INSTITUTO FEDERAL
Ceará

TELECOM WEEK

UMA BREVE INTRODUÇÃO À CIÊNCIA DE DADOS PARA (FUTUROS) PROGRAMADORES/CIENTISTAS

VI Semana da Engenharia de Telecomunicações

IFCE *campus* Fortaleza

7 de outubro de 2019

Prof. Me. Saulo Oliveira
saulo.oliveira@ifce.edu.br

Motivação

Dados podem fazer muito

E aí? Você ainda quer ser um cientista de dados?

Cientista de dados… a profissão *sexy*… do futuro

Só ter dados resolve tudo?

01

MOTIVAÇÃO

Hoje existem muitos dados

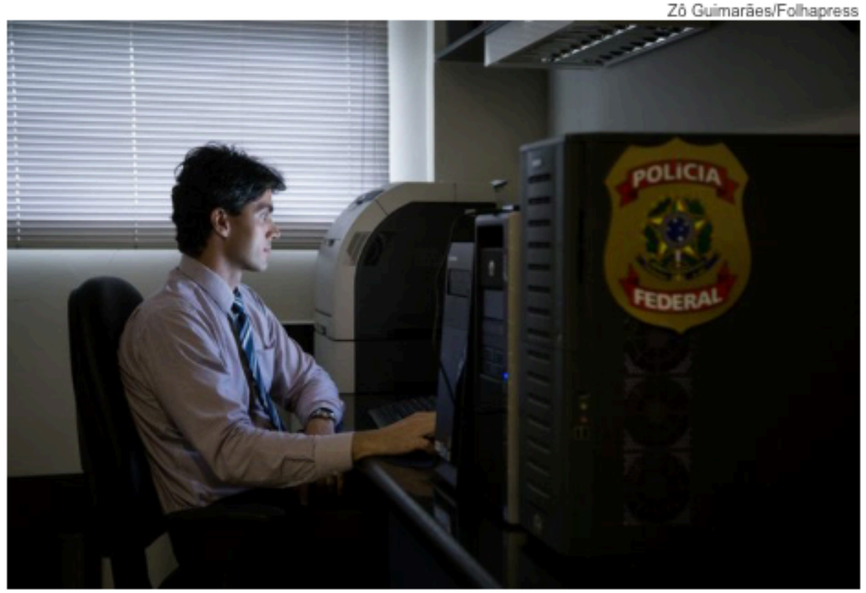
FOLHA DE S.PAULO

★ ★ ★ UM JORNAL A SERVIÇO DO BRASIL

SEXTA-FEIRA, 2 DE NOVEMBRO DE 2018 22:24

Volume de dados da Lava Jato leva PF a criar novo sistema

Zô Guimarães/Folhapress



Luís Filipe da Cruz Nassif, perito da PF que desenvolveu software para processar dados

FLÁVIO FERREIRA
DE ENVIADO ESPECIAL AO RIO DE JANEIRO

02/01/2017 02h00

Forbes

Big Data In Banking: How Citibank Delivers Real Business Benefits With Its Data-First Approach

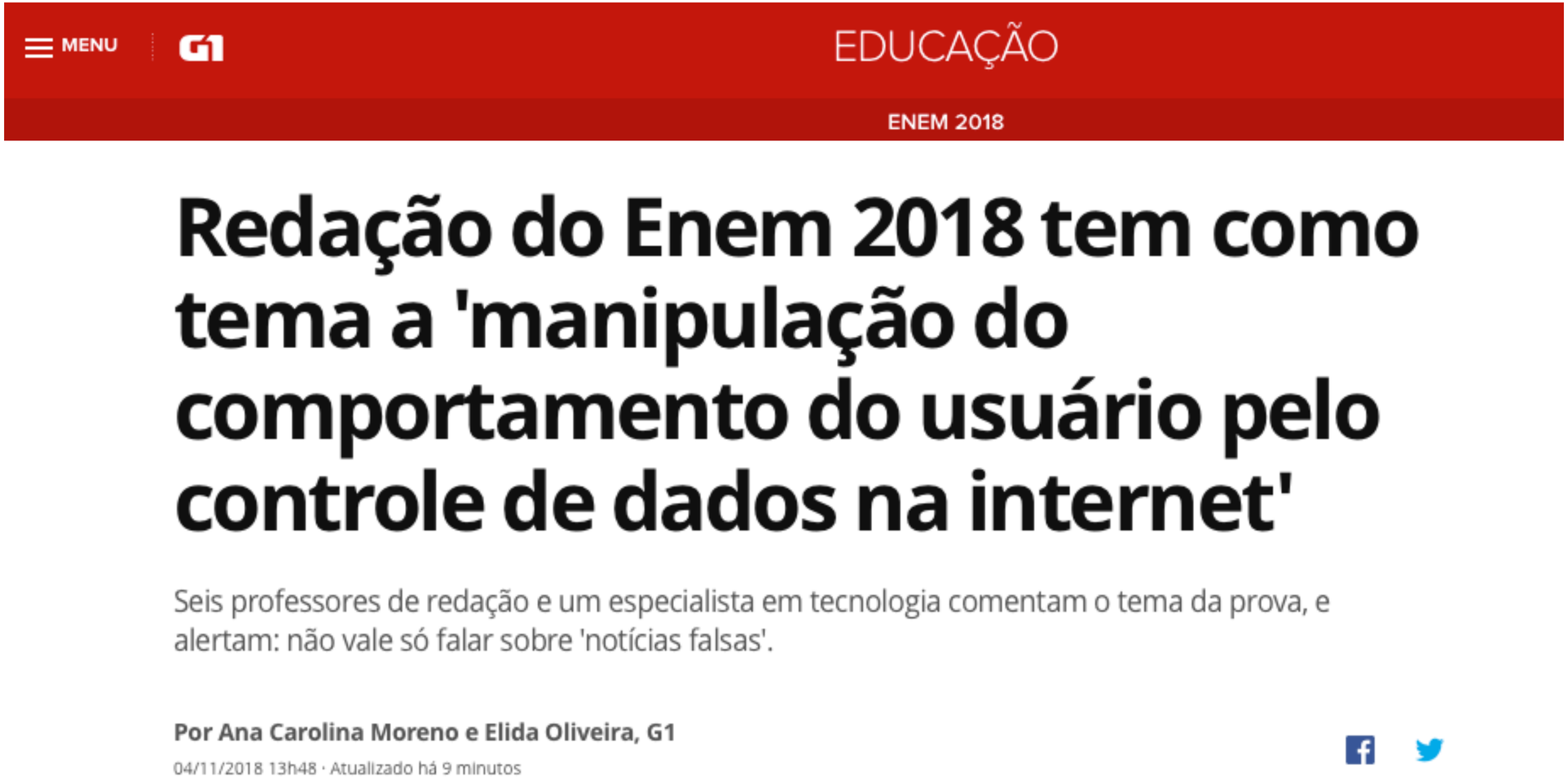
Bernard Marr Contributor

The Guardian International edition

How big data is changing how we study languages

Big data is enriching the field of language study, but data access needs to be opened up more for academics to scrunitise the figures properly

Relevância dos dados



MENU G1 EDUCAÇÃO

ENEM 2018

Redação do Enem 2018 tem como tema a 'manipulação do comportamento do usuário pelo controle de dados na internet'

Seis professores de redação e um especialista em tecnologia comentam o tema da prova, e alertam: não vale só falar sobre 'notícias falsas'.

Por Ana Carolina Moreno e Elida Oliveira, G1

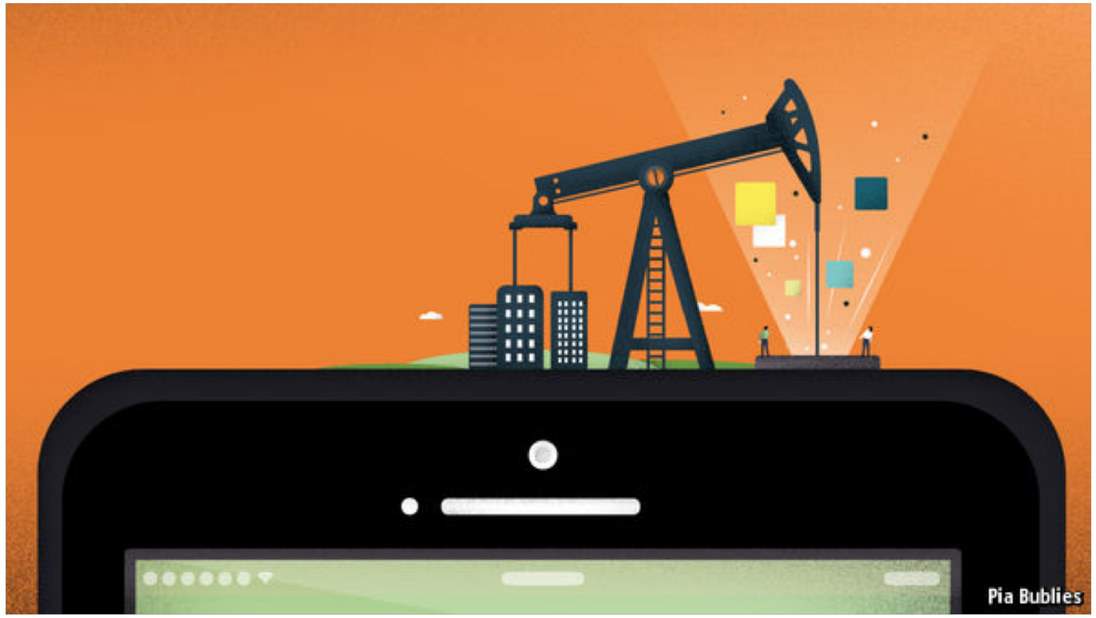
04/11/2018 13h48 · Atualizado há 9 minutos

https://g1.globo.com/educacao/enem/2018/noticia/2018/11/04/redacao-do-enem-2018-tem-como-tema-manipulacao-do-comportamento-do-usuario-pelo-controle-de-dados-na-internet.ghtml

Fuel of the future

Data is giving rise to a new economy

How is it shaping up?



Print edition | Briefing ›
May 6th 2017

“Dados são para este século o que o petróleo foi para o século passado: um impulsionador de crescimento e mudanças.”

https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

Visão de futuro…

Chefe da equipe econômica do Google, Hal Varian, para Statistics and Data

February 25, 2009

Topic

Quotes, Statistics

> *A capacidade de utilizar dados – de ser capaz de entendê-los, processá-los, extrair valor deles, visualizá-los, comunicá-los – Essa será uma habilidade importantíssima nas próximas décadas.*

Hal Varian, The McKinsey Quarterly, January 2009

https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/

02

CIENTISTA DE DADOS… A PROFISSÃO SEXY… DO FUTURO

O emprego dos sonhos...

Chefe da equipe econômica do Google, Hal Varian, para Statistics and Data

February 25, 2009

Topic

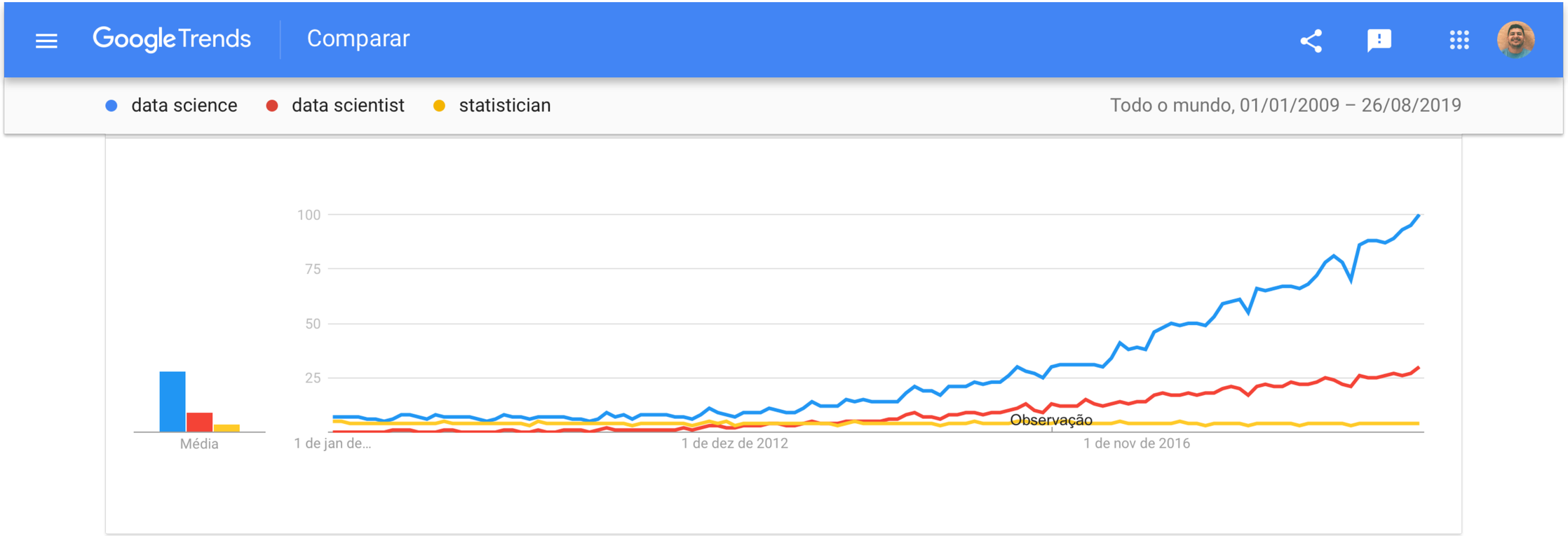
Quotes, Statistics

> *Eu fico dizendo que o emprego dos sonhos nos próximos dez anos será a do Estatístico. Pessoas acham que eu estou fazendo algum tipo de brincadeira, mas quem imaginaria que os engenheiros da computação teriam o emprego dos sonhos nos anos 90?*

Hal Varian, The McKinsey Quarterly, January 2009

https://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/

Data science: Tendência



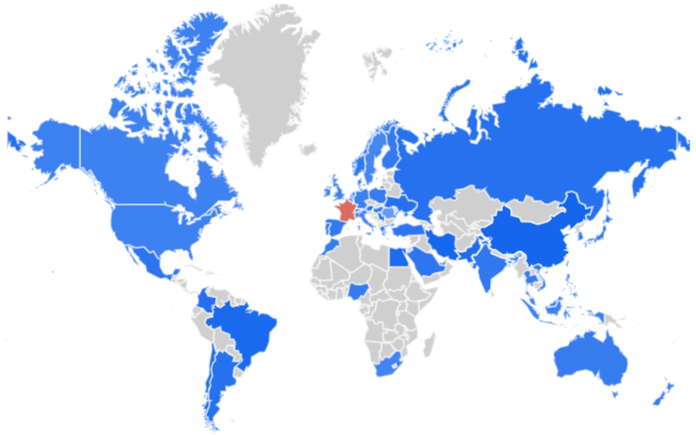
https://trends.google.com.br/trends/explore?date=2009-01-01%202019-08-26&q=data%20science,data%20scientist,statistician

Data science: Tendência

Detalhamento comparado por região

Região

data science data scientist statistician



A intensidade da cor representa o percentual de pesquisas SAIBA MAIS

Classificar: Interesse em data science

1	China	
2	Irã	
3	Brasil	
4	Egito	
5	Chile	

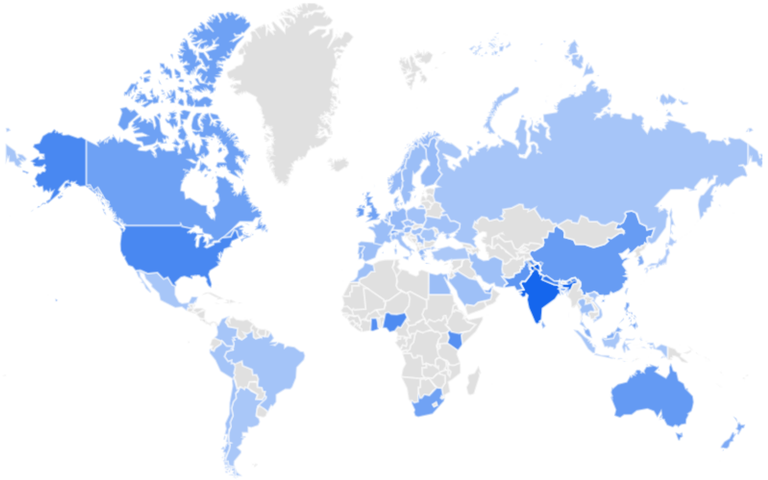
Incluir regiões com baixo volume de pesquisa

Mostrando 1 a 5 de 54 regiões

https://trends.google.com.br/trends/explore?date=2009-01-01%202019-08-26&q=data%20science,data%20scientist,statistician

Data science: Tendência

Interesse por região | Região



☐ Incluir regiões com baixo volume de pesquisa

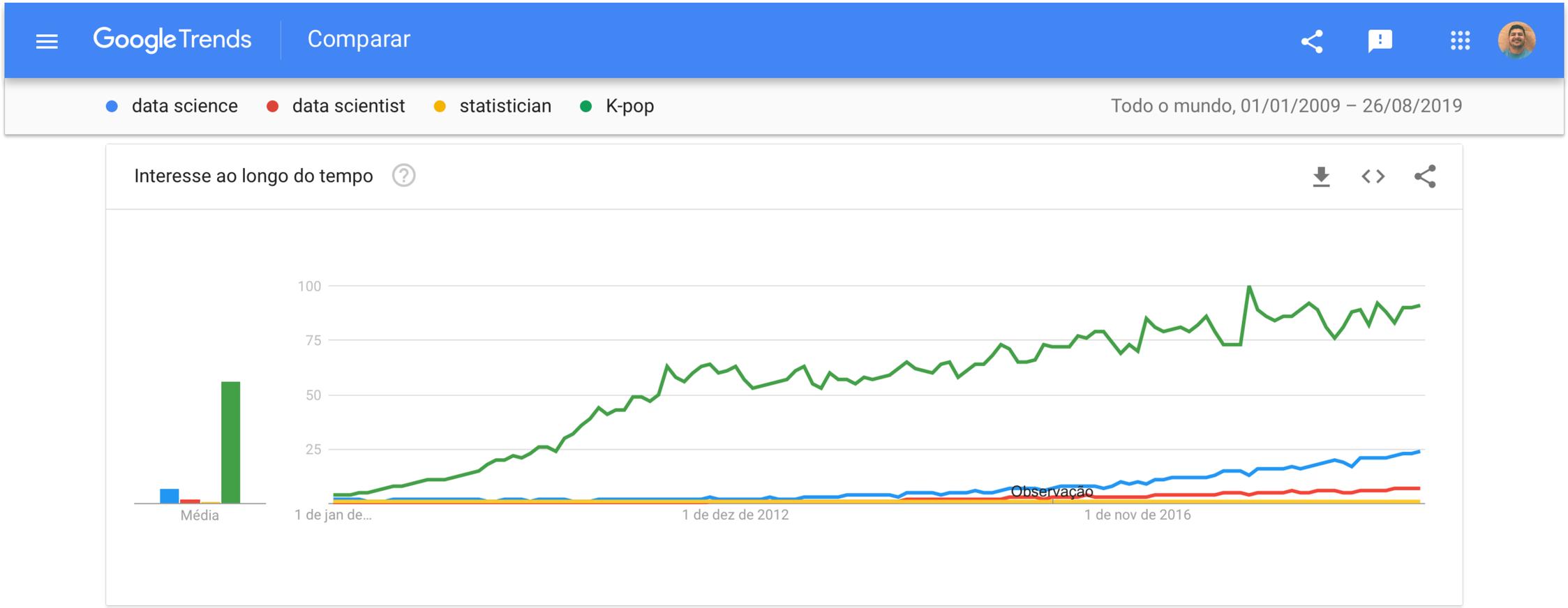
Consultas relacionadas | Em ascensão

1	edx	Aumento repentino
2	udemy data science	Aumento repentino
3	data science from scratch	Aumento repentino
4	udemy	Aumento repentino
5	r programming for data science	Aumento repentino

Mostrando 1 a 5 de 25 consultas

https://trends.google.com.br/trends/explore?date=2009-01-01%202019-08-26&q=data%20science,data%20scientist,statistician

Em perspectiva...



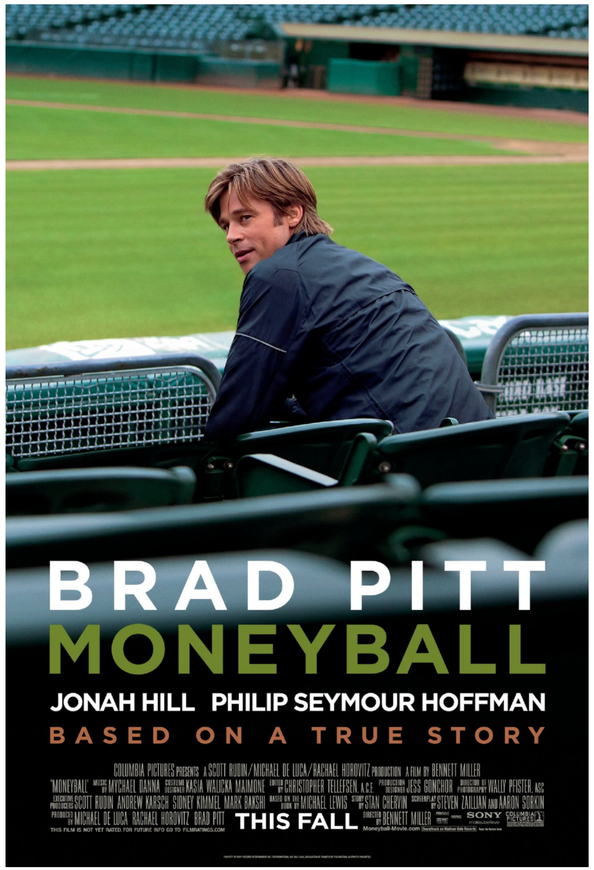
https://trends.google.com.br/trends/explore?date=2009-01-01%202019-08-26&q=data%20science,data%20scientist,statistician,%2Fm%2F02yh8l

03

DADOS PODEM FAZER MUITO

Dados podem fazer muito: Esportes

O filme é baseado no livro Moneyball: The Art of Winning an Unfair Game.



https://www.nytimes.com/video/movies/100000001341145/clip-moneyball.html

Dados podem fazer muito: Esportes

THE WALL STREET JOURNAL.

CIO JOURNAL.

The Morning Download: How German Soccer Team Scored With Big Data

By *Steve Rosenbush*

0 COMMENTS

Jul 11, 2014 7:43 am ET

The Morning Download comes from the editors of CIO Journal and cues up the most important news in business technology every weekday morning. Send us your tips, compliments and complaints. You can get The Morning Download emailed to you each weekday morning by clicking here.

Good morning. There's no shortage of explanations for Germany's stunning victory over Brazil, which will send it to the World Cup final on Sunday in a match vs. Argentina. But we know this: to gain a competitive edge, the team partnered with German software giant **SAP AG** to create a custom match analysis tool that collects and analyzes massive amounts of player performance data.

http://blogs.wsj.com/cio/2014/07/10/germanys-12th-man-at-the-world-cup-big-data/

Dados podem fazer muito: Saúde

https://exame.abril.com.br/tecnologia/5-coisas-que-o-big-data-faz-pela-sua-saude/

Dados podem fazer muito: Recomendação

- **Outubro de 2006:** A Netflix ofereceu US $1 milhão por um algoritmo de recomendação aprimorado.

- **Dados de treinamento:**
  - 100M avaliações;
  - 480K usuários;
  - 17.770 filmes;
  - 6 anos de dados: 2000-2005.

- **Dados de teste:**
  - Últimos avaliações de cada usuário (2.8M);
  - Avaliação via RMSE: raiz do erro quadrático médio;
  - RMSE da ferramenta da Netflix: 0,9514.

- **Concorrência:**
  - Grande prêmio de US $ 1 milhão para 10% de melhoria;
  - Se 10% não forem cumpridos, $ 50K anuais para melhor melhoria**.**

https://www.netflixprize.com

Dados podem fazer muito: Descobrir viés

Stanford | News

Search Stanford news...

Home Find Stories For Journalists Contact

JUNE 15, 2016

Stanford big data study finds racial disparities in Oakland, Calif., police behavior, offers solutions

Stanford researchers analyzing thousands of data points found racial disparities in how Oakland Police Department officers treated African Americans on routine traffic and pedestrian stops. The researchers suggest 50 measures to improve police-community relations, such as better data collection, bias training and changes in cultures and systems.

BY CLIFTON B. PARKER

New Stanford research on thousands of police interactions found significant racial differences in Oakland, California, police conduct toward African Americans in traffic and pedestrian stops, while offering a big data approach to improving police-community relationships there and elsewhere.

https://news.stanford.edu/2016/06/15/stanford-big-data-study-finds-racial-disparities-oakland-calif-police-behavior-offers-solutions/

Dados podem fazer muito: Descobrir viés

INTELIGÊNCIA ARTIFICIAL ›

Se está na cozinha, é uma mulher: como os algoritmos reforçam preconceitos

As máquinas inteligentes consolidam os vieses sexistas, racistas e classistas que prometiam resolver

https://brasil.elpais.com/brasil/2017/09/19/ciencia/1505818015_847097.html

Dados podem fazer muito: Descobrir viés

EL PAÍS TECNOLOGIA

O robô racista, sexista e xenófobo da Microsoft acaba silenciado

Projetado para o mercado dos 'millennials' nos Estados Unidos, Tay não foi capaz de lidar com piadas e perguntas controvertidas

https://brasil.elpais.com/brasil/2016/03/24/tecnologia/1458855274_096966.html

Dados podem fazer muito: Política

NEWS | BRASIL

Notícias | Brasil | Internacional | Economia | Saúde | Ciência | Tecnologia | Aprenda Inglês | #SalaSocial | Galeria de Fotos

Robôs e 'big data': as armas do marketing político para as eleições de 2018

Camilla Veras Mota - @cavmota
Da BBC Brasil em São Paulo

26 setembro 2017

Compartilhar

https://www.bbc.com/portuguese/brasil-41328015

04

SÓ TER DADOS RESOLVE TUDO?

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

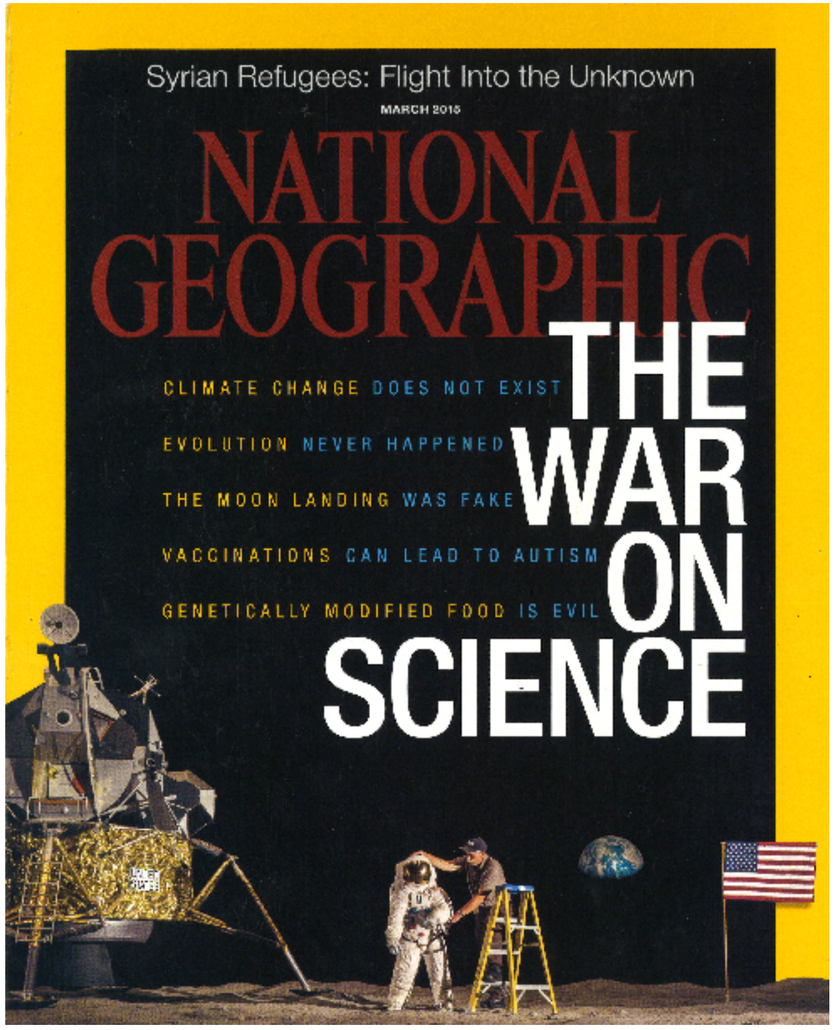
Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

https://www.wired.com/2008/06/pb-theory/

A Ciência está perdendo sua relevância?



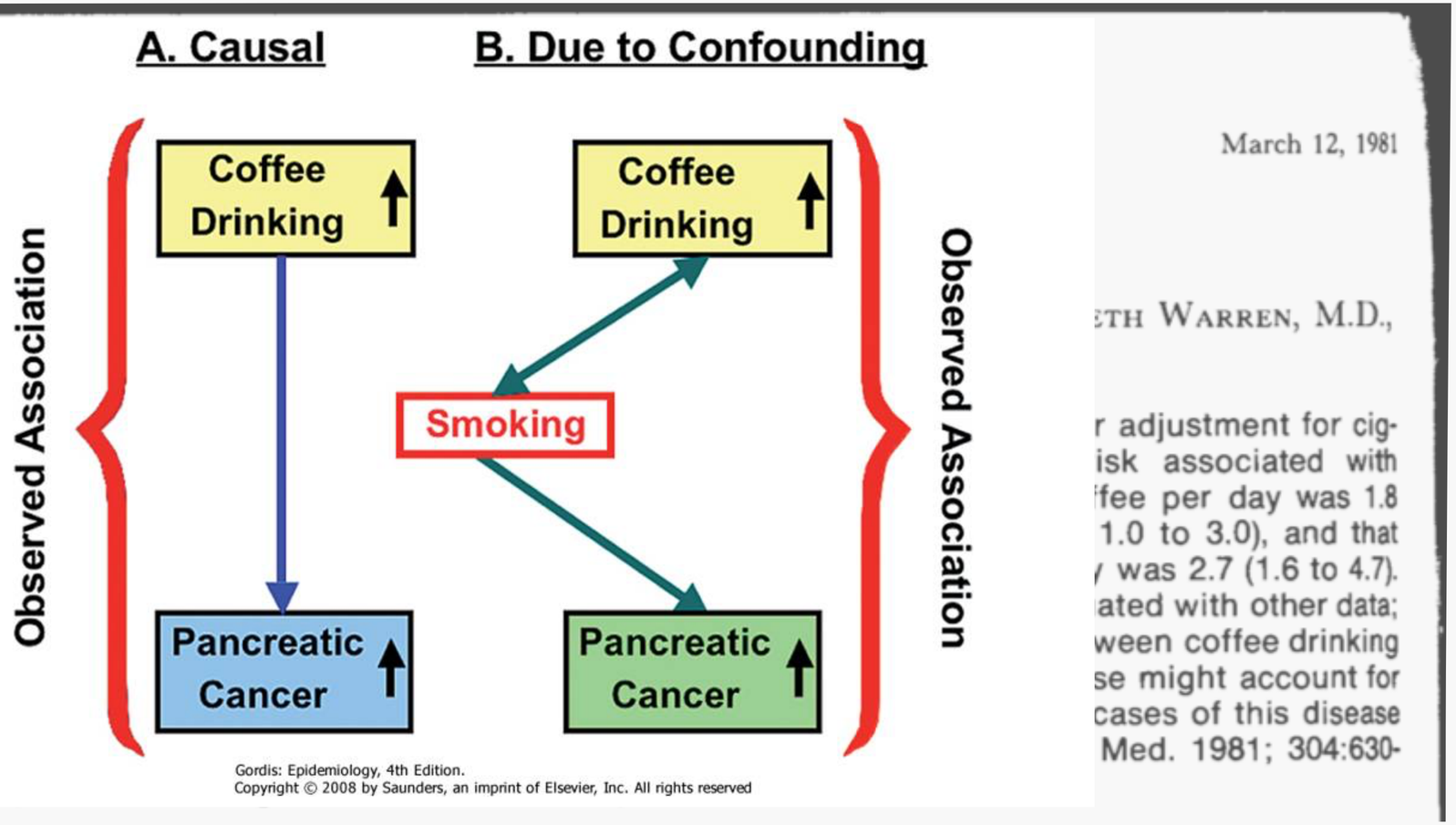
The NEW ENGLAND JOURNAL of MEDICINE

BRIAN MAC

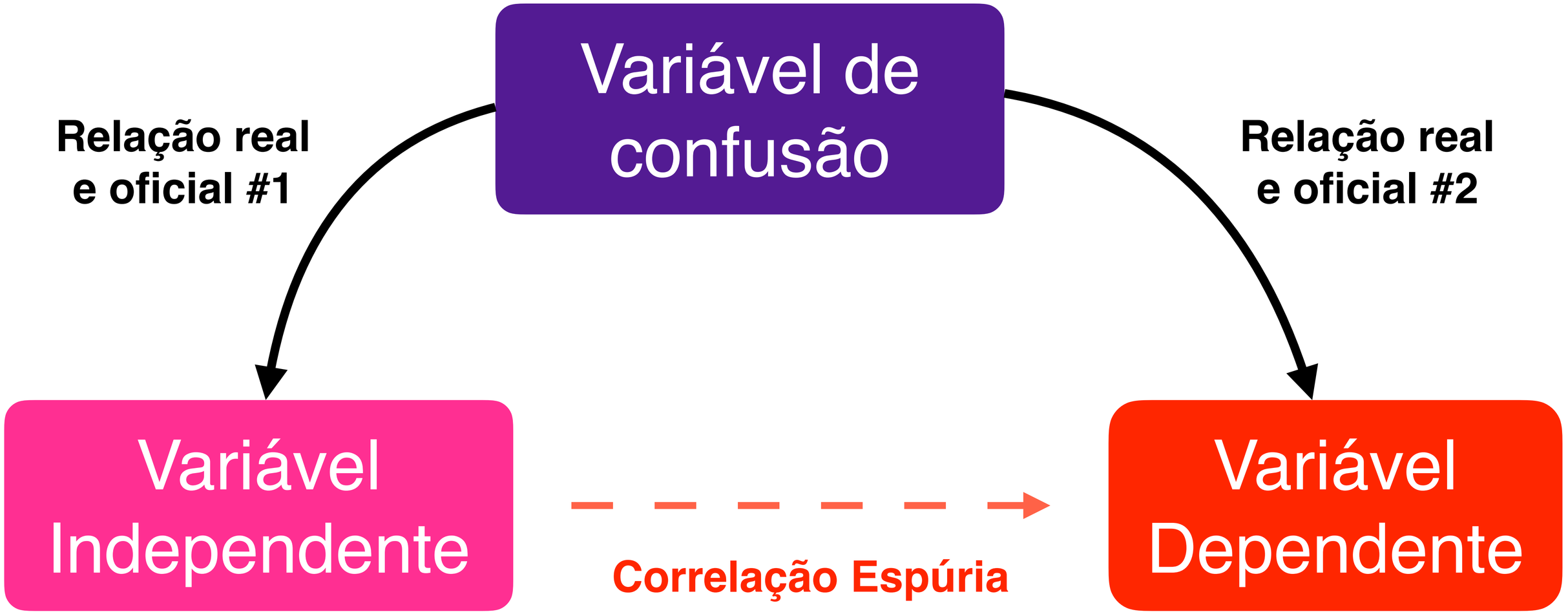
ETH WARREN, M.D.,

Abstract W
logically pr
control patie
tea, and cof
tion betwee
ing, but we
pipe tobacc
sociation be
ic cancer wa
was not affe
the sexes co

r adjustment for cig-
isk associated with
fee per day was 1.8
1.0 to 3.0), and that
was 2.7 (1.6 to 4.7).
ated with other data;
ween coffee drinking
se might account for
cases of this disease
Med. 1981; 304:630-



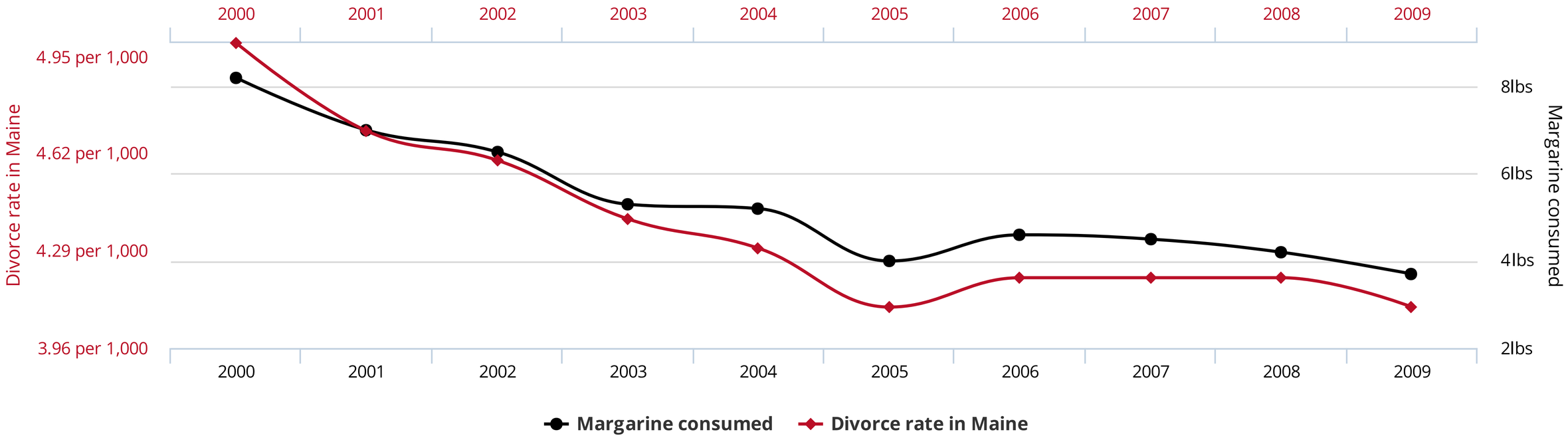
https://www.nejm.org/doi/full/10.1056/NEJM198103123041102


Variável de confusão
Relação real e oficial #1
Relação real e oficial #2
Variável Independente
Variável Dependente
Correlação Espúria

Divorce rate in Maine

correlates with

Per capita consumption of margarine



tylervigen.com

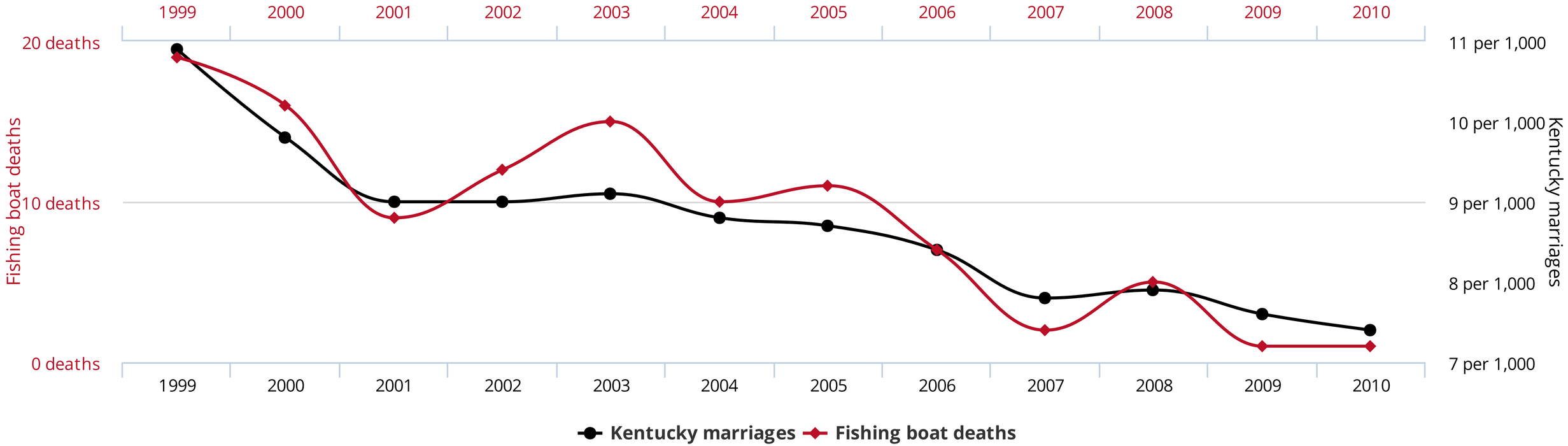
Índice de correlação: **99,79%**

https://www.tylervigen.com/spurious-correlations

People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky



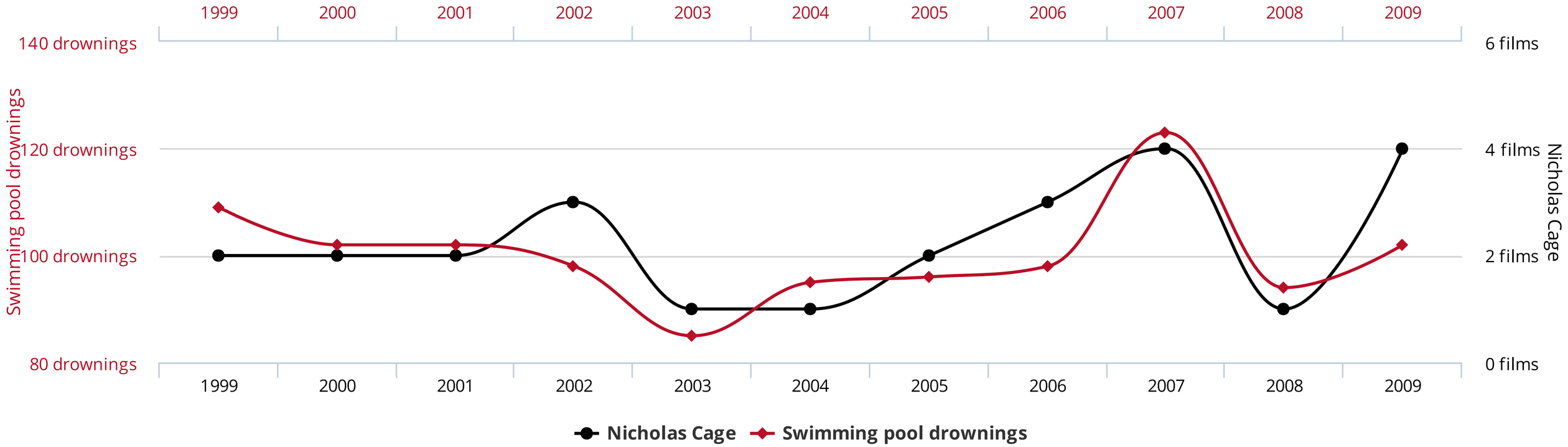
tylervigen.com

Índice de correlação: **95,24%**

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



tylervigen.com

Índice de correlação: **66,60%**

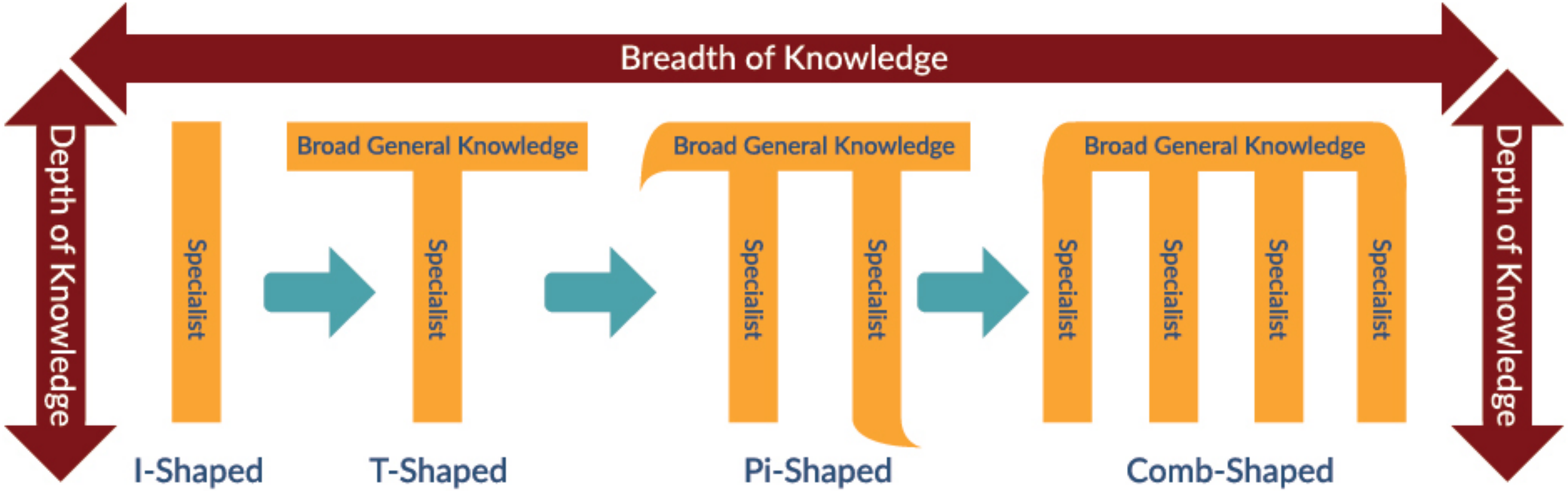
https://www.tylervigen.com/spurious-correlations

É fácil encontrar padrões “interessantes” onde não existem!

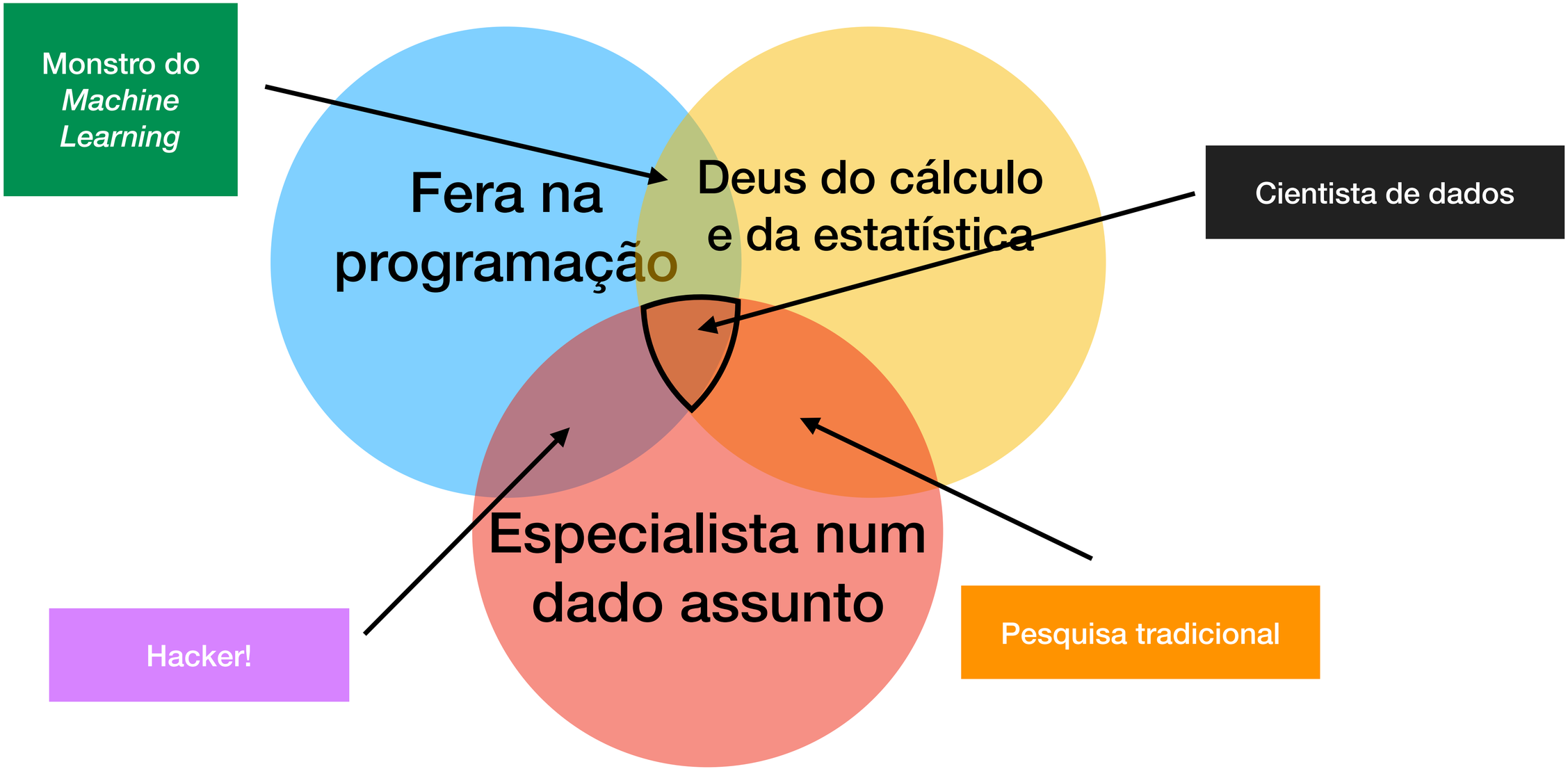
05

E AÍ? VOCÊ AINDA QUER SER UM CIENTISTA DE DADOS?

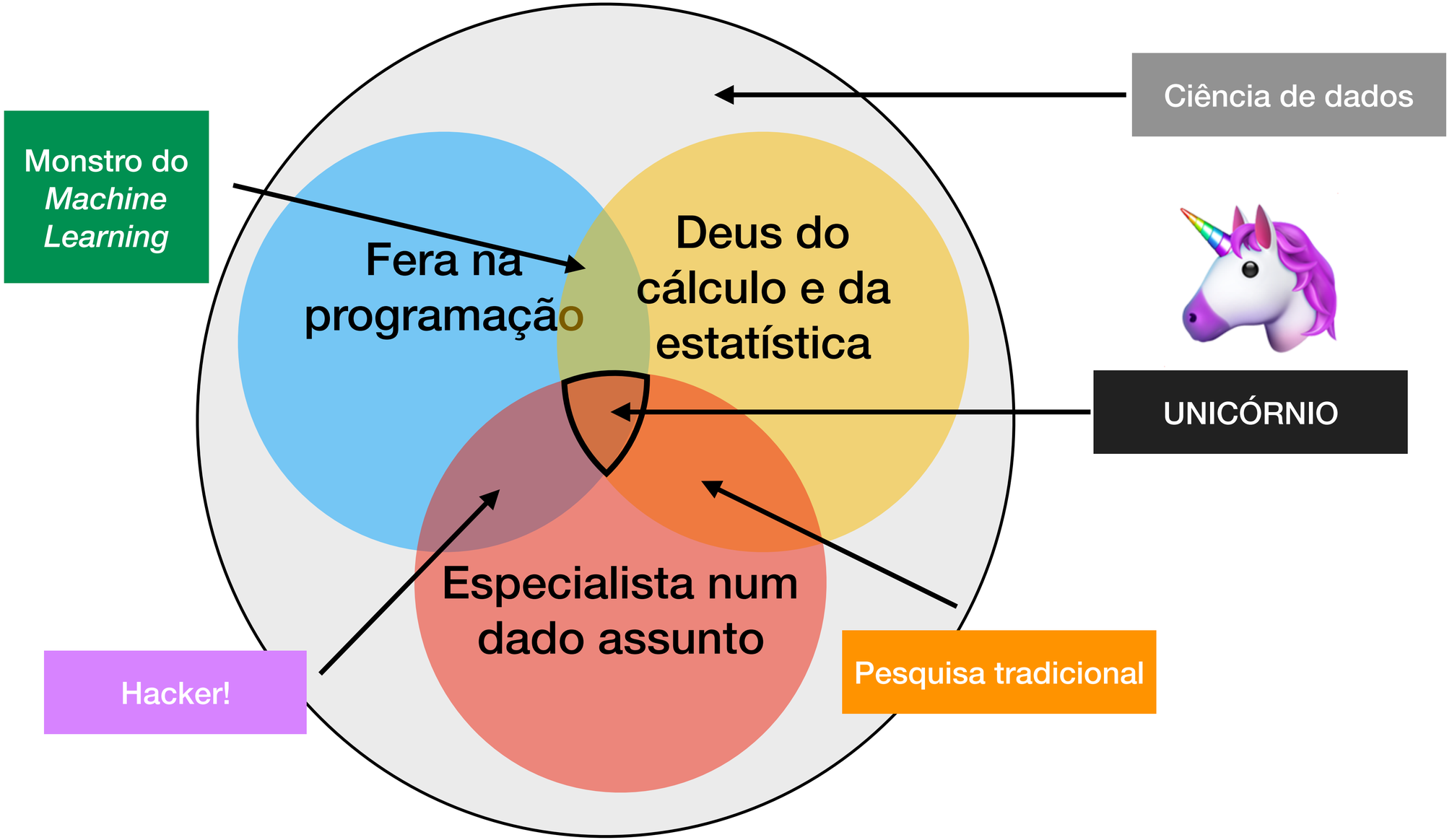
Mas o que é um cientista de dados?



Mas o que é um cientista de dados?



Mas o que é um cientista de dados?



E AÍ? VOCÊ AINDA QUER SER UM CIENTISTA DE DADOS?

...PROVAVELMENTE VOCÊ JÁ TEM POR ONDE COMEÇAR!

TENHA ACESSO A COLEÇÕES DE DADOS TEMÁTICOS EM DIFERENTES GRAUS DE ORGANIZAÇÃO:

UCI Machine Learning Repository

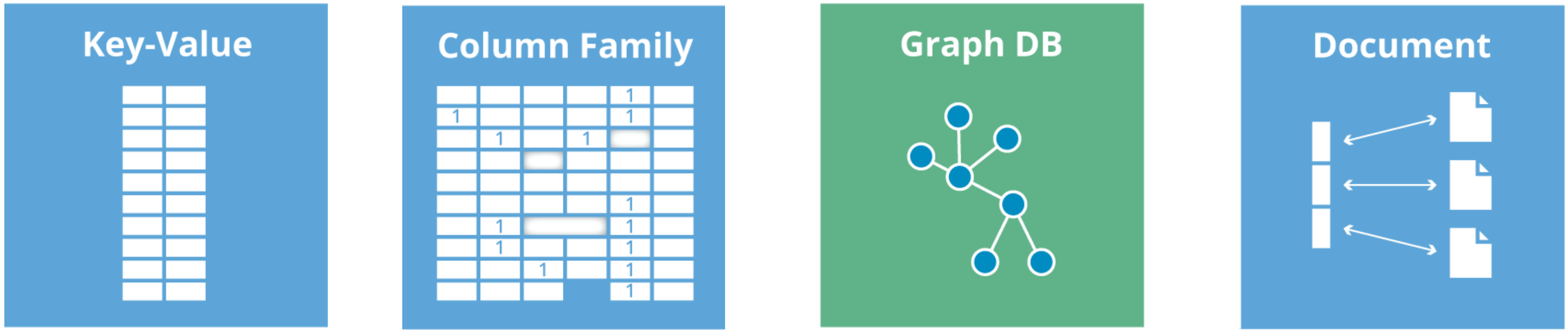
https://archive.ics.uci.edu/ml/index.php

kaggle

https://www.kaggle.com

NA MEDIDA EM QUE A ESTRUTURA DO DADO É MAIS VERSÁTIL AS TABELAS FICAM MAIS COMPLEXAS...

CONHEÇA ALGUNS BANCOS DE DADOS NOSQL, JÁ QUE ELES PODEM SER MAIS FLEXÍVEIS PARA DADOS COM ESTRUTURAS DIFERENTES



https://neo4j.com/blog/why-nosql-databases/

BUSQUE APRENDER AS LINGUAGENS NOVAS QUE FORNECEM SUPORTE À FERRAMENTAS INCRÍVEIS (BIBLIOTECAS E FRAMEWORKS)

https://www.r-project.org

https://www.python.org

https://www.scala-lang.org

BUSQUE POSSUIR UMA LISTA DE CONHECIMENTOS E CAPACIDADES...

- **NÃO EXCLUSIVAS:** NOVAS TECNOLOGIAS APARECEM O TEMPO TODO;
- **SEM VIÉS:** É SAUDÁVEL QUESTIONAR ALGUMAS IDEIAS;
- **POTENCIALMENTE REDUNDANTES:** O CIENTISTA DE DADOS TEM QUE SABER COMO JOGAR EM VÁRIAS POSIÇÕES EM VÁRIOS TIMES;
- **NÃO NECESSARIAMENTE TÉCNICOS:** CIÊNCIA DE DADOS DEVE ENVOLVER ASPECTOS DO MUNDO REAL.

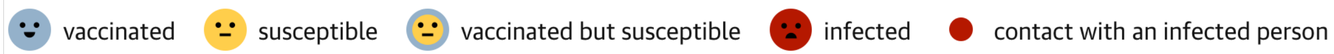
ONDE COMEÇAR A ESTUDAR CIÊNCIA DE DADOS E ESTATÍSTICA?

VOCÊ + DADOS + COMUNICAÇÃO = VISUALIZAÇÃO 🖤

OUTRA ÁREA INTERDISCIPLINAR:

- **VISUALIZAÇÃO:** ARTE E CIÊNCIA.
- **DESIGN:** SIGNIFICADO PARA USUÁRIOS.

Watch how the measles outbreak spreads when kids get vaccinated – and when they don't



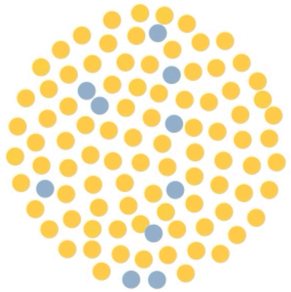
TAXA DE VACINAÇÃO CONTRA SARAMPO

	10%	30%	50%	58,5%	68,9%	74,4%	83,8%	86%	90%	99,7%
PROTEGIDO	10%	30%	**49,5%**	**57,5%**	**68%**	**73,2%**	**83%**	**85%**	**89%**	**99%**
SUSCETÍVEL	9%	7%	5%	**42,5%**	3%	3%	2%	1%	11%	1%
INFECTADO	**81%**	**63%**	**45,5%**	0%	29%	23,8%	15%	14%	0%	0%

https://www.theguardian.com/society/ng-interactive/2015/feb/05/-sp-watch-how-measles-outbreak-spreads-when-kids-get-vaccinated

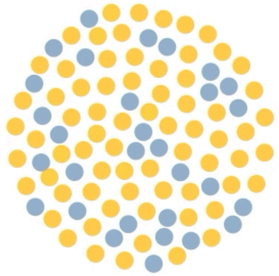
Watch how the measles outbreak spreads when kids get vaccinated – and when they don't

vaccinated · susceptible · vaccinated but susceptible · infected · contact with an infected person



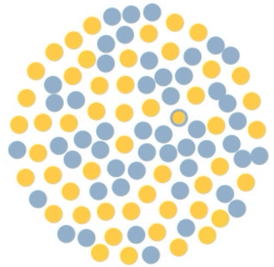
NOT PROTECTED

10.0% vax rate



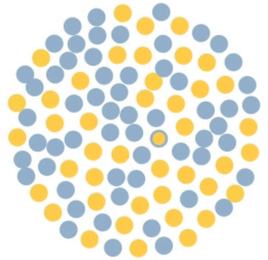
NOT PROTECTED

30.0% vax rate



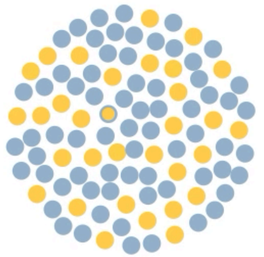
NOT PROTECTED

50.0% vax rate



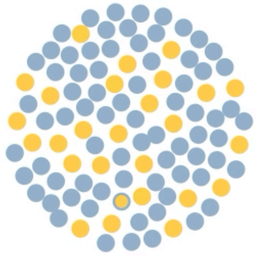
NOT PROTECTED

58.5% vax rate, similar to **Okanagan County, WA**



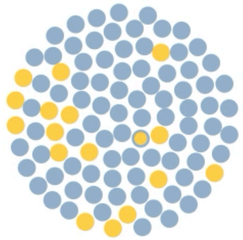
NOT PROTECTED

68.9% vax rate, similar to **Thurston County, WA**



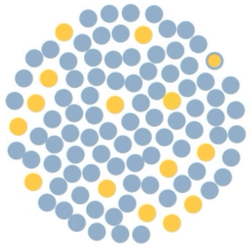
NOT PROTECTED

74.4% vax rate, similar to **Island County, WA**



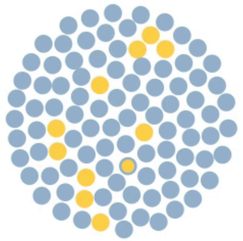
PROTECTED

83.8% vax rate, similar to **Santa Cruz County, CA**



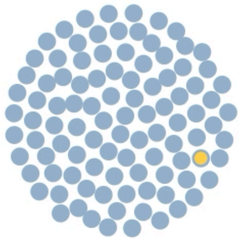
NOT PROTECTED

86.0% vax rate, similar to **Los Angeles County, CA**



PROTECTED

90.0% vax rate, similar to **Orange County, CA**



PROTECTED

99.7% vax rate, similar to **Gadsden County, FL**

https://www.theguardian.com/society/ng-interactive/2015/feb/05/-sp-watch-how-measles-outbreak-spreads-when-kids-get-vaccinated

“A capacidade de utilizar dados – de ser capaz de entendê-los, processá-los, extrair valor deles, visualizá-los, comunicá-los – Essa será uma habilidade importantíssima nas próximas décadas.”

—Hal Varian, Google’s Chief Economist

Referências

Fora os links, duas apresentações foram importantíssimas para a elaboração desse material.

“CONCEITOS DE DATA SCIENCE”, Rafael Santos. Link: http://www.lac.inpe.br/~rafael.santos/Docs/WorCAP/IntroDataScience.pdf

Data Science 101: an introduction. Data Science Team. Stanford University, Department of Statistics. Link: https://web.stanford.edu/class/stats101/intro/intro-lecture01.pdf

Muito obrigado pela atenção!

Prof. Me. **Saulo Oliveira**
IFCE - Campus Tauá
Email: saulo.oliveira@ifce.edu.br

INSTITUTO FEDERAL
Ceará

TELECOM WEEK

UMA BREVE INTRODUÇÃO À CIÊNCIA DE DADOS PARA (FUTUROS) PROGRAMADORES/CIENTISTAS

VI Semana da Engenharia de Telecomunicações

IFCE *campus* Fortaleza

7 de outubro de 2019

Prof. Me. Saulo Oliveira
saulo.oliveira@ifce.edu.br