



Sistemas Inteligentes e Robótica

Curso 2025/2026

📅 Fev - Jun 2026 🏛️ Instituto Superior de Agronomia

Gestão de dados

Rui Figueira

ruifigueira@isa.ulisboa.pt

Sumário

- Dados, metadados, conjuntos de dados
- Dados privados, dados abertos, licenças, FAIR
- Tipos de dados, infraestruturas de dados, repositórios, serviços de dados
- Análise de dados
- Trabalhar com dados em falta
- Remoção de ruído
- Agregar dados

Desenvolvimento e Validação de Armadilha Inteligente de Baixo Custo para a Identificação da Mosca-da-Azeitona por Machine Learning

Rafael Valente Rodrigues

Dissertação para a obtenção do Grau de Mestre em
Ciência de Dados em Agricultura, Alimentação, Floresta e Ambiente

Orientadores: Rui Paulo Nóbrega Figueira
Livia Bischof Pian

Júri
Presidente: Doutor Manuel Lameiras de Figueiredo Campagnolo, Professor
Associado com agregação, Instituto Superior de Agronomia da
Universidade de Lisboa

Vogais: Doutor José Carlos Franco Santos Silva, Professor Associado, Instituto
Superior de Agronomia da Universidade de Lisboa
Doutora, Livia Bischof Pian, Investigadora, Smart Farm Colab



Figura 18 - Armadilha inteligente de Moura (à esquerda) e armadilha inteligente de Beja (à direita)



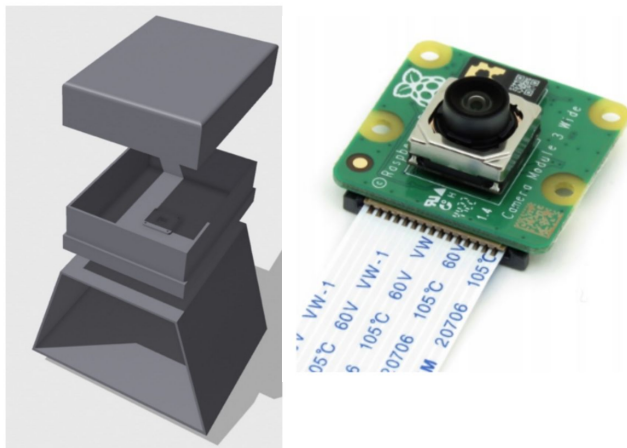


Figura 10 - Desenho 3D da estrutura da armadilha

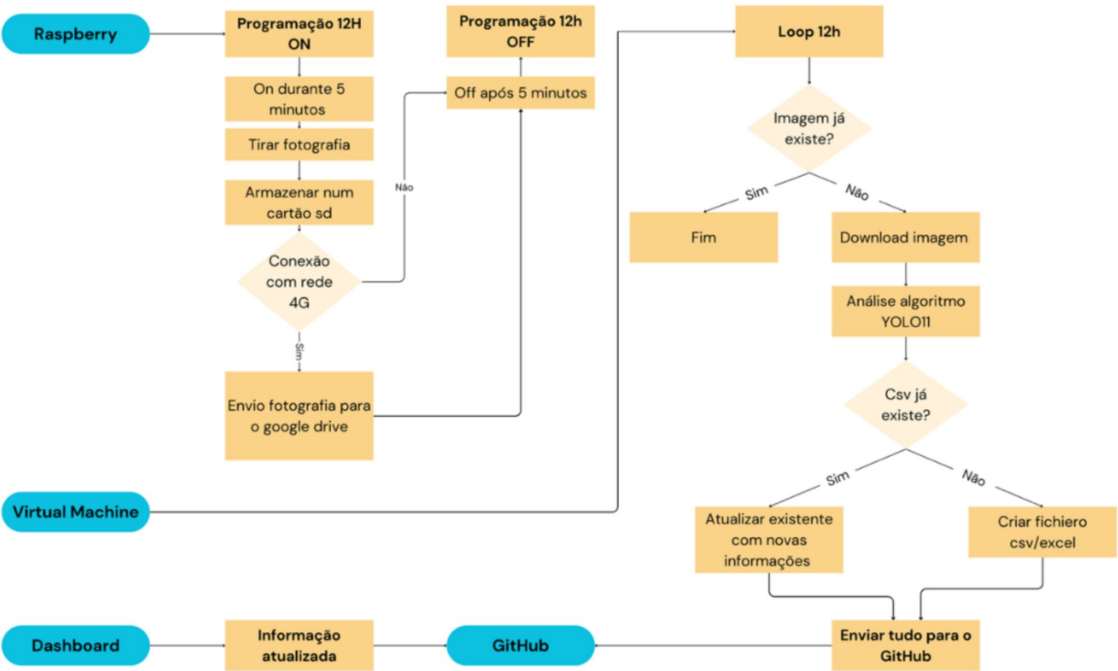


Figura 9 – Workflow de todo processo de automatização da armadilha inteligente, dados e informações

Filtros

Filtrar por localização

Choose options

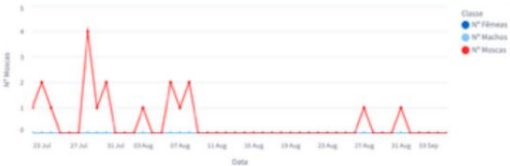
Filtrar por intervalo de datas

YYYY/MM/DD - YYYY/

Dashboard - Capturas da Mosca da Azeitona

Curva de Voo

Alerta: 1 dia com mais de 3 moscas capturadas.



Resumo Diário de Moscas

Data	Nº Fêmeas	Nº Machos	Nº Moscas	Acumulado
45 2025-09-05	0	0	0	19
44 2025-09-04	0	0	0	19
43 2025-09-03	0	0	0	19
42 2025-09-02	0	0	0	19
41 2025-09-01	0	0	0	19
40 2025-09-31	0	0	1	19
39 2025-09-30	0	0	0	18
38 2025-09-29	0	0	0	18
37 2025-09-28	0	0	0	18
36 2025-09-27	0	0	0	18

Total de Moscas por Classe



Moscas Capturadas por Semana

Semana	semana	macho	mosca	
30		0	0	4
31		0	0	6
32		0	0	5
33		0	0	2

Moscas Capturadas por Mês

Mês	semana	macho	mosca	
2025-07 (July)		0	0	11
2025-08 (August)		0	0	8

Total de Moscas Capturadas por Placa

Placa ID	semana	macho	mosca	
PLACA 20250901092745		0	0	19

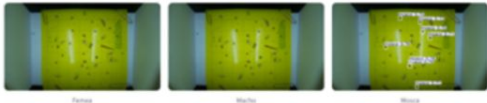
Mapa de Armadilhas com Detecções



image_20250827211929_ppp0.jpg

Localização: Beja

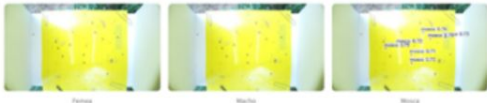
Detecções: F: 0 | M: 0 | Mos: 1



image_20250808100711_ppp0.jpg

Localização: Beja

Detecções: F: 0 | M: 0 | Mos: 1



image_20250808221835_ppp0.jpg

Localização: Beja

Detecções: F: 0 | M: 0 | Mos: 1



image_20250807111344_ppp0.jpg

Localização: Beja

Detecções: F: 0 | M: 0 | Mos: 1

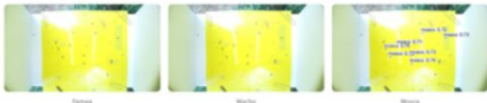


Figura 17 – Três capturas de ecrã sequenciais na vertical do *Dashboard* interativo

Exemplo 2



Preenchimento inteligente de lacunas de dados sensoriais: Aplicação de *Machine Learning* em Sistemas de Rega

Diogo Eugénio Pinto

Relatório de estágio para a obtenção do grau de Mestre em
Ciência de Dados em Agricultura, Alimentação, Floresta e Ambiente

Orientador: Rui Paulo Nóbrega Figueira

Júri:

Presidente: Doutor Manuel Lameiras de Figueiredo Campagnolo, Professor associado com
agregação do(a) Instituto Superior de Agronomia da Universidade de Lisboa.

Vogais: Doutor Rui Paulo Nóbrega Figueira, Professor auxiliar do(a) Instituto Superior de
Agronomia da Universidade de Lisboa, orientador;

Doutor João Rui Rolim Fernandes Machado Lopes, Professor auxiliar do(a) Instituto Superior
de Agronomia da Universidade de Lisboa.

2024

U LISBOA | UNIVERSIDADE
DE LISBOA



Figura 14 - Sonda EnviroPro® antes da instalação (foto pelo autor).

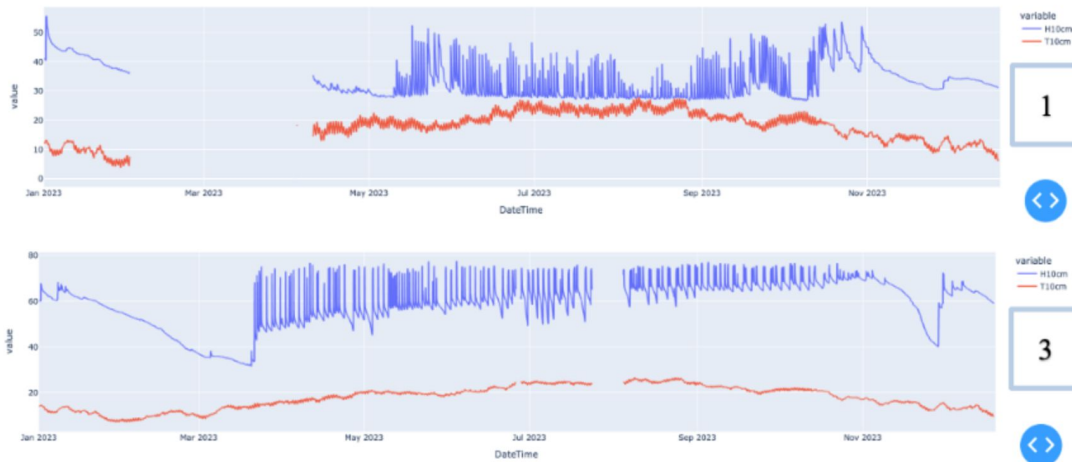


Figura 20 - Variação da humidade do solo (curva azul) e temperatura do solo (linha laranja) ao longo do ano de 2023, respectivo ao campo 1 e 3.

**Preenchimento inteligente de lacunas de dados
sensoriais: Aplicação de *Machine Learning* em Sistemas
de Rega**

Diogo Eugénio Pinto

Relatório de estágio para a obtenção do grau de Mestre em
Ciência de Dados em Agricultura, Alimentação, Floresta e Ambiente

Orientador: Rui Paulo Nóbrega Figueira

Júri:

Presidente: Doutor Manuel Lameiras de Figueiredo Campagnolo, Professor associado com
agregação do(a) Instituto Superior de Agronomia da Universidade de Lisboa.

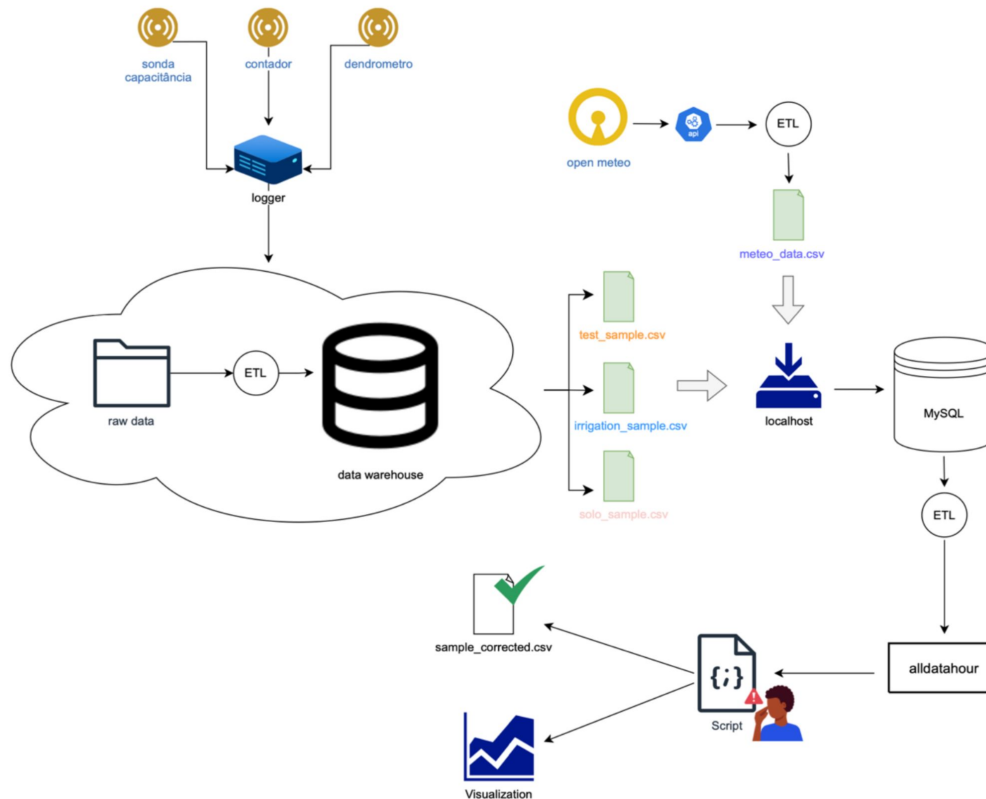
Vogais: Doutor Rui Paulo Nóbrega Figueira, Professor auxiliar do(a) Instituto Superior de
Agronomia da Universidade de Lisboa, orientador;

Doutor João Rui Rolim Fernandes Machado Lopes, Professor auxiliar do(a) Instituto Superior
de Agronomia da Universidade de Lisboa.

2024



Figura 14 - Sonda EnviroPro® antes da instalação (foto pelo autor).



