

Semantic Computing and Smart Cyberinfrastructure

Tony Hey

Corporate Vice President

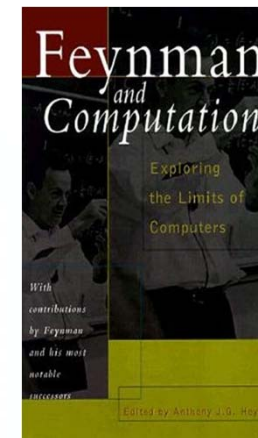
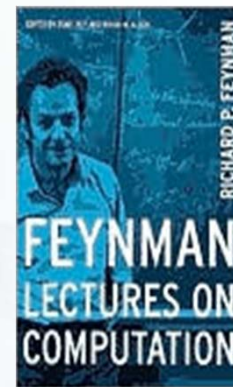
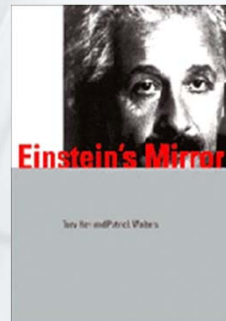
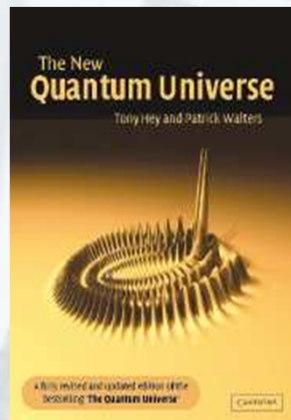
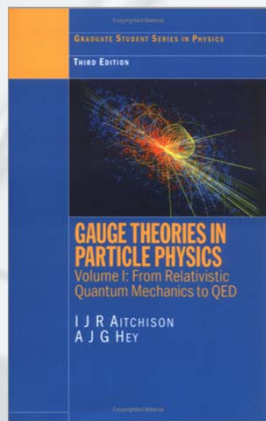
Microsoft Research



Tony Hey – My Background




UNIVERSITY OF
Southampton



UK National Centre for Text Mining

the national centre for text mining



The National Centre for Text Mining

You are in: [Home](#) | [Welcome to NaCTeM](#)

- [Home](#)
- [Aims & Objectives](#)
- [NaCTeM Services](#)
- [Text Mining Tools](#)
- [Resources](#)
- [Terms & Conditions](#)
- [FAQ](#)
- [News & Events](#)
- [People](#)
- [Projects](#)
- [Publications](#)
- [Community](#)
- [External Collaboration](#)
- [Vacancies](#)
- [Teaching & Tutorials](#)
- [Feedback](#)
- [How to Find Us](#)
- [Site Map](#)
- [Search](#)

Welcome to NaCTeM

The National Centre for Text Mining (NaCTeM) is the first publicly-funded text mining centre in the world. We provide text mining services in response to the requirements of the UK academic community. NaCTeM is operated by the University of Manchester with close collaboration with the University of Tokyo.

On our website, you can find pointers to sources of information about text mining such as links to

- text mining services provided by NaCTeM
- software tools, both those developed by the NaCTeM team and by other text mining groups
- seminars, general events, conferences and workshops
- tutorials and demonstrations
- text mining publications

Let us know if you would like to include any of the above in our website.

What text mining can do for you

Text mining offers a solution to the challenge of 'data deluge', information overload and information overlook. For more information, please see:

- [NaCTeM Brochure](#),
- [Text Mining Briefing Paper](#),
- [National Centre for Text Mining: an introduction to tools for researchers](#),
- [Vision for the Future](#),
- [Mining Biomedical Literature](#).
- **NEW!** [Event extraction for systems biology by text mining the literature](#)
- **NEW!** [Supporting the education evidence portal via text mining](#)

NaCTeM has developed text mining services and service exemplars for the UK academic community. Our services are underpinned by a number of generic natural language processing tools:

- [TerMine](#) is a Term Management System which identifies key phrases in text.

Featured News

- [Biomedical Text Mining Training, 27th-29th October 2010](#)
- [BioNLP Shared Task 2011](#)
- [Release of Taverna Plugin for U-Compare](#)
- [Text mining enhances Educational Evidence Portal - new article and demo site](#)
- [Medal of honour awarded to Professor Tsujii](#)
- [Improved acronym disambiguation - release of updated software service and paper](#)

Featured News Feed

Other News & Events

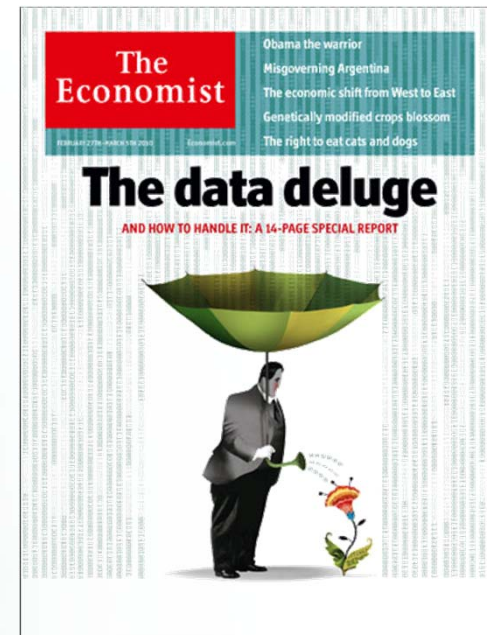
- [Invited lecture at the Institute of Information Science of Academia Sinica](#)



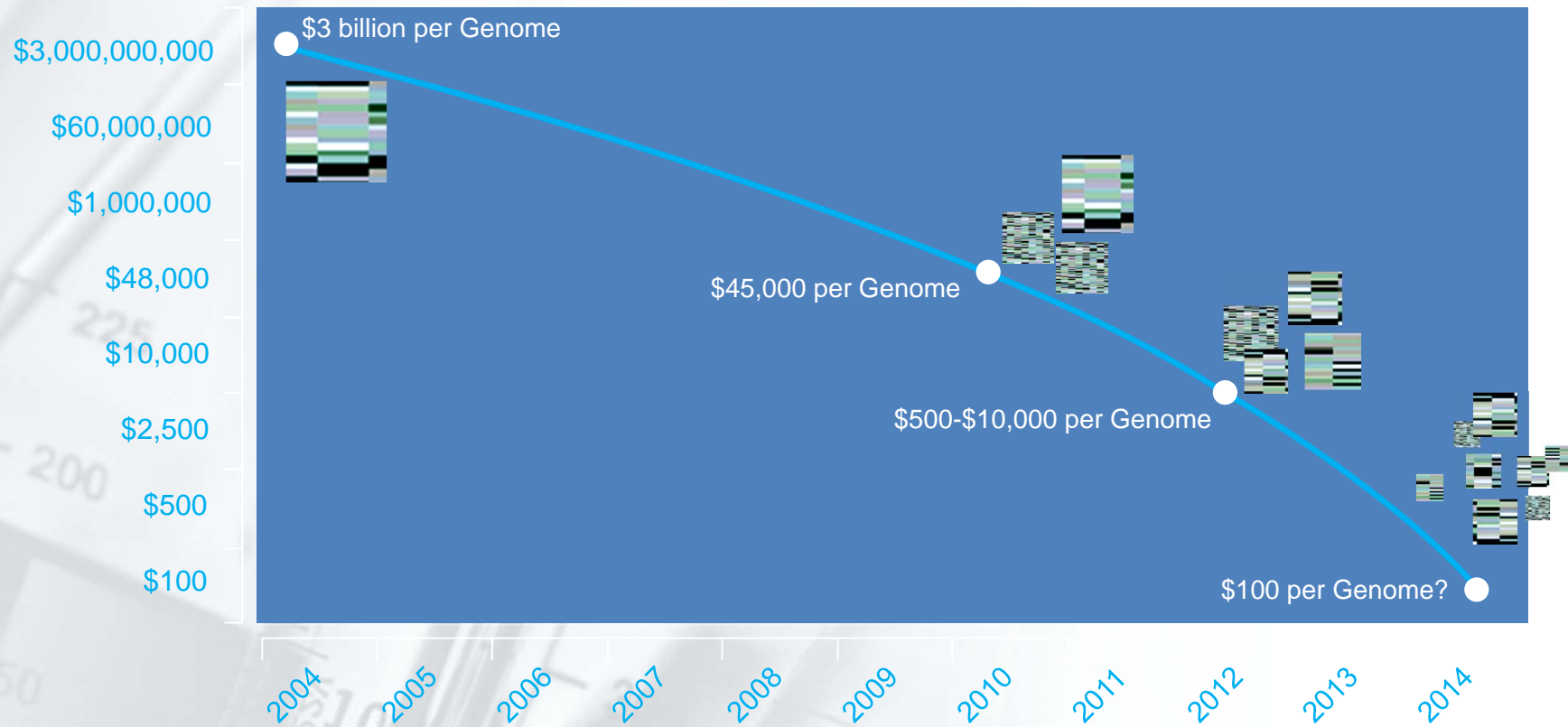
The Data Deluge



A Tidal Wave of Scientific Data



Gene Sequencing Explosion



Source: George Church, Harvard Medical School, as reported in IEEE Spectrum, Feb '10, Figure 1.1



Genomics and Personalized Medicine

Adapting treatments to a person's specific genetic make-up:

- Targeting patients who **can benefit** (e.g. 10% of people cannot respond to codeine), and **not develop toxicities** (e.g. Abacavir for HIV).
- Appropriate **dosage** of a drug by using genetic variants to understand drug metabolism (e.g. anti-depressants, beta blockers, opioid analgesics)
- More **drug approvals (re-approvals)** because can now target the right sub-group based on genetics.



Astronomy and Particle Physics

In 2000 the Sloan Digital Sky Survey collected more data in its 1st week than was collected in the entire history of Astronomy

By 2016 the New Large Synoptic Survey Telescope in Chile will acquire 140 terabytes in 5 days - more than Sloan acquired in 10 years

The Large Hadron Collider at CERN generates 40 terabytes of data every second

Sources: *The Economist*, Feb '10; IDC

Citizen Science: GalaxyZoo

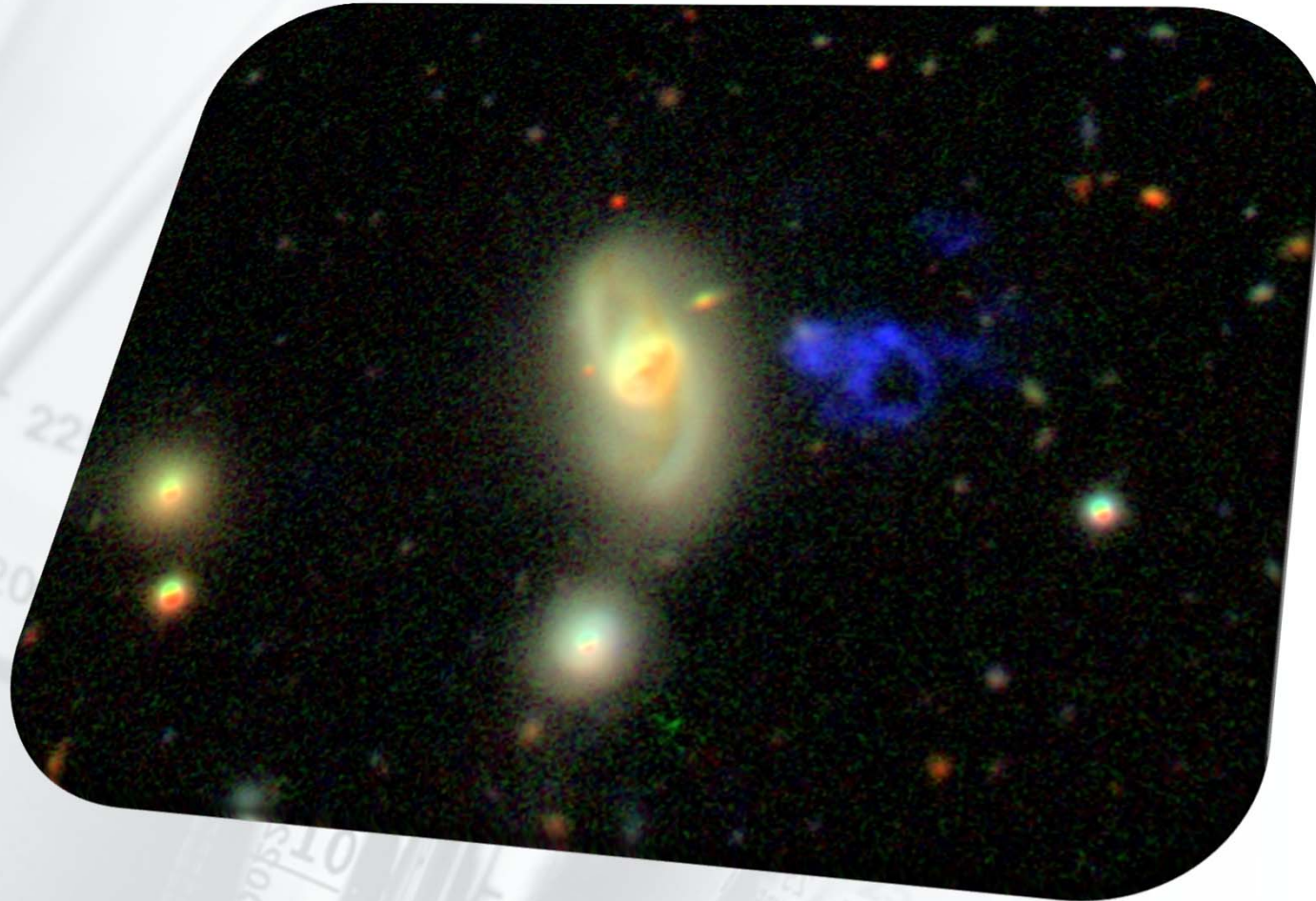
- Goal of 1 million visual galaxy classifications by the public
- Enormous publicity (CNN, Times, Washington Post, BBC)
- 200,000 people participating, blogs, poems ...



