Universiti Malaysia PAHANG
Engineering • Technology • Creativity

Chapter 1
Introduction to Data Mining

DATA MINING (BSD3533)

DR. KU MUHAMMAD NA'IM KU KHALIF

5-STAR WORLD CLASS TECHNOLOGICAL UNIVERSITY

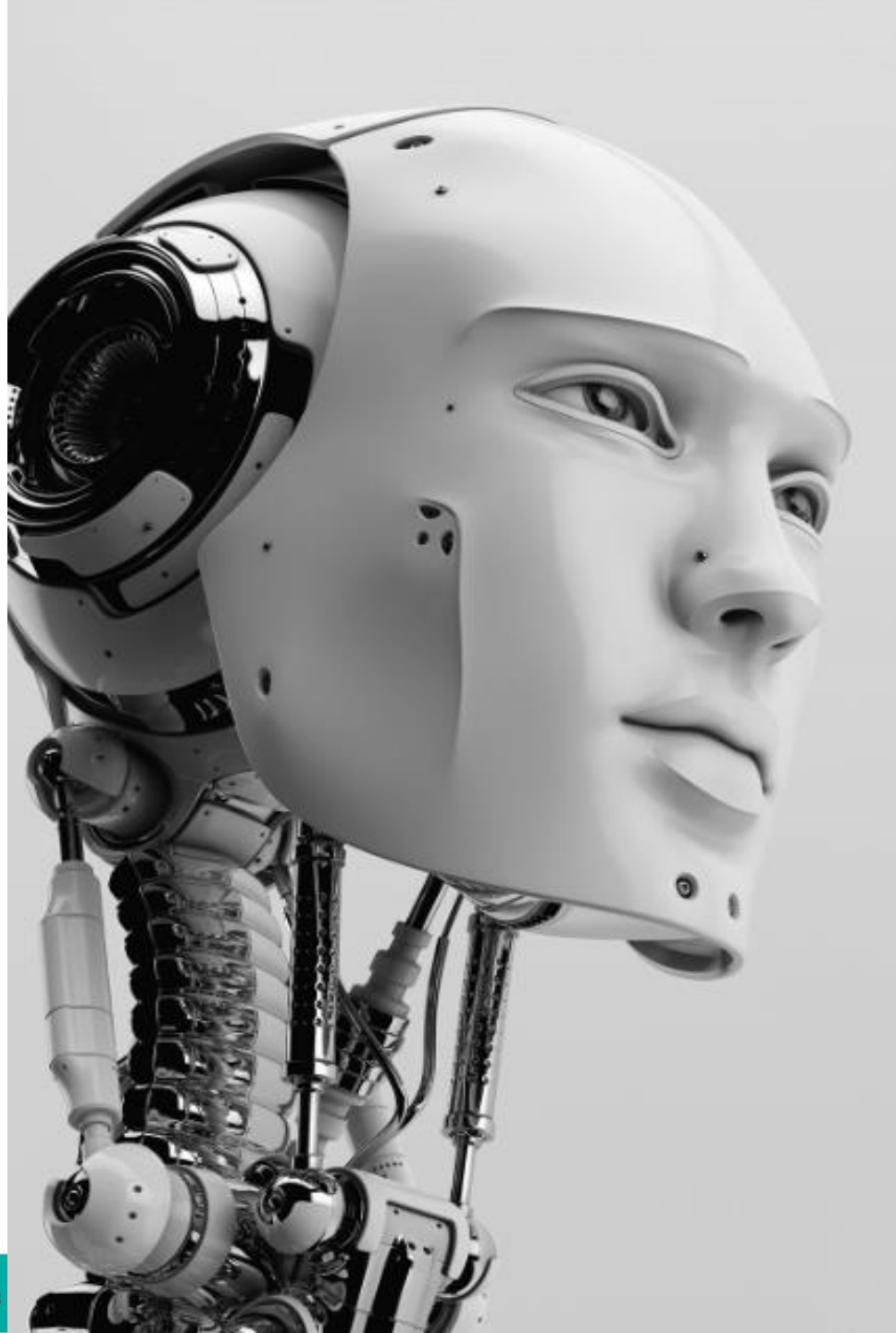
UMPMalaysia

Content

Chapter 1.1: Basic Notions for Data Mining

By the end of this topic, you should be able to:

- understand the concept of data mining and how it's applied in the real world.
- understand the needs why to study data mining.
- acquire the ethics in dealing with data.



An Overview

- Data mining is not a new invention that come with the digital age. The concept has been around for over a century, but came into greater public focus in the 1930s.
- According to Hacker Bits, one of the first modern moments of data mining occurred in 1936, when Alan Turing introduced the idea of a universal machine that could perform computations similar to those of modern-day computers.

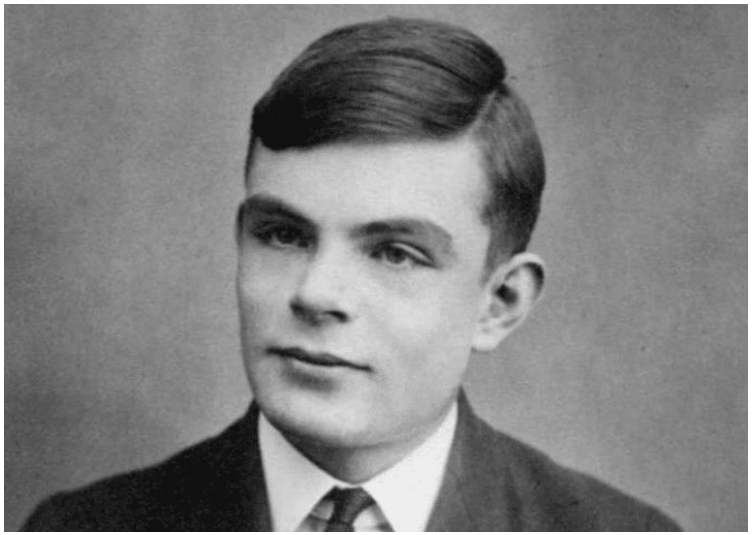


Figure 1.1: The British mathematician Alan Turing was one of the more unquantifiable original minds of the twentieth century.

Data Mining Definitions

"Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical and statistical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data."

(Microsoft,2019)

"Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more."

(SAS, 2020)

human Interact

- The process must be automatic or (more usually) semiautomatic. The pattern discovered must be meaningful in that they lead to some advantages in many real world applications.
- Data mining assists users to find the **insight** from the data/ information.
- **Insight** may refer to the capability to gain deep understanding of something. The user gains it after understanding what it all means for data.

Data Mining Concepts

- Data mining can also be stated as knowledge discovery in data (KDD). KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.
- The volume of data produced doubles every two years. Unstructured data alone makes up 90% of the digital universe. But more information does not necessarily mean more knowledge.
- Data mining is an interdisciplinary subfield based on the usefulness of data in artificial intelligence, statistics and machine learning.

