Constructing Domain Specific Knowledge Graphs

Mayank Kejriwal, Craig Knoblock and Pedro Szekely

Information Sciences Institute,

University of Southern California

Domain-specific search (DSS)

The Massive YouTube Ecosystem

Y-Clad's Hidden Gem

Venues & Event Spaces, Jewelry, Accessories

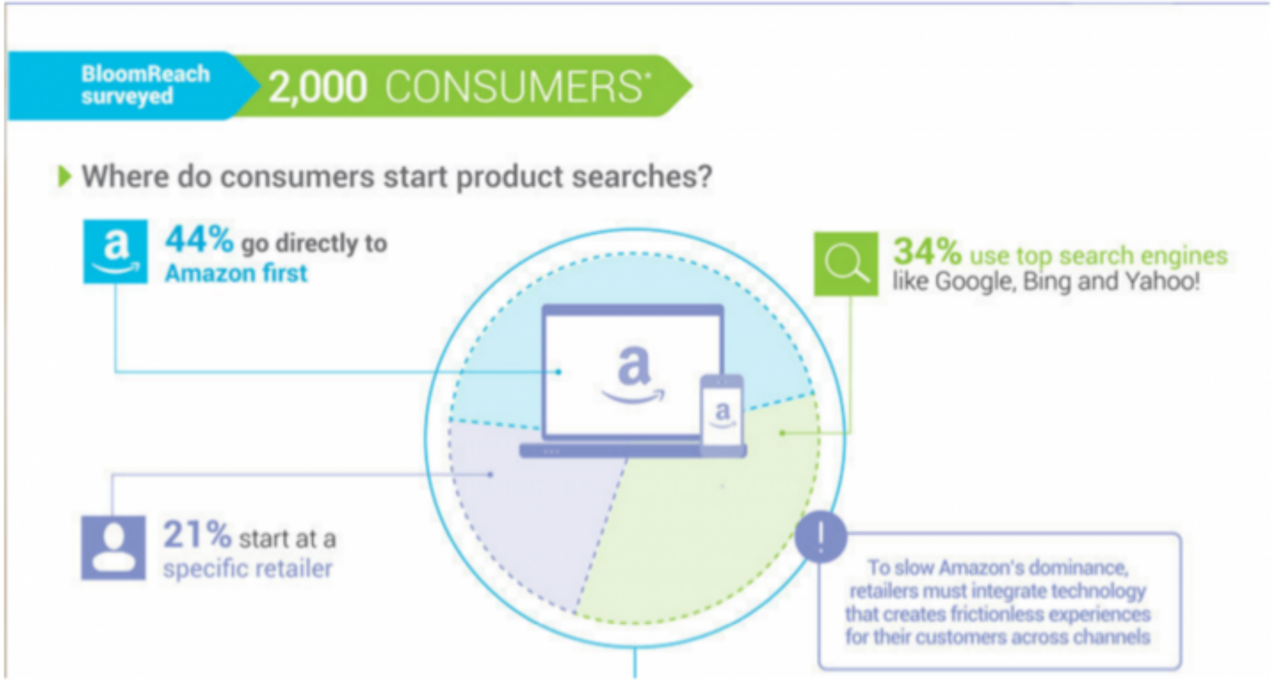
Enjoy Free Shipping on All Orders Over $150!

From the business

Specialties

Reg. Price	Discount	Savings
$100	25%	$25

Buy Gift Certificate


BloomReach surveyed 2,000 CONSUMERS*

Where do consumers start product searches?

44% go directly to Amazon first

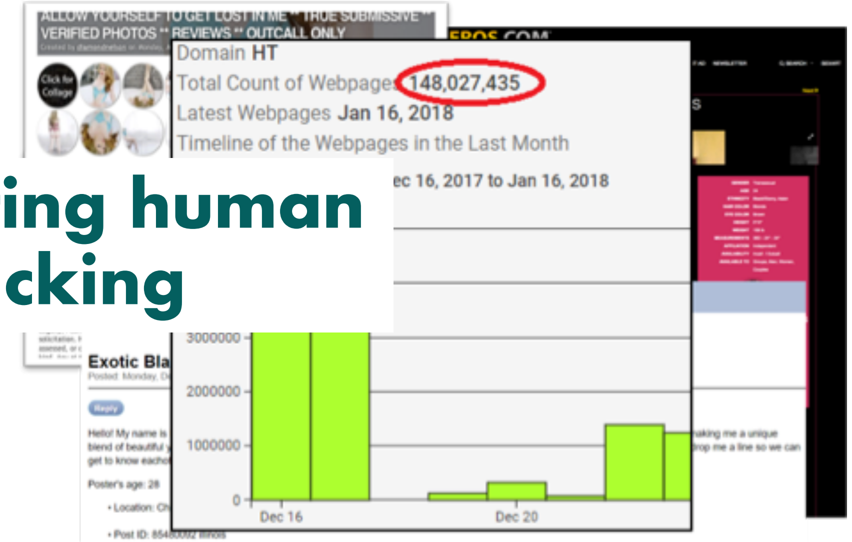
34% use top search engines like Google, Bing and Yahoo!

21% start at a specific retailer

To slow Amazon's dominance, retailers must integrate technology that creates frictionless experiences for their customers across channels

source: https://photos.prnewswire.com/prnfull/20151006/274273-INFO

Emerging opportunities for DSS


ALLOW YOURSELF TO GET LOST IN ME ** TRUE SUBMISSIVE ** VERIFIED PHOTOS ** REVIEWS ** OUTCALL ONLY

Domain HT

Total Count of Webpages 148,027,435

Latest Webpages Jan 16, 2018

Timeline of the Webpages in the Last Month

ec 16, 2017 to Jan 16, 2018

Exotic Bla

Posted: Monday, D

Hello! My name is … blend of beautiful y… get to know eacho… aking me a unique … rop me a line so we can

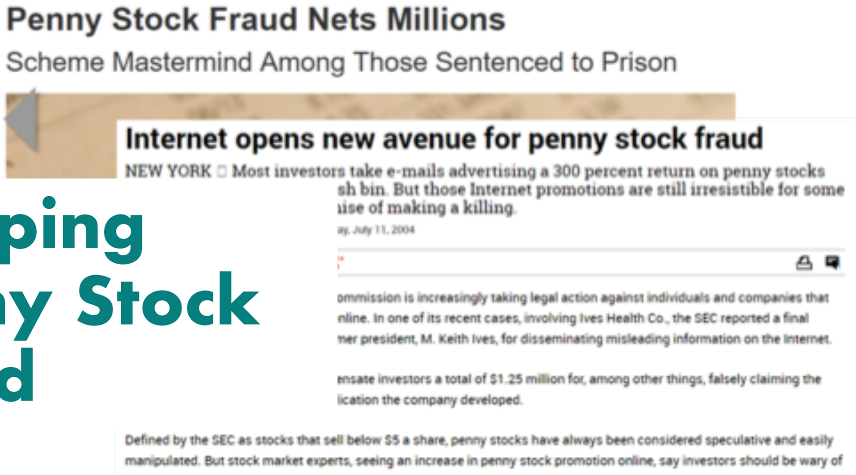
Poster's age: 28

- Location: Ch
- Post ID: 85480092 illinois

Fighting human trafficking

Terrifying interactive map shows global cyber attacks happening in real time

Predicting cyberattacks


Penny Stock Fraud Nets Millions

Scheme Mastermind Among Those Sentenced to Prison

Internet opens new avenue for penny stock fraud

NEW YORK □ Most investors take e-mails advertising a 300 percent return on penny stocks … sh bin. But those Internet promotions are still irresistible for some … ise of making a killing.

ay, July 11, 2004

ommission is increasingly taking legal action against individuals and companies that … nline. In one of its recent cases, involving Ives Health Co., the SEC reported a final … ner president, M. Keith Ives, for disseminating misleading information on the Internet.

ensate investors a total of $1.25 million for, among other things, falsely claiming the … lication the company developed.

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

Stopping Penny Stock Fraud

GEOPOLITICAL FORECASTING CHALLENGE

Can you create a method to predict future

SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION

PHILIP E. TETLOCK, DAN GARDNER

5 TRAITS

perforecasters begin by gathering as much information possible.

perforecasters nurture and develop the habit of thinking terms of probabilities when exploring the likelihood of ecific events.

recasting improves when individuals work in teams.

perforecasters ensure that they are regularly keeping re of their projections.

5. The most successful forecasters are willing to admit error and quickly change course on their projections.

Accurate geopolitical forecasting

DARPA/IARPA programs

Domain HT
Total Count of Webpages 148,027,435
Webpages from Dec 16, 2017 to Jan 16, 2018
Fighting human trafficking

Terrifying interactive map shows global cyber attacks happening in real time

Penny Stock Fraud Nets Millions

Scheme Mastermind Among Those Sentenced to Prison

Internet opens new avenue for penny stock fraud

NEW YORK □ Most investors take e-mails advertising a 300 percent return on penny stocks and relega… …istible for some beginners…

GEOPOLITICAL FORECASTING CHALLENGE
SUPERFORECASTING
THE ART AND SCIENCE OF PREDICTION
PHILIP E. TETLOCK
DAN GARDNER
Accurate geopolitical forecasting

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?

Indicator Mining

List all ads that have high probability of movement

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

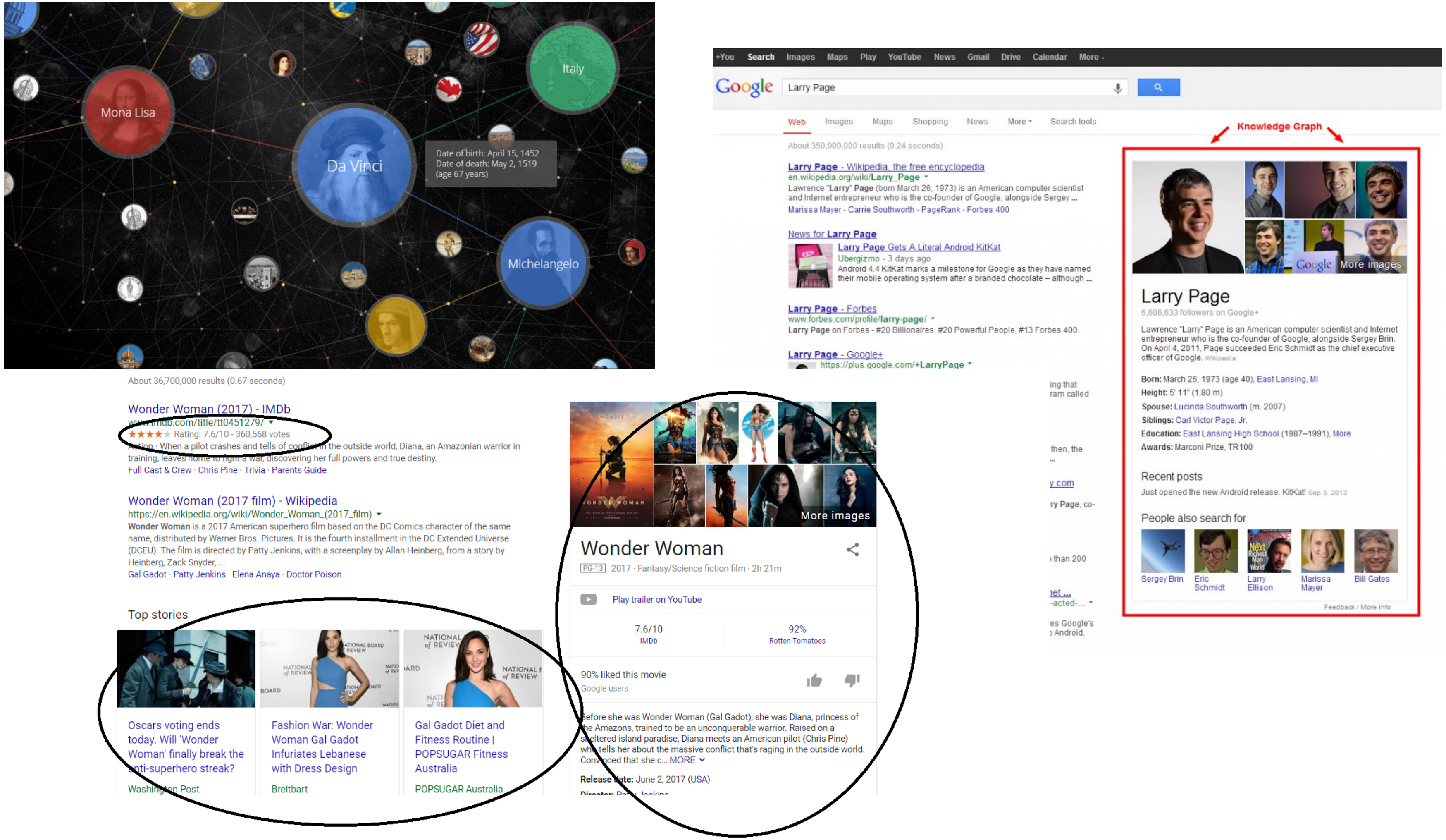
Aggregations/Lists

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

Dossier Generation

Collect and show me all information on the phone number 7075610282

Google Knowledge Graph



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 KitKat
Ubergizmo - 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

ing that
ram called

then, the
...

y.com
ry Page, co-

than 200

net ...
-acted-...

es Google's
Android.