- **Allows general public to search for photographs and classify different types of galaxies**



Hanny van Arkle's Voorwerp



The Fourth Paradigm: Data-Intensive Science

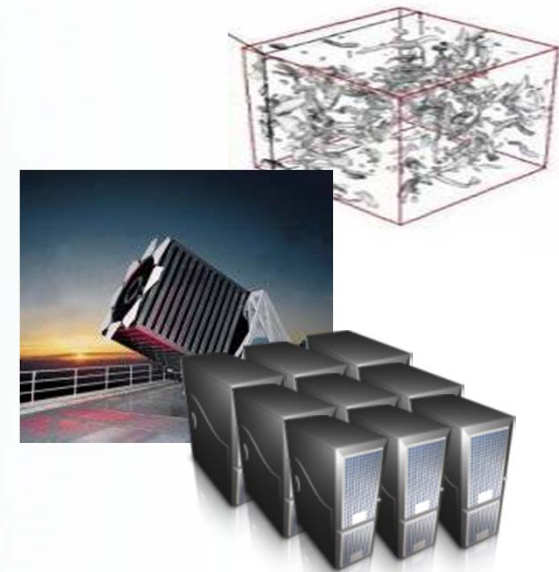


Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
 - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
 - Newton's Laws, Maxwell's Equations...
3. Last few decades – **Computational Science**
 - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
 - Scientists overwhelmed with data sets from many different sources
 - Data captured by instruments
 - Data generated by simulations
 - Data generated by sensor networks
 - **eScience is the set of tools and technologies to support data federation and collaboration**
 - For analysis and data mining
 - For data visualization and exploration
 - For scholarly communication and dissemination



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

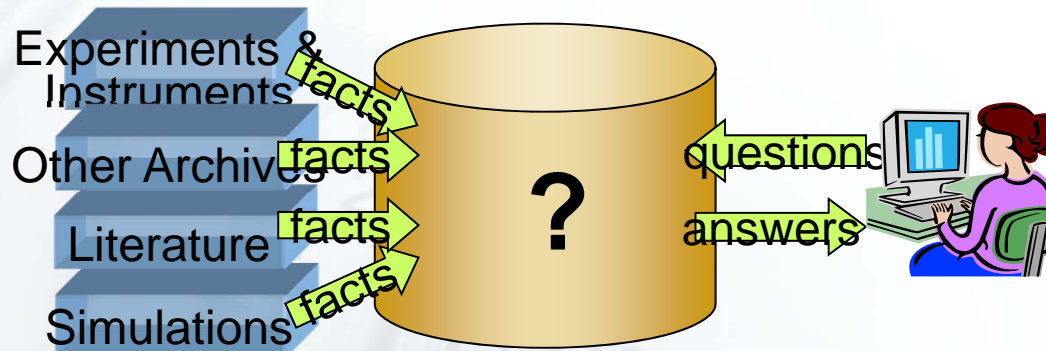


(With thanks to Jim Gray)



X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

With thanks to Jim Gray



World Wide Telescope

www.worldwidetelescope.org



Seamless Rich Social Media Virtual Sky
Web application for science and education

Participants

- Alyssa Goodman; Harvard University
- Alex Szalay; Johns Hopkins University
- Curtis Wong, Jonathan Fay; Microsoft Research
- Integration of data sets and one-click contextual access
- Easy access and use
- As of 1/23/2009: 1,606,950 unique users (someone that has downloaded, installed, and successfully used WWT)
- There have been 4,089,898 sessions for an average of 2.55 sessions per user
- The average number of new users that have installed and used WWT has been 3,773 per day



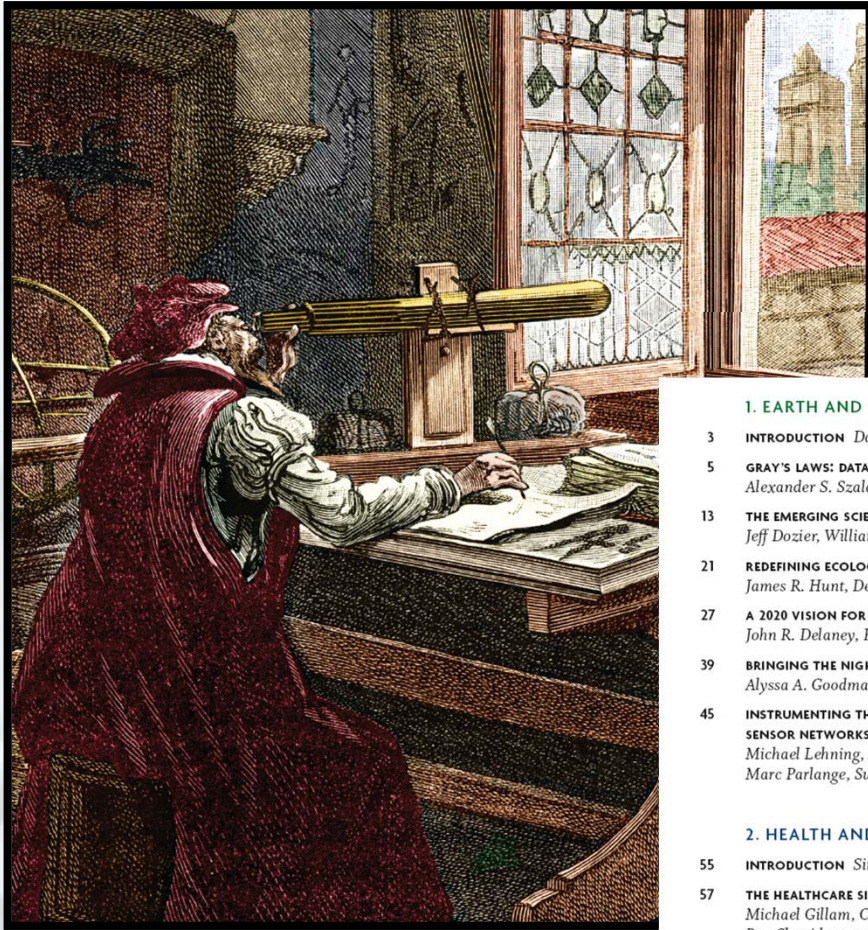


The
F O U R T H
P A R A D I G M

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE





An edited collection of 26 short technical essays, divided into 4 sections

1. EARTH AND ENVIRONMENT

- 3 INTRODUCTION *Dan Fay*
- 5 GRAY'S LAWS: DATABASE-CENTRIC COMPUTING IN SCIENCE
Alexander S. Szalay, José A. Blakeley
- 13 THE EMERGING SCIENCE OF ENVIRONMENTAL APPLICATIONS
Jeff Dozier, William B. Gail
- 21 REDEFINING ECOLOGICAL SCIENCE USING DATA
James R. Hunt, Dennis D. Baldocchi, Catharine van Ingen
- 27 A 2020 VISION FOR OCEAN SCIENCE
John R. Delaney, Roger S. Barga
- 39 BRINGING THE NIGHT SKY CLOSER: DISCOVERIES IN THE DATA DELUGE
Alyssa A. Goodman, Curtis G. Wong
- 45 INSTRUMENTING THE EARTH: NEXT-GENERATION SENSOR NETWORKS AND ENVIRONMENTAL SCIENCE
Michael Lehning, Nicholas Dawes, Mathias Bavay, Marc Parlange, Suman Nath, Feng Zhao

2. HEALTH AND WELLBEING

- 55 INTRODUCTION *Simon Mercer*
- 57 THE HEALTHCARE SINGULARITY AND THE AGE OF SEMANTIC MEDICINE
Michael Gillam, Craig Feied, Jonathan Handler, Eliza Moody, Ben Shneiderman, Catherine Plaisant, Mark Smith, John Dickason
- 65 HEALTHCARE DELIVERY IN DEVELOPING COUNTRIES: CHALLENGES AND POTENTIAL SOLUTIONS
Joel Robertson, Del DeHart, Kristin Tolle, David Heckerman
- 75 DISCOVERING THE WIRING DIAGRAM OF THE BRAIN
Jeff W. Lichtman, R. Clay Reid, Hanspeter Pfister, Michael F. Cohen
- 83 TOWARD A COMPUTATIONAL MICROSCOPE FOR NEUROBIOLOGY
Eric Horvitz, William Kristan
- 91 A UNIFIED MODELING APPROACH TO DATA-INTENSIVE HEALTHCARE
Iain Buchan, John Winn, Chris Bishop
- 99 VISUALIZATION IN PROCESS ALGEBRA MODELS OF BIOLOGICAL SYSTEMS
Luca Cardelli, Corrado Priami

3. SCIENTIFIC INFRASTRUCTURE

- 109 INTRODUCTION *Daron Green*
- 111 A NEW PATH FOR SCIENCE? *Mark R. Abbott*
- 117 BEYOND THE TSUNAMI: DEVELOPING THE INFRASTRUCTURE TO DEAL WITH LIFE SCIENCES DATA
Christopher Southan, Graham Cameron
- 125 MULTICORE COMPUTING AND SCIENTIFIC DISCOVERY
James Larus, Dennis Gannon
- 131 PARALLELISM AND THE CLOUD *Dennis Gannon, Dan Reed*
- 137 THE IMPACT OF WORKFLOW TOOLS ON DATA-CENTRIC RESEARCH
Carole Goble, David De Roure
- 147 SEMANTIC eSCIENCE: ENCODING MEANING IN NEXT-GENERATION DIGITALLY ENHANCED SCIENCE
Peter Fox, James Hendler
- 153 VISUALIZATION FOR DATA-INTENSIVE SCIENCE
Charles Hansen, Chris R. Johnson, Valerio Pascucci, Claudio T. Silva
- 165 A PLATFORM FOR ALL THAT WE KNOW: CREATING A KNOWLEDGE-DRIVEN RESEARCH INFRASTRUCTURE
Savas Parastatidis

4. SCHOLARLY COMMUNICATION

- 175 INTRODUCTION *Lee Dirks*
- 177 JIM GRAY'S FOURTH PARADIGM AND THE CONSTRUCTION OF THE SCIENTIFIC RECORD
Clifford Lynch
- 185 TEXT IN A DATA-CENTRIC WORLD *Paul Ginsparg*
- 193 ALL ABOARD: TOWARD A MACHINE-FRIENDLY SCHOLARLY COMMUNICATION SYSTEM
Herbert Van de Sompel, Carl Lagoze
- 201 THE FUTURE OF DATA POLICY
Anne Fitzgerald, Brian Fitzgerald, Kylie Pappalardo
- 209 I HAVE SEEN THE PARADIGM SHIFT, AND IT IS US *John Wilbanks*
- 215 FROM WEB 2.0 TO THE GLOBAL DATABASE *Timo Hannay*



Free PDF Download

Amazon Kindle version; Paperback print on demand

<http://research.microsoft.com/fourthparadigm/>

- “The impact of Jim Gray’s thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science.”
 - **Bill Gates**, Chairman, Microsoft Corporation
- “One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working.”
 - **Douglas Kell**, University of Manchester
- “The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives.”
 - **Gordon Bell**, Microsoft Research

The screenshot shows the Microsoft Research website. At the top, there is a search bar and navigation links for Projects, Publications, People, and Downloads. Below that, there are tabs for Home, Our Research, Collaboration, and Careers. The main content area features the title "The Fourth Paradigm: Data-Intensive Scientific Discovery" and a sub-headline "Presenting the first broad look at the rapidly emerging field of data-intensive science". A central image shows a book cover for "The Fourth Paradigm" with the subtitle "DATA-INTENSIVE SCIENTIFIC DISCOVERY". To the right of the image, there is a section titled "The Fourth Paradigm Now Available in Paperback and On Demand" with text explaining that the book is available as a free PDF download, or as a printed paperback or Kindle version. Below this, there are links to "Order the paperback from Amazon.com" and "Order the Kindle version from Amazon.com". On the far right, there is a "In the News" section with a link to "A Deluge of Data Shapes a New Era in Computing", a "Download The Fourth Paradigm" section with links for "Full text, low resolution (6 MB)", "Full text, high resolution (93 MB)", and "By chapter and essay", and a "Related Resources" section with links to "Microsoft Research collaborative projects" and "eScience Workshop 2009".



It's a data-driven world

Data and Data Services
as Innovation Enablers



It's a data-driven world

- Machine Translation (MT)
 - From rules to statistics
- Spell Checking as MT
 - Search queries + click through

Banko and Brill (2001)

Effectiveness of statistical NLP techniques is highly susceptible to the **data size** used to develop them

Norvig (2008)

It is the **size of data**, not the sophistication of the algorithms that ultimately play the central role in modern NLP

Machine Translation



[Home](#) | [Tools](#) | [Help](#)

Free online translation service for a truly *worldwide* web

Languages

English (Auto-Detected)



French

[Translate](#)

[Clear All](#)

[Add to Favorites](#)

Enter text or webpage URL

The main goal of the conference is to foster the dialog between experts in each sub-discipline. Therefore we especially encourage submissions of work that is interesting to multiple areas, such as multimodal approaches.



[Report offensive translations](#)

L'objectif principal de la conférence est de favoriser le dialogue entre les experts dans chaque sous discipline. Nous encourageons donc particulièrement soumissions de travail qui est intéressant à plusieurs domaines, tels que les approches multimodales.

Rate this translation: ☆☆☆☆☆

[Copy Translation](#)

[New: Translator V2 API & Widget announced at MIX 2010](#)

Powered by **Microsoft®** Translator



A timeline

1950

1970

1990

2010

1960

1980

2000

1954: 60 Russian sentences successfully translated into English



A timeline

1950



1960

1954: 60 Russian sentences successfully translated into English



A timeline

1950

1970

1990

2010

1960

1980

2000

1954: 60 Russian sentences successfully translated into English

1968: Systran begins work in Russian-English



A timeline

1950

1970

1990

2010

1960

1980

2000

1954: 60 Russian sentences successfully translated into English

1968: Systran begins work in Russian-English

1997: Thirty years later, Systran powers Babelfish, the first web translator



Broad-Domain, High-Quality MT?

- Since the 1950's: “coming soon”
 - Hand-coded systems require decades of intensive work
 - Adequate for narrow domains
 - But in the general domain, can't improve beyond “gisting” quality
- But things started to change 15 years ago
 - Shift to data-driven, machine-learned approach
 - Rapid progress in quality
 - Exploding consumer interest over last few years
 - Gains driven by more/better data, better algorithms



The Statistical Revolution

Instead of hand-coding rules

- Exploit large volumes of existing parallel text
- Learn how words, phrases, and structures translate in context

The Rosetta Stone > The British Museum

THE BRITISH MUSEUM

Home Visiting What's on Explore Research Learning The Museum Join in Shop online

Introduction Themes Highlights World cultures Online tours Galleries Young explorers

Home > Explore > Highlights

The Rosetta Stone

From Fort St Julien, el-Rashid (Rosetta), Egypt, Ptolemaic Period, 196 BC

A valuable key to the decipherment of hieroglyphs, the inscription on the Rosetta Stone is a decree passed by a council of priests. It is one of a series that affirm the royal cult of the 13-year-old Ptolemy V on the first anniversary of his coronation.

In previous years the family of the Ptolemies had lost control of certain parts of the country. It had taken their armies some time to put down opposition in the Delta, and parts of southern Upper Egypt, particularly Thebes, were not yet back under the government's control.

Before the Ptolemaic era (that is before about 332 BC), decrees in hieroglyphs such as this were usually set up by the king. It shows how much things had changed from Pharaonic times that the priests, the only people who had kept the knowledge of writing hieroglyphs, were now issuing such decrees. The list of good deeds done by the king for the temples hints at the way in which the support of the priests was ensured.

The decree is inscribed on the stone three times, in hieroglyphic (suitable for a priestly decree), demotic (the native script used for daily purposes), and Greek (the language of the administration). The importance of this to Egyptology is immense.

Soon after the end of the fourth century AD, when hieroglyphs had gone out of use, the knowledge of how to read and write them disappeared. In the early years of the nineteenth century, some 1400 years later, scholars were able to use the Greek inscription on this stone as the key to decipher them.

Thomas Young, an English physicist, was the first to show that some of the hieroglyphs on the Rosetta Stone wrote the sounds of a royal name, that of Ptolemy. The French scholar Jean-François Champollion then realized that hieroglyphs recorded the sound of the Egyptian language and laid the foundations of our knowledge of ancient Egyptian language and culture.

Soldiers in Napoleon's army discovered the Rosetta Stone in 1799 while digging the foundations of an addition to a fort near the town of el-Rashid (Rosetta). On Napoleon's defeat, the stone became the property of the British under the terms of the Treaty of

On display

8 4 Egyptian sculpture Room 4 View floorplan

British Museum - Piedra Rosetta

THE BRITISH MUSEUM

Home Visiting What's on Explore Research Learning The Museum Join in Shop online

Introduction Themes Highlights World cultures Online tours Galleries Young explorers

Home > Explore > Highlights

Explore / Highlights

English | Français | Italiano

Piedra Rosetta

Origen: Fuerte de San Julián, el-Rashid (Rosetta), Egipto
Período ptolemaico, 196 a.C.

Pieza clave para descifrar jeroglíficos

El texto contenido en la Piedra Rosetta corresponde a un decreto dictado por un consejo de sacerdotes e integra una serie de decretos que ratifican el culto real de Ptolomeo V, de 13 años de edad, en el primer aniversario de su coronación.

En años anteriores, la dinastía ptolemaica había perdido el control de ciertas zonas del país. Después de un largo tiempo, su ejército logró derrocar a la oposición en el Delta, pero la región sur del Alto Egipto, Tebas en especial, no había sido aun recuperada por el gobierno.

Antes de la era ptolemaica (hasta cerca del año 332 a.C.), el rey solía emitir decretos en jeroglíficos como el de esta pieza. Este dato da cuenta de cómo cambiaron las cosas desde los tiempos faraónicos, ya que los sacerdotes, las únicas personas que conocían la escritura jeroglífica, pasaron a emitir dichos decretos. La cantidad de actos reales condescendientes con los templos nos ilustra la forma en la cual se garantizaba el apoyo de los sacerdotes.

El decreto está escrito en la piedra por partida triple, en jeroglífico (acorde a un decreto sacerdotal), en demótico (la escritura nativa de uso diario) y en griego (el idioma del gobierno). Su importancia para la etimología es enorme. Al poco tiempo del final del s. IV a.C., cuando se dejaron de utilizar jeroglíficos, el conocimiento sobre cómo leerlos y escribirlos se perdió. A comienzos del s. XIX, unos 1400 años después, los científicos lograron descifrarlos utilizando

Highlights

Search:

Browse or search over 4,000 highlights from the Museum collection

Related objects

Busto colosal de Ramsés II

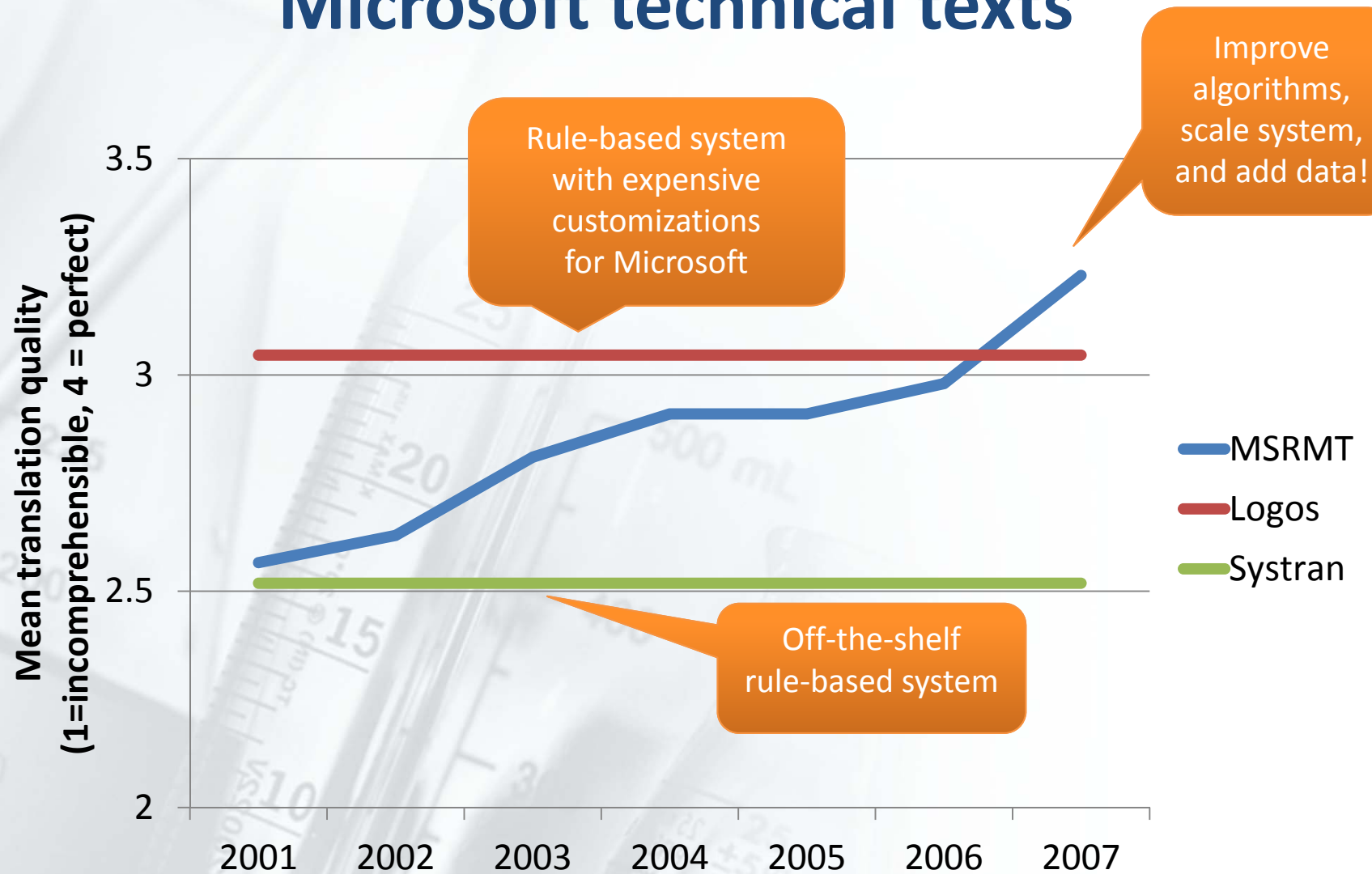
1 of 10 Objects See

Shop online

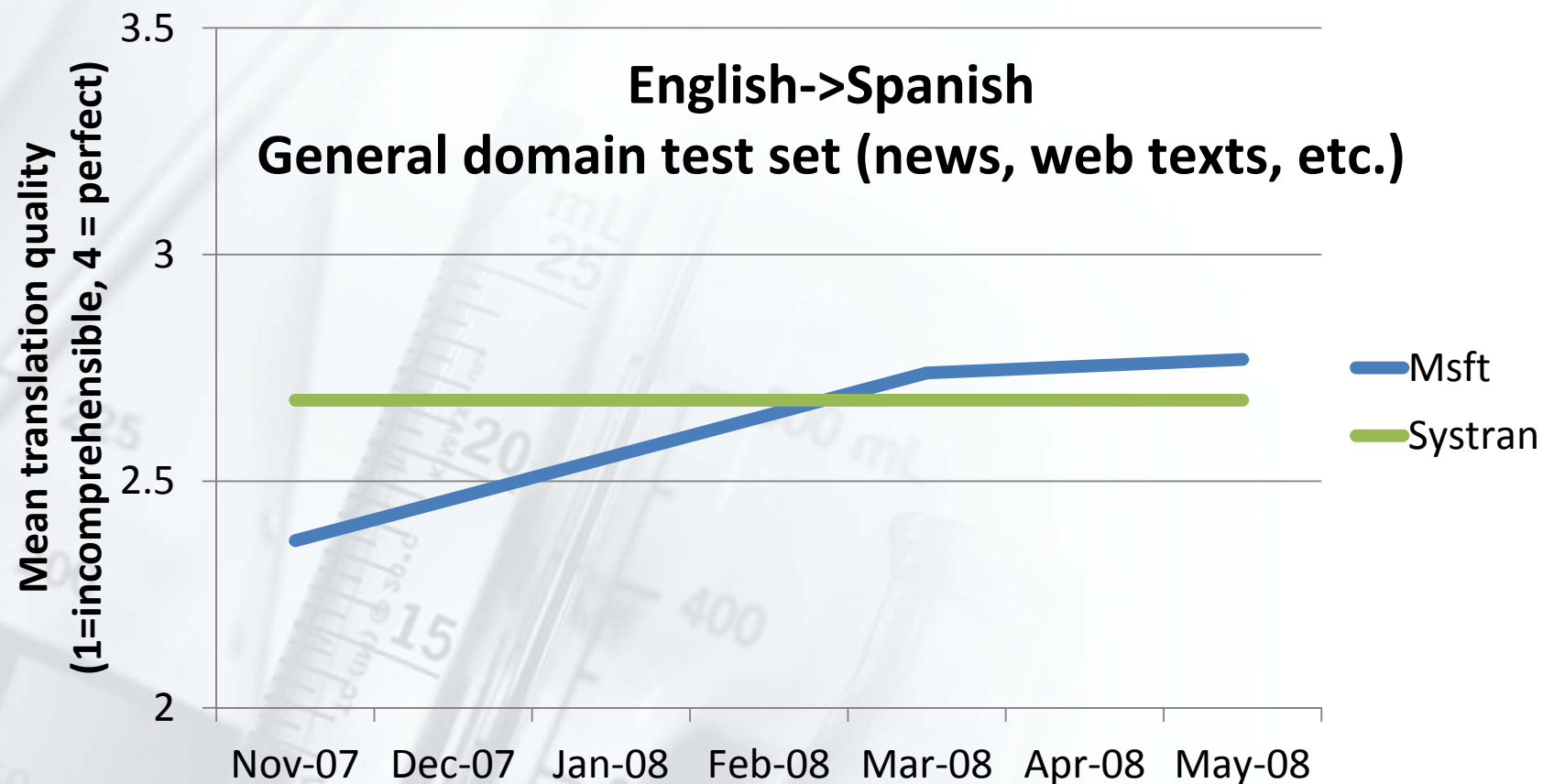
Rosetta Stone plaque, €35.00

Visit the online sh

English-Spanish translation quality: Microsoft technical texts



Bigger and Better Data → Better Translations



- Dramatic gains in Bing index size, quality over this period
- MT quality directly benefited from increasing volume of parallel pages



Haitian Creole

Developing MT for a Low Data Language



Haitian Creole

- One of two official languages in Haiti
- Evolved from French, Spanish, and several African languages (large % French-like)
- Spoken natively by most of Haiti's 8M people
- Recent as a written language (first literature dates to late 18th century), growing literature base
- Semi-literate population, with preference for French (until recently)
- Somewhat inconsistent orthography
- Limited (but growing) Web presence



Tranbleman tè nan Pòtoprens, kapital Ayiti!



