

A Nova Ciência Baseada em Dados



Kelly Rosa Braghetto

Departamento de Ciência da Computação

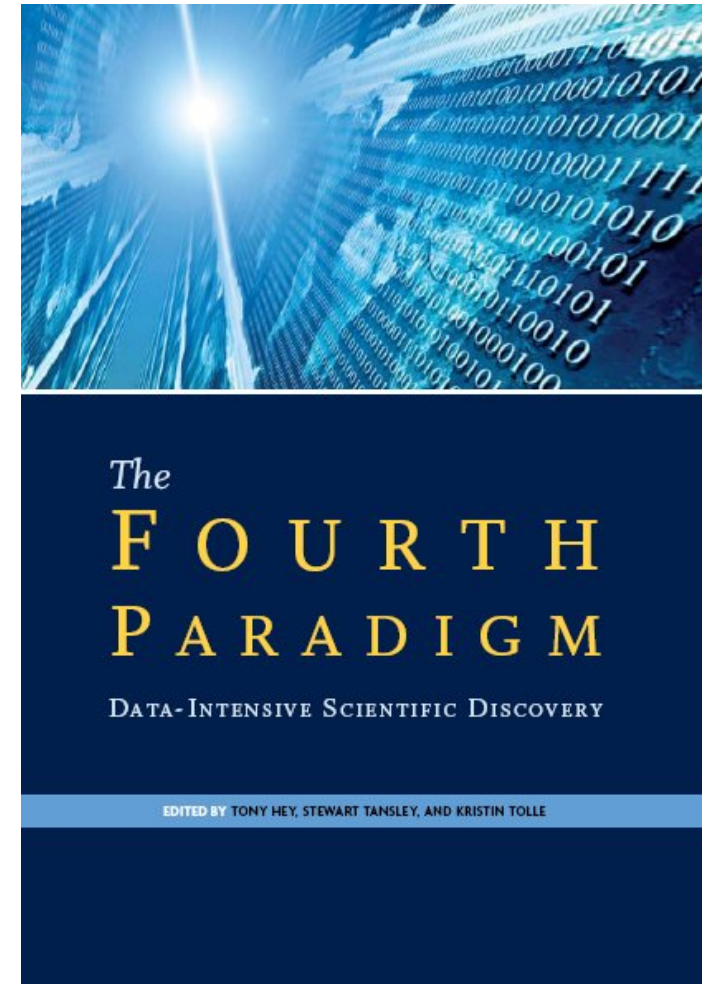
Instituto de Matemática e Estatística – Universidade de São Paulo

Dia do Bibliotecário 2015 – SIBi USP

PARADIGMAS DA CIÊNCIA

Segundo Jim Gray, estamos vivenciando um novo paradigma da ciência:

- Há milhares de anos, a ciência era **empírica**
- Há alguns séculos, a ciência passou a ser também **teórica** (modelos, generalizações, etc.)
- Nas últimas décadas, cientistas passaram a validar seus modelos teóricos com o uso de **simulações**
- **Quarto paradigma (atualidade): exploração de dados**



Um Novo Modo de Fazer Ciência

- Unifica teoria, experimentação e simulação
- “Toda ciência é ciência da computação” (NYT, 2001)
- **e-Science** (John Taylor, 2000) – *“colaboração global em áreas chaves da ciência e a próxima geração de infraestrutura que vai habilitá-la”*
 - Hoje, é entendida como: “pesquisa científica moderna feita por meio do uso intensivo da computação”

The New York Times

March 25, 2001

The World: In Silica Fertilization; All Science Is Computer Science

By GEORGE JOHNSON

EXCEPT for the fact that everything, including DNA and proteins, is made from quarks, particle physics and biology don't seem to have a lot in common. One science uses mammoth particle accelerators to explore the subatomic world; the other uses petri dishes, centrifuges and other laboratory paraphernalia to study the chemistry of life. But there is one tool both have come to find indispensable: supercomputers powerful enough to sift through piles of data that would crush the unaided mind.

