Data mining in unusual domains with information-rich knowledge graph construction, inference and search

Mayank Kejriwal, Pedro Szekely

Information Sciences Institute,

USC Viterbi School of Engineering

KDD2017

Agenda

Unusual domains

Knowledge graphs

KG construction

Case study

Knowledge graph completion

Entity resolution

Probabilistic soft logic

KGs in latent space

Searching knowledge graphs

Unusual domains

Disaster Assistance In Regions Using Low Resource Languages

Data

Social media and news in low resource languages
Akan, Amharic, Hausa, Tagalog, Uyghur, Wolof, Yoruba, …

Example Questions

Identify areas where people are in greatest need
Identify threats to relief personnel
Characterize the evolution of the disaster

Technical Challenges

Very noisy translation
Clustering documents according to need, location and entities
Streaming data

Identify Illegal Firearm Sales

Data

Classified ads and forums
Open and dark web

Example Questions

Identify buyers who purchase on behalf of others
Identify people who buy and sell without proper licenses
Identify vendors who illegally sell across state lines
Identify stolen firearms for sale

Technical Challenges

Most firearm sales are legal
Huge volume of pages and web sites
Unusual language model

Identify Counterfeit Electronics Vendors

Data

Online catalogs and forums
Internal sources containing suspicious vendors, shipping addresses

Example Questions

Identify clusters of companies under control of one organization
Identify fraudulent ads across vendor sites

Technical Challenges

Many pages in Chinese
Fake images on catalogs
Frequent creation of new shell vendor companies

Identify Narcotics Vendors In The Dark Web

Data

Dark web marketplace and forum pages
Open web social media sites

Example Questions

Identify dark web personas in the open web
???

Technical Challenges

Pages in multiple languages
Unusual language model
Deception, fake vendors, fake reviews
Careful concealment of identifying information

Combat Fraud In The Penny Stock Market

Data

Web pages and social media about companies in the Over The Counter (OTC) market

Example Questions

Identify in-progress pump-and-dump scams
Identify chain of shell companies and individuals involved
Identify promoters, attorneys, etc.
Identify evidence of unlawful behavior (e.g., false statements in promotional materials)

Technical Challenges

Pump-and-dump scams are carefully planned and controlled to look legal
Wide variety of page genres: company profiles, press releases, promotions, financial data

Identify Human-Trafficking Victims & Prosecute Traffickers

Data

100,000,000 escort ads published on the web
50,000 images

Example Questions

Identify all ads for an escort given soft identifies
Tag ads as high-risk for human trafficking
Identify stables of escorts controlled by one trafficker
Identify fake images
Identify networks of phone numbers

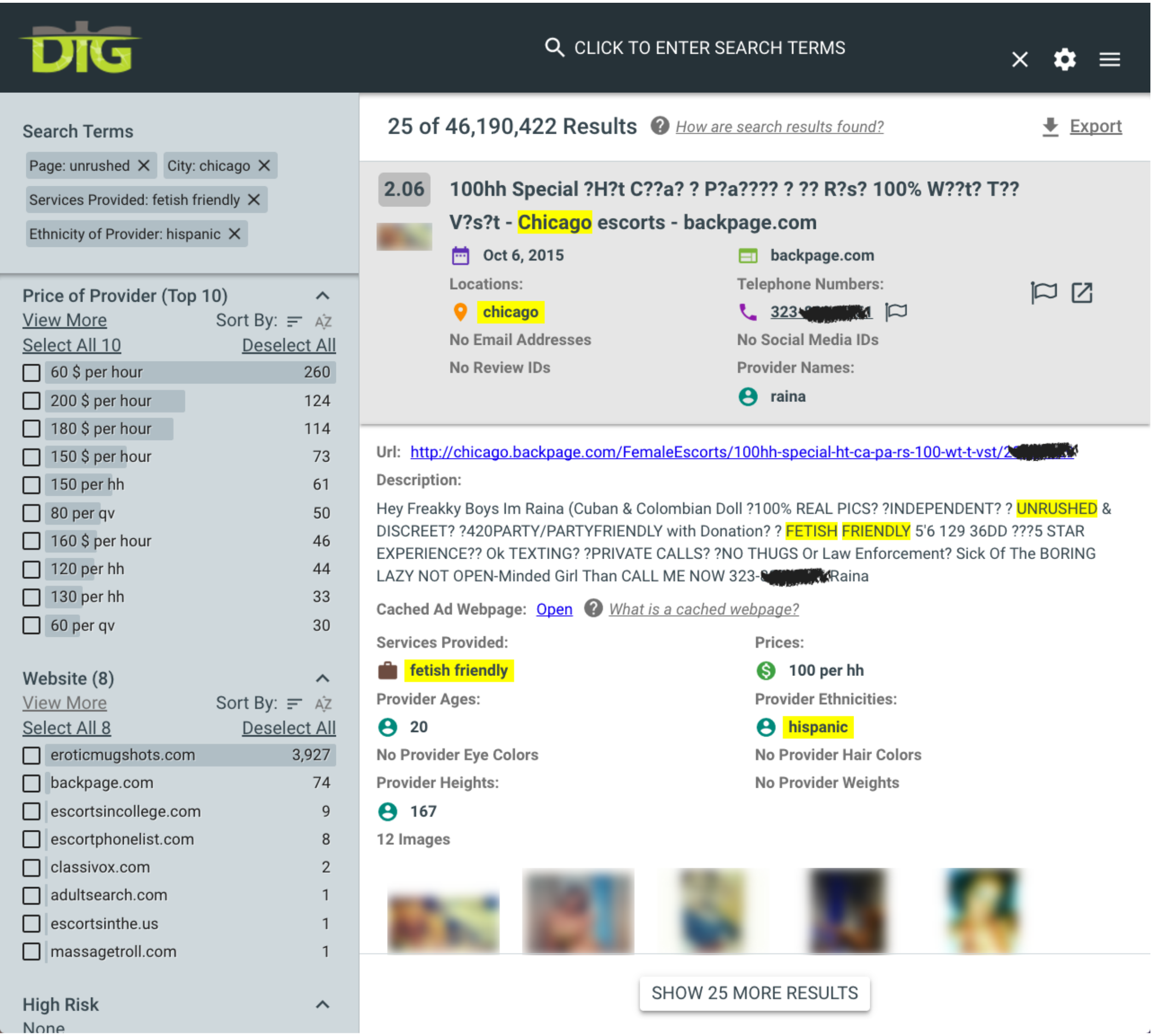
Technical Challenges

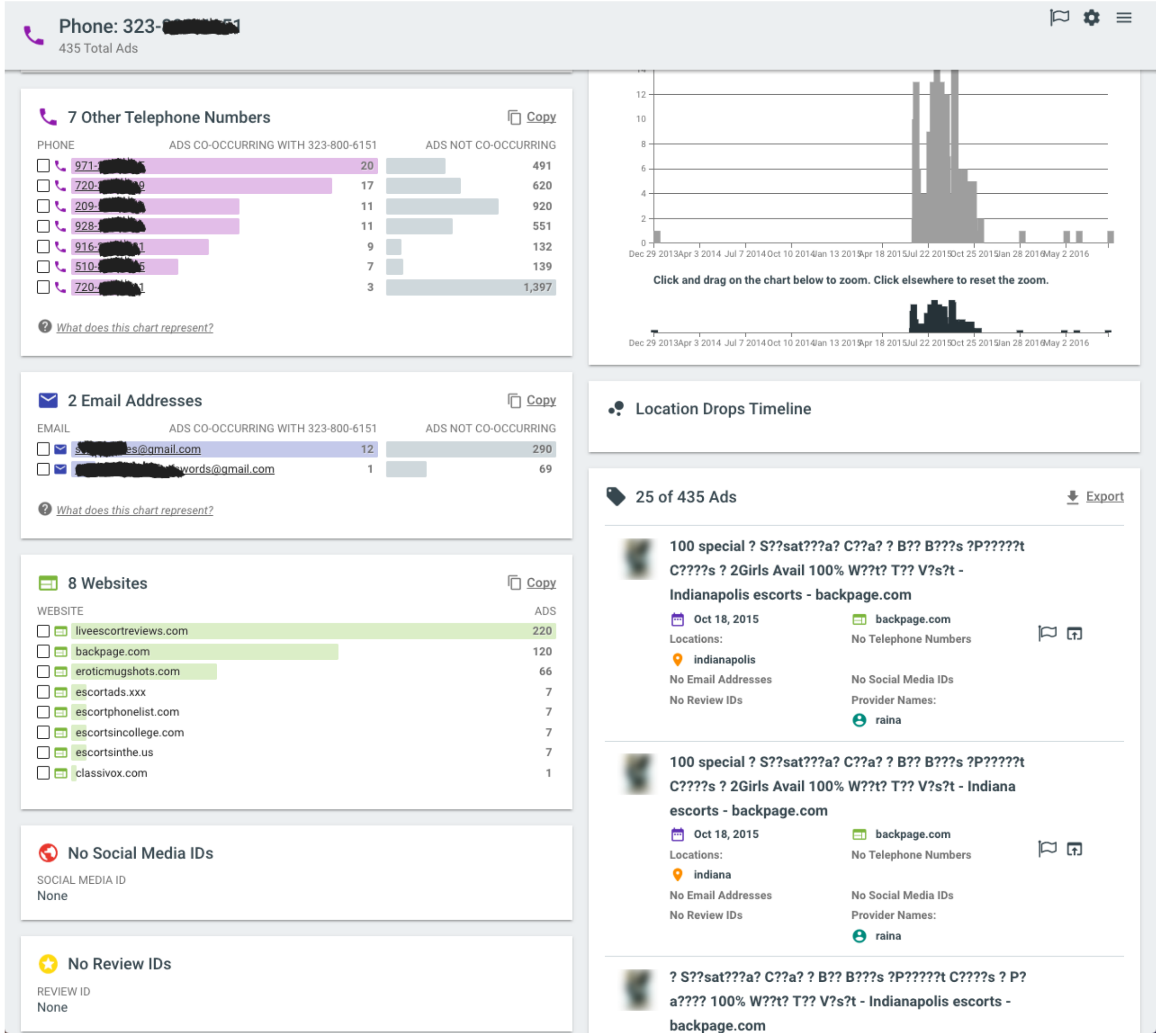
Noise, obfuscation, deception and large size
Unique language model
Long-tailed set of sources
Lack of identifiers and reference data

Work on unusual domains has significant

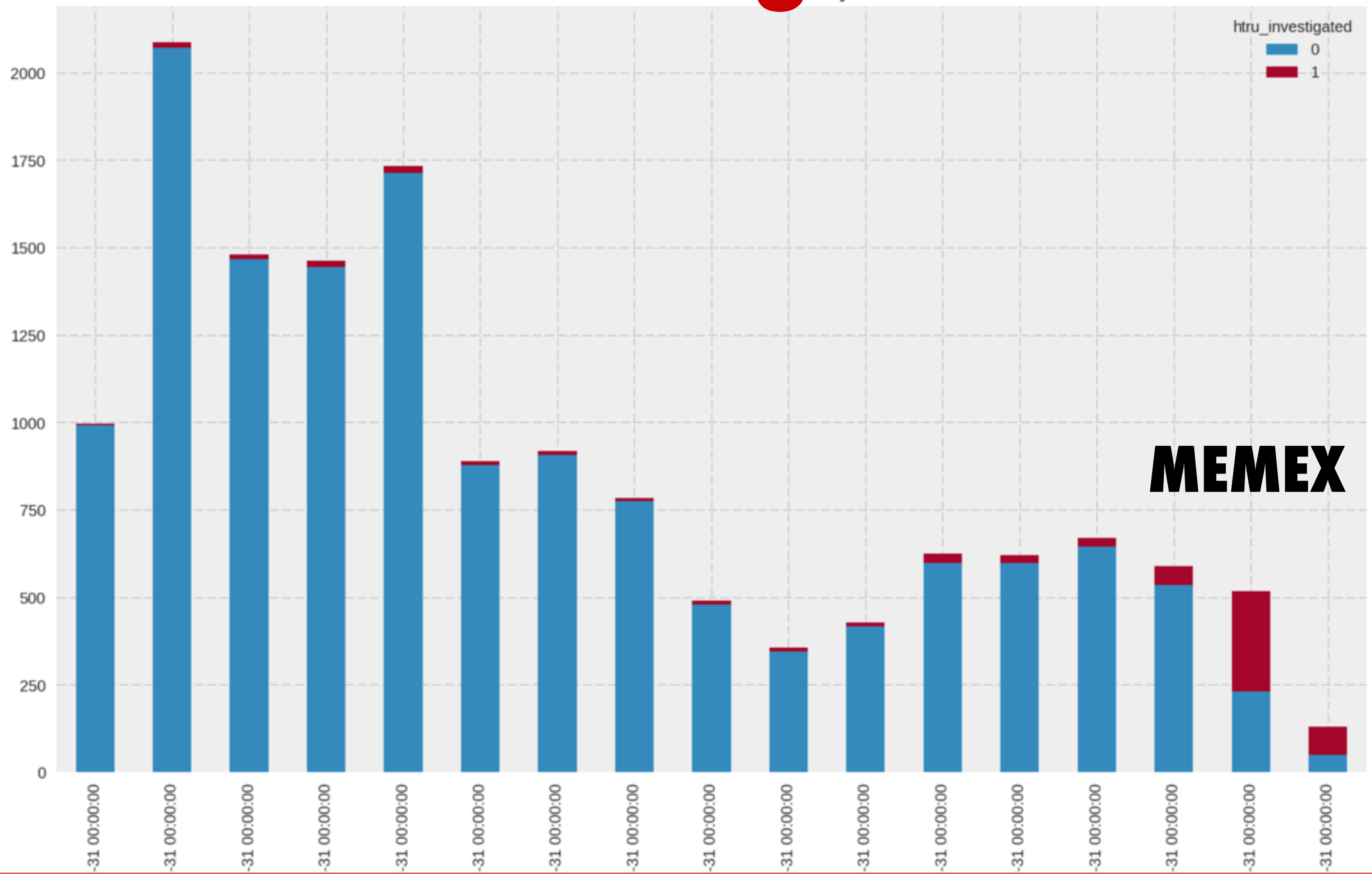
social impact

Find Locations Where An Escort Was Advertised





Number Of HT Investigations Increased



FROM <1% TO 62%

SFGATE LOCAL NEWS SPORTS BUSINESS A&E FOOD LIVING TRAVEL REAL ESTATE C

Man sentenced to 97 years in human trafficking case

By Vivian Ho Updated 7:07 am, Friday, April 22, 2016

123

97-TO-LIFE

Deshawn Birden

- Pimp
- Present in victim "B" case
- Las Vegas, Oakland, SF

Brazil Harris

- Pimp
- Los Angeles, San Diego, SF
- Arr. 4/2016 for possession for sale
- Co-de said she was forced

Jermaine Fulgham

- Self-admitted pimp
- Assisted Geeter in jail
- Present for victim 1 "Z"

Jhontay Wills

- Pimp
- Prior arrest for 653.23 on Capp St.
- Friend of Geeter
- Was assisting in bringing in character witnesses (minors)

4 ACCOMPLICES

Victims

Candise Burns

- Mother of Geeter's first child
- Reached out post verdict to SFDA about victimization
- Sgt. Flores has been in contact

Yasmine Malone

- Missing 15 yo Victimized in 12/2014
- MH issues; assaulted SFSD officer
- Birden/Walls in contact post-trial and assisting in exploitation (2016)

Shakarri Miller

- Missing 15 yo Victimized
- Arrested in Broadmoor, CA with Geeter

Courtney Taylor

- Missing 17 yo Victimized
- Arrested in Garden Grove, CA with Geeter and Birden

Zurline Hurst

Belinda Carson

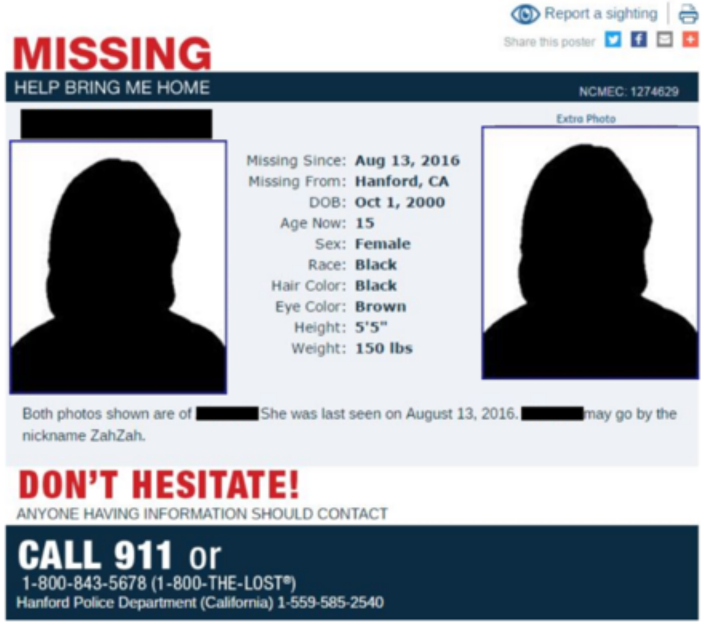
+20 additional, possible victims of ring

>25 VICTIMS

Case Study 4: Recovery of Missing Juvenile

Problem: Missing juvenile via NCMEC who reached out to friends, saying she was being trafficked throughout CA

Outcome: A. Green was recovered in Atlanta, Georgia by the FBI after I provided NCMEC and local law enforcement with her most recent sex ad and location.



CALIFORNIA VICTIM

Usual	Unusual
Good English	Jargon, ungrammatical
Mostly correct information	Obfuscation & deception
Large reference datasets	No reference datasets
Small number of sources	Long tail
Sources readily available	Ephemeral pages, dark web, anti-crawling measures

Two guiding questions

How do we model, and represent knowledge in, unusual domains?

Are there general lessons to be learned across domains?

How do we do search and inference in unusual domains?

how to **represent** KGs?

