

# EEE356 - Data Analytics (R)

## Week 2: Introduction to Data Analytics



**ADANA ALPARSLAN TÜRKES**  
**SCIENCE AND TECHNOLOGY UNIVERSITY**

Dr Kasım Zor

Department of Electrical and Electronic Engineering

Spring 2024

# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics
- 4 Types of Data Analytics
- 5 Data Scientist



# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics
- 4 Types of Data Analytics
- 5 Data Scientist



# Data

**Datum:** A piece of information or a fixed starting point of a scale or operation. (Singular) [Ref: Google]

## Data:

- Information, especially facts or numbers, collected to be examined and considered and used to help with making decisions (Plural) [Ref: Cambridge English Dictionary]
- Data, in the information age, are a large set of bits encoding numbers, texts, images, sounds, videos, and so on. Unless we add information to data, they are meaningless. When we add information, giving a meaning to them, these data become knowledge [1].



# Data Analytics

**Data Science:** Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge [2].

**Data Analytics:** The science that analyse crude (raw) data to extract useful knowledge (patterns or insights) from them. This process can also include data collection, organisation, preprocessing, transformation, modelling and interpretation [1].



# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics
- 4 Types of Data Analytics
- 5 Data Scientist



# A Short Taxonomy of Data Analytics – Part 1 [1]

**Instance or Object:** Examples of the concept that will be characterised.

Contact Person	Phone Number	E-Mail Address
Person 1	Number 1	Address 1
Person 2	Number 2	Address 2
Person 3	Number 3	Address 3

**Table 1:** Data set of a contact list

**Attribute or Feature:** Attributes, also called features, are characteristics of the instances.



## A Short Taxonomy of Data Analytics – Part 2 [1]

**Descriptive Analytics:** Summarise or condense data to extract patterns. In descriptive analytics tasks, the result of a given method or technique is obtained directly by applying an algorithm to the data. The result can be a statistic, such as an average, a plot, or a set of groups with similar instances.

**Predictive Analytics:** Extract models from data to be used for future predictions.





## A Short Taxonomy of Data Analytics – Part 3 [1]

**Method or Technique:** A method or technique is a systematic procedure that allows us to achieve an intended goal. A method shows how to perform a given task. But in order to use a language closer to the language computers can understand, it is necessary to describe the method/technique through an algorithm.

**Algorithm:** An algorithm is a self-contained, step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method. They are self-contained in order to be easily translated to an arbitrary programming language.



# A Short Taxonomy of Data Analytics – Part 4 [1]

**Model:** A model in data analytics is a generalisation obtained from data that can be used afterwards to generate predictions for new given instances. It can be seen as a prototype that can be used to make predictions. Thus, model induction is a predictive task.



# A Short Taxonomy of Data Analytics – Part 5 [1]

**Hyperparameters and Parameters:** Assume that in the induction of a model, there are both hyperparameters and parameters whose values are set. The values of the hyperparameters are set by the user, or some external optimisation method. The parameter values, on the other hand, are model parameters whose values are set by a modelling or learning algorithm in its internal procedure. When the distinction is not clear, the term parameter is used.



# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics**
- 4 Types of Data Analytics
- 5 Data Scientist



# KDD Process (Academia) [1]

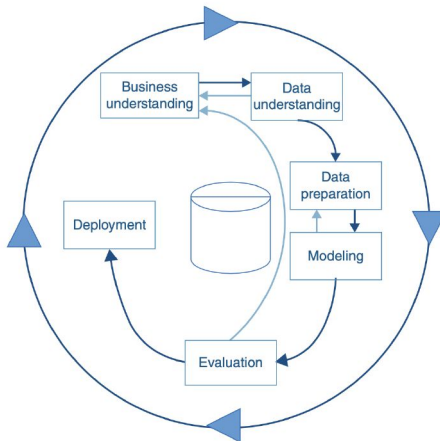
## Knowledge Discovery in Databases

- 1 Learning the application domain
- 2 Creating a target dataset
- 3 Data cleaning and preprocessing
- 4 Data reduction and projection
- 5 Choosing the data mining function
- 6 Choosing the data mining algorithm(s)
- 7 Data mining
- 8 Interpretation
- 9 Using discovered knowledge



# CRISP-DM Methodology (Industry) [1]

## CRoss-Industry Standard Process for Data Mining




# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics
- 4 Types of Data Analytics**
- 5 Data Scientist