<http://www.nytimes.com/2001/03/25/weekinreview/the-world-in-silica-fertilization-all-science-is-computer-science.html>

Um Novo Modo de Pensamento Científico

- A **exploração de dados** promove uma mudança importante no processo de pensamento científico

Antes:

**formulação de hipótese → experimentação →
análise de resultados**

Agora, “*data-driven hypothesis*”:

**formulação de hipótese →
busca da confirmação no banco de dados**

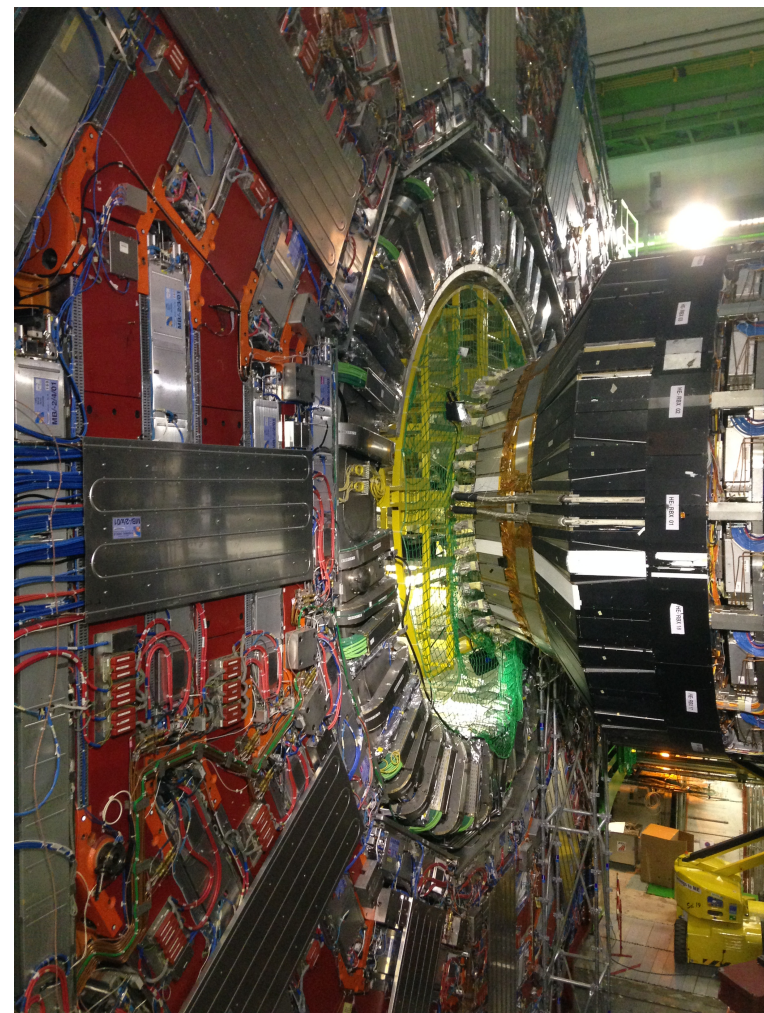
- Essa nova metodologia tem permitido a criação de novos “tipos” de pesquisa em diversas áreas do conhecimento (notadamente nas ciências naturais)

Volumes de Dados Gigantescos

- **Coletados nos mais diversos domínios:** astronomia, meteorologia, física, genômica, ciências sociais, ...
- Exemplo:

Large Hadron Collider (LHC)

- Produz cerca de 25 petabytes/ano
- Processa cerca de um petabyte de dados todos os dias (o equivalente a cerca de 210.000 DVDs)
- Já possui mais de 100 petabytes armazenados desde o início do projeto (= 700 anos de filmes armazenados com qualidade full HD)



"View inside detector at the CMS cavern LHC CERN" by Tighef - Own work.
Licensed under CC BY-SA 3.0 via Wikimedia Commons -

http://commons.wikimedia.org/wiki/File:View_inside_detector_at_the_CMS_cavern_LHC_CERN.jpg

Desafios para a Computação

- Como **curar** volumes tão grandes de dados?
 - Armazenar, organizar, preservar, compartilhar, ...
- Como **processar** volumes tão grandes de dados?
 - Filtrar, corrigir, computar, analisar, ...

