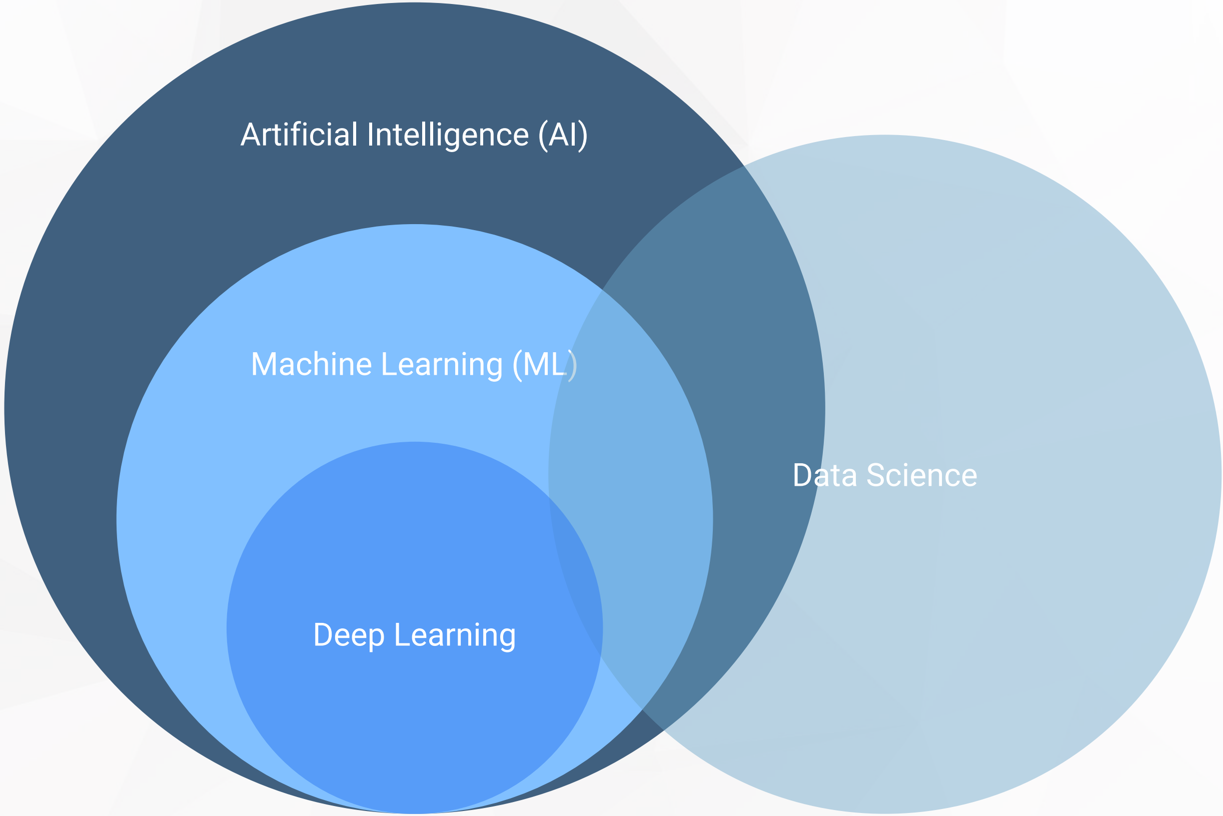
Getting Started with Data Science

analytiks

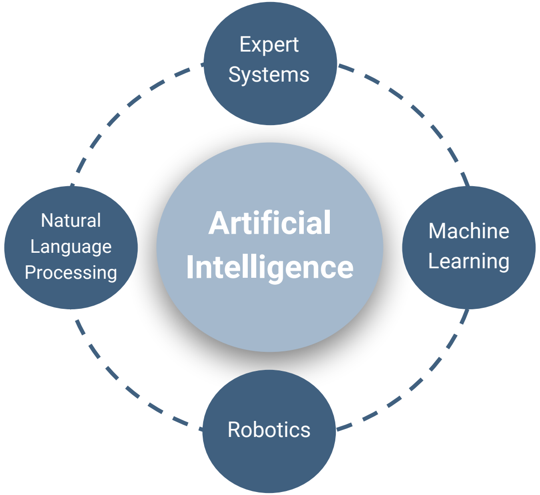
Introduction to Data Science

The Big Picture



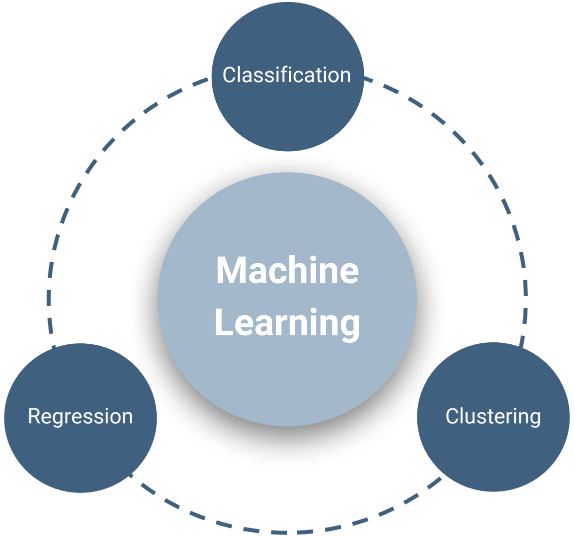
What is Artificial Intelligence?

The broader concept of machines being able to carry out tasks in a way that we would consider intelligent. Machines that are programmed to "think" like a human and mimic the way a person acts.



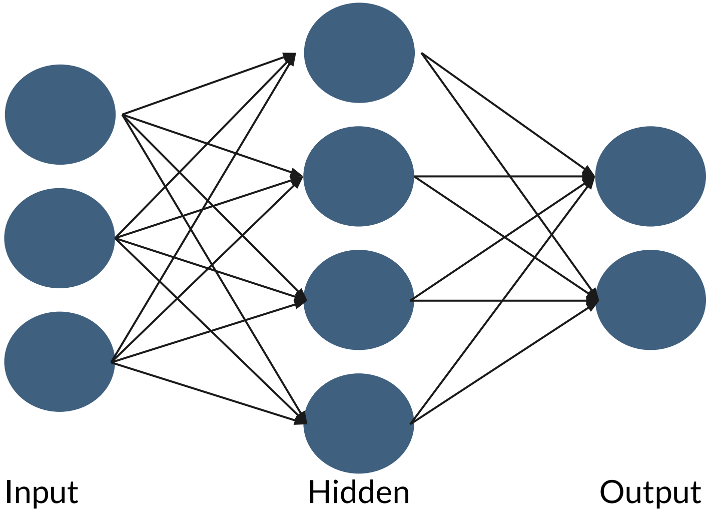
What is Machine Learning?

It is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.



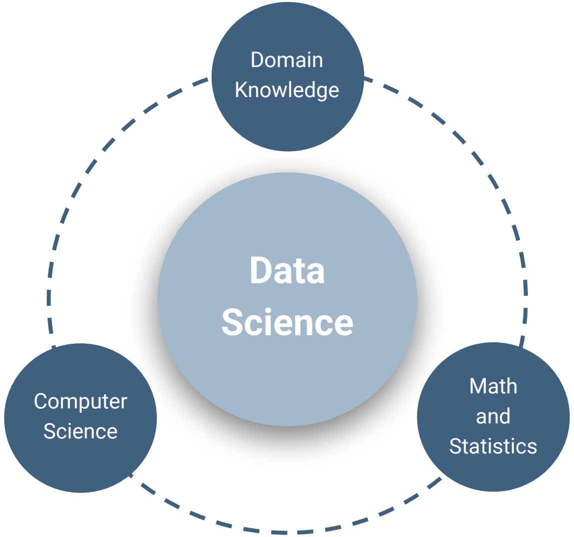
What is Deep Learning?

A subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.



What is Data Science?

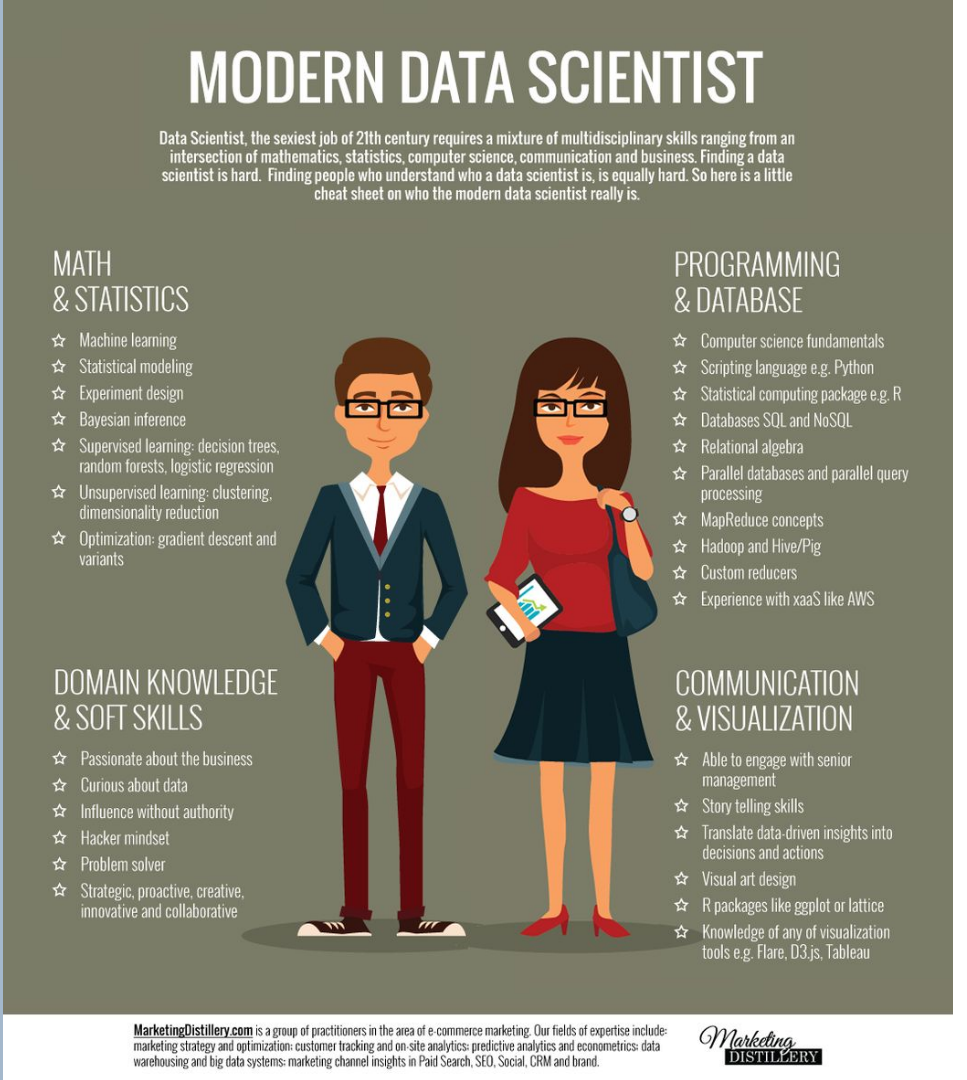
An interdisciplinary field concerned with the study of processes and systems needed to extract knowledge and insights from the vast amount of data that is being collected.



Data Science vs. Business Intelligence

	Business Intelligence	Data Science
Type of Analysis	Descriptive	Descriptive + Prescriptive
Focus	Past and Present	Future
Method	Analytic	Scientific
Data Sources	Structured	Structured + Unstructured
Tools	Statistics, Visualization	Statistics + Machine Learning + AI

What it Takes to Be a Data Scientist



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

What it Takes to Be a Data Scientist

Curiosity	Skepticism	Technical Acumen
Clarity	Humility	Statistical Thinking
Creativity	Grit	Data Intuition

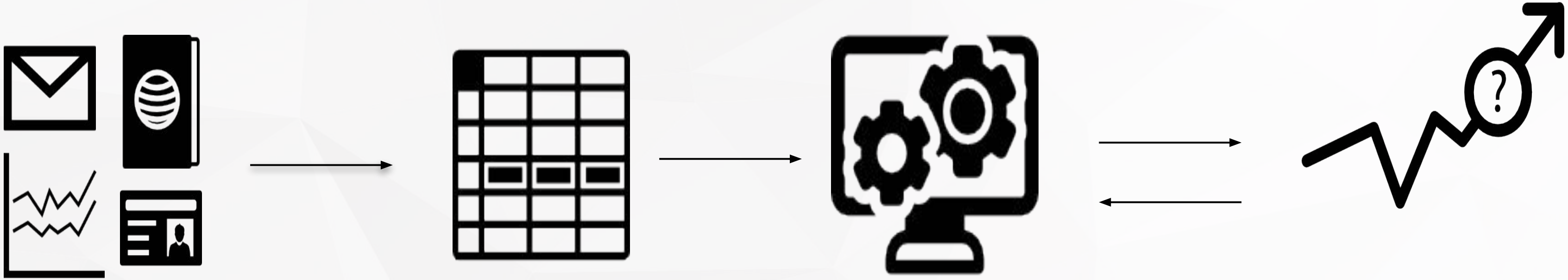
Machine Learning Overview

How does a Machine Learn?

Data Preparation

Model Building
(Training)

Test + Predict



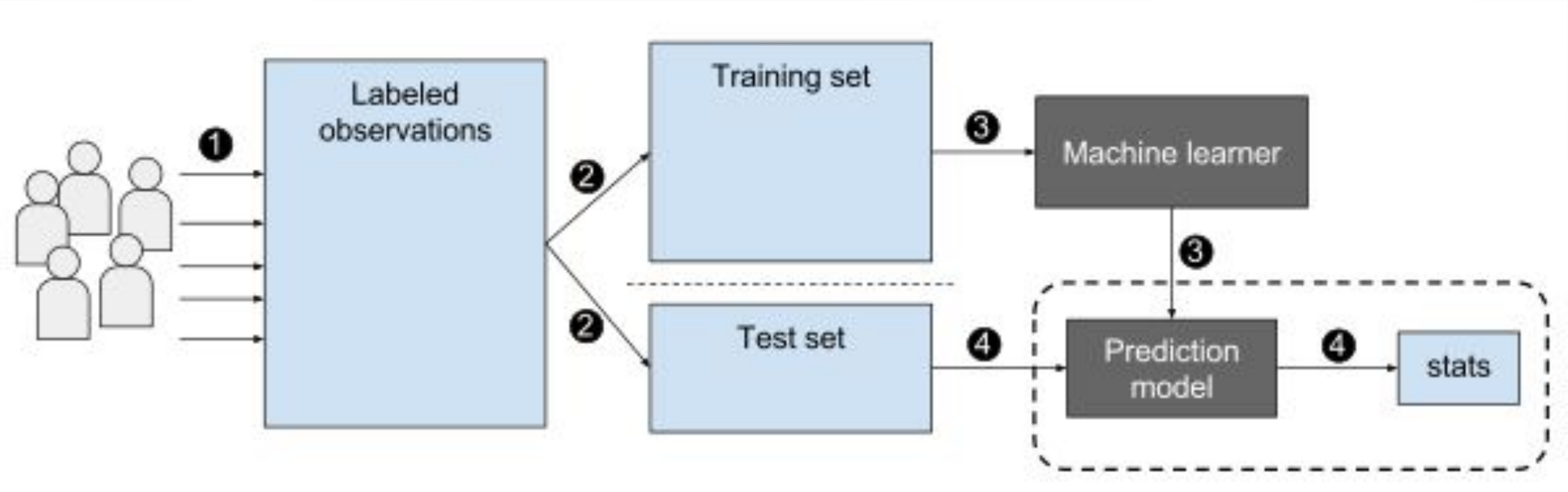
Raw Data
Structured, Unstructured,
From Various Sources

Training Dataset / Curriculum
Clean, Consistent, Complete

Model(s)
Specific to Data Science Problem

Prediction
+ associated Metrics to determine whether if Model is "good"

How does a Machine Learn?



Supervised vs. Unsupervised Learning

	Supervised	Unsupervised
Description	Supervised Learning Algorithms Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time.	Input data is not labeled and does not have a known result.
Process	A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.	A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.
Problems	Classification, Regression	Clustering, Dimensionality Reduction

Supervised: Classification

Which of a set of categories a new observation belongs to, based on a training dataset containing observations whose category membership is known

Two Types: 2-Class / Binary | Multi-Class

Examples:

- Spam Filtering
- Risk Analysis
- Churn Analysis
- Medical Diagnosis
- Fraud Detection
- Employee Retention

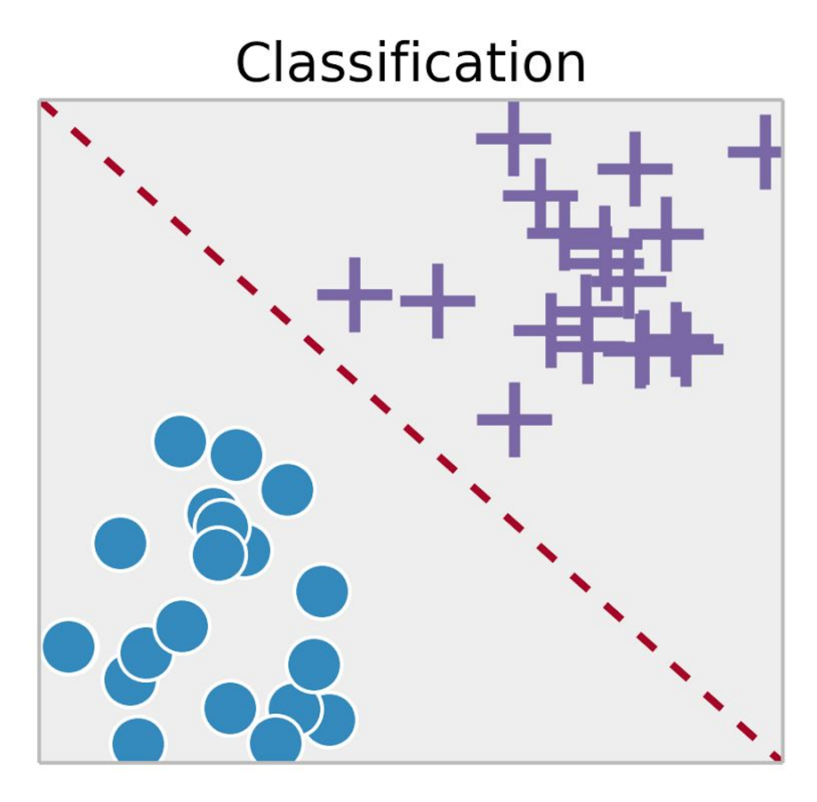
Key Metrics:

- Accuracy Rate
- Error Rate
- Gini Coefficient
- True Positive
True Negative
False Positive
False Negative

%

Class predictions are usually not black or white - tagged to each class prediction is usually a **probability** that the observation belongs to the class.

Classification: Illustration



Supervised: Regression

Predicting a real-numbered value + need to understand relationship between predictor and target variables

Some Types:

- Linear
- Polynomial

Examples:

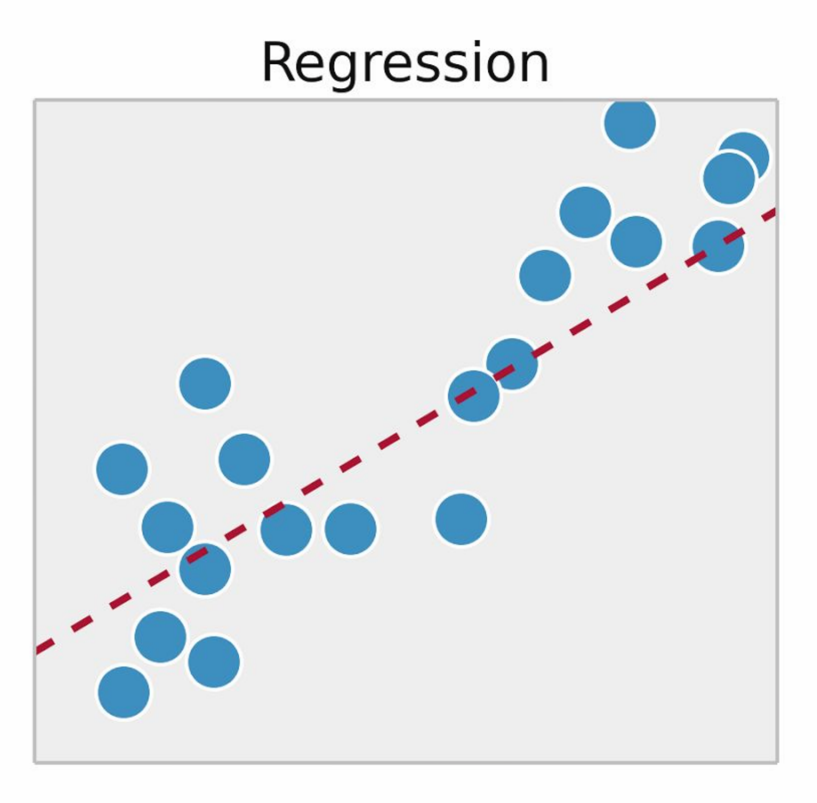
- Customer Lifetime Value
- Energy Consumption
- Insurance Pricing
- Predicting Property Prices
- Flight Time Prediction
- Predicting Profitability

Key Metrics:

- R-Squared
- Mean Absolute Error
- Root Mean Squared Error
- Weighted Mean Absolute Error

Regression is a classic statistical method that is being used for Machine Learning - with ML, more complex relationships between variables can be found.

Regression: Illustration



Unsupervised: Clustering

Segregate groups with similar traits, assign them into clusters.
An Unsupervised learning algorithm - no need for target variable!

Some Types:

K-Means

Hierarchical

Examples:

Customer Profiling

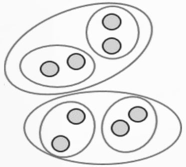
Anomaly Detection

Land Use (Geospatial Images)

Identifying Groups of High Claimants (Insurance)

City Planning: Groups of Houses

Pre-processing for Classification



Key Criterion:

High Inter-Class Similarity

Low Intra-Class Similarity

Look for:

Number of clusters that produce tighter clusters

Number of clusters that are actionable / makes sense to the business

Unsupervised: Clustering

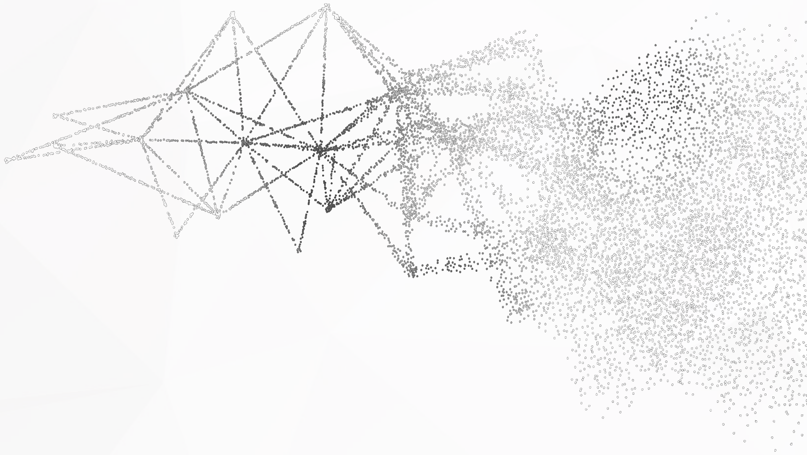
Exploiting the inherent structure in the data in order to summarise or describe data using less information.

Some Types:

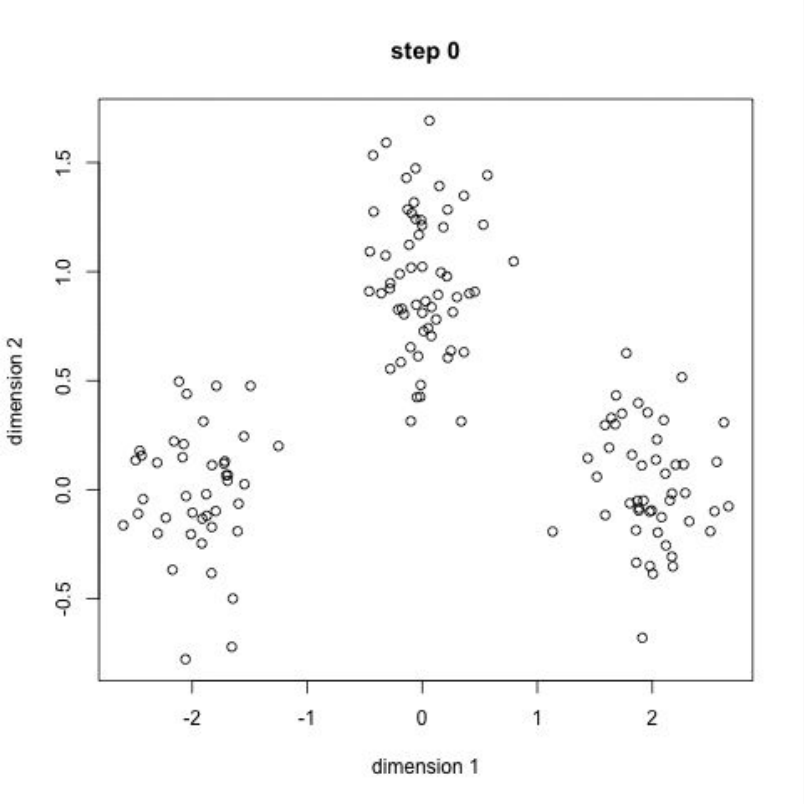
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

Examples:

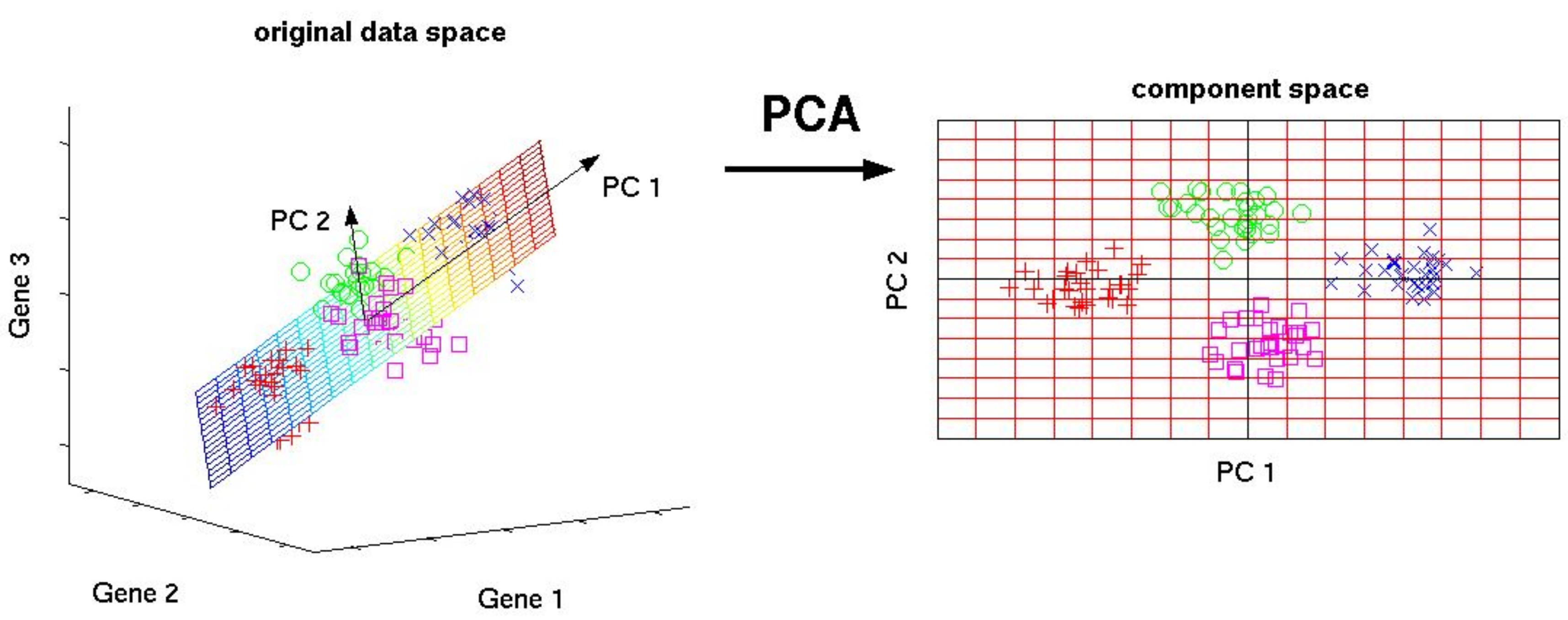
- Visualization and Interpretation
- Feature Extraction
- Feature Reduction
- Feature Selection
- Performance Improvement
- Reduction in Computational Cost



Clustering: Illustration



PCA: Illustration



Advanced Data Science Technology & Techniques

Text Processing

Processing and conversion of text data into usable numerical data for machine learning.

Methods: Bag of Words Word2Vec

Examples:

Sentiment Analysis

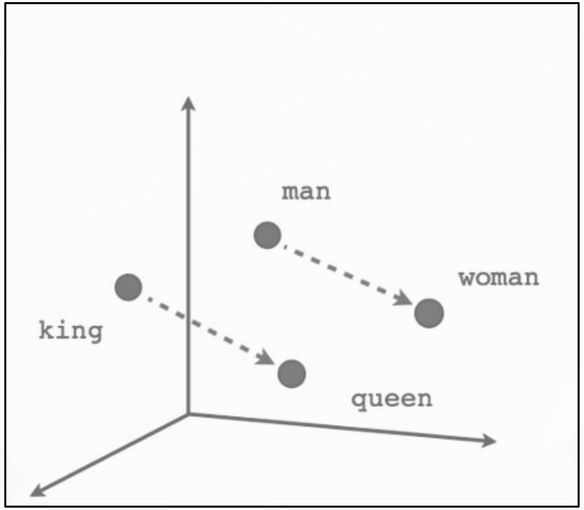
NLP

Document Classification

Chatbots

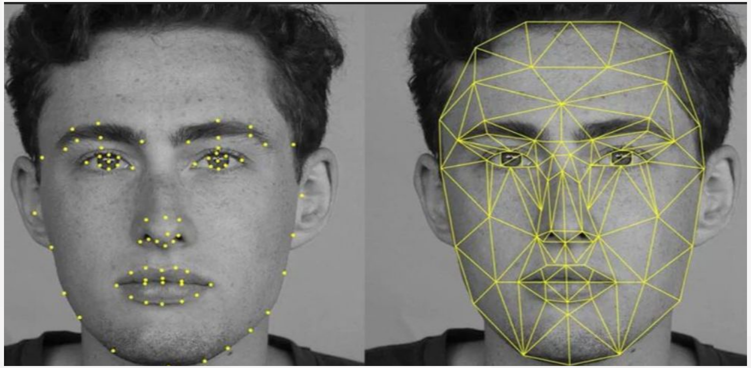
Category Extraction

Fuzzy Matching

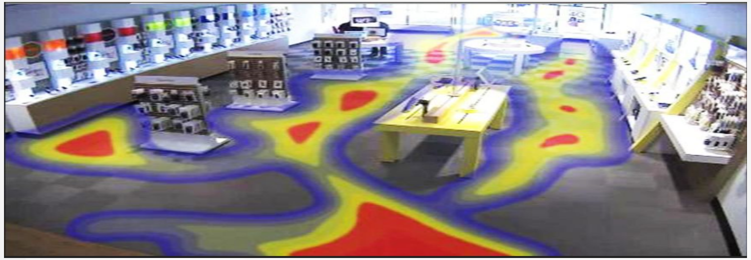


Signal Processing

Processing and conversion of signal data into usable numerical data for machine learning.



Smartphone Detection
Bluetooth
Wifi
SHOP
DEPARTMENT STORE
WI-FI



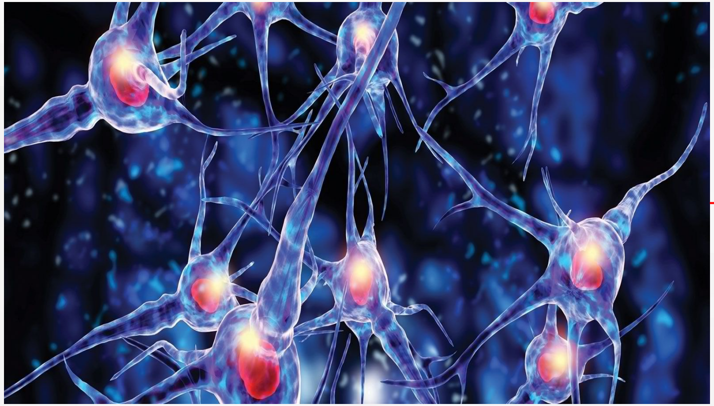
Examples:

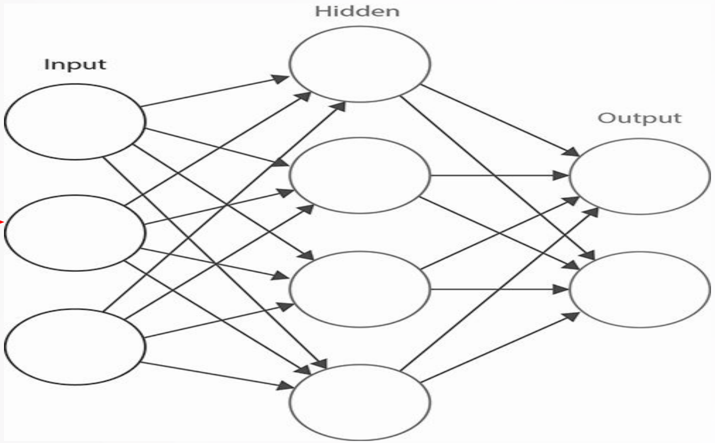
- Image Recognition
- NLP
- Location and Movement Analysis

A.I. / Cognitive Applications

Algorithms to mimic the way humans process information

Methods: Neural Networks Deep Learning





A.I. / Cognitive Applications

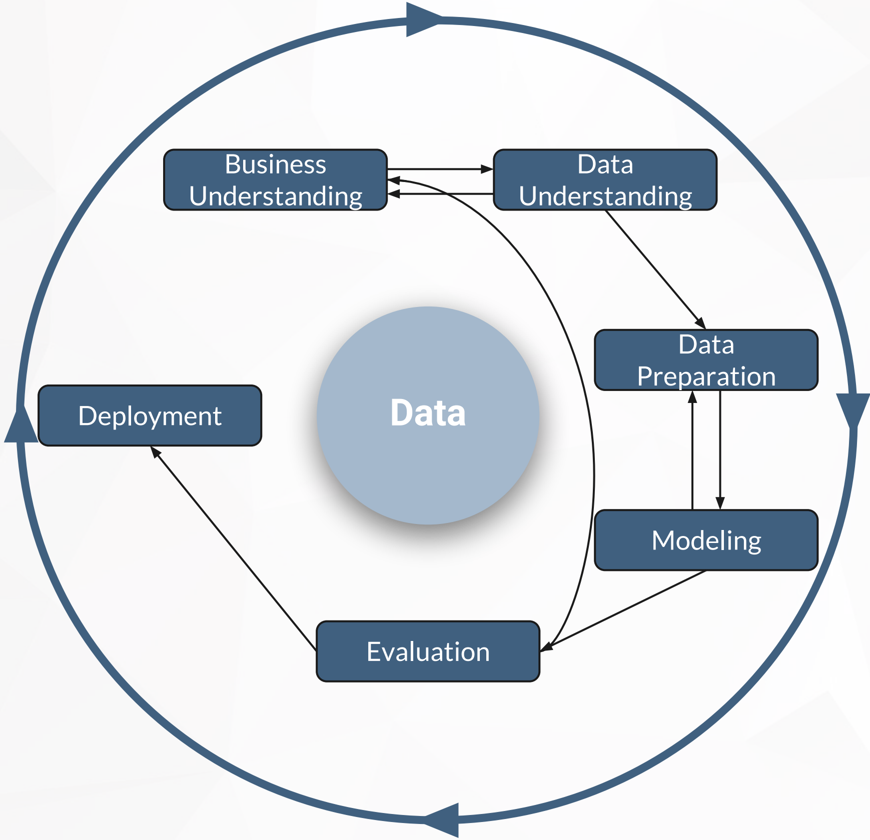
Algorithms to mimic the way humans process information

Methods: Neural Networks Deep Learning

Data Science Methodology

DS Methodology: CRISP-DM

- CRoss-Industry Standard Process for Data Mining
- Leading methodology used for data science projects
- Provides a structured approach to planning a data mining project allowing for reasonable consistency, repeatability, and objectiveness.
- Iterative by design. Often every iteration leads to new discoveries, insights and opportunities.

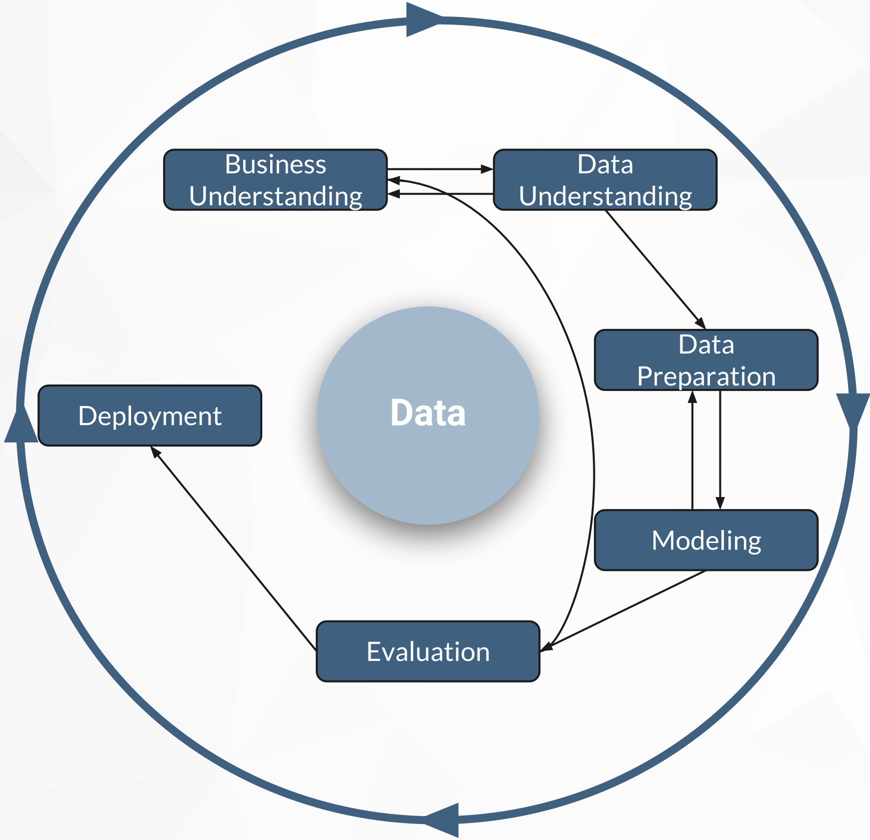


Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Tasks:

- Identify the problem/question that you are trying to solve/answer
- Determine appropriate ML task (classification,clustering,regression)

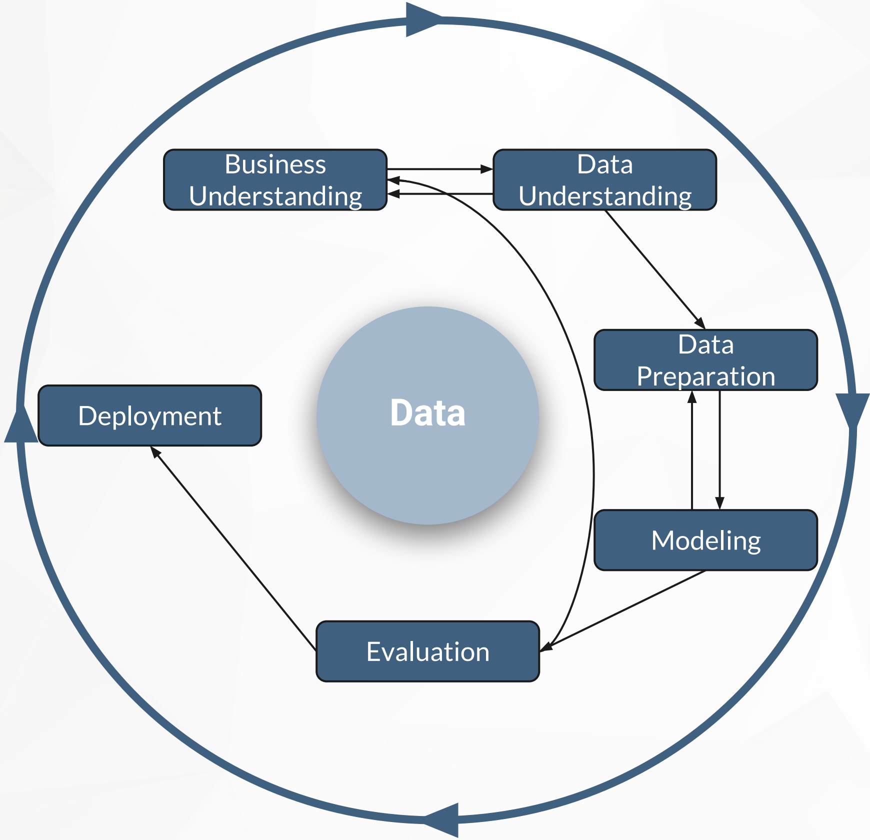


Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Tasks:

- Identifying the available sources of data
- Assessing the data quality, validity, and usability

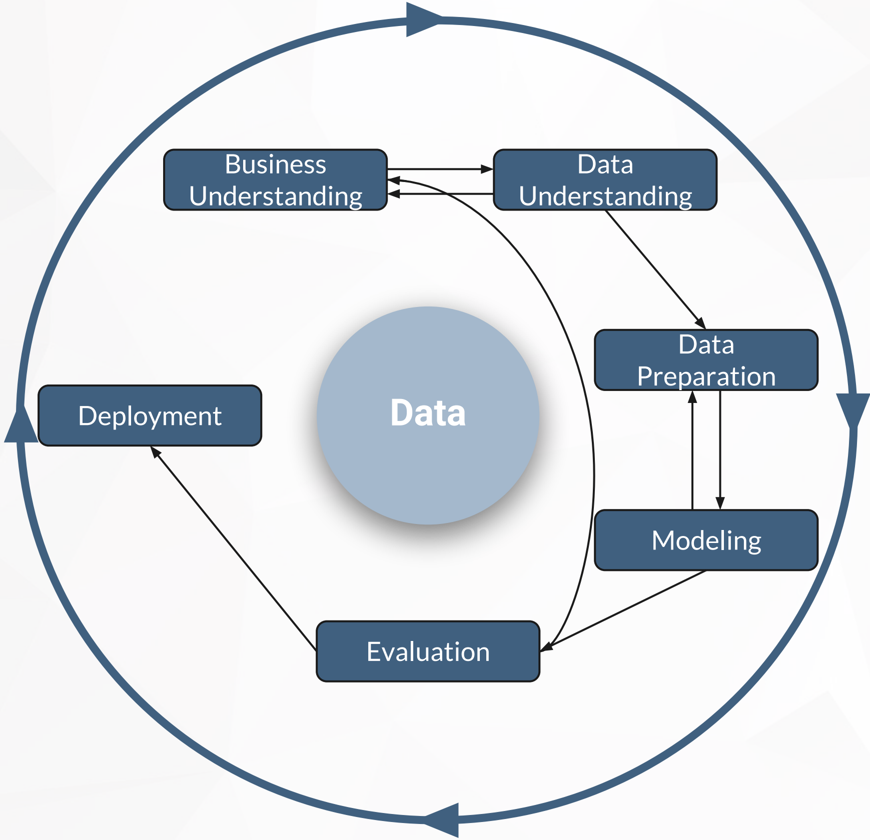


Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order.

Tasks:

- Table, record, and attribute selection and transformation
- Cleaning of data for modeling tools

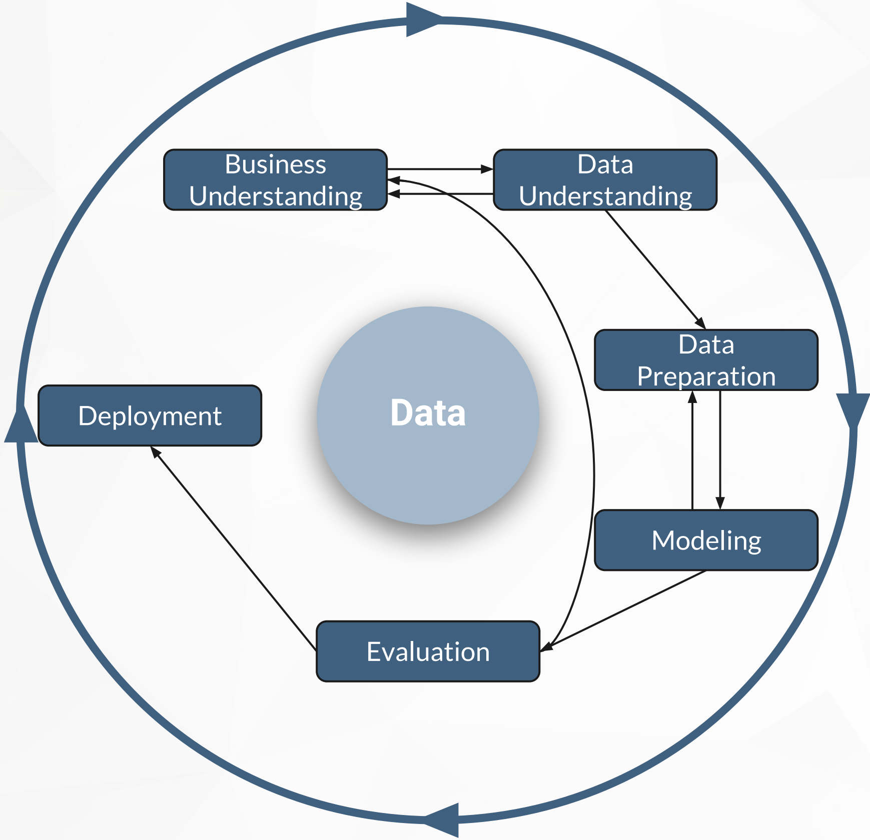


Modeling

Various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type.

Tasks:

- Choosing the appropriate ML algorithms
- Building, evaluating, and optimizing the model

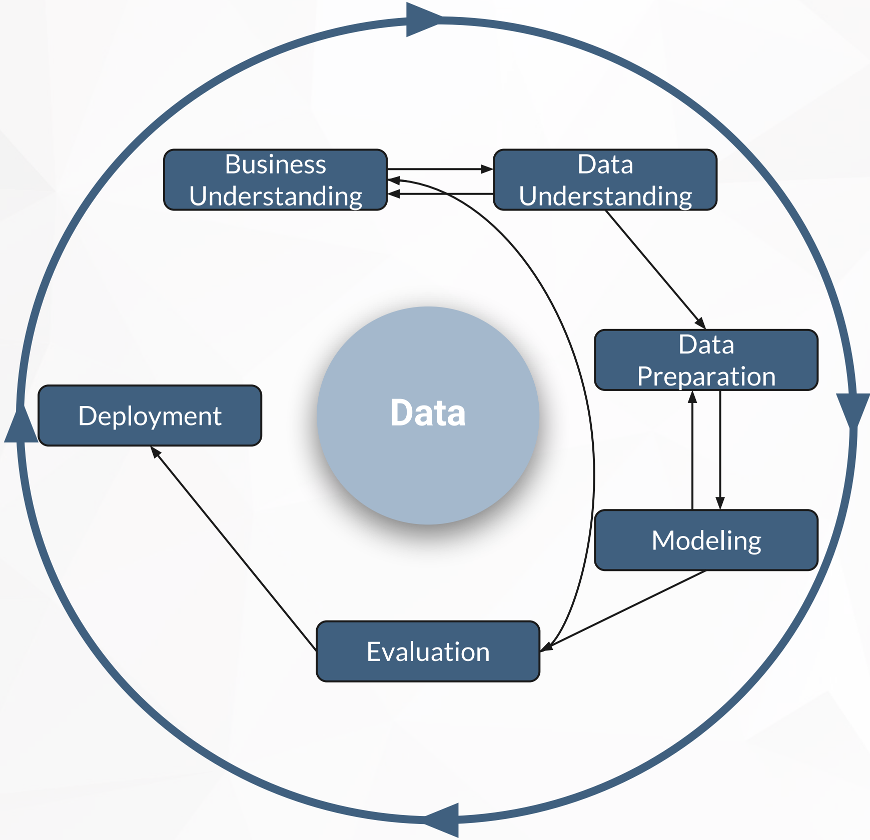


Evaluation

In this phase, you assess the degree to which the model meets your business objectives and determine if there is some business reason why this model is deficient. At the end of this phase, a decision on the use of the data mining results should be reached.

Tasks:

- Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives.
- Determine next steps



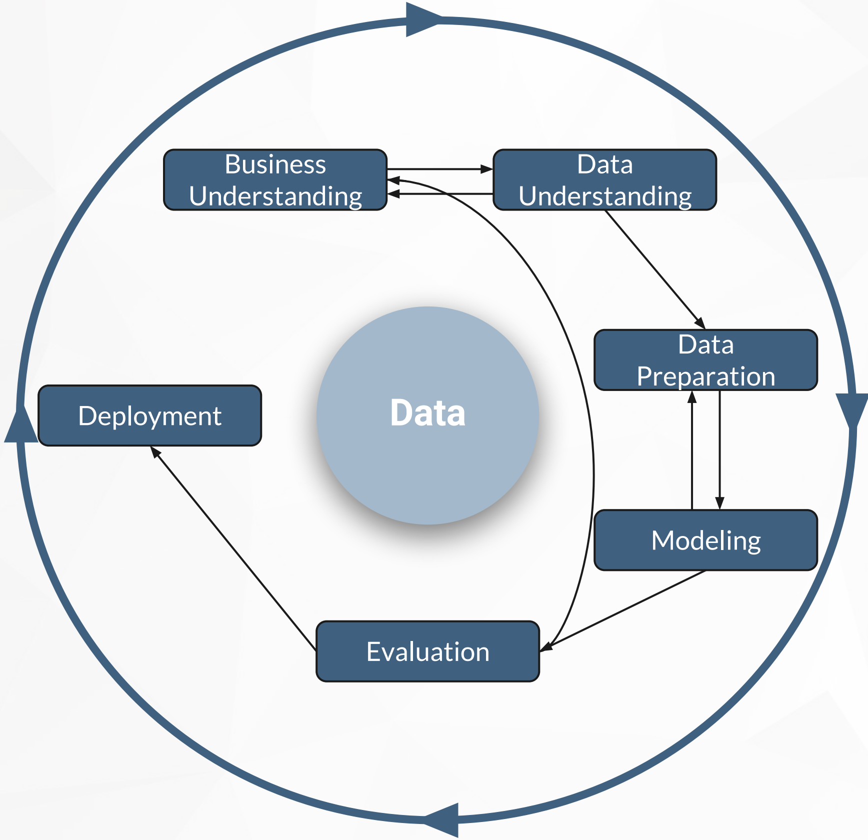
Deployment

At this stage, Evaluation results are taken and a strategy is determined for their deployment. The knowledge gained will need to be organized and presented in a way that is useful to the customer.

Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

Tasks:

- Visualization
- Presentation of Results
- Application for Business Use Case



Data Science Toolkit

Open Source Software and Tools

Python 2.7

Python Libraries

Jupyter/IPython Notebook

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its core libraries combined with its strength as a general purpose programming language make it ideal for Data Science.

Open Source Software and Tools

Python 2.7

Python Libraries

Jupyter/IPython Notebook

Numpy - package for scientific computing with Python

Pandas - library providing high-performance, easy-to-use data structures and data analysis tools

Scipy - provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization.

Sci-Kit Learn - Simple and efficient tools for data mining and data analysis

Matplotlib - 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms

Seaborn - visualization library based on matplotlib

Tensorflow - library that is used for machine learning applications such as neural networks.

** Anaconda - a distribution of Python containing Python, software packages for data analytics, data science, and scientific computing, and the Ipython notebook

Open Source Software and Tools

Python 2.7

Python Libraries

Jupyter/IPython Notebook

The Jupyter (or IPython) Notebook is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media.

Allows you to create and share documents that contain live code, equations, visualizations and explanatory text.

Uses/Features:

- data cleaning and transformation
- numerical simulation
- statistical modeling
- machine learning

LAB:
GETTING STARTED WITH JUPYTER/IPYTHON NOTEBOOK

analytiks