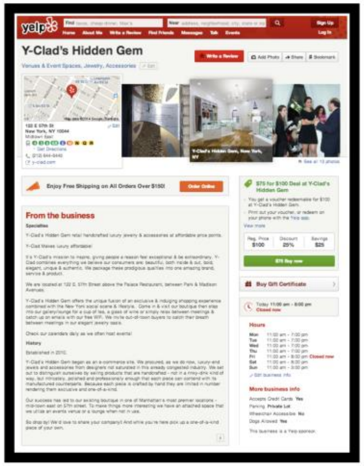
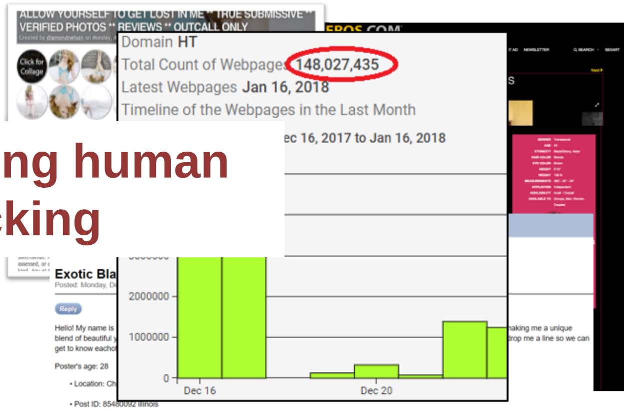
Conclusion and review

Domain-specific search (DSS)



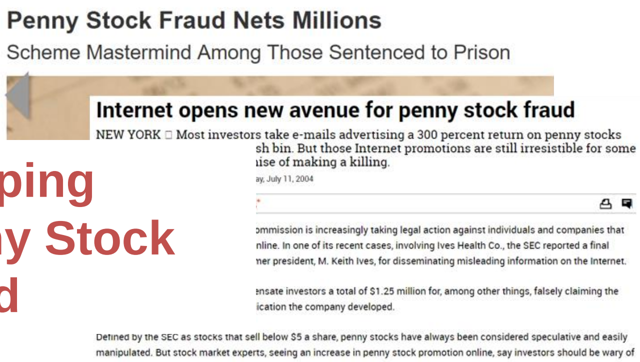
source: https://photos.prnewswire.com/prnfull/20151006/274273-INFO

