Intro to Data Science

Connor Watson

Outline

I. Define Data Science
II. Data
    A. Big Data and Distributed Computing
III. Visualization
IV. Machine Learning
    A. Supervised
    B. Unsupervised

Data Science

- Data Science (DS) - The study of using data efficiently allowing humans/machines to make informed decisions.
  - Businesses make important product decisions.
  - Computers learn their environments.

Applications

- Finance - Using stock data to buy/sell
- Gaming - Determine which characters are least popular
- Recommendation Systems
- Chatbots
- Healthcare - Predicting diagnosis

Data

- Structured Data - Table based information
  - Rows - observations
  - Columns - features / attributes

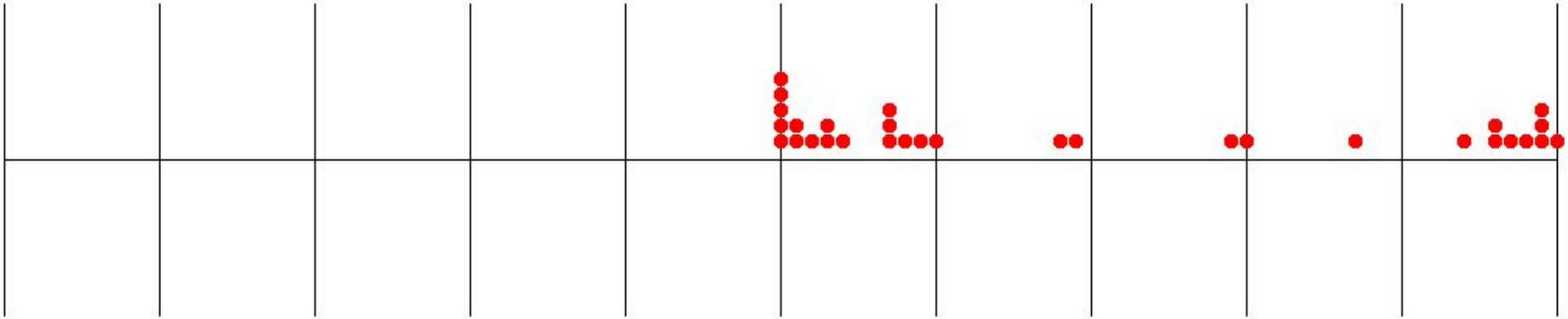
	First	Last	Age
0	Jane	Doe	23
1	Terell	Smith	24
2	Elizabeth	Chen	22
3	Nishant	Patel	25

Stats Review

- grades = [99, 75, 77, 57, 58, 59, 51, 68, 69, 60, 90, 92, 50, 51, 54, 53, 52, 53, 57, 50, 50, 57, 96, 98, 99, 100, 87, 94, 97, 50, 50]
- Average grade is a 70 (C) ...decent
- Median grade is a 59 (F) ...failing
  - Average is sensitive to outliers (weighted)
  - Median shows measure of center/distribution
  - At least 50% of students failed → 17/31 students failed

5 Number Summary

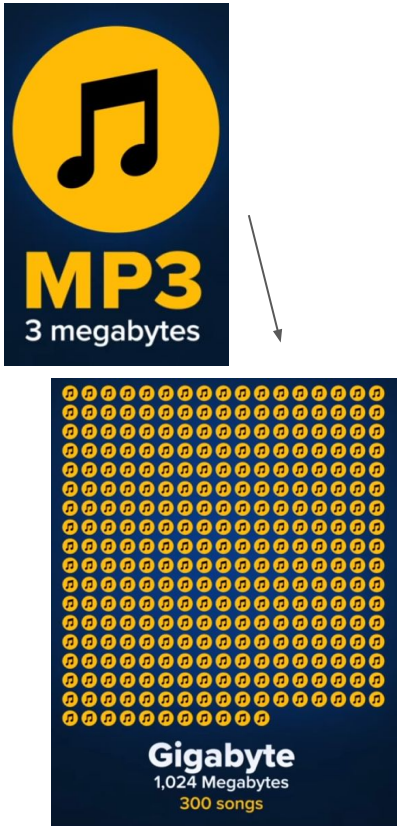
- Max = 100
- Q3 = 87
- Q2 (Median) = 59
- Q1 = 52
- Min = 50



Big Data

- How “big” is Big Data?
  - MB? GB? TB?
- Companies have LOTS of data
- How do we efficiently handle it all?

