UNIVERSITY OF TEXAS ARLINGTON

iDiR

Restoring Trust by Computing: Data-driven Fact-checking and Exceptional Fact Finding

Chengkai Li
Associate Professor and Associate Chair
Director, Innovative Database and Information Systems (IDIR) Lab
CSE Department, UT-Arlington

IBM Research - Almaden
March 13, 2019

Outline

- **A Brief History of Our Computational Journalism Research**
- Computational Journalism
  - Data-driven fact-checking (ClaimBuster)
  - Other ongoing fact-checking projects
  - Exceptional fact finding (FactWatcher, Maverick)
- Graph Data Usability (Orion, GQBE, TableView, Maverick)

How it Started

Chris Paul had 16 points, 10 rebounds, 13 assists and five steals..... The only other active player to have such a game is Jason Kidd...

2009

How did they come up with that?

How it Started



"**Developing the Field of Computational Journalism**" Workshop, July 27-31, 2009


Let's work on this ***fact-finding*** problem.

This is **computational journalism**.

Summer 2010

#@%&!#!%


Computational Journalism

By Sarah Cohen, James T. Hamilton, Fred Turner
Communications of the ACM, October 2011, Vol. 54 No. 10, Pages 66-71

CIDR 2011

The **CCC Outrageous Ideas and Vision Award** papers (chosen by audience voting) were:

- **3rd Place:** Computational Journalism: A Call to Arms to Database Researchers by Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu

Computational Journalism: A Call to Arms to Database Researchers

Sarah Cohen
School of Public Policy
Duke Univ.
sarah.cohen@duke.edu

Chengkai Li
Dept. of Comp. Sci. & Eng.
Univ. of Texas at Arlington
cli@uta.edu

Jun Yang
Dept. of Comp. Sci.
Duke Univ.
junyang@cs.duke.edu

Cong Yu
Google Inc.
congy@umich.edu

fact-finding

Interestingly, there is one area of news where a specialized reporter's black box has been immensely successful—sports. It is amazing how commentaries of the form "player X is the second since year Y to record, as a reserve, at least α points, β rebounds, γ assists, and δ blocks in a game" can be generated seemingly instantaneously. Replicating this success for investigative

fact-checking

We will be able to *fact-check* stories quickly. Fact-checking exposes misinformation by politicians, corporations, and special-interest groups, and guards against errors and shady practices in reporting. Consider the following example from FactCheckED.org

TIMING IS EVERYTHING

The US stock market is off to its best start of the year since the early 1990s

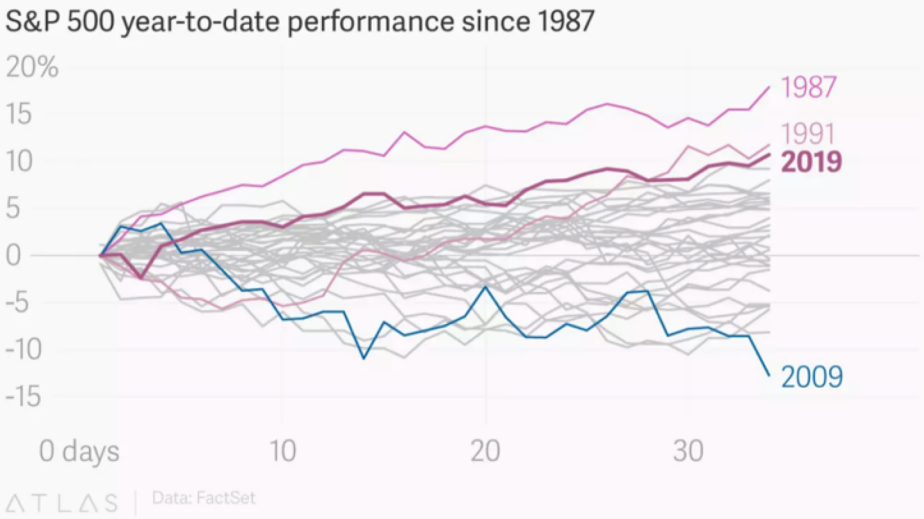
By Jason Karaian • February 17, 2019

Sign up for the Quartz Obsession email

Enter your email

Sign me up

Stay updated about Quartz products and events.



Stocks book their worst year since the financial crisis and worst December since the Great Depression

Jonathan Garber

Dec. 31, 2018, 08:30 AM

SHARE

Universal

The “Fake News” Problem

It was always there, since “Yellow Journalism” started in 1890s.

- Exaggerated headlines, clickbait, computational propaganda, misinformation, disinformation
- Snopes.com (1994), Glenn Kessler (1996), FactCheck.org (2003), PolitiFact (2007)

But it became a daily talking point since 2016.

- Pizzagate
- “filter bubble” and “echo chamber” exacerbated by social media
- The many false claims in political discourses
- Russian meddling with U.S. election
- Twitter bots, Facebook ads, trendy topic algorithms

Outline

- A Brief History of our Computation + Journalism Research
- Computational Journalism
  - **Data-driven fact-checking (ClaimBuster)**
  - Other ongoing fact-checking projects
  - Exceptional fact finding (FactWatcher, Maverick)
- Graph Data Usability (Orion, GQBE, TableView, Maverick)

Toward the Holy Grail of Automated Fact-checking

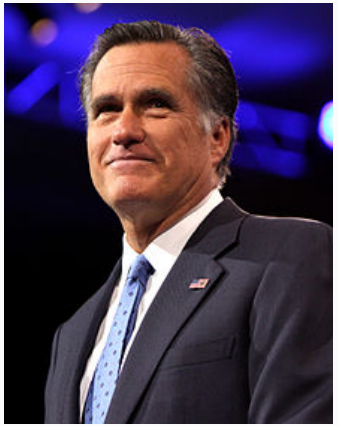
idir.uta.edu/claimbuster

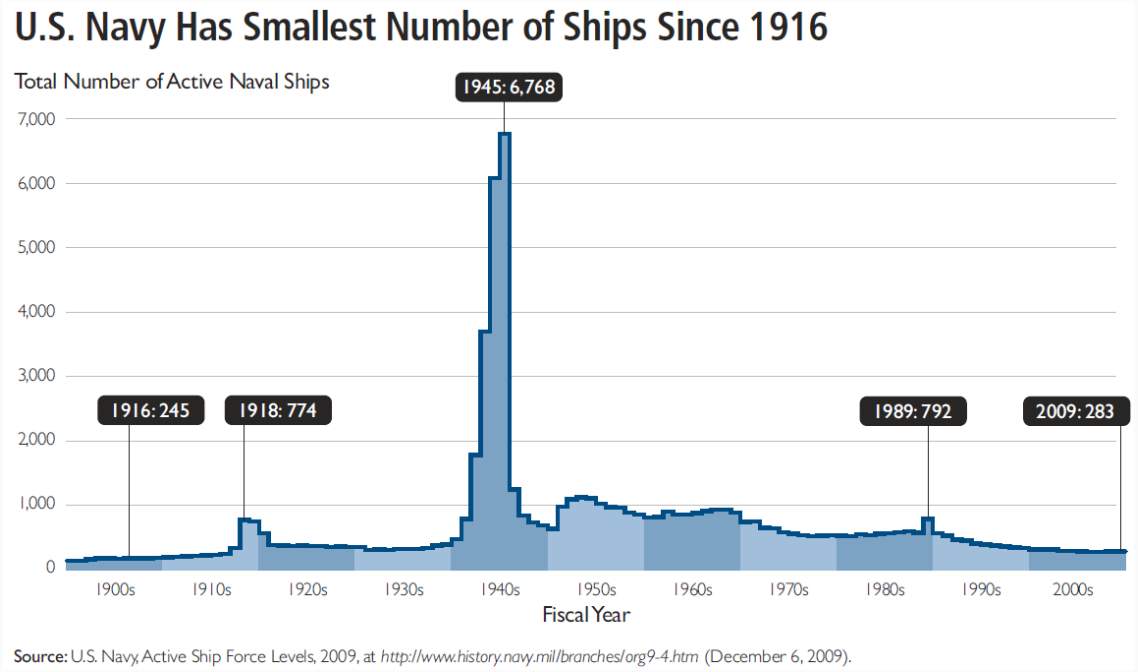
The Holy Grail: Automated, Live Fact-Checking

Source: Bill Adair

Fact-checking is Hard

"… our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in third presidential debate in 2012.

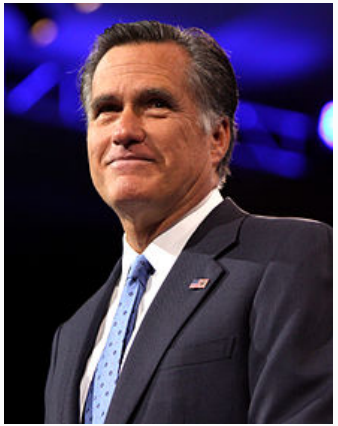




http://en.wikipedia.org/wiki/Mitt_Romney
http://s3.amazonaws.com/thf_media/2010/pdf/Military_chartbook.pdf

Fact-checking is Hard

"… our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in a Republican presidential debate in 2012.



The U.S. military is at risk of losing its "military superiority" because "our Navy is smaller than it's been since 1917. Our Air Force is smaller and older than any time since 1947."

— *Mitt Romney on Monday, January 16th, 2012 in a Republican presidential debate in Myrtle Beach, S.C.*

PANTS ON FIRE!

POLITIFACT

AT THE POYNTER INSTITUTE | WINNER OF THE PULITZER PRIZE

VS

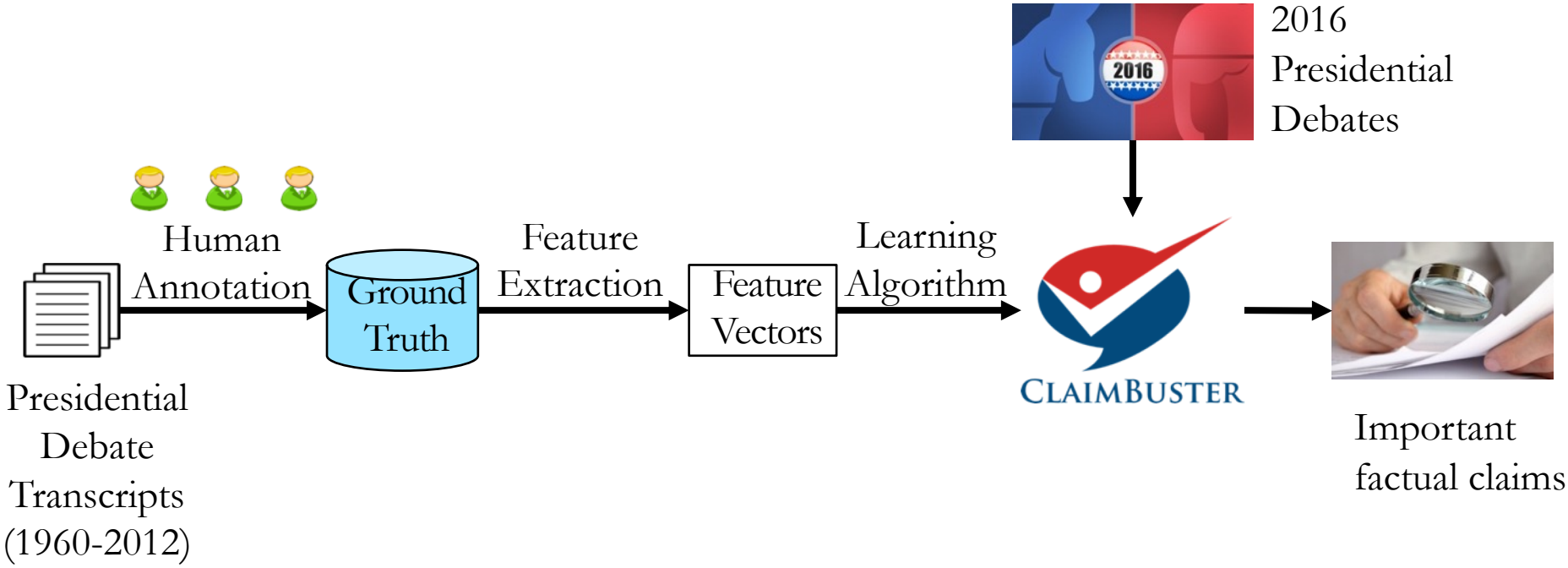
http://en.wikipedia.org/wiki/Mitt_Romney
http://s3.amazonaws.com/thf_media/2010/pdf/Military_chartbook.pdf
http://en.wikipedia.org/wiki/United_States_Navy

PolitiFact "Buffet" of Factual Claims

	A	B	C	D	E
1	**Statement**	**Speaker**	**Date**	**Location/Source**	**Link**
232	I haven't heard one (republican) say that they're in	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
233	concentrated wealth and capital have been put into	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
234	it is not true that regulation holds poor people down or	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
235	Every three weeks we bring online as much as solar	Obama	4/18/2015	Weekly address	https://www.whitehouse.gov
236	Our carbon polution has fallen by 10 percent since	Obama	4/19/2015	Weekly address	https://www.whitehouse.gov
237	"Benghazi has had more hearings, more documents	Claire McCaskill	4/19/2015	ABC this week	http://thehill.com/policy/inte
238	With fast track trade " I would be receiving the same	Obama	4/17/2015	remarks joint press	https://www.whitehouse.gov
239	Loretta Lynch "has been now sitting there longer than	Obama	4/17/2015	remarks joint press	https://www.whitehouse.gov
240	I just counted 19 Republicans who are likely to run for	chris cillizza	4/20/2015	the twitters	https://twitter.com/TheFix/st
241	He [Rand Paul] wanted to cut all our defense	John McCain	4/20/2015	Fox News	https://www.youtube.com/w

Bill Adair is the founder of the Pulitzer Prize-winning website PolitiFact and Knight Professor of the Practice of Journalism and Public Policy at Duke University, where

Finding Important Factual Claims: A Supervised Learning Task



Classification and Ranking by Check-worthiness

CFS: Important factual claims

"We spend less on the military today than at any time in our history." "The President's position on gay marriage has changed." "More people are unemployed today than four years ago."

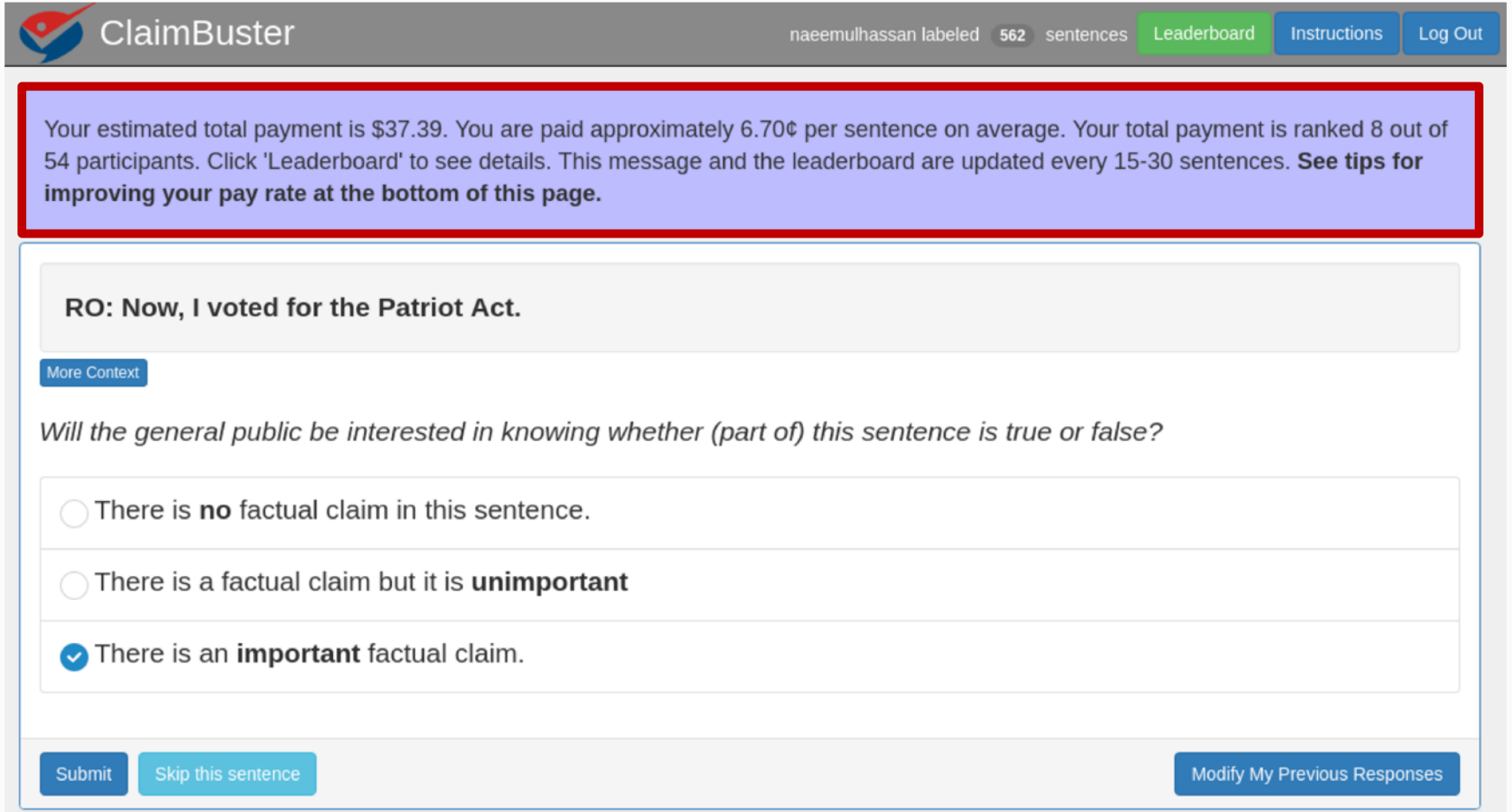
UFS: Unimportant factual claims

"I was in Iowa yesterday." "My mother enjoys cooking." "I ran for President once before."

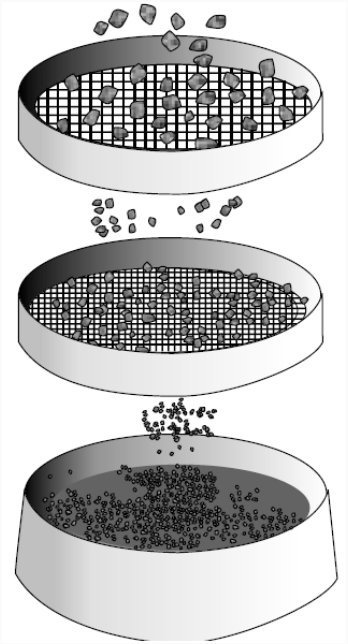
NFS: No factual claims (opinions, questions & declarations)

"Iran must not get nuclear weapons." "7% unemployment is too high." "My opponent is wishy-washy." "I will be tough on crime." "Why should we do that?" "Hello, New Hampshire!" "Our plan is to reduce tax rate by 10%."

Our Own CrowdFlower



Ground Truth Collection



20788 sentences
374 coders
76552 labels

86 top-quality coders
52333 labels

Majority voting
20617 admitted sentences

- 20 months, 374 coders, ~$4,000 paid
- 30 training sentences
- 1032 screening sentences (731 NFS, 63 UFS, 238 CFS) to detect spammers & low-quality coders

Class	Count	Percentage
CFS	4849	23.52%
UFS	2097	10.17%
NFS	13671	66.31%

http://www.engineersdaily.com/2014/03/basics-of-soil-compaction.html


SECURITY
SAMSUNG

ON AIR

2016 Third Presidential Debate Live. Oct. 19, 2016, 9 p.m. EST

Paticipants: Donald Trump, Hillary Clinton
Moderators: Chris Wallace

2016 Third Presidential Debate. Oct. 19, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Chris Wallace

2016 Second Presidential Debate. Oct. 9, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Martha Raddatz, Anderson Cooper, Chris Wallace

2016 Vice Presidential Debate. Oct. 4, 2016, 9 p.m. EST

Paticipants: Tim Kaine, Mike Pence, Hillary Clinton, Donald Trump
Moderators: Elaine Quijano, Martha Raddatz, Anderson Cooper

2016 First Presidential Debate. Sep. 26, 2016, 9 p.m. EST

Paticipants: Hillary Clinton, Donald Trump
Moderators: Lester Holt

2016 Democratic Party Presidential Debate. April 14, 2016, 9 p.m. EST

Paticipants: Bernard Sanders, Hillary Clinton, Donald Trump

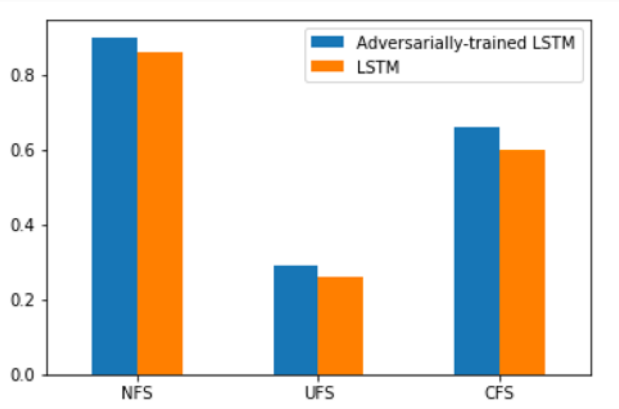
LSTM-based Model for Claim Spotting

- **Model based on an LSTM (long short-term memory) network**
- Tested several different word embeddings but found no significant difference in overall performance
  - Different embeddings did have different affinities (i.e., some produce models which score text with numbers higher on average).
- We used nonsensical sentences to make the model more resistant to attacks. The neural models outperformed the SVM model by 29% in precision and 12% in recall.

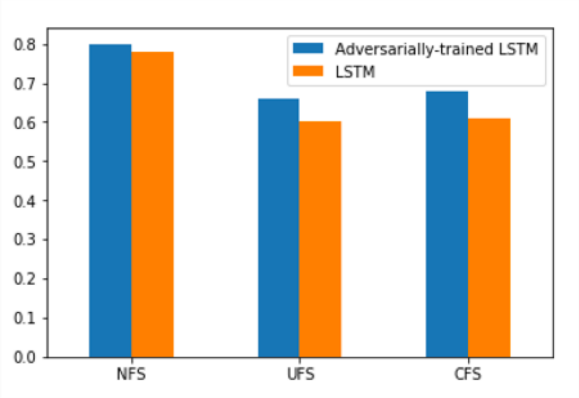
Adversarially Trained LSTM

- Training the LSTM model on adversarial (modified) examples.
- Adversarial noise (perturbations), designed to maximize the chance of misclassification by the network, was calculated using back-propagation.
- Perturbation was applied to word-embeddings rather than directly to the input.

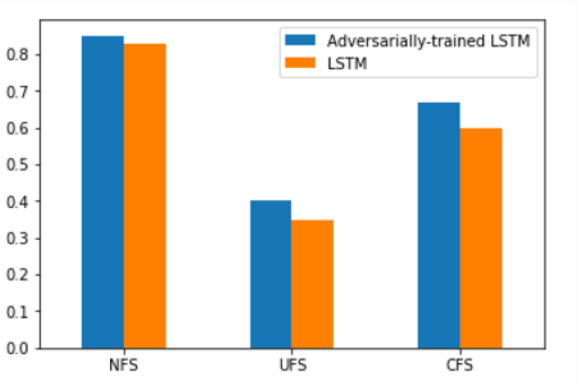
Adversarially Trained LSTM



Recall



Precision



F1-score

ClaimBuster API

apidocs : ClaimBuster API | Show/Hide | List Operations | Expand Operations | Raw

GET	/all/{claim}	All
GET	/se/{claim}	Claim Checker - Search Engine
GET	/fm/{claim}	Claim Matcher
GET	/kb/{claim}	Claim Checker - Knowledge Bases
GET	/score_text/{text}	Claim Spotter - Get Score
GET	/score_url/{url}	Claim Spotter - Get Score URL

From the Congressional Record

Statements by members of the House and Senate as officially transcribed from floor speeches and debates.

Speaker	Claim	Chamber	Link
Mr. REED	On June 4, 1919, citizens of Tioga County joined together and signed a charter to create the Owego Chamber of Commerce.	House	Link
Ms. ESCOBAR	According to Mr. Bradford's son, for more than 30 days, Mr. Bradford avoided capture and eventually was able to find his way back to the American forces.	House	Link
Ms. ESCOBAR	Madam Speaker, I rise today to honor the life of an American hero, Bloyce Bradford, who passed away at 90 years old on Tuesday, January 29, 2019.	House	Link
Ms. ESCOBAR	While in Korea, Mr. Bradford was wounded multiple times and, at one point, was considered missing in action.	House	Link
The PRESIDING OFFICER	The senior assistant legislative clerk read the following letter: U.S. Senate, President pro tempore, Washington, DC, February 15, 2019.	Senate	Link
Mr. REED	Given the above, I ask that this Legislative Body pause in its deliberations and join me to celebrate the Tioga County Chamber of Commerce's one hundred years of service.	House	Link
The PRESIDING OFFICER	To the Senate: Under the provisions of rule I, paragraph 3, of the Standing Rules of the Senate, I hereby appoint the Honorable Ben Sasse, a Senator from the State of Nebrask[illegible]	Senate	Link
Mr. REED	The business show brings new bus[illegible]	House	Link
Mr. REED	By purchasing properties and then improving those properties, the Chamber helped to create environments that welcomed new businesses.	House	Link
Ms. ESCOBAR	His legacy lives on through his sist[illegible] his companion of 43 years and her family.	House	Link
Mr. REED	During World War II, the Chamber assisted with the [illegible]	House	Link
Ms. ESCOBAR	He went on to enroll in the Army, where he served during the Korean War.	House	Link
Mr. REED	The Owego Chamber of Com[illegible] management of real estate.	House	Link
Mr. REED	I congratulate the Tioga Cou[illegible] supporting businesses in our	House	Link
The PRESIDING OFFICER	Mr. SASSE thereupon assume[illegible]	Senate	Link

PolitiFact "Buffet" of Factual Claims

	A	B	C	D	E
1	Statement	Speaker	Date	Location/Source	Link
232	I haven't heard one (republican) say that they're in	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
233	concentrated wealth and capital have been put into	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
234	it is not true that regulation holds poor people down or	o'malley	4/20/2015	morning edition	http://www.npr.org/blogs/its
235	Every three weeks we bring online as much as solar	Obama	4/18/2015	Weekly address	https://www.whitehouse.gov
236	Our carbon polution has fallen by 10 percent since	Obama	4/19/2015	Weekly address	https://www.whitehouse.gov
237	"Benghazi has had more hearings, more documents	Claire McCaskill	4/19/2015	ABC this week	http://thehill.com/policy/inte
238	With fast track trade " I would be receiving the same	Obama	4/17/2015	remarks joint press	https://www.whitehouse.gov
239	Loretta Lynch "has been now sitting there longer than	Obama	4/17/2015	remarks joint press	https://www.whitehouse.gov
240	I just counted 19 Republicans who are likely to run for	chris cillizza	4/20/2015	the twitters	https://twitter.com/TheFix/st
241	He [Rand Paul] wanted to cut all our defense	John McCain	4/20/2015	Fox News	https://www.youtube.com/w

Duke Reporters' Lab uses ClaimBuster API in creating daily newsletters that recommend to The Washington Post, New York Times, PolitiFact, and other fact-checkers the most check-worthy claims in CNN programs, Tweets, Facebook ports, and the Congressional Record.

This Washington Post fact check was chosen by a bot

BY DANIEL FUNKE · JANUARY 30, 2018

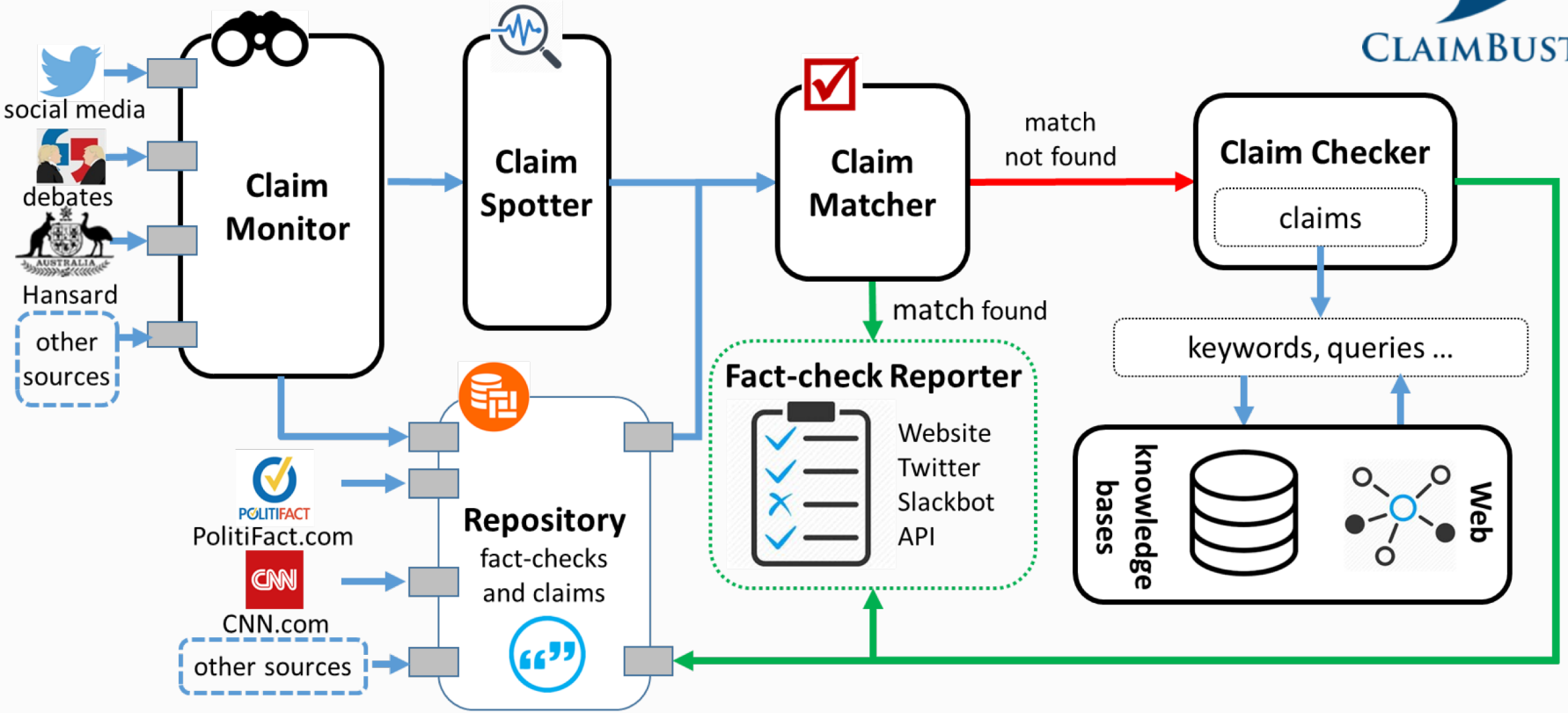
"There were 10 million people that would have been on some path to staying in this country for indefinitely in the proposal that [Sens.] Dick Durbin and Lindsey Graham put forward."
— Former senator Rick Santorum (R-Pa.), in remarks on CNN's "State of the Union," Jan. 21, 2018

"I had actually looked at that transcript but I missed that claim," said Glenn Kessler, who runs The Fact Checker. "So it would have been lost to history if it had not been for ClaimBuster."


Claim
Spotter

The First-ever End-to-end Fact-checking System

[CIKM15a, KDD17, VLDB17demo, IJCNN18]
*** First Runner-Up, SIGMOD17 Student Research Competition**



ClaimBuster: Automated Live Fact-checking

iDiR

Enter Your Own Text

2016 U.S. Presidential Debates

Hansard: Parliament of Australia

ClaimBuster API

ClaimBuster Slackbot

End-to-end Fact-checking (beta)

Tweets by @ClaimBusterTM

ClaimBuster Retweeted

Linda McMahon
@SBALinda

The U.S. has 29 million small businesses & nearly 867,000 of them are in #Michigan. Almost ALL businesses in Michigan, 99.6% are #SmallBiz.

Aug 8, 2017

ClaimBuster Retweeted

Governor Walker
@GovWalker

Since 2013, Kenosha County has seen nearly $1 billion in capital investment – creating over 6,000 jobs!

Embed View on Twitter

2016 Third Presidential Debate. Oct. 19, 2016, 9 p.m. EST

Chronological Order | Order by Score

Least Check-worthy >=0.1 >=0.2 >=0.3 >=0.4 >=0.5 >=0.6 >=0.7 >=0.8 >=0.9 >=1.0 Most Check-worthy

is tougher. But they know what's going on. They know it better than anybody. They want strong borders. They feel we have to have strong borders. I was up in New Hampshire the other day. The biggest complaint they have -- it's with all of the problems going on in the world, many of the problems caused by Hillary Clinton and by Barack Obama. All of the problems -- the single biggest problem is heroin that pours across our southern border. It's just pouring and destroying their youth. It's poisoning the blood of their youth and plenty of other people. We have to [Fact-check this] ders. We have to keep the drugs out of our country. We are -- right now, we're getting the drugs, they're getting the cash. We need strong borders. We need absolute -- we cannot give amnesty. Now, I want to build the wall. We need the wall.

Claim Checker - Knowledge Bases

Consulting the knowledge bases produced the following results:

Truth Rating Indeterminable

Question Asked
What is all of the problems-- the single biggest problem?

Response Recieved
The single biggest problem in communication is the illusion that it has ...

Claim Matcher

We found the following claims which have been professionally fact-checked. Check them out!

Truth Rating True

Claim
"Heroin .. pours across our southern borders."

Speaker Donald Trump

URL politifact

Truth Rating True

Claim Checker - Search Engine

We found the following information after processing some search engine results:

All of the problems -- the single biggest problem is heroin that pours across our southern border. It's just pouring and destroying their youth.

Similarity Rating 0.8320502943378437

URL source

"I was up in New Hampshire the other day," Trump said in the debate. "The biggest

GENIUS

FILTER BY All Annotations

The biggest complaint they have -- it's with all of the problems going on in the world, many of the problems caused by Hillary Clinton and by Barack Obama

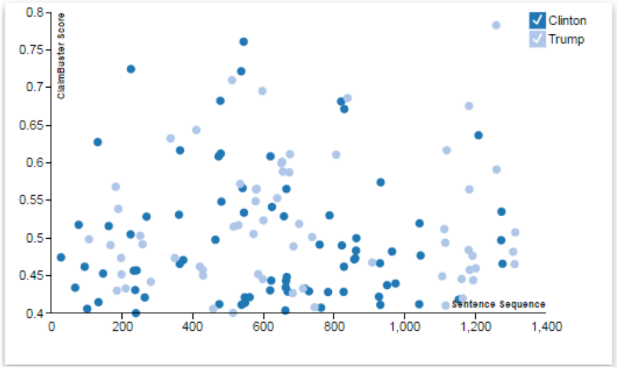
JosueCaraballo 17d

Is that true? What is the source?

report abuse

Upvote

Add a comment

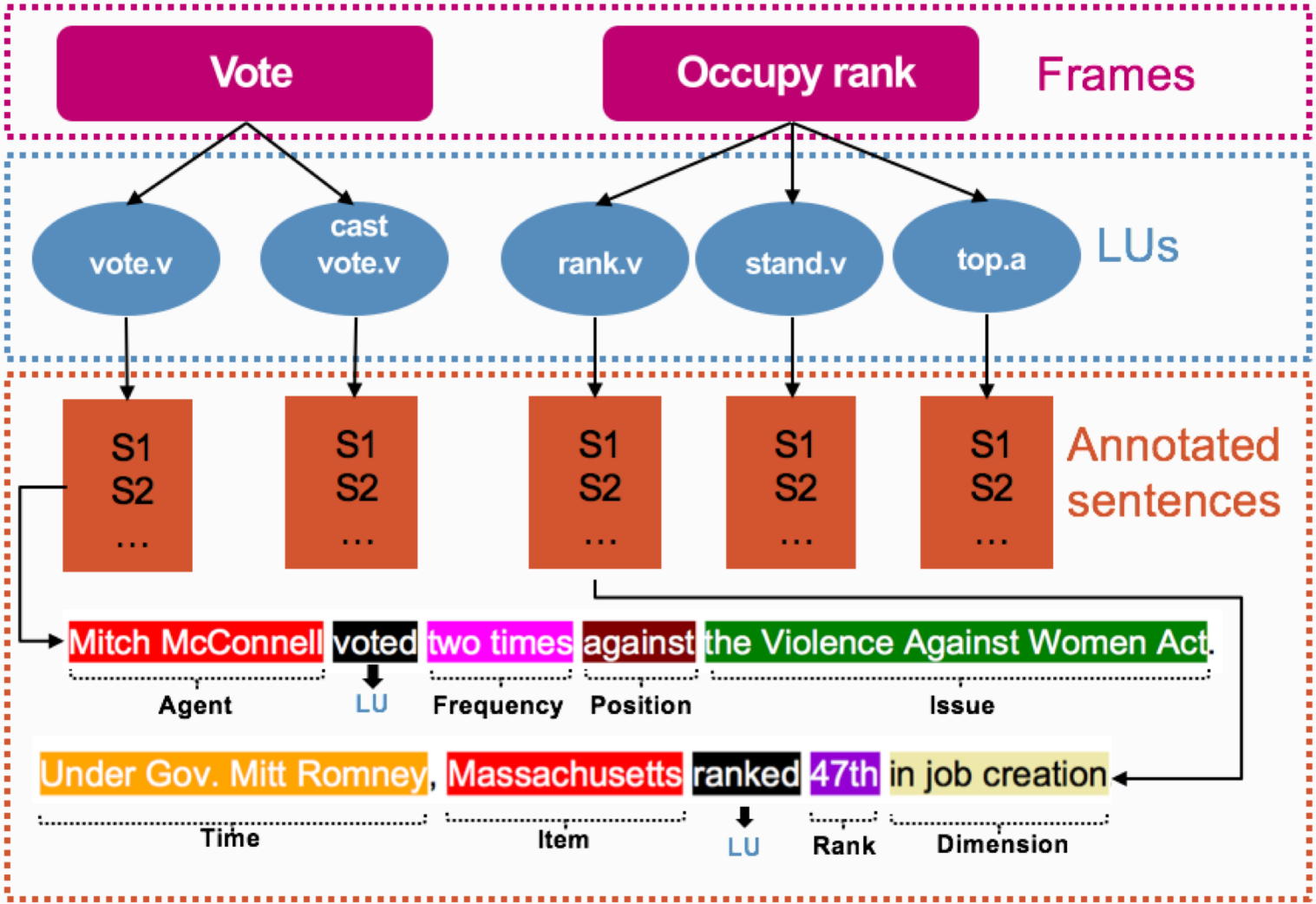


Outline

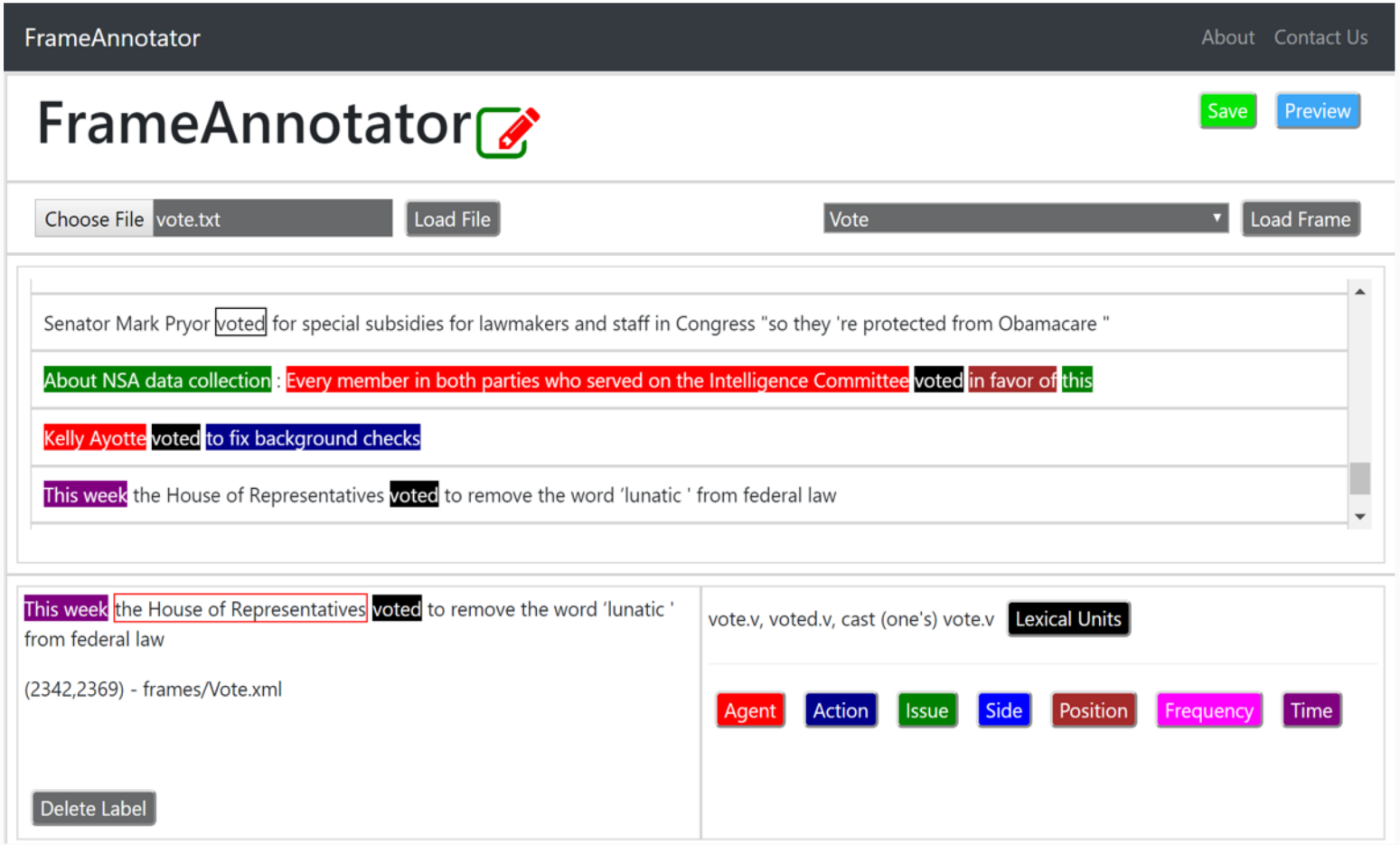
- A Brief History of our Computation + Journalism Research
- Computational Journalism
  - Data-driven fact-checking (ClaimBuster)
  - **Other ongoing fact-checking projects**
  - Exceptional fact finding (FactWatcher, Maverick)
- Graph Data Usability (Orion, GQBE, TableView, Maverick)

Using Frames to Model Factual Claims

- An extension of FrameNet (Baker et. al., 1998) for structured and semantic modeling of factual claims.
- Capture various aspects of a factual claim: domain and topic, entities involved and their relationships, quantities, comparisons, aggregate structures ...
- Useful for various steps in automating fact-checking: detecting factual claims, matching claims with fact-checks, translating claims to structured queries, ...



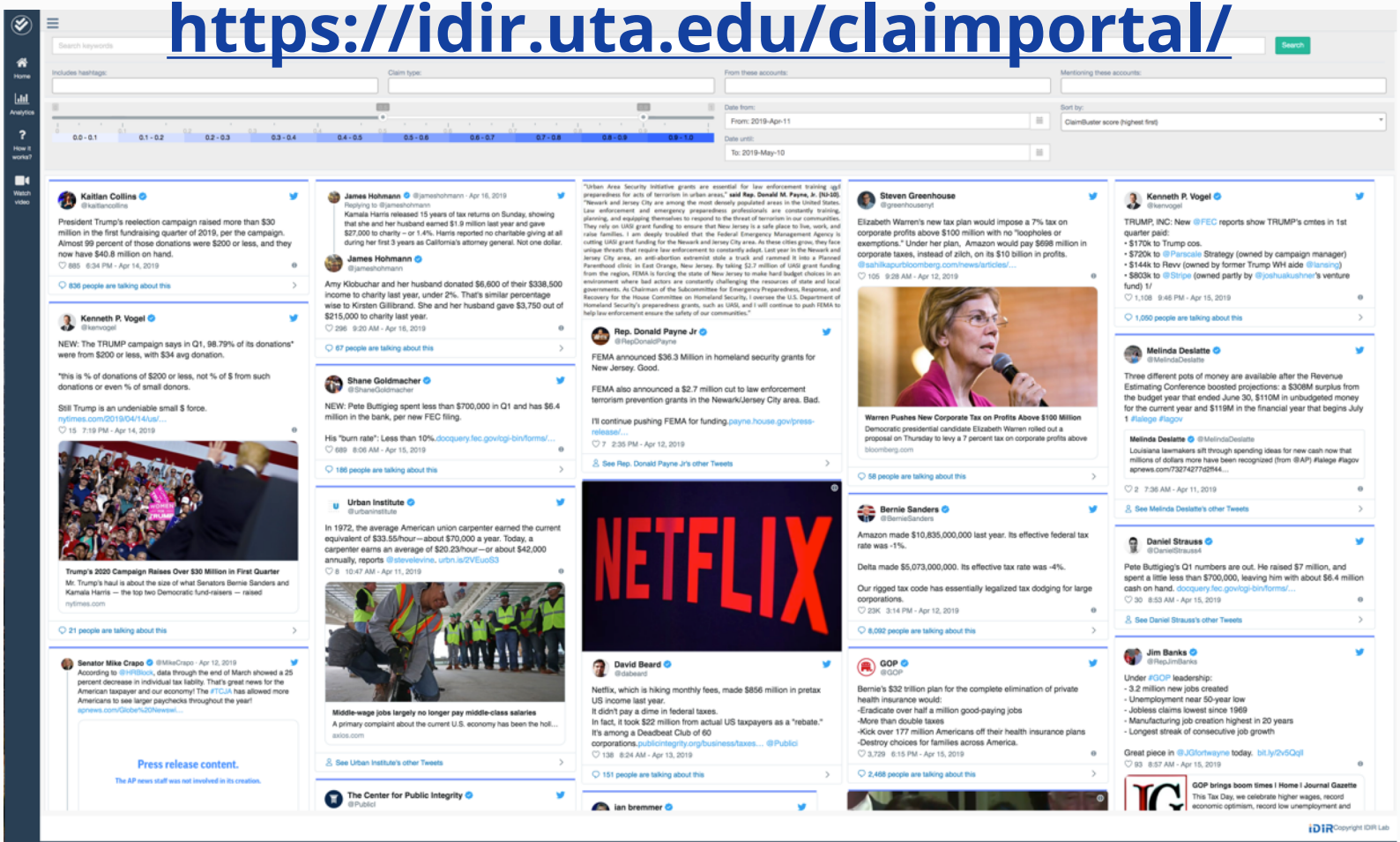
Vote and *Occupy rank* frames along with their *LUs.*



https://idir.uta.edu/frameannotator/

ClaimPortal: Integrated Monitoring, Searching, Checking, and Analytics of Factual Claims on Twitter

https://idir.uta.edu/claimportal/



Outline

- A Brief History of our Computation + Journalism Research
- Computational Journalism
  - Data-driven fact-checking (ClaimBuster)
  - Other ongoing fact-checking projects
  - Exceptional fact finding (**FactWatcher**, Maverick)
- Graph Data Usability (Orion, GQBE, TableView, Maverick)

How it Started

Chris Paul had 16 points, 10 rebounds, 13 assists and five steals..... The only other active player to have such a game is Jason Kidd...

2009

How did they come up with that?

Exceptional Fact Finding

Real-world event (sports, transportation, crime, weather, finance, social media)

New tuple appended to database

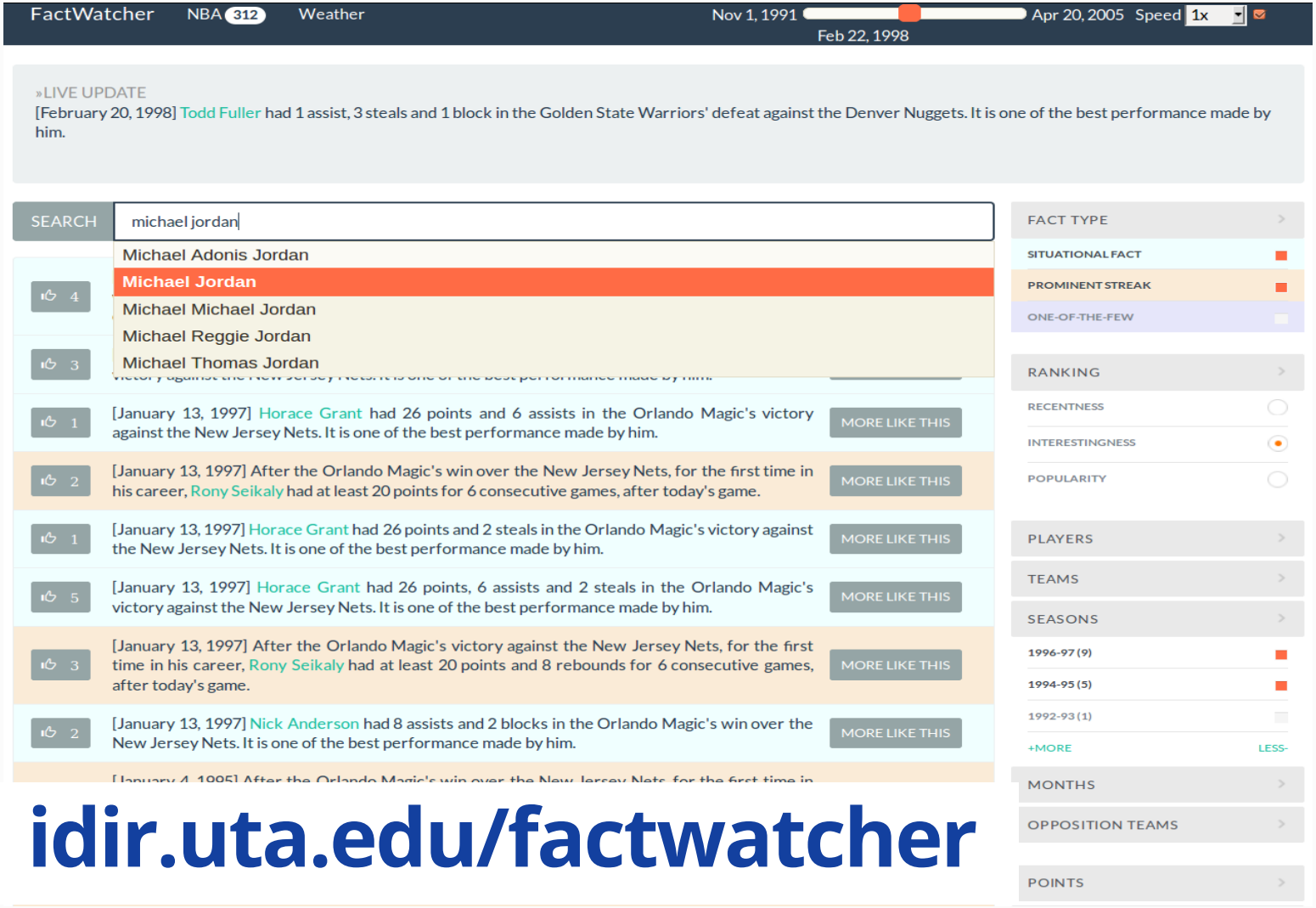
id	player	day	month	season	team	opp_team	pts	ast	reb
t_1	Bogues	11	Feb.	1991-92	Hornets	Hawks	4	12	5
t_2	Seikaly	13	Feb.	1991-92	Heat	Hawks	24	5	15
t_3	Sherman	7	Dec.	1993-94	Celtics	Nets	13	13	5
t_4	Wesley	4	Feb.	1994-95	Celtics	Nets	2	5	2
t_5	Wesley	5	Feb.	1994-95	Celtics	Timberwolves	3	5	3
t_6	Strictland	3	Jan.	1995-96	Blazers	Celtics	27	18	8
t_7	Wesley	25	Feb.	1995-96	Celtics	Nets	12	13	5

Fact-finding algorithms

Wesley had 12 points, 13 assists and 5 rebounds on February 25, 1996 to become the first player with a 12/13/5 (points/assists/rebounds) in February.

Number-based facts make news stories more engaging

http://en.wikipedia.org/wiki/Basketball



idir.uta.edu/factwatcher

VLDB 2014

Excellent Demo Award

Fact-finding Algorithms

FactWatcher [VLDB14 demo]

- [ICDE14] Situational Facts: "No other player scored more pts and reb against DAL than Jordan."
- [KDD12] One-of-the-Few: "Jordan scored 10 pts & 10 reb. Only 3 others had similar performance."
- [KDD11] Prominent Streaks: "The Nikkei 225 closed below 10000 for the 12th consecutive week, the longest such streak since June 2009."
- [TKDD14] General Prominent Streaks: "James has scored at least 20 points and handed out 10 or more assists in each of his last five games against the Hawks, the longest streak he has ever had against one team."

Frequent Episode Mining [ICDE15]

Fact-finding Algorithms

- Many interesting facts in the real-world can be modeled as various types of skyline points.
- The gist of fact-finding is to efficiently, incrementally maintain skyline points over ever-changing data, while considering constraints such as selection conditions.

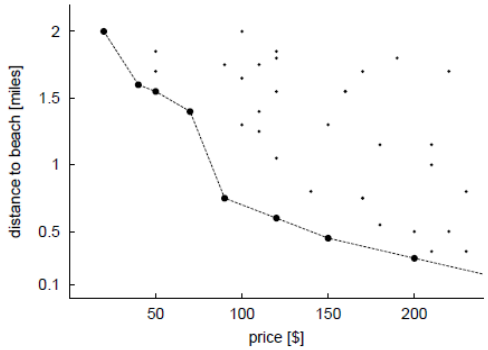


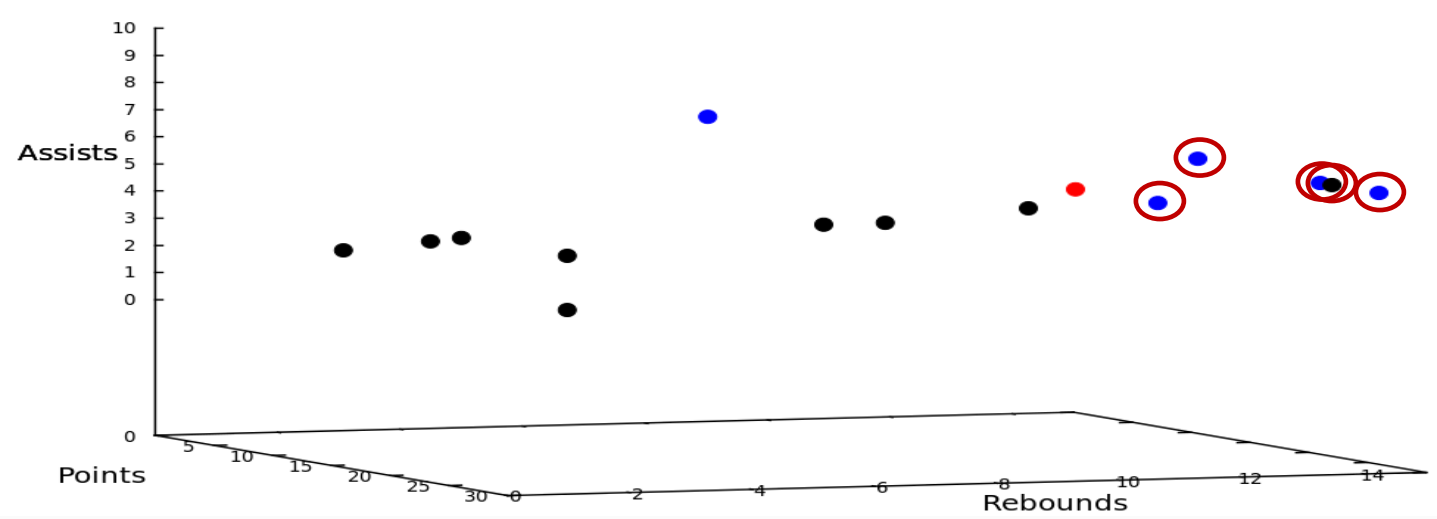
Figure 1: Skyline of Hotels

Figure 2: Skyline of Manhattan

Stephan Börzsönyi, Donald Kossmann, Konrad Stocker: The Skyline Operator. ICDE 2001: 421-430

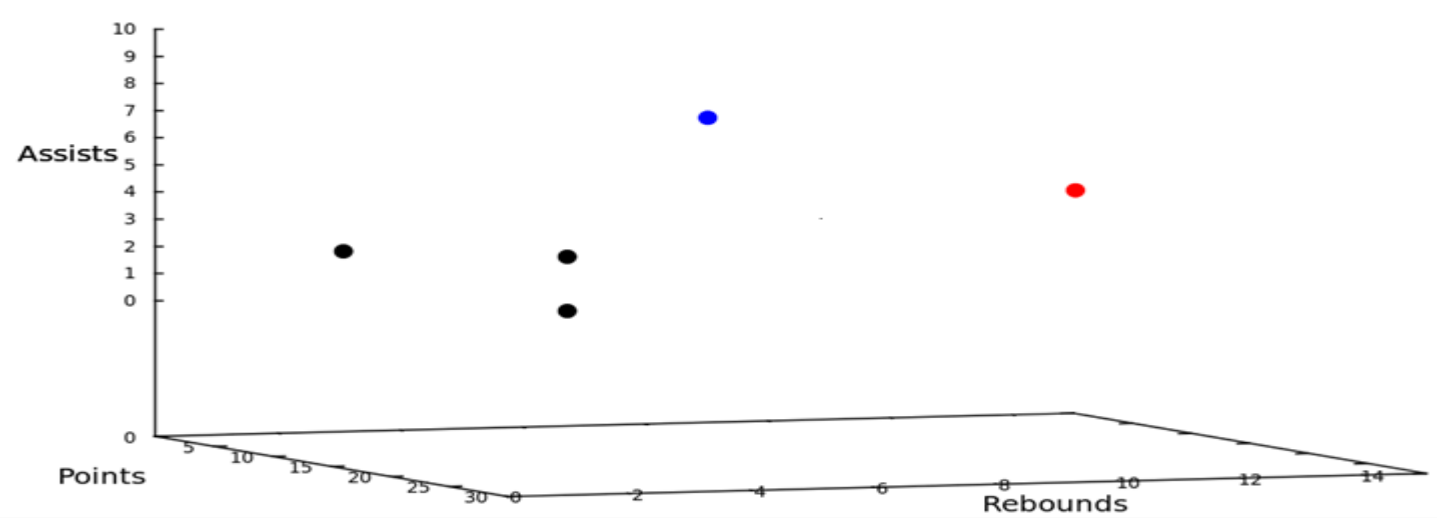
Modeling Situational Facts

"Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992."
(http://espn.go.com/espn/elias?date=20130205)



Modeling Situational Facts

"Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992."
(http://espn.go.com/espn/elias?date=20130205)



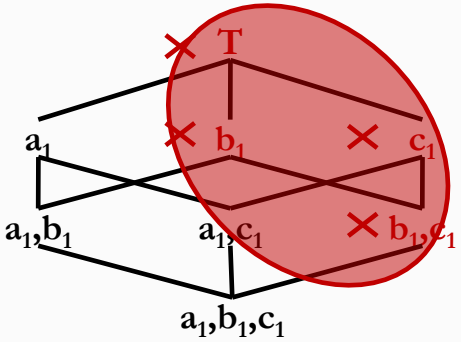
Modeling Situational Facts

Dimension space: $\mathcal{D}=\{d_1,\ldots,d_n\}$

Measure space: $\mathcal{M}=\{m_1,\ldots,m_s\}$

id	player	day	month	season	team	opp_team	pts	ast	reb
t_1	Bogues	11	Feb.	1991-92	Hornets	Hawks	4	12	5
t_2	Seikaly	13	Feb.	1991-92	Heat	Hawks	24	5	15
t_3	Sherman	7	Dec.	1993-94	Celtics	Nets	13	13	5
t_4	Wesley	4	Feb.	1994-95	Celtics	Nets	2	5	2
t_5	Wesley	5	Feb.	1994-95	Celtics	Timberwolves	3	5	3
t_6	Strictland	3	Jan.	1995-96	Blazers	Celtics	27	18	8

Skyline maintenance +
Data Cube/OLAP



Outline

- A Brief History of our Computation + Journalism Research
- Computational Journalism
  - Data-driven fact-checking (ClaimBuster)
  - Other ongoing fact-checking projects
  - Exceptional fact finding (FactWatcher, Maverick)
- **Graph Data Usability (Orion, GQBE, TableView, Maverick)**

Tackling Graph Data Usability Challenges

Challenges

- Massive, complex graphs; millions of entities; billions of edges.
- Requires substantial understanding of schema and data and complex pre-processing, before one can fetch information or gain insights from data.

Objectives

- Make it easy to understand, query, explore, and clean graph data.

Systems

- GQBE [TKDE15, ICDE14demo]: graph query by example
- Orion [VLDB15demo, SIGMOD17tutorial]: auto-suggestion for interactive query formulation
- TableView [SIGMOD16, ICDE18demo]: generating preview tables for knowledge graphs
  (* SIGMOD Most Reproducible Paper Award)
- Maverick [SIGMOD18, VLDB18demo]: finding outliers and errors in graphs

Outline

- A Brief History of our Computation + Journalism Research
- Computational Journalism
  - Data-driven fact-checking (ClaimBuster)
  - Other ongoing fact-checking projects
  - Exceptional fact finding (FactWatcher, **Maverick**)
- Graph Data Usability (Orion, GQBE, TableView, Maverick)

Maverick: Discovering Exceptional Facts from Knowledge Graphs

[SIGMOD18, VLDB18demo]

Exceptional Facts

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

Entity of Interest	Denzel Washington
Context	Best Actor award winners
Attributes	Ethnicity
Peculiar value	African American (only two satisfy)

Exceptional Facts

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

Entity of Interest	Denzel Washington	Given an entity **x** **find**
Context	Best Actor award winners	A context
Attributes	Ethnicity	A set of attributes (subspace)
Peculiar value	African American (only two satisfy)	

Exceptional Facts

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

Entity of Interest	Denzel Washington	Given an entity **x**	
		find	**such that**
Context	Best Actor award winners	A context	the context has many entities, including **x**
Attributes	Ethnicity	A set of attributes (subspace)	
Peculiar value	African American (only two satisfy)		**x** bears a peculiar value w.r.t. the subspace (few in the context have the value)

Exceptional Facts

Denzel Washington followed Sidney Poitier as only the second black to win the Best Actor award.

This was Brazil's first own goal in World Cup history.

Hillary Clinton becomes first female presidential nominee.

Applications

Computational Journalism

- Fact-finding
- Fact-checking
  - The first female presidential nominee was Victoria Woodhull, not Hillary Clinton (snopes.com)

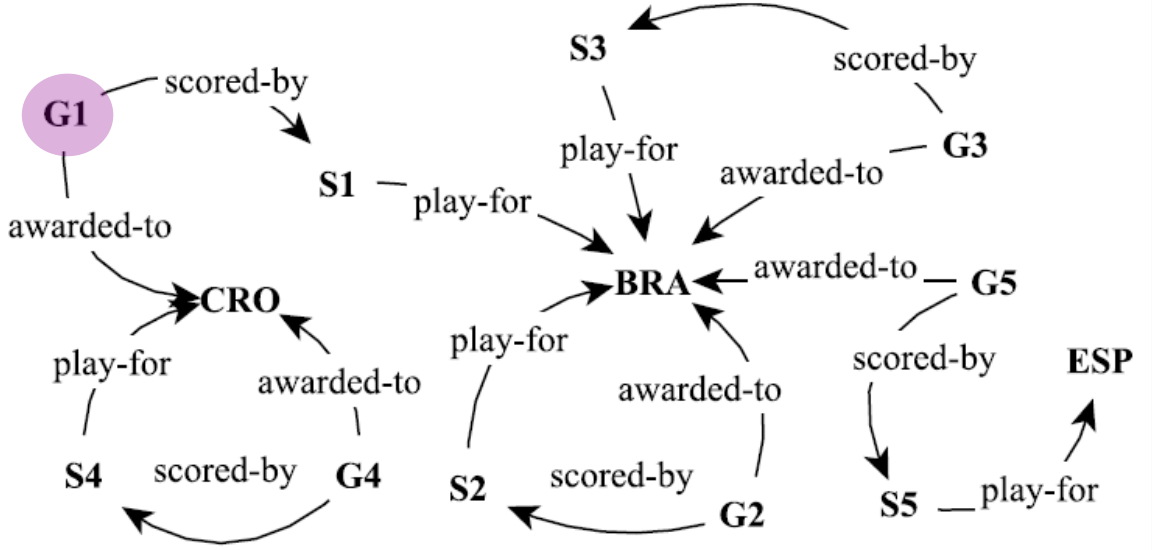
Data Cleaning

Recommendation Systems

- Friends, news, and product promotion

Exceptional Facts from Knowledge Graphs

What is exceptional about G1?



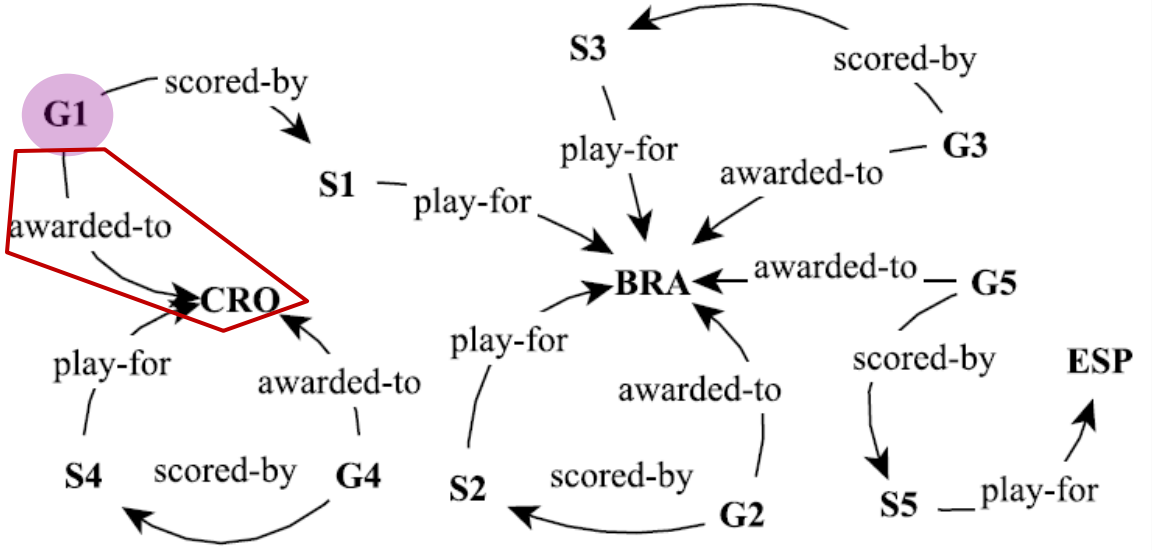
Modeling

Attributes: labels of incoming/outgoing edges

Values: direct neighbors

Subspace: a subset of attributes

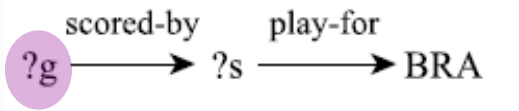
G1. awarded-to = CRO



Modeling

Context: entities sharing some common characteristics

Defined by a pattern-variable pair

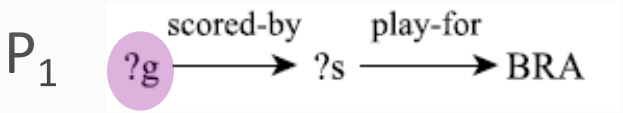
P_1 

Goals scored by Brazilian players

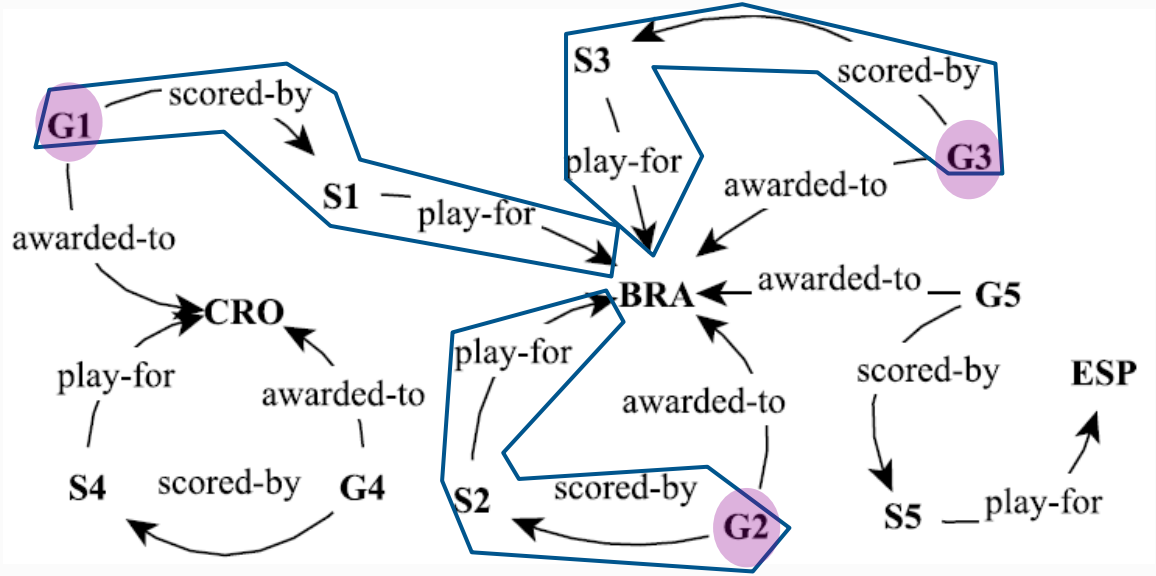
Modeling

Context: entities sharing some common characteristics

Defined by a pattern-variable pair



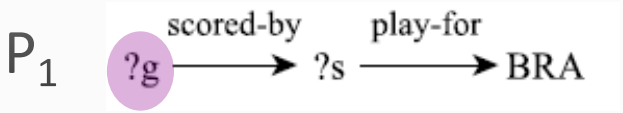
Goals scored by Brazilian players



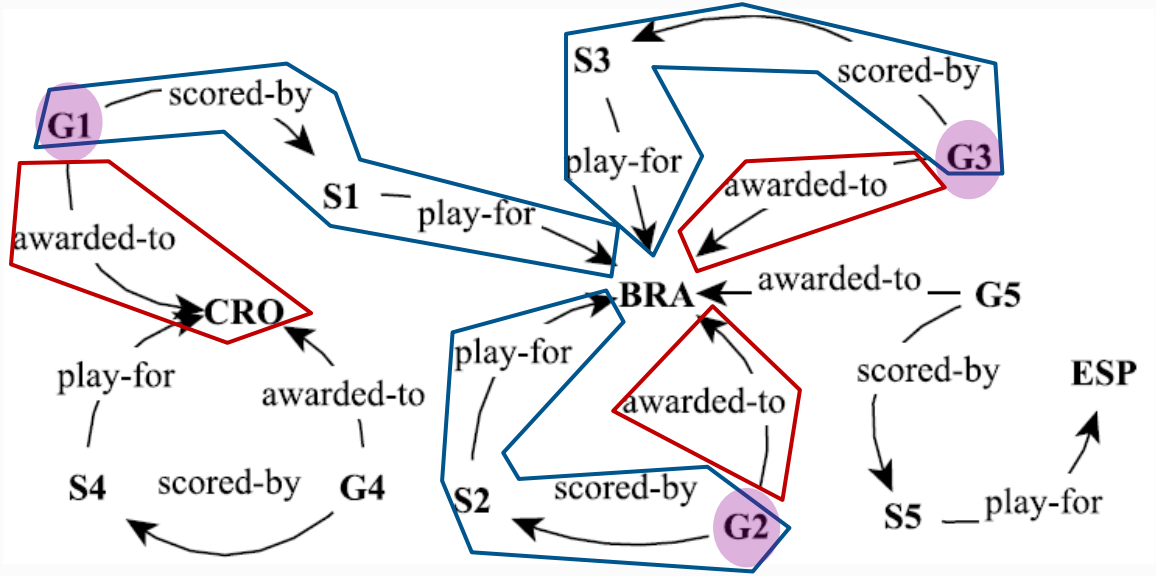
Modeling

Context: entities sharing some common characteristics

Defined by a pattern-variable pair



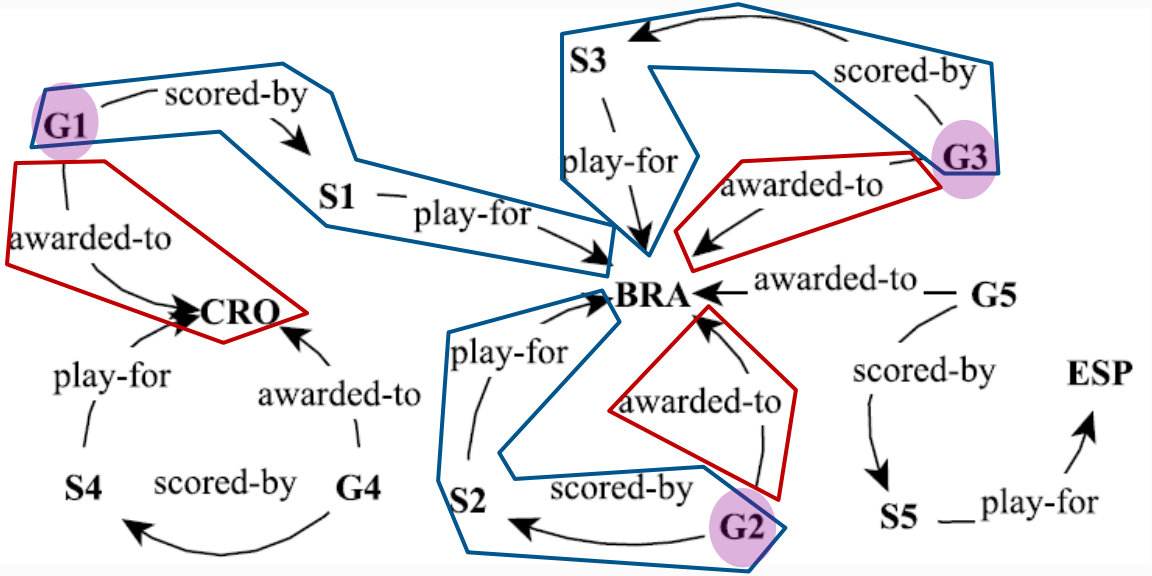
Goals scored by Brazilian players



Modeling

What is exceptional about G1?

Among all the goals scored by BRA players, G1 is the only own goal.



Problem Formulation

Input

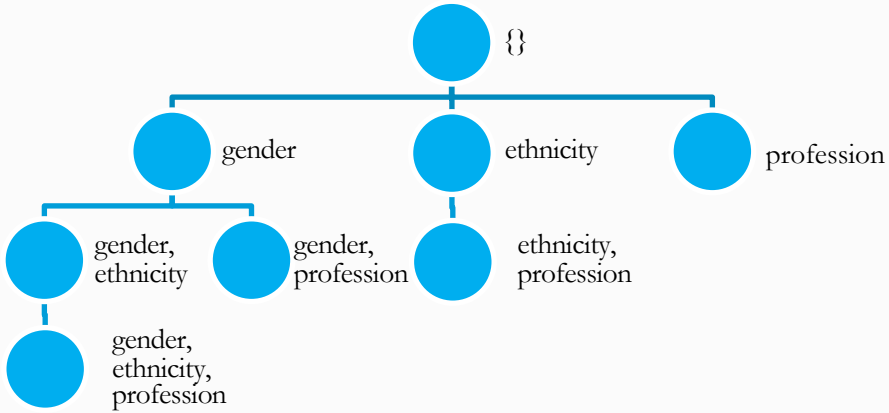
- Entity of interest v_0
- Exceptionality function χ
- Result size k

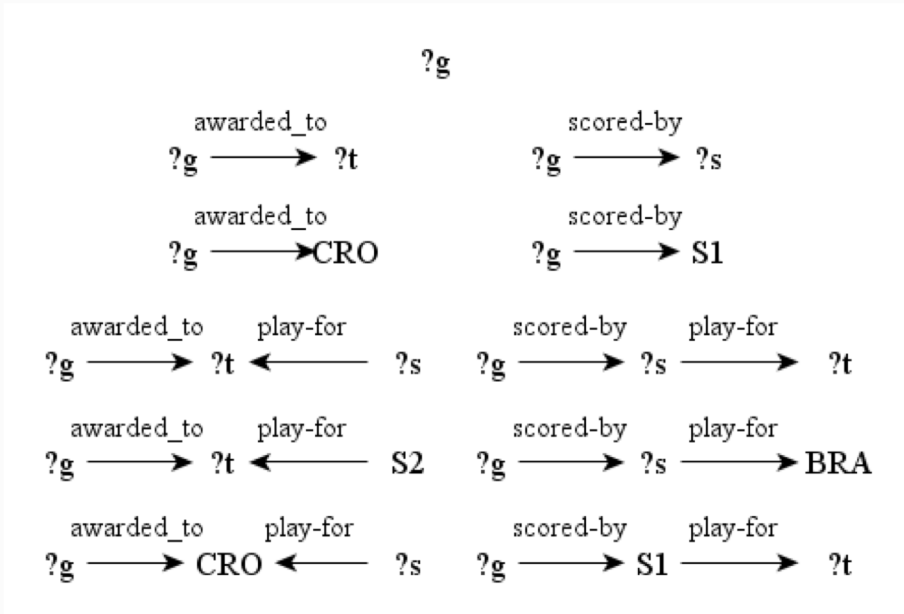
Output

- Top-k (context, subspace) pairs with regard to χ, in which v_0 stands out

Challenges

- **Space of attributes:** $O(2^{|A_{v_0}|})$
- **Space of patterns:** $\Omega(2^{|V_G|})$





Related work

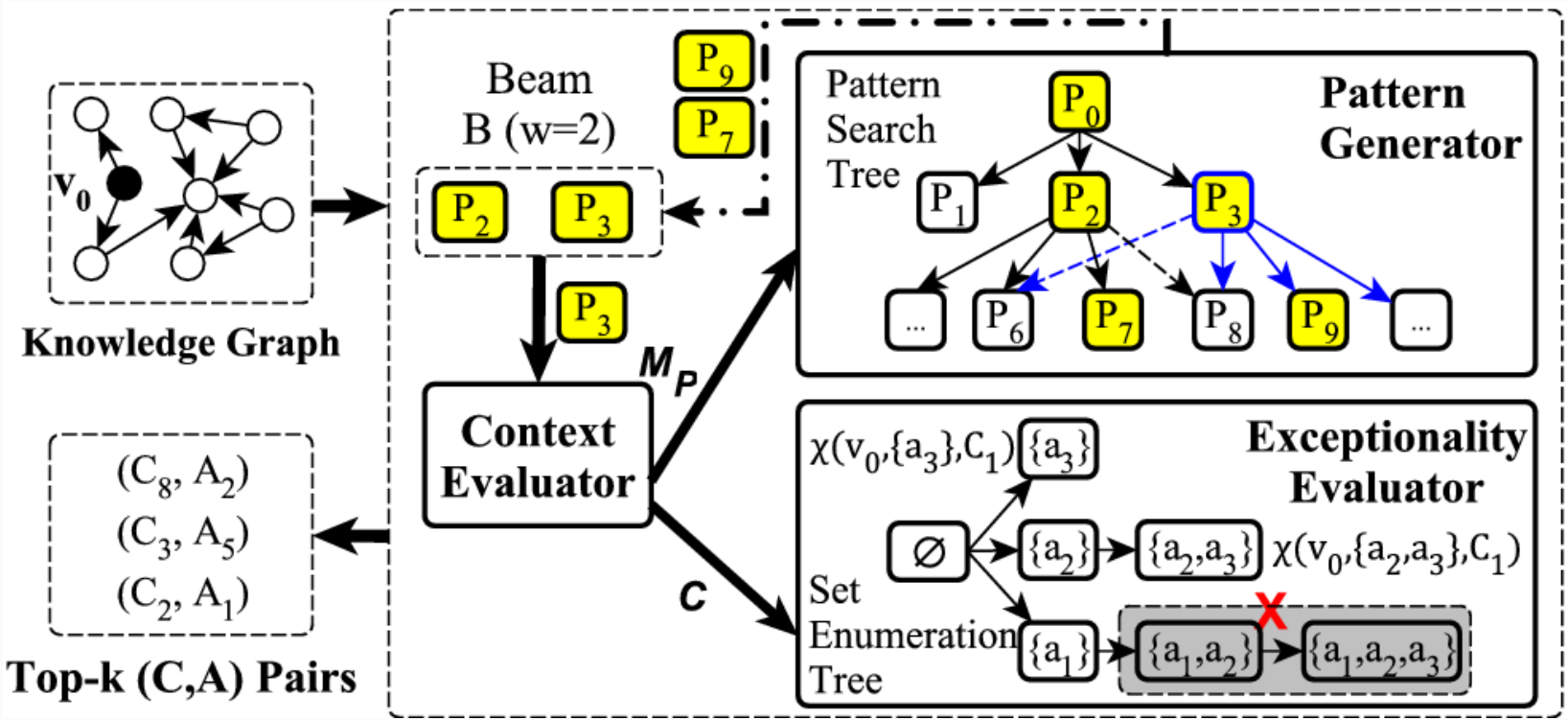
Outlier detection

- Different goal: find a set of objects from a dataset
- Outlying explanation is not a focus

Outlying aspect mining

- Single table model, not suitable for graph-based data
    - Extremely large and sparse table
    - Conjunctive queries ≠ Pattern queries
    - Set values

Maverick [SIGMOD2018]



Exceptionality Function χ

$\chi(v, A, C) \in \mathcal{R}$

outlierness (χ_o) [Angiuli2009TODS], one-of-the-few(χ_f) [Wu2012KDD], isolation scores (χ_i) [Liu2008ICDM]

Upper bound function

Theorem 4.2 $$upper_o(v, A, C) = \sum_{S \in \mathcal{S}_A} (p_S)^2 - \frac{(2\, p_{v.A} + 1) \times |\, C\,| - 2}{|\, C\,|^2}$$

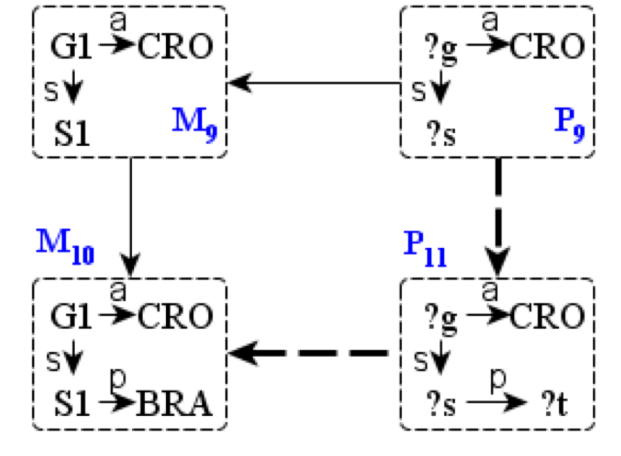
Theorem 4.3 $$upper_f(v, A, C) = |\, \{u \mid u \in \overline{C_v},\ p^A_{u.A} > 1/|\, C\,|\}\,| \Big/ |\, C\,|$$

Theorem 4.4 $$upper_i(v, A, C) = 1 - 2^{-\frac{-\log_2 \frac{1}{|\, C\,|}}{-q_{v.A} - sum_{S \in \mathcal{S}_A \setminus \{v.A\}}\, p_S \times \log_2 p_S}}$$

Match–based Pattern Generation

- **Construct Partial Order of Valid Patterns**

THEOREM 5.4. *Suppose P' is a child of $P \in \mathbb{P}$, i.e., $(P, P') \in E_{\mathbb{P}}$ and thus P' is a valid pattern with matches. Given any match M' to P', there exists a match M to P that is a subgraph of M', i.e., $\forall M' \in \mathcal{M}_{P'}, \exists M \in \mathcal{M}_P$ s.t. $V_M \subseteq V_{M'}$ and $E_M \subseteq E_{M'}$.*



Datasets and Experiments

WCGoals

Created based on FIFA.com

11 node types, 13 edge types

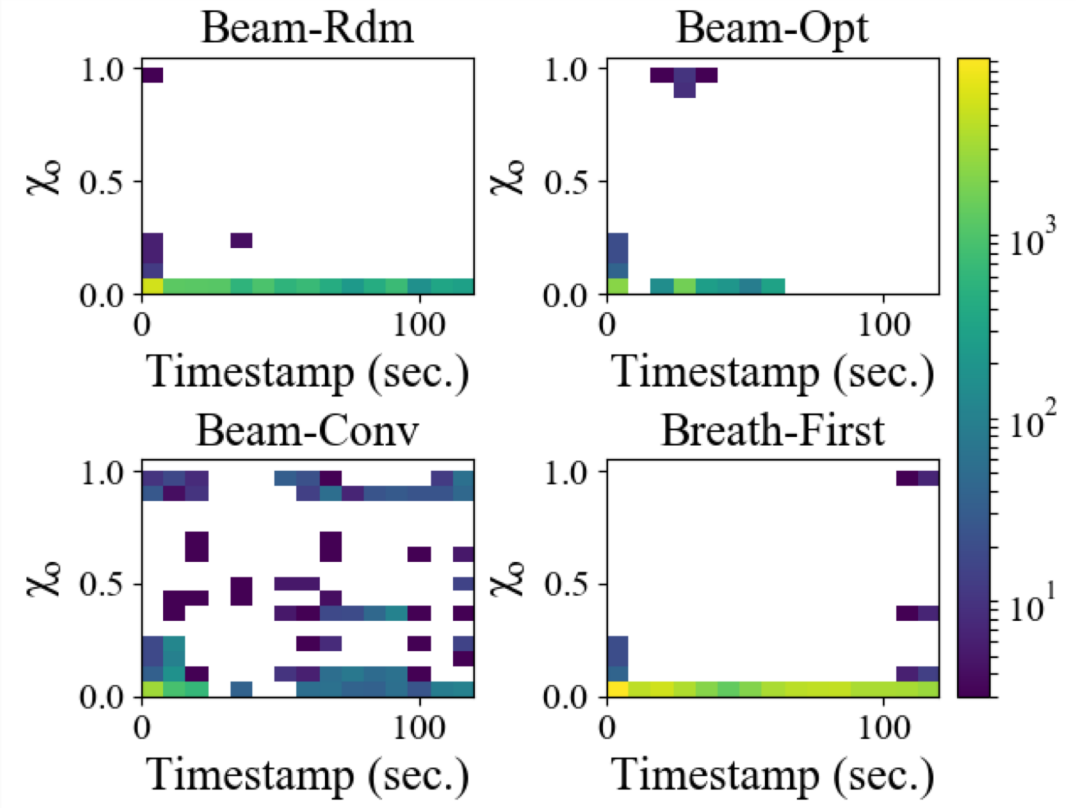
49,078 nodes, 158,114 edges

Film-Award

A subgraph of Freebase

95 node types, 117 edge types

5,437,628 nodes, 10,879,448 edges



IDIR Projects and Demos

Demonstration Videos https://vimeo.com/channels/1024406

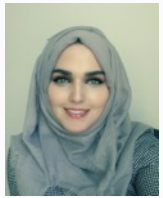
- ClaimBuster (idir.uta.edu/claimbuster) Automated, live fact-checking
- ClaimPortal (idir.uta.edu/claimportal) Monitoring factual-claims on social media
- CrewScout (idir.uta.edu/crewscout) Expert team finding by skyline groups
- ERQ: Entity-Relationship Query (idir.uta.edu/erq) Structured query on Wikipedia
- Facetedpedia (idir.uta.edu/facetedpedia) Faceted search interface for Wikipedia
- FactWatcher (idir.uta.edu/factwatcher) Fact-finding from real-world events
- FrameAnnotator (idir.uta.edu/frameannotator) Frame annotation tool
- GQBE (idir.uta.edu/gqbe) Graph query by example
- Orion (idir.uta.edu/orion) Auto-suggestion for interactive graph query formulation
- TableView Generating preview tables for knowledge graphs
- Maverick Exceptional fact finding from knowledge graphs

Current IDIR Students

Ph.D. students

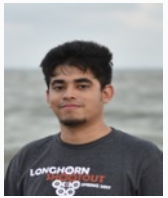

Farahnaz Akrami

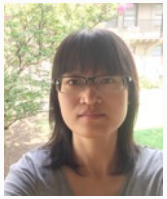

Fatma Arslan


Israa Jaradat


Damian Jimenez


Shadekur Rahman


Samiul Saeef


Xiao Shi


Theodora Toutountzi


Zeyu Zhang

M.S. students


Priyank Arora

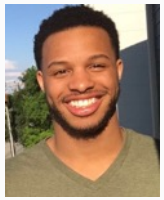

Sumeet Lubal

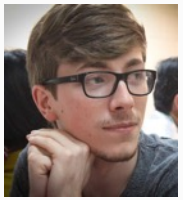

Sarthak Majithia


Daniel Obembe


Sarbajit Roy

B.S. students


Kyrell Dixon


Jacob Devasier

Graduated Students

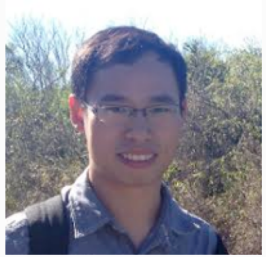
Ph.D. students

Aditya Telang
2011 (co-advised)
IBM Research India

Ning Yan
2013
Research Scientist
Huawei, Santa Clara

Naeemul Hassan
2016
Assistant Professor
University of Mississippi

Nandish Jayaram
2016
Pivotal



Gensheng Zhang
2017
Google

Afroza Sultana
2018
Teradata Lab

M.S. Thesis students

Tulsi Chandwani (2017, Red Hat)
Abu Ayub Ansari Syed (2017)
Ishwor Timilsina (2017, Fidelity)
Nigesh Shakya (2017)
Rohit Bhoopalam (2016, Akamai)
Fatma Dogan (2015, UTA Ph.D.)
Minumol Joseph (2015, Capital One)
Ramesh Venkataraman (2014, Amazon)
Mahesh Gupta (2012, Electronic Arts)
Jijo Philip (2012, Cerner Corporation)
Avinash Bharadwaj (2011, Copper Labs)
Quazi (Sunny) Hasan (2010, Dematic)
Jared Ashman (2010, Ambit Energy)
Muhammad Safiullah (2008, Microsoft)

B.S. students

Josue Caraballo (2017)
Damian Jimenez (2017, UTA Ph.D.)
Long Ly (2015)
Sidharth Goyal (2015)
Huadong Feng (2014)
Raju Karki (2012)
Angus Helm (2010)
Aakash Tuli (2010)

Collaborators on Current and Past IDIR Projects

- Bill Adair (Duke, Public Policy)
- Pankaj Agarwal (Duke)
- Xiang Ao (Chinese Academy of Sciences)
- Vassilis Athitsos (UTA)
- Sourav Bhowmick (Nanyang Technological University)
- Sharma Chakravarthy (UTA)
- Gong Cheng (Nanjing University)
- Byron Choi (Hong Kong Baptist University)
- Christoph Csallner (UTA)
- Gautam Das (UTA)
- Chris Ding (UTA)
- Ramez Elmasri (UTA)
- Leonidas Fegaras (UTA
- Peter Fray (University of Technology Sydney, Journalism)
- James Hamilton (Stanford, Communication)
- Naeemul Hassan (Mississippi)
- Wei Hu (Nanjing University)
- Arijit Khan (Nanyang Technological University)
- Angela Lee (UTD, Communication)
- Zhiqiang Lin (Ohio State)
- Ping Luo (Chinese Academy of Sciences)
- Mark Stancel (Duke, Public Policy)
- Mark Tremayne (UTA, Communication)
- Min Wang (Google)
- Xifeng Yan (UCSB)
- Jun Yang (Duke)
- Cong Yu (Google Research)
- Nan Zhang (Penn State)

Funding Sponsors


Entrepreneurship

International Workshop on Misinformation, Computational Fact-Checking and Credible Web

May 14, 2019, San Francisco, CA, USA

Co-located with The Web Conference 2019

Thank you!