NOVA IMS
Information Management School

1

INTRODUCTION TO DATA SCIENCE

Data Science for Marketing

© 2021-2022 Nuno António

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Summary

1. Introduction
2. Tasks in Data Science
3. Applications in Marketing

NOVA IMS Information Management School

1.1

Introduction

Introduction to Data Science

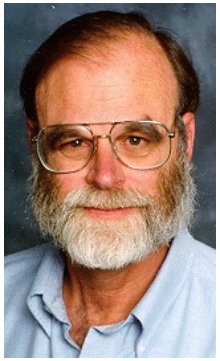
What is Data Science

Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.

- Interdisciplinary field
- Uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data
- Applies knowledge and actionable insights from data across a broad range of application domains
- Relates to other fields such as **Data Mining**, **Machine Learning**, **Network theory**, and **Big Data**

Data Science – The fourth paradigm in Science

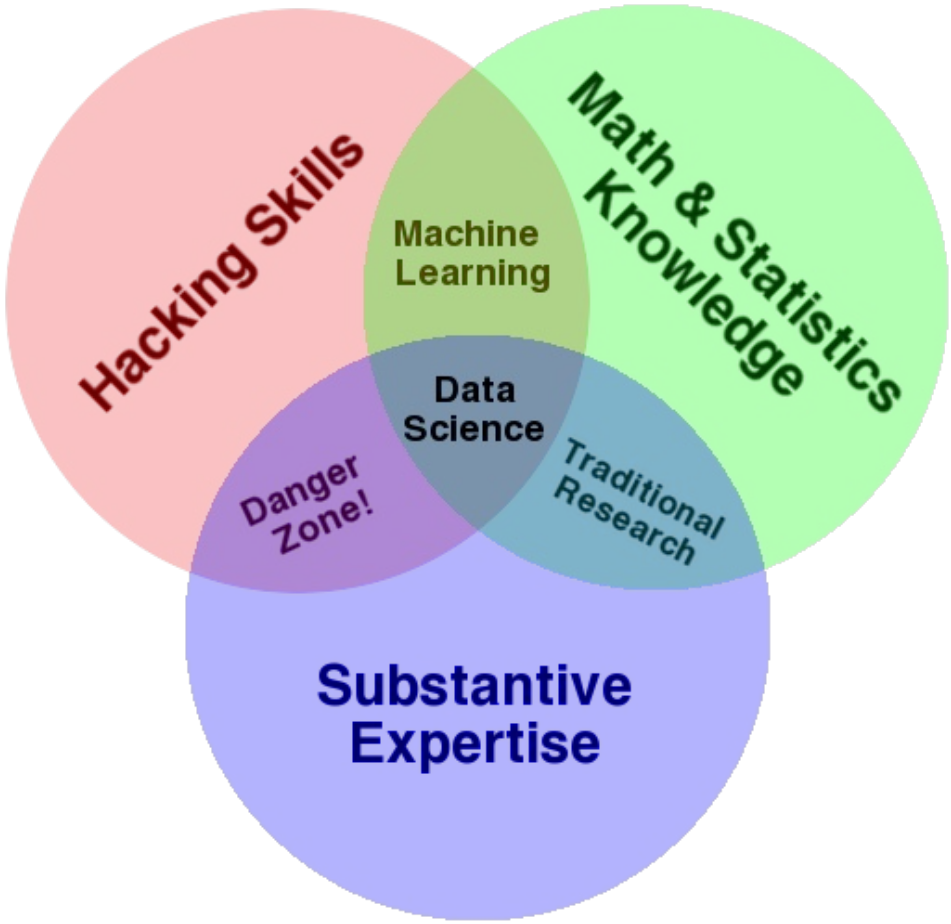
- **Experimental**: empirical measurement science
- **Theory**: causal interpretation
  - Explains experiments
  - Calculates measurements that would confirm the theoretical models
- **Simulation**: performing theory (model)- driven experiments that are not empirically possible
- **Data Science**: an empirical analysis of all types of data produced by processes



Jim Gray

Data Science combines

- Data-driven approach of statistical data analysis
- The computational power and programming acumen of computer science
- Domain-specific expertise and business intelligence



[http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram]

