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

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, and Stanford. Collaboration with HP Labs China and Chinese Academy of Sciences.

- 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)

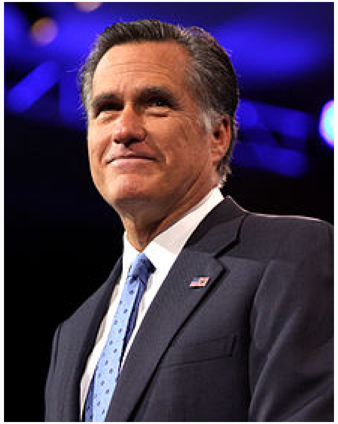
- On "One of the Few" Objects. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. KDD 2012, pages 1487-1495.
- Prominent Streak Discovery in Sequence Data. Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, Yong Yu. KDD 2011, pages 1280-1288.
- Computational Journalism: A Call to Arms to Database Researchers. 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**)

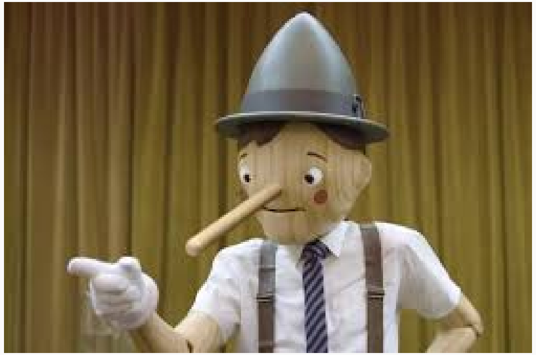
ClaimBuster

The Quest to Automate Fact-Checking

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.

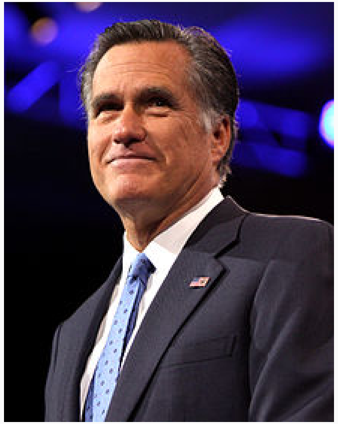


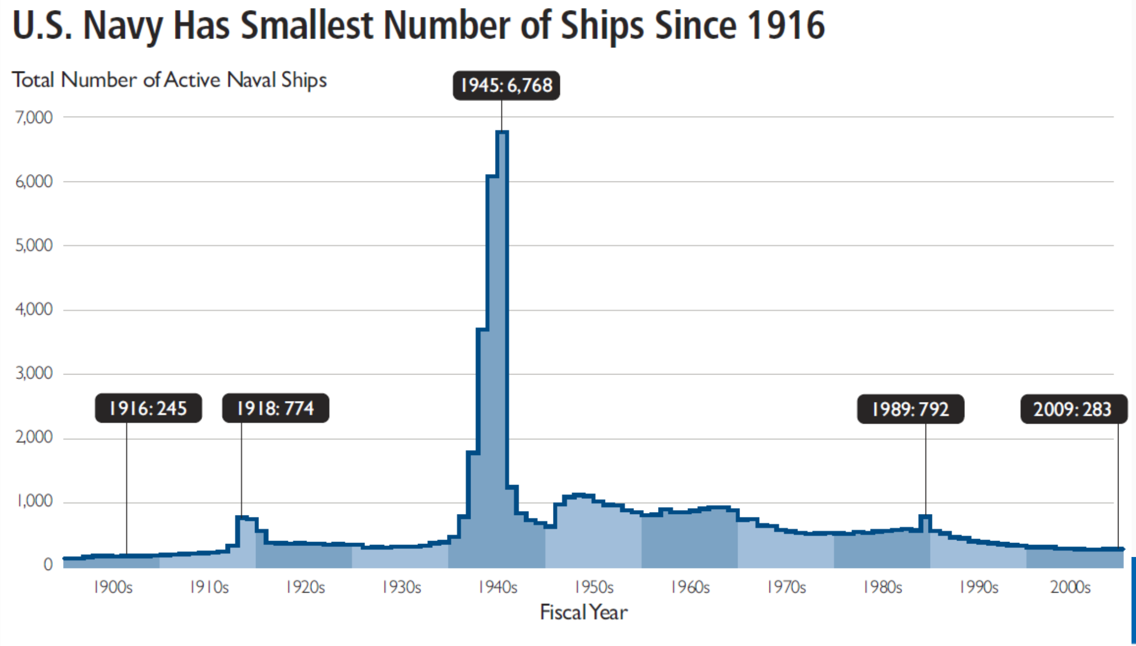


http://en.wikipedia.org/wiki/Mitt_Romney
http://www.thebrainchildgroup.com/

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.





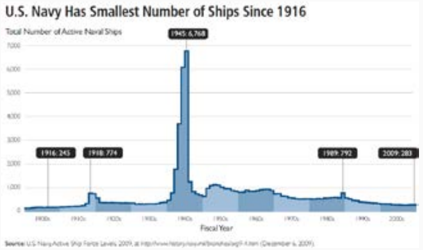
U.S. Navy Has Smallest Number of Ships Since 1916

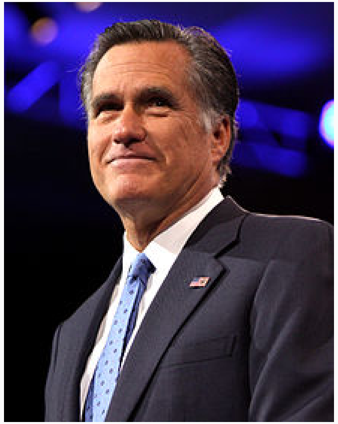
Source: U.S. Navy, Active Ship Force Levels, 2009, at *http://www.history.navy.mil/branches/org9-4.htm* (December 6, 2009).

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

Fact Checking is not Easy

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



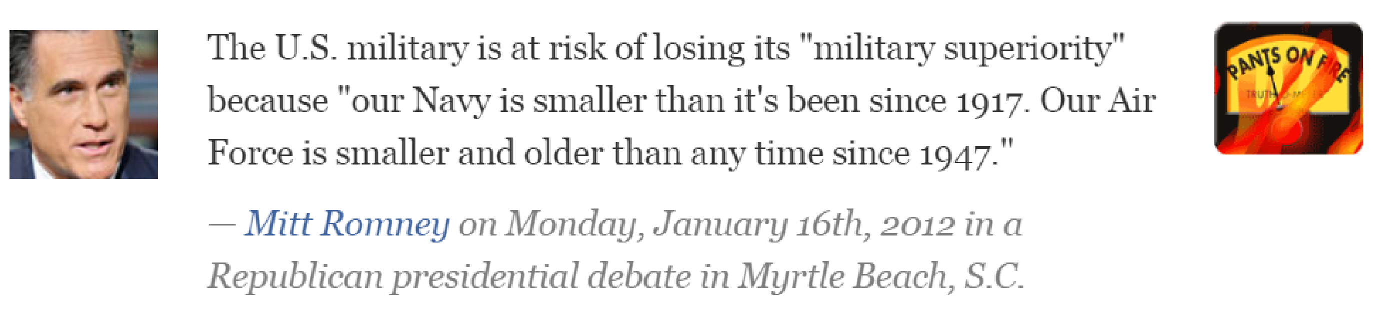


VS

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

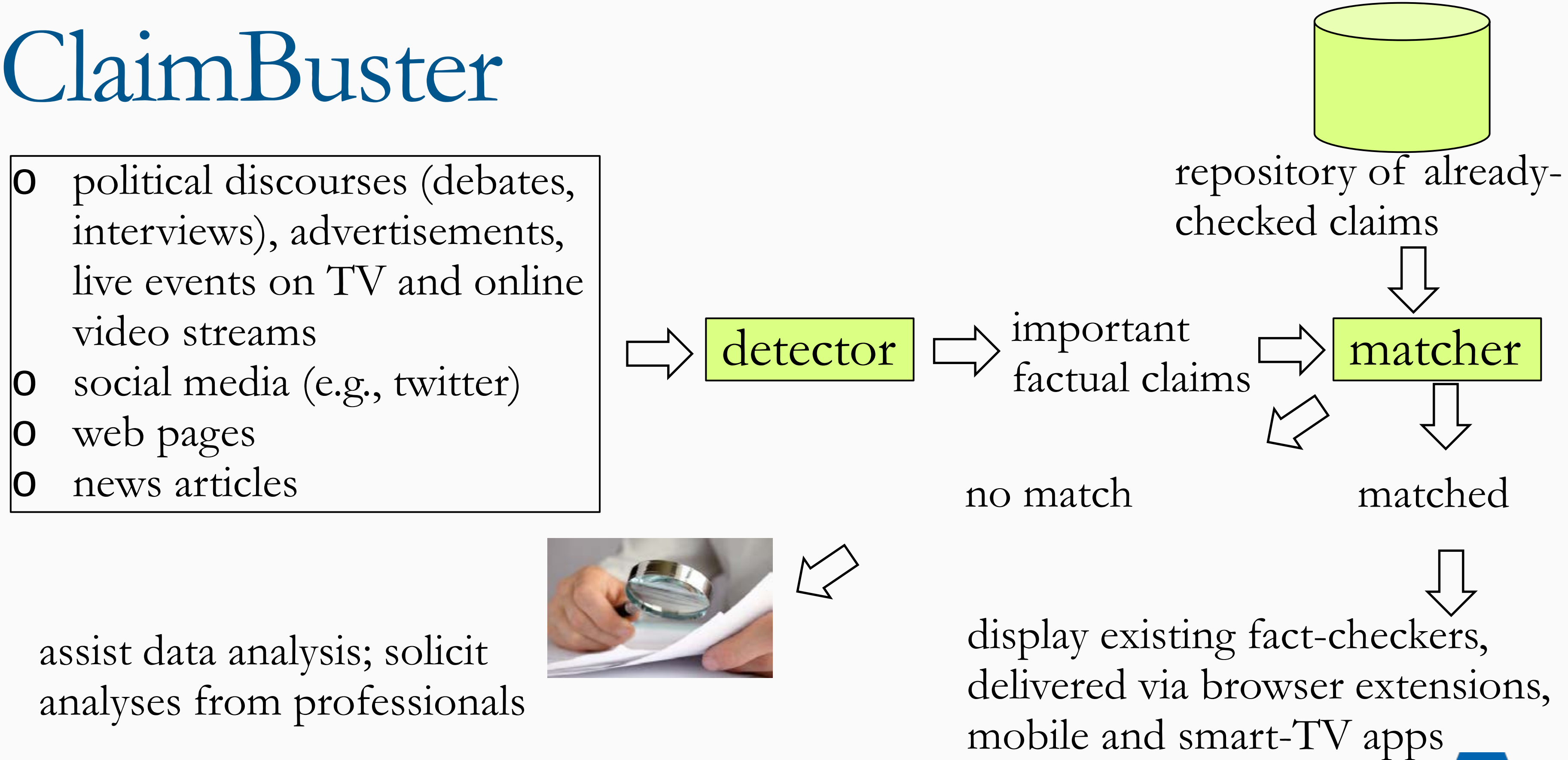
The Holy Grail

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

detector

important factual claims

repository of already-checked claims

matcher

no match

matched

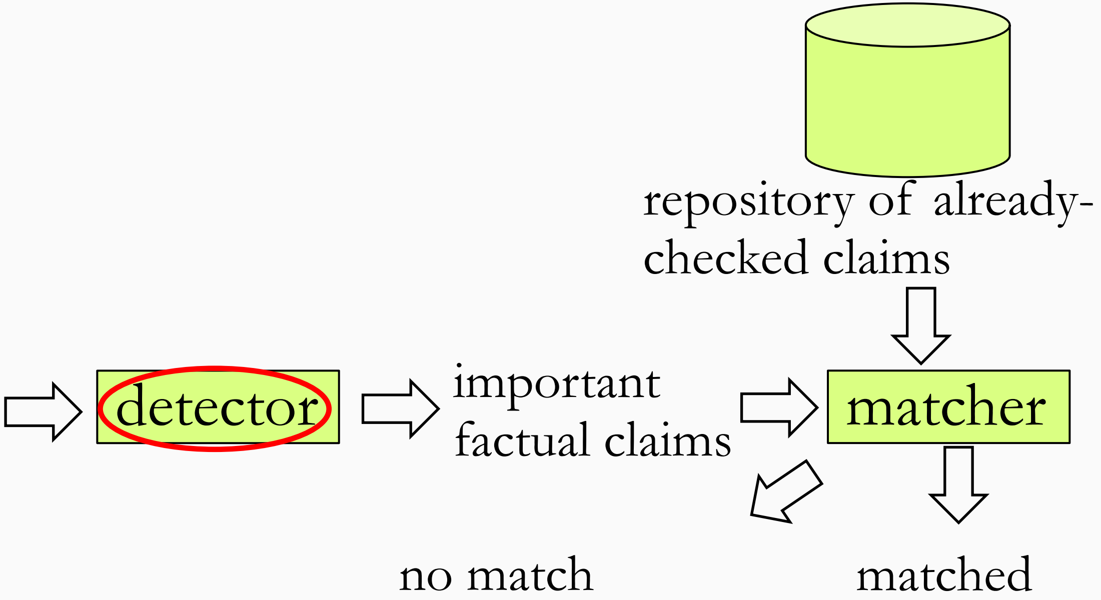
assist data analysis; solicit analyses from professionals

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



detector

important factual claims

repository of already-checked claims

matcher

no match

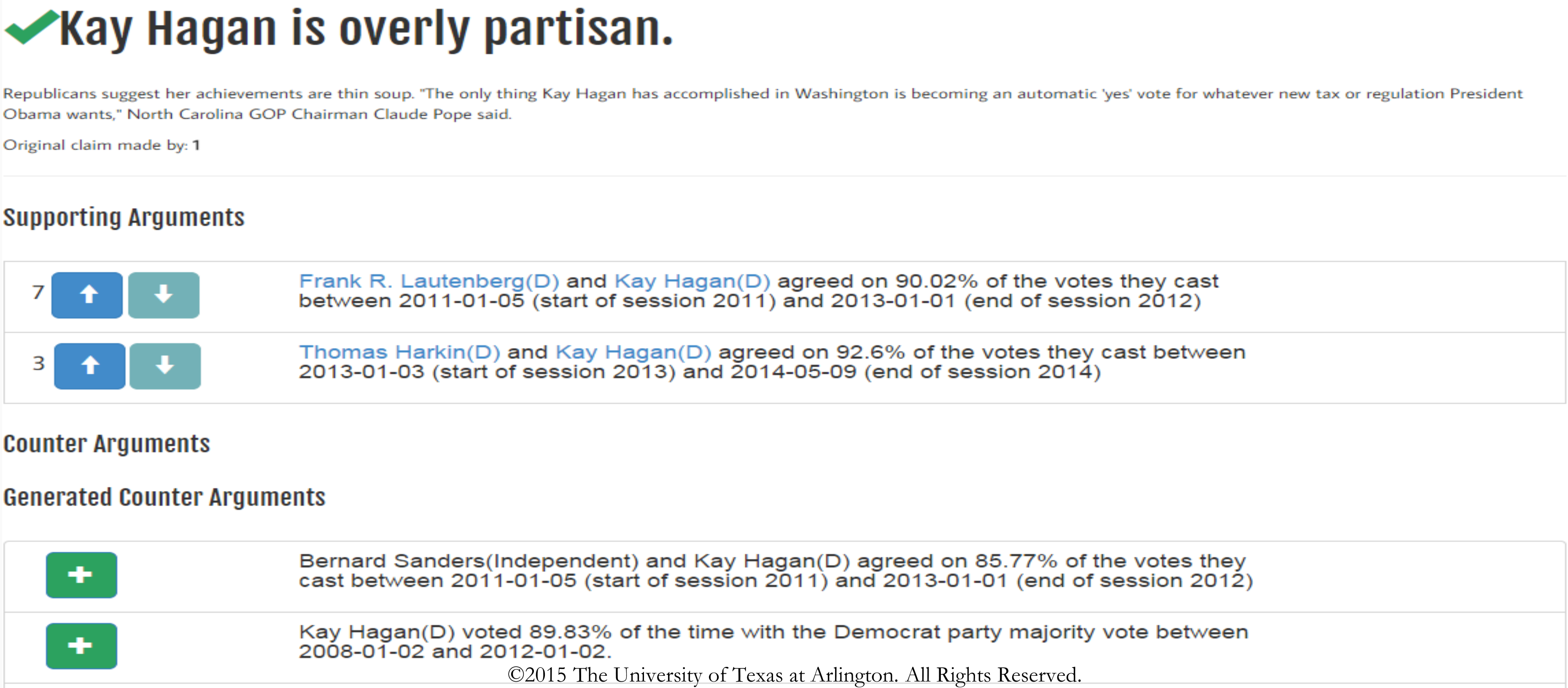
matched

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

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

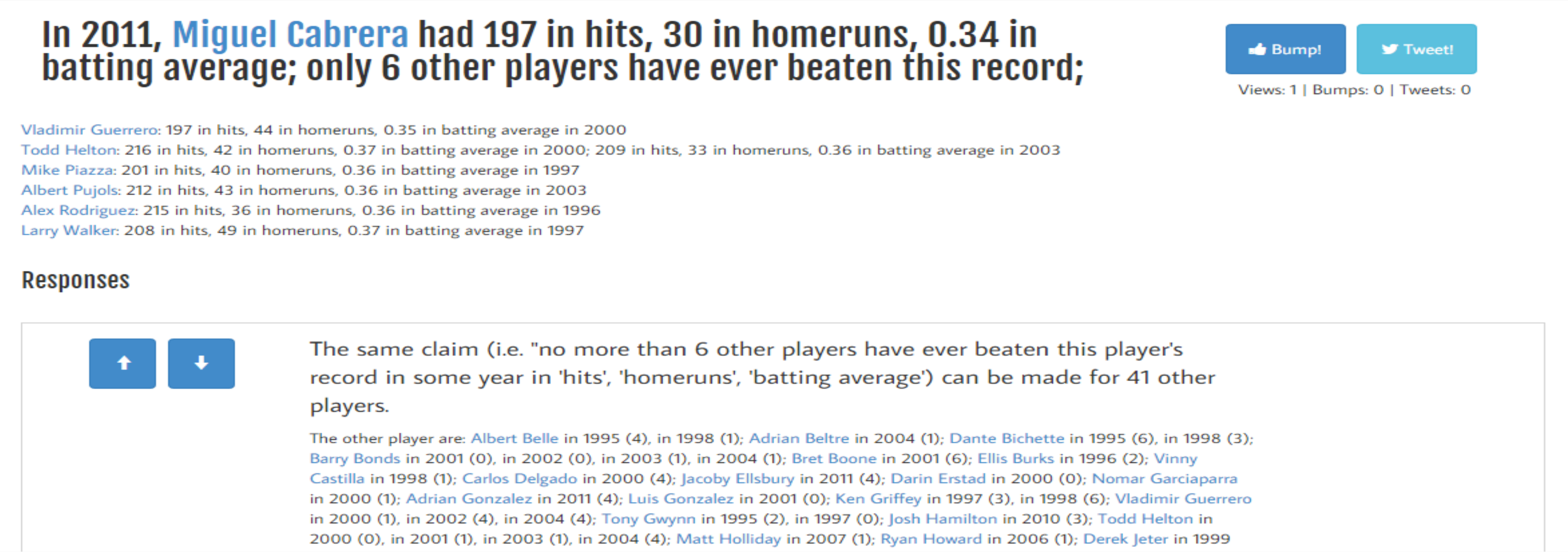
7	⬆ ⬇	Frank R. Lautenberg(D) and Kay Hagan(D) agreed on 90.02% of the votes they cast between 2011-01-05 (start of session 2011) and 2013-01-01 (end of session 2012)
3	⬆ ⬇	Thomas Harkin(D) and Kay Hagan(D) agreed on 92.6% of the votes they cast between 2013-01-03 (start of session 2013) and 2014-05-09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

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

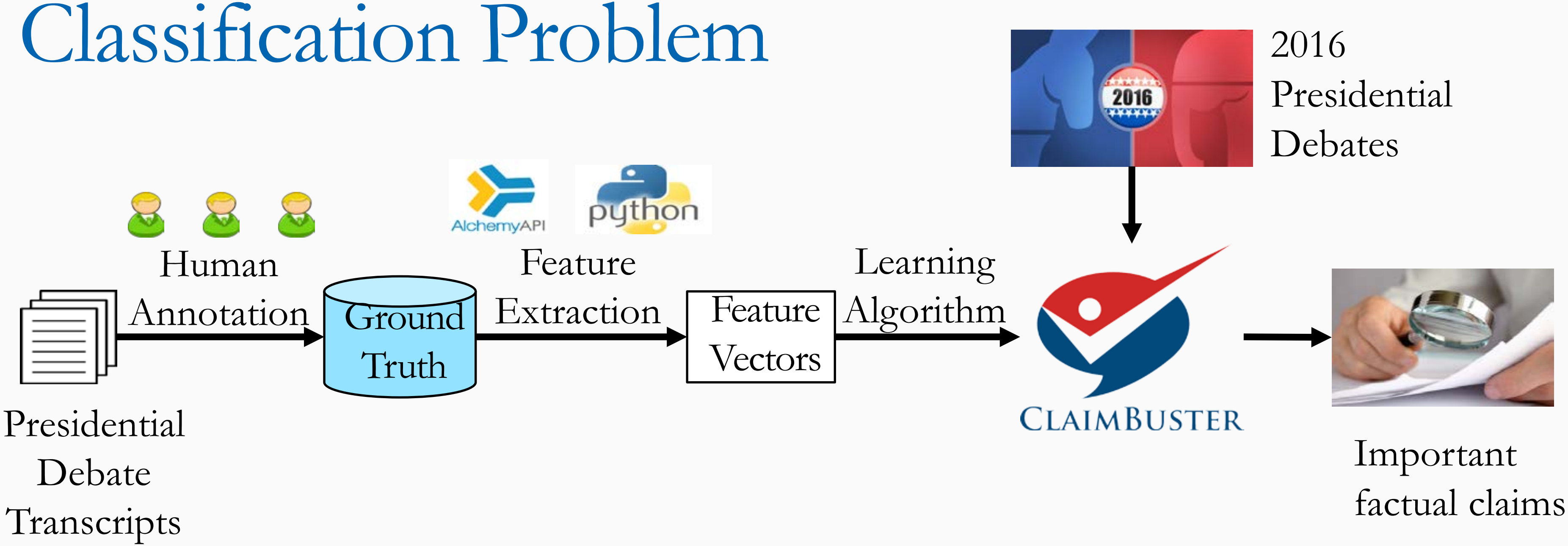
→

Journalists investigate the claims

(or checked by algorithms, citizens, crowd)

iDiR

Finding Important Factual Claims: A Classification Problem



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%.”

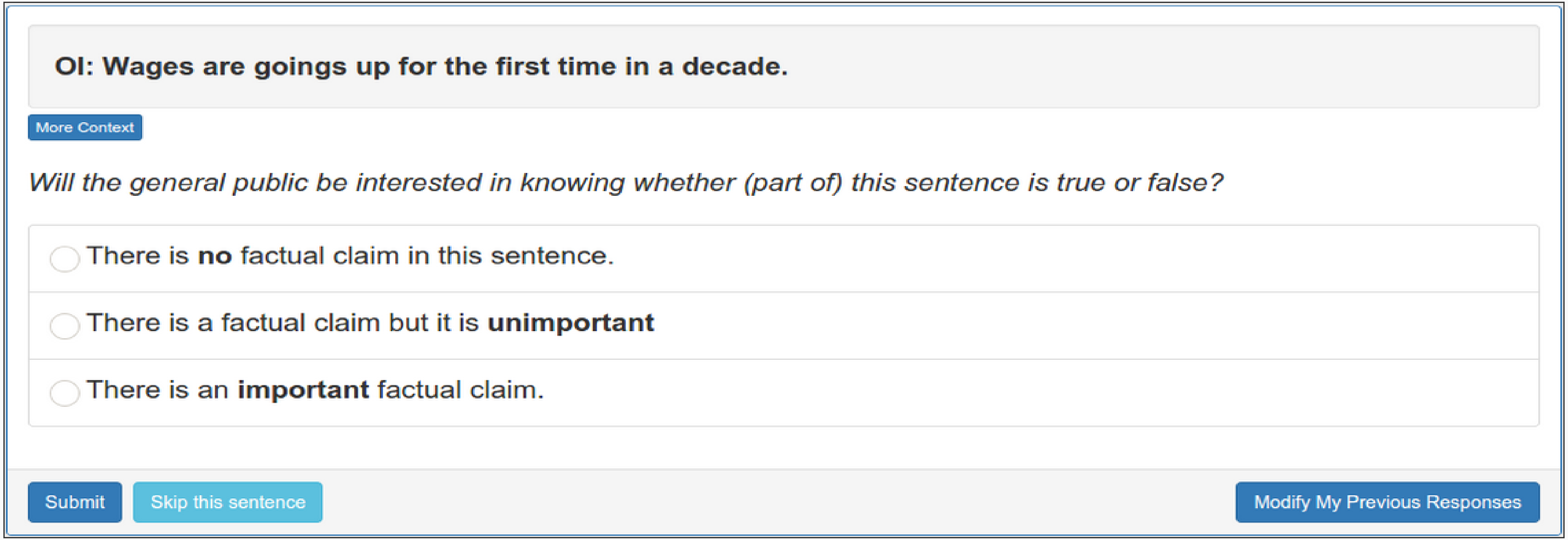
Goal: Given a future sentence, find the class it belongs to.

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



OI: Wages are goings up for the first time in a decade.