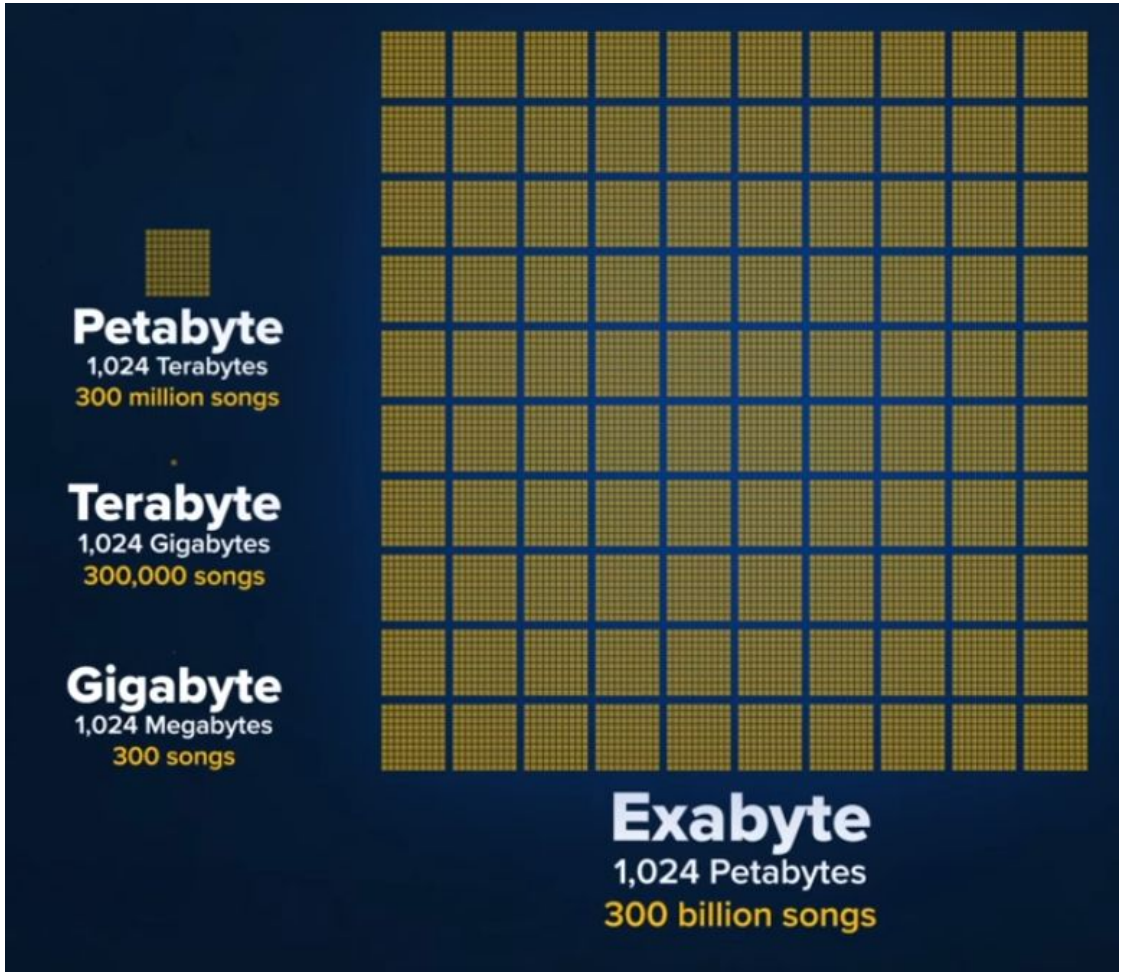
Big Data



Big Data

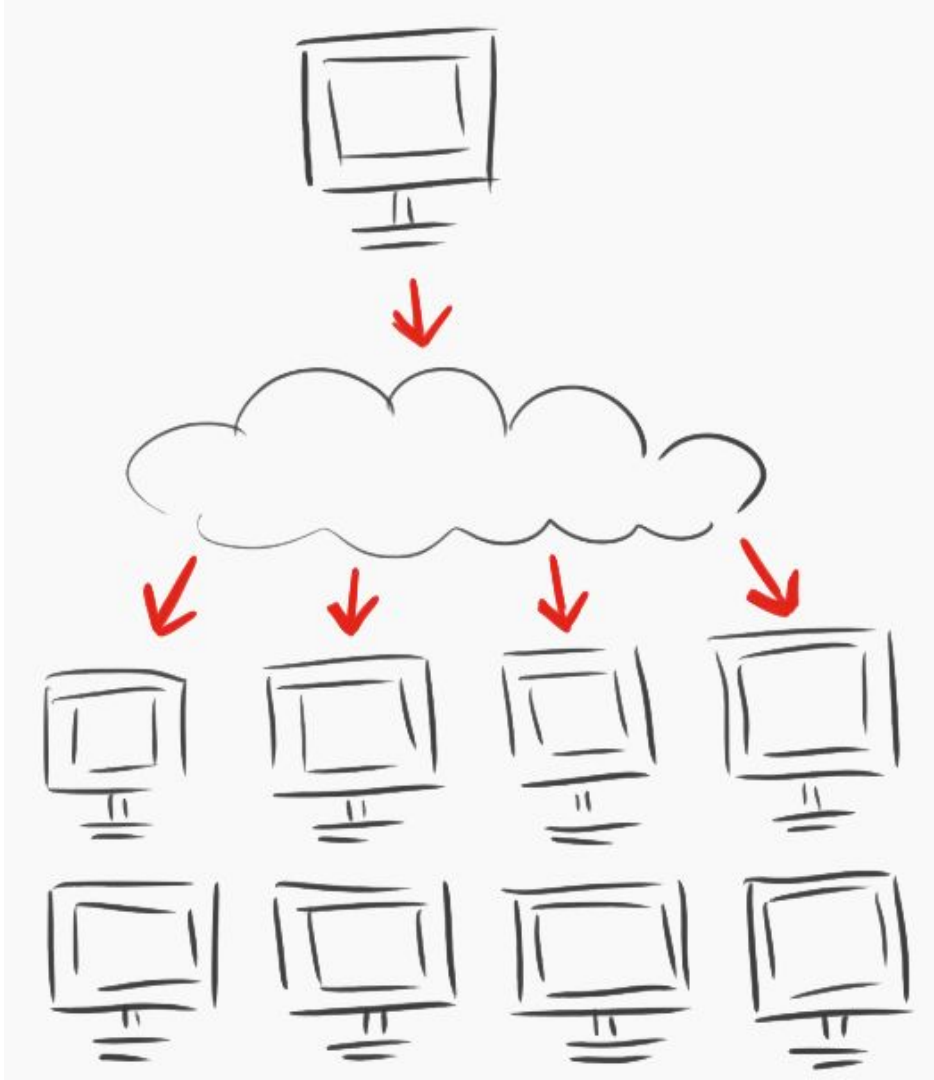

MP3
3 megabytes


Gigabyte
1,024 Megabytes
300 songs

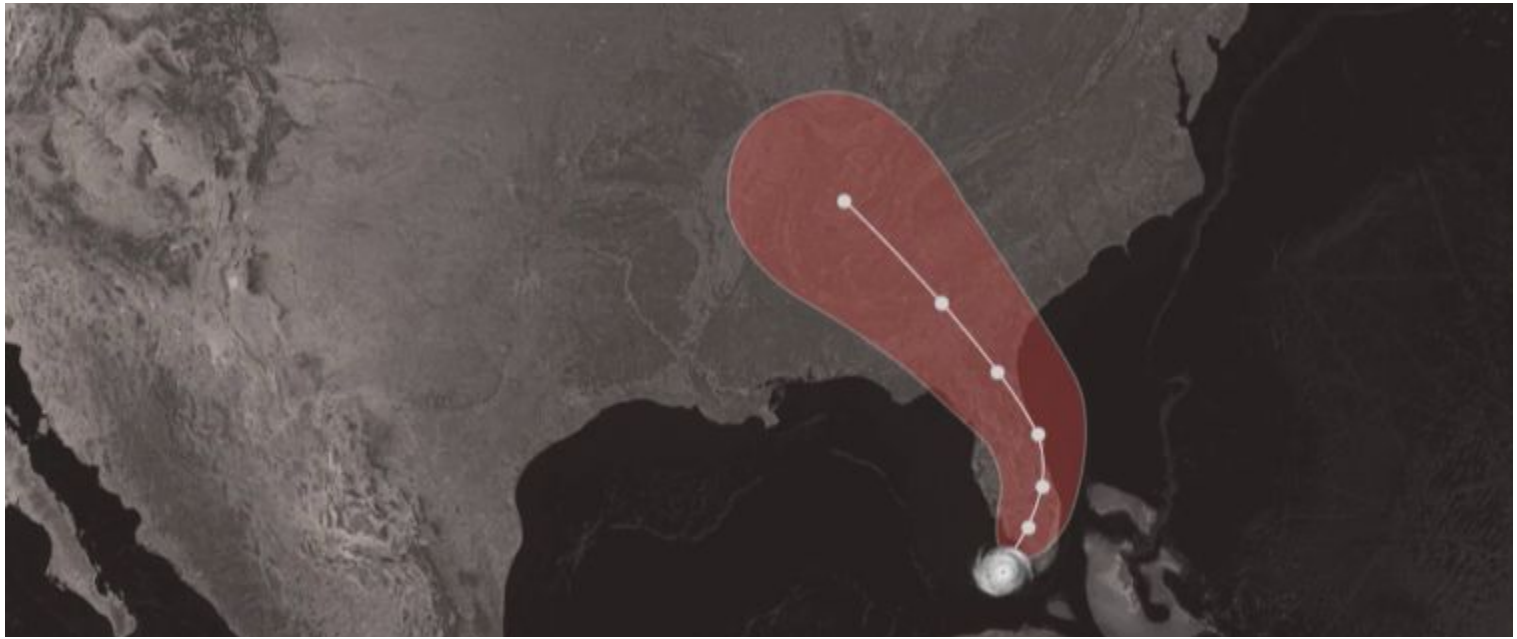

Petabyte
1,024 Terabytes
300 million songs
Terabyte
1,024 Gigabytes
300,000 songs
Gigabyte
1,024 Megabytes
300 songs
Exabyte
1,024 Petabytes
300 billion songs

Distributed Computing

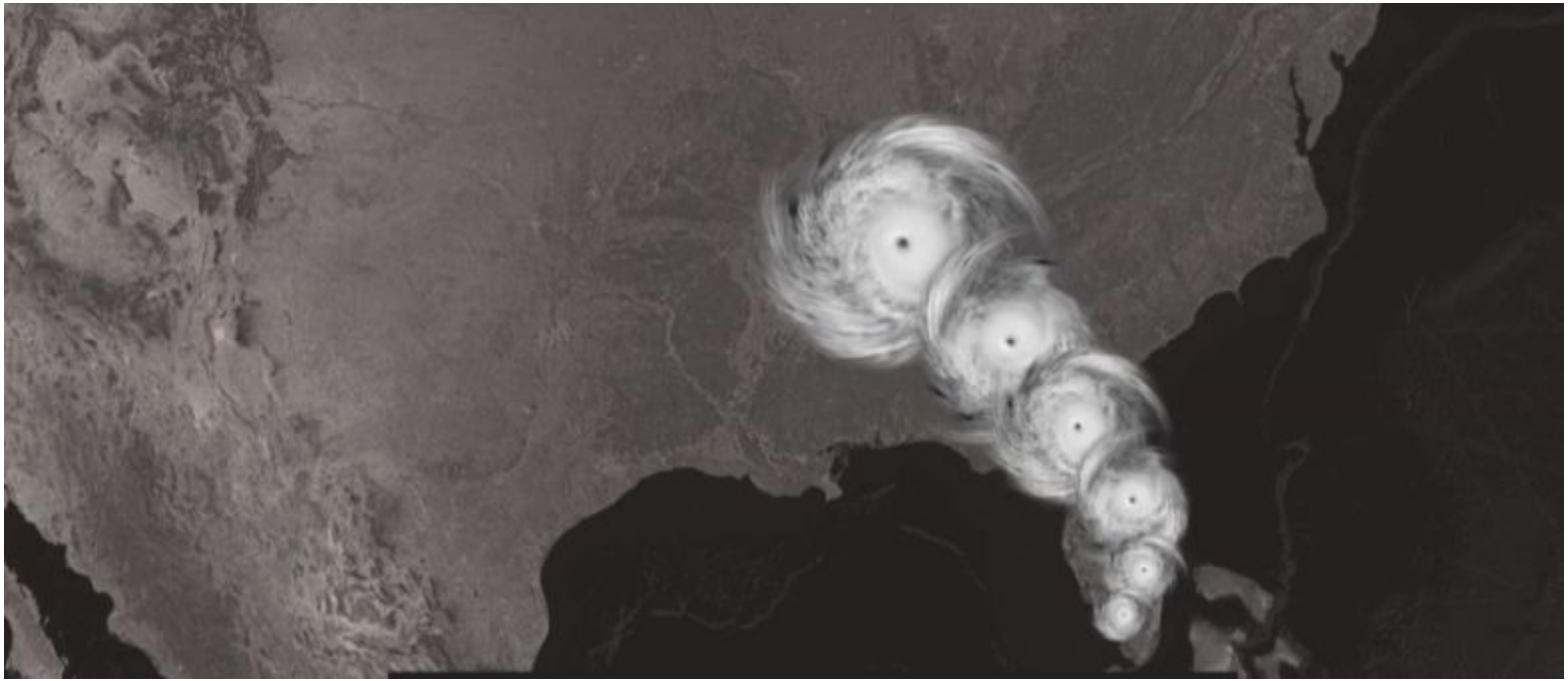
- NJIT Kong
  - https://ist.njit.edu/high-performance-computing-kong/
- Worker nodes handle small tasks
- Head node combines result



Data Visualization



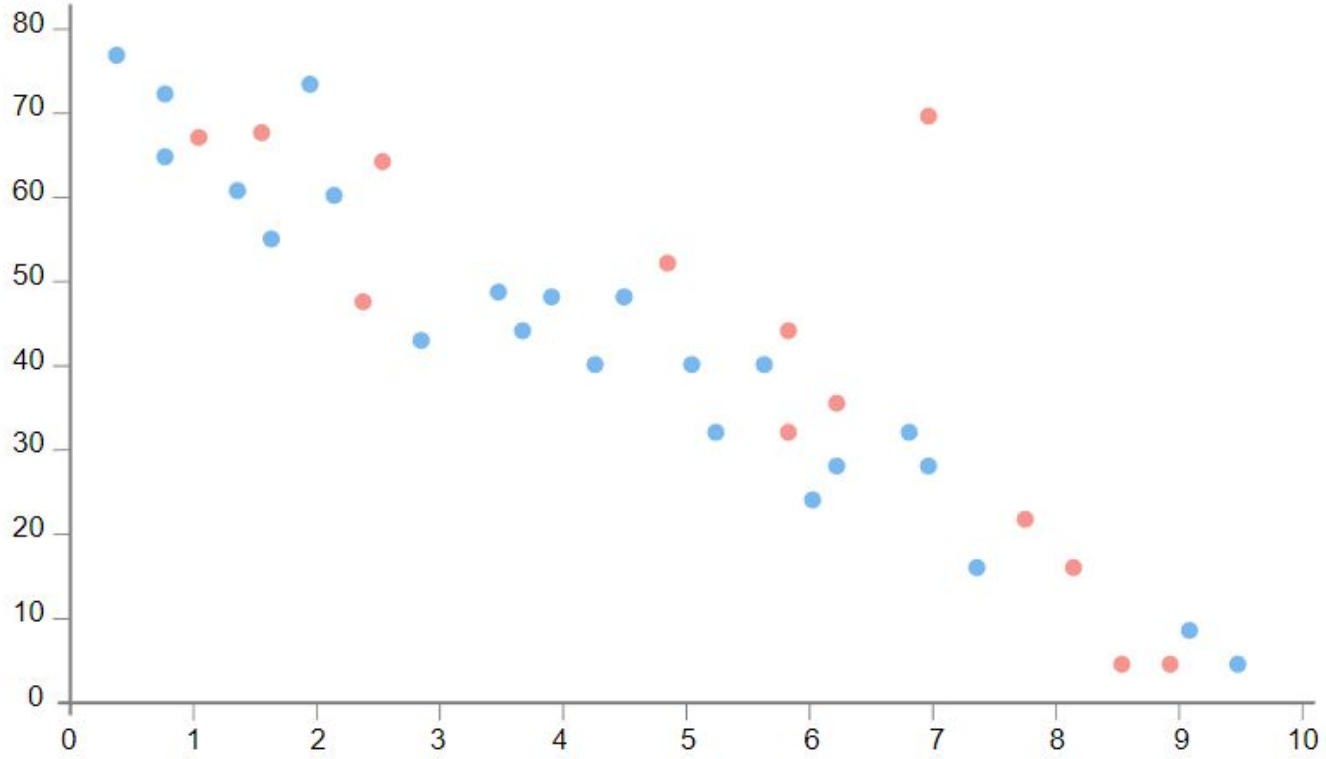
Data Visualization



Data Visualization

Data Visualization

Scatter Plot



Bar Chart

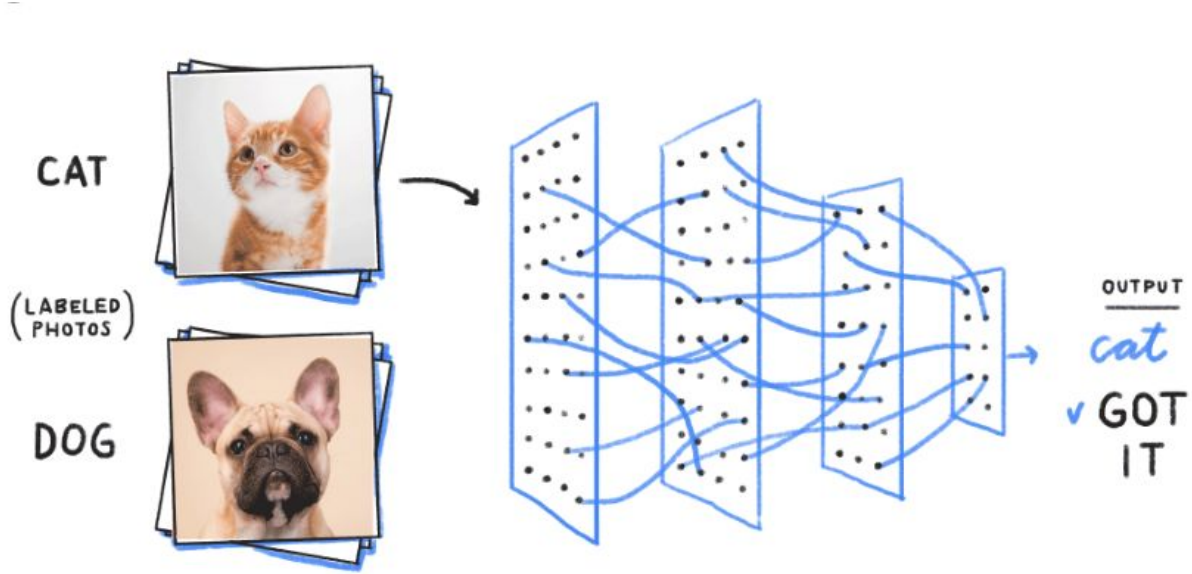

Excuses for being late to class
I had no clean pants to wear
I thought it was Saturday
I forgot to set my alarm
It was still too dark, I thought it was still nighttime
I got stuck in traffic
0
5
10
15
20
25
30

Learning

- Supervised
  - Hyperplane
  - SVM
  - Linear Regression
- Unsupervised
  - K Nearest Means

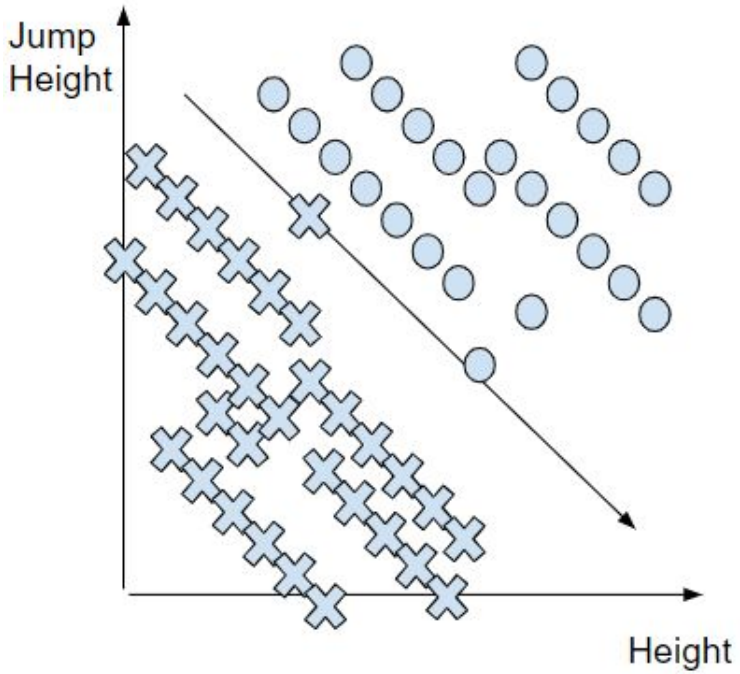
Supervised Learning

- Knowledge about the data is known…
  - Can we **classify** new data points?



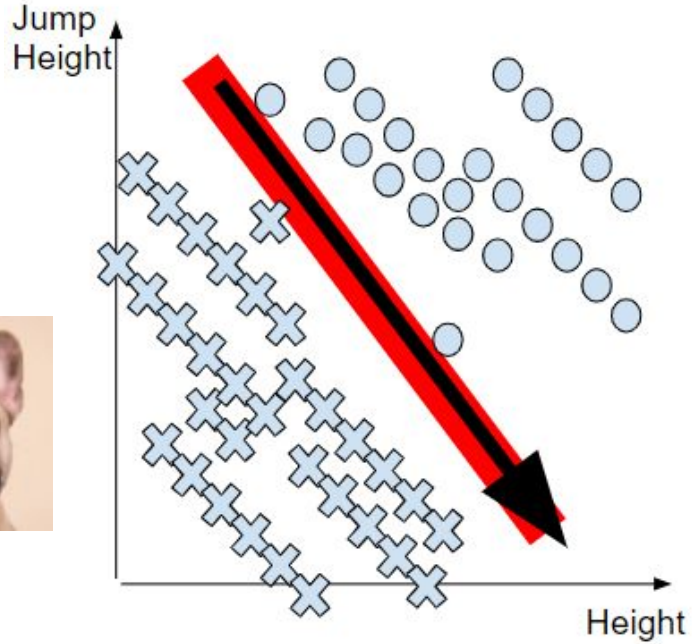
Classification - Hyperplane

- Draw a line between two classes
- Left side is class 1 (dog)
- Right side is class 2 (cat)



