Conclusion and review
Domain-specific search (DSS)
Emerging opportunities for DSS

Fighting human trafficking

Predicting cyberattacks

Stopping Penny Stock Fraud

Accurate geopolitical forecasting
How do we construct domain specific knowledge graphs over web data for powerful DSS applications
Challenges
Many Document Features

- **Text paragraphs without formatting**

  Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

- **Grammatical sentences plus some formatting & links**

  Dr. Steven Minton - Founder/CTO
  Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC’s Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.
  Frank Huybrechts - COO
  Mr. Huybrechts has over 20 years of

- **Non-grammatical snippets, rich formatting & links**

  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.

- **Tables**

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Email</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baro, Andrew G.</td>
<td>(413) 545-2109</td>
<td><a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a></td>
<td>CS276</td>
</tr>
<tr>
<td>Berger, Emory D.</td>
<td>(413) 577-4211</td>
<td><a href="mailto:emory@cs.umass.edu">emory@cs.umass.edu</a></td>
<td>CS444</td>
</tr>
<tr>
<td>Brock, Oliver</td>
<td>(413) 577-0334</td>
<td><a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a></td>
<td>CS246</td>
</tr>
<tr>
<td>Clarke, Lori A.</td>
<td>(413) 545-1328</td>
<td><a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a></td>
<td>CS304</td>
</tr>
<tr>
<td>Cohen, Paul R.</td>
<td>(413) 545-3638</td>
<td><a href="mailto:cohen@cs.umass.edu">cohen@cs.umass.edu</a></td>
<td>CS278</td>
</tr>
</tbody>
</table>

  Professor.
  Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.
  Assistant Professor.
  Software verification, testing, and analysis; software architecture and design.
  Professor.
  Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

- **Charts**

  [Graphical chart showing data]
Scope

Short tail

Genre specific (e.g., forums)

Long tail
## Pattern Complexity

### Closed set
- **U.S. states**
  - He was born in *Alabama*...
  - The big *Wyoming* sky...

### Regular set
- **U.S. phone numbers**
  - Phone: *(413) 545-1323*
  - The CALD main office can be reached at *412-268-1299*

### Complex
- **U.S. postal addresses**
  - University of Arkansas
    - P.O. Box 140
    - Hope, AR 71802
  - Headquarters:
    - 1128 Main Street, 4th Floor
    - Cincinnati, Ohio 45210

### Ambiguous, needing context
- **Person names**
  - ...was among the six houses sold by *Hope Feldman* that
  - Pawel Opalinski, Software Engineer at WhizBang Labs.

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

--

*Courtesy of Andrew McCallum*
small amount of relevant content
irrelevant content very similar to relevant content
Spreadsheets Created For Human Consumption

Cross-Country Indexes, Period-over-Period Change

<table>
<thead>
<tr>
<th>Country</th>
<th>Index</th>
<th>2002M11</th>
<th>2002M12</th>
<th>2003Q1</th>
<th>2003M01</th>
<th>2003M02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecuador</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egypt</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>El Salvador</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equatorial Guinea</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eritrea</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonia</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethiopia</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faeroe Islands</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiji</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French Southern Terri</td>
<td>Consumer Price Index, All Items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Databases with PDF Code Books

<table>
<thead>
<tr>
<th>Event Date</th>
<th>Year</th>
<th>Time</th>
<th>Event Type</th>
<th>Actor 1</th>
<th>Assoc Actor</th>
<th>Actor 2</th>
<th>Assoc Actor</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</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/13/18</td>
<td>2018</td>
<td></td>
<td>Battle-No</td>
<td>Military Forces of Democracy</td>
<td>1</td>
<td>ADF: Allied Democratic Forces</td>
<td>2</td>
<td>Central Afric Democratic (Nord-Kivu)</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

PDF
Data In Web Tables

Security Council Resolutions

<table>
<thead>
<tr>
<th>Resolutions adopted by the Security Council in 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S/RES/2387 (2017)</strong> 22 December 2017</td>
</tr>
<tr>
<td><strong>S/RES/2386 (2017)</strong> 21 December 2017</td>
</tr>
<tr>
<td><strong>S/RES/2385 (2017)</strong> 21 December 2017</td>
</tr>
<tr>
<td><strong>S/RES/2384 (2017)</strong> 21 December 2017</td>
</tr>
<tr>
<td><strong>S/RES/2383 (2017)</strong> 19 December 2017</td>
</tr>
</tbody>
</table>

Information Sciences Institute
Knowledge Graph Construction: Long-tail vs. Short-tail
Extracting Data from Semi-structured Sources

Wrappers

NAME: Casablanca Restaurant
STREET: 220 Lincoln Boulevard
CITY: Venice
PHONE: (310) 392-5751
Information Integration in Karma

Karma supports multiple integration regimes

Karma semi-automatically generates Source Mappings

Domain Model

Samples of Source Data

Karma

Data Warehousing

Virtual Integration
Karma semi-automatically builds semantic models ... and provides a nice GUI to edit them.

Karma uses semantic models to create knowledge graphs.
Practical Considerations for Extractions

Long tail vs. short tail

• How good (precision/recall) is necessary?
  – High precision when showing KG nodes to users
  – High recall when used for ranking results
• How long does it take to construct?
  – Minutes, hours, days, months
• What expertise do I need?
  – None (domain expertise), patience (annotation), scripting, machine learning guru
• What tools can I use?
  – Many …
## Data Tables

### Entity Table

**Arsène Wenger**

<table>
<thead>
<tr>
<th>Personal information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full name</strong></td>
<td>Arsène Wenger[^1]</td>
</tr>
<tr>
<td><strong>Date of birth</strong></td>
<td>22 October 1949 (age 67)</td>
</tr>
<tr>
<td><strong>Place of birth</strong></td>
<td>Strasbourg, Alsace, France</td>
</tr>
<tr>
<td><strong>Height</strong></td>
<td>6 ft 3 in (1.91 m)[^2]</td>
</tr>
<tr>
<td><strong>Playing position</strong></td>
<td>Midfielder</td>
</tr>
</tbody>
</table>

### Matrix Table

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

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

### List 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
Table Type Classification

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

![Semantic + Structure Embedding Diagram]

- 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 Techniques

- Glossary
- Regular expressions
- Natural language rules
- Named entity recognition
- Sequence labeling (Conditional Random Fields)
Searching Knowledge Graphs
Many problems with ‘strict’ execution

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

- Synonyms “red”
- Typos “brunette”
- Not present
- Numbers hard to match
- Claire is a common name
- Gold is a domain word
- Slang, e.g., “FOB” for Asian
- Inference, e.g., “Japanese”

No results
Candidate Generation

Keyword expansion • Context broadening • Constraint relaxation

```
SELECT ?ad ?ethnicity WHERE {
    ?ad a :Ad ;
    :hair_color 'Auburn' ;
    :review_site_id 'cg9469f' ;
    :price_per_hour '500' ;
    :name 'Claire Gold' ;
    :ethnicity ?ethnicity .
}
```
Knowledge Graph Completion
For Web extractions, noise is inevitable

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

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

Many nodes refer to the same underlying entity
Entity Resolution is fundamentally non-linear

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

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

Knowledge graph → Execute blocking → Candidate set → Execute similarity → Resolved entities
Other research frontiers for KG Completion

• Other things explored in the literature:

  • Domain knowledge
    - Collective ER methods have tried to exploit these systematically

  • Multi-type Entity Resolution
    - Extremely useful for knowledge graphs, lots more work to be done

  • Entity Resolution+Ontologies+IE Confidences:
    - Probabilistic Graphical Models like Probabilistic Soft Logic

  • Knowledge graph embeddings
    - Useful for link prediction and triples classification
    - Recall the Microsoft-founded_in-Seattle example earlier
THANK YOU! QUESTIONS...