Emerging opportunities for DSS

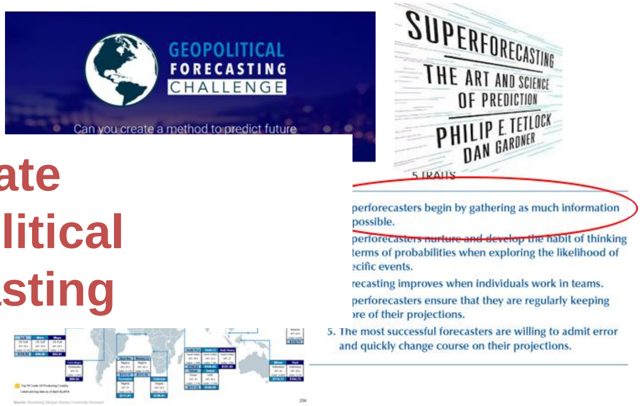


Fighting human trafficking

Predicting cyberattacks



Stopping Penny Stock Fraud

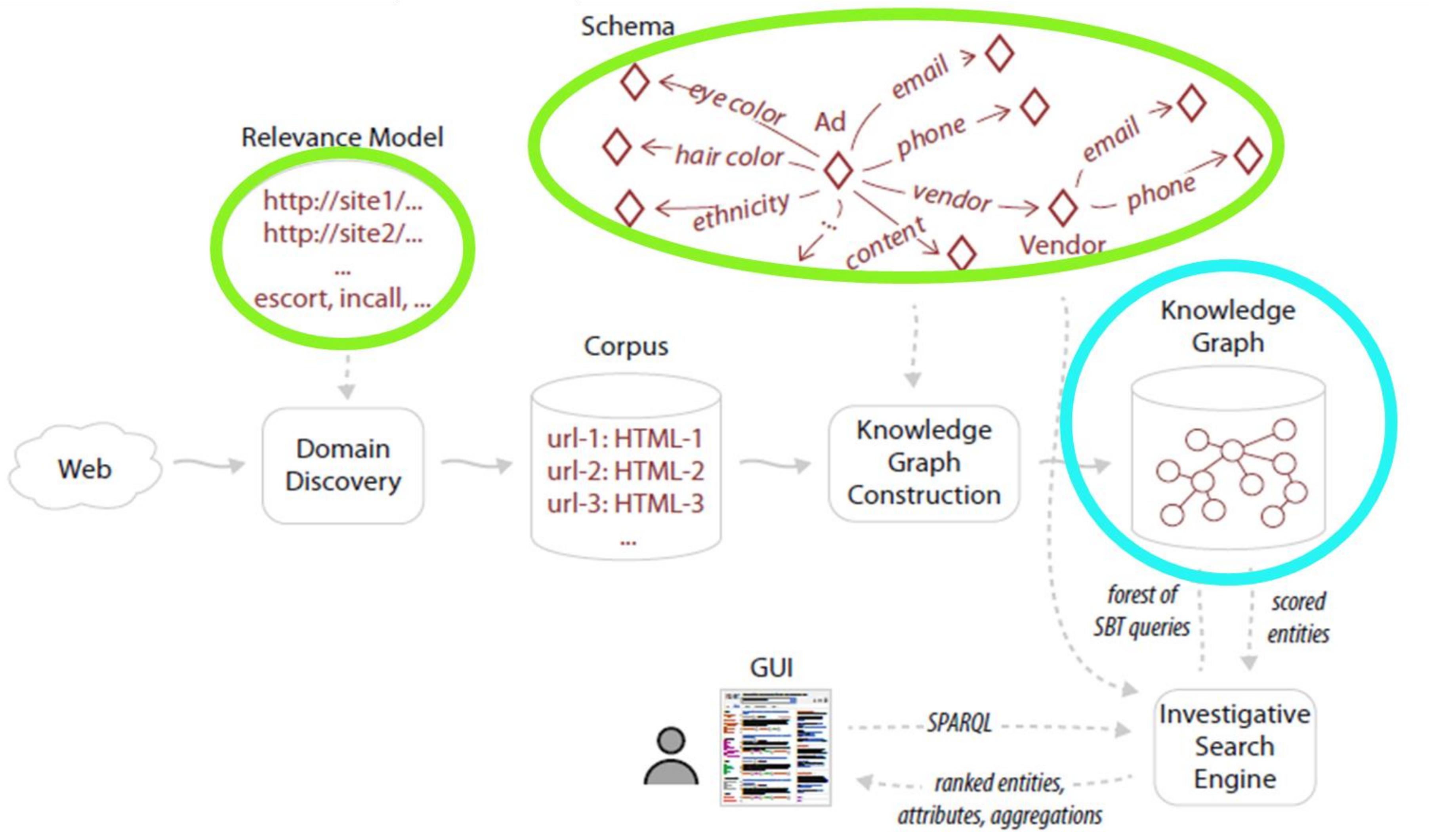


Accurate geopolitical forecasting

General Search	Google Knowledge Graph
DSS	Domain-Specific Knowledge Graphs

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

Knowledge Graphs for DSS



Challenges

Many 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

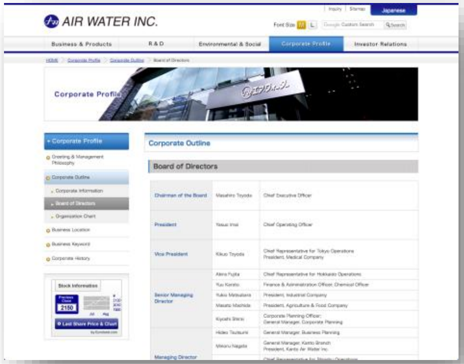
Contact

- 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.			

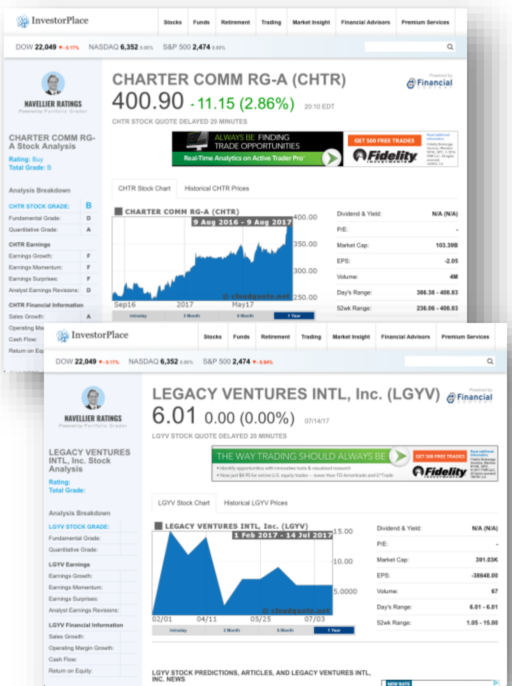
Tables



Charts

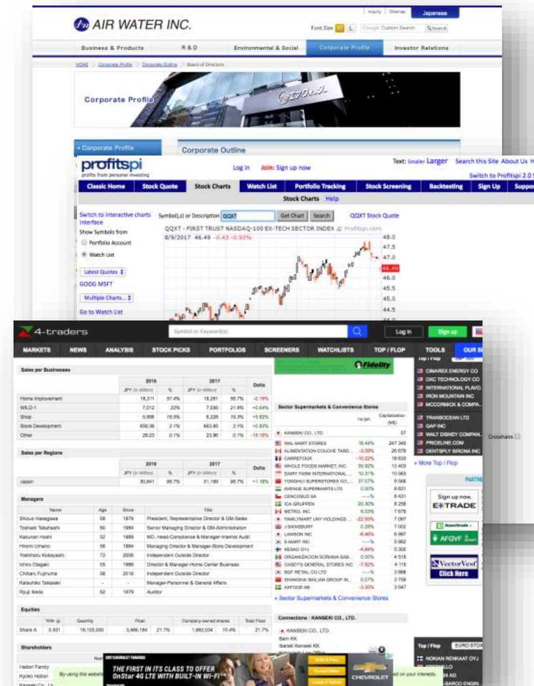
Scope

Short tail



Genre specific (e.g., forums)

Long tail



Pattern Complexity

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

U.S. postal addresses

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

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

Ambiguous, needing context

Person names

…was among the six houses sold by Hope Feldman that

Pawel Opalinski, Software Engineer at WhizBang Labs.

