LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

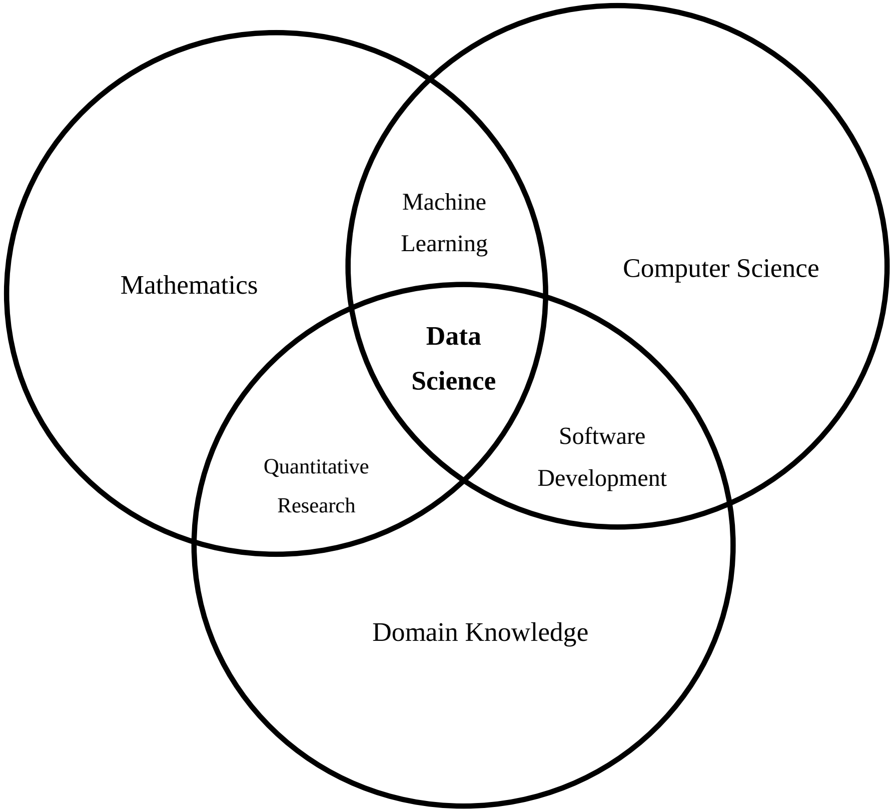
Carlos J. Costa

PROGRAMMING FOR DATA SCIENCE

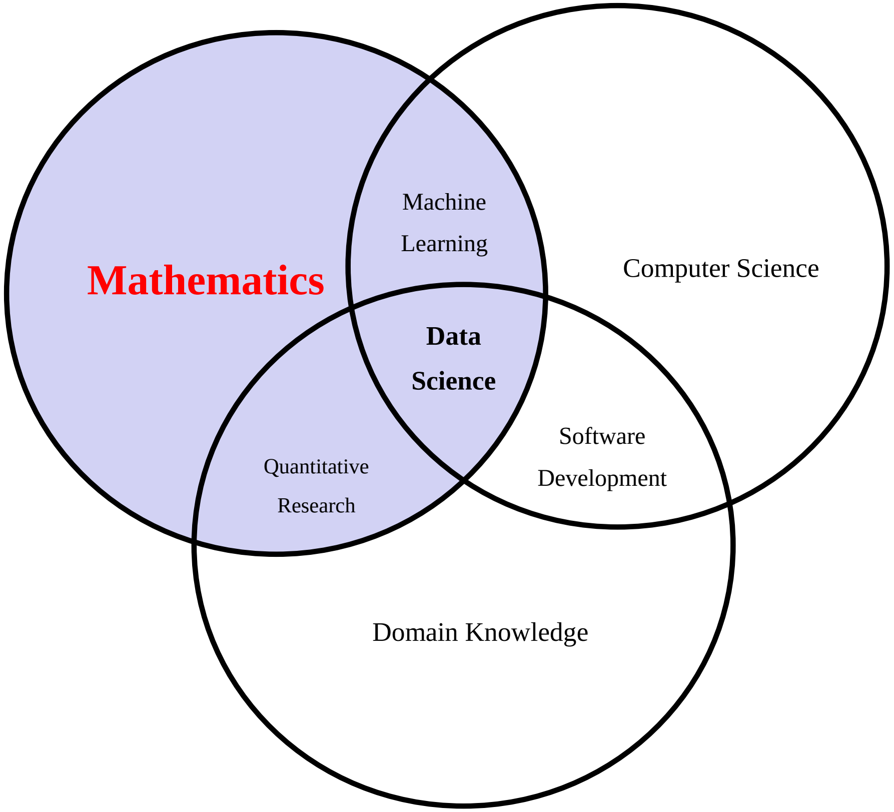
Data Science

- data science is a set of fundamental principles that support and guide the extraction of information and knowledge from data.

