On “One-of-the-few” Objects

You Wu, Pankaj K. Agarwal, Jun Yang Duke University
Chengkai Li University of Texas, Arlington
Cong Yu Google, Inc.

"One of the Few" Claims

Sports: *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*

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

- Do these claims really hold water?
- How do we find truly interesting claims or individuals?

Applications

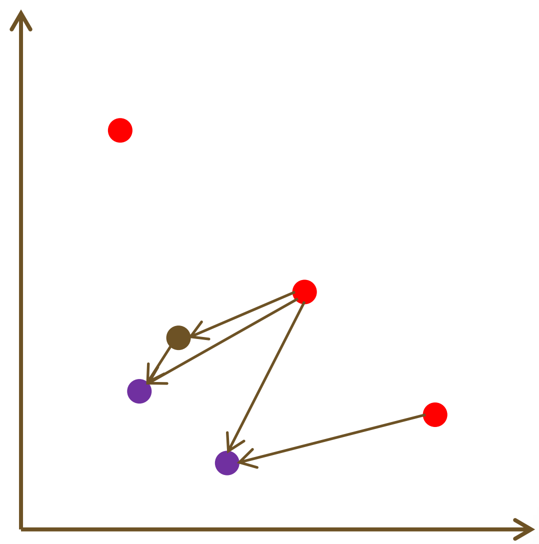
- Computational journalism: use computing to help
  - Increase effectiveness and reduce cost
  - Improve understanding and broaden participation
  - Guard against "lies, damned lies, and statistics"
- Usability is key!
- We target "one of the few" claims in this paper
- Domains include
  - Sports; election campaign finance; government, education, and business performance indexes
  - Or in general, wherever objects are compared across many dimensions

Key Challenges

- General claim: *fewer than k objects <u>dominate</u> X in <u>subspace</u> of attributes $S \subseteq \{A_1, A_2, \dots, A_d\}$*

*He is **ONE OF THE ONLY TWO** players with 25,000 points, 12,000 rebounds, and 5,000 assists*

Point p dominates q if p is no worse than q in all attributes , and strictly better in at least one of them



Key Challenges

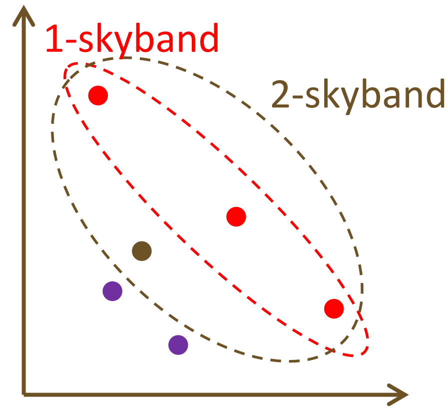
- General claim: *fewer than k objects dominate X in subspace of attributes $S \subseteq \{A_1, A_2, \ldots, A_d\}$*
- Is it interesting?
  - Small k $\neq$ interesting
- Finding interesting claims/individuals
  - Where to look for? – All subspaces
  - Who determines k? – Not the users!
  - How to find interesting claims? – Brute force is too slow

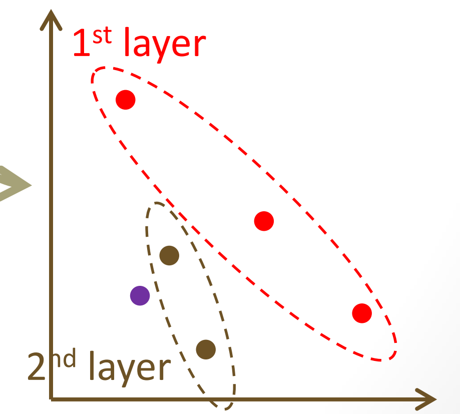
Roadmap

- Introduction
- **Identifying Interesting Claims**
  - **“Uniqueness” of Claims**
  - **Top-τ Skyband Problem**
  - **Algorithms**
- Ranking Objects
- Conclusion and Future Work

Data Model and Preliminaries

- Objects are points in d-dimensional space
- 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 is also known as "skyline"
  - Different from skyline layer by layer

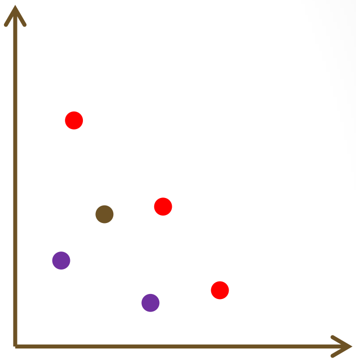




- ***X is one of k in S*** means ***X ∈ k-skyband in S***
  - *Recall general form: fewer than k objects dominate X in subspace S*

Small k ≠ Interesting

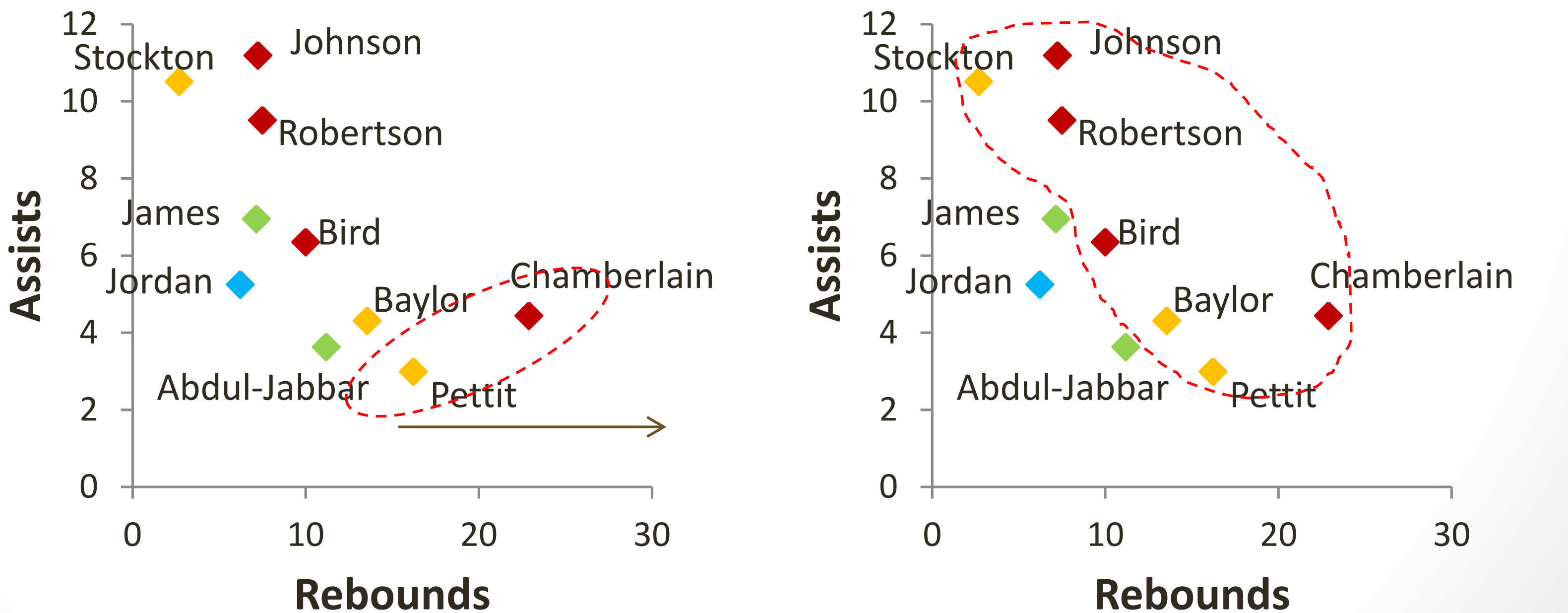
- E.g., *X is dominated in S by no others*
  - 3 on the right, or as many as you'd like



