Introduçãoa a Data Science em Epidemiologia

Desafios e Perspectivas

Oswaldo G Cruz [1]

[1]Programa de Computação Cientifica
Fundação Oswaldo Cruz

Disciplina Data Science aplicada a Epidemiologia

Apresentação

1. O que é:
   - Conceitos Fundamentais
   - Big Data
   - Data Mining
   - Data Science
2. Resultados/Aplicações
   - o que tem sido publicado?
   - Algumas Aplicações
3. Considerações finais

Para enteder o que é DataSciene

- Big Data / Data Mining
- Analise Estatística / Machine Learning / Deep Learning / Predição (forecast)
- Linguagens de programação R, Python, Julia, Java, JS, etc...
- Banco de Dados / Estrutura de Dados /SQL vs NoSQL
- Linguagens de Dados SQL
- Técnicas de visualização de dados
- Reprodutibilidade científica

Big Data está na mídias e nas revistas científicas

Definition

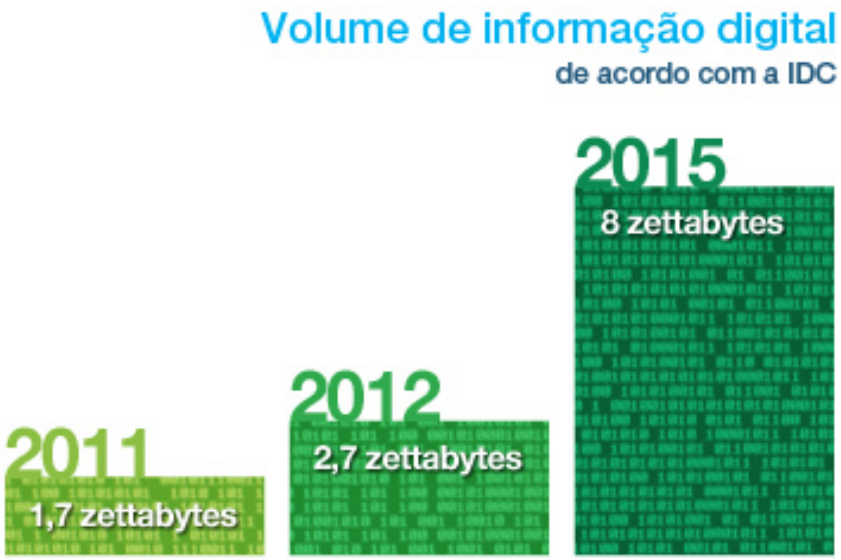
Big data is an all encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. (wikipédia)

O que é Big Data

IBM

O que é Big Data?

Big Data é um termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia.
As dificuldades em armazenar, analisar e utilizar grandes conjuntos de dados têm sido um considerável gargalo para as companhias.



o que é um grande volume de dados?

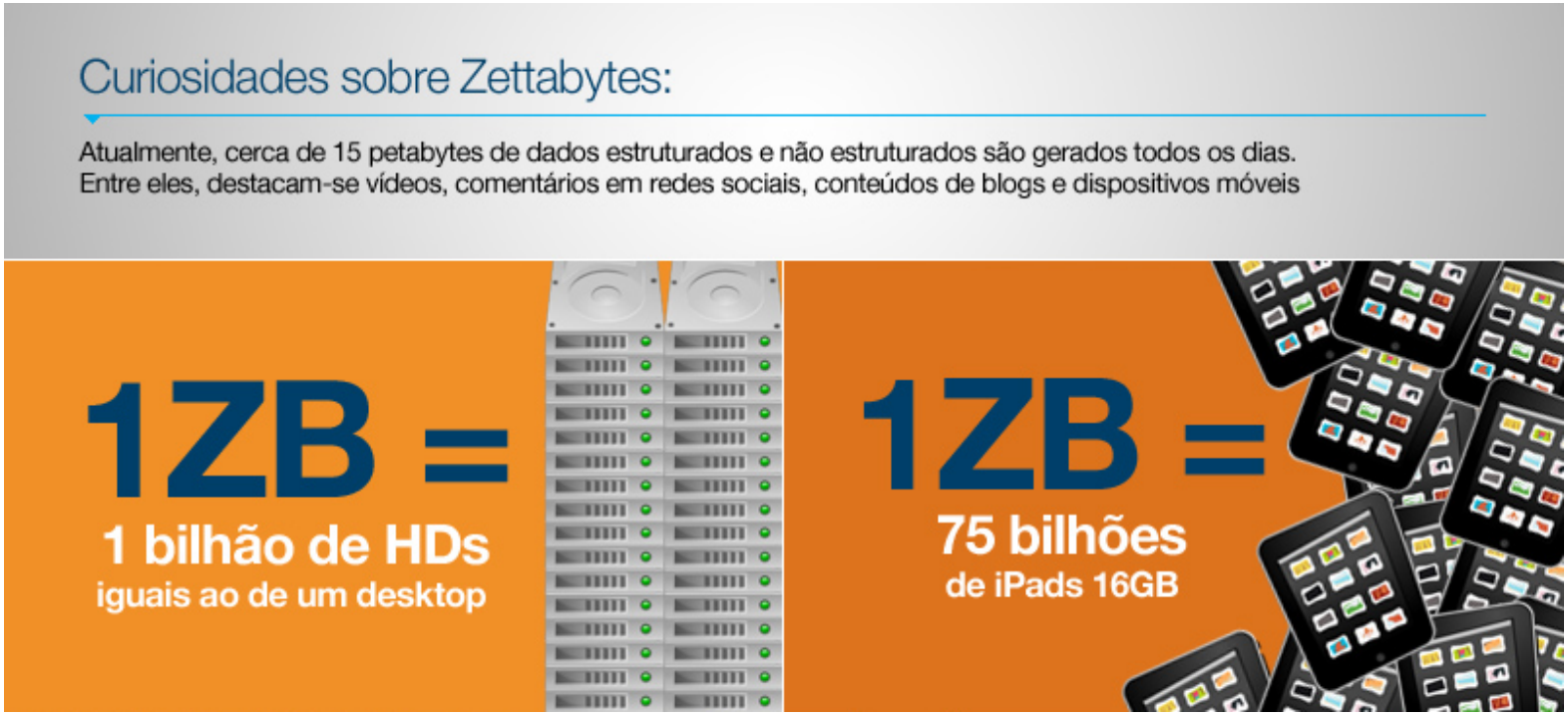
- um Pentabyte são 10^{15}Bytes ou seja 1.000.000.000.000.000
- para melhor comparação vamos usar o Gigabye (10^{9}byte). 1 minuto de vídeo HD ocupa 1 Gigabyte ,10 metros de livros em uma estante
- 1 pentabyte equivale a 13 anos de video HD, 20 milhões de fichários , 20 pentabytes é o volume de dados processado por dia pelo Google

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone...” IBM, 2012

fonte: www.mozy.com

o que é um grande volume de dados?

- um zetabyte é um milhão de pentabytes



fonte: IBM

Tipos de Dados

- dados estruturados (Banco de Dados tradicionais)
- dados semi estruturados (XML, RSS, Planilhas)
- dados não estruturados (emails, blogs,redes sociais)
- grafos (rede sociais, web semantica,RDF)

os 3 Vs

Volume

Geramos um número gigantesco de dados diariamente, e estima-se que esse volume dobre a cada 18 meses.

Variedade

Esses dados vêm de sistemas estruturados e não estruturados gerados por emails, postagens em mídias sociais, mensagens instantâneas, etiquetas RFID, câmeras de vídeo, etc.

Velocidade

Muitas vezes precisamos agir em tempo real para lidar com essa imensa quantidade de dados.

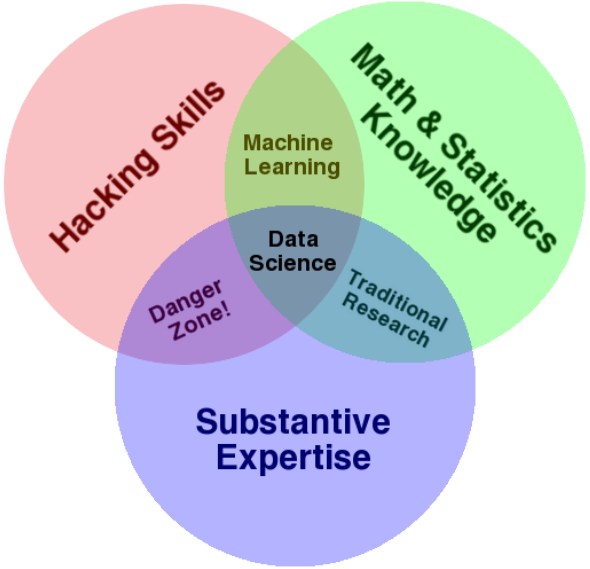
fonte: IBM

o que é Data Mining

- Processo automático de descoberta de informações uteis em grandes repositórios de dados
- Processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequência temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.
- Processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de grandes conjuntos de dados ;varias outras definições
- Data Mining pode auxiliar os pesquisadores Identificando padrões, classificando, segmentando os dados, contribuindo na formulação de hipóteses , etc...

o que é Data Science

- Data Science é uma denominação recente que tem sido aplicada a um campo interdisciplinar é que envolve a interseção de áreas da estatística, ciência da computação e domínios de técnicas de trabalho com big data, data mining, machine learning etc...
- consiste na a extração de conhecimento diretamente dos dados por meio de um processo de descoberta, hipótese, e análise.

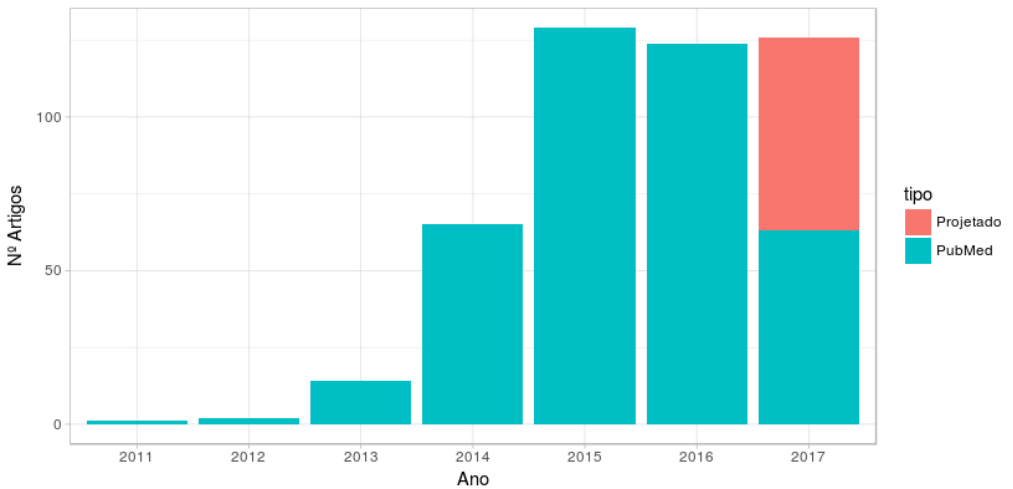


Porque Aprender ?

O uso da metodologia de Data Science oferece a Epidemiologia moderna m conjunto de ferramentas altamente eficientes para a rápida análise, detecção , “predição” e tomada de decisões sobre epidemias , padrões de doenças, planejamento em saúde , etc. . .

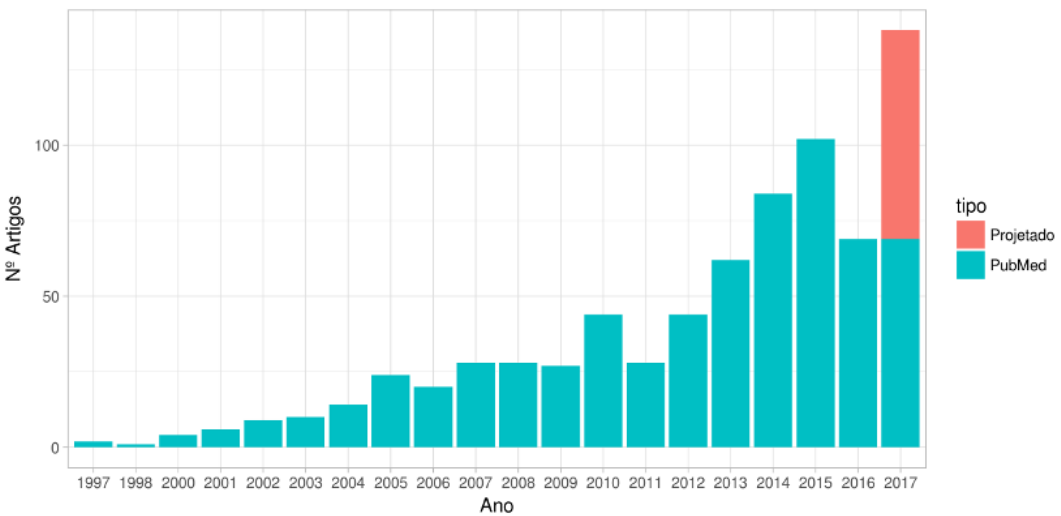
Artigos publicados Big Data no Pubmed

- Query: '("big data"[TIAB] AND ("epidemiology" OR "public health")'
- foram recuperados 400 artigos



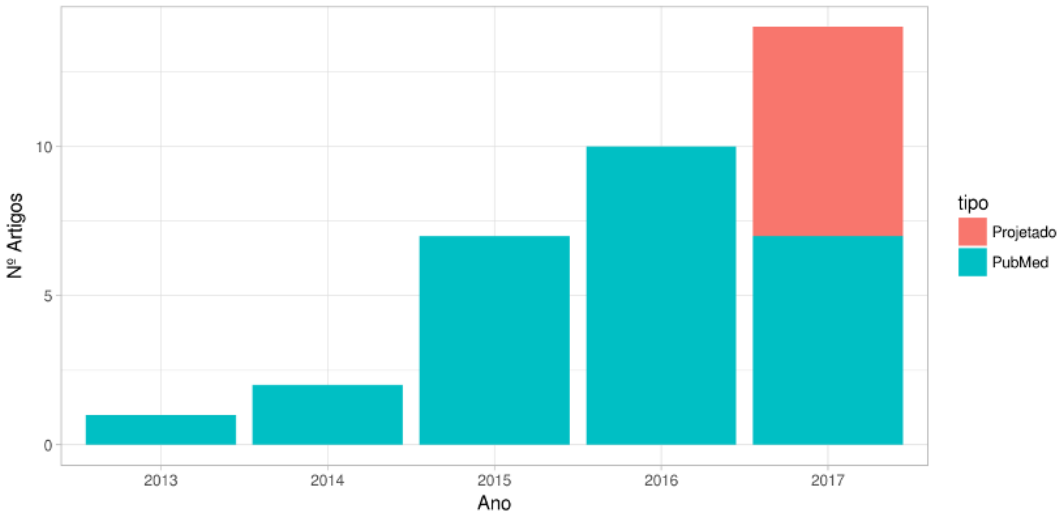
Artigos publicados Data Mining no Pubmed

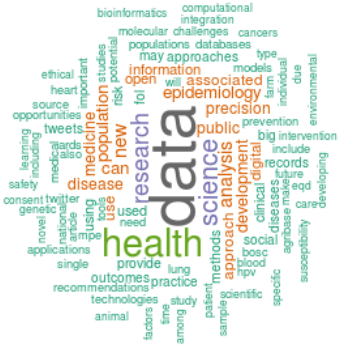
- Query: '("data mining"[TIAB] AND ("epidemiology" OR "public health")'
- foram recuperados 676 artigos



Artigos publicados Data Science no Pubmed

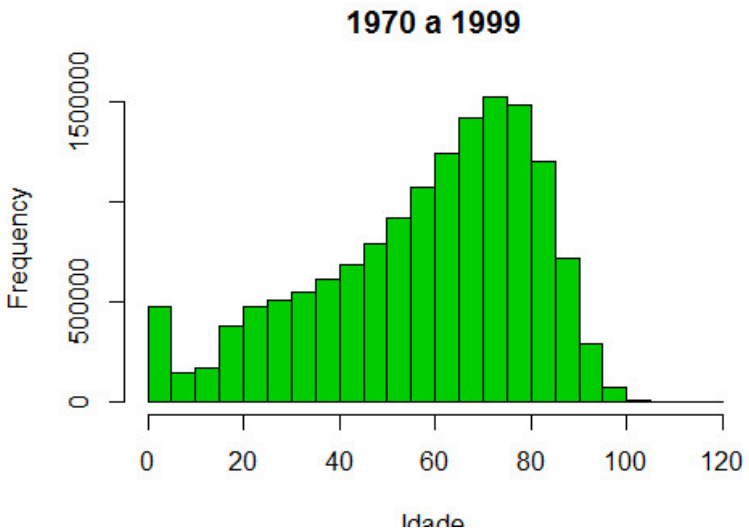
- Query: '("data science"[TIAB] AND ("epidemiology" OR "public health")'
- foram recuperados 27 artigos

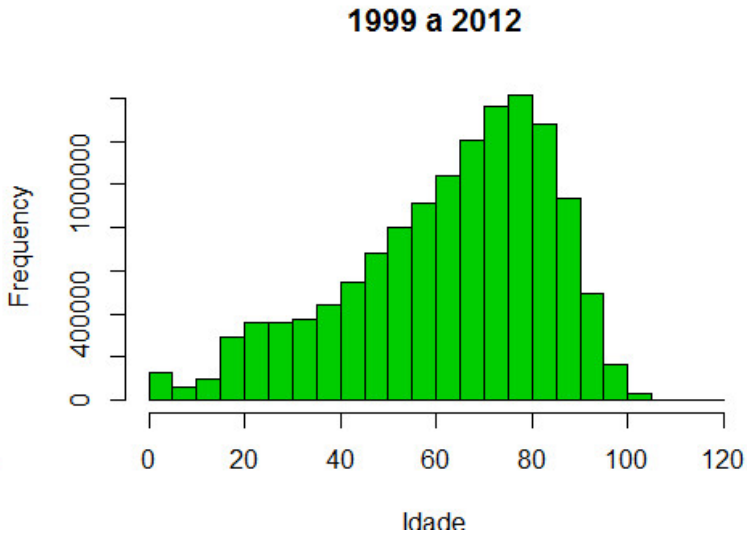




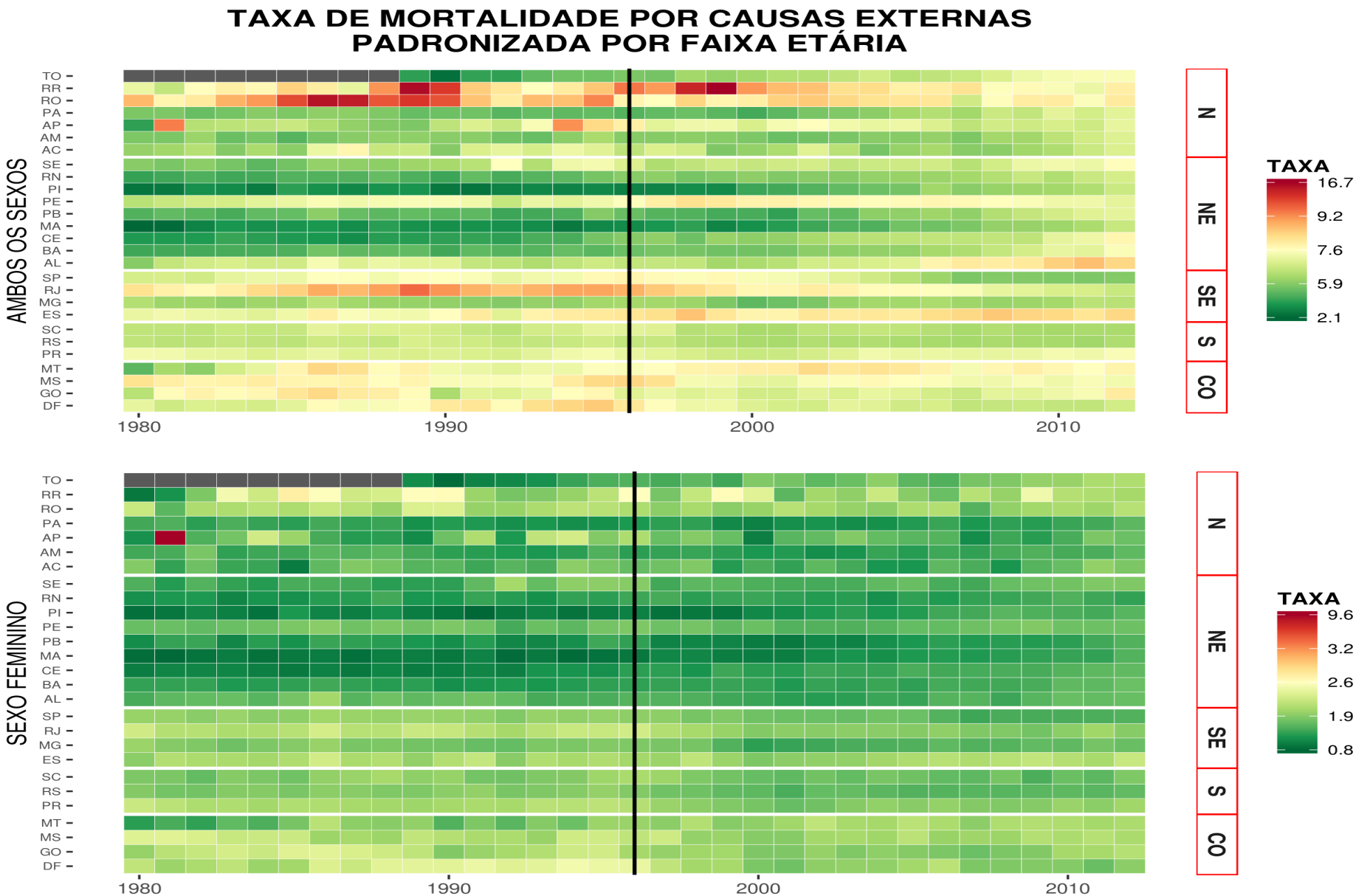
“Big Data” Mortalidade Brasil

- Autores: Davi Barroso Alves & Oswaldo G. Cruz
- Banco em postgresSQL criado a partit do SIM Brasil de 1979 a 2014
- criando em desktop core i7 com 16 GB
- Linhas: $\sim$ 35 milhões colunas:20, espaço em memoria $\sim$ 4GB





“Big Data” Mortalidade Brasil

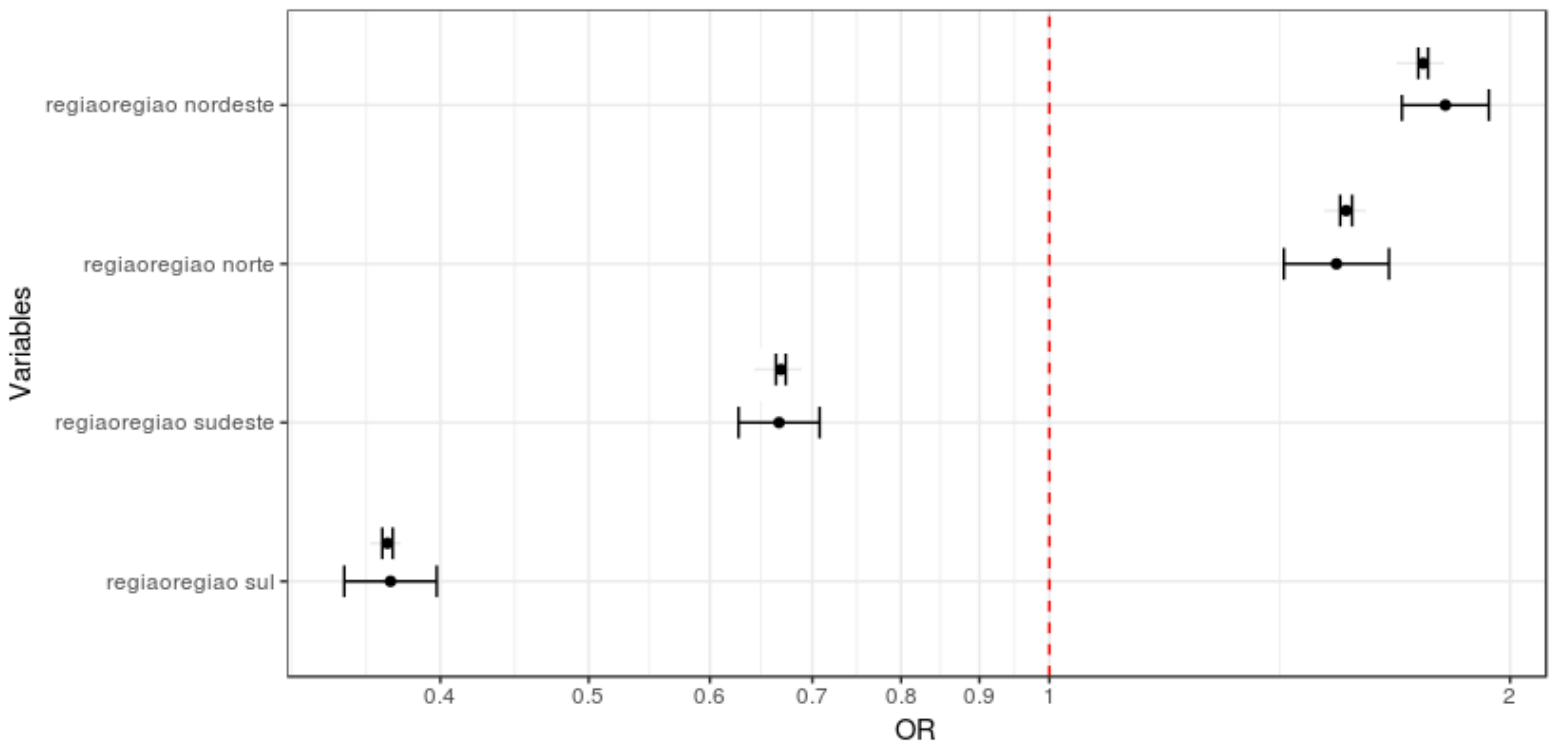


Estudo Peso ao Nascer SINAN

- Trabalho em andamento, de 3 anos de Nascidos Vivos +7 milhões
- comparando o Universo com a amostra de 1% ~ 70.000 usando uma regressão logistica

Região	OR Universo	OR Amostra 1%
NE	1.703 [1.691-1.713]	1.814 [1.670-1.937]
N	1.562 [1.550-1.574]	1.540 [1.423-1.667]
SE	0.652 [0.648-0.656]	0.665 [0.626-0.708]
S	0.368 [0.366-0.371]	0.371 [0.346-0.398]

Estudo Peso ao Nascer SINAN



Mineiração de dados do Twitter

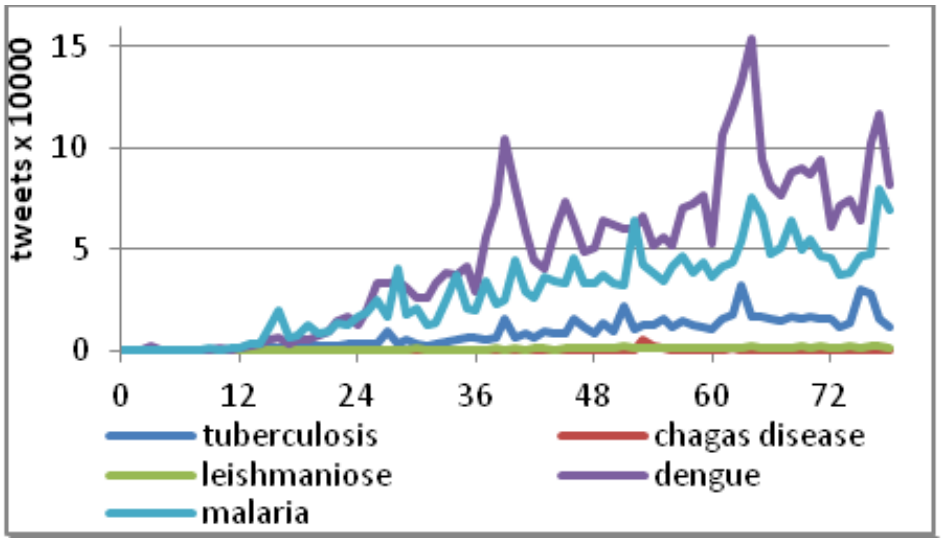
Mining for adverse drug events on twitter

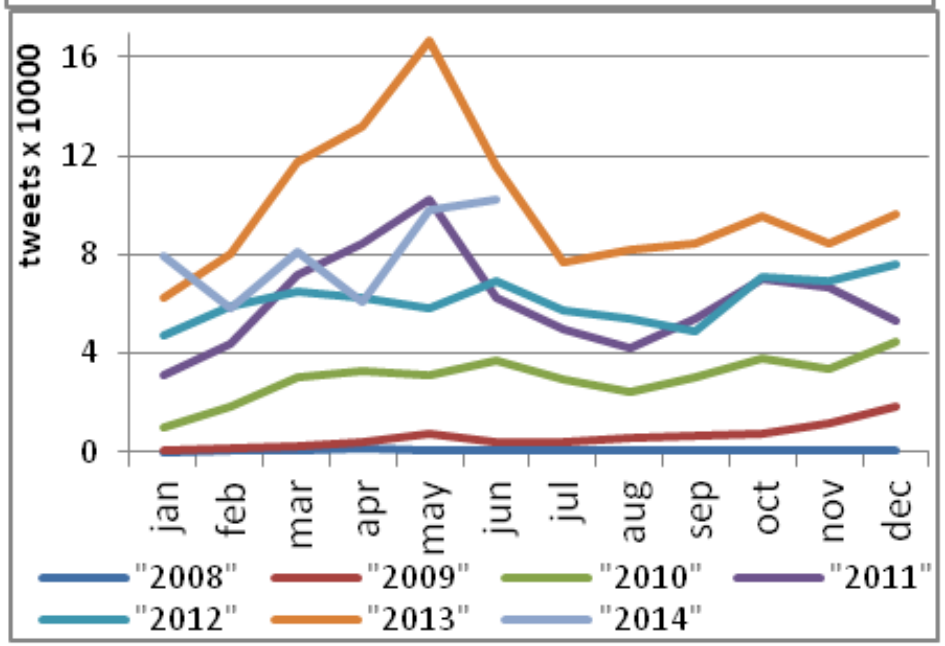
Felipe Duval[1], Ernesto Caffarena[1], Oswaldo Cruz[1] and Fabrício Silva[1]
[1]*Fundação Oswaldo Cruz - Instituto Oswaldo Cruz, Av. Brasil 4365, Rio de Janeiro, Brazil*
felipe.duval@ioc.fiocruz.br,{ernesto,oswaldo,fabs}@fiocruz.br

Keywords: Adverse event, data mining, text mining, big data, pharmacovigilance, neglected diseases, UMLS, Twitter, tweet, topsy, noSQL, disproportionality analysis, malaria, dengue, ruby, cTAKES.

Abstract: At the post marketing phase when drugs are used by large populations and for long periods of time, unexpected adverse events may occur altering the risk-benefit relation of drugs, requiring sometimes a regulatory action. These events at the post marketing phase require a significant increase in health care since they result in unnecessary damage, often fatal, to patients. Therefore, the early discovery of adverse events in the post marketing phase is a primary goal of the health system, in particular for pharmacovigilance systems. The main goal of this paper is to prove that twitter can be used as source to find new and already known adverse drug events. This proposal has a big social relevance, as it will help pharmacovigilance systems.

Mineiração de dados do Twitter





Considerações para fazer datascince & bigdata

- torma-se um datascientist é um processo longo e exige dedicação
- o uso de grandes bancos de dados exigem uma combinação de software e hardware muito diferentes do que os usuários estão acostumados a usar.
- é necessário conhecimentos de programação e bancos de Dados SQL (ou NoSQL)
- nenhum pacote estatístico por si só vai dar conta de criar , gerenciar e analisar dados
- é necessario conhecimentos aprofundados de estatistica

é muito importante invistir na capacitação em R, BD,SQL etc... e programação para usarem Data Science !

FIM

- Obrigado!

Meu Email é

oswaldo.cruz@fiocruz.br

- Para refletir...

Nate Silver

"...if the quantity of information is increasing by 2.5 quintillion bytes per day, the amount of useful information almost certainly isn't. Most of it is just noise, and the noise is increasing faster than the signal. There are so many hypotheses to test, so many data sets to mine but a relatively constant amount of objective truth."