Domain-Specific Corpora

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.

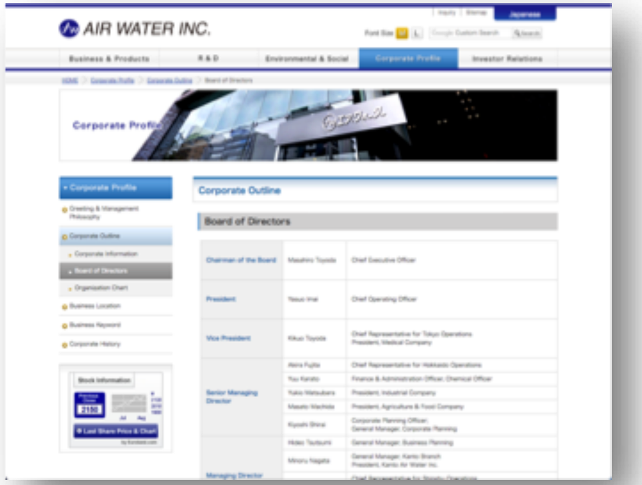
Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

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

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 year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

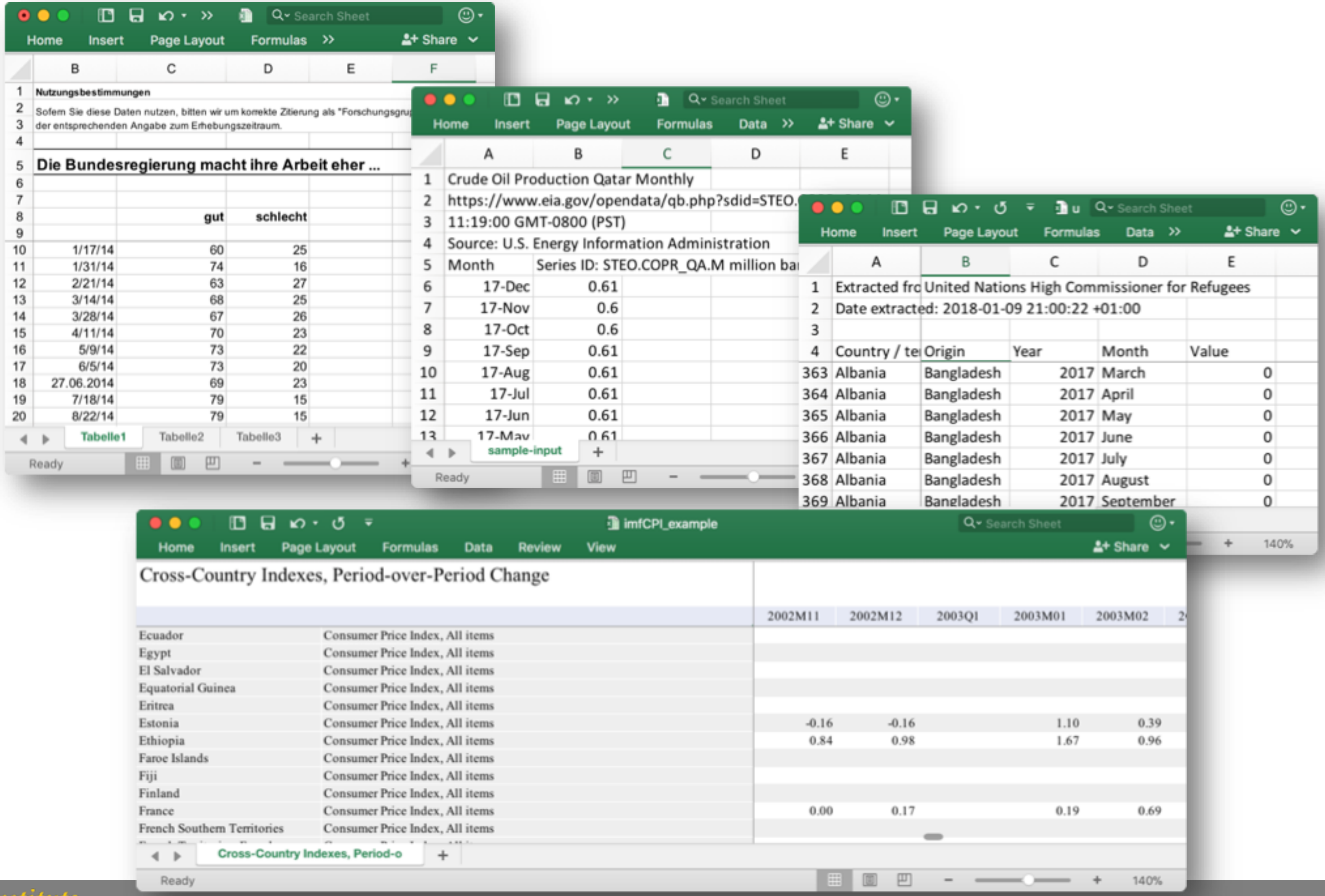
Unusual language models

"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"

small amount of relevant content

irrelevant content very similar to relevant content

Spreadsheets Created For Human Consumption



Databases with PDF Code Books

acled_raw_data

Home Insert Page Layout Formulas Data Review View

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	event_date	year	time_precisi	event_type	actor1	assoc_actor_	inter1	actor2	assoc_actor_	inter2	interaction	region	country	admin1
2	1/13/18	2018	1	Battle-No ch	Military Forces of Democ		1	ADF: Allied Democratic Fc		2	12	Central Afric	Democratic	Nord-Ki
3	1/13/18	2018	1	Battle-No ch	Military Forces of Democ		1	ADF: Allied Democratic Fc		2	12	Central Afric	Democratic	Nord-Ki
4	1/13/18	2018	1	Battle-No ch	Military Forces of Democ		1	ADF: Allied Democratic Fc		2	12	Central Afric	Democratic	Nord-Ki
5	1/13/18	2018	1	Battle-No ch	Al Shabaab		2	Police Forces of Kenya (20		1	12	Eastern Afric	Kenya	Lamu
6	1/13/18	2018	1	Riots/Protest	Rioters (Kenya)		5	Police Forces of Kenya (20		1	15	Eastern Afric	Kenya	Marsab
7	1/13/18	2018	1	Riots/Protest	Protesters (L	Tawergha Cc	6			0	60	Northern Afr	Libya	Sabha
8	1/13/18	2018	1	Violence aga	Unidentified Armed Grou		3	Civilians (Int	IOM: Interna		7	37	Northern Afr	Libya	Sabha
9	1/13/18	2018	1	Riots/Protest	Protesters (Morocco)		6			0	60	Northern Afr	Morocco	Orienta
10	1/13/18	2018	1	Riots/Protest	Protesters (Nigeria)		6			0	60	Western Afri	Nigeria	Benue
11	1/13/18	2018	1	V							7	Western Afri	Nigeria	Kaduna
12	1/13/18	2018	2	R							0	Western Afri	Nigeria	Lagos
13	1/13/18	2018	1	B							3	Eastern Afric	Somalia	Banaac
14	1/13/18	2018	1	B							2	Eastern Afric	Somalia	Shabee

acled_raw_data

Ready

130%

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

（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

Practical Considerations

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

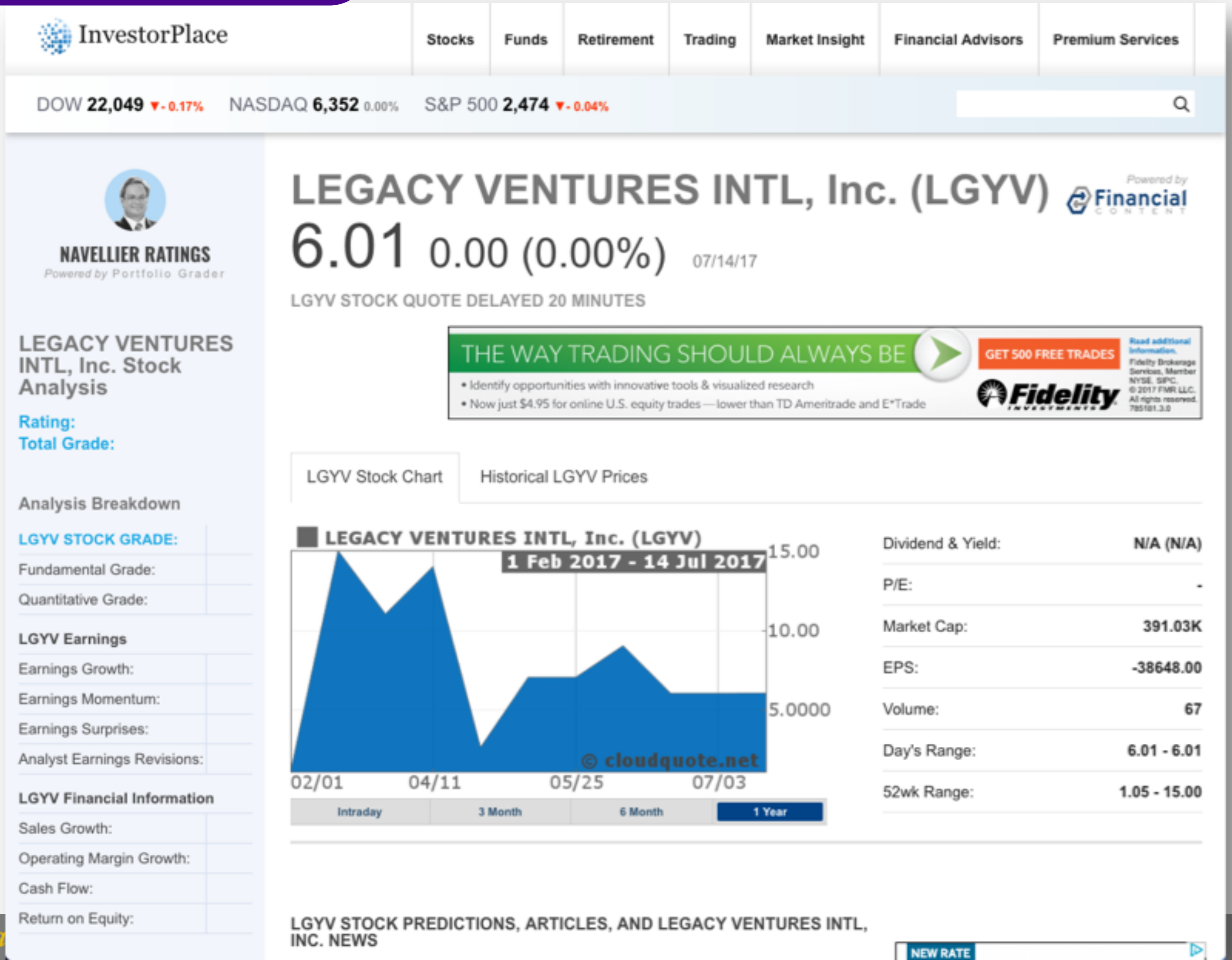
What tools can I use?

