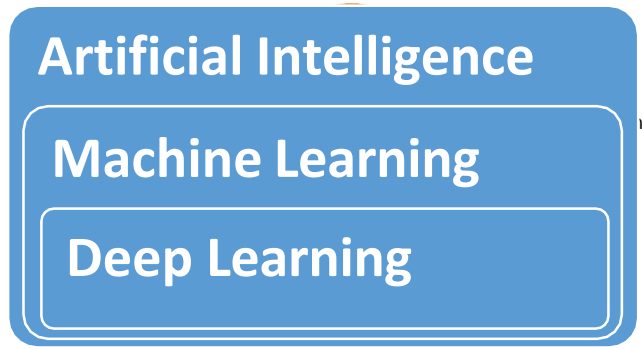
Python-Machine Learning

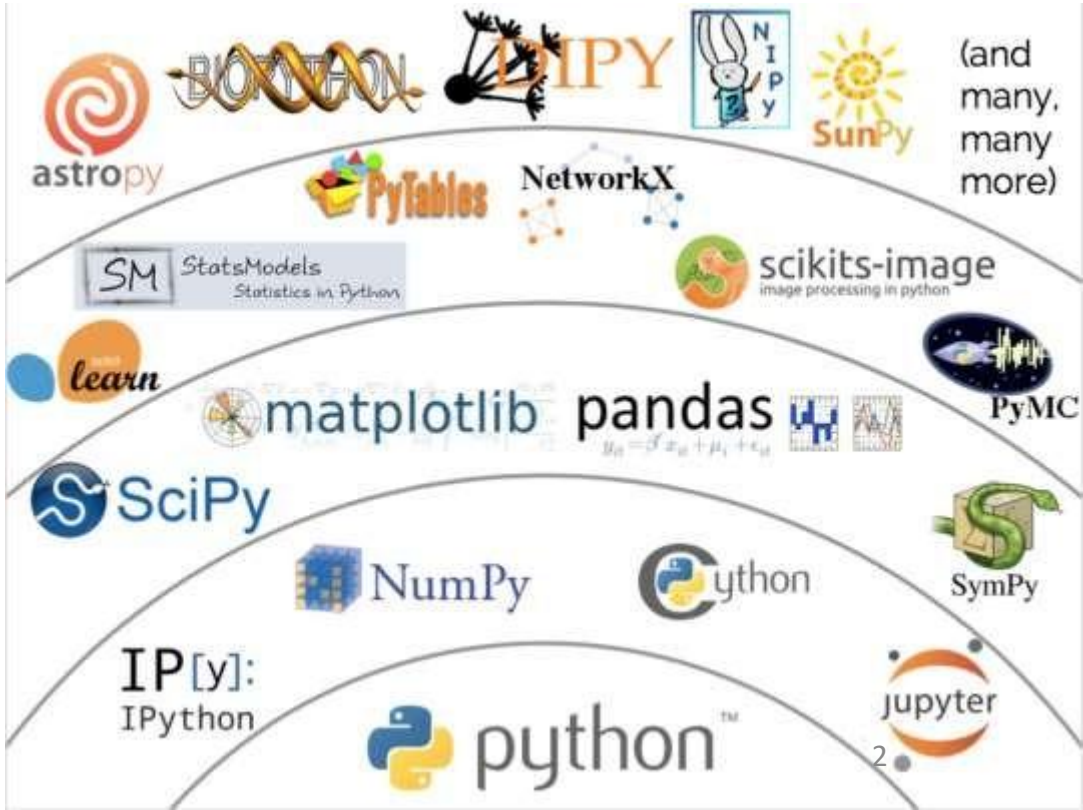
using Scikit-Learn package

Dr. Sarwan Singh

Agenda

- Introduction
- History, need
- Why Machine Learning Matters
- Type of ML-Supervised vs unsupervised
- Classification, Regression, Clustering
- Cheat sheet
- Machine learning flow





Introduction

- Machine learning is where computational and algorithmic skills of data science meet the statistical thinking of data science,
- The result is a collection of approaches to inference and data exploration that are *not about effective theory* so much as *effective computation*.
- Better to think of machine learning as a *means of building models of Data*
- Machine learning along with entire Data Science ecosystem is trying to make this mathematical, model-based "learning" as same as "learning" exhibited by the human brain.

- “A computer program is said to learn

  from experience *E* with respect to some class of tasks *T* and performance measure *P*

  if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.” *by Tom M. Mitchell*

"~~Can machines think~~?"

is replaced with the question

"Can machines do what we (as thinking entities) can do?"

Alan Turing

Source : wikipedia

History

- Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM
- In earlier times scientist attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks“
- Probabilistic reasoning was also employed in various automated medical diagnosis programs

Why Machine Learning Matters

With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

- **Computational finance**, for credit scoring and algorithmic trading
- **Image processing and computer vision**, for face recognition, motion detection, and object detection
- **Computational biology**, for tumor detection, drug discovery, and DNA sequencing
- **Energy production**, for price and load forecasting
- **Automotive, aerospace, and manufacturing**, for predictive maintenance
- **Natural language processing**, for voice recognition applications

Source: Mathworks.com

Why Machine Learning is needed

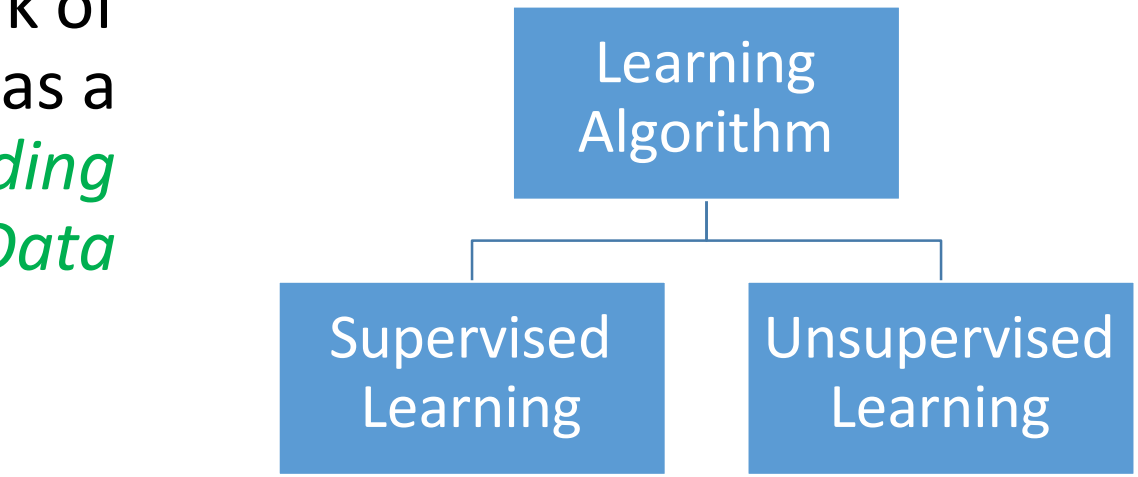
- If Programmer start making use cases / rules for complex system, then it will result in a large number of rules and exceptions.
- Machine Learning is needed in cases where humans cannot directly write a program to handle each and every case.
- So it's better to have a machine (~~rather than human~~) that learns from a large training set.