Larry Page

6,606,633 followers on Google+

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. Wikipedia

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. KitKat! Sep 3, 2013

People also search for

Sergey Brin · Eric Schmidt · Larry Ellison · Marissa Mayer · Bill Gates

Feedback / More info

About 36,700,000 results (0.67 seconds)

Wonder Woman (2017) - IMDb
www.imdb.com/title/tt0451279/
★★★★★ Rating: 7.6/10 - 360,568 votes
Action · When a pilot crashes and tells of conflict in the outside world, Diana, an Amazonian warrior in training, leaves home to fight a war, discovering 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

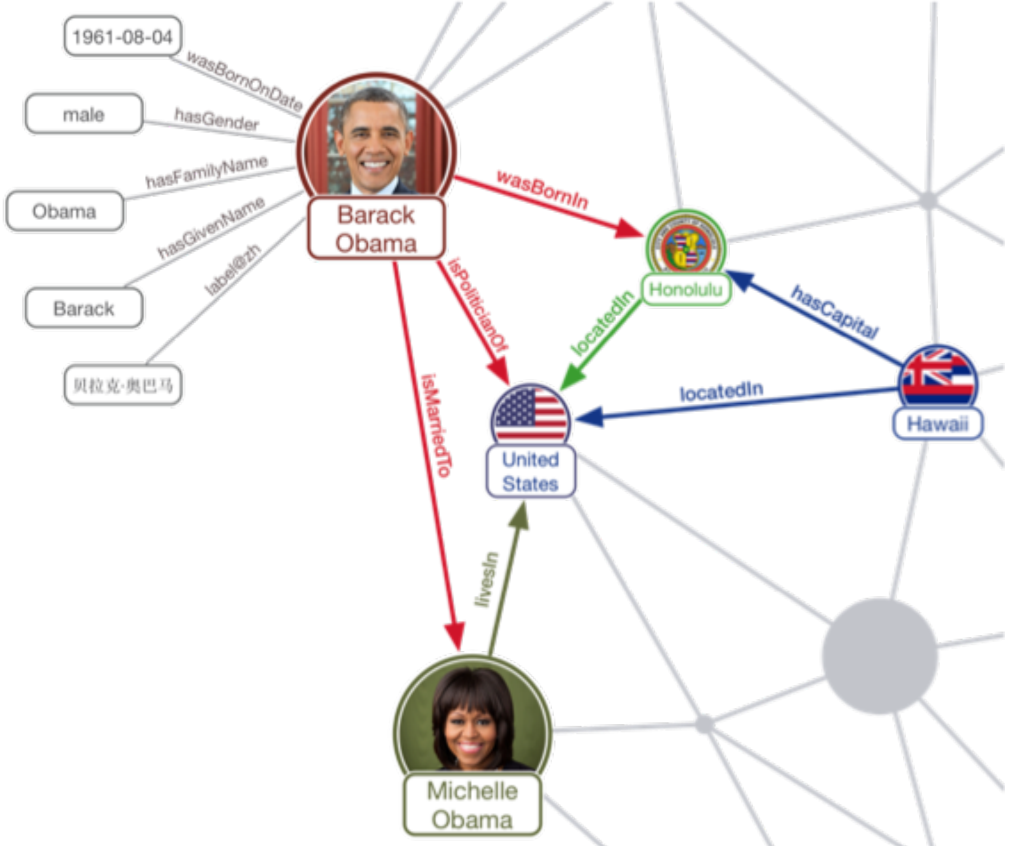
Before she was Wonder Woman (Gal Gadot), she was Diana, princess of the Amazons, trained to be an unconquerable warrior. Raised on a sheltered island paradise, Diana meets an American pilot (Chris Pine) who tells her about the massive conflict that's raging in the outside world. Convinced that she c...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

