BST 261: Data Science II
Lecture 1
Course Introduction, Deep Learning Background and Machine Learning Review

Heather Mattie
Harvard T.H. Chan School of Public Health
Spring 2 2019

Teaching Staff

Instructor

Heather Mattie

Instructor of Data Science

hemattie@hsph.harvard.edu

Building 1 Room 421A

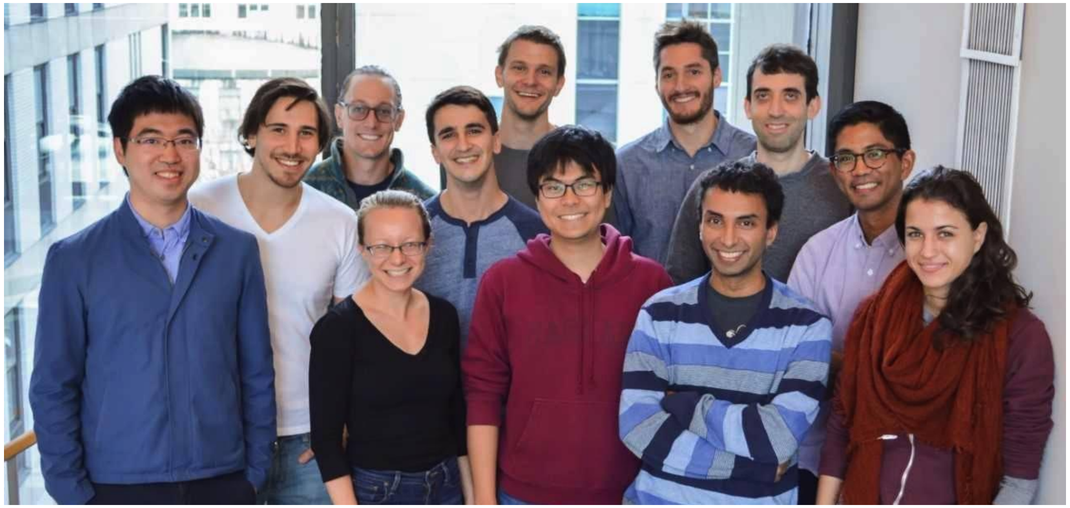
(617) 432-5308

Office Hours: Wed 1-3pm

Teaching Staff - TAs

Matt Ploenzke
Biostatistics PhD Candidate
ploenzke@g.harvard.edu
Office Hour: Fri 11:15-12:15pm
except 3/29 and 5/17
Kresge 202A
Labs: Fri 9:45-11:15am
in LL6

Aaron Sonabend
Biostatistics PhD Candidate
asonabend@g.harvard.edu
Office Hour TBD
Room TBD



2018 Biostatistics Certificate of Distinction in Teaching Awardees

Teaching Staff - TAs

Matt Ploenzke
Biostatistics PhD Candidate
ploenzke@g.harvard.edu
Office Hour: Fri 11:15-12:15pm
except 3/29 and 5/17
Kresge 202A
Labs: Fri 9:45-11:15am
in LL6

Aaron Sonabend
Biostatistics PhD Candidate
asonabend@g.harvard.edu
Office Hour TBD
Room TBD

2018 Biostatistics Certificate of Distinction in Teaching Awardees

Course Overview

Course content:

- Computational and mathematical foundations of deep learning
- Deep learning workflow
- Bias/variance trade off
- Feedforward networks (Multilayer perceptrons - MLPs)
- Convolutional neural networks (CNNs)
- Recursive/Recurrent neural networks (RNNs)
- Deep learning research
- Advanced topics in deep learning (GANs, VAEs, adversarial attacks)
- Cloud computing with Google Cloud Platform (GCP)

- A lot of content will come from: Deep Learning. Ian Goodfellow, Yoshua Bengio, Aaron Courville. MIT Press, 2016. Book available at http://www.deeplearningbook.org
- As well as: Deep Learning with Python. François Chollet. Manning Publications, 2017.
- The first few chapters are available at https://www.manning.com/books/deep-learning-with-python

Course Overview

Lectures:

- Total of 16 lectures and 6 lab sessions; see syllabus for schedule
- Lectures on Mondays & Wednesdays from 9:45 - 11:15am in Kresge G2
- Slides will be available on the course website and GitHub repo before each lecture
- Lectures will be a mixture of theory and application
- All in-class coding examples will be in Python and use Keras with TensorFlow backend
- Pytorch will be introduced in the labs
- Labs will be held on Fridays 9:45 - 11:15am in LL6 (sorry - it was the only room available!)

Course Overview

Homework assignments:

- Homework 1 (25% of grade) due Monday **April 8** by 11:59pm
- Homework 2 (25% of grade) due Monday **April 22** by 11:59pm
- Homework 3 (25% of grade) due Monday **May 10** by 11:59pm

- Output is a Jupyter notebook consisting of code and text
- A notebook template will be available on the course Canvas and GitHub site
- All assignments should be submitted on Canvas

Course Overview

Group project proposal:

- Larger assignment that brings together different course themes
- Output is a Word doc or pdf file that should be submitted on Canvas
- Only one file needs to be submitted per group
- Groups may contain 1 - 4 students
- Due **May 15** by 11:59pm, 25% of grade
- See syllabus for details

Deep Learning

What is *Deep Learning*?

- Computers can solve problems that are intellectually difficult for humans
  - Ex: multiplication with large numbers and decimals, chess, etc.
  - Problems that can be described by a list of formal, mathematical rules

- Humans can solve problems that are intuitive, but difficult to describe formally and thus difficult for computers to solve
  - Ex: handwriting recognition, speech recognition, etc.

- Deep learning is a solution to allow computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts

- The hierarchy of concepts builds on itself, producing a deep graph with many layers, leading to the concept of ***deep learning***

What is *Deep Learning*?

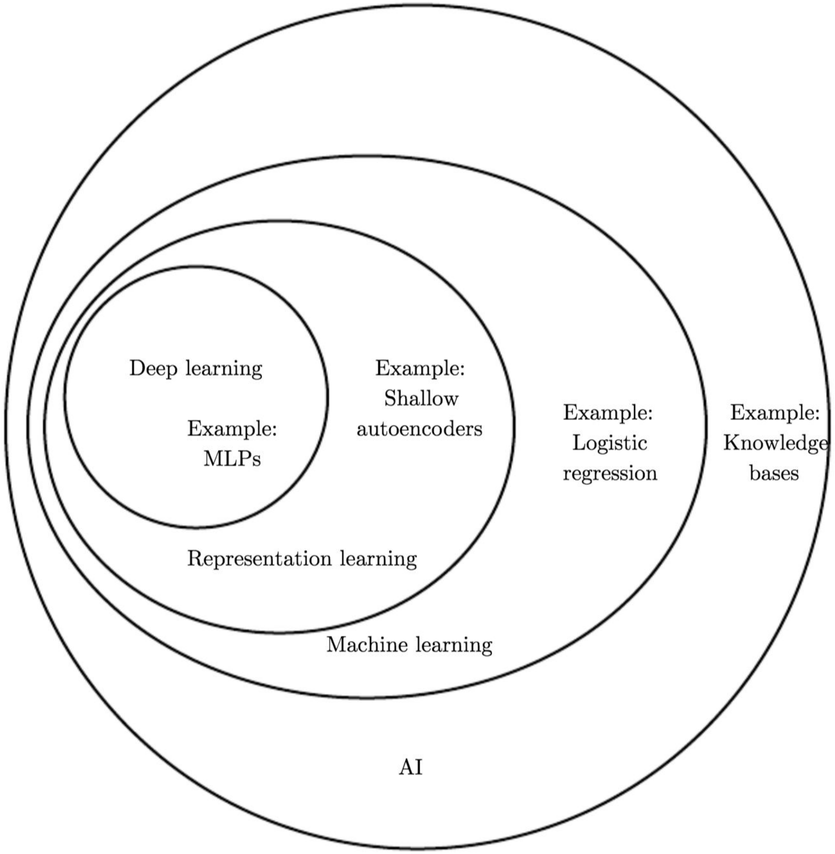
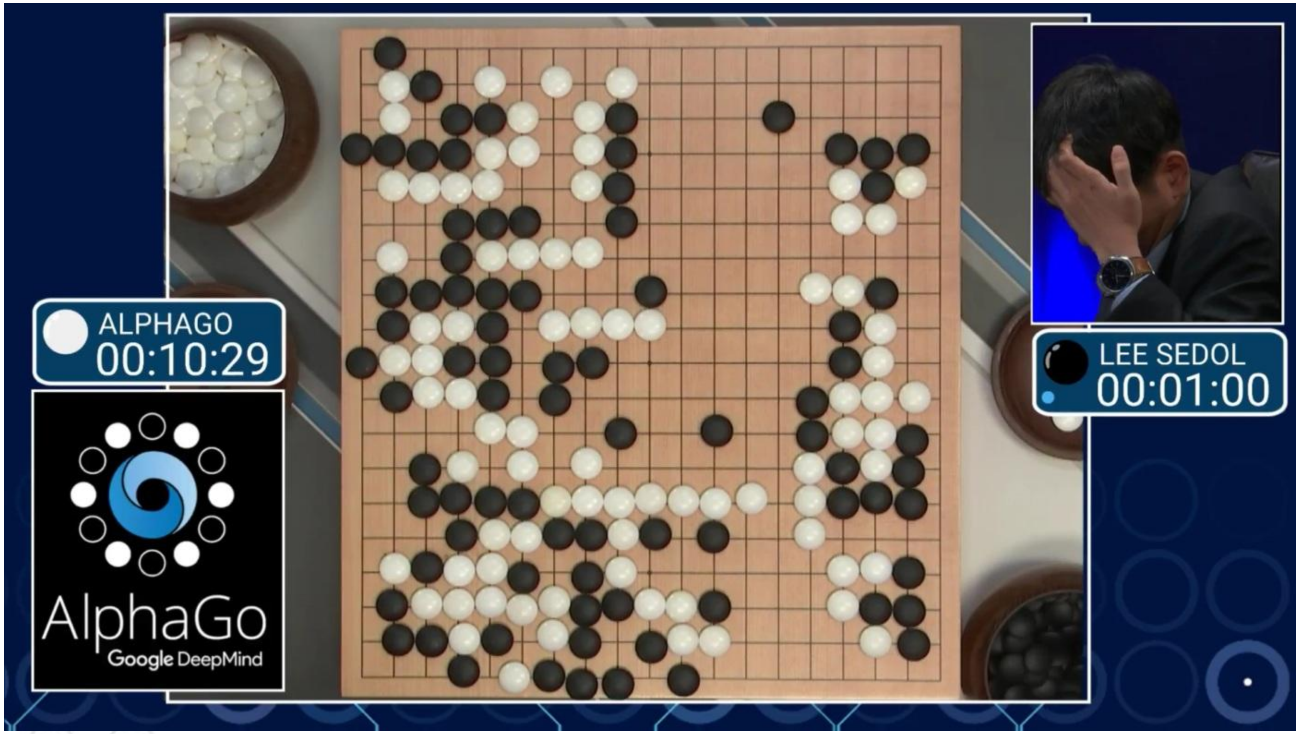


Figure 1.4: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.

http://www.deeplearningbook.org/

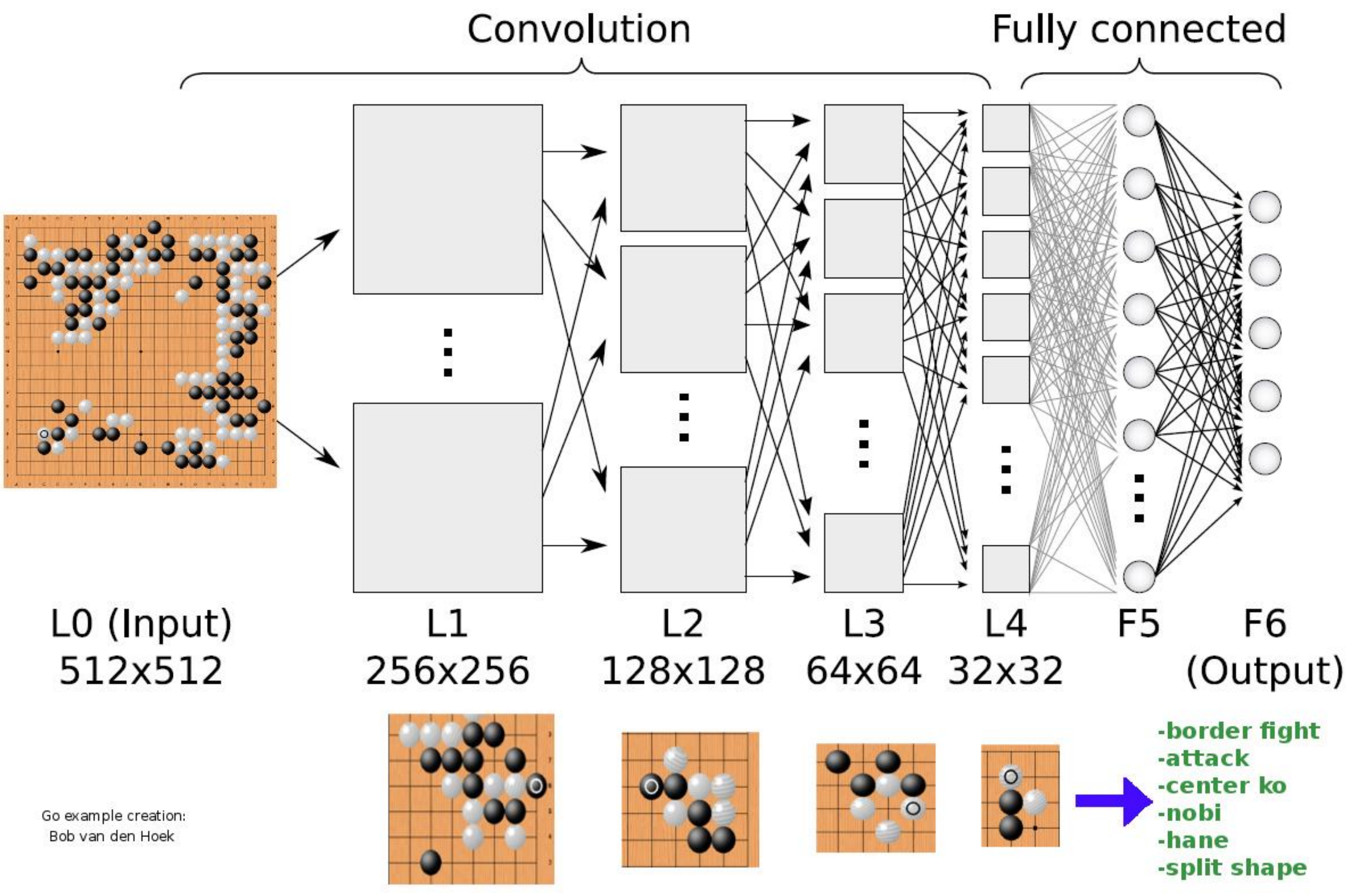
Deep Learning Successes

Games



Deep Learning Successes

Games



Go example creation:
Bob van den Hoek

Deep Learning Successes

Art

https://blog.udacity.com/2018/04/how-to-process-images-with-tensorflow.html

Deep Learning Successes

Art

(a) ***Input****: a casual face photo* (b) ***Outputs****: new headshots with the styles transferred from the examples. The insets show the examples.*

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand, Style Transfer for Headshot Portraits, to appear in *SIGGRAPH 2014*

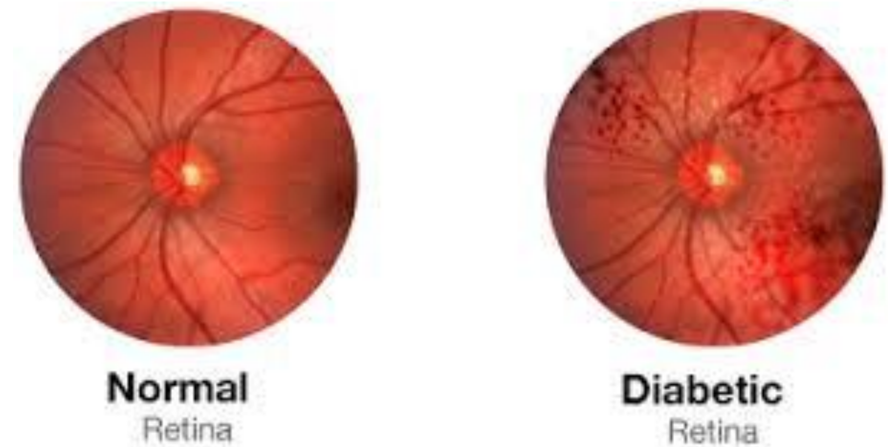
Deep Learning Successes

[The AI-Art Gold Rush]

AI-generated "faceless portraits" by Ahmed Elgammal and AICAN. Photo: Artrendex Inc./The Atlantic

Deep Learning Successes

Medicine



JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

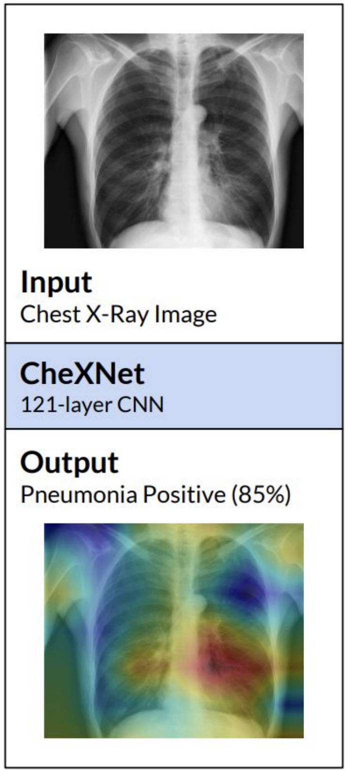
Deep Learning Successes

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar [*1] Jeremy Irvin [*1] Kaylie Zhu [1] Brandon Yang [1] Hershel Mehta [1]
Tony Duan [1] Daisy Ding [1] Aarti Bagul [1] Robyn L. Ball [2] Curtis Langlotz [3] Katie Shpanskaya [3]
Matthew P. Lungren [3] Andrew Y. Ng [1]

Abstract

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Our algorithm, CheXNet, is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.



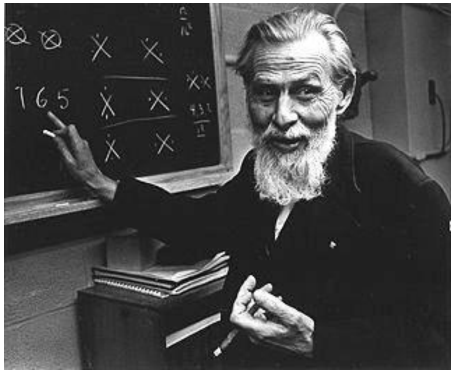
1. Introduction

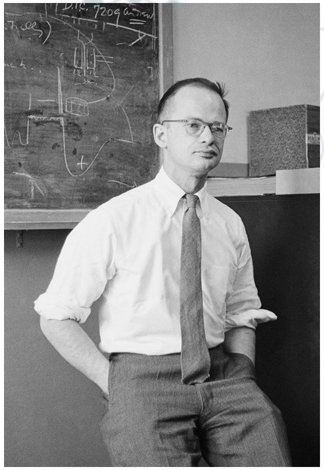
More than 1 million adults are hospitalized with pneumonia and around 50,000 die from the disease every year in the US alone (CDC, 2017). Chest X-rays

History of Neural Nets and Deep Learning

Neural Nets and Deep Learning Not New

- Date back to the 1940s
- Walter Pitts and McCulloch
  - First notion of an artificial neuron
  - Designed to mimic the way a neuron was thought to work
  - 1943 paper



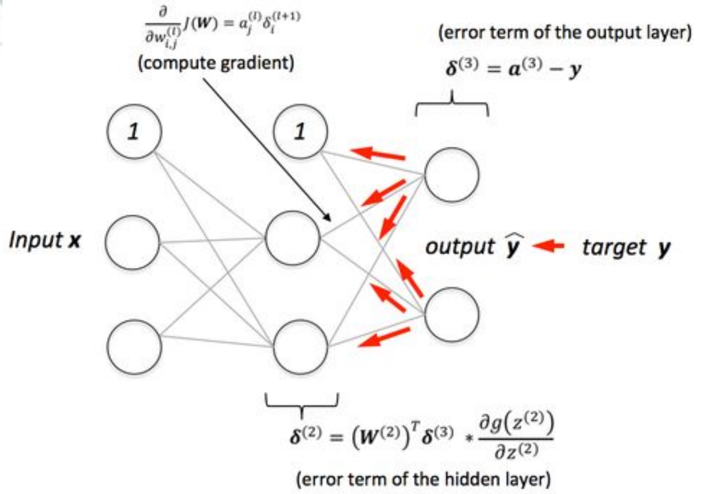


- Frank Rosenblatt
  - “Perceptron” algorithm 1950s
  - Could recognize letters and numbers

AI Winter

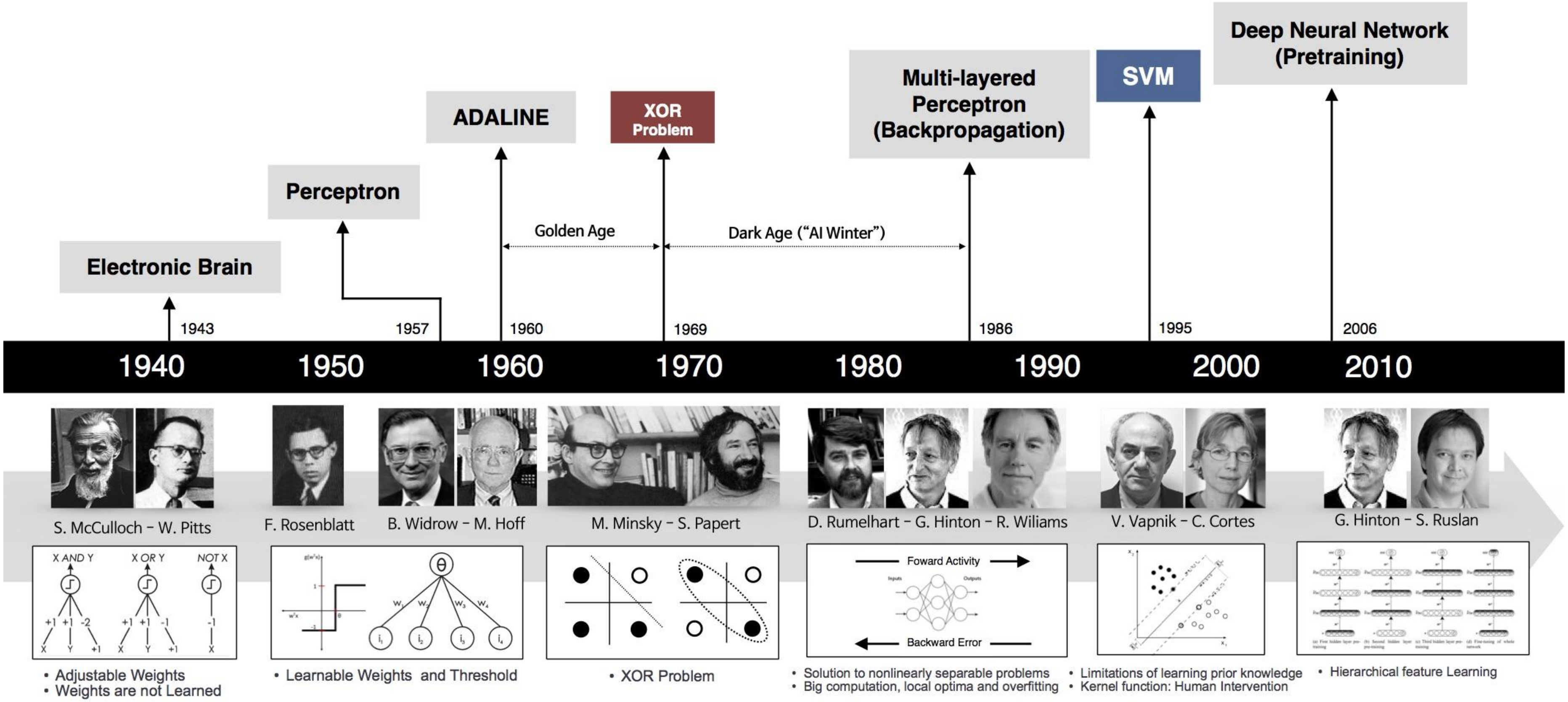
- Many cycles of boom and bust
- Repeated promises of “true AI” that were unfulfilled and followed by “AI winters” - the first in 1969
- Marvin Minsky and Seymor Papert write book about shortcomings of perceptrons and effectively kill all research on neural nets

Return of the Neural Net



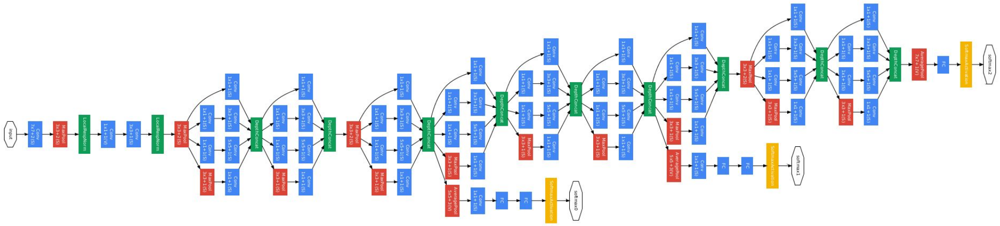
- Geoff Hinton, David Rumelhart and Ronald Williams discover back-propagation (1980s)
  - Allows neural nets to move past the limitations of perceptrons
  - Lead to convolutional neural nets (CNNs) and handwritten digits recognition
  - Problem: didn't scale ⟶ another 10-15 year AI winter
- Rebranding as "Deep Learning" (2006)
  - Unsupervised pretraining and deep belief networks
  - Could create "deeper" neural nets ⟶ "deep" learning
- Great AI Awakening (where we are now!)
  - Alexnet (2012)
  - Availability of GPUs (and TPUs) and larger data sets
  - Neural nets start surpassing humans

Deep Learning Timeline



AI winters are probably over

- We now have large, high-quality, labeled data sets
- GPUs and TPUs abound
  - Allows for deeper models and an increase in accuracy
- Improved functions needed for learning
  - ReLU
  - tanh
- Improved architectures
  - Resnets
  - Inception modules
- New regularization techniques
  - Dropout
  - Batch normalization
- Robust optimizers
- Software platforms
  - Tensorflow
  - Theano



How to stay current

- Advances in deep learning, and AI in general, are happening every day - it isn’t possible to keep track of everything, but below are some good sources to check out
- Read papers on arXiv
- Subscribe to Medium
- Google AI Blog
- Keras Blog
- OpenAI Blog
- Twitter
  - Follow deep learning gods like Ian Goodfellow, Yann Lecun, Fei-Fei Li, Francois Chollet and our own HMS professor Andrew Beam
- Talking machines podcast

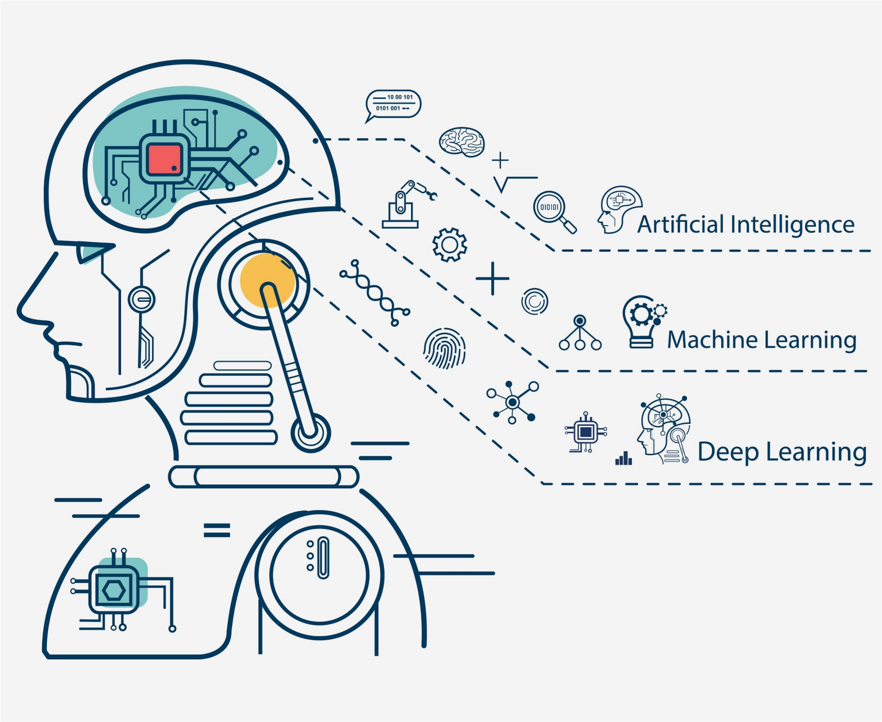
Takeaways

- Deep learning is real and probably here to stay
- Could potentially impact many fields → understand concepts so you have deep learning "insurance"
- Long history and connections to other models and fields
- Prereqs: Data (lots) + GPUs (more = better)
- Deep learning models are like legos, but you need to know what blocks you have and how they fit together
- Need to have a sense of sensible default parameter values to get started
- "Babysitting" the learning process is a skill

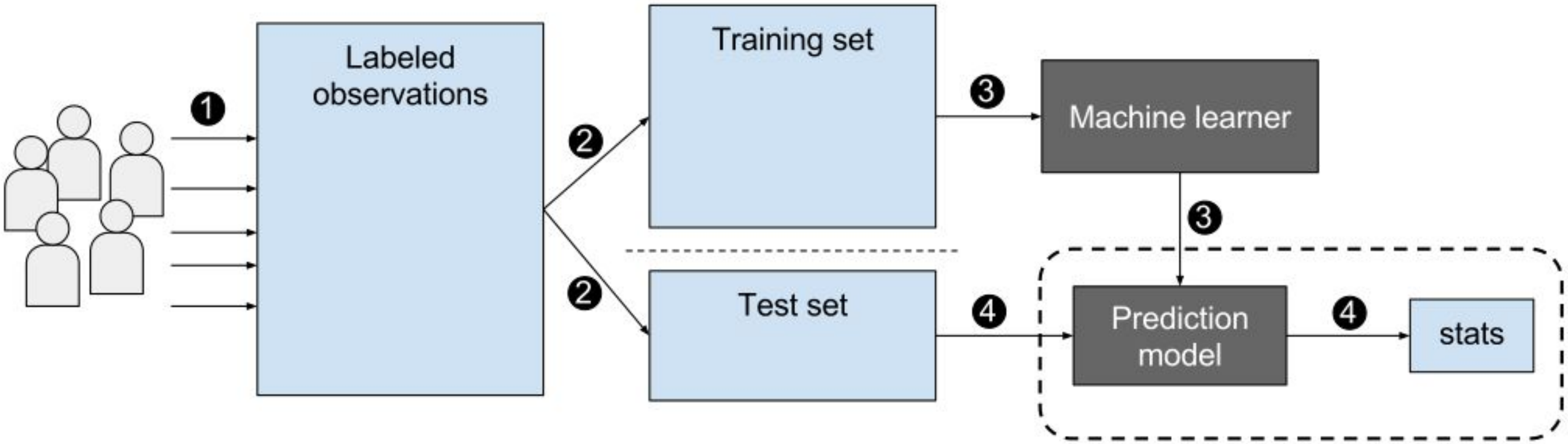
Machine Learning Review

Machine Learning

- Gives machines the ability to learn without being explicitly programmed
- Why is it useful?
  - Some tasks are difficult to program
- Types
  - Supervised
  - Unsupervised
  - Semi-supervised
  - Reinforcement learning



Supervised ML



https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

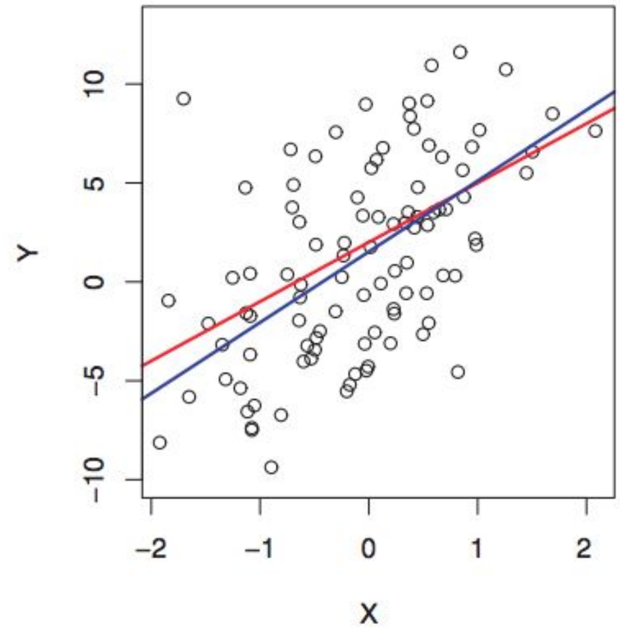
Regression and Classification

- Regression
  - Output: continuous variable
  - Ex: serum cholesterol level, gene expression levels, etc.
- Classification
  - Output: discrete variable (categories = “classes”)
  - Ex: disease types, genotypes

Supervised Learning Algorithms

Linear Regression

- Outcome (Y) is continuous
- Fitting a line to a cloud of data
- Using ordinary least squares (OLS) to find the best line
- Features (predictors) can be continuous, categorical, binary



Dependent Variable

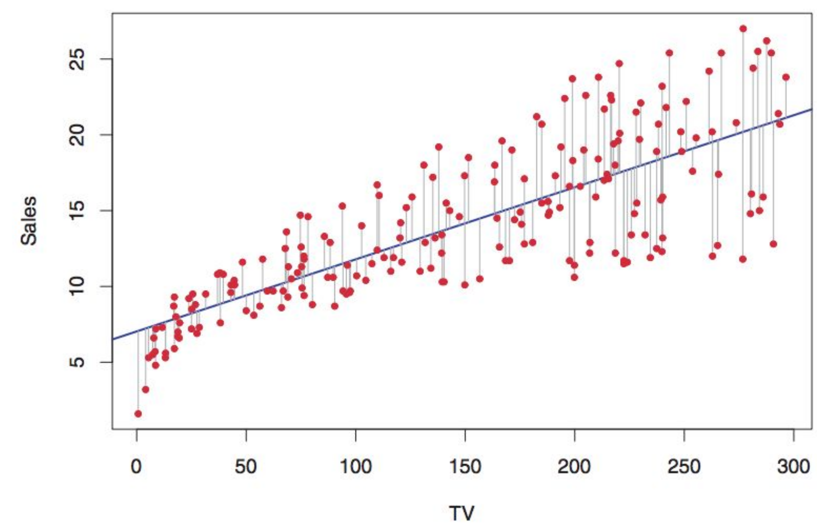
Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

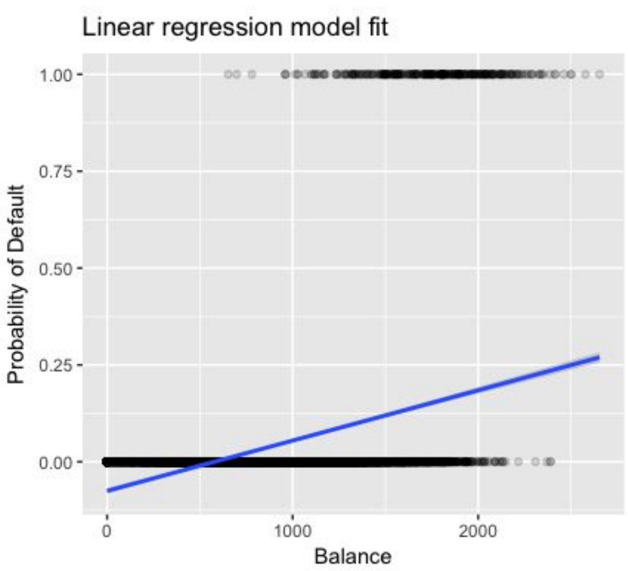


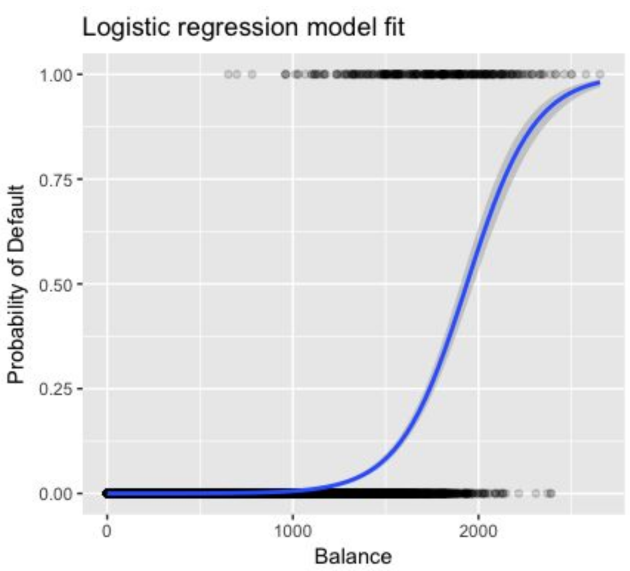
Logistic Regression

$$p(X) = \Pr(Y = 1|X)$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

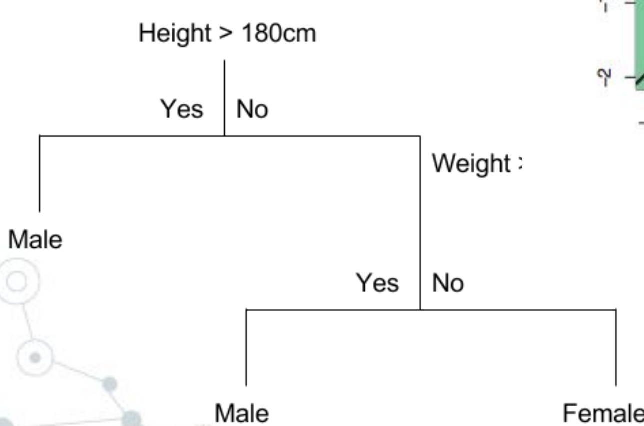
- Outcome (Y) is **binary**
- Linear regression no longer appropriate
- Need to transform the equation we saw for linear regression
- Are now predicting the **probability** of someone being classified as a 1
- Since these are probabilities, you can control the threshold of classifying someone as a 1

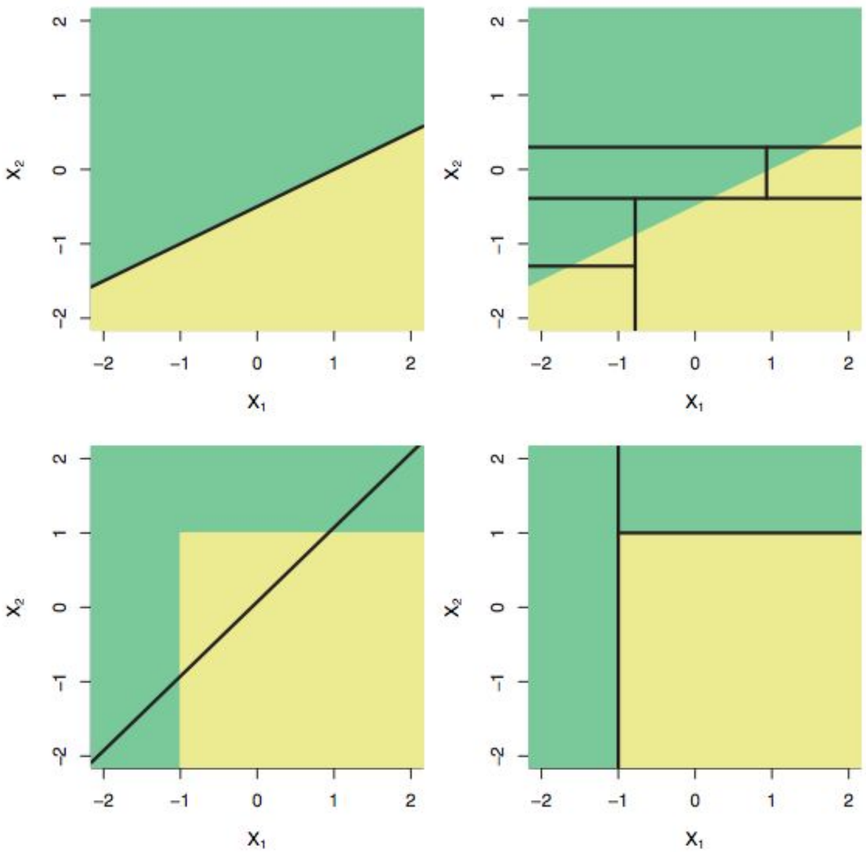




Decision Trees

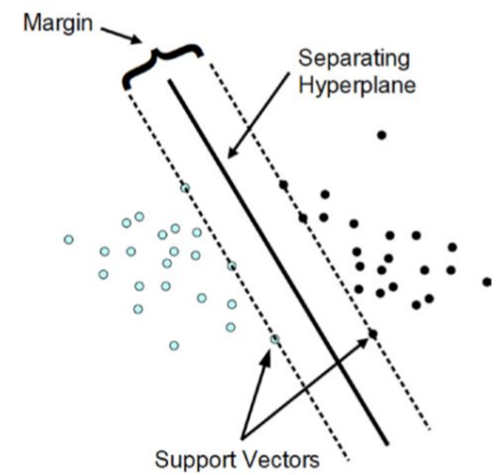
- Outcome (Y) is discrete or continuous
- Drawing a series of boundaries using the features (predictors, Xs)
- On the right: the colors are the truth. Suppose green means you are a 0 and yellow means you are a 1
- Classification trees find the best lines (boundaries) to split the predictions
- Perform best when the truth is linear

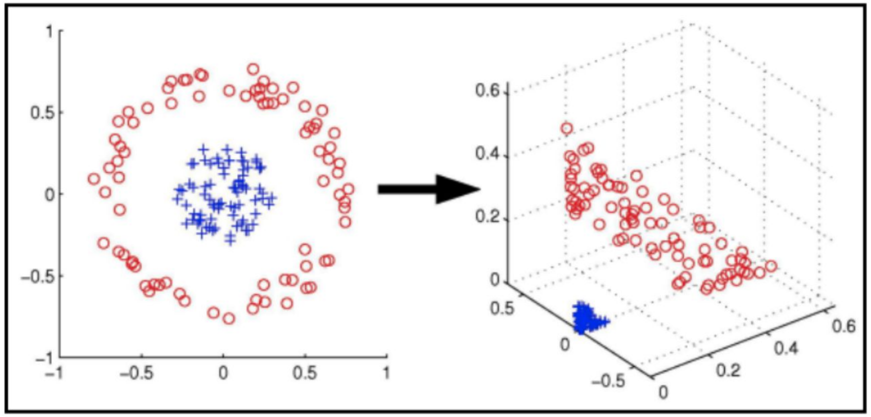




Support Vector Machines (SVM)

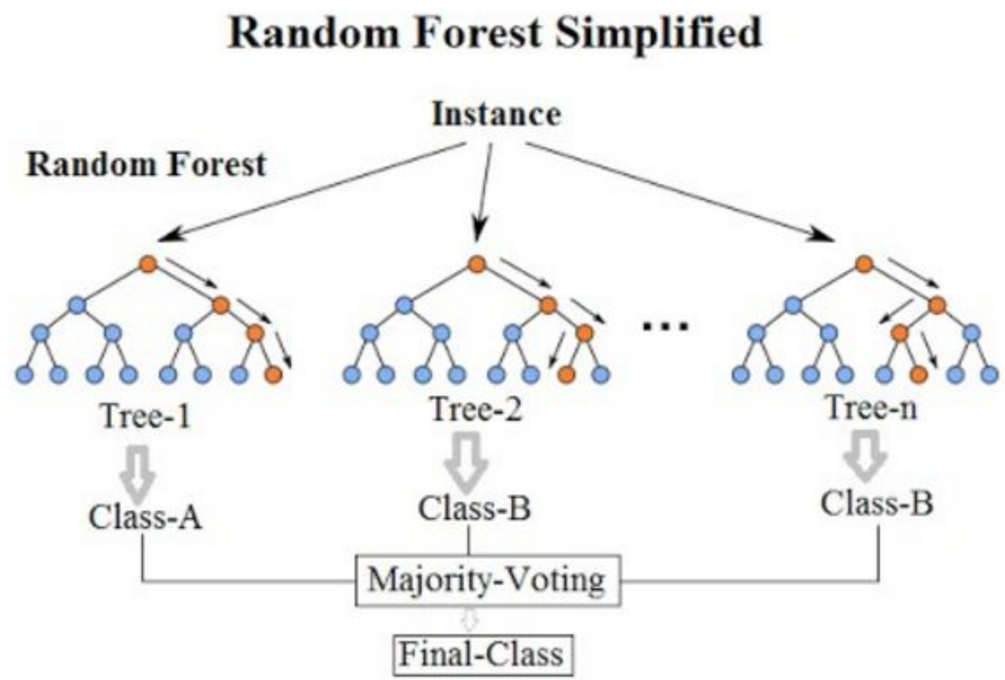
- Identifies a hyperplane that has the largest distance to the nearest training data point of any class
- Uses this hyperplane to separate the classes and classify new data





Random Forests

- An ensemble method
- Generate multiple decision trees
- Aggregate the votes from the trees
  - Majority for classification
  - Mean for regression

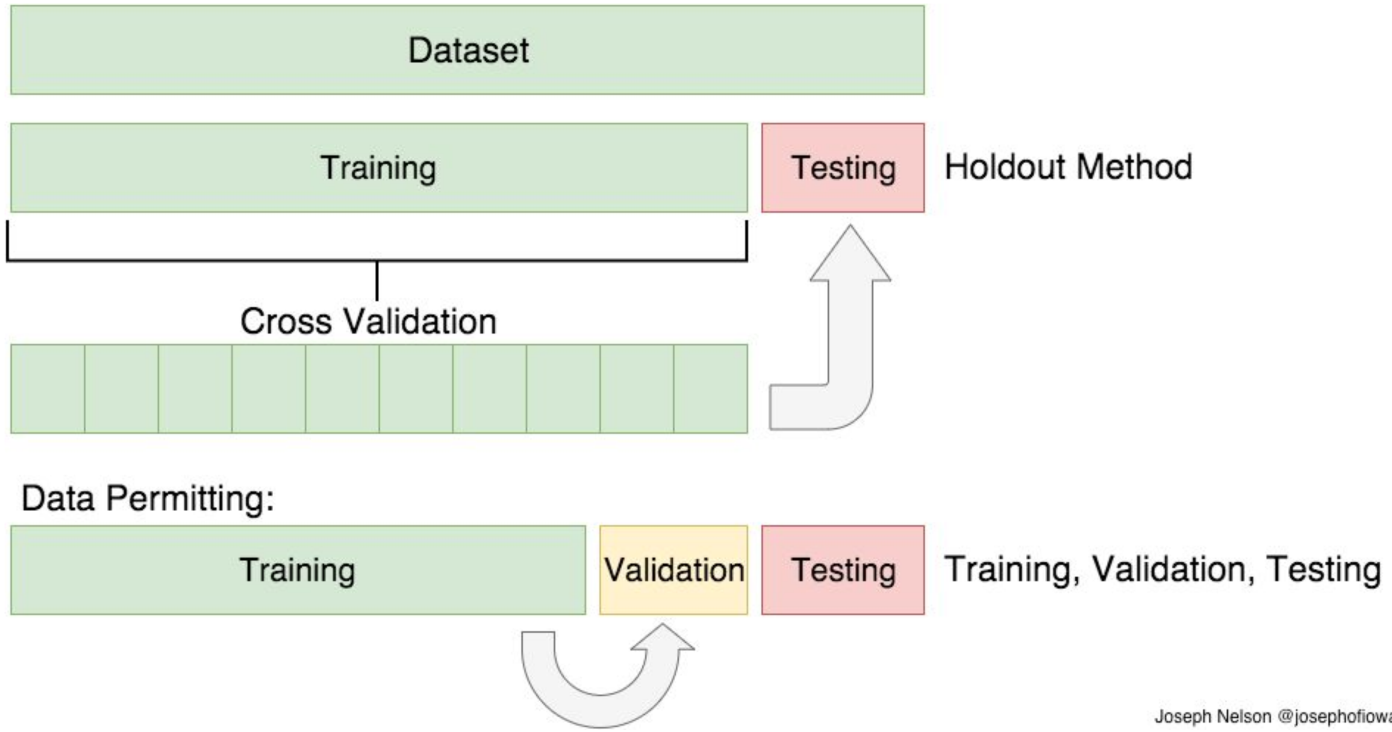


Model Evaluation

Evaluating ML Models

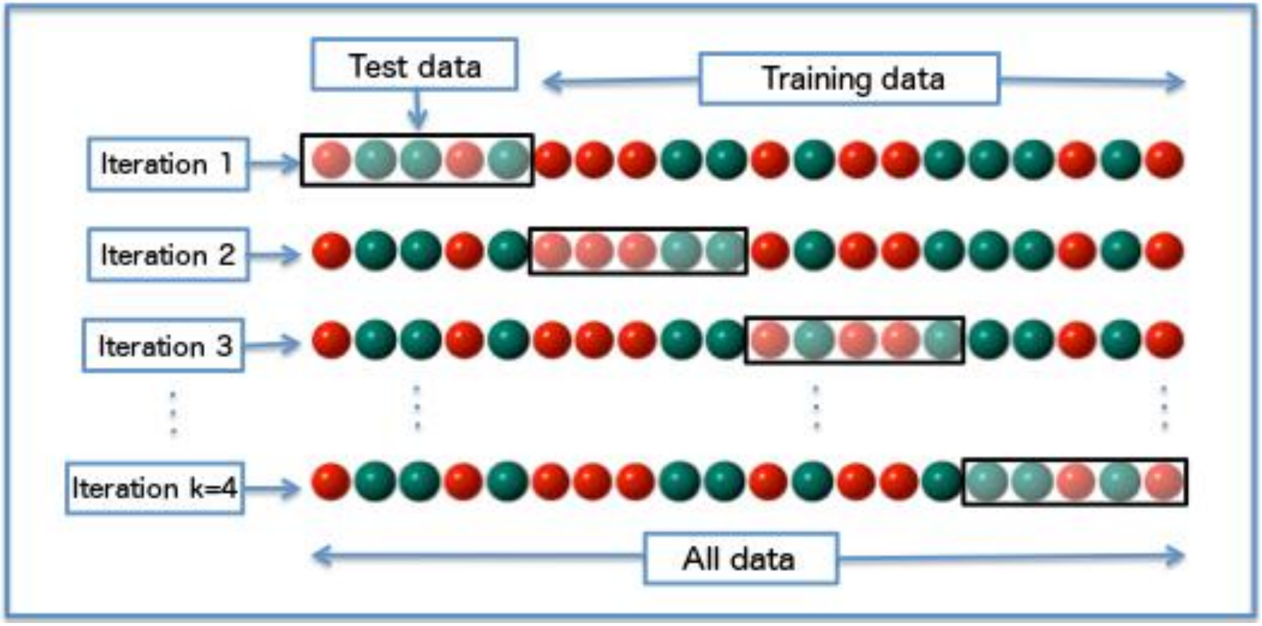
- How do we know what "good" is?
- It depends
  - Overfitting?
  - Underfitting?
  - Class imbalance?
  - Trivial task?

Leave One Out Cross-Validation



https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

K-Fold Cross-Validation

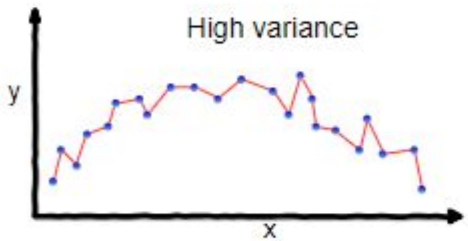


Wikipedia contributors. (2019, March 22). Cross-validation (statistics). In *Wikipedia*, *The Free Encyclopedia*.

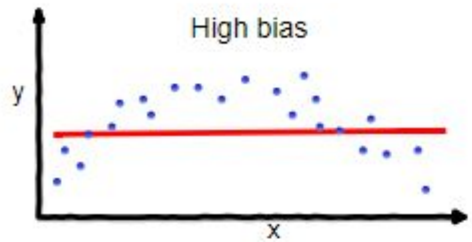
Bias-Variance Tradeoff

Bias-Variance Tradeoff

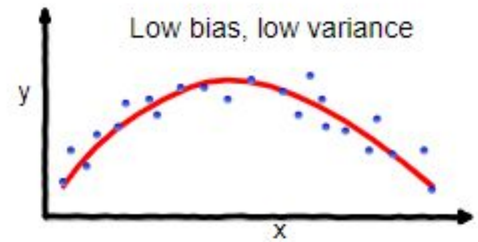
- Bias
  - The expected generalization error - even if fitting to a very large data set
  - Think “underfitting”
- Variance
  - Error due to hypersensitivity to small fluctuations in the training set
  - Think “overfitting”



overfitting



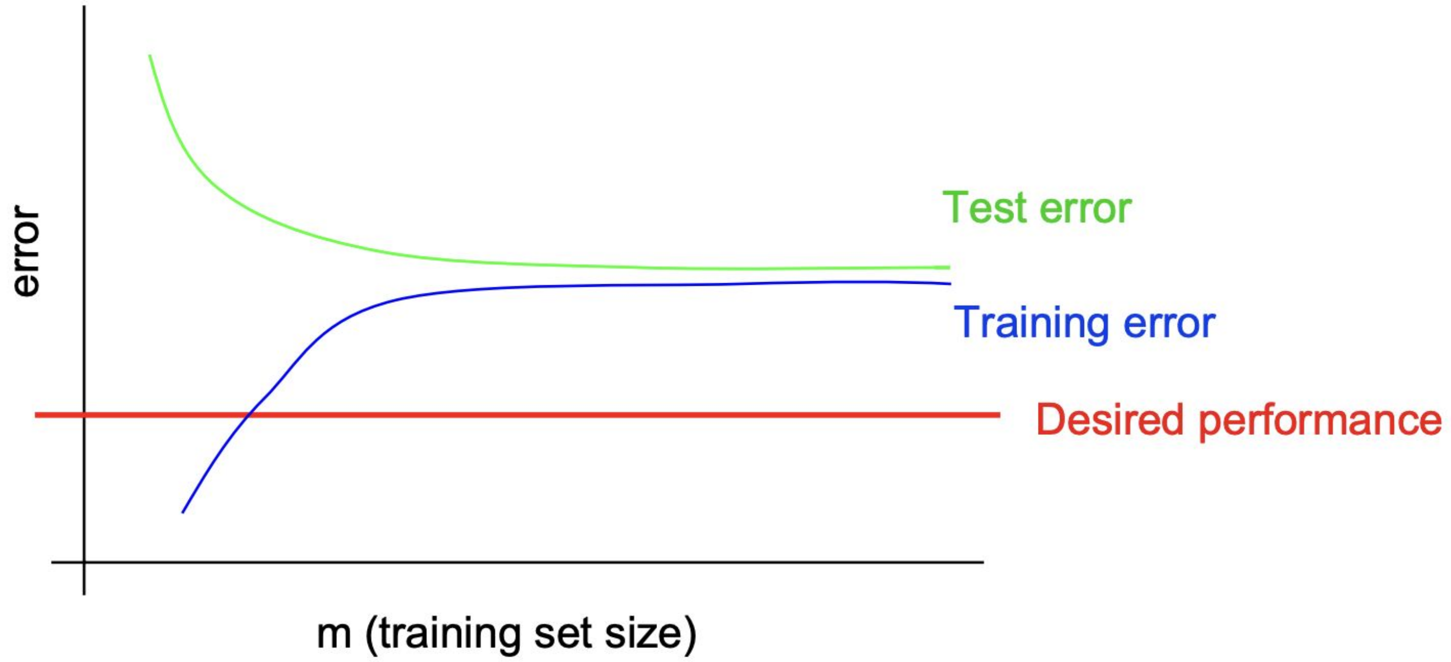
underfitting



Good balance

High Bias

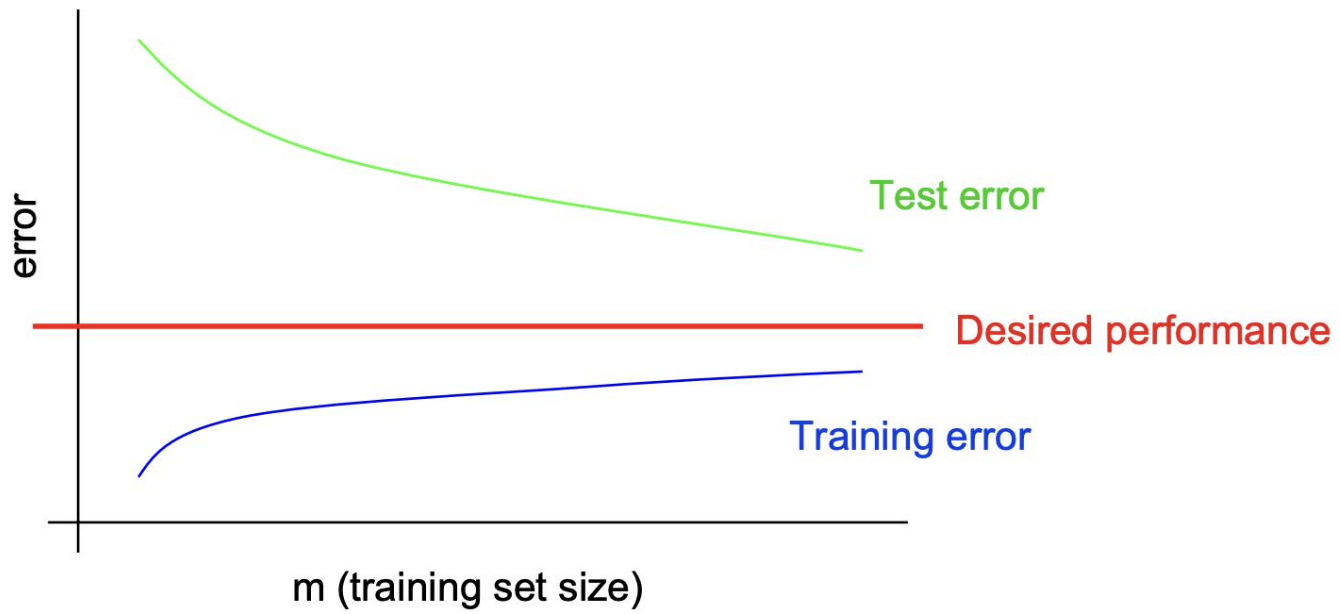
Typical learning curve for high bias:



Andrew Ng's ML advice Notes: http://cs229.stanford.edu/materials/ML-advice.pdf

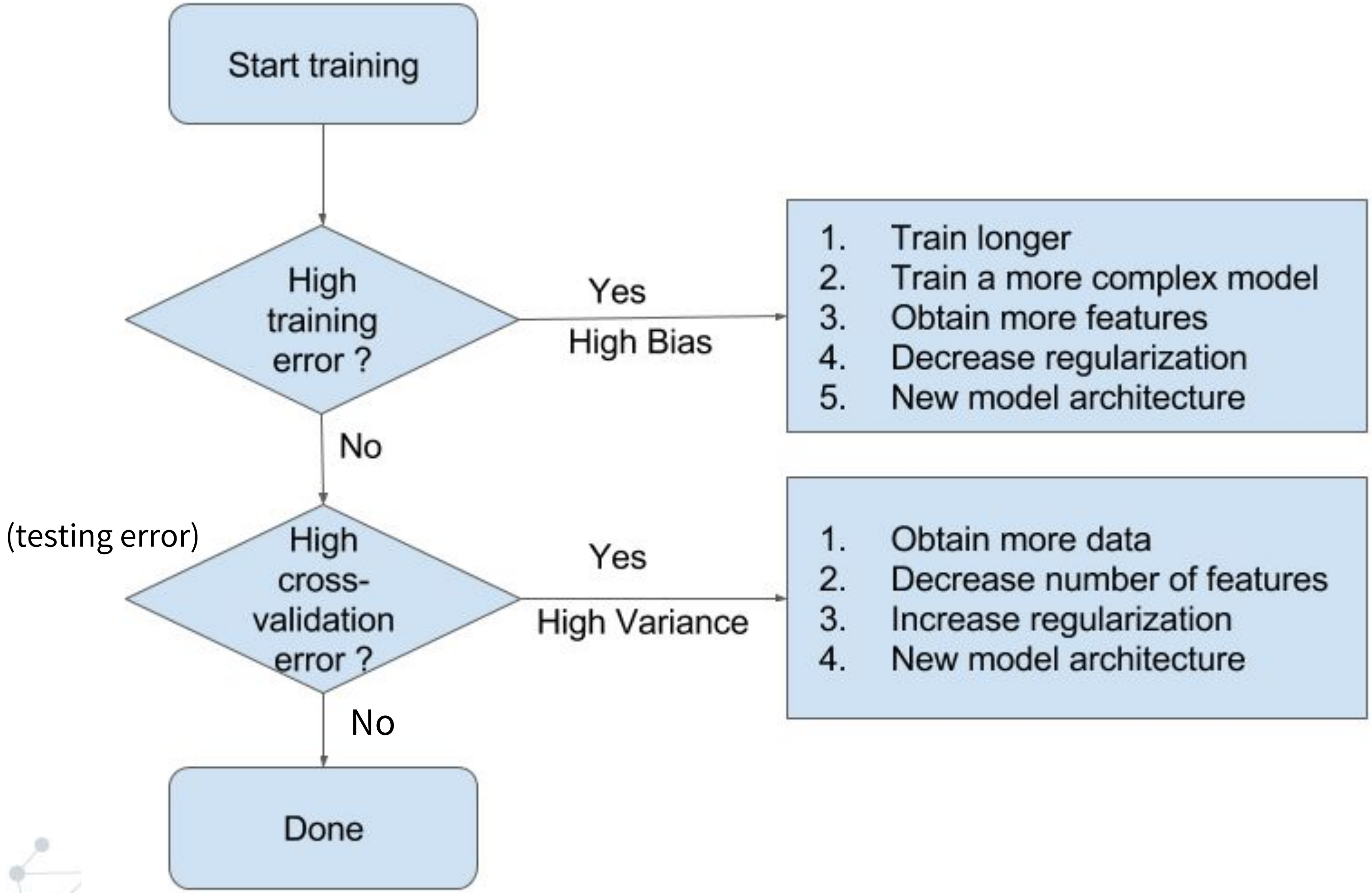
High Variance

Typical learning curve for high variance:



Andrew Ng's ML Advice Notes: http://cs229.stanford.edu/materials/ML-advice.pdf

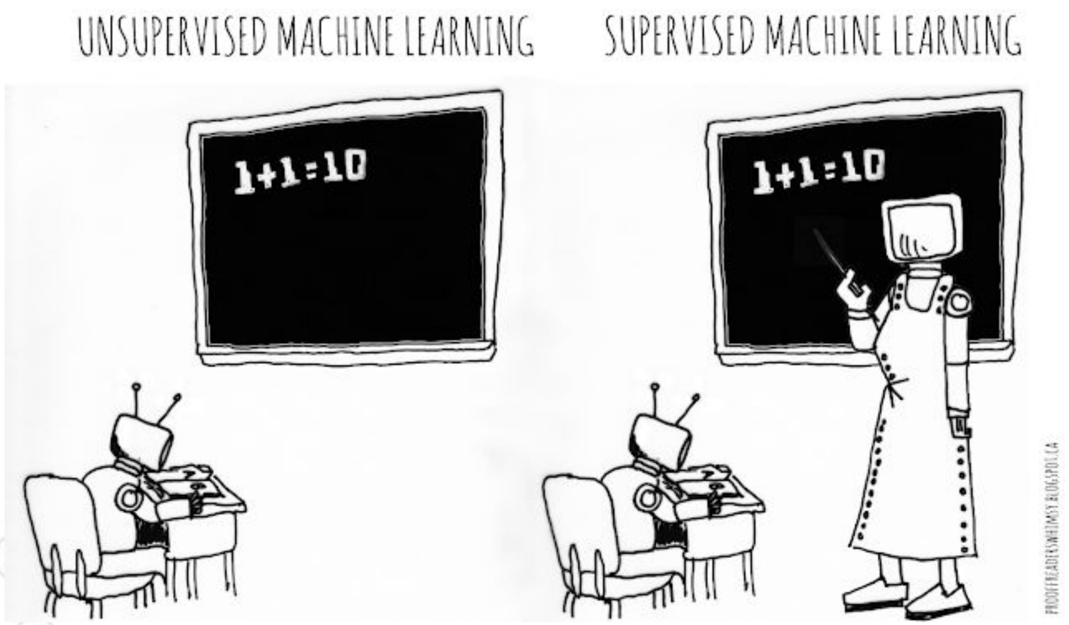
Addressing High Bias and Variance



Unsupervised Learning Algorithms

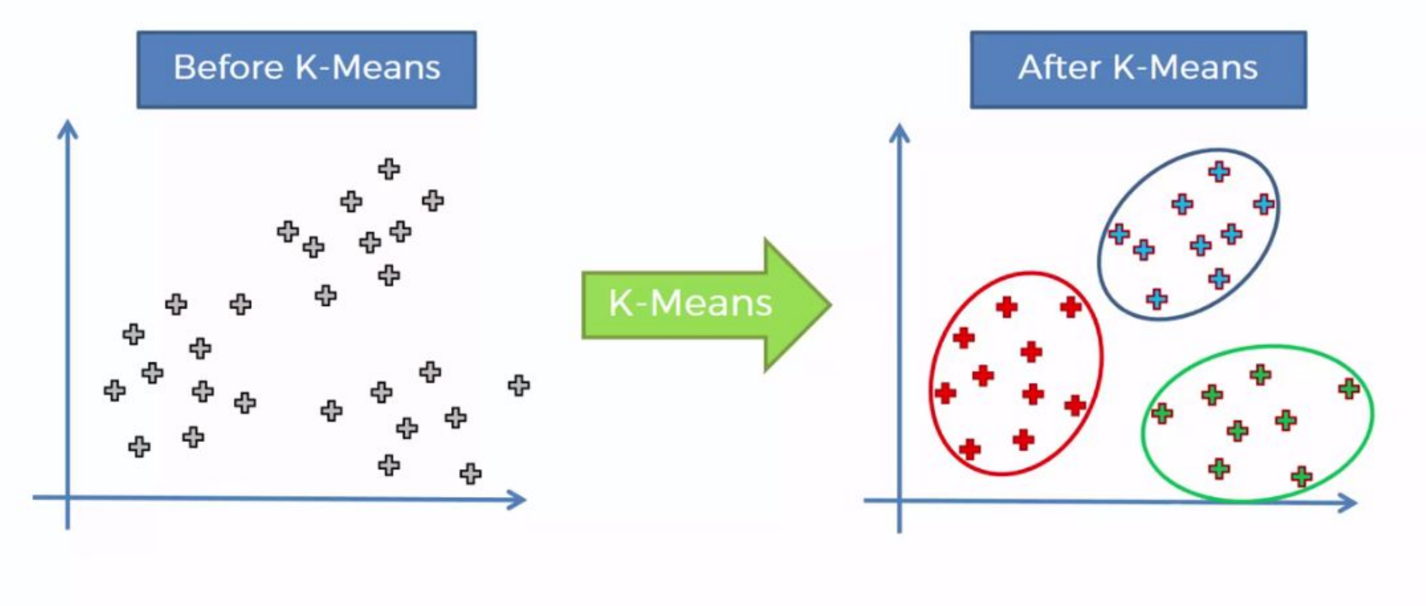
Unsupervised ML

- Training examples do not have ground truth labels
- Model identifies structure, such as clusters
- New data is assigned to clusters



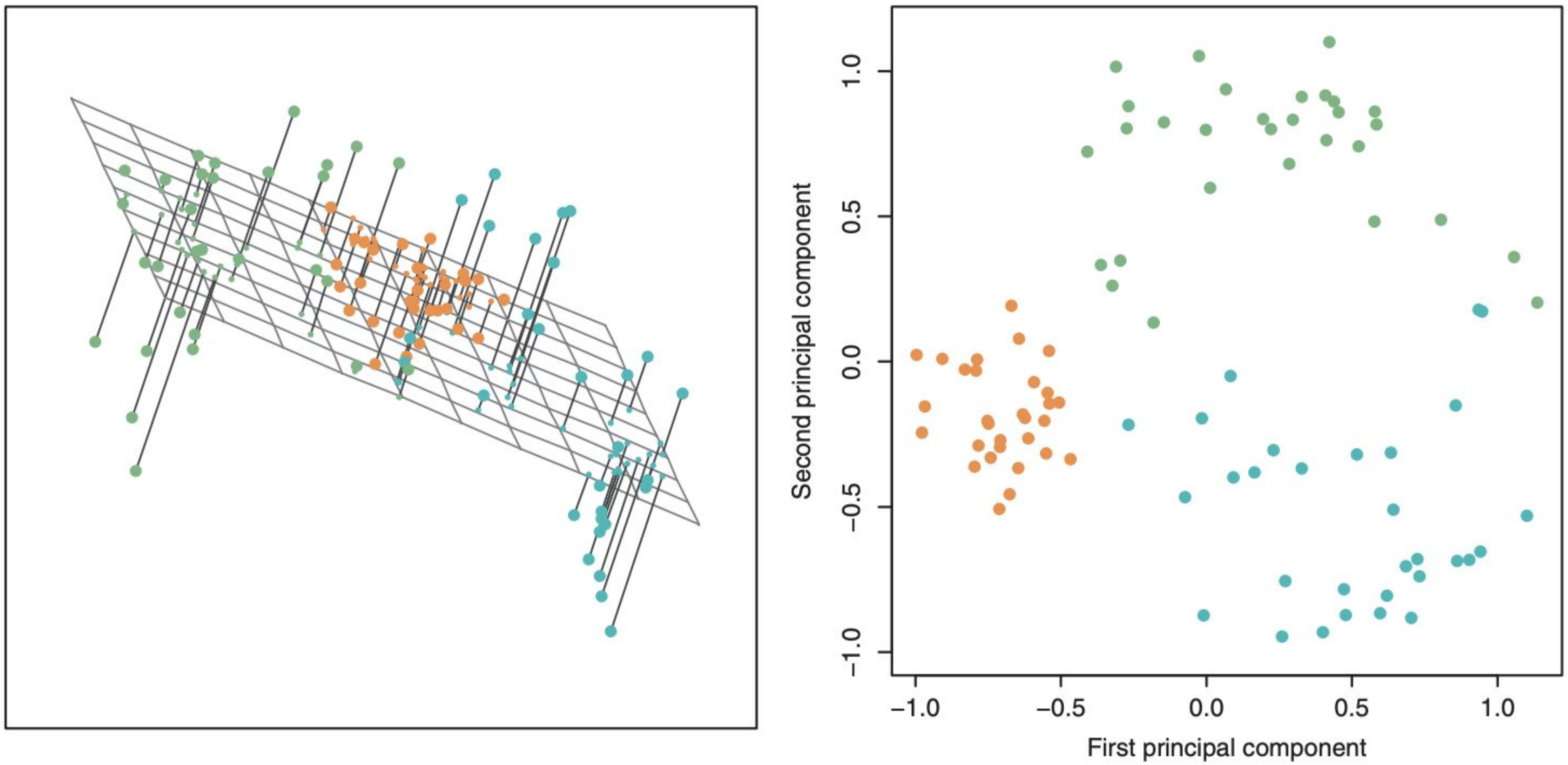
Clustering

- Ex: k-means clustering



https://towardsdatascience.com/k-means-clustering-identifying-f-r-i-e-n-d-s-in-the-world-of-strangers-695537505d

Principal Components Analysis (PCA)



Introduction to Statistical Learning, Springer 2017

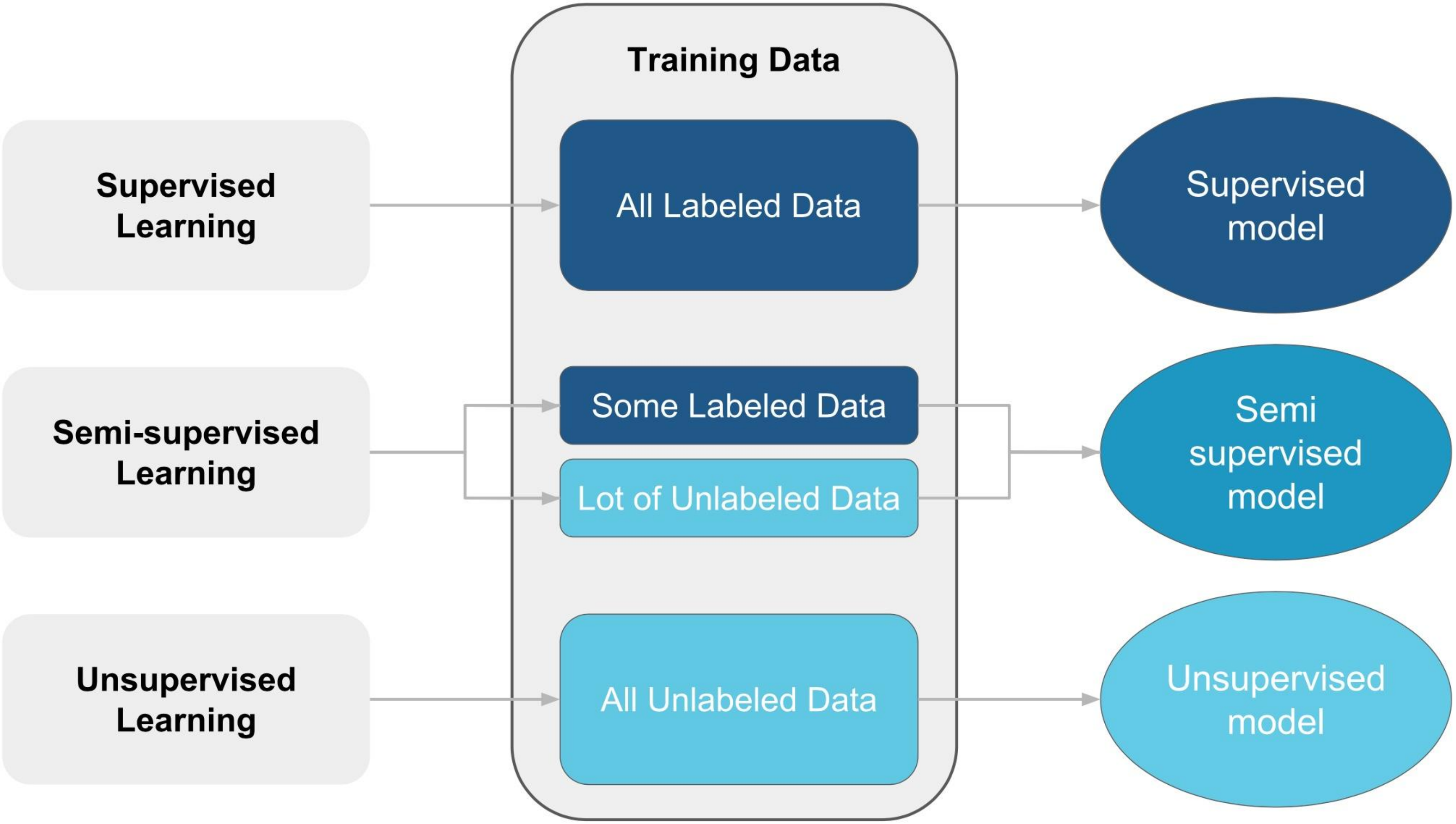
Semi-Supervised Machine Learning

Semi-Supervised Machine Learning

- Supervised machine learning tasks that also utilize unlabeled data for training
  - Ex: performing dimension reduction with PCA as a preprocessing step before applying a classifier

- Some assumptions are needed for the data distribution
  - Continuity
    - Points which are close to each other are more likely to share a label
  - Clustering
    - The data tend to form discrete clusters, and points in the same cluster are more likely to share a label
  - Manifold (topological space)
    - The data lie approximately on a manifold of much lower dimension than the input space (avoiding the curse of dimensionality)

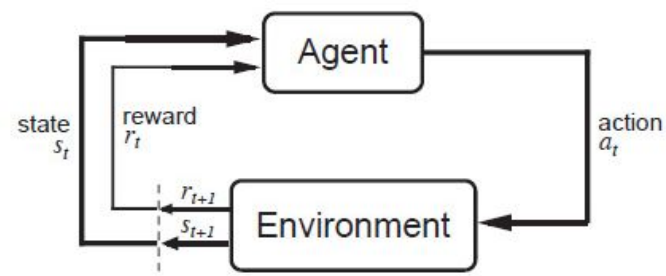
Recap


Training Data
Supervised Learning
All Labeled Data
Supervised model
Semi-supervised Learning
Some Labeled Data
Lot of Unlabeled Data
Semi supervised model
Unsupervised Learning
All Unlabeled Data
Unsupervised model

Reinforcement Learning

Reinforcement Learning

- Goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps; for example, maximize the points won in a game over many moves.
- Interact with the environment
  - Feedback loop between the learning system and its experiences
  - Should take actions to maximize some kind of “reward”
  - Learns from mistakes



AlphaGo

https://skymind.ai/wiki/deep-reinforcement-learning

https://adeshpande3.github.io/Deep-Learning-Research-Review-Week-2-Reinforcement-Learning

https://towardsdatascience.com/a-review-of-recent-reinforcment-learning-applications-to-healthcare-1f8357600407

Limitations of Machine Learning

Limitations of Machine Learning

- External validity
  - Representativeness of training data
  - Generalizability
  - The labeling of classes should evolve over time
- Correlation, not causation
  - How do we determine causation?
- Interpretability
  - Deep learning models viewed as “black boxes”
- Bias (in terms of healthcare, aka algorithm “fairness”)
  - Have we introduced any societal bias?
  - Does the data reflect societal bias?

Action Items

- Sign-up for GitHub account (if you don't yet have one)
- Request MIMIC III data access (if you don't yet have access)
  - Please use school email
- Review linear algebra and Python review slides if you're feeling rusty
  - Available on the course Canvas and GitHub repo
  - Reviewing Calculus would also be useful if you're feeling rusty about how to derive partial derivatives (they will show up in lecture 2 and on homework #1)
- Find group members for the group project proposal
  - Can also work individually
- Email me if you don't have an @hsph.harvard.edu email
  - I'll need to give you access to the Google cloud computing resources