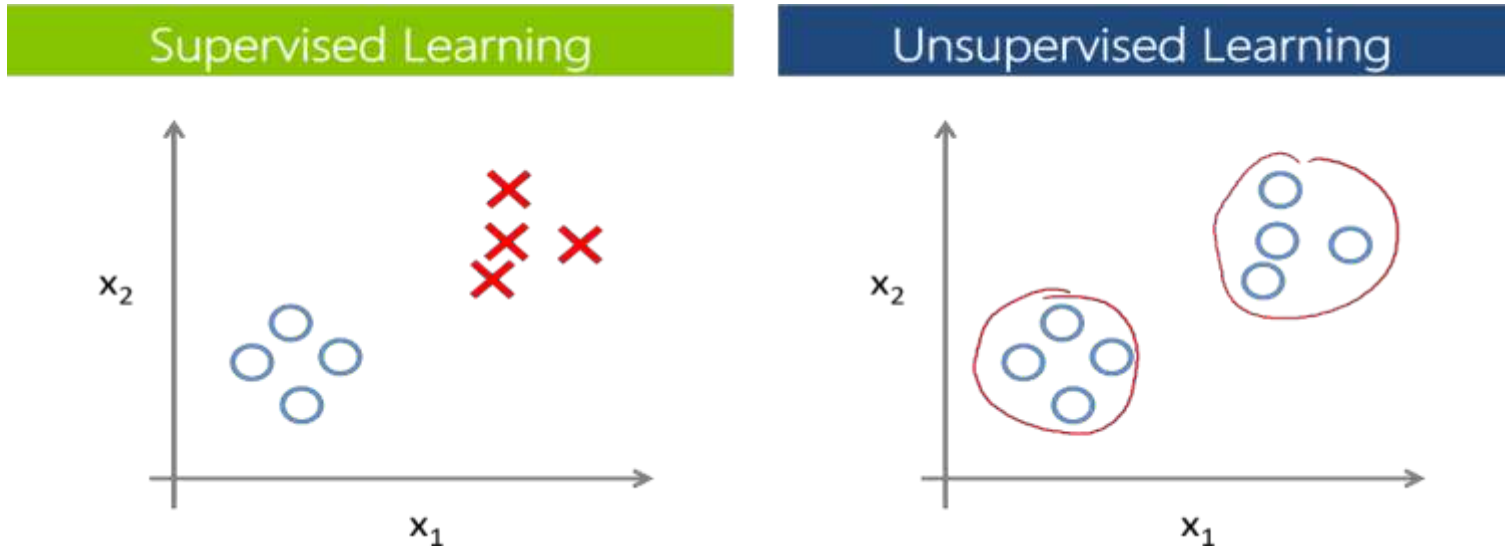
according to the definition earlier:

- Task (T): recognizing and classifying handwritten words within images
- Performance measure (P): percent of words correctly classified
- Training experience (E): a database of handwritten words with given classifications

Major Classes of Learning Algorithms

Better to think of machine learning as a *means of building models of Data*

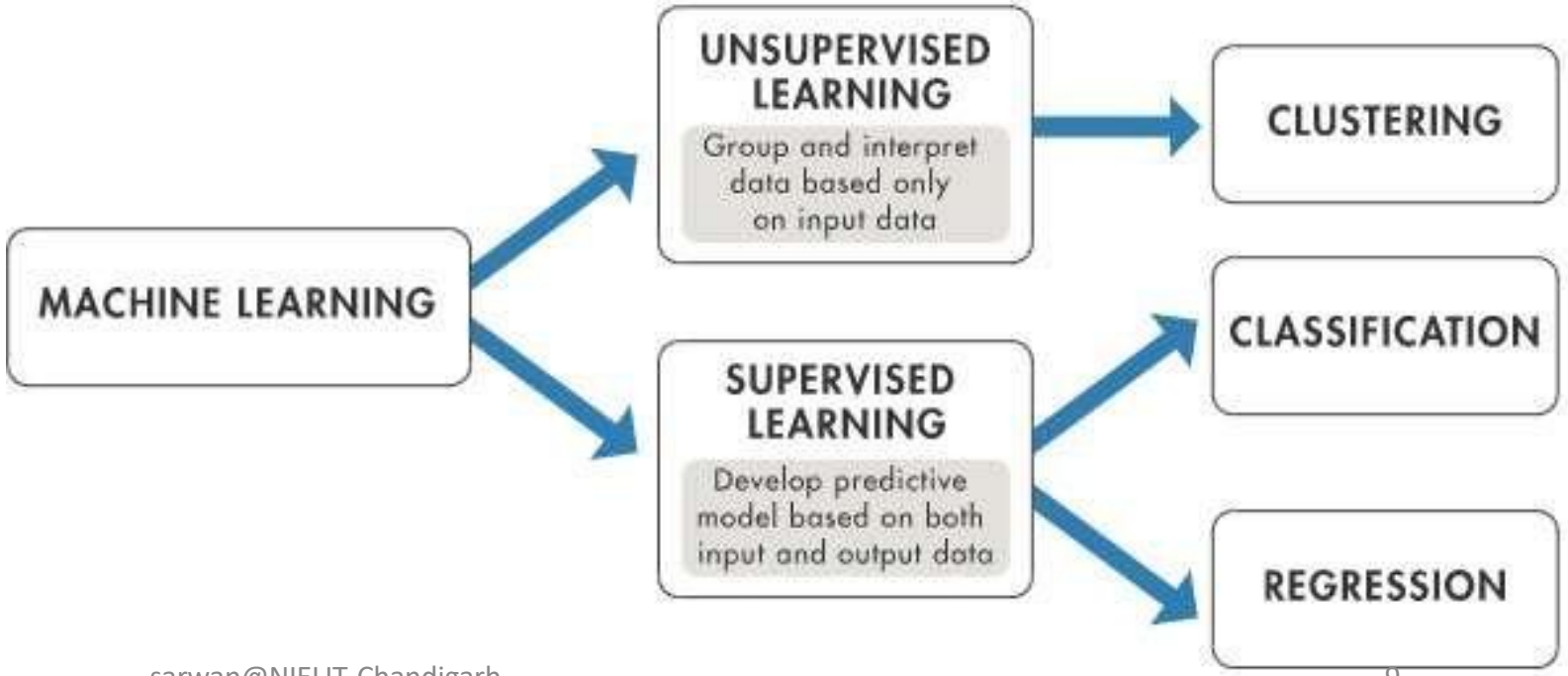




Supervised learning

learning from exemplars

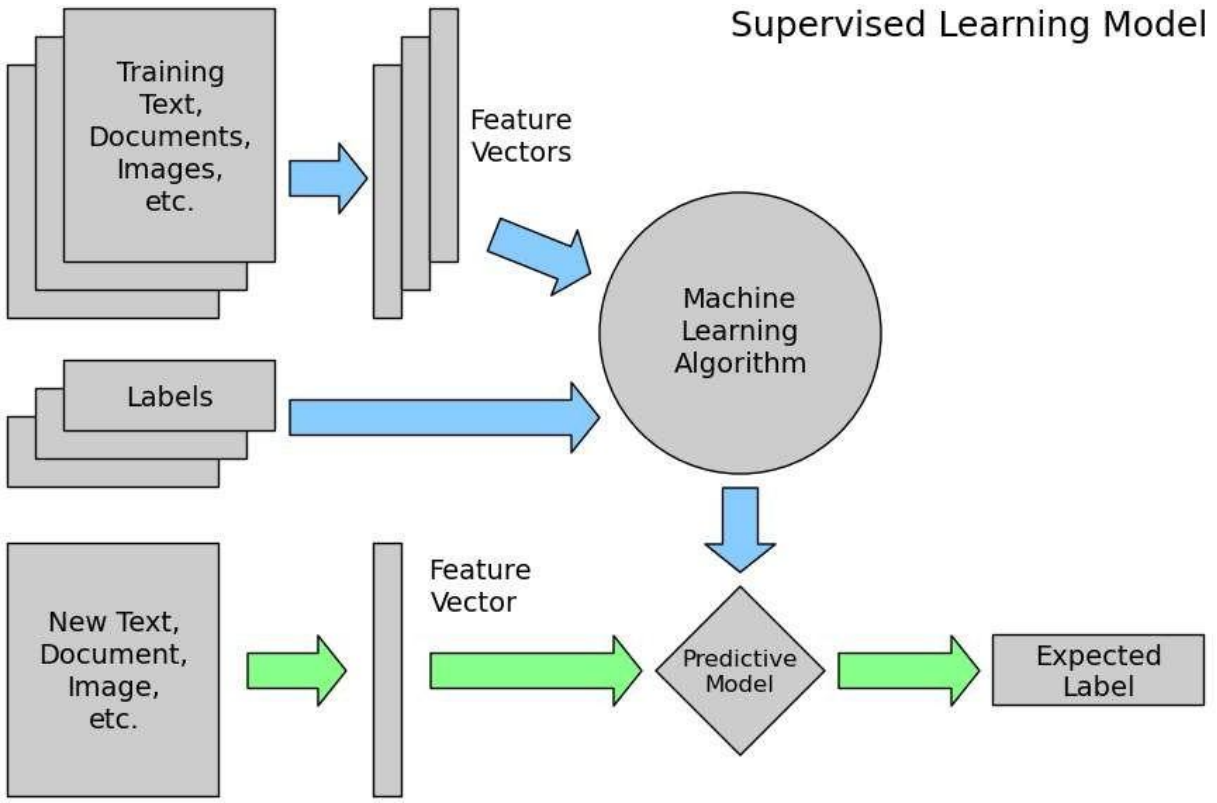
- The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- The training process continues until the model achieves a desired level of accuracy on the training data. once this model is determined, it can be used to apply labels to new, unknown data.



Source : wikipedia

Supervised learning

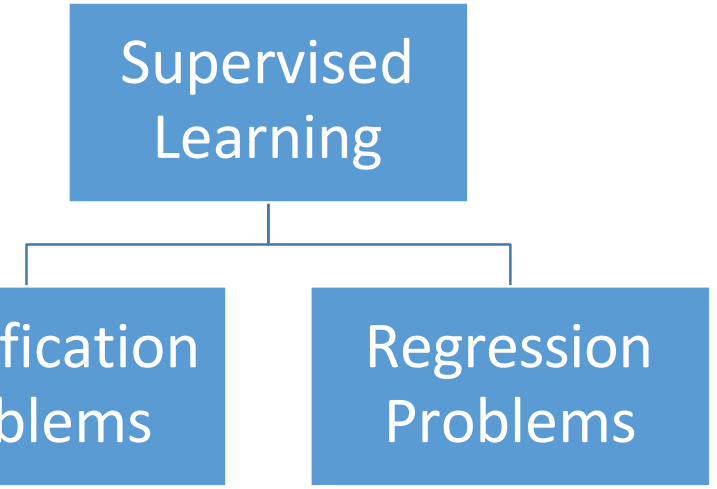
- Semi-supervised learning:
  - the computer is given only an incomplete training signal.
- Active learning:
  - the computer can only obtain training labels for a limited set of instances (based on a budget)
- Reinforcement learning:
  - training data (in form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle, game

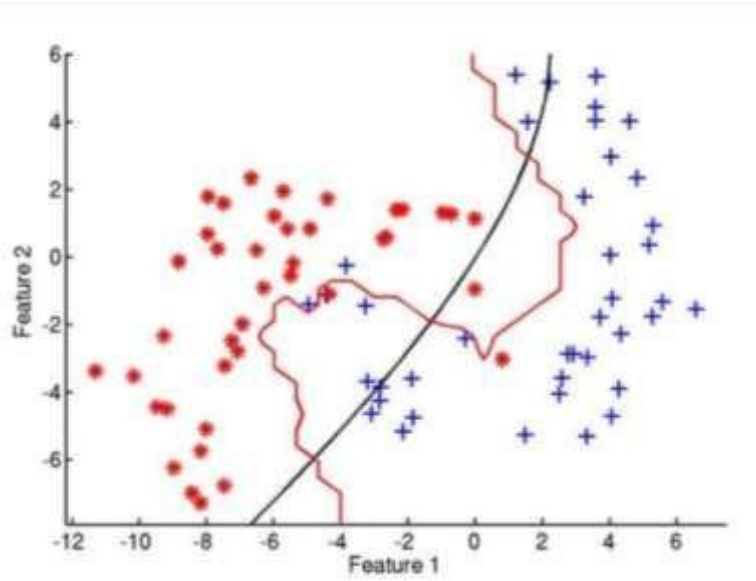


Source : wikipedia

Supervised Learning: Classification Problems

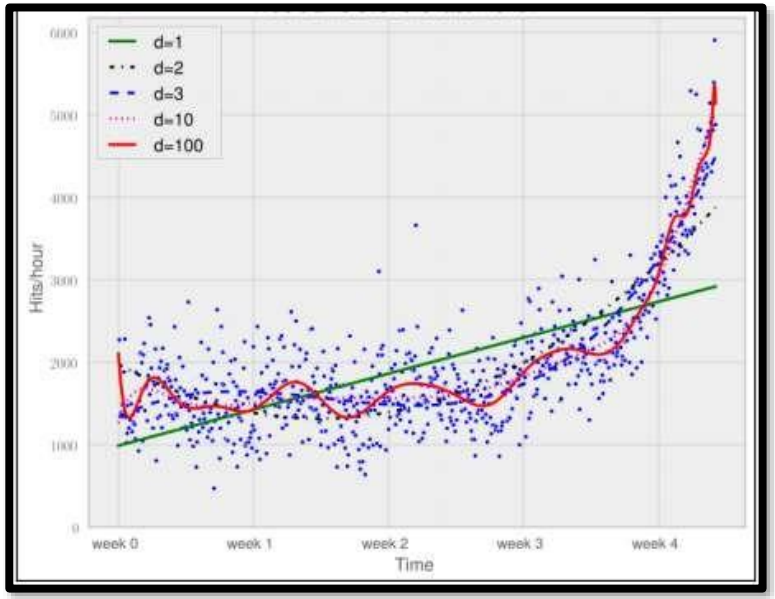
- Consists of taking input vectors and deciding which of the N classes they belong to, based on training from exemplars of each class
- Find 'decision boundaries' that can be used to separate out the different classes.
- It is to decide which class the current input belongs to.





Supervised Learning: Regression Problems

- Given some data, you assume that those values come from some sort of function and try to find out what the function is.
- Try to fit a mathematical function that describes a curve, such that the curve passes as close as possible to all the data points.
- Regression is essentially a problem of function approximation or interpolation



Unsupervised learning

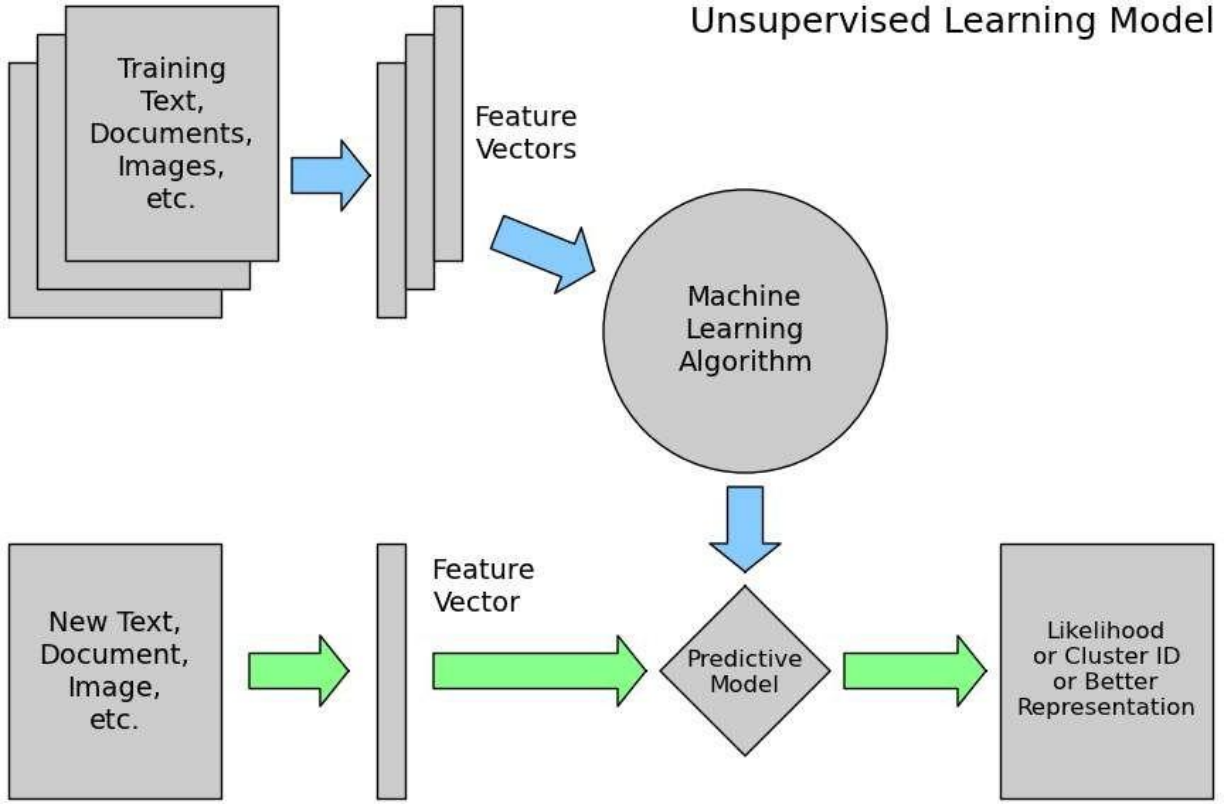
" letting the dataset speak for itself "

- 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

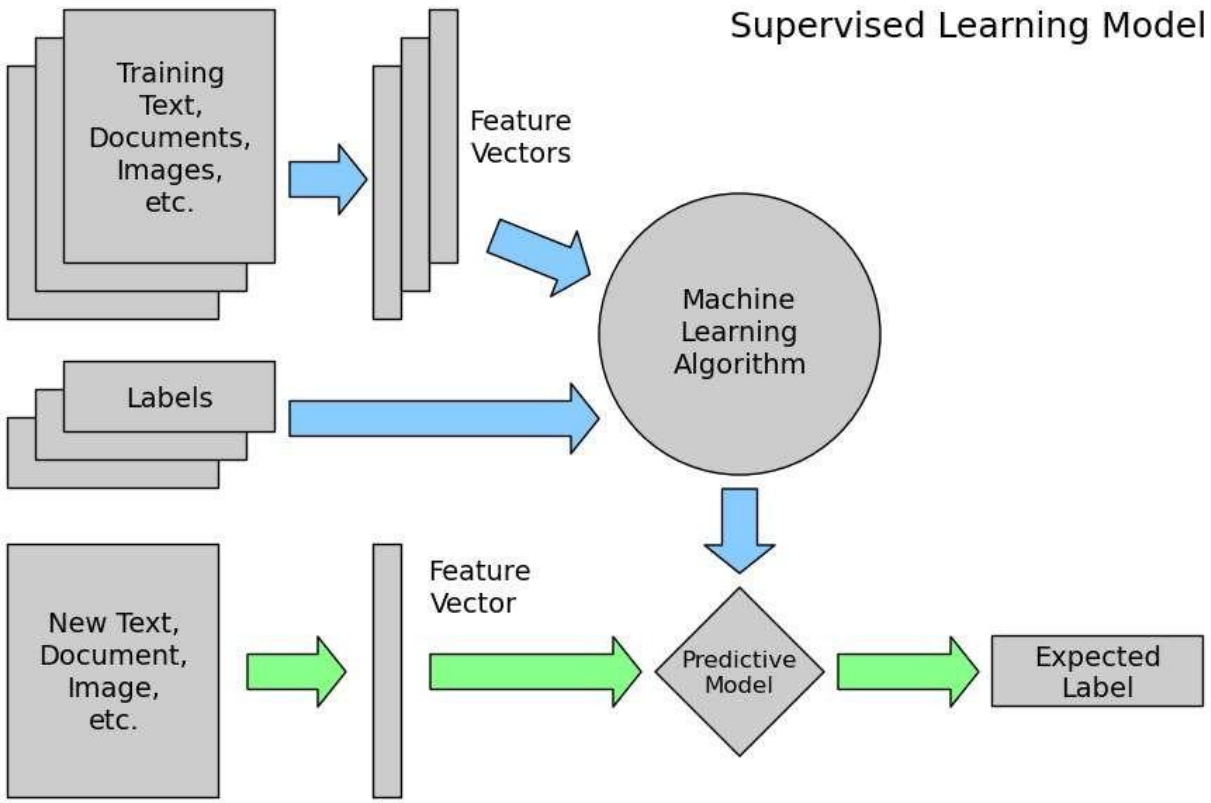
  a means towards an end (feature learning).
- It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.
- These models include tasks such as clustering and dimensionality reduction.
  - Clustering algorithms identify distinct groups of data, while
  - Dimensionality reduction algorithms search for more succinct representations of the data.

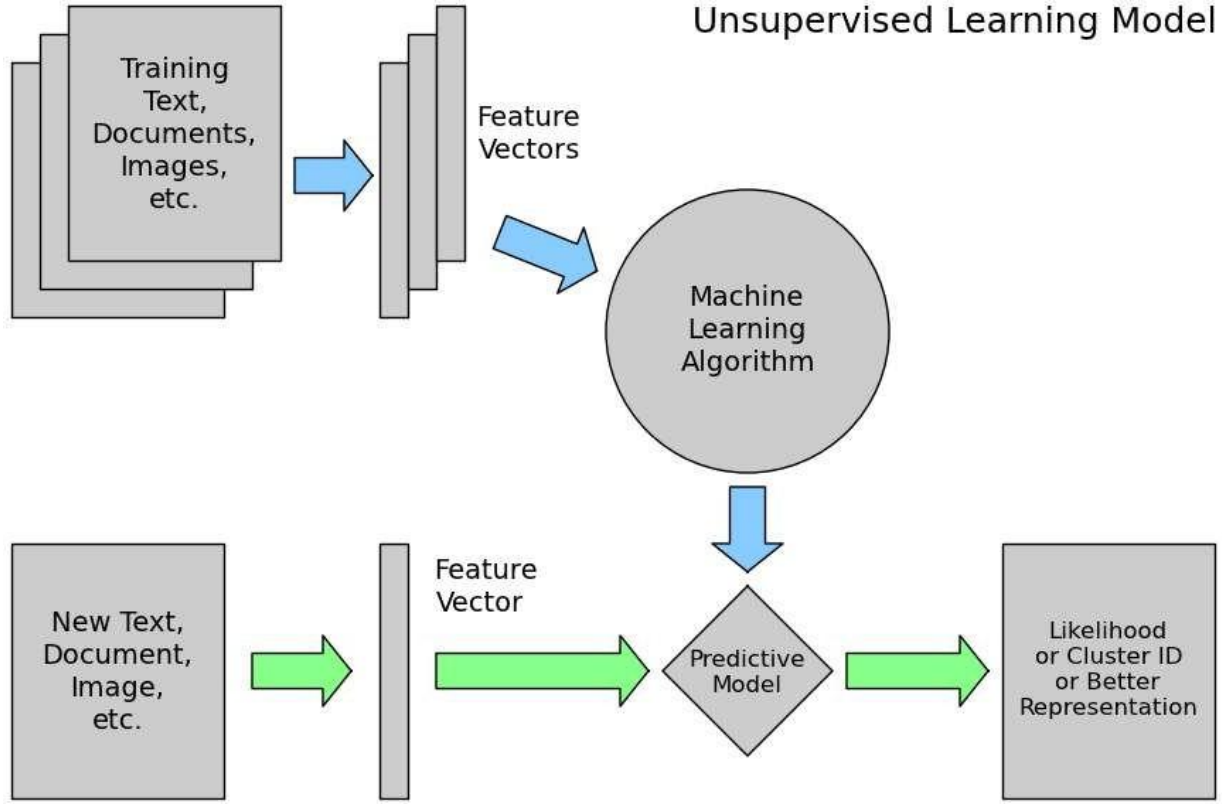
Unsupervised learning - Clustering

- The aim of unsupervised learning is to find clusters of similar inputs in the data without being explicitly told that some datapoints belong to one class and the other in other classes.
- The algorithm has to discover this similarity by itself

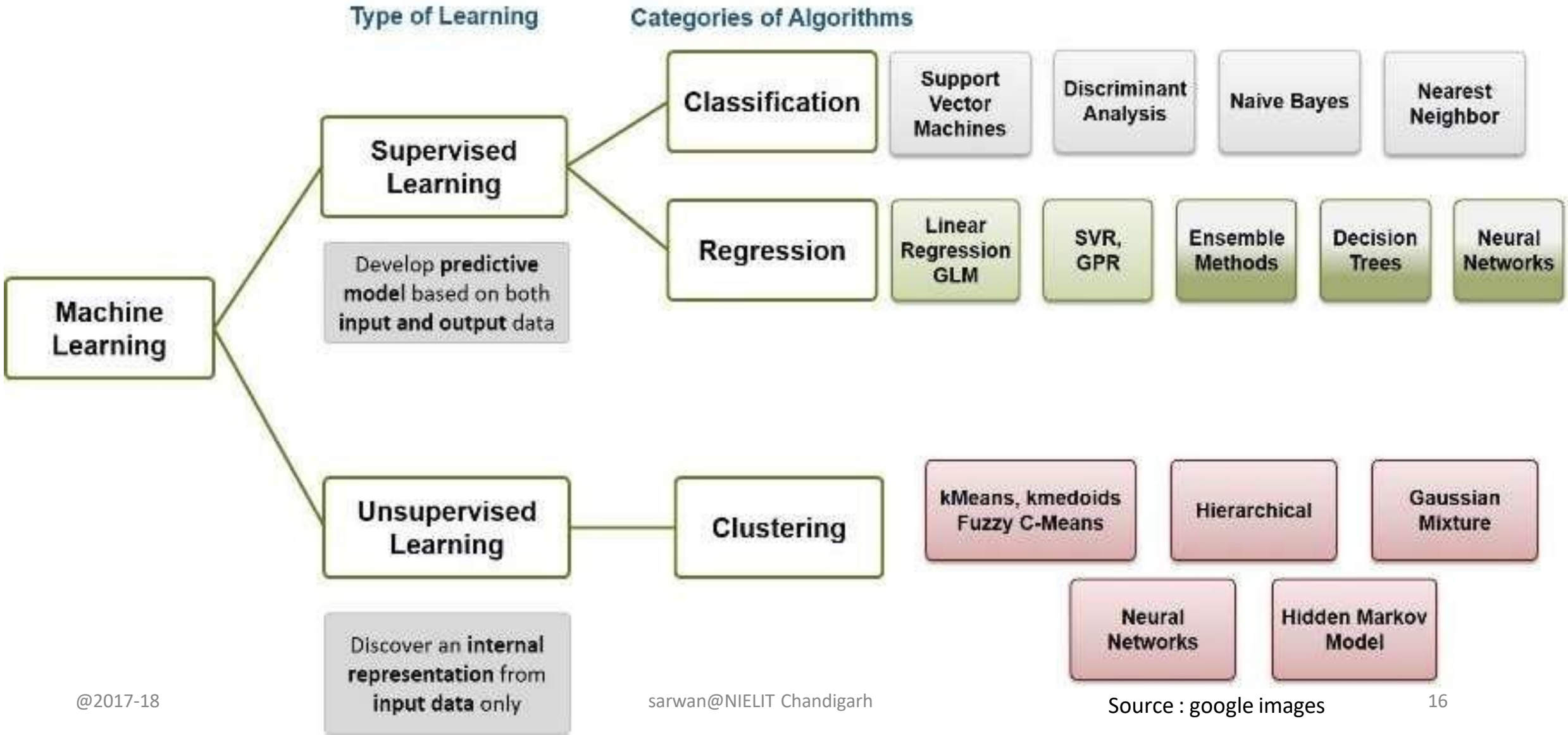


Supervised vs Unsupervised learning

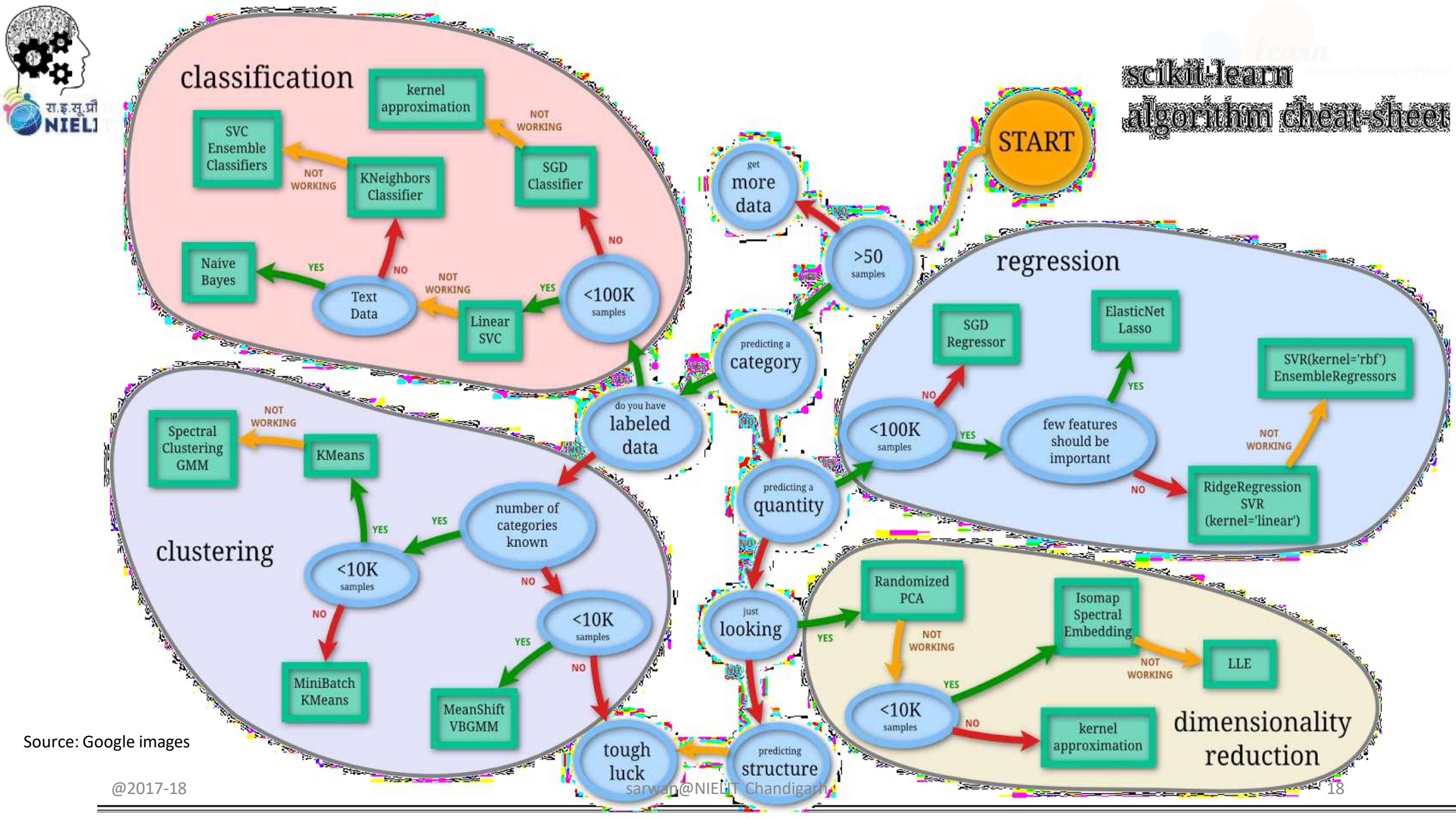




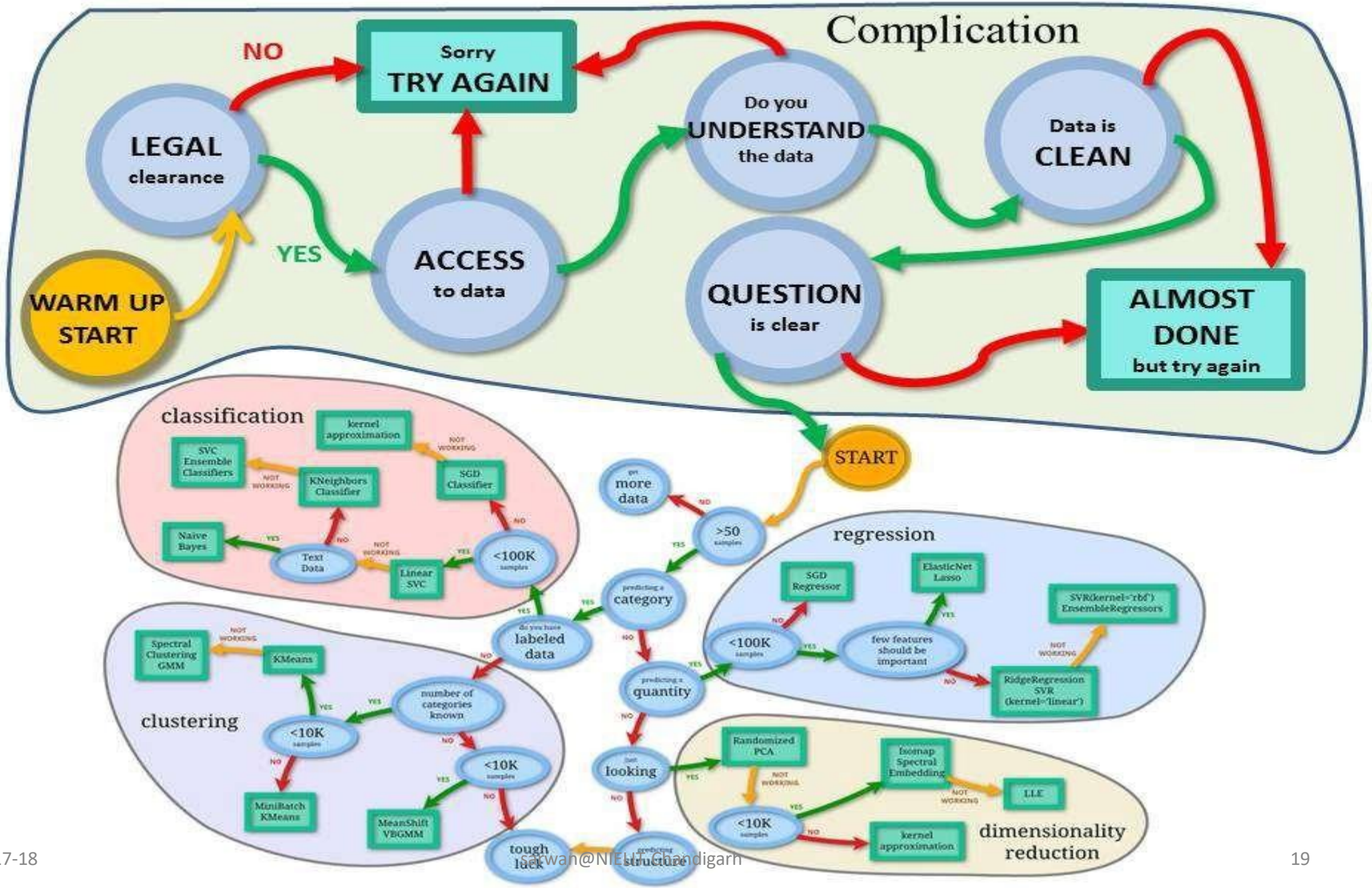
Category of Algorithms

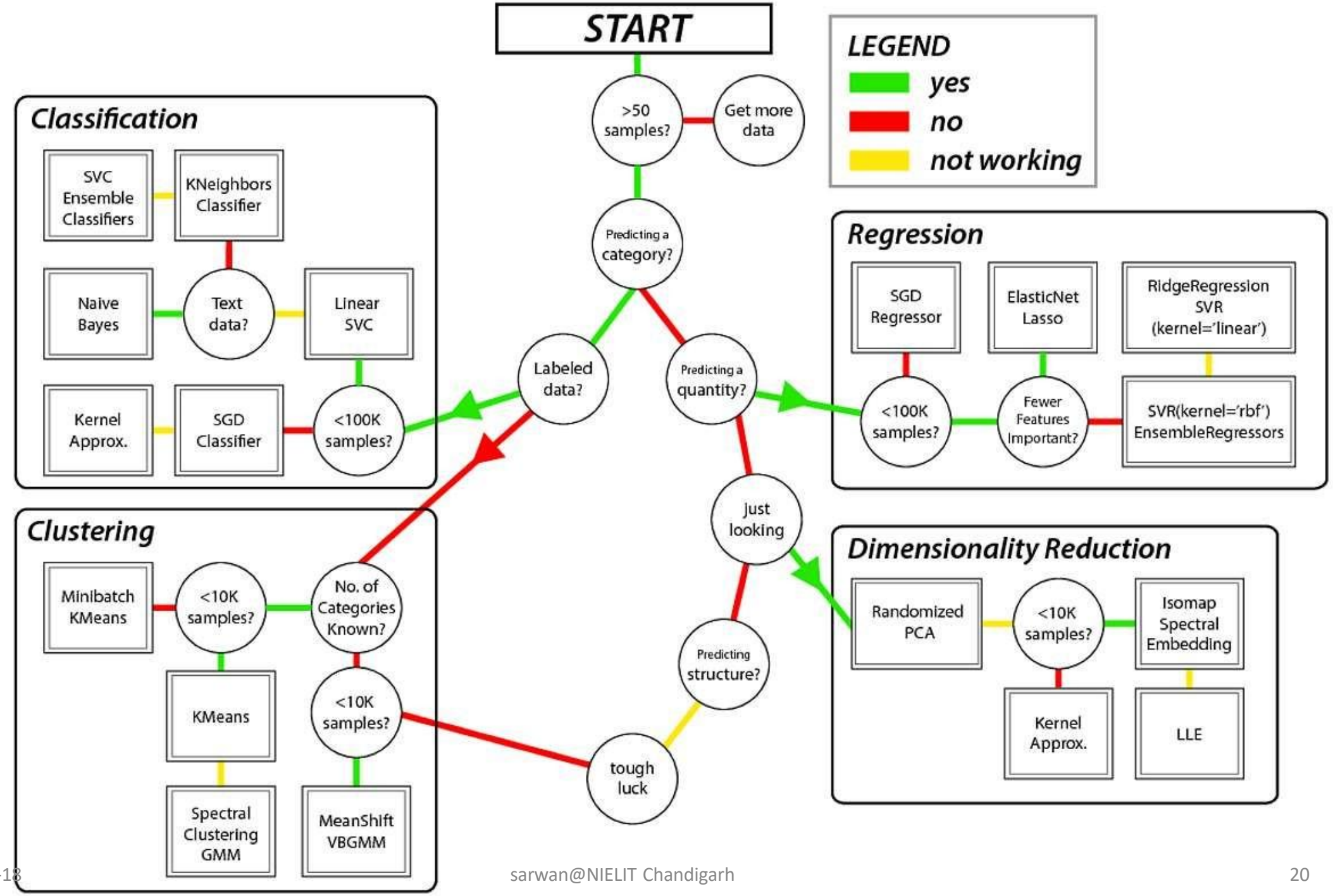


- Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- Note that *scikit-learn is used to build models*. It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)



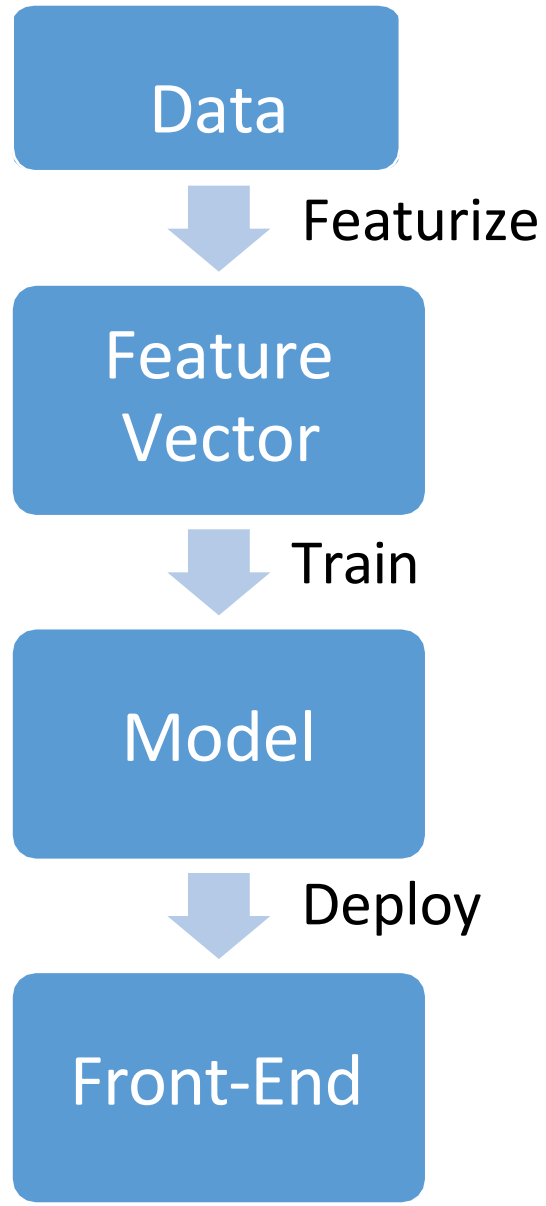
Source: Google images


Complication
NO
Sorry
TRY AGAIN
LEGAL
clearance
YES
ACCESS
to data
WARM UP
START
Do you
UNDERSTAND
the data
Data is
CLEAN
QUESTION
is clear
ALMOST
DONE
but try again
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble
Classifiers
NOT WORKING
KNeighbors
Classifier
SGD
Classifier
Naive
Bayes
YES
NO
Text
Data
NOT WORKING
Linear
SVC
<100K
samples
NO
YES
more
data
NO
>50
samples
YES
predicting a
category
YES
NO
do you have
labeled
data
NO
YES
regression
SGD
Regressor
NO
ElasticNet
Lasso
YES
<100K
samples
YES
few features
should be
important
NO
SVR(kernel='rbf')
EnsembleRegressors
NOT WORKING
RidgeRegression
SVR
(kernel='linear')
predicting a
quantity
YES
NO
Spectral
Clustering
GMM
NOT WORKING
KMeans
clustering
YES
number of
categories
known
YES
NO
<10K
samples
NO
<10K
samples
YES
NO
MiniBatch
KMeans
MeanShift
VBGMM
looking
NO
YES
Randomized
PCA
NOT WORKING
Isomap
Spectral
Embedding
NOT WORKING
LLE
<10K
samples
YES
NO
kernel
approximation
dimensionality
reduction
tough
luck
predicting a
structure


START
>50 samples?
Get more data
LEGEND
yes
no
not working
Predicting a category?
Labeled data?
Predicting a quantity?
Just looking
Predicting structure?
tough luck
Classification
SVC Ensemble Classifiers
KNeighbors Classifier
Naive Bayes
Text data?
Linear SVC
Kernel Approx.
SGD Classifier
<100K samples?
Regression
SGD Regressor
ElasticNet Lasso
RidgeRegression SVR (kernel='linear')
<100K samples?
Fewer Features Important?
SVR(kernel='rbf') EnsembleRegressors
Clustering
Minibatch KMeans
<10K samples?
No. of Categories Known?
KMeans
<10K samples?
Spectral Clustering GMM
MeanShift VBGMM
Dimensionality Reduction
Randomized PCA
<10K samples?
Isomap Spectral Embedding
Kernel Approx.
LLE

Machine Learning Workflow

- 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).
- It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.
- These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data.



Where ML is used

ML in Healthcare

- **Disease Diagnosis:** Detect cancer, diabetes, and heart conditions.
- **Medical Imaging:** Analyze X-rays, MRIs, and CT scans.
- **Personalized Medicine:** Tailored treatments and drug recommendations.
- **Predictive Analytics:** Forecast outbreaks and recovery rates.

Where ML is used

ML in Entertainment and Media

- **Streaming Services:** Recommend movies, shows, and music.
- **Content Creation:** Generate captions and scripts.
- **Ad Targeting:** Personalized advertisements.
- **Game Development:** Adaptive AI for non-player characters.

Where ML is used

ML in Manufacturing

- **Predictive Maintenance:** Forecast equipment failures.
- **Quality Control:** Detect defects with image recognition.
- **Supply Chain Optimization:** Improve logistics and inventory.
- **Robotics:** Enable autonomous tasks in factories.

Where ML is used

ML in Agriculture

- **Crop Monitoring:** Analyze soil and weather conditions.
- **Yield Prediction:** Forecast crop production.
- **Pest Control:** Monitor and mitigate infestations.
- **Smart Irrigation:** Optimize water usage.

Where ML is used

ML in Energy

- **Smart Grids:** Optimize energy distribution.
- **Renewable Energy Forecasting:** Predict wind and solar generation.
- **Energy Consumption Prediction:** Forecast demands.
- **Anomaly Detection:** Identify faults in grids.

Where ML is used

ML in Cybersecurity

- **Threat Detection:** Real-time malware and phishing detection.
- **User Authentication:** Biometric enhancements.
- **Behavioral Analysis:** Detect unusual activities.
- **Spam Filtering:** Block unwanted messages.

Where ML is used

ML in Human Resources

- **Talent Acquisition:** Automate resume screening.
- **Employee Retention:** Predict turnover rates.
- **Training Programs:** Personalize modules.
- **Workforce Optimization:** Improve team efficiency.

Where ML is used

ML in Environment

- **Climate Modeling:** Predict weather and climate changes.
- **Wildlife Monitoring:** Track and protect species.
- **Disaster Management:** Forecast floods and earthquakes.
- **Pollution Control:** Monitor air and water quality.

Where ML is used

ML in Space Exploration

- **Satellite Data Analysis:** Process vast space datasets.
- **Rover Navigation:** Autonomous extraterrestrial movement.
- **Asteroid Tracking:** Predict celestial paths.
- **Spacecraft Operations:** Optimize energy and resources.

Where ML is used

ML in Social Media

- **Content Moderation:** Detect harmful content.
- **Sentiment Analysis:** Gauge public opinions.
- **Trend Analysis:** Identify trending topics.
- **Chatbots:** Automate customer service.

Where ML is used

ML in Research and Development

- **Drug Discovery:** Accelerate medicine development.
- **Scientific Research:** Analyze datasets in multiple domains.
- **Language Translation:** Accurate multilingual tools.

Where ML is used

ML in Retail and E-Commerce
- **Recommendation Systems:** Suggest products based on preferences.
- **Dynamic Pricing:** Optimize prices dynamically.
- **Inventory Management:** Predict stock requirements.
- **Customer Segmentation:** Target specific demographics.

ML in Transportation
- **Autonomous Vehicles:** Self-driving cars using sensor data.
- **Traffic Prediction:** Optimize routes and reduce congestion.
- **Fleet Management:** Monitor vehicle performance.
- **Ride-Sharing:** Efficient driver-passenger matching.

ML in Education
- **Personalized Learning:** Adapt content to individual paces.
- **Automated Grading:** Evaluate assignments and quizzes.
- **Dropout Prediction:** Identify and intervene for at-risk students.
- **Content Recommendation:** Suggest tailored resources.

Conclusion

Transforming Industries

- ML is making systems smarter and more efficient.
- It addresses complex problems previously unsolvable.
- Revolutionizing industries globally.