- An interesting claim should be sufficiently **unique**—it cannot be made for many other objects
- Size of the k-skyband measures uniqueness of one-of-k claims; k itself does not

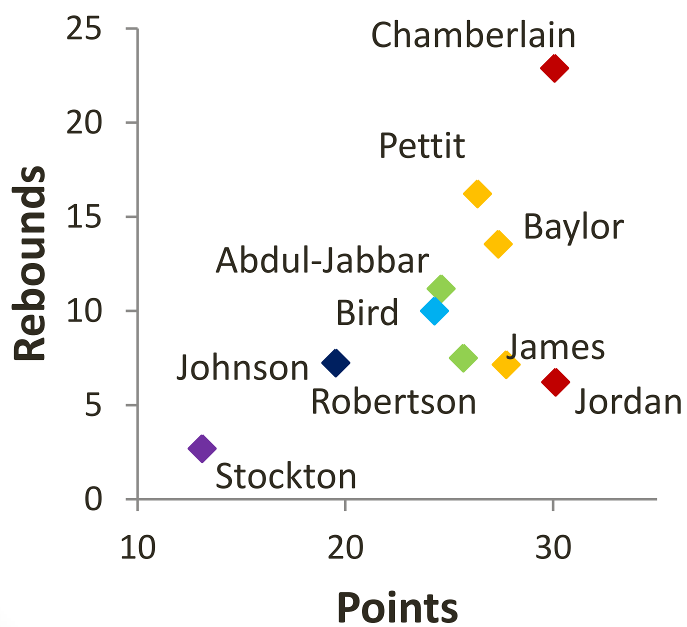
Finding Unique Claims: Challenges

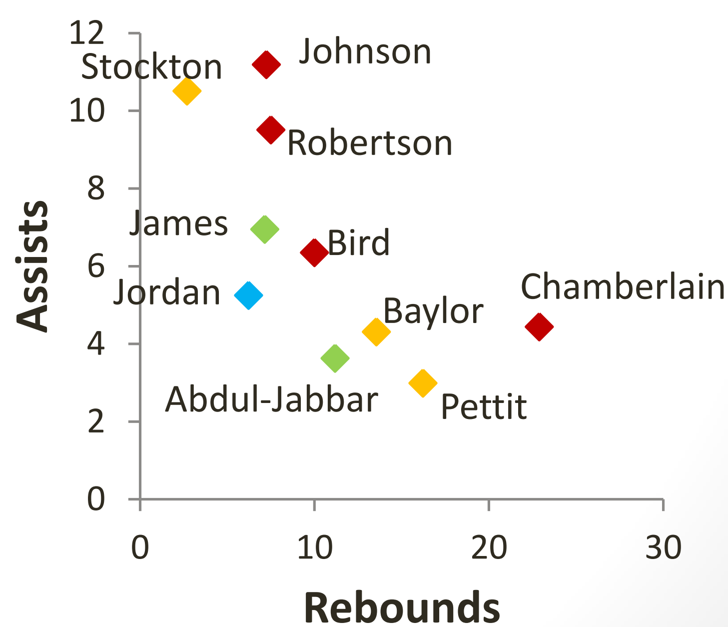
- Existing skyband algorithms require user to pick k
- But to ensure uniqueness, choice of k depends on subspace dimensionality
  - E.g. 2-skyand in {rebounds} vs. in {rebounds, assists}



Finding Unique Claims: Challenges

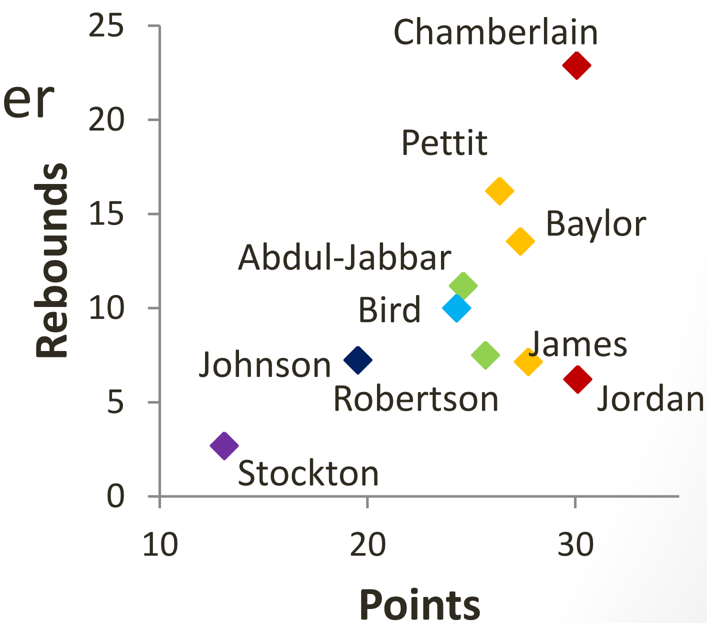
- To ensure uniqueness, choice of k also depends on data distribution
  - Anti-correlated attribute values make skybands bigger
  - E.g.: (Correlated) vs. (Anti-Correlated)





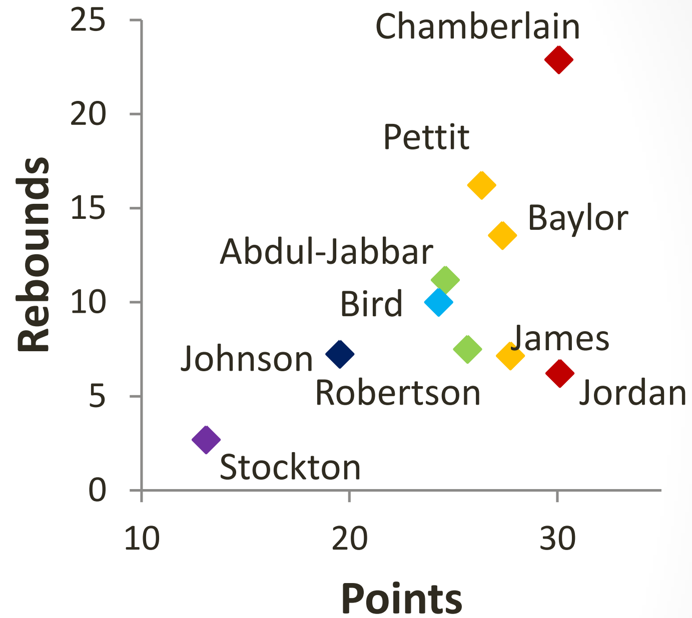
Finding Unique Claims: Solution

- Using same k for all subspaces doesn't work
- Making user pick k for each subspace is infeasible
- Our solution: **top-τ skyband**
  - User specifies a single parameter $\boldsymbol{\tau}$ 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}:
    - $\tau = \mathbf{2}$ → 1-skyband (size 2)
    - $\tau = \mathbf{6}$ → 2-skyband (size 5; 3-skyband would be too big)



Advantages of Top-τ Formulation

- Easy to use and interpret
- A single τ to pick $\rightarrow k$ automatically adapts based on subspace dimensionality and data distribution
  - E.g., 10 2-d points; let $\tau = 3$
  - Automatically detects subspaces with no "unique" claims
- Each claim found comes with the guarantee that the same cannot be said for more than τ objects

