McGill **Artificial Intelligence** Society

# Lecture 1: Introduction to ML and Data Manipulation

Slides based off of Machine Learning at Berkeley
https://github.com/mlberkeley/Machine-Learning-Decal-Fall-2018

# Overview

Who are we?

What is machine learning?

Class Logistics

General Overview and Context

Machine Learning Pipeline

Python/Numpy/Scikit-Learn

Questions

# Who are we?

# The Instructors

## Frank Ye

U4 Electrical Eng.

- Former Data Science/ML Intern at Splunk (Incoming Full-Time)
- Former Software Engineering Intern at Under Ericsson

## Isaac Chan

U4 Electrical Eng.

- Current ML Research Intern at Huawei Technologies
- Former Research Intern at the Graphics & Imaging Lab

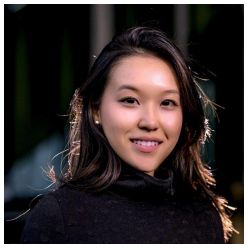
# The Teaching Assistants (TAs)


Aanika


John


Tiff


Claudia


Jenny


David


Hisham


Meg


Nabil


Ketan


Daoud


Josh

# What is Machine Learning?

# Age Old Question

# Can AI Compose Music

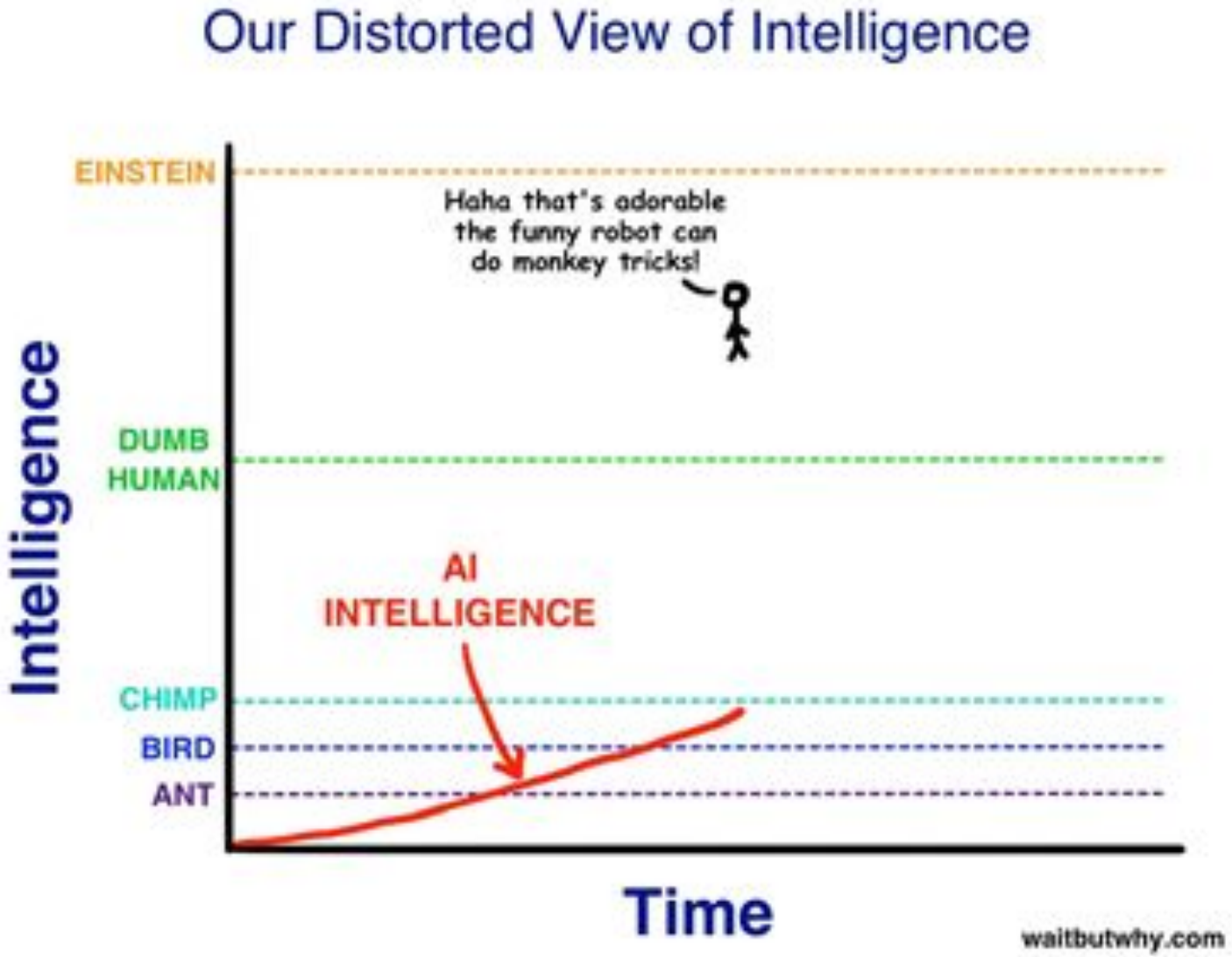
# Can AI Paint A Canvas?

