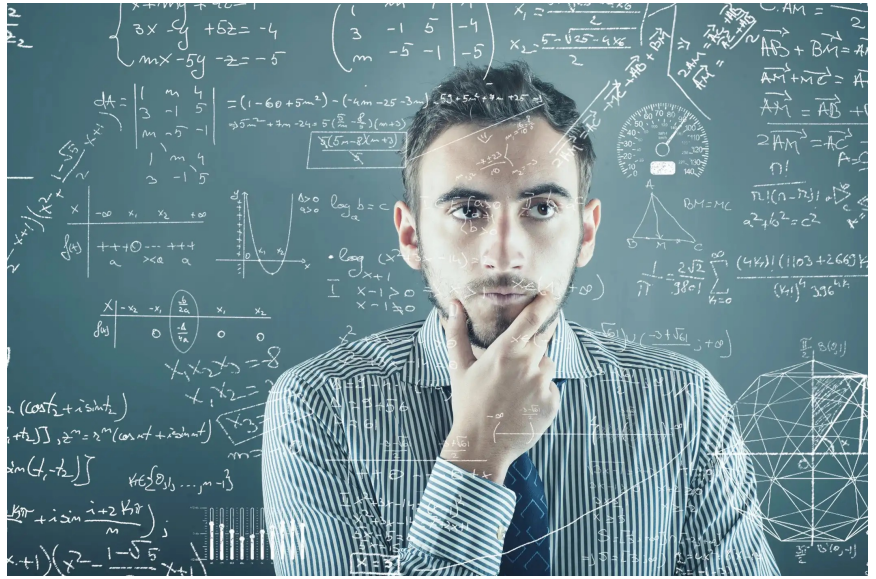
Introduction to:
CS 301 Introduction to Data Science

Michael Renda

Introduction

- Need for Data Science
- What is Data Science?
- Some components of Data Science
- What does a Data Scientist do?
- Data Science lifecycle
- Definitions of some common terms
- Tools of the trade
  - Jupyter notebook
  - Python
  - numpy and pandas
  - matplotlib and seaborn

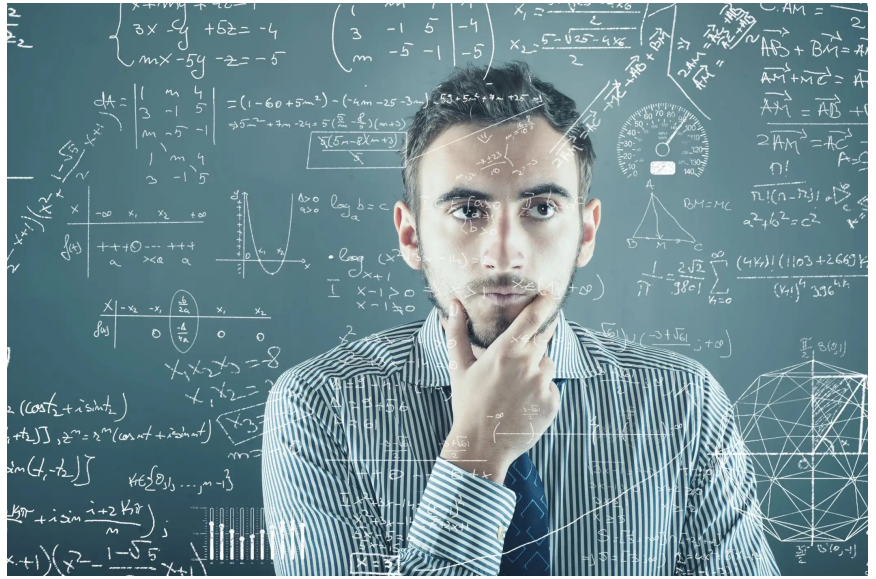


Changing random stuff until your program works is “hacky” and “bad coding practice.”

But if you do it fast enough it is “Machine Learning” and pays 4x your current salary.

Need for Data Science?

- Data science is the most in-demand technology in today's market.



- The term 'data scientist' was coined as recently as 2008 when companies realized the need for data professionals who are skilled in organizing and analyzing massive amounts of data.

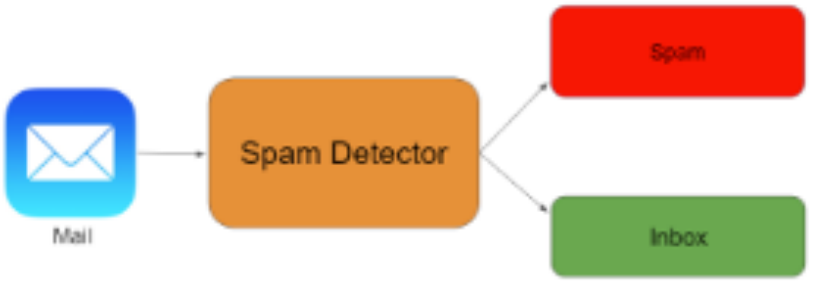
- It's applications range from self-driving cars to predicting deadly diseases.

