

# **Introduction to Machine Learning**

Digital Transformation of Healthcare

---

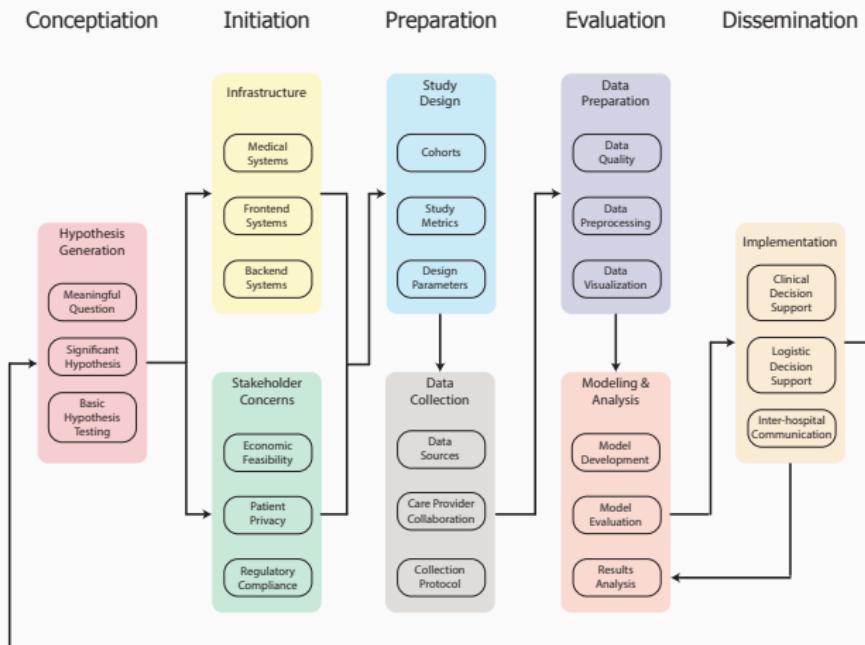
Michoel Snow M.D. Ph.D. and Glen Ferguson Ph.D.

Center for Health Data Innovations

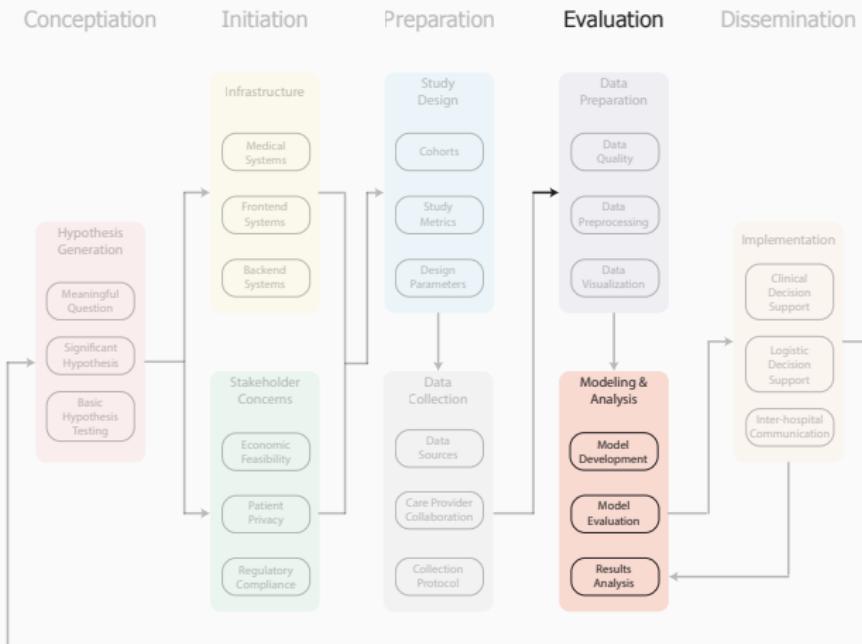
# **Intro to Machine Learning**

---

# Bioinformatics Pipeline

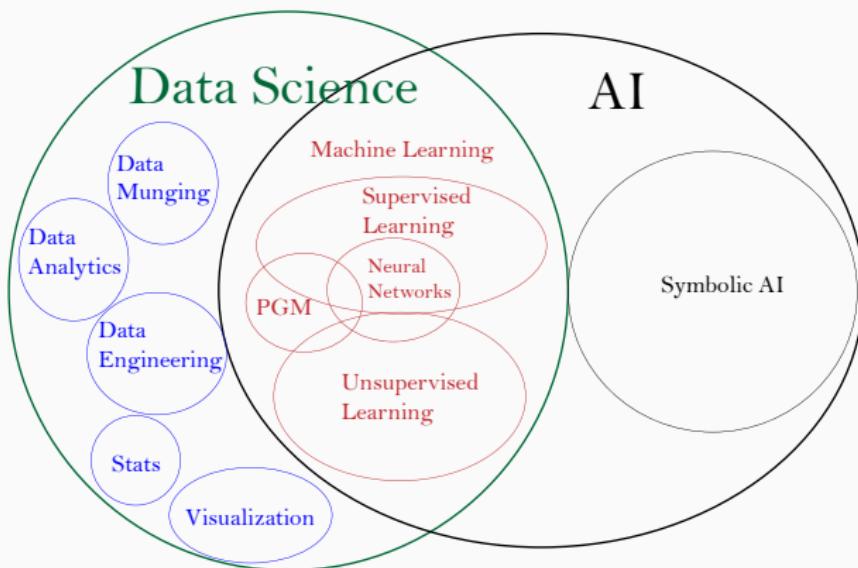


# Data Analysis



# What is Machine Learning?

Machine Learning is part of multiple overlapping disciplines and is often associated with both Data Science and Artificial intelligence.



# What is Machine Learning?

# What is Machine Learning?

- Machine learning is a discipline that uses data to improve a mathematical model

# What is Machine Learning?

- Machine learning is a discipline that uses data to improve a mathematical model
  - Algorithm Examples
    - Logistic Regression
    - Random Forest
    - Neural Networks

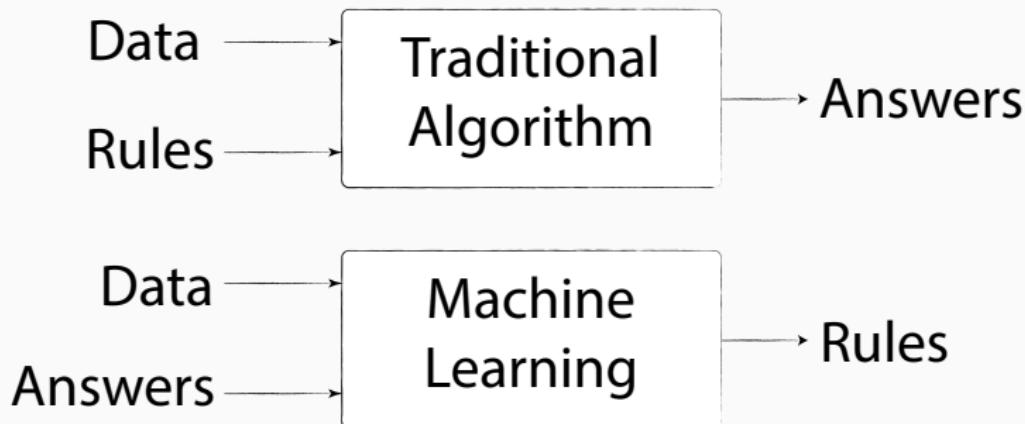
# What is Machine Learning?

- Machine learning is a discipline that uses data to improve a mathematical model
  - Algorithm Examples
    - Logistic Regression
    - Random Forest
    - Neural Networks
- Machine learning contains multiple classes of learning algorithms
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

# What is Machine Learning?

- Machine learning is a discipline that uses data to improve a mathematical model
  - Algorithm Examples
    - Logistic Regression
    - Random Forest
    - Neural Networks
- Machine learning contains multiple classes of learning algorithms
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning
- Machine learning models perform tasks, such as language translation and identifying objects, that present extraordinary difficulty for standard algorithms
  - ImageNet Large Scale Visual Recognition Challenge

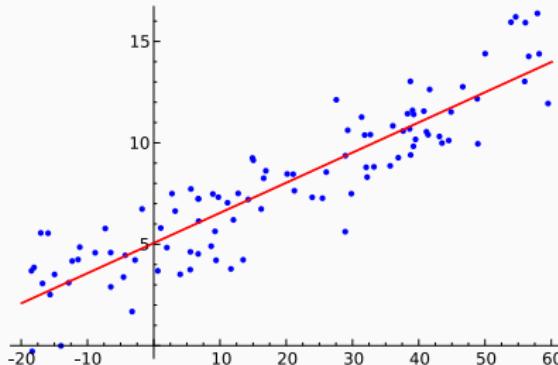
# Traditional Programming vs. Machine Learning



# What is Machine Learning?

Machine learning is a discipline in AI where algorithms use data to improve a model

- Algorithm is a specification of how to solve a problem
- Data is information about the problem we are trying to solve
- Model
  - Scientific - a simplified and idealized understanding of physical systems
  - Computer Science - a simulation to reproduce the behavior of a system



# Types of Learning

- Supervised Learning
  - Prediction of specific value for given data
    - Algorithm learns the values (labels) from the data
    - Learning requires the data to the desired prediction
    - Inference is prediction of these labels from *new* data

# Types of Learning

- Supervised Learning
  - Prediction of specific value for given data
    - Algorithm learns the values (labels) from the data
    - Learning requires the data to the desired prediction
    - Inference is prediction of these labels from *new* data
- Unsupervised Learning
  - Understanding the structure of the data
    - Algorithm learns the structure from the data
    - Learning/inference requires only the data

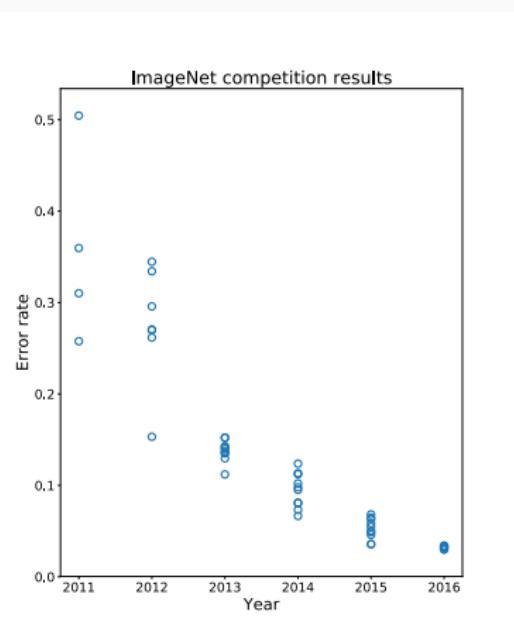
# Types of Learning

- Supervised Learning
  - Prediction of specific value for given data
    - Algorithm learns the values (labels) from the data
    - Learning requires the data to the desired prediction
    - Inference is prediction of these labels from *new* data
- Unsupervised Learning
  - Understanding the structure of the data
    - Algorithm learns the structure from the data
    - Learning/inference requires only the data
- Reinforcement Learning
  - Predicting optimal actions for the environment based on rewards
    - Algorithm learns the actions that maximize a reward for an environment
    - Learning requires information about the environment, rewards and actions
    - Inference is prediction of optimal actions for environment

# What is Machine Learning?

- ImageNet Large Scale Visual Recognition Challenge
  - Algorithmic recognition of 1000 items in 150,000 photos

**Figure 1:** ImageNet Competition Results, By Gkrusze CC BY-SA 4.0



# What is Machine Learning?

- Exercise
  - Using your domain expertise answer the following

# What is Machine Learning?

- Exercise
  - Using your domain expertise answer the following
  - Determine two problems in both Supervised and Unsupervised Learning

# What is Machine Learning?

- Exercise
  - Using your domain expertise answer the following
  - Determine two problems in both Supervised and Unsupervised Learning
  - What types of data are required for each?

# What is Machine Learning?

- Exercise
  - Using your domain expertise answer the following
  - Determine two problems in both Supervised and Unsupervised Learning
  - What types of data are required for each?
  - How does ML fit into the overall health care informatics pipeline?

# What is Machine Learning?

- Exercise
  - Using your domain expertise answer the following
  - Determine two problems in both Supervised and Unsupervised Learning
  - What types of data are required for each?
  - How does ML fit into the overall health care informatics pipeline?
  - How can you used both in a single project?

# Supervised Learning

---

# Types of Supervised Learning

---

*What are the type of supervised learning?*

- Regression - predicting a continuous numerical value

# Types of Supervised Learning

---

*What are the type of supervised learning?*

- Regression - predicting a continuous numerical value
- Classification - prediction of discrete class labels

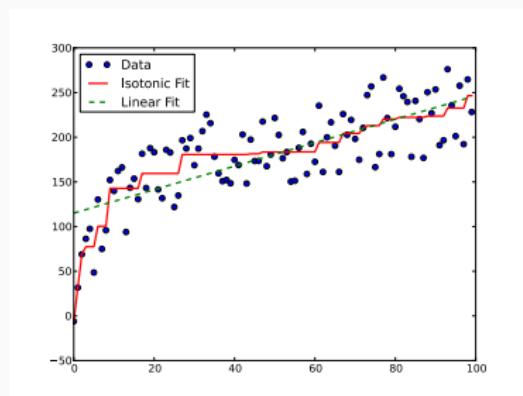
# Types of Supervised Learning

*What are the type of supervised learning?*

- Regression - predicting a continuous numerical value
- Classification - prediction of discrete class labels

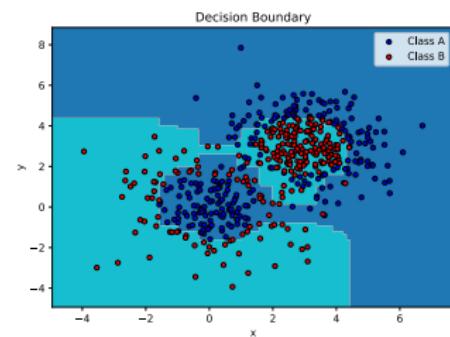
**Figure 2:** Non-linear Regression, By

Alexeicolin - Own work, CC BY-SA 3.0



**Figure 3:** Classification, By Alisneaky Own work,

CC BY-SA 4.0



# Types of Predictions

---

- Exercise
  - Determine the types of predictions for the Supervised Learning problem above

# Types of Predictions

---

- Exercise
  - Determine the types of predictions for the Supervised Learning problem above
  - How many classes do the Classifications problems have?

# Basic Methods

## *Linear and Logistic Regression*

- Linear Regression - Predicting continuous values from linearly separable data

# Basic Methods

## *Linear and Logistic Regression*

- Linear Regression - Predicting continuous values from linearly separable data
- Logistic Regression - Subclass of Generalized Linear Models for binary class prediction

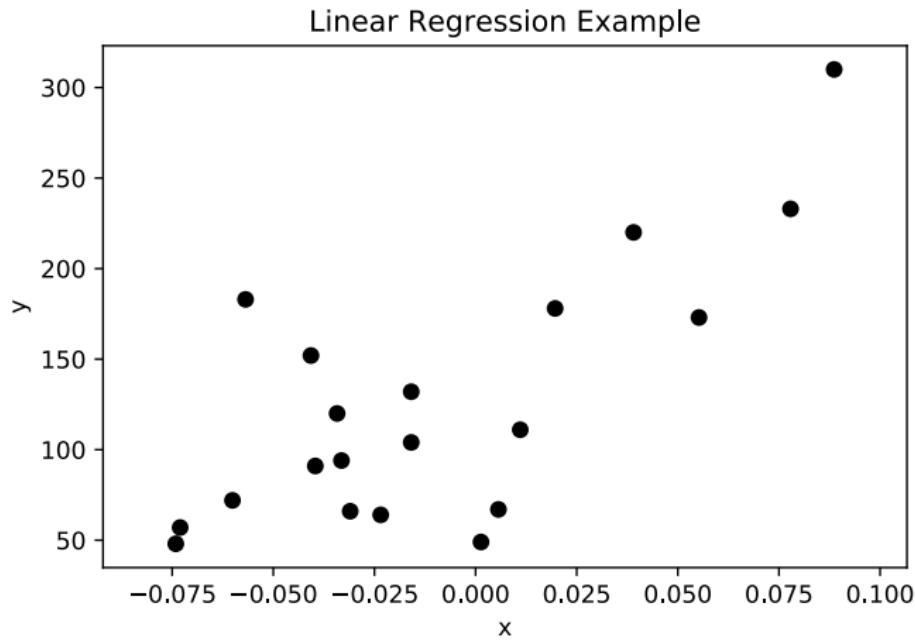
# Basic Methods

## *Linear and Logistic Regression*

- Linear Regression - Predicting continuous values from linearly separable data
- Logistic Regression - Subclass of Generalized Linear Models for binary class prediction
- Data **should** be linearly separable or variables linearly correlated

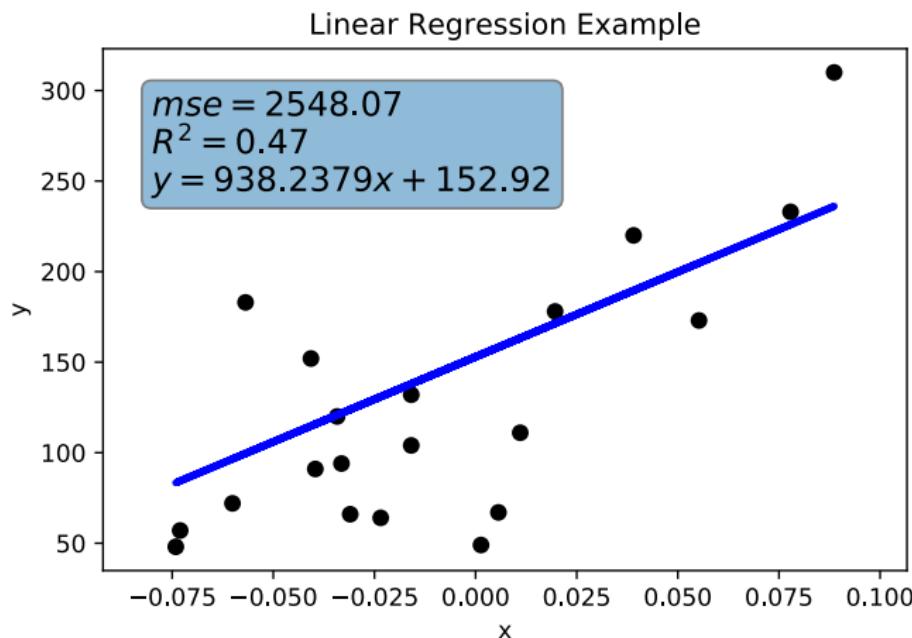
# Linear Regression

Predicting continuous values



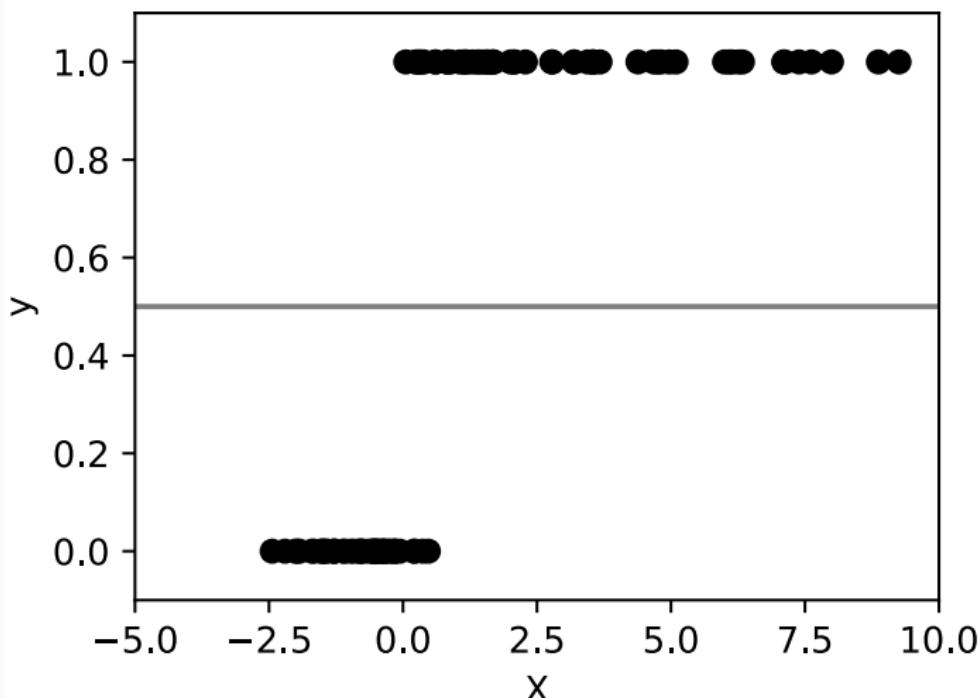
# Linear Regression

Predicting continuous values



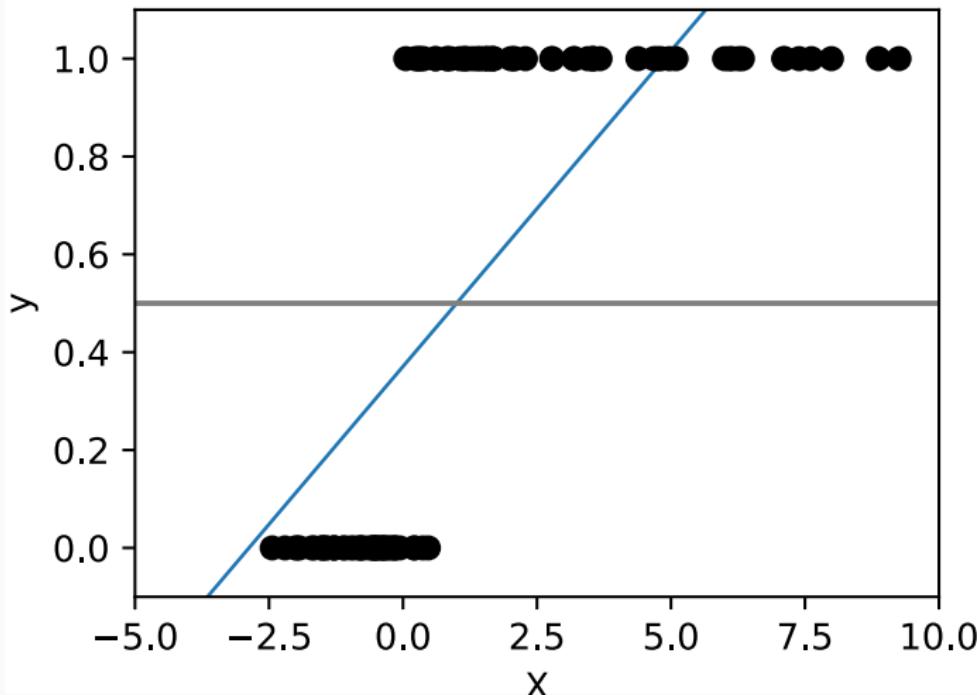
# Logistic Regression

What if the data is binary?



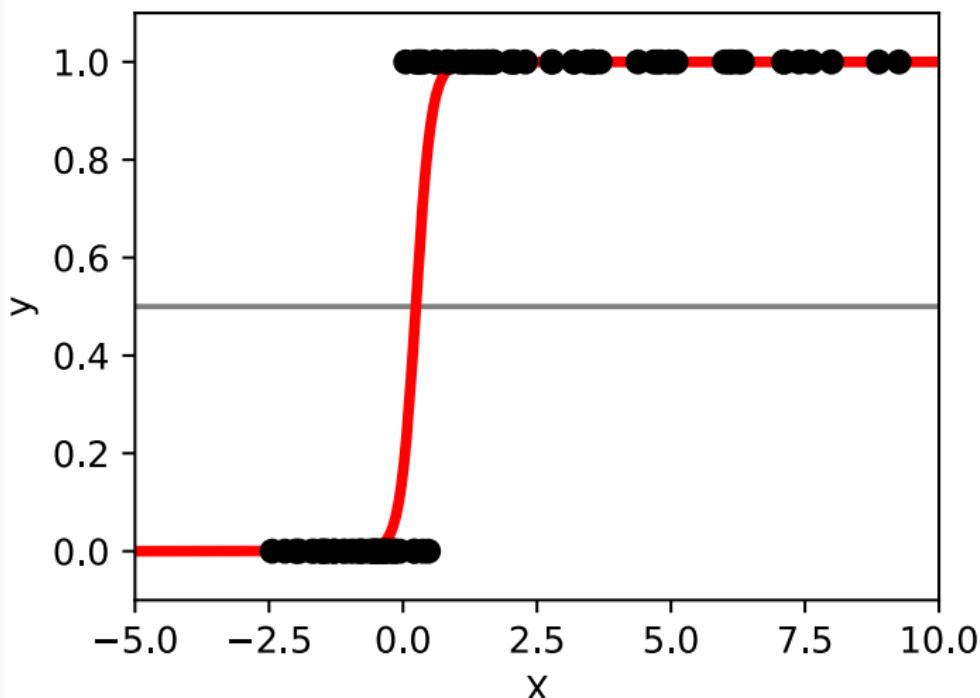
# Logistic Regression

The linear fit does not work



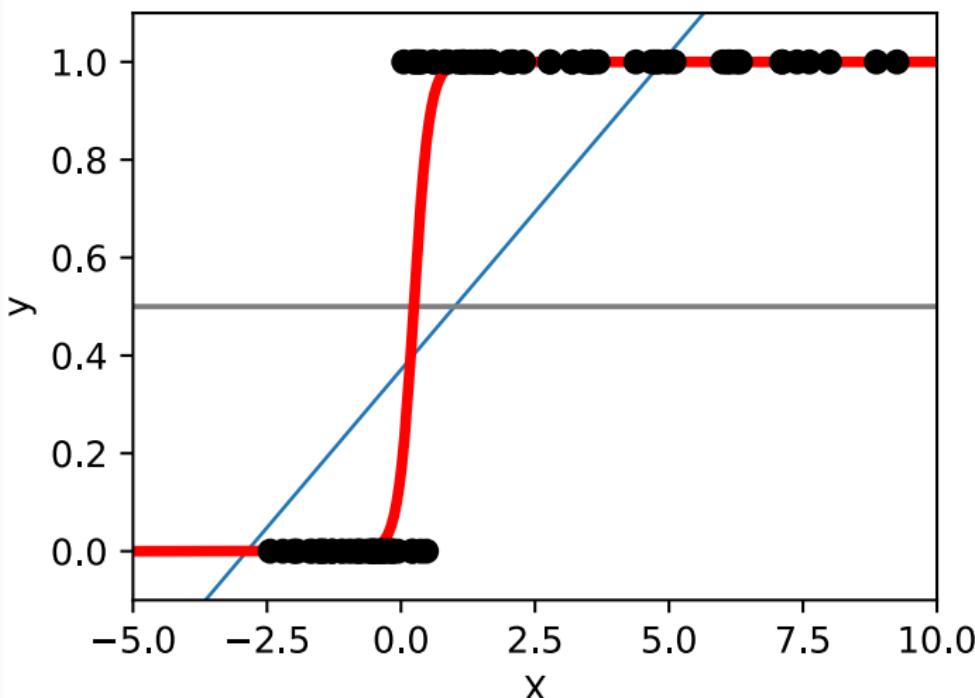
# Logistic Regression

Sigmoid function fits the data well



# Logistic Regression

Logistic regression fits uses a linear fit to a sigmoid function



# Iris Data

*The iris data set is commonly used for teaching machine learning*

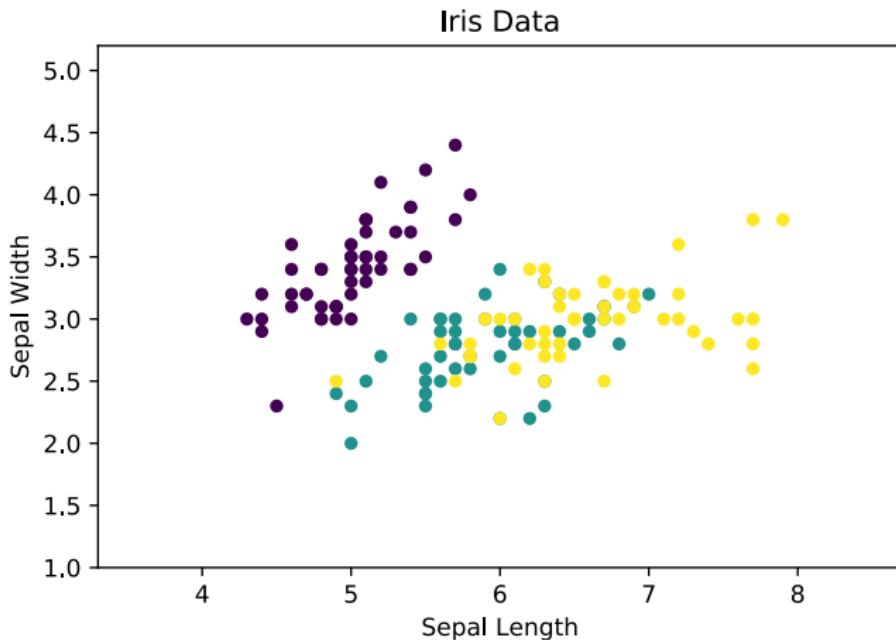
- Data from 1930's to identify flowers from petal and sepal measurements
- Three species of iris flowers: Setosa, Versicolor and Virginica
- Only a few variables, Sepal width, Sepal length, Petal width and Petal length
- 150 sets of measurements

**Figure 4:** Versicolor Iris flower, sepal labeled



# Iris Data

Dataset of Iris Flower Types



## *Support Vector Machines and Decision Trees*

- Support Vector Machine - Prediction by finding the best (hyper)planes that separate the data
- Tree-Based Method - Prediction of using decision-trees
  - Decision Trees
  - Random Forest
  - Boosted Trees

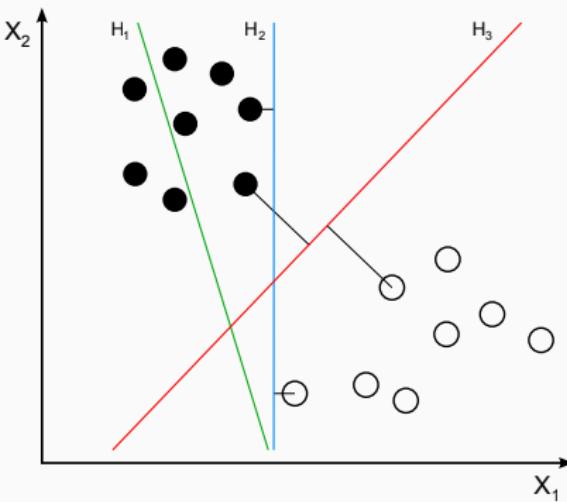
# Support Vector Machines

*SVC and SVR are a common methods of fitting data*

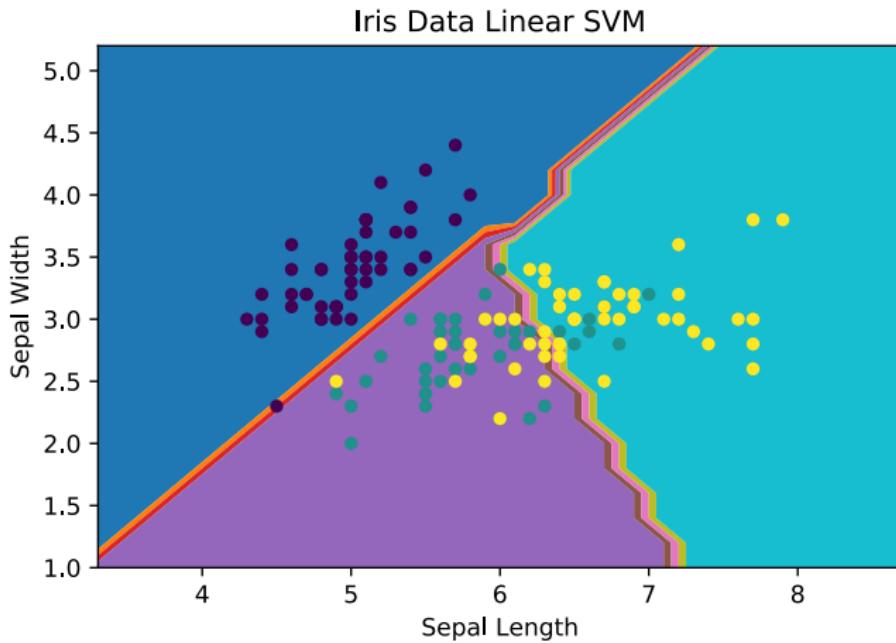
- Finds the plane that *best* divides the data
- Can fit highly non-linear data
- Kernels allow for flexibility but must be chosen by the user
- Does not fit data with more than two classes
- Does not give class probabilities just class membership
- Parameters can be difficult to choose

**Figure 5:** Best plane example, By

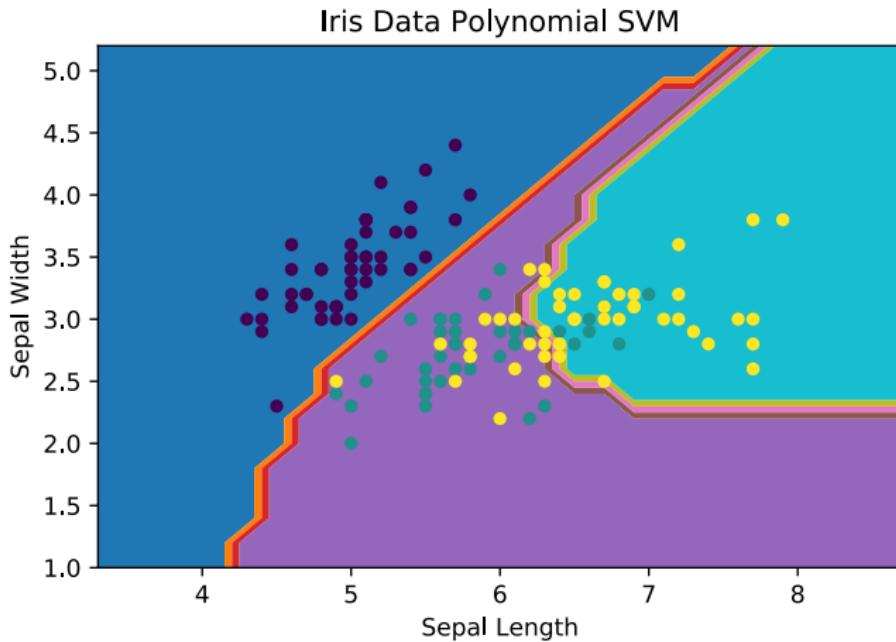
User:ZackWeinberg, based on PNG by User:Cyc derived from: Svm separating hyperplanes.png, CC BY-SA 3.0



# Support Vector Machine



# Support Vector Machine

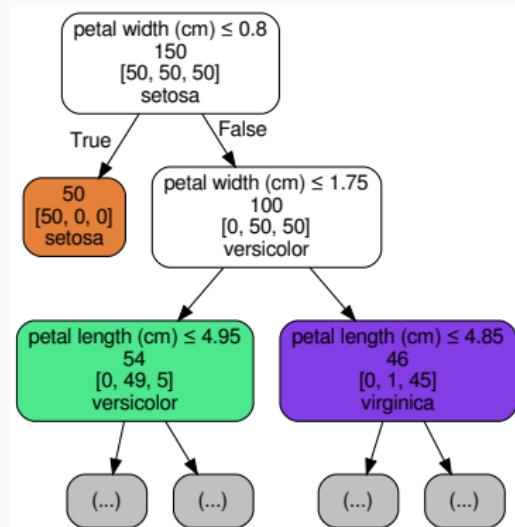


# Tree-based Methods

*Methods using decision trees or ensembles of decision trees for classification and regression problems*

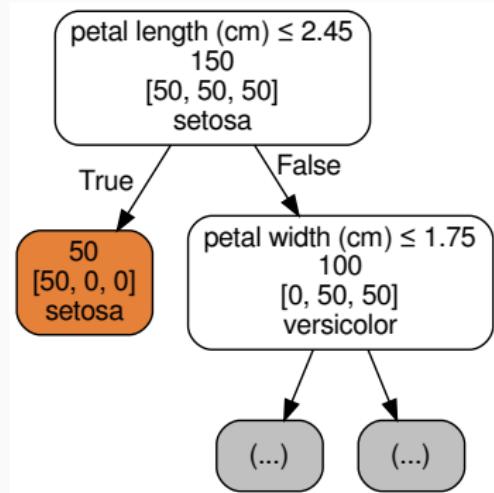
- Fits a decision tree based on the data to predict results
- Can fit highly non-linear data with *high* accuracy
- Models can be very large
- Some models, Random Forest, do not fit sparse data well
- Other models, Boosting, have many parameters that can be difficult to choose

Figure 6: Decision Tree Example

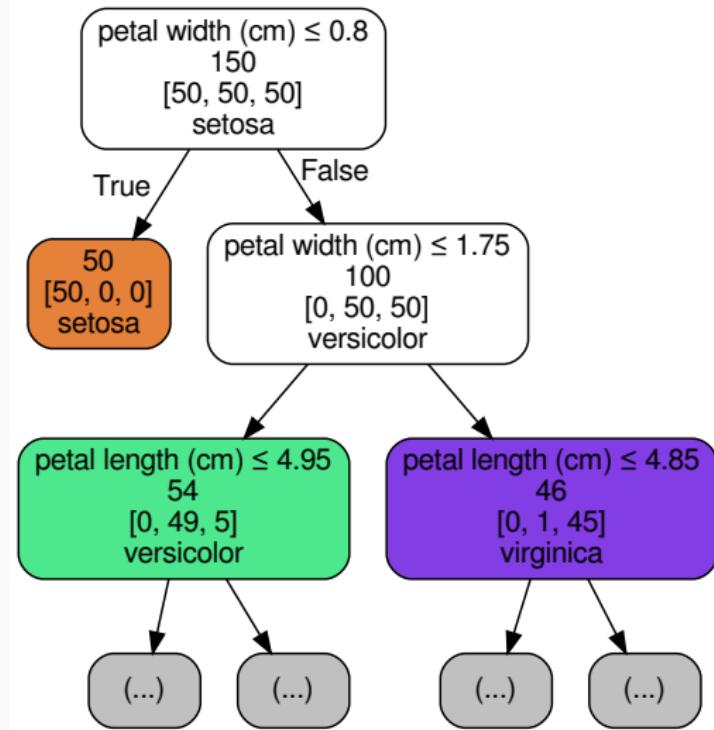


# Building a Decision Tree

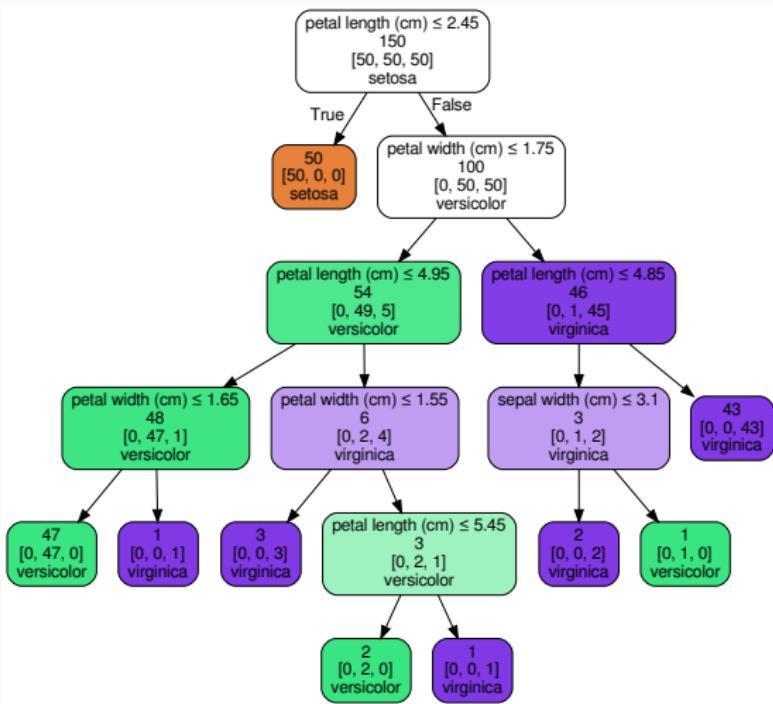
Make splits in the data in such a way that it separates the classes or that minimizes the distance between the predictions in a region.



# Building a Decision Tree

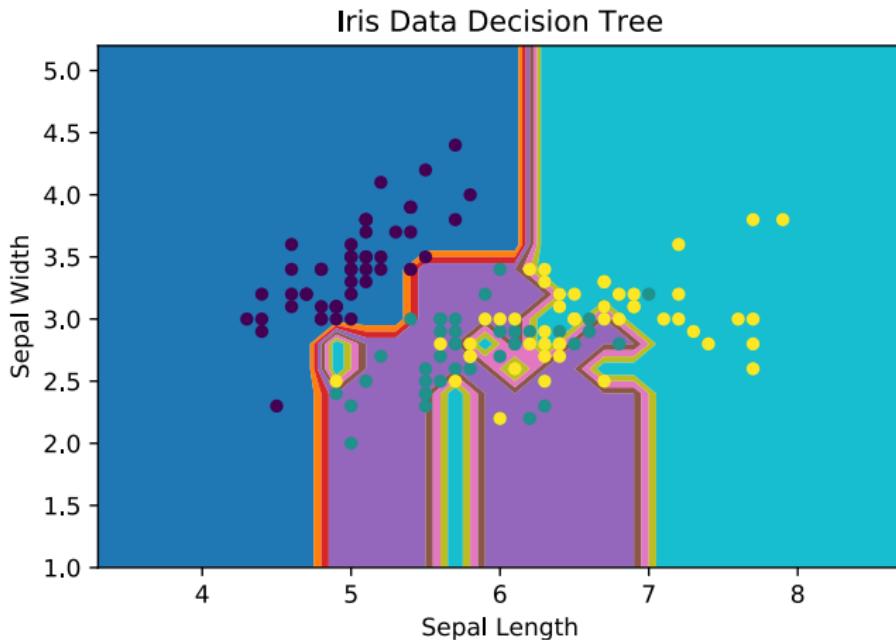


# Building a Decision Tree



# Decision Tree

## Decision Tree Decision Boundary



# Decision Tree

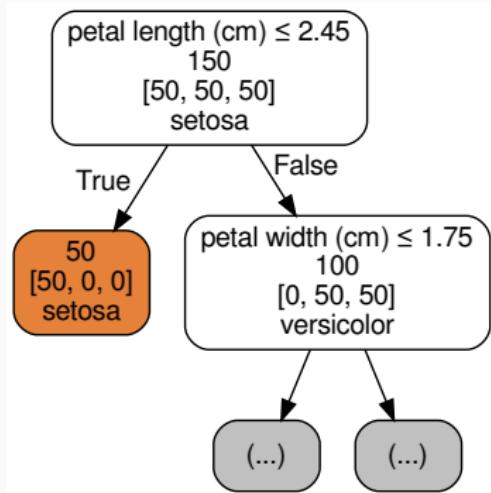
---

A method to fit data that generates reasonable decision boundaries

- Relatively fast and easy to fit
- Intuitive to use
- Often over fits data
  - Over fitting may be controlled with pruning
  - Pruning may reduce accuracy
- Large trees are difficult to interpret

# Building a Forest of Decision Trees

Make splits in the data in such a way that it separates the classes or that minimizes the distance between the predictions in a region.

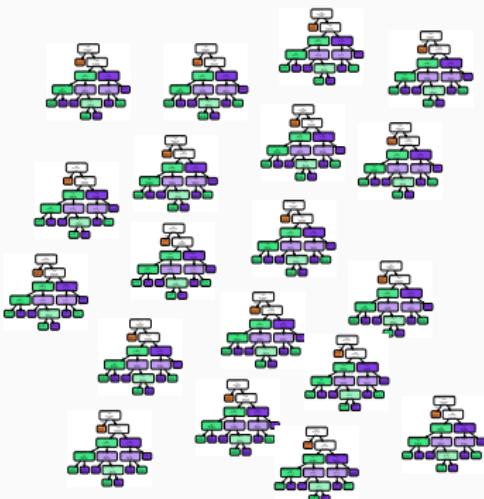


# Random Forest

*Combine decision trees into an ensemble*

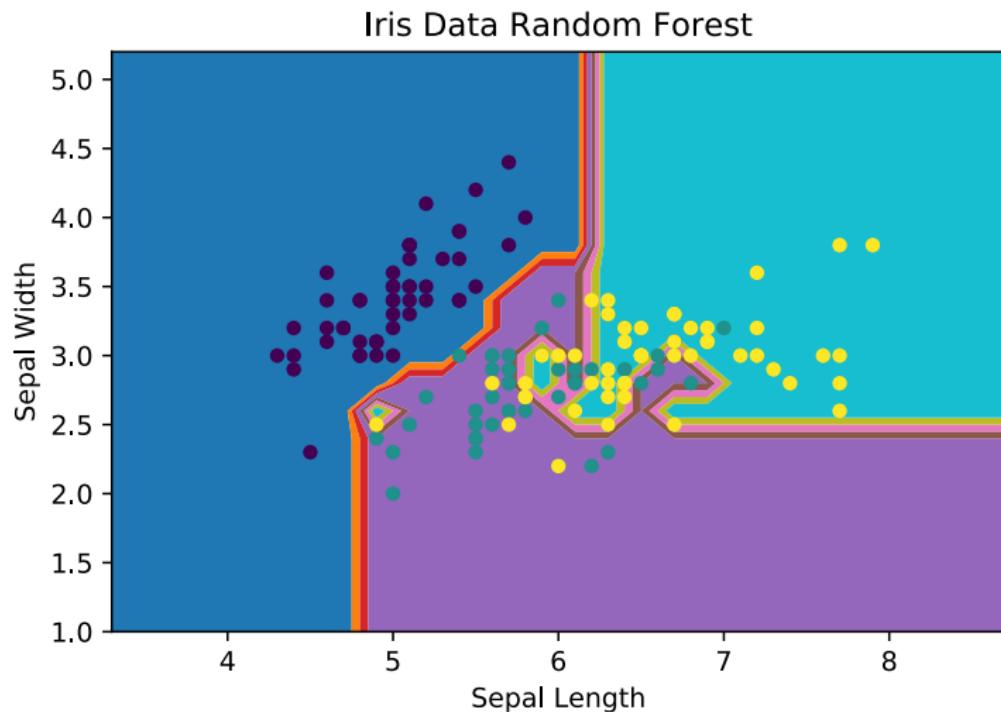
- Work by randomly splitting predictors and data (bagging)
- Make a large number of trees and average their predictions
- Can be optimized with relatively few parameters
- Resulting models can be very large making inference slow
- Models do not fit sparse data well
- Results are difficult to interpret

**Figure 7:** Decision Tree Example



# Random Forest

## Random Forest Decision Boundary



## Boosted Trees

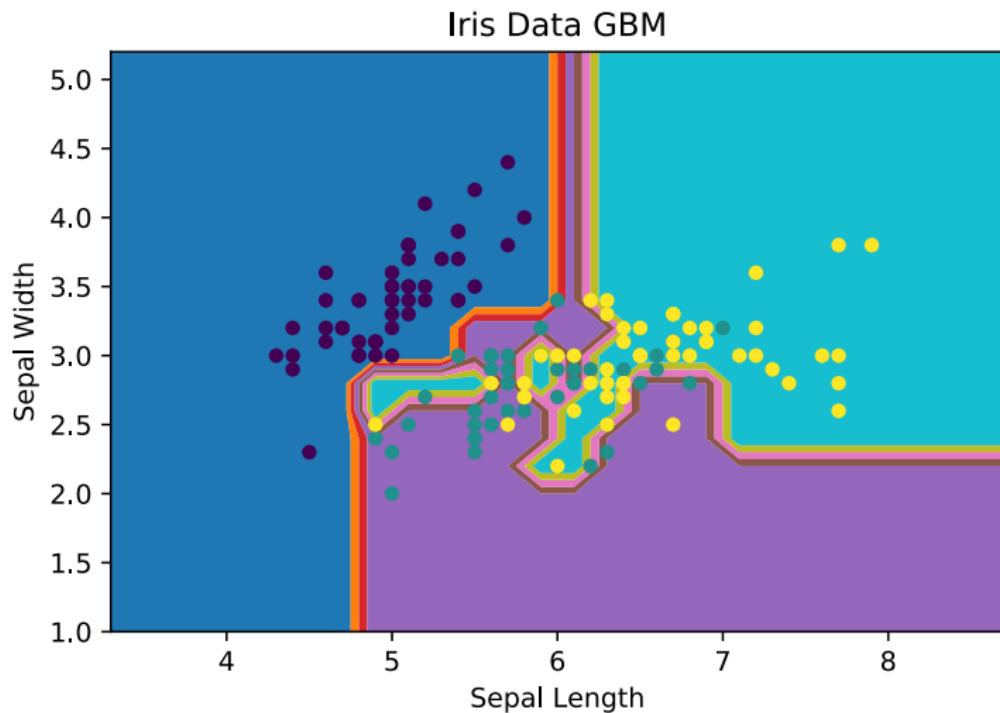
---

*Combine short decision trees fitting to the error of the previous tree*

- Make an ensemble of weak learners (small decision trees)
- Fit to decision trees to reduce the error of the previous tree
- Models generally have very high accuracy
- Over fitting is common
- Large number of difficult to optimize parameters

# Boosted Trees

## GBM Decision Boundary



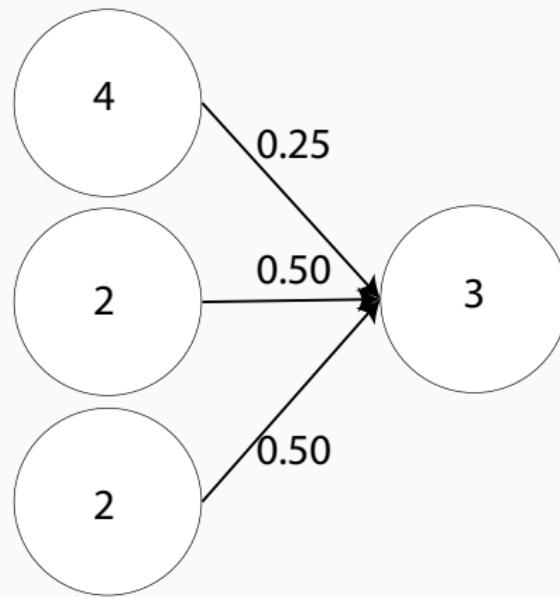
# Neural Networks

---

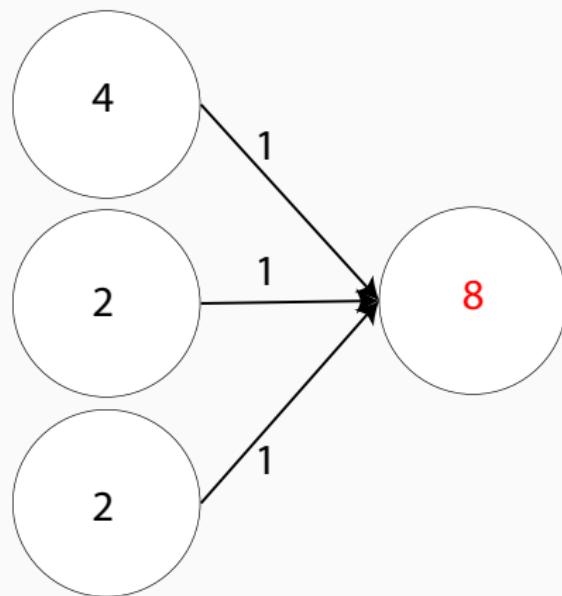
*A class of models using collections of linear algebra operations to make predictions, classify data, so called deep learning and perform complex tasks*

- Wide-range of applications from classification to self-driving cars
- Automatic variable selection and engineering
- Require extensive fitting
- Computationally intensive
- Many architectures for different applications
- Applications include computer vision, natural language processing, predictions, task learning and speech recognition
- Fundamentally simple operations that when combined solve complex problems

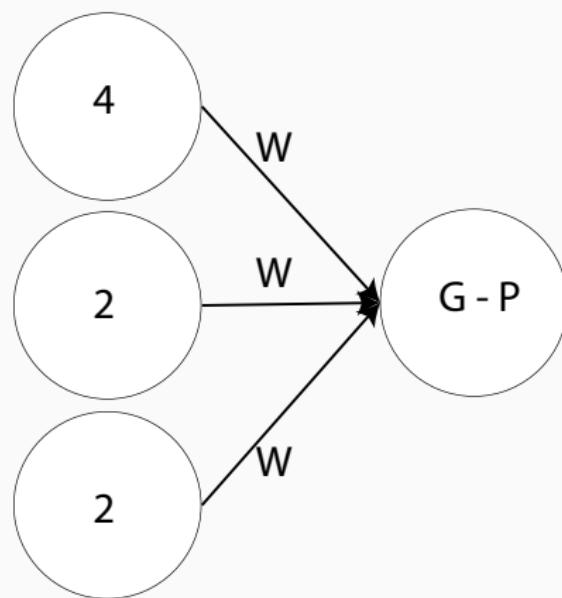
# Simplest Neural Network



# Simplest Neural Network

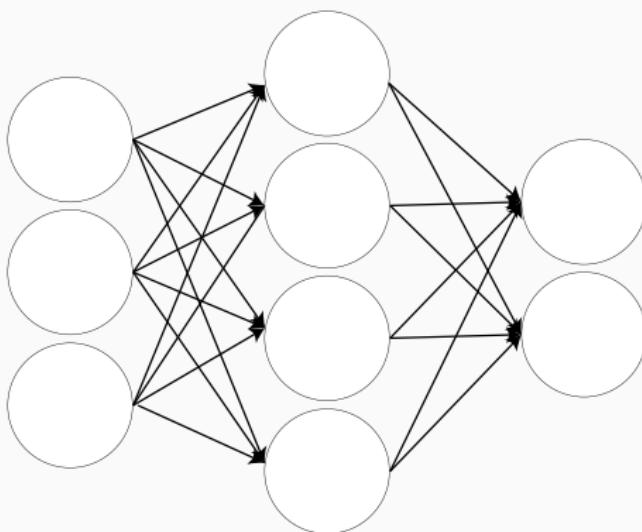


# Simplest Neural Network



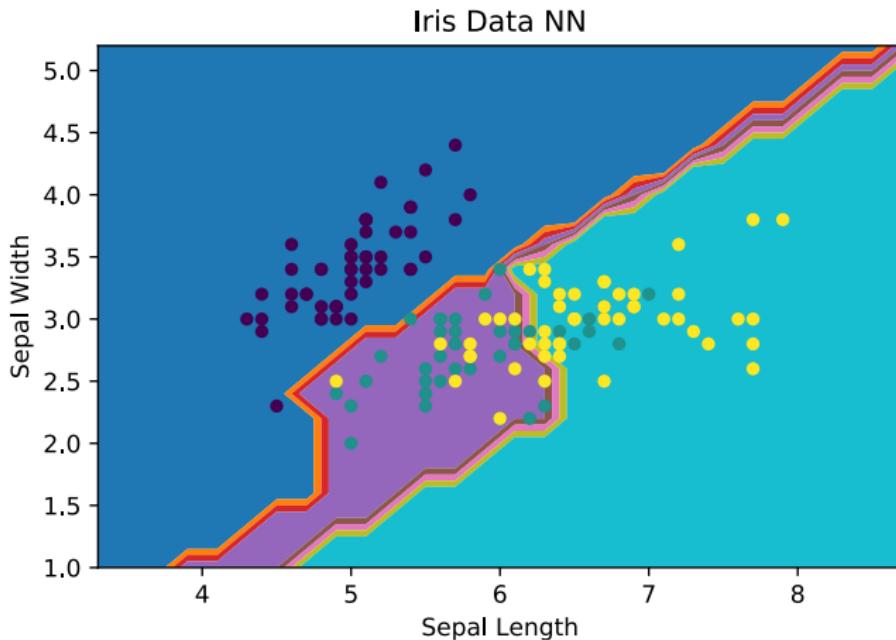
# Simplest Neural Network

Input              Hidden Layer              Output



# Neural Network

## Neural Network Decision Boundary



# Other Supervised Learning Methods

---

*Many other methods available*

- Regularized regression method to control overfitting and reduce variance
- Naive Bayes
- Gaussian Process
- Linear Discriminant Analysis
- k-Nearest Neighbors

## **Unsupervised Learning**

---

# Inferring the structure of unlabeled data

---

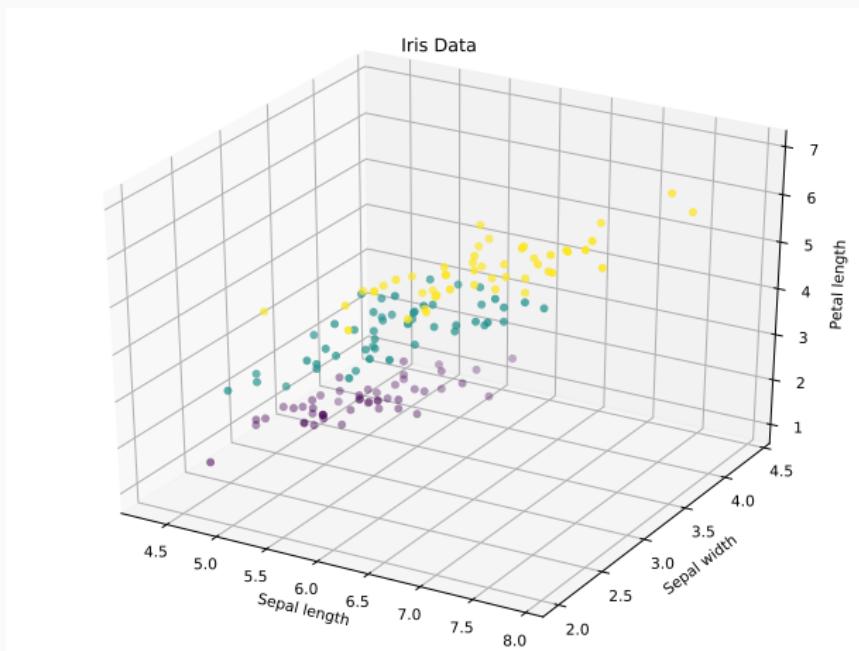
## *Support Vector Machines and Decision Trees*

- Dimensionality Reduction
  - Principle Component Analysis
  - Non-negative Matrix Factorization
  - Latent Dirichlet Allocation
- Clustering
  - K-means
  - Gaussian Mixture Models
  - Manifold Learning

# PCA Dimensionality Reduction

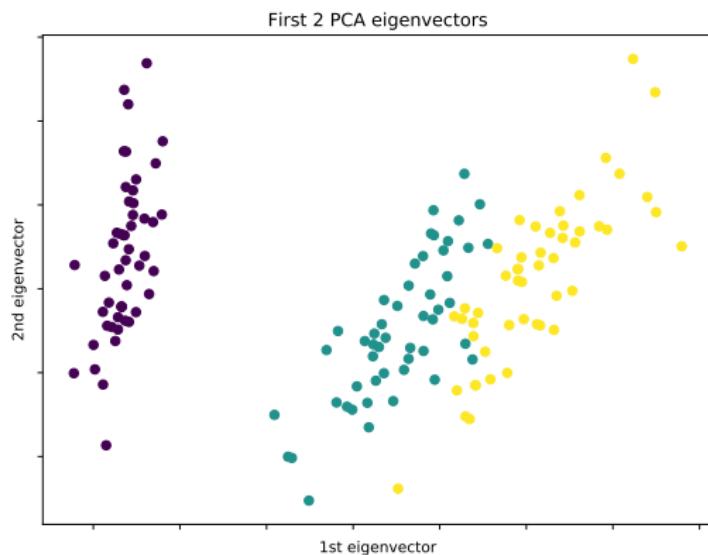
*Find the best low dimensional representation of the data*

**Figure 8:** Iris Data with Three Components



# PCA

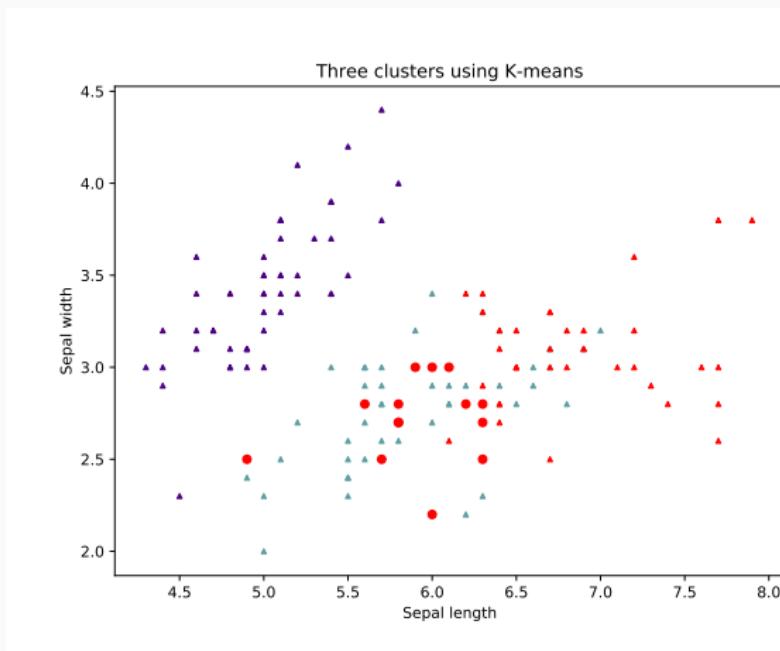
First two PCA eigenvectors



# K-means

*Find K groupings based on variable similarity.*

**Figure 9:** Iris Data with 3 cluster



# Reinforcement Learning

---

## New problem types

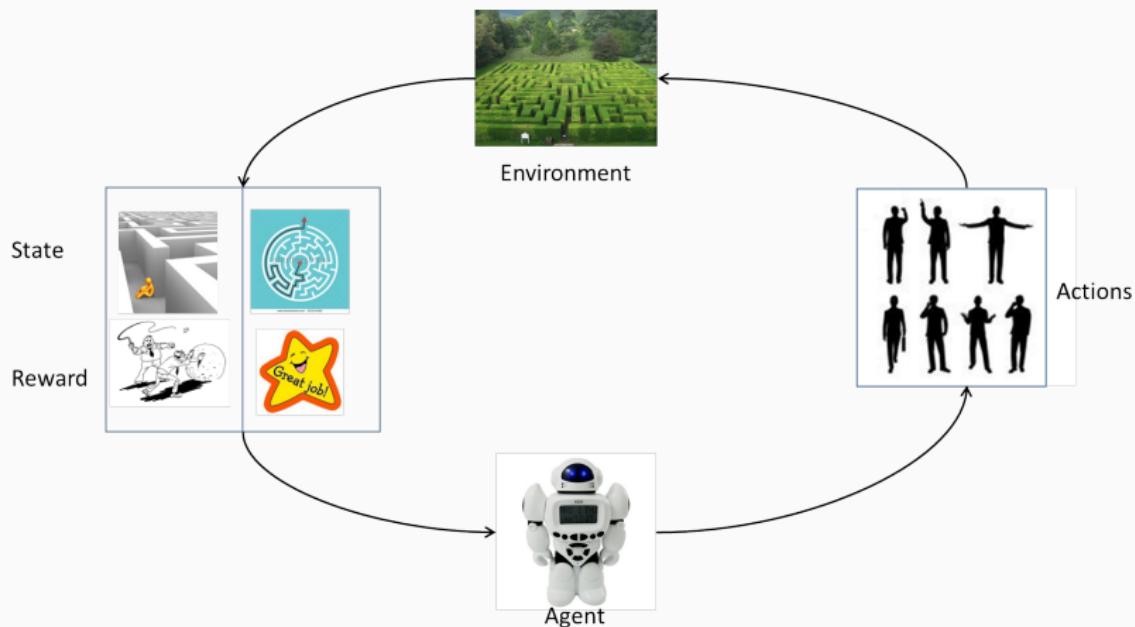
---

*What if we know the desired outcome but the data needed to model every possible situation is vast or there the decision is dependent on many changing factors?*

- Reinforcement learning is a method to learn a policy to arrive at an outcome rather than learning a labeled value
- Successes include self-driving cars, beating the best in the world at the game of Go and playing most Atari games at superhuman levels
- Must define the actions, the rewards and the states
- Very data intense

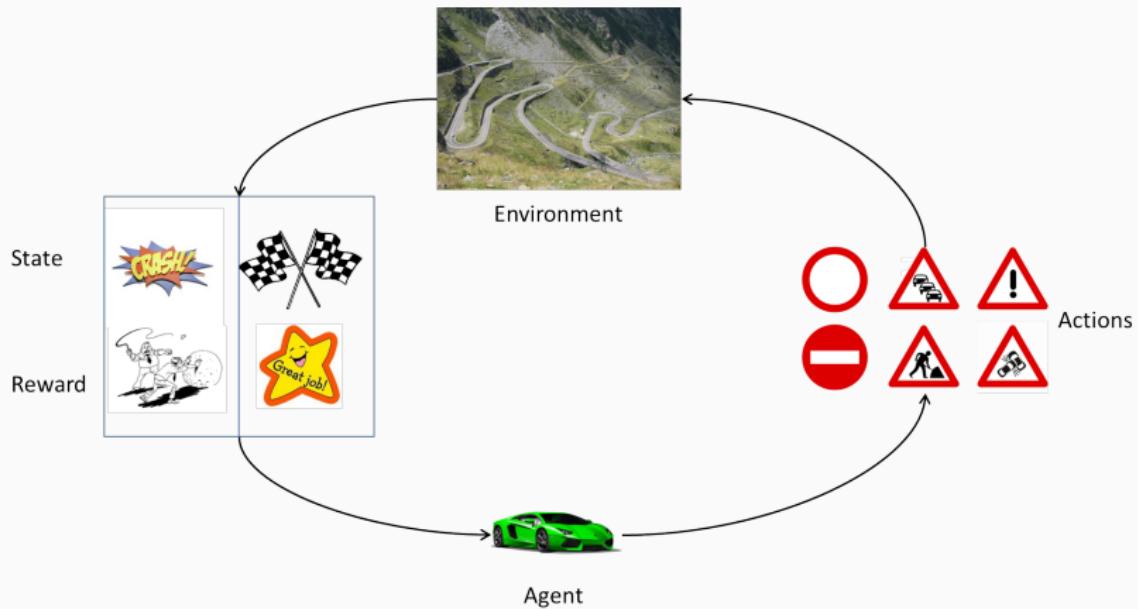
# Reinforcement Learning Concept

*What if we know the desired outcome but the data needed to model every possible situation is vast or there the decision is dependent on many changing factors?*



# Reinforcement Learning Example

## *Self-Driving Race Car*



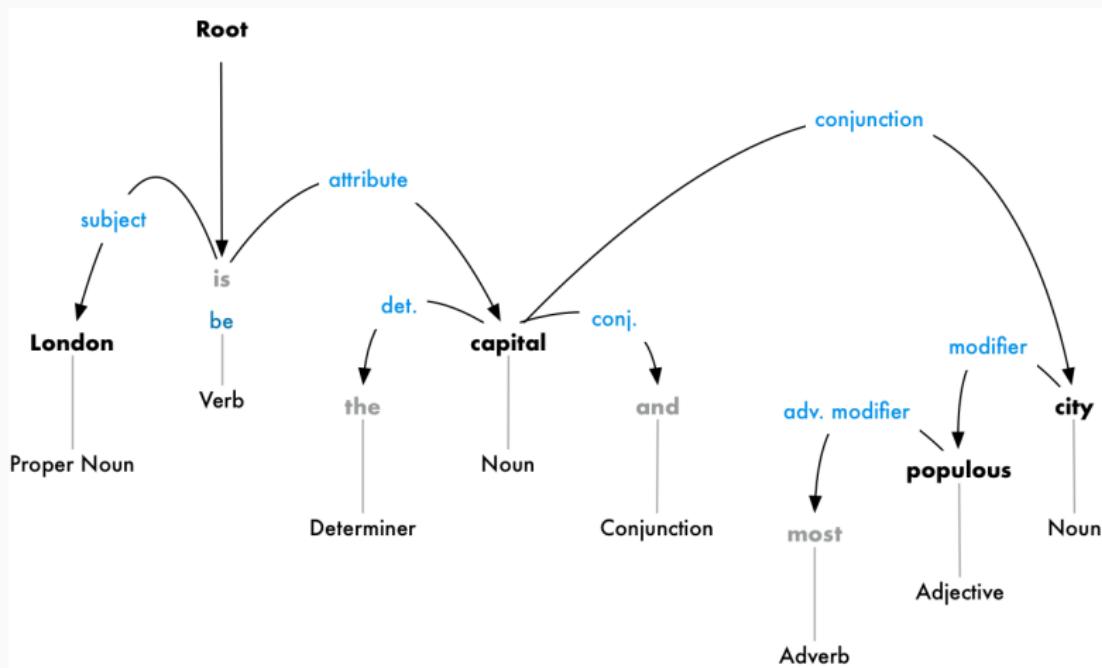
## Special Topics

---

# Teaching a computer to read

*Computers can't read! Wait can they?*

London is the capital and most populous city.



# How do computers see?

*What about images? What would a machine learning algorithm use to add the labels below?*

