From Data Fusion to Knowledge Fusion

Xin Luna Dong, Google Inc.
9/6/2014 @ APWeb ’14
SONYA: A Big Project, A Fancy Machine, And A Cute Little Girl
Knowledge Is Power

- Many Knowledge Bases (KB)

![Logos of various knowledge bases and related projects]
The most important Google story this year was the launch of the *Knowledge Graph*. This marked the shift from a first-generation Google that merely indexed the words and metadata of the Web to a next-generation Google that recognizes discrete things and the relationships between them.

- ReadWrite 12/27/2012
Using KG in Search

岳麓书院-首页
ylsy.hnu.cn 转为繁體網頁 Yuelu Academy

岳麓书院位于中国湖南省长沙市岳麓山东麓，是中国古代四大书院之一，始建于北宋开宝九年（976年），历经宋、元、明、清各个朝代，迄至晚清(1903年)改为湖南高等 ... 历史沿革 - 大事年表 - 书院学规 - 历代山长

岳麓书院- 维基百科，自由的百科全书
岳麓书院位于中国湖南省长沙市岳麓山东麓，是中国古代四大书院之一，始建于北宋开宝九年（976年），历经宋、元、明、清各个朝代，迄至晚清(1903年)改为湖南高等 ... 历史沿革 - 大事年表 - 书院学规 - 历代山长

News for 岳麓

“小社区大民生”领创和谐岳麓
光明网 - 5 hours ago
在坐拥15个街道2个镇、81个社区的岳麓区，其中有1个街道、2个社区荣获全国和谐社区建设示范单位，有3个街道、24个社区分别荣获省、市精品、... 做客乘坐缆车高空抛垃圾 岳麓山很“受伤”
红网 - 21 hours ago

More news for 岳麓

岳麓书院_百度百科
baike.baidu.com/view/7288.htm 转为繁體網頁 Baidu Baike
北宋开宝九年（976），潭州太守朱洞在僧人办学的基础上，正式创立岳麓书院，嗣后，历经宋、元、明、清各代，至清末光绪二十九年（1903）改为湖南高等学堂，尔后相继 ...
But—

- KG requires 99% accuracy for knowledge
- Web data is noisy and extraction is hard
- How to balance coverage and accuracy?
Google's Knowledge Vault already contains 1.6 billion facts

FELICITY NELSON
SATURDAY, 23 AUGUST 2014

The automated, fact-harvesting bot will build up a collection of all human knowledge.
Outline

I. Knowledge extraction

II. Knowledge fusion

III. Interesting applications

IV. Future directions
Knowledge Extraction I–Knowledge

- **Triple:** (subject, predicate, object)
  - e.g., (Tom Cruise, date_of_birth, 7/3/1962)
    - **Subject**– a Freebase mid
      - e.g., /m/07r1h
    - **Predicate**– predefined in Freebase; e.g., people/person/date_of_birth
    - **Object**– a Freebase mid, a string, a number, or a date.
Statistics for Extracted Triples

- A large knowledge base  

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#Triples</td>
<td>1.6B (now 2.8B)</td>
</tr>
<tr>
<td>#Subjects (Entities)</td>
<td>43M</td>
</tr>
<tr>
<td>#Types</td>
<td>1.1K</td>
</tr>
<tr>
<td>#Predicates</td>
<td>4.5K</td>
</tr>
<tr>
<td>#Objects</td>
<td>102M</td>
</tr>
</tbody>
</table>

- Highly skewed data—fat heads, long tail
  - #Triples/type: 1–14M  
    (location, organization, business)
  - #Triples/entity: 1–2M  
    (USA, UK, CA, NYC, TX)

As of 11/2013
Knowledge Extraction II—Sources

Web

TXT   DOM   TBL   ANO

Free texts

Synopsis
Born on April 15, 1452, in Vinci, Italy, Leonardo da Vinci was concerned with the laws of science and nature, which greatly informed his work as a painter, sculptor, architect, draftsman, and engineer. His ideas and body of work -- which in his own time was little known, even in Italy -- influenced countless artists and made da Vinci a forerunner of the Italian Renaissance.

Web tables & Lists

<table>
<thead>
<tr>
<th>Name and (party)</th>
<th>Term</th>
<th>State of birth</th>
<th>Born</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Washington (F)</td>
<td>1789–1797</td>
<td>Va.</td>
<td>2/22/1732</td>
<td>12/14/1826</td>
</tr>
<tr>
<td>2. J. Adams (F)</td>
<td>1797–1801</td>
<td>Mass.</td>
<td>10/30/1735</td>
<td>7/4/1826</td>
</tr>
</tbody>
</table>

Annotations

<h1 itemprop="name">Tom Cruise</h1>

<span itemprop="birthDate">7/3/1962</span>

<span itemprop="gender">Male</span>
1B+ Webpages over the Web

Contribution is skewed: 1-50K
Knowledge Extraction III–Extractors

- Three tasks (any order, maybe combined)
  I. Triple identification
     Tom Cruise
     Thomas Cruise Mapother IV; (July 3, 1962),
     is an American film actor and producer. He has been
     /people/person/person/date_of_birth
  II. Entity linkage
  III. Predicate linkage
Knowledge Extraction III–Extractors

- **Texts/DOM**: distant supervision
  - Example:
    - **Tom Cruise** (born Thomas Cruise Mapother IV; July 3, 1962), is an American film actor and producer. He has been...
    - **Pattern 1**: \( X \) “born” \( Y \) → \((X, /people/person/date_of_birth, Y)\)

- **Web tables/lists**: schema mapping

- **Annotations**: semi-automatic mapping
Statistics for Extractors

- 12 extractors; high variety

<table>
<thead>
<tr>
<th></th>
<th>#Triples</th>
<th>#Webpages</th>
<th>#Patterns</th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT1</td>
<td>274M</td>
<td>202M</td>
<td>4.8M</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td>TXT2</td>
<td>31M</td>
<td>46M</td>
<td>3.7M</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>TXT3</td>
<td>8.8M</td>
<td>16M</td>
<td>1.5M</td>
<td>0.25</td>
<td>0.81</td>
</tr>
<tr>
<td>TXT4</td>
<td>2.9M</td>
<td>1.2M</td>
<td>0.1M</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td>DOM1</td>
<td>804M</td>
<td>344M</td>
<td>25.7M</td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td>DOM2</td>
<td>431M</td>
<td>925M</td>
<td>No pat.</td>
<td>0.09</td>
<td>0.62</td>
</tr>
<tr>
<td>DOM3</td>
<td>45M</td>
<td>N/A</td>
<td>No pat.</td>
<td>0.58</td>
<td>0.93</td>
</tr>
<tr>
<td>DOM4</td>
<td>52M</td>
<td>7.8M</td>
<td>No pat.</td>
<td>0.26</td>
<td>0.34</td>
</tr>
<tr>
<td>DOM5</td>
<td>0.7M</td>
<td>0.5M</td>
<td>No pat.</td>
<td>0.13</td>
<td>No conf.</td>
</tr>
<tr>
<td>TBL1</td>
<td>3.1M</td>
<td>0.4M</td>
<td>No pat.</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>TBL2</td>
<td>7.4M</td>
<td>0.1M</td>
<td>No pat.</td>
<td>0.69</td>
<td>No conf.</td>
</tr>
<tr>
<td>ANO</td>
<td>145M</td>
<td>53M</td>
<td>No pat.</td>
<td>0.28</td>
<td>0.30</td>
</tr>
</tbody>
</table>

As of 11/2013
Errors Can Creep in at Every Stage

Extraction error: (Obama, nationality, Chicago)
Reconciliation error: (Obama, nationality, North America)
Source data error: (Obama, nationality, Kenya)
Knowledge Extraction IV–Quality

- Gold standard: Freebase
- LCWA (Local Closed-World Assumption)
  - If $(s, p, o)$ exists in FB: true
  - Otherwise,
    - If $(s, p)$ exists in FB: false (Freebase knowledge is locally complete)
    - Otherwise: UNKNOWN
- The gold standard contains about 40% of the triples
Statistics for Triple Correctness

- Overall accuracy: 30%
- Random sample on 25 false triples
  - Triple-identification errors: 11 (44%)
  - Entity-linkage errors: 11 (44%)
  - Predicate-linkage errors: 5 (20%)
  - Source-data errors: 1 (4%)
Statistics for Triple Correctness

As of 11/2013

- Bar graph: Triple Accuracy by #Extractors
- Line graph: Triple Accuracy by #URLs
### Statistics for Extractors

- **12 extractors; high variety**

<table>
<thead>
<tr>
<th></th>
<th>#Triples</th>
<th>#Webpages</th>
<th>#Patterns</th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT1</td>
<td>274M</td>
<td>202M</td>
<td>4.8M</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td>TXT2</td>
<td>31M</td>
<td>46M</td>
<td>3.7M</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>TXT3</td>
<td>8.8M</td>
<td>16M</td>
<td>1.5M</td>
<td>0.25</td>
<td>0.81</td>
</tr>
<tr>
<td>TXT4</td>
<td>2.9M</td>
<td>1.2M</td>
<td>0.1M</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td>DOM1</td>
<td>804M</td>
<td>344M</td>
<td>25.7M</td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td>DOM2</td>
<td>431M</td>
<td>925M</td>
<td>No pat.</td>
<td>0.09</td>
<td>0.62</td>
</tr>
<tr>
<td>DOM3</td>
<td>45M</td>
<td>N/A</td>
<td>No pat.</td>
<td>0.58</td>
<td>0.93</td>
</tr>
<tr>
<td>DOM4</td>
<td>52M</td>
<td>7.8M</td>
<td>No pat.</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>DOM5</td>
<td>0.7M</td>
<td>0.5M</td>
<td>No pat.</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>TBL1</td>
<td>3.1M</td>
<td>0.4M</td>
<td>No pat.</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>TBL2</td>
<td>7.4M</td>
<td>0.1M</td>
<td>No pat.</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>ANO</td>
<td>145M</td>
<td>53M</td>
<td>No pat.</td>
<td>0.28</td>
<td></td>
</tr>
</tbody>
</table>

As of 11/2013
I. Knowledge extraction

II. Knowledge fusion

III. Interesting applications

IV. Future directions
Goal: Judge Triple Correctness

- Input: Knowledge triples and their provenances (i.e., which extractor extracts from which source)
- Output: a probability in \([0,1]\) for each triple
  - Probabilistic decisions vs. deterministic decisions
Usage of Probabilistic Knowledge

- Active learning, probabilistic inference, etc.
- Negative training examples, and
- MANY EXCITING APPLICATIONS!!
- Upload to KG
## Data Fusion—Definition

<table>
<thead>
<tr>
<th>Data items</th>
<th>Sources</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td>( S_1 )</td>
<td></td>
</tr>
<tr>
<td>( D_2 )</td>
<td>( S_2 )</td>
<td></td>
</tr>
<tr>
<td>( D_3 )</td>
<td>( \ldots )</td>
<td></td>
</tr>
<tr>
<td>( D_M )</td>
<td>( S_N )</td>
<td></td>
</tr>
</tbody>
</table>

**Input**

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( \ldots )</th>
<th>( S_N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_M )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Output**

<table>
<thead>
<tr>
<th></th>
<th>Truths</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td></td>
</tr>
<tr>
<td>( D_2 )</td>
<td></td>
</tr>
<tr>
<td>( D_3 )</td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
</tr>
<tr>
<td>( D_M )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Src1</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Jagadish</td>
<td>UM</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
</tr>
<tr>
<td>Franklin</td>
<td>UCB</td>
</tr>
</tbody>
</table>
## Data Fusion–Intuition

<table>
<thead>
<tr>
<th></th>
<th>Src1</th>
<th>Src2</th>
<th>Src3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jagadish</td>
<td>UM</td>
<td>ATT</td>
<td>UM</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UW</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>ATT</td>
<td>BEA</td>
</tr>
<tr>
<td>Franklin</td>
<td>UCB</td>
<td>UCB</td>
<td>UMD</td>
</tr>
</tbody>
</table>

**Voting--Trust the majority.**
## Data Fusion–Intuition

<table>
<thead>
<tr>
<th></th>
<th>Src1</th>
<th>Src2</th>
<th>Src3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jagadish</td>
<td>UM</td>
<td>ATT</td>
<td>UM</td>
</tr>
<tr>
<td>Dewitt</td>
<td>MSR</td>
<td>MSR</td>
<td>UW</td>
</tr>
<tr>
<td>Bernstein</td>
<td>MSR</td>
<td>MSR</td>
<td>MSR</td>
</tr>
<tr>
<td>Carey</td>
<td>UCI</td>
<td>ATT</td>
<td>BEA</td>
</tr>
<tr>
<td>Franklin</td>
<td>UCB</td>
<td>UCB</td>
<td>UMD</td>
</tr>
</tbody>
</table>
Quality-based--Give higher votes to more accurate sources.
Q1. How to compute source accuracy?

- Source Accuracy: $A(S)$
  
  $A(S) = \text{Avg} P(v)_{v \in \tilde{V}(S)}$

- $\tilde{V}(S)$ - values provided by S
- $P(v)$ - pr of value v being true
Data Fusion–A Bayesian Model

According to the Bayes Rule, we need to know $Pr(\Phi|v_i \text{ true})$

- Assuming independence of sources, we need to know $Pr(\Phi(S) |v_i \text{ true})$
- If $S$ provides $v_i$: $Pr(\Phi(S) |v_i \text{ true}) = A(S)$
- If $S$ does not provide $v_i$: $Pr(\Phi(S) |v_i \text{ true}) = (1-A(S))/n$

Q2. How to leverage accuracy in voting?

[Q2. How to leverage accuracy in voting?]

Input:
- Data item $D$
- $\text{Dom}(D) = \{v_0, v_1, \ldots, v_n\}$
- Observation $\Phi$ on $D$

Output:
$Pr(v_i \text{ true}|\Phi)$ for each $i=0, \ldots, n$ (sum up to 1)
Q3. How to handle interdependence between source accuracy and value pr?

- Continue until source accuracy converges

**Source Accuracy**

\[ A(S) = \text{Avg} \ Pr(v(D) \mid \Phi) \]

**Value Probability**

\[ Pr(v(D) \mid \Phi) = \frac{e^{C(v(D))}}{\sum_{v_0 \in \text{rel}(D)} e^{C(v_0(D))}} \]

**Source Vote Count**

\[ A'(S) = \ln \frac{nA(S)}{1 - A(S)} \]

**Value Vote Count**

\[ C(v(D)) = \sum_{S \in S(v(D))} A'(S) \]
Knowledge Fusion Challenges

I. Input is *three-dimensional*

(a) Data fusion input

(b) Knowledge fusion input
Knowledge Fusion Challenges

II. Output prs should be *well-calibrated*
Knowledge Fusion Challenges

III. Data are of *Web-scale*

- Three orders of magnitude larger than currently published data-fusion applications
  - Size: 1.1TB
  - Sources: 170K→1B+
  - Data items: 400K→375M
  - Values: 18M→6.4B (1.6B unique)

- Data are highly skewed
  - #Triples/Data-item: 1 - 2.7M
  - #Triples/Source: 1 - 50K
Knowledge Fusion Solutions

- Treat each (URL, Extractor) as a whole (provenance) for accuracy evaluation
- A series of refinements to improve probability calibration
- MapReduce Based Framework
  - Terminate in 5 rounds
  - Sample for too big data items or provenances
Basic Sonya Solution vs. Voting

The curve is closer to the ideal than naive voting
Refinement I. Ignore Low-Coverage Provenances

Calibration Curves for Cumulative Improvements

Coverage: 1→ .918

Smoother curve
Refinement II. Granularity (URL->Site, Extractor->Pattern, Predicate)

Coverage: .918 → .993

Much higher accuracy

Much higher coverage
Refinement III. Ignore Low-Accuracy Provenances

Calibration Curves for Cumulative Improvements

Real probability (accuracy) vs. Predicted probability

Closer to ideal curve
Refinement IV. Initiate Provenance Accuracy by FB

Calibration Curves for Cumulative Improvements

- Basic
- + FilterByCov
- + AccuGranularity
- + FilterByAccu
- + InitAccuByGS

Coverage: .993 → .994

- Closer to ideal curve
- Smoother curve
Precision-Recall Curve

PR-Curves for Various Models

- Sonya
- Sonya+
- Sonya+(FB)
- Vote
Analysis of Errors

Future Directions!!!
One Inherent Limitation

Cannot distinguish errors from extractors and from sources
Other Fusion Techniques

- **Ex:** Adaboost learning from extractions
- **Prior:** (A, parent_of, C), (B, parent_of, C) → (A, spouse_of, B)
Outline

I. Knowledge extraction
   - Extractor
   - Extractor
   - Extractor
   - Web
   - TXT
   - DOM
   - TBL
   - ANO

II. Knowledge fusion
   - prKB
   - Fusion

III. Interesting applications

IV. Future directions
Usage of Probabilistic Knowledge

- Source errors: trustworthiness evaluation
- Extraction errors: data abnormality diagnosis

Negative training examples, and MANY EXCITING APPLICATIONS!!
Application I. A New Angle to Evaluate Web Source Quality

- What we have now
  - Page Rank: links between Websites/Webpages
  - Log based: search log and click-through rate
  - Web spam
  - etc.
Popular Sources w. High Page Rank May Spread Gossip

14 out of 15 Gossip Websites have high page rank

http://www.ebizmba.com/articles/gossip-websites
Tale Sources w. Low Page Rank May Provide Valuable Info
Tale Sources w. Low Page Rank May Provide Valuable Info

Good WebAnswer for an award-winning song
Tale Sources w. Low Page Rank
May Provide Valuable Info

Missing WebAnswer for a not-so-popular song
Tale Sources w. Low Page Rank May Provide Valuable Info

Very precise info on guitar players but low Page Rank
Application I. A New Angle to Evaluate Web Source Quality

Fact 1
Fact 2
Fact 3 ✘
Fact 4
Fact 5 ✘
Fact 6
Fact 7
Fact 8
Fact 9
Fact 10 ✘

Accu: 0.7
Application I. A New Angle to Evaluate Web Source Quality

How to decide if a triple is indeed claimed by the source instead of an extraction error?
**Application I. A New Angle to Evaluate Web Source Quality**

<table>
<thead>
<tr>
<th>Triple</th>
<th>Extraction Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>0.3</td>
<td>1.0</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Accu 0.73**
Extraction and Triple Correctness

Example. (Obama, nationality, ?)

(Obama, nationality, Bolivarianism) (many many such objects)

- 3 extractions (Pr_extCorr=0.01)
  
  http://mathaba.net/news/?x=631316
  http://www.laht.com/article.asp?ArticleId=329187&CategoryId=10717

- Pr_tripleCorr=0
Extraction and Triple Correctness

Example. (Obama, nationality, ?)

(Obama, nationality, Kenya)

● 2087 extractions:
  ○ Example of a correct extraction (Pr_extCorr=0.792):
    
    ![Correct Extraction Example](http://beforeitsnews.com/obama-birthplace-controversy/2013/04/alabama-supreme-court-chief-justice-roy-moore-to-preside-over-obama-eligibility-case-2458624.html)

  ○ Example of a wrong extraction (Pr_extCorr=0.130):
    
    ![Wrong Extraction Example](http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxaqg/-/index.html)

● Pr_tripleCorr=0 (not enough support)
Extraction and Triple Correctness Example. (Obama, nationality, ?)

(Obama, nationality, USA)

- 2481 extractions:
  - Example of a correct extraction (Pr_extCorr=0.999):
  - Example of a wrong extraction (Pr_extCorr=0.261):

- Pr_tripleCorr=1 (Higher support)
Distribution of providers for Kenya and USA
### Sonya Trustworthiness Score

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Triples</th>
<th>Sonya Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.eonline.com">www.eonline.com</a></td>
<td>12,871</td>
<td>0.363</td>
</tr>
<tr>
<td>perezhilton.com</td>
<td>46,912</td>
<td>0.427</td>
</tr>
<tr>
<td>radaronline.com</td>
<td>3,530</td>
<td>0.489</td>
</tr>
<tr>
<td><a href="http://www.zimbio.com">www.zimbio.com</a></td>
<td>2,464,452</td>
<td>0.530</td>
</tr>
<tr>
<td>mediatakeout.com</td>
<td>131</td>
<td>0.531</td>
</tr>
<tr>
<td>gawker.com</td>
<td>6,055</td>
<td>0.567</td>
</tr>
<tr>
<td><a href="http://www.popsugar.com">www.popsugar.com</a></td>
<td>1,805</td>
<td>0.576</td>
</tr>
<tr>
<td><a href="http://www.people.com">www.people.com</a></td>
<td>16,886</td>
<td>0.585</td>
</tr>
<tr>
<td><a href="http://www.tmz.com">www.tmz.com</a></td>
<td>8,149</td>
<td>0.621</td>
</tr>
<tr>
<td><a href="http://www.fishwrapper.com">www.fishwrapper.com</a></td>
<td>14</td>
<td>0.622</td>
</tr>
<tr>
<td>celebrity.yahoo.com</td>
<td>11,187</td>
<td>0.677</td>
</tr>
<tr>
<td>wonderwall.msn.com</td>
<td>2,524</td>
<td>0.684</td>
</tr>
<tr>
<td>hollywoodlife.com</td>
<td>4,536</td>
<td>0.689</td>
</tr>
<tr>
<td><a href="http://www.wetpaint.com">www.wetpaint.com</a></td>
<td>19,284</td>
<td>0.730</td>
</tr>
</tbody>
</table>

Many gossip Web sites DO provide quite a lot of wrong factual information.
Sonya Trustworthiness Score

● Example for (URL, Predicate)

  URL: https://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html
  Predicate: date_of_birth
  - #Facts = 42; Trustworthiness = 0.95
Celebrity Birthdays On March 22

-Marcel Marceau

Below are 124 famous people born on March 22.

Browse Gift Ideas - Browse Ecards

The names in brackets below are duplicate entries.

Aaron North was born on March 22, 1979. American guitarist.
Amy Stud was born on March 22, 1986. English singer-songwriter and musician.
Andreas Johnson was born on March 22, 1970. Swedish pop and rock singer-songwriter and musician.
Andrew Lloyd Webber was born on March 22, 1948. British composer of musicals.
Angelo Badalamenti was born on March 22, 1937. American composer.
Anja Kling was born on March 22, 1970. German actress.
Annabelle Apsion was born on March 22, 1963. English actress.
Anne Hyde was born on March 22, 1638. Wife of James II of England.
Anthony van Dyck was born on March 1599. Flemish Baroque artist.
Armin Hary was born on March 22, 1937. German athlete.
Avraham Fried was born on March 22, 1959. American singer-songwriter and musical entertainer.
Example for (URL, Predicate)

URL: https://ibirthdayworld.blogspot.com/2010/03/celebrity-birthdays-on-march-22.html
Predicate: date_of_birth
- #Facts = 42; Trustworthiness = 0.95
- Mistake: Anne Hyde (the URL says: 3/22/1638; Wiki/KG says: 3/12/1637)

Anne Hyde

Anne Hyde was Duchess of York and Albany as the first wife of James, Duke of York, later King James II and VII. Originally Anglican, her father was a lawyer. Wikipedia

Born: March 12, 1637, Windsor, United Kingdom
Died: March 31, 1671, London, United Kingdom
Spouse: James II of England (m. 1660–1671)
Children: Anne, Queen of Great Britain, Mary II of England, More
Parents: Edward Hyde, 1st Earl of Clarendon, Frances Hyde, Countess of Clarendon
Siblings: Laurence Hyde, 1st Earl of Rochester, Henry Hyde, 2nd Earl of Clarendon
Application II. Provide An X-Ray for Extracted Data

- **Goal:** Help users analyze errors, changes and abnormalities in data
- **Intuitions:** cluster errors by features and return clusters with top error rates
Application II. Provide An X-Ray for Extracted Data

● Cluster I.
  ○ Feature: (besoccor.com, date_of_birth, 1986_02_18)
  ○ #Triples: 630; Errs: 100%
  ○ Reason: default value

● Cluster 2.
  ○ Feature: (ExtractorX, pred: namesakes, obj:the county)
  ○ #Triples: 4878; Errs: 99.8%
  ○ E.g., [Salmon P. Chase, namesakes, The County]
  ○ Contexts: The county was named for Salmon P. Chase, former senator and governor of Ohio
  ○ Reason: Unresolved coreference
Outline

I. Knowledge extraction
   - Extractor
   - Extractor
   - Extractor
   - Web
   - TXT
   - DOM
   - TBL
   - ANO

II. Knowledge fusion
   - prKB
   - Fusion

III. Interesting applications

IV. Future directions
Future Directions: Remove the Assumptions One by One

Assumption I. Independence between pairs of provenances (i.e., (URL, extractor))
Future Directions: Remove the Assumptions One by One

Assumption II. Single true object for each (sub, pred)
Future Directions: Remove the Assumptions One by One

Assumption III. Extractions are deterministic
Future Directions: Remove the Assumptions One by One

Assumption IV. Values (objects) are categorical
Assumption V. We have enough data to judge accuracy of each source
Assumption VI. Local closed-world assumption in evaluation
Future Directions: Remove the Assumptions One by One

Assumption VII. Global closed-world assumption—consider only existing entities and predicates in FB

WE NEED SOMETHING NEW!!!
TAKE AWAYS

● A new area--Knowledge Fusion
● We can solve KF problem fairly well by adapting DF methods
● Many interesting future directions for KF!
● Many exciting applications for the prKB!!
Acknowledgement

Evgeniy Gabrilovich (Manager, need to say anything?)
Geremy Heitz (Strongest supporter)
Wilko Horn (Strictest code reviewer)
Kevin Murphy (Intelligent consultant)
Shaohua Sun (Critical representer to the outside world)
Wei Zhang (Fearless explorer of new ideas)
THANK YOU!

Questions?