Data Science & Business Analytics

Machine Learning Models

David Issá
davidribeiro.issa@gmail.com

Introdução

Objetivos

1. **Adquirir** uma compreensão fundamental dos **principais conceitos de Machine Learning.**
2. **Compreender** os **conceitos-chave** na aquisição, preparação, exploração, visualização e modelização de dados.
3. **A teoria** dos modelos e metodologias fundamentais de Machine Learning **será combinada com casos práticos**.
4. **Os casos práticos** serão orientados para a aplicação de **como construir modelos ML**, bem em **como obter insights** a partir dos seus resultados, **sempre usando Python**.

No final do modulo, o aluno deverá:

1. **Ter uma compreensão** das diferentes tarefas em Data Science e **dos algoritmos mais adequados** para as executar.
2. **Compreender e aplicar uma vasta gama de algoritmos**, incluindo árvores de classificação e regressão, clustering, entre outros.
3. **Avaliar a precisão e performance** dos modelos/algoritmos.
4. **Demonstrar capacidade para realizar um trabalho prático** que exija a aplicação de técnicas de ML.

Programa

1. Introdução a Machine Learning
2. Metodologias de Data Science
3. Compreensão e Pré-Processamento de dados
4. Unsupervised Learning - Clustering
5. Supervised Learning – Avaliação
6. Supervised Learning – Feature Selection
7. Supervised Learning – Modelos

Avaliação

Assiduidade e Pontualidade (15%)

Motivação e Participação (15%)

Quiz Individual (30%)

- 16 questões de escolha múltipla (1 hora)
- A realizar até dia 13 de Abril.

Trabalho de Grupo (40%)

- Exercício prático de clustering.
- Será partilhado com os alunos dia 2 de Abril.
- Os alunos terão até ao fim do dia 20 de Abril para entregar o trabalho.

E o mais importante…

Vamos ter um intervalo!
Entre as 20:45 e as 21:15

Alguma questão?

1. O que é *Data*?

1. O que é *Data*?

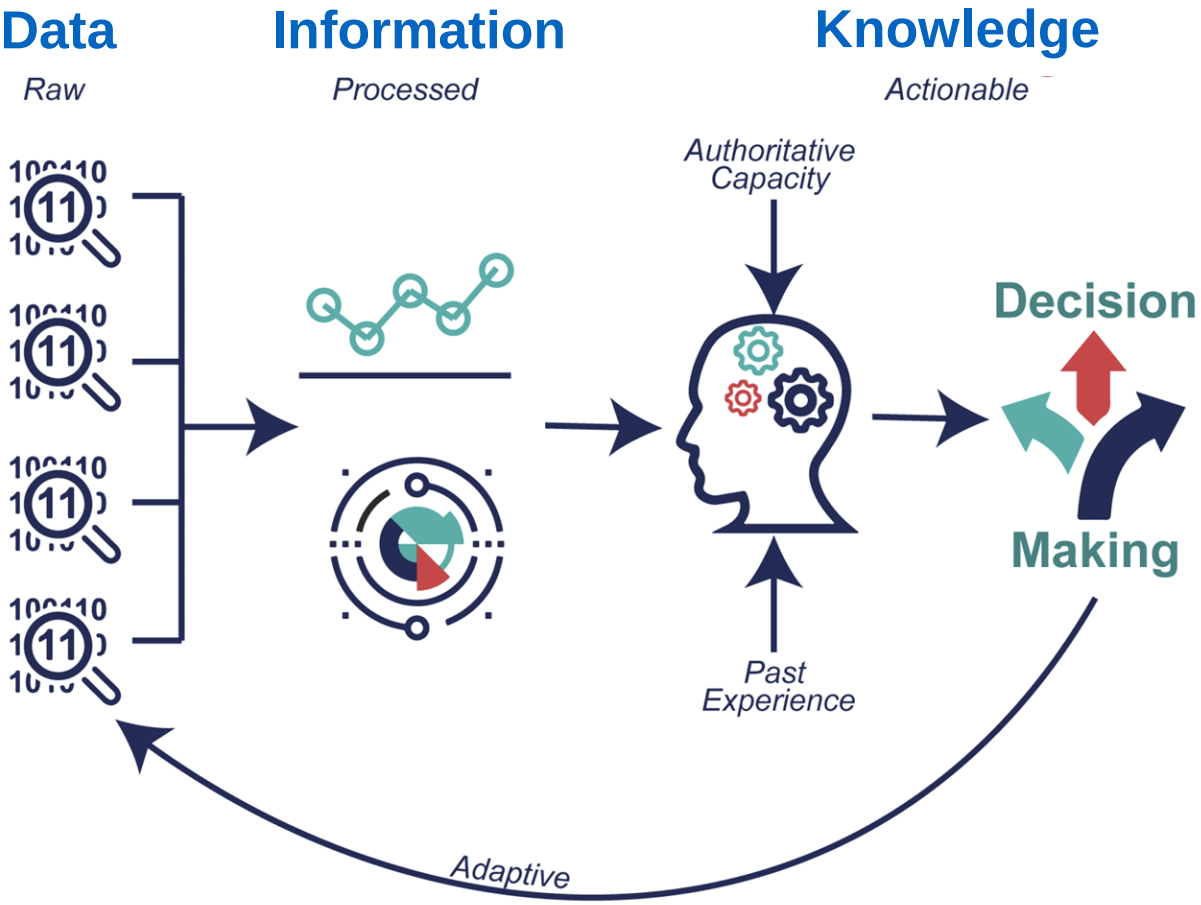
"Data is Information, especially facts or numbers, collected to be examined and considered and used to help decision making, or information in an electronic form that can be stored and used by a computer."

Source: https://dictionary.cambridge.org/dictionary/english/data

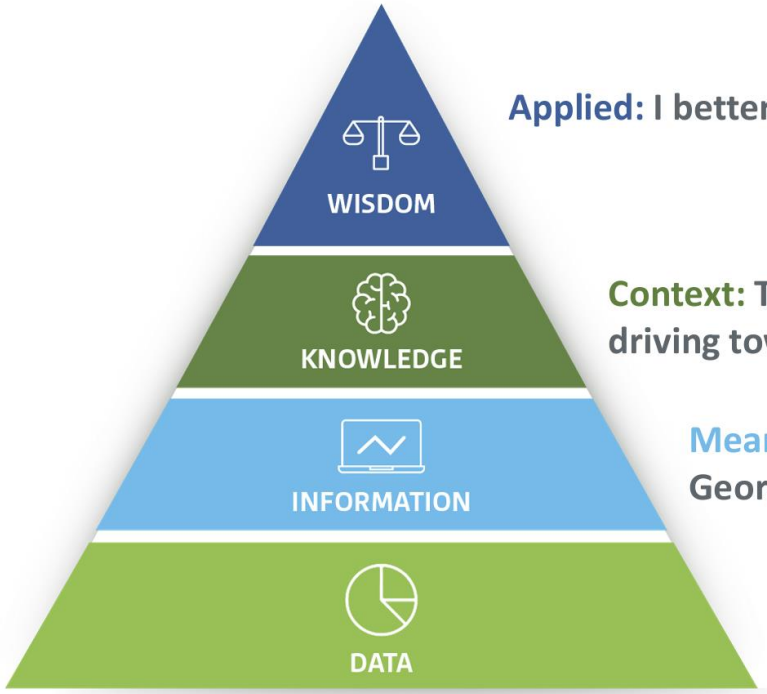
"Data is a collection of facts, numbers, words, observations or other useful information. Through data processing and data analysis, organizations transform raw data points into valuable insights that improve decision-making and drive better business outcomes."

Source: https://www.ibm.com/think/topics/data

1. O que é *Data*?


Data
Raw
Information
Processed
Knowledge
Actionable
Authoritative Capacity
Decision Making
Past Experience
Adaptive

1. O que é *Data*?



Applied: I better stop the car!

Context: The traffic light I am driving towards has turned red

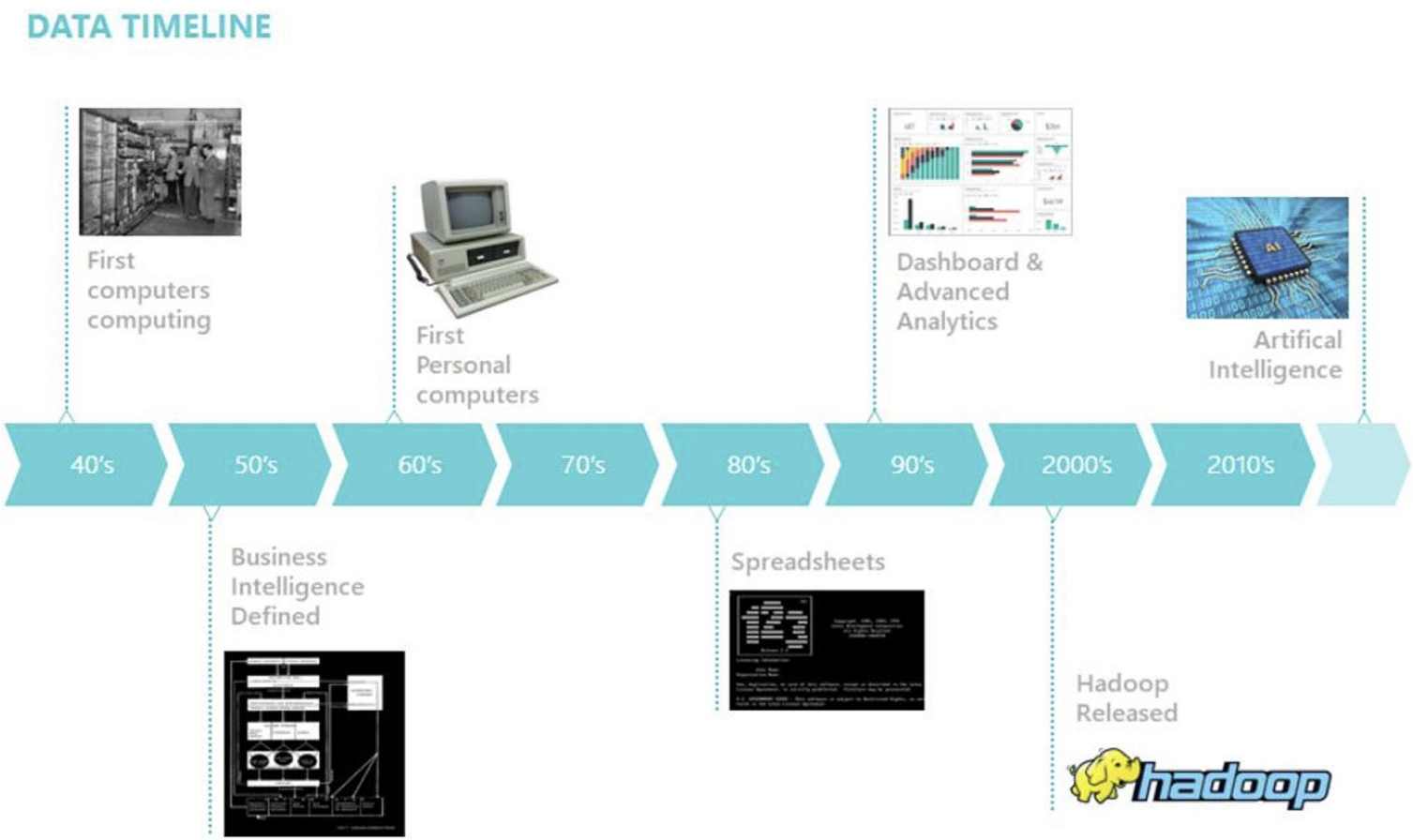
Meaning: South facing traffic light on corner of Pitt and George Streets has turned red

Raw: Red, 192.234.235.245.678, v2.0

DIKW Pyramid


The Economist
MAY 6TH–12TH 2017
Crunch time in France
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
The world's most valuable resource
amazon
Microsoft
Google
TESLA
UBER
Data and the new rules of competition

1. O que é *Data*?


DATA TIMELINE
First computers computing
First Personal computers
Dashboard & Advanced Analytics
Artifical Intelligence
40's
50's
60's
70's
80's
90's
2000's
2010's
Business Intelligence Defined
Spreadsheets
Hadoop Released
hadoop

1. O que é *Data*?

DATA

Informações ou factos sob qualquer forma, como números, texto, imagens ou áudio, que **podem ser processados e analisados.**

Os dados **podem ser armazenados em bases de dados, folhas de cálculo ou outras estruturas.**

BIG DATA

Refere-se a **conjuntos de dados extremamente grandes e complexos** que são gerados e recolhidos a uma escala sem precedentes.

O termo “big data” é frequentemente utilizado para descrever conjuntos de **dados demasiado grandes, em rápida mudança ou complexos para serem processados e analisados por ferramentas tradicionais de gestão de dados.**

1. O que é *Data*? - Big Data

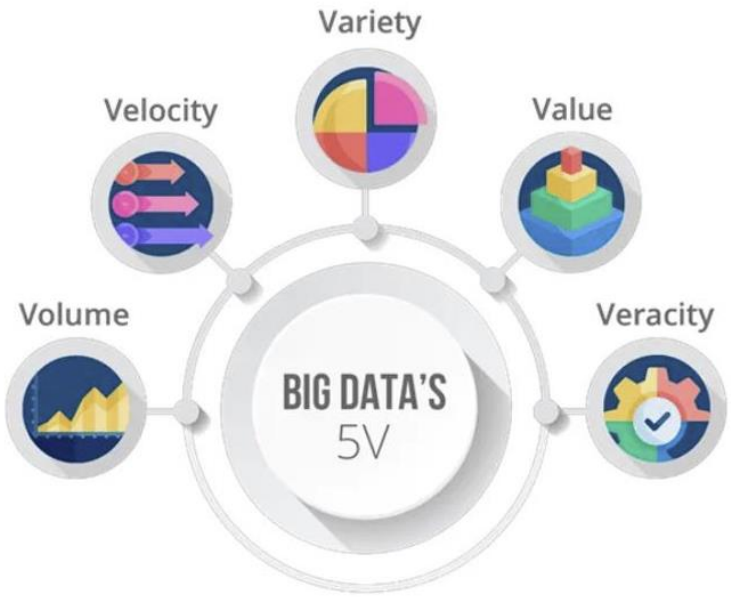
Big data é uma **combinação de dados estruturados, semiestruturados e não estruturados**

recolhidos por empresas e organizações

que **podem ser explorados para obter informações**

e **utilizados em projectos de machine learning, modelação preditiva e outras aplicações analíticas** avançadas

1. O que é *Data*? - Os 5 V's de Big Data



Volume: A quantidade de dados provenientes de inúmeras fontes

Velocidade: A velocidade a que os dados são gerados

Variedade: Os tipos de dados: estruturados, semi-estruturados, não estruturados

Valor: O valor comercial dos dados recolhidos

Veracidade: O grau de fiabilidade dos dados

1. O que é *Data*? - As fases…

Data is changing…

BIG DATA PHASE 1
Period: 1970 - 2000

DBMS-based, structured content:

- RDBMS & data warehousing
- ETL
- Online Analytical Processing
- Dashboards & Scorecards
- Data Mining & Statistical analysis

BIG DATA PHASE 2
Period: 2000 - 2010

Web-based, unstructured content

- Information retrieval and extraction
- Opinion mining
- Question answering
- Web analytics and web intelligence
- Social media analytics
- Social network analysis
- Spatial-temporal analysis

BIG DATA PHASE 3
Period: 2010 - present

Mobile and sensor-based content

- Location-aware analysis
- Person-centered analysis
- Context-relevant analysis
- Mobile visualization
- Human-Computer Interaction

1. O que é *Data*? - As fases…

The web is changing…

WEB 1.0
Period: 1990 - 2005

- Basic Web pages
- HTML
- Ecommerce
- JAVA

WEB 2.0
Period: 2006 - Present

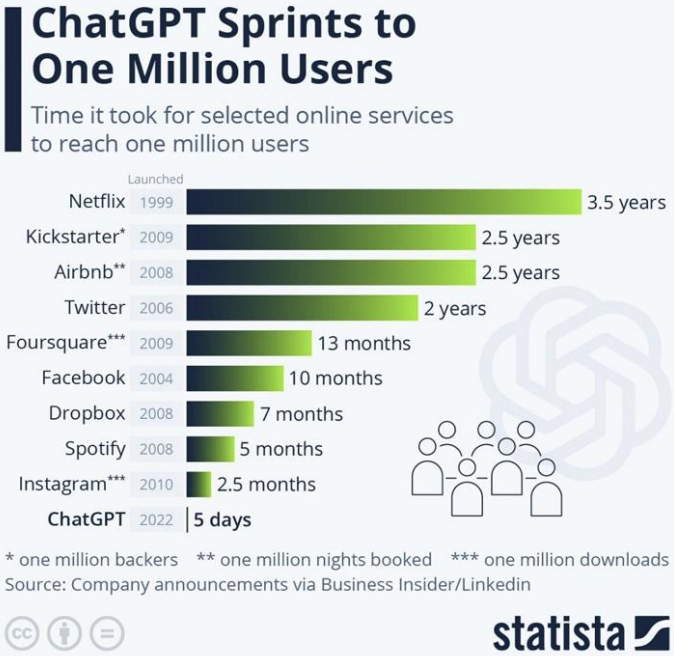
- Social Media
- Global Internet access
- Web Apps
- Data monetization

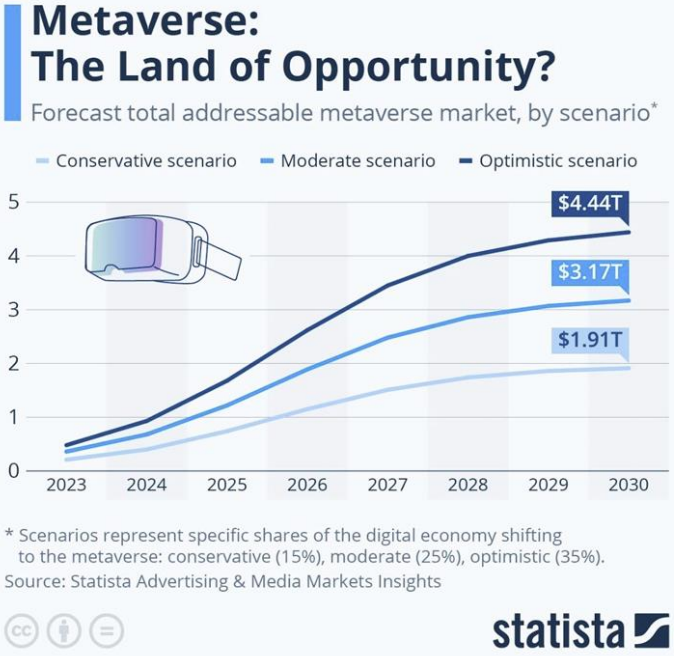
WEB 3.0
Period: Forthcoming

- NFTs
- Semantic Web
- Metaverse (AR & VR)
- Blockchains
- Artificial Intelligence
- Interoperability

1. O que é *Data*? - As fases…

The web is changing…

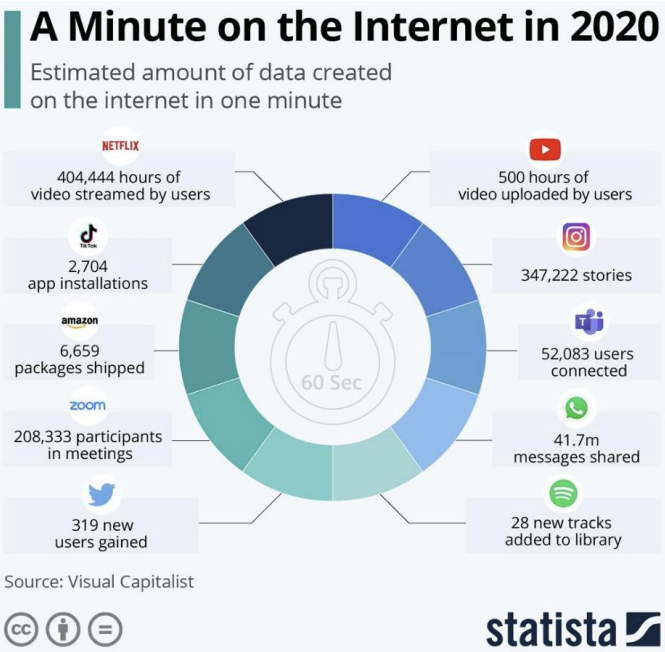


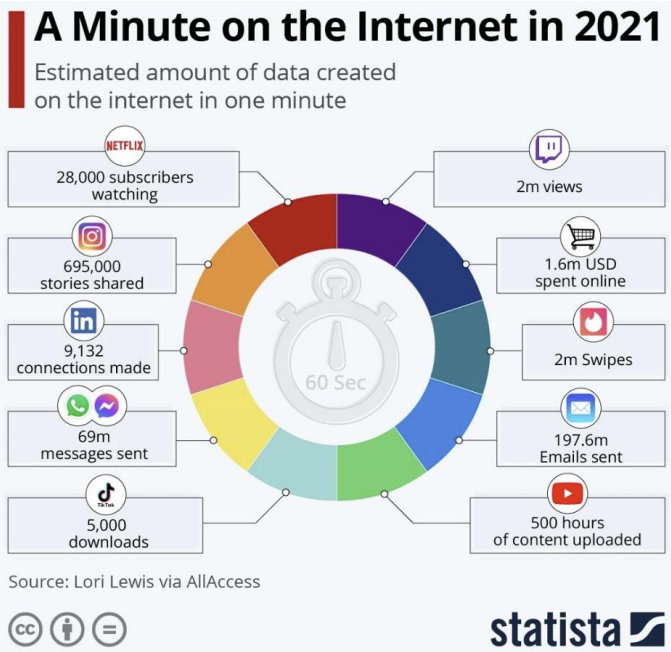


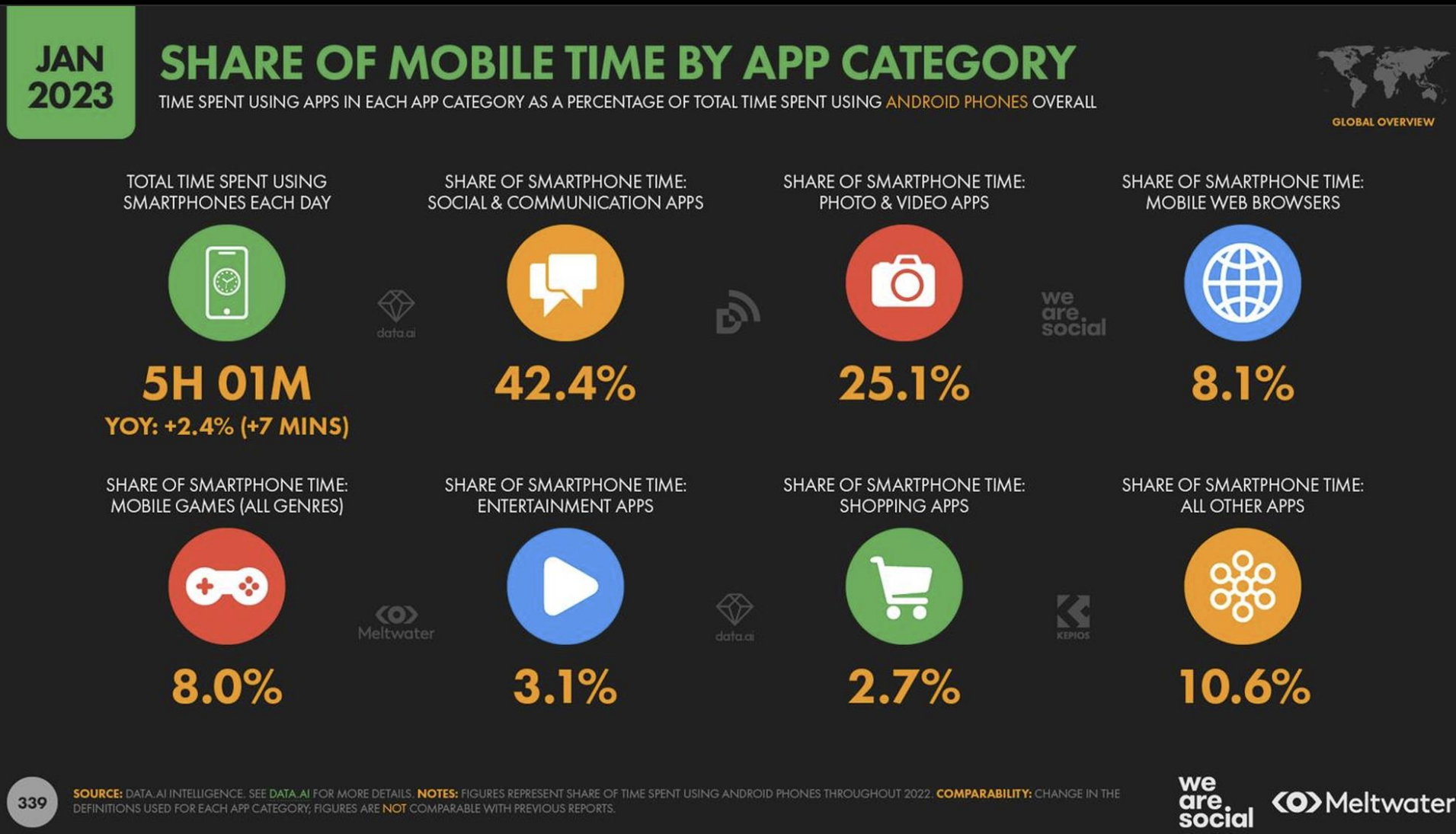
1. O que é *Data*? - Fontes de dados

1. O que é *Data*? - Fontes de dados

How much data?







Source: https://datareportal.com/reports/digital-2023-global-overview-repor

JAN 2023

OVERVIEW OF CONSUMER GOODS ECOMMERCE

HEADLINES FOR THE ADOPTION AND USE OF CONSUMER GOODS ECOMMERCE (B2C ONLY)

GLOBAL OVERVIEW

4.11 BILLION	$3.59 TRILLION	$873	59.8%	17.1%
YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE	YEAR-ON-YEAR CHANGE
+8.3% (+315 MILLION)	-6.5% (-$250 BILLION)	-13.7% (-$138)	+1.2% (+71 BPS)	+4.4% (+72 BPS)

362

SOURCE: STATISTA DIGITAL MARKET OUTLOOK. SEE STATISTA.COM FOR MORE DETAILS. **NOTES:** "CONSUMER GOODS" INCLUDE: ELECTRONICS, FASHION, FURNITURE, TOYS, HOBBY, DIY, BEAUTY, CONSUMER HEALTHCARE, PERSONAL CARE, HOUSEHOLD CARE, FOOD, BEVERAGES, AND PHYSICAL MEDIA. FIGURES REPRESENT ESTIMATES FOR FULL-YEAR 2022, AND COMPARISONS WITH EQUIVALENT VALUES FOR THE PREVIOUS CALENDAR YEAR. FINANCIAL VALUES ARE IN U.S. DOLLARS. PERCENTAGE CHANGE VALUES ARE RELATIVE (I.E. AN INCREASE OF 20% FROM A STARTING VALUE OF 50% WOULD EQUAL 60%, NOT 70%). "BPS" VALUES REPRESENT BASIS POINTS, AND INDICATE ABSOLUTE CHANGE. **COMPARABILITY:** BASE AND CATEGORY DEFINITION CHANGES. FIGURES ARE NOT COMPARABLE WITH PREVIOUS REPORTS.

we are social Meltwater

Source: https://datareportal.com/reports/digital-2023-global-overview-repor

2. Machine Learning – alguns exemplos

target

TECH | 2/16/2012 @ 11:02AM | 2.316.057 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

2. Machine Learning – alguns exemplos

BARNES & NOBLE

Descobriu que os leitores abandonam frequentemente os livros de não-ficção a menos de metade a meio. Introduziu uma nova série de livros curtos de grande sucesso sobre temas da atualidade.

amazon

Originalmente utilizava um painel de avaliadores especializados para livros. O excesso de dados permitiu-lhes que criassem recomendações cada vez mais preditivas. Desde então, o painel foi foi dissolvido e 1/3 das vendas são agora efectuadas pelo sistema de recomendação.

2. Machine Learning – alguns exemplos

Analisou os preços dos bilhetes para voos específicos com base em dados históricos, depois aconselhou os utilizadores a comprar ou esperar de acordo com a trajetória prevista dos custos dos bilhetes.

Utiliza uma série de dados de tráfego para calcular os itinerários mais eficientes em termos de tempo/combustível de acordo com algoritmos complexos.

3. Machine Learning

3. Machine Learning – Conceitos

3. Machine Learning – Conceitos

Data Mining

O processo de **descobrir padrões, relações e tendências** em grandes conjuntos de dados, normalmente **através da utilização de algoritmos estatísticos e de ML.**

O **objetivo é identificar e extrair informações valiosas** de conjuntos de dados e transformá-los em conhecimento acionável.

Data Science

Domínio multidisciplinar que envolve a utilização de métodos estatísticos, matemáticos e computacionais para **extrair conhecimentos dos dados**, transformando os dados em informações acionáveis.

São utilizadas **técnicas como ML, visualização de dados e análise de dados para extrair informações dos dados e criar modelos preditivos.**

3. Machine Learning – Conceitos

Machine Learning

É um campo de estudo da área de Inteligência Artificial, que utiliza algoritmos que aprendem com os dados para fazer previsões.

Estas previsões podem ser geradas através de **supervised learning, em que os algoritmos aprendem padrões a partir de dados existentes**, ou de **unsupervised learning, em que descobrem padrões nos dados**.

3. Machine Learning – Conceitos

Data Analysis

Processo de **avaliação e modelação para extrair conhecimentos e orientar a tomada de decisões.**

Envolve a recolha, limpeza, transformação e modelação de dados para identificar tendências e relações que podem ajudar as organizações a tomar decisões informadas

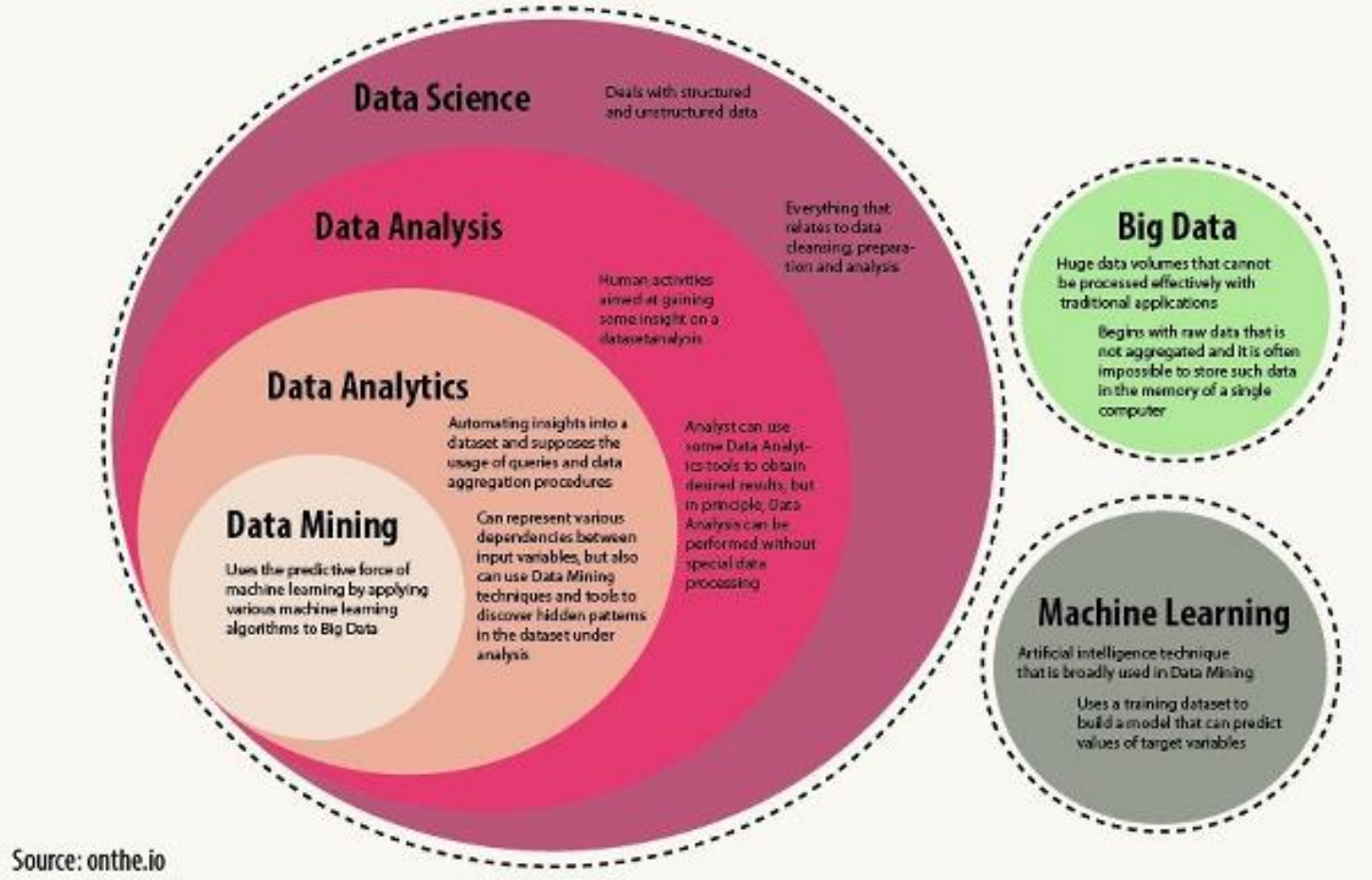
Pode ser aplicada a vários tipos de dados, incluindo dados estruturados e não estruturados.

Data Analytics

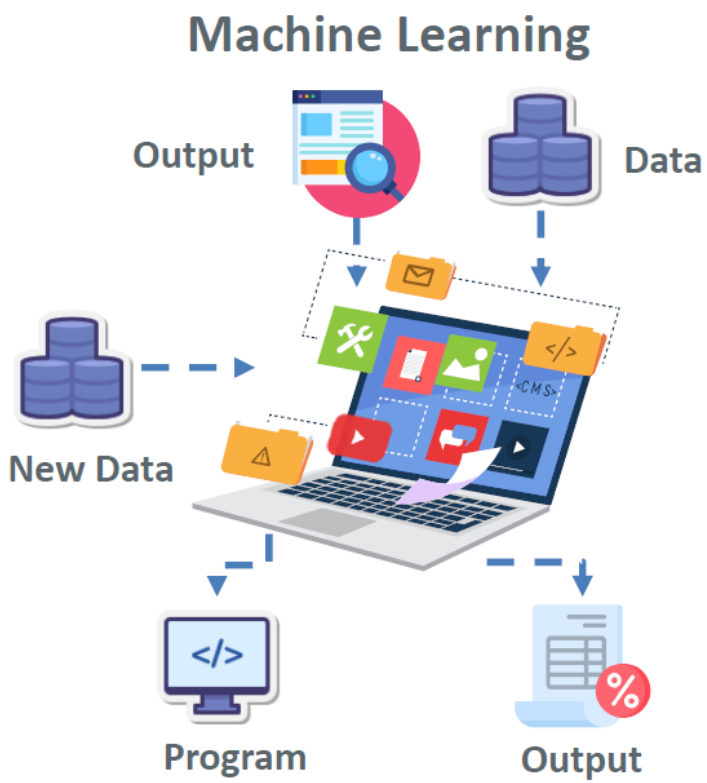
Consiste na **aplicação de um processo mecânico ou algorítmico para obter informações**, por exemplo, analisando vários conjuntos de dados à procura de correlações significativas entre eles.

Trata-se de **automatizar o conhecimento de um conjunto de dados** e pressupõe a utilização de queries e procedimentos de agregação de dados.

3. Machine Learning – Conceitos



3. Machine Learning – Conceitos



3. Machine Learning – Algoritmos

Como aprender todos os algoritmos de ML? É possível?

Objetivo: perceber as diferentes etapas no processo de treinar o modelo.

Não vale a pena decorar modelos…

3. Machine Learning – Algoritmos

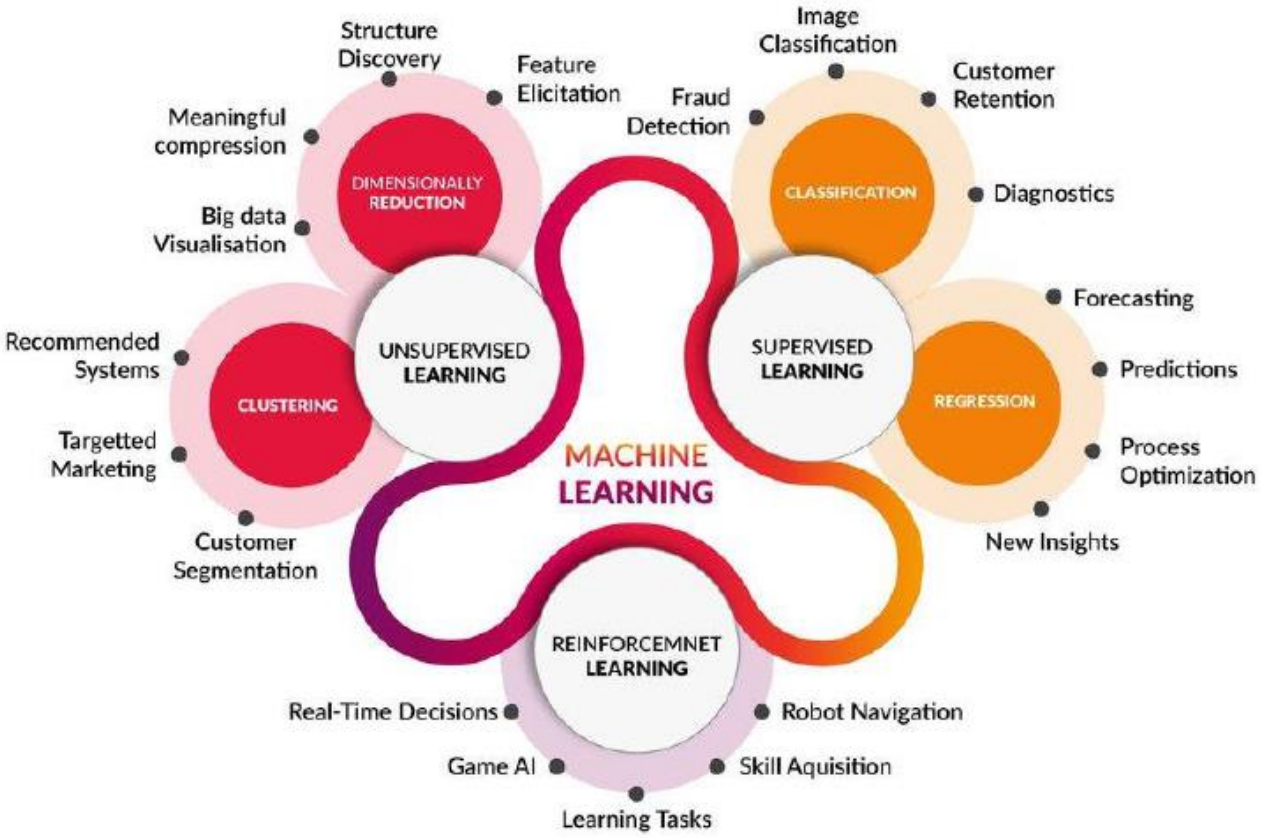
3. Machine Learning – Avaliação de Algoritmos

A escolha do algoritmo mais adequado ao nosso problema requer a **escolha de um método de avaliação.**

Alguns deles aplicam-se a problemas de classificação, outros a problemas de regressão… alguns são mais indicativos quando os nossos dados são equilibrados (balanced), outros quando os dados são desiquilibrados (unbalanced).

Depende do tipo de problema e do tipo de tipo de dados de que dispomos!

3. Machine Learning – Tipos de Aprendizagem



3. Machine Learning – Tipos de Aprendizagem

Supervised Learning

Tipo de aprendizagem em que o algoritmo é **treinado num conjunto de dados categorizados**, o que significa que **o resultado desejado é fornecido para cada exemplo nos dados de treino**.

O algoritmo utiliza estes dados de treino para **aprender a fazer previsões sobre novos dados não vistos.**

Unsupervised Learning

Tipo de aprendizagem em que o algoritmo é **treinado num conjunto de dados não categorizados.**

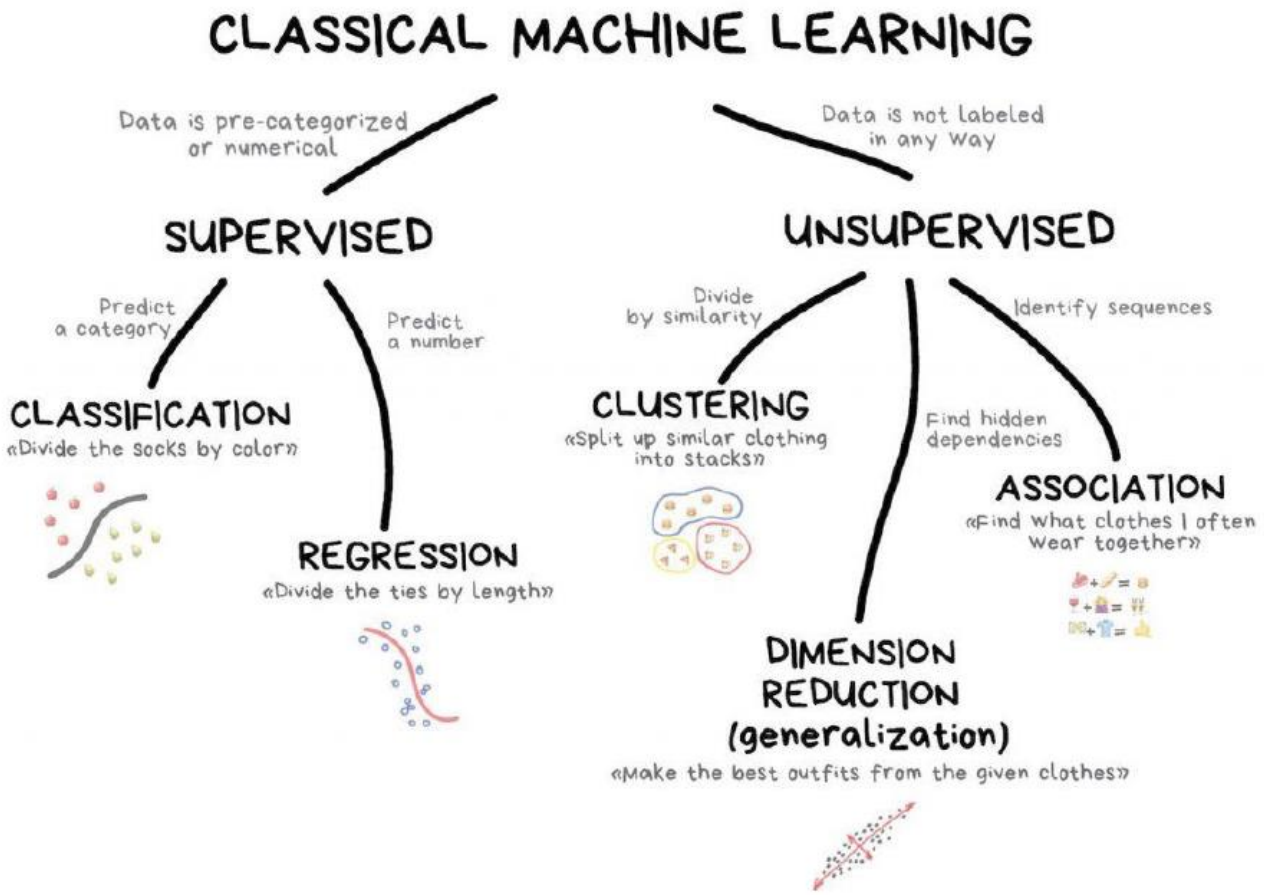
O objetivo é **identificar padrões, relações e estruturas nos dados**, sem que seja dito qual o objetivo ou resultado.

Reinforcement Learning

Aprendizagem centrada **na tomada de decisões para maximizar as recompensas** acumuladas numa determinada situação.

Envolve a **aprendizagem através da experiência**. Um agente aprende a atingir um objetivo, realizando acções e recebendo **feedback através de recompensas ou penalizações**.

3. Machine Learning – Tipos de Aprendizagem



3. Machine Learning – Unsupervised

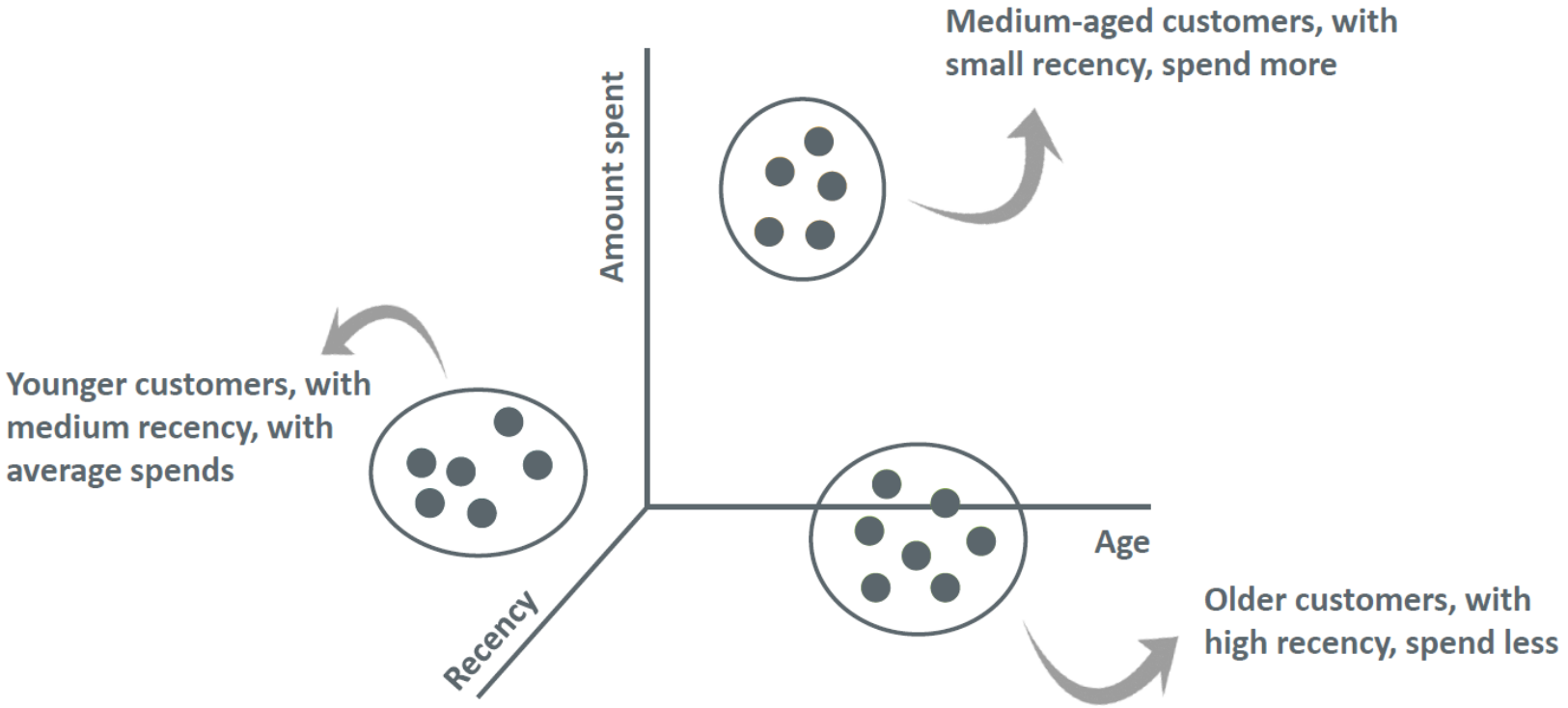
Clustering

Clustering refere-se à tarefa de agrupar observações em classes de objectos semelhantes.

- Um cluster é uma coleção de observações semelhantes entre si, mas diferentes das observações de outros clusters;
- Os algoritmos de clustering procuram segmentar um conjunto de dados em subgrupos homogéneos;
- Nenhuma variável-alvo não é especificada;
- O clustering não tenta classificar / estimar / prever a variável-alvo;

3. Machine Learning – Unsupervised

Clustering



3. Machine Learning – Unsupervised

Association

Association refere-se à tarefa de encontrar associações e relações entre grandes conjuntos de atributos de dados. Este tipo de algoritmos calcula a frequência com que um conjunto de atributos ocorre numa transação.

- Normalmente utilizado para a análise do cabaz de compras
- Quantificar as relações entre dois ou mais atributos sob a forma de regras como:

$$IF\ antecedent\ THEN\ consequent$$

3. Machine Learning – Unsupervised

Association

Um determinado supermercado:

- Na quinta-feira à noite, 200 dos 1.000 clientes compraram fraldas, e 50 dos que compraram fraldas compraram cerveja;
- Regra de associação: “SE comprar fraldas, ENTÃO compra cerveja”.
- Medidas usadas na definição da regra:
  Suporte = 50/1000 = 5%
  Confiança = 50/200 = 25%

3. Machine Learning – Supervised

Classificação

Semelhante aos modelos de regressão, mas em que a variável alvo é categórica.

- Modelos construídos a partir de registos de dados completos: os registos incluem valores para cada variável independente e para a variável alvo categórica, no conjunto de dados de treino;
- Para novas observações, é estimada a variável alvo;
- Exemplo: Estimar a probabilidade de um paciente ter diabetes tendo em conta o género, peso, altura e número de gravidezes.

3. Machine Learning – Supervised

Classificação

Prever probabilidade de ter diabetes

ID	Género	Peso	Altura	Gravidezes	Status
1	M	78	175	0	Sem Diabetes
2	F	66	155	3	Diabetes
3	F	91	165	1	Diabetes
4	M	89	187	0	Sem Diabetes
5	M	101	172	0	Diabetes
6	M	81	179	0	Sem Diabetes
7	F	72	169	0	Sem Diabetes
8	F	93	169	0	?

3. Machine Learning – Supervised

Regressão

Semelhante aos modelos de classificação, mas em que a variável alvo é numérica.

- Modelos construídos a partir de registos de dados completos: os registos incluem valores para cada variável independente e para a variável alvo categórica, no conjunto de dados de treino;
- Para novas observações, é estimada a variável alvo;
- Exemplo: Estimar o preço de uma casa com base nos metros quadrados da casa, localização, tipologia e número de casas de banhocasas de banho

3. Machine Learning – Supervised

Regressão

Prever o preço de uma casa

ID	m2	Localização	Tipo	WCs	Preço
1	234	Restelo	T5	3	1.112.000
2	107	Campolide	T2	2	365.000
3	67	Alfama	T1	1	240.000
4	86	Alavalade	T2	2	320.000
5	102	Campolide	T3	1	330.000
6	78	Benfica	T2	1	295.000
7	104	Areeiro	T2	2	367.000
8	122	Benfica	T3	1	?

Obrigado!