Data In, Facts Out:

Automated Monitoring of Facts by FactWatcher

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

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
…	…

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

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.

Fact Finding

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

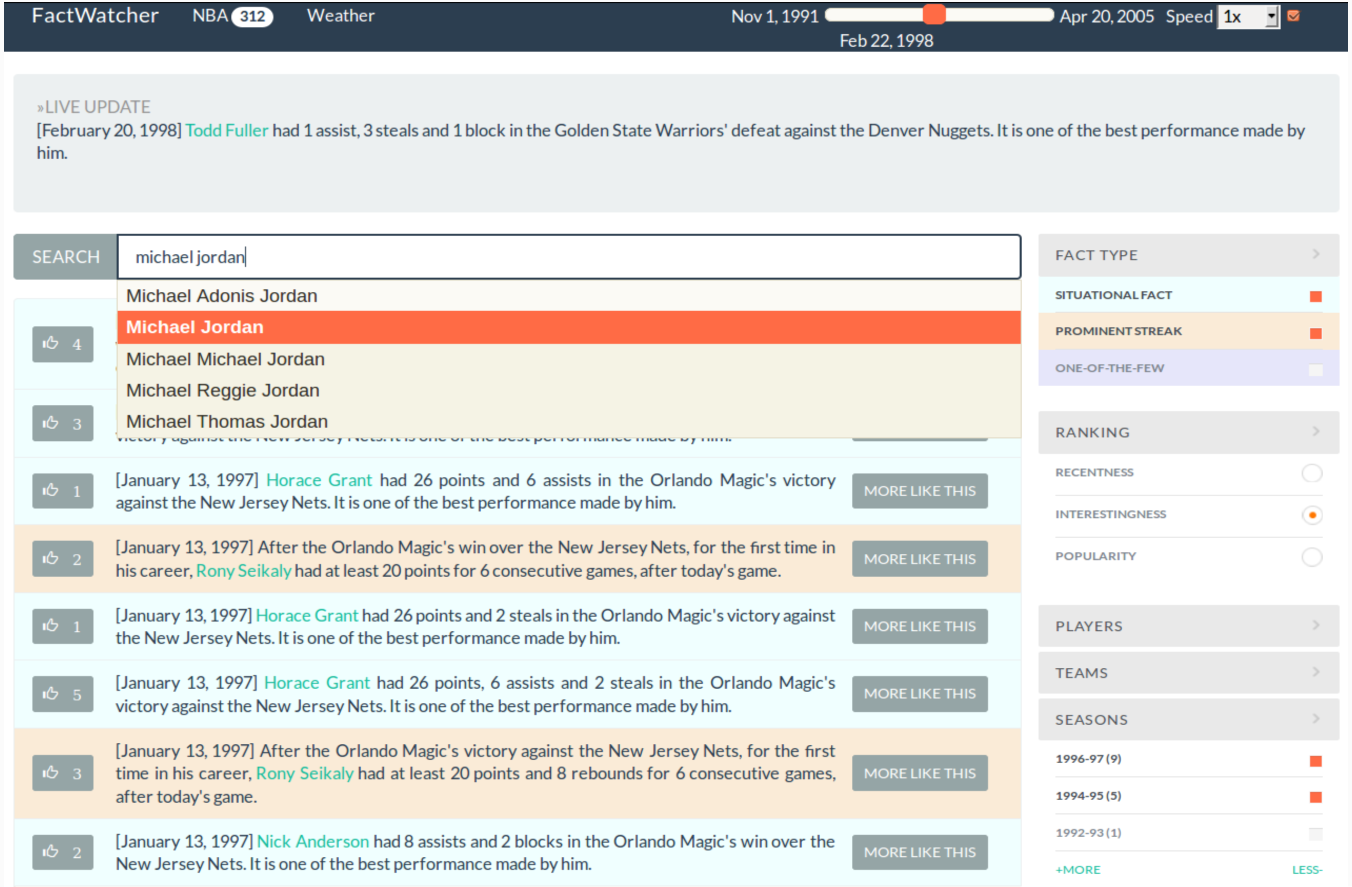
- "Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992."
  http://espn.go.com/espn/elias?date=20130205

- "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

FactWatcher Demo

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

https://vimeo.com/user48311227



Presented In

VLDB 2014

Excellent Demo Award

COMPUTATION + JOURNALISM SYMPOSIUM 2014

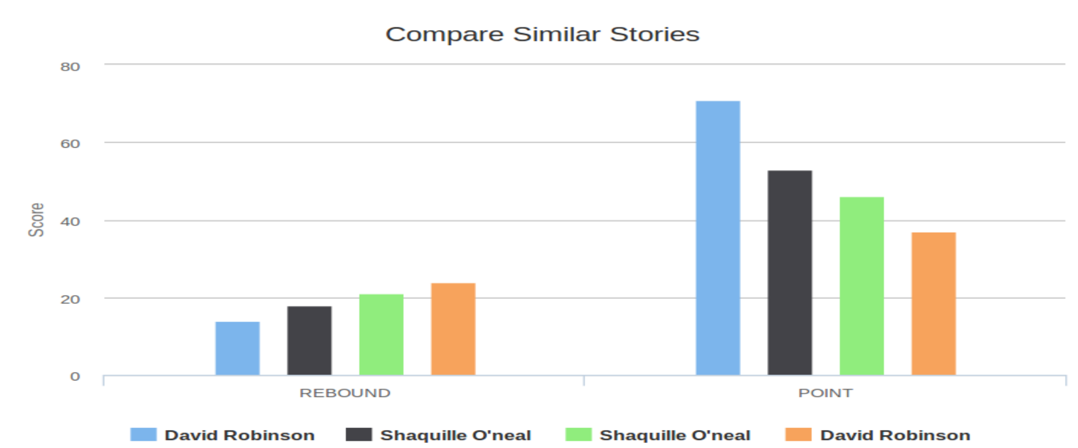
http://idir.uta.edu/factwatcher/

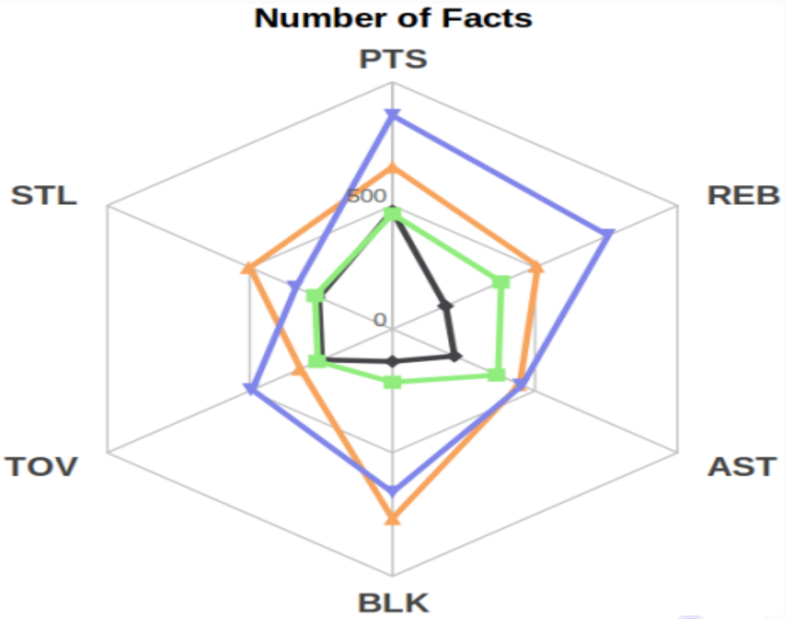
[April 24, 1994] David Robinson had 71 points and 14 rebounds in the San Antonio Spurs' victory against the Los Angeles Clippers. No one before had a better performance in NBA history.

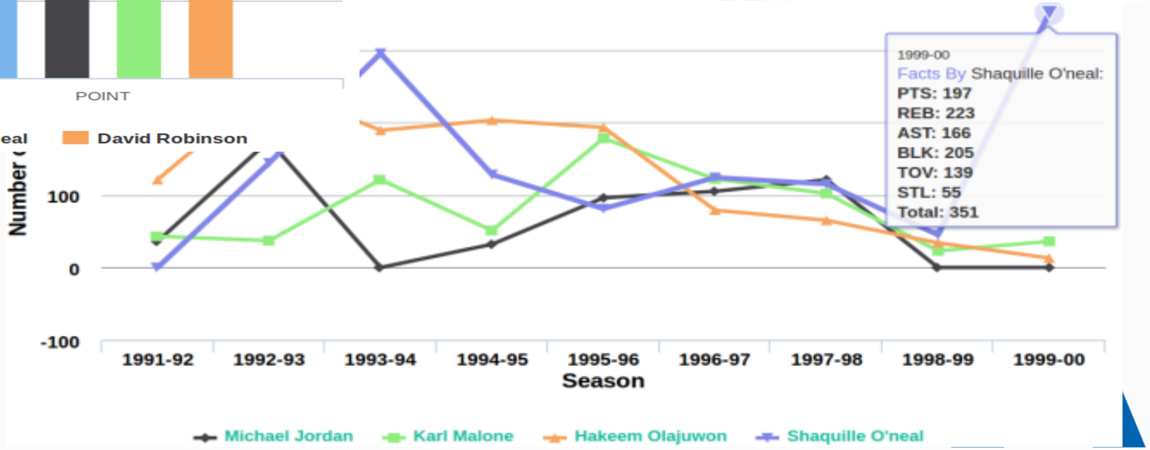
[April 20, 1994] Shaquille O'neal had 53 points and 18 rebounds in the Orlando Magic's win over the Minnesota Timberwolves. No one before had a better performance in NBA history.

