Microsoft® Research

FacultySummit 2011

Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Microsoft® Research

FacultySummit 2011

Cartagena, Colombia | May 18-20 | In partnership with COLCIENCIAS

Semantic Computing for eScience

Evelyne Viegas
Director Semantic Computing

Outline

- My Background
  - Academia
  - Industry

- Data to drive innovation
  - Next Generation Innovations
  - From Data Release to Data Services

- From Data and Information to Knowledge
  - Towards a Knowledge World
  - From Data Services to Knowledge Services

Dr. Evelyne Viegas

Microsoft Research

Computational Semantics
Natural Language Processing
Intelligent Planning

Studies in NATURAL LANGUAGE PROCESSING
COMPUTATIONAL LEXICAL SEMANTICS
Edited by Patrick Saint-Dizier and Evelyne Viegas
Editions Duculot
AUPELF UREF
Traitement automatiq
des langues naturell
BREADTH AND DEPTH OF SEMANTIC LEXICONS

Microsoft

Semantic Search

Language Processing for Help
Natural User Interactions

Microsoft Research

Semantic Computing

Data - Information - Knowledge - Intelligence

Beyond Data

Vision – Enable the *Next Generation Internet* by working with Academia, stakeholders from industry, government, and internet consumers/innovators to build an Intelligent Web, making sense of data via open innovation

DATA - INFORMATION - KNOWLEDGE - INTELLIGENCE

Data has become a first class citizen

IT’S A DATA-DRIVEN WORLD

It’s a data-driven world

- Spell Checking
- Machine Translation
- Search queries + click through
- Online games skill matching

Data logs behaviours in more reliable ways than demographic studies or surveys to study/predict trends

(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

Challenge in Data-driven Research

- Lot of the data needed for data-driven research in industry
  - Reason: scale; privacy, business sensitivity

How to make real world large scale data available to researchers to nurture innovation and perform valid experimentation, while maintaining privacy?

DATA RELEASE

DATA COMPUTE

DATA SERVICES

DATA ACADEMIC ENGAGEMENTS

Data for Open Innovation - Promises

Innovation

- By having access to real world data at scale, researchers can unveil **new** analysis or **research directions** based on shared assets and explore new questions

Science

- By allowing wider use of data, **repeatability of experiments** can be performed and data misrepresentations or faulty results avoided

Training

- Last but not least, real world large scale data is a powerful tool for **training the next generation of researchers**

Data for Open Innovation - Challenges

With web users becoming producers of information, leaving the footprint of their lives in digital trails, it is becoming easier for “data snoopers” to reconstruct the identity of an individual or an organization by cross linking information from different sources

A Face Is Exposed for Searcher No. 4417749

The New York Times, Aug 2006
Thelma Arnold's identity was betrayed by the records of her Web searches

"Search query data can contain the sum total of our work, interests, associations, desires, dreams, fantasies, and even darkest fears" said, Lauren Weinstein, a privacy advocate.

Accelerating Search in Academic Research Request for Proposals (RFPs)

Accelerating Search in Academic Research

Search RFP Awards

- Search assets (15 million search queries + click through)
  - PII (including inadvertent) removed
  - Provided under a limited data licensing agreement
- Increased quota to the Search API

Search Summit 2007

- Search RFP06 projects review
- The Quest for Assets – the Good the Bad and the Wanted

RFPs Program Feedback

- Search Summit 2007 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 up to *petabytes* of real world data

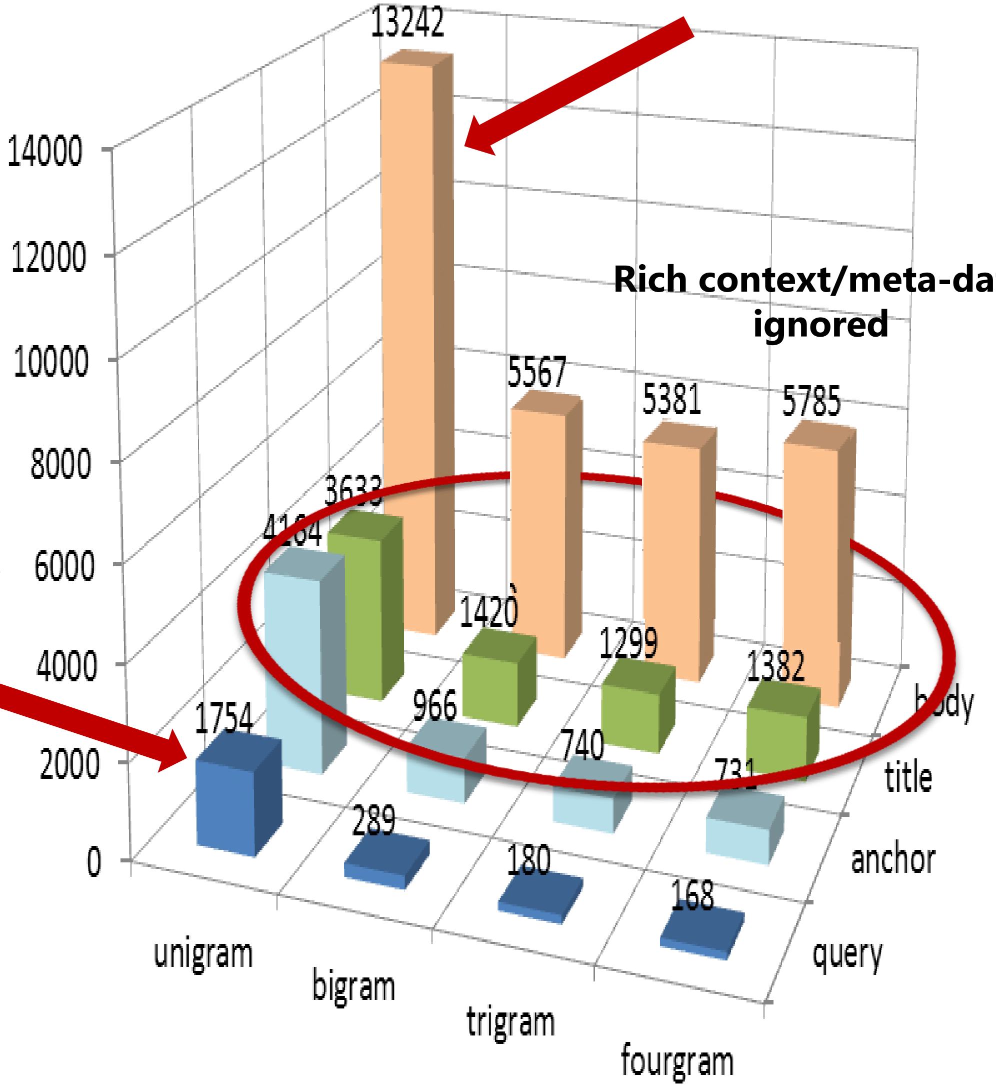
http://research.microsoft.com/web-ngram

Leading technology in Search, Machine Translation,
Speech, Learning

Web N-Gram in Public Beta

Web data has structure…

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



Search engines rely on unigram body …

Rich context/meta-data ignored

Users form ‘query’

Exploring Web Scale Language models for Search Query Processing, in *WWW’2010*

Web N-gram Services

Content types

- Document Body, **Document Title, Anchor Texts**

Training size (Body)

- ***All* documents** for en-us indexed by Bing (no cut off)

Access

- **Hosted Services** by Microsoft

Updates

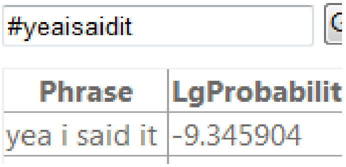
- **Periodical updates**

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

Word Breaking examples

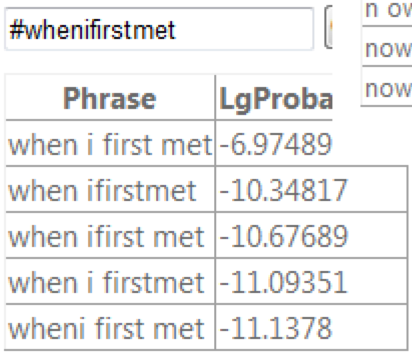
Enter a hash-tag phrase, and we will show the likely breakdown of sub-words. For instance, enter #nowplaying. More examples...

#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



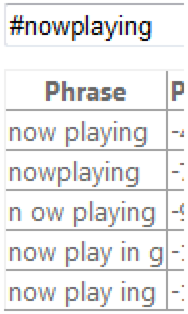
#yeaisaidit

Phrase	LgProbabilit
yea i said it	-9.345904



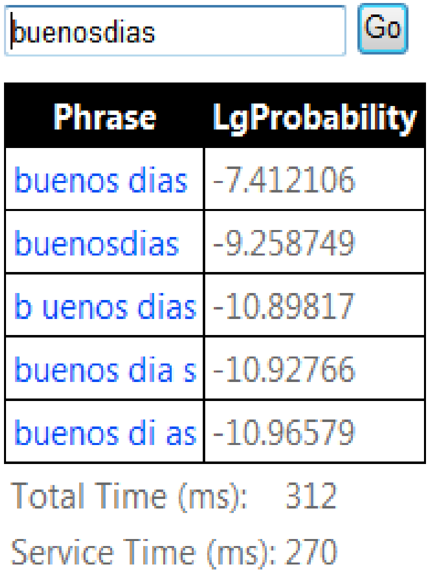
#whenifirstmet

Phrase	LgProba
when i first met	-6.97489
when ifirstmet	-10.34817
when ifirst met	-10.67689
when i firstmet	-11.09351
wheni first met	-11.1378



#nowplaying

Phrase	P
now playing	-
nowplaying	-
n ow playing	-
now play in g	-
now play ing	-

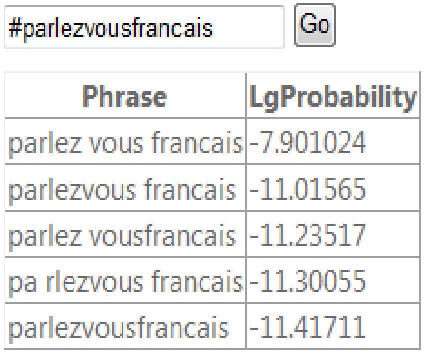


buenosdias

Phrase	LgProbability
buenos dias	-7.412106
buenosdias	-9.258749
b uenos dias	-10.89817
buenos dia s	-10.92766
buenos di as	-10.96579

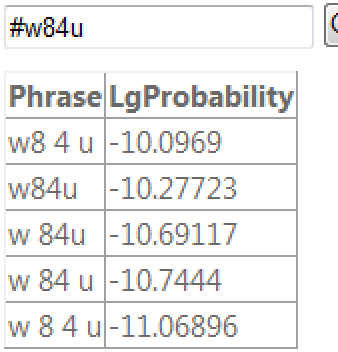
Total Time (ms): 312

Service Time (ms): 270



#parlezvousfrancais

Phrase	LgProbability
parlez vous francais	-7.901024
parlezvous francais	-11.01565
parlez vousfrancais	-11.23517
pa rlezvous francais	-11.30055
parlezvousfrancais	-11.41711



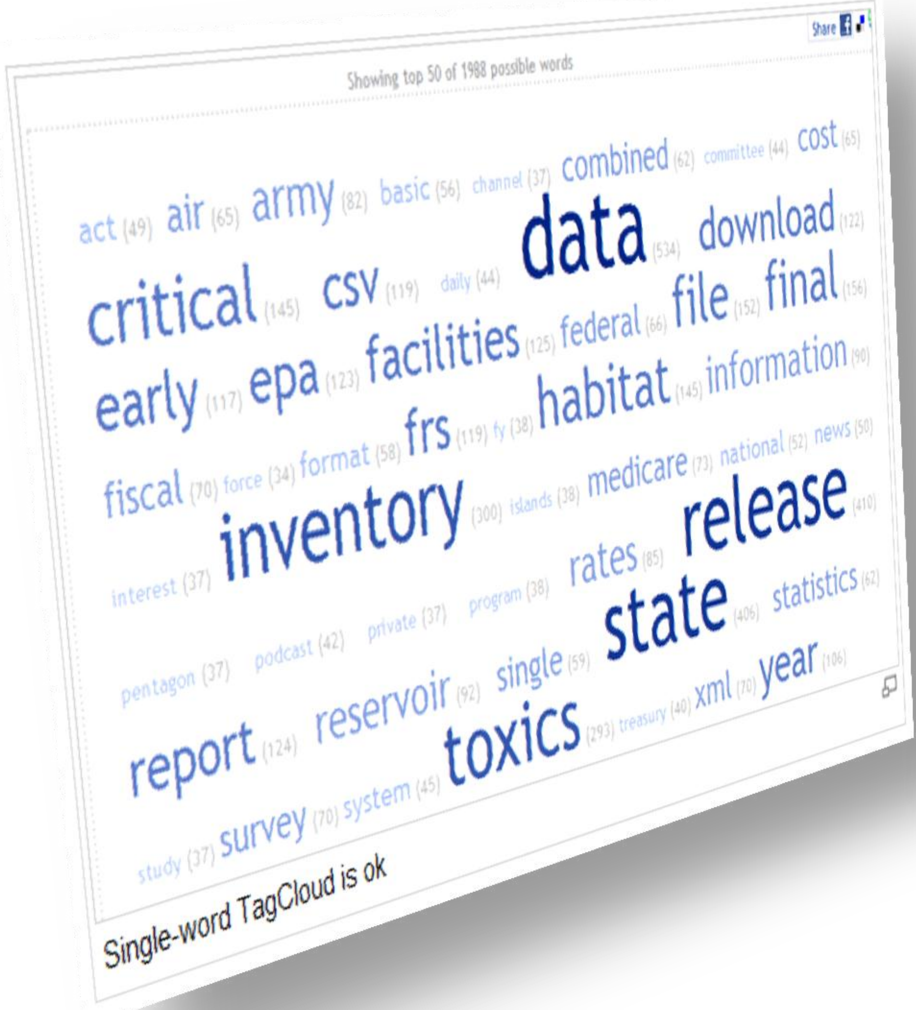
#w84u

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

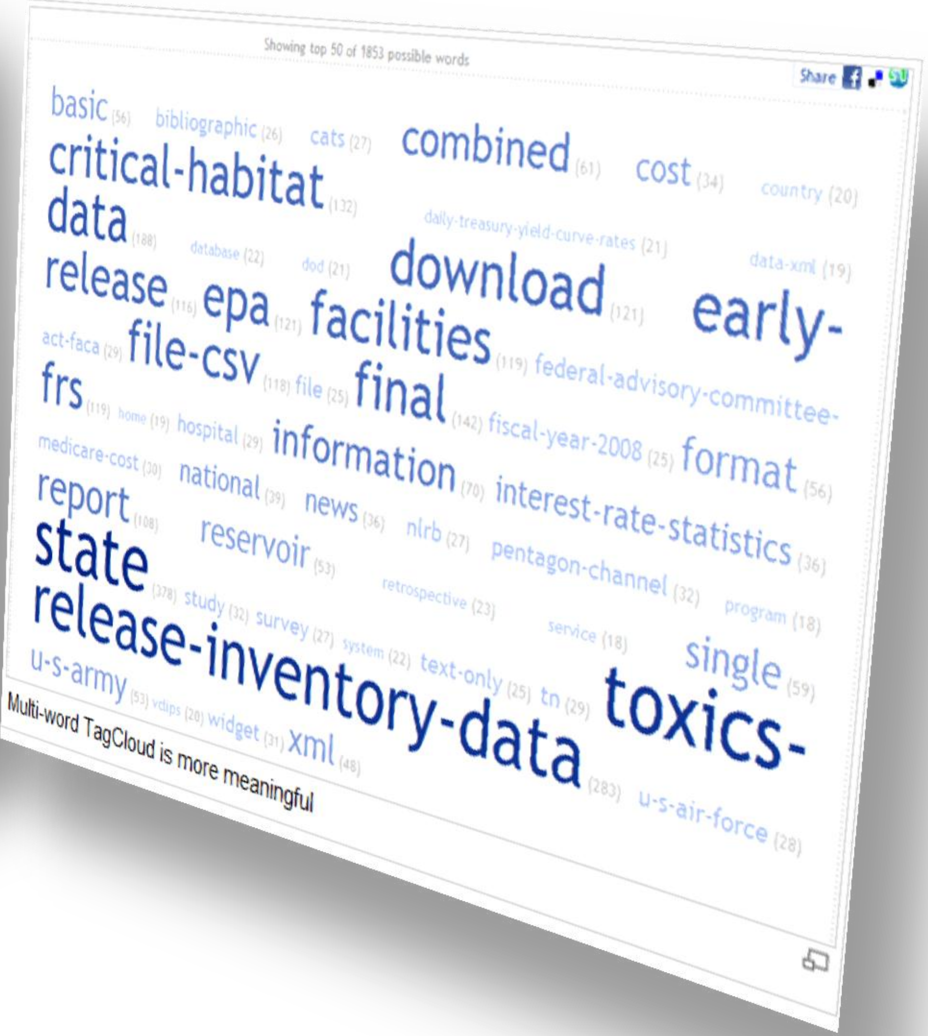
Single Tag Cloud

Multi Tag Cloud


Share
act (49) air (65) army (82) basic (56) channel (37) combined (62) committee (44) cost (65)
data (534) download (122)
critical (145) csv (119) daily (44) file (152) final (156)
early (117) epa (123) facilities (125) federal (66) habitat (145) information (90)
frs (119) fy (38)
fiscal (70) force (34) format (58) inventory (300) islands (38) medicare (73) national (52) news (50)
interest (37) release (410)
rates (85) program (38) private (37) state (406) statistics (62)
pentagon (37) podcast (42) single (59) year (106)
report (124) reservoir (92) toxics (293) treasury (40) xml (70)
study (37) survey (70) system (45)

Showing top 50 of 1988 possible words

Single-word TagCloud is ok


Share
basic (56) bibliographic (26) cats (27) combined (61) cost (34) country (20)
critical-habitat (132)
data (188) database (22) dod (21) daily-treasury-yield-curve-rates (21) data-xml (19)
release (116) epa (121) facilities (119) download (121) early-
act-faca (29) file-csv (118) file (25) final (142) federal-advisory-committee-
frs (119) home (19) hospital (29) information (70) fiscal-year-2008 (25) format (56)
medicare-cost (30) national (39) news (36) nlrb (27) interest-rate-statistics (36)
report (108) reservoir (53) pentagon-channel (32) retrospective (23) program (18)
state (378) study (32) survey (27) system (22) text-only (25) tn (29) service (18) single (59)
release-inventory-data (283) toxics-
u-s-army (53) vdips (20) widget (31) xml (48) u-s-air-force (28)

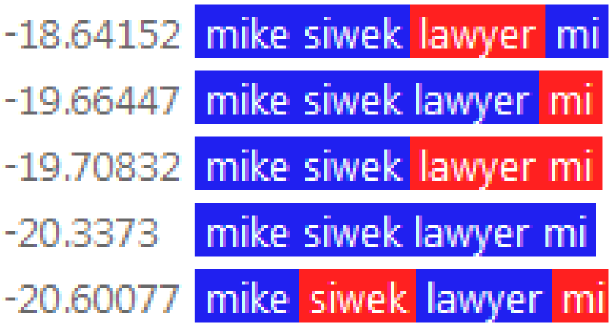
Showing top 50 of 1853 possible words

Multi-word TagCloud is more meaningful

Ref: Dr. Li Ding, Rensselaer Polytechnic Institute

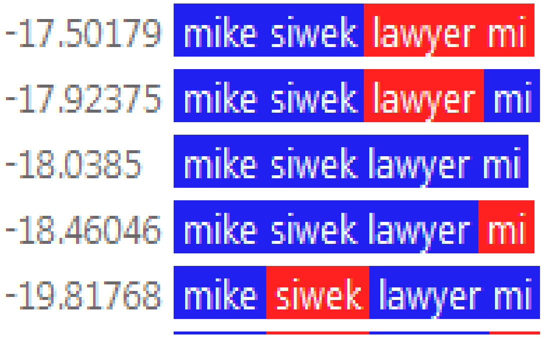
Query Segmentation

Body:



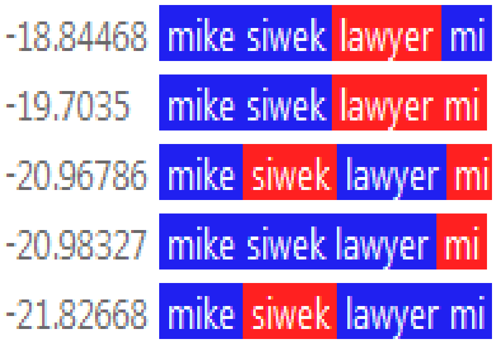
-18.64152	mike siwek | lawyer | mi
-19.66447	mike siwek lawyer | mi
-19.70832	mike siwek | lawyer mi
-20.3373	mike siwek lawyer mi
-20.60077	mike | siwek | lawyer | mi

Title:



-17.50179	mike siwek | lawyer mi
-17.92375	mike siwek | lawyer | mi
-18.0385	mike siwek lawyer mi
-18.46046	mike siwek lawyer | mi
-19.81768	mike | siwek | lawyer mi

Anchor:



-18.84468	mike siwek | lawyer | mi
-19.7035	mike siwek | lawyer mi
-20.96786	mike | siwek | lawyer | mi
-20.98327	mike siwek lawyer | mi
-21.82668	mike | siwek | lawyer mi

Impact with Microsoft Web N-gram Service

- Sheer power of data
  - Cross lingual documents are a way of life. N-grams seem to work on other languages

- Documents have structure and styles
  - A single document is written in many languages, with the document body, title and anchor text being all different languages that should be treated separately
  - Web has other languages such as those used for SMS. The N-gram Service works on this kind of language which opens up a lot of interesting research questions

Are we revisiting the concept of “language identification” as a means of identifying languages of different styles, and not so much on national languages (Wang et al., NAACL-HLT 2010)

Use Microsoft Web N-gram Services and get to Webscale

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

webngram@microsoft.com

Free for *non* commercial research
Scaling to TeraBytes and PetaBytes
Regular data/feature updates
ISRC Research team to engage with

Available on Azure for the awardees of the NSF Program Solicitation Computing in the Cloud

Research papers (SIGIR 2010)

SIGIR 2010
Geneva, Switzerland
July 19-23, 2010

Web N-gram Workshop

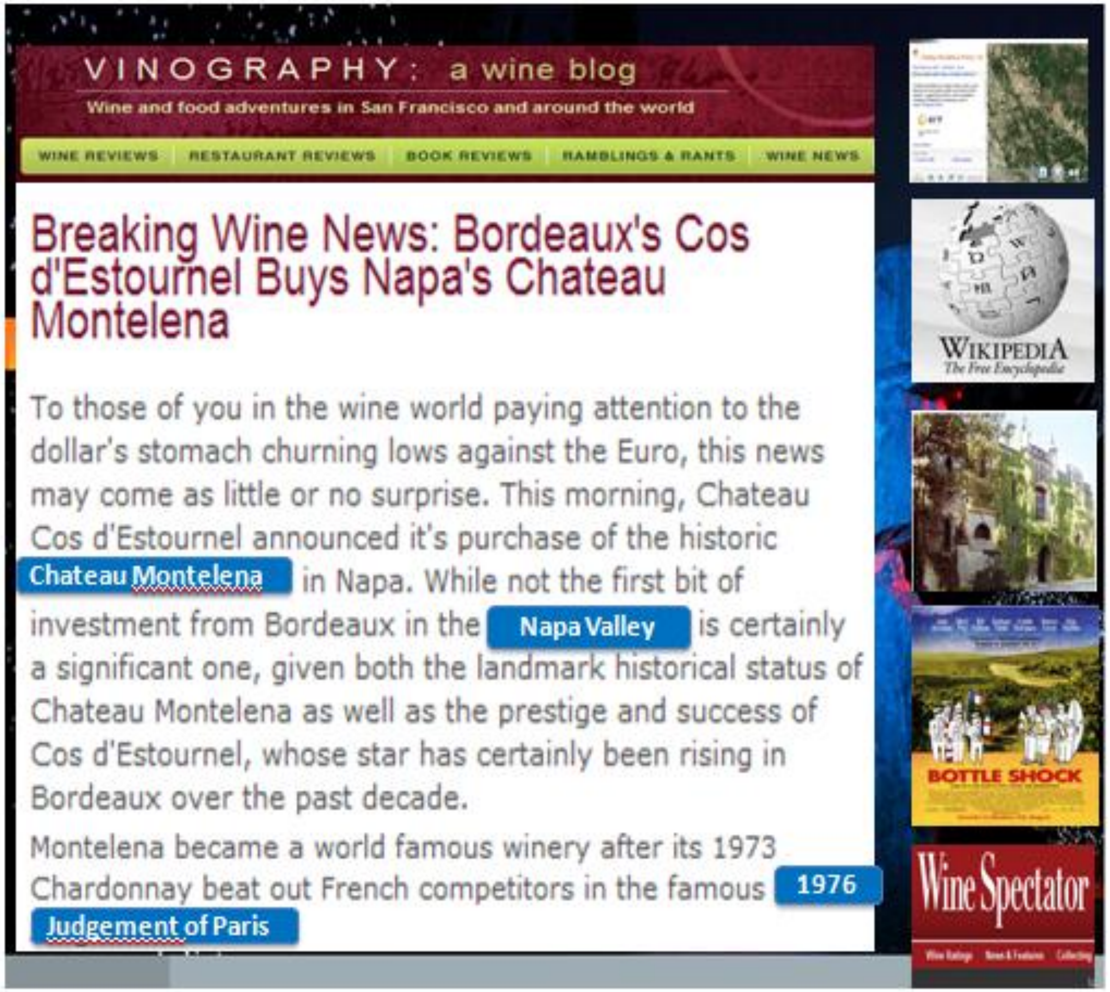
Workshop of the 33rd Annual International
ACM SIGIR Conference
on Research and Development
in Information Retrieval

Organised by
Chengxiang Zhai
David Yarowsky
Evelyne Viegas
Kuansan Wang
Stephan Vogel

Association for Computing Machinery

SIGIR Geneva 2010

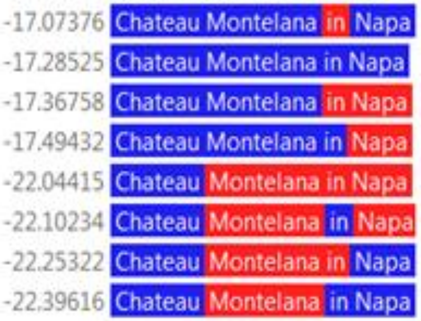
Implicit Search



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

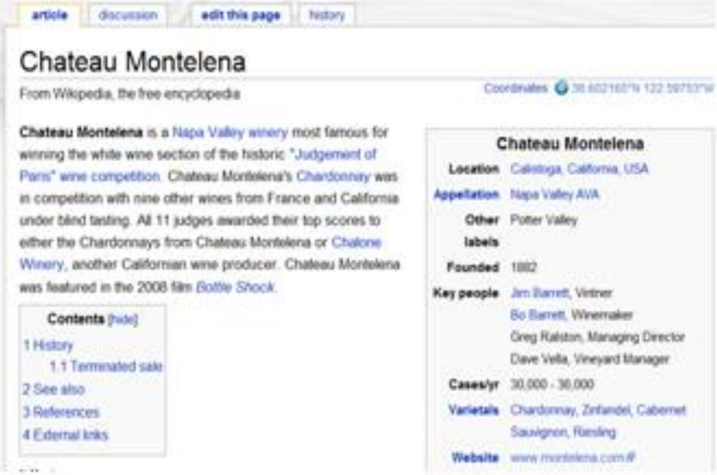
To those of you in the wine world paying attention to the dollar's stomach churning lows against the Euro, this news may come as little or no surprise. This morning, Chateau Cos d'Estournel announced it's purchase of the historic Chateau Montelena in Napa. While not the first bit of investment from Bordeaux 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 1976 Judgement of Paris



-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

From Wikipedia, the free encyclopedia

Coordinates: 38.602165°N 122.597537°W

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*.

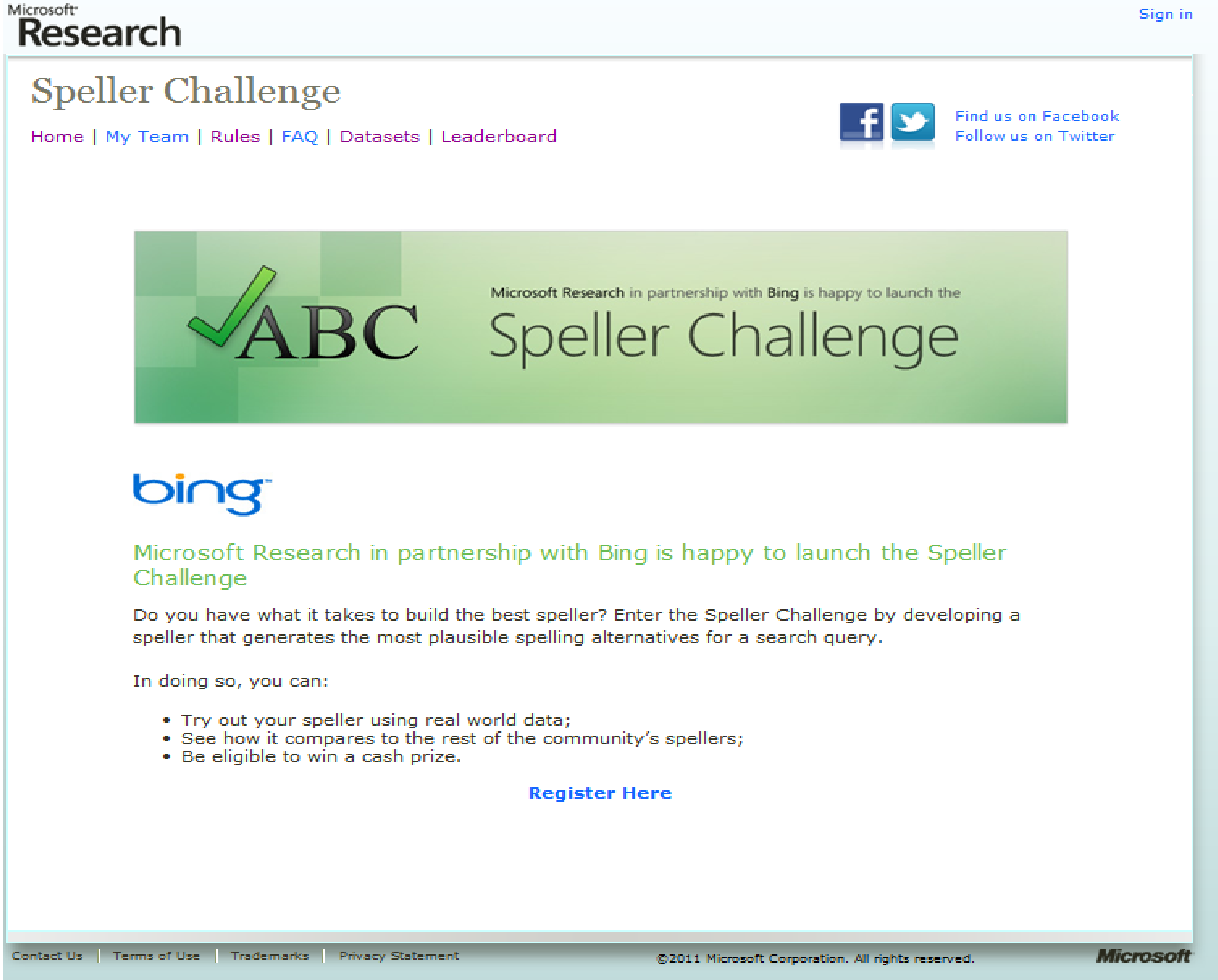
Contents [hide]

Chateau Montelena	
Location	Calistoga, California, USA
Appellation	Napa Valley AVA
Other labels	Potter Valley
Founded	1882
Key people	Jim Barrett, Vintner Bo Barrett, Winemaker Greg Ralston, 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

Spelling Alteration for Web Search
http://spellerchallenge.com

SPELLER CHALLENGE
BING – MICROSOFT RESEARCH PARTNERSHIP



Microsoft Research

Sign in

Speller Challenge

Home | My Team | Rules | FAQ | Datasets | Leaderboard

Find us on Facebook
Follow us on Twitter

Microsoft Research in partnership with Bing is happy to launch the

Speller Challenge

bing

Microsoft Research in partnership with Bing is happy to launch the Speller Challenge

Do you have what it takes to build the best speller? Enter the Speller Challenge by developing a speller that generates the most plausible spelling alternatives for a search query.

In doing so, you can:

- Try out your speller using real world data;
- See how it compares to the rest of the community's spellers;
- Be eligible to win a cash prize.

Register Here

Contact Us | Terms of Use | Trademarks | Privacy Statement

Microsoft

Speller Challenge

Start date: Dec 15, 2010

End date: May 27, 2011

Web Ngram Services used to create data set by participants

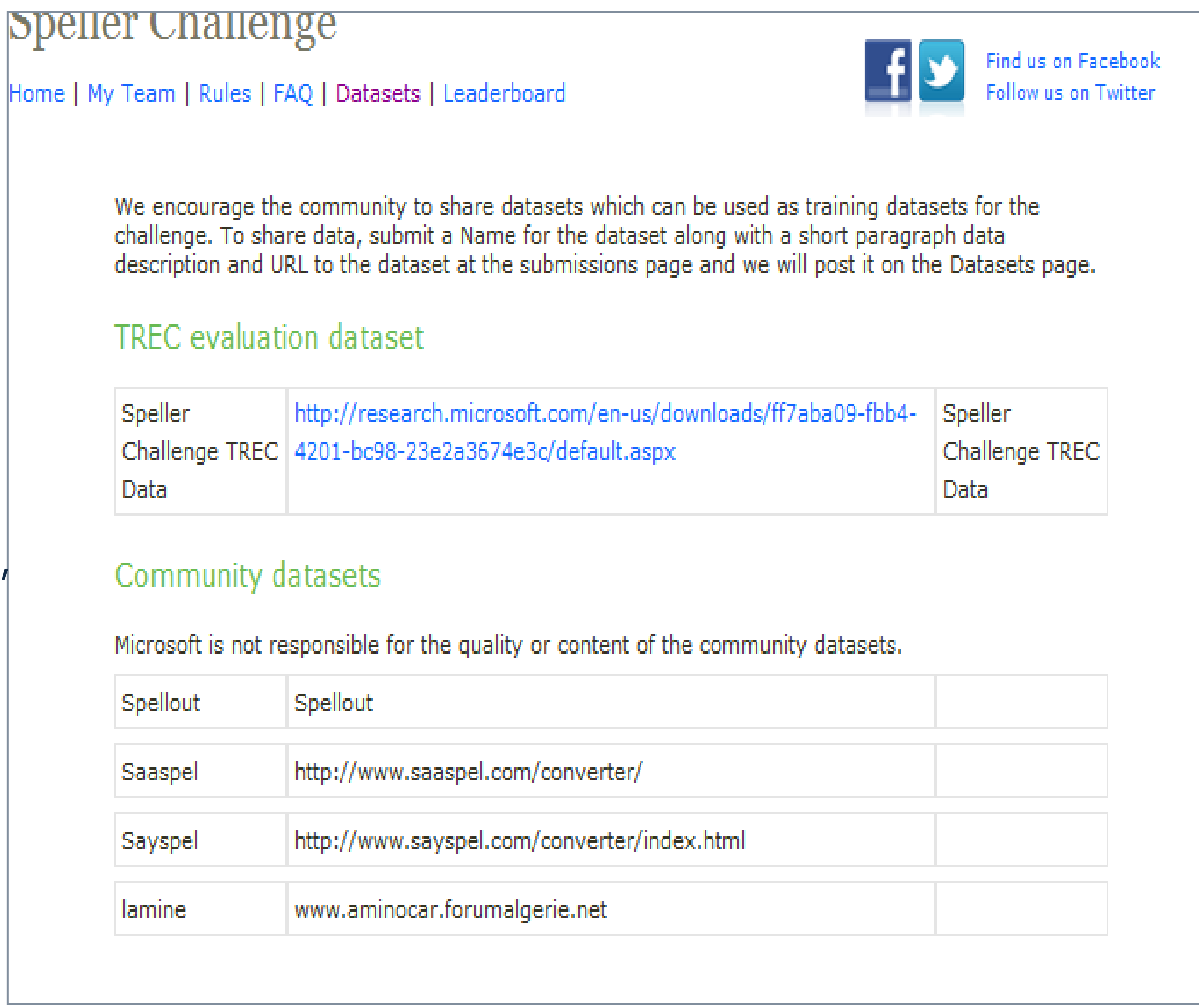
Automatic evaluation of the participants’ spellers

Five Prizes to win

Learning - Community “shared data sets”

Speller Challenge TREC Data

Participants can make their data set available to the rest of the research community by providing a link to their data set, on the challenge “community datasets” page.



We encourage the community to share datasets which can be used as training datasets for the challenge. To share data, submit a Name for the dataset along with a short paragraph data description and URL to the dataset at the submissions page and we will post it on the Datasets page.

TREC evaluation dataset

Speller Challenge TREC Data	http://research.microsoft.com/en-us/downloads/ff7aba09-fbb4-4201-bc98-23e2a3674e3c/default.aspx	Speller Challenge TREC Data

Community datasets

Microsoft is not responsible for the quality or content of the community datasets.

Spellout	Spellout	
Saaspel	http://www.saaspel.com/converter/	
Sayspel	http://www.sayspel.com/converter/index.html	
lamine	www.aminocar.forumalgerie.net	

Summary and Questions

- From data release to data services
  - Allows to handle much bigger data sets
  - Allows to provide abstractions on the data (e.g. language models)
  - Allows to provide data compute capabilities
  - Allows for agile experimentation

- How to better engage with academia to drive data-driven research?
  - How does a cloud-based data service approach change research?

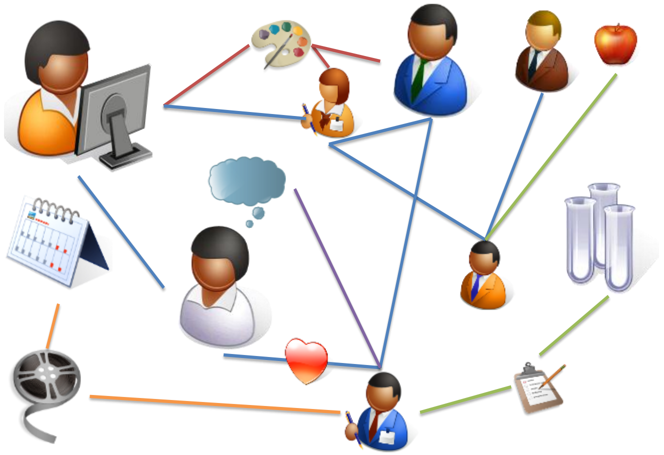
- What else can industries do to help democratize large scale data-driven research?

Contact: evelynev@microsoft.com

The world has become more connected

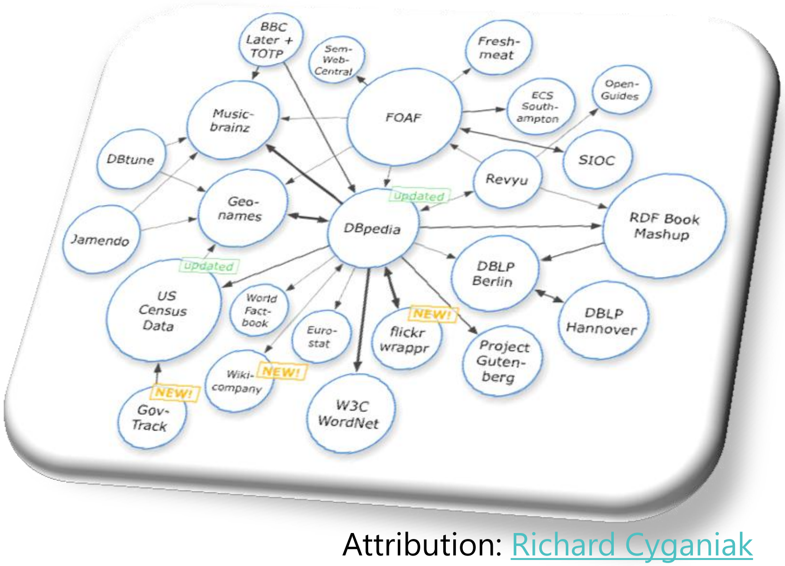
TOWARDS A KNOWLEDGE WORLD

A world where all data is linked…



- Information inter-connected through machine-interpretable information (e.g. paper X **is about** star Y)

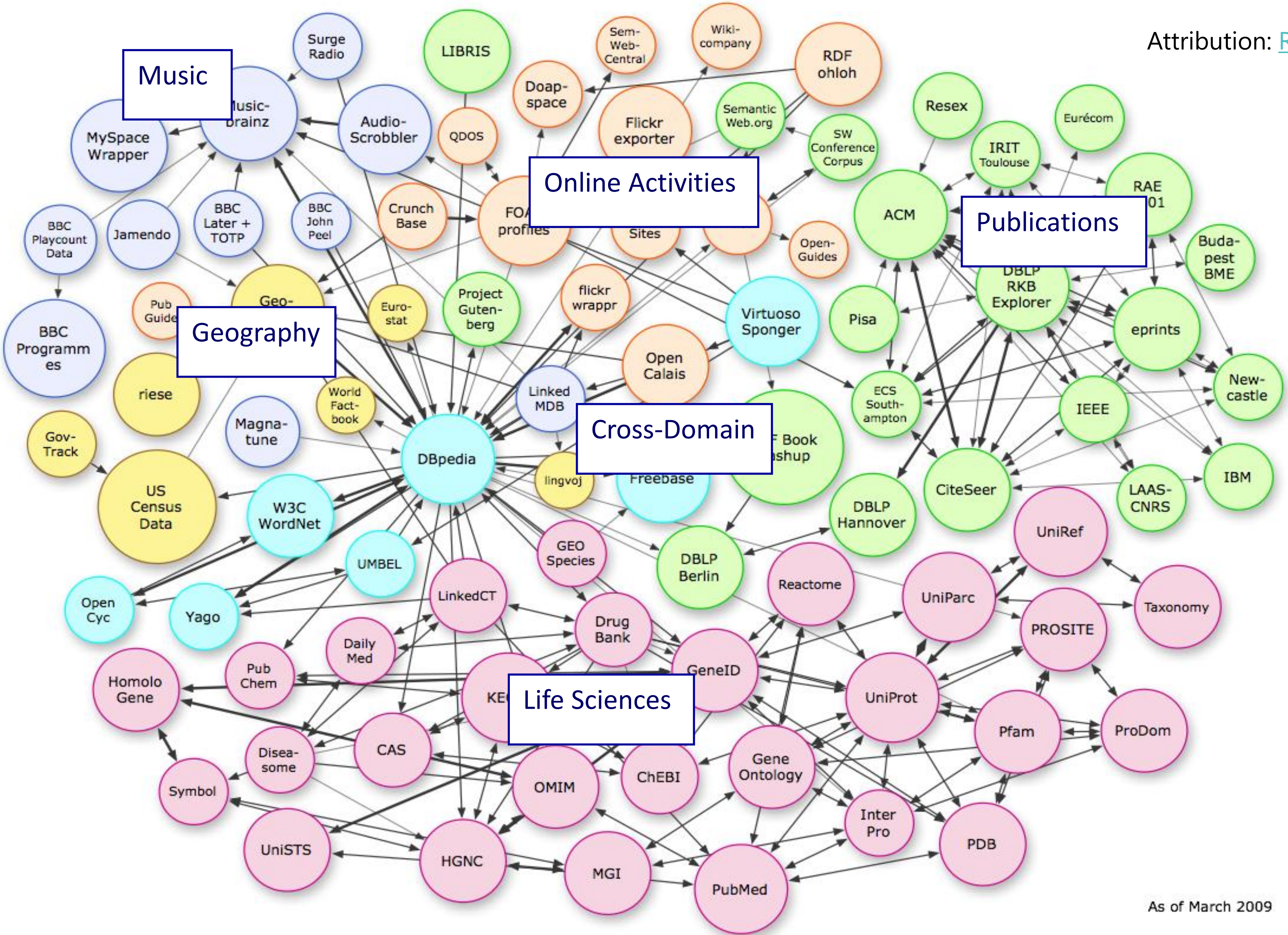
- Formats or “well-known” representations of data/information
- Pervasive access protocols are key (e.g. HTTP)
- Data/information is uniquely identified (e.g. URIs)
- Links/associations between data/information



Attribution: Richard Cyganiak

Linked Data

Attribution: Richard Cyganiak



As of March 2009

Semantics at Web Scale

- **Data, information is dynamic**
  - 450,000 changes per day in Wikipedia (en)
  - One new word created every 98 minutes (14.7 a day)

- **Is there a "right ontology?"**
  - Ontologies are abstractions
    - Ontologists make design choices all the time
  - Ontologies are application dependent
    - Machine Translation: *eat* vs. *comer* (INGEST concept)
    - Robotics: *eat* (task planning)

Probabilistic Modeling for Merging Knowledge Bases

- Yago - Ontology based on data from Wikipedia, WordNet and GeoNames with 400 M facts, 108 relations
- IMDb - Ontology based on data from the IMDb TV & movie site with 23M facts and 10 relations

- Challenges
  - Differences in information coverage
  - Text Mismatches (e.g. Kim_Novak vs. Novak, Kim)
  - Concept Mismatches (e.g. wasCreatedOnDate vs. hasProductionYear)
  - Granularity differences (e.g. one entity for one TV series vs. multiple entities per episode)

~20% entities exact match

Ontology Merging as a Probabilistic Inference and Learning Problem

- Are these two entities the same? Are these two relations the same?
- Can we learn the string transformations in an unsupervised manner from data?
- Given examples, can we learn a good cost function / probability model for matching?
- If two ontologies disagree, which is more reliable?
- How do we represent uncertainty in our ontologies?

Vision: a tool that automatically merges 2 ontologies

Zoubin Ghahramani, Univ. of Cambridge

- Capture concepts (in our mental world)
- Quantify uncertainty (for reasoning)

Probase – 2.7 M concepts

automatically harnessed

Freebase – 2 K concepts

built by community effort

Cyc – 120 K concepts

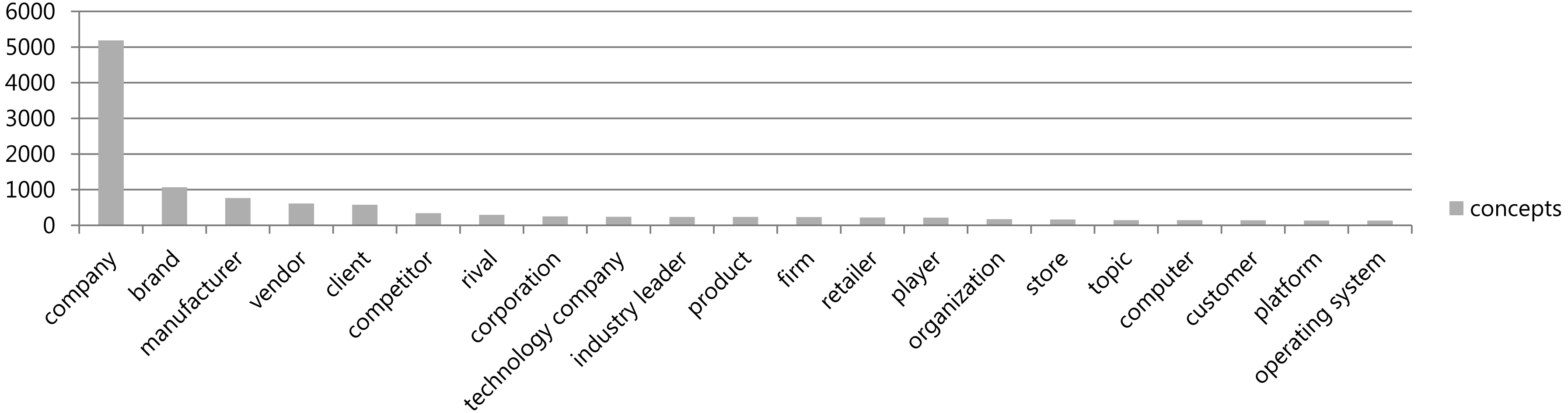
25 years human labor

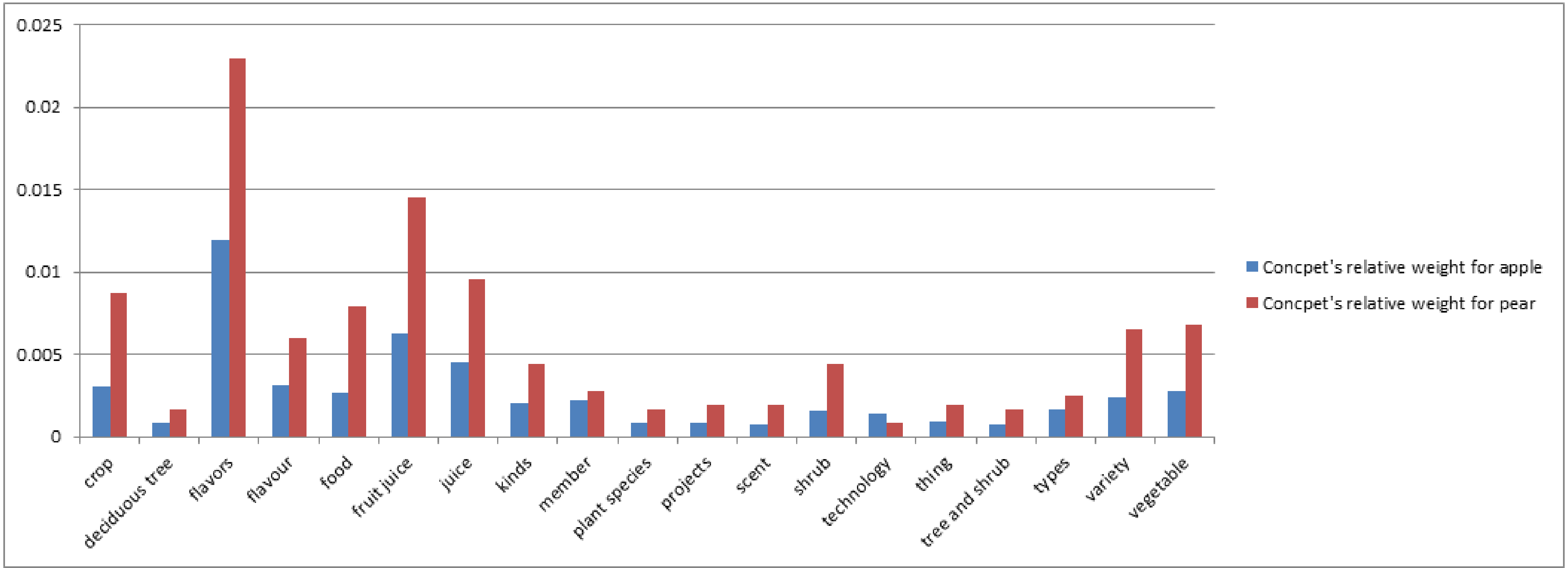
Uncertainty

Probase vs. Free**base**

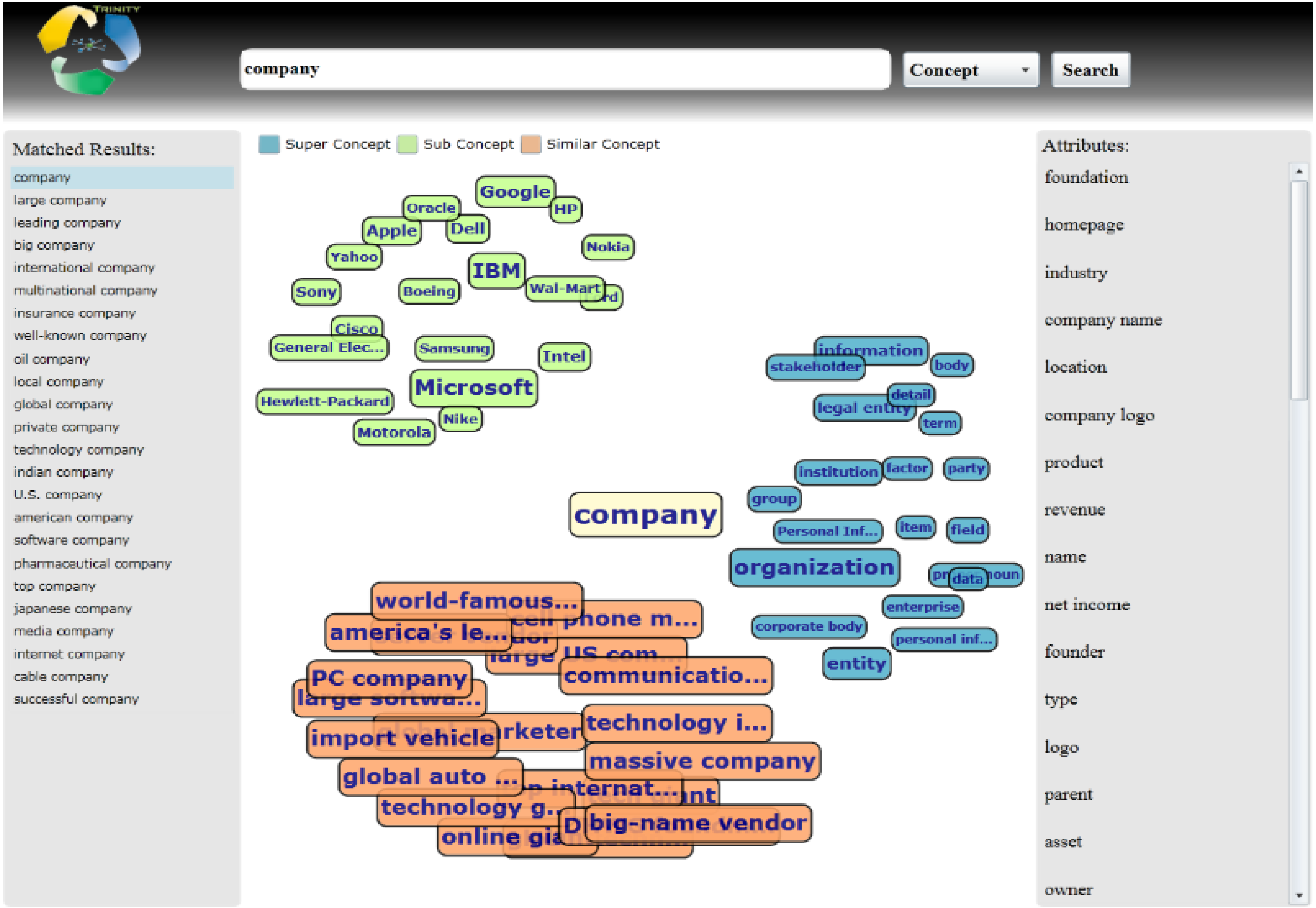
Correctness is a probability.	Knowledge is black and white.
Live with dirty data.	Clean up everything.
Dirty data is very useful.	Dirty data is unusable.

Meaning of words in context (apple, pear)





Concept Search



company

Concept ▾ | Search

Matched Results:

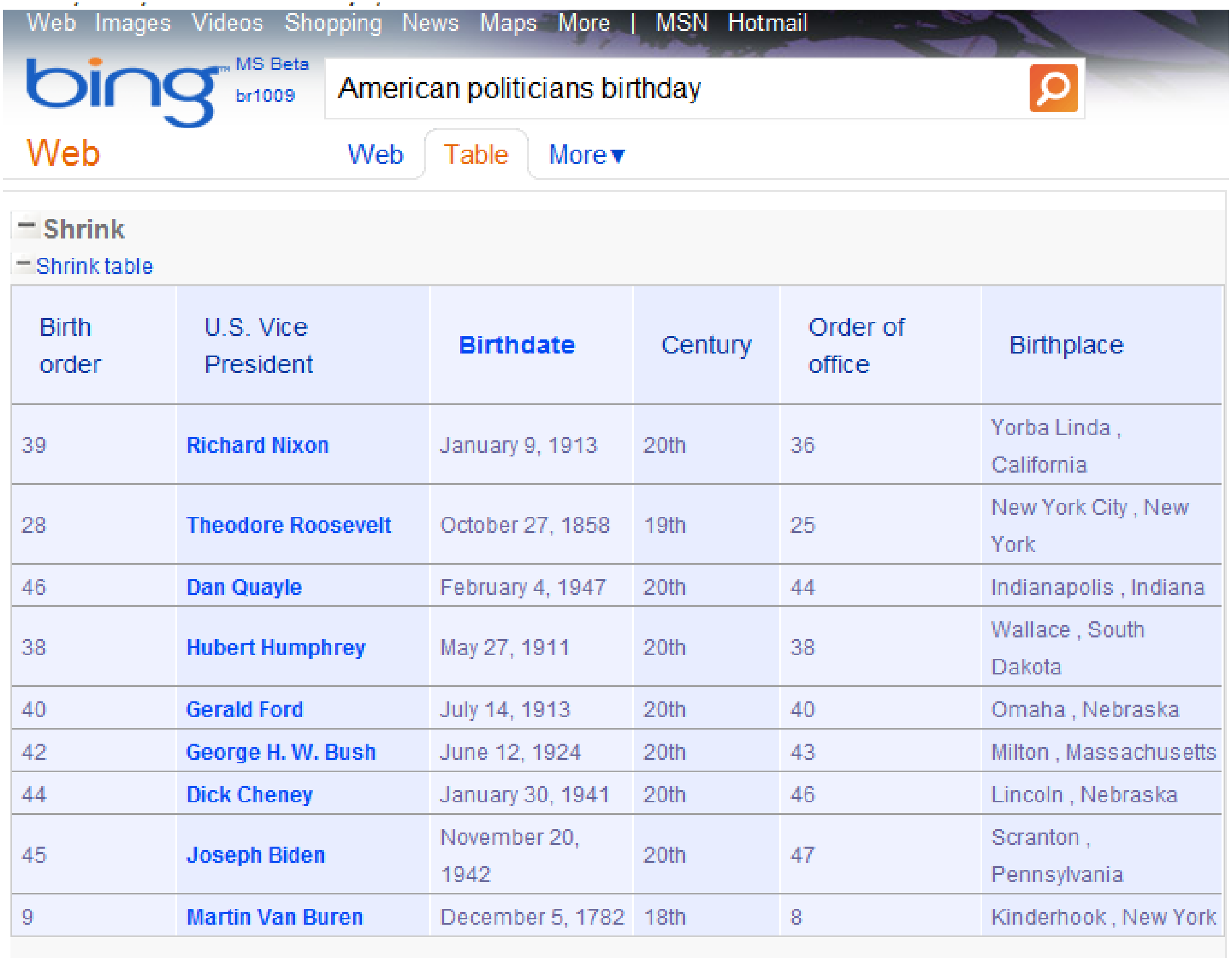
- company
- large company
- leading company
- big company
- international company
- multinational company
- insurance company
- well-known company
- oil company
- local company
- global company
- private company
- technology company
- indian company
- U.S. company
- american company
- software company
- pharmaceutical company
- top company
- japanese company
- media company
- internet company
- cable company
- successful company

🟦 Super Concept 🟩 Sub Concept 🟧 Similar Concept

Attributes:

- foundation
- homepage
- industry
- company name
- location
- company logo
- product
- revenue
- name
- net income
- founder
- type
- logo
- parent
- asset
- owner

Table understanding



American politicians birthday

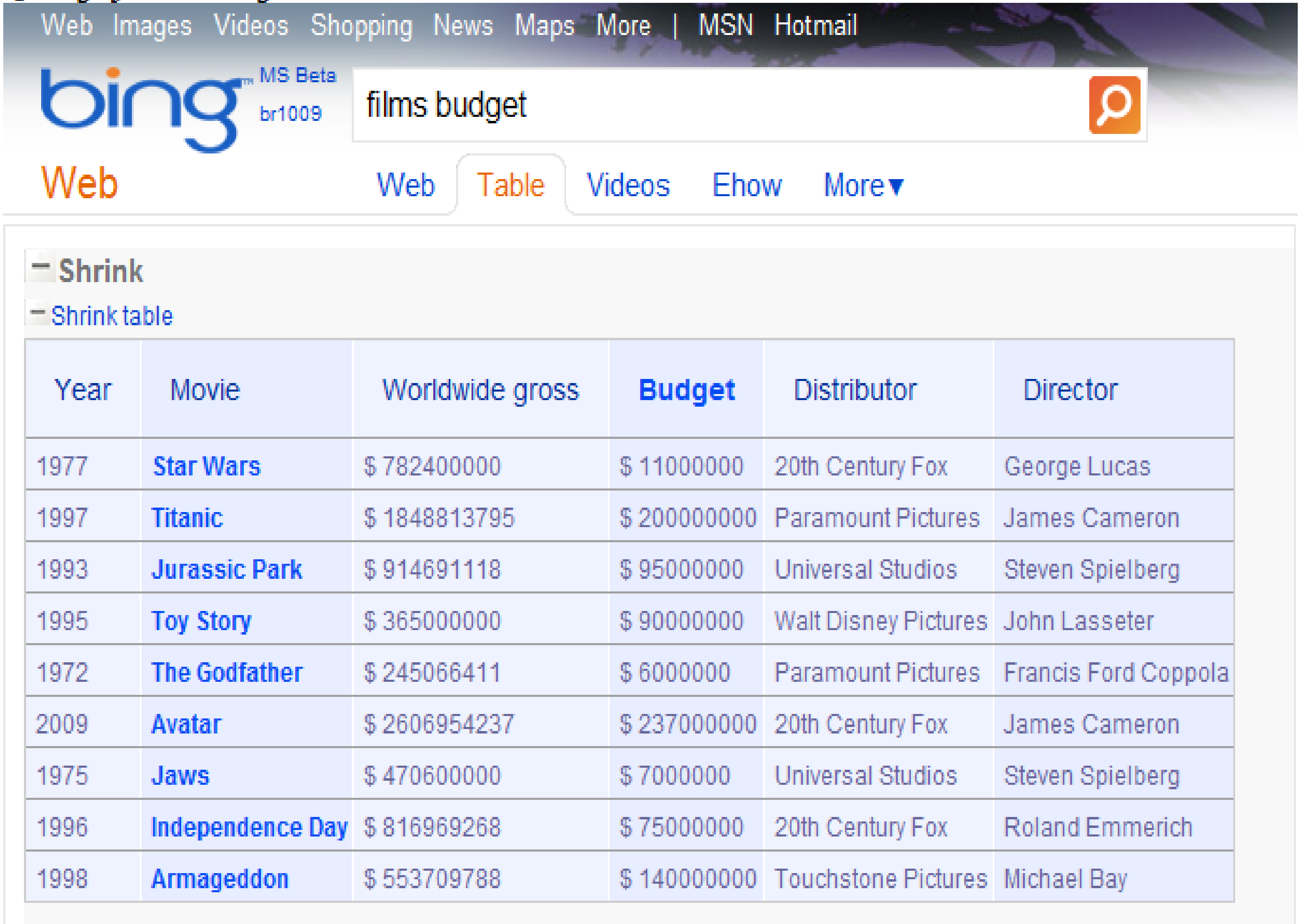
− Shrink

− Shrink table

Birth order	U.S. Vice President	Birthdate	Century	Order of office	Birthplace
39	Richard Nixon	January 9, 1913	20th	36	Yorba Linda , California
28	Theodore Roosevelt	October 27, 1858	19th	25	New York City , New York
46	Dan Quayle	February 4, 1947	20th	44	Indianapolis , Indiana
38	Hubert Humphrey	May 27, 1911	20th	38	Wallace , South Dakota
40	Gerald Ford	July 14, 1913	20th	40	Omaha , Nebraska
42	George H. W. Bush	June 12, 1924	20th	43	Milton , Massachusetts
44	Dick Cheney	January 30, 1941	20th	46	Lincoln , Nebraska
45	Joseph Biden	November 20, 1942	20th	47	Scranton , Pennsylvania
9	Martin Van Buren	December 5, 1782	18th	8	Kinderhook , New York

Table understanding

Query: films budget


Web Images Videos Shopping News Maps More | MSN Hotmail
bing MS Beta
br1009
films budget
Web
Web Table Videos Ehow More▼
Shrink
Shrink table

Year	Movie	Worldwide gross	**Budget**	Distributor	Director
1977	**Star Wars**	$ 782400000	$ 11000000	20th Century Fox	George Lucas
1997	**Titanic**	$ 1848813795	$ 200000000	Paramount Pictures	James Cameron
1993	**Jurassic Park**	$ 914691118	$ 95000000	Universal Studios	Steven Spielberg
1995	**Toy Story**	$ 365000000	$ 90000000	Walt Disney Pictures	John Lasseter
1972	**The Godfather**	$ 245066411	$ 6000000	Paramount Pictures	Francis Ford Coppola
2009	**Avatar**	$ 2606954237	$ 237000000	20th Century Fox	James Cameron
1975	**Jaws**	$ 470600000	$ 7000000	Universal Studios	Steven Spielberg
1996	**Independence Day**	$ 816969268	$ 75000000	20th Century Fox	Roland Emmerich
1998	**Armageddon**	$ 553709788	$ 140000000	Touchstone Pictures	Michael Bay

Summary

- Data to drive innovation
  - From Data Release to Data Services
  - Next Generation Innovations

- From Data and Information to Knowledge
  - Towards a Knowledge World
  - From Data Services to Knowledge Services

- How to better engage with academia to drive semantic computing research?
  - How does a cloud-based data and knowledge service approach change research?

- What else can industries do to help democratize semantic computing research?

Contact: evelynev@microsoft.com

Microsoft®

Your potential. Our passion.™

Acknowledgments

Jianfeng Gao, Zoubin Ghahramani, Mark Greaves, Savas Parastatidis, Chris Thrasher, Haixun Wang, Kuansan Wang, …

Microsoft®

The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation.

MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.