



INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Rio de Janeiro



CIÊNCIA DE DADOS: DESAFIOS E OPORTUNIDADES

Eduardo Ogasawara
eogasawara@ieee.org
<https://eic.cefet-rj.br/~eogasawara>

Ciência de Dados

- Ciência de dados é a Ciência que estuda a análise de dados
 - A extração de conhecimento a partir de dados (Big Data)
- Envolve
 - Análise de dados
 - Aprendizagem estatística e de máquina
 - Gerência de grandes coleções de dados
 - Mineração de dados
 - Computação paralela e distribuída
 - Visualização de dados
- Característica multidisciplinar
- Aplicações na ciência/indústria/governo
 - Bioinformática, petróleo, energia, finanças, astronomia, Internet, mobilidade urbana, defesa cibernética, educação, etc



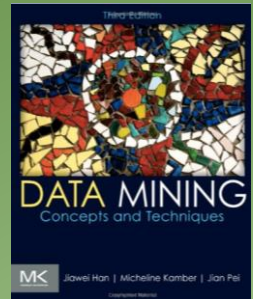
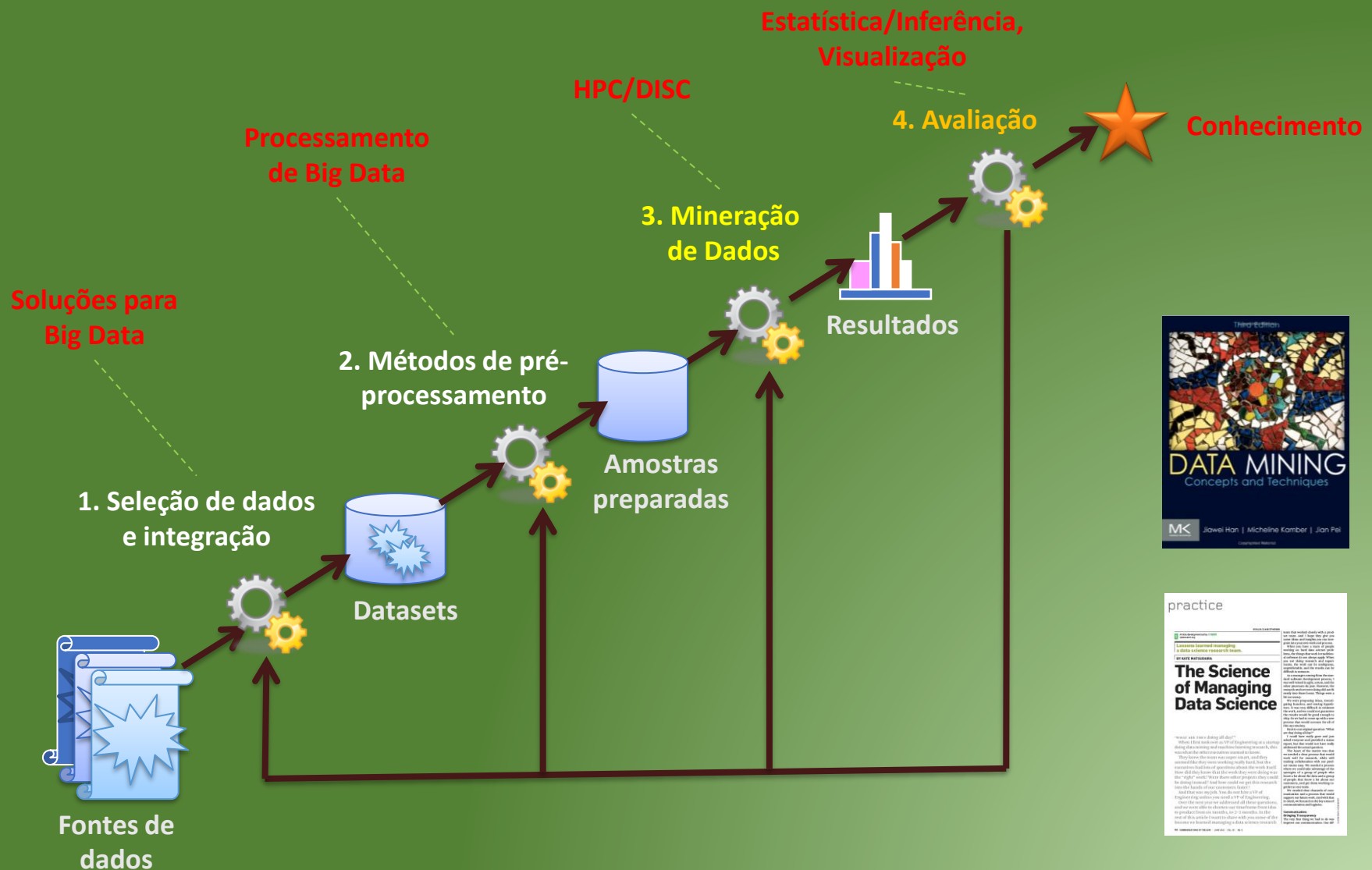
[1] V. Dhar, 2013, Data science and prediction, *Communications of the ACM*, v. 56, n. 12, p. 64–73.
[2] L. Cao, 2017, Data science: A comprehensive overview, *ACM Computing Surveys*, v. 50, n. 3
[3] L. Cao, 2017, Data science: Challenges and directions, *Communications of the ACM*, v. 60, n. 8, p. 59–68.

Por que Ciência de Dados?

- Cenário de Big Data:
 - O crescimento explosivo dos dados: de terabytes a petabytes
 - Coleta de dados e disponibilidade de dados
- Principais fontes de dados abundantes e diversos
 - Negócios: comércio eletrônico, transações
 - Ciência: sensores, astronomia, bioinformática, simulação
 - Sociedade e todos: dados abertos, IoT, sensoriamento populacional
- Estamos nos afogando em dados, mas famintos por conhecimento!
 - Mineração de Dados: Análise automatizada de grandes conjuntos de dados
 - Ciência de Dados: traz a Ciência para a Análise dos Dados