Pòtoprens te catastrophically afekte 12 janvyè 2010 tranbleman tè a.

- The earthquake of January 12th, 2010 a significant humanitarian crisis.
- Aid agencies, foreign governments, a variety of NGOs, all responded *en masse*

- Need for translated materials critical, especially those related to medicine and the relief effort.
- Mission: 4636 text messages from the field (up to 5K/hour at peak) require rapid translation



AFP Photo

Moun ap fouye pami debri yon bilding ki kraze nan tranblemann' tè 12 Janvyè a.



The Plan

- Identify as much parallel data as we can find; start with
 - Bible
 - Data from Carnegie Mellon University (CMU)
 - Haitisurf.com
 - Official government documents, including constitution
 - Data identified by CrisisCommons
 - Parallel sentences from Creole-English Wiki pages
- Rally team to help process the data (and everything else!)
- Find linguistic experts in Creole to advise and help
- Find native speakers to review output and translate content
- Engage the relief community involved in the Haiti effort



Message Translation

Mwen rele FIRST LAST mwen

My name is FIRST LAST. I

se yon bòs mason

work in construction,

kay mwen kraze mwen gen

and I have four children.

kat pitit numero mwen

My number is 99999999.

se 99999999

Ki sa pou nou f? ak timoun

What can we do with the

yo kos?nan lekòl la e pui

children regarding school

kile moun duval nan croi

and when will the people

des bouket ap jwen manje

of duval in croix des

pou met nan vant yo

bouquets get food to put

in their bellies?

Voye kÄk konsÄy pou

Send me some advice.



Example of the Power of Data-Driven MT

- Systems improve by learning from human-produced translations
 - Adding more parallel data yields a better system
 - As the web grows, translation quality improves
 - Quality already exceeds best rule-based systems
- Given data, new language pairs can be launched very quickly
 - Haitian Creole <-> English: deployed in 4 days and 17 hours
 - A rule-based system would have taken months to build



Spell-Checking for Explicit Query-Level Dialog

Web Images Videos Shopping News Maps More | MSN Hotmail

bing

les angeles

Web Maps Weather Events Videos News Images

RELATED SEARCHES

- Los Angeles Galaxy
- Ncis Los Angeles
- Angeles Y Arcangeles
- Angeles Celestiales
- Angeles Del Cielo
- Angeles De La Guardia
- Fotos De Angeles
- Dibujos De Angeles

SEARCH HISTORY

See all
Clear all · Turn off

Web Images Videos Shopping News Maps More | MSN Hotmail

Web

Web Maps Weather Events Videos News Images

ALL RESULTS 1-20 of 179,000,000 results · [Advanced](#)

[Los Angeles Hotel Deals](#) - [www.priceline.com](#) Sponsored sites

The all new Priceline now lets you shop and compare before you buy!

Sponsored sites

[Travelodge Los Angeles](#)
Book at the official Travelodge® site for our best rates guaranteed.
[www.Travelodge.com/Los Angeles](#)

[los angeles](#)
Photos, Customer Ratings & Reviews. Save on Trips to Los Angeles, CA.
[www.expedia.com](#)

[Taxi Service 888-904-2345](#)
LA LAX South Bay Santa Monica Hollywood & Other Cities
[www.BellCab.com](#)

[California Luxury Hotels](#)
Indulge Yourself At A Luxury Resort Bid Online At Special Prices Now!
[LuxuryLink.com](#)

[Super 8® Fall Promotion](#)
Buy 2 & Get the Equivalent of Third Night Free at Super 8. Book Now.
[www.Super8.com/LosAngeles](#)

[See your message here](#)

Los Angeles, California

City Population: 3,694,820
Median household income: \$36,687
Median age: 31.6

Slideshow

Maps Weather Flight deals Attractions

64°F
75° / 61°
Cloudy

\$229
SEA > SNA
[See all deals](#)

Walt Disney Concert Hall
Norton Simon Museum of Art
Venice Beach's Ocean Front Walk
Peninsula Spa

SHARE [Facebook](#) [Twitter](#) [Messenger](#) [Email](#)

[Los Angeles - Wikipedia, the free encyclopedia](#)
History · Cityscape · Geography · Economy
Los Angeles is the second largest city in the United States, the largest city in the state of California and the Western United States, with a population of 3.83 million ...
[en.wikipedia.org/wiki/Los_Angeles](#) · [Cached page](#)

[Los Angeles Convention and Visitors Bureau](#)
The Official Guide to Los Angeles. News on great deals, free stuff and fun things to do. The most complete list of everything happening in LA.
[www.discoverlosangeles.com](#) · [Cached page](#)

[Los Angeles City Guide | Hotels, Restaurants & Nightlife | Attractions ...](#)
A comprehensive city guide for Los Angeles hotels, attractions, restaurants, nightlife, real estate and local business yellow page listings.
[www.losangeles.com](#) · [Cached page](#)

Lebron's bi
See what pe

Elements of Search Quality

Relevance



Ensuring that
best results
rank at top

Completeness

Freshness

Speed



How fast do
result pages
render?

Ease of Use



Simple
interface

Query & click



Semantic Impact to Go Beyond Search

Relevance



Selection and ranking based on **meaning** and **concepts**, not keywords

Direct answers

Speed



Reduce efforts to **task completion**

Direct answers

Fewer clicks

Ease of Use



Intuitive queries

Simplified tasks

Information aggregation & classification



Spatial Dialog

All Bing Vertical Services

The screenshot shows a Bing search result for "book of eli". The search bar at the top contains "book of eli" and the Bing logo. Below the search bar are navigation tabs for "Web", "Showtimes", "News", and "Wikipedia". The main content area displays a search result for "The Book of Eli near Redmond, Washington" with a movie poster and details: "Rated R · Action, Drama · 1 hrs 58 min · Overview · Review · ★★★★★", "Regal East Valley 13 · Renton · Map", and "M-Th 1:10PM, 4:05, 7:20, 10:10". To the right of the main result is a "Sponsored sites" section featuring "Book Of Eli at Amazon" with the text "Low Prices on Book of eli" and "Free 2-Day Shipping w/ Amazon Prime". On the left side, there are two boxes: "RELATED SEARCHES" listing items like "Books on Eli Whitney" and "The Book of Eli Movie", and "SEARCH HISTORY" showing "book of eli". Below the main result is a section for "The Book of Eli (2010 film)" with a poster, release date (2010), running time (118 minutes), genre (Action), and rating (R). It includes "Reviews" from IMDb (7 out of 10) and Rotten Tomatoes (46 out of 100), a "Cast" list (Denzel Washington, Gary Oldman, Mila Kunis, Ray Stevenson), and "Buy/Rent" options (Buy: Amazon, Rent: Netflix, Rent: Blockbuster). At the bottom, there is a Wikipedia link "The Book of Eli - Wikipedia, the free encyclopedia" and a detailed description of the film. A red box highlights the search bar and navigation tabs, and another red box highlights the sponsored site. Blue arrows point from the text boxes on the right to these elements.

Web Images Videos Shopping News Maps More | MSN Hotmail

bing MS Beta 10553

book of eli

Web Showtimes News Wikipedia

RELATED SEARCHES

- Books on Eli Whitney
- NVCC Alexandria Home
- Definition of Press
- NVCC My Nova
- NVCC Eli Bookstore
- Sons of Eli
- Samuel Eli
- The Book of Eli Movie

SEARCH HISTORY

book of eli

See all

Clear all · Turn off

[The Book of Eli near Redmond, Washington](#) change location

Rated R · Action, Drama · 1 hrs 58 min · Overview · Review · ★★★★★

Regal East Valley 13 · Renton · Map

M-Th 1:10PM, 4:05, 7:20, 10:10

Source: MSN Movies

SHARE Facebook Twitter Email

DELIVER US

The Book of Eli (2010 film)

Released: 2010 · Running time: 118 minutes [1] · Genre: Action · Rated: R

See more from: [Wikipedia](#) · [IMDb](#) · [Rotten Tomatoes](#)

Reviews	Cast	Buy/Rent
7 out of 10 IMDb (27542 reviews)	Denzel Washington Gary Oldman Mila Kunis Ray Stevenson	Buy: Amazon Rent: Netflix Rent: Blockbuster
46 out of 100 Rotten Tomatoes (172 reviews)		

[The Book of Eli - Wikipedia, the free encyclopedia](#)

Plot · Production · Reception · Home media

The **Book of Eli** is a 2010 American post-apocalyptic action film directed by the Hughes brothers, written by Gary Whitta, and starring Denzel Washington, Gary Oldman, Jennifer Beals ...

en.wikipedia.org/wiki/The_Book_of_Eli · Wikipedia on Bing · Mark as spam

[The Book of Eli \(2010\)](#)

User rating: 7/10 · Action/Adventure/Drama/Thriller/ · R · 118 min

A post-apocalyptic tale, in which a lone man fights his way across America in order to protect a sacred **book** that holds the secrets to saving humankind.

www.imdb.com/title/tt1037705 · Cached page · Mark as spam

Sponsored sites

[Book Of Eli at Amazon](#)

Low Prices on **Book of eli**

Free 2-Day Shipping w/ Amazon Prime

www.Amazon.com/Books

Quick Tabs for relevant Bing verticals, domains, answers

Entity-based result summary

Spatial-Temporal Dialog: Re-Rank with Session

The screenshot shows a Bing search results page for the query "toyota prius". The page is annotated with three blue callout boxes on the right side, each pointing to a specific feature:

- History-aware Autosuggest:** Points to the left sidebar, which displays search suggestions such as "honda civic", "2010 Honda Civic", "Honda Civic Problems", "Honda Civic Si", "Honda Accord", "Honda Fit", "2010 Honda Civic", "1979 Honda Civic", "Cadillac Escalade", and "New Honda Civic Cars".
- History-aware Compare:** Points to the "Advanced" link in the top right corner of the search results area.
- History-aware Implicit Compare:** Points to a search result titled "honda civic vs. toyota prius | Hybrid Cars" at the bottom of the page, which includes a snippet: "Have anyone done research comparing the civic and the prius? Which is he better vehicle? www.hybridcars.com/forums.honda-civic-vs-t421.html".

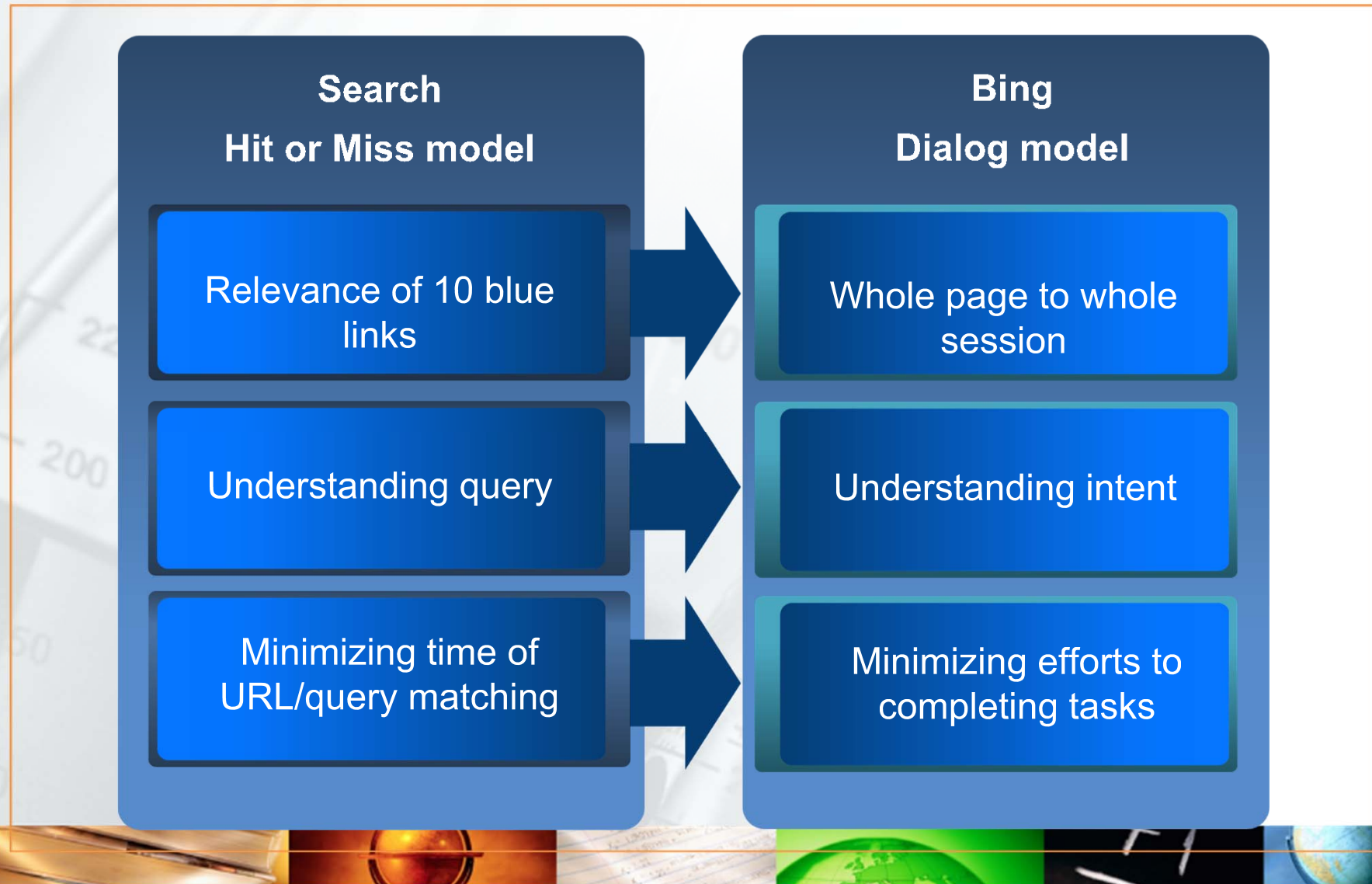
The main search results area displays "ALL RESULTS" for "toyota prius" with 1-10 of 2,270,000 results. It includes sponsored sites, news articles, videos, and related searches. The "RELATED SEARCHES" section lists items like "Toyota Prius Honda", "Toyota Prius Hybrid", "Toyota Camry", "Toyota Matrix", "Toyota Tundra", "Toyota Car Models", and "2010 Toyota Prius". The "SEARCH HISTORY" section shows previous searches for "toyota prius", "honda civic", "jason kidd", "carmelo anthony", "You've also tried toyota", and "See all".