Figure 1.2: Data mining as an interdisciplinary subfield from artificial intelligence, statistics and machine learning.

Data Analysis vs Data Analytics vs Data Mining

Data analysis refers to the process of separating a whole problem/ data/ information into its parts so that the parts can be critically examined at the granular level (Delen, 2015). It involves the process of inspecting, cleaning, transforming and modelling data focusing on the objective of discovering the useful information.

Data analytics is a process of dealing with data using variety of methods, technologies and associated tools for creating new knowledge/ insight to solve complex problems and make better and faster decisions (Delen, 2015). It comprises the algorithms, the specific software to be used, the process of data science implementation and so forth.

Data mining is the process of discovering actionable information from large sets of data. Data can be in any form either structured or unstructured or semi-structured. Data mining techniques convert data into one single form that can be easily retrieved from the server to any user and can be easy processed.

Data Important

- In many niche areas, most commercial organizations have realized that there is **huge value hiding** in the data and are employing the techniques we ask about to realize that value.
- Eventually, these works produce insights, things that we may not have known otherwise.
- Insights are the items of information that cause a behaviour change.
- These insights can help shape the goals of entire organizations.

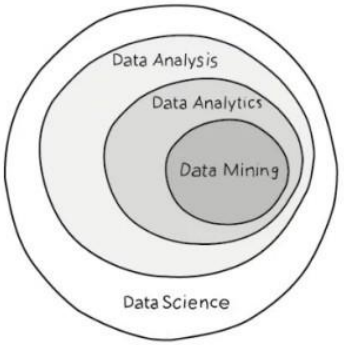


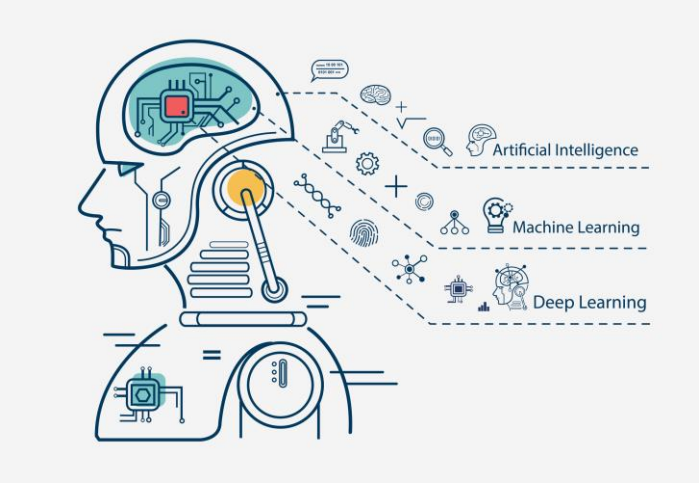
Figure 1.3: Data mining as a subset of data science knowledge.

Data Mining and Machine Learning

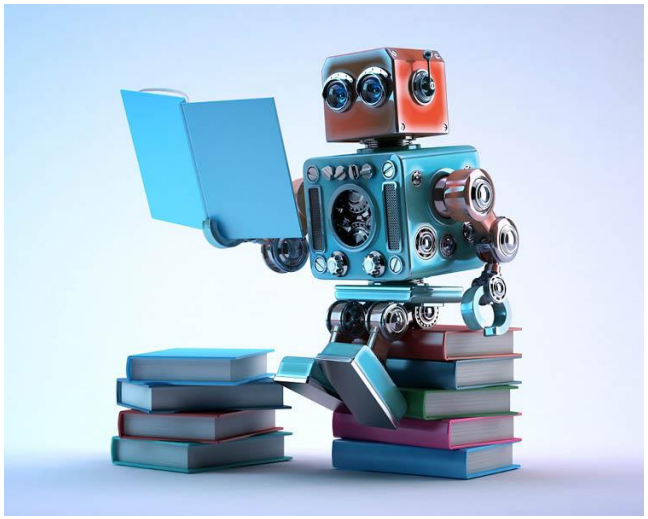
- Both data mining and machine learning fall under the aegis/ support of Data Science, which makes sense since they both use data.
- Both processes are used for solving complex problems, so consequently, many people (erroneously) use the two terms interchangeably. This isn't so surprising, considering that machine learning is sometimes used as a means of conducting useful data mining. While data gathered from data mining can be used to teach machines, the lines between the two concepts become a bit blurred.
- Furthermore, both processes employ the same critical algorithms for discovering data patterns. Although their desired results ultimately differ, something which will become clear as you read on.

Machine Learning

- Machine learning is related to the development and designing of a machine that can learn itself from a specified set of data to obtain a desirable result without it being explicitly coded. Hence, machine learning implies 'a machine which learns on its own.
- Arthur Samuel invented the term machine learning, an American pioneer in the area of computer gaming and artificial intelligence in 1959. He said that "it gives computers the ability to learn without being explicitly programmed."



- Machine learning is a technique that creates complex algorithms for large data processing and provides outcomes to its users. It utilizes complex programs that can learn through experience and make predictions.
- The algorithms are enhanced by themselves by frequent input of training data.
- The aim of machine learning is to understand information and build models from data that can be understood and used by humans.



Data Miner Influencer

Figure 1.4: Kirk Borne, Principal Data Scientist Booz Allen Hamilton. Global Speaker. PhD Astrophysicist. https://www.linkedin.com/in/kirkdborne/

Dr. Kirk Borne is a data scientist and an astrophysicist. He is Principal Data Scientist in the Strategic Innovation Group at Booz-Allen Hamilton since 2015. He was Professor of Astrophysics and Computational Science in the George Mason University (GMU) School of Physics, Astronomy, and Computational Sciences during 2003-2015.

Prior to that, he spent nearly 20 years supporting NASA projects, including NASA's Hubble Space Telescope as Data Archive Project Scientist, NASA's Astronomy Data Centre, and NASA's Space Science Data Operations Office. He has extensive experience in large scientific databases and information systems, including expertise in scientific data mining.

He was a contributor to the design and development of the new Large Synoptic Survey Telescope (LSST), for which he contributed in the areas of science data management, informatics and statistical science research, galaxies research, and education and public outreach.

Figure 1.5: DJ Patil, Former US Chief Data Scientist. https://www.linkedin.com/in/dpatil/

DJ Patil is the Head of Technology for Devoted Health, a Senior Fellow at the Belfer Centre at the Harvard Kennedy School, and an Advisor to Venrock Partners.

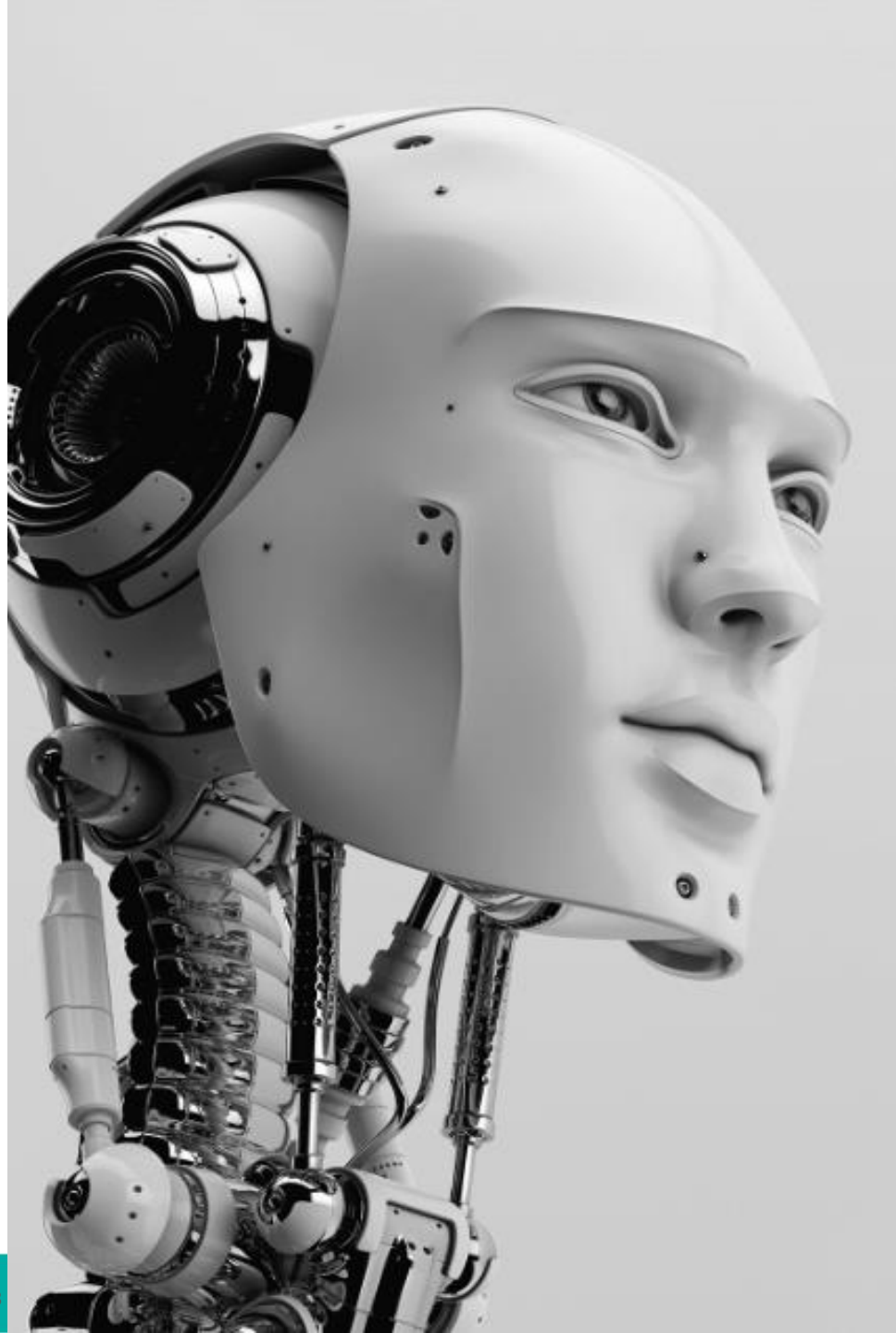
Dr. Patil was appointed by President Obama to be the first U.S. Chief Data Scientist where his efforts led to the establishment of nearly 40 Chief Data Officer roles across the Federal government. He also established new health care programs including the Precision Medicine Initiative and the Cancer Moonshot, new criminal justice reforms including the Data-Driven Justice and Police Data Initiatives that cover more than 94 million Americans, as well as leading the national data efforts.

He also has been active in national security and for his efforts was awarded by Secretary Carter the Department of Defence Medal for Distinguished Public Service for his efforts which the highest honour the department bestows on a civilian.

Chapter 1.2: Stages of Data Mining Process

By the end of this topic, you should be able to:

- understand the stages of data mining process.
- understand the application of CRISP-DM in the real world.



Data Mining Process

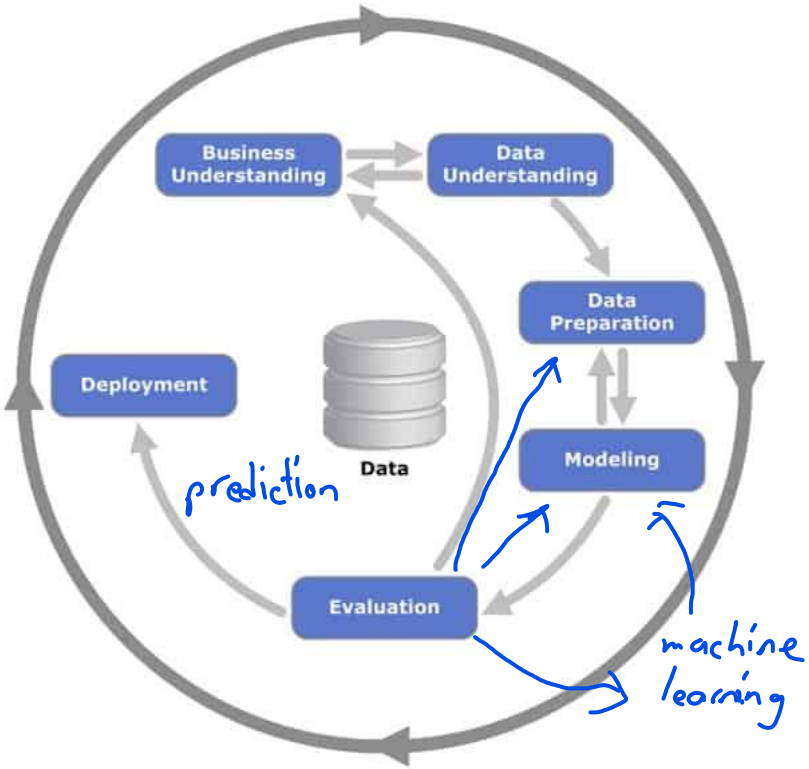


Figure 1.6 : CRISP-DM - CRoss Industry Standard Process for Data Mining

- A process is a general strategy that guides the processes and activities within a given domain.
- Process does not depend on particular technologies or tools, nor is it a set of techniques or recipes.
- Rather, the process provides the data miner with a framework on how to proceed with whatever methods, processes and heuristics will be used to obtain answers or results.

What is CRISP-DM?

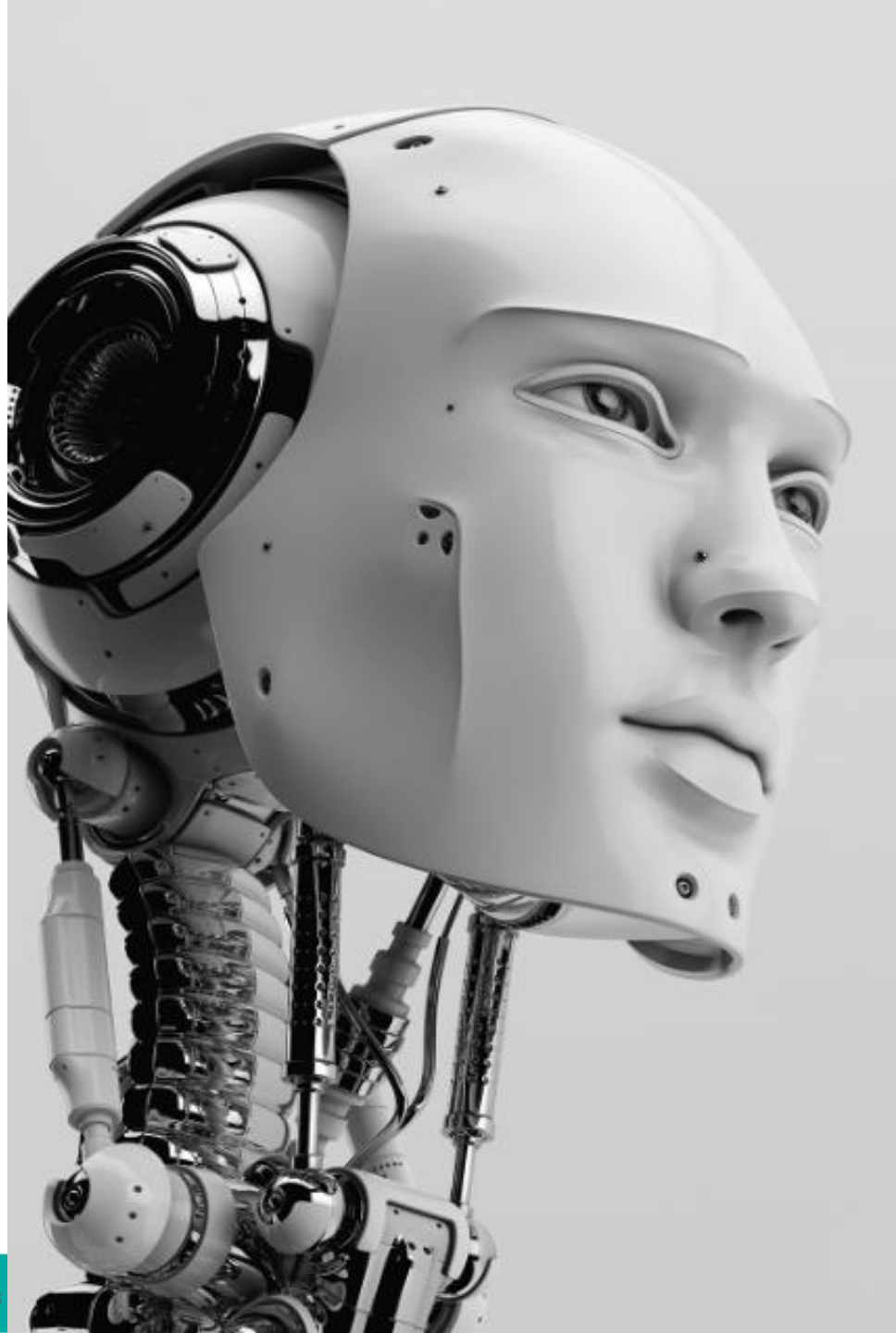
- The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model with six phases that naturally describes the Data Science life cycle.
- It is like a set of guardrails to help you plan, organize, and implement your data science project.

a) Business understanding – What does the business need?
b) Data understanding – What data do we have/ need? Is it clean?
c) Data preparation – How do we organize the data for modelling?
d) Modelling – What modelling techniques should we apply?
e) Evaluation – Which model best meets the business objectives?
f) Deployment – How do stakeholders access the results?

Chapter 1.3:
Data Mining Techniques

By the end of this topic, you should be able to:

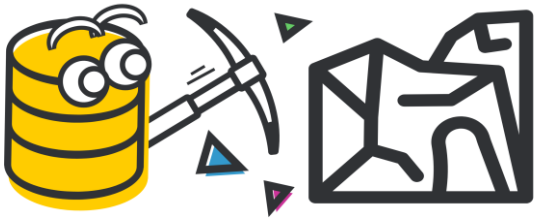
- understand the concept of data mining techniques.
- understand the needs why to study data mining techniques.
- apply the data mining techniques in real world phenomena.



Data Mining Techniques and Algorithms

Data mining uses already build tools to get out useful hidden patterns trends and predictions of future can be obtained using techniques:

a. Predictive Modelling
b. Database Segmentation
c. Link Analysis
d. Deviation Detection



Predictive Modelling

- Predictive modeling is based on predicting the outcome of an event. It is designed on a pattern similar to the human learning experience in using observations to form a model of the important characteristics of some task. It is developed using a supervised learning approach, where we have some labeled data and we use this data to predict the outcome of unknown instances.
- It can be of two types, i.e., regression and classification.
- Some of the applications of predictive modeling are: predicting the outcome of an event, predicting the sale price of a property, predicting placement of students, predicting the score of any team during a cricket match and so on.

Regression Analysis

- Regression analysis tries to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable. Independent variables can be affected by each other, but it does not mean that this dependency is both ways as is the case with correlation analysis. A regression analysis can show that one variable is dependent on another but not vice-versa.
- Data mining algorithm – Linear regression (single and multiple), Logistic regression.
- Applications: to determine different levels of customer satisfactions and how they affect customer loyalty and how service levels can be affected by for example the weather.

Classification Analysis

- Classification Analysis is a systematic process for obtaining important and relevant information about data, and metadata - data about data. The classification analysis helps to identify the categories the data belongs. Classification analysis is closely linked to cluster analysis as the classification can be used to cluster data.
- Data mining algorithm– Decision tree, Logistic Regression, Naïve Bayes, Neural Network
- Applications: Your email provider performs a well-known example of classification analysis: they use algorithms that can classify your email as legitimate or mark it as spam. This is done based on data that is linked with the email or the information that is in the email, for example, certain words or attachments that indicate spam.

Databased Segmentation

- Database segmentation is based on the concept of clustering of data and it falls under unsupervised learning, where data is not labelled.
- This data is segmented into groups or clusters based on its features or attributes. Segmentation is creating a group of similar records that share a number of properties.
- Applications of database segmentation include customer segmentation, customer churn, direct marketing, and cross-selling.

Clustering Analysis

- Clustering analysis is the process of identifying data sets that are similar to each other, to understand the differences as well as similarities within the data. Clusters have certain traits in common that can be used to improve targeting algorithms. For example, clusters of customers with similar buying behaviour can be targeted with similar products and services to increase the conversation rate.
- Data mining algorithm – K means, Fuzzy c-Means, Hierarchical Clustering, Density Based
- Applications: A result of a clustering analysis can be the creation of personas. Personas are fictional characters created to represent the different user types within a targeted demographic, attitude and/or behaviour set that might use a site, brand or product similarly.

Link Analysis

- Link analysis aims to establish links, called associations, between the individual record, or sets of records, in a database.
- There are three specialisations of link analysis.
    a. Associations discovery
    b. Sequential pattern discovery
    c. Similar time sequence discovery

Association Rule Learning

- Association rule learning enables the discovery of interesting relations (interdependencies) between different variables in large databases. Association rule learning uncovers hidden patterns in the data that can be used to identify variables within the data and the co-occurrences of different variables that appear with the greatest frequencies.
- Data mining algorithm – Apriori
- Applications: Association rule learning is often used in the retail industry when to find patterns in point-of-sales data. Wall-Mart uses massive data bank to predict what America wants to buy. The experts mined the data and found that the stores would indeed need certain products – and not just the usual flashlights. “We didn’t know in the past the strawberry Pop-Tarts increase in sales, like seven times their normal sale rate, ahead of a hurricane,” Dillman said in a interview. And the pre-hurricane top-selling item was beer.

Deviation Detection

- Deviation detection is a relatively new technique in terms of commercially available data mining tools.
- It is based on identifying the outliers in the database, which indicates deviation from some previously known expectations and norm.
- This operation can be performed using statistics and visualization techniques.
- Applications of deviation detection include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing.

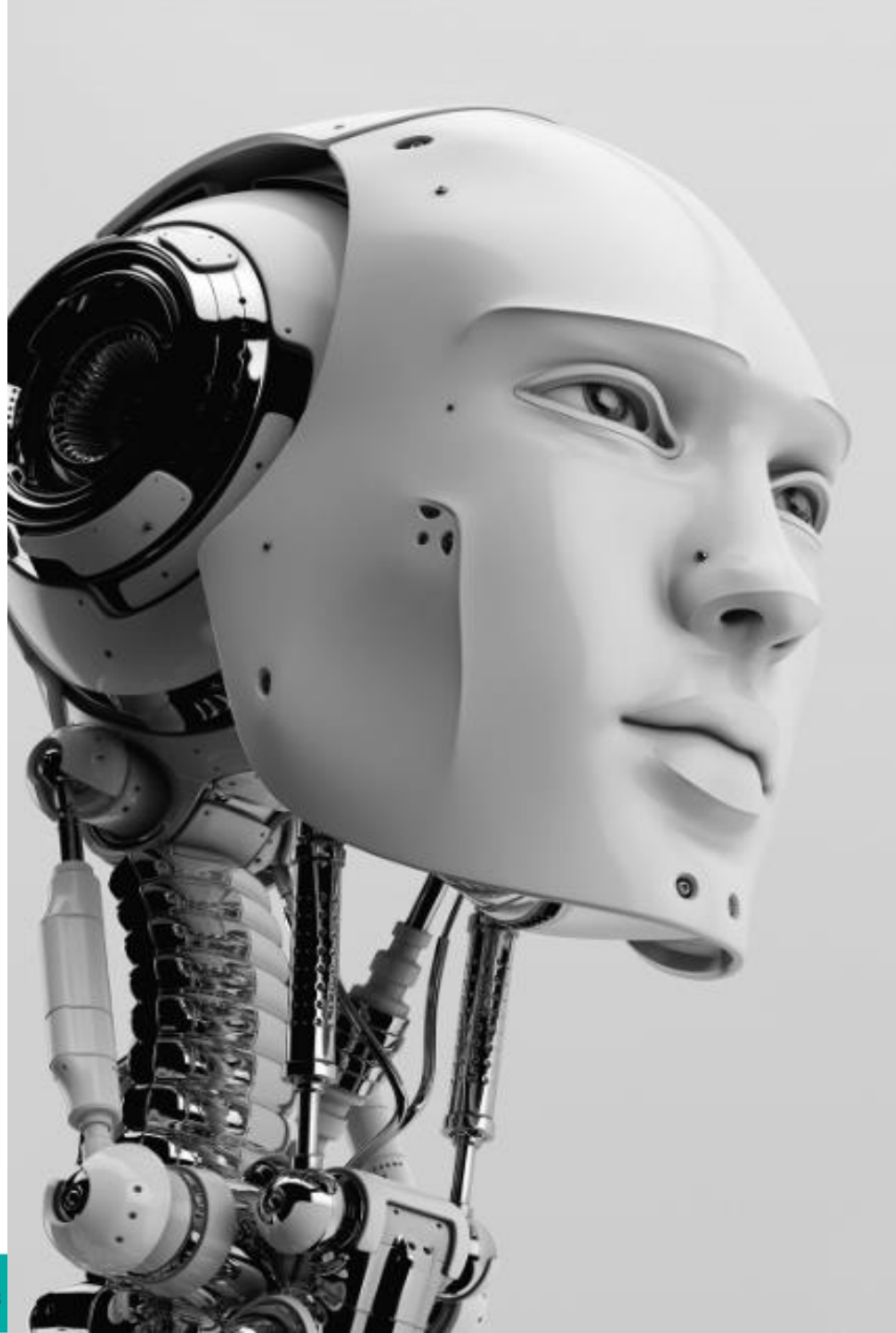
Anomaly/ Outlier Detection

- Anomaly detection refers to the search for data items in a dataset that do not match a projected pattern or expected behaviour. Anomalies are also called outliers, exceptions, surprises or contaminants and they often provide critical and actionable information. An outlier is an object that deviates significantly from the general average within a dataset or a combination of data. It is numerically distant from the rest of the data and, therefore, the outlier indicates that something is out of the ordinary and requires additional analysis.
- Data mining algorithm – Regression Analysis, Neural network, SVM, Naïve Bayes, Fuzzy c-Means
- Applications: in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, detecting ecosystem disturbances, and detection crime investigation detection.

Chapter 1.4: Applications

By the end of this topic, you should be able to:

- understand the applications of data mining on real world phenomena.
- recognize that have been applied in the real world phenomena.



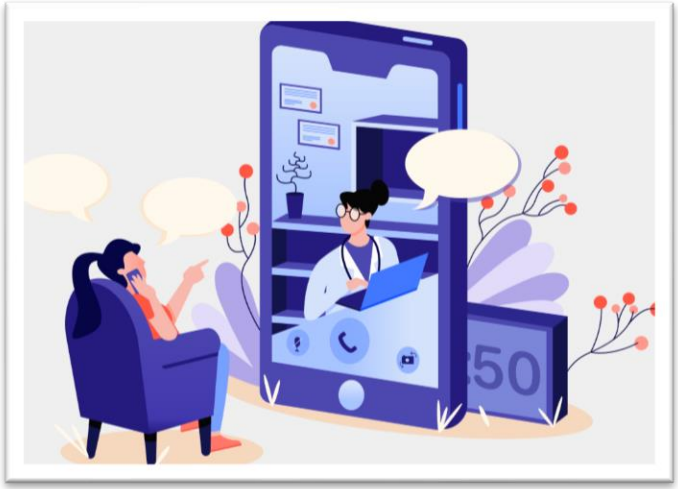
Data Mining Applications

- These patterns and trends can be collected and defined as a data mining model.
- Mining models can be applied to some applications, such as:

1. medicine
2. risk and probability
3. health care and insurance domain
4. financial firms, banks, and their analysis
5. transportation
6. education

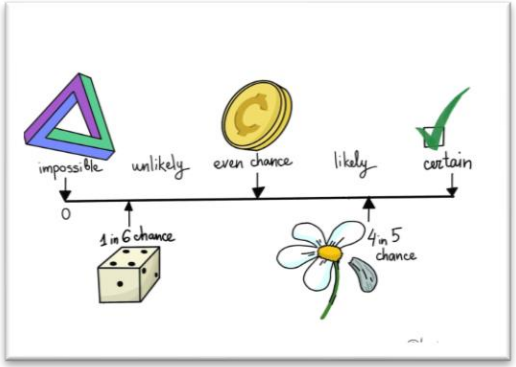
1. Medicine

Data mining helps in the identification of patterns that have successful medical therapies for various kinds of illnesses to ensure that the patients receive appropriate attention whenever needed. It helps to diagnose diseases and also plays a very important role in the treatment of different diseases. It also helps medicine and healthcare based insurers detect fake and fraud cases.



2. Risk and probability

Data Mining assists in choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios and assigning probabilities to diagnoses or other outcomes.



3. Health care domain and insurance domain

Data mining is used to efficiently track and monitor a patient's health condition and also can help in efficient diagnosis based on past sickness records. It can help insurance firms make crucial business decisions. The insurance sector is primarily dependent on the customer base.

4. Financial firms, banks, and their analysis

Data mining helps in tracking suspicious activities or any kind of mischievous or fraudulent transactions, be it related to credit card or net banking or any other banking service. The business success of the banking industry depends strongly on the credit risk evaluation of potential debtors.

5. Transportation

The historical data will help to identify the transport mode of a particular customer who generally opts for going to a particular place, thereby providing the customer with alluring offers and heavy discounts on new products and launched services. Data mining adopted in the transportation industry attempt to overcome the direct and indirect traffic issues on humanity and societies.

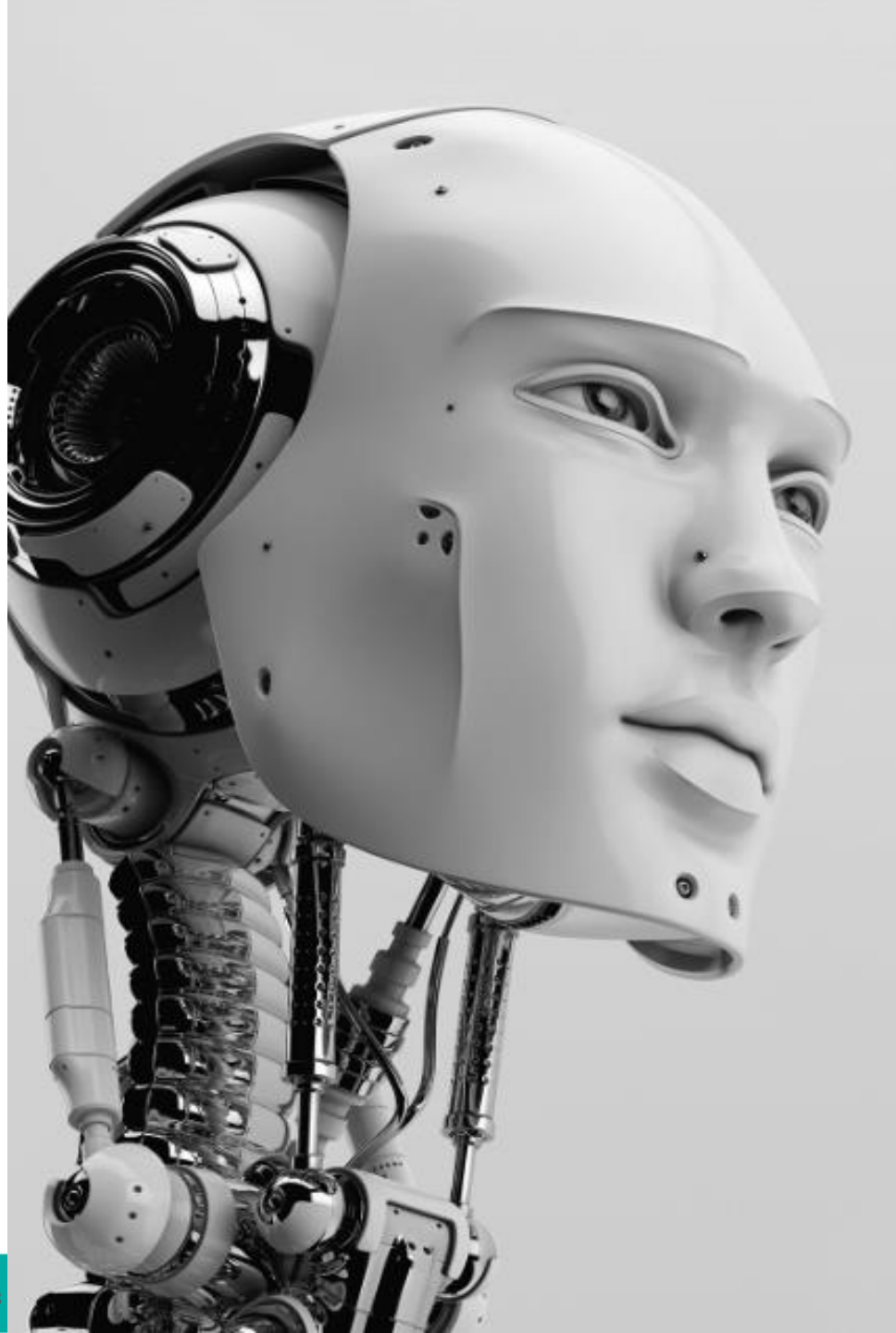
6. Education

The application of data mining has been prevalent where the emerging field of educational data mining focuses mainly on the ways and methods by which the data can be extracted from age-old processes and systems of educational institutions.

Chapter 1.5: Software for Data Mining

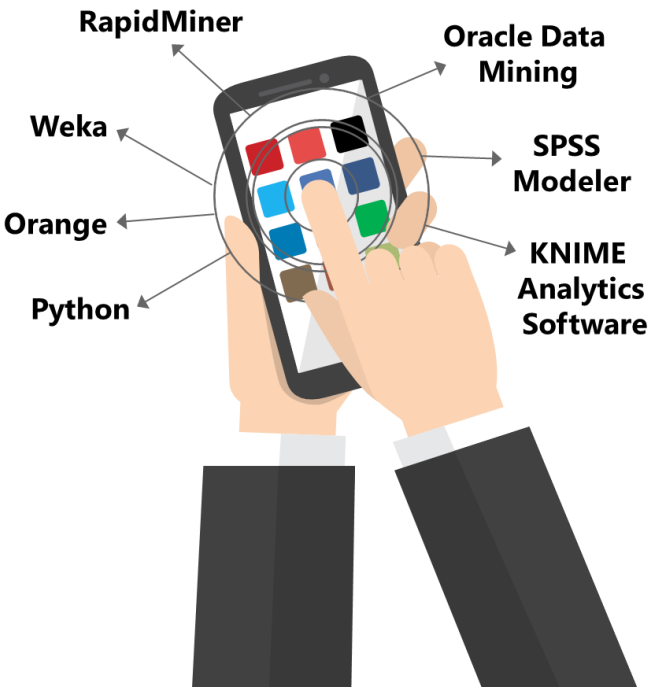
By the end of this topic, you should be able to:

- explore some data mining tools .
- utilise the data mining tools in solving data science problems.



Data Mining Tools

- Data mining serves the primary purpose of discovering patterns among large volumes of data and transforming data into more refined and actionable information.
- This technique utilizes specific algorithms, statistical analysis, artificial intelligence and database systems to juice out the information from huge datasets and convert them into an understandable form.
- There are many useful tools available for data mining. This lists out 7 comprehensive data mining tools widely used in the big data industry are given in the next pages.



RapidMiner

- Rapid Miner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis.
- It is one of the apex leading open source system for data mining. The program is written entirely in Java programming language.
- The program provides an option to try around with a huge number of arbitrarily nestable operators which are detailed in XML files and are made with graphical user interference of rapid miner.

Oracle Data Mining

- It is a representative of the Oracle's Advanced Analytics Database.
- Market leading companies use it to maximize the potential of their data to make accurate predictions.
- The system works with a powerful data algorithm to target best customers.
- Also, it identifies both anomalies and cross-selling opportunities and enables users to apply a different predictive model based on their need.
- Further, it customizes customer profiles in desired way.

IBM SPSS Modeler

- When it comes to large-scale projects IBM SPSS Modeler turns out to be the best fit.
- In this modeler, text analytics and its state-of-the-art visual interface prove to be extremely valuable.
- It helps to generate data mining algorithms with minimal or no programming.
- It can be widely used in anomaly detection, Bayesian networks, CARMA, Cox regression and basic neural networks that use multilayer perceptron with back-propagation learning.

KNIME Analytics Software

- Konstanz Information Miner is an open source data analysis platform.
- In this, you can deploy, scale and familiarize data in short time. In the business intelligent world, KNIME is known as the platform that helps to make predictive intelligence accessible to inexperienced users.
- Moreover, the data-driven innovation system helps uncover data potential.
- Also, it includes more than thousands of modules and ready-to-use examples and an array of integrated tools and algorithms.

Python

- Python holds an important at the top tools for Data Science and is often the go-to choice for a range of tasks for domains such as Machine Learning, Deep Learning, Artificial Intelligence, and more. It is object-oriented, easy to use and extremely developer-friendly thanks to its high code readability.
- Python's vast ecosystem of rich libraries and implementation for various purposes makes it a genuinely multi-faceted option. It also support for powerful Data Science libraries such as Keras, Scikit-Learn, matplotlib, TensorFlow and etc. It's perfectly suited for tasks like data collection, analysis, modelling, and visualization.

Orange

- Orange is an open source data visualization, machine learning and data mining toolkit.
- It features a visual programming front-end for exploratory data analysis and interactive data visualization. Orange is a component-based visual programming software package for data visualization, machine learning, data mining and data analysis.
- Orange components are called widgets and they range from simple data visualization, subset selection and pre-processing, to evaluation of learning algorithms and predictive modelling.
- Visual programming in orange is performed through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration.

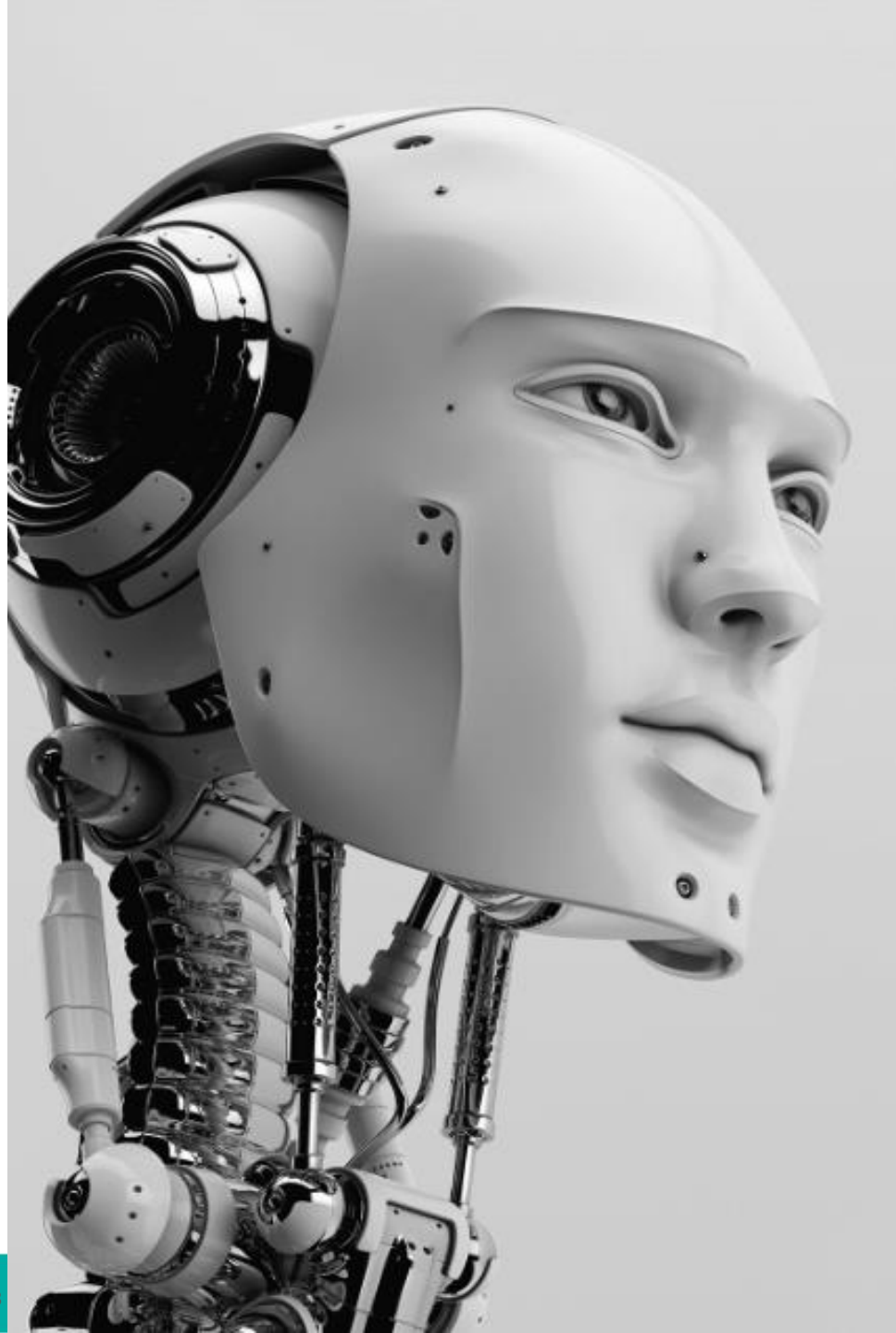
Weka

- Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software developed at the University of Waikato, New Zealand.
- The program is written in Java. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling coupled with graphical user interface.
- Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection.

Chapter 1.6:
Data Mining for Ethics

By the end of this topic, you should be able to:

- concern of some ethics in data mining practices.



Data Mining Ethics

What are Ethics?

Ethics is a code of behaviour that represents what is right and wrong.

- They have shared values / societal rules.
- Ethics is not law.
- No philosophical questions, but the ethical practice of data science.

Data ethics encompass the following:

i. data handling: generation, recording, curation, processing, dissemination, sharing, and use.

ii. algorithms: artificial intelligence (AI), artificial agents, machine learning, and robots.

Ethical issues arise in practical applications

a) Anonymizing data is difficult.

Example: 85% of Americans can be identified from just zip code, birth date and sex.

b) Data mining often used to discriminate.

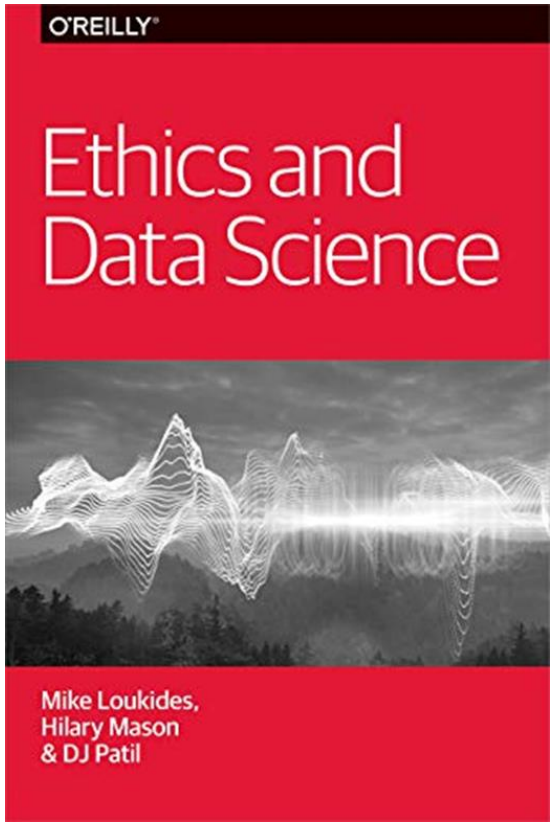
Example: loan applications using some information (e.g.,sex, religion, race) is unethical.

c) Ethical situation depends on applications.

Example: same information is okay to be used in medical application.

d) Input output/ attributes may contain problematic information.

Example: area code may correlate with race.



Links to Read

Data Mining in Brief

https://towardsdatascience.com/data-mining-in-brief-26483437f178

Data Mining Tools

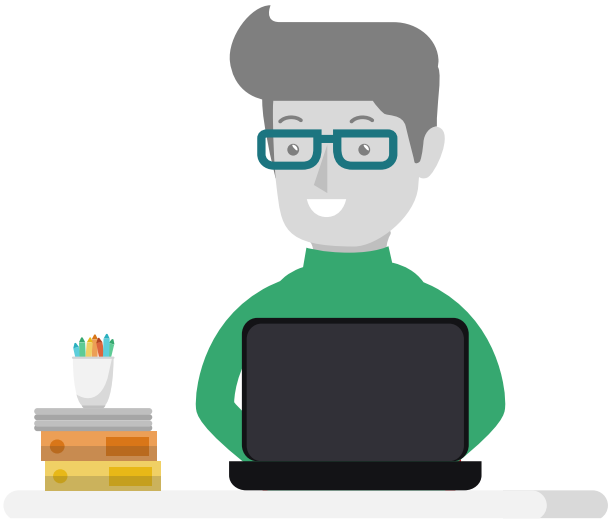
https://towardsdatascience.com/data-mining-tools-f701645e0f4c

Data Mining (Introduction to Data Mining and Tools)

https://towardsdatascience.com/data-mining-bc7feca95887

Top Influencers in Data Mining to Follow

https://bigdata-madesimple.com/top-influencers-in-data-mining-to-follow/



Links to Read

5 Data Mining Techniques Businesses Need To Know About

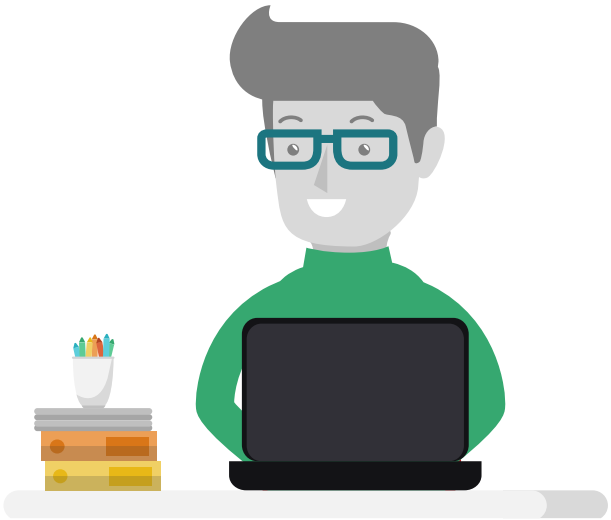
https://towardsdatascience.com/5-data-mining-techniques-businesses-need-to-know-about-20fd723800b2

12 Most Useful Data Mining Applications of 2021

https://www.upgrad.com/blog/12-most-useful-data-mining-applications-of-2020/

The Balancing Act of Data Mining Ethics: The Challenges of Ethical Data Mining

https://www.information-age.com/data-mining-123481736/#:~:text=Data%20mining%20ethics%3A%20the%20responsibility,ethical%20and%20transparent%20as%20well



Links to Read

Data Mining Techniques: Algorithm, Methods & Top Data Mining Tools

https://www.softwaretestinghelp.com/data-mining-techniques/#:~:text=These%20techniques%20are%20basically%20in%20the%20form%20of,techniques%20like%20Statistical%2C%20Visual%20and%20Audio%20data%20mining.

Data Mining Vs. Machine Learning: The Key Difference

https://www.simplilearn.com/data-mining-vs-machine-learning-article