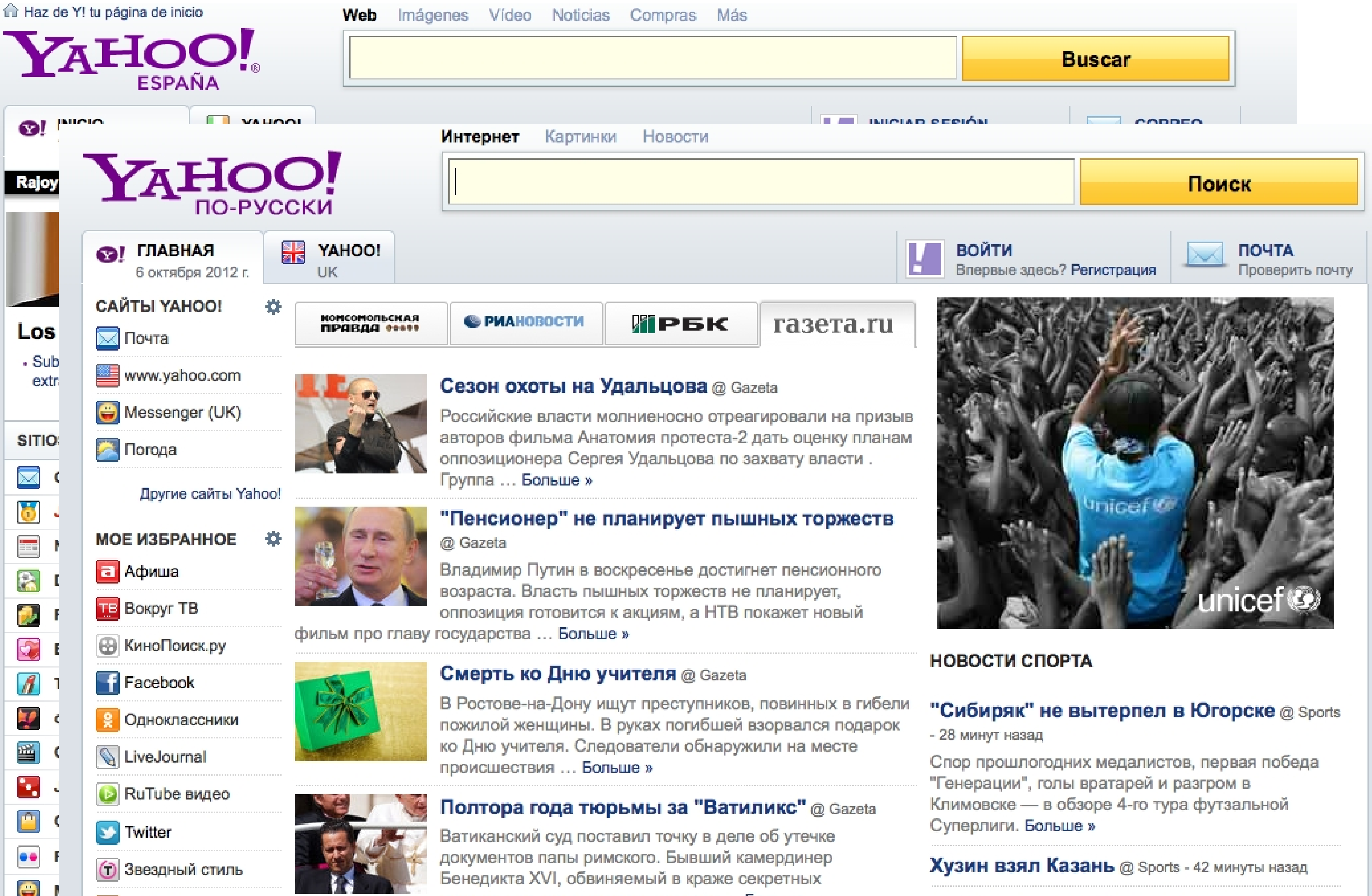
Semantic Search Evaluation

Peter Mika

Senior Research Scientist

Yahoo! Research

Yahoo! serves over 700 million users in 25 countries



Haz de Y! tu página de inicio

YAHOO! ESPAÑA

Web Imágenes Vídeo Noticias Compras Más

Buscar

Интернет Картинки Новости

YAHOO! ПО-РУССКИ

Поиск

ГЛАВНАЯ
6 октября 2012 г.

YAHOO!
UK

ВОЙТИ
Впервые здесь? Регистрация

ПОЧТА
Проверить почту

САЙТЫ YAHOO!
- Почта
- www.yahoo.com
- Messenger (UK)
- Погода

Другие сайты Yahoo!

МОЕ ИЗБРАННОЕ
- Афиша
- Вокруг ТВ
- КиноПоиск.ру
- Facebook
- Одноклассники
- LiveJournal
- RuTube видео
- Twitter
- Звездный стиль

Комсомольская правда | РИАНОВОСТИ | РБК | газета.ru

Сезон охоты на Удальцова @ Gazeta
Российские власти молниеносно отреагировали на призыв авторов фильма Анатомия протеста-2 дать оценку планам оппозиционера Сергея Удальцова по захвату власти . Группа … Больше »

"Пенсионер" не планирует пышных торжеств @ Gazeta
Владимир Путин в воскресенье достигнет пенсионного возраста. Власть пышных торжеств не планирует, оппозиция готовится к акциям, а НТВ покажет новый фильм про главу государства … Больше »

Смерть ко Дню учителя @ Gazeta
В Ростове-на-Дону ищут преступников, повинных в гибели пожилой женщины. В руках погибшей взорвался подарок ко Дню учителя. Следователи обнаружили на месте происшествия … Больше »

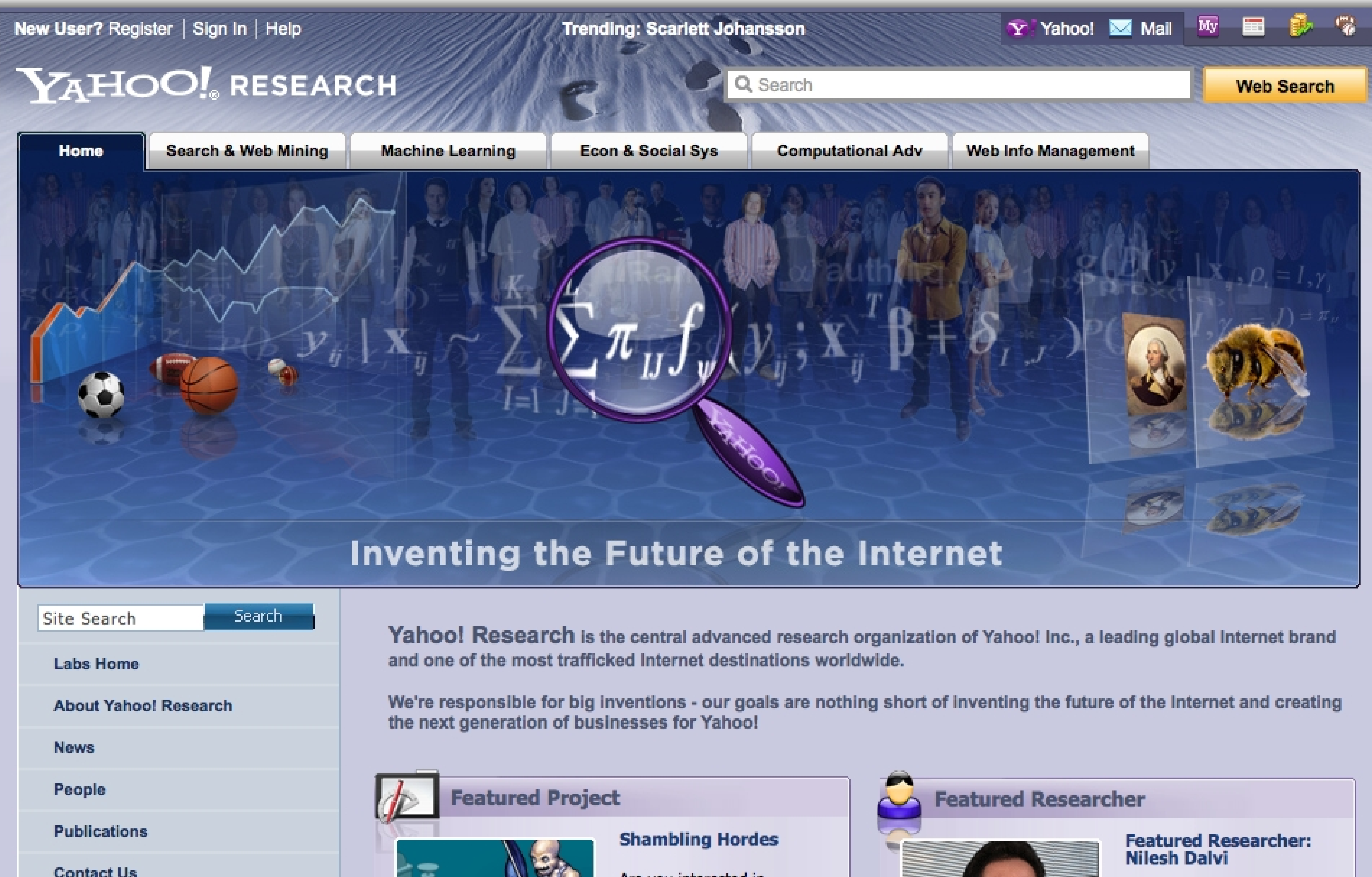
Полтора года тюрьмы за "Ватиликс" @ Gazeta
Ватиканский суд поставил точку в деле об утечке документов папы римского. Бывший камердинер Бенедикта XVI, обвиняемый в краже секретных

НОВОСТИ СПОРТА

"Сибиряк" не вытерпел в Югорске @ Sports - 28 минут назад
Спор прошлогодних медалистов, первая победа "Генерации", голы вратарей и разгром в Климовске — в обзоре 4-го тура футзальной Суперлиги. Больше »

Хузин взял Казань @ Sports - 42 минуты назад

Yahoo! Research: visit us at research.yahoo.com



Yahoo! Research is the central advanced research organization of Yahoo! Inc., a leading global Internet brand and one of the most trafficked Internet destinations worldwide.

We're responsible for big inventions - our goals are nothing short of inventing the future of the Internet and creating the next generation of businesses for Yahoo!

Shambling Hordes

Are you interested in

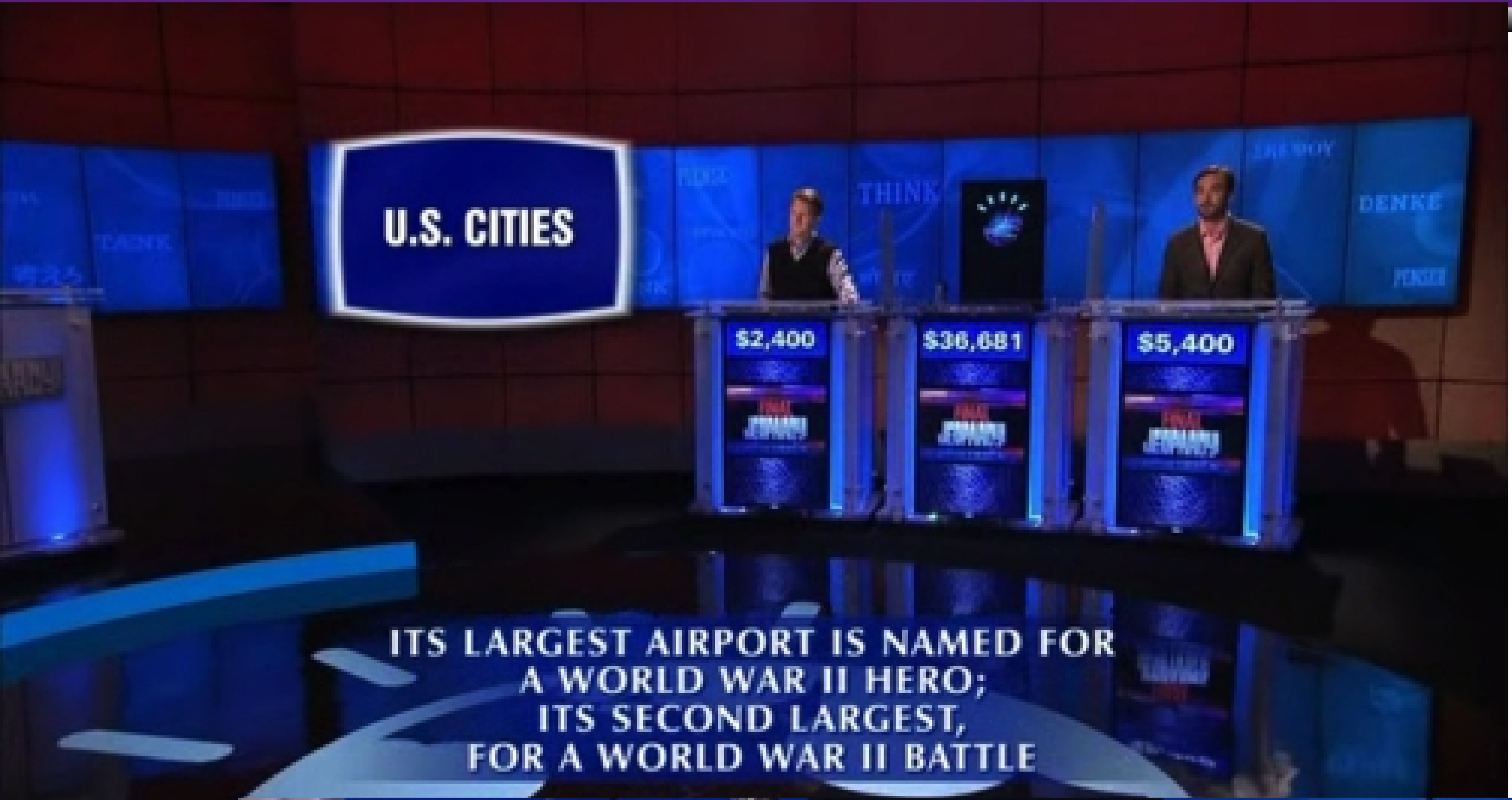
Featured Researcher: Nilesh Dalvi

Yahoo! Research Barcelona

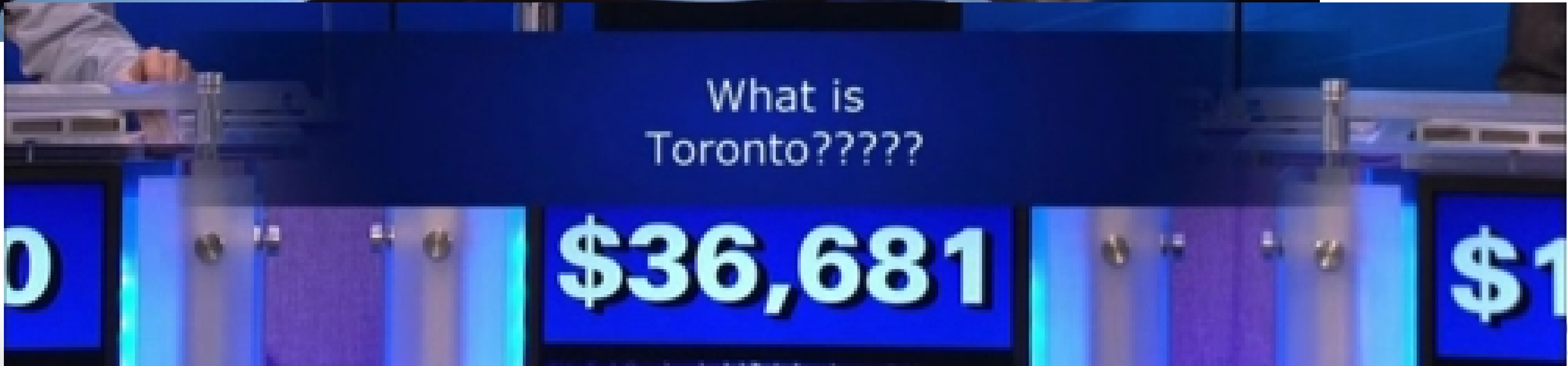
- Established January, 2006
- Led by Ricardo Baeza-Yates
- Research areas
  - Web Mining
    - content, structure, usage
  - Social Media
  - Distributed Systems
  - Semantic Search

http://www.flickr.com/photos/bcnbits

Search is really fast, without necessarily being intelligent



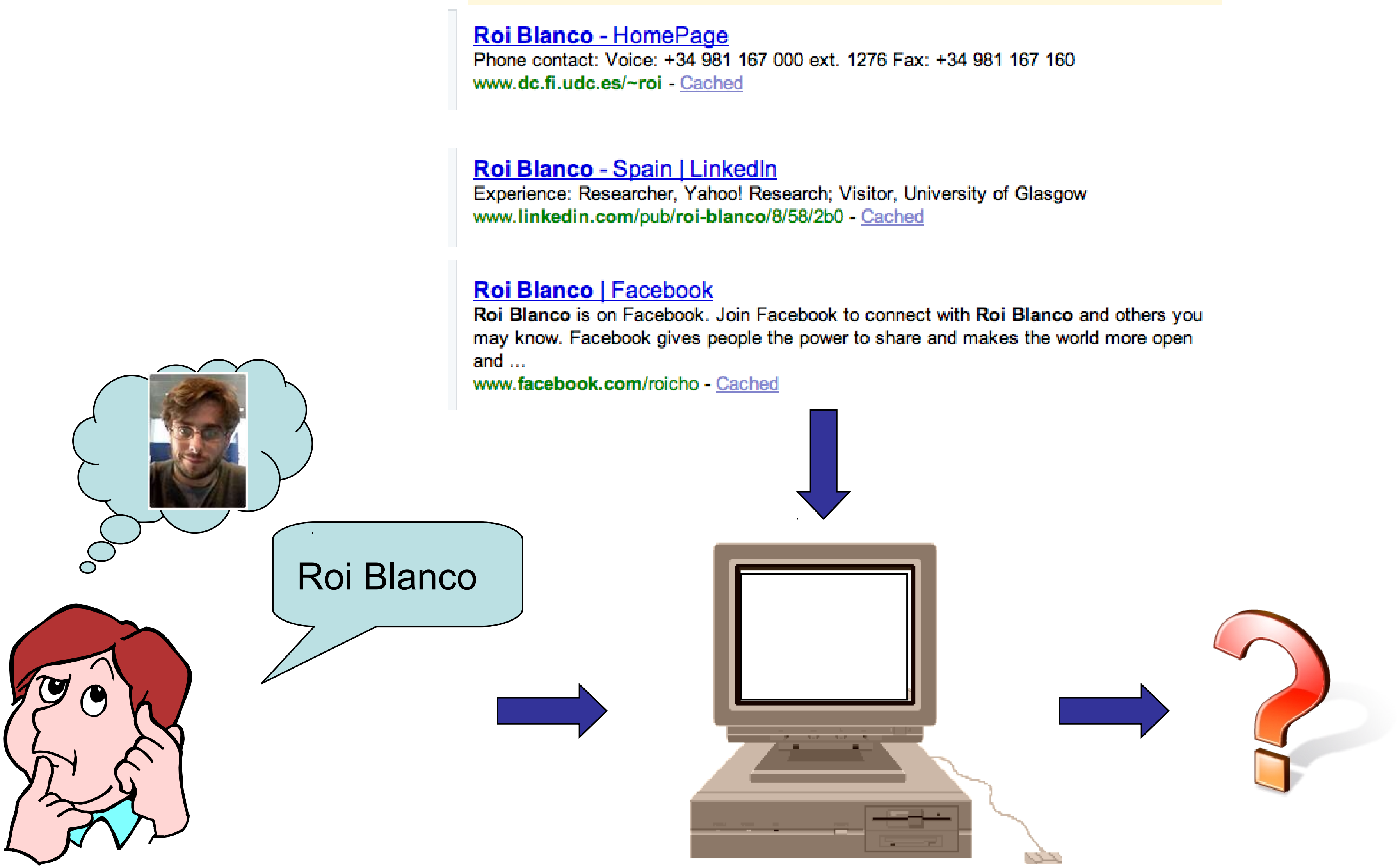
ITS LARGEST AIRPORT IS NAMED FOR A WORLD WAR II HERO; ITS SECOND LARGEST, FOR A WORLD WAR II BATTLE


What is Toronto?????

Why Semantic Search? Part I

- Improvements in IR are harder and harder to come by
  - Machine learning using hundreds of features
    - Text-based features for matching
    - Graph-based features provide authority
  - Heavy investment in computational power, e.g. real-time indexing and instant search
- Remaining challenges are not computational, but in modeling user cognition
  - Need a deeper understanding of the query, the content and/or the world at large
  - *Could Watson explain why the answer is Toronto?*

What it’s like to be a machine?

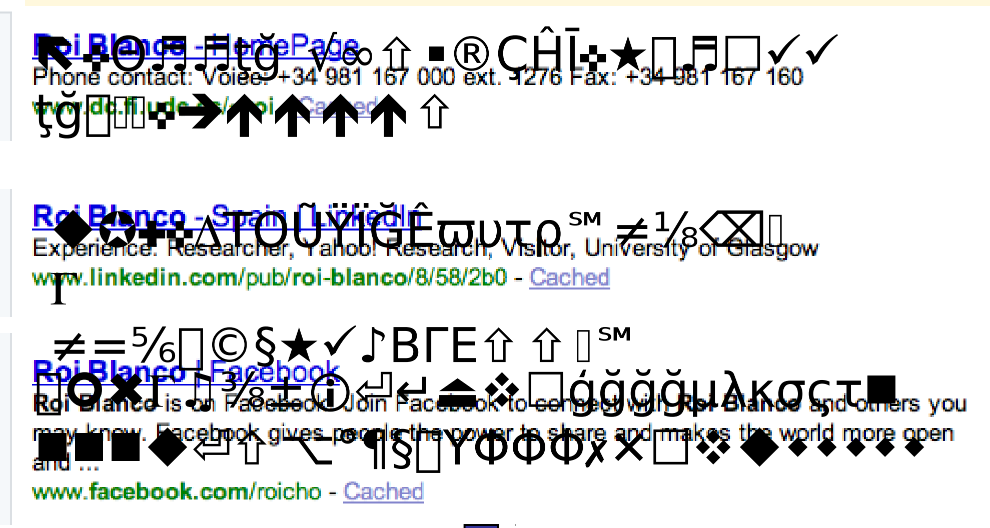


Roi Blanco - HomePage
Phone contact: Voice: +34 981 167 000 ext. 1276 Fax: +34 981 167 160
www.dc.fi.udc.es/~roi - Cached

Roi Blanco - Spain | LinkedIn
Experience: Researcher, Yahoo! Research; Visitor, University of Glasgow
www.linkedin.com/pub/roi-blanco/8/58/2b0 - Cached

Roi Blanco | Facebook
Roi Blanco is on Facebook. Join Facebook to connect with **Roi Blanco** and others you may know. Facebook gives people the power to share and makes the world more open and ...
www.facebook.com/roicho - Cached

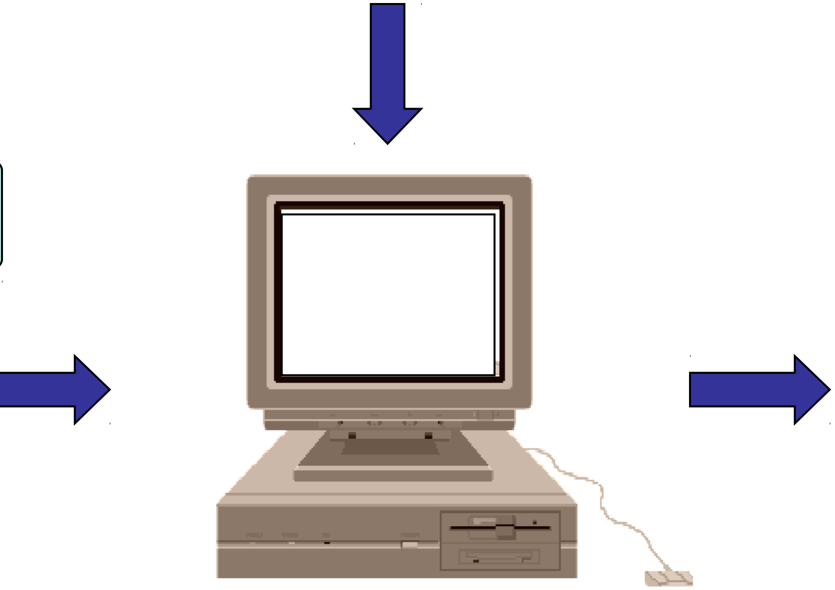
What it's like to be a machine?



Roi Blanco - HomePage
Phone contact: Voice: +34 981 167 000 ext. 1276 Fax: +34 981 167 160
www.dc.fi.udc.es/~roi - Cached

Roi Blanco - Spain | LinkedIn
Experience: Researcher, Yahoo! Research, Visitor, University of Glasgow
www.linkedin.com/pub/roi-blanco/8/58/2b0 - Cached

Roi Blanco | Facebook
Roi Blanco is on Facebook. Join Facebook to connect with Roi Blanco and others you may know. Facebook gives people the power to share and makes the world more open and ...
www.facebook.com/roicho - Cached

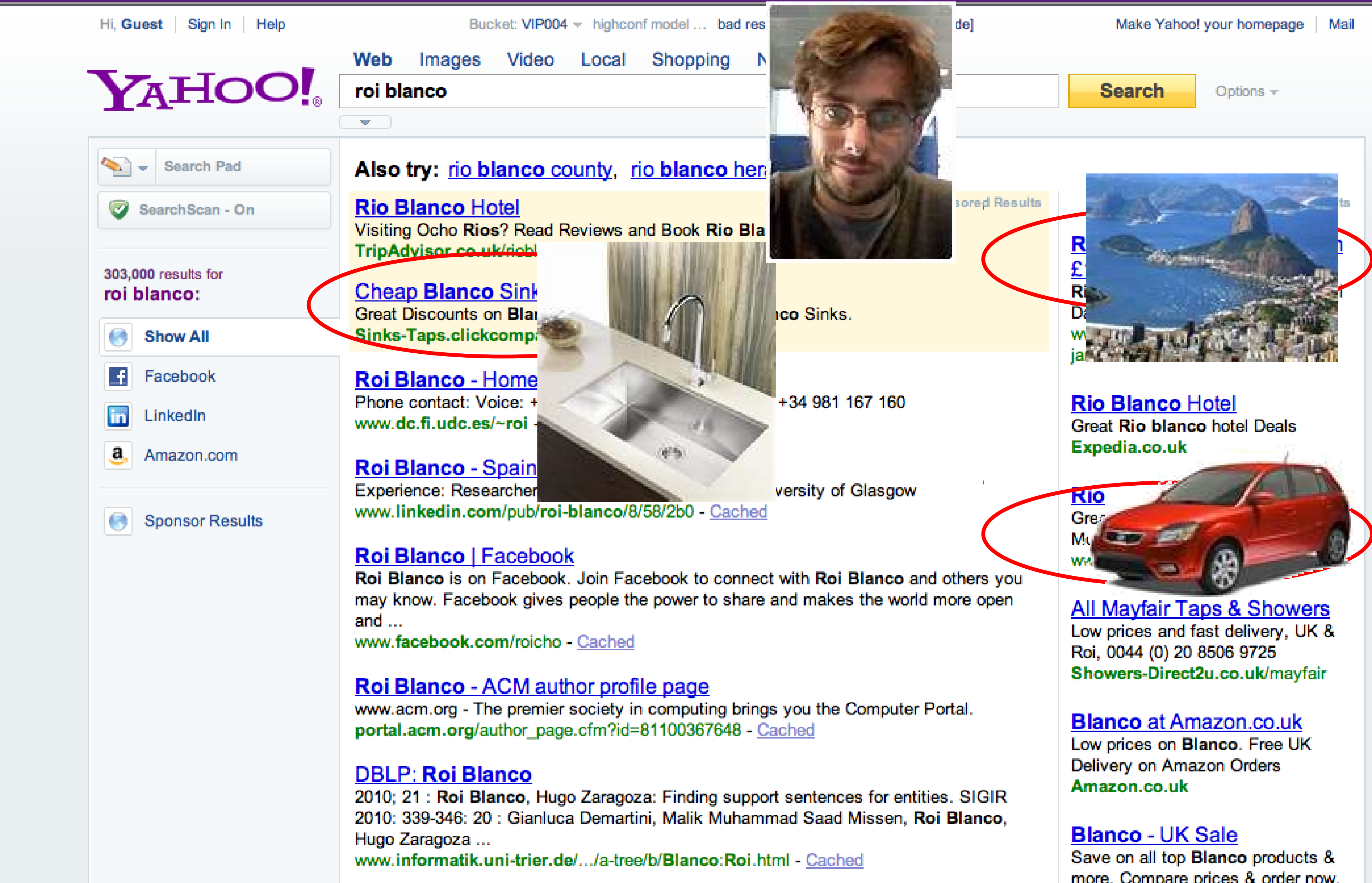


Poorly solved information needs

- Multiple interpretations
  - paris hilton
- Long tail queries
  - george bush (and I mean the beer brewer in
- Multimedia search
  - paris hilton sexy
- Imprecise or overly precise searches
  - jim hendler
  - pictures of strong adventures people
- Searches for descriptions
  - countries in africa
  - 32 year old computer scientist living in barcelona
  - reliable digital camera under 300 dollars

Many of these queries would not be asked by users, who learned over time what search technology can and can not do.

Ambiguity



Hi, **Guest** | Sign In | Help

Bucket: VIP004 ▾ highconf model ... bad res de]

Make Yahoo! your homepage | Mail

Web Images Video Local Shopping N

roi blanco

Search

Options ▾

Search Pad

SearchScan - On

303,000 results for **roi blanco:**

Show All

Facebook

LinkedIn

Amazon.com

Sponsor Results

Also try: rio blanco county, rio blanco her

sored Results

Rio Blanco Hotel
Visiting Ocho **Rios**? Read Reviews and Book **Rio Bla**
TripAdvisor.co.uk/riobl

Cheap **Blanco** Sink
Great Discounts on **Bla** co Sinks.
Sinks-Taps.clickcompa

Roi Blanco - Home
Phone contact: Voice: + +34 981 167 160
www.dc.fi.udc.es/~roi

Roi Blanco - Spain
Experience: Researcher versity of Glasgow
www.linkedin.com/pub/**roi-blanco**/8/58/2b0 - Cached

Roi Blanco | Facebook
Roi Blanco is on Facebook. Join Facebook to connect with **Roi Blanco** and others you may know. Facebook gives people the power to share and makes the world more open and ...
www.**facebook.com**/roicho - Cached

Roi Blanco - ACM author profile page
www.acm.org - The premier society in computing brings you the Computer Portal.
portal.acm.org/author_page.cfm?id=81100367648 - Cached

DBLP: **Roi Blanco**
2010; 21 : **Roi Blanco**, Hugo Zaragoza: Finding support sentences for entities. SIGIR 2010: 339-346: 20 : Gianluca Demartini, Malik Muhammad Saad Missen, **Roi Blanco**, Hugo Zaragoza ...
www.**informatik.uni-trier.de**/.../a-tree/b/**Blanco**:**Roi**.html - Cached

Rio Blanco Hotel
Great **Rio blanco** hotel Deals
Expedia.co.uk

Rio

All Mayfair Taps & Showers
Low prices and fast delivery, UK & Roi, 0044 (0) 20 8506 9725
Showers-Direct2u.co.uk/mayfair

Blanco at Amazon.co.uk
Low prices on **Blanco**. Free UK Delivery on Amazon Orders
Amazon.co.uk

Blanco - UK Sale
Save on all top **Blanco** products & more. Compare prices & order now.

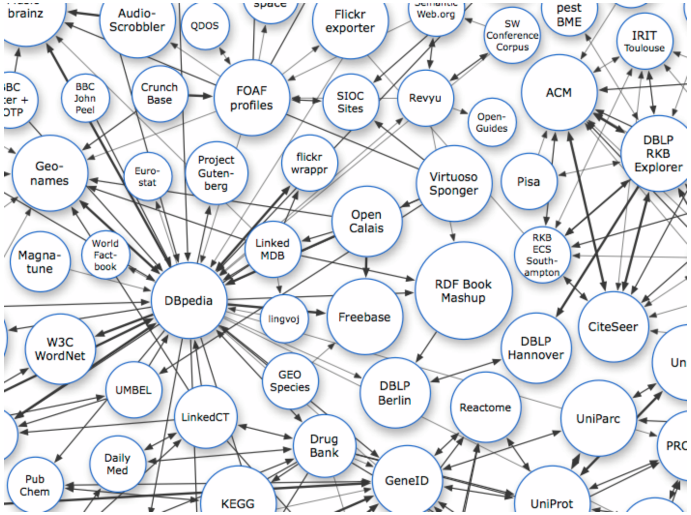
Why Semantic Search? Part II

- The Semantic Web is here
  - Data
    - Large amounts of RDF data
    - Heterogeneous schemas
    - Diverse quality
  - End users
    - Not skilled in writing complex queries (e.g. SPARQL)
    - Not familiar with the data
- Novel applications
  - Complementing document search
    - Rich Snippets, related entities, direct answers
  - Other novel search tasks

Searching over 10,000 ontologies

Semantic Web data

- Linked Data
  - Data published as RDF documents linked to other RDF documents and/or using SPARQL end-points
  - Community effort to re-publish large public datasets (e.g. Dbpedia, open government data)



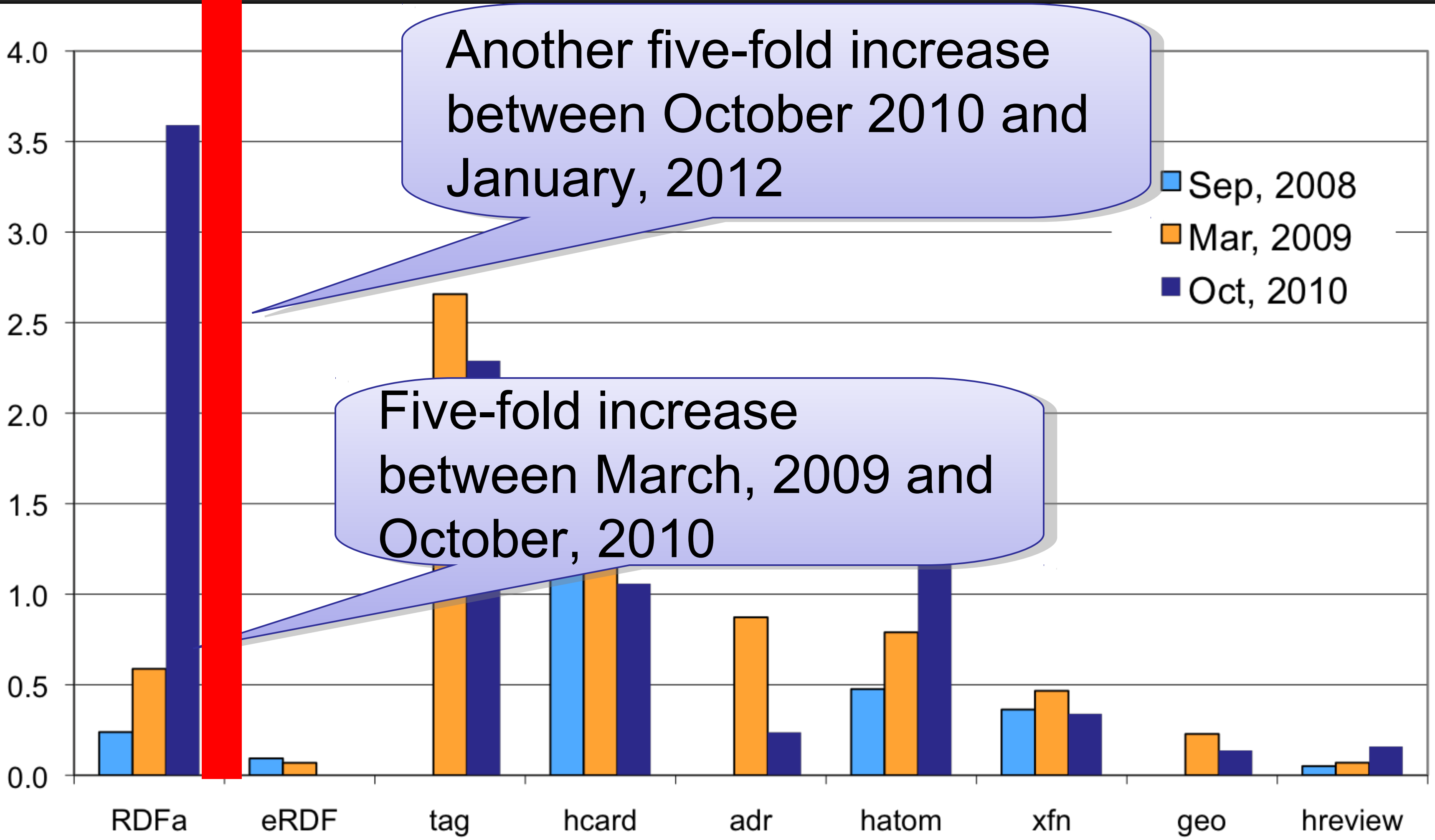
- RDFa
  - Data embedded inside HTML pages
  - Recommended for site owners by Yahoo, Google, Facebook

RDFa example: Facebook’s Open Graph Protocol

- RDF vocabulary to be used in conjunction with RDFa
  - Simplify the work of developers by restricting the freedom in RDFa
- Activities, Businesses, Groups, Organizations, People, Places, Products and Entertainment
- Only HTML <head> accepted
- http://opengraphprotocol.org/

```
<html xmlns:og="http://opengraphprotocol.org/schema/">
<head>
    <title>The Rock (1996)</title>
    <meta property="og:title" content="The Rock" />
    <meta property="og:type" content="movie" />
    <meta property="og:url"
    content="http://www.imdb.com/title/tt0117500/" />
    <meta property="og:image" content="http://ia.media-
    imdb.com/images/rock.jpg" /> …
</head> ...
```

Exponent[redacted]l growth in RDFa data

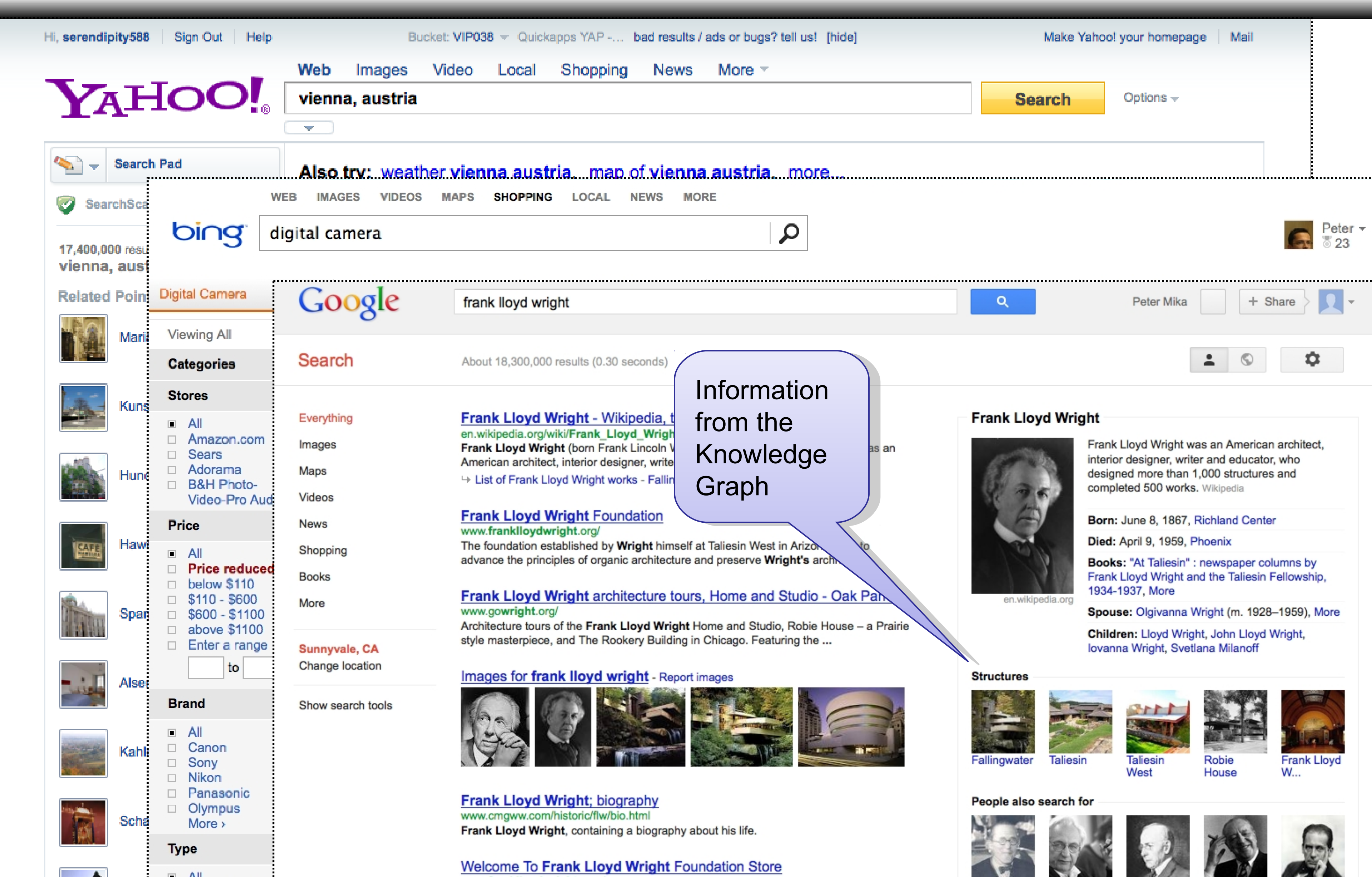


Another five-fold increase between October 2010 and January, 2012

Five-fold increase between March, 2009 and October, 2010

Percentage of URLs with embedded metadata in various formats

Application: direct answers and entity suggestions



Hi, serendipity588 | Sign Out | Help   Bucket: VIP038 ▾ Quickapps YAP -... bad results / ads or bugs? tell us! [hide]   Make Yahoo! your homepage | Mail

Web Images Video Local Shopping News More ▾

vienna, austria

Search   Options ▾

Search Pad

Also try: weather vienna austria, map of vienna austria, more...

SearchSca

17,400,000 resu

vienna, aus

Related Poin

Mari

Kuns

Hun

Haw

Spa

Alse

Kahl

Sch

WEB IMAGES VIDEOS MAPS SHOPPING LOCAL NEWS MORE

digital camera

Peter ▾ 23

Digital Camera

Viewing All

Categories

Stores

- ■ All
- □ Amazon.com
- □ Sears
- □ Adorama
- □ B&H Photo-Video-Pro Aud

Price

- ■ All
- □ Price reduced
- □ below $110
- □ $110 - $600
- □ $600 - $1100
- □ above $1100
- □ Enter a range

□ to □

Brand

- ■ All
- □ Canon
- □ Sony
- □ Nikon
- □ Panasonic
- □ Olympus

More ›

Type

frank lloyd wright

Peter Mika + Share

Search   About 18,300,000 results (0.30 seconds)

Everything
Images
Maps
Videos
News
Shopping
Books
More

Sunnyvale, CA
Change location

Show search tools

Frank Lloyd Wright - Wikipedia, t
en.wikipedia.org/wiki/Frank_Lloyd_Wrigh
Frank Lloyd Wright (born Frank Lincoln W ... as an American architect, interior designer, write
↳ List of Frank Lloyd Wright works - Fallin

Frank Lloyd Wright Foundation
www.franklloydwright.org/
The foundation established by Wright himself at Taliesin West in Arizo ... to advance the principles of organic architecture and preserve Wright's arch

Frank Lloyd Wright architecture tours, Home and Studio - Oak Par
www.gowright.org/
Architecture tours of the Frank Lloyd Wright Home and Studio, Robie House – a Prairie style masterpiece, and The Rookery Building in Chicago. Featuring the ...

Images for frank lloyd wright - Report images

Frank Lloyd Wright; biography
www.cmgwm.com/historic/flw/bio.html
Frank Lloyd Wright, containing a biography about his life.

Welcome To Frank Lloyd Wright Foundation Store

Information from the Knowledge Graph

Frank Lloyd Wright

Frank Lloyd Wright was an American architect, interior designer, writer and educator, who designed more than 1,000 structures and completed 500 works. Wikipedia

en.wikipedia.org

Born: June 8, 1867, Richland Center

Died: April 9, 1959, Phoenix

Books: "At Taliesin" : newspaper columns by Frank Lloyd Wright and the Taliesin Fellowship, 1934-1937, More

Spouse: Olgivanna Wright (m. 1928–1959), More

Children: Lloyd Wright, John Lloyd Wright, Iovanna Wright, Svetlana Milanoff

Structures

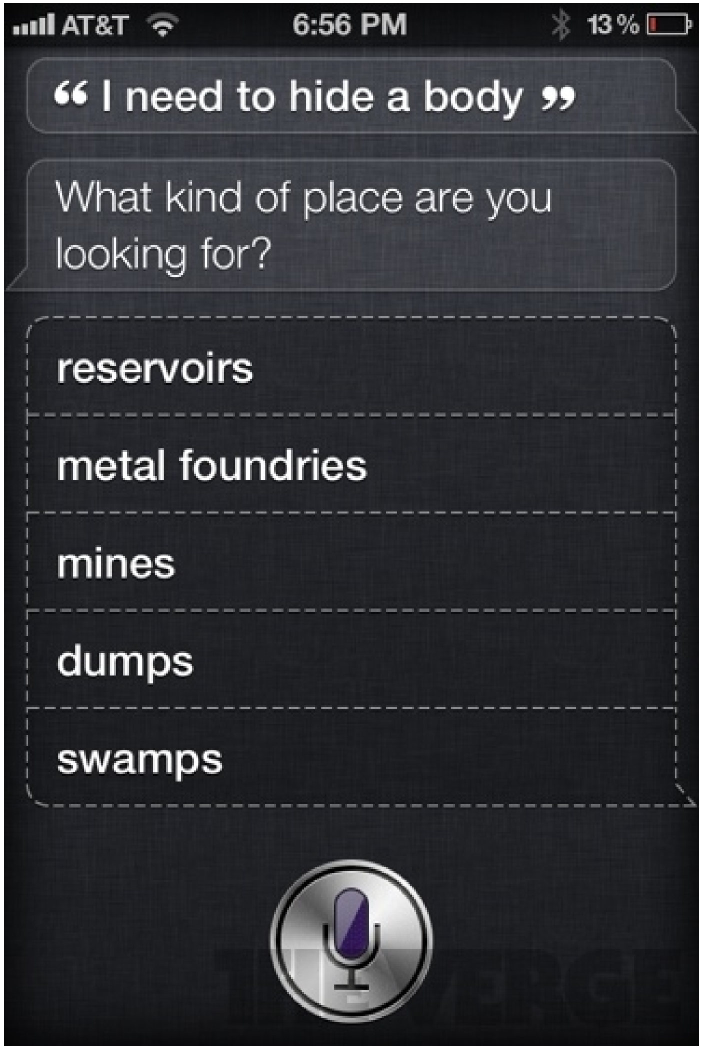
Fallingwater Taliesin Taliesin West Robie House Frank Lloyd W...

People also search for

Other novel applications

- Aggregation of search results
  - e.g. price comparison across websites
- Analysis and prediction
  - e.g. world temperature by 2020
- Semantic profiling
  - Ontology-based modeling of user interests
- Semantic log analysis
  - Linking query and navigation logs to ontologies
- Task completion
  - e.g. booking a vacation using a combination of services
- Conversational search
  - e.g. PARLANCE EU FP7 project

Interactive search and task completion



“ I need to hide a body ”

What kind of place are you looking for?

- reservoirs
- metal foundries
- mines
- dumps
- swamps

Ok, I can send a text to Ka Egan for you…what would you like it to say?

“ Be there in 30 minutes

Ok, I'll send your message

To: Kate Egan

Be there in 30 minute

“ On May 19 remind me it's dad's birthday ”

Here's your reminder for May 19, 2012 at 9 am:

19 Saturday May 2012

□ Dad's birthday 5/19/12

ll it be sunny this eekend in Miami ”

your sunglasses, it's to be nice in Miami.

Weekly Forecast

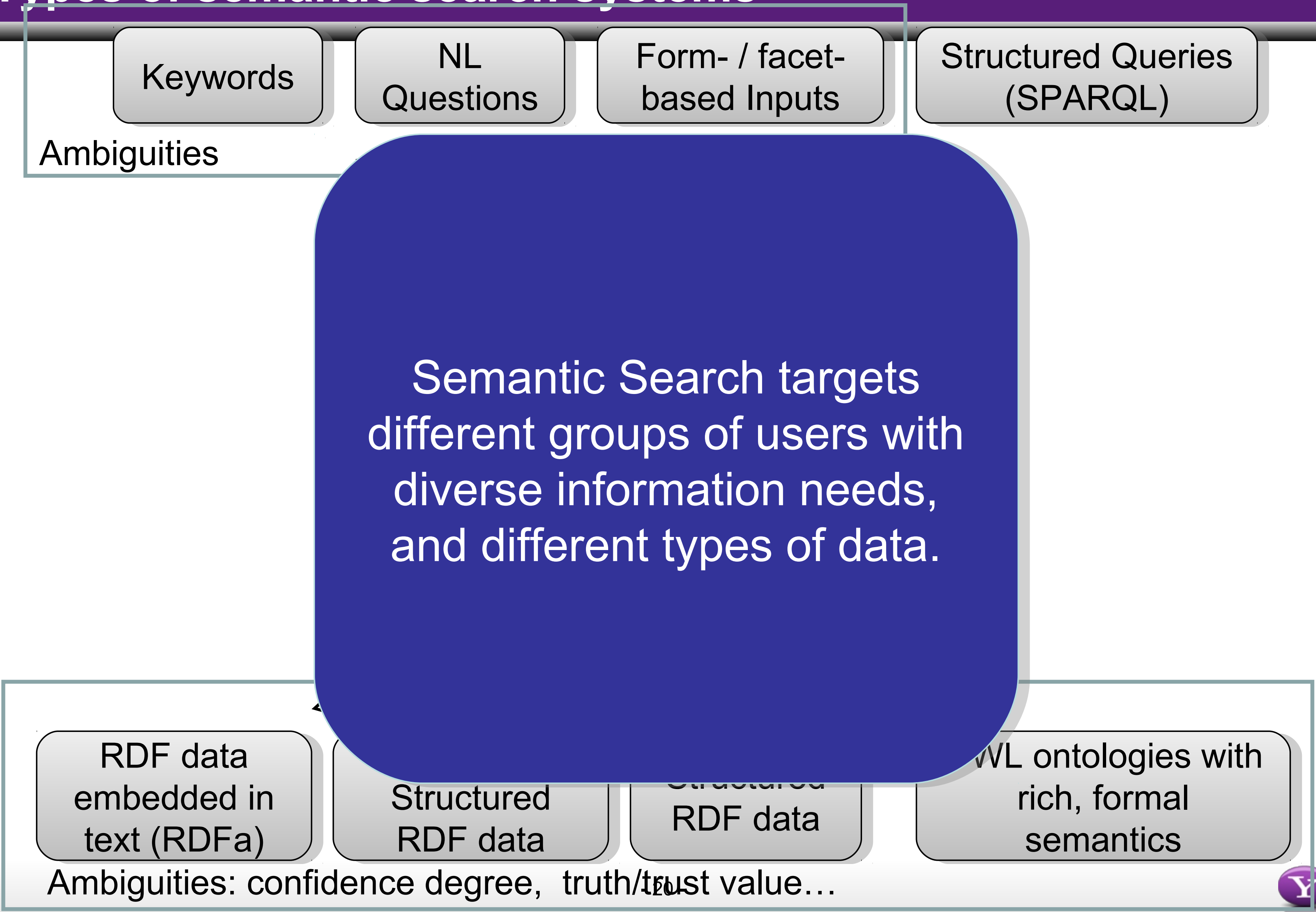
WED	THU	FRI	SAT	SUN
97°	97°	93°	93°	93
63°	63°	61°	61	61

Semantic Search

Semantic Search: a definition

- Semantic search is a retrieval paradigm that
  - **Makes use of the structure of the data or explicit schemas to understand user intent and the meaning of content**
  - **Exploits this understanding at some part of the search process**
- Emerging field of research
  - Exploiting Semantic Annotations in Information Retrieval (2008-2012)
  - Semantic Search (SemSearch) workshop series (2008-2011)
  - Entity-oriented search workshop (2010-2011)
  - Joint Intl. Workshop on Semantic and Entity-oriented Search (2012)
  - SIGIR 2012 tracks on Structured Data and Entities
- Related fields:
  - XML retrieval, Keyword search in databases, NL retrieval

Types of semantic search systems



Keywords

NL Questions

Form- / facet-based Inputs

Structured Queries (SPARQL)

Ambiguities

Semantic Search targets different groups of users with diverse information needs, and different types of data.

RDF data embedded in text (RDFa)

Structured RDF data

Structured RDF data

WL ontologies with rich, formal semantics

Ambiguities: confidence degree, truth/trust value…

Taxonomy of evaluation

- What is being measured?
  - Goal
- What granularity? How much?
  - Scale
- How is it evaluated? Who is doing the evaluation?
  - Methodology
- Who is participating?
  - Scope
- What is shared at the end?
  - Outcomes

Goal: what is being measured?

- Efficiency: how fast?
  - Performance of the system
  - Time spent by the user
- Effectiveness: how good?
  - Relevance
  - Freshness
  - Diversity
  - …

Scale: At what level? How much?

- Granularity
  - Individual results
    - *Can not capture e.g. diversity of results*
  - Complete or partial result sets
    - Side-by-side (SBS) comparison
    - User modeling
      - *Assuming a model of user behaviour*
  - Task-based evaluation
    - User-defined or pre-defined tasks
    - *Example: PARLANCE evaluation*
  - (Online) usage testing
    - Bucket testing
    - Historical analysis
- Size
  - Number of items to be evaluated
  - Typically a **very small** subset of potential inputs

Methodology: how it is evaluated?

- Method of collecting judgments
  - Explicit feedback
    - Relevance assessments on a binary or multi-valued scale, for example Perfect/Excellent/Good/Fair/Bad (PEGFB)
  - Implicit feedback
    - *e.g. result clicks, long-dwell, abandonment*
- Subjects
  - Experts
    - *Examples: internal search editorial team at Yahoo! or retired intelligence analysts used by NIST.*
  - Users
  - Crowd-sourcing
    - *Not necessarily experts, nor users*
- Setting
  - Lab vs. natural setting

Methodology: confounding factors

- **Rendering of results**
  - Result may be perceived better just because of rendering
    - e.g. quality of text snippets or images
  - Relevance before and after clicking
    - Perceived vs. actual relevance
    - Good clicks vs. bad clicks: a bad click is a click on an irrelevant result, e.g. because the result seemed relevant or did not contain enough information
  - Results in the context of other results
    - Eye tracking studies show users inspect results near other relevant results
  - Position bias
    - User starts reading from top of results, and get progressively more tired
    - Pagination: result #11 is clicked much less than Result #10
  - Results in the context of other elements in the SERP (search engine result page)
- **Efficiency**
  - Users may conflate efficiency with effectiveness. A faster search engine is perceived to be better

Scope: who is participating?

- Public vs. private
  - Public competitions
    - Shared task, radically different solutions
    - *Ideally, participants should represent the state-of-the-art*
  - Private, internal evaluation
    - *Comparing different versions of the same system, e.g. for parameter tuning, measuring the effects of data quality etc.*
  - Comparison to previous state-of-the-art or a strong baseline
    - *Typical for one-off academic papers*

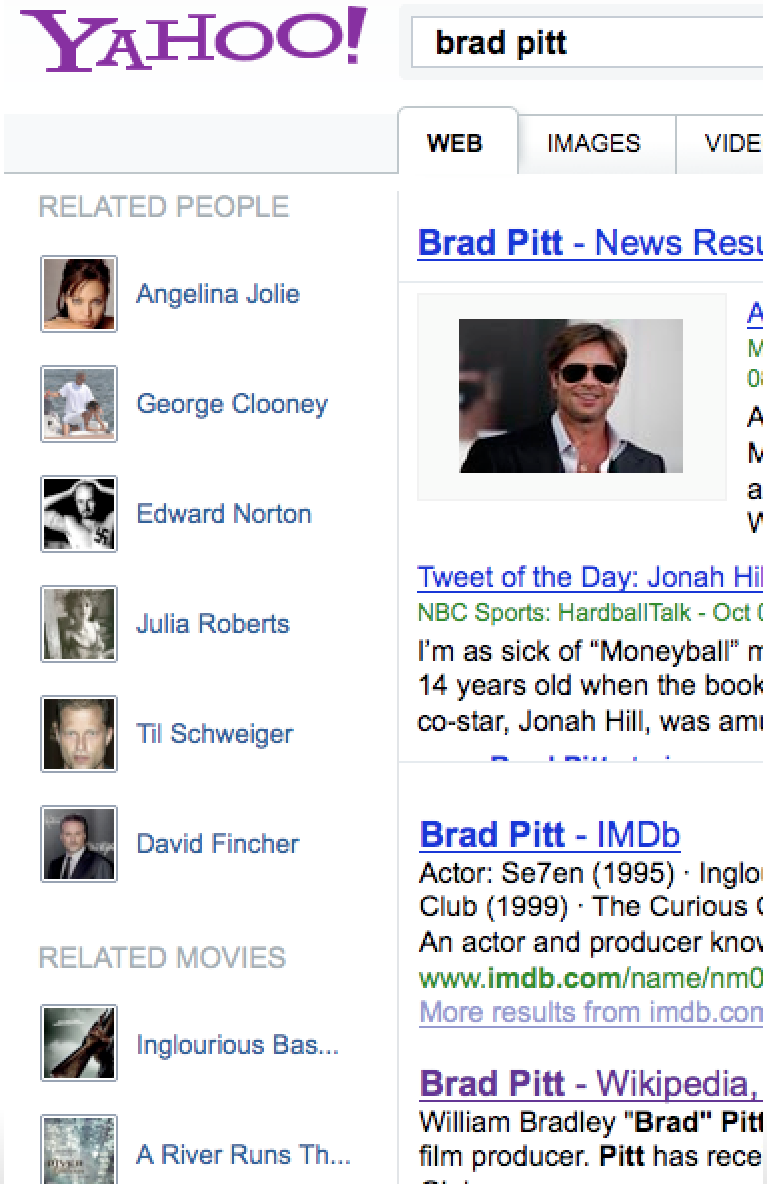
What is the form of the evaluation result?

- Evaluation guidelines
  - Set of guidelines developed before and typically refined during the evaluation
- Evaluation data
  - Set of queries, data and relevance assessments
  - Information on how the data was produced, e.g. what was the agreement
  - Code for working with the data and computing metrics
- Rankings of systems
  - By one or more metric
- Complete evaluation solution
  - Evaluation system capable to assess new submissions containing new results

Related entity suggestions in web search

Related entity suggestions

- Task: given a keyword query suggest a ranked list of related entities
  - For presentation purposes, the results are grouped by type
  - The order of groups and the minimal number of items per group is fixed
- Private evaluation
  - Novel task, no baselines
  - Setting is too specific, task has many subtasks
  - Proprietary data
    - Knowledge Graph
    - Usage data
      - Flickr, Twitter, query logs



RELATED PEOPLE

Angelina Jolie

George Clooney

Edward Norton

Julia Roberts

Til Schweiger

David Fincher

RELATED MOVIES

Inglourious Bas...

A River Runs Th...

Brad Pitt - News Resu

Tweet of the Day: Jonah Hil

NBC Sports: HardballTalk - Oct

I'm as sick of "Moneyball" m
14 years old when the book
co-star, Jonah Hill, was amu

Brad Pitt - IMDb

Actor: Se7en (1995) · Inglo
Club (1999) · The Curious C
An actor and producer know
www.imdb.com/name/nm0
More results from imdb.com

Brad Pitt - Wikipedia,

William Bradley "Brad" Pitt
film producer. Pitt has rece

Evaluation using experts

- Relevance assessment using experts
  - Yahoo! Search editorial team
  - Size of the data, agreement among judges are reported
  - PEGFB scale
  - Metrics:

$$DCG@p = gain_1 + \sum_{i=0}^{p} \frac{gain_i}{log_2(i)}$$

$$nDCG@p = \frac{DCG_{run}@p}{DCG_{ideal}@p}$$

  - *van Zwol et al.: Faceted exploration of image search results. WWW 2010: 961-970*
  - *Kang et al.: Ranking related entities for web search queries. WWW (Companion Volume) 2011: 67-68*

Evaluation based on usage data

- Implicit relevance assessment using usage logs
  - Clicks turned into labels or preferences
  - Size of the data is not a concern
  - Agreement can not be computed
  - Metrics
    - nDCG
    - Gains are computed from normalized CTR/COEC

$$ctr_{e,f} = \frac{clicks_{e,f}}{views_{e,f}} \qquad coec_{e,f} = \frac{clicks_{e,f}}{\sum_{p=1}^{P} views_{e,f_p} \times ctr_p}$$

van Zwol et al. Ranking Entity Facets Based on User Click Feedback. ICSC 2010: 192-199.

Side-by-side testing

- Comparing two systems
  - A/B comparison, e.g. current system under development and production system
  - Scale: A is better, B is better
- Separate tests for relevance and image quality
  - Image quality can significantly influence user perceptions
  - Images can violate safe search rules
- Classification of errors
  - Results: missing important results/contains irrelevant results, too few results, entities are not fresh, more/less diverse, should not have triggered
  - Images: bad photo choice, blurry, group shots, nude/racy etc.
- Notes
  - *Borderline, set one entities relate to the movie Psy but the query is most likely about Gangnam style*
  - *Blondie and Mickey Gilley are 70's performers and do not belong on a list of 60's musicians.*
  - *There is absolutely no relation between Finland and California.*

Bucket testing

- Also called online evaluation
  - Comparing against baseline version of the system
  - Baseline does not change during the test
- Small % of search traffic redirected to test system, another small % to the baseline system
- Data collection over at least a week, looking for stat. significant differences that are also stable over time
- Metrics in web search
  - Searches per browser-cookie (SPBC)
  - Other key metrics should not impacted negatively, e.g. Abandonment and retry rate, Daily Active Users (DAU), Revenue Per Search (RPS), etc.

Semantic Search Challenge 2010/2011

Harry Halpin, Daniel Herzig, Peter Mika, Jeff Pound,
Henry Thompson, Roi Blanco, Thanh Tran Duc

Evaluation campaigns in IR

- Long tradition of comparative system evaluations in IR
  - Cranfield studies
  - Fixed queries (topics) data sets (corpus) and assessments
- Public evaluations
  - Evaluating the results submitted by the participants
    - “Pooling”: the results that are returned by multiple systems need to be evaluated only once
  - Expert judgments
  - Standard metrics
    - TRECEval software commonly used for computing metrics
- TREC, CLEF, INEX
  - Multiple tracks (~tasks) at each competition
  - Organized yearly, with some tasks repeating to measure progress
  - Publicly funded evaluations, data made publicly available

Semantic Search evaluation campaigns

- Limited interest in Semantic Search prior to 2010
  - Small scale datasets in Semantic Web
  - No heterogeneous web data
  - Focus on expert users who can
    - Formulate SPARQL queries
    - Familiar with the ontology of the data
  - Benchmarking
    - Lehigh University Benchmark (LUBM)
    - Berlin SPARQL Benchmark (BSBM)
    - Most recent: Linked Data Benchmark Council (LDBC)
- Evaluation campaigns starting from 2010
  - SemSearch Challenge 2010/2011
  - TREC Entity Track 2011/2012
  - Question Answering over Linked Data 2011/2012/2013

SemSearch Challenge 2011: Entity Search Track

- Entity Search
  - retrieval of data related to a single entity
- Queries
  - Selected from the Search Query Tiny Sample v1.0 dataset, provided by the Yahoo! Webscope program
  - Real web search queries sampled from the US query log of January, 2009
  - Queries asked by at least three different users and with long number sequence removed (privacy reasons)
  - 50 selected queries that name an entity explicitly (but may also provide context)
  - Last year: same type of queries, but a mix of Microsoft and Yahoo! Logs

SemSearch Challenge 2011: List query track

- List queries
  - Queries that describe a set of entities
  - The answer is a closed set
  - Relatively small number of possible answers
  - The answer is not likely to change
- Hand-picked but not hand-written
  - Yahoo! Search logs
    - Queries from the Tiny Sample v1.0 dataset
    - Queries with clicks on Wikipedia
  - TrueKnowledge
    - Recent queries

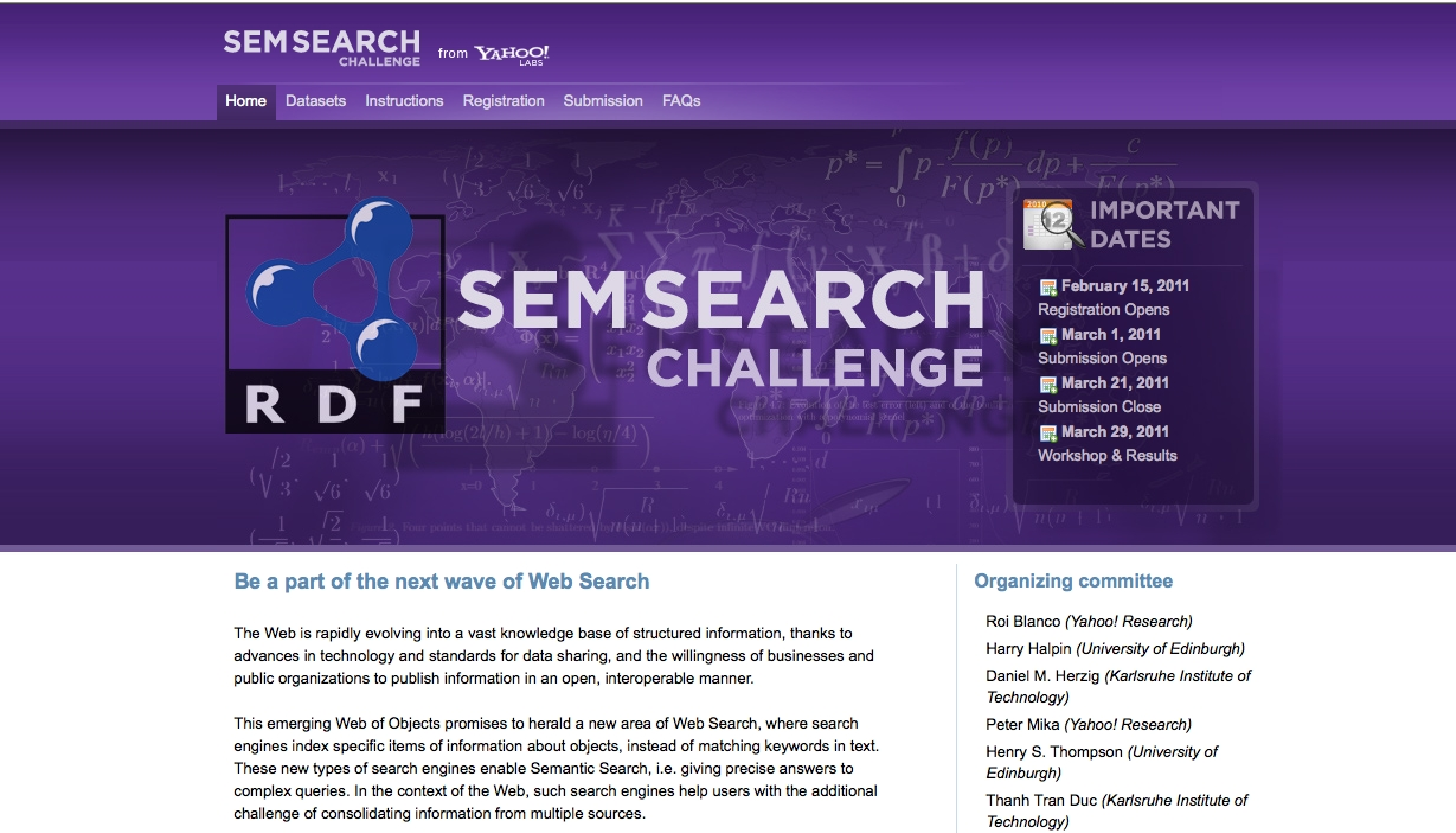
Data set

- Same as Billion Triples Challenge 2009 data set
  - Blank nodes are encoded as URIs
- A data set combining crawls of multiple semantic search engines
  - doesn’t necessarily match the current state of the Web
  - doesn’t necessarily match the coverage of any particular search engine
- Final dataset

Collecting the results

- Submissions via semsearch.yahoo.com
  - max. 3 submissions per team per track
- Pooling of results
  - Top 20 results are evaluated
  - Despite validation, still problems
    - e.g. N-Triples encoded URIs, lowercased URIs
- Collecting triples for each result
  - All triples where the URI is the subject
  - Discarded URIs that didn’t appear as subject
- Rendering result display
  - Values are clipped at 300 chars (last # or / for object-properties)
  - RDF built-ins shown first
  - Preference to English language values

semsearch.yahoo.com



February 15, 2011
Registration Opens

March 1, 2011
Submission Opens

March 21, 2011
Submission Close

March 29, 2011
Workshop & Results

Be a part of the next wave of Web Search

The Web is rapidly evolving into a vast knowledge base of structured information, thanks to advances in technology and standards for data sharing, and the willingness of businesses and public organizations to publish information in an open, interoperable manner.

This emerging Web of Objects promises to herald a new area of Web Search, where search engines index specific items of information about objects, instead of matching keywords in text. These new types of search engines enable Semantic Search, i.e. giving precise answers to complex queries. In the context of the Web, such search engines help users with the additional challenge of consolidating information from multiple sources.

Organizing committee

Roi Blanco *(Yahoo! Research)*

Harry Halpin *(University of Edinburgh)*

Daniel M. Herzig *(Karlsruhe Institute of Technology)*

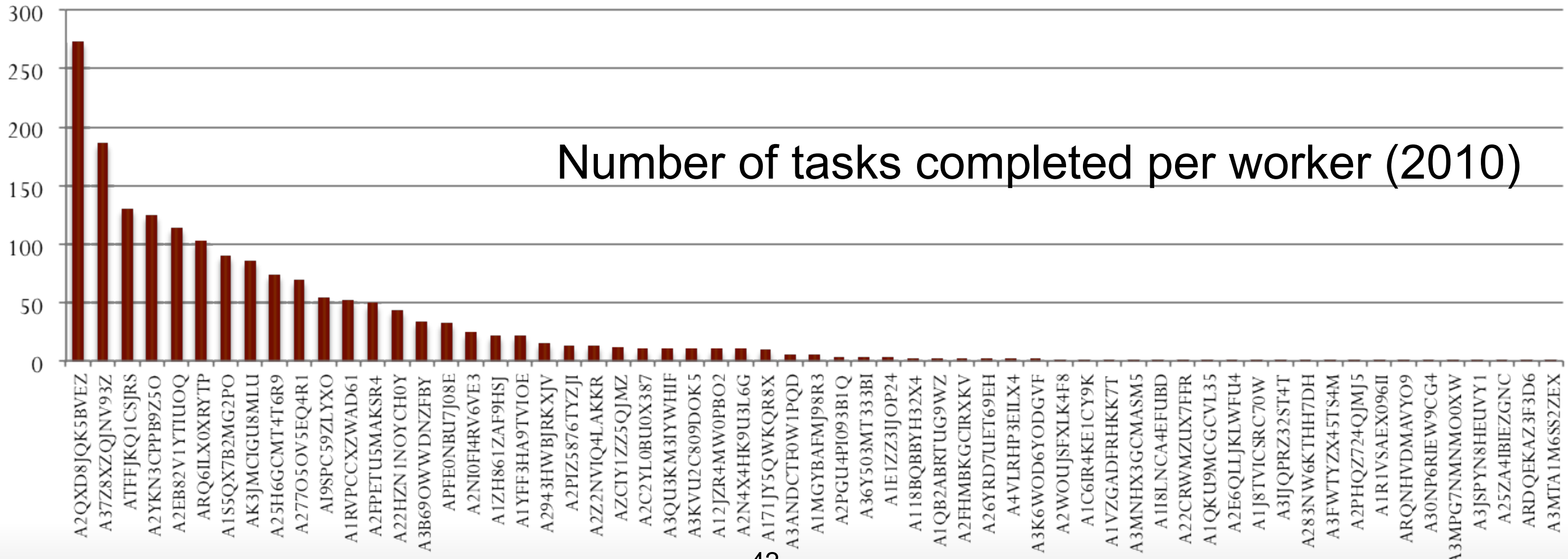
Peter Mika *(Yahoo! Research)*

Henry S. Thompson *(University of Edinburgh)*

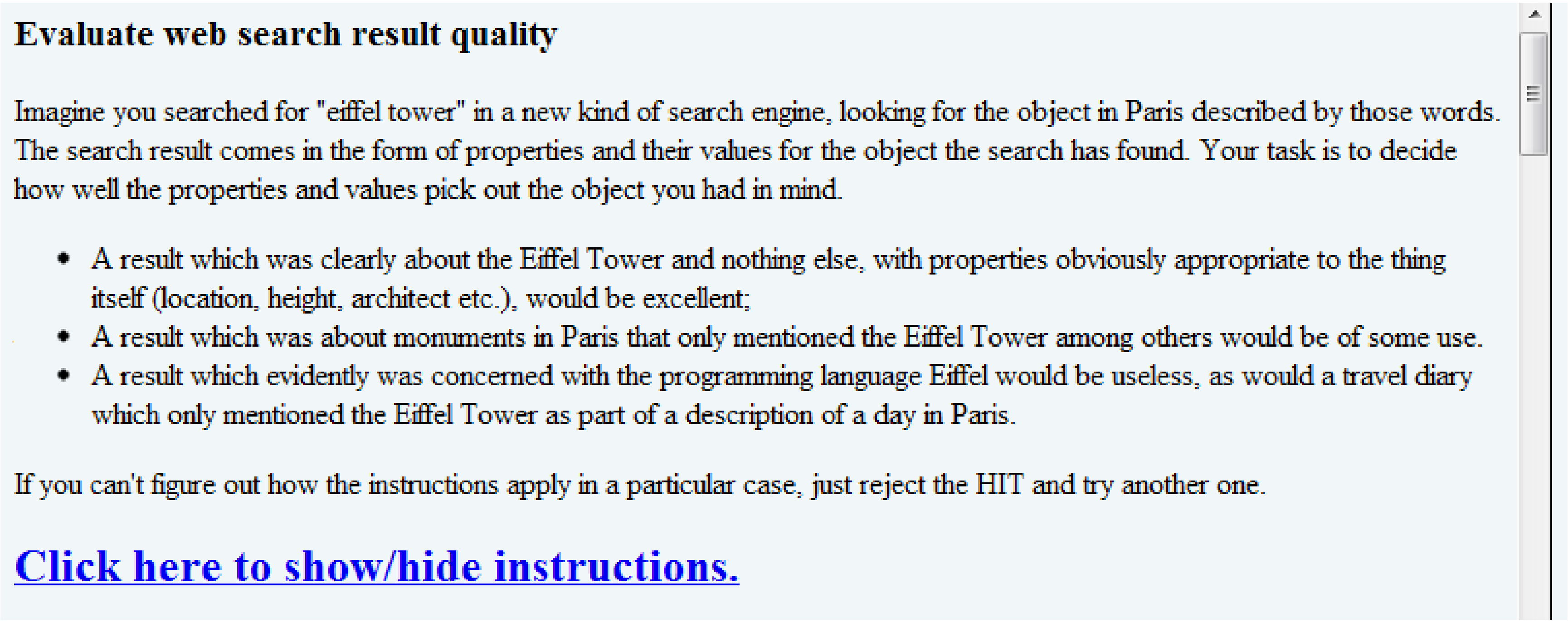
Thanh Tran Duc *(Karlsruhe Institute of Technology)*

Assessment with Amazon Mechanical Turk

- Evaluation using non-expert judges
  - Paid $0.2 per 12 results
    - Typically done in 1-2 minutes (~ $6-$12 an hour)
    - Sponsored by the European SEALS project
  - Each result is evaluated by 5 workers
- *Blanco et al. Repeatable and Reliable Search System Evaluation using Crowd-Sourcing, SIGIR2011*


Number of tasks completed per worker (2010)

Evaluation form



Evaluate web search result quality

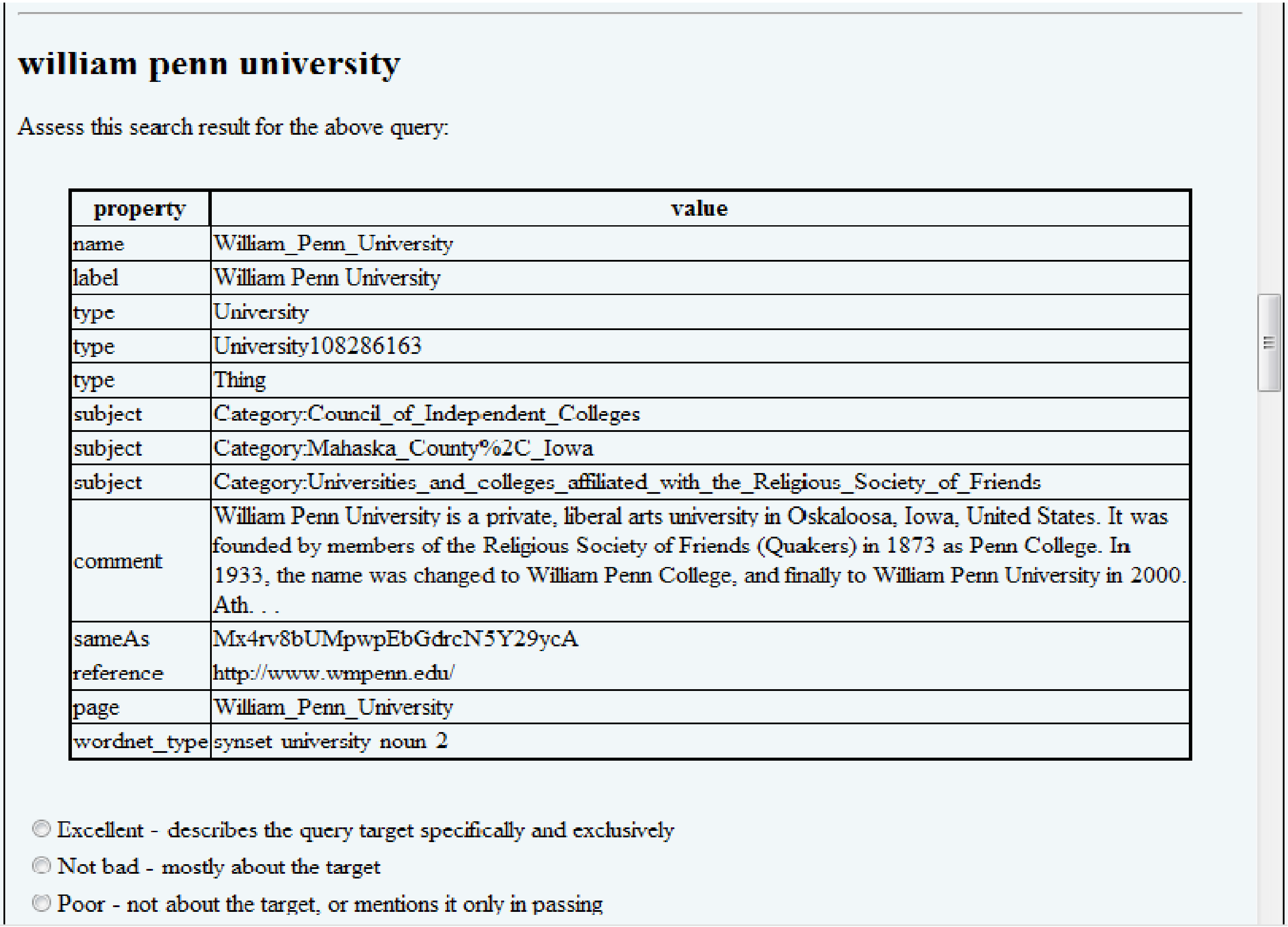
Imagine you searched for "eiffel tower" in a new kind of search engine, looking for the object in Paris described by those words. The search result comes in the form of properties and their values for the object the search has found. Your task is to decide how well the properties and values pick out the object you had in mind.

- A result which was clearly about the Eiffel Tower and nothing else, with properties obviously appropriate to the thing itself (location, height, architect etc.), would be excellent;
- A result which was about monuments in Paris that only mentioned the Eiffel Tower among others would be of some use.
- A result which evidently was concerned with the programming language Eiffel would be useless, as would a travel diary which only mentioned the Eiffel Tower as part of a description of a day in Paris.

If you can't figure out how the instructions apply in a particular case, just reject the HIT and try another one.

Click here to show/hide instructions.

Evaluation form



william penn university

Assess this search result for the above query:

property	value
name	William_Penn_University
label	William Penn University
type	University
type	University108286163
type	Thing
subject	Category:Council_of_Independent_Colleges
subject	Category:Mahaska_County%2C_Iowa
subject	Category:Universities_and_colleges_affiliated_with_the_Religious_Society_of_Friends
comment	William Penn University is a private, liberal arts university in Oskaloosa, Iowa, United States. It was founded by members of the Religious Society of Friends (Quakers) in 1873 as Penn College. In 1933, the name was changed to William Penn College, and finally to William Penn University in 2000. Ath. . .
sameAs	Mx4rv8bUMpwpEbGdrcN5Y29ycA
reference	http://www.wmpenn.edu/
page	William_Penn_University
wordnet_type	synset university noun 2

○ Excellent - describes the query target specifically and exclusively

○ Not bad - mostly about the target

○ Poor - not about the target, or mentions it only in passing

Catching the bad guys

- Payment can be rejected for workers who try to game the system
  - An explanation is commonly expected, though cheaters rarely complain
- We opted to mix control questions into the real results
  - Gold-win cases that are known to be perfect
  - Gold-loose cases that are known to be bad
- Metrics
  - Avg. and std. dev on gold-win and gold-loose results

Worker	Known bad		Real		Known Good		Total N	Time to complete (sec)
	N	Mean	N	Mean	N	Mean		
badguy	20	2.556	200	2.738	20	2.684	240	29.6
goodguy	13	1	130	2.038	13	3	156	95
whoknows	1	1	21	1.571	2	3	24	83.5

Results

- See workshop reports
  - Halpin et al. Evaluating ad-hoc object retrieval. IWEST 2010 workshop proceedings. PDF
  - Blanco et al. Entity search evaluation over structured web data. EOS 2011 workshop proceedings. PDF

Other evaluations: entity search in documents

- Web People Search (WePS) 2008-2010
  - WePS-1: name ambiguity problem: cluster web search results for a given person name based on whether they belong to the same person
  - WePS-2: same task + attribute extraction task
  - WePS-3:
    - Task 1: combination: extract attributes and cluster
    - Task 2: name ambiguity resolution in Twitter data
- TREC Entity Track
  - Related Entity Finding
    - Entities related to a given entity through a particular relationship
    - Retrieval over documents (ClueWeb 09 collection)
    - Example: (Homepages of) airlines that fly Boeing 747
  - Entity List Completion
    - Given some elements of a list of entities, complete the list

Other evaluations: Q & A

- Question Answering over Linked Data
  - First two editions at ESWC 2011 and 2012
  - Data
    - Dbpedia and MusicBrainz in RDF
  - Questions
    - Full natural language questions of different forms, written by the organizers
    - *Give me all actors starring in Batman Begins*
  - Results are defined by an equivalent SPARQL query
    - Systems are free to return list of results or a SPARQL query
  - Third edition at CLEF 2013
    - Multilingual Q&A
    - Ontology lexicalization

Summary

- Semantic Search is a new field
  - Emerged because of available semantic data and the needs of IR
  - Intersection of IR, DB, Semantic Web
- Evaluation
  - Both efficiency (DB-style) and effectiveness (IR-style)
  - Multiple evaluations addressing different types of query and data
  - Mechanical Turk based evaluation an alternative to experts or users
- Future work
  - Evaluating complex conversational search systems such as Siri

The End

- Many thanks to members of the SemSearch group at Yahoo! Research in Barcelona
- Contact
  - pmika@yahoo-inc.com
  - Internships available for PhD students (deadline in January)