Many …

Information Extraction Process

Segmentation

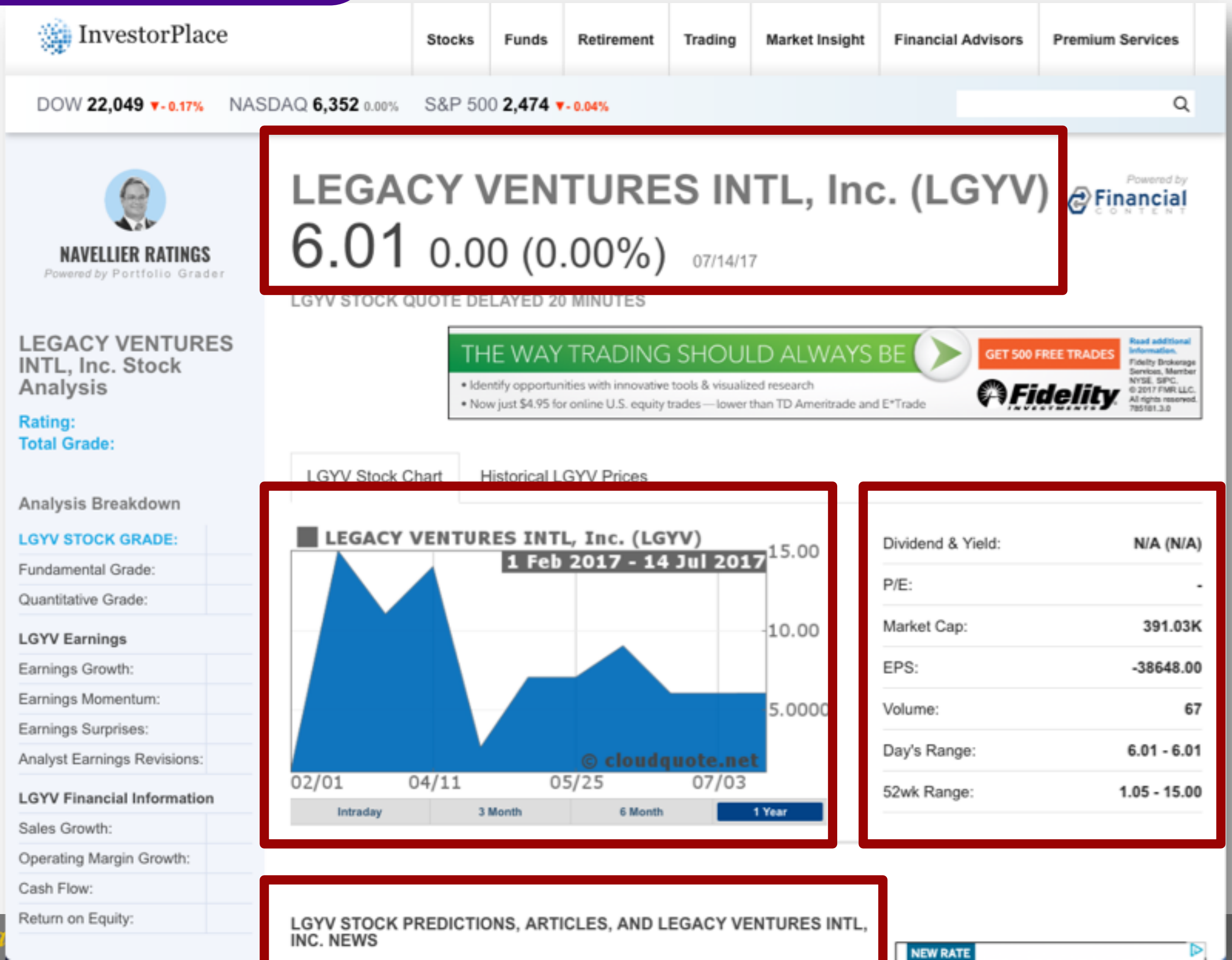
Data Extraction



Information Extraction Process

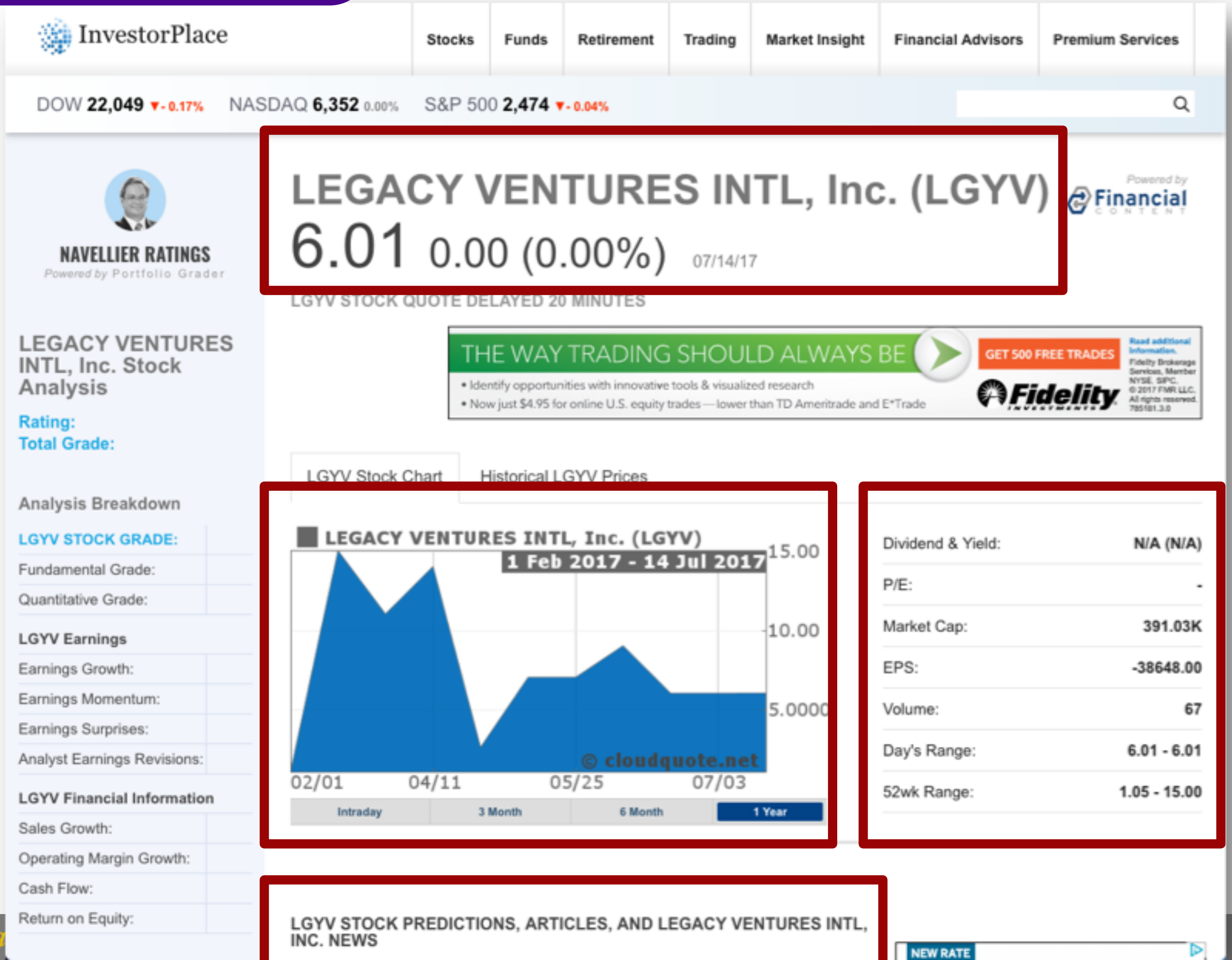
Segmentation

Data Extraction



Information Extraction Process

Segmentation



Data Extraction

Name:
Legacy Ventures Intl, Inc.

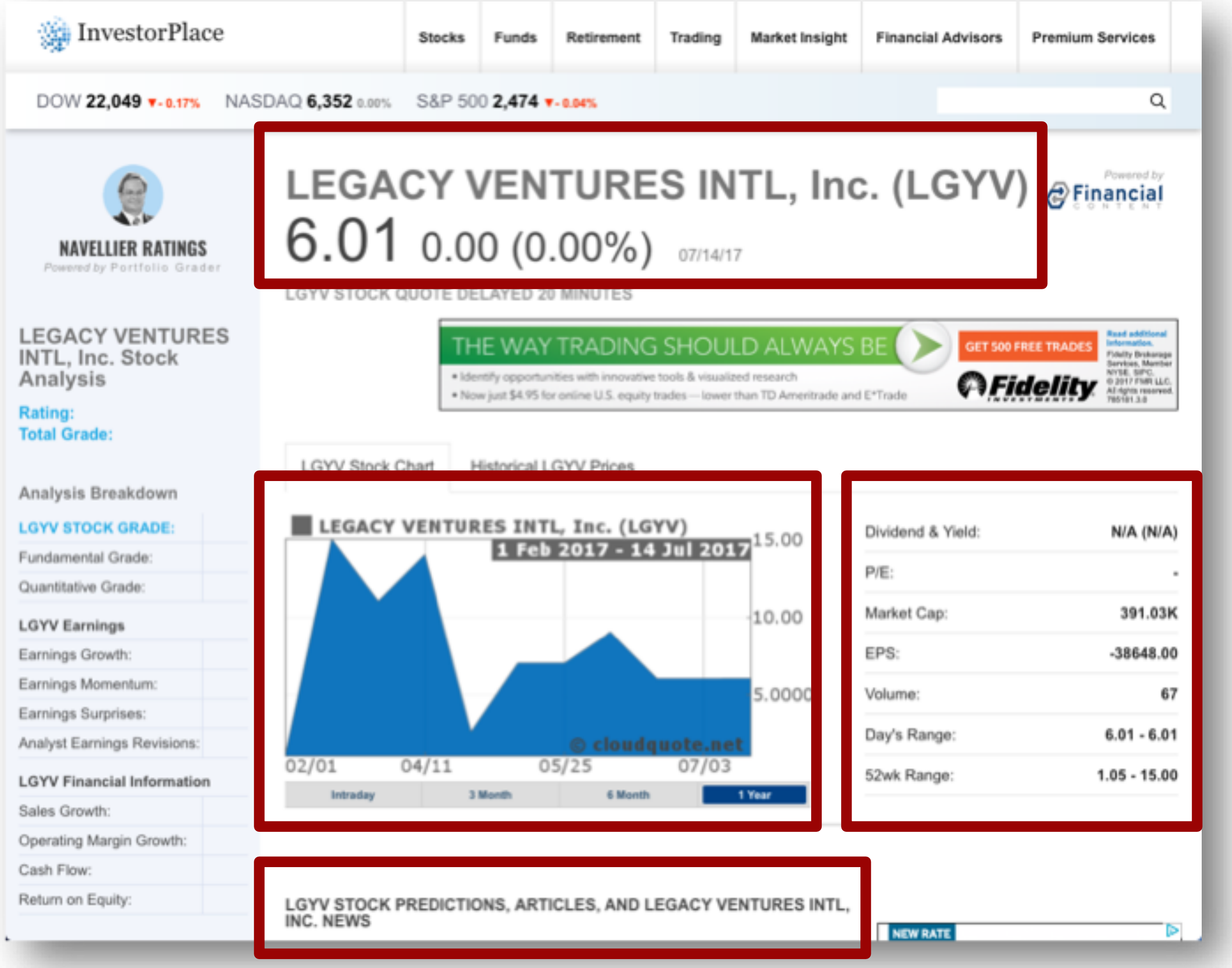
Stock:
LGYV

Date:
2017-07-14

Market Cap:
391,030

Segmentation

Segmentation

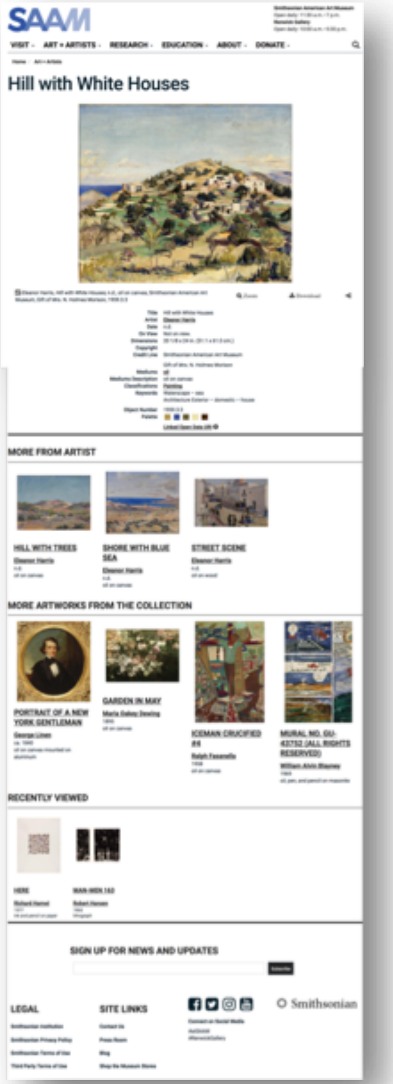


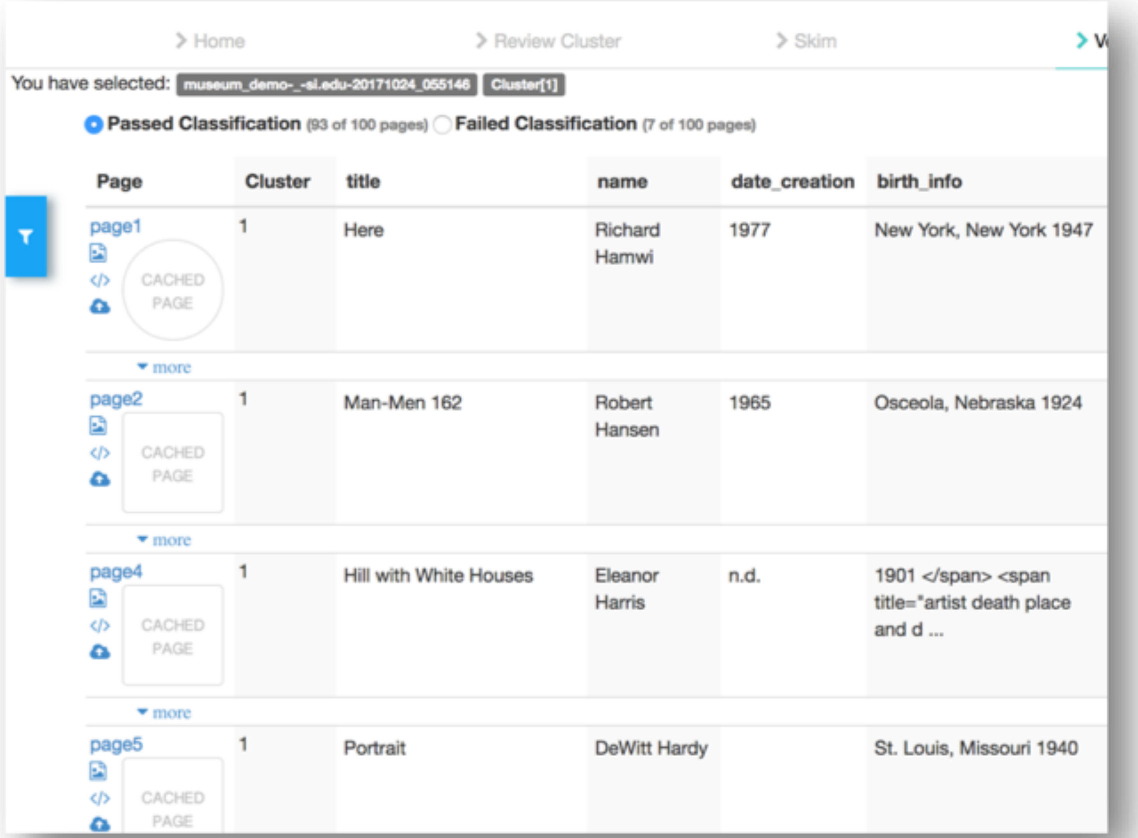
Homogeneous blocks

Segmentation

Block Type	Tool
Repeating blocks (short tail)	Web wrappers
Tables (long tail)	Data table extractors
Main content (long tail)	https://code.google.com/archive/p/arc90labs-readability/ https://github.com/kohlschutter/boilerpipe
Microdata (long tail)	https://github.com/namsral/microdata

Web Wrappers





Page	Cluster	title	name	date_creation	birth_info
page1	1	Here	Richard Hamwi	1977	New York, New York 1947
page2	1	Man-Men 162	Robert Hansen	1965	Osceola, Nebraska 1924
page4	1	Hill with White Houses	Eleanor Harris	n.d.	1901 <span title="artist death place and d ...
page5	1	Portrait	DeWitt Hardy		St. Louis, Missouri 1940

myDIG Demo

Focusing On Inferlink Web Wrapper

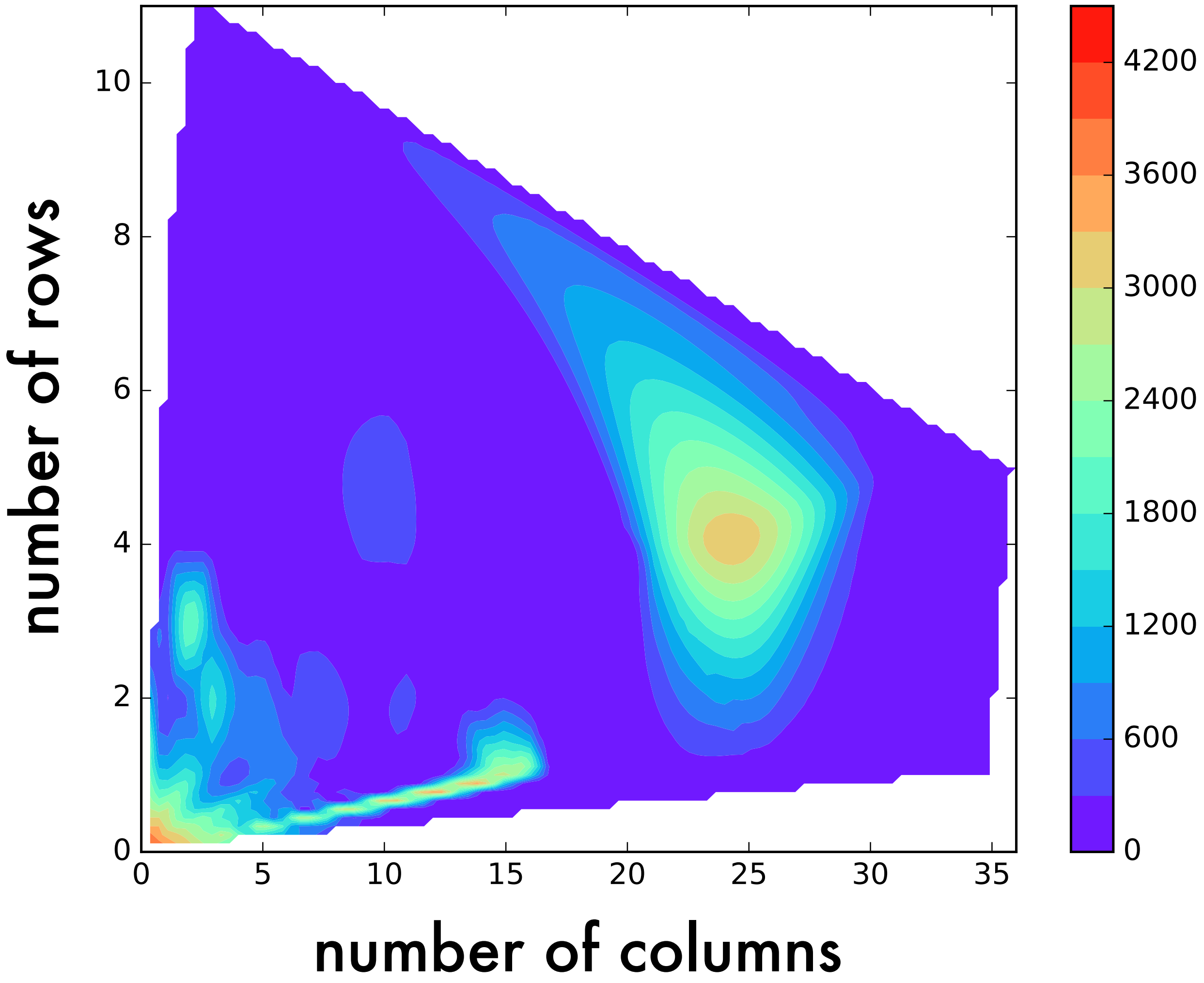
Table Extraction

Classification Of Web Tables

Table type	% total	count
"Tiny" tables	88.06	12.34B
HTML forms	1.34	187.37M
Calendars	0.04	5.50M
Filtered Non-relational, total	**89.44**	**12.53B**
Other non-rel (est.)	9.46	1.33B
Relational (est.)	1.10	154.15M