NOVA IMS Information Management School

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 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 packages, 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
(c) Krzysztof Zawadzki

But remember…

You do not need to know everything by memory nor to be an expert in every analytical tool. You need to know:

- How to make questions and to whom
- What techniques and algorithms apply to the business problem
- What tools to use
- Search for information and interpret the documentation

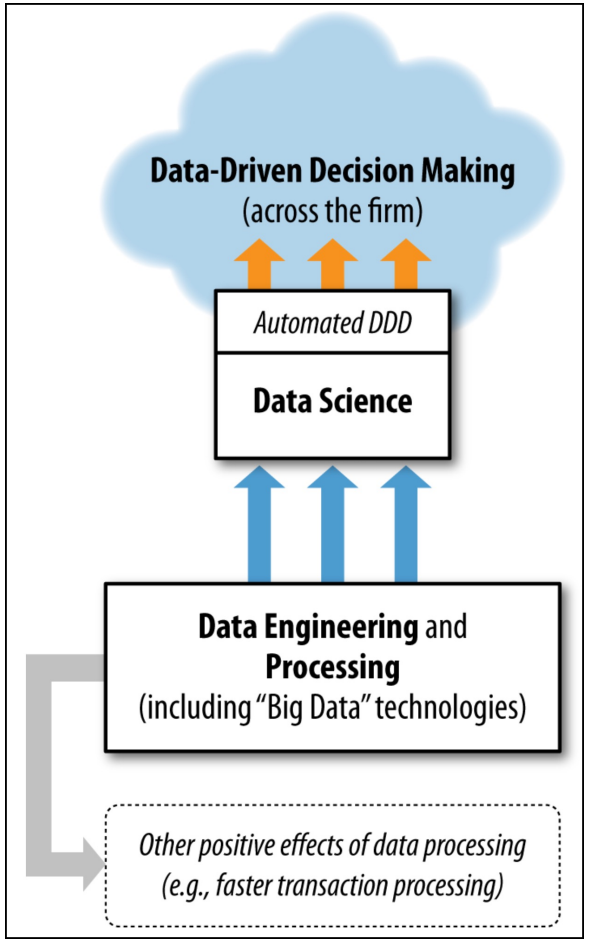
Data-driven decision-making

Data Science is the base for Data-Driven Decision-making (DDD). DDD refers to the practice of basing decisions on the analysis of data, rather than purely on intuition.

According to Brynjolfsson, Hitt, & Kim (2011) in "Strength in Numbers: How Does Data-Driven Decision-making Affect Firm Performance?" firms who apply DDD have an output and performance that is 5-6% higher than other firms

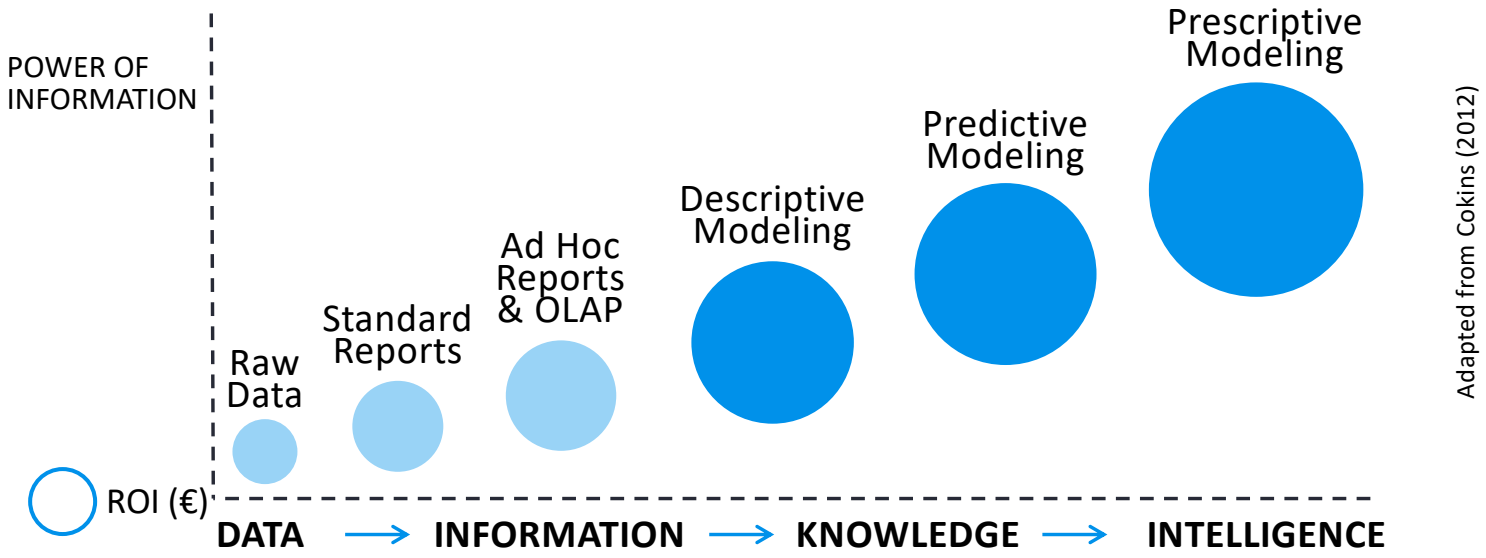
How should a marketer select advertisements:

- Based on his/her long experience in the field and eye for what will work, or
- **Based on the analysis of data of how consumers react to the different type of ads**

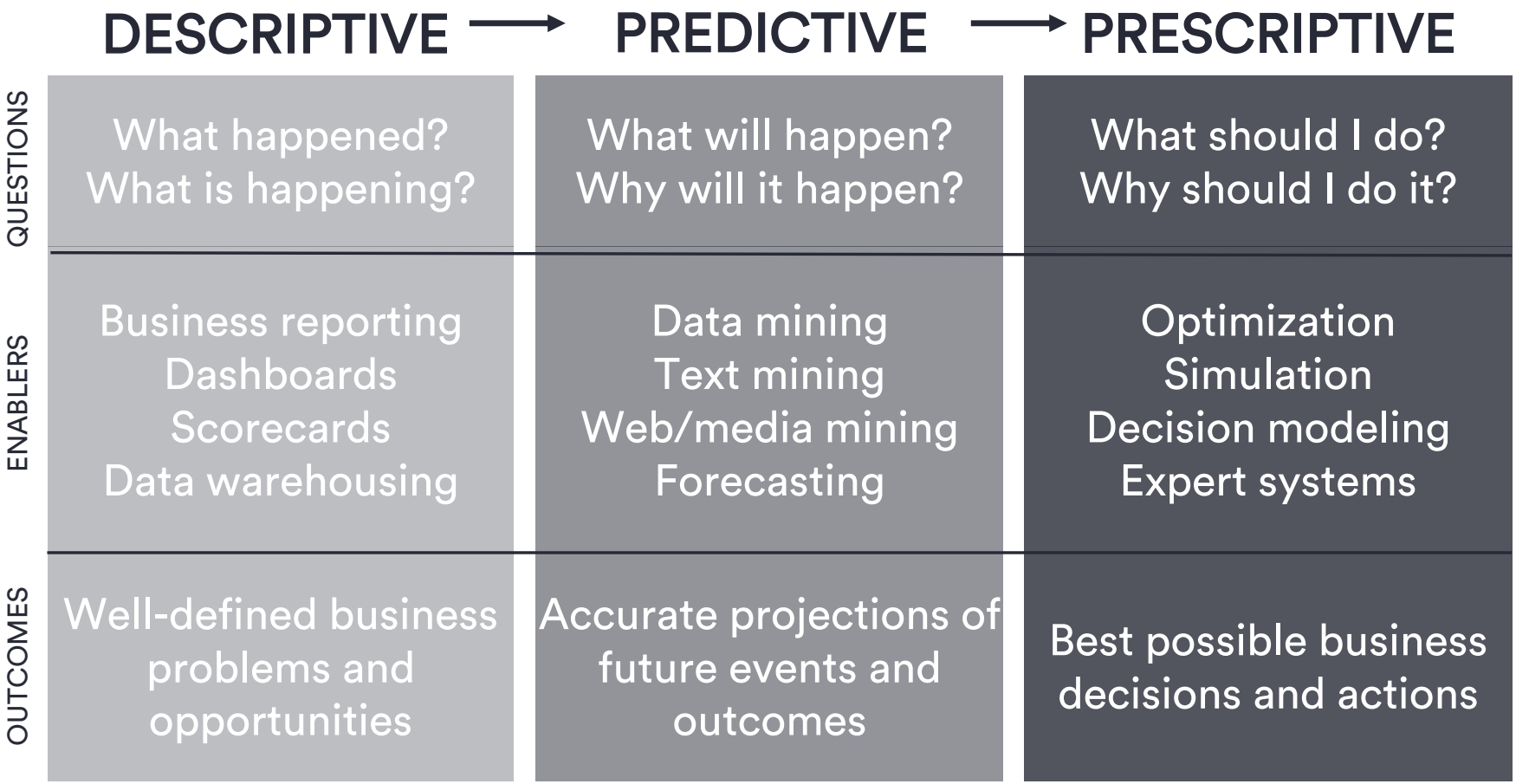


[Provost & Fawcett 2013]

The intelligence hierarchy



What is today known as “Analytics”



NOVA IMS Information Management School

1.2

Tasks in Data Science

Introduction to Data Science

Main goal in Data Science

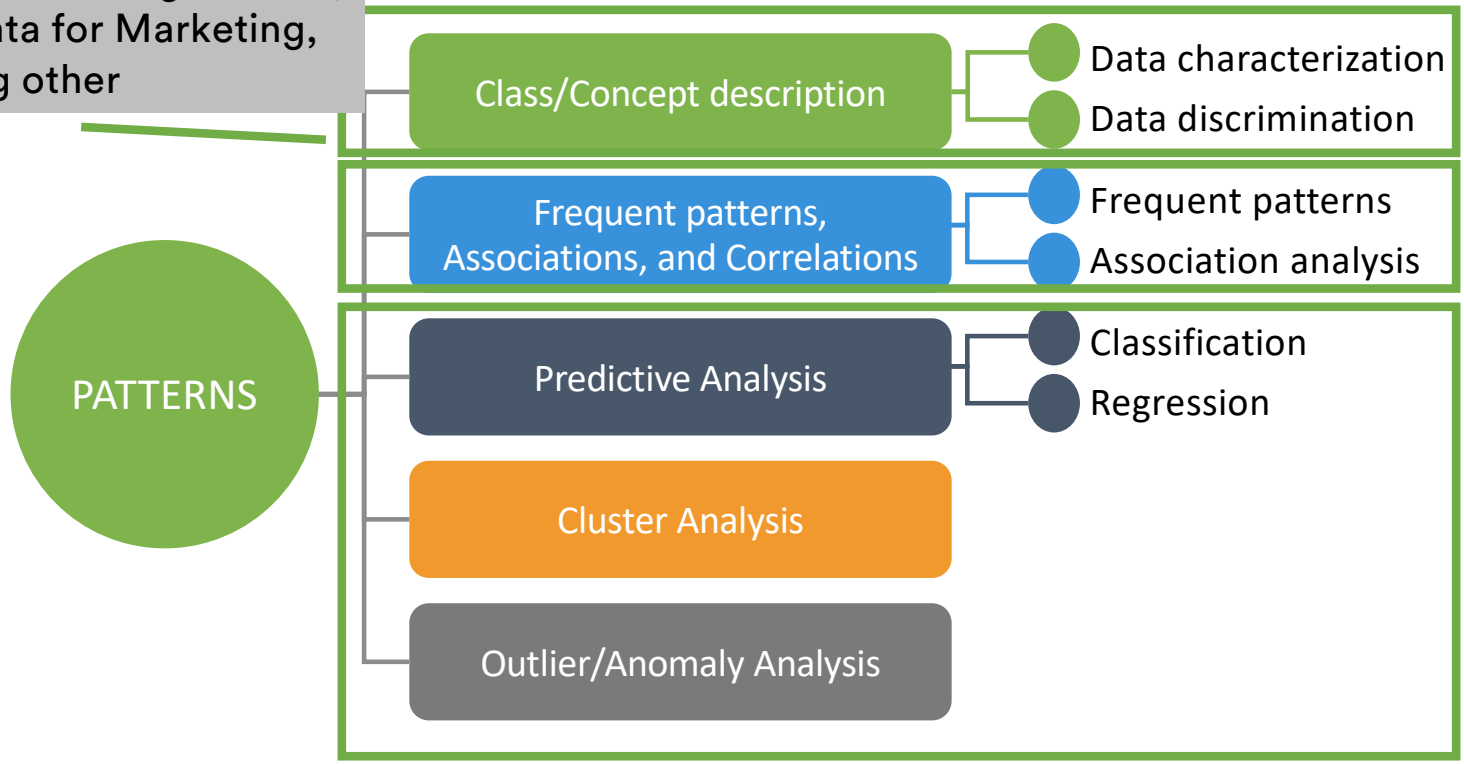
Like in Data Mining, the main goal of Data Science is to uncover interesting patterns in data. To be interesting a pattern must:

1. Easily understood by humans
2. Valid on new or test data with some degree of certainty
3. Potentially useful
4. Novel

or, validate a hypotheses the user is sought to confirm

Types of patterns

Additional information covered in Descriptive Analytics in Marketing, Business Intelligence I/II, Big Data for Marketing, among other



Networks covered in Social Network Analysis, Social Media Analytics, among other

Covered in Machine Learning for Marketing, Big Data for Marketing, Text Mining, among other

Class/Concept descriptions

Useful to describe classes or concepts (e.g., segments of customers) in summarized, concise, and yet precise terms

Data characterization

- Summarization of the general characteristics of a target class of data (for example the characteristics of products with sales that increased by 10% in the previous year)
- Data obtained from a transactional or OLAP database
- The output is usually pie charts, bar charts, line charts, pivot tables, or crosstabs

Data discrimination

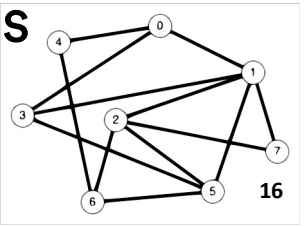
- Comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes (e.g., products with sales that increased by 10% in the previous year against those that decreased 30%)
- The type of data and outputs are similar to data characterization

Frequent patterns, associations, and correlations

Frequent pattern mining investigates recurring relationships in a dataset, like the co-occurrence of two or more objects of interest.

Types of frequent patterns:

