

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

Chengkai Li

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

Texas Health Resources
August 29, 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



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

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

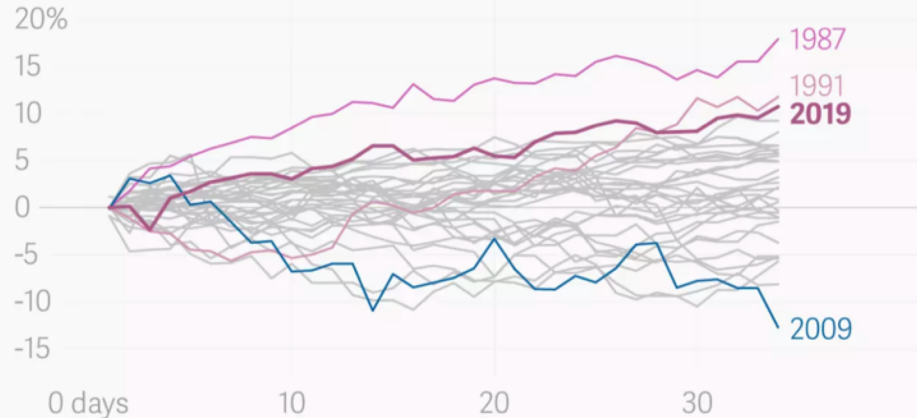
By [Jason Karaian](#) • February 17, 2019

Sign up for the Quartz Obsession email

Sign me up

☐ Stay updated about Quartz products and events.

S&P 500 year-to-date performance since 1987



ATLAS | Data: FactSet

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

Ad closed by Google

Report this ad

Why this ad? ⓘ

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



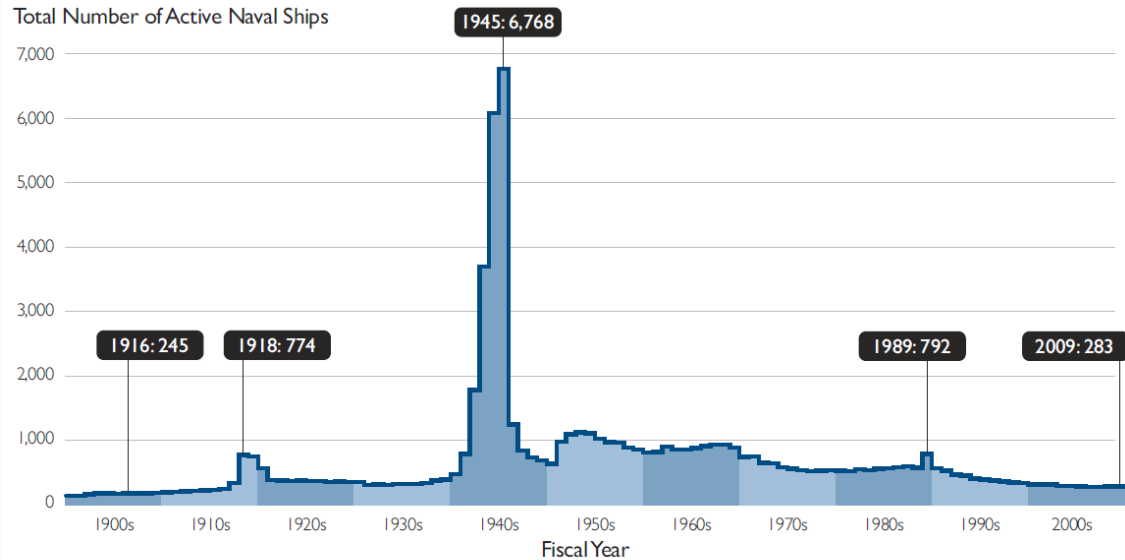
Source: Bill Adair

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.



U.S. Navy Has Smallest Number of Ships Since 1916



Source: U.S. Navy, Active Ship Force Levels, 2009, at <http://www.history.navy.mil/branches/org9-4.htm> (December 6, 2009).

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.



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

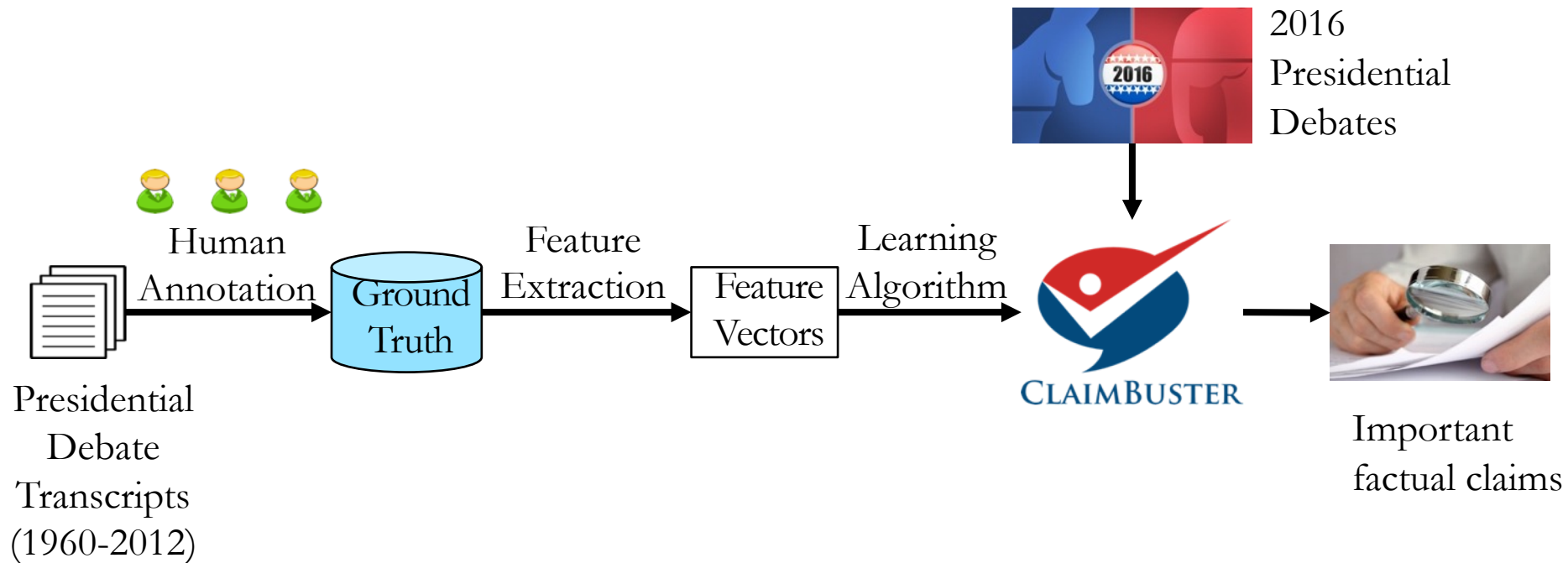
PolitiFact “Buffet” of Factual Claims

	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 pollution 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

Your estimated total payment is \$37.39. You are paid approximately 6.70¢ per sentence on average. Your total payment is ranked 8 out of 54 participants. Click 'Leaderboard' to see details. This message and the leaderboard are updated every 15-30 sentences. **See tips for improving your pay rate at the bottom of this page.**

RO: Now, I voted for the Patriot Act.

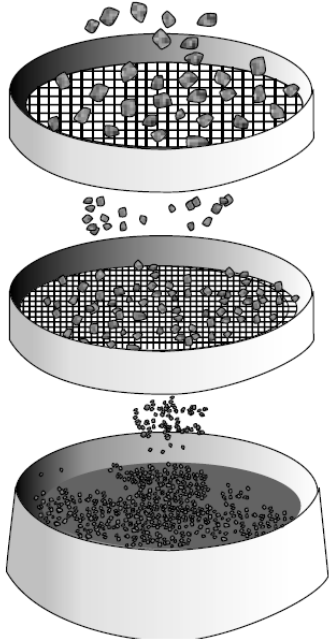
[More Context](#)

Will the general public be interested in knowing whether (part of) this sentence is true or false?

- ☐ There is **no** factual claim in this sentence.
- ☐ There is a factual claim but it is **unimportant**
- ☒ There is an **important** factual claim.

[Submit](#)[Skip this sentence](#)[Modify My Previous Responses](#)

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%





ON AIR

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

Participants: Donald Trump, Hillary Clinton

Moderators: Chris Wallace



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

Participants: Hillary Clinton, Donald Trump

Moderators: Chris Wallace



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

Participants: Hillary Clinton, Donald Trump

Moderators: Martha Raddatz, Anderson Cooper, Chris Wallace



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

Participants: 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

Participants: Hillary Clinton, Donald Trump

Moderators: Lester Holt



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

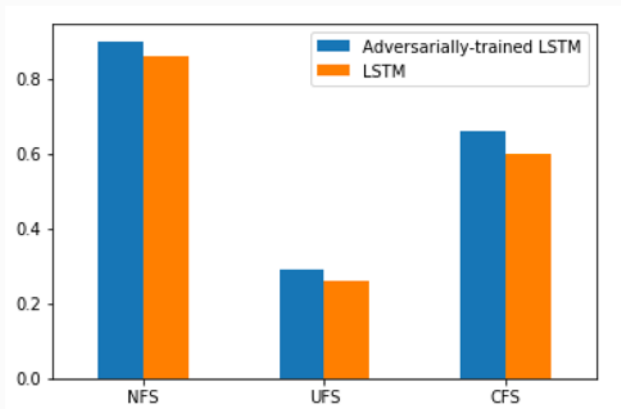
Participants: Bernard Sanders, Hillary Clinton, Donald Trump

- **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.

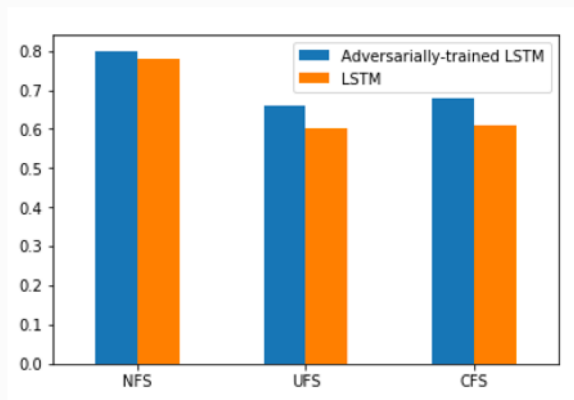
Adversarially Trained LSTM

- 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.

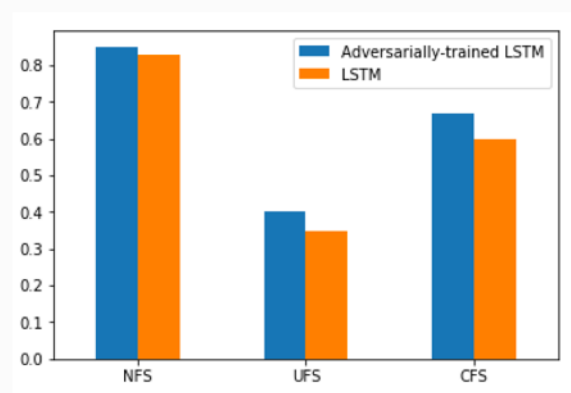
Adversarially Trained LSTM



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 Nebraska, to perform the duties of the Chair.	Senate	Link
Mr. REED	The business show brings new businesses and existing businesses together, putting them in contact with each other and the general public.	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 sister Gwendolyn, his five children and their families as well as his companion of 43 years and her family.	House	Link
Mr. REED	During World War II, the Chamber assisted with the selling of war bonds to aid in the war effort.	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 Commerce manages the management of real estate.	House	Link
Mr. REED	I congratulate the Tioga County Chamber of Commerce for supporting businesses in our community.	House	Link
The PRESIDING OFFICER	Mr. SASSE thereupon assumes the Chair.	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 pollution 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

This Washington Post fact check was chosen by a bot

BY DANIEL FUNKE · JANUARY 30, 2018

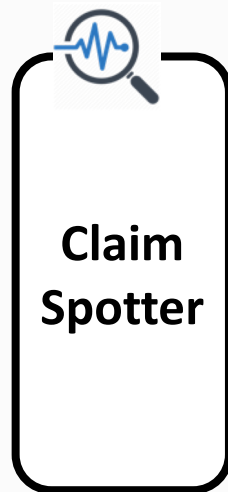


Duke Reporters' Lab uses ClaimBuster API in creating daily newsletters that recommend to The Washington Post, New York Times, PolitiFact, and other fact-checkers the most check-worthy claims in CNN programs, Tweets, Facebook posts, and the Congressional Record.

"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

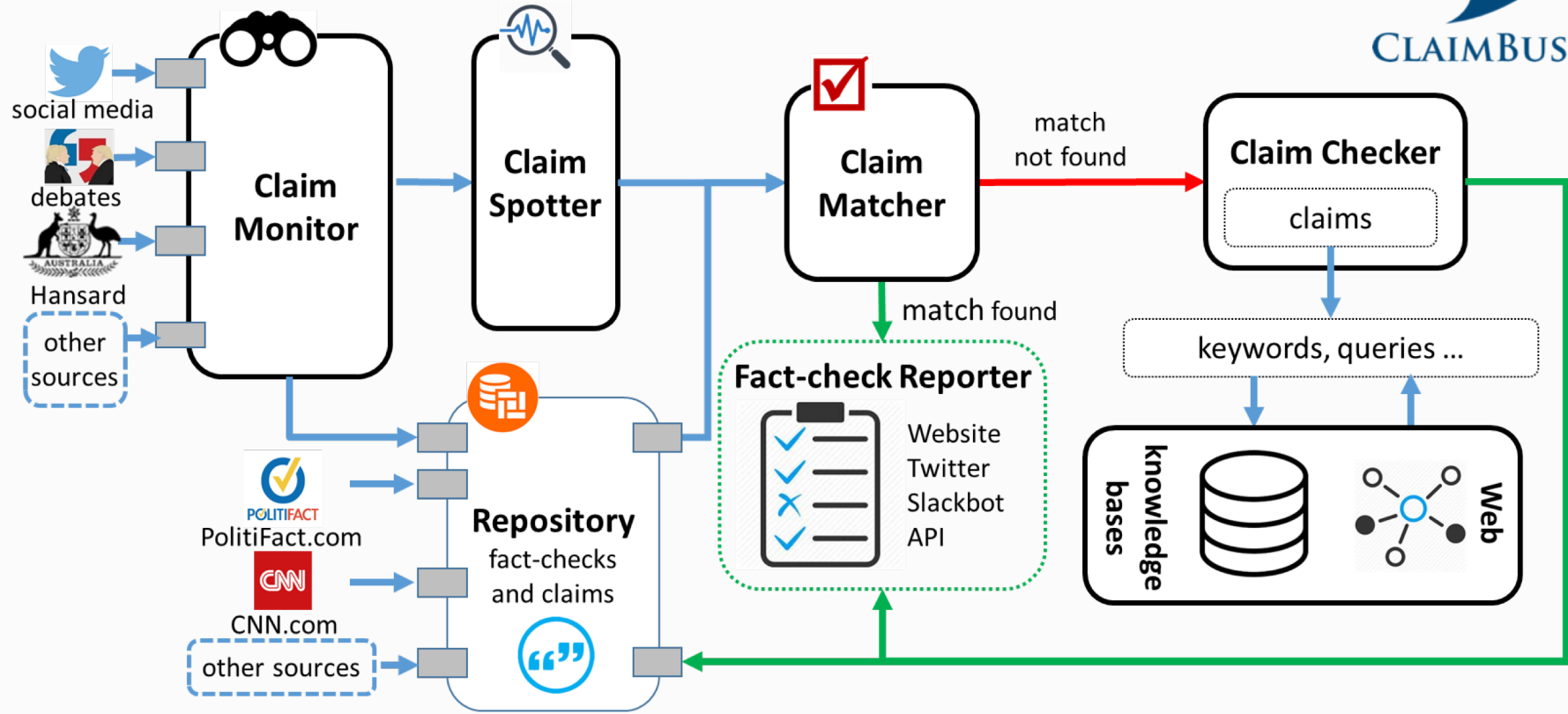
"I had actually looked at that transcript but I missed that claim," said Glenn Kessler, who runs The Fact Checker. "So it would have been lost to history if it had not been for ClaimBuster."



The First-ever End-to-end Fact-checking System ^{iDiR}

[CIKM15a, KDD17, VLDB17demo, IJCNN18]

* First Runner-Up, SIGMOD17 Student Research Competition





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](#)

Chronological Order Order by Score

Least Check-worthy $\geq 0.1 \geq 0.2 \geq 0.3 \geq 0.4 \geq 0.5 \geq 0.6 \geq 0.7 \geq 0.8 \geq 0.9 \geq 1.0$ 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** borders. 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 Received

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

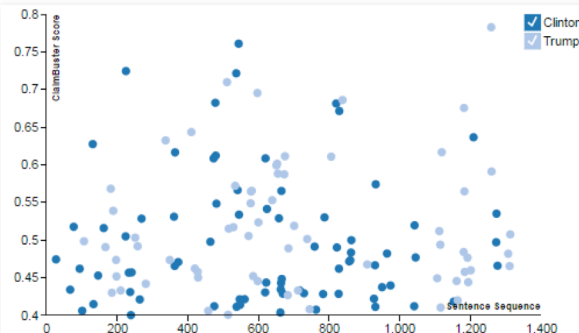
Is that true? What is the source?

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)

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

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

Home

Analytics

How it works?

Watch video

Search keywords

Industries/hashtags

Claim type

From these accounts

Monitoring these accounts

Date from: 2019-Apr-11

Date until: 2019-May-10

Sort by: ClaimBuster score (highest first)

Search

0.0 - 0.1

0.1 - 0.2

0.2 - 0.3

0.3 - 0.4

0.4 - 0.5

0.5 - 0.6

0.6 - 0.7

0.7 - 0.8

0.8 - 0.9

0.9 - 1.0

Kaifan Collins
@kaifanollins

President Trump's reelection campaign raised more than \$30 million in the first fundraising quarter of 2019, per the campaign. Almost 99 percent of those donations were \$200 or less, and they now have \$40.8 million on hand.
885 6:34 PM - Apr 14, 2019

826 people are talking about this

Kenneth P. Vogel
@kenvogel

NEW: The TRUMP campaign says in Q1, 98.79% of its donations* were from \$200 or less, with \$34 avg donation.

*this is % of donations of \$200 or less, not % of \$ from such donations or even % of small donors.

Still Trump is an undeniable small \$ force.
nytimes.com/2019/04/14/us/...
15 7:19 PM - Apr 14, 2019

186 people are talking about this

Trump's 2020 Campaign Raises Over \$30 Million in First Quarter
Mr. Trump's haul is about the size of what Senators Bernie Sanders and Kamala Harris — the top two Democratic fund-raisers — raised
nytimes.com

21 people are talking about this

Senator Mike Crapo
@MikeCrapo

According to [this](#) [NYTimes](#) data through the end of March showed a 25 percent decrease in individual tax liability. That's great news for the American taxpayer and our economy! The [ETCJA](#) has allowed more Americans to see larger paychecks throughout the year!
apnews.com/2019/04/20/news/...

See Senator Crapo's other Tweets

Press release content.
The AP news staff was not involved in its creation.

James Hohmann
@jameshohmann

Replying to @jameshohmann
Kamala Harris released 15 years of tax returns on Sunday, showing that she and her husband earned \$1.8 million last year and gave \$27,000 to charity — or 1.4%. Harris reported no charitable giving at all during her first 3 years as California's attorney general. Not one dollar.
@jameshohmann

67 people are talking about this

Amy Klobuchar
@AmyKlobuchar

AMY KLOBUCHAR and her husband donated \$6,600 of their \$338,500 income to charity last year, under 2%. That's a similar percentage to what Kristin Gilliland, her and her husband gave \$3,750 out of \$215,000 to charity last year.
296 9:20 AM - Apr 15, 2019

67 people are talking about this

Shane Goldmacher
@ShaneGoldmacher

NEW: Pete Buttigieg spent less than \$700,000 in Q1 and has \$6.4 million in the bank, per new FEC filing.

His "burn rate." Less than 10% [discovery.fec.gov/govinfo/form/...](#)
689 8:06 AM - Apr 15, 2019

186 people are talking about this

Urban Institute
@urbaninstitute

In 1972, the average American urban carpenter earned the current equivalent of \$33.55/hour—about \$70,000 a year. Today, a carpenter earns an average of \$20.23/hour—or about \$42,000 annually, reports [this](#) [Urban Institute](#) [report](#).
8 10:47 AM - Apr 11, 2019

See Urban Institute's other Tweets

Middle-wage jobs largely no longer pay middle-class salaries
A primary complaint about the current U.S. economy has been the loss...
aol.com

See Urban Institute's other Tweets

David Beard
@DavidBeard

Netflix, which is hiking monthly fees, made \$856 million in pretax U.S. income last year. It didn't pay a dime in federal taxes. In fact, it took \$22 million from actual U.S. taxpayers as a "rebate." It's among a [Deadbeat Club](#) of 60 corporate tax dodgers, [publications.nytimes.com/business/faxes...@Publici](#)
136 6:24 AM - Apr 13, 2019

151 people are talking about this

Ian Bremmer
@ianbremmer

Urban Area Security Initiative grants are essential for law enforcement training and preparedness for acts of terrorism in urban areas," said Rep. Donald M. Payne, Jr. (NJ-08), "Newark and Jersey City are among the most densely populated areas in the United States. New residents and emergency preparedness professionals are constantly training, planning, and equipping themselves to respond to the threat of terrorism in our communities. They rely on USIA grant funding to ensure that New Jersey is a safe place to live, work, and raise families. I am deeply troubled that the Federal Emergency Management Agency is cutting USIA grant funding for the Newark and Jersey City area. As these cities grow, they face unique threats that require law enforcement to constantly adapt. Last year in the Newark and Jersey City area, an anti-abortion extremist stole a truck and rammed it into a Planned Parenthood clinic in East Orange, New Jersey. By taking \$2.7 million of USIA grant funding from the region, FEMA is forcing the state of New Jersey to make hard budget choices in an environment where local actors are constantly challenging the resources of state and local governments. As Chairman of the Subcommittee on Emergency Preparedness, Response, and Recovery for the House Committee on Homeland Security, I oversee the U.S. Department of Homeland Security's emergency grants, such as USIA, and I will continue to push FEMA to help law enforcement ensure the safety of our communities."

See Rep. Donald Payne Jr's other Tweets

Steven Greenhouse
@greenhouseyt

Elizabeth Warren's new tax plan would impose a 7% tax on corporate profits above \$100 million with no "loopholes or exemptions." Under her plan, Amazon would pay \$668 million in corporate taxes, instead of \$100, on its \$10 billion in profits.
[@ashleypurtonbloomberg.com/news/articles/...](#)
105 9:28 AM - Apr 12, 2019

58 people are talking about this

Bernie Sanders
@BernieSanders

Amazon made \$10,835,000,000 last year. Its effective federal tax rate was -1%.

Delta made \$5,073,000,000. Its effective tax rate was -4%.

Our rigged tax code has essentially legalized tax dodging for large corporations.
23K 3:14 PM - Apr 12, 2019

8,092 people are talking about this

GOP
@GOP

Bernie's \$32 trillion plan for the complete elimination of private health insurance would:
-Eradicate over half a million good-paying jobs
-More than double taxes
-Kick over 177 million Americans off their health insurance plans
Destroy choices for families across America.
3,729 6:15 PM - Apr 15, 2019

2,468 people are talking about this

Jim Banks
@RepJimBanks

Under [H.R. 1](#) leadership:
- 3.2 million new jobs created
- Unemployment near 50-year low
- Jobless claims lowest since 1969
- Manufacturing job creation highest in 20 years
- Longest streak of consecutive job growth

Great piece in [@JONLeary](#) today. [bit.ly/2vQdgl](#)
83 8:57 AM - Apr 15, 2019

See Jim Banks's other Tweets

IC
@IC

GOP brings boom times I Home I Journal Gazette
This Tax Day, we celebrate higher wages, record economic optimism, record low unemployment and

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

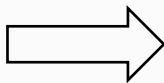
Chris Paul had 16 points, 10 rebounds, 13 assists and five steals..... The only other active player to have such a game is Jason Kidd...

2009

How did they come up with that?

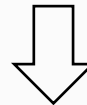


Exceptional Fact Finding

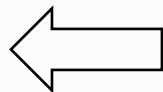


New tuple appended to database

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	Strickland	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.

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

Number-based facts make news stories more engaging

»LIVE UPDATE

[February 20, 1998] **Todd Fuller** had 1 assist, 3 steals and 1 block in the Golden State Warriors' defeat against the Denver Nuggets. It is one of the best performance made by him.

SEARCH michael jordan

Michael Adonis Jordan

Michael Jordan

Michael Michael Jordan

Michael Reggie Jordan

Michael Thomas Jordan

[January 13, 1997] **Horace Grant** had 26 points and 6 assists in the Orlando Magic's victory against the New Jersey Nets. It is one of the best performance made by him.

MORE LIKE THIS

[January 13, 1997] After the Orlando Magic's win over the New Jersey Nets, for the first time in his career, **Rony Seikaly** had at least 20 points for 6 consecutive games, after today's game.

MORE LIKE THIS

[January 13, 1997] **Horace Grant** had 26 points and 2 steals in the Orlando Magic's victory against the New Jersey Nets. It is one of the best performance made by him.

MORE LIKE THIS

[January 13, 1997] **Horace Grant** had 26 points, 6 assists and 2 steals in the Orlando Magic's victory against the New Jersey Nets. It is one of the best performance made by him.

MORE LIKE THIS

[January 13, 1997] After the Orlando Magic's victory against the New Jersey Nets, for the first time in his career, **Rony Seikaly** had at least 20 points and 8 rebounds for 6 consecutive games, after today's game.

MORE LIKE THIS

[January 13, 1997] **Nick Anderson** had 8 assists and 2 blocks in the Orlando Magic's win over the New Jersey Nets. It is one of the best performance made by him.

MORE LIKE THIS

[January 4, 1995] After the Orlando Magic's win over the New Jersey Nets, for the first time in

FACT TYPE

SITUATIONAL FACT

PROMINENT STREAK

ONE-OF-THE-FEW

RANKING

RECENTNESS

INTERESTINGNESS

POPULARITY

PLAYERS

TEAMS

SEASONS

1996-97 (9)

1994-95 (5)

1992-93 (1)

+MORE

LESS-

MONTHS

OPPOSITION TEAMS

POINTS



Excellent
Demo Award

idir.uta.edu/factwatcher

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.

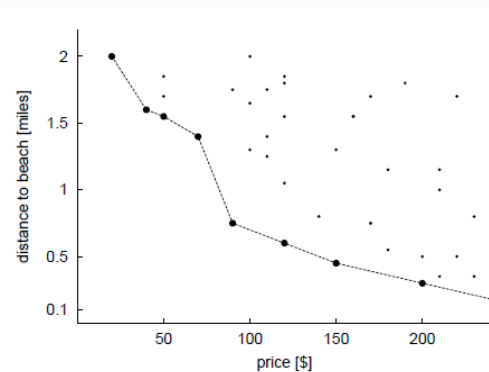


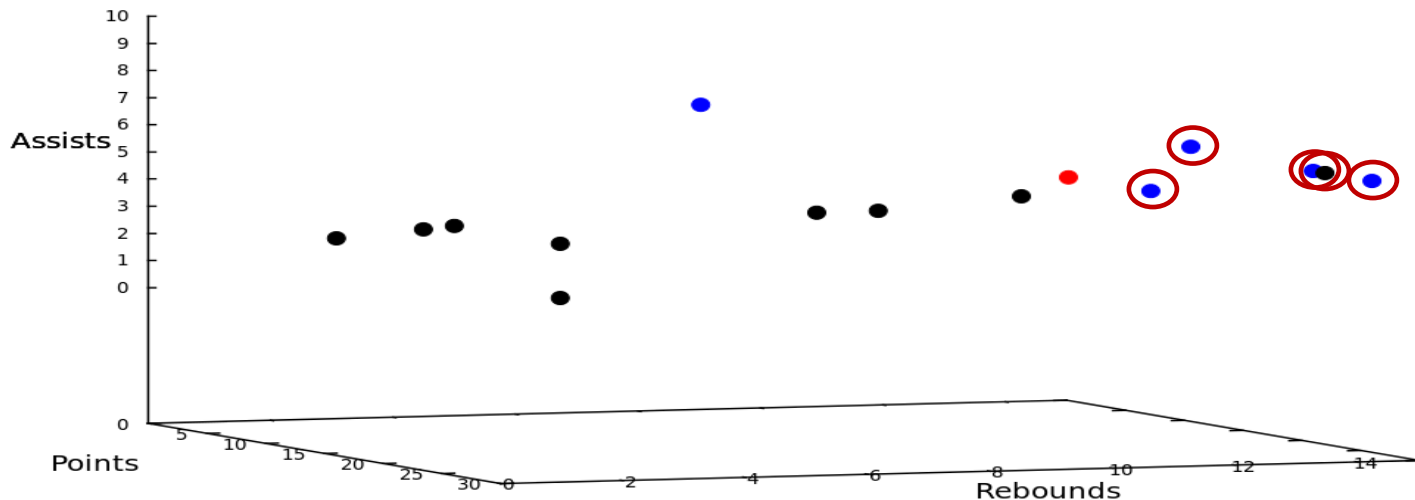
Figure 1: Skyline of Hotels



Figure 2: Skyline of Manhattan

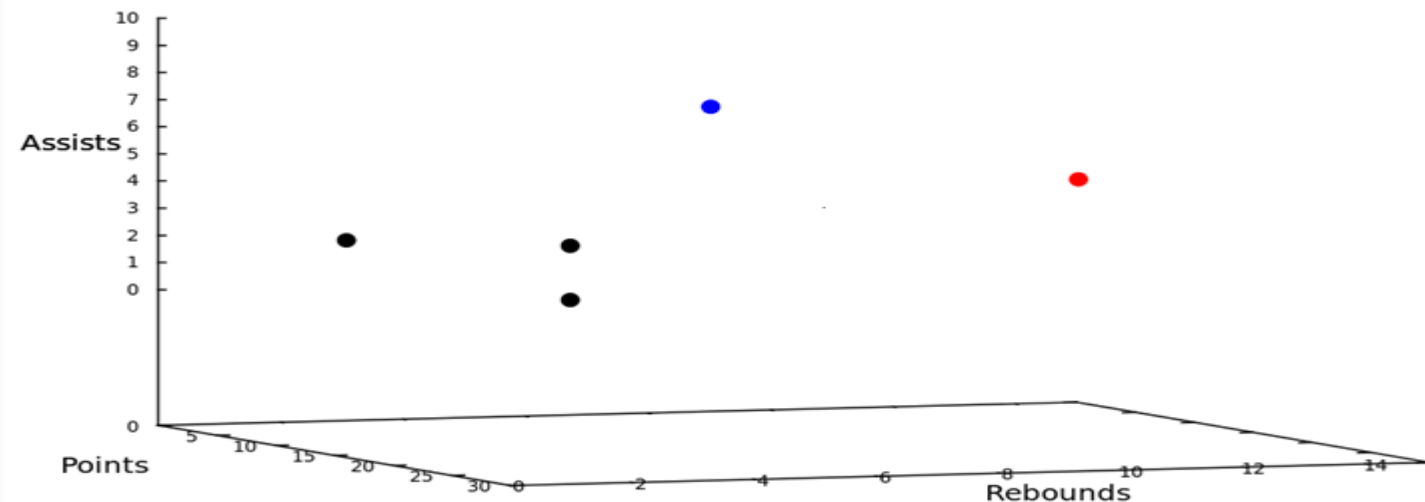
Modeling Situational Facts

“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>)



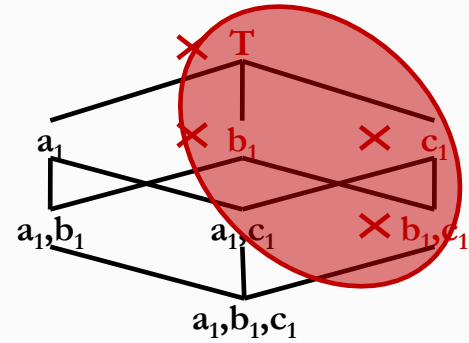
Modeling Situational Facts

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

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

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

IDiR Projects and Demos

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

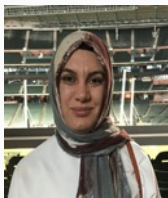
- [ClaimBuster](http://idir.uta.edu/claimbuster) (idir.uta.edu/claimbuster) Automated, live fact-checking
- [ClaimPortal](http://idir.uta.edu/claimportal) (idir.uta.edu/claimportal) Monitoring factual-claims on social media
- [CrewScout](http://idir.uta.edu/crewscout) (idir.uta.edu/crewscout) Expert team finding by skyline groups
- [ERQ: Entity-Relationship Query](http://idir.uta.edu/erq) (idir.uta.edu/erq) Structured query on Wikipedia
- [Facetedpedia](http://idir.uta.edu/facetedpedia) (idir.uta.edu/facetedpedia) Faceted search interface for Wikipedia
- [FactWatcher](http://idir.uta.edu/factwatcher) (idir.uta.edu/factwatcher) Fact-finding from real-world events
- [FrameAnnotator](http://idir.uta.edu/frameannotator) (idir.uta.edu/frameannotator) Frame annotation tool
- [GQBE](http://idir.uta.edu/gqbe) (idir.uta.edu/gqbe) Graph query by example
- [Orion](http://idir.uta.edu/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

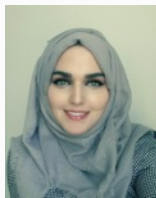
Ph.D. students



Farahnaz
Akrami



Fatma
Arslan



Israa
Jaradat



Damian
Jimenez



Shadekur
Rahman



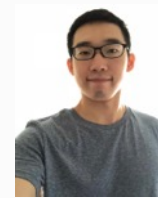
Samiul
Saeef



Xiao
Shi



Theodora
Toutountzi



Zeyu
Zhang

M.S. students



Priyank
Arora



Sumeet
Lubal



Sarthak
Majithia

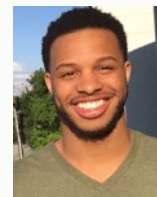


Daniel
Obembe

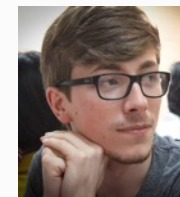


Sarbajit
Roy

B.S. students



Kyrell
Dixon



Jacob
Devasier

Graduated Students

Ph.D. students



Aditya Telang
2011 (co-advised)
IBM Research India



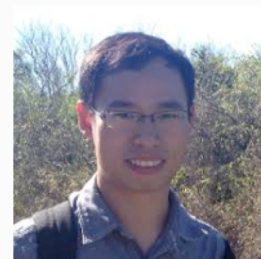
Ning Yan
2013
Research Scientist
Huawei, Santa Clara



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



CORPS
NSF Innovation Corps

Thank you!