Cafarella'08

Tables In The Human Trafficking Domain



Data Tables

Name	Nationality	From	To	M	W	D	L	GF	GA	Win %	Honour
Arsène Wenger	France	1 October 1996	*Present*	1,188	684	271	233	2,063	1,088	57.58	Premier League champions: 1997–98, 2 FA Cup winners: 1997–98, 2001–02, 200 15, 2016–17 Charity/Community Shield winners: 1998 2017
Pat Rice †	Northern Ireland	13 September 1996	30 September 1996	4	3	0	1	10	4	75.00	
Stewart Houston †	Scotland	12 August 1996	13 September 1996	6	2	2	2	11	10	33.33	
Bruce Rioch	Scotland	15 June 1995	12 August 1996	47	22	15	10	67	37	46.81	
Stewart Houston †	Scotland	21 February 1995	15 June 1995	19	7	3	9	29	25	36.84	
George Graham	Scotland	14 May 1986	21 February 1995	460	225	133	102	711	403	48.91	First Division champions: 1988–89, 1990 FA Cup winners: 1992–93 Football League Cup winners: 1986–87, Charity Shield winners: 1991 (shared) UEFA Cup Winners' Cup winners: 1993–
Steve Burtenshaw †	England	23 March 1986	14 May 1986	11	3	2	6	7	15	27.27	

Relational

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

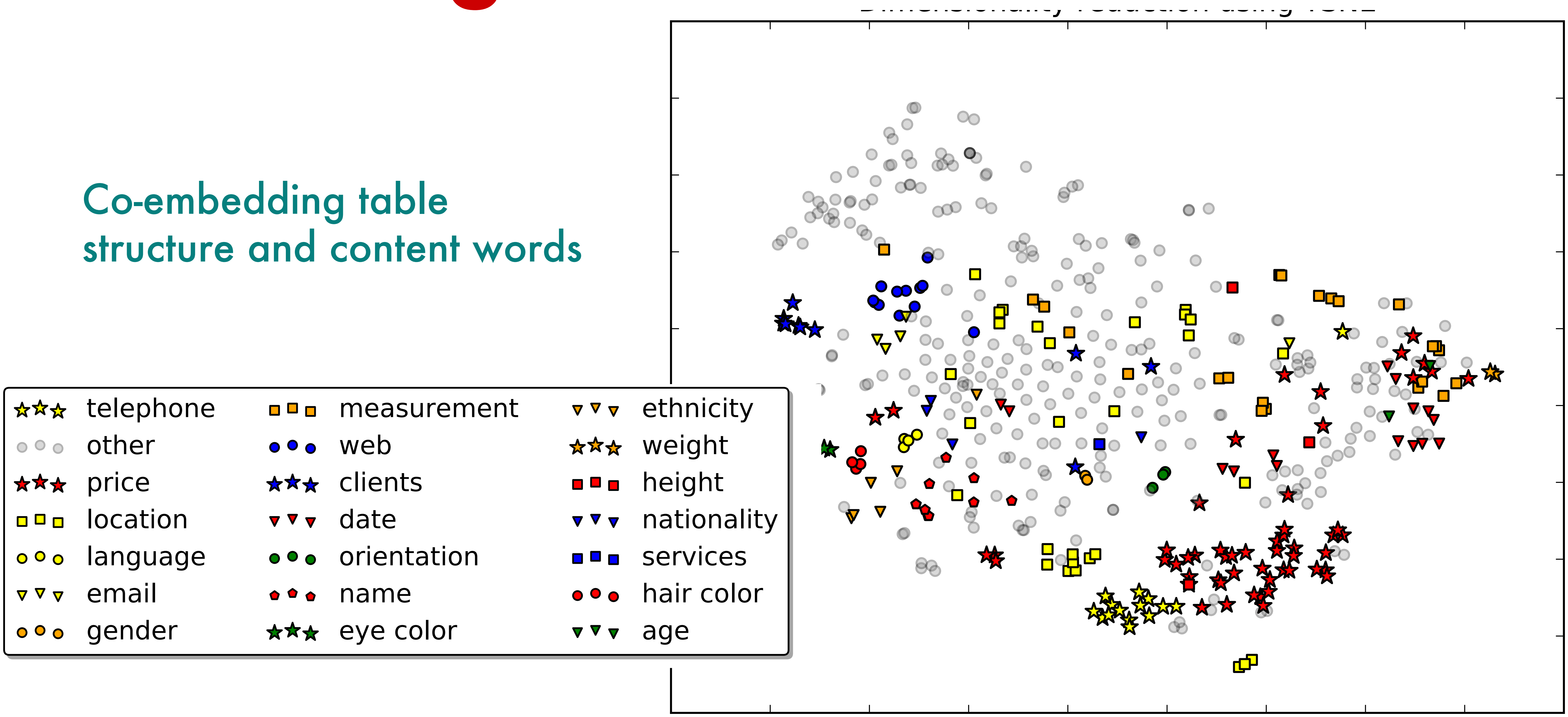
Identifying Data Tables

Heuristic

HTML tables that don’t contain nested tables and contain at least 2 rows and 2 columns

Extracting Data From Tables

Co-embedding table
structure and content words



Data Extraction

Data Extraction Techniques

Glossary

Regular expressions

Natural language rules

Named entity recognition

Sequence labeling (Conditional Random Fields)

Glossary Extraction

Glossary Extraction

Simple

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 using context

Creating/extending glossaries automatically

Regex Extraction

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 specific domains

unusual nomenclature and short-hands

obfuscation

NLP Rule-Based Extraction

Rule-based matching

spaCy features a rule-matching engine that operates over tokens, similar to regular expressions. The rules can refer to token annotations and flags, and matches support callbacks to accept, modify and/or act on the match. The rule matcher also allows you to associate patterns with entity IDs, to allow some basic entity linking or disambiguation.

Here's a minimal example. We first add a pattern that specifies three tokens:

1. A token whose lower-case form matches "hello"
2. A token whose `is_punct` flag is set to `True`
3. A token whose lower-case form matches "world"

Once we've added the pattern, we can use the `matcher` as a callable, to receive a list of `(ent_id, start, end)` tuples.

```
from spacy.matcher import Matcher
from spacy.attrs import IS_PUNCT, LOWER

matcher = Matcher(nlp.vocab)
matcher.add_pattern("HelloWorld", [{LOWER: "hello"}, {IS_PUNCT: True}, {L
```

spaCy chat

NLP Rule-Based Extraction

Tokenization

Pattern Matching

Tokenization matters, a lot

My name is Pedro	My	name	is	Pedro	
310-822-1511	310-822-1511				
	310	-	822	-	1511
♥Candy♥ is here	♥	Candy	♥	is	here
	♥Candy♥	is	here		

Token Properties

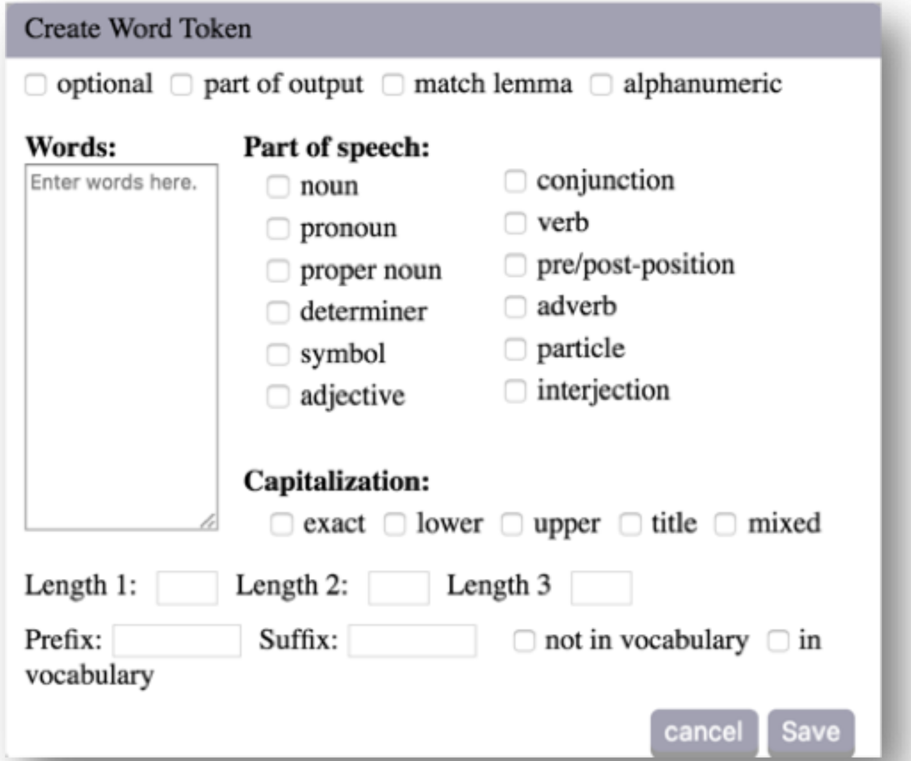
Surface properties

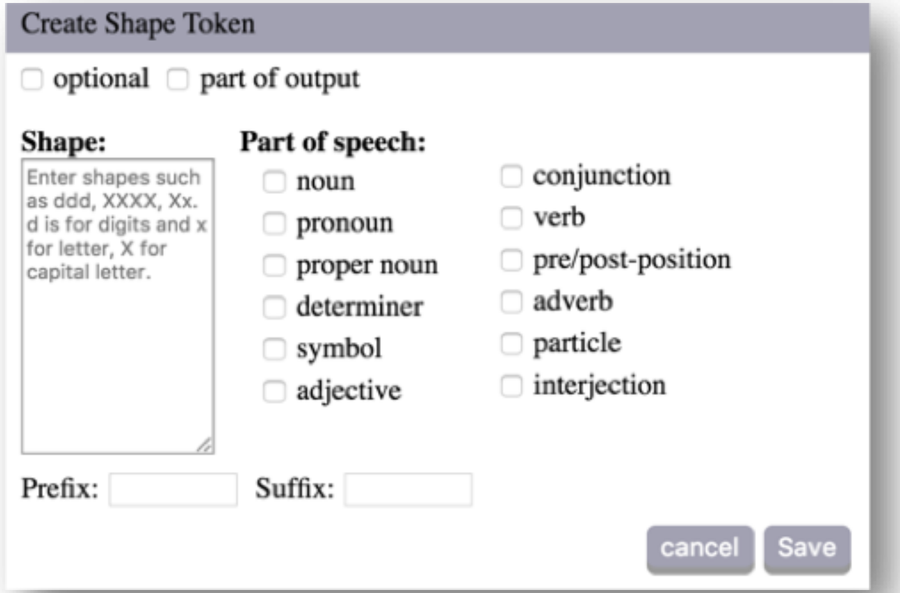
Literal, type, shape, capitalization, length, prefix, suffix, minimum, maximum

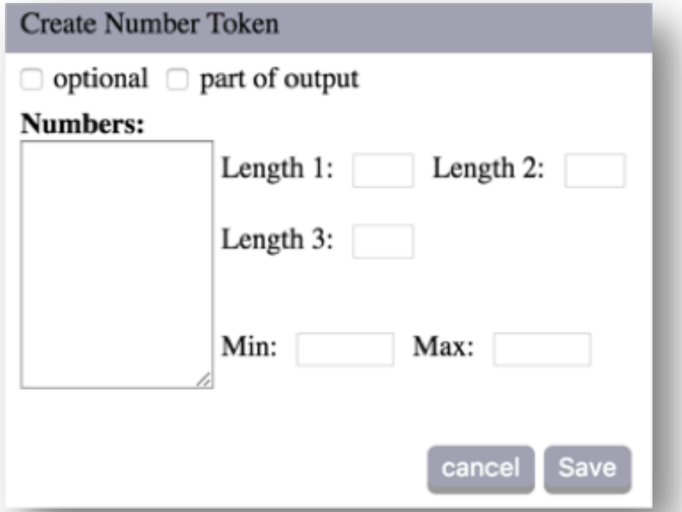
Language properties

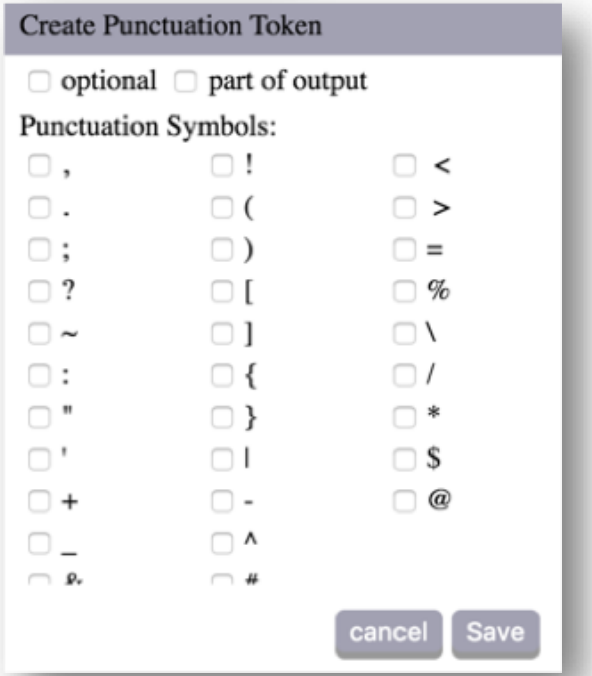
Part of speech tag, lemma, dependency

Token Types


Create Word Token
optional part of output match lemma alphanumeric
Words:
Enter words here.
Part of speech:
noun
pronoun
proper noun
determiner
symbol
adjective
conjunction
verb
pre/post-position
adverb
particle
interjection
Capitalization:
exact lower upper title mixed
Length 1: Length 2: Length 3
Prefix: Suffix: not in vocabulary in vocabulary
cancel Save


Create Shape Token
optional part of output
Shape:
Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.
Part of speech:
noun
pronoun
proper noun
determiner
symbol
adjective
conjunction
verb
pre/post-position
adverb
particle
interjection
Prefix: Suffix:
cancel Save


Create Number Token
optional part of output
Numbers:
Length 1: Length 2:
Length 3:
Min: Max:
cancel Save


Create Punctuation Token
optional part of output
Punctuation Symbols:
, . ; ? ~ : " ' + _
! ( ) [ ] { } | - ^
< > = % \ / * $ @
cancel Save

Patterns

```
Pattern := Token-Spec

           [Token-Spec]      Optional

           Token-Spec +      One or more

           Token-Spec  Pattern
```

Positive/Negative Patterns

General **Positive**

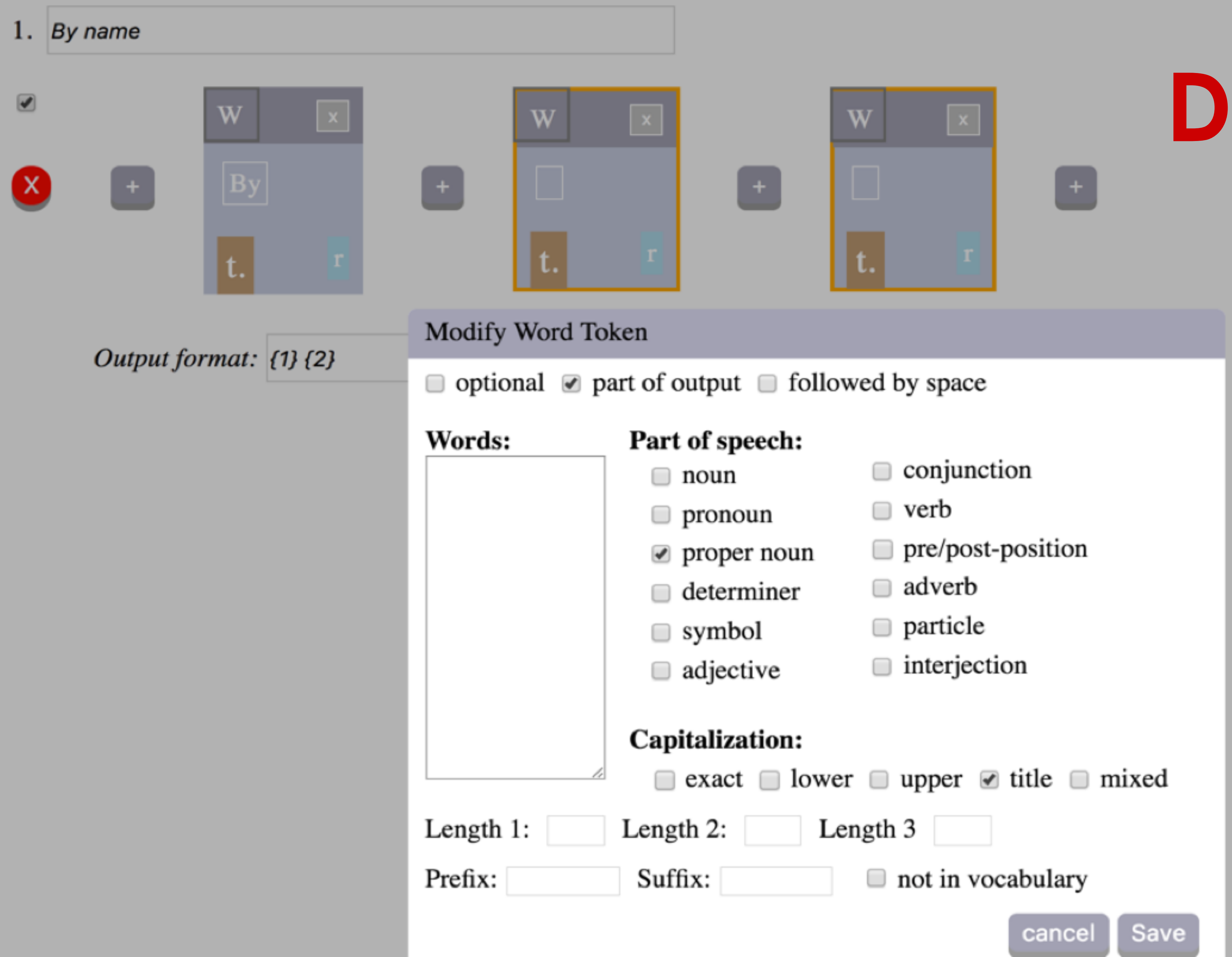
Generate candidates

Specific **Negative**

Remove candidates

Output overlaps positive candidates

DIG Demo


1. By name
W
By
t.
r
W
t.
r
W
t.
r
Output format: {1} {2}
Modify Word Token
optional
part of output
followed by space
Words:
Part of speech:
noun
pronoun
proper noun
determiner
symbol
adjective
conjunction
verb
pre/post-position
adverb
particle
interjection
Capitalization:
exact
lower
upper
title
mixed
Length 1:
Length 2:
Length 3
Prefix:
Suffix:
not in vocabulary
cancel
Save

NLP Rule-Based Extraction

Advantages

Easy to define

High precision

Recall increases with number of rules

Disadvantages

Text must follow strict patterns

Named-Entity Recognizers

Named Entity Recognizers

Machine learning models

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

Entity recognition

spaCy features an extremely fast statistical entity recognition system, that assigns labels to contiguous spans of tokens. The default model identifies a variety of named and numeric entities, including companies, locations, organizations and products. You can add arbitrary classes to the entity recognition system, and update the model with new examples.

EXAMPLE

```
import spacy
nlp = spacy.load('en')
doc = nlp(u'London is a big city in the United Kingdom
for ent in doc.ents:
    print(ent.label_, ent.text)
    # GPE London
    # GPE United Kingdom
```

The standard way to access entity annotations is the `doc.ents` property, which produces a sequence of `Span` objects. The entity type is accessible either as an integer ID or as a string, using the attributes `ent.label` and `ent.label_`. The `Span` object acts as a sequence of tokens, so you can iterate over the entity or index into it. You can also get the text form of the whole entity, as though it were a single token. See the API reference for more details.

You can access token entity annotations using the `token.ent_iob` and `token.ent_type` attributes. The `token.ent_iob` attribute indicates whether an entity starts, continues or ends on the tag (In, Begin, Out).

EXAMPLE

```
doc = nlp(u'London is a big city in the United Kingdom.')
print(doc[0].text, doc[0].ent_iob, doc[0].ent_type_)
```

spaCy chat

https://demos.explosion.ai/displacy-ent

EXPLOSION

displaCy

Dependency Visualizer

Named Entity Visualizer
Visualise spaCy's guess at the named entities in the document. You can filter the displayed types, to only show the annotations you're interested in.

Similarity

Sentence Similarity

sense2vec: Semantic Analysis of the Reddit Hivemind

View on GitHub

displaCy Named Entity Visualizer

Enter your text below to explore spaCy's default entity recognition model. You can use the drop-down menu to select the entity types you're interested in.

2 April 2016 Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum The Nigeria Labour Congress (NLC) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union, Ayuba Wabba, gave the Union's unalloyed support in the fight against corruption during a visit to the chairman of the EFCC, Ibrahim Magu his Abuja office. A statement yesterday from the EFCC spokesman Wilson Uwujaren

Entities ⌄ Model ⌄

2 April 2016 DATE Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum PERSON The Nigeria Labour Congress ORG (NLC ORG) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union, Ayuba Wabba PERSON , gave the Union's unalloyed support in the fight against corruption during a visit to the chairman of the EFCC ORG , Ibrahim Magu PERSON his Abuja ORG office. A statement yesterday DATE from the EFCC ORG spokesman Wilson Uwujaren PERSON quoted Wabba PERSON as

Named Entity Recognizers

Advantages

Easy to use

Tolerant of some noise

Easy to train

Disadvantages

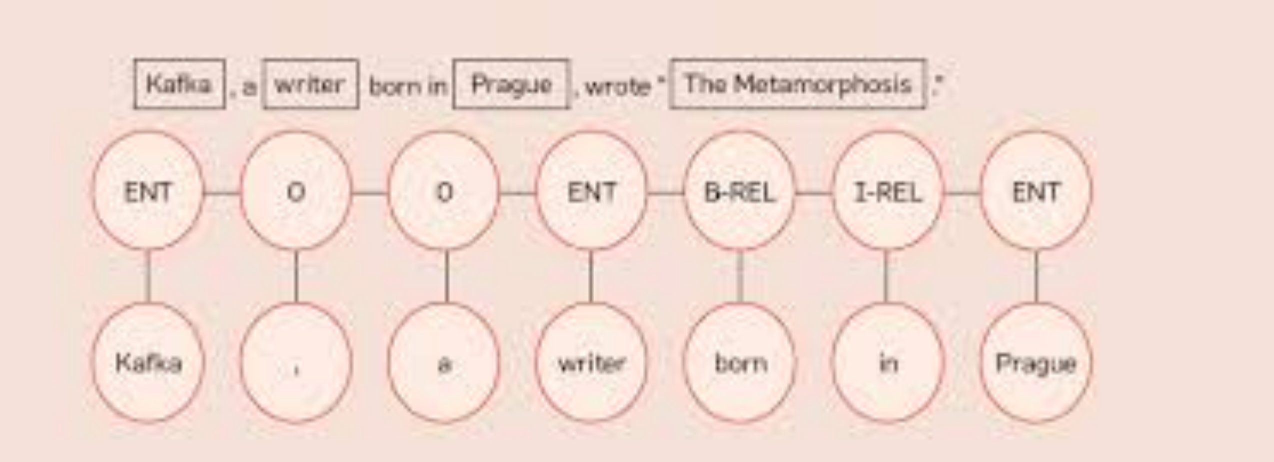
Performance degrades rapidly for new genres, language models

Requires hundreds to thousands of training examples

Conditional Random Fields

Conditional Random Fields (CRF)

Good for fields that have regular text structure/context



In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Modeling Problems With CRF

i	X1 (word)	X2 (capitalized)	X3 (POS Tag)	Y (entity)
1	My	1	Possessive Pron	Other
2	name	0	Noun	Other
3	is	0	Verb	Other
4	Pedro	1	Proper Noun	Person-Name
5	Szekely	1	Proper Noun	Person-Name

Other common features:

lemma, prefix, suffix, length

CRF Advantages/Disadvantages

Advantages

Expressive

Tolerant of noise

Stood test of time

Software packages available

Disadvantages

Requires feature engineering

Requires thousands of training examples

Open Information Extraction

http://openie.allenai.org/

Open Information Extraction

Hosted by

AI2

Created at

Example Queries:

- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

Typed Example Queries:

- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?

Argument 1:	Relation:	Argument 2:	Corpus:	
what/who	verb phrase	what/who	All	Search

AI2 proudly announces the launch of Semantic Scholar, an AI-based academic search engine.

To learn more about Open IE, watch our YouTube video!

Powered by ReVerb, our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

NEW! **Open IE 4.0**, the successor to ReVerb and Ollie, has been released. Download it from GitHub!

Publications:

- Search Needs a Shake-up (Nature 2011)
- Open Information Extraction (IJCAI 2011)
- Ollie (EMNLP 2012)
- Reverb (EMNLP 2011)
- TextRunner (IJCAI 2007)

Public resources based on Open IE:

- 18 million question-paraphrases (Fader et al. ACL 2013)

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	genre	narrow

how to **represent** KGs?

KG Definition

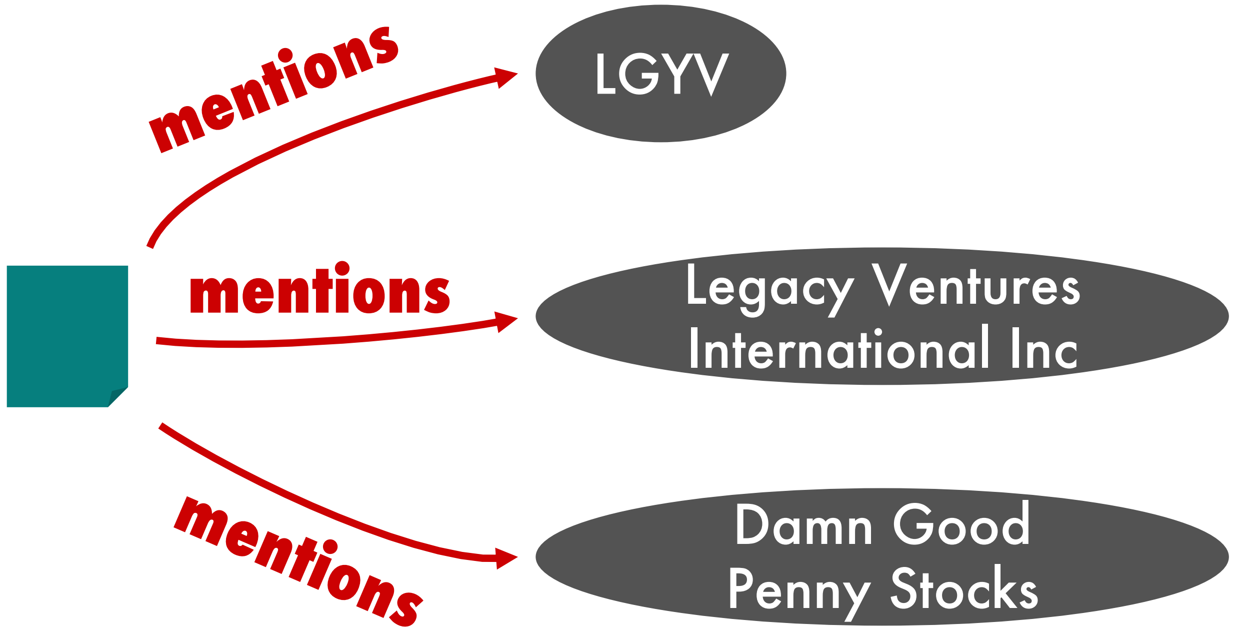
a directed, labeled multi-relational graph representing facts/assertions as triples

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

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

Simplest Knowledge Graph

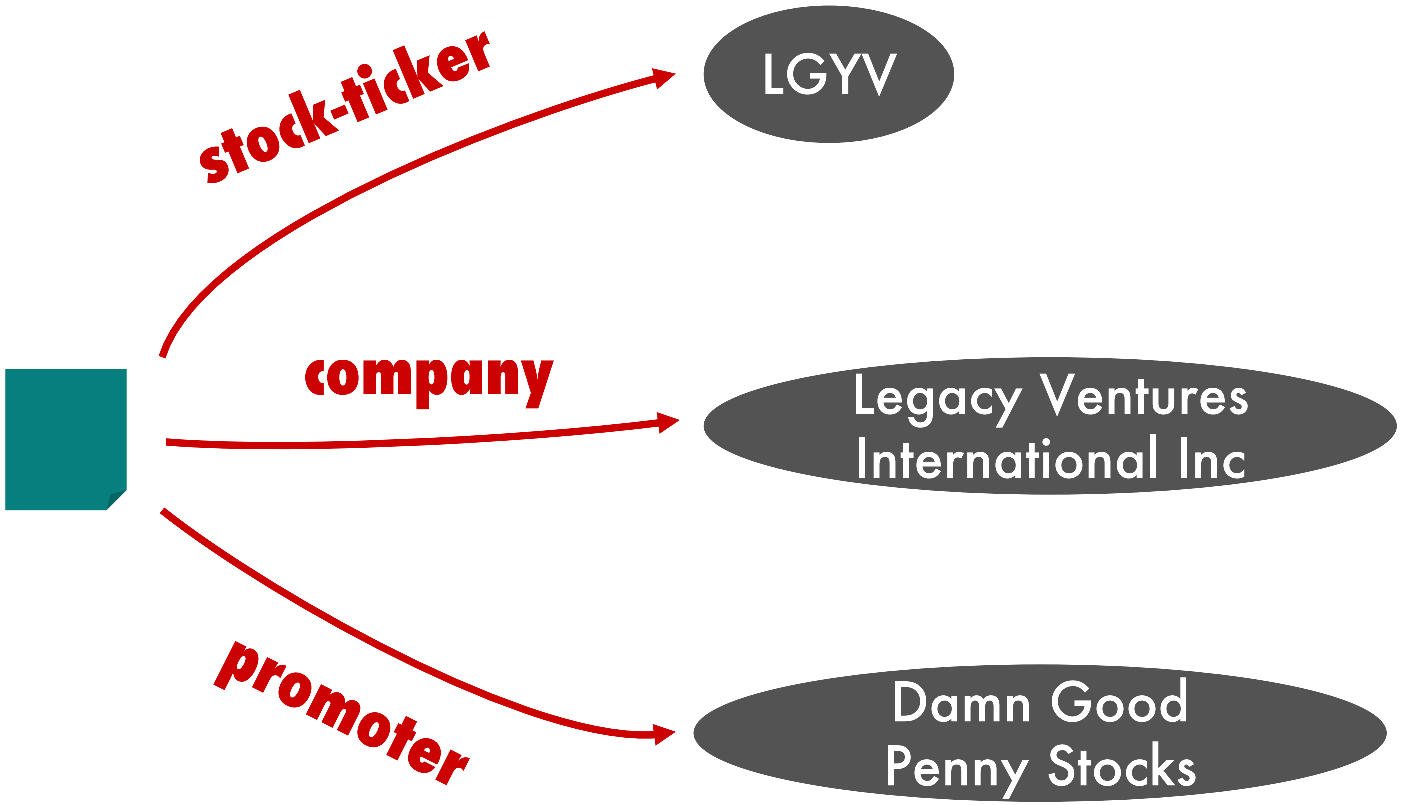
Entities



Easiest to build

Simple, But Useful KG

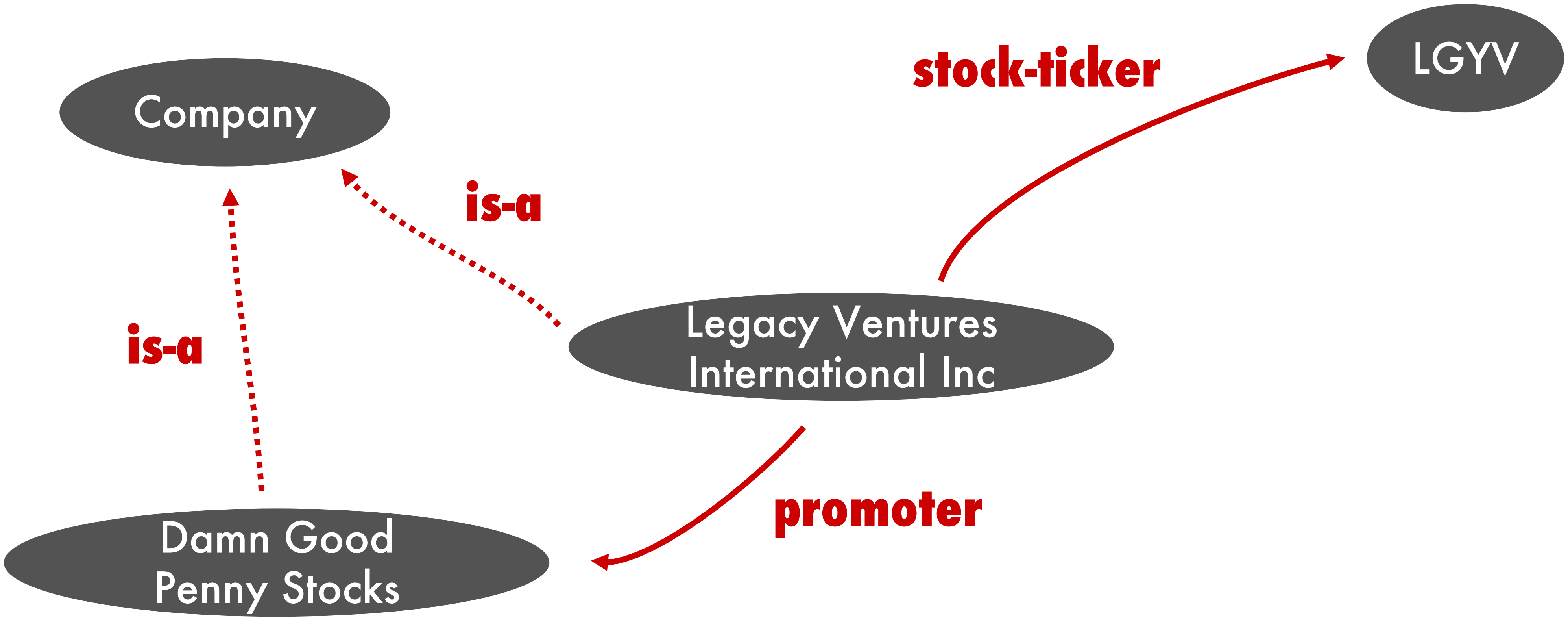
Entities + properties



"Easy" to build

Semantic Web KG (RDF/OWL)

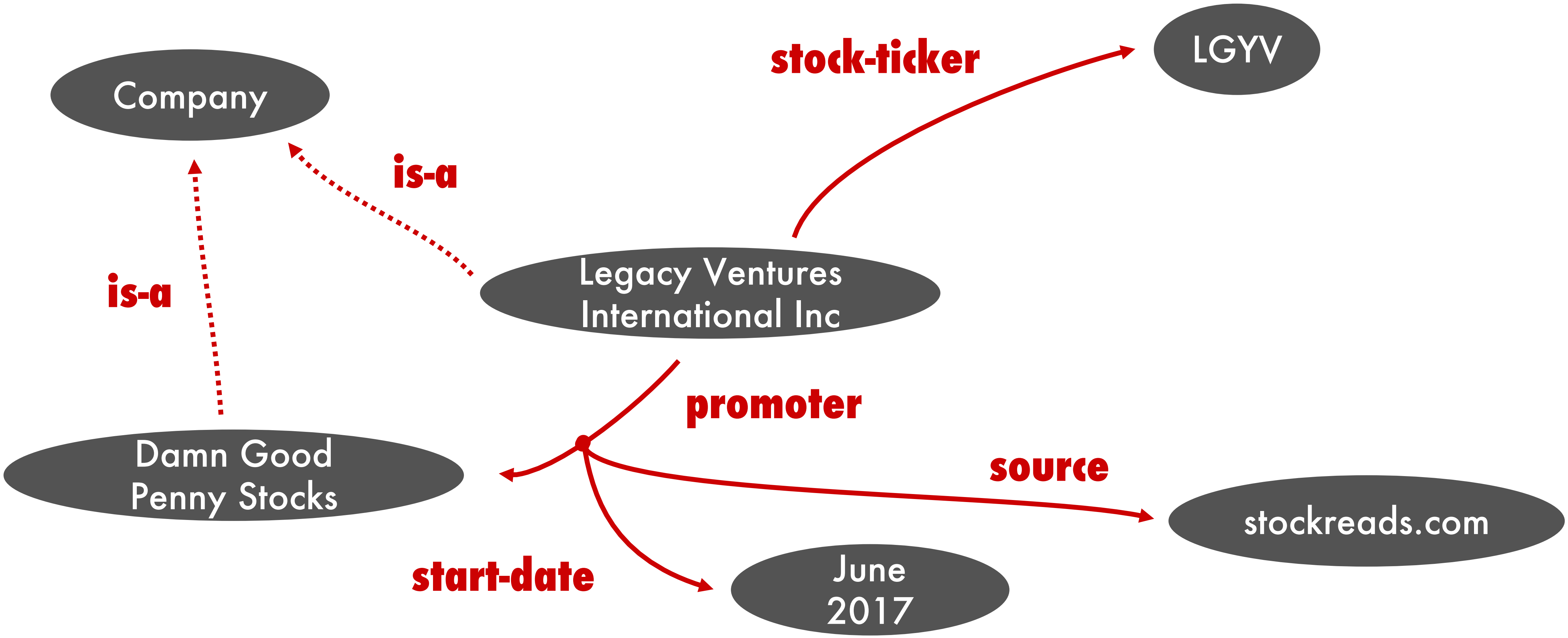
Entities + properties + classes



Very hard to build

“Ideal” KG

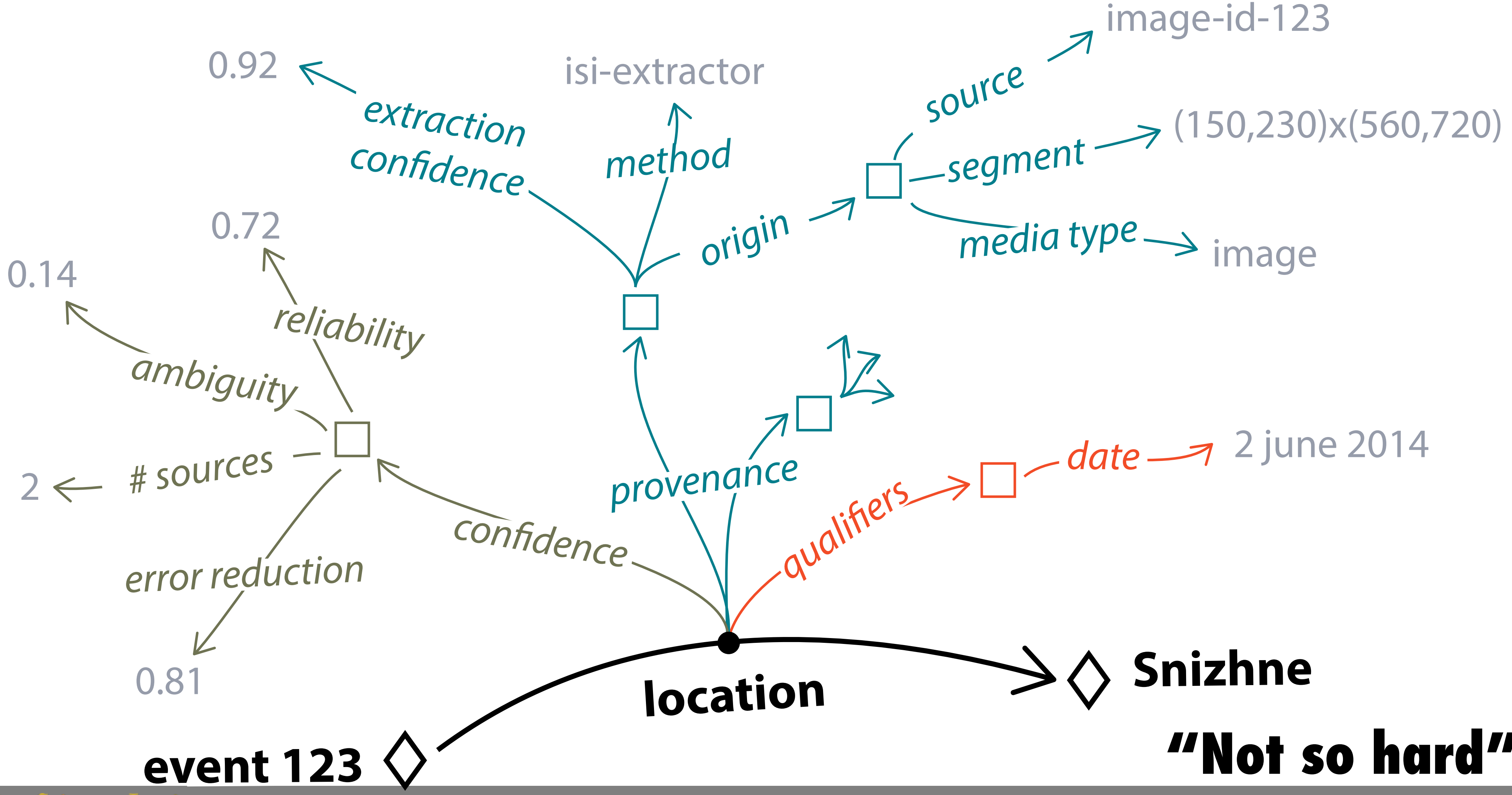
Entities + properties + classes + qualifiers



Very very hard to build

Semi-Structured KG

Entities + properties + text + provenance + confidence



"Not so hard" to build

Where to **Store KGs?**

Serializing Knowledge Graphs

Resource Description Framework (RDF)

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

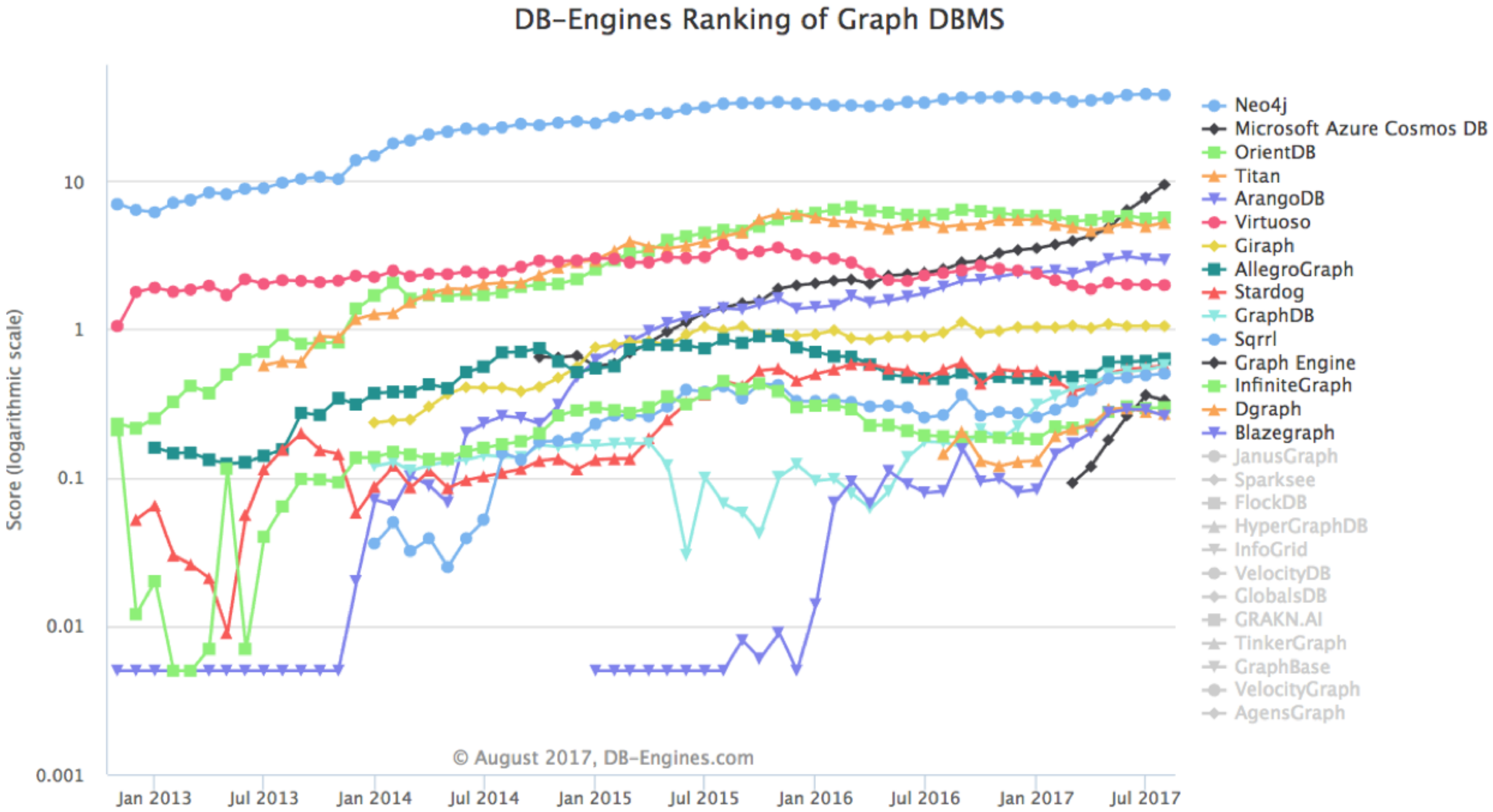
Key-Value, Document Stores

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

Graph Databases

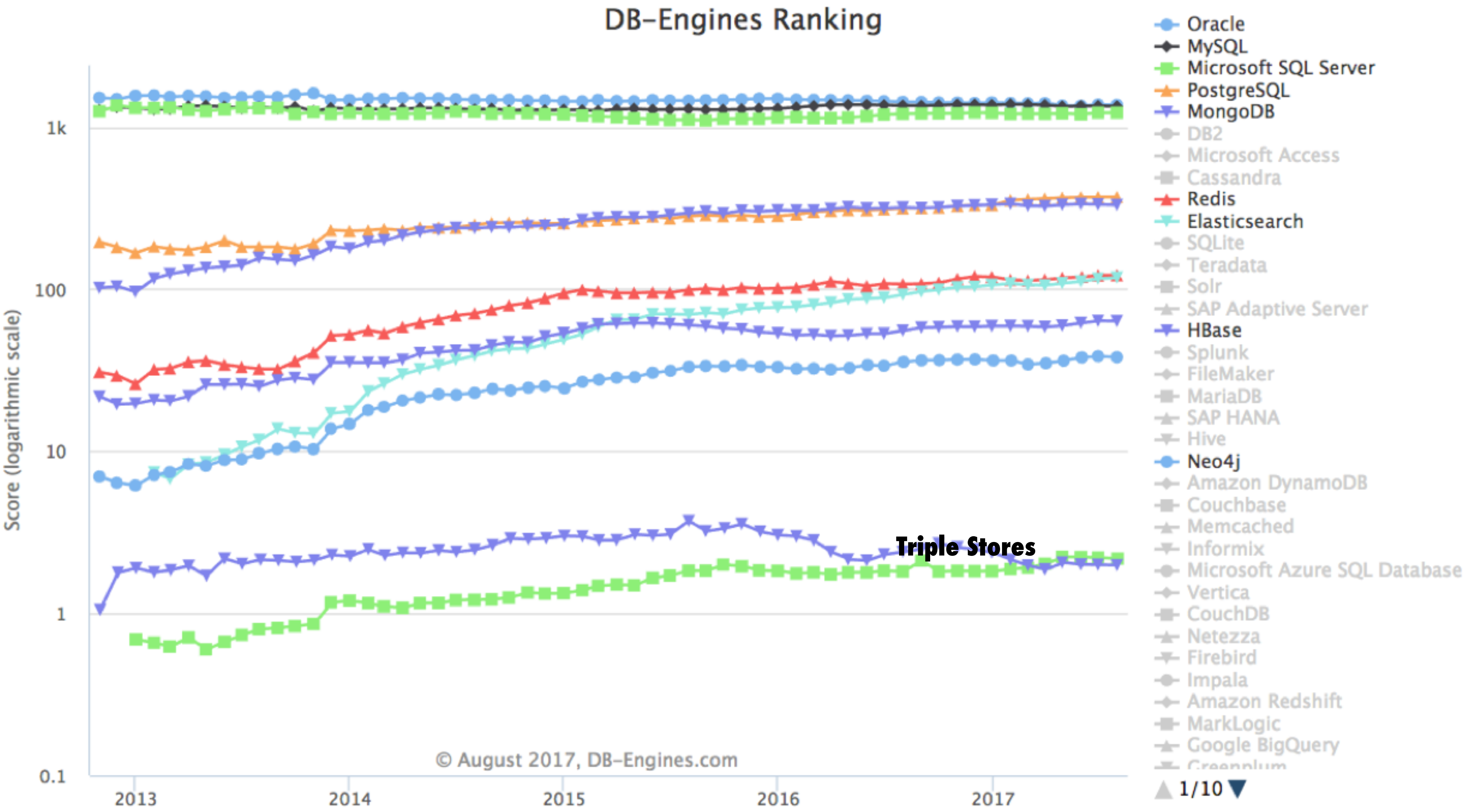
Data model: graph
Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, …
Query: GraphQL, Gremlin, Cypher

Popularity Ranking Of Graph Databases



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

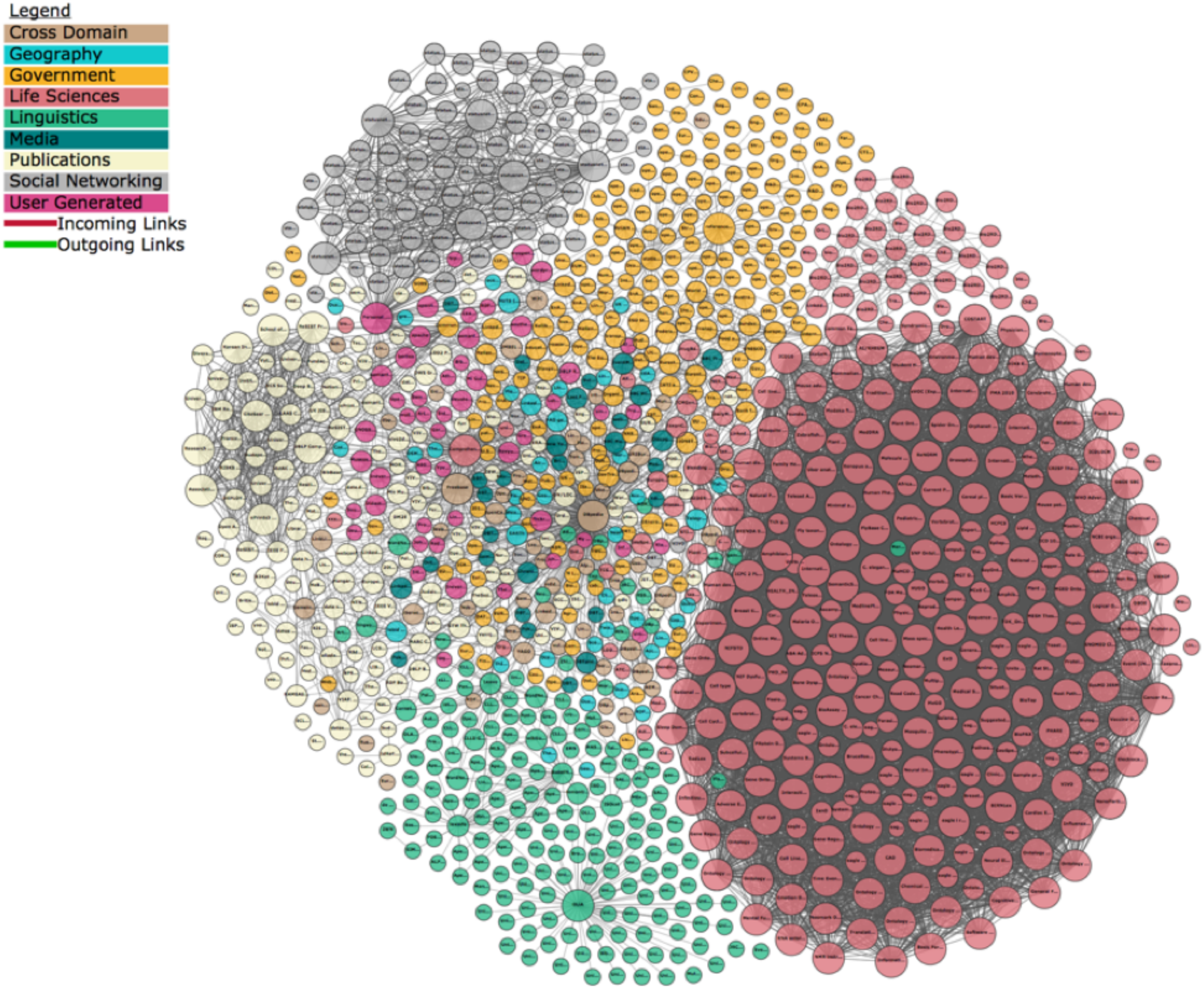
ElasticSearch, MongoDB & Neo4J Have Wide Adoption



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

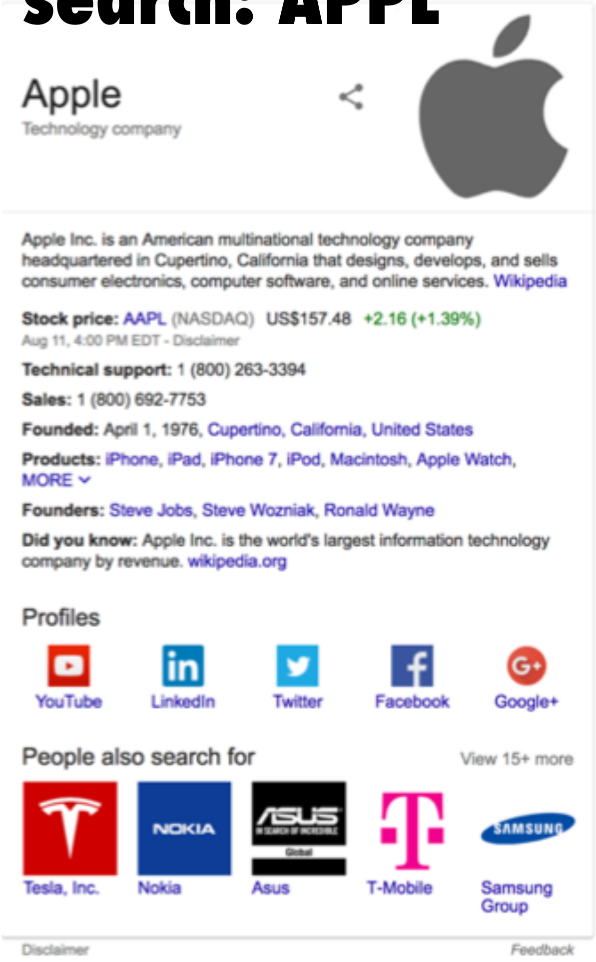
31,419,072 items

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/

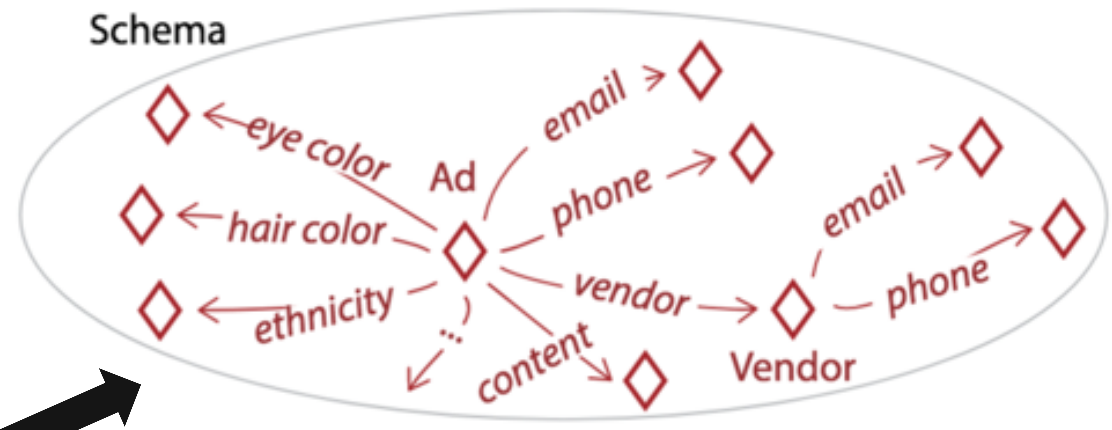
LinkedIn Knowledge Graph

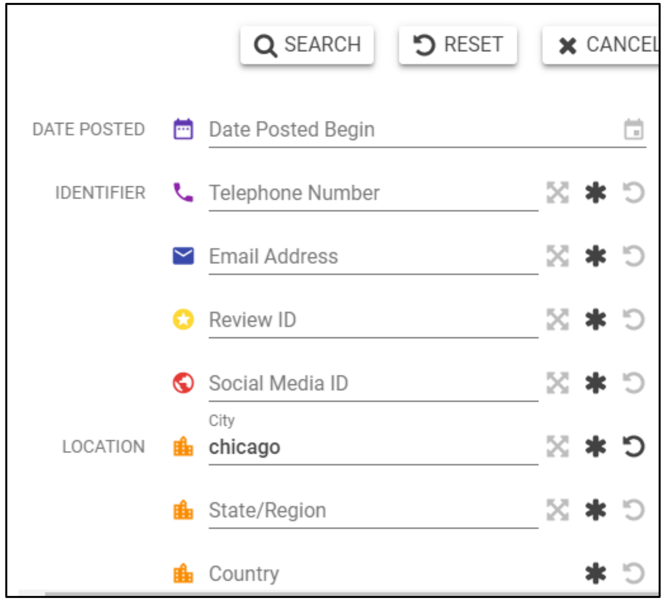
450M members, 190M historical job listings, 9M companies, 28K schools,
1.5K fields of study, 600+ degrees, 24K titles and 35K skills in 19 languages

Querying Knowledge Graphs

Knowledge Graph Query

*What is the **ethnicity** listed in the **ad** that contains the **phone number** **6135019502**, **located in** **Toronto Ontario**, with the **title** **'the millionaires mistress'**?*






```
SELECT ?ad ?ethnicity WHERE {
  ?ad a :Ad ;
    :phone '6135019502' ;
    :location 'Toronto, Ontario' ;
    :title 'the millionaires mistress' ;
    :ethnicity ?ethnicity . }
```

Why can't I just 'execute' the query?

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

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

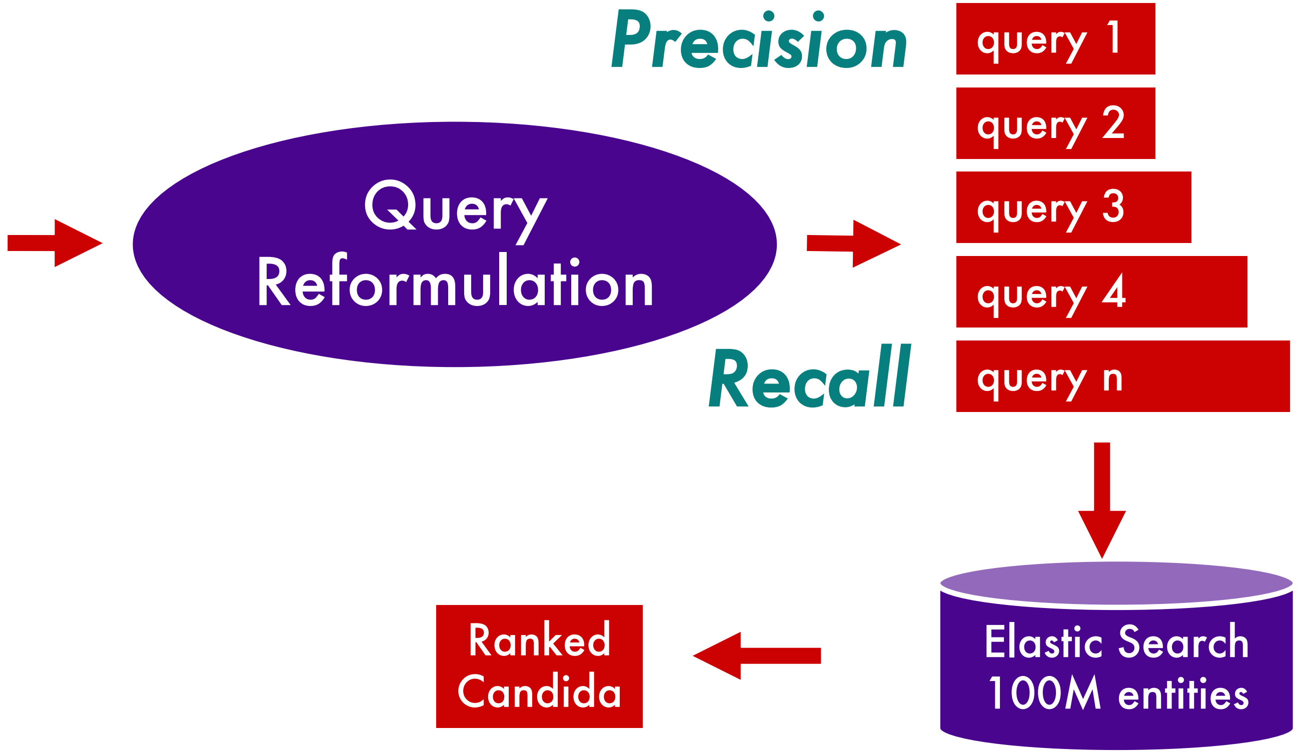
slang, e.g., "FOB" for Asian
inference, e.g., "Japanese"

No results

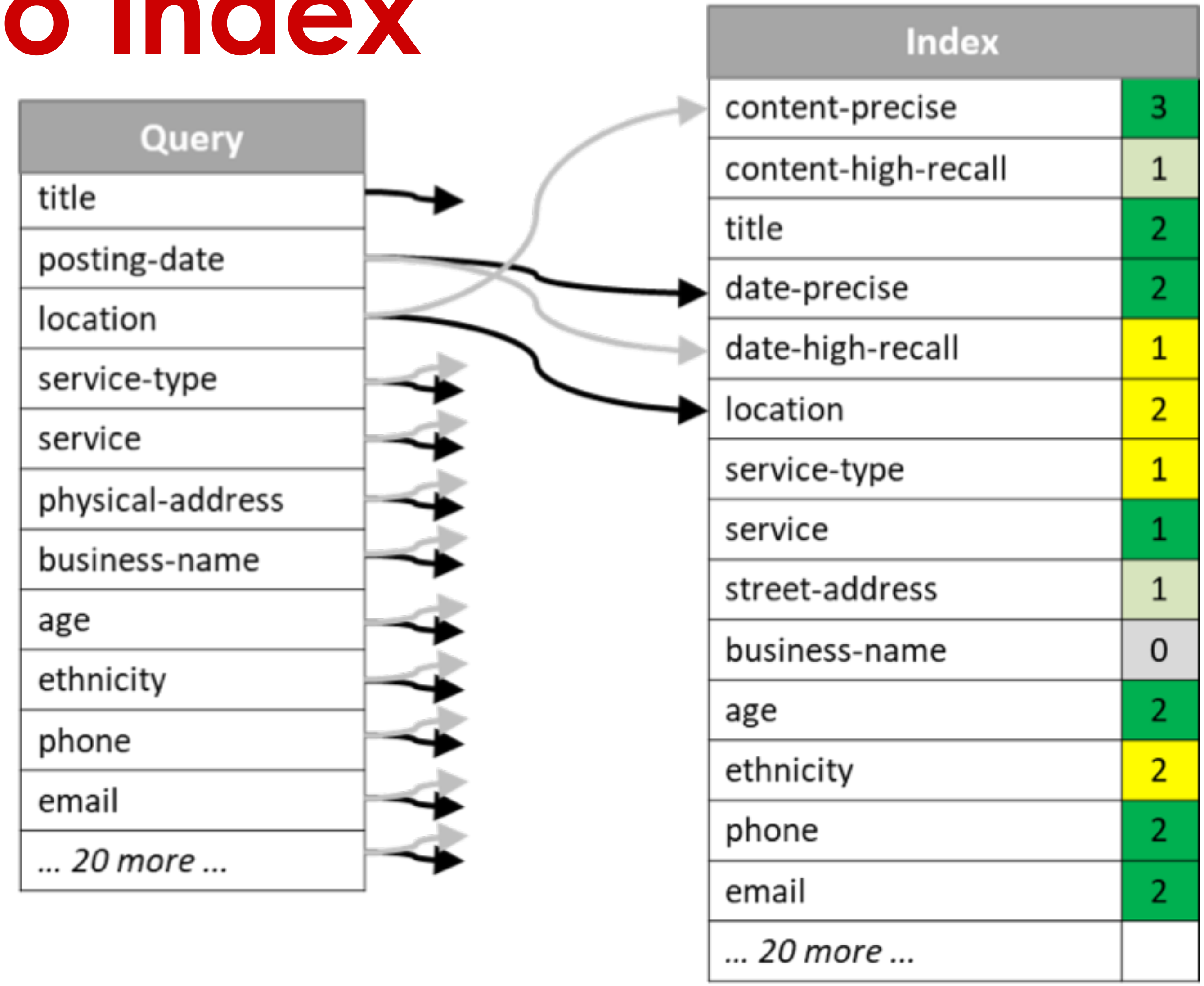
Candidate Generation

Keyword expansion • Context broadening • Constraint relaxation

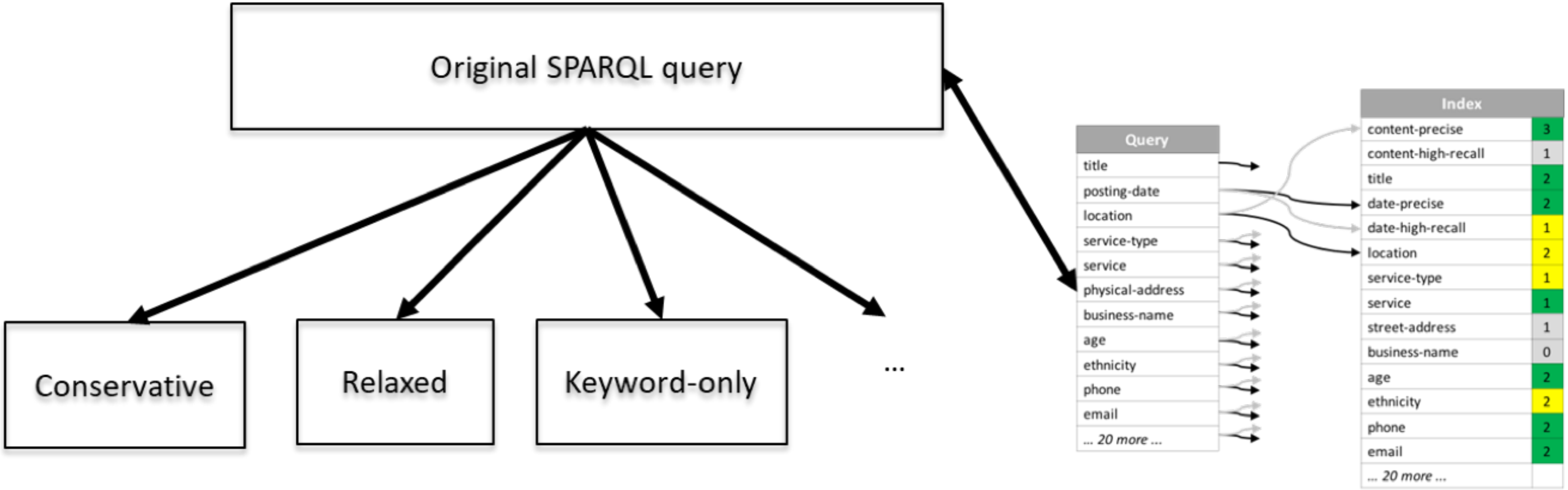
```
SELECT ?ad ?ethnicity WHERE
{
    ?ad a :Ad ;
        :hair_color 'Auburn' ;
        :review_site_id 'cg9469f' ;
        :price_per_hour '500' ;
        :name ’Claire Gold’ ;
        :ethnicity ?ethnicity .
}
```



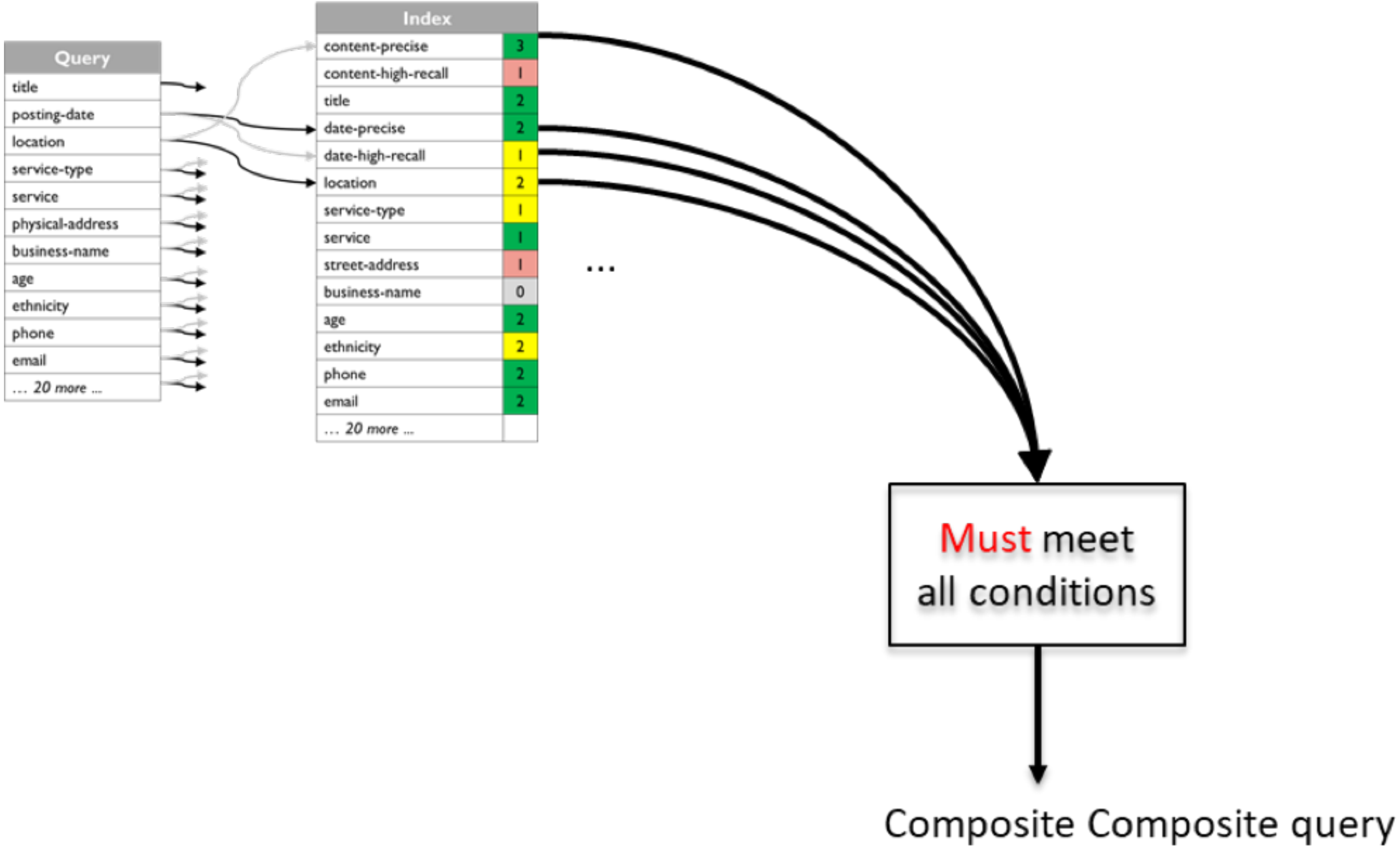
Offline step: Weighted Mapping Of Query To Index



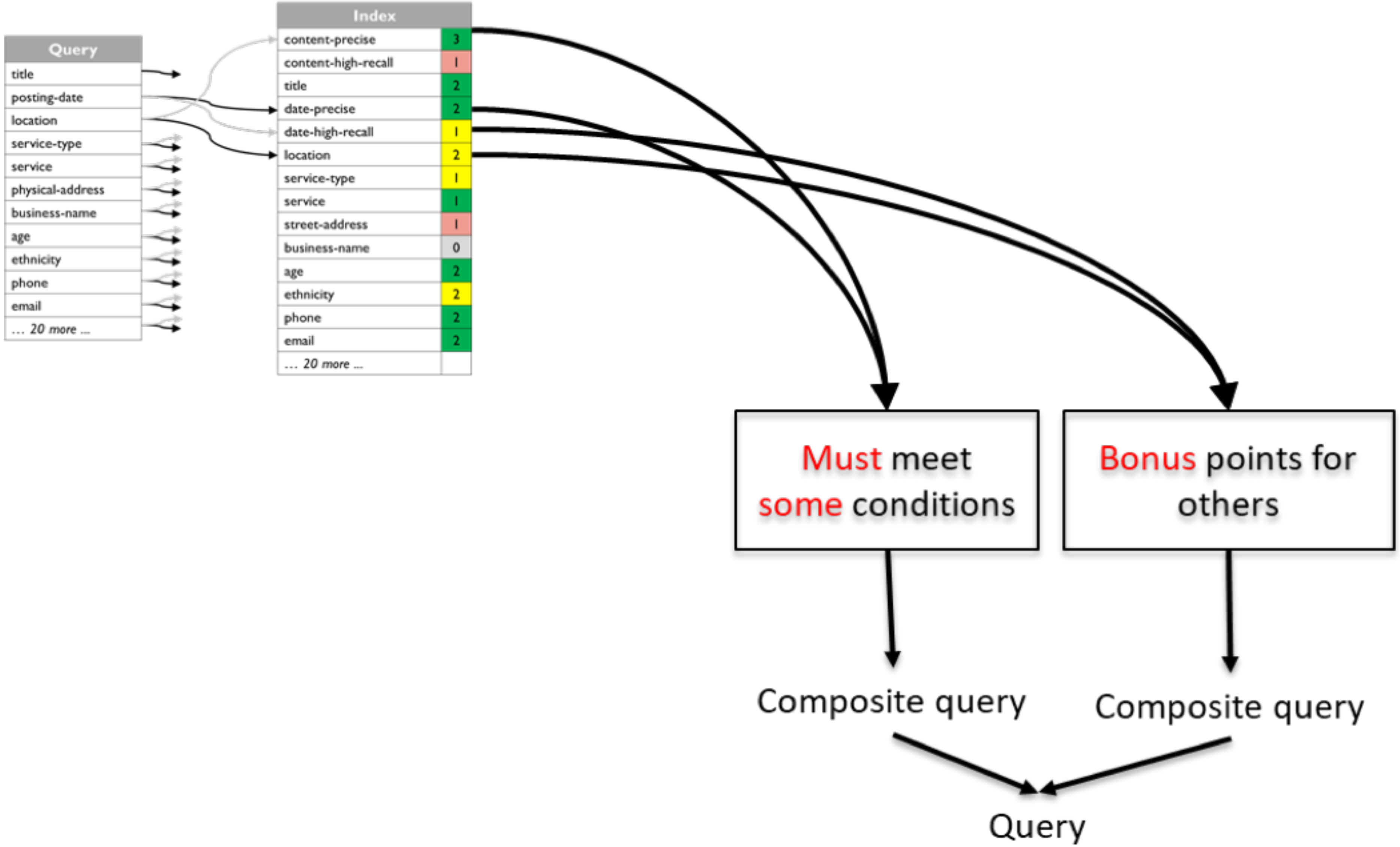
Online Step: Query reformulation using Semantic Strategies



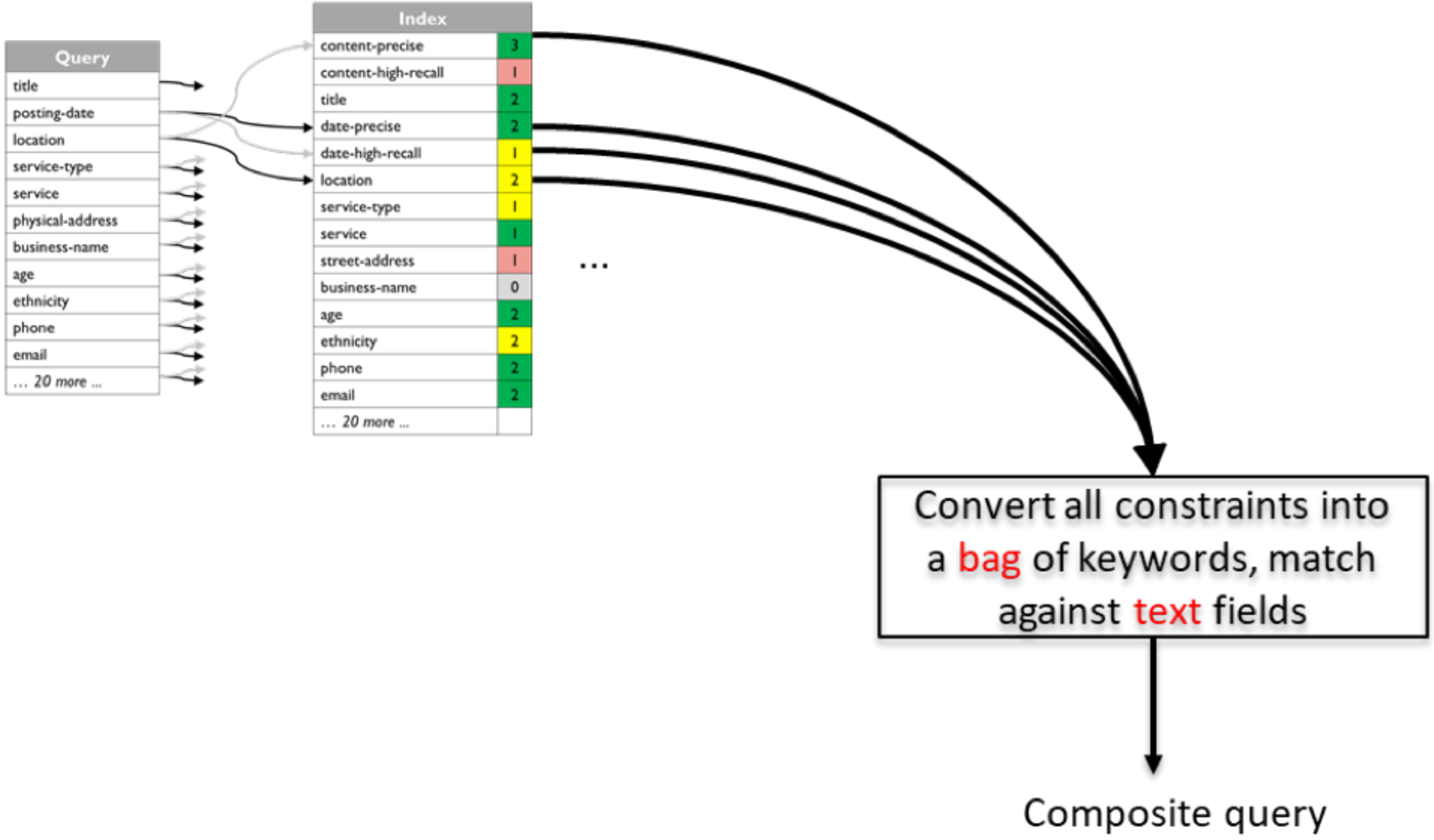
Conservative Query



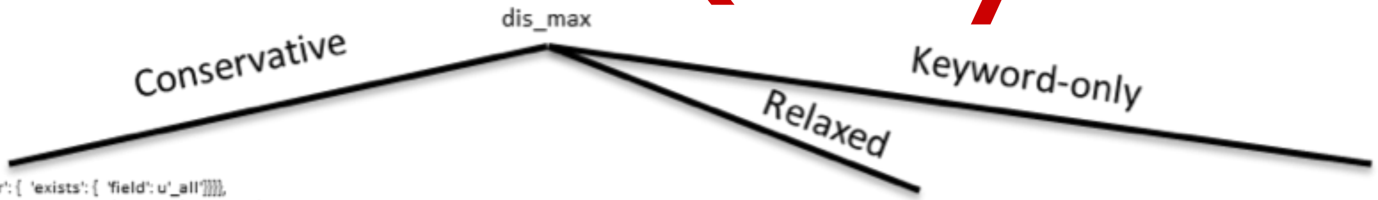
Relaxed Query



Keyword-only Query



Example of 'Final' Query



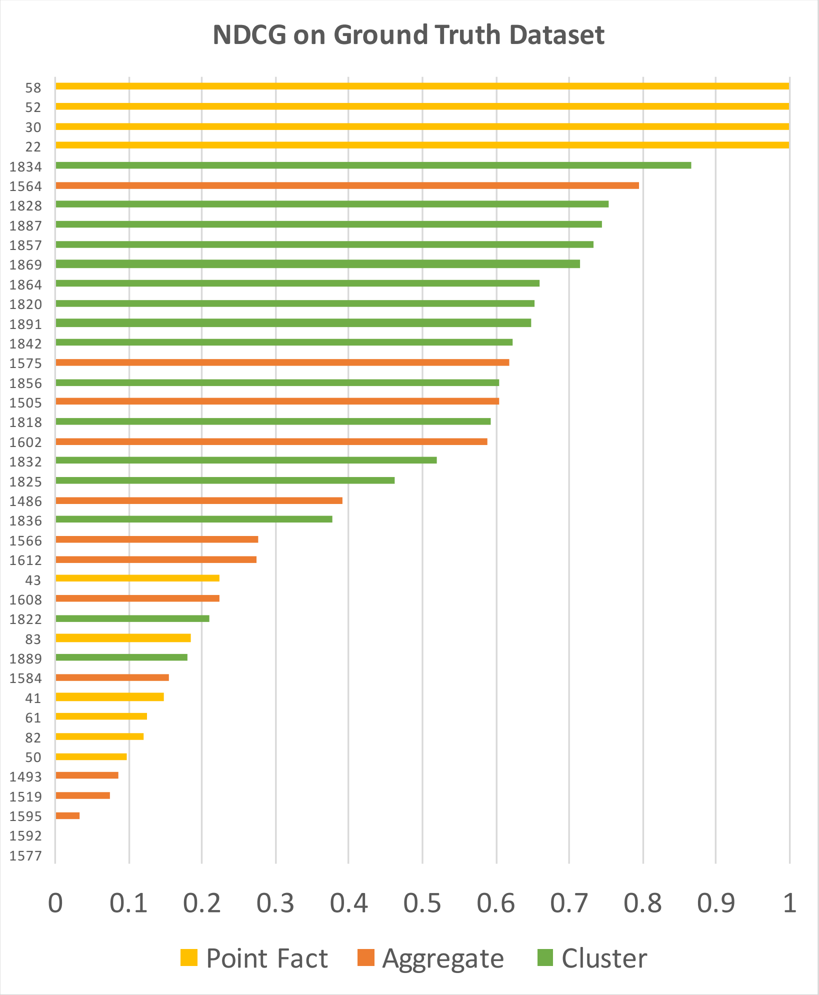
'bool': { 'filter': [{ 'constant_score': { 'filter': { 'exists': { 'field': u'_all'}}}},
{ 'constant_score': { 'filter': { 'exists': { 'field': u'url'}}}},
{ 'constant_score': { 'filter': { 'exists': { 'field': u'telephone.name'}}}},
{ 'constant_score': { 'filter': { 'exists': { 'field': 'identifier'}}}},
{ 'constant_score': { 'filter': { 'exists': { 'field': u'telephone.name.raw'}}}},
{ 'constant_score': { 'filter': { 'exists': { 'field': u'telephone_count'}}}}],
'must': [{ 'bool': { 'should': [{ 'match_phrase': { u'readability_date': u'2016-06-30'}},
{ 'match_phrase': { u'inferlink_date': u'2016-06-30'}},
{ 'match': { u'addressLocality': u'Bengaluru, India'}},
{ 'match': { u'name': u'Vishwas'}},
{ 'match': { u'name_count': 5}},
{ 'bool': { 'should': [{ 'match': { u'readability_text': u'Bengaluru, India'}},
{ 'match': { u'_all': u'Bengaluru, India'}},
{ 'match': { u'readability_text': u'Vishwas'}},
{ 'match': { u'_all': u'Vishwas'}}]}}]}}],
'should': [{ 'match_phrase': { u'readability_date': u'2016-06-30'}},
{ 'match_phrase': { u'inferlink_date': u'2016-06-30'}},
{ 'match': { u'addressLocality': u'Bengaluru, India'}},
{ 'match': { u'name': u'Vishwas'}},
{ 'match': { u'name_count': 5}},
{ 'bool': { 'should': [{ 'match': { u'readability_text': u'Bengaluru, India'}},
{ 'match': { u'_all': u'Bengaluru, India'}},
{ 'match': { u'readability_text': u'Vishwas'}},
{ 'match': { u'_all': u'Vishwas'}}]}}]}},

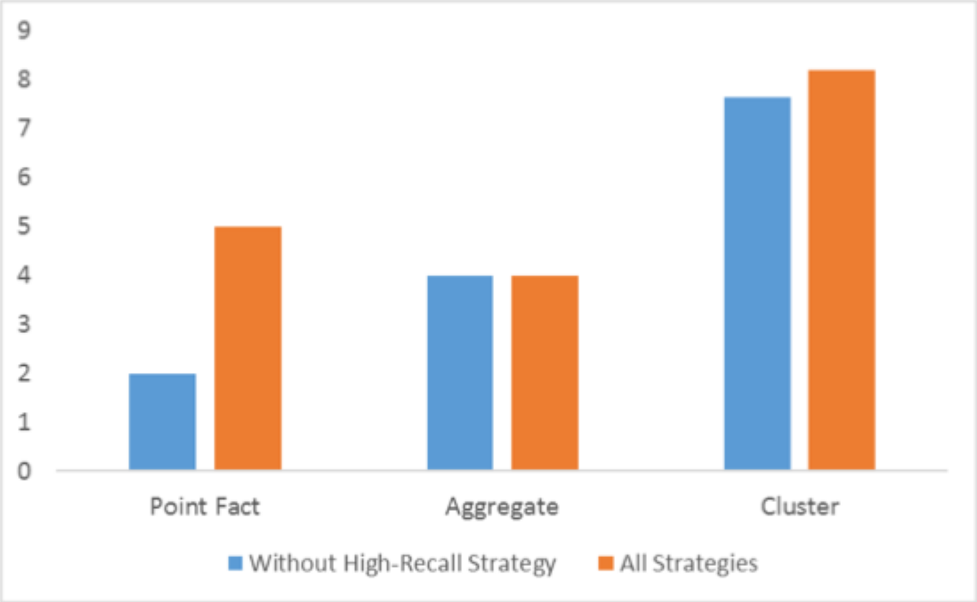
{ 'bool': { 'filter': [{ 'bool': { 'should': [{ 'exists': { 'field': u'_all'}},
{ 'exists': { 'field': u'url'}},
{ 'exists': { 'field': u'telephone.name'}},
{ 'exists': { 'field': 'identifier'}},
{ 'exists': { 'field': u'telephone.name.raw'}},
{ 'exists': { 'field': u'telephone_count'}}]}},
{ 'bool': { 'filter': [{ 'exists': { 'field': 'identifier'}}]}},
{ 'bool': { 'filter': [{ 'exists': { 'field': u'_all'}},
{ 'exists': { 'field': u'telephone.name'}},
{ 'exists': { 'field': u'url'}},
{ 'exists': { 'field': u'telephone.name.raw'}}]}},
{ 'bool': { 'filter': [{ 'exists': { 'field': u'telephone_count'}}]}}],
'must': [{ 'bool': { 'should': [{ 'match_phrase': { u'readability_date': u'2016-06-30'}},
{ 'match_phrase': { u'inferlink_date': u'2016-06-30'}},
{ 'match': { u'addressLocality': u'Bengaluru, India'}},
{ 'match': { u'name': u'Vishwas'}},
{ 'match': { u'name_count': 5}},
{ 'bool': { 'should': [{ 'match': { u'readability_text': u'Bengaluru, India'}},
{ 'match': { u'_all': u'Bengaluru, India'}},
{ 'match': { u'readability_text': u'Vishwas'}},
{ 'match': { u'_all': u'Vishwas'}}]}}]}}],
'should': [{ 'match_phrase': { u'readability_date': u'2016-06-30'}},
{ 'match_phrase': { u'inferlink_date': u'2016-06-30'}},
{ 'match': { u'addressLocality': u'Bengaluru, India'}},
{ 'match': { u'name': u'Vishwas'}},
{ 'match': { u'name_count': 5}},
{ 'bool': { 'should': [{ 'match': { u'readability_text': u'Bengaluru, India'}},
{ 'match': { u'_all': u'Bengaluru, India'}},
{ 'match': { u'readability_text': u'Vishwas'}},
{ 'match': { u'_all': u'Vishwas'}}]}}]}}

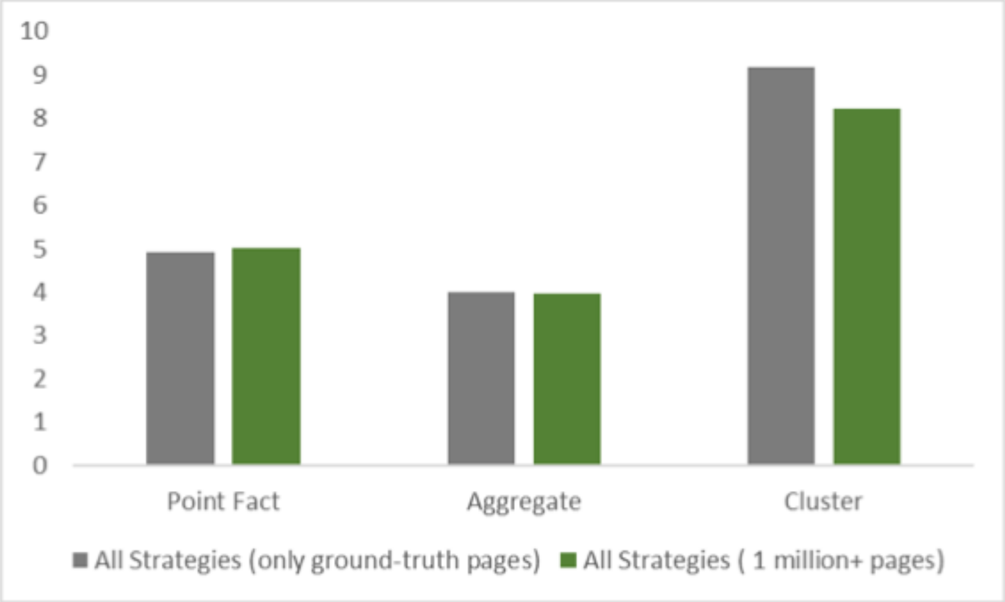
Example: query execution/ranking

name	hair color	price	review site id	ethnicity
Claire Gold	Auburn	500	cg9469f	?
Claire title/dict Rosa content/dict June content/dict	Red content/dict Black content/dict Auburn content/CRF	500 content/regex 400 content/regex 2016 content/regex		Asian content/dict Japanese content/dict Korean content/CRF
Clara content/dict June content/dict			cg9469f content/ES	Japanese content/dict
			cg9469f content/ES	Asian content/dict Japanese content/dict
Claire Gold content/ES	Auburn content/ES	150 title/regex 125 title/regex 100 content/regex		Caramel content/dict
…	…	…	…	…

Results


NDCG on Ground Truth Dataset
58
52
30
22
1834
1564
1828
1887
1857
1869
1864
1820
1891
1842
1575
1856
1505
1818
1602
1832
1825
1486
1836
1566
1612
43
1608
1822
83
1889
1584
41
61
82
50
1493
1519
1595
1592
1577
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Point Fact Aggregate Cluster


9
8
7
6
5
4
3
2
1
0
Point Fact
Aggregate
Cluster
Without High-Recall Strategy
All Strategies


10
9
8
7
6
5
4
3
2
1
0
Point Fact
Aggregate
Cluster
All Strategies (only ground-truth pages)
All Strategies (1 million+ pages)

myDIG: A KG Construction Toolkit

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

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

Docker deployment with single “docker compose up” installation