Banco de Dados

Fundamentos

André Santanchè e Patrícia Cavoto
Instituto de Computação – UNICAMP
Agosto 2015

Fantasia à Constantinople por Felix Ziem

Começando com Banco de Dados

Motivação

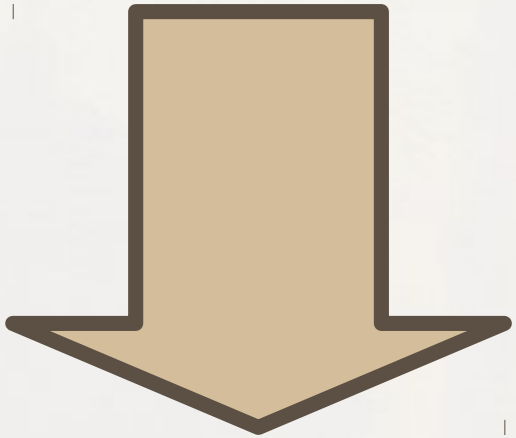
- Aplicações computacionais de todos os portes trabalham com grandes volumes de dados
  - Gerenciamento de uma farmácia
  - Sistema bibliotecário
  - Sistema bancário

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
  - etc.

Banco de Dados
Aplicação Exemplo

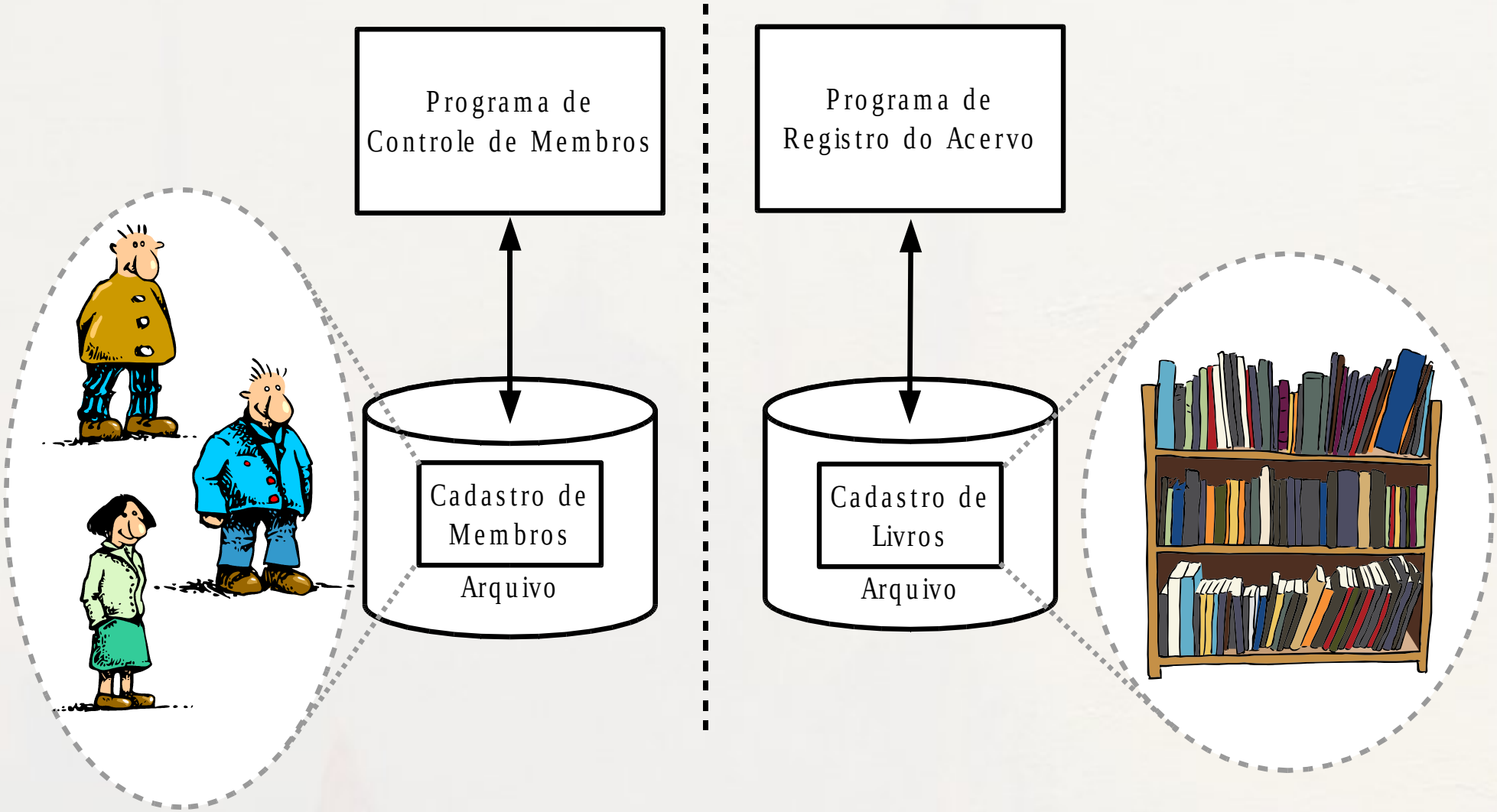
- Gerenciamento de uma biblioteca
- Serviços:
  - Cadastro de membros associados
  - Registro do acervo (ex.: livros, revistas etc.)
  - Controle de empréstimos

Banco de Dados
Perspectivas

- Arquitetura
- Abstração

Banco de Dados
Arquitetura

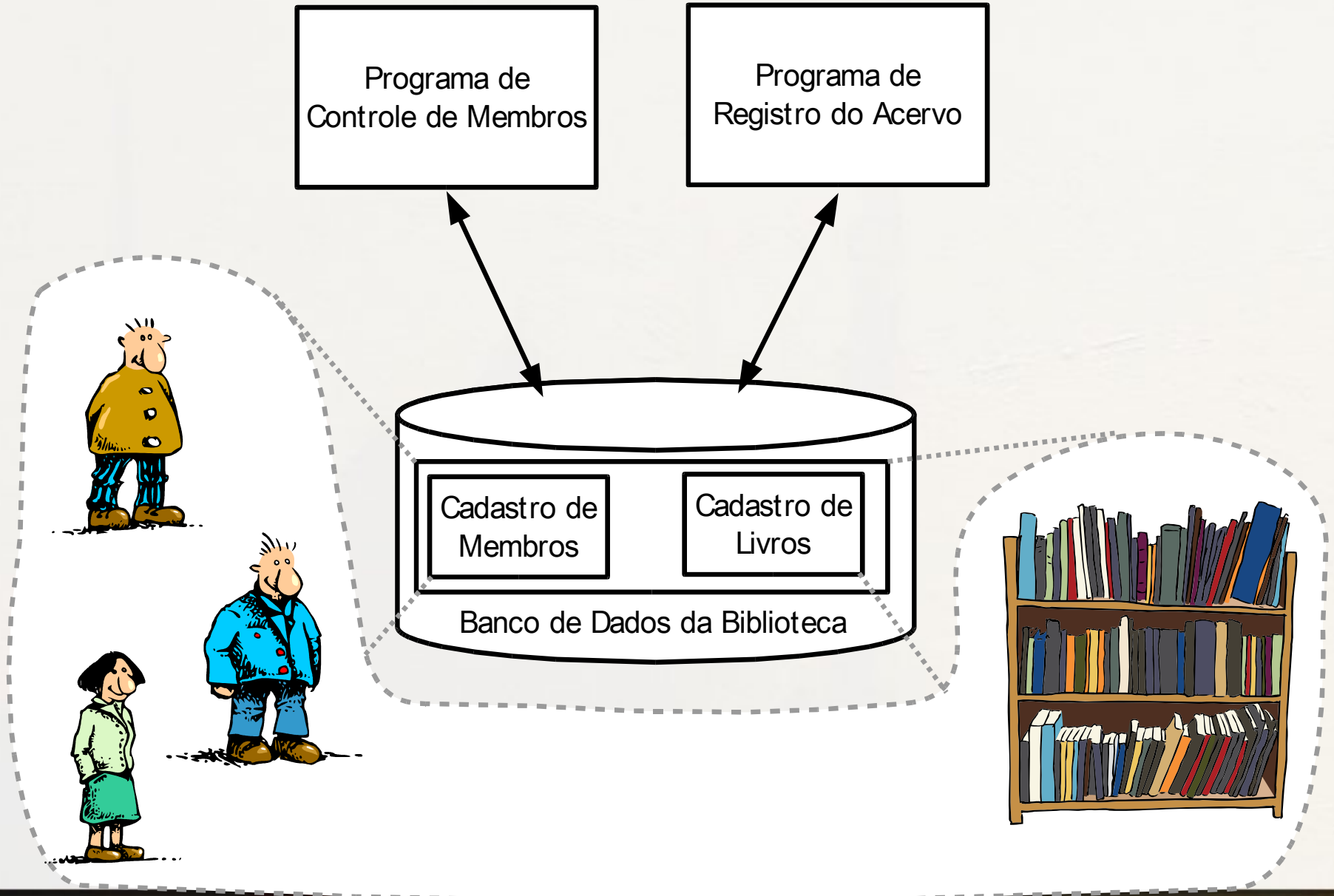
Sistemas Isolados



Sistemas/Arquivos Isolados

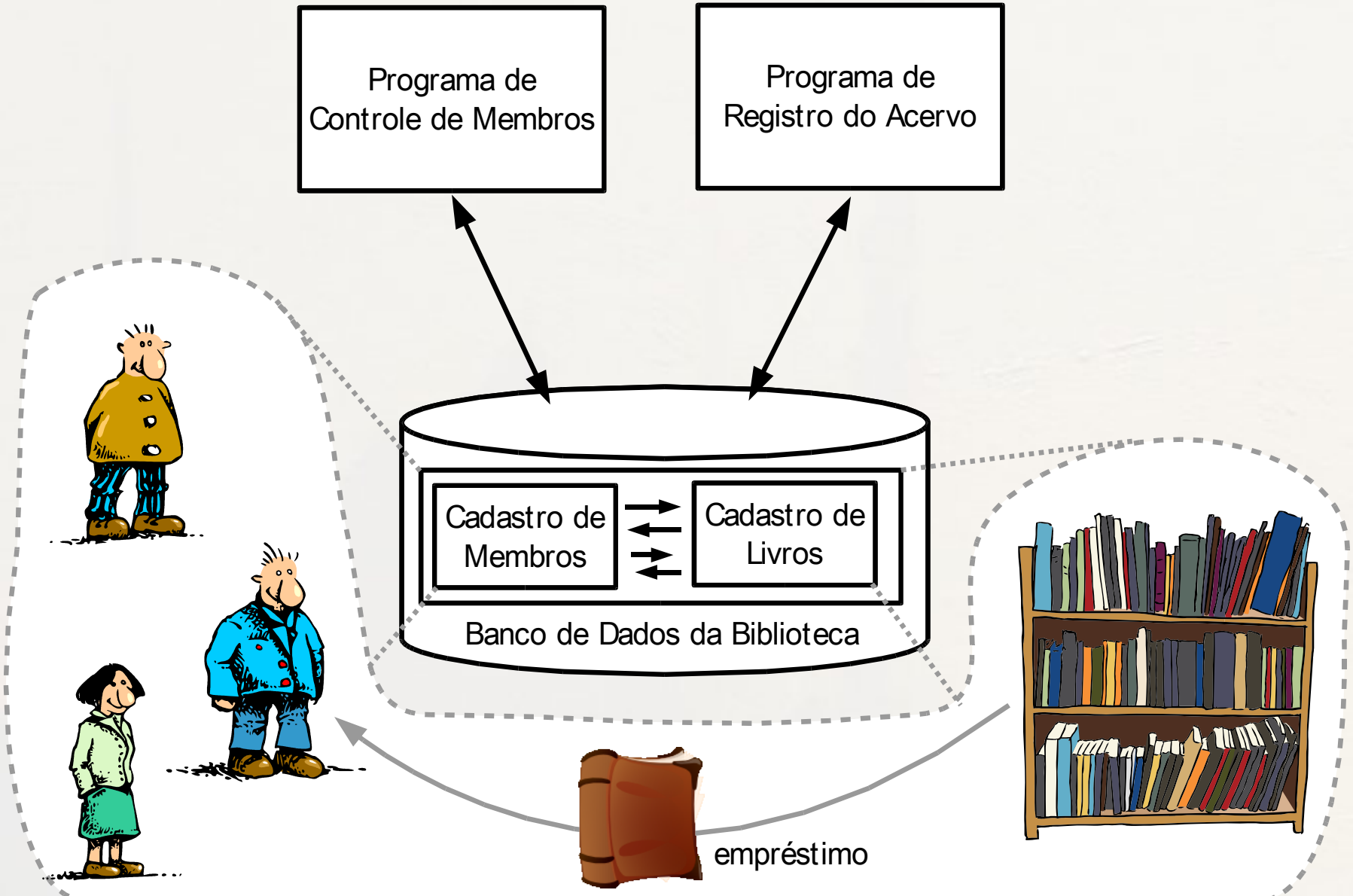
- Redundância não controlada
  - Repetição
  - Inconsistência
- Barreiras para relacionamento entre arquivos
- Dificuldades com:
  - acesso concorrente
  - integridade e recuperação em caso de crash
  - segurança e controle de acesso

Banco de Dados
Compartilhamento de Dados



Banco de Dados

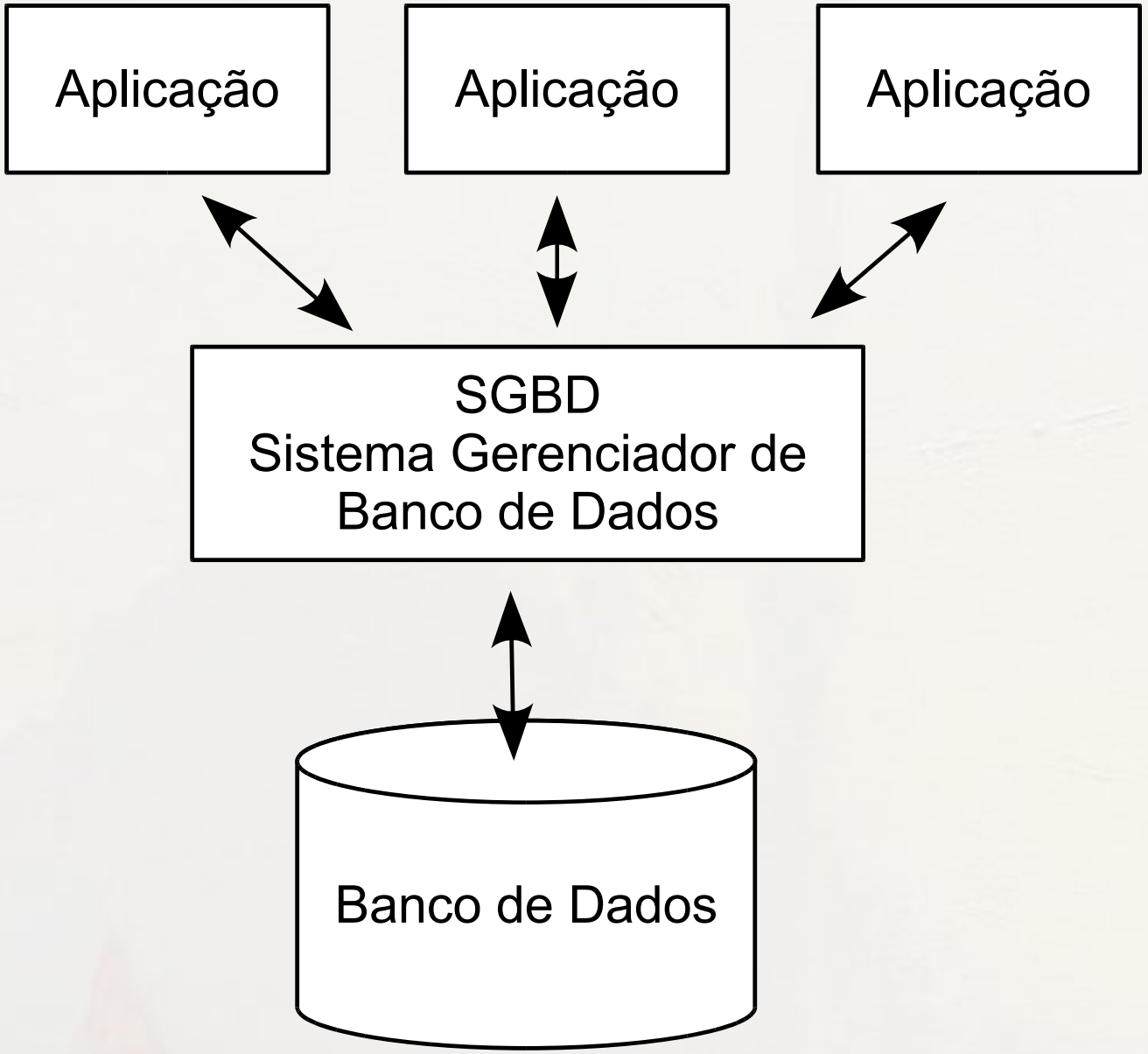
Relações entre os Dados



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


Aplicação
Aplicação
Aplicação
SGBD
Sistema Gerenciador de
Banco de Dados
Banco de Dados

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
  - 30/06/2015 – 1,49 bilhões de usuários ativos
    - http://newsroom.fb.com/company-info/

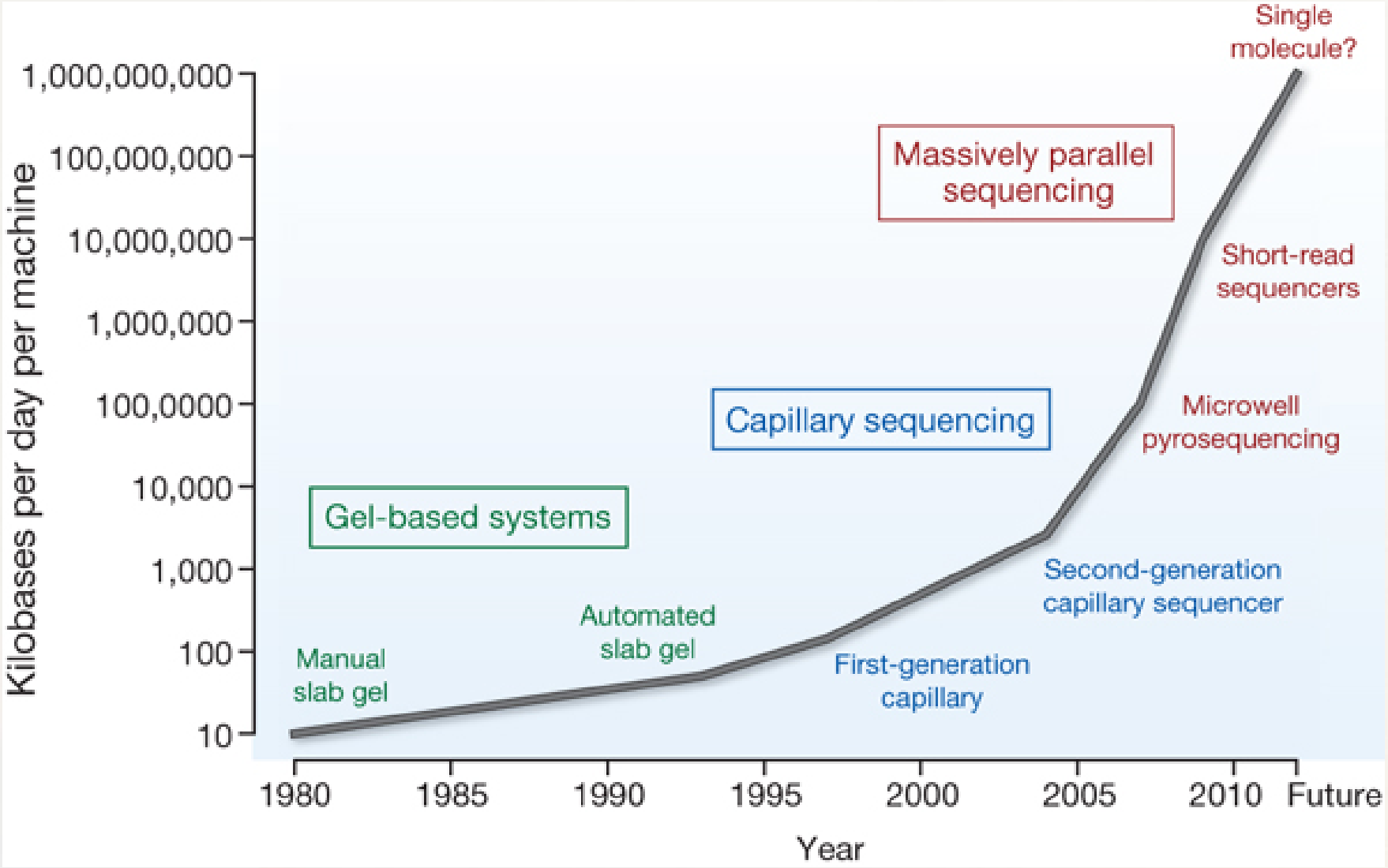
Lei de Moore

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

Como crescem os dados?

