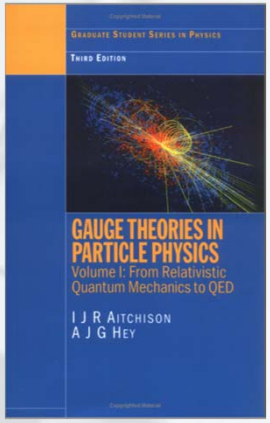
Semantic Computing and Smart Cyberinfrastructure

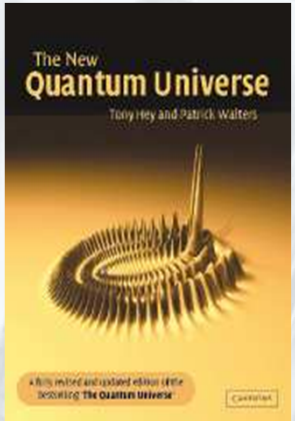
Tony Hey

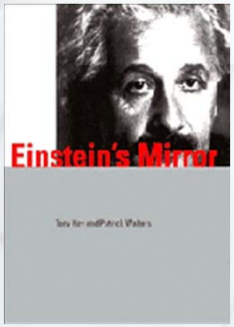
Corporate Vice President

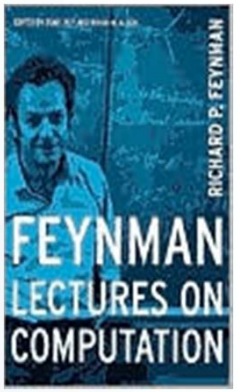
Microsoft Research

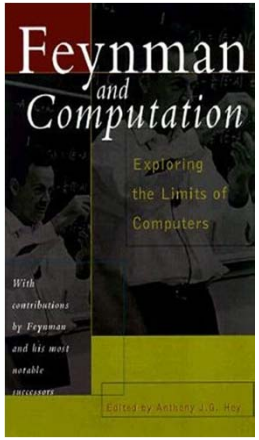
Tony Hey – My Background











UK National Centre for Text Mining

the national centre for text mining

NaCTeM
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

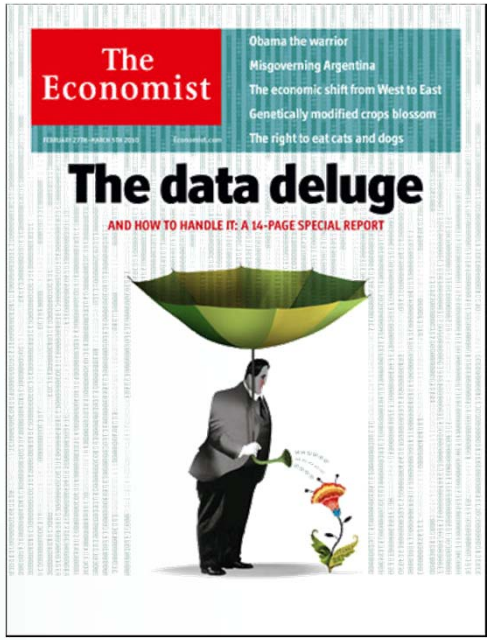
Featured News Feed

Other News & Events

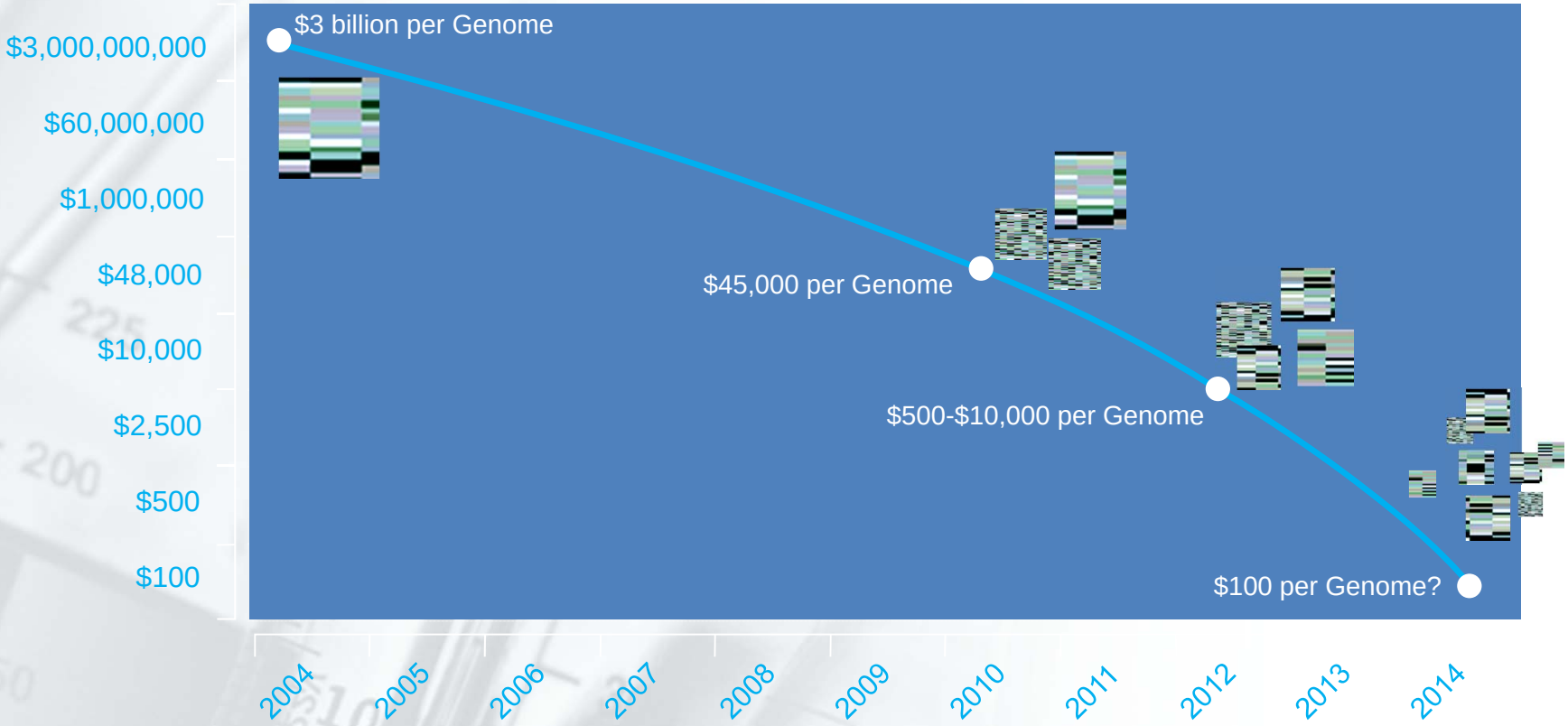
- Invited lecture at the Institute of Information Science of Academia Sinica

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, Fig. 2

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

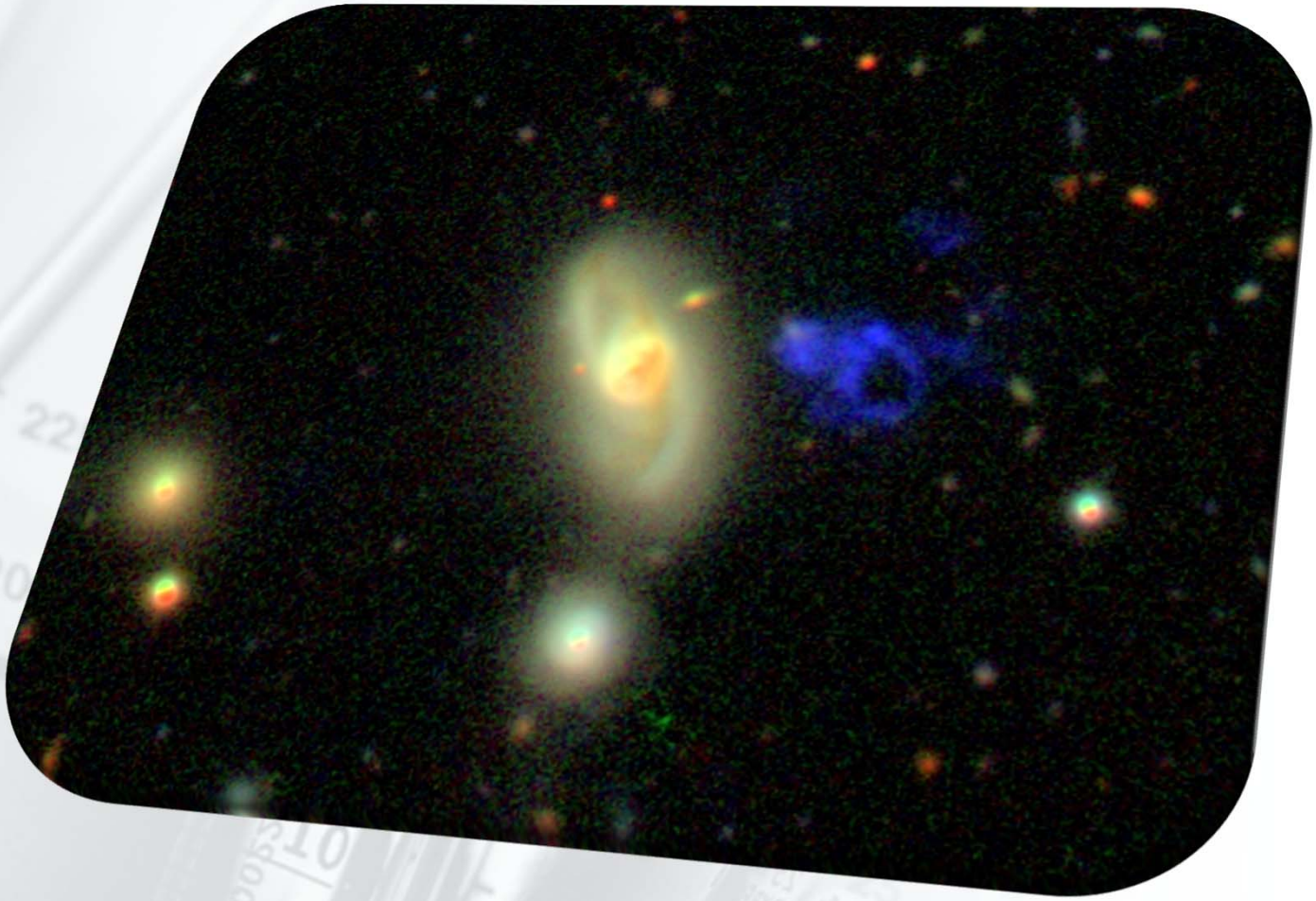
Sources: The Economist, Feb '10; IDC

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

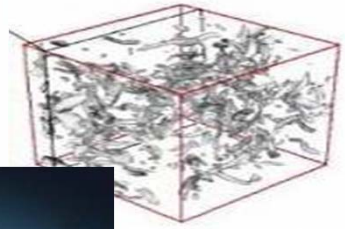


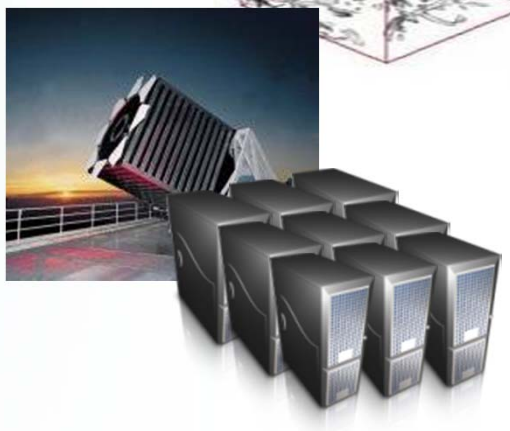
The Fourth Paradigm: Data-Intensive Science

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

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

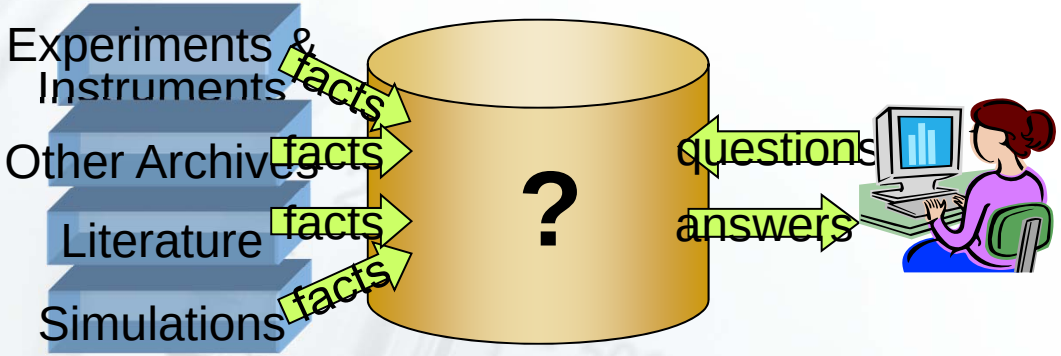




(With thanks to Jim Gray)

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 *re*organize 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

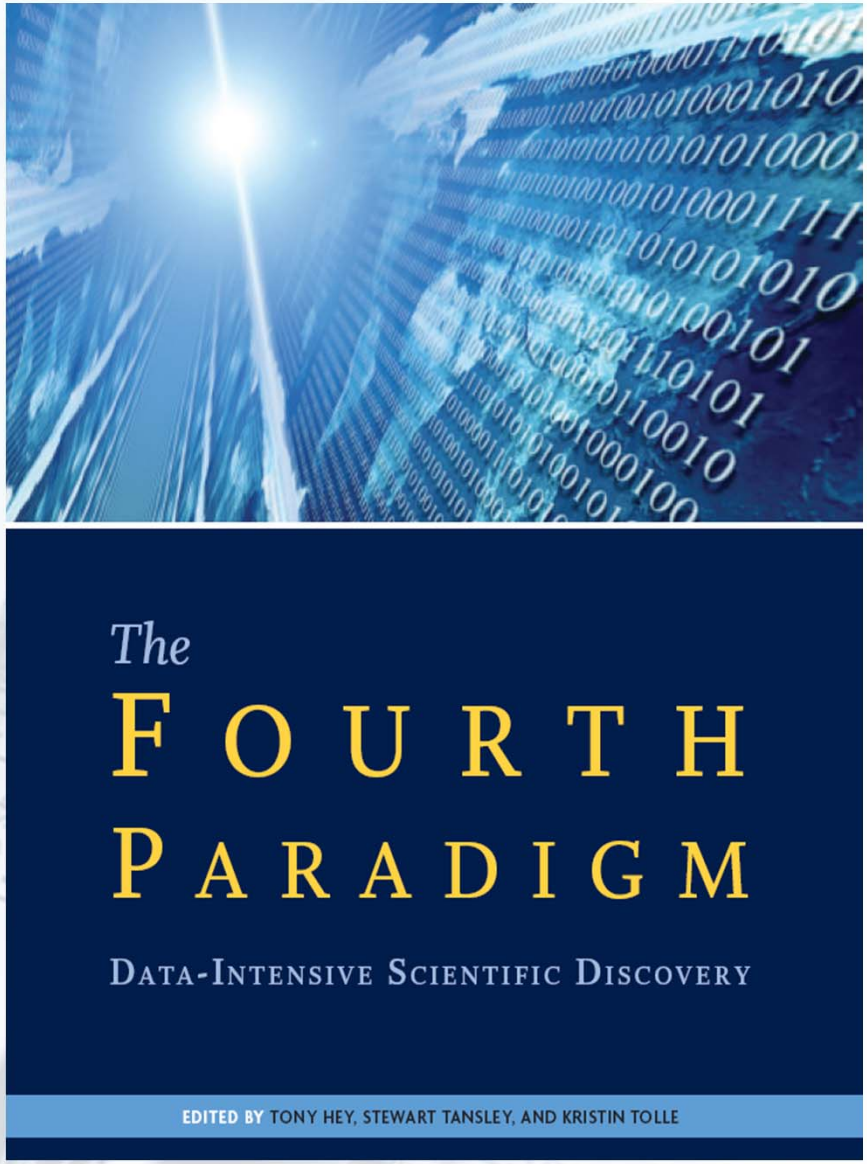
Seamless Rich Social Media Virtual Sky
Web application for science and education

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
FOURTH
PARADIGM
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

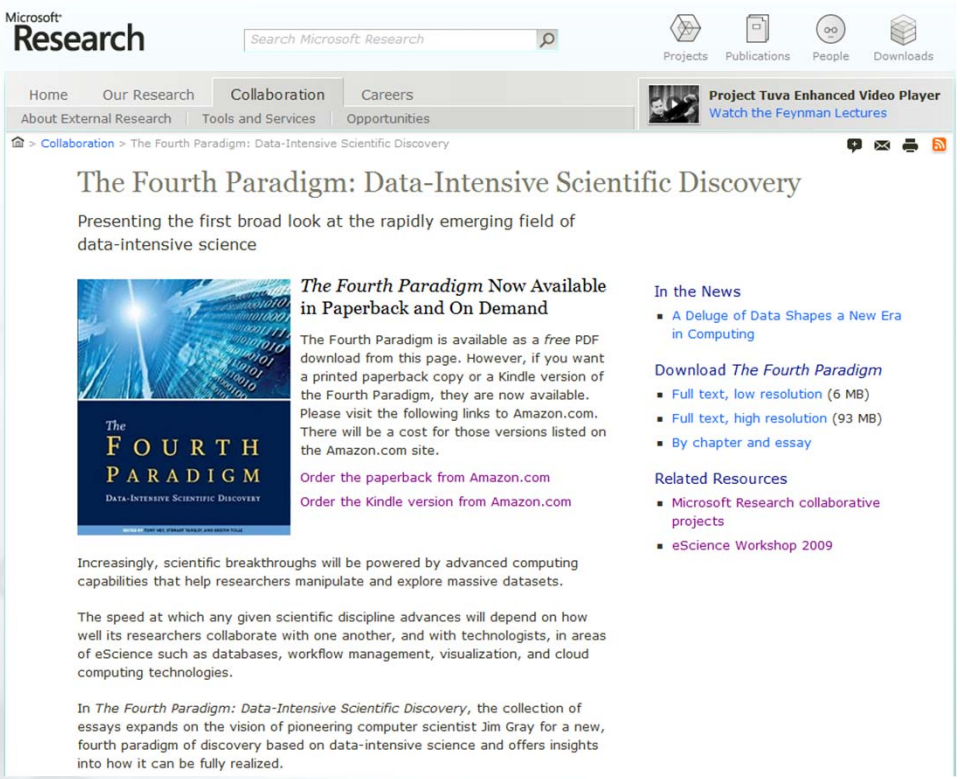
Free PDF Download
Amazon Kindle version; Paperback print on demand

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



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

bing Translator

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

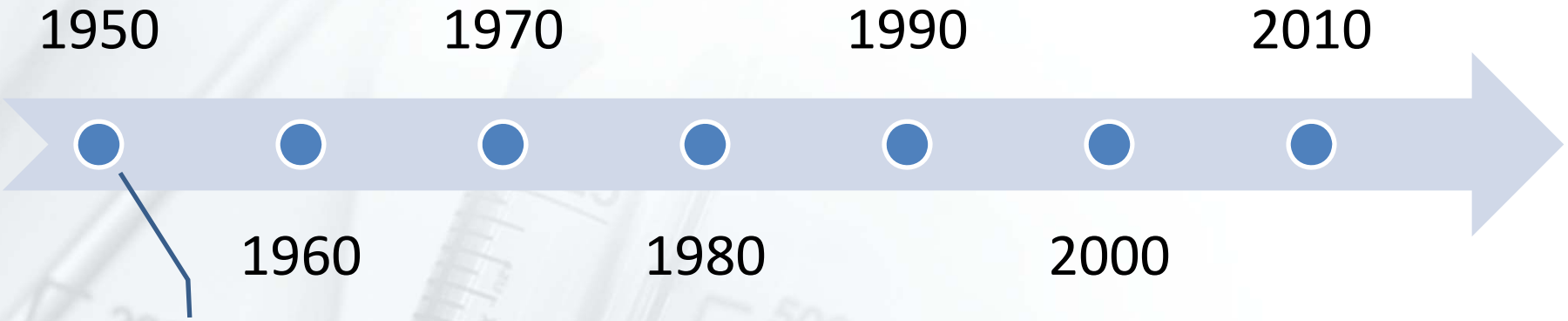
L'objectif principal de la conférence est de favoriser le dialogue entre les experts dans chaque sous discipline. Nous encourageons donc particulièrement soumissions de travail qui est intéressant à plusieurs domaines, tels que les approches multimodales.

Rate this translation: ☆☆☆☆☆ Copy Translation

New: Translator V2 API & Widget announced at MIX 2010

Powered by Microsoft® Translator

A timeline



1954: 60 Russian sentences successfully translated into English

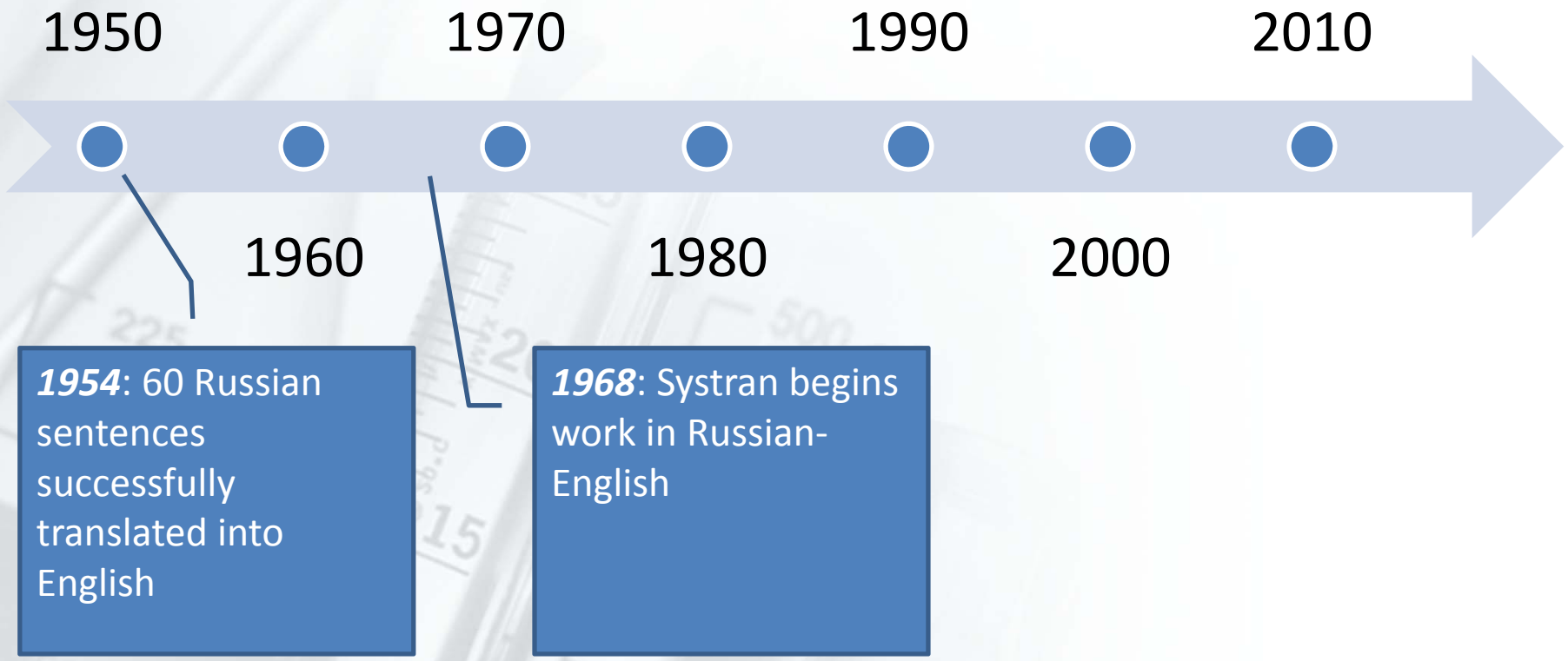
A timeline

1950

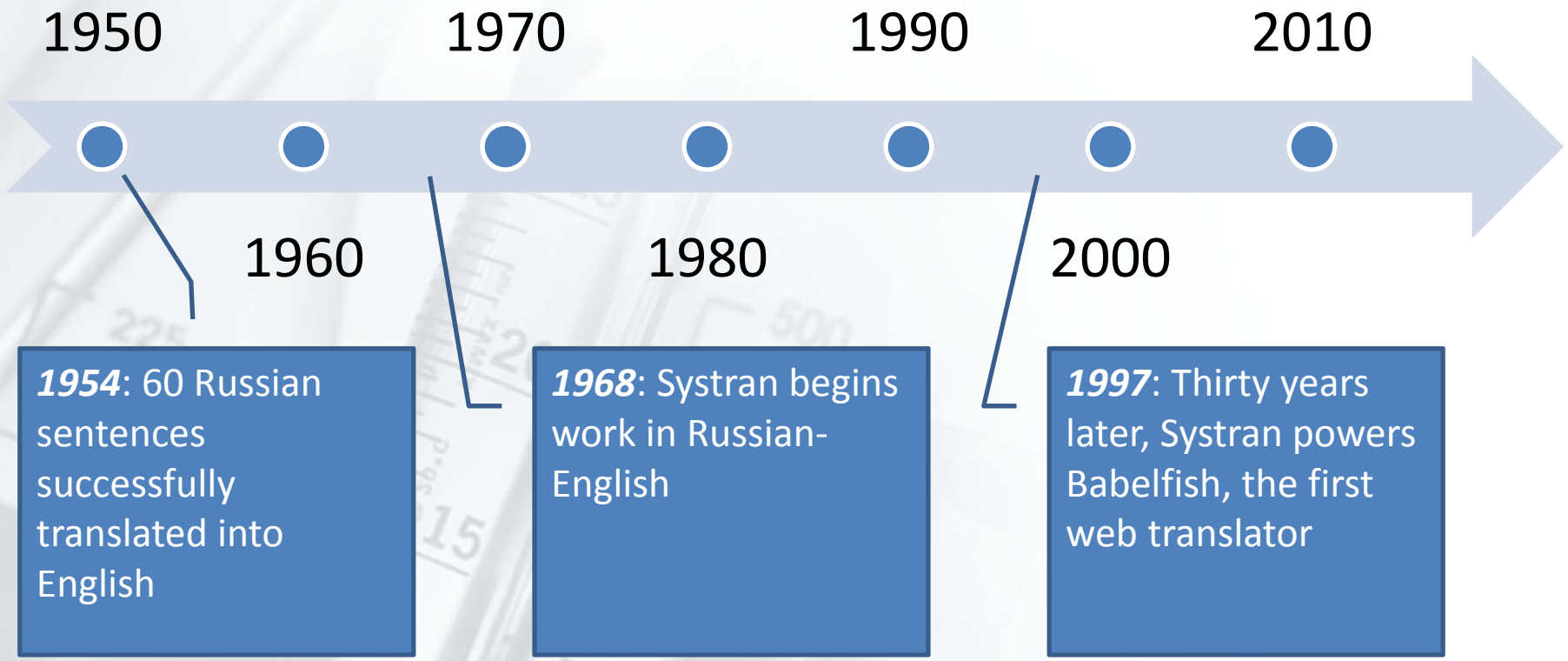
1960

1954: 60 Russian sentences successfully translated into English

A timeline


1950
1960
1970
1980
1990
2000
2010
1954: 60 Russian sentences successfully translated into English
1968: Systran begins work in Russian-English

A timeline


1950
1960
1970
1980
1990
2000
2010
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

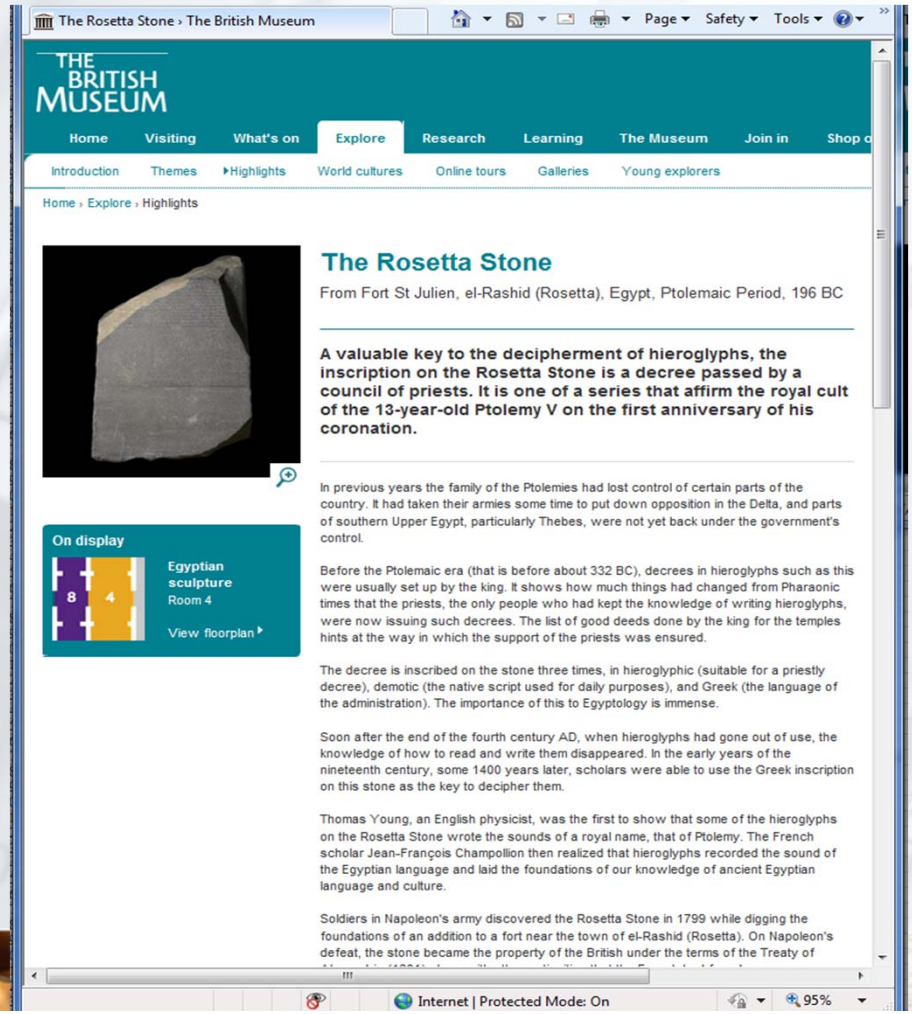
Broad-Domain, High-Quality MT?

- 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

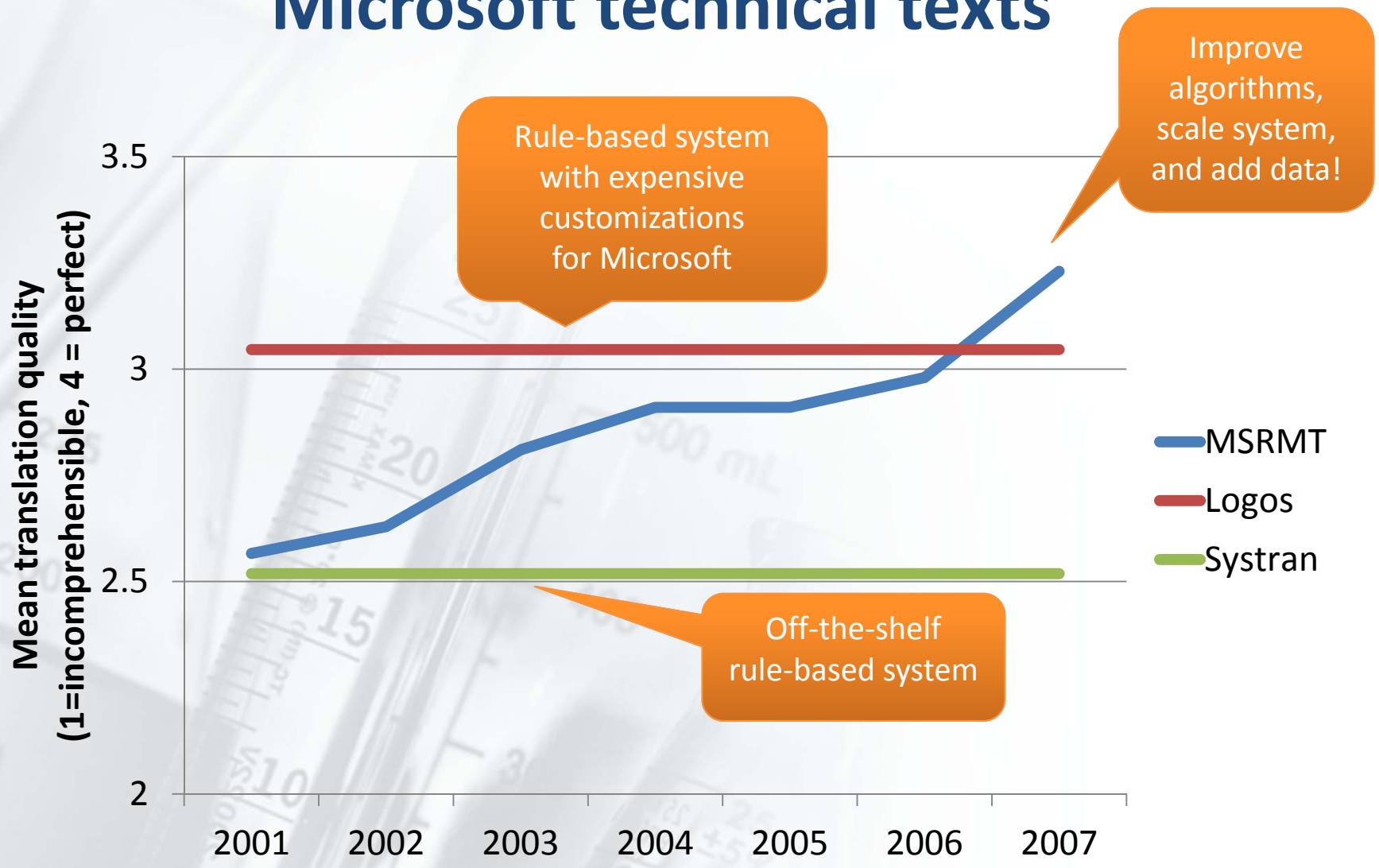
The Statistical Revolution

Instead of hand-coding rules

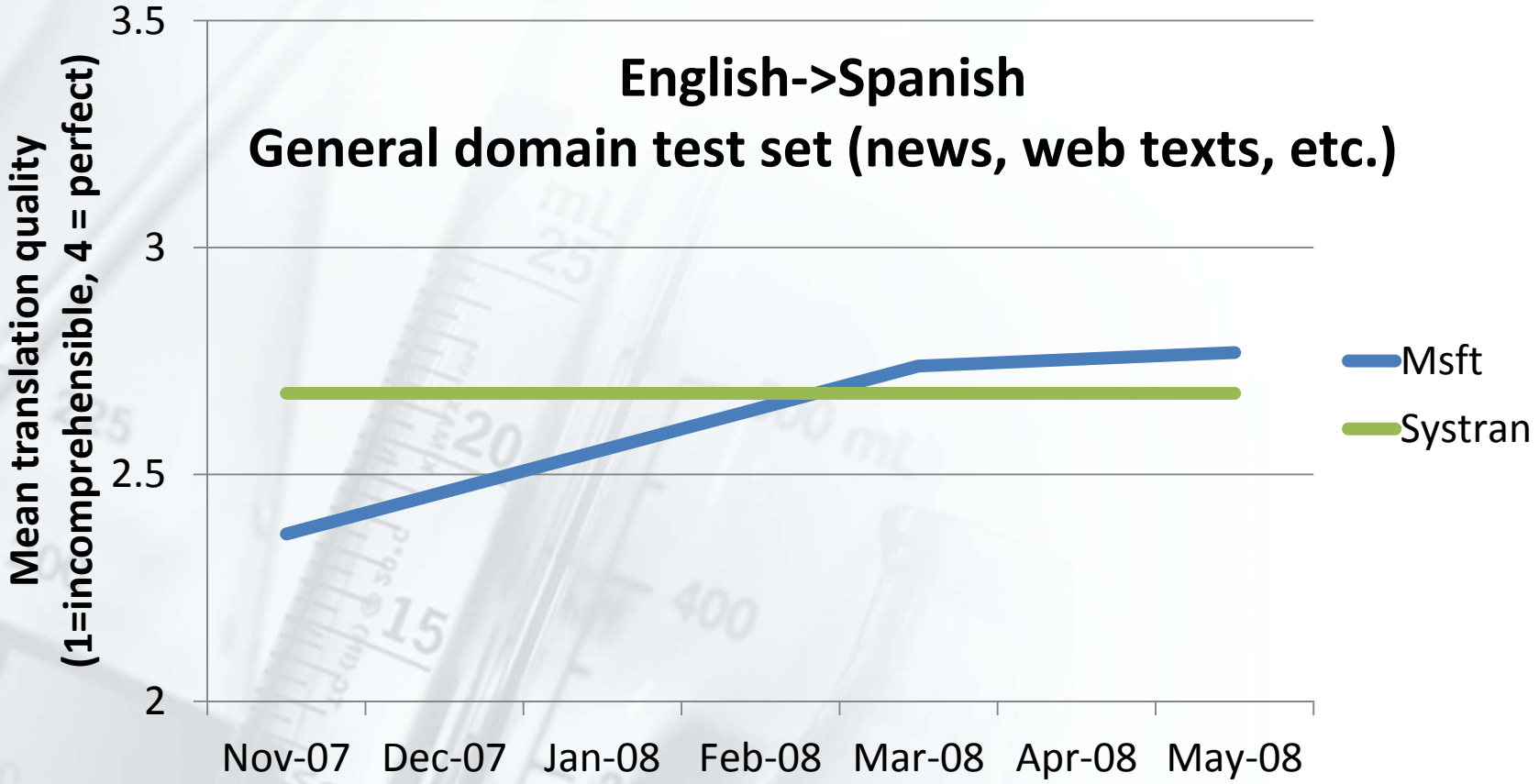
- Exploit large volumes of existing parallel text
- Learn how words, phrases, and structures translate in context



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 janvye 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

Moun ap fouye pami debri yon bilding ki kraze nan tranblemann' tè 12 Janvye 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

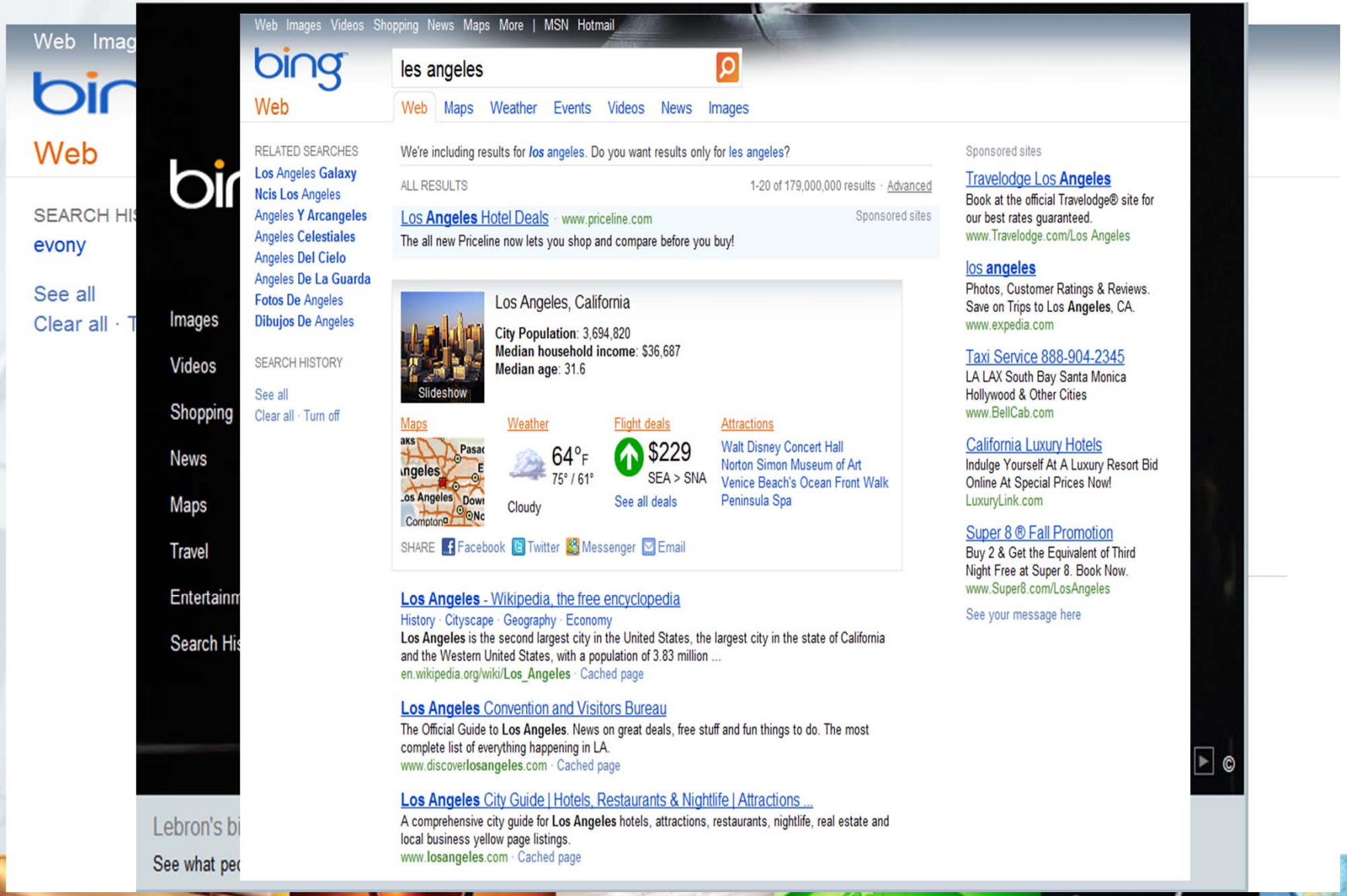
Message Translation

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 lekol 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



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

Web Images Videos Shopping News Maps More | MSN Hotmail

bing MS Beta 10553

book of eli

Web

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

Sponsored sites

Book Of Eli at Amazon
Low Prices on Book of eli
Free 2-Day Shipping w/ Amazon Prime
www.Amazon.com/Books

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	**Buy**: Amazon
46 out of 100	Rotten Tomatoes (172 reviews)	Gary Oldman	**Rent**: Netflix
		Mila Kunis	**Rent**: Blockbuster
		Ray Stevenson	

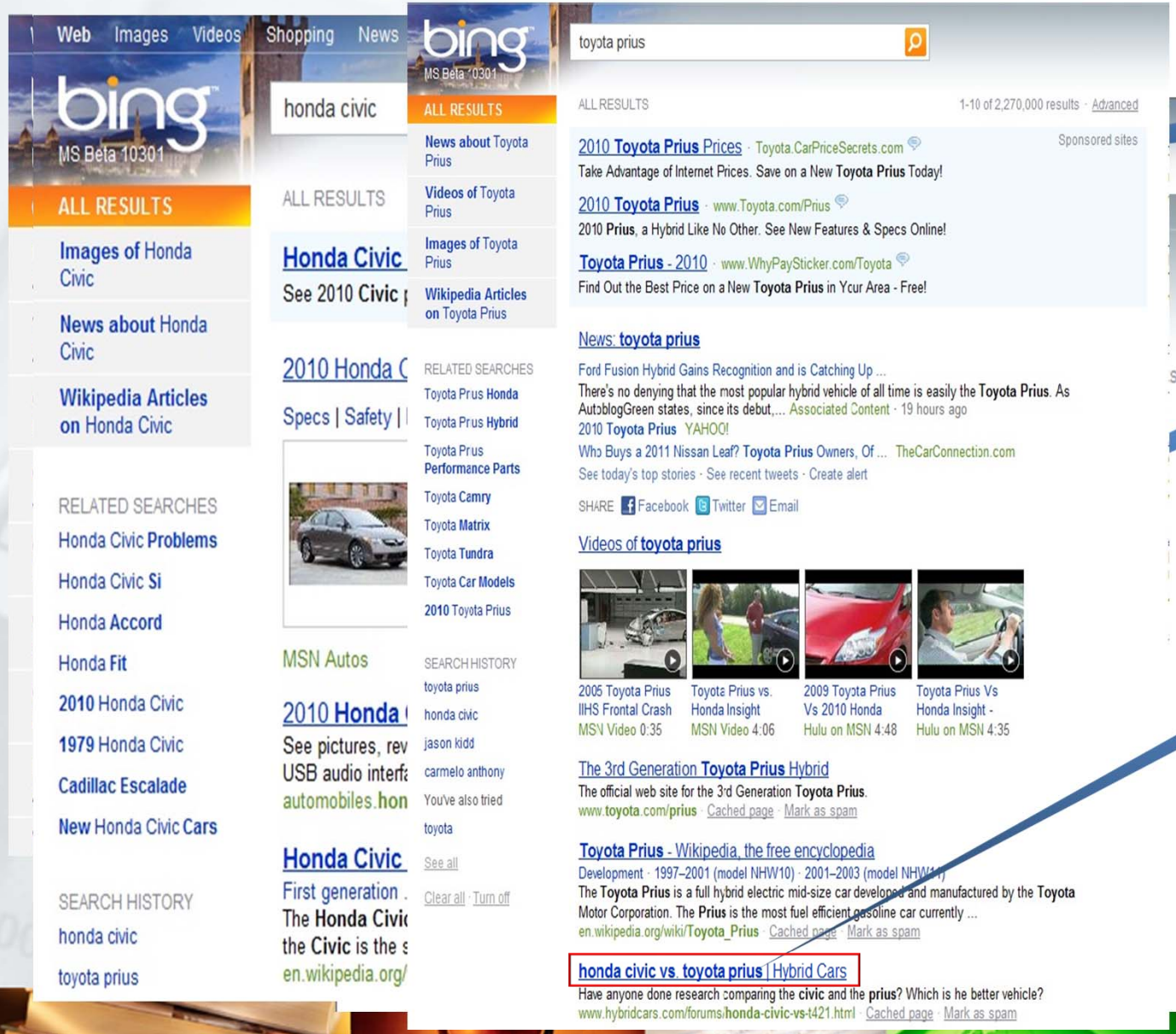
The Book of Eli - Wikipedia, the free encyclopedia
Plot · Production · Reception · Home media
The **Book of Eli** is a 2010 American post-apocalyptic action film directed by the Hughes brothers, written by Gary Whitta, and starring Denzel Washington, Gary Oldman, Jennifer Beals ...
en.wikipedia.org/wiki/The_Book_of_Eli · 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

Quick Tabs for relevant Bing verticals, domains, answers

Entity-based result summary

Spatial-Temporal Dialog: Re-Rank with Session

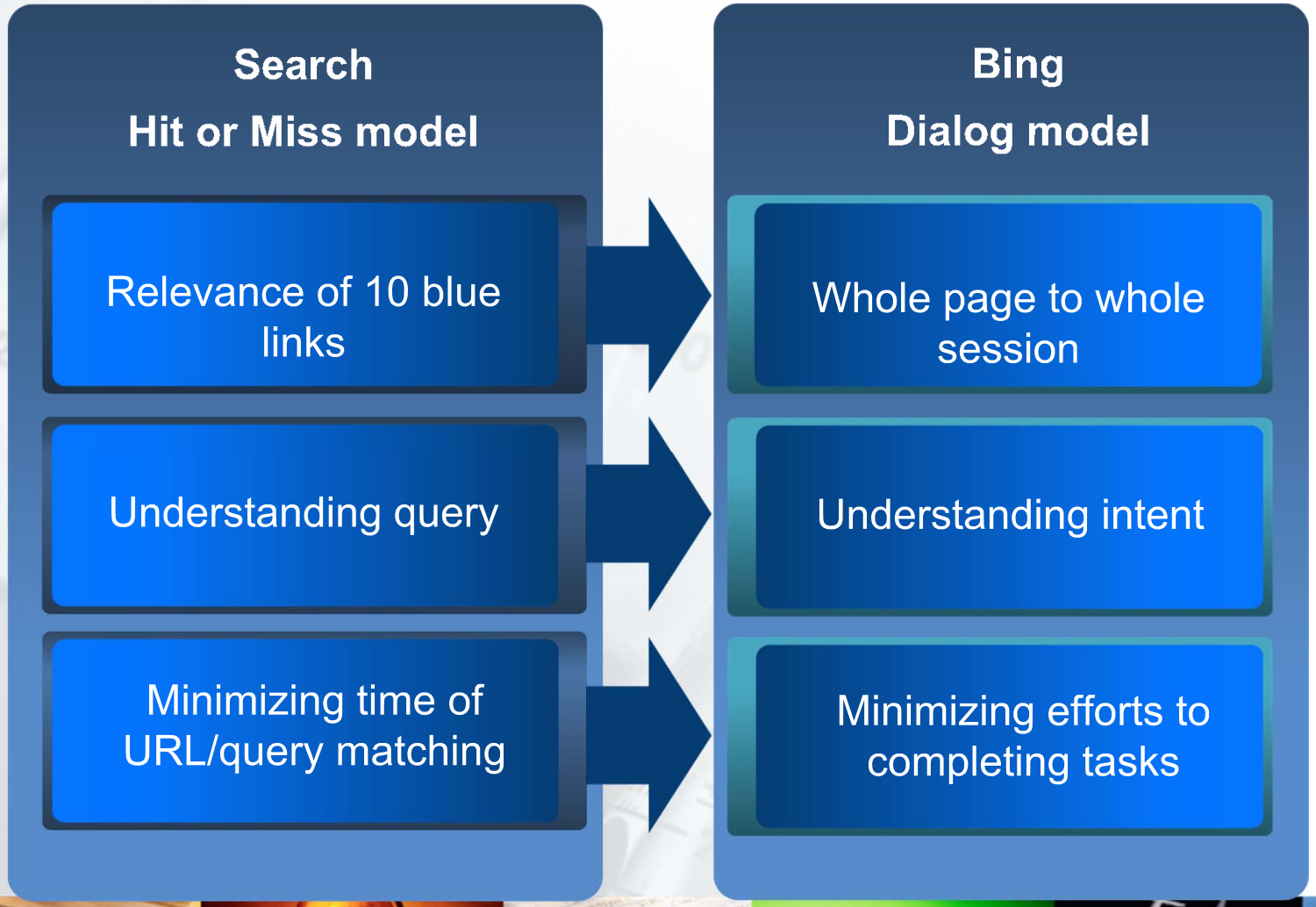


History-aware Autosuggest

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?

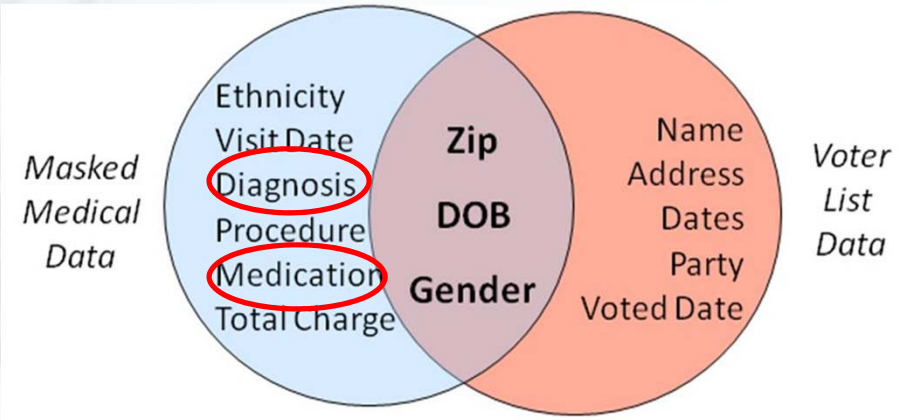
- Much of the data needed for data-driven research is locked away in industry vaults
  - Reasons: privacy, scale, business sensitivity

- 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]

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

http://www.nytimes.com/2006/08/09/technology/09aol.html

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

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS

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,"

http://www.nytimes.com/2006/08/09/technology/09aol.html

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

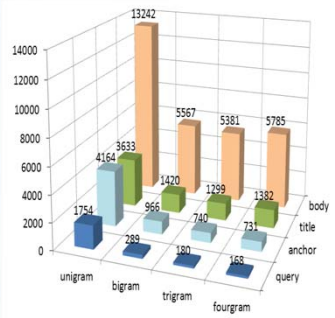
Web N-gram Services

Access to *petabytes* of real world data

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

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

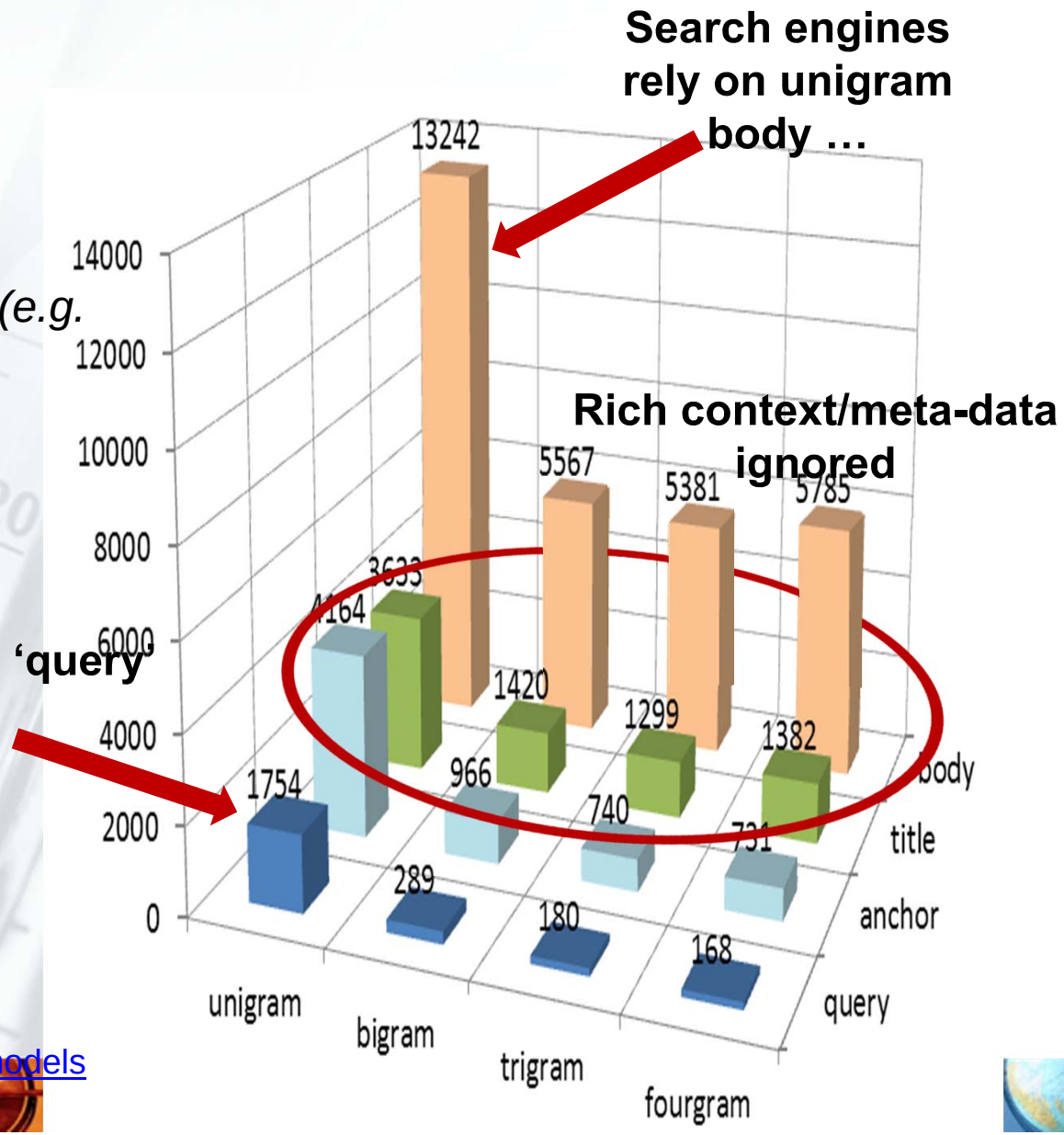
Web data has structure – and that difference counts



Web N-Gram in Public Beta

Web data has structure…

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



[Exploring Web Scale Language models for Search Query Processing](), in

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 #riptherunway #complimentgonebad #SMHyoureghetto #letmefindout #idoit2 #itaintmyfault #FlavoredCondoms #jayparkaom #ChrisBrownRocks #thingthatihate #nowplaying #whenifirstmet #hereugo #stpatricksday #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

#yeaisaidit Go

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

#whenifirstmet Go

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

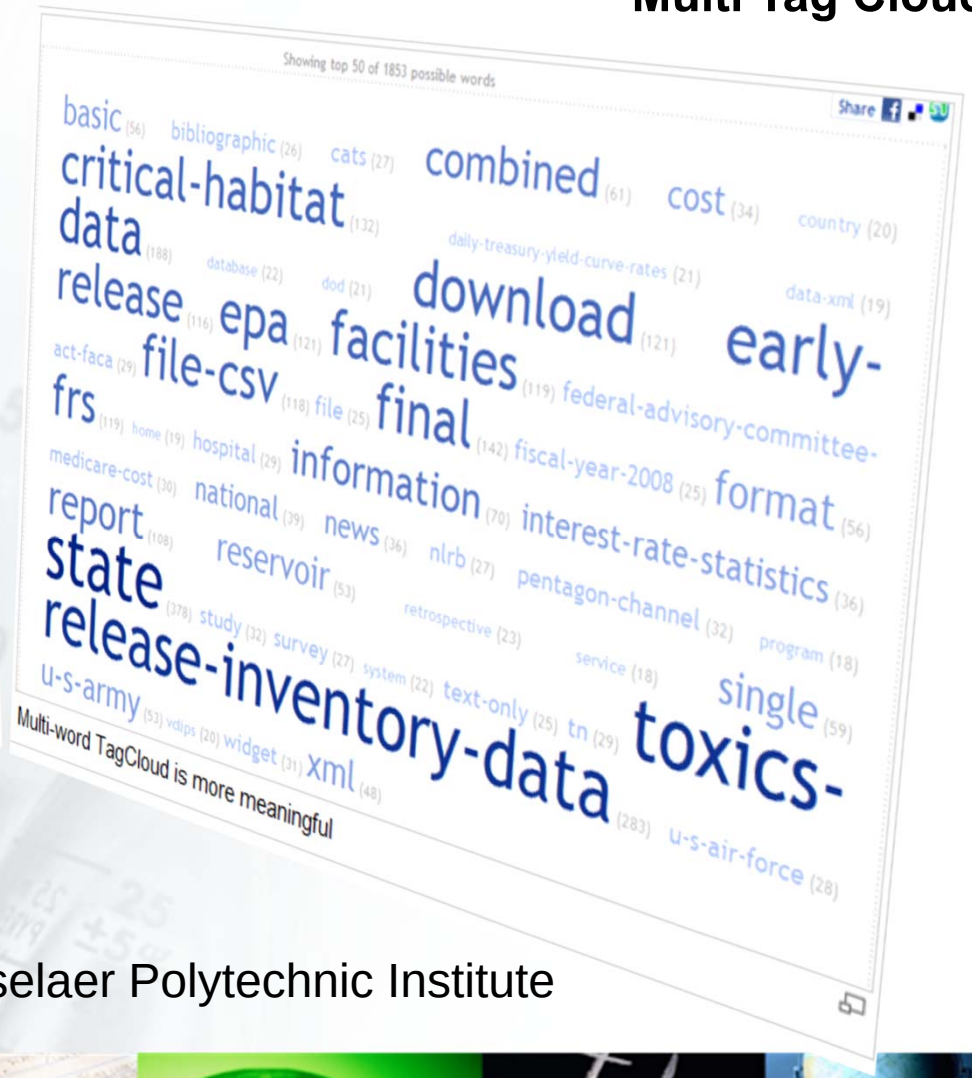
#w84u

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

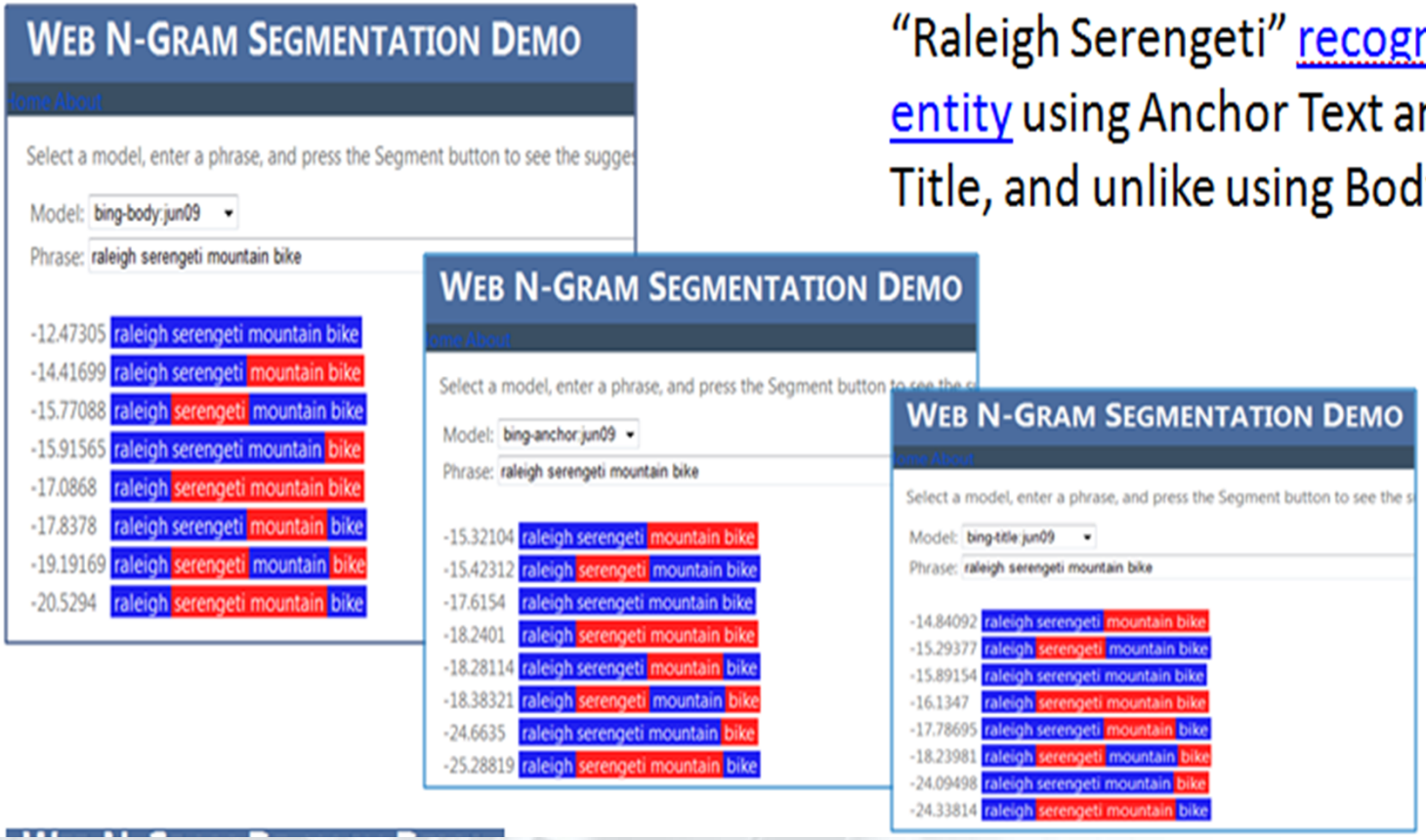
Multi Tag Cloud



Ref: Dr. Li Ding, Rensselaer Polytechnic Institute

INFORMATION

Language Models



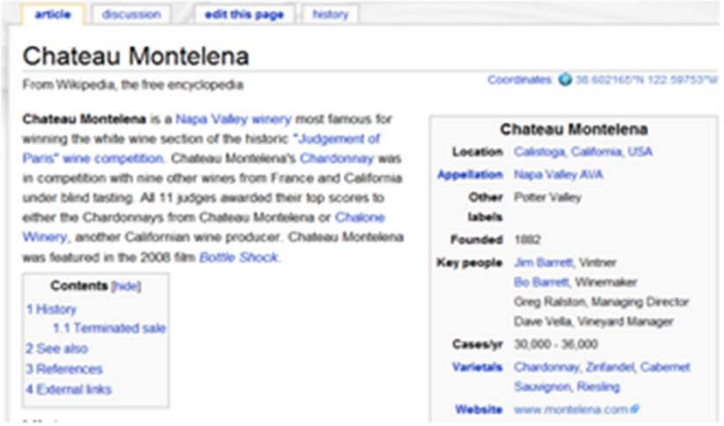
"Raleigh Serengeti" recognised as an entity using Anchor Text and Document Title, and unlike using Body

Semantic Computing to enable

Implicit Search

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

'Chateau Montelena in Napa' segmentation



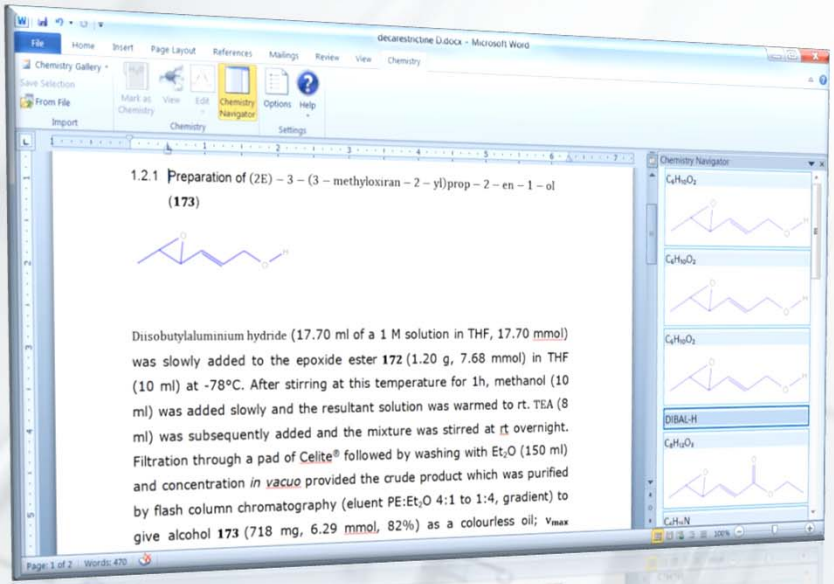
'Chateau Montelena' as an *entity* in Wikipedia

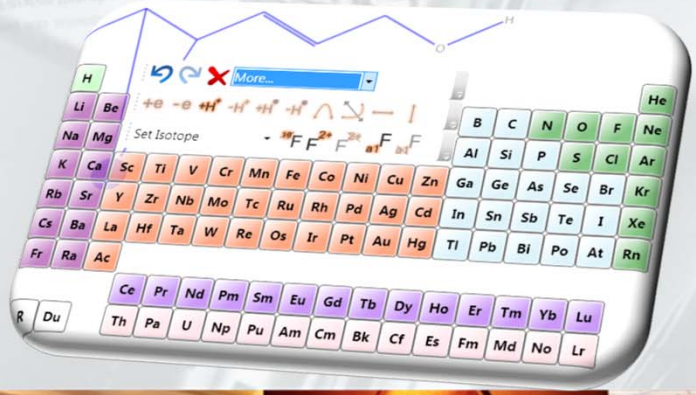
Semantic Computing and Science Applications

Science Examples

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

Chemistry Add-in for Word





- 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

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

science commons

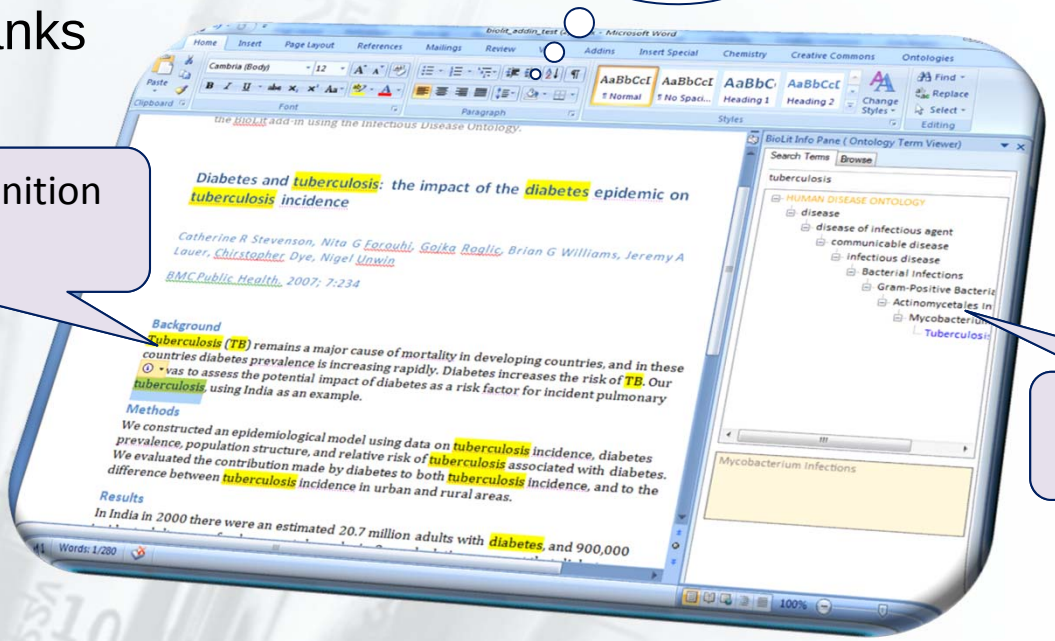
- John Wilbanks

Services: Ontology download web service

University of California San Diego

- Phil Bourne
- Lynn Fink

Intent: Term recognition & disambiguation



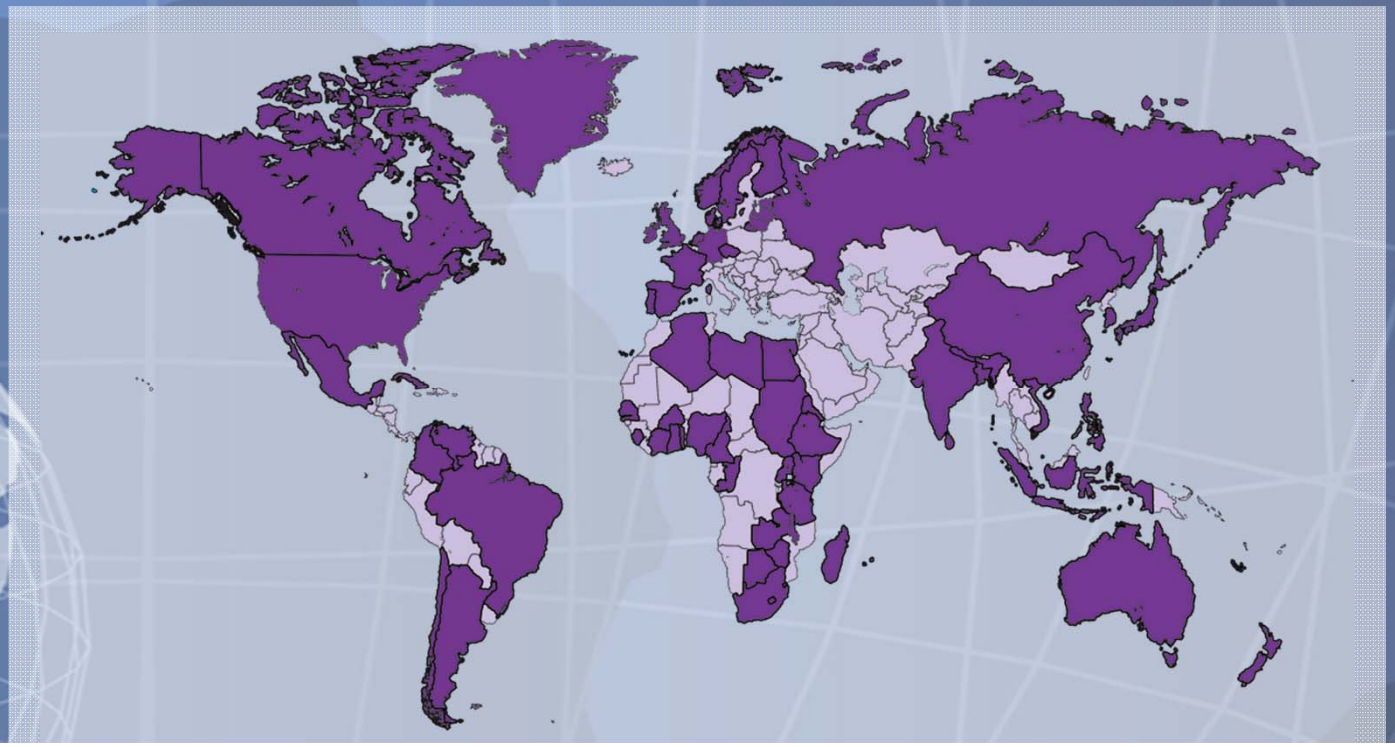
Relationships: Ontology browser

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*

WORLDWIDE SCIENCE*Alliance*

by Deep Web TECHNOLOGIES

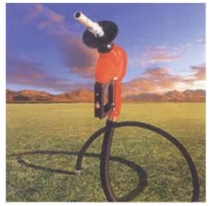
Translations powered by Microsoft Translator

Enabling Science and Innovation
ICSTI
International Council for Scientific and Technical Information

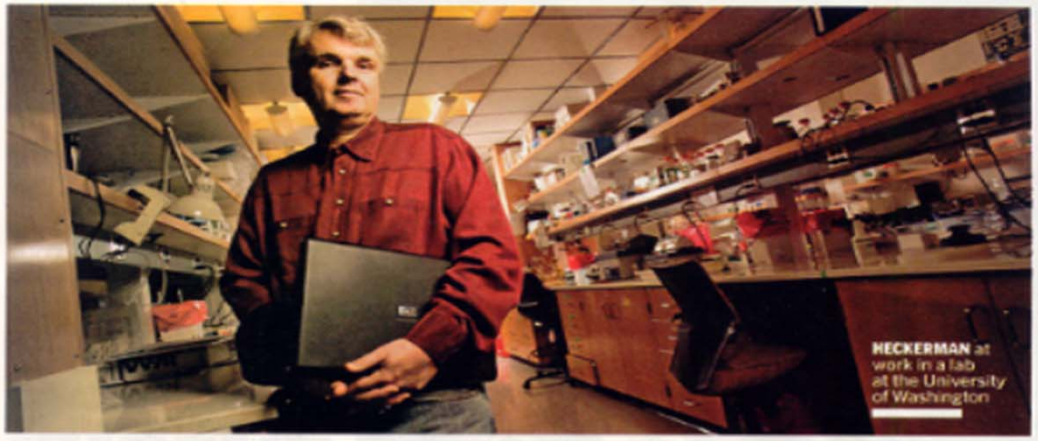
Tackling societal challenges with ML

- 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.

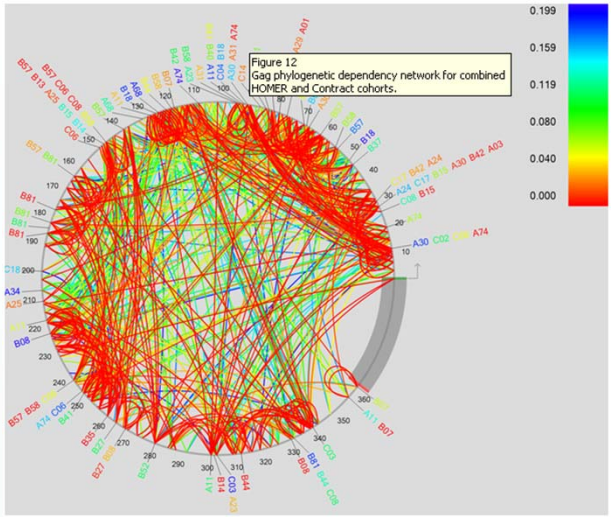
Similar mutations may crop up in computer and medical viruses

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 wilier 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

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



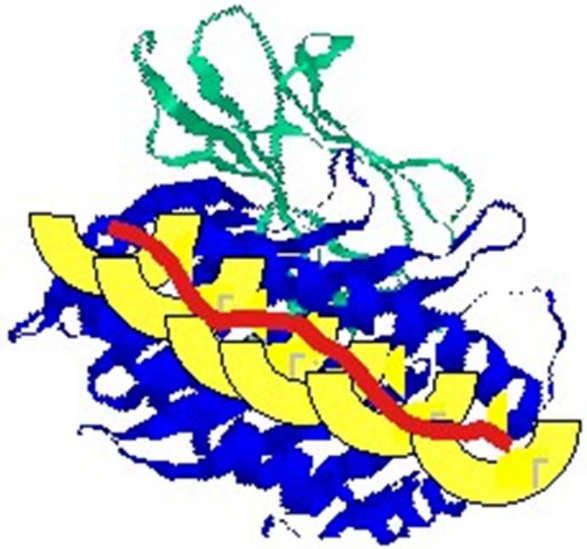
PhyloD.Net on cover of *PLoS Comp Bio*, Nov 2008 Carlson, Kadie, Heckerman et al.

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

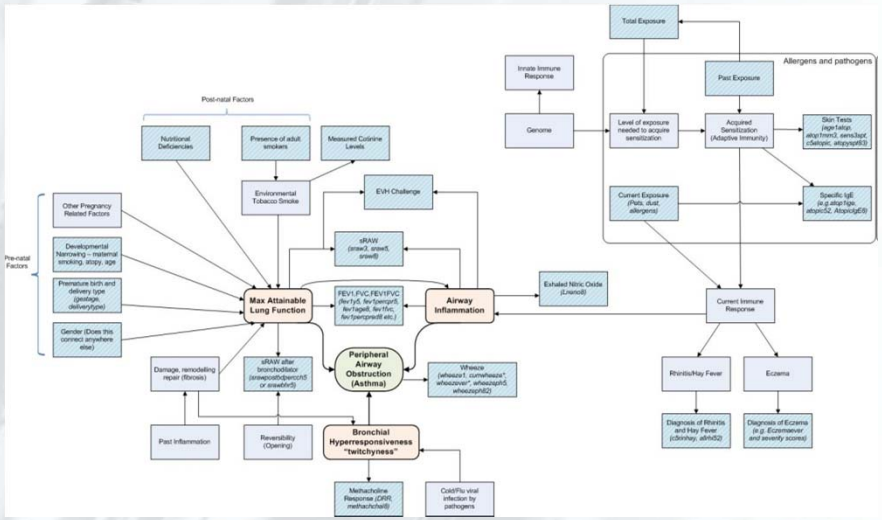
Collaborations with University of Manchester (MSR Cambridge and eScience group) and with Sanger (MSR Cambridge)

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 our 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

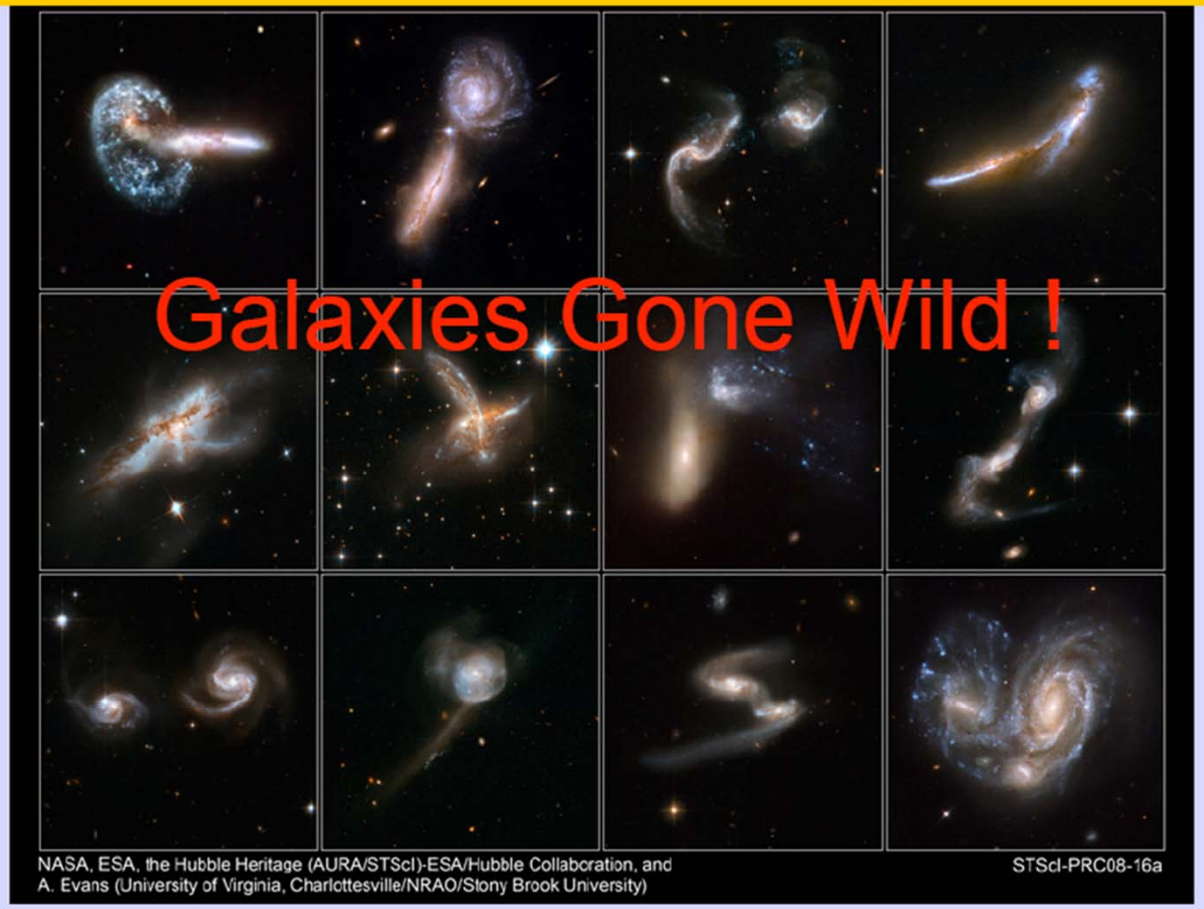
http://zooniverse.org/

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

ZOONIVERSE
REAL SCIENCE ONLINE

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



Semantic Computing and Healthcare Applications

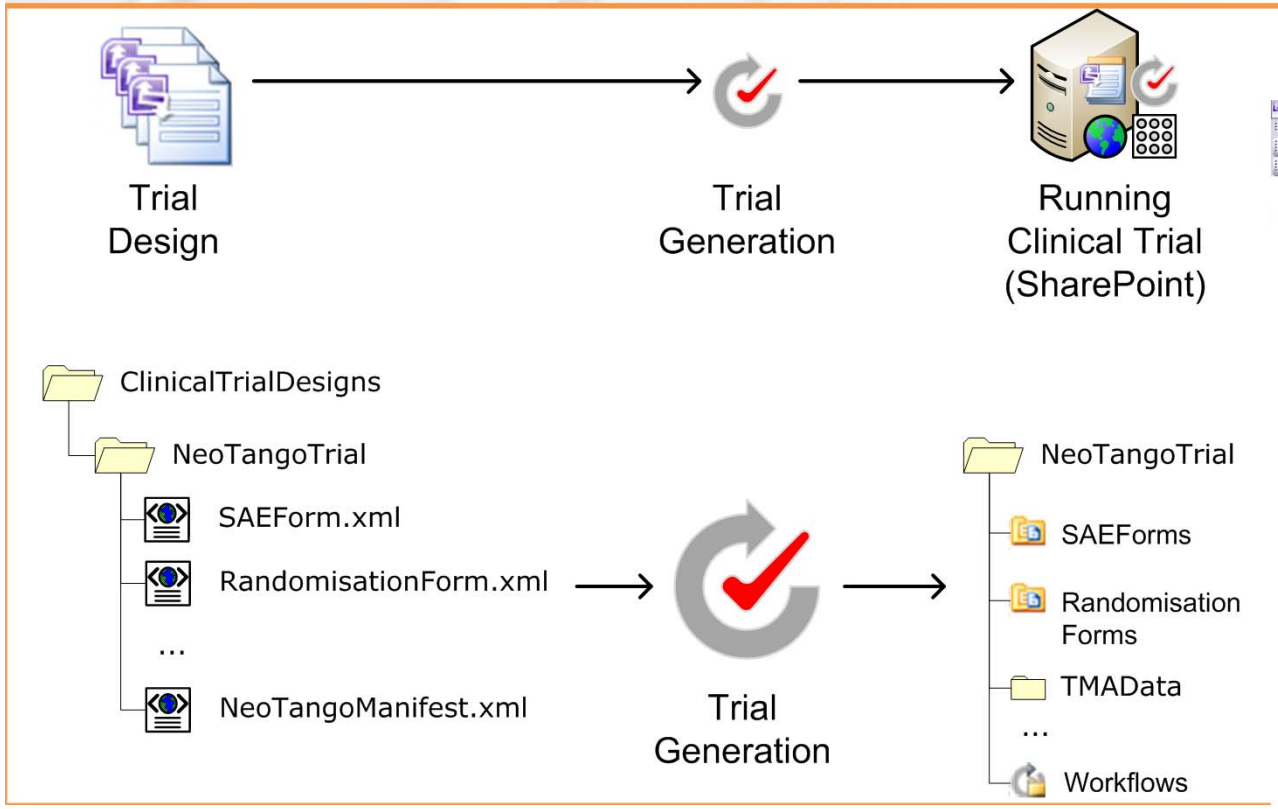
Healthcare Examples

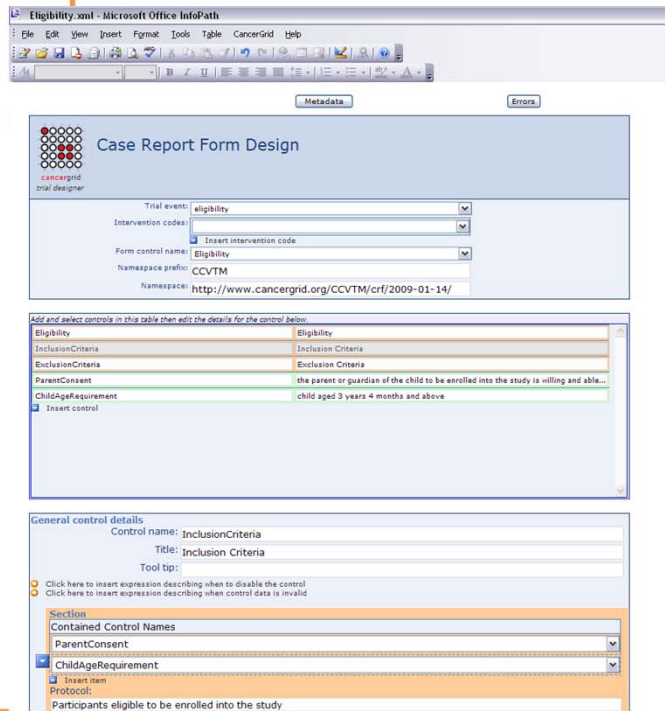
- Clinical Trials software project
- ML and Hospital Readmissions
- Amalga Life Science platform

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 Villella 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

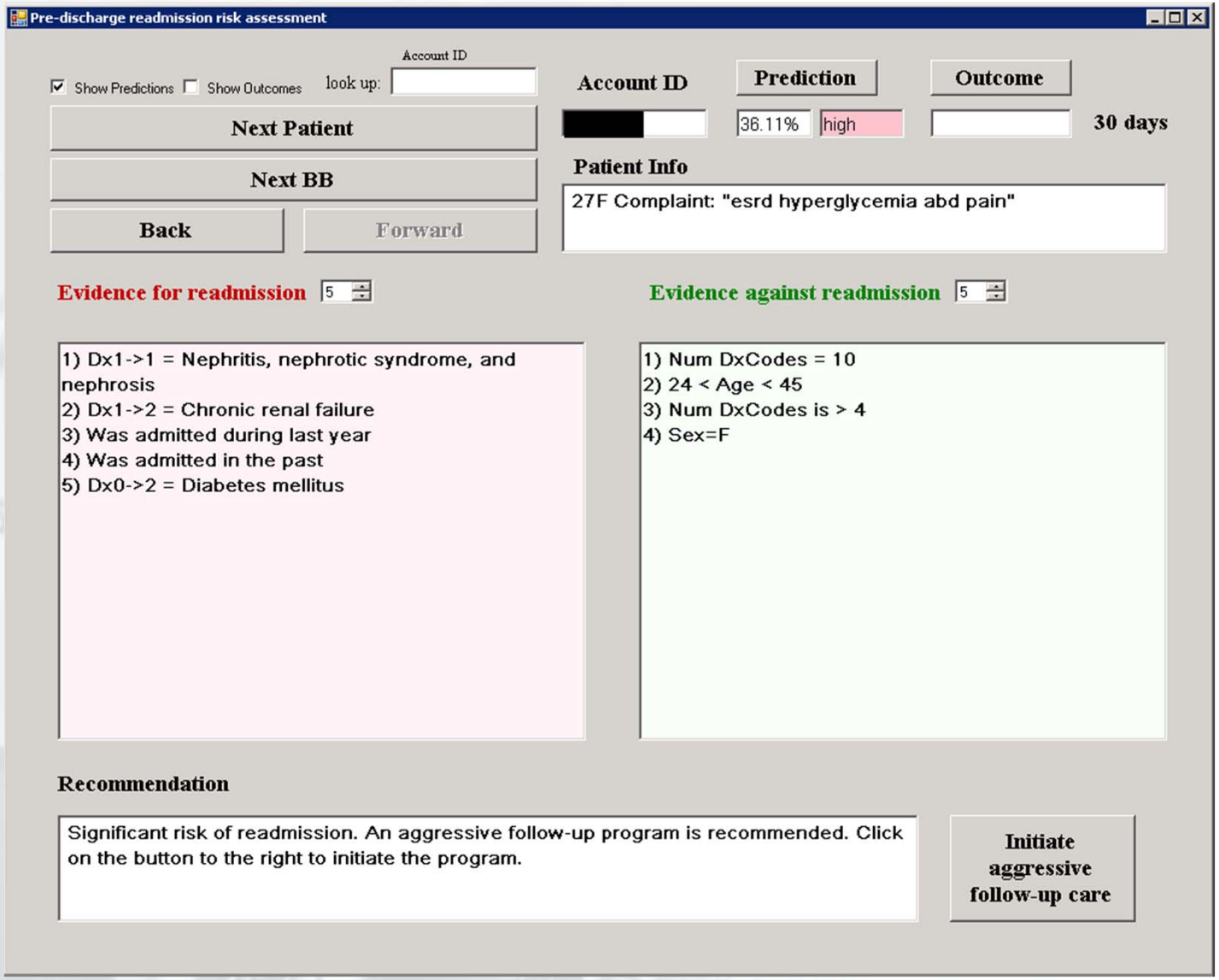
Learning from a Rich Case Library

- 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.
  - …

Microsoft Amalga Unified Intelligence System

Bayati, Braverman, and Horvitz

Towards Fielding an Advisory Tool



Bayati, Braverman, and Horvitz

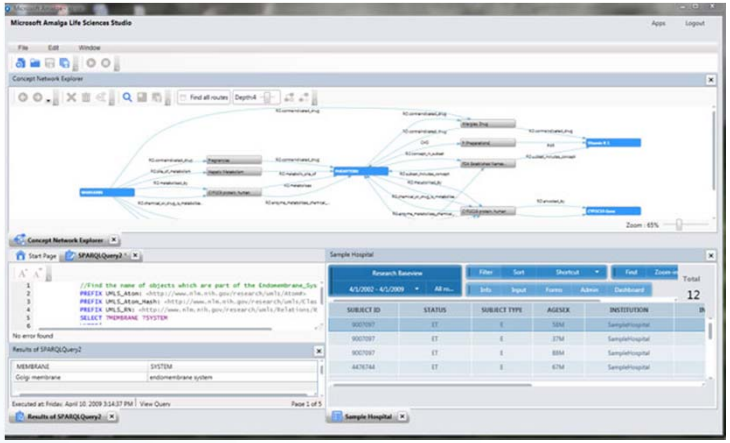
Amalga Life Sciences

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.

- 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

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

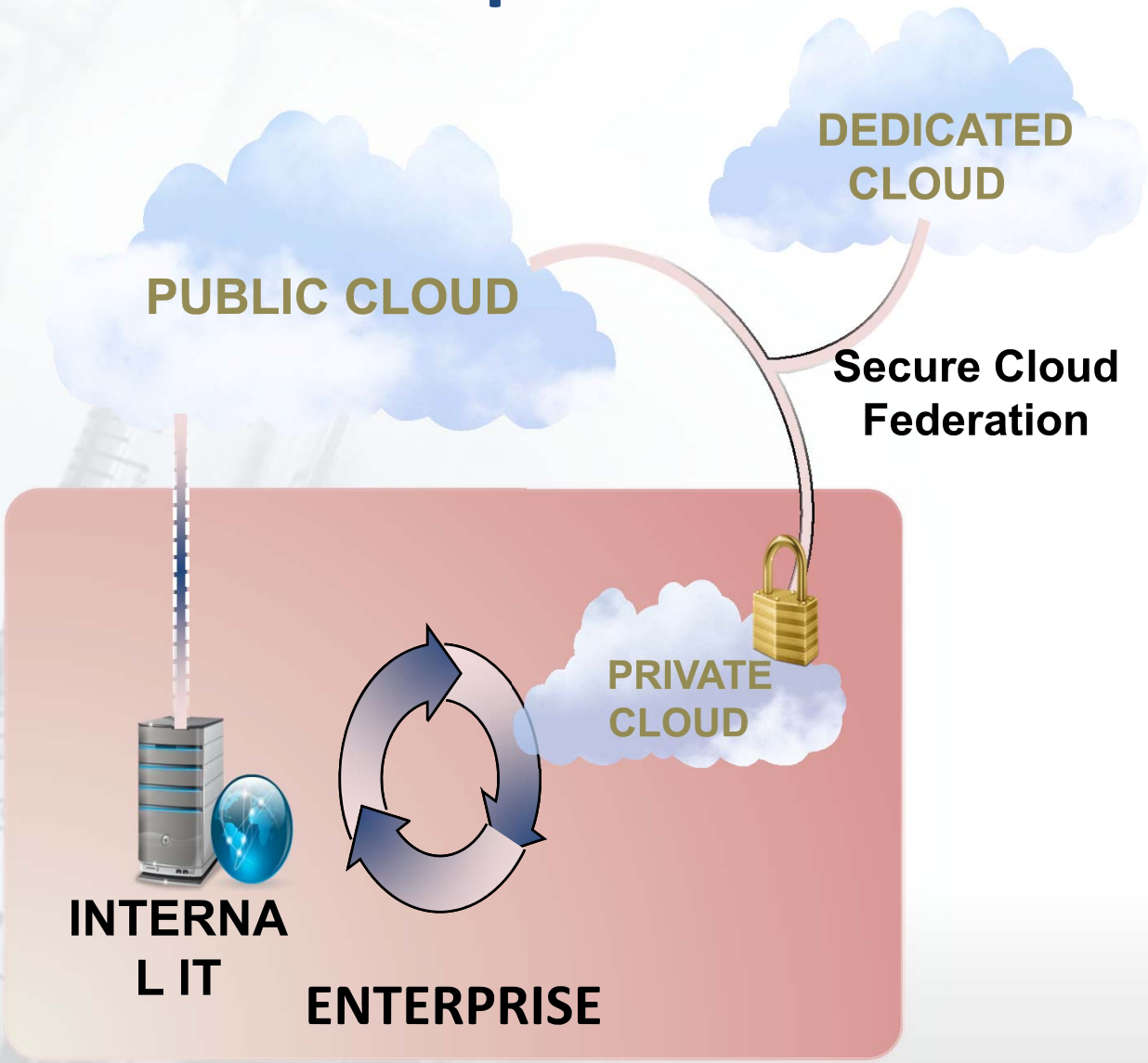
Towards a Smart Cyberinfrastructure

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

storing	computing
managing	indexing

huge amounts of **data**

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

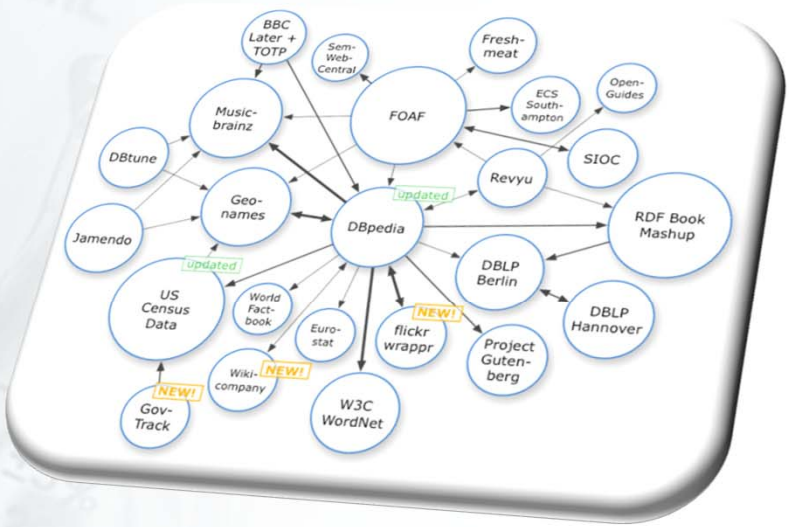
acquisition	discovery
aggregation	organization
correlation	analysis
interpretation	inference

of the world's **information**

A world where all data is linked…

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

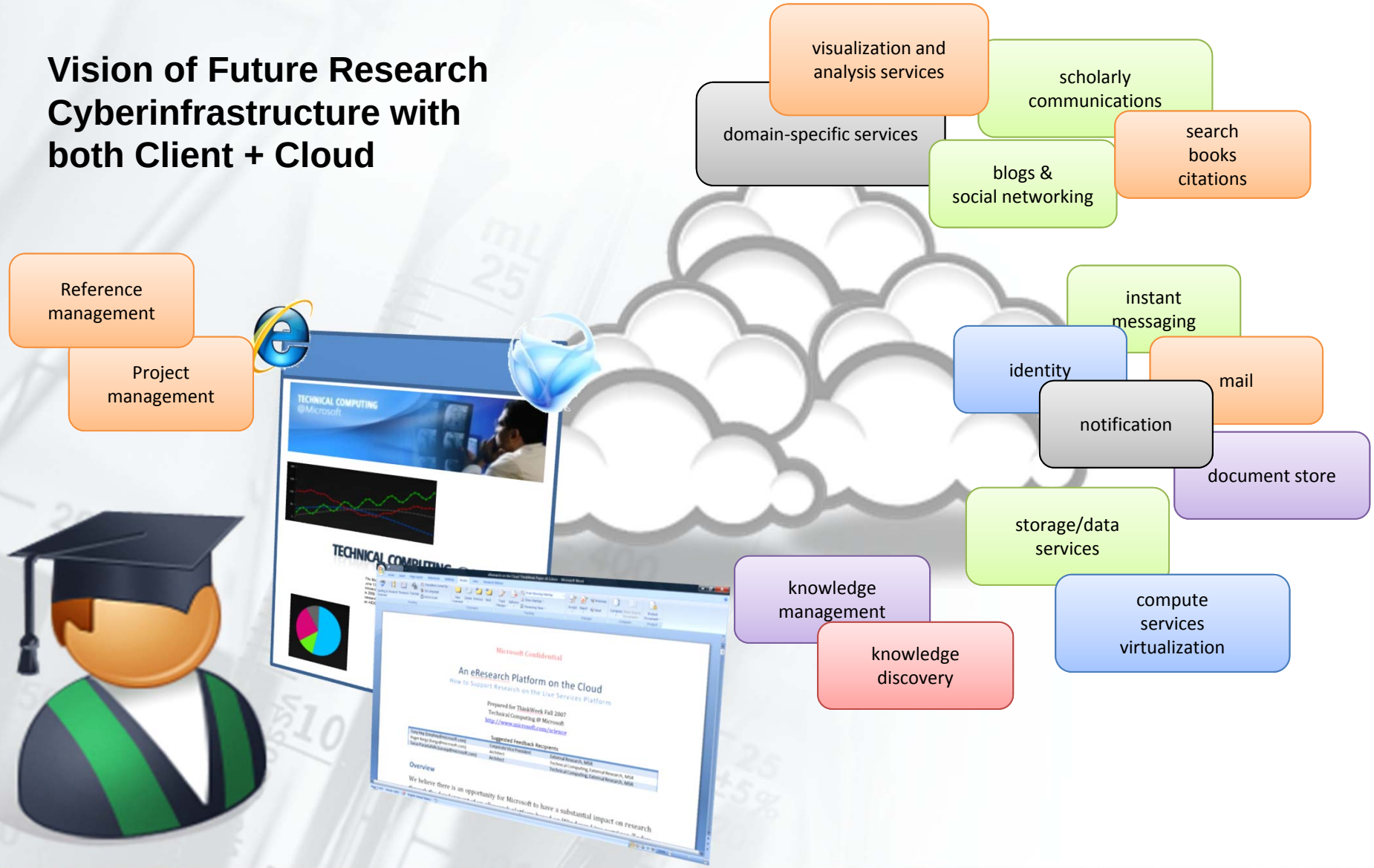
- **Important/key considerations**
  - Formats or "well-known" representations of data/information
  - Pervasive access protocols are key (e.g. HTTP)
  - Data/information is uniquely identified (e.g. URIs)
  - Links/associations between data/information



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.™