

# 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*



# 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 identities**

**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



**Search Terms**

Page: unrushed X City: chicago X  
 Services Provided: fetish friendly X  
 Ethnicity of Provider: hispanic X

**Price of Provider (Top 10)**

Price Range	Count
\$0-\$100 per hour	260
\$100-\$199 per hour	124
\$199-\$299 per hour	114
\$299-\$399 per hour	73
\$399-\$499 per hour	61
\$499-\$599 per hour	50
\$599-\$699 per hour	46
\$699-\$799 per hour	44
\$799-\$899 per hour	33
\$899-\$999 per hour	30

**Website (8)**

Website	Count
eroticmugshots.com	3,927
backpage.com	74
escortsincollege.com	9
escortphonelist.com	8
classivox.com	2
adultsearch.com	1
escortsinthe.us	1
massagetroll.com	1

**High Risk**

None

**CLICK TO ENTER SEARCH TERMS**

**25 of 46,190,422 Results** [How are search results found?](#) [Export](#)

**2.06** 100hh Special ?H?t C??a? ? P?a???? ? ?? R?s? 100% W??t? T??  
**V?s?t - Chicago escorts - backpage.com**

**Oct 6, 2015** [backpage.com](#)  
**Locations:** [chicago](#) **Telephone Numbers:** [323-\[REDACTED\]](#) [\[REDACTED\]](#)  
**No Email Addresses** **Provider Names:** [raina](#)

**Url:** [http://chicago.backpage.com/FemaleEscorts/100hh-special-ht-ca-pa-rs-100-wt-t-vst/2-\[REDACTED\]](http://chicago.backpage.com/FemaleEscorts/100hh-special-ht-ca-pa-rs-100-wt-t-vst/2-[REDACTED])

**Description:**  
 Hey Freaky Boys Im Raina (Cuban & Colombian Doll 2100% REAL PICS? ?INDEPENDENT? ? **UNRUSHED** & DISCREET? ?420PARTY/PARTYFRIENDLY with Donation? ? **FETISH FRIENDLY** 5'6 129 36DD ???5 STAR EXPERIENCE? ?OK TEXTING? ?PRIVATE CALLS? ?NO THUGS Or Law Enforcement? Sick Of The BORING LAZY NOT OPEN-Minded Girl Than CALL ME NOW 323-[REDACTED] Raina

**Cached Ad Webpage:** [Open](#) [What is a cached webpage?](#)

**Services Provided:** [fetish friendly](#) **Prices:** [\\$ 100 per hh](#)  
**Provider Ages:** [20](#) **Provider Ethnicities:** [hispanic](#)  
**Provider Eye Colors:** No Provider Eye Colors **Provider Hair Colors:** No Provider Hair Colors  
**Provider Heights:** [167](#) **Provider Weights:** No Provider Weights

**12 Images**

[SHOW 25 MORE RESULTS](#)

**Phone: 323-[REDACTED]**  
 435 Total Ads

**7 Other Telephone Numbers** [Copy](#)

PHONE	ADS CO-OCCURRING WITH 323-800-6151	ADS NOT CO-OCCURRING
971-[REDACTED]	20	491
720-[REDACTED]	17	620
209-[REDACTED]	11	920
928-[REDACTED]	11	551
916-[REDACTED]	9	132
510-[REDACTED]	7	139
720-[REDACTED]	3	1,397

[What does this chart represent?](#)

**2 Email Addresses** [Copy](#)

EMAIL	ADS CO-OCCURRING WITH 323-800-6151	ADS NOT CO-OCCURRING
[REDACTED]@gmail.com	12	290
[REDACTED]@words@gmail.com	1	69

[What does this chart represent?](#)

**8 Websites** [Copy](#)

WEBSITE	ADS
liveescortreviews.com	220
backpage.com	120
eroticmugshots.com	66
escortads.xxx	7
escortphonelist.com	7
escortsincollege.com	7
escortsinthe.us	7
classivox.com	1

**No Social Media IDs**

SOCIAL MEDIA ID  
 None

**No Review IDs**

REVIEW ID  
 None

**Location Drops Timeline**

**Click and drag on the chart below to zoom. Click elsewhere to reset the zoom.**

**25 of 435 Ads** [Export](#)

**100 special ? S??sat???a? C??a? ? B?? B??s? P?????t C????s ? 2Girls Avail 100% W??t? T?? V?s?t - Indianapolis escorts - backpage.com**

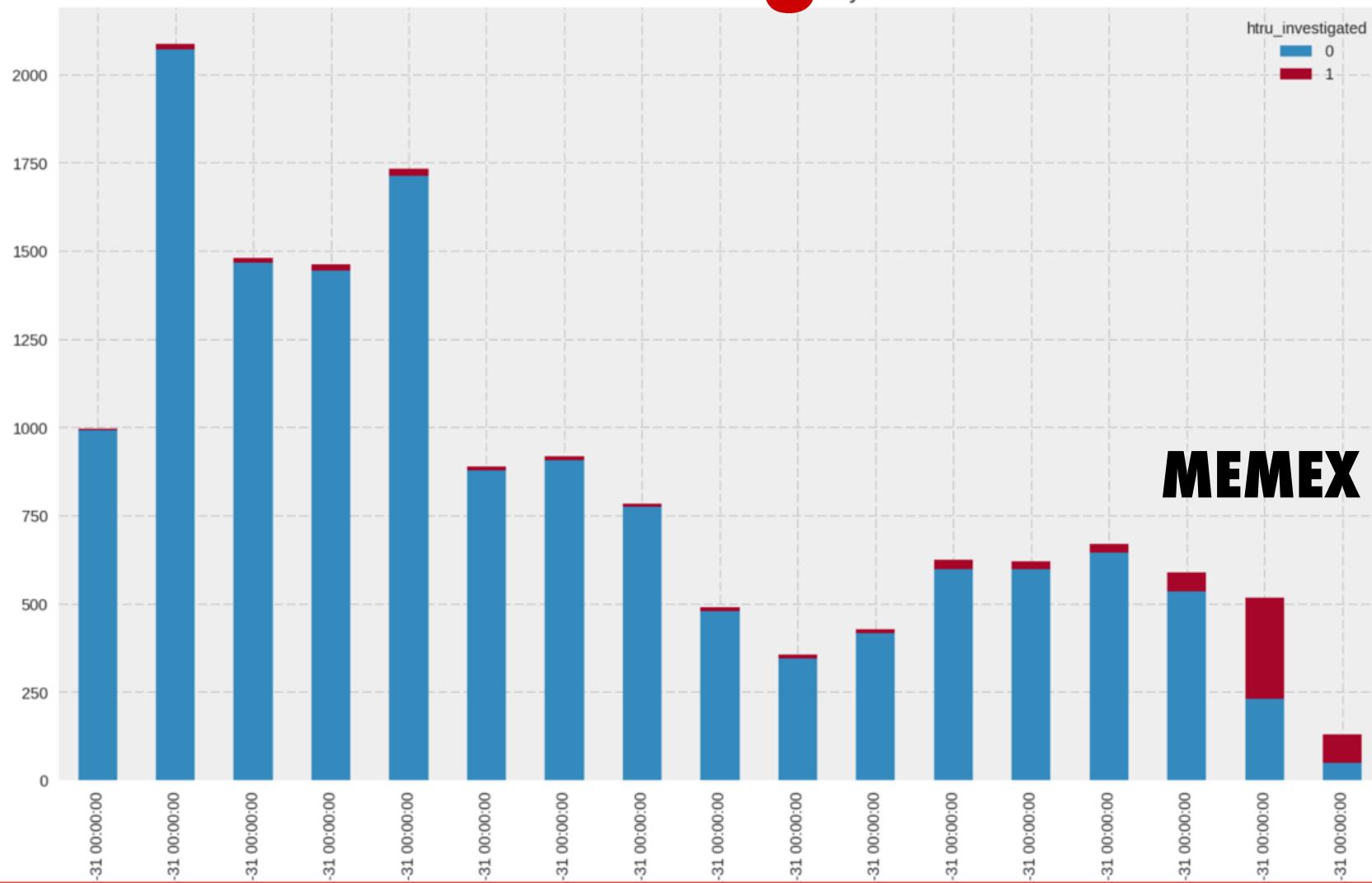
**Oct 18, 2015** [backpage.com](#)  
**Locations:** [indianapolis](#) **No Telephone Numbers**  
**No Email Addresses** **No Social Media IDs**  
**No Review IDs** **Provider Names:** [raina](#)

**100 special ? S??sat???a? C??a? ? B?? B??s? P?????t C????s ? 2Girls Avail 100% W??t? T?? V?s?t - Indiana escorts - backpage.com**

**Oct 18, 2015** [backpage.com](#)  
**Locations:** [indiana](#) **No Telephone Numbers**  
**No Email Addresses** **No Social Media IDs**  
**No Review IDs** **Provider Names:** [raina](#)

**? S??sat???a? C??a? ? B?? B??s? P?????t C????s ? P? a???? 100% W??t? T?? V?s?t - Indianapolis escorts - backpage.com**

# Number Of HT Investigations Increased



FROM <1% TO 62%

TIMBUK2

SAN FRANCISCO



\$219

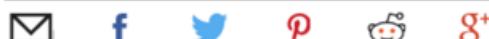


\$118

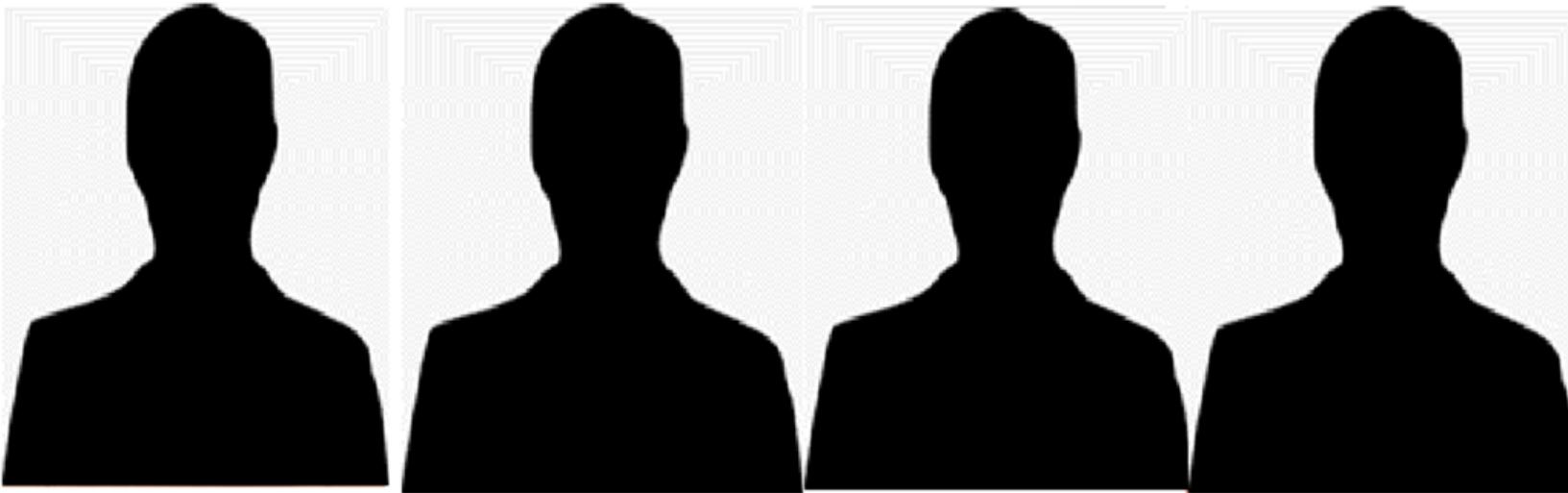


# Man sentenced to 97 years in human trafficking case

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



## 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)



Shakari 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.



A screenshot of a NCMEC "MISSING" poster for a juvenile named A. Green. The poster includes a "HELP BRING ME HOME" section with two blacked-out profile photos. Below this, detailed information is provided:

Missing Since:	Aug 13, 2016
Missing From:	Hanford, CA
DOB:	Oct 1, 2000
Age Now:	15
Sex:	Female
Race:	Black
Hair Color:	Black
Eye Color:	Brown
Height:	5'5"
Weight:	150 lbs

Text below the details states: "Both photos shown are of [REDACTED] She was last seen on August 13, 2016. [REDACTED] may go by the nickname ZahZah." A "DON'T HESITATE!" section encourages reporting information to 911 or the Hanford Police Department.

# CALIFORNIA VICTIM

## **Usual**

**Good English**

**Mostly correct information**

**Large reference datasets**

**Small number of sources**

**Sources readily available**

## **Unusual**

**Jargon, ungrammatical**

**Obfuscation & deception**

**No reference datasets**

**Long tail**

**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

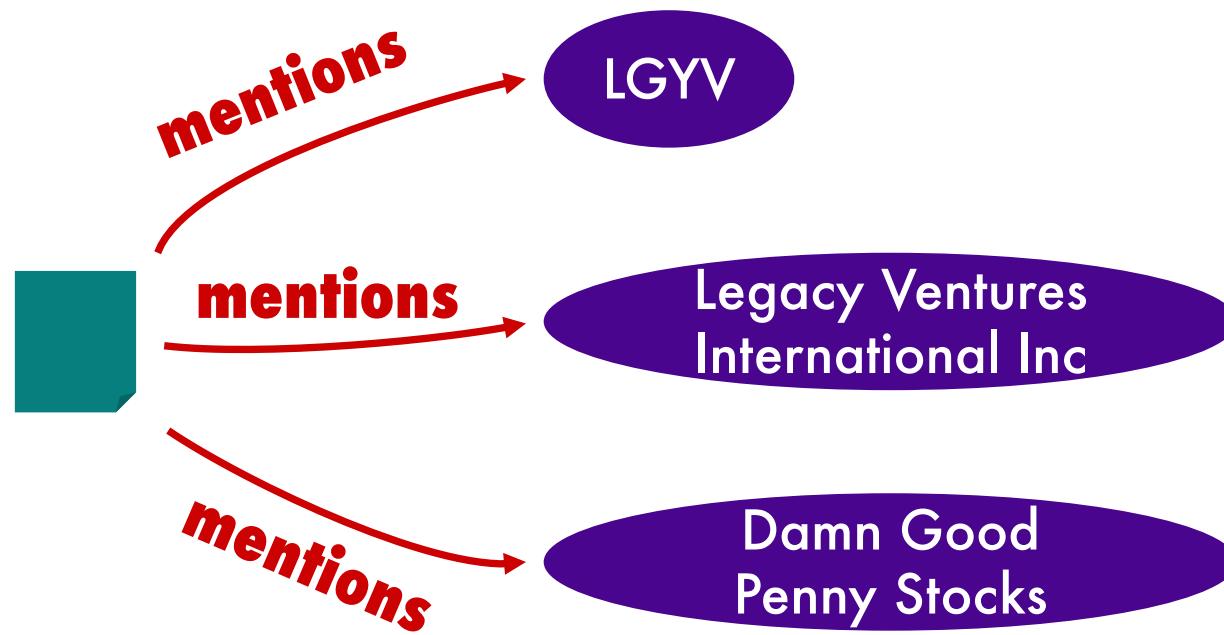
**a directed, labeled multi-relational graph  
representing facts/assertions as triples**

**(h, r, t)**      **head entity, relation, tail entity**

**(s, p, o)**      **subject, predicate, object**

# Simplest Knowledge Graph

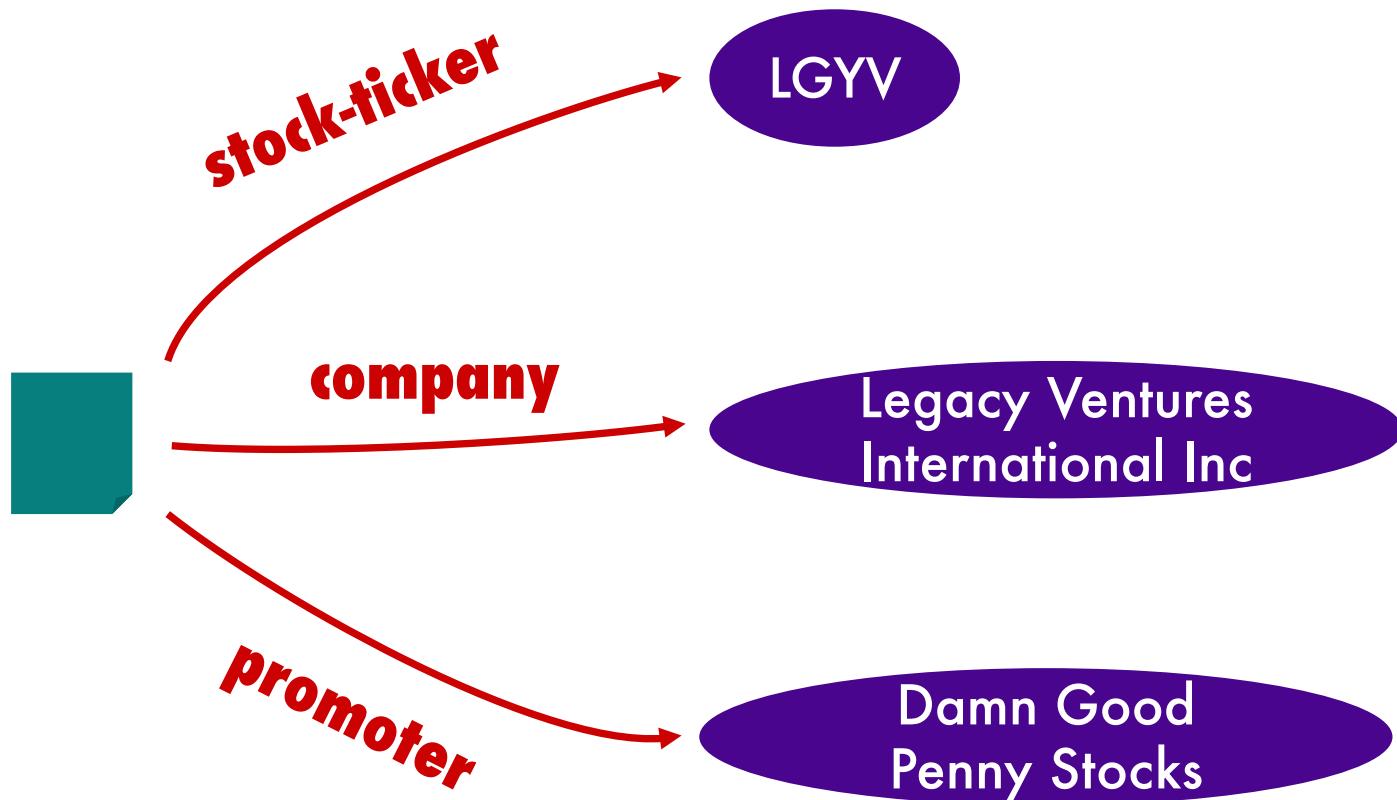
Entities



**Easiest to build**

# Simple, But Useful KG

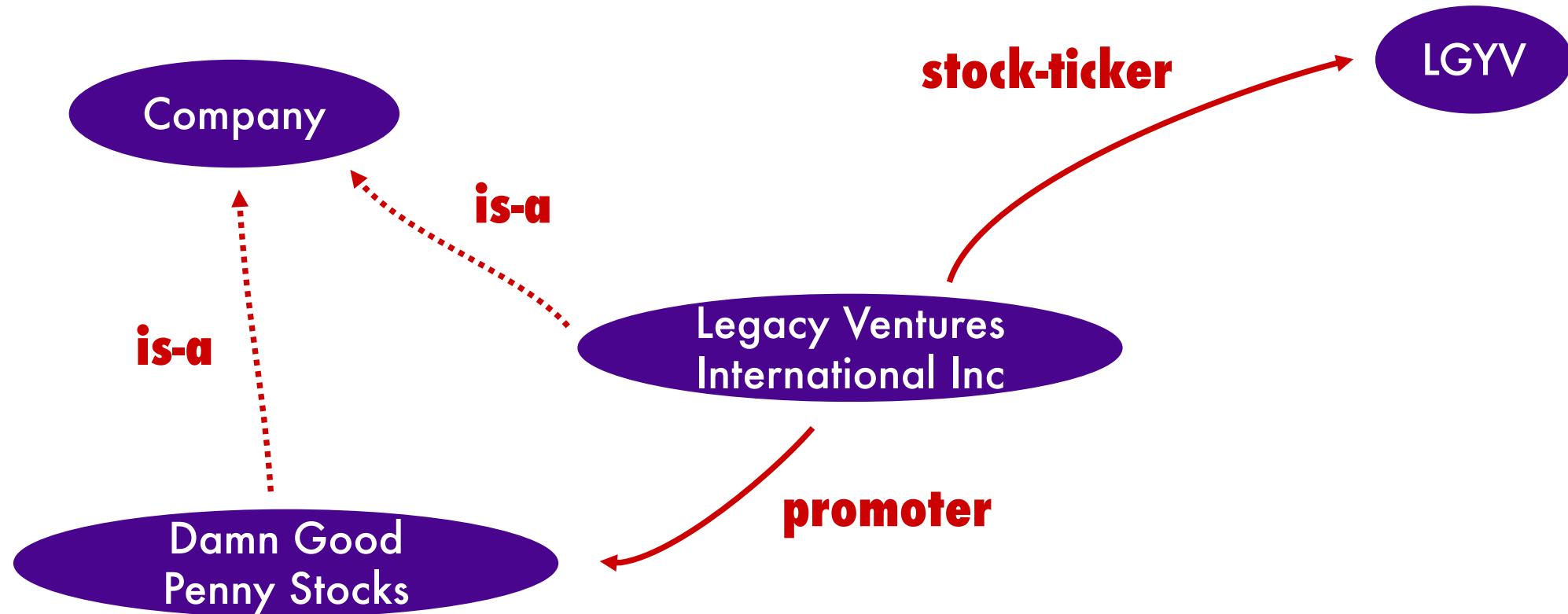
Entities + properties



**"Easy" to build**

# Semantic Web KG (RDF / OWL)

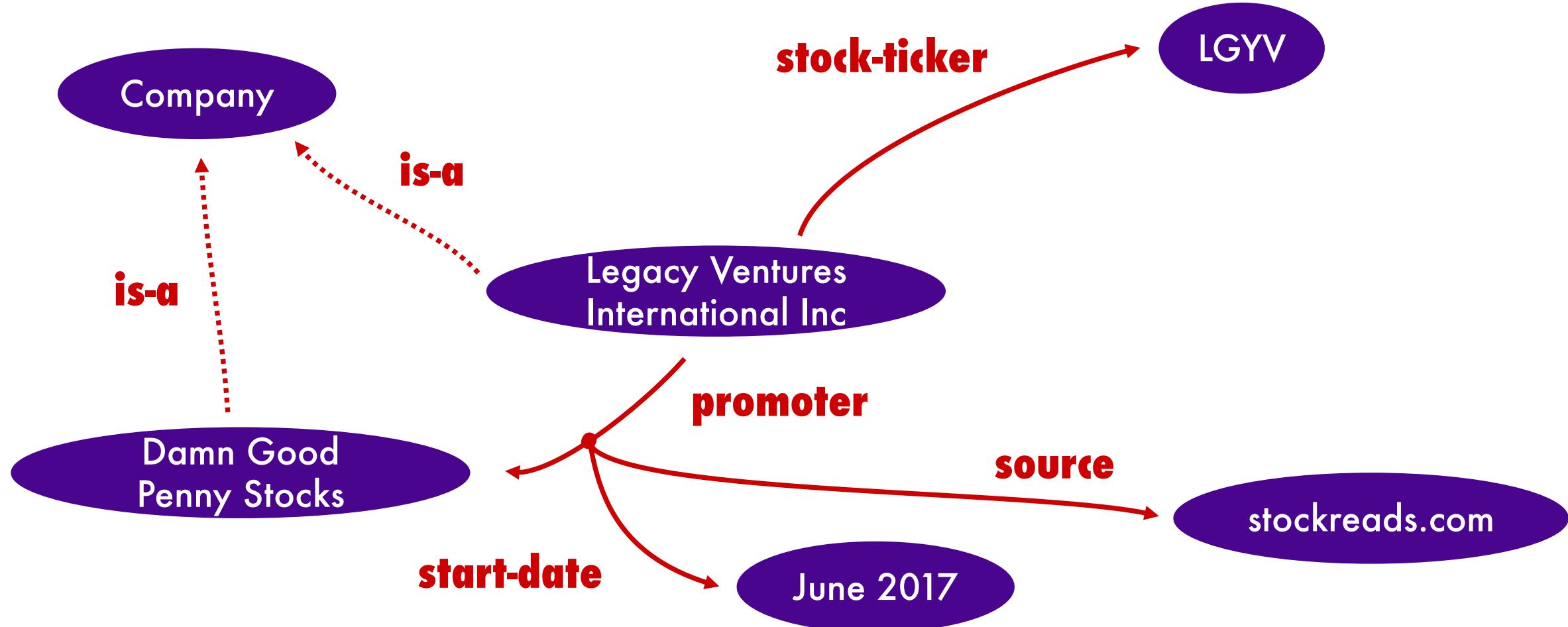
Entities + properties + classes



**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 unusual 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

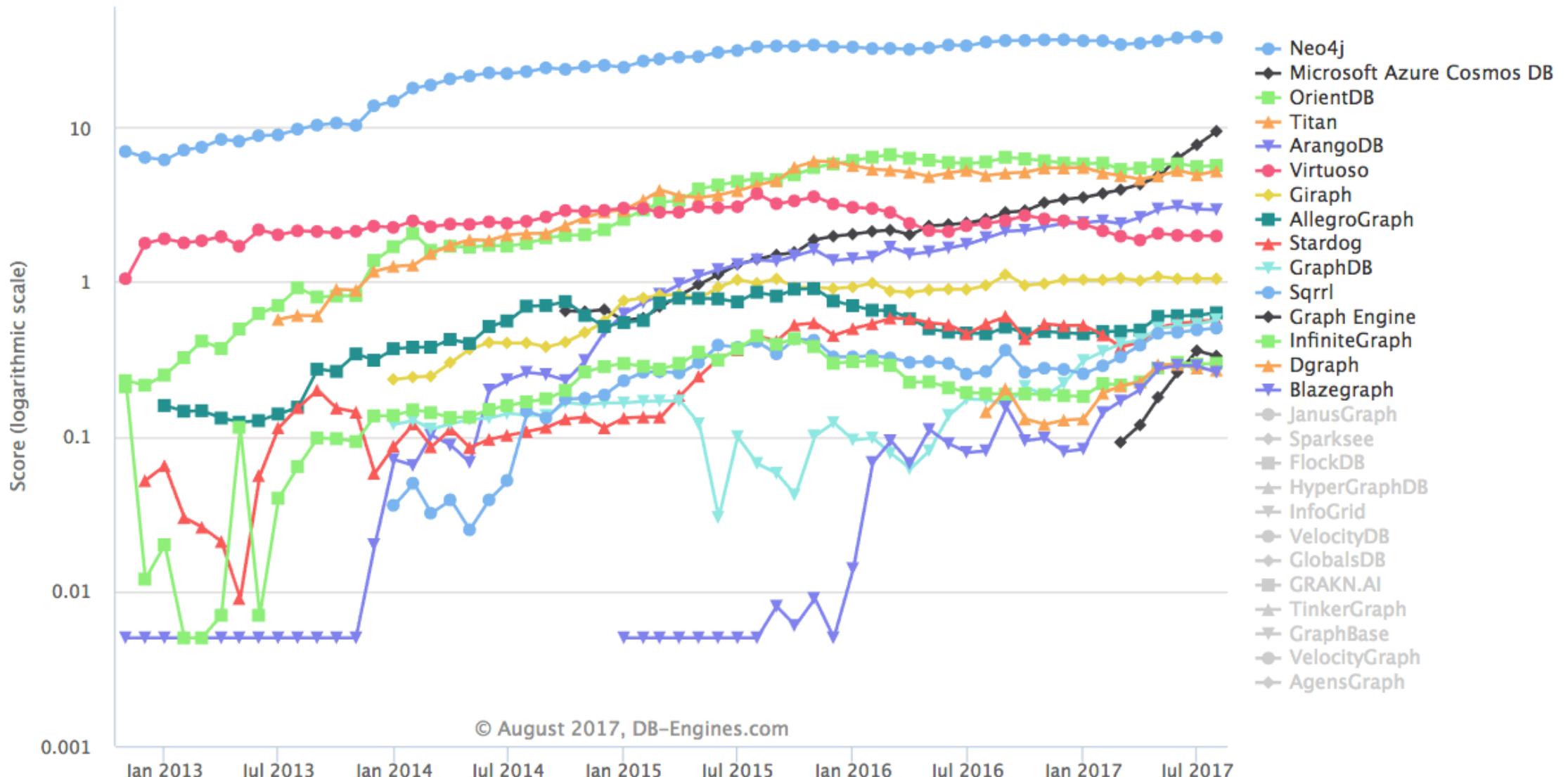
**Data model: graph**

**Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, ...**

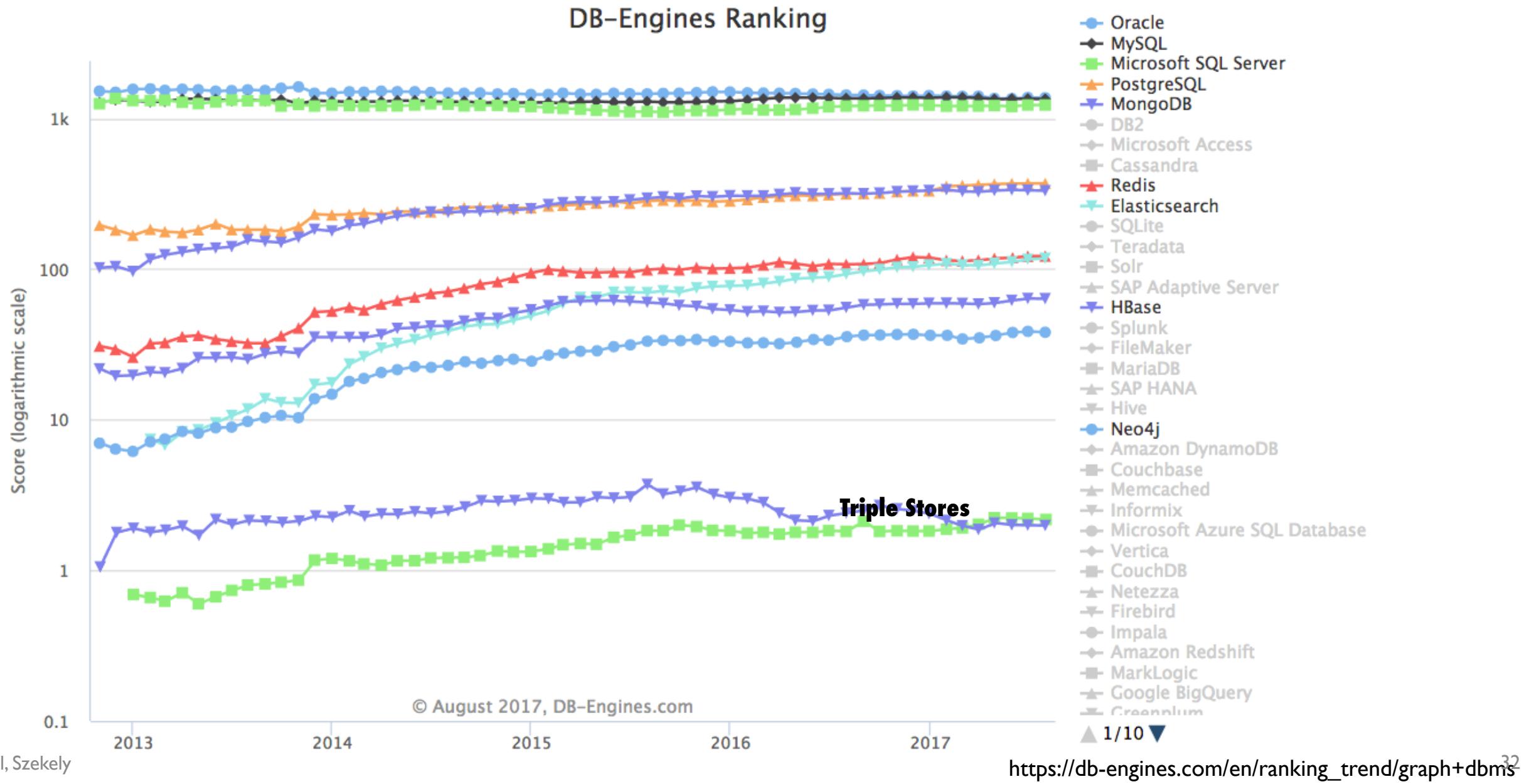
**Query: GraphQL, Gremlin, Cypher**

# Popularity Ranking Of Graph Databases

DB-Engines Ranking of Graph DBMS

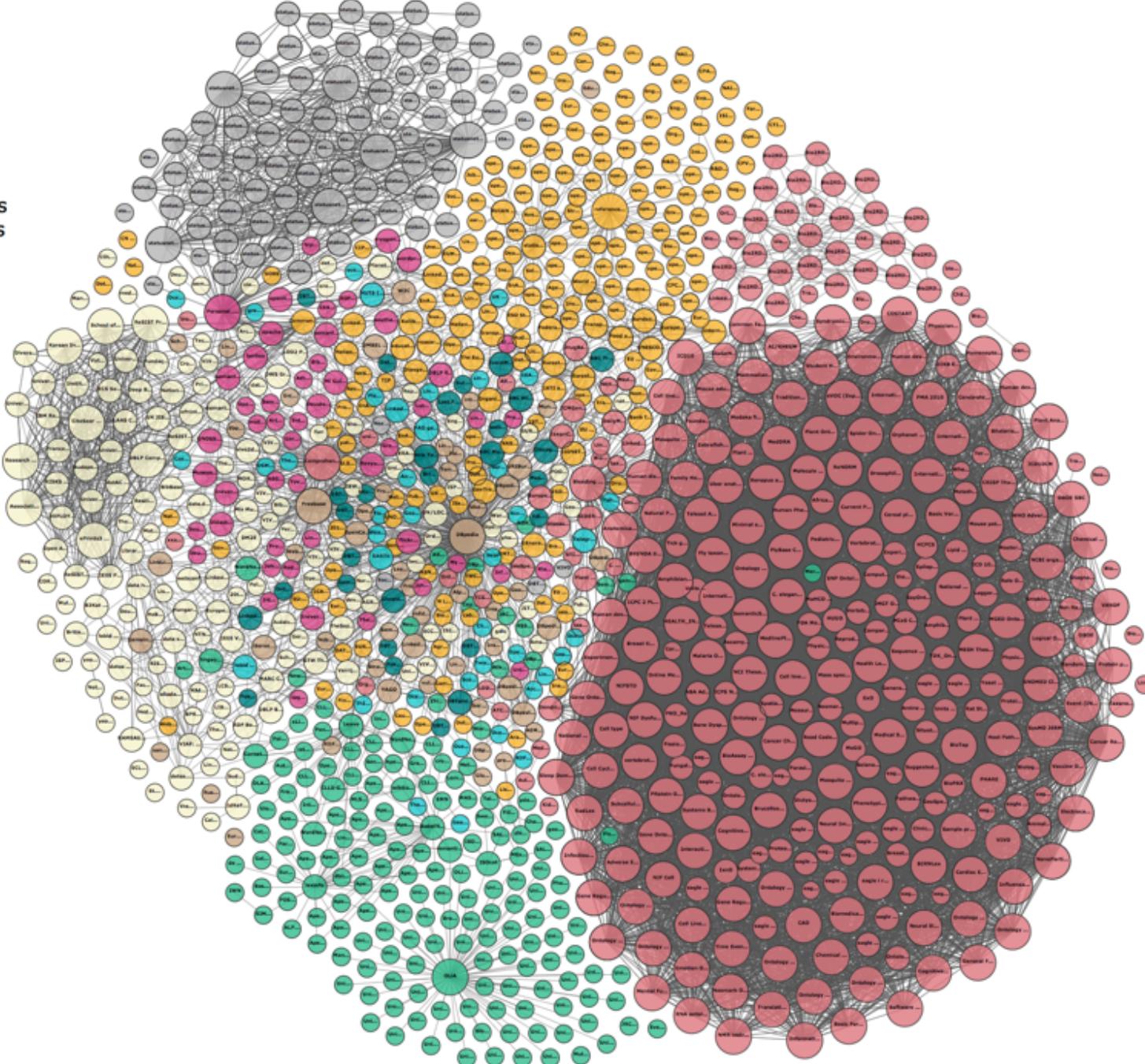


# ElasticSearch, MongoDB & Neo4J Have Wide Adoption



**KGs I can Reuse**

Legend	
Cross Domain	
Geography	
Government	
Life Sciences	
Linguistics	
Media	
Publications	
Social Networking	
User Generated	
Incoming Links	—
Outgoing Links	—



# 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](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>

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

**powers a "knowledge panel"**

**the Knowledge Graph now holds  
70 billion facts**

**search: APPL**

Apple  
Technology company



Apple Inc. is an American multinational technology company headquartered in Cupertino, California that designs, develops, and sells consumer electronics, computer software, and online services. [Wikipedia](#)

**Stock price:** AAPL (NASDAQ) US\$157.48 +2.16 (+1.39%)  
Aug 11, 4:00 PM EDT - Disclaimer

**Technical support:** 1 (800) 263-3394

**Sales:** 1 (800) 692-7753

**Founded:** April 1, 1976, Cupertino, California, United States

**Products:** iPhone, iPad, iPhone 7, iPod, Macintosh, Apple Watch,  
[MORE ▾](#)

**Founders:** Steve Jobs, Steve Wozniak, Ronald Wayne

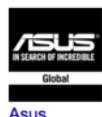
**Did you know:** Apple Inc. is the world's largest information technology company by revenue. [wikipedia.org](#)

**Profiles**



People also search for

[View 15+ more](#)



# 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**

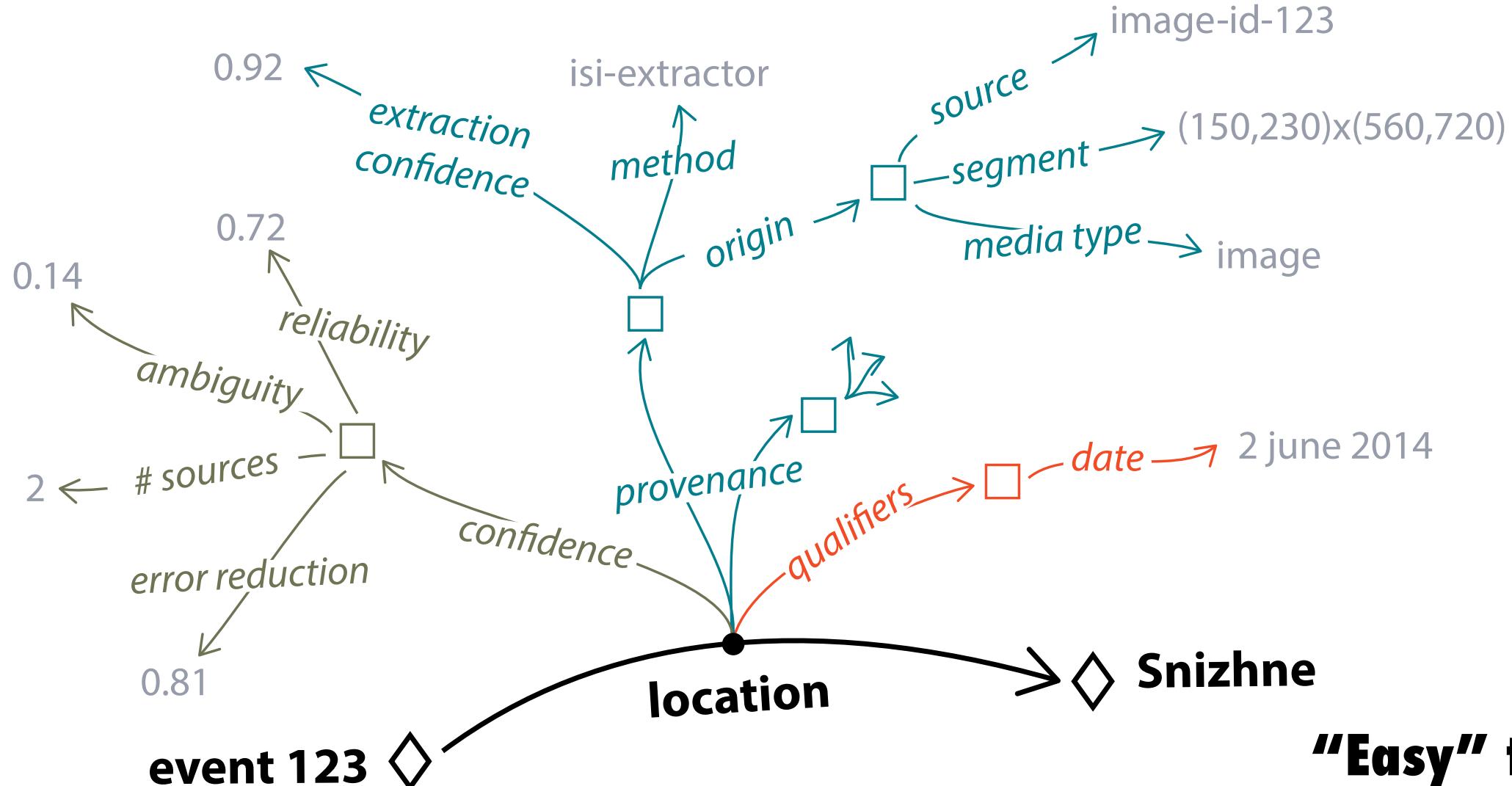
**Hybrid text/structured representation**

**keep the text, essential for machine learning and search**

**“Easy” to build**

# KG For Unusual Domains

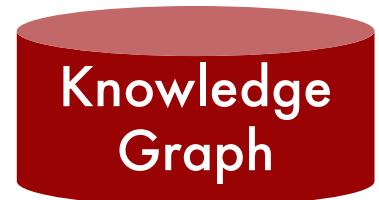
Entities + properties + provenance + confidence + qualifiers



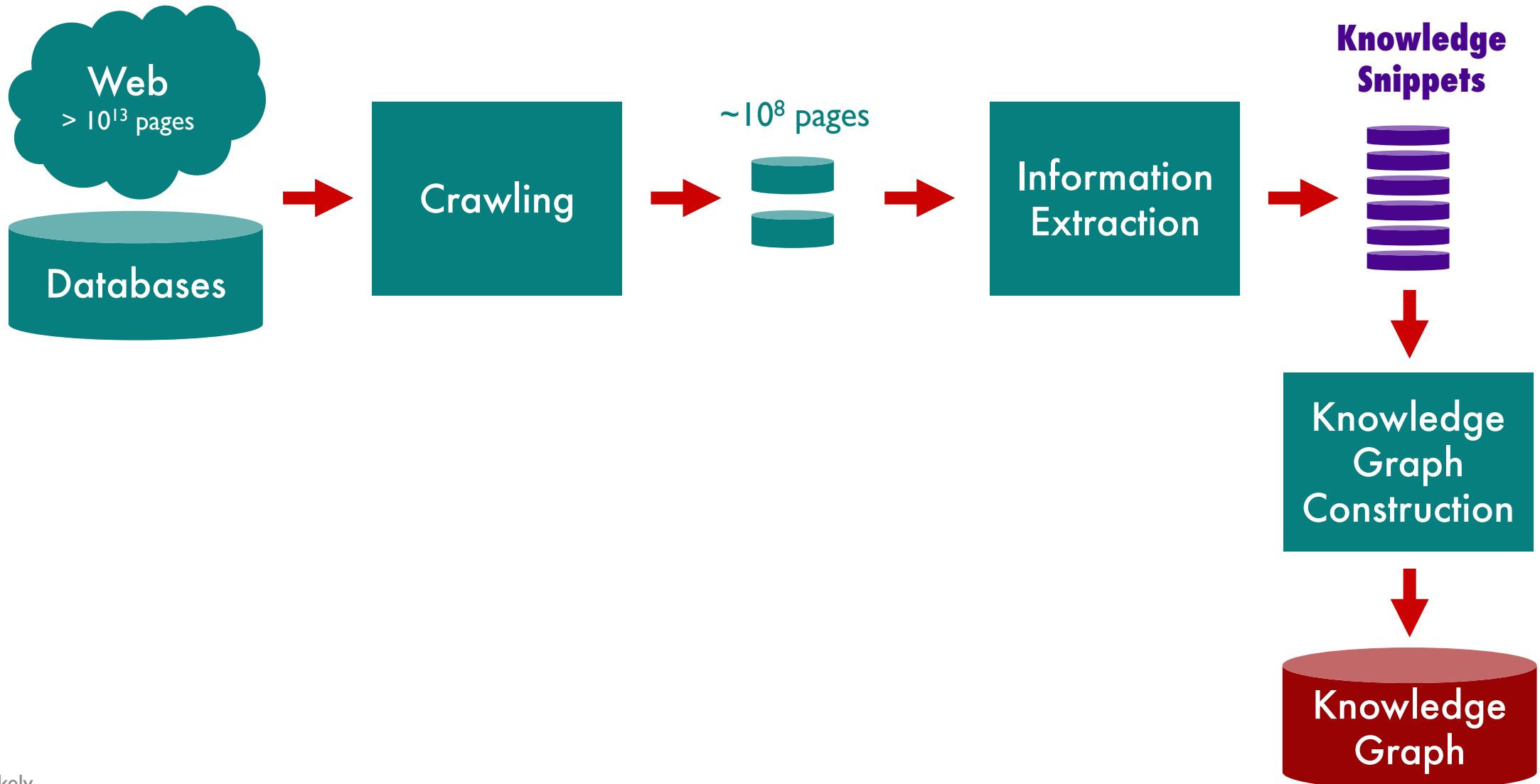
# Knowledge Graph Construction

**How can I **build** a KG?**

# KG Construction Problem

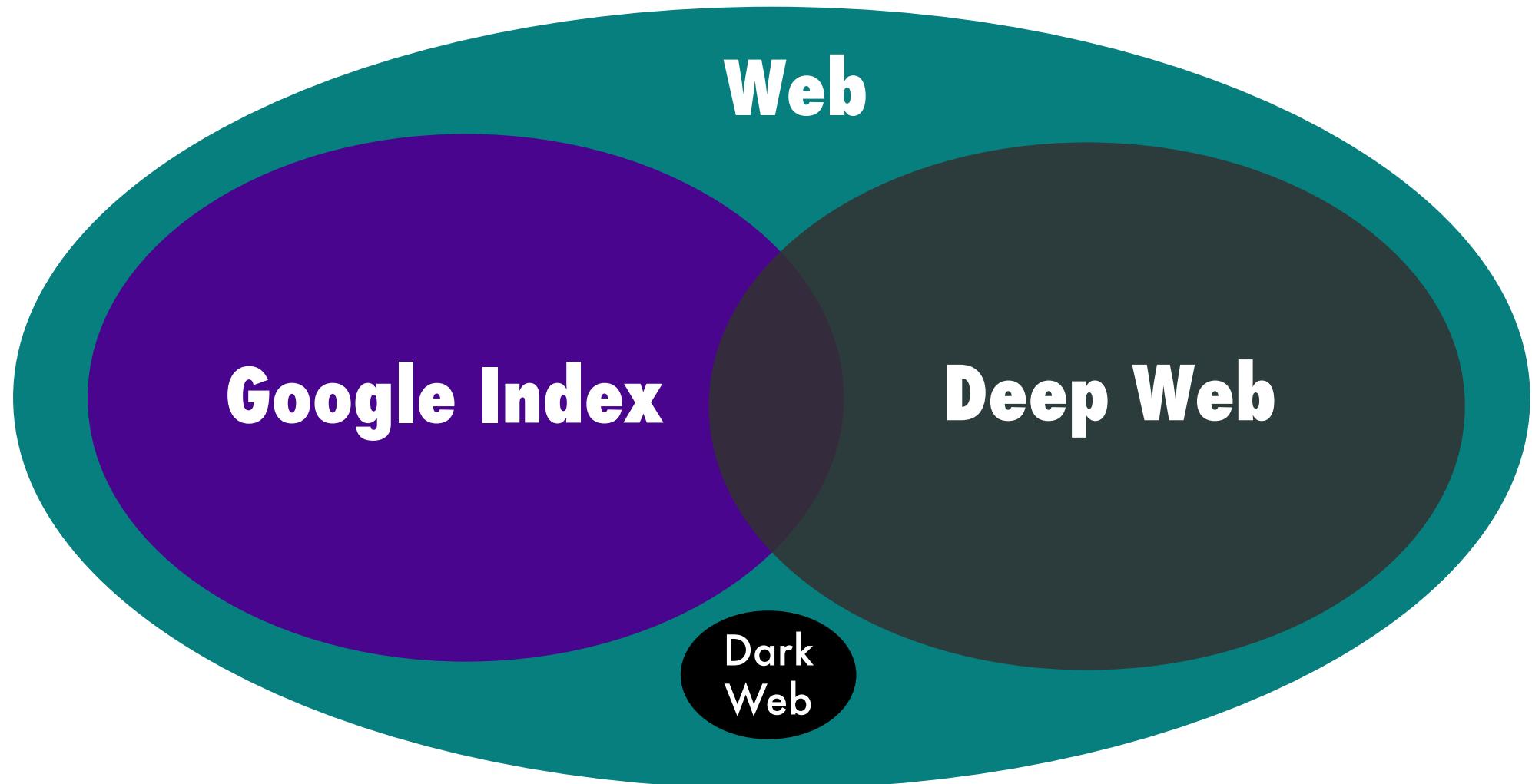


# Steps To Build a KG

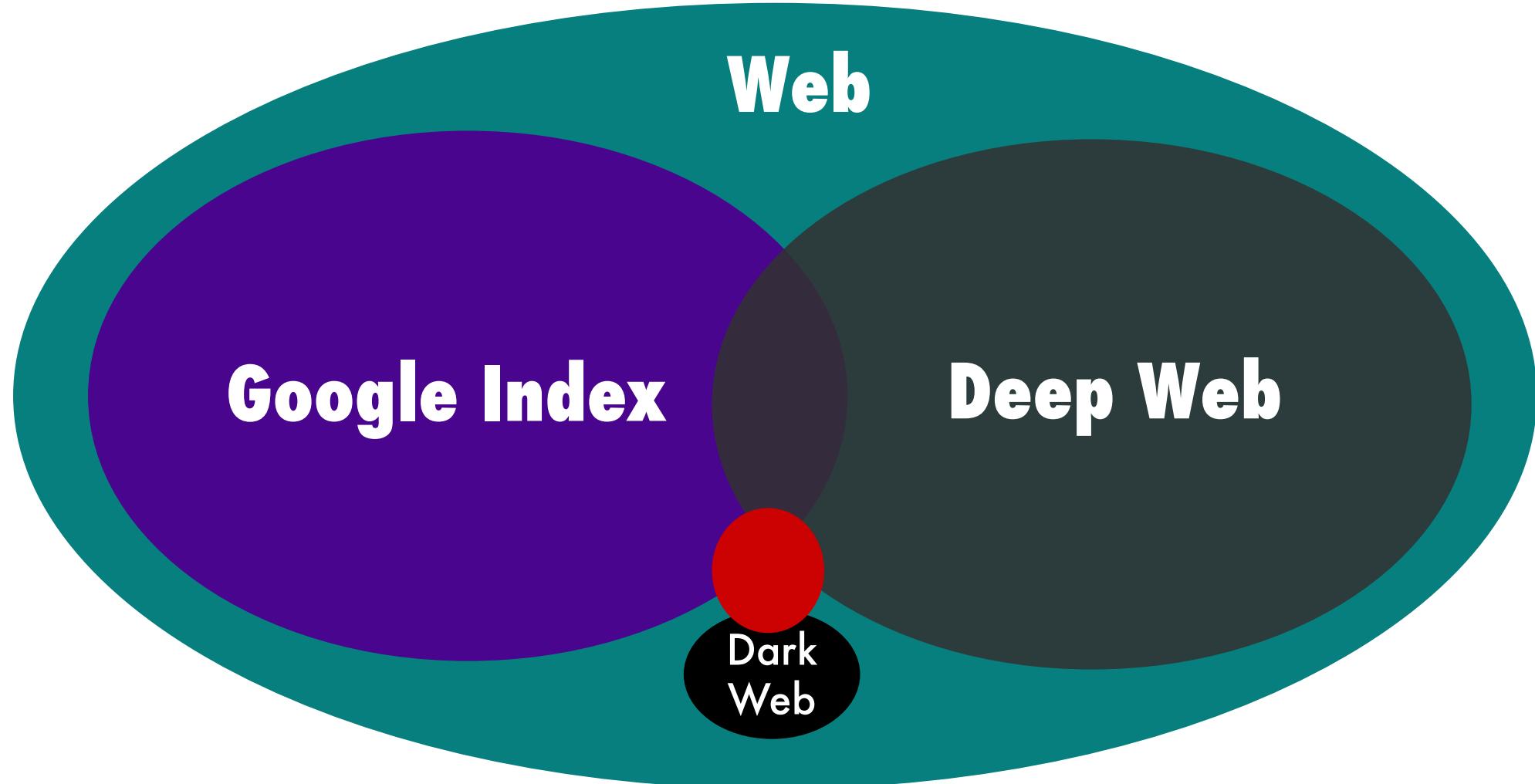


# Crawling the web

# 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>

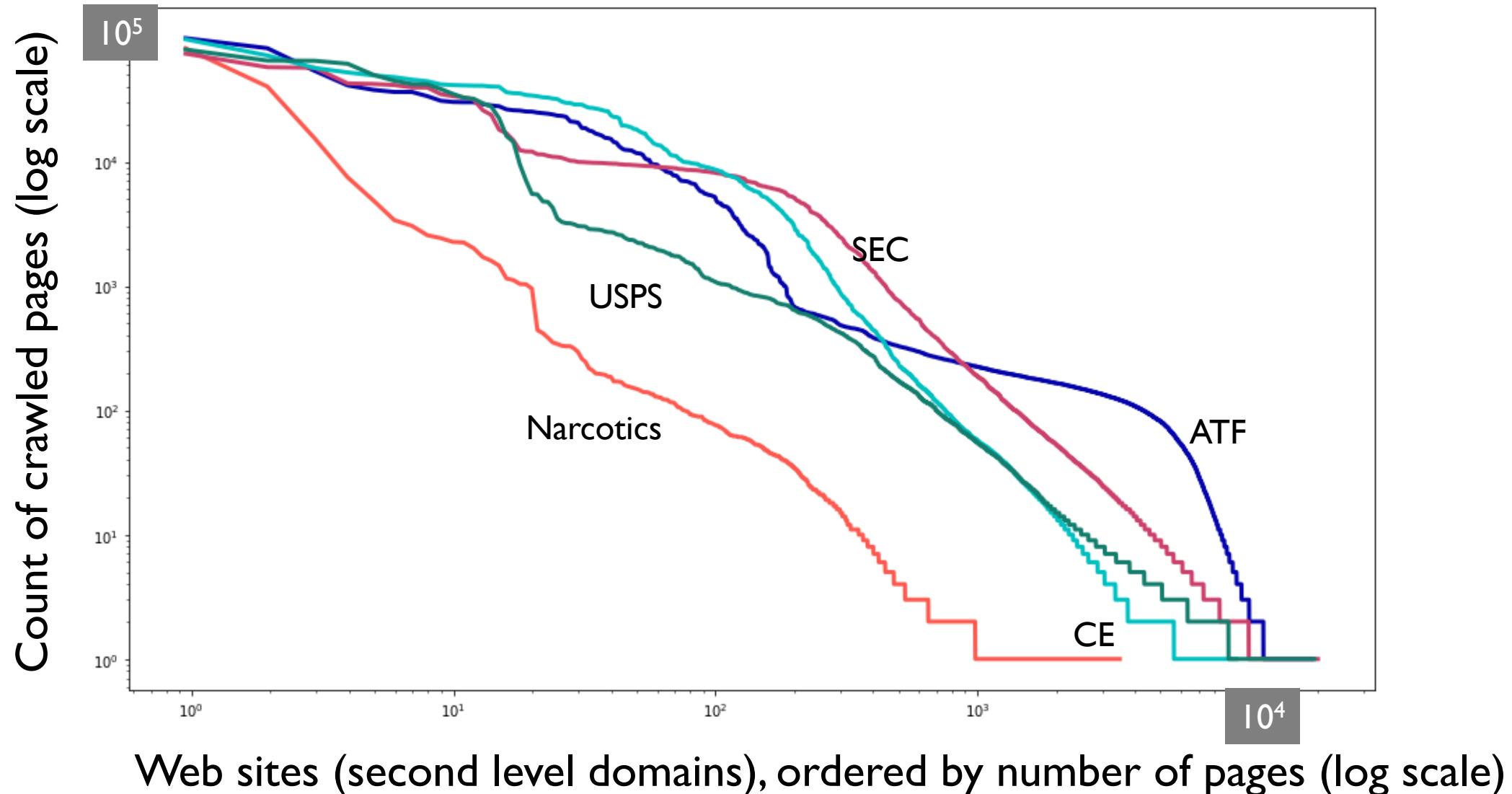
## **ACHE (focused crawler)**

<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

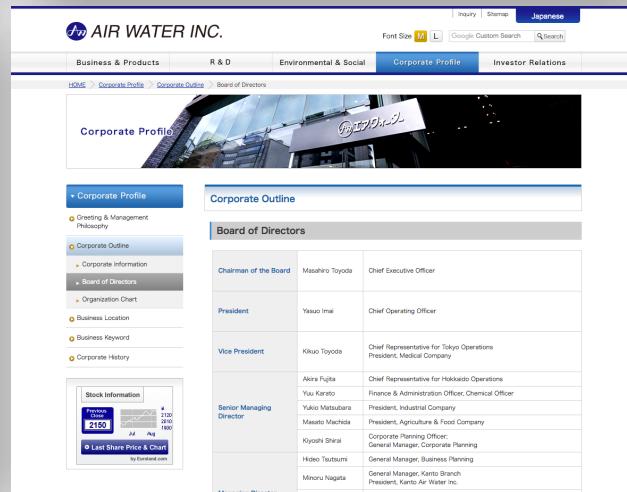
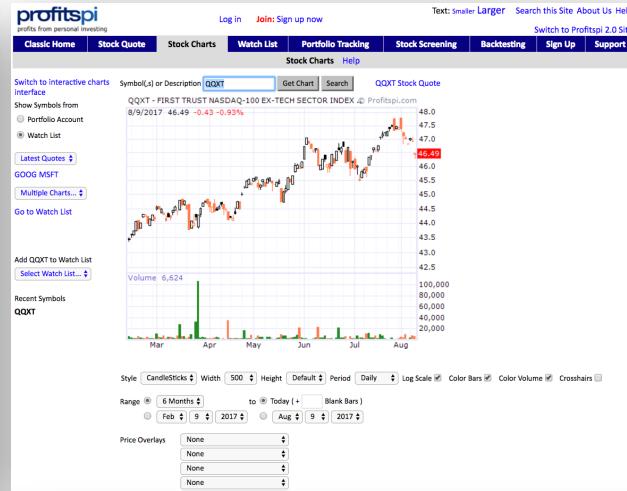
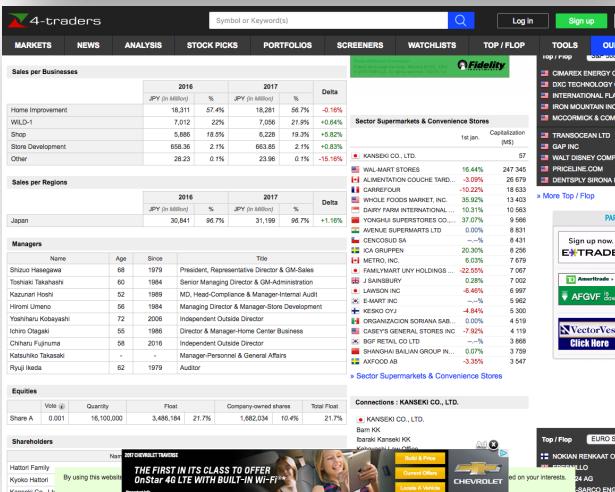
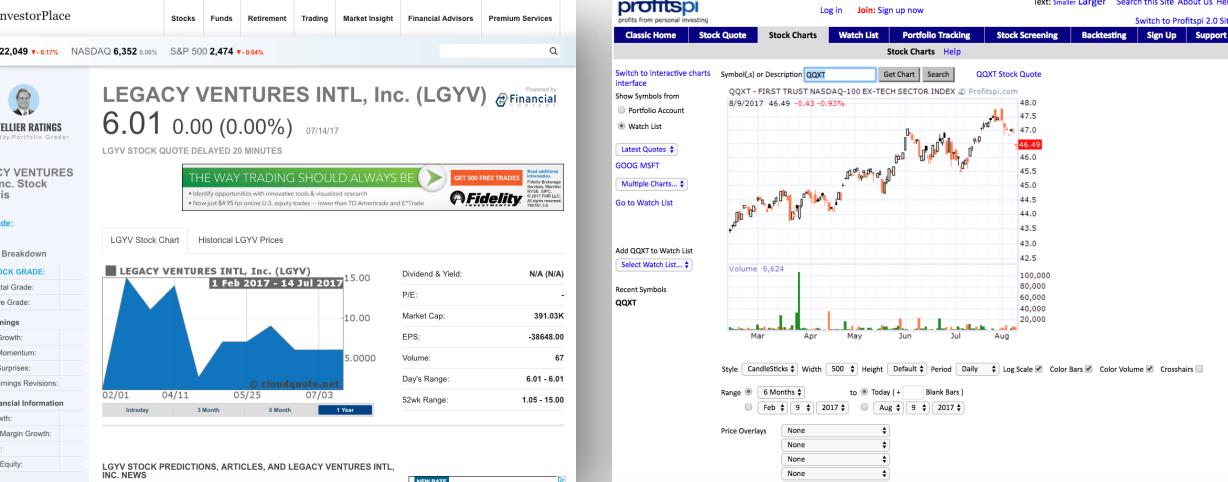
Less grammar, more formatting & linking

## 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."



# **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.

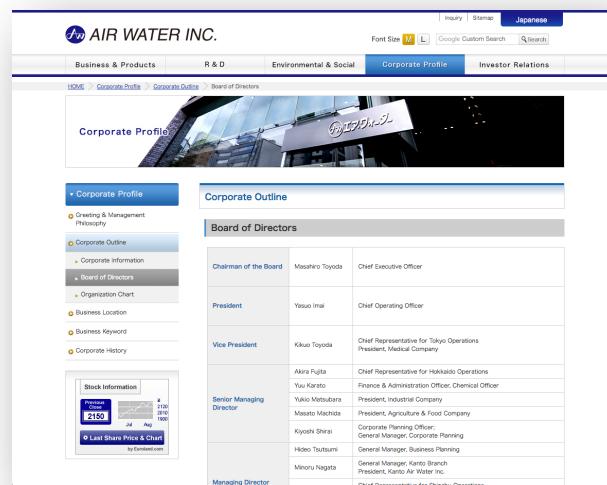
**Frank Huybrechts** - COO  
Mr. Huybrechts has over 20 years of

- Press Contact
  - General information
  - Directions maps

## Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	<a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a>	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	<a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a>	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	<a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a>	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	<a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a>	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			
Cohen, Paul R.	(413) 545-3638	<a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a>	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			

## Tables



Chairman of the Board	Masahiro Toyoda	Chief Executive Officer
President	Yasuo Imai	Chief Operating Officer
Vice President	Kikuo Toyoda	Chief Representative for Tokyo Operations President, Medical Company
Senior Managing Director	Akira Fujita Yuu Karato Yukio Matsubara Masato Machida Kiyoshi Shirai Hideo Totsunari Minoru Nagata	Chief Representative for Hokkaido Operations Finance & Administration Officer, Chemical Officer President, Industrial Company President, Agriculture & Food Company Corporate Planning Officer General Manager, Building Planning General Manager, Sales & Marketing President, Kansai Air Water Inc.
Managing Director	Chief Representative for Shoboku Operations	

## Charts



# Scope

## Web site specific

**CHARTER COMM RG-A (CHTR)**  
400.90 -11.15 (2.86%) 20:10 EDT  
CHTR STOCK QUOTE DELAYED 20 MINUTES

**Analysis Breakdown**

CHTR STOCK GRADE:	B
Fundamental Grade:	D
Quantitative Grade:	A
CHTR Earnings	
Earnings Growth:	F
Earnings Momentum:	F
Earnings Surprises:	F
Analyst Earnings Revisions:	D
CHTR Financial Information	
Sales Growth:	A
Operating M	
Cash Flow:	
Return on E	

**LEGACY VENTURES INTL, Inc. (LGYV)**  
6.01 0.00 (0.00%) 07/14/2017  
LGYV STOCK QUOTE DELAYED 20 MINUTES

**Analysis Breakdown**

LGYV STOCK GRADE:	B
Fundamental Grade:	D
Quantitative Grade:	B
LGYV Earnings	
Earnings Growth:	F
Earnings Momentum:	F
Earnings Surprises:	F
Analyst Earnings Revisions:	D
LGYV Financial Information	
Sales Growth:	
Operating Margin Growth:	
Cash Flow:	
Return on Equity:	

## Genre specific (e.g., forums)

**TheLion.com Central**  
Welcome Stranger! Please sign up or log in to enable additional features.  
Forum - TheLion.com Central like 1.1K  
Discuss everything about TheLion.com

**Stockaholics**  
MAIN FORUMS GROUPS LIVE CHARTS CHAT TOOLS Page 1 of 1  
If You Ignore This \$0.22 Stock - You'll Never Forgive Yourself  
Do Not Sit By And Watch Your Neighbor Get Rich Off This Stock. See For Yourself personal message on  
200% Gains in 2 Hours - Our Stock Picks are Insane  
Buy These Stocks Now  
Morn Raises Credit Score

**InvestorsHub**  
Boards Hot Tools Streamer Level 2 Follow Feed  
Opportunity is Everywhere if you know where to look. Get Started at ET TRADE.

## Wide, non-specific

**AIR WATER INC.**  
Business & Products R & D Environmental & Social Corporate Profile Investor Relations

**profitspsi**  
Symbol(s) or Description QQXT Get Chart Search QQXT Stock Quote  
QQXT - FIRST TRUST NASDAQ-100 EX-TECH SECTOR INDEX 48.0  
8/9/2017 46.49 -0.43 -0.93%  
GOOG MSFT  
Multiple Charts... Go to Watch List

**4-traders**  
MARKETS NEWS ANALYSIS STOCK PICKS PORTFOLIOS SCREENSERS WATCHLISTS TOP / FLOP  
Sales by Businesses  
2016 2017 Delta  
JPY (in billions) % JPY (in billions) %  
Home Improvement 18.311 57.4% 18.281 56.7% -0.15%  
WLD-1 7.012 22% 7.056 21.9% +0.04%  
Shop 5,666 18.5% 6,228 19.3% +0.82%  
Store Development 656.26 2.1% 683.85 2.1% +0.03%  
Other 28.23 0.1% 23.96 0.1% -0.16%  
  
Sales by Regions  
2016 2017 Delta  
JPY (in billions) % JPY (in billions) %  
Japan 30,841 99.7% 31,199 99.7% +1.16% 9,966  
  
Managers  
Name Age Since Title  
Uchimura Koji 60 1984 President, Representative Director & CEO  
Tsurumi Toshiharo 60 1984 Senior Managing Director & GM Administration  
Kikuchi Keishi 52 1989 MD Head Compliance & Manager-Internal Audit  
Hirata Umeno 56 1988 Managing Director & Manager-Store Development  
Yoshitoku Kobayashi 72 2004 Independent Outside Director  
Ishii Osaki 55 1988 Director & Manager-Home Center Business  
Chihara Fujunaga 58 2016 Independent Outside Director  
Katsuhiro Takasaki - - Manager-Personnel & General Affairs  
Ryuji Ieda 62 1979 Auditor  
  
Equities  
Name Sector Supermarkets & Convenience Stores  
KANEKI CO., LTD. 5.02% 7.07%  
FAMILIA-MART UNY HOLDINGS -25.55% 7.00%  
CARREFOUR -0.05% 26.79%  
WORLDCLOUD MARKET INC. -0.22% 18.633  
YAMAZAKI FARM INTERNATIONAL -0.31% 14.602  
YONKUJI SUPERSTORES CO., LTD. 37.07% 9.966  
AVENUE SUPERMARKETS LTD 0.00% 8.431  
CA GRUPPI 0.00% 4.811  
METRO, INC. 5.02% 9.296  
FAMILIA-MART UNY HOLDINGS -25.55% 7.00%  
CARREFOUR -0.05% 26.79%  
WORLDCLOUD MARKET INC. -0.22% 18.633  
YAMAZAKI FARM INTERNATIONAL -0.31% 14.602  
YONKUJI SUPERSTORES CO., LTD. 37.07% 9.966  
AVENUE SUPERMARKETS LTD 0.00% 8.431  
CA GRUPPI 0.00% 4.811  
METRO, INC. 5.02% 9.296  
FAMILIA-MART UNY HOLDINGS -25.55% 7.00%  
CARREFOUR -0.05% 26.79%  
WORLDCLOUD MARKET INC. -0.22% 18.633  
YAMAZAKI FARM INTERNATIONAL -0.31% 14.602  
YONKUJI SUPERSTORES CO., LTD. 37.07% 9.966  
AVENUE SUPERMARKETS LTD 0.00% 8.431  
CA GRUPPI 0.00% 4.811  
METRO, INC. 5.02% 9.296  
  
Connections : KANEKI CO., LTD.  
KANEKI CO., LTD. 0.00% 0.00%  
Bain KK 0.00% 0.00%  
Isaruk Kansai inc. 0.00% 0.00%  
Kaneko Co., Ltd. 0.00% 0.00%  
  
Shareholders  
Name Sector Supermarkets & Convenience Stores  
NOKIAN REKAATTI OYJ 0.00% 0.00%  
Top / Flop EURO STOCK  
NOKIAN REKAATTI OYJ 0.00% 0.00%  
CHEVROLET 0.00% 0.00%  
SARCO ENGI. 0.00% 0.00%

# 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"

[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 Providerdatches ck dl6472fp 411 p98690

phone:773 431 8174 REFERENCES REQUIREDDBDSM, Domme, & Fetishes Available | www.delialondon.com | Call **647-241-1986**. See my menu of services on my profile  
[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 01, 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 06, 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 ; \) L A T I N A - 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](#)
- Chicago, Illinois (312-699-8787)  
[P A R T Y ! ! !](#)
- Las Vegas, Florida (609-100-0000)  
[I M M E R S I O N S ! ! !](#)
- Atlanta, Georgia (401-324-9388)  
[WoW. MuSt TaKe A LoOk At ThIs. - 21](#)
- Atlanta, Georgia (347-940-1982)  
[SMOKING HOT Specials BuSTa BaBe \(\( 5 SeRvICe \)\) Pretty 36DDDS \(\) \(](#)
- Atlanta, Georgia (404-955-1000)  
[The Best Companion in ALL M... Curvy, Curvy Girl](#)
- Atlanta, Georgia (404-955-1000)  
[Beautiful Salvadoran 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)  
[\\*\\*21 years old \\* \\$80 \\*\\*real pictures \\*\\* A sian Kathy \\*\\*\\* - 21](#)

Top Escort Cities

- [New York, New York](#)
- [Toronto, Ontario](#)
- [Dallas, Texas](#)
- [Chicago, Illinois](#)
- [Atlanta, Georgia](#)
- [North Jersey, New Jersey](#)
- [Detroit, Michigan](#)
- [Las Vegas, Nevada](#)
- [Los Angeles, California](#)
- [San Jose, California](#)
- [Houston, Texas](#)
- [Phoenix, Arizona](#)
- [Philadelphia, Pennsylvania](#)
- [Boston, Massachusetts](#)
- [Washington, DC](#)
- [Chicago, Illinois](#)
- [Las Vegas, Nevada](#)
- [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 Unusual Domains

Full spectrum of in all dimensions

Document Features	Scope	Pattern Complexity	Relevance
<ul style="list-style-type: none"><li>text paragraphs</li><li>grammatical, some formatting</li><li>ungrammatical, rich formatting</li><li>tables</li><li>charts</li></ul>	<ul style="list-style-type: none"><li>website specific</li><li>genre-specific</li><li>wide</li></ul>	<ul style="list-style-type: none"><li>closed set</li><li>regular set</li><li>complex pattern</li><li>ambiguous pattern</li><li>unusual language model</li></ul>	<ul style="list-style-type: none"><li>all relevant</li><li>significant irrelevant content</li><li>active deception</li><li>purposeful obfuscation</li></ul>

**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

# 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

thehotpennystocks.com  
thelion.com  
theotc.todtraders.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

Actions

Tables

Fields

Tags

Glossaries

Sample DigApp

## Get Sample Pages

[Link to Inferlink Tool](#)

Retrieve sample pages from the CDR for testing. This will collect pages from each of your TLDs as well as a random selection of pages from other TLDs

Number of Pages per TLD

# DEMO

[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

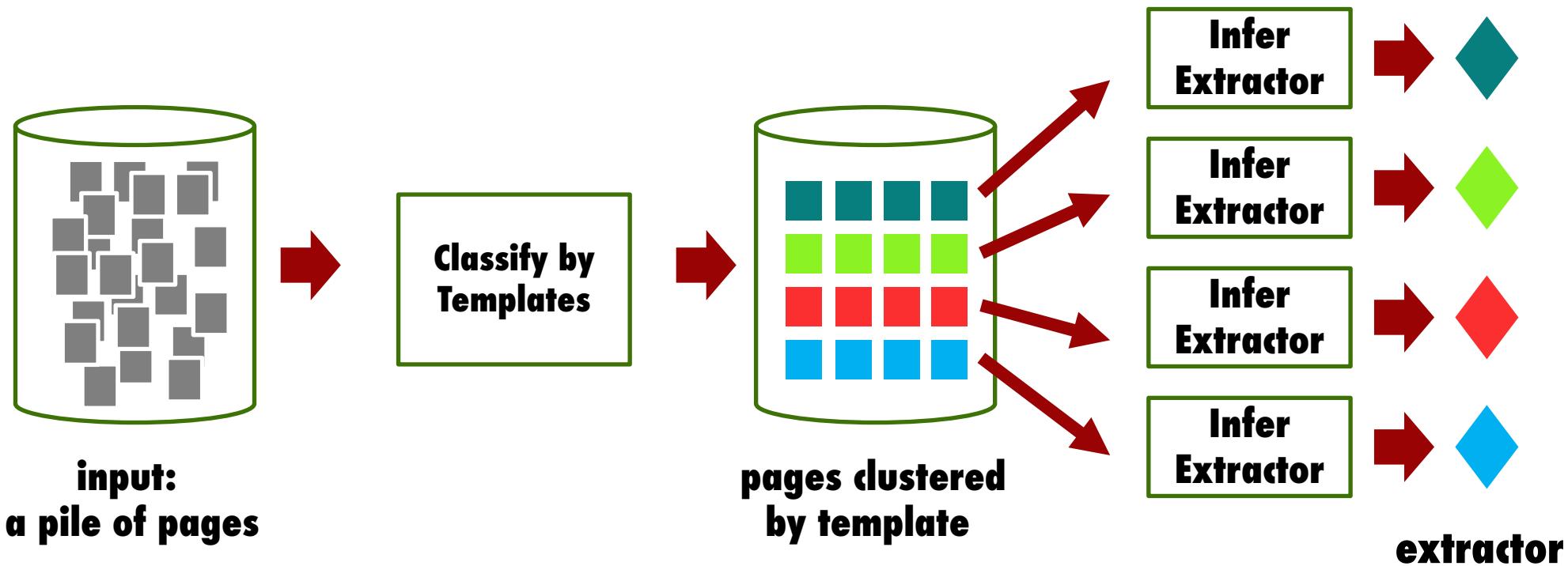
Run all extractors currently defined and load the extracted data into a new ElasticSearch index. You need to perform the Update To New Index command to replace your current index with the new index.

# 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	<b>1.0</b> (50/50)	<b>.76</b> (37/49)	<b>.95</b> (40/42)	<b>.83</b> (40/48)	<b>.87</b> (39/45)	<b>.51</b> (23/45)	<b>.68</b> (34/50)	<b>1.0</b> (35/35)	<b>.52</b> (15/29)	<b>.76</b> (19/25)	<b>.97</b> (35/36)	
Pretty Good	<b>1.0</b> (50/50)	<b>.98</b> (48/49)	<b>.95</b> (40/42)	<b>.83</b> (40/48)	<b>.98</b> (44/45)	<b>.84</b> (38/45)	<b>.88</b> (44/50)	<b>1.0</b> (35/35)	<b>.55</b> (16/29)	<b>1.0</b> (25/25)	<b>1.0</b> (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 compose 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)**

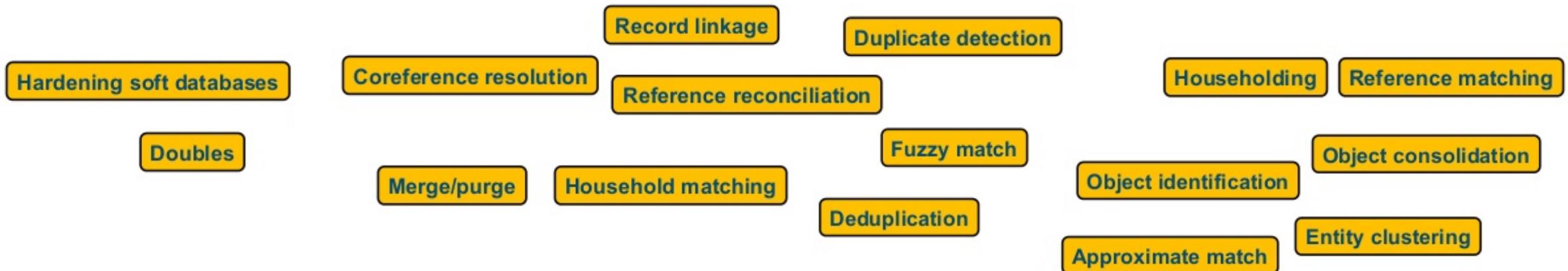
**Probabilistic Soft Logic (PSL)**

**Knowledge Graphs in Latent Space aka knowledge  
graph embeddings**

# **Entity Resolution (ER)**

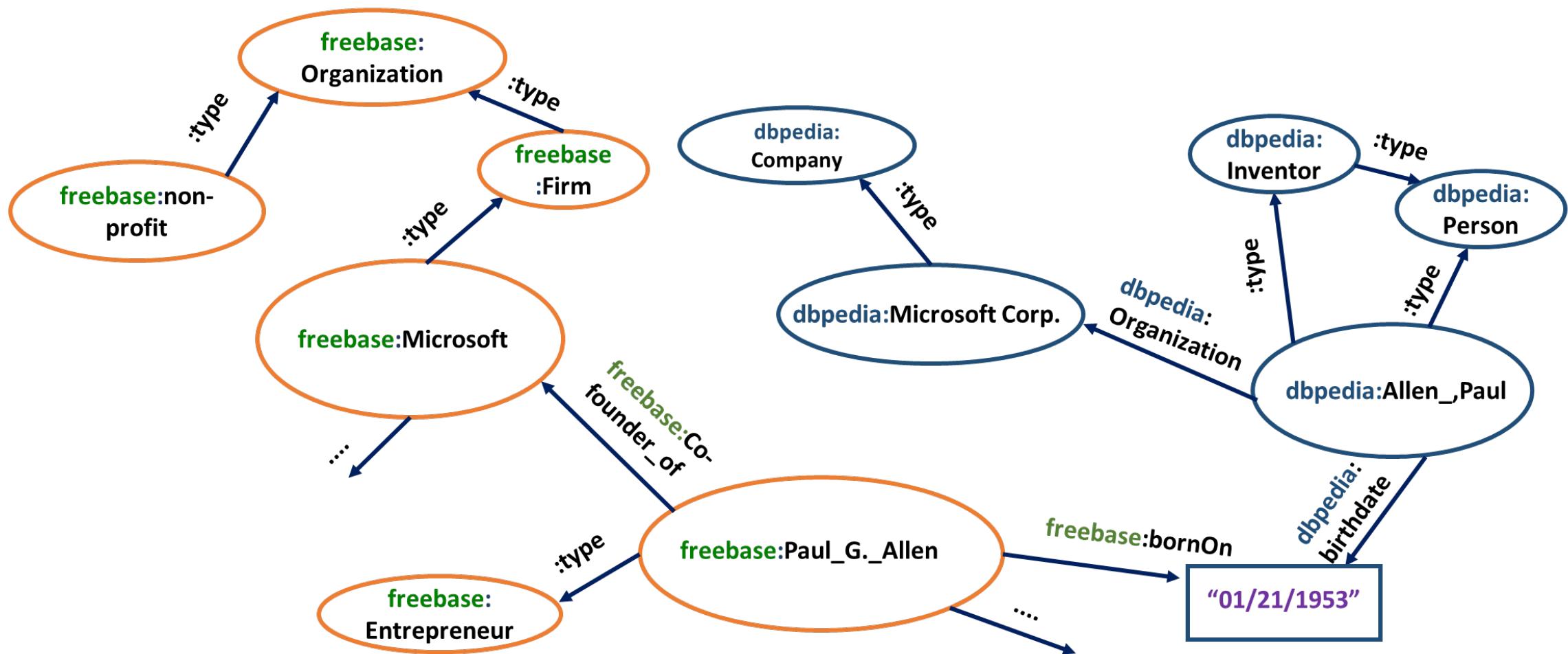
# Entity Resolution (ER)

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

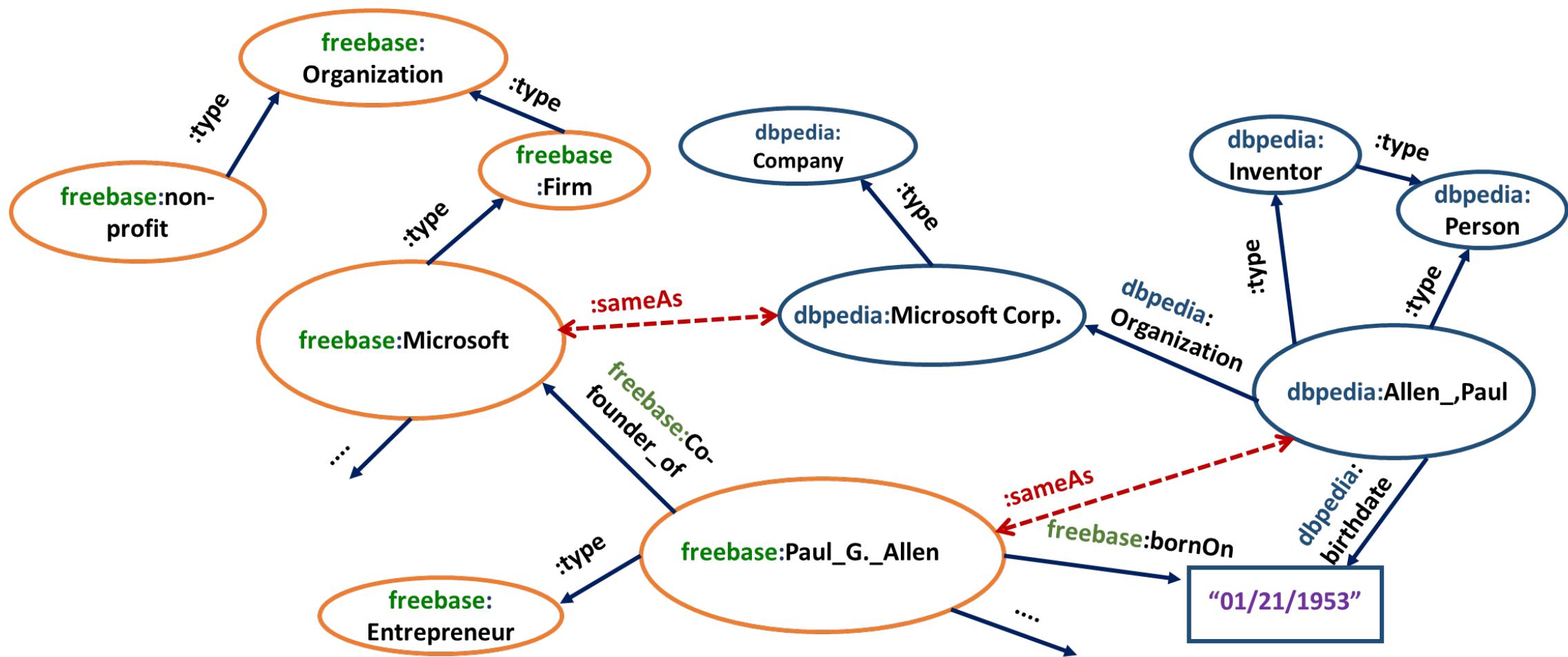


# Example: Linking Dbpedia to Freebase

3

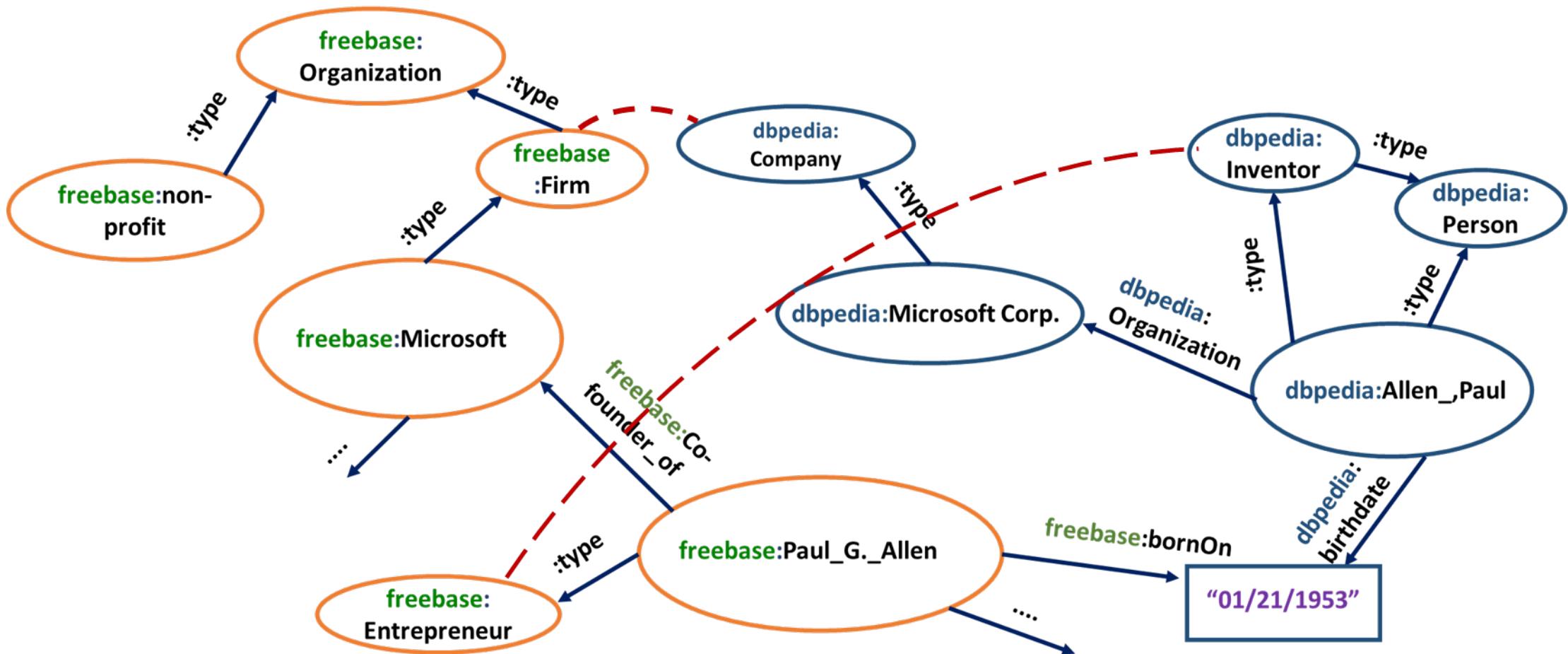


# Knowledge graphs contain duplicate entities



# Knowledge graph nodes are multi-type

(c)



# 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)**

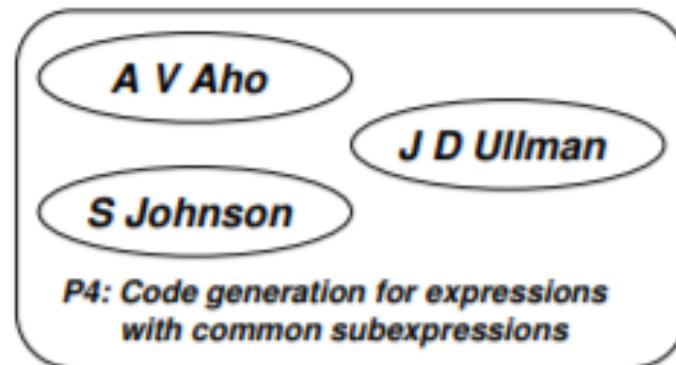
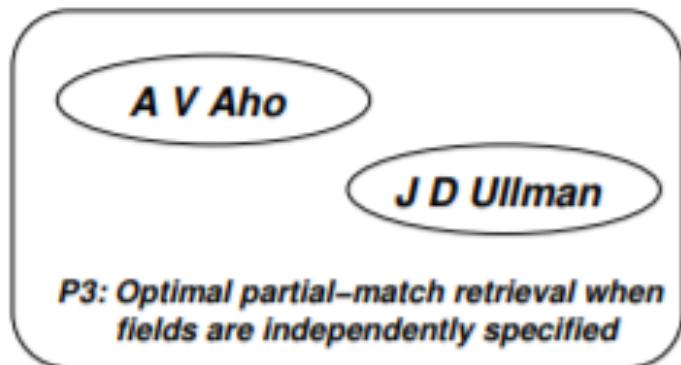
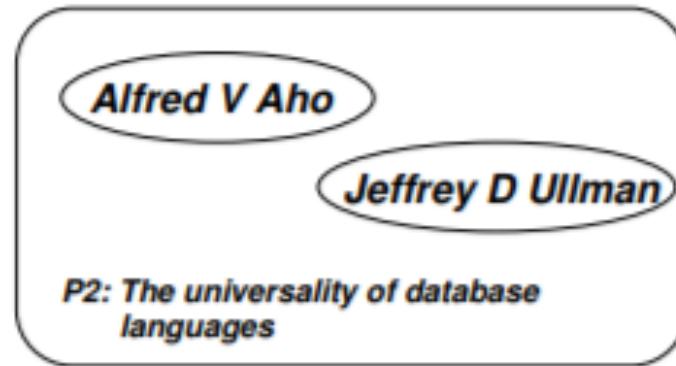
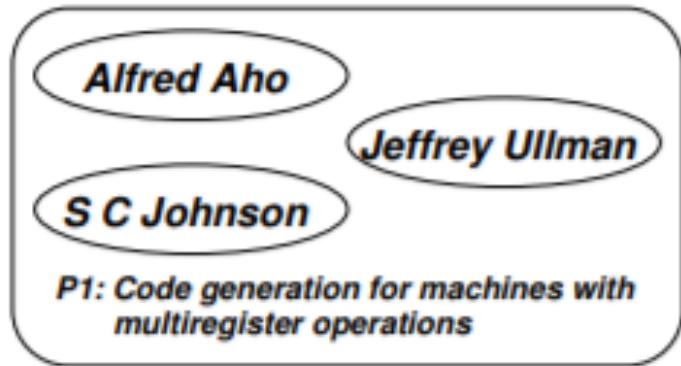
More expensive but better performance

**Can be generic or use domain knowledge e.g.,**

**citation/bibliography domain Bhattacharya and Getoor**

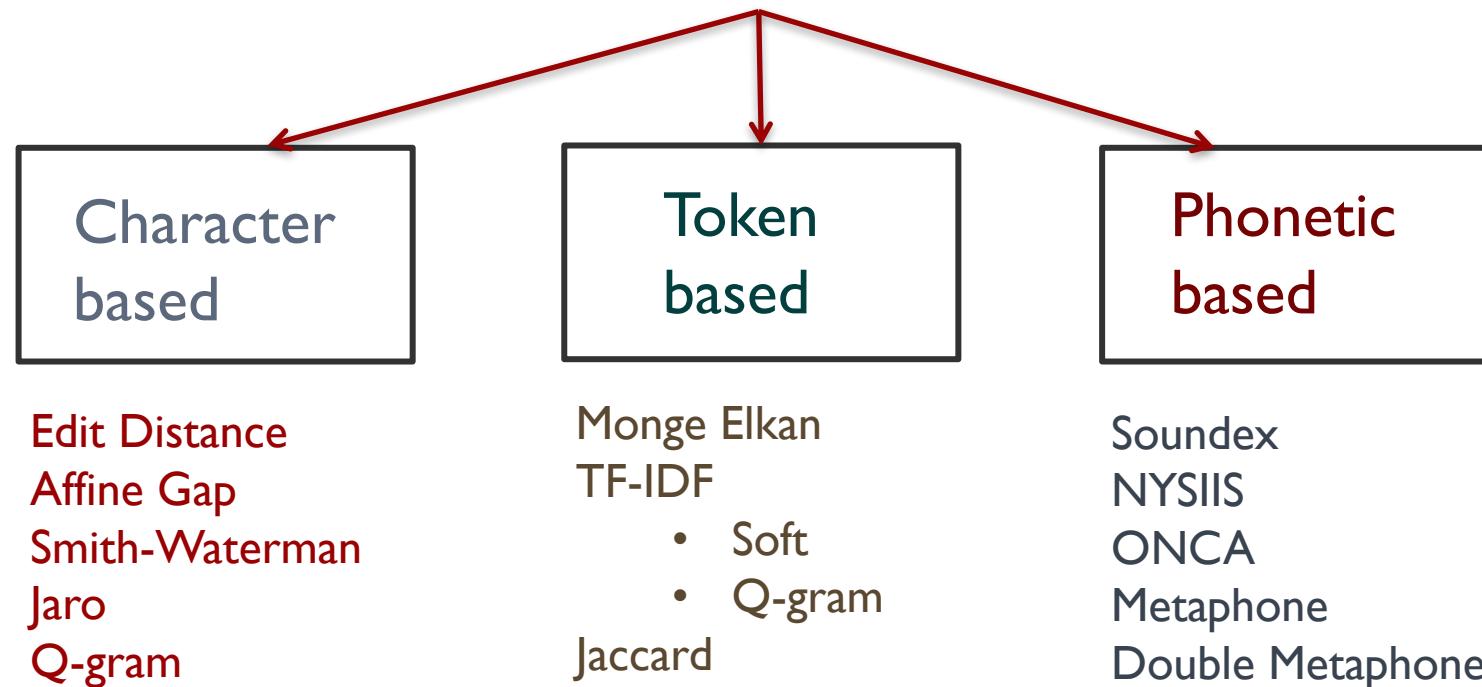
**(2006,2007)**

# Example (co-authorship)



# Feature functions - I

**First line of attack is *string matching***



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

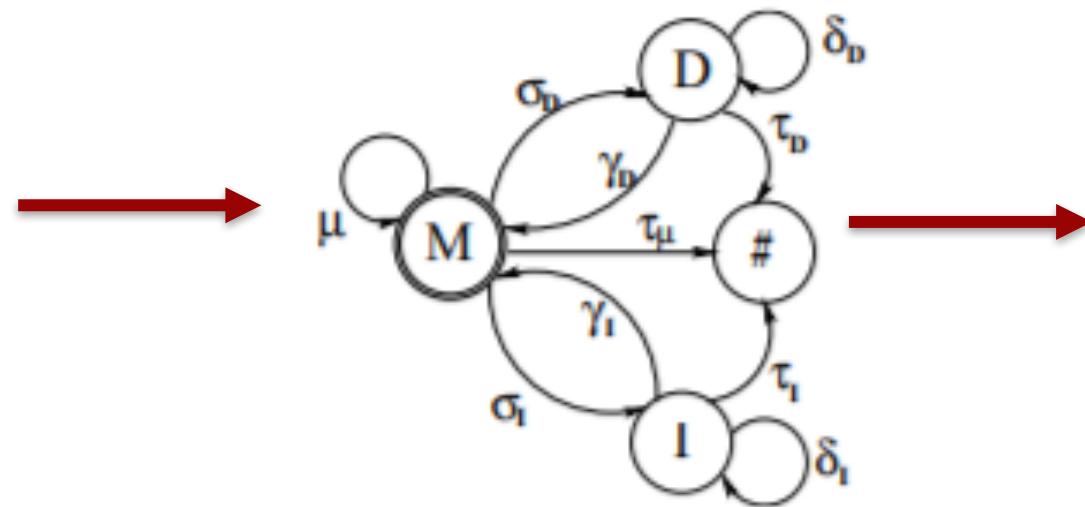
Review IDs for human trafficking

Stock tickers for SEC

# Learnable string similarity

## Example: adaptive edit distance

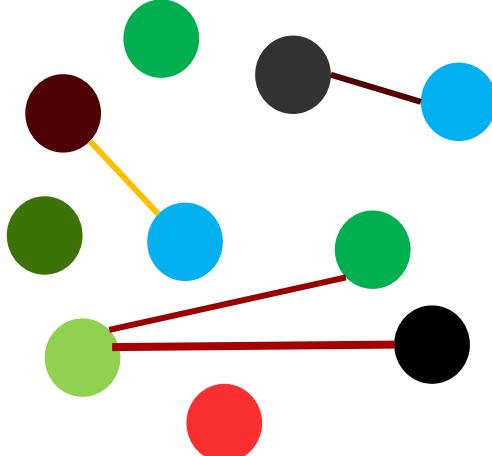
Sets of equivalent  
string pairs (e.g.,  
**<Suite 1001, Ste.  
1001>**



Bilenko and Mooney (2003)

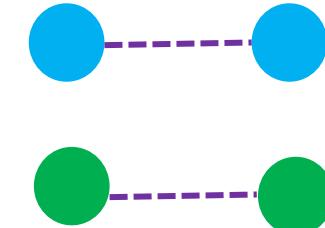
# After training...

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



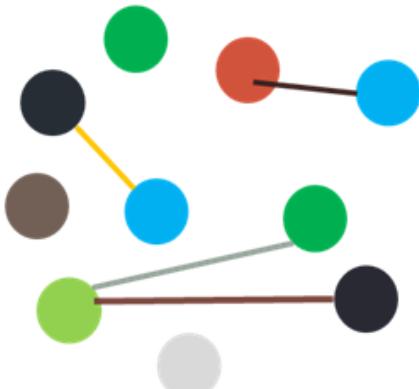
$O(|V|^2)$   
**applications of  
similarity  
function**

**Linked mentions**

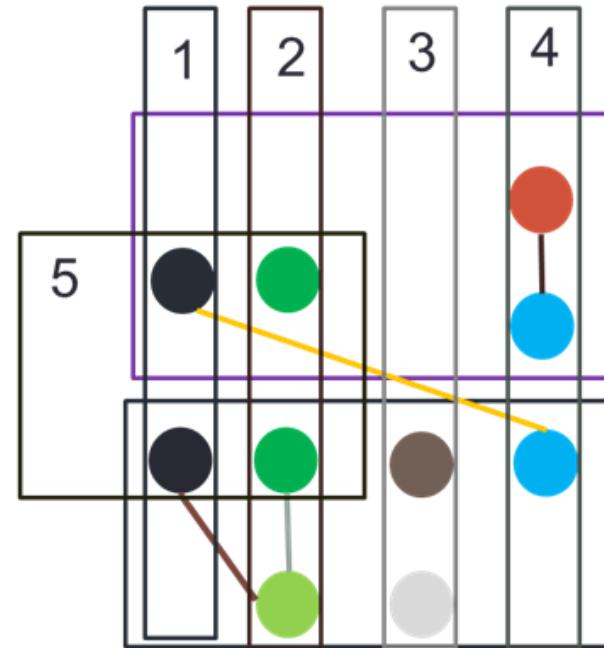


# Blocking trick

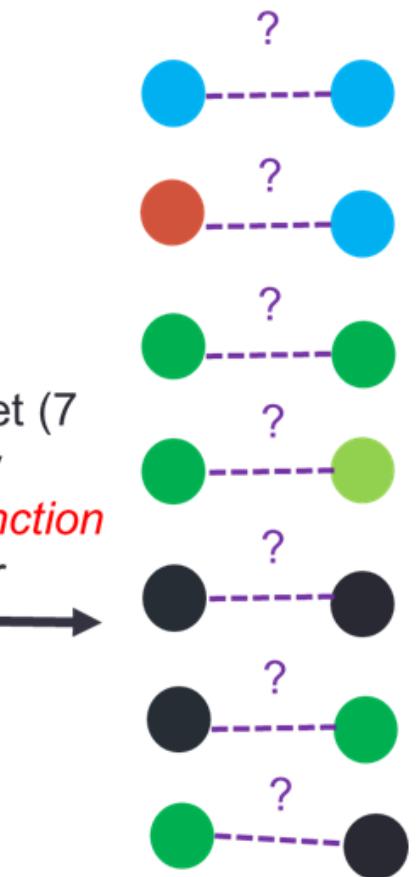
Like a **configurable inverted index function**



Apply *blocking key*  
e.g. *Tokens(LastName)*



Generate candidate set (7 pairs), apply *similarity function* on each pair



# **What is a good blocking key?**

**Achieves high recall**

**Achieves high reduction**

**Good survey on blocking: Christen (2012)**

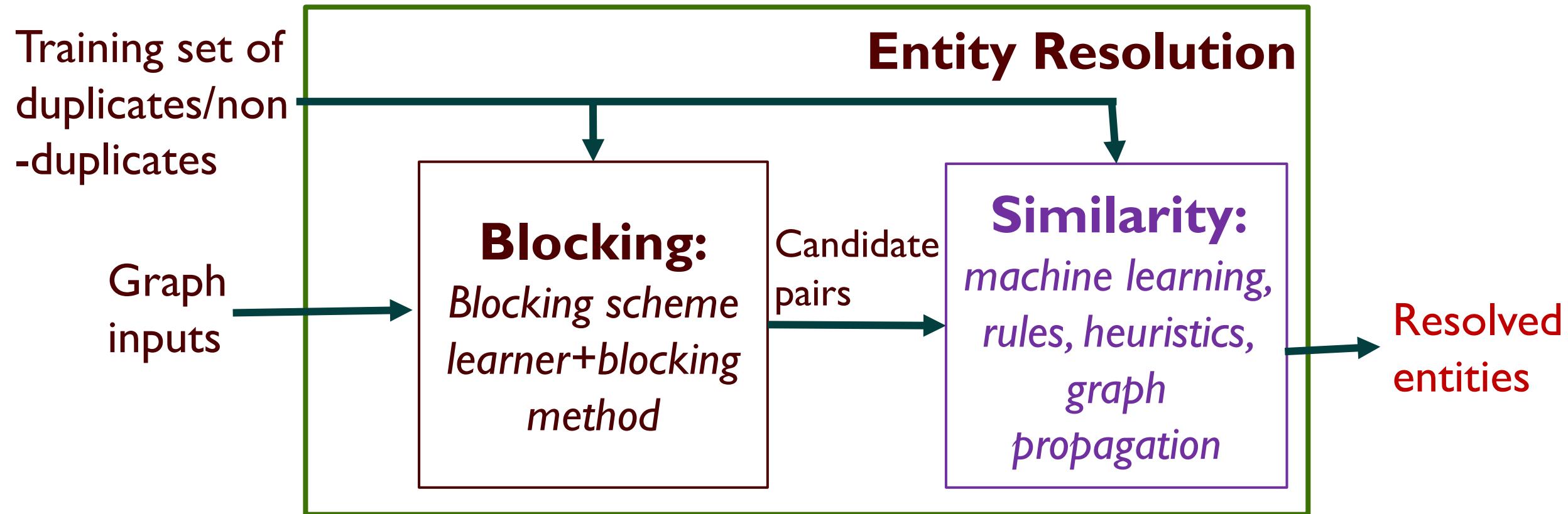
# How do we learn a good blocking key?

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

CharTriGrams(Last\_Name)  $\cup$  (Numbers(Address)  $\times$  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)**

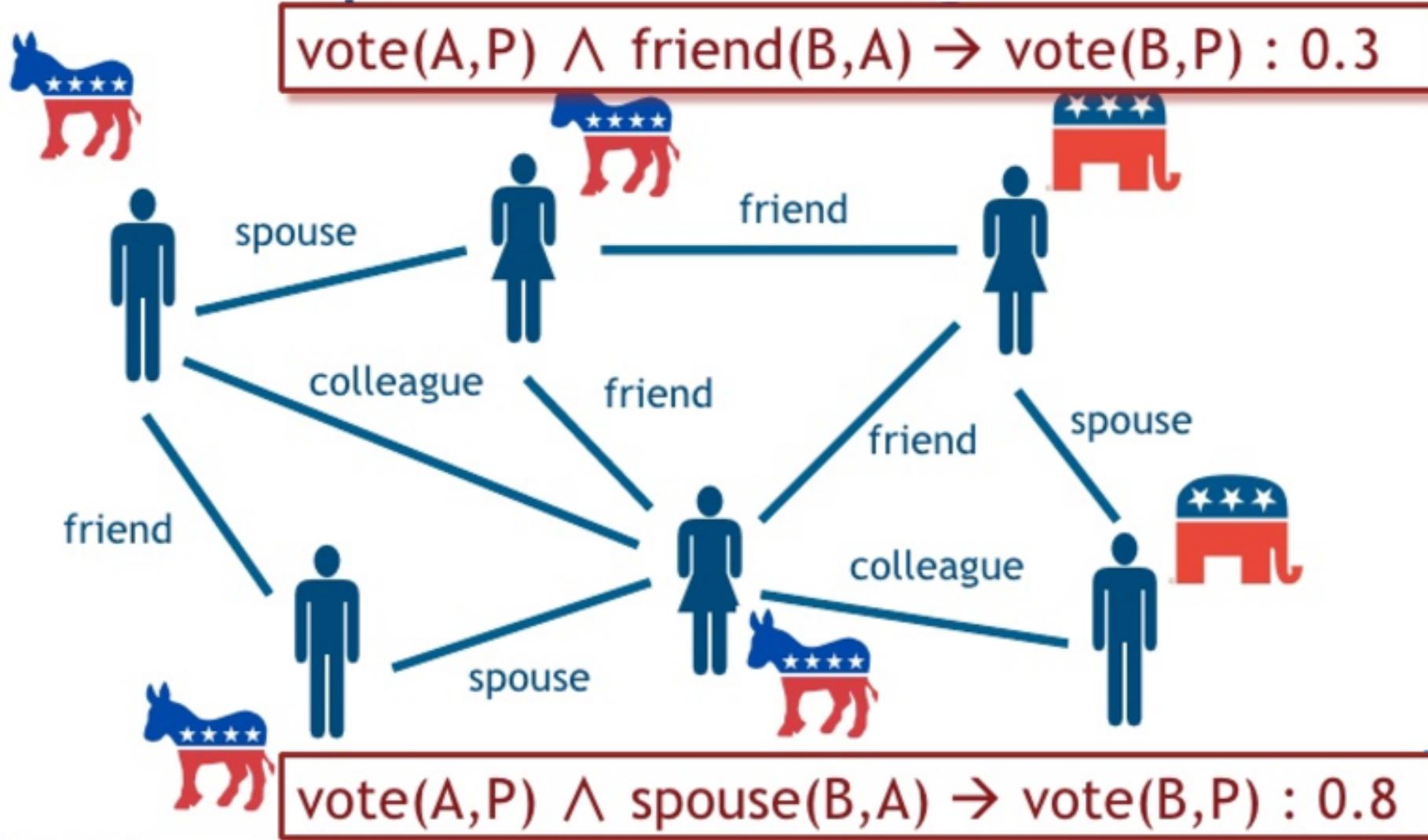
Probabilistic Relational Models (PRMs) **Koller (1999)**

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**

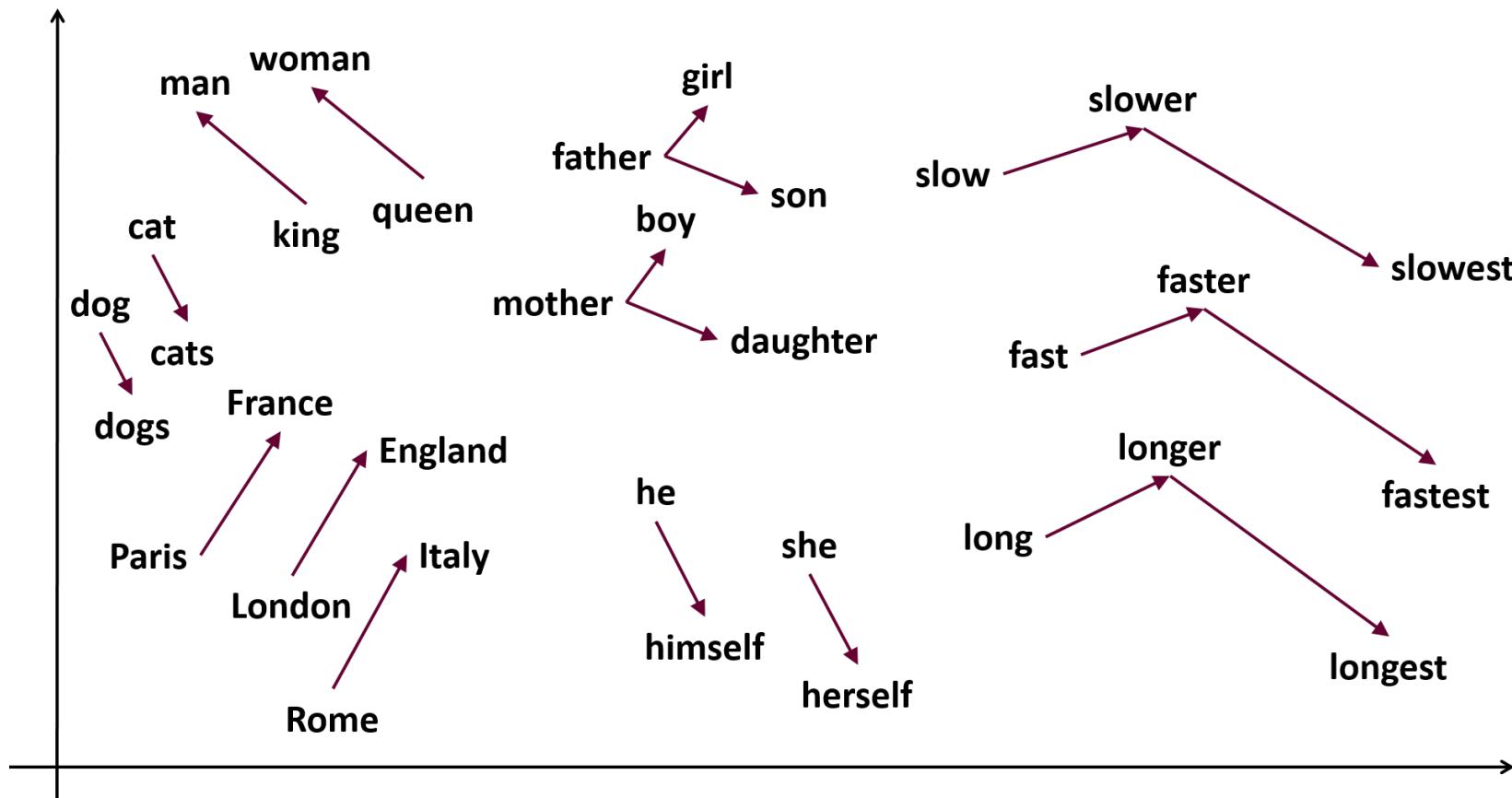
**Can easily encode rules (in PSL) such as a city can only be in one state and one country, a city must actually exist in a state and country...**

**Use elements like population to assign weights**

# **Knowledge Graphs in Latent Space**

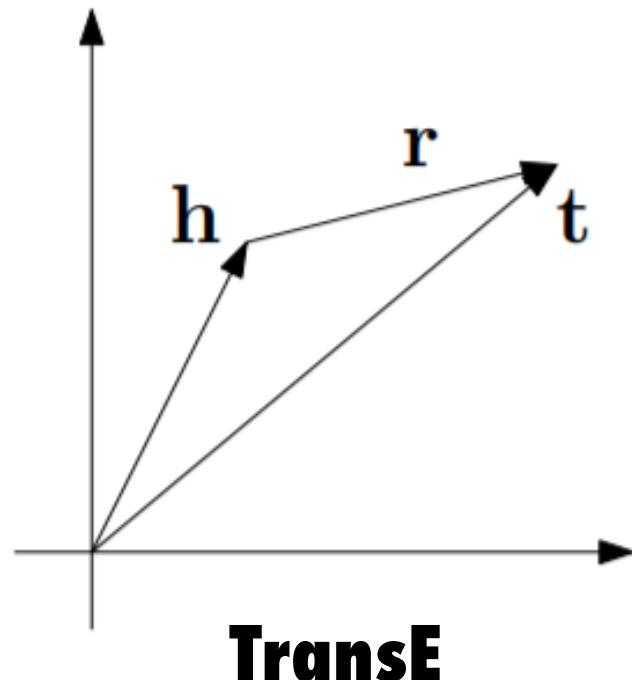
# Low-dimensional vector spaces

Very popular for documents, graphs, words...

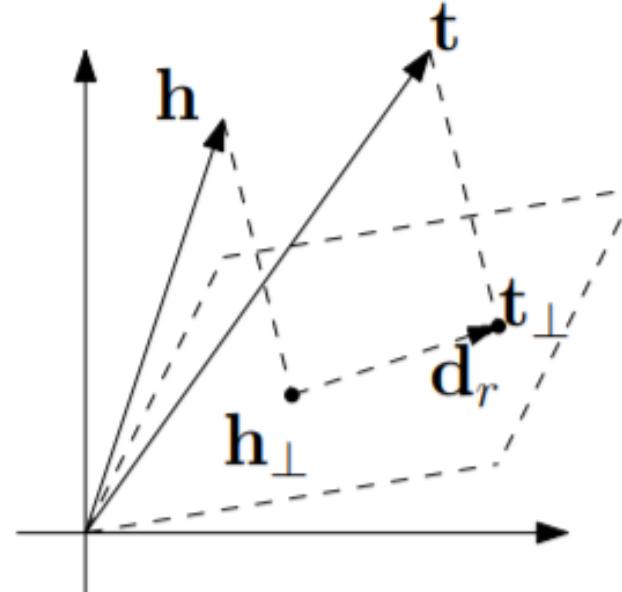


# Knowledge graph embeddings

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	$\mathbf{u}_r^\top f(W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r(s k + s))$
NTN (Socher et al. 2013)	$\mathbf{u}_r^\top f(\mathbf{h}^\top \mathbf{W}_r \mathbf{t} + W_{rh}\mathbf{h} + W_{rt}\mathbf{t} + \mathbf{b}_r)$ $\mathbf{u}_r, \mathbf{b}_r \in \mathbb{R}^s, \mathbf{W}_r \in \mathbb{R}^{k \times k \times s}, W_{rh}, W_{rt} \in \mathbb{R}^{s \times k}$	$O(n_e k + n_r(s k^2 + 2s k + 2s))$
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			
	MEAN		HITS@10		MEAN		HITS@10	
Metric	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	<b>251</b>	75.4	<b>89.2</b>	243	125	34.9	47.1
TransH (unif.)	318	<b>303</b>	75.4	<b>86.7</b>	211	<b>84</b>	42.5	<b>58.5</b>
TransH (bern.)	400.8	388	73.0	82.3	212	<b>87</b>	45.7	<b>64.4</b>

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	<b>85.3</b>	-
Bilinear Model	73.8	84.3	-
NTN	70.4	<b>87.1</b>	66.5 ( $\approx 40h$ )
TransE (unif.)	75.85	70.9	79.7 ( $\approx 5m$ )
TransE (bern.)	75.87	81.5	<b>87.3</b> ( $\approx 5m$ )
TransH (unif.)	<b>77.68</b>	76.5	80.2 ( $\approx 30m$ )
TransH (bern.)	<b>78.80</b>	83.3	<b>87.7</b> ( $\approx 30m$ )

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. Converting ‘locations’ into feature vectors**

**Relevant for toponym resolution, building rich graphs...**

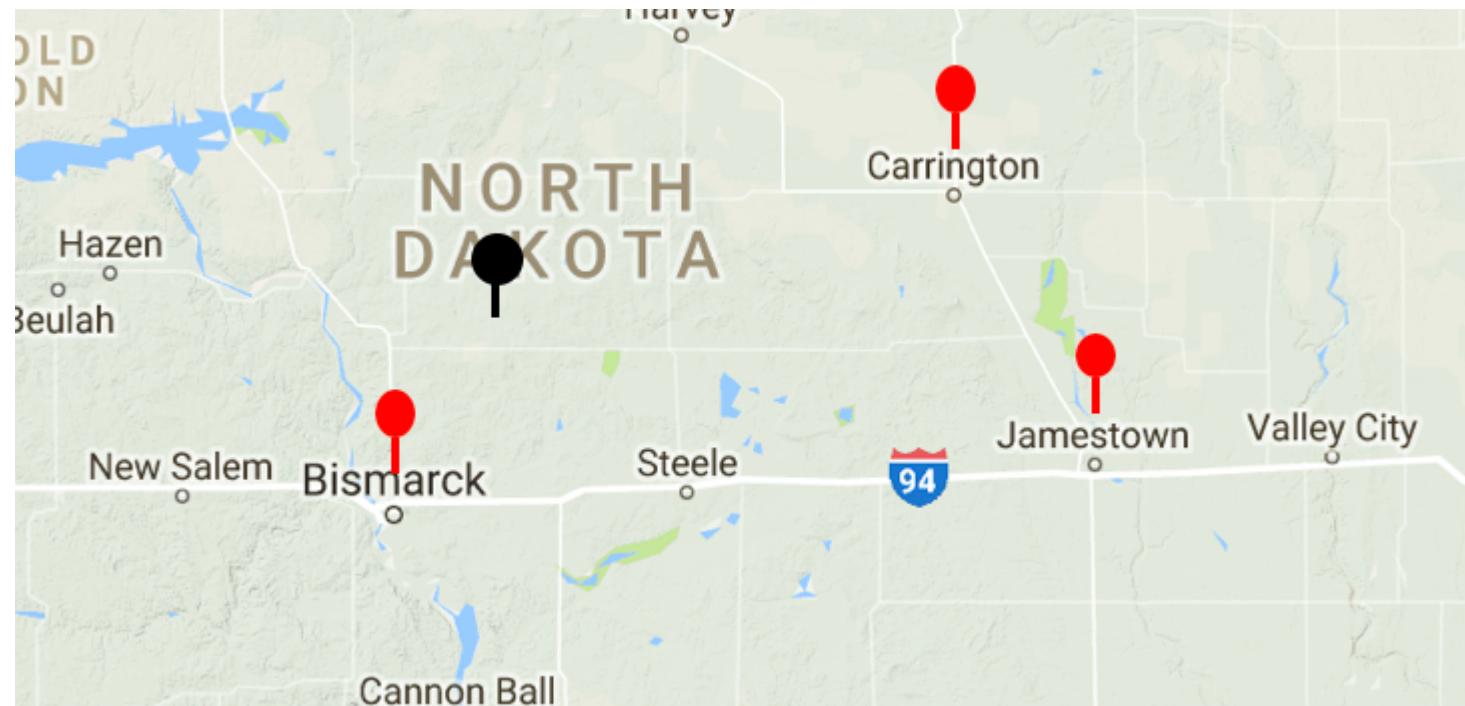
Kejriwal, Mayank; Szekely, Pedro (2017): Neural Embeddings for Populated GeoNames Locations. figshare.

<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

Enter Keywords (Advanced Search On)  
brianna elite 🔍 × ⚙️ ≡

25 of 851,393 Results ? [How are search results found?](#) ⬇️ Export

---

 **Brianna Sky**   Elite Companion   Long Legged, Green Eyed, Blonde Bombshell   Specials  - Maryland  
escorts - backpage.com  

 **Brianna Sky**   Elite Companion   Long Legged, Green Eyed, Blonde Bombshell  - Baltimore escorts -  
backpage.com  

## Search Terms

- City: chicago
- Services Provided: fetish friendly
- Services Provided: unrushed
- Ethnicity of Provider: hispanic

[What do facets represent?](#)

## Telephone Number (Top 10)

<input type="checkbox"/>	41	18
<input type="checkbox"/>	41	9
<input type="checkbox"/>	41	9
<input type="checkbox"/>	20	8
<input type="checkbox"/>	61	5
<input type="checkbox"/>	50	4
<input type="checkbox"/>	85	3
<input type="checkbox"/>	20	2
<input type="checkbox"/>	21	1
<input type="checkbox"/>	26	1

## Email Address (7)

<input type="checkbox"/>	al	o.com	1
<input type="checkbox"/>	dc	.com	1
<input type="checkbox"/>	ha	.com	1
<input type="checkbox"/>	ih	.mail.com	1
<input type="checkbox"/>	m	400.my	1
<input type="checkbox"/>	m	@yahoo.com	1
<input type="checkbox"/>	se	.il.com	1

## Social Media ID

[View More](#)

None

## Review ID

[View More](#)

None

## City (Top 10)

<input type="checkbox"/>	Kejriwal, Szekey	Sort By:
--------------------------	------------------	----------

25 of 368,663 Results [How are search results found?](#)**3.51** \$\$\$\$ :) (Brand new hot **hispanic fetish friendly** on board with great special :) :) - **chicago**

escorts - backpage.com

Website

backpage.com

Post Date

Dec 9, 2013

Locations

**chicago, illinois**

Telephone Numbers

87 88

**3.13** 36 H NaTuRaL BrEaStS \*SiNgLe MoM\* NEED HELP ASAP! (**FeTiSh FrIeNdLy & 420 friendly**)- **chicago** escorts - backpage.com

Website

backpage.com

Post Date

Jan 27, 2014

Locations

**portage, indiana**

Telephone Numbers

27 10

Url <http://chicago.backpage.com/FemaleEscorts/36-h-natural-breasts-single-mom-need-help-asapfetish-friendly-and-420-friendly-26/19126439>

## Description

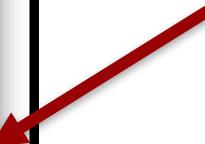
Posted: Monday, January 27, 2014 Im Hispanic, 26, 5'4, 140, size 36 h natural breasts, n I have hidden talents. im fetish in 420 friendly (n trained domme) if ur lookin for fun or really the truth of my situation is im a single mother behind in bills. I could really use the help ASAP! call or txt 219 708 1000, ask for Hobie (located in portage/hobart area) Poster's age: 26 • Location: Northwest Indiana, incall only hobart/portage area • Post ID: 19

Cached Ad Webpage [Open](#)

DATES	Click to set start date	Click to set end date	
IDENTIFIERS	Telephone Number	Email Address	Social Media ID
LOCATION	Review ID	City	State/Region
PROVIDER	Name of Provider	Age of Provider	Ethnicity of Provider
	Eye Color of Provider	Hair Color of Provider	Height of Provider
	Weight of Provider	Services Provided	Price of Provider
AD	Website	Title	Page

[SHOW 25 MORE RESULTS](#)

# 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**

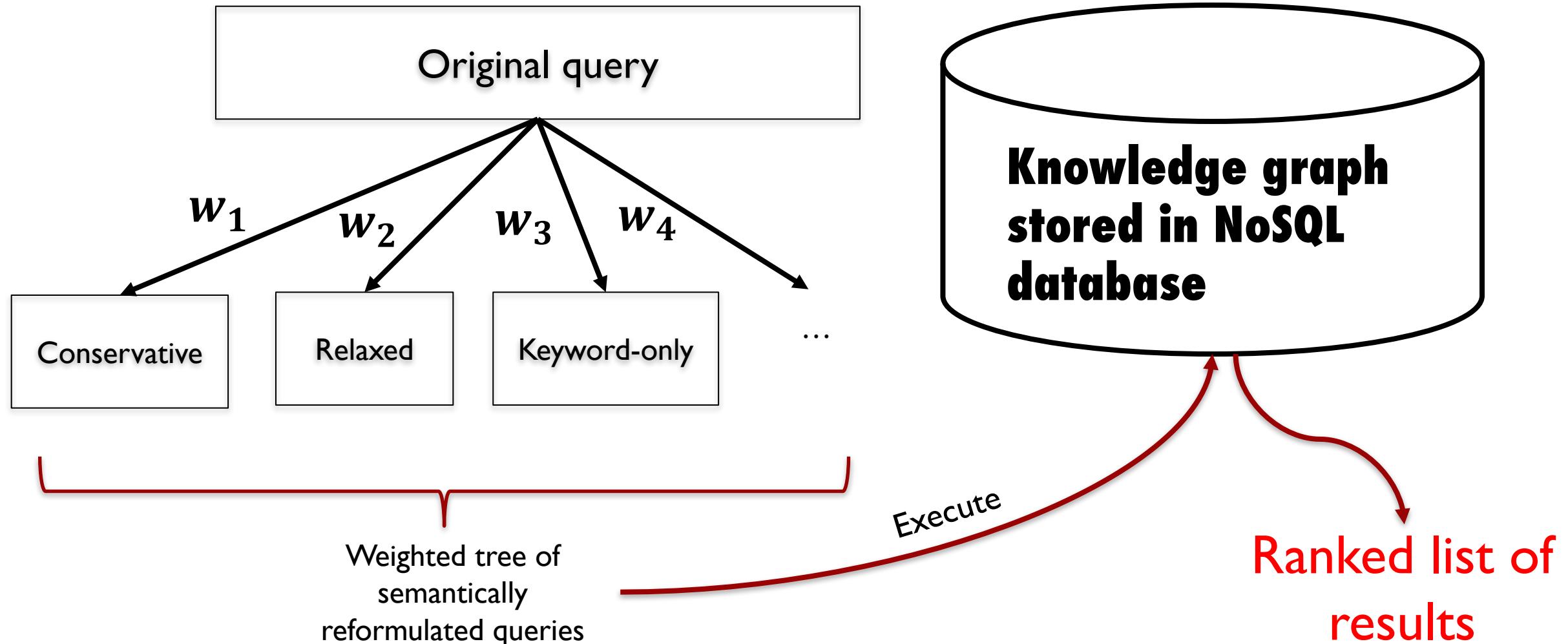
# **Key technique: Query reformulation**

**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

Michelangelo

Date of birth: April 15, 1452  
Date of death: May 2, 1519  
(age 67 years)

Italy