Release date: June 2, 2017 (USA)
Director: Patty Jenkins

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

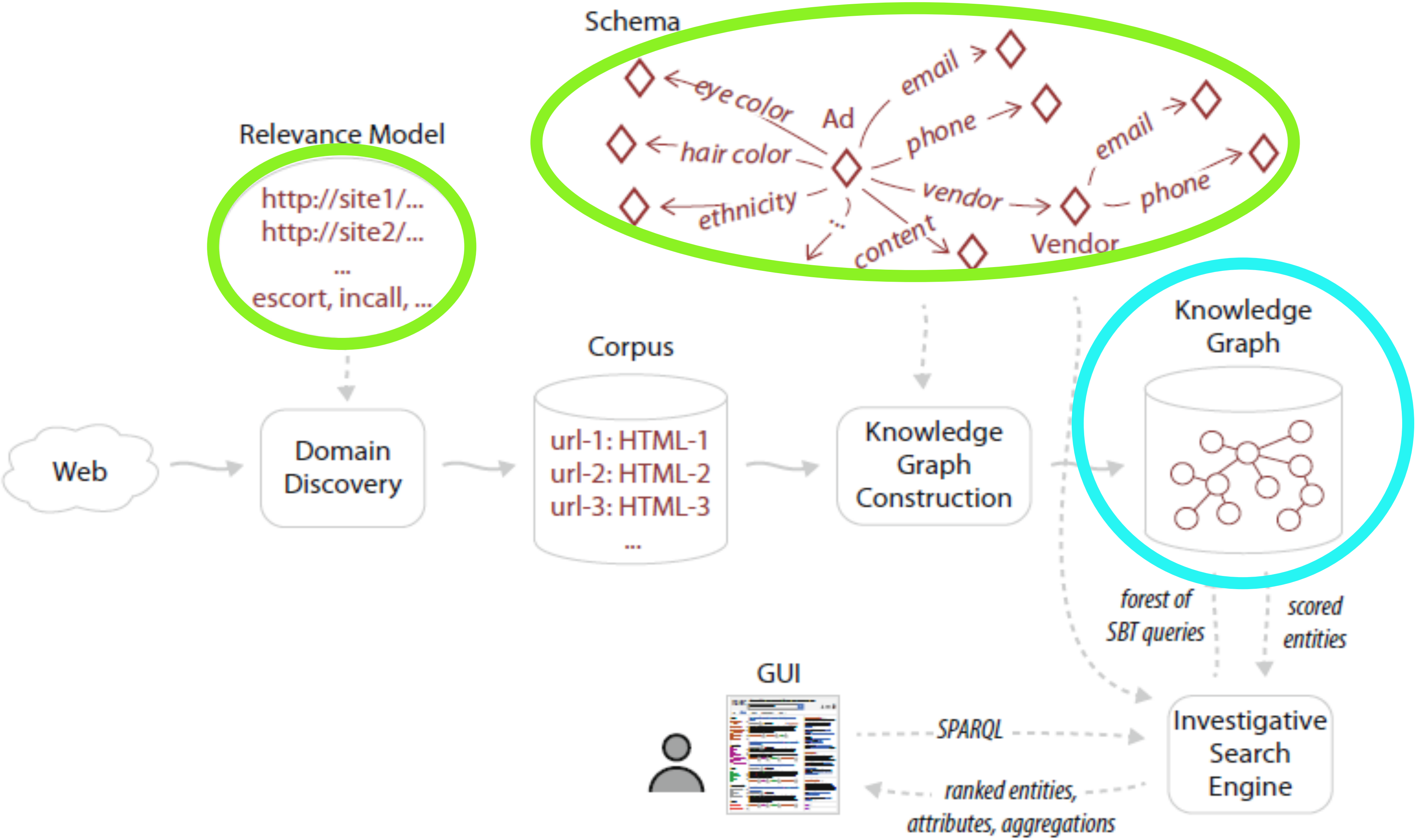
(Barack Obama, wasBornOnDate, 1961-08-04),
(Barack Obama, hasGender, male),
…
(Hawaii, hasCapital, Honolulu),
…
(Michelle Obama, livesIn, United States)



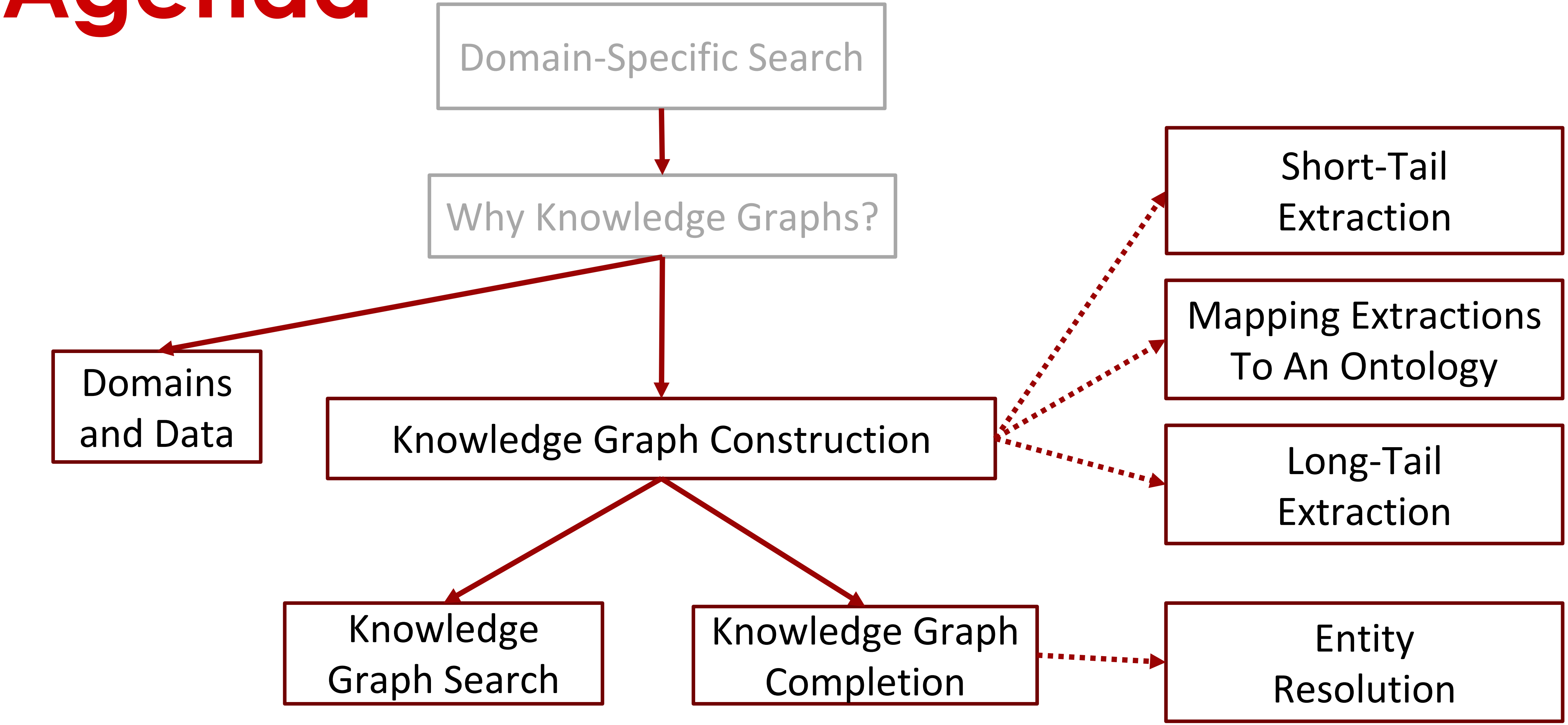
General Search	Google Knowledge Graph
DSS	Domain-Specific Knowledge Graphs