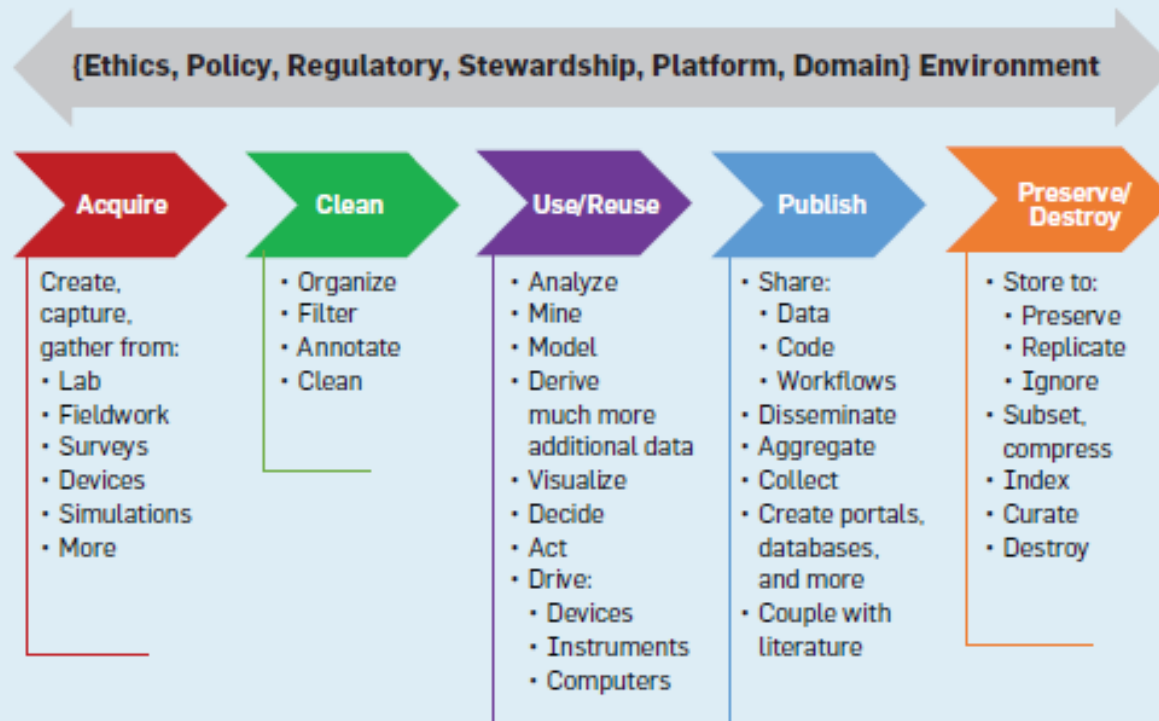


Processo de Descoberta de Conhecimento

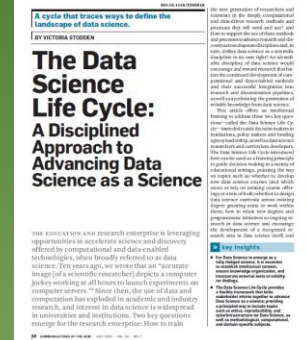


Ciclo de Vida da Ciência de Dados

Figure 1. Example of a data life cycle and surrounding data ecosystem (reprinted with permission).¹



contributed articles



Contexto de Ciência de Dados



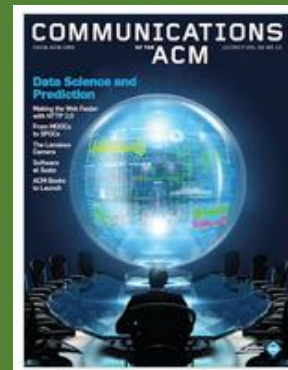
ago/2002



dez/2008



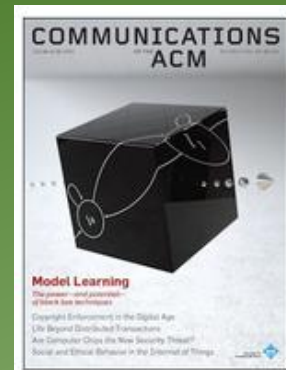
fev/2012



dez/2013



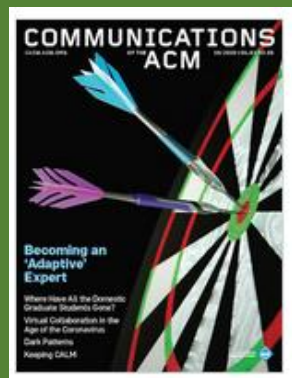
jul/2015



fev/2017



jul/2018



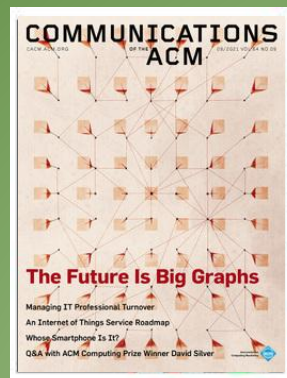
set/2020



mar/2021



jul/2021

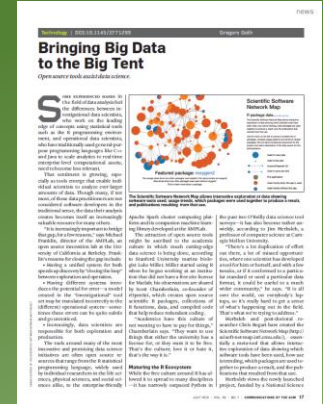
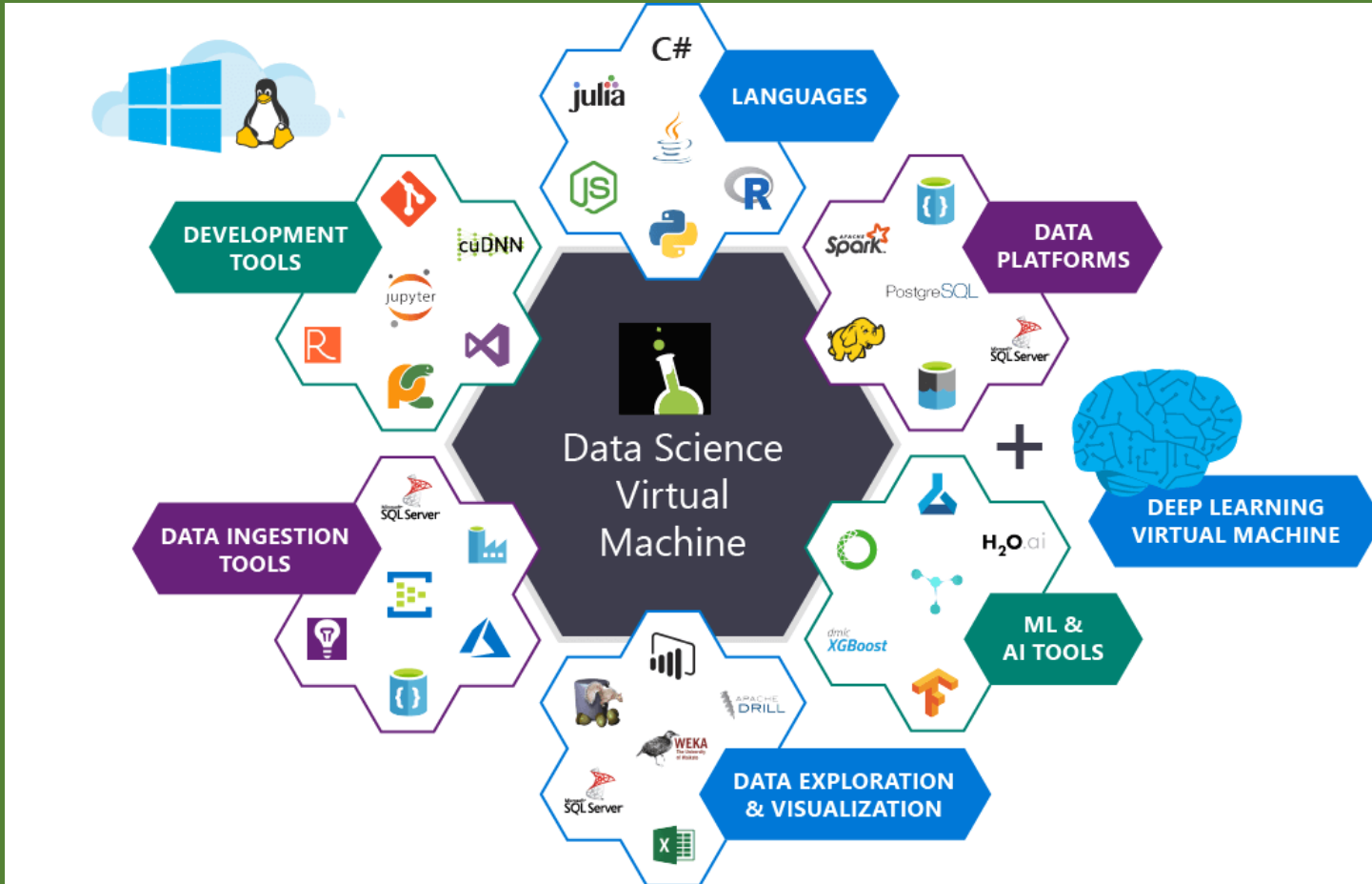


