

Ciência de Dados

Apresentação

Profa. Dra. Roseli A. F. Romero

Material organizado: Dr. André C. P.
L. F. de Carvalho e Dr. Isvani Frias-
Blanco
ICMC-USP



© André de Carvalho - ICMC/USP



Objetivo

- *Introduzir os principais conceitos, técnicas e ferramentas referentes a ciência de dados.*
- *O curso visa prover teoria e prática a fim de que os alunos possam aplicar as novas técnicas e ferramentas estudadas em problemas reais.*



Ementa

- Ciência de Dados
- Descoberta de Conhecimento em Bases de Dados
- Data Mining
- Preparação de dados
- Pré-processamento de dados
- Modelagem de dados
- Estudo de algoritmo preditivo simples (k-NN)
- planejamento de experimentos
- Análise de resultados experimentais



Ciência de Dados

- Estuda princípios, métodos e sistemas computacionais para extrair conhecimento de dados
- Pergunta chave da área:
 - Como encontrar de forma eficiente padrões em (grandes) conjuntos (fluxos) de dados

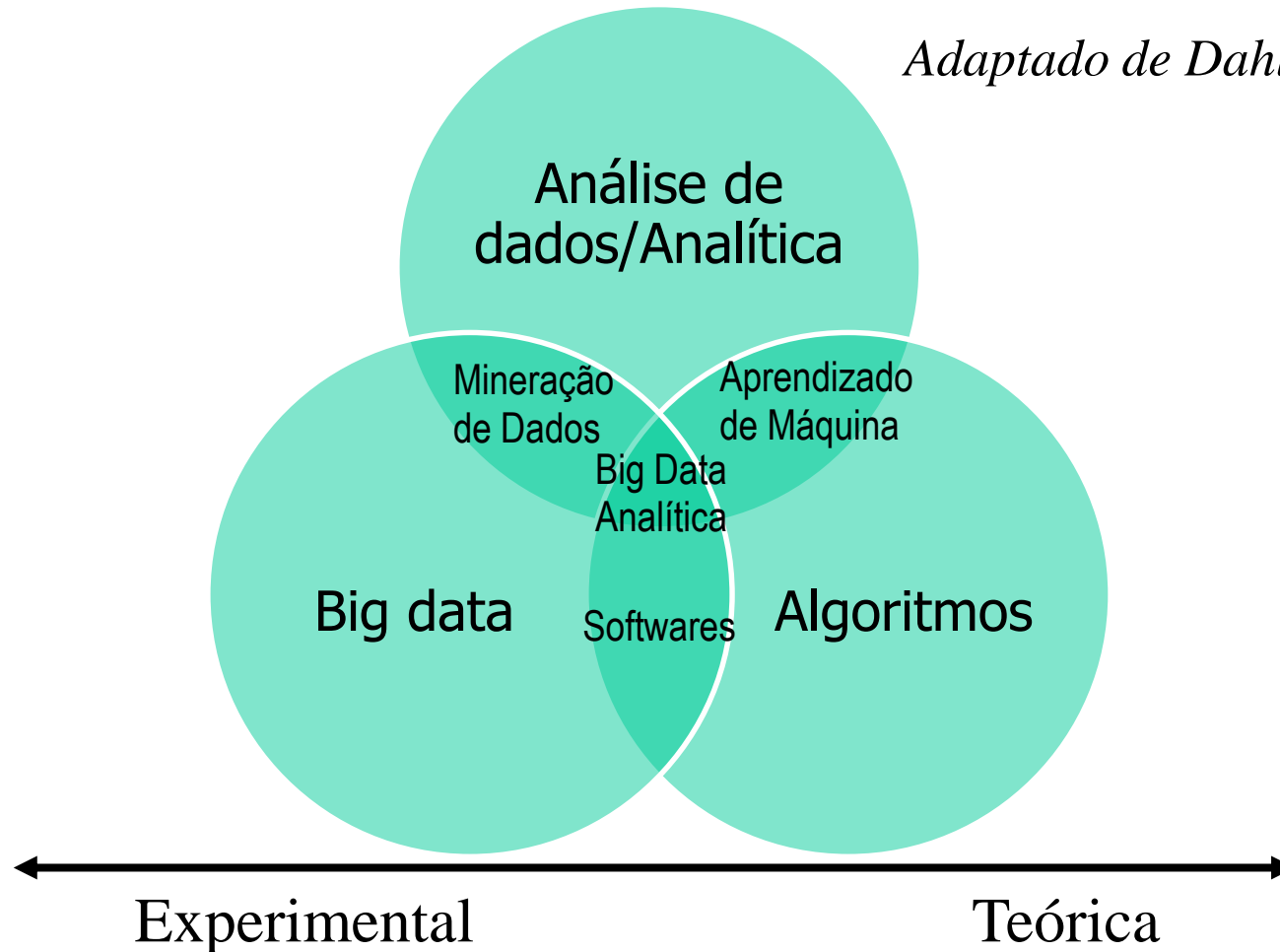


© André de Carvalho - ICMC/USP



Áreas da Ciência de Dados

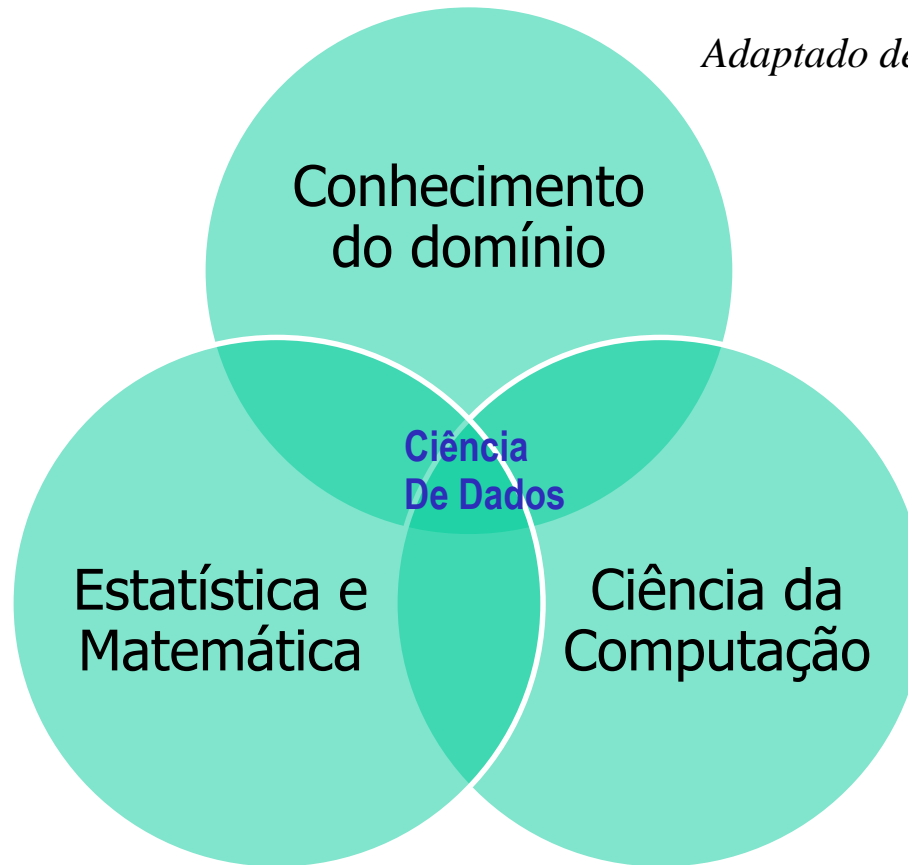
Adaptado de Dahl Winters, 2015





Pilares da Ciência de Dados

Adaptado de Anand Ramanathan, 2016



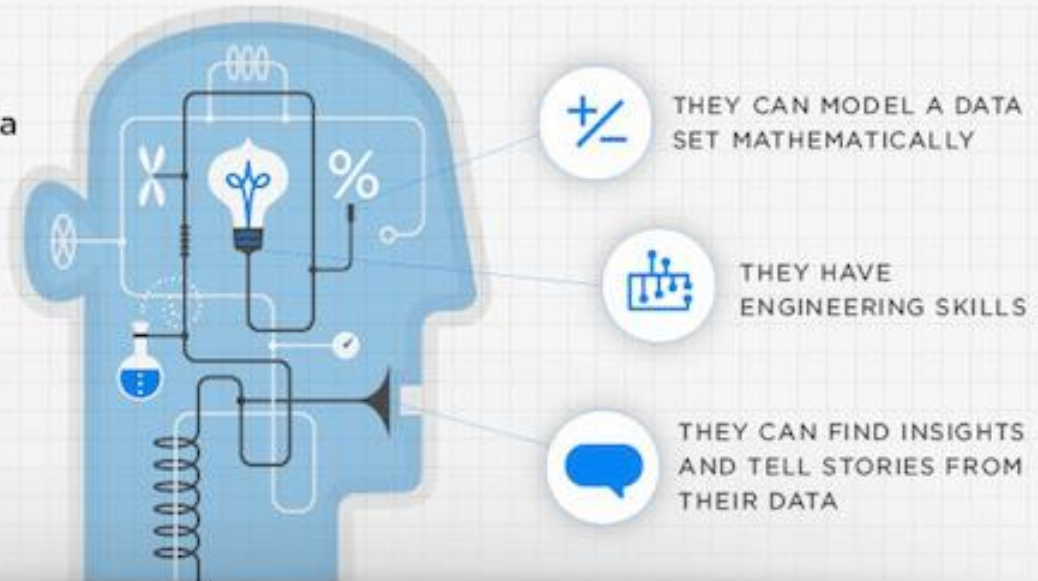


Cientista de Dados

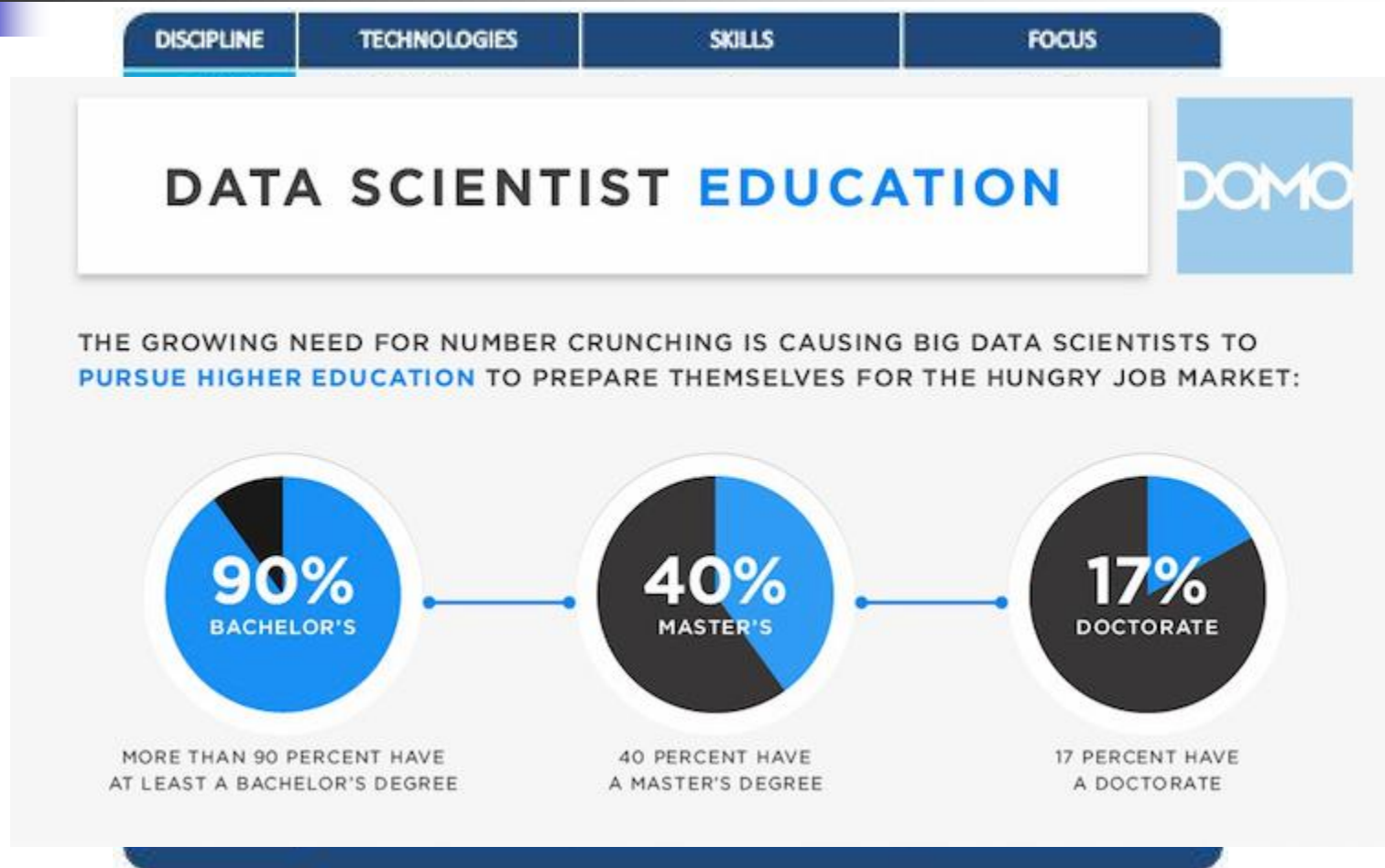
- Precisa conhecer
 - Estatística
 - Álgebra Linear
 - Aprendizado de Máquina
 - Banco de Dados
 - Visualização