Unusual language models

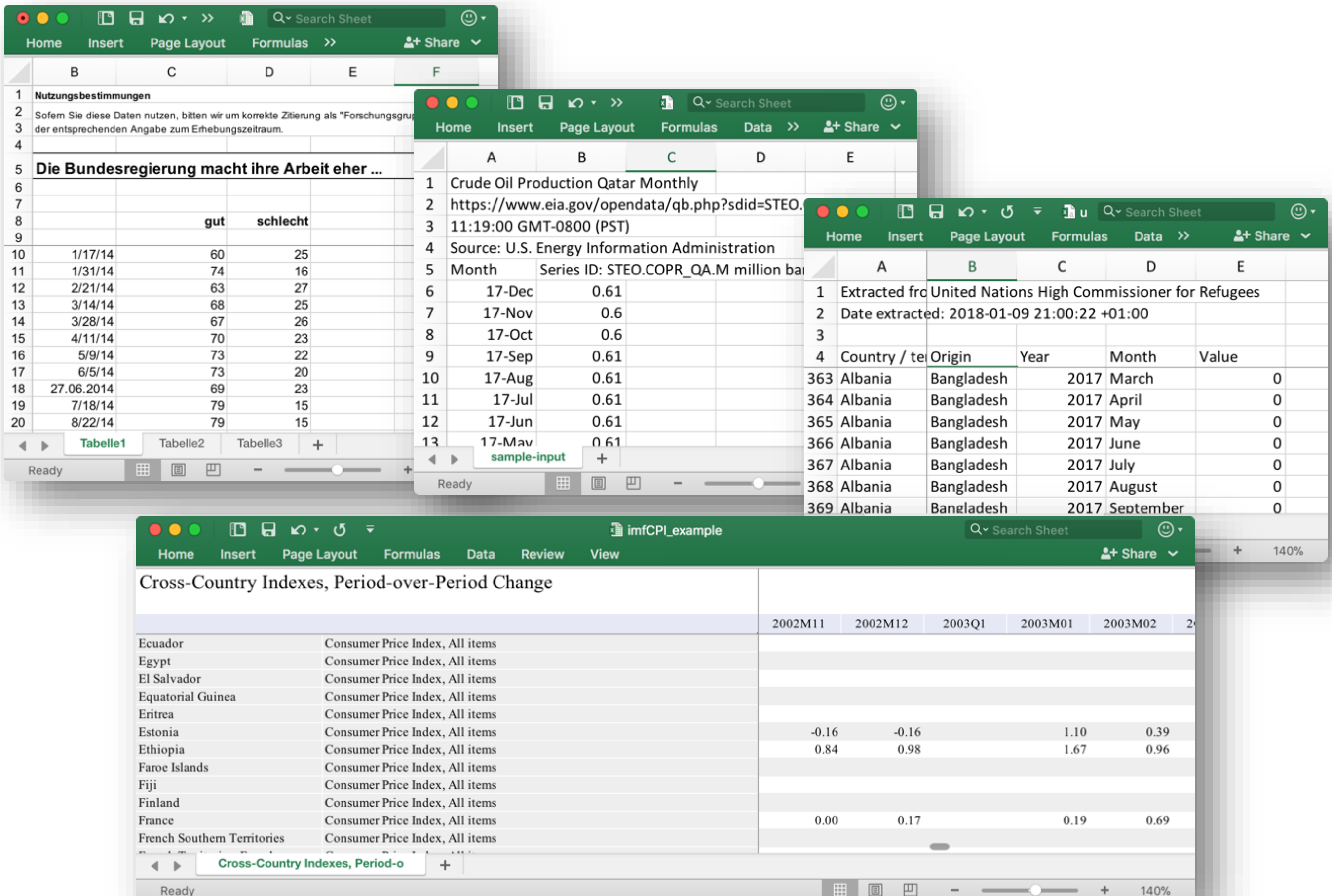
"YOU don't wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7○7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses"

Courtesy of Andrew McCallum

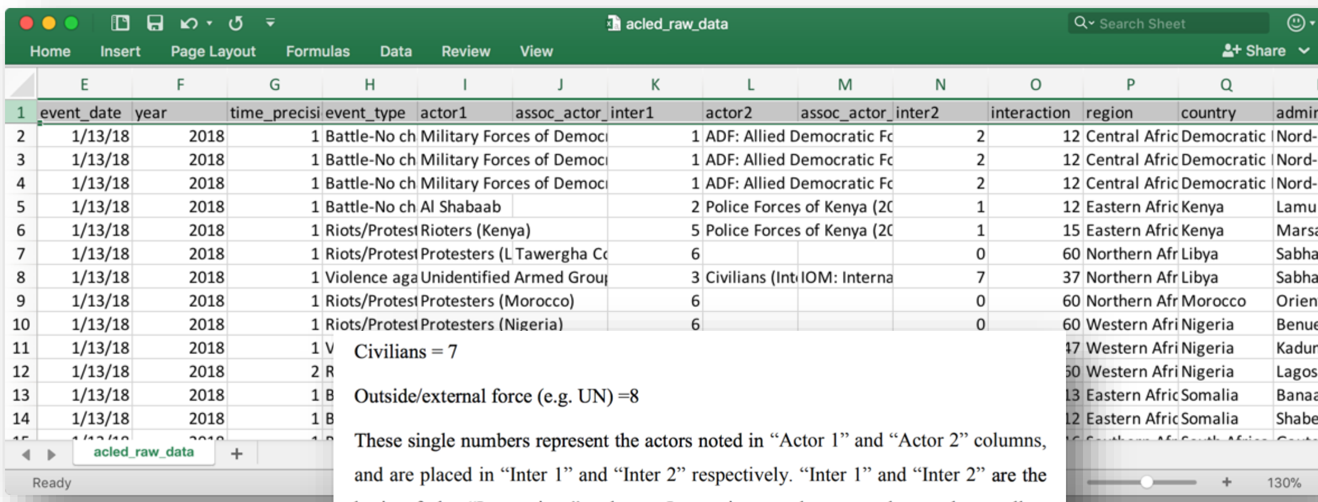
small amount of relevant content

irrelevant content very similar to relevant content

Spreadsheets Created For Human Consumption



Databases with PDF Code Books



Civilians = 7

Outside/external force (e.g. UN) =8

These single numbers represent the actors noted in "Actor 1" and "Actor 2" columns, and are placed in "Inter 1" and "Inter 2" respectively. "Inter 1" and "Inter 2" are the basis of the "Interactions" column. Interaction numbers are always the smallest possible number (for example, 37 instead of 73), regardless of the order of "Actor 1" and "Actor 2". For single actor events, the empty second actor category is coded as "0".

Interaction codes include:

10- SOLE MILITARY ACTION

11- MILITARY VERSUS MILITARY

12- MILITARY VERSUS REBELS

13- MILITARY VERSUS POLITICAL MILITIA

PDF

Data In Web Tables

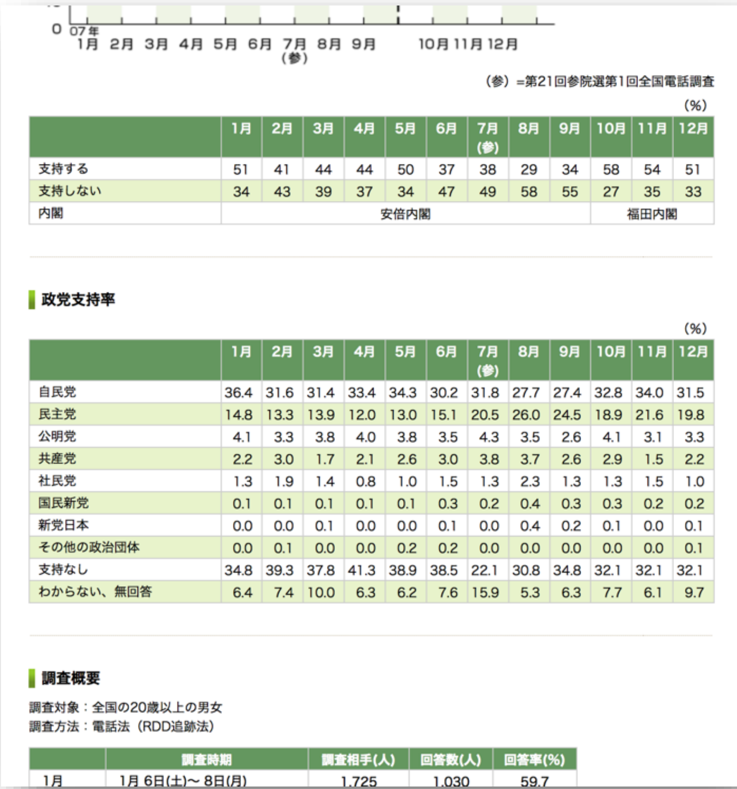
UNITED NATIONS
SECURITY COUNCIL

ABOUT | PRESIDENCY | MEMBERS | PROGRAMME OF WORK | DOCUMENTS | MEETINGS | SUBSIDIARY ORGANS | 2231 (2015) | REPERTOIRE

- About SC Documents
- Resolutions
- Presidential Statements
- Press Statements
- Notes by the President
- Exchange of Letters
- Reports submitted by / transmitted by the Secretary-General
- Reports of the Security Council Missions
- Annual Reports
- Volumes of Resolutions
- Round-ups
- Highlights of Security Council Practice
  - Annual Highlights

Security Council Resolutions

Resolutions adopted by the Security Council in 2017		
S/RES/2397 (2017)	22 December 2017	Non-proliferation/Democratic People's Republic of Korea
S/RES/2396 (2017)	21 December 2017	Threats to international peace and security caused by terrorist acts
S/RES/2395 (2017)	21 December 2017	Threats to international peace and security caused by terrorist acts
S/RES/2394 (2017)	21 December 2017	The situation in the Middle East
S/RES/2393 (2017)	19 December	The situation in the Middle East



0 07年 1月 2月 3月 4月 5月 6月 7月(参) 8月 9月 10月 11月 12月

（参）=第21回参院選第1回全国電話調査

（%）

	1月	2月	3月	4月	5月	6月	7月(参)	8月	9月	10月	11月	12月
支持する	51	41	44	44	50	37	38	29	34	58	54	51
支持しない	34	43	39	37	34	47	49	58	55	27	35	33
内閣	安倍内閣									福田内閣		

政党支持率

（%）

	1月	2月	3月	4月	5月	6月	7月(参)	8月	9月	10月	11月	12月
自民党	36.4	31.6	31.4	33.4	34.3	30.2	31.8	27.7	27.4	32.8	34.0	31.5
民主党	14.8	13.3	13.9	12.0	13.0	15.1	20.5	26.0	24.5	18.9	21.6	19.8
公明党	4.1	3.3	3.8	4.0	3.8	3.5	4.3	3.5	2.6	4.1	3.1	3.3
共産党	2.2	3.0	1.7	2.1	2.6	3.0	3.8	3.7	2.6	2.9	1.5	2.2
社民党	1.3	1.9	1.4	0.8	1.0	1.5	1.3	2.3	1.3	1.3	1.5	1.0
国民新党	0.1	0.1	0.1	0.1	0.1	0.3	0.2	0.4	0.3	0.3	0.2	0.2
新党日本	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.4	0.2	0.1	0.0	0.1
その他の政治団体	0.0	0.1	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.1
支持なし	34.8	39.3	37.8	41.3	38.9	38.5	22.1	30.8	34.8	32.1	32.1	32.1
わからない、無回答	6.4	7.4	10.0	6.3	6.2	7.6	15.9	5.3	6.3	7.7	6.1	9.7