# Types of Data Analytics [3]

## Descriptive, Predictive, and Prescriptive Analytics

	Descriptive	Predictive	Prescriptive
	What <b>HAS</b> happened?	What <b>COULD</b> happen?	What <b>SHOULD</b> happen?
What the user needs to <b>DO</b>	<ul style="list-style-type: none"> <li>• <b>Increase</b> asset reliability</li> <li>• <b>Reduce</b> labor and inventory costs</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Predict</b> infrastructure failures</li> <li>• <b>Forecast</b> facilities space demands</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Increase</b> asset utilization</li> <li>• <b>Optimize</b> resource schedules</li> </ul>
What the user needs to <b>KNOW</b>	<ul style="list-style-type: none"> <li>• The <b>number and types</b> of asset failures</li> <li>• Why <b>maintenance costs</b> are high</li> <li>• The value of the <b>materials inventory</b></li> </ul>	<ul style="list-style-type: none"> <li>• How to <b>anticipate failures</b> for specific asset types</li> <li>• When to <b>consolidate underutilized</b> facilities</li> <li>• How to <b>determine costs</b> to improve service levels</li> </ul>	<ul style="list-style-type: none"> <li>• How to <b>increase</b> asset production</li> <li>• Where to <b>optimally route</b> service technicians</li> <li>• Which strategic facilities plan provides the <b>highest long-term utilization</b></li> </ul>
How analytics gets <b>ANSWERS</b>	<ul style="list-style-type: none"> <li>• <b>Standard reporting</b> - What happened?</li> <li>• <b>Query/drill down</b> - Where exactly is the problem?</li> <li>• <b>Ad hoc reporting</b> - How many, how often, where?</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Predictive modeling</b> - What will happen next?</li> <li>• <b>Forecasting</b> - What if these trends continue?</li> <li>• <b>Simulation</b> - What could happen?</li> <li>• <b>Alerts</b> - What actions are needed?</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Optimization</b> - What is the best possible outcome?</li> <li>• <b>Random variable optimization</b> - What is the best outcome given the variability in specified areas?</li> </ul>
What makes this analysis <b>POSSIBLE</b>	<ul style="list-style-type: none"> <li>• Alerts, reports, dashboards, <b>business intelligence</b></li> </ul>	<ul style="list-style-type: none"> <li>• Predictive <b>models</b>, forecasts, statistical analysis, scoring</li> </ul>	<ul style="list-style-type: none"> <li>• Business rules, organization <b>models</b>, comparisons, <b>optimization</b></li> </ul>
			





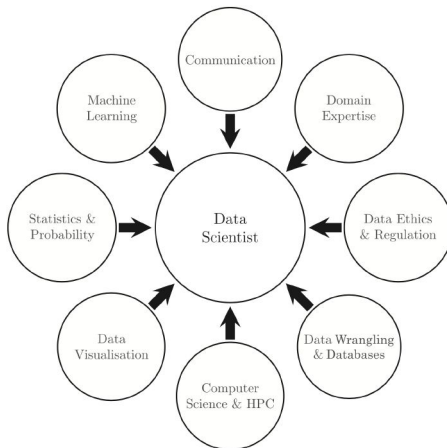
# Outline

- 1 Introduction
- 2 A Short Taxonomy of Data Analytics
- 3 Methodologies for Data Analytics
- 4 Types of Data Analytics
- 5 Data Scientist**



# Data Scientist [4]

## A Skills-Set Desideratum for a Data Scientist



# Data Scientist [5]

## Who is a Data Scientist?

### MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st 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 Hadoop'ing
- ☆ 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



# References I

- [1] Joao Mendes Moreira, Andre C. P. L. F. de Carvalho, and Tomas Horvath. *A General Introduction to Data Analytics*. John Wiley & Sons, 2019. ISBN 978-1-119-29625-6.
- [2] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017. URL <https://r4ds.had.co.nz>.
- [3] IBM Watson. Descriptive, predictive, and prescriptive analytics, 2015. URL <https://gemba.nl/wp-content/uploads/2015/10/watsonbusinessvalue.png>.
- [4] John D. Kelleher and Brendan Tierney. *Data Science*. The MIT Press, 2018. ISBN 978-0-262-53543-4.
- [5] Marketing Distillery. Who is a data scientist?, 2018. URL [https://miro.medium.com/max/1040/1\\*T5GfsoZ-IWK3rcVkZ7R2bw.png](https://miro.medium.com/max/1040/1*T5GfsoZ-IWK3rcVkZ7R2bw.png).