History-aware Autosuggest

History-aware Compare

History-aware Implicit Compare

Evolution of Search Evolution: Organizing the Web for Tasks



Breaking Down Data Barriers: Working with the NSF

Data and Data Services
as Innovation Enablers



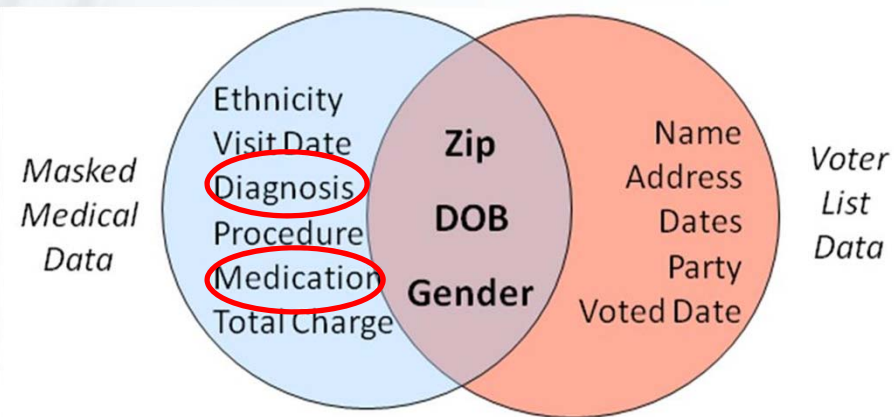
Challenges for Data-driven Research?

- Much of the data needed for data-driven research is locked away in industry vaults
 - Reasons: privacy, scale, business sensitivity
- How can we help drive innovative research in the world of cloud-based data services?
- Can semantic computing help ?

Cautionary Example #1: Privacy loss with Public Cross-linked Data Sets



The Massachusetts Governor Case



He was the only male in his zipcode with his Date of Birth



Online Privacy?

We leave our traces online at multiple sites such as social networks, blogs, forums etc.

- E.g. Re-identify users from movie mentions in forums to user ratings of movies [Frankowski'06]

BobZ	Some Recommendations Reply Reply with Quote
Your predictions:	Sep 21, 2005 5:29:45 PM
Life Aquatic .. <input type="checkbox"/> ★★★★★ 4.5 stars	I've enjoyed some great movies from Netflix recently. Last night I watched The Life Aquatic with Steve Zissou , which was quirky and funny. The best comedy I've seen in years.
Finding Never .. <input checked="" type="checkbox"/> ★★★★★ Not seen	Earlier in the week I saw Finding Neverland . Also a great movie.



Search Queries and Privacy

Why not just release Search Logs to researchers?

- **Problem is that search Logs can be privacy revealing**
 - Search queries (free text) themselves may contain Personally Identifiable Information
 - Sanitizing all search entries is not possible

Cautionary Example #2: Search Queries and Privacy

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.




Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches,"

 SIGN IN TO E-MAIL THIS

 PRINT

 SINGLE PAGE

 REPRINTS

ARTICLE TOOLS
SPONSORED BY



Data Workshops – Program Solicitation (Evelyne Viegas)

Goals

- Break Down Data barriers
- Enable data-driven research

NSF Directorates

Directorate for Computer and Information Science and Engineering (CISE)

Directorate of CyberInfrastructure (OCI)

Directorate for Social, Behavioral, and Economic Sciences (SBE)

Data Confidentiality 2007

<http://dcws.stat.cmu.edu/index.html>

- Participation from 13 federal agencies; 7 industries; 18 universities
- [Trustworthy Program Solicitation](#)

Confidential Data Collection for Innovation Analysis in Organizations 2009

<http://www.lrdc.pitt.edu/schunn/cdi2009/home.html>

- Cross-disciplinary between computer science & social sciences (cognitive psychology, economy)

NSF Program Solicitation Computing in the Cloud 2010

- Web N-gram Service (access to Bing Index)
<http://research.microsoft.com/web-ngram>
 - 10 to 15 awards in FY 2010



RFPs Program Feedback

- Researchers in Academia need **access** to large scale real world data, and infrastructure to drive innovation, enable science (repeatability)
 - [Search Summit 2007](#) new asks:
 - Need more data, larger scale;
 - Need to follow a user (privacy!)
 - [Beyond Search – Semantic Computing and Internet Economics 2009](#) new asks:
 - Need data access (as opposed to data release);
 - Compute power

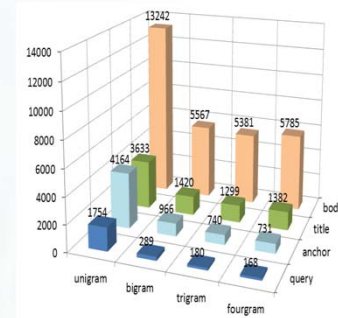
Web N-gram Services

Access to *petabytes* of real world data

<http://research.microsoft.com/web-ngram>

Leading technology in Search, Machine Translation,
Speech, Learning, ...

Web data has structure – and that difference counts



Web N-Gram in Public Beta

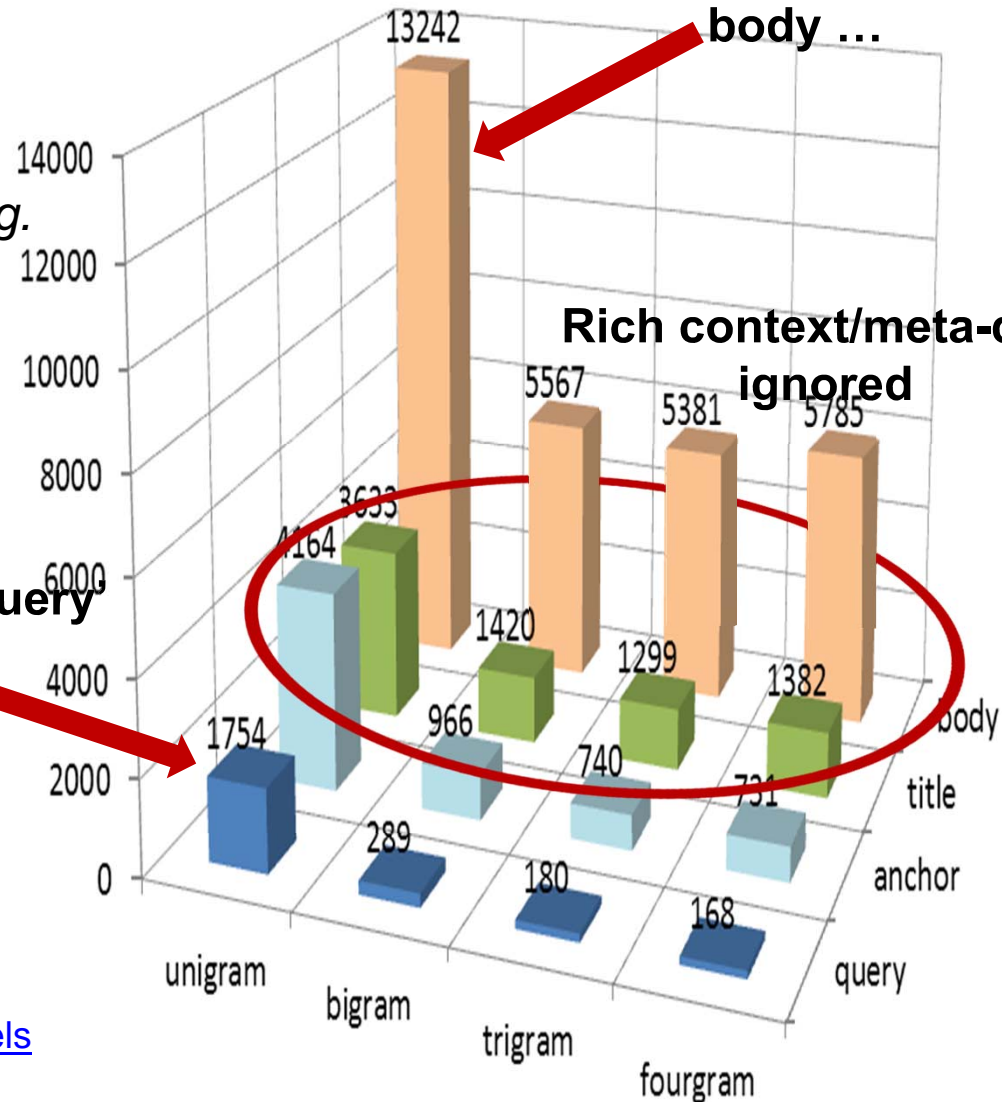
Web data has structure...

...and that counts (e.g. Body, Title, Anchor)

Users form 'query'

Search engines rely on unigram body ...

Rich context/meta-data ignored



Web N-gram Offering

<http://research.microsoft.com/web-ngram>

- Content types
 - Document Body, **Document Title, Anchor Texts**
- Model types
 - **Smoothed models**
- N-gram availability
 - unigram, bigram, trigram, N-gram with N=4, 5
- Training size (Body):
 - **All documents** indexed by Bing (no cut off)
- Access
 - **Hosted Services** by Microsoft
- Updates
 - **Periodical updates**

Word Breaking Made Easier?

Enter a hash-tag phrase, and we will show the likely breakdown of sub-words. For instance, enter #nowplaying. More examples...

#whenifirstmet #nowplaying #wtfyoumean #thissummer #enoughisenough #ifirstmet #riphtherunway #complimentgonebad
 #SMHyoureghetto #letmefindout #idoit2 #itaintmyfault #FlavoredCondoms #jayparkaom #ChrisBrownRocks #thingthatihate #nowplaying
 #whenifirstmet #hereugo #stpatriksday #thissummer #hiphopaintdead #idoit2 #sexisthebest #Lupequotes #WillYouEver
 #flavoredcondoms #whenimeetjustin #hcr #FF #Nowplaying #followfriday #howyouathug #youaintforme #OhJustLikeMe #NotMeThough
 #HCR #idoit2 #yeaisaidit #Advice #iloveitwhentrey #MarchMadness #TLS #ihatequotes #s1battle #nowplaying #howyouathug
 #uaintforme #youaintforme #WhenIfirstmet #whatsworse #WhenTwitterWasDown #howuathug #ChrisBrownonUstream #hereugo #TLS
 #justinbiebermyspace #idoit2 #HCR #willyouever #marchmadness #Hereyougo #nowplaying #imthekindofperson #FF #6wordstory
 #whitecusswords #whoelsenoticed #yeaisaidit #hcr #idoit2 #ss3forindonesia #Ohjustlikeme #blackcusswords #theboltonnews
 #ss2malaysia #FollowFriday #arashi #StopHatingDemi #mucoreSNSD #nowplaying #imthekindofperson #MJis #whitecusswords
 #OhJustLikeMe #idoit2 #thankstwitter4 #YourUnderArrest #hcr #BounceBackTeuk #inschool I #Imliableto #DontBeMadBut
 #becauseofbieber #ChrisBrownonUstream #hbu #nowplaying #dearfuturewife #imthekindofperson #musicmonday #Isitjustme
 #goseethedoctor #hcr #idoit2 #thankstwitter4 #MM #OhJustLikeMe #TLS #ohmySiWon #thatisall #ihatequotes #afmlmoment
 #biebermemories #tellmewhyumad

Phrase	LgProbability
yea i said it	-9.345904
yeaisaidit	-10.42589
yeai said it	-11.31242
yea isaid it	-12.04566
ye a i said it	-13.61018

Phrase	LgProbability
when i first met	-6.974892
when ifirstmet	-10.34817
when ifirst met	-10.67689
when i firstmet	-11.09351
wheni first met	-11.1378

Phrase	LgProbability
w8 4 u	-10.0969
w84u	-10.27723
w 84u	-10.69117
w 84 u	-10.7444
w 8 4 u	-11.06896

Multi-word Tag Cloud from Government Dataset Titles

Single Tag Cloud

Multi Tag Cloud



Ref: Dr. Li Ding, Rensselaer Polytechnic Institute



INFORMATION

Language Models

WEB N-GRAM SEGMENTATION DEMO

Home About

Select a model, enter a phrase, and press the Segment button to see the suggestions.

Model:

Phrase: raleigh serengeti mountain bike

-12.47305 raleigh serengeti mountain bike
-14.41699 raleigh serengeti mountain bike
-15.77088 raleigh serengeti mountain bike
-15.91565 raleigh serengeti mountain bike
-17.0868 raleigh serengeti mountain bike
-17.8378 raleigh serengeti mountain bike
-19.19169 raleigh serengeti mountain bike
-20.5294 raleigh serengeti mountain bike

WEB N-GRAM SEGMENTATION DEMO

Home About

Select a model, enter a phrase, and press the Segment button to see the suggestions.

Model:

Phrase: raleigh serengeti mountain bike

-15.32104 raleigh serengeti mountain bike
-15.42312 raleigh serengeti mountain bike
-17.6154 raleigh serengeti mountain bike
-18.2401 raleigh serengeti mountain bike
-18.28114 raleigh serengeti mountain bike
-18.38321 raleigh serengeti mountain bike
-24.6635 raleigh serengeti mountain bike
-25.28819 raleigh serengeti mountain bike

“Raleigh Serengeti” recognised as an entity using Anchor Text and Document Title, and unlike using Body

WEB N-GRAM SEGMENTATION DEMO

Home About

Select a model, enter a phrase, and press the Segment button to see the suggestions.

Model:

Phrase: raleigh serengeti mountain bike

-14.84092 raleigh serengeti mountain bike
-15.29377 raleigh serengeti mountain bike
-15.89154 raleigh serengeti mountain bike
-16.1347 raleigh serengeti mountain bike
-17.78695 raleigh serengeti mountain bike
-18.23981 raleigh serengeti mountain bike
-24.09498 raleigh serengeti mountain bike
-24.33814 raleigh serengeti mountain bike

Semantic Computing to enable Implicit Search

VINOGRAPHY: a wine blog
Wine and food adventures in San Francisco and around the world

WINE REVIEWS | RESTAURANT REVIEWS | BOOK REVIEWS | RAMBLINGS & RANTS | WINE NEWS

Breaking Wine News: Bordeaux's Cos d'Estournel Buys Napa's Chateau Montelena

To those of you in the wine world paying attention to the dollar's stomach churning lows against the Euro, this news may come as little or no surprise. This morning, Chateau Cos d'Estournel announced its purchase of the historic **Chateau Montelena** in Napa. While not the first bit of investment from Bordeaux in the **Napa Valley** is certainly a significant one, given both the landmark historical status of Chateau Montelena as well as the prestige and success of Cos d'Estournel, whose star has certainly been rising in Bordeaux over the past decade.

Montelena became a world famous winery after its 1973 Chardonnay beat out French competitors in the famous **Judgement of Paris**

Wikipedia: The Free Encyclopedia

BOTTLE SHOCK

Wine Spectator

- 17.07376 Chateau Montelena in Napa
- 17.28525 Chateau Montelena in Napa
- 17.36758 Chateau Montelena in Napa
- 17.49432 Chateau Montelena in Napa
- 22.04415 Chateau Montelena in Napa
- 22.10234 Chateau Montelena in Napa
- 22.25322 Chateau Montelena in Napa
- 22.39616 Chateau Montelena in Napa

'Chateau Montelena in Napa'
segmentation

article | discussion | edit this page | history

Chateau Montelena

From Wikipedia, the free encyclopedia

Coordinates: 38°40′16.5″N 122°09′53″W﻿ / ﻿38.67125°N 122.16472°W﻿ / 38.67125; -122.16472

Chateau Montelena is a Napa Valley winery most famous for winning the white wine section of the historic "Judgement of Paris" wine competition. Chateau Montelena's Chardonnay was in competition with nine other wines from France and California under blind tasting. All 11 judges awarded their top scores to either the Chardonnays from Chateau Montelena or Chalone Winery, another Californian wine producer. Chateau Montelena was featured in the 2008 film *Bottle Shock*.

Contents [hide]

- History
 - 1.1 Terminated sale
- See also
- References
- External links

Chateau Montelena

Location	Calistoga, California, USA
Appellation	Napa Valley AVA
Other labels	Potter Valley
Founded	1882
Key people	Jim Barrett, Winemaker Bo Barrett, Winemaker Greg Rabston, Managing Director Dave Vella, Vineyard Manager
Cases/yr	30,000 - 36,000
Varietals	Chardonnay, Zinfandel, Cabernet Sauvignon, Riesling
Website	www.montelena.com

'Chateau Montelena' as an
entity
in Wikipedia

Semantic Computing and Science Applications



Science Examples

- Add-ins for Word
- MT and WorldWideScience.org
- ML and Diseases
- Tagging and Astronomy

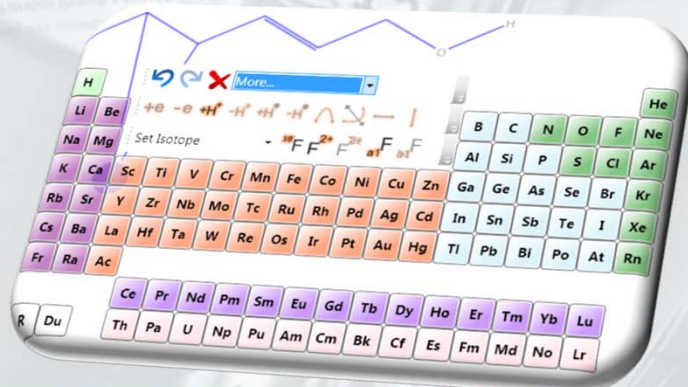


Chemistry Add-in for Word

1.2.1 Preparation of (2E)-3-(3-methyloxiran-2-yl)prop-2-en-1-ol (173)

Diisobutylaluminium hydride (17.70 ml of a 1 M solution in THF, 17.70 mmol) was slowly added to the epoxide ester **172** (1.20 g, 7.68 mmol) in THF (10 ml) at -78°C. After stirring at this temperature for 1h, methanol (10 ml) was added slowly and the resultant solution was warmed to rt. TEA (8 ml) was subsequently added and the mixture was stirred at rt overnight. Filtration through a pad of Celite® followed by washing with Et₂O (150 ml) and concentration *in vacuo* provided the crude product which was purified by flash column chromatography (eluent PE:Et₂O 4:1 to 1:4, gradient) to give alcohol **173** (718 mg, 6.29 mmol, 82%) as a colourless oil; ν_{max}

- Authoring and rendering of semantic-rich chemical information ([CML](#))
- In partnership with the University of Cambridge
- Support for Office 2007 and Office 2010
- Available under Apache 2.0
- Over **200K** [downloads](#) since March 22nd, 2010



Ontology Add-in for Word



- John Wilbanks

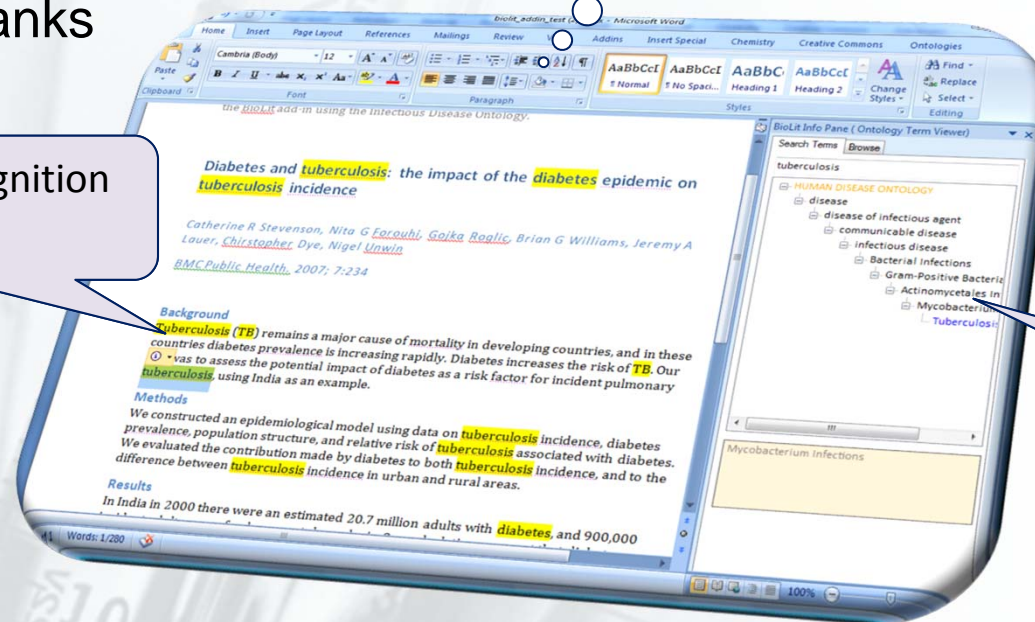
Services: Ontology download web service



University of California
San Diego

- Phil Bourne
- Lynn Fink

Intent: Term recognition & disambiguation



Relationships:
Ontology browser

Downloads = 4,000+

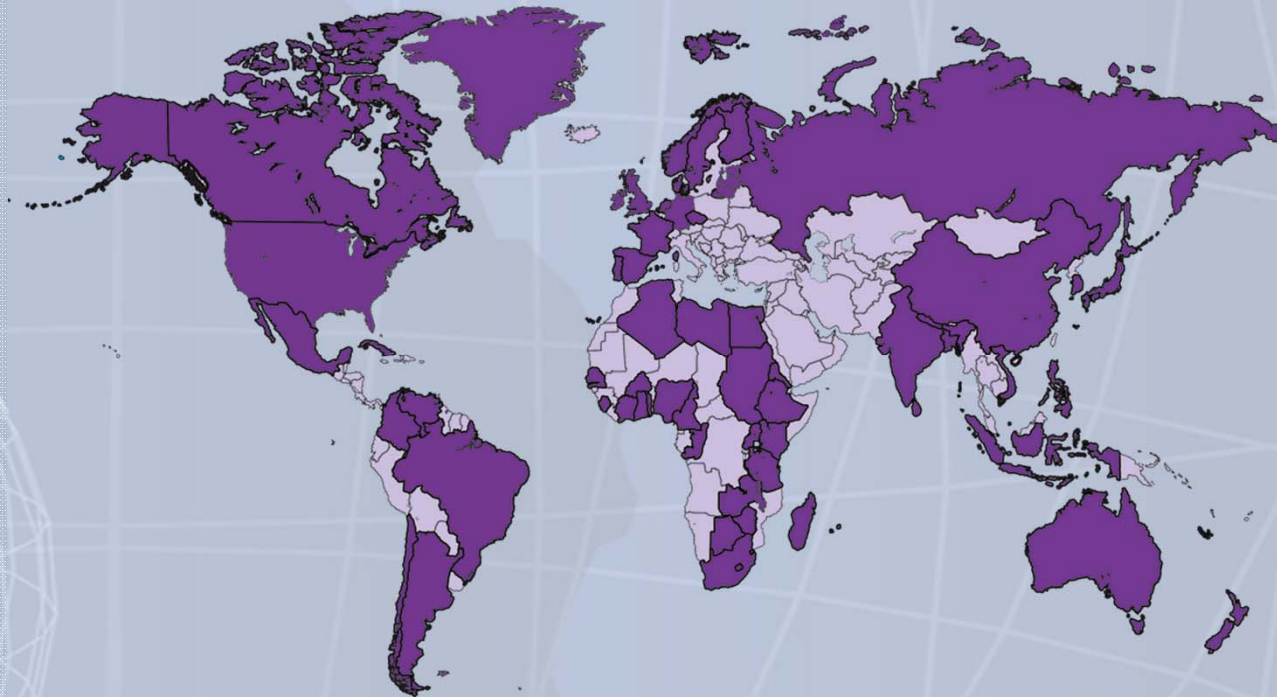
Source code + binary:

<http://research.microsoft.com/ontology/>



WorldWideScience – Facts and Figures

- Tremendous growth in search content: from 10 nations to 65 nations in 3 years
- > 400 million pages
 - From well-known sources: e.g., PubMed, CERN, KoreaScience
 - To more obscure sources: e.g., Bangladesh Journals Online



Now, we have the essential ingredients for real-time translation of science

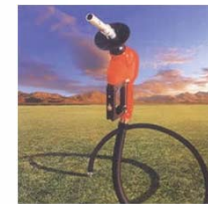
- National science databases in multiple languages
- Federated search
- Multilingual translation on both front and back end of the user experience

A public-private partnership, introduced as ***Multilingual WorldWideScience.org^{Beta}***



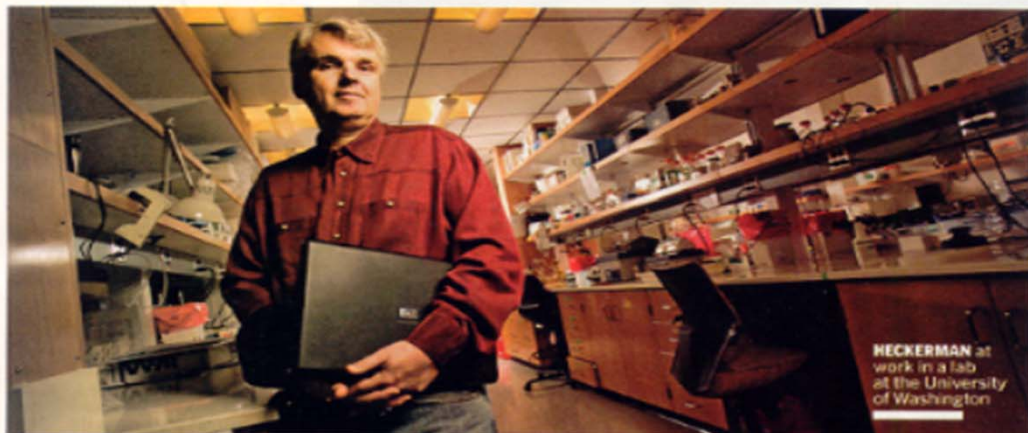
Tackling societal challenges with ML

- Fighting HIV/AIDS, H1N1, etc.
- Identifying genetic and environmental causes of disease
 - Diabetes, asthma, ALS, aging



David Heckerman and eScience
Research Group





HECKERMAN at work in a lab at the University of Washington

Using Spam Blockers To Target HIV, Too

A Microsoft researcher and his team make a surprising new assault on the AIDS epidemic

BY STEPHEN BAKER AND JAY GREENE

CUT-RATE PAINKILLERS! Unclaimed riches in Nigeria! Most of us quickly identify such e-mail messages as spam. But how would you teach that skill to a machine? David Heckerman needed to know. Early this decade, Heckerman was leading a spam-blocking team at Microsoft Research. To build their tool, team members meticulously mapped out thousands of signals that a message might be junk. An e-mail featuring "Viagra," for example, was a good bet to be spam—but things got complicated in a hurry.

If spammers saw that "Viagra" messages were getting zapped, they switched to Viagra, or Vi agra. It was almost as if spam, like a living thing, were mutating.

This parallel between spam and biology resonated for Heckerman, a physician as well as a PhD in computer science. It didn't take him long to realize that his spam-blocking tool could extend far beyond junk e-mail, into the realm of life science. In 2003, he surprised colleagues in Redmond, Wash., by refocusing the spam-blocking technology on one of the world's deadliest, fastest-mutating conundrums: HIV, the virus that leads to AIDS.

Heckerman was plunging into medicine—and carrying Microsoft with him. When he brought his plan to Bill Gates, the company chairman "got really excited," Heckerman says. Well versed on HIV

from his philanthropy work, Gates lined up Heckerman with AIDS researchers at Massachusetts General Hospital, the University of Washington, and elsewhere.

Since then, the 50-year-old Heckerman and two colleagues have created their own biology niche at Microsoft, where they build HIV-detecting software. These are research tools to spot infected cells and correlate the viral mutations with the individual's genetic profile. Heckerman's team runs mountains of data through enormous clusters of 320 computers, operating in parallel. Thanks to smarter algorithms and more powerful machines, they're sifting through the data 480 times faster than a year ago. In June, the team released its first batch of tools for free on the Internet.

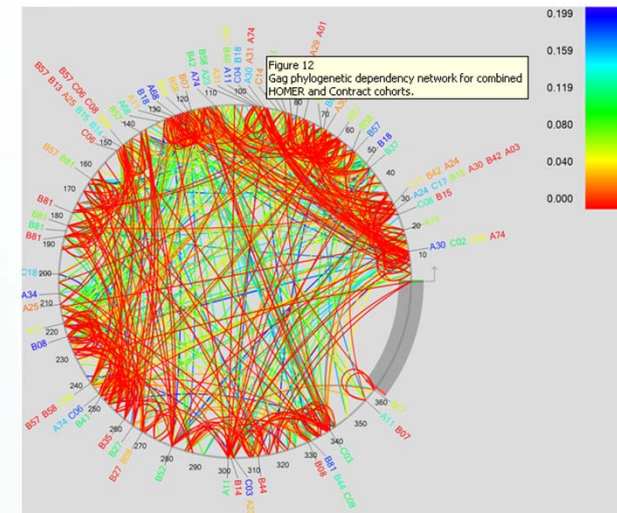
A new industry for the behemoth to conquer? Not exactly. Heckerman's nook in Redmond represents just one small node in a global AIDS research effort marked largely by cooperation. "The Microsoft group has a different perspective and a good statistical background," says Bette Korber, an HIV researcher at Los Alamos National Laboratories. The key quarry they all face is the virus itself, which is proving willier than any of Microsoft's corporate foes. While Heckerman has high hopes that his tools will lead to vaccines that can be tested on humans within three years, his research

Similar mutations may crop up in computer and medical viruses



Fighting HIV with ML and HPC

- PhyloD.Net is a Bayes-net-based tool that deciphers evolution of HIV within a patient
- Developed by eScience research group and published in *Science*, March 2007
- Now used by dozens of HIV research groups
- Led to discovery of two key insights to fight HIV:
 - Our immune system attacks frameshift epitopes, which may be useful to include in a vaccine (*JEM*, 2010)
 - Natural killer cells directly attack HIV (*Nature Medicine*, in review)
- Typical runs require CPU years, but delightfully parallel and runs well on our HPC servers
- Can also now run PhyloD in the Cloud as an Azure application



PhyloD.Net on cover of *PLoS Comp Bio*, Nov 2008 Carlson, Kadie, Heckerman et al.

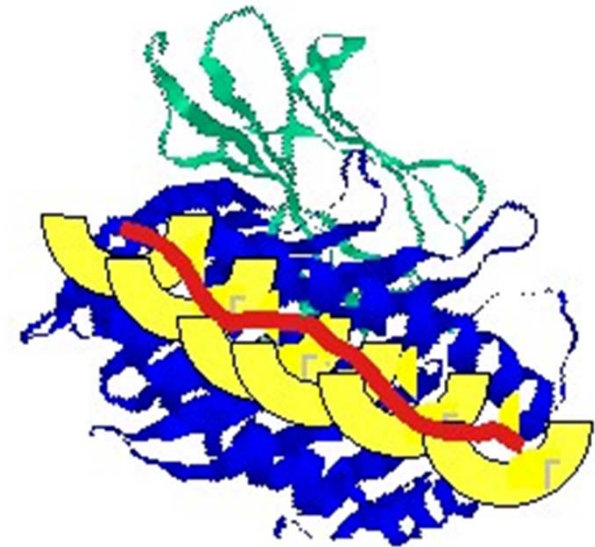


Better understanding of viruses through ML

We have discovered that DNA patterns shared among viruses are more readily attacked by our immune systems

This helps to explain

- Why H1N1 killed more people in Mexico (collaboration with Fred Hutchinson)
- Why only some patients get Dengue Hemorrhagic fever, and why some HIV patients have higher HIV viral load (collaboration with Perth Royal Hospital)
- The relationship between Rubella and other viruses (with CDC)



Understanding Asthma and Diabetes

Goal:

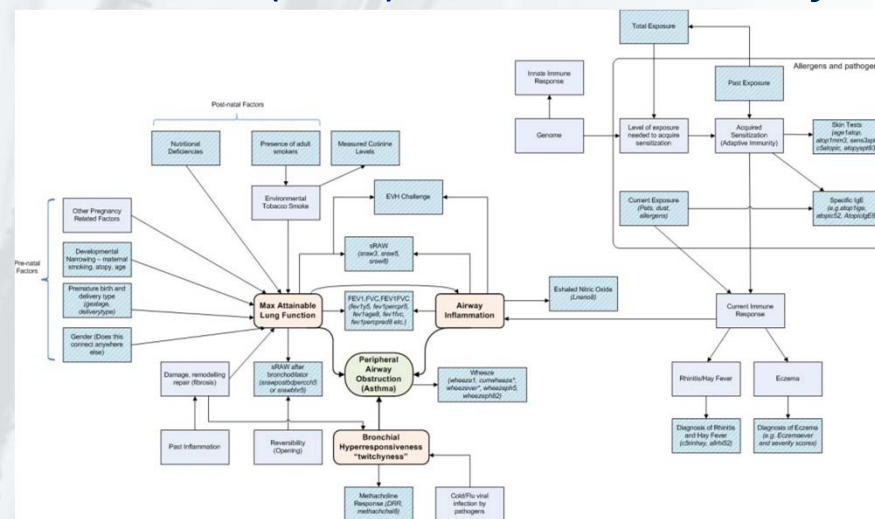
Understand environmental and genetic factors responsible for asthma and diabetes

Collaborations with University of Manchester (MSR Cambridge and eScience group) and with Sanger (MSR Cambridge)

ML challenge:

Hundreds of thousands of variables → graphical models, infer.net toolkit
Discovered new analysis algorithms that can find three times as much signal as previously and filter out spurious associations

A Simpson, V. Y. F. Tan V, J. Winn, M. Svensén, C. M. Bishop, D. E. Heckerman, I. Buchan, and A. Custovic (2009). *American Journal of Respiratory and Critical Care Medicine*



Machine Learning in the Galaxy Zoo Database

Kirk Borne

George Mason University



GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Merger Zoo (release November 2009)

- <http://mergers.galaxyzoo.org/>
- Run N-body simulations to find best model to match a real merger
- One new merger every day



ZOONIVERSE
REAL SCIENCE ONLINE

<http://zooniverse.org/>



Key Feature of Zooniverse: Data mining from the volunteer-contributed labels

- Train the automated pipeline classifiers with:
 - Improved classification algorithms
 - Better identification of anomalies
 - Fewer classification errors
- Millions of training examples
- Hundreds of millions of class labels
- Statistics deluxe! ...
 - Users (see paper: <http://arxiv.org/abs/0909.2925>)
 - Uncertainty quantification
 - Classification certainty vs. Classification dispersion



First Case Study: test SDSS science catalog attributes to find which attributes correlate most strongly with user-classified mergers.



NASA, ESA, the Hubble Heritage (AURA/STScI)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University)

STScI-PRC08-16a



Semantic Computing and Healthcare Applications



Healthcare Examples

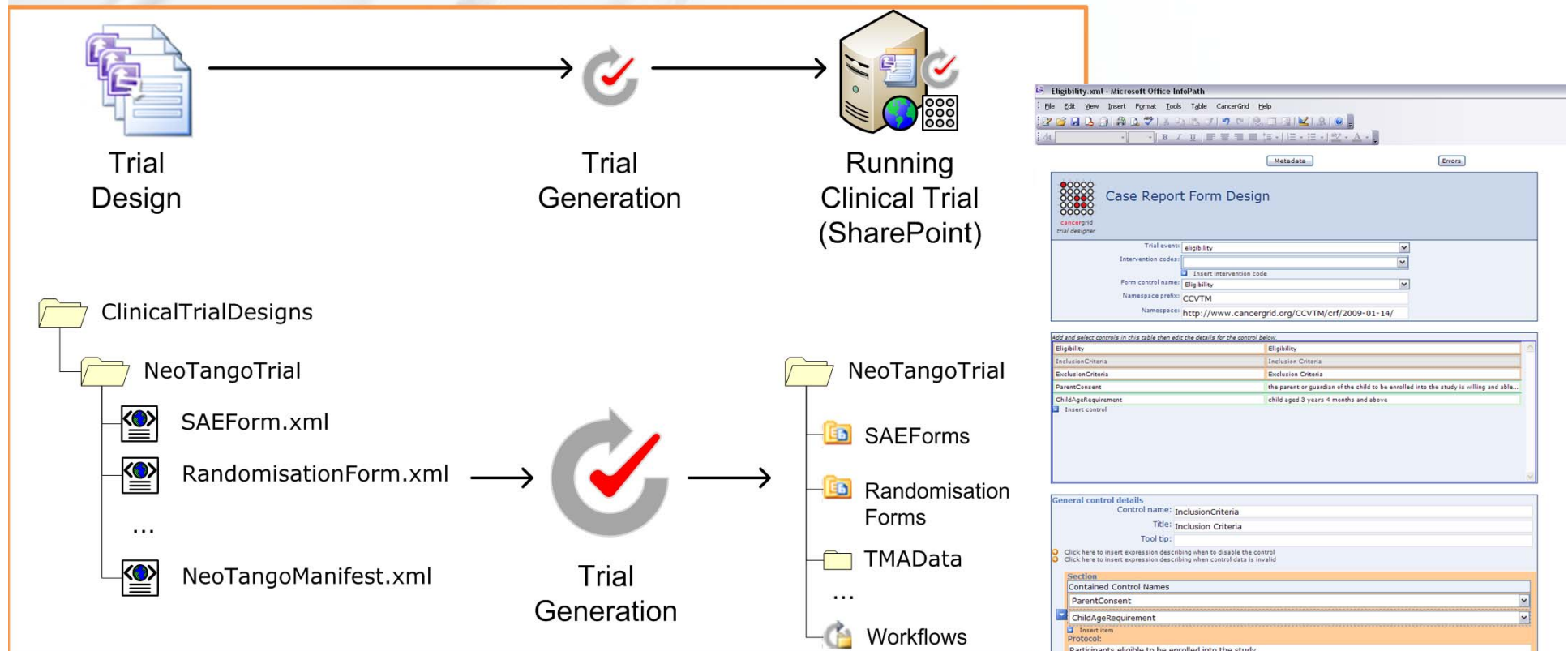
- Clinical Trials software project
- ML and Hospital Readmissions
- Amalga Life Science platform



Semantically Driven Forms for Clinical Trials

(With thanks to Jim Davies and Jeremy Gibbons Oxford University)

- Supports rapid online development of clinical trials
- Ends the “reinvention” process for individual trials by systemizing the data into a trials ontology
- Users select only relevant pieces and forms are auto-generated



Hospital Readmissions

Hospitals Pay for Cutting Costly Readmissions

By REED ABELSON

Published: May 8, 2009

It is one of the biggest avoidable costs on the nation's medical bill.

 [Enlarge This Image](#)



Dawn Vilella for The New York Times
Adeline and Chester Patyk of Plymouth, Minn., log their weight at home.

Millions of patients each year leave the hospital only to return within weeks or months for lack of proper follow-up care. One in five [Medicare](#) patients, for example, [returns to the hospital within 30 days](#). Over all, readmissions cost the federal government an estimated \$17 billion a year.

But even when [hospitals](#) find ways to



Learning from a Rich Case Library

- Microsoft Research project (MSR-Redmond & MSR-NE)
- Data from Washington Health Center hospitals (DC)
- All ED visits during the years 2001 to 2009 (~300,000 visits)
 - Patient's complaint (A string with 3-4 words on average).
 - Age, gender
 - Admitting and attending MD's code.
 - Length of stay in the ED.
 - Class of visit ("Emergency" or "Inpatient").
 - Date and time of discharge.
 - Diagnosis (Up to 10 diagnosis codes that are sorted with decreasing priority.)
The codes are based on ICD9 coding system.
 - Lab results.
 - ...

