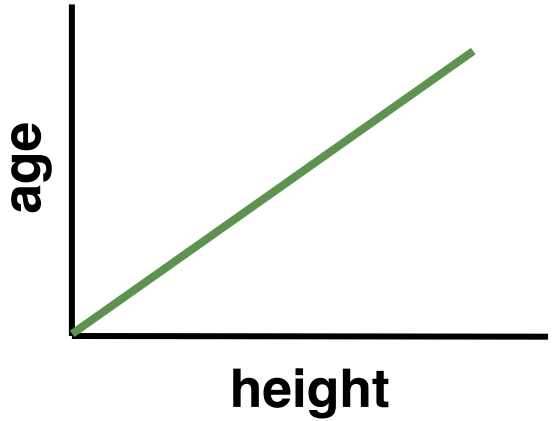
Supervised Learning

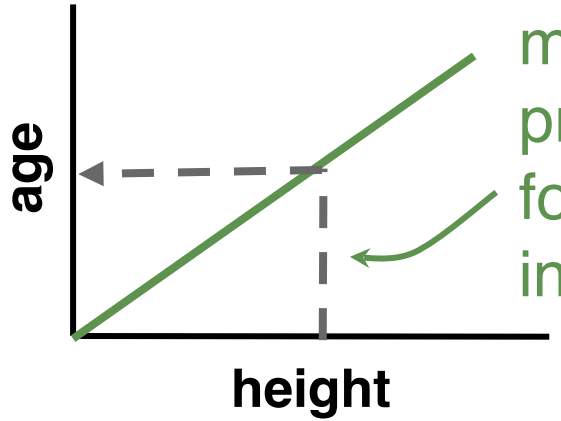
regression

continuous variable prediction







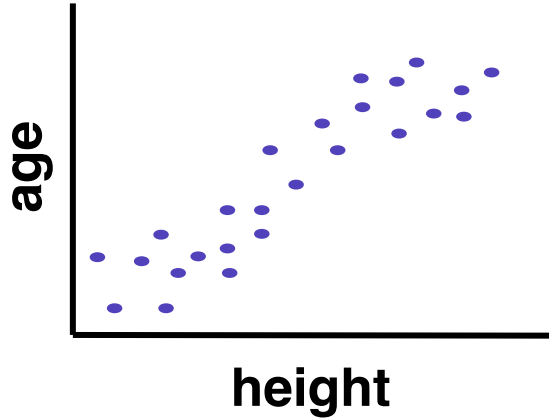


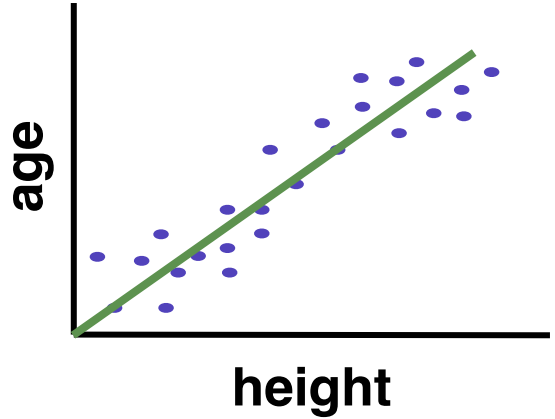
how we'll make predictions for a future individual

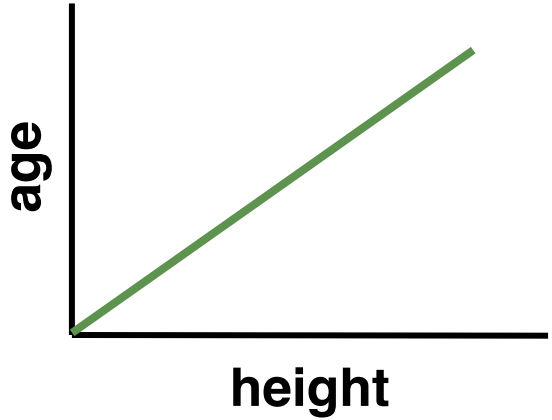
Supervised Learning

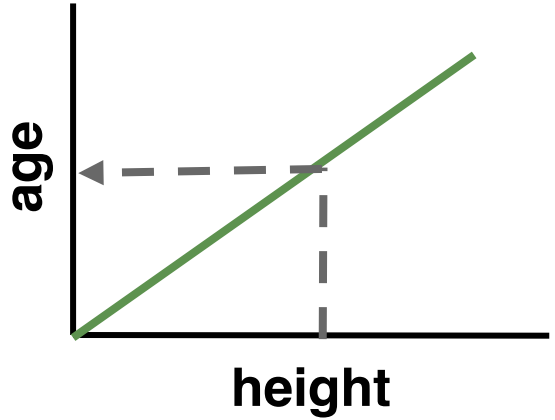
regression

continuous variable prediction

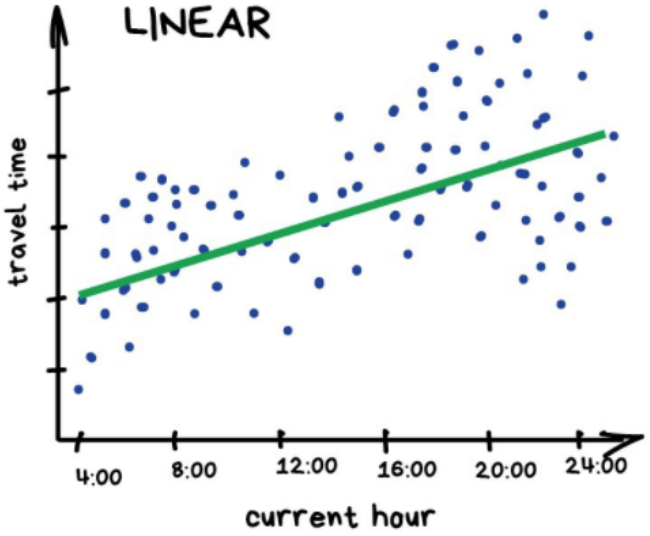


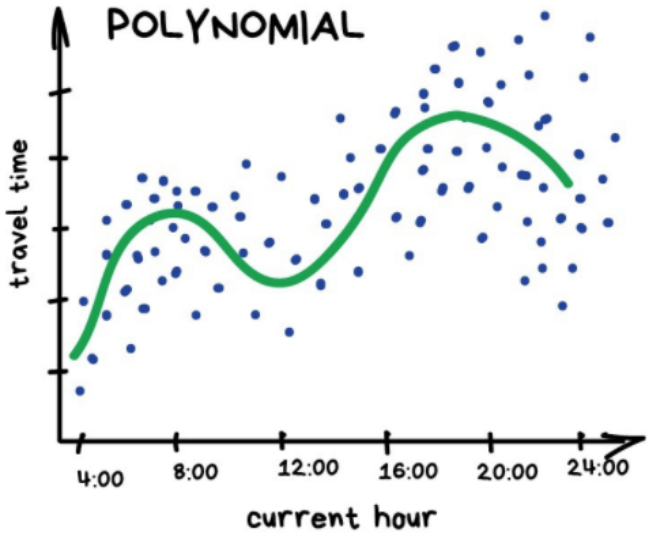






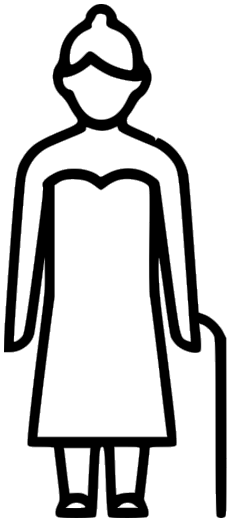
PREDICT TRAFFIC JAMS





REGRESSION

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



Regression:

predicting continuous variables

(i.e. Age)

Classification:

predicting categorical variables

(i.e. give a loan?)

