Application of Big Data in Social Science

week 12

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	Machine Learning: Supervised & Unsupervised Learning
12.1	14	
12.8	15	Data Visualization
12.15	16	Final Exam

Week 12 (Today):

- ML: Supervised

Week 13:

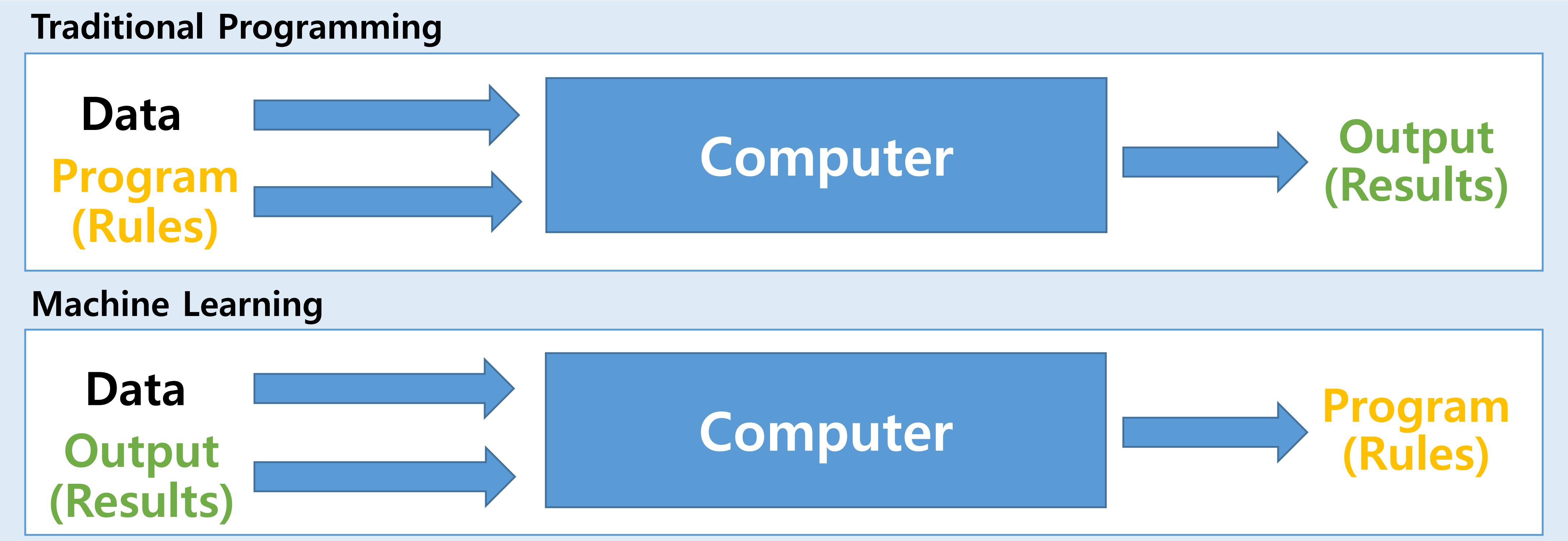
- ML: Supervised

Week 14:

- ML: Unsupervised

Traditional Programming vs Machine Learning

- Traditional programming paradigm:
  - the programmer(user) write a set of instructions using code that makes the computers to perform specific computations on data.
- Machine learning paradigm:
  - computers use input data and expected outputs to try to learn the inherent patterns in the data, which would help in making data driven decisions in the future.



