Course Announcements

Major screwups on my part:

- Q6 is now due on Wed of this week not tonight
- Weekly Project Survey (*optional*)

Due Friday (11:59 PM):

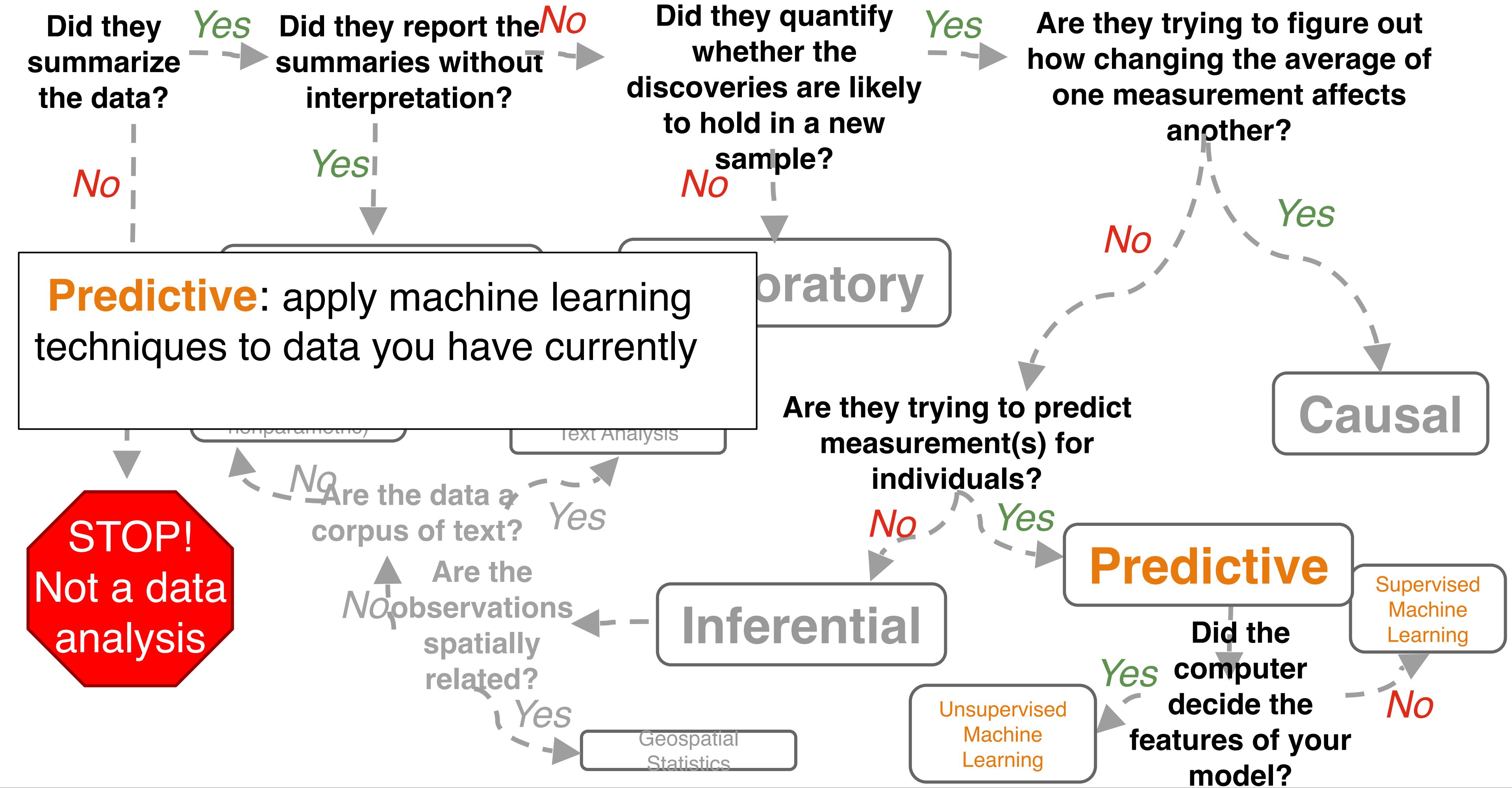
- D6
- A3

Machine learning

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

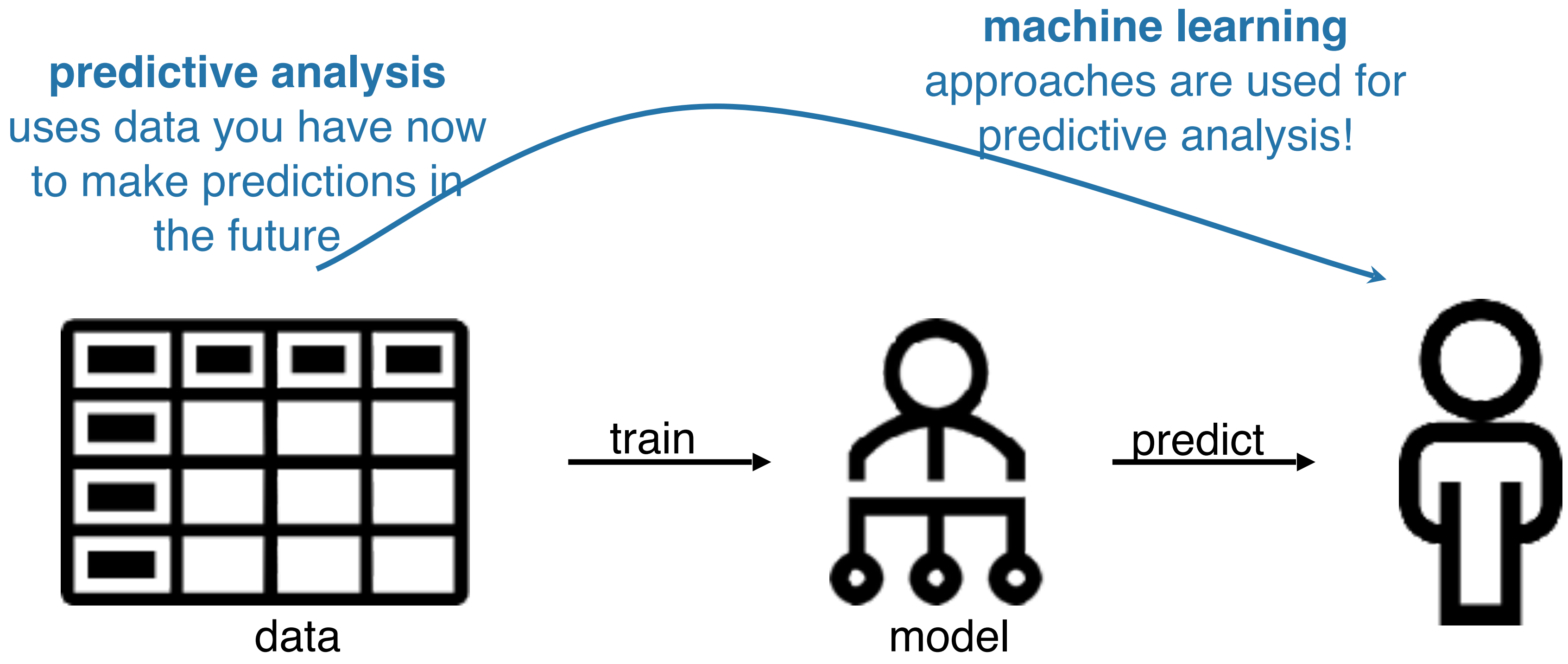
@jasongfleischer

https://jgfleischer.com



Did they summarize the data?
Yes
Did they report the summaries without interpretation?
No
Did they quantify whether the discoveries are likely to hold in a new sample?
Yes
Are they trying to figure out how changing the average of one measurement affects another?
No
Yes
No
oratory
No
Yes
Predictive: apply machine learning techniques to data you have currently
Causal
Are they trying to predict measurement(s) for individuals?
Text Analysis
No
Are the data a corpus of text?
Yes
No
Yes
STOP! Not a data analysis
Predictive
Supervised Machine Learning
Are the observations spatially related?
No
Inferential
Did the computer decide the features of your model?
Yes
No
Yes
Geospatial Statistics
Unsupervised Machine Learning

- **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!
train
predict
data
model

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

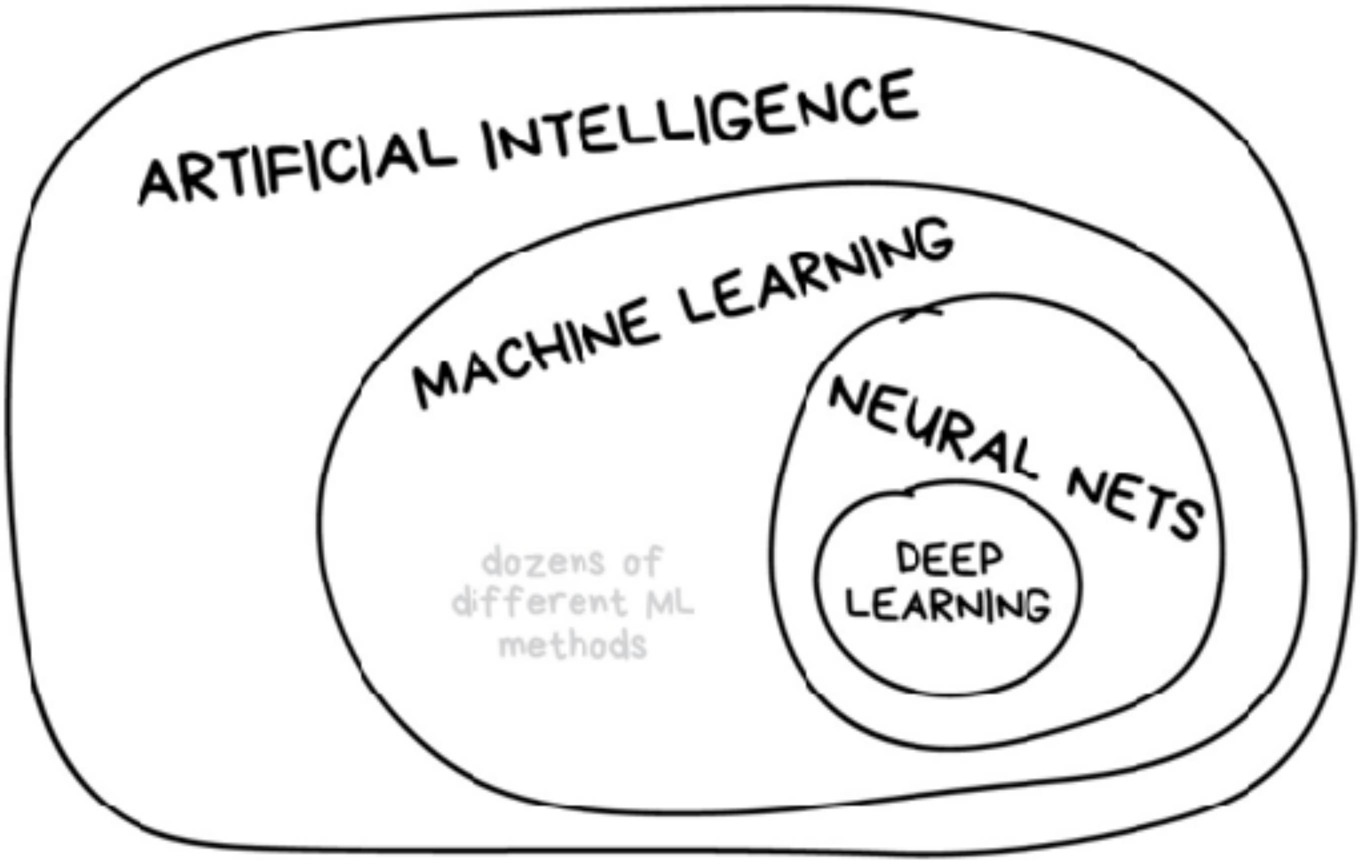


Image source: https://vas3k.com/blog/machine_learning/

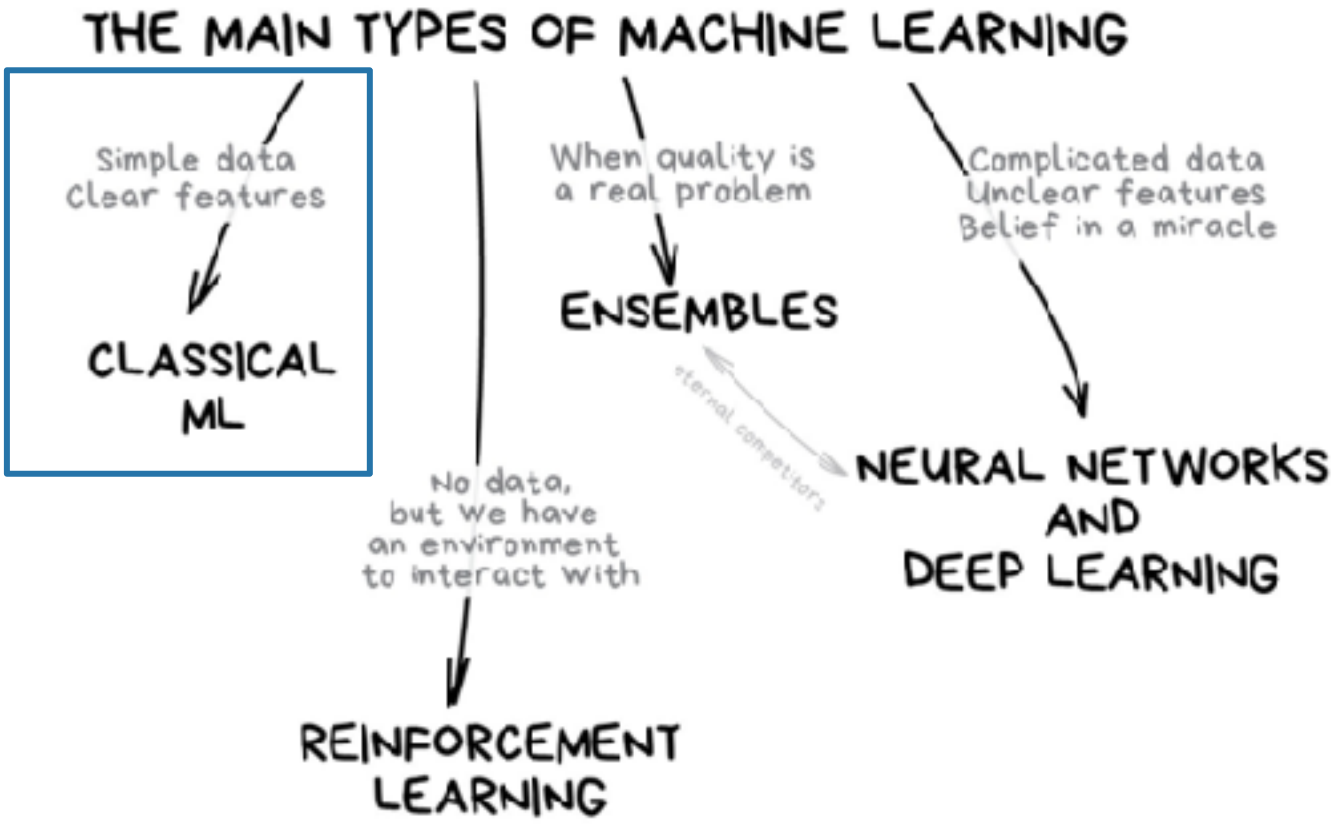


Image source: https://vas3k.com/blog/machine_learning/

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?

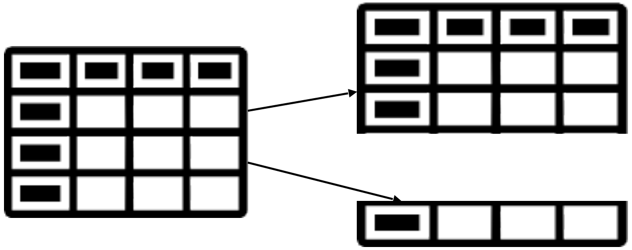
C What is the relationship between early childhood educational programming and success in elementary school?

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

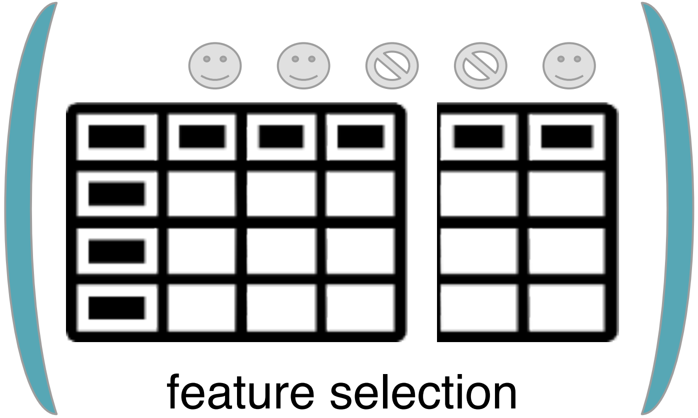
E How does Sesame Street cause an increase in educational attainment?

Machine Learning Generalizations

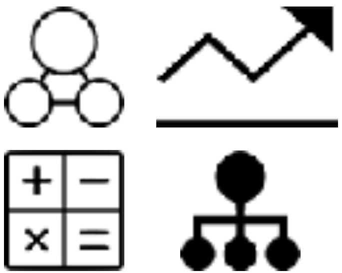
Basic Steps to Prediction



data partitioning

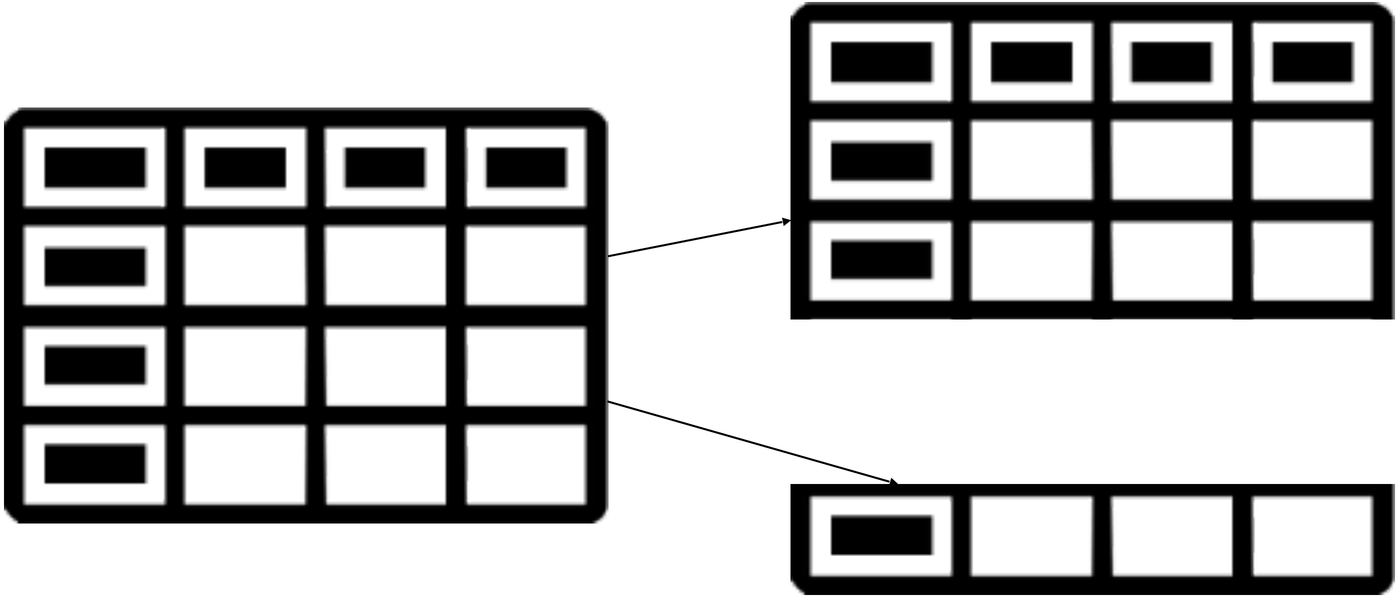


feature selection

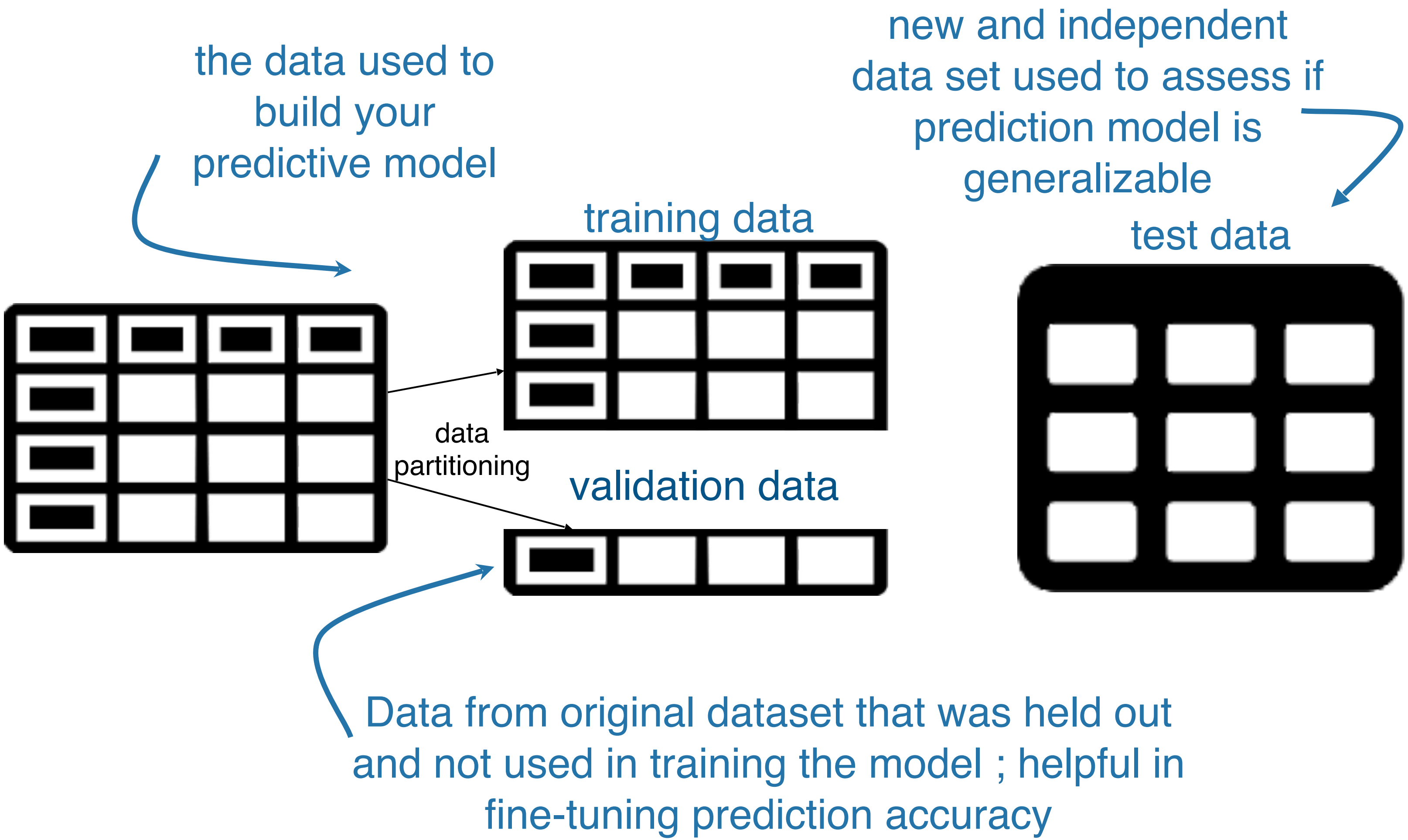


model selection

model assessment



data partitioning


the data used to build your predictive model
training data
data partitioning
validation data
new and independent data set used to assess if prediction model is generalizable
test data
Data from original dataset that was held out and not used in training the model ; helpful in fine-tuning prediction accuracy

Data Partitioning

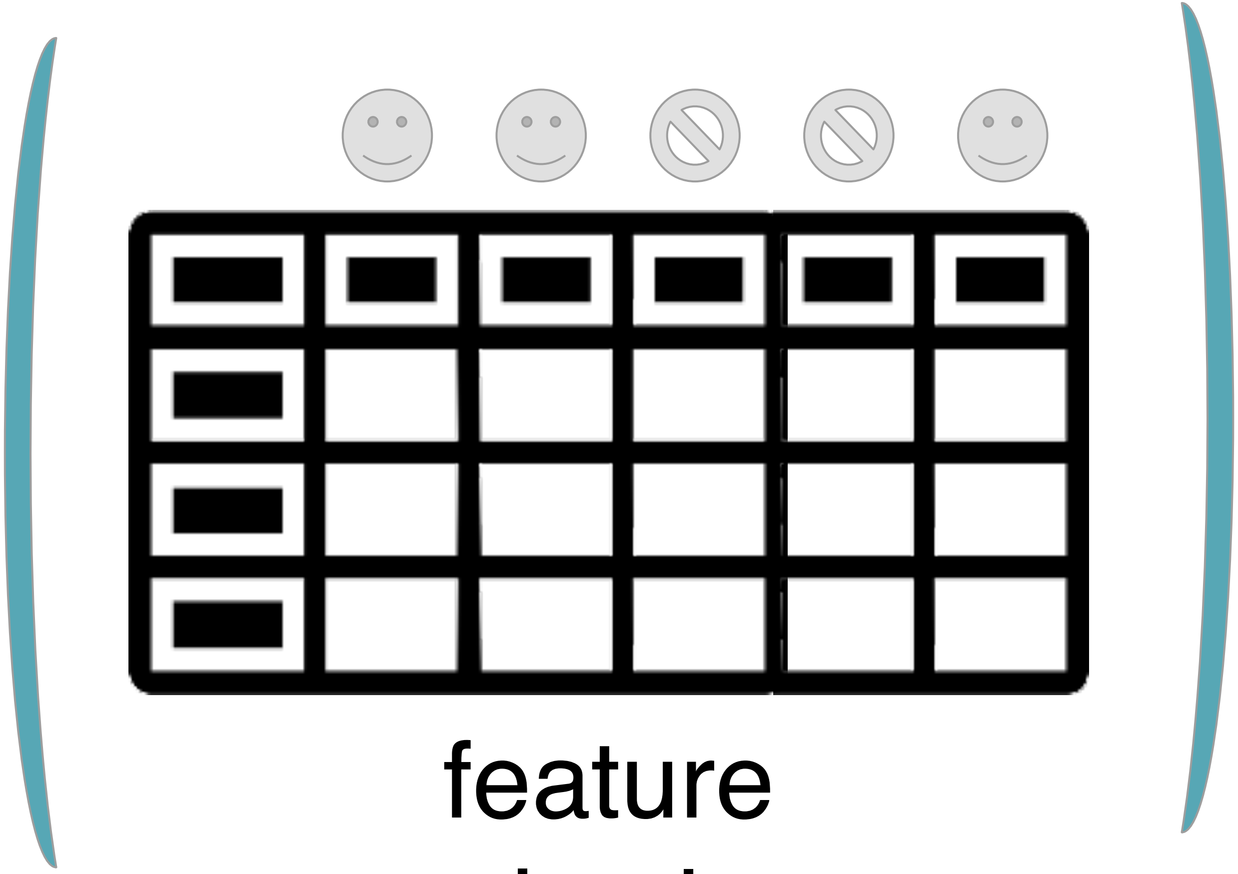
What portion of the data are typically used for generating the model?

