Application of Big Data in Social Science

week 14

Heidi Hyeseung Choi
Fall 2022
HYSIS
heidichoi@hanyang.ac.kr

Announcements

~~9.1~~	~~1~~	~~Introduction~~
~~9.8~~	~~2~~	~~Web Scraping~~
~~9.15~~	~~3~~	
~~9.22~~	~~4~~	~~Natural Language Processing~~
~~9.29~~	~~5~~	
~~10.6~~	~~6~~	~~Text Analysis~~ ~~(recorded lecture on week 7)~~
~~10.13~~	~~7~~	
~~10.20~~	~~8~~	~~Mid term exam (as school schedule)~~
~~10.27~~	~~9~~	~~Midterm review & word cloud~~
~~11.3~~	~~10~~	~~Social Network Analysis~~
~~11.10~~	~~11~~	~~COVID-19 T-T~~
~~11.17~~	~~12~~	~~Machine Learning: Supervised Learning~~
~~11.24~~	~~13~~	~~Supervised Learning~~
12.1	14	Machine Learning: Unsupervised Learning
12.8	15	Data Visualization
12.15	16	Final Exam

Homework assignment 2

- Due 22nd December.

You have two options to choose from:

1. Devising a marketing strategy.
2. Visualization.

Homework assignment 2: working with raw data

Option 1: Devising a marketing strategy

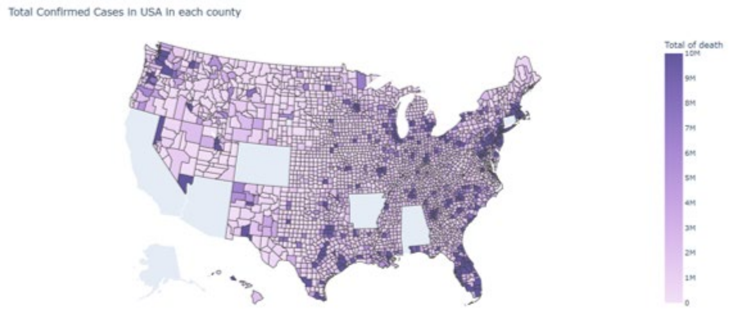
- You are at a marketing team and you want to do a target marketing
- Using raw data given, you are to process data and using analysis that we have done during the class, try to come up with target marketing.
- For each target group, you can write a short paragraph describing their consuming behavior, and how you would conduct target marketing for each group.
- Within data, there are a lot of info available so whichever data you choose, it is up to you.
- Hand in your 1)jupyter notebook file and 2)a word file with figures and written descriptions. (up to 4 pages including everything would be fine)

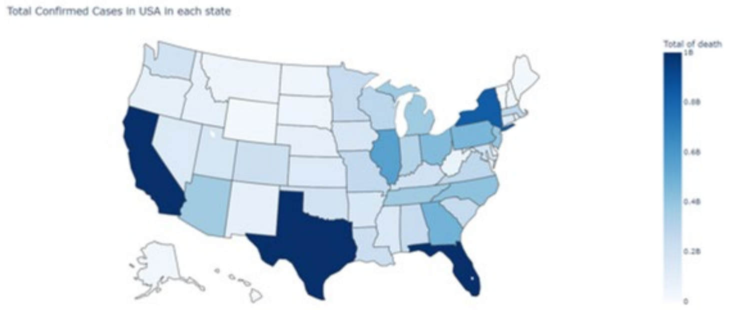
Homework assignment 2: working with raw data (choose one!)

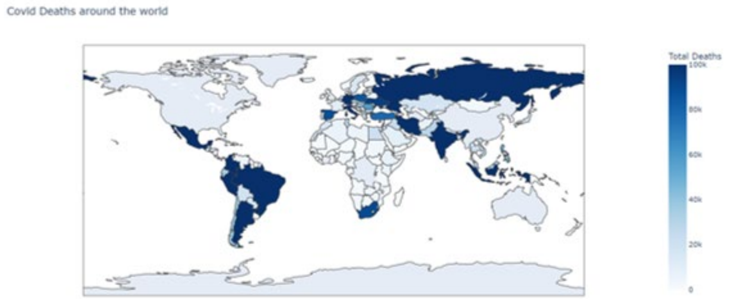
Option 2. Visualization: COVID-19 data released by JHU

- Process data using raw data.
- You would need to work on data manipulation.
- Work on both US and World data.
- For US data, do both county level and state level.
- For world data, just do the country level, not province nor state.
- Hand in your 1) jupyter notebook file and 2) word file with figures in them.
- # of figures, etc does not matter as long as you are able to get the figures out.

Example output for each option

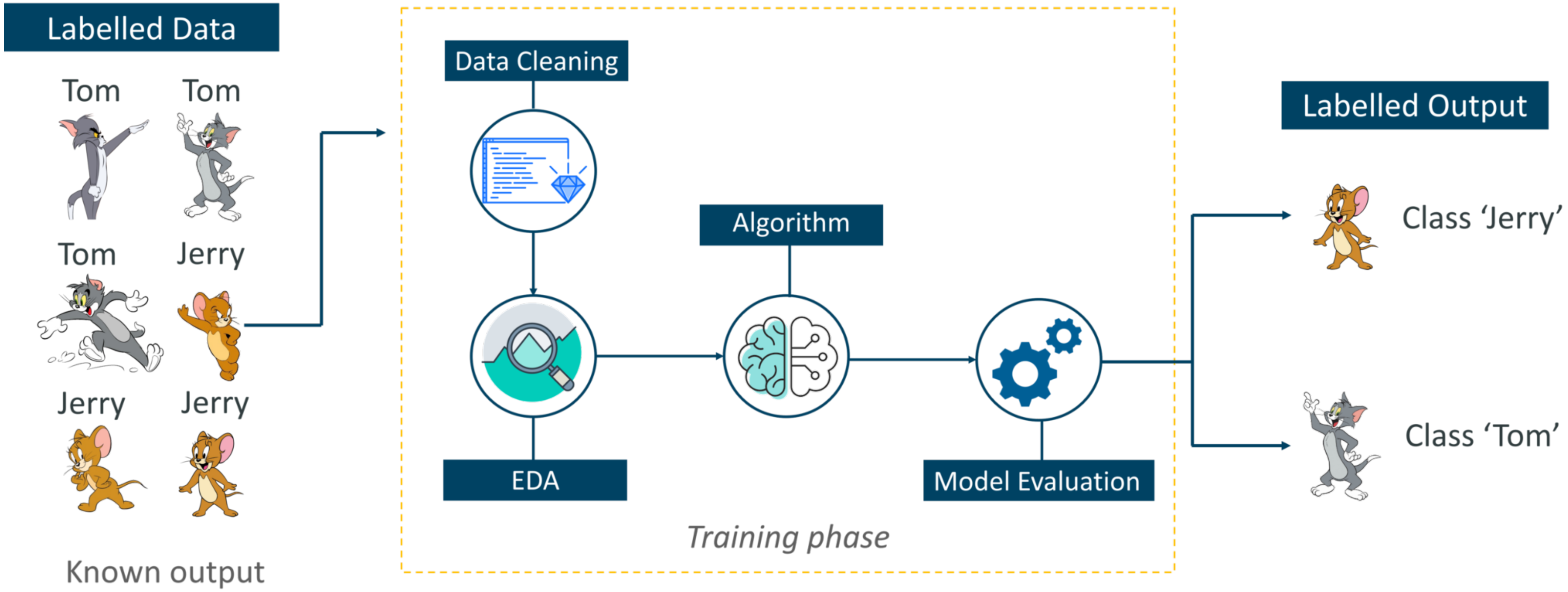

Total Confirmed Cases in USA in each county


Total Confirmed Cases in USA in each state


Covid Deaths around the world

Machine Learning (ML)

