Domain-Specific Corpora
Many Document Features

Text paragraphs without formatting

Non-grammatical snippets, rich formatting & links

Grammatical sentences plus some formatting & links

Tables

Charts

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.

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 experience in the software development industry.

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 ollie@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.
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”
Spreadsheets Created For Human Consumption
Databases with PDF Code Books

<table>
<thead>
<tr>
<th>Event Date</th>
<th>Year</th>
<th>Time Precision</th>
<th>Event Type</th>
<th>Actor 1</th>
<th>Actor 2</th>
<th>Interaction</th>
<th>Region</th>
<th>Country</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No ch Military Forces of Democracy</td>
<td>1 ADF: Allied Democratic FC</td>
<td>2 Central African Democratic (Nord-K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No ch Military Forces of Democracy</td>
<td>1 ADF: Allied Democratic FC</td>
<td>2 Central African Democratic (Nord-K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No ch Military Forces of Democracy</td>
<td>1 ADF: Allied Democratic FC</td>
<td>2 Central African Democratic (Nord-K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No ch Al Shabaab</td>
<td>2 Police Forces of Kenya (2C)</td>
<td>1 Eastern Africa (Kenya) Lamu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Riots/Protests (Kenya)</td>
<td>3 Police Forces of Kenya (2C)</td>
<td>1 Eastern Africa (Kenya) Mombasa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Riots/Protests (L. Tawartha, Co)</td>
<td>6</td>
<td>0 Northern AfLibya Sabha</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Violence against Unidentified Armed Group</td>
<td>3 Civilian, International</td>
<td>7 Northern AfLibya Sabha</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Riots/Protests (Morocco)</td>
<td>6</td>
<td>0 Northern AfLibya Sahara</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Riots/Protests (Nigeria)</td>
<td>6</td>
<td>0 Northern AfLibya Benue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Civilian = 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
Data In Web Tables

Security Council Resolutions

Resolutions adopted by the Security Council in 2017

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/RES/2297 (2017)</td>
<td>22 December 2017</td>
<td>Non-proliferation/Democratic People's Republic of Korea</td>
</tr>
<tr>
<td>S/RES/2298 (2017)</td>
<td>21 December 2017</td>
<td>Threats to international peace and security caused by terrorist acts</td>
</tr>
<tr>
<td>S/RES/2299 (2017)</td>
<td>21 December 2017</td>
<td>Threats to international peace and security caused by terrorist acts</td>
</tr>
<tr>
<td>S/RES/2304 (2017)</td>
<td>21 December 2017</td>
<td>The situation in the Middle East</td>
</tr>
<tr>
<td>S/RES/2393 (2017)</td>
<td>19 December</td>
<td>The situation in the Middle East</td>
</tr>
</tbody>
</table>

[Table image with data]
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

LEGACY VENTURES INTL, Inc. (LGYV)
6.01 0.00 (0.00%) 07/14/17

LEGACY VENTURES INTL, Inc. (LGYV)

- Dividend & Yield: N/A (N/A)
- P/E: -
- Market Cap: 391,03K
- EPS: -26648.00
- Volume: 67
- Day's Range: 6.01 - 6.01
- 52wk Range: 1.05 - 15.00

LEGXY Stock Quote Delayed 20 Minutes
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

<table>
<thead>
<tr>
<th>Block Type</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repeating blocks</strong></td>
<td>Web wrappers</td>
</tr>
<tr>
<td>(short tail)</td>
<td></td>
</tr>
<tr>
<td><strong>Tables</strong></td>
<td>Data table extractors</td>
</tr>
<tr>
<td>(long tail)</td>
<td></td>
</tr>
<tr>
<td><strong>Main content</strong></td>
<td><a href="https://code.google.com/archive/p/arc90labs-readability/">https://code.google.com/archive/p/arc90labs-readability/</a></td>
</tr>
<tr>
<td>(long tail)</td>
<td><a href="https://github.com/kohlschutter/boilerpipe">https://github.com/kohlschutter/boilerpipe</a></td>
</tr>
<tr>
<td><strong>Microdata</strong></td>
<td><a href="https://github.com/namsral/microdata">https://github.com/namsral/microdata</a></td>
</tr>
<tr>
<td>(long tail)</td>
<td></td>
</tr>
</tbody>
</table>
Web Wrappers
myDIG Demo

Focusing On Inferlink Web Wrapper
Table Extraction
## Classification Of Web Tables

<table>
<thead>
<tr>
<th>Table type</th>
<th>% total</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Tiny” tables</td>
<td>88.06</td>
<td>12.34B</td>
</tr>
<tr>
<td>HTML forms</td>
<td>1.34</td>
<td>187.37M</td>
</tr>
<tr>
<td>Calendars</td>
<td>0.04</td>
<td>5.50M</td>
</tr>
<tr>
<td>Filtered Non-relational, total</td>
<td>89.44</td>
<td>12.53B</td>
</tr>
<tr>
<td>Other non-rel (est.)</td>
<td>9.46</td>
<td>1.33B</td>
</tr>
<tr>
<td>Relational (est.)</td>
<td>1.10</td>
<td>154.15M</td>
</tr>
</tbody>
</table>

*Cafarella’08*
Tables In The Human Trafficking Domain
# Data Tables

<table>
<thead>
<tr>
<th>Name</th>
<th>Nationality</th>
<th>From</th>
<th>To</th>
<th>M</th>
<th>W</th>
<th>D</th>
<th>L</th>
<th>GF</th>
<th>GA</th>
<th>Win %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsène Wenger</td>
<td>France</td>
<td>1 October 1996</td>
<td>Present</td>
<td>1,188</td>
<td>684</td>
<td>271</td>
<td>233</td>
<td>2,063</td>
<td>1,088</td>
<td>57.58</td>
</tr>
<tr>
<td>Pat Rice †</td>
<td>Northern Ireland</td>
<td>13 September 1996</td>
<td>30 September 1996</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>75.00</td>
</tr>
<tr>
<td>Stewart Houston †</td>
<td>Scotland</td>
<td>12 August 1996</td>
<td>13 September 1996</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>11</td>
<td>10</td>
<td>33.33</td>
</tr>
<tr>
<td>Bruce Rioch</td>
<td>Scotland</td>
<td>15 June 1995</td>
<td>12 August 1996</td>
<td>47</td>
<td>22</td>
<td>15</td>
<td>10</td>
<td>67</td>
<td>37</td>
<td>46.81</td>
</tr>
<tr>
<td>Stewart Houston †</td>
<td>Scotland</td>
<td>21 February 1995</td>
<td>15 June 1995</td>
<td>19</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>29</td>
<td>25</td>
<td>36.84</td>
</tr>
<tr>
<td>George Graham</td>
<td>Scotland</td>
<td>14 May 1986</td>
<td>21 February 1995</td>
<td>460</td>
<td>225</td>
<td>133</td>
<td>102</td>
<td>711</td>
<td>403</td>
<td>48.91</td>
</tr>
<tr>
<td>Steve Burtenshaw †</td>
<td>England</td>
<td>23 March 1986</td>
<td>14 May 1986</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>15</td>
<td>27.27</td>
</tr>
</tbody>
</table>

Honours:
- **Charity/Community Shield winners**: 1996–97, 2017

Honours:
- **First Division champions**: 1988–89, 1990–91
- **FA Cup winners**: 1992–93
- **Football League Cup winners**: 1986–87
- **Charity Shield winners**: 1991 (shared)
- **UEFA Cup Winners’ Cup winners**: 1993–94
Data Tables

**Entity Table**

**Matrix Table**

**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/extendings glossaries automatically
Regex Extraction
Extraction Using Regular Expressions

Too difficult for non-programmers

regex for North American phone numbers:

```
^\(?:\d+\)?\s*\(?\d{3}\)?\s*\d{3}\s*\d{4}\s*(?:\s*\#|x\.|ext\.|extension)\s*\(?\d{1,4}\)?\s*$
```

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.

```python
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"world"]
```
NLP Rule-Based Extraction

Tokenization

Pattern Matching
Tokenization matters, a lot

My name is Pedro

310-822-1511

❤️ 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:
Part of speech:
- Noun
- Pronoun
- Proper noun
- Determiner
- Symbol
- Adjective

Capitalization:
- Exact
- Lower
- Upper
- Title
- Mixed

Length 1: [ ] Length 2: [ ] Length 3: [ ]
Prefix: [ ] Suffix: [ ] Not in vocabulary

Create Shape Token
- Optional
- Part of output

Shape:
Enter shapes such as dots, XXXX, Xv.
d is for digits and x for letter, X for capital letter.

Part of speech:
- Noun
- Pronoun
- Proper noun
- Determiner
- Symbol
- Adjective

Prefix: [ ] Suffix: [ ]

Create Number Token
- Optional
- Part of output

Numbers:
- Length 1: [ ] Length 2: [ ]
- Length 3: [ ]
- Min: [ ] Max: [ ]

Create Punctuation Token
- Optional
- Part of output

Punctuation Symbols:
- ,
- .
- (,)
- [ ]
- \n
- ;
- ?
- ~
- :,
- "
- '"'

- 
- +
- -
- ^
- &
- @
- <
- >
- =
- %
- \n
- 
- $
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
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.

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).

```
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
```
2 April 2016: 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 Uwuujaren quoted Wabba as saying: "Corruption is a monster that has done much harm to our country and everyone else."
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 US 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

<table>
<thead>
<tr>
<th>i</th>
<th><strong>X1</strong> (word)</th>
<th><strong>X2</strong> (capitalized)</th>
<th><strong>X3</strong> (POS Tag)</th>
<th><strong>Y</strong> (entity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>My</td>
<td>1</td>
<td>Possessive Pron</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>name</td>
<td>0</td>
<td>Noun</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>0</td>
<td>Verb</td>
<td>Other</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
<tr>
<td>5</td>
<td>Szekely</td>
<td>1</td>
<td>Proper Noun</td>
<td>Person-Name</td>
</tr>
</tbody>
</table>

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

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?

A12 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-answermatches (Fader et al. ACL 2013)
# Practical IE Technologies

<table>
<thead>
<tr>
<th>Effort</th>
<th>Glossary</th>
<th>Regex</th>
<th>NLP Rules</th>
<th>Semi-Structured</th>
<th>CRF</th>
<th>NER</th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>assemble glossary</td>
<td>hours</td>
<td>hours</td>
<td>minutes</td>
<td>$O(1000)$ annotations</td>
<td>zero</td>
<td>$O(10)$ annotations</td>
</tr>
<tr>
<td>Expertise</td>
<td>minimal</td>
<td>high, programmer</td>
<td>low</td>
<td>minimal</td>
<td>low-medium</td>
<td>zero</td>
<td>minimal</td>
</tr>
<tr>
<td>Precision</td>
<td>medium (ambiguity)</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>medium-high</td>
<td>medium-high</td>
<td>high</td>
</tr>
<tr>
<td>Recall</td>
<td>medium (formatting)</td>
<td>low f(# regex)</td>
<td>medium f(# rules)</td>
<td>high</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Coverage</td>
<td>wide</td>
<td>wide</td>
<td>wide</td>
<td>single site</td>
<td>genre</td>
<td>genre</td>
<td>narrow</td>
</tr>
</tbody>
</table>
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

- LGYV
- Legacy Ventures International Inc
- Damn Good Penny Stocks

Easiest to build
Simple, But Useful KG

Entities + properties

stock-ticker

company

promoter

LGYV

Legacy Ventures International Inc

Damn Good Penny Stocks

“Easy” to build
Semantic Web KG (RDF/OWL)

Entities + properties + classes

Very hard to build
“Ideal” KG
Entities + properties + classes + qualifiers

Company

is-a

Legacy Ventures International Inc

stock-ticker

LGYV

Damn Good Penny Stocks

is-a

promoter

source

stockreads.com

start-date

June 2017

Very very hard to build
Semi-Structured KG

Entities + properties + text + provenance + confidence

- extraction confidence
- reliability
- ambiguity
- # sources
- error reduction
- confidence
- method
- origin
- source
- segment
- media type
- image
- # of sources
- reliability
- error reduction

 ISI-extractor

(150,230)x(560,720)

image-id-123

0.92

0.72

0.14

0.81

2 june 2014

2 june 2014

Snizhne

location

event 123

“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

DB-Engines Ranking of Graph DBMS

© August 2017, DB-Engines.com

https://db-engines.com/en/ranking_trend/graph+dbms
ElasticSearch, MongoDB & Neo4J Have Wide Adoption

![DB-Engines Ranking](https://db-engines.com/en/ranking_trend/graph/dbms)
KGs I can Reuse
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


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
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
http://engineering.linkedin.com/2016/10/building-the-linkedin-knowledge-graph
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”

NoSQL store → 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**

<table>
<thead>
<tr>
<th>Query</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>content-precise 3</td>
</tr>
<tr>
<td>posting-date</td>
<td>content-high-recall 1</td>
</tr>
<tr>
<td>location</td>
<td>title 2</td>
</tr>
<tr>
<td>service-type</td>
<td>date-precise 2</td>
</tr>
<tr>
<td>service</td>
<td>date-high-recall 1</td>
</tr>
<tr>
<td>physical-address</td>
<td>location 2</td>
</tr>
<tr>
<td>business-name</td>
<td>service-type 1</td>
</tr>
<tr>
<td>age</td>
<td>service 1</td>
</tr>
<tr>
<td>ethnicity</td>
<td>street-address 1</td>
</tr>
<tr>
<td>phone</td>
<td>business-name 0</td>
</tr>
<tr>
<td>email</td>
<td>age 2</td>
</tr>
<tr>
<td>... 20 more ...</td>
<td>ethnicity 2</td>
</tr>
<tr>
<td></td>
<td>phone 2</td>
</tr>
<tr>
<td></td>
<td>email 2</td>
</tr>
<tr>
<td></td>
<td>... 20 more ...</td>
</tr>
</tbody>
</table>
Online Step: Query reformulation using Semantic Strategies

Original SPARQL query

- Conservative
- Relaxed
- Keyword-only

Index

<table>
<thead>
<tr>
<th>Query</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>content-precise</td>
<td>3</td>
</tr>
<tr>
<td>content-high-recall</td>
<td>1</td>
</tr>
<tr>
<td>title</td>
<td>2</td>
</tr>
<tr>
<td>posting-date</td>
<td>2</td>
</tr>
<tr>
<td>location</td>
<td>1</td>
</tr>
<tr>
<td>service-type</td>
<td>1</td>
</tr>
<tr>
<td>service</td>
<td>1</td>
</tr>
<tr>
<td>physical-address</td>
<td>1</td>
</tr>
<tr>
<td>business-name</td>
<td>1</td>
</tr>
<tr>
<td>age</td>
<td>0</td>
</tr>
<tr>
<td>ethnicity</td>
<td>2</td>
</tr>
<tr>
<td>phone</td>
<td>2</td>
</tr>
<tr>
<td>email</td>
<td>2</td>
</tr>
</tbody>
</table>

... 20 more...
Conservative Query

Composite Composite query

Must meet all conditions

Index
- content-precise: 0
- content-high-recall: 1
- title: 2
- date-precise: 1
- date-high-recall: 1
- location: 2
- service-type: 0
- service: 2
- street-address: 1
- business-name: 0
- age: 2
- ethnicity: 2
- phone: 2
- email: 2
- ... 20 more ...

Query
- title
- posting-date
- location
- service-type
- service
- physical-address
- business-name
- age
- ethnicity
- phone
- email
- ... 20 more ...
Relaxed Query

Index

| Query       | content-precise | content-high-recall | title | date-precise | date-high-recall | location | service-type | service | street-address | business-name | age | ethnicity | phone | email | ... 20 more ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posting-date</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>service-type</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>service</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>street-name</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>business-name</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnicity</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phone</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>email</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>... 20 more</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Must meet some conditions**

**Bonus points for others**

Composite query

Query
Keyword-only Query

Convert all constraints into a **bag** of keywords, match against **text** fields

Composite query
Example of ‘Final’ Query
## Example: query execution/ranking

<table>
<thead>
<tr>
<th>name</th>
<th>hair color</th>
<th>price</th>
<th>review site id</th>
<th>ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claire Gold</td>
<td>Auburn</td>
<td>500</td>
<td>cg9469f</td>
<td></td>
</tr>
<tr>
<td>Claire title/dict</td>
<td>Red content/dict</td>
<td>500 content/regex</td>
<td></td>
<td>Asian content/dict</td>
</tr>
<tr>
<td>Rosa content/dict</td>
<td>Black content/dict</td>
<td>400 content/regex</td>
<td></td>
<td>Japanese content/dict</td>
</tr>
<tr>
<td>June content/dict</td>
<td>Auburn content/CRF</td>
<td>2016 content/regex</td>
<td></td>
<td>Korean content/CRF</td>
</tr>
<tr>
<td>Clara content/dict</td>
<td></td>
<td></td>
<td>cg9469f content/ES</td>
<td>Japanese content/dict</td>
</tr>
<tr>
<td>June content/dict</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Claire Gold content/ES</td>
<td>Auburn content/ES</td>
<td>150 title/regex</td>
<td>cg9469f content/ES</td>
<td>Asian content/dict</td>
</tr>
<tr>
<td></td>
<td></td>
<td>125 title/regex</td>
<td></td>
<td>Japanese content/dict</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 content/regex</td>
<td></td>
<td>Caramel content/dict</td>
</tr>
</tbody>
</table>

---

*Information Sciences Institute*
Results

NDCG on Ground Truth Dataset

Point Fact vs Aggregate vs Cluster

Without High-Recall Strategy vs All Strategies

All Strategies (only ground-truth pages) vs All Strategies (1 million+ pages)
myDIG: A KG Construction Toolkit

Python, MIT license, https://github.com/usc-isisi2/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