sep/2021



mar/2022

Ecosystema de Ciência de Dados



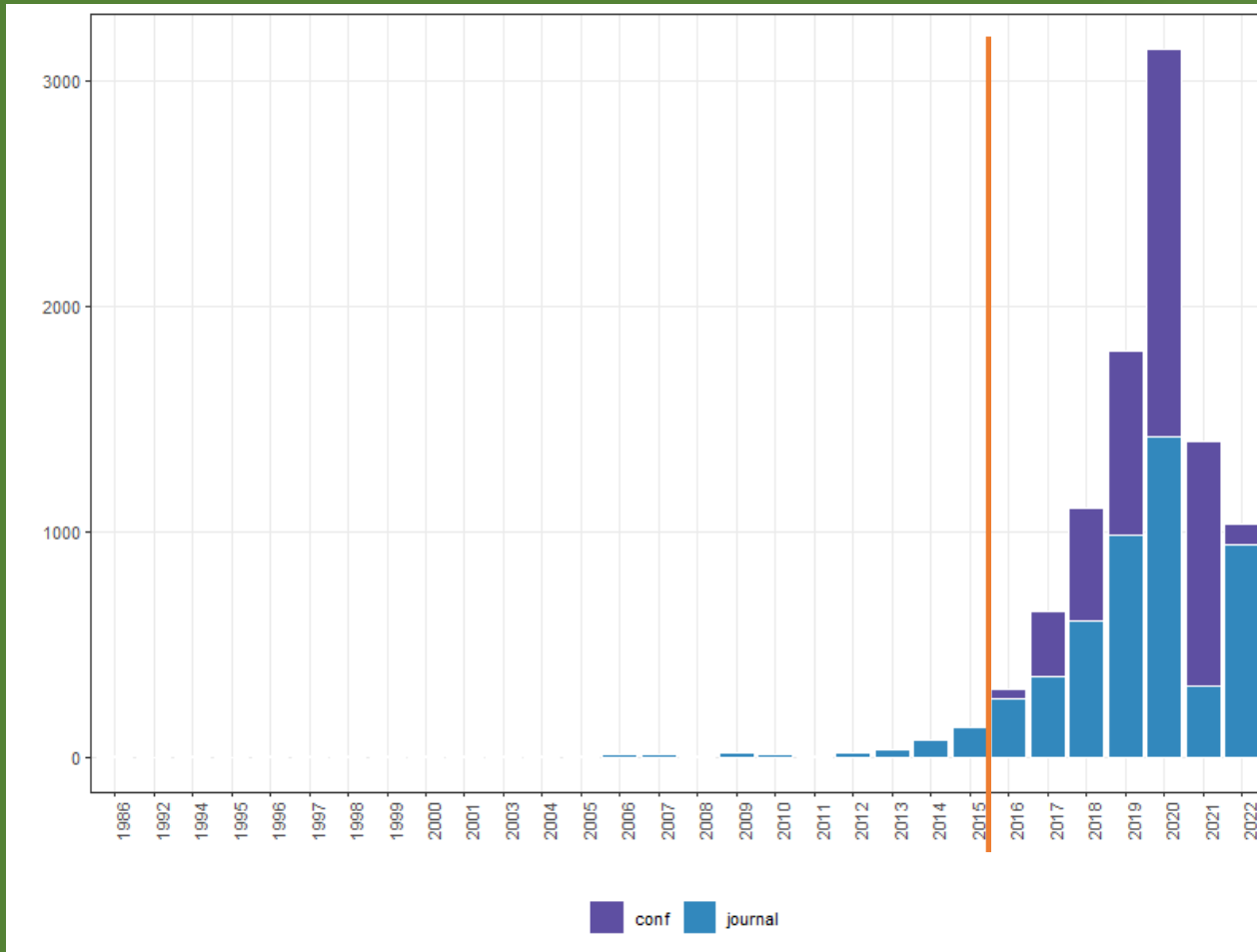
[1] G. Goth, 2015, Bringing big data to the big tent, *Communications of the ACM*, v. 58, n. 7, p. 17–19.

[2] <https://www.datasciencecentral.com/>

Perspectiva das Ciências

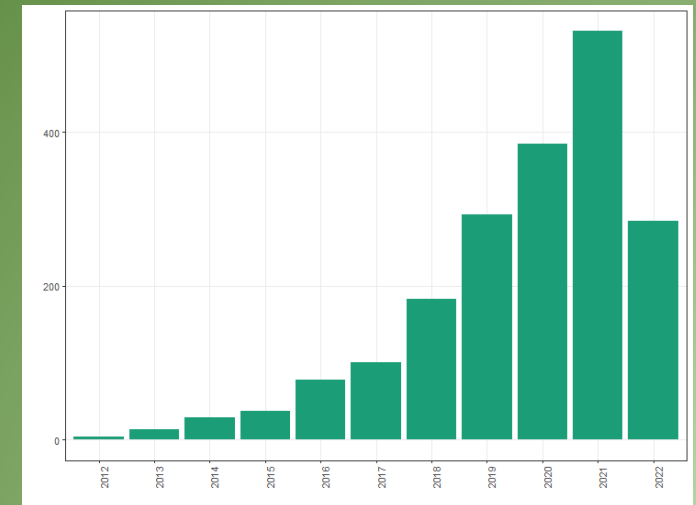
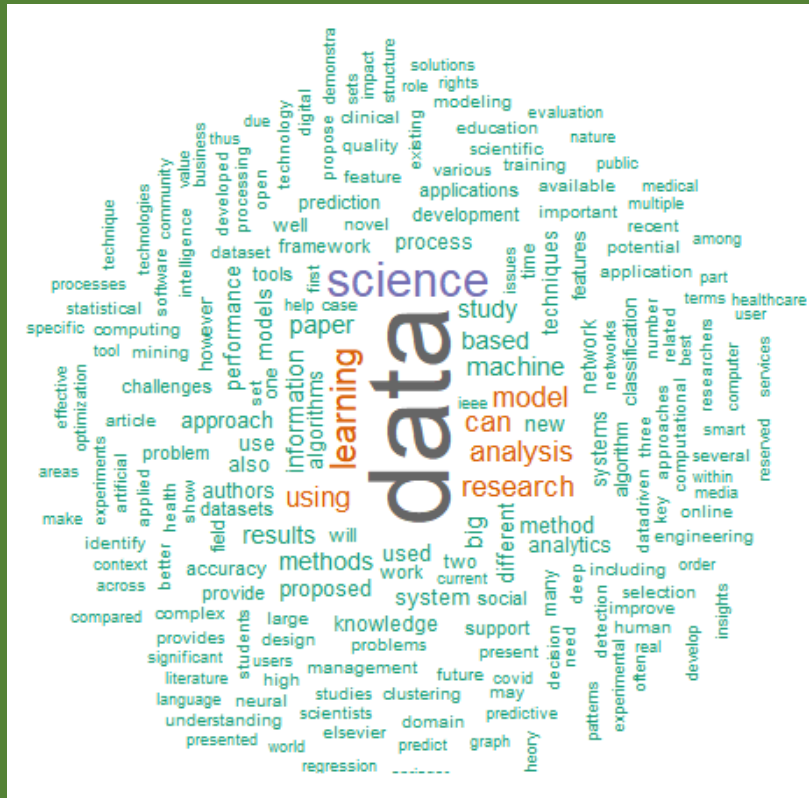
Como está área?

- Pesquisa básica e aplicada
- 9788 artigos publicados entre 1986 e 2022



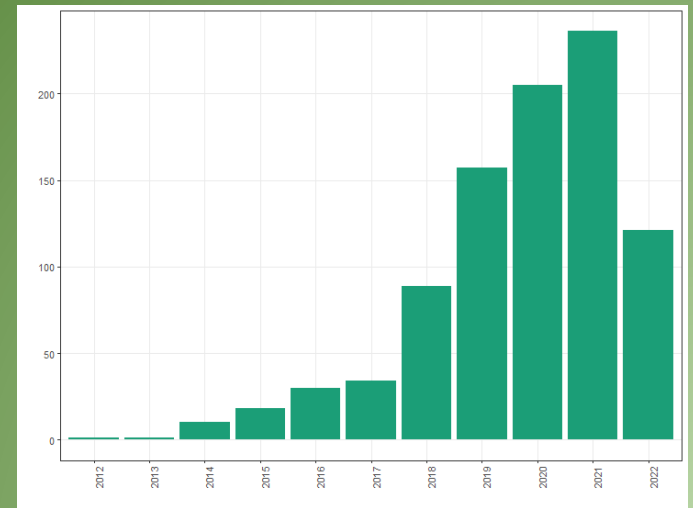
Ciência de Dados na Computação

- Termos de Mineração de Dados, modelo e análise, Aprendizado de Máquina
- Área fim saúde (maior intensidade)



Ciência de Dados na Saúde

- Termos médicos e foco no valor para área
- Análise de dados e dataset
 - Ex.: medicina precisa, centrada em Ciência de Dados



- Educação, área de humanas e sociedade
- Áreas como logística buscando melhorias, área de humanas avaliando a ética



- Termos Estatísticos e Matemáticos
 - Modelos, Análise e Ciência
- Ciência para a Análise dos Dados (John Tukey – 50 anos)



Quais são os desafios?

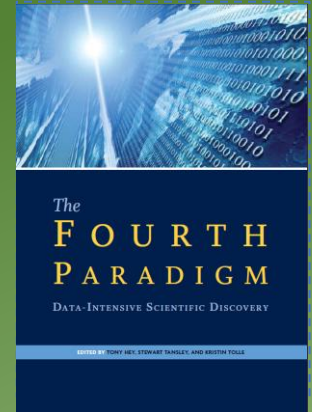
Automatização

- A automatização em ciência de dados visa facilitar e transformar o trabalho dos cientistas de dados
- Mecanização
 - Ocorre quando uma tarefa é tão bem especificada que não há necessidade de envolvimento humano
 - AutoML / AutoWeka
- Composição
 - lida com o sequenciamento estratégico de tarefas ou integração de diferentes partes de uma tarefa
 - Workflows
- Assistência
 - deriva da produção de elementos como visualizações, padrões, explicações, entre outros, que são especificamente direcionados para apoiar a eficiência humana
- Na engenharia de dados há várias oportunidades
 - Organização, Qualidade e Transformação de Dados



Modelos Data-Driven versus Theory-Driven

- Ciência da Computação: terceiro pilar da ciência
 - Teoria
 - Experimentação
 - Simulação (in-silico)
- Descoberta Científica baseadas em Dados
- Modelos orientados a dados
- *Data-driven e theory-driven*
 - Concorrência
 - Realimentação
 - Associação
 - Construção de modelos centrado em dados
 - Que respeitem “leis científicas”
 - *Physical informed Neural Networks*

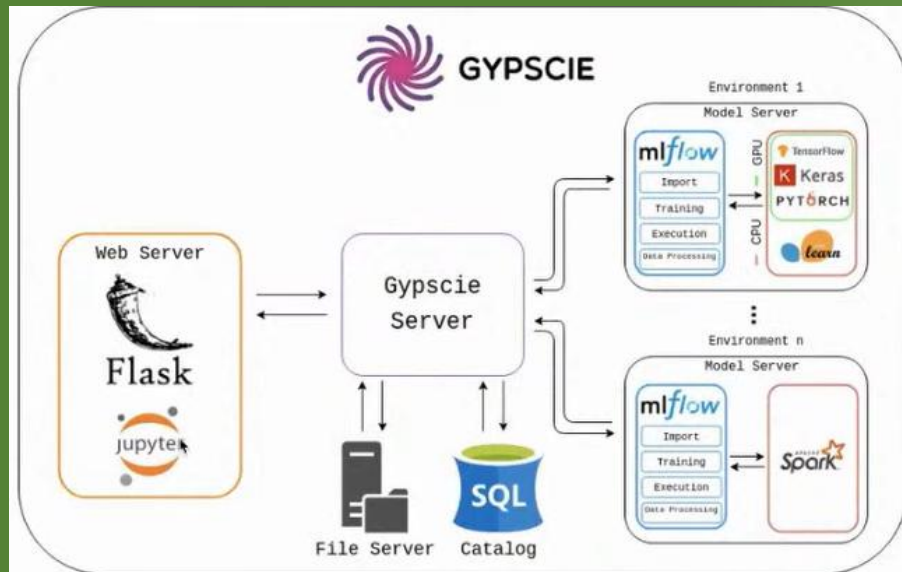
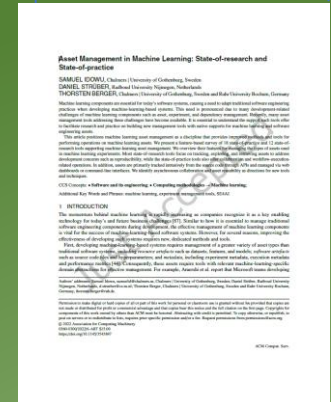


[1] T. Hey, 2009, *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft Research.

[2] J.G. Hoffer, A.B. Ofner, F.M. Rohrhofer, M. Lovric, R. Kern, S. Lindstaedt, and B.C. Geiger, 2022, Theory-inspired machine learning - towards a synergy between knowledge and data, *Welding in the World*, v. 66, n. 7, p. 1291-1304.

Gerência de Modelos

- Gerência de Modelos
 - Construção
 - Produção
 - Reutilização
 - Transferência de conhecimento
 - Seleção dinâmica
 - Distribuição
 - Espaço-tempo



Ciclo de Seminários Online Instituto de Computação

Gypscie: Um Ambiente para Gerência de Dados e Modelos de Aprendizado de Máquina

Fabio Porto
LNCC

15 de Outubro
17:00

TRANSMISSÃO AO VIVO

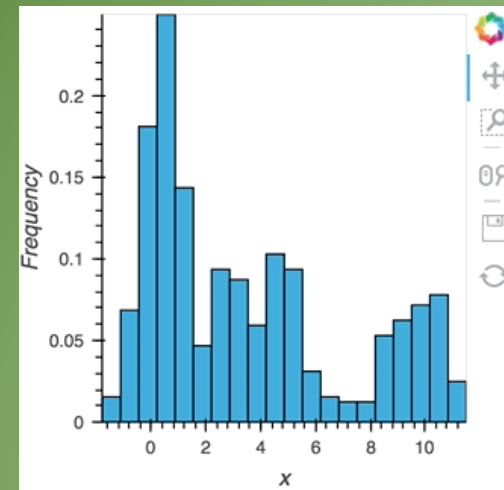
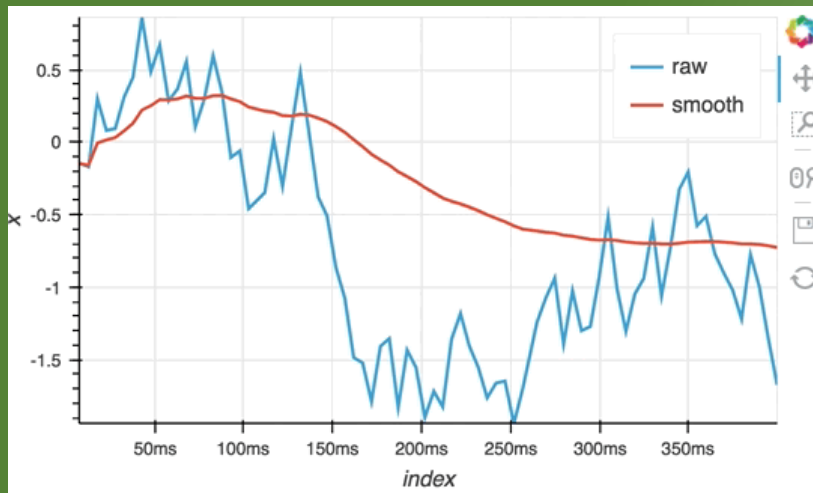
DEXL
DATA EXTREME LAB
Laboratório Nacional de Computação Científica

YouTube http://tiny.cc/ucff_youtube **f** <https://fb.me/informeic>

#InstitutoDeComputação **#UFFContraoCoronaVirus** **#FiqueEmCasa**

As múltiplas faces da não-estacionariedade

- Mudança de conceito (*Concept Drift*)
 - visão de Mineração de Dados
- Dilema da Plasticidade e Estabilidade
 - visão de Aprendizado de Máquina
- Padrões emergentes (*Emerging patterns*)
 - visão de Banco de Dados
- Crítica de Lucas
 - visão de Econometria



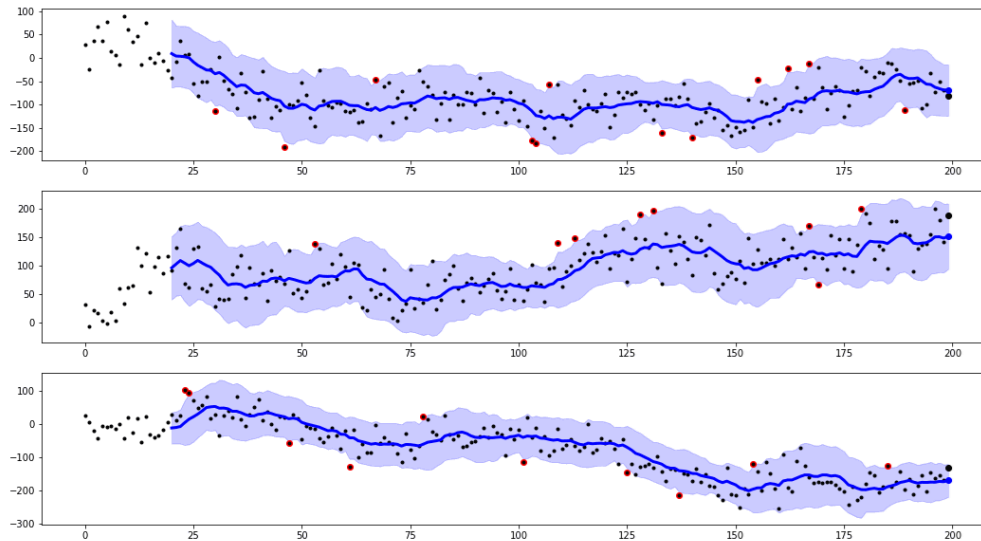
[1] A.S. Iwashita and J.P. Papa, 2019, An Overview on Concept Drift Learning, *IEEE Access*, v. 7, p. 1532–1547.

[2] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, 2015, Learning in Nonstationary Environments: A Survey, *IEEE Computational Intelligence Magazine*, v. 10, n. 4, p. 12–25.

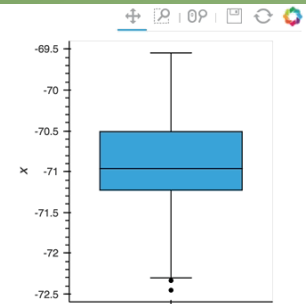
[3] D. Gujarati, 2002, *Basic Econometrics*. 4 ed. Boston; Montreal, McGraw-Hill/Irwin. / [4] S.O. Haykin, 2008, *Neural Networks and Learning Machines*. 3 ed. New York, Prentice Hall.

Detecção de Eventos

- Processamento de stream
 - Velocidade
 - Volume de dados
- Detecção de eventos
 - Offline versus Online

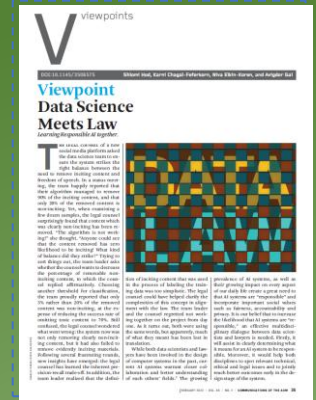


#	index	x
0	2017-10-27 19:43:04	-70.87262710547351
1	2017-10-27 19:43:04	-71.00788295730518
2	2017-10-27 19:43:04	-71.01297392873352
3	2017-10-27 19:43:04	-71.4148637783796
4	2017-10-27 19:43:04	-71.60890969520968
5	2017-10-27 19:43:04	-71.32610485802545
6	2017-10-27 19:43:04	-71.1935680343768
7	2017-10-27 19:43:04	-70.76579342173753
8	2017-10-27 19:43:04	-70.59743524950701
9	2017-10-27 19:43:04	-70.99300214112863



Proteção de Dados, Ética, Legislação

- Desenvolver sistemas que tenham bom balanço entre remoção de conteúdos de incitação e liberdade de expressão
- Tanto advogados quando cientistas de dados precisam estabelecer um dialogo
 - Conceitos como precisão e revocação deixam os advogados confusos
 - Conceitos como justiça, privacidade, responsabilidade, transparência e explicabilidade podem ser vagos para Cientistas de Dados
- Formação
 - Conceitos básicos de leis, ética e sociedade
 - Classes integradas na graduação são um potencial caminho
 - <https://teach.responsibly.ai>



[1] S. Hod, K. Chagal-Ferferkorn, N. Elkin-Koren, and A. Gal, 2022, Data science meets law, Communications of the ACM, v. 65, n. 2, p. 35–39.

[2] L. Floridi and M. Taddeo, 2016, What is data ethics?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, v. 374, n. 2083

Empresas

Características e áreas

- Características das empresas

- Contexto de Big Data
- Business Intelligence & Data Analytics
- Necessidade de Deployment

- Áreas

- Financeira (análise de risco)
- Saúde (melhoria de diagnósticos)
- Turismo (perfis, segmentação)
- Energia (reduzir risco, otimizar investimentos)
- Manufatura (Industria 4.0, logística, previsão)
- Jogos (perfil do consumidor)
- Farmacêutica (planejamento clínico)

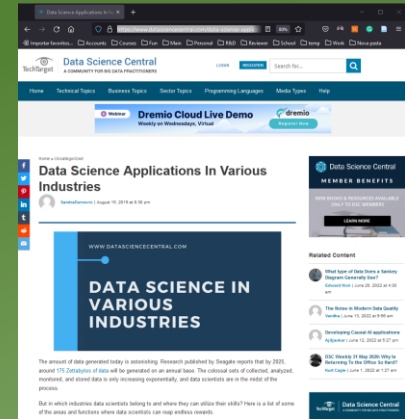


Figure 8: What Is the Main Driver for Your Data and/or Analytics Program?



Source: State of Data in Travel Survey, 2017

Ciência de Dados nas empresas e indústrias

- Construir modelos úteis
- Desafios das soluções caixa preta
 - algoritmos, parâmetros de ajuste, efeitos sutis e montagens não são acessíveis
 - podem ser adequadas para soluções de curto prazo, mas a robustez à mudança não é certa
 - habilidades para compreender os métodos analíticos não são valorizadas e desenvolvidas



2012

In 2012, Harvard Business Review named data scientist the "[sexiest job of the 21st century.](#)" More recently, Glassdoor named it the "[best job of the year](#)" for 2016.

"It isn't a big surprise," Dr. Andrew Chamberlain, [Glassdoor's](#) chief economist, told Business Insider. "It's one of the hottest and fastest growing jobs we're seeing right now."

According to Glassdoor, data scientists earn a base pay of **\$116,840 a year**, on average.



VFS Digital Design/Flickr

Facebook: \$133,841

Apple: \$149,963

Airbnb: \$117,229

Twitter: \$134,861

Microsoft: \$119,129

LinkedIn: \$138,798

O que se espera do Cientista de Dados?

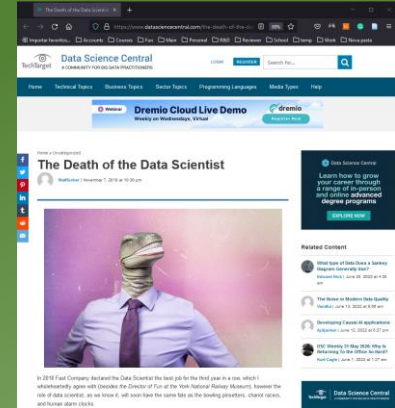
- Teoria e Prática
 - Fundamentos de matemática e estatística
 - Banco de dados
 - Programação paralela e distribuída
 - Mineração de dados
 - Visualização
 - Conhecimento do domínio
 - Inovação
- Desenvolvimento de software
 - Diversas linguagens de programação
 - Diversas ferramentas



Ciência de Dados: A morte do Cientista de Dados

2018

- Existem um conjunto de atividades de Ciência de Dados que podem ser automatizadas
- Ferramentas vêm sendo desenvolvidas com boa usabilidade
 - ETL
 - Construção de Modelos
- Não é “apenas” produzir código básico em Python e R



Fundamentos...



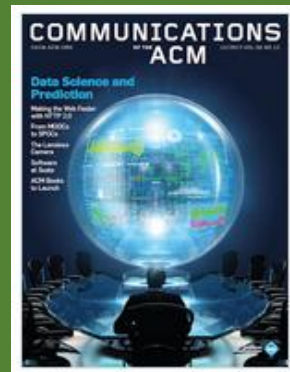
ago/2002



dez/2008



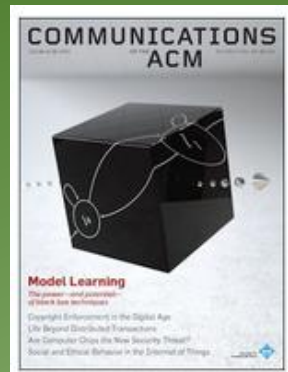
fev/2012



dez/2013



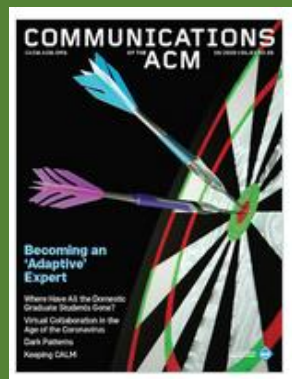
jul/2015



fev/2017



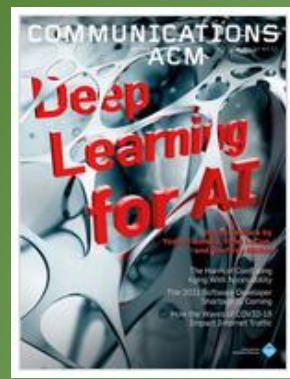
jul/2018



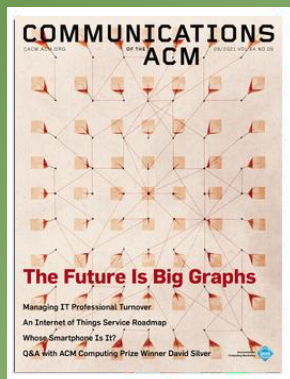
set/2020



mar/2021



jul/2021



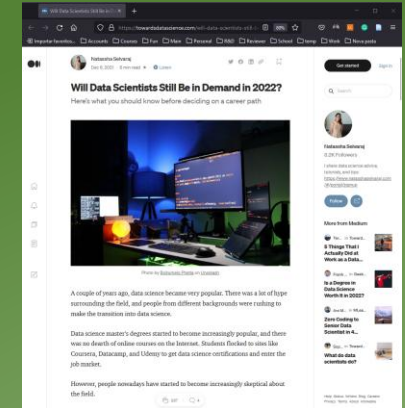
sep/2021



mar/2022

Ciência de Dados: Cenário atual

- Nos EUA e em outros países as oportunidades são boas
 - Oportunidades vêm sem mantendo
 - Perspectivas é de alta até 2030
- No Brasil existe uma defasagem
 - Em 2014 havia poucas ofertas
 - Hoje o número vagas é razoável



Changes in Base Salaries by Job Level for **Data Science Individual Contributors**

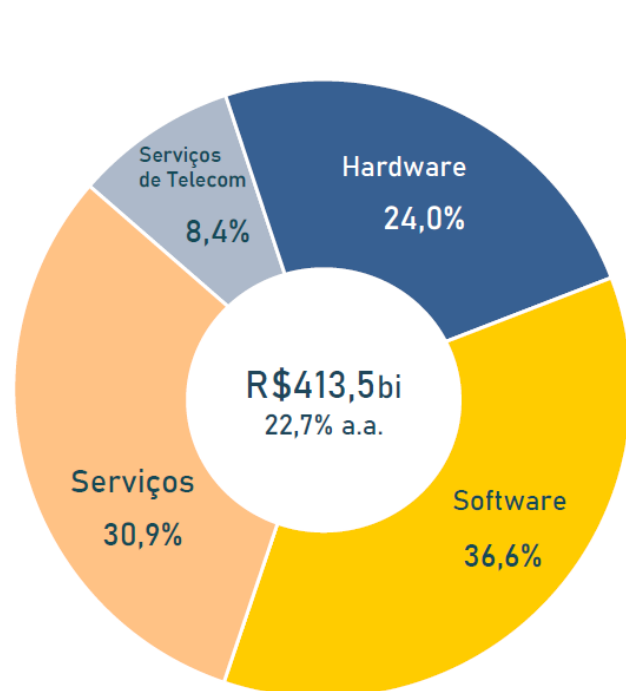
Job Level	Year	25%	Median	Mean	75%	N
Individual Contributor Level 1	2021	\$87,750	\$95,000	\$100,618	\$115,000	34
	2020	\$85,000	\$95,500	\$97,394	\$110,000	124
	Change	+3%	-0.5%	+3%	+5%	
Individual Contributor Level 2	2021	\$110,000	\$130,000	\$130,853	\$150,000	95
	2020	\$120,000	\$130,000	\$130,750	\$140,000	132
	Change	-8%	0%	0%	+7%	
Individual Contributor Level 3	2021	\$149,250	\$160,000	\$162,995	\$175,000	62
	2020	\$148,000	\$165,000	\$172,309	\$185,000	55
	Change	+1%	-3%	-5%	-5%	

Perspectivas de Investimentos em TI no Brasil

Perspectivas de Investimentos de 2021-2024 (R\$ bilhões)



Tecnologias de Transformação Digital



Nuvem
R\$ 181,9 bi | 28% a.a.



Robótica
R\$ 31,4 bi | 6% a.a.



Internet das Coisas
R\$ 74,3 bi | 29% a.a.



Realidade Virtual
R\$ 2,2 bi | 40% a.a.



Big Data & Analytics
R\$ 77,2 bi | 13% a.a.



Redes Sociais
R\$ 14,1 bi | 19% a.a.



Segurança da Informação
R\$ 16,2 bi | 13% a.a.



Impressão 3D
R\$ 0,7 bi | 15% a.a.



Inteligência Artificial
R\$ 11,5 bi | 31% a.a.



Blockchain
R\$ 3,8 bi | 65% a.a.

Mobilidade e Conectividade

Mobile, Dados e Banda Larga



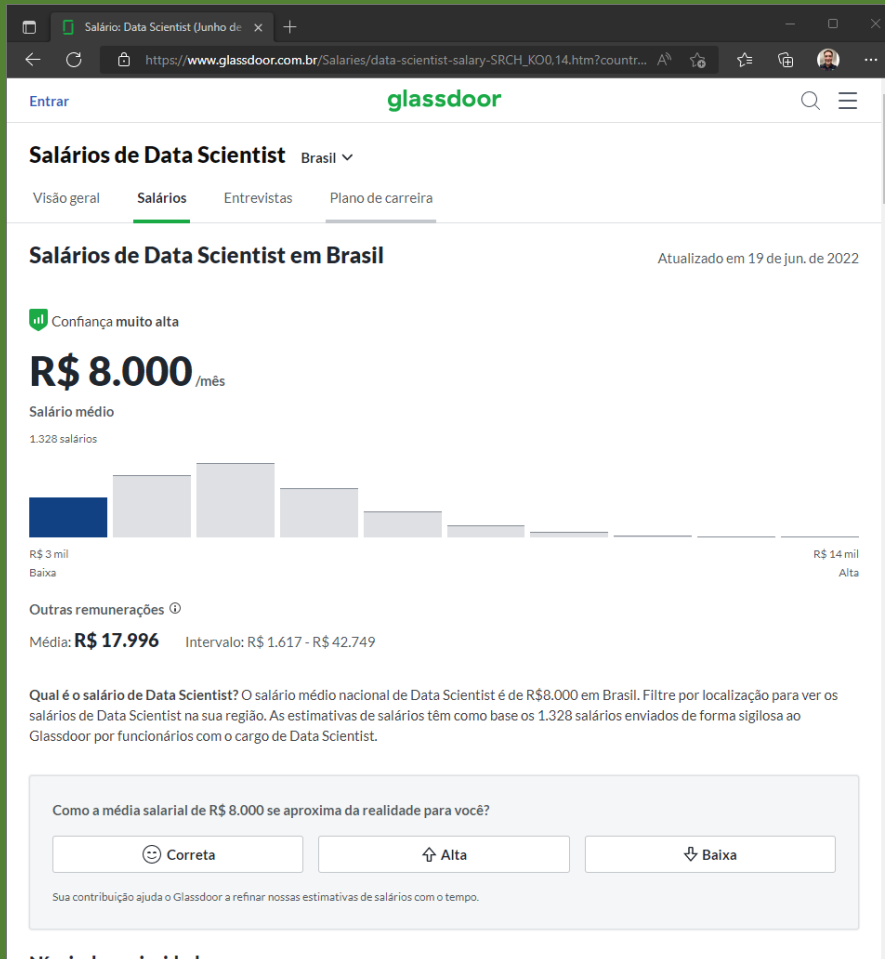
R\$ 431,5 bi

4,6% a.a.



Taxa de câmbio: R\$/US\$ 3,95 (2019)

Carreira no Brasil



(7) Vagas de ciência de dados

ciência de dados / Bra...

Vagas Data do anúncio Nível de experiência Todos os filtros

Ciência de dados em: Brasil

1.462 resultados Configurar alerta

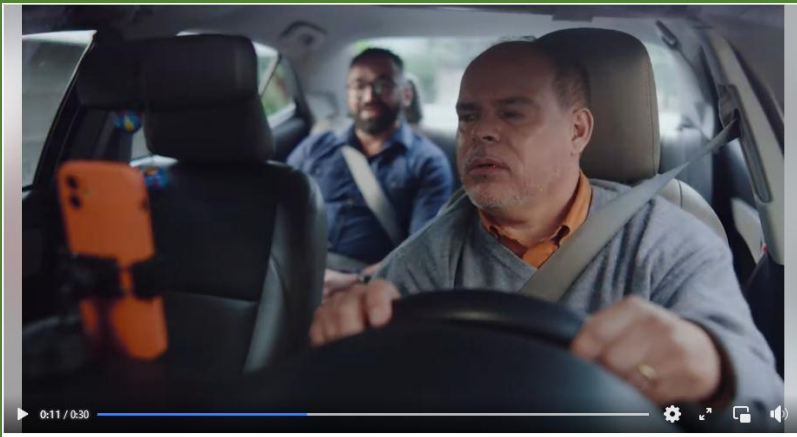
- Engenharia de Machine Learning**
Porto
São Paulo, Brasil (Híbrido)
17 ex-estudantes trabalham aqui
Há 1 hora
- Cientista de Dados (remoto)**
Data Science Brigade
Porto Alegre, Rio Grande do Sul, Brasil (Remoto)
Seu perfil corresponde a esta vaga
Há 1 dia • Candidatura simplificada
- Cientista de Dados - Sênior**
Mutant
Rio de Janeiro, Brasil (Presencial)
1 ex-estudante trabalha aqui
Há 2 horas
- Pessoa Cientista de Dados**
Zup Innovation
Brasil (Remoto)
1 conexão trabalha aqui
Há 4 horas • Candidatura simplificada
- Cientista de Dados Sênior**
banco BV
São Paulo, São Paulo, Brasil
1 conexão trabalha aqui
Há 6 dias
- Staff Data Scientist**
PicPay
Brasil (Remoto)

Mensagens

Oportunidades

- Oferta aparece com várias nomenclaturas diferentes
 - Data Science
 - Machine Learning
 - Big Data, Business Intelligence, Data Analytics, Data Engineering

Propaganda do Itaú

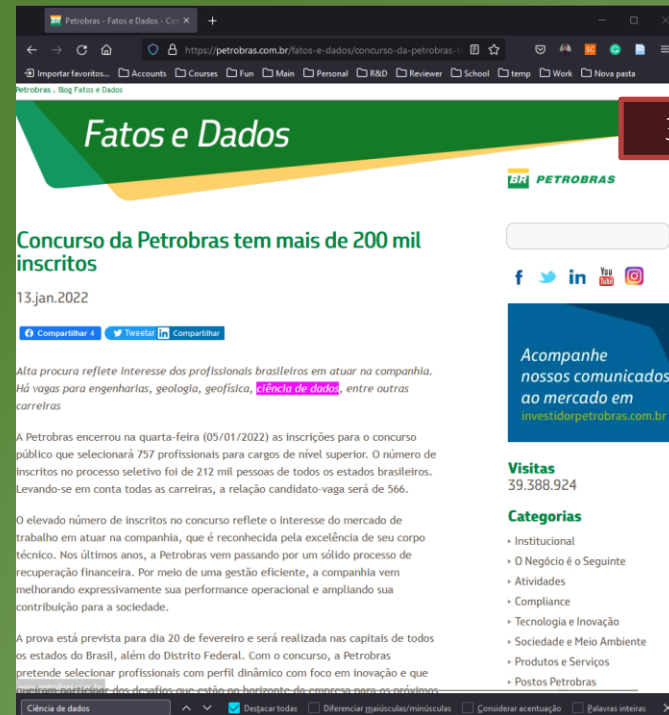


Não é uma panaceia

Estabilidade versus volatilidade

Expectativa versus realidade

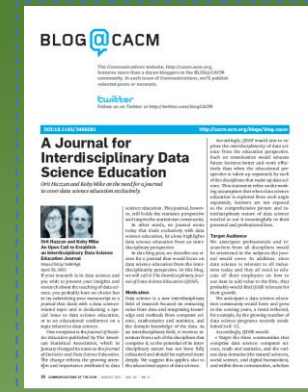
É um nicho, portanto, deve ser encarado como um diferencial



33 vagas

Educação no Brasil

- INEP – Indicadores CINE (2021- Inclusão de área detalhada e rótulos na área geral 06- Computação e Tecnologias da Informação e Comunicação)
- Referencias da SBC para Graduação em Ciência de Dados (jul/2023)
 - Graduação em Ciência de Dados
 - Angelo Brayner (UFC), Eduardo Ogasawara (CEFET/RJ), André de Carvalho (USP), Dulcan Ruiz (UFRGS)



Eixos propostos para graduação

- Fundamentos de Matemática, Estatística e Computação para Ciência de Dados
- Resolução de Problemas
- Desenvolvimento de Sistemas
- Engenharia e Exploração de Dados
- Gerência de Dados em Larga Escala
- Mineração de Dados e Aprendizado de Máquina
- Aprendizado Contínuo e Autônomo
- Ciência, Tecnologia, Inovação e Empreendedorismo

Análise da Pós-graduação em Ciência de Dados

- Estudo realizado em 2015
 - Predominância nos EUA
 - Foram estudados mais de vinte programas
 - Organização e disciplinas oferecidas
- Área Interdisciplinar
 - Interação com diversos departamentos
- Sólida ligação com
 - Ciência da Computação
 - Estatística
 - Área de Aplicação (interface)

CEFET/RJ – Programa de Pós-graduação em Ciência da Computação (PPCIC)

- Mestrado acadêmico criado desde 2016
 - Primeiro com o foco em Ciência de Dados
- Perfil do Egresso:
 - Sólida formação em Ciência da Computação
- Diferencial do Programa:
 - Formação de Cientistas de Dados
- Egresso deve estar apto a:
 - Atuar em setores onde há necessidade de extração de conhecimento a partir de dados
 - Tratar de questões teórico-práticas relacionadas aos problemas mais atuais relacionados à Computação
 - Propagar conhecimento no meio acadêmico-científico

Disciplinas

Básicas

- Análise e Projeto de Algoritmos
- Arquitetura de Computadores
- Banco de Dados
- Computação Paralela e Distribuída
- Metodologia Científica em Computação
- Métodos Estatísticos

Específicas

- Álgebra Linear Computacional
- Aprendizado de Máquina
- Ciência de Redes
- Gerência de Dados em Larga Escala
- Mineração de Dados
- Mineração de Textos
- Otimização por Metaheurísticas
- Visualização de Dados
- ...

Novidades

The image shows a screenshot of a web browser displaying the YouTube channel page for Eduardo Ogasawara. The browser's address bar shows the URL <https://www.youtube.com/channel/UCAm1hAXWEqYJfXz4EzzBhVg>. The YouTube interface includes a search bar, a 'SIGN IN' button, and a navigation sidebar on the left with options like Home, Explore, Shorts, Subscriptions, Library, and History. The channel header features Eduardo Ogasawara's profile picture, name, and '445 subscribers'. Below this are tabs for HOME, VIDEOS, PLAYLISTS, CHANNELS, and ABOUT. The 'HOME' tab is selected, showing a 'Destaque' (Featured) section with a 'PLAY ALL' button. Three video thumbnails are displayed: 'VISÃO GERAL DA PESQUISA' (15:38), 'INTRODUÇÃO' (25:31), and 'INTRODUÇÃO À METODOLOGIA CIENTÍFICA' (29:41). Each video is by Eduardo Ogasawara and includes view counts and upload dates. A green checkmark is placed over the 'ABOUT' tab.

YouTube BR

Search

SIGN IN

Eduardo Ogasawara
445 subscribers

SUBSCRIBE

HOME VIDEOS PLAYLISTS CHANNELS ABOUT

Destaque ▶ PLAY ALL

Pesquisas em andamento
Eduardo Ogasawara
249 views • 8 months ago

Introdução ao R
Eduardo Ogasawara
71 views • 3 weeks ago

Metodologia Científica - Introdução
Eduardo Ogasawara
237 views • 8 months ago

