

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.

Non-grammatical snippets, rich formatting & links

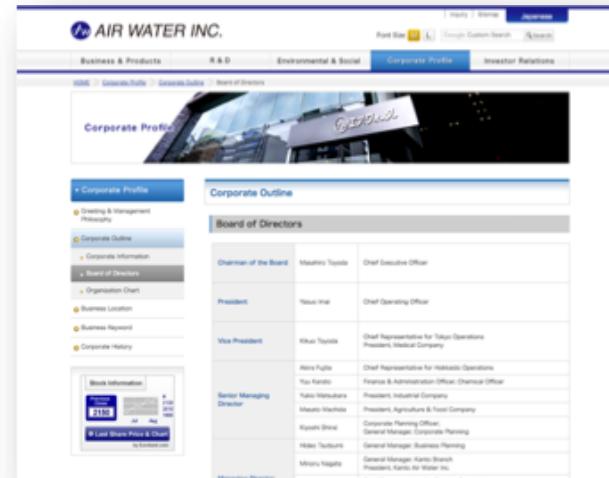
Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			 
Berger, Emery D.	(413) 577-4211	emory@cs.umass.edu	CS344
Assistant Professor.			 
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.			 

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

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"

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

Recent Escort Classifieds

- North Jersey, New Jersey (732-621-4443)
[*: GOOD GIRL *: GONE *: BAD ; \) LATINA - 21](#)
- Chicago, Illinois (773-412-2044)
[\(LAxE NIGHT\) UNRUSHEd \(ULTIMAtE\) PLEASURE \(*AmAZing Azz*\) CHOOSE..W...](#)
- Chicago, Illinois (414-914-3777)
[Petite, and Sweet. Super new and Ready... in out call -](#)
- Chicago, Illinois (312-600-8628)
[\(LAxE NIGHT\) UNRUSHEd \(ULTIMAtE\) PLEASURE \(*AmAZing Azz*\) CHOOSE..W...](#)
- Ft Lauderdale, Florida (954-518-1860)
[you'll M... GEoGraPhy - 21](#)
- Atlanta, Georgia (404-244-9386)
[WoW, MuSt TaKe A LoOk At ThIs. - 21](#)
- Atlanta, Georgia (347-940-1982)
[SMOKING HOT! Specials BuSTy Ba\(\(5 SeRviCe \)\) Pretty 36DDDs \(\)\(...](#)
- Atlanta, Georgia (404-244-8142)
[here's a casual outfit in AL, ME 100% real. White - 21](#)
- Houston, Texas (281-444-6660)
[Beautiful Salvadorean The One And Only\(- - 21](#)
- Phoenix, Arizona (623-500-7076)
[NEW GIRL PERSIAN Gem EXotIC Blend - 21](#)
- Toronto, Ontario (416-554-3337)
[\(L\)\(L\) ---Special 80 for 20 min:\) 22YeAr of d \\$Sexyy LaTiNa BoMbSheLL---\(L...](#)
- Toronto, Ontario (416-520-5198)
[**21 years old * \\$80 **real pictures ** A sian Kathy *** - 21](#)
- Toronto, Ontario (647-702-6825)
[2016-10-14 - 01:14:54 - 100% real - 21](#)

Top Escort Cities

- [New York, New York](#)
- [Toronto, Ontario](#)
- [Dallas, Texas](#)
- [Chicago, Illinois](#)
- [Atlanta, Georgia](#)
- [North Jersey, New Jersey](#)
- [Detroit, Michigan](#)
- [Austin, Texas](#)
- [Orlando, Florida](#)
- [Cincinnati, Ohio](#)
- [Houston, Texas](#)
- [Phoenix, Arizona](#)
- [Philadelphia, Pennsylvania](#)
- [Boston, Massachusetts](#)
- [Washington, DC](#)
- [Portland, Oregon](#)
- [Las Vegas, Nevada](#)
- [Miami, Florida](#)

Recent Blog Posts

- [Sheriff candidate Minister and Detective Reno Fell 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

Spreadsheets Created For Human Consumption

The image displays four separate screenshots of spreadsheet software interfaces, likely Microsoft Excel, demonstrating different types of data handling:

- Screenshot 1:** A spreadsheet titled "Nutzungsbestimmungen" showing survey results. It includes a header row and two data columns labeled "gut" and "schlecht". Rows 10 through 20 contain data points with dates from 1/17/14 to 8/22/14.
- Screenshot 2:** A screenshot of a search results page for "Crude Oil Production Qatar Monthly" from the U.S. Energy Information Administration. It shows monthly production data from December 2012 to May 2013.
- Screenshot 3:** A screenshot of a table showing refugee data extracted from the United Nations High Commissioner for Refugees. It lists countries (Albania) and months (March to September 2017) with corresponding values.
- Screenshot 4:** A screenshot of a spreadsheet titled "imfCPI_example" showing "Cross-Country Indexes, Period-over-Period Change". It lists countries and their CPI index values for specific periods (e.g., 2002M11, 2002M12, 2003Q1, 2003M01, 2003M02).

Databases with PDF Code Books

event_date	year	time_precision	event_type	actor1	assoc_actor_inter1	actor2	assoc_actor_inter2	interaction	region	country	admin1
1/13/18	2018	1	Battle-No ch	Military Forces of Democ	1 ADF: Allied Democratic Fc	2	12 Central Afric Democratic	I Nord-Ki			
1/13/18	2018	1	Battle-No ch	Military Forces of Democ	1 ADF: Allied Democratic Fc	2	12 Central Afric Democratic	I Nord-Ki			
1/13/18	2018	1	Battle-No ch	Military Forces of Democ	1 ADF: Allied Democratic Fc	2	12 Central Afric Democratic	I Nord-Ki			
1/13/18	2018	1	Battle-No ch	Al Shabaab	2 Police Forces of Kenya (2C	1	12 Eastern Afric Kenya	Lamu			
1/13/18	2018	1	Riots/Protest	Rioters (Kenya)	5 Police Forces of Kenya (2C	1	15 Eastern Afric Kenya	Marsab			
1/13/18	2018	1	Riots/Protest	Protesters (L Tawergha Co	6	0	60 Northern Afr Libya	Sabha			
1/13/18	2018	1	Violence aga	Unidentified Armed Group	3 Civilians (Int'l) OIM: Interna	7	37 Northern Afr Libya	Sabha			
1/13/18	2018	1	Riots/Protest	Protesters (Morocco)	6	0	60 Northern Afr Morocco	Oriental			
1/13/18	2018	1	Riots/Protest	Protesters (Nigeria)	6	0	60 Western Afr Nigeria	Benue			
1/13/18	2018	1	V	Civilians = 7				17 Western Afr Nigeria	Kaduna		
1/13/18	2018	2	R					50 Western Afr Nigeria	Lagos		
1/13/18	2018	1	B	Outside/external force (e.g. UN) =8				13 Eastern Afric Somalia	Banaad		
1/13/18	2018	1	B					12 Eastern Afric Somalia	Shabee		
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



The screenshot shows the official website of the United Nations Security Council. The header features the UN logo and the text "UNITED NATIONS SECURITY COUNCIL". Below the header, there is a navigation bar with links: ABOUT, PRESIDENCY, MEMBERS, PROGRAMME OF WORK, DOCUMENTS (which is the active tab), MEETINGS, SUBSIDIARY ORGANS, 2231 (2015), and REPERTOIRE. A large photograph of the Security Council meeting room is displayed above the navigation bar. On the left side, there is a sidebar with various links related to Security Council documents and practices.

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				



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

InvestorPlace

DOW 22,049 ▼ -0.17% NASDAQ 6,352 0.00% S&P 500 2,474 ▼ -0.04%

NAVELLIER RATINGS
Powered by Portfolio Grader

LEGACY VENTURES INTL, Inc. (LGYV) Powered by Financial Content

6.01 0.00 (0.00%) 07/14/17

LGYV STOCK QUOTE DELAYED 20 MINUTES

THE WAY TRADING SHOULD ALWAYS BE  GET 500 FREE TRADES

Identify opportunities with innovative tools & visualized research
Now just \$4.95 for online U.S. equity trades — lower than TD Ameritrade and E*Trade

Fidelity Investments

Read additional disclosure
Fidelity Brokerage Services, Member
NYSE, SIPC.
© 2017 FMR LLC.
All rights reserved.
785101.3.0

LGYV Stock Chart Historical LGYV Prices

LEGACY VENTURES INTL, Inc. (LGYV)
1 Feb 2017 - 14 Jul 2017



Dividend & Yield: N/A (N/A)

P/E:

Market Cap: 391.03K

EPS: -38648.00

Volume: 67

Day's Range: 6.01 - 6.01

52wk Range: 1.05 - 15.00

Intraday 3 Month 6 Month 1 Year

LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS

NEW RATE

Data Extraction

Information Extraction Process

Segmentation

The screenshot shows the InvestorPlace website interface. At the top, there's a navigation bar with links for Stocks, Funds, Retirement, Trading, Market Insight, Financial Advisors, and Premium Services. Below the bar, market indices are displayed: DOW 22,049 (▼ -0.17%), NASDAQ 6,352 (0.00%), and S&P 500 2,474 (▼ -0.04%). A search bar is also present.

The main content area features a large red box highlighting the stock information for LEGACY VENTURES INTL, Inc. (LGYV). It shows the current price of 6.01, the previous close of 0.00, and the change of 0.00% (0.00%) from the previous day (07/14/17).

Below this, a banner for Fidelity Investments promotes "THE WAY TRADING SHOULD ALWAYS BE" with a call to action "GET 500 FREE TRADES".

On the left side, there's a sidebar titled "LEGACY VENTURES INTL, Inc. Stock Analysis" which includes a "Rating:" section and a "Total Grade:". Further down, there's an "Analysis Breakdown" section with various metrics like Earnings Growth, Earnings Momentum, and Earnings Surprises.

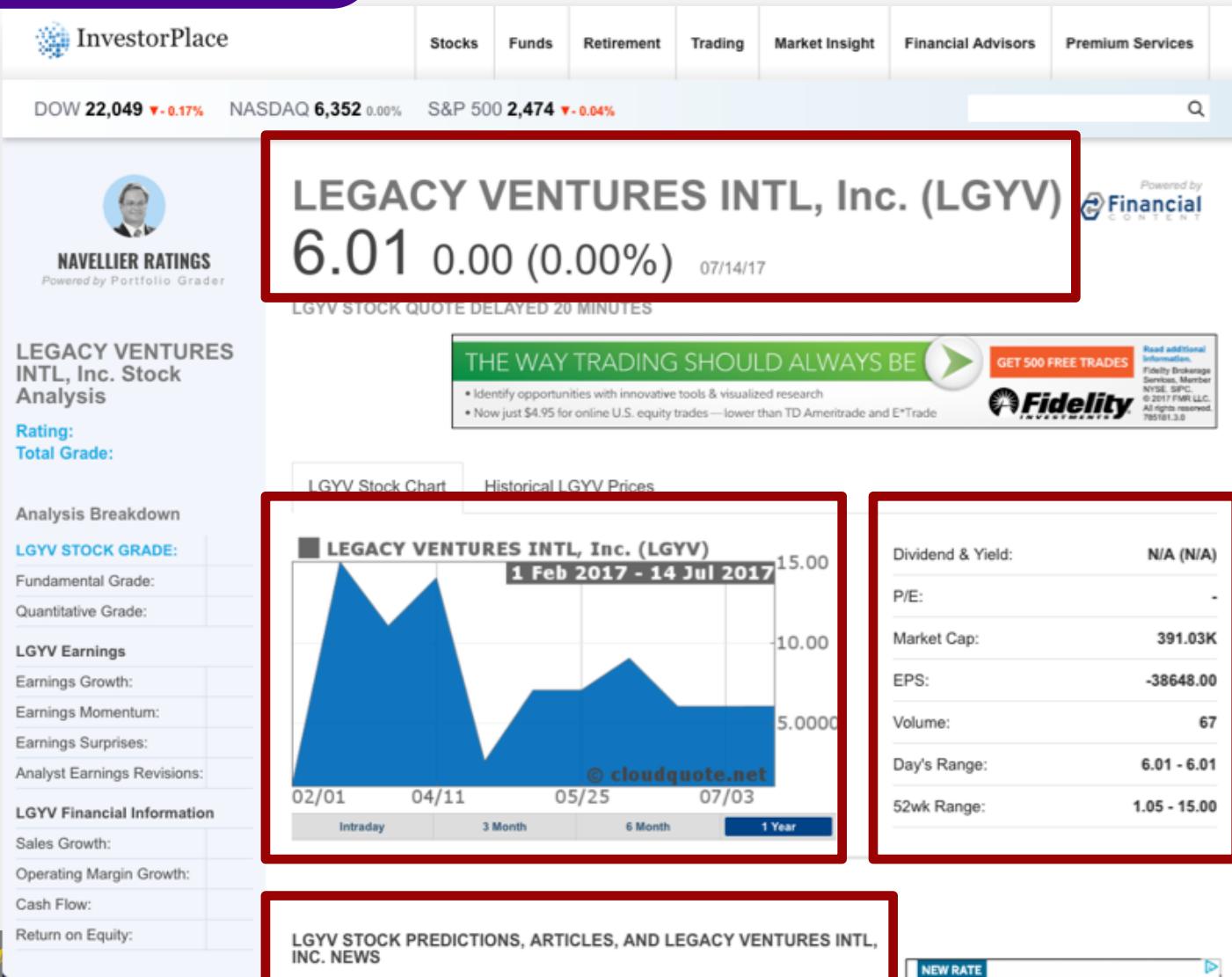
The central part of the page contains two red boxes: one showing a "LGYV Stock Chart" with a line graph from 1 Feb 2017 to 14 Jul 2017, and another showing "Historical LGYV Prices" with a table of financial data including Dividend & Yield, P/E, Market Cap, EPS, Volume, Day's Range, and 52wk Range.

At the bottom, there's a section for "LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS" with a "NEW RATE" button.

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

The screenshot shows a detailed stock analysis page for LGYV. At the top, there's a navigation bar with links for Stocks, Funds, Retirement, Trading, Market Insight, Financial Advisors, and Premium Services. Below that, the current market values are displayed: DOW 22,049 (down 0.17%), NASDAQ 6,352 (0.00%), and S&P 500 2,474 (down 0.04%). A search bar is also present.

The main content area features a large box for "LEGACY VENTURES INTL, Inc. (LGYV)" with a rating of 6.01 and a price of 0.00 (0.00%) from 07/14/17. This box is highlighted with a red border. To the right of this box is a "Powered by Fidelity Content" logo.

Below this, there's a banner for "THE WAY TRADING SHOULD ALWAYS BE" from Fidelity Investments, offering 500 free trades. The page also includes a "LGYV STOCK QUOTE DELAYED 20 MINUTES" section with a chart and historical prices, and a "LGYV Stock Chart" section with a line graph from 1 Feb 2017 to 14 Jul 2017.

The sidebar on the left contains various analysis breakdowns and financial information, such as "NAVELLIER RATINGS" (Powered by Portfolio Grader), "LGYV STOCK GRADE", "Fundamental Grade", "Quantitative Grade", "LGYV Earnings", "Earnings Growth", "Earnings Momentum", "Earnings Surprises", "Analyst Earnings Revisions", "LGYV Financial Information", "Sales Growth", "Operating Margin Growth", "Cash Flow", and "Return on Equity".

At the bottom, there's a section for "LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS" with a "NEW RATE" button.

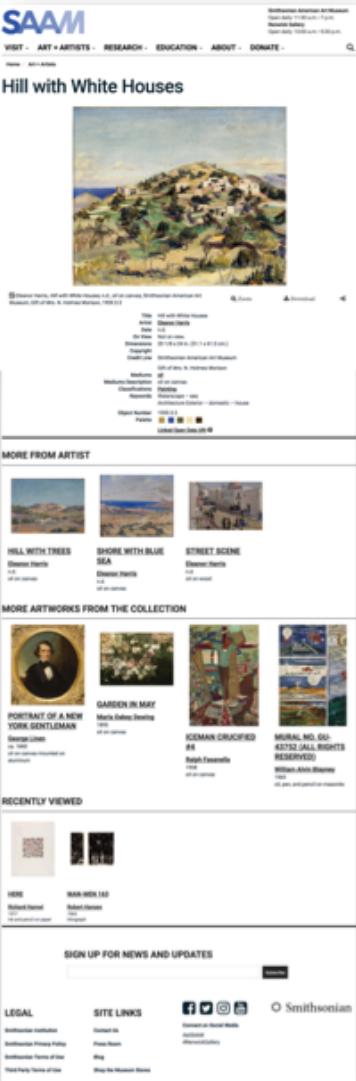


Homogeneous
blocks

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



You have selected: museum_demo_--si.edu-20171024_055146 Cluster[1]					
<input checked="" type="radio"/> Passed Classification (93 of 100 pages) <input type="radio"/> Failed Classification (7 of 100 pages)					
Page	Cluster	title	name	date_creation	birth_info
page1	1	Here	Richard Hamwi	1977	New York, New York 1947
		CACHED PAGE			
more					
page2	1	Man-Men 162	Robert Hansen	1965	Osceola, Nebraska 1924
		CACHED PAGE			
more					
page4	1	Hill with White Houses	Eleanor Harris	n.d.	1901 <span title="artist death place and d ..."
		CACHED PAGE			
more					
page5	1	Portrait	DeWitt Hardy		St. Louis, Missouri 1940
		CACHED PAGE			

myDIG Demo

Focusing On Inferlink Web Wrapper

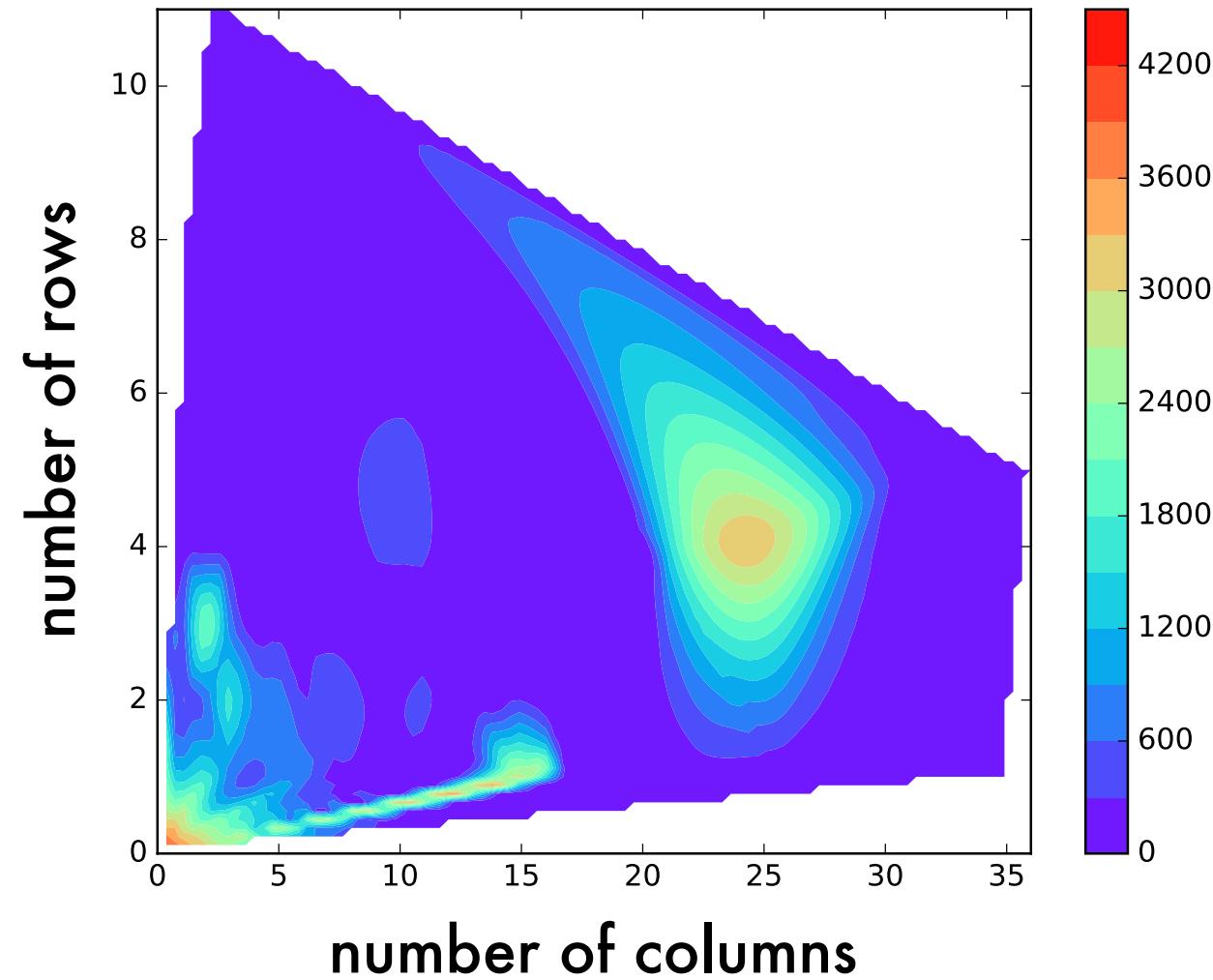
Table Extraction

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, 2003–04, 2004–05, 2005–06, 2006–07, 2007–08, 2008–09, 2009–10, 2010–11, 2011–12, 2012–13, 2013–14, 2014–15 FA Cup winners: 1997–98, 2001–02, 2003–04, 2004–05, 2005–06, 2006–07, 2007–08, 2008–09, 2009–10, 2010–11, 2011–12, 2012–13, 2013–14, 2014–15, 2015–16 15, 2016–17 Charity/Community Shield winners: 1998, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 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, 1990–91 FA Cup winners: 1992–93 Football League Cup winners: 1986–87, 1987–88 Charity Shield winners: 1991 (shared) UEFA Cup Winners' Cup winners: 1993–94
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

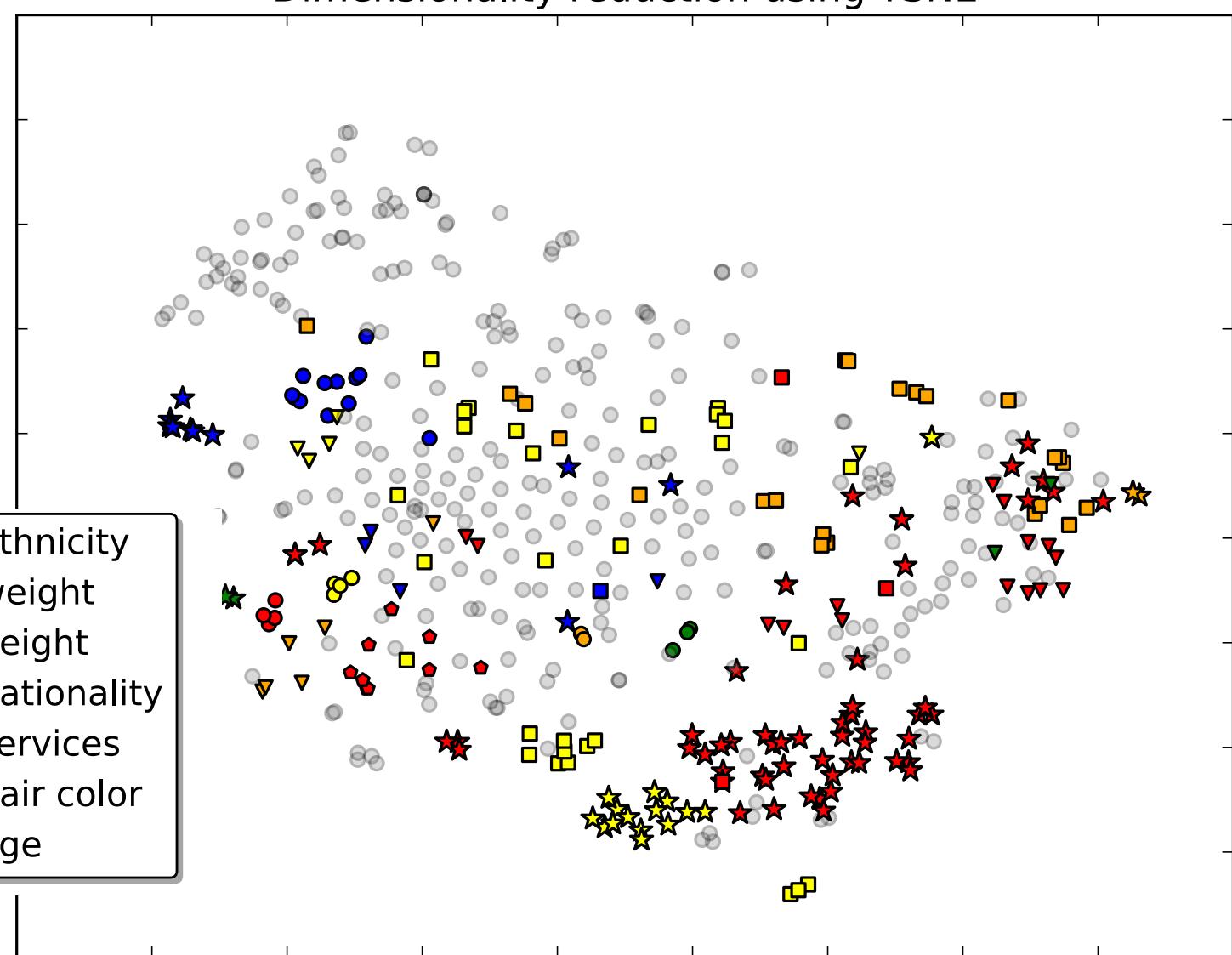
Heuristic

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

★★★ telephone	■■■ measurement	▼▼▼ ethnicity
●●● other	●●● web	★★★ weight
★★★ price	★★★ clients	■■■ height
■■■ location	▼▼▼ date	▼▼▼ nationality
○○○ language	●●● orientation	■■■ services
▼▼▼ email	●●● name	●●● hair color
○○○ gender	★★★ eye color	▼▼▼ age



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:

```
^(?:(?:\+\d{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\d{1} | ext\d{1} | extension)\s*(\d+))?)$
```

Brittle and difficult to adapt to specific domains

unusual nomenclature and short-hands

obfuscation

NLP Rule-Based Extraction

GET STARTED

Installation

Models

Lightning tour

Command line

Troubleshooting

Resources

WORKFLOWS

Loading the pipeline

Processing text

spaCy's data model

POS tagging

Using the parse

Entity recognition

Custom pipelines

Rule-based matching

Word vectors

Deep learning

Custom tokenization

Adding languages

Training

Training NER

Saving & loading

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}, {LOWER: "world"}])
```



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 conjunction
- pronoun verb
- proper noun pre/post-position
- determiner adverb
- symbol particle
- adjective 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 conjunction
- pronoun verb
- proper noun pre/post-position
- determiner adverb
- symbol particle
- adjective 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:

<input type="checkbox"/> ,	<input type="checkbox"/> !	<input type="checkbox"/> <
<input type="checkbox"/> .	<input type="checkbox"/> (<input type="checkbox"/> >
<input type="checkbox"/> ;	<input type="checkbox"/>)	<input type="checkbox"/> =
<input type="checkbox"/> ?	<input type="checkbox"/> [<input type="checkbox"/> %
<input type="checkbox"/> ~	<input type="checkbox"/>]	<input type="checkbox"/> \
<input type="checkbox"/> :	<input type="checkbox"/> {	<input type="checkbox"/> /
<input type="checkbox"/> "	<input type="checkbox"/> }	<input type="checkbox"/> *
<input type="checkbox"/> '	<input type="checkbox"/>	<input type="checkbox"/> \$
<input type="checkbox"/> +	<input type="checkbox"/> -	<input type="checkbox"/> @
<input type="checkbox"/> _	<input type="checkbox"/> ^	
<input type="checkbox"/> p.	<input type="checkbox"/> #	

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

1. By name

The DIG Demo interface displays three stages of word token modification:

- Stage 1:** A row of tokens: "W", "x", "By", "+", "t.", and "r". The "By" token is highlighted with a red circle containing an "X".
- Stage 2:** The "By" token has been removed, resulting in tokens: "W", "x", "+", "t.", and "r". The "t." token is highlighted with a red circle containing a plus sign.
- Stage 3:** The "t." token has been modified, resulting in tokens: "W", "x", "+", "t.", and "r". The "t." token is highlighted with a red circle containing a plus sign.

Output format: {1} {2}

Modify Word Token

optional part of output followed by space

Words:

Part of speech:

- noun
- conjunction
- pronoun
- verb
- proper noun
- pre/post-position
- determiner
- adverb
- symbol
- particle
- adjective
- interjection

Capitalization:

- exact
- lower
- upper
- title
- mixed

Length 1: Length 2: Length 3:

Prefix: Suffix: not in vocabulary

cancel **Save**

DIG Demo

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>

GET STARTED

Installation

Models

Lightning tour

Command line

Troubleshooting

Resources

WORKFLOWS

Loading the pipeline

Processing text

spaCy's data model

POS tagging

Using the parse

Entity recognition

Custom pipelines

Rule-based matching

Word vectors

Deep learning

Custom tokenization

Adding languages

Training

Training NER

Saving & loading

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.

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_)
```

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
```



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

displaCy Named Entity Visualizer

View on GitHub

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

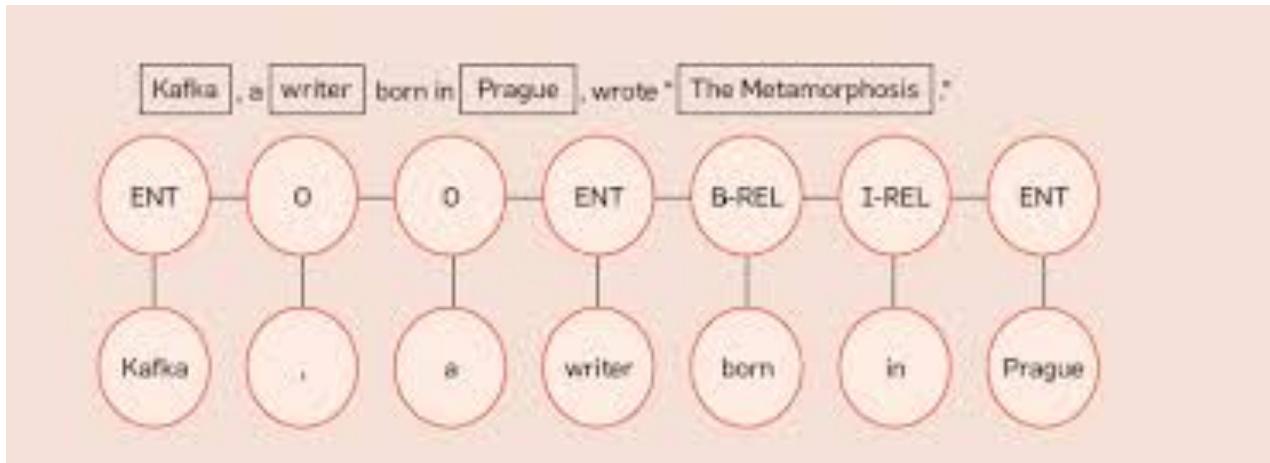
Performance degrades rapidly for new genres, language models

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	I	Possessive Pron	Other
2	name	0	Noun	Other
3	is	0	Verb	Other
4	Pedro	I	Proper Noun	Person-Name
5	Szekely	I	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



Open Information Extraction

Hosted by



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:

what/who

Relation:

verb phrase

Argument 2:

what/who

Corpus:

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 $f(\# \text{ regex})$	medium $f(\# \text{ rules})$	high	medium	medium	high
Coverage	wide	wide	wide	single site	genre	genre	narrow

how to represent KGs?

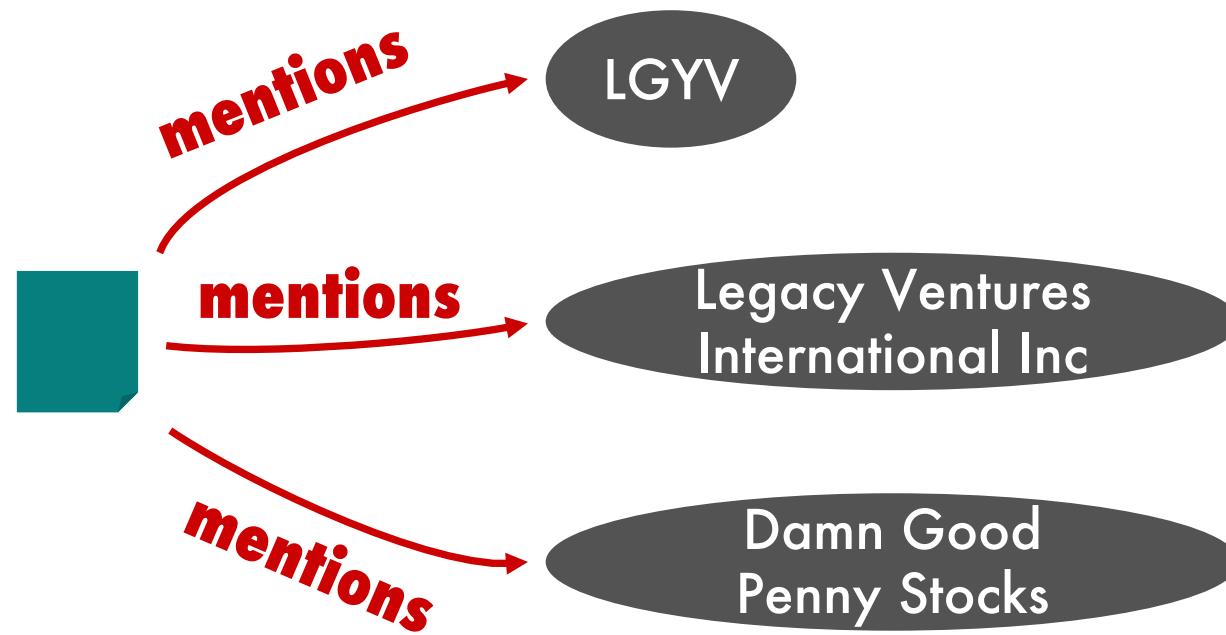
KG Definition

**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

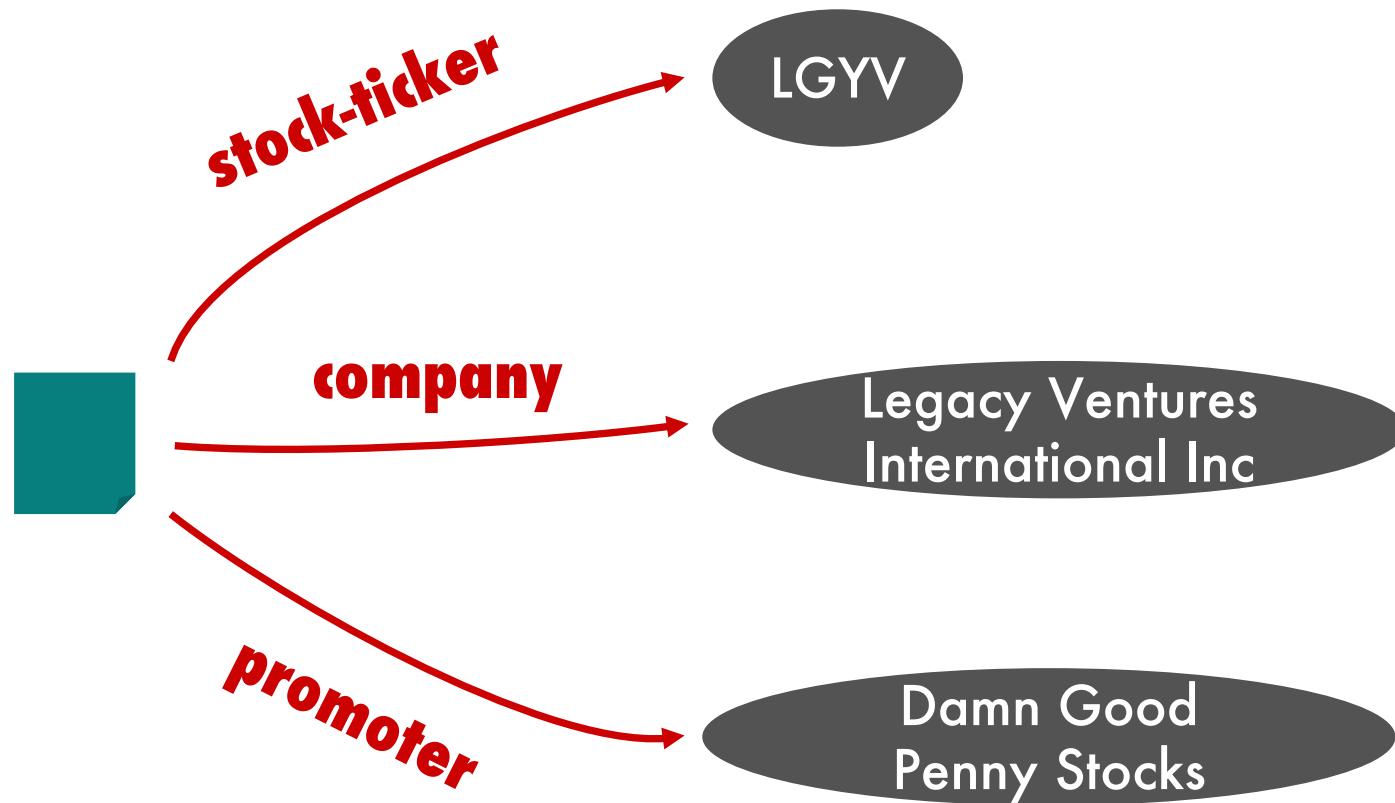
Entities



Easiest to build

Simple, But Useful KG

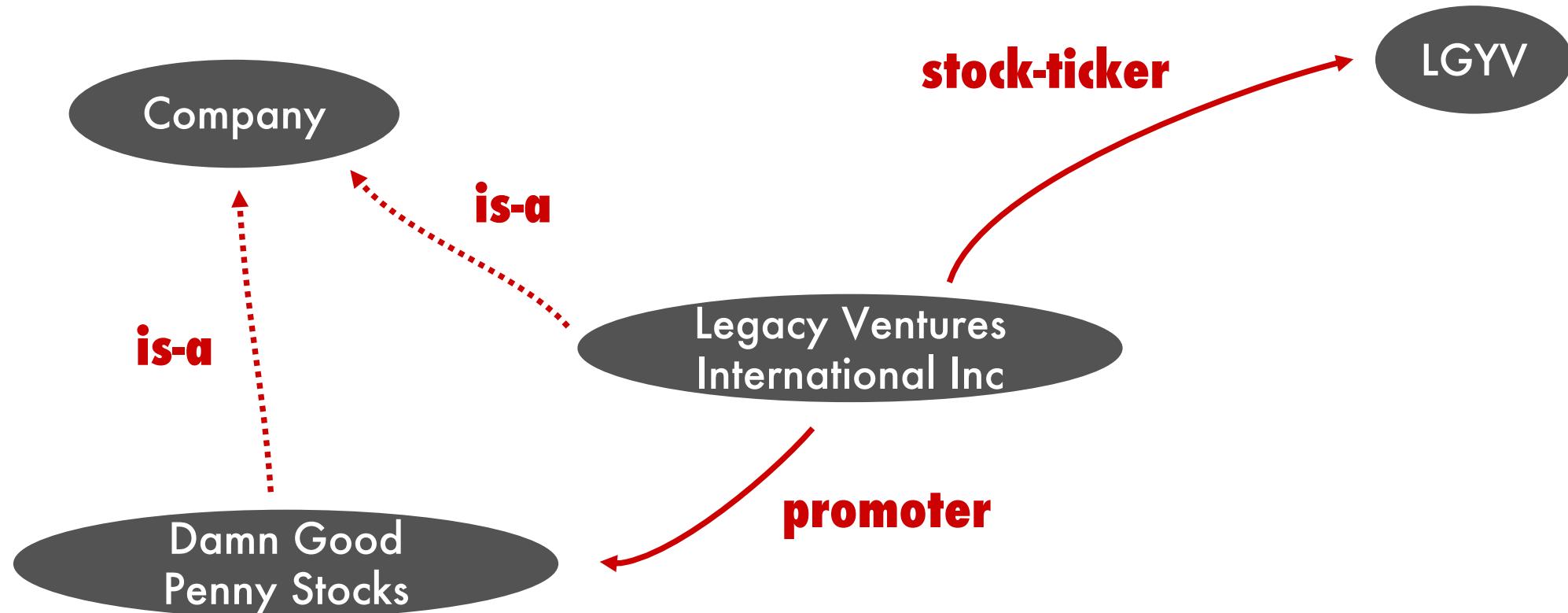
Entities + properties



"Easy" to build

Semantic Web KG (RDF/OWL)

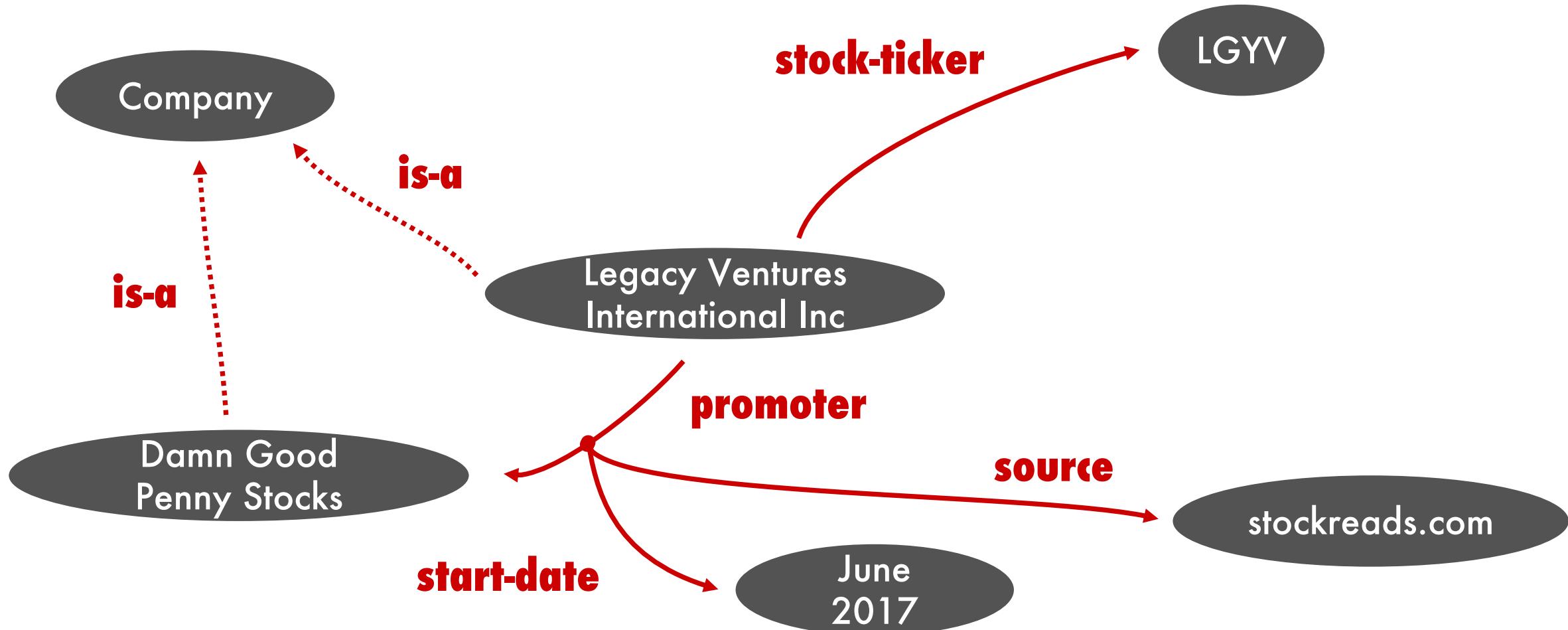
Entities + properties + classes



Very hard to build

“Ideal” KG

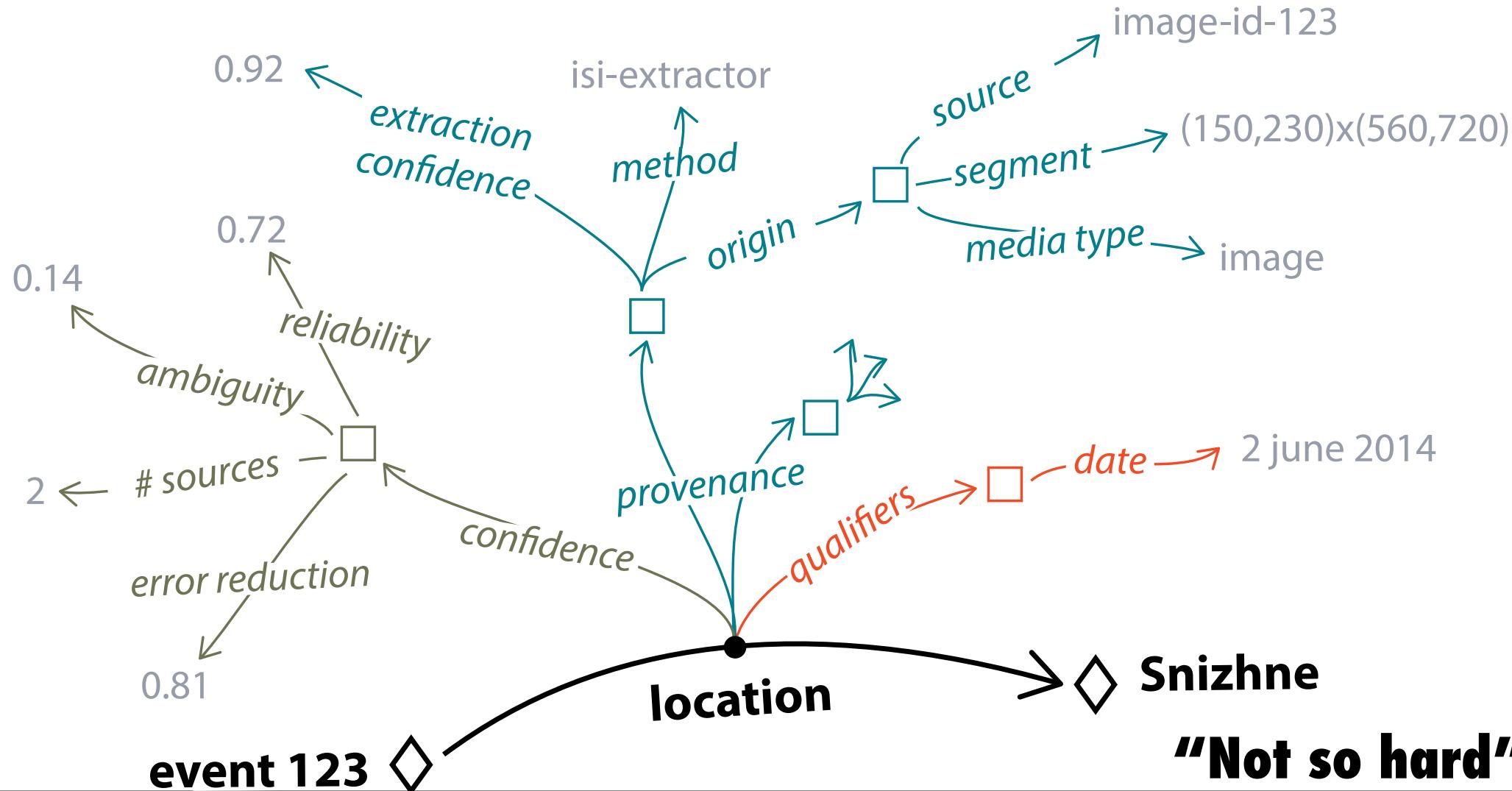
Entities + properties + classes + qualifiers



Very very hard to build

Semi-Structured KG

Entities + properties + text + provenance + confidence



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

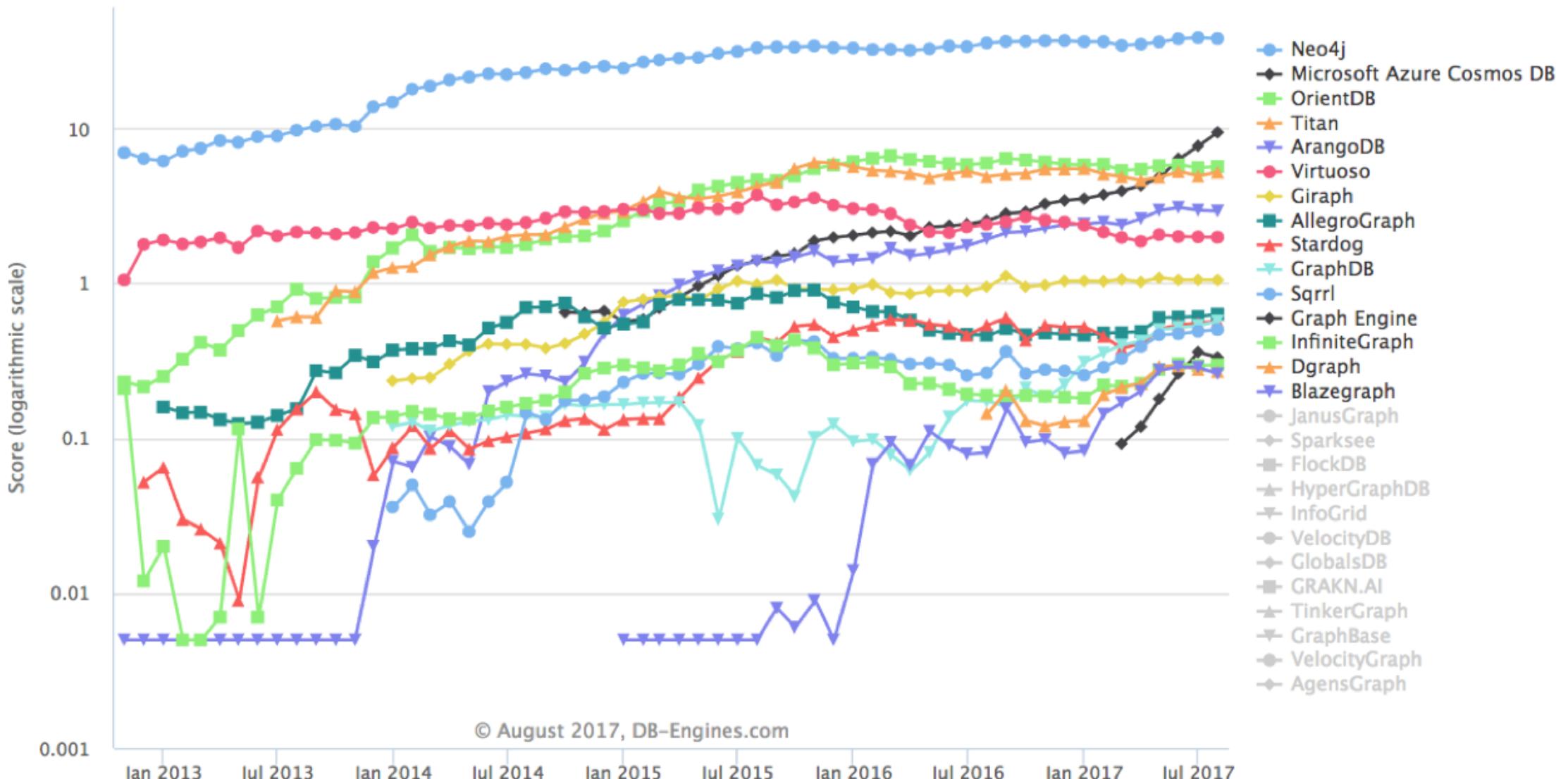
Data model: graph

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

Query: GraphQL, Gremlin, Cypher

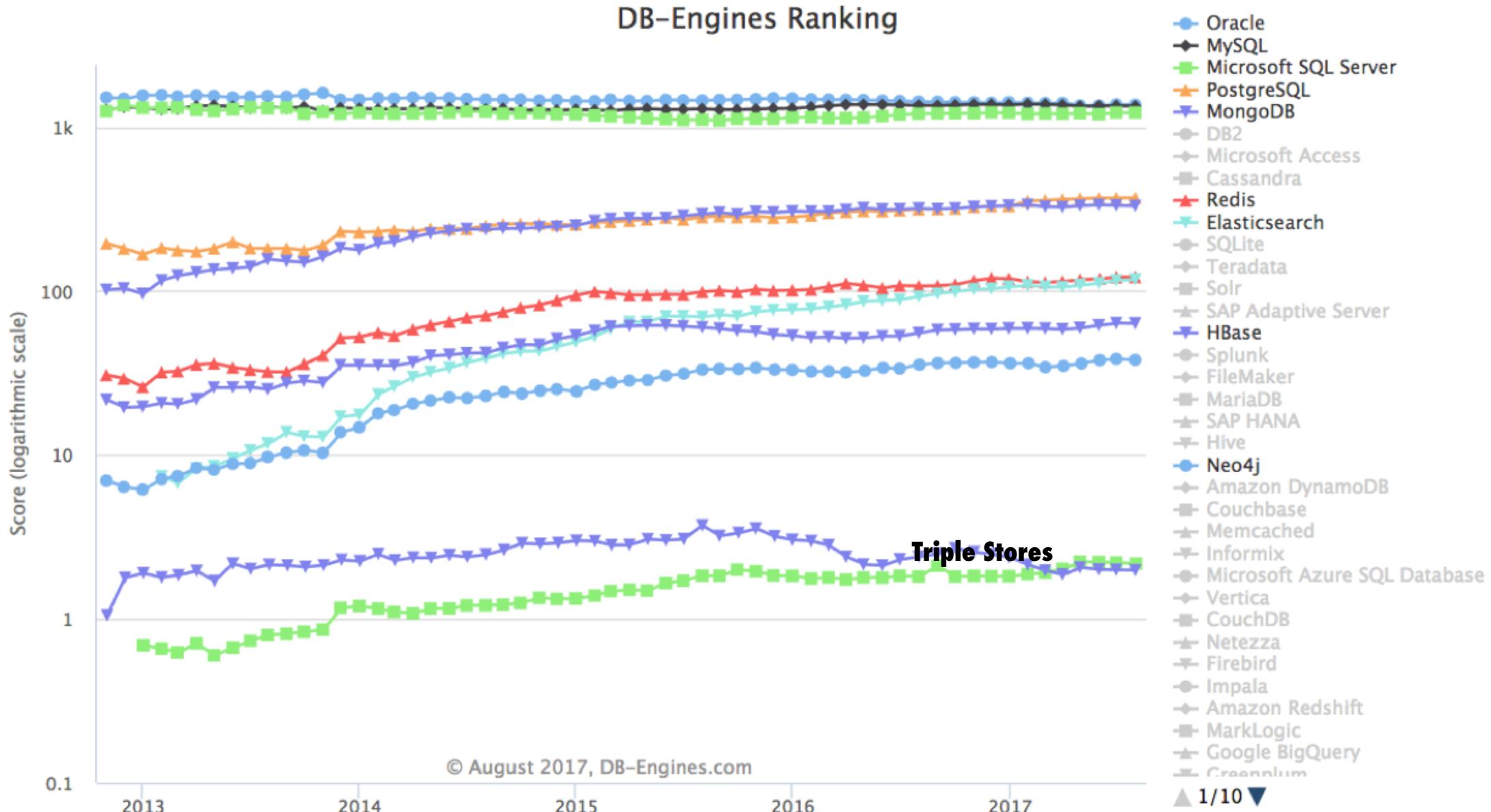
Popularity Ranking Of Graph Databases

DB-Engines Ranking of Graph DBMS



© August 2017, DB-Engines.com

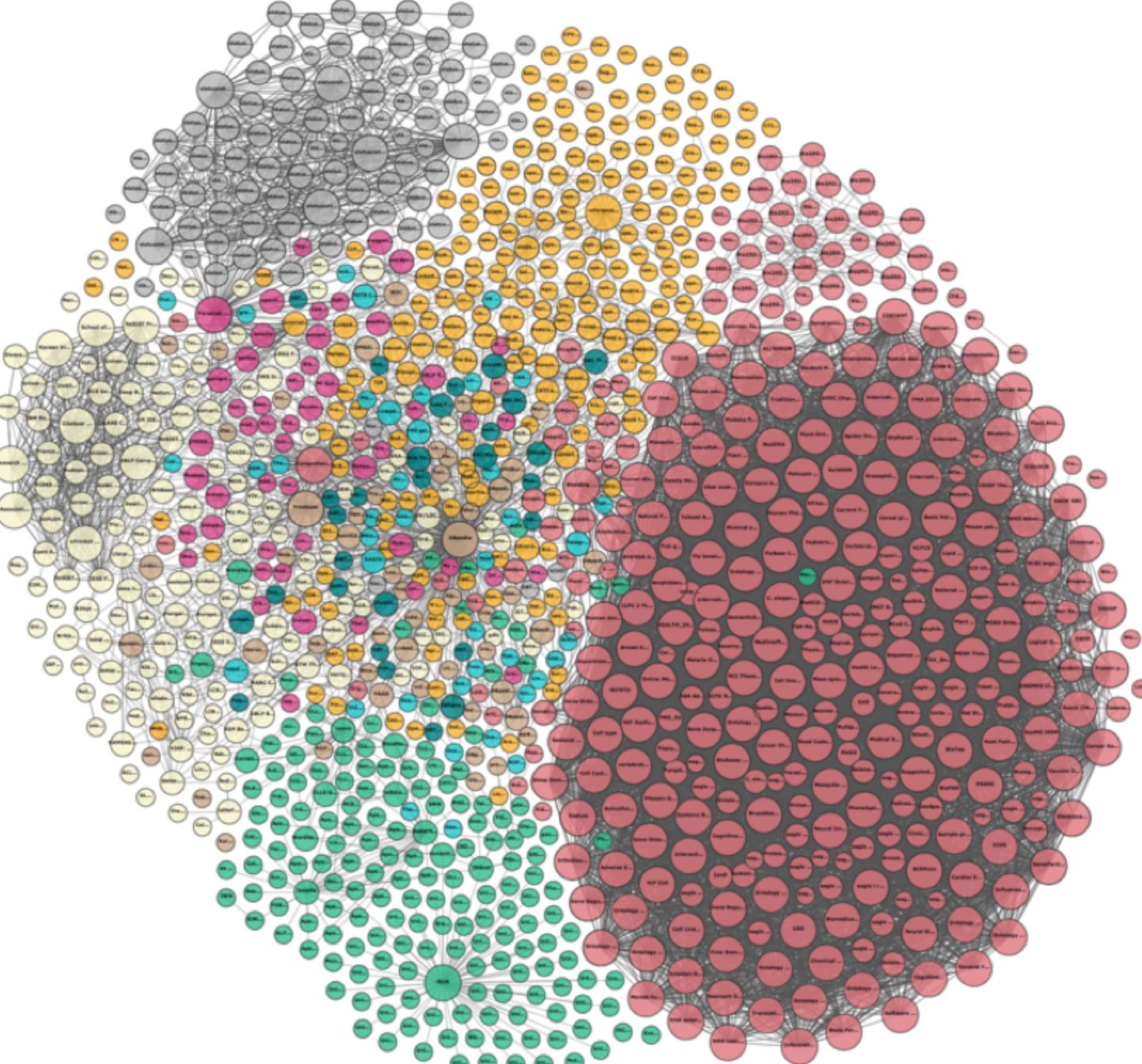
ElasticSearch, MongoDB & Neo4J Have Wide Adoption



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

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



YouTube



LinkedIn



Twitter



Facebook



Google+

People also search for



Tesla, Inc.



Nokia



Asus



T-Mobile



Samsung Group

[View 15+ more](#)

[Disclaimer](#)

[Feedback](#)

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

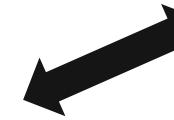
Querying Knowledge Graphs

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'**?



Schema



SEARCH RESET CANCEL

DATE POSTED	Date Posted Begin
IDENTIFIER	Telephone Number Email Address Review ID Social Media ID
LOCATION	City: chicago State/Region Country

```
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?

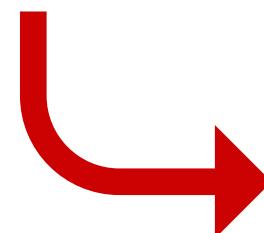
```
SELECT ?ad WHERE
{
  ?ad a :Ad ;
    :hair_color 'Auburn' ;
    :review_site_id 'cg9469f' ;
    :price_per_hour '500' ;
    :name 'Claire Gold' ;
    :ethnicity 'Asian'.
}
```



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”

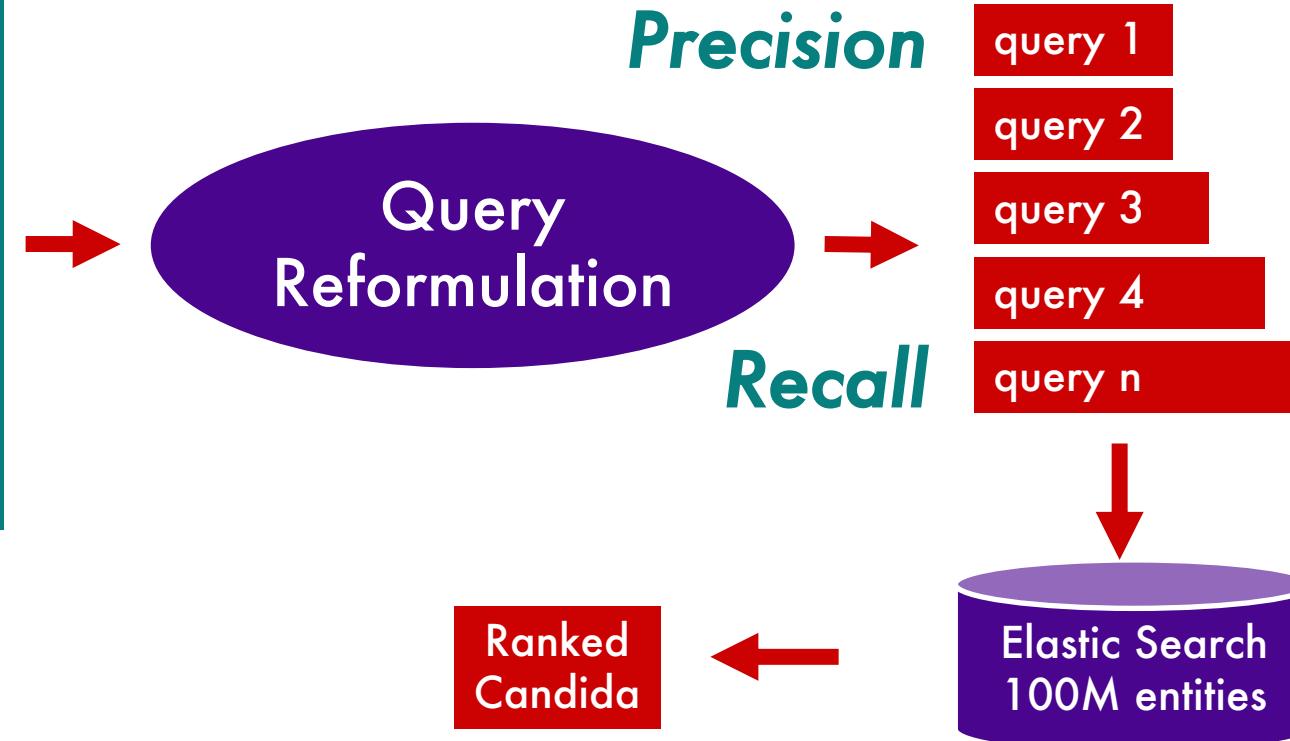


**No
results**

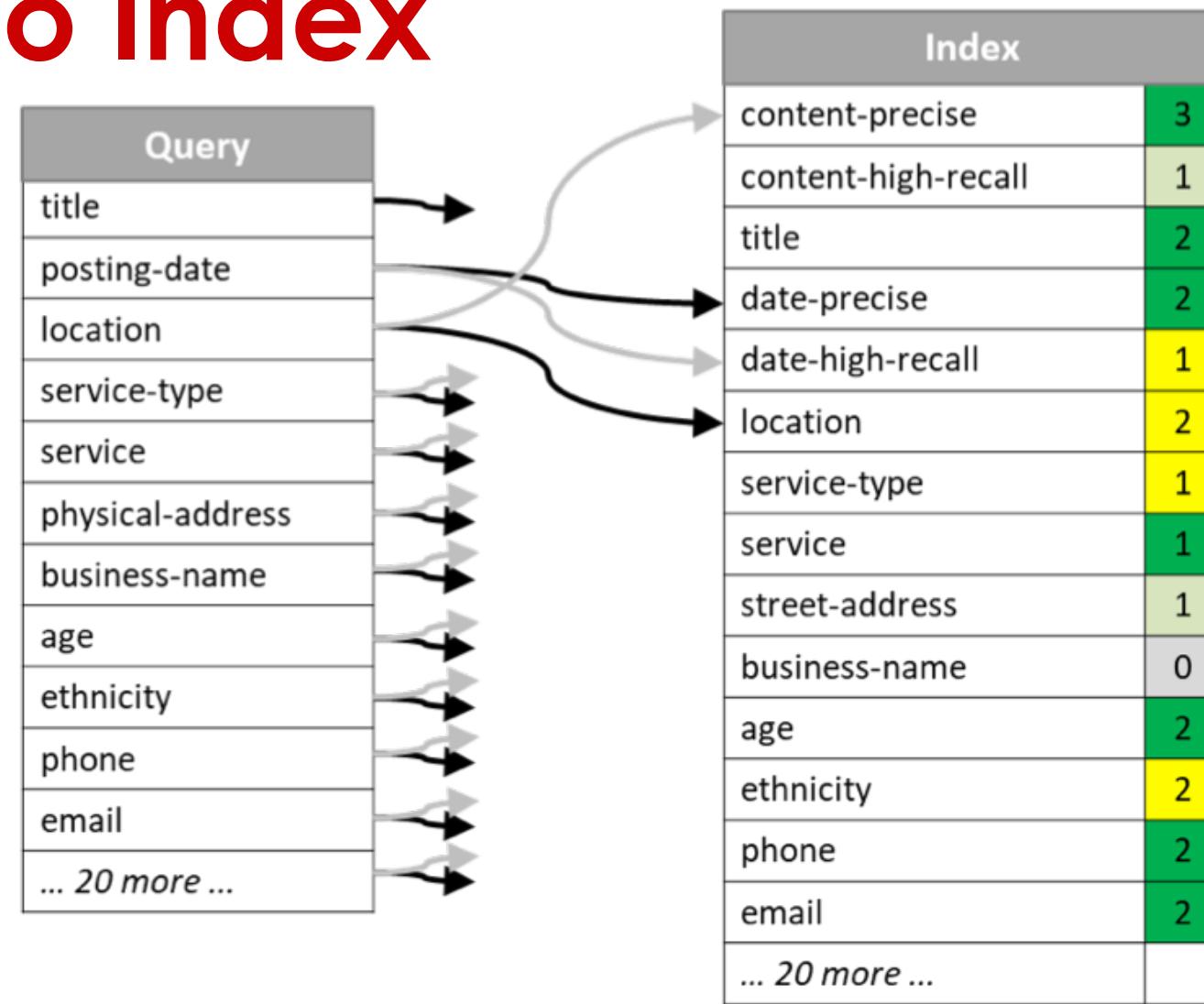
Candidate Generation

Keyword expansion • Context broadening • Constraint relaxation

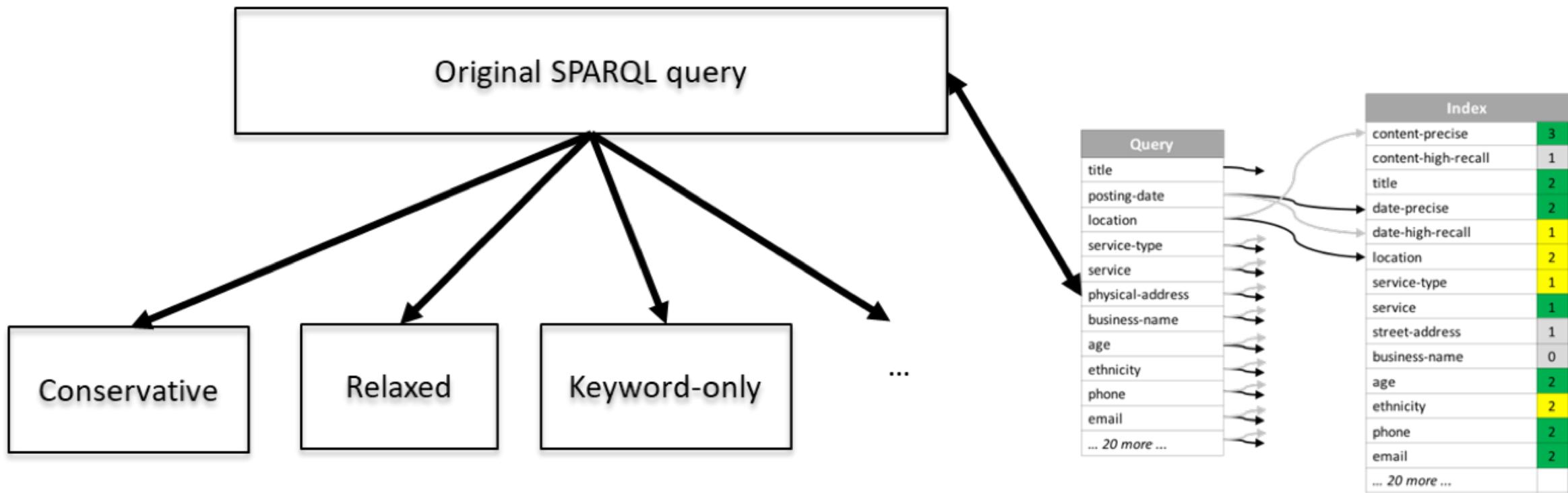
```
SELECT ?ad ?ethnicity WHERE
{
  ?ad a :Ad ;
  :hair_color 'Auburn' ;
  :review_site_id 'cg9469f' ;
  :price_per_hour '500' ;
  :name 'Claire Gold' ;
  :ethnicity ?ethnicity .
}
```



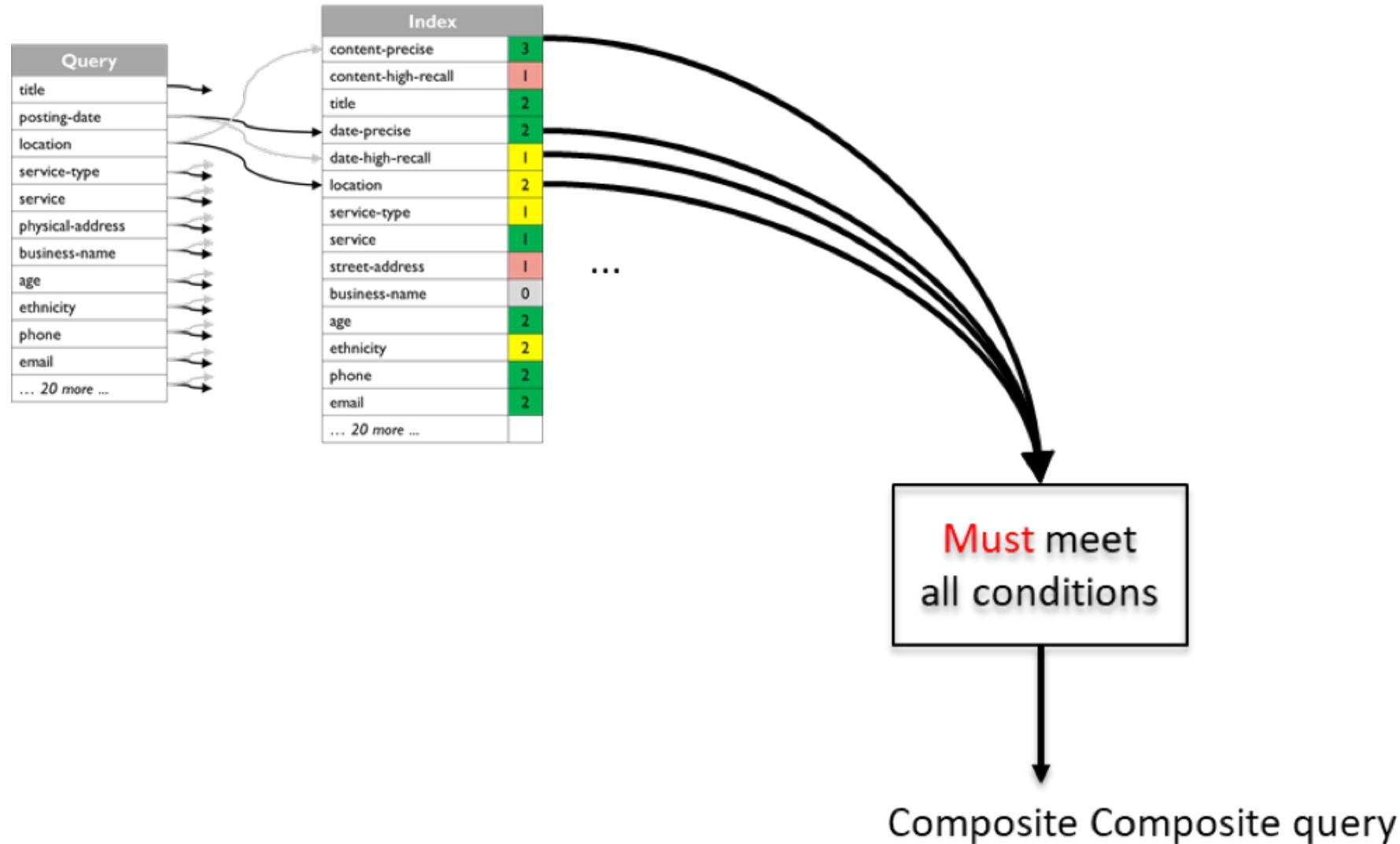
Offline step: Weighted Mapping Of Query To Index



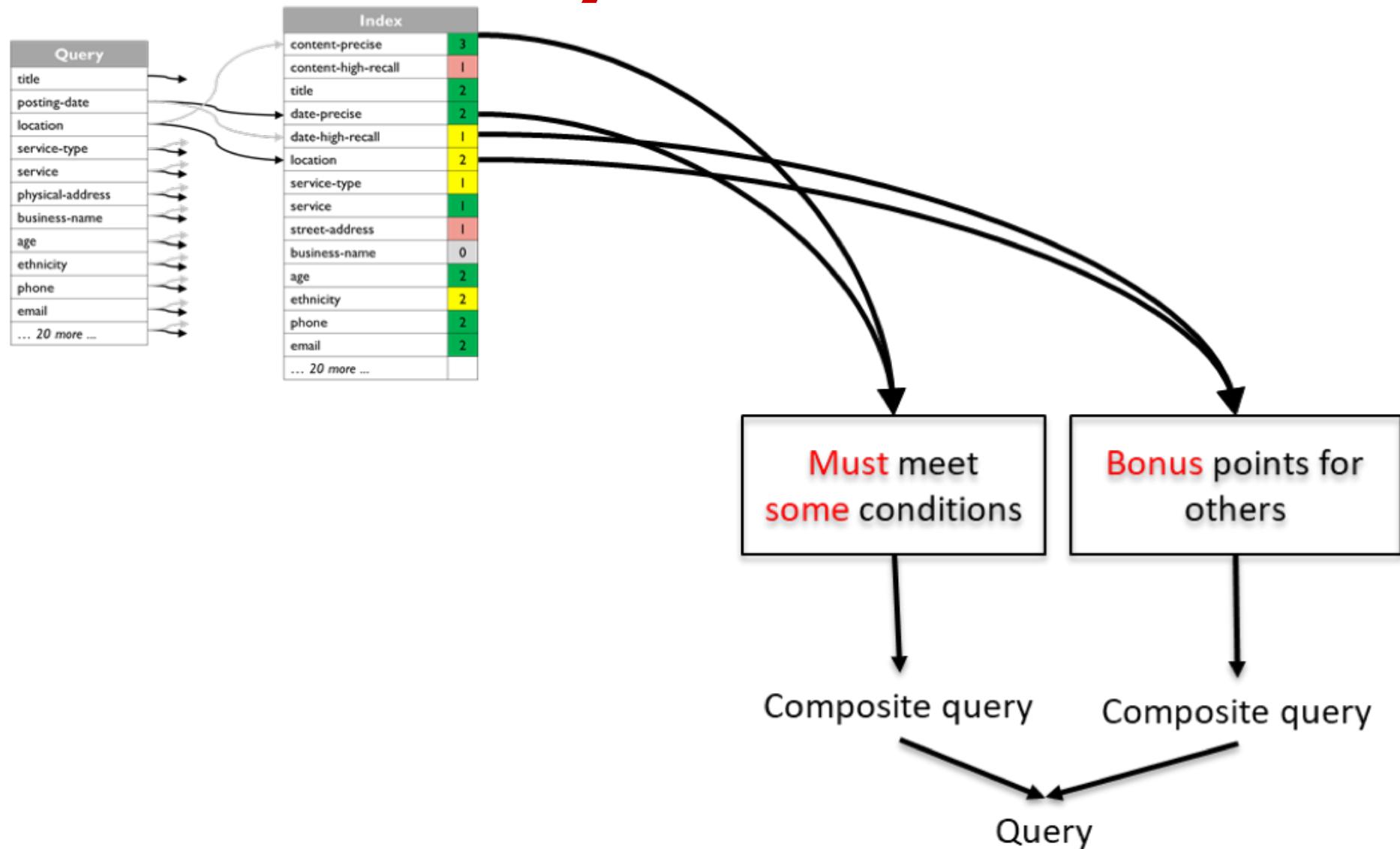
Online Step: Query reformulation using Semantic Strategies



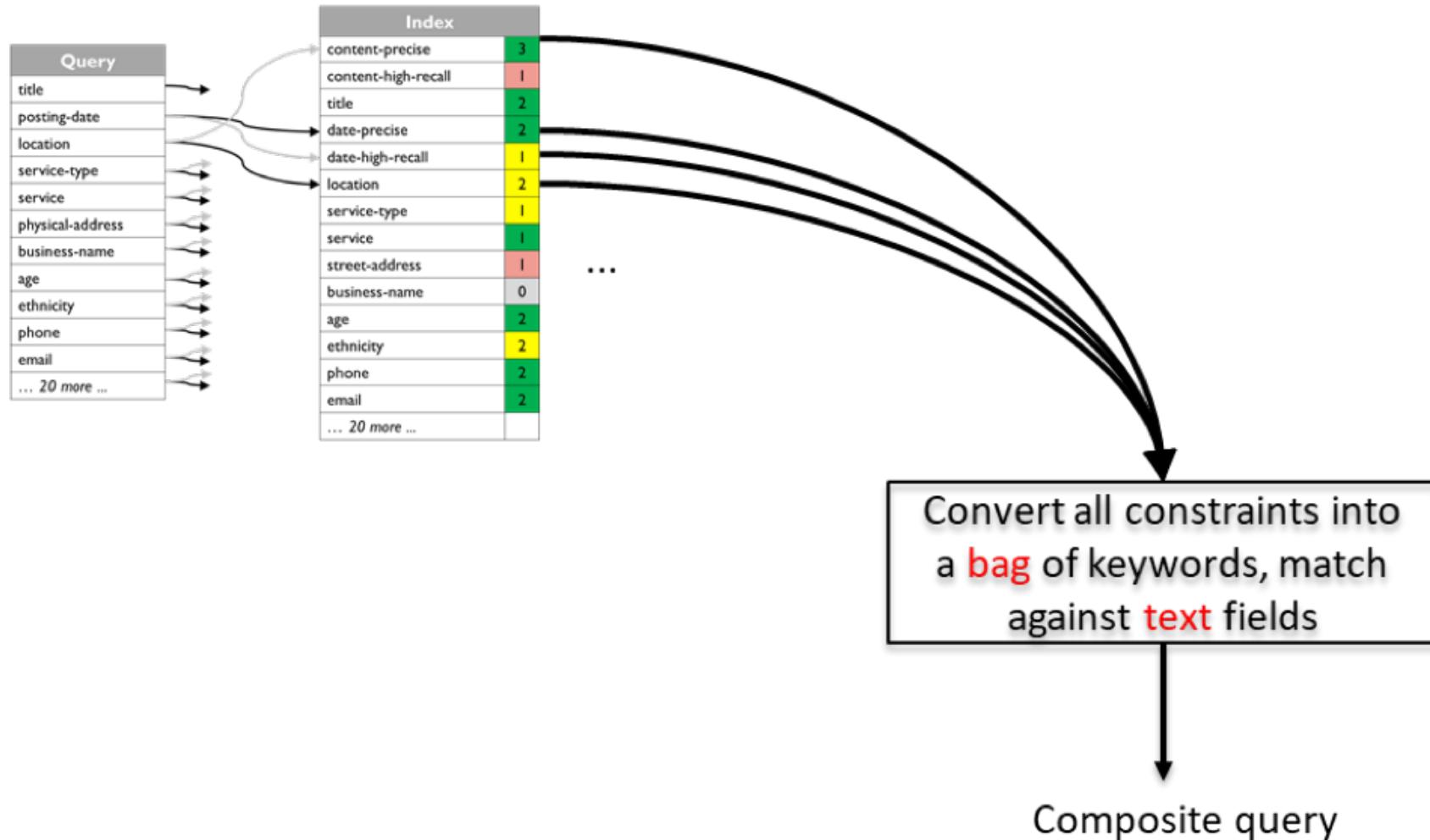
Conservative Query



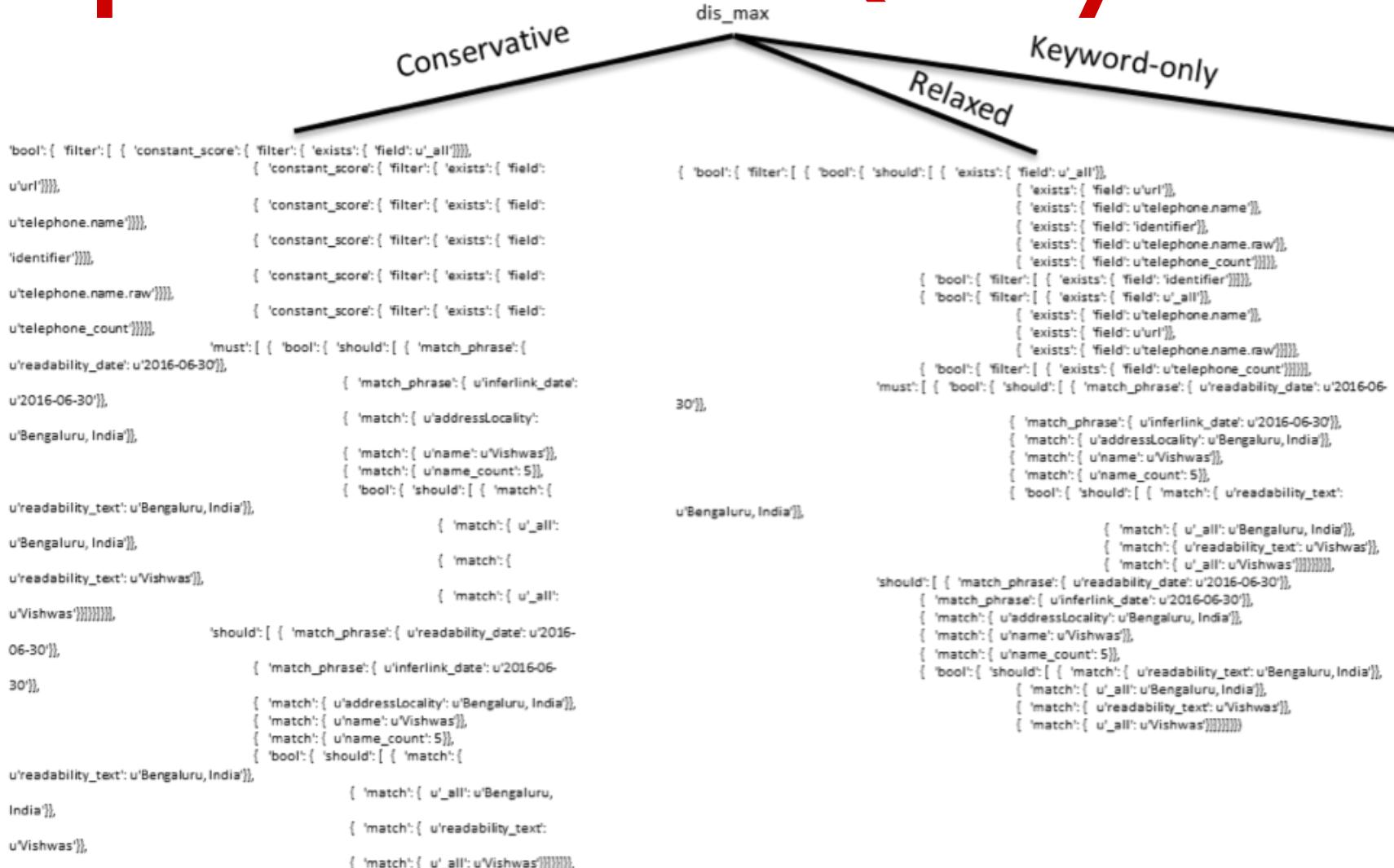
Relaxed Query



Keyword-only Query



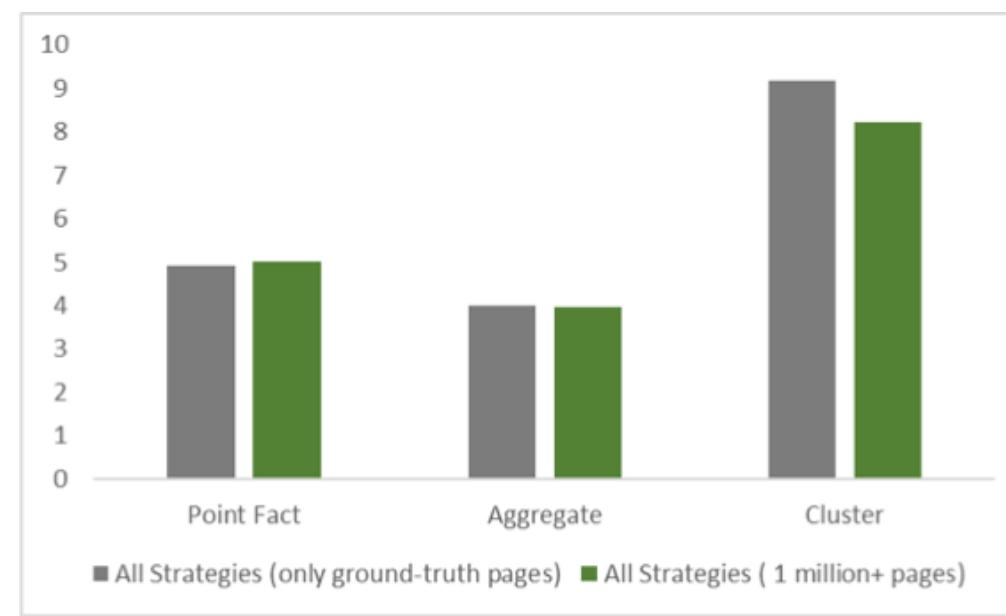
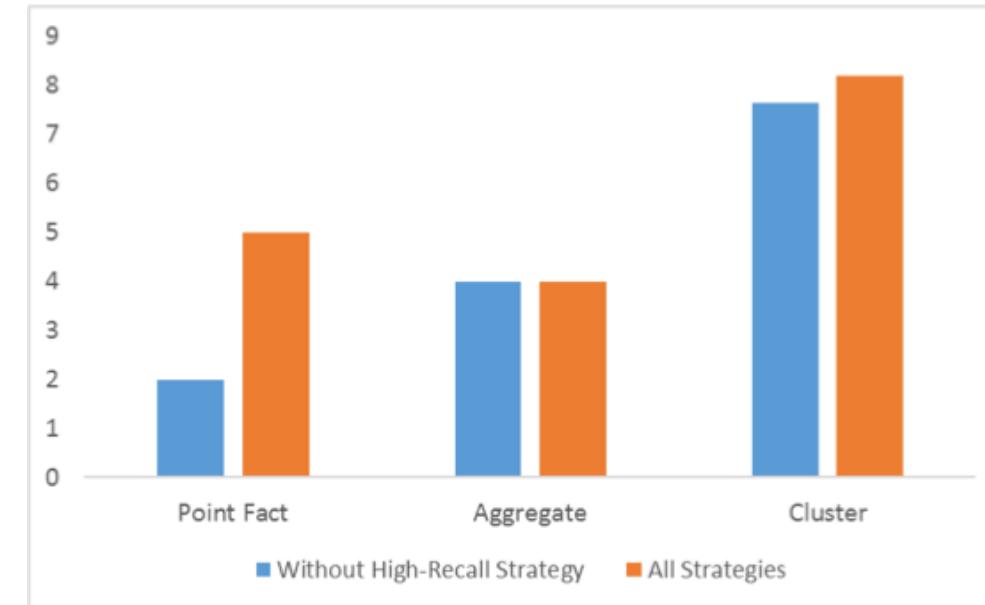
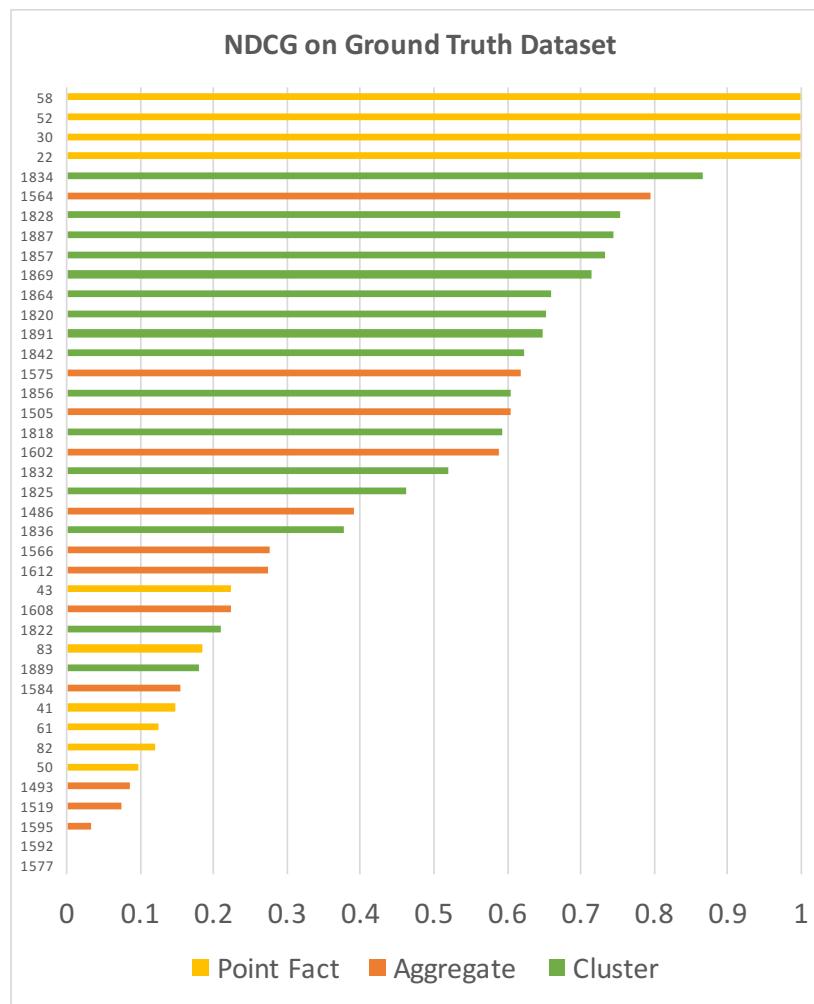
Example of ‘Final’ Query



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



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