Teste vocacional

If you've already been through school, but still want to be a data scientist? Not to worry, [Hilary Mason](#), chief scientist at Bitly says that good data scientists have three essential traits:



Nível de formação





Nível de carreira

- Cientista iniciante
 - Maior parte do tempo preparando e processando dados
- Cientista chefe
 - Menos tempo processando dados
 - Mais tempo em projeto de Analítica e interpretação de modelos e resultados

Oportunidades

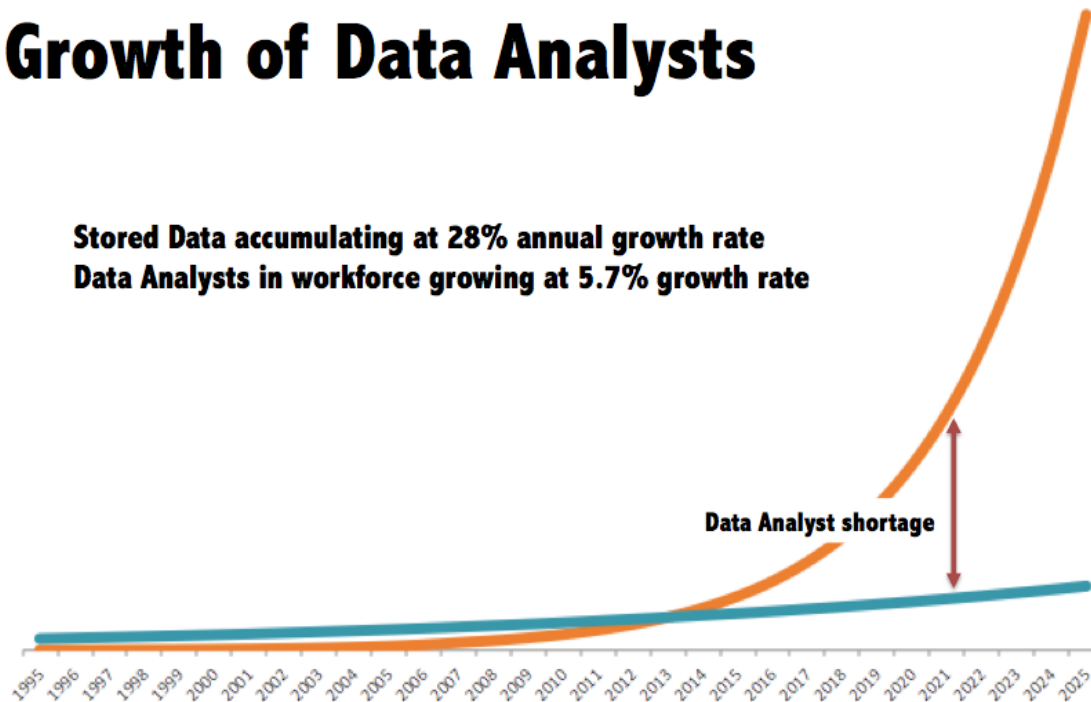
- “Data Scientist: The Sexiest Job of the 21st Century”
 - *Harvard Business Review*, Outubro de 2012
- Ajuda tomadores de decisão a mudar análise subjetiva para análise baseada em dados