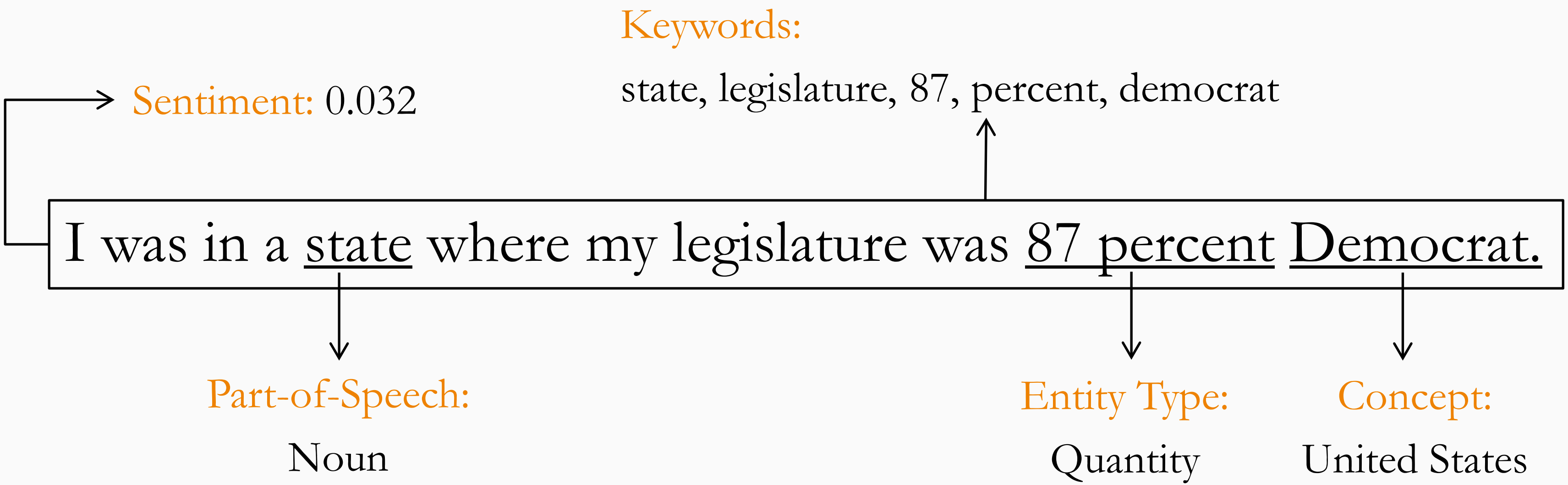
Will the general public be interested in knowing whether (part of) this sentence is true or false?

○ There is **no** factual claim in this sentence.

○ There is a factual claim but it is **unimportant**

○ There is an **important** factual claim.

Feature Extraction



Keywords:

state, legislature, 87, percent, democrat

Sentiment: 0.032

I was in a state where my legislature was 87 percent Democrat.

Part-of-Speech:

Noun

Entity Type:

Quantity

Concept:

United States

Sentiment: [-1.0 to 1.0]

Words: tf-idf scores of 6130 words (excluding rare words)

POS Tag: 43 tags

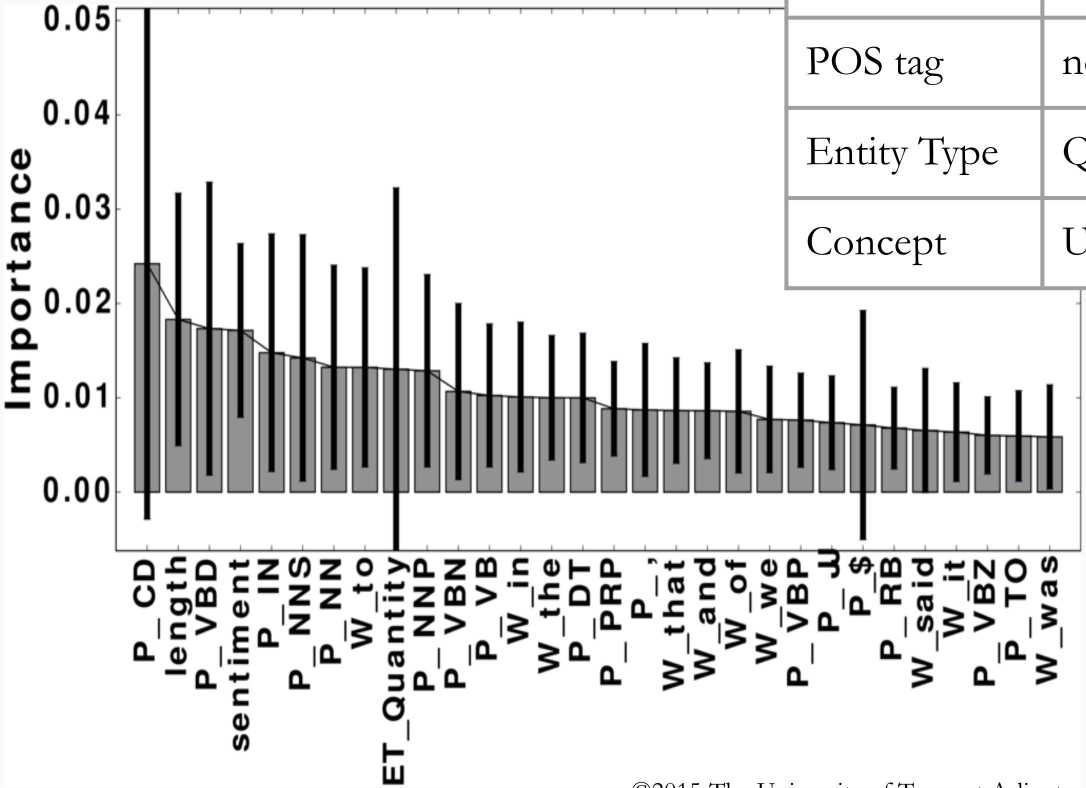
Entity Type: 26 types

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

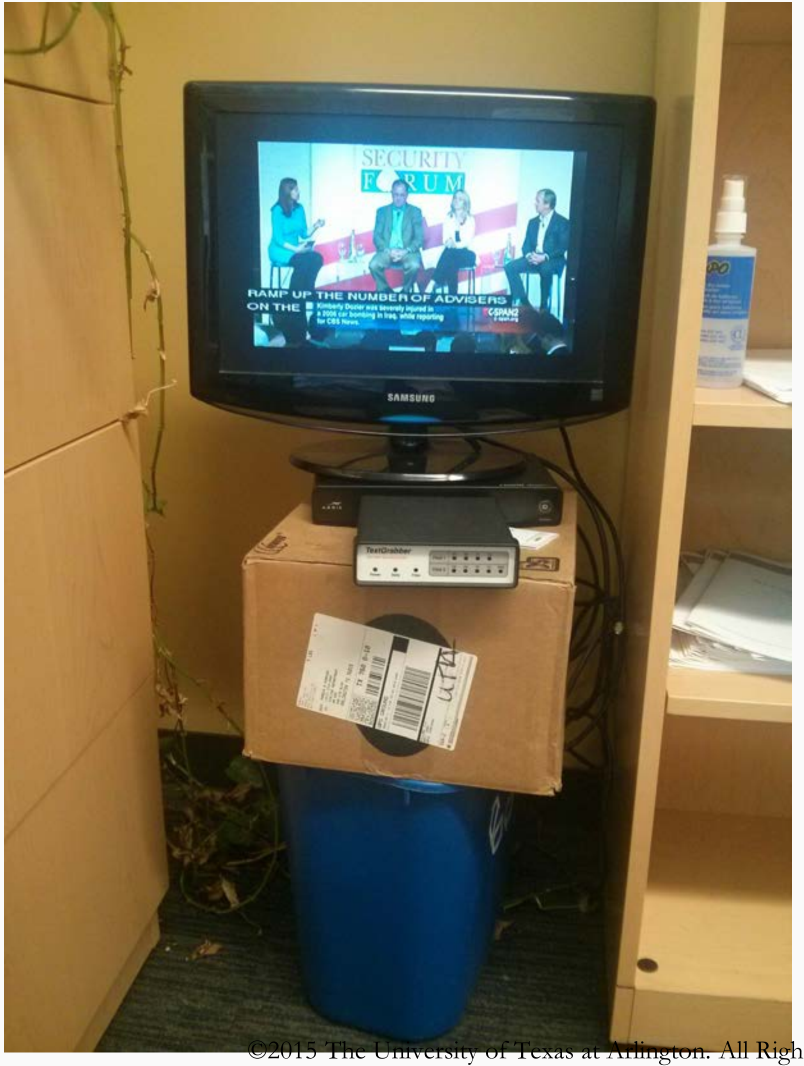
- Near real-time experiment with 2015 first Republican primary debate
- Transcript grabbed from closed captions of the Fox News channel using TextGrabber
- 1393 sentences
- 71% of the fact-checks from CNN, factcheck.org & PolitiFact were ranked by ClaimBuster within top 18%.

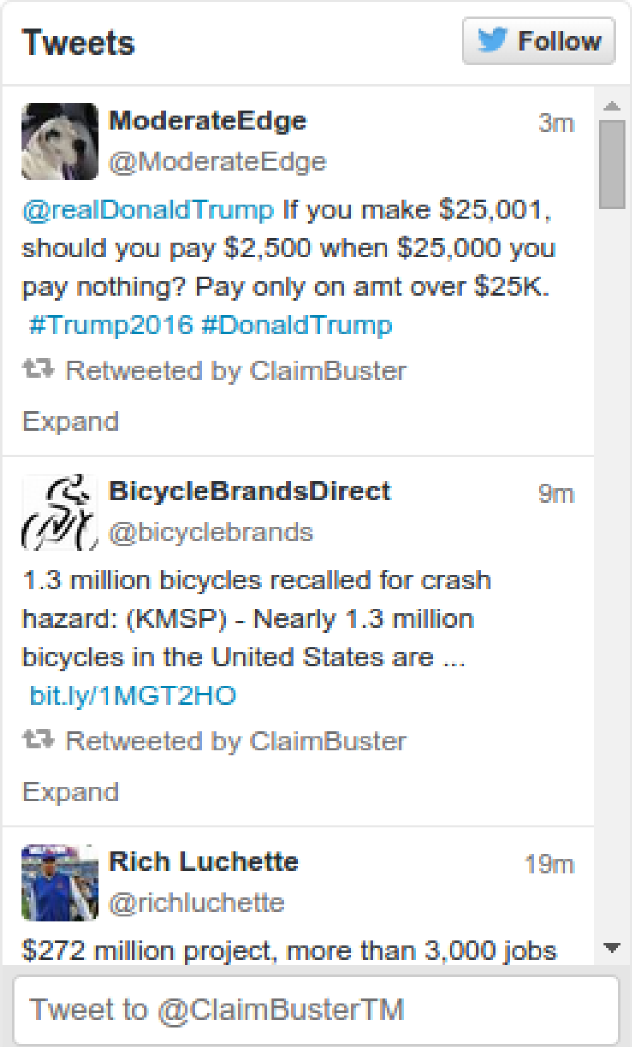
Case Study: #GOPDebate2015

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!


SECURITY FORUM
RAMP UP THE NUMBER OF ADVISERS ON THE
Kimberly Dozier was severely injured in a 2006 car bombing in Iraq, while reporting for CBS News.
C-SPAN2
SAMSUNG
TextGrabber



ModerateEdge 3m
@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

BicycleBrandsDirect 9m
@bicyclebrands

1.3 million bicycles recalled for crash hazard: (KMSP) - Nearly 1.3 million bicycles in the United States are ... bit.ly/1MGT2HO

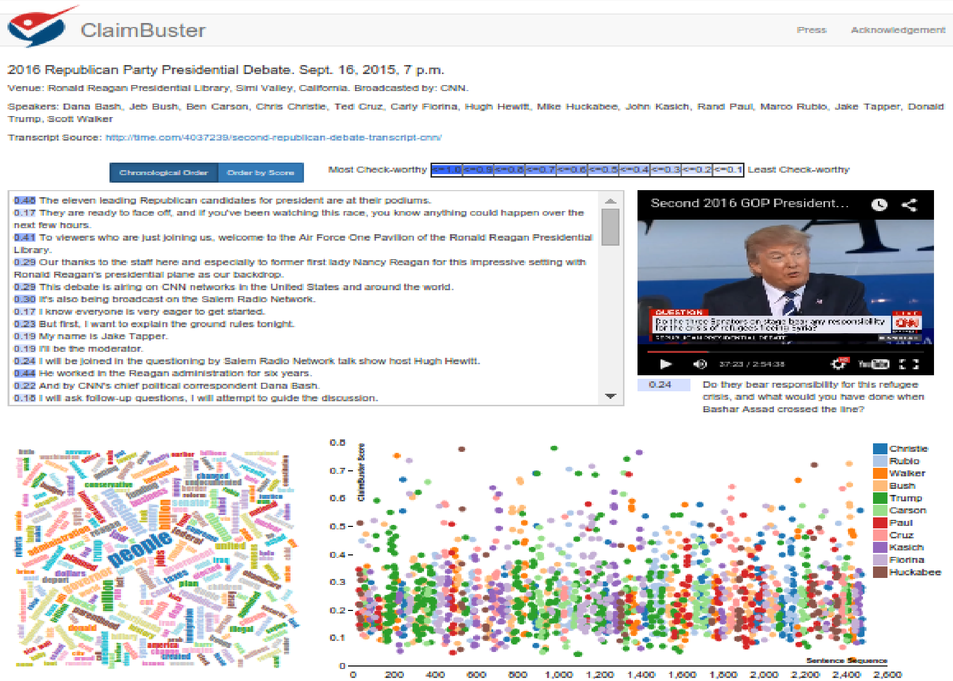
Retweeted by ClaimBuster

Rich Luchette 19m
@richluchette

$272 million project, more than 3,000 jobs

Demo

http://idir.uta.edu/claimbuster



ClaimBuster

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

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

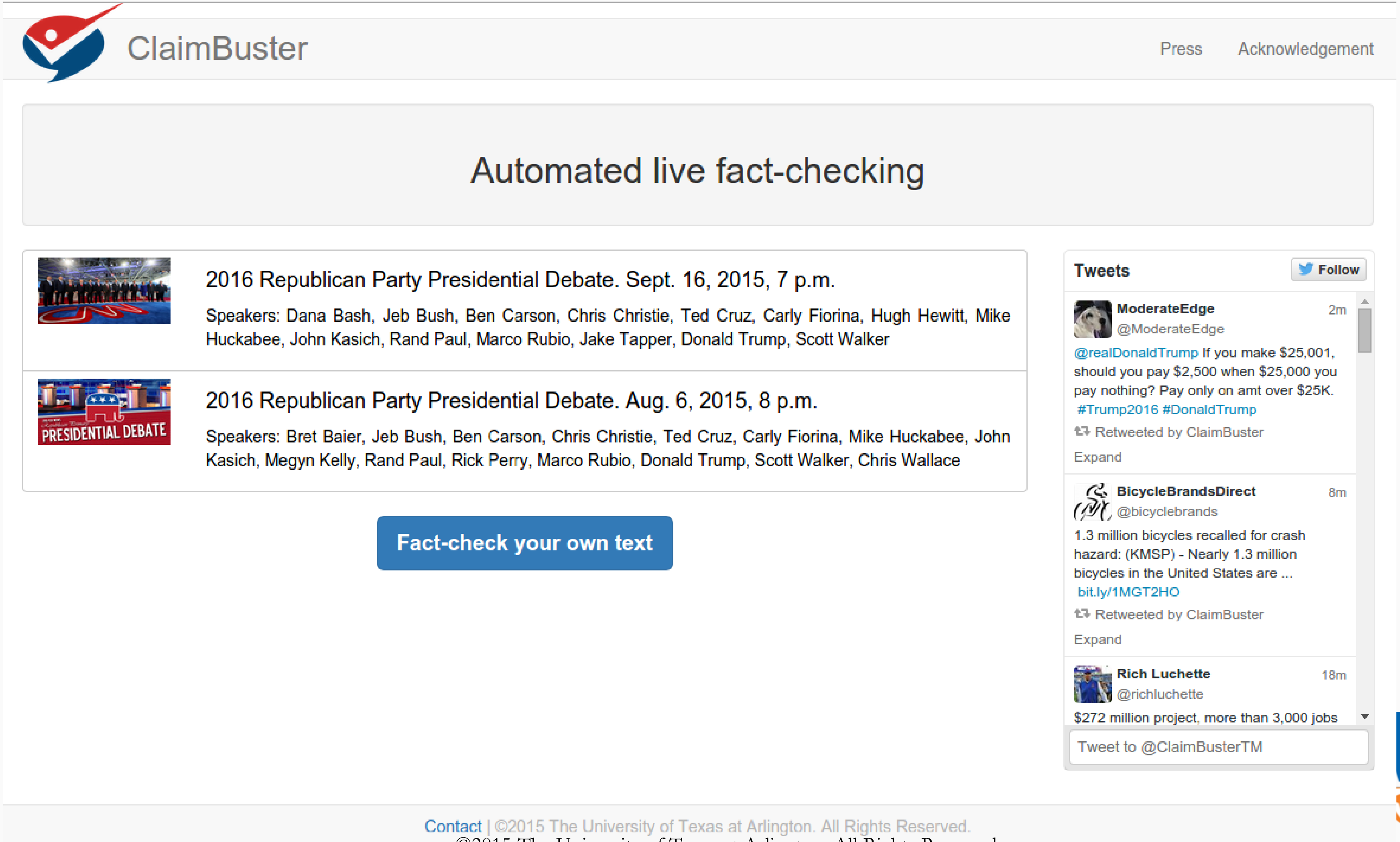
Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker

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

0.46 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.
0.16 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?



ClaimBuster

Press Acknowledgement

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

ModerateEdge @ModerateEdge 2m

@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 @bicyclebrands 8m

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

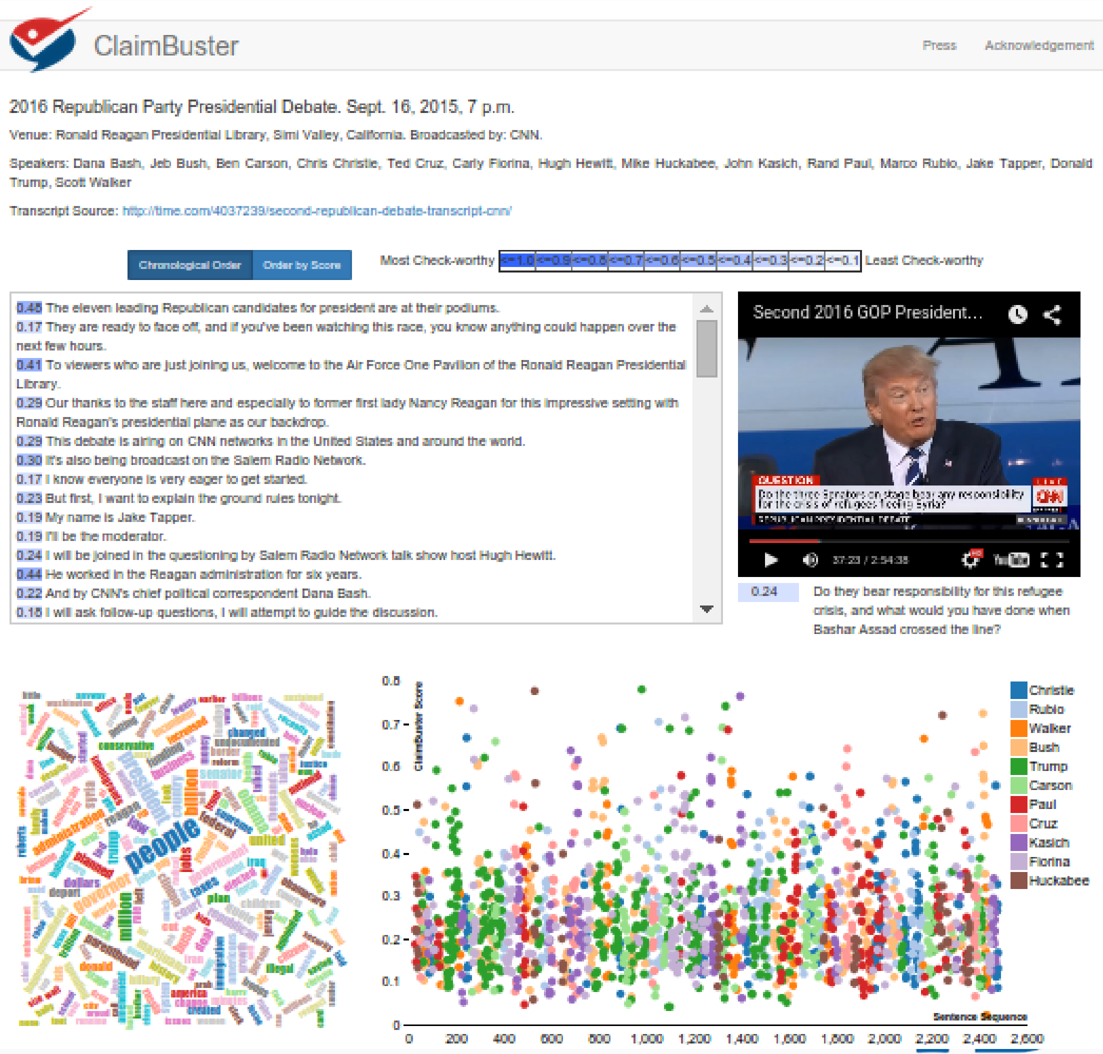
Expand

Rich Luchette @richluchette 18m

$272 million project, more than 3,000 jobs

Tweet to @ClaimBusterTM

Contact | ©2015 The University of Texas at Arlington. All Rights Reserved.



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

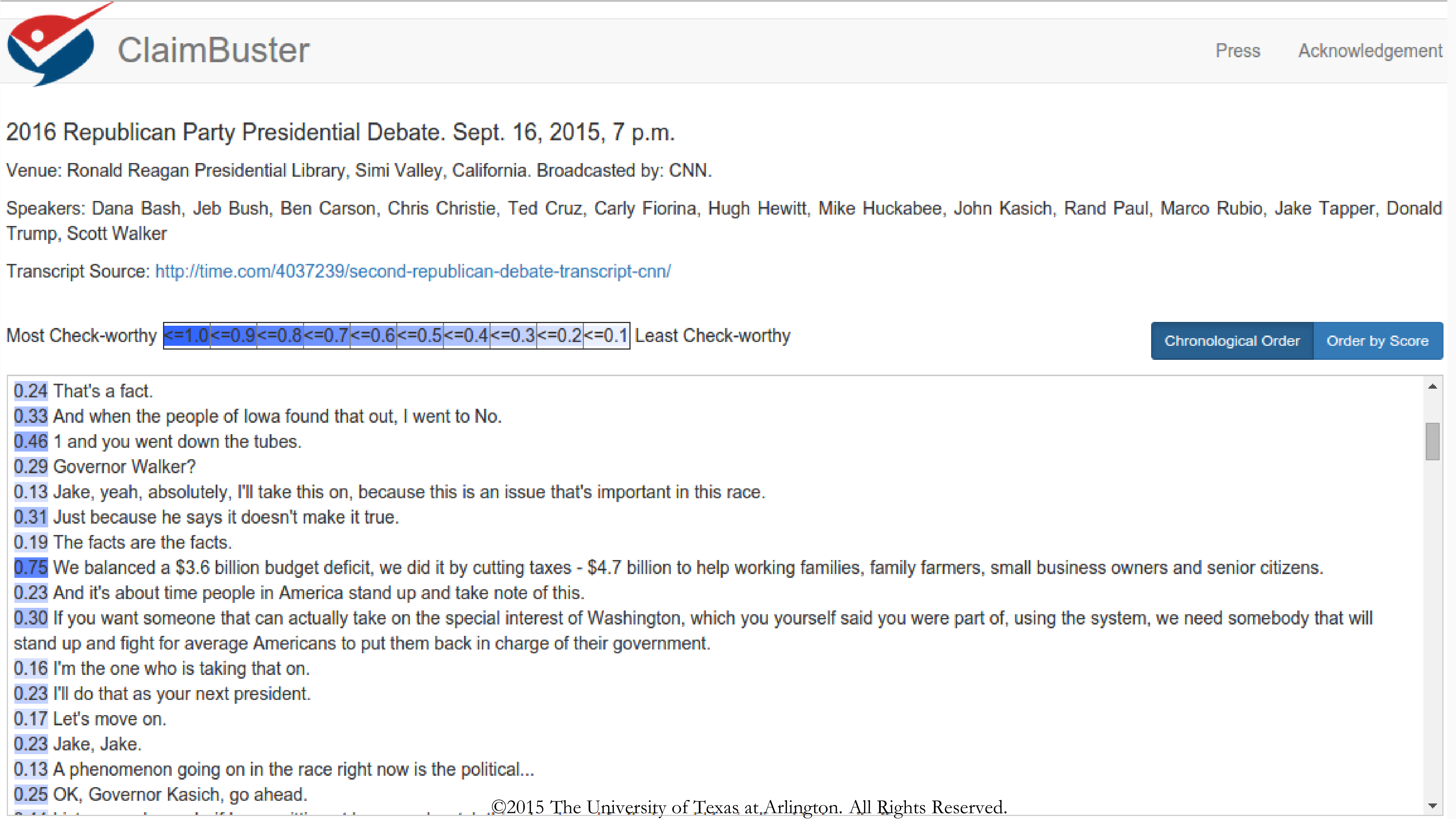
Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.

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/

0.48 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.
0.18 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.

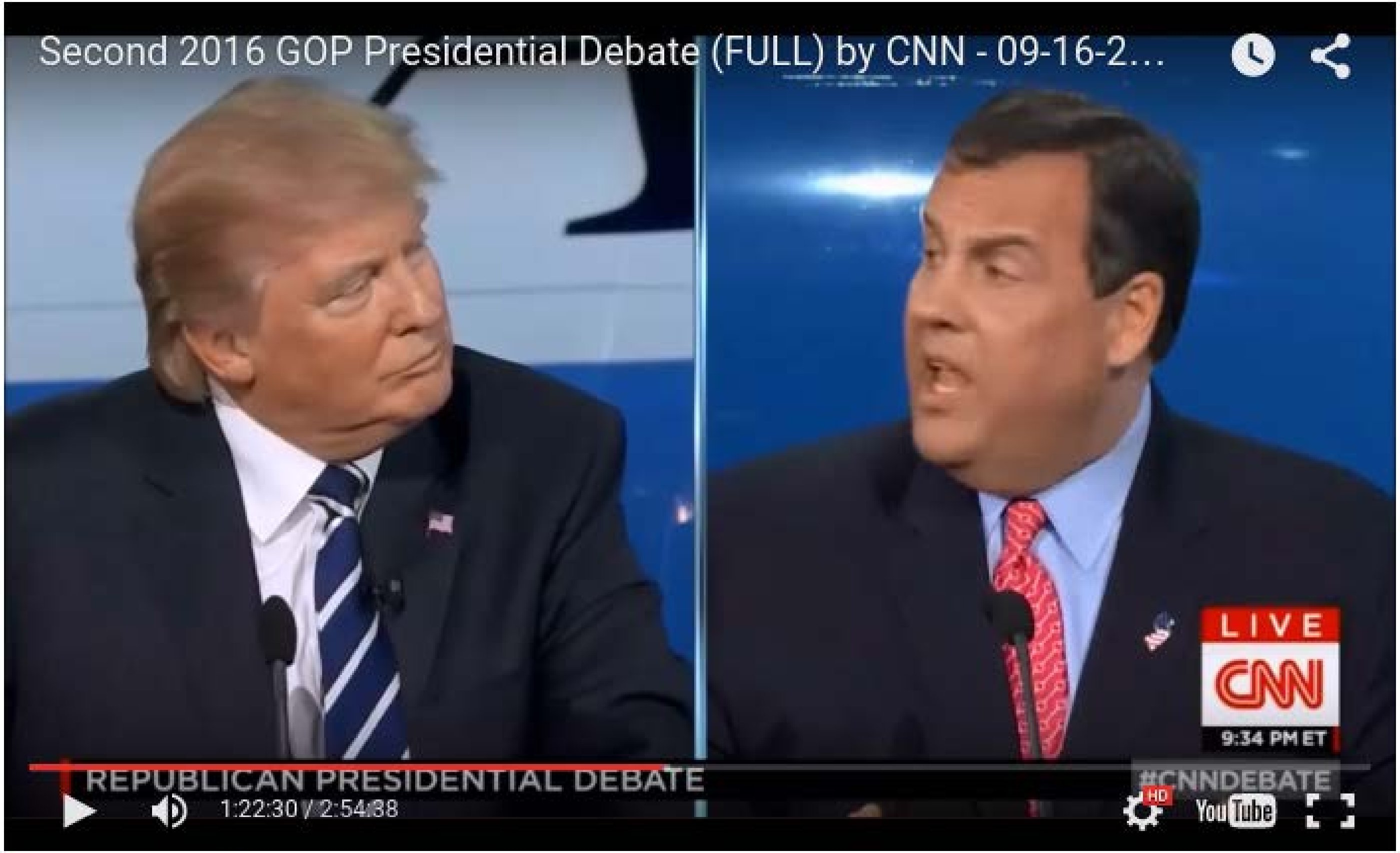
Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.

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

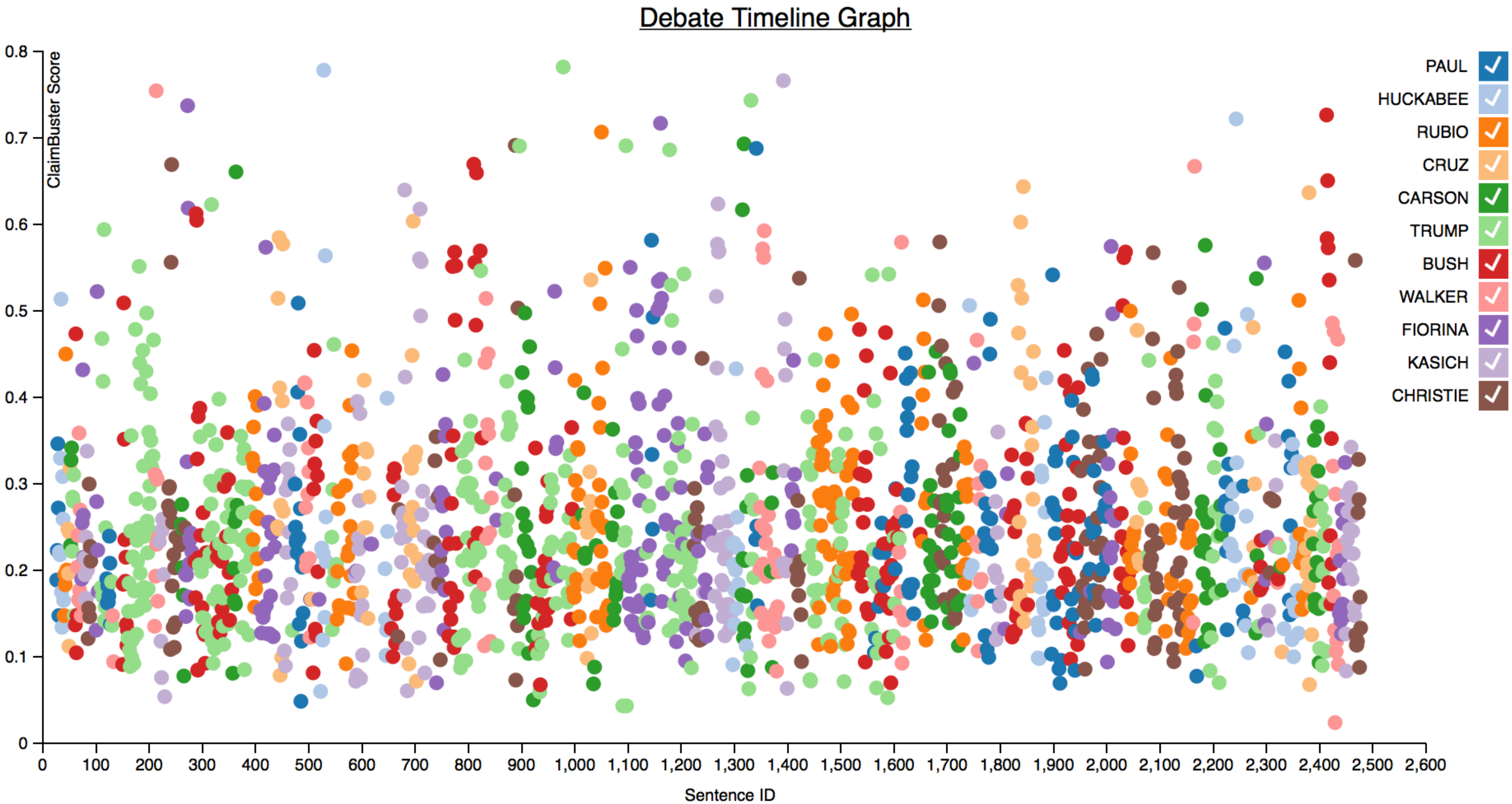
Transcript Source: http://time.com/4037239/second-republican-debate-transcript-cnn/

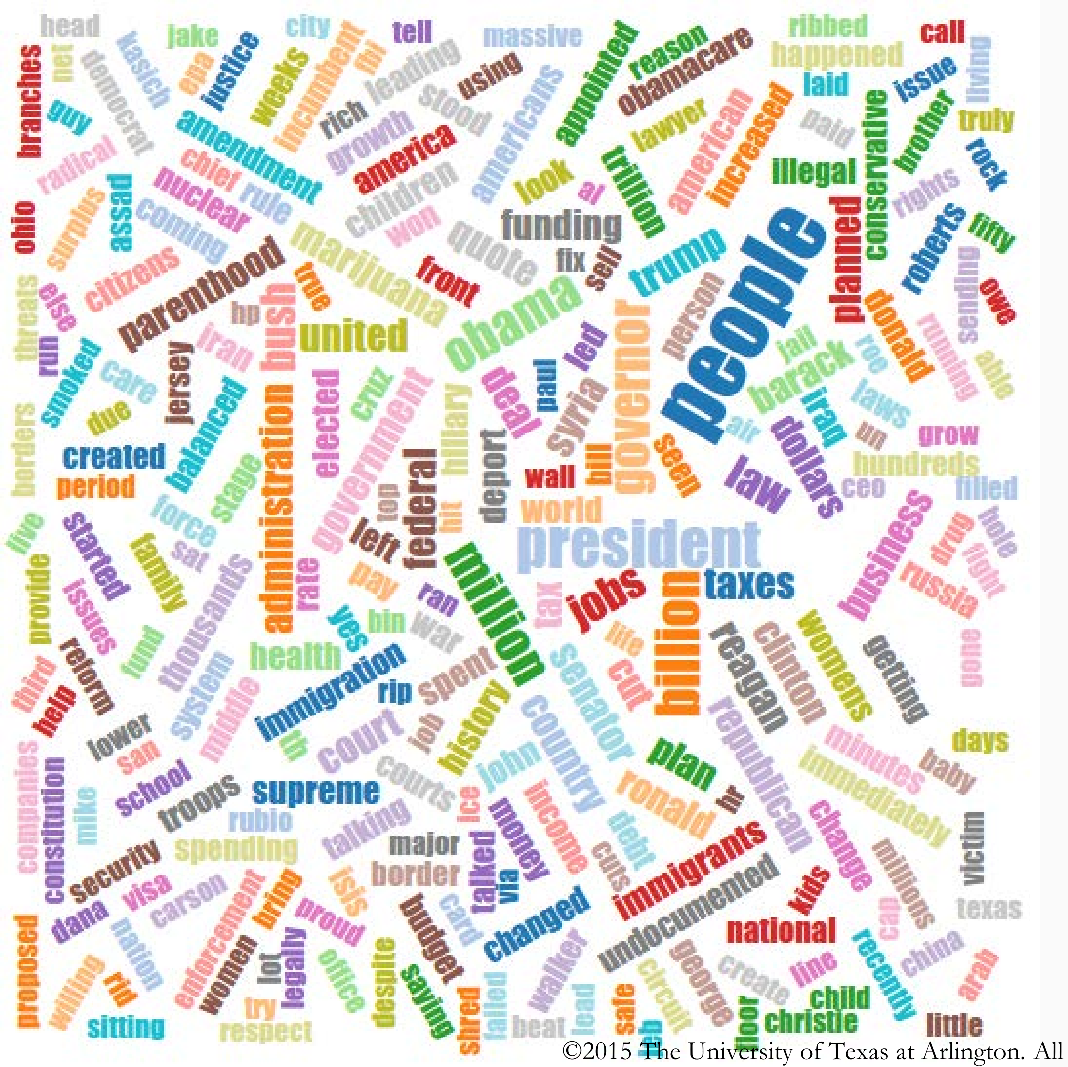
Most Check-worthy <=1.0 <=0.9 <=0.8 <=0.7 <=0.6 <=0.5 <=0.4 <=0.3 <=0.2 <=0.1 Least Check-worthy

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.


Second 2016 GOP Presidential Debate (FULL) by CNN - 09-16-2...
REPUBLICAN PRESIDENTIAL DEBATE
1:22:30 / 2:54:38
LIVE
CNN
9:34 PM ET
#CNNDEBATE
HD
YouTube

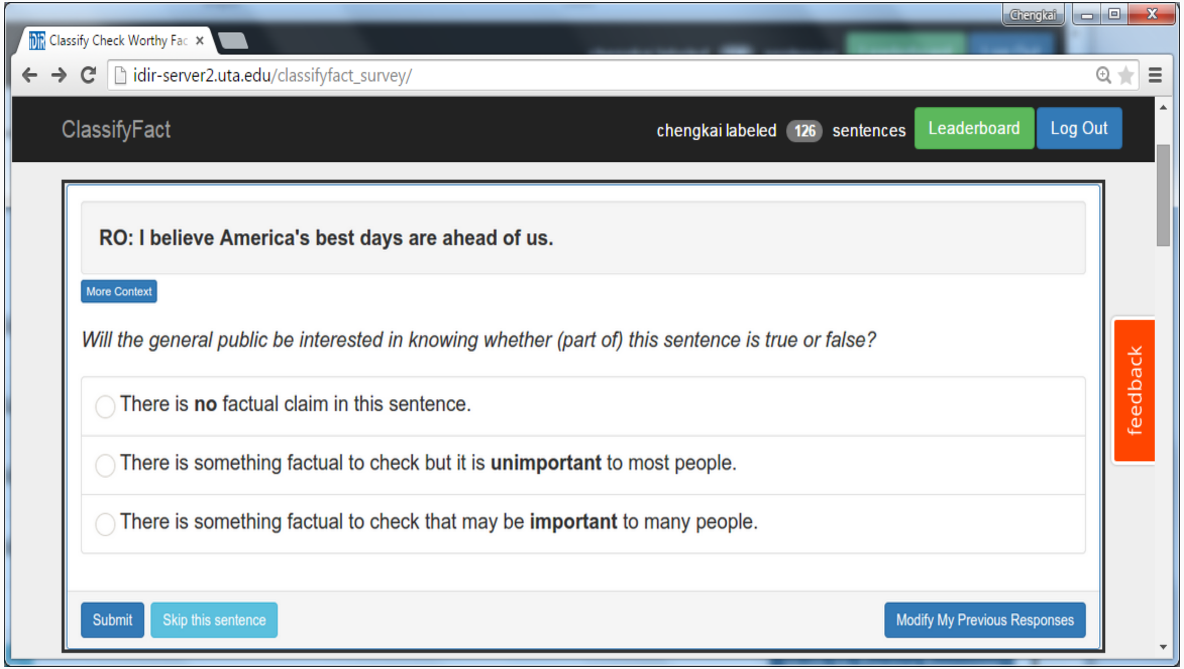
0.53	In fact, today, on the front page of the Wall Street Journal, they fired another 25 or 30,000 people saying we still haven't recovered from the catastrophe.


Debate Timeline Graph
ClaimBuster Score
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300
1,400
1,500
1,600
1,700
1,800
1,900
2,000
2,100
2,200
2,300
2,400
2,500
2,600
Sentence ID
PAUL ✓
HUCKABEE ✓
RUBIO ✓
CRUZ ✓
CARSON ✓
TRUMP ✓
BUSH ✓
WALKER ✓
FIORINA ✓
KASICH ✓
CHRISTIE ✓


head
jake
city
tell
massive
ribbed
call
branches
net
kasich
epa
justice
weeks
incumbent
fbi
leading
using
appointed
reason
obamacare
happened
laid
issue
living
guy
democrat
rich
stood
americans
lawyer
paid
truly
amendment
growth
america
american
increased
conservative
brother
rock
radical
chief
nuclear
rule
children
look
al
trillion
illegal
rights
ohio
surplus
assad
coming
won
funding
quote
fix
sell
trump
planned
roberts
fifty
citizens
parenthood
marijuana
front
people
sending
owe
threats
else
hp
bush
true
united
obama
governor
person
jail
barack
roe
donald
running
able
run
smoked
iran
care
jersey
due
balanced
administration
elected
cruz
government
hillary
deal
paul
led
syria
iraq
laws
un
grow
borders
created
period
force
stage
federal
deport
wall
bill
seen
law
dollars
air
hundreds
ceo
filled
live
started
family
sat
top
left
hit
world
president
business
drug
role
fight
provide
issues
thousands
rate
pay
ran
million
tax
jobs
billion
taxes
russia
health
bin
war
yes
life
cut
reagan
clinton
womens
getting
gone
third
return
fund
system
middle
immigration
rip
spent
history
senator
country
plan
republican
minutes
days
help
lower
san
up
court
job
john
ronald
hr
immediately
baby
companies
constitution
mike
school
troops
supreme
courts
ice
rubio
talking
money
income
debt
immigrants
change
millions
victim
security
spending
major
border
talked
via
cuts
kids
texas
dana
visa
carson
bring
isis
proud
enforcement
budget
card
changed
undocumented
national
cap
china
proposed
nation
women
lot
legally
office
despite
walker
george
create
line
child
recently
arab
willing
rid
try
respect
sitting
shred
failed
beat
lead
safe
circuit
floor
christie
little

You are Invited

http://bit.ly/claimbusters


Classify Check Worthy Fac
Chengkai
idir-server2.uta.edu/classifyfact_survey/
ClassifyFact
Leaderboard
Log Out
More Context
feedback
Submit
Skip this sentence
Modify My Previous Responses

chengkai labeled 126 sentences

RO: I believe America's best days are ahead of us.

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 something factual to check but it is **unimportant** to most people.

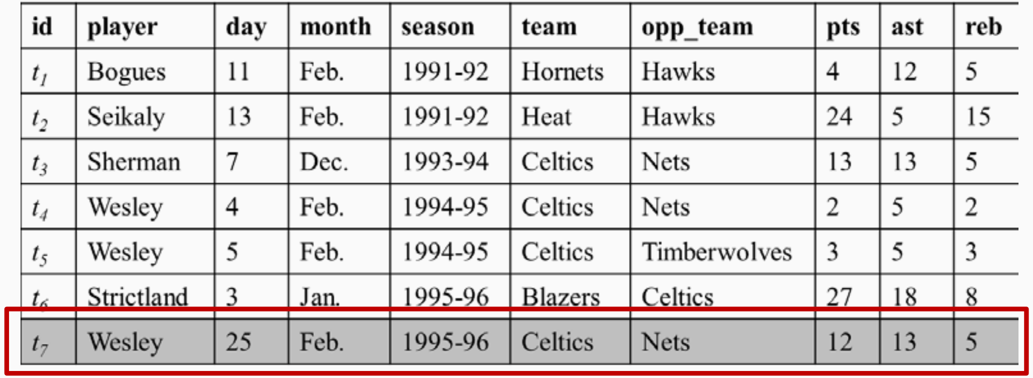
○ There is something factual to check that may be **important** to many people.

FactWatcher

Automated Monitoring of Facts from Real-World Events

FactWatcher

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

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

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

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

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

Prominent streaks

Long consecutive subsequence of high values in a sequence

One-of-the-few objects

Qualifying statements that can only be made for very few objects

Situational facts

Comparison contexts and spaces that make a given object stand out

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

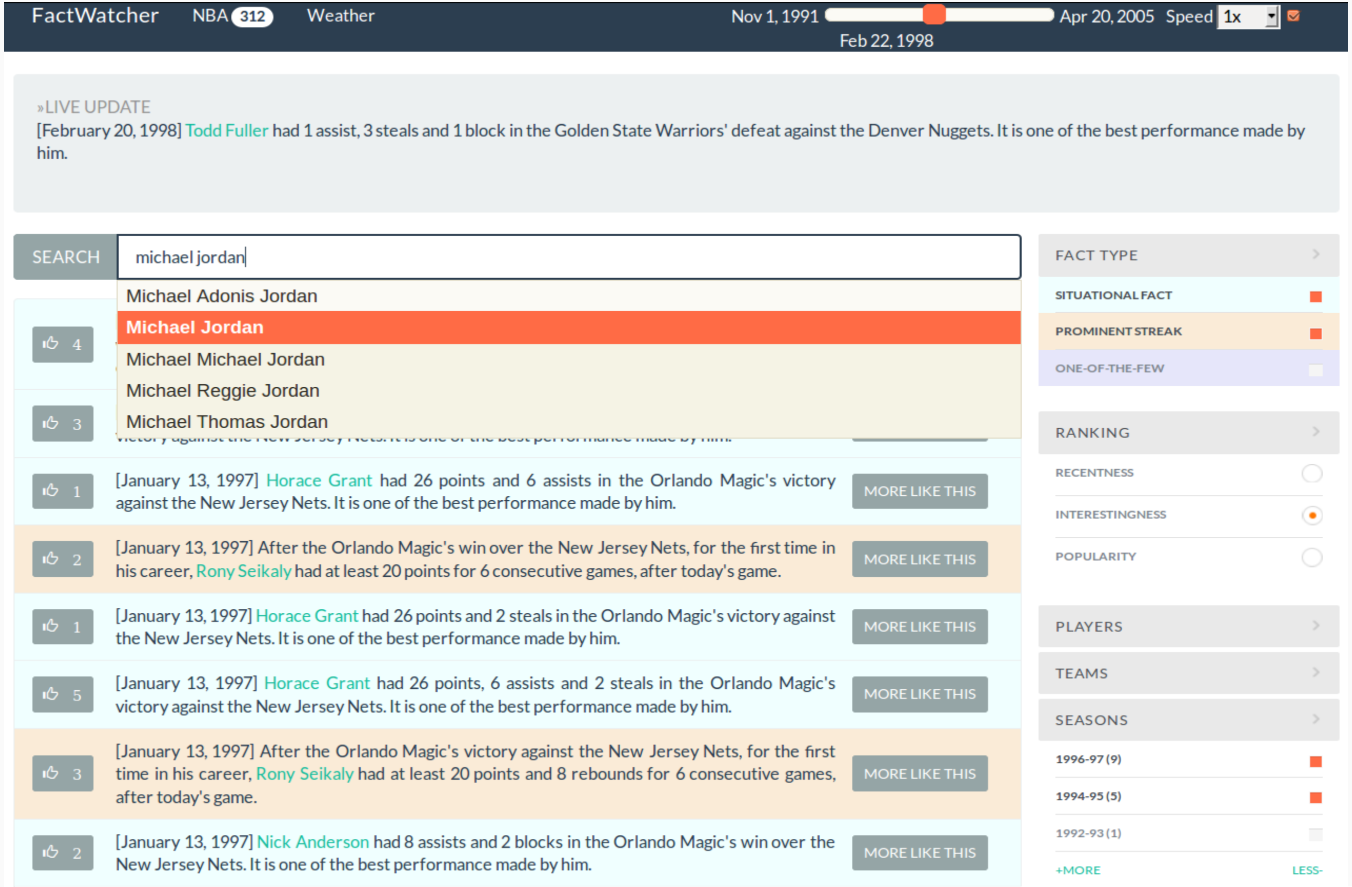
FactWatcher Finds Three Types of Facts (and can be Extended)

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



»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

4

3 victory against the New Jersey Nets. It is one of the best performance made by him.

1 [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

2 [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

1 [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

5 [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

3 [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

2 [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

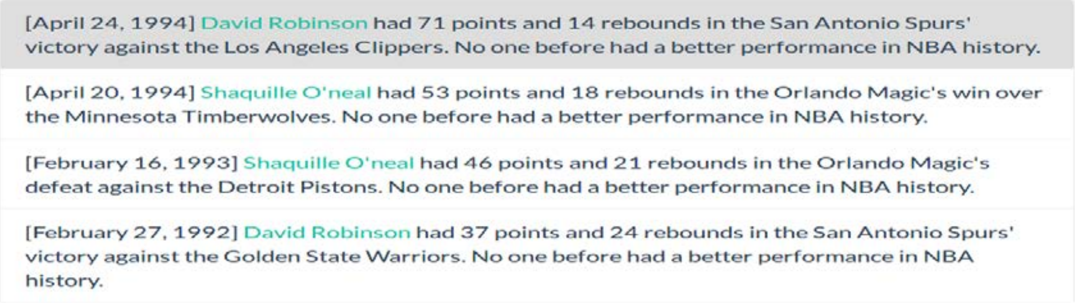
Presented In

VLDB 2014

Excellent Demo Award

COMPUTATION + JOURNALISM SYMPOSIUM 2014

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

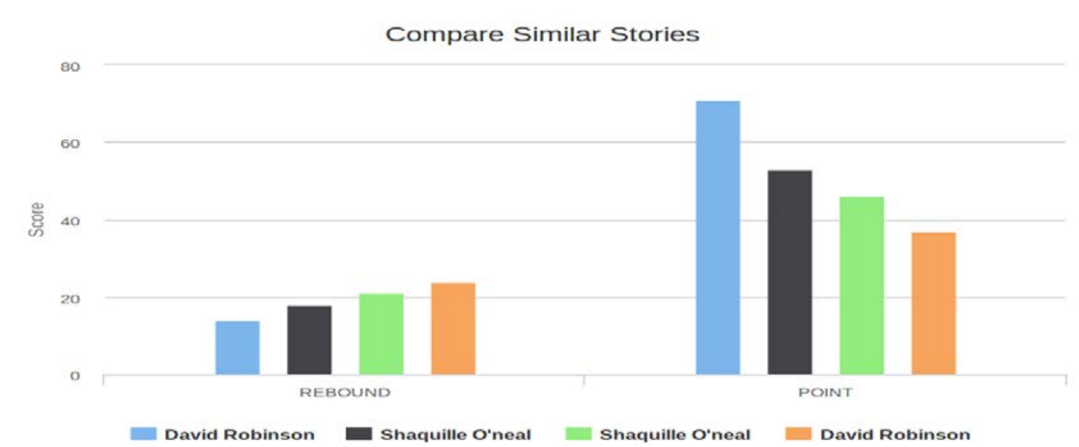


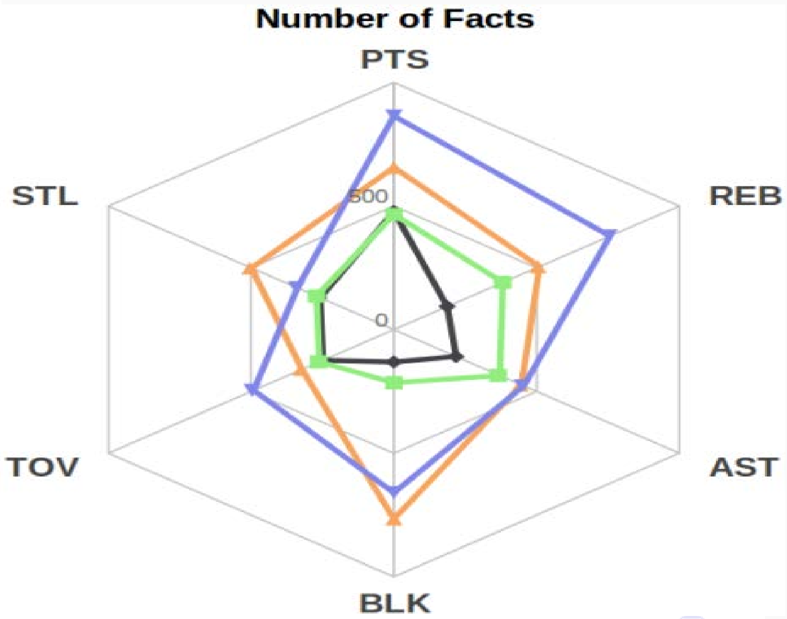
[April 24, 1994] David Robinson had 71 points and 14 rebounds in the San Antonio Spurs' victory against the Los Angeles Clippers. No one before had a better performance in NBA history.

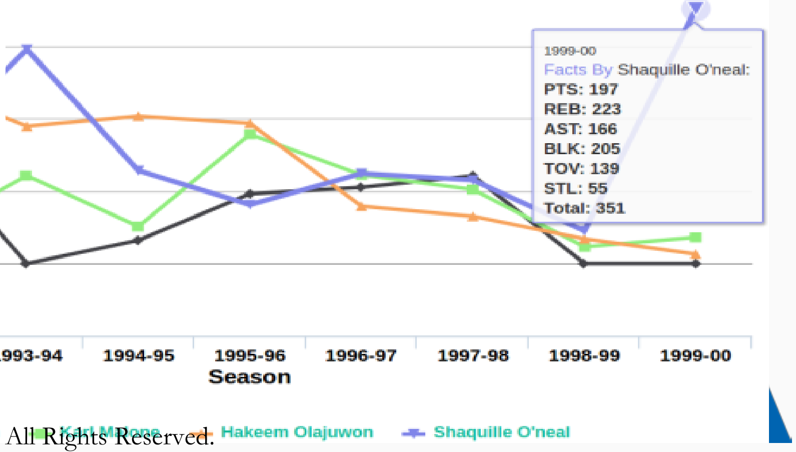
[April 20, 1994] Shaquille O'neal had 53 points and 18 rebounds in the Orlando Magic's win over the Minnesota Timberwolves. No one before had a better performance in NBA history.

[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.







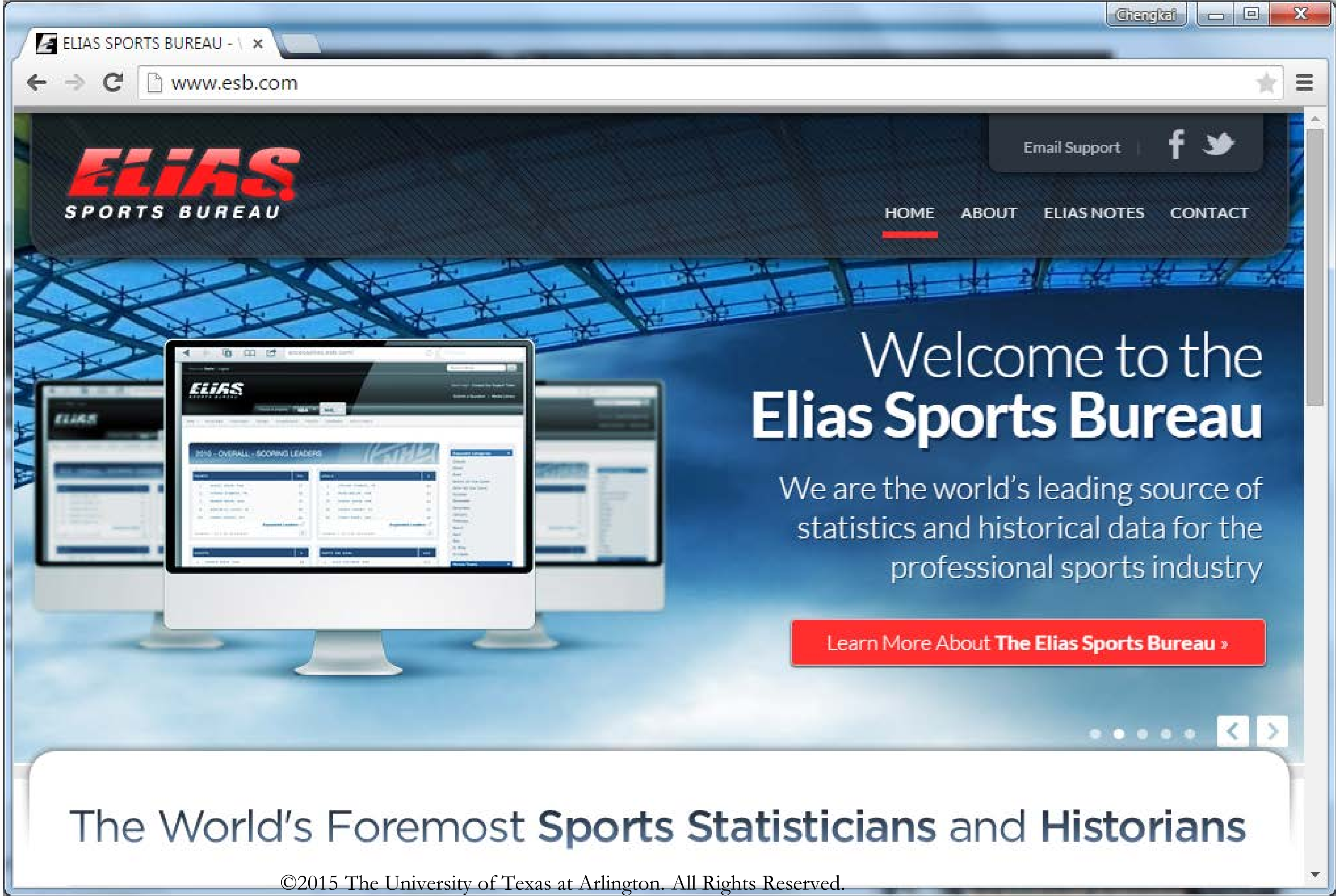
1999-00
Facts By Shaquille O'neal:
PTS: 197
REB: 223
AST: 166
BLK: 205
TOV: 139
STL: 55
Total: 351

How were such Facts Discovered in Current Systems?

Our (educated?) guess

- Experts monitor real-world events (e.g., watching an NBA game), have a gut-feeling, issue database queries, check out or not
- Prepared facts-to-be (e.g., Nowitzki only needs 477 more points to surpass O'Neal. Perhaps will happen around Christmas 2015)
- Predefined templates of facts/database queries
- Perhaps in-house systems/algorithms similar to FactWatcher

Elias Sports Bureau



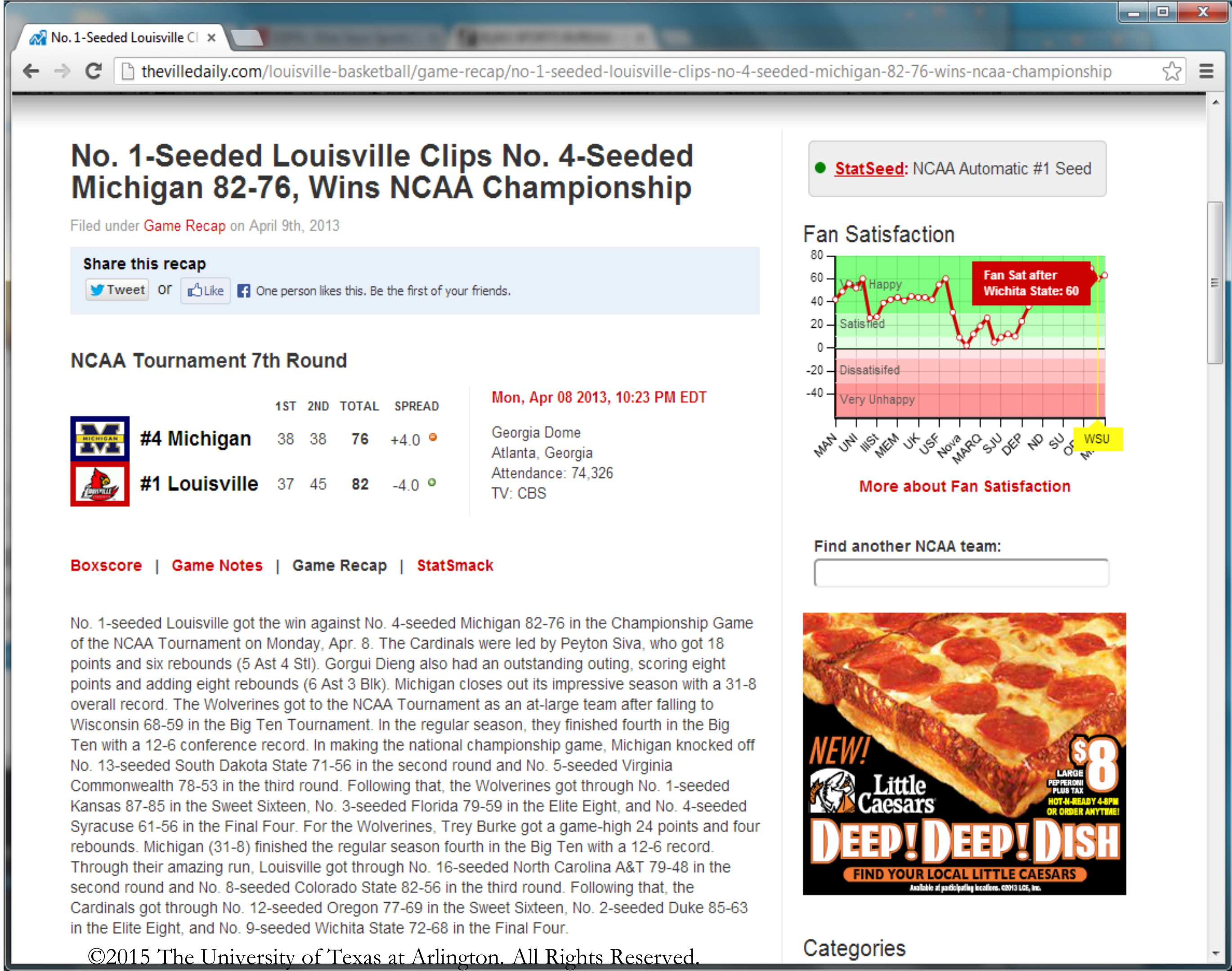
Welcome to the
Elias Sports Bureau

We are the world's leading source of statistics and historical data for the professional sports industry

Learn More About **The Elias Sports Bureau** »

The World's Foremost **Sports Statisticians** and **Historians**

StatSheet



No. 1-Seeded Louisville Cl ×
thevilledaily.com/louisville-basketball/game-recap/no-1-seeded-louisville-clips-no-4-seeded-michigan-82-76-wins-ncaa-championship

No. 1-Seeded Louisville Clips No. 4-Seeded Michigan 82-76, Wins NCAA Championship

Filed under Game Recap on April 9th, 2013

Share this recap

Tweet or Like One person likes this. Be the first of your friends.

NCAA Tournament 7th Round

	1ST	2ND	TOTAL	SPREAD
#4 Michigan	38	38	76	+4.0 ●
#1 Louisville	37	45	82	-4.0 ●

Mon, Apr 08 2013, 10:23 PM EDT

Georgia Dome
Atlanta, Georgia
Attendance: 74,326
TV: CBS

Boxscore | Game Notes | Game Recap | StatSmack

No. 1-seeded Louisville got the win against No. 4-seeded Michigan 82-76 in the Championship Game of the NCAA Tournament on Monday, Apr. 8. The Cardinals were led by Peyton Siva, who got 18 points and six rebounds (5 Ast 4 Stl). Gorgui Dieng also had an outstanding outing, scoring eight points and adding eight rebounds (6 Ast 3 Blk). Michigan closes out its impressive season with a 31-8 overall record. The Wolverines got to the NCAA Tournament as an at-large team after falling to Wisconsin 68-59 in the Big Ten Tournament. In the regular season, they finished fourth in the Big Ten with a 12-6 conference record. In making the national championship game, Michigan knocked off No. 13-seeded South Dakota State 71-56 in the second round and No. 5-seeded Virginia Commonwealth 78-53 in the third round. Following that, the Wolverines got through No. 1-seeded Kansas 87-85 in the Sweet Sixteen, No. 3-seeded Florida 79-59 in the Elite Eight, and No. 4-seeded Syracuse 61-56 in the Final Four. For the Wolverines, Trey Burke got a game-high 24 points and four rebounds. Michigan (31-8) finished the regular season fourth in the Big Ten with a 12-6 record. Through their amazing run, Louisville got through No. 16-seeded North Carolina A&T 79-48 in the second round and No. 8-seeded Colorado State 82-56 in the third round. Following that, the Cardinals got through No. 12-seeded Oregon 77-69 in the Sweet Sixteen, No. 2-seeded Duke 85-63 in the Elite Eight, and No. 9-seeded Wichita State 72-68 in the Final Four.

● StatSeed: NCAA Automatic #1 Seed

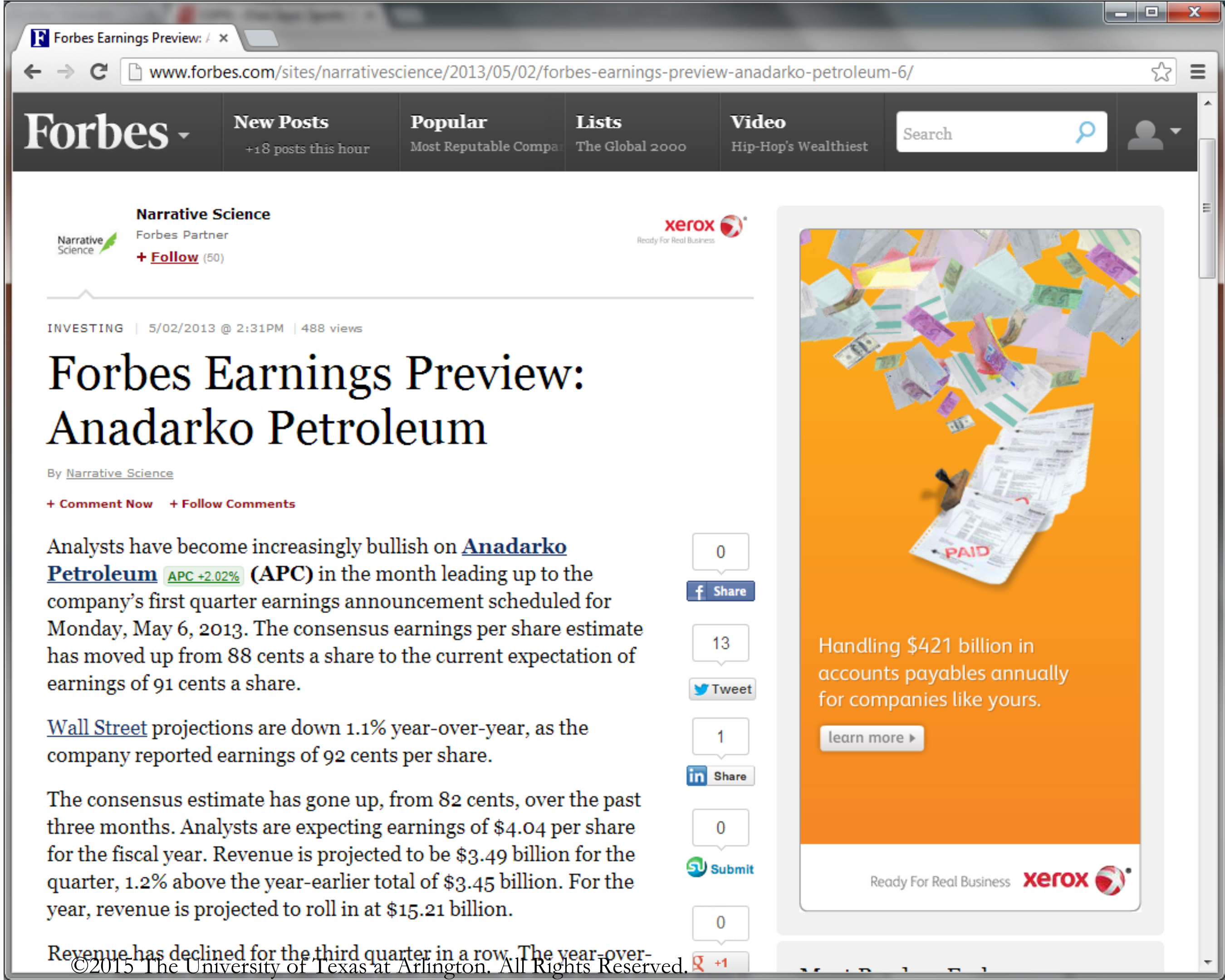
Fan Satisfaction

More about Fan Satisfaction

Find another NCAA team:

Categories

Narrative Science



Narrative Science
Forbes Partner
+ Follow (50)

INVESTING | 5/02/2013 @ 2:31PM | 488 views

Forbes Earnings Preview: Anadarko Petroleum

By Narrative Science

+ Comment Now + Follow Comments

Analysts have become increasingly bullish on Anadarko Petroleum APC +2.02% (APC) in the month leading up to the company's first quarter earnings announcement scheduled for Monday, May 6, 2013. The consensus earnings per share estimate has moved up from 88 cents a share to the current expectation of earnings of 91 cents a share.

Wall Street projections are down 1.1% year-over-year, as the company reported earnings of 92 cents per share.

The consensus estimate has gone up, from 82 cents, over the past three months. Analysts are expecting earnings of $4.04 per share for the fiscal year. Revenue is projected to be $3.49 billion for the quarter, 1.2% above the year-earlier total of $3.45 billion. For the year, revenue is projected to roll in at $15.21 billion.

Revenue has declined for the third quarter in a row. The year-over-

Incremental Discovery of Prominent Situational Facts. Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, Cong Yu. ICDE 2014, pages 112-123.

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)

Skyline

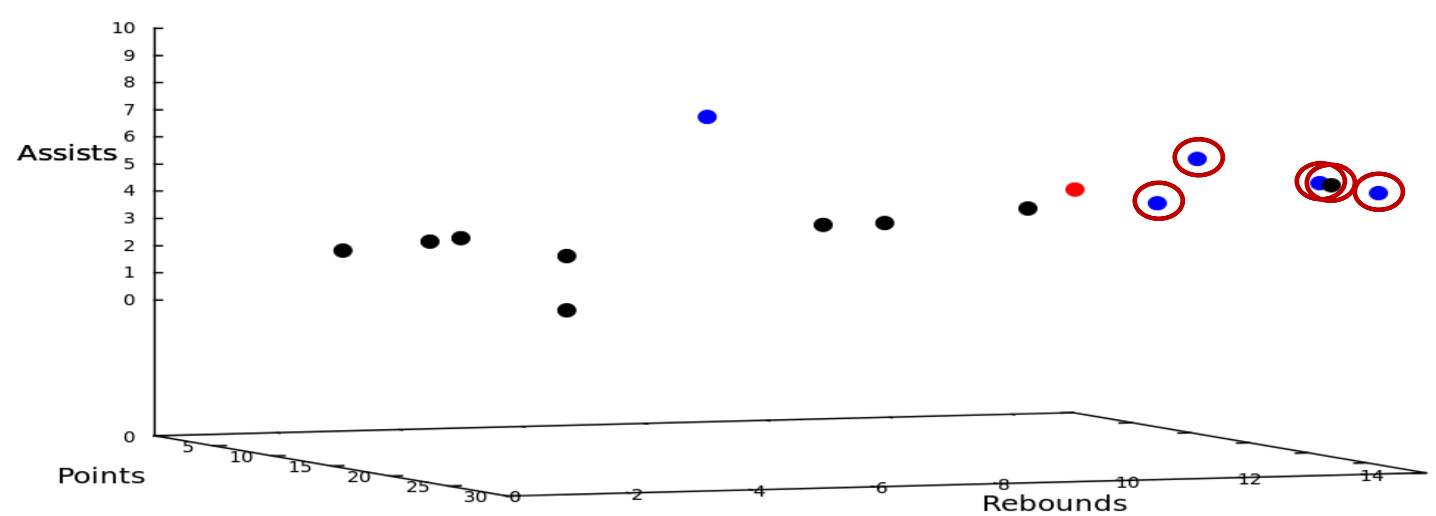
www.rtkl.com

jansport.com

www.utepprintstore.com

Situational Facts

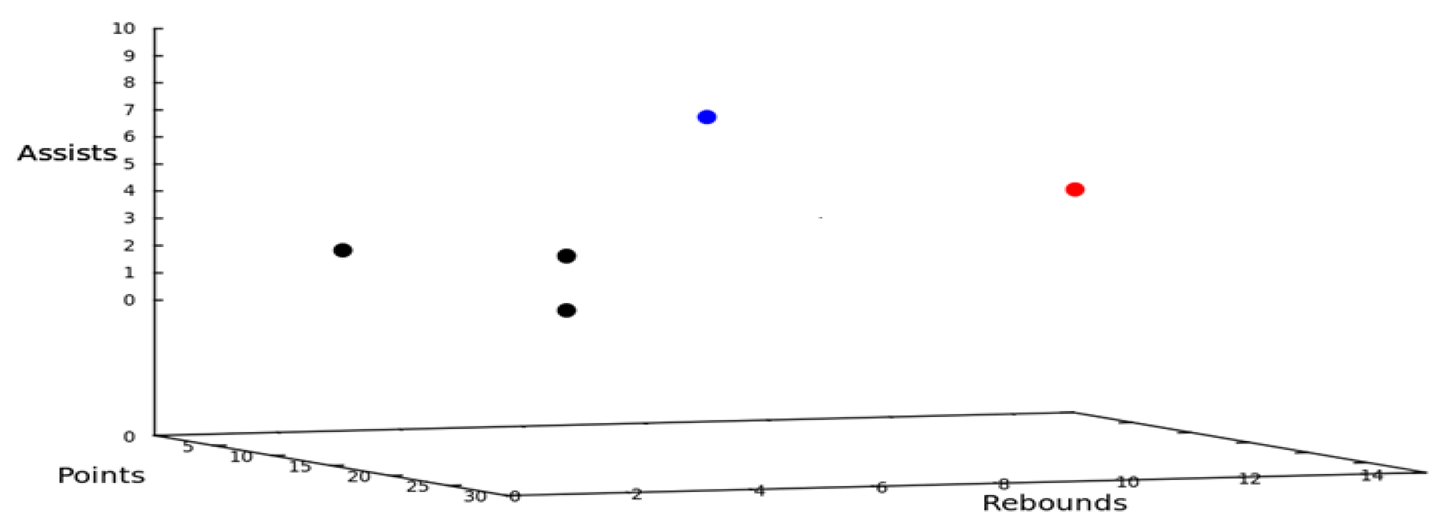
“Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992.” (http://espn.go.com/espn/elias?date=20130205)



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."

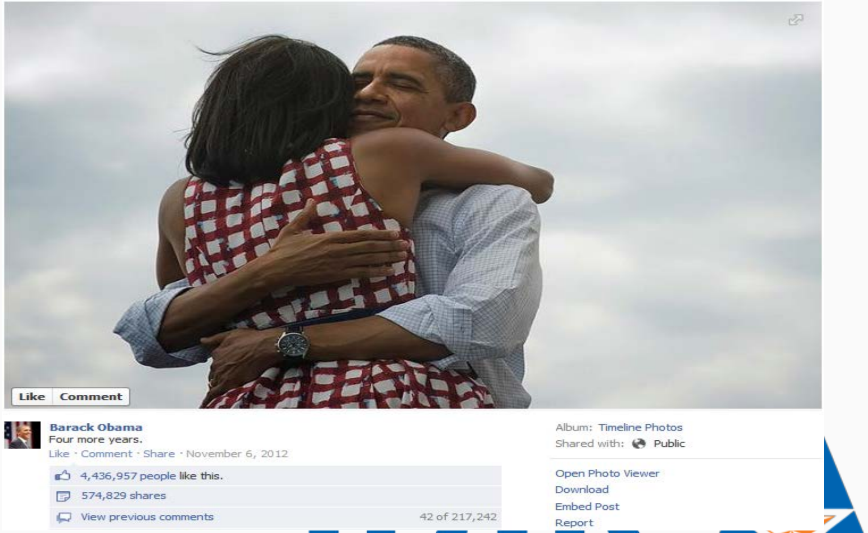
(http://espn.go.com/espn/elias?date=20130205)



Situational Facts

“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/President Obama Sets New Social Media Record)



Barack Obama
Four more years.
Like · Comment · Share · November 6, 2012

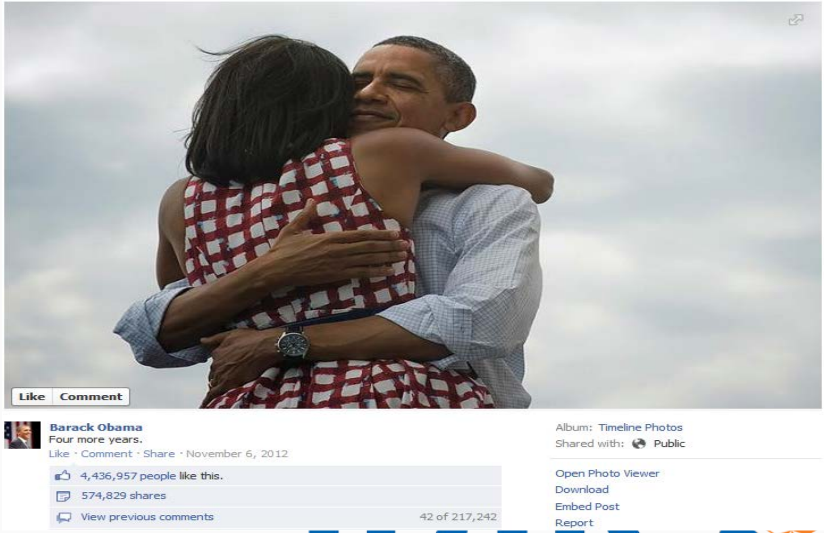
4,436,957 people like this.
574,829 shares
View previous comments 42 of 217,242

Album: Timeline Photos
Shared with: Public

Situational Facts

“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/President Obama Sets New Social Media Record)



Barack Obama
Four more years.
Like · Comment · Share · November 6, 2012

4,436,957 people like this.

574,829 shares

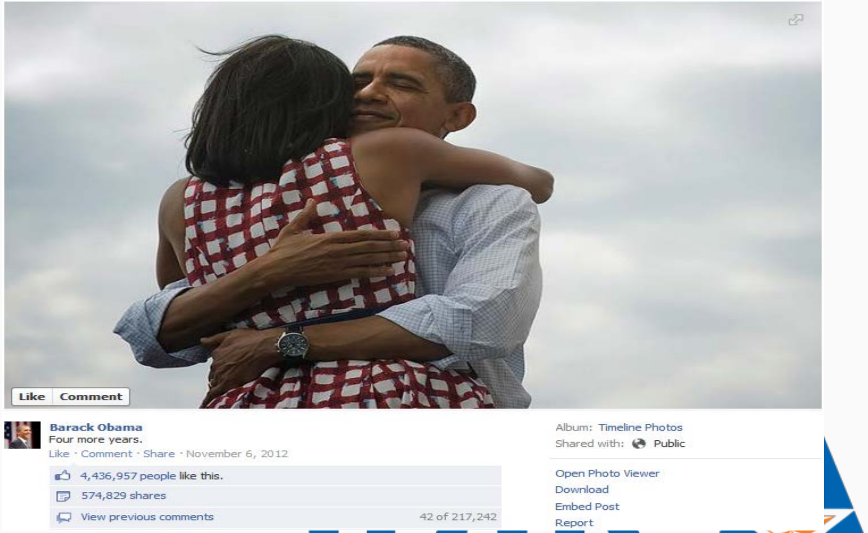
View previous comments 42 of 217,242

Album: Timeline Photos
Shared with: Public

Situational Facts

“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/President Obama Sets New Social Media Record)



Barack Obama
Four more years.
Like · Comment · Share · November 6, 2012

4,436,957 people like this.

574,829 shares

View previous comments 42 of 217,242

Album: Timeline Photos
Shared with: Public

Situational Facts

- Stock Data: Stock A becomes the first stock in history with price over $300 and market cap over $400 billion.
- Weather Data: Today’s measures of wind speed and humidity are x and y, respectively. City B has never encountered such high wind speed and humidity in March.
- Criminal Records: There were 50 DUI arrests and 20 collisions in city C yesterday, the first time in 2013.

Financial Analyst

Journalists

Scientists

Citizens

A Mini-world of Basketball Gamelogs

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

Last tuple appended to table

A Mini-world of Basketball Gamelogs

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

A Mini-world of Basketball Gamelogs

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			Feb.				12	13	5

A Mini-world of Basketball Gamelogs

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			Feb.				12	13	5

- 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.

A Mini-world of Basketball Gamelogs

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				1995-96			12	13	5

A Mini-world of Basketball Gamelogs

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					Celtics	Nets		13	5

- Wesley had 13 assists and 5 rebounds on February 25, 1996 to become the second Celtics player with a 13/5 (assists/rebounds) game against the Nets.

Problem Definition

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

append-only table

Problem Definition

❑Constraint (*C*): $d_1=v_1 \wedge d_2=v_2 \wedge . . . \wedge d_n=v_n,\ v_i \in dom(d_i) \cup \{*\}$

- team=*Celtics* ∧ opp_team=*Nets*

id	**player**	**day**	**month**	**season**	**team**	**opp_team**	**pts**	**ast**	**rb**
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

Problem Definition

❑Constraint-Measure Pair (C, M): Combination of a constraint and measure subspace

- (team=*Celtics* ∧ opp_team=*Nets*,{assists,rebounds})

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

Problem Definition

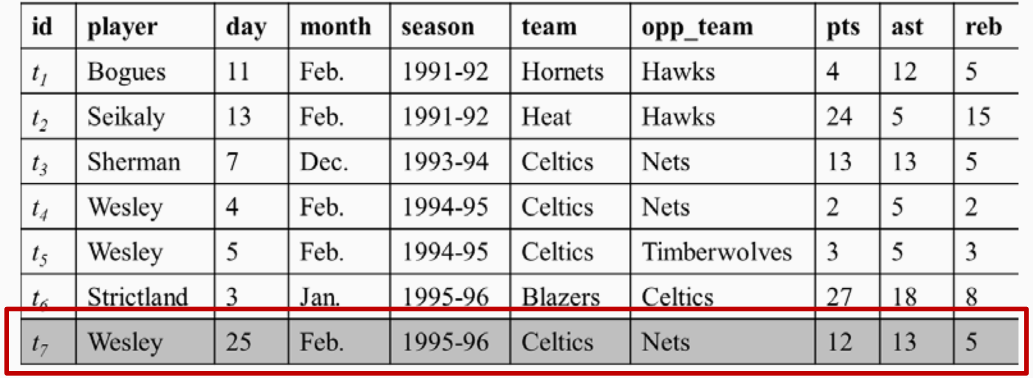
- Contextual skyline: skyline regarding (C, M)
  - $\sigma_{\text{team}=\mathit{Celtics}\ \wedge\ \text{opp\_team}=\mathit{Nets}}(R)$, M={assists,rebounds}
    - $\{t_3\}$

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

FactWatcher

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

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

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

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

Generate factual claim

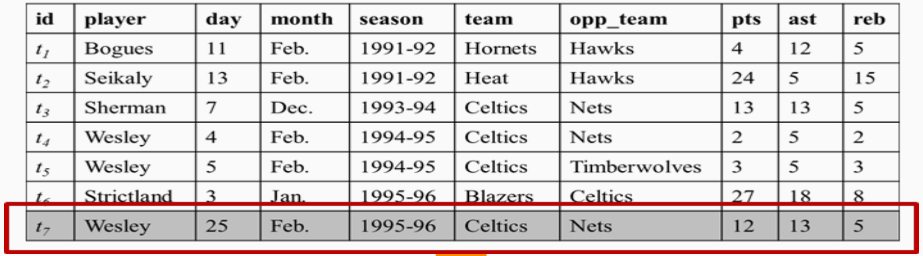
Wesley had 12 points, 13 assists and 5 rebounds on February 25, 1996 to become the first player with a 12/13/5 (points/assists/rebounds) in February.

Related Work

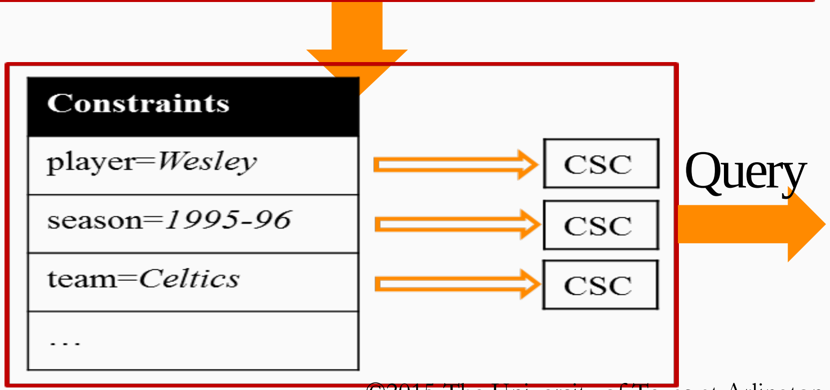
- Conventional skyline analysis (Borzsonyi et al. ICDE 2001)
  - Q: context, measure subspace ⟹ A: contextual skyline tuples
    - ✓ Our focus--- A: tuple ⟹ Q: constraint-measure pairs

Related Works

- Compressed Skycube (Xia et al. SIGMOD 2006)
  - Update compressed skycube in monitoring fashion
    - ✓We adapted CSC for each constraint: Constraint-CSC



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



Constraints

player=*Wesley*

season=*1995-96*

team=*Celtics*

…

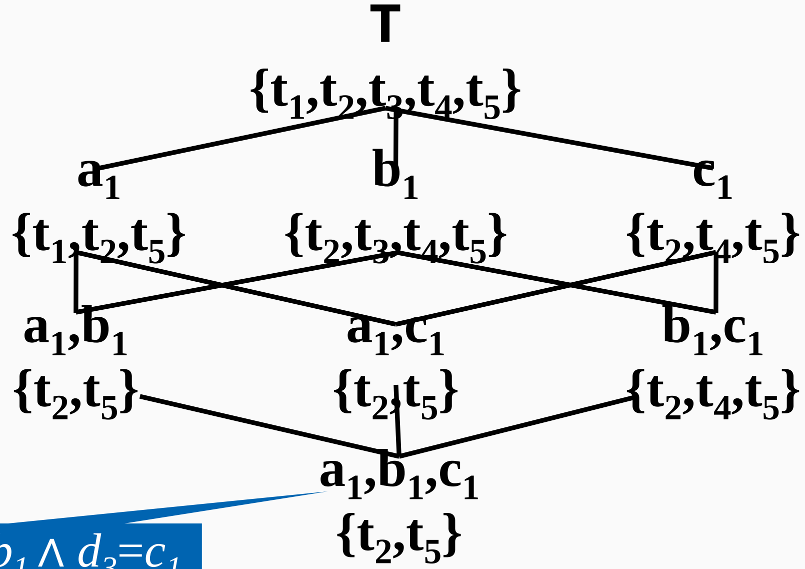
Constraint	**Measure**
month=*Feb*	pts, ast, reb
opp_team=*Nets*	ast, reb
team=*Celtics* & opp_team=*Nets*	ast, reb
…	…

Related Works

- Prominent Analysis by Ranking (Wu et. Al. VLDB 2009)
  - Static data, onetime query
    - ✓ We dealt on continuous data, standing query
  - Find the contexts where an object is ranked high in a single scoring attribute
    - ✓ We considered skyline on multiple measure subspaces

Modeling

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

$\{t_1,t_2,t_3,t_4,t_5\}$

a_1

$\{t_1,t_2,t_5\}$

b_1

$\{t_2,t_3,t_4,t_5\}$

c_1

$\{t_2,t_4,t_5\}$

a_1,b_1

$\{t_2,t_5\}$

a_1,c_1

$\{t_2,t_5\}$

b_1,c_1

$\{t_2,t_4,t_5\}$

a_1,b_1,c_1

$\{t_2,t_5\}$

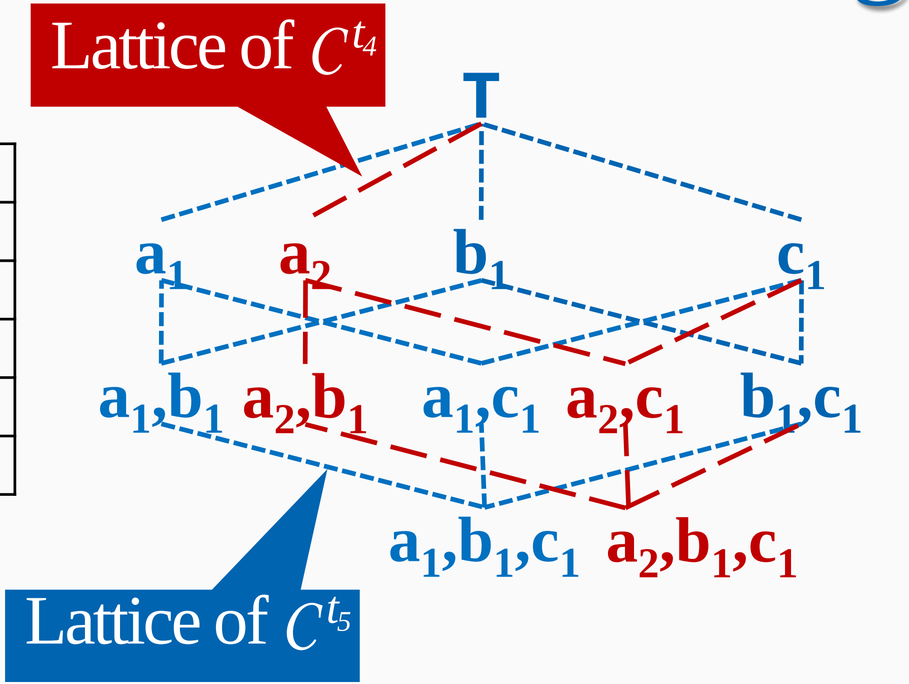
$d_1=a_1 \wedge d_2=b_1 \wedge d_3=c_1$

Lattice of C^{t_5}

Tuple Satisfied Constraint C^{t}: If $\forall d_i \in \mathcal{D}$, $C.d_i = *$ or $C.d_i = t.d_i$, t satisfies C.

Modeling

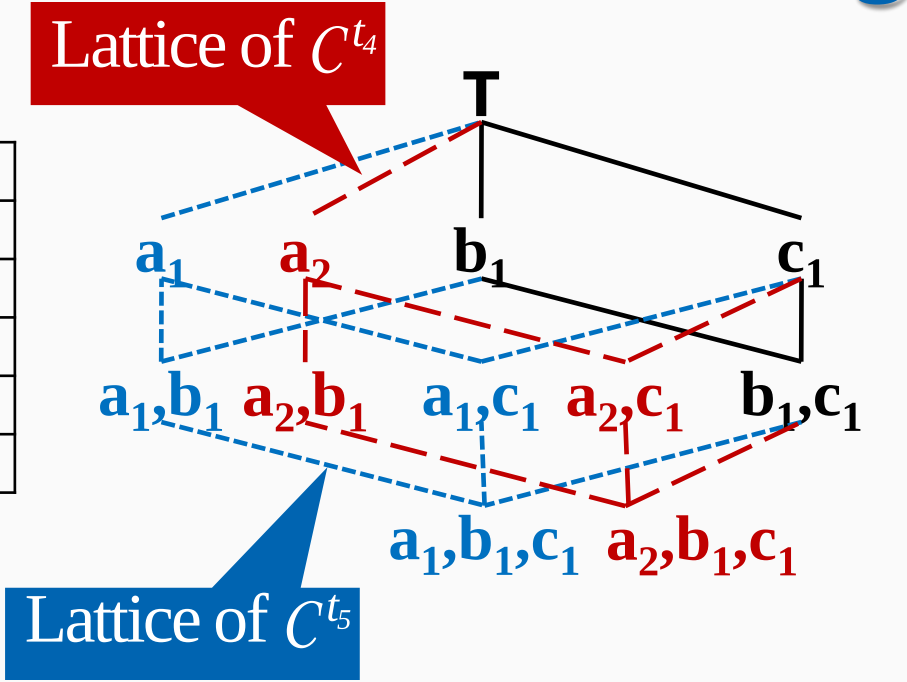
id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15


Lattice of C^{t_4}

Lattice of C^{t_5}

Modeling

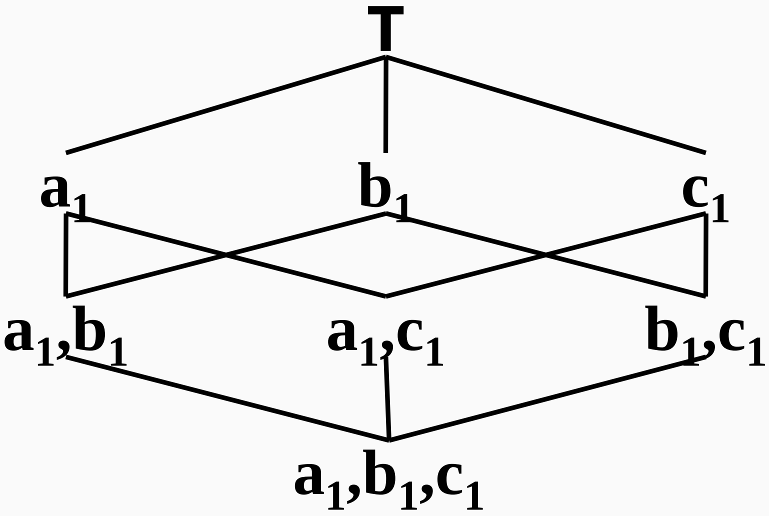
id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Lattice Intersection: $C^{t_4,t_5}=C^{t_4} \cap C^{t_5}$

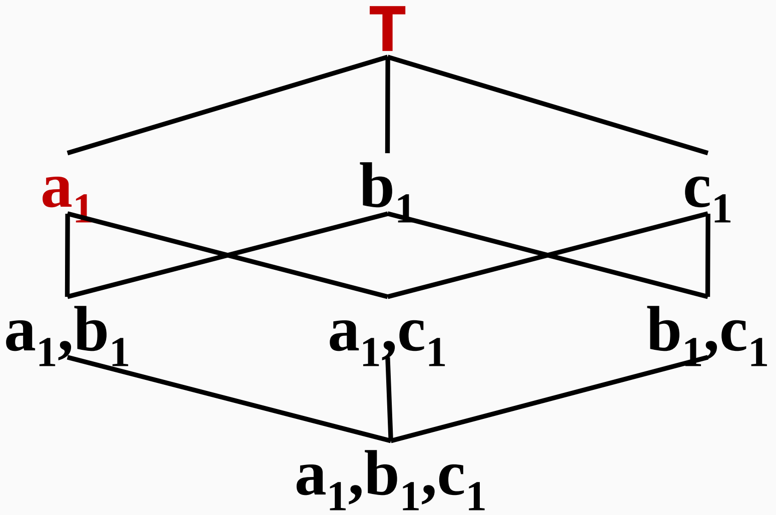
Brute-Force Approach

id	***d_1***	***d_2***	***d_3***	***m_1***	***m_2***
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



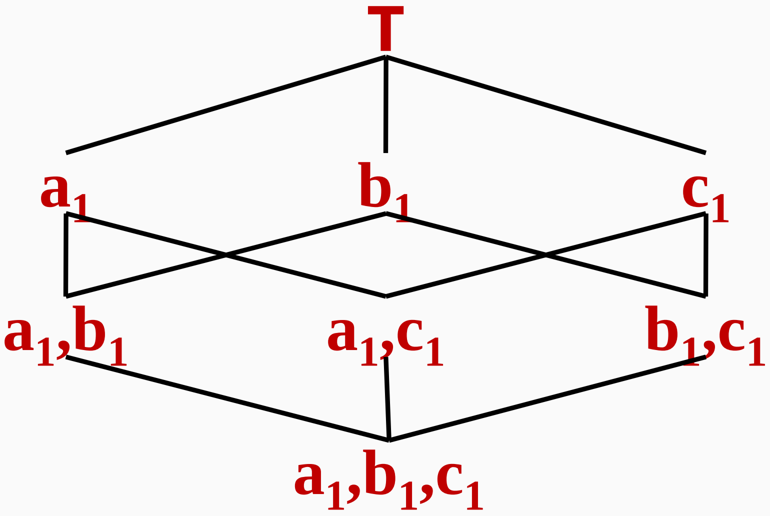
Brute-Force Approach

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



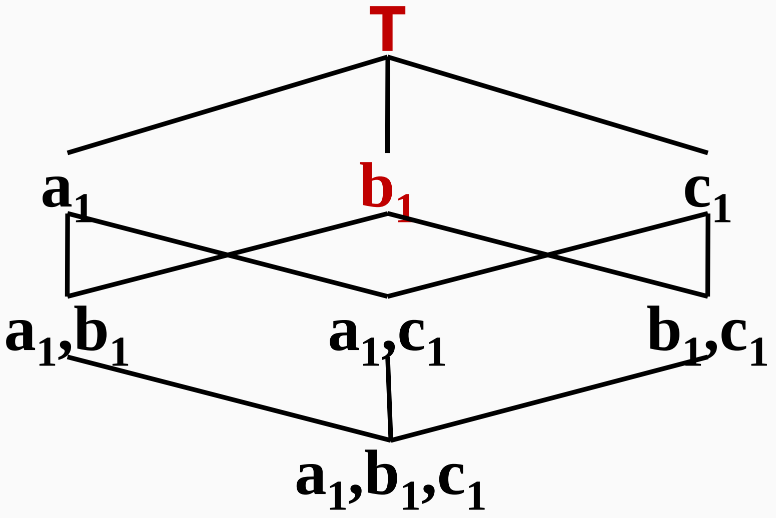
Brute-Force Approach

id	***d_1***	***d_2***	***d_3***	***m_1***	***m_2***
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



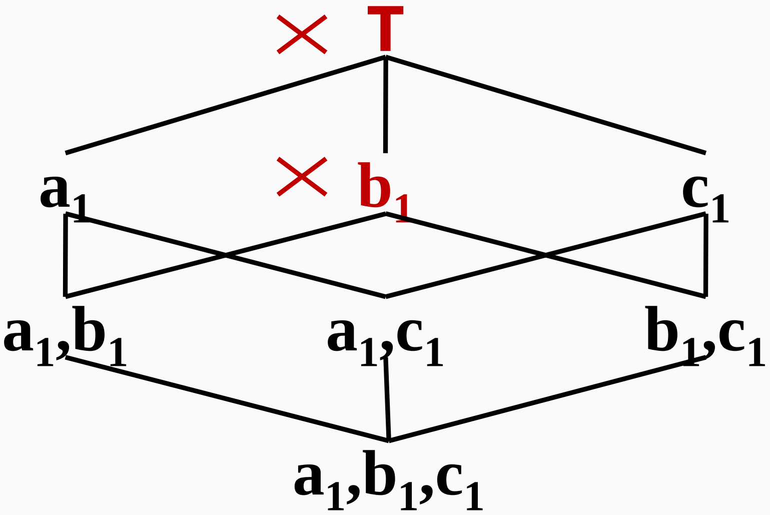
Brute-Force Approach

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



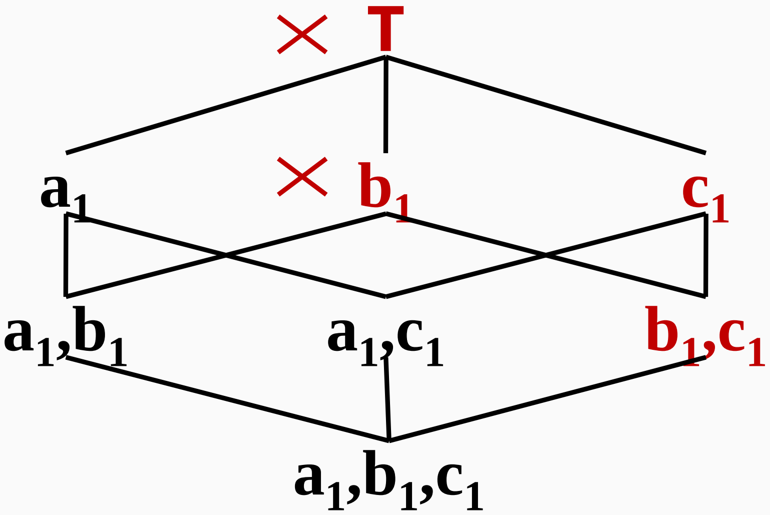
Brute-Force Approach

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



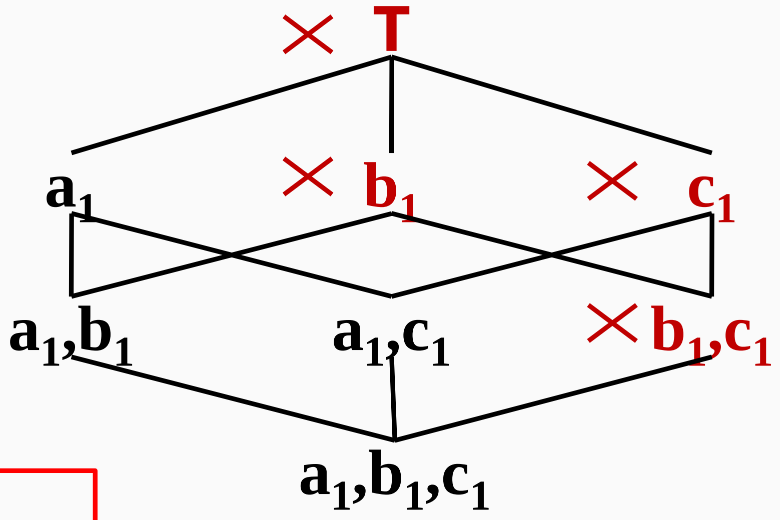
Brute-Force Approach

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Brute-Force Approach

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Total $|R|*(2^{|\mathcal{D}|+|\mathcal{M}|}-1)$ comparisons!

Total 16 comparisons in this case!

Challenges

- Exhaustive comparison with every tuple
- Under every constraint
- Over every measure subspace

Challenges and Ideas

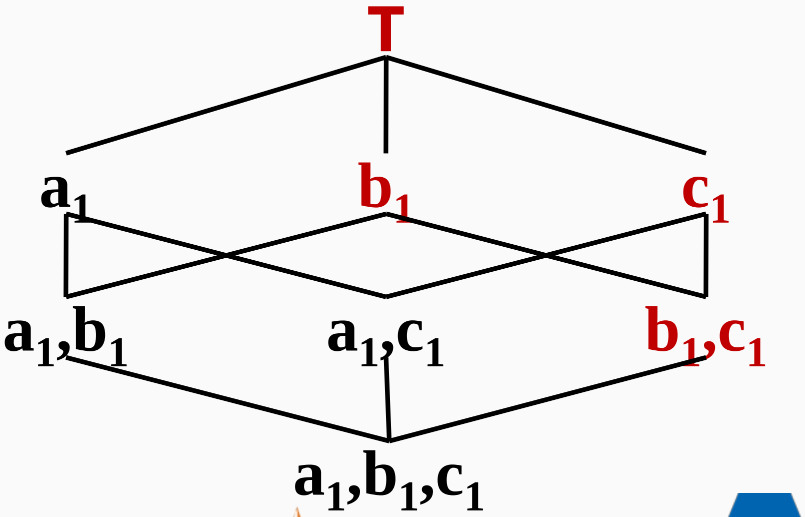
- Exhaustive comparison with every tuple
  - ✓ Tuple reduction
    - Comparison with skyline tuples is enough
    - $t_4 \succ_{\{m_1,m_2\}} t_3 \succ_{\{m_1,m_2\}} t_5 \Rightarrow t_4 \succ_{\{m_1,m_2\}} t_5$

id	d_1	$\boldsymbol{d_2}$	d_3	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5		b_1		11	15

Challenges and Ideas

- Under every constraint
  - ✓Constraint pruning
    - In $C^{t,t'}$, one comparison on t and t' is enough

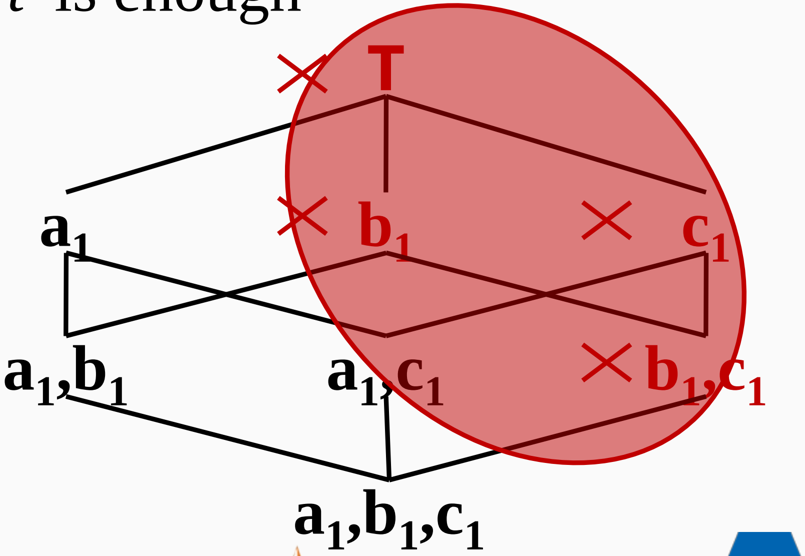
id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Challenges and Ideas

- Under every constraint
  - ✓ Constraint pruning
    - In $C^{t,t'}$, one comparison on t and t' is enough

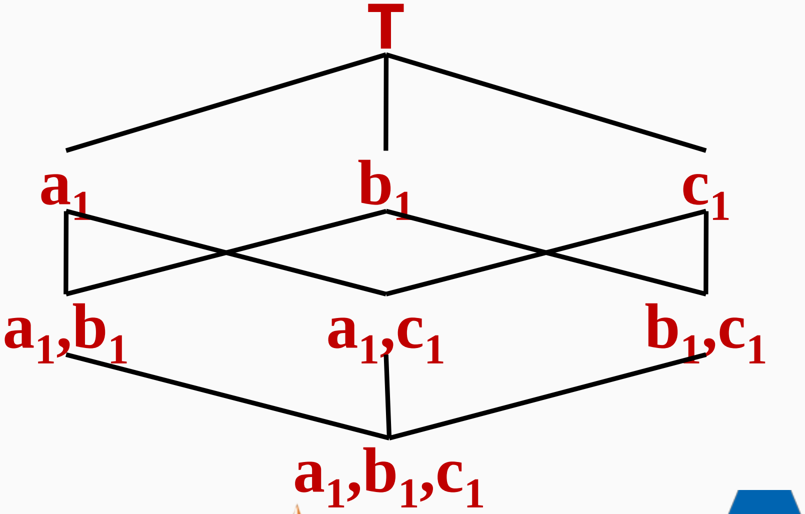
id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Challenges and Ideas

- Over every measure subspace
  - ✓ Sharing computation across measure subspaces
    - Reusing computations on full space in subspaces

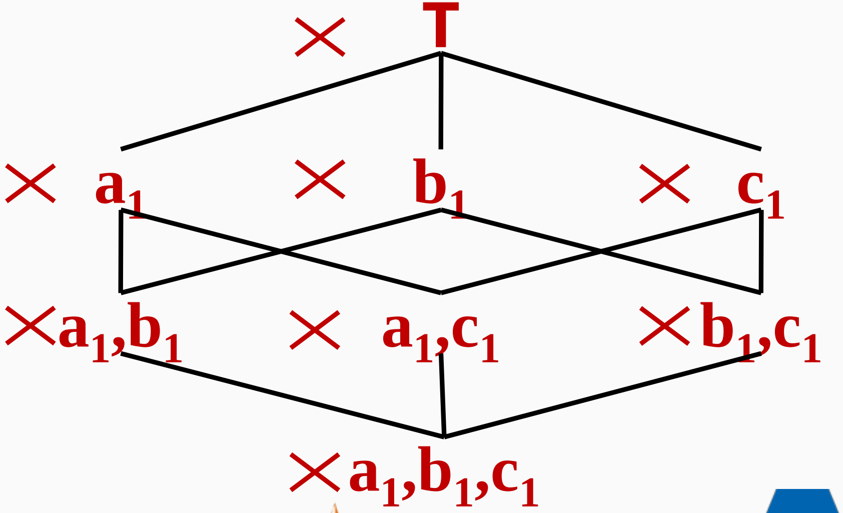
id	***d_1***	***d_2***	***d_3***	***m_1***	***m_2***
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Challenges and Ideas

- Over every measure subspace
  - ✓ Sharing computation across measure subspaces
    - Reusing computations on full space in subspaces

id	*d₁*	*d₂*	*d₃*	*m₁*	*m₂*
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	



Our Algorithms

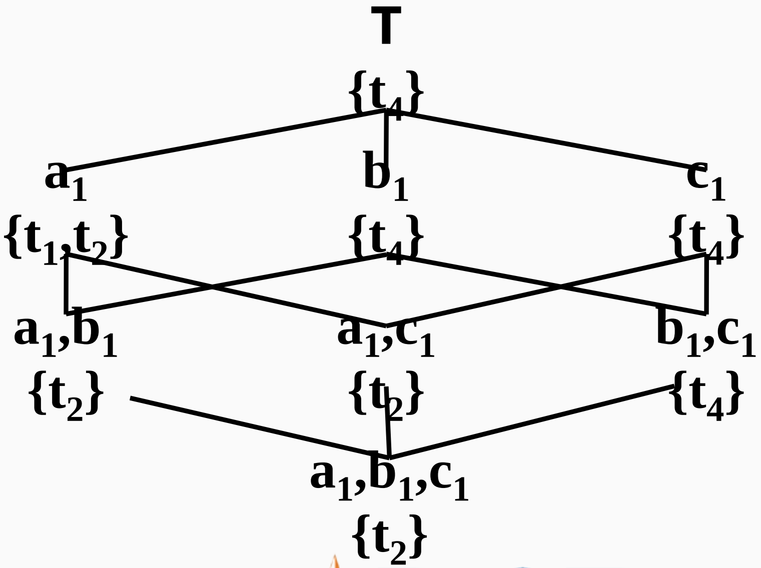
- Tuple reduction + Constraint pruning
  - BottomUp
  - TopDown
- Tuple reduction + Constraint pruning + Sharing computation
  - SBottomUp
  - STopDown

BottomUp

- Stores a tuple for every such constraint that qualifies it as a contextual skyline tuple
- Traverses the constraints in C^t in a bottom-up, breadth-first manner

BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

$\{t_4\}$

a_1 $\{t_1,t_2\}$

b_1 $\{t_4\}$

c_1 $\{t_4\}$

a_1,b_1 $\{t_2\}$

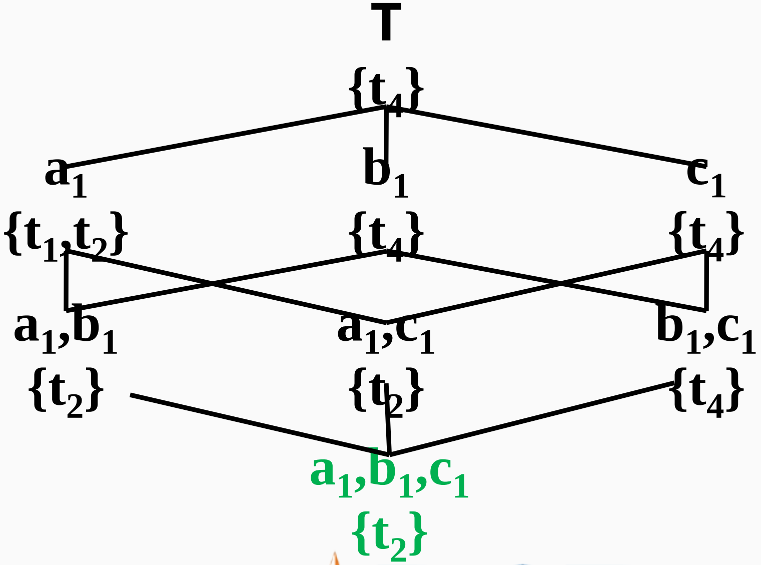
a_1,c_1 $\{t_2\}$

b_1,c_1 $\{t_4\}$

a_1,b_1,c_1 $\{t_2\}$

BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

$\{t_4\}$

a_1

$\{t_1,t_2\}$

b_1

$\{t_4\}$

c_1

$\{t_4\}$

a_1,b_1

$\{t_2\}$

a_1,c_1

$\{t_2\}$

b_1,c_1

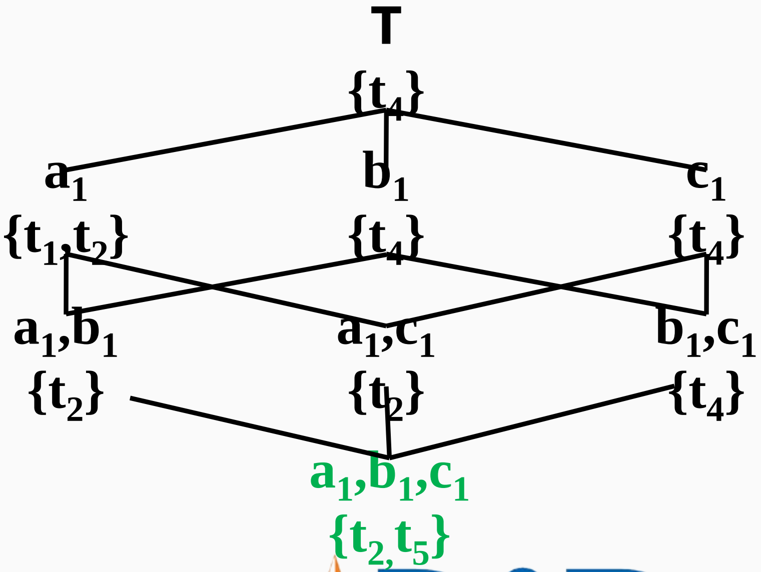
$\{t_4\}$

a_1,b_1,c_1

$\{t_2\}$

BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

$\{t_4\}$

a_1 $\{t_1,t_2\}$

b_1 $\{t_4\}$

c_1 $\{t_4\}$

a_1,b_1 $\{t_2\}$

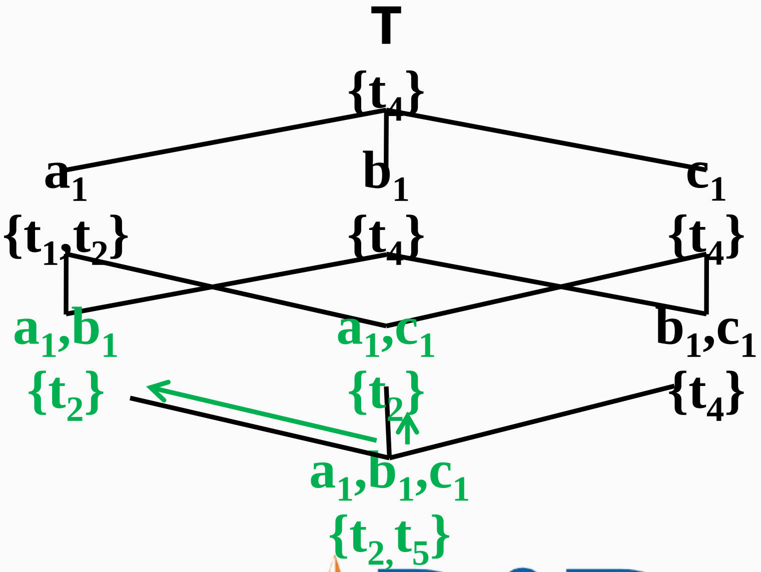
a_1,c_1 $\{t_2\}$

b_1,c_1 $\{t_4\}$

a_1,b_1,c_1 $\{t_2,t_5\}$

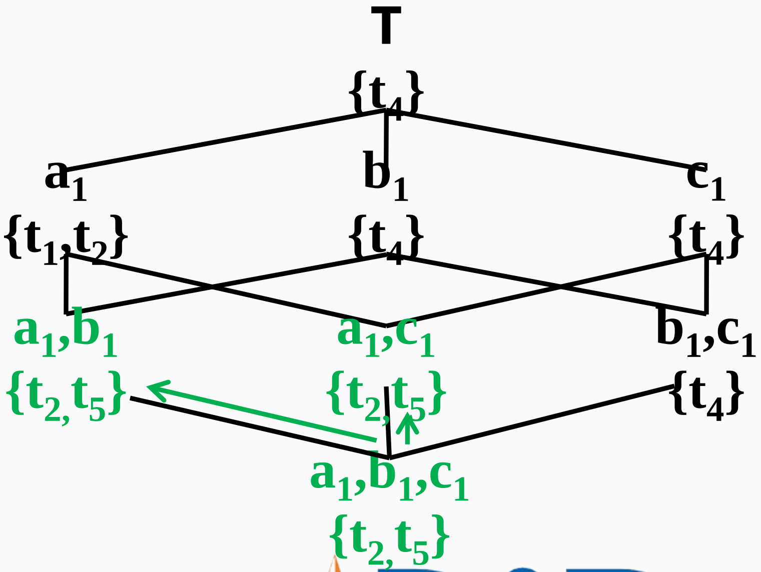
BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



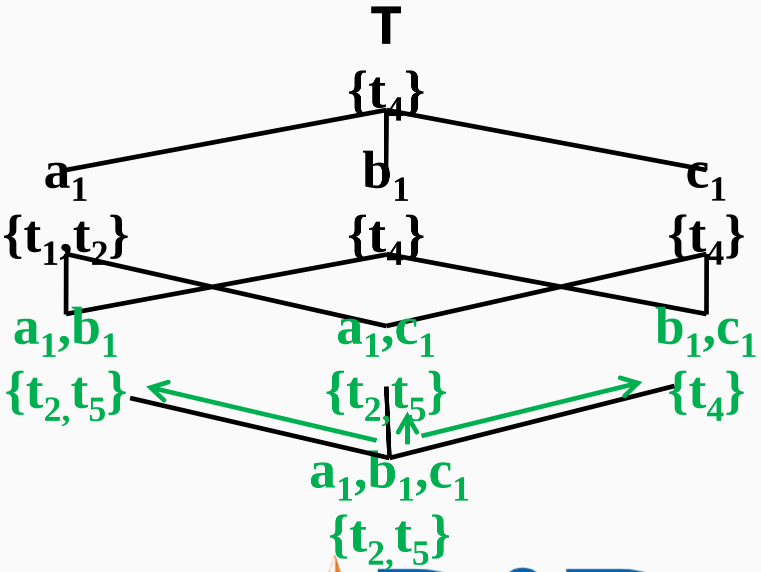
BottomUp

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



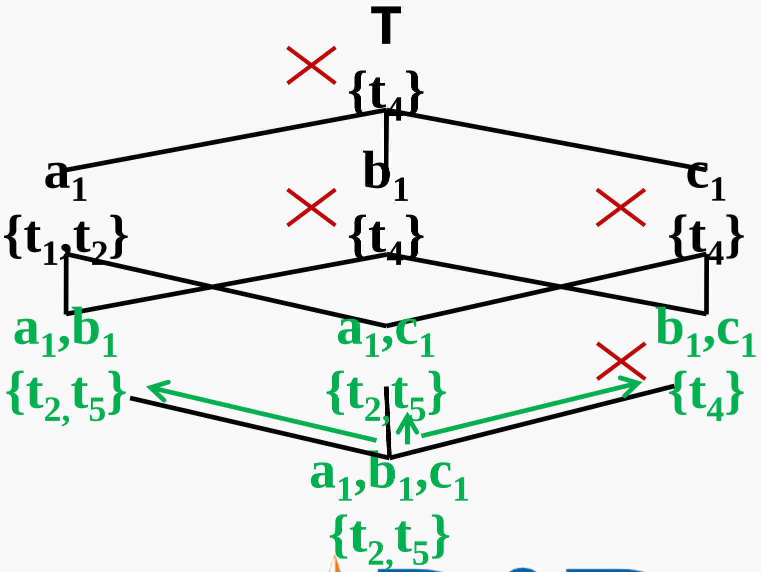
BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5		b_1	c_1	11	15



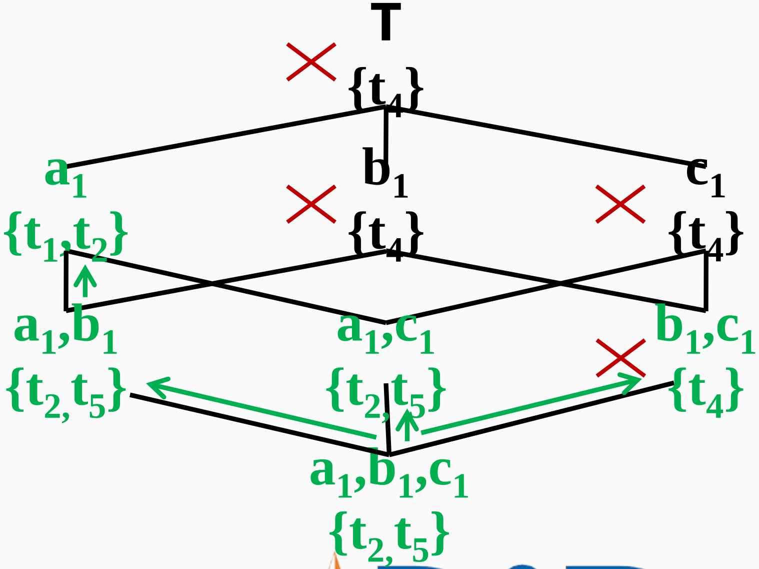
BottomUp

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5		b_1	c_1	11	15



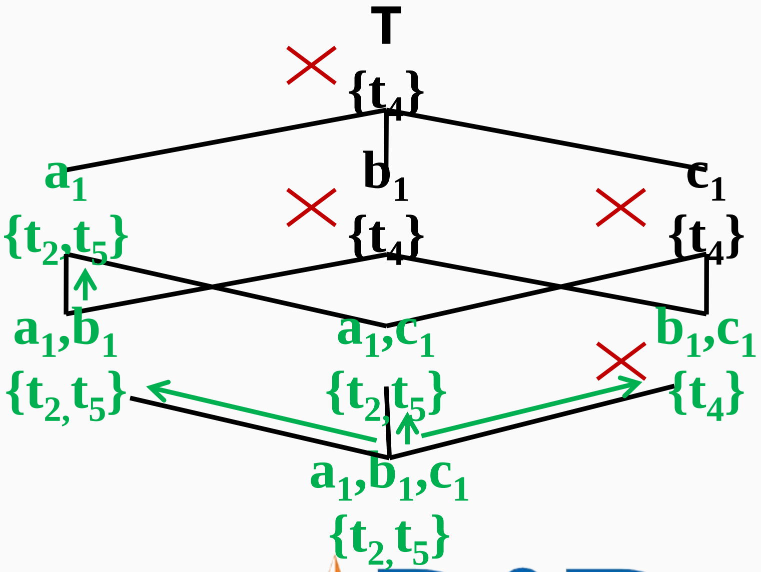
BottomUp

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1			11	15



BottomUp

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	*a_1*	*b_2*	*c_2*	10	15
t_2	*a_1*	*b_1*	*c_1*	15	10
t_3	*a_2*	*b_1*	*c_2*	17	17
t_4	*a_2*	*b_1*	*c_1*	20	20
t_5	*a_1*			11	15



6 comparisons in this case

BottomUp

- Cons of BottomUp
  - Repetitive storage: space complexity
  - Repetitive comparisons: time complexity

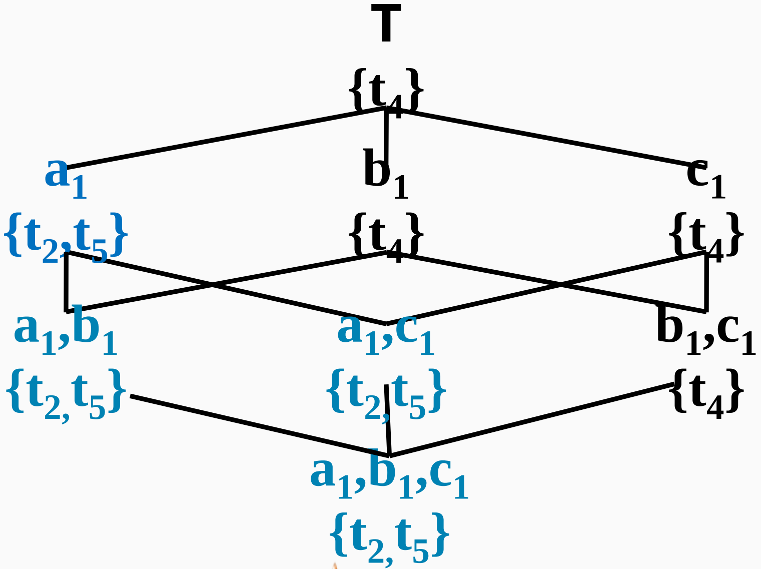
TopDown stores a tuple for its maximal skyline constraints only.

TopDown

Skyline Constraints

Constraints whose contextual skylines include *t*.

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15

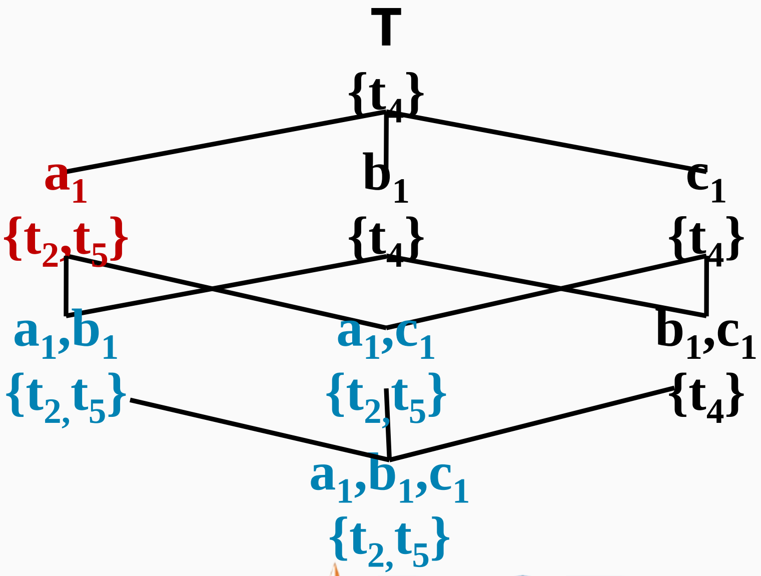


TopDown

Maximal Skyline Constraints

Constraints not subsumed by any other skyline constraints of *t*.

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15

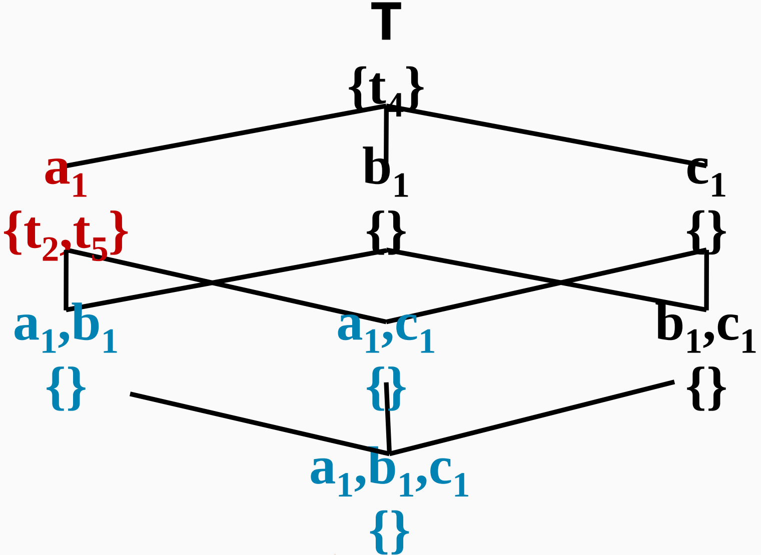


TopDown

Maximal Skyline Constraints

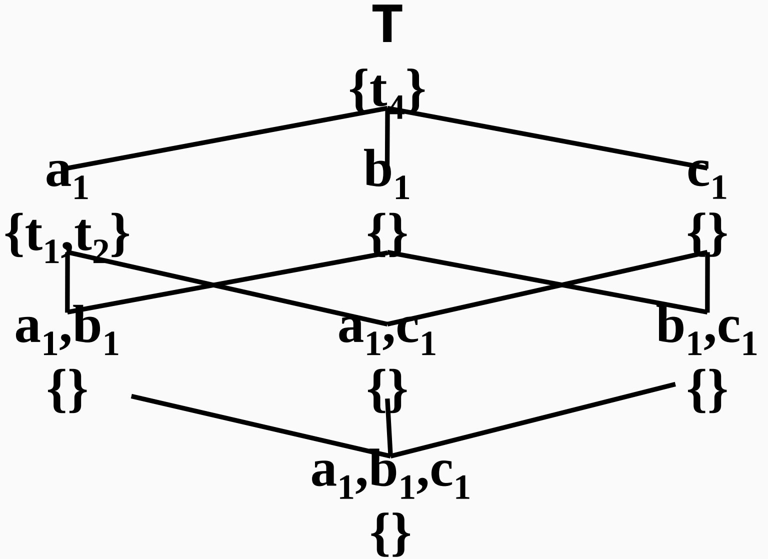
Constraints not subsumed by any other skyline constraints of *t*.

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



TopDown

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

{t_4}

a_1

{t_1,t_2}

b_1

{}

c_1

{}

a_1,b_1

{}

a_1,c_1

{}

b_1,c_1

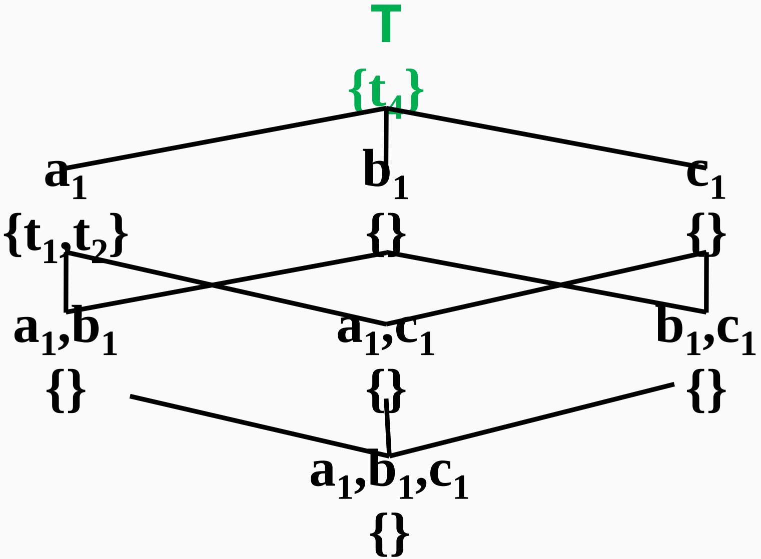
{}

a_1,b_1,c_1

{}

TopDown

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



T

$\{t_4\}$

a_1

$\{t_1,t_2\}$

b_1

$\{\}$

c_1

$\{\}$

a_1,b_1

$\{\}$

a_1,c_1

$\{\}$

b_1,c_1

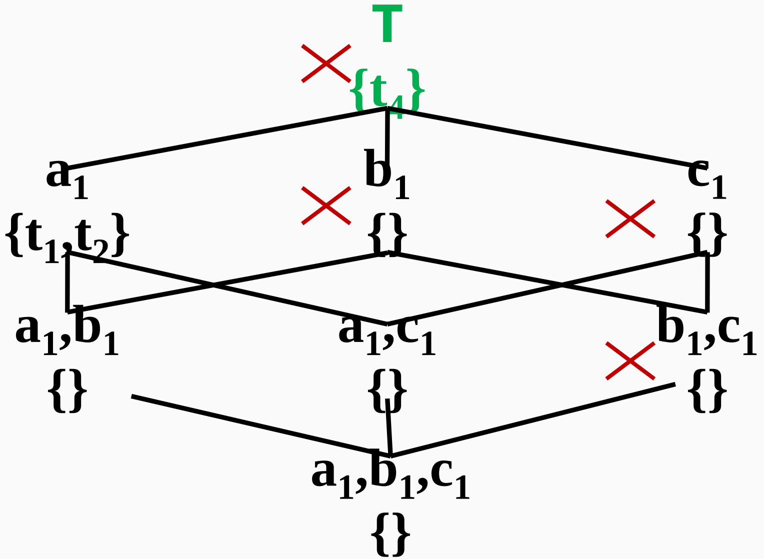
$\{\}$

a_1,b_1,c_1

$\{\}$

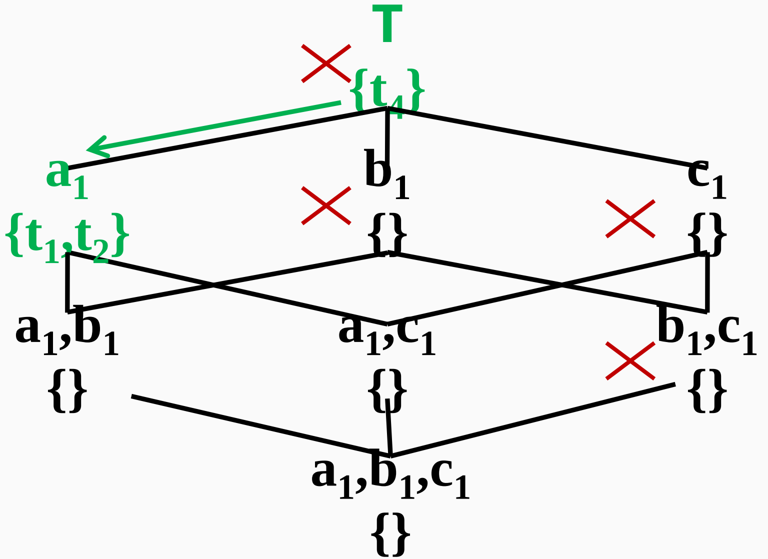
TopDown

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



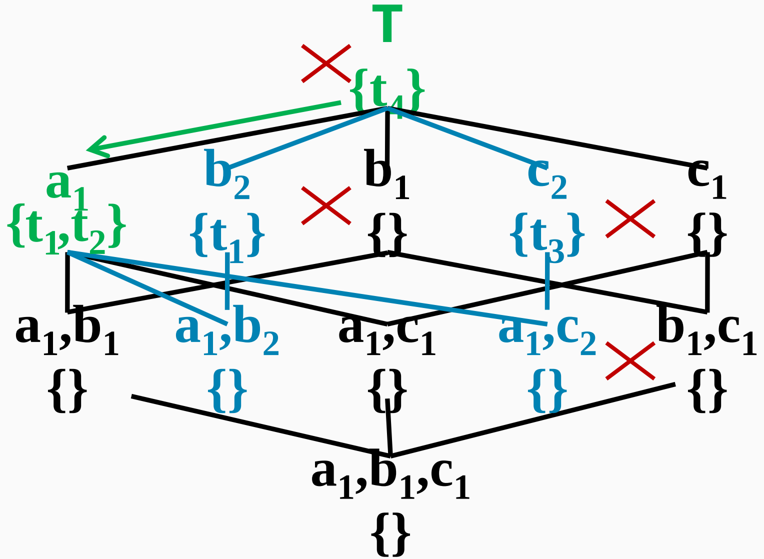
TopDown

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	*a_1*	*b_2*	*c_2*	10	15
t_2	*a_1*	*b_1*	*c_1*	15	10
t_3	*a_2*	*b_1*	*c_2*	17	17
t_4	*a_2*	*b_1*	*c_1*	20	20
t_5	*a_1*			11	15



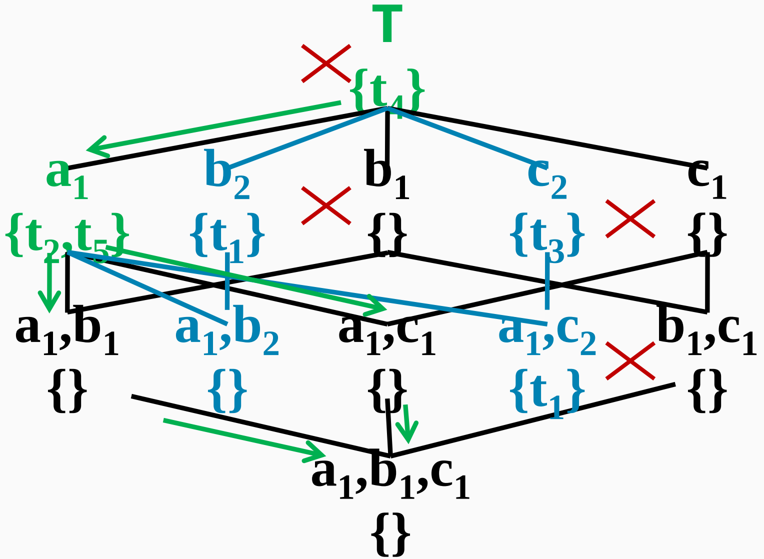
TopDown

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	*a_1*	*b_2*	*c_2*	10	15
t_2	*a_1*	*b_1*	*c_1*	15	10
t_3	*a_2*	*b_1*	*c_2*	17	17
t_4	*a_2*	*b_1*	*c_1*	20	20
t_5	*a_1*			11	15



TopDown

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1			11	15



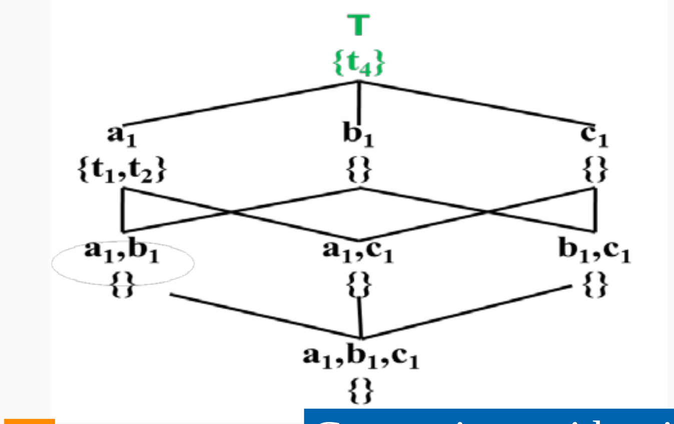
3 comparisons in this case

STopDown and SBottomUp

- Con of BottomUp and TopDown
  - Need to compute over every measure subspace separately
    - STopDown and SBottomUp share computation across different subspaces

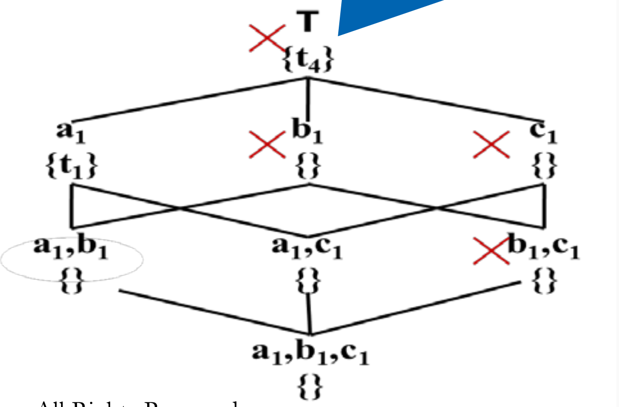
STopDown

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



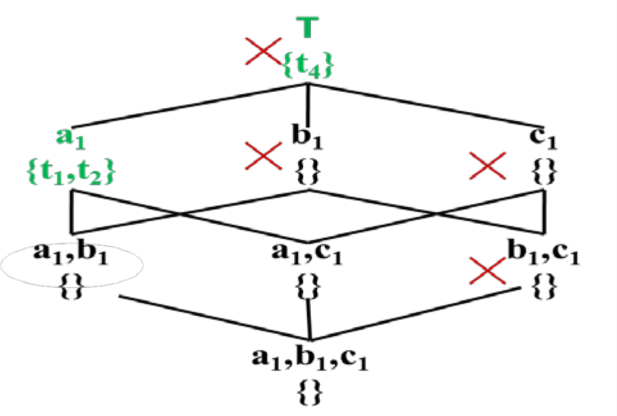
id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1		15

Comparison with t_4 is skipped



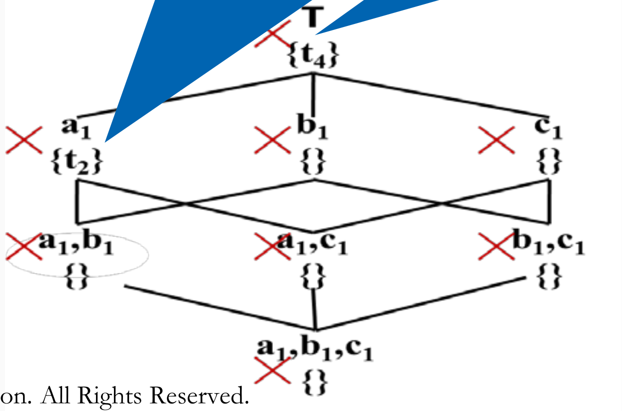
STopDown

id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	*a_1*	*b_2*	*c_2*	10	15
t_2	*a_1*	*b_1*	*c_1*	15	10
t_3	*a_2*	*b_1*	*c_2*	17	17
t_4	*a_2*	*b_1*	*c_1*	20	20
t_5	*a_1*			11	15



Comparisons with t_2 & t_4 are skipped

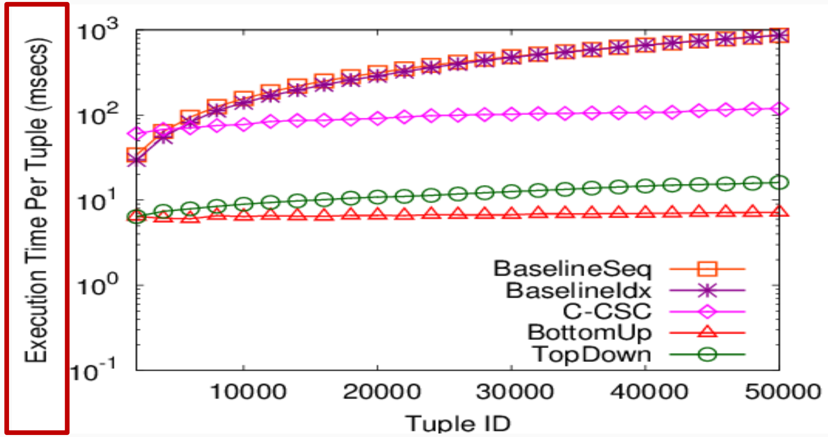
id	*d_1*	*d_2*	*d_3*	*m_1*	*m_2*
t_1	*a_1*	*b_2*	*c_2*	10	15
t_2	*a_1*	*b_1*	*c_1*	15	10
t_3	*a_2*	*b_1*	*c_2*	17	17
t_4	*a_2*	*b_1*	*c_1*	20	20
t_5	*a_1*	*b_1*	*c_1*	11	



Experiment Setup

- ❑ NBA Dataset
  - 317,371 tuples of NBA box scores from 1991-2004 seasons
  - 8 dimension attributes
  - 7 measure attributes
- ❑ Weather Dataset
  - 7.8 million tuples of weather forecast from different locations of six countries & regions of UK
  - 7 dimension attributes
  - 7 measure attributes

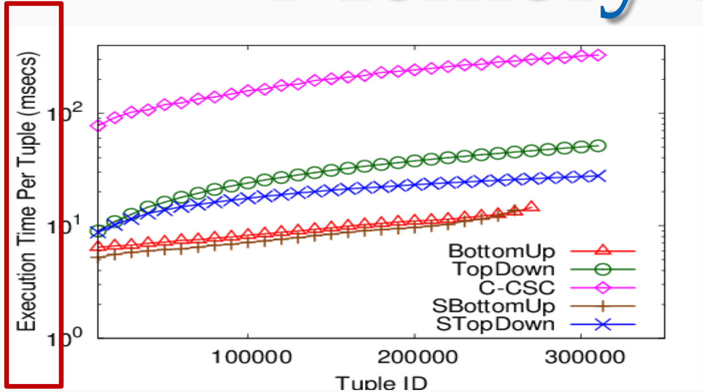
Memory-Based Implementation



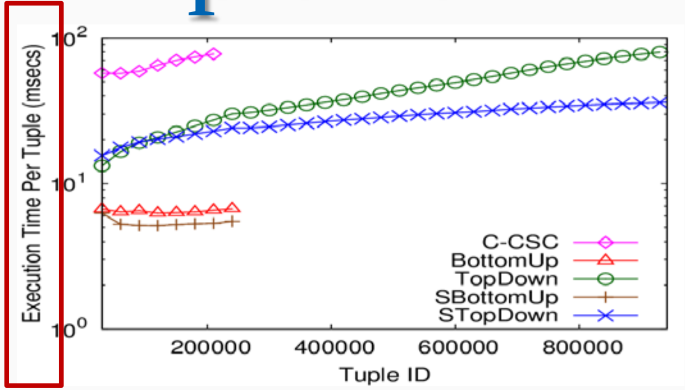
NBA Dataset

- ❑ Maintaining CSC for each constraint causes overhead (Xia et al. SIGMOD 2006)
  - Can’t take advantage of constraint pruning

Memory-Based Implementation



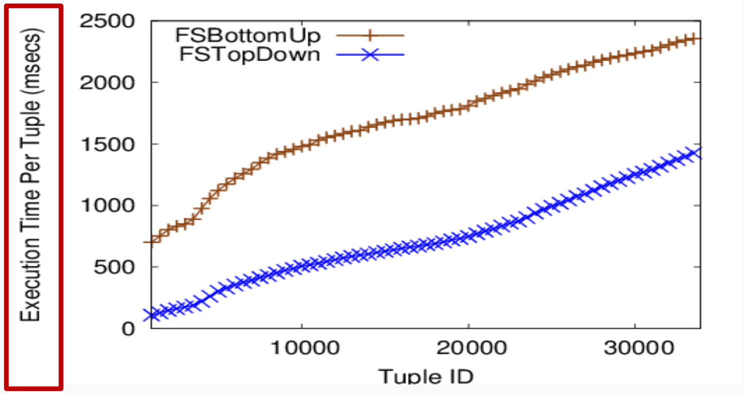
NBA Dataset



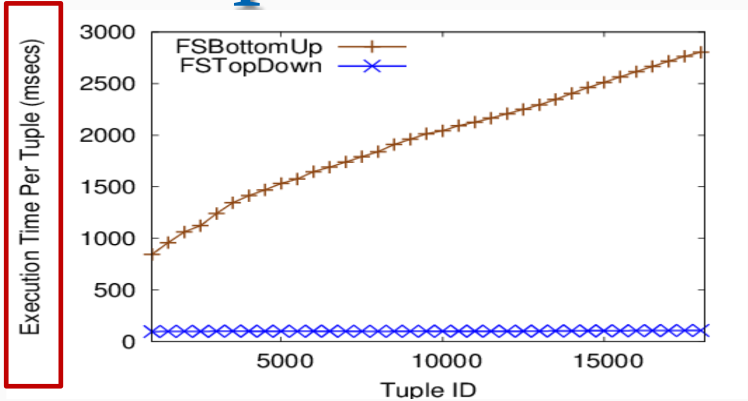
Weather Dataset

- ❑ BottomUp/SBottomUp exhausted available JVM heap
  - ▪ memory overflow
- ❑ TopDown / STopDown was outperformed by BottomUp/ SBottomUp
  - ▪ Updating maximal skyline constraints causes overhead

File-Based Implementation



NBA Dataset



Weather Dataset

- ❑ Each (*C*,*M*) is stored in a binary file
- ❑ While traversing, file-read operation occurs if file is non-empty: FSTopDown encounters many empty files
- ❑ For updating, file-write operation occurs: FSTopDown stores fewer tuples
- ❑ I/O-cost dominates in-memory computation

Discovered Facts

- Lamar Odom had 30 points, 19 rebounds and 11 assists on March 6, 2004. No one before had a better or equal performance in NBA history.
- Allen Iverson had 38 points and 16 assists on April 14, 2004 to become the first player with a 38/16 (points/assists) game in the 2004-2005 season.
- Damon Stoudamire scored 54 points on January 14, 2005. It is the highest score in history made by any Trail Blazers.

Prominent Streak Discovery in Sequence Data. Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, Yong Yu. KDD 2011, pages 1280-1288.

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.

Prominent Streaks

Prominent streaks stated in news articles:

“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.”

“The Nikkei 225 closed below 10000 for the 12th consecutive week, the longest such streak since June 2009.”

“He (LeBron James) scored 35 or more points in nine consecutive games and joined Michael Jordan and Kobe Bryant as the only players since 1970 to accomplish the feat.”

Concepts

Streak

Input: a sequence of values

Streak <[l, r], v> is a triple: left-end (l), right-end (r), minimum value in interval [l,r]

3 1 7 7 2 $\underline{5\ \ 4\ \ 6}$ 7 3

<[6, 8], 4>

Streak dominance relation

s1=<[l1, r1], v1> dominates s2=<[l2, r2], v2> iff

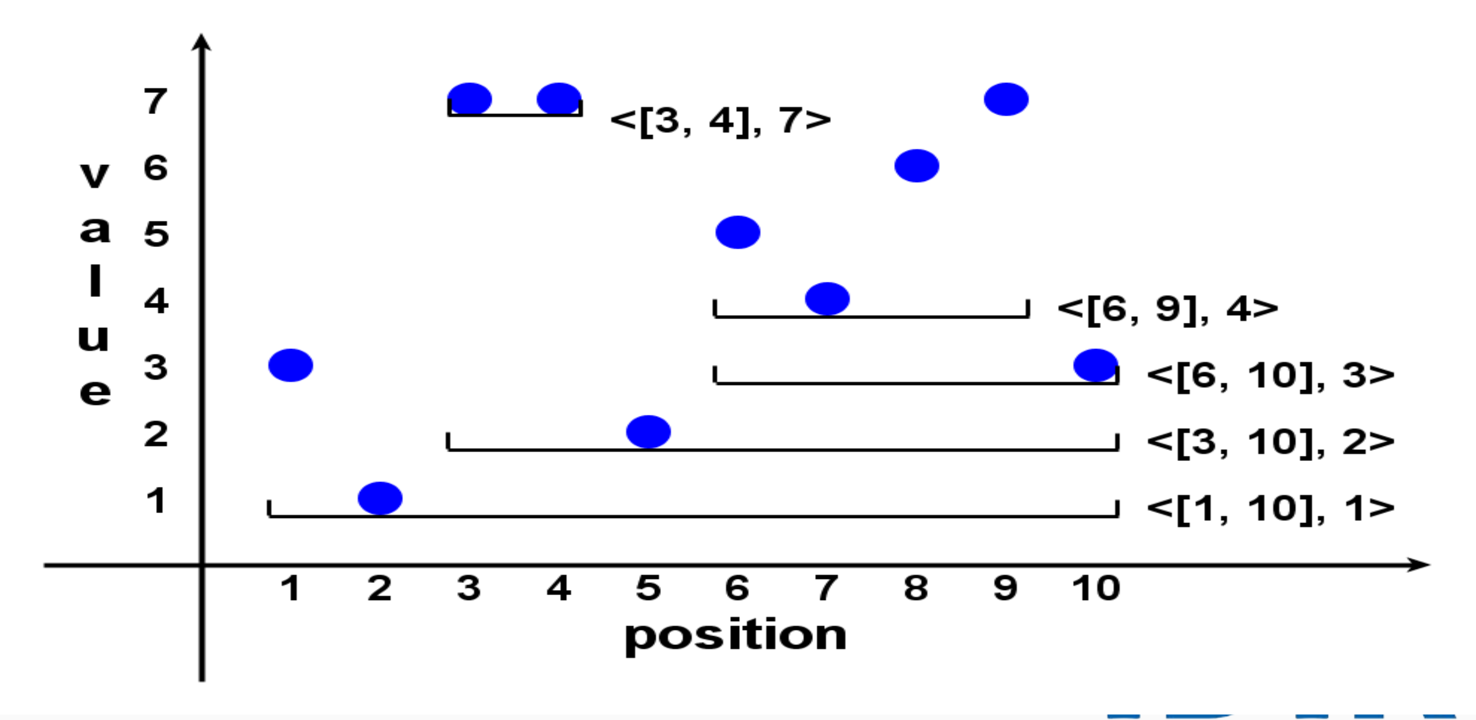
r1 - l1 > r2 - l2, v1 >= v2 or r1 - l1 >= r2 - l2, v1 >v2

Prominent streaks (PS)

A streak is prominent if it is not dominated by any other streaks.

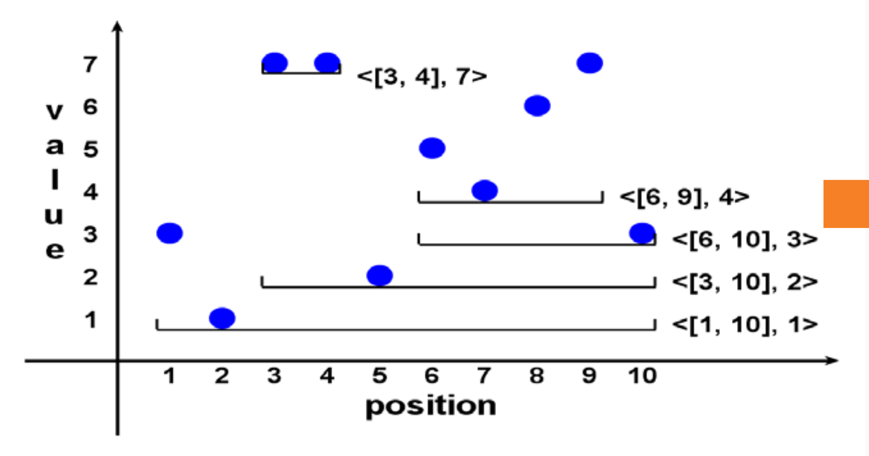
Example

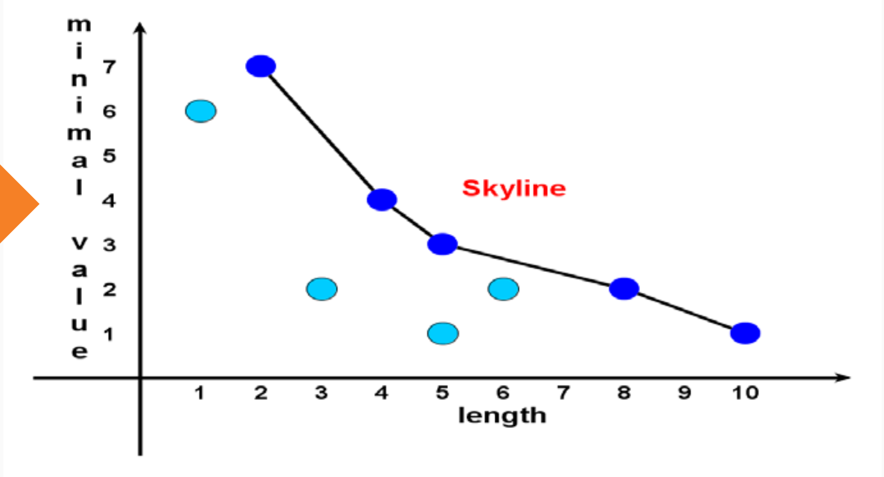
3 1 7 7 2 5 4 6 7 3



Prominent Streaks are Skyline Points in 2-d Space

3 1 7 7 2 5 4 6 7 3





Tasks

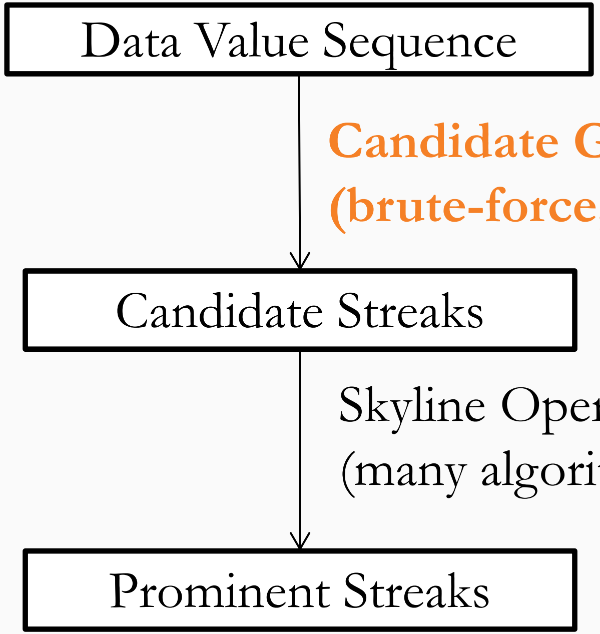
Task 1: discovery

Find all prominent streaks in a sequence

Task 2: monitoring

Always keep prominent streaks up-to-date, when sequence grows (real-world sequences often grow)

Solution Framework



Data Value Sequence

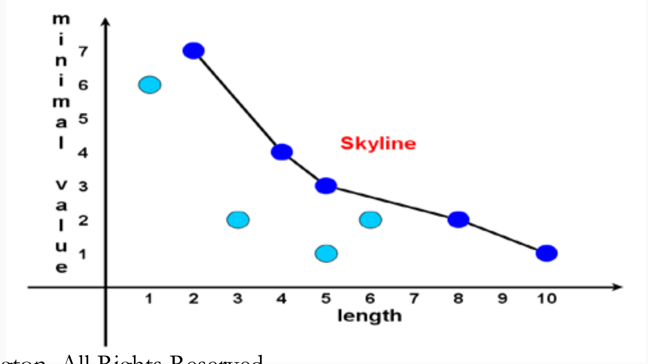
3 1 7 7 2 5 4 6 7 3

Candidate Generation Algorithms (brute-force, NLPS, LLPS)

Candidate Streaks

Skyline Operation [Börzsönyi et al. 2001] (many algorithms)

Prominent Streaks



Candidate Generation: Number Of Candidates

Brute-force

Quadratic

NLPS

Superlinear

LLPS

Linear

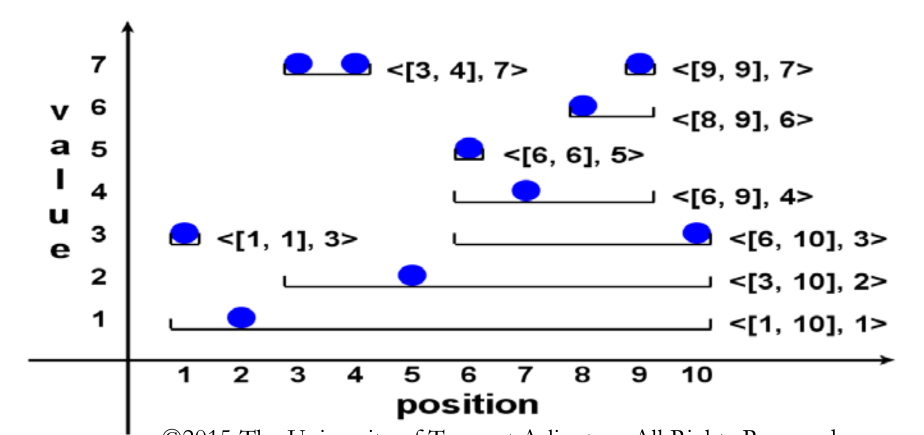
Local Prominent Streak

Local dominance relation

s1=<[l1, r1], v1> locally dominates s2=<[l2, r2], v2> iff
s1 dominates s2 and [l1, r1] ⊃ [l2, r2]

Local prominent streak (LPS)

A streak is locally prominent if it is not locally dominated by any other streaks.



Important Properties

(1) LPS is sufficient

A prominent streak must be an LPS.

(2) LPS is small

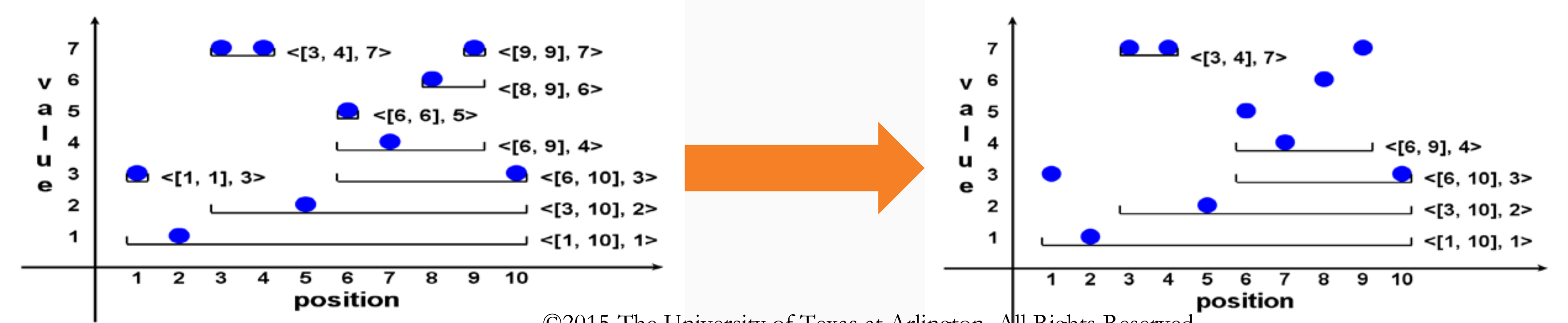
The number of LPSs is less than or equal to the sequence length.

(Hint: The number of LPSs getting min value at position k is at most 1.)

Conclusion

LPS is an excellent set of candidate streaks, of linear size.

Candidate generation problem => finding local prominent streaks



Linear LPS (LLPS) Method

Sequence $p_1, p_2, \ldots, p_n$.

1. Maintain a list of candidate streaks when scanning the sequence rightward.

2. After p_k, right-ends of candidates are all k.

3. At p_{k+1}, try to extend the candidates rightward.

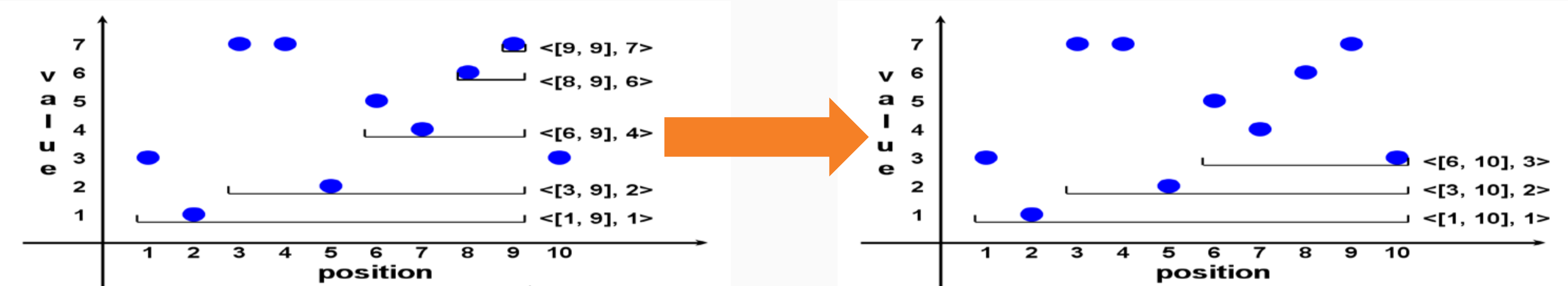
Candidates s:

(3.a) s.v < p_{k+1} : extend.

(3.b) s.v > p_{k+1} : belong to LPS.

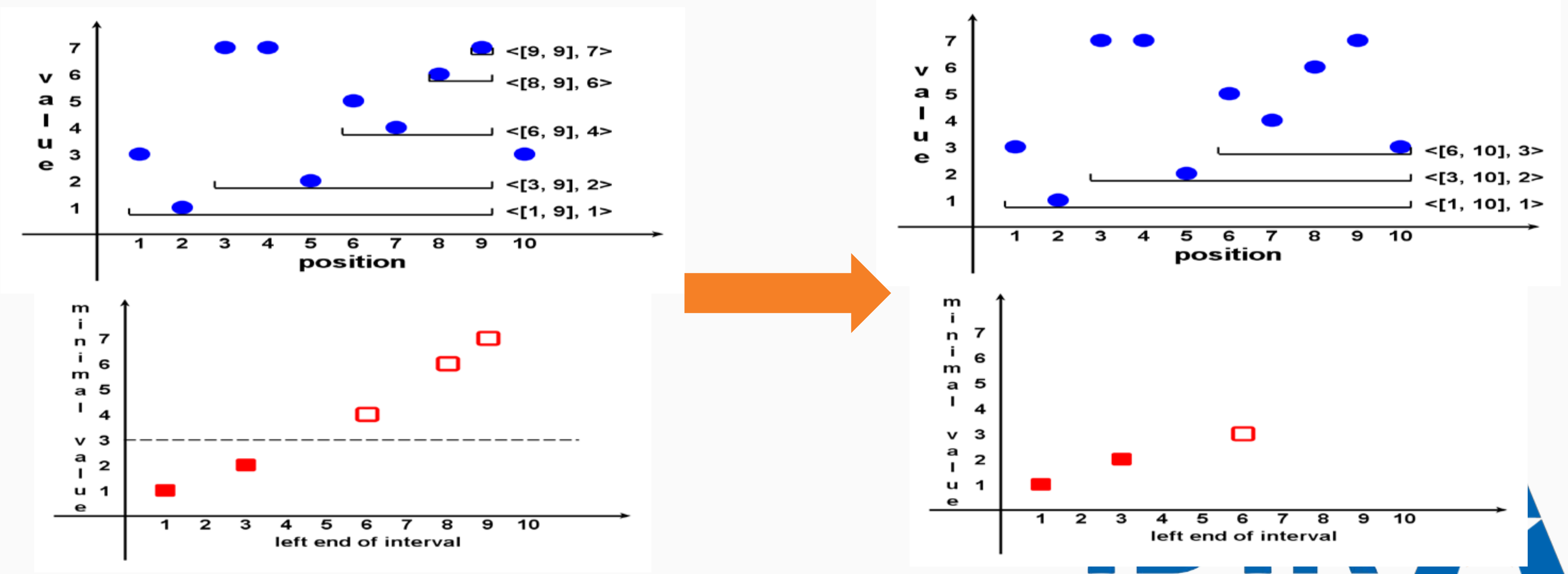
(3.c) s.v >= p_{k+1} : extend the leftmost (longest) such s.

4. After p_n all remaining candidates are LPS.



Linear LPS (LLPS) Method

Candidates share the same right-end, their minimum values monotonically increase, if they are listed in the increasing order of left-ends.



Linear LPS (LLPS) Method

After p_k , it has found:

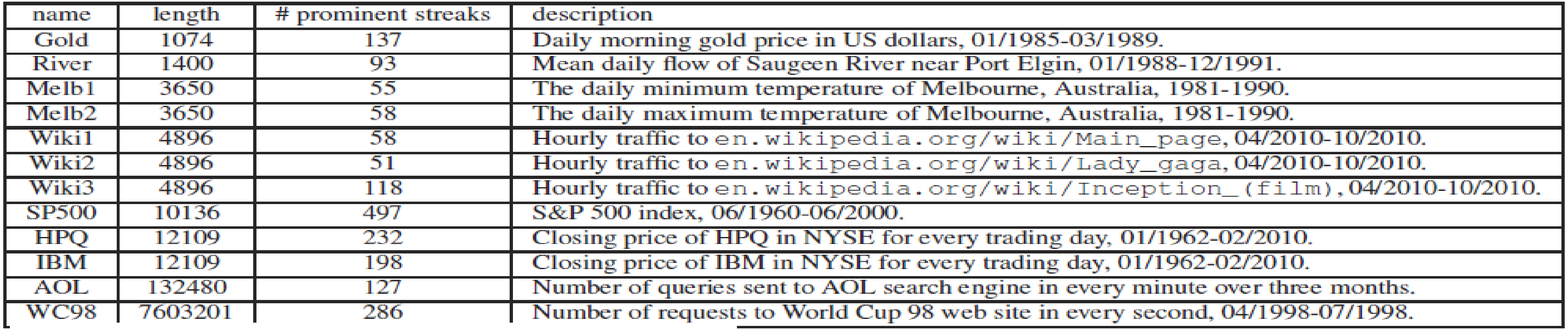
All LPSs ending before k

Candidates ending at k either are LPSs or can be grown to LPSs ending after k.

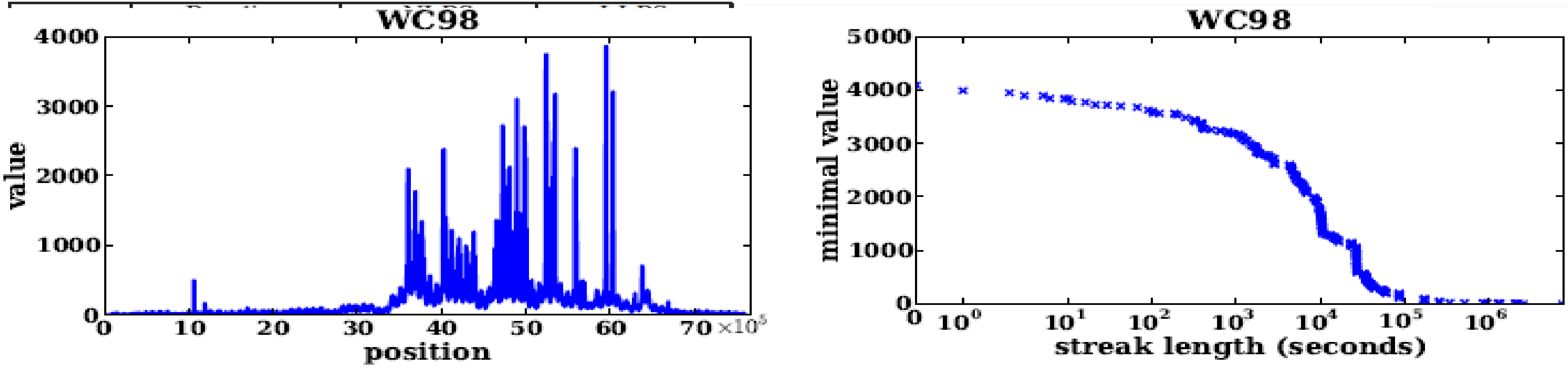
Monitoring (keeping prominent streaks up-to-date) is simple:

If PSs till k are requested, compare all found LPSs and all remaining candidates.

Datasets In Experiments



name	length	# prominent streaks	description
Gold	1074	137	Daily morning gold price in US dollars, 01/1985-03/1989.
River	1400	93	Mean daily flow of Saugeen River near Port Elgin, 01/1988-12/1991.
Melb1	3650	55	The daily minimum temperature of Melbourne, Australia, 1981-1990.
Melb2	3650	58	The daily maximum temperature of Melbourne, Australia, 1981-1990.
Wiki1	4896	58	Hourly traffic to `en.wikipedia.org/wiki/Main_page`, 04/2010-10/2010.
Wiki2	4896	51	Hourly traffic to `en.wikipedia.org/wiki/Lady_gaga`, 04/2010-10/2010.
Wiki3	4896	118	Hourly traffic to `en.wikipedia.org/wiki/Inception_(film)`, 04/2010-10/2010.
SP500	10136	497	S&P 500 index, 06/1960-06/2000.
HPQ	12109	232	Closing price of HPQ in NYSE for every trading day, 01/1962-02/2010.
IBM	12109	198	Closing price of IBM in NYSE for every trading day, 01/1962-02/2010.
AOL	132480	127	Number of queries sent to AOL search engine in every minute over three months.
WC98	7603201	286	Number of requests to World Cup 98 web site in every second, 04/1998-07/1998.



(a) Data Sequence

(b) Prominent Streaks

Sample Prominent Streaks

Melbourne daily min/max temperature between 1981 and 1990 (Melb1 & Melb2)

More than 2000 days with min temperature above zero

6 days: the longest streak above 35 degrees Celsius

Traffic count of Wikipedia page of Lady Gaga (Wiki2)

More than half of the prominent streaks are around Sep. 12th (VMA 2010)

at least 2000 hourly visits lasting for almost 4 days

General Prominent Streaks

Top-k, multi-dimensional and multi-sequence PS

“He (LeBron James) scored 35 or more points in nine consecutive games and joined Michael Jordan and Kobe Bryant as the only players since 1970 to accomplish the feat.”

“Only player in NBA history to average at least 20 points, 10 rebounds and 5 assists per game for 6 consecutive seasons.” (http://en.wikipedia.org/wiki/Kevin Garnett)

NLPS/LLPS extended to such general PSs

Experiments On Multi-Sequence PSs

Tab

nu

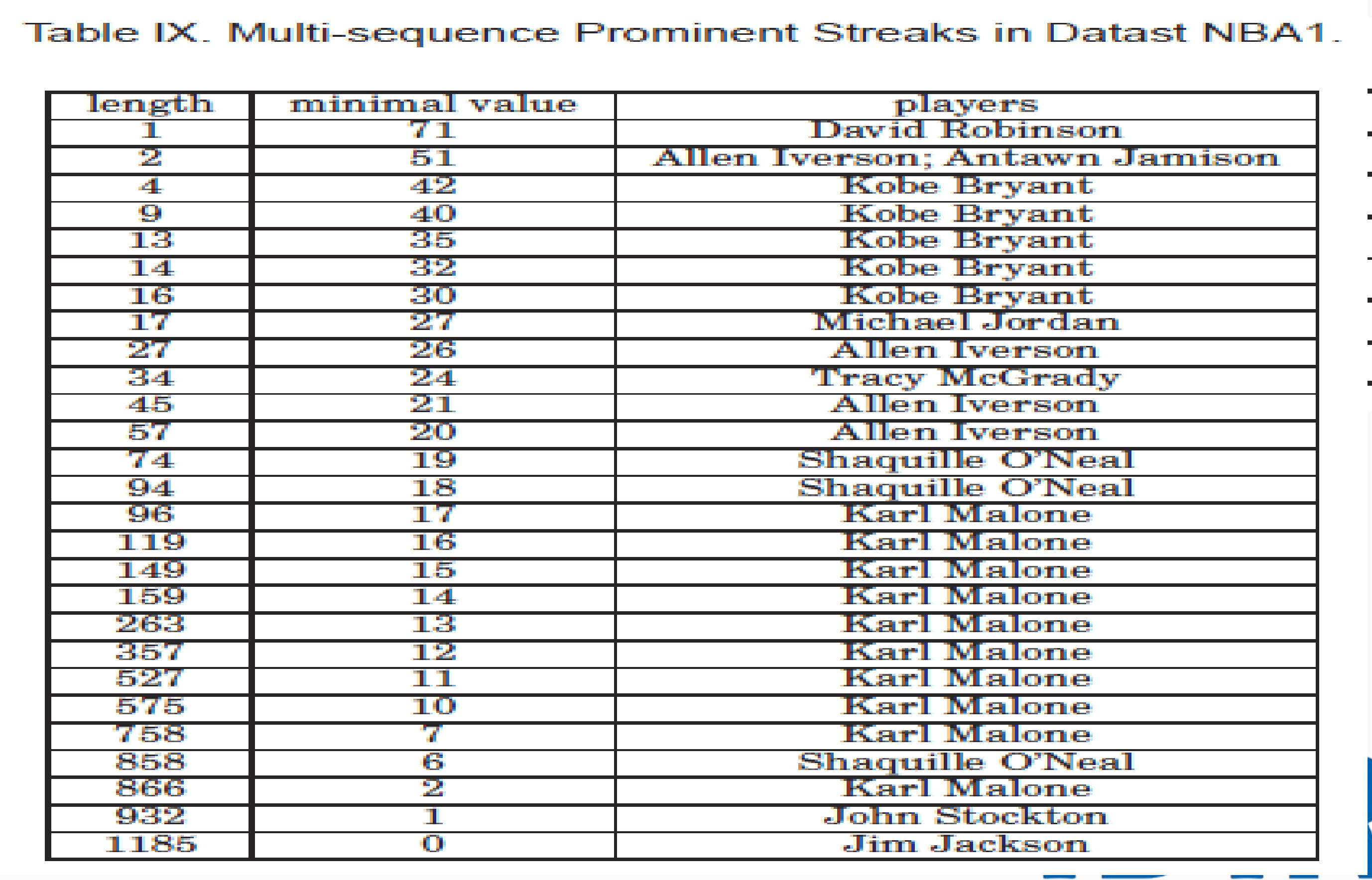


Table IX. Multi-sequence Prominent Streaks in Datast NBA1.

length	minimal value	players
1	71	David Robinson
2	51	Allen Iverson; Antawn Jamison
4	42	Kobe Bryant
9	40	Kobe Bryant
13	35	Kobe Bryant
14	32	Kobe Bryant
16	30	Kobe Bryant
17	27	Michael Jordan
27	26	Allen Iverson
34	24	Tracy McGrady
45	21	Allen Iverson
57	20	Allen Iverson
74	19	Shaquille O'Neal
94	18	Shaquille O'Neal
96	17	Karl Malone
119	16	Karl Malone
149	15	Karl Malone
159	14	Karl Malone
263	13	Karl Malone
357	12	Karl Malone
527	11	Karl Malone
575	10	Karl Malone
758	7	Karl Malone
858	6	Shaquille O'Neal
866	2	Karl Malone
932	1	John Stockton
1185	0	Jim Jackson

Experiments On Multi-Dim PSs

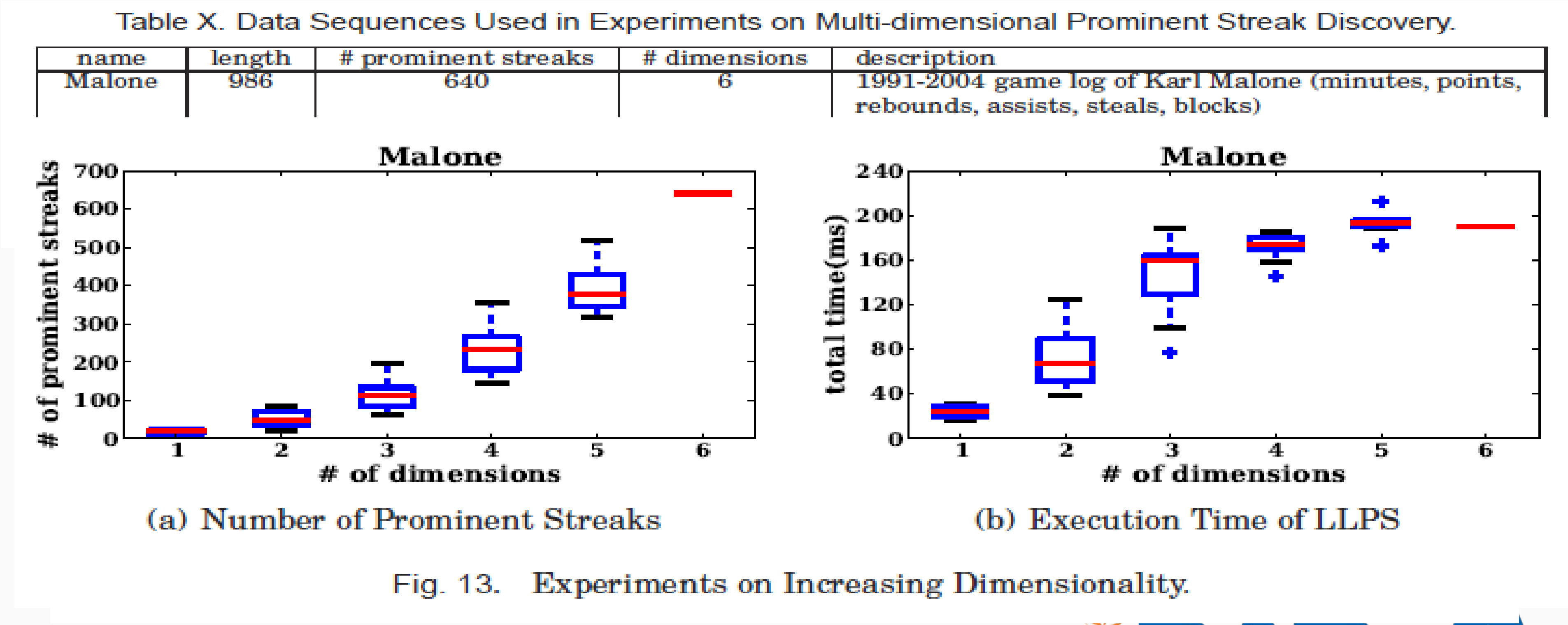


Table X. Data Sequences Used in Experiments on Multi-dimensional Prominent Streak Discovery.

name	length	# prominent streaks	# dimensions	description
Malone	986	640	6	1991-2004 game log of Karl Malone (minutes, points, rebounds, assists, steals, blocks)

(a) Number of Prominent Streaks

(b) Execution Time of LLPS

Fig. 13. Experiments on Increasing Dimensionality.

Experiments On General PSs

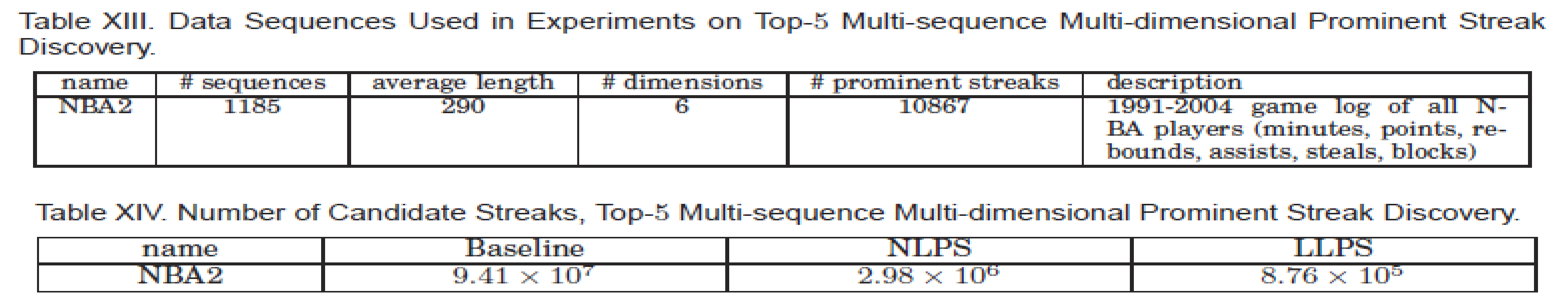


Table XIII. Data Sequences Used in Experiments on Top-5 Multi-sequence Multi-dimensional Prominent Streak Discovery.

name	# sequences	average length	# dimensions	# prominent streaks	description
NBA2	1185	290	6	10867	1991-2004 game log of all NBA players (minutes, points, rebounds, assists, steals, blocks)

Table XIV. Number of Candidate Streaks, Top-5 Multi-sequence Multi-dimensional Prominent Streak Discovery.

name	Baseline	NLPS	LLPS
NBA2	9.41×10^7	2.98×10^6	8.76×10^5

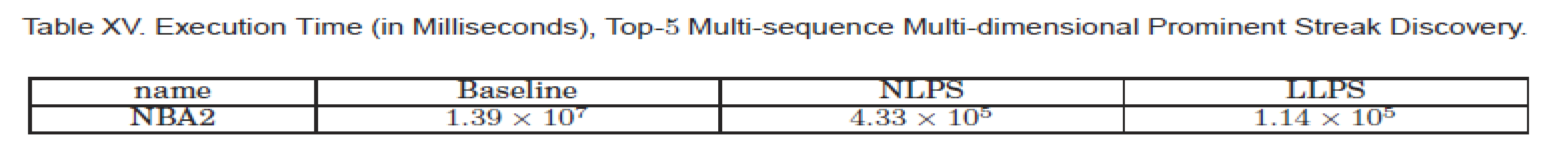


Table XV. Execution Time (in Milliseconds), Top-5 Multi-sequence Multi-dimensional Prominent Streak Discovery.

name	Baseline	NLPS	LLPS
NBA2	1.39×10^7	4.33×10^5	1.14×10^5

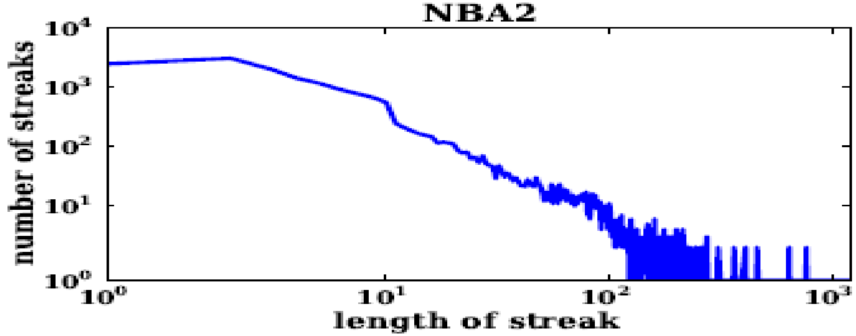


Fig. 14. Distribution of Prominent Streaks by Length.

Acknowledgment

UTA Students

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

Collaborators

- Bill Adair
- Pankaj Agarwal
- James Hamilton
- Ping Luo
- Mark Tremayne
- Min Wang
- Jun Yang
- Cong Yu

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

- Demos

ClaimBuster   idir.uta.edu/claimbuster

FactWatcher   idir.uta.edu/factwatcher

- Please help us label the data

http://bit.ly/claimbusters