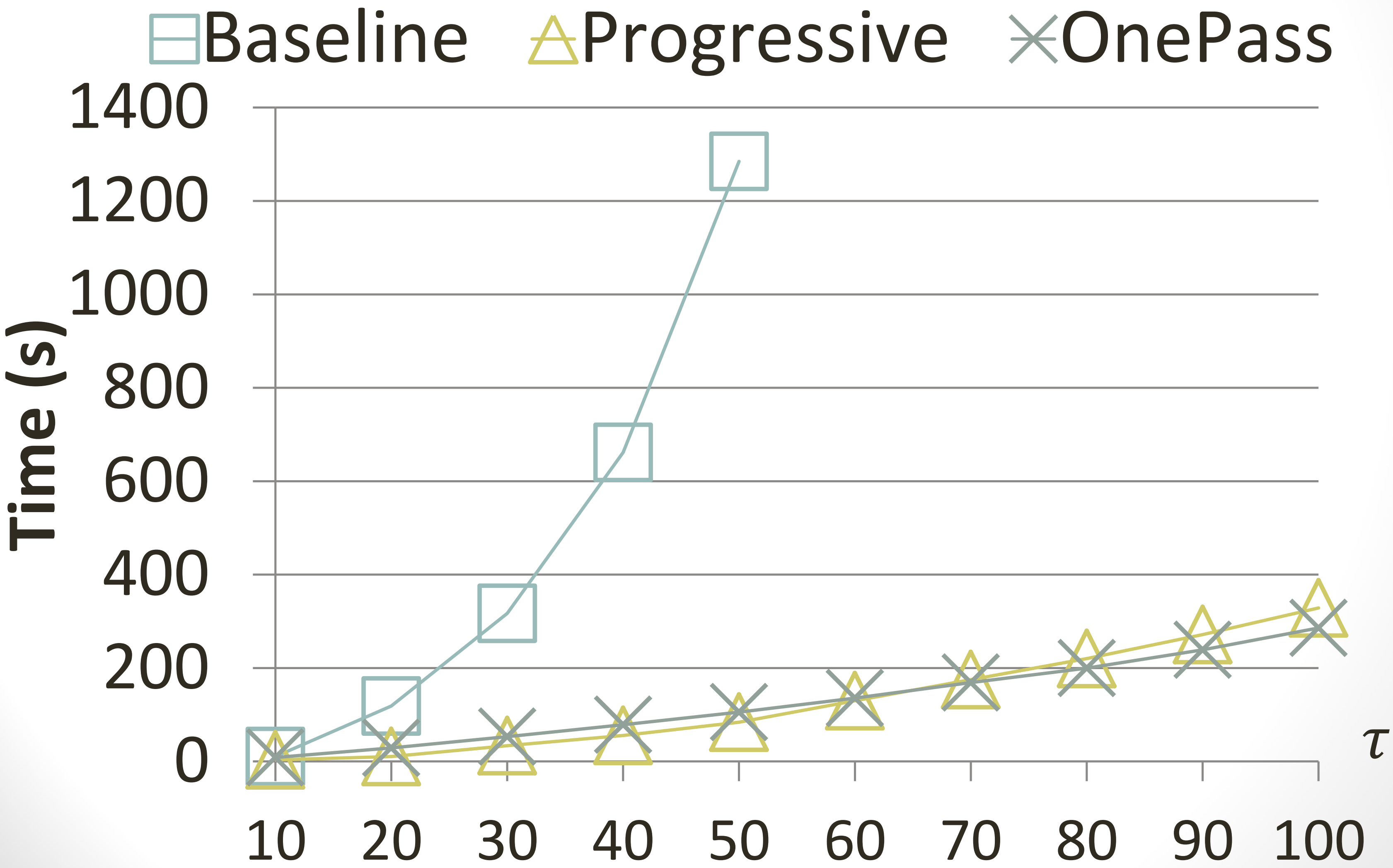


Computing Top-τ Skybands

- Computing top-τ skyband in an individual subspace
  - ***Progressive***: grow the skyband tier by tier until it is too big; the next tier is always contained in the skyline of non-skyband objects
  - ***OnePass***: bound the size of "working set" by τ by processing objects in a particular order to avoid full exploration of a tier that is too large

- Computing top-τ skybands in all subspaces
  - Bottom-up (subspace) lattice traversal [Pei et al. 2006]
  - sharing computation, new pruning techniques

Performance on NBA career total data



Roadmap

- Introduction
- Identifying Interesting Claims
- **Ranking Objects**
  - **Existing Solutions**
  - **Adjustable Positional Score with Ties**
- Conclusion and Future Work

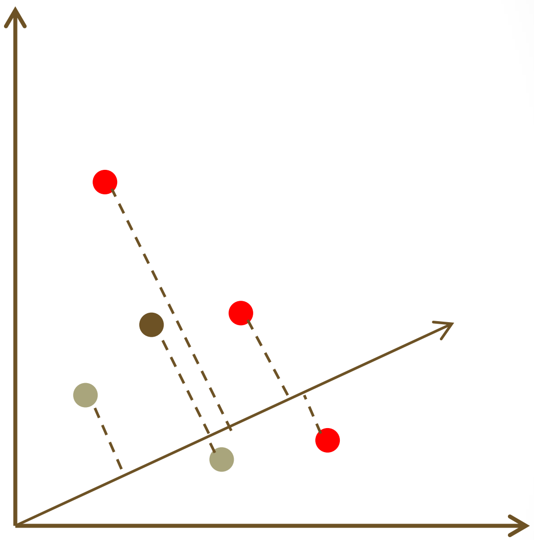
Ranking Objects

- Sometimes we are more interested in investigating objects that show up in claims than individual claims per se
  - Need to rank objects by their "interestingness"
- Grouping claims by the objects they mention also helps user navigate through numerous claims

Existing Methods: Valued-Based

Weighted Sum

- User specifies a weight vector (one weight for each attribute)
  - I.e., a direction in the d-dim space
- Objects are ranked based on the weighted sum of their attribute values
  - I.e., their projections onto the weight vector
- Difficult to use: too many knobs ($d - 1$) to set and tune



Existing Methods: Rank-Based

Kemeny Optimal Aggregation [Dwork et al. 2001]

- Given a number of input rankings of all objects, find a ranking that minimizes the total number of pairwise disagreements with the input rankings
  - Natural to use d input rankings, one for each attribute
- NP-hard to compute
- Inflexible to use: no knob at all
  - Some tuning is often needed; e.g., which of the following players would you prefer?
    - John Stockton (specialized):
      $404^{th}/1622^{nd}/2^{nd}$ in points/rebounds/assists
    - Larry Bird (well-rounded):
      $17^{th}/60^{th}/44^{th}$ in points/rebounds/assists

Our Approach

- Extend the uniqueness-based interestingness measure of claims to objects
  - Score an object in each subspace S by the uniqueness of the one-of-few claim involving this object in S
  - Sum up object scores across all subspaces
  - Rank objects by their aggregate scores
- Provide one (and only one) knob to tune preference towards specialized vs. well-rounded objects
  - This knob is a parameter (α) in the per-subspace object scoring function

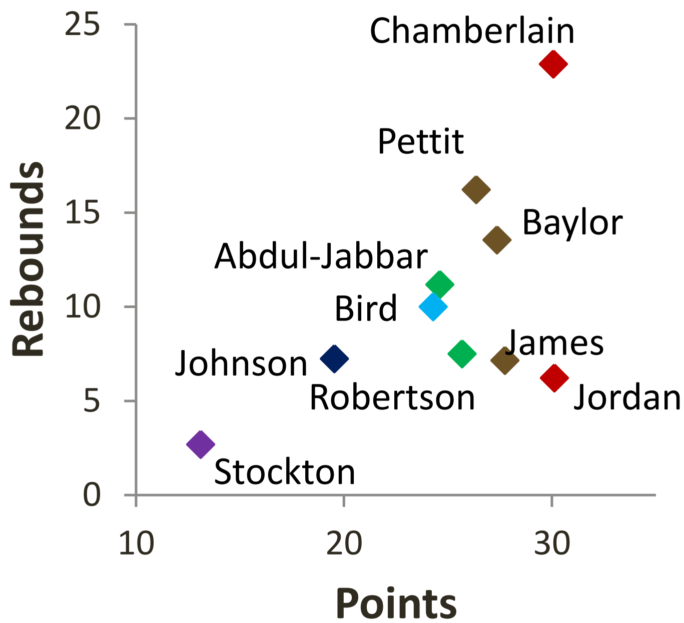
APST-α

All-Subspace Positional Score with Ties

- In each subspace, order & score objects by skyband tiers
  - Score drops exponentially: position i gets α^{i-1} $(0 < \alpha < 1)$
  - Objects in the same skyband tier (i.e., ties) divide up the total score for the tier equally
- For each object★, sum up its scores across all subspaces

Subspace	$1, \alpha, \alpha^2, \alpha^3, \alpha^4, \ldots$			
…	…			
$\{B, C\}$	1-skyband	2-skyband ★	4-skyband	5-skyband
…	…			
$\{A, B, C\}$	1-skyband ★	3-skyband		

Scoring Example

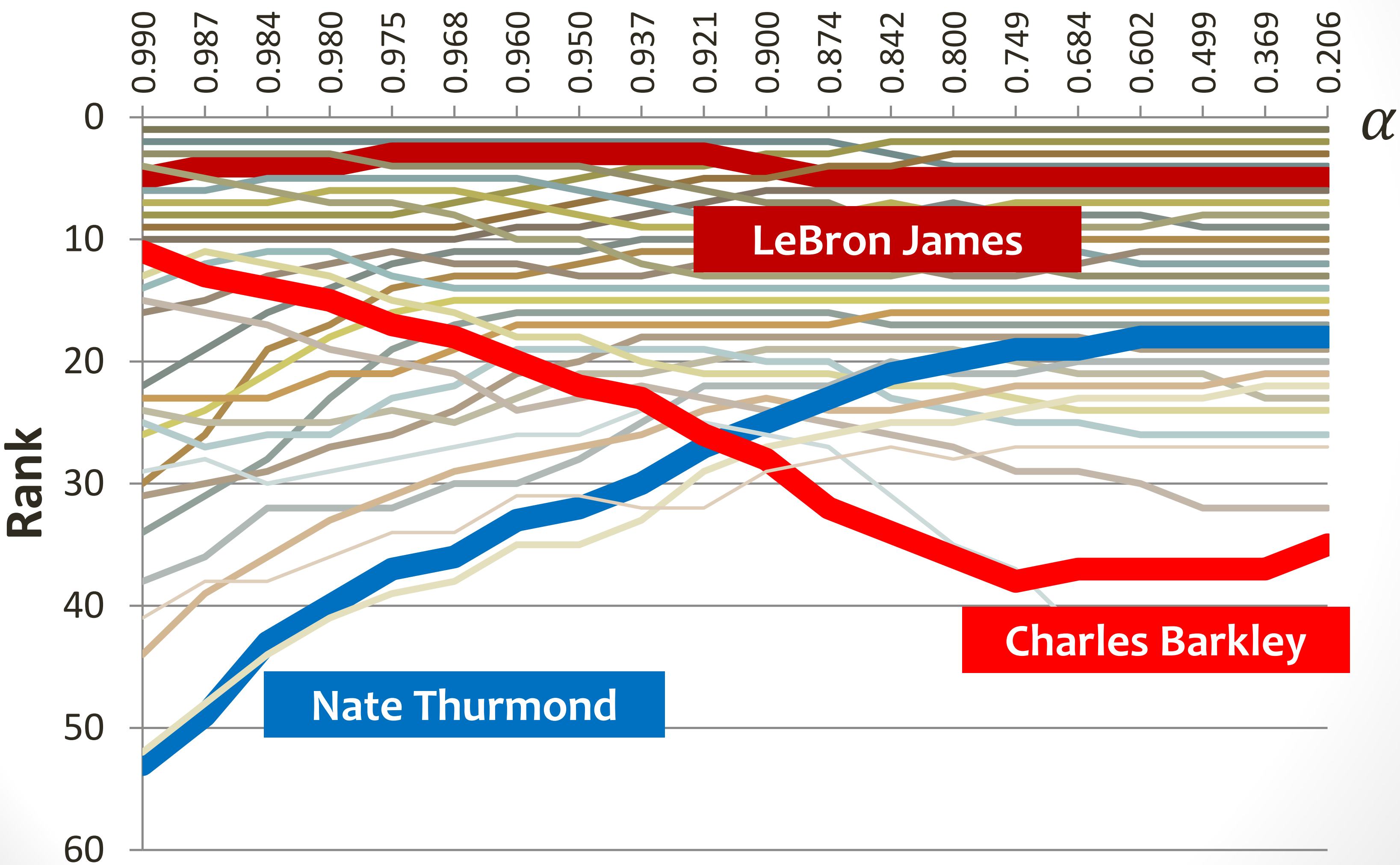


Rank	Score	Object
1	2^{0}	Chamberlain
2	2^{-1}	Jordan
3	2^{-2}	Baylor
4	2^{-3}	James
5	2^{-4}	Pettit
6	2^{-5}	Abdul-Jabbar
7	2^{-6}	Robertson
8	2^{-7}	Bird
9	2^{-8}	Johnson
10	2^{-9}	Stockton

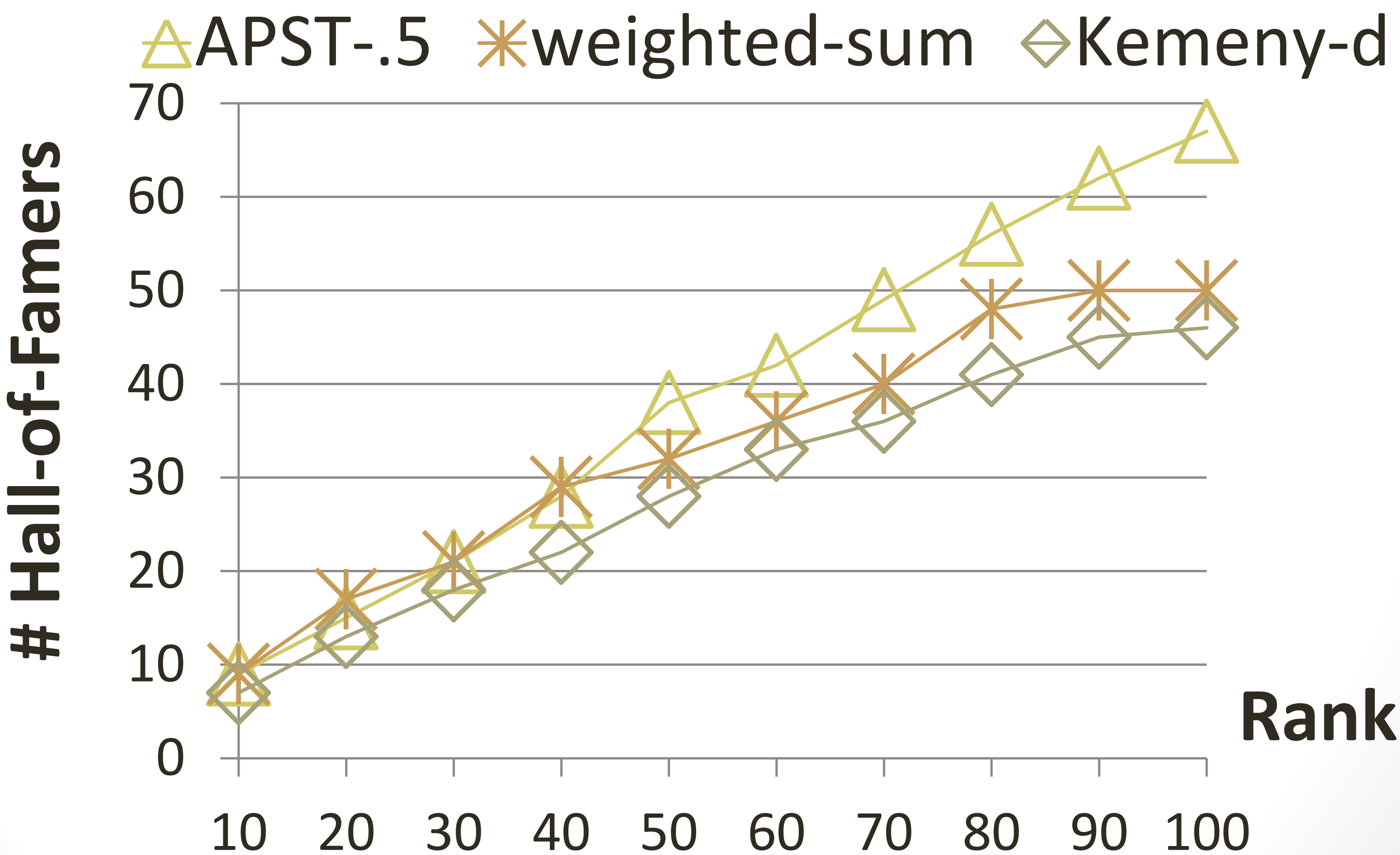
- In subspace {points, rebounds}
  - Chamberlain and Jordan each scores $(2^0 + 2^{-1})/2$
  - Baylor, James, and Pettit each scores $(2^{-2} + 2^{-3} + 2^{-4})/3$
  - …

- Scores for Chamberlain
  - 2^{-1} in {points},
  - 2^0 in {rebounds},
  - $(2^0 + 2^{-1})/2$ in {points, rebounds},
  - $.5 + 1 + 1.5/2 = 2.25$ in total.

Rank of NBA players by APST-α



Quality of Ranking



Roadmap

- Introduction
- Identifying Interesting Claims
- Ranking Objects
- **Conclusions and Future Work**

Main Contributions

- An intuitive, uniqueness-based measure of interestingness for one-of-the-few claims
- Finding interesting one-of-the-few claims from high-dimensional data
  - User-friendly problem formulation with one parameter (τ) that works for all subspaces and data distributions
  - Efficient algorithms
- A method for ranking high-dimensional objects
  - Natural: builds on the notion of claim uniqueness
  - User-friendly: a single knob (α) for effectively tuning preferences

Future Work

- Other criteria of interestingness
  - How many objects are begin considered?
    - E.g., all NBA players vs. point guards since 2000
  - How sensitive is the claim to perturbation in its parameters?
- Other types of statements
- Computational journalism project aimed at automating fact-checking and fact-finding

Thank You!

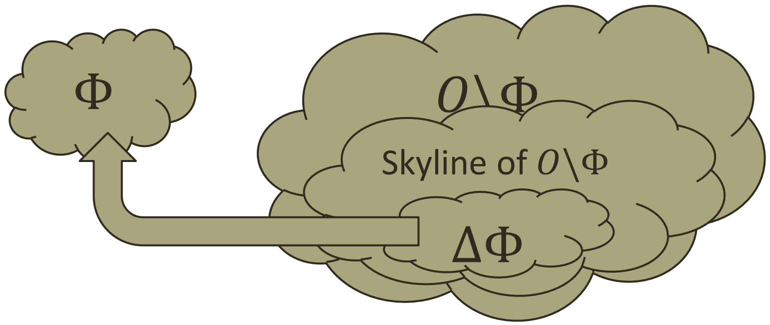
Questions?

Mining Interesting Statements

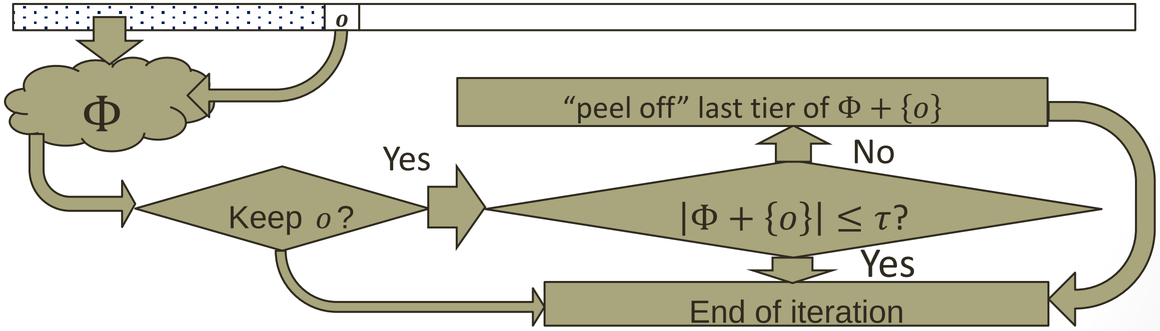
- Subgroup discovery
  - A person who *smokes* and *has family history* has a high chance of *having coronary heart disease.* [Atzmueller, 2005]
- Redescription mining
  - Russia and China are the only two countries which “have land area > 3,000,000 square miles outside of the America” or “are Permanent members of the UN security council who have a history of communism” [Parida et al., 2005]
- Prominent streak discovery
  - “LeBron James scored 35 points in nine consecutive games and joined Michael Jordan and Kobe Bryant as the only players since 1970 to accomplish the feat” [Jiang et al., 2011]

Computing top-τ skyband

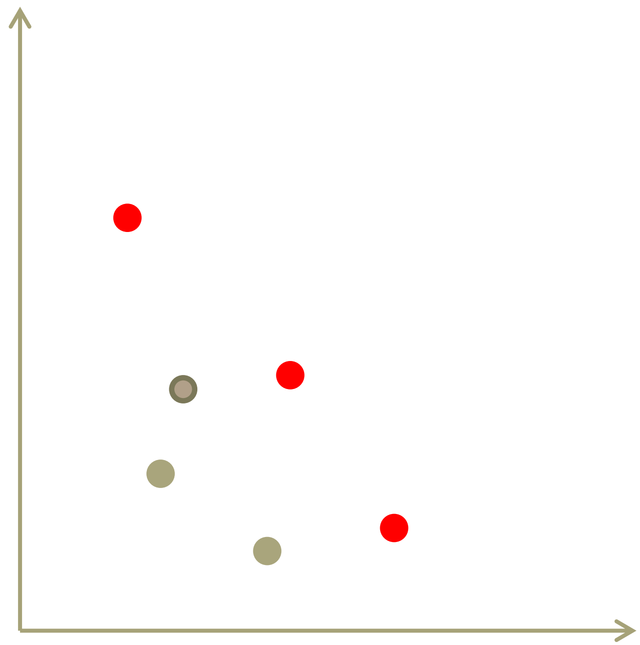
- Flow charts for the two algorithms
  - Progressive



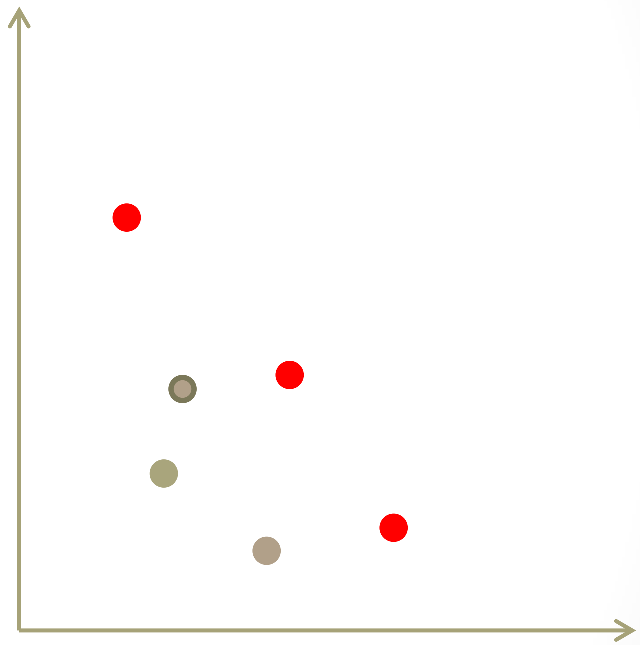
  - OnePass



Skyline Layers vs. Skyband tiers



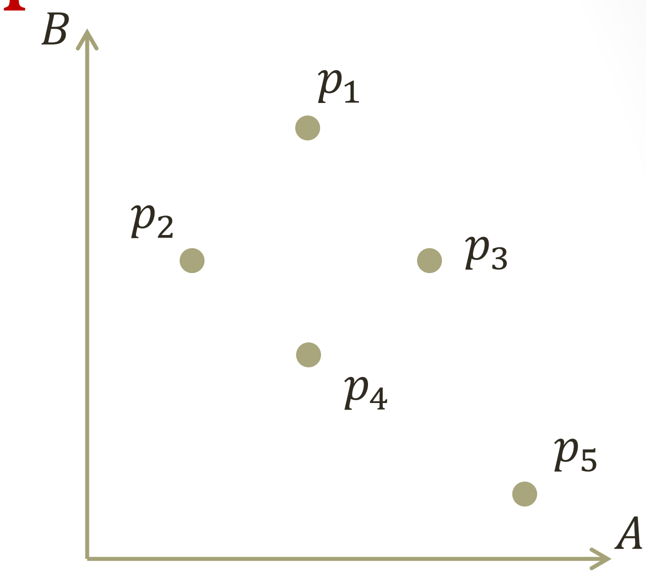
Color by skyband tiers



Color by skyline layers

APST-α Scoring Example

- $\alpha = 0.5$
- E.g. score for p_1
  - $s_{p_1,\{A\}} = (0.5^2 + 0.5^3)/2$
  - $s_{p_1,\{B\}} = 1$
  - $s_{p_1,\{A,B\}} = (1 + 0.5 + 0.5^2)/3$
  - Total score $Score_{p_1} = \sum_{S \subseteq \{A,B\}, S \neq \emptyset} s_{p_1,S}$



<table>
<tr><th>Subspace</th><th>1</th><th>0.5</th><th>0.5^2</th><th>0.5^3</th><th>0.5^4</th></tr>
<tr><td>$\{A\}$</td><td>p_5</td><td>p_3</td><td colspan="2">★p_1, p_4</td><td>p_2</td></tr>
<tr><td>$\{B\}$</td><td>★p_1</td><td colspan="2">p_2, p_3</td><td>p_4</td><td>p_5</td></tr>
<tr><td>$\{A, B\}$</td><td colspan="3">★p_1, p_3, p_5</td><td colspan="2">p_2, p_4</td></tr>
</table>

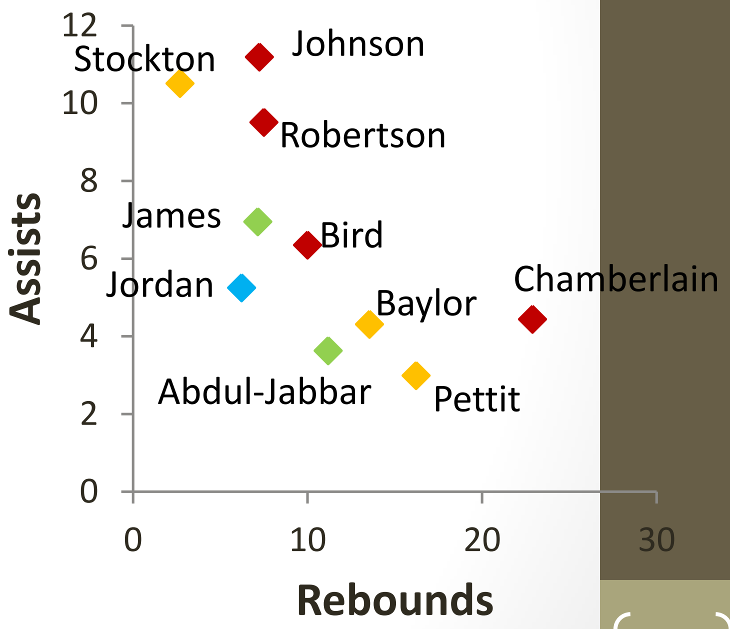
Algorithms (for one subspace)

- Progressive
  - In each iteration
    - Compute the skyline of non-top-τ skyband
    - Select objects from the skyline and add to top-τ skyband
  - Worse case running time $O(n^2)$

- OnePass
  - Examine objects in a "safe order"
  - Test dominance with current top-τ skyband
  - Worse case running time $O(n\tau)$

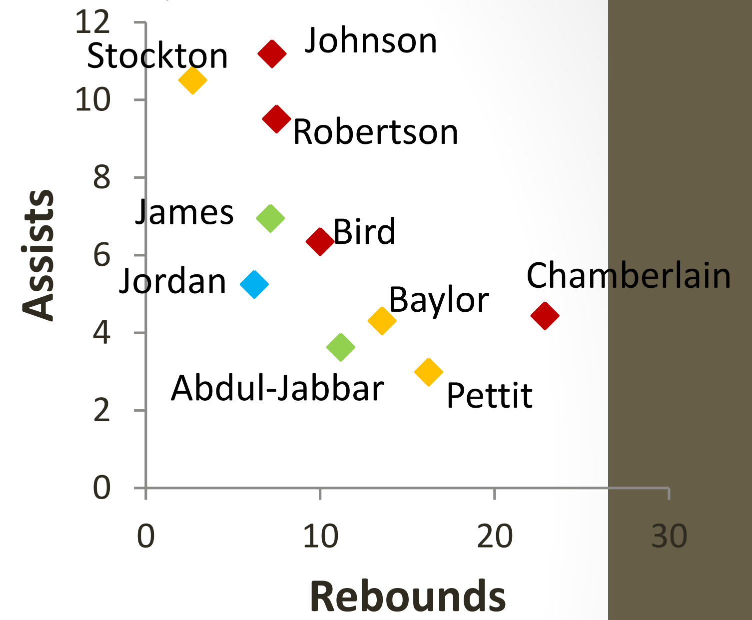
Progressive example ($\tau = 8$)