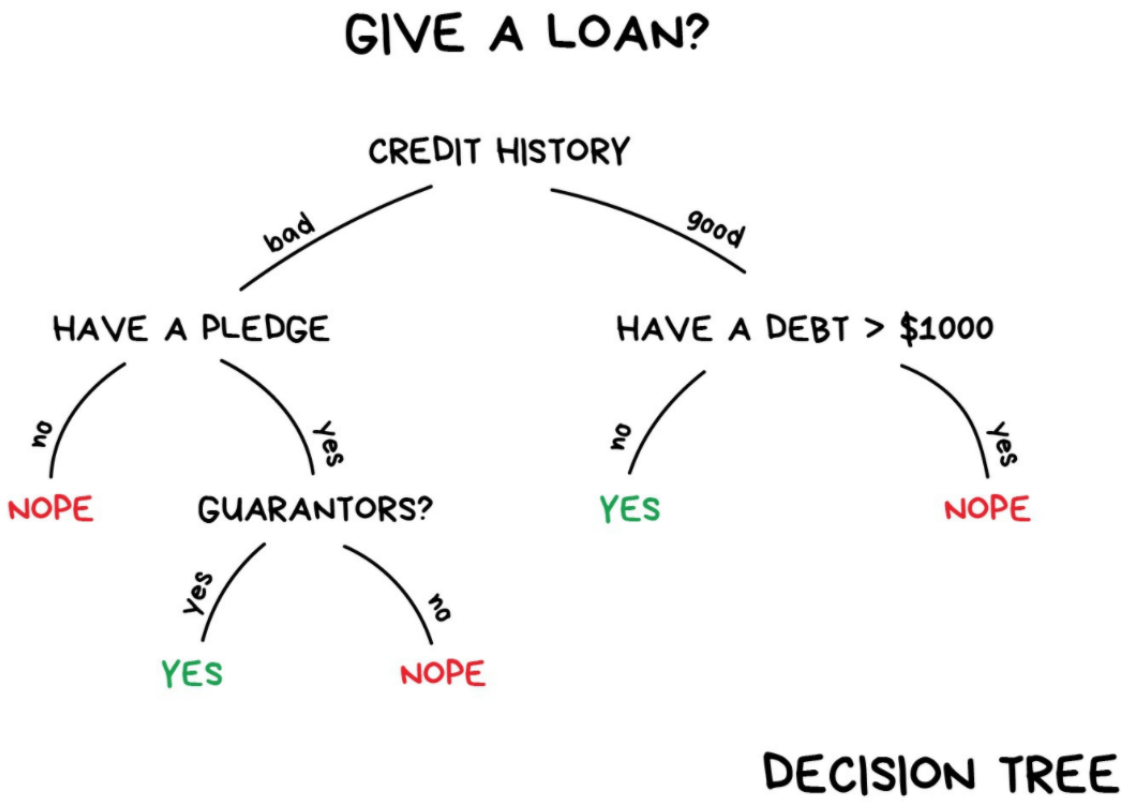
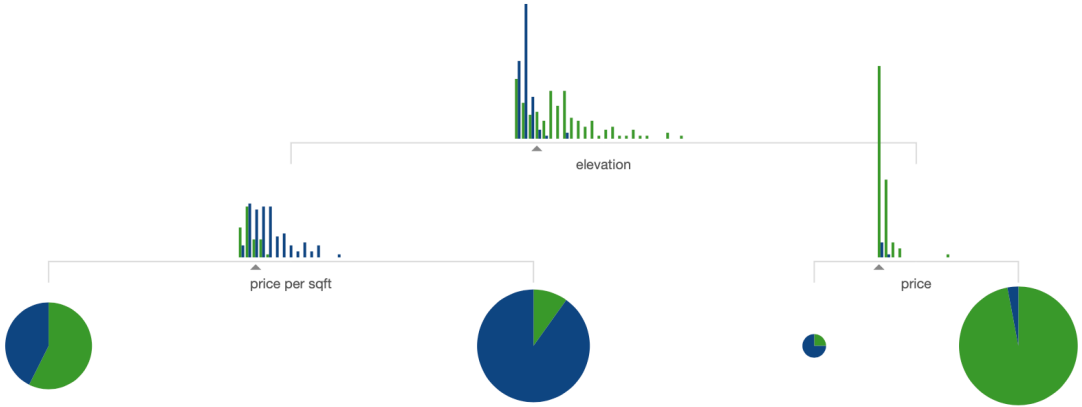