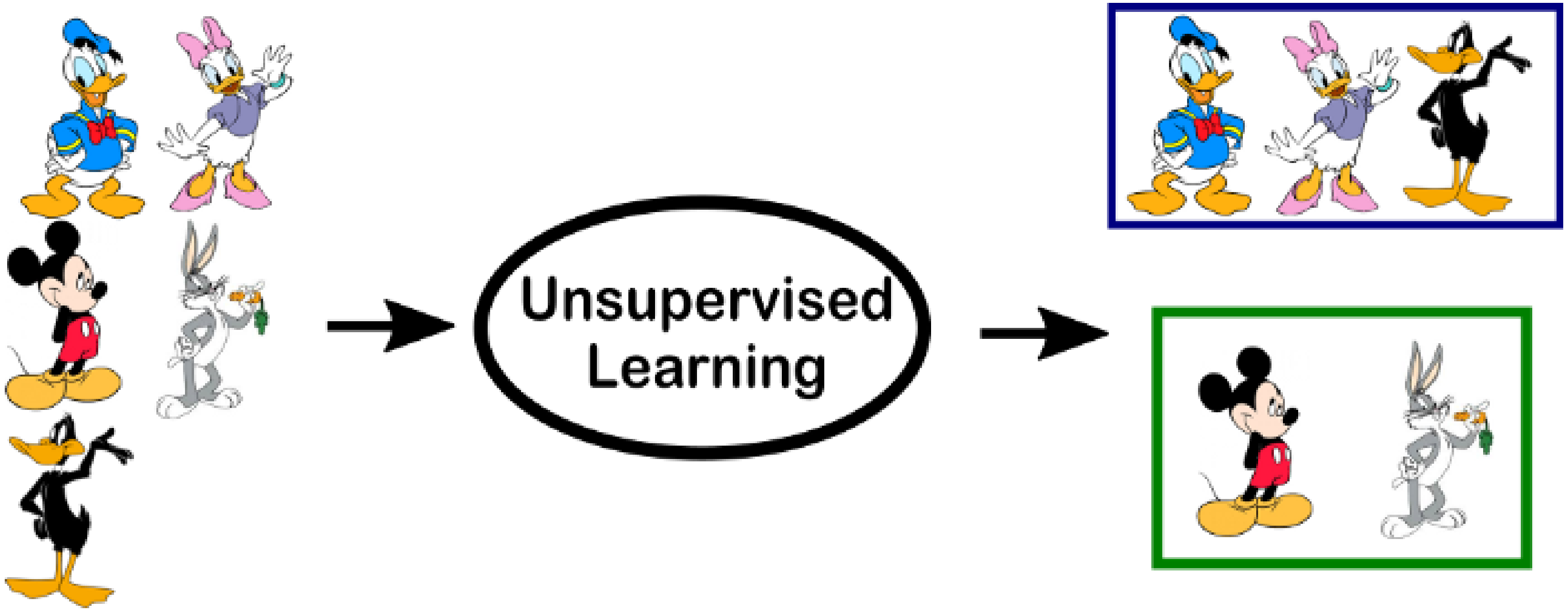
- supervised machine learning:
  - algorithm is provided with set of data where it has the answer for each of the input values.
  - apply what has been learned in the past to new data to predict future events.
  - Types of supervised learning: Regression, classification



source: https://becominghuman.ai/an-introduction-to-machine-learning-33a1b5d3a560

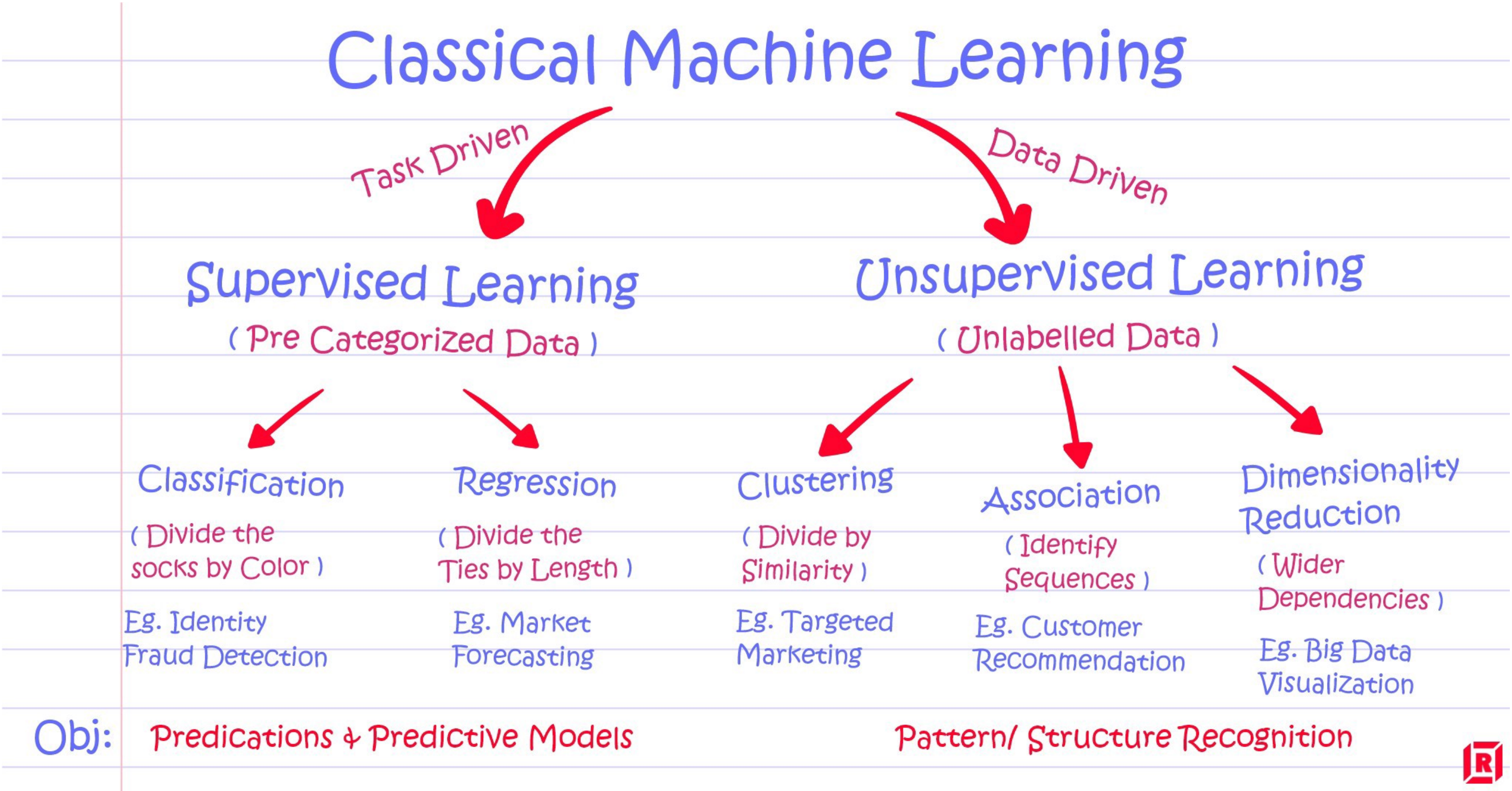
Machine Learning (ML)