- Effective data scientists are able to identify relevant questions, collect data from a multitude of different data sources, organize the information, translate results into solutions, and communicate their findings in a way that positively affects business decisions. These skills are required in almost all industries, causing skilled data scientists to be increasingly valuable to companies.

Need for data science

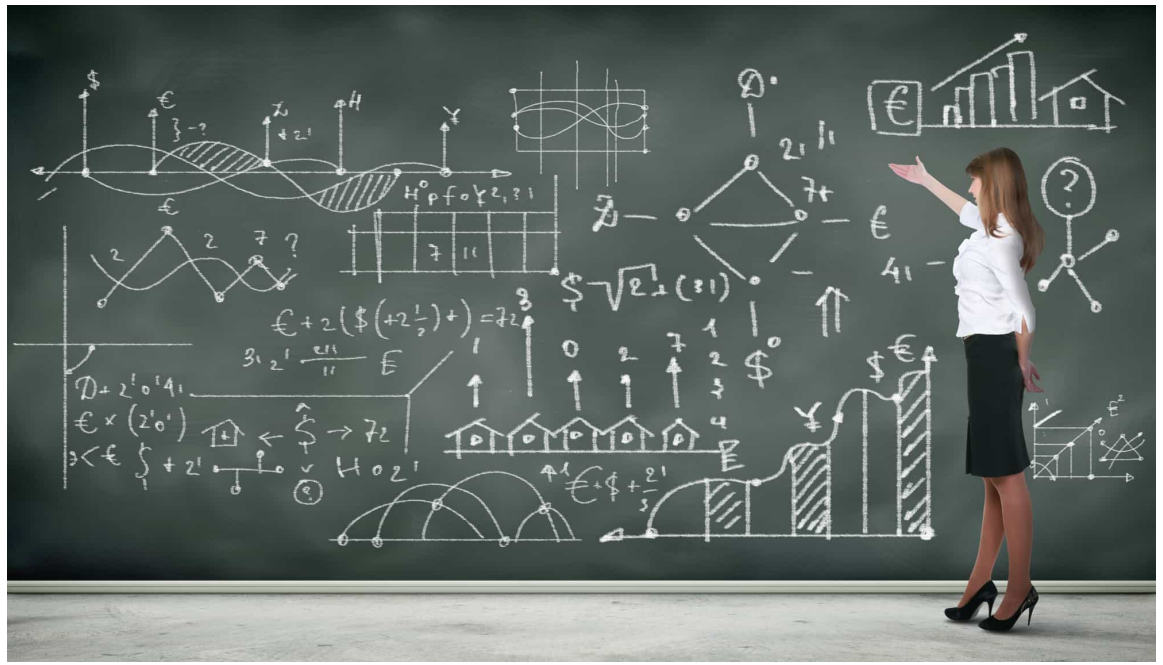
- Increase in data generation: due to the exponential increase in production of data, we need methods that can be used to structure, analyze, and draw useful insights from data.
- Data driven companies such as Netflix and Amazon build data science models by using tons of data in order to identify profitable opportunities or avoid unwanted risk.
- Through these learning models we strive to:
  - predict risk
  - identify profits
  - identify opportunities which will help you grow your business.

Data Science in Our Lives



What is Data Science?

Data science is a multidisciplinary field that combines techniques from statistics, mathematics, computer science, and domain expertise to extract knowledge and insights from structured and unstructured data.

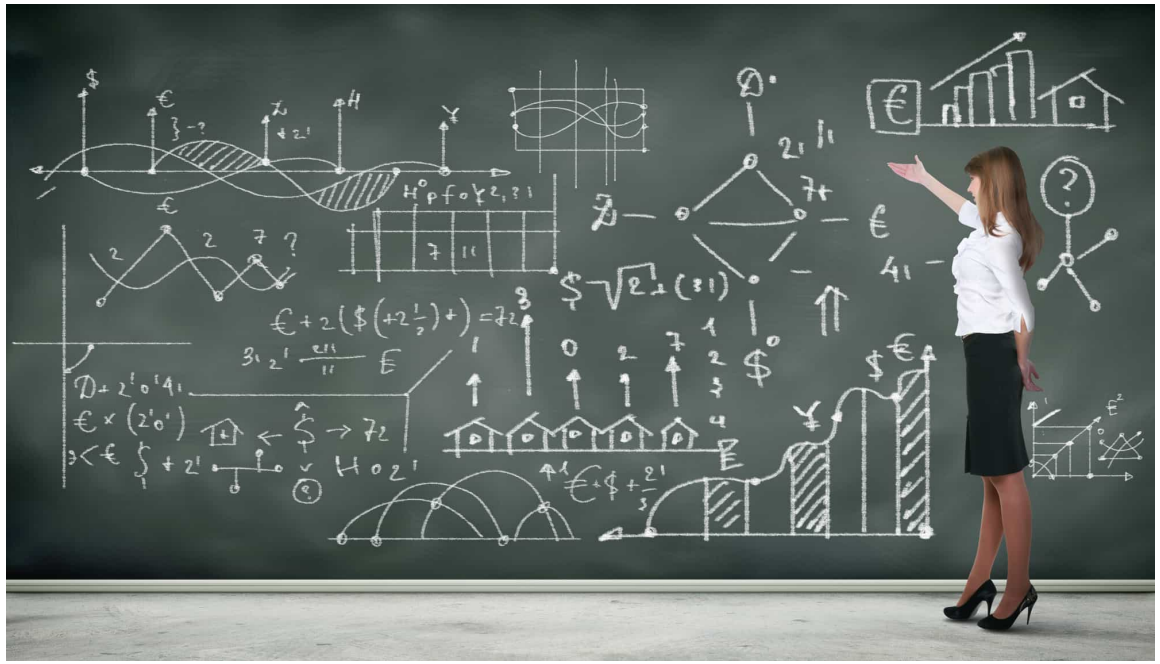


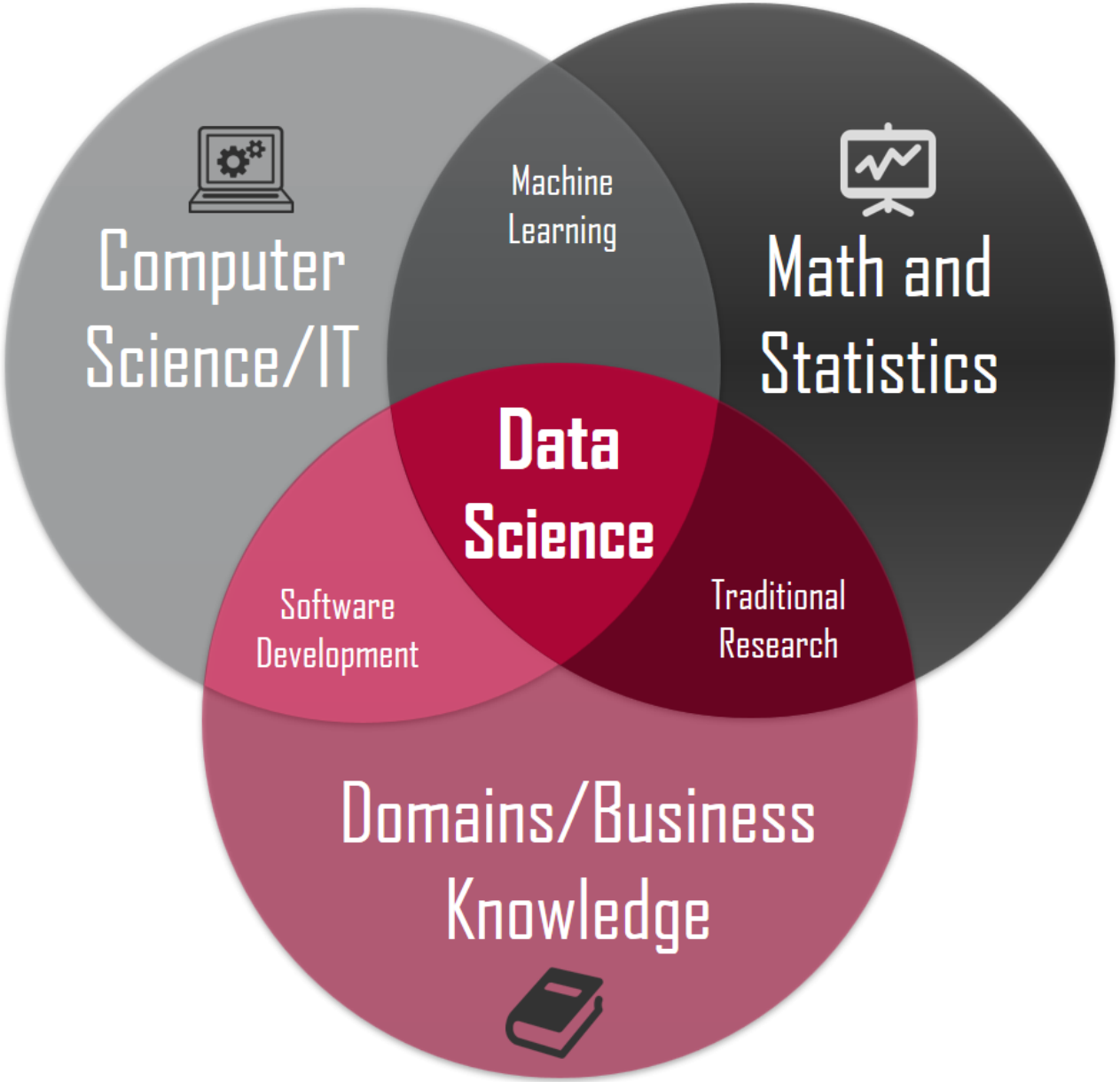
ChatGPT

What is Data Science?

Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured, and unstructured data.

Wikipedia





Source: https://thedatascientist.com/data-science-considered-own-discipline/

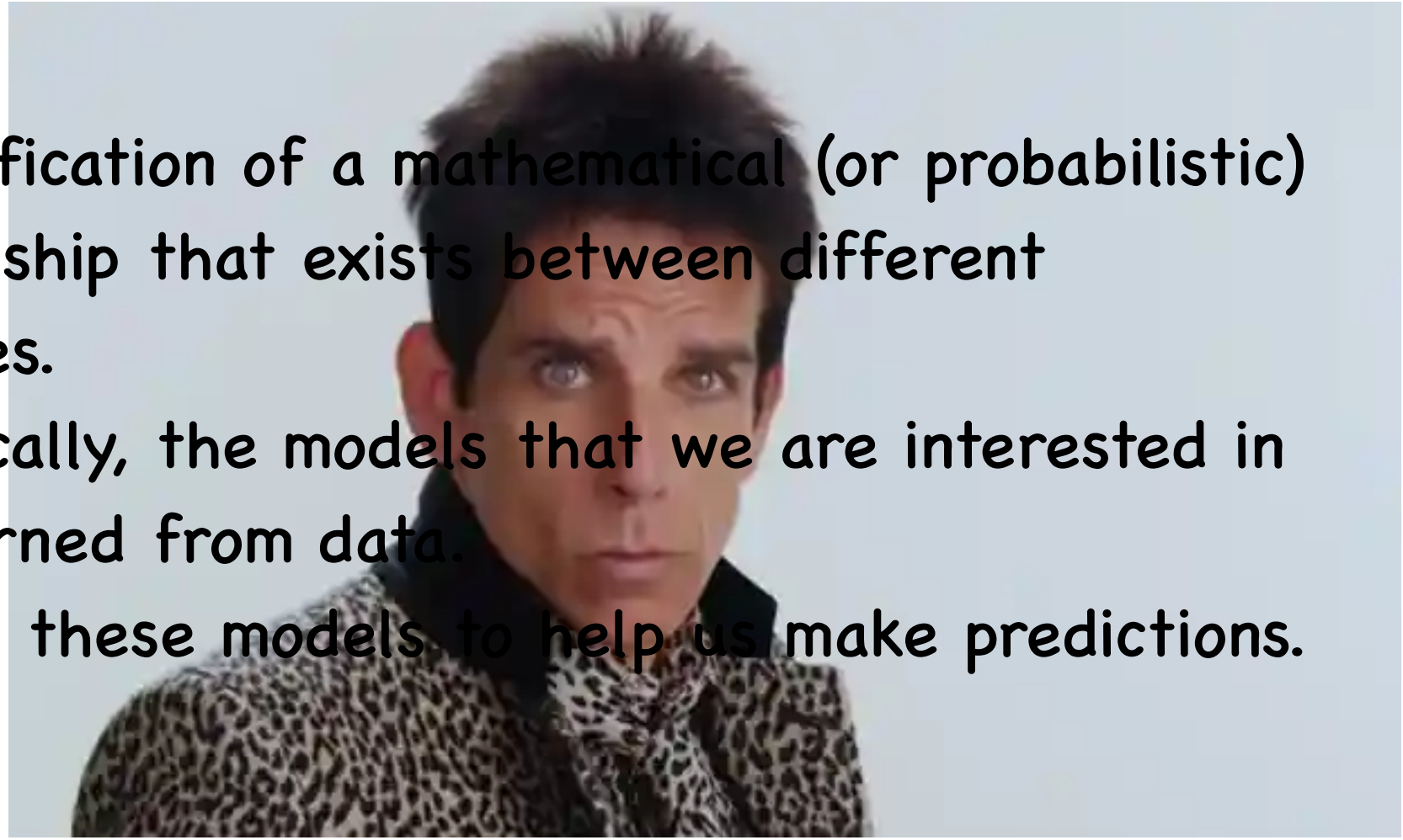
Data Mining

- Seeks to explore and extract valuable patterns and trends from vast datasets.
- Predominantly relies on algorithms and statistical techniques to extract insights.
    - These include big data, cluster analysis, and rule mining and artificial intelligence.

Artificial Intelligence & Machine Learning

- **Machine learning (ML)** is the subset of artificial intelligence (AI) that focuses on building systems that learn—or improve performance—based on the data they consume.
- **Artificial intelligence** is a broad term that refers to systems or machines that mimic human intelligence.
- Creating and using **models** that are learned from data.

What is a Model?

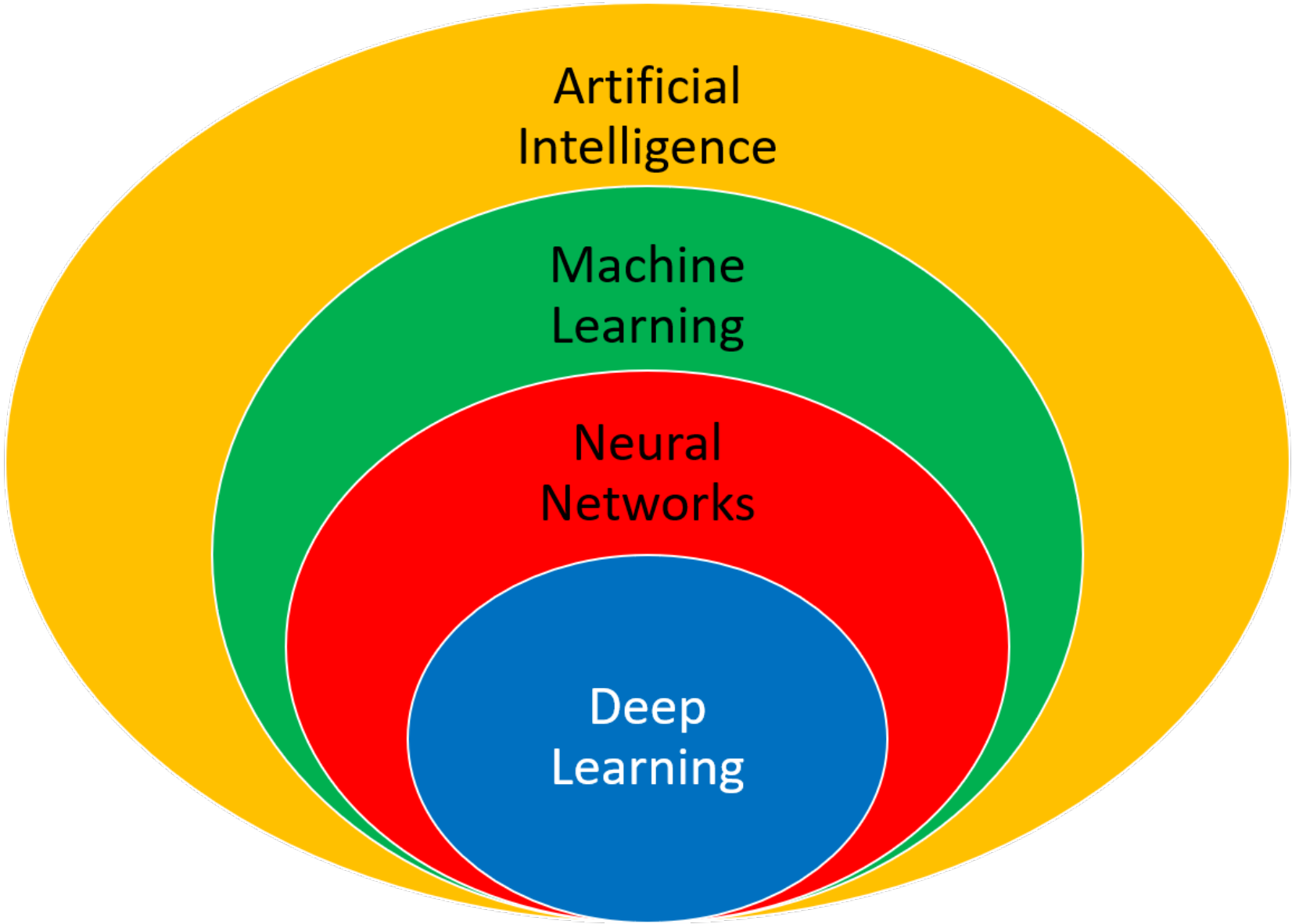


- A specification of a mathematical (or probabilistic) relationship that exists between different variables.
- Specifically, the models that we are interested in are learned from data.
- We use these models to help us make predictions.

Joel Grus, *Data Science from Scratch*

Thinking About Models

- **What we are trying to predict:**
  - **Regression models:** predict real number values.
  - **Classification models:** predict class membership.
- **How a model learns:**
  - **Supervised learning:** the dataset used for training is labeled.
  - **Unsupervised learning:** the dataset used for training is unlabeled.