Image source: https://vas3k.com/blog/machine_learning/

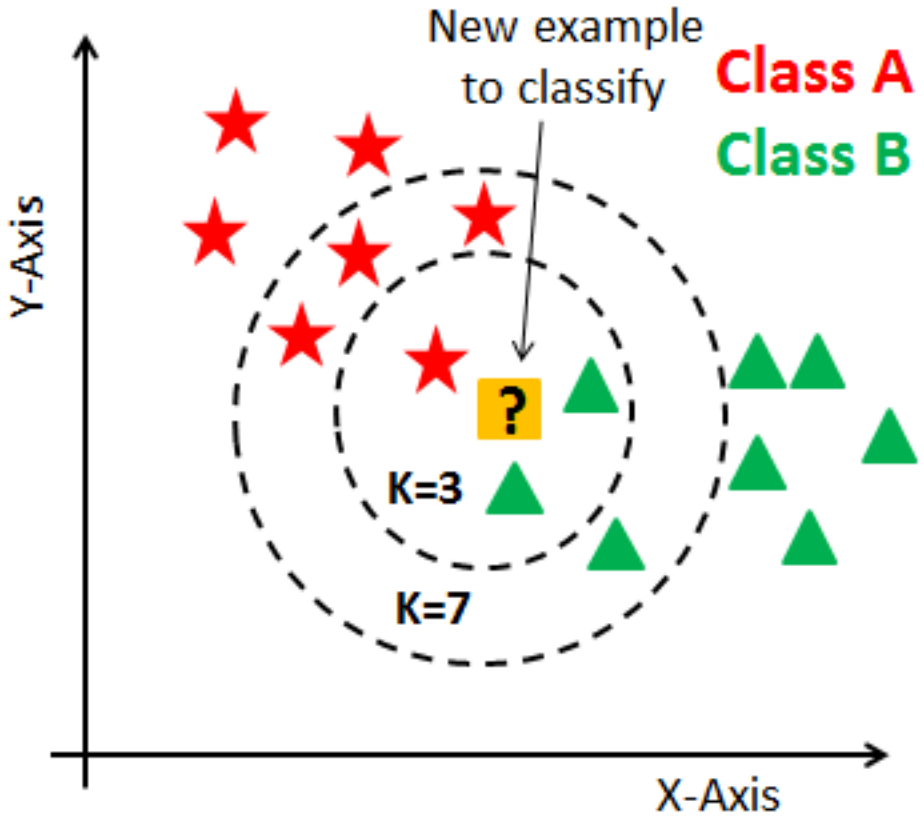
Growing a tree

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

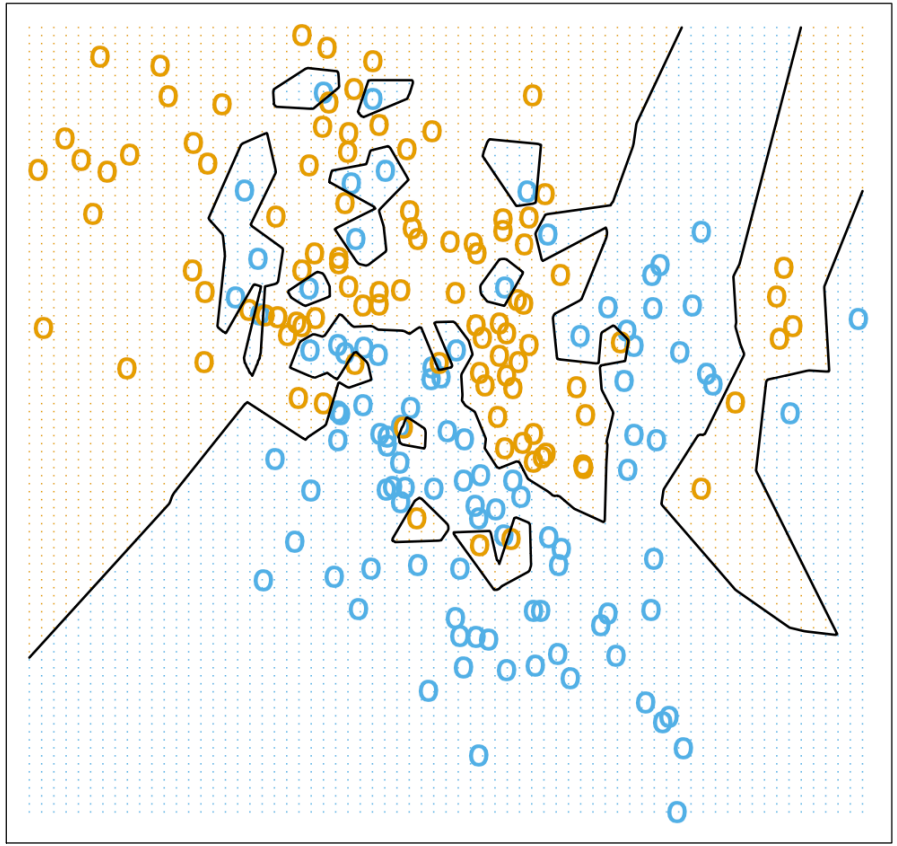


http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

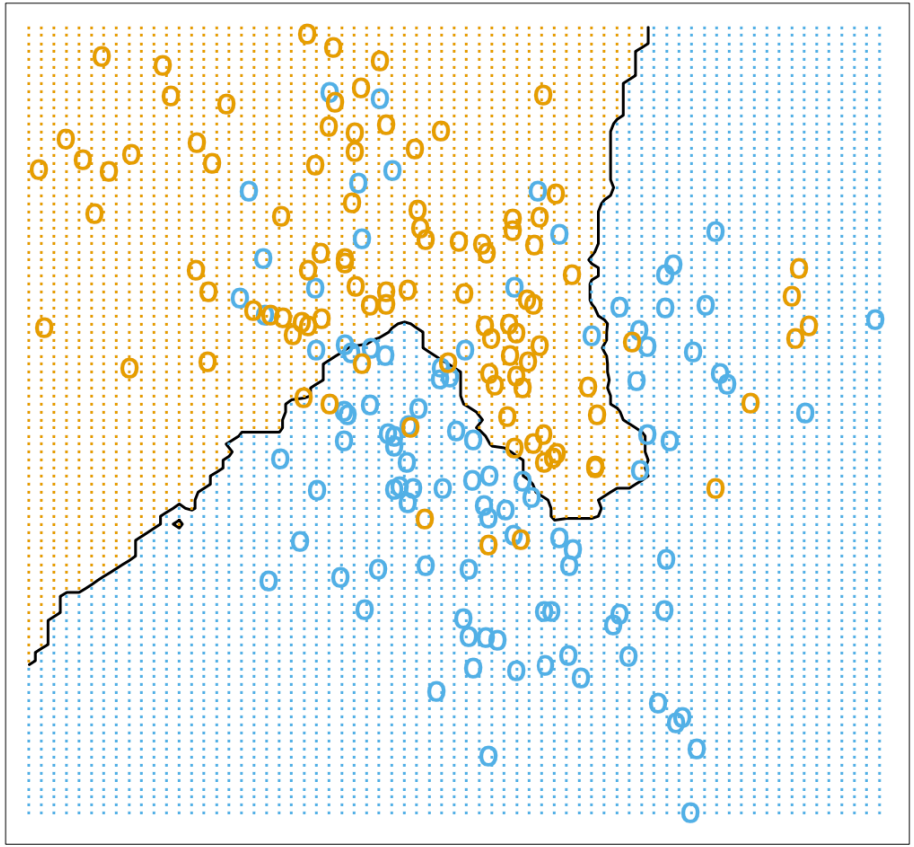
K-nearest neighbors

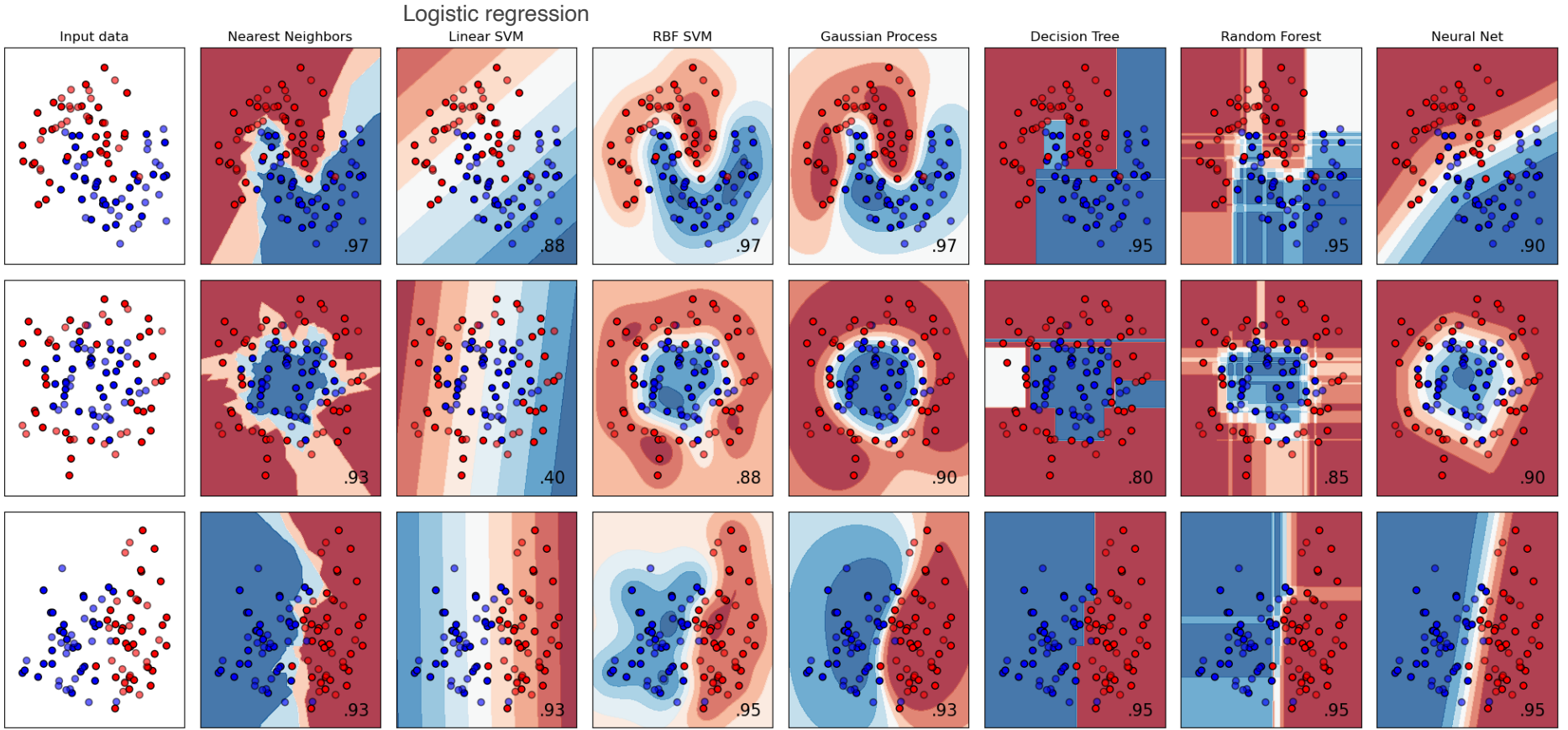


1–Nearest Neighbor Classifier



15-Nearest Neighbor Classifier



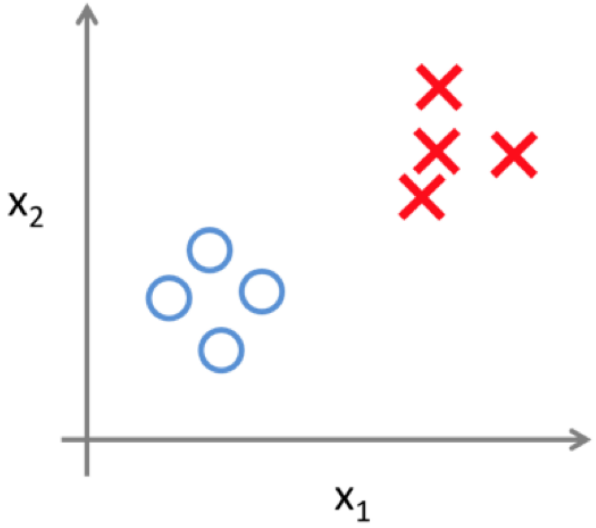


https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

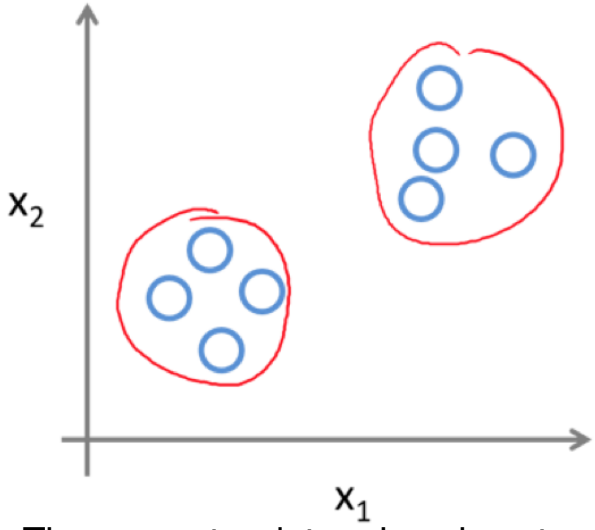
Unsupervised Learning

To modes of machine learning

Supervised Learning



Unsupervised Learning



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

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

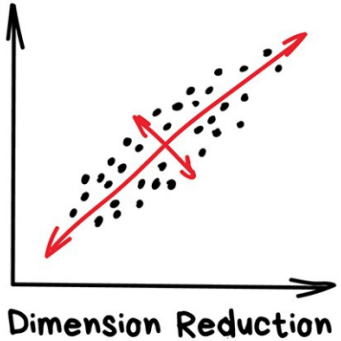
Popular algorithms: K-means_clustering, Mean-Shift, DBSCAN

Dimensionality Reduction (Generalization)

"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



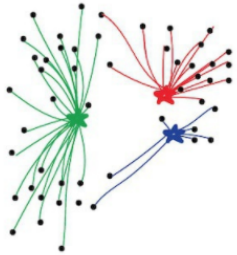
Popular algorithms: Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA, pLSA, GLSA), t-SNE (for visualization)

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

PUT KEBAB KIOSKS IN THE OPTIMAL WAY

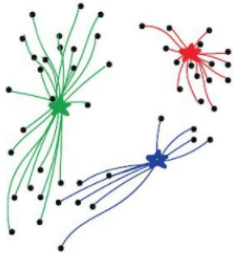
(also illustrating the K-means method)

1. Put kebab kiosks in random places in city



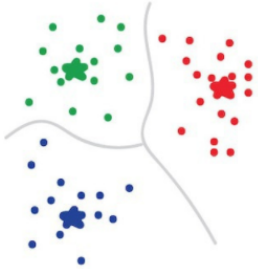
2. Watch how buyers choose the nearest one

3. Move kiosks closer to the centers of their popularity



4. Watch and move again

5. Repeat a million times



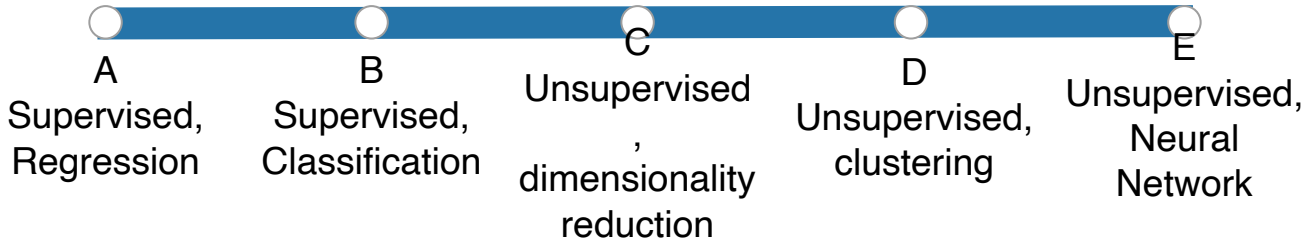
6. Done!
You're god of kebabs!