Exemplo 3



INSTITUTO
SUPERIOR DE
AGRONOMIA
Universidade de Lisboa

DEVELOPMENT OF A FOREST CLEAR-CUT REFERENCE DATABASE FOR CONTINENTAL PORTUGAL AT SENTINEL-2 RESOLUTION

INÊS INGHAM BARROS DA SILVEIRA

DISSERTATION FOR A MASTER'S DEGREE IN
CIÊNCIAS DE DADOS EM AGRICULTURA, ALIMENTAÇÃO, FLORESTA E
AMBIENTE

ORIENTADOR: DOUTOR MANUEL LAMEIRAS DE FIGUEIREDO CAMPAGNOLO

JÚRI:

PRESIDENTE: DOUTOR RUI PAULO NÓBREGA FIGUEIRA, PROFESSOR AUXILIAR DO
INSTITUTO SUPERIOR DE AGRONOMIA

VOGAIS: DOUTOR MANUEL LAMEIRAS DE FIGUEIREDO CAMPAGNOLO, PROFESSOR
ASSOCIADO COM AGREGAÇÃO DO INSTITUTO SUPERIOR DE AGRONOMIA, ORIENTADOR

DOUTOR HUGO ALEXANDRE GOMES DA COSTA, TÉCNICO SUPERIOR DA DIREÇÃO GERAL
DO TERRITÓRIO

2024

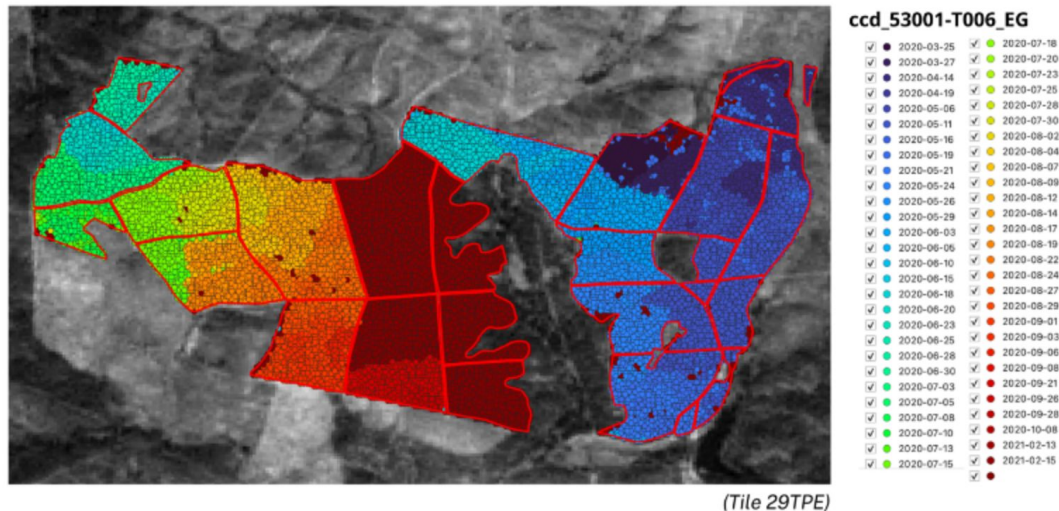


Figure 20 - Temporal distribution of the estimated clear-cut dates for parcel 53001-T006_EG, using CCDC algorithm. This parcel is shown as an example of a correct identification of clear-cutting activities by the CCDC. The pixels of dark red color represent sub-parcels that were not cut. The basemap is the NDVI derived from Sentinel-2 data for 10th of November 2020.

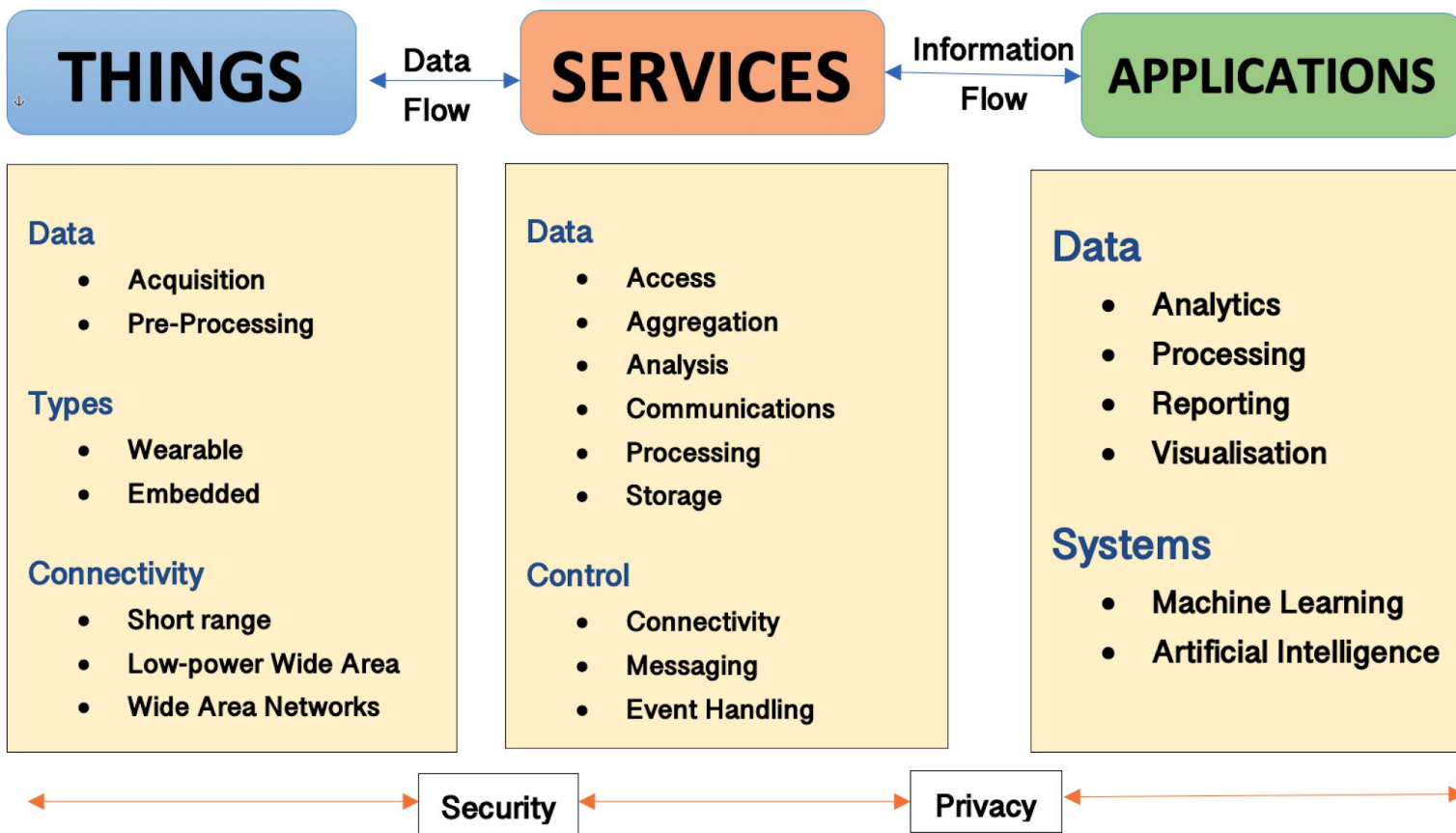


Figure 2. IoT's operational framework.
<https://doi.org/10.3390/electronics9030511>

O que são **dados**?

DADOS:

- Factos sobre coisas (texto, números)
 - propriedades, medições, descrições, etc...
- Sons, imagens, videos

INFORMAÇÃO:

- dados com contexto (dados organizados e interpretáveis)

**Algés
é a localidade
onde eu resido.**

- Os dados são uma coleção de factos, a informação contextualiza-os.
- Os dados não estão organizados, a informação está.
- Os pontos de dados são independentes, a informação pode fornecer ligações.
- Os dados não têm significado por si só. Após serem analisados, podem tornar-se informação.
- Os dados não dependem de informação. As informações dependem de dados.
- Os dados não auxiliam na tomada de decisões, a informação pode auxiliá-la.

O que são **dados**?

Algés
é a localidade
onde eu resido.

- Os dados são uma coleção de factos, a informação contextualiza-os.
- Os dados não estão organizados, a informação está.
- Os pontos de dados são independentes, a informação pode fornecer ligações.
- Os dados não têm significado por si só. Após serem analisados, podem tornar-se informação.
- Os dados não dependem de informação. As informações dependem de dados.
- Os dados não auxiliam na tomada de decisões, a informação pode auxiliá-la.

O que são **metadados**?

São dados sobre os dados

- informações estruturadas que descrevem, contextualizam e facilitam a gestão de outros dados, sem revelar o seu conteúdo
- Funcionam como etiquetas organizacionais (ex: autor, data de criação, formato, localização) essenciais para pesquisar, encontrar e utilizar informação digital.

O que é um conjunto de dados?

É o conjunto Dados + Metadados

Exemplo:
[link](#)

Dados

	A	B	C	D	E	F	G	H	I	J
1	Data	Fonte de	Link fonte	Ano de re	Idigo_seto	Descrição	Idigo_seto	Descrição	Indicador	Valor
2	3/15/2024	Submissão	https://apan	1990	1.	Energia	A.1.	Combustão	*kt CO2 eq	16,415
3	3/15/2024	Submissão	https://apan	1990	1.	Energia	A.2.	Combustão	*kt CO2 eq	8,994
4	3/15/2024	Submissão	https://apan	1990	1.	Energia	A.3.	Combustão	*kt CO2 eq	10,820
5	3/15/2024	Submissão	https://apan	1990	1.	Energia	A.4.	Combustão	*kt CO2 eq	4,139
6	3/15/2024	Submissão	https://apan	1990	1.	Energia	A.5.	Combustão	*kt CO2 eq	97
7	3/15/2024	Submissão	https://apan	1990	1.	Energia	B.	Emissões fu	*kt CO2 eq	218

Metadados

Emissões nacionais de gases com efeito de estufa (GEE) por setor - Inventário Nacional

Descrição

São apresentadas as estimativas de emissões de gases com efeito de estufa (GEE) conforme submetidas pelo Inventário Nacional Emissões por Fontes e Remoção por Sumidouros de Poluentes Atmosféricos (INERPA), desde 1990, para os setores de energia, processos industriais, agricultura, usos e ocupação de solos e resíduos.

Produtor



Agência Portuguesa do Ambiente

Última atualização

28 de maio de 2024

Licença

[Creative Commons Attribution 4.0 - CC BY 4.0](#)

Qualidade dos metadados

- Formatos de ficheiro não normalizados
- Cobertura de tempo não fornecida
- Cobertura espacial não fornecida

Dados para a Ciência de Dados

Propriedade dos dados

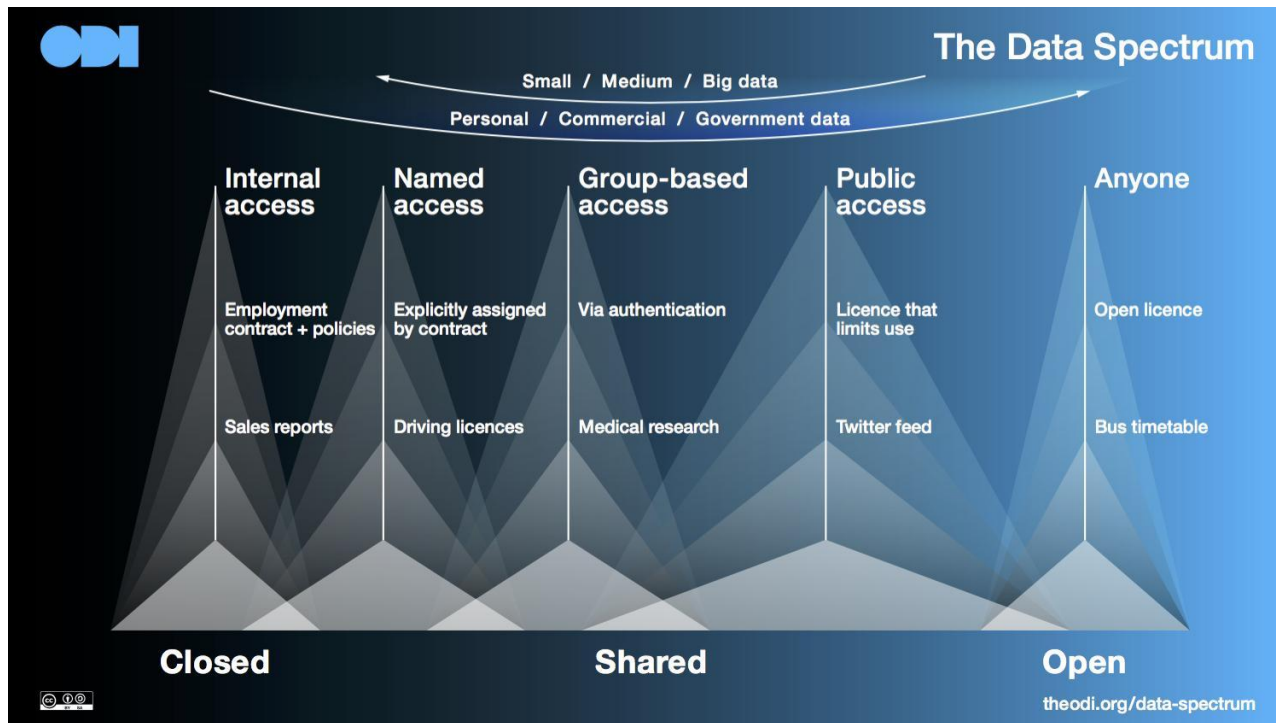
- Dados privados
 - Confidenciais
 - Informações privadas ou pessoais
 - Dados comercialmente sensíveis
- Dados abertos
 - Instituições científicas
 - Governos
 - Organizações
 - Empresas