Data Science



Mathematics & Statistics

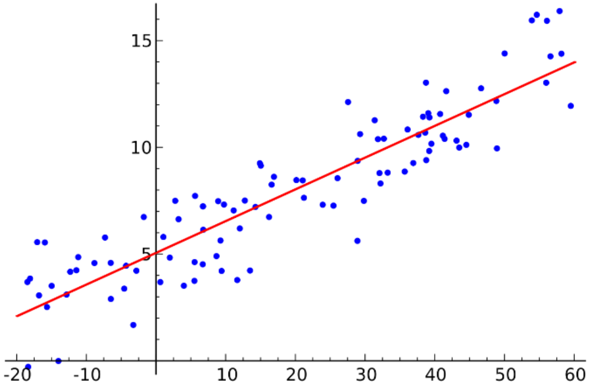


Mathematics & Statistics

- Regressions
- Logistics Regression
- Random forest
- Cluster Analysis
- Social Network Analysis

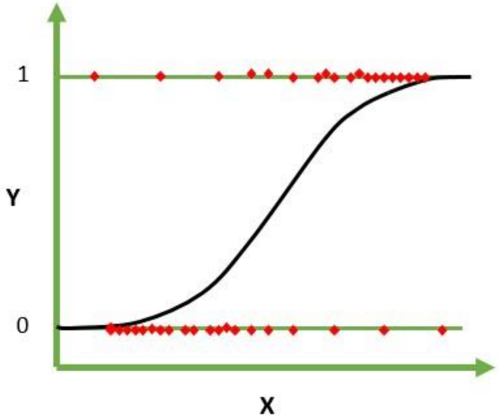
Mathematics & Statistics

- Regression analysis is a is a set of statistical processes for estimating the relationships among variables.



Mathematics & Statistics

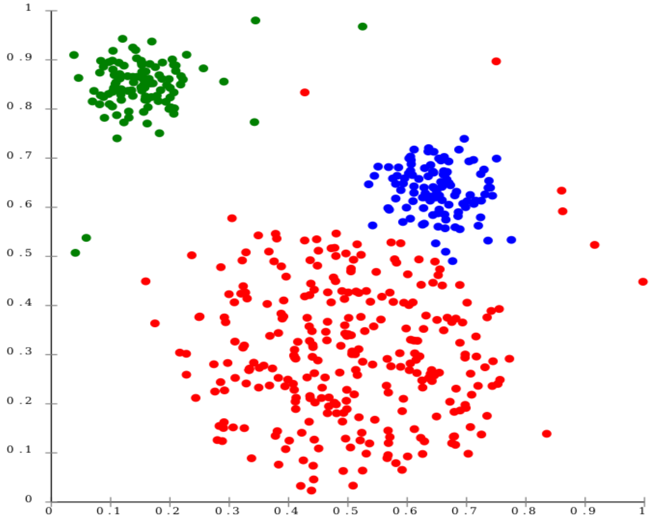
- Logistics Regression
  - A regression that having binary dependent variable
  - in its basic form, uses a logistic function to model a binary dependent variable



Mathematics & Statistics

- Random Forest
  - are an ensemble learning method for classification, regression and other tasks
  - operates by constructing a multitude of decision trees at training time
  - outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Mathematics & Statistics

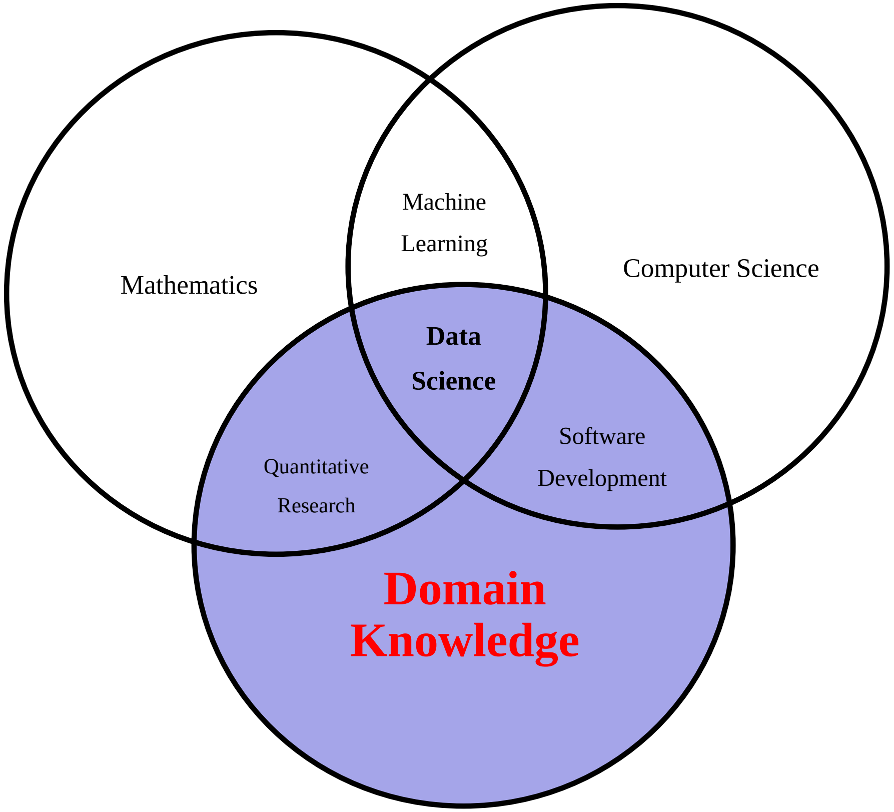


- Cluster Analysis
  - Cluster analysis is a multivariate method
  - aims to classify a sample of subjects (or objects) into several different groups such that similar subjects are placed in the same group
  - based on a set of measured variables

Mathematics & Statistics

- Social Network Analysis
  - is not a formal theory in sociology but rather a strategy for investigating social structures.
  - is the process of investigating social structures using networks and graph theory.
  - uses edges and nodes to describe social relations.
  - there is an assumption of non-randomness or locality.

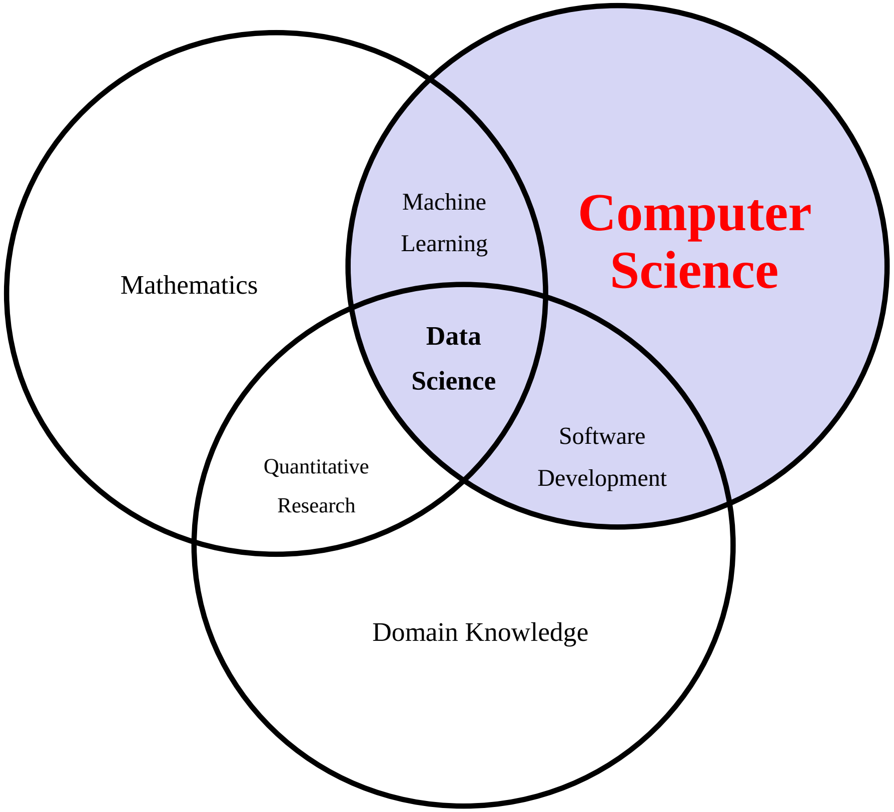
Domain Knowledge



Domain Knowledge

- Functional areas:
  - Marketing
  - Business Strategy
  - Finance
  - Operation Management
  - …
- Industry
  - Manufacturing
  - Utilities
  - Banking
  - …

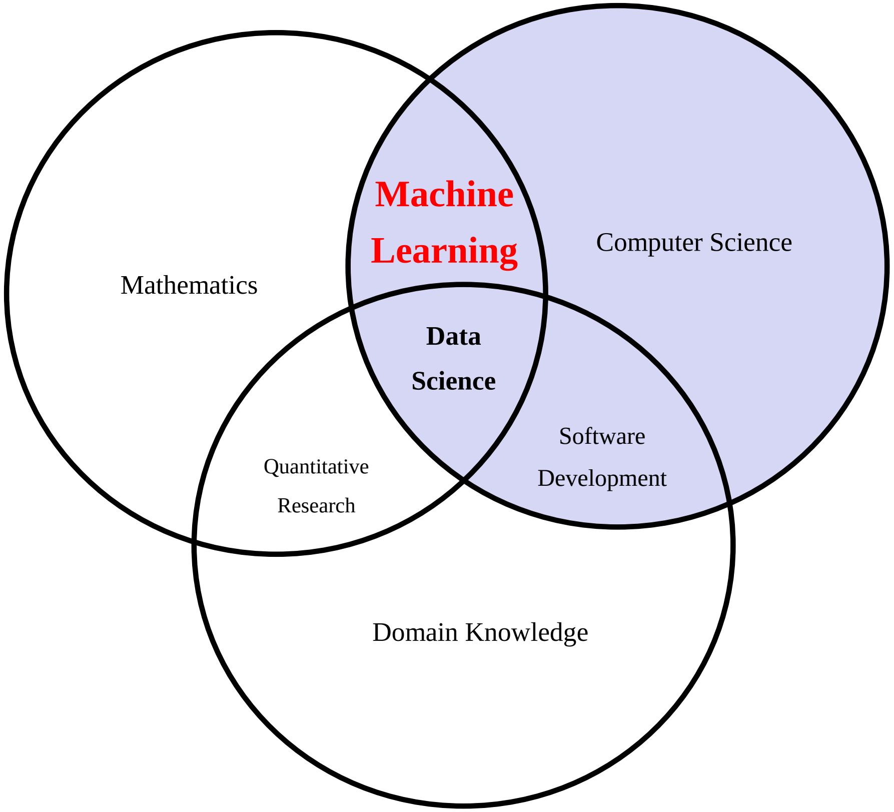
Computer Science Concepts



Computer Science Concepts

- Main Programming Concepts
  - Variables
  - Control Structure
  - Collections
  - Functions
  - Objects

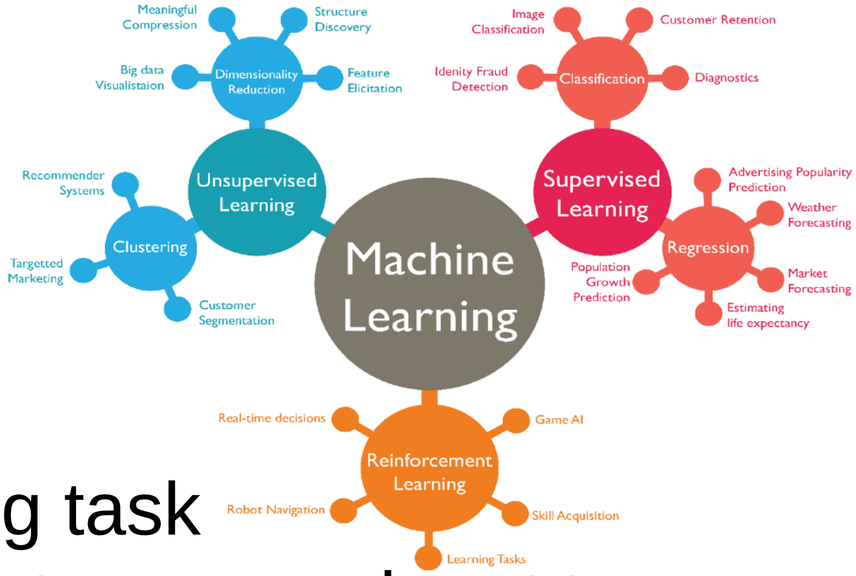
Machine Learning



Machine Learning

- Machine learning
  - It is as a subset of artificial intelligence.
  - It is the scientific study of algorithms that computer systems use to perform a specific task without using explicit instructions
  - study and construction of algorithms that can learn from and make predictions on data

Machine Learning



- Supervised learning
  - It is the machine learning task of learning a function that maps an input to an output based on example input-output pairs
- Unsupervised learning
  - The goal of unsupervised learning is to extract an efficient internal representation of the statistical structure implicit in the inputs. (Hinton & Sejnowski,1999)

Machine Learning

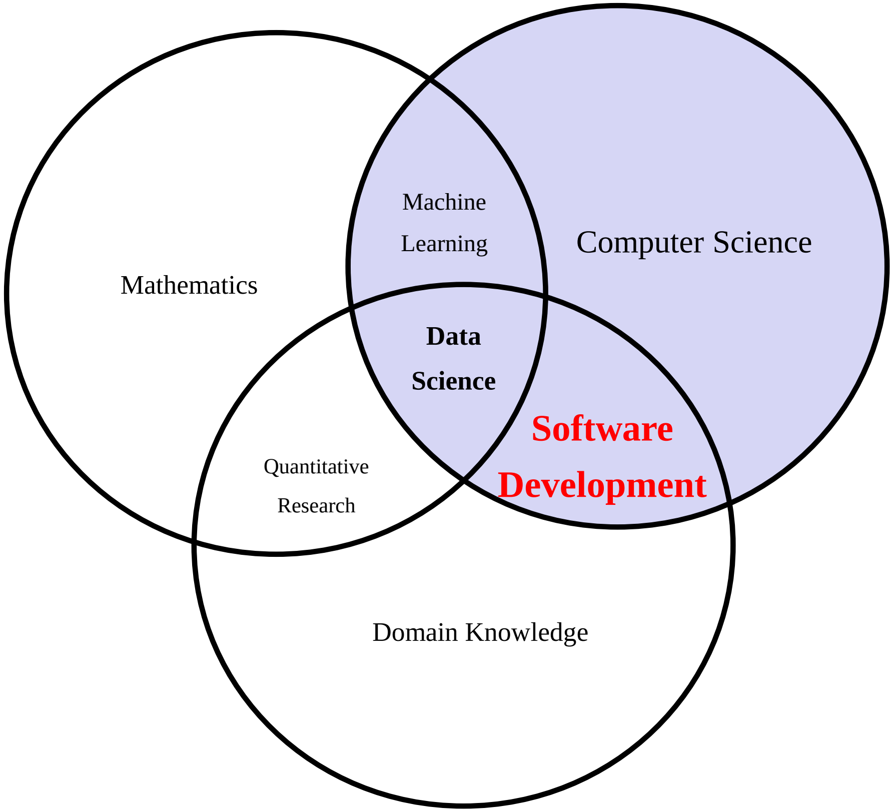
- Train- Validate-Test
- Step 1: Making the model examine data.
- Step 2: Making the model learn from its mistakes.
- Step 3: Making a conclusion on how well the model performs

Machine Learning

- Data Processing and Machine Learning
  - Libraries: Numpy, Pandas, statsmodels, sklearn, networkx
  - Tools: IDE – Jupiter

Integrated Development Environment

Software Development



Software Development

- Web Development
  - Framework: Flask
  - Tool: pythonanywhere

Flask

web development, one drop at a time

References

- Hinton, J.; Sejnowski, T.(1999). *Unsupervised Learning: Foundations of Neural Computation*. MIT Press
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (1 edition). Cambridge, MA: The MIT Press.
- Otte, E.; Rousseau, R. (2002). "Social network analysis: a powerful strategy, also for the information sciences". *Journal of Information Science*. 28 (6): 441–453. doi:10.1177/016555150202800601.
- Stuart J. R., Norvig, P. (2010) *Artificial Intelligence: A Modern Approach*, Third Edition, Prentice Hall ISBN 9780136042594.