[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.





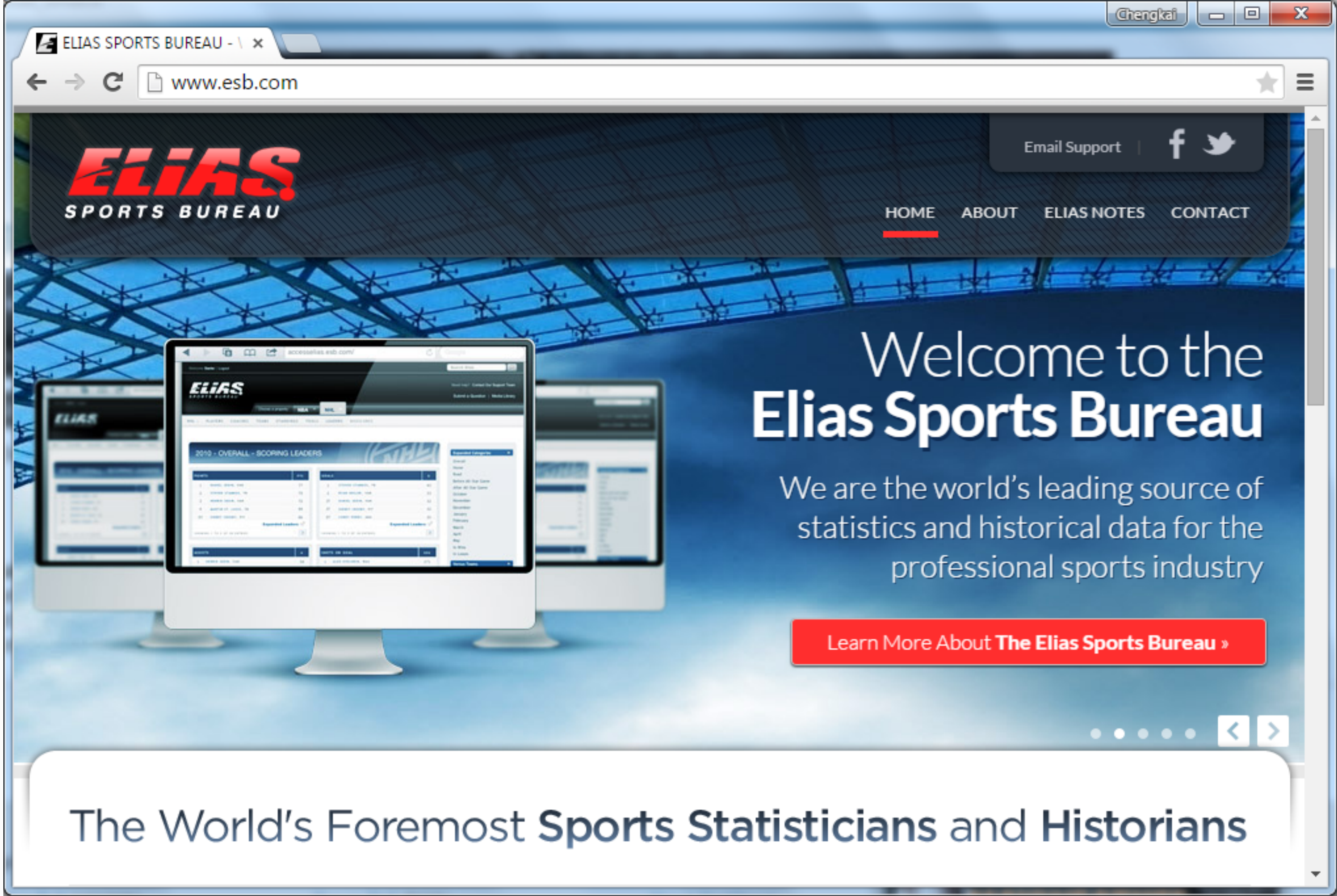


How were these Facts Discovered in Current Systems?