- unsupervised machine learning:
  - when data provided is neither classified nor labeled.
  - used to find hidden structures from unlabeled data.
  - types of unsupervised machine learning: clustering, association



source: https://becominghuman.ai/an-introduction-to-machine-learning-33a1b5d3a560

Supervised vs unsupervised learning



Classical Machine Learning

Task Driven

Data Driven

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

Unsupervised Learning

(Unlabelled Data)

Clustering

(Divide by Similarity)

Eg. Targeted Marketing

Association

(Identify Sequences)

Eg. Customer Recommendation

Dimensionality Reduction

(Wider Dependencies)

Eg. Big Data Visualization

Obj: Predications & Predictive Models

Pattern/ Structure Recognition

Unsupervised Learning

- Unsupervised Learning: **No ‘target’**
  No labels are given to the learning algorithm, leaving it on its own to find structure in its input.
- Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- The goal is…
  - to discover groups of similar patterns embedded in the data (clustering)
  - to determine how the data is distributed (density estimation)

On Unsupervised Learning..

Issues:

- Unsupervised Learning is difficult as compared to Supervised Learning.
  - Accuracy rate?
  - Meaningful result?
- External evaluation of the field expert to derive meaning is required.
- Need to define an objective function on clustering.

Why do we do this?:

- There are cases where we do not know how many classes data is divided into.
- We may use clustering to gain insights of the patterns and structure of the data.

Parametric vs non-parametric Unsupervised Learning

Parametric model assumes that sample data comes from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters.*

*(any measured quantity of a poplation that describes an aspect of the population. (eg. Mean, sd)

Non-parametric model does not assume an explicit (finite-parametric) mathematical form for the distribution when modeling the data.

Parametric Unsupervised Learning

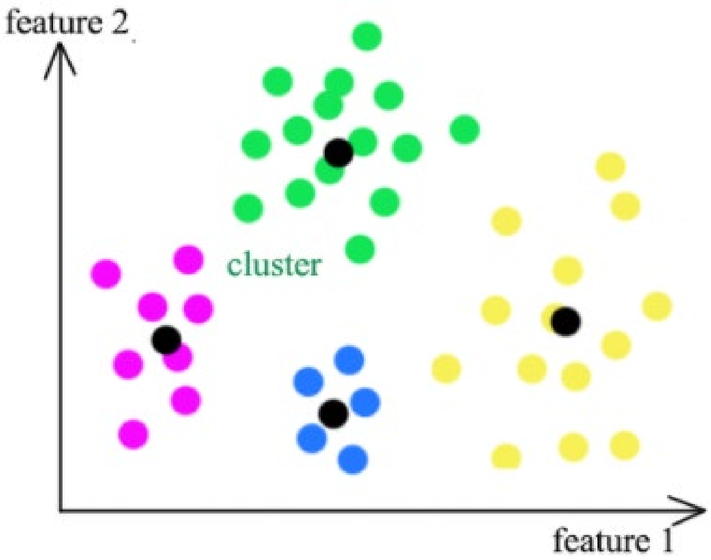
- We assume a parametric distribution of data.
- The sample data comes from a population that follows a probability distribution based on a fixed set of parameters.
- Parametric Unsupervised Learning involves building Gaussian Mixture Model and using Expectation-Maximization algorithm to predict the class of the sample.
- This case is much harder than the standard supervised learning because we cannot check for accuracy of our data.

Non-parametric Unsupervised Learning

- Data is grouped into clusters, where each cluster shows something about categories and classes present in the data.
- This method is commonly used to model and analyze data with small sample sizes.
- Nonparametric models do not require the modeler to make any assumptions about the distribution of the population.

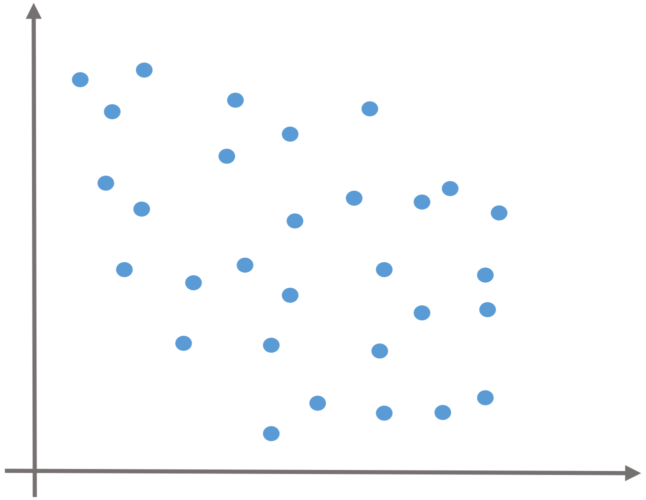
Clustering

- Clustering is the "process of organizing objects into groups that are similar in some way."
- It is considered as one of the most important unsupervised learning problem.
- Finding a collection of objects which are similar between them and dissimilar to the other objects in the other clusters.
  - Internal distance within the cluster should be small (close/similar to one another)
  - External distance should be large (far/dissimilar to those in other groups)
- There is no 'best' criterion for a good clustering. It should be tailored so that it would derive a meaning that is hidden behind data.
- Clustering algorithms
  - K-means
  - Hierarchical clustering
  - Mixture of Gaussians

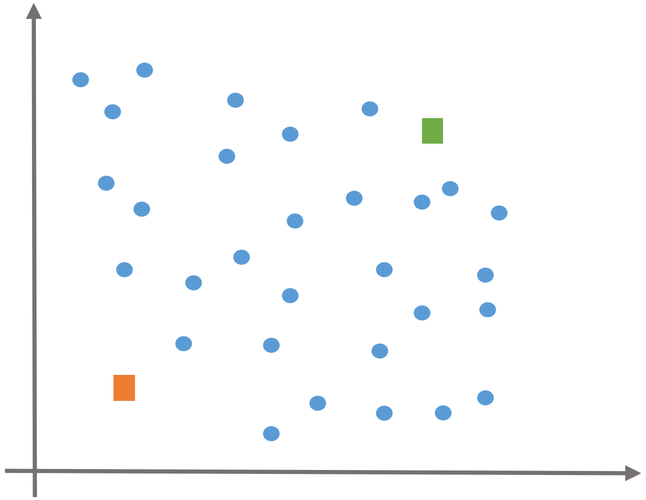


K-means clustering

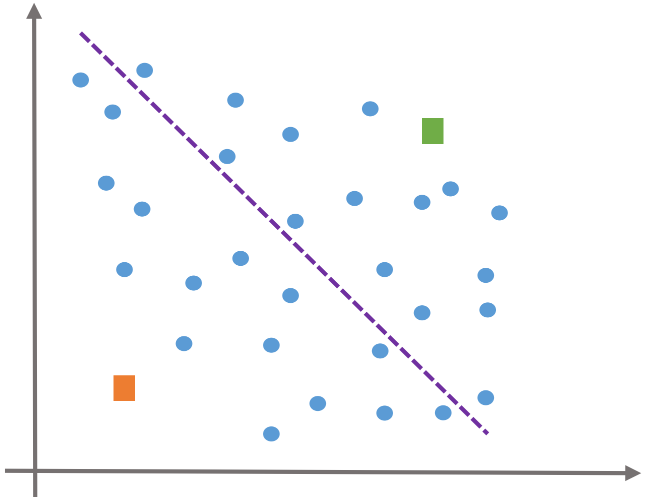
1. original data



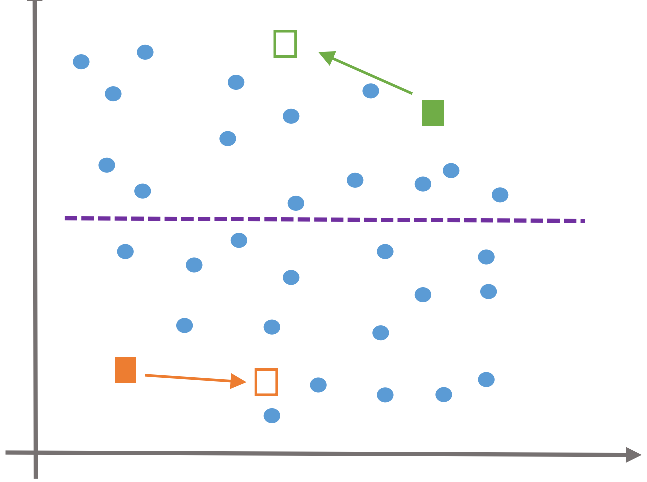
2. Select K random points as cluster centers called ‘centroids’



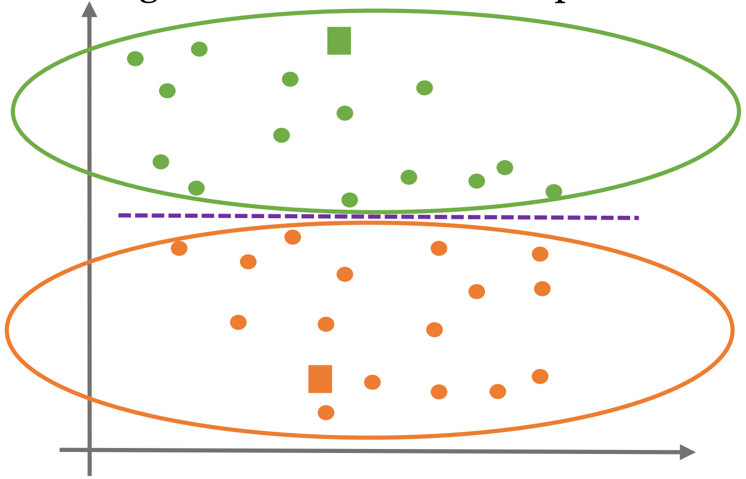
3. Assign each data point to the closest centroid



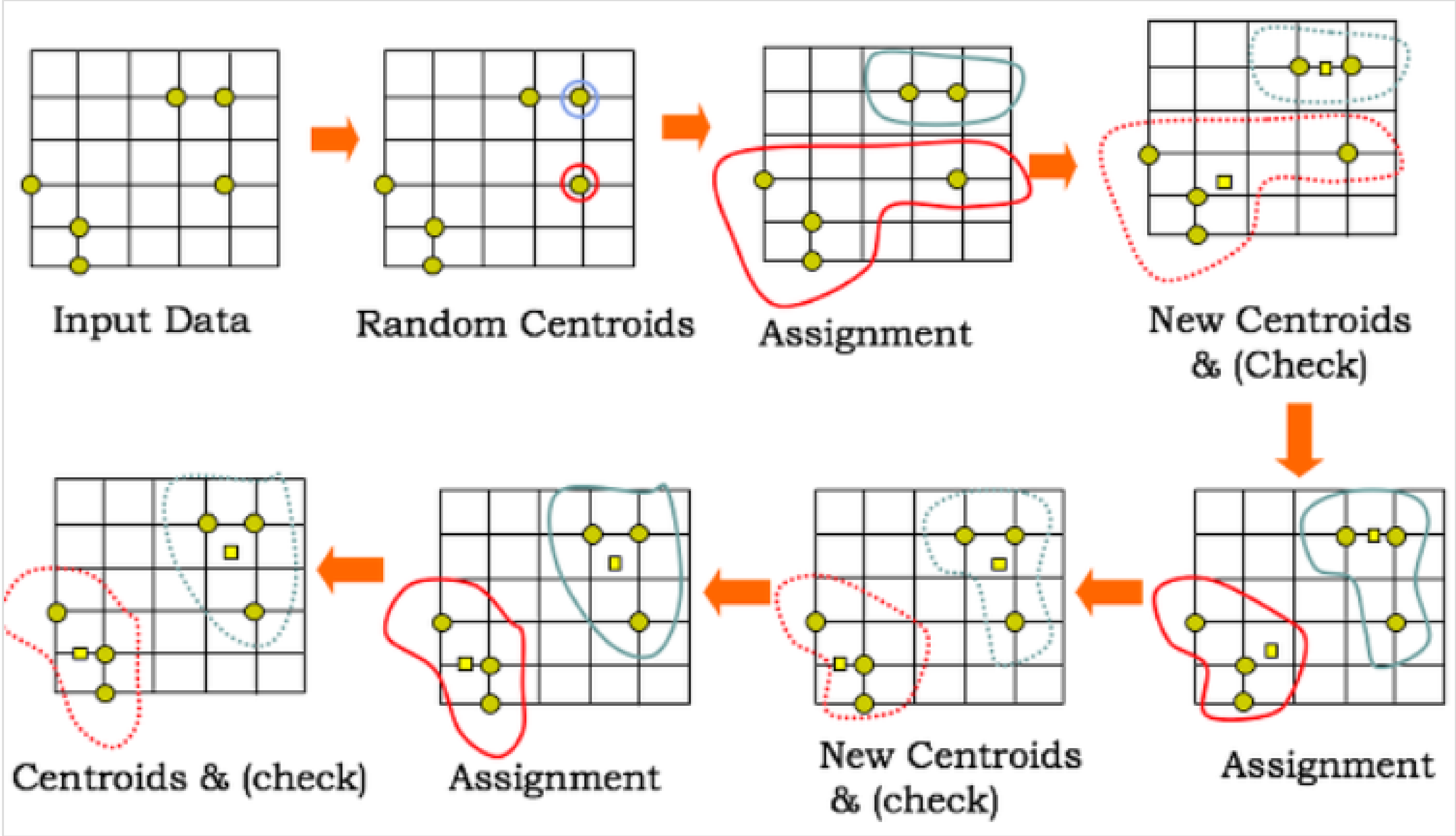
4. Determine new centroids by computing the average of the assigned points



5. Repeat until all the data points are perfectly organized within a cluster space



K-means clustering



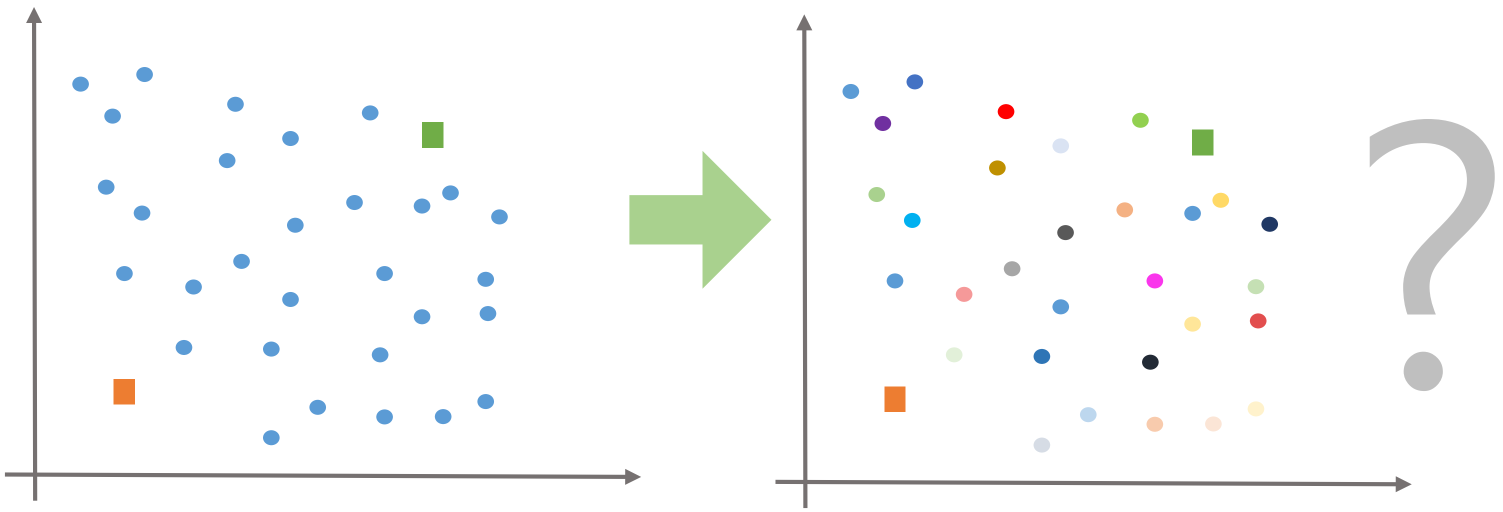
How many clusters to make?

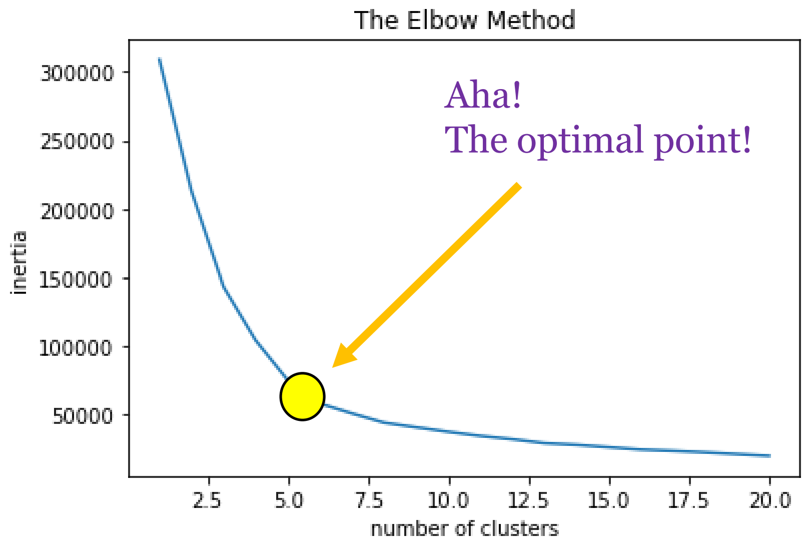
- What is a GOOD or WELL clustered group?
  - If data points are clustered well together, we will say it is well done.
  - To do so, we can refer to “inertia”.
- Inertia:
  - Measures how well a dataset was clustered by K-means.
  - calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.
- Within Cluster Sum of Squares (WCSS)
  - WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids.

How many clusters to make?

- A good model is one with low inertia AND a low number of clusters (k).
- There is a tradeoff because as K increases, inertia decreases.
- Smallest inertia would be 0, where each dataset belong to each cluster…?

- To find the optimal K for a dataset, we can use the Elbow method !
- It helps us to find the point where the decrease in inertia begins to slow.



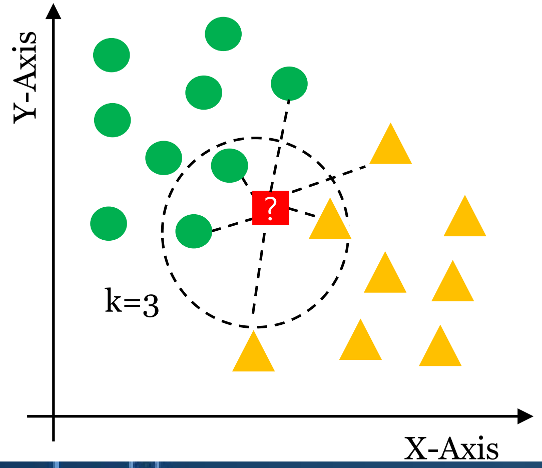


Aha!
The optimal point!

K-Nearest Neighbor vs K-means algorithm?!!?

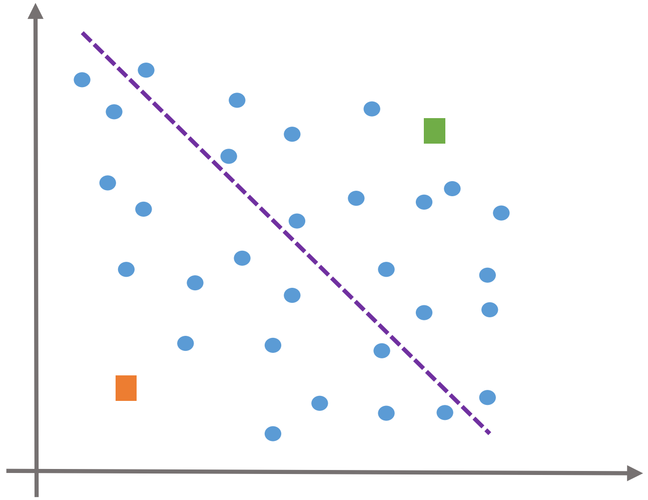
KNN algorithm

- Supervised Learning algorithm
- Have a specific target to predict
- “K” is the number of neighbors to consider.
- Non-parametric
- Classification: majority vote.

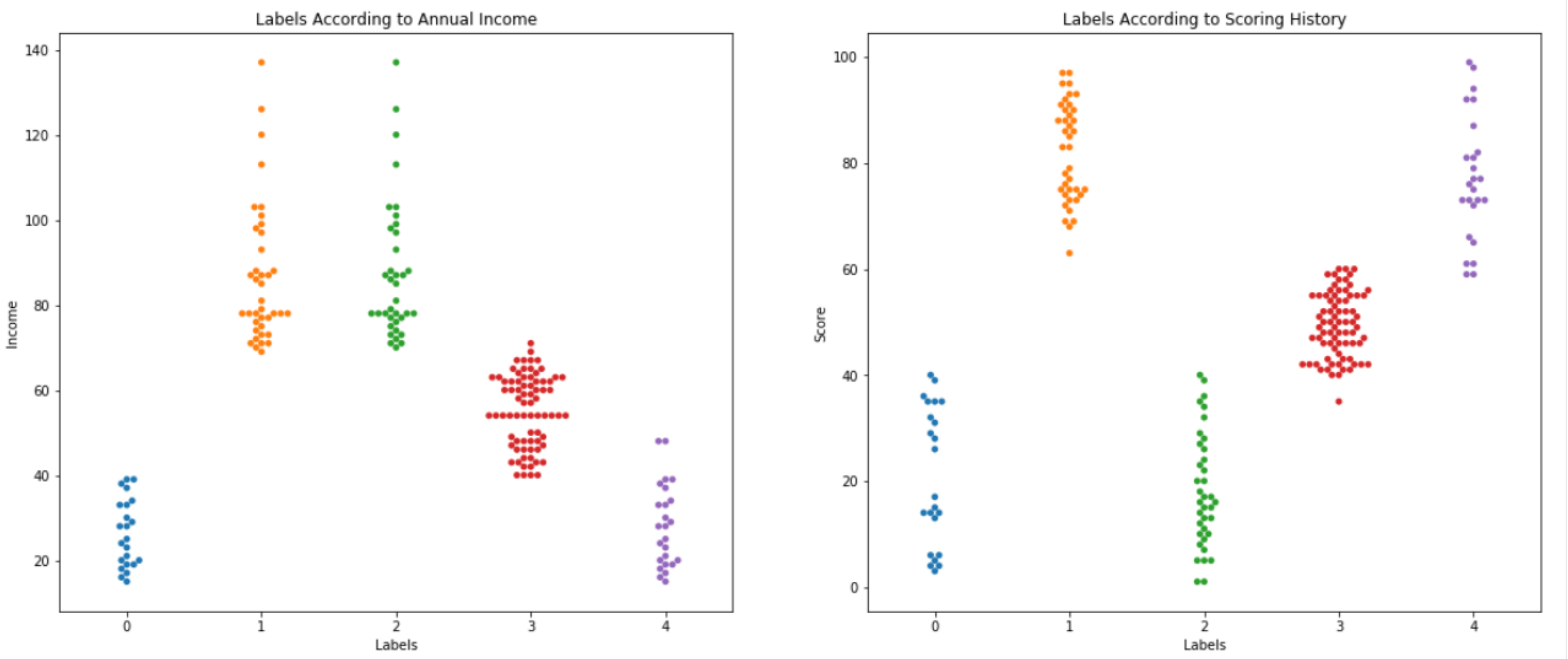


K-means algorithm

- Unsupervised Learning algorithm
- Use unlabeled points and tries to group them into “k” number of cluster.
- “K” refers to the number of clusters you want to make.
- Non-parametric
- Cluster analysis



a little tip of matplotlib: subplots



when we draw a graph using plt:

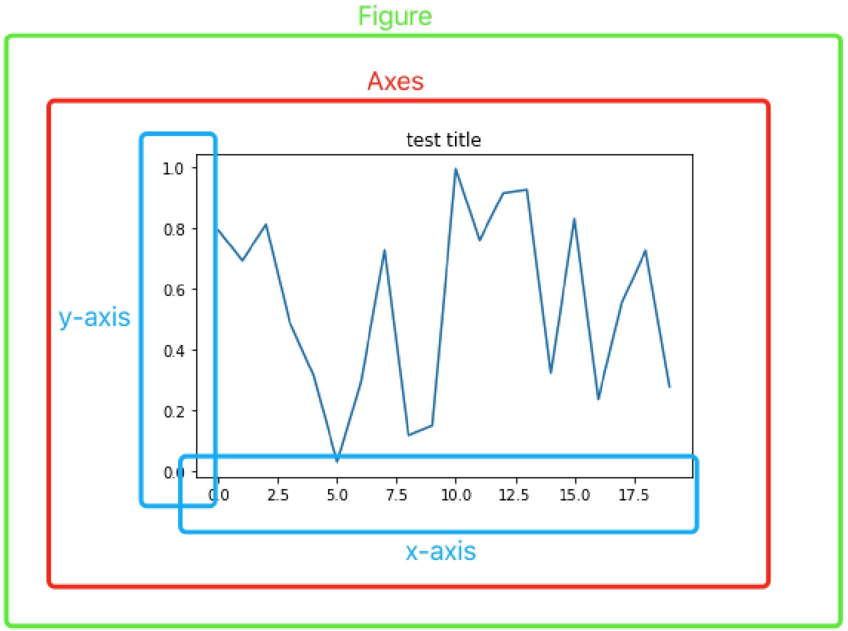
1. A **Figure** object is generated (green) :
   - paper that you can draw anything you want.

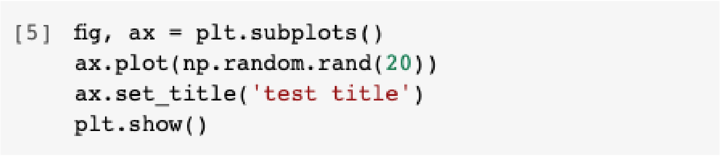
2. An **Axes** object is generated implicitly with the plotted line chart (red):
   - drawing a chart in a ‘cell’.

3. All the elements of the plot (eg.x , y-axis) are rendered inside the Axes object (blue)

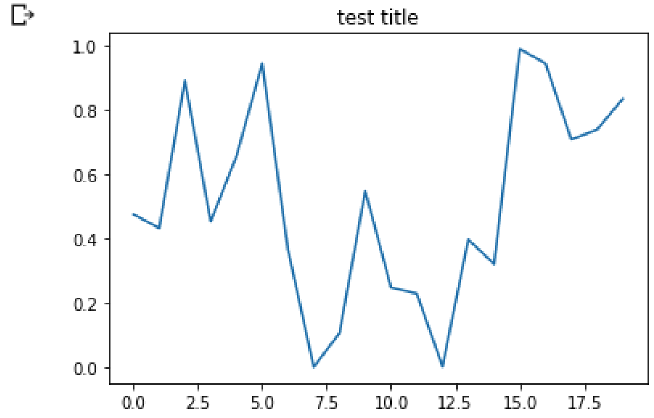
If we’re drawing only one graph, we don’t have to draw a “cell” first, just simply draw on the paper. So, we can use plt.plot(...).

We can explicitly draw a “cell” on the “paper”, to tell Matplotlib that we will be drawing a chart inside this cell.





```
[5] fig, ax = plt.subplots()
    ax.plot(np.random.rand(20))
    ax.set_title('test title')
    plt.show()
```



matplotlib: subplots

- Subplots are used when we want to display two graphs side-by-side.
- This is different from having two lines in the same graphs.
- We call each set of axes a subplot.
- The picture or object that contains all of the subplots is called a *figure*.
- We can have many different *subplots* in the same figure, and we can lay them out in many different ways.
- We can think of our layouts as having rows and columns of subplots.

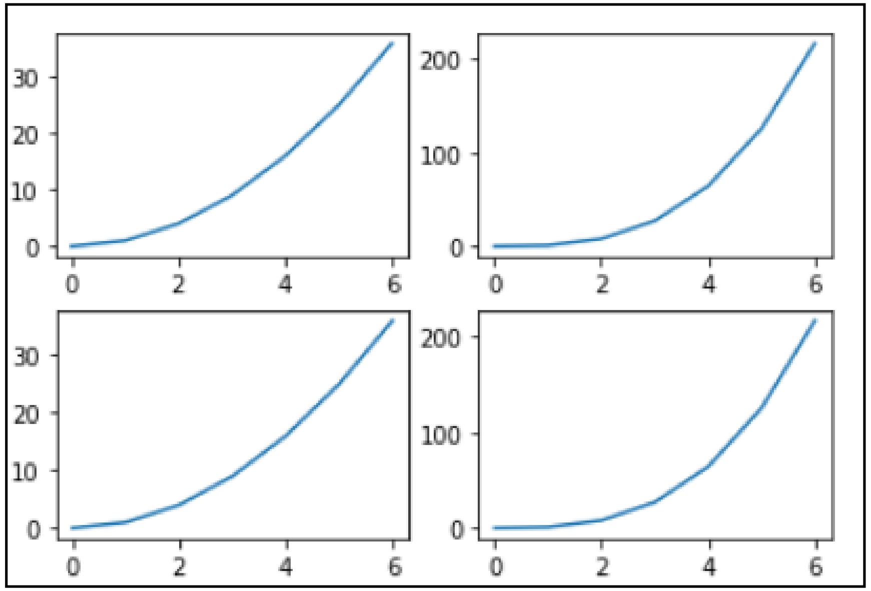
matplotlib: subplots

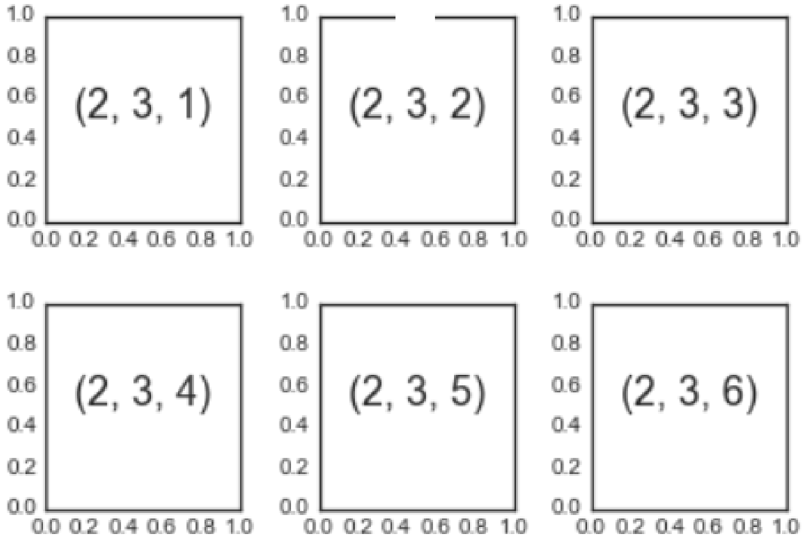
The command plt.subplot() needs three arguments to be passed into it:

- The number of rows of subplots
- The number of columns of subplots
- The index of the subplot we want to create

plt.subplot(2,2,1) # or just (221)

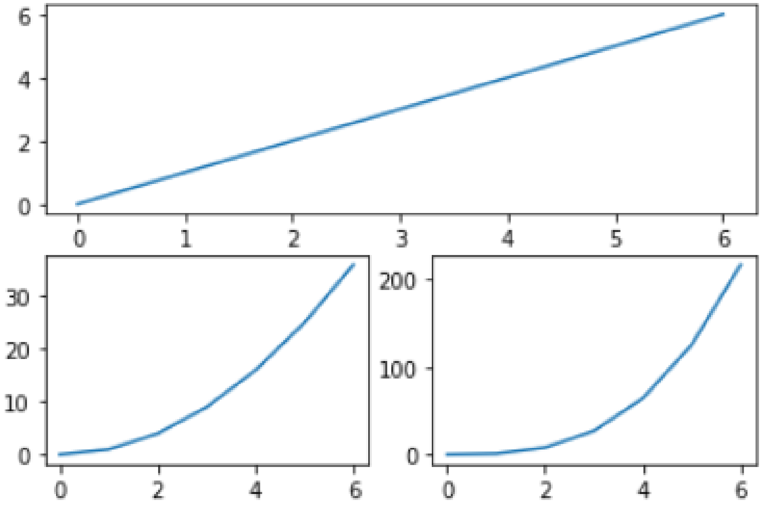
plt.subplot(2,2,3)





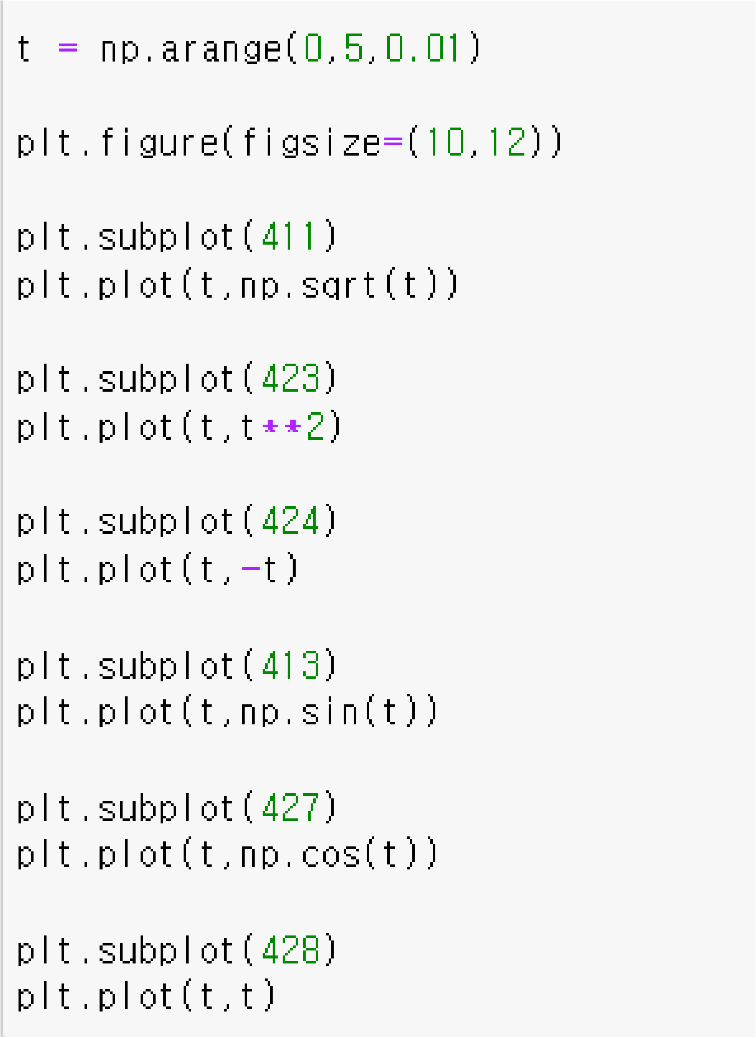
plt.subplot(2,1,1)

plt.subplot(2,2,3)

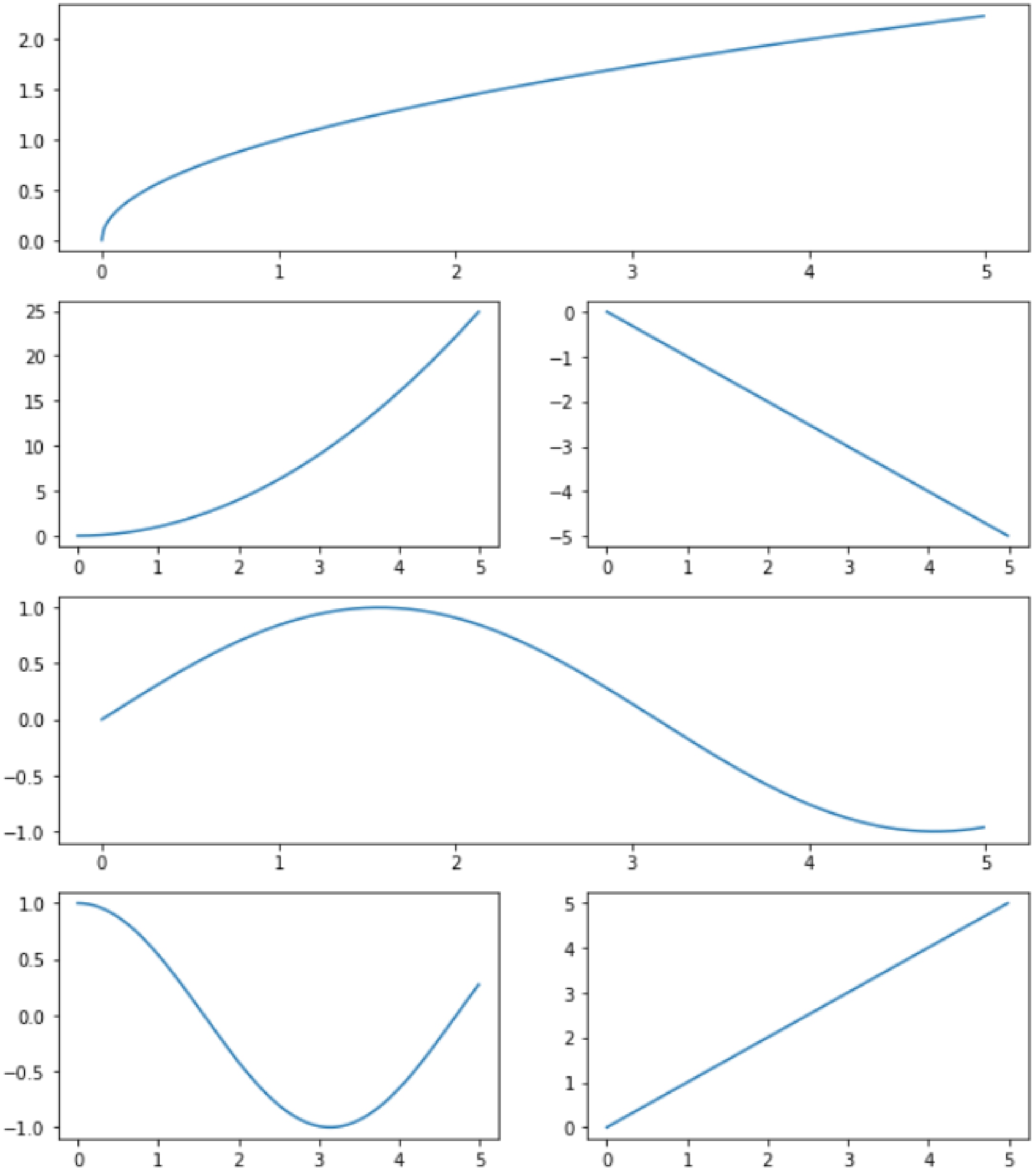


matplotlib: subplots

>>> plt.subplot(row, column, index)



```
t = np.arange(0,5,0.01)

plt.figure(figsize=(10,12))

plt.subplot(411)
plt.plot(t,np.sqrt(t))

plt.subplot(423)
plt.plot(t,t**2)

plt.subplot(424)
plt.plot(t,-t)

plt.subplot(413)
plt.plot(t,np.sin(t))

plt.subplot(427)
plt.plot(t,np.cos(t))

plt.subplot(428)
plt.plot(t,t)
```



Let’s launch our Jupyter notebook