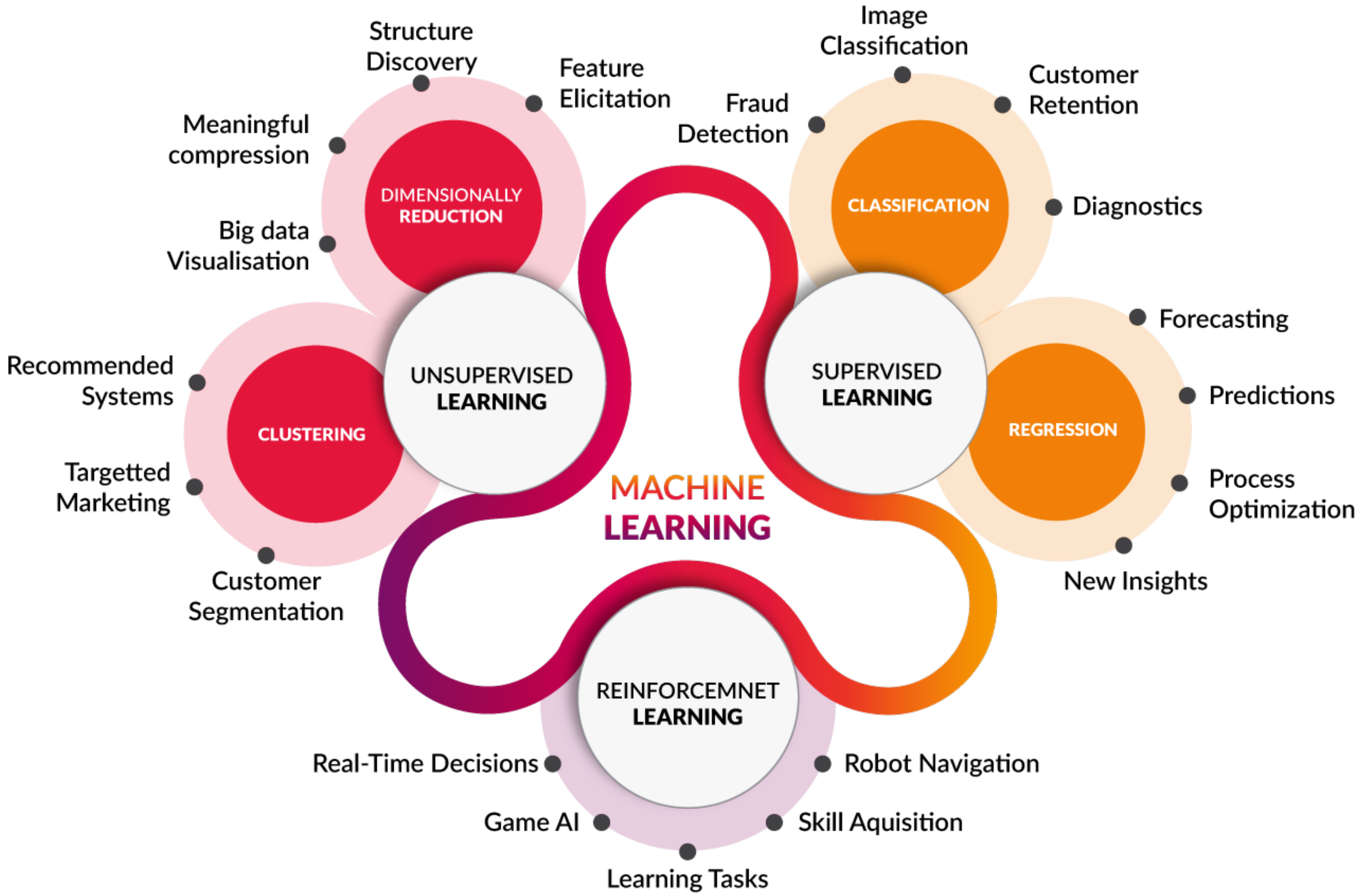
Machine Learning

“Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed.” - Arthur Samuel (1901-1990)

“ To make data-driven decisions at scale”

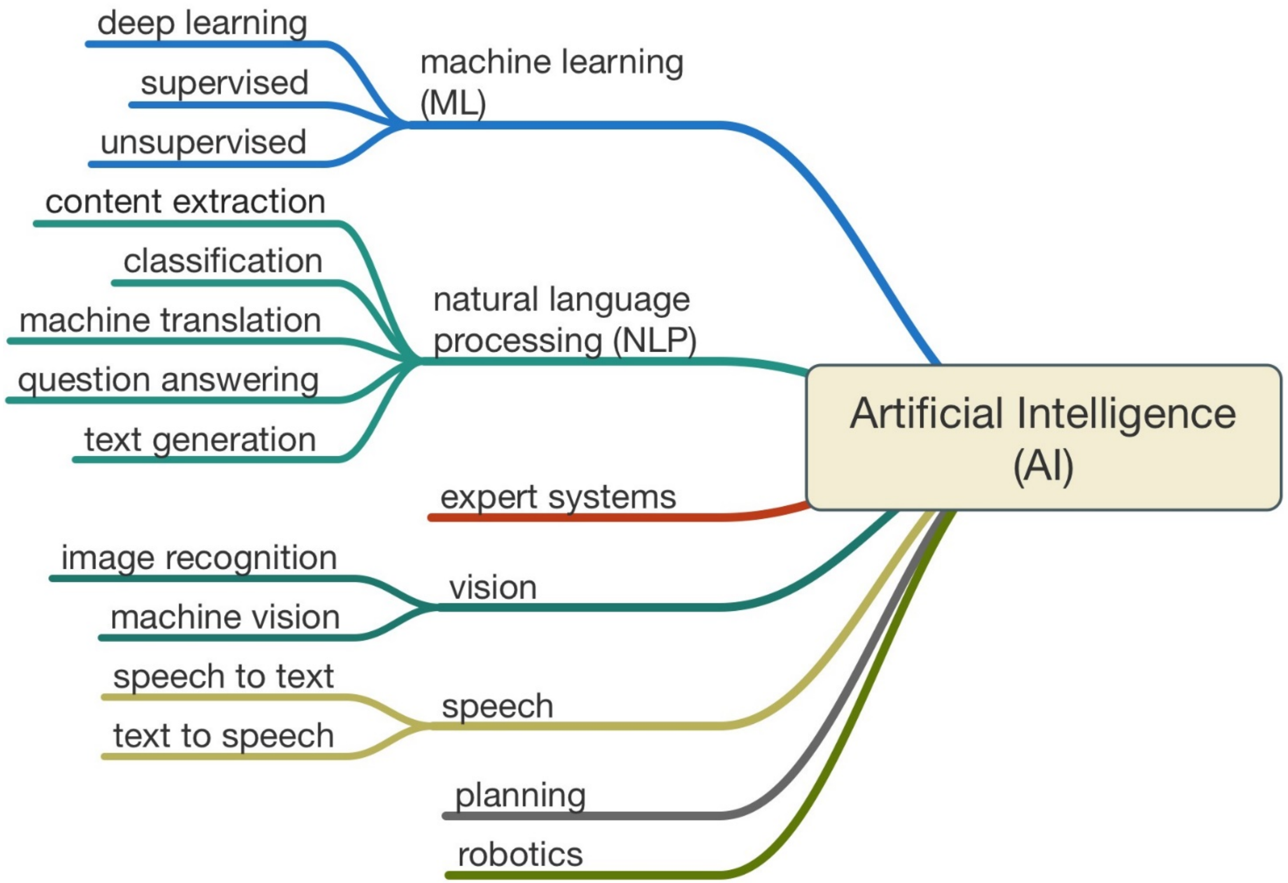
- Making data-driven decisions are not a new concept.
- Machine learning is where a computer system is fed large amounts of data, which it then uses to learn how to carry out a specific task, such as understanding speech or captioning a photograph.
- Machine Learning is not constrained by domains, and we can apply the techniques to solve problems from diverse domains, businesses and industries.
- Machine Learning is largely divided into:
  - supervised learning
  - unsupervised learning
  - reinforcement learning