Falta de Cientistas de Dados

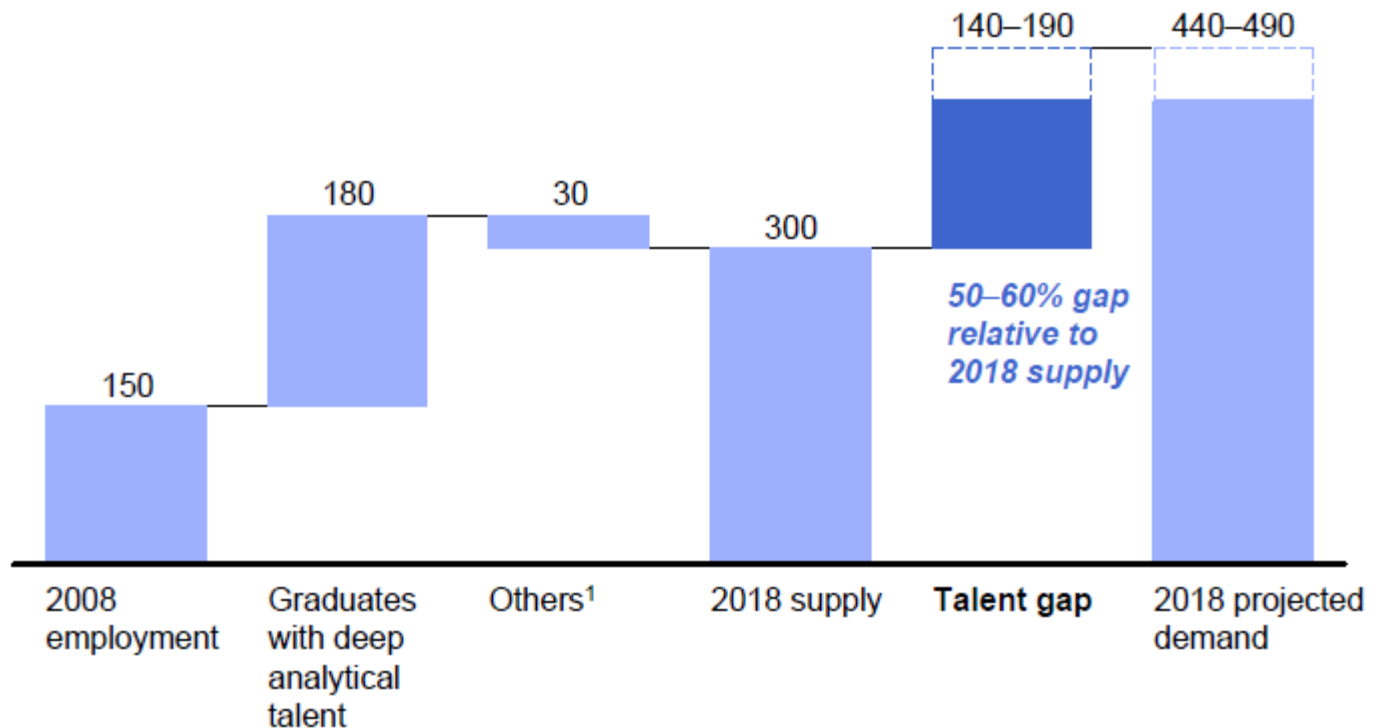
Growth of Data vs. Growth of Data Analysts

Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate



www.delphianalytics.net

Falta de cientistas de dados



Haverá falta de especialistas em data science.

Em 2018, faltarão nos EUA 140.000 a 190.000 analistas com capacidade para análises detalhadas de dados.

Falta de Cientistas de Dados





Quem esta contratando CD

- Apple
- Booking.com
- Disney
- Google
- Greepeace
- Mercedes-Benz
- Red Bull F1
- Vários Bancos (inclusive do Brasil)



Red Bull F1

- ***Team Leader Simulation Development & Data Science at Red Bull F1***
 - *In this role you will lead, develop and implement analysis algorithms and techniques to optimise car design and on track performance.*
 - *This is a multi-disciplinary role where you will have responsibility for:*
 - *Development of new modelling and simulation methods to enhance the understanding of vehicle performance;*
 - *Development of high level car optimisation strategies based on simulation and measured data;*
 - *Offer the engineering department guidelines in the areas of model calibration, analytical decision making, and uncertainty propagation;*
 - *Extend the capabilities of our existing simulation and real-time models by leveraging parallel processing techniques.*



Booking.com

- ***Booking.com BV (the company behind [Booking.com](https://www.booking.com)™, the market leading online hotel reservation service in the world) is looking for the world's best Data Scientists!***
- *Would you like to translate terabytes of data into unforgettable holidays for millions of people around the globe? Booking.com, the world's largest accommodation booking website, is looking for rock star Data Scientists to add to join our highly successful Personalization Team within the Front End department.*
- *This product development team crunches endless amounts of data to provide our customers with the best possible experience. They focus on anything from understanding and predicting market data, to ranking all properties on our website, and providing our customers with the most relevant personalized recommendations.*
- *As a Data Scientist you'll work side by side with Developers, Designers and Product Owners, and take full ownership of your work – from the initial idea-generation phase to the implementation of the final product on our website. Our ideal candidate is result-focused, innovative and has solid quantitative background and a good business understanding.*



- ***Google - Data Science Manager, People Analytics Team (Mountain View, CA)***
- *For immediate consideration, please apply to the following URL:
<https://goo.gl/6J91Yz>*
- *Note: Resumes submitted by email will not be considered.*
- *The area:*
- *Great just isn't good enough for our People Operations team (you probably know us better as "Human Resources"). Made up of equal parts HR professionals, former consultants and analysts, we're the champions of Google's colorful culture. In People Ops, we "find them, grow them, and keep them" - we bring the world's most innovative people to Google and provide the programs that help them thrive. Whether recruiting the next great Googler, refining our core programs, developing talent or simply looking for ways to inject some more fun into the lives of our Googlers, we bring a data-driven approach that is reinventing the human resources field.*



Apple

- *Want to improve Apple's music recommendation and playlisting services, and have a chance to influence the next generation of Apple products?*
- *We'd like to hear from strong scientific engineers who'd be interested in joining us in London.*
- *You'll need to have a good knowledge of machine learning (but hopefully that's why you're reading this list) and experience of working at scale. And to be a serious music lover.*



Greenpeace

- ***Greenpeace - Decision Support Analyst (Washington, DC)***
- *The Decision Support Analyst will help evaluate and improve Greenpeace's constituent engagement efforts. The Analyst will collaborate with various organizational departments and will work on a wide range of problems. This role will bring analytical rigor to challenges such as measuring the impact of an email test, optimizing donor retention, optimizing channels of constituent growth and building models of donor behavior.*
- ***RESPONSIBILITIES***
 - *Act as a subject matter expert in cross functional teams across the organization*
 - *Develop a deep understanding of the goals and strategies of the teams/departments being supported*
 - *Along with IT staff, develop systems for measuring the success of engagement efforts*
 - *Report regularly on latest recruitment and retention efforts*
 - *Design tests to provide insight into complex constituent behavior*
 - *Ensure shared learning from testing results across the organization*
 - *Present recommendations on best practices*
 - *Write complex queries for interacting with data from a variety of sources*