A
The entire dataset

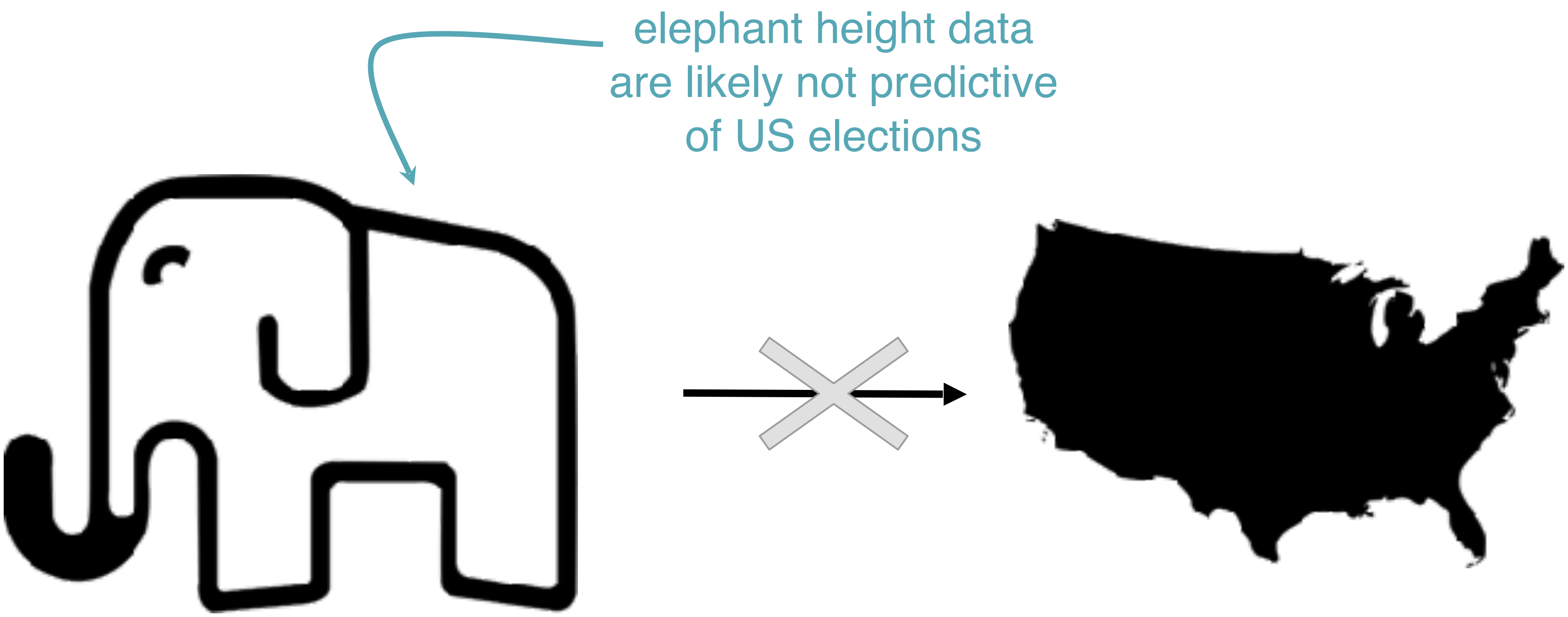
B
The training data

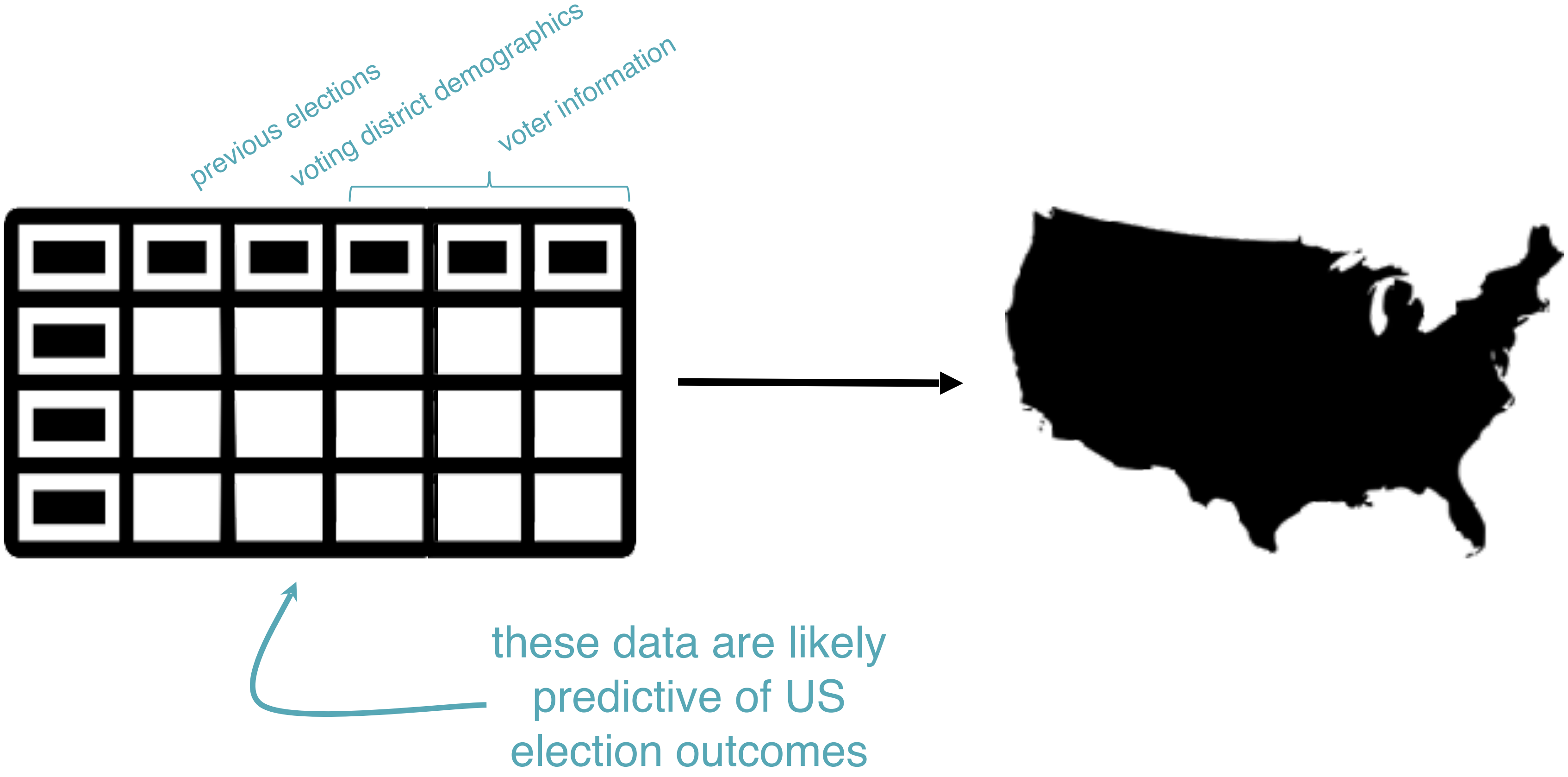
C
The testing data

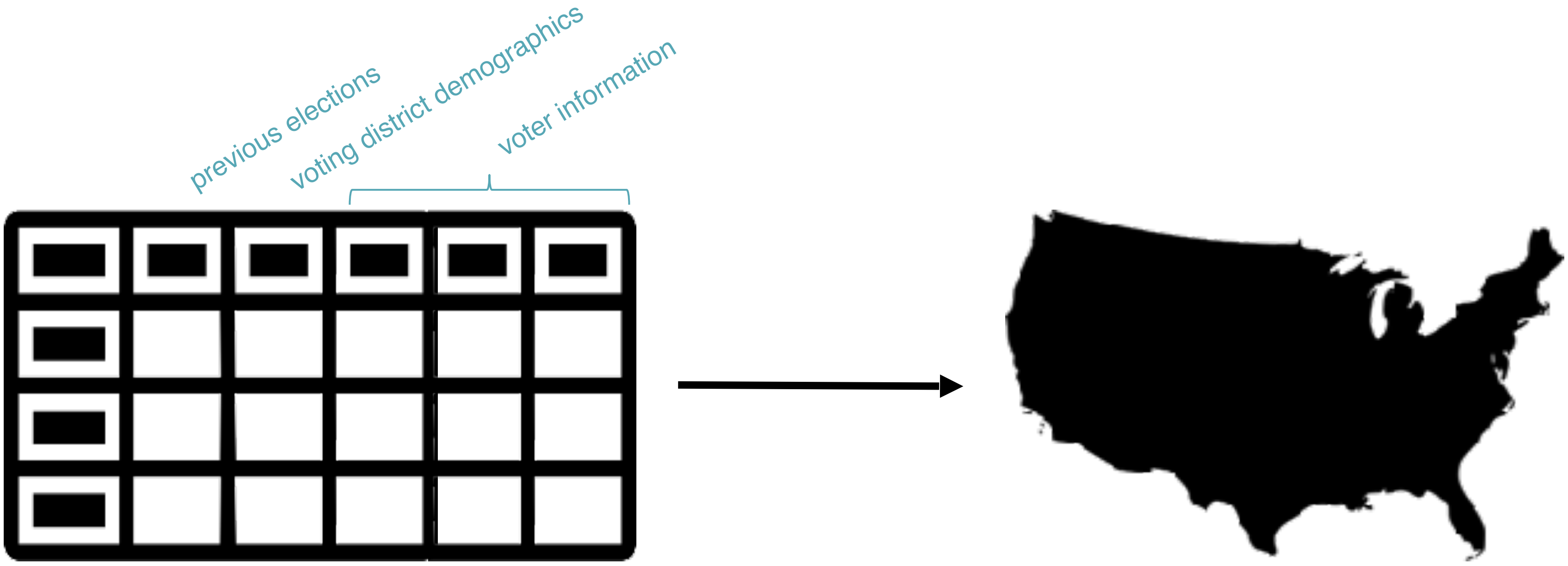
D
The validation data



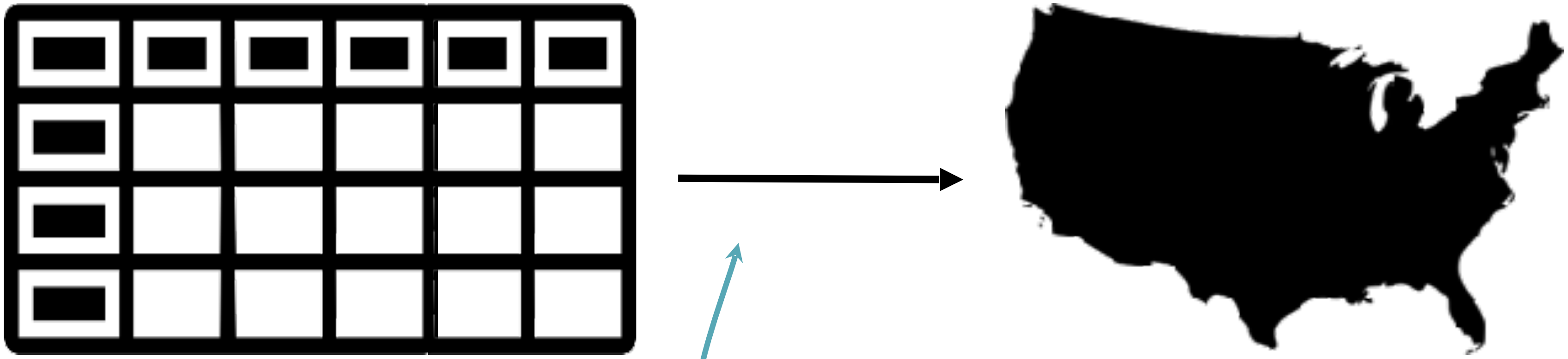
feature
selection


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

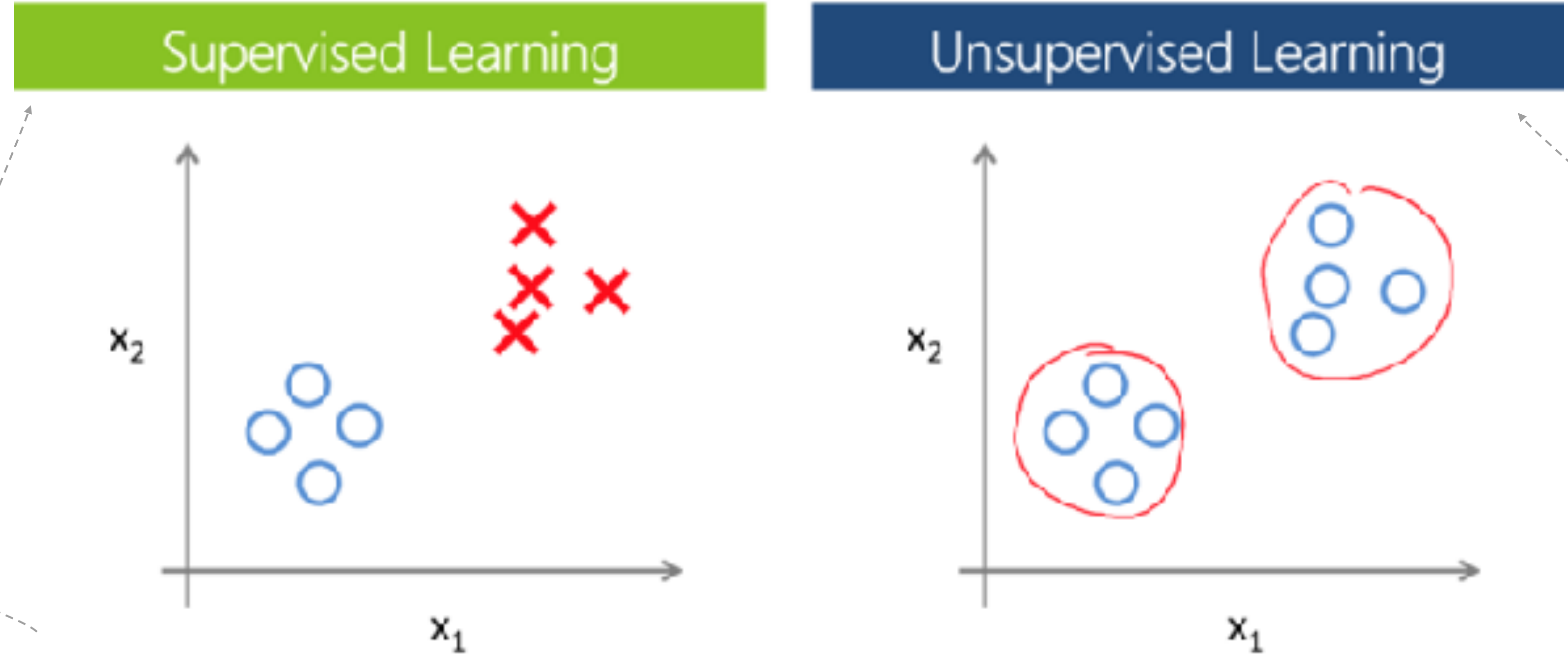


feature selection determines which variables are most predictive and includes them in the model



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

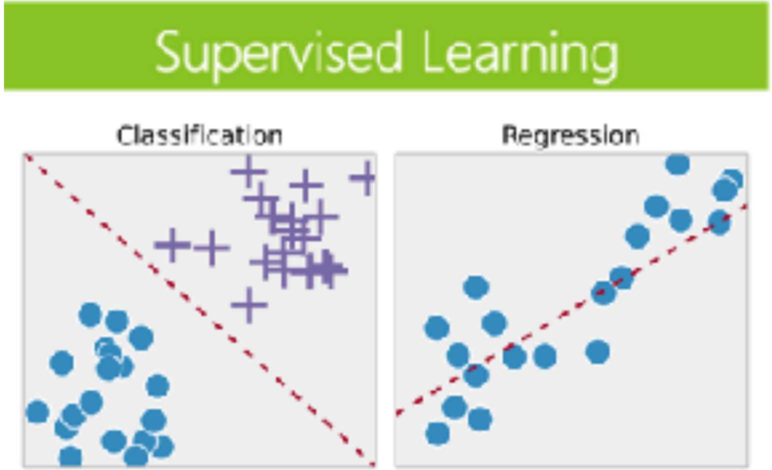
Two modes of machine learning



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

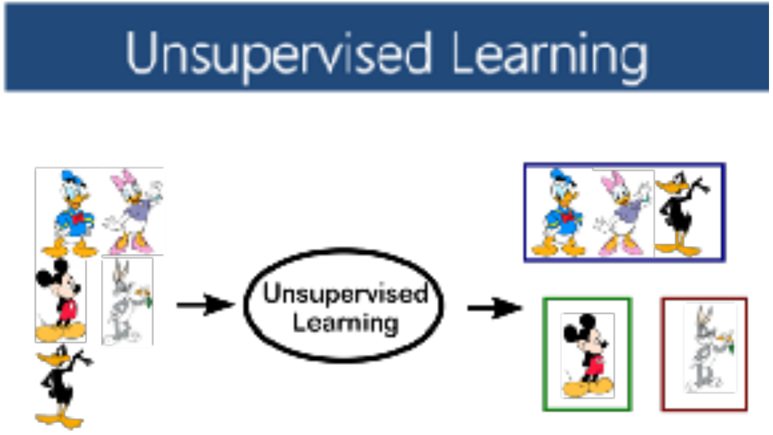
The computer determines how to classify based on properties within the data

Approaches to machine learning



categorical
variables

continuous
variables



Clustering (categorical)
& dimensionality reduction (continuous)

can automatically identify
structure in data

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

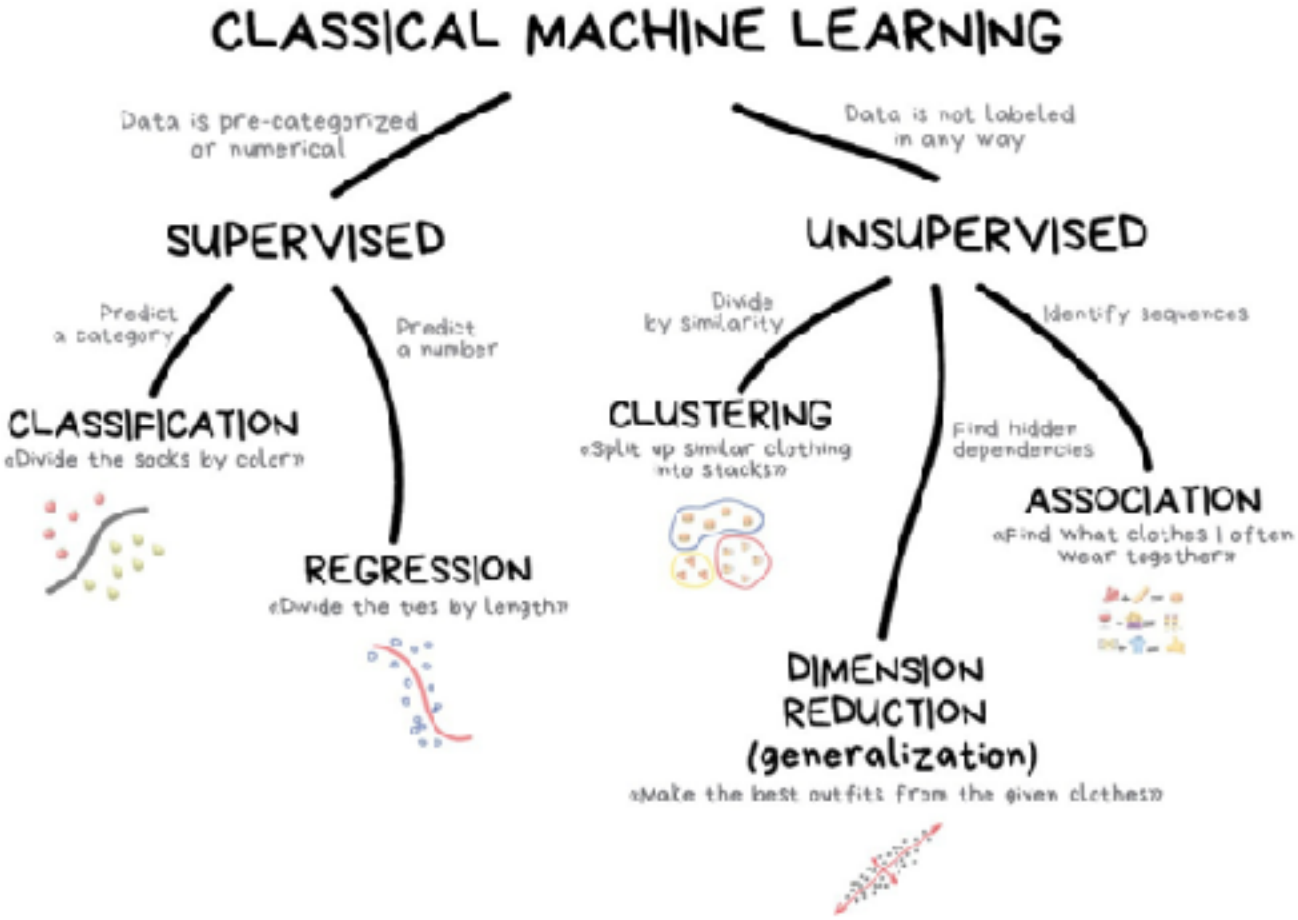


Image source: https://vas3k.com/blog/machine_learning/

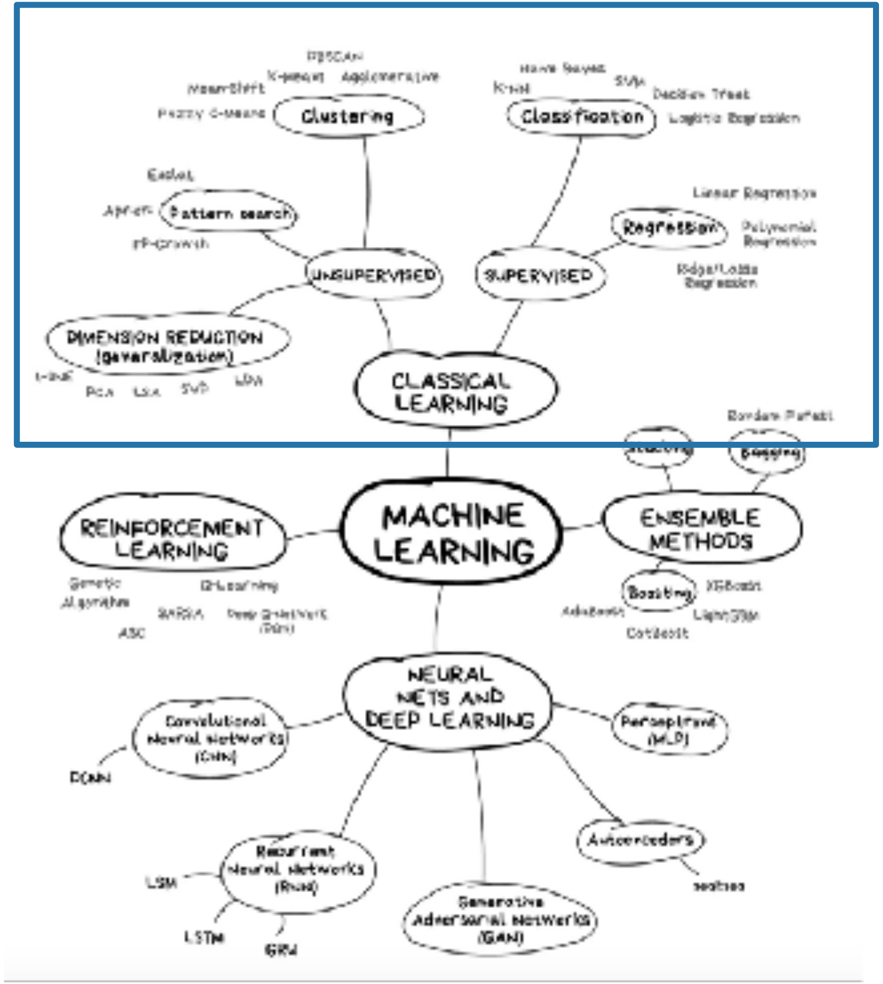


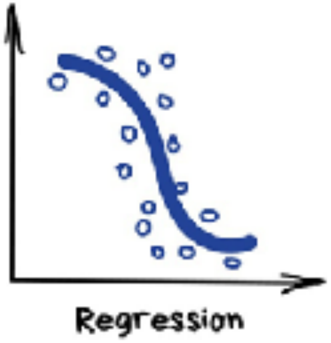
Image source: https://vas3k.com/blog/machine_learning/

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

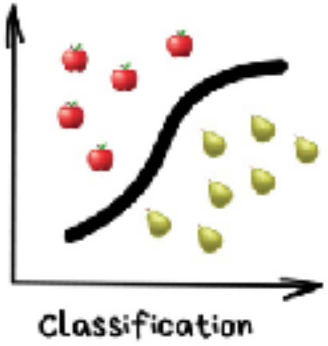


Popular algorithms are Linear and Polynomial regressions.