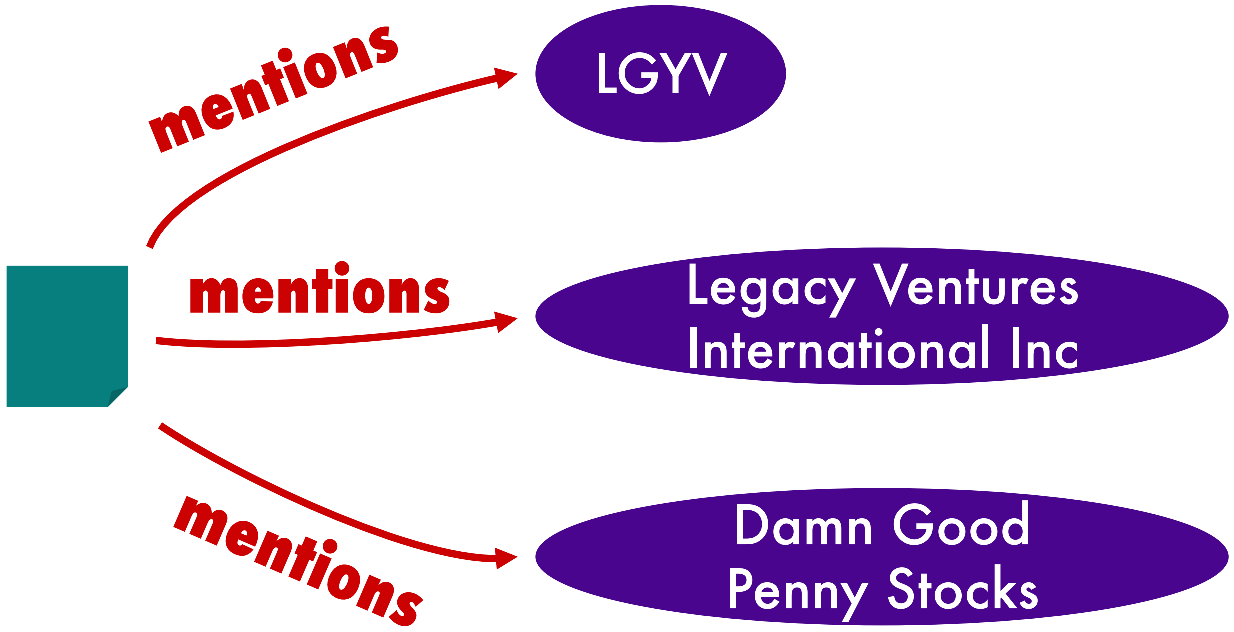
KG Definition

a directed, labeled multi-relational graph representing facts/assertions as **triples**

(h, r, t)	head entity, relation, tail entity
(s, p, o)	subject, predicate, object

Simplest Knowledge Graph

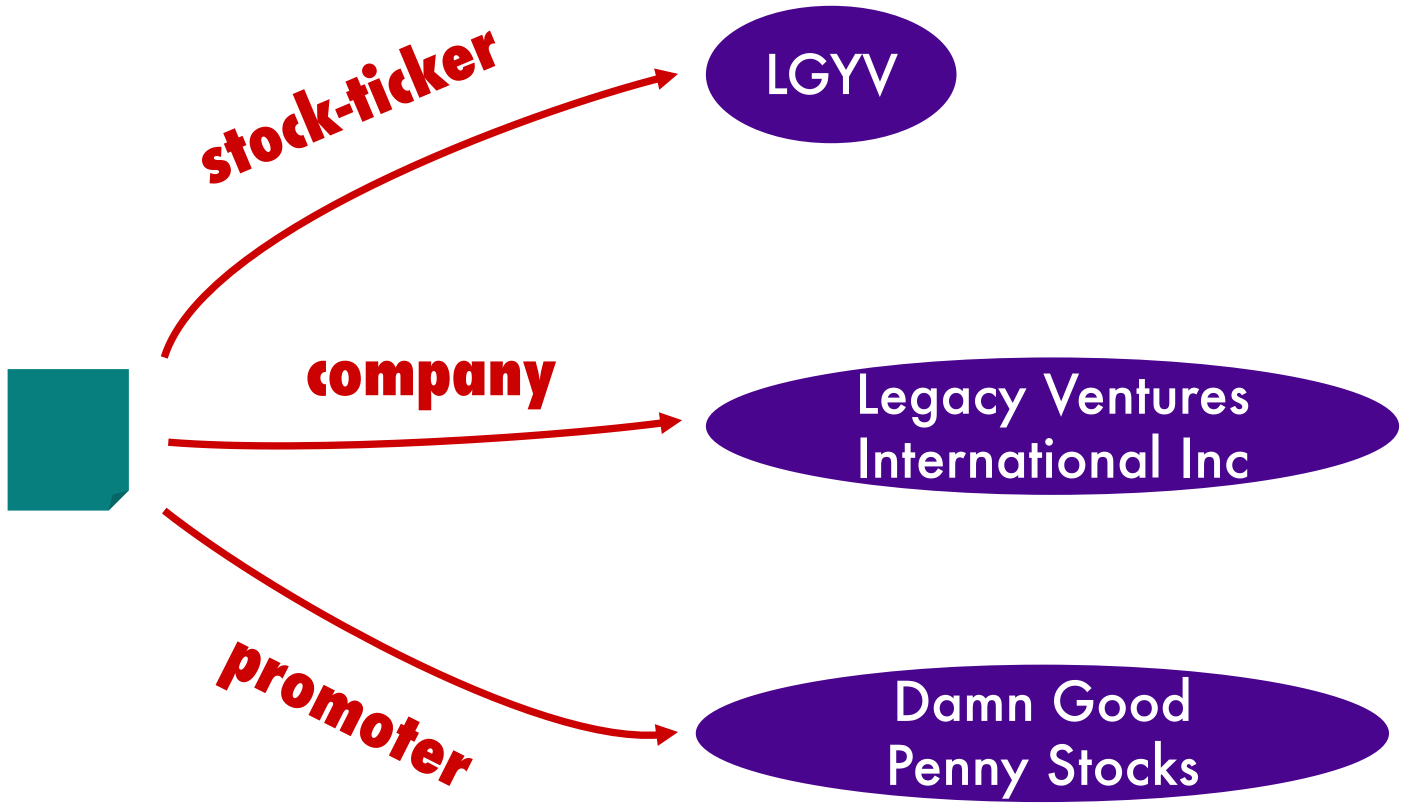
Entities



Easiest to build

Simple, But Useful KG

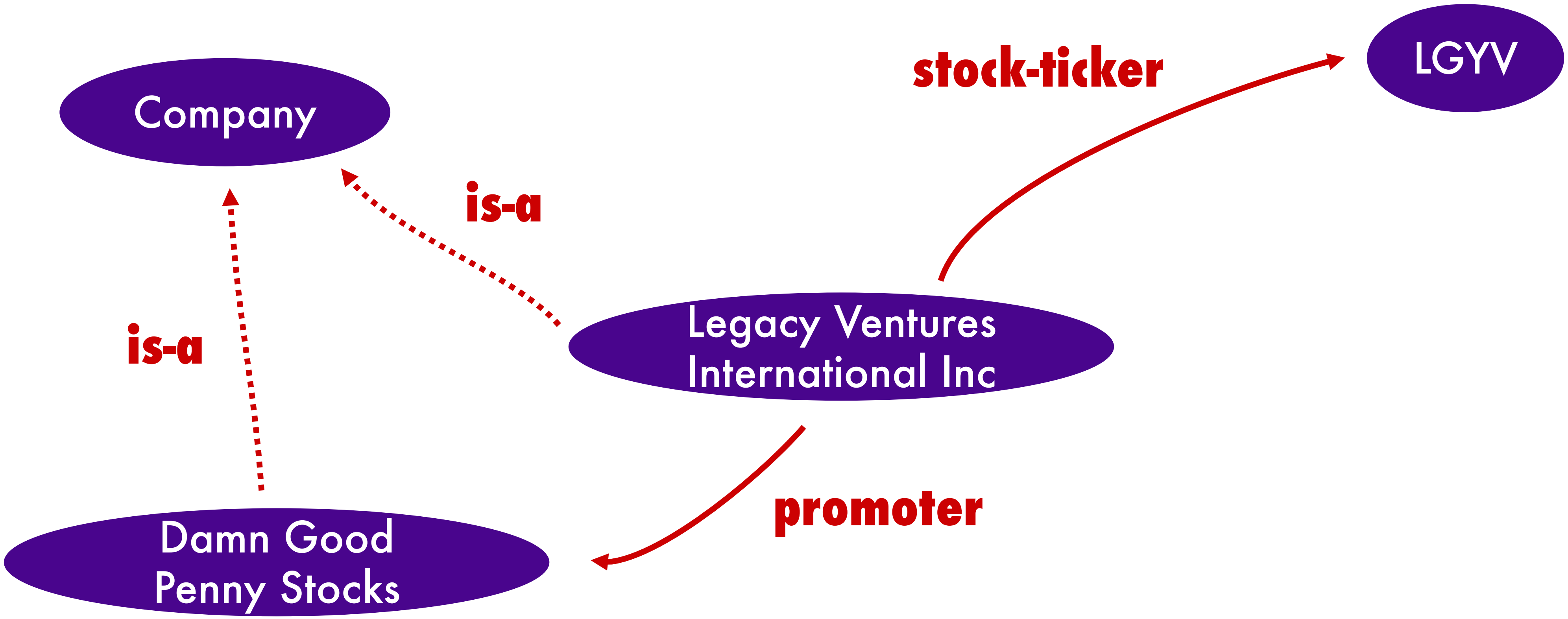
Entities + properties



"Easy" to build

Semantic Web KG (RDF/OWL)

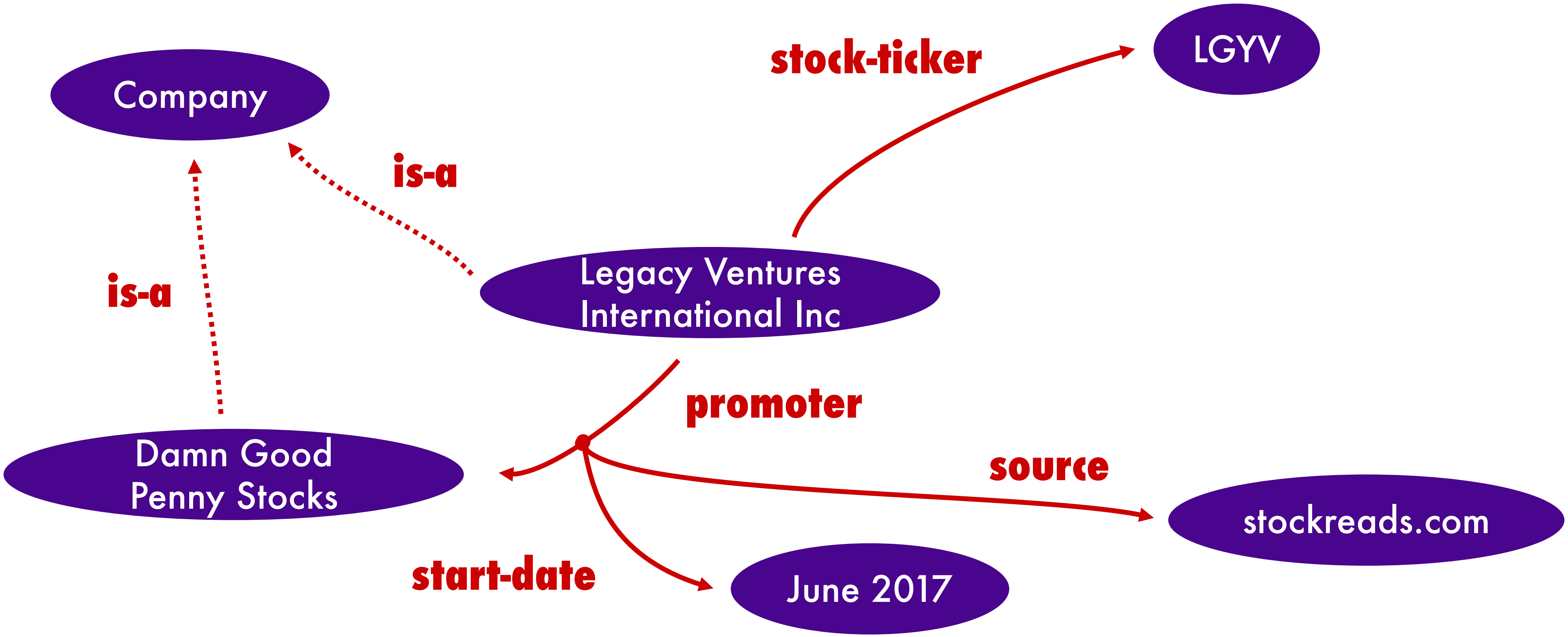
Entities + properties + classes



Very hard to build

"Ideal" KG

Entities + properties + classes + qualifiers



Very very hard to build

How about **Ontologies?**

Ontologies?

Domain ontologies don't exist for unusual domains

Build new one or extend existing one?

Deep or shallow?

Ontologies?

Goal is to help users solve problems

Let users tell you what entities matter to them

Build the shallowest, simplest ontology that captures the entities users care about

Ontologies?

Domain ontologies don't exist for unusal domains

Build new one or extend existing one?

Deep or shallow?

Where to Store KGs?

Serializing Knowledge Graphs

Resource Description Framework (RDF)

Database (triple store): AllegroGraph, Virtuoso,
Query: SPARQL (SQL-like)

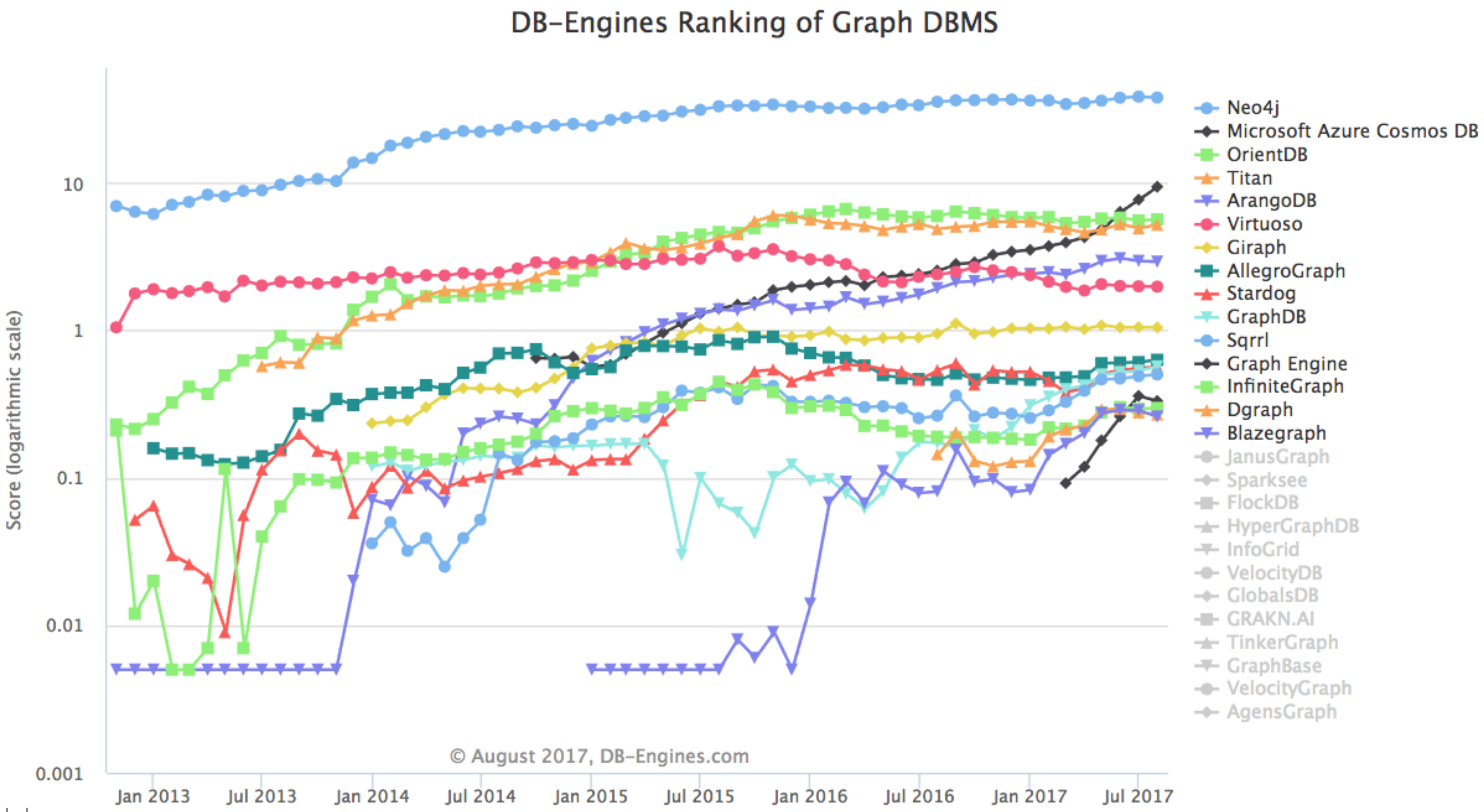
Key-Value, Document Stores

Data model: Node-centric
Databases: Hbase, MongoDB, Elastic Search, …
Query: filters, keywords, aggregation (no joins)

Graph Databases

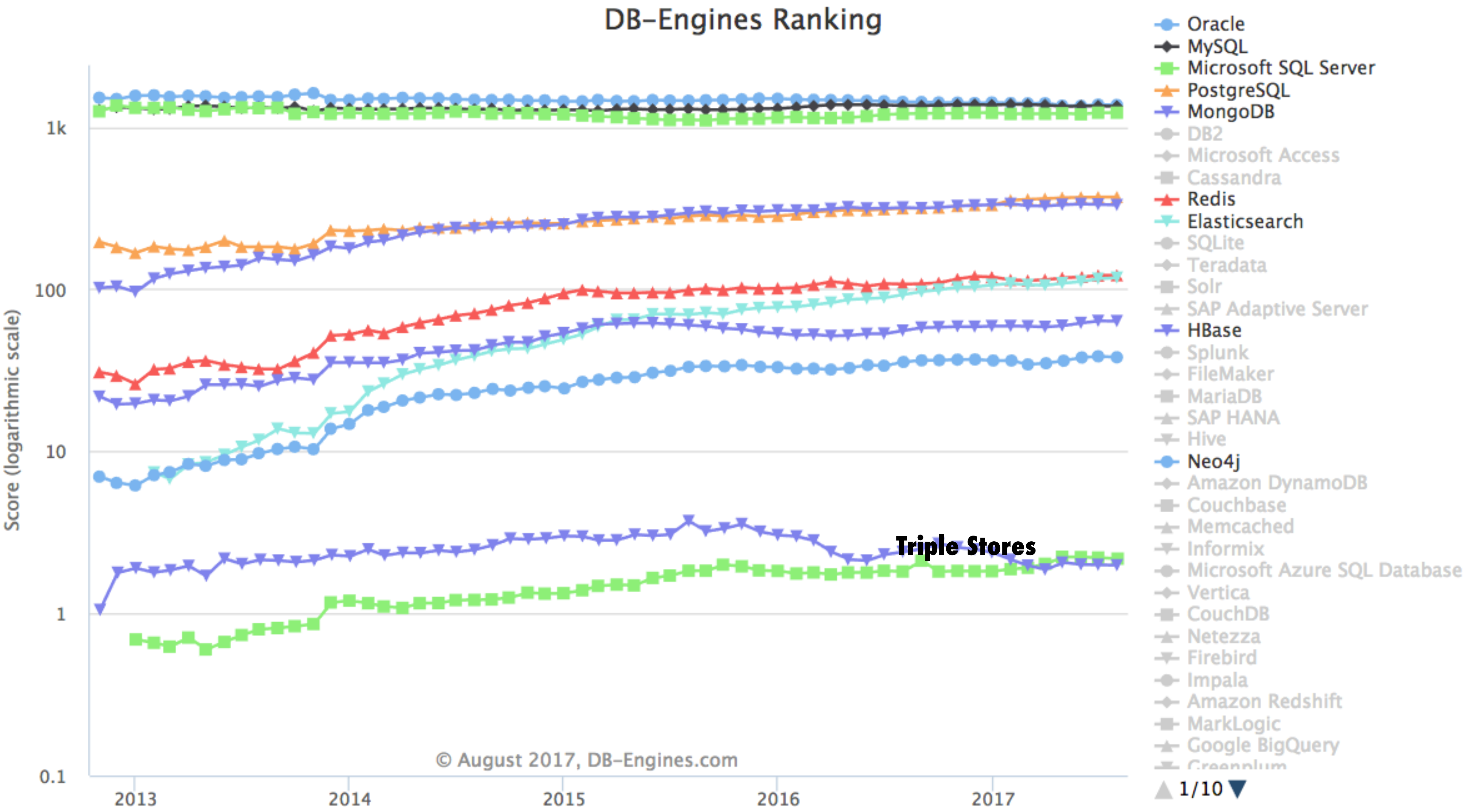
Data model: graph
Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, …
Query: GraphQL, Gremlin, Cypher

Popularity Ranking Of Graph Databases



https://db-engines.com/en/ranking_trend/graph+dbms

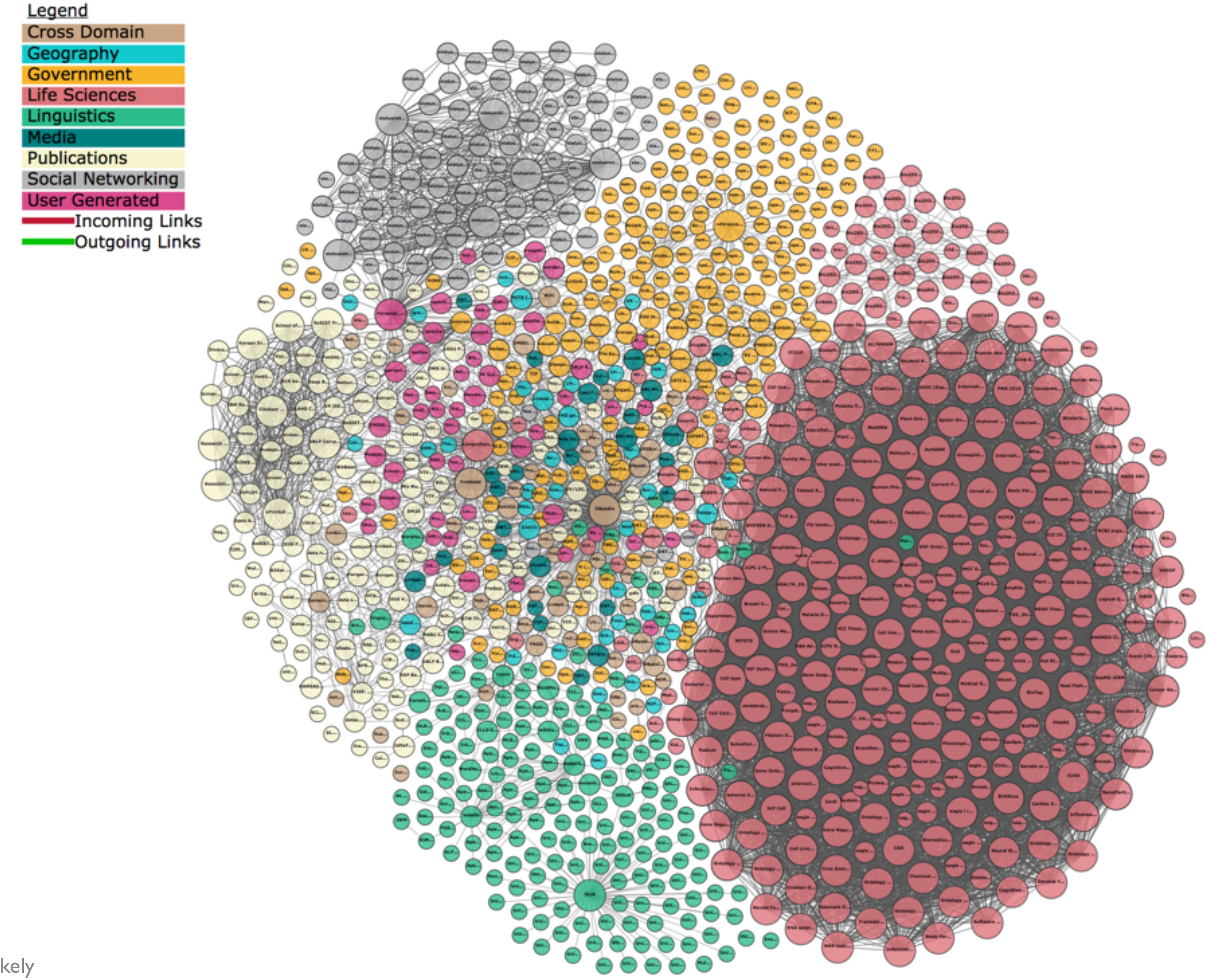
ElasticSearch, MongoDB & Neo4J Have Wide Adoption



https://db-engines.com/en/ranking_trend/graph+dbms

KGs I can **Reuse**

Linked Open Data Cloud



DBpedia

RDF graph derived from Wikipedia
http://wiki.dbpedia.org/

4.58 million things

4.22 million are classified in a consistent ontology

1,445,000 persons

735,000 places

478,000 populated places),

411,000 creative works

123,000 music albums, 87,000 films and 19,000 video games

241,000 organizations

58,000 companies and 49,000 educational institutions

251,000 species

6,000 diseases

YAGO Knowledge Base

http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads

Derived from Wikipedia WordNet and GeoNames

10 million entities

120 million assertions

persons, organizations, cities, etc.

350,000 classes

many fine grained classes, inferred from the data

Wikidata

The "wikipedia" of data
https://www.wikidata.org/wiki/Wikidata:Main_Page

Collaborative, multilingual

collecting structured data to provide support for Wikipedia

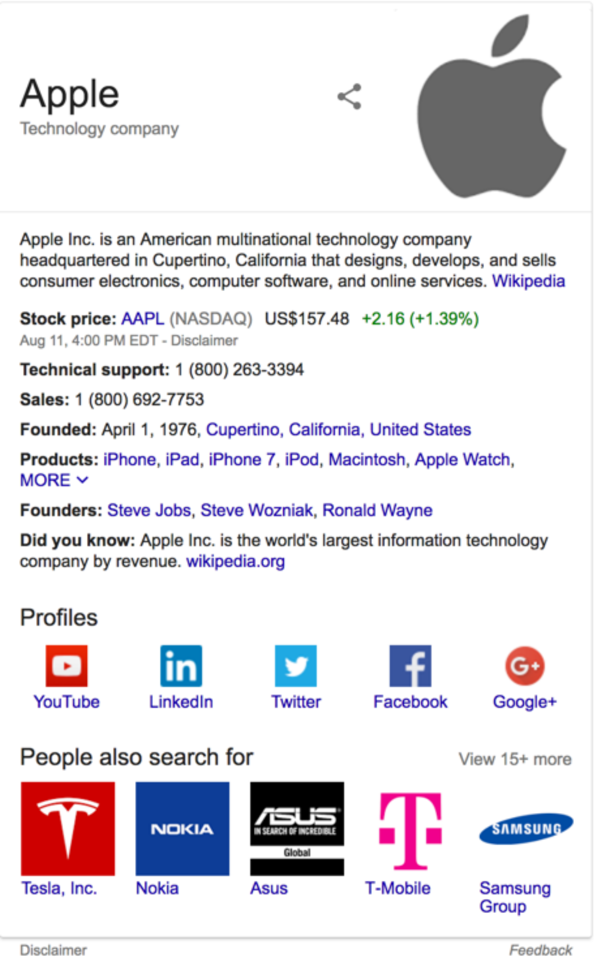
31,419,072 items

534,615,360 edits since the project launch

Google Knowledge Graph

https://developers.google.com/knowledge-graph/how-tos/search-widget-example

search: APPL



derived from many sources, including the CIA World Factbook, Wikidata, and Wikipedia

powers a "knowledge panel"

the Knowledge Graph now holds 70 billion facts

Other Knowledge Graphs

Internet Movie Firearms Database

Firearms used or featured in movies, television shows, video games, and anime
22,159 articles, extensive coverage and ontology
http://www.imfdb.org/wiki/Category:Gun

Microsoft Satori

Large knowledge graph similar to Google KG, e.g., 1.8 million bottles of wine
Many streaming channels of real-time data, e.g., bitcoin, transportation, …
https://www.satori.com/

LinkedIn Knowledge Graph

450M members, 190M historical job listings, 9M companies, 28K schools,
1.5K fields of study, 600+ degrees, 24K titles and 35K skills in 19 languages
https://engineering.linkedin.com/blog/2016/10/building-the-linkedin-knowledge-graph

Knowledge ‘base’ vs ‘graph’

Terms often used interchangeably in the literature

Knowledge base is more of a catch-all e.g., Wikipedia vs. DBpedia

We use the term ‘knowledge graph’ where possible

Why Knowledge Graphs?

Combine advantages of databases and unstructured text

Machine and human understandable

Useful in search and data mining

Many interesting extensions

Interest in multiple communities incl. NLP, data mining…

Many more!

KGs in Unusual domains

Knowledge Graphs In Unusual Domains

Many challenges (and frontier research questions)!

Data skew

Lack of training data

Importance of data exploration

Keeping domain experts in the loop

Multi-faceted prediction and inference

Scalability

…

KG For Unusual Domains

Entities + properties + provenance + confidence + qualifiers

Simple, shallow ontology

customized to domain, if not needed, leave it out

Rich provenance and confidences

essential for end-users, useful for knowledge graph improvement

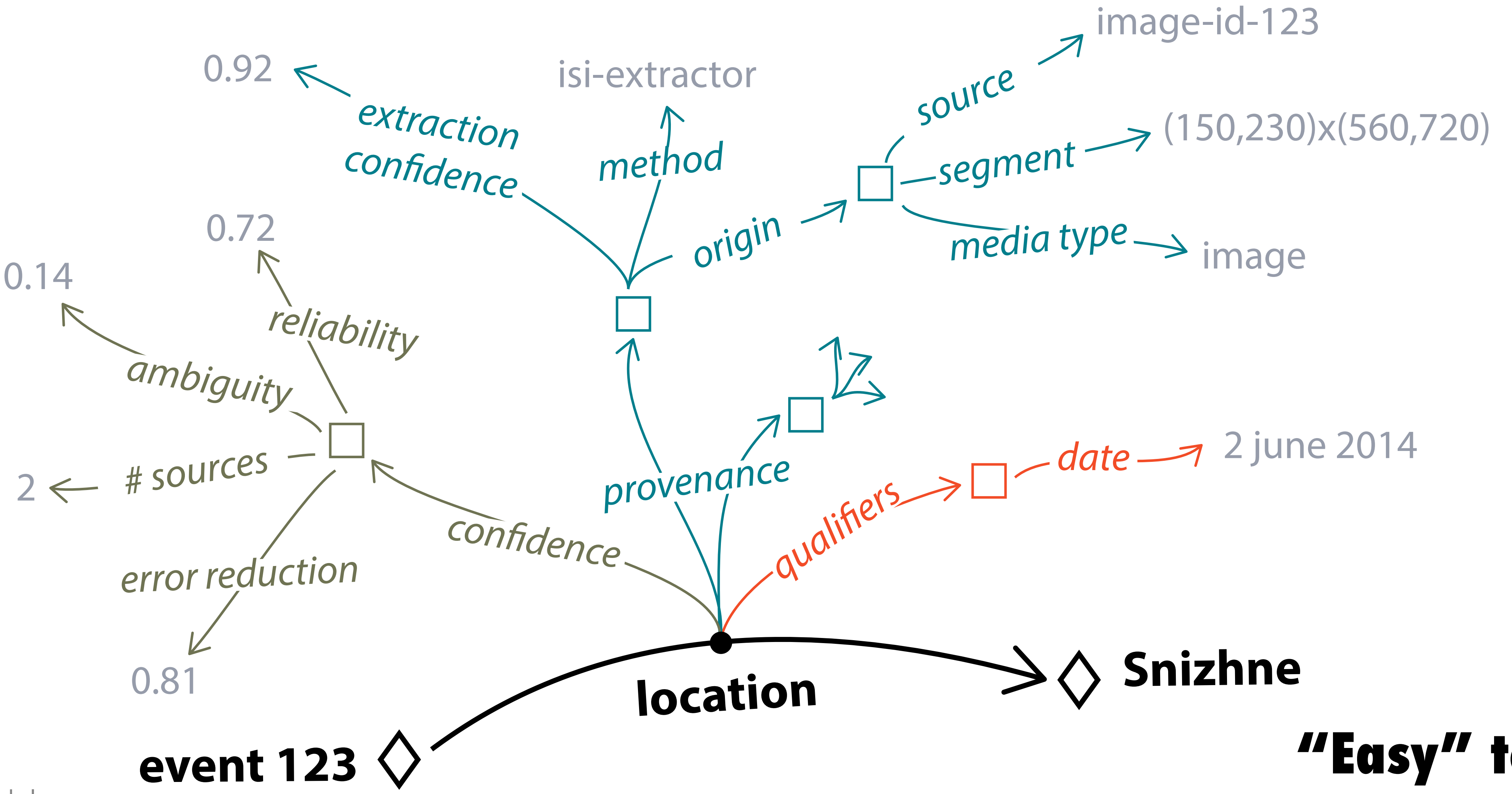
Hybrid text/structured representation

keep the text, essential for machine learning and search

"Easy" to build

KG For Unusual Domains

Entities + properties + provenance + confidence + qualifiers

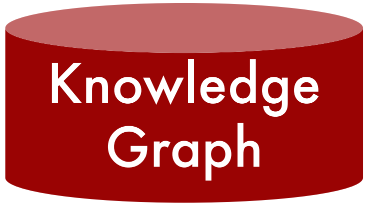


"Easy" to build

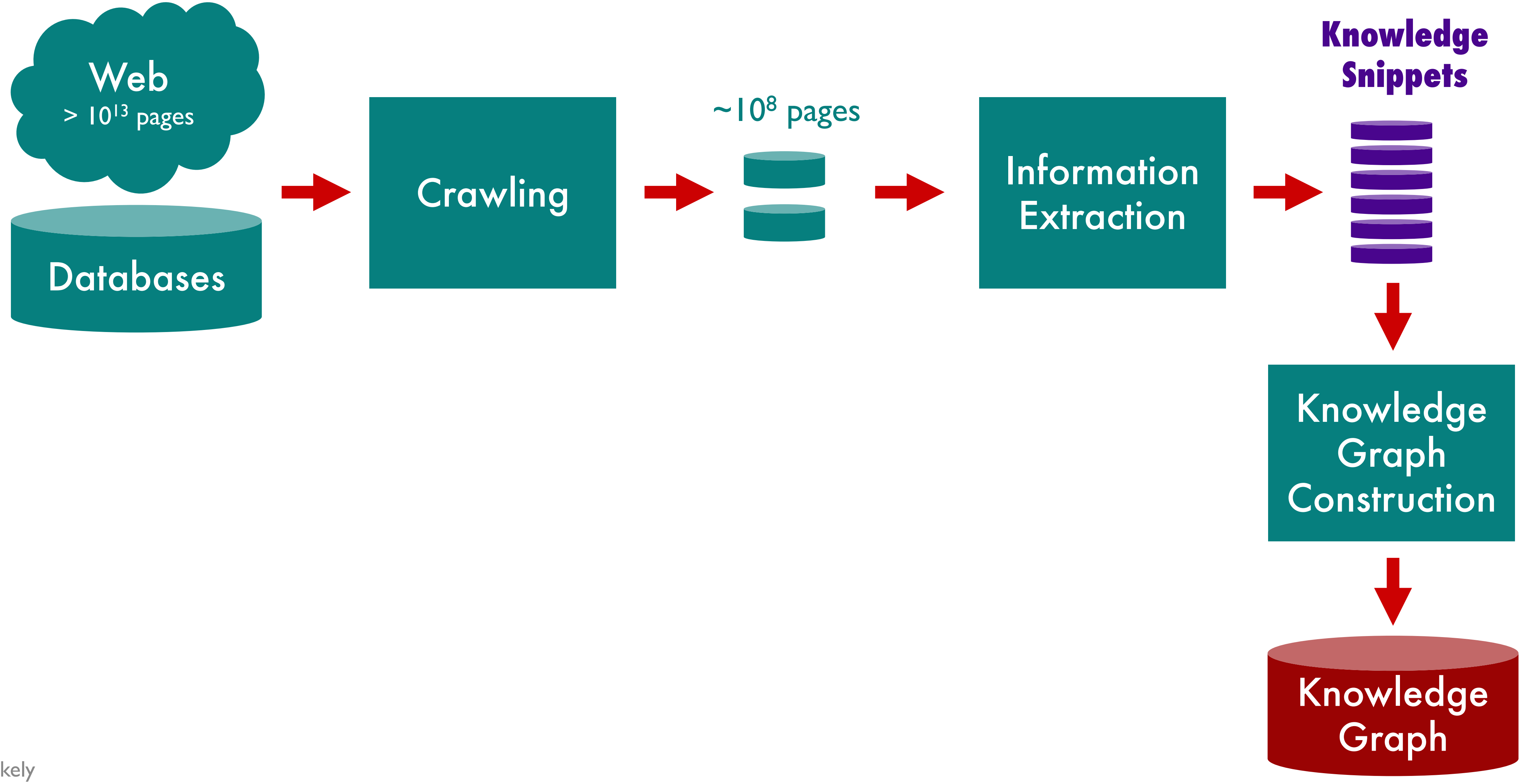
Knowledge Graph Construction

How can I **build** a KG?

KG Construction Problem

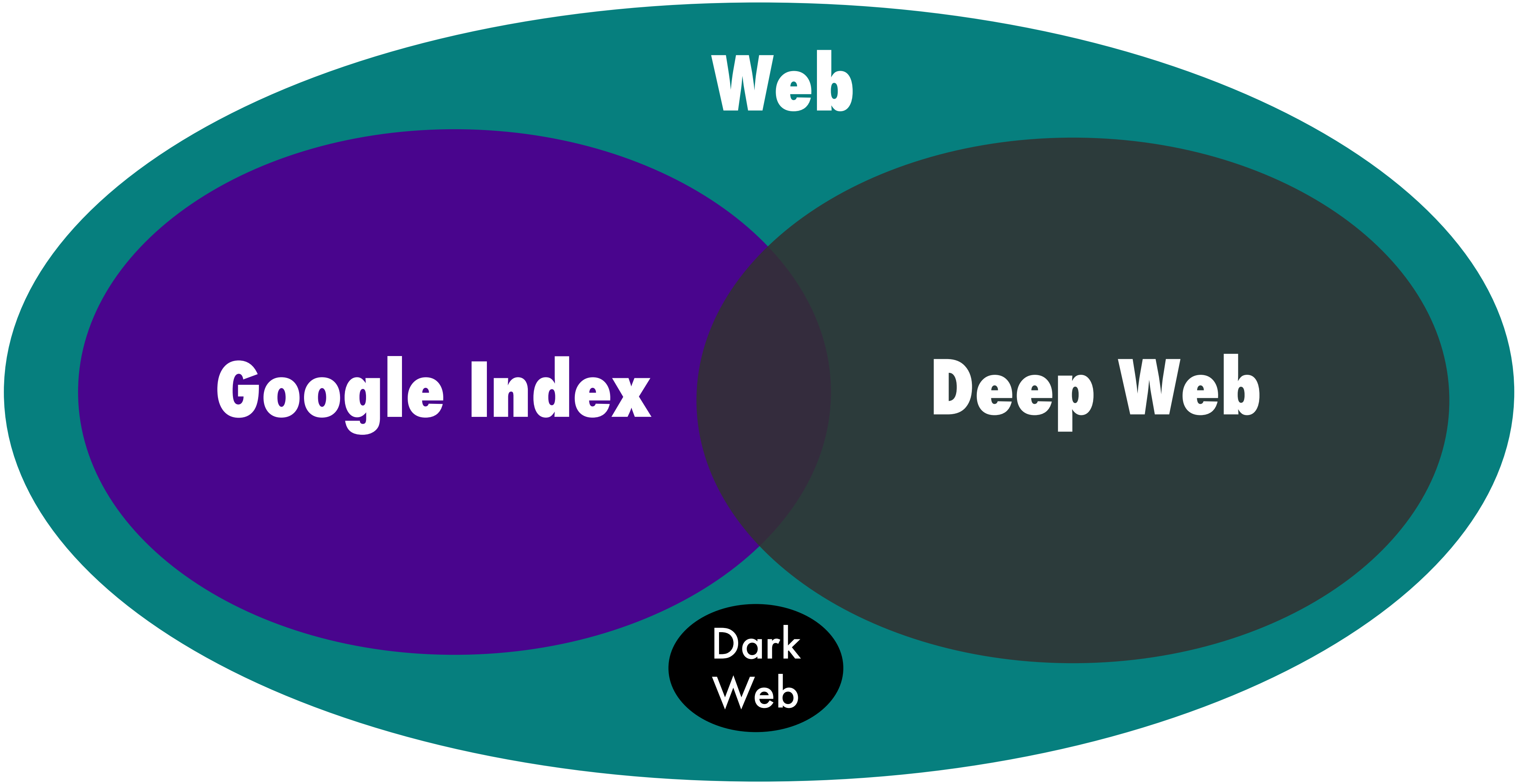


Steps To Build a KG

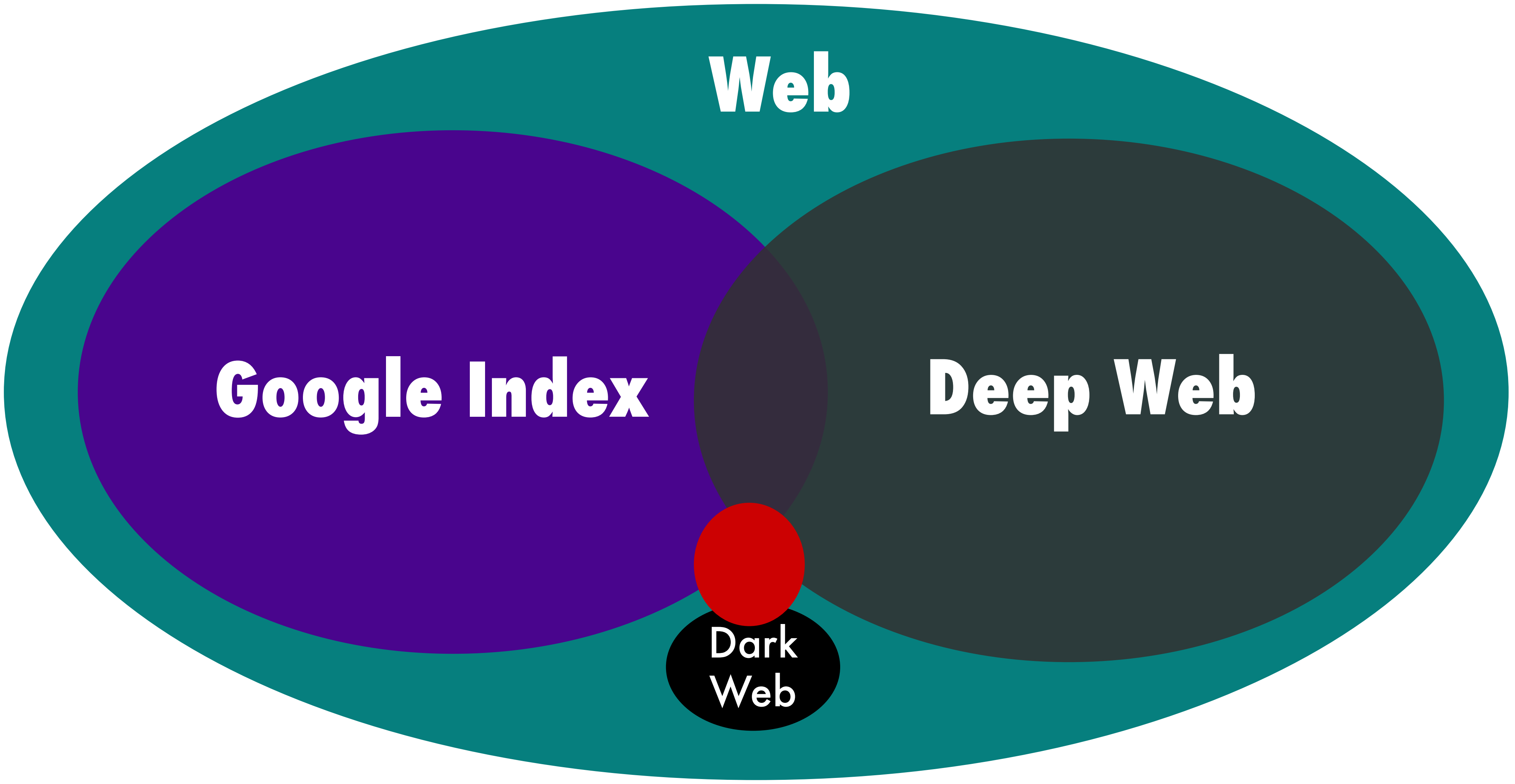


Crawling the web

The web is very BIG



We can only get a tiny part



Crawling Research Challenges

Domain discovery

Identifying relevant sites, datasets and pages

Crawling

Building models of relevant content
Identifying new content
Downloading dynamic content
Overcoming anti-crawling measures
Exhibiting human-like behavior

Crawling Tools

Scrapy (targeted crawling)

Open source and collaborative framework for extracting data from websites
https://scrapy.org/ACHE

Apache Nutch (massive crawling)

highly extensible, highly scalable Web crawler
http://nutch.apache.org

Deep Deep (Adaptive crawler)

reinforcement learning to learn which links to follow
https://github.com/TeamHG-Memex/deep-deep

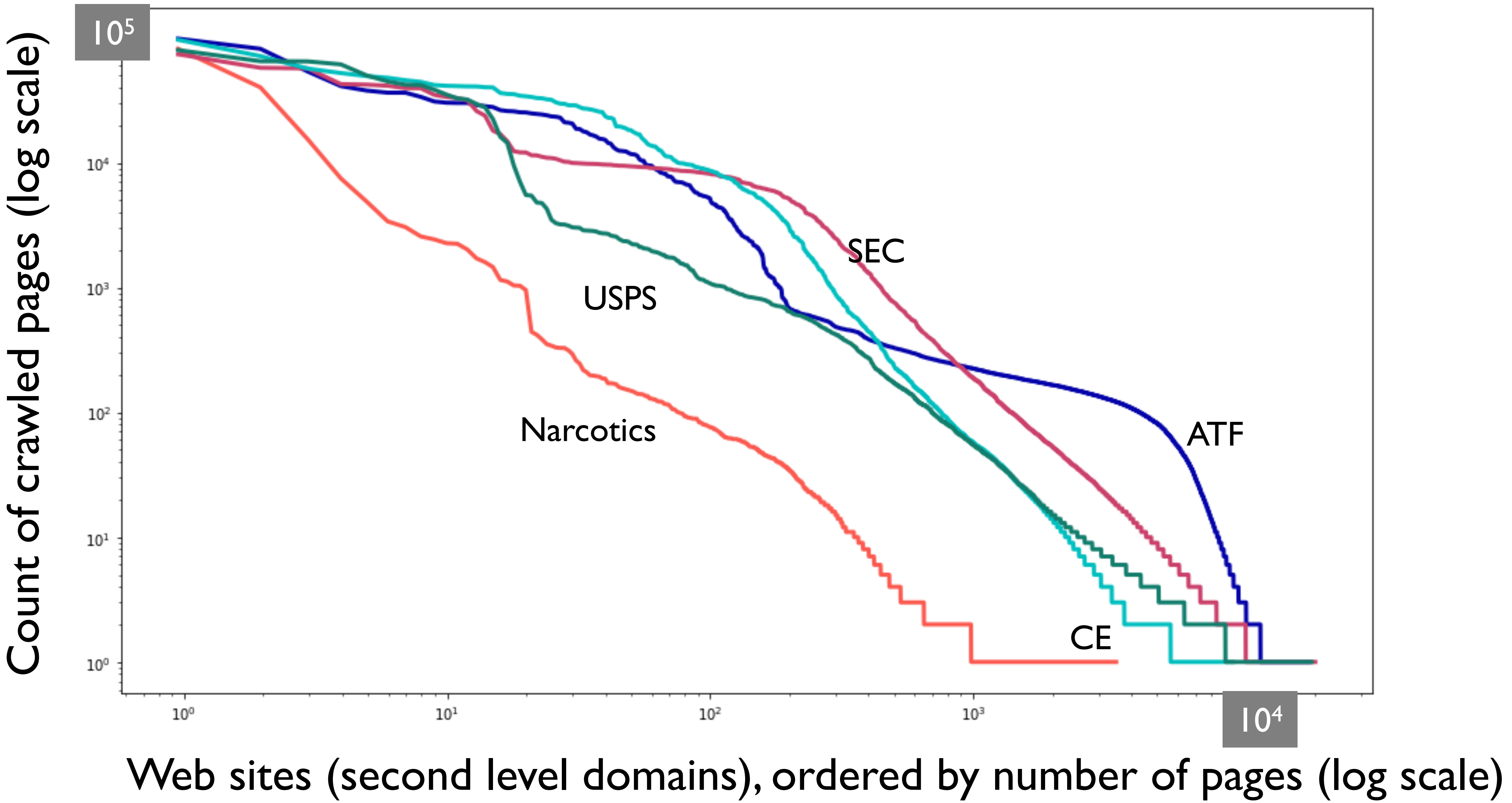
ACHE (focused crawler)

https://github.com/ViDA-NYU/ache

ScrapingHub (service)

https://scrapinghub.com/data-on-demand

Unusual Domains Have Long Tails



principles and challenges in

Information Extraction

Information Extraction (IE)

IE On The Web Is Challenging

Less grammar, more formatting & linking

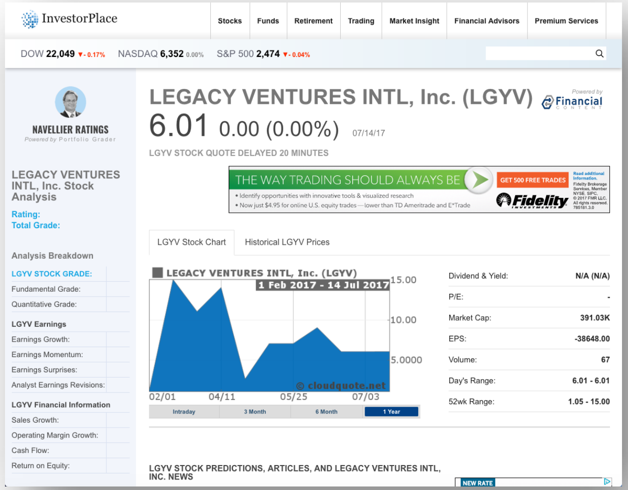
Newswire

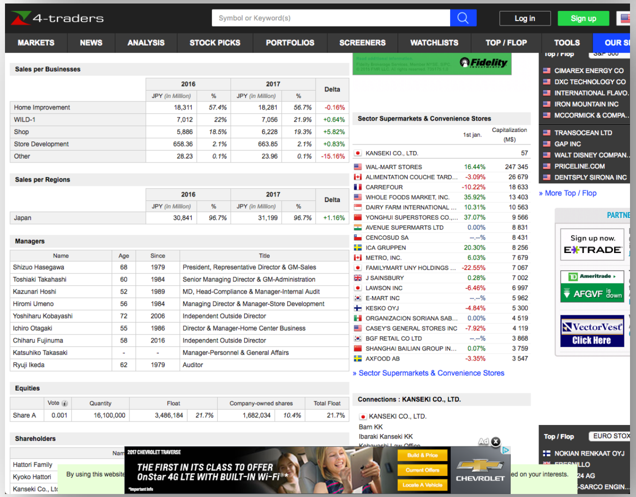
Apple to Open Its First Retail Store in New York City

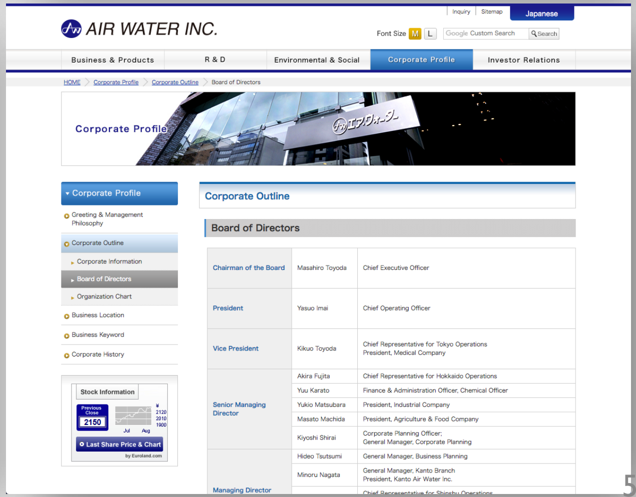
MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web







Dimensions of IE

Document features

Scope

Pattern complexity

Relevance

Document Features

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences plus some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

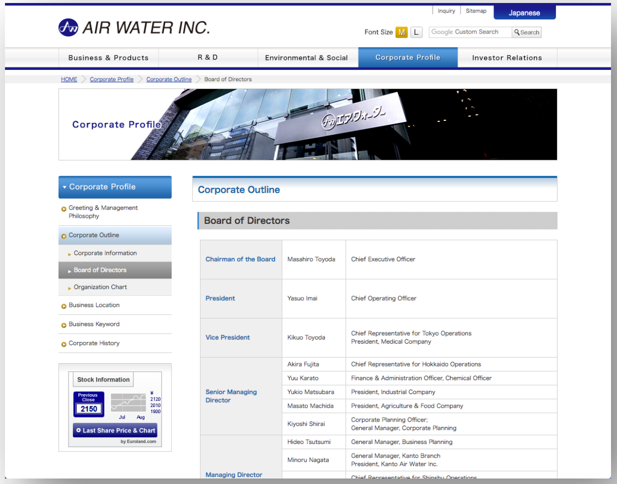
Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

Brock, Oliver (413) 577-0334 oli@cs.umass.edu CS246
Assistant Professor.

Clarke, Lori A. (413) 545-1328 clarke@cs.umass.edu CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

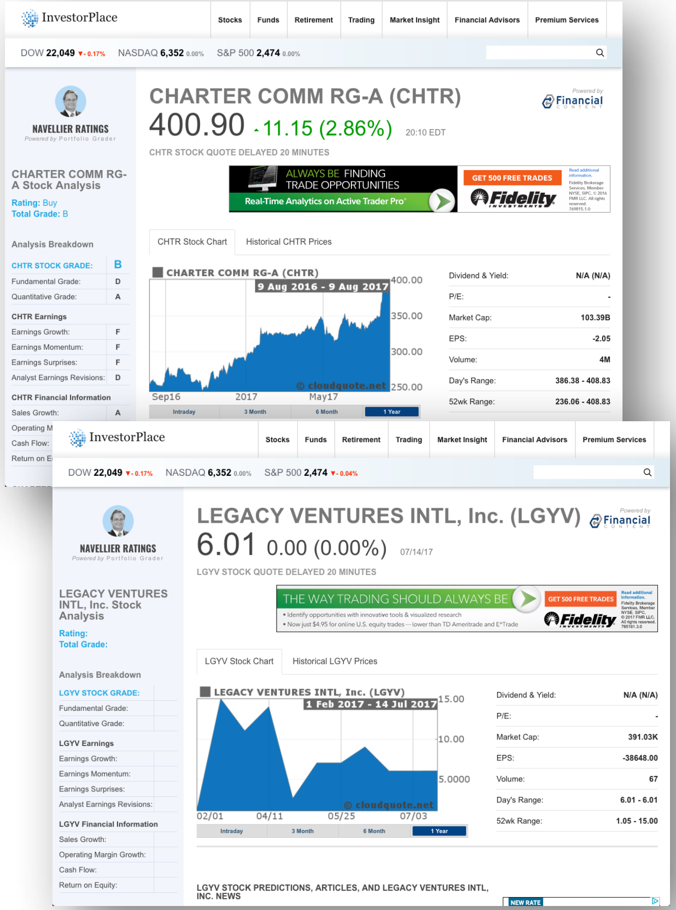
Tables



Charts

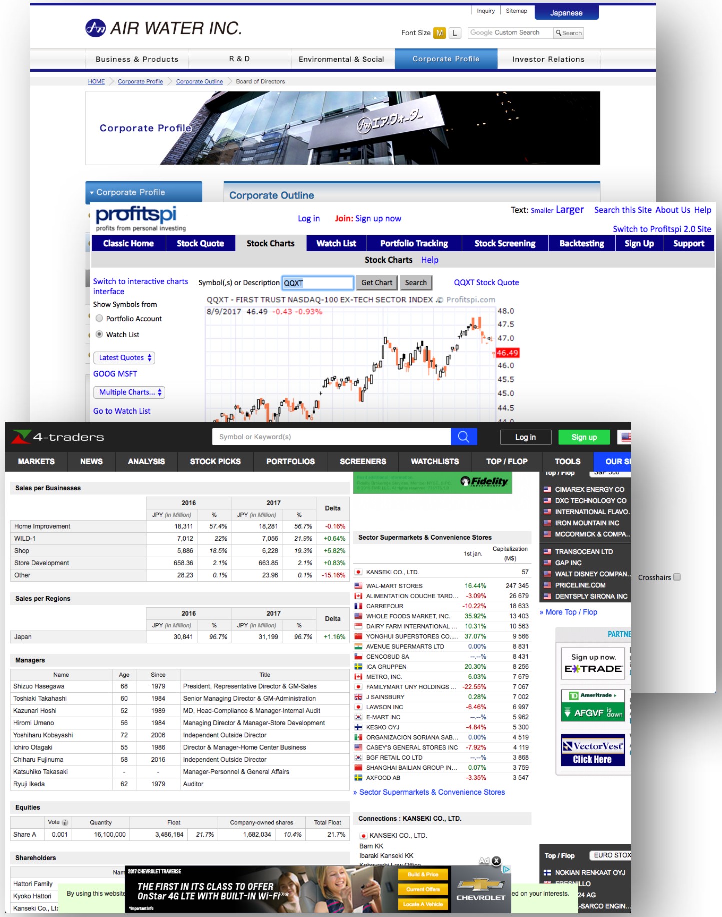
Scope

Web site specific



Genre specific (e.g., forums)

Wide, non-specific



Pattern Complexity

E.g., word patterns

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

"YOU don't wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7o7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses"