- ***CBS Corporation/ showtime Networks - Data Scientist (New York City)***
- *This Data Scientist role requires experience in predictive and exploratory analytics, querying data systems, performing research and data manipulation supporting Showtime Networks. The Data Scientist will work closely with Digital Services, Research, Marketing, IT and Strategy & Analysis teams to use data to help build research and digital analytic platform. We are looking for someone with strong statistical background who is passionate about analytics, thrives in an evolving environment, and brings an enthusiastic and collaborative attitude.*
- ***Responsibilities***
 - *Work closely with the digital services and research group in understanding the current process and help define technology roadmap for digital data analytics platform.*
 - *Responsible for gathering business requirements, process solutioning both functional and technical, delivery management and managing consultants.*
 - *Build and execute analytics and reporting across platforms to identify user behavior and analyze trends, patterns, and shifts in user behavior, both independently and in collaboration with product managers and data analytics resources.*
 - *Develop best practices for configuring analytics technology, for analyzing user behavior on multiple platforms, and for collecting and interpreting data from multiple sources.*



Centros de Ciência de Dados

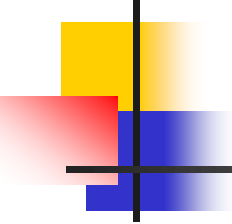
- Columbia University, EUA
- Eindhoven University of Technology, Holanda
- Imperial College, Reino Unido
- Leiden University, Holanda
- New York University, EUA
- Tilburg University, Alemanha
- University of Edinburgh, Reino Unido
- University of Massachusetts at Amherst, EUA



Cursos em Universidades

- Graduações, mestrados e doutorados
- Graduações
 - Eindhoven University of Technology, Holanda
 - Tilburg University, Alemanha
 - University of Nottingham, Reino Unido
 - University of Warwick, Reino Unido
 - University of Essex, Reino Unido

Cientistas de Dados

- 
- Cientistas de dados são os grandes mineradores de dados. Eles recebem uma enorme massa de dados desorganizados (estruturados e não estruturados) e usam suas habilidades em matemática, estatística e programação para **limpar, tratar e organizá-los**. Em seguida, eles aplicam suas capacidades analíticas – conhecimento da indústria, compreensão contextual, ceticismo de suposições existentes – para descobrir soluções para os desafios de negócios ocultos. Entre suas principais responsabilidades estão:
 - Realizar pesquisas sem direção e formular perguntas abertas aos dados
 - Extrair grandes volumes de dados de múltiplas fontes internas e externas
 - Empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e prescritiva



Cientistas de Dados

- Explorar e analisar dados de uma variedade de ângulos para determinar fraquezas escondidas, tendências e / ou oportunidades
- Conceber soluções orientadas a dados para os desafios mais prementes
- Inventar novos algoritmos para resolver problemas e criar novas ferramentas para automatizar o trabalho
- Comunicar previsões e resultados para a gestão e os departamentos de TI através de visualizações de dados eficazes
- Recomendar mudanças econômicas aos procedimentos e estratégias existentes



Ciência de Dados para o Bem

- Movimento sem fins lucrativos
 - Trazer benefícios sociais para as pessoas e comunidades
 - Alguns programas são adotados por empresas
- Como isso ocorre?
 - Reuniões
 - Eventos
 - Estágios acadêmicos
 - Redes sociais



Ciência de Dados para o Bem

- Traz benefícios sociais para pessoas e comunidades
 - Bons serviços de saúde para todos
 - Desenvolvimento econômico de países pobres
 - Educação pública de qualidade
 - Energia limpa e barata
 - Melhor exercício da cidadania
 - Proteção ambiental
 - Meios de transportes mais seguros, rápidos e limpos



Exercícios

- Por em prática o que for visto durante o curso
 - Preparação de dados
 - Implementação
 - Realização de experimentos
 - Análise de resultados
 - Relatório bem escrito



Projeto

- Utilizar técnicas vistas em sala de aula para resolver problema real
 - Dados públicos ou privados
 - Relatório, descrevendo todas as características observadas e extraída dos dados, técnicas utilizadas para análise e interpretação dos dados



Avaliação (PCCMC)

- $NF = (0.5 * M_{pv} + 0.5 * MT)$

onde:

- $M_{pv} = 0.4 * P1 + 0.6 * P2$ (2 provas: P1, P2)
- $MT = m.\text{aritm. de projetos}$
- Se algumas das notas < 5
 - $MF = \text{menor valor entre as notas}$

- **Não haverá prova substitutiva**



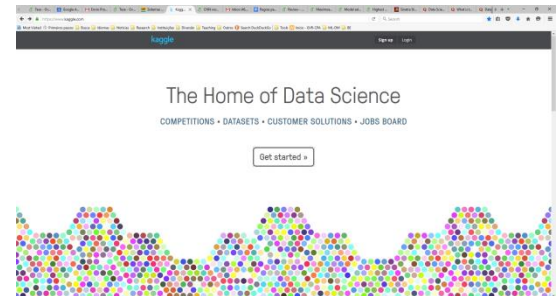
Práticas

- Grupos de até 2 pessoas
 - Usar Python
 - Aula pratica semanalmente



Práticas (PCCMC)

- Aplicar conceitos vistos em conjunto de dados do Kaggle
 - <https://www.kaggle.com/>
 - Problema de classificação
 - Práticas em todas as aulas
 - Relatórios semanais
 - Até duas páginas
 - Falar o que foi feito



Kaggle

The screenshot shows the Kaggle website homepage. At the top, there's a navigation bar with links to Host, Competitions, Datasets, Scripts, Jobs, and Community. A user profile for 'Andre' is visible in the top right corner. Below the navigation bar, a light blue banner reads: 'We are making our URLs prettier -- Claim your personal URL now!'. The main content area features a welcome message: 'Hi Andre! We'd like to welcome you to Kaggle. Since you're new, here's just a few ways to get started:'. This message is accompanied by three icons: a starburst with a 'k' for 'Explore the competitions', a folder with 'kaggle?' for 'Learn from great code', and a speech bubble with 'Hi' for 'Visit the jobs board'. Below these, there are three sections: 'Explore the competitions' (Download data from one of the active competitions listed below), 'Learn from great code' (Check out best practice code from top Kagglers on our [scripts page](#)), and 'Visit the jobs board' (See who's hiring on our [jobs board](#)). To the right of the welcome message, there's a profile card for 'Andre' with a bird icon and links to 'View / Edit Profile'. Below the profile card, there's a section titled 'Recent Jobs' listing several job openings. At the bottom, there's a section titled 'Active Competitions' listing three competitions: 'Second Annual Data Science Bowl', 'Santander Customer Satisfaction', and 'Home Depot Product Search Relevance'.

Hi Andre! We'd like to welcome you to Kaggle.

Since you're new, here's just a few ways to get started:

Explore the competitions
Download data from one of the active competitions listed below.

Learn from great code
Check out best practice code from top Kagglers on our [scripts page](#).

Visit the jobs board
See who's hiring on our [jobs board](#).

Recent Jobs

- CEP America - Sr. Systems Engineer (Emeryville, California, United States, 94608)
- CEP America - Sr. Network Engineer (Emeryville, California, United States, 94608)
- Booking.com - Data Scientist - Machine Learning (Amsterdam)
- Booking.com - Data Scientist (Amsterdam)
- trivago - Junior Data Analyst - Quality Assurance - Business Intelligence (Düsseldorf)
- Even Responsible Finance - Data Scientist (Oakland, CA)

Active Competitions

Competition	Days	Teams	Scripts
Second Annual Data Science Bowl Transforming How We Diagnose Heart Disease	5.2 days	33 teams	\$200,000
Santander Customer Satisfaction Which customers are happy customers?	54 days	874 teams	323 scripts \$60,000
Home Depot Product Search Relevance Predict the relevance of search results on homedepot.com	47 days	1415 teams	1034 scripts

On the Forums

- lectures in Machine Learning
- Petals Around the Rose challenge for ML
- Teaming up for kaggle projects
- What is the highest accuracy



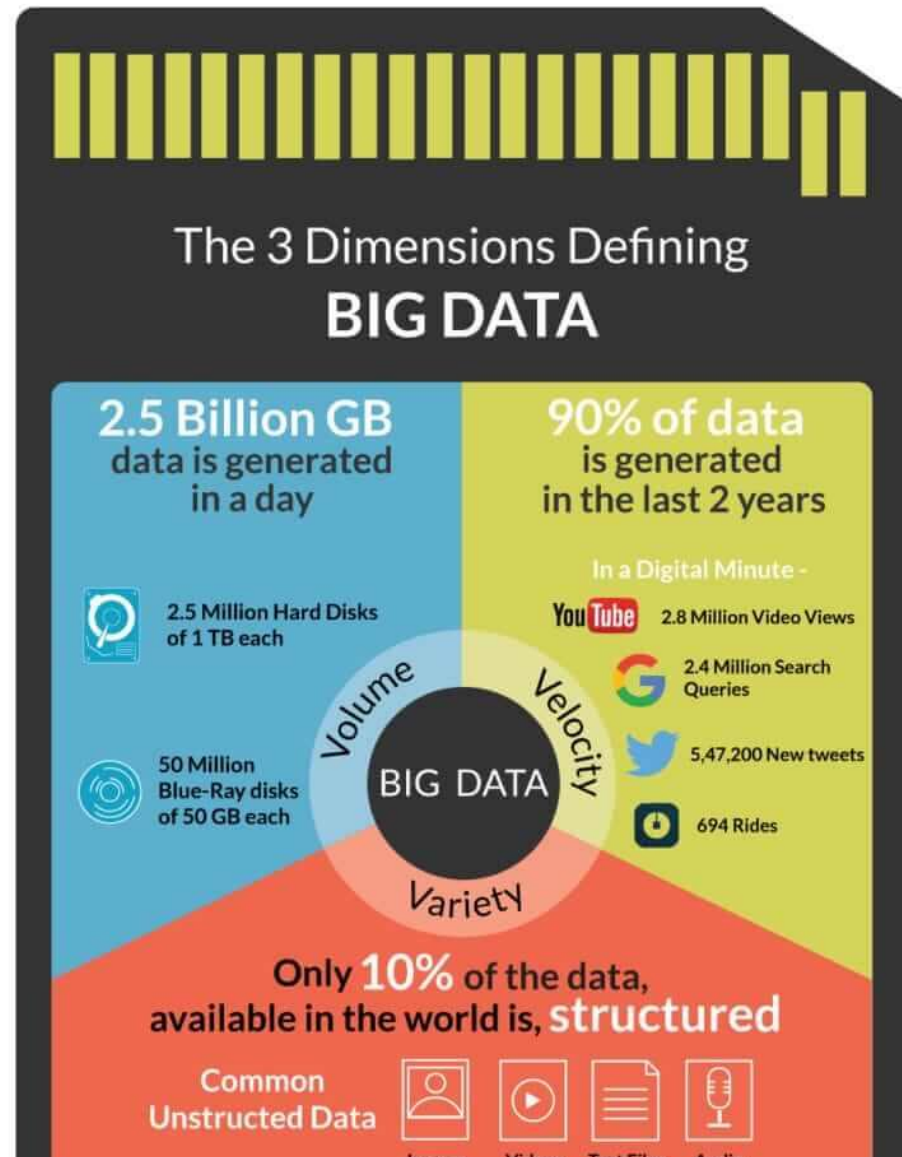
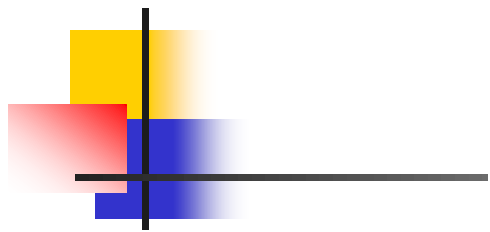
Bibliografia