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

Knowledge Graphs for DSS



Agenda



Domain-Specific Search

Why Knowledge Graphs?

Domains and Data

Knowledge Graph Construction

Short-Tail Extraction

Mapping Extractions To An Ontology

Long-Tail Extraction

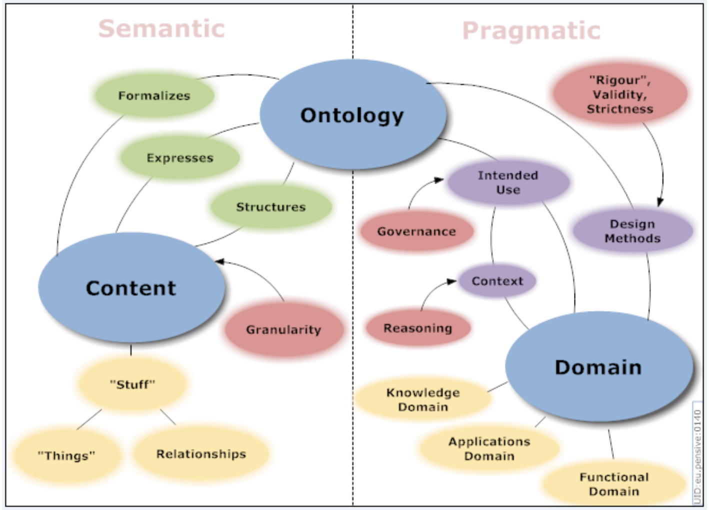
Knowledge Graph Search

Knowledge Graph Completion

Entity Resolution

What is (or even isn’t) a domain?

WHAT DID WE COVER IN CLASS LAST WEEK?
IT'S IN THE SYLLABUS.
WHAT'S YOUR LATE HOMEWORK POLICY?
IT'S IN THE SYLLABUS.
WHEN ARE YOUR OFFICE HOURS?
IT'S IN THE SYLLABUS.
HOW WILL MY GRADE BE COMPUTED?
IT'S IN THE SYLLABUS
JORGE CHAM © 2013
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

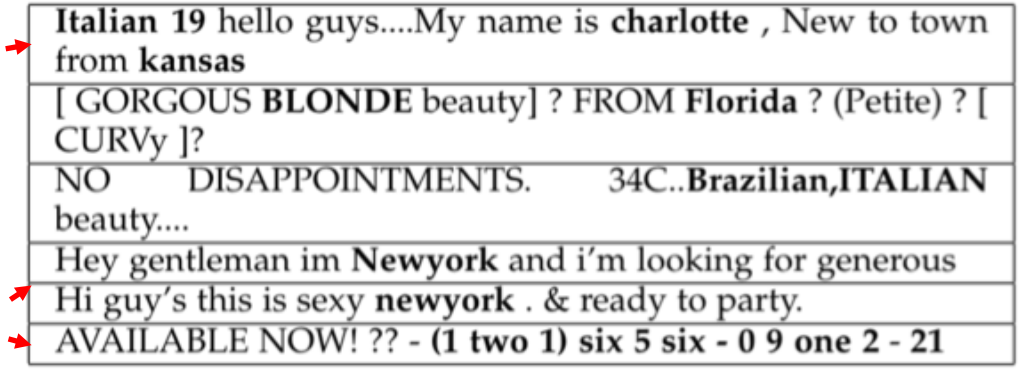
How do domain experts specify the sphere?

Examples

Ontology

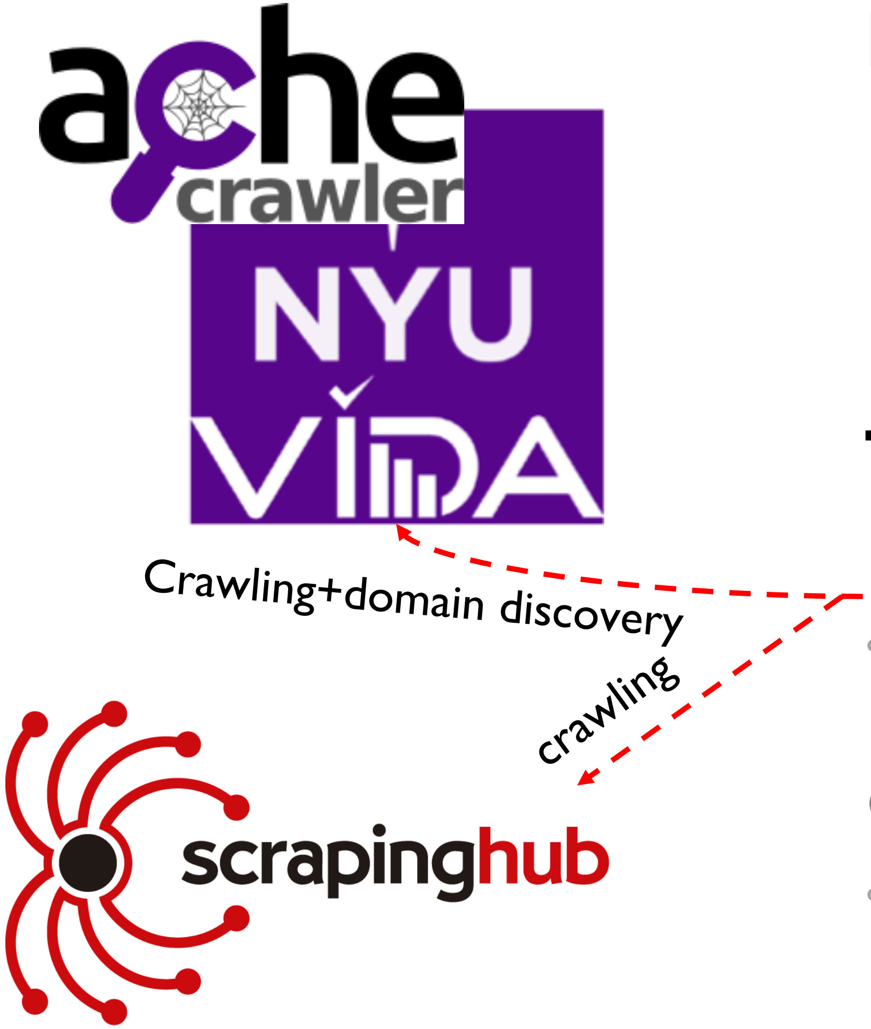
Domain-Specific Challenges

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



Italian 19 hello guys....My name is **charlotte** , New to town from **kansas**
[GORGOUS **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



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

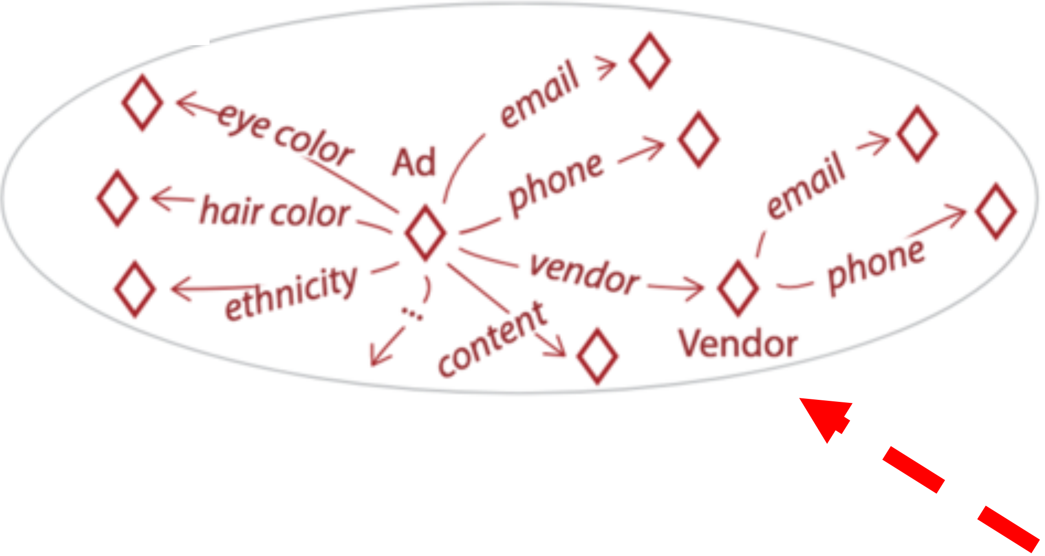
Data Acquisition

- 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

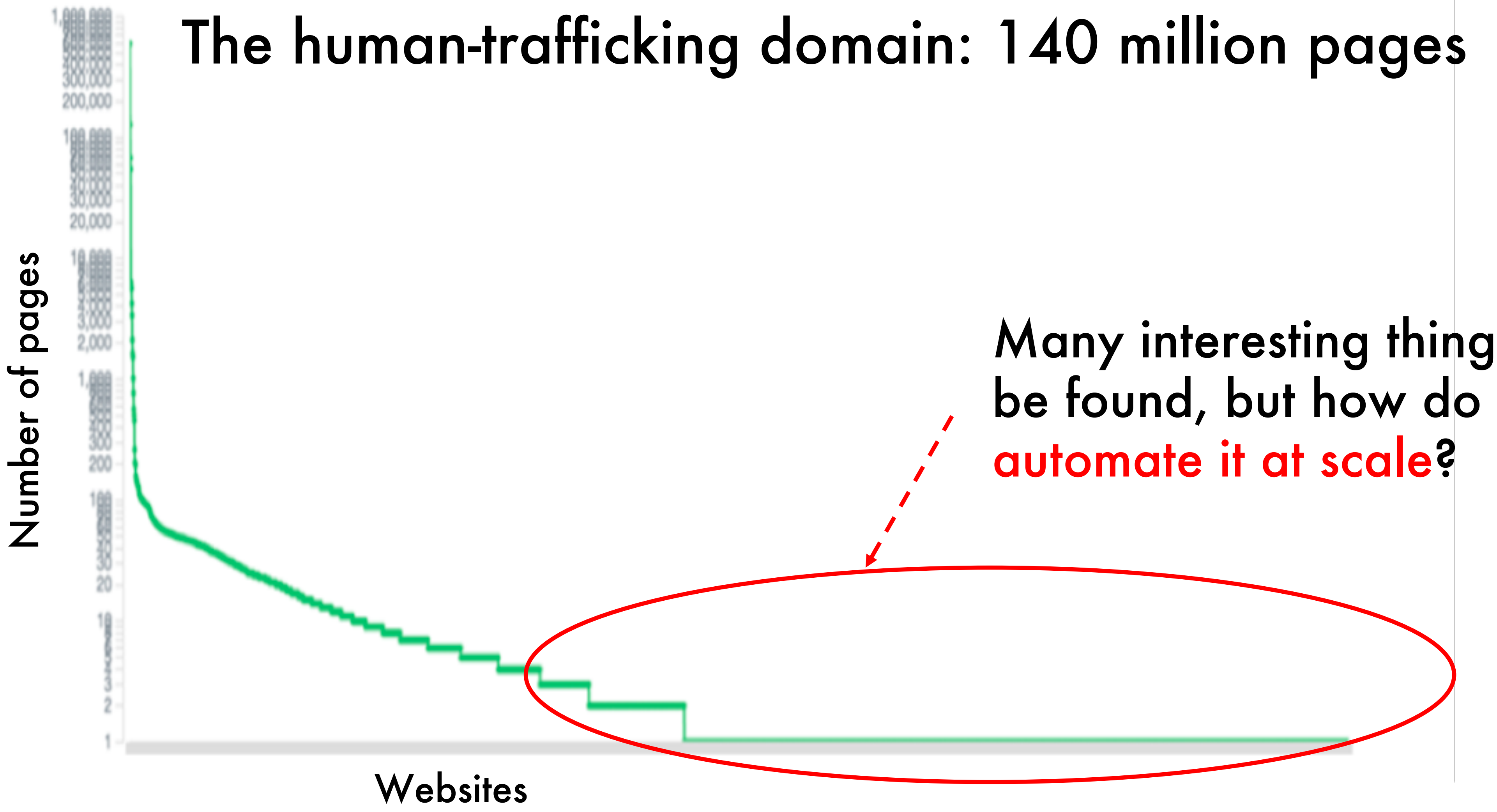
Login, captchas, trap, fake errors, banning

Freshness and deduplication

Identify and re-crawl new content

Domains have a long tail

The human-trafficking domain: 140 million pages



Many interesting things to be found, but how do we automate it at scale?