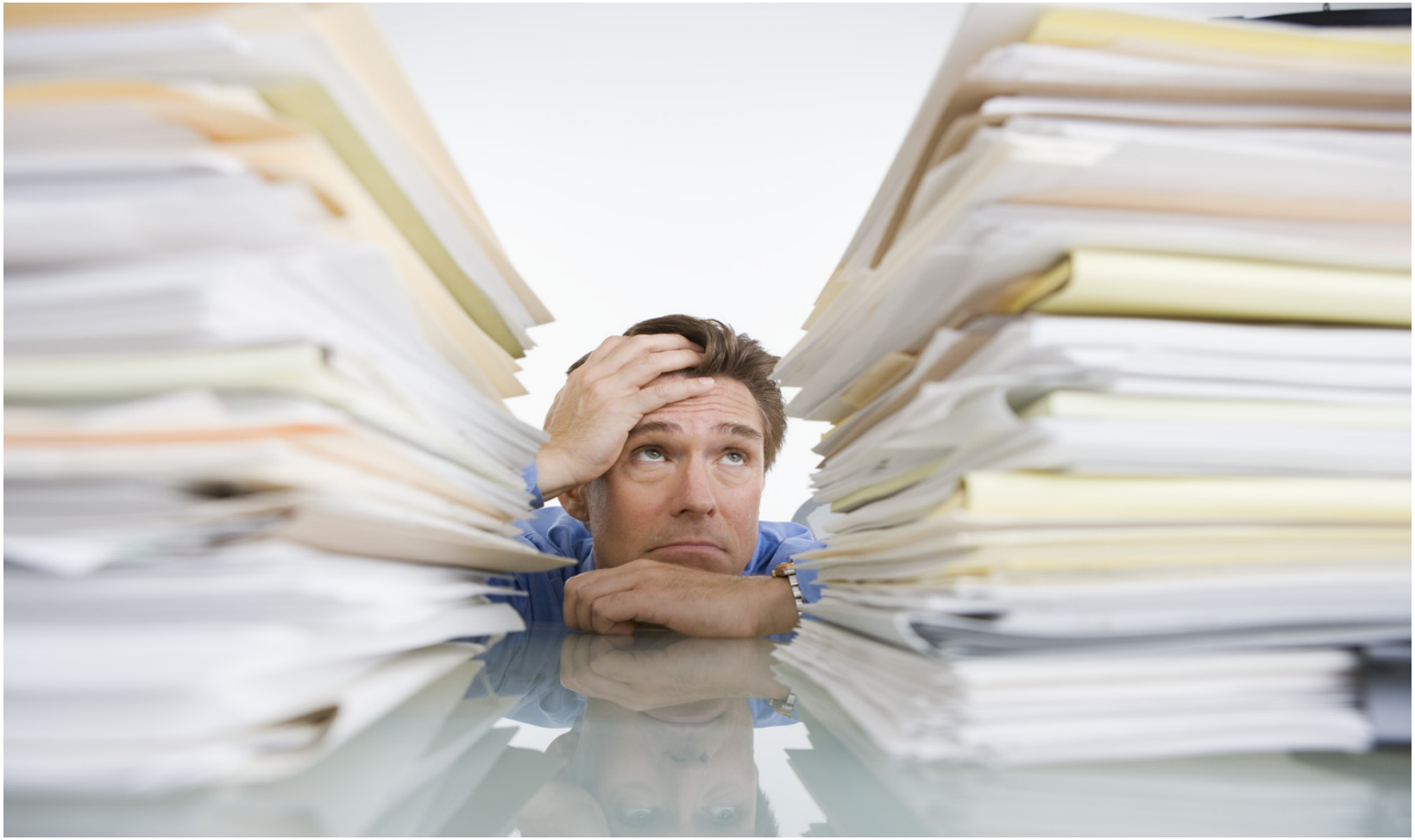
Web Data Mining com R: contexto

Fabrício Jailson Barth

Faculdade BandTec e VAGAS Tecnologia

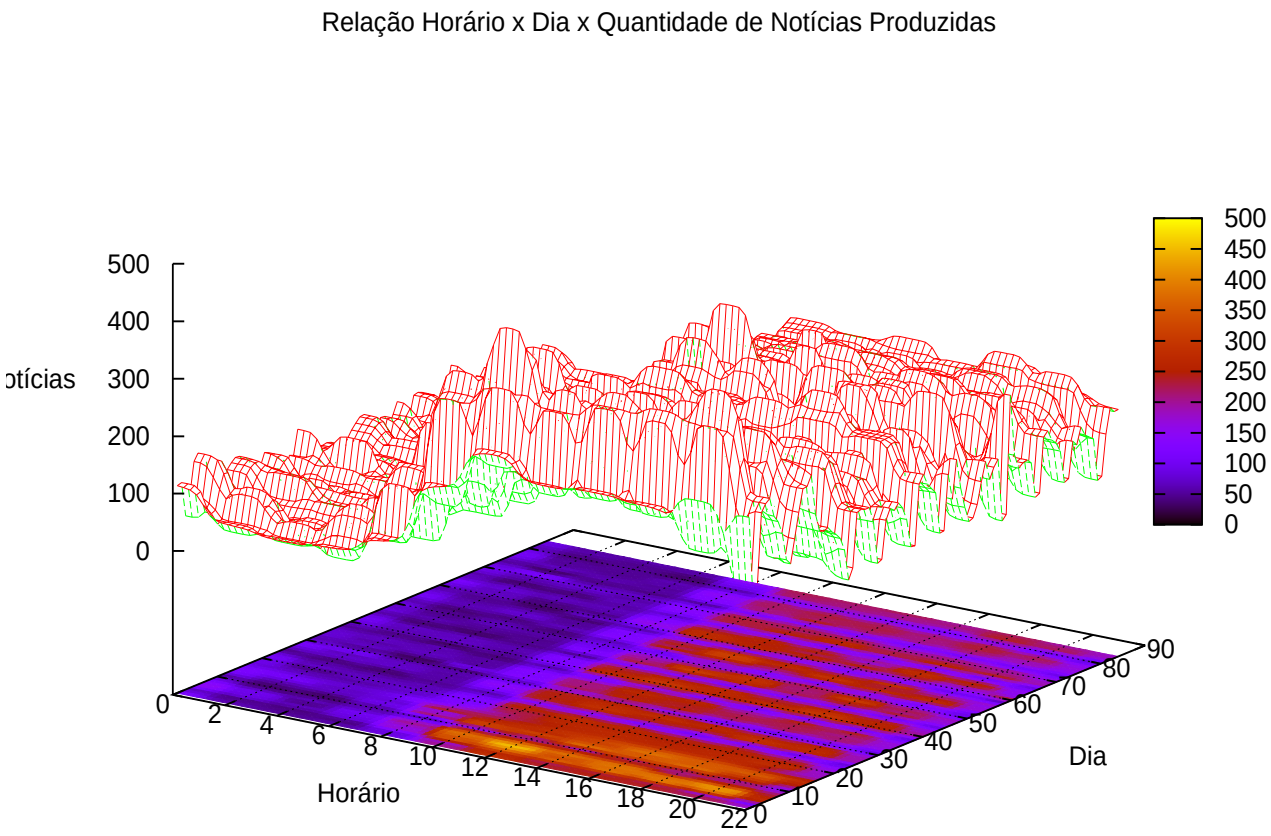
Junho de 2013

Problema



http://investingcaffeine.com/2010/01/07/tmi-the-age-of-information-overload/

Alguns dados...



Quantidade de notícias publicadas na Web por apenas seis veículos brasileiros de notícias ($D_0 = 17/07/2007$)

Dados mais atuais

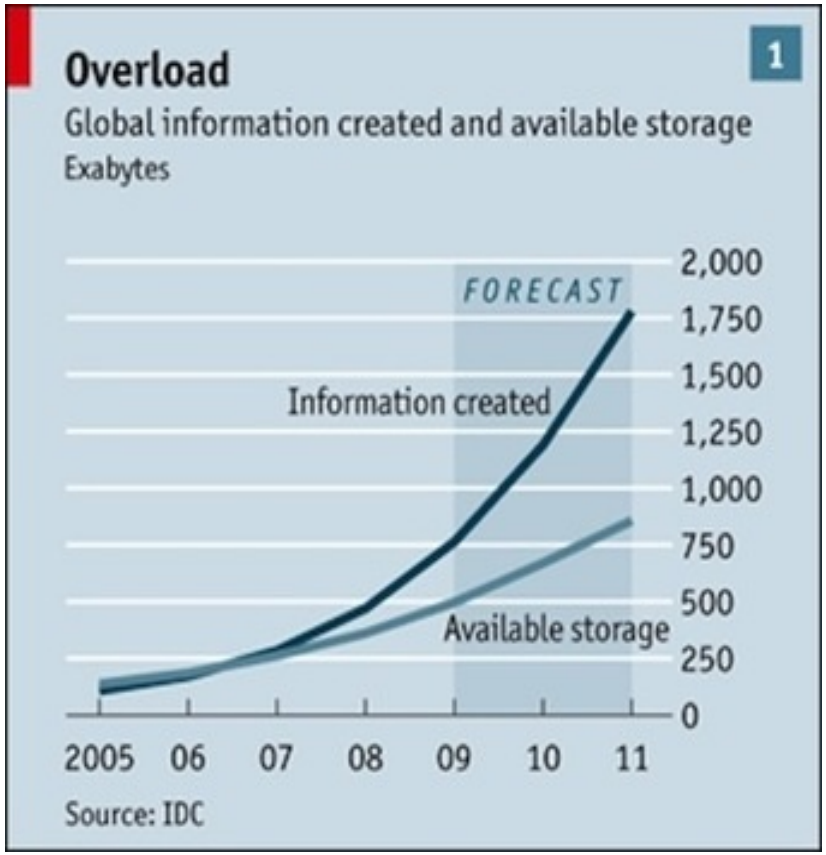
- A380: Heathrow $\rightarrow$ JFK: 640 TBs de log
- Twitter: 12+ TBs of tweet every day
- Facebook: 25+ TBs of log data every day
- Sistemas baseados em RFID
- Smartphones com GPS, acelerómetro, ...

http://www.ibmbigdatahub.com/
Mitchell. Mining our reality. Science. 2009

De onde vem os dados?

Summer School on Big Data - EMC

Big Data



"We collect an astonishing amount of digital information... ...we've long since surpassed our ability to store and process it all. Big data is here, and it's causing big problems..."[1]

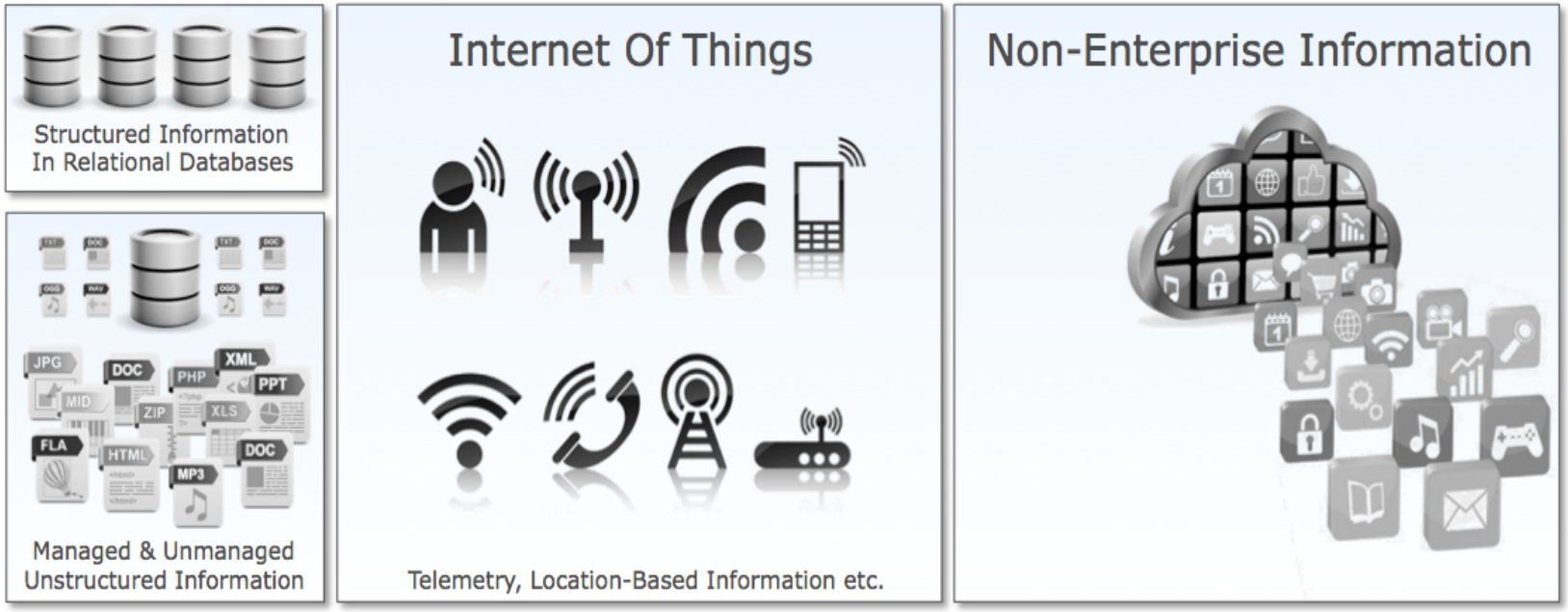
Big Data é um conceito relativo

- O que é **grande hoje**...
- Pode não ser **grande amanhã**...

Big Data refers to...

- All data that comes at high **Volume**
- All data that comes at high **Velocity**
- All data that comes from a **Variety of Sources** (structured + unstructured data)

Variety of sources



Summer School on Big Data - EMC

Por que manipular este tipo de dado?

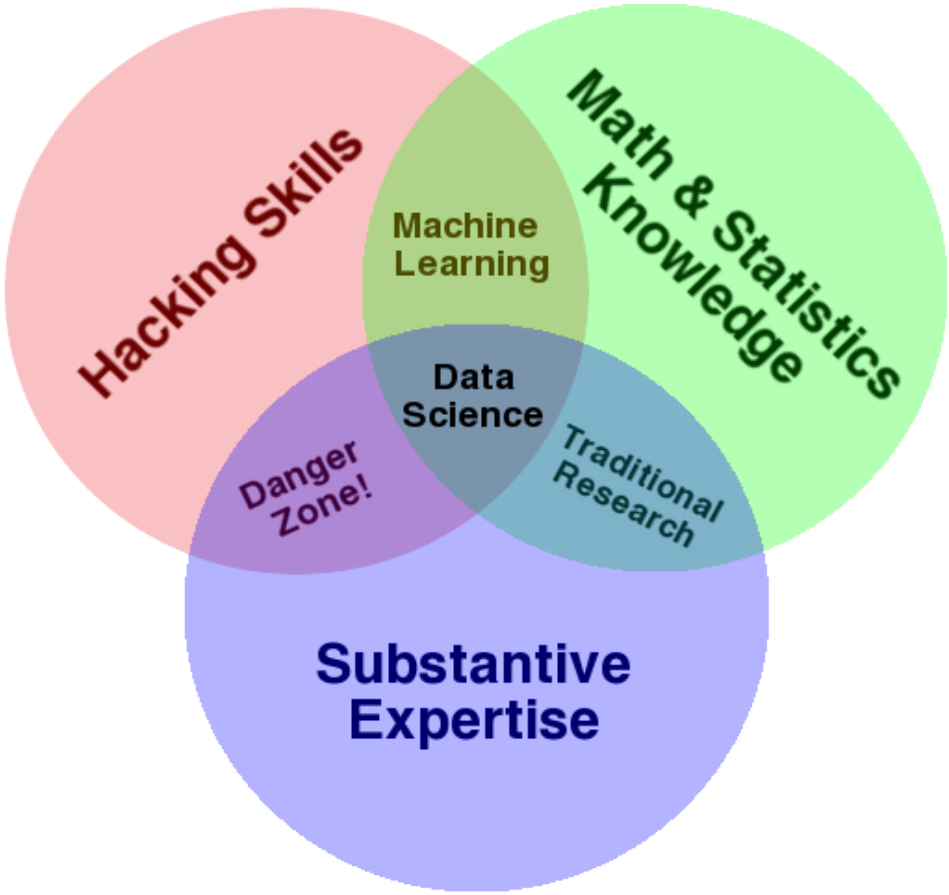
- Identificar comportamento anômalo (i.e., fraudes, falhas)
- Sumarizar tendências de publicações de artigos e patentes sobre um determinado tema.
- Sumarizar e filtrar notícias relevantes.

- Sumarizar a opinião expressa na Web sobre a sua empresa.
- Identificar padrões de navegação em sites.
- Identificar conteúdo impróprio em sites.
- Recomendação de livros, filmes, restaurantes e empregos.

Data Scientist

- *Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review.*
- **Data Scientist** applies advanced **analytical** tools and algorithms to generate **predictive insights** and **new** product **innovations** that are a direct result of the data.

Data Science Venn Diagram



http://www.drewconway.com/zia/?p=2378

Web Data Mining

A área de Web Data Mining tem como objetivo descobrir **conhecimento útil** a partir da estrutura dos **hyperlinks da Web**, **conteúdo das páginas** e **log de utilização dos sites** [2].

References

[1] Data, data everywhere. a special report on managing information. *The Economist*, pages 1–16, February 2010.

[2] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.