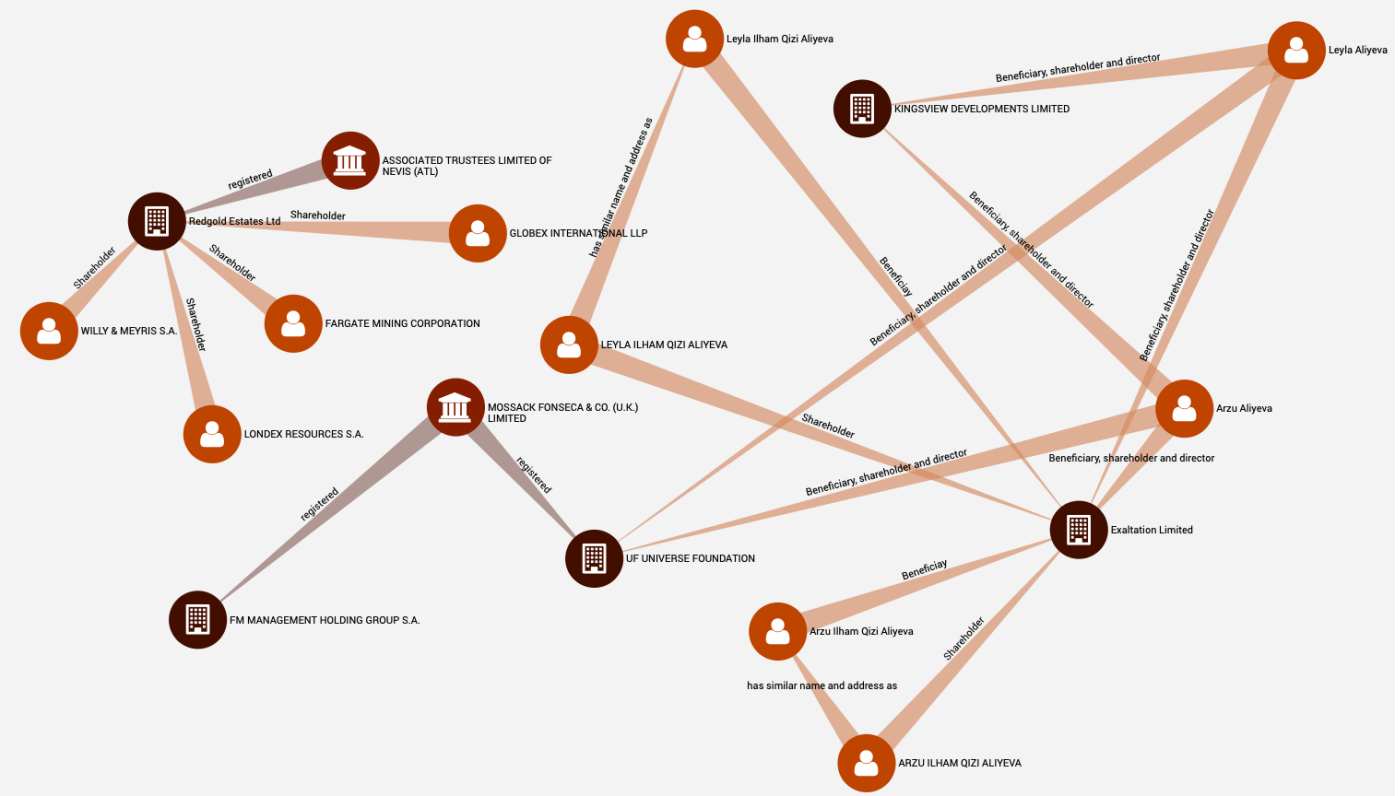
- **Frequent itemsets:** typically refers to a set of items that often appear together in a transactional dataset (e.g., milk and bread in a grocery store dataset – "Market basket analysis")
- **Sequential pattern:** refers to patterns that occur across time or positions in a dataset (e.g., a customer first buys a laptop, followed by a webcam, and then an external hard disk)
- **Structured pattern:** structural patterns (graphs, trees, or lattices) that can be combined with itemsets or sequential patterns



Frequent patterns, associations, and correlations

Frequent structured patterns example

"The Panama Papers" – The International Consortium of Investigative Journalists Exposed highly connected networks of offshore tax structures



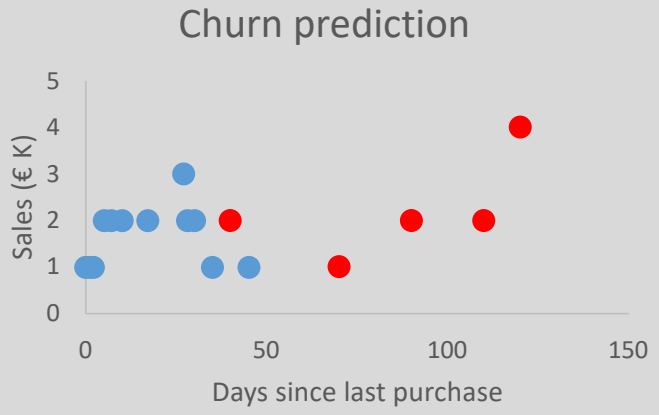
source: https://neo4j.com/blog/analyzing-panama-papers-neo4j/

Predictive analysis

Uses labeled input attributes to predict an outcome

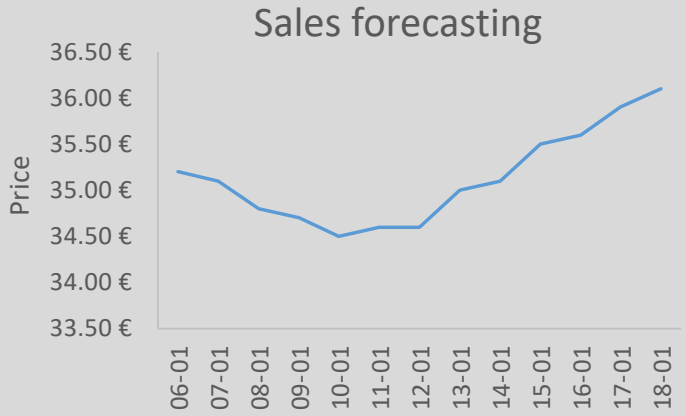
Classification

Process of finding a model to predict data classes or concepts (the outcome is categorical)



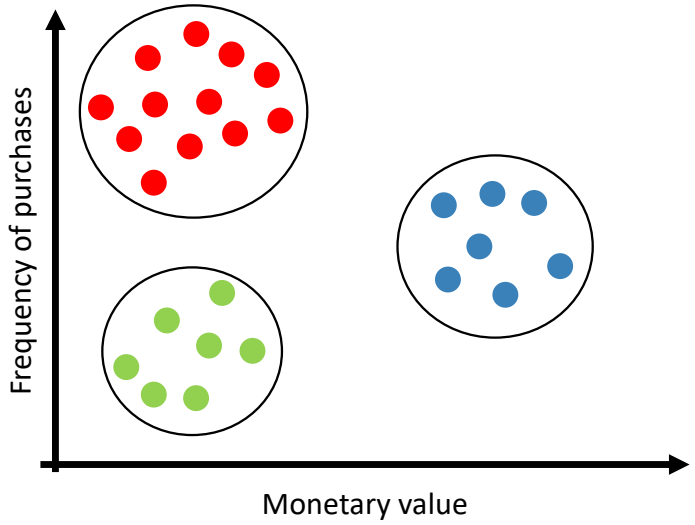
Regression

Process of finding a model to predict numeric outcomes (the outcome is continuous)



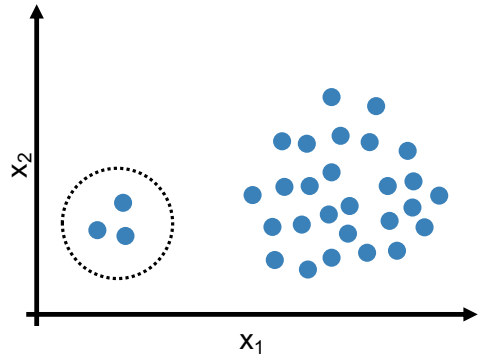
Cluster analysis

- Method for the **partition of multiple entities into groups** so that, within the same group, entities share a certain degree of similarity, but are ideally very dissimilar to the entities in the other groups (e.g., customer segmentation)
- Unlike classification and regression, clustering analyzes non-labeled data



Outlier/anomaly analysis

- An outlier (or anomaly) is a data point that deviates significantly from the dataset's remaining points. Thus, the objective of an outlier analysis is the **discovery of unusual data points** in the dataset
- These data points can occur due to human errors (data entry errors), experimental errors (errors related to data extraction or preparation), processing errors (data manipulation errors), sampling errors (incorrect sampling or data sources selection), or novelty errors (indicating, for example, new trends)
- Many Data Mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) **the rare events can be more interesting** than the more regularly occurring ones



NOVA IMS Information Management School

1.3

Applications in Marketing

Introduction to Data Science

Customer-oriented

- Targeting current customers
  - Segmentation based on touchpoint engagement
  - Segmentation based on purchase patterns
  - Micro-segmentation/personalization
- Finding new customers
  - Lead targeting
  - Lead scoring
- Retaining customers
  - Churn prediction
- Predicting sales
  - Demand forecast

Product-oriented

- Understanding markets
  - Understanding customers' likes and dislikes
- Positioning products
  - Budget optimization
- Developing new products
  - Real-time experimentation
- Promoting products
  - Optimize campaigns
- Recommending products
  - Market basket analysis
- Assessing brands and prices
  - Pricing analysis
  - Competitor's analysis
- Predicting sales
  - Demand forecasting

Algorithmic marketing

The advancement of digital marketing channels changed the game and created an environment that requires millions of *micro-decisions* to be made, which simply cannot be done efficiently without intelligent marketing software and algorithms:

- Targeted sales promotions
- Dynamic pricing in brick-and-mortar and online stores
- E-Commerce search and recommendations services
- Online advertising

Questions?

Data Science for Marketing

© 2021-2022 Nuno António (Rev. 2022-08-10)

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa