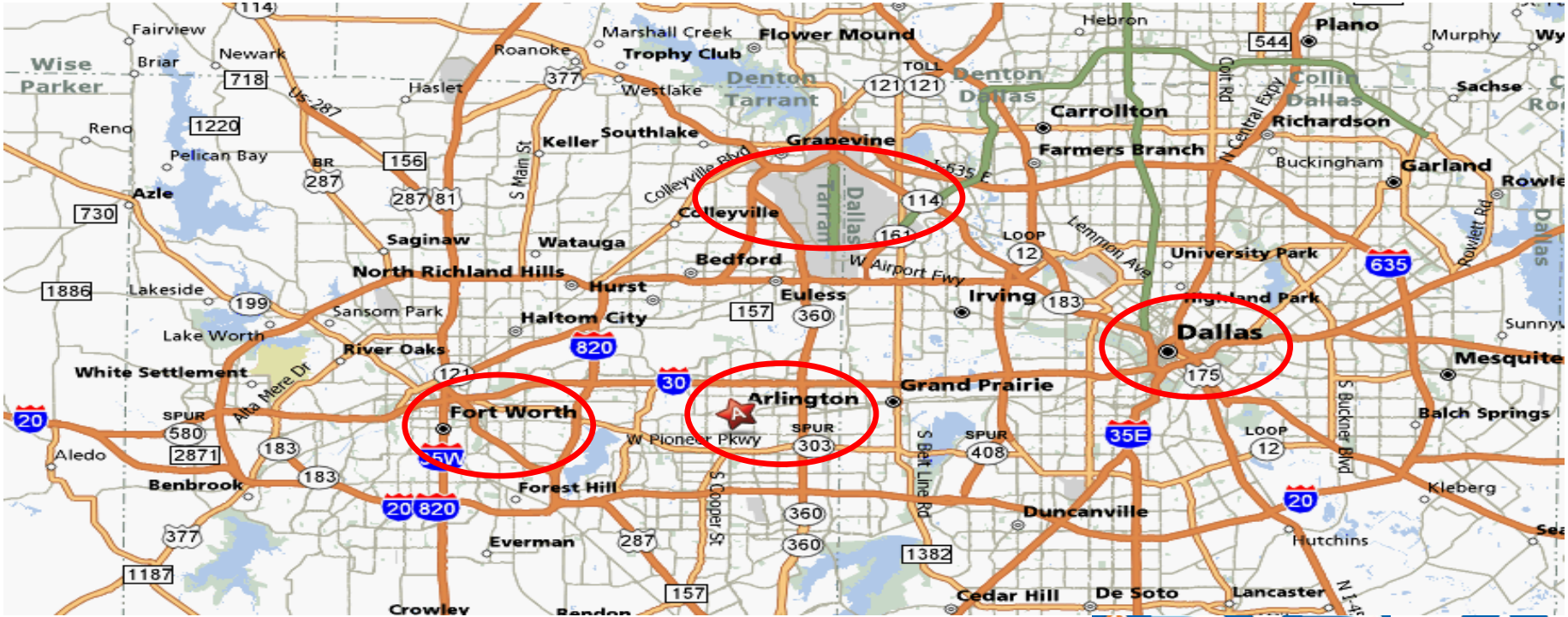
Enabling Computational Journalism:
Automated Fact-Checking and Story-Finding

Chengkai Li
Associate Professor, Department of Computer Science and Engineering
Director, Innovative Database and Information Systems Research (IDIR) Laboratory
University of Texas at Arlington

Shandong University, Oct. 9th, 2015
Nanjing University, Oct. 13th, 2015

Dallas-Fort Worth Metroplex, Texas



The University of Texas at Arlington

"The University of Texas at Arlington is a growing research powerhouse committed to life-enhancing discovery, innovative instruction, and caring community engagement. An educational leader in the heart of the thriving North Texas region, UT Arlington nurtures minds within an environment that values excellence, ingenuity, and diversity."

- Founded in 1895
- One of the 9 academic institutions in The University of Texas System
- 51,000 students (37,008 on-campus), second-largest institution within the UT System
- Classified by the Carnegie Foundation as a "Research University/High Research Activity " institution
- One of the fastest growing public research universities in the nation

Department of Computer Science and Engineering

- 29 tenured and tenure-track faculty, 11 non-tenure track faculty, 17 adjunct faculty
- ~110 Ph.D. students, ~800 M.S. students, ~1000 B.S. students

Ranking

- #19 Data Mining, #19 Machine Learning and Pattern Recognition, #35 Databases (source: Microsoft Academic Search, 2014)
- Computer Engineering (Graduate) #83 (source: U.S. News, 2016)
- Computer Science (Graduate) #90 (source: U.S. News, 2014)

Major areas of research: Big Data and Large-Scale Computing, Biocomputing and Health Informatics, Computer Networks, Computer Vision and Multimedia, Database and Information Systems, Embedded Systems and Mobile Computing, Machine Learning and Data Mining, Robotics and AI, Security and Privacy, Software Engineering, Sustainable Computing

The Innovative Database and Information Systems Research (IDIR) Laboratory

Research areas

- Big Data and Data Science (Database, Data Mining, Web Data Management, Information Retrieval)

Theme of current research

- building large-scale human-assisting and human-assisted data and information systems with high usability, low cost and applications for social good

Research directions

- computational journalism
- crowdsourcing and human computation
- data exploration by ranking/skyline/preference queries
- database testing
- entity search and entity query
- graph database usability

Our Computational Journalism Project

Started in 2010. Collaborative project with Duke, Google Research, HP Labs, Stanford

- Story finding: finding and monitoring number-based facts pertinent to real-world events. The facts are leads to news stories.
- Fact checking: discovering and checking factual claims in political discourses, social media, and news.

Publications

- Detecting Check-worthy Factual Claims in Presidential Debates. Naeemul Hassan, Chengkai Li, Mark Tremayne. CIKM 2015, pages 1835-1838.
- The Quest to Automate Fact-Checking. Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang and Cong Yu. 2015 Computation+Journalism Symposium.
- Online Frequent Episode Mining. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, and Qing He. ICDE 2015, pages 891-902.
- Data In, Fact Out: Automated Monitoring of Facts by FactWatcher. Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. VLDB 2014, pages 1557-1560. Demonstration description. (**excellent demonstration award**)
- Finding, Monitoring, and Checking Claims Computationally Based on Structured Data. Brett Walenz, You (Will) Wu, Seokhyun (Alex) Song, Emre Sonmez, Eric Wu, Kevin Wu, Pankaj K. Agarwal, Jun Yang, Naeemul Hassan, Afroza Sultana, Gensheng Zhang, Chengkai Li, Cong Yu. 2014 Computation+Journalism Symposium.

Publications (cont'd)

- Toward Computational Fact-Checking. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. VLDB 2014, pages 589-600.
- iCheck: computationally combating "lies, d-ned lies, and statistics". You Wu, Brett Walenz, Peggy Li, Andrew Shim, Emre Sonmez, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. SIGMOD 2014, pages 1063-1066.
- Incremental Discovery of Prominent Situational Facts. Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, Cong Yu. ICDE 2014, pages 112-123.
- Discovering General Prominent Streaks in Sequence Data. Gensheng Zhang, Xiao Jiang, Ping Luo, Min Wang, Chengkai Li. ACM TKDD, 8(2):article 9, June 2014.
- Discovering and Learning Sensational Episodes of News Events. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, Qing He, and Zhongzhi Shi. WWW 2014, pages 217-218.

Publications (cont'd)

- On "One of the Few" Objects. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. KDD 2012, pages 1487-1495.
- Prominent Streak Discovery in Sequence Data. Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, Yong Yu. KDD 2011, pages 1280-1288.
- Computational Journalism: A Call to Arms to Database Researchers. Sarah Cohen, Chengkai Li, Jun Yang, Cong Yu. CIDR 2011, pages 148-151. (**3rd place in best Outrageous Ideas and Vision (OIV) Track paper competition**)

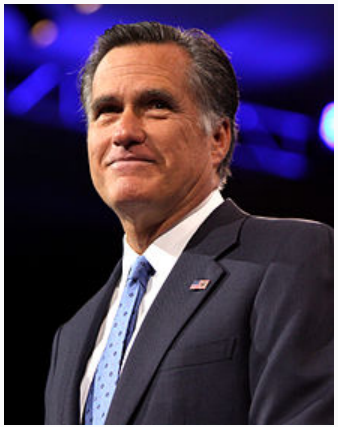
CLAIMBUSTER

The Quest to Automate Fact-Checking

iDiR

People Make Claims All The Time

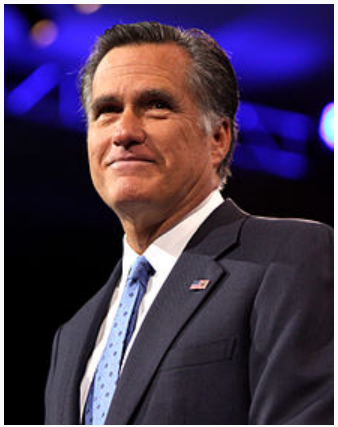
"… our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in third presidential debate in 2012.

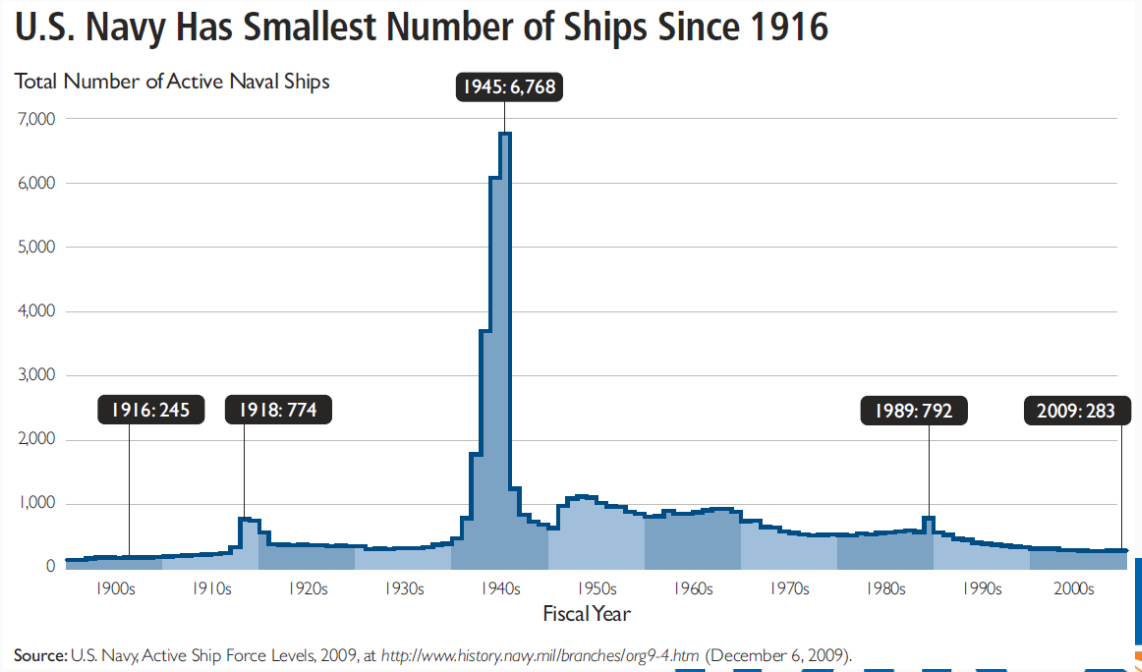


http://en.wikipedia.org/wiki/Mitt_Romney
http://www.thebrainchildgroup.com/

Fact Checking is not Easy

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

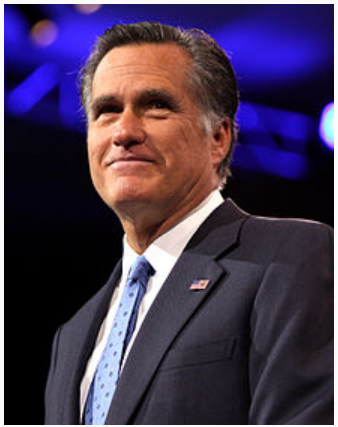


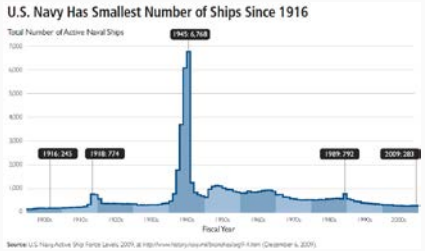


http://en.wikipedia.org/wiki/Mitt_Romney
http://s3.amazonaws.com/thf_media/2010/pdf/Military_chartbook.pdf

Fact Checking is not Easy

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





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

Existing Fact Checking Projects

Journalists and reporters spend good amount of time on fact checking

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 http://www.politifact.com/

FactCheckEU https://factcheckeu.org/

FullFact http://fullfact.org/

Snopes http://www.snopes.com/info/whatsnew.asp

Factcheck http://www.factcheck.org/

Numerous Claims to Check. Rise of Fact-Checkers

Republican candidate debate, August 6, 2015.[1]

- 9 facts checked by factcheck.org
- 8 facts checked by CNN
- 24 facts checked by PolitiFact

64 active fact-checking sites in 2015, 44 in 2014. [2]

1. http://time.com/3988276/republican-debate-primetime-transcript-full-text/
2. http://reporterslab.org/snapshot-of-fact-checking-around-the-world-july-2015/

Limitations of Current Fact-Checking Practices

- Journalists spend hours going through documents to identify claims.
- Significant time gap between speech and reporting times. Audience doesn't get correct information.
- Requires advanced writing skills to persuade readers. Such skilled writers are sparse.
- Lack of Structured Journalism and use of old publishing frameworks hinders Semantic Web applications.

The Holy Grail: Automated, Live Fact-Checking

The Holy Grail

Source: Bill Adair

The Holy Grail

Source: Bill Adair

The Holy Grail

Source: Bill Adair

ClaimBuster

- political discourses (debates, interviews), advertisements, live events on TV and online video streams
- social media (e.g., twitter)
- web pages
- news articles

repository of already-checked claims

detector → important factual claims → matcher

matcher → no match

matcher → matched

assist data analysis; solicit analyses from professionals

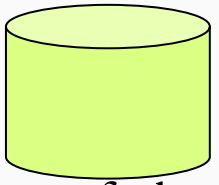
display existing fact-checkers, delivered via browser extensions, mobile and smart-TV apps

www.wsandb.co.uk/wsb/news/2405401/european-court-rules-insurers-need-to-include-economic-consequences-of-policies#
www.outsidethebeltway.com

ClaimBuster

- political discourses (debates, interviews), advertisements, live events on TV and online video streams
- social media (e.g., twitter)
- web pages
- news articles

assist data analysis; solicit analyses from professionals



repository of already-checked claims

detector → important factual claims → matcher

no match

matched

display existing fact-checkers, delivered via browser extensions, mobile and smart-TV apps

www.wsandb.co.uk/wsb/news/2405401/european-court-rules-insurers-need-to-include-economic-consequences-of-policies#
www.outsidethebeltway.com

iCheck (Led by Duke)

✔Kay Hagan is overly partisan.

Republicans suggest her achievements are thin soup. "The only thing Kay Hagan has accomplished in Washington is becoming an automatic 'yes' vote for whatever new tax or regulation President Obama wants," North Carolina GOP Chairman Claude Pope said.

Original claim made by: 1

Supporting Arguments

7 ⬆ ⬇	Frank R. Lautenberg(D) and Kay Hagan(D) agreed on 90.02% of the votes they cast between 2011-01-05 (start of session 2011) and 2013-01-01 (end of session 2012)
3 ⬆ ⬇	Thomas Harkin(D) and Kay Hagan(D) agreed on 92.6% of the votes they cast between 2013-01-03 (start of session 2013) and 2014-05-09 (end of session 2014)

Counter Arguments

Generated Counter Arguments

+	Bernard Sanders(Independent) and Kay Hagan(D) agreed on 85.77% of the votes they cast between 2011-01-05 (start of session 2011) and 2013-01-01 (end of session 2012)
+	Kay Hagan(D) voted 89.83% of the time with the Democrat party majority vote between 2008-01-02 and 2012-01-02.

iCheck (Led by Duke)

In 2011, Miguel Cabrera had 197 in hits, 30 in homeruns, 0.34 in batting average; only 6 other players have ever beaten this record;

Bump! Tweet!

Views: 1 | Bumps: 0 | Tweets: 0

Vladimir Guerrero: 197 in hits, 44 in homeruns, 0.35 in batting average in 2000
Todd Helton: 216 in hits, 42 in homeruns, 0.37 in batting average in 2000; 209 in hits, 33 in homeruns, 0.36 in batting average in 2003
Mike Piazza: 201 in hits, 40 in homeruns, 0.36 in batting average in 1997
Albert Pujols: 212 in hits, 43 in homeruns, 0.36 in batting average in 2003
Alex Rodriguez: 215 in hits, 36 in homeruns, 0.36 in batting average in 1996
Larry Walker: 208 in hits, 49 in homeruns, 0.37 in batting average in 1997

Responses

The same claim (i.e. "no more than 6 other players have ever beaten this player's record in some year in 'hits', 'homeruns', 'batting average') can be made for 41 other players.

The other player are: Albert Belle in 1995 (4), in 1998 (1); Adrian Beltre in 2004 (1); Dante Bichette in 1995 (6), in 1998 (3); Barry Bonds in 2001 (0), in 2002 (0), in 2003 (1), in 2004 (1); Bret Boone in 2001 (6); Ellis Burks in 1996 (2); Vinny Castilla in 1998 (1); Carlos Delgado in 2000 (4); Jacoby Ellsbury in 2011 (4); Darin Erstad in 2000 (0); Nomar Garciaparra in 2000 (1); Adrian Gonzalez in 2011 (4); Luis Gonzalez in 2001 (0); Ken Griffey in 1997 (3), in 1998 (6); Vladimir Guerrero in 2000 (1), in 2002 (4), in 2004 (4); Tony Gwynn in 1995 (2), in 1997 (0); Josh Hamilton in 2010 (3); Todd Helton in 2000 (0), in 2001 (1), in 2003 (1), in 2004 (4); Matt Holliday in 2007 (1); Ryan Howard in 2006 (1); Derek Jeter in 1999

ClaimBuster to be 2016-Ready

2016 Presidential Debates

(Speeches, debates, interviews, social media, news)

CLAIMBUSTER

Factual claims recommended to be checked

Journalists investigate the claims

(or checked by algorithms, citizens, crowd)

Finding Important Factual Claims: A Classification Problem

2016 Presidential Debates

Human Annotation

Feature Extraction

Learning Algorithm

Presidential Debate Transcripts

Ground Truth

Feature Vectors

CLAIMBUSTER

Important factual claims

Dataset: Presidential Debate Transcripts

- Source: http://www.debates.org/index.php?page=debate-transcripts
- All 30 debates (11 elections) in history: 1960, 1976—2012
- 20k sentences by presidential candidates: removed very short (< 5 words) sentences

3 Classes of Sentences

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."

Unimportant factual claims

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

Sentences with no factual claims (just 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%."

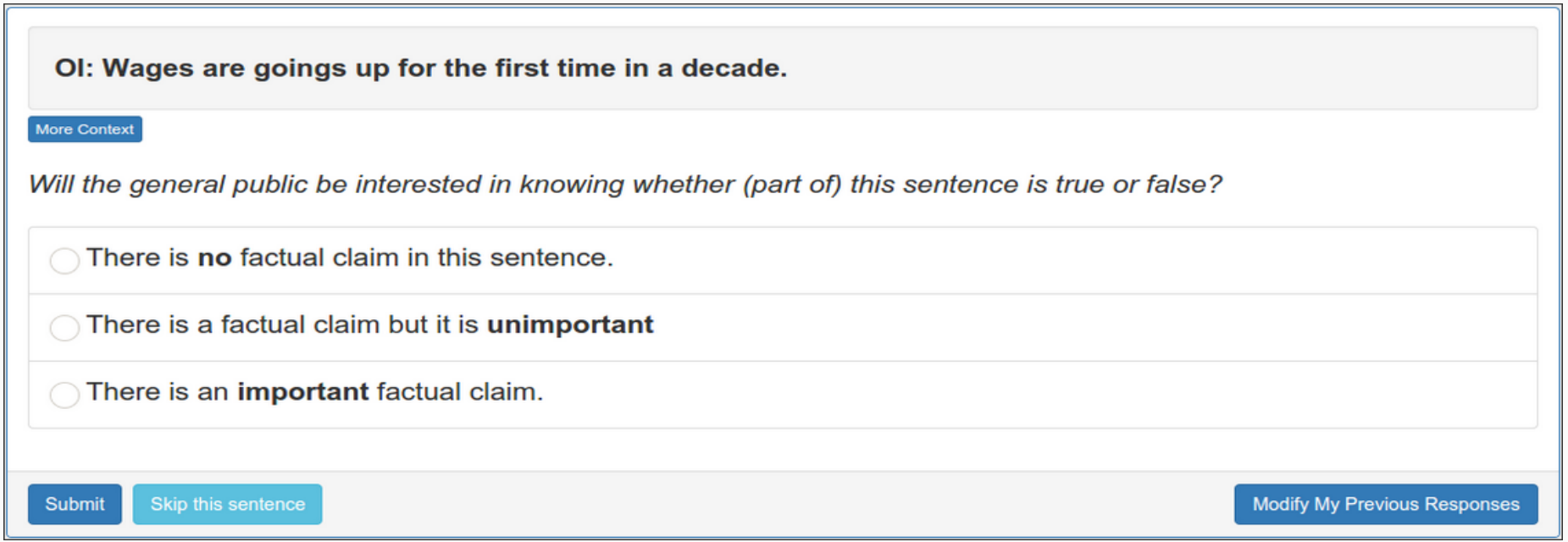
Goal: Given a future sentence, find the class it belongs to.

Ground Truth Collection

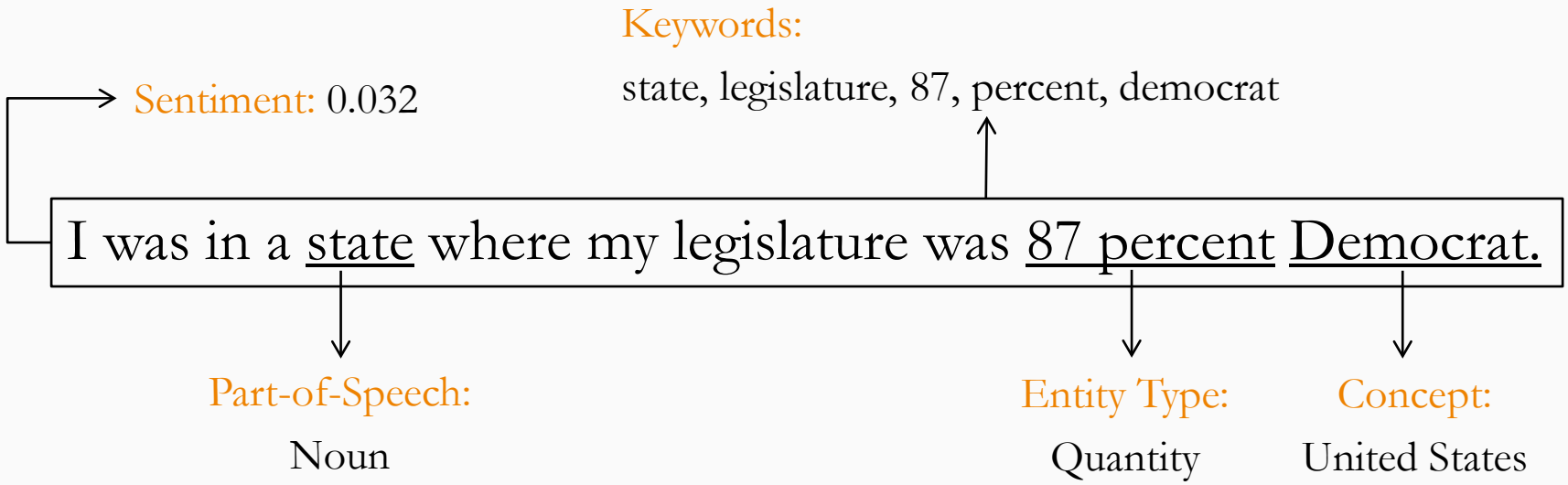
- Developed a data collection platform bit.ly/claimbusters.
- In 3 months, we accumulated 226 participants.
- Used 600 screening sentences to detect spammers & low-quality participants.
- Admitted sentences which are agreed by at least 2 top-quality participants.
- 8015 such sentences.

Class	Count
CFS	1673
UFS	482
NFS	5860

Ground Truth Collection Website



Feature Extraction

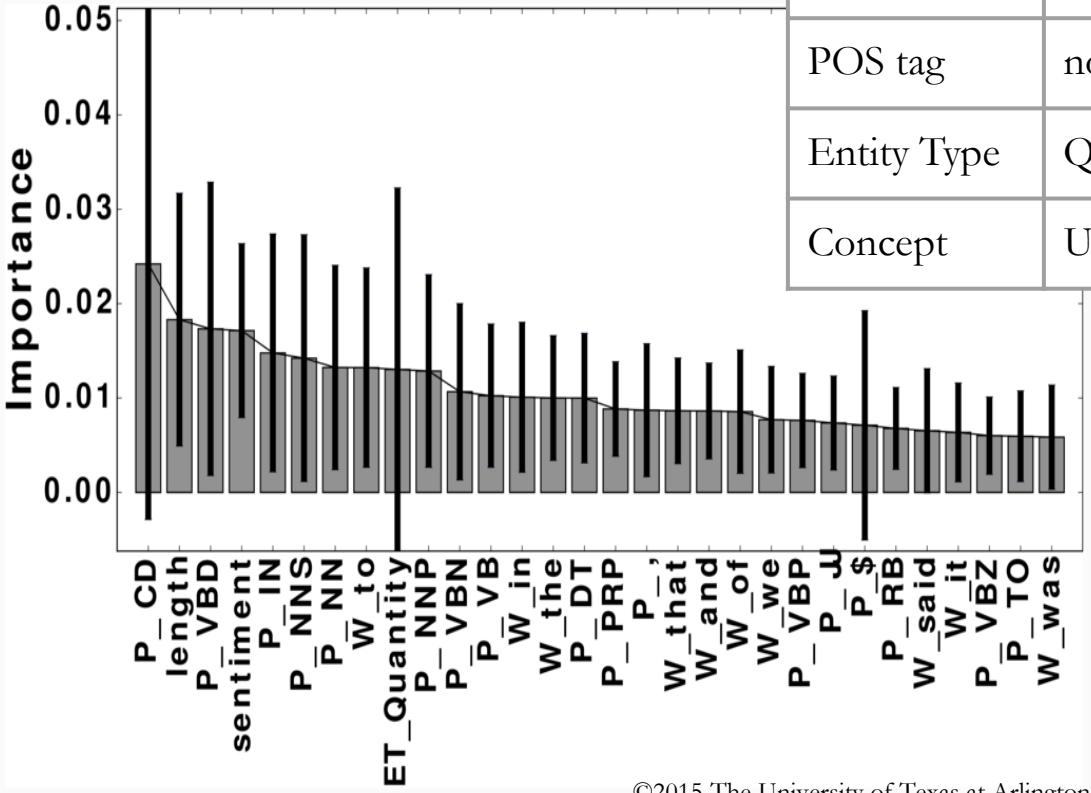


Sentiment: [-1.0 to 1.0]
Words: tf-idf scores of 6130 words (excluding rare words)
POS Tag: 43 tags
Entity Type: 26 types

Feature Selection

- 6201 features in total
- Used a Random Forest Classifier to calculate importance of each feature.
- Most Important Feature: POS tag ‘Cardinal Number’

Important Features



Word	percent; people; jobs
POS tag	noun; cardinal number; past tense; preposition
Entity Type	Quantity; Country; FieldTerminology; Person
Concept	United States Senate; Barack Obama

Implementation: Python NLP/ML Tools

Data wrangling

- Use NLTK (Natural Language Toolkit) to transform debate files into structured data format
- Use mysql-python-connector to store extracted features into an MySQL database
- Use matplotlib to plot classifiers' performance.

Feature extraction

- Use AlchemyAPI (Python wrapper) to extract rich features of sentences

Classification

- Use scikit-learn to build classification models

Evaluation: Classification

- 4-fold cross validation
- Algorithms: Naive Bayes, Random Forest & Support Vector Machine
- Support Vector Machine performed better than others in general.

	Precision	Recall	F-measure
NFS	0.90	0.96	0.93
UFS	0.65	0.26	0.37
CFS	0.79	0.74	0.77

Evaluation: Ranking

- Measured accuracy of top-K sentences.
- ClaimBuster has a strong agreement with high-quality human coders on the check-worthiness of sentences

K	P@K	NDCG@K
25	1	1
50	1	1
100	0.960	0.970
200	0.940	0.951
300	0.853	0.881
500	0.690	0.840

Case Study: #GOPDebate2015

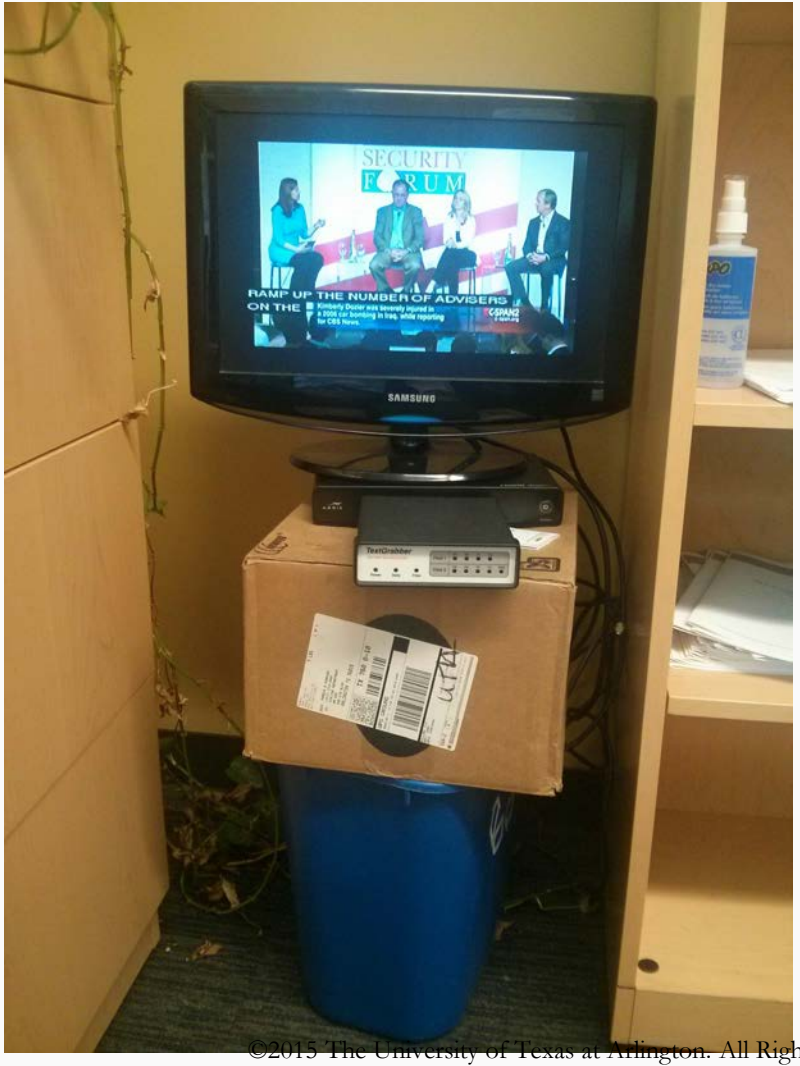
- Near real-time experiment with 2015 first Republican primary debate
- Transcript grabbed from closed captions of the Fox News channel using TextGrabber
- 1393 sentences
- 71% of the fact-checks from CNN, factcheck.org & PolitiFact were ranked by ClaimBuster within top 18%.