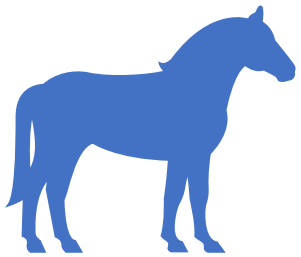
Source: https://apmonitor.com/do/index.php/Main/DeepLearning

Data Sources

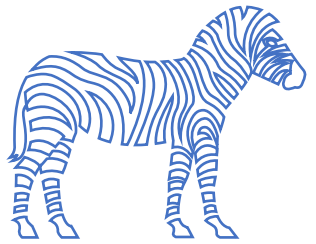
- Transactional databases
- Data warehouses
- Spatial databases
- Multimedia databases
- External databases
- Time-series databases
- WWW
- Streaming data sources

Types of Data

- Structured
- Unstructured
- Semi-structured
- Labeled vs unlabeled



Horse



Zebra

Supervised Learning

Unsupervised Learning

Some Patterns that can be mined

- Characterization/Discrimination
- Frequent patterns
  - Association analysis
- Cluster analysis
- Outlier Analysis
- Classification & Regression for Predictive Analysis

Data Science Activities from 30,000 feet

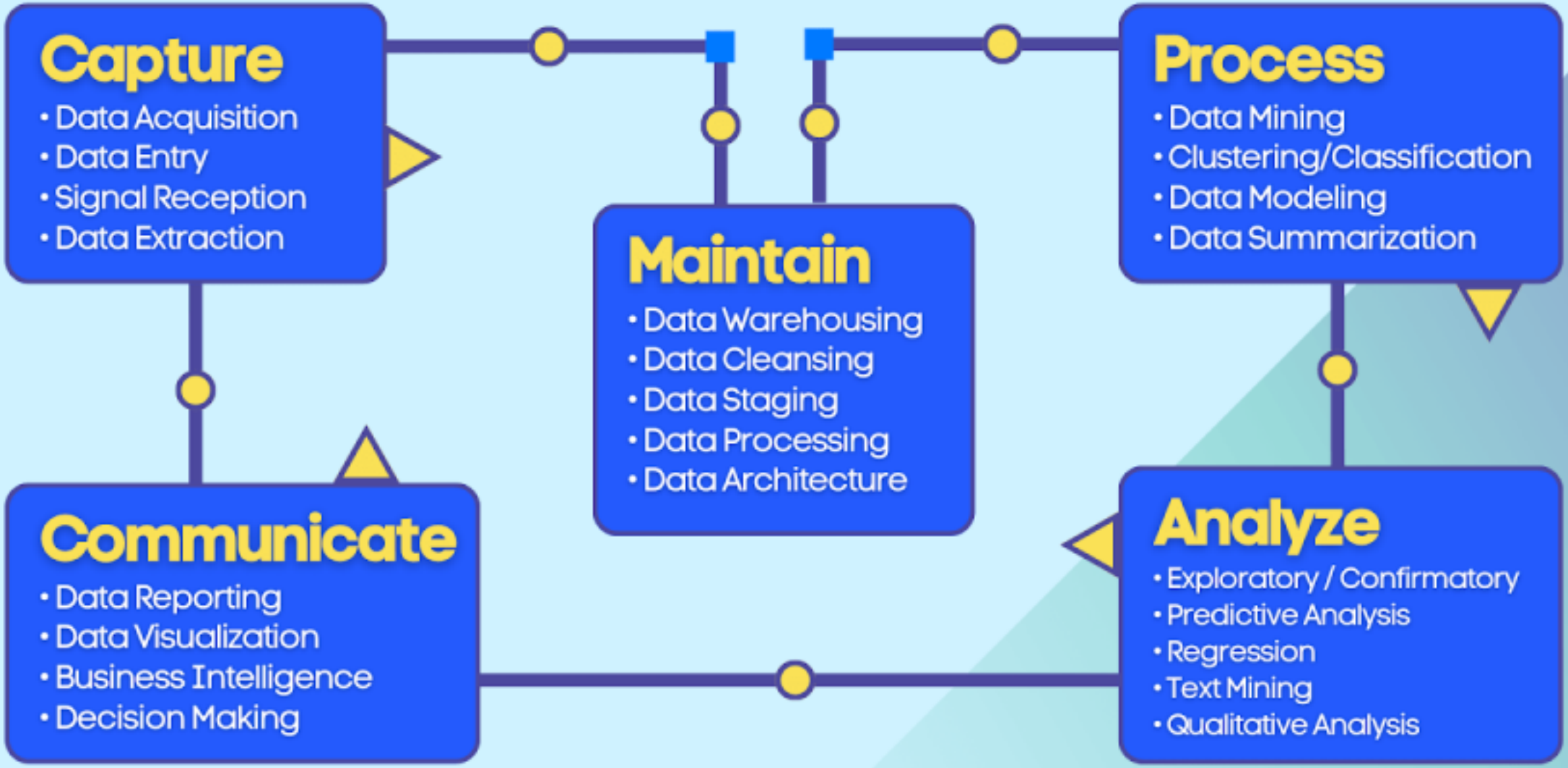
- Turning business problems into data problems
- Collecting data and understanding data
- Cleaning data and formatting data
- Building a model (machine learning / neural networks, deep learning, LLMs)

Joel Grus, *Data Science from Scratch*

Machine learning process:

- Machine learning process always begins with defining the objective or defining the problem that you're trying to solve.
- Next step is data gathering (or data collection). The data that you need to solve this problem is collected at this stage.
- Data preparation (or data processing).
- Data exploration and analysis.
- Build a machine learning model
- Model tuning and evaluation
- Finally, you have predictions (or your output).

Data Science Life Cycle



>hackr.io

Source: https://hackr.io/blog/what-is-data-science

Benefits from Data Science:

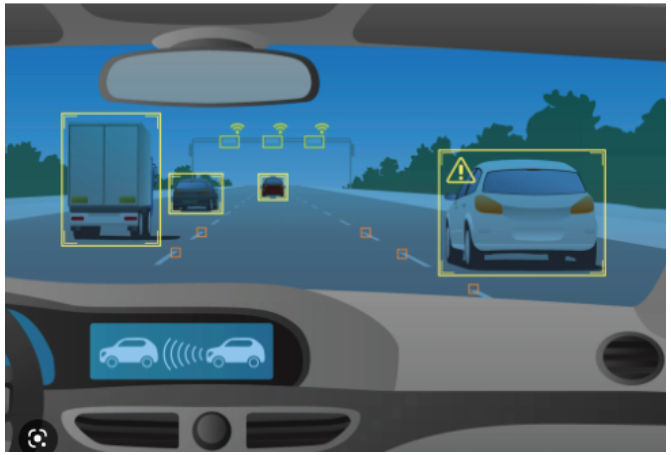
- **Improve decision making:** by making use of various algorithms, machine learning can be used to make better business decisions.
- **Uncover patterns & trends in data:** Finding hidden patterns and extracting key insights from data is the most essential part of machine learning.
- **Solves complex problems:** from detecting diseases to building self-driving cars, and building face detection systems; data science can be used to solve the most complex problems.

Commonly used terms in Data Science:

- **Algorithm:** is a set of rules or statistical techniques that are used to learn patterns from data. An algorithm is the logic behind a machine learning model. An example of a machine learning algorithm is linear regression.
- **Predictor variable (or independent variable)**: is a feature of the data that can be used to predict the output.
- **Response variable (or target variable or dependent variable):** it is the feature or the output variable that will be predicted based on the predictor variables.
- **Training data:** the data that is fed to a machine learning model is always split into two parts: **training data** and **testing data**. Training data is basically used to building the machine learning model. So usually training data is much larger than the testing data because training results generally improve with more data. **Testing data** is used to evaluate the accuracy and efficiency of the model.

Which machine learning type should be used?

YouTube

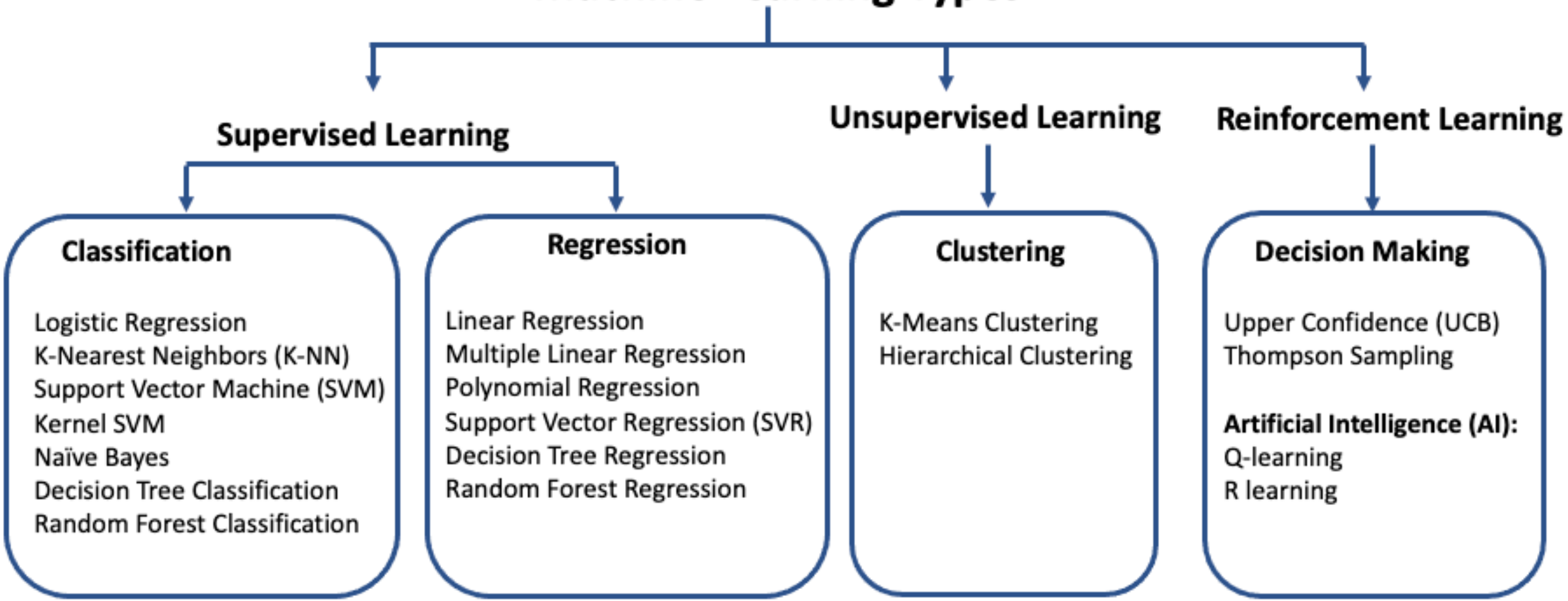


Machine Learning Types:

- **Supervised Learning:** Develop predictive model based on both input and output data

- **Unsupervised Learning:** Group and interpret data based only on input data

Machine Learning Types



Predicting a categorical variable
Input: labeled data set
Output: Discrete values

Predicting a numeric variable
Input: Label data set
Output: continuous values

Identify a pattern or groups of similar objects

solve interacting problems where the data observed up to time t is considered to decide which action to take at time t + 1. It is also used for Artificial Intelligence

Machine Learning Algorithm Classification

https://chart-studio.plotly.com/create/?fid=SolClover:200#/

Machine Learning - The Bottom Line

https://xkcd.com/1838/

Prepare your desktop environment for the class:

- Install Miniconda. Navigate to https://docs.conda.io/en/latest/miniconda.html and select the appropriate installer for your operating system.
- Open the downloaded installation executable and start the installation. Accept all of the defaults.

Latest Miniconda installer links

This list of installers is for the latest release of Python: 3.11.3. For installers for older versions of Python, see Other installer links. For an archive of Miniconda versions, see https://repo.anaconda.com/miniconda/.

Latest - Conda 23.5.2 Python 3.11.3 released July 13, 2023

Platform	Name	SHA256 hash
Windows	Miniconda3 Windows 64-bit	00e8370542836862d4c790aa8966f1d7344a8addd4b766004febcb23f40e2914
	Miniconda3 Windows 32-bit	4fb64e6c9c28b88beab16994bfba4829110ea3145baa60bda5344174ab65d462
macOS	Miniconda3 macOS Intel x86 64-bit bash	1622e7a0fa60a7d3d892c2d8153b54cd6ffe3e6b979d931320ba56bd52581d4b
	Miniconda3 macOS Intel x86 64-bit pkg	2236a243b6cbe6f16ec324ecc9e631102494c031d41791b44612bbb6a7a1a6b4
	Miniconda3 macOS Apple M1 64-bit bash	c8f436dbde130f171d39dd7b4fca669c223f130ba7789b83959adc1611a35644
	Miniconda3 macOS Apple M1 64-bit pkg	837371f3b6e8ae2b65bdfc8370e6be812b564ff9f40bcd4eb0b22f84bf9b4fe5
Linux	Miniconda3 Linux 64-bit	634d76df5e489c44ade4085552b97bebc786d49245ed1a830022b0b406de5817
	Miniconda3 Linux-aarch64 64-bit	3962738cfac270ae4ff30da0e382aecf6b3305a12064b196457747b157749a7a
	Miniconda3 Linux-ppc64le 64-bit	92237cb2a443dd15005ec004f2f744b14de02cd5513a00983c2f191eb43d1b29
	Miniconda3 Linux-s390x 64-bit	221a4cd7f0a9275c3263efa07fa37385746de884f4306bb5d1fe5733ca770550

Prepare your desktop environment for the class:

- Open the command prompt (*Anaconda Prompt* on Windows or *terminal* on Mac).
- Enter the command to update *conda* to the latest release:

  *(base) C:\Users\michaelrenda\mlprojects>**conda update conda***

Prepare your desktop environment for the class:

- Enter the command to create a new virtual environment named ‘ds’:

  (base) C:\Users\michaelrenda\mlprojects>***conda create -n ds python=3.11***

- Activate the new virtual environment:

  *(base) C:\Users\michaelrenda\mlprojects>****conda activate ds***

  (ds) C:\Users\michaelrenda\mlprojects>

- Install the initial set of python packages:

  *(ds) C:\Users\michaelrenda\mlprojects>****conda install -y numpy pandas scipy matplotlib seaborn jupyter scikit-learn statsmodels keras***

Prepare your desktop environment for the class:

- Create and navigate to the directory where you will keep your projects. This is the directory where you will place all of your Jupyter notebook files.

  (ds) C:\Users\michaelrenda\dsprojects>

- Create a subdirectory where you will place the data files used in your Jupyter notebooks.

  (ds) C:\Users\michaelrenda\dsprojects\data

- Start *Jupyter notebook*:

  (ds) C:\Users\michaelrenda\dsprojects>***jupyter notebook***