# Pose Tracking!

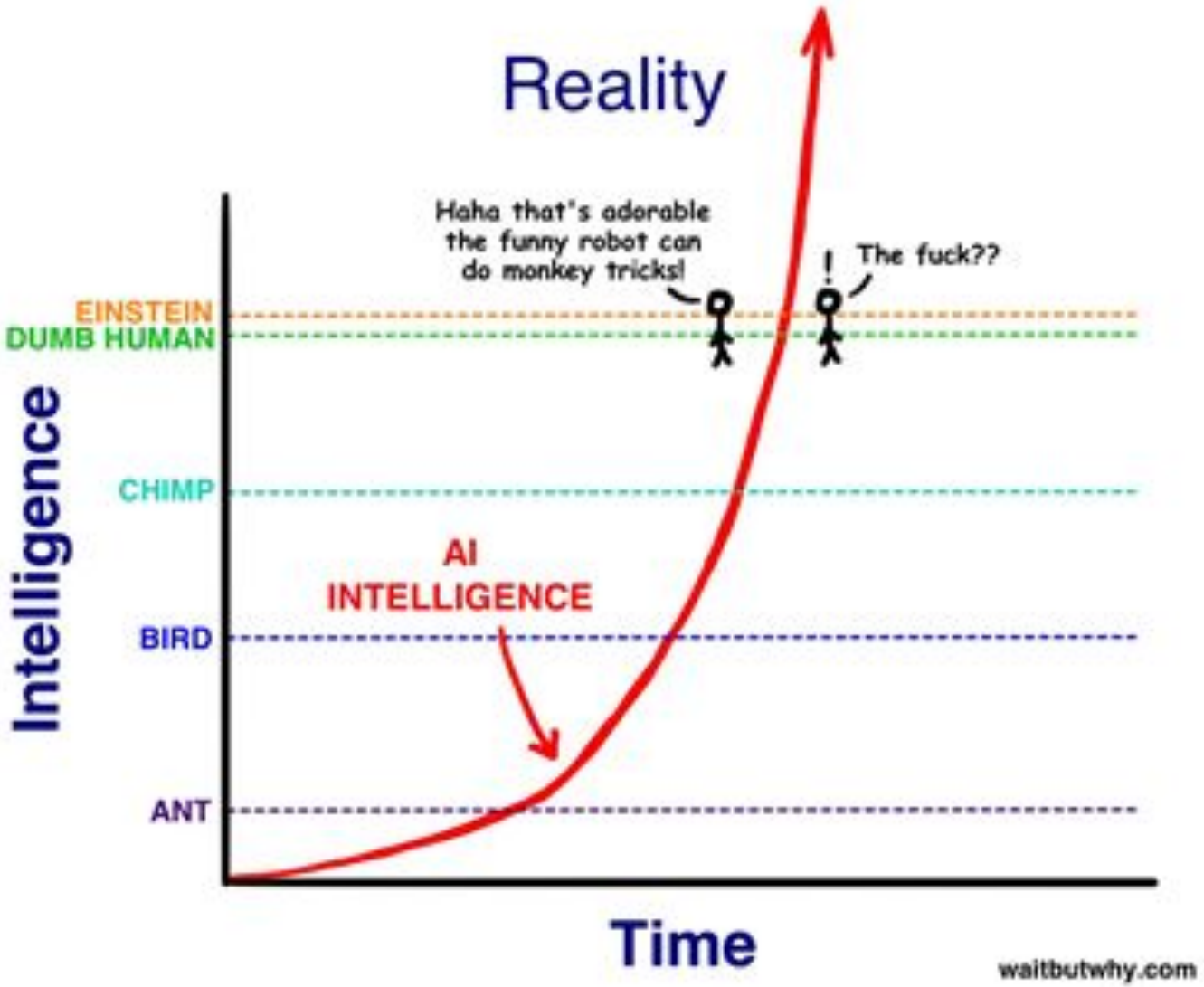
# Superhuman Reasoning!

# Self-Driving Cars

# Our View Of Intelligence



# What Intelligence Is Actually Like



# Class Logistics

# Goals

Understand major concepts in machine learning

Understand trade-offs between different approaches (what do I use - when and why?)

- Statistical Learning Methods
- Deep Learning Methods

Gain familiarity to solve ML problems

Develop the skills for your first data science/machine learning  internship

Build a community and have fun!

# How we accomplish this

**Lectures (8 total)**

- 2 hours/week
- Theoretical introduction
- Hands on coding tutorial

**Homework (5 total)**

- 3-6 hours/week
- Practice implementing material taught in lecture
- Due before the next lecture after being assigned

**Office Hours**

- 2 hours per week
- Dedicate time for in person support

# How we accomplish this (Ct'd)

**Final Project**

- 3 deliverables + final presentation
- Find a dataset on kaggle and create any real life application using the data (web app, mobile app, IoT, robot, etc)

**Final Blog**

- Submit on medium at end of course
- Summarize project and reflect on learning

**Guest Lectures and Social Events (Optional)**

- Gain extra insight from industry professionals
- Work hard and play hard!

# Logistical

**Join the Slack**
https://slack-link-here.com

**Clone the github**
https://github.com/McGillAISociety/mais-bootcamp-w2019

**Share your Repository**

All assignments and deliverables will be marked from your github repository. Learning git is crucial for your career!

# Attendance

Attendance is mandatory and will be taken at every lecture.

You may miss up to 1 lecture (there are only 8!) if you have midterms or other commitments.

*"Studying for midterms or busy with assignments"* is not an excuse. Conflicting schedule with midterm time is excusable.

We have put in a HUGE COMMITMENT to be here for YOU.

If you are not gaining value out of lectures, tell us why!

# Evaluation

Homeworks and deliverables are marked based on completion with feedback given to help you improve.

You may have up to 1 incompletion for homework assignments, but all deliverables must be completed.

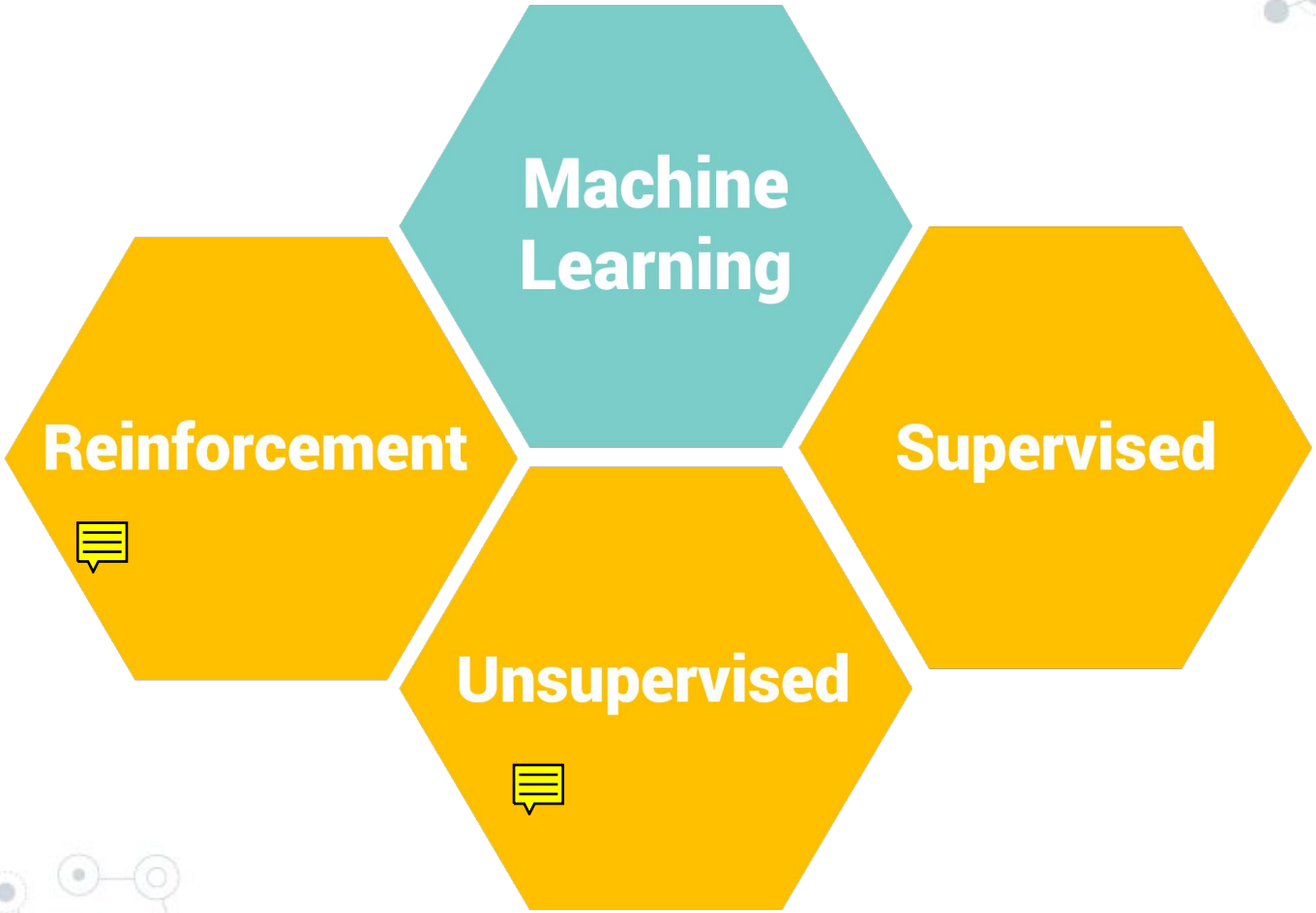
Final project and blogs must be completed for certificate.

We aren't paid for teaching and you do not need this for graduation. We all get out what we put into it.

We will not be enforcing plagiarism and we encourage collaboration, because learning together helps AND you get to make new friends!

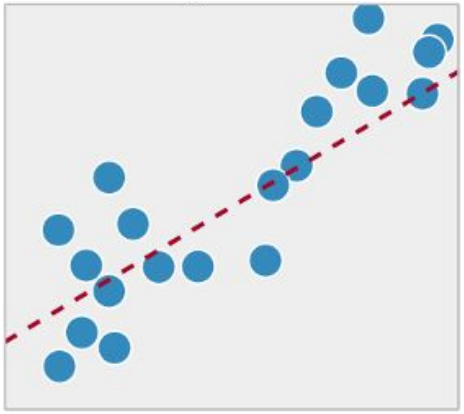
# General Overview And Context

# 3 different classes of machine learning problems



# 1- Supervised Learning

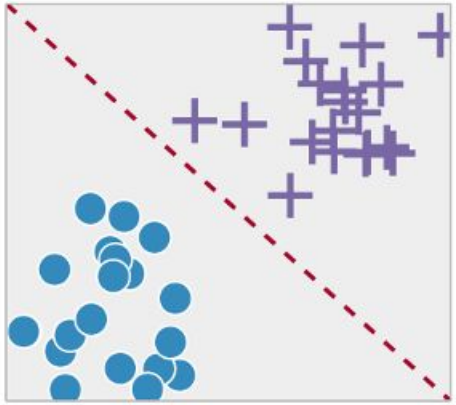
**Regression**



Learning a function for a **continuous** output

Eg. Predicting sales price of house.

**Classification**



Learning a function for a **categorical** output

Eg. Classifying cats vs dogs in images.

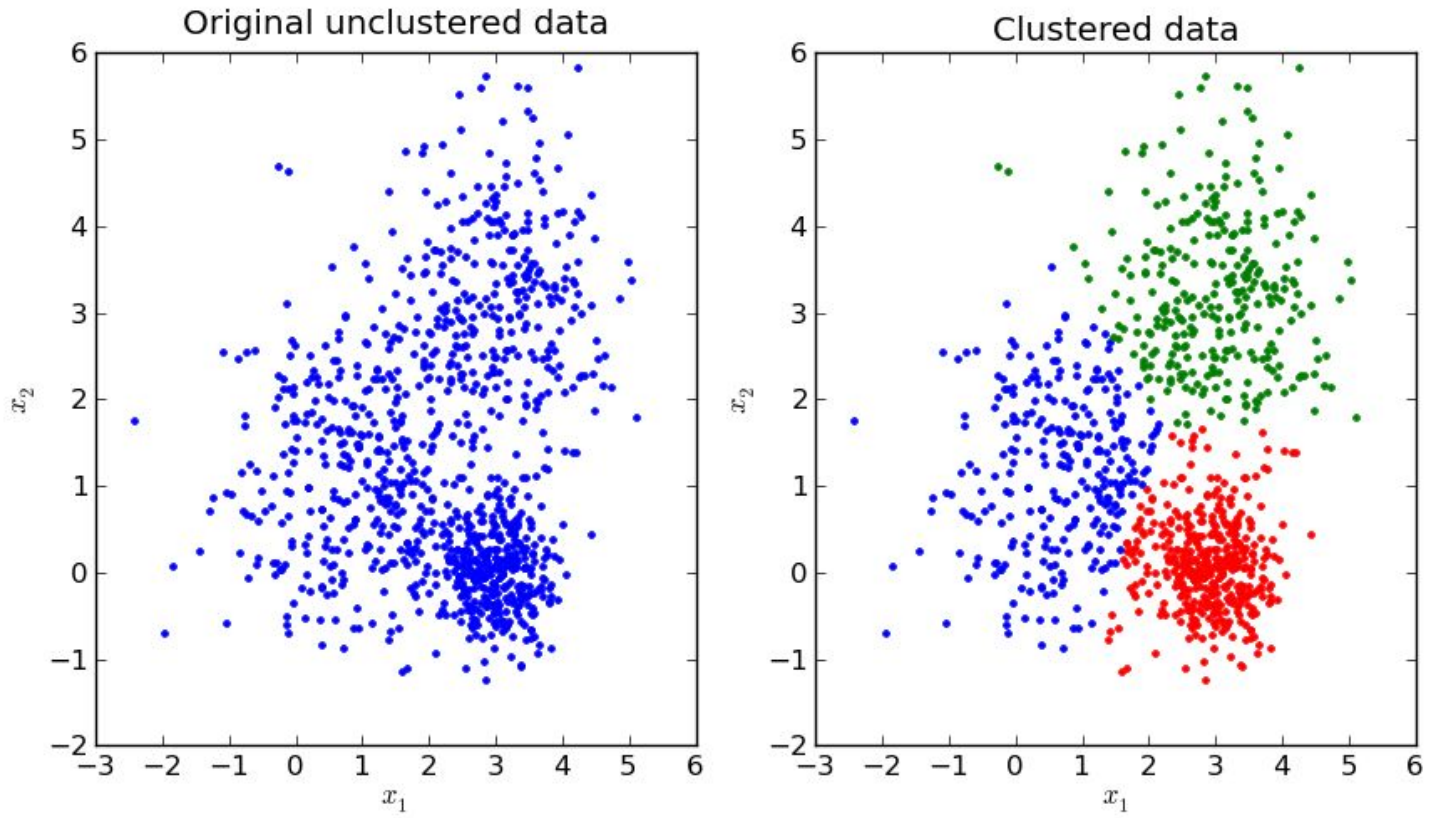
# Some Basic Terminology

**Features/ Attributes**

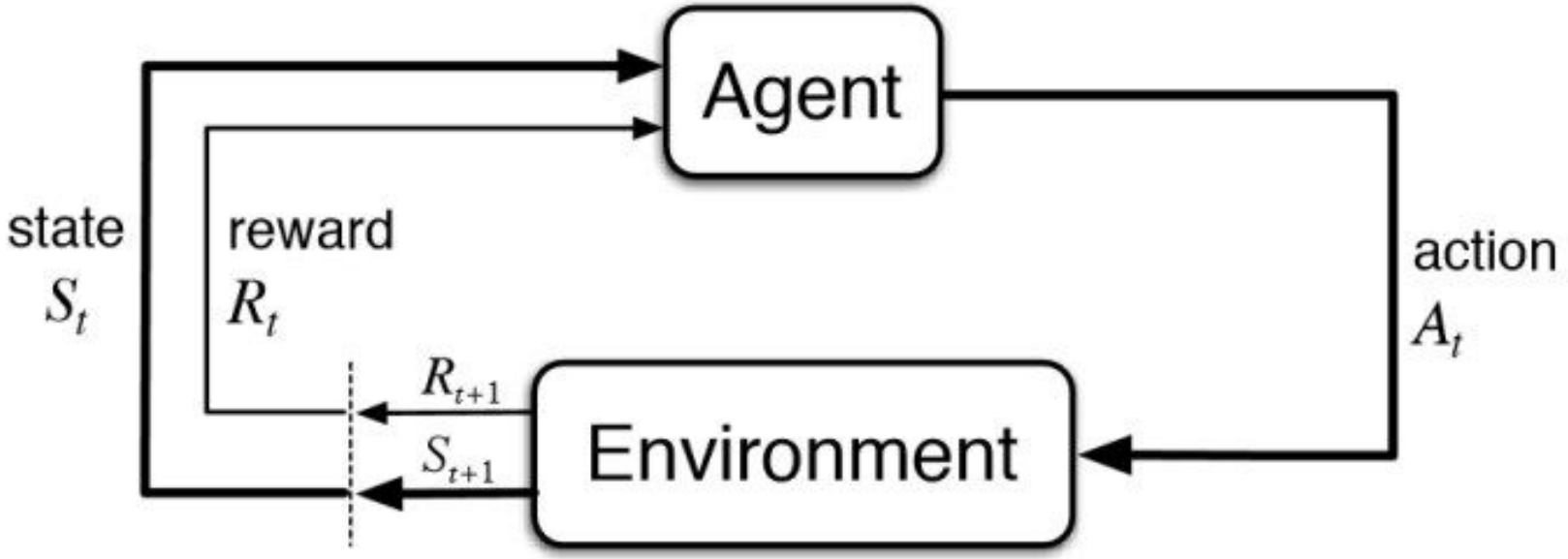
**Target Variable**

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 | setosa |
| 15 | 5.8 | 4.0 | 1.2 | 0.2 | setosa |

# 2) Unsupervised Learning



# 3) Reinforcement Learning



# Machine Learning Pipeline

# The ML Process

## 1 Identify Problem

Carefully define the problem you want to solve. What specific question are you trying to answer?

## 2 Gather Data

Figure out what data is needed and where to retrieve it. Does similar data exist or do we need to generate it?

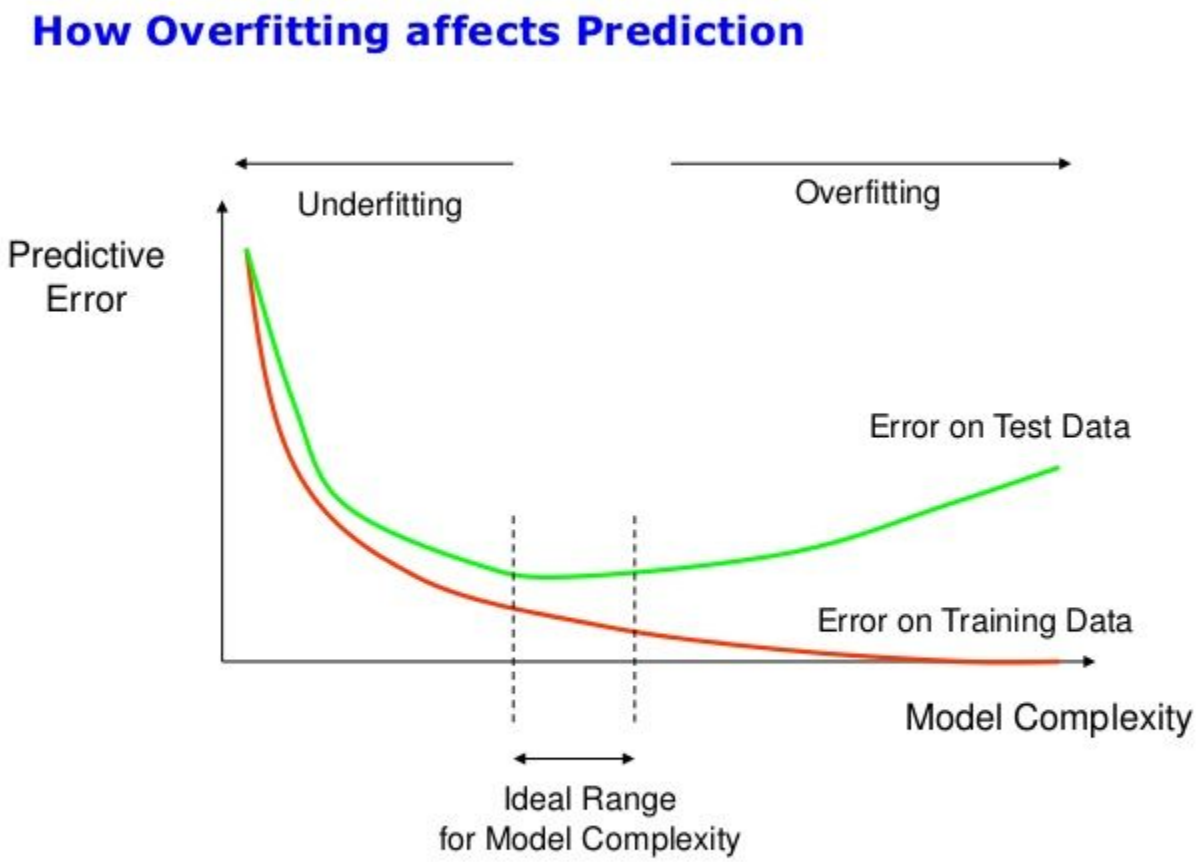
## 3 Process Data

Format data that can be interpreted by a computer. That includes cleaning, manipulating and extracting important features to feed into the training model.

# Train-test Split

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 13 | 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 14 | 4.3 | 3.0 | 1.1 | 0.1 | setosa |
| 15 | 5.8 | 4.0 | 1.2 | 0.2 | setosa |

# Overfitting vs Underfitting




https://www.google.ca/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjP3cC_-73dAhXq8YMKHbCjDhUQjRx6BAgBEAU&url=https%3A%2F%2Fstats.stackexchange.com%2Fquestions%2F292283%2Fgeneral-question-regarding-over-fitting-vs-complexity-of-models&psig=AOvVaw2w0fUI83QmZj2Wm8HNcyXd&ust=1537133719302285

# The ML Process (Continued)

### 4 Train Model

Training the dataset on your selected model. In practice, datasets are split into train, validation and test sets in order to measure model performance.

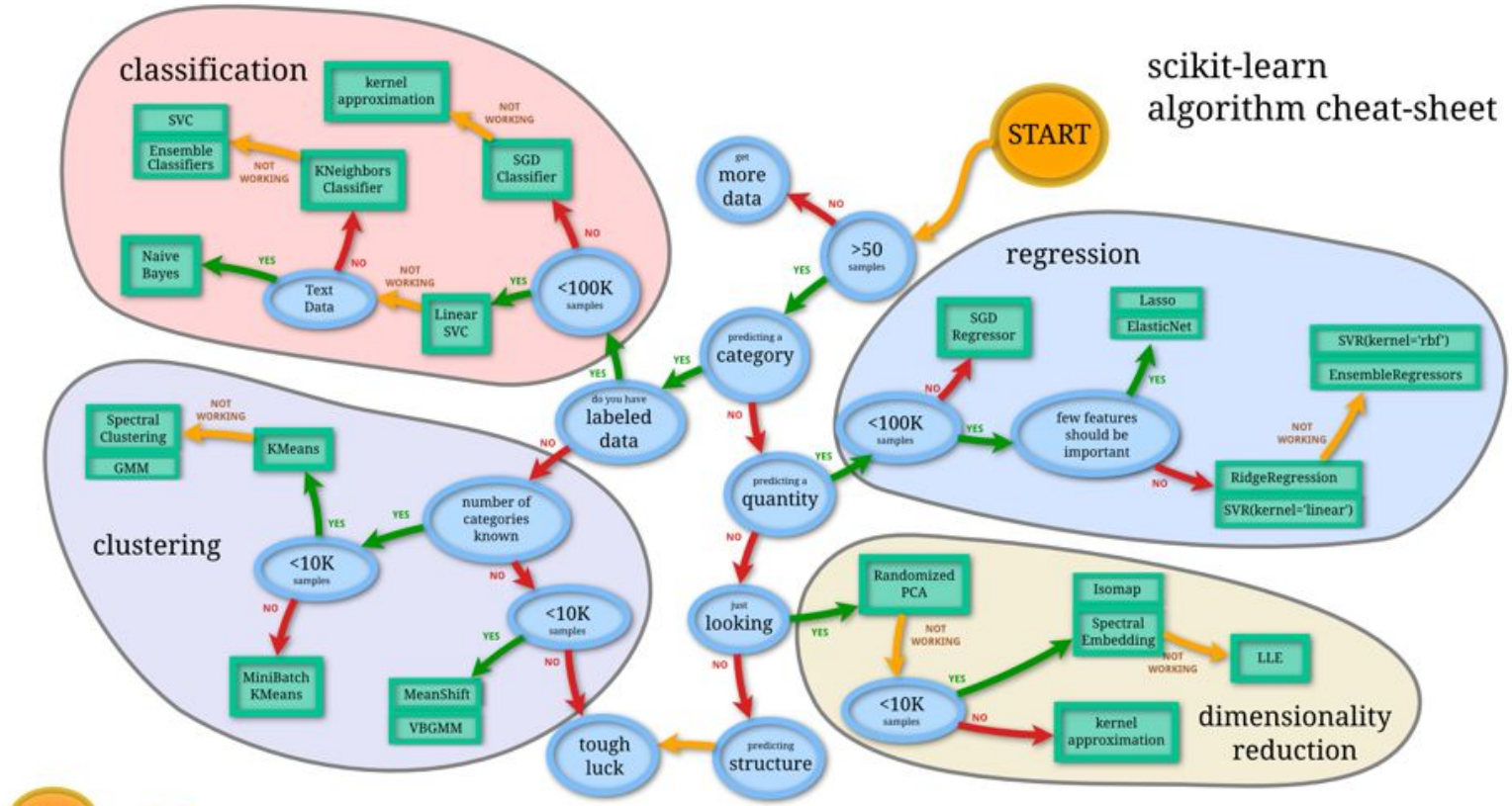
### 5 Evaluate Results

Does the trained model solve your initial problem? Does it satisfy your performance requirements?

### 6 Repeat!

Improve your model by reiterating the process!

# Choosing A Model



# K-Fold Cross Validation

# Python and Numpy Introduction/Demo

# Scikit Learn Introduction/Demo

# Thanks!

## Any questions?

Reminders:

Homework 1 and deliverable 1 due before next lecture.