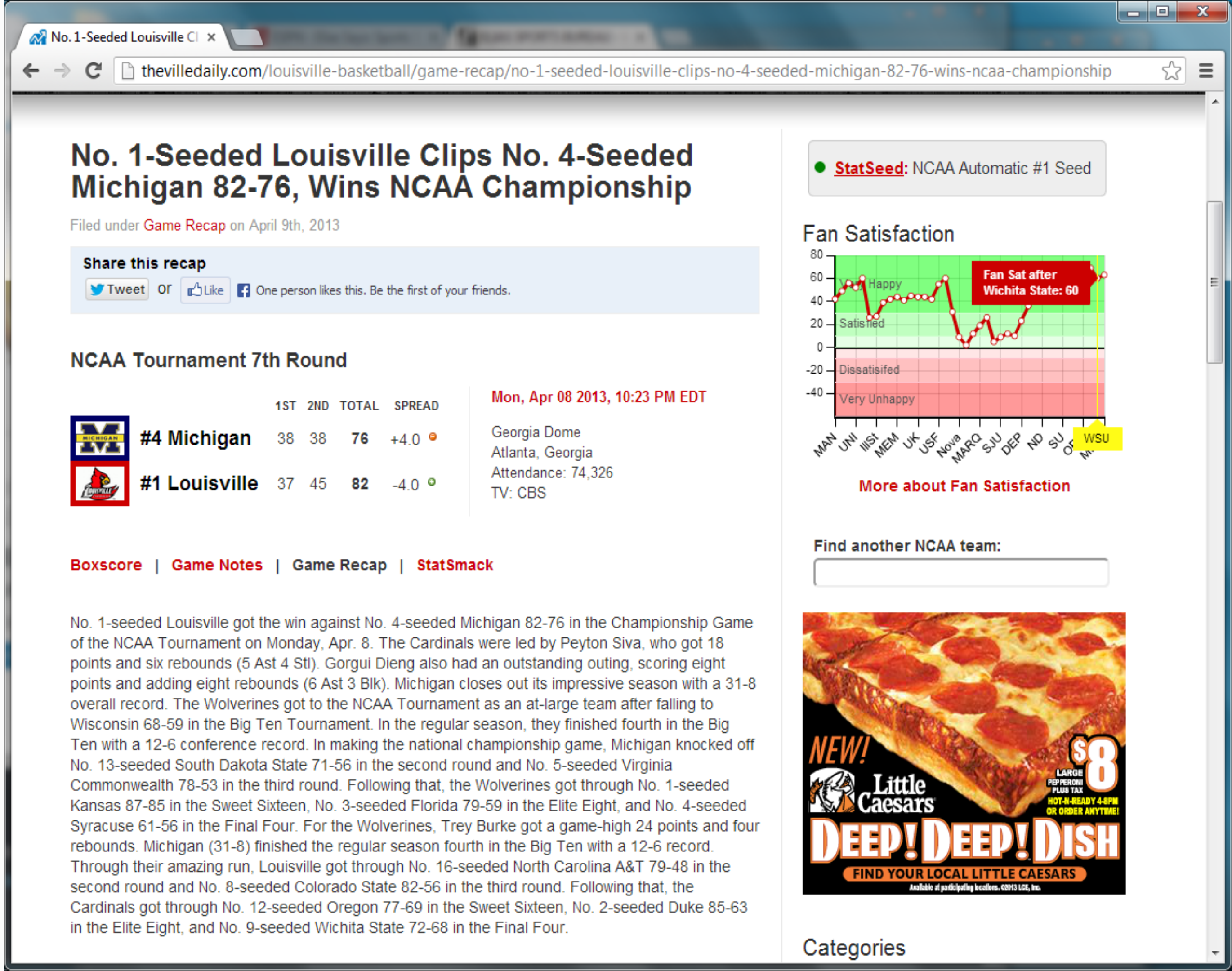
Our (educated?) guess

- 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



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

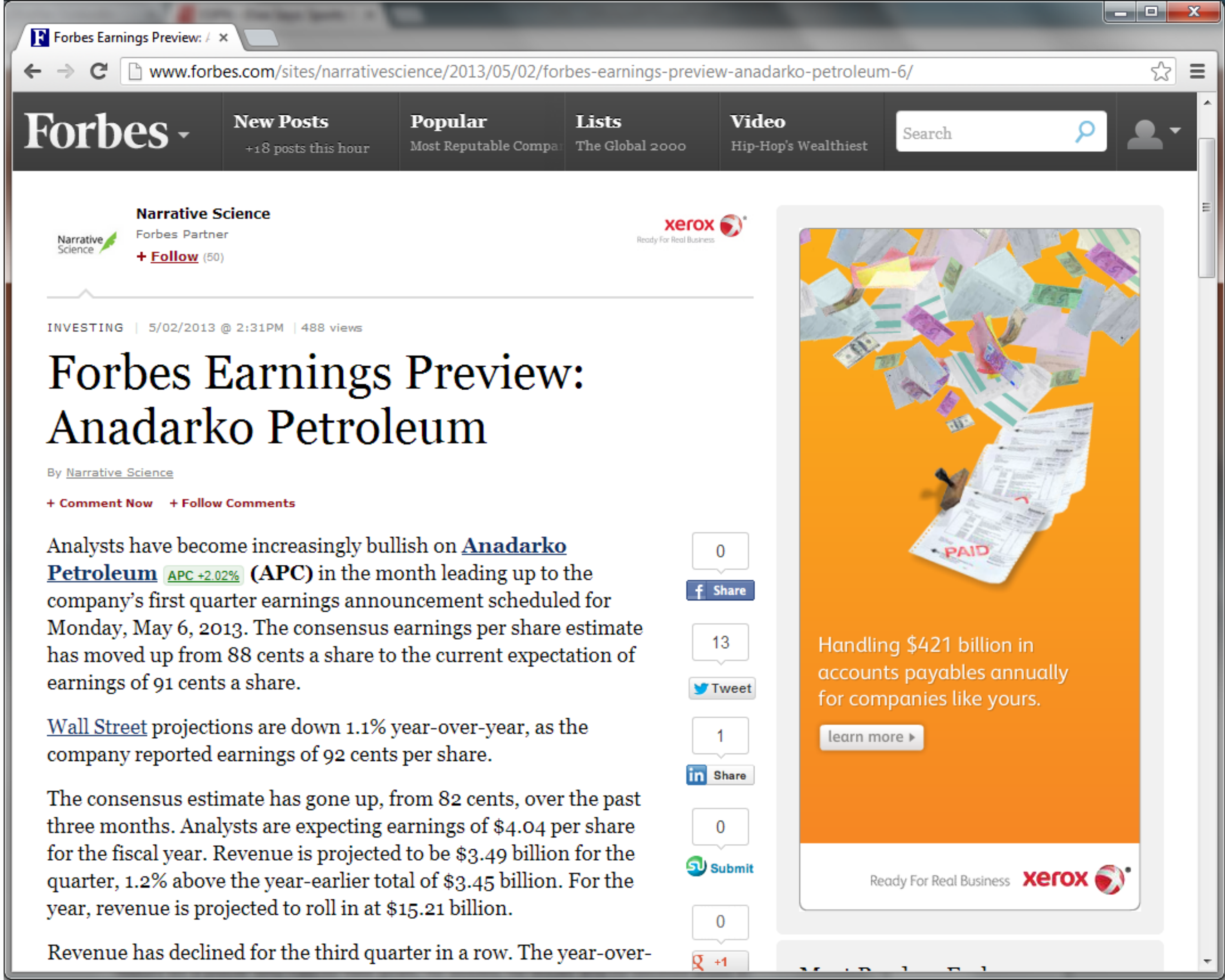
Fan Satisfaction

More about Fan Satisfaction

Find another NCAA team:

Categories

Narrative Science



Forbes Earnings Preview: /

www.forbes.com/sites/narrativescience/2013/05/02/forbes-earnings-preview-anadarko-petroleum-6/

Forbes | New Posts +18 posts this hour | Popular Most Reputable Compa | Lists The Global 2000 | Video Hip-Hop's Wealthiest | Search

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-

Publications

- 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.
- 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.
- 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.

Situational Facts

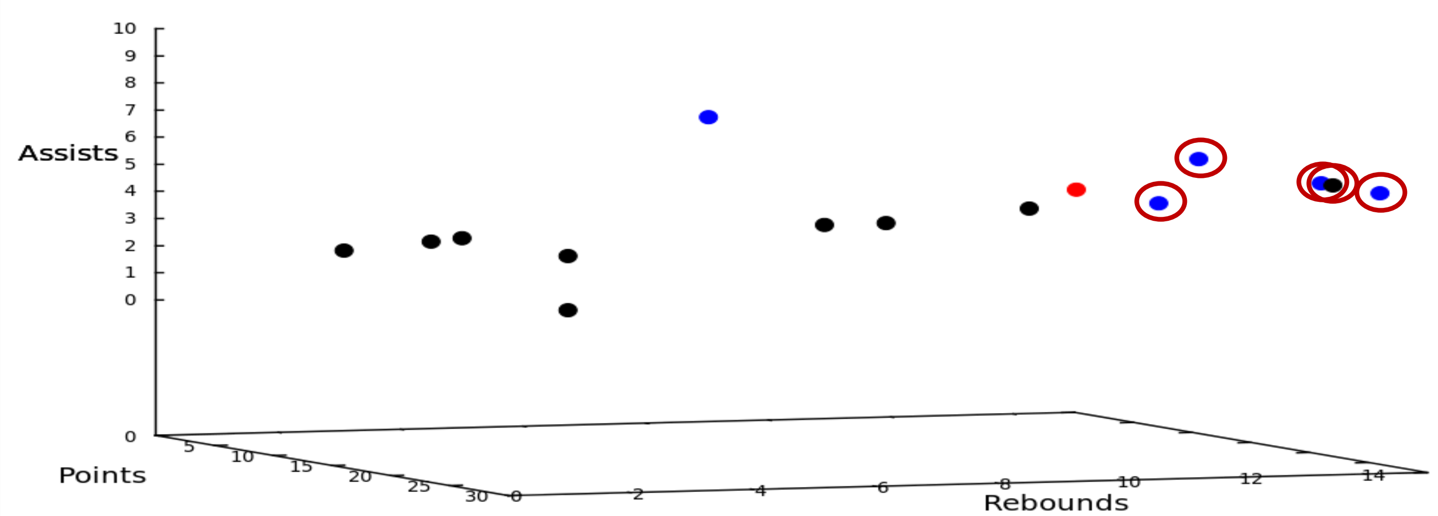
Incremental Discovery of Prominent Situational Facts. Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, Cong Yu. ICDE 2014, pages 112-123.

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)

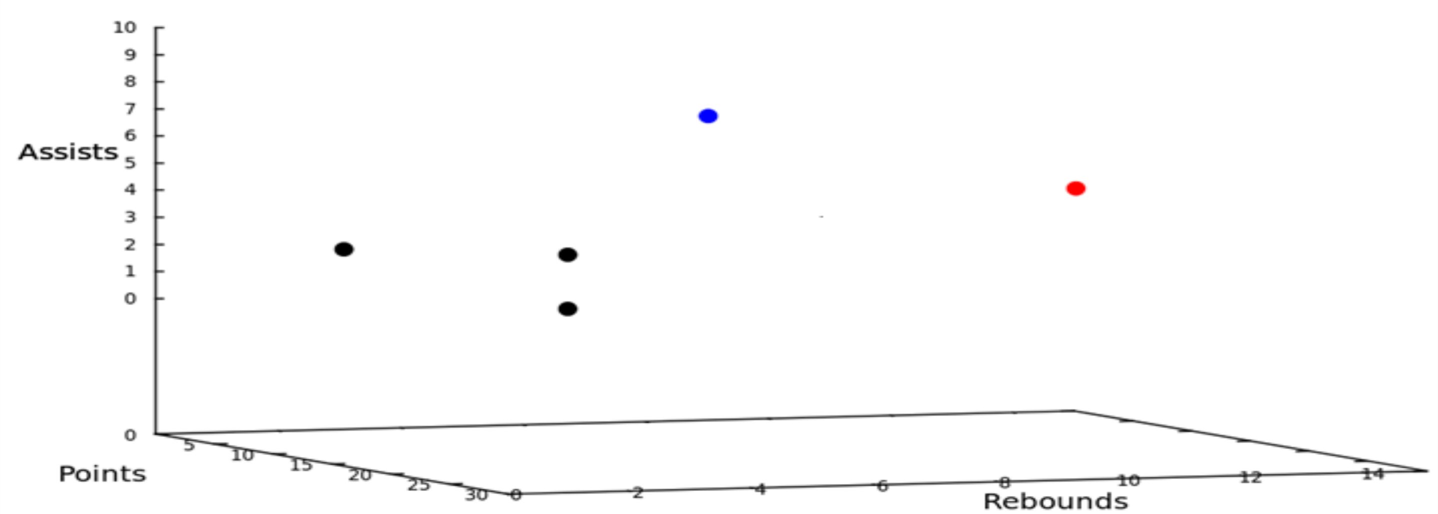
Situational Facts

"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)

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)

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)

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
Scientists Journalists Citizens

IDIRA

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):** $d_1 = v_1 \wedge d_2 = v_2 \wedge \ldots \wedge d_n = v_n$, $v_i \in dom(d_i) \cup \{*\}$
  - 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

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
…	…

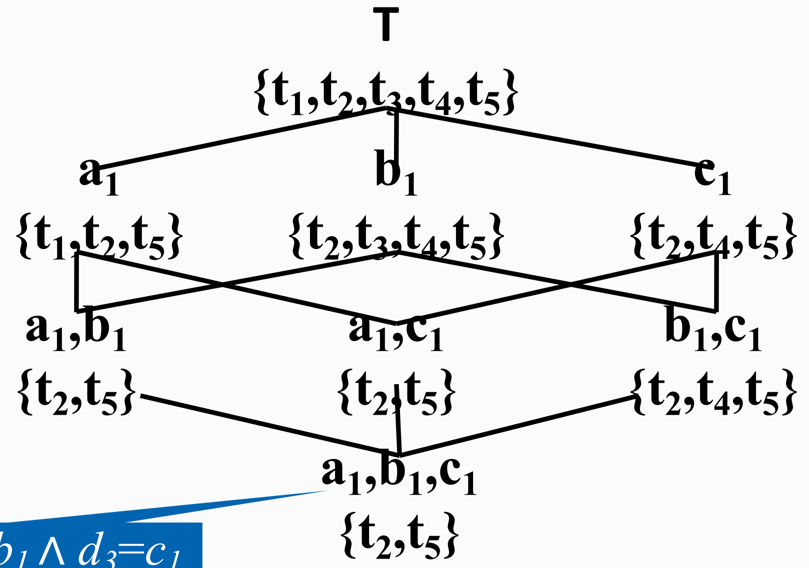
http://en.wikipedia.org/wiki/Basketball

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.

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



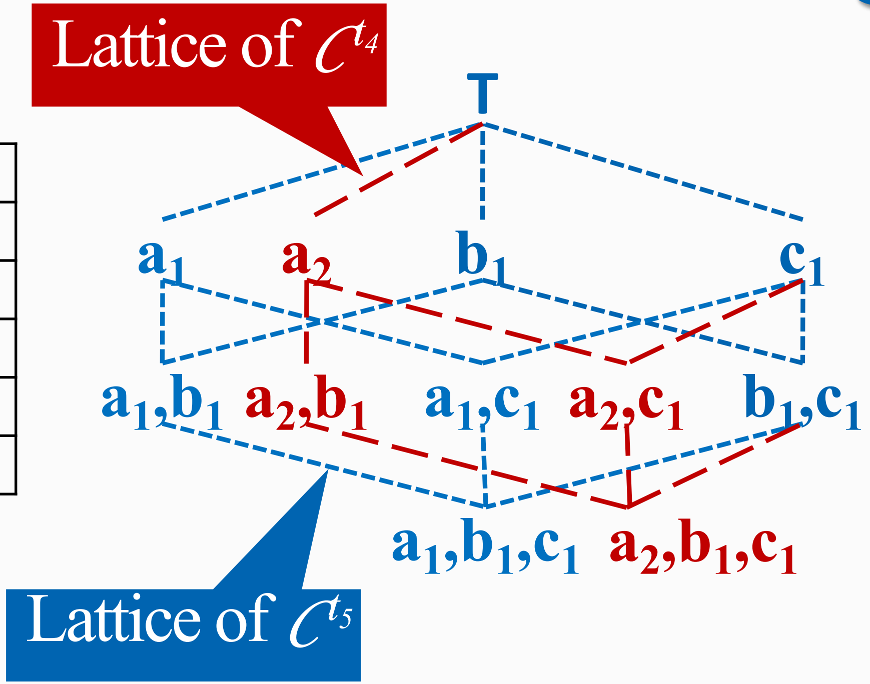
$d_1=a_1 \wedge d_2=b_1 \wedge d_3=c_1$

Lattice of $\mathcal{C}^{t_5}$

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

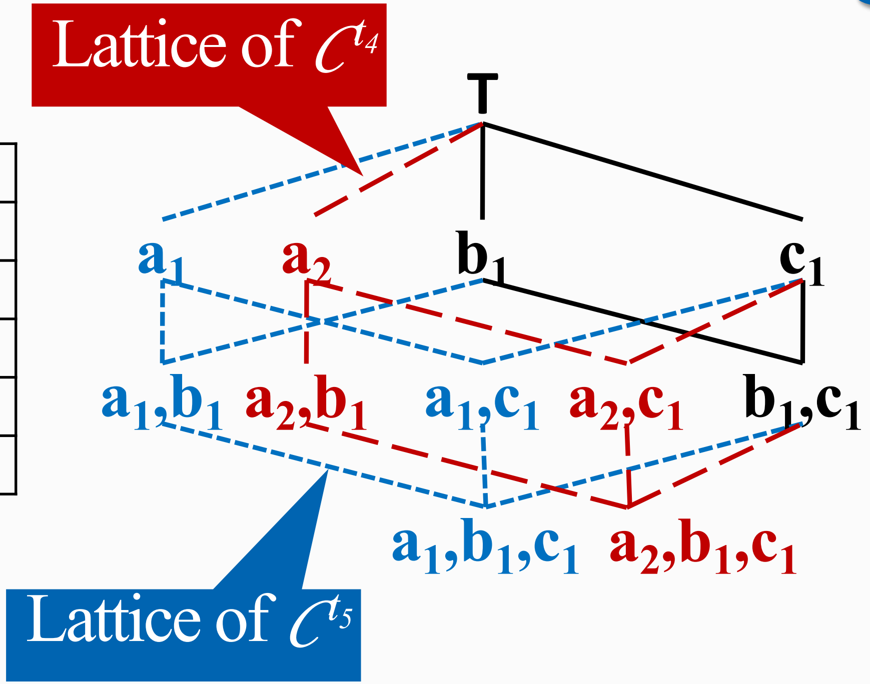
Modeling

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



Modeling

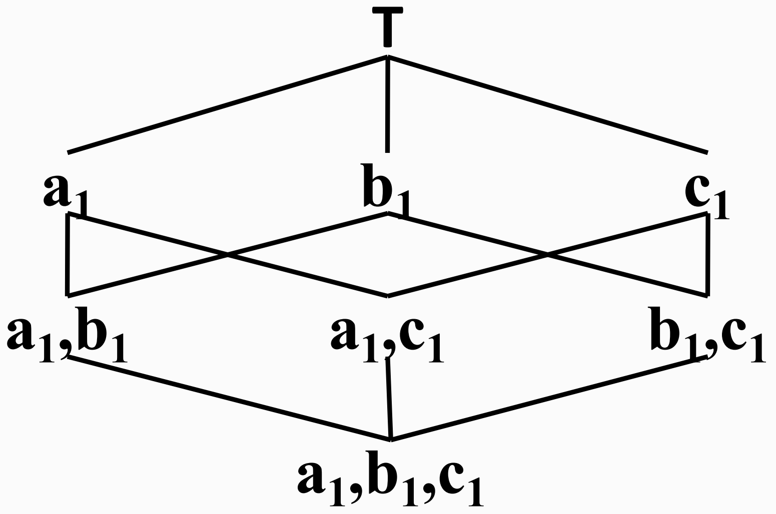
id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



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

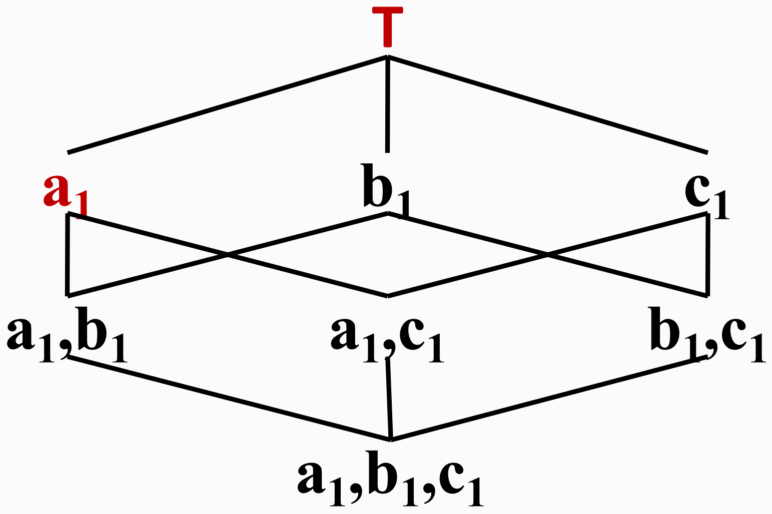
Brute-Force Approach

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



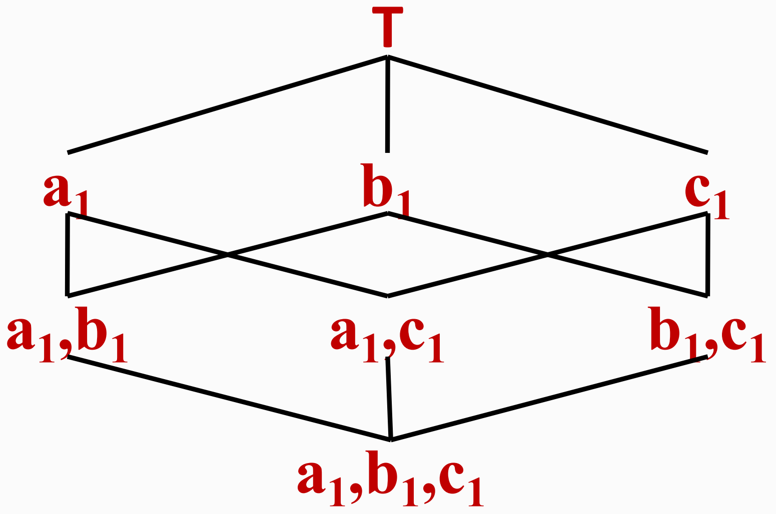
Brute-Force Approach

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



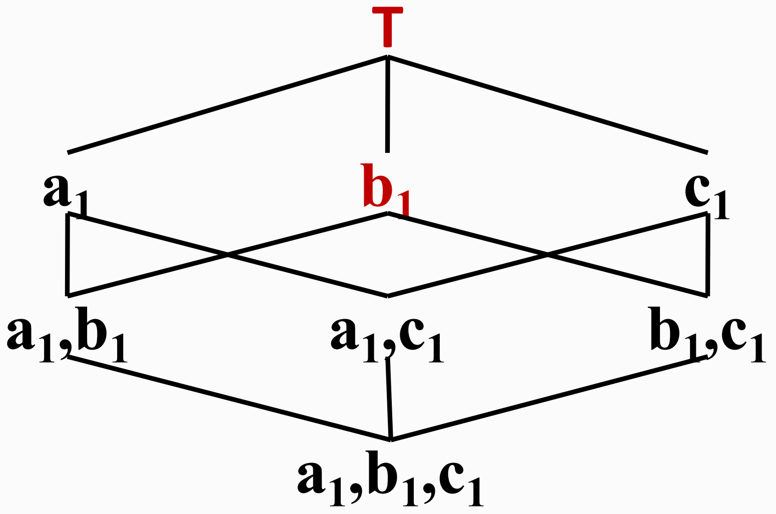
Brute-Force Approach

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



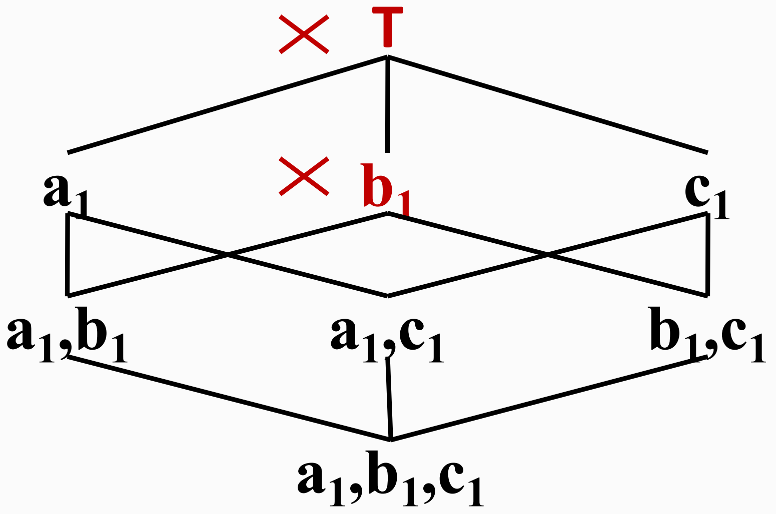
Brute-Force Approach

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



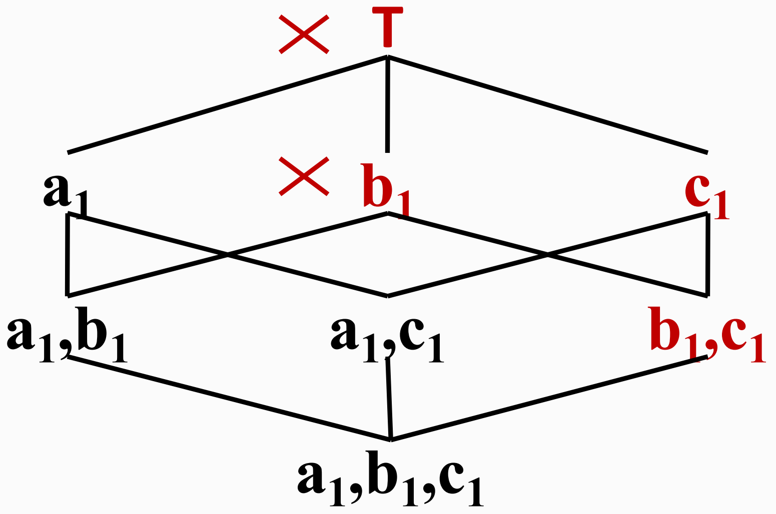
Brute-Force Approach

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



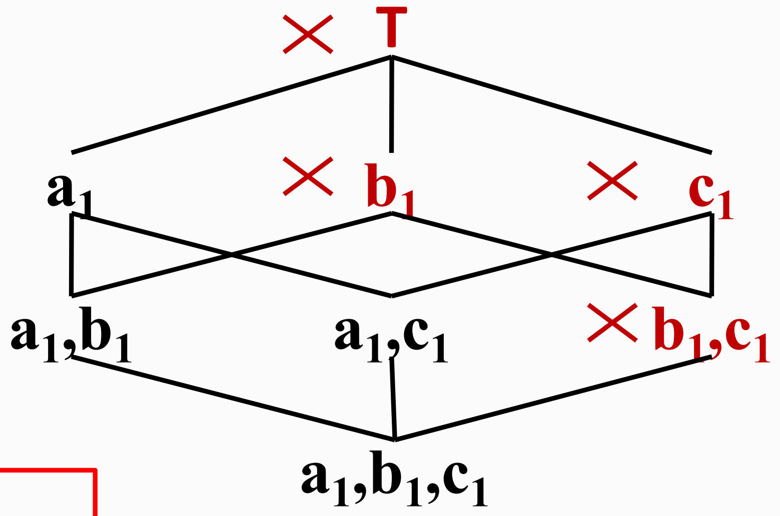
Brute-Force Approach

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

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

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 Streaks

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 5 4 6 7 3

<[6, 8], 4>

Streak dominance relation

s1=<[l1, r1], v1> dominates s2=<[l2, r2], v2> if and only if

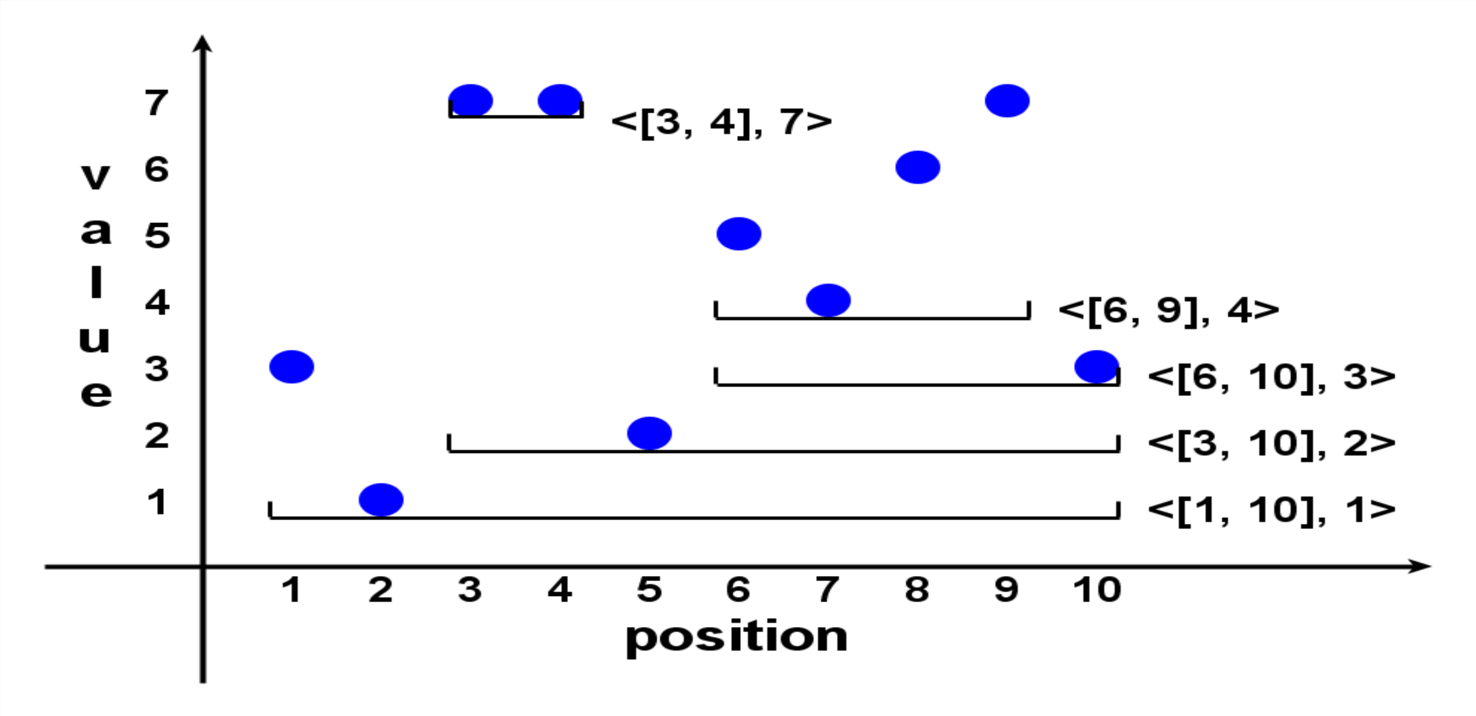
r1 - l1 > r2 - l2, v1 >= v2 or r1 - l1 >= r2 - l2, v1 > v2

