

# Unsupervised **Slot Filling Validation** based on Multi-dimensional **Truth-Finding**

Dian Yu<sup>1</sup>, Hongzhao Huang<sup>1</sup>, Taylor Cassidy<sup>2,3</sup>, Heng Ji<sup>1</sup>  
Chi Wang<sup>4</sup>, Shi Zhi<sup>4</sup>, Jiawei Han<sup>4</sup>, Clare Voss<sup>2</sup>, Malik Magdon-Ismael<sup>1</sup>

<sup>1</sup>Computer Science Department, Rensselaer Polytechnic Institute

<sup>2</sup>U.S. Army Research Lab<sup>2</sup> <sup>3</sup>IBM T. J. Watson Research Center

<sup>4</sup>Computer Science Department, University of Illinois at Urbana-Champaign



Rensselaer

# Outline



- Introduction
- Limitations of State-of-the-art
- MTM: A Multi-dimensional Truth-Finding Model
- Response Credibility Initialization
  - Linguistic Indicators
- Experiments



# Outline



- ➔ ■ Introduction
- Limitations of State-of-the-art
- MTM: A Multi-dimensional Truth-Finding Model
- Response Credibility Initialization
  - Linguistic Indicators
- Experiments



# Knowledge Base Population (Slot Filling)

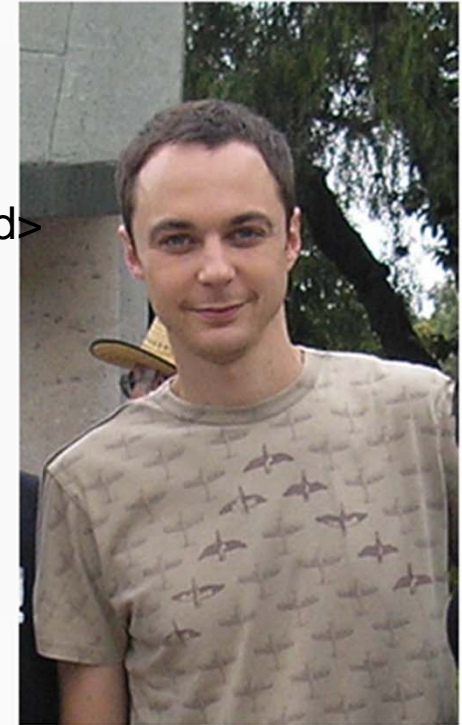


Rensselaer



```
<query id="SF114">  
  <name>Jim Parsons</name>  
  <docid>eng-WL-11-174592-12943233</docid>  
  <enttype>PER</enttype>  
  <nodeid>E0300113</nodeid>  
  <ignore>per:date_of_birth  
    per:age per:country_of_birth  
    per:city_of_birth</ignore>  
</query>
```

Jim Parsons



Parsons in 2008

<b>Born</b>	James Joseph Parsons March 24, 1973 (age 37) <a href="#">Houston, Texas, U.S.</a>
<b>Occupation</b>	Actor
<b>Years active</b>	2000–present

Jim Parsons, a graduate of the **School Attended: University of Houston** and Dance, won the Emmy on Sunday for Lead Actor in a Comedy Series for his work on The Big Bang Theory.



# KB Slots



Person		Organization
per:alternate_names	per:title	org:alternate_names
per:date_of_birth	per:member_of	org:political/religious_affiliation
per:age	per:employee_of	org:top_members/employees
per:country_of_birth	per:religion	org:number_of_employees/members
per:stateorprovince_of_birth	per:spouse	org:members
per:city_of_birth	per:children	org:member_of
per:origin	per:parents	org:subsidiaries
per:date_of_death	per:siblings	org:parents
per:country_of_death	per:other_family	org:founded_by
per:stateorprovince_of_death	per:charges	org:founded
per:city_of_death		org:dissolved
per:cause_of_death		org:country_of_headquarters
per:countries_of_residence		org:stateorprovince_of_headquarters
per:stateorprovinces_of_residence		org:city_of_headquarters
per:cities_of_residence		org:shareholders
per:schools_attended		org:website



# Slot Filling & Slot filler Validation






Rensselaer


- Slot Filling (SF)
  - **Definition:** The slot filling task is to search a document collection to fill in values for predefined slots (attributes) for a given entity to populate a reference KB.
  - **Queries:** 50 person queries and 50 organization queries such as “*Marc Bolland*” and “Public Library of Science”
  - **Response:** Claim + Evidence
  - 41 slot types: single or multiple attribute values
- Slot Filling Validation (SFV)
  - 52 runs from 18 SF teams



# Extracting true claims from multiple sources



- Problems:
  - different information sources may generate claims with varied trustability   
  - various SF systems may generate *erroneous, conflicting, redundant, complementary, ambiguously worded, or inter-dependent* claims from the same set of documents

System	Source	Slot Filler	Evidence
A	Agence France- Presse, News	Los Angeles	The statement was confirmed by publicist Maureen O'Connor, who said <b>Dio</b> died in <b>Los Angeles</b> .
B	New York Times, News	Los Angeles	<b>Ronnie James Dio</b> , a singer with the heavy-metal bands Rainbow, died on Sunday in <b>Los Angeles</b> .
C	Discussion Forum	Atlantic City	<b>Dio</b> revealed last summer that he was suffering from stomach cancer shortly after wrapping up a tour in <b>Atlantic City</b> .
D 	Associated Press Worldstream, News	Los Angeles	<b>LOS ANGELES</b> 2010-05-16 20:31:18 UTC <b>Ronnie James Dio</b> ... has died, according to his wife.

# Solution



## Truth Finding:

**Determine the veracity of multiple conflicting claims from various sources and providers (i.e. systems or humans)**

**TRUTH**





# Outline

- Introduction
- ➔ ▪ Limitations of State-of-the-art
- MTM: A Multi-dimensional Truth-Finding Model
- Response Credibility Initialization
  - Linguistic Indicators
- Experiments



# Truth Finding Problem



- We require not only high-confidence claims but also trustworthy evidence to verify them. → deep understanding is needed.
- Previous truth finding work assumed most claims are likely to be true. Most of them relied on the “wisdom of the crowd”.
  - In SF, 72.02% responses are false.
  - Certain truths might only be discovered by a minority of systems or from a few sources (62% from 1 or 2 systems)



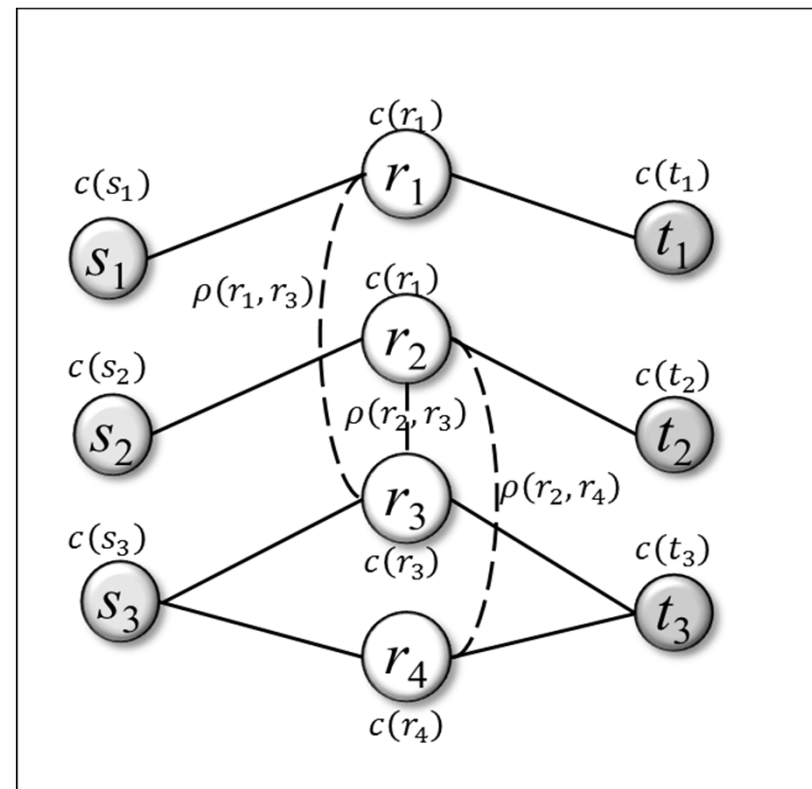
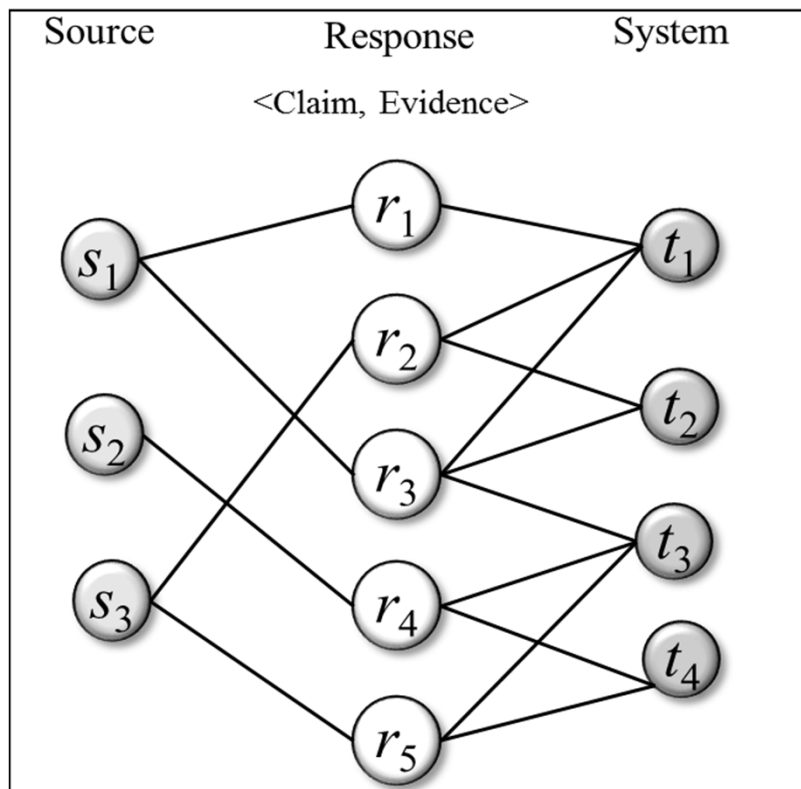
# Outline



- Introduction
- Limitations of State-of-the-art
- ➔ ▪ **MTM: A Multi-dimensional Truth-Finding Model**
- Response Credibility Initialization
  - Linguistic Indicators
- Experiments



# Multi-dimensional truth-finding model (MTM)



# Heuristics Explored in MTM



- **Heuristic 1:**

*A response is more likely to be true if derived from many trustworthy sources. A source is more likely to be trustworthy if many responses derived from it are true.*

- **Heuristic 2:**

*A response is more likely to be true if it is extracted by many trustworthy systems. A system is more likely to be trustworthy if many responses generated by it are true.*



# Credibility Initialization



- Source ( $S$ ):
  - a combination of publication venue and genre
  - initialized uniformly as  $1/n$  ( $n$  is the number of sources)
- System ( $T = \{t_1, \dots, t_l\}$ ):
  - Each system  $t_i$  generates a set of responses  $R_{t_i}$ .
  - Similarity between system  $t_i$  and  $t_j$  is 
$$\frac{|R_{t_i} \cap R_{t_j}|}{\log(|R_{t_i}|) + \log(|R_{t_j}|)}$$
 (Mihalcea, 2004).
  - Construct a weighted undirected graph  $G = \langle T, E \rangle$ ,  $T(G) = T$ ,  $E(G) = \{ \langle t_i, t_j \rangle \}$ ,  $\langle t_i, t_j \rangle = \text{similarity}(t_i, t_j)$
  - Apply TextRank to obtain the initial score.
- Response ( $R$ ):
  - Rely on deep linguistic analysis of the evidence sentences and semantic clues. We will introduce it later.



# Credibility Propagation



- Extension of Co-HITS (Deng et al., 2009)
- Given the initial credibility scores  $c^0(r)$ ,  $c^0(s)$ , and  $c^0(t)$ , we aim to obtain the refined credibility scores  $c(r)$ ,  $c(s)$  and  $c(t)$ .

- Propagation:

- Sources: Consider both the initial score for source and the propagation from connected responses.

$$c(s_i) = (1 - \lambda_{rs})c^0(s_i) + \lambda_{rs} \sum_{r_j \in R} p_{ji}^{rs} c(r_j)$$

- System: Consider both the initial score for system and the propagation from responses to systems

$$c(t_k) = (1 - \lambda_{rt})c^0(t_k) + \lambda_{rt} \sum_{r_j \in R} p_{jk}^{rt} c(r_j)$$

- Response: Each response's score is influenced by both linked sources and systems.

$$c(r_j) = (1 - \lambda_{sr} - \lambda_{tr})c^0(r_j) + \lambda_{sr} \sum_{s_i \in S} p_{ij}^{sr} c(s_i) + \lambda_{tr} \sum_{t_k \in T} p_{kj}^{tr} c(t_k)$$

- Converges and a similar proof to HITS (Peserico and Pretto, 2009)



# Outline

- Introduction
- Limitations of State-of-the-art
- MTM: A Multi-dimensional Truth-Finding Model
- ➔ ▪ **Response Credibility Initialization**
  - Linguistic Indicators
- Experiments
- Remaining Challenges





# Bottleneck: Low Coverage of Patterns



Rensselaer

- Manually crafted/edited patterns: low coverage; expensive
- Bootstrapping: hard to generalize; long-tail distribution

- Typical Dependency patterns for per:place\_of\_birth

- `<Query_PER> nsubjpass-1 born prep_in <Filler_LOC>`
- `<Query_PER> partmod born prep_in <Filler_LOC>`
- `<Query_PER> nsubjpass-1 born prep_on <Filler_LOC>`
- `<Query_PER> rcmmod born prep_in <Filler_LOC>`

- Missing some simple cases

- **Charles Gwathmey [1]** was born on June 19 , 1938 , in **Charlotte [2]** , N.C..

- Dependency path between [1] and [2]:

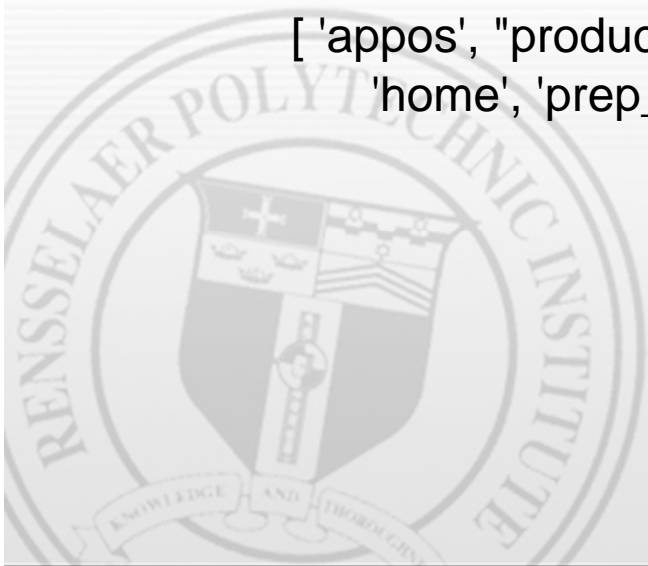
[ 'nsubjpass', 'born', 'prep\_on', 'June', 'prep\_in', 'N.C', 'nn' ) ]



# Bottleneck: Low Coverage of Patterns



- Typical Dependency Patterns for per:place\_of\_death
  - <Q\_PER> nsubj-1 dies prep\_in <A\_LOC>
  - <Q\_PER> nsubj-1 died prep\_in <A\_LOC>
  - <Q\_PER> nsubj-1 died prep\_on <A\_LOC>
  - <Q\_PER> nsubj-1 died prep\_in hospital nn <A\_LOC>
- Missing some simple cases
  - ``60 Minutes'' was the brainchild of [Don Hewitt \[1\]](#), the show 's longtime executive producer who died Wednesday of pancreatic cancer at his home in Bridgehampton, [N.Y. \[2\]](#) , at age 86 .
  - Dependency path between [1] and [2]:  
[ 'appos', "producer", 'nsubj', 'died', "who", 'rcmod', 'died', 'prep\_at', 'home', 'prep\_in' ]



# Knowledge Gap 1



- Deep Knowledge Acquisition: Nominal Coreference
  - Almost overnight, he became fabulously rich, with a \$3-million book deal, a \$100,000 speech making fee, and a lucrative multifaceted consulting business, **Giuliani Partners**. As a celebrity rainmaker and lawyer, his income last year exceeded \$17 million. His consulting partners included seven of those who were with him on 9/11, and in 2002 Alan Placa, his boyhood pal, went to work at **the firm**.
  - After successful karting career in Europe, Perera became part of the **Toyota** F1 Young Drivers Development Program and was a Formula One test driver for **the Japanese company** in 2006.
  - “**Alexandra Burke** is out with the video for her second single ... taken from **the British artist**’s debut album”
  - “a woman charged with running **a prostitution ring** ... **her business, Pamela Martin and Associates**”
- Our Solution: Online knowledge graph construction; enrich paths with semantic annotations and Information Extraction (coreference/relation/event)

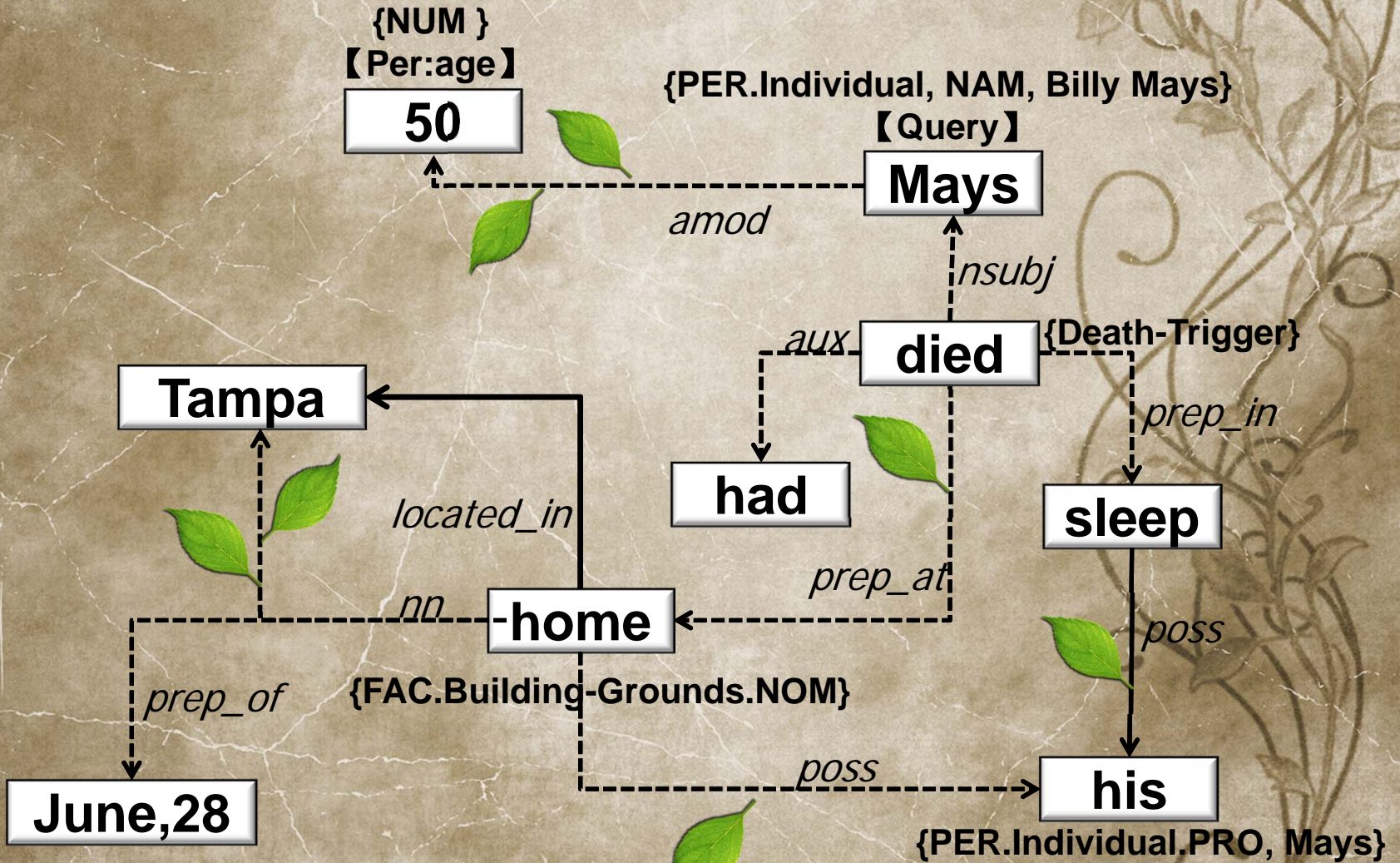
# Knowledge Gap 2



- ❑ Deep Knowledge Acquisition: Implicit paraphrases & long-tail distribution
- ❑ “employee/member”:
  - ❑ **Sutil**, a trained pianist, **tested** for **Midland** in 2006 and raced for Spyker in 2007 where he scored one point in the Japanese Grand Prix.
  - ❑ **Daimler Chrysler** reports 2004 profits of \$3.3 billion; **Chrysler earns** \$1.9 billion.
  - ❑ In her second term, **she received a seat** on the powerful **Ways and Means Committee**
  - ❑ **Jennifer Dunn** was the **face** of the **Washington state Republican Party** for more than two decades
  - ❑ State of Residence: **Davis** became **Virginia's first Republican woman** elected to Congress in 2000, and she was a member of the House Armed Services Committee and the Foreign Affairs Committee
  - ❑ **Buchwald** lied about his age and **escaped into** the **Marine Corps**.
  - ❑ By 1942, Peterson was **performing** with one of Canada's leading big bands, the Johnny Holmes Orchestra.
  - ❑ Even more: “would join”, “would be appointed”, “will start at”, “went to work”, “was transferred to”, “was recruited by”, “took over as”, “succeeded PERSON”, “began to teach piano”, ...
- ❑ “spouse”:
  - ❑ **Buchwald** 's 1952 **wedding** -- **Lena Horne arranged** for it to be held in London 's Westminster Cathedral -- was attended by Gene Kelly , John Huston , Jose Ferrer , Perle Mesta and Rosemary Clooney , to name a few

# Linguistic Indicators:

## Knowledge Graph Construction



# Linguistic Indicators



- Linguistic Indicators: **(binary classification result)**

*Linguistic indicators make use of linguistic features on varying levels - surface form, sentential syntax, semantics, and pragmatics.*

- Node Indicators
- Path Indicators
- Interdependent Claims



# Node Indicators

- Surface: stop words, lowercased
- Entity type, subtype and mention type
  - Fillers for org:top\_employee
  - Fillers for org:website
- Entity attributes mined by the NELL system (Carlson et al., 2010)



# Path Indicators



- Trigger phrases
  - Examples:
    - "top-employees": chief executive officer, chief financial officer, chief operating officer, chief strategy and development officer, chief information officer, e-commerce and security officer,...
    - "headquarters": based, headquarter, headquarters, 's
    - Disease list from medical ontology
- Relations and events:
  - e.g. "Start-Position" indicates slot type: `per:employee_or_member_of`
- Path length:
  - e.g. the path length for `per:title` is usually 1.

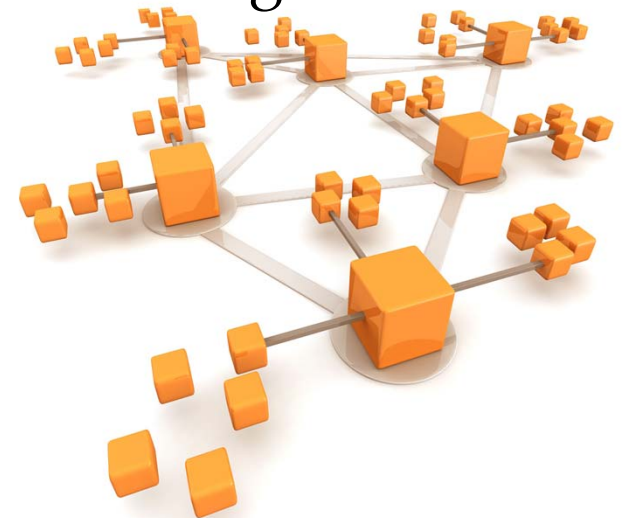




# Independent Claims Indicators

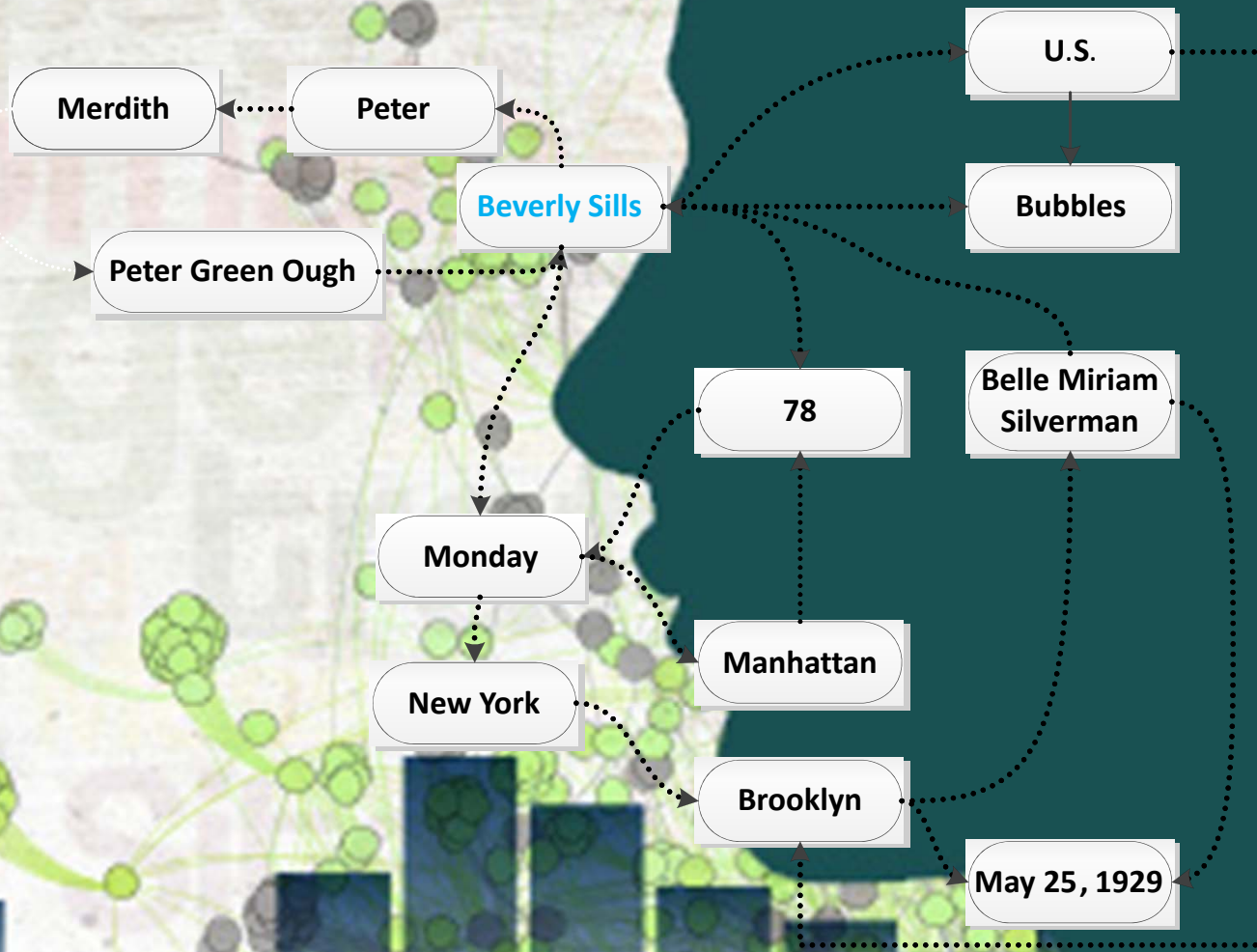


- Conflicting slot fillers
- Inter-dependent slot types:
  - After initial credibility scores for each response, we check whether evidence exists for any *implied* claims.
  - e.g.: Given A is B's son and C is A's sibling brother- $\rightarrow$  A is C's parent.



# Inter-dependent Slots

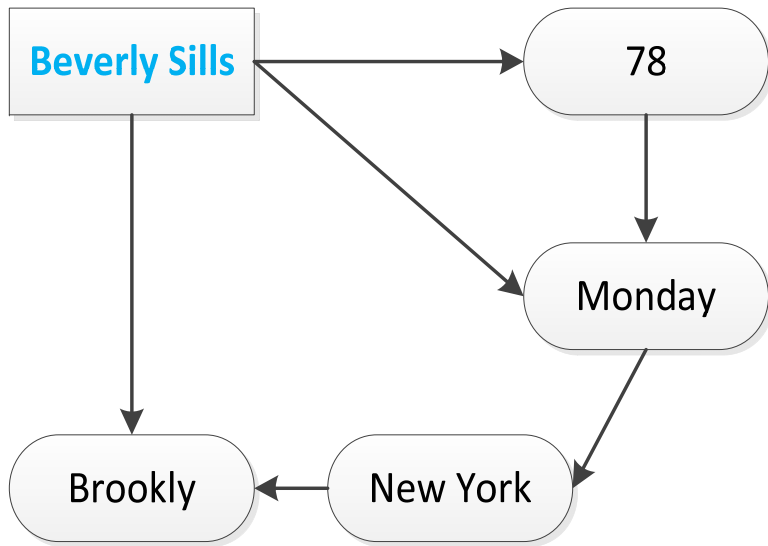
- Query: Beverly Sills



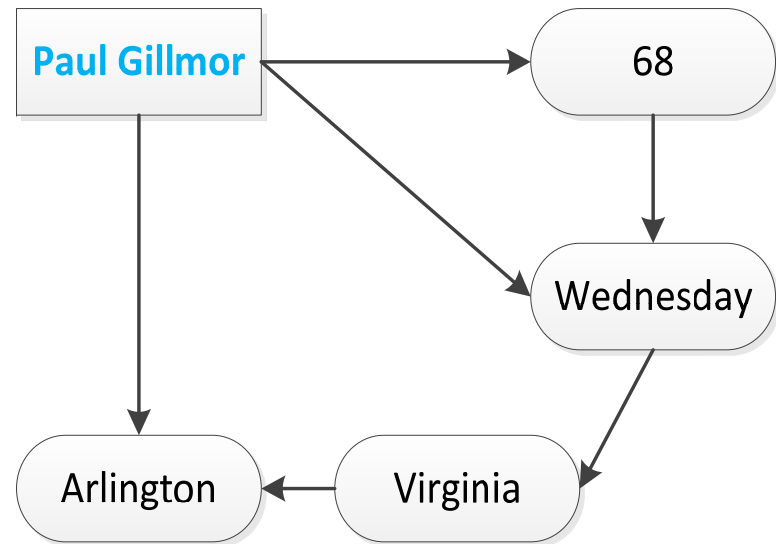


# Example: local structure for death related slots

We already know Beverly Sills, 78, died on Monday in Brookly, NY.



Given the knowledge graph of Paul Gillmor and a similar local structure, we can predict the slot types of nodes .



# Outline



- Introduction
- Limitations of State-of-the-art
- MTM: A Multi-dimensional Truth-Finding Model
- Response Credibility Initialization
  - Hard Constraints

 Experiments



# Truth Finding Overall Performance Rensselaer

Methods	Precision	Recall	F-measure	Accuracy	MAP*
1. Random	28.64%	50.48%	36.54%	50.54%	34%
2. Voting	42.16%	70.18%	52.68%	62.54%	62%
3. Linguistic Indicators	50.24%	70.69%	58.73%	72.29%	60%
4. SVM (3+system+source)	56.59%	48.72%	52.36%	75.86%	56%
5. MTM (3+system+source)	53.94%	72.11%	61.72%	81.57%	70%

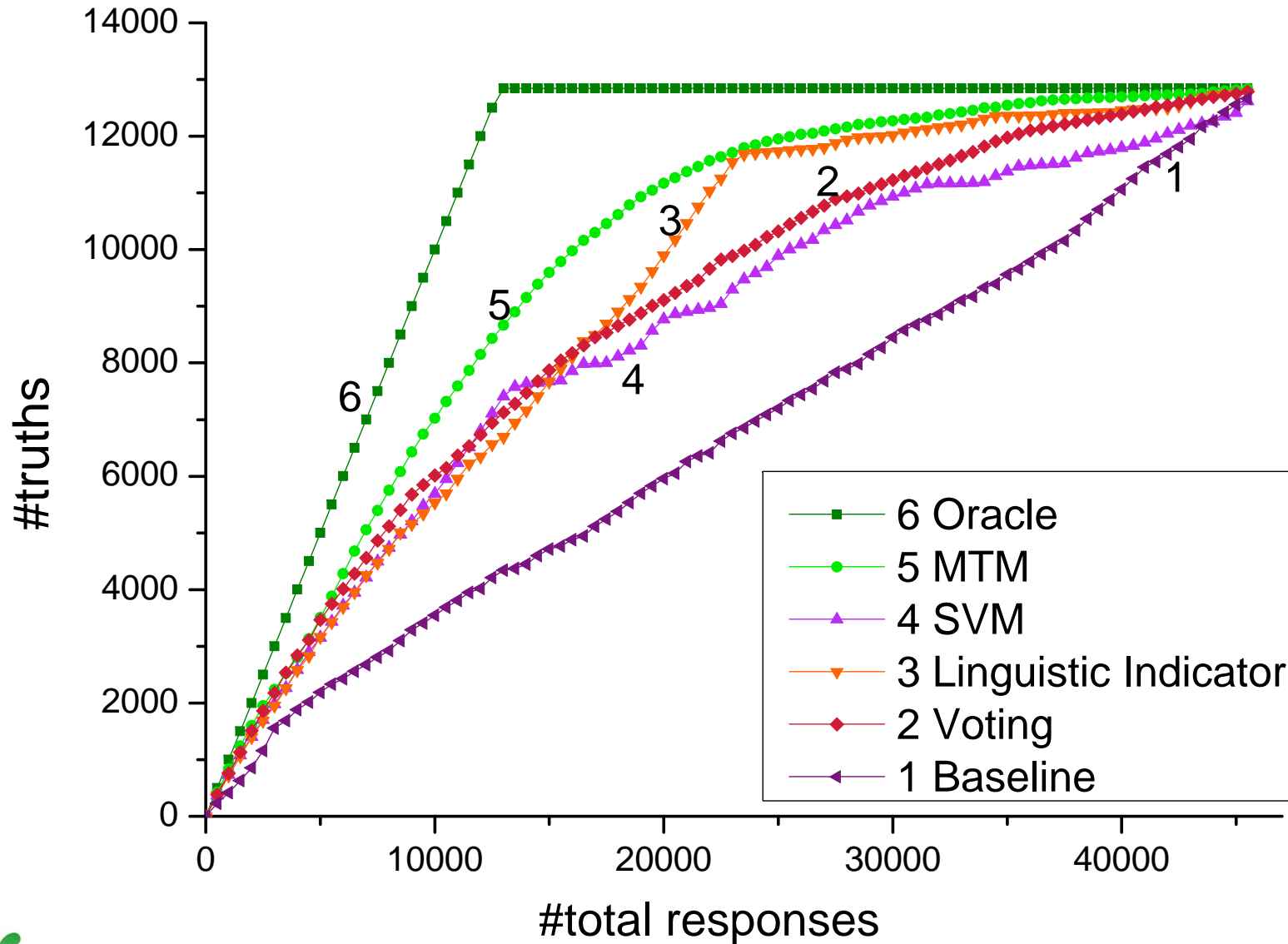


\* MAP: Mean Average Precision

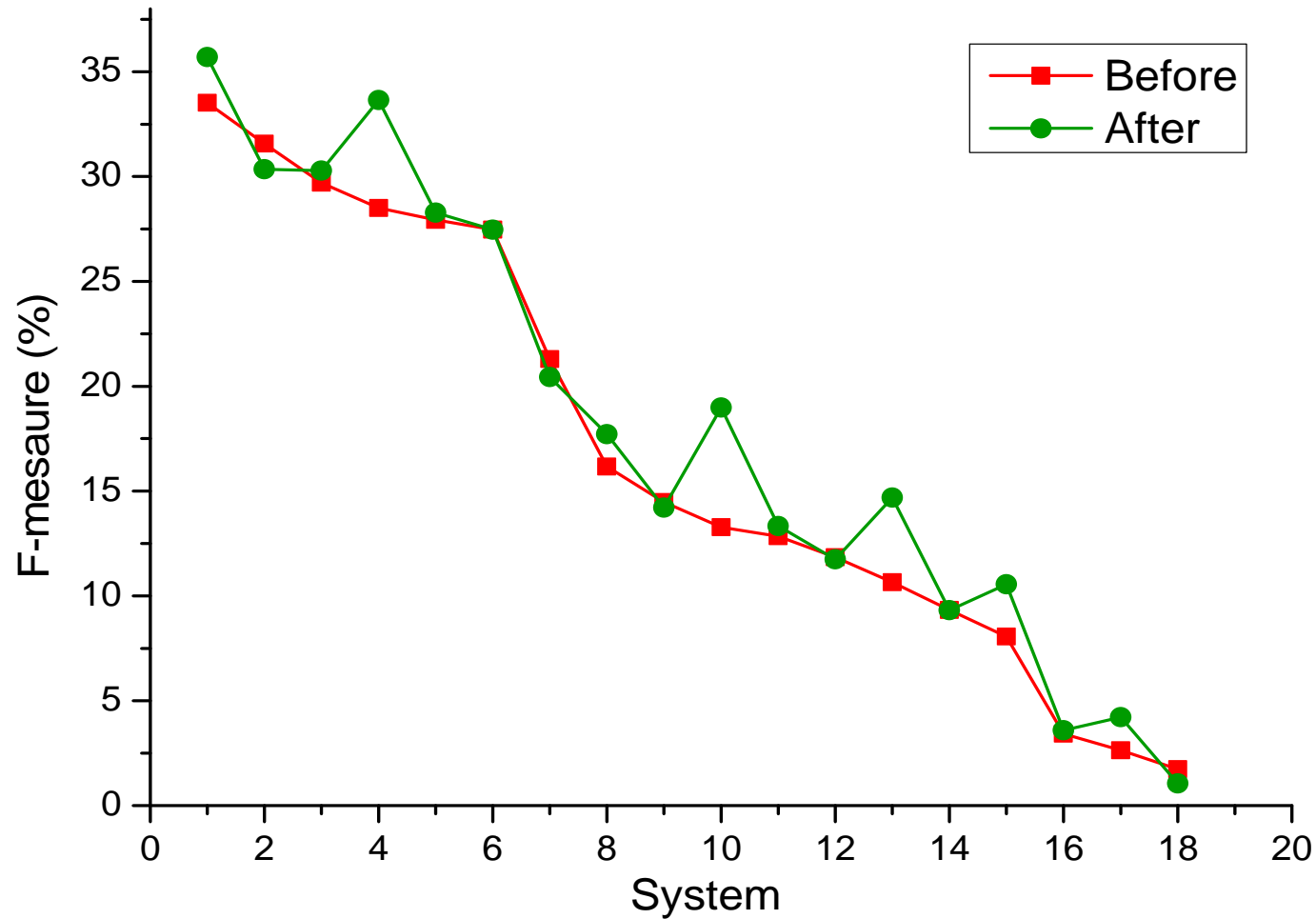
# Truth Finding Efficiency



Rensselaer



# Enhance Individual SF Systems



# Remaining Challenges



- Name Tagging Errors
- Coreference Resolution Errors
  - **He** worked his way up the organization under founder Ted Arison and his son Micky , who now leads **Carnival Corp.** and called **Dickinson**, `` one of the most influential people in the development of the modern-day cruise industry.
  - Indiana Muslim running for Congress wants to combat ignorance about **his** [Andre Carson] faith INDIANAPOLIS -- A convert to Islam stands an election victory away from becoming the second **Muslim** elected to **Congress** and a **role model** for a faith community seeking to make its mark in national politics.
- Vague Justification
  - It was in December 1970 that **Anderson** criticized Hoover 's pretrial attack on two **Roman Catholic** priests , Daniel J. and Philip F. Berrigan , who were later convicted of destroying draft board records. → religion filler?
- Fuzzy Definition
  - **She** and **Russell Simmons**, 50, **have two daughters**: 8-year-old Ming Lee and 5-year-old Aoki Lee.



# Remaining Challenges



- Distinguish Slot Directions
  - Organization parent/subsidiary; members/member\_of
- Implicit Relations
  - He [**Pascal Yoadimnadj**] has been evacuated to France on Wednesday after falling ill and slipping into a coma in Chad, Ambassador Moukhtar Wawa Dahab told The Associated Press. His wife, who accompanied Yoadimnadj to **Paris**, will repatriate his body to Chad, the amba. → is he dead? in Paris?
  - Until last week, **Palin** was relatively unknown outside **Alaska**, and as facts have dribbled out about her, the McCain campaign has insisted that its examination of her background was thorough and that nothing that has come out about her was a surprise. → does she live in Alaska?
  - The list says that the state is owed \$2,665,305 in personal income taxes by singer **Dionne Warwick** of South Orange, **N.J.**, with the tax lien dating back to 1997. → does she live in NJ?
  - **Vernon Bellecourt** -- whose Ojibwe name, WaBun-Inini, means "Man of Dawn" or "Daybreak" -- was born on the White Earth Indian Reservation in **Minnesota**. He left home at 15 after finding work in a carnival. → did he live in Minnesota?

# Thank You

