Introdução à Análise Exploratória de Dados

Noções Gerais sobre Data Science

João Pedro Albino

Departamento de Computação / Faculdade de Ciências

PPG-MiT / Faculdade de Artes, Arquitetura, Comunicação e Design

Oportunidade

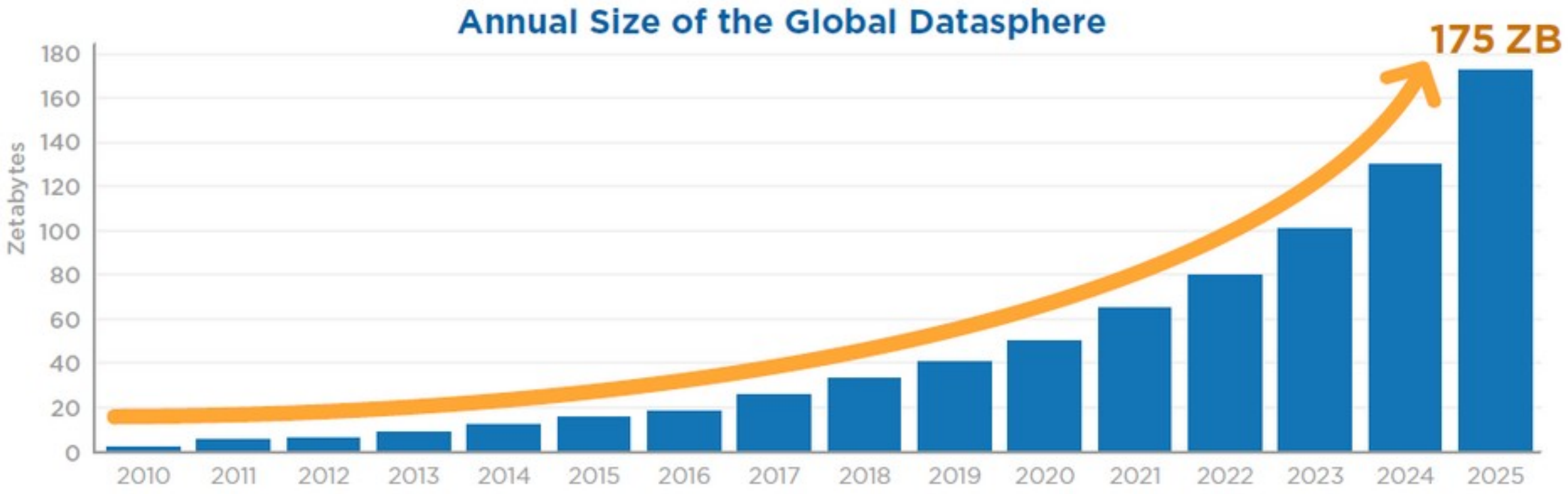
Data, data everywhere

Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier — but also big headaches.

The Economist, Feb. 27th, 2010

Era do Big Data

Figure 1 – Annual Size of the Global Datasphere



https://www.csa.gov.sg/singcert/-/media/Singcert/CyberSense/IDC_DataSphere.png

O que é Big Data?

- Conjunto de dados brutos
- Coletados e armazenados, por vários meios
- Podem ser analisados computacionalmente
  - Revelar padrões
  - Tendências
  - Associações
  - Especialmente em relação ao comportamento humano e suas interações.

https://www.integrity-ux.com.br/blog/2020/06/12/big-data-o-que-e-e-como-funciona/

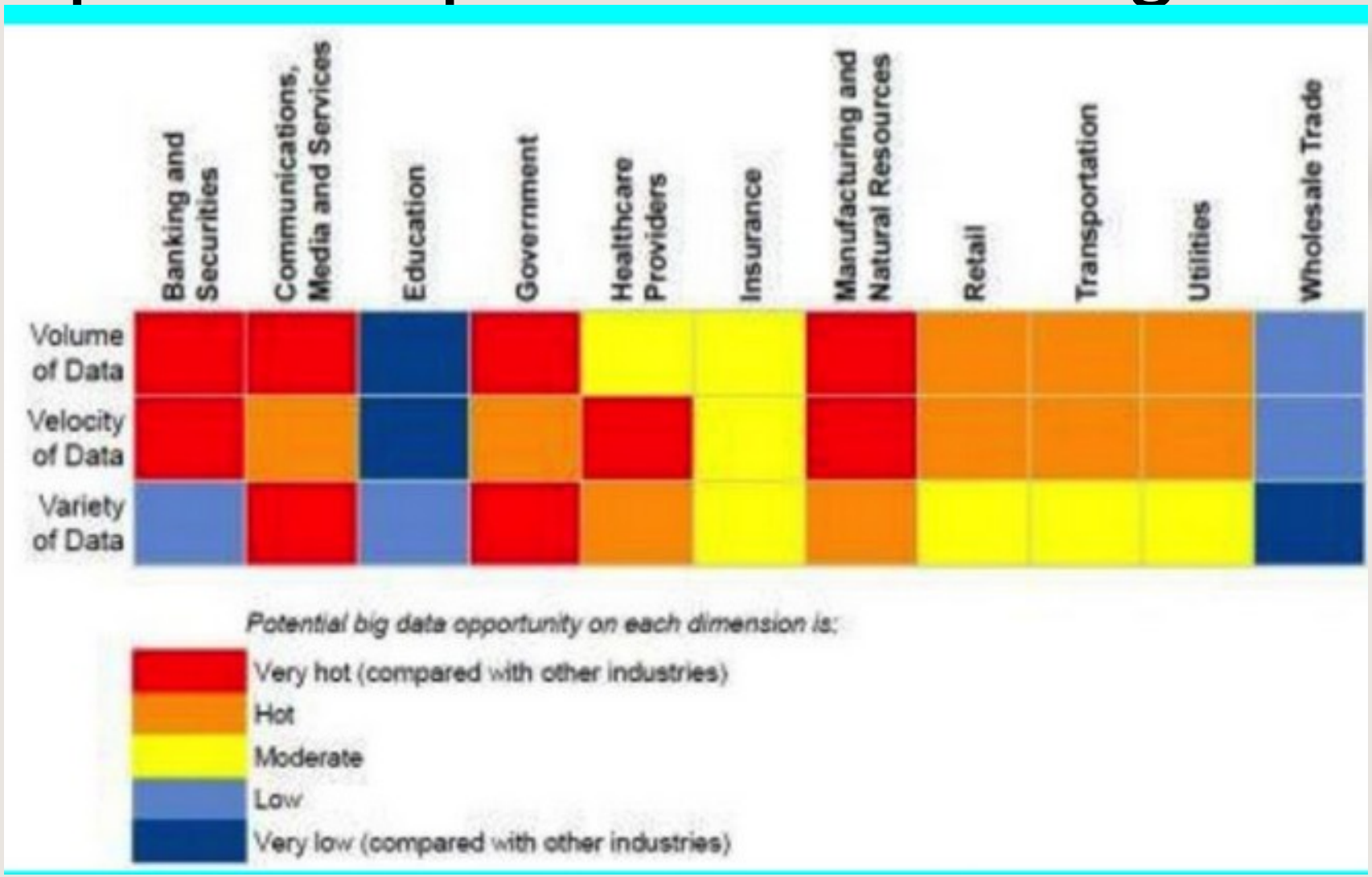
Os 5 V’s do Big Data

- **Volume**
  - É quantidade imensurável de dados que existe em todo o mundo hoje.
- **Velocidade**
  - Com o avanço das tecnologias, a produção de dados é mais veloz e a tomada de decisão mais rápida torna-se cada vez mais importante.
- **Variedade**
  - Com as diversas plataformas e meios de comunicação, as fontes de dados são mais variadas
- **Veracidade**
  - Garantia de que os dados utilizados estão corretos e são válidos
- **Valor**
  - Garantia de que os dados utilizados agreguem valor
  - O valor que os dados geram para os usuários e para os negócios

https://www.researchgate.net/figure/Os-5-Vs-do-big-data-e-suas-associacoes_fig2_342758257