調査概要

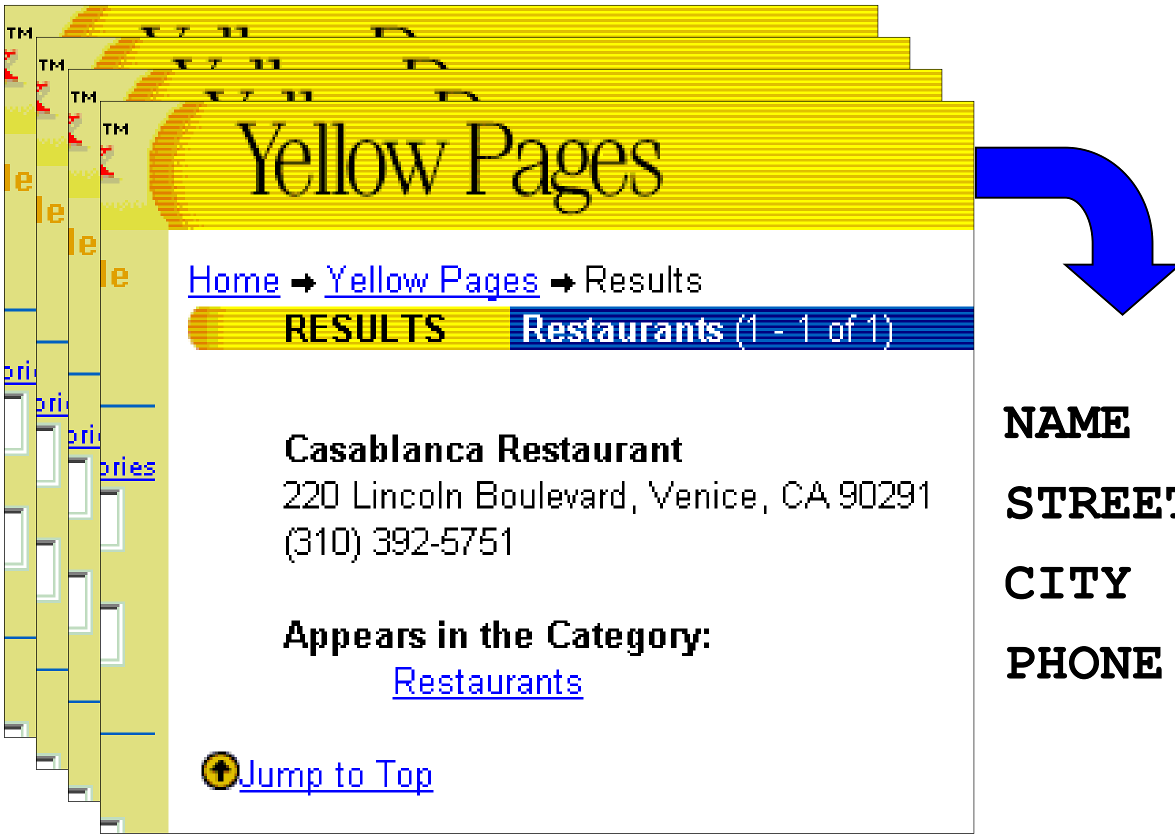
調査対象：全国の20歳以上の男女
調査方法：電話法（RDD追跡法）

	調査時期	調査相手(人)	回答数(人)	回答率(%)
1月	1月 6日(土)~ 8日(月)	1,725	1,030	59.7

Knowledge Graph Construction: Long-tail vs. Short-tail

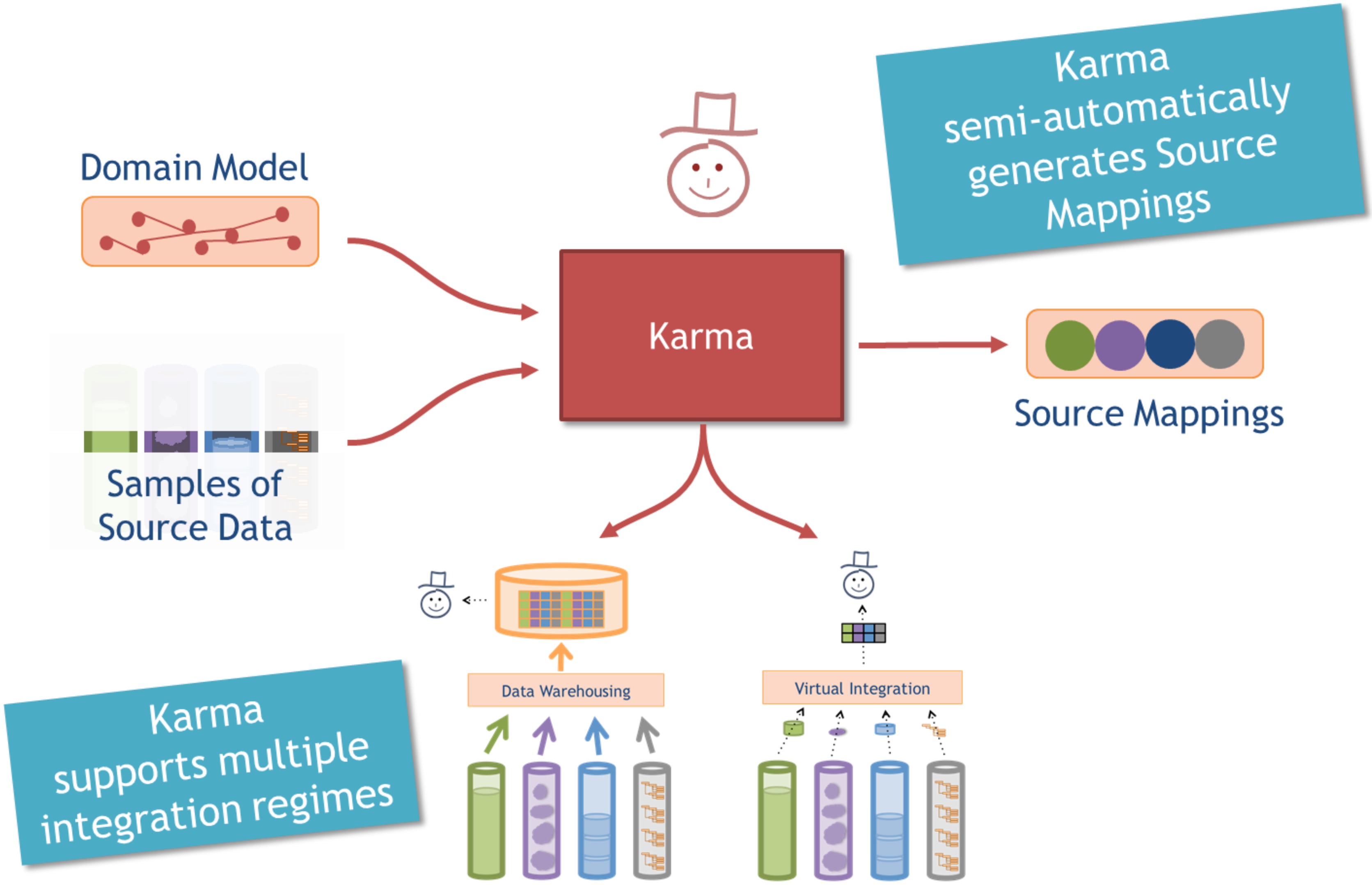
Extracting Data from Semi-structured Sources

Wrappers



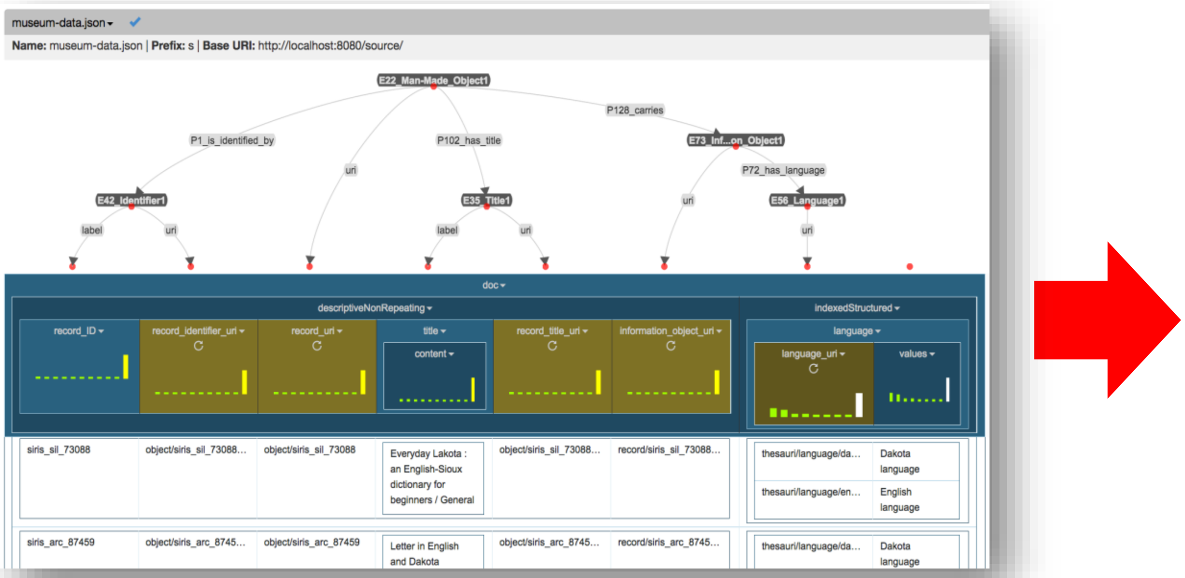
NAME	Casablanca Restaurant
STREET	220 Lincoln Boulevard
CITY	Venice
PHONE	(310) 392-5751

Information Integration in Karma



Karma semi-automatically builds semantic models

… and provides a nice GUI to edit them



Knowledge Graphs

Karma uses semantic models to create knowledge graphs

Practical Considerations for Extractions

Long tail vs. short tail

- How good (precision/recall) is necessary?
  - High precision when showing KG nodes 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), scripting, machine learning guru
- What tools can I use?
  - Many …

Data Tables

Arsène Wenger

Wenger in July 2015

Personal information	
Full name	Arsène Wenger[1]
Date of birth	22 October 1949 (age 67)
Place of birth	Strasbourg, Alsace, France
Height	6 ft 3 in (1.91 m)[2]
Playing position	Midfielder
Club information	

Entity Table

Table 4: Average (mean) earnings (£) of UK employees by 2010

	Women F/T £	Women P/T £	Men F/T £	Men P/T £
Managers and senior officials	18.66	15.74	24.67	xxx
Professional occupations	20.43	22.82	22.47	27.55
Associate professional and technical	14.85	14.77	16.84	15.41
Administrative and secretarial	10.80	9.54	12.05	9.73
Skilled trades	8.86	7.89	11.59	10.63

Matrix Table

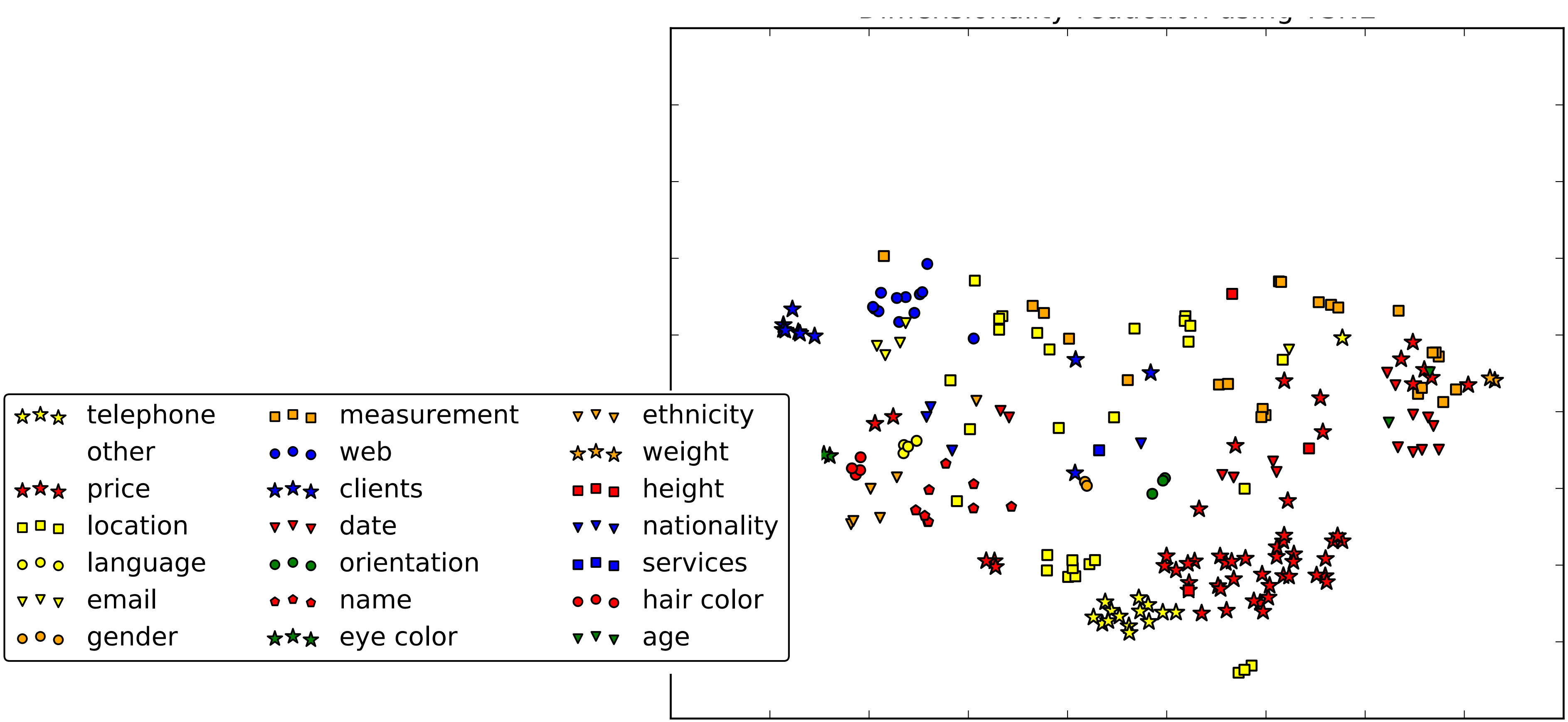
20 Strongest Performing Metro Areas
1. San Antonio, TX
2. Oklahoma City, OK
3. Austin, TX
4. Houston, TX
5. Dallas, TX
6. McAllen, TX
7. Little Rock, AR
8. Baton Rouge, LA
9. Tulsa, OK
10. Omaha, NE-IA
11. El Paso, TX
12. Wichita, KS
13. Washington, DC-VA-MD-WV
14. Des Moines, IA
15. Albuquerque, NM
16. Virginia Beach, VA-NC

List Table

Table Type Classification

- Feature-based supervised classification
  - Cafarella’08
  - Crestan’11
  - Eberius’15
- Deep Learning
  - Nishida’2017

Semantic + Structure Embedding



Data Extraction Techniques

- Glossary
- Regular expressions
- Natural language rules
- Named entity recognition
- Sequence labeling (Conditional Random Fields)

Searching Knowledge Graphs

Many problems with ‘strict’ execution

```
SELECT ?ad  WHERE
{
   ?ad a :Ad ;
      :hair_color 'Auburn' ;
      :review_site_id 'cg9469f' ;
      :price_per_hour '500' ;
      :name 'Claire Gold' ;
      :ethnicity ’Asian'.
}
```

synonyms “red”
typos “brunette”

not present

numbers hard to match

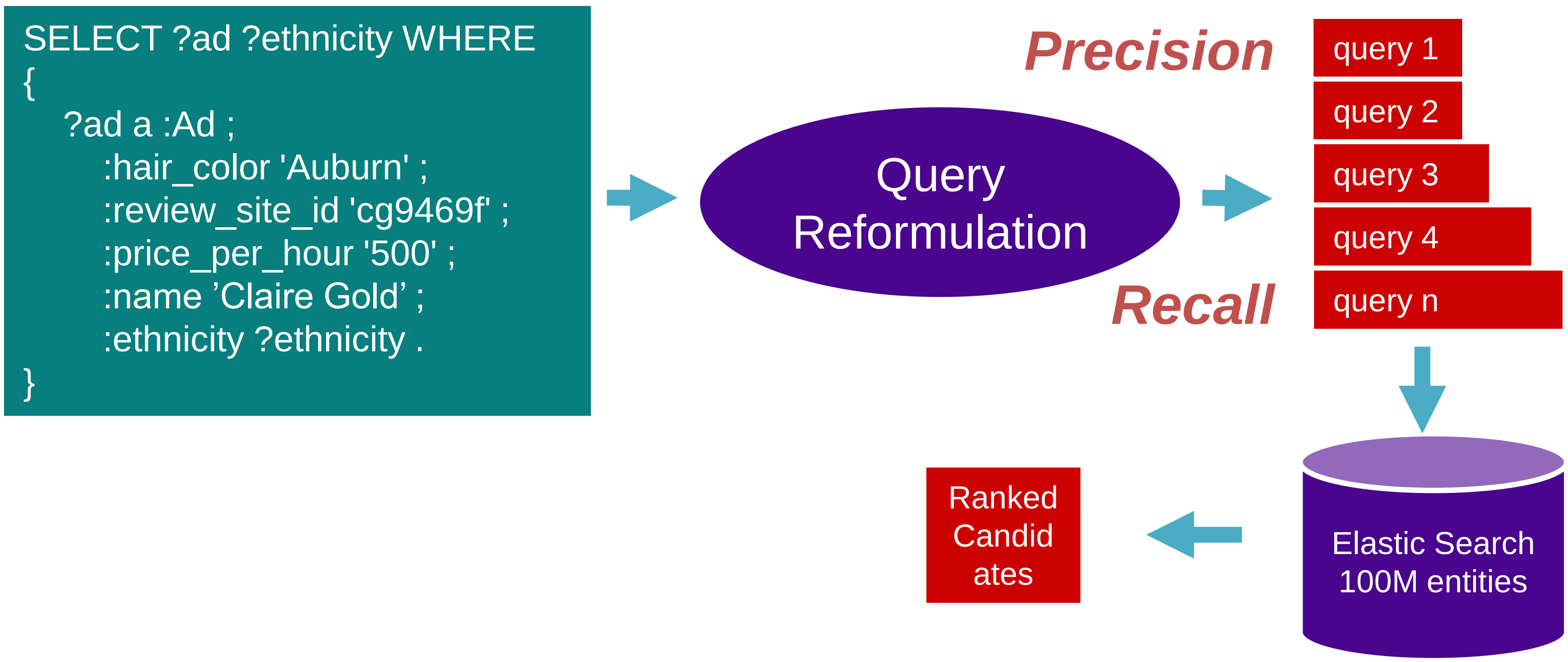
Claire is a common name
Gold is a domain word

slang, e.g., “FOB” for Asian
inference, e.g., “Japanese”

No results

Candidate Generation

Keyword expansion • Context broadening • Constraint relaxation



Knowledge Graph Completion

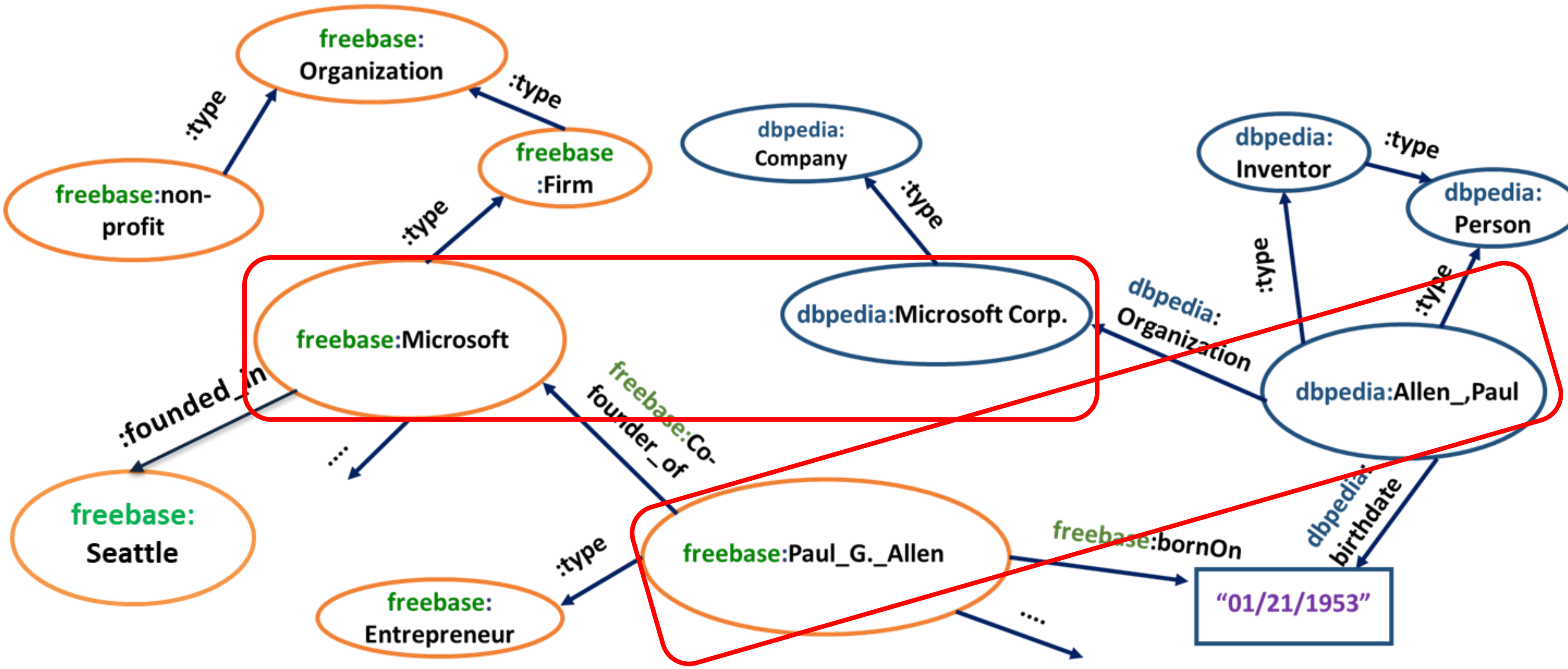
For Web extractions, noise is inevitable

- Thousands of web domains
- Many page formats
- Distracting & irrelevant content
- Purposeful obfuscation
- Poor grammar & spelling
- Tables

To reach its potential, a **constructed** KG must be **completed or identified**

Entity Resolution

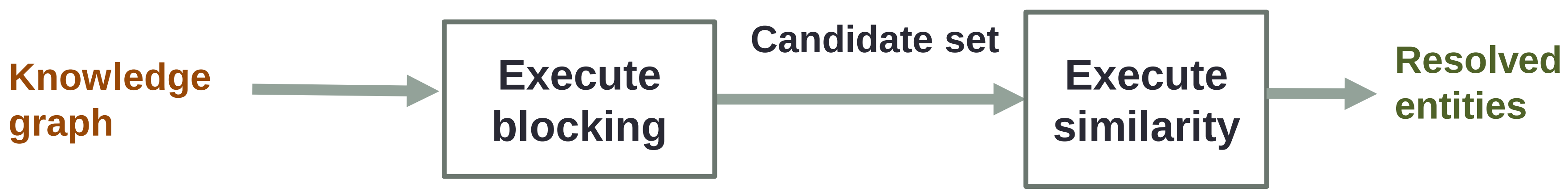
Many nodes refer to the same underlying entity



Entity Resolution is fundamentally non-linear

- Theoretically quadratic in the number of nodes, even if ‘resolution rule’ was known
- In practice, number of ‘duplicates’ tends to grow linearly, and duplicates overlap in non-trivial ways
- How to devise efficient algorithms?

50 years of research has agreed on a two-step solutions



Other research frontiers for KG Completion

- Other things explored in the literature:
  - **Domain knowledge**
    - **Collective ER methods have tried to exploit these systematically**
  - **Multi-type Entity Resolution**
    - **Extremely useful for knowledge graphs, lots more work to be done**
  - **Entity Resolution+Ontologies+IE Confidences:**
    - **Probabilistic Graphical Models like Probabilistic Soft Logic**
  - **Knowledge graph embeddings**
    - **Useful for link prediction and triples classification**
    - **Recall the Microsoft-founded_in-Seattle example earlier**

USC University of Southern California

Information Sciences Institute

DIG

THANK YOU! QUESTIONS...