

# Banco de Dados

## Fundamentos

André Santanchè e Patrícia Cavoto  
Instituto de Computação - UNICAMP  
Agosto 2016

Prêmio de US\$ 1 milhão

# Exercício 1: Netflix Prize

## Recomendar um Filme

- Considerando que você vai recomendar filmes para usuários do Netflix, detalhe:
  - Que dados você levaria em consideração?
  - Que passos você seguiria para recomendar um filme para um usuário.

# Filtragem Colaborativa

- Técnica de recomendação

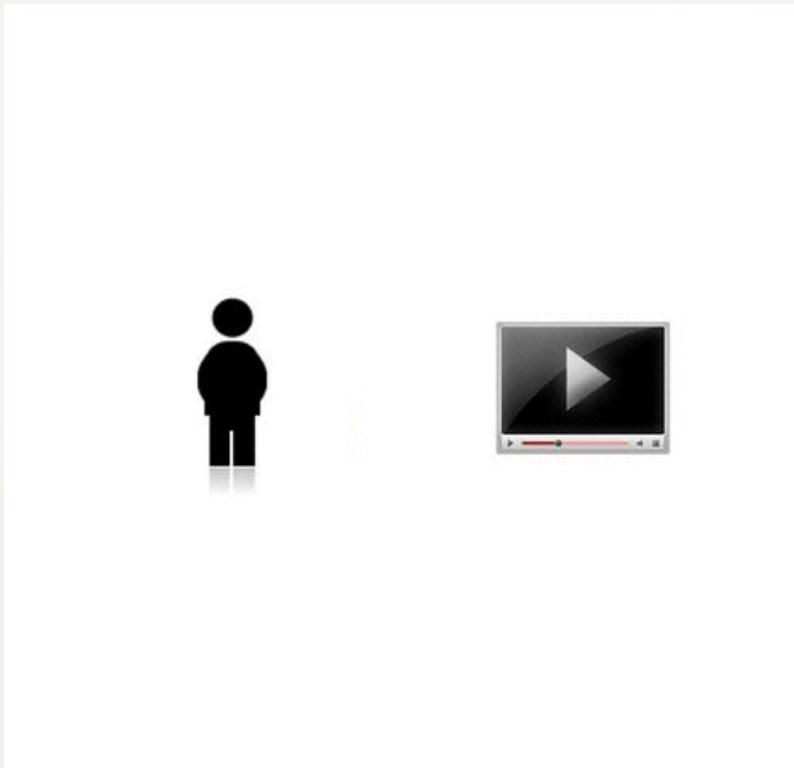
- Exemplo baseado em

[https://en.wikipedia.org/wiki/Collaborative\\_filtering](https://en.wikipedia.org/wiki/Collaborative_filtering)

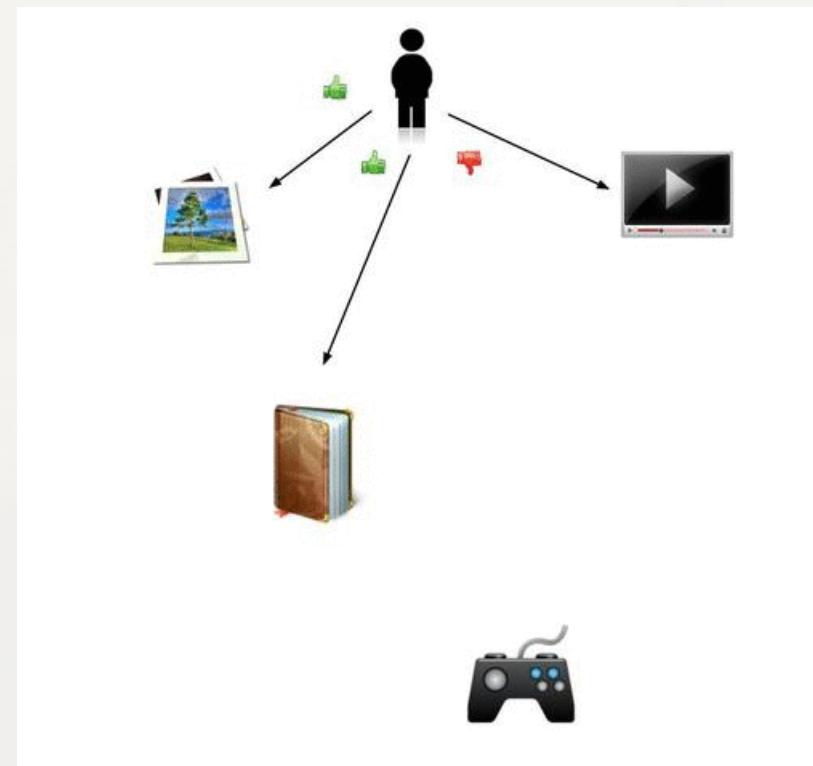
# Itens candidatos a recomendação

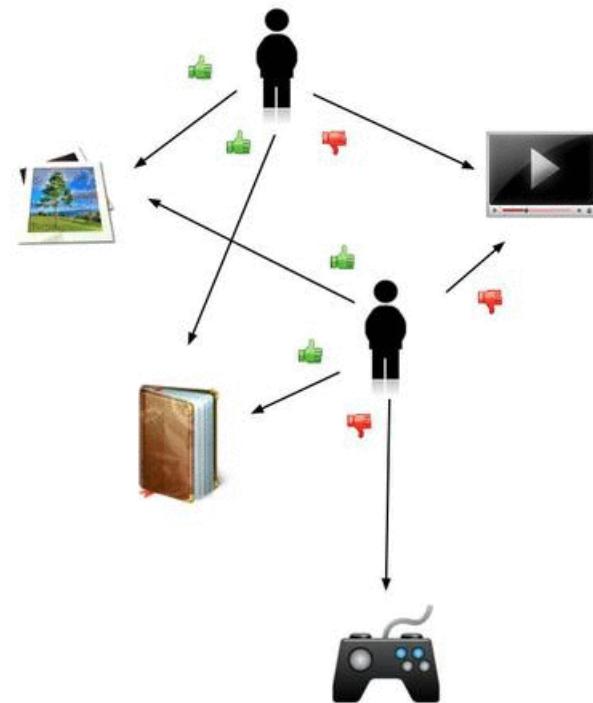


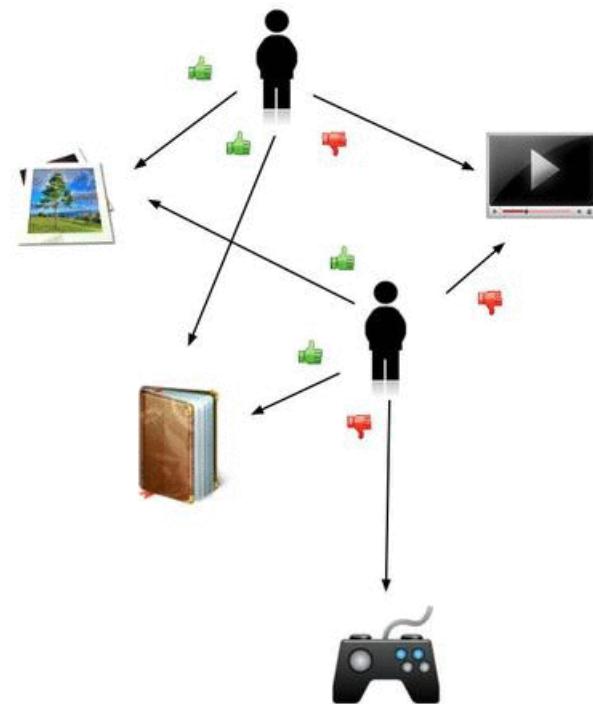
# Devo recomendar para Asdrubal?

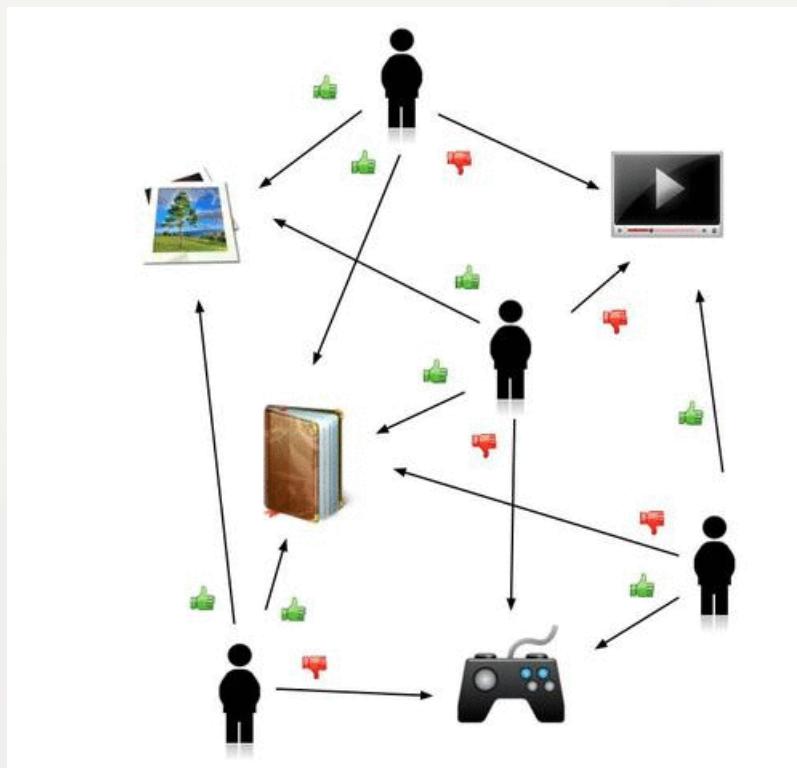


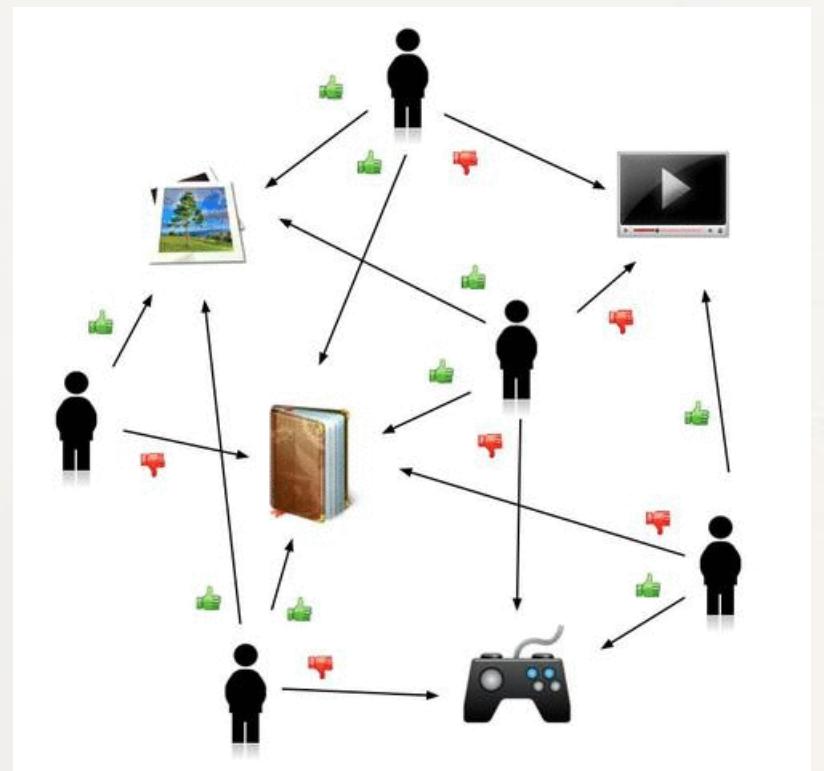
# Usuários e suas avaliações

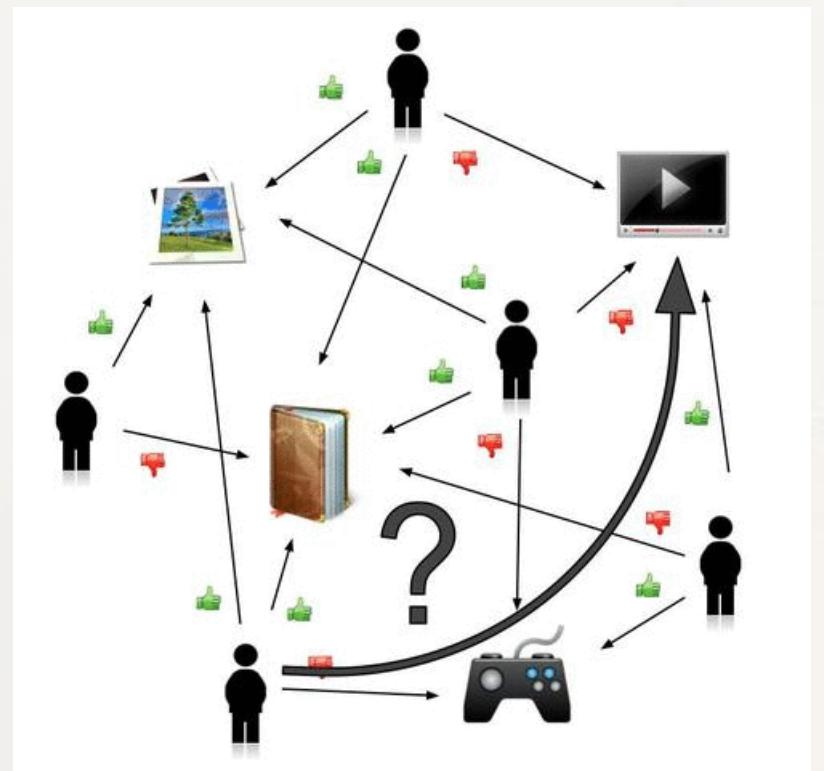




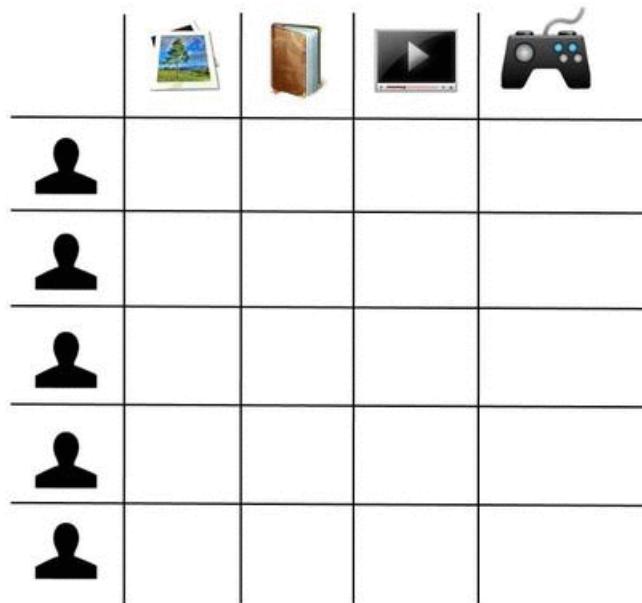








# Vendo sob outra perspectiva



# Usuários e suas avaliações





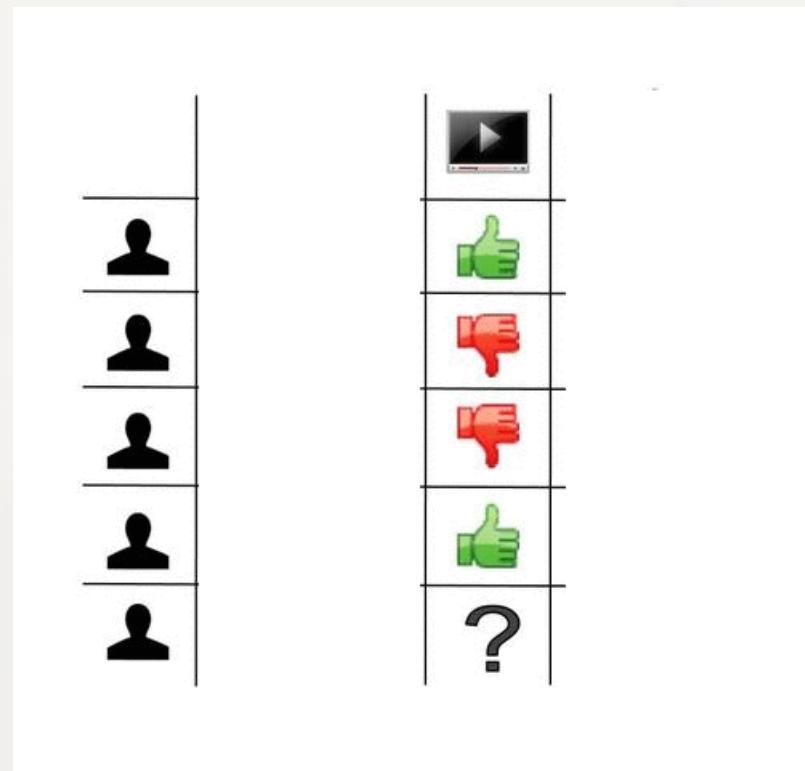


# O que recomendar para Asdrúbal?

# Olhar a avaliação dos demais?



E se os demais não pensarem como  
Asdrúbal?

# Exercício 2: Netflix Prize

## Refinando a Recomendação

- Refine a recomendação para:
  - Que dados você levaria em consideração?
  - Que passos você seguiria para recomendar um filme para um usuário.

