INTRODUÇÃO À CIÊNCIA DE DADOS

Hemilyn Stephanye

QUEM SOU EU

- UFSCar - Ciência da Computação
- Cientista de dados
- PyLadies
- PANDAs
- linkedin
- @Teclad0

QUEM SOU EU

- UFSCar - Ciência da Computação
- Cientista de dados
- PyLadies
  - Grupos de estudos
  - Eventos
  - Minicursos
- PANDAs
- linkedin
- @Teclad0

QUEM SOU EU

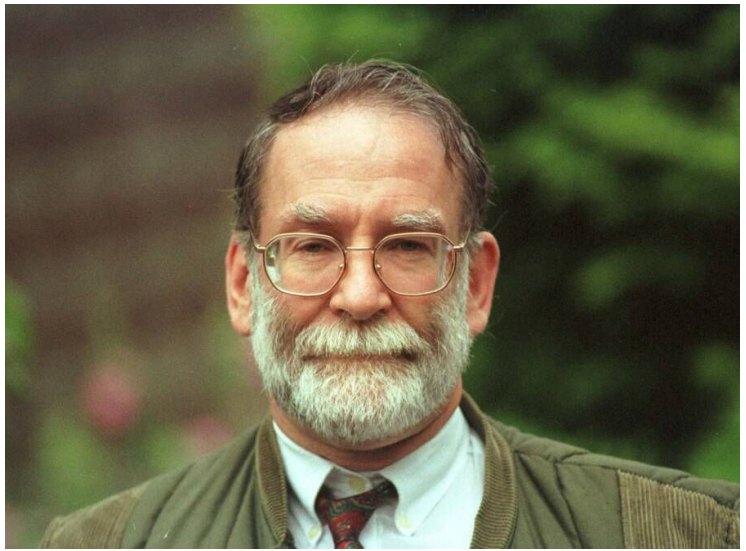
- UFSCar - Ciência da Computação
- Cientista de dados
- PyLadies
- PANDAs
- linkedin
- @Teclad0

ROTEIRO

- Exemplos
- O que são dados?
- Ciência de dados
- Métodos de obtenção de dados

Harold Shipman

- Assassino em série na Grã-Bretanha
- 1975 - 1998
- 15 pacientes
- Enorme dose de opiáceos
- Inquérito público



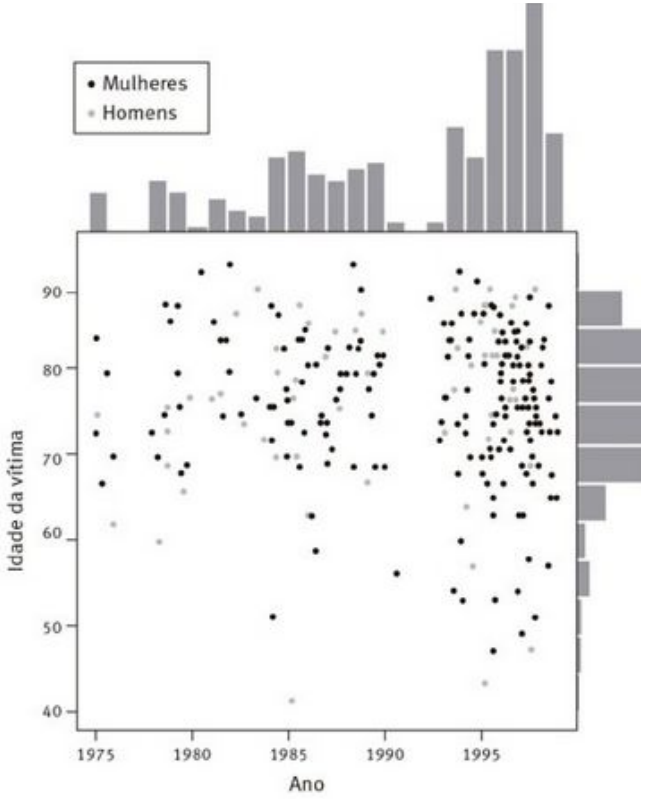


FIGURA 0.1 Um gráfico de dispersão mostrando a idade e o ano de morte das 215 vítimas confirmadas de Harold Shipman. Os gráficos de barras foram acrescentados aos eixos para revelar o padrão das idades e o padrão dos anos nos quais ele cometeu os assassinatos.

Que tipo de pessoas Harold Shipman assassinou, e quando morreram?

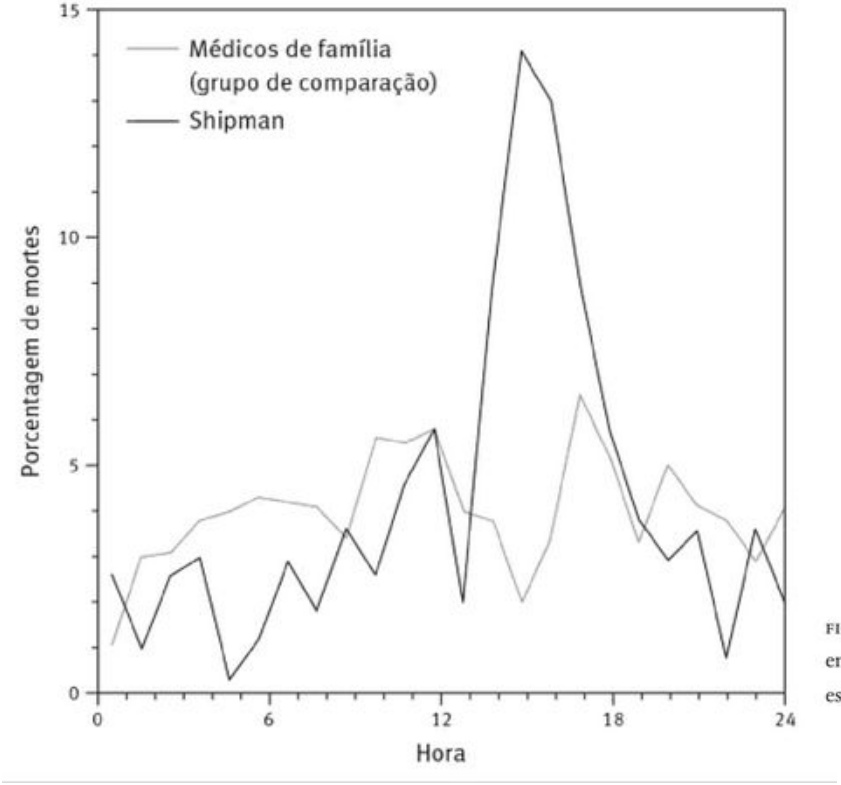


FIGURA 0.2 As horas nas quais os pacientes de Harold Shipman morreram, em comparação com as horas em que morreram pacientes de outros médicos de família locais. O padrão não requer uma análise estatística sofisticada.