- Iteration 0:
  - Top-τ skyband = Ø
- Iteration 1:
  - Skyline of the rest = {Johnson, Robertson, Bird, Chamberlain}
  - Top-τ skyband = {Johnson, Robertson, Bird, Chamberlain} (4 players)
- Iteration 2:
  - Skyline of the rest = {Stockton, James, Baylor, Pettit}
  - Top-τ skyband = {Johnson, Robertson, Bird, Chamberlain; Stockton, Baylor, Pettit} (7 players)
- Iteration 3:
  - Skyline of the rest = {James, Abdul-Jabbar}
  - Top-τ skyband = {Johnson, Robertson, Bird, Chamberlain; Stockton, Baylor, Pettit; James, Abdul-Jabbar} (9 players)
  - Excessing $\tau = 8$, return Top-τ skyband at the end of previous iteration (iter. 2)



OnePass example ($\tau = 8$)

- Examine points in order: Johnson, Stockton, Robertson, James, Bird, Jordan, Chamberlain, Baylor, Abdul-Jabbar, Pettit



Player	# dom.
Johnson	0

Player	# dom.
Johnson	0
Stockton	1

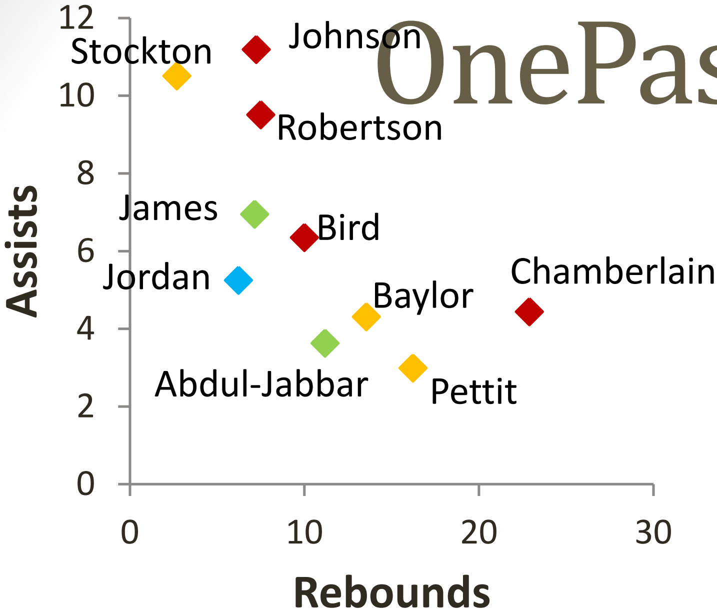
Player	# dom.
Johnson	0
Stockton	1
Robertson	0

Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2

Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2
Bird	0

Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2
Bird	0
Jordan	3

OnePass example ($\tau = 8$)



Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2
Bird	0
Jordan	3
Chamberlain	0

Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2
Bird	0
Jordan	3
Chamberlain	0
Baylor	1

Player	# dom.
Johnson	0
Stockton	1
Robertson	0
James	2
Bird	0
~~Jordan~~	~~3~~
Chamberlain	0
Baylor	1
Abdul-Jabbar	2

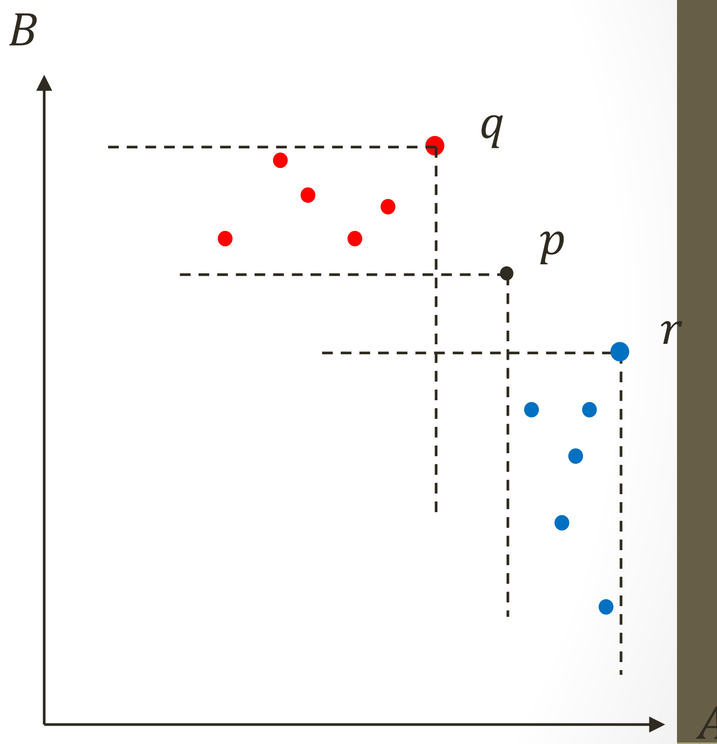
Player	# dom.
Johnson	0
Stockton	1
Robertson	0
~~James~~	~~2~~
Bird	0
Chamberlain	0
Baylor	1
~~Abdul-Jabbar~~	~~2~~
Pettit	1

Lattice Traversal

- Going from low dimension to high dimension…
  - Skyline points of $\{A\}$ also go in to skyline of $\{A, B\}$ (with distinct value condition)
  - If skyline of $\{A\}$ has more than τ distinct points, any subspace containing A must have empty top-τ skyband
  - If the union of skylines from subspaces $\{A\}$ and $\{B\}$ contains more than τ distinct points, top-τ skyband of subspace $\{A, B\}$ is empty

Special case for kemeny

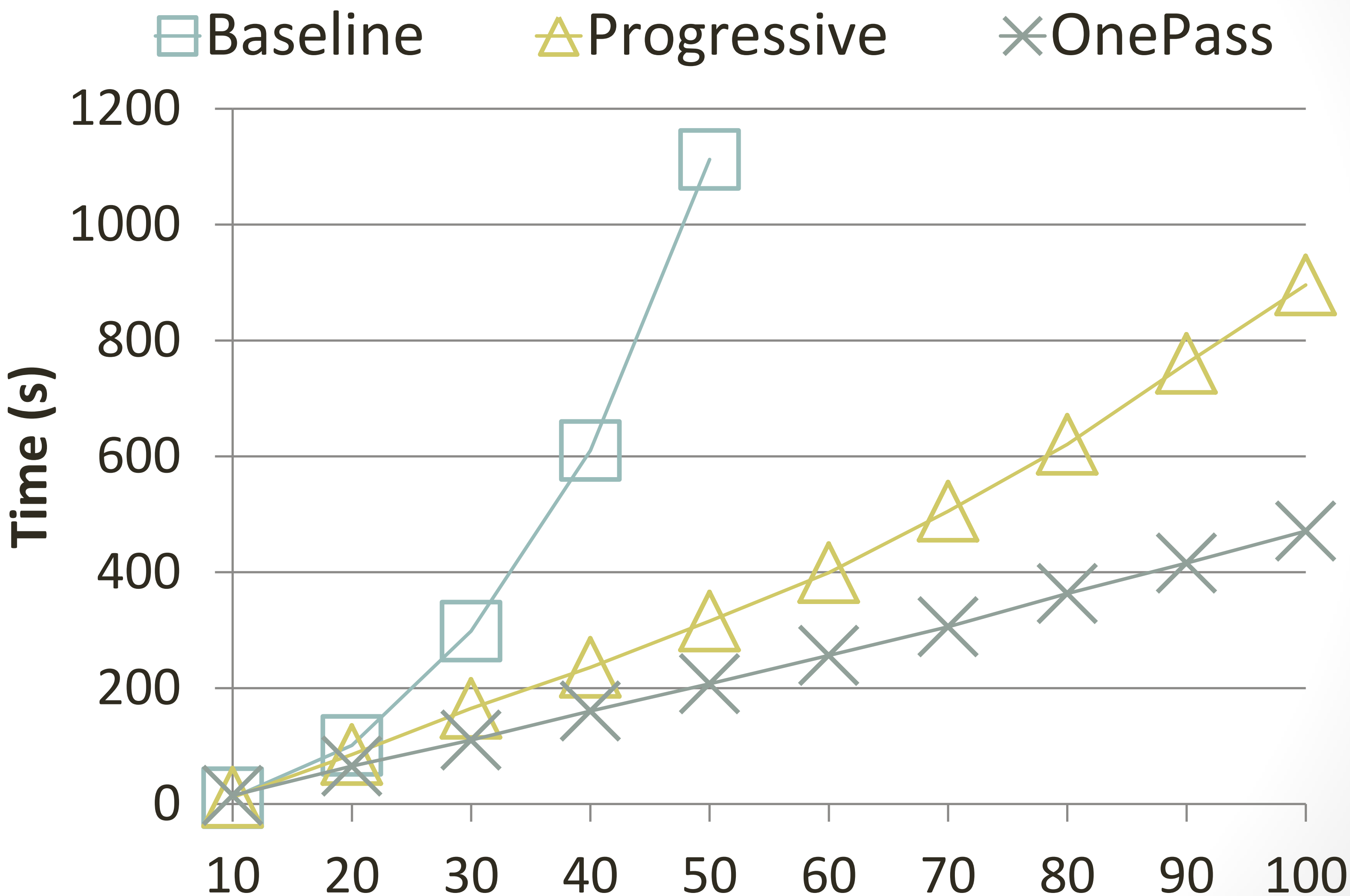
- Sort on A
  - rBBBBBpqRRRRR
- Sort on B
  - qRRRRRprBBBBB
- Kemeny cannot rank p properly



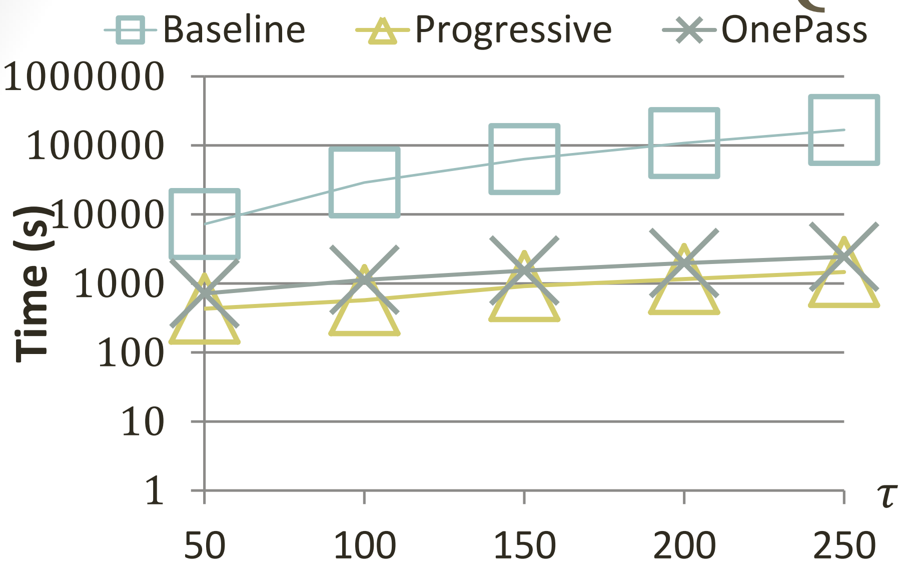
Computing APST scores

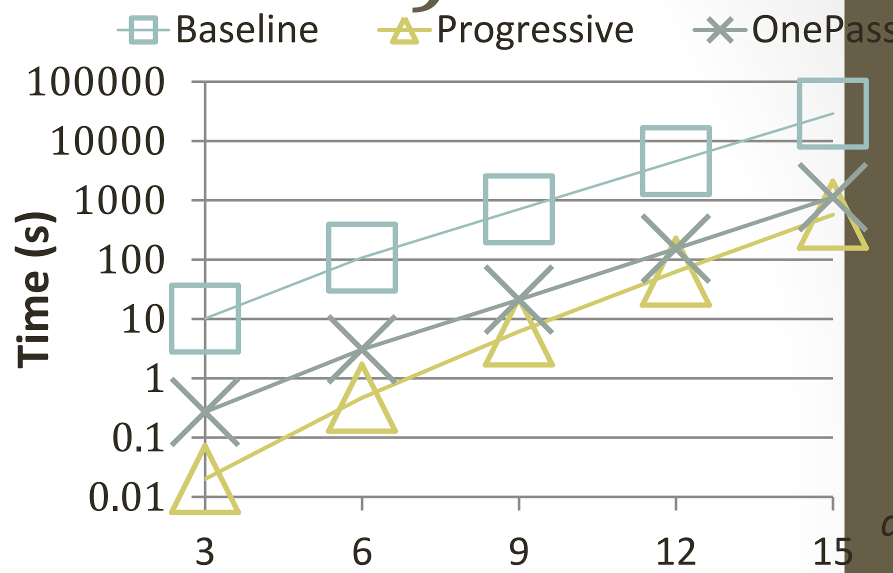
- Computing exact scores take $O(n^2)$ time for each subspace
- Given any error $\epsilon > 0$, computes the score of each object such that $\hat{\Gamma}(o) \in (\Gamma(o) - \epsilon, \Gamma(\mathrm{o})]$.
- Approximated using Progressive or OnePass
- Intuition: in a subspace, if the score of each object in the next tier of skyband is small enough, there's no need to compute any successive layers

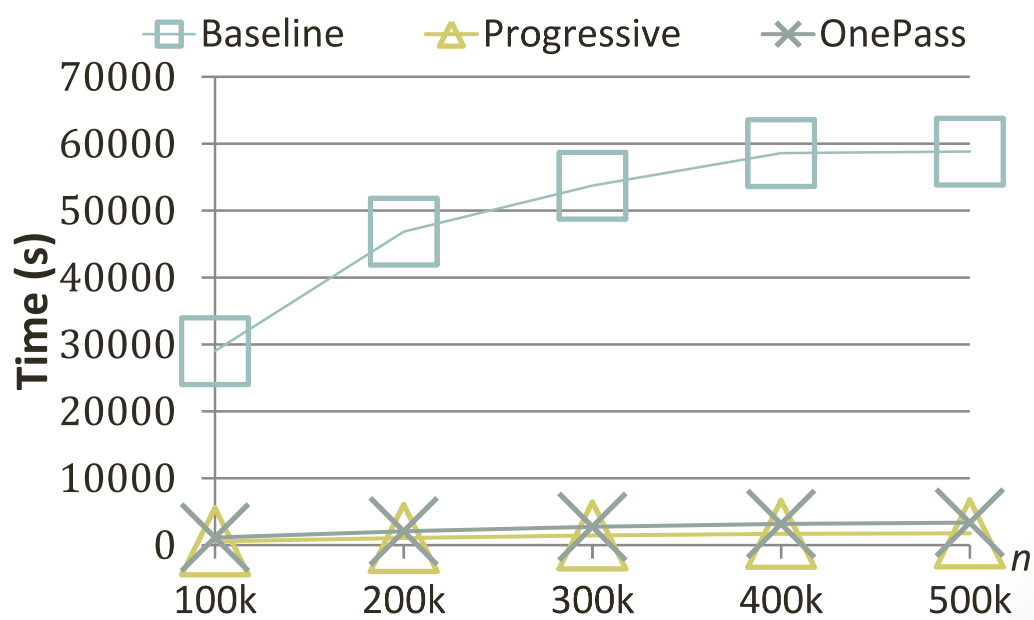
Performance (NBA)



Performance (correlated)







Performance (Independent)