Plan for Today's Lecture(s)

- Getting beyond "10 blue links" (holdover)
- Multi-platform and Mobile Information Retrieval
- Using Context in Information Retrieval
- Using Text and Non-Text Descriptions to Retrieve Non-Text Resources
- Delphi

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

19 November 2013
Lecture 24.4 –Getting Beyond "10 Blue Links"

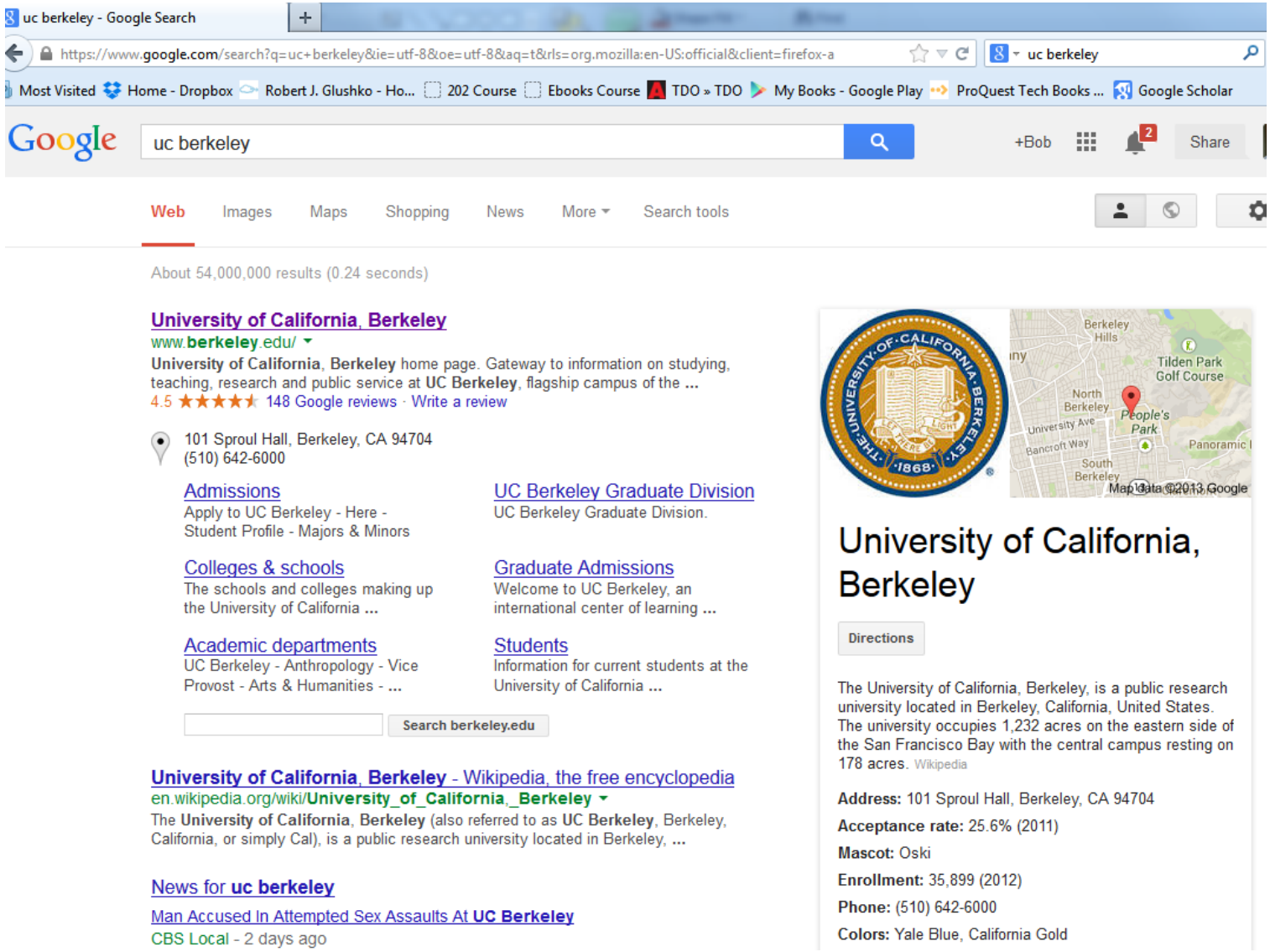
IR Models and “Information Needs”

- In classical IR, an "information need" is formulated as a query submitted to some information collection
- "Documents" in the information collection are returned that "satisfy" the query
- The "IR model" specifies how the queries and documents are represented and how "satisfaction" is calculated
- "Satisfaction" can be an exact match, but more often the returned documents are ranked according to the statistical similarity between their representations and that of the query

Changes in Information Needs

- Web search initially emphasized discovery and "finding out about" --> retrieving a ranked list of relevant sites is appropriate for that information need
- More recently, other types of information needs and queries are more common:
  - "navigational" - the user wants a specific page that they know exists ("UC Berkeley")
  - "informational" - user wants an answer to a well-defined question ("Berkeley temperature")
  - "transactional" - user wants to buy something

“Navigation” with “Content Assembly”



Getting Beyond the "10 Blue Links"

- IR systems locate relevant documents in response to a query, but the user must extract the actual answer to his or her question
- This challenge has been described as "getting beyond the 10 blue links" – here are some of the ways to do that:
  - Semantic search
  - Structured data search
  - Structure / graph search

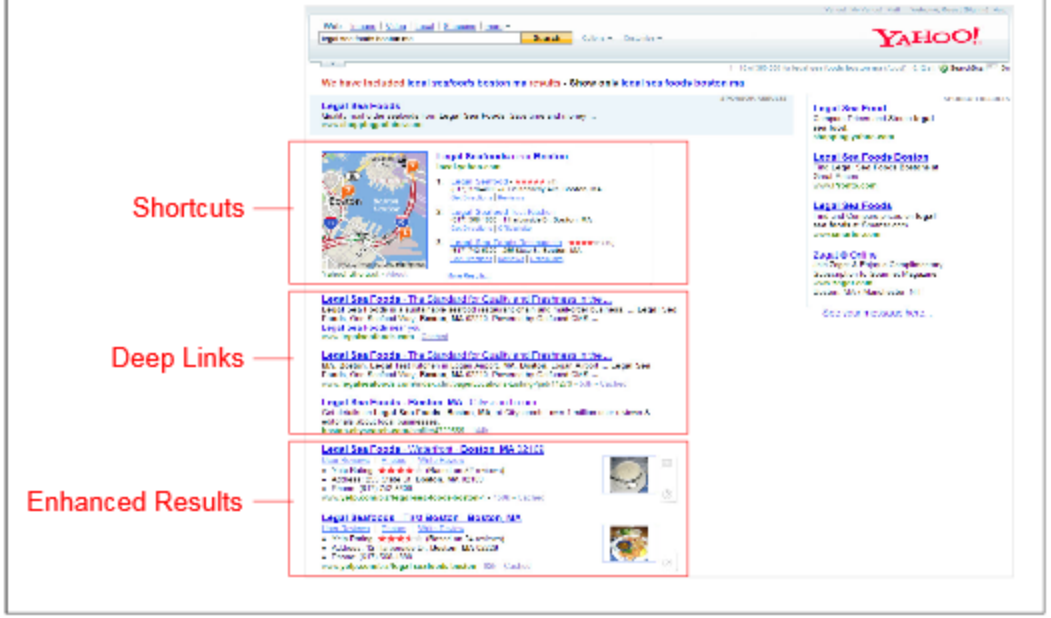
Getting Beyond the "10 Blue Links"

- Search engines improve their responses to these kinds of queries not by improving relevance ranking but by becoming better at identifying user intent
- "Identifying user intent" doesn't require "understanding" the query or the retrieved documents, but it involves recognizing query patterns and matching them with document / data structures likely to contain the answers to the query
- Structured data is most valuable when natural language processing techniques enable queries to be deconstructed, transformed, and joined to find information from more than one source

Structured Data

- Structured data can be extracted from a wide variety of documents and Web sources depending on the consistency and predictability of structure, presentation, and content
  - Microformats, schema.org templates
  - “Info boxes” in Wikipedia and elsewhere
  - “Linked data” sources, other “knowledge bases”

More Information in One Search



The new frontiers of Web search: going beyond the 10 blue links

Ricardo Baeza-Yates, Andrei Broder, Yoelle Maarek and Prabhakar Raghavan

Yahoo! Labs

- We move from a Web of Pages to a Web of Objects
- Objects are people, places, businesses, restaurants … (named entities)
- Objects have attributes
  - Missing, noisy, etc.
- Intents are satisfied by presenting objects and attributes
- Attributes define faceted search

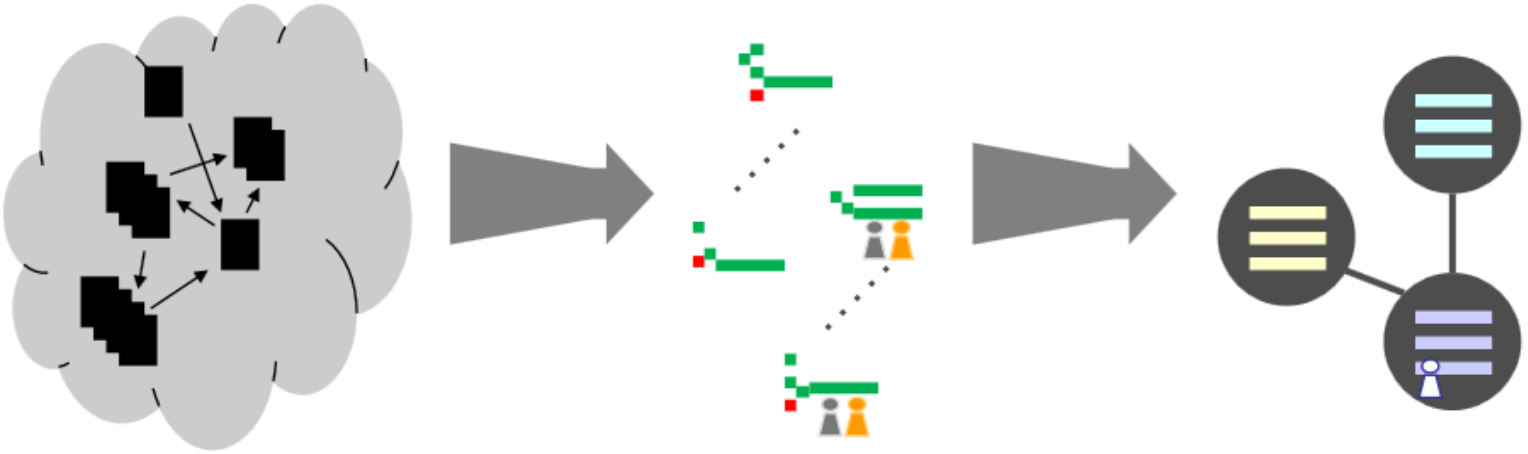
Old: Find relevant pages → "web of pages"

The new frontiers of Web search: going beyond the 10 blue links

Ricardo Baeza-Yates, Andrei Broder, Yoelle Maarek and Prabhakar Raghavan

Yahoo! Labs

Deconstructing the Web & creating a web of objects



Crawls, feeds: extract content/data

Identify potential objects, their attributes & possible relationships

Normalize to structured objects with relationships

The new frontiers of Web search: going beyond the 10 blue links

Ricardo Baeza-Yates, Andrei Broder, Yoelle Maarek and Prabhakar Raghavan

Yahoo! Labs

Find relevant facts/entities & combine → "web of things"

The new frontiers of Web search: going beyond the 10 blue links

Ricardo Baeza-Yates, Andrei Broder, Yoelle Maarek and Prabhakar Raghavan

Yahoo! Labs

Wolfram Alpha

- Wolfram Alpha is a computational “answer engine” developed by Wolfram Research (and based on Mathematica, a widely-used computational toolkit)
- It answers factual queries directly by computing the answer from externally sourced "curated” data from the “deep” web
- It has substantial natural language capability

Smarter Search!

WolframAlpha computational... knowledge engine

unemployment oakland california

Examples Random

Input interpretation:

Oakland, California | unemployment rate

Result:

11.4% (August 2013)

History: Log scale

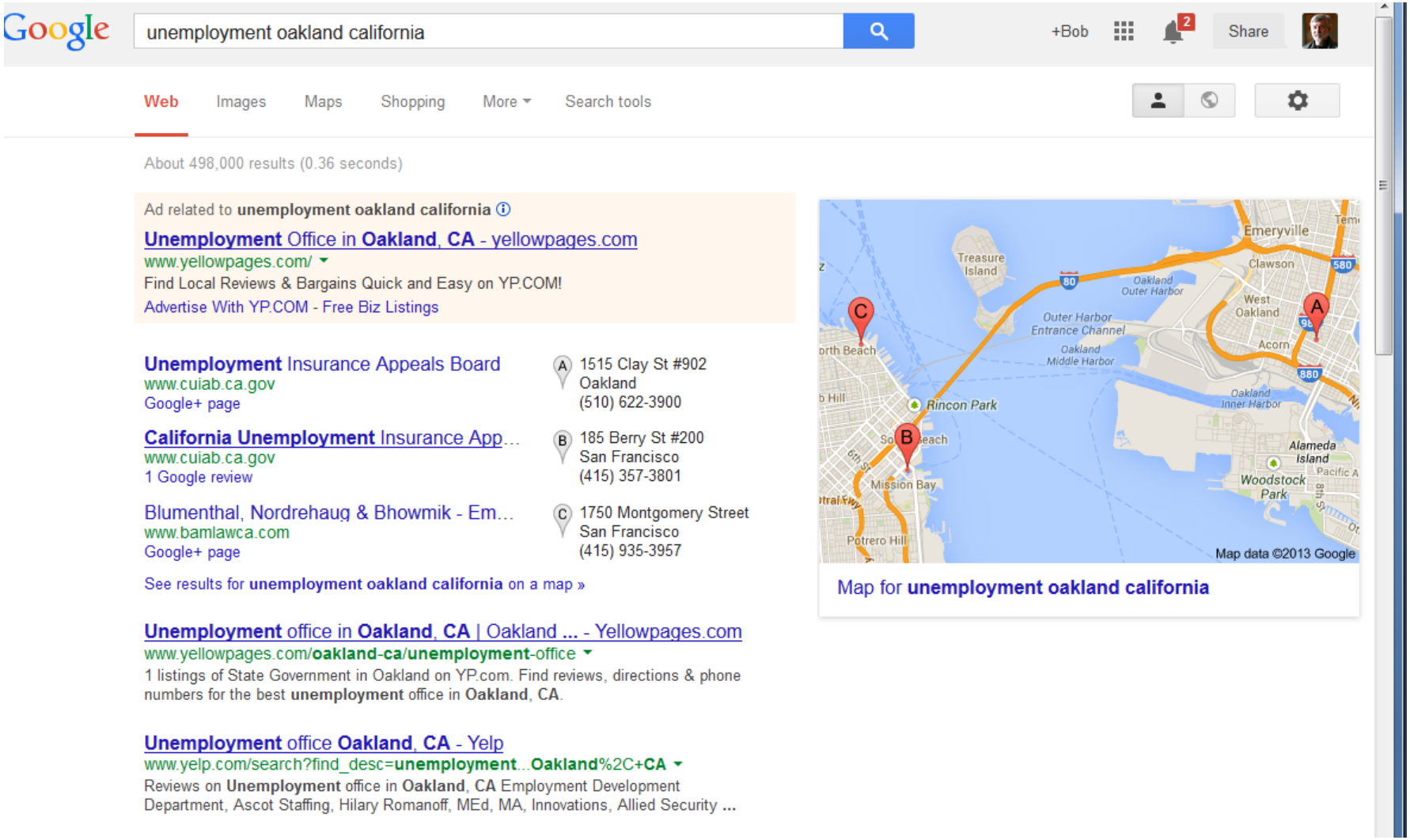
(from Jan 1990 to Aug 2013) (in percent)

(not seasonally adjusted)

Computed by Wolfram Mathematica

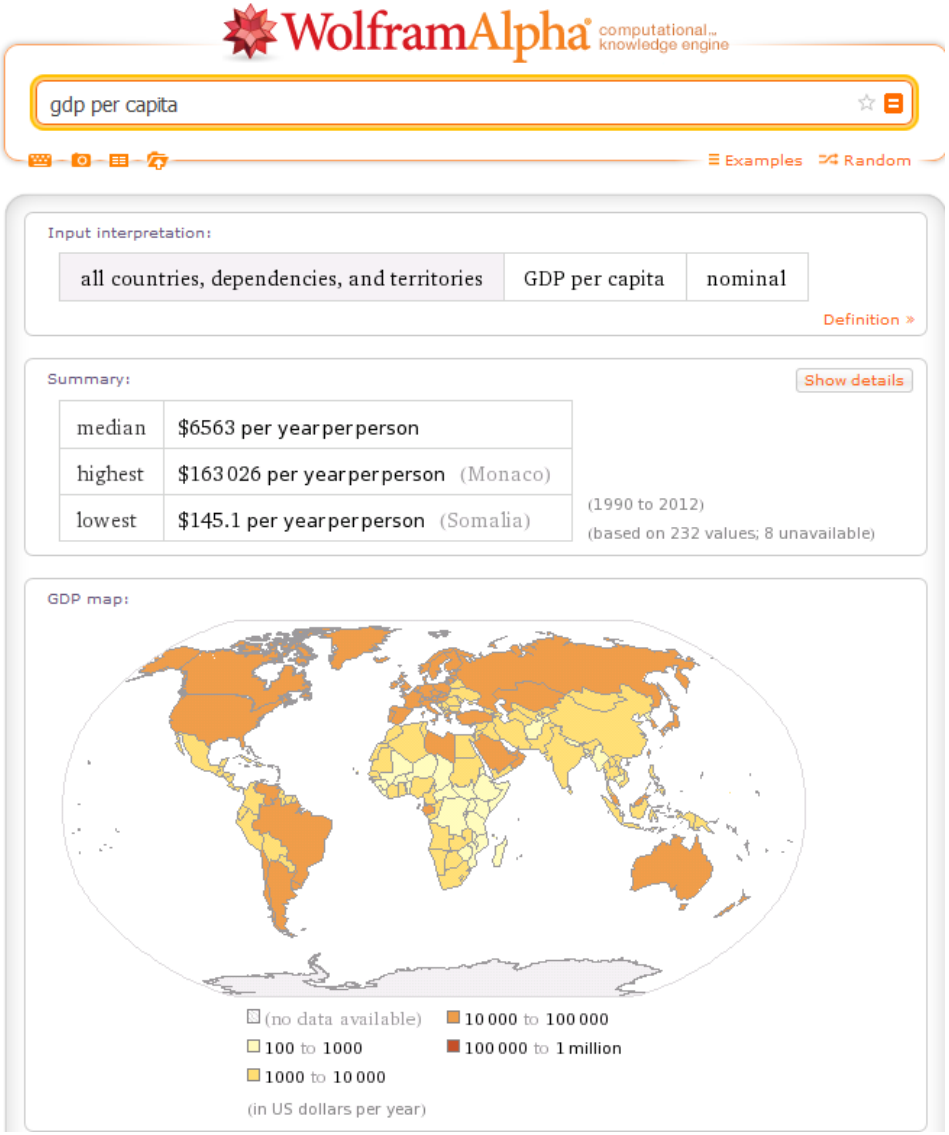
Sources Download page

(Compare to…)


Google
unemployment oakland california
+Bob
Share
Web
Images
Maps
Shopping
More
Search tools
About 498,000 results (0.36 seconds)
Ad related to unemployment oakland california
Unemployment Office in Oakland, CA - yellowpages.com
www.yellowpages.com/
Find Local Reviews & Bargains Quick and Easy on YP.COM!
Advertise With YP.COM - Free Biz Listings
Unemployment Insurance Appeals Board
www.cuiab.ca.gov
Google+ page
A 1515 Clay St #902
Oakland
(510) 622-3900
California Unemployment Insurance App…
www.cuiab.ca.gov
1 Google review
B 185 Berry St #200
San Francisco
(415) 357-3801
Blumenthal, Nordrehaug & Bhowmik - Em…
www.bamlawca.com
Google+ page
C 1750 Montgomery Street
San Francisco
(415) 935-3957
See results for unemployment oakland california on a map »
Unemployment office in Oakland, CA | Oakland ... - Yellowpages.com
www.yellowpages.com/oakland-ca/unemployment-office
1 listings of State Government in Oakland on YP.com. Find reviews, directions & phone numbers for the best unemployment office in Oakland, CA.
Unemployment office Oakland, CA - Yelp
www.yelp.com/search?find_desc=unemployment...Oakland%2C+CA
Reviews on Unemployment office in Oakland, CA Employment Development Department, Ascot Staffing, Hilary Romanoff, MEd, MA, Innovations, Allied Security ...
Emeryville
Treasure Island
Clawson
Oakland Outer Harbor
West Oakland
Outer Harbor Entrance Channel
Acorn
Oakland Middle Harbor
Oakland Inner Harbor
Rincon Park
Alameda Island
Woodstock Park
Mission Bay
Potrero Hill
Map data ©2013 Google
Map for unemployment oakland california

Smarter Search!

Smarter Search!



[What is the fifteenth highest country by GDP per capita?]

[AND WHAT'S COMING NEXT FROM WOLFRAM IS TRULY MIND-BOGGLING – A PROGRAMMING LANGUAGE THAT CONTAINS ALL THE KNOWLEDGE BUILT INTO WOLFRAM ALPHA!]

Facebook Graph Search

- Searches the information created by the billion users of Facebook
- Queries consist of multiple phrases that map to different parts of the structured data in profiles
  - Locations
  - Time
  - Objects (that are mentioned, tagged, liked…)
- Results are based on content matches constrained by the relationships between a user and the people in his or her social network

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

21 November 2013
Lecture 25.1 – Multi-platform and Mobile IR

Multi-platform and Mobile Information Retrieval

- How are queries affected?
- How are user interfaces affected?
- How are query results affected?

The Multi-{Device, Platform} Design Imperative

- Many people routinely use more than one platform or device for different services, but this can create expectations about using the same service on different platforms
- (How) can a service provider offer the "same" service on multiple platforms or devices?

The Mobile "Phone" Platform

- The mobile phone was invented in the 1970s, but widespread commercial use didn't begin until the 1990s; SMS was added in 1991
- Today, many times more people in the world use mobile phones than personal computers
- There are several hundred models of mobile phones on the market today-
- A rapidly growing percentage of "phones" are multifunction devices with substantial computing capability

The Simplest Mobile IR Services

- On even the "dumbest" phones a query and response can be performed using an SMS search service)
- But since your dumb phone doesn't know its location, you have to explicitly localize your query or you'll not get the results you want for many queries

Comparing Devices / Platforms

- Devices and platforms differ on many different capability dimensions
  - Computing power
  - Storage capacity
  - Portability
  - Display size and resolution
  - Multimedia processing
  - Bandwidth
  - GPS capability
  - …

Comparing Devices / Platforms

- These capabilities are not always correlated and bundled into devices in the same combinations (for both technical and business reasons)
- --> user interface design tradeoffs across devices

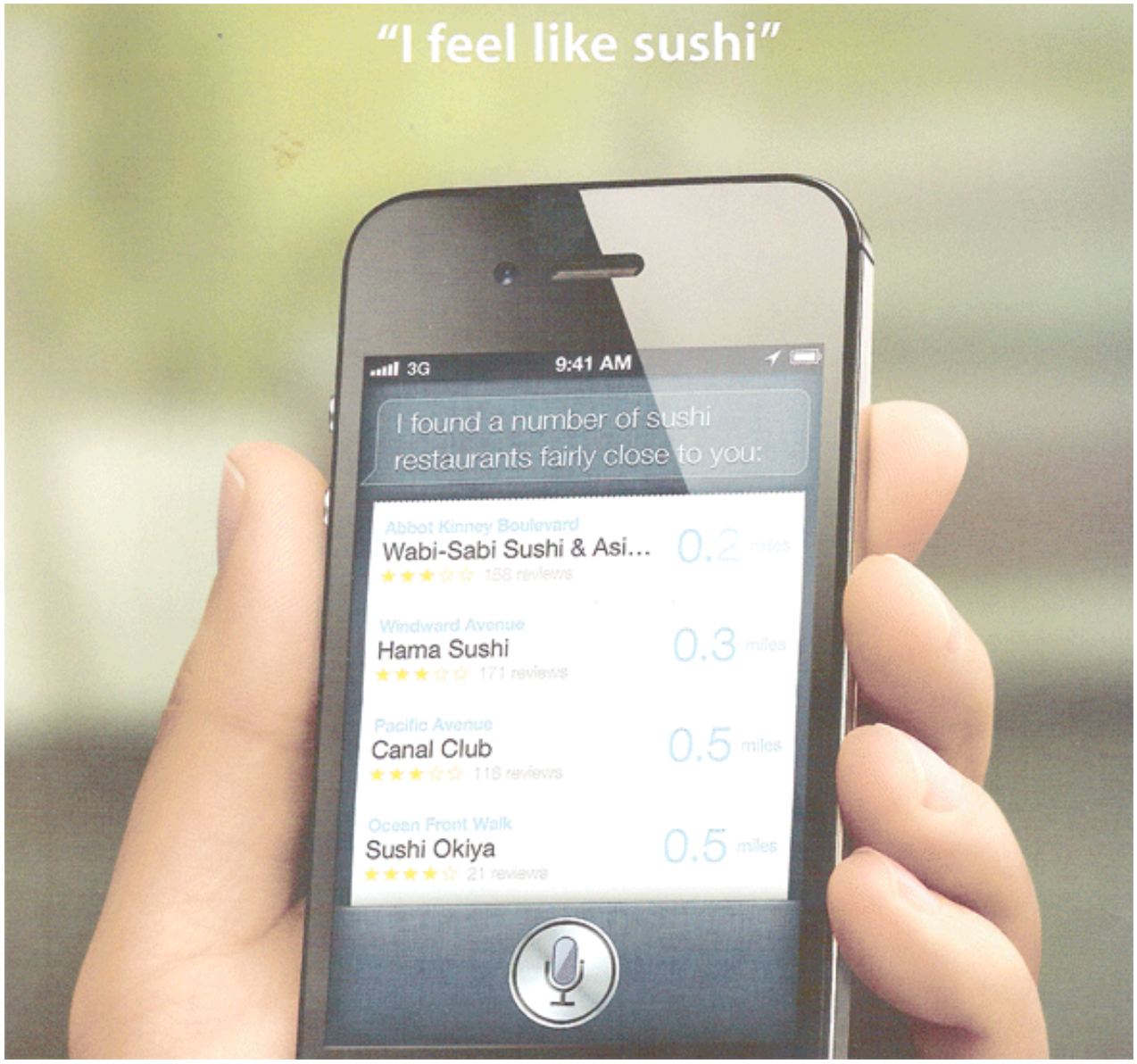
Smart Phone Capability Checklist

- Wi-fi
- Touchscreen
- Keypad
- Music (MP3) playback
- Audio recording
- Video (MP4) playback
- Video recording
- Camera
- GPS (possibly with navigation)
- Bluetooth
- USB
- Memory card reader
- Alarms
- FM Radio
- Picture calling
- Video calling
- …
- ***An application developer community***

User Interfaces on Mobile Devices

- Mobile devices lack full-size keyboards; entering text-based queries can be tedious
- But mobile devices have input capabilities that larger computers lack
  - Voice recognition (e.g., SIRI)
  - Audio processing capability of mobile phone to create content-based query (e.g., Shazam)
  - Image or bar/QR code processing to identify resources

Voice Input with SIRI



Bar Code Input

QR Code Input

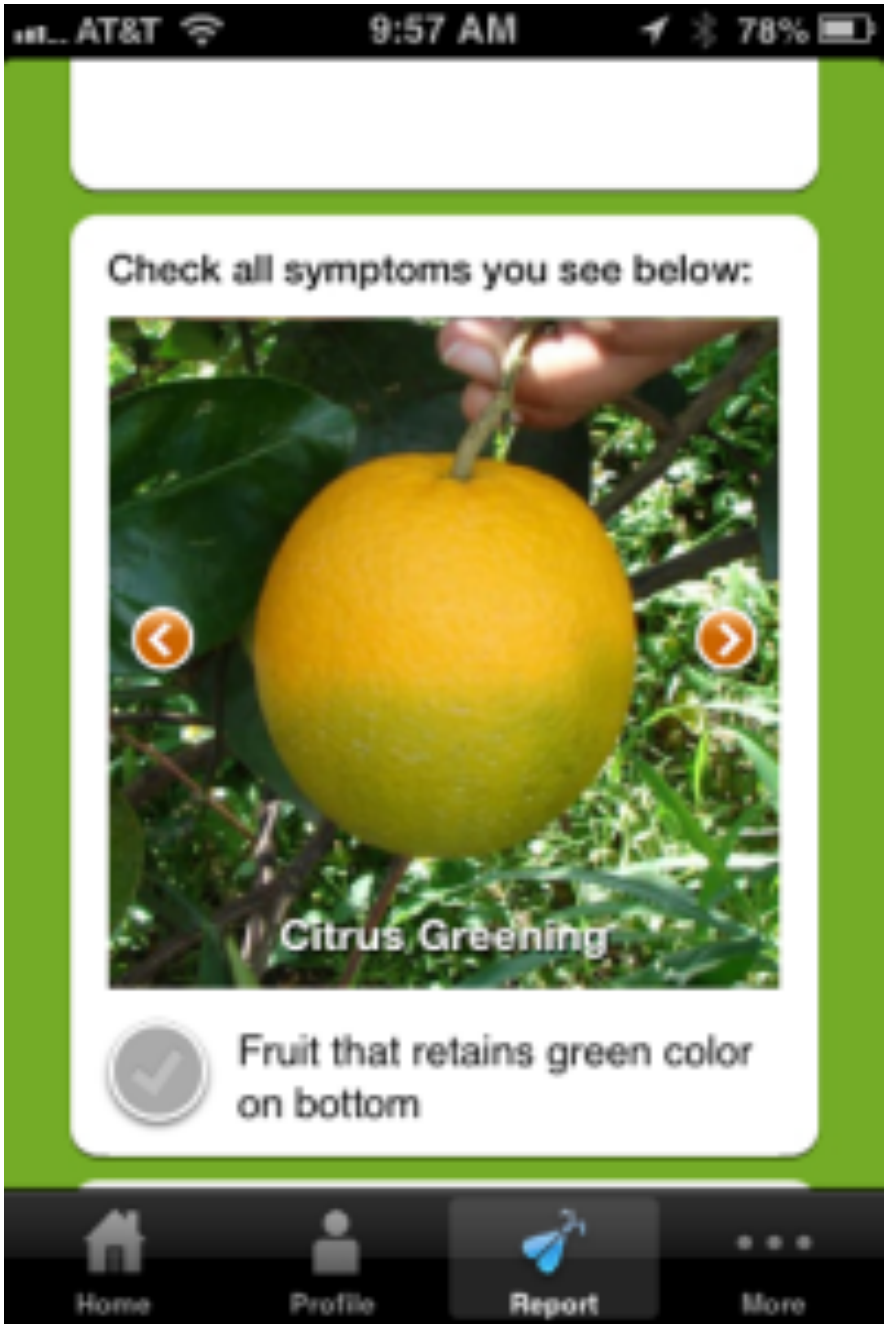


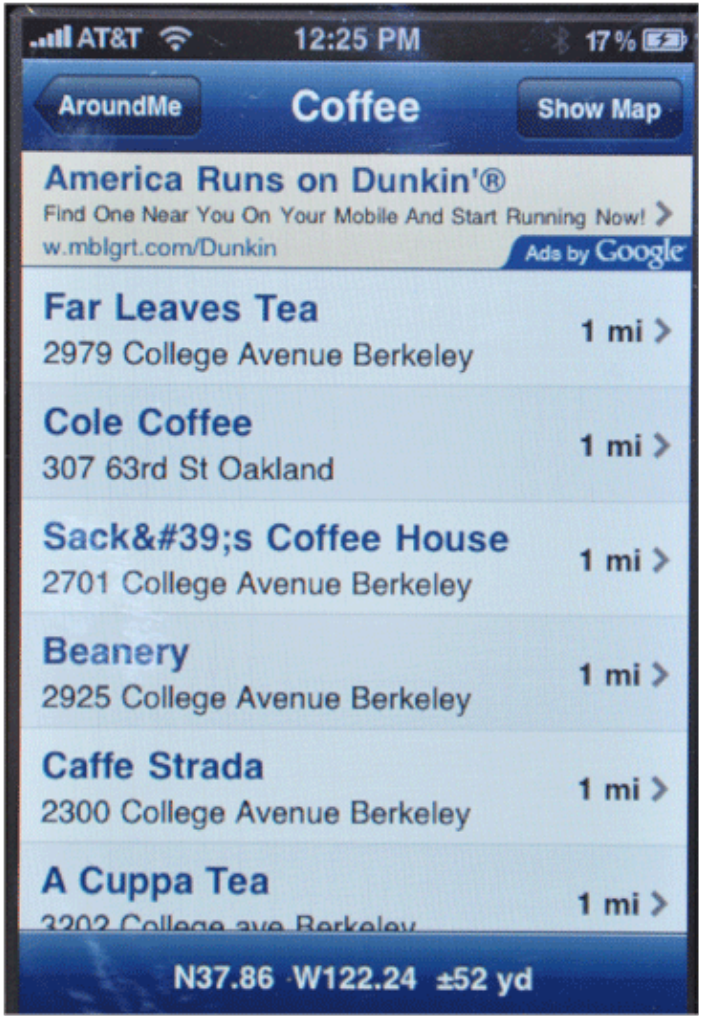
Photo Input - "Save Our Citris"

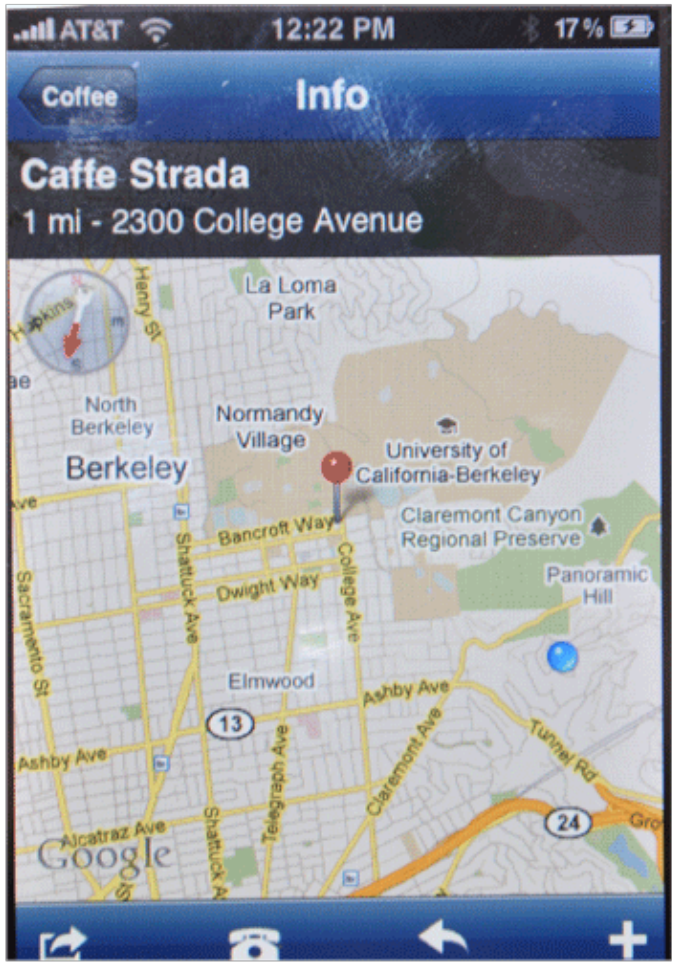
View images of diseased fruit and then take photos of lemons and oranges and send them in to US Dept. of Agriculture for a (sometimes automated) analysis

The Simplest Mobile IR Service on a Smart Phone

- A search service on a smart phone can use just one piece of information - current location - to personalize or adapt the outputs
- Many providers assume local search for "weather" and many other domains that are location-based
- For other services that have local providers (restaurant, transportation, etc.) many "Local Search" functions list only "nearby" results pinpointed on a local map

Search Results as List & Map ("Around Me")





Presenting Results for Mobile Queries

- Mobile phones have smaller screens than non-mobile computers so conventional search results presentation won't work as well
- But information about location makes for more precise results, so results set is smaller
- Display limits imply reformatted and abbreviated text content

Presenting Results for Mobile Queries

- Bandwidth limits might imply transcoding to lower quality images, audio, or video content
- Channel translation the right solution in some contexts; voice synthesis instead of text display
- Maps are natural UI elements because most mobile services are location-based

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

21 November 2013
Lecture 25.2 – Using Context in Information Retrieval

Defining "Context Awareness"

If devices can exploit emerging technologies

to infer the current *activity state* of the user

and the *characteristics* of the environment

they can then intelligently manage both the *information content*

and the means of *information distribution*

Design Issues for Context-Awareness

- What information is defined as the "context" ?
- How is the information sensed or obtained from the environment?
- When is the context information obtained?
- Is this information always collected in the same place, or is locating and arranging the sensors or collection points a task in its own right?
- What initiates the sending of context information to the service?

Design Issues for Context-Awareness

- How "smart" is the information when the service receives it?
- Is the contextual information shared with other contexts and services?
- How does the service process the information it obtains?

Context Awareness: Substituting Information for Interaction

- Context awareness means a system or service doesn't need to ask:
  - Where are you?
  - What are you doing, looking at, looking for?
  - Who are you with?
  - What's it like there?
- And the service can eliminate irrelevant information from its responses

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

21 November 2013
Lecture 25.3 – Using Text and Non-text Descriptions to Retrieve Non-text Resources

Describing Non-Text With Text - By People

- People can always assign names and other textual metadata to non-text objects
- Professional cataloguers of "museum objects," images/paintings and other "cultural works" often use the Getty CDWA or CDWA-Lite
- ID3 tags on MP3 audio files contain a very limited amount of song metadata
- MPEG-7 is the newest, most standard, and most complicated specification for "semantic" image and video metadata

Describing Non-Text With Text - Automated

- Other textual metadata can be assigned by the devices or mechanisms that created the non-text objects (e.g.EXIF (Exchangeable Image File Format) is used in digital cameras)
- But this level of description isn't what most people want to use when they search for non-text resources

Exploiting Multi-Modal Information

- Video often has audio content and text transcriptions
- Correlations or anomalies with these different content types can be exploited "semi-automatically"
- Text overlays (captions) can be used to identify people or places in videos
- Location information (e.g, GPS) attached to images or video can be used to infer content descriptions

Exploiting Multi-Modal Information

- Distinguish speech from non-speech sounds; attempt speech recognition for the former
- Excited speech, crowd noise suggest important events
- Multi-modal annotation visualization as user interface

Crossing the Semantic Gap Through Computation

- A consequence of the semantic gap for mulitimedia is that there are a very large number of low-level features that can be reliably identified
- Any description using these features will be "sparse" - lots of missing values
- Dimensionality reduction techniques can exploit correlations between low-level features

"

A Music Search Engine Built upon Audio-based and Web-based Similarity Measures"

- Use "artist" / "album" / "title" metadata in ID3 tags as search terms to find web content about a song (excluding the lyrics)
- Create a tf/idf "document vector" for each song from the terms located in the web content
- Compute 19 different audio characteristics from each song

"

A Music Search Engine Built upon Audio-based and Web-based Similarity Measures"

- Use audio characteristics to reduce the dimensionality of the document vector
- Queries using the descriptive terms can rank order songs by acoustic similarity

Ignoring the Semantic Gap in Search

- But maybe we don't need to cross the semantic gap to have effective multimedia IR
- We can use low-level features that can be extracted automatically to index the multimedia collection and then extract the same ones from a multimedia query
- Shazam uses audio "fingerprinting" for identification: (try Test Song)
- Midomi uses different features to enable query by humming (broader “equivalence class”)

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

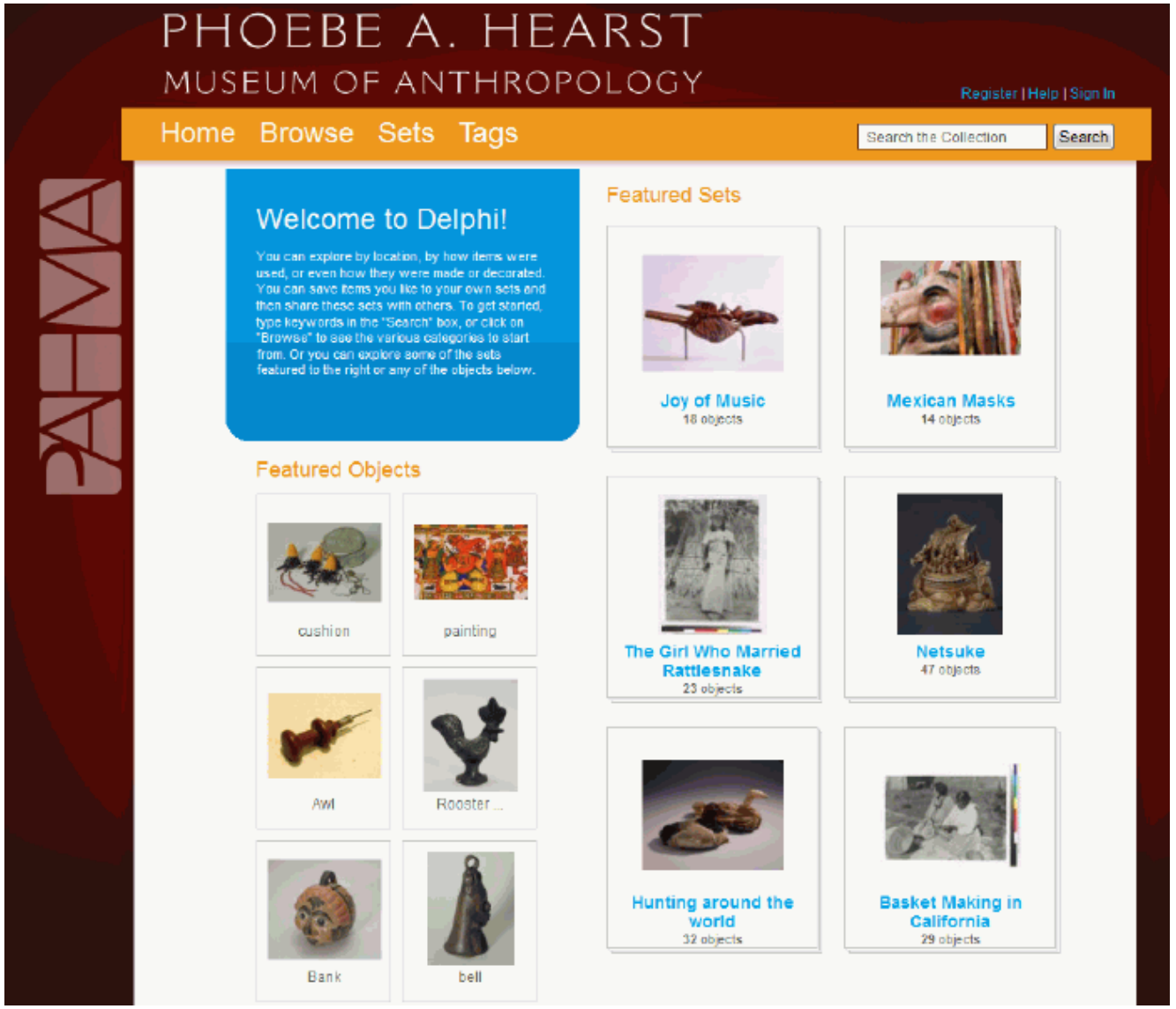
INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

21 November 2013
Lecture 25.4 – Delphi

"The Delphi Toolkit: Enabling Semantic Search in Museum Collections"

Delphi Home Page



Motivation for Delphi

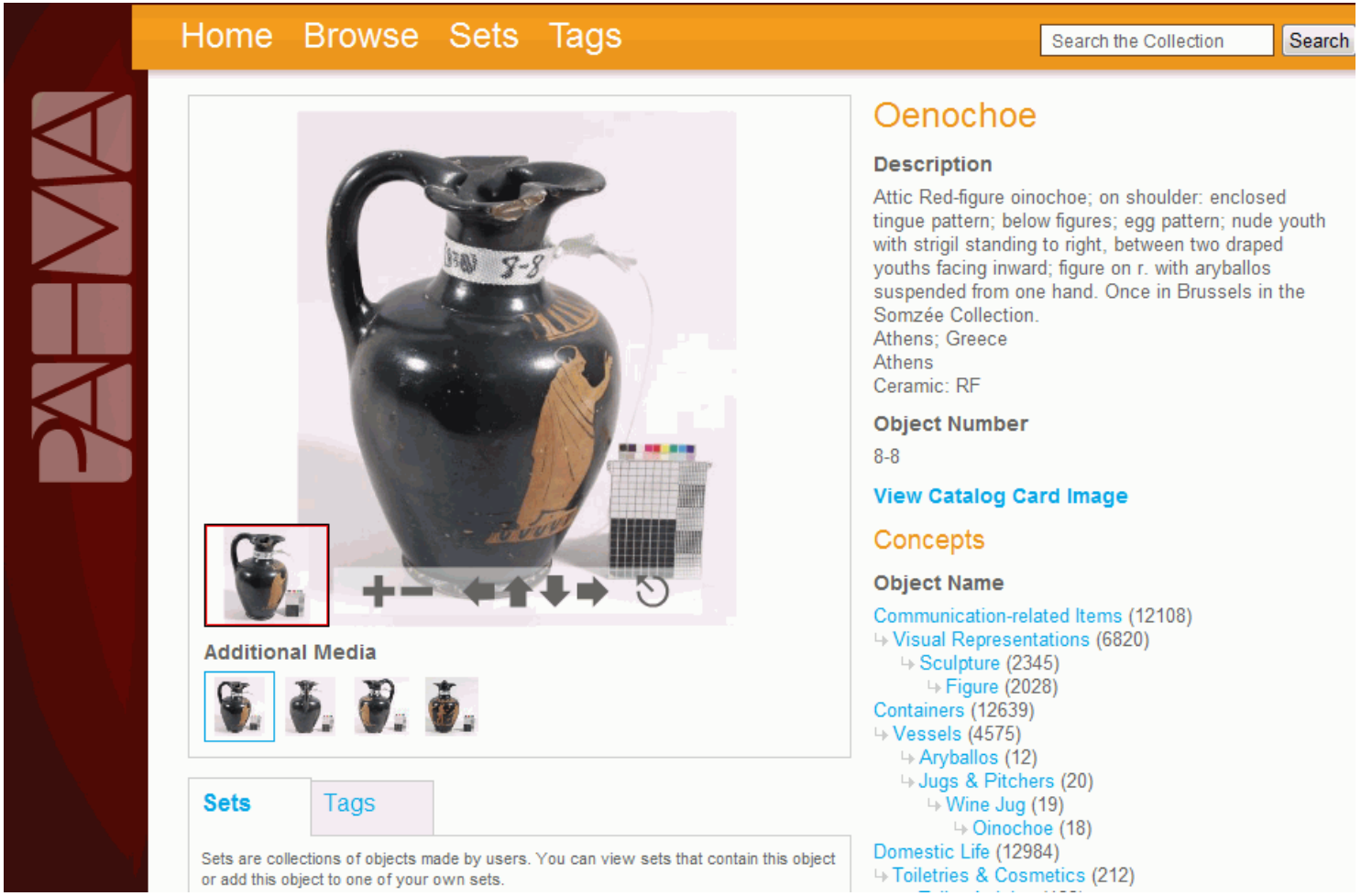
- The Museum of Anthropology (founded in 1901 by Phoebe Apperson Hearst) has an estimated 2-3 million objects in its collection
- While the museum is relatively well known among researchers, a small public exhibition space (relative to the total collection size) means that the public has no clue about the richness of the collection
- The Museum uses a collection management system (TMS) with roughly over 600,000 records describing the entire collection

Delphi Faceted Browsing

What Delphi Does

- Uses NLP techniques to enhance "scientific anthropological" metadata and descriptions of objects
- Organizes the metadata and descriptions into an ontology
- Ontology facets generate a dynamic searching and browsing UI
- Supports personal user organization and annotation of the collections by allowing visitors to create sets and tag individual items

Delphi Resource



Readings for Next Lecture

- TDO 6.5, 7.6
- Li, Jiexun, Rong Zheng, and Hsinchun Chen. "From fingerprint to writeprint." Communications of the ACM 49, no. 4 (2006): 76-82. dl.acm.org/citation.cfm?id=1121951
- Blanzieri, Enrico, and Anton Bryl. "A survey of learning-based techniques of email spam filtering." Artificial Intelligence Review 29, no. 1 (2008): 63-92. (Section 3, 67-75). link.springer.com/article/10.1007/s10462-009-9109-6
- Linden, Greg, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." Internet Computing, IEEE 7, no. 1 (2003): 76-80.