

Scalable Construction and Querying of Massive Knowledge Bases

Xiang Ren¹ Yu Su² Pedro Szekely³ Xifeng Yan²

¹University of Southern California

²University of California, Santa Barbara

³Information Sciences Institute, University of Southern California



Tutorial website:

http://usc-isi-i2.github.io/WWW18_1/

Slides, code, datasets, references



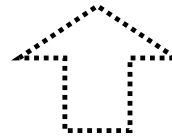
Growing Gap between Human and Data



What disease does the patient have?

- (EMR) Similar patients?
- (Literature) New findings?
- (Gene sequence) Suspicious mutations?
-

Ad-hoc information needs for on-demand decision making



Massive, heterogeneous data

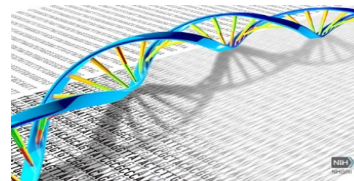
86.9% adoption
(NEHRS 2015)



27M+ papers, >1M
new/year (PubMed)



\$1000 gene sequencing



24x7 monitoring

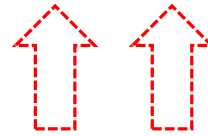


How can AI Bridge the Gap?



Insights
Discoveries
Solutions

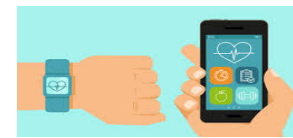
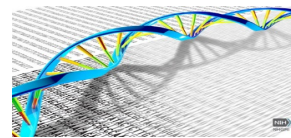
Bottleneck #2: Access



Bottleneck #3: Reasoning



Bottleneck #1: Knowledge

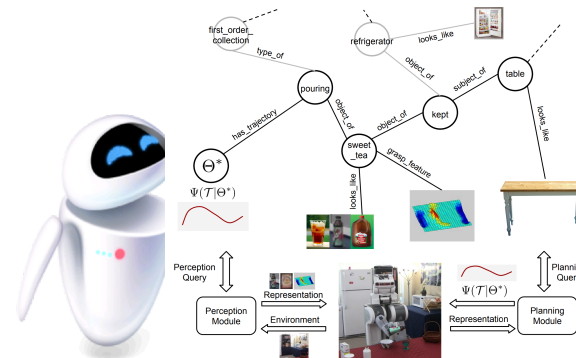


Broad Applications



“Which cement stocks go up the most when a Category 3 hurricane hits Florida?”

KENSHO



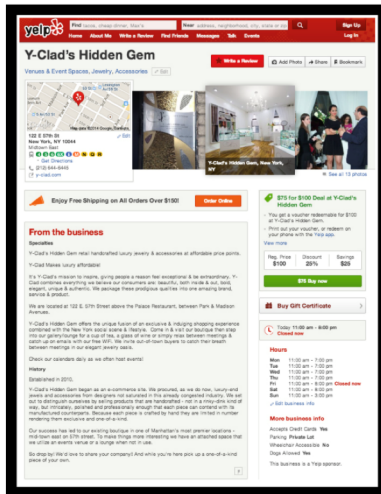
Constructing Domain Specific Knowledge Graphs

Pedro Szekely

**Information Sciences Institute,
University of Southern California**

Domain-specific search (DSS)

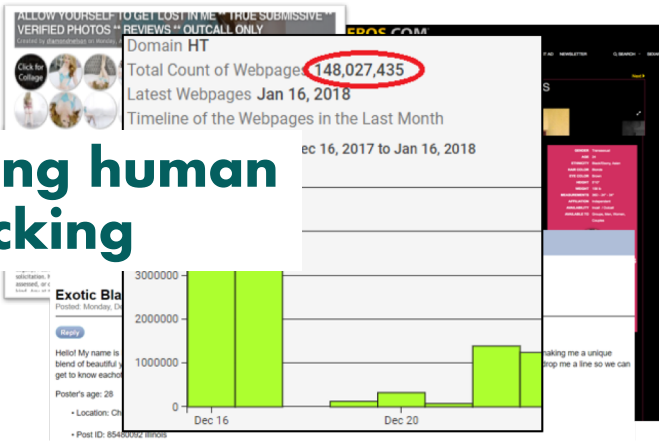
The Massive YouTube Ecosystem



source: <https://photos.pnnewswire.com/prnfull/20151006/274273-INFO>

Emerging opportunities for DSS

Fighting human trafficking



Predicting cyberattacks



Penny Stock Fraud Nets Millions

Scheme Mastermind Among Those Sentenced to Prison

Stopping Penny Stock Fraud

Internet opens new avenue for penny stock fraud
NEW YORK □ Most investors take e-mails advertising a 300 percent return on penny stocks as a sure thing. But those Internet promotions are still irresistible to some investors, leading to a string of making a killing.

July 11, 2004

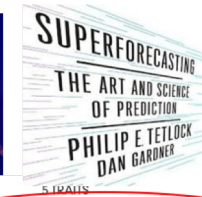
The SEC is increasingly taking legal action against individuals and companies that engage in penny stock fraud. In one of its recent cases, involving Ives Health Co., the SEC reported a final order against M. Keith Ives, for disseminating misleading information on the Internet.

In the Ives case, Ives sold investors a total of \$1.25 million for, among other things, falsely claiming that Ives Health had developed a new drug.

Defined by the SEC as stocks that sell below \$5 a share, penny stocks have always been considered speculative and easily manipulated. But stock market experts, seeing an increase in penny stock promotion online, say investors should be wary of



Accurate geopolitical forecasting



5. Traits
- perforecasters begin by gathering as much information as possible.
 - perforecasters nurture and develop the habit of thinking in terms of probabilities when exploring the likelihood of specific events.
 - perforecasting improves when individuals work in teams.
 - perforecasters ensure that they are regularly keeping track of their projections.
5. The most successful forecasters are willing to admit error and quickly change course on their projections.



DARPA/IARPA programs

DARPA Memex

IARPA Hybrid Forecasting Competition

DARPA AIDA

DARPA Causal Exploration

DARPA LORELEI

IARPA CAUSE

DSS is more than keyword search

Lead Investigation

What is the ad with the earliest post date containing number 7075610282?

Aggregations/Lists

List all ads in Seattle, WA that include an ethnicity in the ad text. In the answer field, concatenate and list ethnicities

Indicator Mining

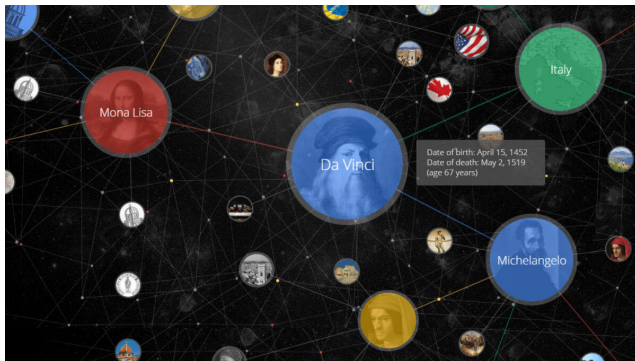
List all ads that have high probability of movement

List all ads in the Chicago area advertising multiple people at once

Dossier Generation

Collect and show me all information on the phone number 7075610282

Google Knowledge Graph



Google Larry Page

Web Images Maps Shopping News More Search tools

About 350,000,000 results (0.24 seconds)

Larry Page - Wikipedia - the free encyclopedia
en.wikipedia.org/wiki/Larry_Page
Lawrence "Larry" Page (born March 26, 1973) is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey ... Marissa Mayer - Carrie Southworth - PageRank - Forbes 400

News for Larry Page
Larry Page Gets A Literal Android KIKat
Uberzmo - 3 days ago
Android 4.4 KitKat marks a milestone for Google as they have named their mobile operating system after a branded chocolate - although ...

Larry Page - Forbes
www.forbes.com/profile/larry-page/ -
Larry Page on Forbes - #20 Billionaires, #20 Powerful People, #13 Forbes 400.

Larry Page - Google+
https://plus.google.com/+LarryPage/

Knowledge Graph

Larry Page
1973, a.k.a. Lawrence "Larry" Page

Lawrence "Larry" Page is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey Brin. On April 4, 2011, Page succeeded Eric Schmidt as the chief executive officer of Google. [More info](#)

Born: March 26, 1973 (age 40), East Lansing, MI
Height: 5' 11" (1.80 m)
Spouse: Lucinda Southworth (m. 2007)
Siblings: Carl Victor Page, Jr.
Education: East Lansing High School (1987–1991), More
Awards: Marconi Prize, TR100

Recent posts
Just opened the new Android release, KikKat [Sep 3, 2013](#)

People also search for
Sergey Brin, Eric Schmidt, Larry Ellison, Marissa Mayer, Bill Gates

About 36,700,000 results (0.67 seconds)

Wonder Woman (2017) - IMDb

www.imdb.com/title/tt0451279/ -
★★★★★ Rating: 7.6/10 - 360,568 votes
When a pilot crashes and tells of secret life in the outside world, Diana, an Amazonian warrior in training, leaves home to fight evil and discover her full powers and true destiny.
Full Cast & Crew · Chris Pine · Trivia · Parents Guide

Wonder Woman (2017 film) - Wikipedia

https://en.wikipedia.org/wiki/Wonder_Woman_(2017_film) -
Wonder Woman is a 2017 American superhero film based on the DC Comics character of the same name, distributed by Warner Bros. Pictures. It is the fourth installment in the DC Extended Universe (DCEU). The film is directed by Patty Jenkins, with a screenplay by Allan Heinberg, from a story by Heinberg, Zack Snyder, Gal Gadot · Patty Jenkins · Elena Anaya · Doctor Poison

Top stories

Oscars voting ends today: Will 'Wonder Woman' finally break the anti-superhero streak? [Washington Post](#)

Fashion War: Wonder Woman Gal Gadot Infuriates Lebanese with Dress Design [Breitbart](#)

Gal Gadot Diet and Fitness Routine | POPSUGAR Fitness Australia [POPSUGAR Australia](#)



Wonder Woman

PG-13 2017 · Fantasy/Science fiction film · 2h 21m

Play trailer on YouTube

7.6/10 IMDb
92% Rotten Tomatoes

90% liked this movie
google users

Before she was Wonder Woman (Gal Gadot), she was Diana, princess of the Amazons, trained to be an unconquerable warrior. Raised on a sheltered island paradise, Diana meets an American pilot (Chris Pine) who tells her about the massive conflict that's raging in the outside world. Convinced that she c... [MORE](#) ▾

Release date: June 2, 2017 (USA)

What is a Knowledge Graph?

set of triples, where each triple (h, r, t) represents a **relationship r** between **head entity h** and **tail entity t**

(Barack Obama, wasBornOnDate, 1961-08-04),

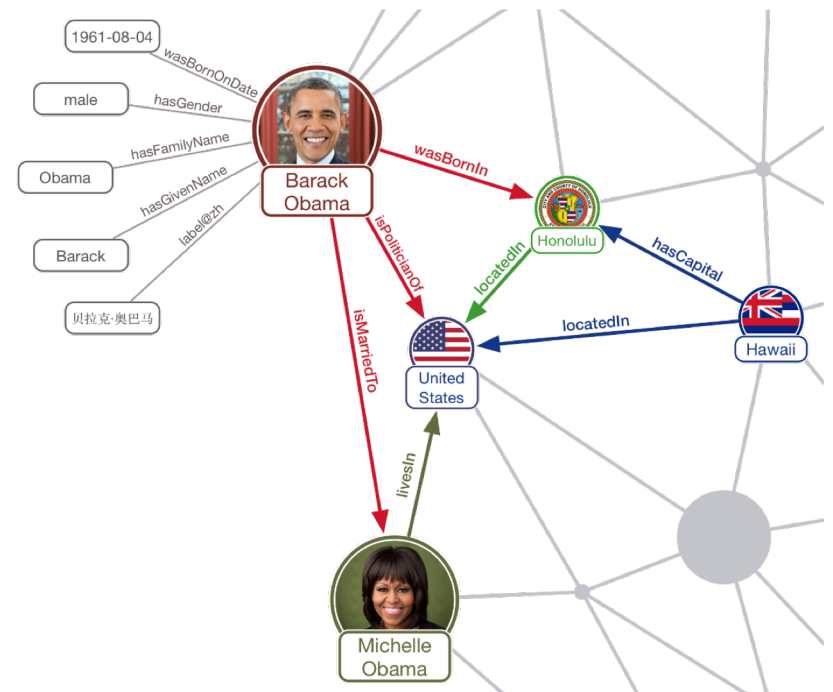
(Barack Obama, hasGender, male),

...

(Hawaii, hasCapital, Honolulu),

...

(Michelle Obama, livesIn, United States)



General Search

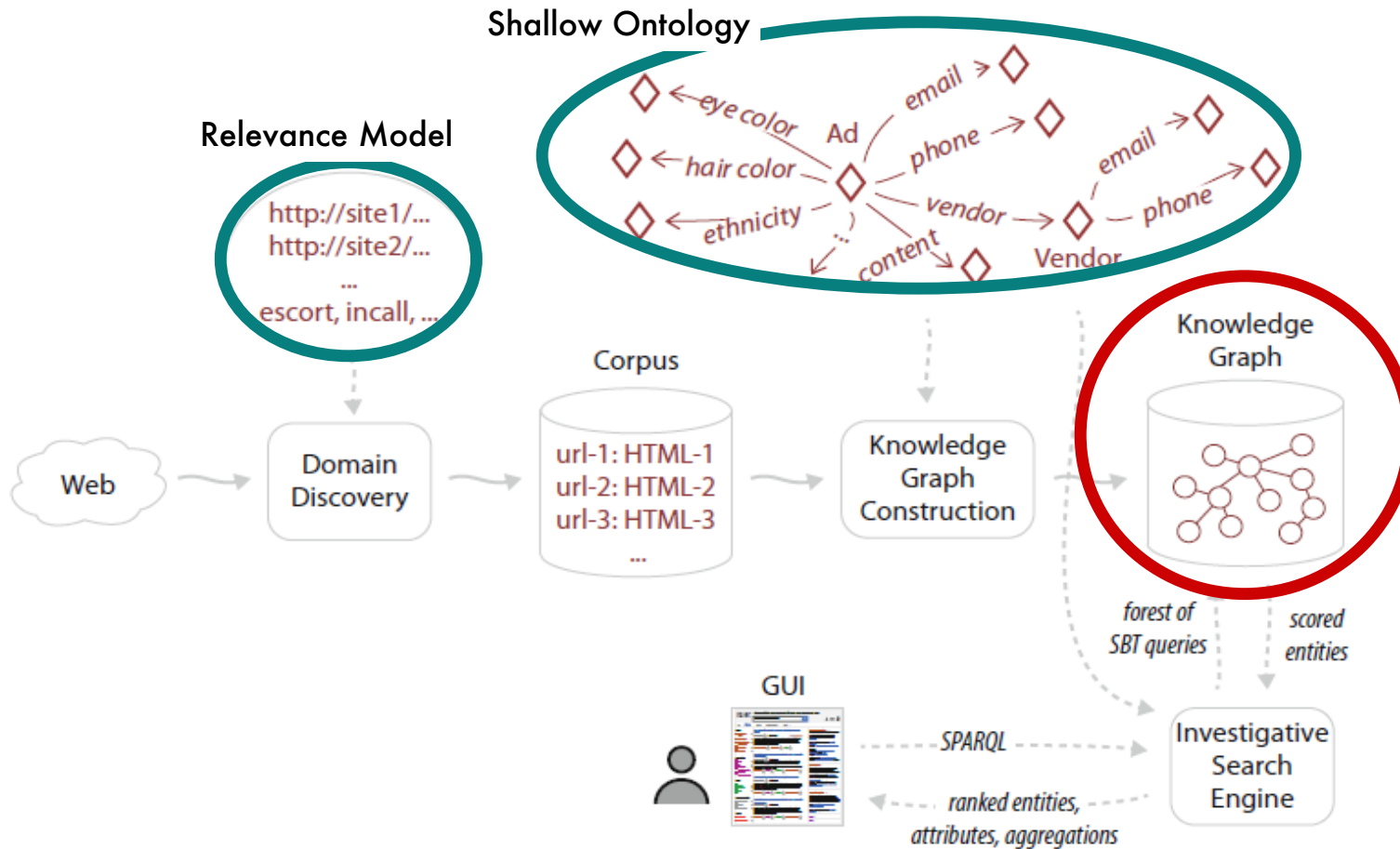
Google Knowledge Graph

DSS

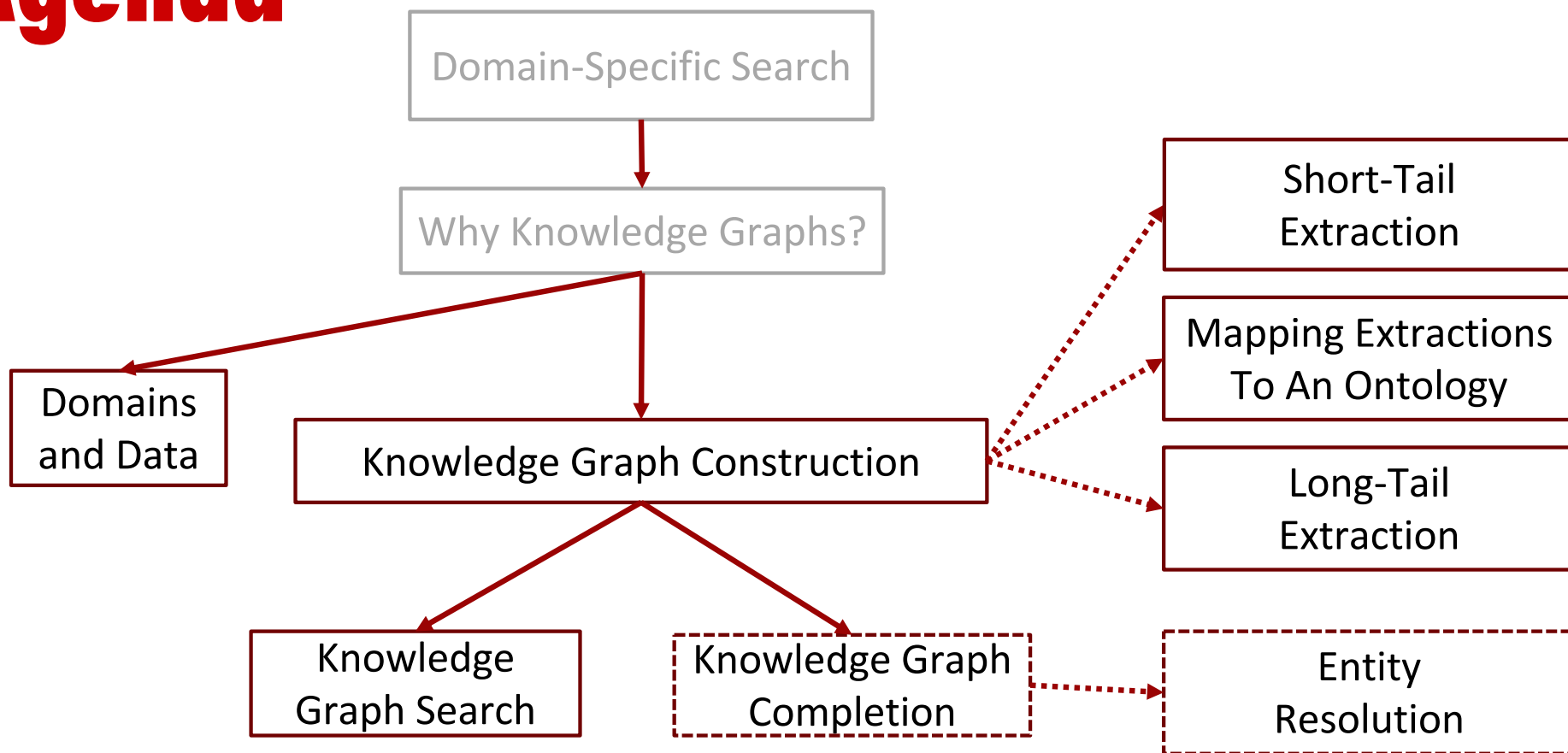
Domain-Specific Knowledge Graphs

How do we construct domain specific knowledge graphs over web data for powerful DSS applications

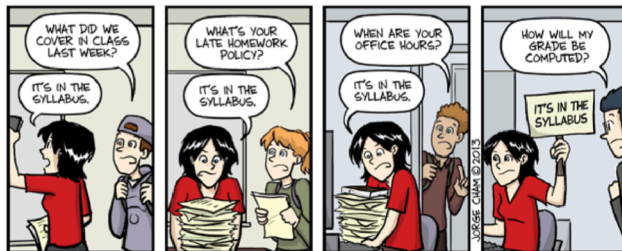
Knowledge Graphs for DSS



Agenda



What is (or even isn't) a domain?

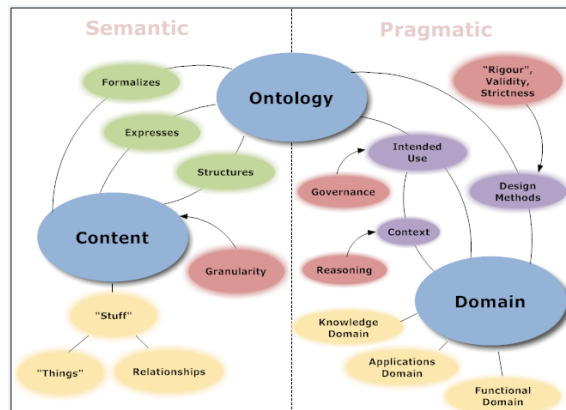


IT'S IN THE SYLLABUS

This message brought to you by every instructor that ever lived.

WWW.PHDCOMICS.COM

"Piled Higher and Deeper" by Jorge Cham



Some dictionary definitions

(Merriam Webster) A sphere of **knowledge, influence** or **activity**

(Oxford) A **specified** sphere of activity or knowledge

Specifying the sphere

Rules

Scope (e.g., the legal system)

Syllabi (for classrooms)

Examples

How do domain experts

specify the sphere?

Examples

Ontology

Domain-Specific Challenges

- Subject matter
- Complex nature
- Ambiguous
- Obfuscation
- How to adapt off-the-shelf tools?

Italian 19 hello guys....My name is charlotte , New to town from kansas
[GORGIOUS BLONDE beauty] ? FROM Florida ? (Petite) ? [CURVy]?
NO DISAPPOINTMENTS. 34C.. Brazilian,ITALIAN beauty....
Hey gentleman im Newyork and i'm looking for generous
Hi guy's this is sexy newyork . & ready to party.
AVAILABLE NOW! ?? - (1 two 1) six 5 six - 0 9 one 2 - 21

Specifying investigative domains



Crawling+domain discovery

crawling



Functional

I have some questions I'd like answers to
Domain is the scope of the answers
Presents interesting cognitive dilemma!
I know what I want but can't define it precisely

Two major functional steps

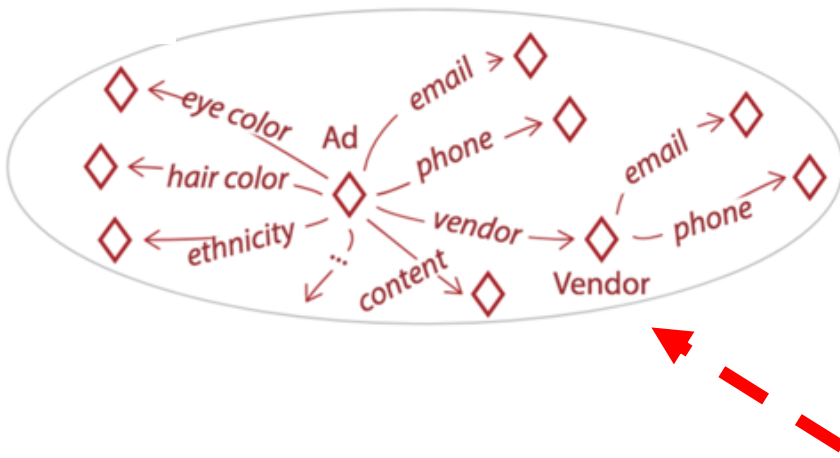
Data Acquisition

- Find me the data from a universe aka the Web that can help me answer my questions

Ontological Specification

- Let me define fields and field properties that will help me unambiguously represent questions and interpret answers

Specifying investigative domains



Functional

I have some questions I'd like answers to

Domain is the scope of the answers

Presents interesting cognitive dilemma!

I know what I want but can't define it precisely

Two major functional steps

Data Acquisition

- The data from a universe aka the Web that can help me answer my questions

Ontological Specification

- The classes and fields that will help me unambiguously represent questions and interpret answers

In practice...

...investigators think of a domain as a **tri-faceted** combination of:

1. Questions
2. Entity types (a **shallow ontology**)
 - Ad, Posting Date, Title, Content, Phone, Email, Review ID, Social Media ID, Price, Location, Service, Hair Color, Eye Color, Ethnicity, Weight, Height**
3. Examples/Annotations

Crawling Challenges

Scale, cost, speed

DNS, fetching, parsing/extracting, memory/disk

Errors, redirects, localization

Need sophisticated software

Deep web, forms, dynamic pages, infinite scrolling

Identify and fill in forms, render pages while crawling (headless browser)

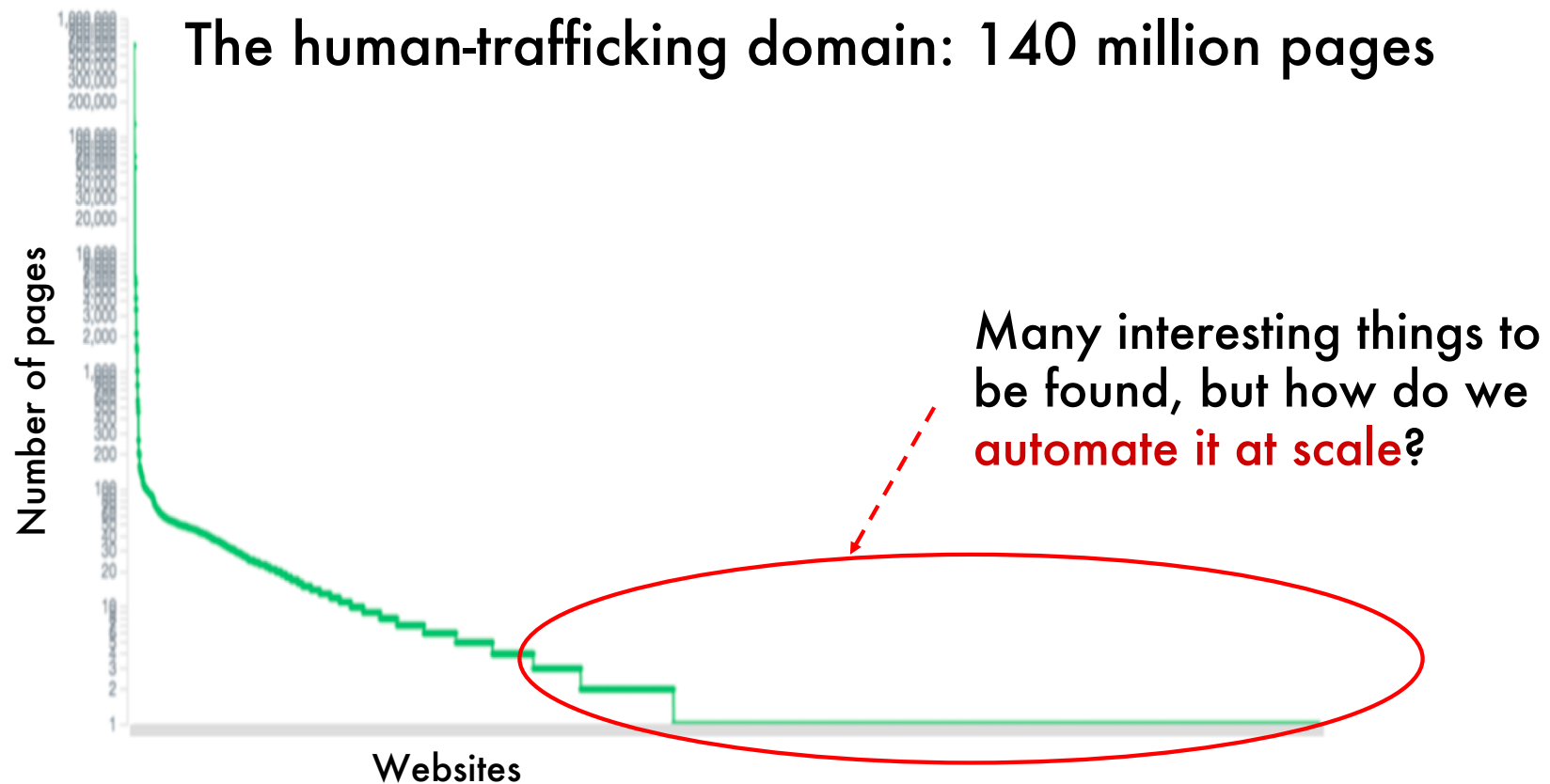
Counter-crawling measures

Login, captchas, traps, fake errors, banning

Freshness and deduplication

Identify and re-crawl new content

Domains have a long tail



Schema-agnostic Knowledge Base Querying

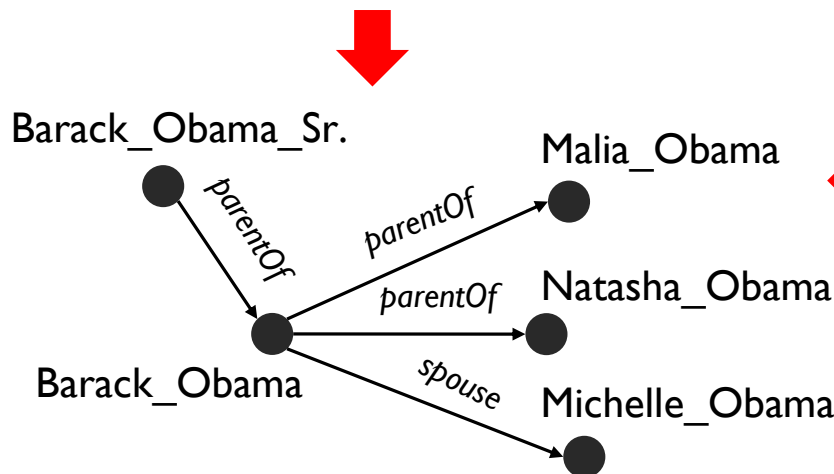
Yu Su

University of California, Santa Barbara

Structured Query: RDF + SPARQL

Triples in an RDF graph

Subject	Predicate	Object
Barack_Obama	parentOf	Malia_Obama
Barack_Obama	parentOf	Natasha_Obama
Barack_Obama	spouse	Michelle_Obama
Barack_Obama_Sr.	parentOf	Barack_Obama



SPARQL query

```
SELECT ?x WHERE
{
  Barack_Obama_Sr. parentOf ?y .
  ?y parentOf ?x .
}
```

Answer

```
<Malia_Obama>
<Natasha_Obama>
```


Why Structured Query Falls Short?

Knowledge Base	# Entities	# Triples	# Classes	# Relations
Freebase	45M	3B	53K	35K
DBpedia	6.6M	13B	760	2.8K
Google Knowledge Graph*	570M	18B	1.5K	35K
YAGO	10M	120M	350K	100
Knowledge Vault	45M	1.6B	1.1K	4.5K

* as of 2014

It's more than large: High heterogeneity of KBs

***If it's hard to write SQL on simple relational tables,
it's only harder to write SPARQL on large knowledge
bases***

Even harder on automatically constructed KBs with a loosely-defined schema

Not Everyone Can Program...



“find all patients diagnosed with eye tumor”

```
WITH Traversed (cls,syn) AS (  
  (SELECT R.cls, R.syn  
   FROM XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/properties  
    [property/name/text()='Synonym' and  
    property/value/text()='Eye Tumor']  
    /property[name/text()='Synonym']/value'  
   COLUMNS  
    cls CHAR(64) PATH './parent::*/  
    /parent::*/  
    parent::*/name',  
    tgt CHAR(64) PATH '.') AS R)  
UNION ALL  
  (SELECT CH.cls, CH.syn  
   FROM Traversed PR,  
    XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/definingConcepts/  
    concept[./text()=$parent]/parent::*/  
    parent::*/  
    properties/property[name/text()='Synonym']/value'  
   PASSING PR.cls AS "parent"  
   COLUMNS  
    cls CHAR(64) PATH './parent::*/  
    parent::*/  
    parent::*/parent::*/name',  
    syn CHAR(64) PATH '.') AS CH))  
SELECT DISTINCT V.*  
FROM Visit V  
WHERE V.diagnosis IN  
  (SELECT DISTINCT syn FROM Traversed)
```

NCIthesaurus

“Semantic queries by example”,

Information Sciences Lim et al., EDBT (2014)

USC Viterbi

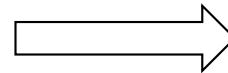
In Pursue of Efficiency

find all patients diagnosed with eye tumor



```
WITH Traversed (cls, syn) AS (  
  (SELECT R.cls, R.syn  
   FROM XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/properties  
    [property/name/text()='Synonym' and  
    property/value/text()='Eye Tumor']  
    /property[name/text()='Synonym']/value  
   COLUMNS  
    cls CHAR(64) PATH '/parent:*/parent:*/  
    /parent:*/name',  
    syn CHAR(64) PATH '/') AS R)  
  UNION ALL  
  (SELECT CH.cls, CH.syn  
   FROM Traversed PR,  
   XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/definingConcepts/  
    concept[isect()='Superior']/parent:*/  
    properties/property[name/text()='Synonym']/value  
   PASSING PR.cls AS 'parent'  
   COLUMNS  
    cls CHAR(64) PATH '/parent:*/  
    /parent:*/parent:*/name',  
    syn CHAR(64) PATH '/') AS CH))  
SELECT DISTINCT V.  
FROM Test V  
WHERE V.Diagnosis IN  
(SELECT DISTINCT syn FROM Traversed)
```

Seconds



Days



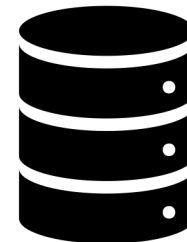
In Pursue of Efficiency

find all patients diagnosed with eye tumor



Schema-agnostic
Querying

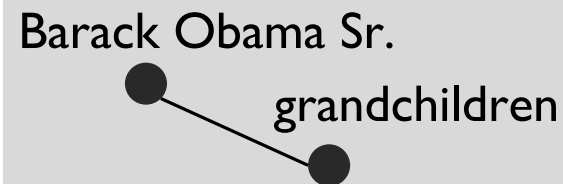
```
WITH Traversed (cls, syn) AS (  
  (SELECT R.cls, R.syn  
   FROM XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/properties  
    [property/name/text()='Synonym' and  
    property/value/text()='Eye Tumor']  
    /property[name/text()='Synonym']/value'  
   COLUMNS  
    cls CHAR(64) PATH './parent:*/parent:*/  
    /parent:*/name',  
    syn CHAR(64) PATH './parent:*/parent:*/  
    /parent:*/name') AS R)  
 UNION ALL  
  (SELECT CH.cls, CH.syn  
   FROM Traversed PR,  
   XMLTABLE ('Document("Thesaurus.xml")  
    /terminology/conceptDef/definingConcepts/  
    concept[isect()='Superior']/parent:*/  
    properties/property[name/text()='Synonym']/value'  
   PASSING PR.cls AS 'parent'  
   COLUMNS  
    cls CHAR(64) PATH './parent:*/parent:*/  
    /parent:*/name',  
    syn CHAR(64) PATH './parent:*/parent:*/  
    /parent:*/name') AS CH))  
 SELECT DISTINCT V.  
 FROM Test V  
 WHERE V.Diagnosis IN  
  (SELECT DISTINCT syn FROM Traversed)
```



Schema-agnostic KB Querying

“Barack Obama Sr. grandchildren”

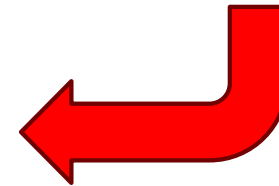
Keyword query: query like search engine



Graph query: add a little structure

“Who are Barack Obama Sr.’s grandchildren?”

Natural language query: the holy grail



Tutorial Outline

Introduction

Part I: Domain-specific KB Construction

Lunch Break

Part II: Schema-agnostic KB Querying

Summary & Future Directions