Ministry of Higher Education and Scientific Research
Djilali BOUNAAMA University - Khemis Miliana(UDBKM)
Faculty of Science and Technology
Department of Mathematics and Computer Science

Chapter 1

Introduction to Data Science

AIBD-M1-UEM112 : Introduction to Data Science

Noureddine AZZOUZA

n.azzouza@univ-dbkm.dz

Course Topics

1. **Definition**

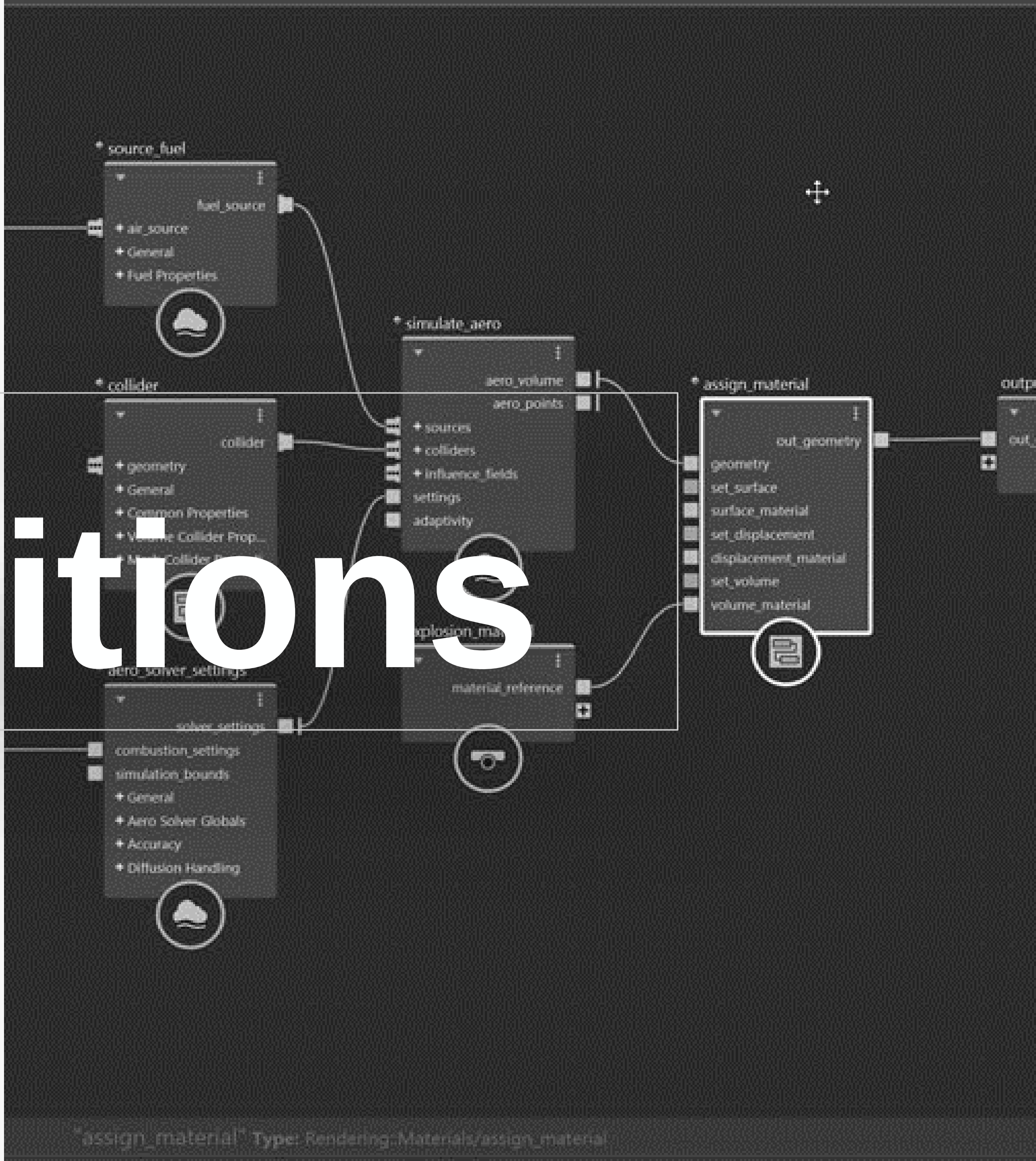
2. **Objectives**

3. **Origin**

4. **References**

3. **Data Science**

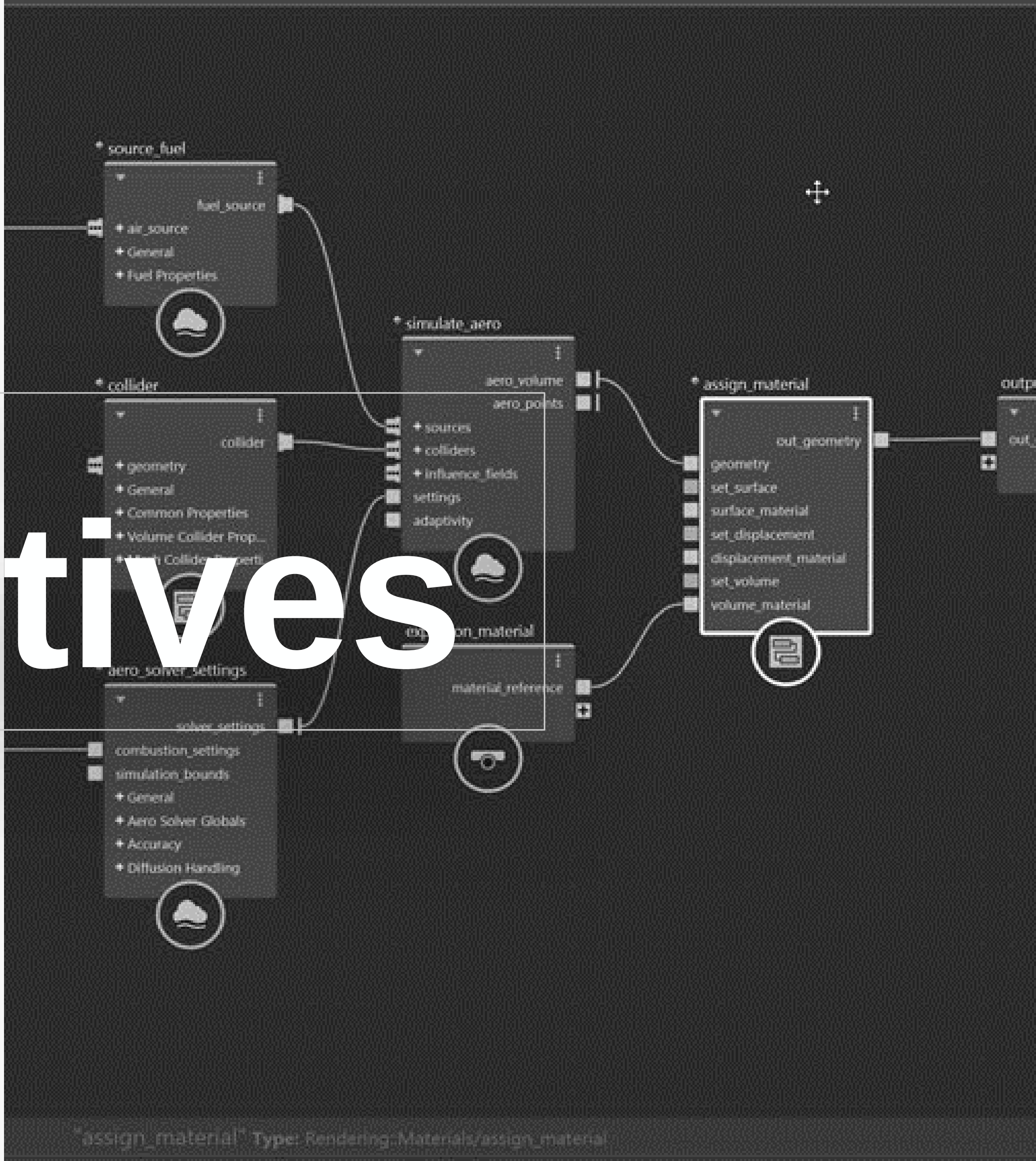
Definitions



Definitions

- ✓ ***Data :*** Data are raw symbols that represent the properties of objects and events and in this sense data has no meaning of itself, it simply exists (Russell L. Acko, 1989).
  - ***Example***: “John”, “Smith”, 30000
- ✓ ***Information :*** Data + meaning.
  - ***Example***: (first name, “John”), (last name, “Smith”), (salary, 30000)
- ✓ ***Knowledge :*** Data + meaning + context → lead to a decision
  - ***Example***: John Smith’s salary will be fixed to 30000 starting from today

Objectives



"assign_material" Type: Rendering::Materials/assign_material

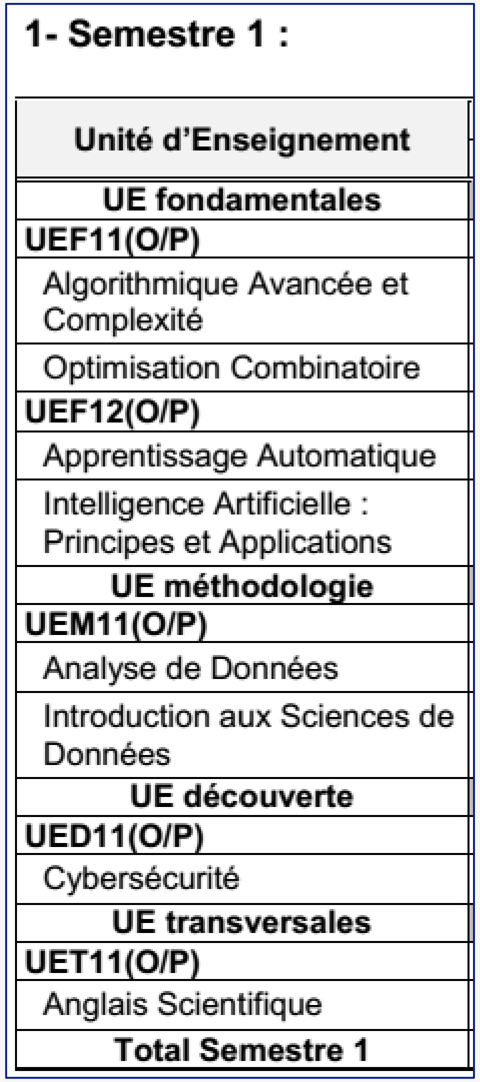
Objectives

- ✓ This module is introductory
- ✓ aims to familiarize students with the concepts relating to Data Science
- ✓ understanding its usefulness through examples.

« ***Know how to manipulate data from its creation to visualization and sharing.*** »

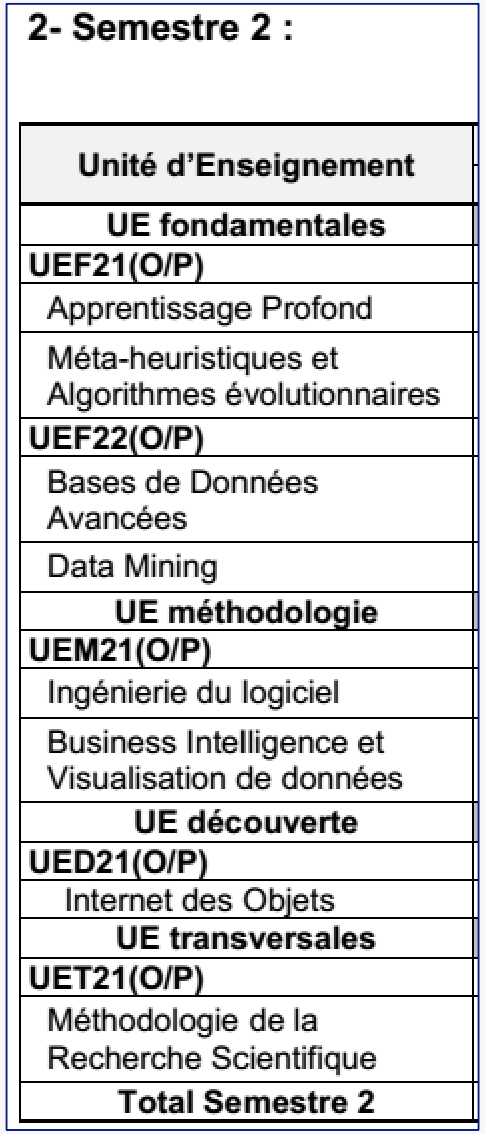
Introduction

Curriculum



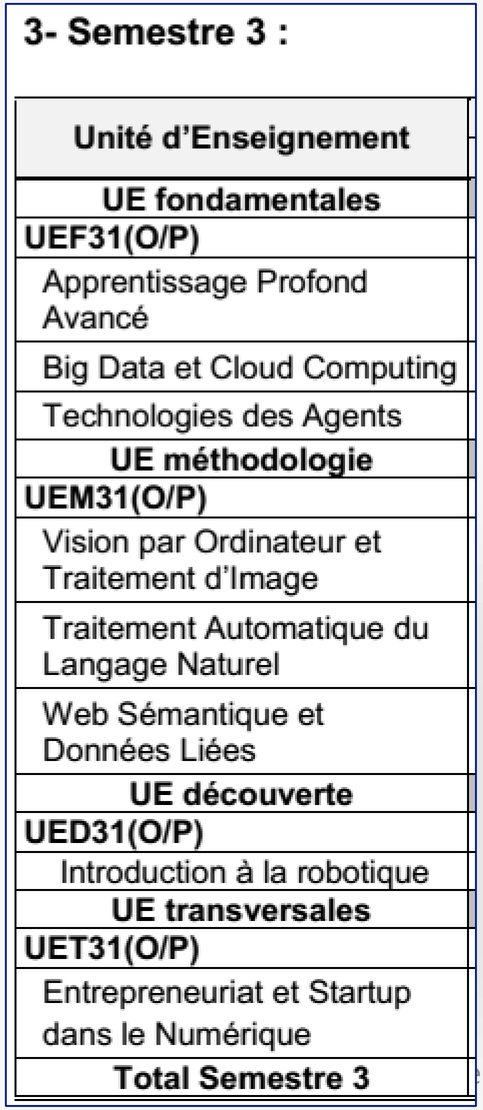
1- Semestre 1 :

Unité d'Enseignement
UE fondamentales
UEF11(O/P)
Algorithmique Avancée et Complexité
Optimisation Combinatoire
UEF12(O/P)
Apprentissage Automatique
Intelligence Artificielle : Principes et Applications
UE méthodologie
UEM11(O/P)
Analyse de Données
Introduction aux Sciences de Données
UE découverte
UED11(O/P)
Cybersécurité
UE transversales
UET11(O/P)
Anglais Scientifique
Total Semestre 1



2- Semestre 2 :

Unité d'Enseignement
UE fondamentales
UEF21(O/P)
Apprentissage Profond
Méta-heuristiques et Algorithmes évolutionnaires
UEF22(O/P)
Bases de Données Avancées
Data Mining
UE méthodologie
UEM21(O/P)
Ingénierie du logiciel
Business Intelligence et Visualisation de données
UE découverte
UED21(O/P)
Internet des Objets
UE transversales
UET21(O/P)
Méthodologie de la Recherche Scientifique
Total Semestre 2



3- Semestre 3 :

Unité d'Enseignement
UE fondamentales
UEF31(O/P)
Apprentissage Profond Avancé
Big Data et Cloud Computing
Technologies des Agents
UE méthodologie
UEM31(O/P)
Vision par Ordinateur et Traitement d'Image
Traitement Automatique du Langage Naturel
Web Sémantique et Données Liées
UE découverte
UED31(O/P)
Introduction à la robotique
UE transversales
UET31(O/P)
Entrepreneuriat et Startup dans le Numérique
Total Semestre 3

Content of this course

1. **Chapter 1. Introduction to Data Science**
   - ➢ data types
   - ➢ The data science process
   - ➢ The big data ecosystem and data science
2. **Chapter 2. The Data Science Process**
   - ➢ Overview of the data science process
   - ➢ Step 1: Define research objectives and create a project charter
   - ➢ Step 2: Data recovery
   - ➢ Step 3: Clean, integrate and transform data
   - ➢ Step 4: Exploratory data analysis
   - ➢ Step 5: Build the models
   - ➢ Step 6: Presenting the results and building applications

Content of this course

3. **Chapter 3: Data science tools**
   - ➢ Data storage tools
   - ➢ Data preparation tools
   - ➢ Data visualization tools
   - ➢ IDE notebook tools
   - ➢ Complete Data science platforms
4. **Chapter 4: Data sources**
   - ➢ Existing data
   - ➢ APIs
   - ➢ Scrapping
   - ➢ The creation of new data

Content of this course

5. **Chapter 5: Data communication**
   - ➢ The interpretability of the data
   - ➢ Data exploitation
   - ➢ Data visualization
   - ➢ Integration with other solutions

References

& Books

References & Books

References & Books

Data science : fondamentaux et études de cas
Data Science and Big Data Analytics
DATA SCIENCE
Introduction to Data Science
Principles of Data Science
Packt

- Morand, Elisabeth. "Data science: fondamentaux et études de cas, Machine learning avec Python et R by Eric Biernat and Michel Lutz." *Population, English edition* 73.2 (2018): 386-387.

Dietrich, David. "Data Science & Big Data Analytics." (2015).

KELLEHER, JOHN D. "Data science/John D. Kelleher and Brendan Tierney. Description: Cambridge, MA: The MIT Press, 2018

Laura, Igual, and Seguí Santi. "Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications." (2017).

Ozdemir, Sinan. *Principles of data science*. Packt Publishing Ltd, 2016.

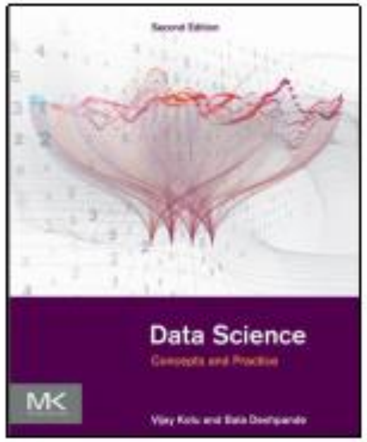
4

References & Books

References & Books

FUNDAMENTALS OF DATA SCIENCE
Sanjeev J. Wagh, Manisha S. Bhende, and Anuradha D. Thakare
A Chapman & Hall Book
CRC Press

Wagh, Sanjeev J., Manisha S. Bhende, and Anuradha D. Thakare. *Fundamentals of Data Science*. Chapman and Hall/CRC, 2021.

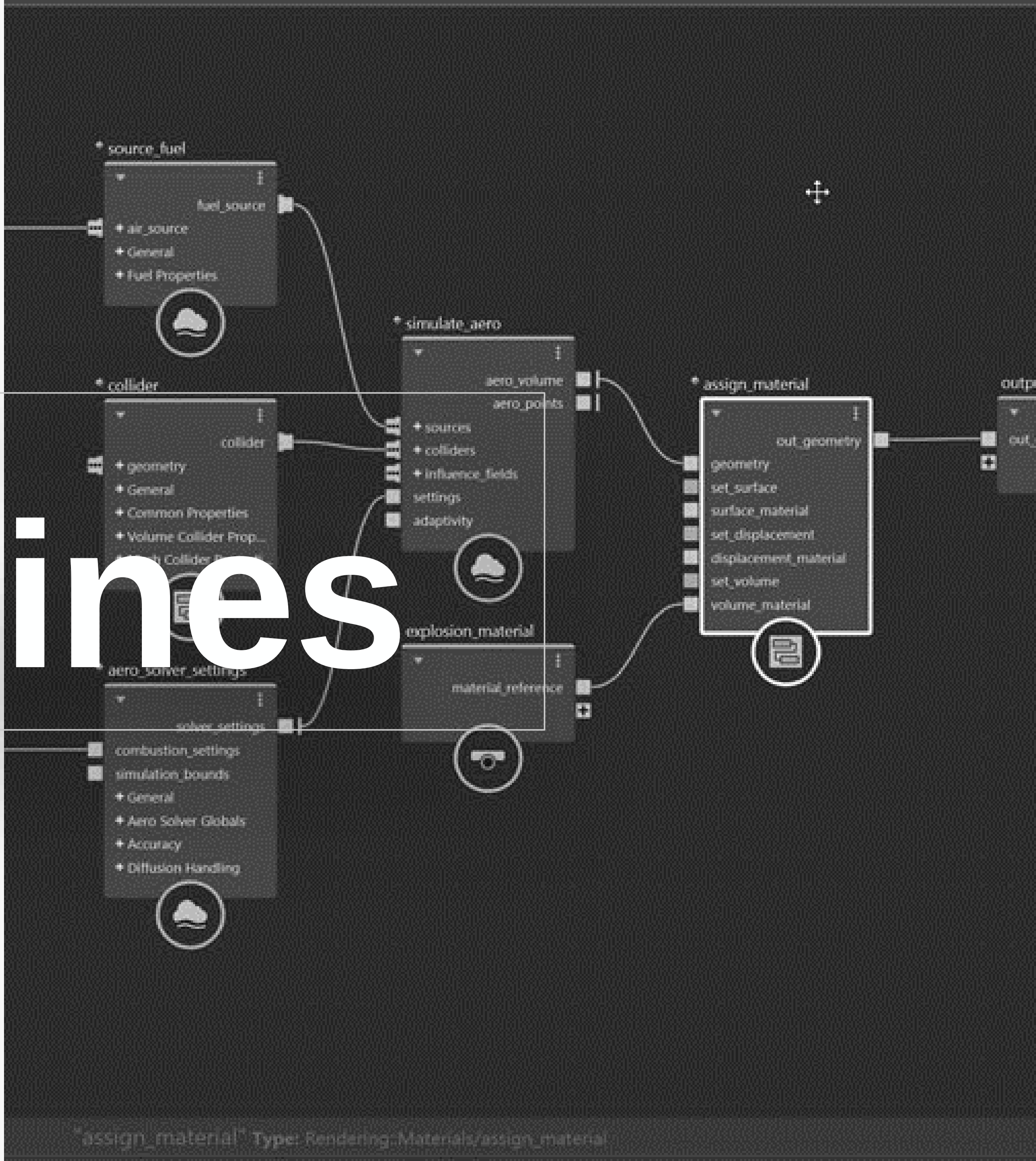


Kotu, Vijay, and Bala Deshpande. *Data science: concepts and practice*. Morgan Kaufmann, 2018.

Joel Grus
Data science par la pratique
Fondamentaux avec Python
EYROLLES
Préparez-vous aux métiers du futur !

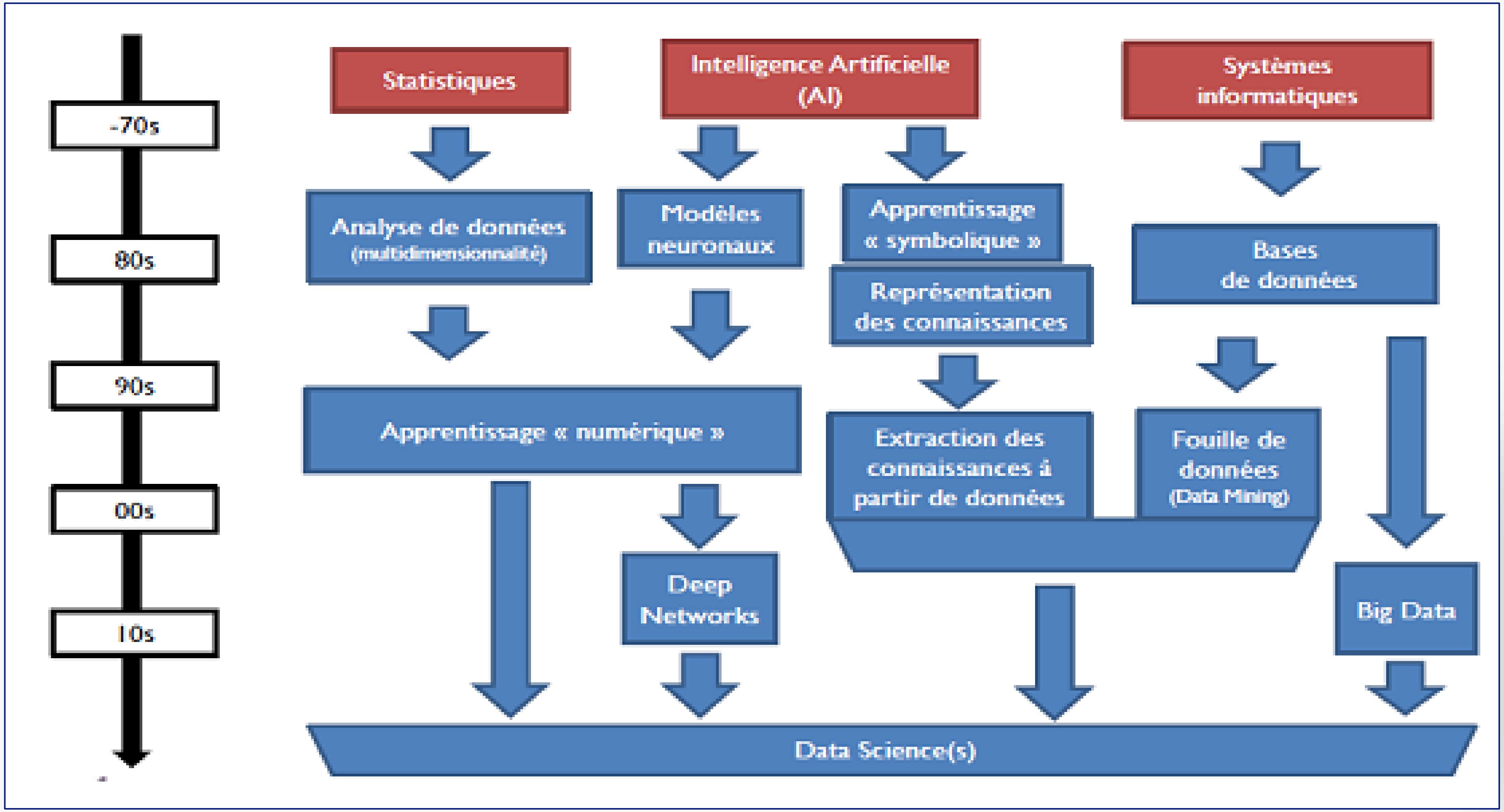
Grus, Joel. *Data science par la pratique: fondamentaux avec Python*. Eyrolles, 2020.

Origines

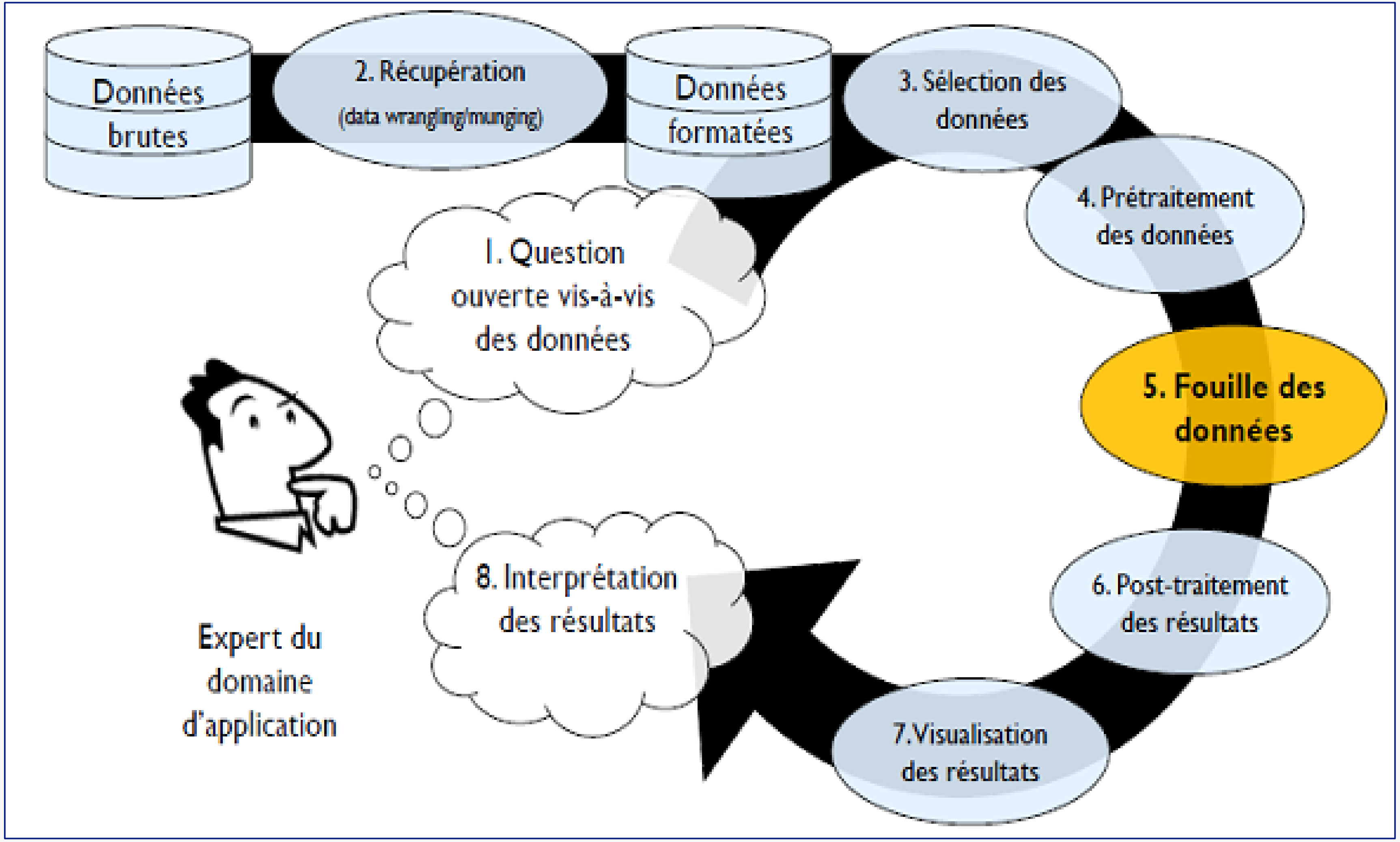


"assign_material" Type: Rendering::Materials/assign_material

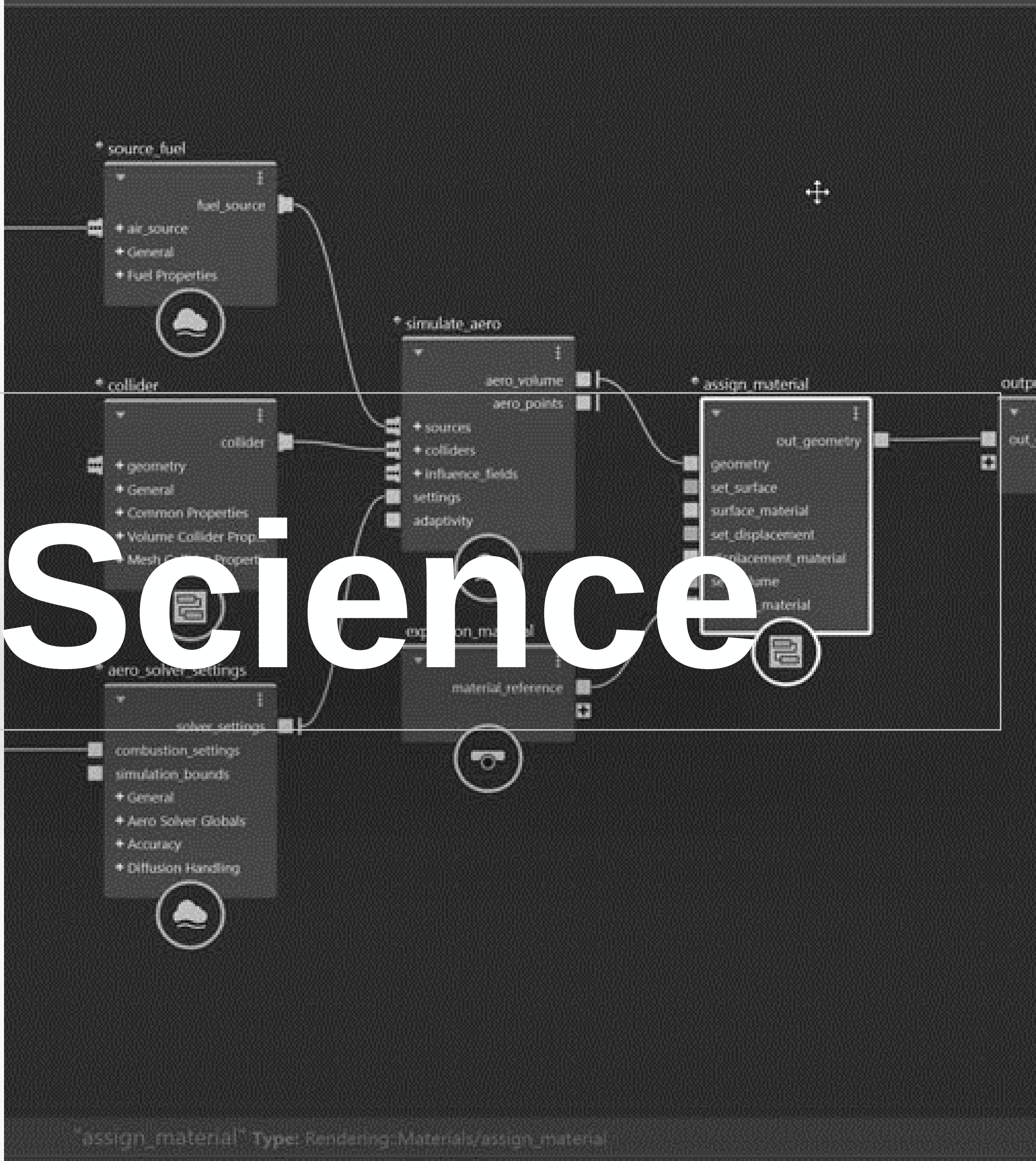
Origines of DS


-70s
80s
90s
00s
10s
Statistiques
Intelligence Artificielle (AI)
Systèmes informatiques
Analyse de données (multidimensionnalité)
Modèles neuronaux
Apprentissage « symbolique »
Représentation des connaissances
Bases de données
Apprentissage « numérique »
Extraction des connaissances à partir de données
Fouille de données (Data Mining)
Deep Networks
Big Data
Data Science(s)

Knowledge Discovery from Data : KDD


Données brutes
2. Récupération (data wrangling/munging)
Données formatées
3. Sélection des données
4. Prétraitement des données
5. Fouille des données
6. Post-traitement des résultats
7.Visualisation des résultats
8. Interprétation des résultats
1. Question ouverte vis-à-vis des données
Expert du domaine d'application

Data Science



Definition

- ✓ Data science is a compilation of techniques that extract value from data.
- ✓ Some of the techniques used in data science have a long history and trace their roots to applied statistics, machine learning, visualization, logic, and computer science

(Kotu, Vijay, and Bala Deshpande. Data science: concepts and practice. Morgan Kaufmann, 2018.)

Definition

- ✓ Data science is commonly defined as a methodology by which actionable insights can be inferred from data….
- ✓ Performing data science is a task with an ambitious objective: the production of beliefs informed by data and to be used as the basis of decision-making.

Laura, Igual, and Seguí Santi. "Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications." (2017).

Definition

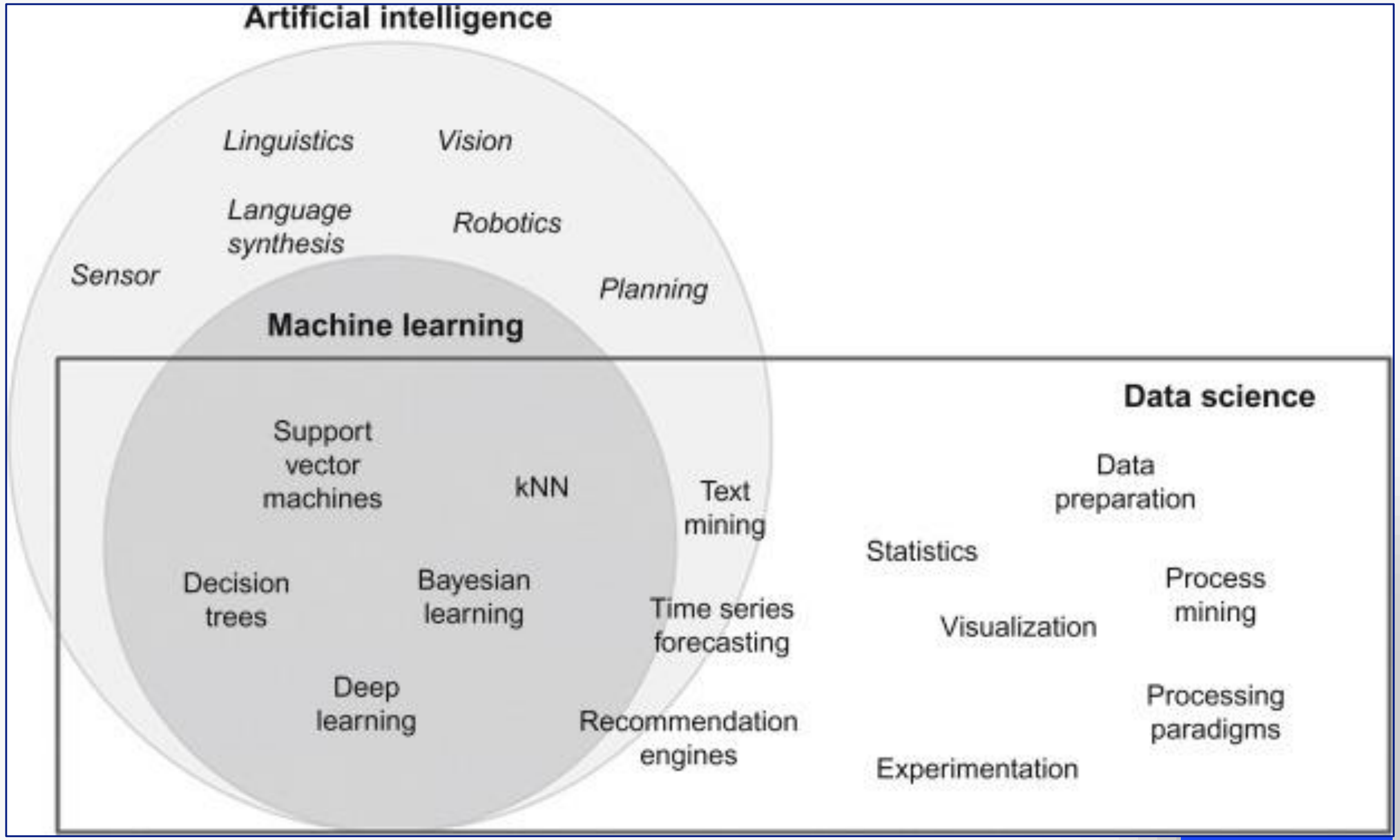
“Data science is an ***interdisciplinary*** *field, which borrows from* ***business****,* ***statistics*** *and* ***computer science*** *various methods, processes and algorithms to extract information from data”.*

New Wave

- ✓ Explosion in the quantity of data produced and collected (big data, etc.)
- ✓ More economical data storage
- ✓ Exponential increase in computing capabilities
- ✓ Improved accessibility to algorithms

Domains

Talking about ***Data Science***, means covering many possible areas of application. The most common areas are:



Artificial intelligence

✓ AI involves giving machines the ability to imitate human behavior, particularly cognitive functions.

➢ ***Examples***: facial recognition, automated driving, mail sorting based on postal code.

✓ In some cases, machines have far exceeded human capabilities

➢ ***Examples :***sorting thousands of postal items in a few seconds

✓ There is a whole range of techniques relating to AI :

➢ ***Examples :*** natural language processing, decision science, robotics, planning, etc.

Machine Learning

✓ Learning is an important part of human ability.

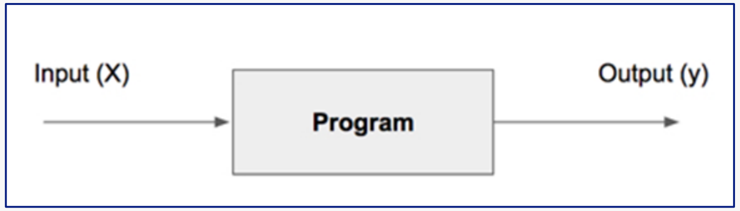
✓ Machine learning can be considered as a sub-field or one of the tools of AI, it provides machines with the ability to learn from experience

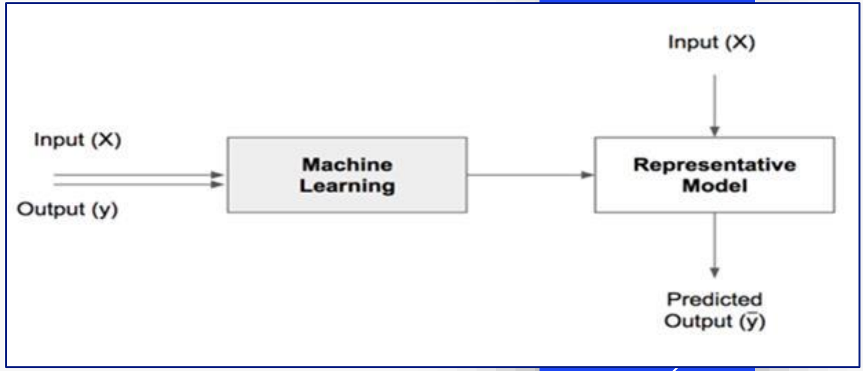
✓ Machine experience comes in the form of data.

✓ The data used to teach machines is called ***training data***.

Machine Learning

- ✓ Machine learning disrupts the traditional programming model:
  - ➢ A program transforms input signals into output signals using predetermined rules and relationships.
  - ➢ Machine learning algorithms, also called "learners", take both known input and output (training data) to determine a model for the program that converts the input to output.





Machine Learning

- ✓ Many organizations such as social media platforms, review sites or forums are required to moderate posts and remove abusive content.
  - ➢ *How can we teach machines to automate the removal of abusive content?*
- ✓ Machines should receive examples of abusive and non-abusive messages with a clear indication of which one is abusive.
- ✓ Learners will generalize a pattern based on certain words or sequences of words in order to conclude whether the overall message is abusive or not.
- ✓ The model can take the form of a set of "if – then - else" rules.
- ✓ Once the rules or data science model is developed, machines can begin to categorize the nature of any new messages.

Data Science

- ✓ Data science is the business application of ML, AI, and other quantitative fields such as statistics, visualization, and mathematics.
- ✓ It is an interdisciplinary field that extracts value from data.
- ✓ It relies heavily on ML and is sometimes called data mining.
  - ➢ ***Examples*** :
    - ✓ recommendation engines that can recommend movies for a particular user,
    - ✓ a fraud alert model that detects fraudulent credit card transactions,
    - ✓ find customers who will most likely unsubscribe next month or predict revenue for the next quarter.

Data Science

- ✓ Data science begins with data, which can range from a simple table of a few numerical observations to a complex matrix of millions of observations with thousands of variables.
- ✓ Data science uses certain specialized computational methods in order to discover meaningful and useful structures in a set of data.
- ✓ The discipline of data science coexists and is closely associated with a number of related fields such as database systems, data engineering, visualization, data analysis, and business intelligence. business (BI).

Data Science

Data Science

- ✓ Data science begins with data, which can range from a simple table of a few numerical observations to a complex matrix of millions of observations with thousands of variables.
- ✓ Data science uses certain specialized computational methods in order to discover meaningful and useful structures in a set of data.
- ✓ The discipline of data science coexists and is closely associated with a number of related fields such as database systems, data engineering, visualization, data analysis, and business intelligence. business (BI).

Characteristics and Motivations

✓ We can more precisely define data science by studying some of its main characteristics and motivations such as:

1. Extracting meaningful patterns
2. Building representative models
3. Combination of statistics, ML and computer science
4. Learning Algorithms
5. Other associated fields

Data Science

Characteristics and Motivations

1. Extracting Meaningful Patterns

- ➢ Data science involves the inference and iteration of many different hypotheses.
- ➢ One of the key aspects of DS is the process of generalizing patterns from a set of data.
- ➢ The generalization must be valid, not only for the dataset used to observe the pattern, but also for the new data
  - ➢ ***Examples*** : Je suis (adjective or noun); Je suis Algérien

    Il lit a (document); He reads a book

Characteristics and Motivations

2. Building representative models

- In statistics, a model describes how one or more variables in data relate to other variables.
- Modeling is a process in which a representative abstraction is constructed from the observed data set. Example 1: •
  - ***Examples*** : Human reads a (document), document = book, newspaper, magazine….

Characteristics and Motivations

3. Combination of statistics, ML and computer science

- In order to extract useful and relevant information, data science borrows computational techniques from the disciplines of statistics, ML, and database theories.
- The algorithms used in DS originate from these disciplines but have since evolved (parallel computing, scalable computing, etc.)
- One of the key ingredients of successful DS is solid prior knowledge about data and business processes that generate the data

Characteristics and Motivations

3. Combination of statistics, ML and computer science

- ➢ Data science also typically works on large data sets that need to be stored, processed, and calculated.
- ➢ This is where database techniques as well as parallel and distributed computing techniques play an important role in data science.

Characteristics and Motivations

4. Learning Algorithms

- Applying sophisticated learning algorithms to extract useful patterns from data differentiates DS from traditional data analysis techniques.
- Many of these algorithms have been developed over the past few decades and are part of ML and AI.
- Some algorithms are based on the foundations of **Bayesian probabilistic** theories and **regression analysis**, dating back hundreds of years. These iterative algorithms automate the process of finding an optimal solution for a given data problem.
- Data Science uses specific learning algorithms such as **decision trees**, **neural networks**, **k-nearest neighbors** (k-NN), and **k-means** clustering, among others

Characteristics and Motivations

5. Associated / Related fields

➢ While data science covers a wide range of techniques, applications, and disciplines, there are a few related areas that data science relies heavily on:

✓ ***Descriptive statistics***: mean calculation, standard deviation, correlation and other descriptive statistics help quantify the aggregate structure of a data set.

✓ ***Exploratory visualization***: The process of expressing data in visual coordinates allows users to find patterns and relationships in data and understand large data sets. • Business Intelligence: Helps organizations use data effectively. It allows querying the data without the need to

Characteristics and Motivations

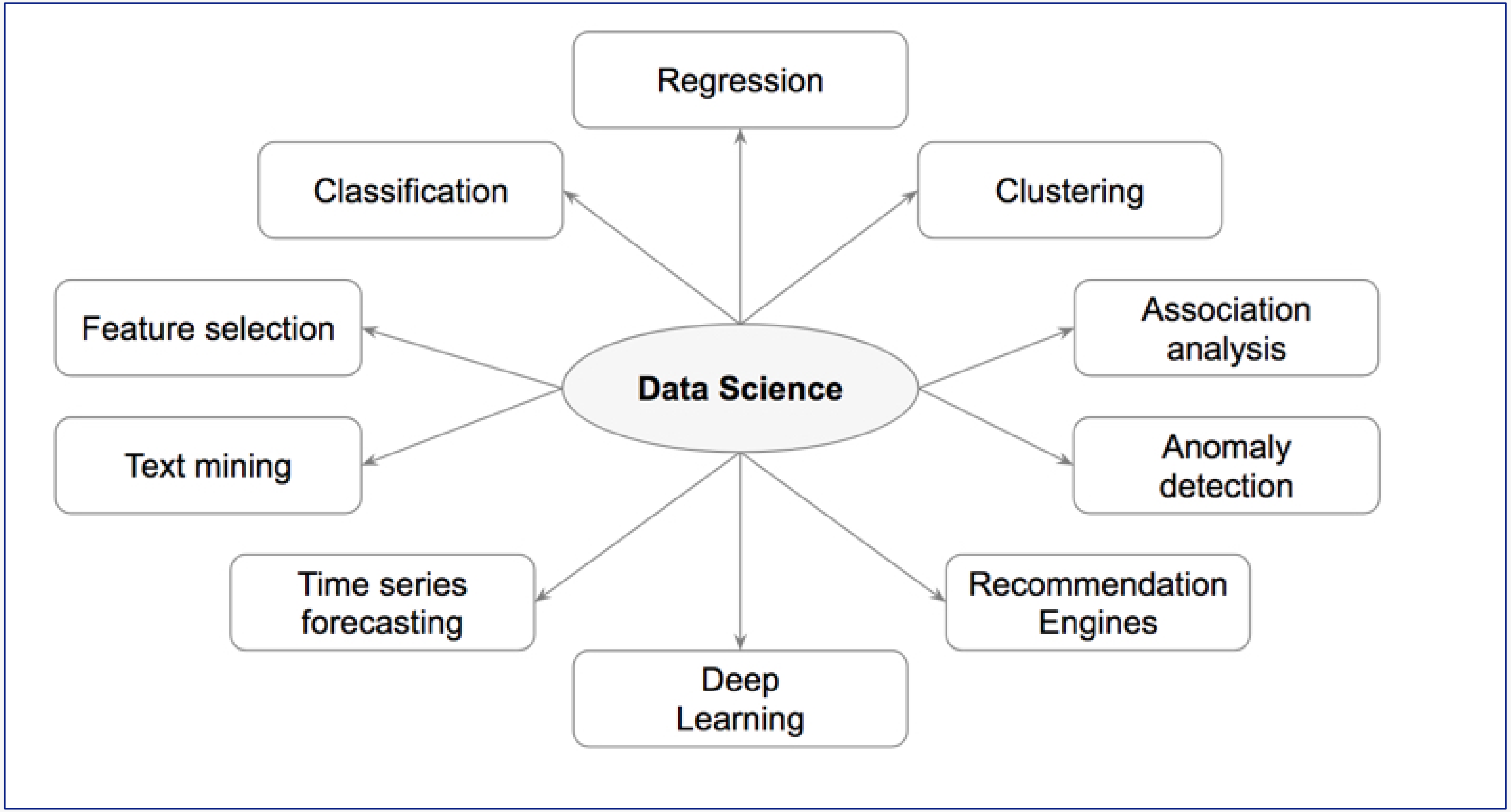
5. Associated / Related fields

➢ there are a few related areas that data science relies heavily on:

- ✓ ***Business Intelligence***: Helps organizations use data effectively. It allows querying the data without the need to write the technical query command,
- ✓ ***Data engineering***: it is the process of searching, organizing, assembling, storing and distributing data for analysis and effective use.

  Database engineering (e.g. Apache Hadoop, Spark, Kafka), parallel computing, data warehousing constitute data engineering techniques.

Types of Data Science



Ministry of Higher Education and Scientific Research
Djilali BOUNAAMA University - Khemis Miliana(UDBKM)
Faculty of Science and Technology
Department of Mathematics and Computer Science

Chapter 1

Introduction to Data Science

AIBD-M1-UEM112 : Introduction to Data Science

Noureddine AZZOUZA

n.azzouza@univ-dbkm.dz