Machine Learning



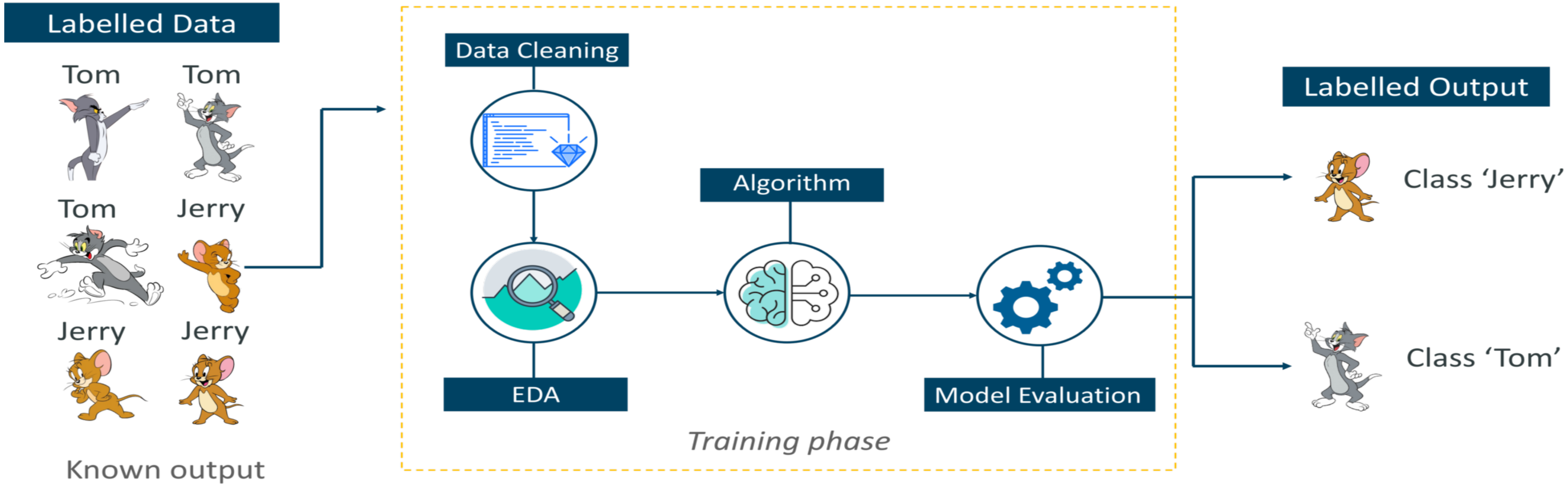
Artificial Intelligence

- Machine Learning came into prominence in the 1990s as a sub-field of AI where techniques were borrowed from AI, probability, statistics etc.
- AI is the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions.
- The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.



Machine Learning (ML)

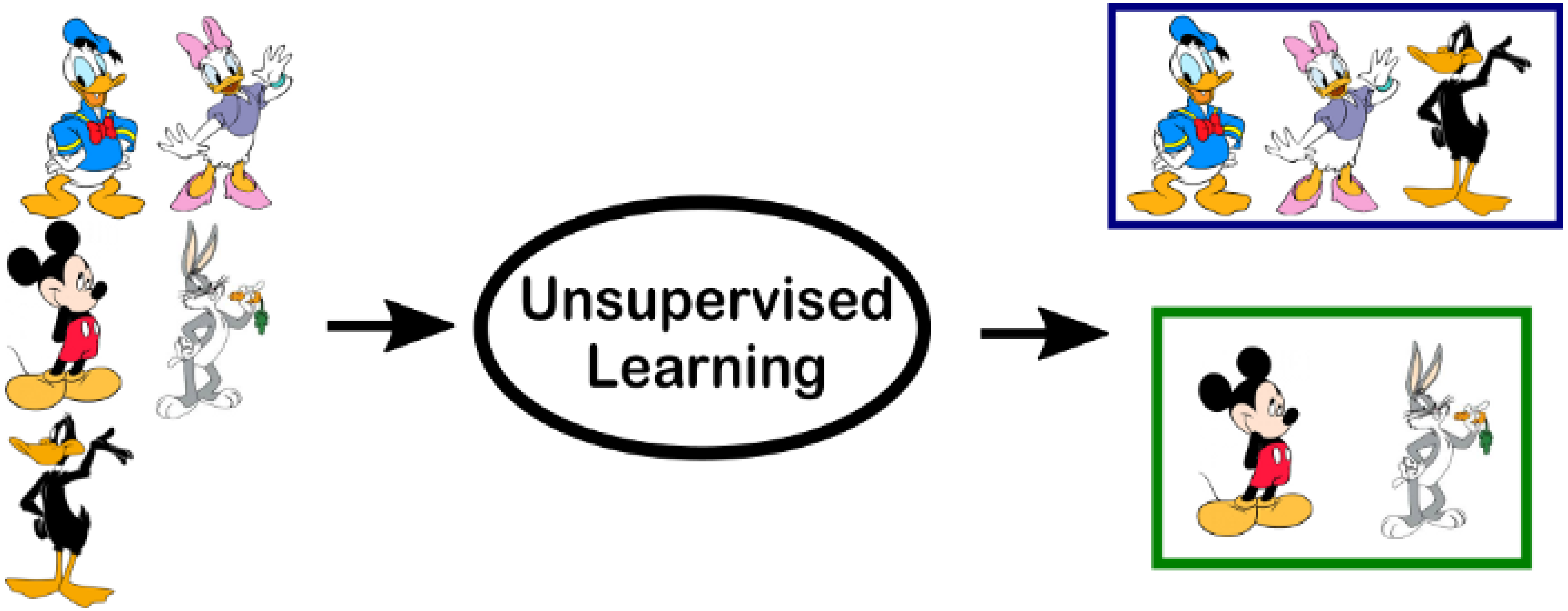
- Supervised learning:
  - uses labeled datasets to train algorithms to classify data or predict outcomes accurately.
  - As input data is fed into the model, it adjusts its weights until the model has been fitted well.
  - Applies what has been learned in the past to new data to predict future events.
  - Types of supervised learning: Regression, classification



