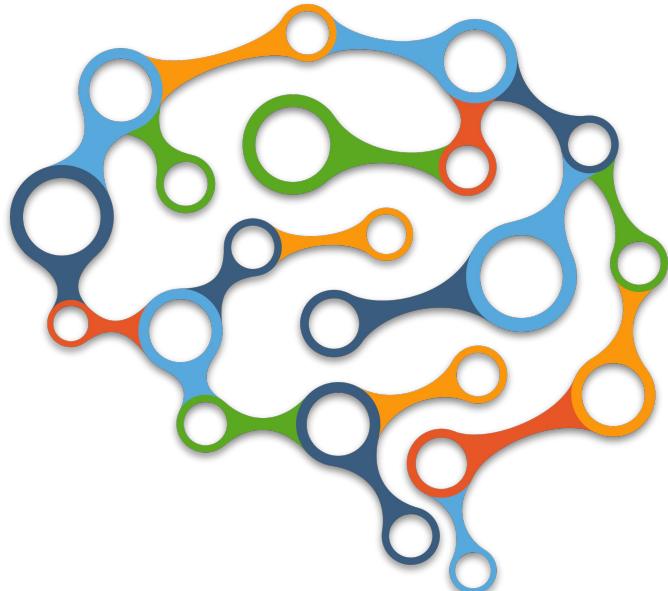


# Coletando, processando e visualizando dados abertos

Usando python para coletar, processar e visualizar dados disponibilizados pelo poder público.

Adolfo Guimarães

@profadolfoguimaraes



# Quem sou eu?

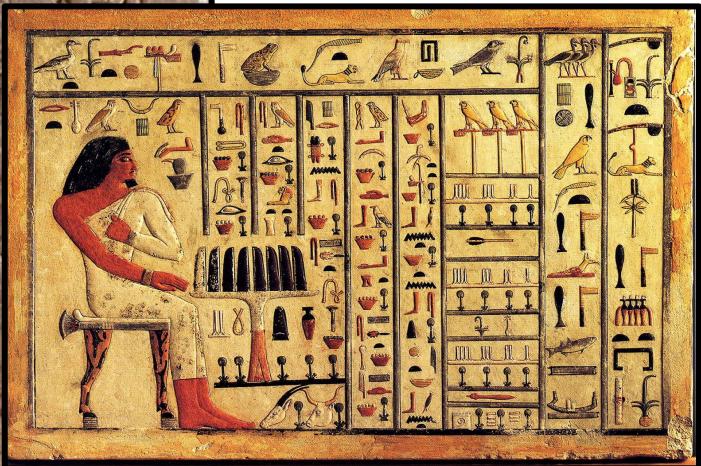
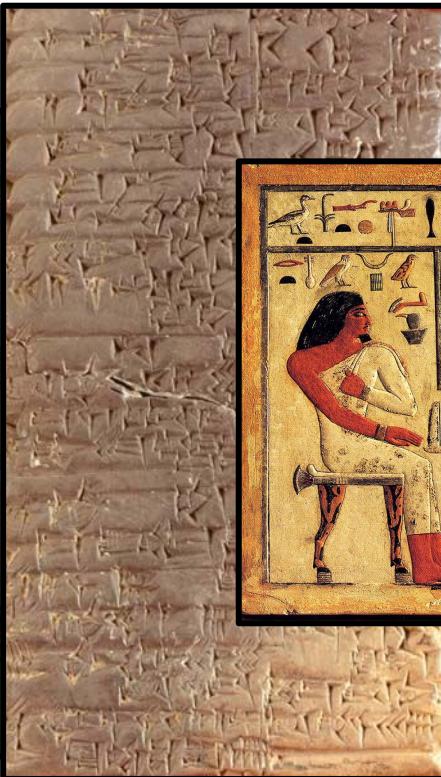


- 🌐 Brasileiro / Nordestino / Sergipano.
- 📚 Mestrado em Ciência da Computação.
- 💻 Python.
- ❤️ Inteligência Artificial.
- ✊ Acredito que a tecnologia pode e deve ser utilizada em prol da sociedade.

<https://github.com/adolfoguimaraes>  
<https://twitter.com/adolfoguimaraes>  
<https://instagram.com/profadolfovoguimaraes>



# Não é de hoje que a humanidade produz dados ...



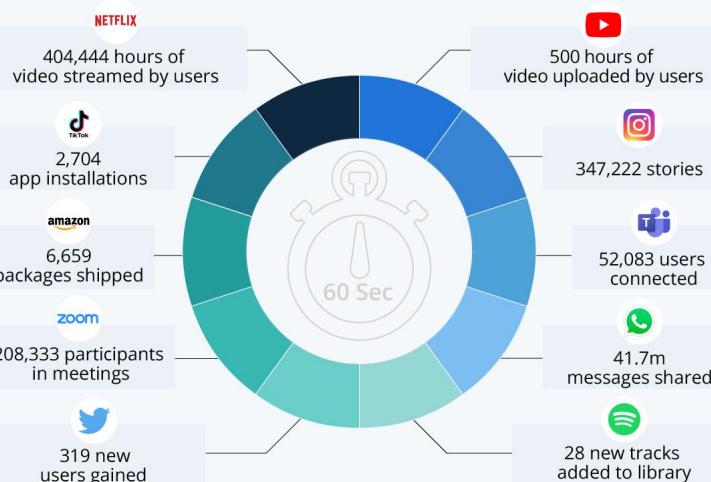
... nos mais variados formatos ...



# ... e continuamos produzindo MUITO dado.

## A Minute on the Internet in 2020

Estimated amount of data created  
on the internet in one minute



Source: Visual Capitalist



statista

The figure displays three screenshots of data portals:

- DATA.GOV**: Shows a search bar and a map filter for datasets ordered by popularity. One dataset listed is "Department for the Aging (DFTA) Geriatric Mental Providers".
- dados.gov.br**: Shows a search bar and a sidebar with organization links like Banco Central do Brasil (3305) and Instituto Nacional (1386). It also lists 10,709 datasets found.
- Brasil.io**: Shows a search bar and a list of datasets, including COVID-19 and Estado da balneabilidade da costa balnear.
- EU Open Data Portal**: Shows a search bar and a list of datasets, including European Union Open Data Portal, Home, Data, Applications, Linked data, Visualisations, Developers' corner, and About.

# COLETA



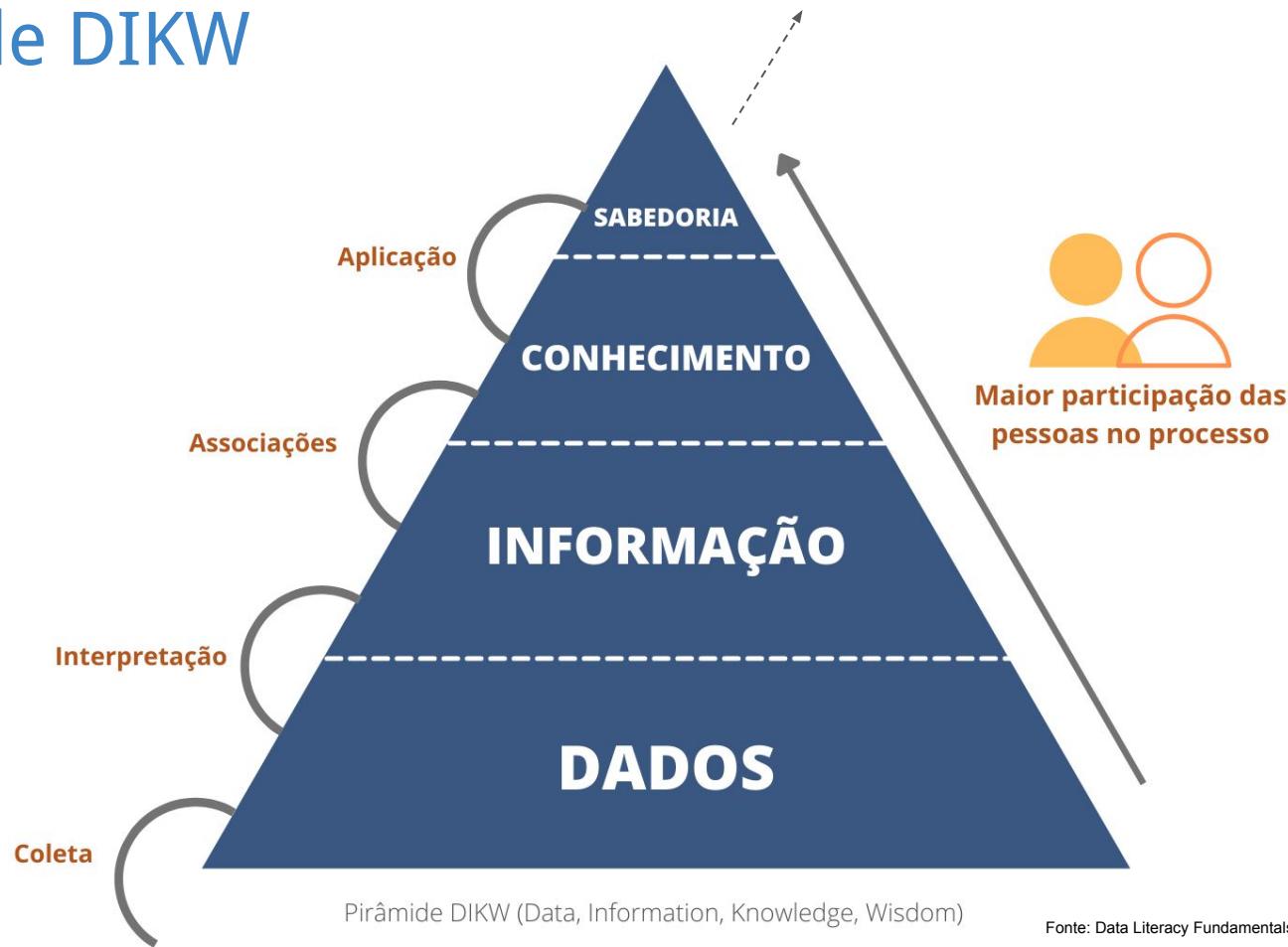
# PROCESSAMENTO E ANÁLISE



# VISUALIZAÇÃO

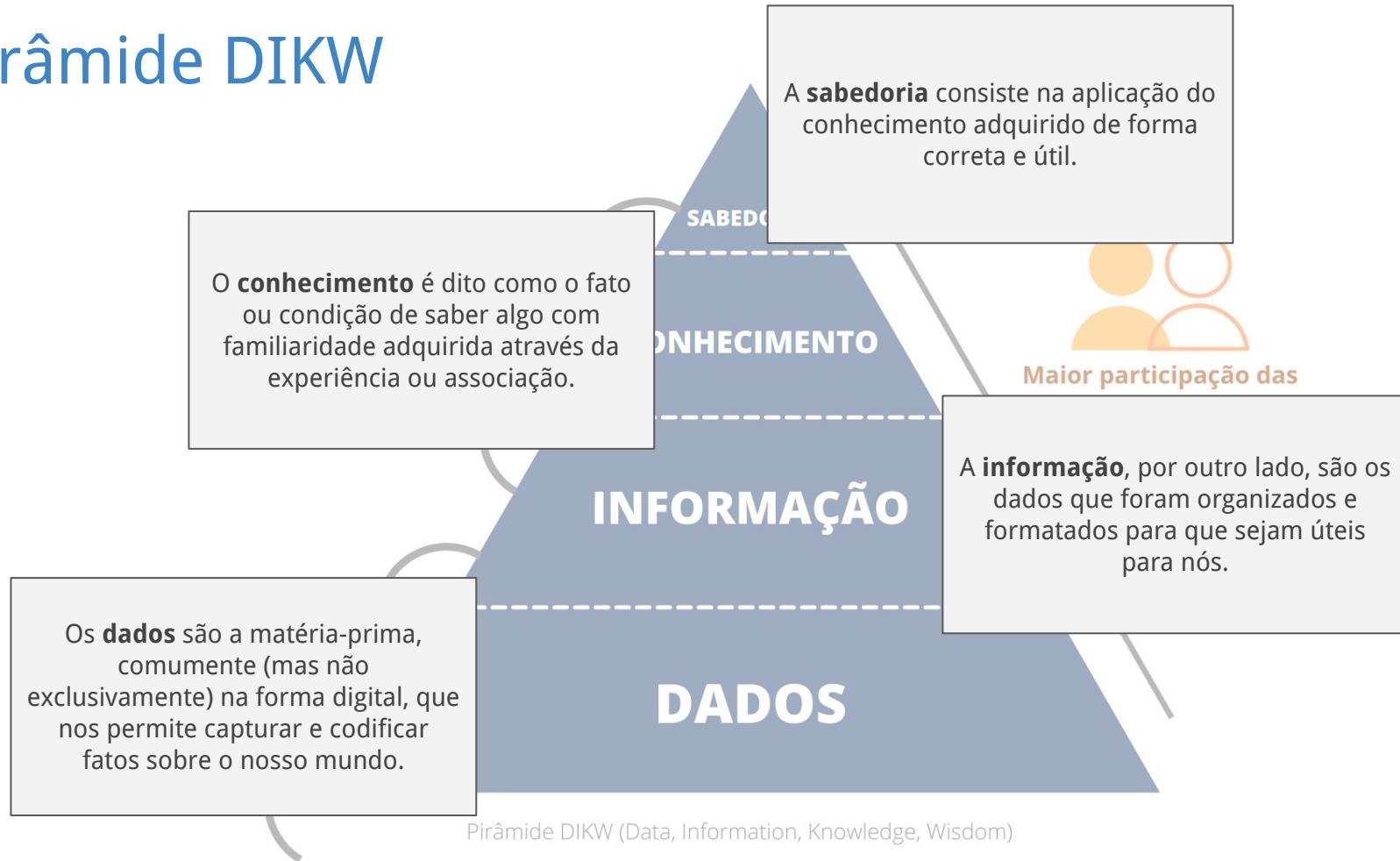


# Pirâmide DIKW

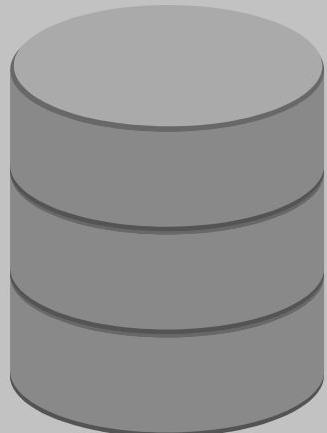


Fonte: Data Literacy Fundamentals: Understanding the Power & Value of Data  
Ben Jones  
<https://www.amazon.com.br/Data-Literacy-Fundamentals-Understanding-English-ebook>

# Pirâmide DIKW



# COLETA



**Onde conseguir os dados?**

# Nosso caminho começa com a LAI



A LAI (**Lei de Acesso à Informação**) foi editada em **2011** e regula o direito de **acesso à informação** já previsto na constituição.

A lei permite que **nós cidadãos façamos o controle social** a partir do conhecimento e da participação da administração pública.

# Nosso caminho começa com a LAI



A **LAI (Lei de Acesso à Informação)** foi editada em **2011** e regula o direito de **acesso à informação** já previsto na constituição.

A lei permite que **nós cidadãos façamos o controle social** a partir do conhecimento e da participação da administração pública.

A transparência pode ser classificada como **transparência ativa** e **transparência passiva**.

**transparência ativa**  
divulgação proativa de informações públicas

**transparência passiva**  
divulgação de informações mediante solicitação

# Quem está sujeito a lei?

Qualquer instituição que receba dinheiro público deve responder à LAI:

- Executivo, Judiciário, Legislativo da União, estados e municípios;
- Sociedades de economia mista e demais entidades controladas direta e indiretamente pelo governo (incluindo estatais e o Sistema S);
- Entidades sem fins lucrativos beneficiadas por recursos públicos.

# Conhecendo um pouco a lei.

**Art. 8º:** É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas.

**§ 1º Na divulgação das informações a que se refere o caput, deverão constar, no mínimo:**

**I** - registro das competências e estrutura organizacional, endereços e telefones das respectivas unidades e horários de atendimento ao público;

**II** - registros de quaisquer repasses ou transferências de recursos financeiros;

**III** - registros das despesas;

**IV** - informações concernentes a procedimentos licitatórios, inclusive os respectivos editais e resultados, bem como a todos os contratos celebrados;

**V** - dados gerais para o acompanhamento de programas, ações, projetos e obras de órgãos e entidades; e

**VI** - respostas a perguntas mais frequentes da sociedade.

# Conhecendo um pouco a lei.

**§ 3º Os sítios de que trata o § 2º deverão, na forma de regulamento, atender, entre outros, aos seguintes requisitos:**

**I** - conter ferramenta de pesquisa de conteúdo que permita o acesso à informação de forma objetiva, transparente, clara e em linguagem de fácil compreensão;

**II** - possibilitar a gravação de relatórios em diversos formatos eletrônicos, inclusive abertos e não proprietários, tais como planilhas e texto, de modo a facilitar a análise das informações;

**III** - possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina;

**IV** - divulgar em detalhes os formatos utilizados para estruturação da informação

**V** - garantir a autenticidade e a integridade das informações disponíveis para acesso;

**VI** - manter atualizadas as informações disponíveis para acesso;

**VII** - indicar local e instruções que permitam ao interessado comunicar-se, por via eletrônica ou telefônica, com o órgão ou entidade detentora do sítio; e

**VIII** - adotar as medidas necessárias para garantir a acessibilidade de conteúdo para pessoas com deficiência, nos termos do art. 17 da Lei nº 10.098, de 19 de dezembro de 2000, e do art. 9º da Convenção sobre os Direitos das Pessoas com Deficiência, aprovada pelo Decreto Legislativo nº 186, de 9 de julho de 2008.

# **Atenção!!!**

A lei aborda mais tópicos além dos que foram citados aqui e deve ser de conhecimento de todo cidadão, mas principalmente para quem vai trabalhar com esse tipo de dados.

Acesse em: [http://www.planalto.gov.br/ccivil\\_03/ato2011-2014/2011/lei/l12527.htm](http://www.planalto.gov.br/ccivil_03/ato2011-2014/2011/lei/l12527.htm)

# O que são dados abertos?

"Por dados abertos governamentais entendem-se os dados públicos, produzidos, coletados ou custodiados por órgãos estatais e disponibilizados na web em formato aberto (não proprietário), estruturado e comprehensível logicamente por máquina, de modo que qualquer um possa livremente acessar, reutilizar, modificar e redistribuir para quaisquer finalidades, estando sujeito a, no máximo, exigências de creditar a sua autoria e compartilhar sob a mesma licença"

# 8 princípios dos dados abertos

## Completos

todos os dados públicos devem ser disponibilizados, ao menos que tenham alguma restrição regulamentada por estatuto e publicamente justificada;

## Atuais

os dados são atualizados periodicamente e rapidamente para manter o seu valor e utilidade;

## Não proprietários

os dados não estão sob o controle de nenhuma entidade e em formatos que necessitem uso de ferramentas exclusivas de uma determinada organização;

## Livres de licença

os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Algumas restrições relacionadas à privacidade, segurança, controle de acesso e outros privilégios são permitidos desde que sejam transparentes, justificadas e regulamentadas por estatuto.

## Acessíveis

os dados são disponibilizados gratuitamente para alcançar o maior número de usuários e propósitos, sem restrições;

## Compreensíveis por máquina

os dados são disponibilizados gratuitamente para alcançar o maior número de usuários e propósitos, sem restrições;

## Primários

os dados são pré-processados, sem agregação ou modificação. Salvo exceções relacionadas ao sigilo de informações pessoais ou qualquer outro tipo de sigilo regulamentado por estatuto e publicamente justificado;

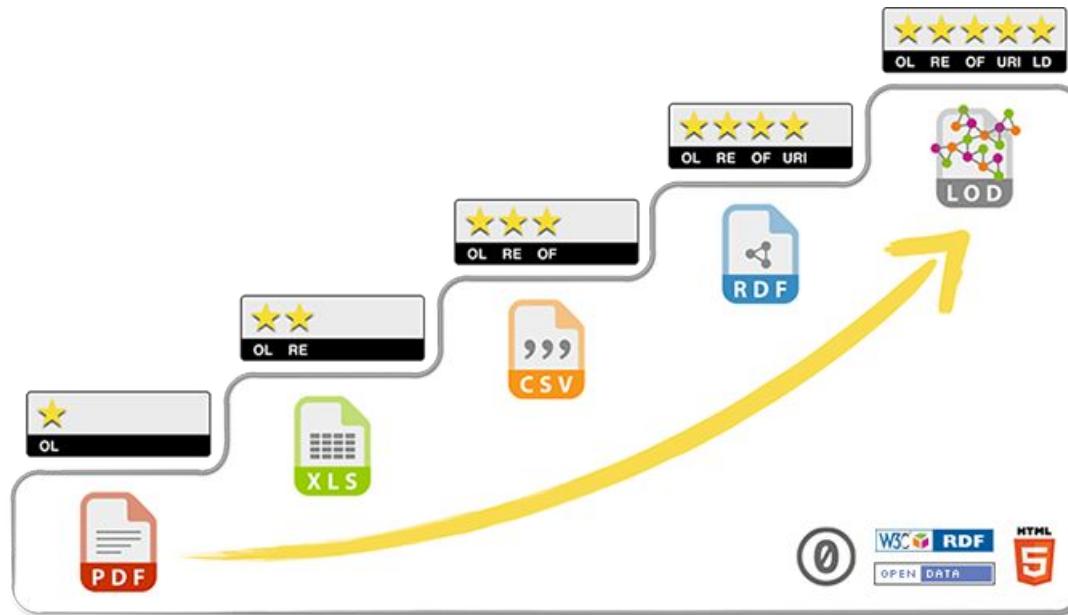
## Não discriminatórios

os dados são atualizados periodicamente e rapidamente para manter o seu valor e utilidade;

Fonte: Possamai, A. J., & De Souza, V. G. (2020). *Transparéncia e Dados Abertos Governamentais: Possibilidades e Desafios a Partir da Lei De Acesso À Informação*. Administração Pública e Gestão Social, 2020(2). <https://doi.org/10.21118/agps.v12i2.5872>

# As 5 estrelas dos dados abertos

Tim Berners-Lee, o inventor da Web e o primeiro a pensar nos Dados Ligados, sugeriu um esquema de implementação das 5 estrelas para Dados Abertos. Aqui, damos exemplos para cada etapa das estrelas e explicamos custos e benefícios que vêm junto com eles.



# Onde buscar os dados?

<http://www.portaltransparencia.gov.br/>

The screenshot shows the main interface of the Portal da Transparéncia. At the top, there's a blue header with the portal's logo and navigation links like 'ACESSIBILIDADE', 'ALTO CONTRASTE', and 'MARCOS SITES'. Below the header, a large green banner highlights the 'PAINEL CORONAVÍRUS' with a link to 'acessar'. The main content area features several cards: 'AUXÍLIO EMERGENCIAL', 'BENEFÍCIOS AO CIDADÃO', 'CARTÕES DE PAGAMENTO', 'CONVÉNIOS', 'DESPESAS E RECEITAS', 'EMENDA PARLAMENTAR', 'ESTADOS E MUNICÍPIOS', 'LICITAÇÕES E CONTRATOS', 'ORÇAMENTO', 'RECURSOS TRANSFERIDOS', and 'SERVIDORES E PENSIONISTAS'. A sidebar on the left contains sections for 'PANORAMA' (with links to 'Governo Federal', 'Pessoas Jurídicas', and 'Pessoas Físicas'), 'Consulte os documentos diários de despesa', 'Sanções' (with links to 'Fazenda', 'Empresas Multinacionais', 'Suspensões', 'Empresas Punitivas', 'Ordinadas', 'Projetos sem Fins Lucrativos Impedidas', and 'Acordos de Família'), and 'Destques - Rede de Transparência' and 'Destques - Combate ao Coronavírus'.

This screenshot shows the PORTAL BRASILEIRO DE DADOS ABERTOS. At the top, there's a search bar with the placeholder 'Buscar no Portal' and a button labeled 'PESQUISAR'. Below the search bar, a message says 'Encontrados 10405 conjuntos de dados com 50202 recursos (jo que é isto?)'. A section titled 'Conjuntos de dados em destaque' lists 'A - Série histórica de preços de combustíveis - revenda' (with a note about weekly sales data) and 'Compras Públicas do Governo Federal' (with a note about the System Integrado de Administração e Serviços Gerais - SIAS-G). Another section, 'Publicações mais recentes', lists several datasets with their respective publishers and dates, such as 'PROJETO DE PESQUISA' by Fundação Universidade Federal de Se... on 22/10/2021, and 'Processos de Licitações e Preços - HC-UFPE' by Hospital das Clínicas da Universidade... on 20/10/2021. The bottom part of the page shows a 'Notícias' section with two articles: 'CGU lança pesquisa sobre demanda do setor privado por dados abertos governamentais' (published 25/08/2021) and 'Governo Federal promove Hackathon Ribeirão Preto' (published 24/08/2021).

This screenshot shows the 'Portal da Transparéncia de estados e municípios' for the city of Aracaju. The top banner features the text 'TRANSPARÊNCIA COVID-19' and 'Aracaju contra o Coronavírus'. The main content area includes a 'Portal da Transparéncia' section with a note 'O que é o portal da transparéncia?' and a 'Acompanhe aqui todas as informações para a maior transparéncia de nosso município.' section. Below this are several boxes for different municipal agencies: 'Prefeitura Municipal de ARACAJU' (with a note 'Acesse às informações. Faça seus questionamentos.'), 'Aracaju Previdência', 'SMTT', 'FUNCAJU', 'EMURB', and 'EMSURB'. Each agency box has its logo and name.

# Onde buscar os dados?

<https://brasil.io/>

The screenshot shows the Brasil.io homepage. At the top, there's a purple header with the Brasil.io logo and navigation links for 'Sua conta', 'Dados', 'Sobre', 'Blog', and 'AJUDE'. Below the header, the main title 'O Brasil em dados libertos' is displayed, followed by a subtitle 'Repositório de dados públicos disponibilizados em formato acessível'. A section titled 'Últimas atualizações' features a COVID-19 dataset update from July 2021. Another section highlights 'ESPECIAL COVID-19: DADOS POR MUNICÍPIO'. A large central area contains text about the COVID-19 data collection effort, mentioning 40 volunteers from 27 states. Below this, there are four cards: 'Manifesto' (with a link to understand the project's motivation), 'Datasets' (with a link to view the catalog), 'Sugira um Dataset' (with a link to contribute), and 'Colabore' (with a link to contribute). A 'VER MAIS' button is located at the bottom of the COVID-19 section.

<https://basedosdados.org/>

The screenshot shows the Basedosdados.org homepage. At the top, there's a navigation bar with 'BD' and links for 'Dados', 'Documentação', 'Contato', 'Sobre', 'Newsletter', 'Apóie', and 'Entrar'. A search bar is positioned above a central call-to-action: 'Um único lugar para buscar e acessar os dados que você precisa'. To the right of the search bar, there's a 'Pesquisar palavras-chave, insta...' input field and a 'Termos populares' dropdown. Below the main heading, there are three icons: a magnifying glass over a database icon, two cylinders with a download arrow, and a computer screen with R code. Text next to these icons explains the search function, the availability of treated and raw datasets, and the use of Python and R for data analysis. A 'Novidades nodatalake BD+' section features a preview of the 'Relação Anual de Informações Sociais (RAIS)' dataset from the Ministry of Economy (ME).

E tantos outros sites de organizações civis e projetos que trabalham com dados abertos. Vale visitar:

- Open Knowledge Brasil:  
<https://ok.org.br/>

- Escola de Dados:  
<https://escoladedados.org/>

- Querido diário:  
<https://queridodiario.ok.org.br/>



**Já tenho os dados.  
O que fazer com eles?**

# Pirâmide DIKW



# 5 formas diferentes para analisar os dados



Descriptiva



Inferencial



Diagnóstica



Preditiva



Prescritiva

# 5 formas diferentes para analisar os dados

## Análise Descritiva

### O que aconteceu no passado?

O foco é descobrir o que os dados podem contar sobre Quem, O quê, quando e onde sobre o tópico de interesse.

# 5 formas diferentes para analisar os dados

## Análise Inferencial

### E sobre o resto?

Esse tipo de análise aparece quando é impraticável ou impossível coletar dados de cada membro da população analisada.

# 5 formas diferentes para analisar os dados

## Análise Diagnóstica

### O que tem sob a superfície?

Esse tipo de análise envolve olhar além das observações iniciais, detalhando as categorias, procurando por anomalias, ou comparando os dados obtidos com outros dados não incluídos no início.

Fonte: Data Literacy Fundamentals: Understanding the Power & Value of Data

Ben Jones

<https://www.amazon.com.br/Data-Literacy-Fundamentals-Understanding-English-ebook>

# 5 formas diferentes para analisar os dados

## Análise Preditiva

### O que acontece em seguida?

Esse tipo de análise envolve usar dados históricos para previsões sobre resultados futuros. Os dados geralmente revelam tendências e padrões que podem ser utilizados para a construção de modelos que podem "prever o futuro".

# 5 formas diferentes para analisar os dados

## Análise Prescritiva

### O que deveríamos fazer?

Esse tipo de análise concentra-se na tarefa de quais ações podemos tomar após as análises prévias apresentadas. Em resumo, permite escolher a melhor alternativa para transformar decisões em decisões baseadas em dados.

# Analogia a um paciente visitando um médico



- **Análise descritiva:** o médico começa observando e verificando os sintomas do paciente.
- **Análise inferencial:** O médico pode fazer algumas inferências a outros pacientes ou estudos.
- **Análise diagnóstica:** O médico pode investigar um pouco mais para entender melhor a causa dos sintomas.
- **Análise preditiva:** A partir do que foi observado, o médico pode fazer algumas previsões de como a doença pode evoluir ao longo do tempo se não for feita alguma intervenção.
- **Análise prescritiva:** E finalmente, o médico prescreve medicações ou procedimento para aumentar as chances de recuperação do paciente.



Dados coletados e  
analisados. Onde entra a  
visualização nisso tudo?

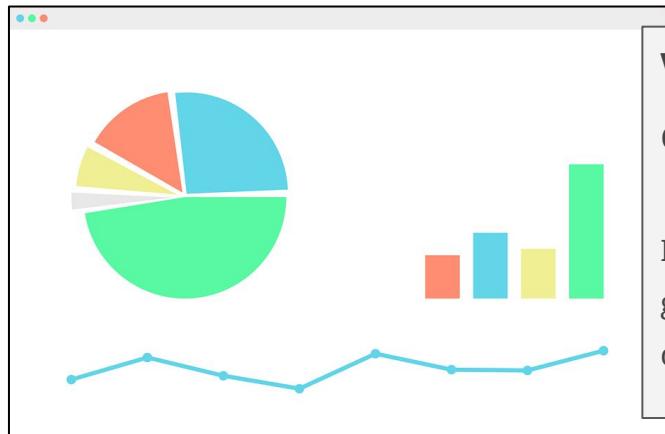


## Dados coletados e analisados. Onde entra a visualização nisso tudo?

A visualização é inerente a diversas etapas do processo de análise de dados. Desde o seu uso para entender os dados a visualização das análises e apresentação dos resultados.

# O que é visualização?

Consiste no estudo e na criação de representações visuais para os dados com o intuito de **ajudar as pessoas a realização tarefas como mais eficácia.**



**Visual Literacy:** capacidade de expor e interpretar dados de forma visual.

Isso vai além de construir gráficos. É preciso saber escolher o melhor gráfico para um determinado tipo de dados e entender a mensagem que os dados querem transmitir.

# Ainda sobre a definição ...

- A visualização é adequada quando há a **necessidade de aumentar as capacidades humanas** ao invés de substituí-las por métodos computacionais de tomada de decisão;
- A maioria das possibilidades de design para construção das visualizações são ineficazes, portanto **avaliar a eficácia** de um design é **muito importante**, portanto **não é uma tarefa simples**;
- Deve-se levar em consideração 3 limitações: os **computadores**, os **humanos** e os **meios de visualização**;
- O uso de uma visualização deve ser analisada em termos de **por que o usuário precisa dela, quais dados vão ser exibidos e como será projetada**.

## Ainda sobre a definição ...

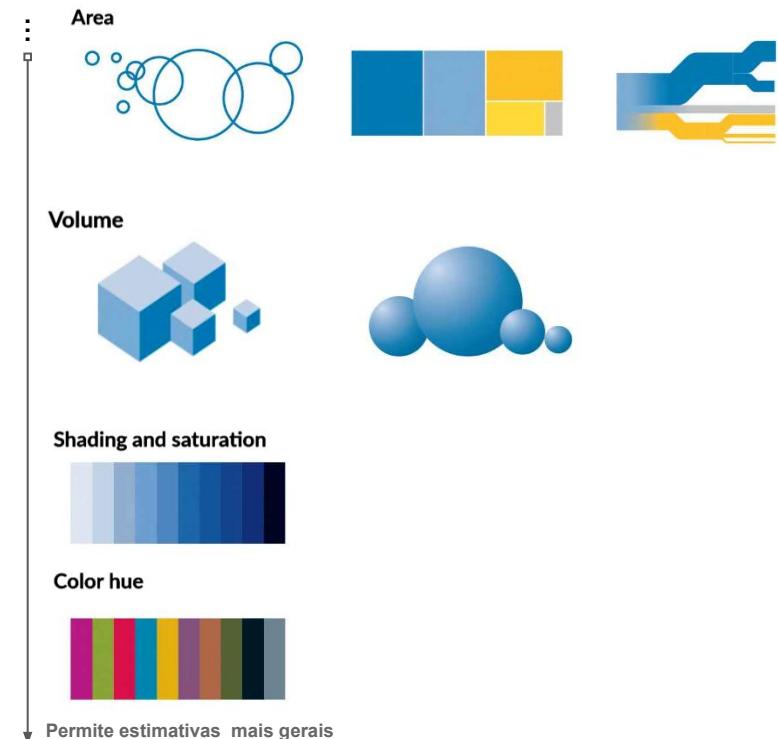
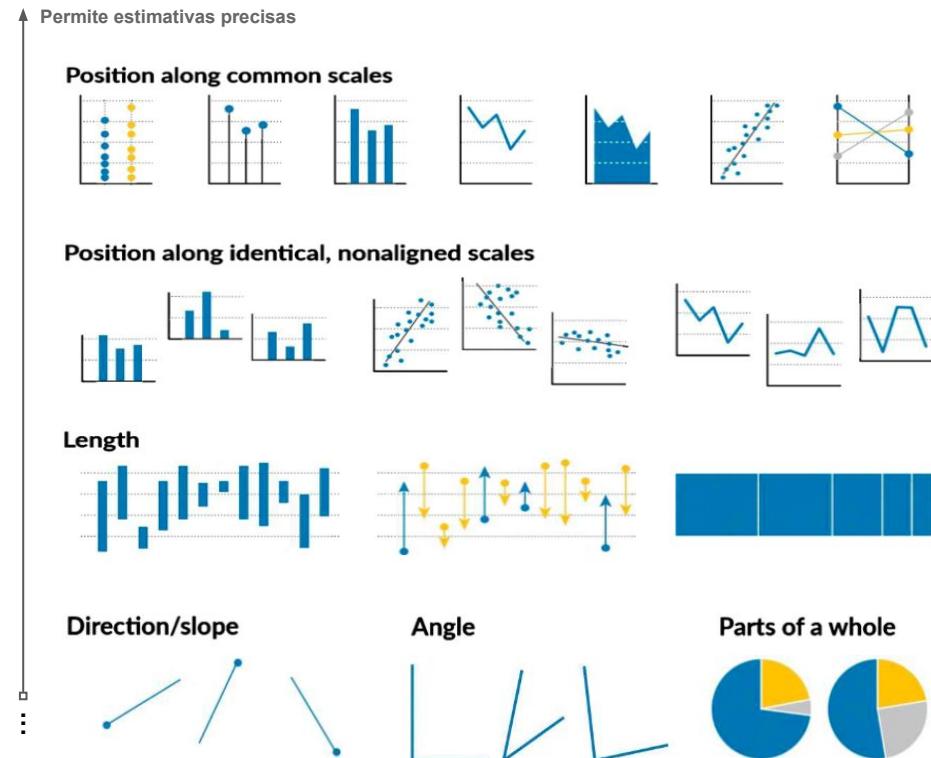
"A escolha do melhor design para as **visualizações** de dados, vai muito **além de questões técnicas**. O **fator humano**, o **público** e os **objetivos** são determinantes na seleção da visualização que será adotada".

# Não é só um gráfico ...

Construir uma visualização não é só construir um gráfico. É pensar nos diversos aspectos que influenciam como percebemos os objetos, as cores e os estímulos visuais.

# Alguns conceitos importantes

## Classificação de Percepção

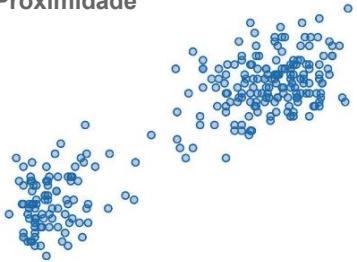


# Alguns conceitos importantes

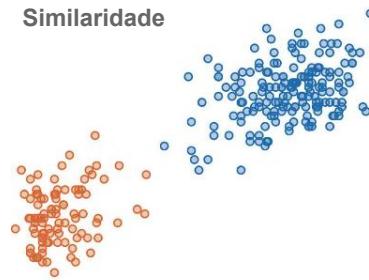
## Princípios de Gestalt

A teoria da Gestalt foi desenvolvida no início do século XX por psicólogos alemães e se refere a como tendemos a organizar os elementos visuais em grupos.

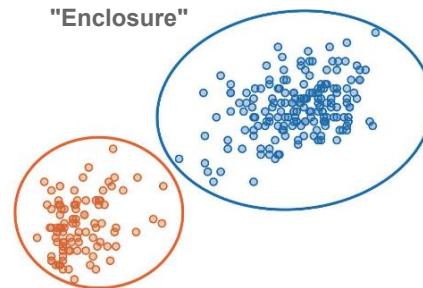
Proximidade



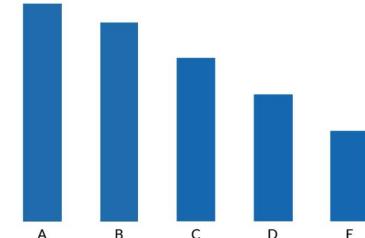
Similaridade



"Enclosure"



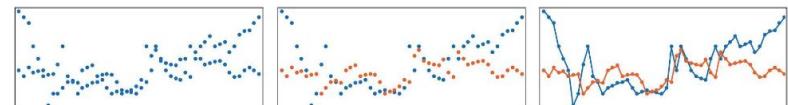
Continuidade



Continuidade



Conexão



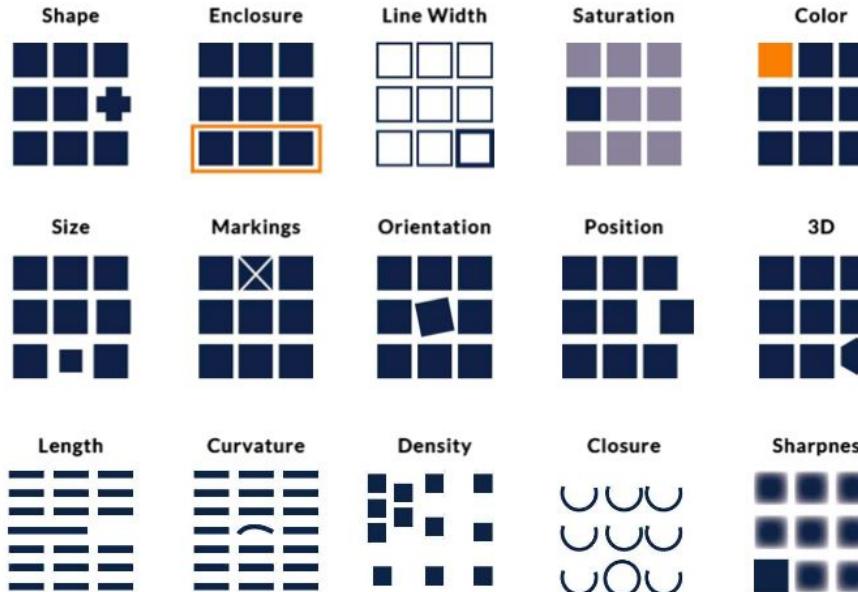
Fonte: Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks

Jonathan Schwabish

<https://www.amazon.com.br/Better-Data-Visualizations-Scholars-Researchers/dp/0231193114/>

# Alguns conceitos importantes

## Atributos pré-atenciosos



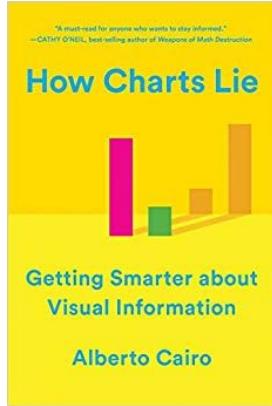
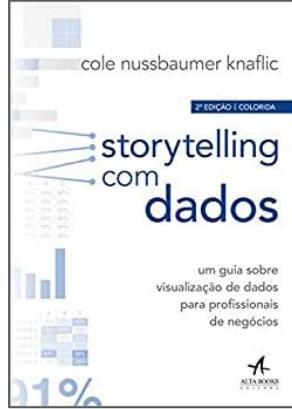
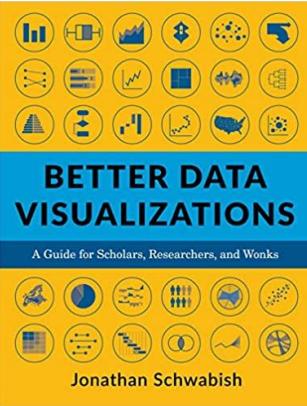
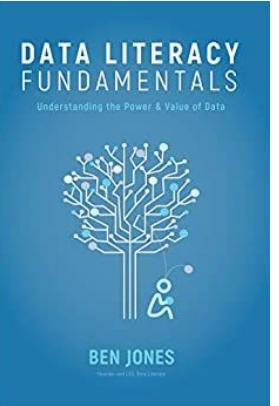
Exemplos de atributos de pré-atenciosos que podemos utilizar em nossas visualizações para direcionar a atenção do leitor

# Vamos ver um pouco disso tudo na prática :)

Acesse o repositório:

[www.github.com/adolfoguimaraes/minicurso\\_dadosabertos](https://www.github.com/adolfoguimaraes/minicurso_dadosabertos)

# Algumas dicas de material ...



**Escola de Dados**  
@EscolaDeDados

Ajudamos pessoas e organizações a transformarem dados em histórias, evidências e mudanças sociais. Junte-se a nós pela democratização dos dados no Brasil.

© Brasil Ⓛ [escoladedados.org/membrosia](http://escoladedados.org/membrosia) Ingressou em junho de 2013

**DataViz Brasil**  
@datavizbr Segue você

Publicação sobre visualização de dados em português: [medium.com/datavizbr](https://medium.com/datavizbr)

Curadoria por Rodrigo Medeiros (@medeiros\_rod) #dataviz

Traduzir bio

© Brazil Ⓛ [linktr.ee/datavizbr](https://linktr.ee/datavizbr) Ⓛ Nascido(a) em 1 de dezembro  
Ingressou em dezembro de 2017

**DataVis Lisboa**  
@DataVisLisboa

Data Visualization Lisboa is both a Meetup and a passion. We see great potential in making data more accessible and usable. #vislis #DataVis #lisboa

Traduzir bio

© Lisbon, Portugal Ⓛ [youtube.com/c/DataVisLisboa](https://youtube.com/c/DataVisLisboa)  
Ingressou em fevereiro de 2019

**PEIXE BABEL**  
#tecnologia, diversidade e sociedade  
@canalpeixebabel

Transmitindo conhecimento um bocado de cada vez! Tecnologia na voz de profissionais diversas 🌈 Por: @milaranjeira e @avivimota, e membro do @svbofficial

© Brasil Ⓛ [youtube.com/CanalPeixeBabel](https://youtube.com/CanalPeixeBabel) Ingressou em setembro de 2018

# E para finalizar ...



... isso não é mágica ...

# E para finalizar ...



... isso não é  
ficção científica ...

# E para finalizar ...



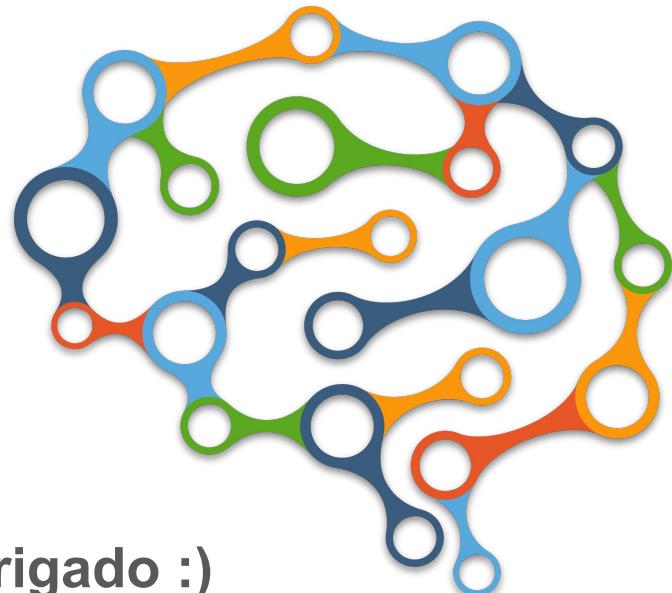
... isso é computação!

# Coletando, processando e visualizando dados abertos

Usando python para coletar, processar e visualizar dados disponibilizados pelo poder público.

Adolfo Guimarães

@profadolfoguimaraes



Obrigado :)