Prominent streaks (PS)

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

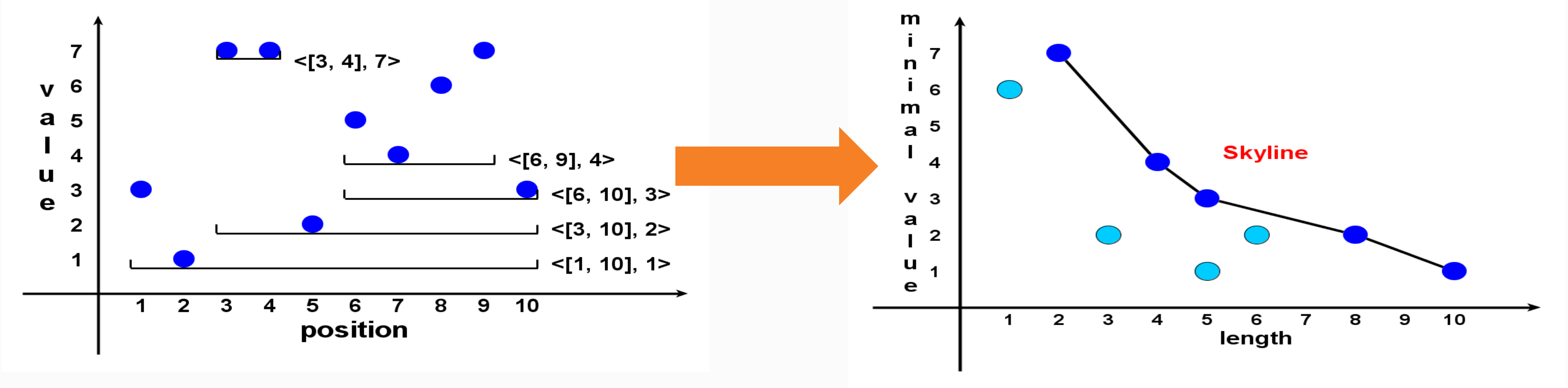
Example

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



Each streak can be viewed as a point in a 2-dimension space (length, minimal_value). Prominent streaks are skyline points in this space.

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

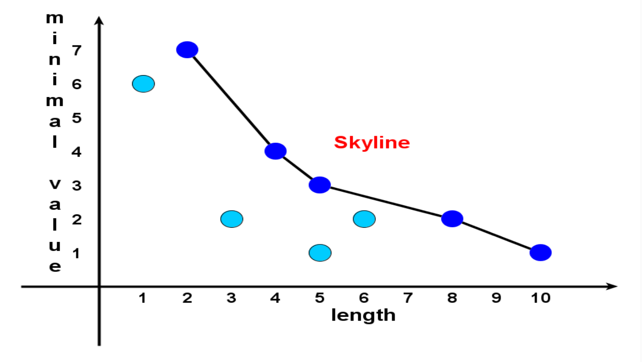
3 1 7 7 2 5 4 6 7 3

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

Candidate Streaks

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

Prominent Streaks



Candidate Generation: Number Of Candidate streaks

Brute-force

Quadratic (How many streaks are in a sequence of n values? These are all candidate streaks.)

NLPS

Superlinear

LLPS

Linear

Given a sequence of n values, how many streaks are there? That's the number of candidate streaks the brute-force algorithm will produce.

$n*(n+1)/2$

- Streaks starting at the 1^{st} position: n
- Streaks start at the 2^{nd} position: n-1
- …
- Streaks start at the nth position: 1

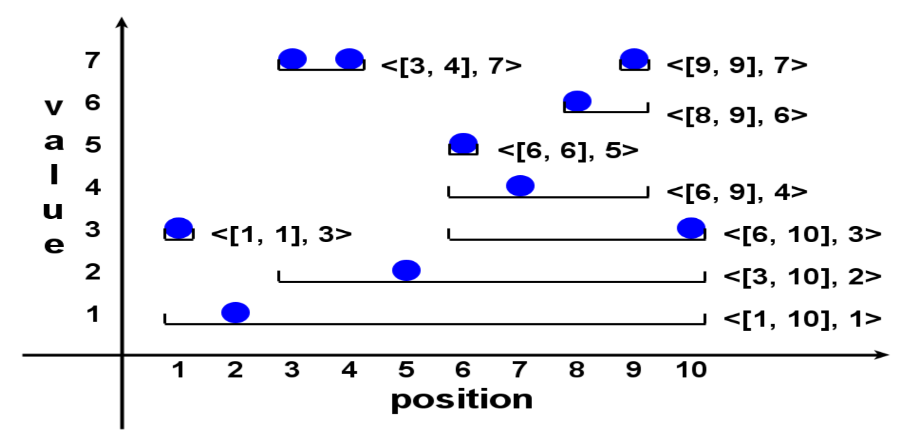
Local Prominent Streak

Local dominance relation

s1=<[l1, r1], v1> locally dominates s2=<[l2, r2], v2> if and only if a) s1 dominates s2 and b) [l1, r1] $\supset$ [l2, r2] (i.e., s1 subsumes/contains s2 and s2 is thus a sub-sequence of s1).

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.

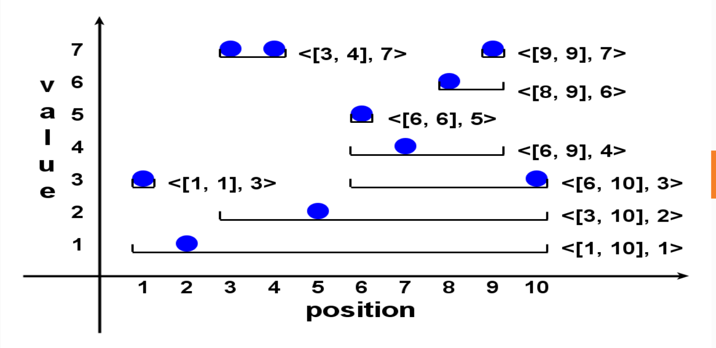
(Hint: The number of LPSs getting min value at position k is at most 1.)

Conclusion

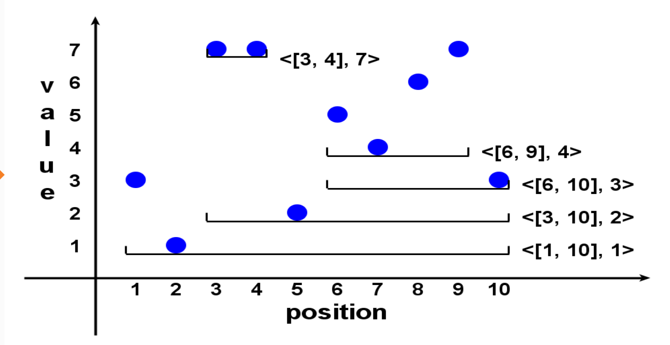
LPS is an excellent set of candidate streaks, of linear size.

The candidate generation problem becomes finding local prominent streaks.

Local prominent streaks (candidates)



prominent streaks (skyline points)

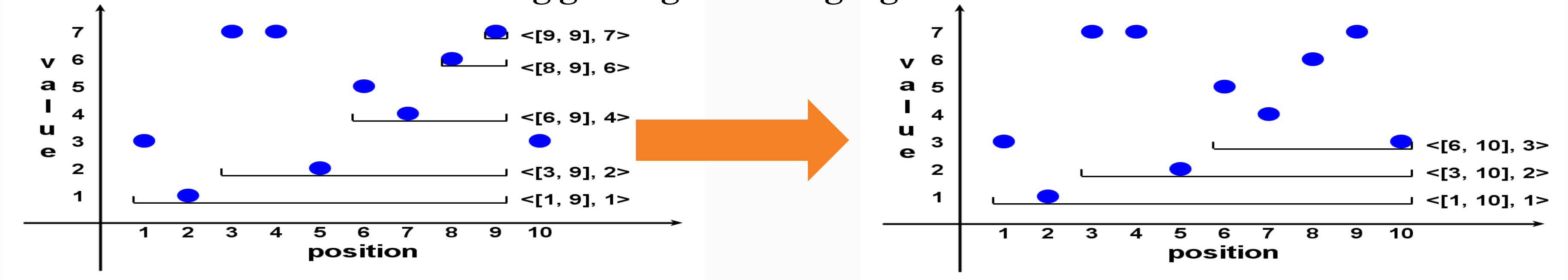


Linear LPS (LLPS) Method

Given a sequence p_1 , p_2 , …, p_n, …, maintain a list of growing streaks when scanning the sequence rightward.

1. At p_1, create a growing streak to include the 1st position.
2. After p_k , right-ends of growing streaks are all at position k.
3. At p_{k+1} , try to extend the growing streaks rightward. The growing streaks are partitioned by p_{k+1} into two groups:
   (3.a) Those growing steaks with s.v <= p_{k+1} : extend s to include position k+1.
   (3.b) Those growing streaks with s.v > p_{k+1} : include s into LPS; remove s from growing streaks as it won't grow anymore.
4. At p_{k+1}, do one of the following, based on which condition is satisfied:
   (4.a) There was a growing streak with s.v = p_{k+1} : nothing more needs to be done for p_{k+1}.
   (4.b) There were one or more growing streaks with s.v > p_{k+1} : create a new growing streak by extending the leftmost (longest) such growing streak s that satisfies s.v > p_{k+1} to include position k+1.
   (4.c) There was no growing streak with s.v >= p_{k+1} : create a new growing streak to include just position k+1.

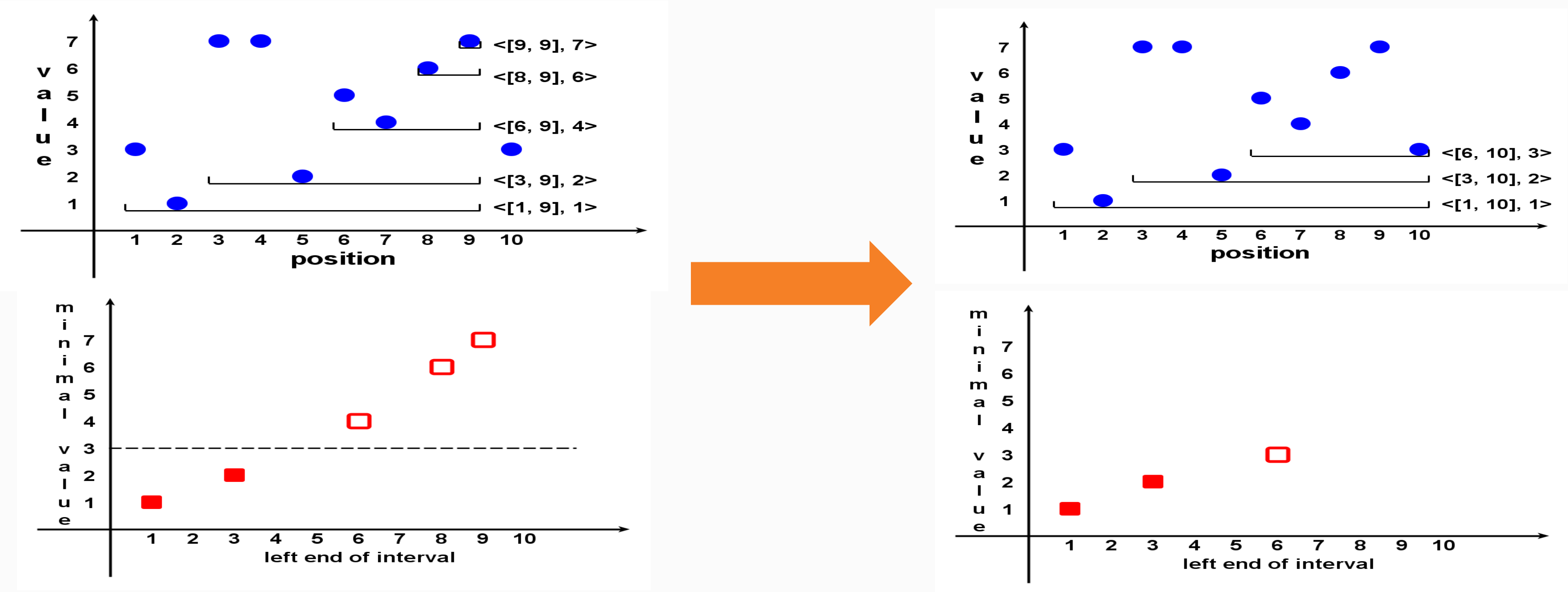
An illustration of the method at p_{10}, with all remaining growing streaks highlighted



Linear LPS (LLPS) Method

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

Two different ways of illustrating the method at p_{10}, with all remaining growing streaks highlighted



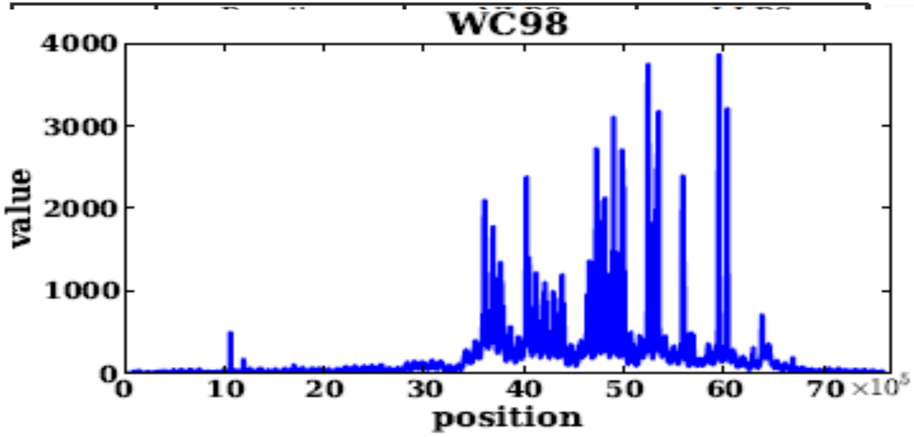
Linear LPS (LLPS) Method

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

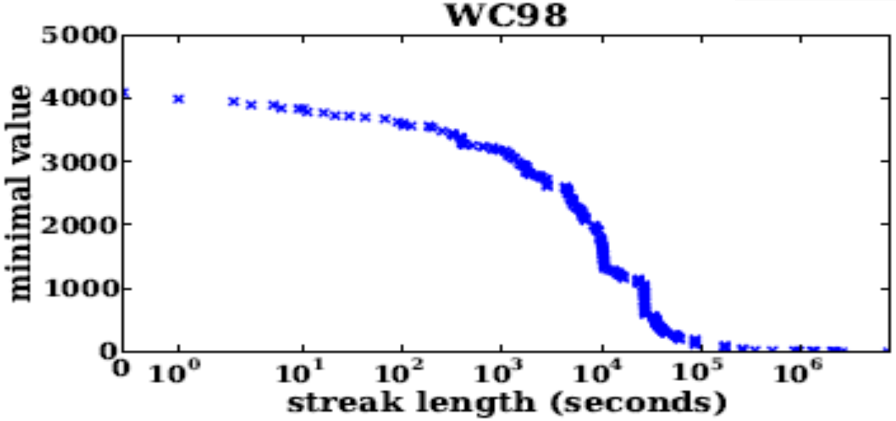
- Keeps applying the algorithm when new values of the sequence are received.
- After the latest value at p_n, it has found all LPSs ending before position n.
- Growing streaks ending at n either will eventually be LPSs or can be grown into LPSs ending at a future position after n.
- If prominent streaks till n are requested, compare all found LPSs and all growing streaks.

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

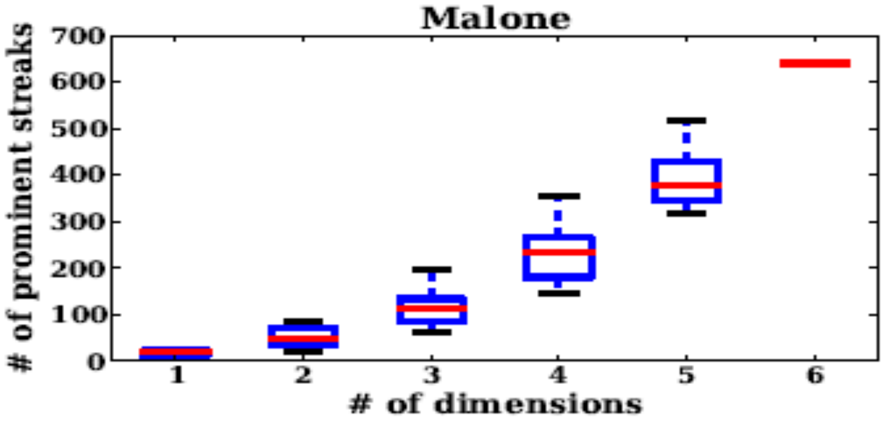
Table IX. Multi-sequence Prominent Streaks in Datast NBA1.

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

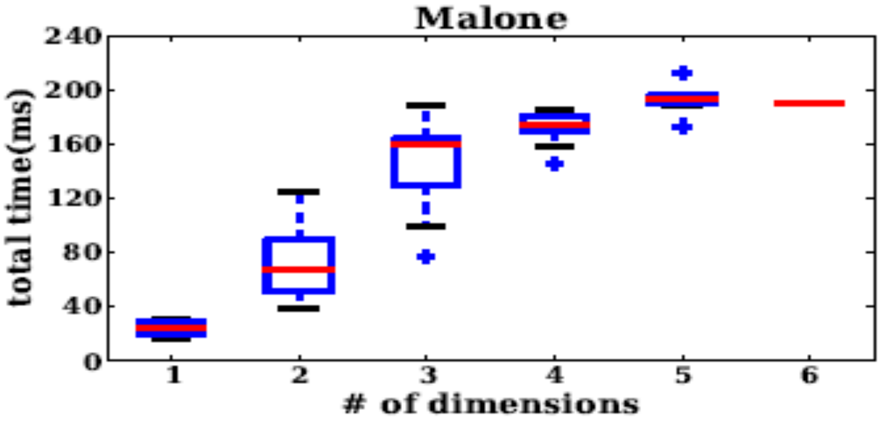
Experiments On Multi-Dim PSs

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 N-BA players (minutes, points, re-bounds, 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

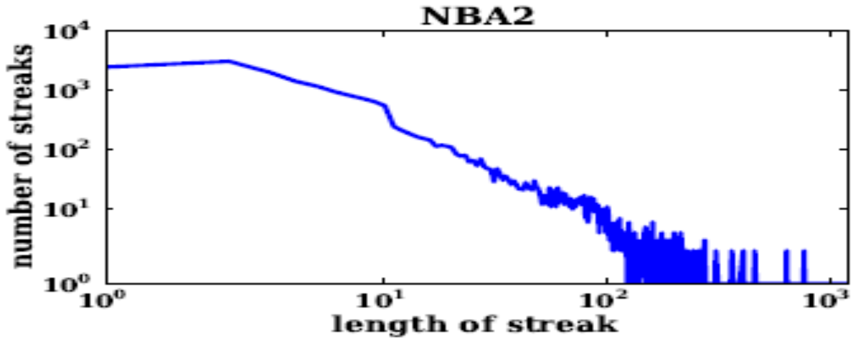


Fig. 14. Distribution of Prominent Streaks by Length.

Exercise

Apply the LLPS algorithm to find prominent streaks in 4 1 3 5 4 7 3 6. Show local prominent streaks and growing streaks at every position.

One-of-the-few Objects

On "One of the Few" Objects. You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. KDD 2012, pages 1487-1495

One-Of-The-Few Claims

Do these claims really hold water?

Karl Malone is ONE OF THE ONLY TWO players in NBA history with 25,000 points, 12,000 rebounds, and 5,000 assists in one's career.

He is ONE OF THE ONLY THREE candidates who have raised more than 25% from PAC contributions and 25% from self-financing.

How do we find truly interesting claims or individuals?

X Is One-Of-K →X Is In K-Skyband

Claim

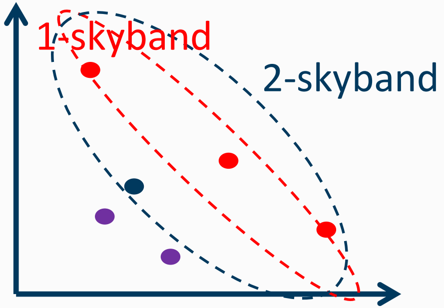
Karl Malone is ONE OF THE ONLY TWO players in NBA history with 25,000 points, 12,000 rebounds, and 5,000 assists in one's career.

General claim

Fewer than k objects dominate X in subspace of attributes $S \subseteq \{A\_1, A\_2, \ldots, A\_d\}$

k-skyband [Papadias et al. 2005] in S is the set of points each dominated by fewer than k other points in S

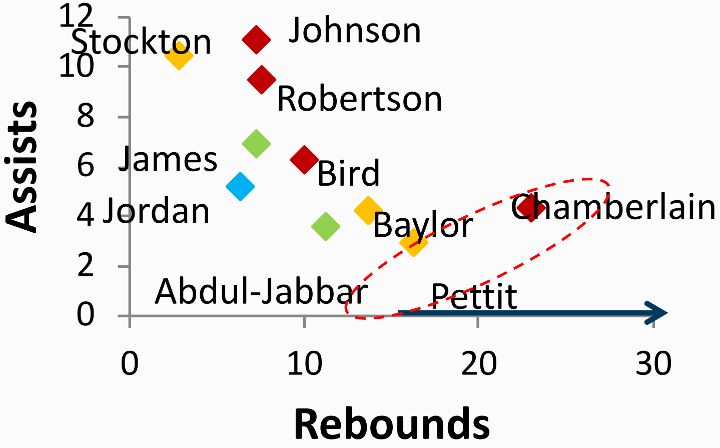
1-skyband : skyline

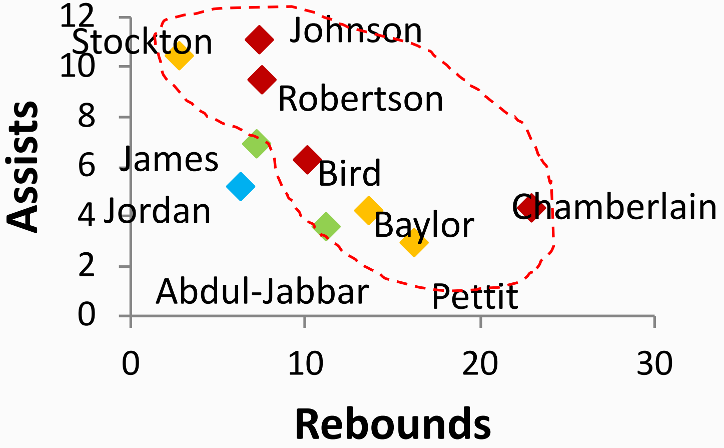


Small K ≠Interesting

Subspaces are different

E.g., 2-skyand in {rebounds} vs. in {rebounds, assists}

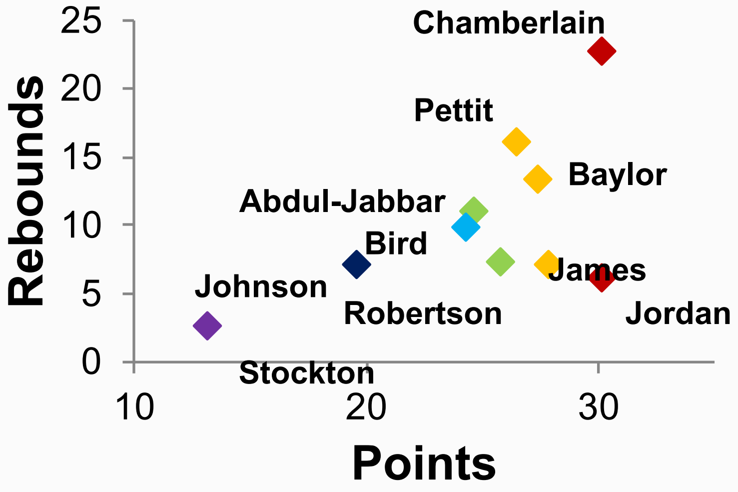


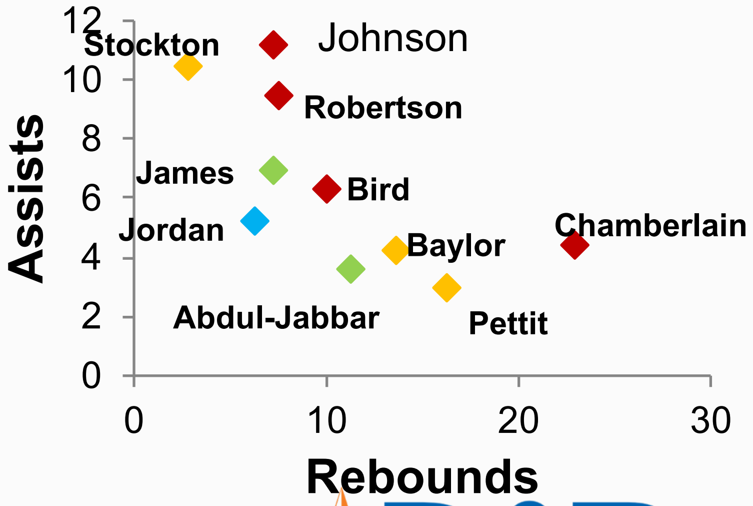


Small K ≠Interesting

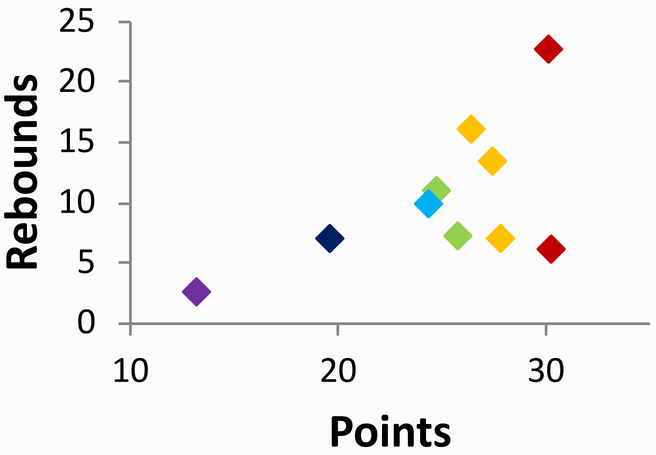
Data distribution matters

E.g., 2-skyand in {points, rebounds} vs. in {rebounds, assists}





Top-τ Skyband



k-Skyband

Using the same k for all subspaces doesn't work

Asking user pick k for each subspace is infeasible

Top-τ Skyband

- User specifies a single parameter τ to cap # skyband objects.
- For each subspace S, find its top-τ skyband, i.e., the largest k-skyband containing no more than τ objects
- E.g., in {points, rebounds}:

τ=**2**→ 1-skyband (size 2)

τ=**6**→ 2-skyband (size 5; 3-skyband would be too big)