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

Prediction Approach

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

How would you approach this with machine learning?



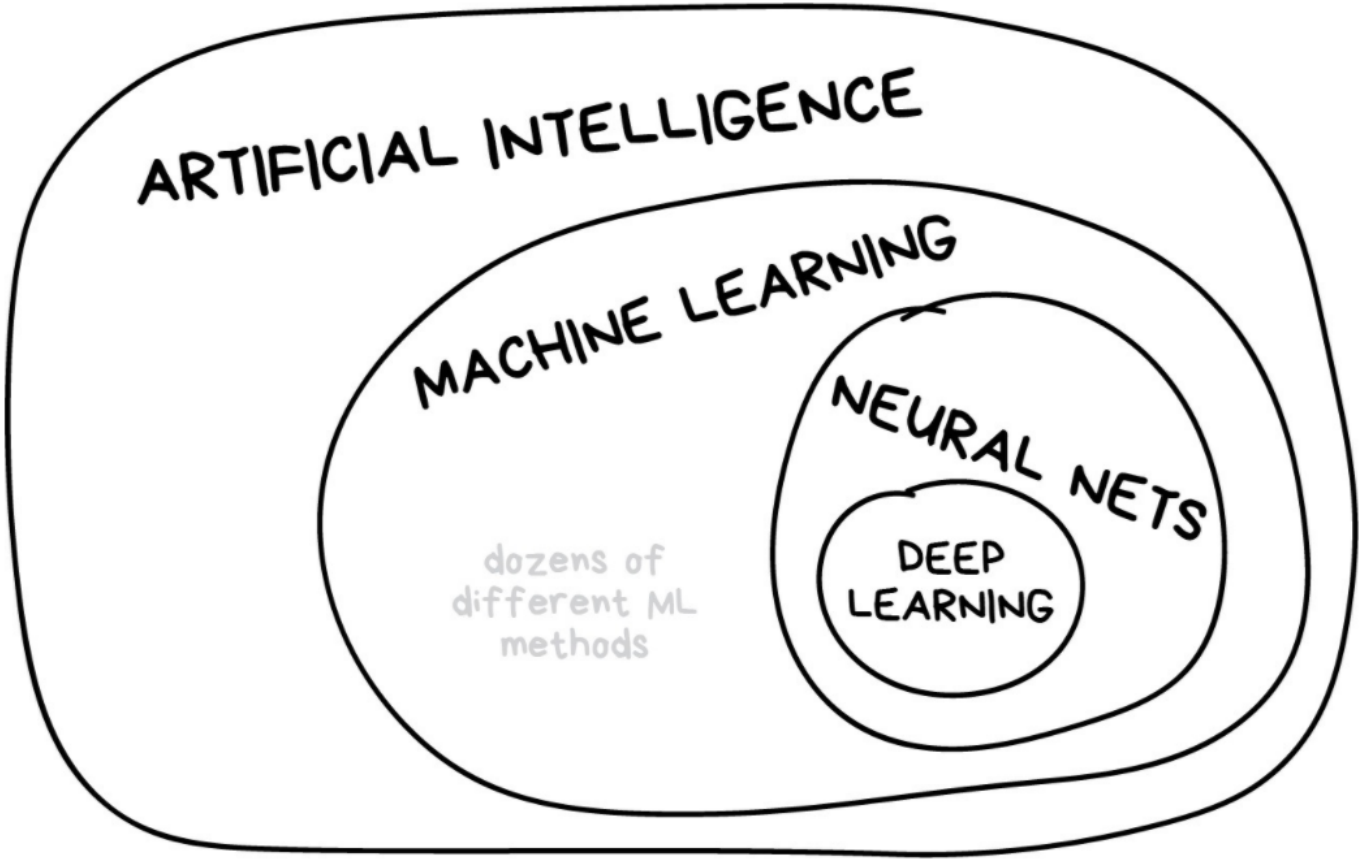


Image source: https://vas3k.com/blog/machine_learning/

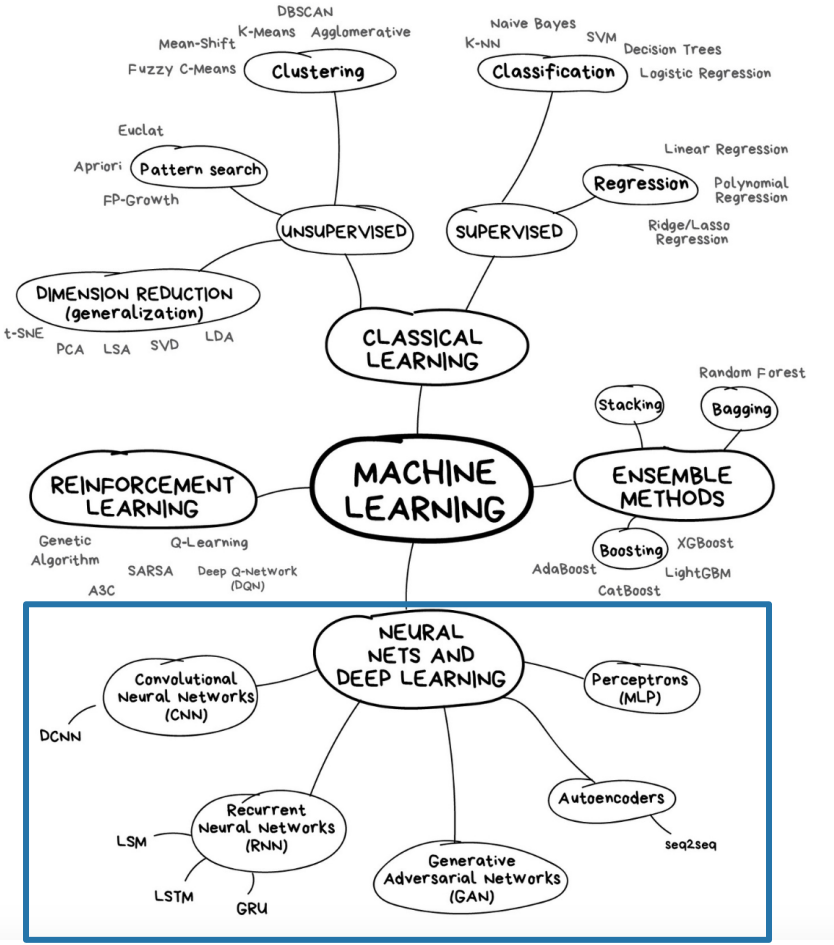
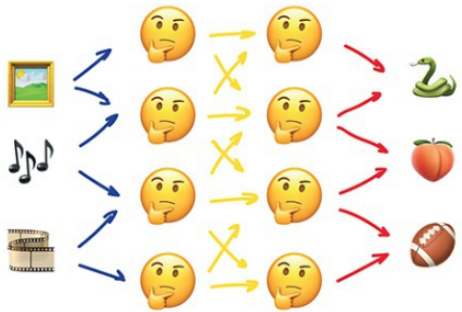


Image source: https://vas3k.com/blog/machine_learning/

"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:

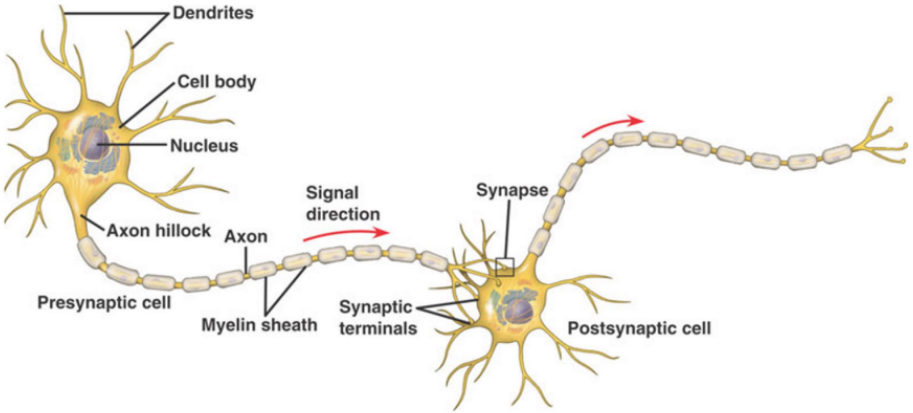
- Replacement of all algorithms above
- Object identification on photos and videos
- Speech recognition and synthesis
- Image processing, style transfer
- Machine translation



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

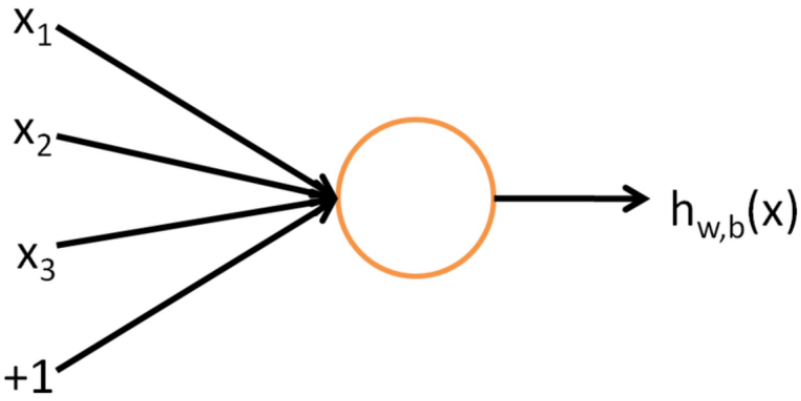
Image source: https://vas3k.com/blog/machine_learning/

WHAT IS A NEURON?