Que tipo de pessoas Harold Shipman assassinou, e quando morreram?

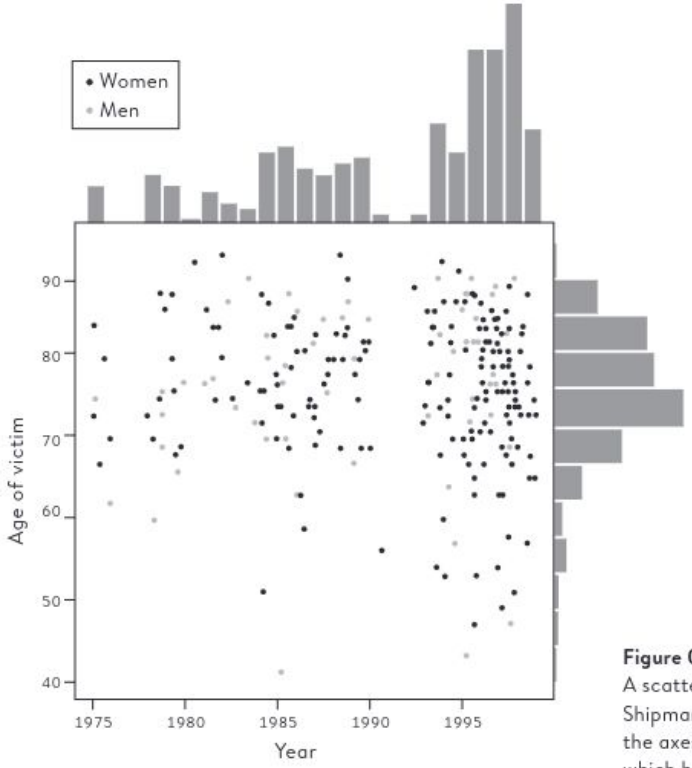


Figure 0.1
A scatter-plot showing the age and the year of death of Harold Shipman's 215 confirmed victims. Bar-charts have been added on the axes to reveal the pattern of ages and the pattern of years in which he committed murders.

Que tipo de pessoas Harold Shipman assassinou, e quando morreram?

Open Access Article

Potential Confounders in the Analysis of Brazilian Adolescent's Health: A Combination of Machine Learning and Graph Theory

by Amanda Yumi Ambriola Oku [1], Guilherme Augusto Zimeo Morais [2], Ana Paula Arantes Bueno [1], André Fujita [3] and João Ricardo Sato [1,*]

[1] Center of Mathematics, Computing and Cognition—Universidade Federal do
Brazil
[2] Big Data—Hospital Israelita Albert Einstein, São Paulo CEP 05652-900, Bra
[3] Institute of Mathematics and Statistics—University of São Paulo, São Paulo
* Author to whom correspondence should be addressed.

Int. J. Environ. Res. Public Health **2020**, *17*(1), 90; https://doi.org/10.3390/ijerp

Submission received: 30 October 2019 / Revised: 9 December 2019 / Accep
Published: 21 December 2019

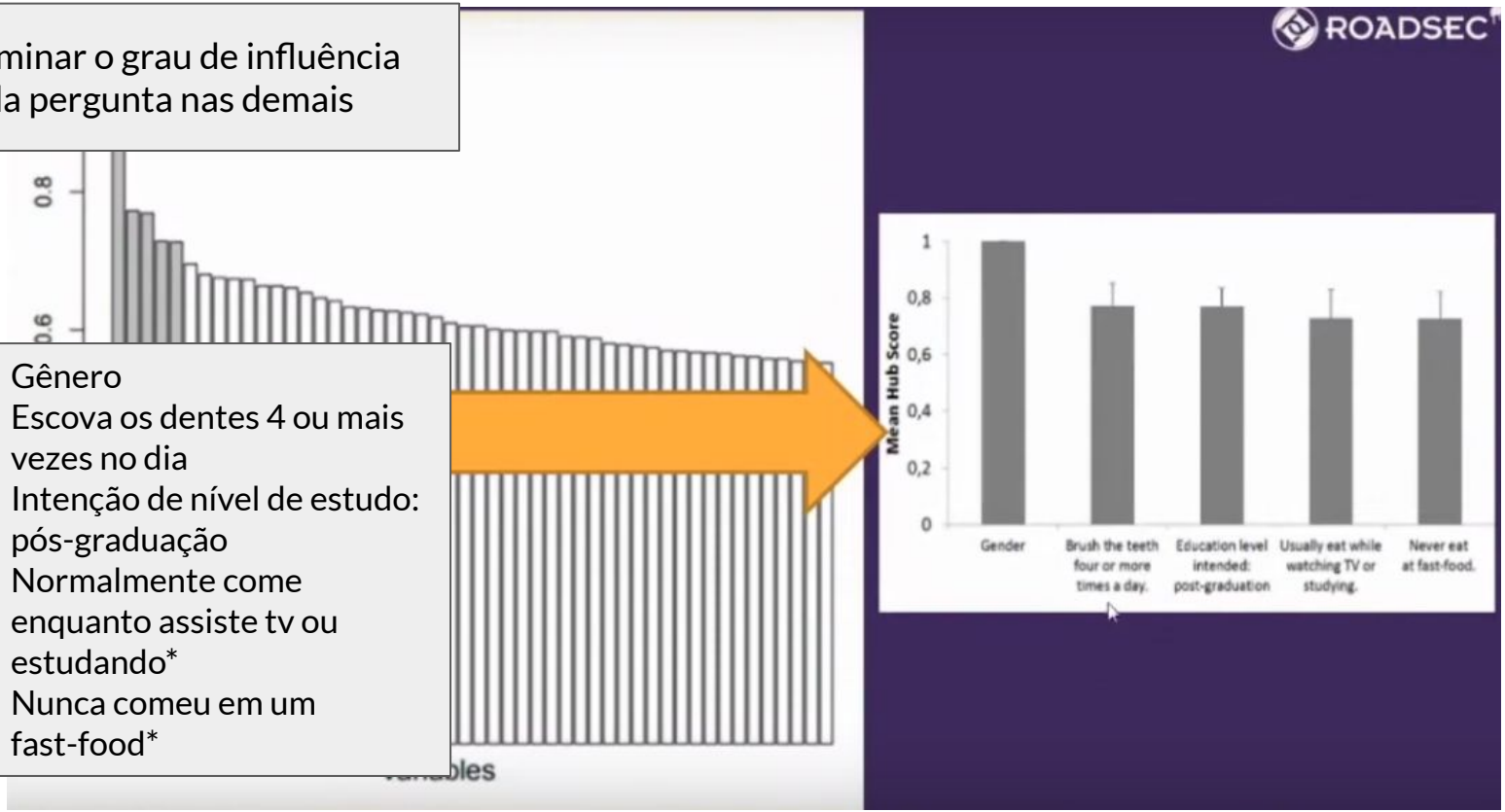
(This article belongs to the Special Issue Network Analytics in Healthcare De

Download Browse Figures Versions Notes

Análise dos dados da Pesquisa Nacional de Saúde Escolar (PeNSE) através de grafos e aprendizagem de máquina por Yumi. Grafos e machine learning para análise de dados públicos | Yumi Ambriola

Determinar o grau de influência de cada pergunta nas demais

- Gênero
- Escova os dentes 4 ou mais vezes no dia
- Intenção de nível de estudo: pós-graduação
- Normalmente come enquanto assiste tv ou estudando*
- Nunca comeu em um fast-food*



Análise dos dados da Pesquisa Nacional de Saúde Escolar (PeNSE) através de grafos e aprendizagem de máquina por Yumi. [Grafos e machine learning para análise de dados públicos | Yumi Ambriola]

Geral

Estudo inédito indica alta chance de fraude em mil provas do Enem

23/04/2018 11h33

Análise estatística inédita aponta alta probabilidade de ter havido fraude, de diferentes formas, em ao menos 1.125 provas do Enem, exame nacional que seleciona estudantes para universidades públicas no país.

Essas provas estão dentro de grupos com padrão de respostas tão semelhantes entre si que, estatisticamente, é improvável que não tenha havido algum tipo de cola nesses casos.

Segundo o modelo estatístico desenvolvido pela Folha, a chance de essas provas serem semelhantes apenas devido ao acaso em uma edição do Enem é de no mínimo 1 em 1.000.

Ou seja, seria necessário repetir o exame mil vezes para que duas provas, sem interferência, fossem tão parecidas como os gabaritos suspeitos.

Estudo realizado pela folha com os microdados do Enem. Fonte: https://www.blogdobg.com.br/estudo-inedito-indica-alta-chance-de-fraude-em-mil-provas-do-enem/

O QUE SÃO DADOS?

O QUE SÃO DADOS? - ATIVIDADES NO DC

ID	NOME	TIPO
1	Introdução à PLN	Aula aberta
2	Introdução à ciência de dados	Aula aberta
3	Plataformização da educação	Mesa de debate
...	....	...
10	Robos	Cine
11	Liberdade de expressão na internet	Podcast

O QUE SÃO DADOS?

TIPO	QUANTIDADE
Aula aberta	3
Podcast	2
Cine	3
…	….

O QUE SÃO DADOS?

- Dados são valores atribuídos a algo
- Descrição básica de objetos
- Estes valores não precisam ser necessariamente números

FORMATOS

SEQUÊNCIAS ORDENADAS

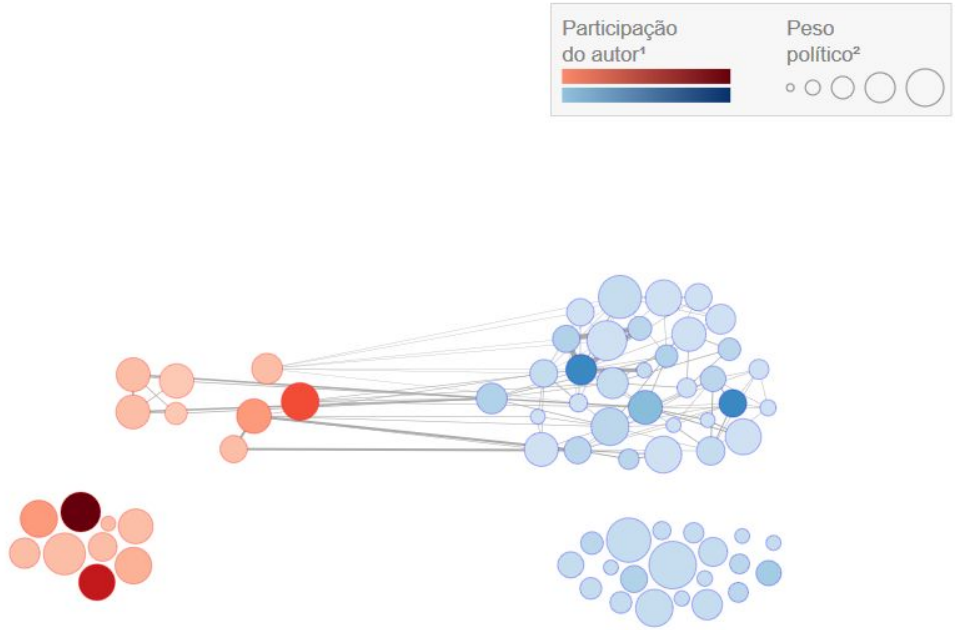
greve = [“apoie”, “a”, “greve”]

TABELAS

Covid-19 - Quantidade de casos confirmados no Acre, por município

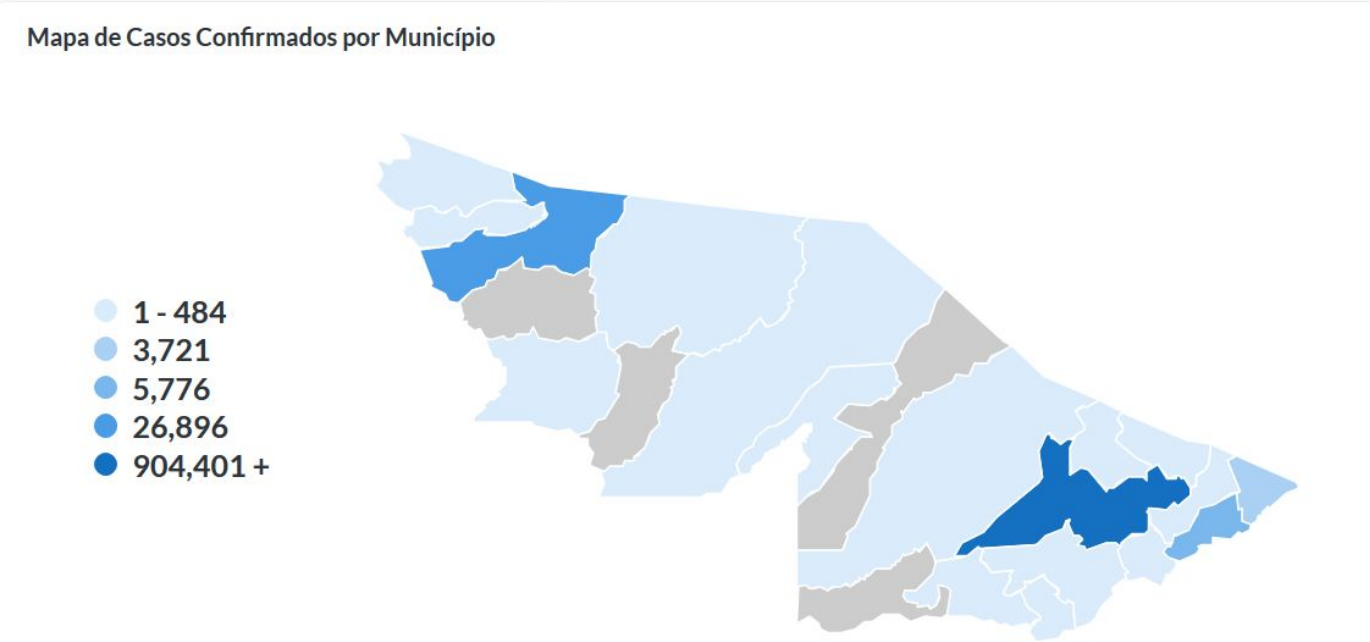
Município	Número de casos
Rio Branco	826
Cruzeiro do Sul	132
Plácido de Castro	69
Acrelândia	40
Tarauacá	20

REDES E GRAFOS



Parlamentares conectados pelos textos em conjunto no PL das Armas. Fonte: https://ok.org.br/projetos/parlametria/

DADOS GEOGRÁFICOS



Casos confirmados de Covid-19 no Acre, por município.. Fonte: https://covid.saude.gov.br/

TEXTO

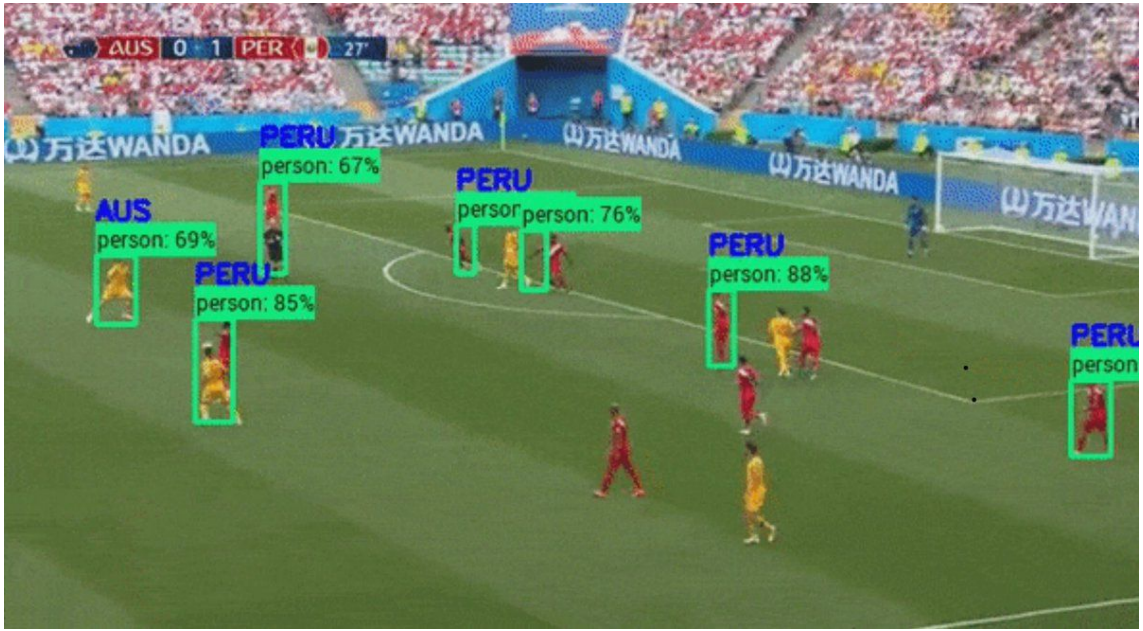
No meio do caminho tinha uma pedra
Tinha uma pedra no meio do caminho
Tinha uma pedra
No meio do caminho tinha uma pedra

Nunca me esquecerei desse acontecimento
Na vida de minhas retinas tão fatigadas

Nunca me esquecerei que no meio do caminho
Tinha uma pedra
Tinha uma pedra no meio do caminho
No meio do caminho tinha uma pedra

No Meio do Caminho
Carlos Drummond de Andrade

IMAGENS

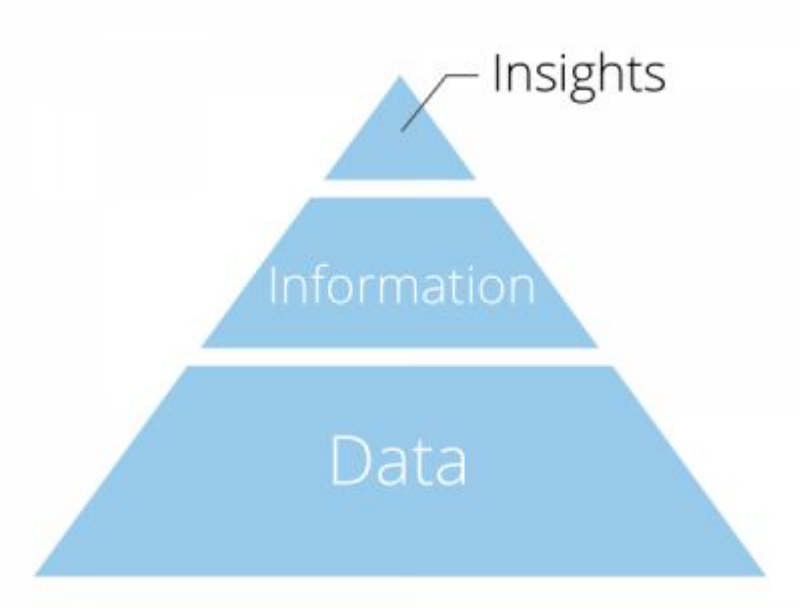


Análise de partida de futebol com OpenCV. Fonte: https://github.com/priya-dwivedi

CIÊNCIA DE DADOS

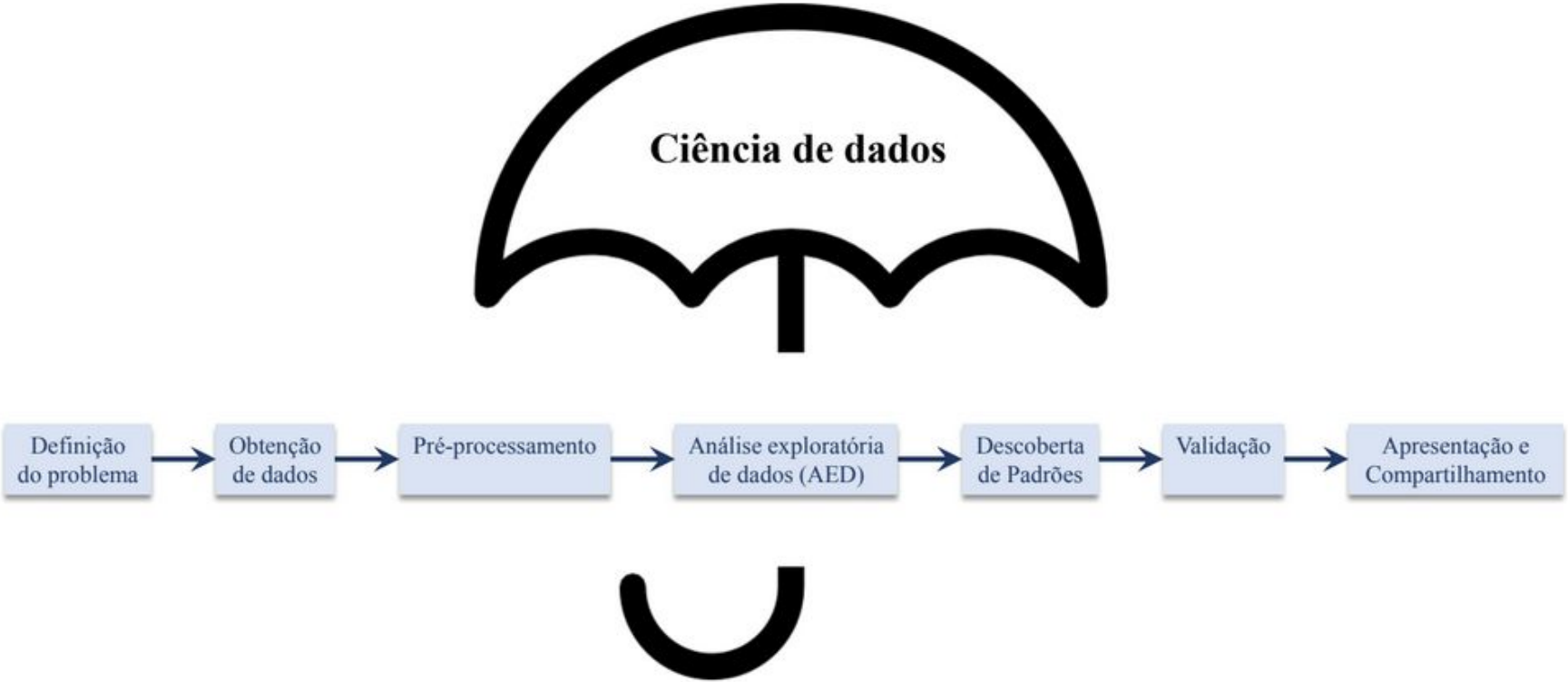
MOTIVAÇÃO

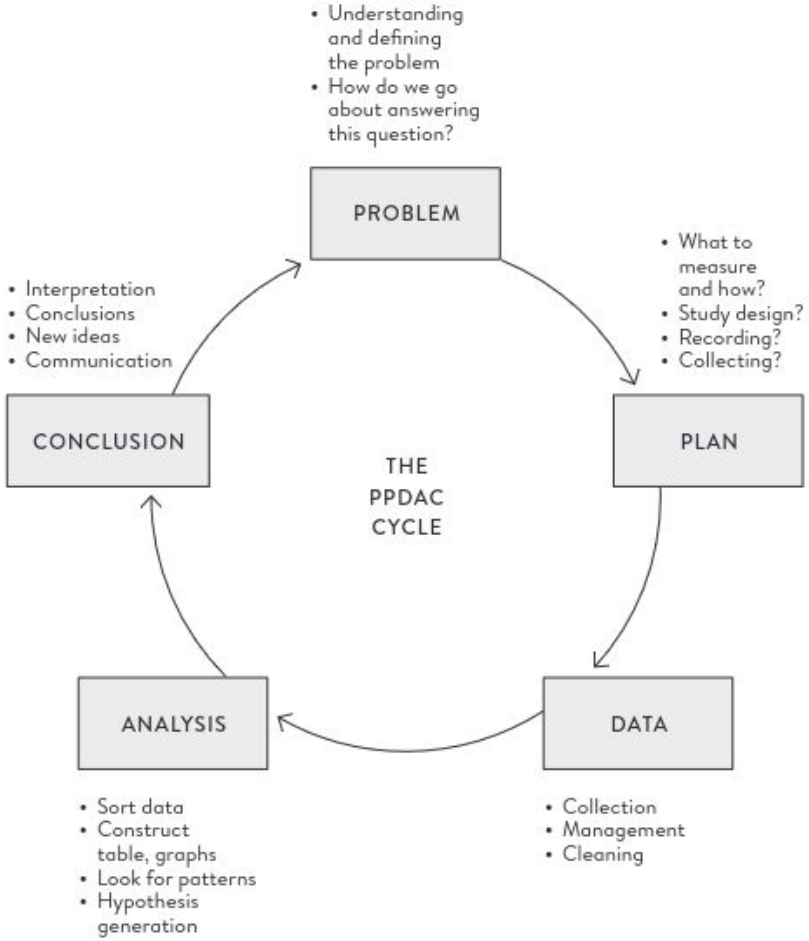
- Os sistemas computacionais
- Valiosa fonte de conhecimento
- **Dificuldade de exploração manual desse conhecimento**
- Necessidade de técnicas automáticas para extrair **padrões** sobre os dados armazenados

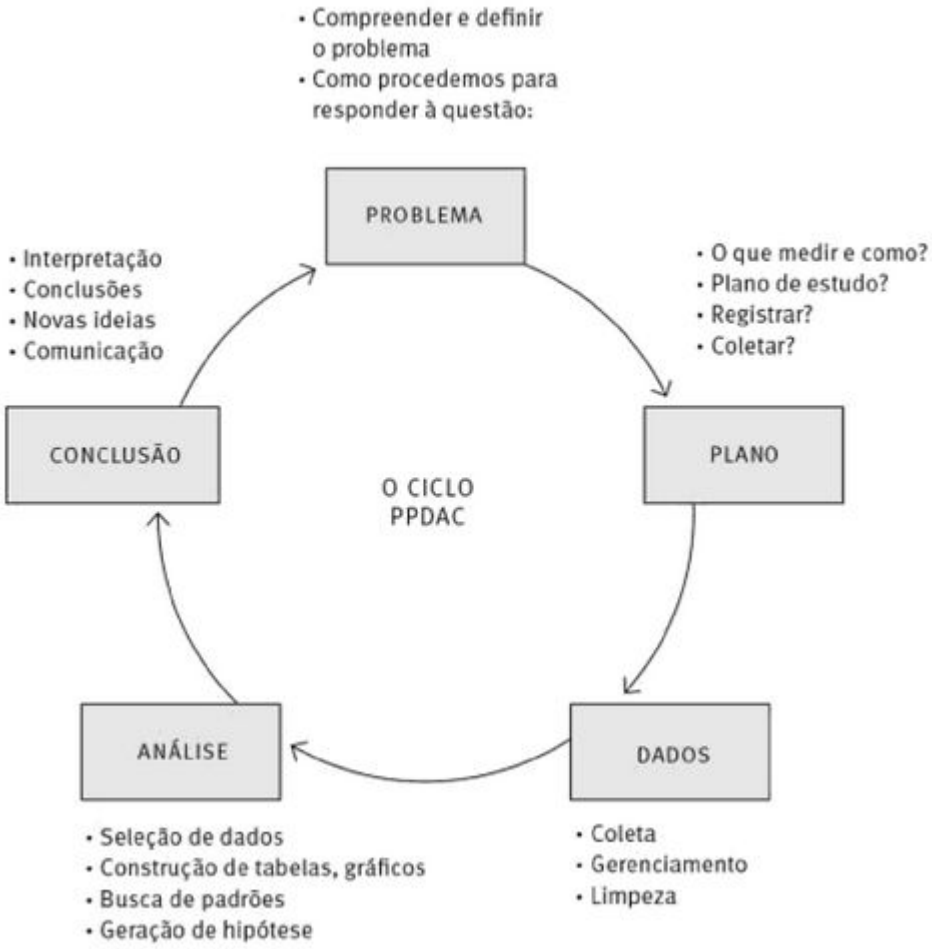


COMO ANALISAMOS TAIS DADOS?

- Processos de ETL (extração, transformação e carga dos dados)
- Tarefas de mineração
- Técnicas de Visualização e Exploração
- Modelos, algoritmos, sistemas, equipamentos


Ciência de dados
Definição do problema
Obtenção de dados
Pré-processamento
Análise exploratória de dados (AED)
Descoberta de Padrões
Validação
Apresentação e Compartilhamento


• Understanding and defining the problem
• How do we go about answering this question?
PROBLEM
• What to measure and how?
• Study design?
• Recording?
• Collecting?
PLAN
DATA
• Collection
• Management
• Cleaning
ANALYSIS
• Sort data
• Construct table, graphs
• Look for patterns
• Hypothesis generation
CONCLUSION
• Interpretation
• Conclusions
• New ideas
• Communication
THE PPDAC CYCLE

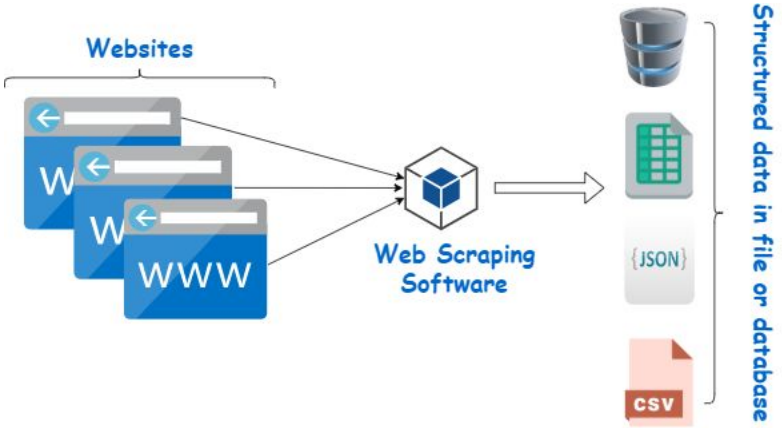

• Compreender e definir o problema
• Como procedemos para responder à questão:
PROBLEMA
• O que medir e como?
• Plano de estudo?
• Registrar?
• Coletar?
PLANO
DADOS
• Coleta
• Gerenciamento
• Limpeza
ANÁLISE
• Seleção de dados
• Construção de tabelas, gráficos
• Busca de padrões
• Geração de hipótese
CONCLUSÃO
• Interpretação
• Conclusões
• Novas ideias
• Comunicação
O CICLO PPDAC

MÉTODOS DE OBTENÇÃO DE DADOS

MÉTODOS DE OBTENÇÃO DE DADOS

	Métodos comuns	Formatos comuns	Dificuldade técnica para acessar	Dificuldade para atualizar ou filtrar os dados	Dificuldade para acessar a base completa
Dados tabulares	Raspagem de dados, download manual	HTML, CSV, XLS	Baixa	Média	Baixa
API	Scripts e programas	JSON, XML	Média	Baixa	Alta
Dados não estruturados	Raspagem de dados, download, OCR	PDF, HTML	Média	Alta	Alta

RASPAGEM DE DADOS



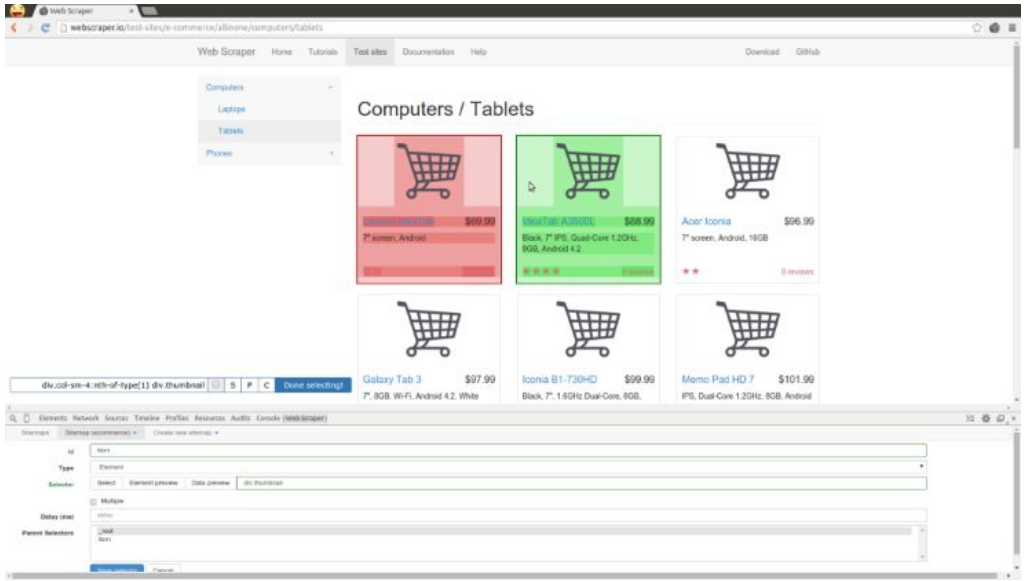
Processo de **coleta automatizada** de conjuntos de dados contidos em websites e outras formas de visualizações

Conhecimentos envolvidos
HTML, CSS, JavaScript, XPath

RASPAGEM DE DADOS

Ferramentas úteis

- WebScraper
- Google Sheets
- IFTTT (robô)
- Beautiful Soup(Python)
  - rvest (R)



RASPAGEM DE DADOS

Alesp - presença nas sessões plenárias

- https://github.com/priscilaportela-zz/opendataday2020/blob/master/presenca-em-plenario-alesp/presenca_em_plenario.ipynb

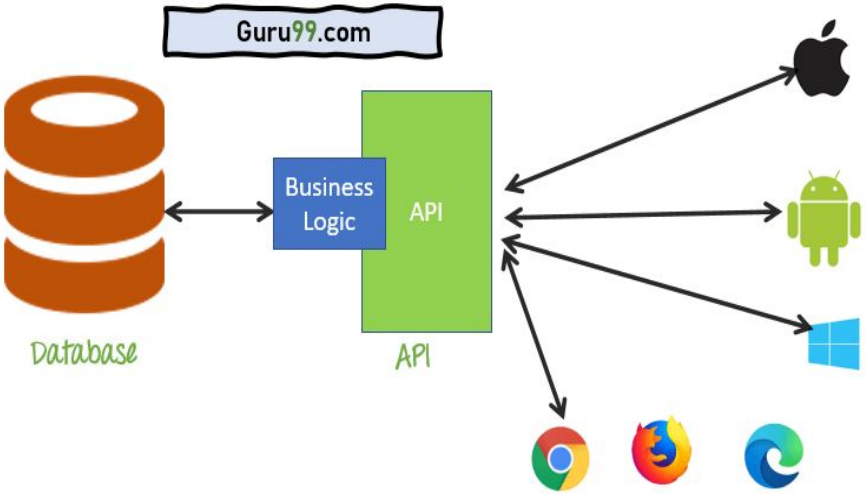
Aviões atrasados ou cancelados (não atualizado)

- =importHTML("http://voos.infraero.gov.br/hstvoos/RelatorioPortal.aspx"; "table"; 1)
- Google Sheets

RASPAGEM DE DADOS

- Dificuldades
  - Sites sem um padrão definido
  - Dados protegidos com CAPTCHA
  - Pequenas alterações no site podem “quebrar” a raspagem
  - São públicos? Se não, quais termos ou condições?