- Faceli, K., Lorena, A., Gama, J. e Carvalho, A., Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina, LTC, 2011
- Provost, F.; Fawcett, T. Data Science for Business: What you need to know about data mining and data-analytic thinking by O'Reilly Media, 2013
- Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann , 2011
- Witten, I.; Frank, E. Third Edition (The Morgan Kaufmann Series in Data Management Systems). 2011
- Tan, P.-N.; Steinbach, M.; Kumar, T. Introduction to Data Mining. Addison Wesley, 2005

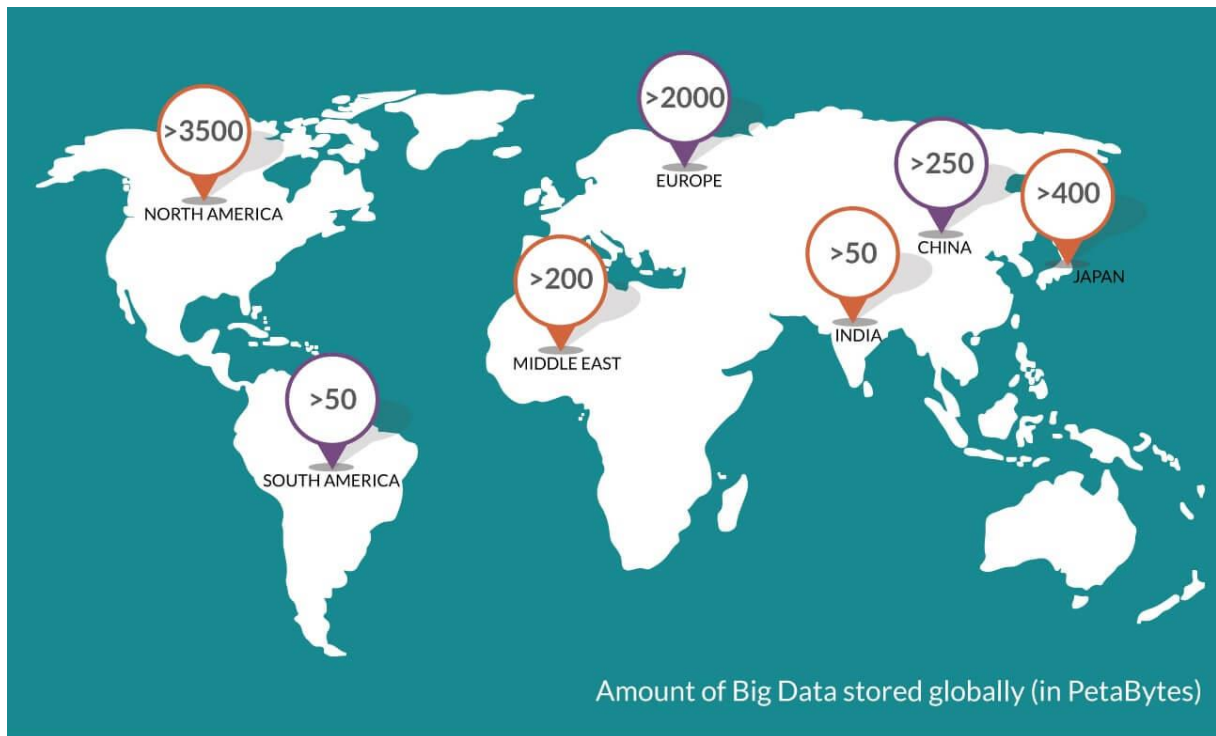


Perguntas





Volume

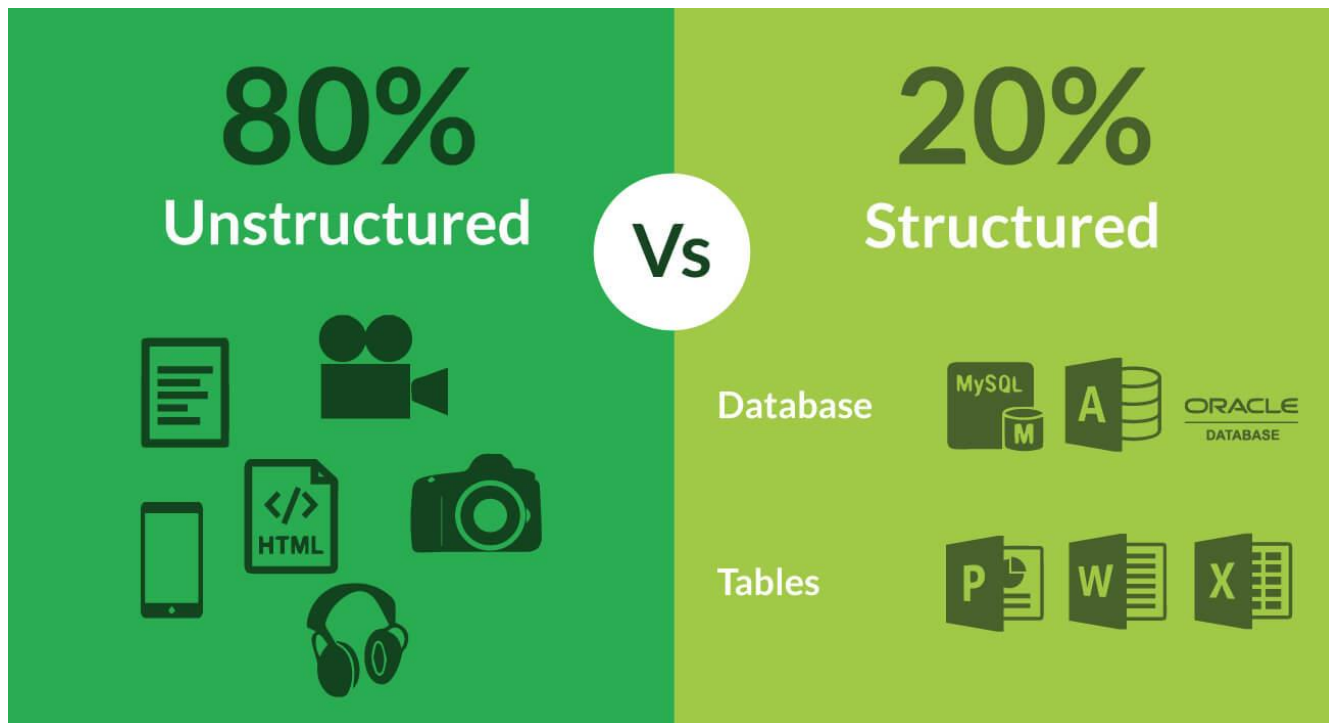


Velocity

What happens in an Internet Minute?



Variety

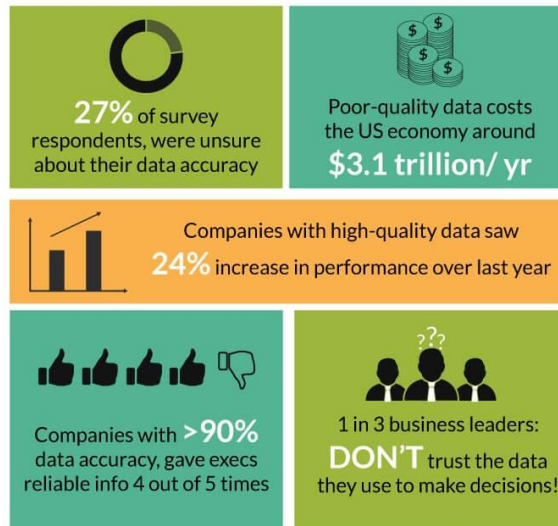


Veracity

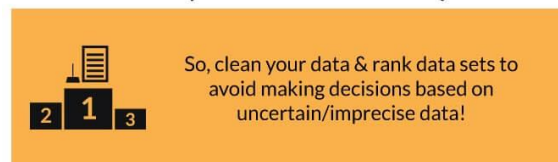
Veracity includes



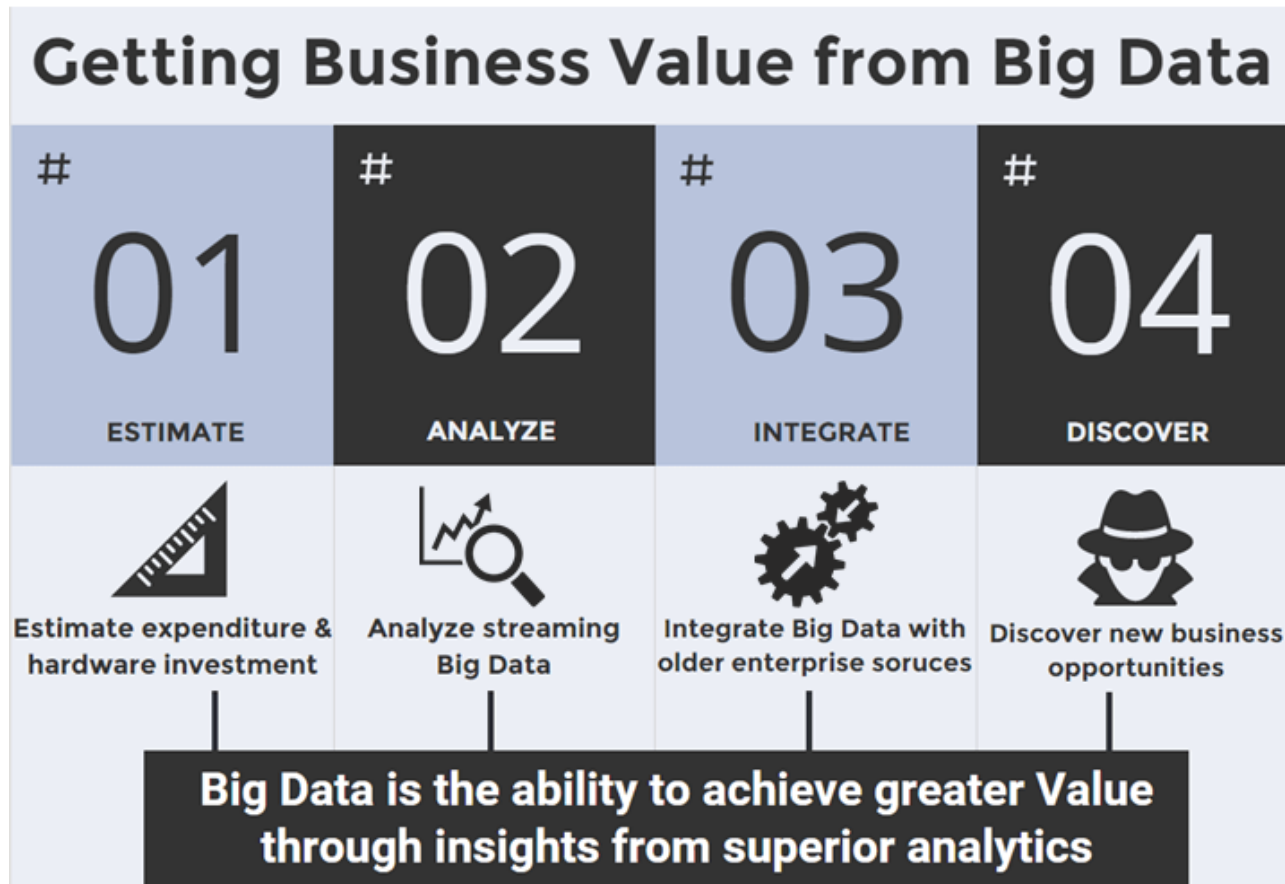
Some numbers on Veracity



A tip to handle Veracity



Value



<https://upxacademy.com/beginners-guide-to-big-data/>



DATA ENGINEER

I gather and store data. I do batch processing or real-time processing on data. I also serve data via an API to a data scientist who can easily query it.



DATA ANALYST

I collect, organize and interpret data to help companies make better business decisions.



DATA SCIENTIST

I have programming skills, knowledge of statistics and domain knowledge. I ask the right questions and try to find out insights from a given data set.