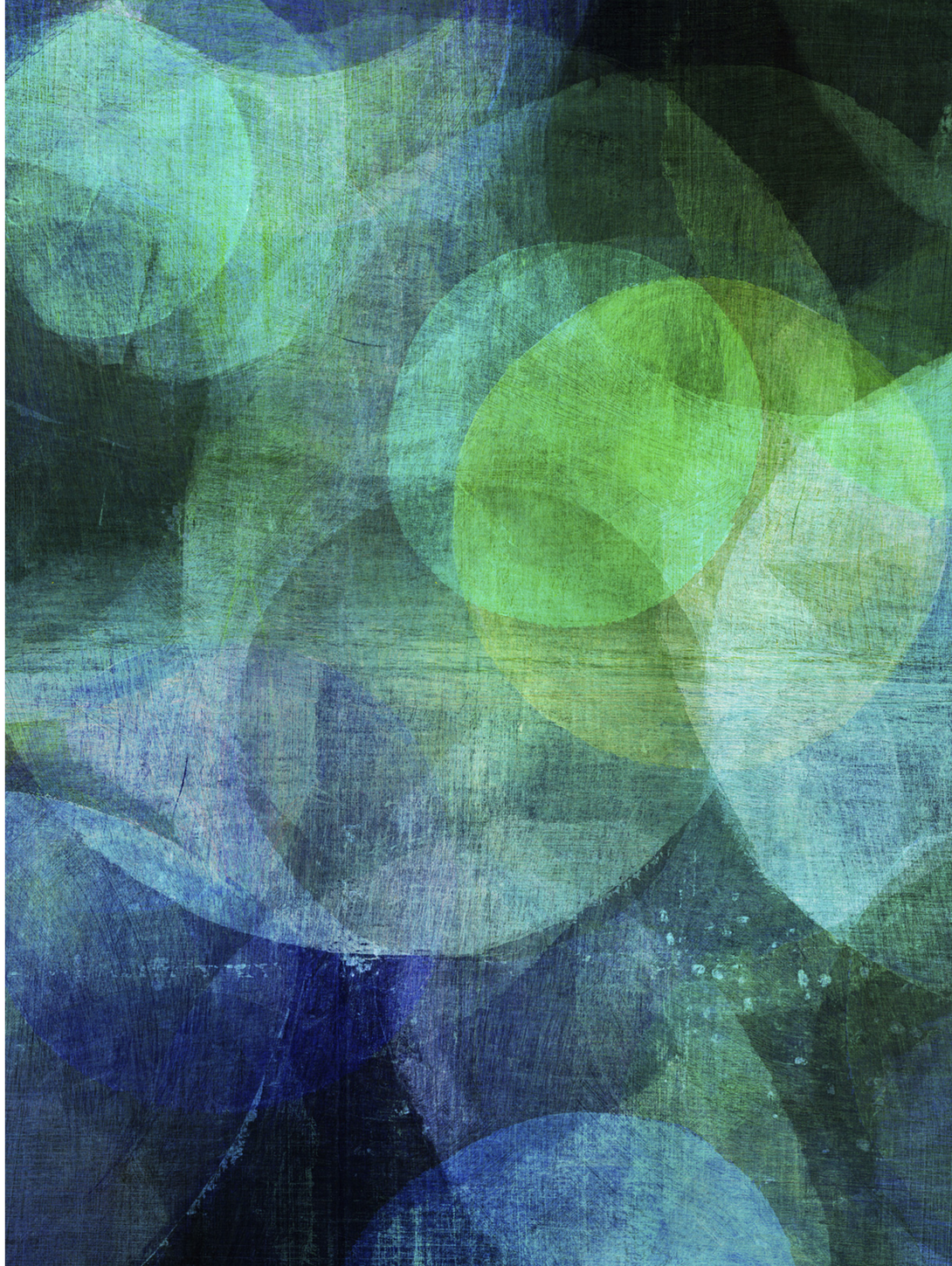
Bootcamp Data Science, Data Analytics & Machine Learning

CIÊNCIA DE DADOS - INTRODUÇÃO

Prof. Charles Prado



CIÊNCIA DE DADOS – INTRODUÇÃO

- Definição (estudo, modelos, arte …)
- Exemplos de aplicações
- Quem é o Cientista de Dados
- Conceitos (Big Data, Data Mining, AI, ML etc)
- Data Science in Business
- Tools for Data Science
- Python

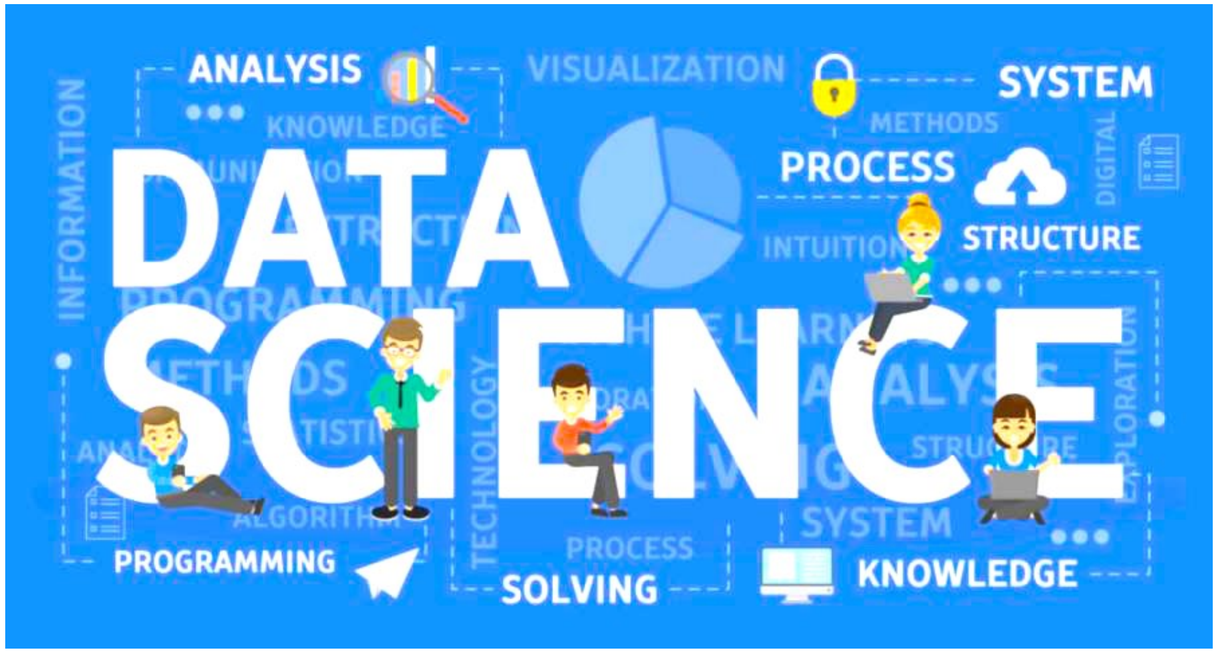
CIÊNCIA DE DADOS – INTRODUÇÃO

- O que é Ciência de Dados ?
  - Estudos, Processos, Modelos, Arte …
  - Existem diversas definições
  - Como você definiria Ciência de Dados ?

CIÊNCIA DE DADOS – INTRODUÇÃO

"**Data science**, also known as **data**-driven **science**, is an interdisciplinary field about **scientific** methods, processes, and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured, similar to **data** mining."

" Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices."



CIÊNCIA DE DADOS - INTRODUÇÃO

Sobre Dados

- Extrair dados
- Armazenar dados
- Explorar dados
- Analisar dados
- Descobrir, revelar informações escondidas
- Interpretar os dados
- Data Science Methodology

CIÊNCIA DE DADOS – INTRODUÇÃO

Quais dados

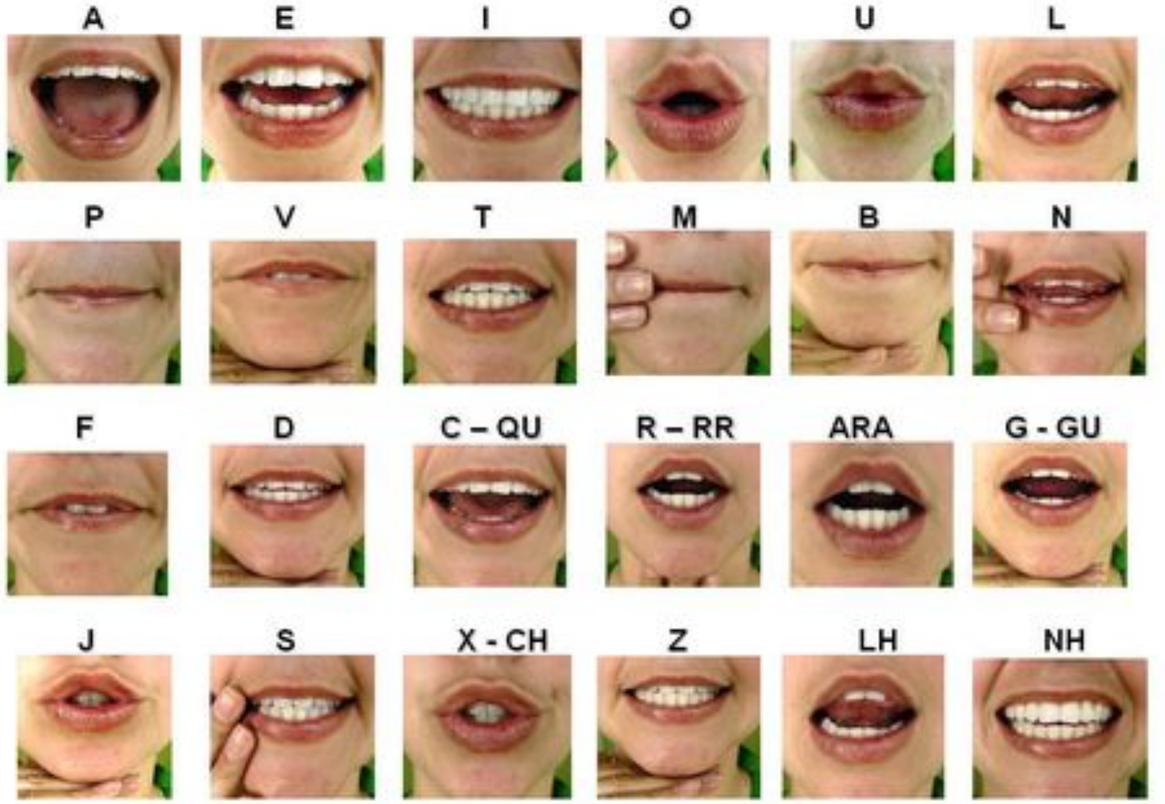
- Arquivos de Log
- E-mails
- Mídia social
- Dados de vendas
- Sensores de fábricas
- Câmeras de segurança
- Etc

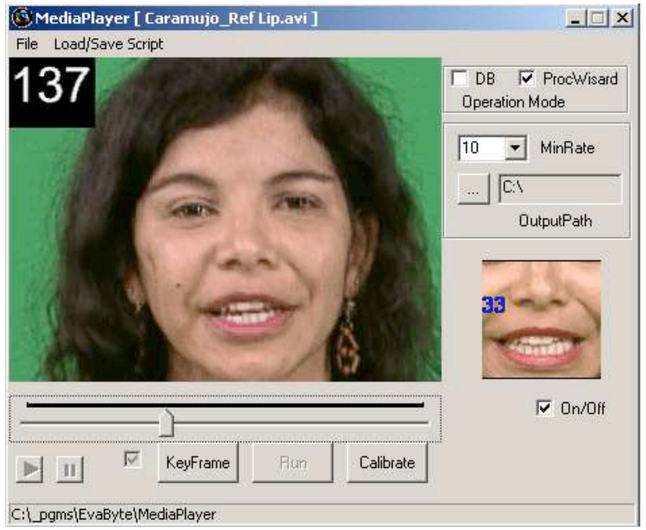
CIÊNCIA DE DADOS – INTRODUÇÃO

Relevância

- Diagnósticos/Prognósticos médicos
- Investimentos Financeiros
- Plantas Industriais
- Defesa e Segurança cibernética
- Outros ?

CIÊNCIA DE DADOS – INTRODUÇÃO





CIÊNCIA DE DADOS – INTRODUÇÃO

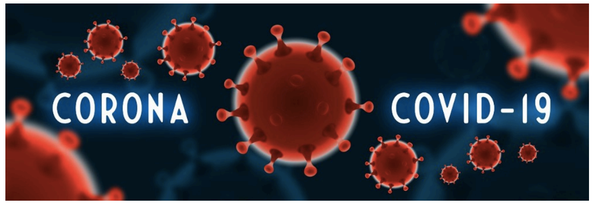
Detecção de fraudes em cartão de créditos

- Exercício em Aula
- Localidade (de uso do cartão, do usuário)
- Valores
- Tipo de compra
- CPF, dados do comprador
- Renda do usuário, Delta de distância e tempo
- Horário da compra
- Compra na Internet - IP da Loja
- Classe de cartões
- Informações da loja física , virtual
- Histórico de fraudes de uma região
- * Mecanismo anti-fraude
- Valor médio de vendas de uma Loja

CIÊNCIA DE DADOS - INTRODUÇÃO

Monitoramento da COVID19

- Mostrar Exemplo



CIENCIA DE DADOS - INTRODUÇAO

- https://colab.research.google.com/

CIENCIA DE DADOS – INTRODUÇAO

Papel do Cientista de Dados

- Contador de histórias
- Apresentador de dados - visualização de dados
- Explorador de dados
- “New data scientists need to be curious, judgemental and argumentative."
- Algumas características : matemática, ciência, **dados**.

CIENCIA DE DADOS - INTRODUÇAO

Cientista de Dados (algumas responsabilidades)

- Identificar o problema
- Conhecer em detalhes o problema (compreensão dos dados)
- Coletar os dados
- Identificar as ferramentas corretas
- Desenvolver uma estratégia de análise de dados

CIENCIA DE DADOS – INTRODUÇAO

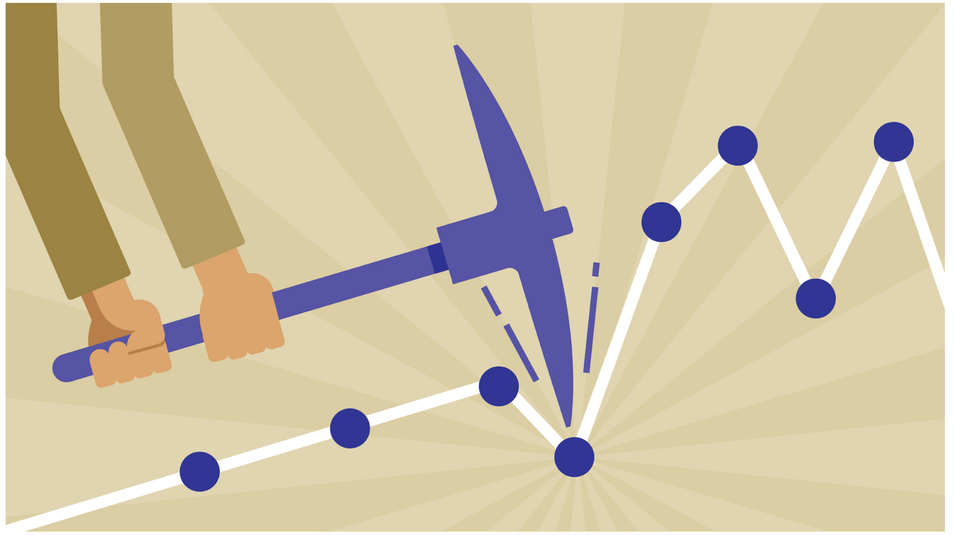
Big Data

- Velocidade
- Volume
- Variedade
- Veracidade
- Valor

CIENCIA DE DADOS – INTRODUÇAO

Data Mining

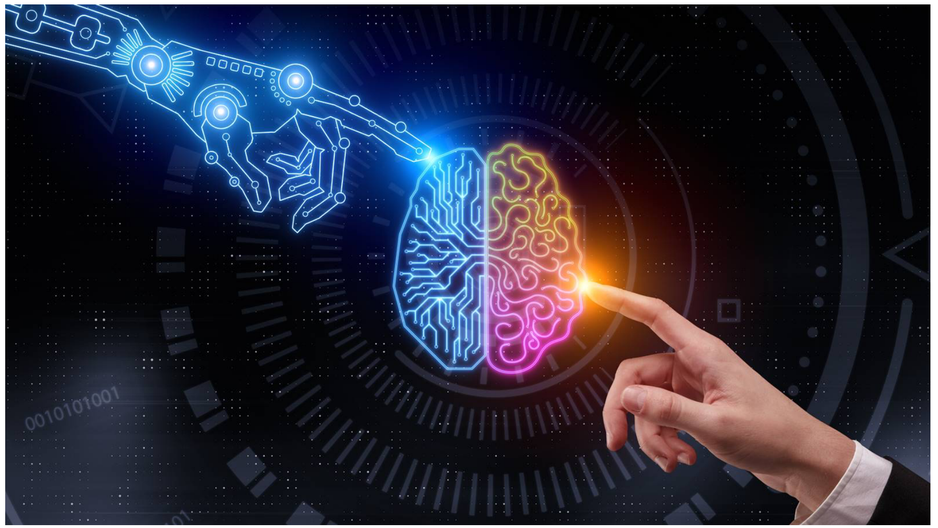
- Data mining is the process of automatically searching and analyzing data, discovering previously unrevealed patterns. It involves preprocessing the data to prepare it and transforming it into an appropriate format. Once this is done, insights and patterns are mined and extracted using various tools and techniques ranging from simple data visualization tools to machine learning and statistical model



CIENCIA DE DADOS – INTRODUÇAO

Artificial Intelligence

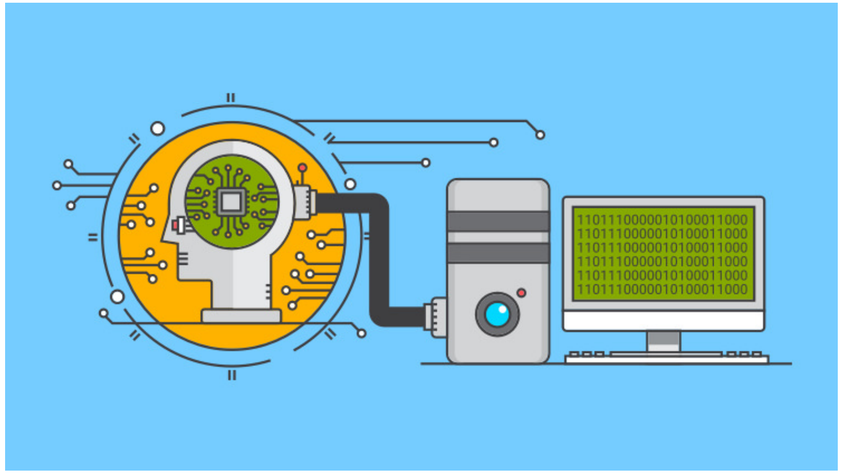
- AI is composed of 2 words Artificial and intelligence. Anything which is not natural and created by humans is artificial. Intelligence means ability to understand, reason, plan etc. So we can say that any code, tech or algorithm that enable machine to mimic, develop or demonstrate the human cognition or behavior is AI.



CIENCIA DE DADOS – INTRODUÇAO

Machine Learning

- Machine learning is a subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it is learned without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms. Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.

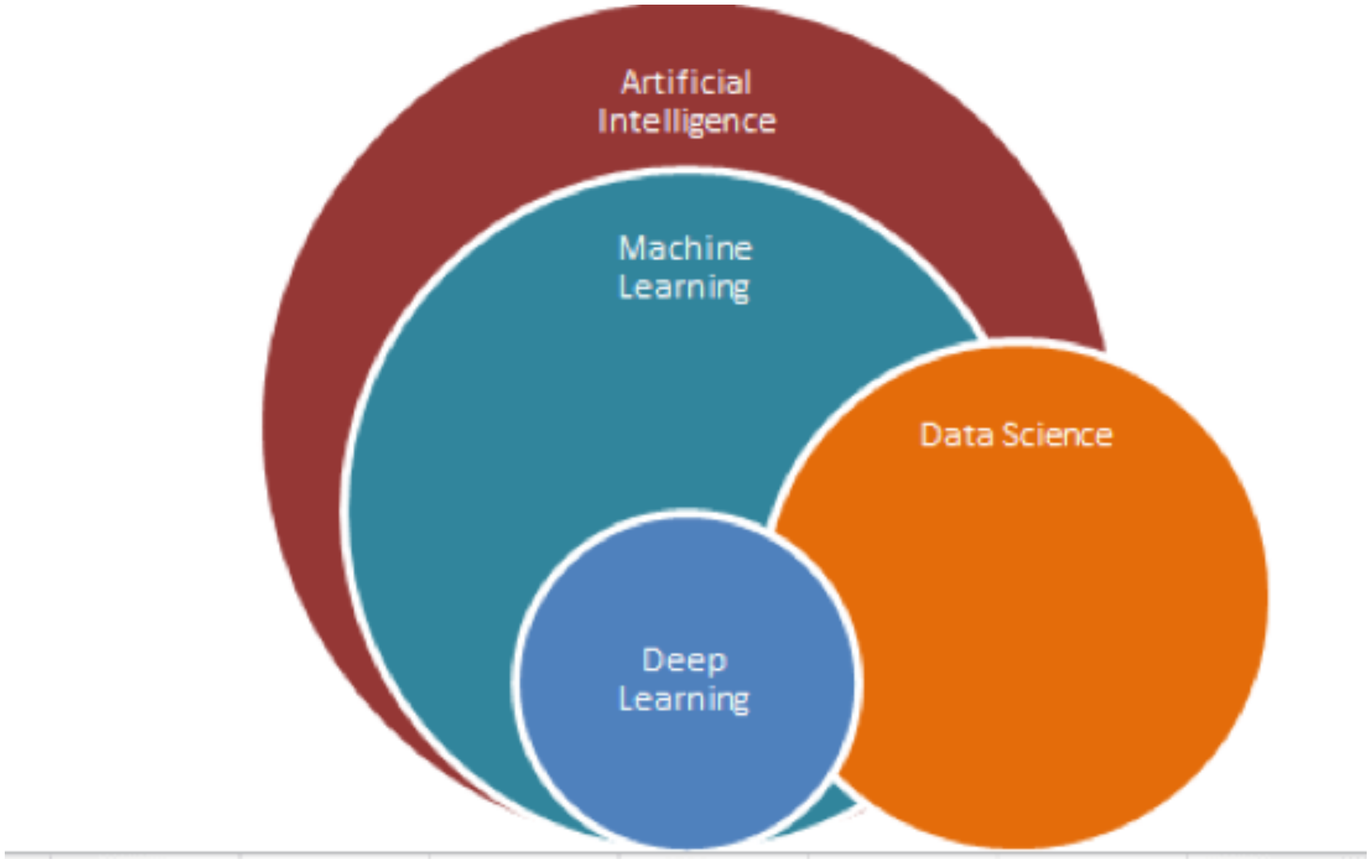


CIENCIA DE DADOS – INTRODUÇAO

Deep Learning

- Deep learning is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can label and categorize information and identify patterns. It is what enables AI systems to continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct.

CIENCIA DE DADOS – INTRODUÇAO



CIENCIA DE DADOS – INTRODUÇAO

Data science in Business

- Data Science helps physicians provide the best treatment for their patients, and helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- How businesses like Netflix, Amazon, UPs, Google, and Apple use the data generated by their consumers and employees.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.

CIENCIA DE DADOS - INTRODUÇAO

Data Scientists

- Data Scientists need programming, mathematics, and database skills, many of which can be gained through self-learning.
- Companies recruiting for a Data Science team need to understand the variety of different roles Data Scientists can play, and look for soft skills like storytelling and relationship building as well as technical skills.

CIENCIA DE DADOS – INTRODUÇAO

Tools for Data Science

- Linguagens de Programação - Python, R, SQL
- Outras linguagens - Scala,Java, C++, Julia
- Gerenciamento de dados - MySQL, PostgreSQL (relacional); NoSQL (MongoDB, CouchDB); Hadoop, Cloud File Systems (File-based);
- Integração e Transformação de dados (“ELT”) - Apache Air Flow, KubeFlow, Apache Nifi, Apache SparkSQL, NodeRED etc
- Visualização de dados - Hue (SQL queries) , Kibana, Apache Superset etc
- Model Monitoring - ModelDB, Prometheus
- Model Performance - IBM AI Fairness 360
- Desenvolvimento de software x Gerenciamento/Controle de versão - GitHub

CIENCIA DE DADOS – INTRODUÇAO

Python

- Python is a high-level general-purpose programming language that can be applied to many different classes of problems.
- It has a large, standard library that provides tools suited to many different tasks, including but not limited to databases, automation, web scraping, text processing, image processing, machine learning, and data analytics.
- For data science, you can use Python's scientific computing libraries such as Pandas, NumPy, SciPy, and Matplotlib.
- For artificial intelligence, it has TensorFlow, PyTorch, Keras, and Scikit-learn.

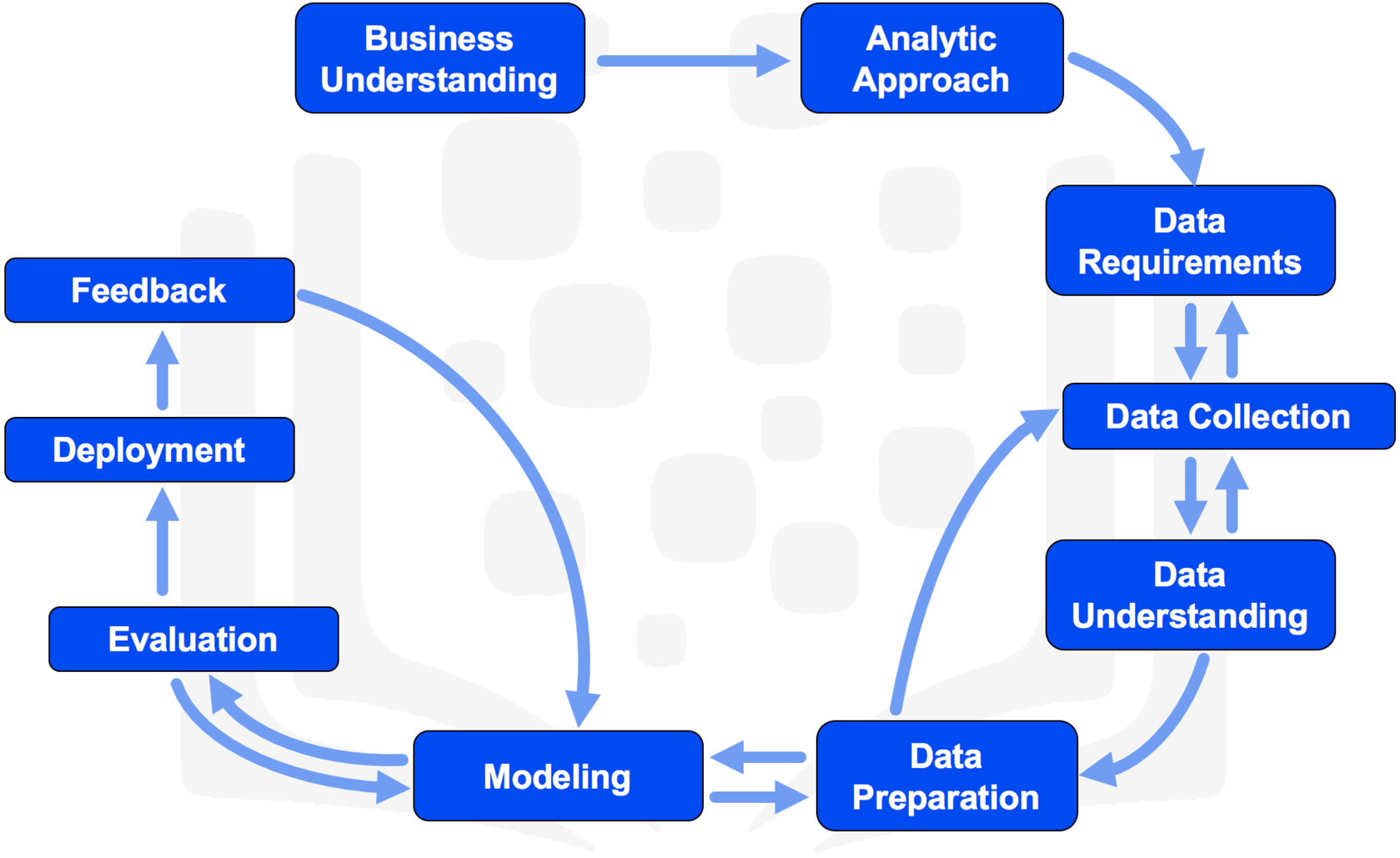
CIENCIA DE DADOS – INTRODUÇAO

Bibliotecas para Data Science (Python)

- Manipulação de dados - Numpy, SciPy, Pandas
- Visualização de dados - Matplotlib, Seaborn
- Machine Learning. - Scikit-learn, TensorFlow (deep learning)

CIENCIA DE DADOS - INTRODUÇAO

Data Science Methodology



CIENCIA DE DADOS – INTRODUÇAO