economia
sociedade
saúde
transportes
meio ambiente

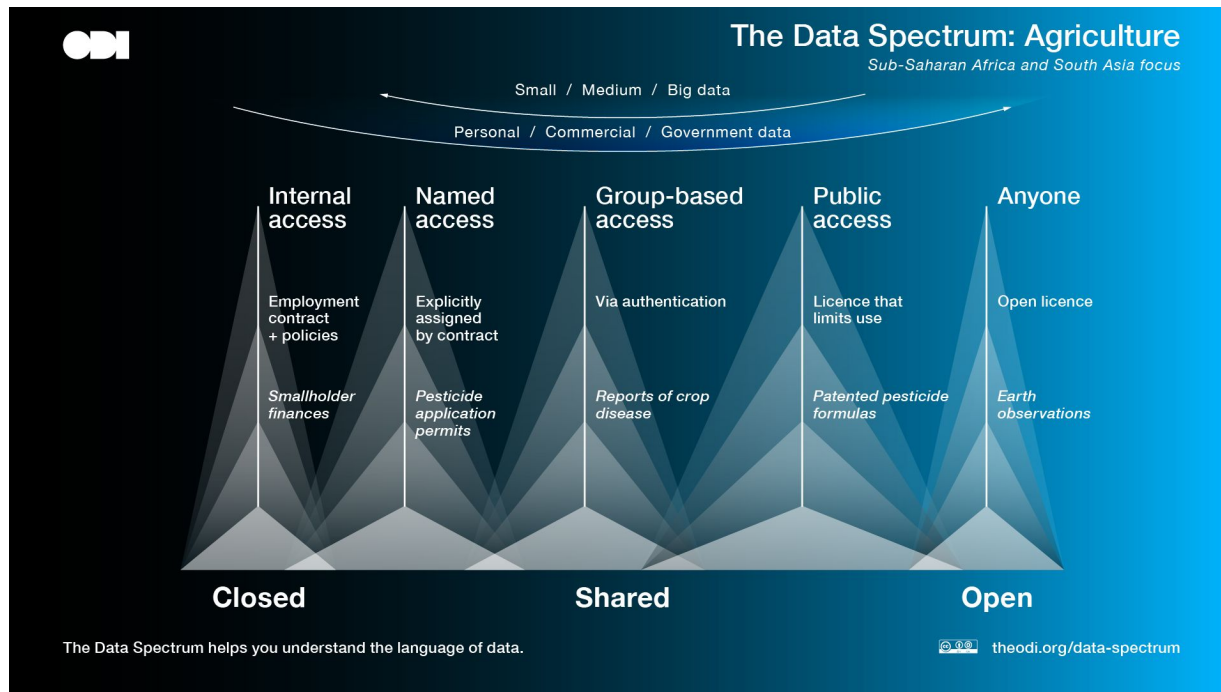
Espectro dos dados

<https://theodi.org/insights/tools/the-data-spectrum/>

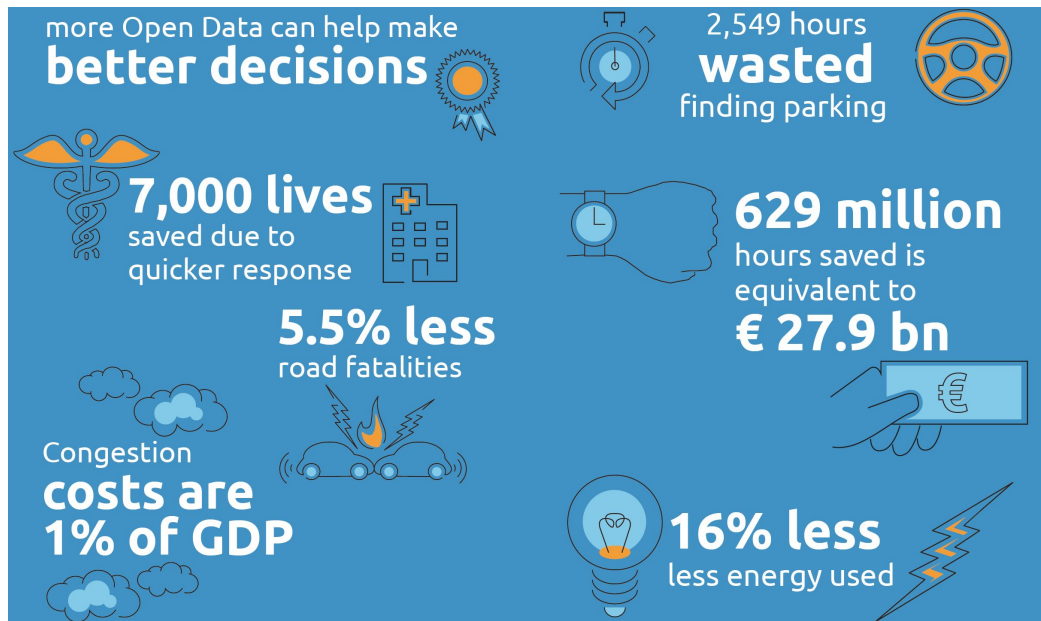


Espectro dos dados

<https://theodi.org/insights/tools/the-data-spectrum/>



Motivações para a adoção de Dados Abertos



data.europa.eu

Licences

Creative Commons (CC)



CC Zero (CC0) –



Attribution (BY) –



Attribution + Noncommercial uses (BY-NC) –



Attribution + No derivatives (BY-ND) –



Attribution + Share under the same terms (BY-SA) –



Attribution + Noncommercial uses + No derivatives (BY-NC-ND) –



Attribution + Noncommercial uses + Share under the same terms (BY-NC-SA) –



5 Star Open Data

<https://5stardata.info/en/>

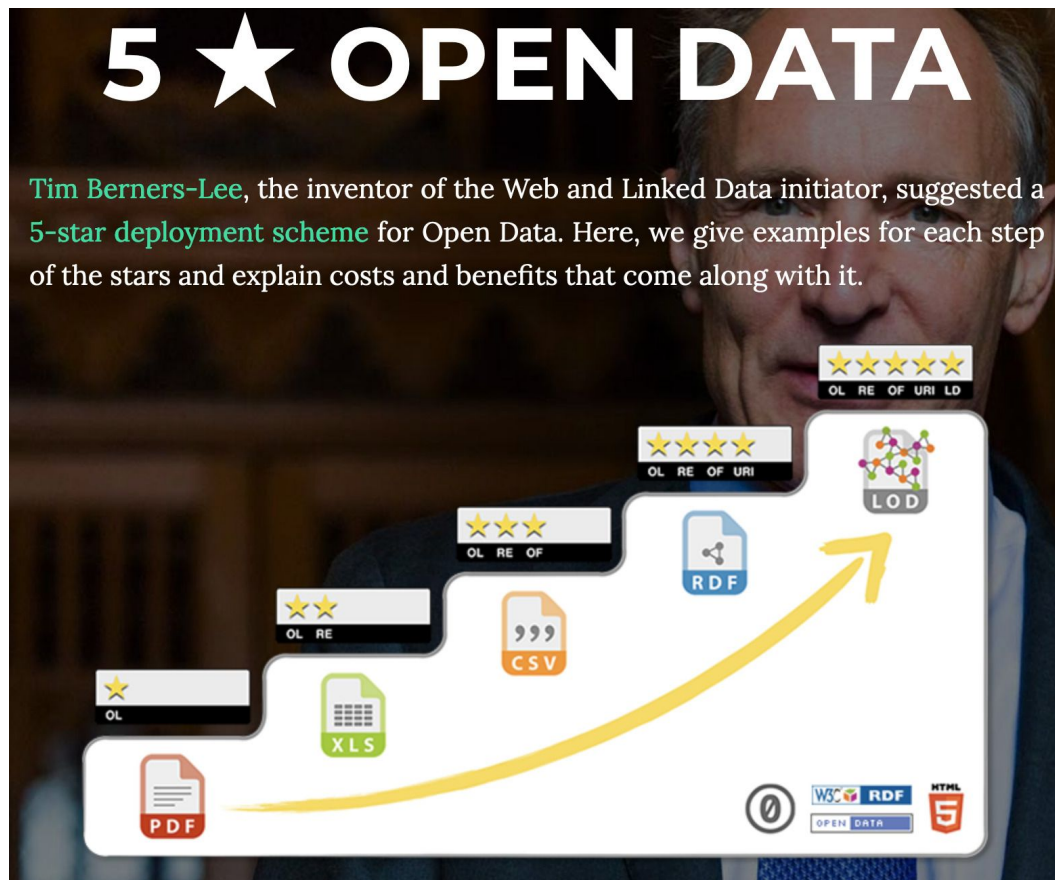
★ - an open licence

★★ - re-usable format (structured)

★★★ - open format (not proprietary, e.g. excel)

★★★★ - use identifiers to reference, use URLs to identify things, so that people find it

★★★★★ - link your data to other people's data to provide context



Dados abertos - wrap up

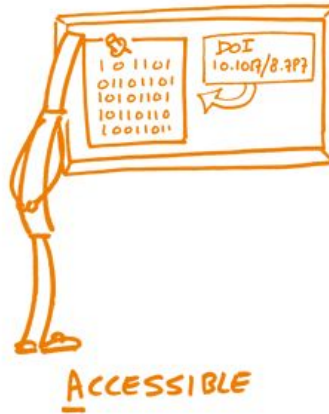
Os dados são abertos se qualquer pessoa puder aceder, utilizar e partilhar.

- Sem limitações de utilização.
- Gratuito para usar, modificar, combinar e partilhar, incluindo comercialmente.
- Gratuito para usar, mas pode haver um custo de acesso.
- Existem custos associados à criação, manutenção e publicação de dados utilizáveis.
- Taxa relacionada com o custo de reprodução, mas é geralmente insignificante.
- Gratuito para usar, reutilizar e distribuir, incluindo comercialmente.
- O valor dos dados abertos é medido pela forma como podem ser utilizados, e não pela forma como são disponibilizados.
- Devem ser estruturados e legíveis por máquina.

European Commission "What is open data?"
<https://data.europa.eu/elearning/en/#/id/co-01>

FAIR data principles

FAIR DATA PRINCIPLES



<https://doi.org/10.5281/zenodo.1212496>

Princípios de dados FAIR

To be **Findable** (Localizáveis)

Para tornar os dados localizáveis, é recomendado:

F.1. A atribuição de um **identificador único global e persistente** aos (meta)dados;

F.2. A descrição dos dados com metadados pormenorizados.

F.3. Os metadados incluem de forma clara e explícita o identificador dos dados que descrevem

F.4. O registo ou a indexação dos (meta)dados num recurso pesquisável.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, sdata201618.
<https://doi.org/10.1038/sdata.2016.18>

Identificadores Persistentes

O que é um Identificador Persistente (PID)?

É um identificador duradouro (persistente) que fornece uma referência única a "algo" - um documento, um ficheiro, um objeto digital, um registo numa base de dados - a ser referenciado de forma inequívoca.

Ex.: o seu número de identidade, um DOI, ORCID IDs, ISBN, GUID

Porque é que os PIDs são úteis?

Eles são inequívocos. E, no contexto da internet, são acionáveis, ou seja, podem ser resolvidos para ligar ao item a que fazem referência. Ex.:

<https://orcid.org/0000-0002-8351-4028>

Identificadores Persistentes

O que significam os PID de valor?

Idealmente, nada. O valor deve ser opaco, sem significado. Desta forma, haverá menos tentação de o alterar quando o contexto mudar. Não há necessidade de ser legível por humanos.

Que propriedades devem ter os PID?

- Deve ser globalmente único.
- Deve existir indefinidamente (persistente).
- Deve ser opaco.
- Deve ser de geração independente.
- Deve ser acionável (ligar ao objeto a que faz referência).

Princípios de dados FAIR

To be **Accessible (Acessíveis)**:

Tornar os dados acessíveis significa que:

A.1. Os (meta)dados são recuperáveis através do seu identificador, mediante um protocolo de comunicações normalizado.

A.1.1. O protocolo de comunicações é aberto, gratuito e universalmente implementável.

A.1.2. O protocolo de comunicações permite um procedimento de autenticação e autorização, quando necessário.

A.2. Os metadados permanecem acessíveis, mesmo se os dados já não estiverem disponíveis.

Princípios de dados FAIR

To be **Interoperable (Interoperáveis)**:

Os dados serão interoperáveis se:

- I.1. Os (meta)dados usam uma **linguagem formal**, acessível, partilhada e de ampla aplicabilidade para a representação do conhecimento.
- I.2. Os (meta)dados usam **vocabulários** que seguem os princípios FAIR.
- I.3. Os (meta)dados incluem referências qualificadas a outros (meta)dados.

Princípios de dados FAIR

To be **Reusable (Reutilizáveis)**:

Assegurar que os dados são reutilizáveis significa que:

R.1. Os (meta)dados têm uma pluralidade de atributos precisos e relevantes.

R.2. Os (meta)dados são disponibilizados com uma licença clara e acessível de uso dos dados.

R.3. Os (meta)dados são associados à sua proveniência.

R.4. Os (meta)dados cumprem normas relevantes da comunidade disciplinar.

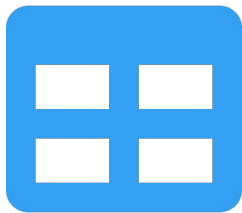
<https://tinyurl.com/mru5wc2r>



Tipos de dados

Estruturados

Todos os registos partilham um **esquema comum**, ou seja, todos têm valores para o **mesmo atributo**.



Exemplo: dados tabulares, csv, tsv

Semi-estruturados

Cada registo pode ter um esquema diferente (também chamado, por vezes, de dados heterogéneos).

```
[ {
  "Indicador": "0011007",
  "IndicadorSig": "Average value of total standard output by agricultural
holding (4) by Geographic localization (NUTS -2013); Decennial - Statistics
Portugal, Agriculture census - 2019",
  "MetaInfo": "https://www.ine.pt/bddXplorer/bddocs/minfo.jsp?
var_cd=011007&imgae=0",
  "DataExtração": "2022-09-13T14:55:06.885+01:00",
  "DataAtualizacao": "2021-10-02",
  "UltimoAtualizacao": "2019",
  "Dados": {
    "1999": {
      "geocod": "11C181318",
      "geodis": "União das freguesias de Freixil e Miondes",
      "valor": "4077.1"
    }, {
      "geocod": "11A010934",
      "geodis": "União das freguesias de Lobão, Gilo, Louredo e Guisande",
      "valor": "5750.1"
    }, {
      "geocod": "11A010935",
      "geodis": "União das freguesias de Santa Maria da Feira, Travanca,
Santinho e Espargoso",
      "valor": "29903.5"
    }, {
      "geocod": "11D180122",
      "geodis": "União das freguesias de São Romão e Santiago",
      "valor": "9240.9"
    }, {

```

Exemplo: email, json, xml, tree data, graph data

Não-estruturados

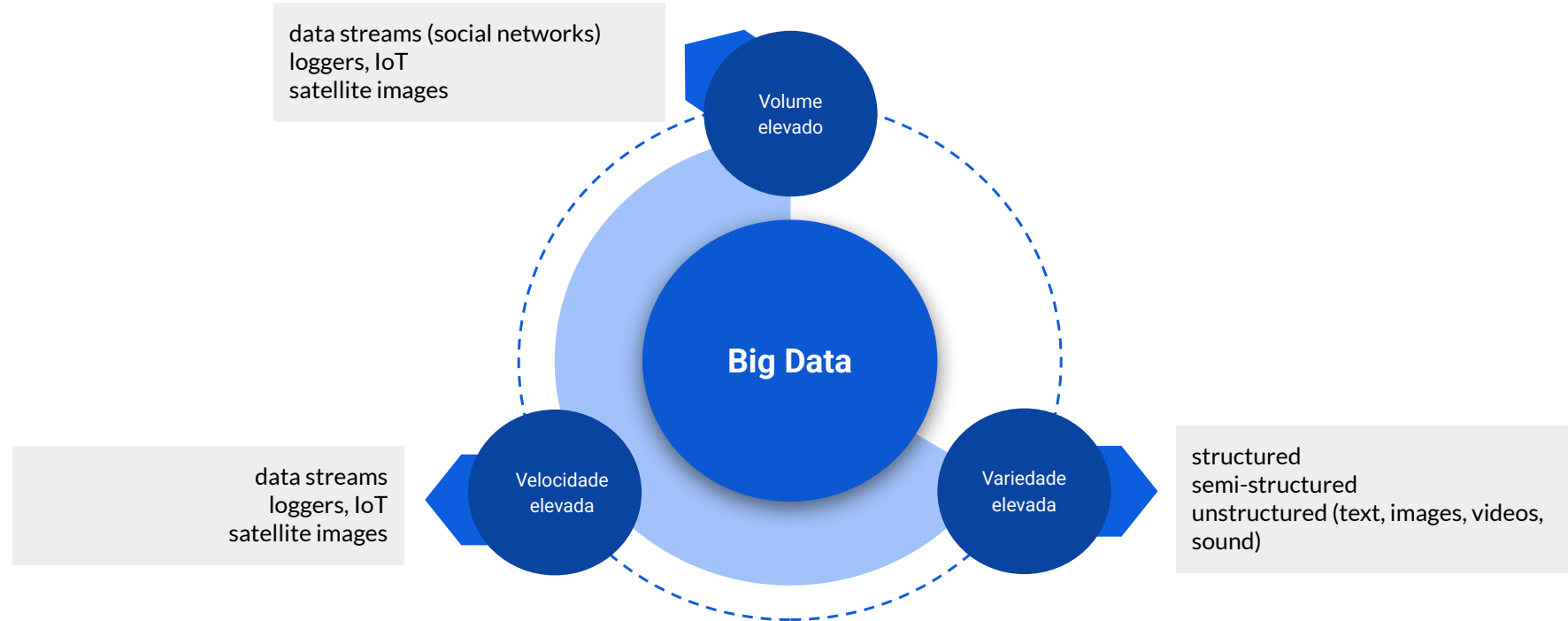
A **estrutura não é explícita**, como acontece nos dados estruturados e semiestruturados (i.e., não existem etiquetas ou marcadores para separar os registos em atributos).



<https://unsplash.com/photos/Wh9ZC4727e4>

Exemplo: news, social network post, etc.

Big Data



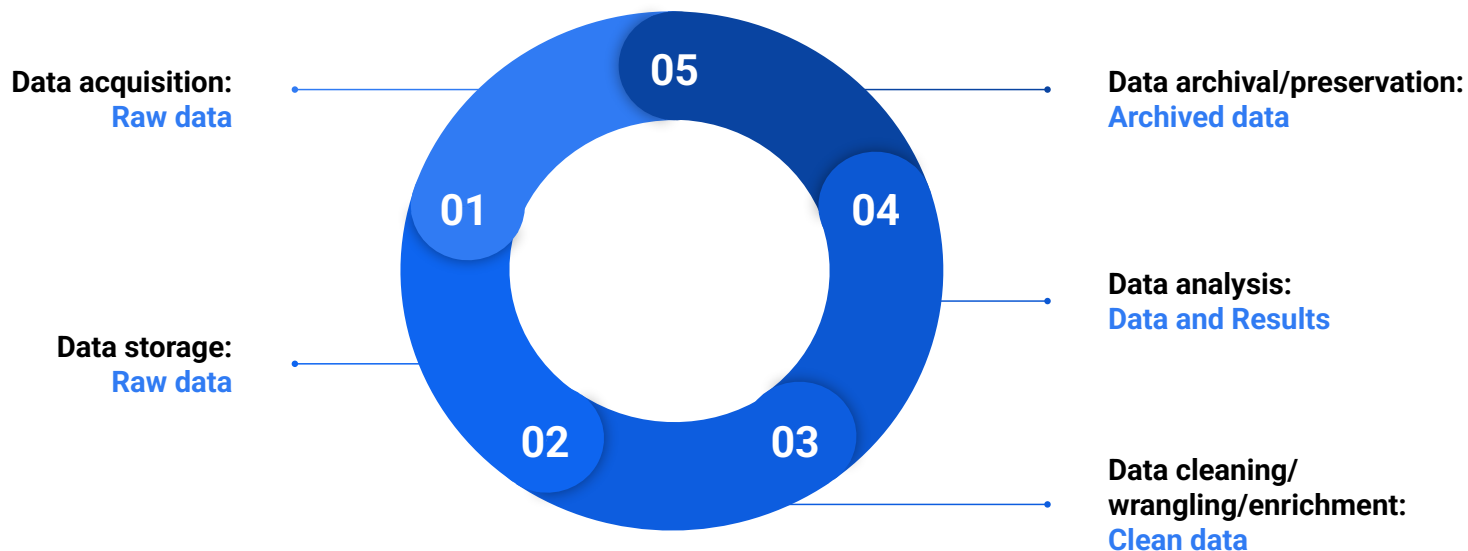
Big Data - exemples

01	Text	• continuous text • structured text • collection of texts, tags, etc
02	Graphics	• road maps • 3D graphics • technical drawing, etc
03	Image	• photographs • satellite images • x-ray image, ect
04	Audio	• language • music • sounds, animal sounds, etc
05	Video	• film • animation • ads, phone conferences, etc



Ciclo de vida dos dados

Ciclo de vida dos dados - transformações aplicadas aos dados e aos estados pelos quais os dados passam como resultado dessas transformações.



Infraestructura computacional

Infraestrutura computacional

As **infraestruturas físicas e na nuvem** fornecem os ambientes de computação e armazenamento necessários para **processar, analisar e gerir dados**.

Cada uma tem as suas próprias características, benefícios e casos de utilização típicos.

As **máquinas virtuais** (VMs) ou **containers** desempenham um papel crucial em ambos os ambientes, permitindo ambientes de computação **flexíveis, isolados e escaláveis**.



Infraestrutura Física

A **infraestrutura física** refere-se ao hardware local (*on-premises*), como servidores, dispositivos de armazenamento e equipamentos de rede pertencentes e mantidos pela organização

Componentes:

- **Servidores:** Executam fluxos de trabalho de análise de dados.
- **Sistemas de armazenamento:** Armazenam conjuntos de dados e resultados (ex.: NAS, SAN).
- **Rede:** Garante a conectividade interna e externa.
- **Arrefecimento e energia:** Mantêm o funcionamento estável.

Pros:

- Controlo total sobre o hardware e a segurança.
- Desempenho previsível (sem recursos partilhados).
- Investimento único em hardware

Cons:

- Elevado custo inicial e manutenção contínua.
- Escalabilidade limitada.
- Requer pessoal de TI para suporte

Infraestrutura na nuvem

A **infraestrutura na nuvem** fornece recursos computacionais virtualizados através da internet. Inclui a Infraestrutura como Serviço (IaaS), a Plataforma como Serviço (PaaS) e o Software como Serviço (SaaS).

Exemplos de Cloud Providers:

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure

Serviços principais:

- Compute instances (e.g., EC2 in AWS): Virtual machines with configurable CPU, memory, etc.
- Armazenamento: Object stores (e.g., S3), block storage, databases.
- Managed services: BigQuery, Databricks, SageMaker.

Pros:

- Escalável e elástico (pay-as-you-go)
- Implantação rápida.
- Disponibilidade e redundância globais.
- Reduz a necessidade de manutenção.

Cons:

- Custos contínuos.
- Dependência de acesso à internet.
- Preocupações com a privacidade e o controle de dados.

Máquinas Virtuais (VMs)

As **máquinas virtuais (VMs)** são emulações definidas por software de computadores físicos que são executadas num ambiente isolado. Podem ser implantadas tanto em servidores físicos como em plataformas de cloud.

Usos na análise de dados:

- Executar diferentes ambientes de software (por exemplo, Python, R, MATLAB).
- Isolar tarefas ou utilizadores.
- Hospedar servidores Jupyter, RStudio ou pipelines de análise.
- Gerir ambientes reproduzíveis para experiências.

Benefícios:

- Isolamento: Cada máquina virtual opera de forma independente, minimizando as interferências.
- Flexibilidade: Fácil de criar snapshots, clonar ou migrar.
- Segurança: Ambientes de sandbox para testar código ou analisar dados sensíveis.

Repositórios de dados

Sistemas de armazenamento centralizados concebidos para recolher, organizar, preservar e fornecer acesso a conjuntos de dados para fins de pesquisa, análise e partilha, com metadados, documentação e controlos de acesso apropriados

Tipos de Repositórios

- **Institucionais:** Universidades e unid de investigação
- **Específicos:** GenBank, ICPSR, dados climáticos
- **Gerais:** Zenodo, Figshare, Dryad
- **Comerciais:** AWS Open Data, Google Dataset

Principais características

- Sistemas de metadados e documentação robustos
- Controlo de versão e rastreio da proveniência dos dados
- Recursos de pesquisa e descoberta
- Controlos de acesso e autenticação
- Identificadores persistentes (DOIs)

Funções principais

- Preservação de dados a longo prazo
- Garantia e validação da qualidade
- Integração com ferramentas de análise
- Facilitação da colaboração
- Migração e manutenção de formatos

Repositórios de dados

Fontes de dados

- United Nations - UNDATA - <http://data.un.org/>
 - many databases from partner organisations: <http://data.un.org/Partners.aspx>
- FAO (several databases) - <https://www.fao.org/statistics/databases/en/>
 - FAOSTAT - <https://www.fao.org/faostat/en/#data>
- Copernicus Data Space Ecosystem- Satellite data - <https://dataspace.copernicus.eu/>
- European data - <https://data.europa.eu/en>
- US Open Data - <https://data.gov/>
- Dados abertos da administração pública - Portugal - <https://dados.gov.pt/pt/>
- Data Portals - <http://datacatalogs.org/>
- Kaggle - online community - find and publish data sets - <https://www.kaggle.com/>
- Google Data Search - <https://datasetsearch.research.google.com/>

Repositórios de dados

Casos de uso para Portugal

O fórum GDI é uma comunidade de prática portuguesa dedicada aos repositórios de dados.

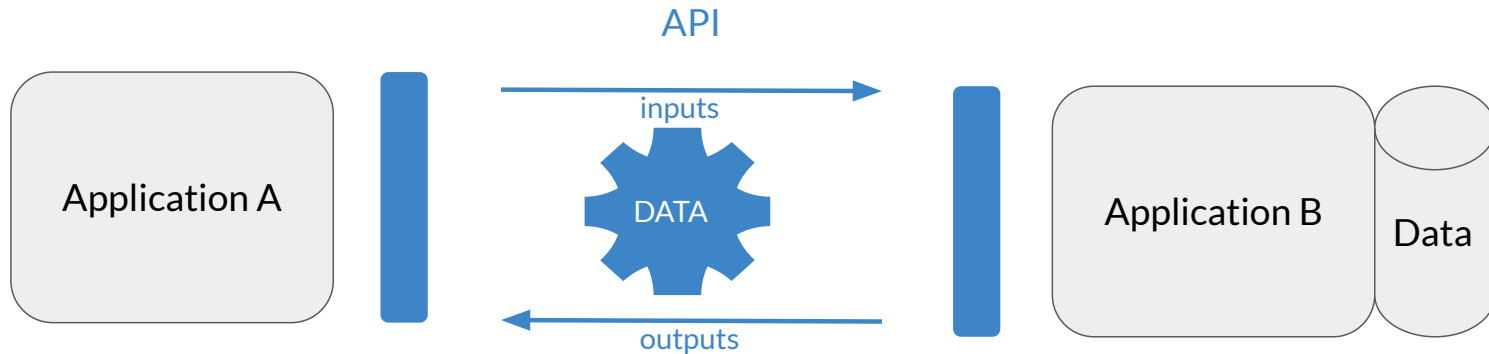
Casos de uso compilados:

<https://zenodo.org/communities/forumgdi/records?q=&f=subject%3ACaso%20de%20estudo&l=list&p=1&s=10&sort=newest>

Serviços de dados Web

API - Application Programming Interfaces

- Definições e protocolos para permitir a comunicação entre dois componentes informáticos.



A aplicação A não tem de "saber" nada sobre a aplicação B e vice-versa...

Como funcionam os APIs

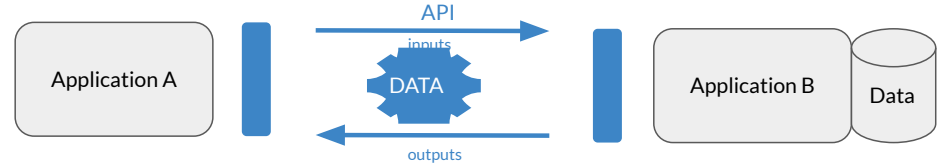
Métodos API comuns:

GET - obter dados (ler/aceder a dados)

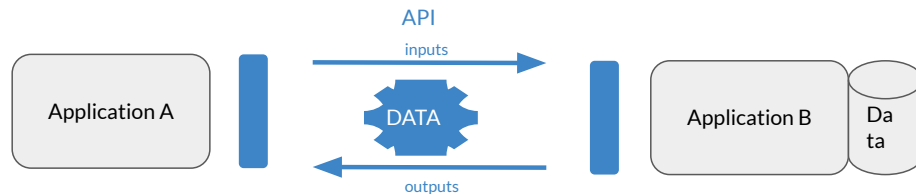
POST - Criar novos dados (escrever dados)

PUT - Actualizar dados existentes (editar dados)

DELETE - Remover dados (apagar dados)



Serviços de dados Web



API - Application Programming Interfaces

- Information is transferred through the internet, using the HTTP protocol. Data is transferred in a structured format (.g. JSON)

Exemplo: obter dados de pinheiro manso para Portugal

<https://api.gbif.org/v1/species/search?q=Pinus%20pineae&rank=SPECIES>

<https://www.gbif.org/species/search?q=Pinus%20pineae&rank=SPECIES>

result

nubKey: 5285165

<https://api.gbif.org/v1/occurrence/search?taxonKey=5285165&country=PT>

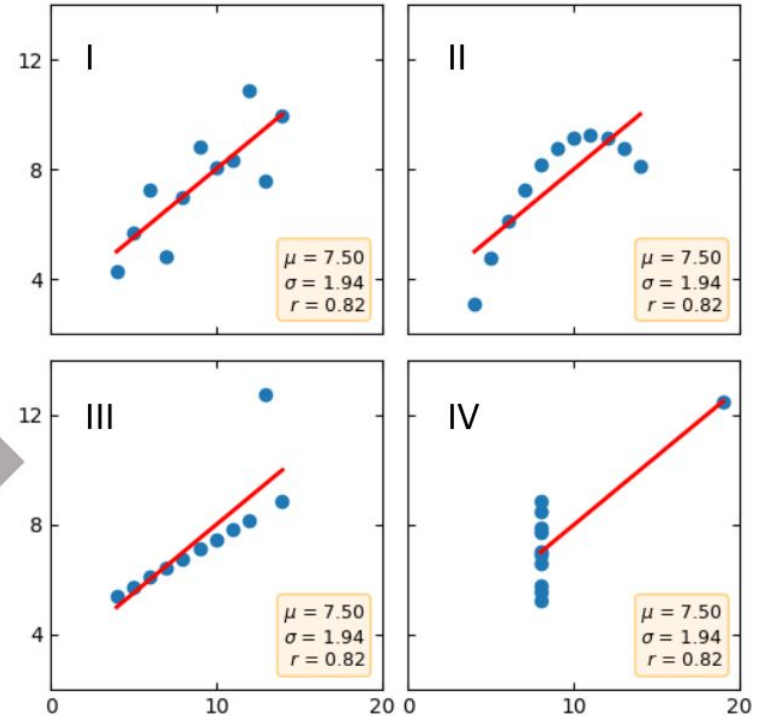
Análise de dados

Why data visualization?