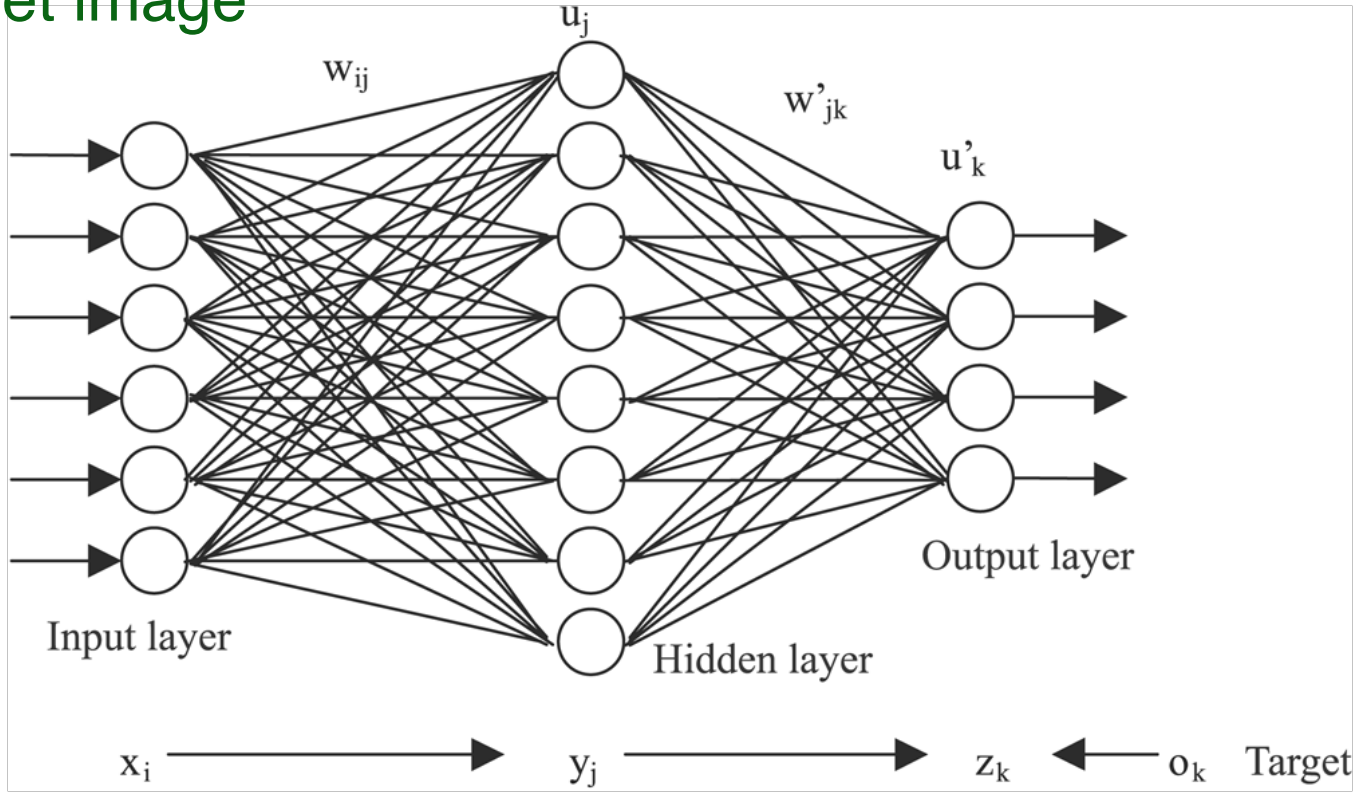
- Receives signal on synapse
- When trigger sends signal on axon

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

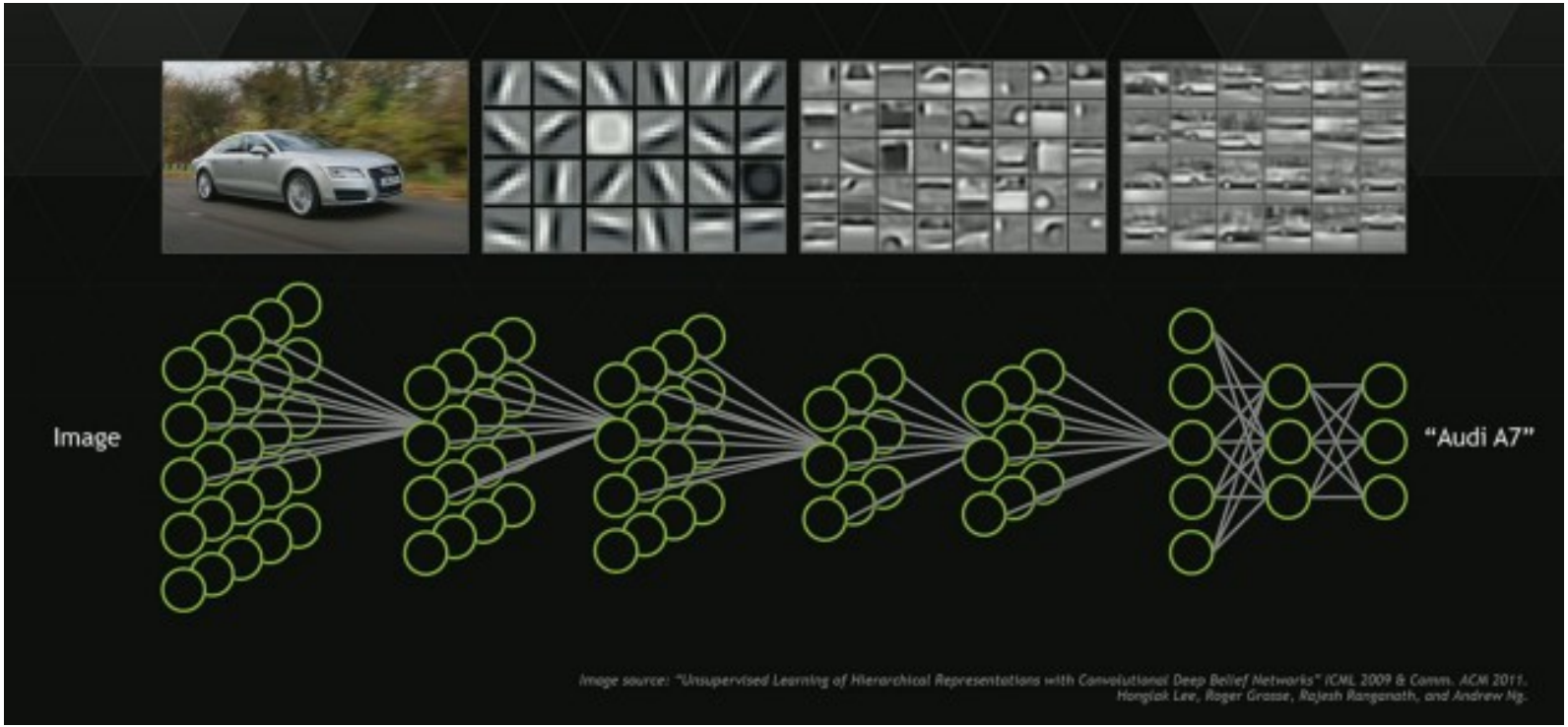


Image Source: https://towardsdatascience.com/understanding-residual-networks-9add4b664b03

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.

