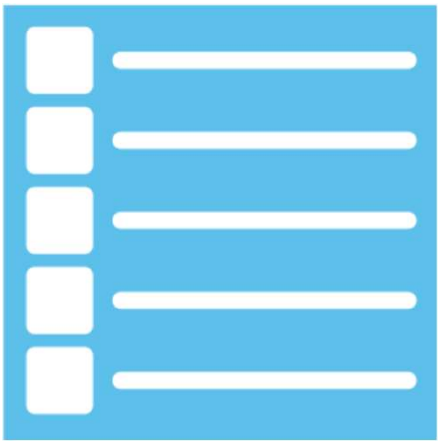
CS2073: Comp. Prog. w Eng. App.

W12: Intro to Data Science (MITRE GenAI Project)

Hamidreza Moradi

Agenda



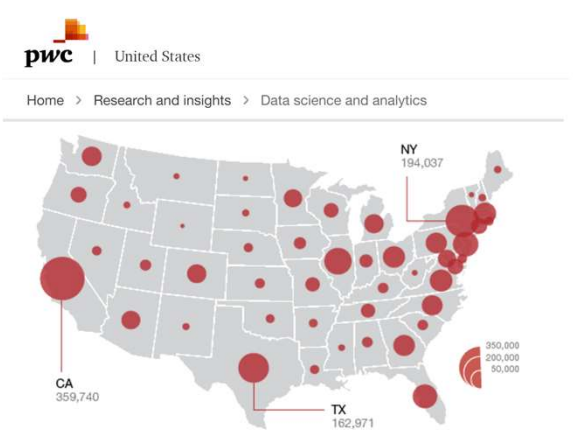
- **Introduction into data science**
- **Python programming language**
- **Machine Learning algorithms**
- **K-means algorithm**
- **Python code to train a K-means model**

MITRE

- Gives students access to AI training, tools, and big data.
  - One session about machine-learning / data science
    - A homework (in-class assignment) in python

https://www.mitre.org/publications/project-stories/preparing-the-next-generation-for-the-fourth-industrial-revolution

Data Scientists are in high demand



Forbes

28,058 views | Dec 11, 2017, 08:32pm

LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning

IBM

THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000
DSA job listings in 2015

By 2020, DSA job openings are projected to grow
15%

364,000
Additional job listings projected in 2020

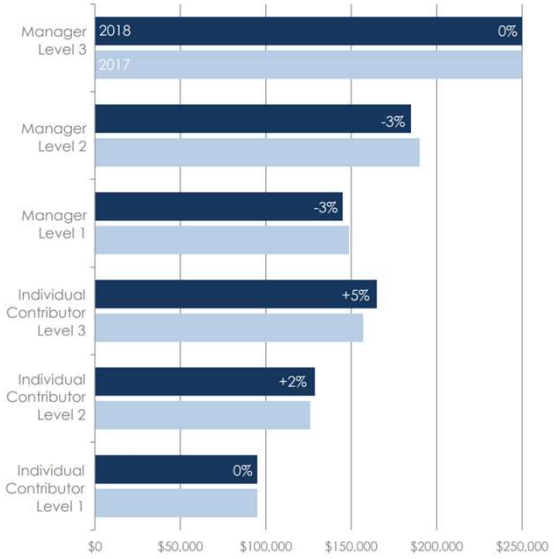
Demand for both Data Scientists and Data Engineers is projected to grow
39%

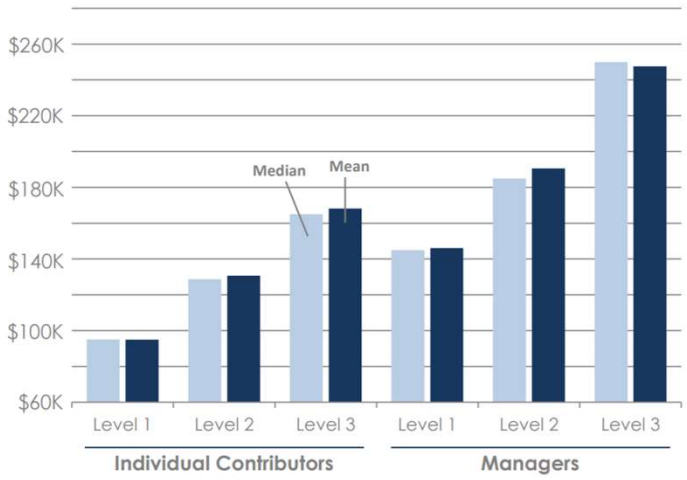
DSA jobs remain open
5 days
longer than average

DSA jobs advertise average salaries of
$80,265
With a premium over all BA+ jobs of
$8,736

81%
Of DSA jobs require workers with 3-5 years of experience or more

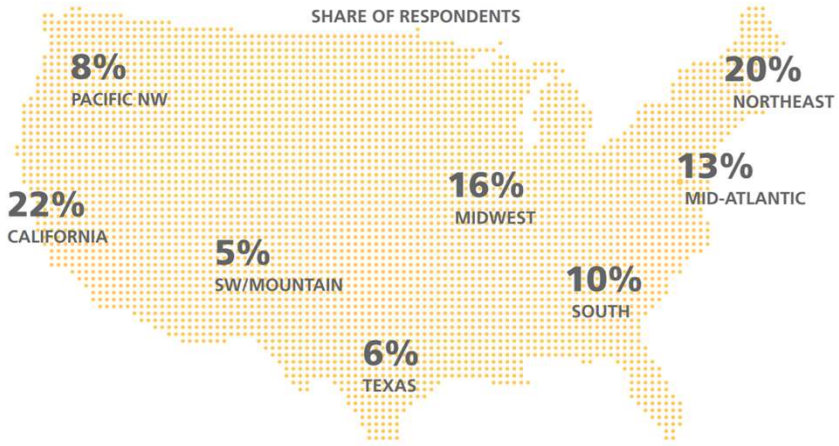
Paid well

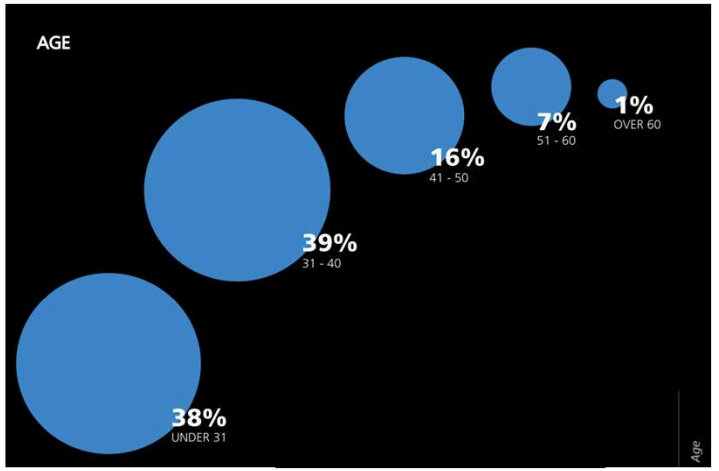




https://www.burtchworks.com/wp-content/uploads/2018/05/Burtch-Works-Study_DS-2018.pdf

...

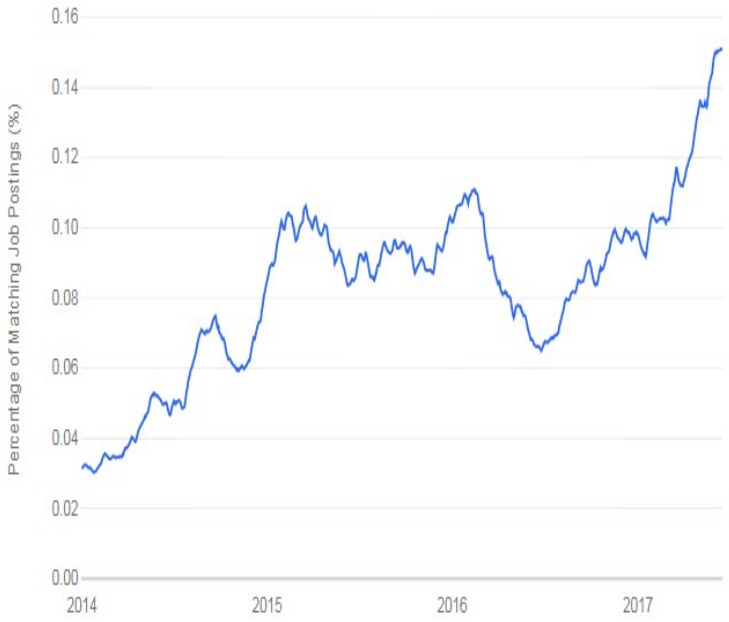




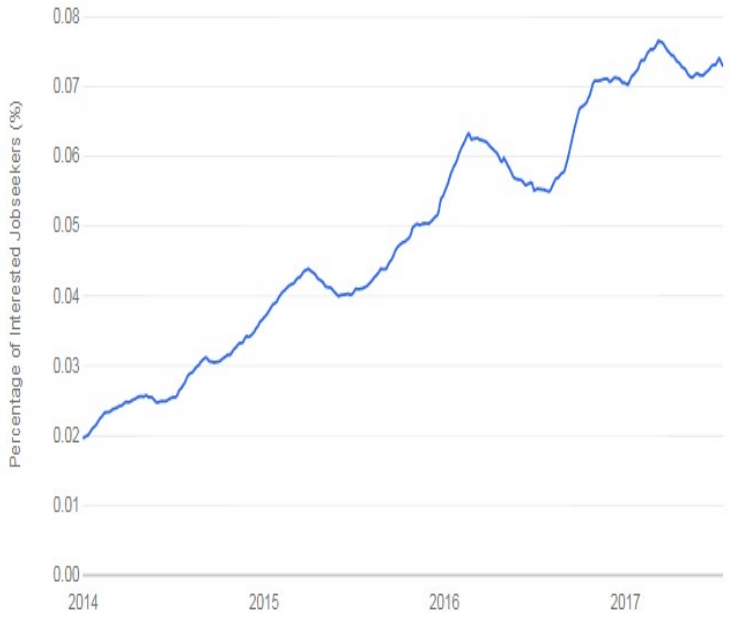
https://www.oreilly.com/data/free/files/2016-data-science-salary-survey.pdf

Data Scientist Job Demand

Job postings



Jobseeker interest



http://blog.indeed.com/

Why the demand is increasing ?

- Increase in the amount of data and computation power
- The need to clean, process, analyze, and generate meaningful insight from existing data.

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of disease .

More Examples

- Recommender systems → NetFlix, Amazon, YouTube.
- Real-time Monitoring → Smart Home, Internet of Things.
- Consumer Satisfaction → Social Media Data (Facebook, Twitter)

Data Science Definition

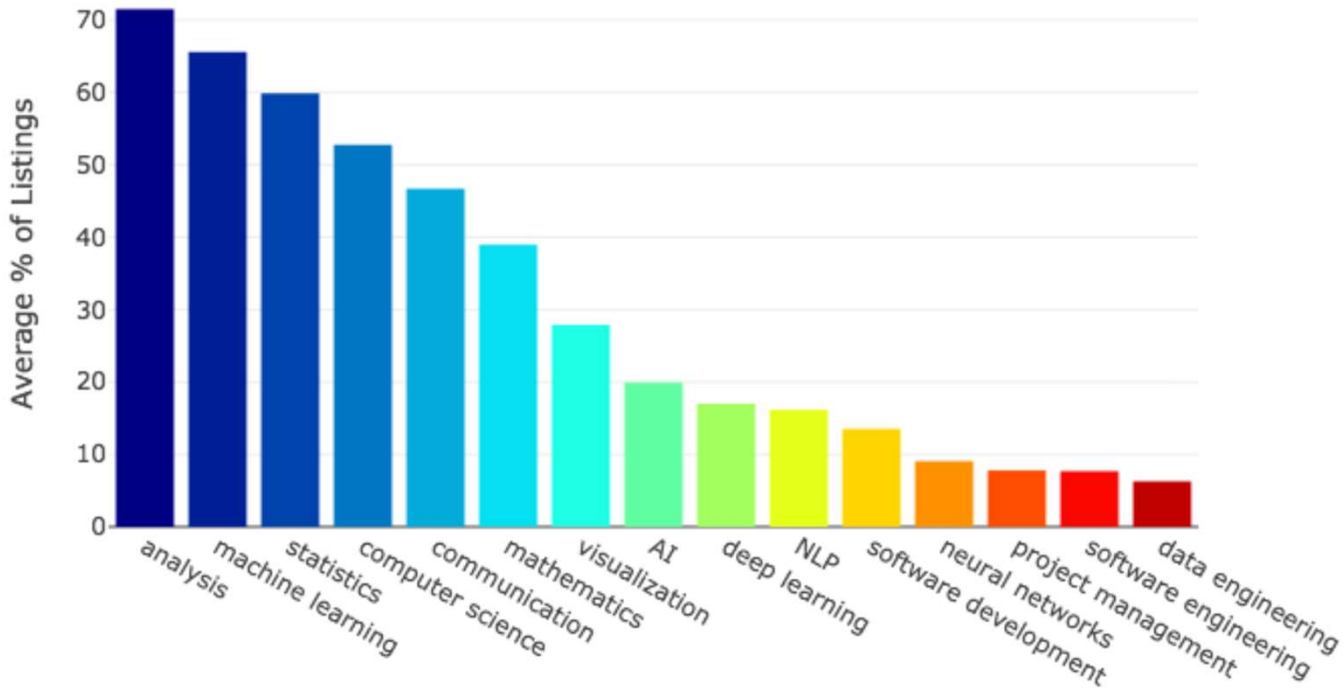
Data Science is the science which:

Uses: computer science, statistics and machine learning, and visualization

To: collect, clean, integrate, analyze, and visualize the data

With the goal of: Extracting knowledge and insights.

Data Scientists Skills



General Skills

https://www.kdnuggets.com/2018/11/most-demand-skills-data-scientists.html

Python language of preference for Data Scientist

- Invented in early 90s by Guido van Rossum
- Open source
- Considered a scripting language
  - No compilation needed
  - Scripts are evaluated by the interpreter, line by line

Variables and objects

- Variables are created the first time it is assigned a value
  - No need to declare type
  - Types are associated at initialization
    - x = 5
    - x = [1, 3, 5]
    - x = ‘python’
  - Assignment creates *references*, not *copies*

    X = [1, 3, 5]

    Y= X

    X[0] = 2

    Print (Y) # Y is [2, 3, 5]

Formatting

- Instead of curly braces to delimit blocks of code, **Python uses indentation.**
  - Incorrect indentation causes error.
- Comments start with #
- Colons start a new block in many constructs, e.g. function definitions, if-elif clause, for, while

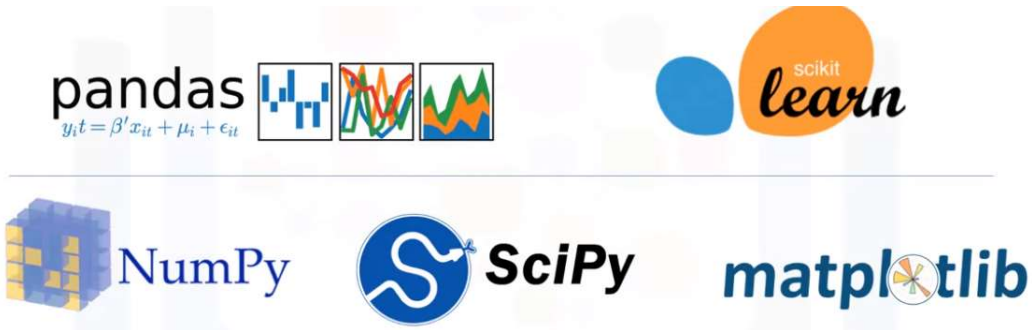
```
1 user_string = ' intro to data science '
2 int_number = 10
3 float_number  = 10.0
4 for i in [0, 1, 2, 3, 4]:
5         for j in range(5):
6                 print (j+i) # print sum of i and j as number
7                 print (j+" "+i) # print a string with values
8
9 print ("done looping")
```

Modules

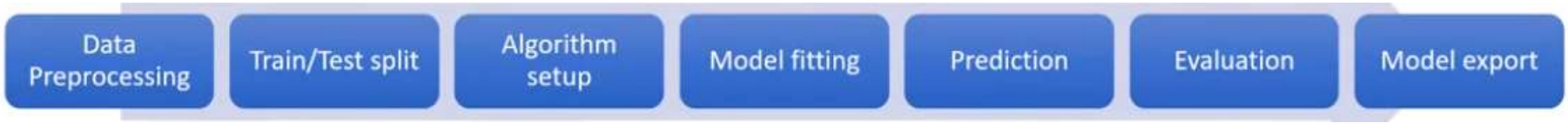
- Certain features of Python are not loaded by default
  - In order to use these features, you'll need to import the modules.
- E.g.

  import matplotlib.pyplot as plt

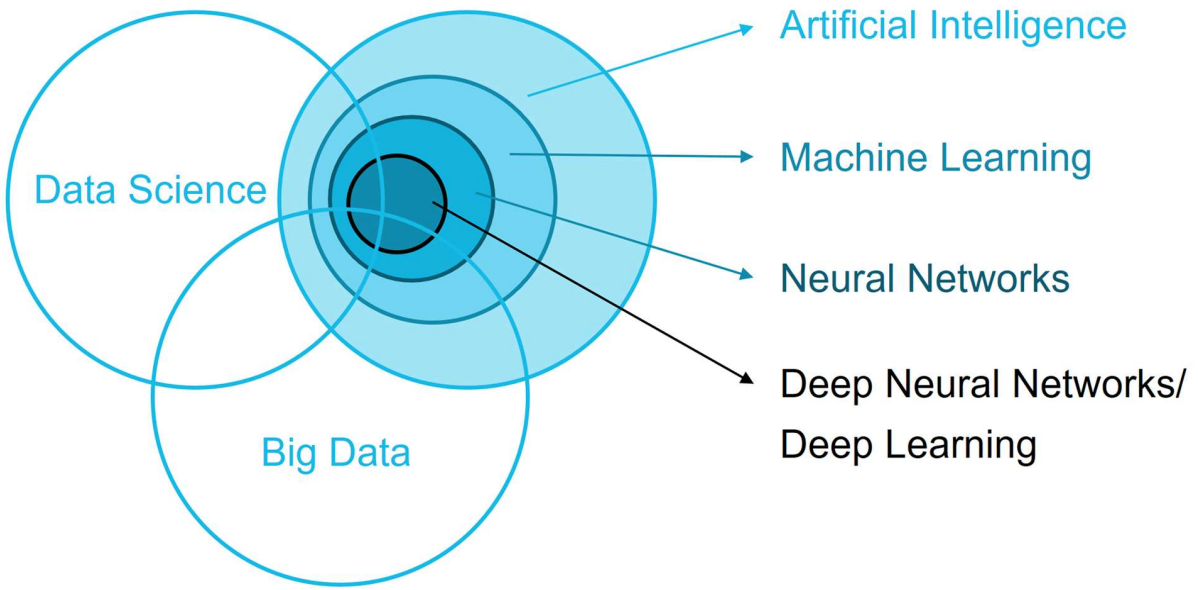
  import numpy as np



Stages of a Data Science Task



Data Scientists Roll, data and AI



https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579

Machine Learning

- The science of getting computers to accomplish a task without being explicitly programmed about how to do the task.
- It applies algorithms that can learn from data to make decisions.
  - Example applications: self-driving cars, face recognition software, voice recognition, autonomous drones, detect credit card frauds, etc.

Machine Learning Types

- Supervised Learning
  - Classification: determining what category something belongs to, after seeing a number of examples of things belonging to several categories.
  - Regression: learn a function that describes the relationship between inputs and outputs and predicting how the output change with change in inputs.

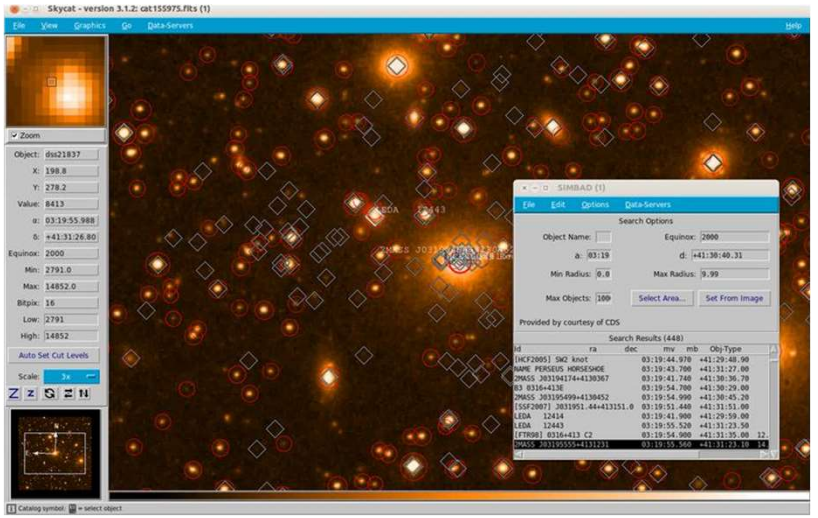
- Unsupervised Learning
  - Finding patterns in data without explicit labels in the training examples

K-Means Clustering

- Algorithm to group objects (data points) based on similarity of attributes/features into k groups.
- If you have label for the data, it will be considered classification algorithm.
  - Cyber security: Detecting suspicious activities (potential cyber attacks) recorded in log files.
  - Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
  - Banking: Identifying fraudulent credit card transactions, risky loan applications, etc.
  - Many more..

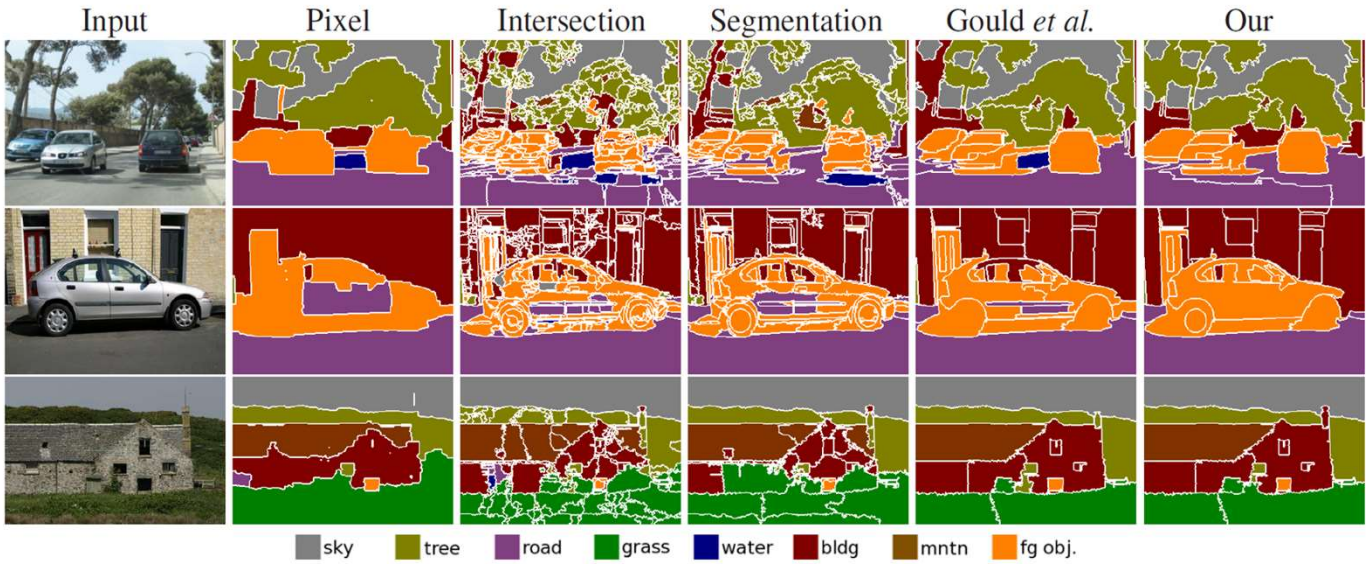
Applications of Clustering

- Astronomy
  - SkyCat: Clustered sky objects into stars, galaxies, quasars, etc based on radiation emitted in different spectrum bands.



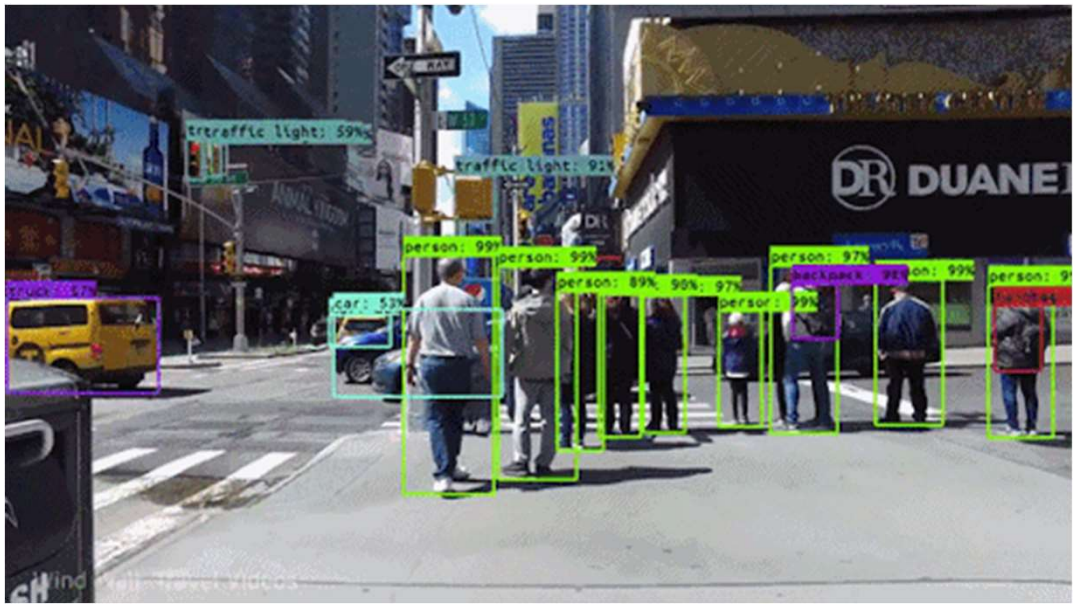
Applications of Clustering

- Image Segmentation
  - Finding “objects” in images to focus on.



Applications of Clustering

- Image Classification
  - distinguishing different “objects” in images/video.



Applications of Clustering

- Loan Application

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Categorical Variable

- Possibility of moving to a new brand

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

- Recommend a drug

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Categorical Variable

- Services to provide

X: Independent variable

Y: Dependent variable

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Basic Idea

- Given *k*, the *k-means* algorithm consists of four steps:
  - Select initial k centroids (cluster centers) at random.
  - Assign each data point to the cluster with the nearest centroid.
  - Compute each centroid as the mean of the data points assigned to it.
  - Repeat previous 2 steps until stopping criterion is met.

How K-means works

A set of measurements

(a)

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

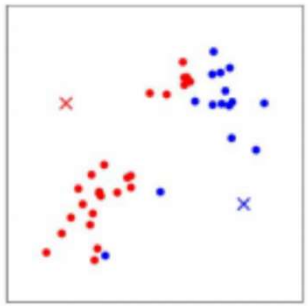
K=2 Random cluster

(b)

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

Calculating the distance of each point from the centers and assigning it to the closer cluser

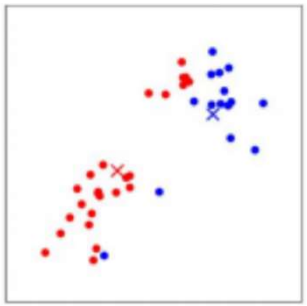


(c)

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

Take the average of points in each group and considering as the new center

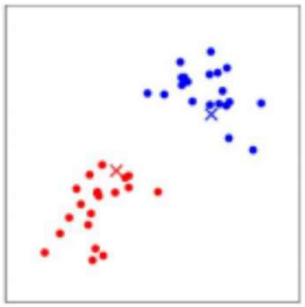


(d)

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

Repeating the process with new centers and assigning each point to the closest cluster

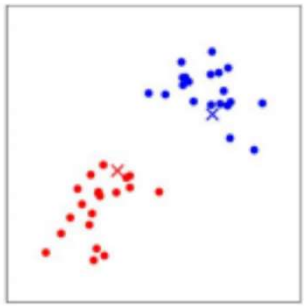


(e)

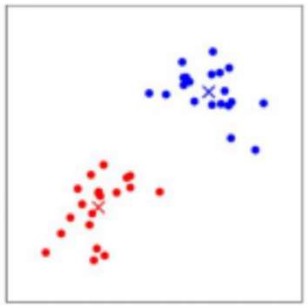
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

After couple of iteration the changes in the cluster centers become negligible



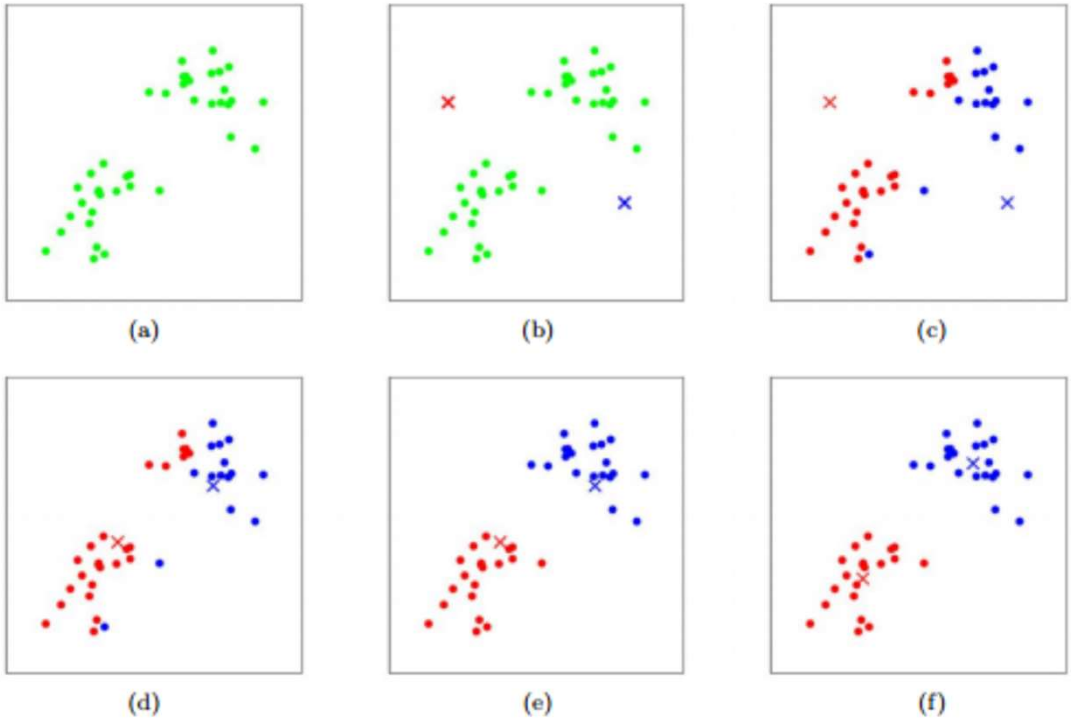
(e)



(f)

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

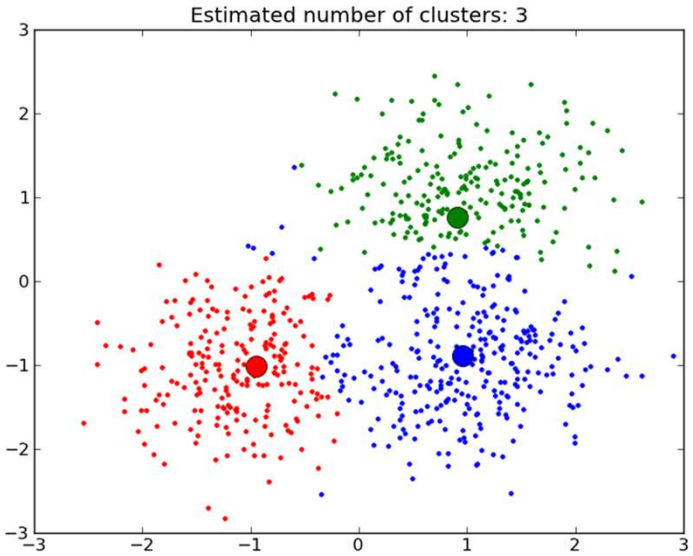


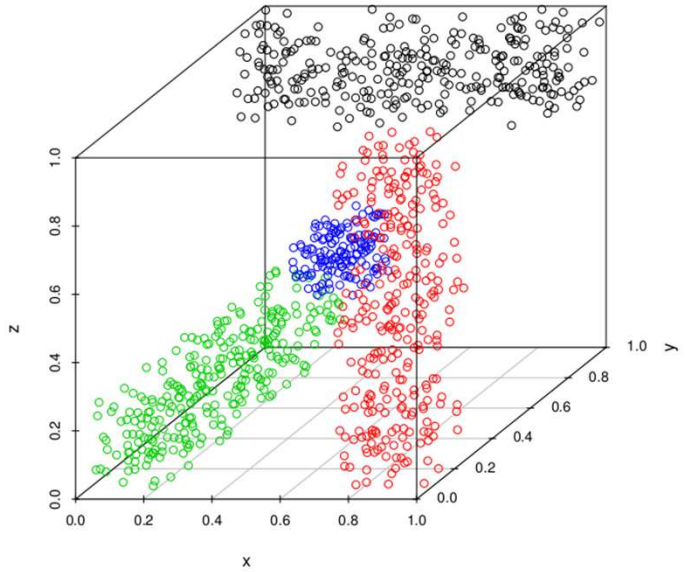
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

How K-means works

http://shabal.in/visuals/kmeans/5.html

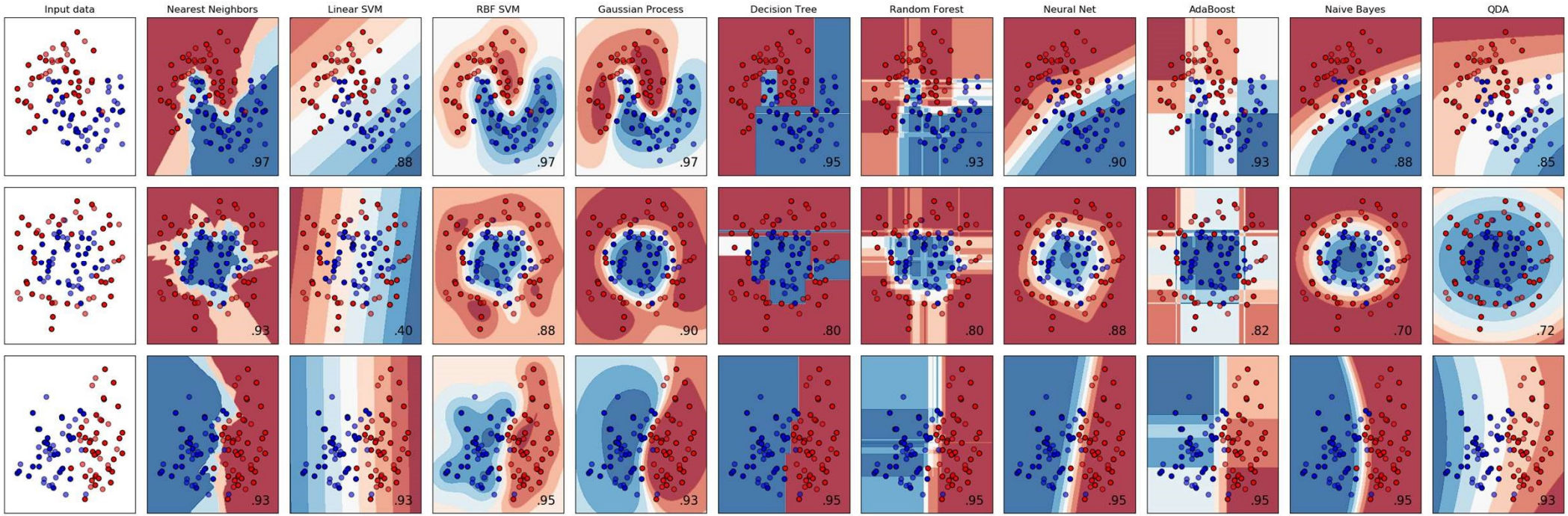
How K-means works





https://clusteval.sdu.dk/1/datasets/3351

Different algorithms and Clustering



Datasets

- Exploratory Analysis
- General Machine Learning
- Deep Learning
- Natural Language Processing
- Cloud-Based Machine Learning
- Time Series Analysis
- Recommender Systems
- Specific Industries
- Streaming Data
- Web Scraping
- Current Events

Textbooks


O'REILLY
Data Science from Scratch
FIRST PRINCIPLES WITH PYTHON
Joel Grus
Data Wrangling with Pandas, NumPy, and IPython
Python for Data Analysis
O'REILLY
Wes McKinney
Probability and Statistics for Programmers
Think Stats
O'REILLY
Allen B. Downey
O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
Jake VanderPlas

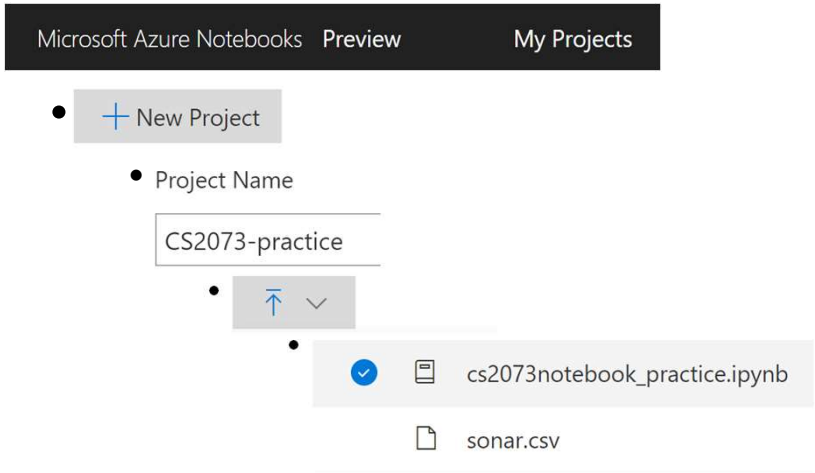
Sample Dataset: Mines vs. Rocks

- This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals.
- Goal : train a model to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.

60 features / measurements

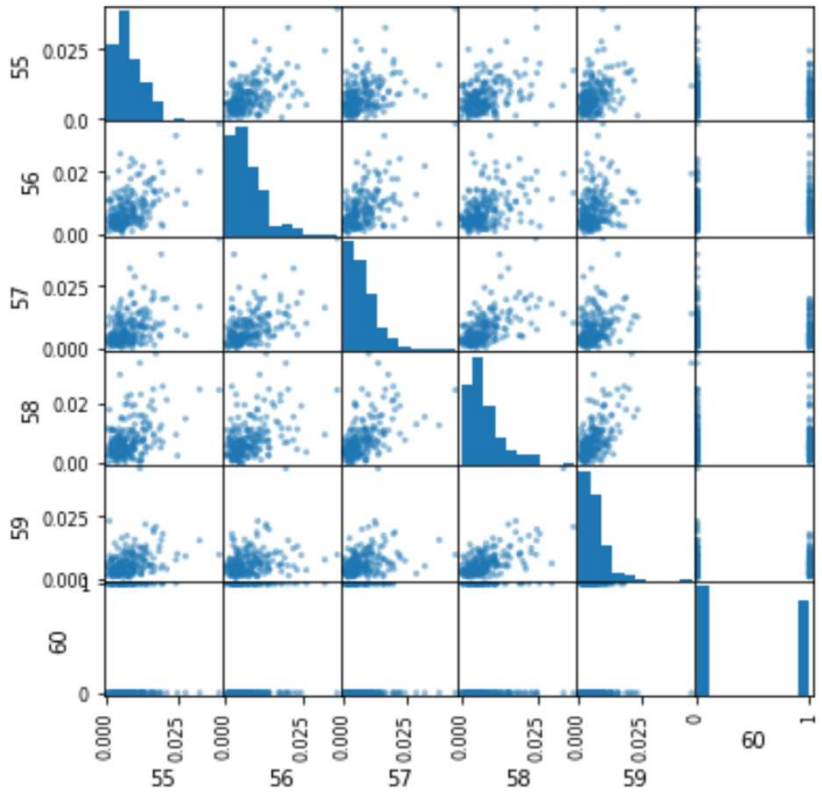
attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	attribu	Class
0.2166	0.1951	0.4947	0.4925	0.4041	0.2402	0.1392	0.1779	0.1946	0.1723	0.1522	0.0929	0.0179	0.0242	0.0083	0.0037	0.0095	0.0105	0.003	0.0132	0.0068	0.0108	0.009	Rock
0.5856	0.4993	0.2866	0.0601	0.1167	0.2737	0.2812	0.2078	0.066	0.0491	0.0345	0.0172	0.0287	0.0027	0.0208	0.0048	0.0199	0.0126	0.0022	0.0037	0.0034	0.0114	0.0077	Rock
0.2299	0.2789	0.3833	0.2933	0.1155	0.1705	0.1294	0.0909	0.08	0.0567	0.0198	0.0114	0.0151	0.0085	0.0178	0.0073	0.0079	0.0038	0.0116	0.0033	0.0039	0.0081	0.0053	Rock
0.2023	0.1794	0.0227	0.1313	0.1775	0.1549	0.1626	0.0708	0.0129	0.0795	0.0762	0.0117	0.0061	0.0257	0.0089	0.0262	0.0108	0.0138	0.0187	0.023	0.0057	0.0113	0.0131	Mine
0.182	0.1815	0.1593	0.0576	0.0954	0.1086	0.0812	0.0784	0.0487	0.0439	0.0586	0.037	0.0185	0.0302	0.0244	0.0232	0.0093	0.0159	0.0193	0.0032	0.0377	0.0126	0.0156	Mine
0.2633	0.3198	0.1933	0.0934	0.0443	0.078	0.0722	0.0405	0.0553	0.1081	0.1139	0.0767	0.0265	0.0215	0.0331	0.0111	0.0088	0.0158	0.0122	0.0038	0.0101	0.0228	0.0124	Mine
0.4029	0.3676	0.151	0.0745	0.1395	0.1552	0.0377	0.0636	0.0443	0.0264	0.0223	0.0187	0.0077	0.0137	0.0071	0.0082	0.0232	0.0198	0.0074	0.0035	0.01	0.0048	0.0019	Mine
0.4374	0.182	0.3376	0.6202	0.4448	0.1863	0.142	0.0589	0.0576	0.0672	0.0269	0.0245	0.019	0.0063	0.0321	0.0189	0.0137	0.0277	0.0152	0.0052	0.0121	0.0124	0.0055	Mine

Runtime Environments

- https://notebooks.azure.com/
- Sign in
  - **Email: abc123@my.utsa.edu**
- You will be redirected to UTSA login page
  - **User: abc123**
- 

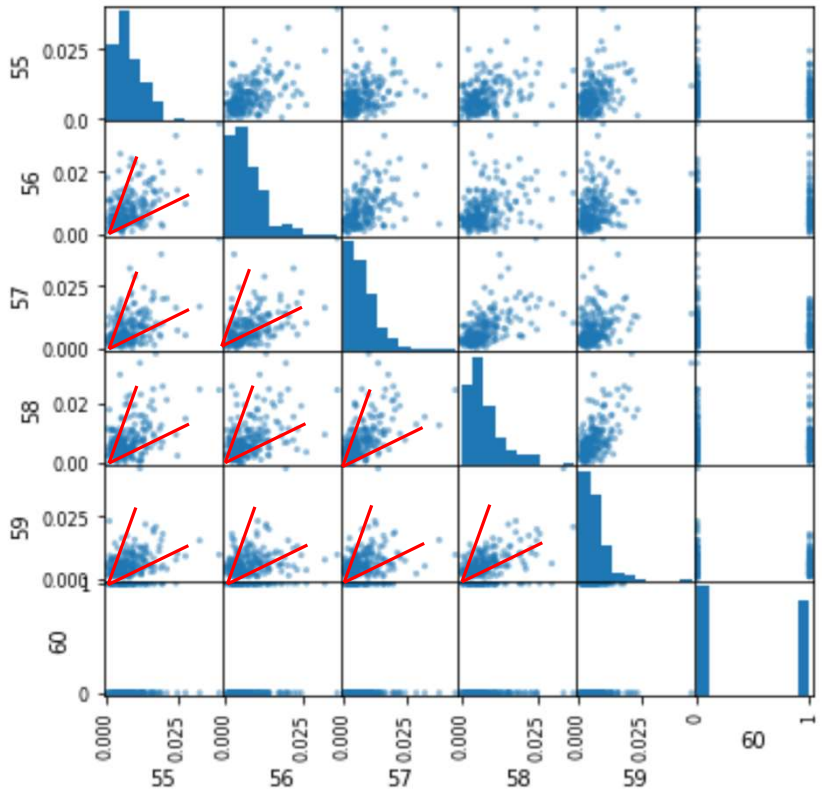
Scatter matrix of the data

- What the graph will tell us



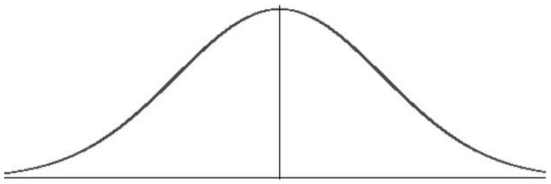
Scatter matrix of the data

- Positive correlation between signals bounced off the object from different angles

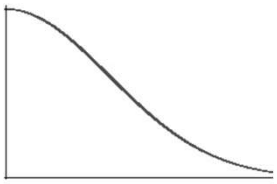


Scatter matric of the data

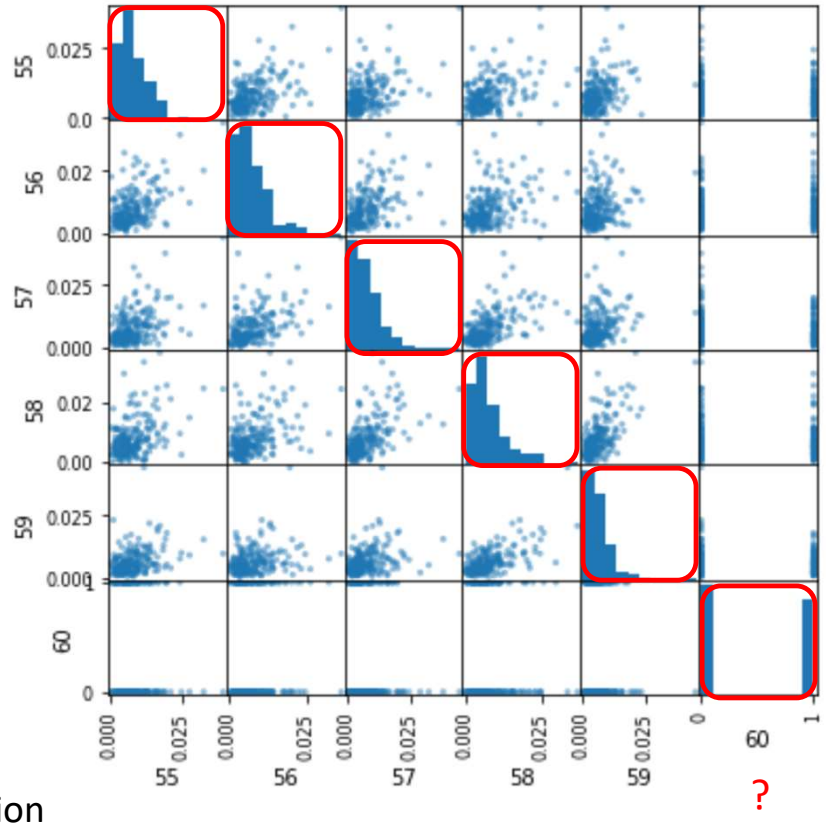
- Normal distribution:
  - The mean, mode and median are all equal.
  - The curve is symmetric at the center (i.e. around the mean, μ).
  - Exactly half of the values are to the left of center and exactly half the values are to the right.
  - The total area under the curve is 1.



Normal distribution



half-Normal distribution



Correlation of the data

- Positive correlation between signals bounced off the object from different angles

	55	56	57	58	59	60
55	1.000000	0.515154	0.463659	0.430804	0.349449	-0.129341
56	0.515154	1.000000	0.509805	0.431295	0.287219	-0.000933
57	0.463659	0.509805	1.000000	0.550235	0.329827	-0.184191
58	0.430804	0.431295	0.550235	1.000000	0.642872	-0.130826
59	0.349449	0.287219	0.329827	0.642872	1.000000	-0.090055
60	-0.129341	-0.000933	-0.184191	-0.130826	-0.090055	1.000000

Correlation of the data

- Negative correlation between signals bounced off the object and its category
  - Stronger signals bounced off are for the category with the smaller number

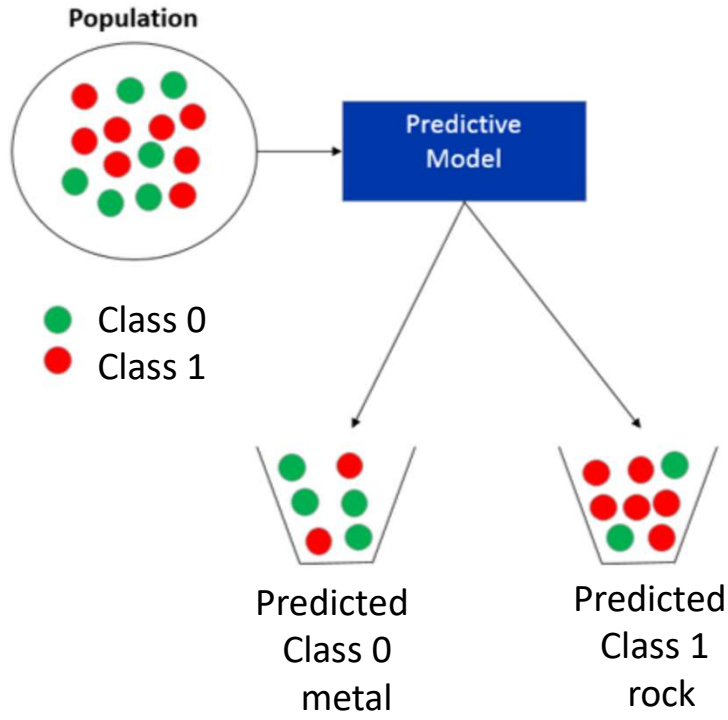
	55	56	57	58	59	60
55	1.000000	0.515154	0.463659	0.430804	0.349449	-0.129341
56	0.515154	1.000000	0.509805	0.431295	0.287219	-0.000933
57	0.463659	0.509805	1.000000	0.550235	0.329827	-0.184191
58	0.430804	0.431295	0.550235	1.000000	0.642872	-0.130826
59	0.349449	0.287219	0.329827	0.642872	1.000000	-0.090055
60	-0.129341	-0.000933	-0.184191	-0.130826	-0.090055	1.000000

Train the model

- Steps
  - Split the data into training and testing sets
  - Standardize features
  - Select model
  - Train the model

Prediction Results

- **accuracy_score**
  - fraction of samples predicted correctly
- **recall_score**
  - fraction of positives events that you predicted correctly
- **precision_score**
  - fraction of predicted positives events that are actually positive
- **f1_score**
  - harmonic mean of recall and precision, with a higher score as a better model



Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3bc336865019

In-class assignment W12 MITRE GenAI

- Download the code and upload it to Blackboard.