Numerical calculations *per se* may be misleading.

=> We need visualization!

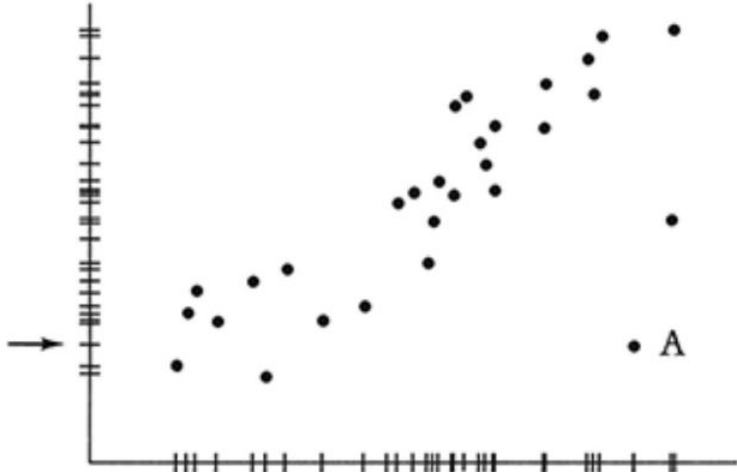
	Set I		Set II		Set III		Set IV	
	X	Y	X	Y	X	Y	X	Y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
corr	0.82		0.82		0.82		0.82	
lin. reg.	$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$	



Credits: slides by P. Segurado, [course AVDCAA](#)

Why data visualization?

Example:



Point A show up in the bivariate scatter plot but not in both marginal distributions

Credits: slides by P. Segurado, [course AVDCAA](#)

Data Analysis

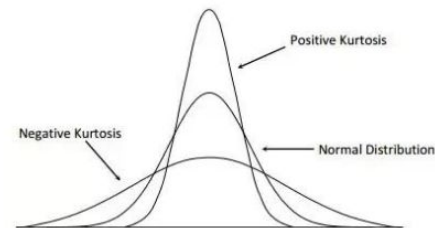
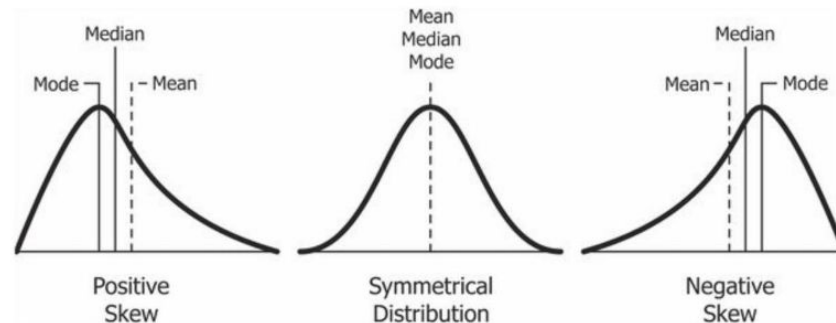
Três grandes áreas de análise de dados:

- **Estatística Descritiva (DS)** - concentra-se em resumir dados por meio de números e gráficos (por exemplo, tendência central, dispersão, distribuição).
- **Análise exploratória de dados (EDA)** - concentra-se na descoberta de novas características nos dados (por exemplo, visualização de dados; ordenação, classificação; encontrar ideias para uma teoria).
- **Análise confirmatória de dados (CDA)** – concentra-se em confirmar ou refutar hipóteses existentes (por exemplo, teste de hipóteses; testar teorias resultantes da EDA).

Estatísticos básicos

Variáveis quantitativas

- Tipo de dados(discreto ou contínuo)
- Medidas de **localização** e **tendência** central:
 - Média
 - Moda (apenas para dados quantitativos discretos)
 - Mediana
 - Percentis/quantis/quartis
- Medidas de **dispersão**
 - Desvio padrão/variância
 - Skewness
 - Kurtosis



Credits: slides by P. Segurado, [course AVDCAA](#)

Estatísticos básicos

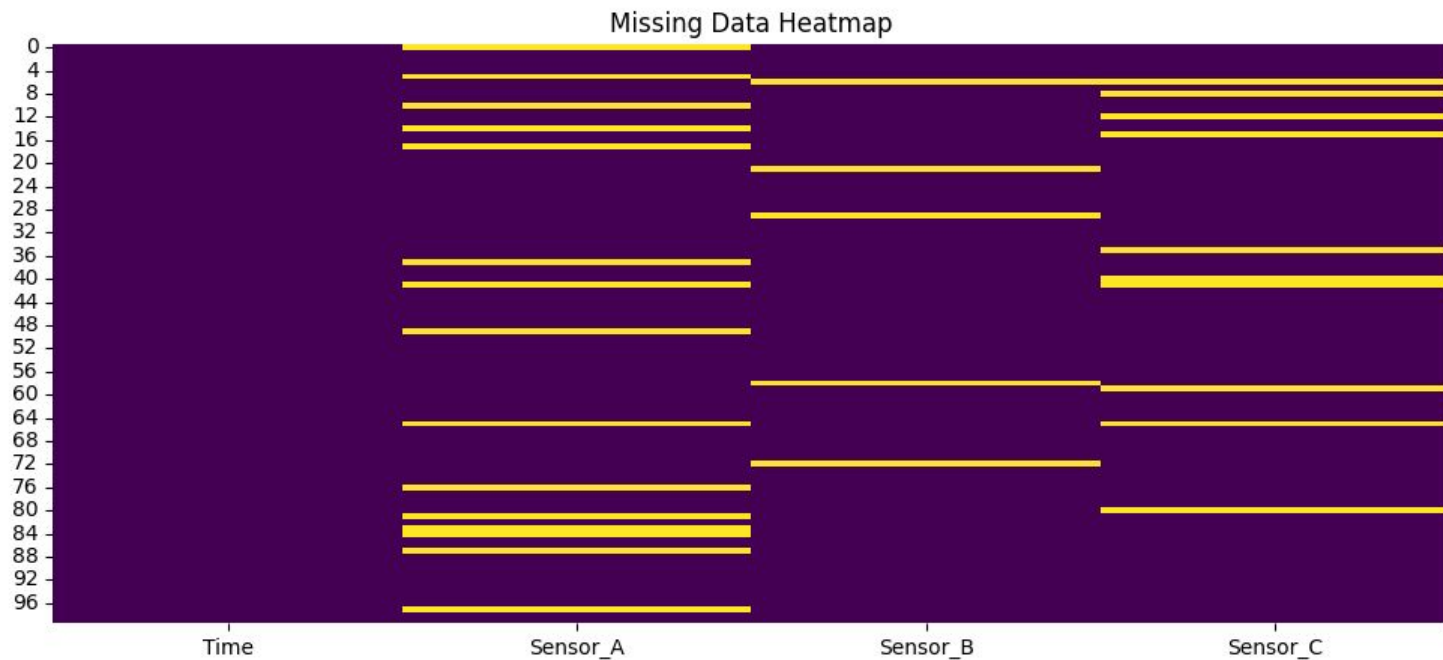
Variáveis categóricas

Estatísticos básicos:

- Tipo de dados
- Número de classes ou níveis
- Medidas de tendência central: Moda (classe ou nível mais frequente)
- Tabela de frequência (número ou proporção de casos por classe)

Trabalhar com dados em falta

Dados em falta



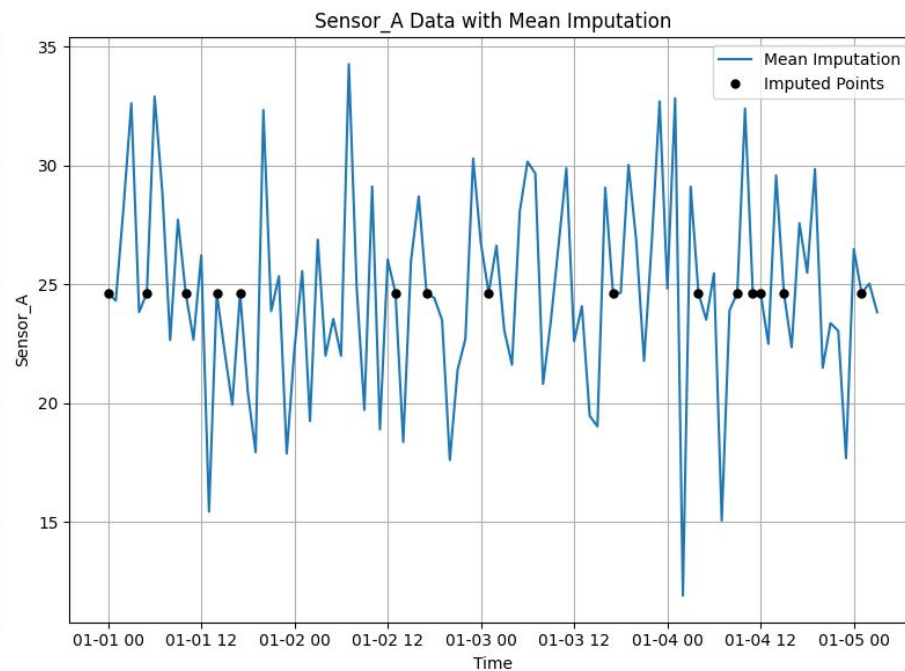
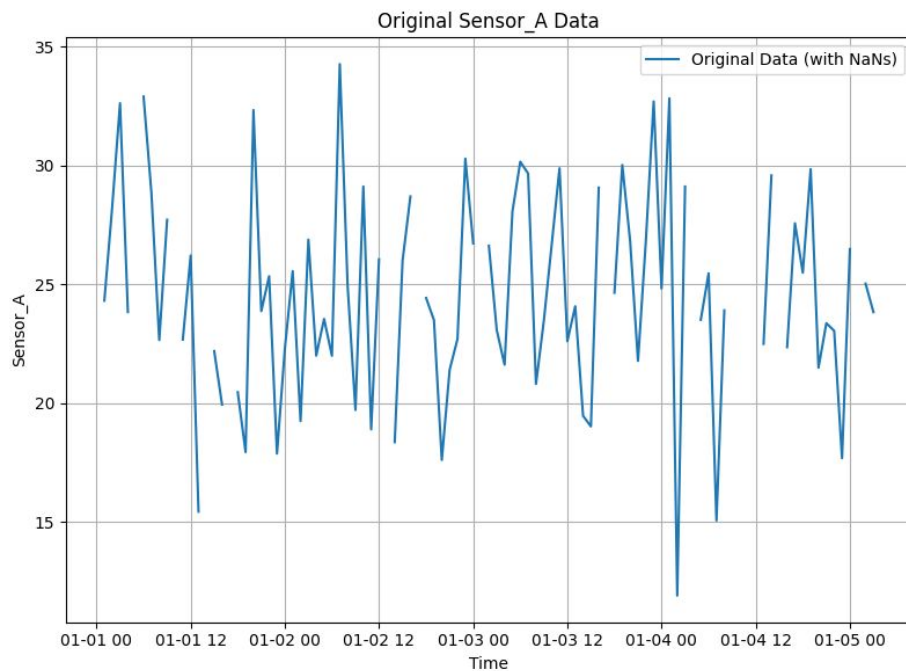
Dados em falta - métodos

Nome do método	Descrição
Imputação da média	Substitui os valores ausentes pela média dos valores não ausentes na mesma coluna.
Imputação da mediana	Substitui os valores ausentes pela mediana dos valores não ausentes na mesma coluna. É mais robusto a valores discrepantes do que a média.
Imputação de moda	Substitui os valores ausentes pelo valor mais frequente (moda) na mesma coluna. Geralmente usado para dados categóricos, mas também aplicável a dados numéricos.
Forward Fill (ffill)	Substitui os valores ausentes pelo último valor não ausente observado na coluna (Last Observation Carried Forward - LOCF). Adequado para séries temporais.

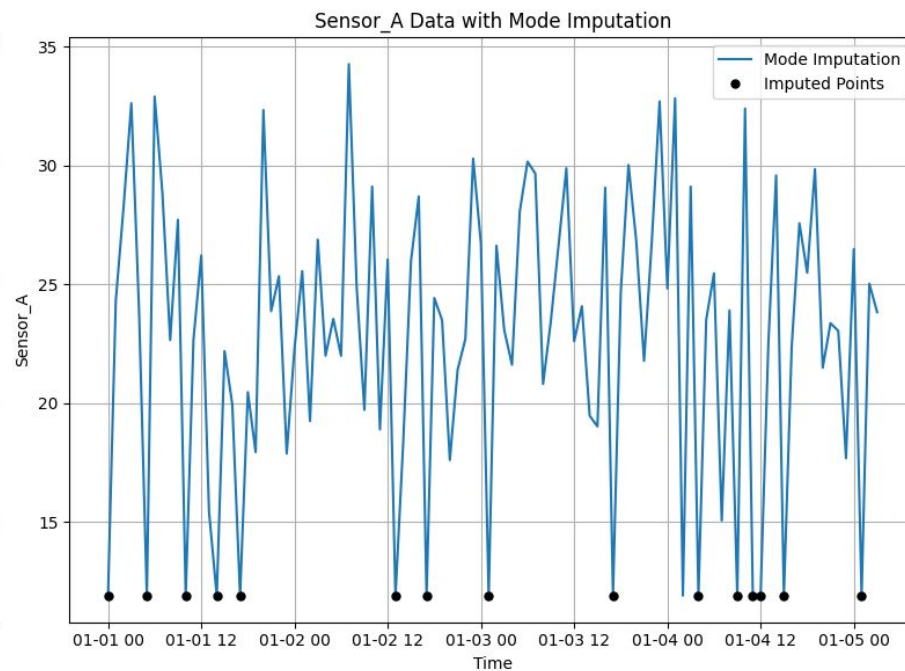
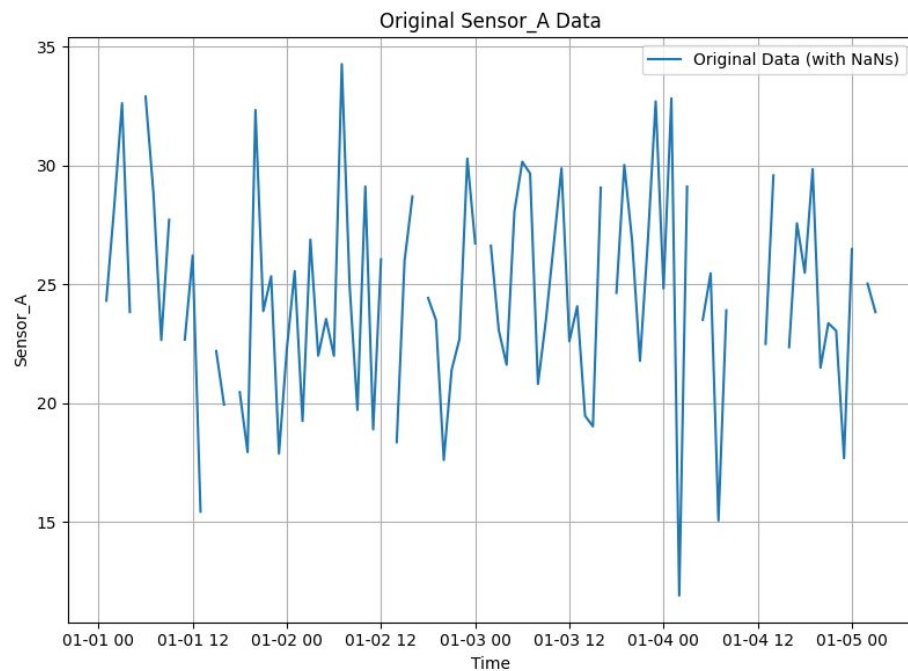
Missing data - methods

Method Name	Description
Backward Fill (bfill)	Substitui os valores ausentes pelo próximo valor não ausente observado na coluna (Next Observation Carried Backward - NOCB). Adequado para séries temporais.
Interpolação Linear	Estima valores ausentes traçando uma linha reta entre os pontos de dados não ausentes mais próximos. Excelente para dados de séries temporais onde os valores variam suavemente.
K-Nearest Neighbors (k-NN) Imputation	Preenche os valores ausentes encontrando os k nearest data points (k pontos mais próximos) com base noutras características e usa os valores para estimar o valor em falta. Considera as relações entre as características.
Iterative Imputer (MICE)	Modela cada característica com valores ausentes como uma função de outras características num processo iterativo, prevendo os valores ausentes usando um modelo de machine learning. Geralmente é mais robusto.

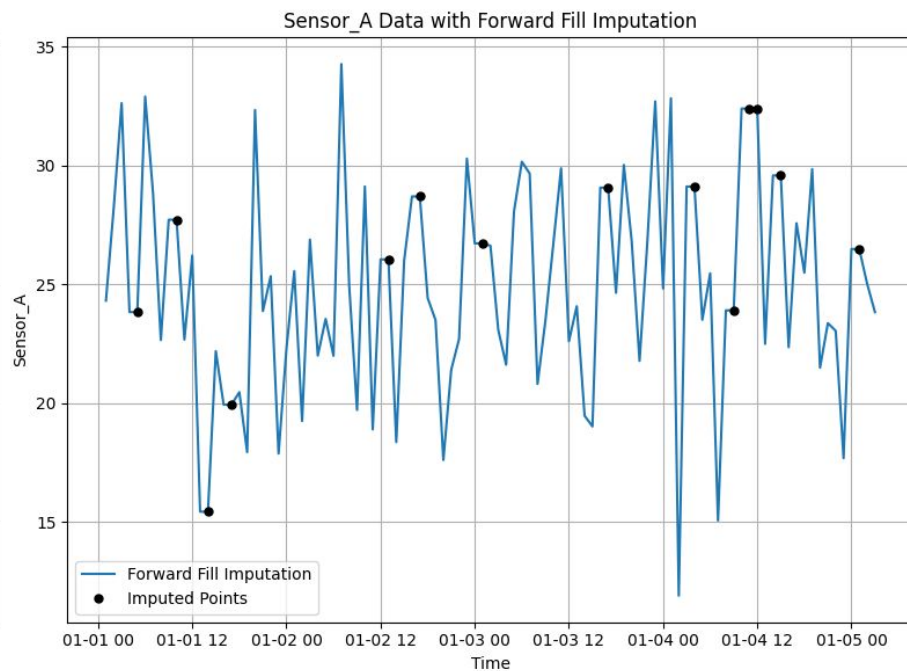
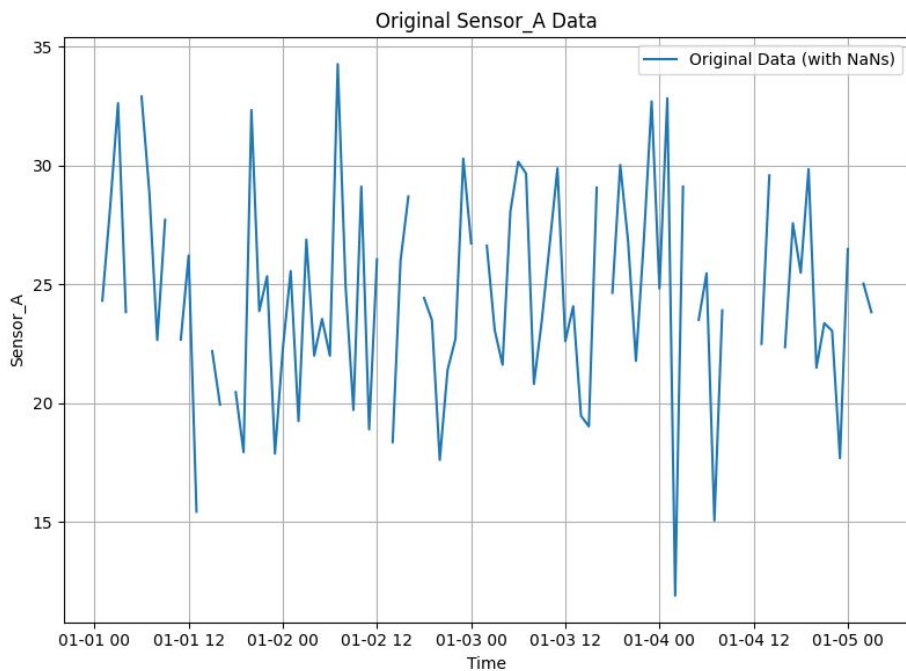
Preenchimento de dados em falta - uso da média



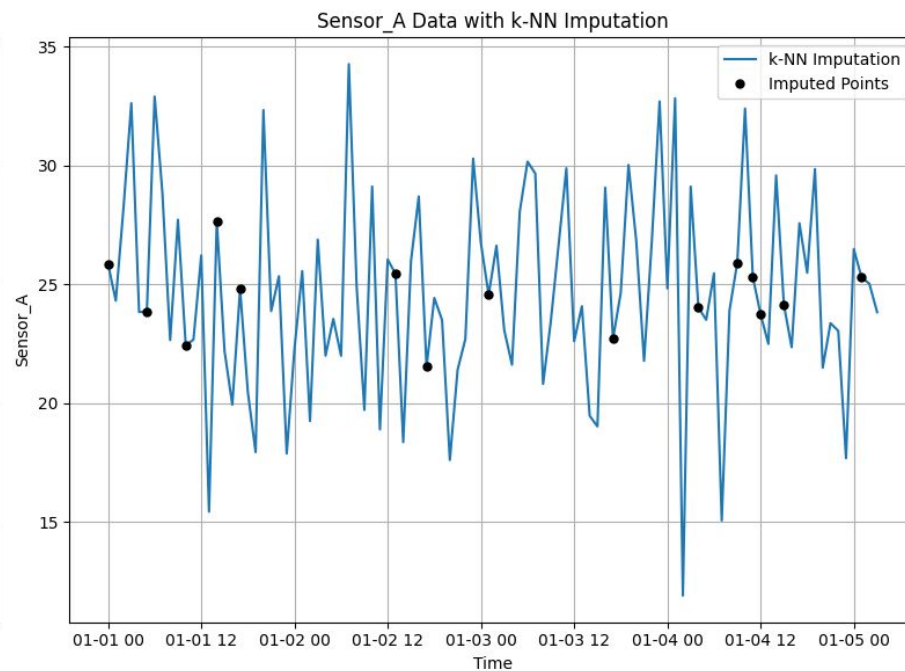
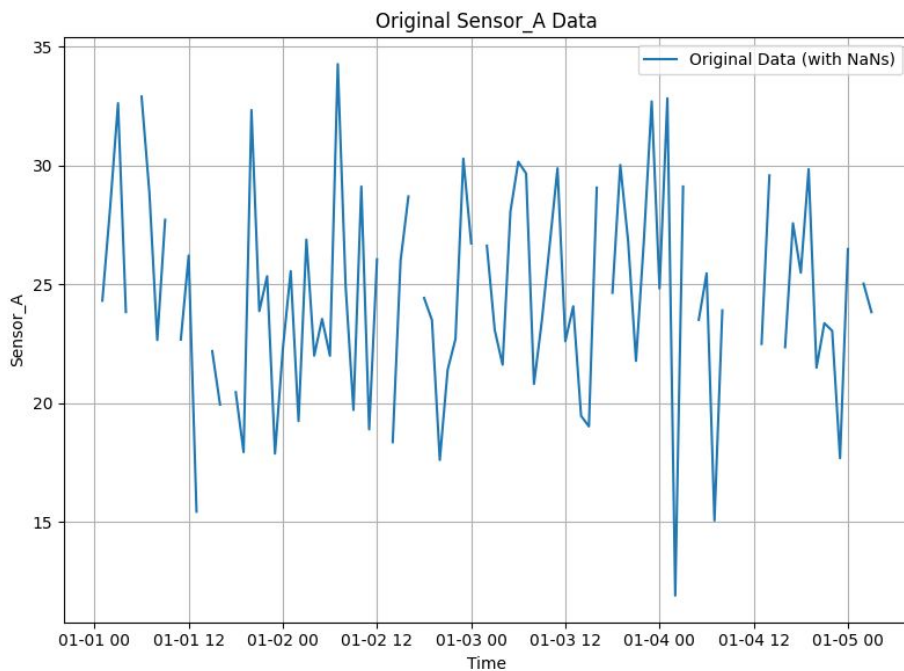
Preencimento de dados em falta - uso da moda



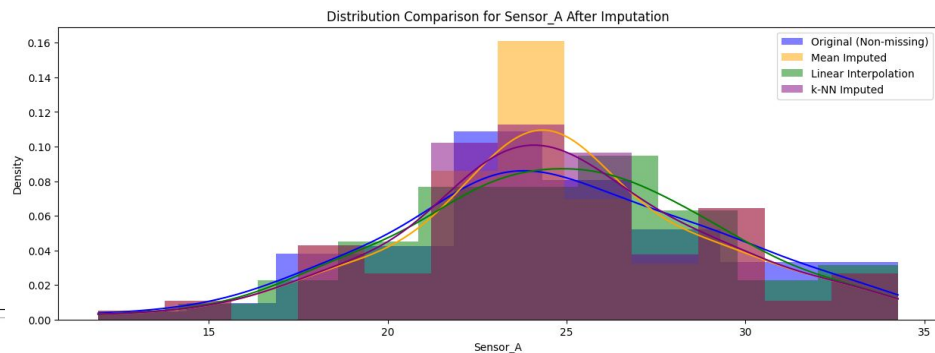
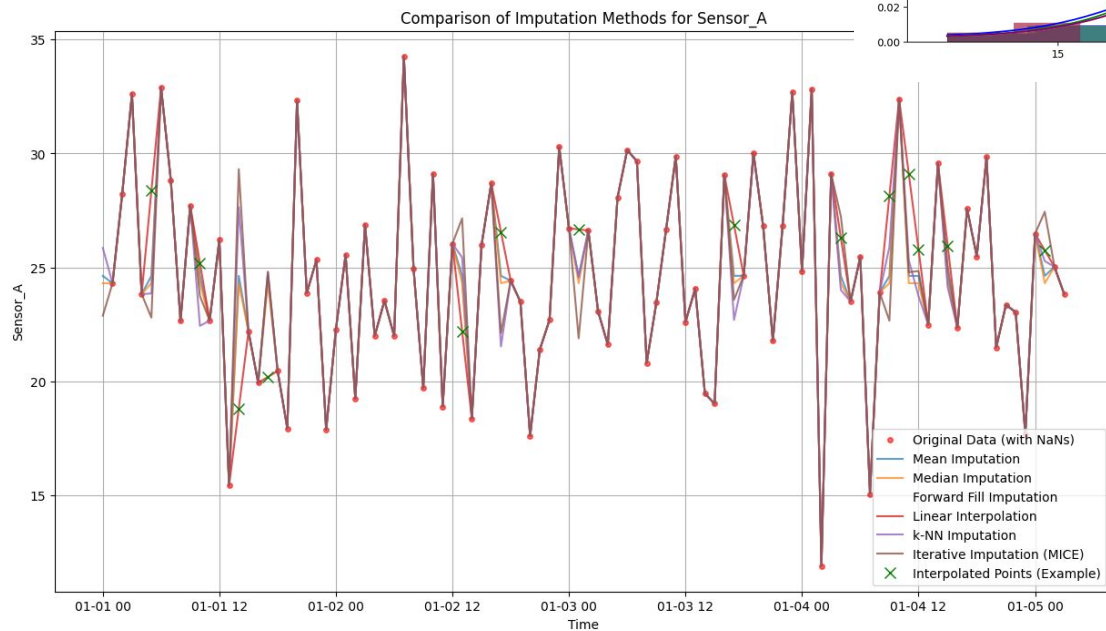
Preencimento de dados em falta - Forward fill



Preencimento de dados em falta - k-NN

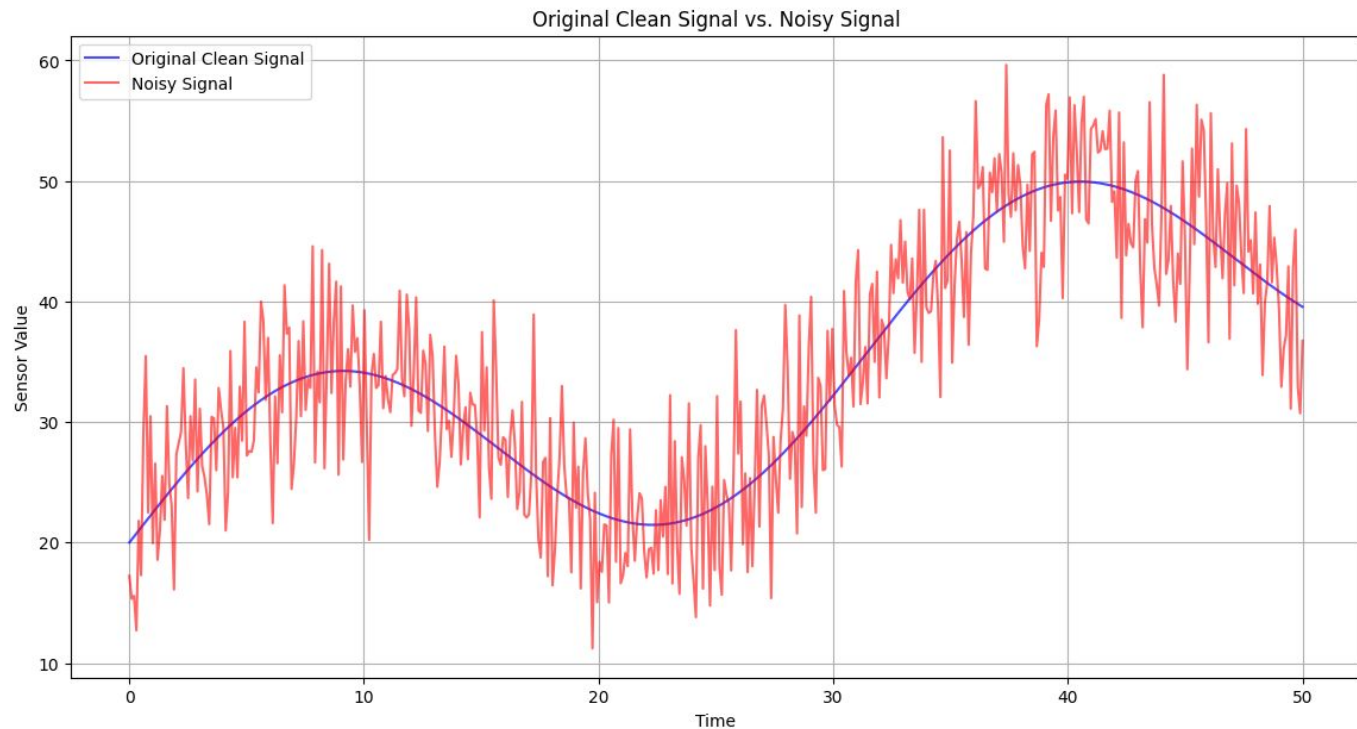


Missing data



Denoise data

Remoção de ruído em dados

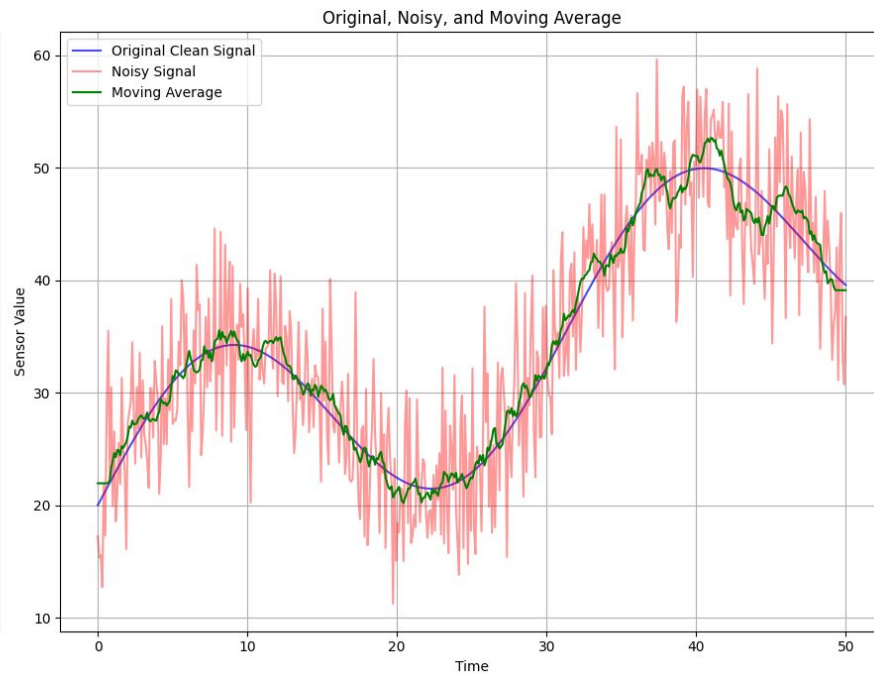
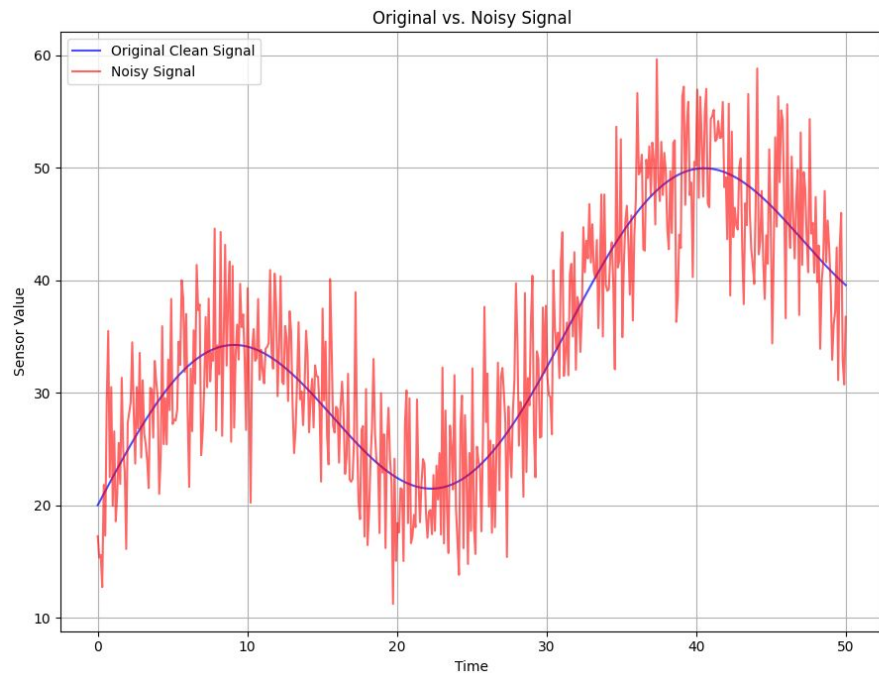


Remoção de ruído em dados

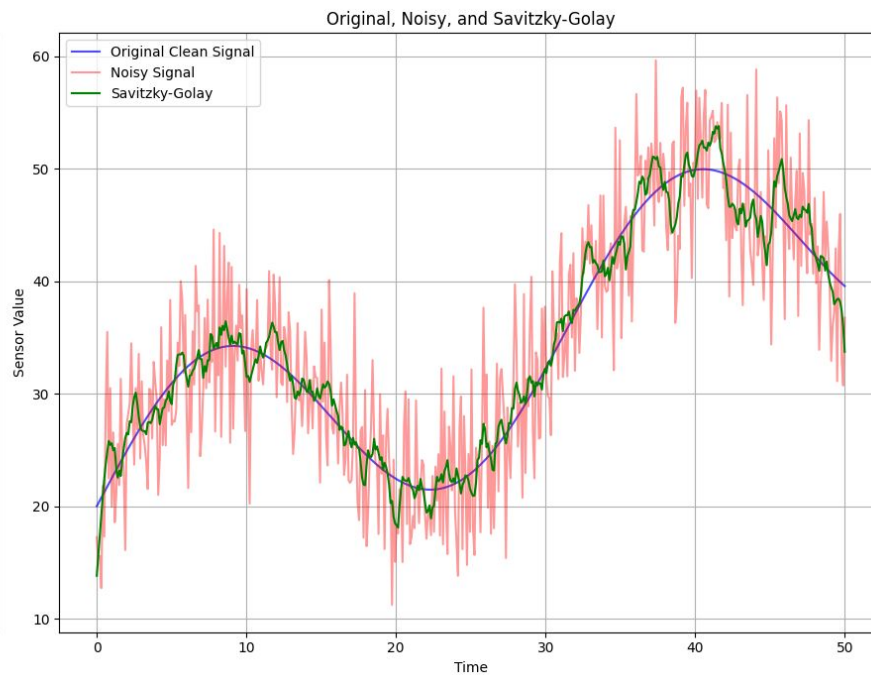
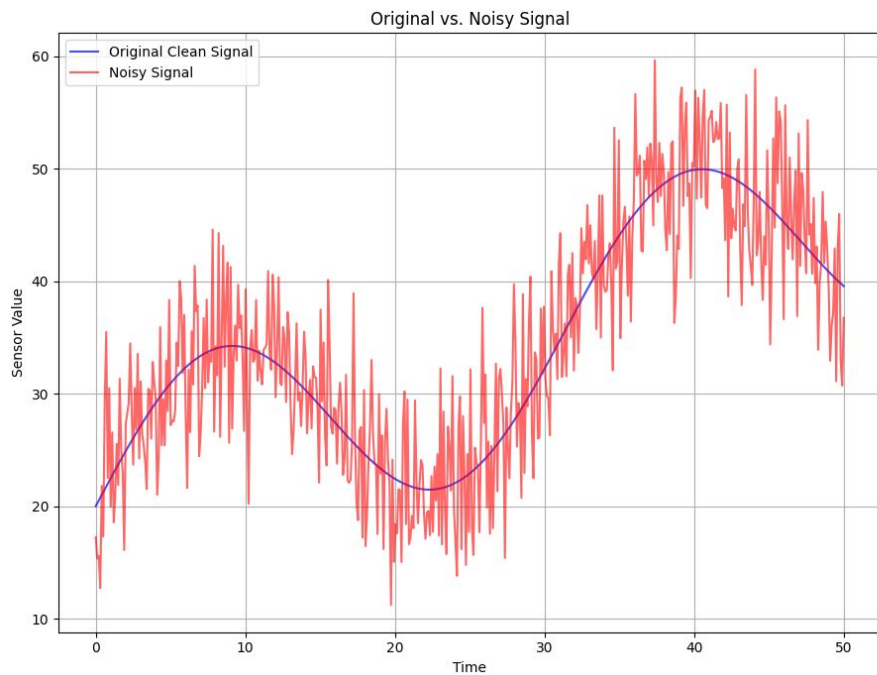
Métodos de redução de ruído:

Method Name	Descrição
Média Móvel	Suaviza os dados substituindo cada ponto pela média dele mesmo e de seus vizinhos dentro de uma janela definida. Simples, mas pode introduzir atraso.
Filtro de Savitzky-Golay	Ajusta um polinómio a uma janela de pontos de dados para estimar o valor suavizado, preservando melhor o formato do sinal do que uma média móvel simples.
Filtro da mediana	Substitui cada ponto pelo valor mediano dentro de uma janela. Excelente para remover picos ou ruído impulsivo sem desfocar as bordas.
Low-pass Butterworth Filter	Um filtro digital que permite a passagem de frequências abaixo de um determinado limite, atenuando as frequências mais altas e removendo ruídos de alta frequência.
Wavelet Denoising	Decompõe o sinal em componentes de frequência (wavelets), aplica um limiar aos componentes de ruído e reconstrói o sinal. Eficaz para ruídos complexos.

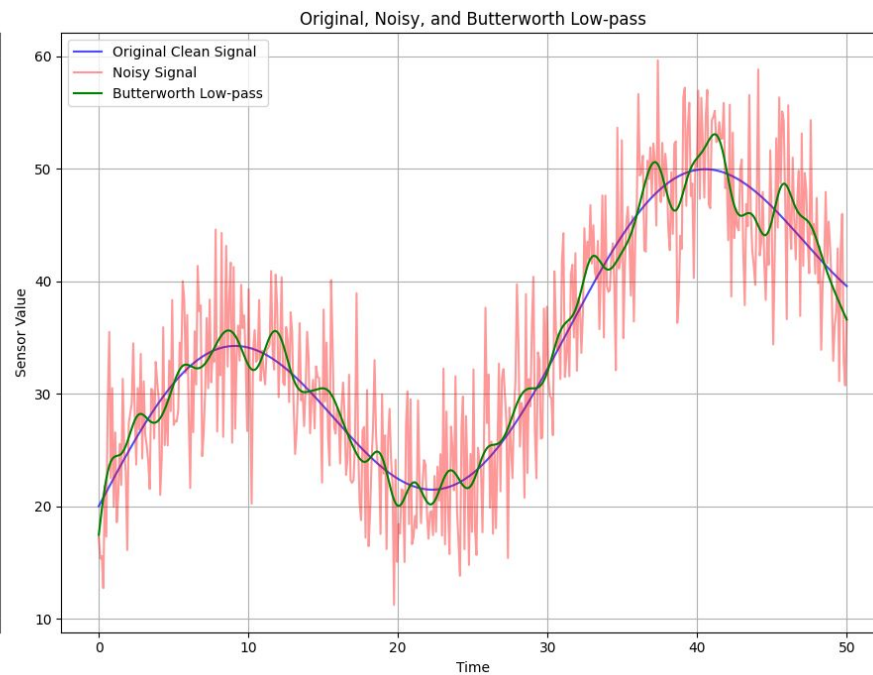
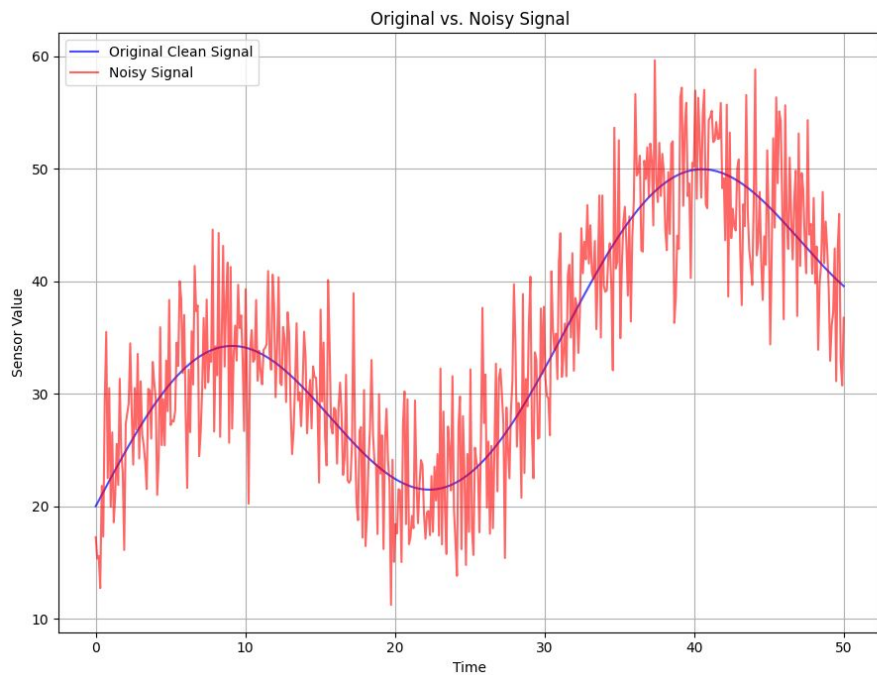
Remoção de ruído em dados



