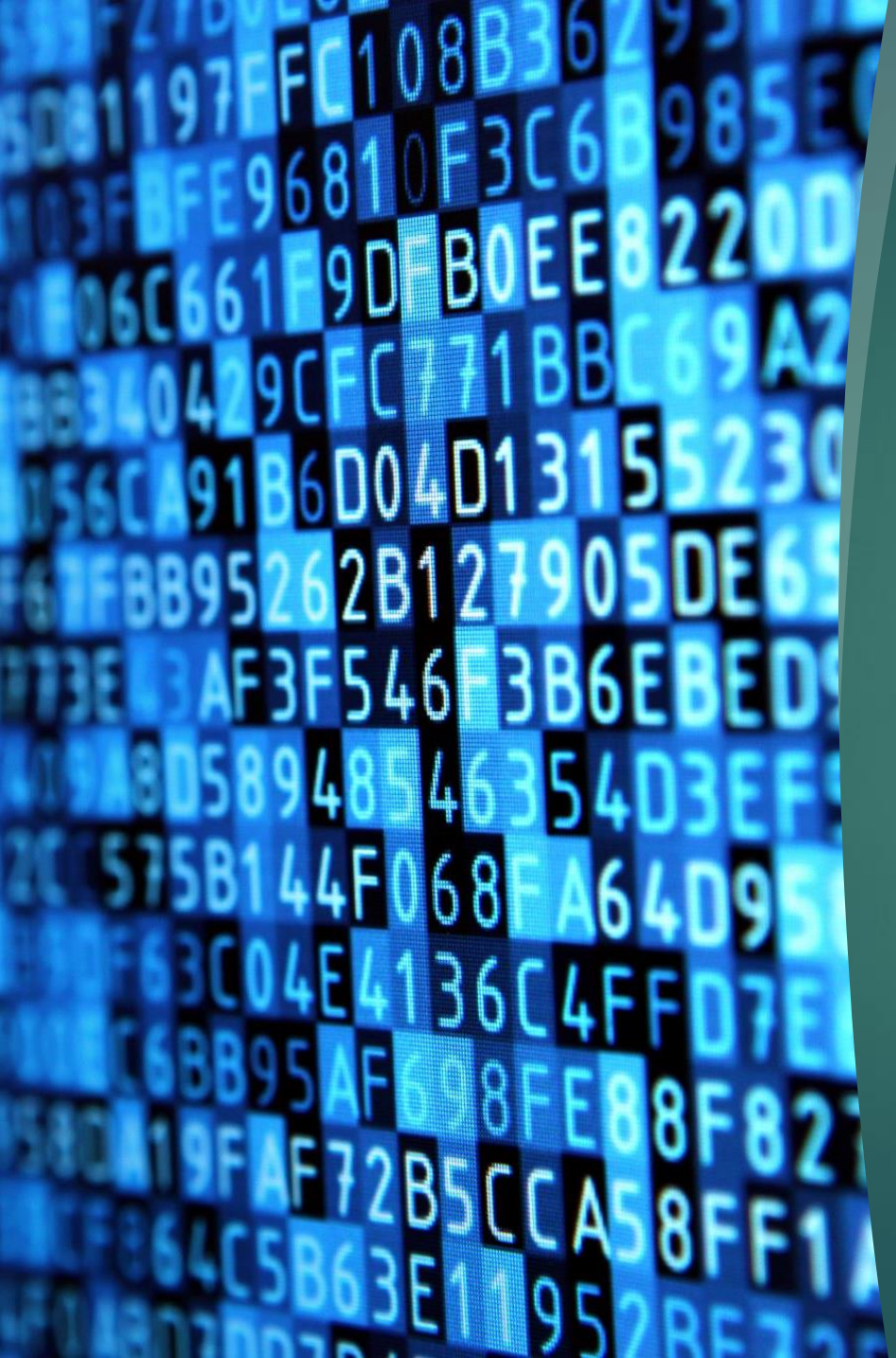

Introduction to Data Science and Machine Learning

COMP-1702

Module 2: Definitions

Module 2: Definitions

Learning Outcomes

- By the end of this module, you should be to:
  - Define data science and machine learning industry terminology.
- What are you going to learn in this module?
  - Common terms used in data science and machine learning.
- Why are you going to learn this?
  - To become familiar with terms and definitions used in industry.

Module 2: Definitions

Data Science

- An interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of applications domains.
- Data science is related to data mining, machine learning and big data.
- Came from the field of statistics and data analysis, which slowly changed over time to incorporate computing.

https://en.wikipedia.org/wiki/Data_science

Module 2: Definitions

Algorithm

- In mathematics and computer science, an algorithm is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation.
- Algorithms are always unambiguous and are used as specifications for performing calculations, data processing, automated reasoning, and other tasks.
- Take 5 minutes and find some examples of algorithms to share with the class.

https://en.wikipedia.org/wiki/Algorithm

Module 2: Definitions

Machine Learning

- The study of computer algorithms that improve automatically through experience and by the use of data.
- It is seen as part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.
- Machine learning algorithms are used in a wide variety of applications, such as medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.
- Take 5 minutes and find some examples of machine learning to share with the class.

https://en.wikipedia.org/wiki/Machine_learning

Module 2: Definitions

Big Data

- A field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
- Often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value.
- Current usage of the term big data tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set.
- Take 5 minutes and find some examples of big data to share with the class.

https://en.wikipedia.org/wiki/Big_data

Module 2: Definitions

Data Mining

- The process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
- An interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.
- The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.

https://en.wikipedia.org/wiki/Data_mining

Module 2: Definitions

Supervised Learning

- The machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- Infers a function from labeled training data consisting of a set of training examples.
- Analyzes the training data and produces an inferred function, which can be used for mapping new examples.
- An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.
- Take 5 minutes and find some examples of supervised learning to share with the class.

https://en.wikipedia.org/wiki/Supervised_learning

Module 2: Definitions

Unsupervised Learning

- A type of algorithm that learns patterns from untagged data.
- In contrast to supervised learning where data is tagged by a human, unsupervised learning exhibits self-organization that captures patterns as neuronal predilections or probability densities.
- The other levels in the supervision spectrum are reinforcement learning where the machine is given only a numerical performance score as its guidance (AWS Deep Racer), and semi-supervised learning where a smaller portion of the data is tagged.
- Two broad methods are neural networks and probabilistic methods.
- Take 5 minutes and find some examples of unsupervised learning to share with the class.

https://en.wikipedia.org/wiki/Unsupervised_learning

Module 2: Definitions

Artificial Intelligence

- Intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality.
- "Strong AI" is usually labeled as artificial general intelligence (AGI) while attempts to emulate "natural" intelligence have been called artificial biological intelligence (ABI).
- Often used to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".
- Take 5 minutes and find some examples of AI to share with the class.

https://en.wikipedia.org/wiki/Artificial_intelligence

Module 2: Definitions

Analytics

- The systematic computational analysis of data or statistics.
- Used for the discovery, interpretation, and communication of meaningful patterns in data.
- Also entails applying data patterns towards effective decision-making.
- Can be valuable in areas rich with recorded information.
- Relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

https://en.wikipedia.org/wiki/Analytics

Module 2: Definitions

Data Wrangling

- Sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
- Typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

https://en.wikipedia.org/wiki/Data_wrangling

Module 2: Definitions

Data Cleansing/Cleaning

- The process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
- May be performed interactively with data wrangling tools, or as a batch process through scripting.

https://en.wikipedia.org/wiki/Data_cleansing

Module 2: Definitions

Data Standardization

- The critical process of bringing data into a common format that allows for collaborative research, large scale analytics, and sharing of sophisticated tools and methodologies.

https://www.ohdsi.org/data-standardization/

Module 2: Definitions

KPI

- Key Performance Indicators (KPI's) are the critical (key) indicators of progress toward an intended result.
- KPI's provides a focus for strategic and operational improvement, create an analytical basis for decision making and help focus attention on what matters most.
- Managing with the use of KPI's includes setting targets (the desired level of performance) and tracking progress against that target.
- Managing with KPI's often means working to improve leading indicators that will later drive lagging benefits.
- Leading indicators are precursors of future success, lagging indicators show how successful the organization was at achieving results in the past.

https://kpi.org/KPI-Basics

Module 2: Definitions

Assignment

- For this module, there will be a short quiz to test your understanding of the definitions we just covered.
- You can find the quiz in Module 2 in Learn.