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

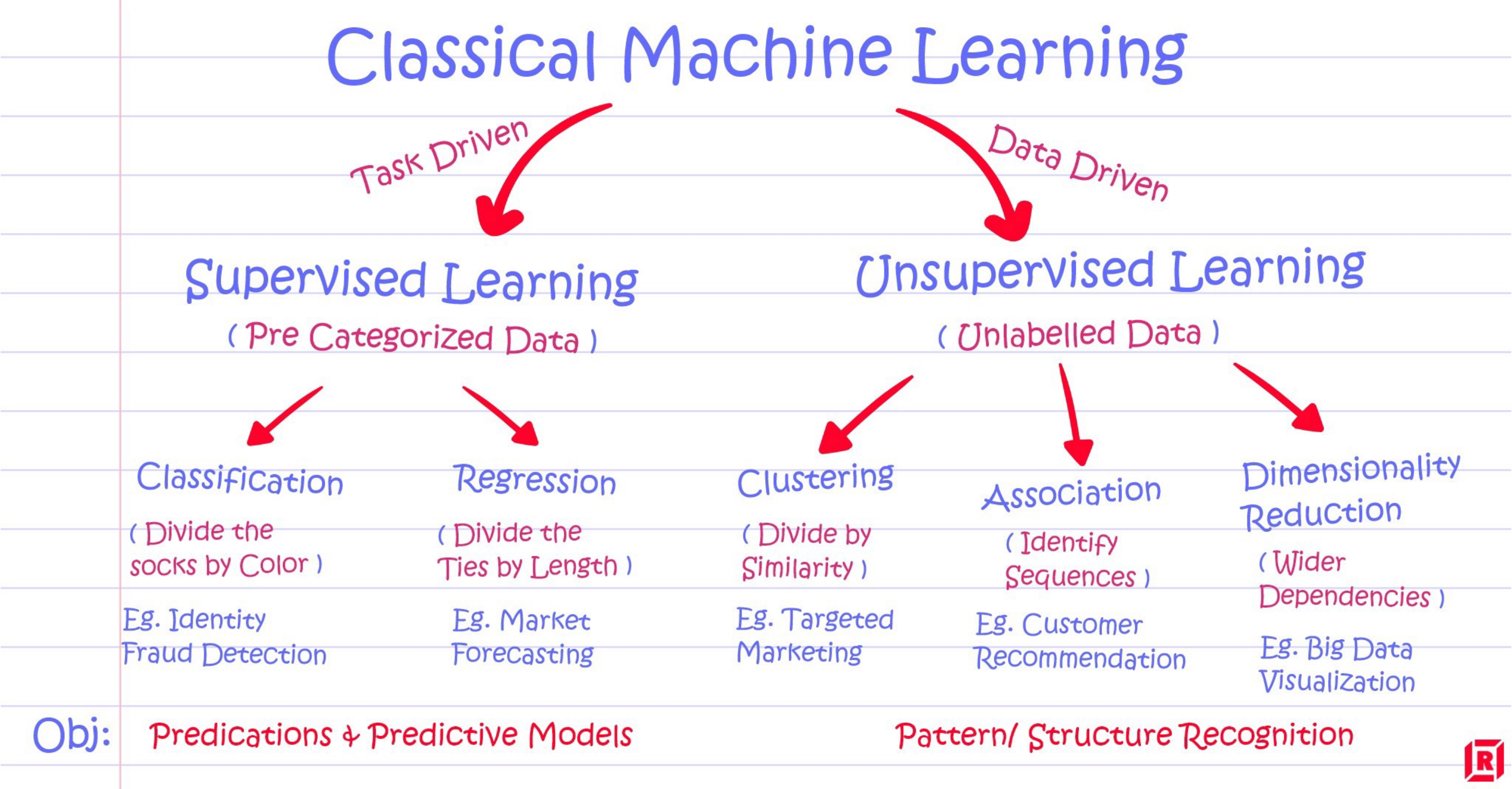
Machine Learning (ML)

- Unsupervised 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)

Unsupervised Learning

(Unlabelled Data)

Classification

(Divide the socks by Color)

Eg. Identity Fraud Detection

Regression

(Divide the Ties by Length)

Eg. Market Forecasting

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

Supervised Learning

- Learning algorithms that take in data samples (training data) and associated outputs(labels or targets) with each data sample during the model training process.
- The main objective is to learn association between input data samples x and their corresponding outputs y based on training data instances.
- The learned knowledge is used in the future to predict an output $\hat{y}$ for any new input data sample x.
- The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.
- It’s called ‘supervised’ because the model learns on data samples where the desired output labels are already known beforehand in the training phase.

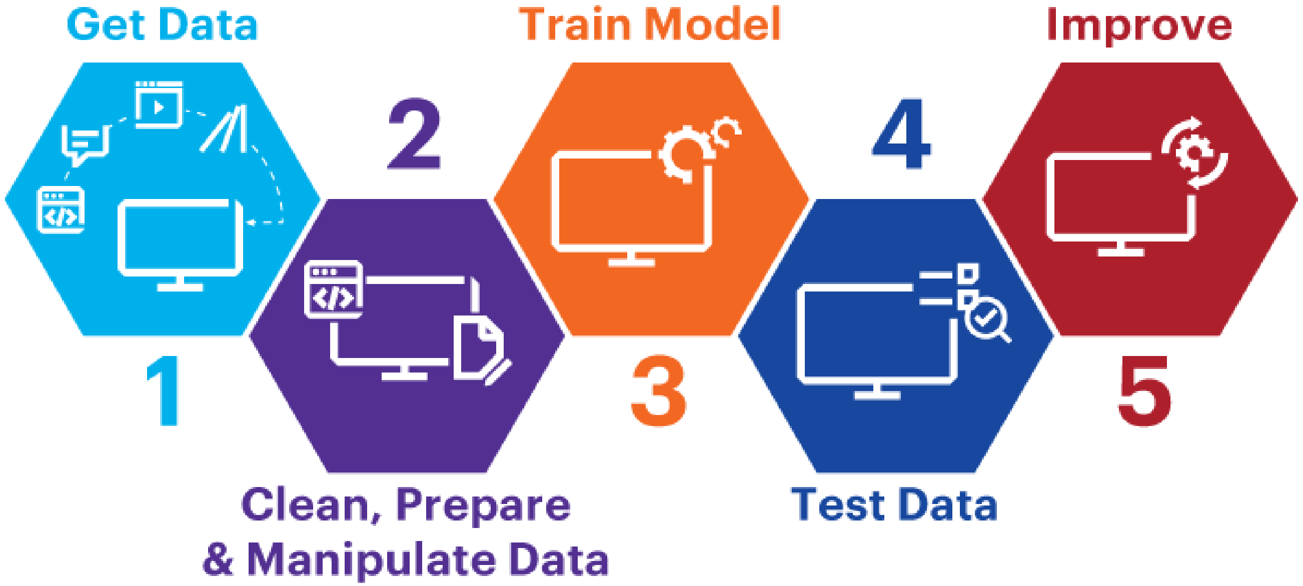
Supervised Learning

- Supervised learning is extensively used in predictive analytics.
    - Classification: uses an algorithm to accurately assign test data into specific categories.
        - ✓ algorithms: linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest etc.
    - Regression: is used to understand the relationship between dependent and independent variables.
        - ✓ algorithms: linear regression, logistical regression, and polynomial regression