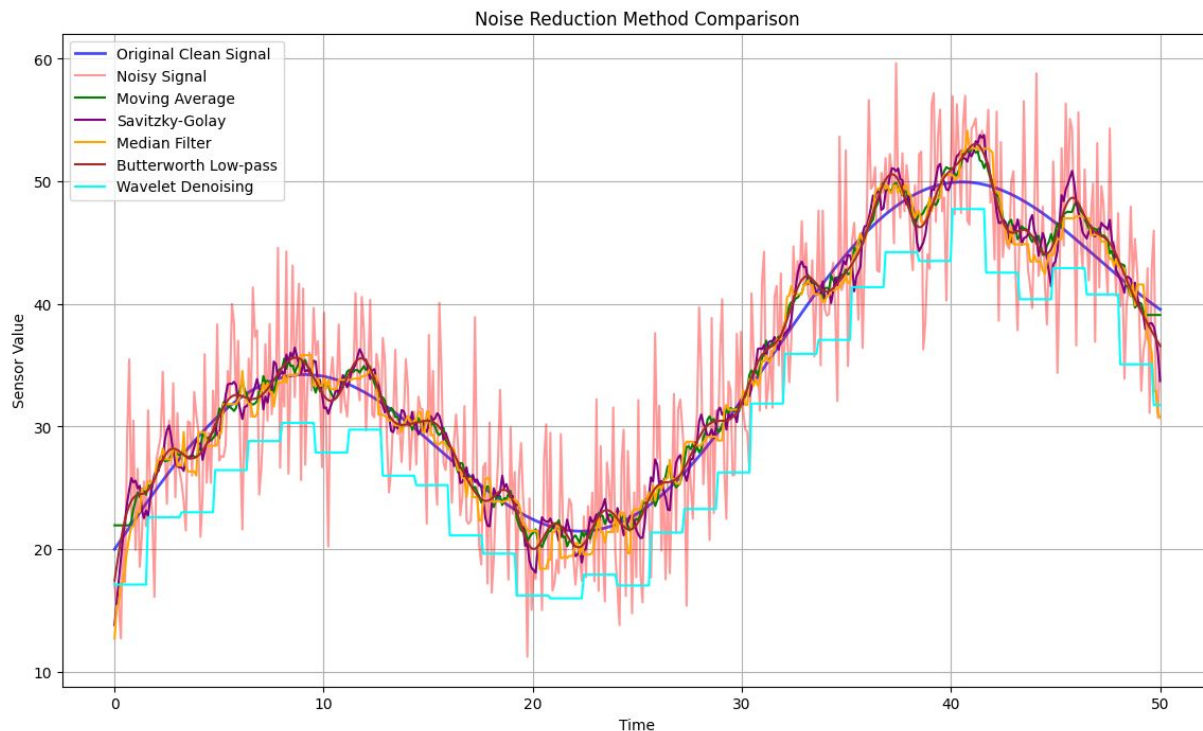
Remoção de ruído em dados



Remoção de ruído em dados



Remoção de ruído em dados



Method	RMSE
Noisy Signal	5.13
Moving Average	1.29
Savitzky-Golay	1.88
Median Filter	1.85
Butterworth Low-pass	1.51
Wavelet Denoising	4.89

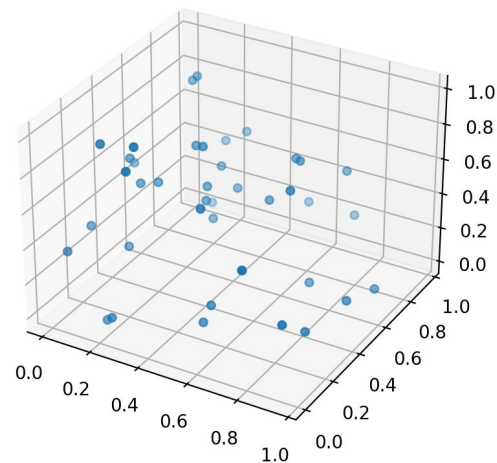
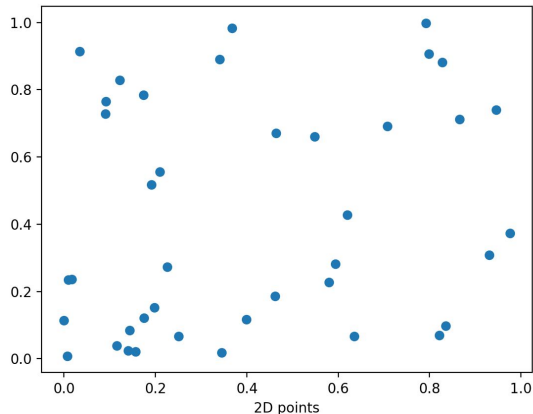
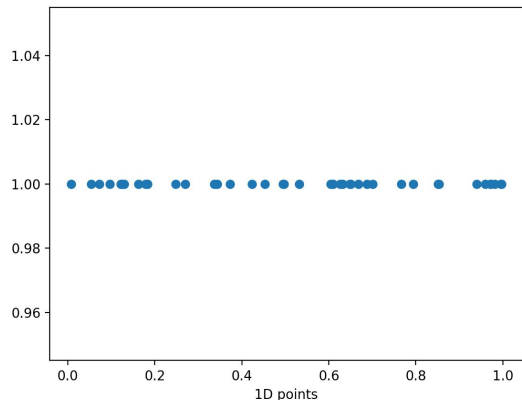
Agregar datos

Maldição da dimensionalidade

Dimensão - número de variáveis

- na teoria: + variáveis $\rightarrow$ + performance
- na prática: demasiadas variáveis, pior performance
(autocorrelações, mais ruído, perda de significado, ...)

Temos de realizar uma redução de dimensionalidade.



Correlação

- **Covariance** $\sigma_{Y_1Y_2}$ - measures how 2 variables covary ($-\infty$ to $+\infty$).

Sum of cross products

Estimator:
$$S_{Y_1Y_2} = \frac{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{n-1}$$
 (absolute magnitude depends on the units of variables)

- **Pearson (product-moment) correlation** $\rho_{Y_1Y_2}$ - covariance standardized by the product of the standard variations of the two variables (-1 to $+1$).

Estimator:
$$r_{Y_1Y_2} = \frac{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^n (Y_{i2} - \bar{Y}_2)^2}}$$
 ($= S_{Y_1Y_2}$ when Y_1 and Y_2 are standardized to mean=0 and SD=1)

Correlação ordinal

Spearman rank correlation coefficient r_s

Computed the same way as Pearsons' correlation but using ranks instead of values

Kendall's rank correlation coefficient or Kendall's tau τ

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})}$$

A pair of observations (x_i, y_i) and (x_j, y_j) are **concordant** if either both $x_i < x_j$ and $y_i < y_j$ or both $x_i > x_j$ and $y_i > y_j$. Otherwise they are **discordant**.

- Usually $r_s > \tau$
- Both measures the strength of monotonic (not necessarily linear) relationships between two variables
- Hypothesis testing: in this case H_0 is ***“there is no monotonic relationship between the two variables”***.

Matrizes de covariância e correlação

Covariance matrix – represents how all pairs of variables in a dataset covary with each other

$$\begin{bmatrix} S_{Y_1}^2 & \cdots & S_{Y_n Y_1} \\ \vdots & \ddots & \vdots \\ S_{Y_1 Y_n} & \cdots & S_{Y_n}^2 \end{bmatrix}$$

Covariance = $S_{Y_n}^2 \times S_{Y_1}^2 \times r_{Y_n Y_1}$
 Variances Pearson's r
 Variance (diagonal)

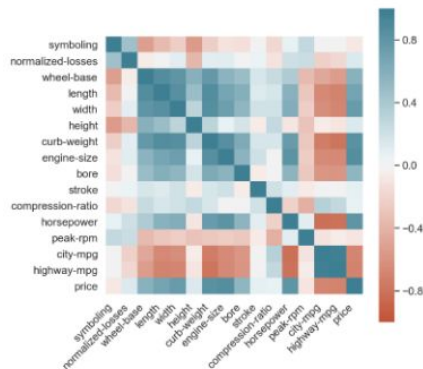
Correlation matrix – represents how all pairs of variables in a dataset correlate with each other

$$\begin{bmatrix} 1 & \cdots & r_{Y_n Y_1} \\ \vdots & \ddots & \vdots \\ r_{Y_1 Y_n} & \cdots & 1 \end{bmatrix}$$

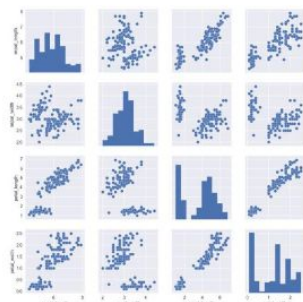
Matrizes de covariância e correlação

Visualization of correlation matrices

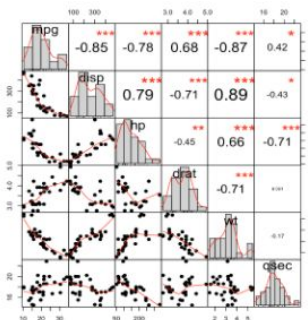
Heatplots
(many variables)



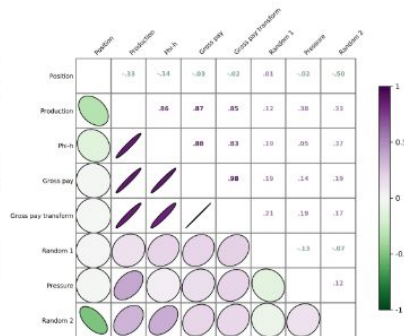
Scatter plot/histogram



Scatter/hist/coefficients



PDF plot/coefficients



Can be triangular

Redução da dimensionalidade

Análise em componentes principais

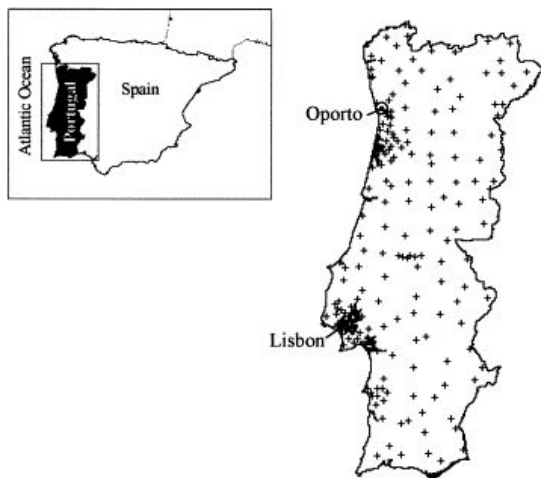


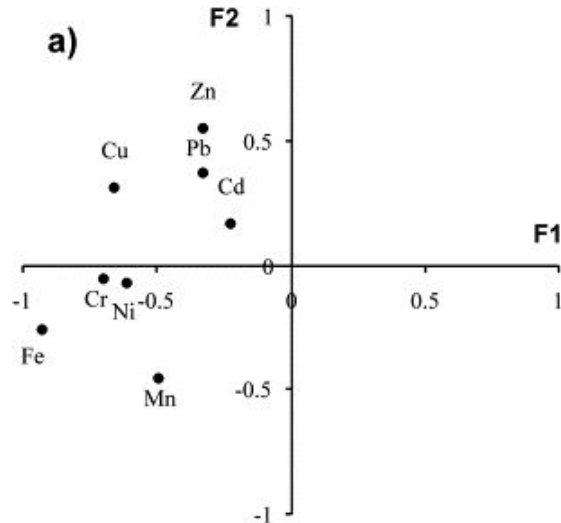
Table 1. Summary of statistics obtained for the concentration ($\mu\text{g/g}$ dry wt.) of all elements determined in Portuguese mosses

	Cd	Cr	Cu	Fe	Mn	Ni	Pb	Zn
Mean	0.79	3.86	9.10	1934.29	187.51	13.54	21.77	52.48
S.D.	0.54	8.99	7.49	2026.65	154.10	23.18	22.80	59.52
Coef. Variation	0.68	2.33	0.82	1.05	0.82	1.71	1.05	1.13
Minimum	0.04	0.04	0.40	82.98	4.03	0.83	1.71	12.97
Maximum	3.53	107.26	52.01	11 193.61	948.76	254.46	191.17	709.79
Median	0.70	1.95	7.38	1204.62	152.14	9.50	16.67	40.43
1st Quartile	0.44	1.00	4.88	663.85	85.58	5.96	10.12	29.02
3rd Quartile	0.98	3.78	10.72	2381.35	246.05	14.50	26.86	58.95
Quartile range	0.54	2.78	5.84	1717.50	160.47	8.54	16.74	29.93
Skewness	2.02	9.20	3.17	2.28	1.81	8.34	4.68	8.13

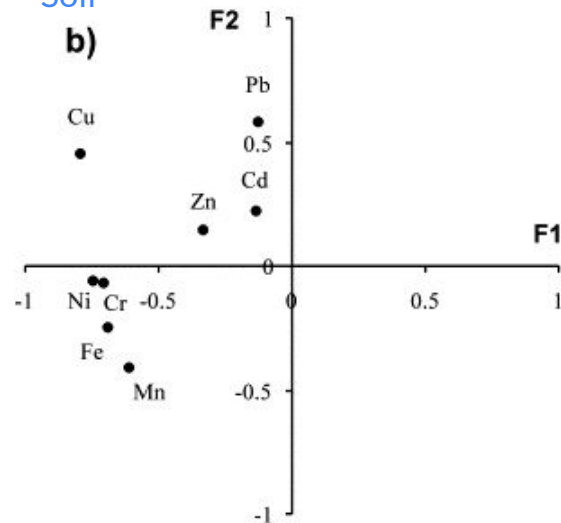
Redução da dimensionalidade

Análise em componentes principais

Moss



Soil



Sumário

- Dados, metadados, conjuntos de dados
- Dados privados, dados abertos, licenças, FAIR
- Tipos de dados, infraestruturas de dados, repositórios, serviços de dados
- Análise de dados
- Trabalhar com dados em falta
- Remoção de ruído
- Agregar dados

<https://tinyurl.com/mru5wc2r>



Exercícios

Aceder a

https://github.com/isa-ulisboa/SIR/blob/main/gestaoDados/SIR_DA_exec01.md