DATA ANALYTICS AND MACHINE LEARNING WITH R

INTRODUCTION TO DATA SCIENCE

LUIS GUSTAVO NARDIN
INTERNET TECHNOLOGY
BRANDENBURG UNIVERSITY OF TECHNOLOGY

HISTORICAL PERSPECTIVE

- The term **Data Science** to designate a new profession is relatively recent
- The concept, however, has a long history
- It can trace back to 1962 with the publication of **Future of Data Analysis** by John W. Tukey, which states

> **Data analysis** include, among other things: procedures to **analyzing** data, techniques to **interpreting** the results of such procedures, ways of **planning the gathering** of data to make its analysis easier, more precise or more accurate, and all the **machinery** and results of (mathematical) **statistics** which apply to analyzing data.
>
> (Turkey, 1962)

HISTORICAL PERSPECTIVE

- Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*. 26:4, 745-766. DOI: **10.1080/10618600.2017.1384734**
- Tukey, John W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*. 33:1, 1-67. DOI: **10.1214/aoms/1177704711**
- Press, G. (2013, May). A Very Short History of Data Science. *Forbes*. Retrieved from **https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/**
- [Data Science Festival]. (2017, June 21). *Sean Owen: What 50 Years of Data Science Leaves Out* [Video File]. Retrieved from **https://youtu.be/Vcvn09Vs5I4**

WHAT IS DATA SCIENCE?

HIGH LEVEL DEFINITION

"Data science is the science of data, or data science is the study of data."

(Cao, 2017, 2018)

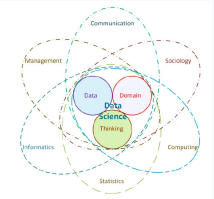
WHAT IS DATA SCIENCE?

TRANS-DISCIPLINARY DEFINITION

Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, such as statistics, informatics, computing, communication, management and sociology, to study data and its domain employing data science thinking.

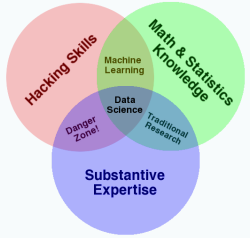
(Cao, 2017, 2018)

data science = def statistics ∩ informatics ∩ computing ∩ communication ∩ sociology ∩ management | data ∩ domain ∩ thinking



WHAT IS DATA SCIENCE?

TRANS-DISCIPLINARY DEFINITION



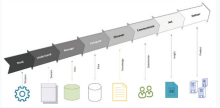
Source: Drew Conway

WHAT IS DATA SCIENCE?

PROCESS-BASED DEFINITION

From the DIKIW-processing perspective, data science is a systematic approach to *thinking with wisdom*, *understanding the domain*, *managing data*, *computing with data*, *discovering knowledge*, *communicating with stakeholders*, *acting on insights*, and *delivering products*.

(Cao, 2018)

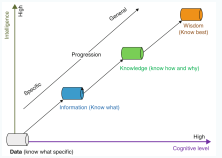


data science = def think ∩ understand ∩ manage ∩ compute ∩ discover ∩ communicate ∩ act ∩ deliver | DIKIW

FROM DATA TO WISDOM

- **Data** are symbols that represent the properties of objects and events
- **Information** consists of processed data (i.e. more compact) that increases its usefulness
- **Knowledge** concerns to how-to questions and answer to why questions
- **Wisdom** is the intelligence to know best about hot to act on the basis of knowledge

(Ackoff, 1989; Cao, 2018)



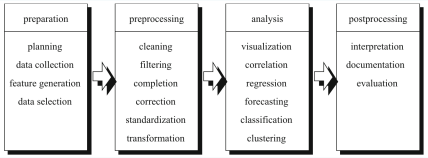
(Cao, 2018)

WHAT IS DATA?

Data are symbols that represent the properties of objects and events

- **Structured data** uses a particular organizational criteria like industrial process data and business data
  - Pre-defined data model
  - Easy to search
- **Unstructured data** does not have a predefined structured like text, image and video
  - No pre-defined data model
  - Difficult to search
- **Semi-structured data** contains semantic tags, but does not conform to an specific pre-defined data model.
  - No pre-defined data model
  - Data annotated with semantic tags
  - Easier to search than unstructured data

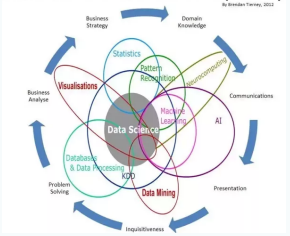
PHASES OF DATA ANALYSIS PROJECTS



(Ruckler, 2016)

DATA SCIENCE TECHNIQUES

Data Science Is Multidisciplinary



Source: Quora

DATA ANALYTICS

Data Analytics refers to the theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insight into data. Data analytics consists of

- **Descriptive** analytics typically uses statistics to describe the data used to gain information, or for other useful purposes.
- **Predictive** analytics makes predictions about unknown future events and discloses the reasons behind them, typically by advanced analytics.
- **Prescriptive** analytics optimizes indications and recommends actions for smart decision-making.

(Cao, 2017)

MACHINE LEARNING

Machine learning is a field of artificial intelligence in computer science that uses statistical techniques and algorithms to give computer systems the ability to "learn" (e.g. progressively improve performance on a specific task) from data, without being explicitly programmed

(Koza et al., 1996)

MACHINE LEARNING

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

(Mitchell, 1997)

MACHINE LEARNING

- **Unsupervised Learning**
  Finding patterns and relationship in data sets without any prior knowledge of the system.
- **Supervised Learning**
  We know the answer to a problem, and let the computer deduce the logic behind it.
- **Reinforcement Learning**
  Learning is achieved by trial-and-error, solely from rewards and punishment.

MACHINE LEARNING

- **Classification**
  Outputs are divided into two or more classes. The goal is to produce a model that assigns inputs into one of these classes.
- **Regression**
  Finding a functional relationship between an input data set and a reference data set. The goal is to construct a function that maps input data to continuous output values.
- **Clustering**
  Data are divided into groups with certain common traits, without knowing the different groups beforehand.

REFERENCES

- Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis*. 16, 3-9.
- Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Survey*. 50:3, Article 43. DOI: **10.1145/3076253**.
- Cao, L. (2018). *Data Science Thinking: The Next Scientific, Technological and Economic Revolution*. Springer: Cham. DOI: **10.1007/978-3-319-95092-1**
- Koza, J. R., Bennett, F. H., Andre, D., Keane, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero J.S., Sudweeks F. (eds) *Artificial Intelligence in Design'96*. Springer: Dordrecht.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Runkler, T. A. (2012). *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Springer Vieweg: Wiwsbaden. DOI: **10.1007/978-3-8348-2589-6**
- Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*. 3:3, 210–229. DOI: **10.1147/rd.33.0210**.