647-241-1986 New Haven Escort Listing

View Escorts in other cities

647-241-1986 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25

Escort's Phone: **647-241-1986**
Escort's Location: New Haven, Connecticut
Escort's Age: 25
Date of Escort Post: Jun 17th 4:49pm
REVIEWS: READ AND CREATE REVIEWS FOR THIS ESCORT

There are 42 girls looking in . VIEW GIRLS

If you are looking for the right combination of Erotic & Sensual then you have come to the right place.Always a great personality, and environment.
NO RUSH SERVICE Discreet & Upscale PLAYFUL 100% REAL PHOTOS.
100% Independent | Dedicated | Verified Providerdateche ck dl6472fp 411 p98690
phone:773 431 8174 ___ REFERENCES REQUIREDBDSM, Domme, & Fetishes Available | www.delialondon.com |. Call 647-241-1986. See my menu of services on my profil
EZsex Find me... BackDoorOpen

Call me on my cell at 647-241-1986.
Date of ad: 2016-06-17 16:49:00

More posts from 647-241-1986

- 647-241-1986 Oct 28, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Oct 25, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN...
- 647-241-1986 Oct 09, 2016 Verified Upscale _ Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Oct 09, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London In town TODA...
- 647-241-1986 Oct 07, 2016 Visiting ..Today Only ::: Verified + Reviewed -- // Delia London ... In town for ...
- 647-241-1986 Oct 05, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN...
- 647-241-1986 Aug 16, 2016 NEW PICS Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Aug 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Aug 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Jun 19, 2016 NOW IN WRJ Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Jun 15, 2016 In & outcalls Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 May 16, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- 647-241-1986 May 02, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- 647-241-1986 Apr 30, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- 647-241-1986 Mar 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN - 24
- 647-241-1986 Feb 26, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- 647-241-1986 Jan 13, 2016 Erotic x Busty Asian Companion Verified + Reviewed + Safe In town now - 24
- 647-241-1986 Dec 21, 2015 Asian American -- Busty Companion + Kinkstress :: New Pics + Verified Provider . - ...
- 647-241-1986 Dec 14, 2015 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 26

Recent Escort Classifieds

- North Jersey, New Jersey (732-621-4443)
  :*: G O O D G I R L :*: G O N E *:**: B A D ;) LATINA - 21
- Chicago, Illinois (773-412-2044)
  (LAtE NiGHt) UNRUShEd (ULTiMAtE) PLEASURE (*AmAziNg Azz*) CHOOSE..W...
- Chicago, Illinois (414-914-3777)
  Petite. and Sweet. Super new and Ready... in call - 21
- Atlanta, Georgia (401-324-9388)
  WoW. MuSt TaKe A LoOk At ThIs. - 21
- Atlanta, Georgia (347-940-1982)
  SMOKiNG HOT Specials BuSTy BaBe ((5 SeRvICe)) Pretty 36DDDs ( ) (
- Beautiful Salvadorean The One And Only(: - 21
- Phoenix, Arizona (623-500-7076)
  NEW GIRL PERSIAN Gem EXotIC Blend - 21
- Toronto, Ontario (416-554-3337)
  (L) (L) ~~~Special 80 for 20 min:) 22YeAr oLd $$exyy LaTiNa BoMbSheLL~~(L...
- Toronto, Ontario (416-520-5198)
  21 years old * $80 **real pictures ** A sian Kathy * - 21
- Toronto, Ontario (647-702-6825)

Top Escort Cities

- New York, New York
- Toronto, Ontario
- Dallas, Texas
- Chicago, Illinois
- Atlanta, Georgia
- North Jersey, New Jersey
- Houston, Texas
- Phoenix, Arizona
- Philadelphia, Pennsylvania
- Miami, Florida

Recent Blog Posts

- Sheriff candidate Minister and Detective Reno Fells arrested in prostitution bust
- Man gets 35 years for impersonating cop to get free sex from hooker
- Alexander Marino: Psychologist by Day, Pimp by Night
- Surfside Beach, SC Prostitution BUST: Video

Search Box

Search For Profiles

- Register Here
- Login to your account
- Non Mobile Version
- Escort Blog
- Key for Escort Acronyms
- Top 10 Escort Practices
- Escort Reviews
- See Escorts on Webcam
- Prostitution Laws

Most Recently Viewed
Today at 5:30pm Pacific

419-283-6378
Detroit

small amount of relevant content

irrelevant content very similar to relevant content

IE In Unusal Domains

Full spectrum of in all dimensions

Document Features	Scope	Pattern Complexity	Relevance
• text paragraphs • grammatical, some formatting • ungrammatical, rich formatting • tables • charts	• website specific • genre-specific • wide	• closed set • regular set • complex pattern • ambiguous pattern • unusual language model	• all relevant • significant irrelevant content • active deception • purposeful obfuscation

How to adapt existing techniques without much supervision?

currently, only one universal extractor

Practical Considerations

How good (precision/recall) is necessary?

High precision when showing extractions to users
High recall when used for ranking results

How long does it take to construct?

Minutes, hours, days, months

What expertise do I need?

None (domain expertise), patience (annotation), simple scripting, machine learning guru

What tools can I use?

Many …

Practical IE Technologies

	Glossary	Regex	NLP Rules	Semi-Structured	CRF	NER	Table
Effort	assemble glossary	hours	hours	minutes	O(1000) annotations	zero	O(10) annotations
Expertise	minimal	high, programmer	low	minimal	low-medium	zero	minimal
Precision	medium (ambiguity)	high	high	high	medium-high	medium-high	high
Recall	medium (formatting)	low f(# regex)	medium f(# rules)	high	medium	medium	high
Coverage	wide	wide	wide	single site	genre	news wire	narrow

Case Study

Combating fraud in the penny stock market

Microcap Stock Fraud

Microcap stock fraud is a form of securities fraud involving stocks of "microcap" companies, generally defined in the United States as those with a market capitalization of under $250 million. Its prevalence has been estimated to run into the billions of dollars a year.

Pump and dump schemes, involving use of false or misleading statements to hype stocks, which are "dumped" on the public at inflated prices. Such schemes involve telemarketing and Internet fraud

https://en.wikipedia.org/wiki/Microcap_stock_fraud

Most Relevant Websites

4-traders.com
advfn.com
analystratings.net
barchart.com
bitcointalk.org
blogspot.com
blogspot.in
businessinsider.com
businessprofiles.com
dividend.com
dynamoo.com
etf.com
facebook.com

fifighter.com
financialcontent.com
finanzen.nl
finanzen100.de
hotstocked.com
index.co
investorshangout.com
marketnewscall.com
marketwatch.com
minyanville.com
moneyhub.net
nasdaq.com
openpr.com

otcmarkets.com
pennystock101.org
pennystocktweets.com
pinkinvesting.com
prnewswire.com
rumas.de
sify.com
siliconinvestor.com
stockguru.com
stockopedia.com
stockreads.com
superstockscreener.com
tdameritrade.com

thehotpennysto
cks.com
thelion.com
theotc.tod
traders.com
trading-treff.de
twitter.com
uservoice.com
wikinvest.com
yahoo.com

Microcap Fraud Ontology

defined by two SEC users in 45 minutes

Property	Description
address	US postal service address
city	specific cities mentioned in a page
compensation_amount	amount a promoter was compensated
counsel	counsel
country	countries mentioned in a page
date_of_post	post date of the blog or article
disclaimer	disclaimer in promotion
email	email addresses mentioned in the page
industry	industry of organization
market	stock market of the issuer
message_board	name of message board
message_board_category	category of the message board
organization_name	name of the organization

Property	Description
org_registration_date	date organization was created
org_registration_number	state registration number for the organization
organization_status	status of the organization's state registration
paying_party	entity that is paying for the promotion
phone	phone numbers mentioned in the page
posted_date	any date mentioned in a page
promoter_name	name of the promoter
state	specific states mentioned in a page
stock_ticker	stock tickers mentioned in the page
twitter_tag	tags in a tweet
user_registration_date	date the user registered for the site
username	user name in a website
website	website where page was published

myDIG Demo

setting up a new domain (after crawling)

Specifying websites

Defining the domain ontology

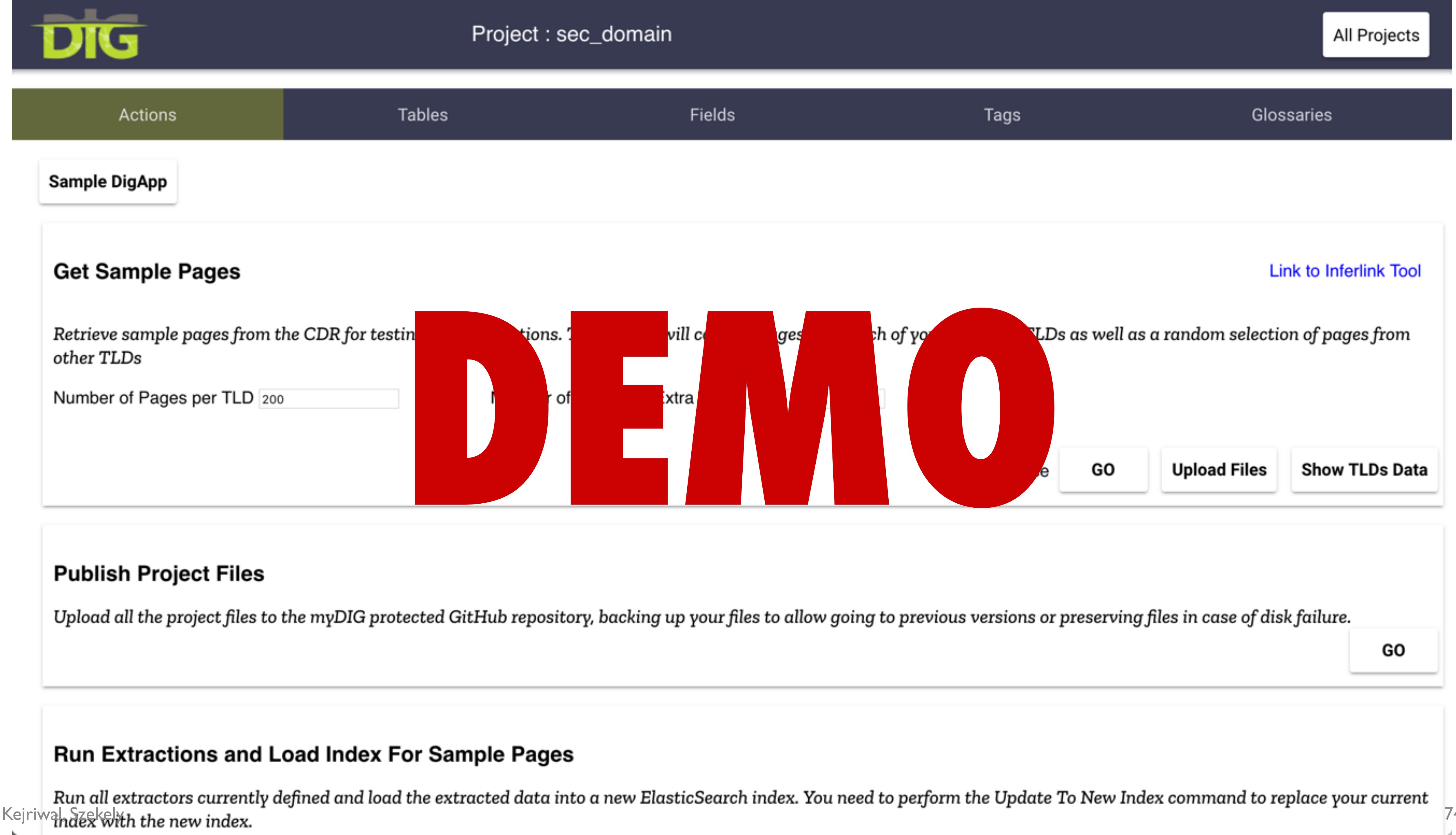
Defining extractors

Building the knowledge graph

Customizing the search engine

Featured Extractors

	Glossary	Regex	NLP Rules	Semi-Structured	CRF	NER	Table
Effort	assemble glossary	hours	hours	minutes	O(1000) annotations	zero	O(10) annotations
Expertise	minimal	high, programmer	low	minimal	low-medium	zero	minimal
Precision	medium (ambiguity)	high	high	high	medium-high	medium-high	high
Recall	medium (formatting)	low f(# regex)	medium f(# rules)	high	medium	medium	high
Coverage	wide	wide	wide	single site	genre	news wire	narrow

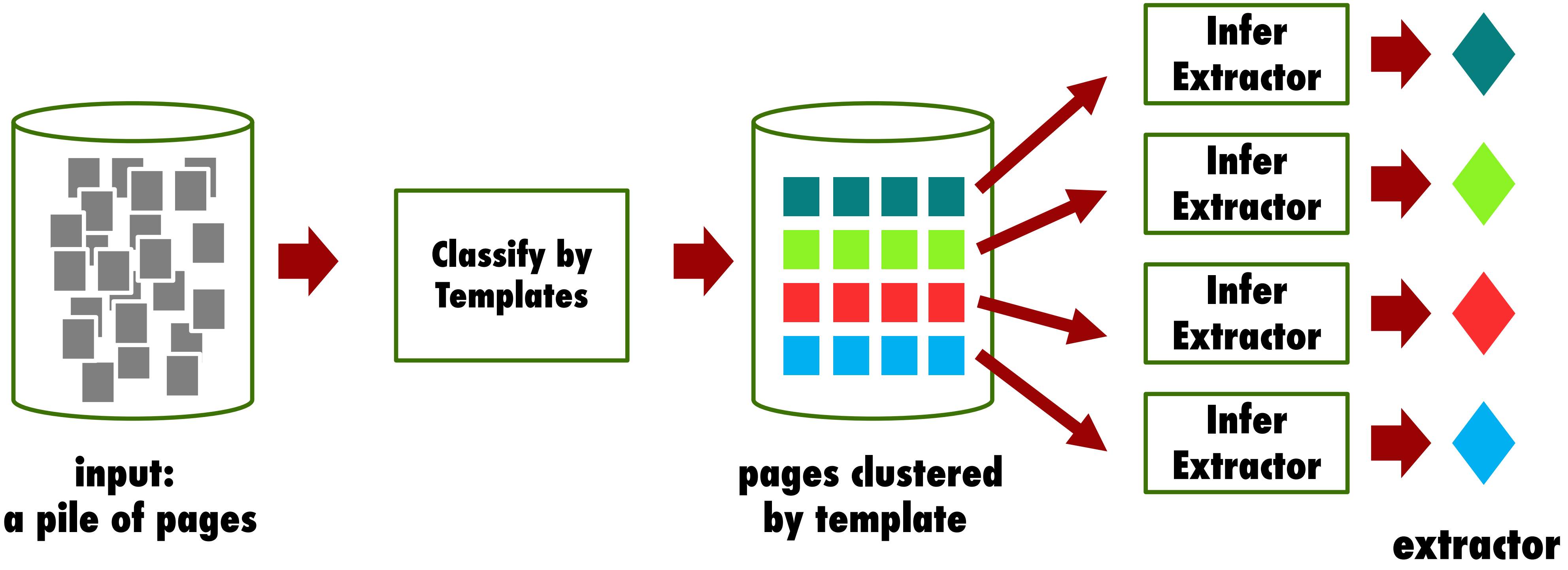

DIG
Project : sec_domain
All Projects
Actions
Tables
Fields
Tags
Glossaries
Sample DigApp
Get Sample Pages
Link to Inferlink Tool
DEMO
Number of Pages per TLD 200
GO
Upload Files
Show TLDs Data
Publish Project Files
Upload all the project files to the myDIG protected GitHub repository, backing up your files to allow going to previous versions or preserving files in case of disk failure.
GO
Run Extractions and Load Index For Sample Pages
Run all extractors currently defined and load the extracted data into a new ElasticSearch index. You need to perform the Update To New Index command to replace your current index with the new index.

Inferlink Extractor

Automatic extraction from semi-structured pages

https://github.com/inferlink

Inferlink Extractor



Inferlink Extraction Accuracy

firearms domain, 10 websites, 5 pages each

	fields										
	Title	Desc	Seller	Date	Price	Loc	Cat	Member Since	Expires	Views	ID
Perfect	1.0 (50/50)	.76 (37/49)	.95 (40/42)	.83 (40/48)	.87 (39/45)	.51 (23/45)	.68 (34/50)	1.0 (35/35)	.52 (15/29)	.76 (19/25)	.97 (35/36)
Pretty Good	1.0 (50/50)	.98 (48/49)	.95 (40/42)	.83 (40/48)	.98 (44/45)	.84 (38/45)	.88 (44/50)	1.0 (35/35)	.55 (16/29)	1.0 (25/25)	1.0 (36/36)

Pretty Good: useful to user

- **extra tokens present**
- **non-essential tokens missing**

Glossary Extraction

Simple in principle

list of words or phrases to extract

Challenges

Ambiguity: Charlotte is a name of a person and a city

Colloquial expressions: "Asia Broadband, Inc." vs "Asia Broadband"

Research

Improving precision of glossary extractions

Extending glossaries automatically

Extraction Using Regular Expressions

Too difficult for non-programmers

regex for North American phone numbers:

```
^(?:(?:\+?1\s*(?:[.-]\s*)?)?(?:\(\s*([2-9]1[02-9]|[2-9][02-8]1|[2-9][02-8][02-9])\s*\)|([2-9]1[02-9]|[2-
9][02-8]1|[2-9][02-8][02-9]))\s*(?:[.-]\s*)?)?([2-9]1[02-9]|[2-9][02-9]1|[2-9][02-9]{2})\s*(?:[.-]\s*)?([
0-9]{4})(?:\s*(?:#|x\.?|ext\.?|extension)\s*(\d+))?$
```

Brittle and difficult to adapt to unusual domains

unusual nomenclature and short-hands

obfuscation

NLP Rule-Based Extraction

Tokenization for unusual domains

tokenize on white-space, punctuation and emojis

Token properties

literal, part of speech tag, lemma, in/out of dictionary
dependency parsing relationships (advanced)
type (alphanumeric, alphabetic, numeric)
shape (pattern of digits and characters), capitalization, prefix and suffix
number of characters, range (numbers)

Pattern

Sequence of required/optional tokens
positive and negative patterns

Named Entity Recognizers

Machine learning models (Conditional Random Field)

people, places, organizations and a few others

SpaCy

complete NLP toolkit, Python (Cython), MIT license
code: https://github.com/explosion/spaCy
demo: http://textanalysisonline.com/spacy-named-entity-recognition-ner

Stanford NER

part of Stanford's NLP software library, Java, GNU license
code: https://nlp.stanford.edu/software/CRF-NER.shtml
demo: http://nlp.stanford.edu:8080/ner/process

myDIG: A KG Construction Toolkit

Python, MIT license, https://github.com/usc-isi-i2

Enable end-users to construct domain-specific KGs

end users from 5 government orgs constructed KGs in less than one day

Suite of extraction techniques

semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)

KG includes provenance and confidences

enable research to improve extractions and KG quality

Scalable

runs on laptop (~100K docs), cluster (> 100M docs)

Robust

Deployed to many law enforcement agencies

Easy to install

Aug 31 2017: Docker deployment with single "docker componse up" installation

Knowledge Graph Completion

Our thanks to Lise Getoor for some slides on Entity Resolution and PSL

Problem

Extractions are noisy

Noise is not random

- Postal code got extracted as phone
- Email ID, social network ID got interchanged
- **Entity Disambiguation**: Charlotte, NC vs. Charlotte the person

Complete the knowledge graph by inferring wrong links, new links

Some solutions we'll cover today

Entity Resolution (ER)

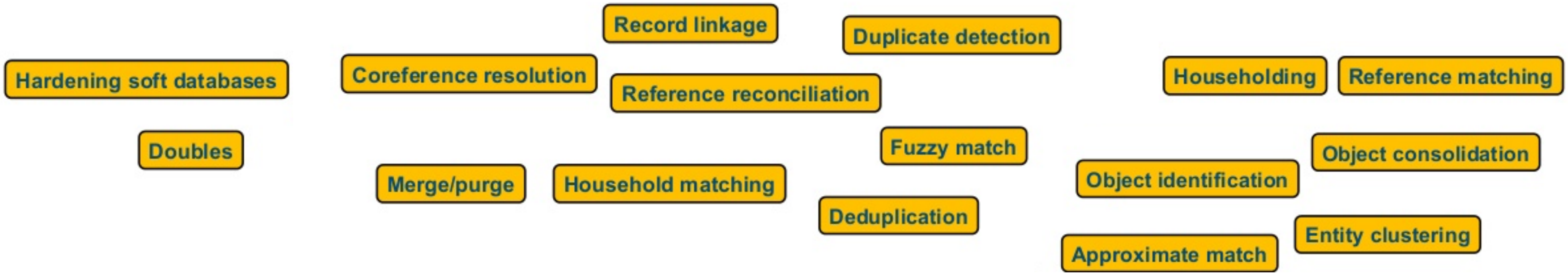
Probabilistic Soft Logic (PSL)

Knowledge Graphs in Latent Space aka knowledge graph embeddings

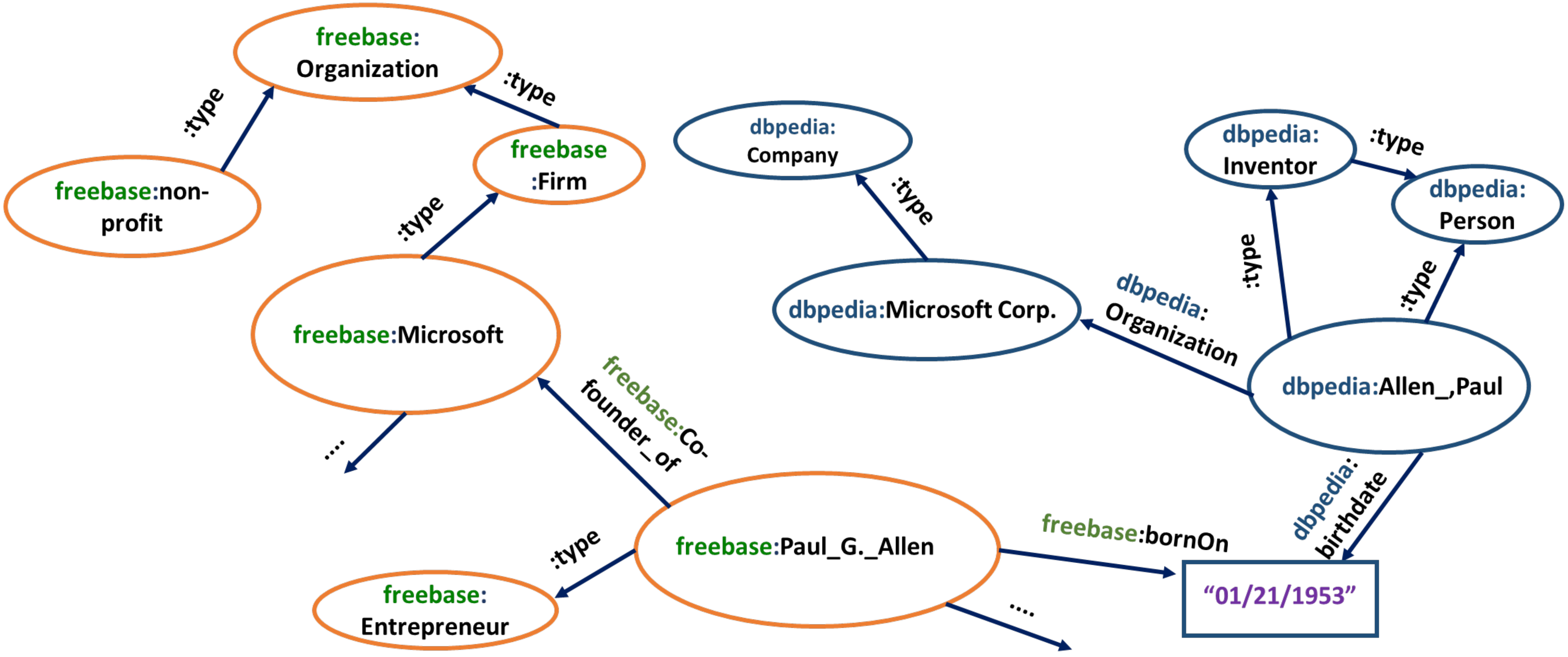
Entity Resolution (ER)

Entity Resolution (ER)

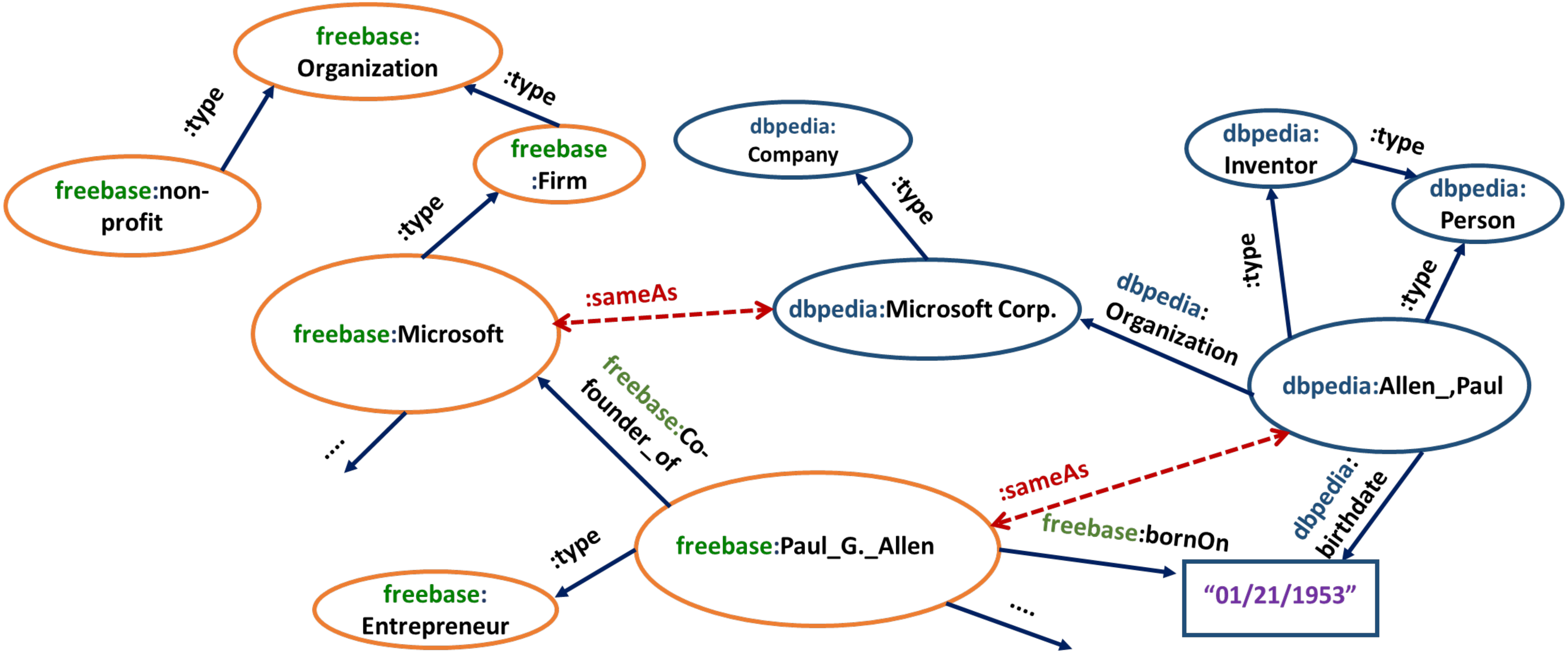
The problem of clustering mentions that refer to the same underlying entity



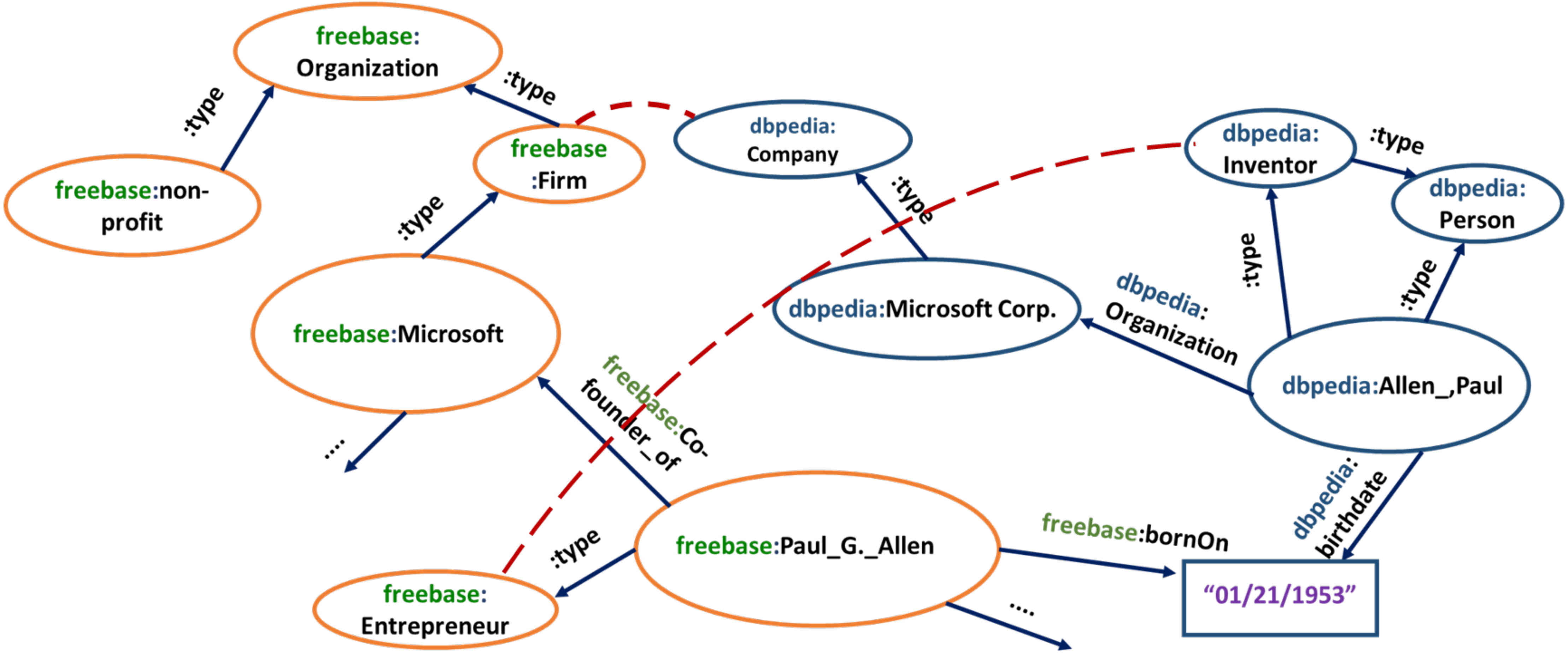
Example: Linking Dbpedia to Freebase



Knowledge graphs contain duplicate entities



Knowledge graph nodes are multi-type



How to do ER?

Popular methods use some form of machine learning; see surveys by Kopcke and Rahm (2010), Elmagarmid et al. (2007), Christophides et al. (2015)

- Probabilistic Matching Methods
- Supervised, Semi-supervised
  - Marlin (SVM based)
  - Bilenko and Mooney (2003)
- Active Learning
- Rule Based
- Distance Based
- Unsupervised
  - EM
  - Winkler (1993)
  - Hierarchical Graphical Models
  - Ravikumar and Cohen (2004)
  - SVM
  - Christen (2008)

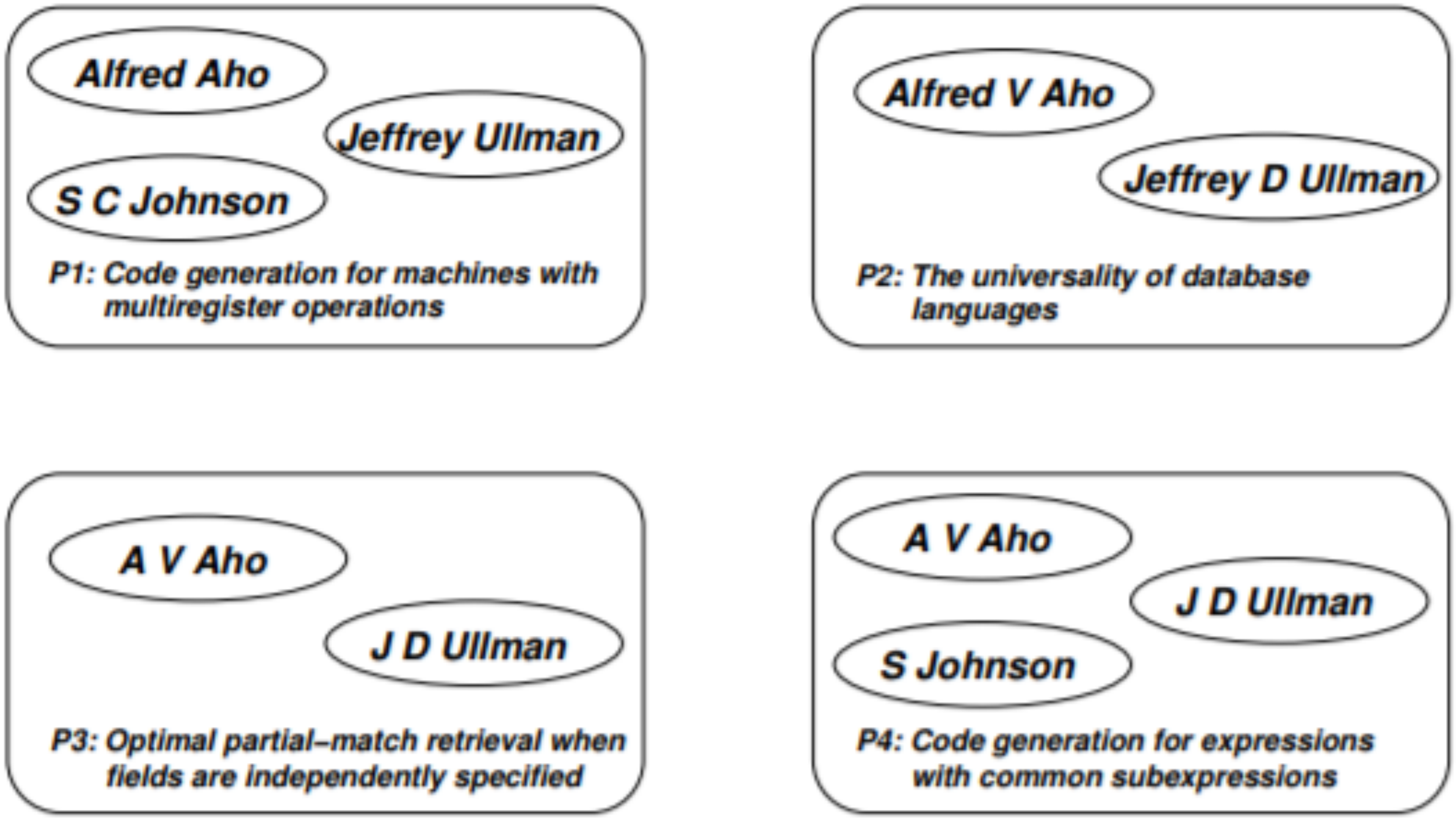
With graph representation

Can propagate similarity decisions Melnik, Garcia-Molina and Rahm (2002)

- More expensive but better performance

Can be generic or use domain knowledge e.g., citation/bibliography domain Bhattacharya and Getoor (2006,2007)

Example (co-authorship)



Bhattacharya and Getoor (2006,2007)

Feature functions - I

First line of attack is *string matching*

Character based

Edit Distance
Affine Gap
Smith-Waterman
Jaro
Q-gram

Token based

Monge Elkan
TF-IDF
- Soft
- Q-gram

Jaccard

Phonetic based

Soundex
NYSIIS
ONCA
Metaphone
Double Metaphone

Available Packages: SecondString, FEBRL, Whirl…

Feature functions - II

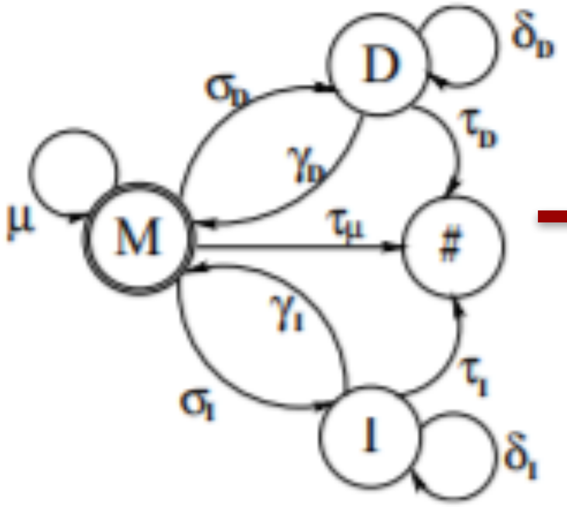
Unusual domains have many non-text fields Feature functions may have to be hand-crafted or learned from data:

- Tracking numbers for mail fraud/narcotics
- Review IDs for human trafficking
- Stock tickers for SEC

Learnable string similarity

Example: adaptive edit distance

Sets of equivalent string pairs (e.g., <Suite 1001, Ste. 1001>

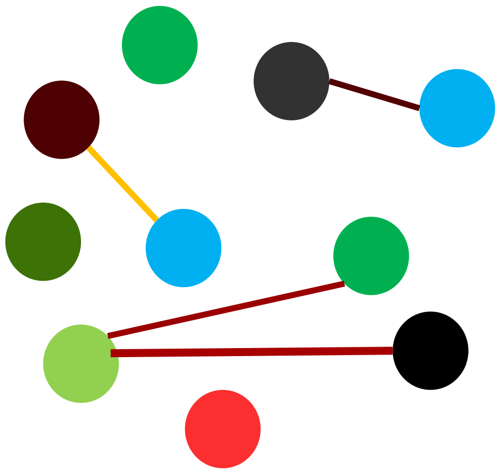


Learned parameters

Bilenko and Mooney (2003)

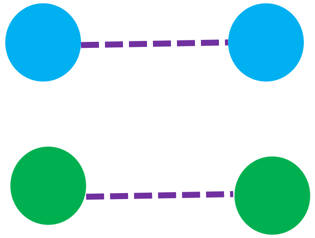
After training...

Apply classifier i.e. similarity function to every pair of nodes? Quadratic complexity!



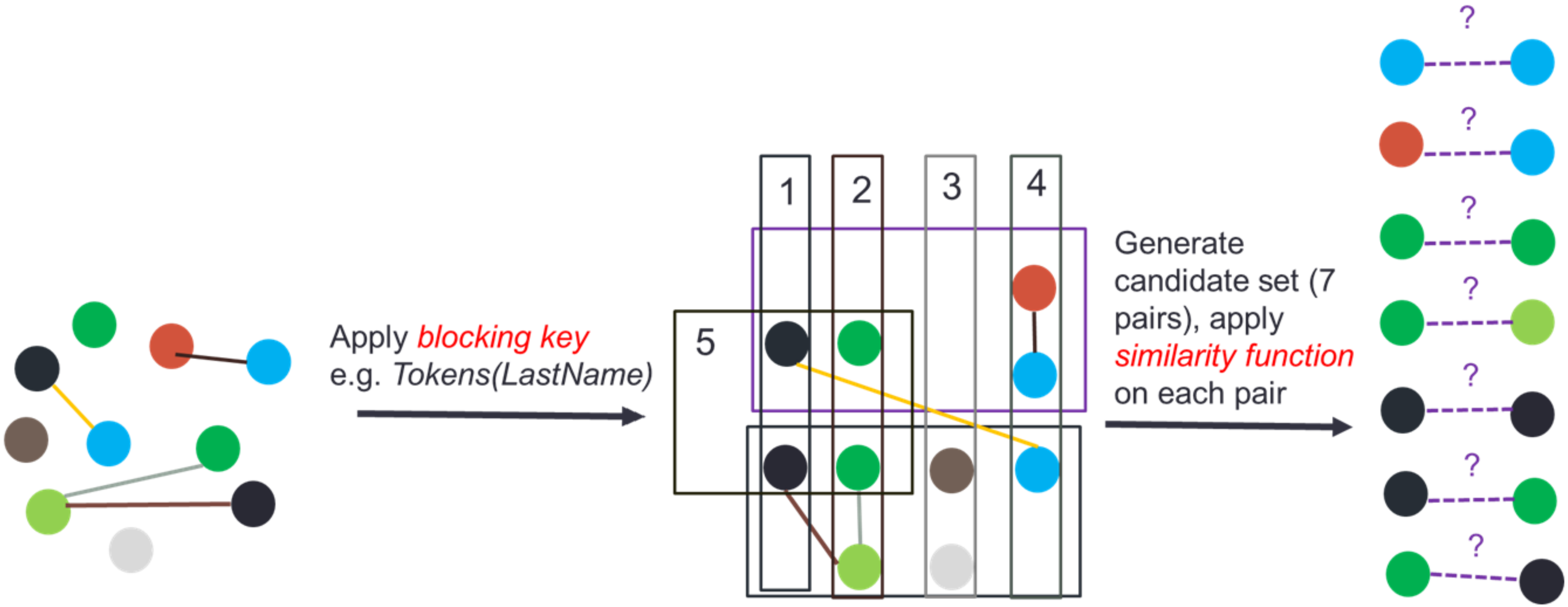
$O(|V|^2)$ applications of similarity function

Linked mentions



Blocking trick

Like a configurable inverted index function



What is a good blocking key?

Achieves high recall

Achieves high reduction

Good survey on blocking: Christen (2012)

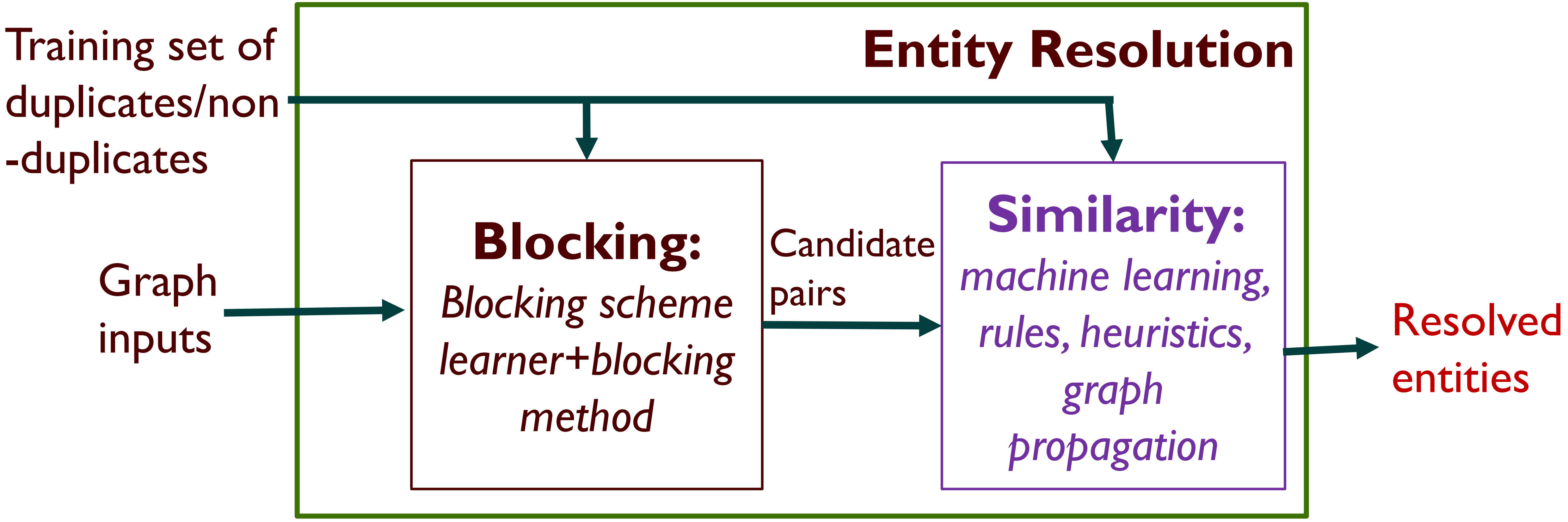
How do we learn a good blocking key?

Key idea in existing work is to learn a DNF rule with indexing functions as atoms

CharTriGrams(Last_Name) U (Numbers(Address) X Last4Chars(SSN))

Michelson and Knoblock (2006), Bilenko, Kamath and Mooney (2006), Kejriwal and Miranker (2013; 2015)...

Putting it together: two-step ER



ER packages

Not many, still tend to be inefficient or for specific domain

FEBRL was designed for **biomedical** record linkage (Christen, 2008)

Dedupe crashes on graphs with fewer than a million nodes
https://github.com/dedupeio/dedupe

LIMES, Silk mostly designed for RDF data, often **require pre-specified similarity functions** (Ngonga Ngomo and Auer, 2008; Isele et al. 2010)

Not all attributes are equal

Phones/emails important in human trafficking

(names are unreliable)

Names can be important in SEC

(nothing special about phones)

How do we use this knowledge?

Domain knowledge

Especially important for unusual domains but how do we express and use it?

Use rules? Too brittle, don't always work!

Use machine learning? Training data hard to come by, how to encode rule-based intuitions?

Probabilistic Soft Logic (PSL)

Practical methods to resolve logic/probability dilemma

Statistical Relational Learning (SRL)

Stochastic Logic Programs (SLPs) Muggleton (1996)

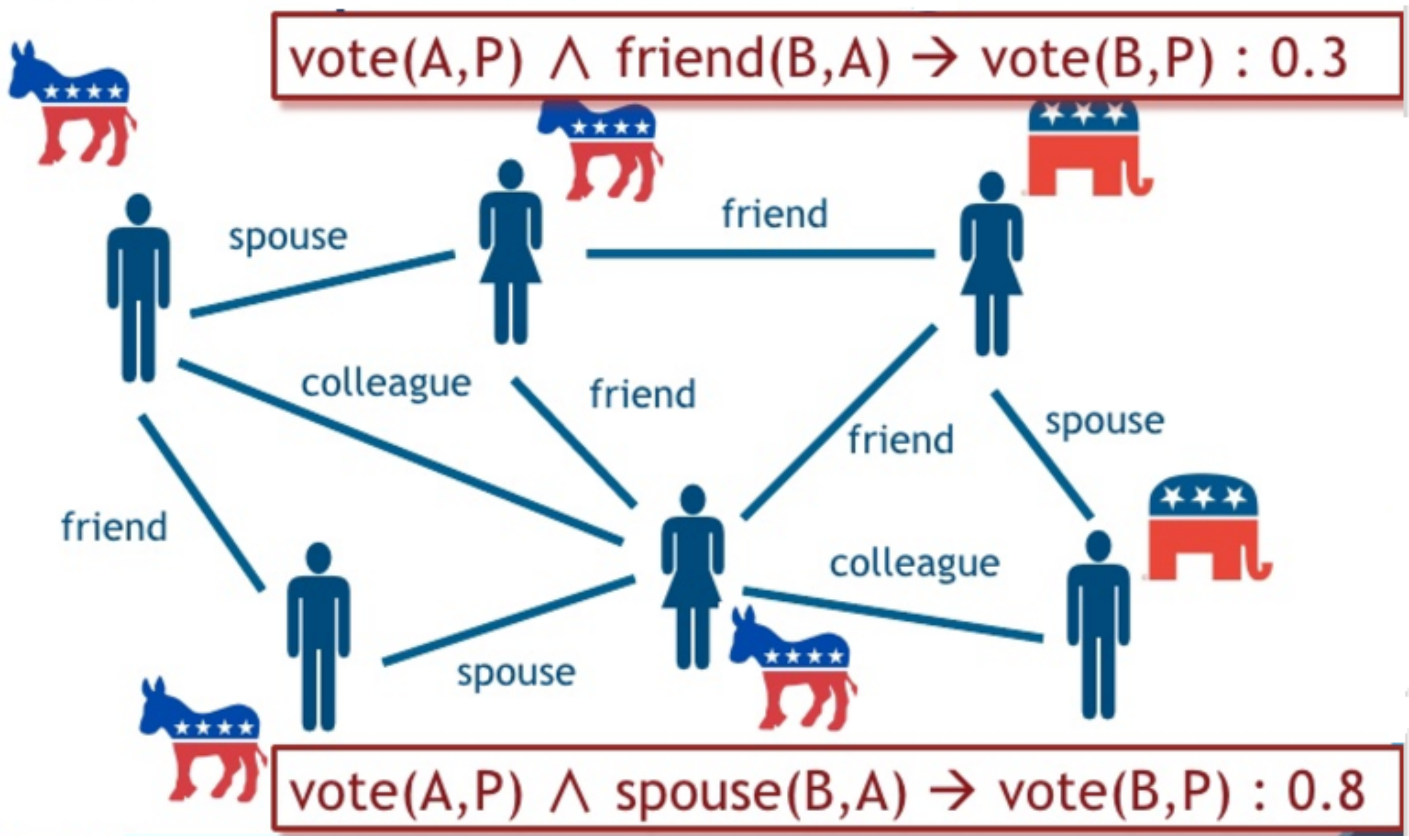
Probabilistic Relational Models (PRMs) Koller (1999)

Bayesian Logic Programs (BLPs) Kersting and De Raedt (2001)

Markov Logic Networks (MLNs) Richardson and Domingos (2006)

Probabilistic Soft Logic (PSL) Kimmig et al., (2012)

Intuitive example (PSL)



Why PSL?

- Continuous Random Variables
  - Mathematical Foundation
  - Logic Foundation
  - Inference & Learning
- Sets and Aggregators
- Extensible
- High Performance

Probabilistic Soft Logic (PSL)

Good framework for expressing domain constraints

Code available, many tutorials and videos on how to get started

http://psl.linqs.org/index.html

Rules need to be pre-specified

But weights can be learned from training data

$$vote(A, P) \ \wedge friend(B, A) \rightarrow vote(B, P): ?$$

$$vote(A, P) \ \wedge spouse(B, A) \rightarrow vote(B, P): ?$$

Some recent work has also tried to learn rules (but may lose interpretability)

Case Study: Toponym Resolution

Toponym resolution: resolving extracted location mention to a canonical geolocation in a KB like **GeoNames**

Applied to human trafficking domain

> …I'm from beautiful downtown **London** in **Ontario, Canada**…

Rules for toponym resolution

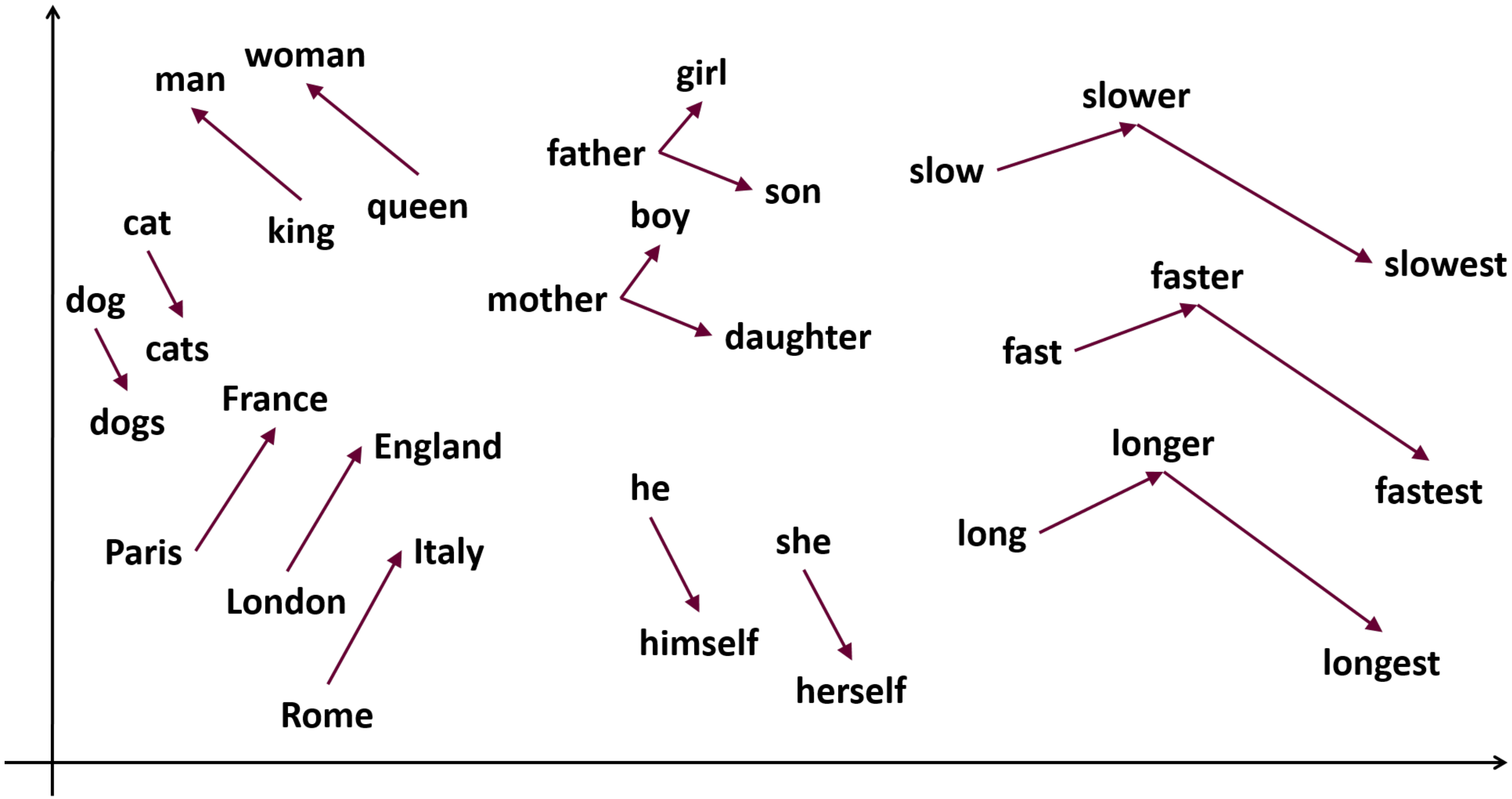
Can easily encode rules (in PSL) such as a city can only be in one state and one country, a city must actually exist in a state and country...

Use elements like population to assign weights

Knowledge Graphs in Latent Space

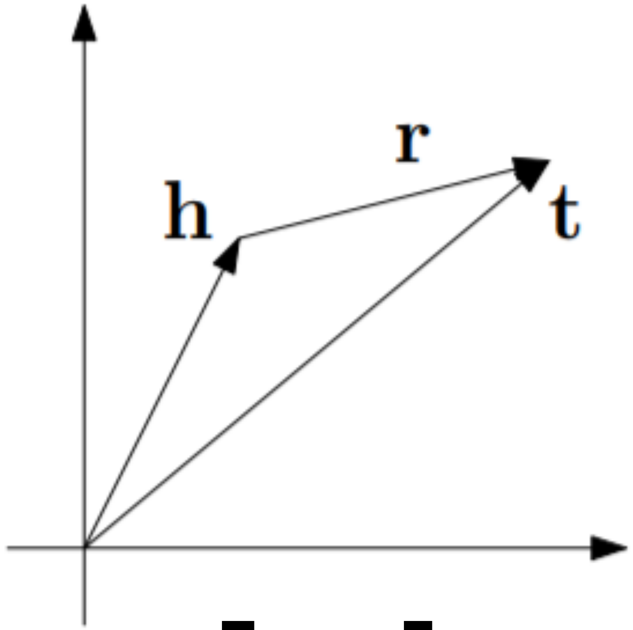
Low-dimensional vector spaces

Very popular for documents, graphs, words…

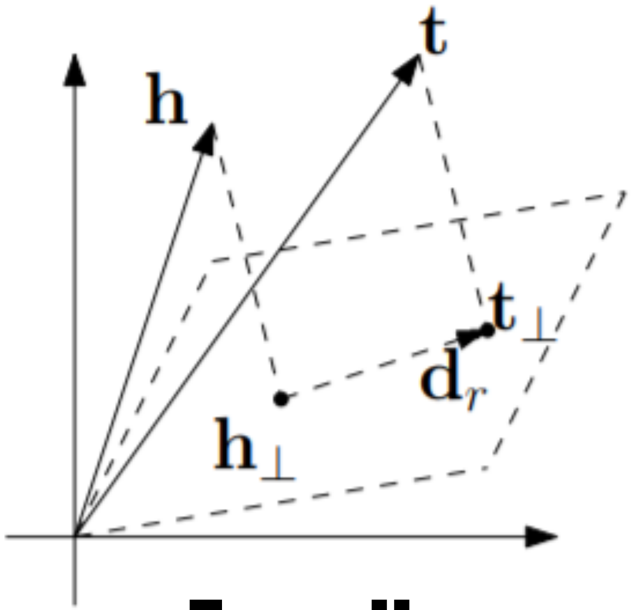


Knowledge graph embeddings

Many ways to model the problem: entities are usually vectors, relations could be vectors or matrices



TransE



TransH

Objective/loss/energy functions

What is an 'optimal' vector/matrix for an entity or relation?

Model	Score function $f_r(\mathbf{h}, \mathbf{t})$	# Parameters
TransE (Bordes et al. 2013b)	$\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{\ell_{1/2}}$, $\mathbf{r} \in \mathbb{R}^k$	$O(n_e k + n_r k)$
Unstructured (Bordes et al. 2012)	$\|\mathbf{h} - \mathbf{t}\|_2^2$	$O(n_e k)$
Distant (Bordes et al. 2011)	$\|W_{rh}\mathbf{h} - W_{rt}\mathbf{t}\|_1$, $W_{rh}, W_{rt} \in \mathbb{R}^{k \times k}$	$O(n_e k + 2n_r k^2)$
Bilinear (Jenatton et al. 2012)	$\mathbf{h}^\top W_r \mathbf{t}$, $W_r \in \mathbb{R}^{k \times k}$	$O(n_e k + n_r k^2)$
Single Layer	$\mathbf{u}_r^\top f(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s$, $W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r(sk + s))$
NTN (Socher et al. 2013)	$\mathbf{u}_r^\top f(\mathbf{h}^\top \mathbf{W}_r \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s$, $\mathbf{W}_r \in \mathbb{R}^{k \times k \times s}$, $W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r(sk^2 + 2sk + 2s))$
TransH (	$\|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r)\|_2^2$ $\mathbf{w}_r, \mathbf{d}_r \in \mathbb{R}^k$	$O(n_e k + 2n_r k)$

Existing work

Typically evaluate on Freebase and WordNet

Data	WN18	FB15K	WN11	FB13
#Rel	18	1,345	11	13
#Ent	40,943	14,951	38,696	75,043
#Train	141,442	483,142	112,581	316,232
#Valid	5,000	50,000	2,609	5,908
#Test	5,000	59,071	10,544	23,733

Wang et al. (2008)

Application 1: Triples completion

Dataset	WN18				FB15k			
Metric	MEAN		HITS@10		MEAN		HITS@10	
	Raw	Filt.	Raw	Filt.	Raw	Filt.	Raw	Filt.
Unstructured (Bordes et al. 2012)	315	304	35.3	38.2	1,074	979	4.5	6.3
RESCAL (Nickel, Tresp, and Kriegel 2011)	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE (Bordes et al. 2011)	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (Linear) (Bordes et al. 2012)	545	533	65.1	74.1	274	154	30.7	40.8
SME (Bilinear) (Bordes et al. 2012)	526	509	54.7	61.3	284	158	31.3	41.3
LFM (Jenatton et al. 2012)	469	456	71.4	81.6	283	164	26.0	33.1
TransE (Bordes et al. 2013b)	263	**251**	75.4	**89.2**	243	125	34.9	47.1
TransH (unif.)	318	**303**	75.4	**86.7**	211	**84**	42.5	**58.5**
TransH (bern.)	400.8	388	73.0	82.3	212	**87**	45.7	**64.4**

Wang et al. (2008)

Application 2: Triples classification

Dataset	WN11	FB13	FB15k
Distant Model	53.0	75.2	-
Hadamard Model	70.0	63.7	-
Single Layer Model	69.9	**85.3**	-
Bilinear Model	73.8	84.3	-
NTN	70.4	**87.1**	66.5 ($\approx 40h$)
TransE (unif.)	75.85	70.9	79.7 ($\approx 5m$)
TransE (bern.)	75.87	81.5	**87.3** ($\approx 5m$)
TransH (unif.)	**77.68**	76.5	80.2 ($\approx 30m$)
TransH (bern.)	**78.80**	83.3	**87.7** ($\approx 30m$)

Wang et al. (2008)

Code availability

Code for replicating experiments can be found at https://github.com/glorotxa/SME ; implemented using both theano/tensorflow backend

Unclear how to extend to new, sparse data, how to scale to much bigger KGs

Application 3: Featurizing nodes

E.g. Convering 'locations' into feature vectors

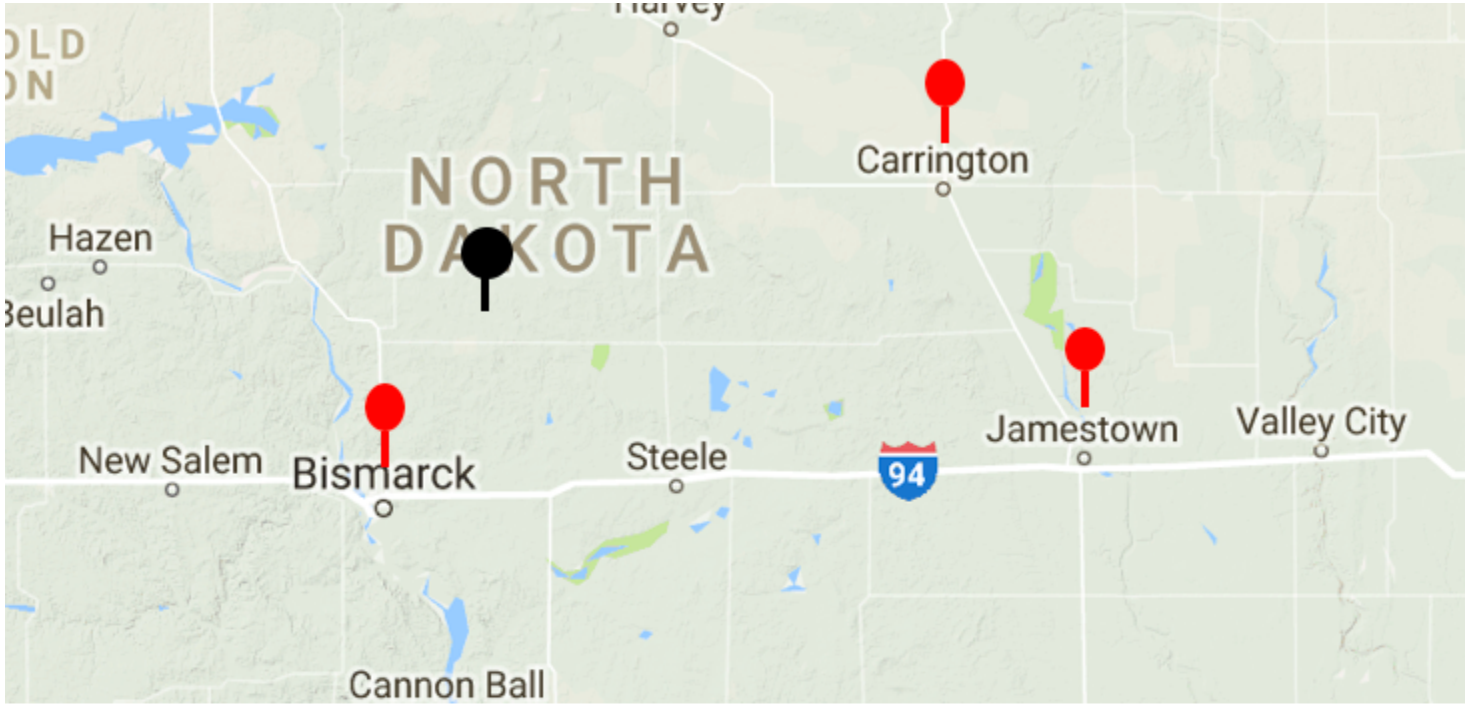
Relevant for toponym resolution, building rich graphs…

Kejriwal, Mayank; Szekely, Pedro (2017): Neural Embeddings for Populated GeoNames Locations. figshare. https://doi.org/10.6084/m9.figshare.5248120

https://github.com/mayankkejriwal/Geonames-embeddings

Features encode spatial proximity

But could encode much else, lots of room for new research!



In unusual domains…

Not clear to what extent these techniques work (which objective function is better?)

Several questions persist

- How to **acquire training data** for triples classification?
- How to **model** the concept of an entity?
- How to **regularize** with sparse mentions?

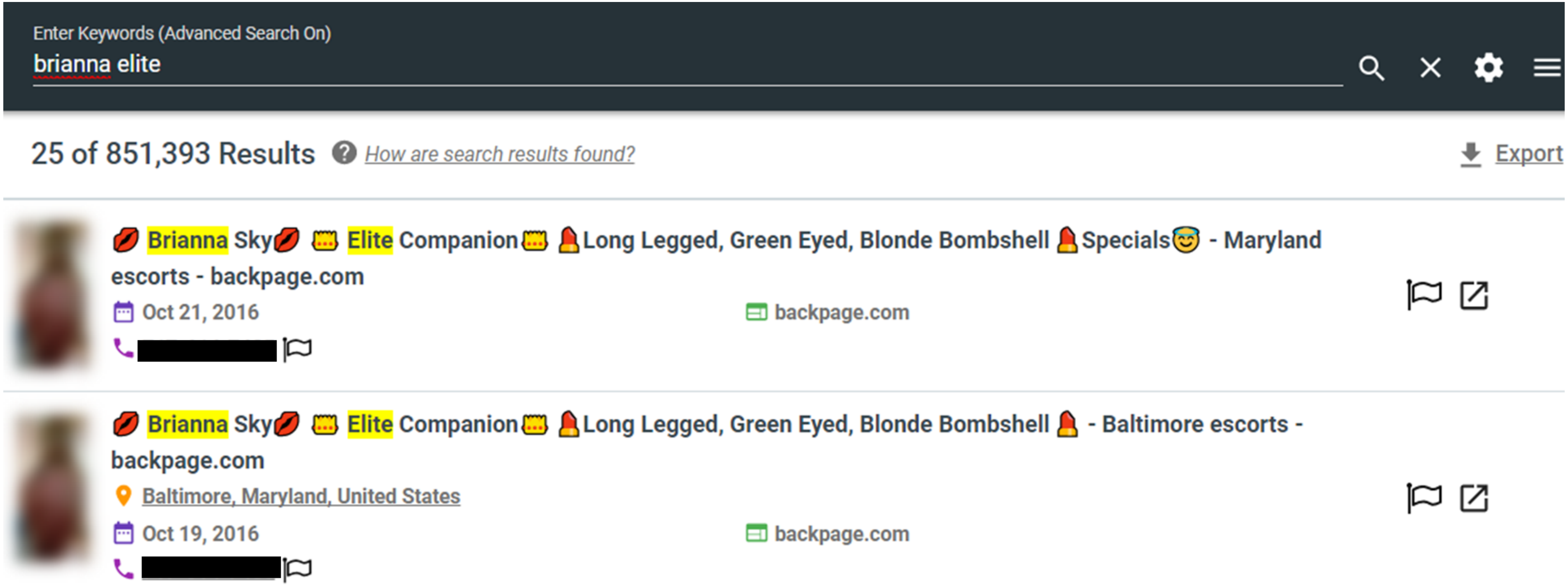
Searching Knowledge Graphs

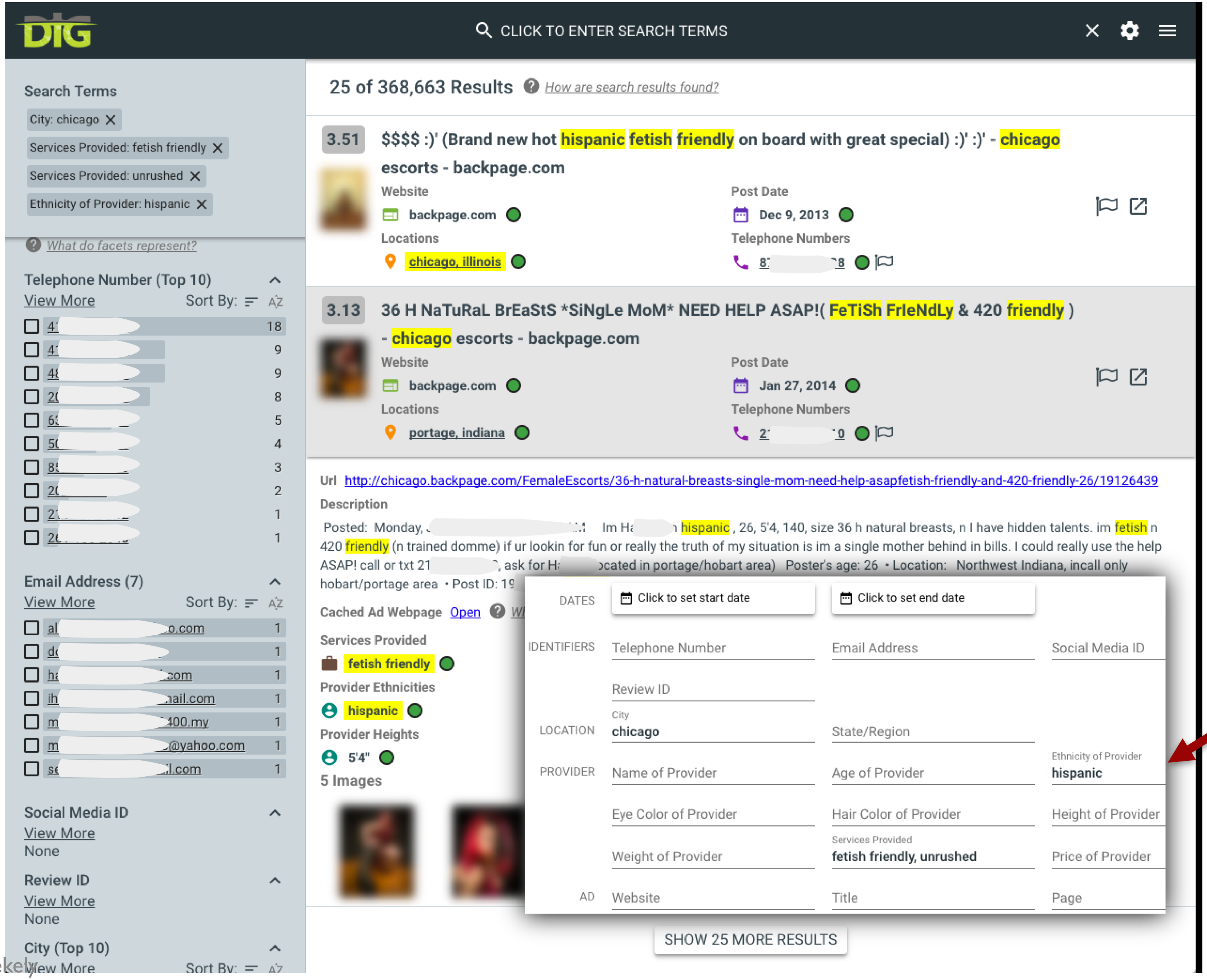
Motivation

Direct query execution does not succeed because of noise/incomplete data

What is the phone number provided in the ad that contains the email address brianna.elite@dummy-email.com, the price $600/hour, and the location manhattan midtown east 33rd & 2nd ave?

Keyword vs. faceted search





Faceted search form

Other kinds of search: many research opportunities

Clustering and network queries

Find all **massage parlor ads** linked either directly or indirectly to **phone** **12345678**

Aggregate queries

Find the **most common ethnicity** in the **massage parlor ads** linked either directly or indirectly to **phone** **12345678**

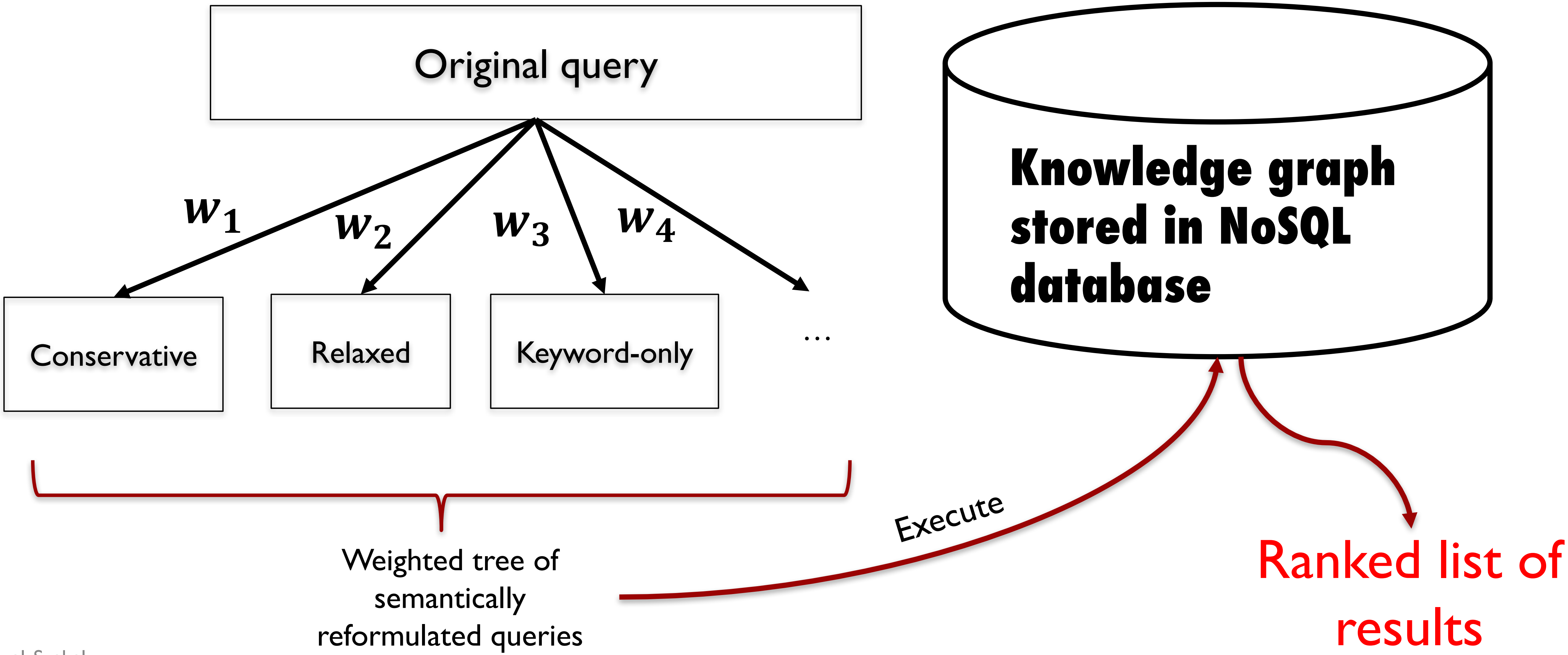
Key technique: Query reformulation

Traditional work tends to reformulate if original query returns **no results**

- Does not work well in presence of noise or in interactive session

Use **weighted combination** of reformulated queries instead, execute in NoSQL database

Semantic Query Reformulation



Other applicable techniques

Entity set learning and expansion

- E.g., can expand **red** to **red, auburn, fiery…**

Use machine learning to learn weights of reformulated sub-queries

- Relatively unexplored research area: similar to **learning to rank** in IR community, but weights strategies, not features

Wrap up and review

Summary

Unusual domains

Interesting, fun, high social impact

Knowledge graphs

Wide-spectrum of representations

Knowledge graph construction

Many tools to help, easy to create interesting graphs

Case study

Tools make it possible to build a KG in a day

Knowledge graph completion

Inferring wrong and missing links

Entity resolution

Clustering/linking mentions referring to the same underlying entity

Probabilistic soft logic

Correcting outputs of KG construction and entity resolution through logical probabilistic rules

KGs in latent space

Performing inference in low-dimensional vector spaces

Searching knowledge graphs

Using semantic query reformulation to answer complex queries over noisy knowledge graphs

Thank you!


Mona Lisa
Da Vinci
Italy
Date of birth: April 15, 1452
Date of death: May 2, 1519
(age 67 years)
Michelangelo