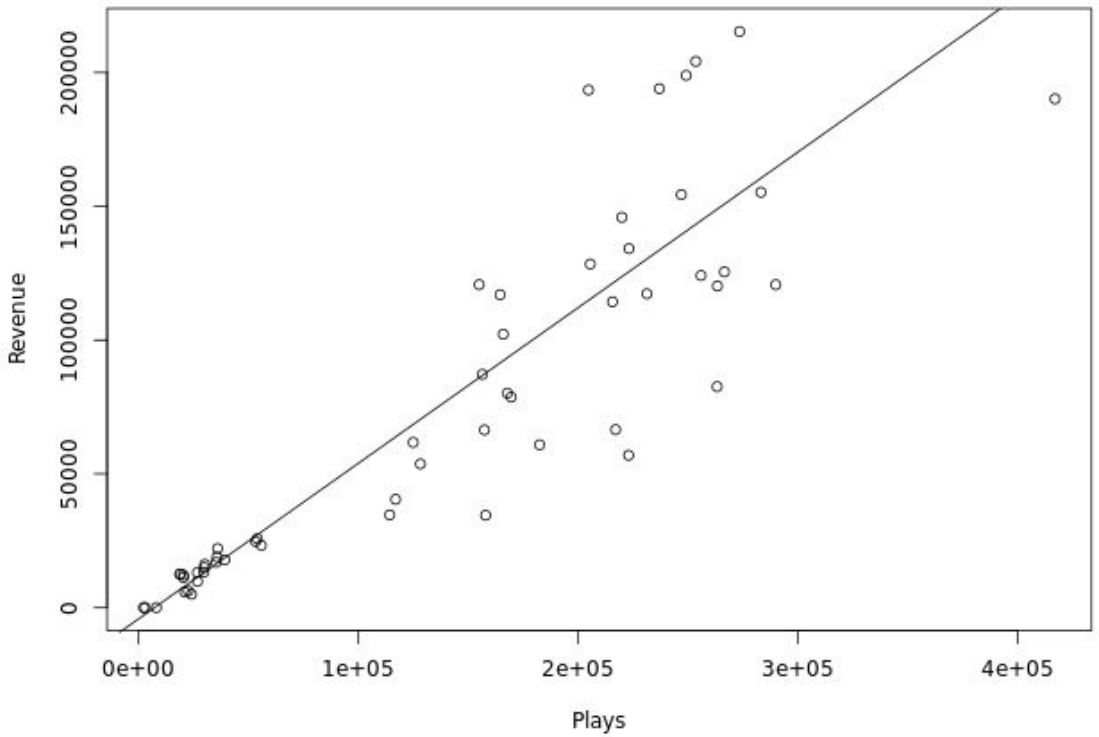
Classification - Support Vector Machine

- Introduce a **Decision Boundary**
- Separates points based on their distance to the hyperplane



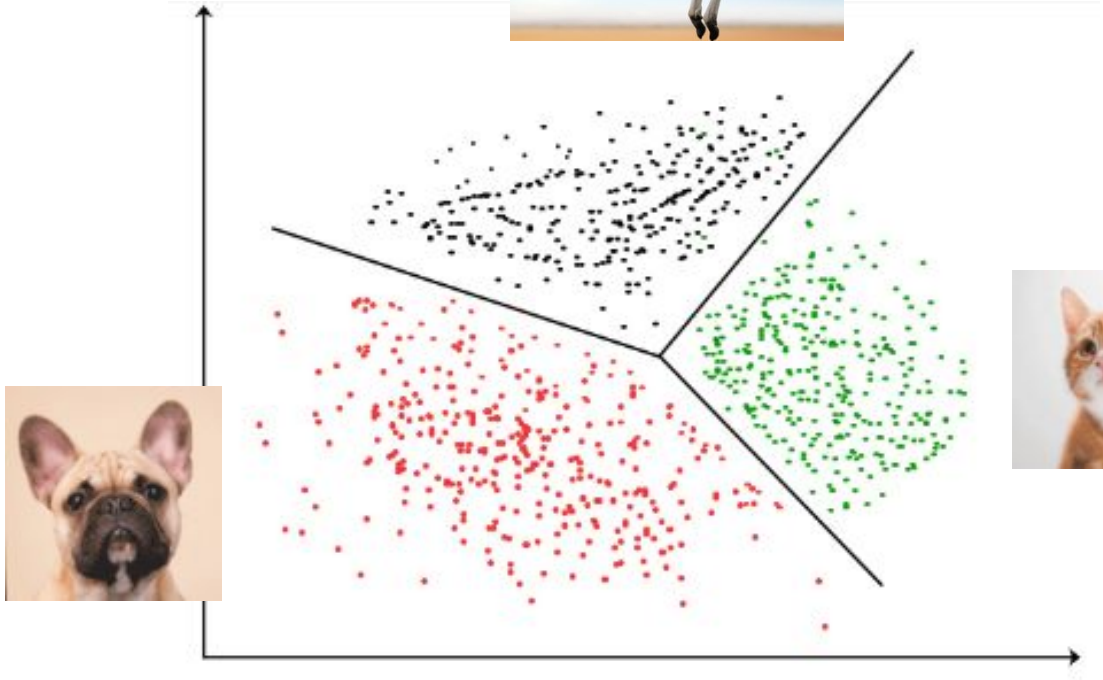
Regression - Predicting Continuous Values

- Given number of plays:
  - Can we predict revenue?
- x axis: explanatory
  - independent
- y axis: response
  - dependent



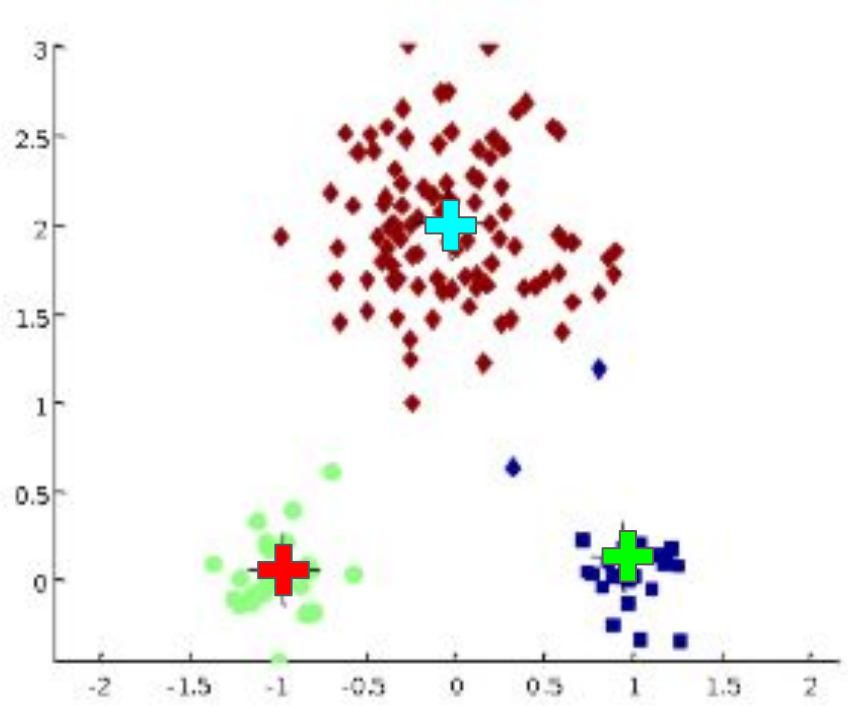
Unsupervised Learning

- Knowledge about the data is NOT known…
  - Can we **cluster** data points to learn their behavior?



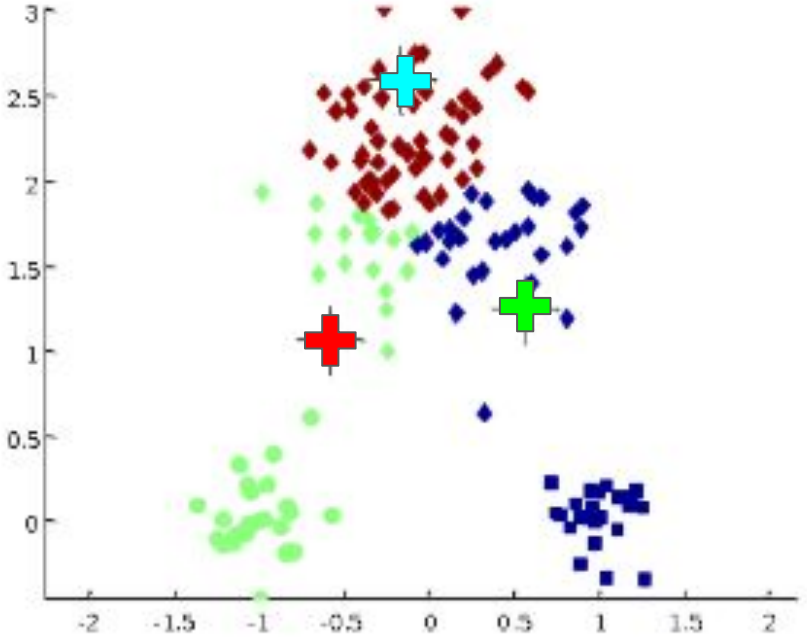
k Means Clustering

- Pick k random **centroids**
- Find the points closest to them
- Recalculate mean (new centroid)
- Mean is **outlier sensitive**



k Means Clustering

- Pick k random **centroids**
- Find the points closest to them
- Recalculate mean (new centroid)
- Mean is **outlier sensitive**



How Does This Help You?

- Build an interactive data viz tool!
  - React, D3.js, Dash by plotly
- Try a Kaggle competition!
  - Classification problems
- Clean a data set and present some statistics!

NJIT Data Science

https://join.slack.com/t/njitdatascienceclub/shared_invite/enQtNzEzMzc4Mjk1ODQ2LWI1ZWQ5NjcyZjJlYTBhM2EyYmY2ODQzZjQ3MmM0NzFhNWY3YTYzNzMyYzYwNDc5ZGNjYmIyOGY3NWVjMGQ1OTc

- Workshops (like this one)
- Mini lectures
- Industry sponsors


THANK YOU!
ANY QUESTIONS?
makeameme.org