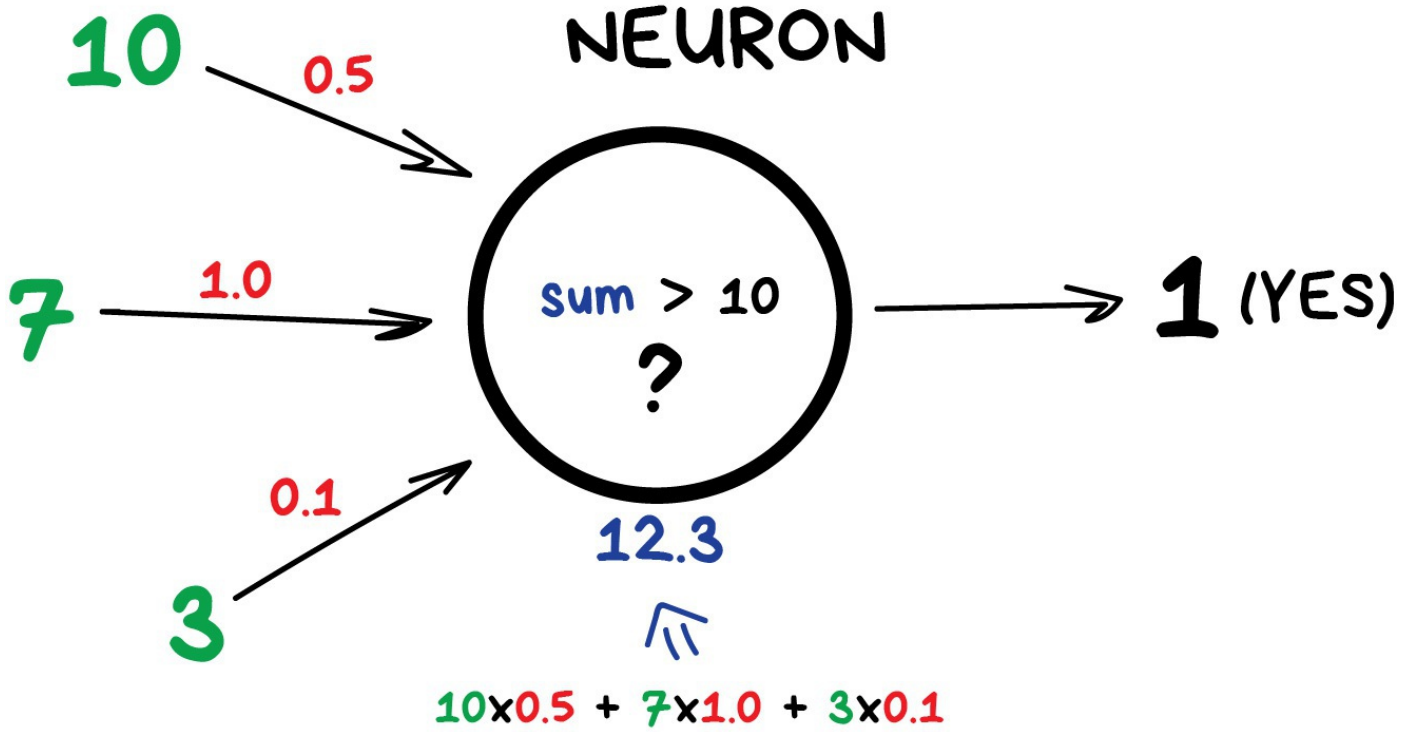


Image source: https://vas3k.com/blog/machine_learning/

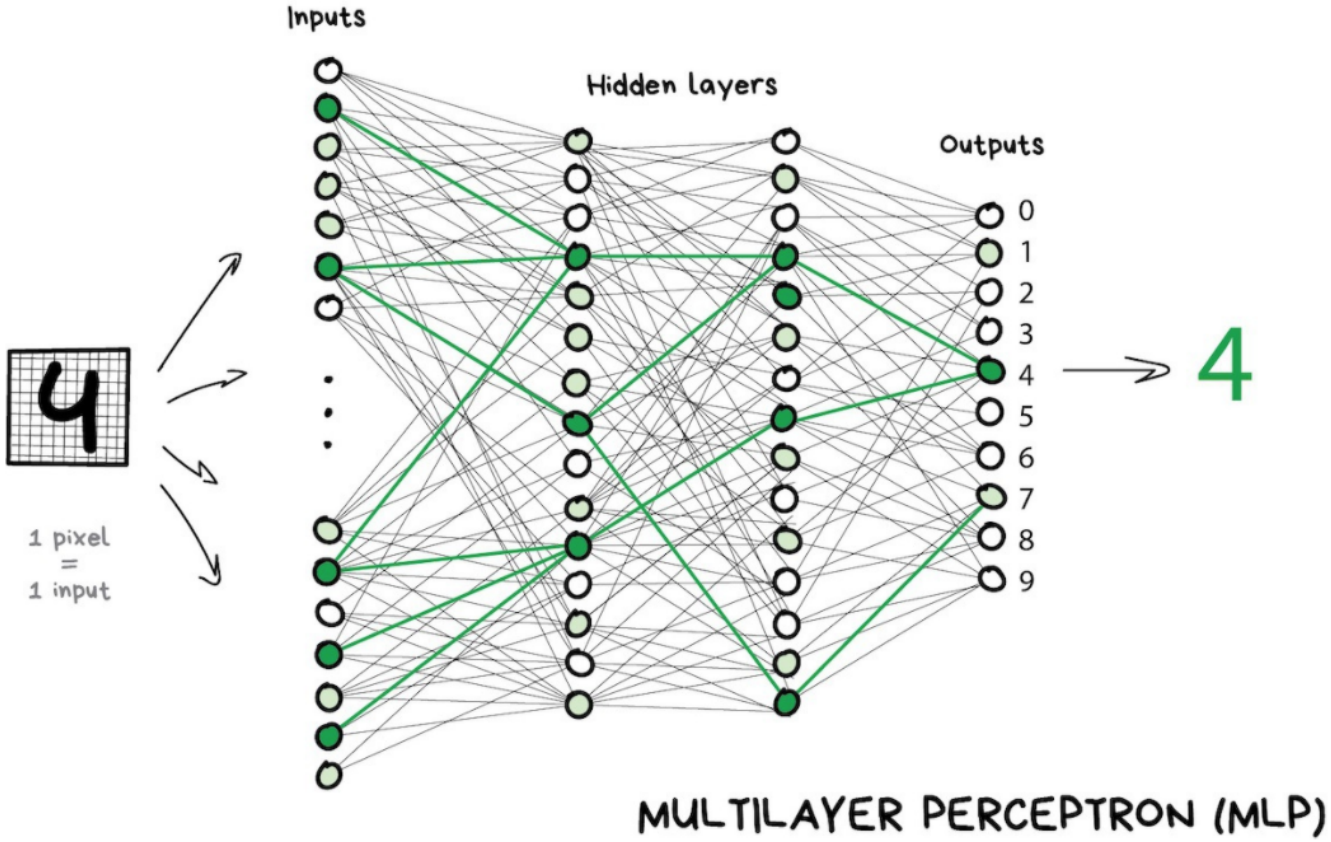


Image source: https://vas3k.com/blog/machine_learning/

Manually labeling used to be the way...

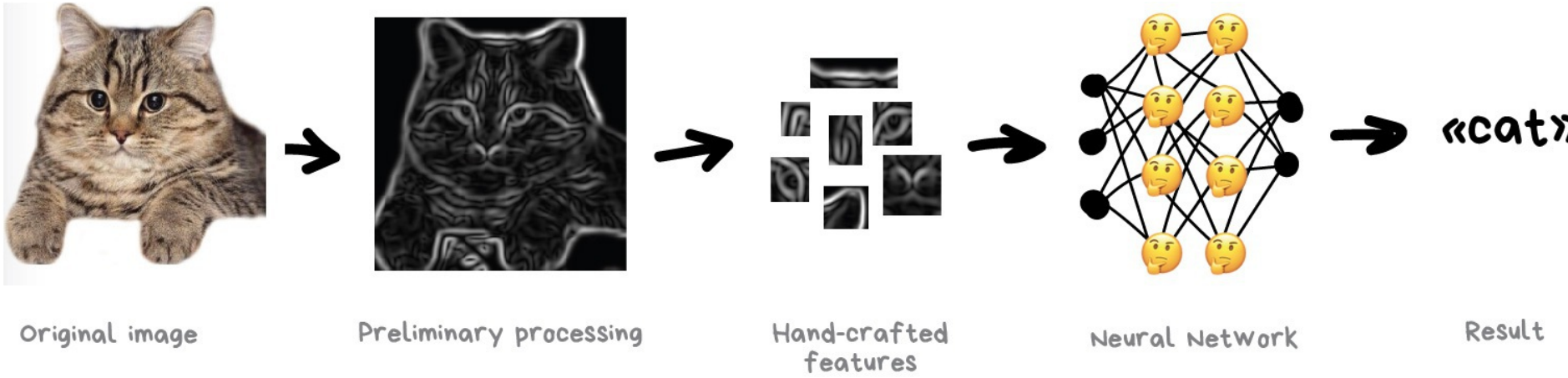
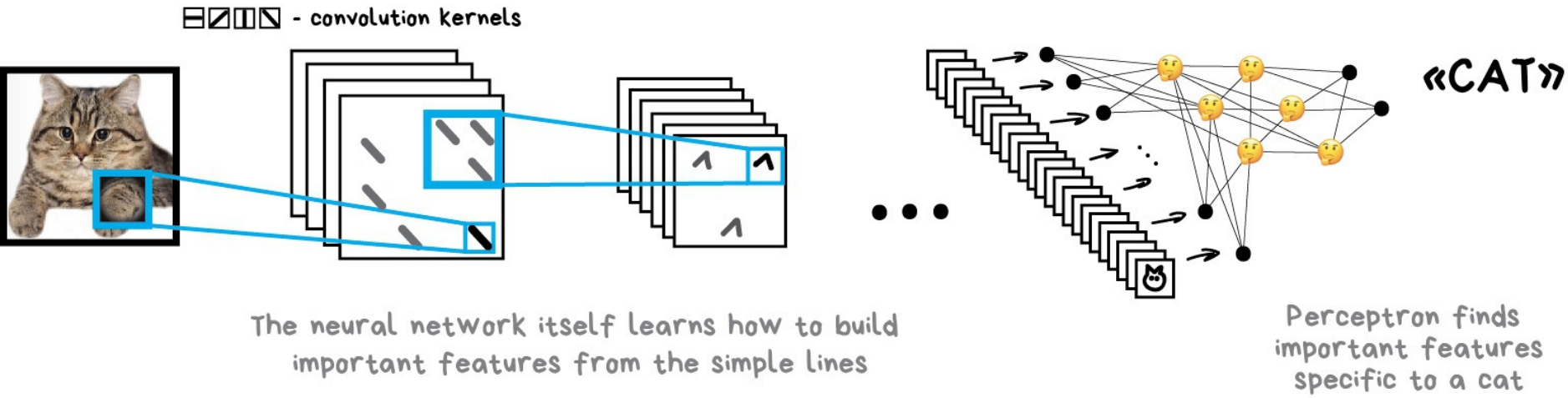


Image source: https://vas3k.com/blog/machine_learning/

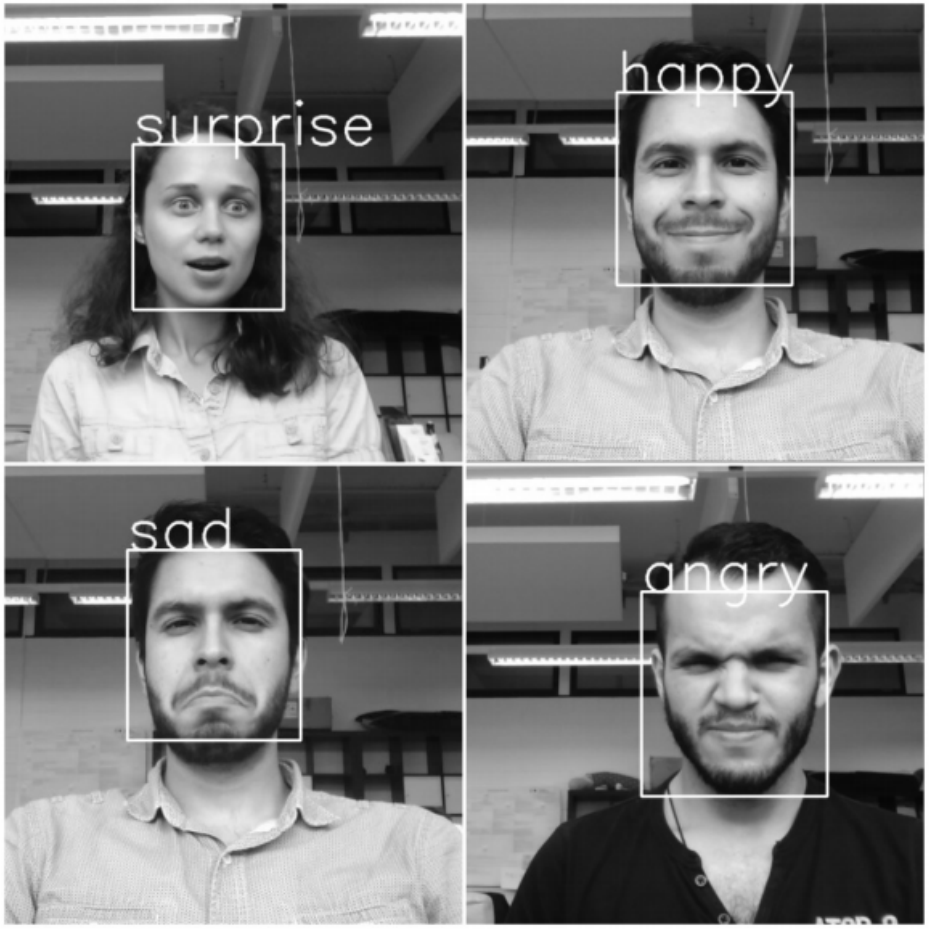
CNNs avoid manual features



“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)

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

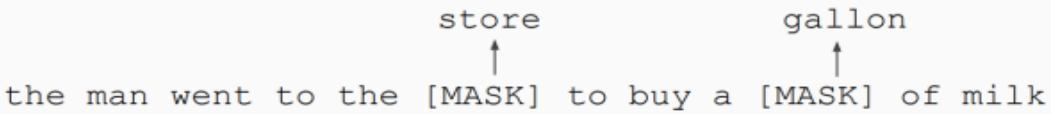


https://github.com/oarriaga/face_classification

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%



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

https://neptune.ai/blog/unmasking-bert-transformer-model-performance

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)

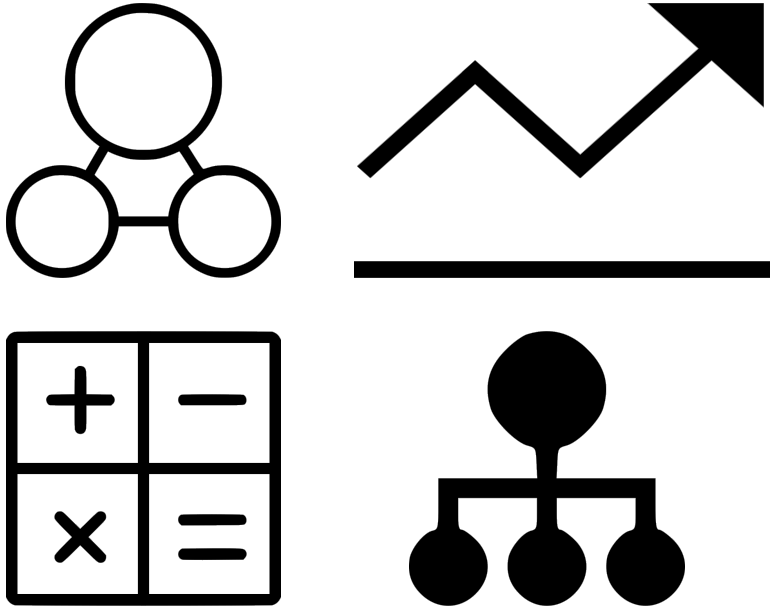
candy theme: normal
messages: BESTIE, CUTIE PIE, SOUL MATE, SWEET PEA, UR CUTE, YOU + ME, BE MINE, PICK ME, KISS ME, LOVE BIRDS, MARRY ME, OOO LA LA, TRUE LOVE, WINK WINK, XOXO
candy theme: programming
messages:

(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

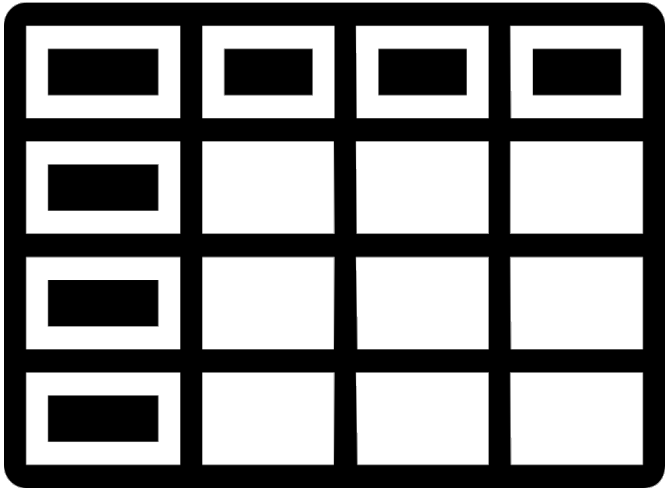
ML jargon

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

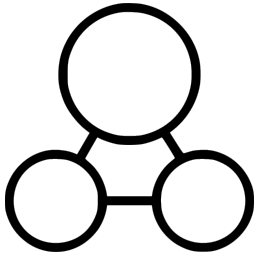
- Algorithm:
    - some math to optimally solve a problem given some training data
- Solver:
    - The software implementation/approximation of the math; many different ways to get the job done
- Hyperparameters:
    - Knobs/Dials that affect how the algorithm solves the problem, the solver is one such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



model selection

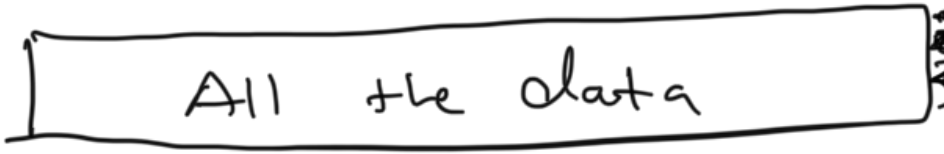


BIGGER
datasets



SIMPLER
models

1

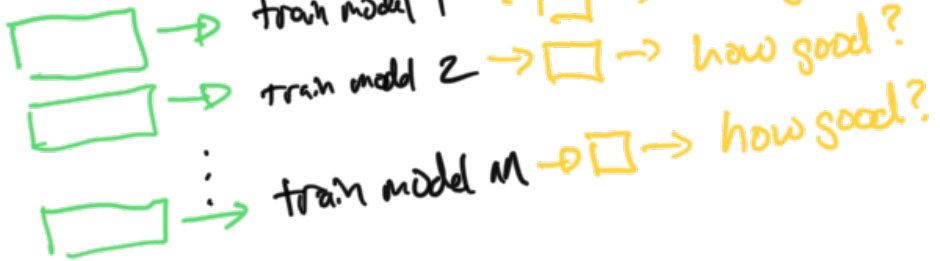


randomly shuffle rows of dataset

2

Split into training, validation and test sets

3



- if you have M models to try do this once for each

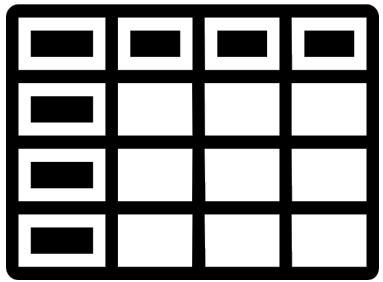
4

Pick best model from validation set performance

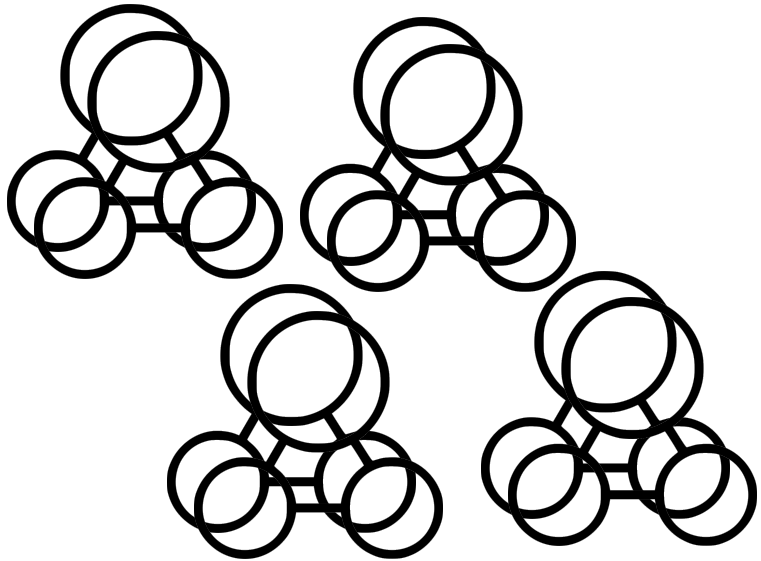
train it on T+V then test it on Test set

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. Test

split off a test set

3. fold 1, fold 2, fold 3, ..., fold K

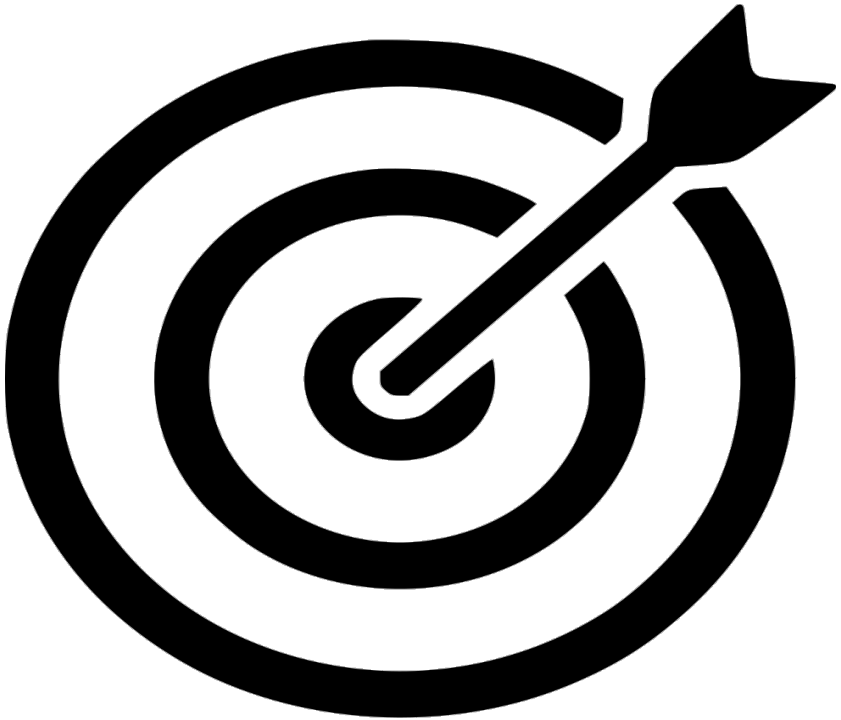
split remaining data into K folds with even(ish) amounts of data w/ same(ish) distributions

4. split 1 → train / validate; split 2 → train / valid; split 3 → train; ... split K; etc ...

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 → trained best model → Estimated performance in future



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

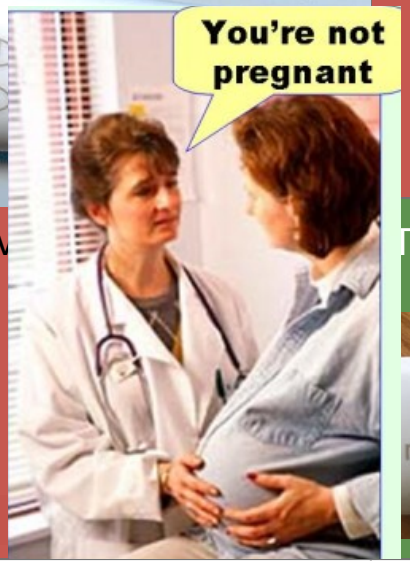
categorical variable prediction

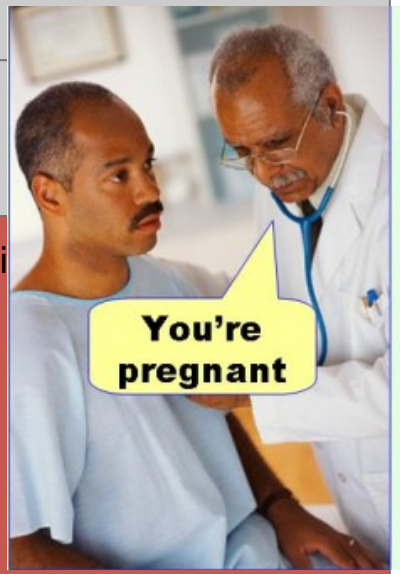
$$\text{Accuracy} = \frac{\#\text{ of samples predicted correctly}}{\#\text{ of samples predicted}} * 100$$

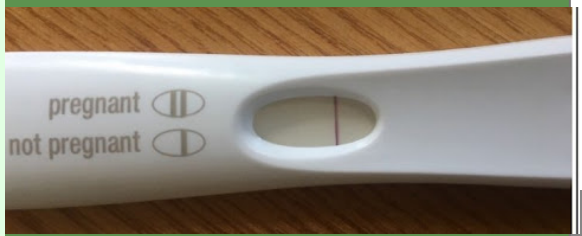
Accuracy can mislead

- 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 Positi
	Negative	False Negativ	True Negative (TN)







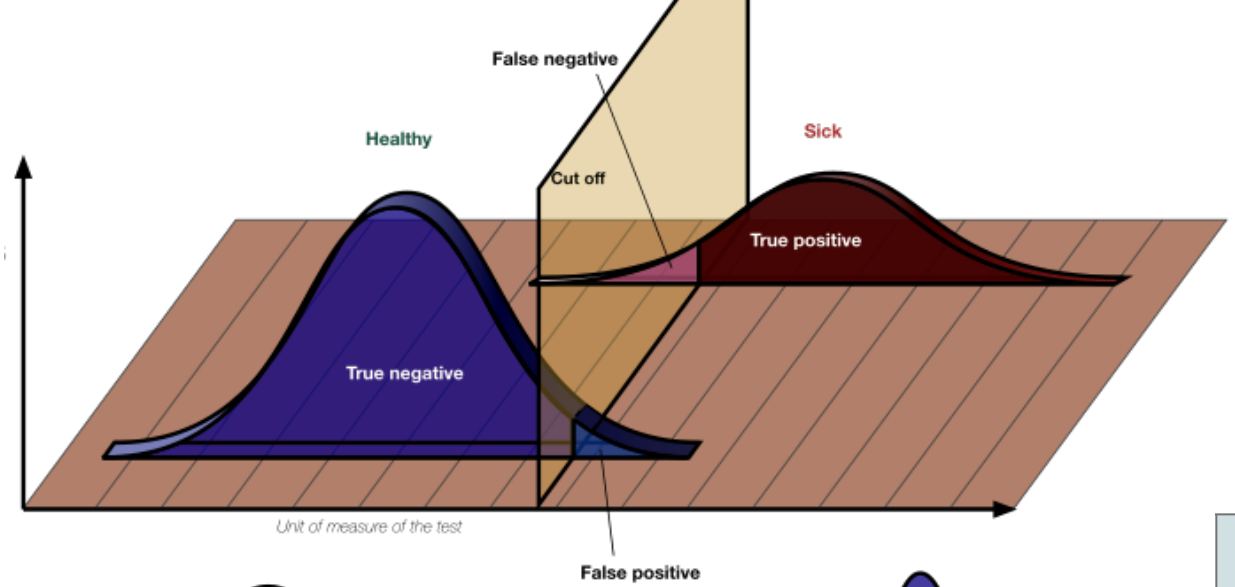
categorical variable prediction

confusion matrix

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

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?

categorical variable prediction



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

PPV =

NPV =

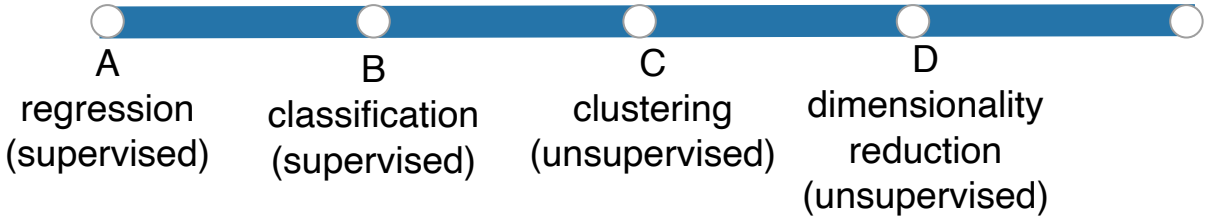
Sensitivity =

Specificity =

Prediction Approach

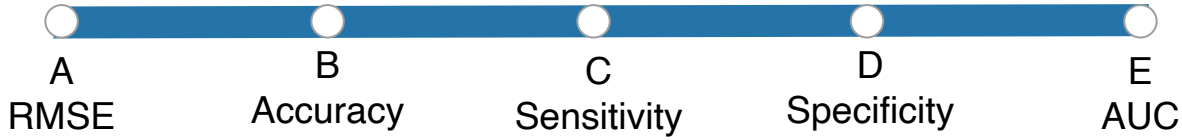
You've been given a dataset with a number of features and have been asked to predict each individual's age.

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?