Sequenciamento de Genoma

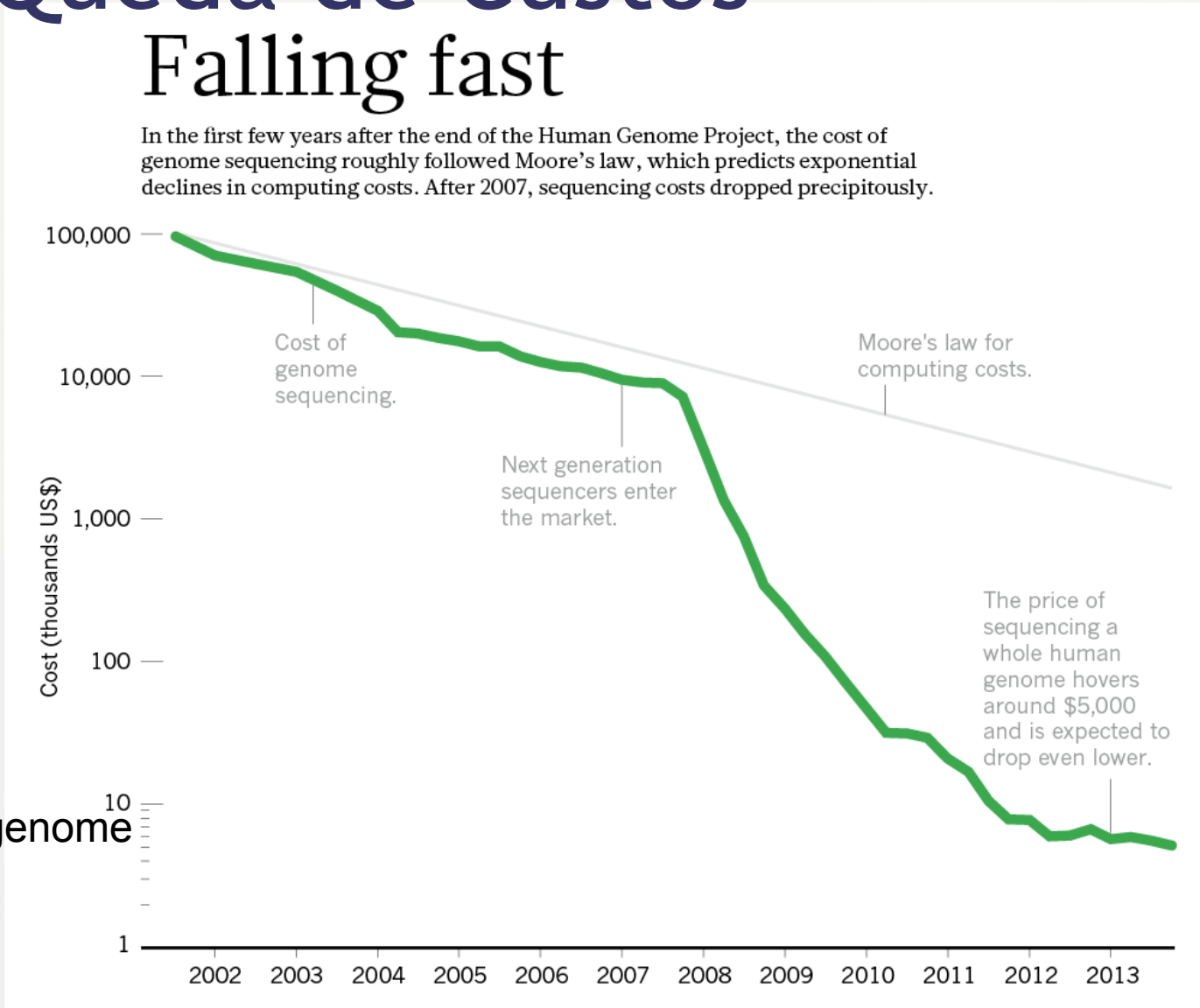
Sequenciamento de Genoma
Aumento na Eficiência



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Sequenciamento de Genoma

Queda de Custos

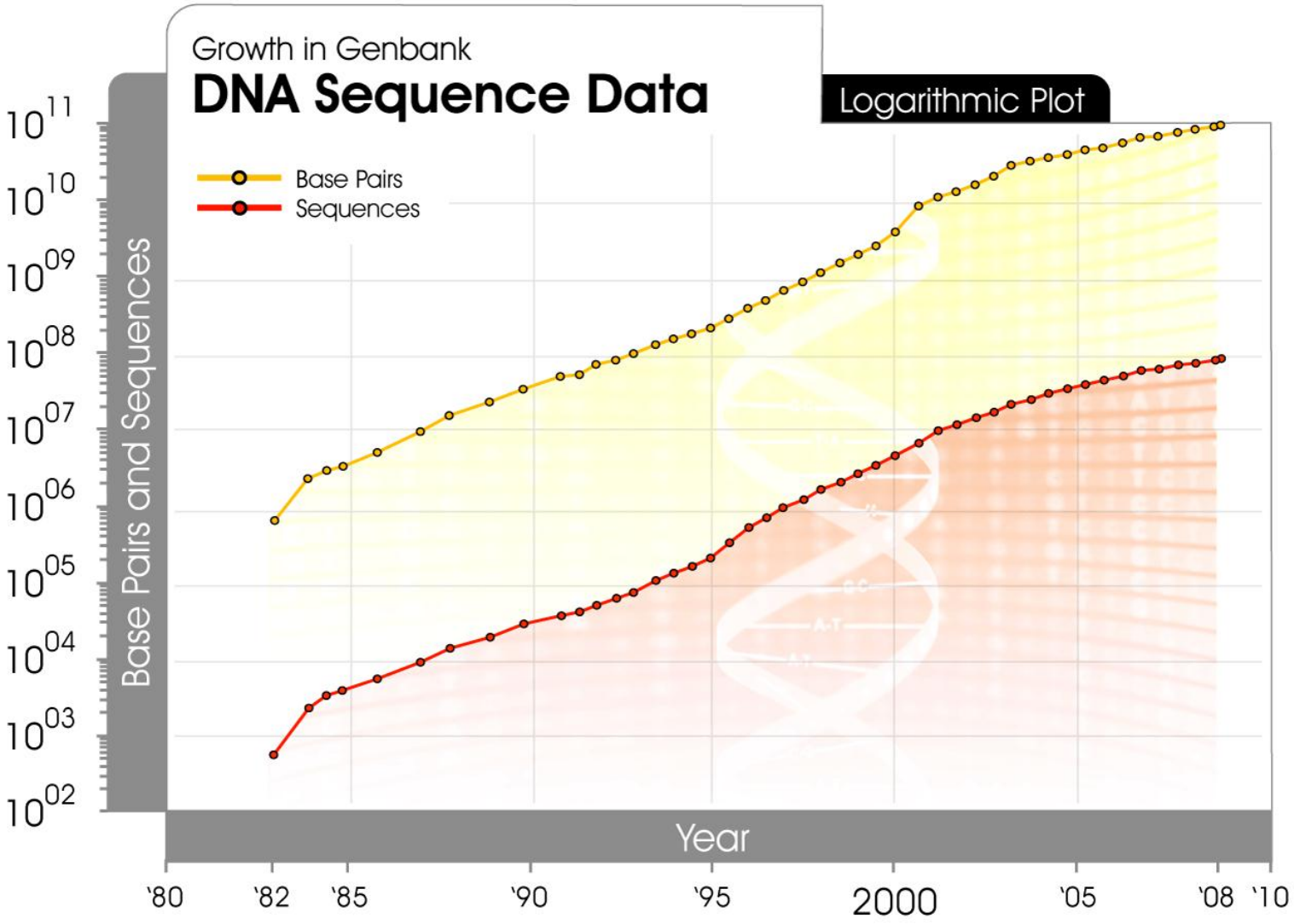


Technology: The $1,000 genome
by Erika Check Hayden
19 March 2014
Nature News

http://www.nature.com/news/technology-the-1-000-genome-1.14901

Sequenciamento de Genoma

Volume de Dados



Raymond Kurzweil

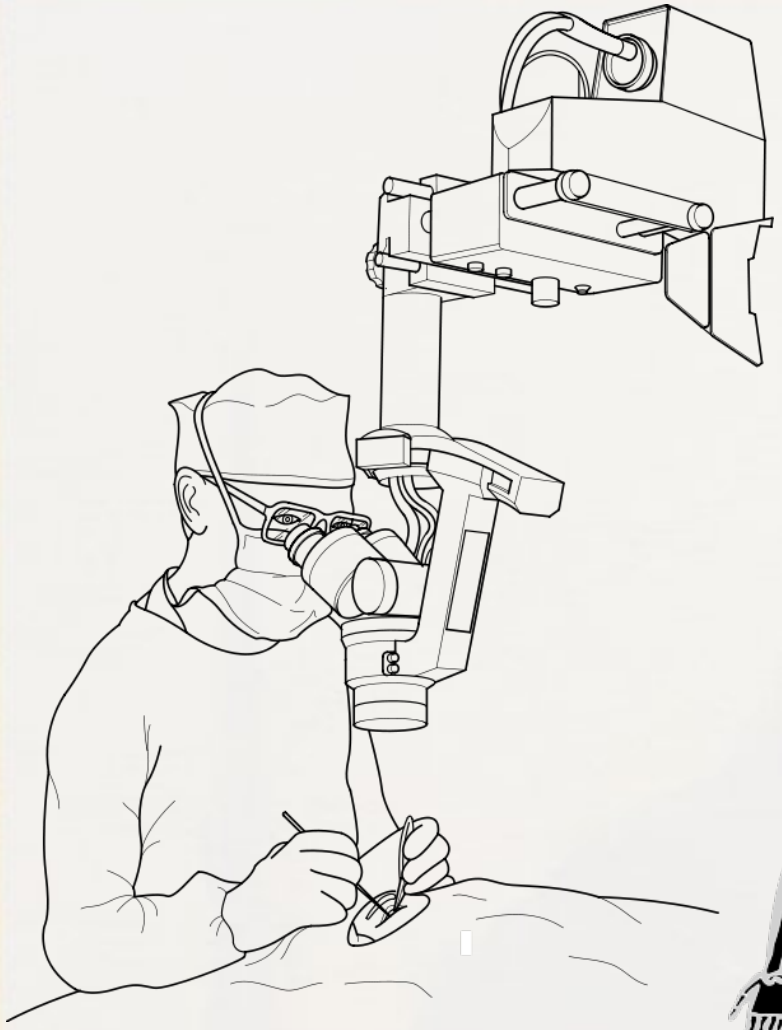
http://www.kurzweilai.net/dna-sequencing-data

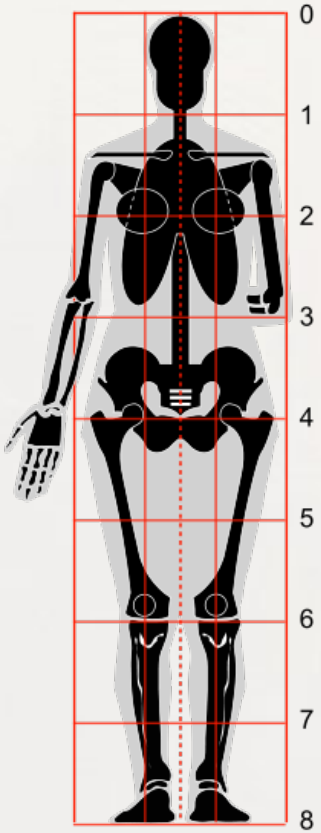
Harvard Business Review
The True Measures of Success
10 Rules for Managing Global Innovation
What Ever Happened To Accountability?
GETTING CONTROL OF BIG DATA
How vast new streams of information are changing the art of management
Science
11 February 2011 | $10
data
The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
COMPUTERWORLD
WHY BIG DATA IS A BIG DEAL
COVER STORY | BI & ANALYTICS
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

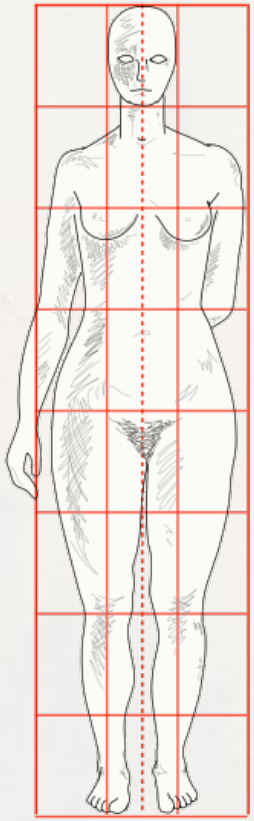
Data Engineering
&
Data Science

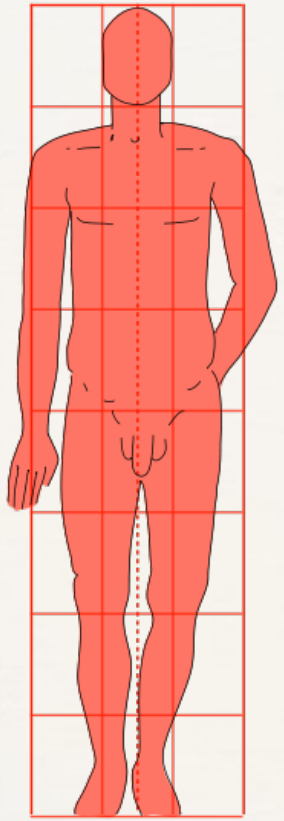
Como Aprendemos Computação?

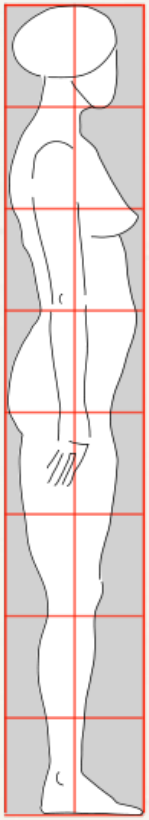
Cirurgia





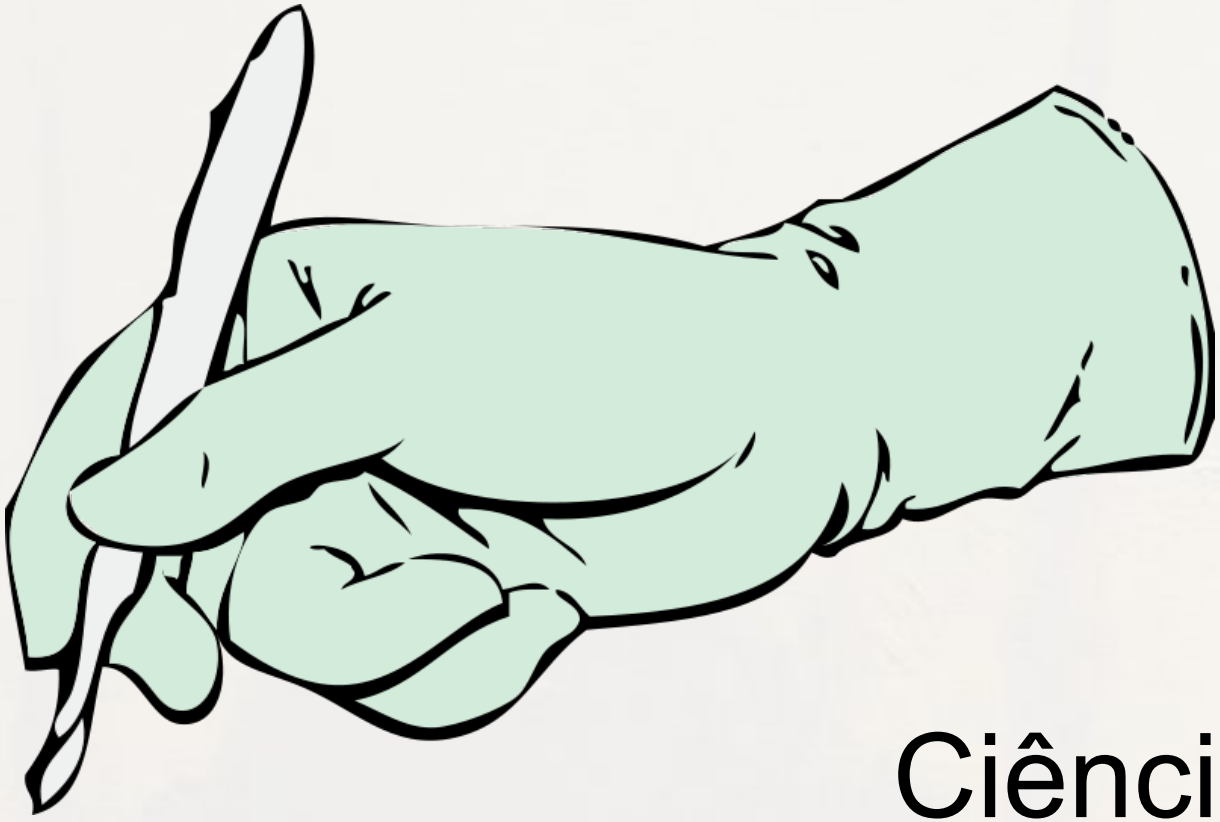






By Original by Schnorch retraced by LadyofHats

Cirurgia



Ciência da Faca

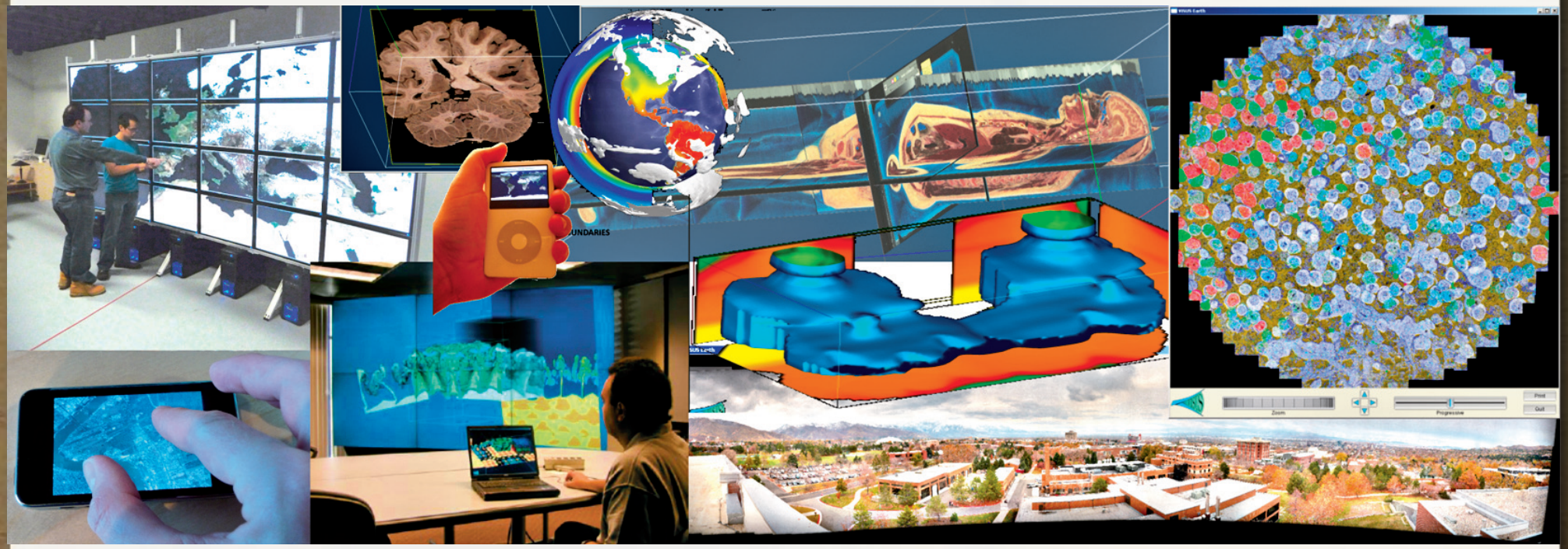
Ciência da Computação?

- 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 Krist in Tolle
Microsoft Research
Redmond, 2009

Fourth Paradigm

- **Thousand years ago:**
  - science was empirical; describing natural phenomena
- **Last few hundred years:**
  - theoretical branch; using models, generalizations
- **Last few decades:**
  - a computational branch; simulating complex phenomena
- **Today: data exploration (eScience)**
  - unify theory, experiment, and simulation

(Jim Gray, 2007)

Mineração de Dados
e
Descoberta de Conhecimento

What Wal-Mart Knows About Customers' Habits

Constance L. Hays

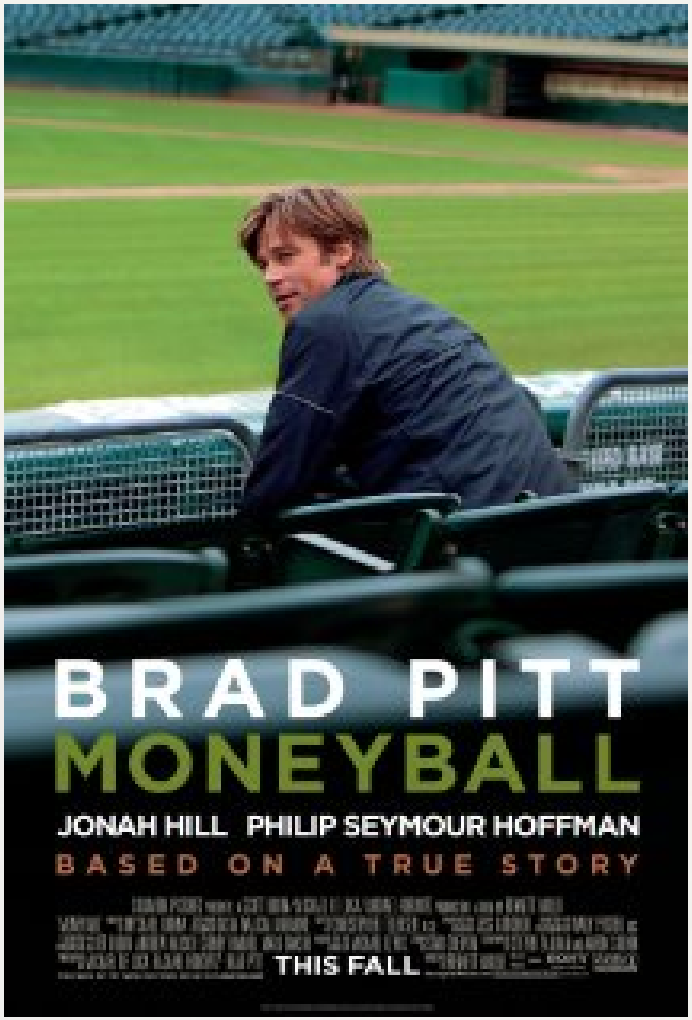
The New York Times, 2004

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



Alemanha 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

Our Brand Is Crisis

by Rachel Boynton

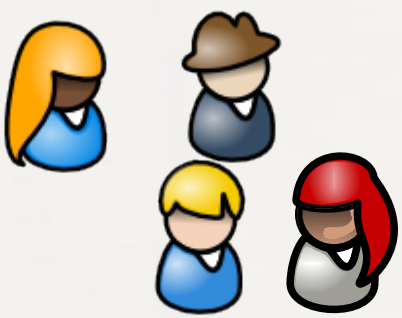
- Documentary of the 2002 Bolivian presidential election
- Gonzalo Sánchez de Lozada x Evo Morales
- Tacts by the Greenberg Carville Shrum (GCS) company

Minerando na Web

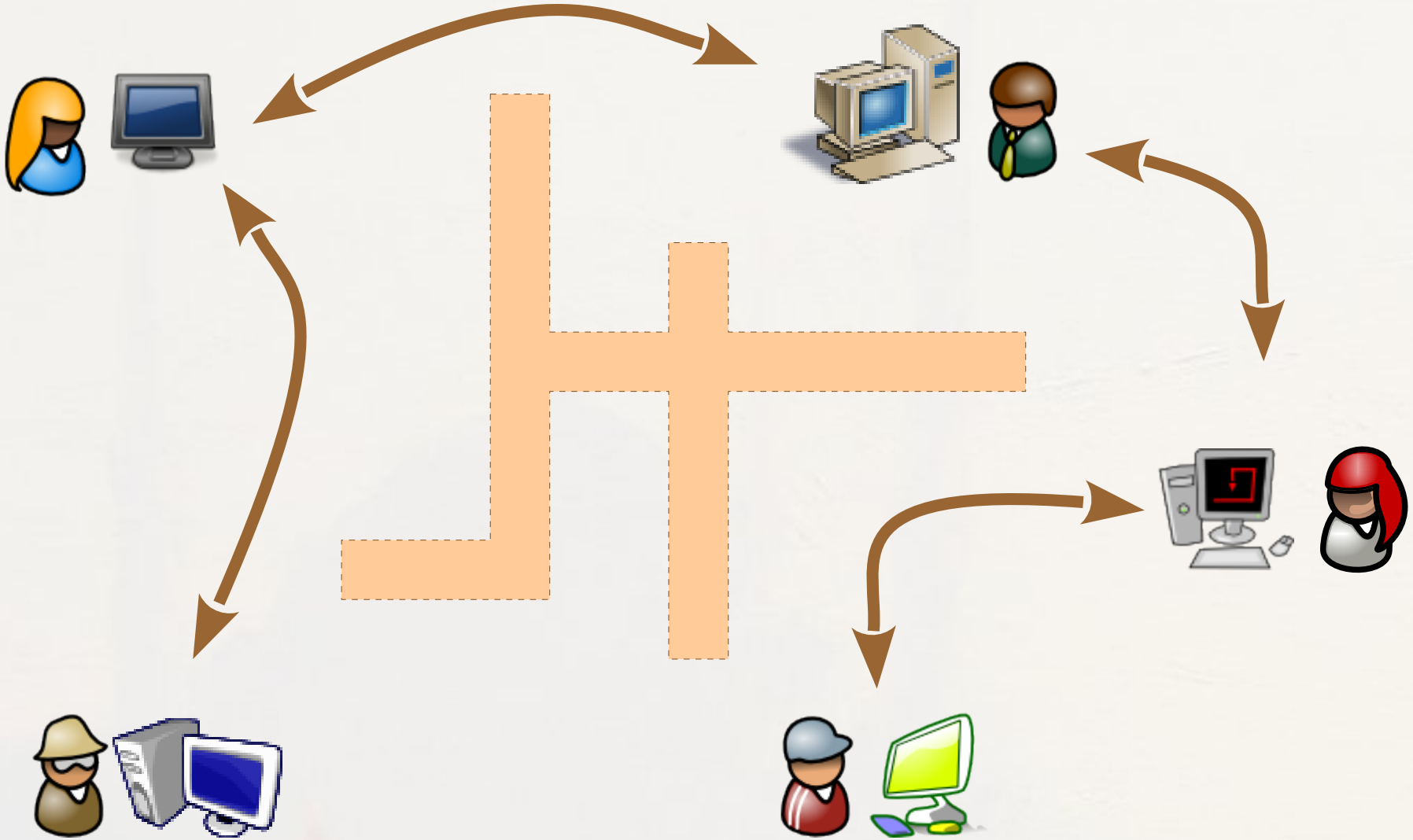
- Information extraction
- Mining
- Searching
- Matching
- Entity resolution
- Deep Web

Redes Sociais e Dados sobre nós

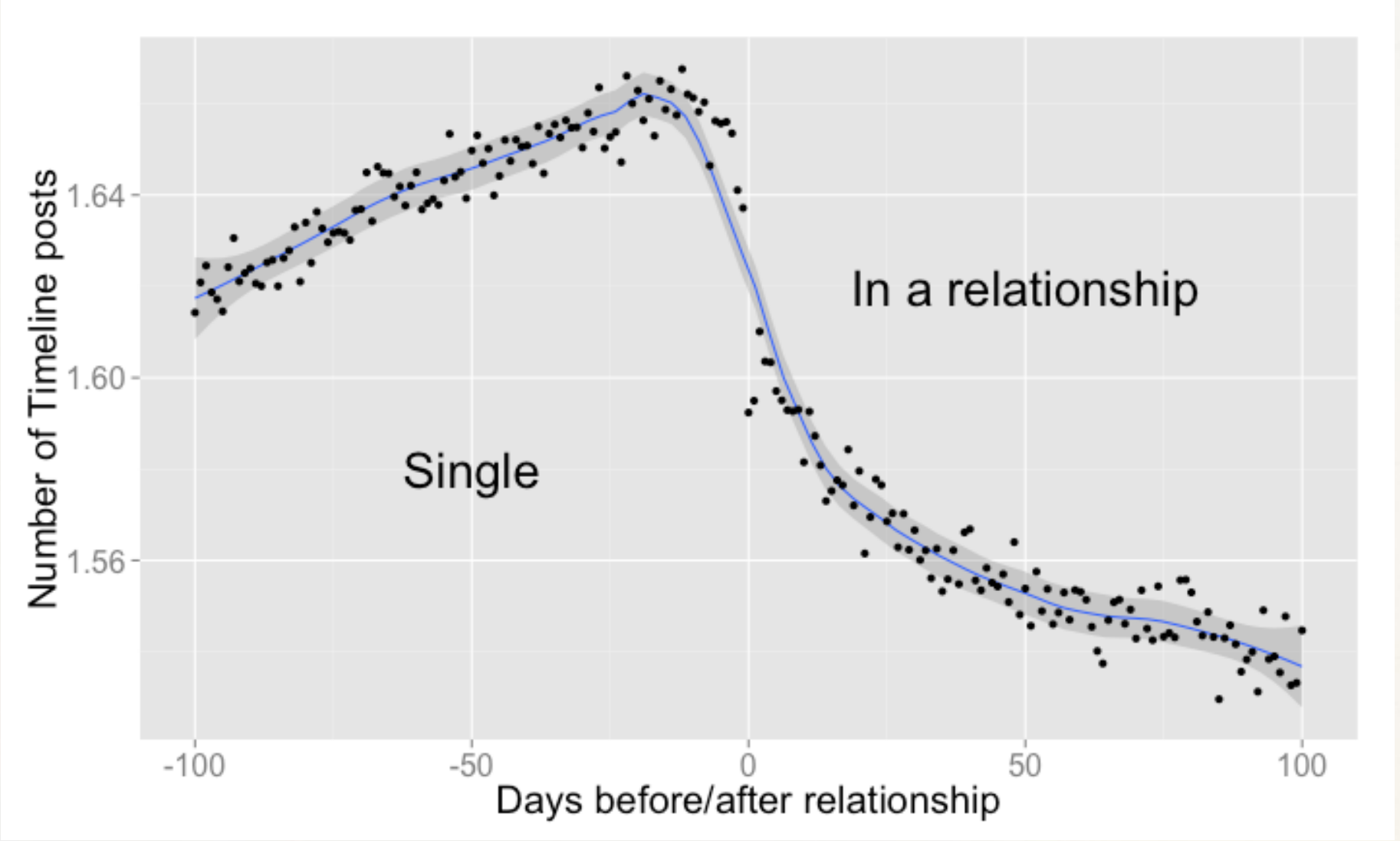
Mundo Físico



Mundo Virtual

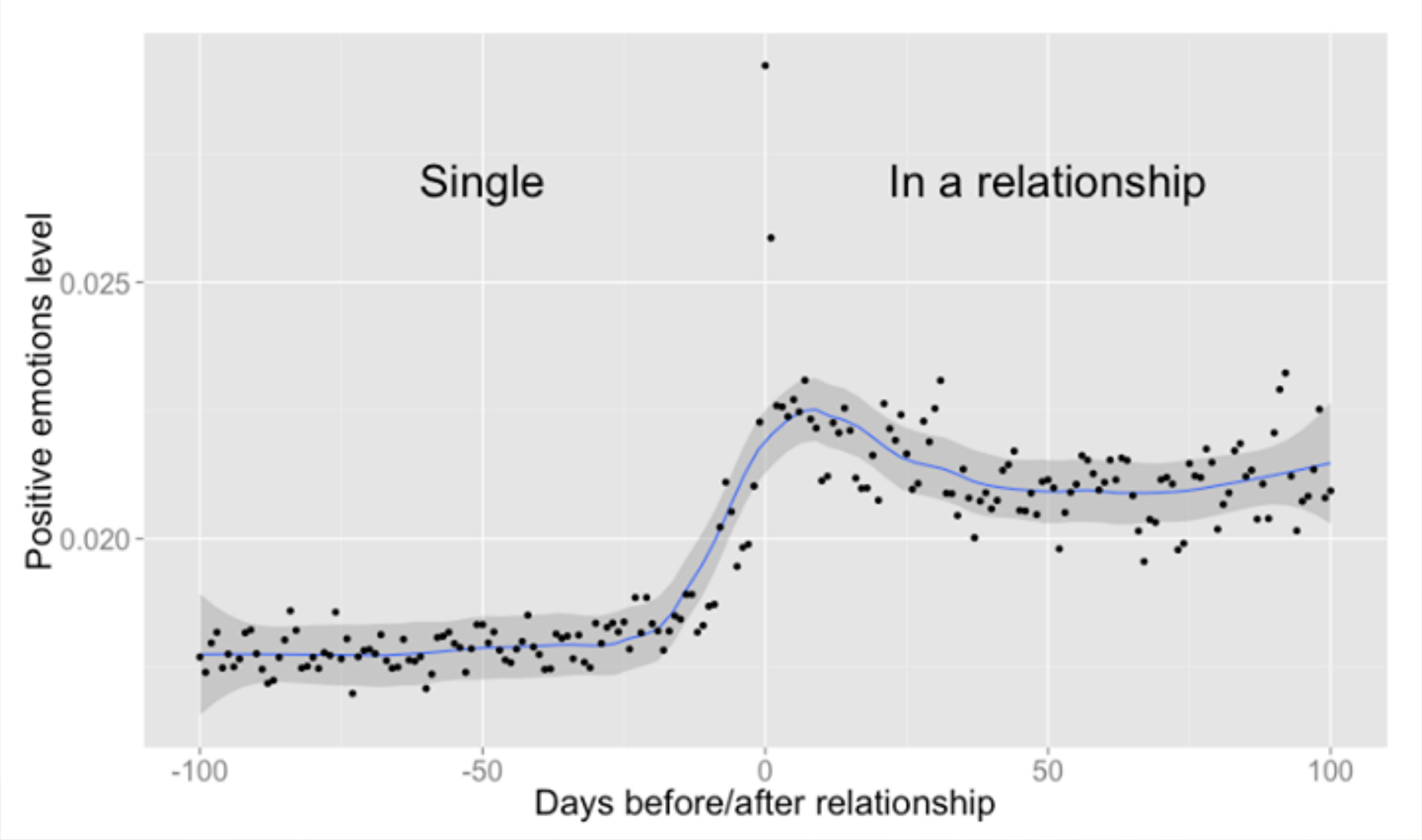


Facebook



(Diuk, 2014)

Facebook



(Diuk, 2014)

The Formation of Love

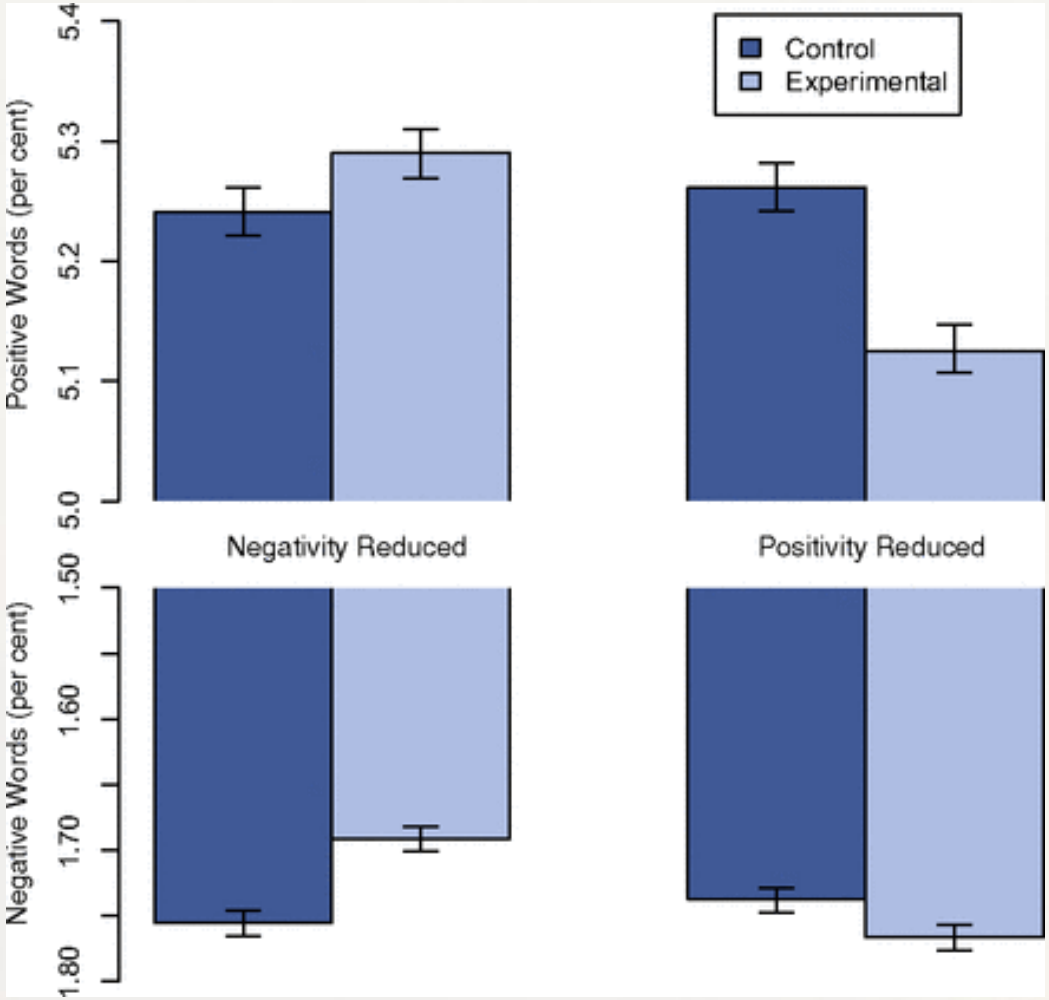
The Formation of Love

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. Kramera (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

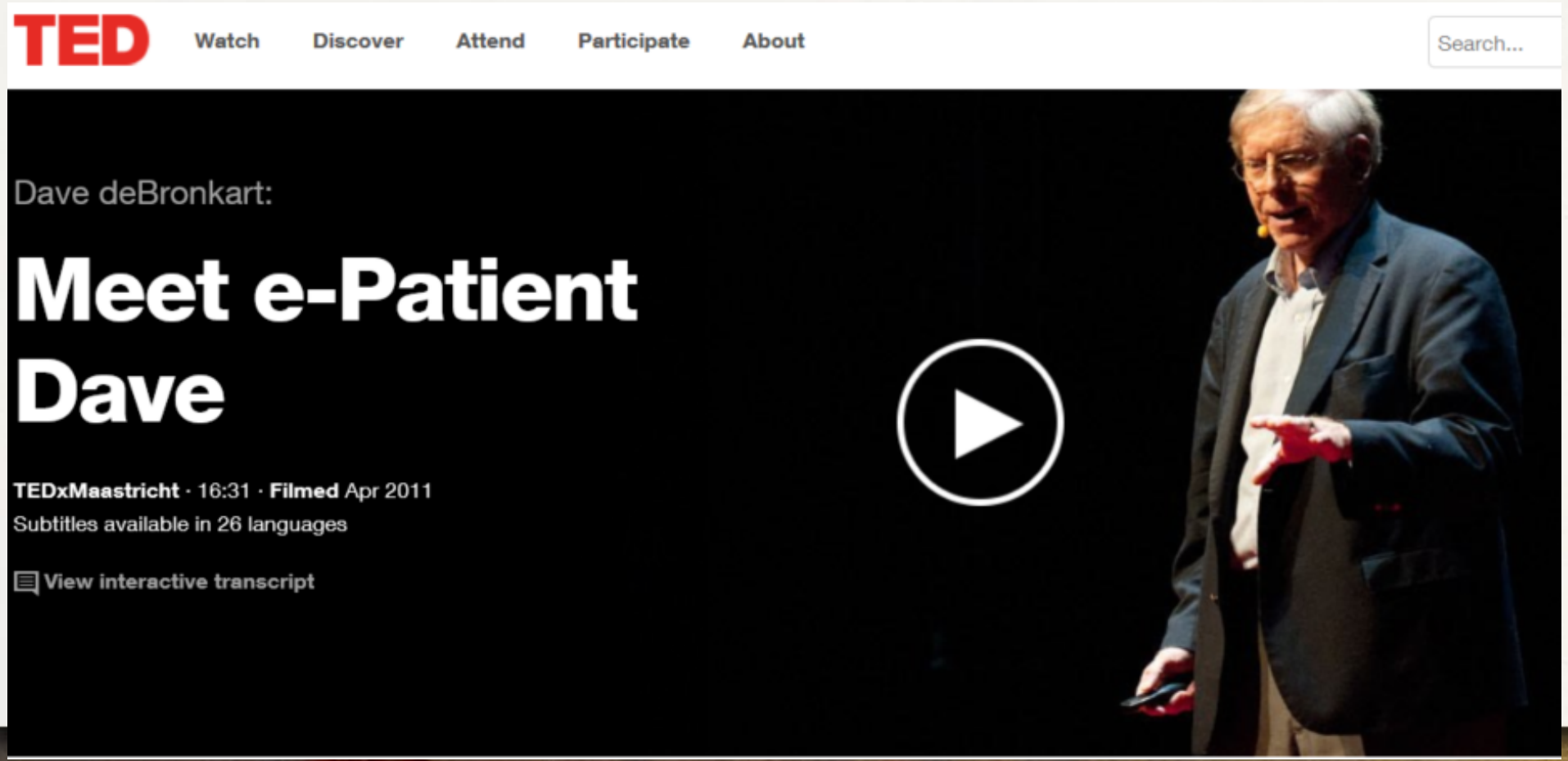
Minerando a Web & Saúde

Mining Data for Better Medicine

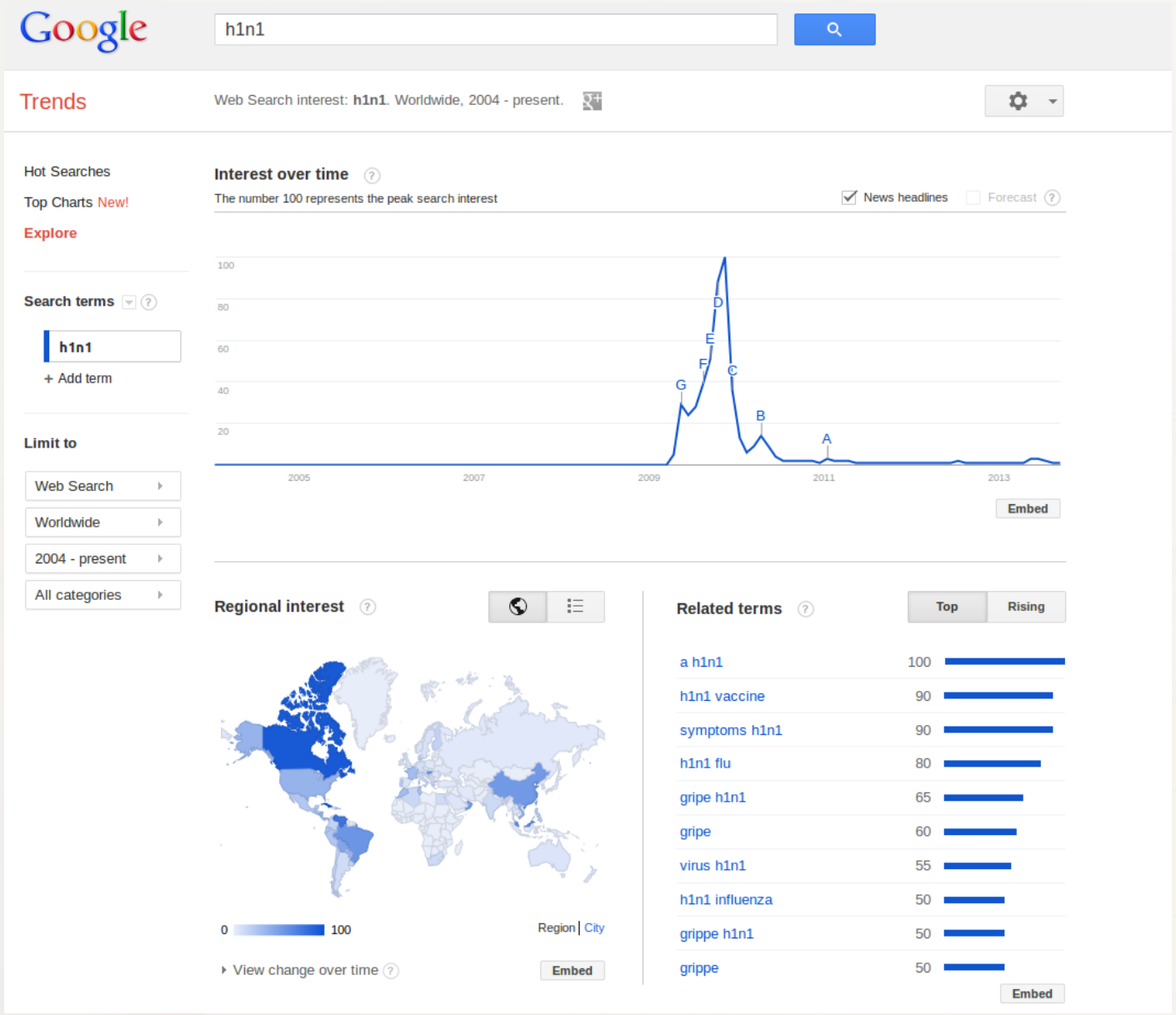
- Mining Data for Better Medicine
  Monday, September 19, 2011
  By Neil Savage
  http://www.technologyreview.com/news/425466/mining-data-for-better-medicine/
- “The health battles of millions, recorded digitally, open a world of virtual research.”

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



Google Trends



Google Flu Trends

https://www.google.org/flutrends/

google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

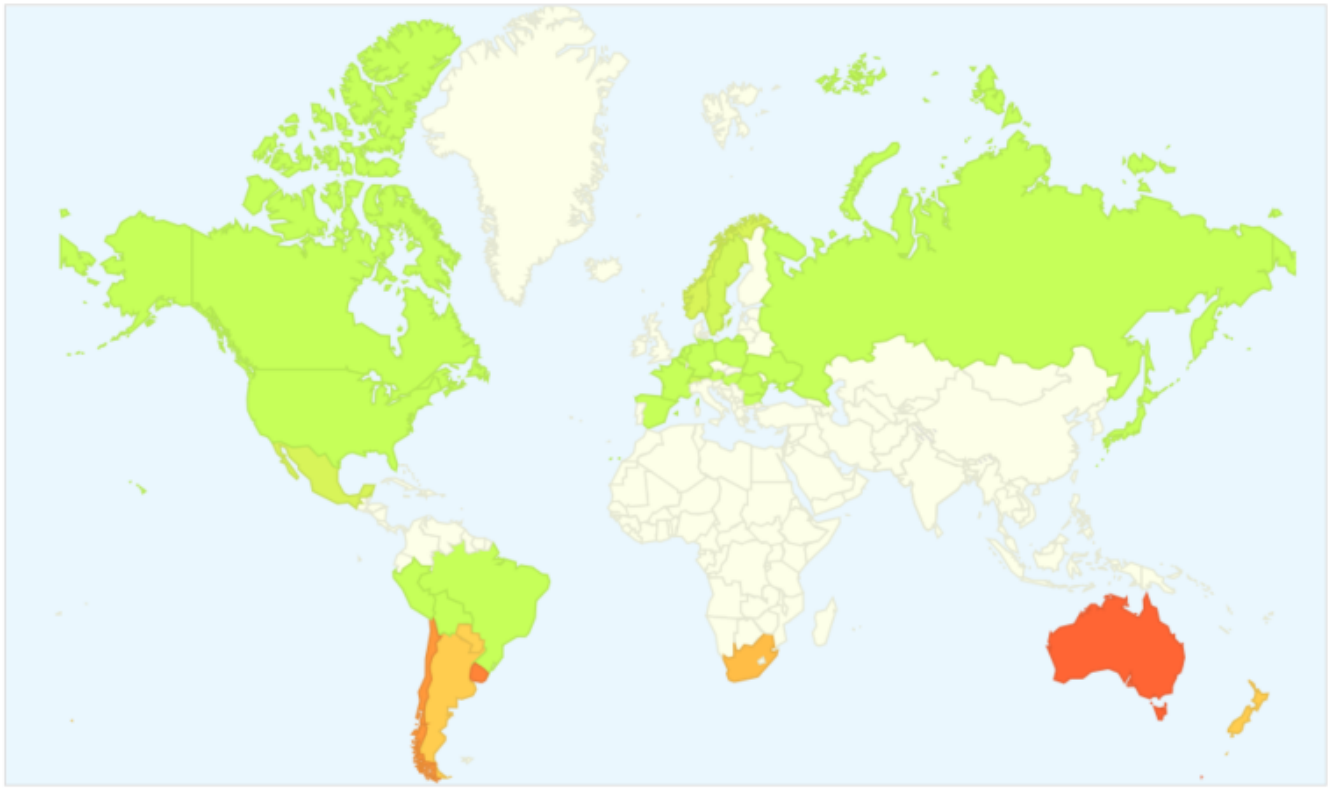
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

Google Dengue Trends

https://www.google.org/denguetrends/

google.org Dengue Trends

Google.org home

Flu Trends

Dengue Trends

Home

Select country/region

How does this work?

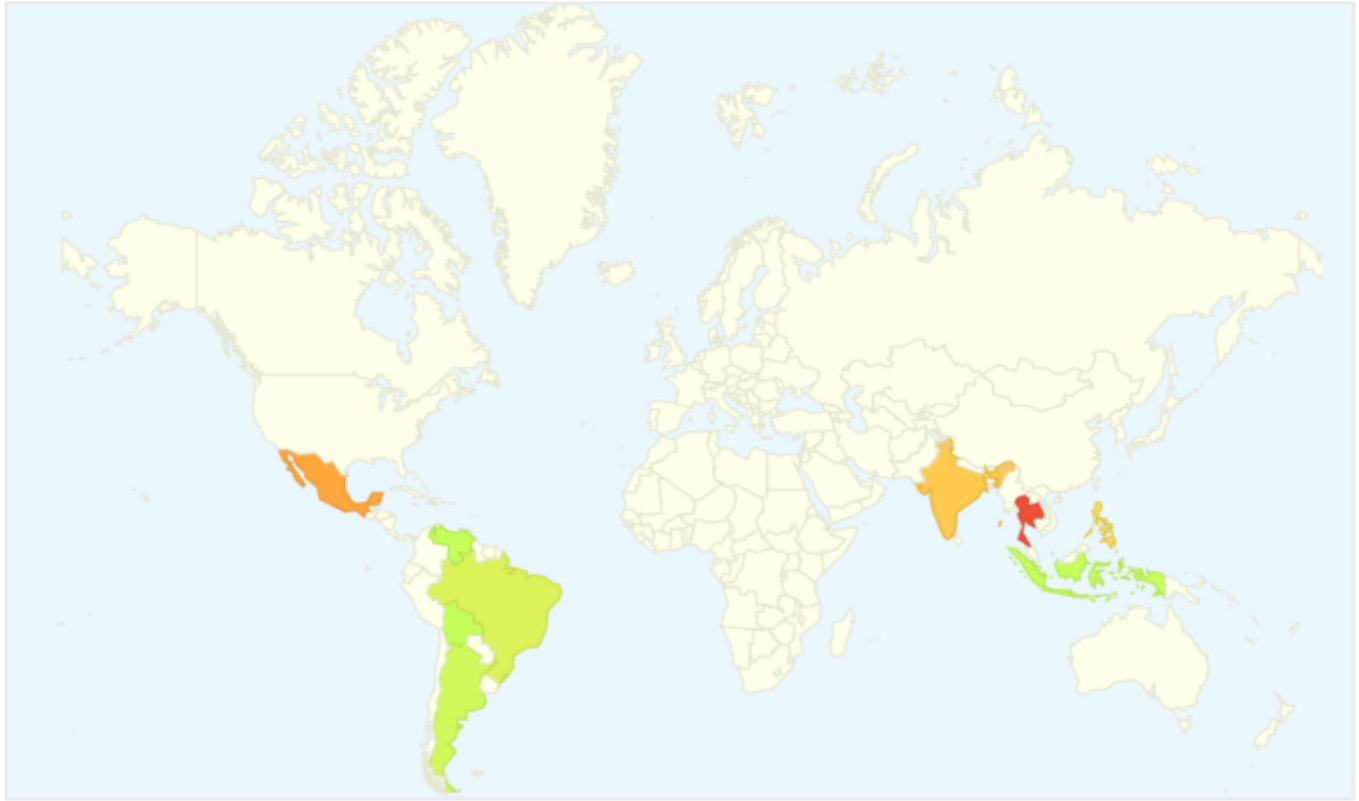
FAQ

Dengue activity

- Intense
- High
- Moderate
- Low
- Minimal

Dengue trends around the world

We've found that certain search terms are good indicators of dengue activity. Google Dengue Trends uses aggregated Google search data to estimate dengue activity. Learn more »

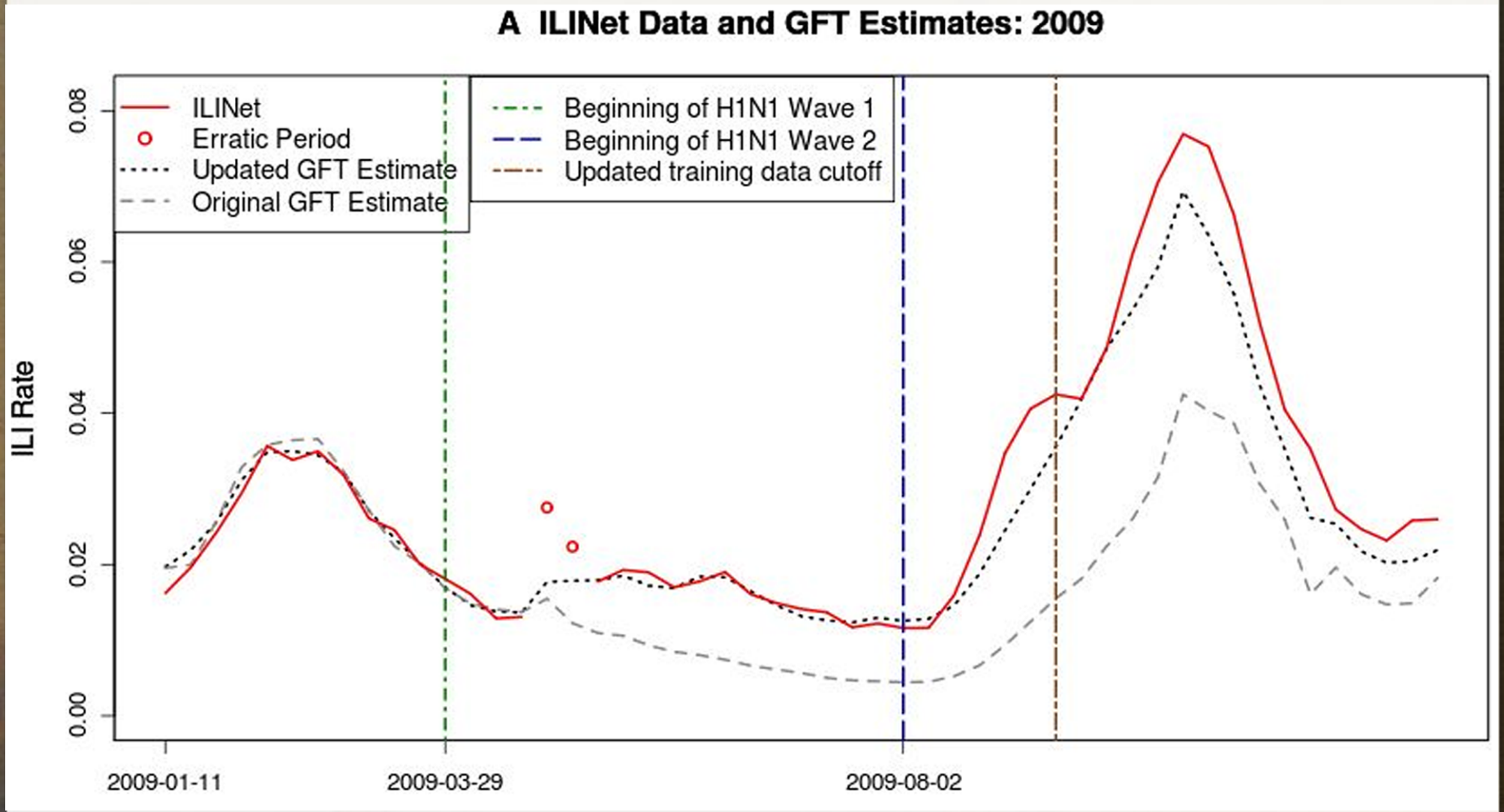


Download world dengue activity data

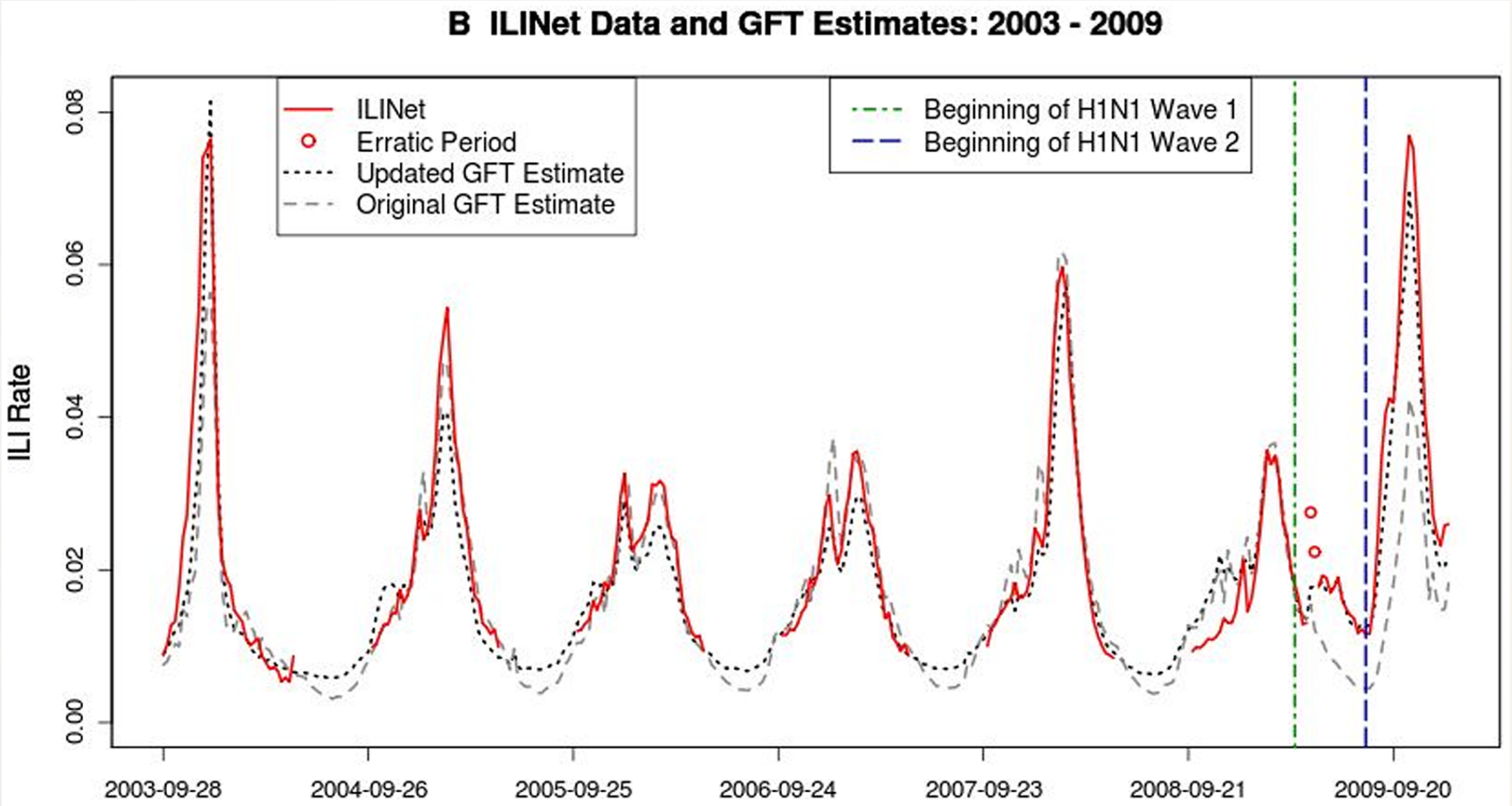
Google Trends & H1N1

- **Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic**
  by Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, Matthew H. Mohebbi
  PLOS, August 19, 2011
  http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023610

Google Trends & H1N1



Google Trends & H1N1



Web Observatory

Web Observatory

- INCT INWeb
  - http://observatorio.inweb.org.br/
  - Elections Observatory
  - Brasileirão Observatory
  - Dengue Observatory

Recomendação

André Santanchè

See On Pinterest

Linked Data

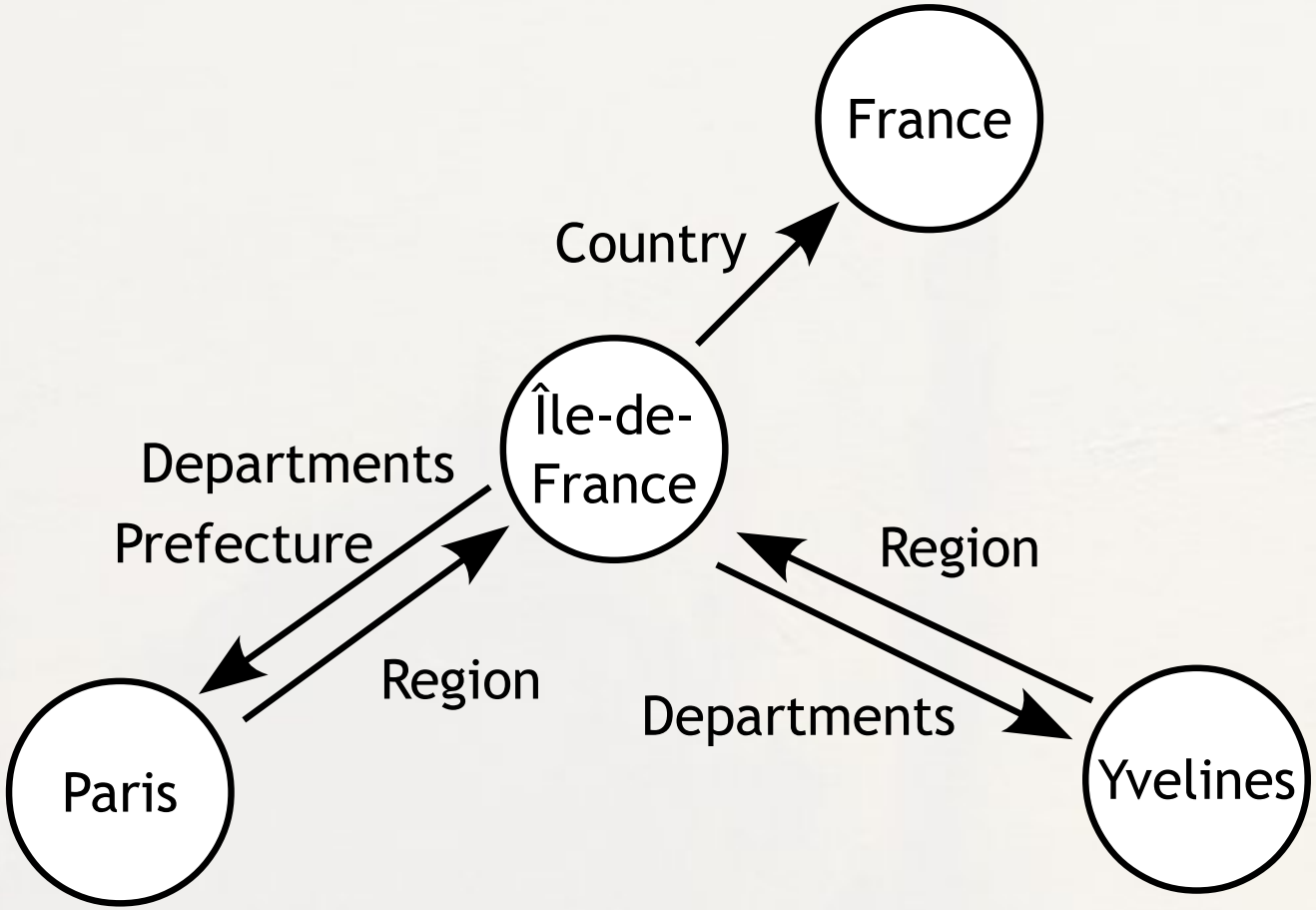
Wikipedia

Country	France
Region	Île-de-France
Department	**Paris**
Subdivisions	20 arrondissements
Government	
• **Mayor** (2008–14)	Bertrand Delanoë (PS)
Area[1]	
• **Urban** (2010)	2,844.8 km² (1,098.4 sq mi)
• **Metro** (2010)	17,174.4 km² (6,631.1 sq mi)
• **Land**[1]	105.4 km² (40.7 sq mi)
Population (2010)[5]	
• **Rank**	1st in France

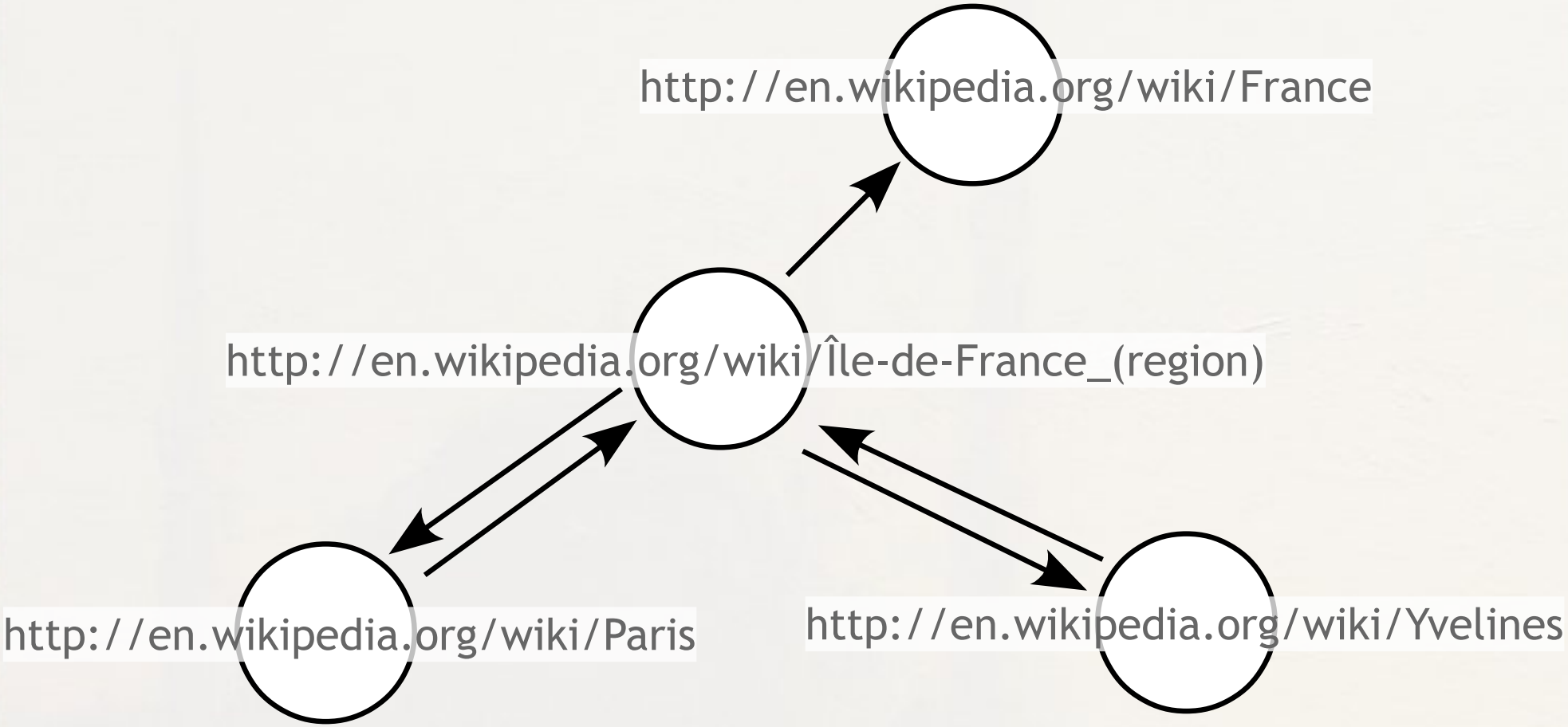
Country	France
Prefecture	Paris
Departments	8 [hide]
	Paris
	Essonne
	Hauts-de-Seine
	Seine-Saint-Denis
	Seine-et-Marne
	Val-de-Marne
	Val-d'Oise
	Yvelines
Government	
• **President**	Jean-Paul Huchon (PS)
Area	
• **Total**	12,012 km² (4,638 sq mi)
Population (2012)[1]	
• **Total**	11,914,812
• **Density**	990/km² (2,600/sq mi)

Infobox

DBPedia


France
Country
Île-de-
France
Departments
Prefecture
Region
Region
Departments
Paris
Yvelines

DBPedia (URIs)



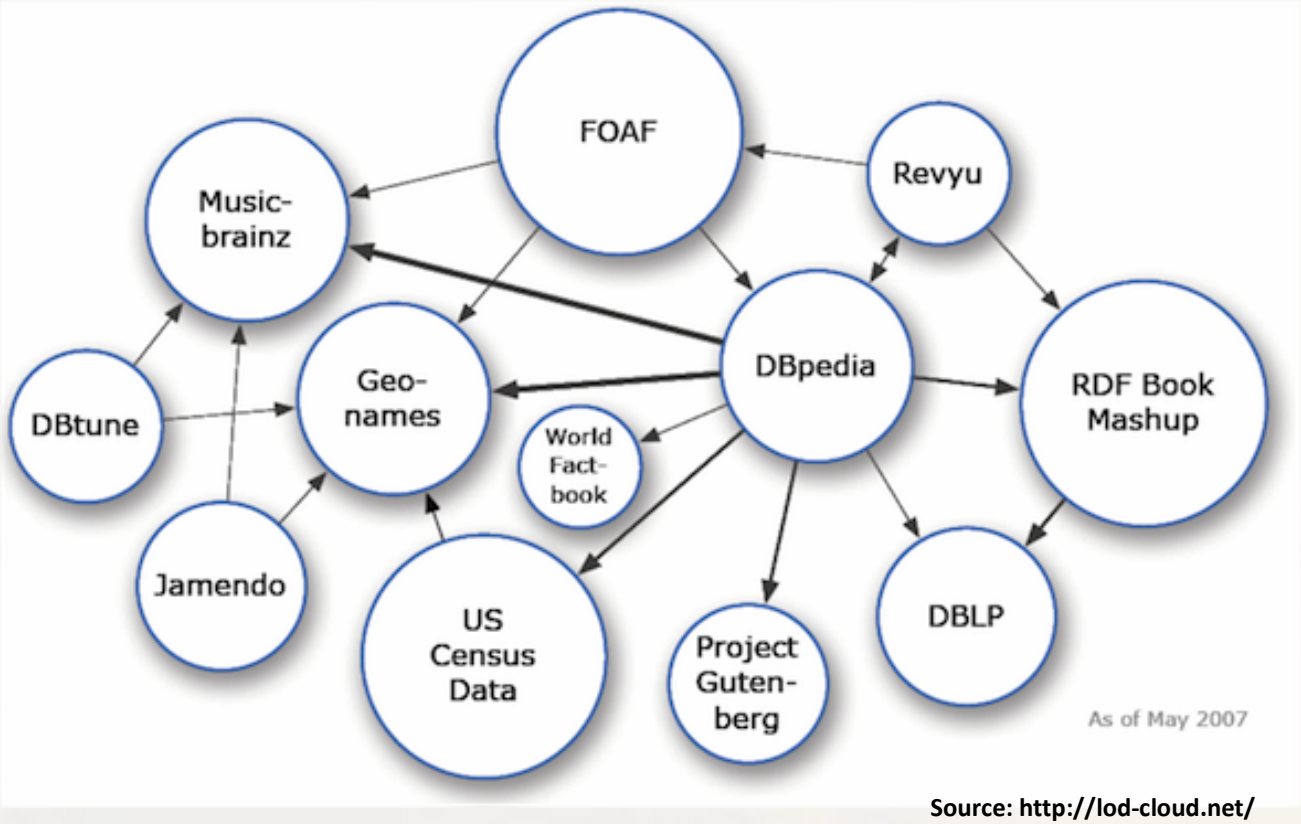
DBPedia – English

- **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

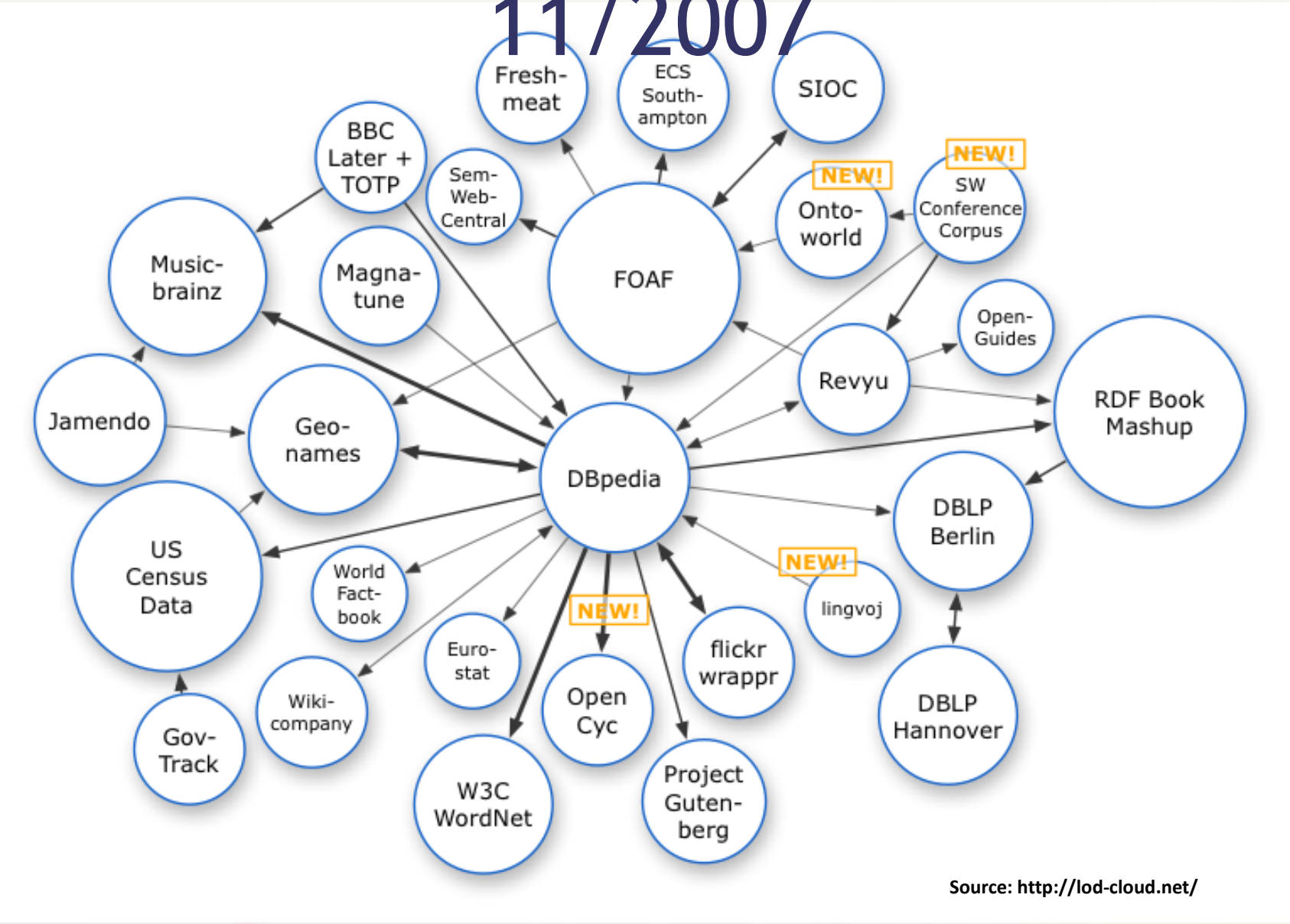
Linked Data
05/2007



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

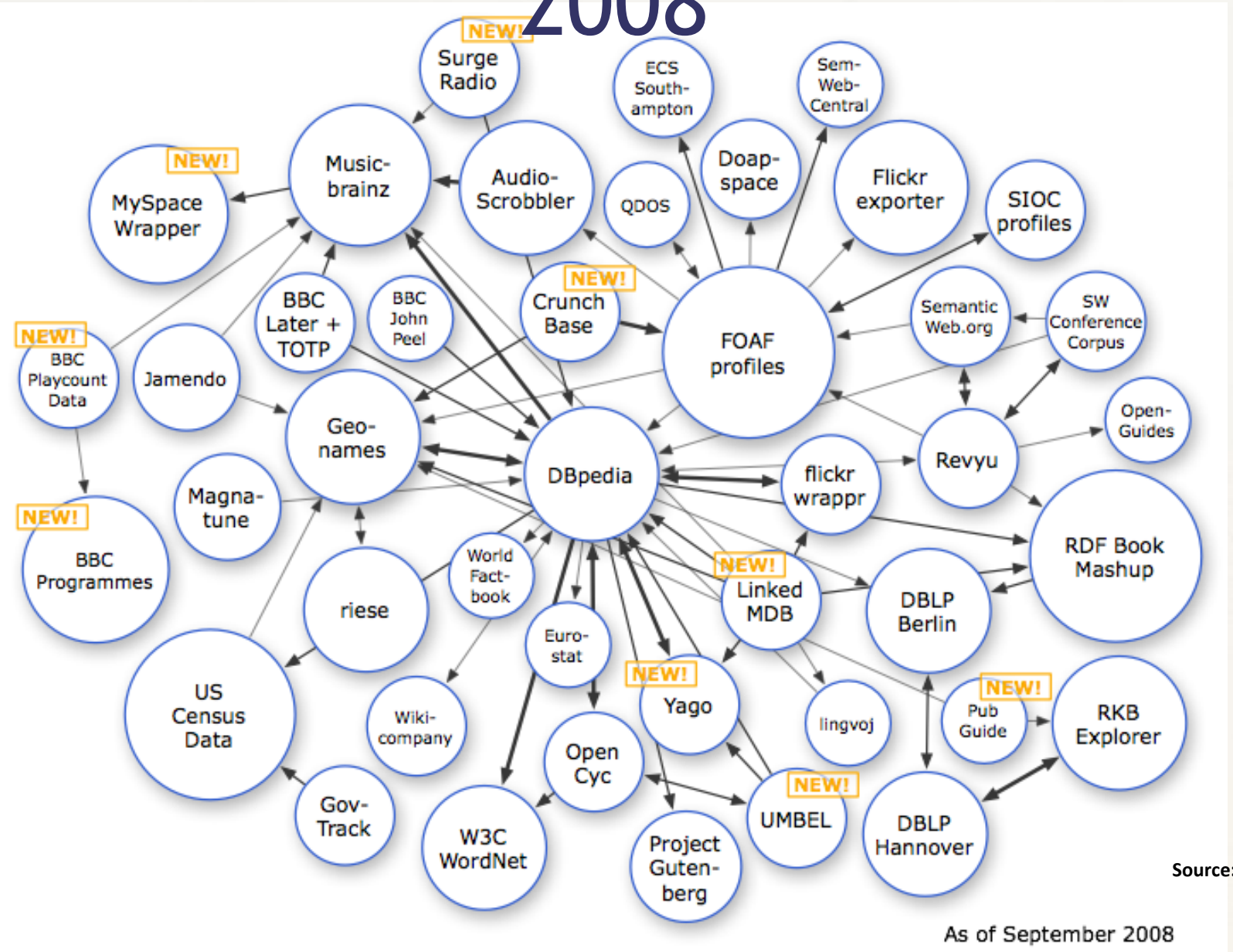
Datasets published following Linked Data 'format': **05/2007**

Linked Data 11/2007



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

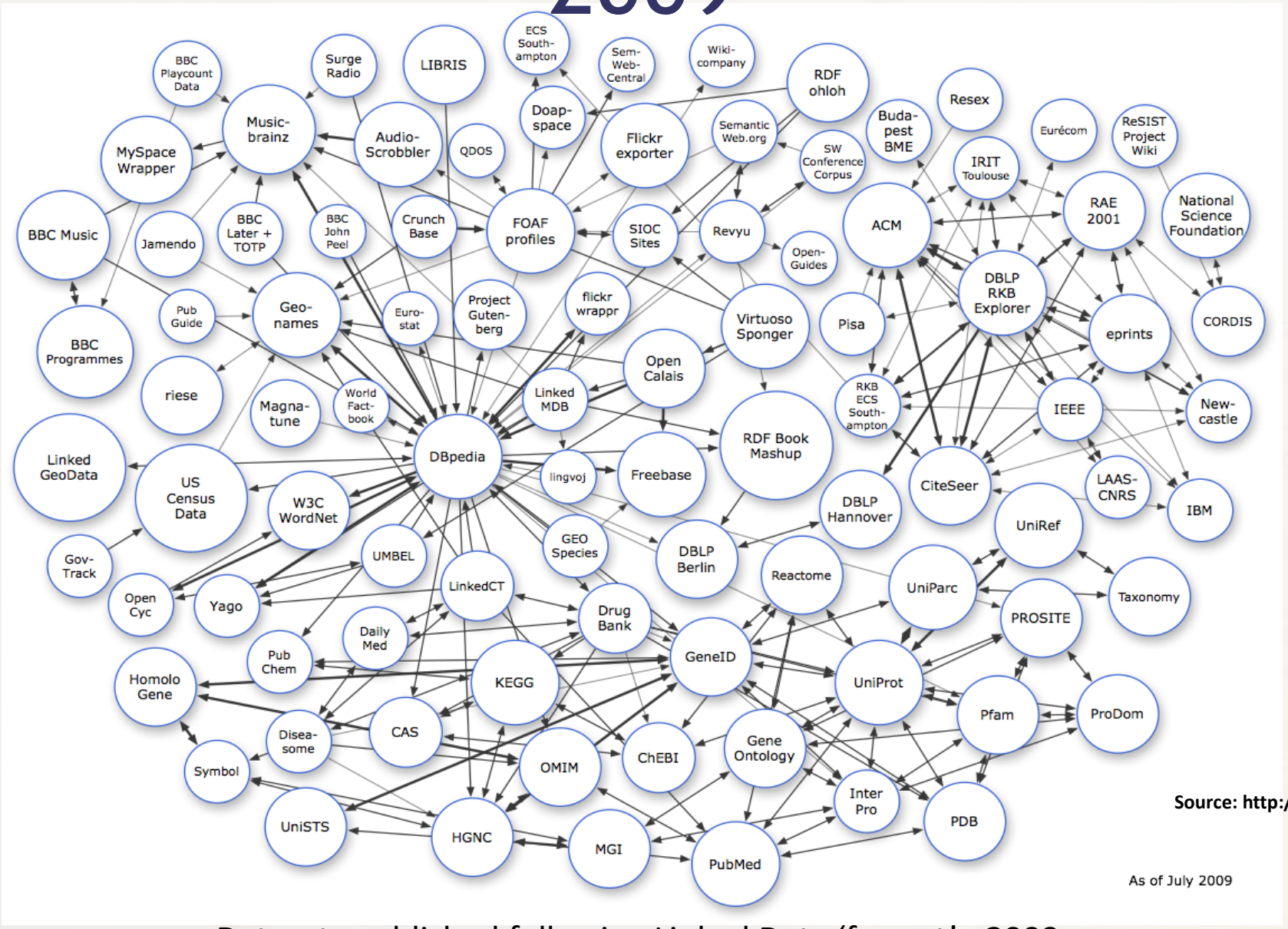
Linked Data 2008



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

Datasets published following Linked Data 'format': **2008**

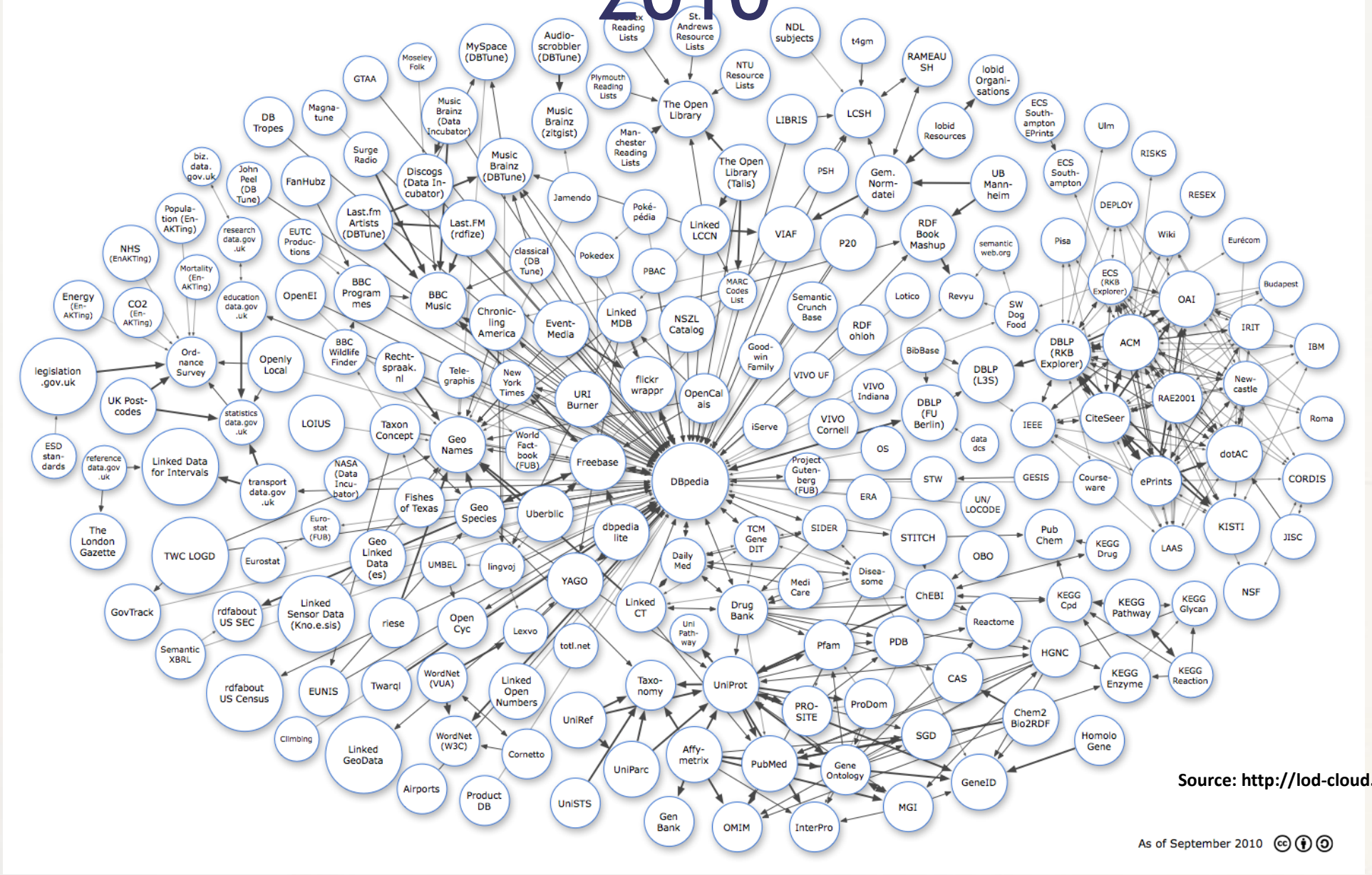
Linked Data
2009



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

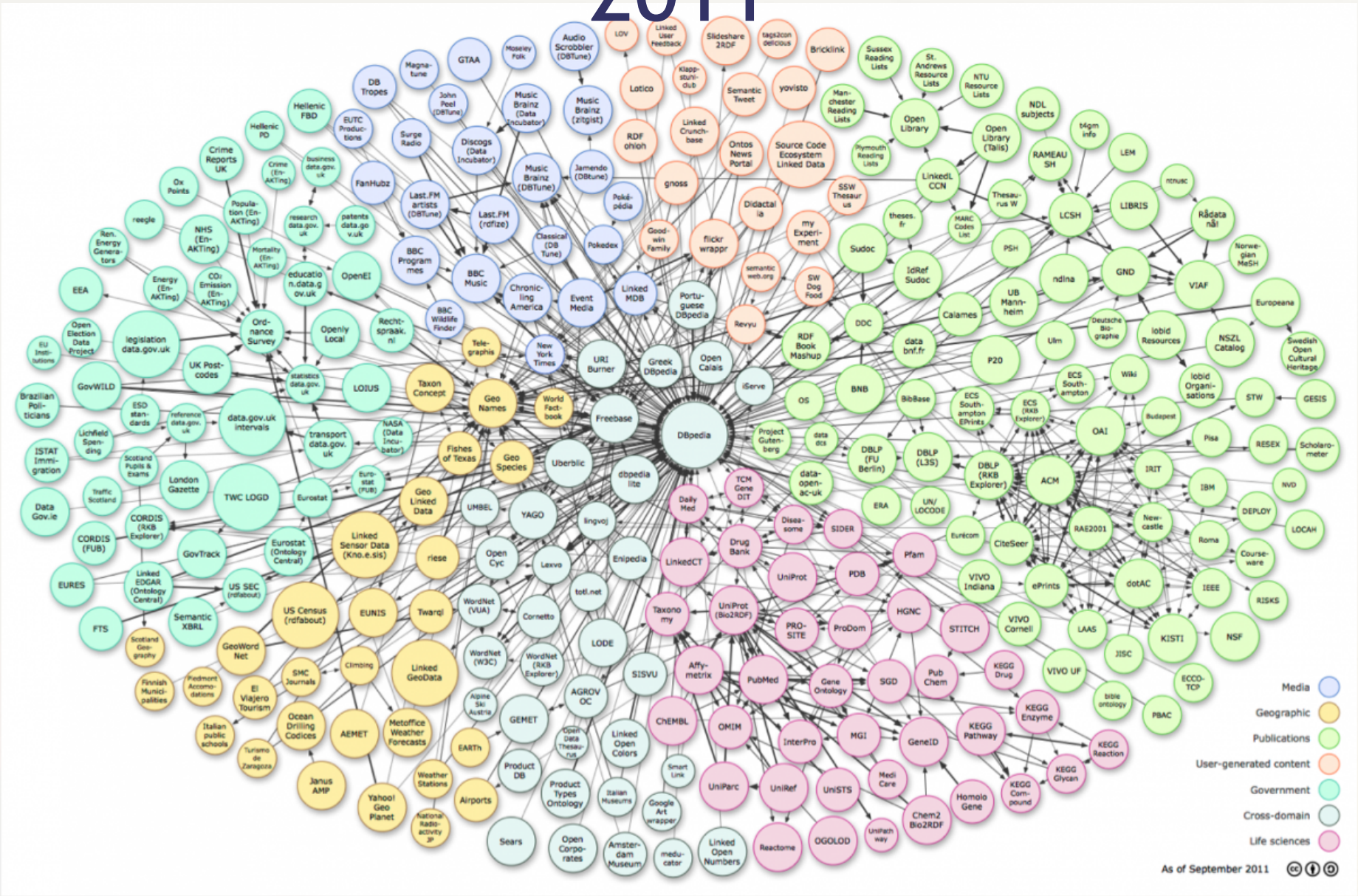
Datasets published following Linked Data ‘format’: **2009**

Linked Data 2010



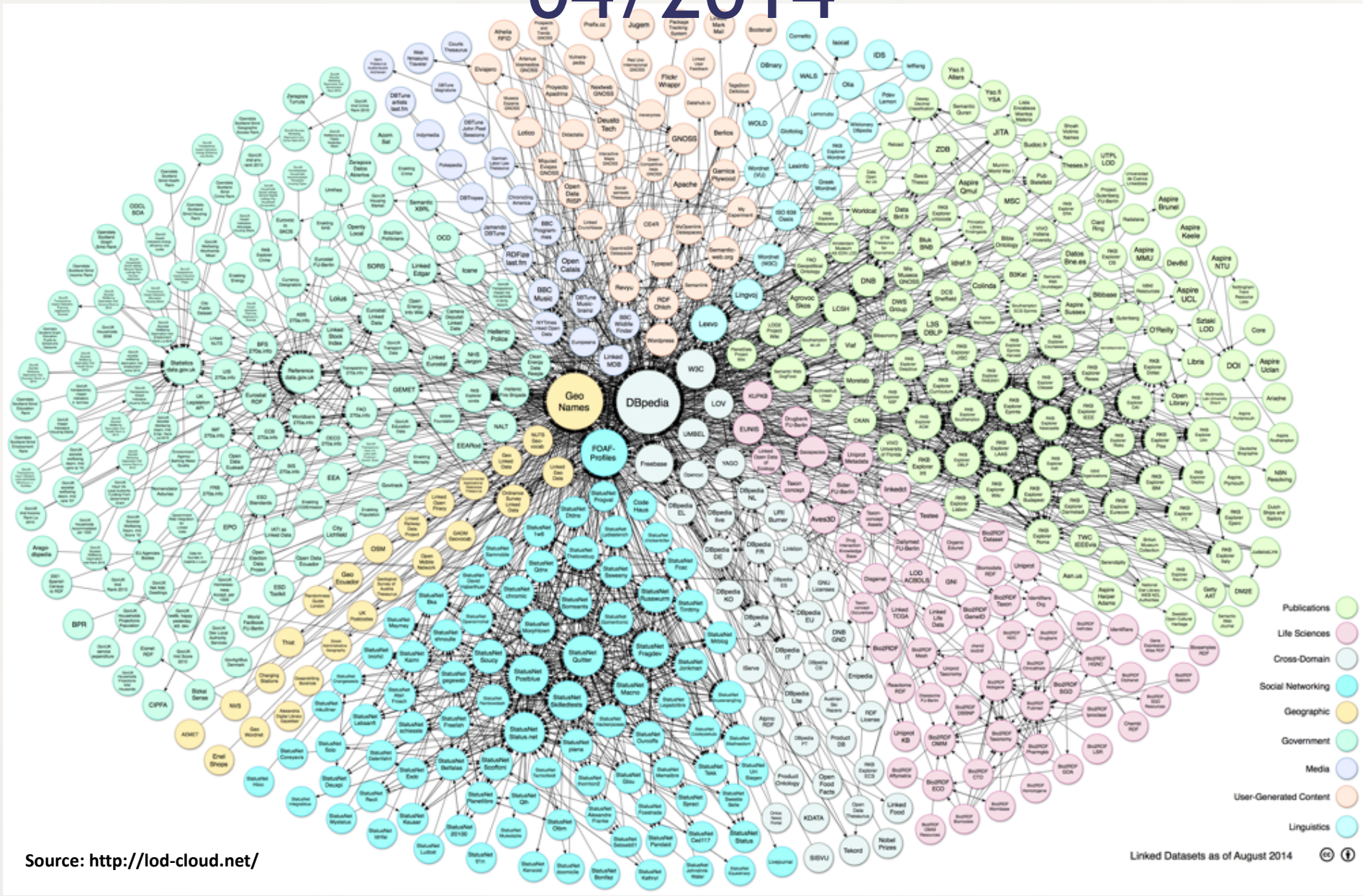
Datasets published following Linked Data ‘format’: **2010**

Linked Data 2011



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

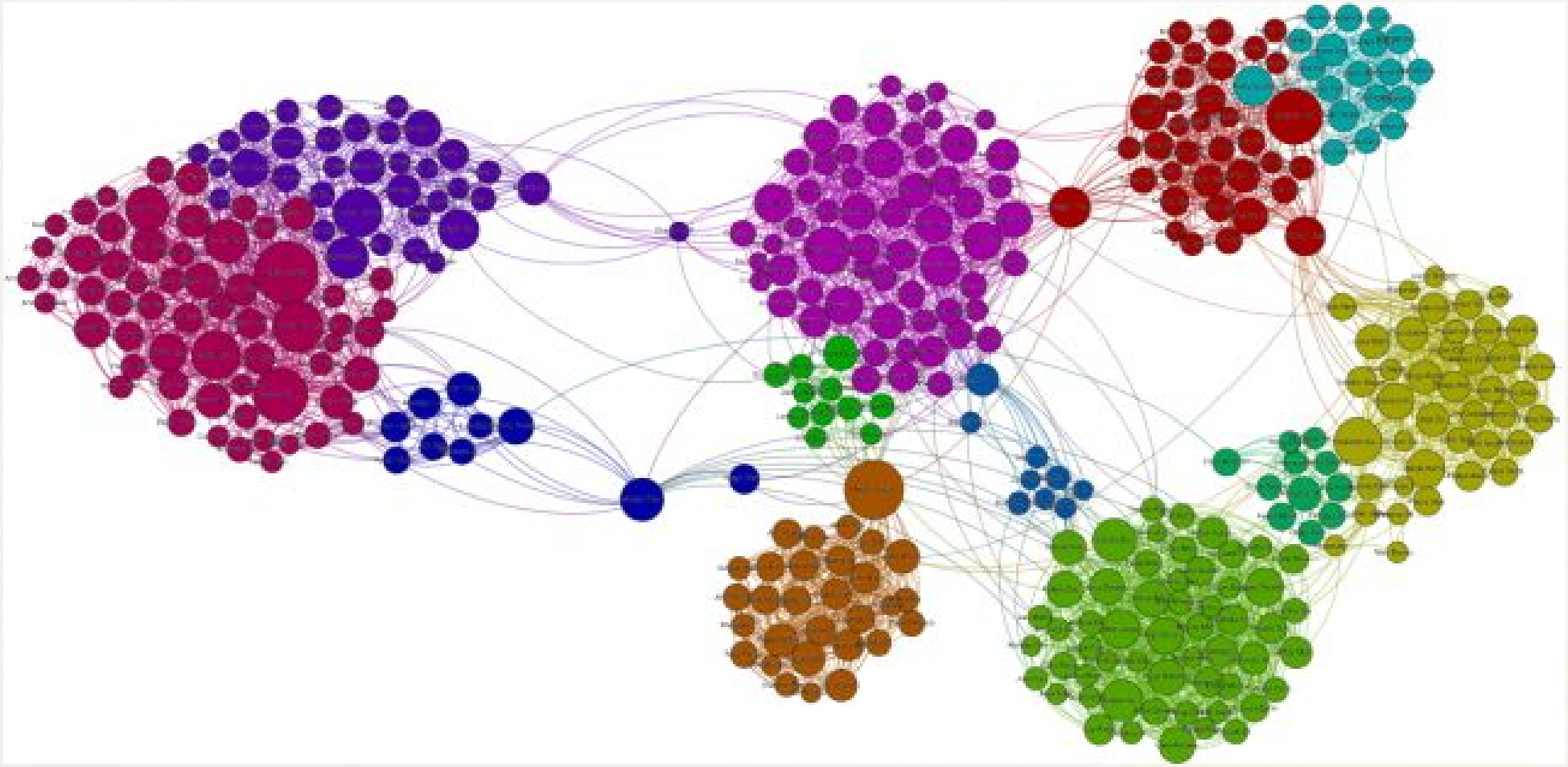
Linked Data
04/2014



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

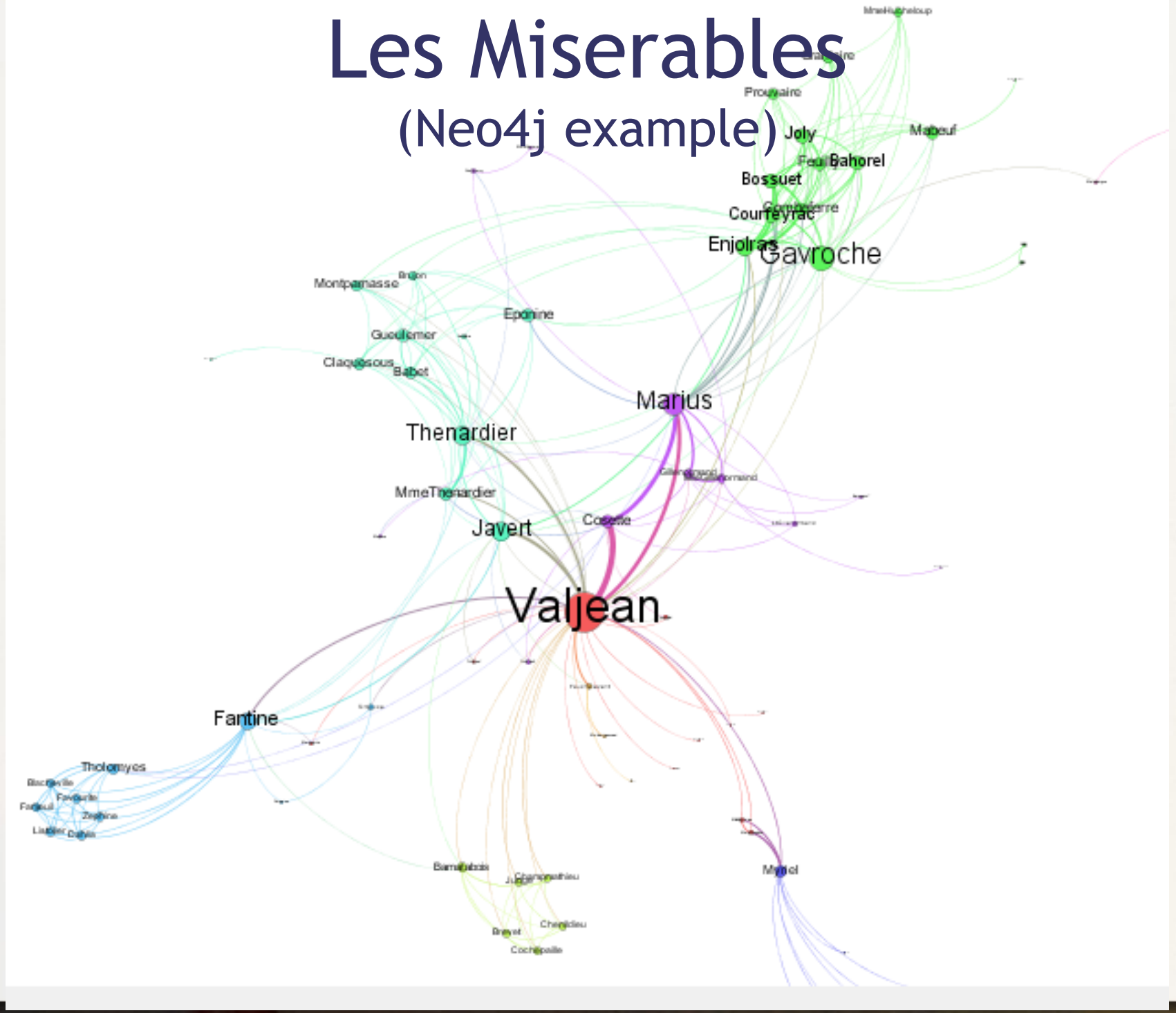
Dados em forma de rede
&
Network Science

Redes Sociais

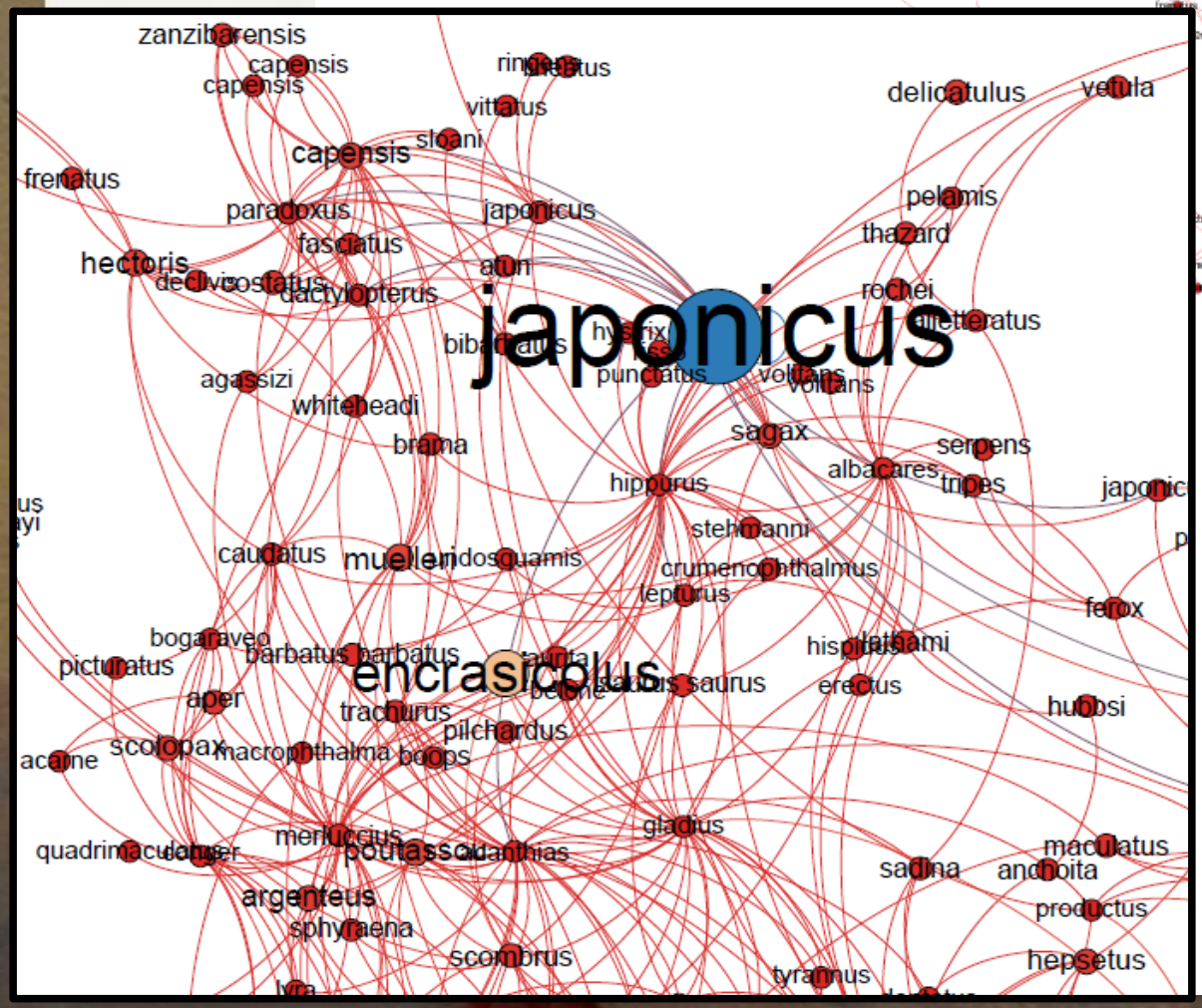


Les Miserables

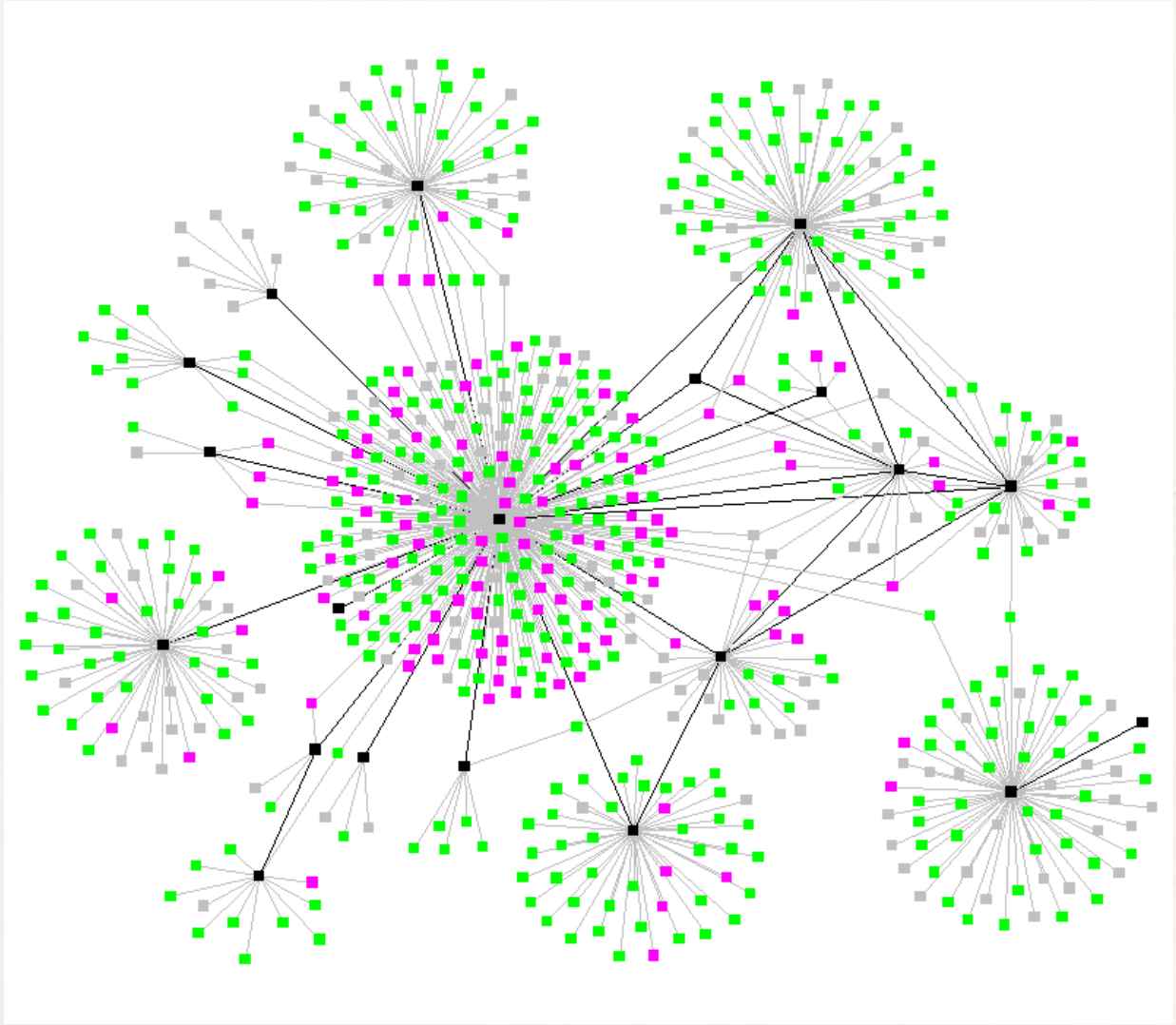
(Neo4j example)



Cadeia Alimentar FishBase

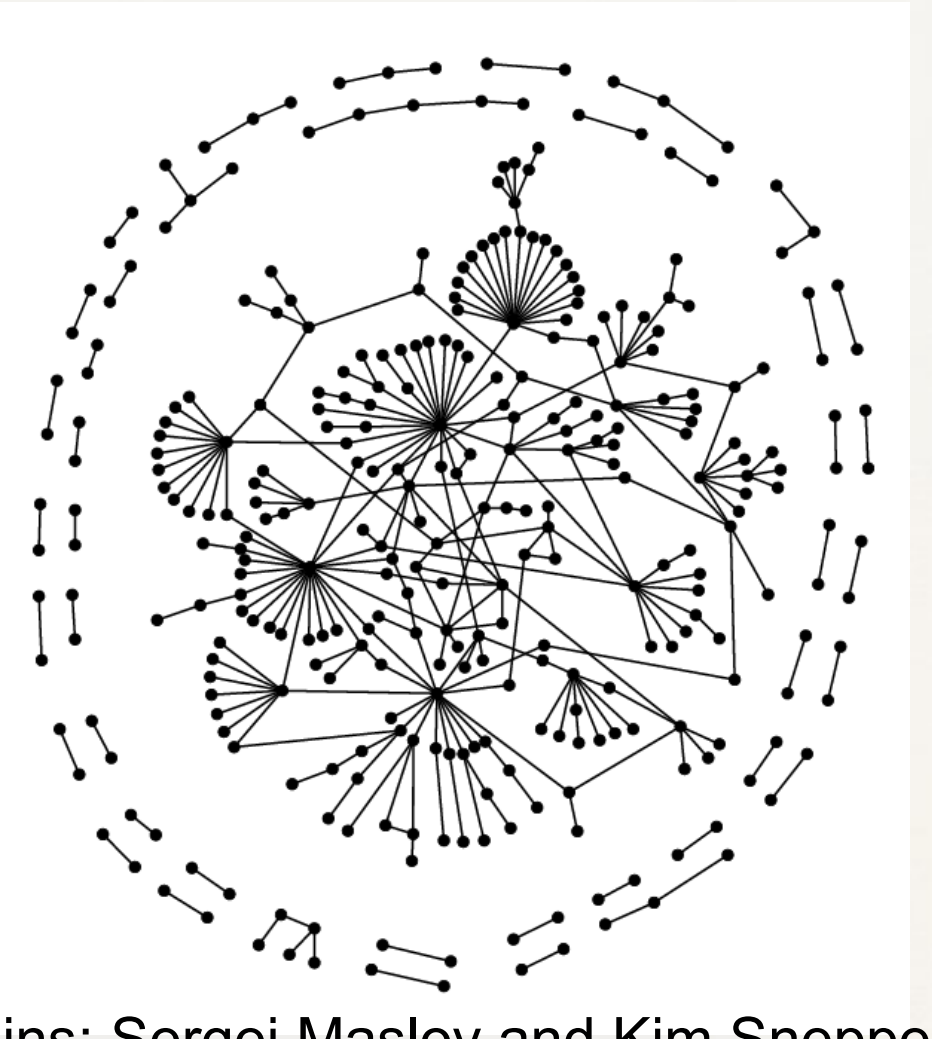


Contágio da TB



Contagion of TB, books on politics: Valdis Krebs, www.orgnet.com.

Proteínas da Levedura

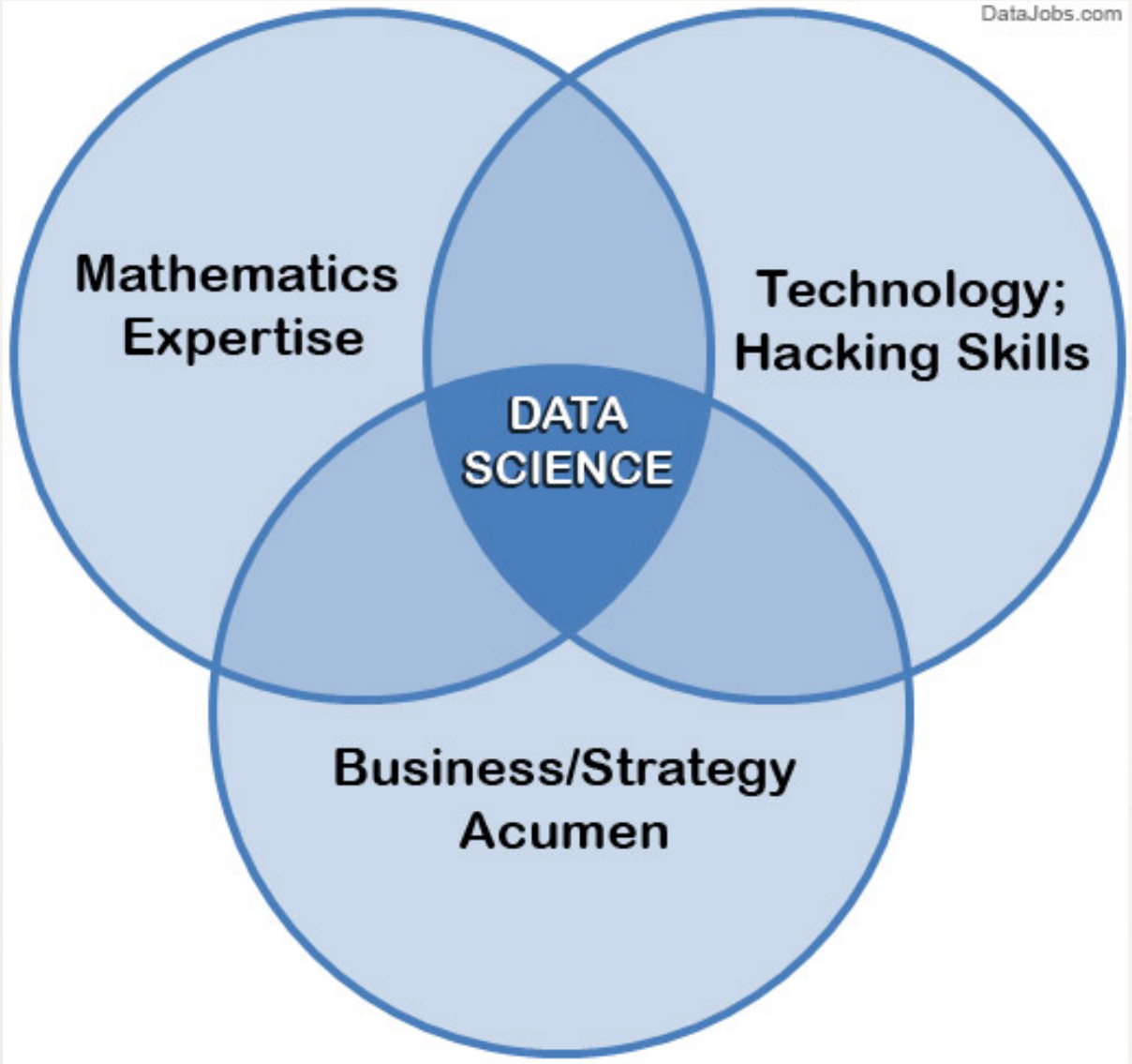


Yeast proteins: Sergei Maslov and Kim Sneppen,
Specificity and stability in topology of protein networks,
Science 296, 910-913 (2002).

Data Scientist

What is Data Science?

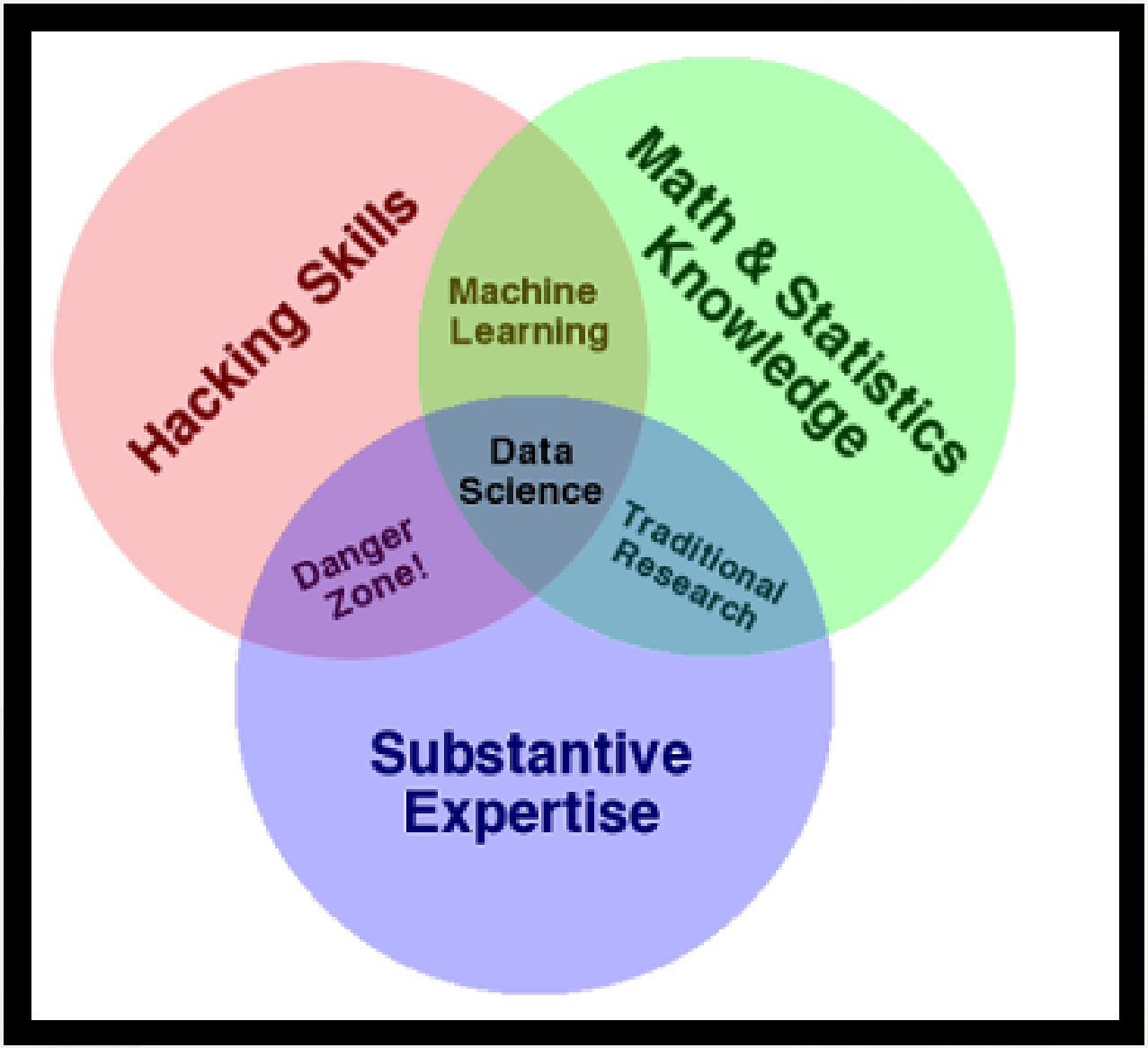
https://datajobs.com/what-is-data-science



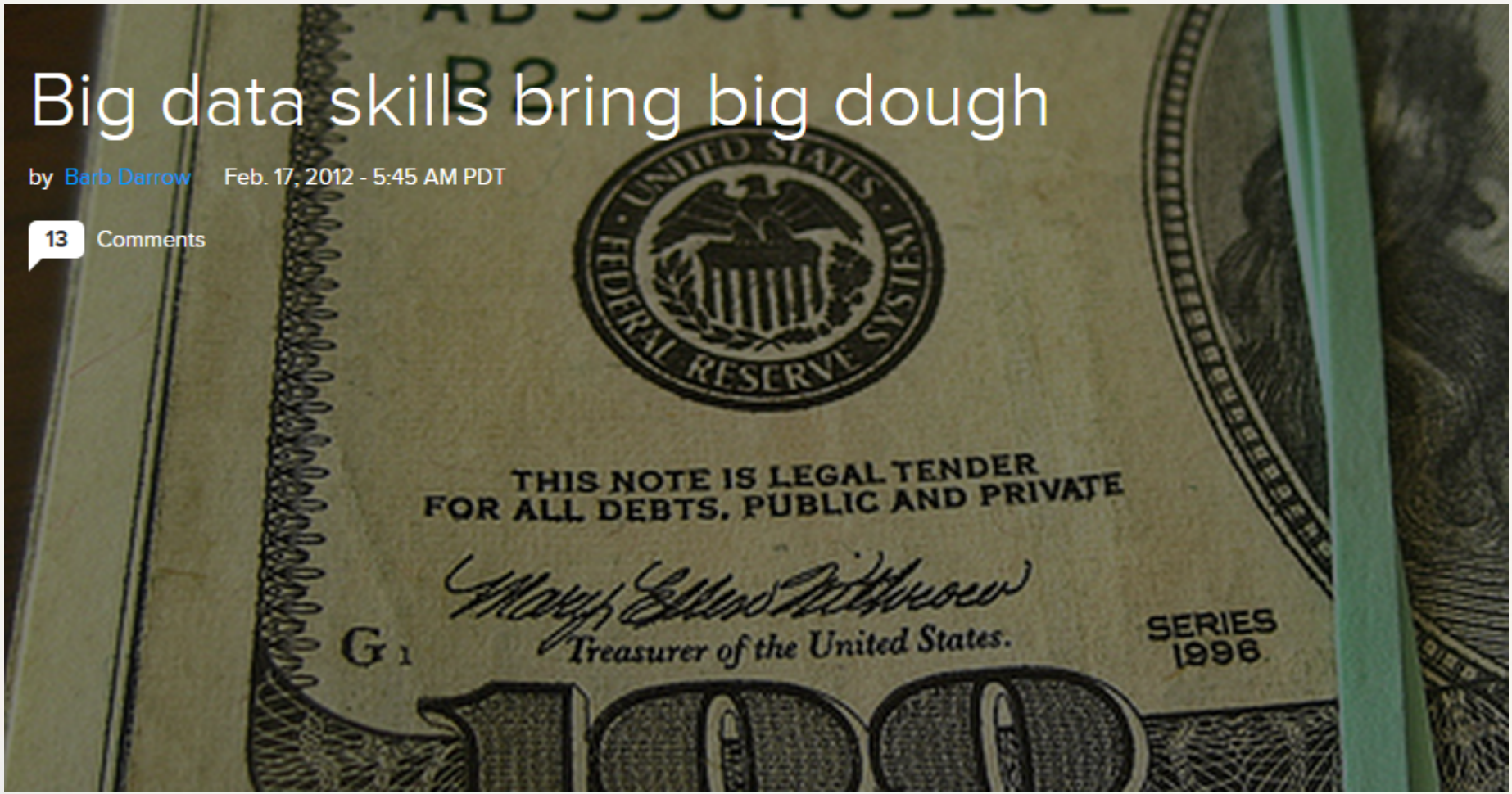
So you Want to be a Data-Scientist

Michael Spencer – 21/07/2015

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



Remuneração



https://gigaom.com/2012/02/17/big-data-skills-bring-big-dough/

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, 3rd edition.

Agradecimentos

- Luiz Celso Gomes Jr (professor desta disciplina em 2014) pela contribuição na disciplina e nos slides.

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