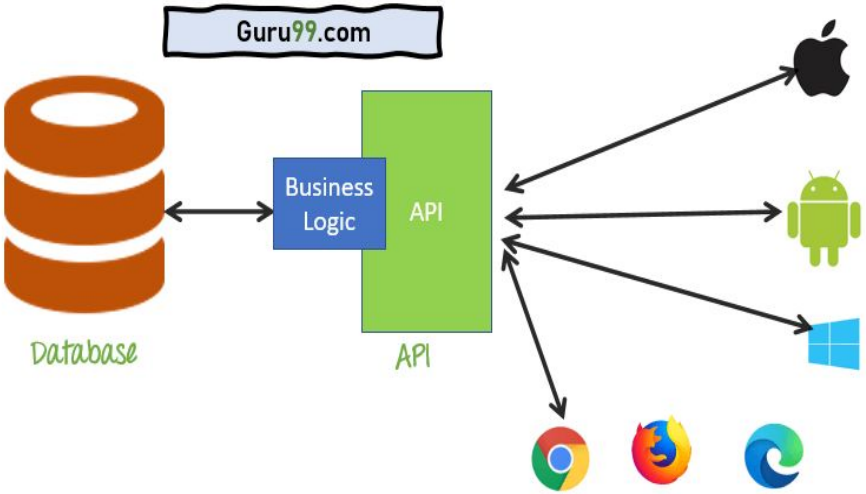
API



Interfaces que facilitam o consumo e obtenção de dados para máquinas

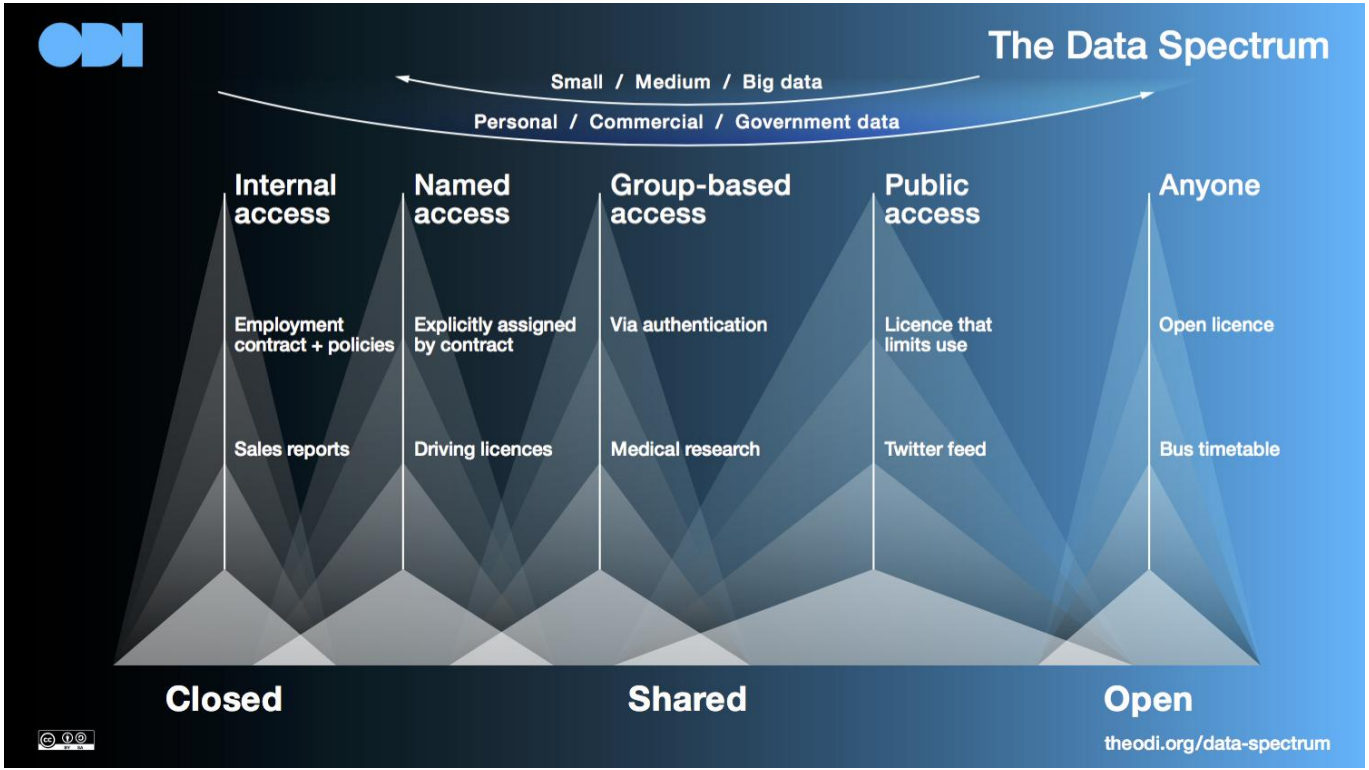
Conhecimentos envolvidos
Python, R ou outra linguagem de programação

API



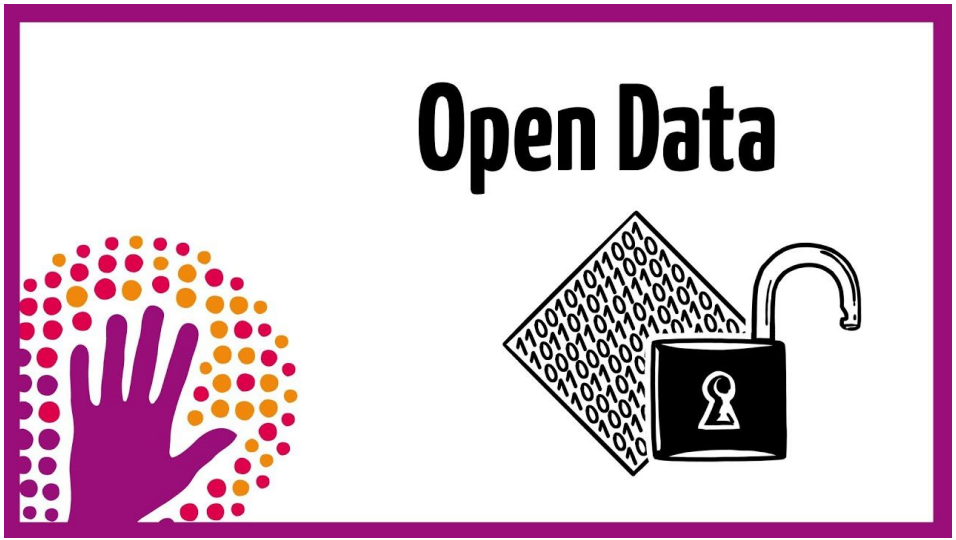
- Referências úteis
  - API do Portal da Transparência
  - API da Câmera e do Senado Federal
  - Dados abertos da Câmera
  - Open Movie Database

PORTAIS DE DADOS ABERTOS



PORTAIS DE DADOS ABERTOS

- **usados, reutilizados e redistribuídos**
- por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras



PORTAIS DE DADOS ABERTOS

Portal Brasileiro de Dados Abertos
- Série Histórica de Preços de Combustíveis

IBGE

Dados Abertos da Educação
- Índice de Desenvolvimento da Educação do Estados de São Paulo(IDESP)

OUTROS

Lei Acesso a Informação

- Solicite dados de órgãos públicos

Base dos Dados

- ONG que disponibiliza dados

Kaggle

- Diversos conjuntos de dados

RESUMO

- Análises de dados
- Pensamento para análise de dados
- Métodos de obtenção de dados

INSPIRAÇÕES

- Estudo inédito indica alta chance de fraude em mil provas do enem
  - https://www.blogdobg.com.br/estudo-inedito-indica-alta-chance-de-fraude-em-mil-provas-do-enem/ (sem paywall)
- Fogo cruzado
  - Dados sobre violência armada
  - https://www.youtube.com/watch?v=tAM1Gce5Quo

- Ciclistas
- The Largest Ever Analysis of Film Dialogue by Gender

- Operação Serenata de Amor
  - Inteligência artificial para controle social da administração pública
- VOCÊ SABE EXTRAIR INFORMAÇÕES DE DADOS? | Análise de Dados #1

REFERÊNCIAS

- A arte da estatística: Como aprender a partir de dados - David Spiegelhalter
- Introduction to Cultural Analytics & Python