# Selecionando Usuários Similares

## Filtragem Colaborativa

	Album	Book	Movie	Game
User 1	Like	Dislike	Like	Like
User 2		Like	Dislike	Dislike
User 3	Like	Like	Dislike	
User 4	Dislike		Like	
User 5	Like	Like	?	Dislike

# Indicando a partir dos similares

## Filtragem Colaborativa

# Indicando a partir dos similares

## Filtragem Colaborativa



# Começando com Banco de Dados

# Motivação

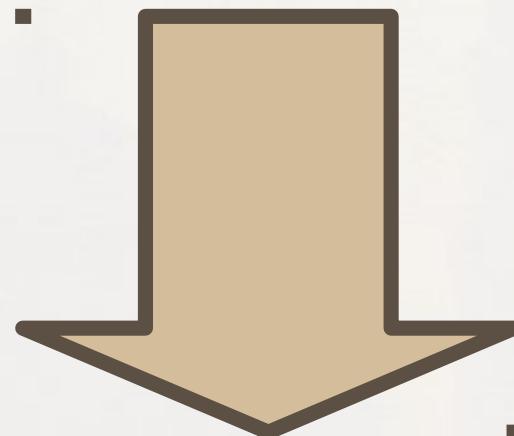
- Aplicações computacionais de todos os portes trabalham com grandes volumes de dados
  - Gerenciamento de uma rede de farmácias
  - Sistema bancário
  - Sequenciamento do Genoma Humano
  - Redes Sociais

# Motivação

- Grandes volumes de dados e suas relações complexas justificam a criação de estratégias específicas para gerenciá-los

# Motivação

- Grandes volumes de dados e suas relações complexas justificam a criação de estratégias específicas para gerenciá-los



Bancos de Dados

# Aplicações Tradicionais

- Bancos de dados numéricos e tradicionais
- Exemplos:
  - Gerenciamento de uma farmácia
  - Sistema bibliotecário
  - Sistema bancário

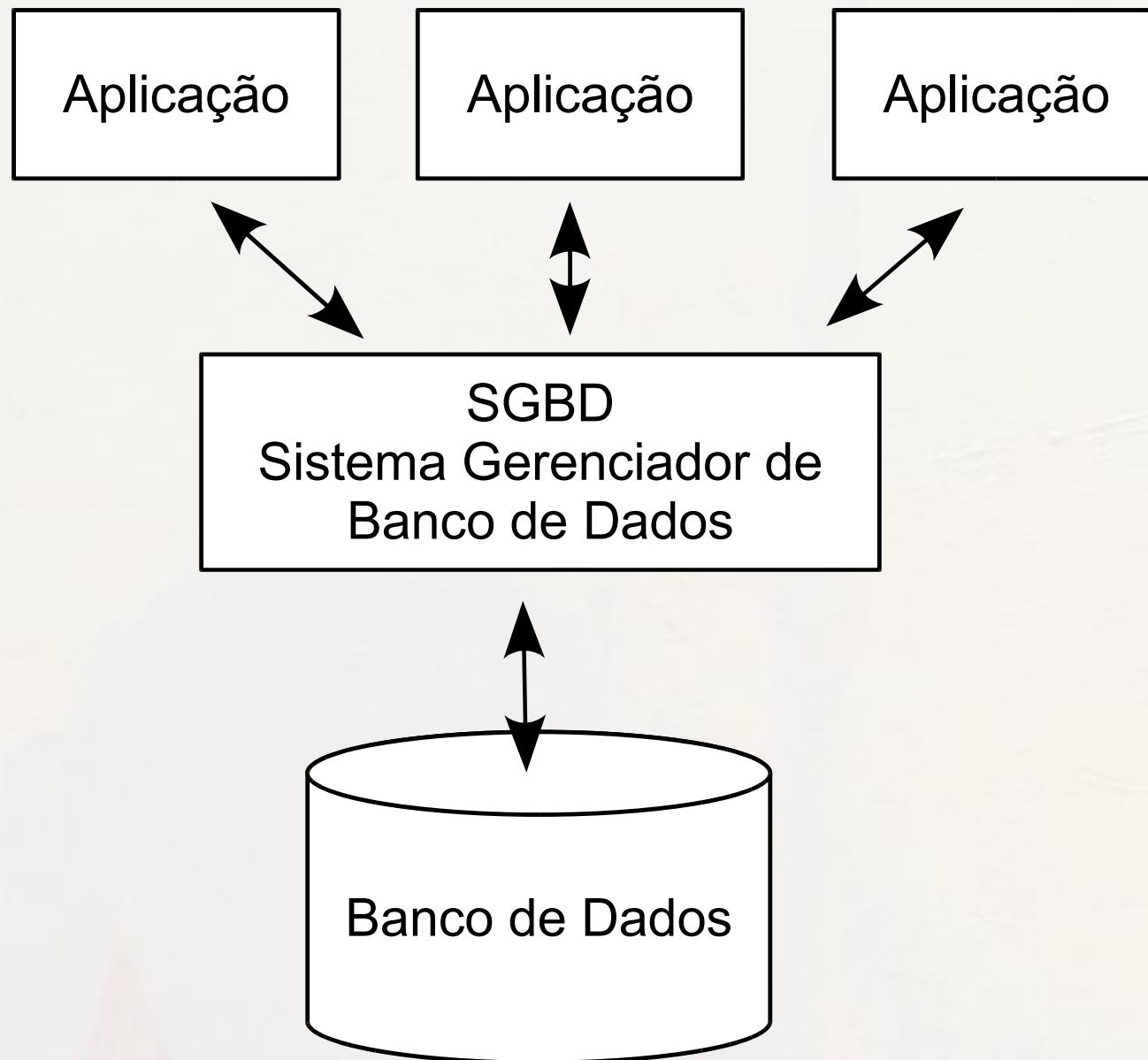
# Aplicações

- Aplicações mais recentes
  - Bancos de Dados Multimídia
  - Sistemas de Informação Geográfica (GIS)
  - Data Warehouses
  - Internet das coisas
  - etc.

# Sistema Gerenciador de Banco de Dados (SGBD)

- Sistema de software com finalidade genérica
- Projetado para a definição, construção e manipulação de bancos de dados
- Pode atender várias aplicações

# SGBD



# Vantagens de um SGBD

- Independência de dados
- Acesso eficiente
- Tempo reduzido no desenvolvimento de aplicações
- Segurança e integridade de dados
- Administração de dados uniforme
- Acesso concorrente
- Recuperação contra *crashes*

(Ramakrishnan, 2003b)

# O que está mudando?

- Dados estão por toda a parte
  - não somente centralizados em um banco
  - produzidos de forma distribuída e interligados
- Modelagem e semântica ganham importância
  - Web Semântica e ontologias
- Data deluge e Big Data
  - novas abordagens (NoSQL)
  - processamento e armazenamento descentralizados
  - bancos de dados em memória

# Data Deluge

- Genoma Humano
  - 3.3 bilhões base-pairs
- Facebook
  - 30/06/2016 - 1,71 bilhões de usuários ativos
    - <http://newsroom.fb.com/company-info/>

# Lei de Moore

- Poder de processamento dobra a cada dois anos
- Como crescem os dados?

# O que acontece na internet em 60 segundos?

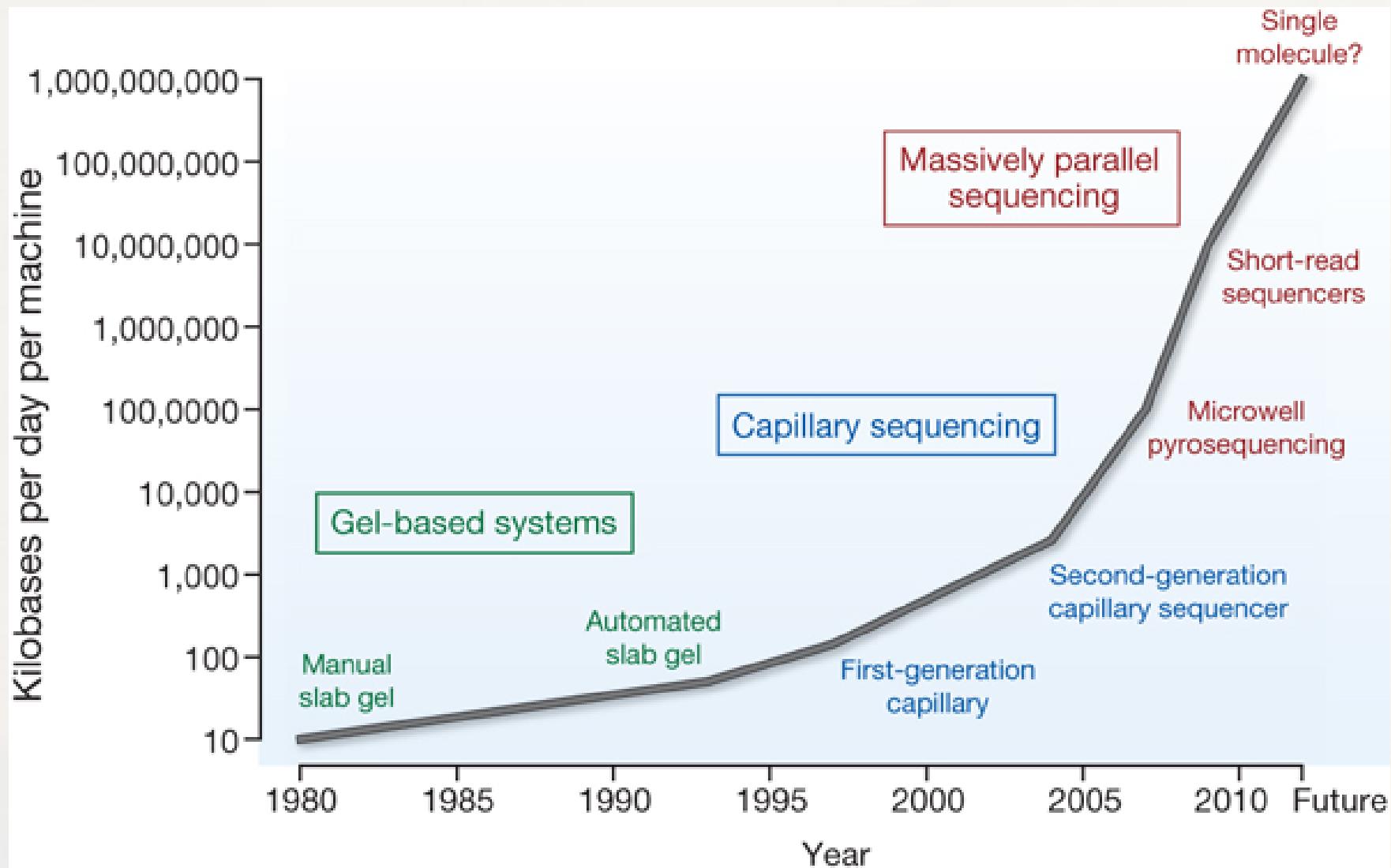


<http://www.dailymail.co.uk/sciencetech/article-2381188/Revealed-happens-just-ONE-minute-internet-216-000-photos-posted-278-000-Tweets-1-8m-Facebook-likes.html>

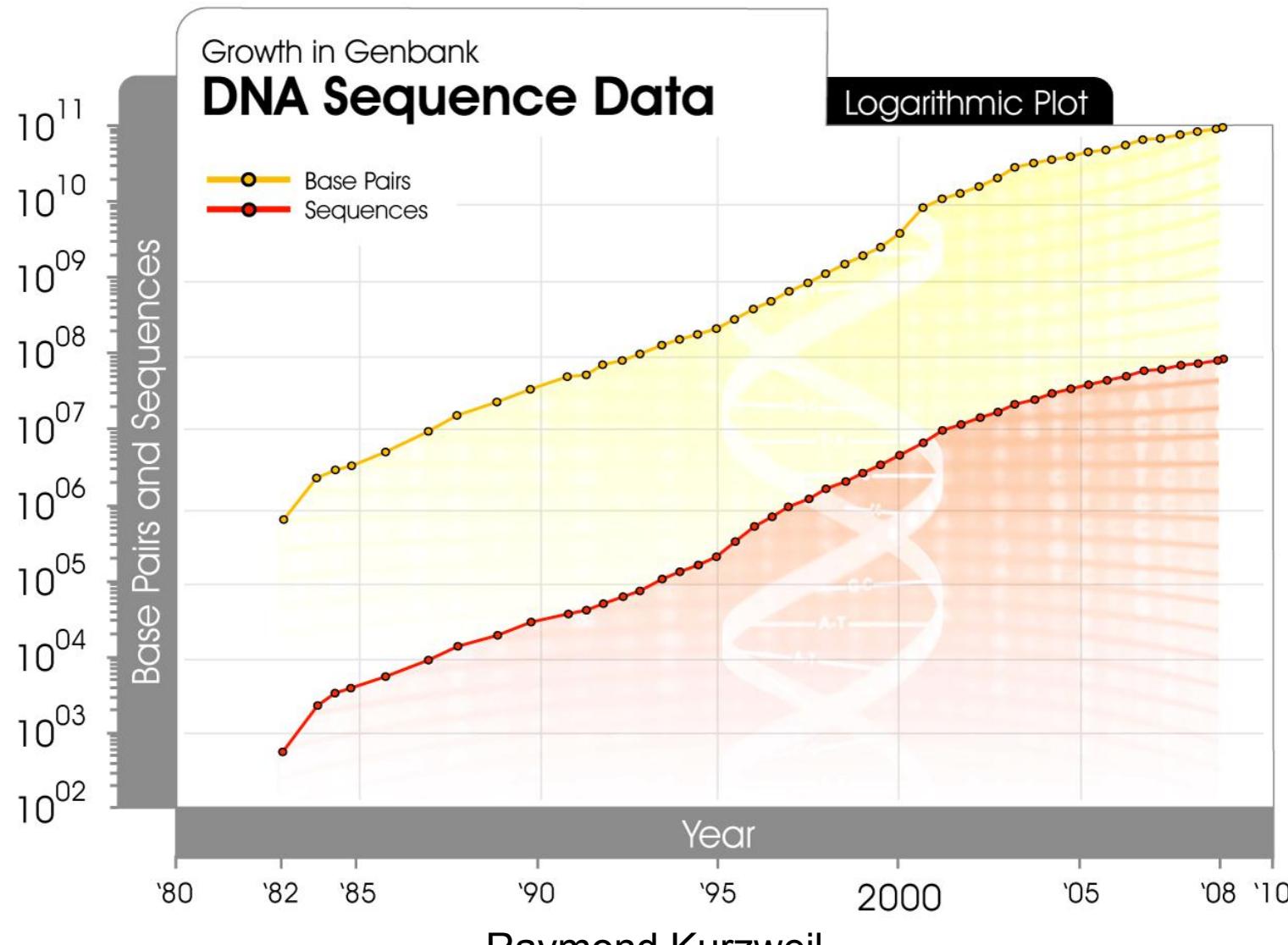
# Como crescem os dados? Sequenciamento de Genoma

# Sequenciamento de Genoma

## Aumento na Eficiência



# Sequenciamento de Genoma Volume de Dados



Raymond Kurzweil

<http://www.kurzweilai.net/dna-sequencing-data>

# Harvard Business Review

GETTING  
CONTROL OF

# BIG DATA

How vast new streams of information are changing the art of management  
PAGE 102

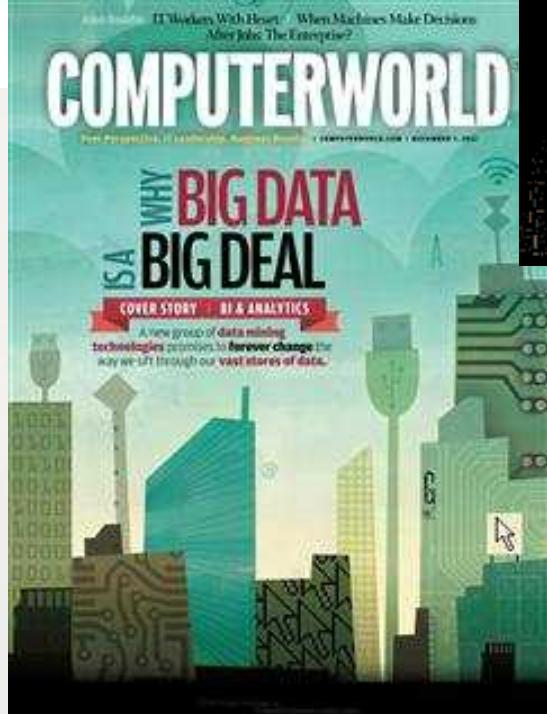


# COMPUTERWORLD

## IS A BIG DATA BIG DEAL

COVER STORY / BI & ANALYTICS

A new group of data mining technologies promises to forever change the way we sift through our vast stores of data.



OCTOBER 2010

are The Big Miss  
The True Measures  
of Success  
Michael J. Moritz

an International Business  
10 Rules for Managing  
Global Innovation  
Kenley Wilson and Peter L. Rea

are Leadership  
What Ever Happened  
To Accountability?  
Thomas G. Jackson



# The Economist

11 February 2011 | £10

Economist.com

# The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT



# nature

4 September 2008 | www.nature.com/nature | £10

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

# BIG DATA

THE BITTER BIT  
Viral infections for viruses  
TROPICAL CYCLONES  
The strong get stronger  
BLACK HOLE PHYSICS  
A new window on the Galactic Centre

NATUREJOBS  
Minnesota musings

## SCIENCE IN THE PETABYTE ERA



Obama the warrior

Misgoverning Argentina

The economic shift from West to East

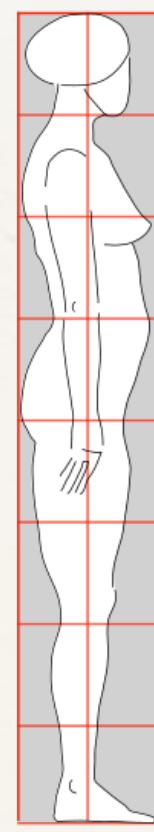
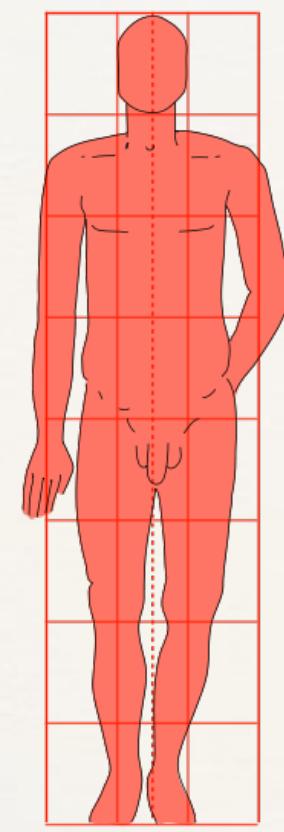
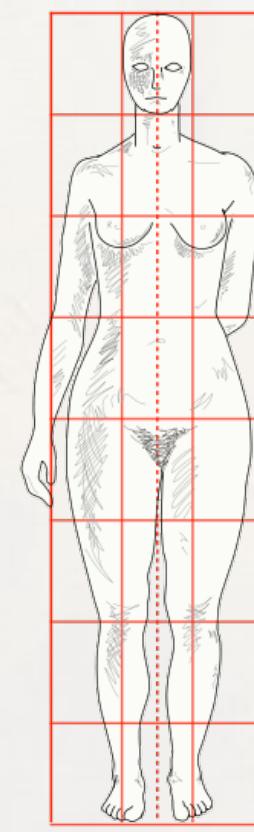
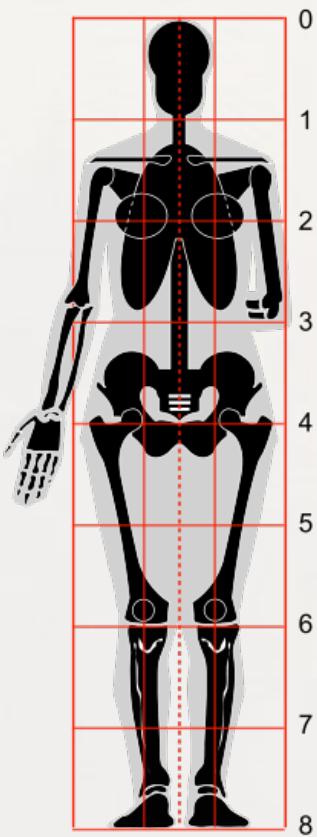
Genetically modified crops blossom

The right to eat cats and dogs

# Data Engineering & Data Science

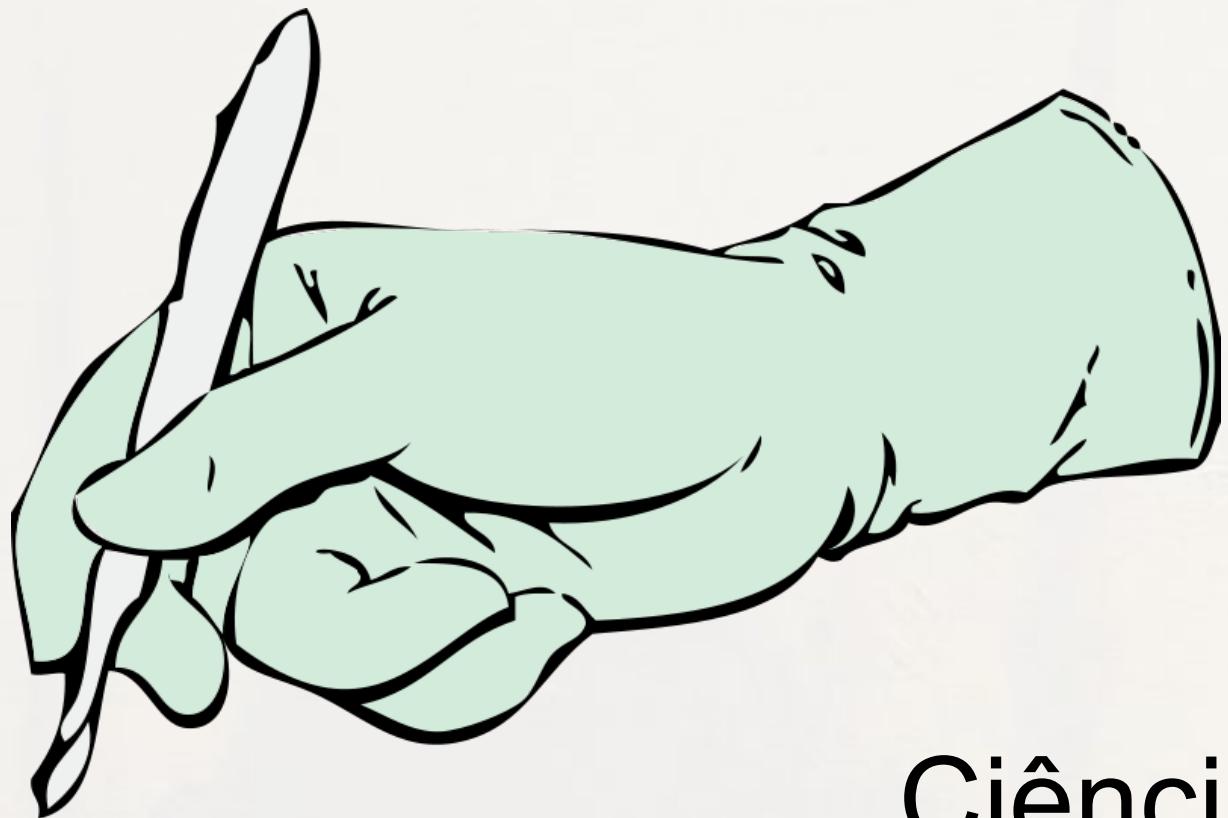
# Como Aprendemos Computação?

Cirurgia



By Original by Schnorch retracted by [LadyofHats](#)

# Cirurgia



## Ciência da Faca

# Ciência da Computação?

- Computer Science like Knife Science

(Dijkstra, 1986)

# Data Science

# Quarto Paradigma



The Fourth Paradigm: Data-Intensive Scientific Discovery  
Editado por Tony Hey, Stewart Tansley, and Kristin Tolle  
Microsoft Research  
Redmond, 2009

Mineração de Dados  
e  
Descoberta de Conhecimento

# What Wal-Mart Knows About Customers' Habits

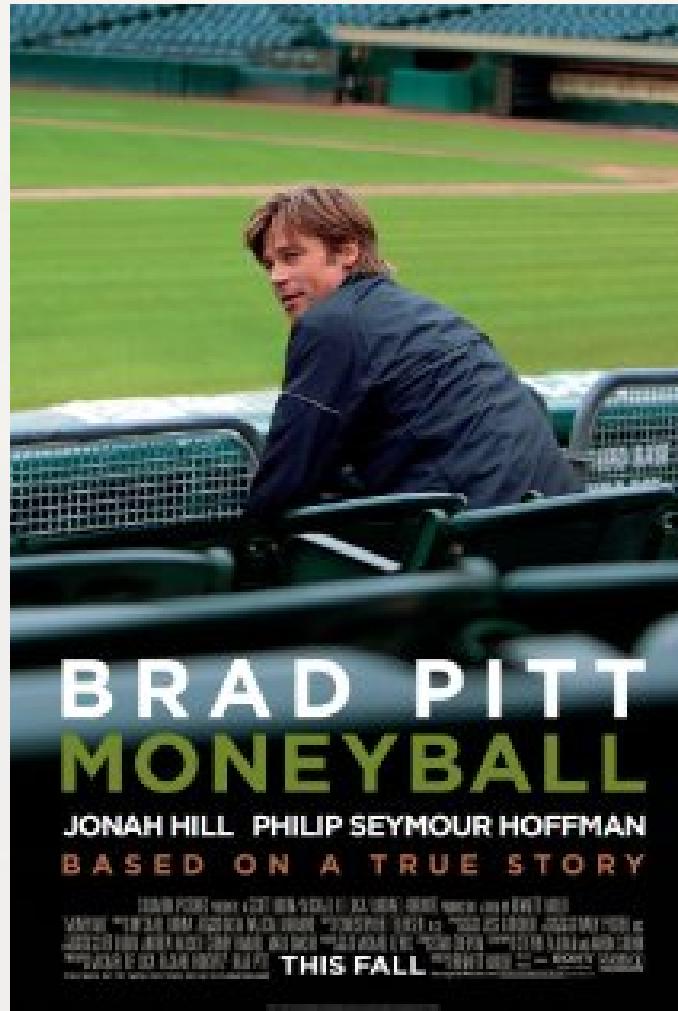
Constance L. Hays  
The New York Times, 2004

# What Wal-Mart Knows About Customers' Habits

- "start predicting what's going to happen, instead of waiting for it to happen"
- "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane"
- "And the pre-hurricane top-selling item was beer"

Linda M. Dillman - Wal Mart

# Dados e Estratégia



# Alemania e Big Data

SAP and Germany Make a Big Data Team at the World Cup

July 8, 2014 By Ben Hammonds

Sporttechie

<http://www.sporttechie.com/2014/07/08/sap-and-germany-make-smart-big-data-choices-at-world-cup/>

# Alemanha e Big Data

- SAP is using Big Data to help the German coaching staff **make smart decisions** on tactics, player fitness, scouting, preparation as well as in game management. SAP has introduced a new concept called **SAP Match Insights** that assists players and coaches to prepare themselves for upcoming matches by dissecting key situations that may present themselves throughout the course of the match.

# Basketball Analytics Database Programmer

- The Boston Celtics are seeking a Basketball Analytics Database Programmer



<http://www.nba.com/celtics/contact/bball/analytics-database-programmer>

# Futebol



Divulgação/Rio 2016/Getty Images

# Our Brand Is Crisis

by Rachel Boynton

- Documentary of the 2002 Bolivian presidential election
- Gonzalo Sánchez de Lozada x Evo Morales
- Tacts by the Greenberg Carville Shrum (GCS) company

# Minerando na Web

- Information extraction
- Mining
- Searching
- Matching
- Entity resolution
- Deep Web



# Redes Sociais e Dados sobre nós



# Facebook - Looking for Love

5 cidades dos EUA com maior percentual de pessoas solteiras:

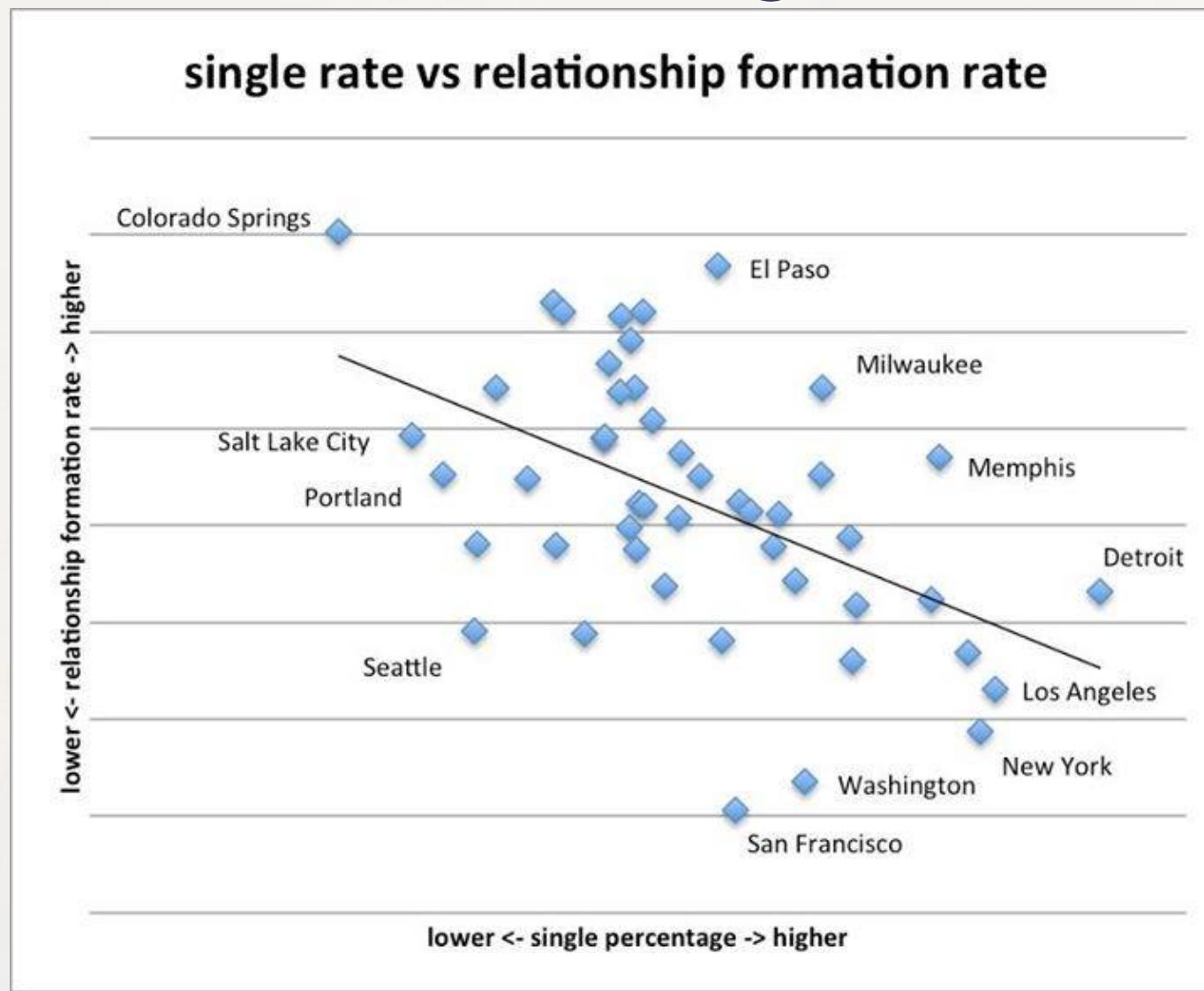
1. Detroit, MI
2. Los Angeles, CA
3. New York, NY
4. Miami, FL
5. Memphis, TN

# Facebook - Looking for Love

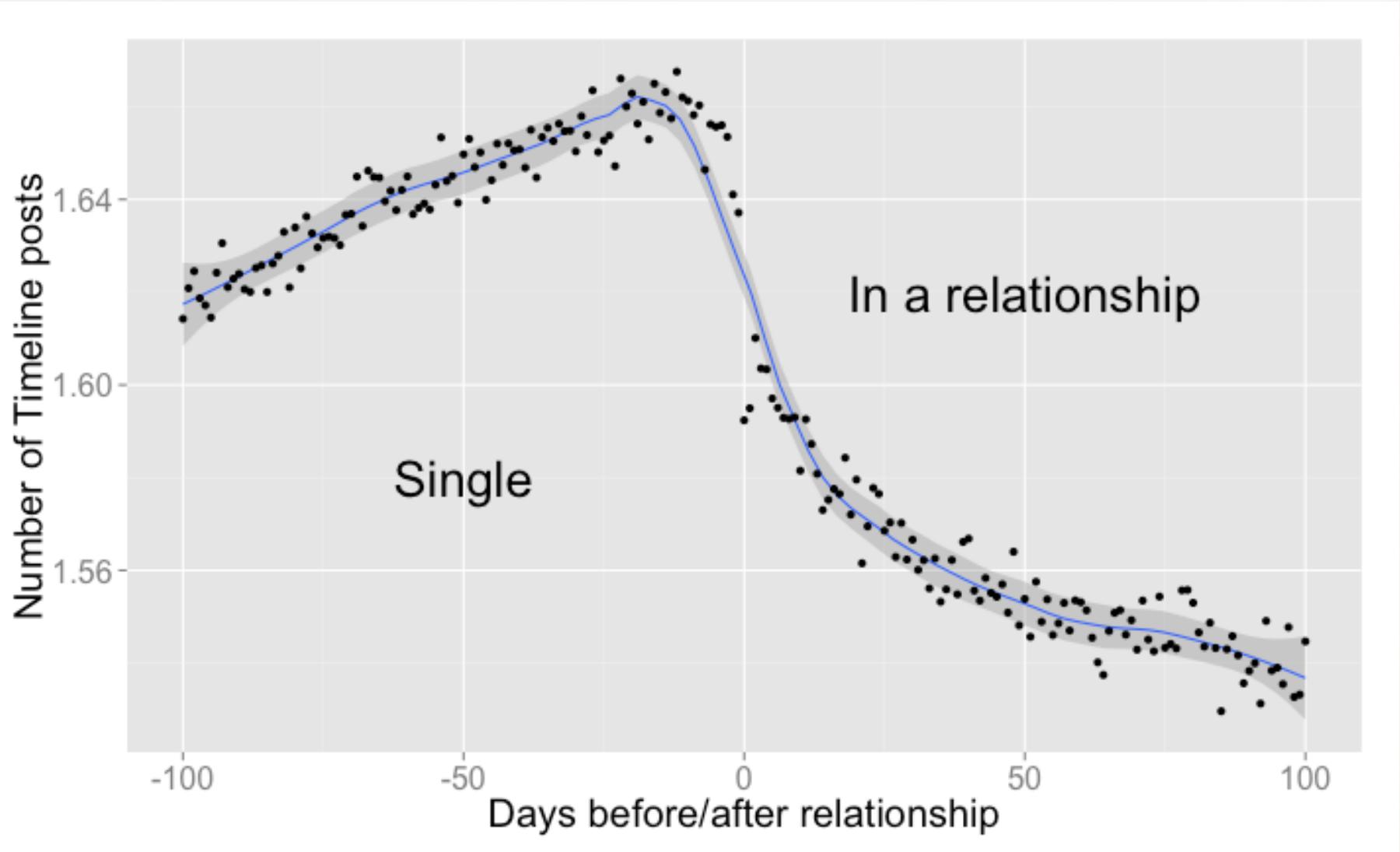
5 cidades dos EUA com maior probabilidade de formar relacionamentos duradouros:

1. Colorado Springs, CO
2. El Paso, TX
3. Louisville, KY
4. Fort Worth, TX
5. San Antonio, TX

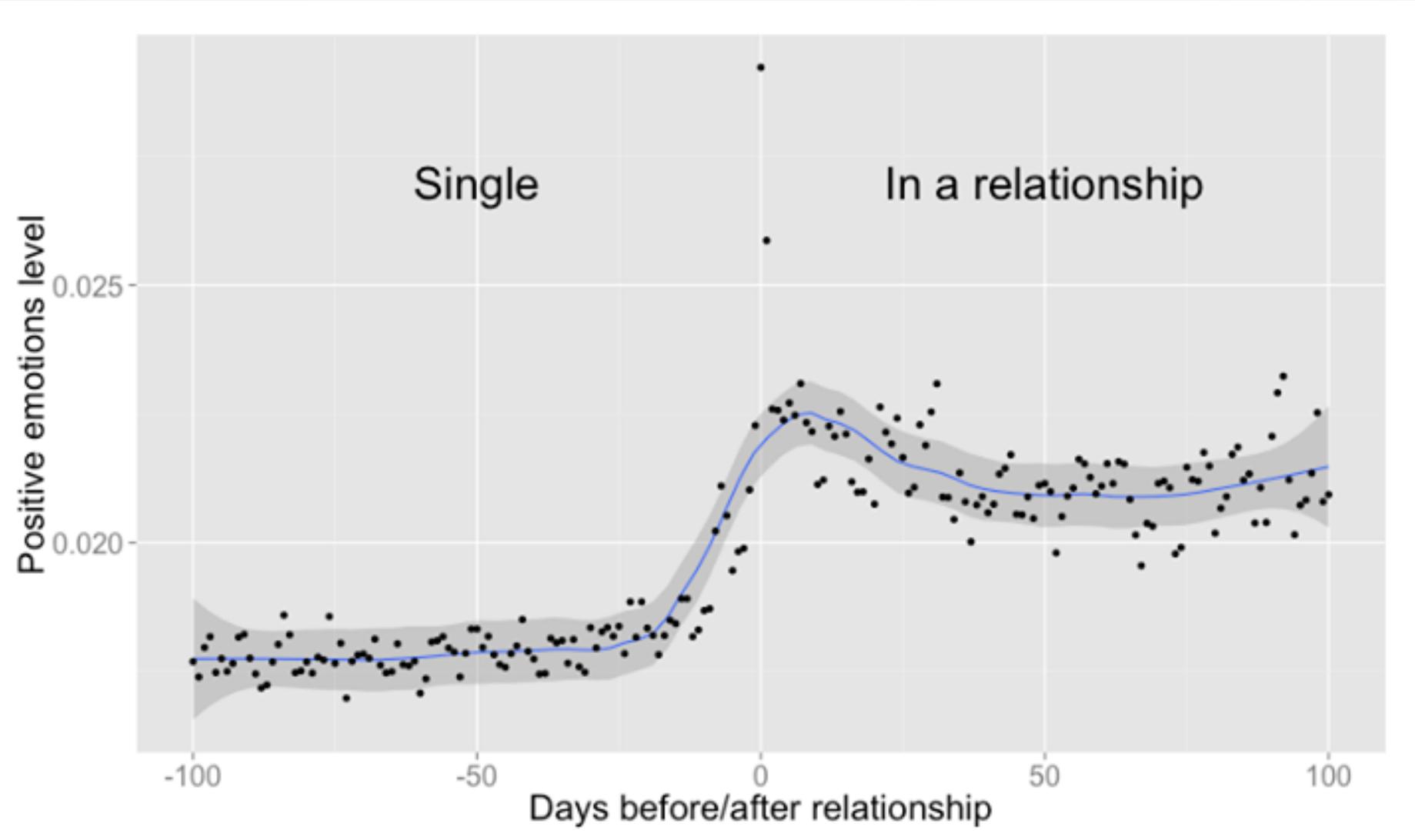
# Facebook - Looking for Love



# Facebook



# Facebook



(Diuk, 2014)

# The Formation of Love

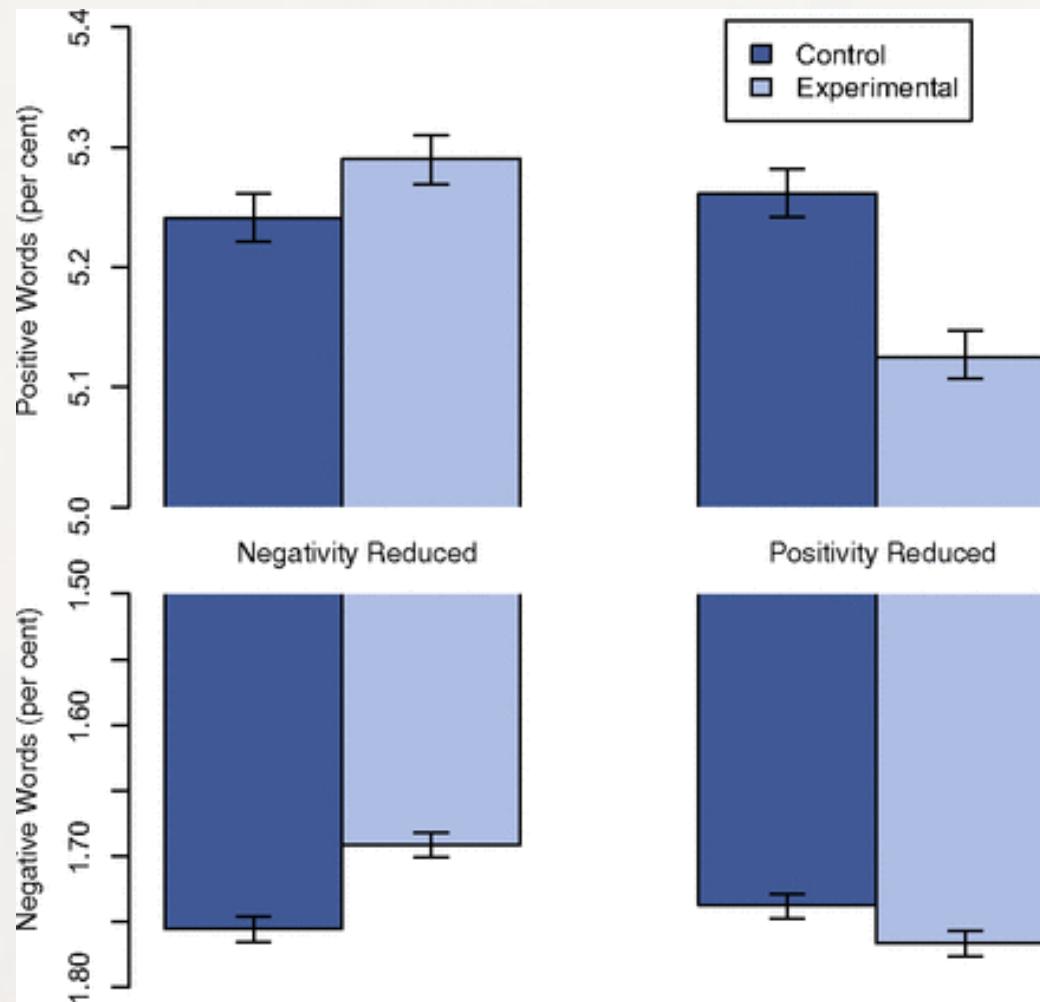
The Formation of Love

By Carlos Greg Diuk on Friday, February 14, 2014  
at 3:59pm

by Carlos Diuk, Facebook Data Science

<https://www.facebook.com/notes/facebook-data-science/the-formation-of-love/10152064609253859>

# Massive-scale Emotional Contagion



(Adam et al., 2014)

# Massive-scale Emotional Contagion

**Experimental evidence of massive-scale emotional contagion through social networks**

By Adam D. I. Kramer (Facebook), Jamie E. Guillory (Cornell), and Jeffrey T. Hancock (Cornell)

Proceedings of the National Academy of Sciences of the United States of America (PNAS)

June 17, 2014 , vol. 111 no. 24

# Minerando a Web & Saúde

# Mining Data for Better Medicine

- Mining Data for Better Medicine

Monday, September 19, 2011

By Neil Savage

<http://www.technologyreview.com/news/425466/mining-data-for-better-medicine/>

- “The health battles of millions, recorded digitally, open a world of virtual research.”

# e-Patient

- When Dave deBronkart learned he had a rare and terminal cancer, he turned to a group of fellow patients online – and found the medical treatment that saved his life.
- [https://www.ted.com/talks/dave\\_debronkart\\_meet\\_e\\_patient\\_dave](https://www.ted.com/talks/dave_debronkart_meet_e_patient_dave)

**TED** Watch Discover Attend Participate About Search...

Dave deBronkart:

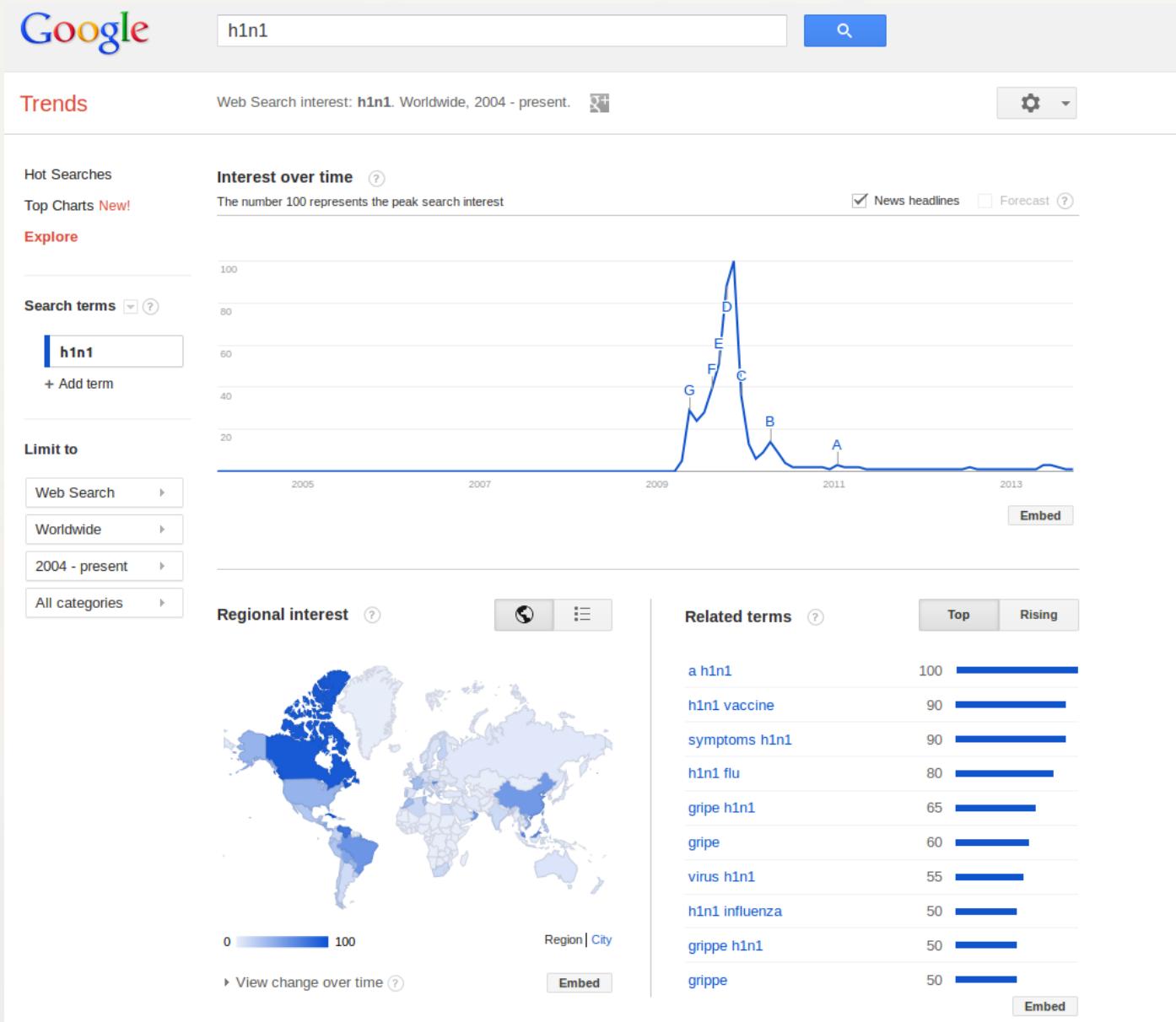
## Meet e-Patient Dave

TEDxMaastricht · 16:31 · Filmed Apr 2011  
Subtitles available in 26 languages

 View interactive transcript



# Google Trends



# Web Observatory

Área Restrita | English Português

# InWeb

Instituto Nacional de Ciência e Tecnologia para a Web

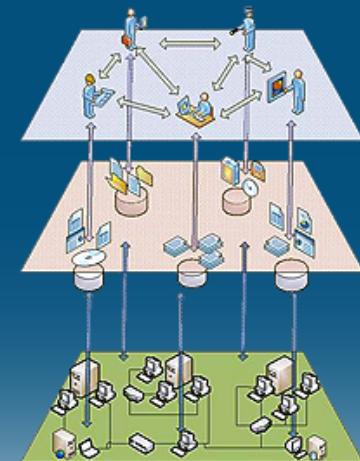
O InWeb Linhas de Pesquisa Projetos Publicações Equipe Eventos Contato blog

## Web em 3 Camadas

O InWeb vê a Web como um sistema composto de múltiplas camadas de redes complexas dinâmicas e interdependentes, pelas quais a informação flui e é disseminada. A pesquisa do InWeb está foca nas camadas de interação, serviços e infra-estrutura.

[ + ]

1 2 3



» Home

**Observatório da Dengue na Mídia**  
28/07/2011 | Notícias | Sem comentários  
Parceria do InWeb com o Instituto Nacional da Dengue, nos últimos dias o

Enquete

Nenhuma enquete aberta no momento.

NewScientist Tech

Twitter to track dengue fever outbreaks in Brazil

# Web Observatory

- INCT INWeb

- <http://observatorio.inweb.org.br/>
- Elections Observatory
- Brasileirão Observatory
- Dengue Observatory

# Recomendação



André Santanchè



See On **Pinterest**

# Linked Data

# Wikipedia

Firefox ▾ W Paris - Wikipedia, the free encyclopedia + en.wikipedia.org/wiki/Paris

Article Talk Read View source View history Search

## Paris

From Wikipedia, the free encyclopedia

This article is about the capital of France. For other uses, see [Paris \(disambiguation\)](#).

**Paris** (English /pærɪs/, /pɛrɪs/; French: [paʁi] (listen)) is the capital and most populous city of France. It is situated on the River Seine, in the north of the country, at the heart of the Île-de-France region. Within its administrative limits (the 20 arrondissements), the city had 2,234,105 inhabitants in 2009 while its metropolitan area is one of the largest population centres in Europe with more than 12 million inhabitants.

**Country** France  
**Region** Île-de-France  
**Department** Paris  
**Subdivisions** 20 arrondissements  
**Government**

- **Mayor (2008–14)** Bertrand Delanoë (PS)

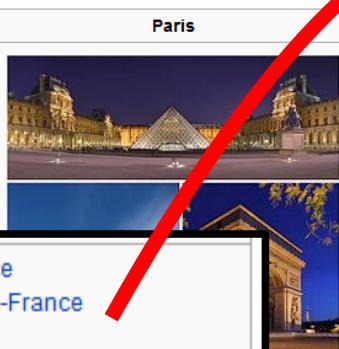
**Area<sup>[1]</sup>**

- **Urban (2010)** 2,844.8 km<sup>2</sup>  
(1,098.4 sq mi)
- **Metro (2010)** 17,174.4 km<sup>2</sup>  
(6,631.1 sq mi)
- **Land<sup>1</sup>** 105.4 km<sup>2</sup> (40.7 sq mi)

**Population (2010)<sup>[5]</sup>**

- **Rank** 1st in France

Coordinates: 48°51'24"N 2°21'03"E



WIKIPEDIA The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia

Interaction Help About Wikipedia Community portal Recent changes Contact page

Tools Print/export

Firefox ▾ W Île-de-France - Wikipedia, the free encyclopedia + en.wikipedia.org/wiki/Île-de-France\_(region)

Article Talk Read Edit View history Search

## Île-de-France

From Wikipedia, the free encyclopedia (Redirected from [Île-de-France \(region\)](#))

For other uses, see [Île-de-France \(disambiguation\)](#).

**Île-de-France** (French)

**Country** France  
**Prefecture** Paris  
**Departments** 8

**Regions**

- Paris
- Essonne
- Hauts-de-Seine
- Seine-Saint-Denis
- Seine-et-Marne
- Val-de-Marne
- Val-d'Oise
- Yvelines

**Government**

- **President** Jean-Paul Huchon (PS)

**Area**

- **Total** 12,012 km<sup>2</sup> (4,638 sq mi)

**Population (2012)<sup>[1]</sup>**

- **Total** 11,914,812
- **Density** 990/km<sup>2</sup> (2,600/sq mi)

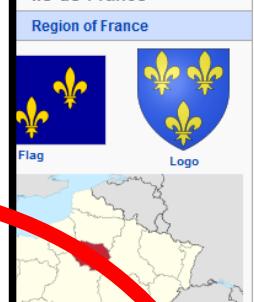
Coordinates: 48°30'N 2°30'E

WIKIPEDIA The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia

Interaction Help About Wikipedia Community portal Recent changes Contact page

Tools Print/export



Île-de-France

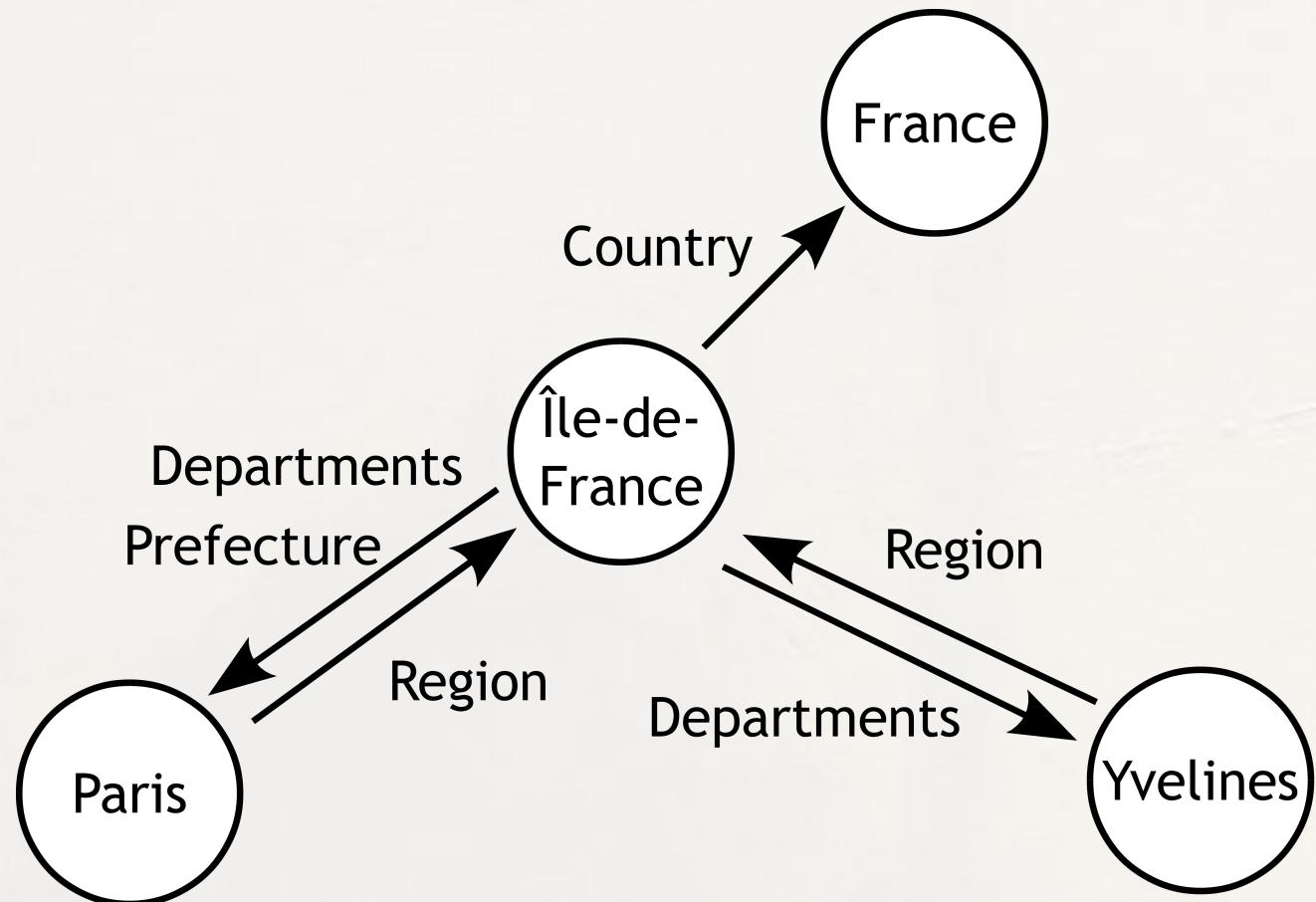
Region of France

Flag Logo

Map of the Île-de-France region.

# Infobox

# DBpedia



# DBpedia - English

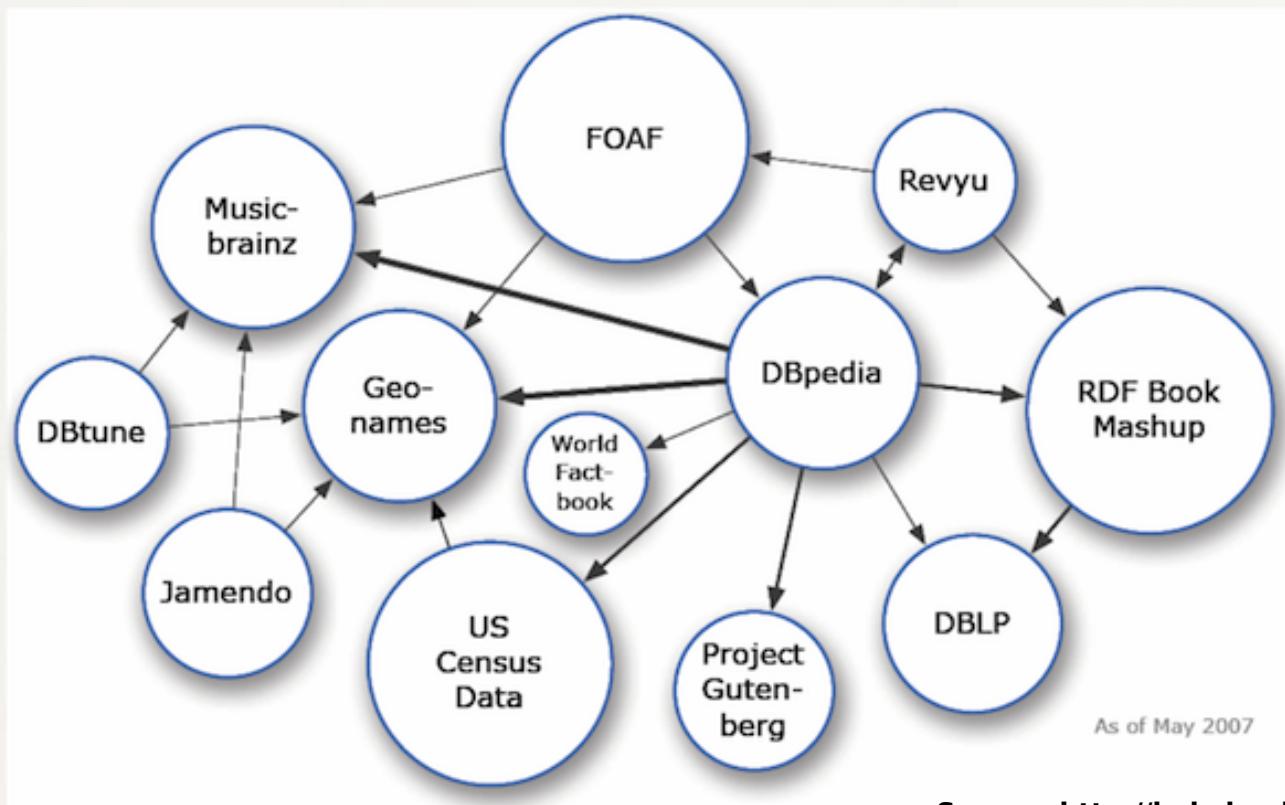
- **4 million things**
- **3.22 million classified in a consistent ontology**
  - 832,000 persons
  - 639,000 places (427,000 populated)
  - 372,000 creative works
    - 116,000 music albums; 78,000 films; 18,500 video games
  - 209,000 organizations
  - 226,000 species
  - 5,600 diseases.

# DBpedia - International

- 119 languages
- 24.9 million things
- 16.8 million interlinked with English
- 12.6 million unique things

# Linked Data

## 05/2007

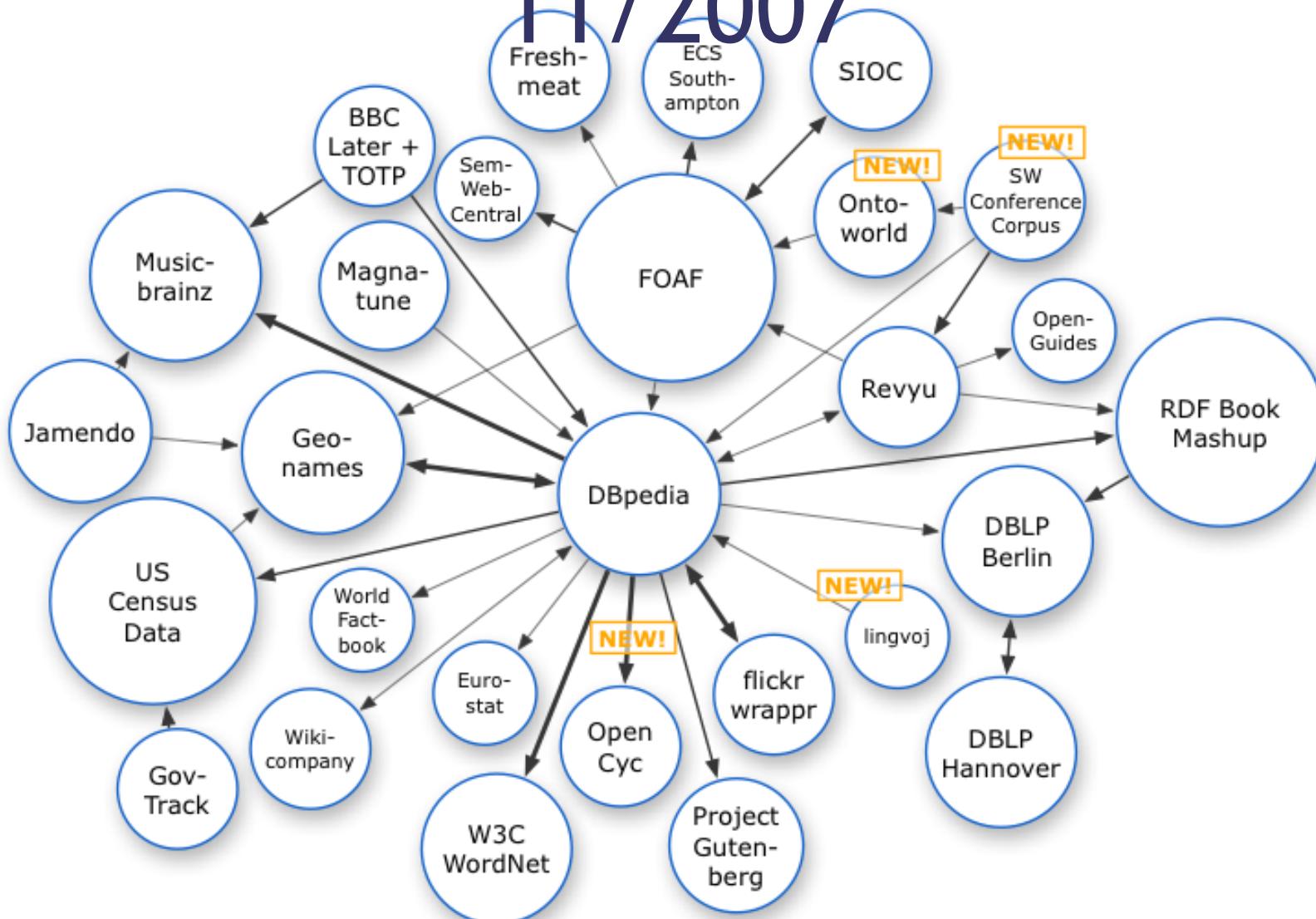


Source: <http://lod-cloud.net/>

Datasets published following Linked Data 'format':  
**05/2007**

# Linked Data

11/2007

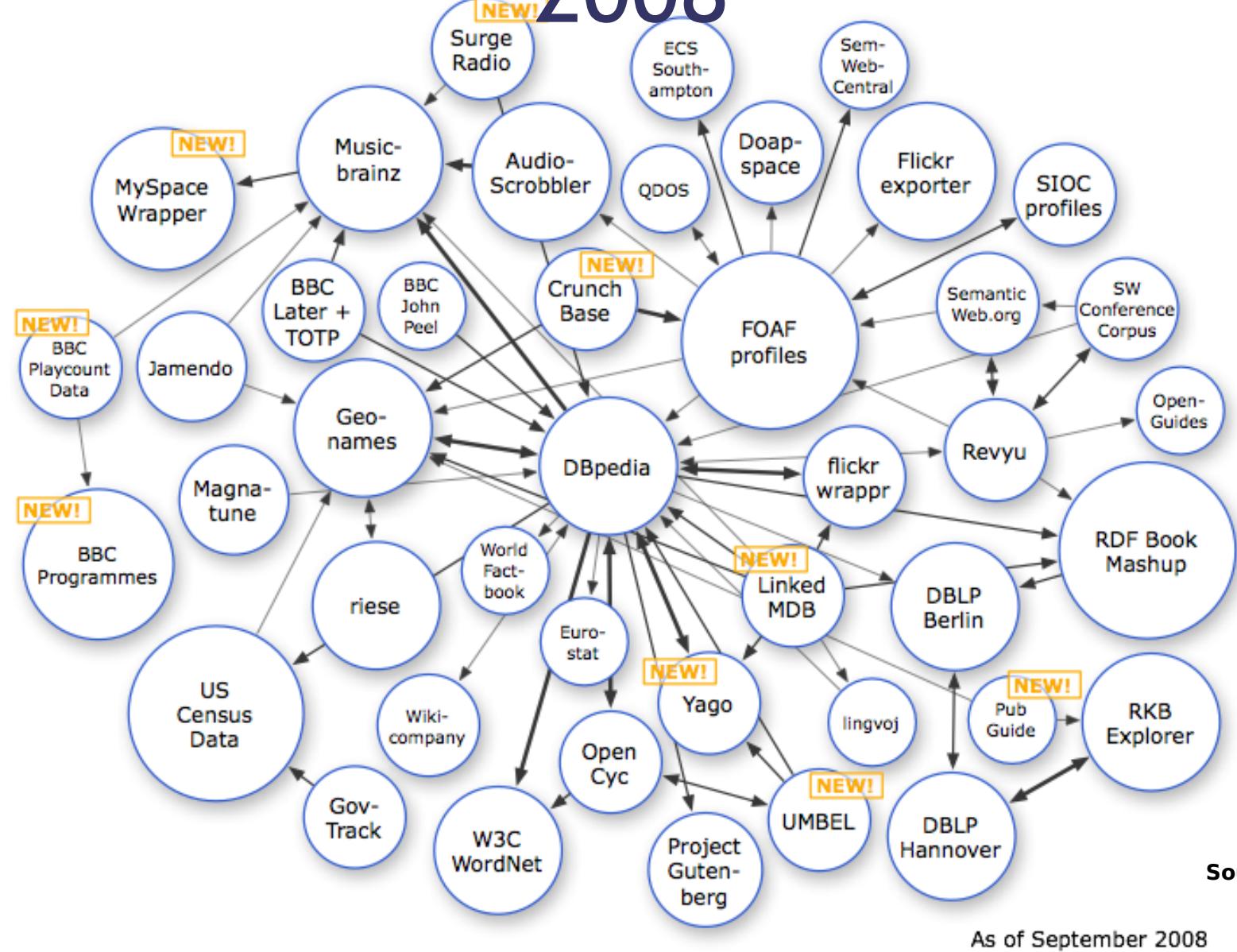


Source: <http://lod-cloud.net/>

Datasets published following Linked Data 'format':  
11/2007

# Linked Data

2008

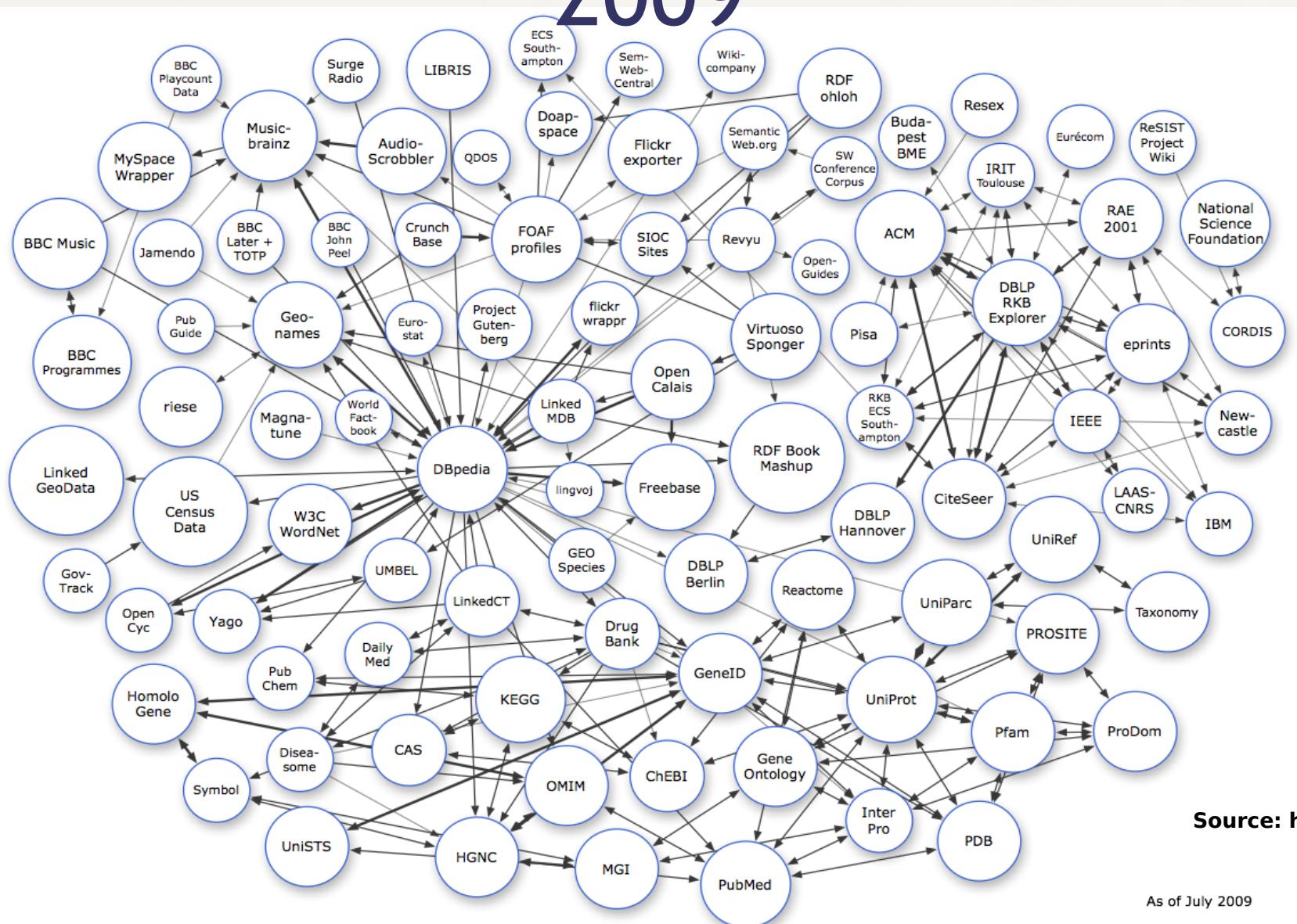


Source: <http://lod-cloud.net>

As of September 2008

## Datasets published following Linked Data ‘format’: **2008**

# Linked Data 2009



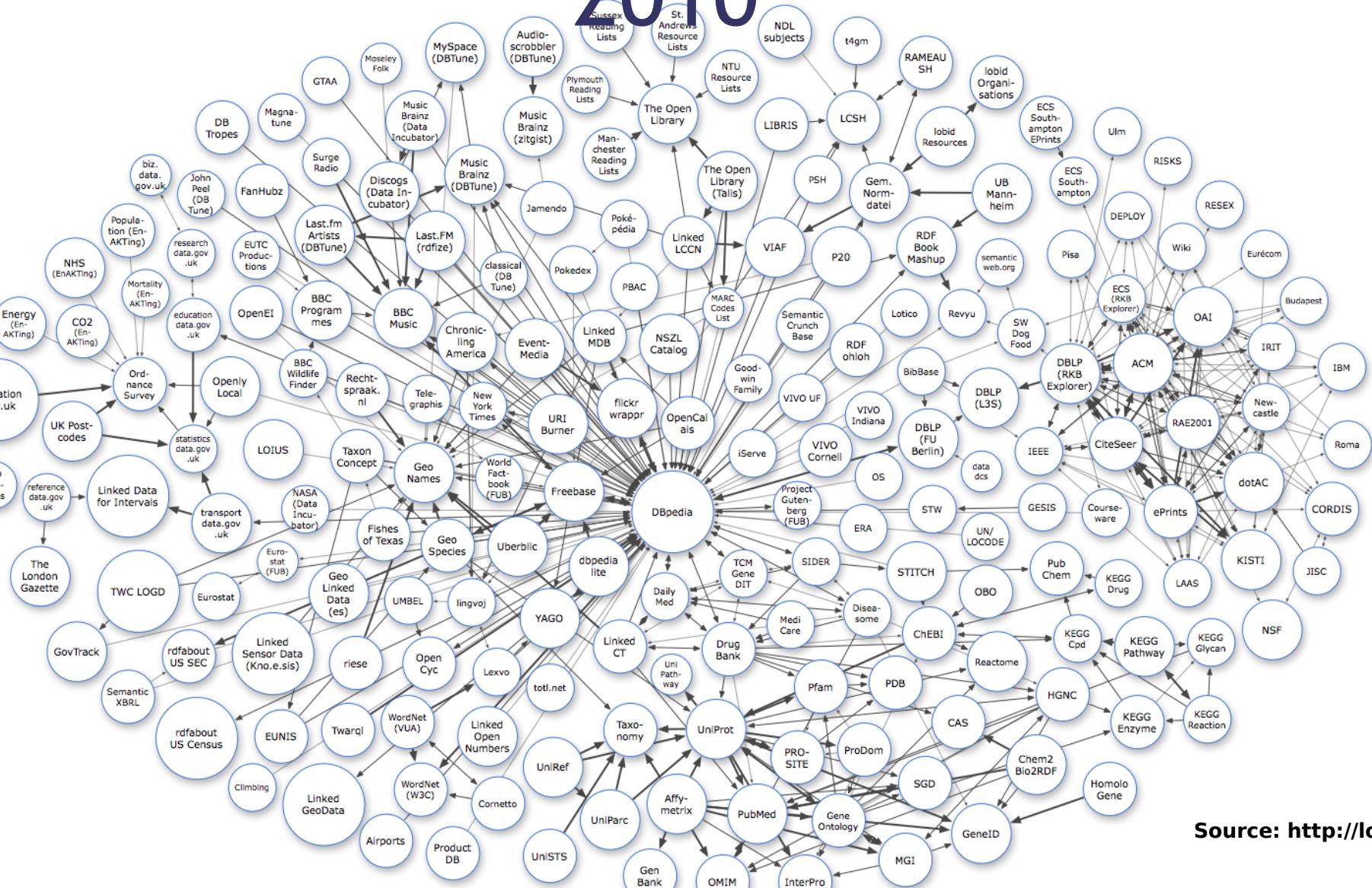
Source: <http://lod-cloud.net>

As of July 2009

## Datasets published following Linked Data ‘format’: **2009**

# Linked Data

**2010**



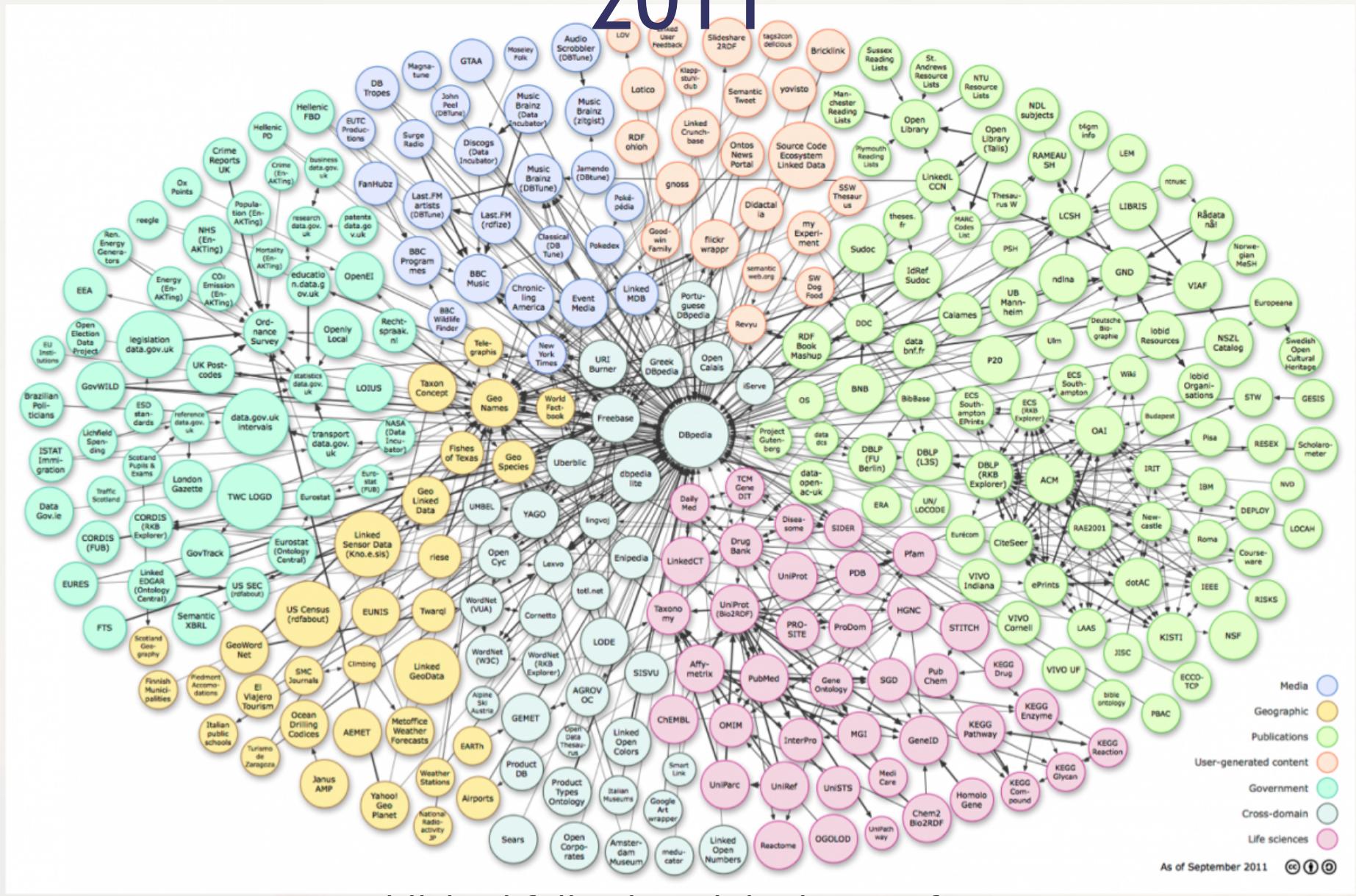
Source: <http://lod-cloud.net>

As of September 2010 

Datasets published following Linked Data ‘format’: **2010**

# Linked Data

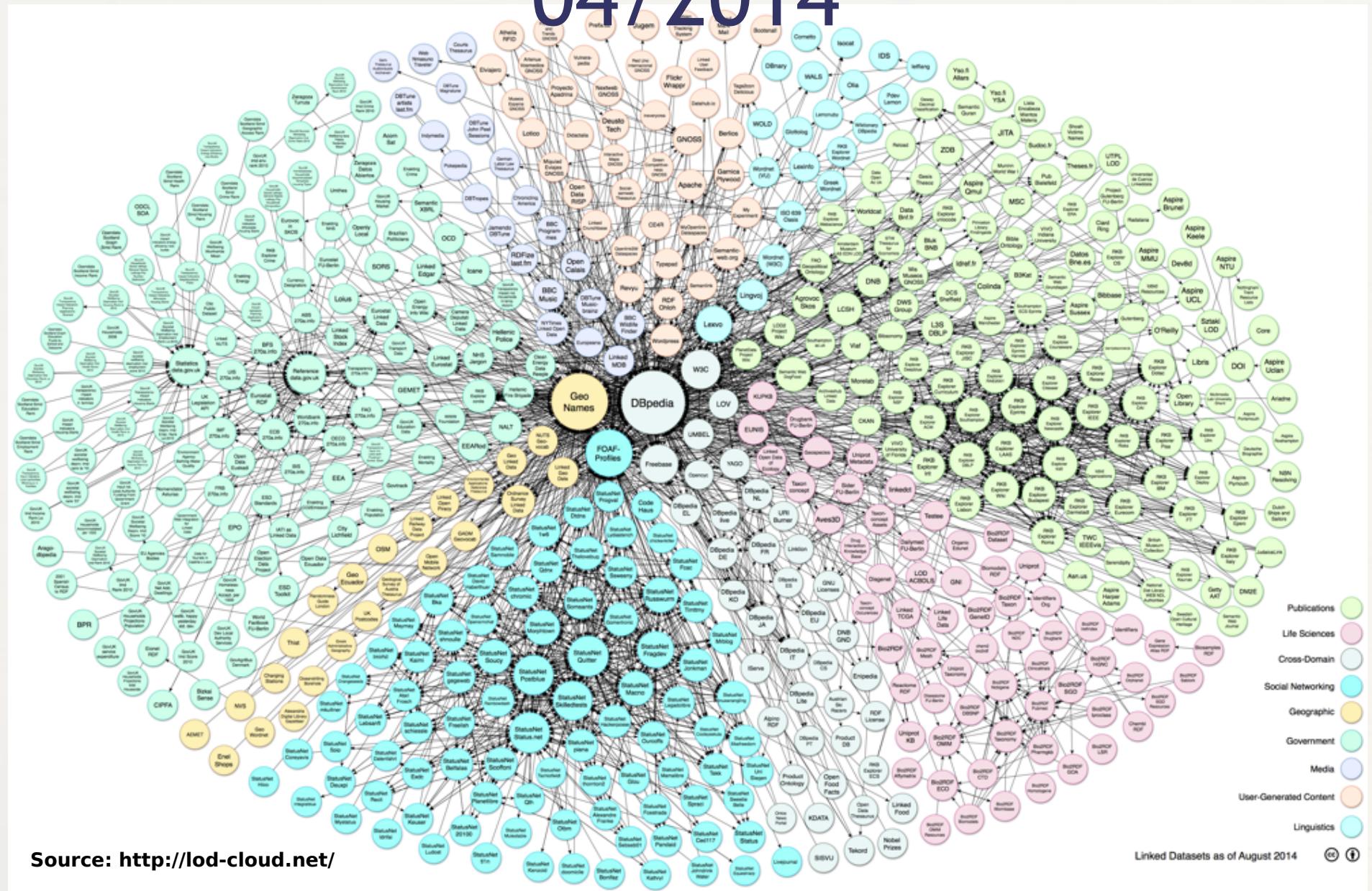
## 2011



Datasets published following Linked Data 'format': **2011**

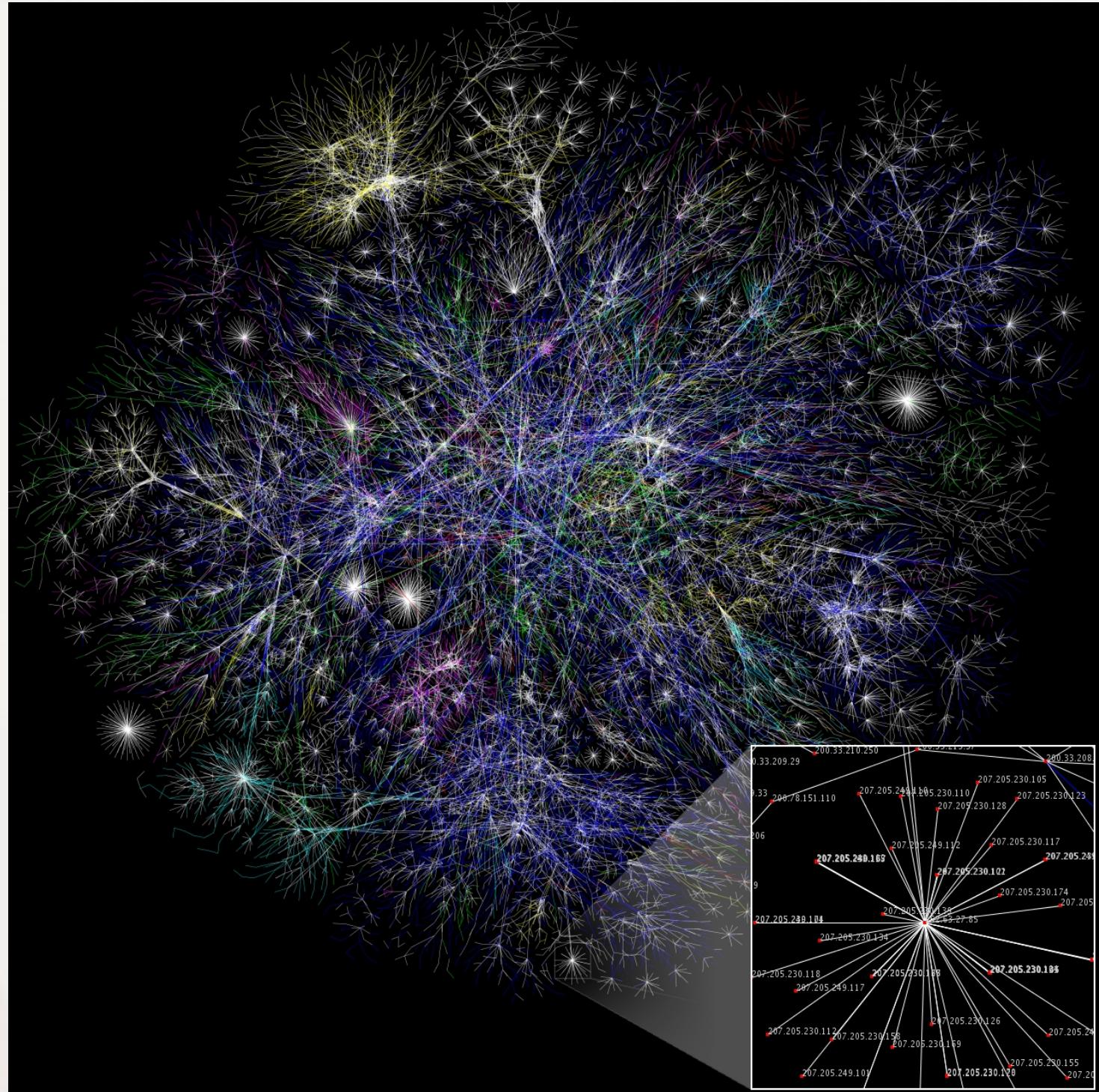
# Linked Data

# 04/2014



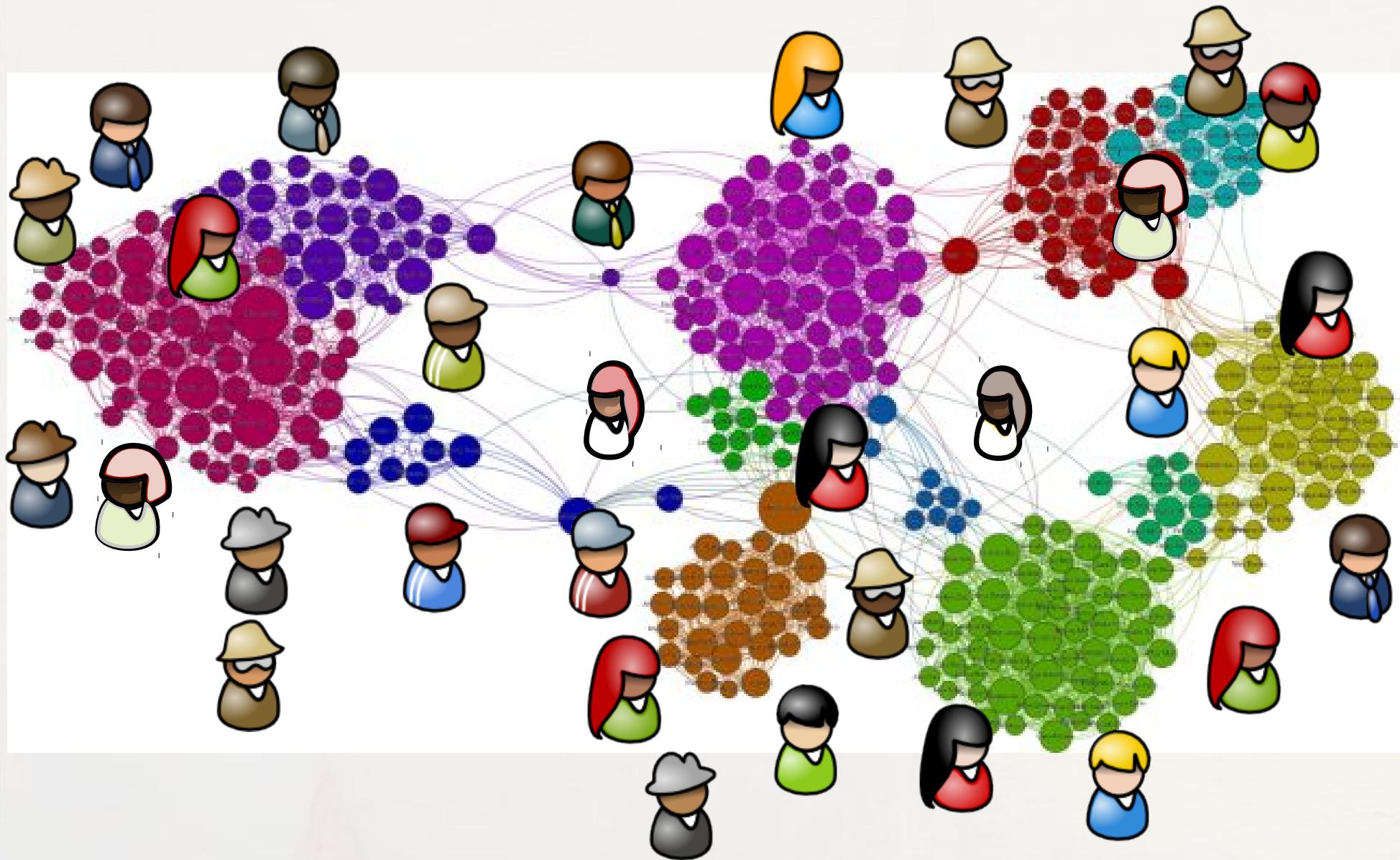
# Dados em forma de rede & Network Science

# Internet

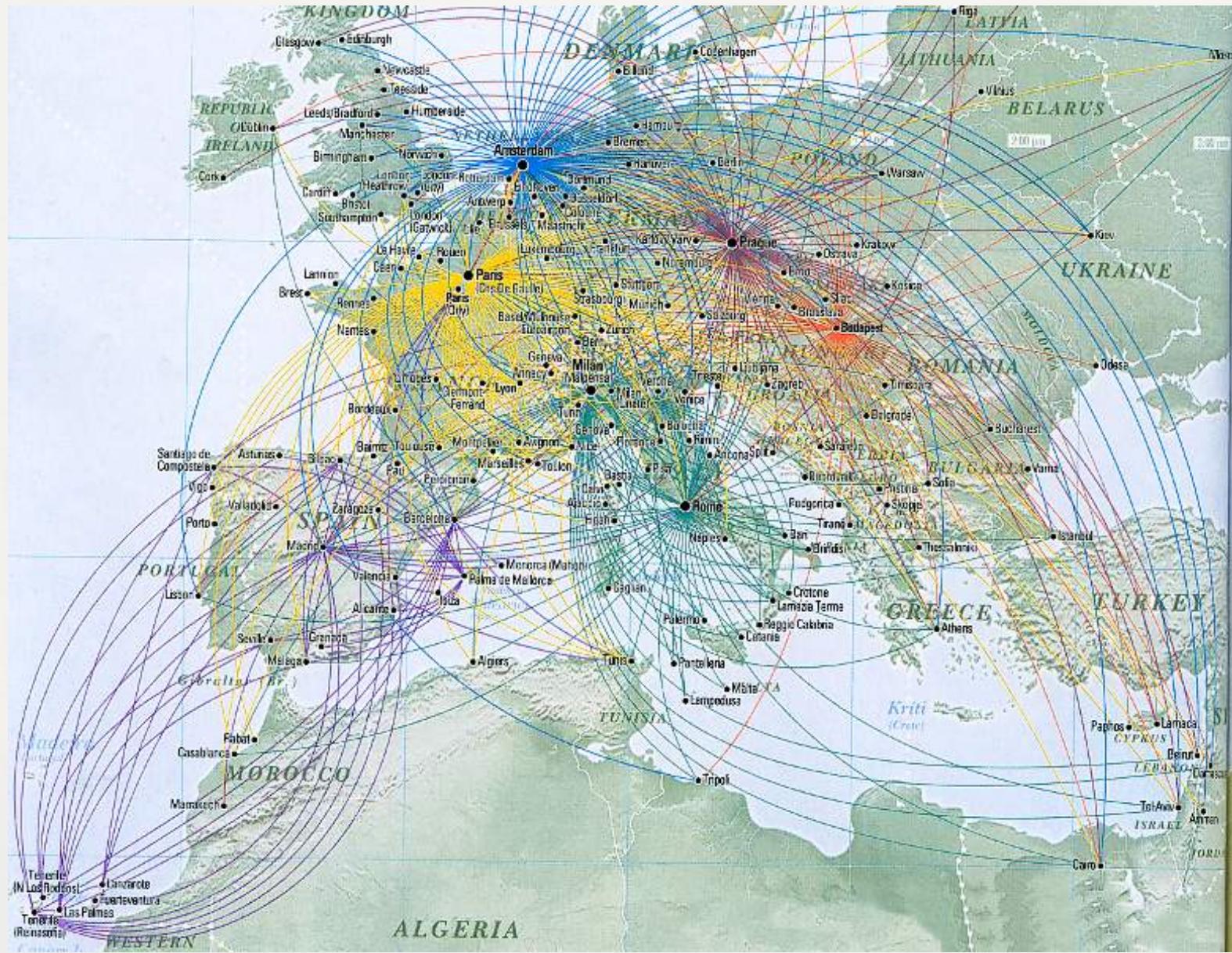


By <https://commons.wikimedia.org/wiki/User:Mmxx>

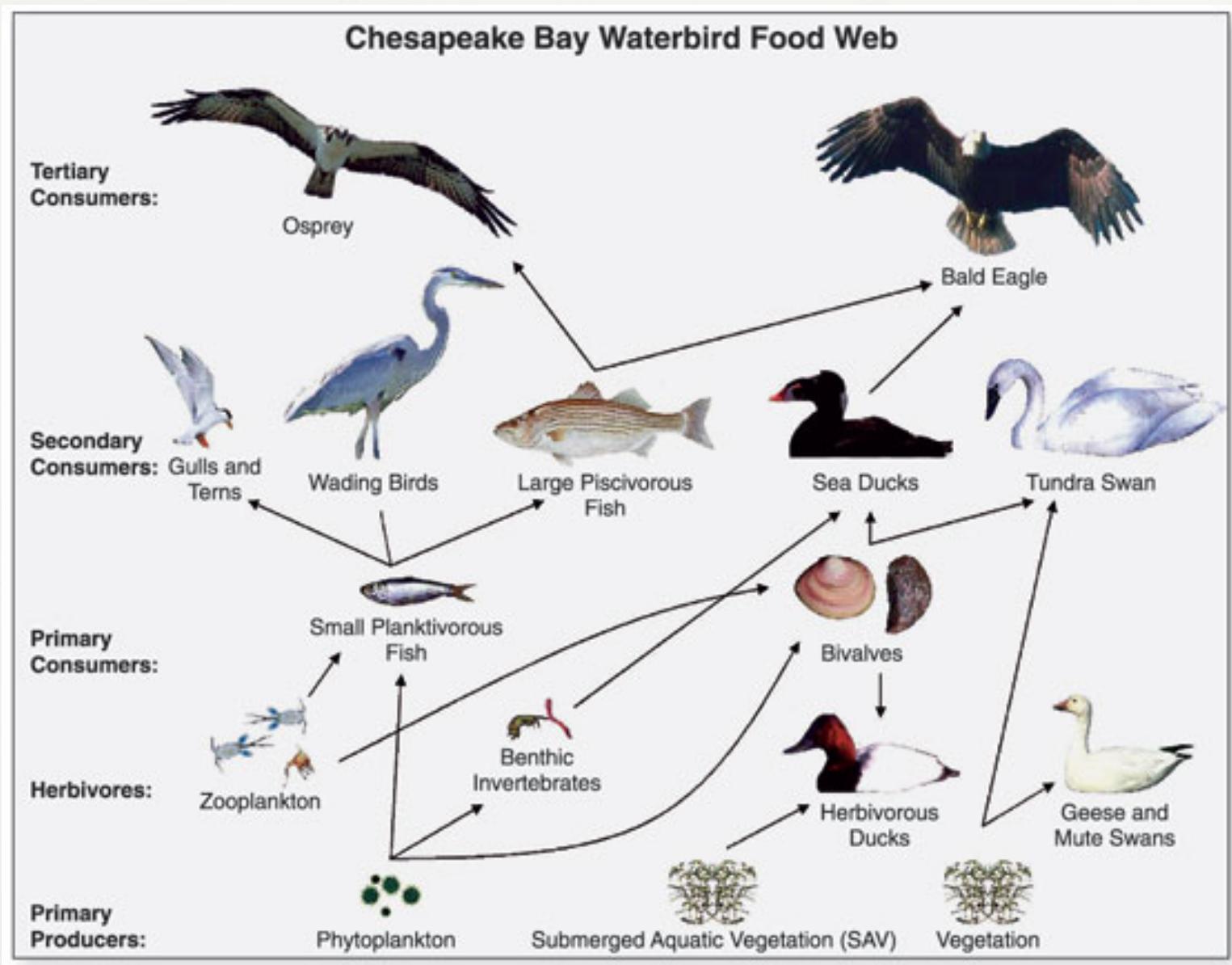
# Redes Sociais



# Redes de Transporte Aéreo

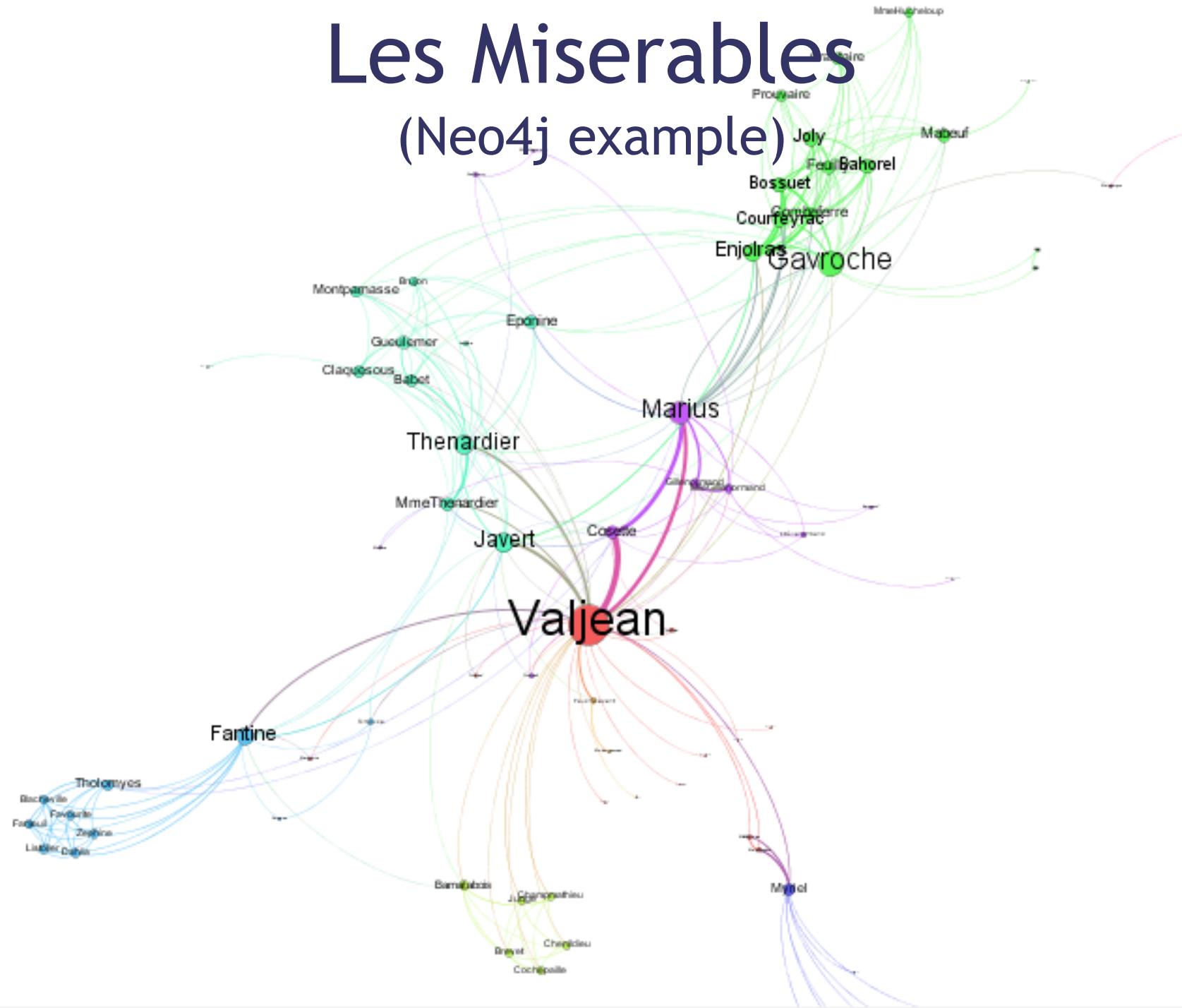


# Cadeia Alimentar

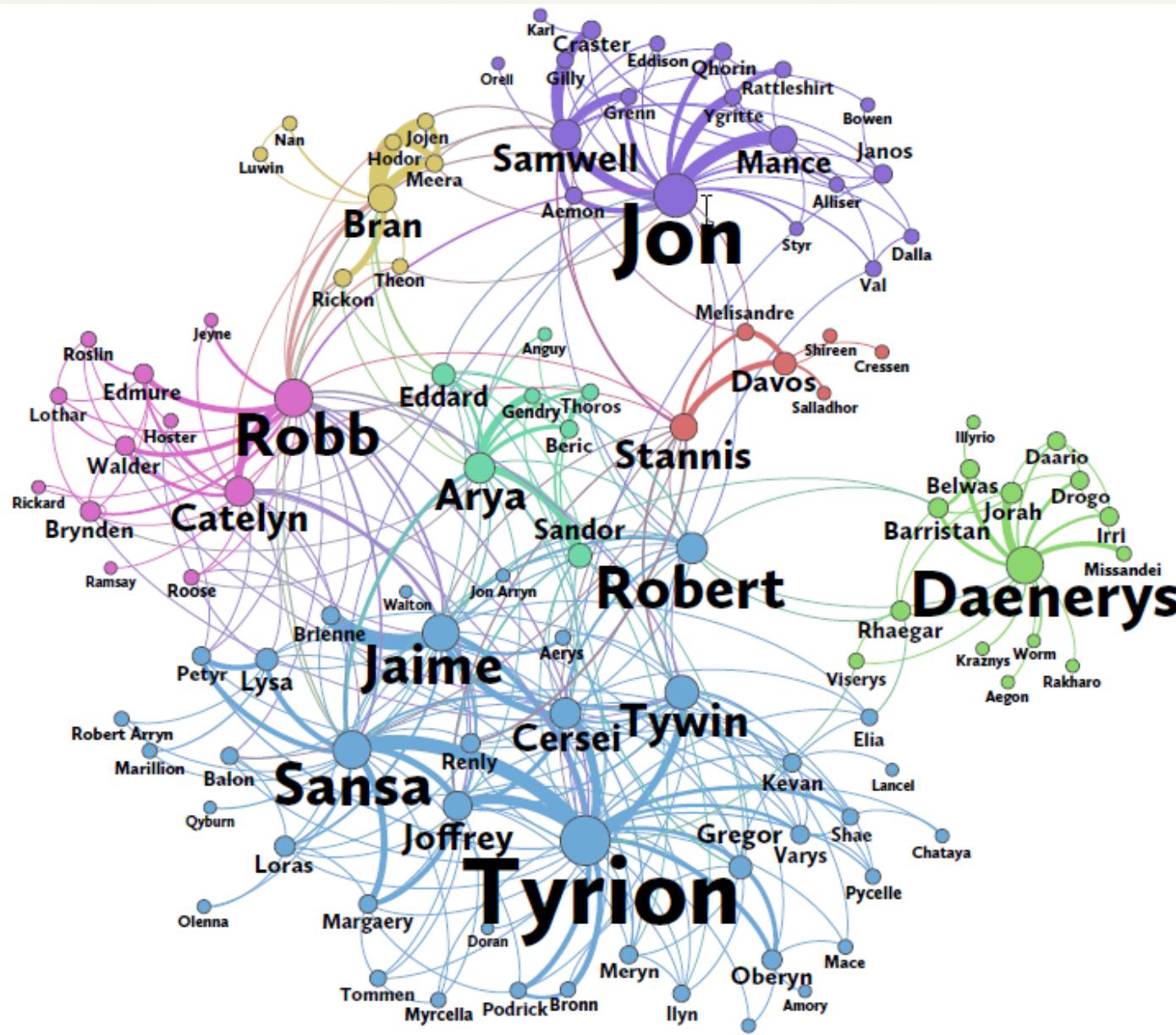


# Les Misérables

# (Neo4j example)



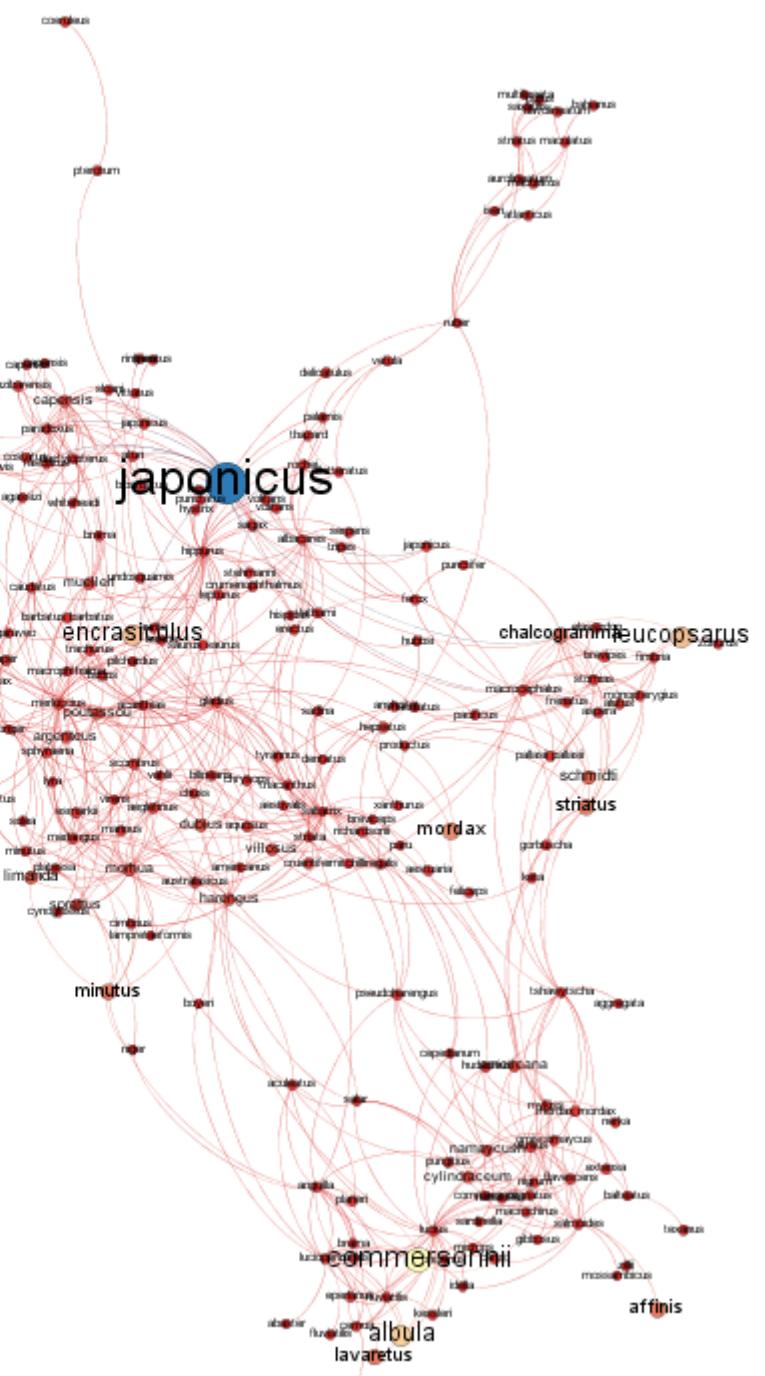
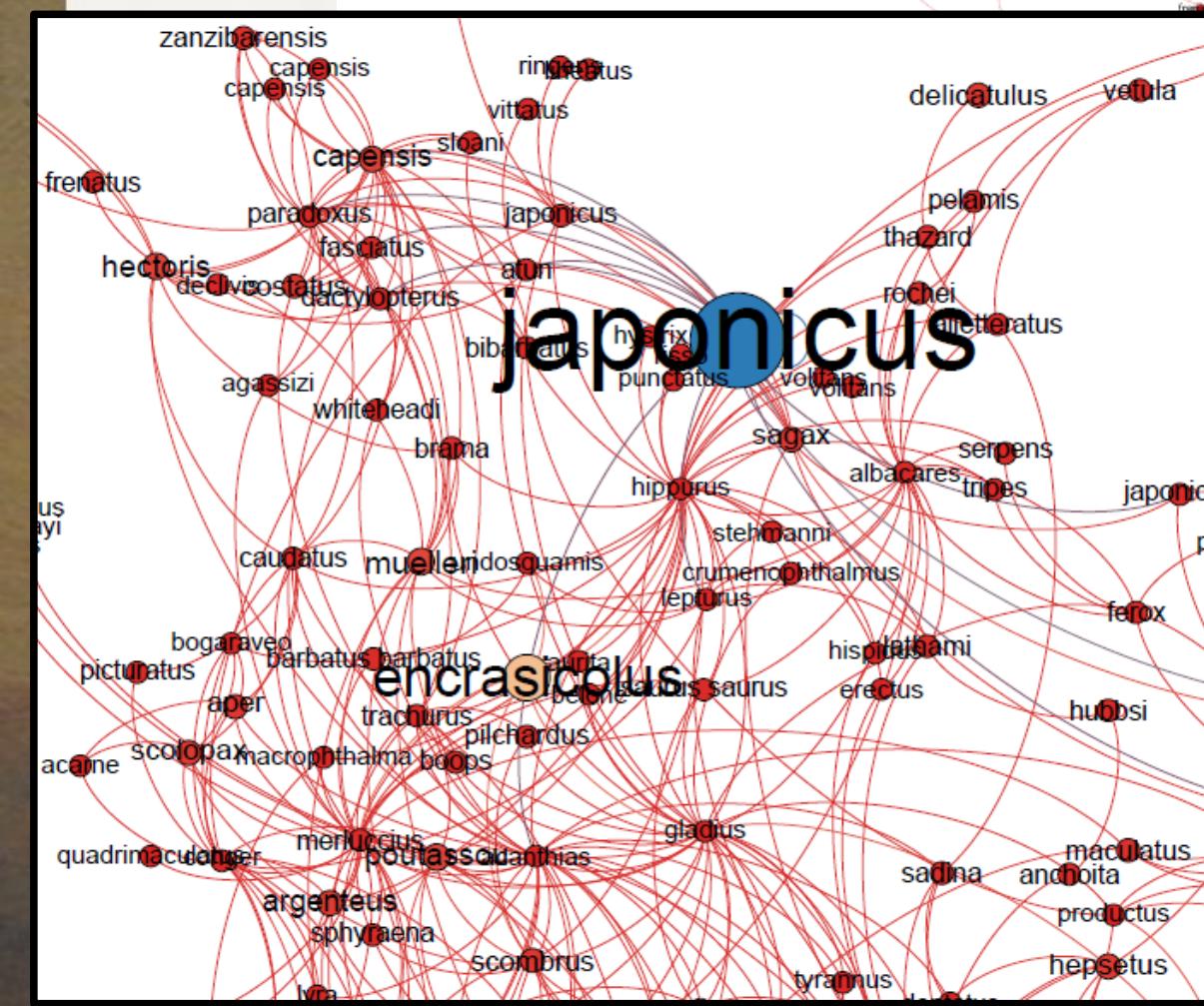
# Network of Thrones



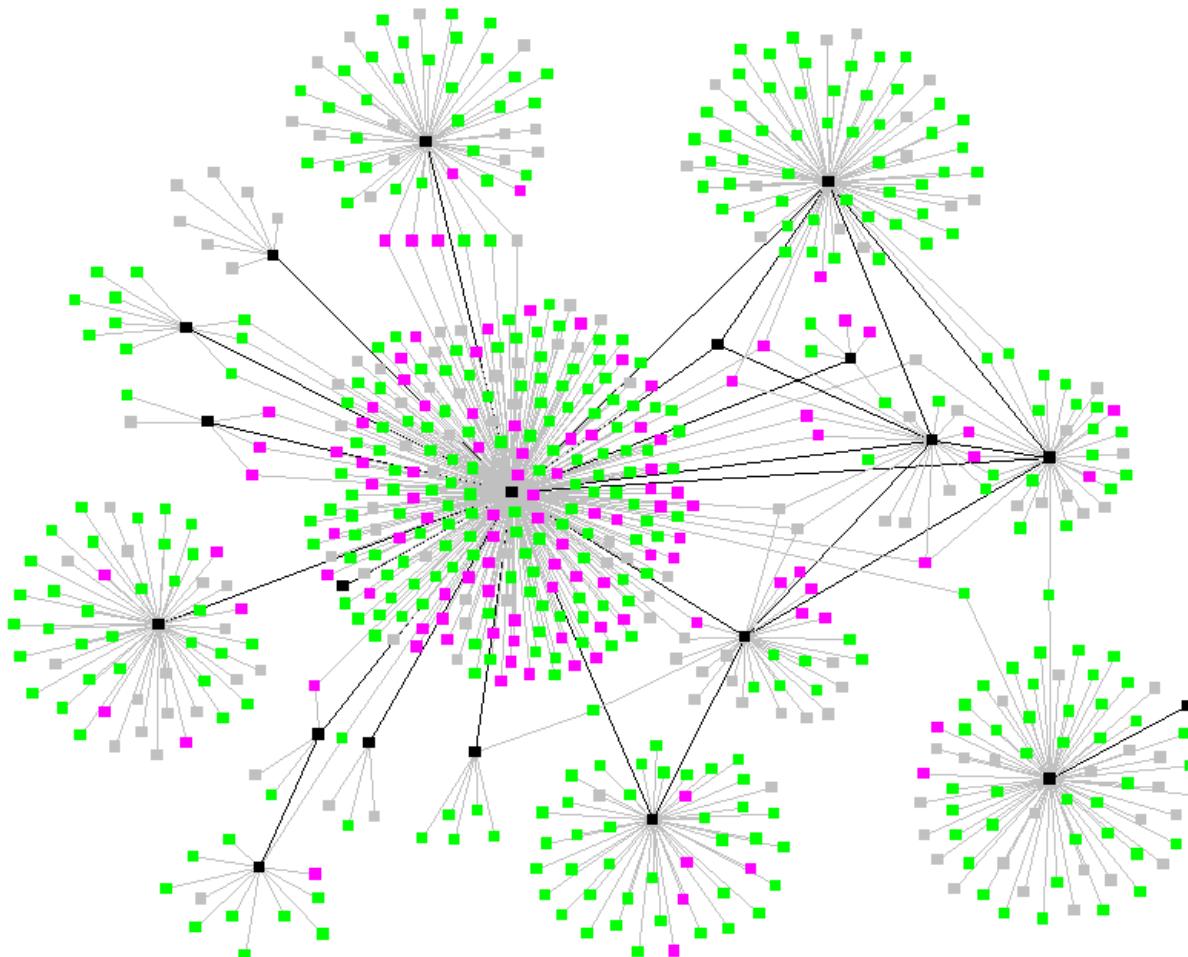
[https://www.maa.org/sites/default/files/pdf/Mathhorizons/NetworkofThrones%20\(1\).pdf](https://www.maa.org/sites/default/files/pdf/Mathhorizons/NetworkofThrones%20(1).pdf)

# Cadeia Alimentar

# FishBase

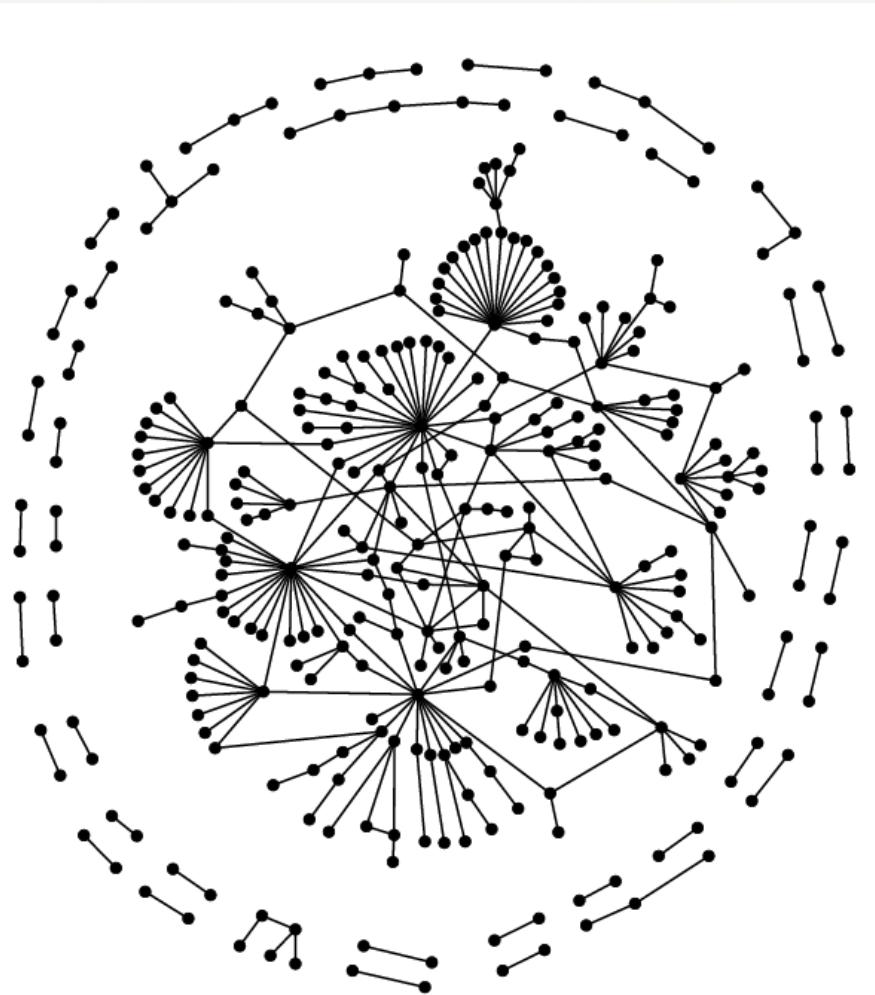


# Contágio da TB



Contagion of TB, books on politics: Valdis Krebs, [www.orgnet.com](http://www.orgnet.com).

# Proteínas da Levedura

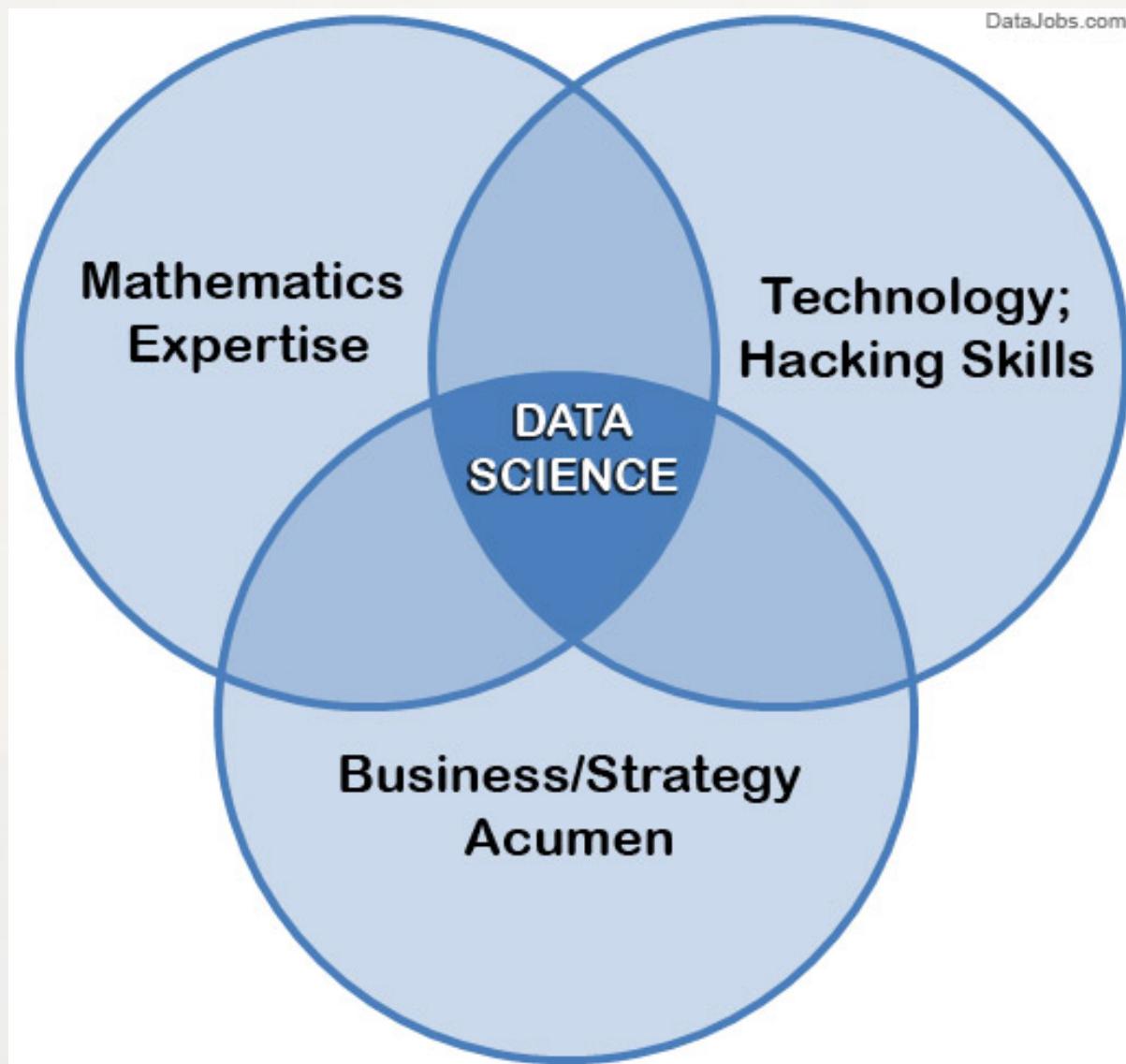


Yeast proteins: Sergei Maslov and Kim Sneppen,  
Specificity and stability in topology of protein networks,  
Science 296, 910-913 (2002).

# Data Scientist

# What is Data Science?

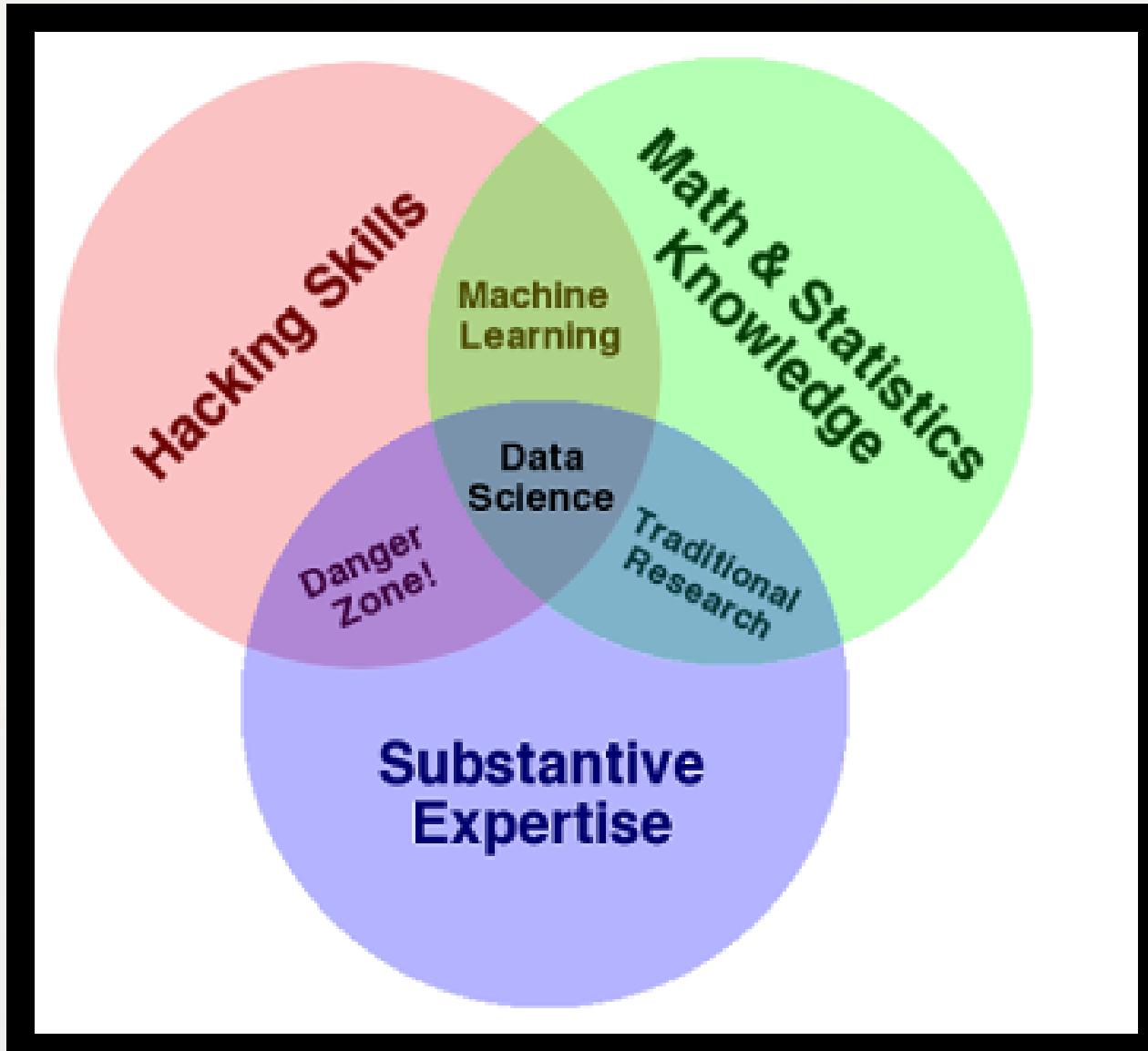
<https://datajobs.com/what-is-data-science>



# So you Want to be a Data-Scientist

Michael Spencer - 21/07/2015

<https://www.linkedin.com/pulse/so-you-want-data-scientist-michael-spencer>



# Você quer ser um Data Scientist?

## MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

## PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS



# Você quer ser um Data Scientist?

## DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative



## COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

# Competências mais buscadas por recrutadores brasileiros - LinkedIn / 2015

Ranking	Categoria de competência
1	Análise estatística e mineração de dados
2	Desenvolvimento mobile
3	Segurança de qualidade (QA) de software e teste de usabilidade
4	Logística
5	Arquitetura da web e frameworks de desenvolvimento

# Competências mais buscadas por recrutadores globais - LinkedIn / 2015

Ranking	Categoria de competência
1	Computação em nuvem e distribuída
2	Análise estatística e mineração de dados
3	Gestão de campanhas de marketing
4	Marketing, SMO e SEO
5	Middleware e softwares de integração

# Vagas na Região de Campinas e São Paulo



## Data Scientist

Big Data Brasil

São Paulo Area, Brazil

Trabalhar com uma equipe de pessoas que envolve DevOps e outros data scientists junior, al m de interlocu o direta com o CEO da empresa e ...



## Data Scientist

TOTVS

São Paulo e Região, Brasil

Strong statistics background, ideally experience with Natural Language Processing techniques; loves building mathematical and ...



## Analista de Big Data com experiência em Cassandra e Hadoop

CI&T

São Paulo e Região, Brasil

Como Analista voc trabalhar com Big Data e ferramentas de processamento de um auto volume de dados com a miss o de tra ar perfil ...



## Senior Data Scientist - Advanced Analytics

McKinsey & Company

Sao Paulo-Brazil

Working on projects and exchanging experiences with your colleagues means you will face new intellectual challenges on a daily basis, ...

<https://www.linkedin.com/vsearch/j?type=all&keywords=data%20scientist>

# Vagas na Região de Campinas e São Paulo



## Data Analyst, Communications

Facebook

São Paulo -Brazil

The Data Comms team mission is to create data stories that highlight the many ways in which people come together on Facebook during those...



## Scientist

Philips

Brazil-São Paulo-São Paulo

For this, we are seeking a research engineer to investigate and drive technical innovations to bring Big Data concepts to clinical ...



## Consultant (Data Scientist)

Accenture

São Paulo, 27, BR

Ph.D. in Econometrics, Statistics, Economics or Mathematics (with sound time series and/or statistical background) - Experienced ... [careercarc.com](http://careercarc.com)



## Analista de Ciencia de Dados Pleno/Senior

Itaú Unibanco

São Paulo, São Paulo, Brazil

Aplicar vis o hol stica e considerar iniciativas atuais de consumo de dados e o ambiente de dados j existente. Racioc nio anal tico.

<https://www.linkedin.com/vsearch/j?type=all&keywords=data%20scientist>

USE THE  
CRS DATA-  
BASE TO  
SIZE THE  
MARKET.

THAT  
DATA IS  
WRONG.



[www.dilbert.com](http://www.dilbert.com)

THEN  
USE THE  
SIBS  
DATA-  
BASE.

THAT  
DATA IS  
ALSO  
WRONG.



5-7-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.

CAN YOU  
AVERAGE  
THEM?

SURE. I CAN  
MULTIPLY  
THEM TOO.



# Referências

- Dijkstra, E. W. (1986) On a cultural gap. *The Mathematical Intelligencer*. vol. 8, no. 1, pp. 48-52.
- Ramakrishnan, Raghu; Gehrke, Johannes (2003) **Database Management Systems**. McGraw-Hill, 3<sup>rd</sup> edition.

# Agradecimentos

- Luiz Celso Gomes Jr (professor desta disciplina em 2014) pela contribuição na disciplina e nos slides. Página do Celso: <http://dainf.ct.utfpr.edu.br/~gomesjr/>

# André Santanchè

<http://www.ic.unicamp.br/~santanche>

# Licença

- Estes slides são concedidos sob uma Licença Creative Commons. Sob as seguintes condições: Atribuição, Uso Não-Comercial e Compartilhamento pela mesma Licença.
- Mais detalhes sobre a referida licença Creative Commons veja no link:  
<http://creativecommons.org/licenses/by-nc-sa/3.0/>
- Fotografia da capa feita por André Santanchè no Petit Palais (Paris) em 17/02/2013 do quadro: Fantasia à Constantinople de Felix Ziem