Bayati, Braverman, and Horvitz



Towards Fielding an Advisory Tool

Pre-discharge readmission risk assessment

Show Predictions Show Outcomes look up:

Account ID: Prediction: 36.11% high Outcome: 30 days

Next Patient

Next BB

Back Forward

Patient Info
27F Complaint: "esrd hyperglycemia abd pain"

Evidence for readmission

- 1) Dx1->1 = Nephritis, nephrotic syndrome, and nephrosis
- 2) Dx1->2 = Chronic renal failure
- 3) Was admitted during last year
- 4) Was admitted in the past
- 5) Dx0->2 = Diabetes mellitus

Evidence against readmission

- 1) Num DxCodes = 10
- 2) 24 < Age < 45
- 3) Num DxCodes is > 4
- 4) Sex=F

Recommendation

Significant risk of readmission. An aggressive follow-up program is recommended. Click on the button to the right to initiate the program.

Initiate aggressive follow-up care

Bayati, Braverman, and Horvitz



Amalga Life Sciences

- Accelerate research velocity
 - Ability to see all data
 - Data provenance via metadata
 - Semantic modeling and analysis of data
- Provides Actionable Knowledge
 - Flexible data integration
 - Semantic querying and reasoning
 - Identification of novel relationships

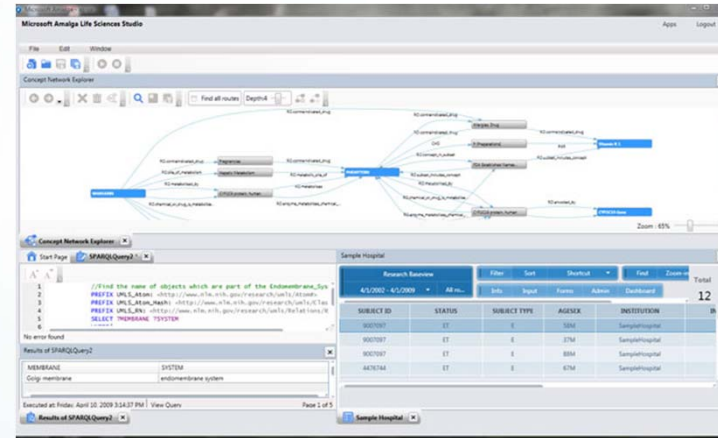


Microsoft Amalga Life Sciences 2009

Microsoft Amalga Life Sciences 2009 allows research and development organizations to aggregate data from disparate systems, both within their own institutions and from partner organizations, helping them move faster, with more agility, and with purpose and direction supported by validated facts. This allows researchers to address many data challenges from a single system and transforms the way they do research.



Amalga Life Sciences



Open platform

- APIs allow access by third-party software, including open source components
- Data and knowledge can be shared among different users, organizations, and applications.





Semantic Computing and the Cloud

Towards a Smart Cyberinfrastructure



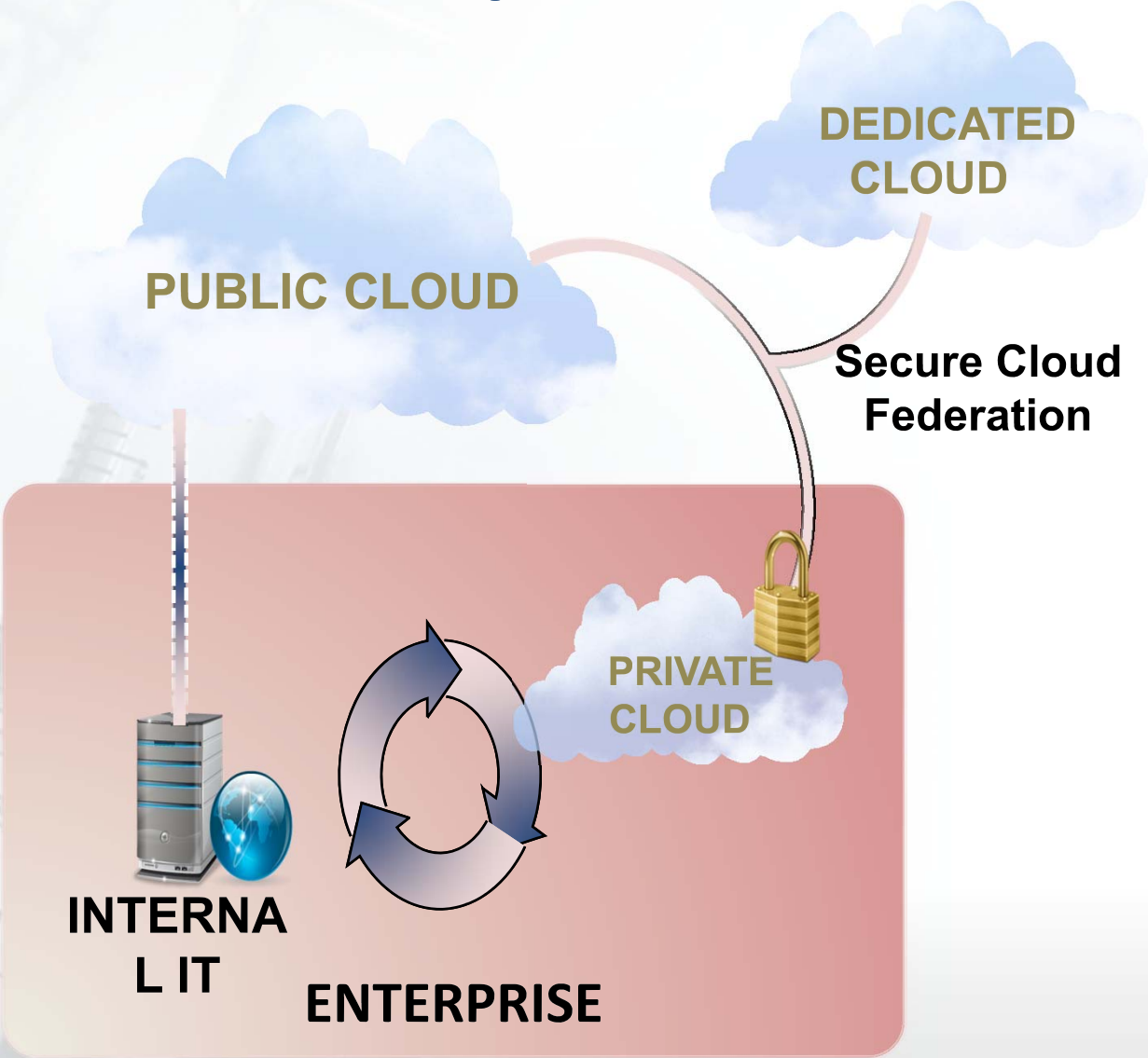
Cloud Computing: One Definition

For the US National Institute of Standards and Technology (NIST), Cloud Computing means:

- On-demand service
- Broad network access
- Resource pooling
- Flexible resource allocation
- Measured service



Cloud Options



Tomorrow...

Computers will still be great **tools** for

huge amounts of **data**



We would like computers to also help with the **automatic**

of the world's **information**



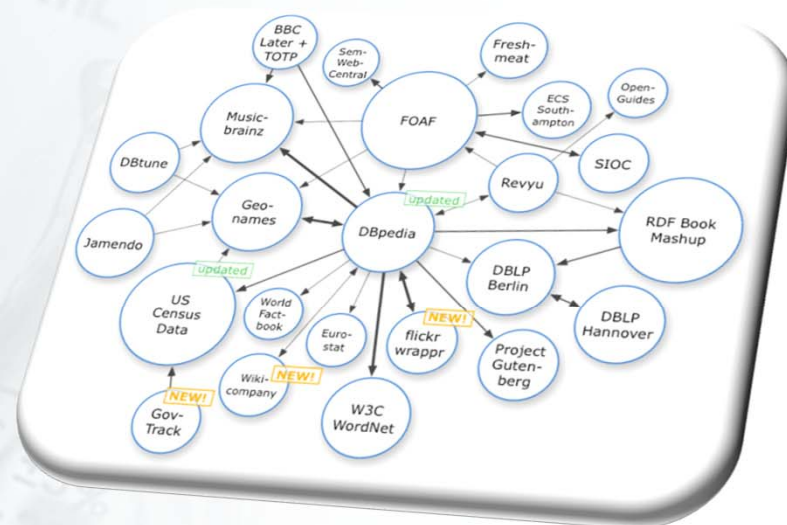
A world where all data is linked...



- Data/information is interconnected through machine-interpretable information (e.g. **paper X is about star Y**)
- Social networks are a special case of 'data meshes'

- **Important/key considerations**

- Formats or "well-known" representations of data/information
- Pervasive access protocols are key (e.g. HTTP)
- Data/information is uniquely identified (e.g. URIs)
- Links/associations between data/information

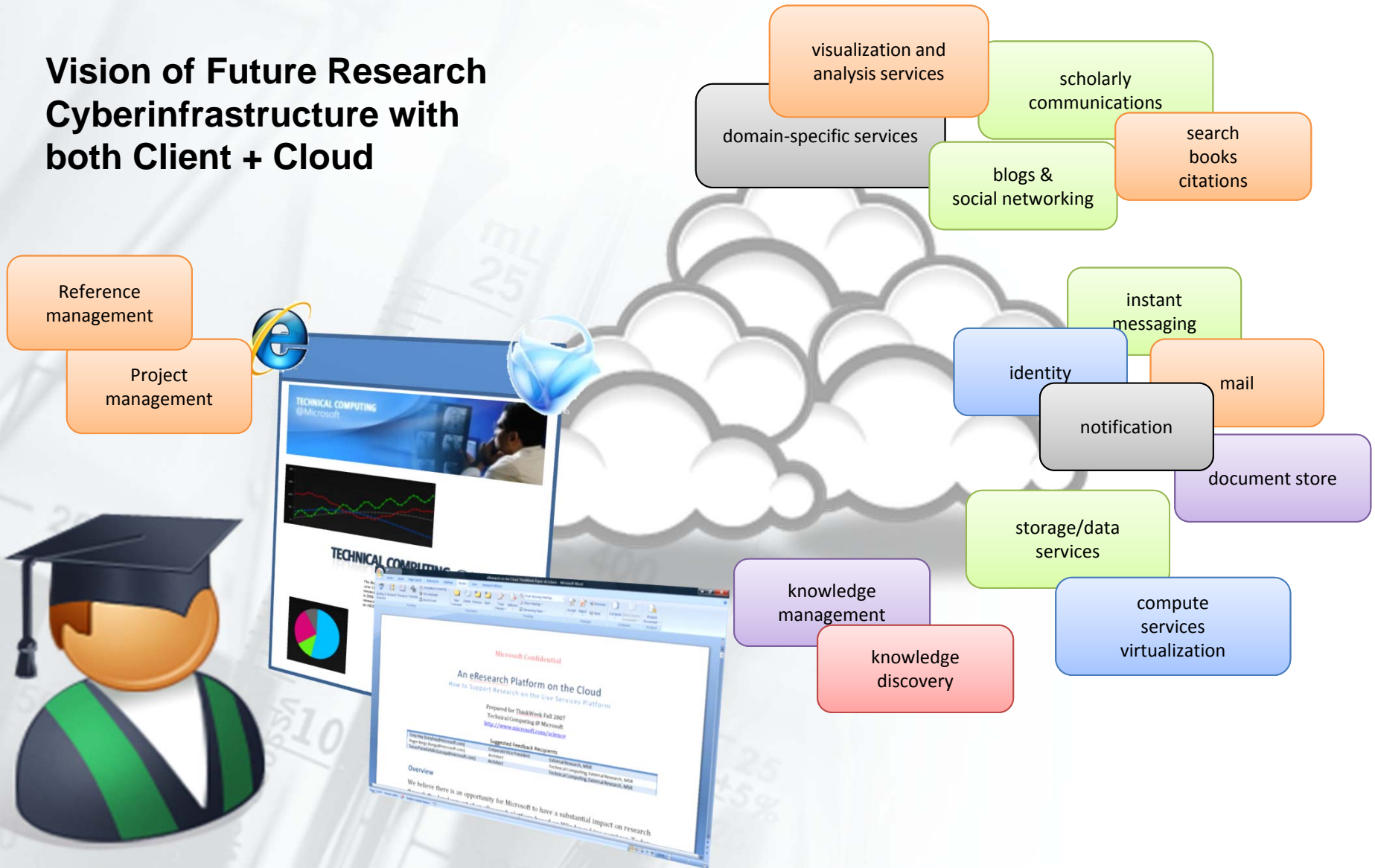


Attribution: [Richard Cyganiak](#)



...and stored/processed/analyzed in the Cloud

Vision of Future Research Cyberinfrastructure with both Client + Cloud



Acknowledgements

My thanks to Bill Dolan, David Heckerman, Eric Horvitz, Oscar Naim, Savas Parastatidis, and especially **Evelyne Viegas** for their help in preparing this talk.



Resources

- Microsoft Research
 - <http://research.microsoft.com>
 - Microsoft Research downloads:
<http://research.microsoft.com/research/downloads>
- Microsoft External Research
 - <http://research.microsoft.com/en-us/collaboration/>
- Science at Microsoft
 - <http://www.microsoft.com/science>
- Scholarly Communications
 - <http://www.microsoft.com/scholarlycomm>
- CodePlex
 - <http://www.codeplex.com>





Microsoft[®]

Your potential. Our passion.[™]