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:

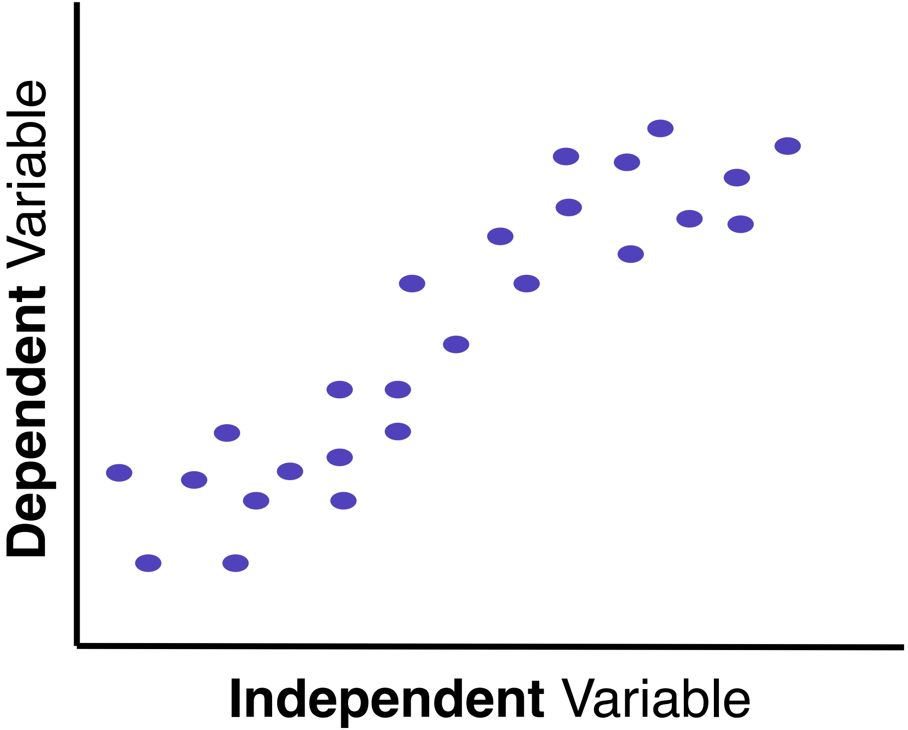
predicting categorical variables (i.e. education level)

categorical variable prediction

Supervised Learning

regression

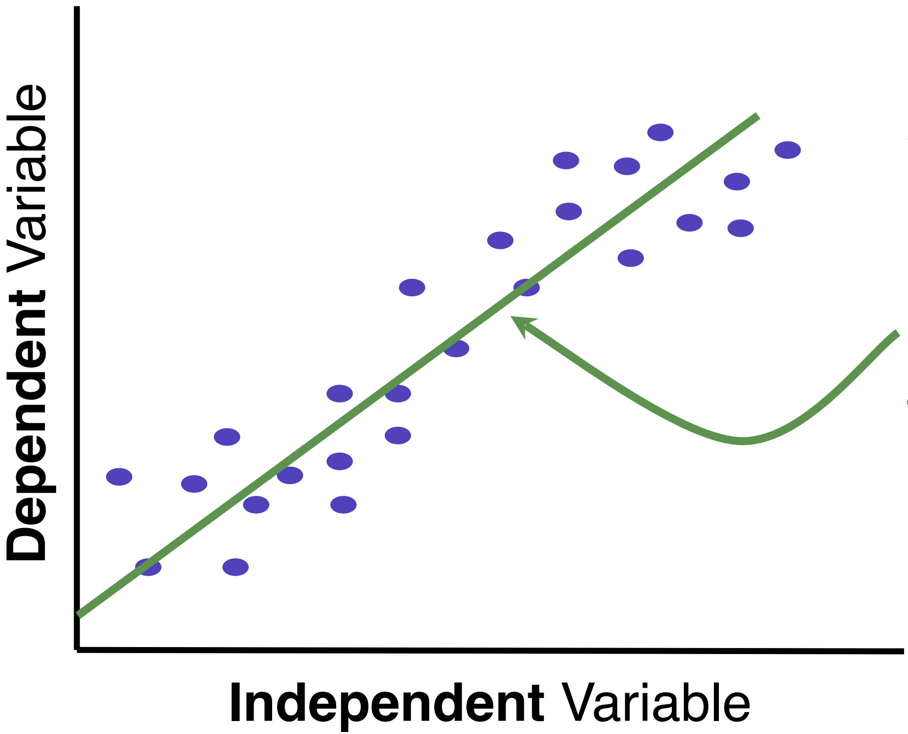
continuous variable prediction



Supervised Learning

regression

continuous variable prediction

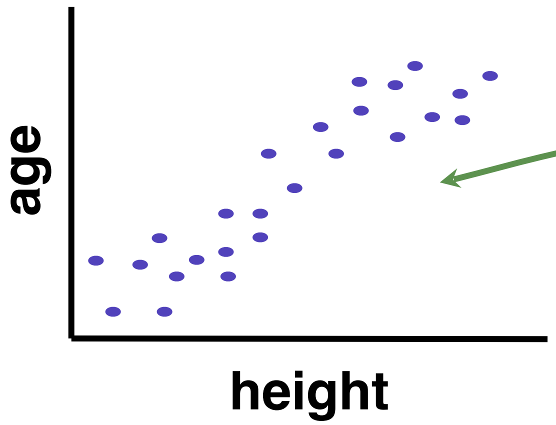


We'll use the linear relationship between variables to generate a **predictive model**

Supervised Learning

regression

continuous variable prediction

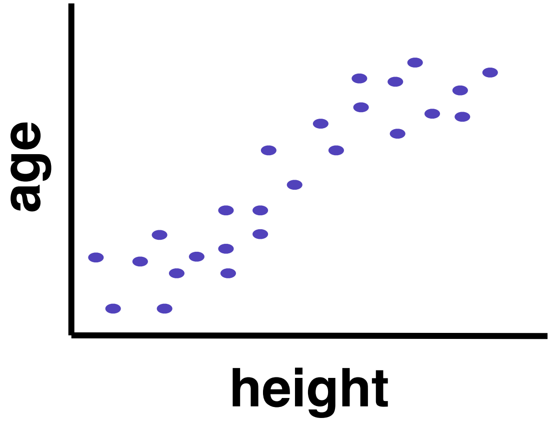


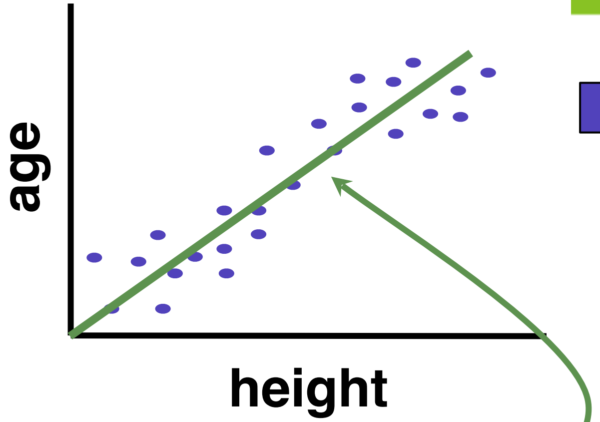
the training data will be used to build the predictive model

Supervised Learning

regression

continuous variable prediction



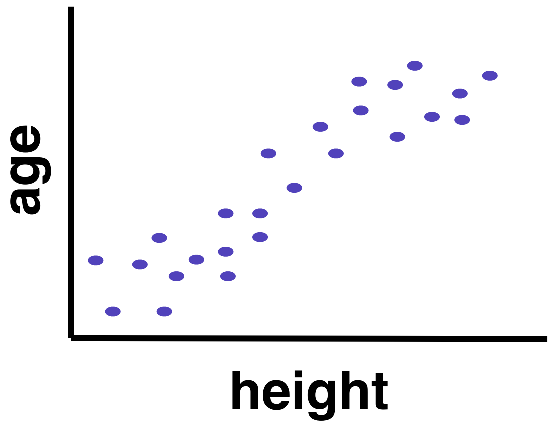


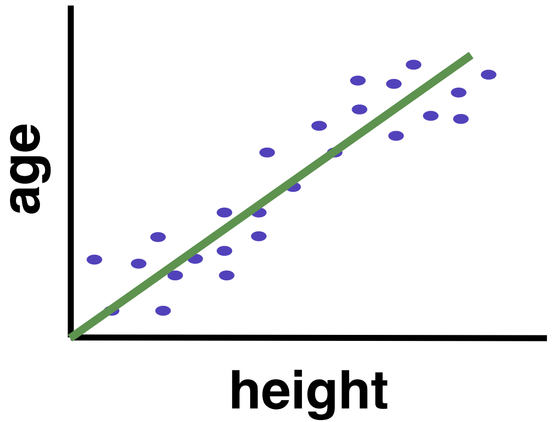
use linear regression to model the relationship

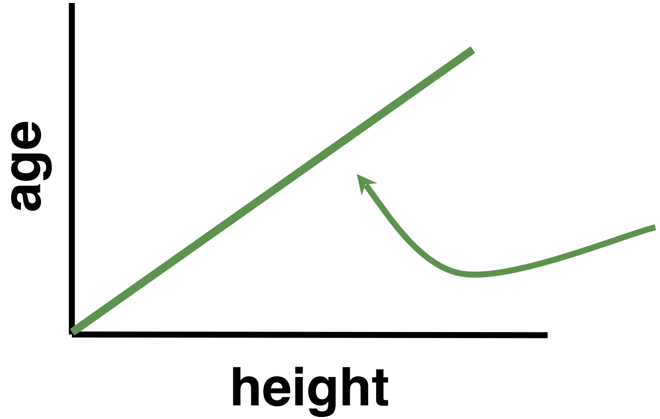
Supervised Learning

regression

continuous variable prediction





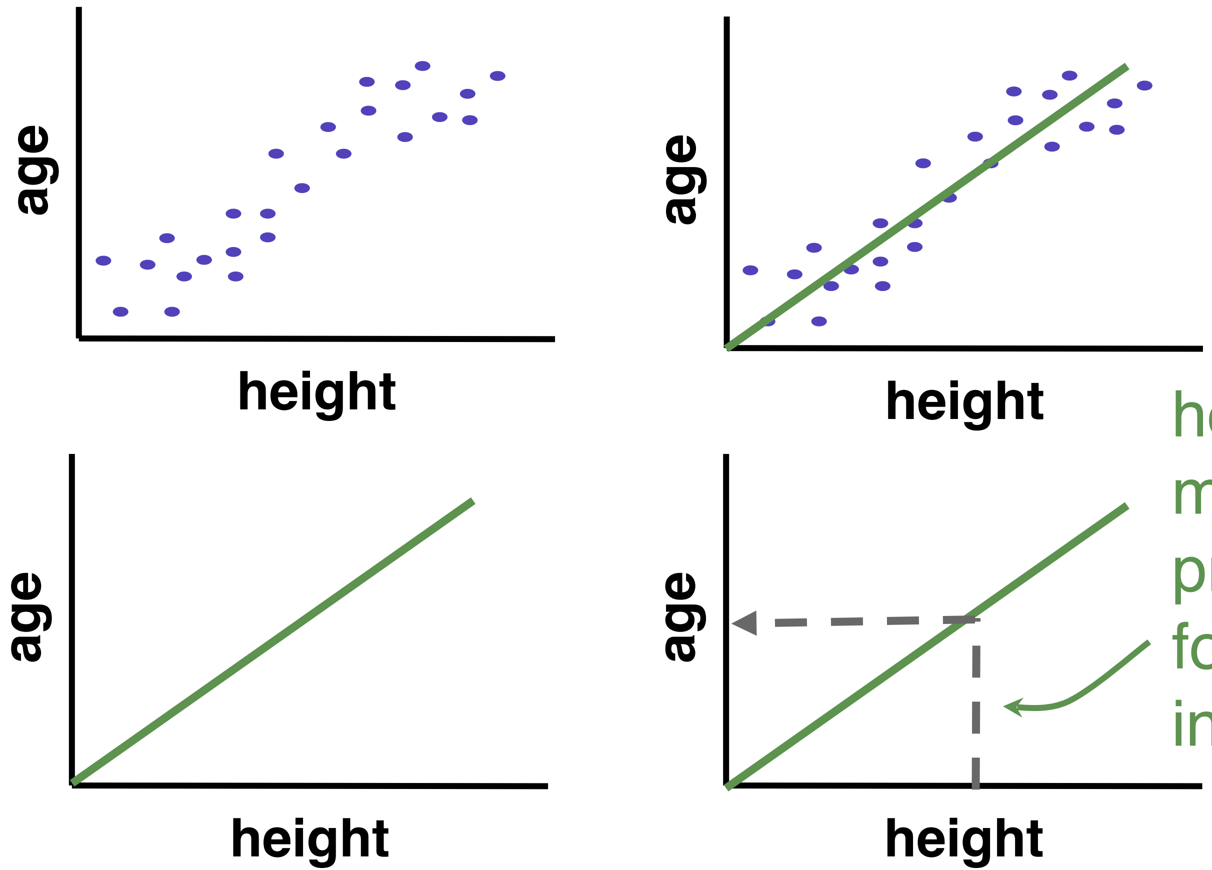


For prediction, the individual values in the training data are *not* important. We only need the model.

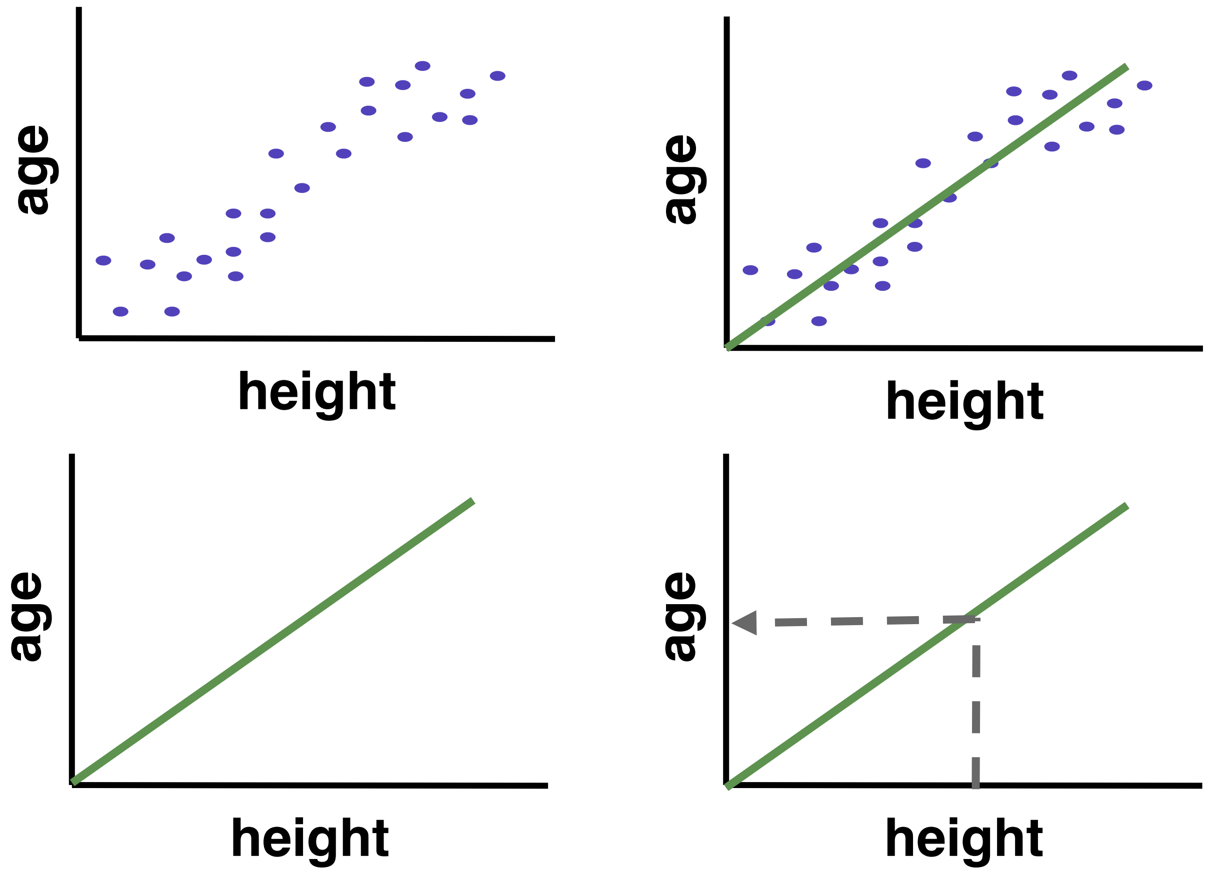
Supervised Learning

regression

continuous variable prediction



how we'll make predictions for a future individual



Supervised Learning

regression

continuous variable prediction

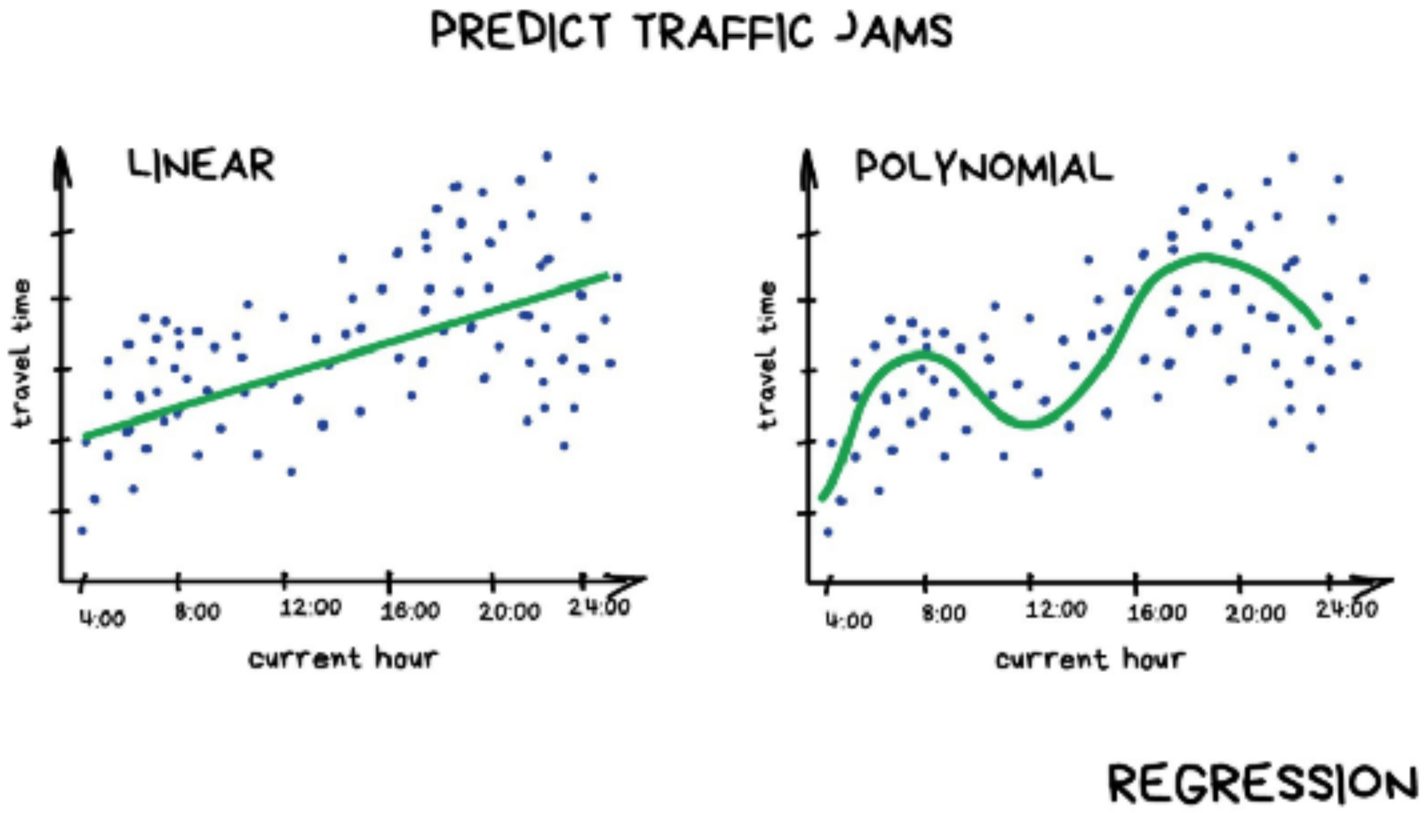


Image source: https://vas3k.com/blog/machine_learning/

Regression:
predicting <u>continuous</u>
variables
(i.e. Age)

Classification:
predicting <u>categorical</u>
variables
(i.e. give a loan?)

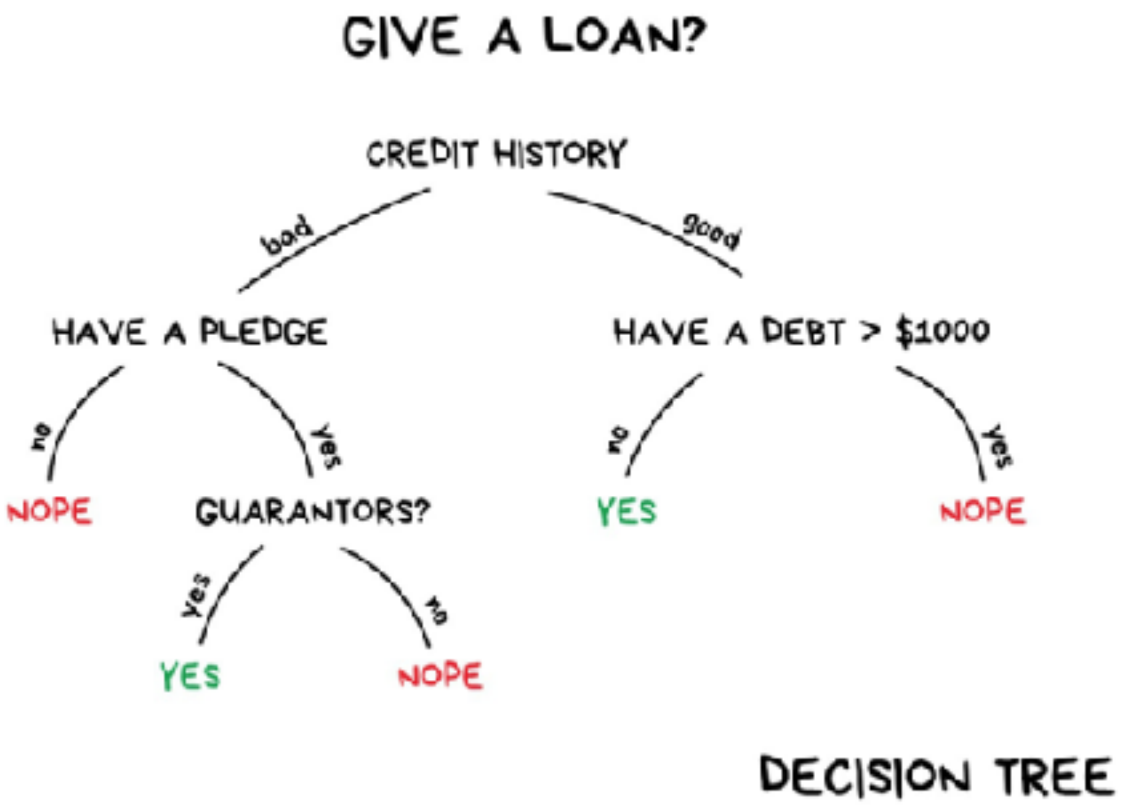
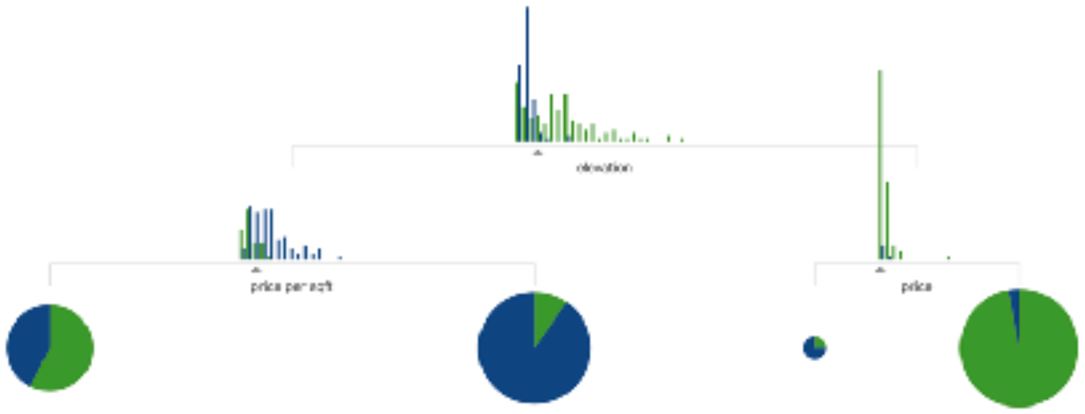


Image source: https://vas3k.com/blog/machine_learning/

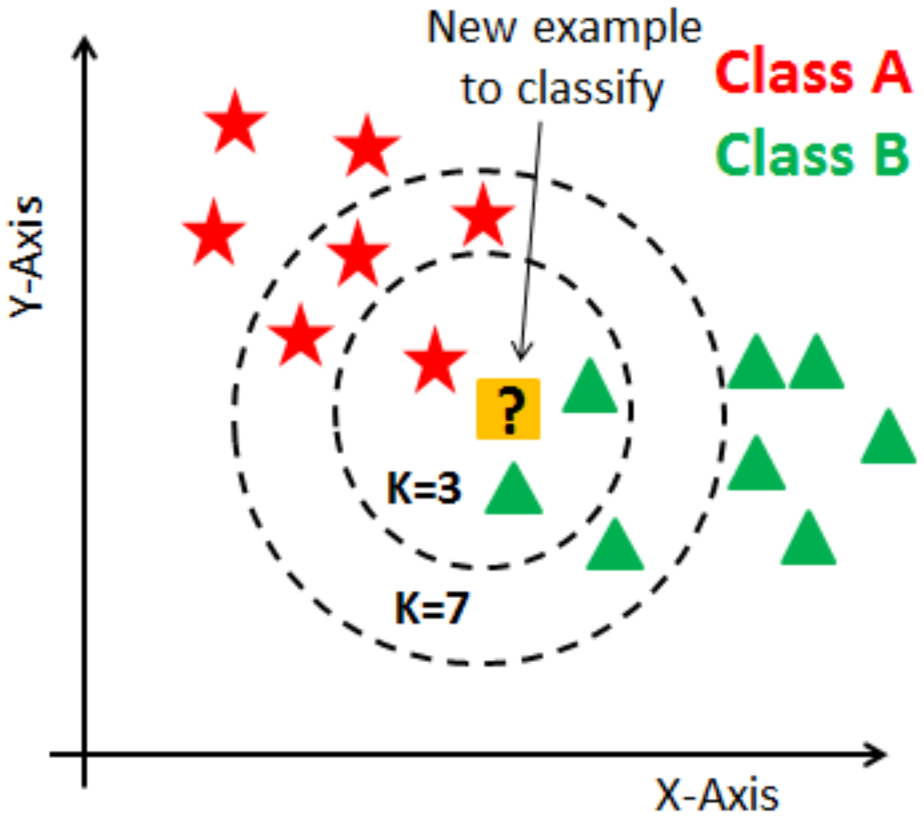
Growing a tree

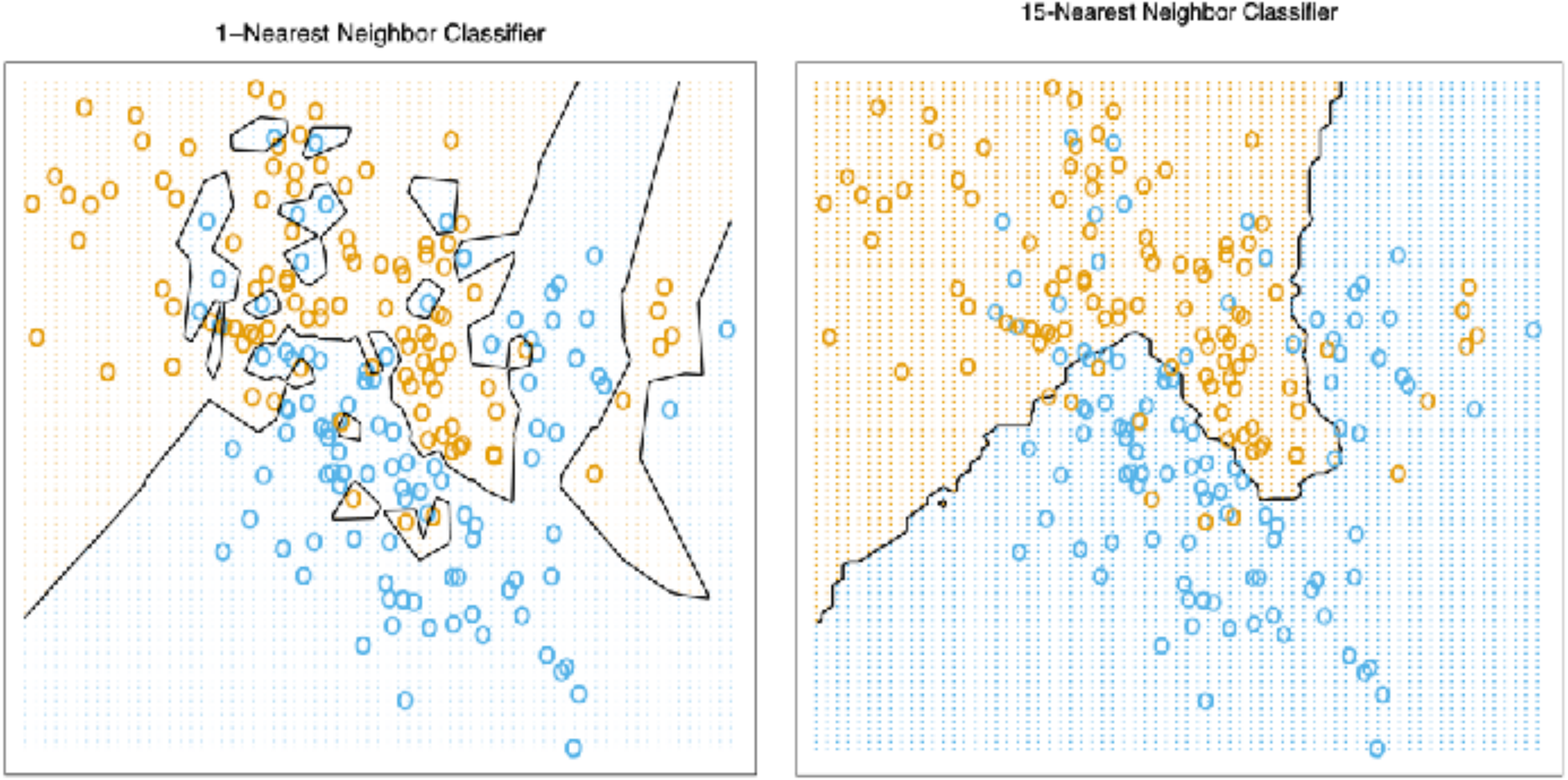
Additional forks will add new information that can increase a tree's **prediction accuracy.**

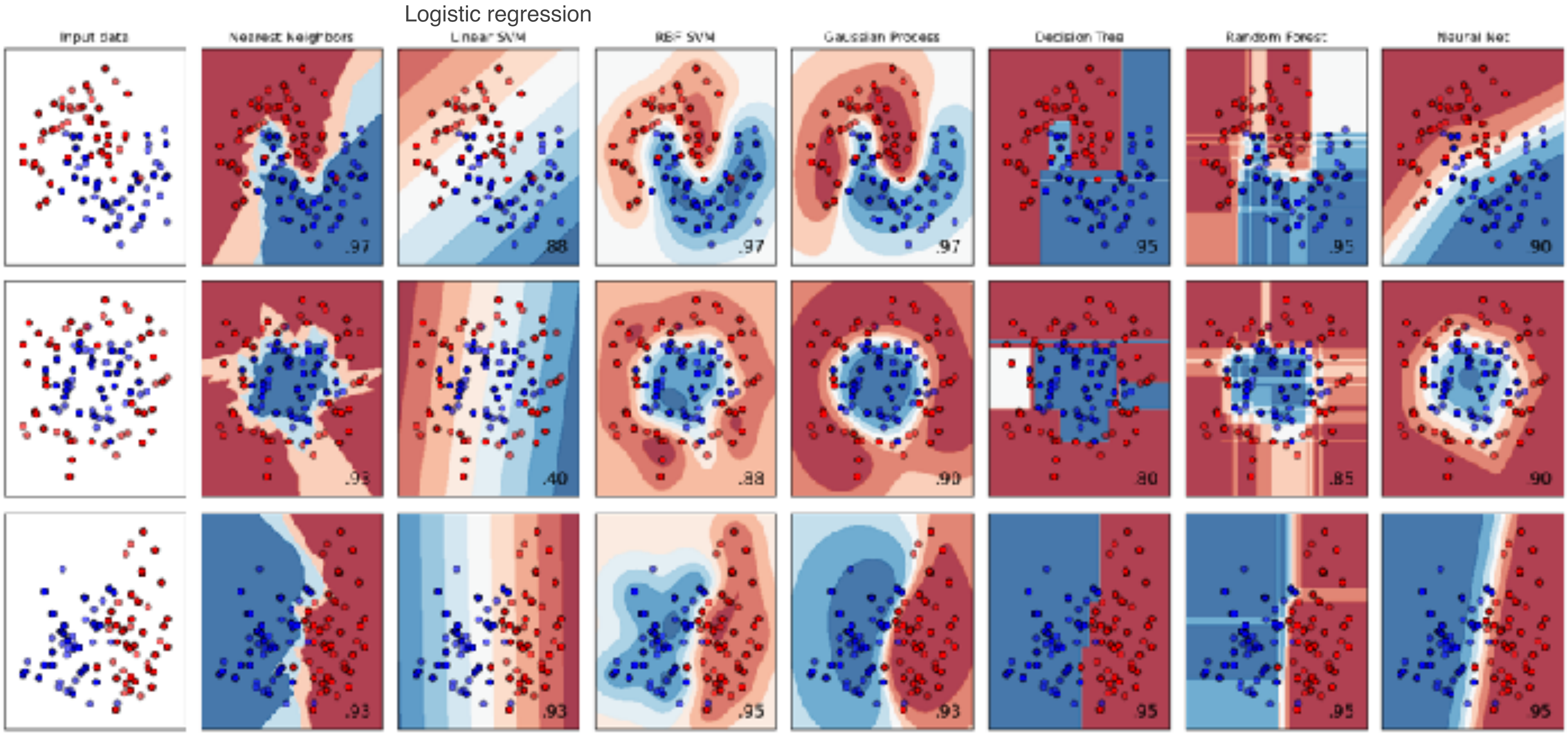


http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

K-nearest neighbors



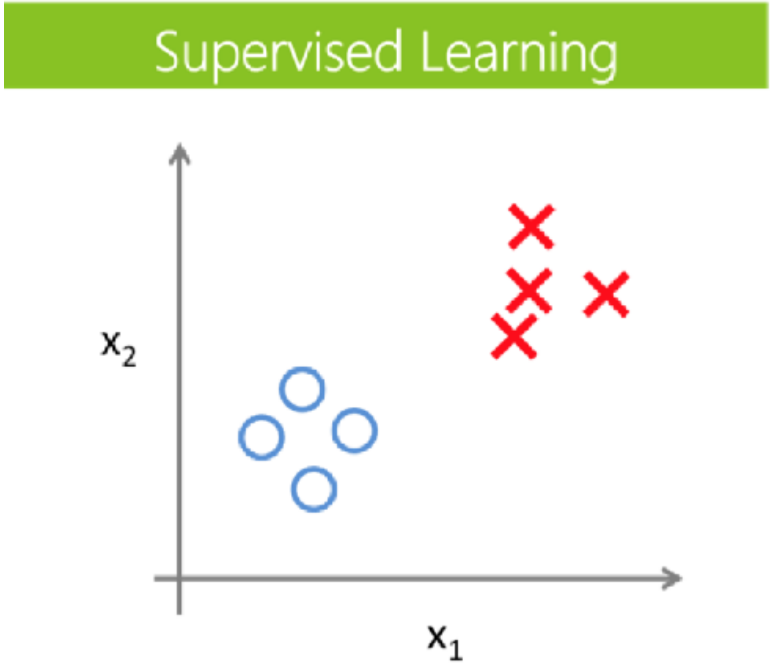

1–Nearest Neighbor Classifier
15-Nearest Neighbor Classifier

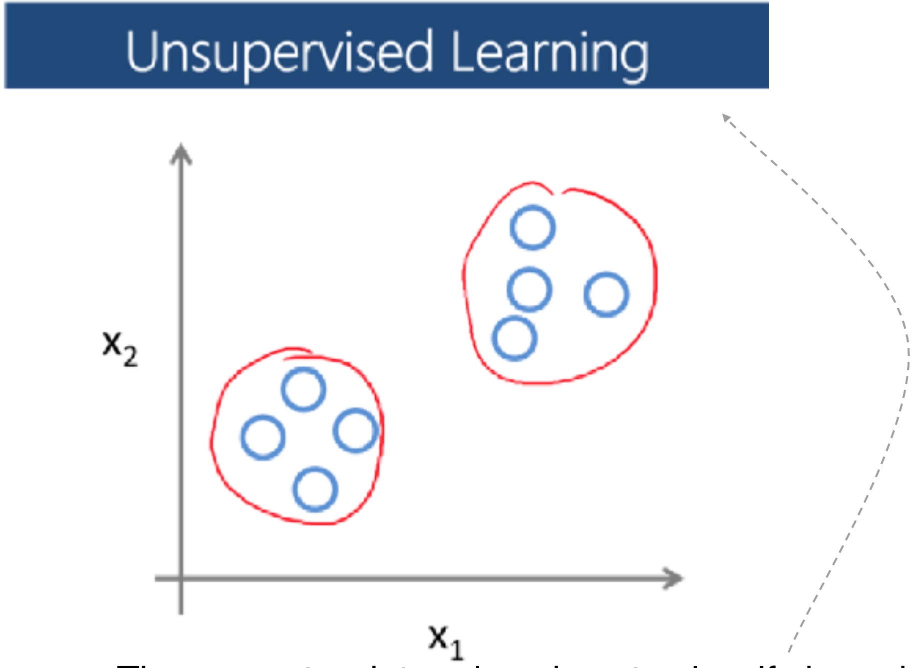


https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Unsupervised Learning

To modes of machine learning





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:

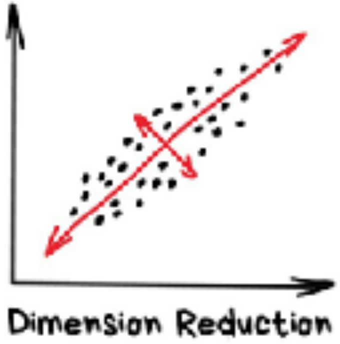
- For market segmentation (types of customers, loyalty)
- To merge close points on a map
- For image compression
- To analyze and label new data
- To detect abnormal behavior

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)

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

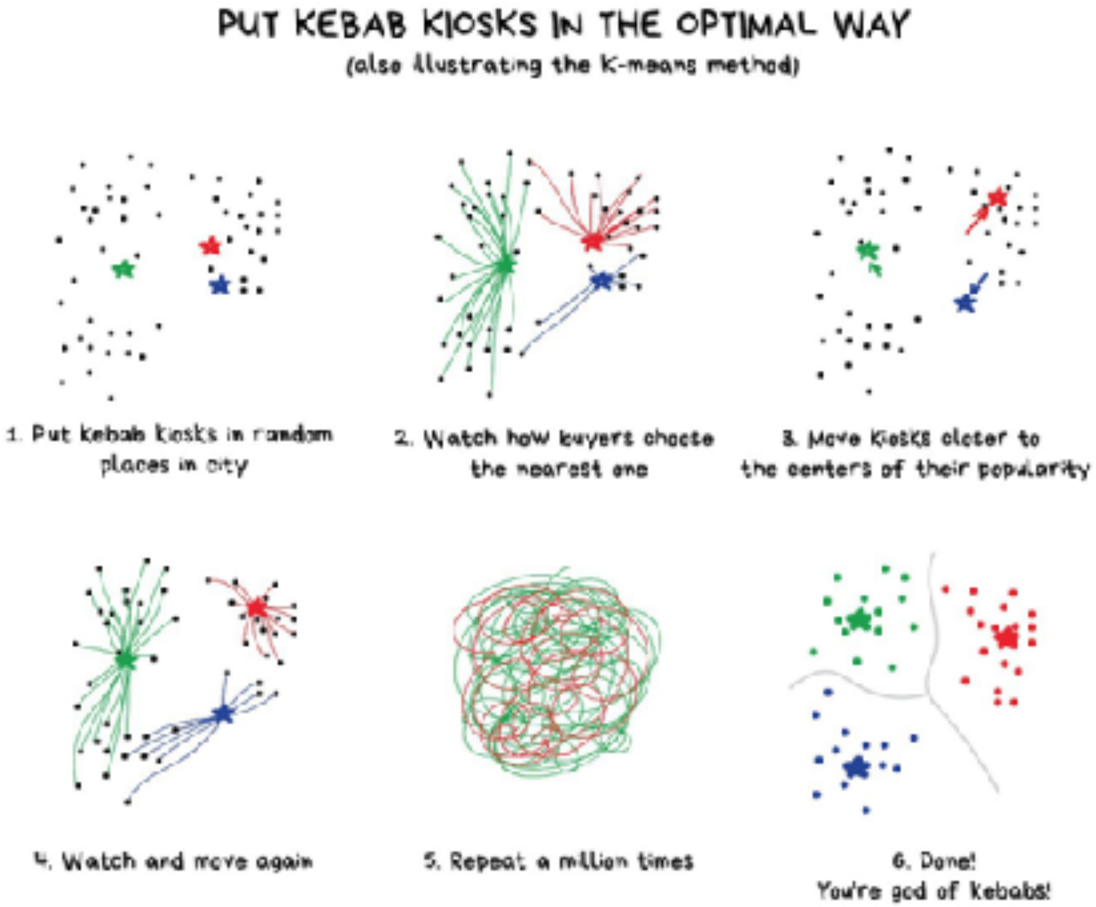


Image source: https://vas3k.com/blog/machine_learning/

Prediction Approach

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

How would you approach this with machine learning?

A
Supervised, Regression

B
Supervised, Classification

C
Unsupervised, dimensionality reduction

D
Unsupervised, clustering

E
Unsupervised, Neural Network

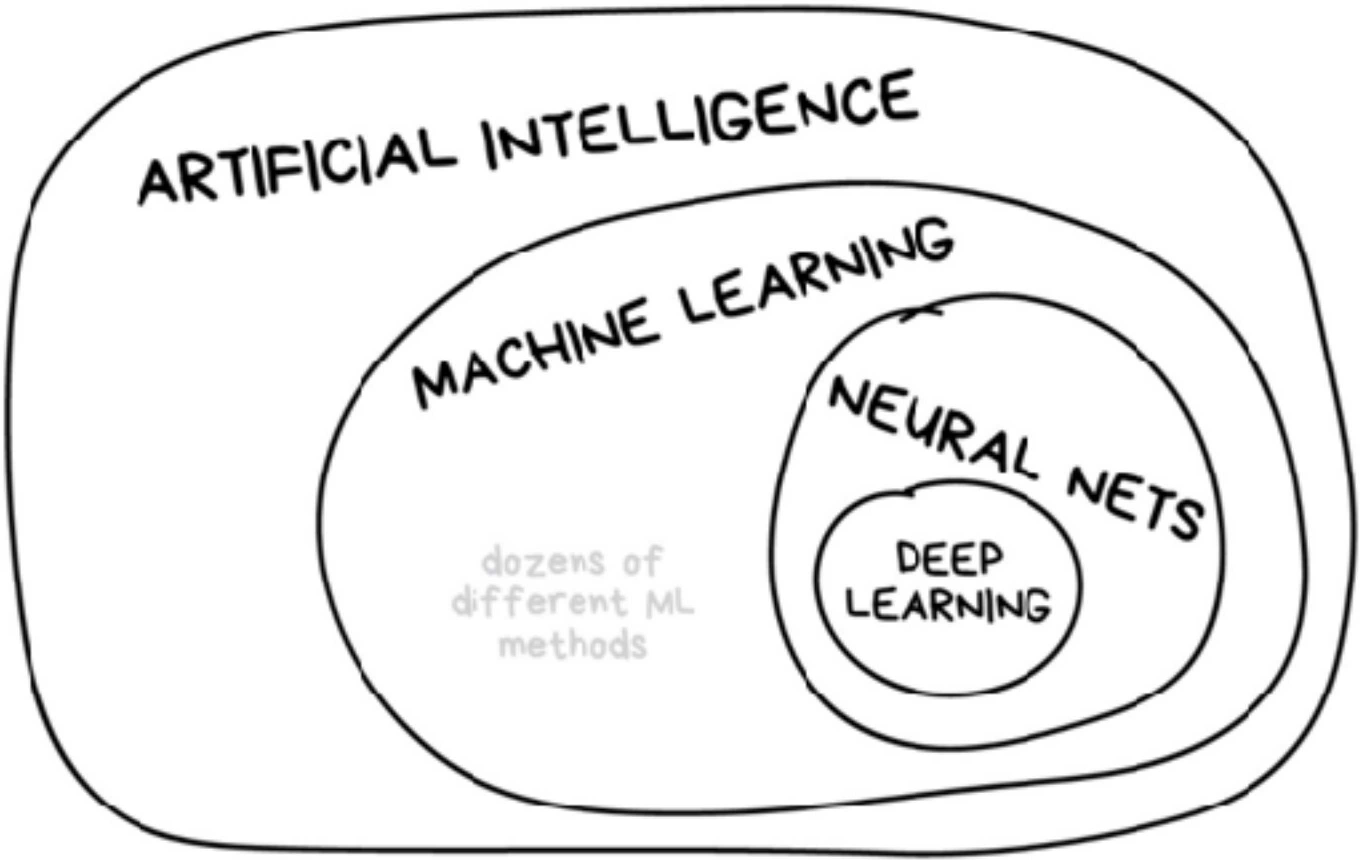


Image source: https://vas3k.com/blog/machine_learning/

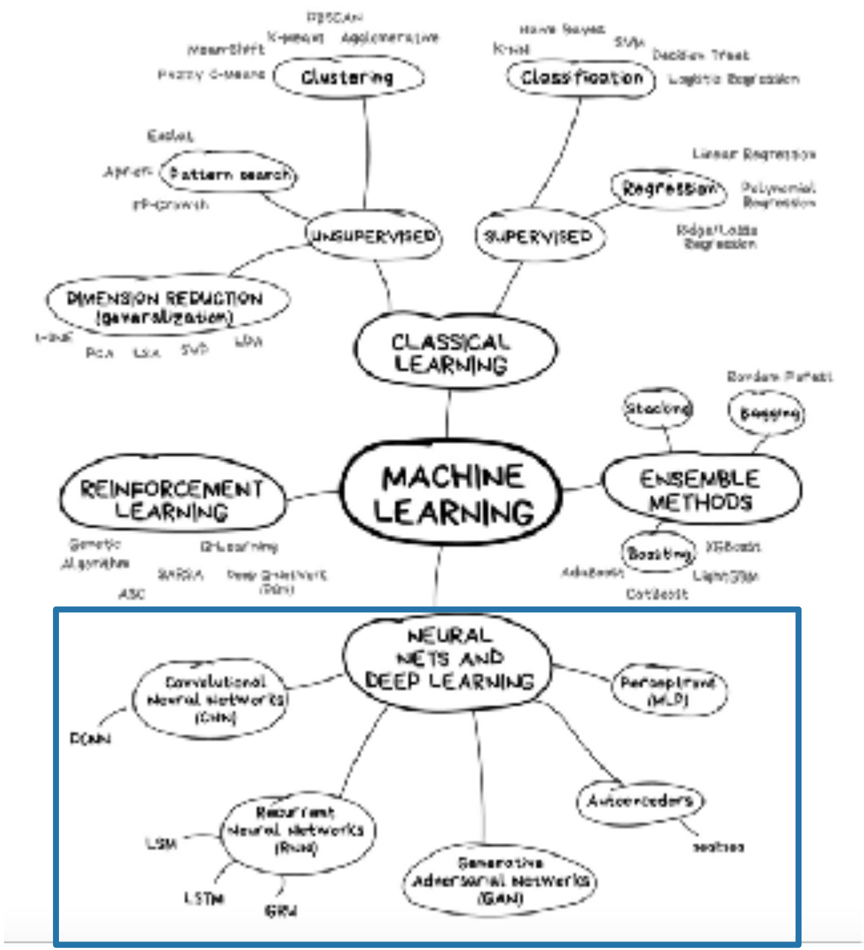
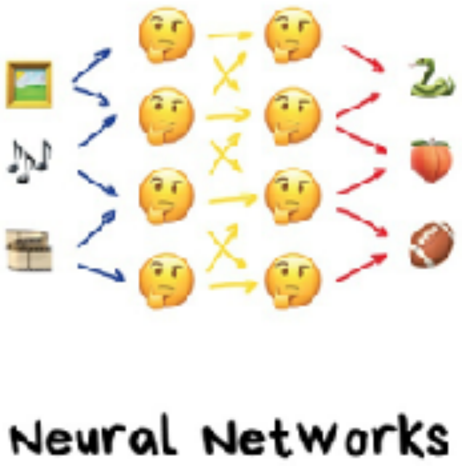


Image source: https://vas3k.com/blog/machine_learning/

"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:

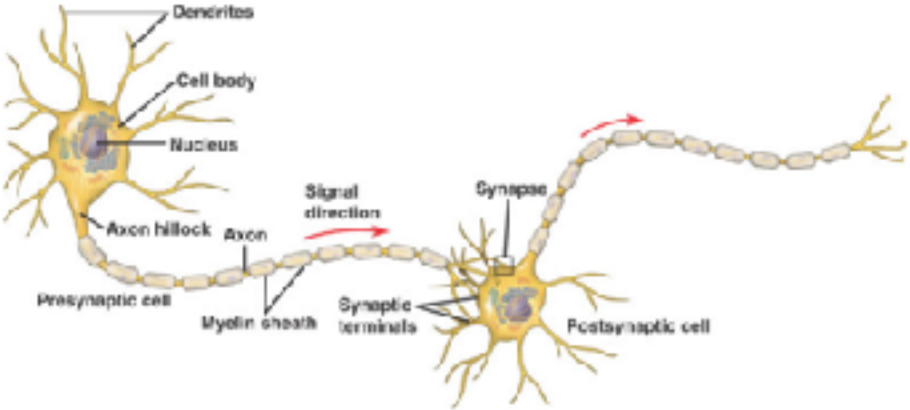
- Replacement of all algorithms above
- Object identification on photos and videos
- Speech recognition and synthesis
- Image processing, style transfer
- Machine translation



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

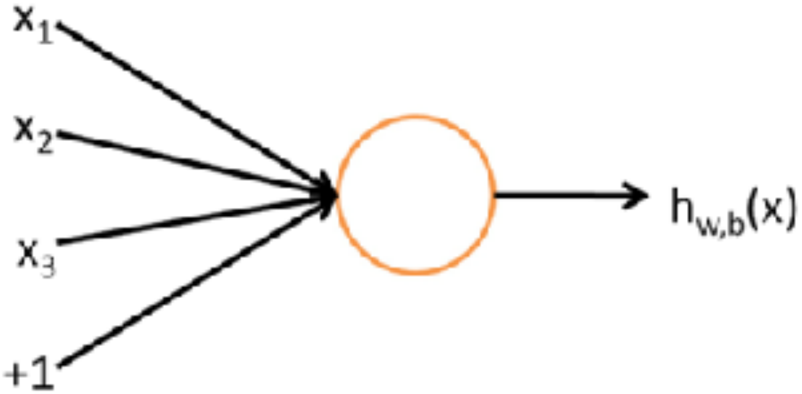
Image source: https://vas3k.com/blog/machine_learning/

WHAT IS A NEURON?



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

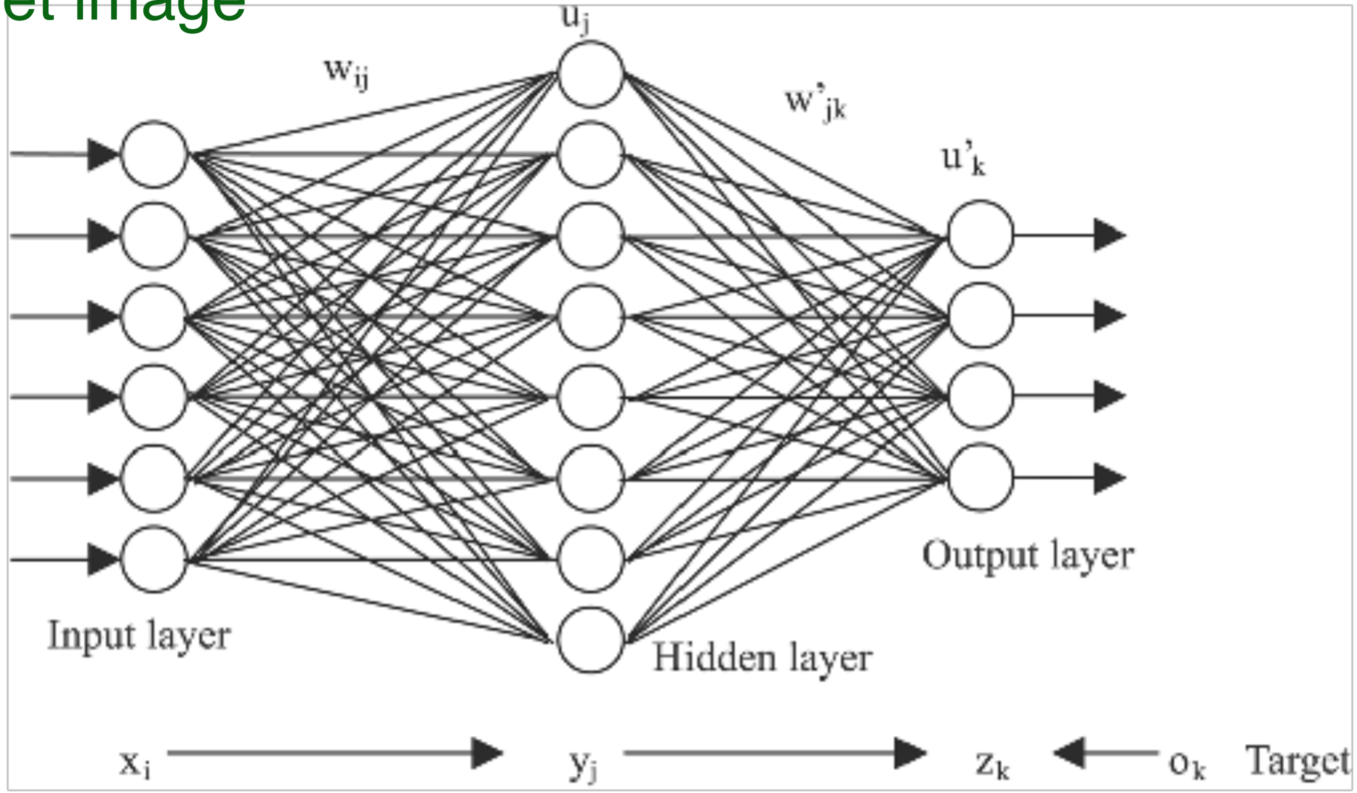
MATHEMATICAL NEURON



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

adapted from Brad Voytek

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



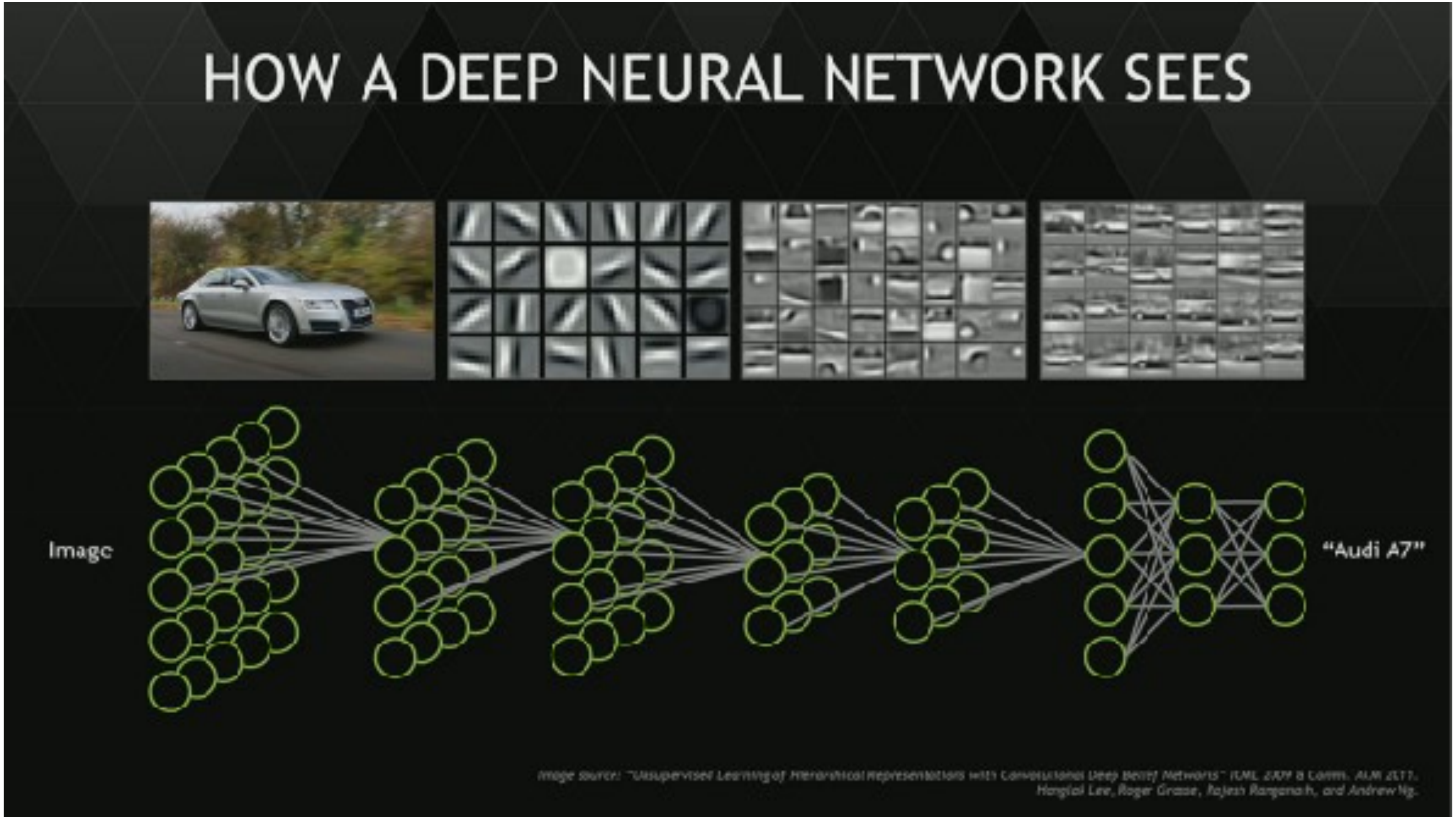


Image Source: https://towardsdatascience.com/understanding-residual-networks-9add4b664b03

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.

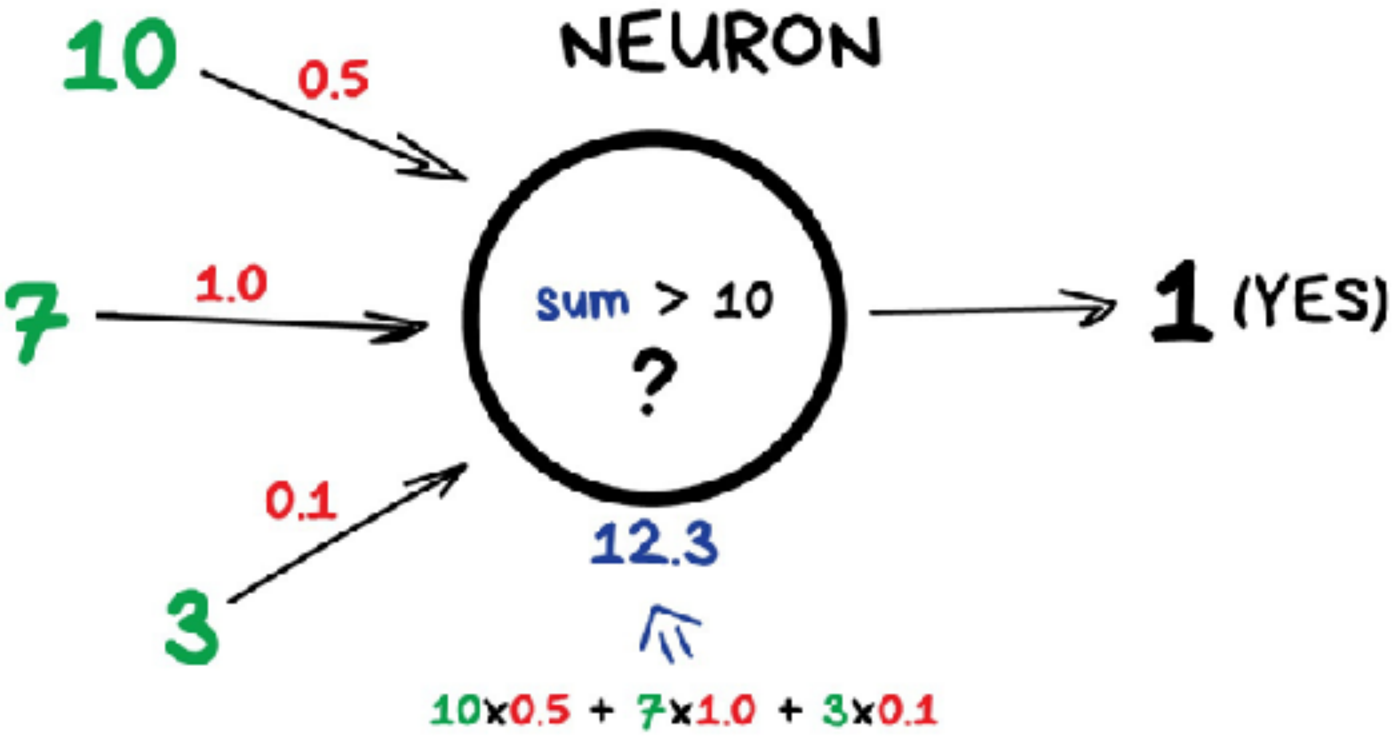


Image source: https://vas3k.com/blog/machine_learning/

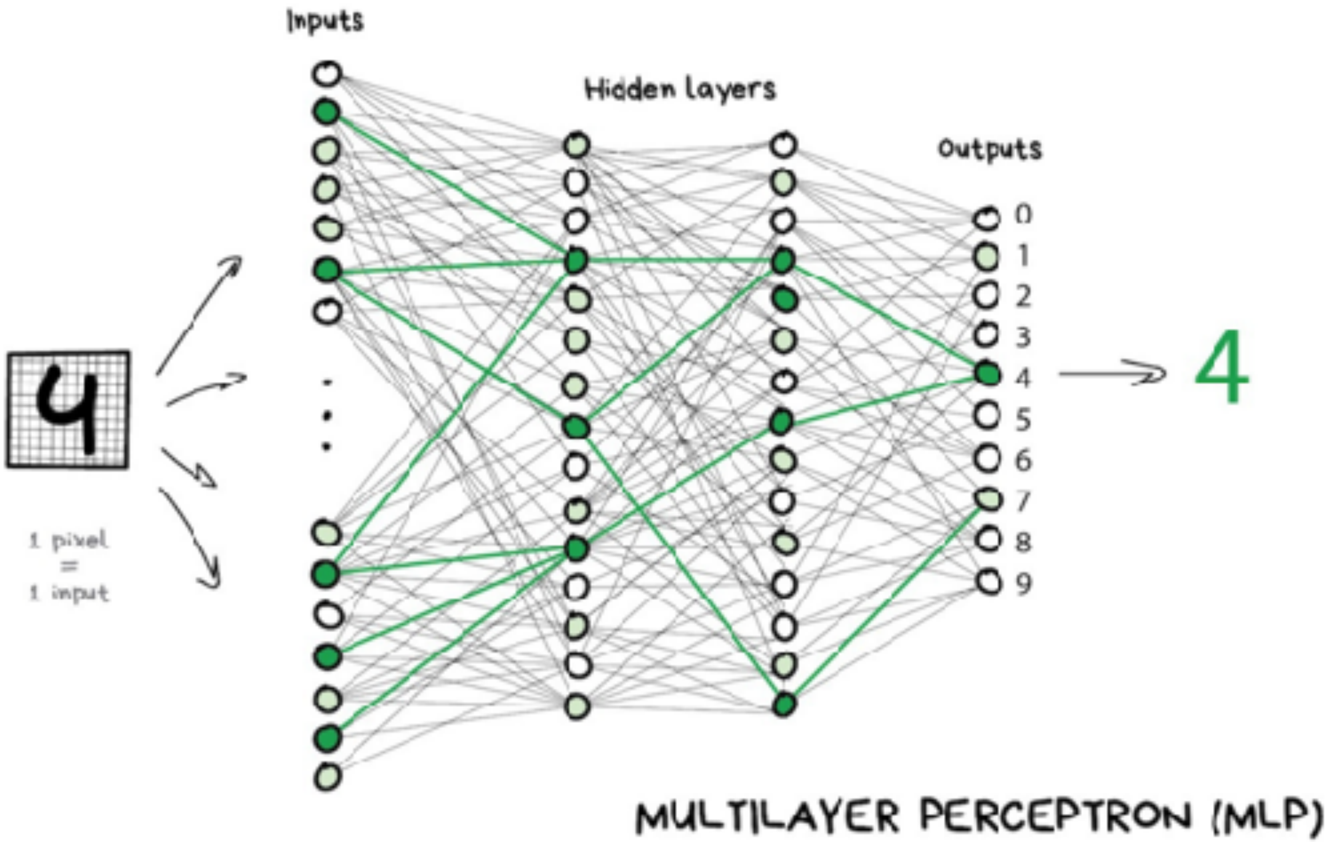


Image source: https://vas3k.com/blog/machine_learning/

Manually labeling used to be the way...

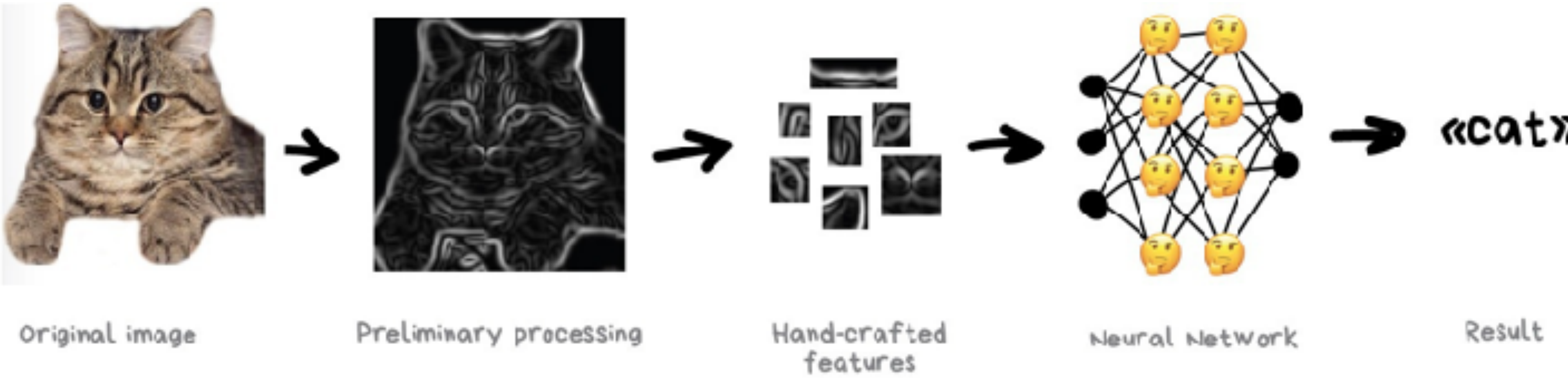
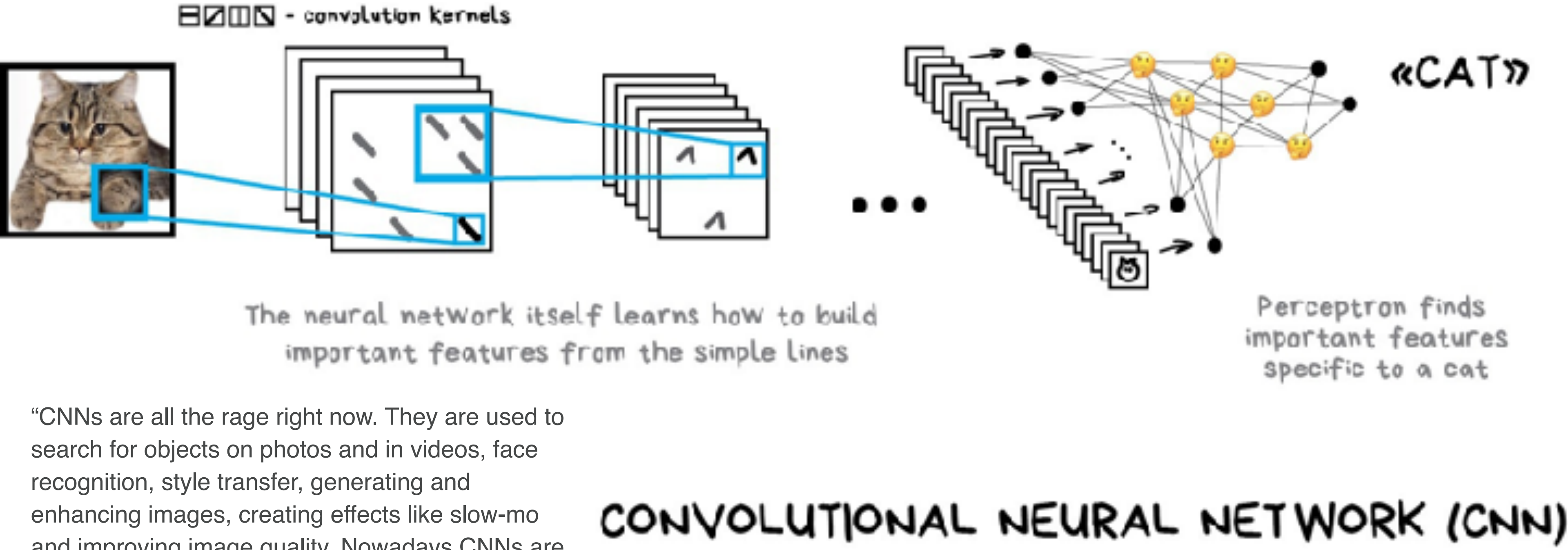


Image source: https://vas3k.com/blog/machine_learning/

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."

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%

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

- 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)

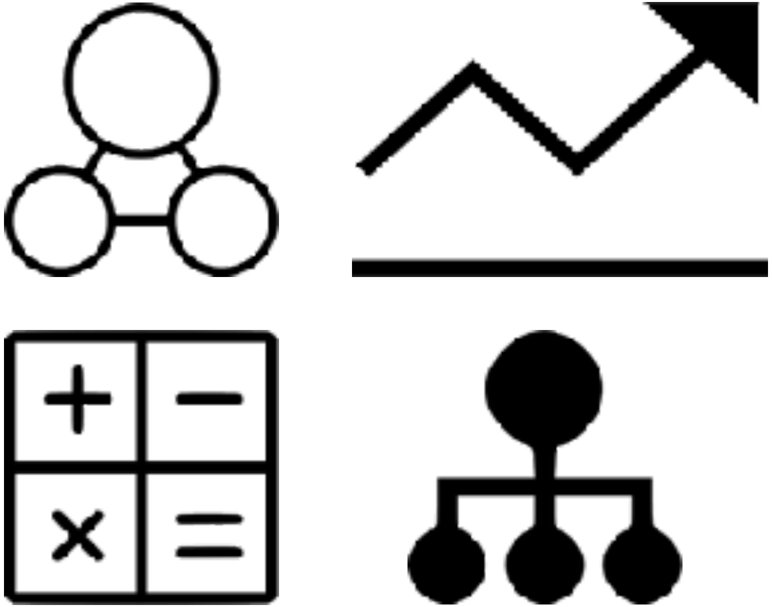
candy theme: normal

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

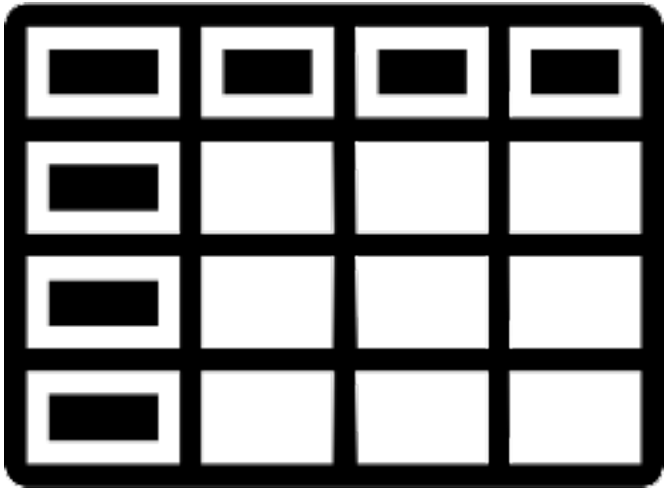
candy theme: programming

messages:

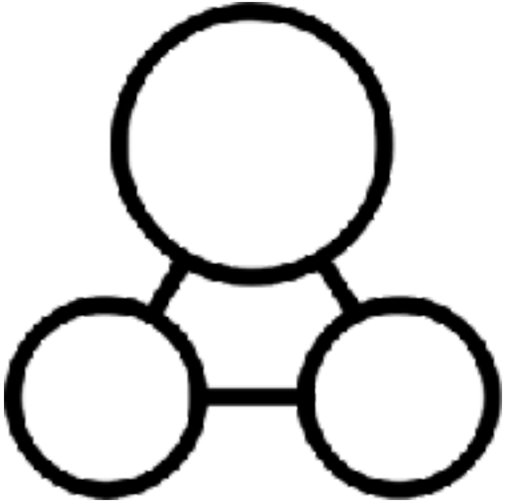
(1) I <3 programmers, (2) Programmers do it better, (3) If you can't code it, you can't date it, (4) Code is poetry, (5) Programmers make the world go round, (6) If you can't code, you can't love, (7) Love is the language of the future, (8) I heart code, (9) Code is life, (10) If you don't code, you



model selection



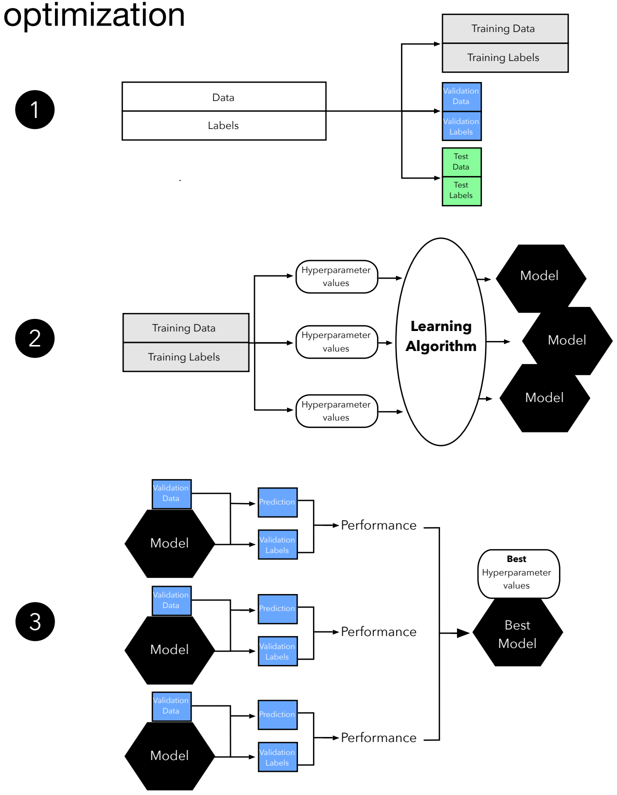
BIG
datasets

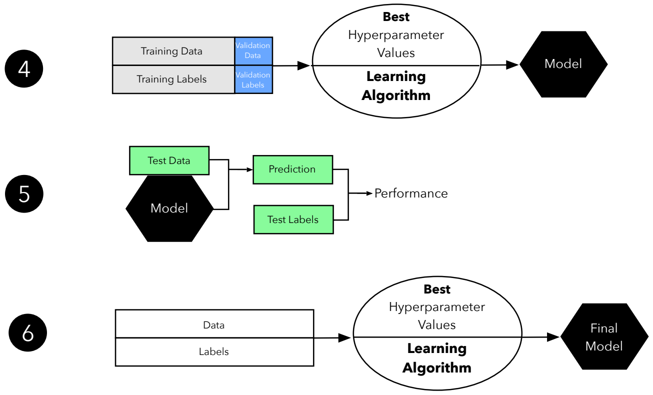


SIMPLE
models

3-Way Holdout

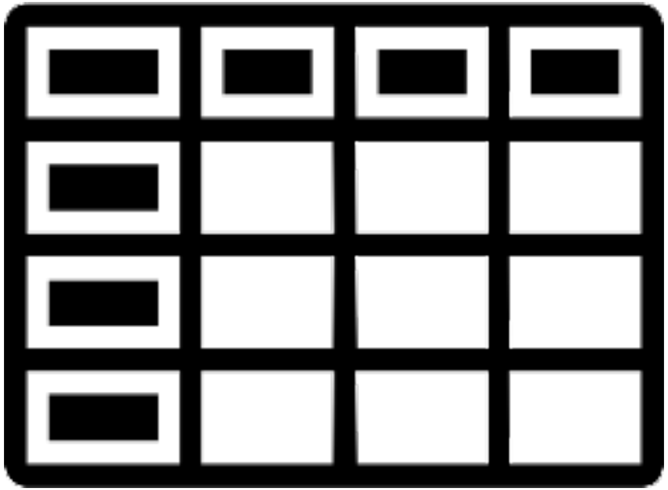
instead of "regular" holdout to avoid "data leakage" during hyperparameter optimization



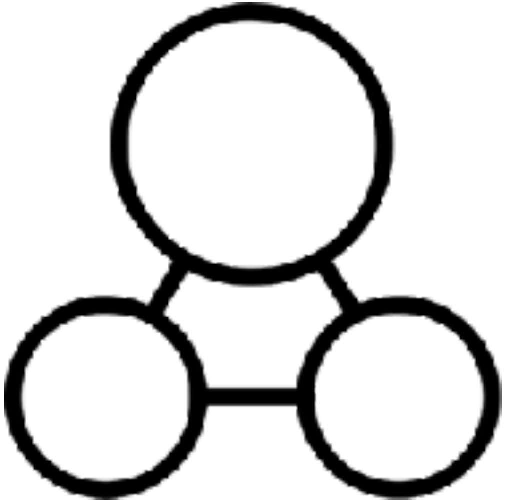


Model, feature, and hyperparameter selection

- 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



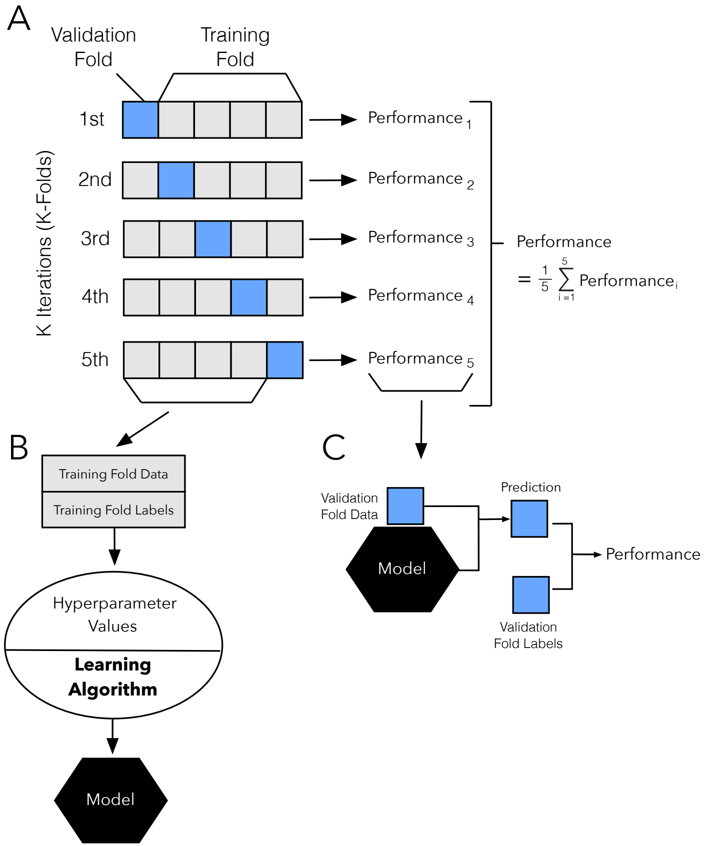
SMALLER
datasets



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 hi variability



model assessment

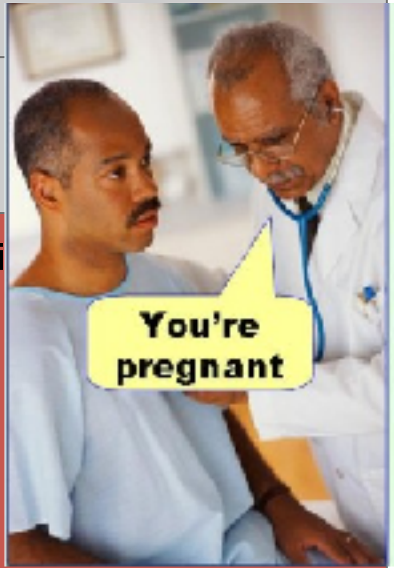
Root Mean Squared Error (RMSE)

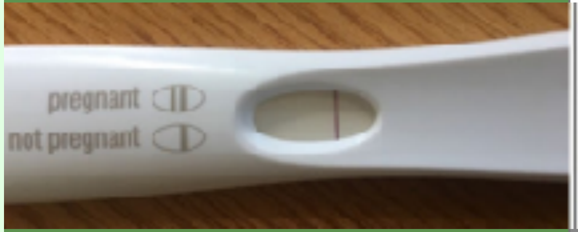
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

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

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

		Actual	
		Positive	
Predicted	Positive	True Positive (TP)	False Positi
	Negative	False Negativ	True Negative (TN)





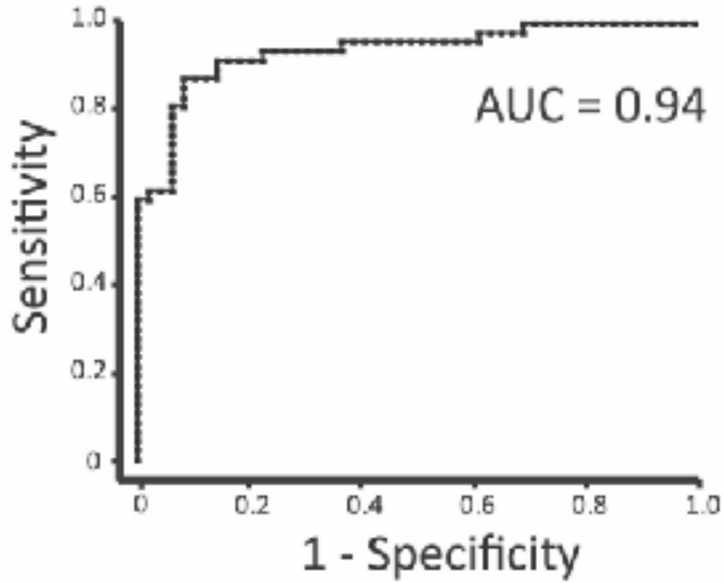
categorical variable prediction

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

A 2x2 table is a type of confusion matrix

Sensitivity

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



Specificity

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

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

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?

A
regression
(supervised)

B
classification
(supervised)

C
clustering
(unsupervised
)

D
dimensionality
reduction
(unsupervised)

Prediction Approach

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

A
RMSE

B
Accuracy

C
Sensitivity

D
Specificity

E
AUC

Prediction Approach

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