- Usages:
    - predictive analysis: what would be the market price for next month? year?
    - customer sentiment analysis: how do customers feel about out product?
    - spam detection: train database to recognize patterns to classify as spam or non-spam.

So how does this work?

- Scikit-learn: machine learning library.
  - classification, regression, clustering etc
  - easily used with pandas and numpy



2. Clean, Prepare & manipulate data

- encode categorical data into numerical data
- LabelEncoder of sklearn.preprocessing

3. Train Model:

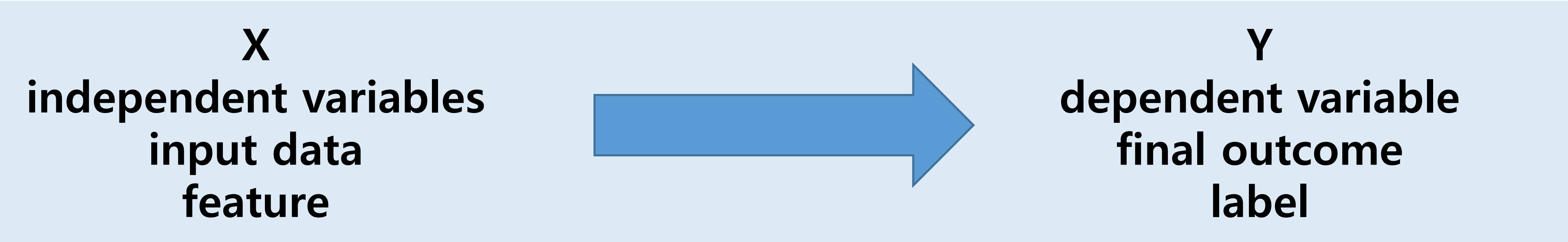
- Divide the dataset into a training set and a test set.
- we split dataset by using scikit-learn function
  train_test_split()

4. Test Data:

– Using the test set, we test our data and see the accuracy between actual and predicted category

Key machine learning definitions

- Training data: data used for training for machine learning.
- Feature:
  - an individual measurable property or characteristic of a phenomenon
  - independent variables that act like an input
  - in prediction models, we use features to make predictions
  - the number of features are called dimensions.
- Label (Target): the final output



X
independent variables
input data
feature

Y
dependent variable
final outcome
label

X_train - independent variables and they are used to train the model.

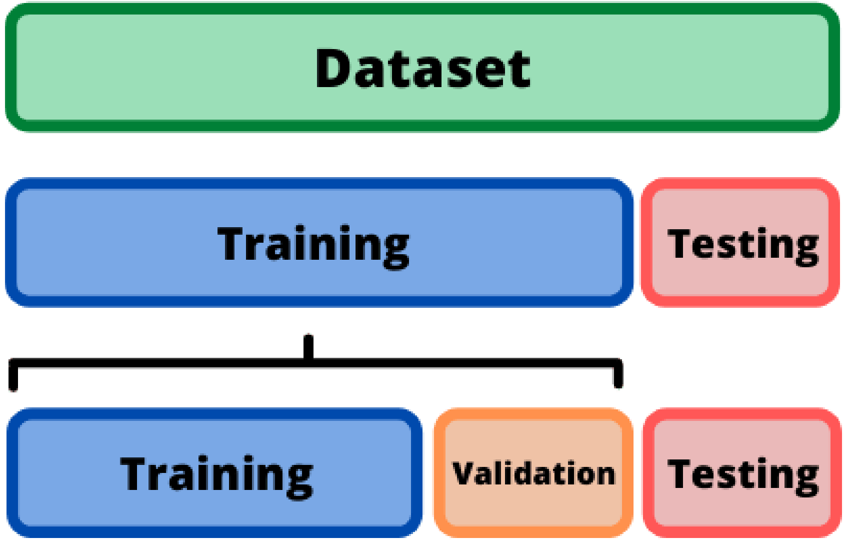
X_test - This is remaining portion of the independent variables which would not be used in the training, but is used to make predictions on the accuracy of the model.

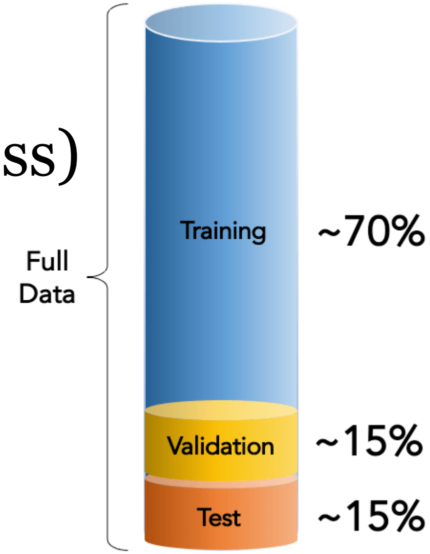
y_train - dependent variable that include category labels against independent variables. Dependent variables should be specified while training and fitting the model.

y_test - Dependent variable to test the accuracy of the model

Data sets

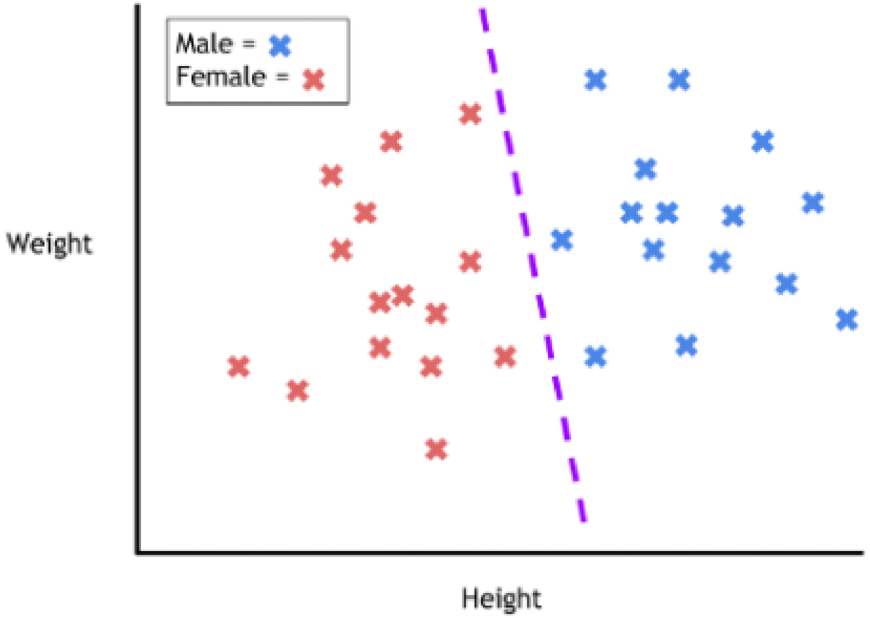
- Train set:
    - dataset used for training (identify patterns):
    - Preparing for the exam
- Test set:
    - used data to evaluate the performance of model:
    - Eg) mid-term exam
- Validation set:
    - to improve the model used during training.(to measure the progress)
    - Eg) Practice questions





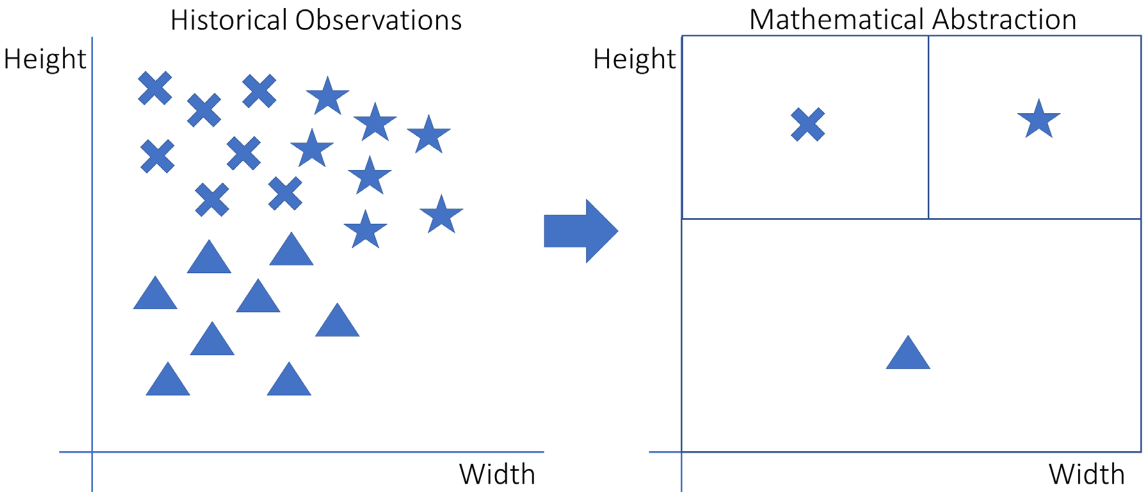
Supervised Learning

- Classification: the output variable is a category instead of values,
  - example 1: “red” or “blue”
  - example 2: “yes” or “no”.
- Take an input value and assign it a class (category) that it fits into based on the training data provided.
- eg) spam email
- Main algorithms:
  - Linear Classifiers
  - Support Vector Machines
  - **K-Nearest Neighbor**
  - Decision Trees
  - Random Forest



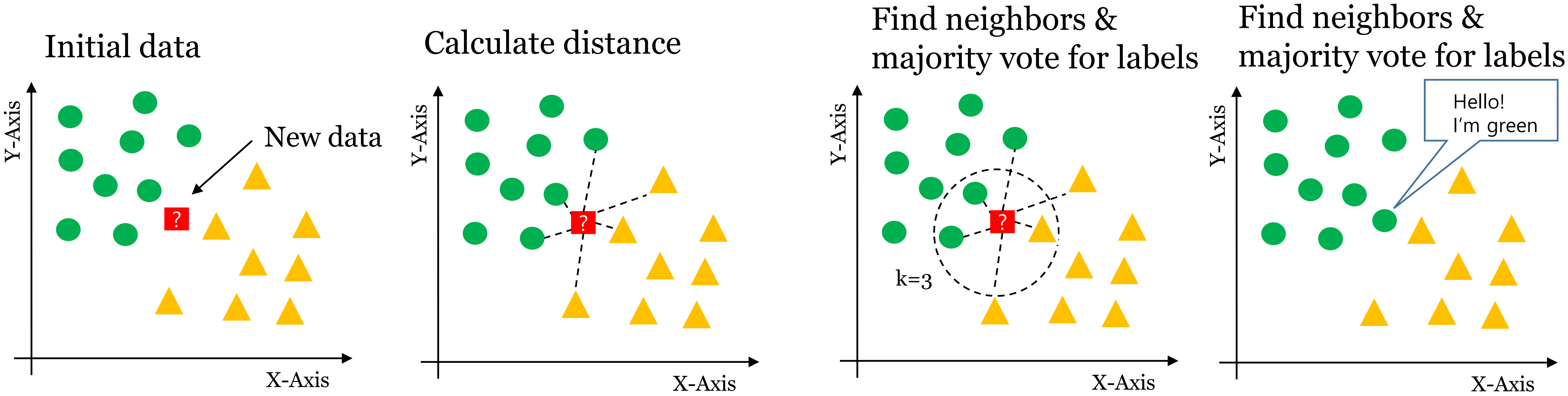
kNN Classification using Scikit-learn

- k-Nearest Neighbor (kNN) is a simple and one of the topmost machine learning algorithm.
- This algorithm is used in finance, political science, image recognition, video recognition, health care etc.
- Political science: classifying potential voters(vote or not vote?)
- Financial institutes can predict the credit ratings of their customers.
- KNN is a non-parametric algorithm (model structure is determined from the dataset)
- It is useful in real world dataset where it does not follow mathematical assumptions.



kNN Classification using Scikit-learn

- A new data point arrives (red square box).
- The kNN algorithm finds the nearest neighbors of this red square. (k closest points)
- It takes the values of those neighbors and uses them as a prediction for the new data point.
- The kNN algorithm is based on the notion that you can predict the features of a data point based on the features of its neighbors.
- To find the data points that are closest to the point of red square, we use Euclidean distance.



Initial data

Calculate distance

Find neighbors & majority vote for labels

Find neighbors & majority vote for labels

A little Python tip: zip( )

```
>>> zip(iterator1, iterator2, iterator3)
```

Zip() function returns a zip object, which is an iterator of tuples where the first item in each passed iterator is paired together, and then the second item in each passed iterator are paired together etc.

If the passed iterators have different lengths, the iterator with the least items decides the length of the new iterator.

```
>>> a = ("John", "Charles", "Mike")
>>> b = ("Jenny", "Christy", "Monica", "Vicky")
>>> x = zip(a, b)
(('John', 'Jenny'), ('Charles', 'Christy'), ('Mike', 'Monica'))
```

A little Python tip: zip and dictionary

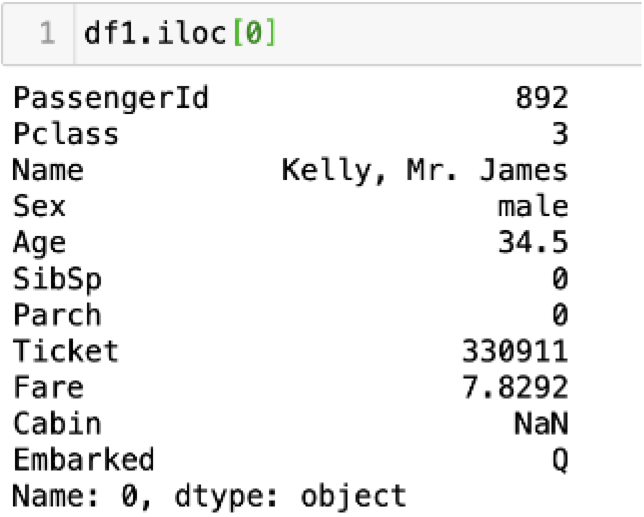
```
dictionary = dict(zip(keys, values))
a = [‘a’, ‘b’, ‘c’]
b = [1, 2, 3]

print(dictionary)
{ ‘a’ : 1, ‘b’ : 2 , ‘c’ : 3 }

>>> numbers =[1, 2, 3]
>>> letters = [‘A’, ‘B’]
>>> list(zip(numbers, letters))
[(‘1’, ‘A’), (2, ‘B’)]
```

Pandas iloc vs loc to locate (index) certain row or column

- loc = location by writing the name of the column or condition
- iloc = integer location.

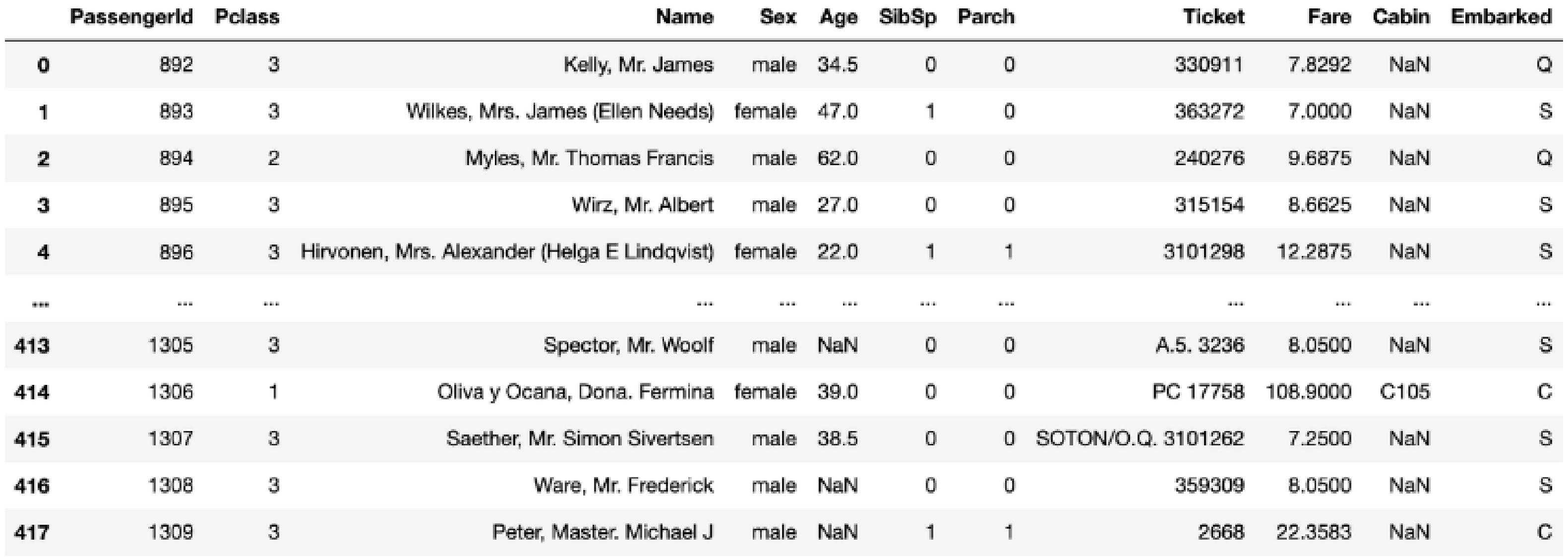
>>> df.iloc[row index, column index]

>>> df1.iloc[0]



```
df1.iloc[0]
```

```
PassengerId               892
Pclass                      3
Name          Kelly, Mr. James
Sex                      male
Age                      34.5
SibSp                       0
Parch                       0
Ticket                 330911
Fare                   7.8292
Cabin                     NaN
Embarked                    Q
Name: 0, dtype: object
```



	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

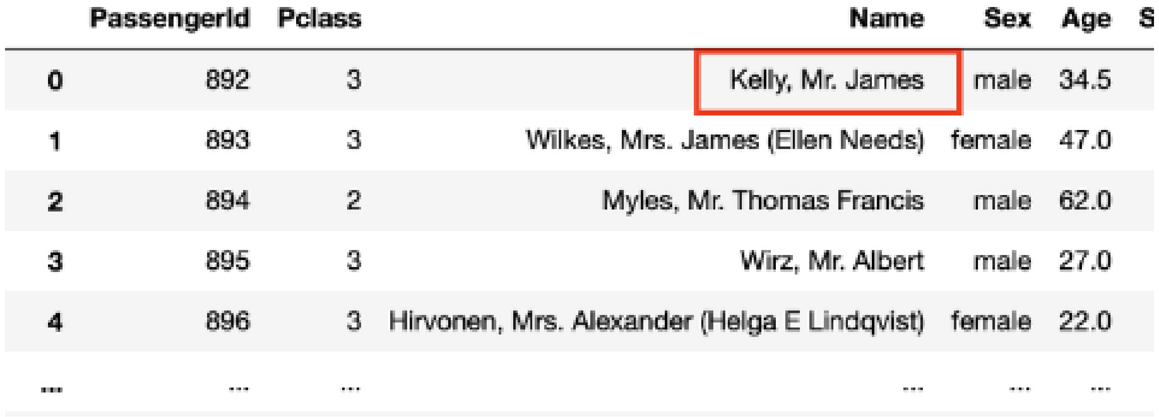
Pandas iloc vs loc to locate (index) certain row or column

- iloc = integer location.

>>> df.iloc[row index, column index]

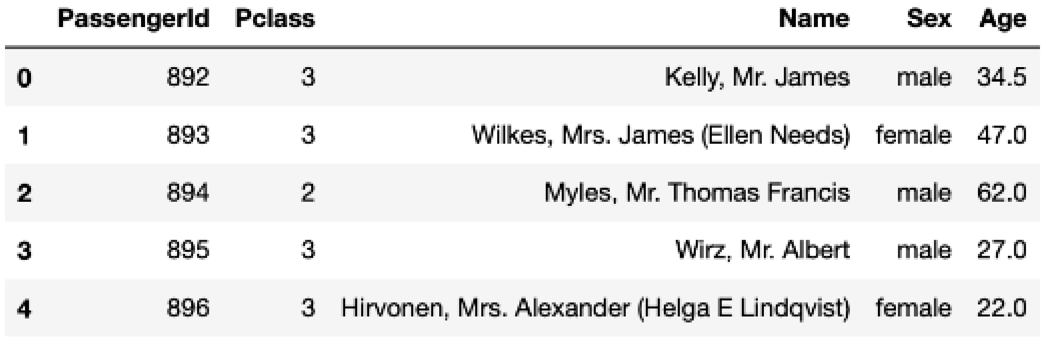
>>> df1.iloc[0,2]

'Kelly, Mr. James'



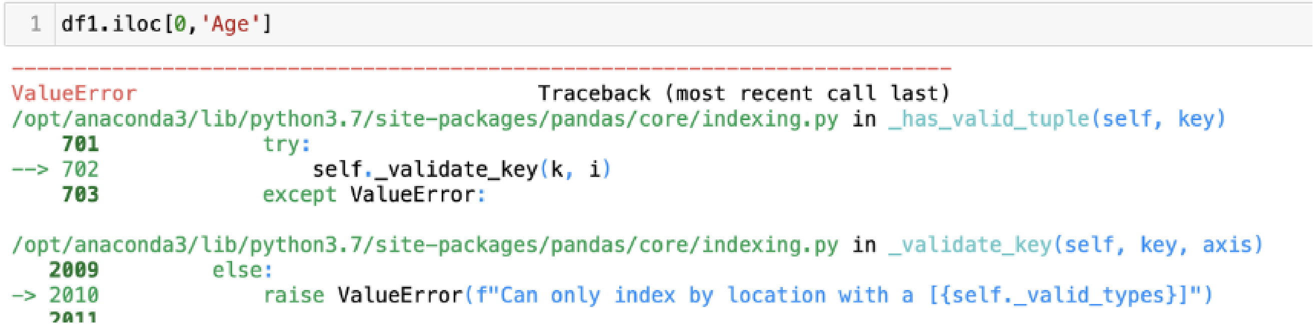
	PassengerId	Pclass	Name	Sex	Age	S
0	892	3	Kelly, Mr. James	male	34.5	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	
3	895	3	Wirz, Mr. Albert	male	27.0	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	
...	...	...	...	...	...	

>>> df1.iloc[:5,:5]



	PassengerId	Pclass	Name	Sex	Age
0	892	3	Kelly, Mr. James	male	34.5
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0
2	894	2	Myles, Mr. Thomas Francis	male	62.0
3	895	3	Wirz, Mr. Albert	male	27.0
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0

Be Careful!!!!!!!



```
df1.iloc[0,'Age']
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
/opt/anaconda3/lib/python3.7/site-packages/pandas/core/indexing.py in _has_valid_tuple(self, key)
    701             try:
--> 702                 self._validate_key(k, i)
    703             except ValueError:

/opt/anaconda3/lib/python3.7/site-packages/pandas/core/indexing.py in _validate_key(self, key, axis)
   2009         else:
-> 2010             raise ValueError(f"Can only index by location with a [{self._valid_types}]")
   2011
```

ValueError: Location based indexing can only have [integer, integer slice (START point is INCLUDED, END point is EXCLUDED), listlike of integers, boolean array] types

Let’s launch our Jupyter notebook