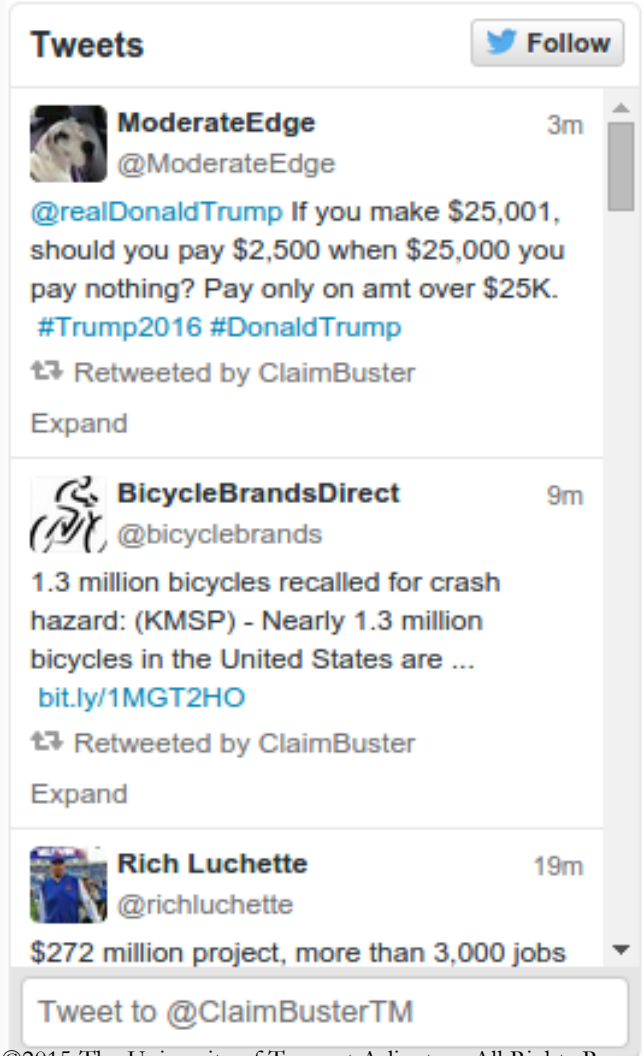
Case Study: #GOPDebate2015

CNN Claim	Associated sentence(s)[From TextGrabber]	Score
1	Part of this iranian deal was lifting the international sanctions on general sulemani.	0.415
2	I would go on to add – >> you don't favor – >> i have never said that.	0.511
3	A majority of the candidates on this stage supported amnesty.	0.295
4	Timely the medicaid is growing at one of the lowest rates in the country.	0.534
4	We went from $8 billion in the hole to $5 million in the black.	0.773
5	And the mexican government is much smarter, much sharper, much more cunning and they send the bad ones over because they don't want to pay for them.	0.215
6	[Not found in the transcript]	N/A

Case Study: #GOPDebate2015

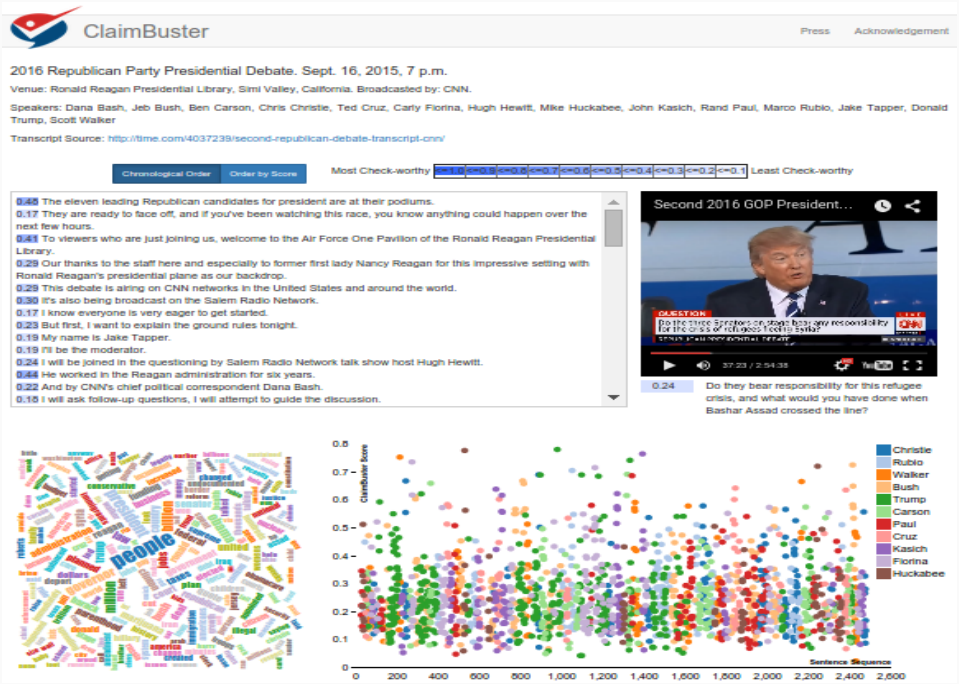
- Real-time experiment with 2015 second Republican primary debate
- Closed Captions from CNN channel
- Tweeted important factual claims to https://twitter.com/ClaimBusterTM, live!





Demo

http://idir.uta.edu/claimbuster



Automated live fact-checking

2016 Republican Party Presidential Debate. Sept. 16, 2015, 7 p.m.

Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker

2016 Republican Party Presidential Debate. Aug. 6, 2015, 8 p.m.

Speakers: Bret Baier, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Mike Huckabee, John Kasich, Megyn Kelly, Rand Paul, Rick Perry, Marco Rubio, Donald Trump, Scott Walker, Chris Wallace

Fact-check your own text

Tweets Follow

ModerateEdge 2m
@ModerateEdge
@realDonaldTrump If you make $25,001, should you pay $2,500 when $25,000 you pay nothing? Pay only on amt over $25K. #Trump2016 #DonaldTrump
Retweeted by ClaimBuster
Expand

BicycleBrandsDirect 8m
@bicyclebrands
1.3 million bicycles recalled for crash hazard: (KMSP) - Nearly 1.3 million bicycles in the United States are ... bit.ly/1MGT2HO
Retweeted by ClaimBuster
Expand

Rich Luchette 18m
@richluchette
$272 million project, more than 3,000 jobs

Tweet to @ClaimBusterTM

ClaimBuster

Press Acknowledgement

2016 Republican Party Presidential Debate. Sept. 16, 2015, 7 p.m.

Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.

Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker

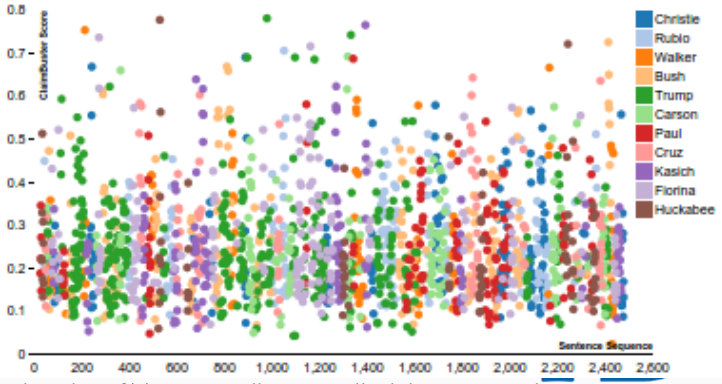
Transcript Source: http://time.com/4037239/second-republican-debate-transcript-cnn/

Chronological Order | Order by Score

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

0.45 The eleven leading Republican candidates for president are at their podiums.
0.17 They are ready to face off, and if you've been watching this race, you know anything could happen over the next few hours.
0.41 To viewers who are just joining us, welcome to the Air Force One Pavilion of the Ronald Reagan Presidential Library.
0.29 Our thanks to the staff here and especially to former first lady Nancy Reagan for this impressive setting with Ronald Reagan's presidential plane as our backdrop.
0.29 This debate is airing on CNN networks in the United States and around the world.
0.30 It's also being broadcast on the Salem Radio Network.
0.17 I know everyone is very eager to get started.
0.23 But first, I want to explain the ground rules tonight.
0.19 My name is Jake Tapper.
0.19 I'll be the moderator.
0.24 I will be joined in the questioning by Salem Radio Network talk show host Hugh Hewitt.
0.44 He worked in the Reagan administration for six years.
0.22 And by CNN's chief political correspondent Dana Bash.
0.15 I will ask follow-up questions, I will attempt to guide the discussion.

0.24 Do they bear responsibility for this refugee crisis, and what would you have done when Bashar Assad crossed the line?



2016 Republican Party Presidential Debate. Sept. 16, 2015, 7 p.m.

Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.

Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker

Transcript Source: http://time.com/4037239/second-republican-debate-transcript-cnn/

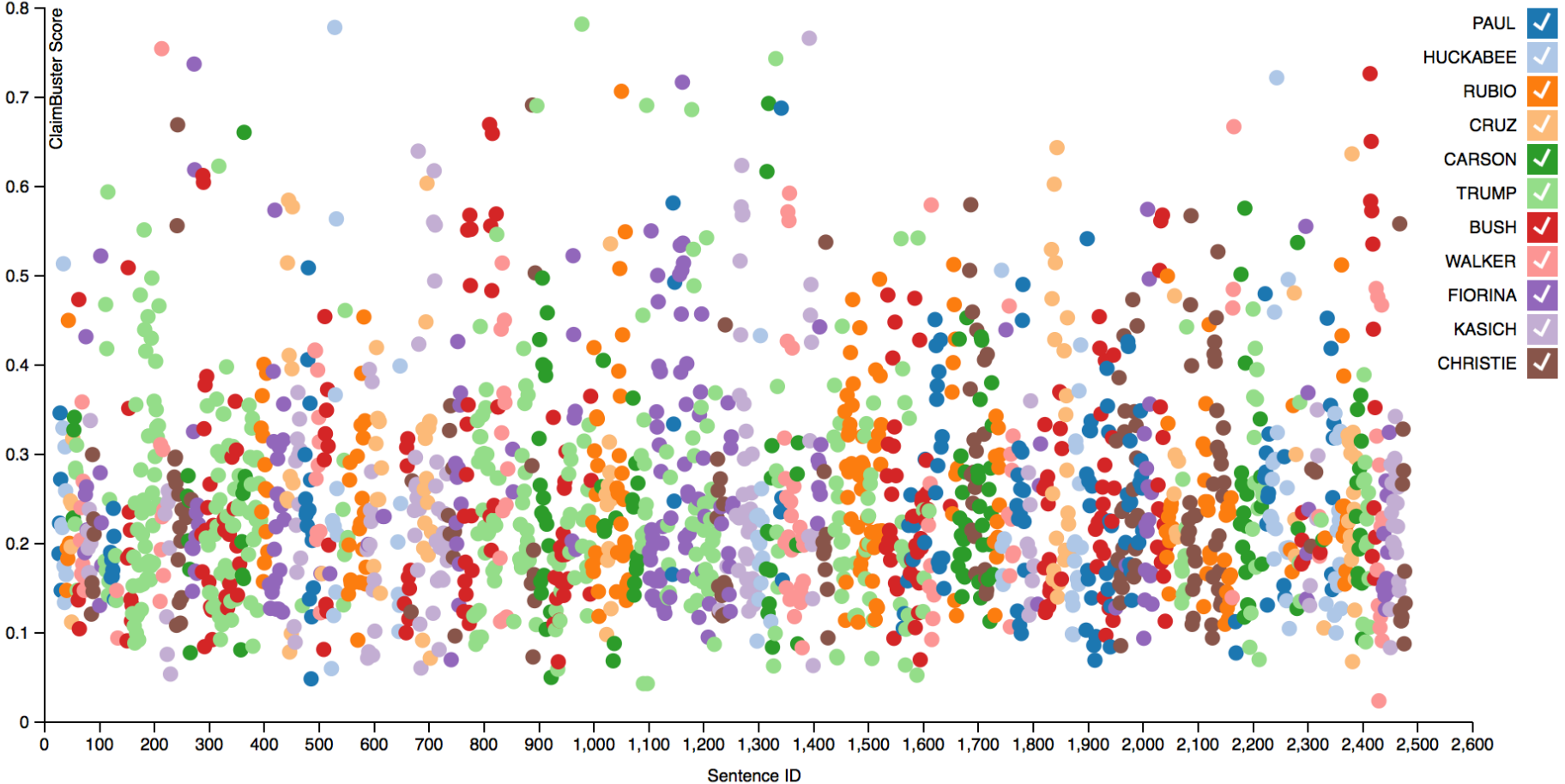
Most Check-worthy <=1.0 <=0.9 <=0.8 <=0.7 <=0.6 <=0.5 <=0.4 <=0.3 <=0.2 <=0.1 Least Check-worthy

Chronological Order | Order by Score

0.24 That's a fact.
0.33 And when the people of Iowa found that out, I went to No.
0.46 1 and you went down the tubes.
0.29 Governor Walker?
0.13 Jake, yeah, absolutely, I'll take this on, because this is an issue that's important in this race.
0.31 Just because he says it doesn't make it true.
0.19 The facts are the facts.
0.75 We balanced a $3.6 billion budget deficit, we did it by cutting taxes - $4.7 billion to help working families, family farmers, small business owners and senior citizens.
0.23 And it's about time people in America stand up and take note of this.
0.30 If you want someone that can actually take on the special interest of Washington, which you yourself said you were part of, using the system, we need somebody that will stand up and fight for average Americans to put them back in charge of their government.
0.16 I'm the one who is taking that on.
0.23 I'll do that as your next president.
0.17 Let's move on.
0.23 Jake, Jake.
0.13 A phenomenon going on in the race right now is the political...
0.25 OK, Governor Kasich, go ahead.

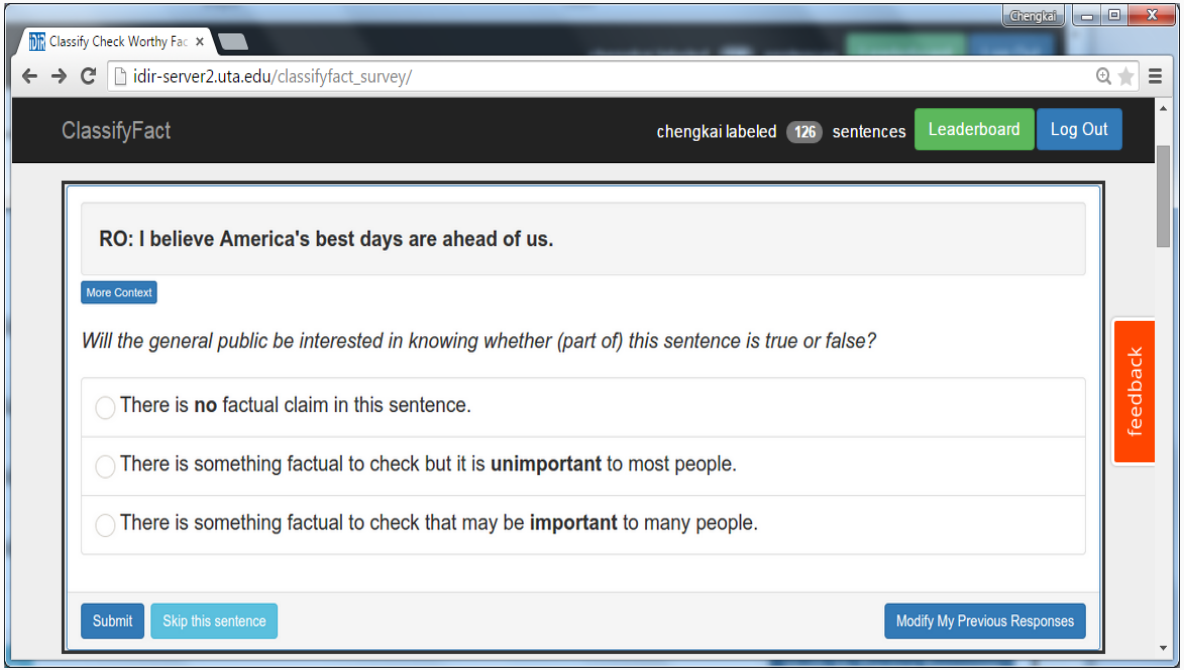
0.53	In fact, today, on the front page of the Wall Street Journal, they fired another 25 or 30,000 people saying we still haven't recovered from the catastrophe.

Debate Timeline Graph



You are Invited

http://bit.ly/claimbusters



FactWatcher

Automated Monitoring of Facts from Real-World Events

FactWatcher (Fact Finding)

Tuple t for new real world event 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

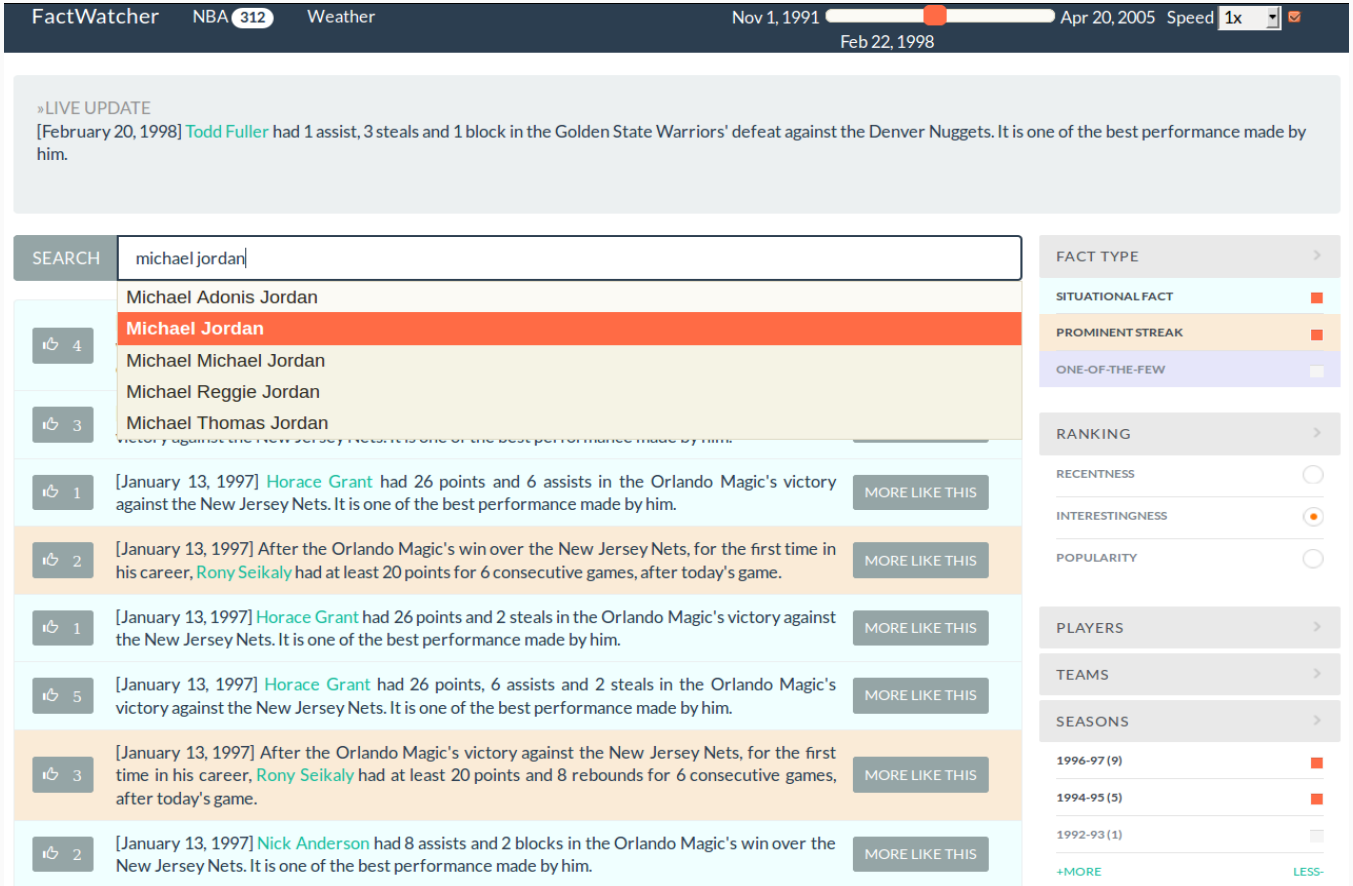
Constraint	Measure
month=*Feb*	pts, ast, reb
opp_team=*Nets*	ast, reb
team=*Celtics* & opp_team=*Nets*	ast, reb
…	…

Find constraint-measure pair (C, M) such that t is in the contextual skyline

Generate factual claim

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.

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



Presented In

VLDB 2014

Excellent Demo Award

COMPUTATION + JOURNALISM SYMPOSIUM 2014

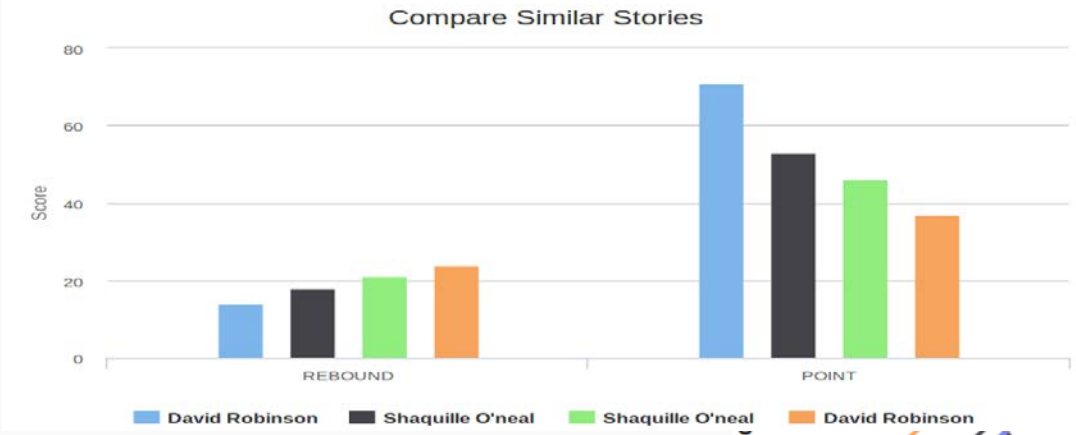
http://idir.uta.edu/factwatcher/

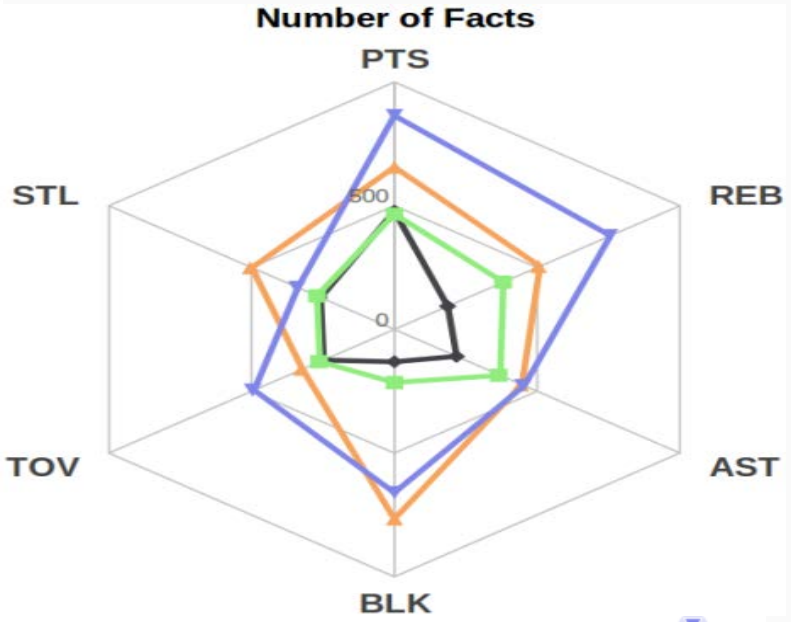
[April 24, 1994] David Robinson had 71 points and 14 rebounds in the San Antonio Spurs' victory against the Los Angeles Clippers. No one before had a better performance in NBA history.

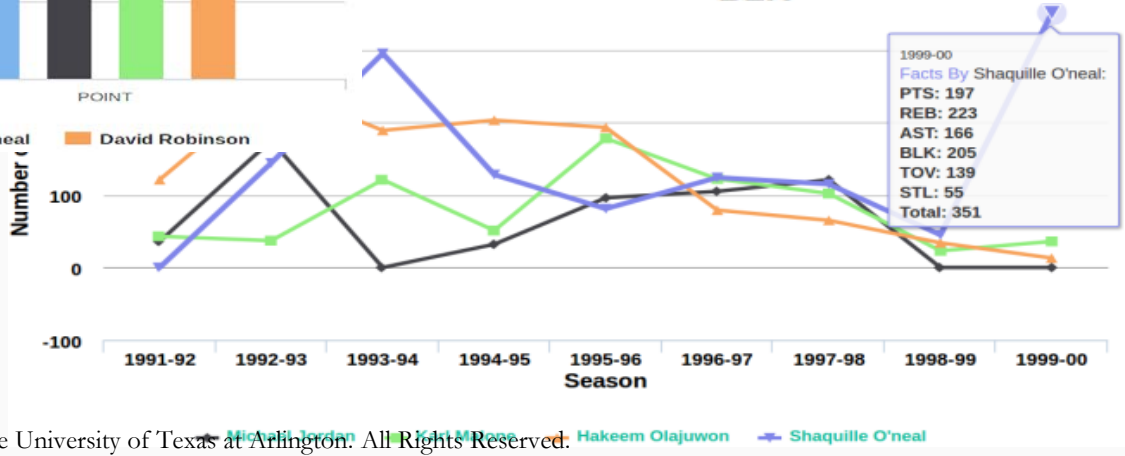
[April 20, 1994] Shaquille O'neal had 53 points and 18 rebounds in the Orlando Magic's win over the Minnesota Timberwolves. No one before had a better performance in NBA history.

[February 16, 1993] Shaquille O'neal had 46 points and 21 rebounds in the Orlando Magic's defeat against the Detroit Pistons. No one before had a better performance in NBA history.

[February 27, 1992] David Robinson had 37 points and 24 rebounds in the San Antonio Spurs' victory against the Golden State Warriors. No one before had a better performance in NBA history.







FactWatcher Finds Three Types of Facts (and can be Extended)

Domains

- sports, weather, crimes, transportation, finance, social media analytics

Examples from Real News Media

Prominent streaks

- "This month the Chinese capital has experienced 10 days with a maximum temperature in around 35 degrees Celsius – the most for the month of July in a decade."
  http://www.chinadaily.com.cn/china/2010-07/27/content_11055675.htm
- "The Nikkei 225 closed below 10000 for the 12th consecutive week, the longest such streak since June 2009."
  http://www.bloomberg.com/news/articles/2010-08-06/japanese-stocks-fall-for-second-day-this-week-on-u-s-jobless-claims-yen

FactWatcher Finds Three Types of Facts (and can be Extended)

Examples from Real News Media

Situational facts, One-of-the-few objects

- "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

- "The social world's most viral photo ever generated 3.5 million likes, 170,000 comments and 460,000 shares by Wednesday afternoon."

  http://www.cnbc.com/id/49728455

Demo

http://idir.uta.edu/factwatcher/

Acknowledgment

UTA Students

- Naeemul Hassan
- Afroza Sultana
- Gensheng Zhang
- Joseph Minumol
- Fatma Dogan

Collaborators

- Bill Adair (Duke)
- Pankaj Agarwal (Duke)
- Sarah Cohen (Columbia)
- James Hamilton (Stanford)
- Ping Luo (Chinese Academy of Sciences)
- Mark Tremayne (UTA)
- Min Wang (Google Research)
- Jun Yang (Duke)
- Cong Yu (Google Research)

Acknowledgment

Funding sponsors

Disclaimer: This material is based upon work partially supported by the National Science Foundation Grants 1018865, 1117369 and 1408928, 2011 and 2012 HP Labs Innovation Research Awards, and the National Natural Science Foundation of China Grant 61370019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Thank You! Questions?

http://ranger.uta.edu/~cli

http://idir.uta.edu

cli@uta.edu

Please help us label the data

http://bit.ly/claimbusters

Other IDIR Projects

CrowdConsensus

Team/organization decision making and event organizer

Humans' Obsession: Comparing Things

Who is the better footballer?

http://visual.ly/messi-vs-ronaldo

Which is the better company to work for?

Facebook-Google Showdown

Facebook vs Google best

www.businessinsider.com/facebook-vs-google-best-employer-2013-11

BUSINESS INSIDER Tech Finance Politics Strategy Life

These Charts Prove Facebook Is A Better Place To Work Than Google

JULIE BORT
NOV. 26, 2013, 9:22 PM 18,234 6

FACEBOOK LINKEDIN TWITTER GOOGLE+

When it comes to desirable places to work in the tech industry, two companies are always at the top of the list: Facebook and Google.

But which one is *really* the better employer?

To answer that question, we compared the two

Mark Zuckerberg and Larry Page.

Facebook vs Google best

www.businessinsider.com/facebook-vs-google-best-employer-2013-11

BUSINESS INSIDER Tech Finance Politics Strategy Life

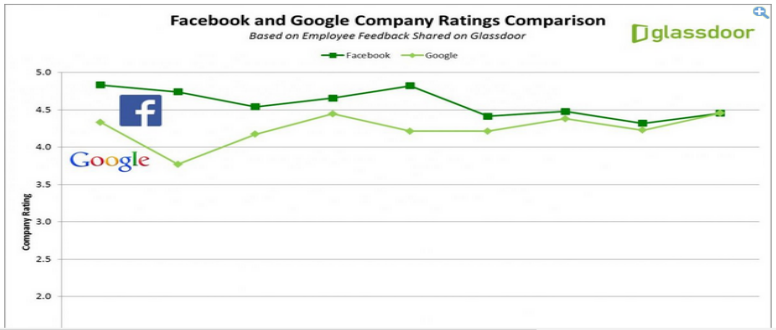
Round 1: Overall satisfaction ratings — Facebook wins

(Score so far: Facebook - 1, Google - 0)

Facebook employees rate their employer slightly better overall (4.6) compared to how Google employees rate Google (4.1), from at least 550 company reviews per company.

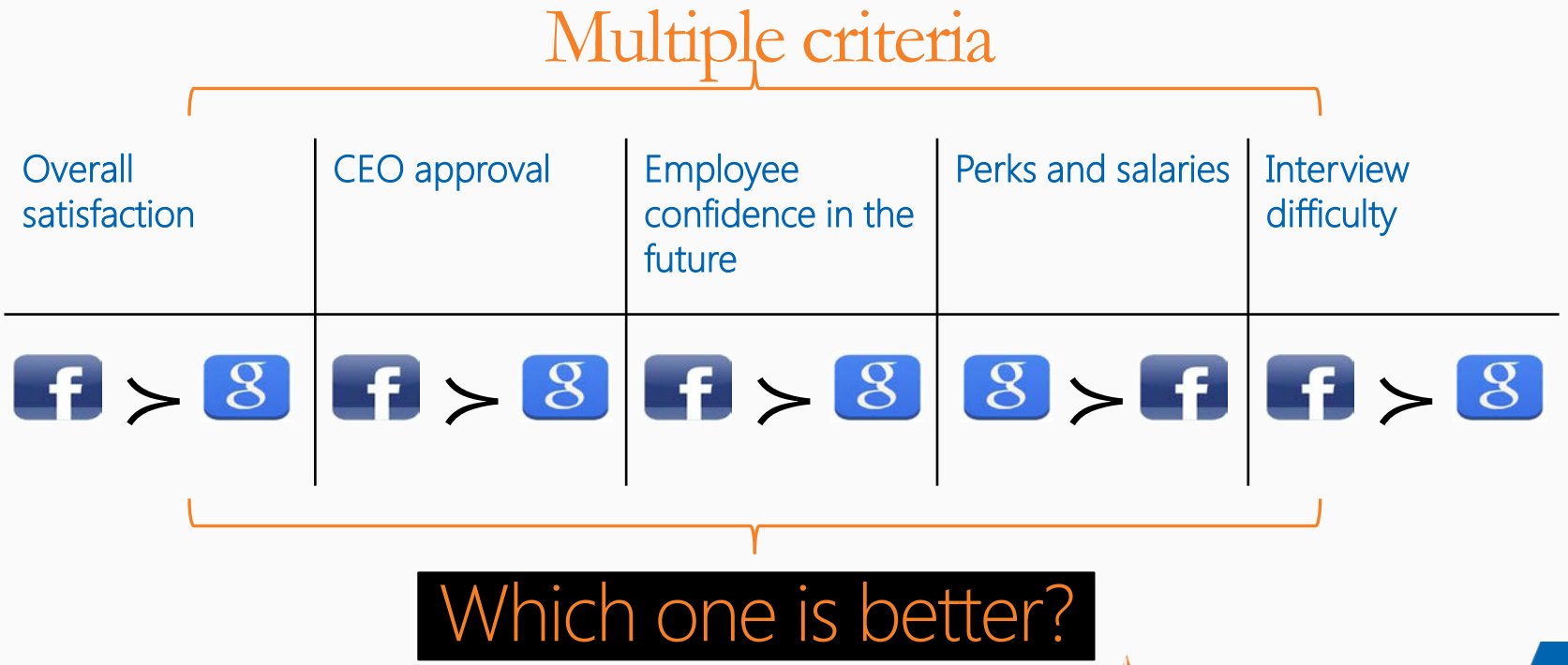
Satisfaction ratings are based on a 5-point scale: 1.0=very dissatisfied, 3.0=OK, 5.0=very satisfied. So, we can see that employees are highly satisfied with both companies.

Interestingly, Google's rating has been climbing, indicating that current employers are getting happier. As of this quarter, the two companies are tied at 4.5 each, as the chart below shows:



http://www.businessinsider.com/facebook-vs-google-best-employer-2013-11

Facebook-Google Showdown



CrowdConsensus

Optimization techniques and algorithms that minimize questions to users for reaching consensus.

Applications

Collecting Public Opinion

Group Decision Making

- Where for lunch, which product to use, which candidate to hire

Information Exploration

- Compare photos by color, sharpness, and landscape

How to Get Preference Relations? Ask the Crowd (mobile apps)

Between A separation(2011) and The big Lebowski(1998) which movie is better with regard to *story*?

- A separation(2011).
- The big Lebowski(1998).
- no preference.

submit | skip this question

Between each pair of cities, which has the better weather overall?

Dallas(TX)	about the same	Chicago(IL)	not sure

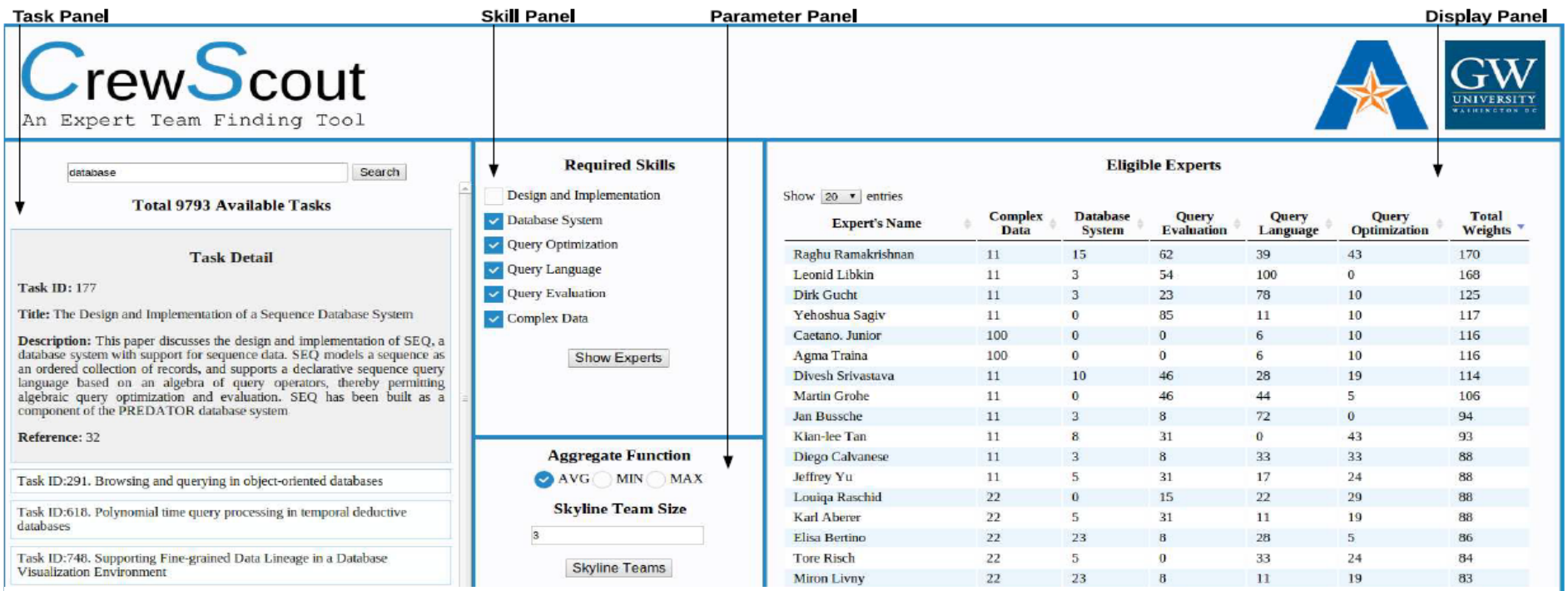
In each following pair(each row), which picture has more interesting colors?

Left Picture	Right Picture	Preference
		Left Picture Right Picture No Preference

CrewScout

Expert Team Formation

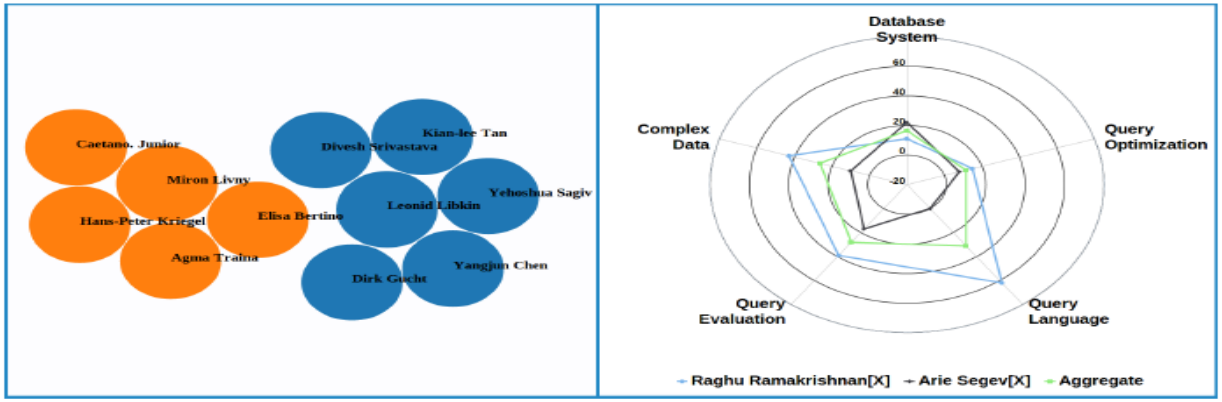
CrewScout --- Expert Team Formation



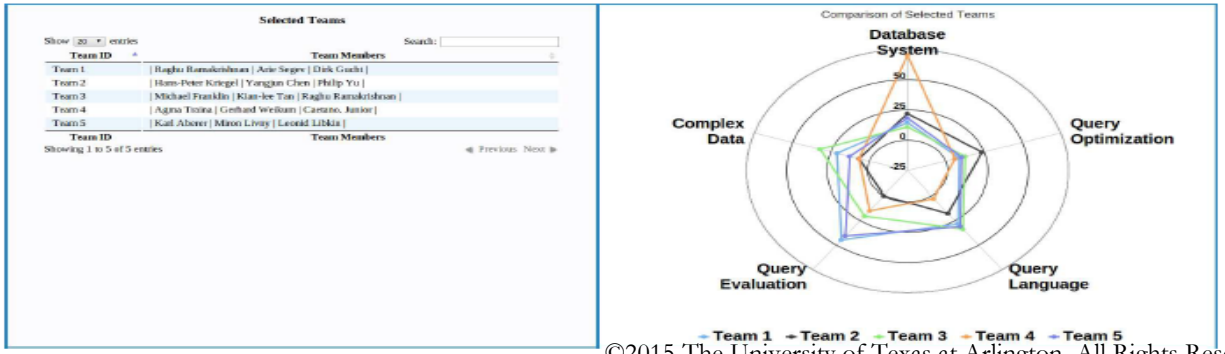
http://idir.uta.edu/crewscout/

CrewScout http://idir.uta.edu/crewscout/

CrewScout http://idir.uta.edu/crewscout/



(b) 2 experts selected



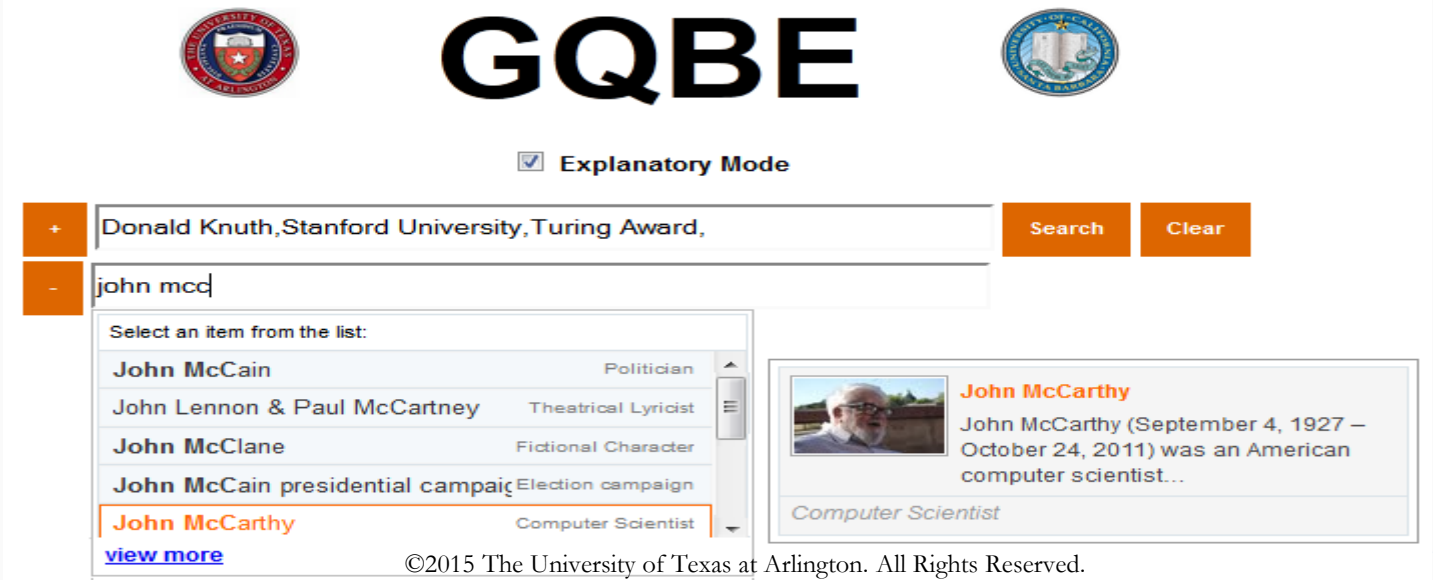
GQBE (Graph Query-by-Example)

http://idir.uta.edu/gqbe

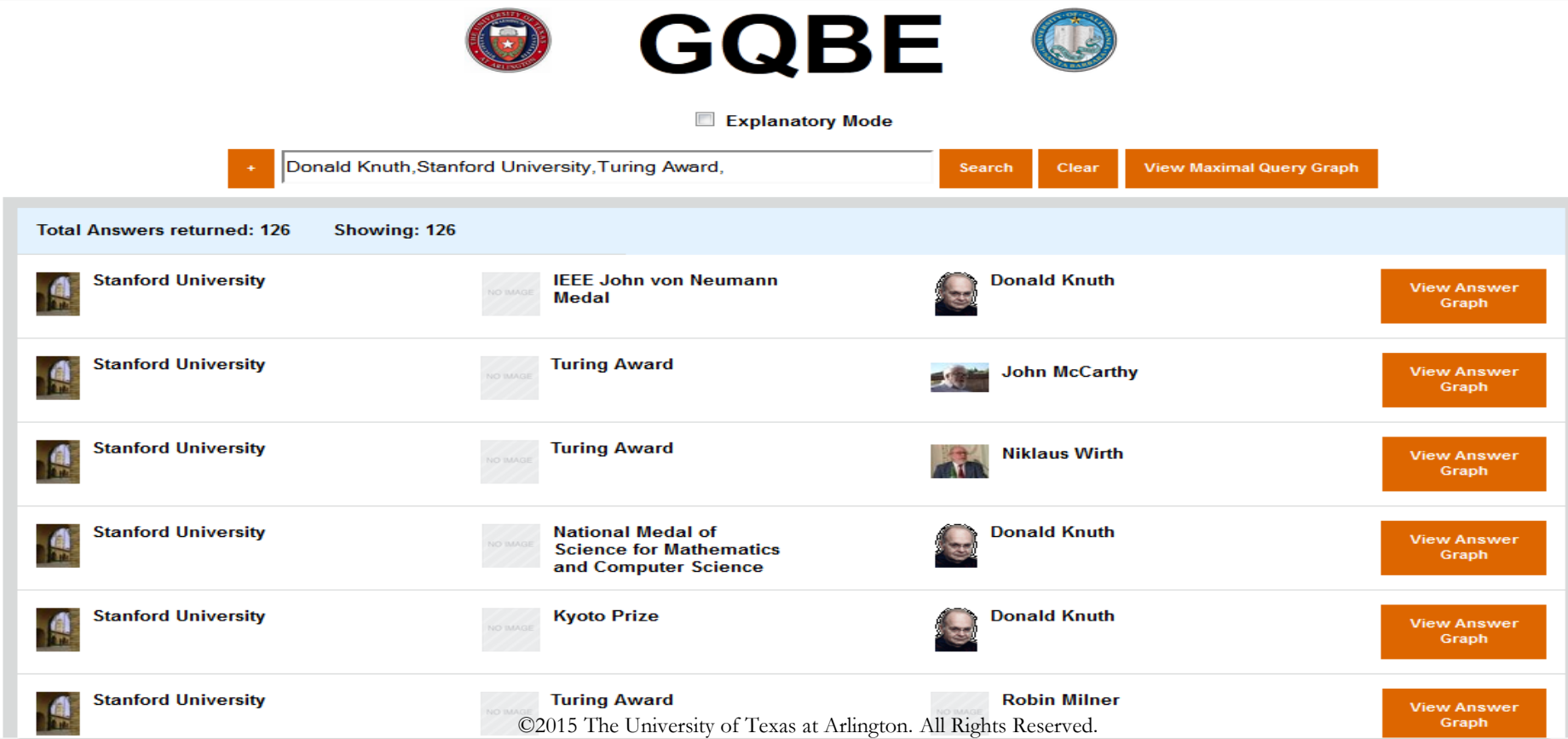
On a knowledge graph:

Input: one or more example n-entity tuples

Output: top-k most similar answer n-tuples.

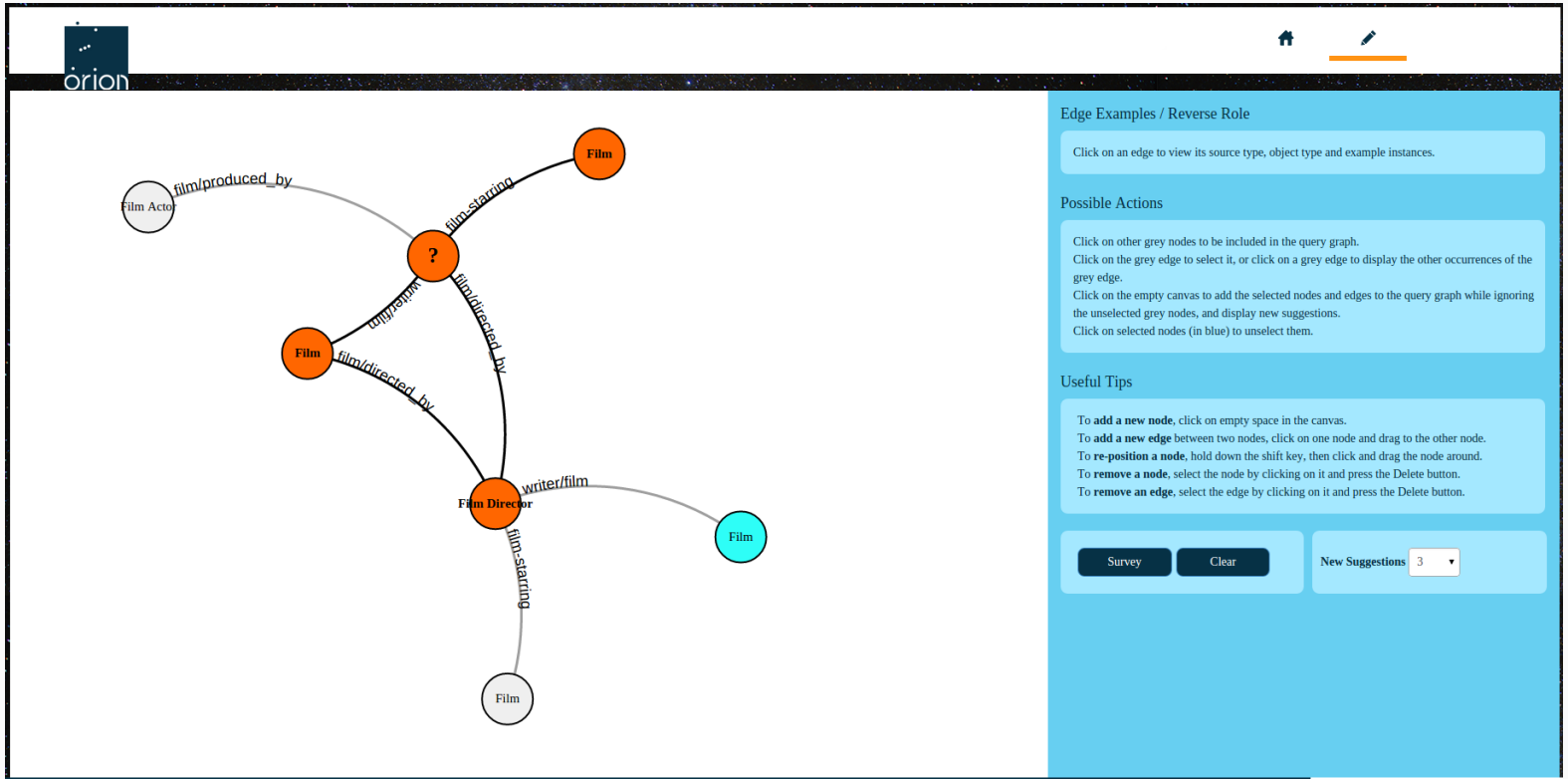


GQBE (Graph Query-by-Example)



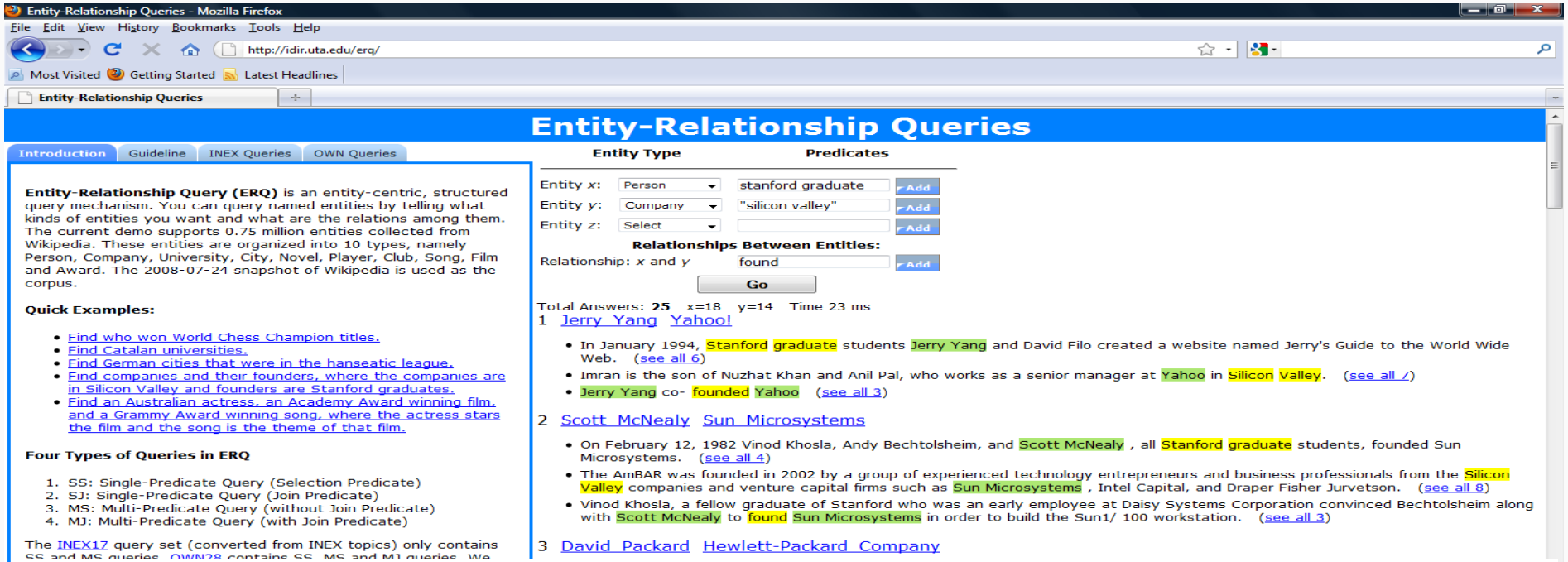
Orion: Auto-suggestion for Interactive Graph Query Formulation

http://idir.uta.edu/orion



UNIVERSITY OF TEXAS ARLINGTON

Entity-Relationship Queries (ERQ)



http://idir.uta.edu/erq

Facetedpedia http://idir.uta.edu/facetedpedia

In contrast, a faceted interface over the result list looks like this: