Introduction to Data Science

Pritha Banerjee

University of Calcutta

banerjee.pritha74@gmail.com

March 2, 2023

Overview

What is Data Science

- A concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data [Chiko Hayashi, 1998].
- Historically: Science evolved through empirical observations, basic theoretical research based on general principles, problem solving through computing machines and now the focus is on data available due to advancement of information technology and internet.
- Earlier data were analyzed to explain the data set
- Analytics is the discovery, interpretation and communication of meaningful patterns in data, which can be used for effective decision making and/or prediction.

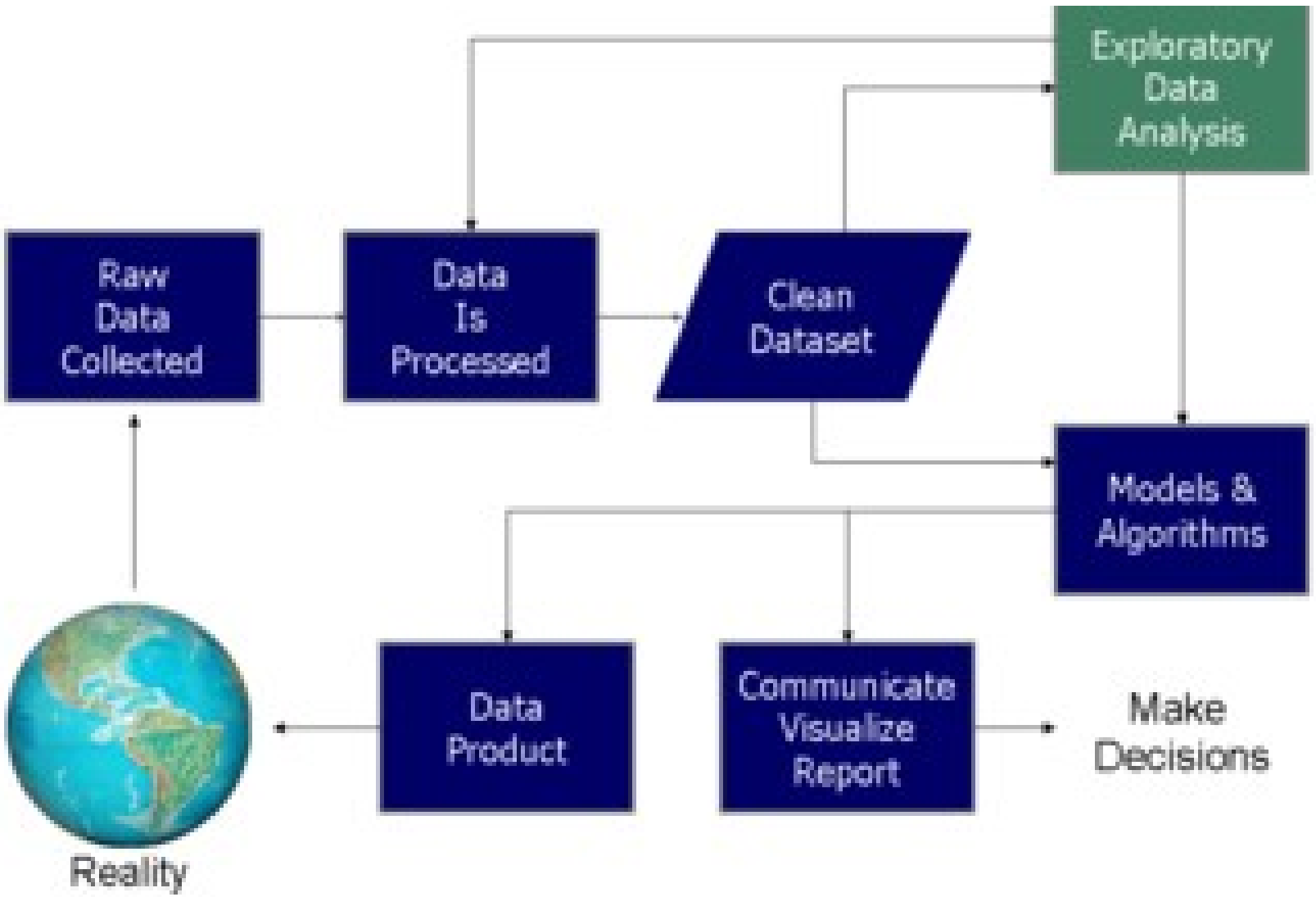
Datafication

- A process of taking all aspects of life and turning them into data. Example:
  - google glass datafies the gaze of a person
  - Twitter datafies thoughts of people
  - LinkedIn datafies professional network
  - sensors, cameras datafies objects, behaviours in its vicinity
- datafication can transform data and information into new form of values, leading to increased efficiency through automatrion
- entrepreneurs gets opportunity to make money out of datafication; there comes the ethics and confidentiality of one's personal data.

Background for Data Science

- Typically Data Science team is multidisciplinary
- Skills required are: Matrix Algebra, Statistics, Machine Learning, Computer Science, Data Visualization, Domain Expertise etc.

Data science Process (1)



Data science Process (2)

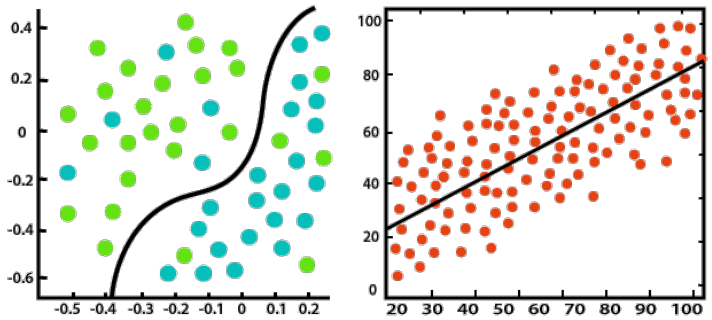
- Pose the question to be answered by data science process: Problem could be a classification problem or a prediction problem or it could be a hypothesis to be validated.
- Raw data is collected from the real world
- Data is processed through pipeline of data wrangling/munging steps that include parsing, scraping, joining to format unstructured raw data into a suitable form for later analytics
- Data is cleaned such that dupicate data is removed, fields/ attributes are well defined, outlier data is removed, missing data is handled properly and finally data is validated (is it the right data for analysis of the question asked)
- Exploratory data analysis and visualization tools helps in identifying and removing data outliers, identifying trends in time and space, uncover patterns related to the target, creating hypotheses and testing them through experiments identifying new sources of data

Data science Process (3)

- Use appropriate Machine Learning algorithm or Statistical models to answer the question posed.
- Decisions can be made from the previous step and communicated through different visualization tools.
- Further the model could be build as a data product and given back to the real world for further data acquision and analysis.

Types of Problems

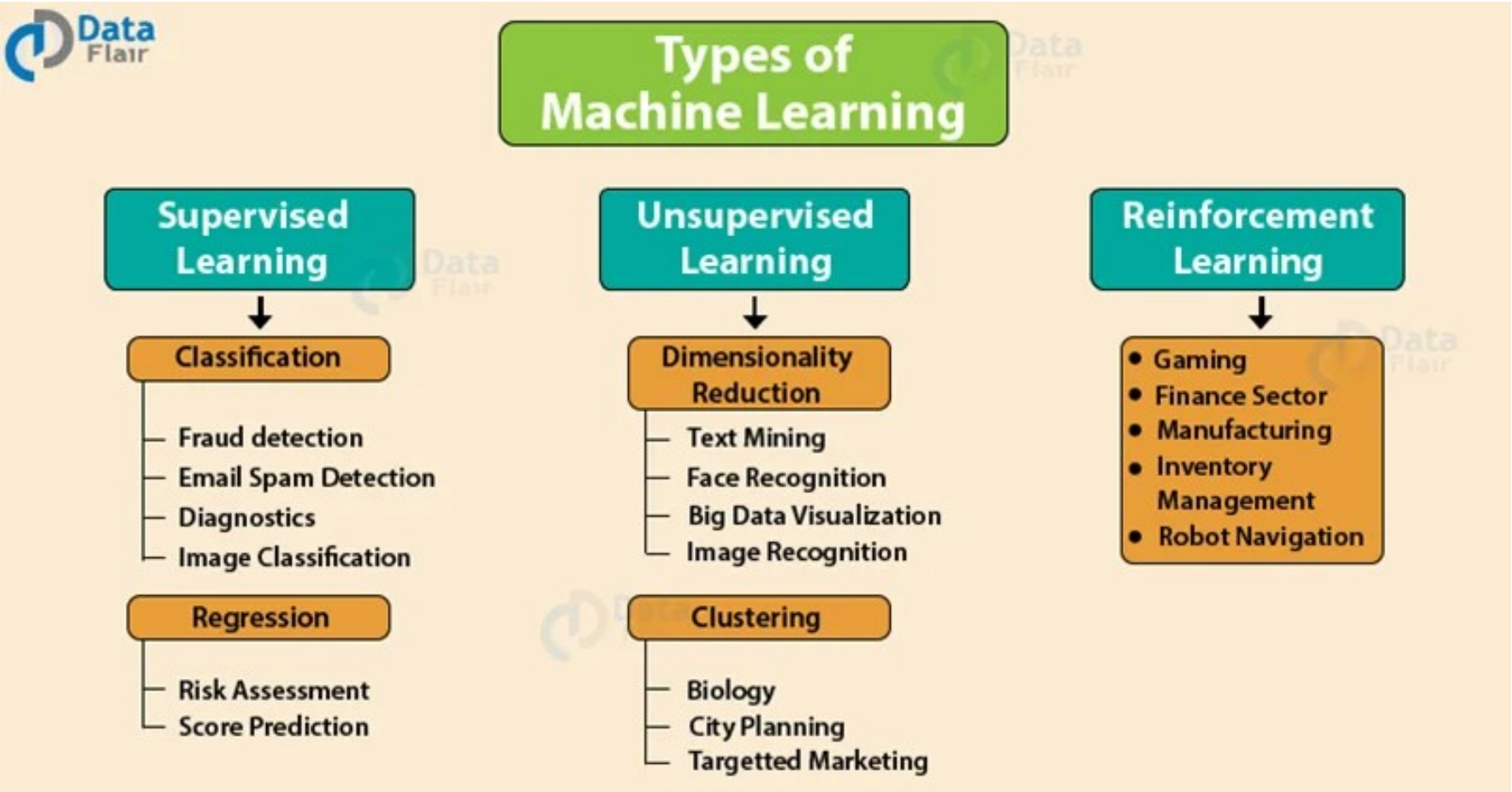
- Classification Problem: classification problems are those which require the given data set to be classified in two or more categories.
- Function approximation Problem: Given a set of dependent variable y_i corresponding to a set of independent variables x_i, function approximation problem is to find a functional form $y = f(x)$ that best represent the data. Once a function is obtained, that can be used to predict an output y for a new set of x_i s.



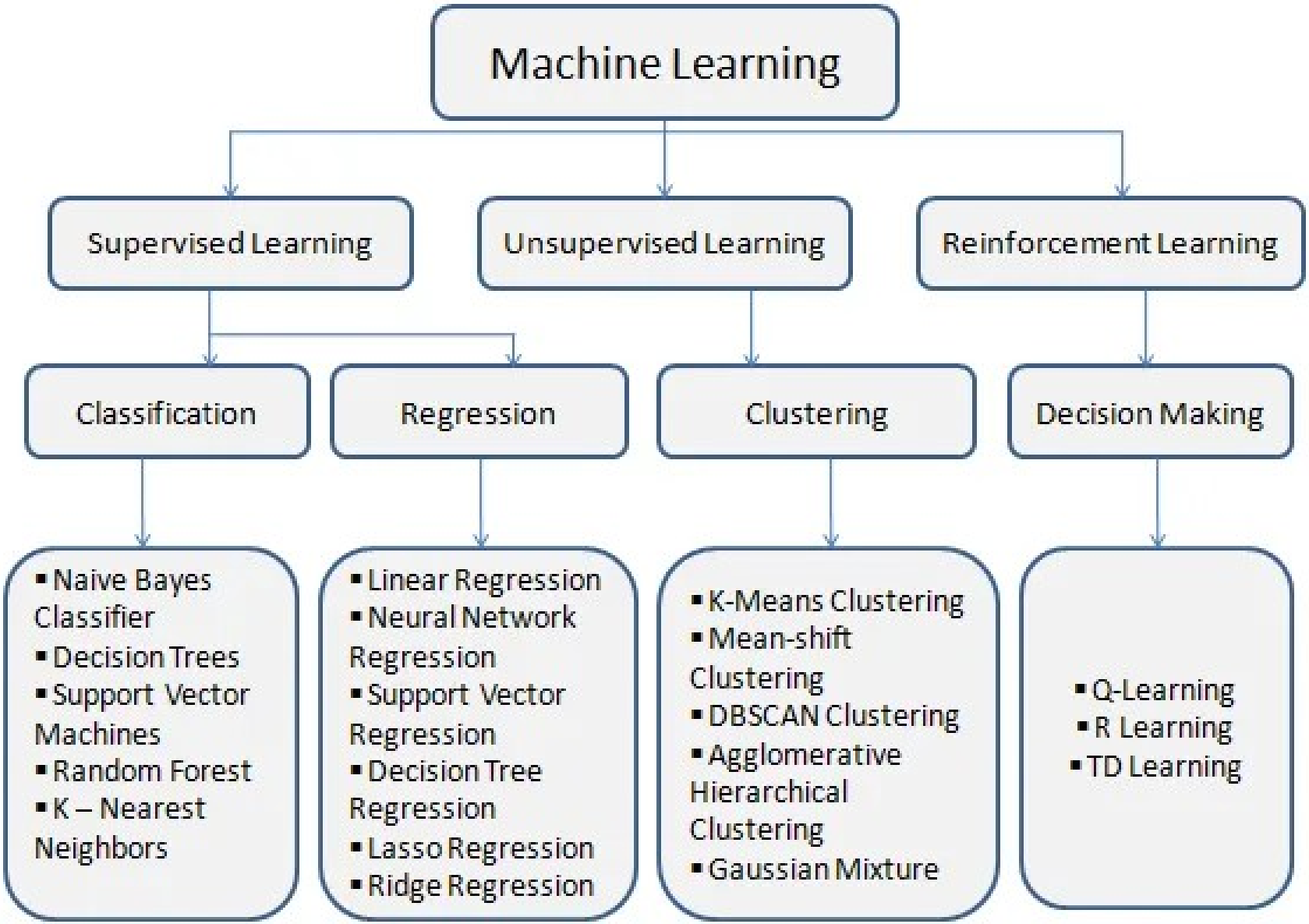
Types of Learning (1)

- Supervised Learning: machine learning approach that uses labeled datasets, that train or "supervise" algorithms into classifying data or predicting outcomes accurately. Classification, Regression
- Unsupervised Learning: uses machine learning algorithms to analyze and cluster unlabeled data sets,discovering hidden patterns in data without the need for human intervention (hence, they are "unsupervised").Clustering, Association, dimensionality reduction
- Reinforcement Learning: intelligent agents take actions from feedback of learning so that the system adjusts to dynamic conditions in order to maximize the notion of cumulative reward.

Types of Learning (2)



Types of Learning (3)



Data Visualization: Purpose

- Classical statistics uses inferences for drawing conclusion about large population from a small samples
- look for patterns, differences, and other features that address the questions we are interested in, check for inconsistencies and identify limitations.
- Simplifies the complex quantitative information
- Analyze and explore data easily
- Identifies areas for improvement
- Identifies relationship between variables and data points
- Explore patterns in the data

Data Types

- unstructured raw data from sensors, images, text, voice, video etc. must be processed to a structured form for analysis.
- structured data are typically represented as a table with rows and columns, called rectangular data
- structured data of two types: Numerical (Discrete and Continuous), Categorical (that takes only specific set of values representing categories)
- Categorical data could be either Binary (True/ False, 0/1) or Ordinal (explicit ordering like gradation

Data Visualization: Types (1)

- Tables: typical represention of data in rows and columns. But can overwhelm users to look for trends in the data
- Pie charts and stacked bar charts: shows data as a part of whole, used to compare the size of each component with other
- line chart and area chart: shows the change in one or more quantities by plotting a series of data point over time, used for predictive analytics
  - Line charts: changes are shown by lines
  - Area charts: changes are shown by connecting data points with line segment and stacking varibles on top of another with different coloring schemes.

Data Visualization: Types (2)

- Histograms: barplot , where each bar represents the frequency or proportion(count/ total count) of cases for a range of values; create uniform sized bins on x axis to count number of items fall in that bin, y axis represent the count. Give the distribution of data; can help understanding central tendency, spread, modality, shape and outliers.
- Scatterplot: used to visualize relationship between two variables.
- Heat Maps: a two-dimensional representation of data in which values are represented by colors
- Tree Maps: used to visualize hierarchical data as a set of nested shapes, typically rectangles; used for comparing the proportions between categories via their area size.
- there are many more... explore.