Áreas de Aplicação do Big Data

Perspectivas para o uso de Big Data



https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article

Big Data: Principais Desafios

Desafios para o Big Data: Comunicações, mídia e entretenimento

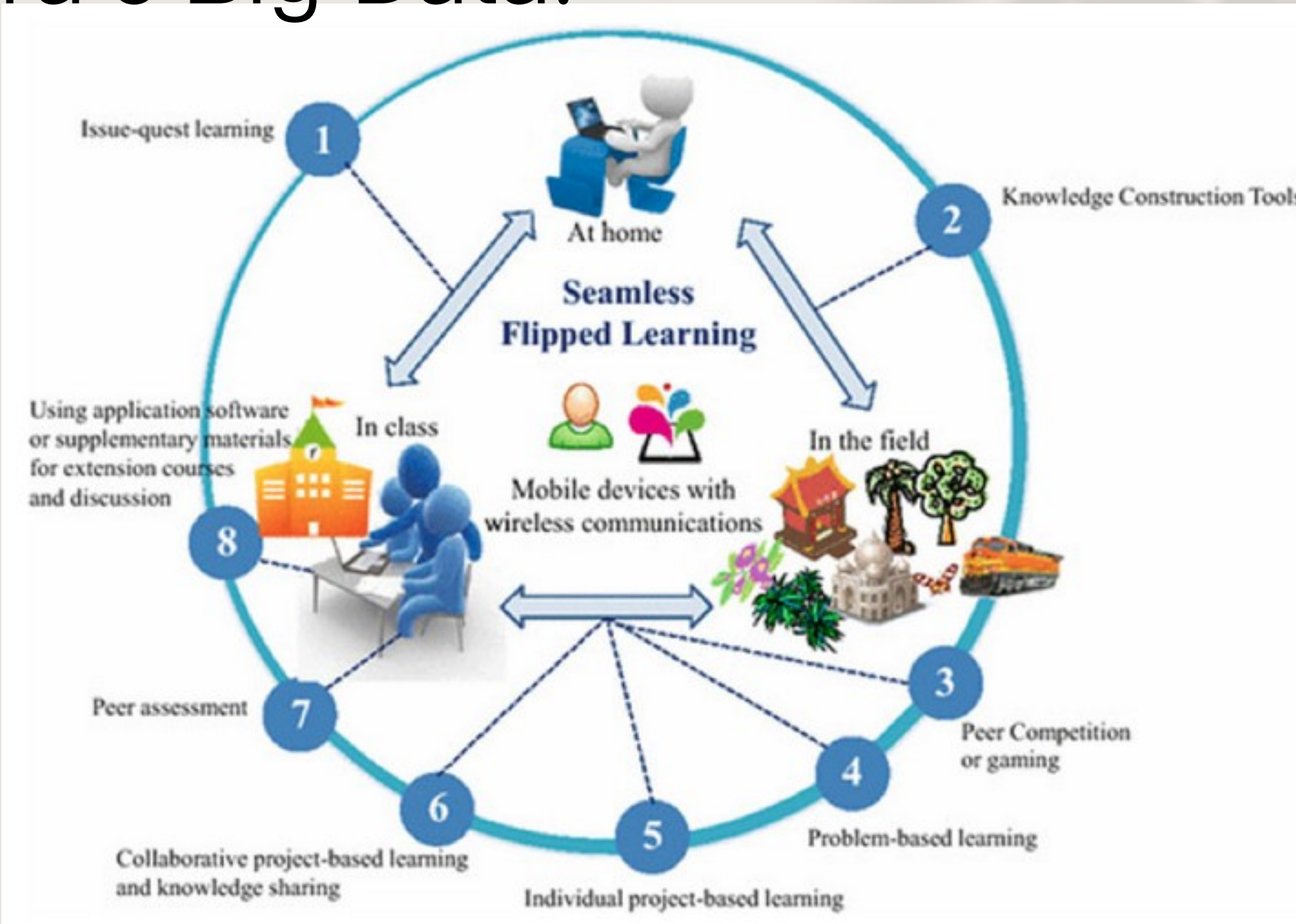
- **Consumidores esperam rich media**
  - *Vídeo, áudio ou outros elementos que incentivam a interação e*
  - *Envolvimento* **on-demand** em diferentes formatos e dispositivos
  - Coleta, análise e uso dos ***insights*** do consumidor
  - Aproveitar o conteúdo de mídia móvel e social
  - Compreensão dos padrões de uso de conteúdo de mídia em tempo real
- **Aplicações de big data**
  - Criar perfis detalhados de clientes
  - Criar conteúdo para diferentes públicos-alvo
  - Recomendar conteúdo sob demanda
  - Medir o desempenho do conteúdo
-

Desafios para o Big Data: Educação

- **Usado de forma significativa no ensino superior**
  - medir eficácia do professor
  - garantir boa experiência para alunos / professores
  - desempenho do professor ajustado / medido
  - variáveis
    - número de estudantes, assunto, demografia estudantil, aspirações estudantis, classificação comportamental
- **Departamento de Educação dos EUA**
  - grandes dados para análises
  - ajudar alunos que se distraem ao usar recursos on-line
  - padrões de clique usados para detectar aborrecimento

Desafios para o Big Data: Educação

- Seamless Learning



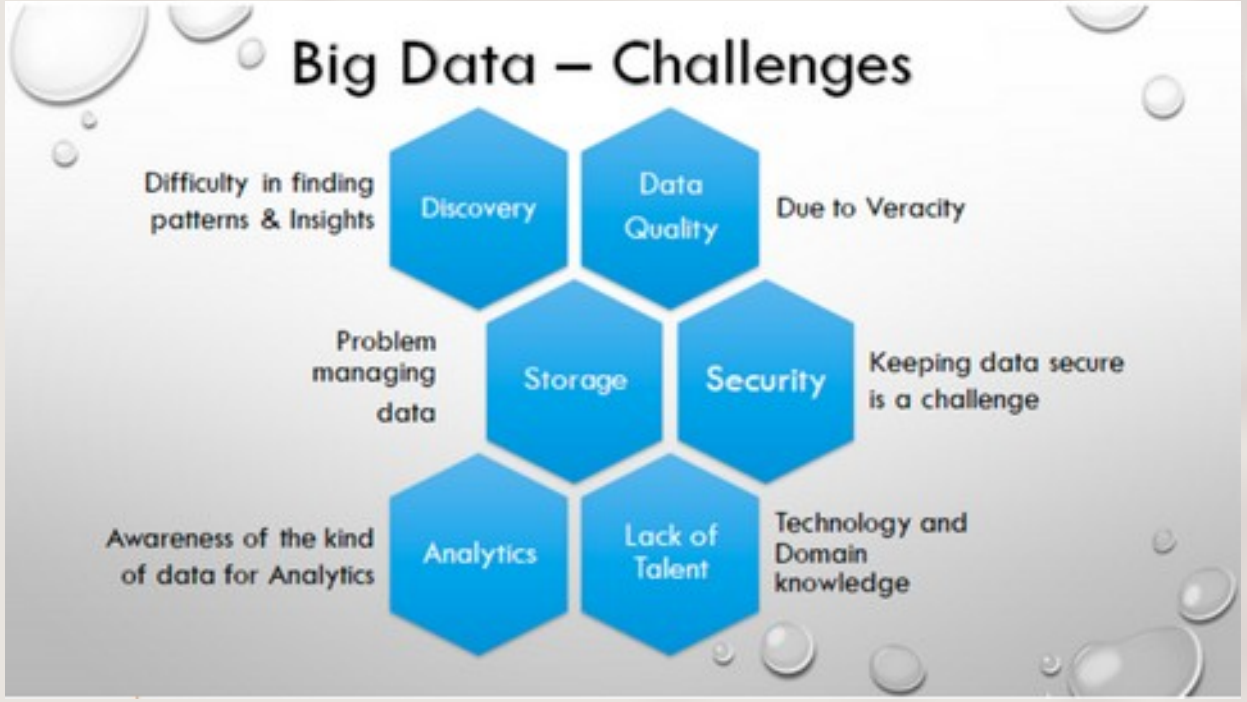
Desafios para o Big Data: Educação

- Learning **Analytics**

https://medium.com/@limaolflavio/a-import%C3%A2ncia-do-learning-analytics-na-avalia%C3%A7%C3%A3o-do-aprendizado-on-line-b43d

Desafios para o Big Data: no geral

- Qualidade e veracidade dos dados
- Segurança: manter os dados seguros ainda é um desafio
- Falta de profissionais com domínio na tecnologia e na área de aplicação
- Consciência do tipo de dado para análise
- Problema para armazenamento e gerenciamento dos dados
- Dificuldade em descobrir padrões e *insights*



Como trabalhar com Big Data

Ciência de Dados / Data Science

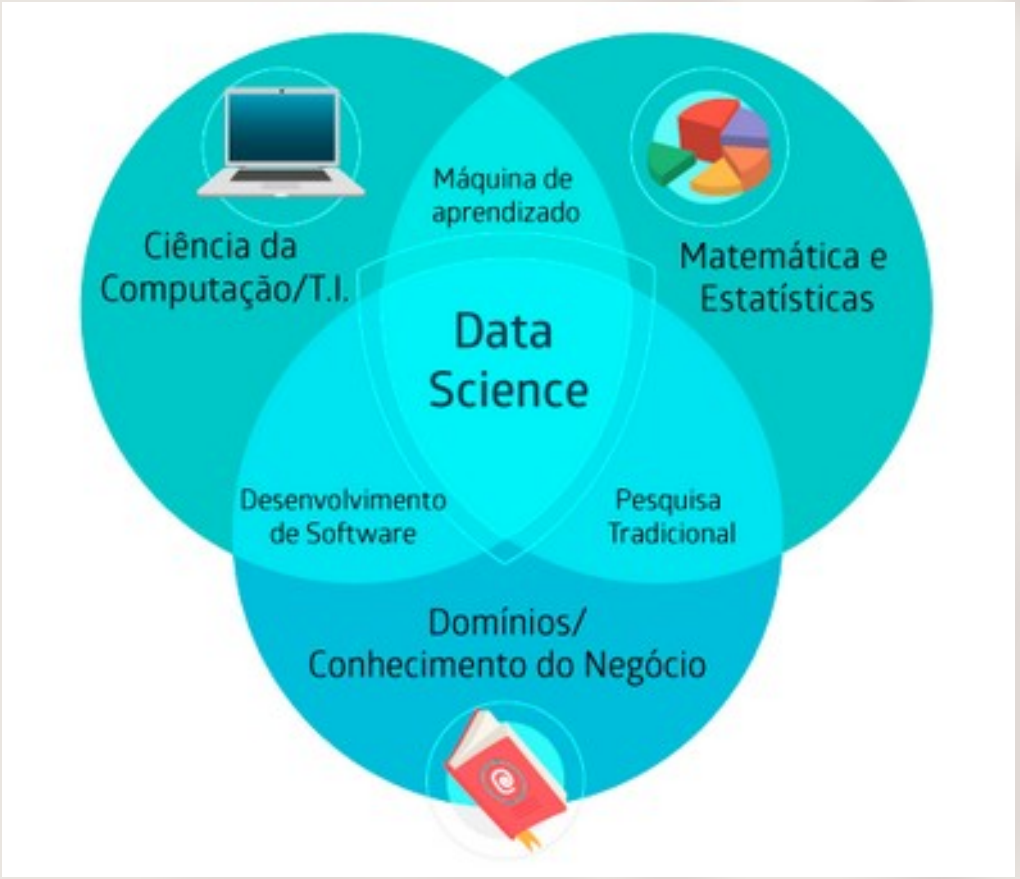
- **Definição**
  - *Termo utilizado para determinar uma combinação de várias ferramentas, algoritmos e princípios de aprendizado de máquina responsáveis pela descoberta de padrões e insights a partir de dados brutos*.
- Organizações enfrentam o desafio de lidar com enorme fluxo de informações geradas pela sociedade (big data)
- Saber utilizar a ciência de dados passou a ser um diferencial de negócios.
- É uma ferramenta valiosa para explorar e processar esses grandes volumes gerados por meio de diversas fontes.

O que é Data Science ?

Campo de estudo multidisciplinar que engloba dados, algoritmos e tecnologias capazes de extrair valor de dados estruturados ou não e resolver problemas analiticamente complexos.

É uma abordagem mais profunda, técnica e especializada sobre os elementos digitais.

Inclui o uso de modelos estatísticos e matemáticos, bem como de outras ferramentas para visualizar os dados.



O que é Ciência de Dados

Extração de conhecimento (insights) de grandes volumes de dados estruturados ou não estruturados

Uma ampliação dos campos de mineração de dados e análise preditiva

Cientista de Dados: Profissão dos anos 20xx?

Artwork: Tamar Co
on a page from a h

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

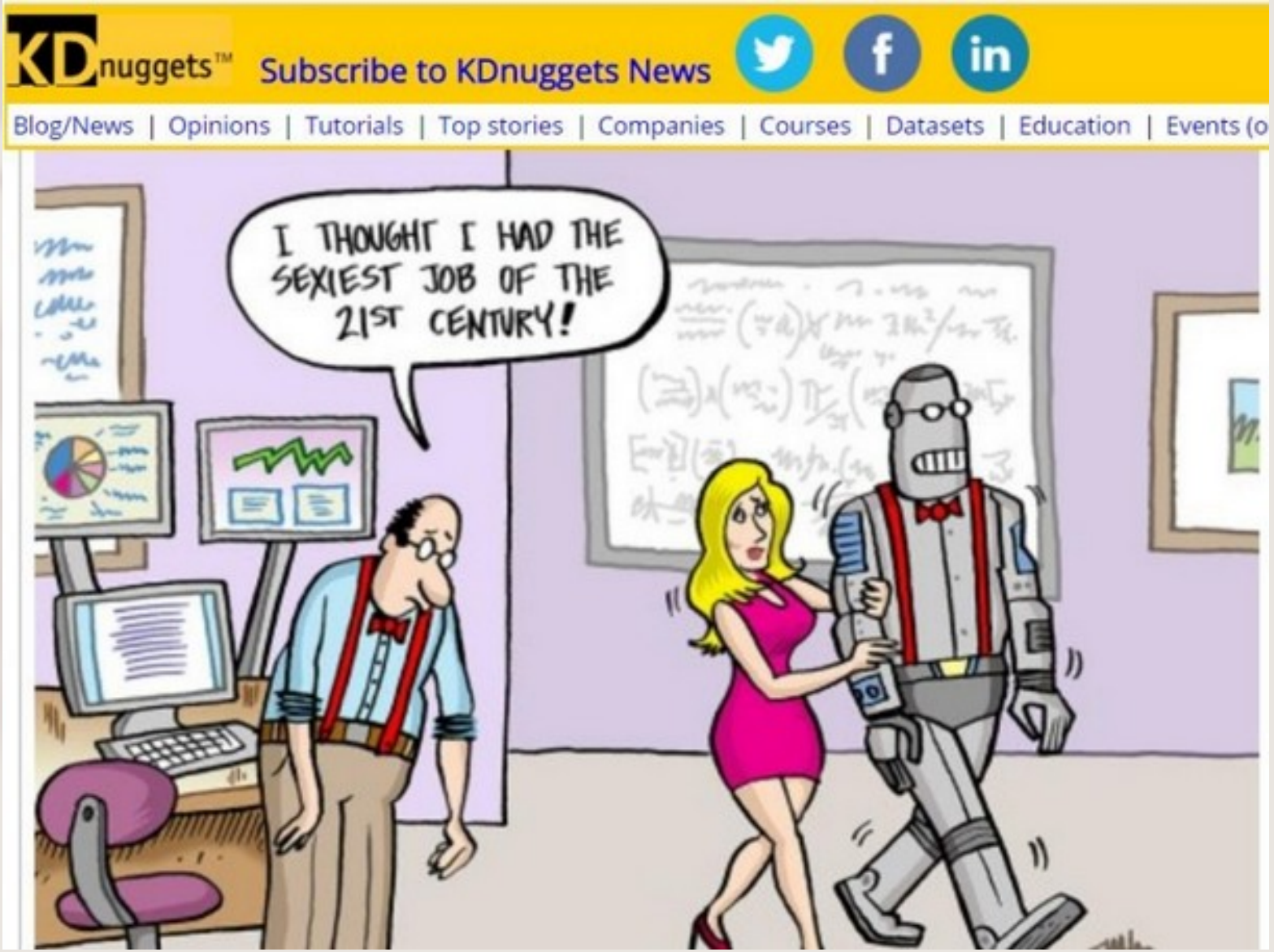
Cientista de Dados: valores fundamentais

Um Cientista de Dados é alguém que sabe obter, tabular, explorar, modelar e interpretar dados, combinando e utilizando estatística e aprendizagem de máquina.

Cientistas de Dados não somente são adeptos a trabalhar com dados, mas apreciam esses dados como um produto de primeira classe.

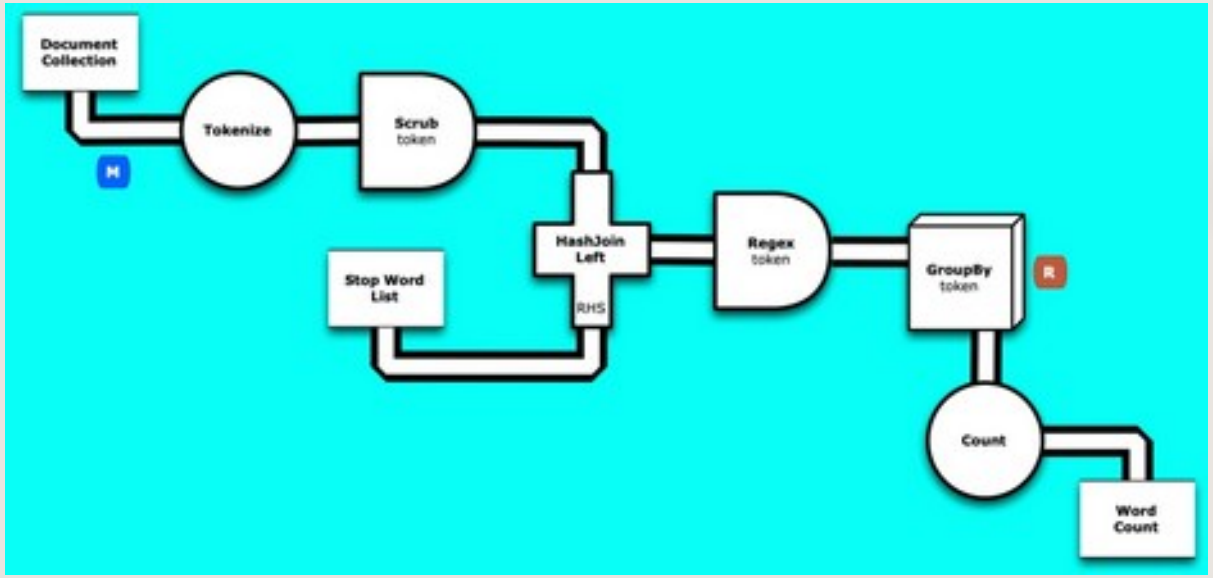
Hillary Mason
Cientista de Dados
Accel

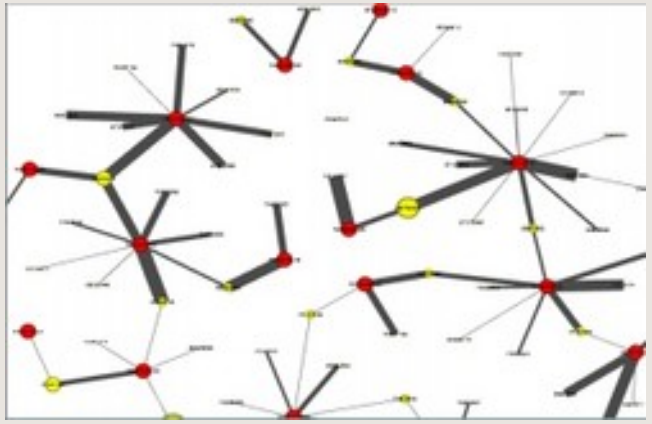
Habilidades de um(a) Cientista de Dados

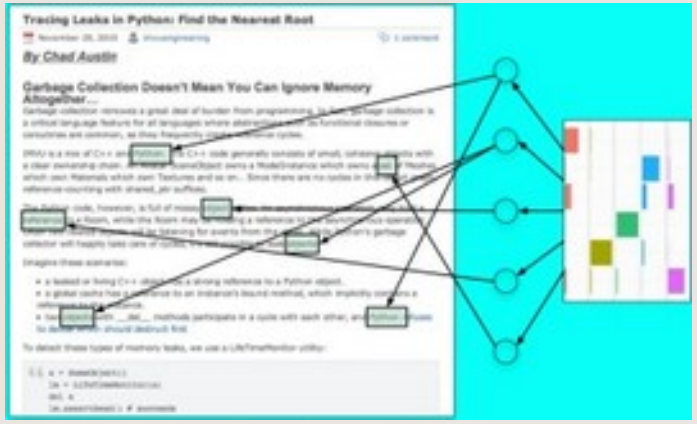


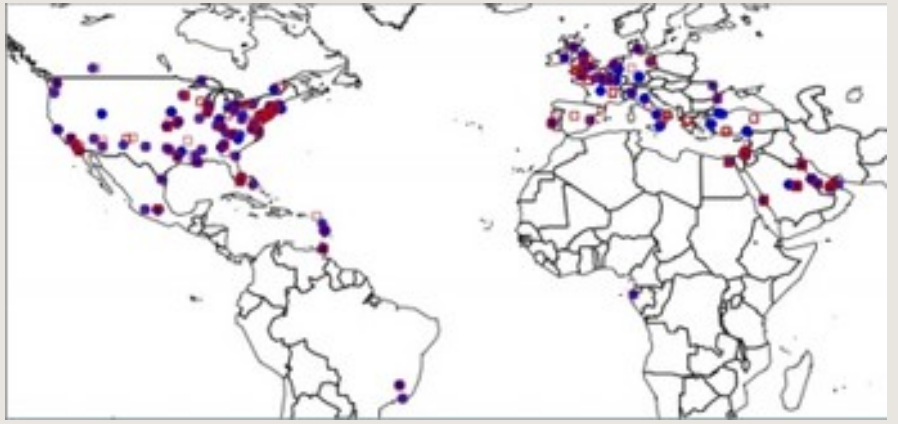
Cientista de Dados: Realidade em 202x?

Ciência de Dados: bastidores









Pensamento Estatístico

um modo de raciocínio que inclui tanto o *raciocínio lógico quanto o analítico* :

- avaliar a **totalidade** de um problema, bem como suas **partes componentes**;
- procura avaliar os efeitos na mudança em uma ou mais variáveis!

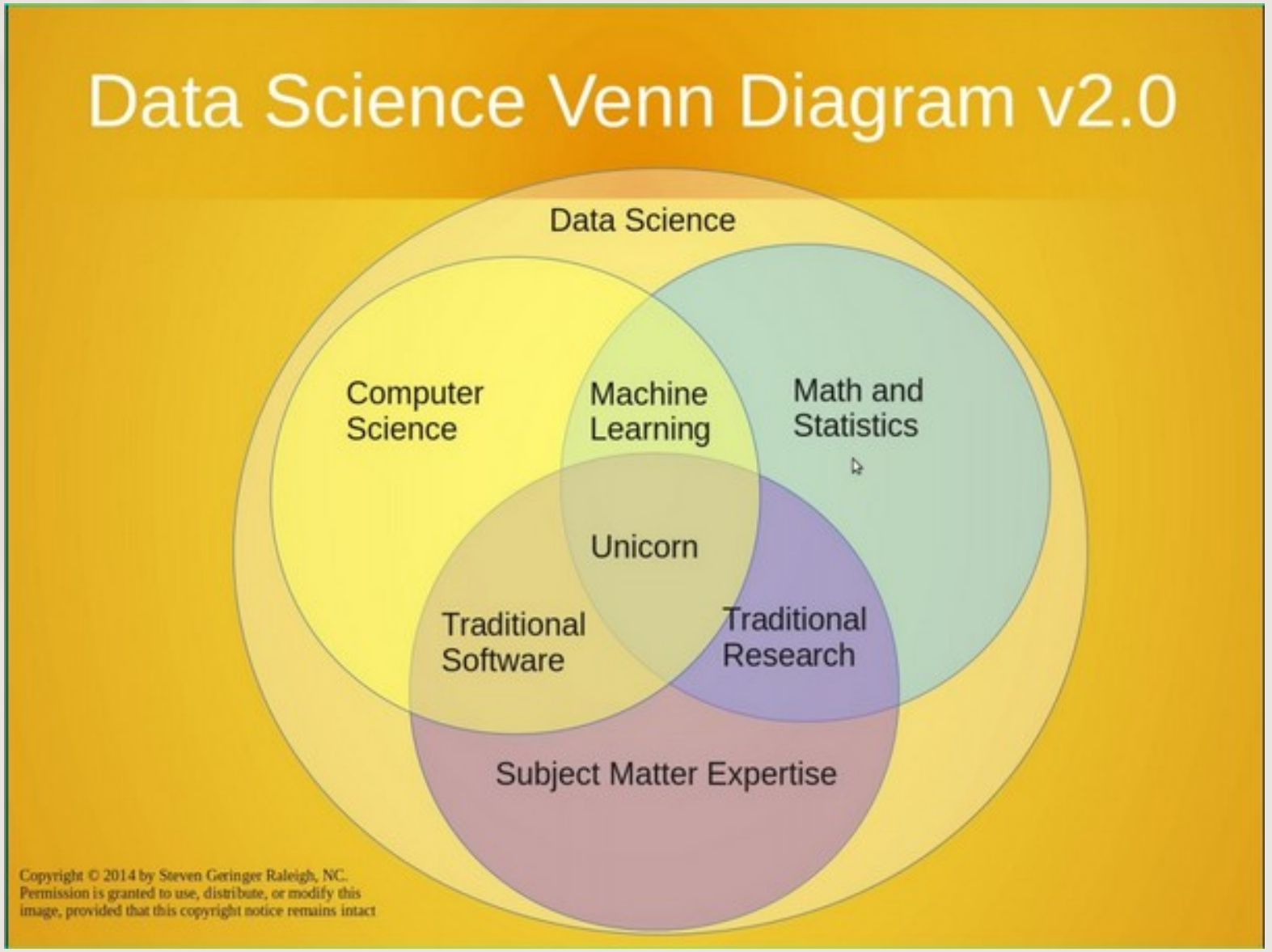
esta abordagem tenta entender não apenas *problemas* e *soluções*,

- mas também os processos envolvidos e suas variações!

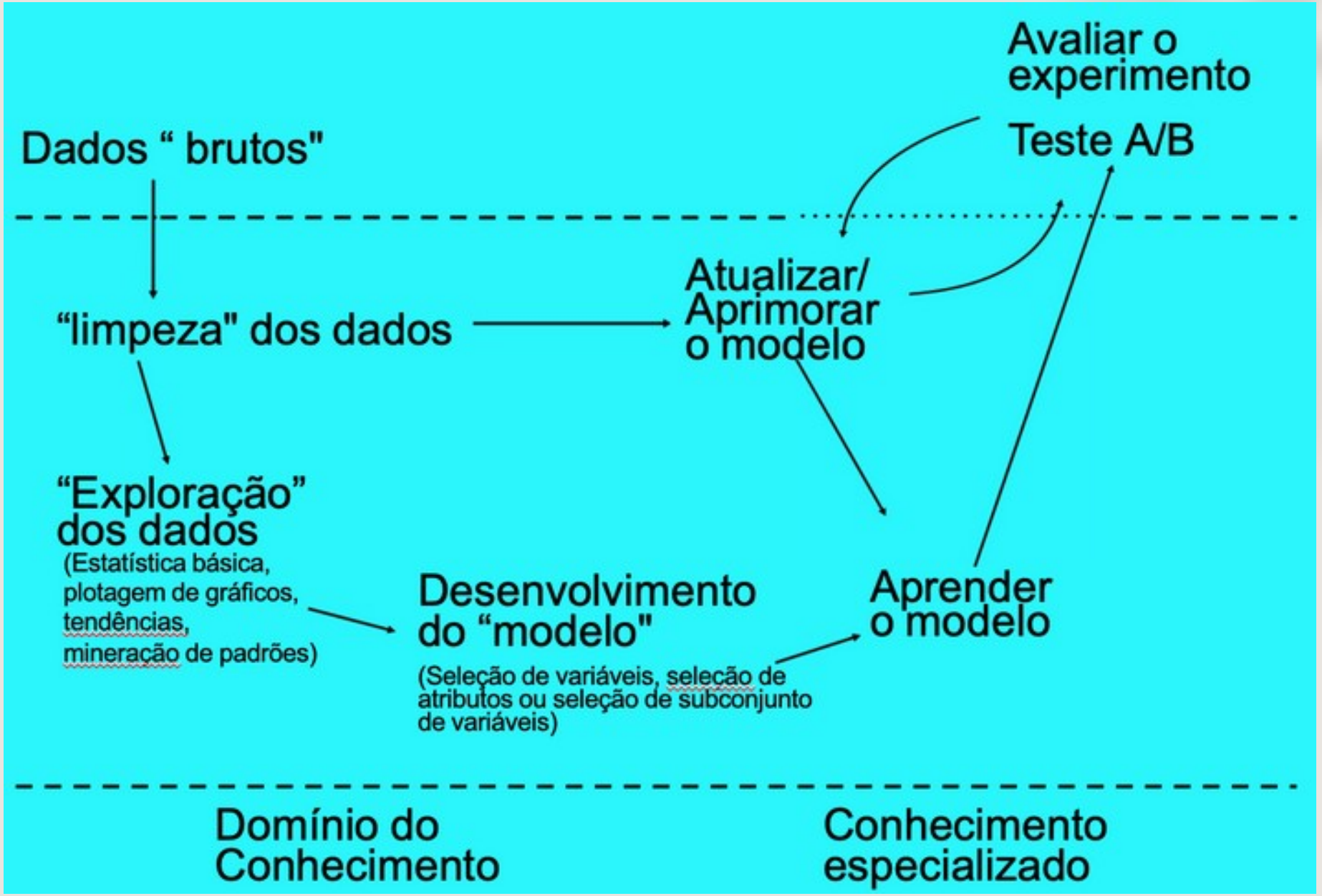
valioso no trabalho do Big Data quando combinado com experiência prática em comunicação

- p.ex.:aproximadamente 50% dos pares vêm do jornalismo ou de rádio e tv ...
- programadores normalmente não pensam assim!

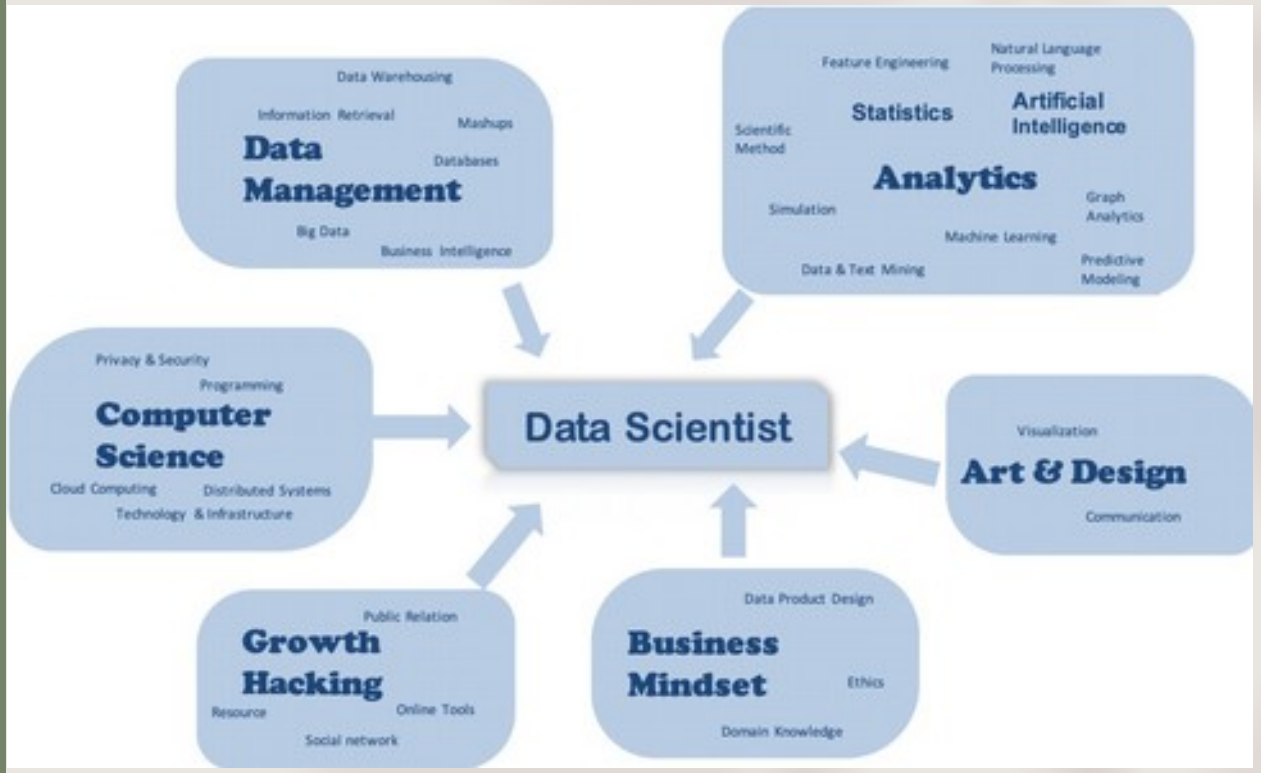
Disciplinas em Ciência de Dados



Fluxo de Ciência de Dados



Quem é o cientista de dados?



Algumas “pedras” no caminho

- **Data wrangling** ou **data preparation** – significa *preparação de dados*.
    - Conceito recente e diz respeito ao ato de **coletar**, **limpar**, **normalizar**, **combinar**, **estruturar** e **organizar** os dados que serão analisados.
- Aproximadamente 80% do custo dos projetos relacionados à análise de dados (**tempo + dinheiro**) são gastos na preparação destes dados (**data wrangling**) - principalmente nas questões de **limpeza** e garantia de **qualidade**.
- Infelizmente, os orçamentos relativos a dados tendem a entrar em frameworks que só podem ser utilizadas após a sua limpeza.

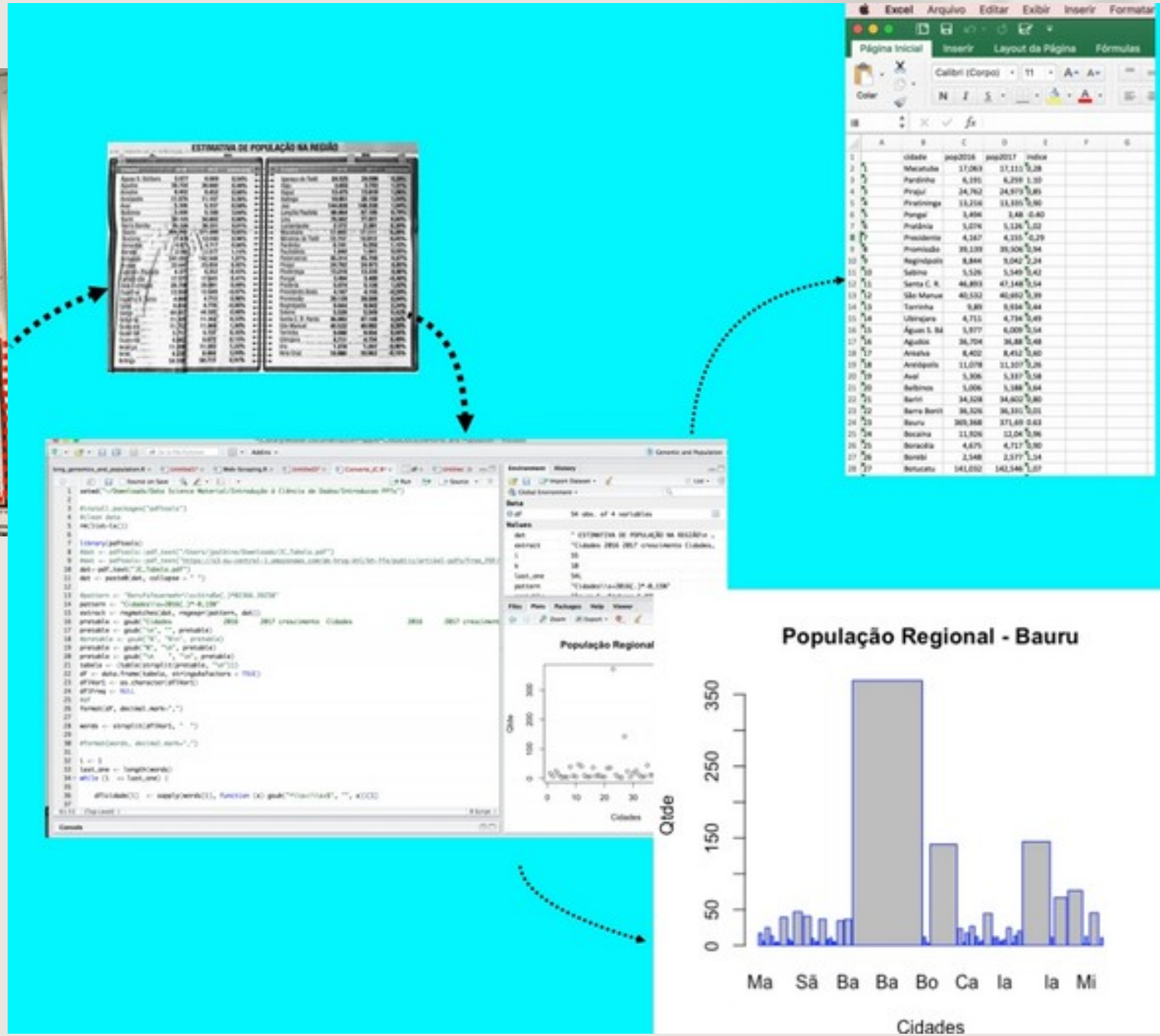
Ultrapassando obstáculos para obter “dados”

- “Criar" (gerar) os próprios dados, através de instrumentos (p.ex. questionários, surveys , etc.)
- Obter os dados através de APIs web, interfaces providas por bases de dados e por várias aplicações web modernas (incluindo Twitter, Facebook, dentre outras).
- Bases de dados "abertas", repositórios de dados, etc. (p.ex.: https://www.cienciaedados.com/o-poder-do-open-data/)
- Extrair as informações de arquivos (p.ex. PDF).
- Extrair informações de telas dos sites (scraping).

https://portal.fiocruz.br/ciencia-aberta-na-fiocruz

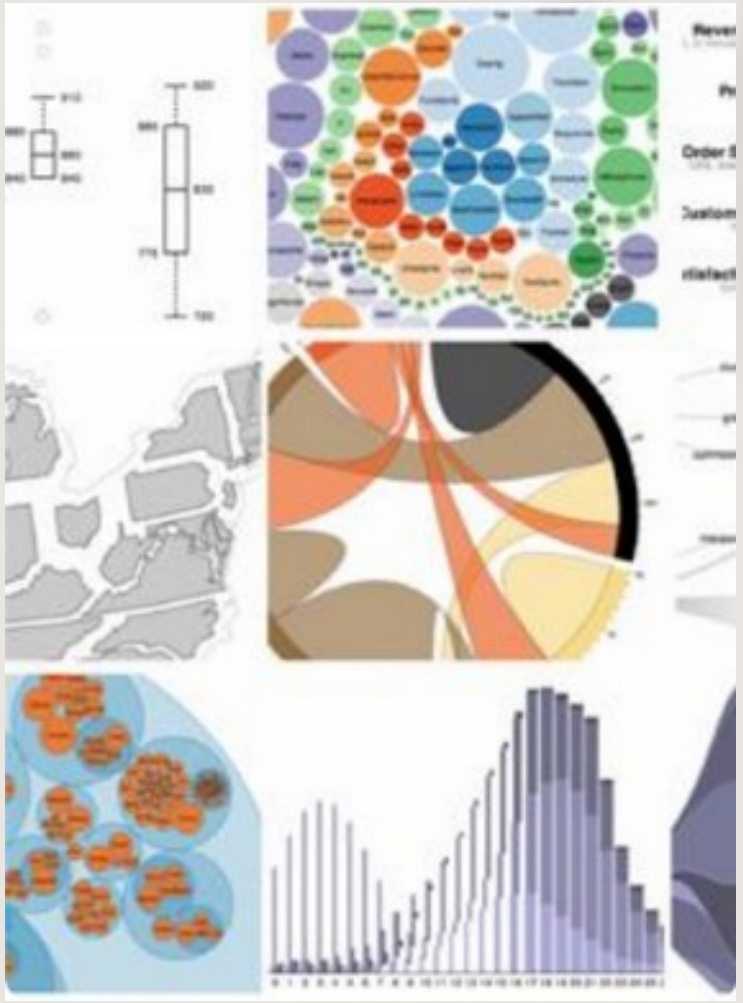
Exemplo


Com um crescimento estável, Bauru
ating 371.690 habitantes, diz IBGE

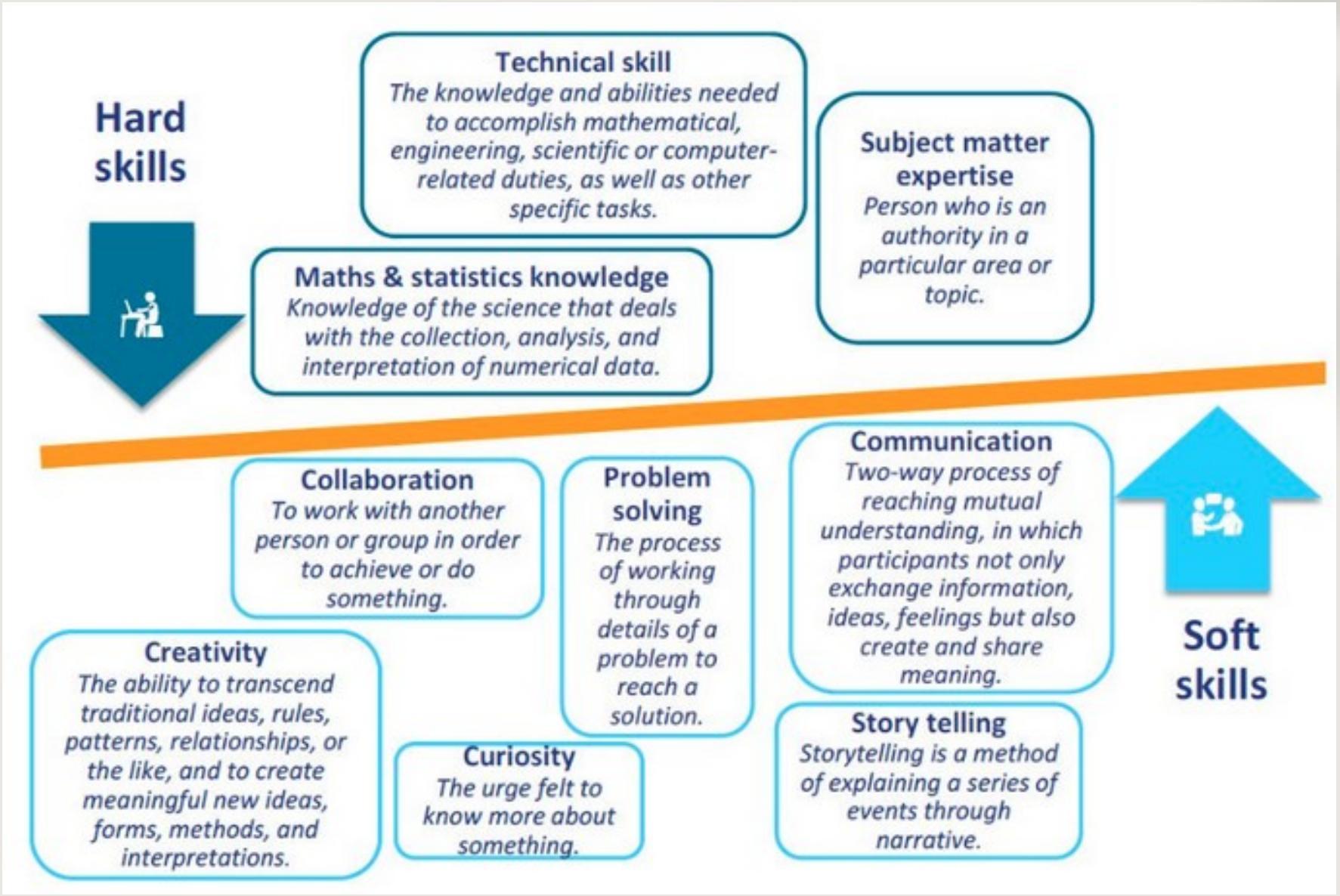

População Regional - Bauru
Qtde
Cidades
0
50
150
250
350
Ma
Sã
Ba
Ba
Bo
Ca
Ia
Ia
Mi

Habilidades mais importantes

- Aprender a utilizar ferramentas programáveis que preparem os dados
- **Aprender a gerar visualizações de dados convincentes**
- Aprender a estimar a confiança dos resultados relatados
- Aprender a automatizar o trabalho, tornando a análise repetível
- Outras importantes habilidades
    - Modelagem
    - Algorítmos
    - Etc.



Balancear “hard skills” e “soft skills”



Porque os cientistas de dados necessitam de ferramentas de visualização?

Porque os cientistas de dados necessitam de ferramentas de visualização?

Dar sentido aos dados ganhou grande importância neste século 21.

Programar é uma das formas de se manusear os dados disponíveis e torná-los utilizável.

Entretanto, nem todo mundo é “programador” e para tais pessoas os cientistas de dados podem usar ferramentas de visualização de dados para “contar estórias” .

Visualização de dados

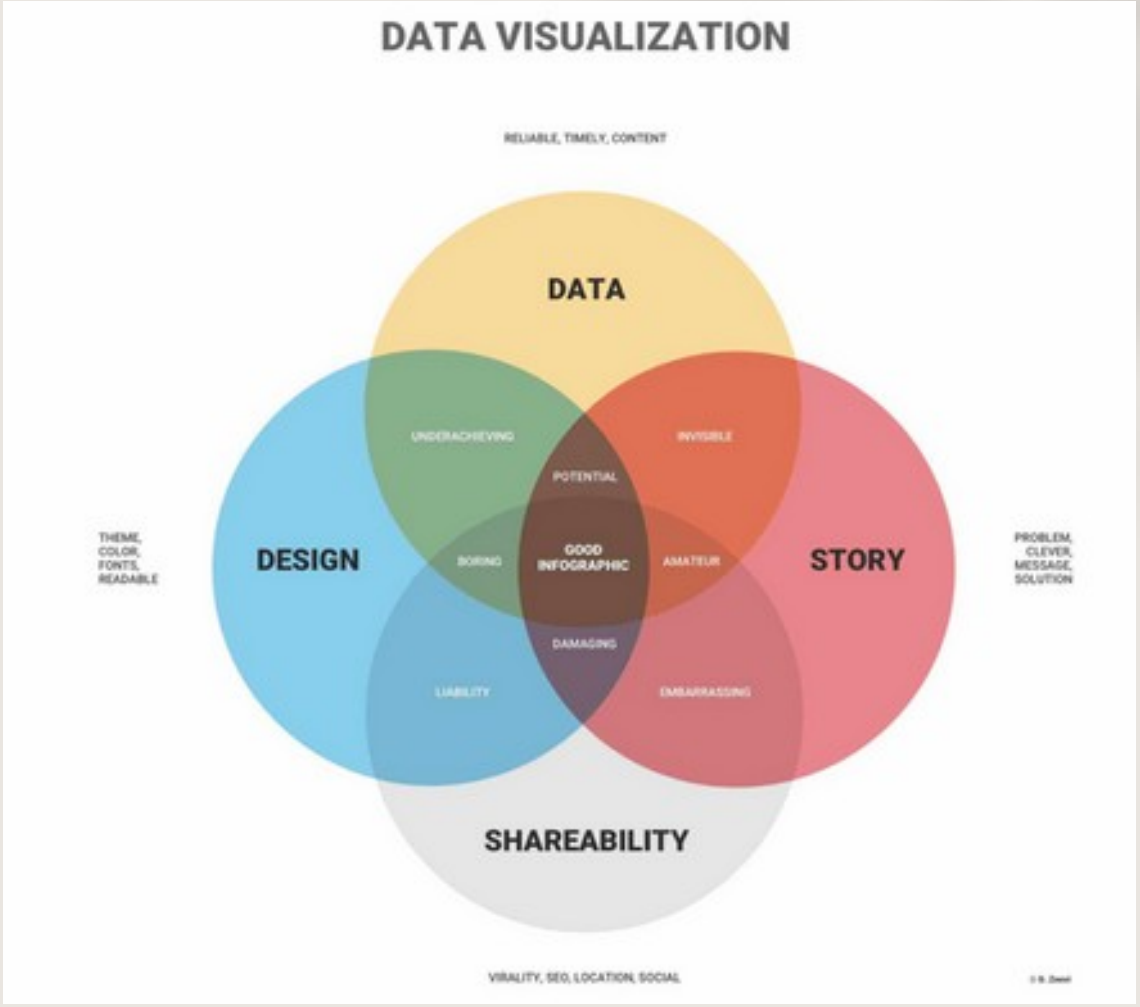
Visualização de dados é a apresentação de informações quantitativas em uma forma gráfica.

As visualizações de dados transformam conjuntos de dados (datasets) em visuais para o cérebro humano entender e processar.

https://infogram.com/pt/pagina/visualizacao-de-dados

Boas visualizações contam boas “estórias”

- Boas visualizações: ***comunicação + ciência de dados + design***
- Utilizadas para descobrir fatos e tendências desconhecidos
- Entregam informações importantes sobre conteúdos complicados de *datasets* de maneiras significativas e intuitivas
- Eduard Tufte
  - ”Ideias complexas comunicadas com clareza, precisão e eficiência"



https://infogram.com/pt/pagina/visualizacao-de-dados

Por que a visualização de dados (dataviz) é importante?

Melhor tomada de decisão

- Organizações fazem melhores perguntas e tomam melhores decisões baseadas em dados.
- Ênfase em monitoramento e desempenho: painéis de dados (dashboards) e KPIs.

Storytelling Significativo

- Data Storytelling: técnicas que orientam a apresentação de informações e insights de dados
- Gráficos de informações (infográficos): essencial para a grande mídia atual no jornalismo de dados.
- Empresas: Cientistas de dados apresentam análises dos dados e os resultados utilizando data storytelling
- Marketing: combinação de dados de qualidade e narrativa emocional

Alfabetização de dados

- Ser capaz de compreender e ler as visualizações de dados tornou-se um requisito necessário para o século XXI.

: https://www.datatelling.eu/what-is-data-storytelling/

Inovação

Um(a) grande cientista de dados deve ser inovador(a) e criativo(a) com as habilidades que possui.

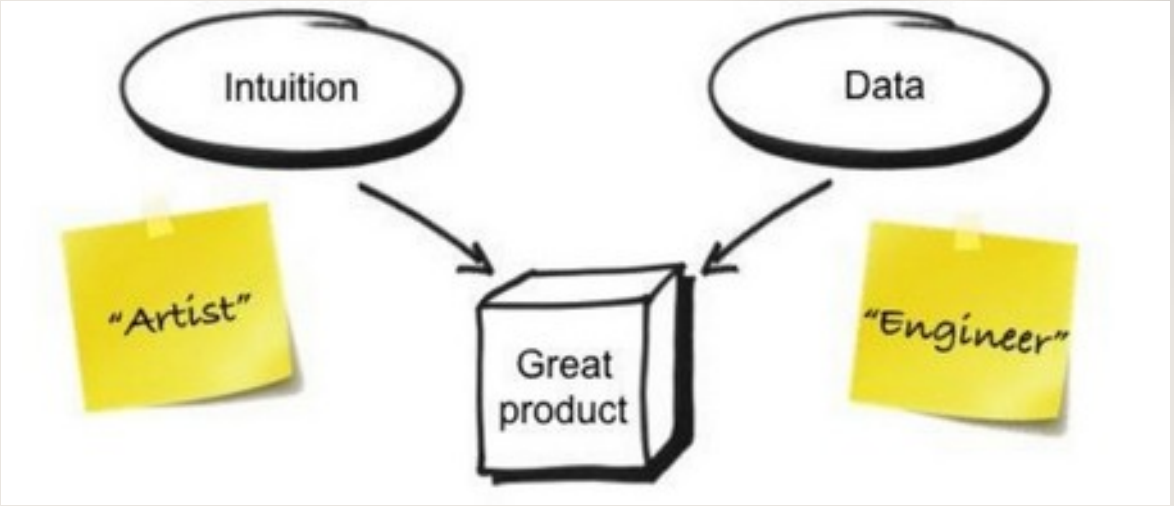
A criatividade de um(a) cientista de dados o(a) ajuda a determinar onde os dados podem agregar valor e trazer resultados lucrativos para a sua “ aplicação".

Intuição sobre os dados

Um bom cientista de dados não é aquele que apenas adiciona todas as possíveis características em um modelo de aprendizagem de máquina e analisa os resultados.

A coisa mais importante que um cientista de dados deve fazer antes de alimentar um modelo de aprendizagem de máquina é pensar se aquele modelo tem sentido!

Um cientista de dados deve ter **intuição sobre os dados**!



If you torture the data long enough, it will confess to anything.

Ronald Coase

https://i.redd.it/ikdymainj6p21.jpg

Introdução à Análise Exploratória de Dados

Noções Gerais de Data Science

João Pedro Albino

Departamento de Computação / Faculdade de Ciências

PPG-MiT / Faculdade de Artes, Arquitetura, Comunicação e Design