- Necessidade:

**Plataformas de
Computação de Alto
Desempenho**



Fonte: <http://home.web.cern.ch/about/computing>

- **Supercomputadores, aglomerados e grades computacionais**
 - Custo de aquisição e manutenção elevados
 - Poucos laboratórios têm acesso
- **Nuvens computacionais**
 - Recursos configuráveis, alugados de acordo com a demanda
 - Recursos acessados via Internet
 - Paga-se somente por aquilo que é consumido
 - **Democratização do poder computacional**
- Ex.: Nuvem USP
<https://nuvem.uspdigital.usp.br/>



Nuvens como Catalisadoras de Pesquisa Transformativa

DA CIÊNCIA À E-CIÊNCIA: PARADIGMAS DA DESCOBERTA DO CONHECIMENTO

DANIEL CORDEIRO

KELLY R. BRAGHETTO

ALFREDO GOLDMAN

FABIO KON

Artigo na edição 97 da Revista USP
(março/abril/maio de 2013)

<http://www.revistas.usp.br/revusp/article/view/61867>

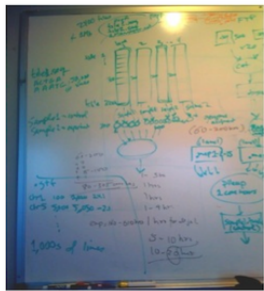


Workflows Científicos

- Automatizam experimentos ou processos científicos

The image displays three screenshots of workflow management systems. On the left is the Kepler interface, featuring a banner with the text 'Kepler Your Science. Enabled.' and a navigation menu. In the center is the Taverna website, showing the 'Taverna Workflow Management System' header, a navigation menu, and a 'RECENT NEWS' section. On the right is the Pegasus website, with a banner that reads 'Pegasus Workflow Management System' and a navigation menu.

If You can draw....



We can make it run....



Epigenomics,
Ben Berman,
USC

Helio-Seismology : Laurent Gizon, Max Planck

PSOCI
(Solar
Fuels)
Jeff
Tilson,
RENCI



Seminar by Frédéric Suter:
SimGrid, Versatile Simulation
of Distributed Systems - Jan
2015

Seminar by Christos Nikolaou:
A Game Theoretic Approach
for Managing Multi-Modal
Urban Mobility Systems - Jan
2015

Pegasus 4.4.1 Released - Dec
2014

Seminar by Dan Katz: Building
and Linking Local, Regional,
and National
Cyberinfrastructure to
Advance Science - Dec 2014

Seminar by Rizos Sakellariou:

Exemplos:

<https://kepler-project.org/>
<http://pegasus.isi.edu/>
<http://www.taverna.org.uk/>

Algumas Ferramentas dos Cientistas da Atualidade

Science Gateways

- Conjuntos de dados, ferramentas e aplicações, desenvolvidos por comunidades e geralmente disponibilizados em portais Web

The image shows two web portals. On the left is the CIPRES Science Gateway, featuring a tree logo and navigation links for 'CIPRES', 'Home', and 'Tools'. Below it is a text box asking for help with missing results. On the right is the NIF (Neuroscience Information Framework) website, which includes a search bar, a 'Search NIF' button, and a 'Search NeuroLex' button. A word cloud at the bottom of the NIF page features terms like 'neuron', 'coeruleus', 'amygdala', and 'hippocampus'. To the right of the search bar is a 'NIF STATISTICS' box with the following data: NIF Version: 6.2, Ontology Version: 2.9, Level 2.5/3.0 Resources: 239, Registry Entries: 12,598, and Total Records: 829,679,866. Below the statistics is a 'NIF NAVIGATOR' box with a 'Powered By NIF' logo.

Exemplos:

<https://www.phylo.org/>

<http://www.neuinfo.org/>

Benefícios do uso dessas ferramentas:

- Facilidade no **compartilhamento** de recursos
- **Automatização** de procedimentos rotineiros
- Documentação dos experimentos científicos, ajudando a garantir a sua **reprodutibilidade**
- Promoção da **colaboração** científica
- Promoção da **Ciência Aberta**

Se baseia em três pilares:

- Acesso aberto
 - Ex.: **SIBi – Sistema Integrado de Bibliotecas, da USP**
<http://www.sibi.usp.br/>
- Software livre
 - Ex.: **SAM (Sequence Alignment/Map) Tools** – ferramenta da área de bioinformática, usada no alinhamento de sequências
<http://www.htslib.org/>
- Dados abertos
 - Ex.: **Corpus Histórico do Português Tycho Brahe**
<http://www.tycho.iel.unicamp.br/>

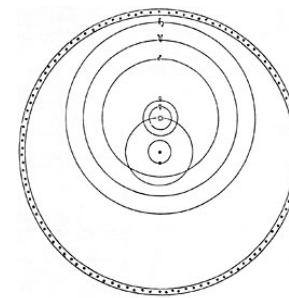


Fig. 16 - Le système de Tycho-Brahe.

Projeto Tycho Brahe

Dados Abertos


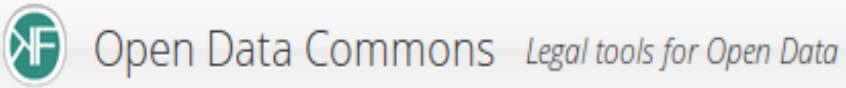
- Precisam ser mais do que “dados de acesso público”
- Segundo a definição da *Open Knowledge Foundation*:
“Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).”
- Para que possam ser reutilizados, é preciso que os dados tenham **boa qualidade** e sejam **compreensíveis**
- Preferencialmente, devem estar acompanhados de **metadados**
 - Informações sobre sua **estrutura**
 - Informações sobre sua **proveniência** (que definem como, quando, onde, por quem e por quê os dados foram gerados)

- Periódicos já condicionam a publicação de um artigo à disponibilização dos dados usados no estudo
 - Ex.: BioMed Central, PLOS One
 - <http://www.biomedcentral.com/about/opendata>
 - <http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/>
- FAPESP já exige que dados e softwares produzidos nos projetos que financia sejam colocados à disposição do público
- **Benefícios**
 - Possibilidade de validação e reprodução dos resultados
 - Produção de ciência de melhor qualidade e maior impacto

Dados Científicos Abertos – Principais Desafios

- Criar **padrões** para a representação dos dados
 - Ausência de consenso entre os cientistas sobre o que é necessário armazenar
- Registrar a **proveniência** completa dos dados
- Vencer a **resistência** da comunidade científica
 - Coletar dados é um trabalho muito dispendioso e pouco reconhecido
- Assegurar que cientistas e instituições recebam **créditos** pelos dados que produzem e que possam **proteger** seu uso futuro
- Financiar a **curadoria** dos dados (que depende de equipamentos e pessoas)

Dados Abertos – Avanços nos Aspectos Legais

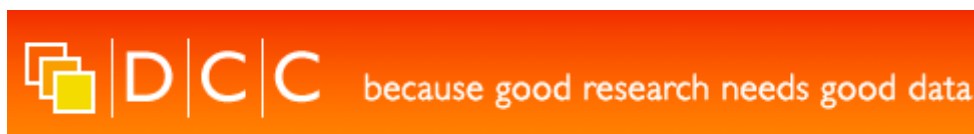
- *Lei de Acesso à Informação* (decreto Nº 7.724 de 16/05/12)
 - Garante acesso às informações produzidas ou custodiadas por órgãos e entidades da União, incluindo as universidades federais e instituições de fomento à pesquisa (CNPq, CAPES)
- Licenças específicas para bancos de dados
 - *Creative Commons* (CC) <http://creativecommons.org/>

 - *Open Data Commons* (ODC) <http://opendatacommons.org/>

- Licenças permitem condicionar o uso dos dados à atribuição de crédito ao seus autores e estabelecer que redistribuições só possam ser feitas sob a mesma licença (ou similar)

- **Grupo de trabalho em Ciência Aberta**

<http://www.cienciaaberta.net/>

- ***Digital Curation Centre***

<http://www.dcc.ac.uk/>



- ***Open Knowledge Foundation***

<https://okfn.org/>

<http://br.okfn.org/>



O Tema “Dados Científicos Abertos” na Mídia

- Bruno de Pierro, “**Uma ciência mais aberta**”, Revista FAPESP – março, 2013

<http://revistapesquisa.fapesp.br/2013/03/15/uma-ciencia-mais-aberta/>

- Claudia Domingues e Fabio Kon, “**Em defesa do compartilhamento público de dados científicos**”, Le Monde Diplomatique Brasil – maio, 2014

<http://www.diplomatique.org.br/artigo.php?id=1653>

- Kelly R. Braghetto, “**A ciência precisa ser aberta**”, Revista ARede – Tecnologia para Inclusão Social – julho, 2014

<http://www.revista.aredes.inf.br/site/edicao-n-99-julho-agosto-2014/7035-opiniao-o-que-kelly-rosa-braghetto-esta-pensando>

Versão estendida disponível em:

<http://neuromat.numec.prp.usp.br/content/open-data-science-neuromat-op-%C2%ADed>

- **Centro de Pesquisa, Inovação e Difusão em Neuromática (NeuroMat)**
 - Integra modelagem matemática com pesquisa básica e aplicada na fronteira da Neurociência.
- ***Neuroscience Experiments System* (NES)** – um software livre para o gerenciamento de dados neurofisiológicos clínicos e experimentais
<https://github.com/neuromat/nas>
- Próximos passos do projeto no que se refere à transferência de tecnologia:
 - Novos módulos no NES
 - **Banco de dados aberto**
 - *Science gateway* (o Portal do NeuroMat)



Um “causo” envolvendo *Big Data*: Google Flu Trends

google.org Flu Trends

Language: English (United States)

[Google.org home](#)

[Dengue Trends](#)

Flu Trends

[Home](#)

United States

National

[Download data](#)

[How does this work?](#)

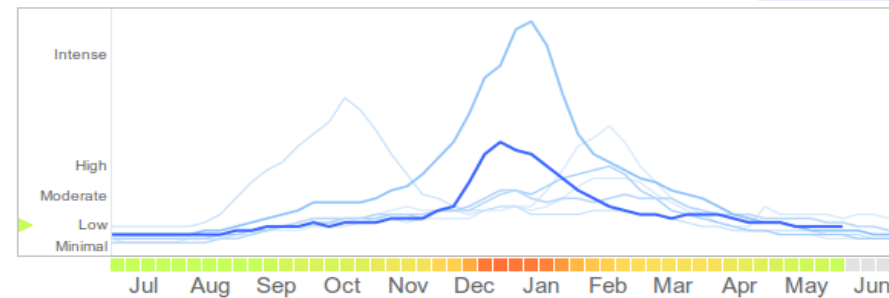
[FAQ](#)

Explore flu trends - United States

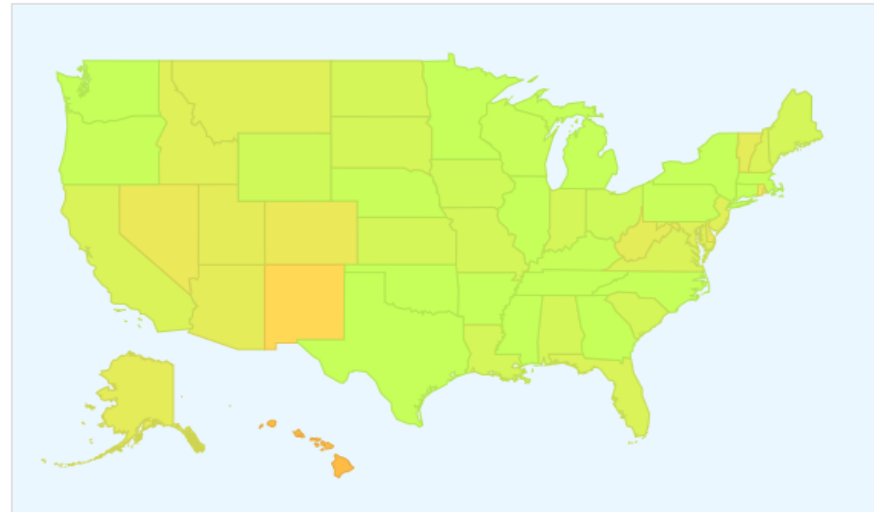
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)

National

● 2013-2014 ● Past years ▼



States | [Cities](#) (Experimental)




Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through June 3, 2014.

Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu:

1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.

 [Centers for Disease Control and Prevention](#)

Animated Flu Trends in Google Earth

[Download and explore](#) Flu Trends data in Google Earth. Need Google Earth? [Download it here.](#)

Embed this chart

Use [this embed code](#) to show this chart on your website.

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year¹. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities². Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza^{3,4}. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t , α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza

<http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html>

A Acurácia das Predições Impressionaram...

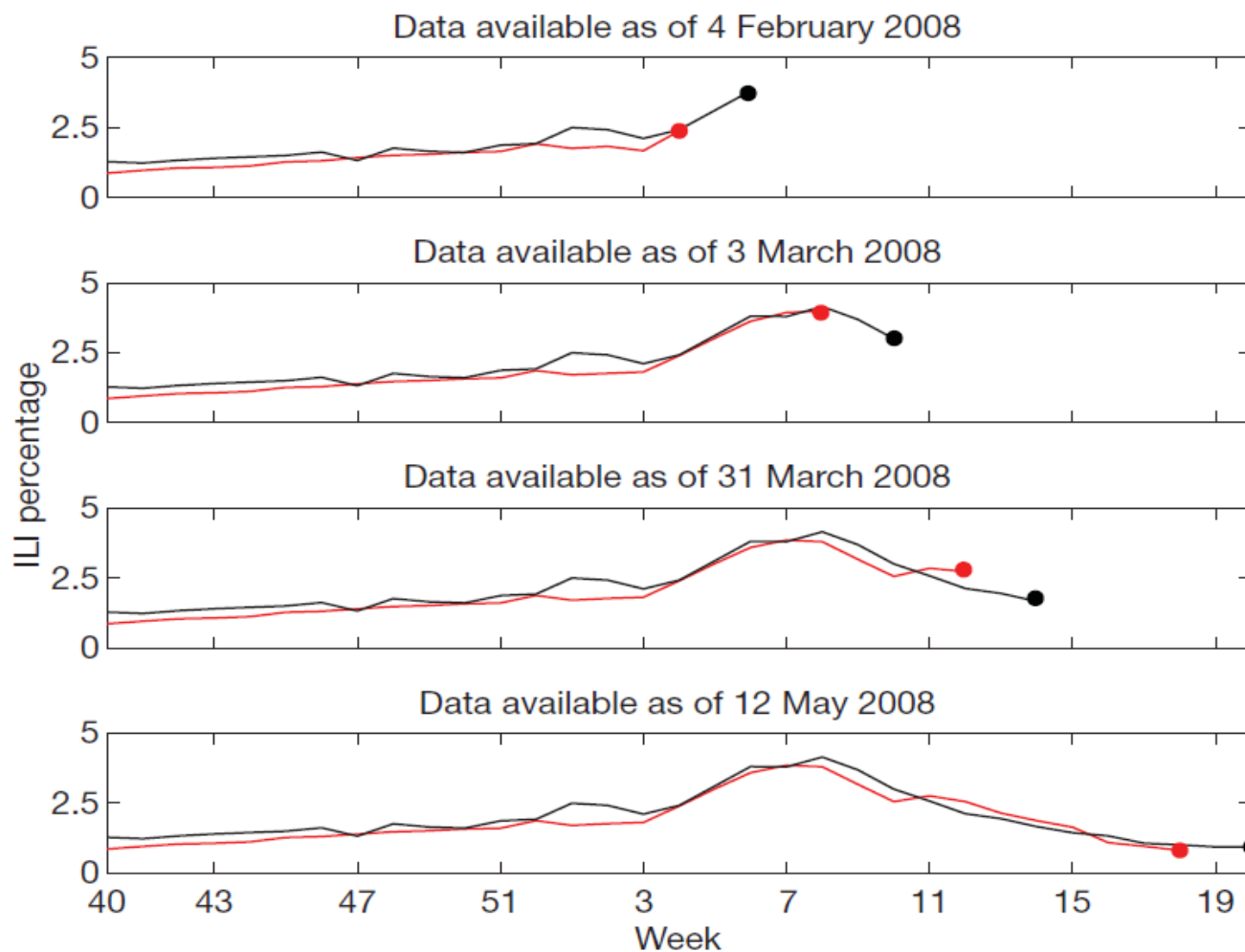


Figure 3 | ILI percentages estimated by our model (black) and provided by the CDC (red) in the mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season. During week 5 we detected a sharply increasing ILI percentage in the mid-Atlantic region; similarly, on 3 March our model indicated that the peak ILI percentage had been reached during week 8, with sharp declines in weeks 9 and 10. Both results were later confirmed by CDC ILI data.

Mas no “dilúvio de dados”, nem tudo é um mar de rosas...

- Apesar das previsões “acertadas” em boa parte do tempo, o modelo do Google errou na epidemia de gripe de 2012-2013 dos EUA (estimou o dobro da quantidade real de casos):



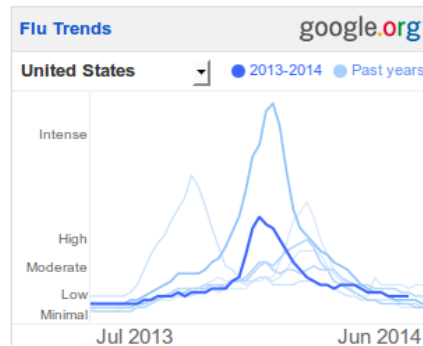
BIG DATA

Google Flu Trends: The Limits of Big Data

By STEVE LOHR MARCH 28, 2014 7:00 AM 14 Comments

- E-MAIL
- f
- CEBOOKER
- SAVE
- MORE

Google Flu Trends, once a poster child for the power of big-data analysis, seems to be under attack.



This month in a



Disruptions: Data Without Context Tells a Misleading Story

By NICK BILTON FEBRUARY 24, 2013 11:00 AM 4 Comments

<http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/>

<http://bits.blogs.nytimes.com/2013/02/24/disruptions-google-flu-trends-shows-problems-of-big-data-without-context/>

<http://www.nature.com/news/when-google-got-flu-wrong-1.12413>

<http://gking.harvard.edu/publications/parable-google-flu%C2%A0traps-big-data-analysis>

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{3,5,6}

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

BY DECLAN BUTLER

complement, but not substitute for, traditional epidemiological surveillance networks.

Ciência de Dados: um Meio ou um Fim?

- O **contexto** em que os dados foram coletados ou criados pode impactar o resultado das análises
 - É sempre possível capturar e considerar informações de contexto?
 - “Contexto” não é o mesmo que “proveniência”
- A análise automatizada de dados deve ser usada como uma **ferramenta de apoio** no processo de descoberta científica
 - Ela não substitui o papel de **crítico** exercido pelo cientista nesse processo

Obrigada por sua atenção.

Kelly Rosa Braghetto

kellyrb@ime.usp.br

Estes slides estão disponíveis em minha página:

<http://www.ime.usp.br/~kellyrb/>



INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO