School of
Information Systems

Opportunities for Spatial Database Research in the Context of Preference Queries

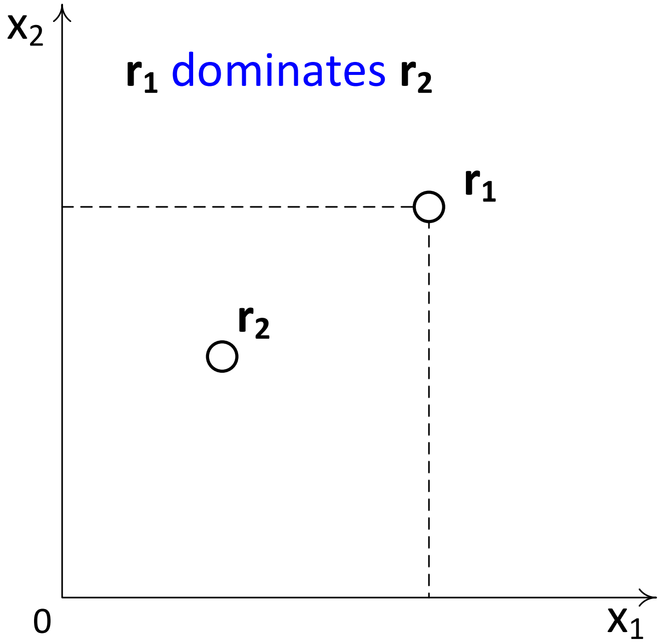
[Keynote Speech]

Kyriakos Mouratidis

Singapore Management University

Introduction

- Paradigms to identify options of preference in a **multi-objective** setting:
  - Dominance-based: **Skyline** (and **k-skyband**)
  - Ranking by utility: Top-k query (input: **preference vector w** of *d* weights; **utility** of an option defined as the weighted sum of its attributes)



Preference vector **w** = (0.2, 0.8)

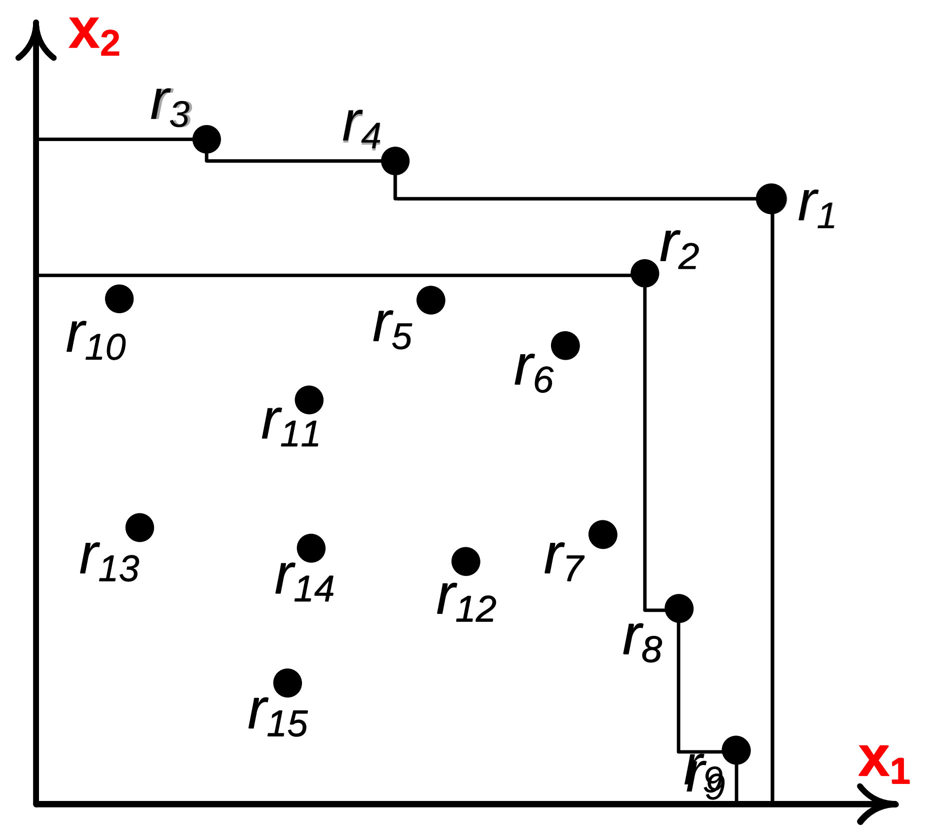
Utility of option **r** = (x_1, x_2) defined as:

$U(\mathbf{r}) = 0.2 \cdot x_1 + 0.8 \cdot x_2$

Top-k: the *k* options with highest utility

Skyline and Skyband

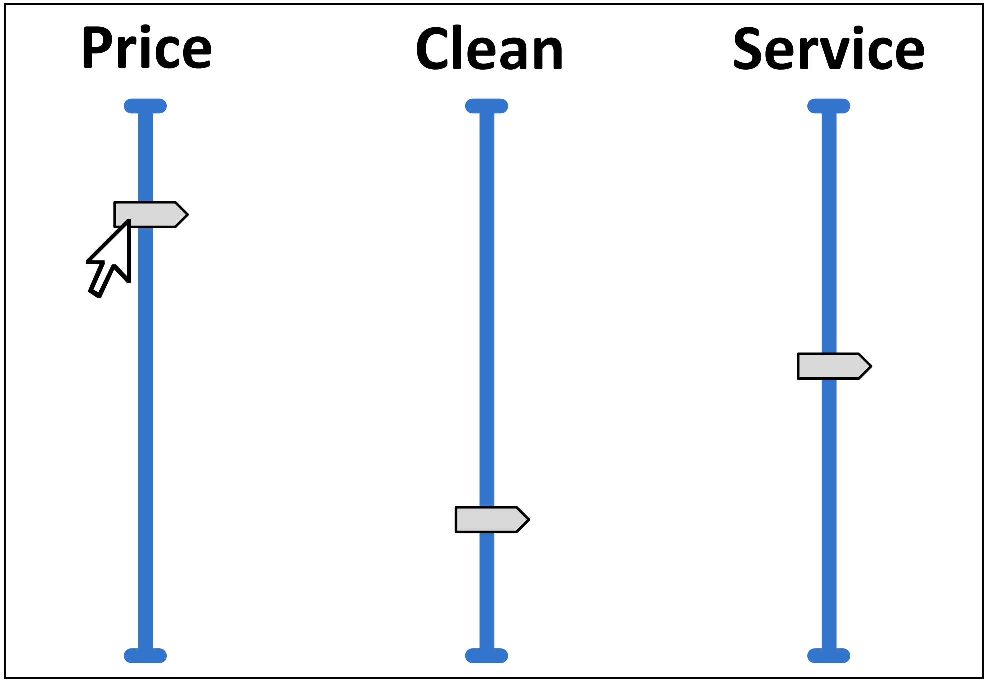
- **Skyline**: all opts. that aren't dominated
- Includes top-1 ∀ **w**
- **k-skyband**: all opts. not dominated by k or more others
- Includes top-k ∀ **w**



Traditional top-k query

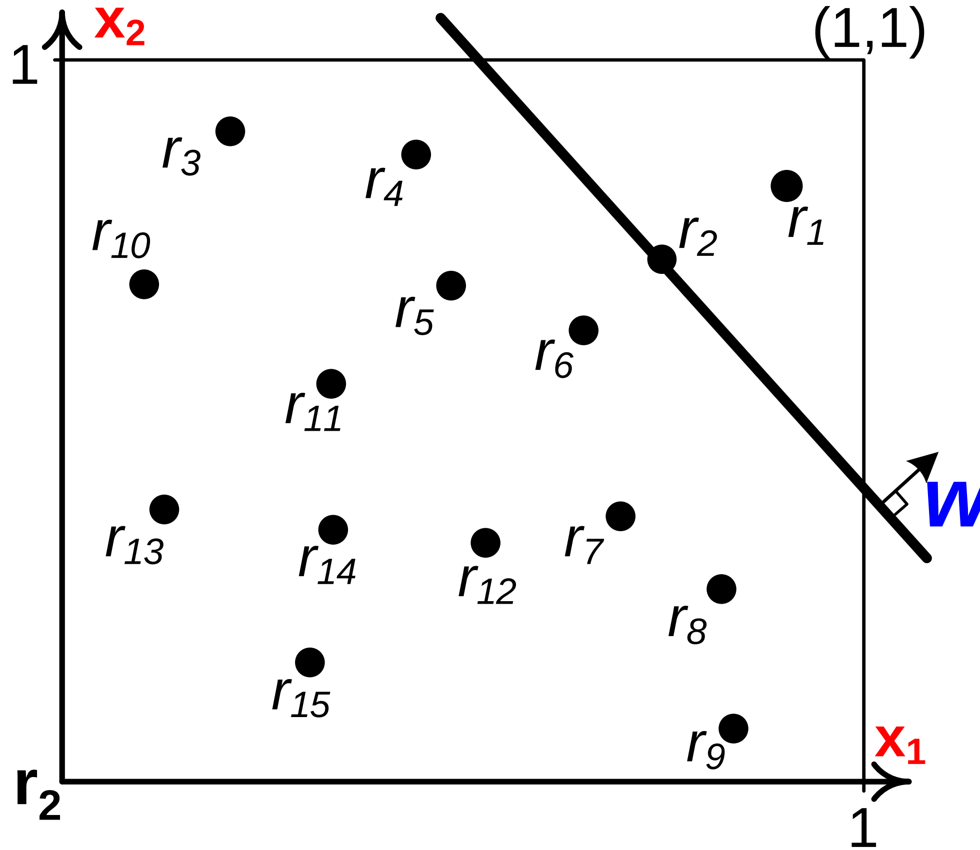
- Top-*k* query: shortlists **top options** from a set of alternatives
- E.g. TripAdvisor.com
  - rate (and browse) hotels according to price, cleanliness, location, service, etc.
- A user's criteria: ***price***, ***cleanliness*** and ***service***, with different ***weights***

Weights could be captured by slide-bars:



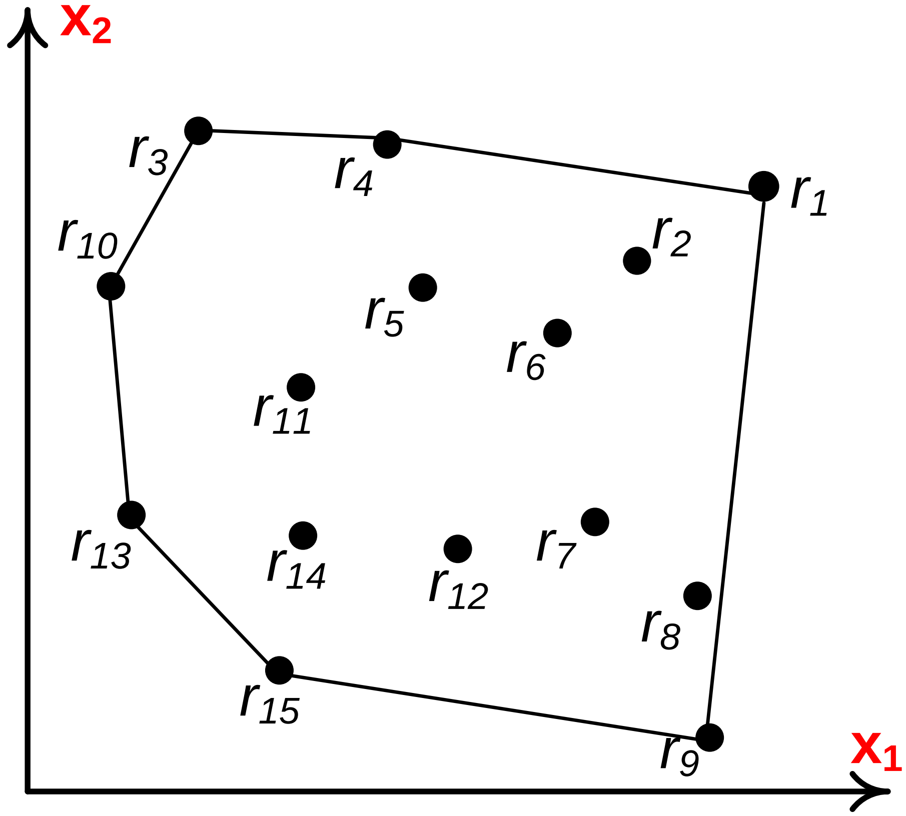
Top-k as sweeping the data space

- Assume all **weights** are **positive**
- …and each **option attribute** is in range [0,1]
- Example for d = 2 (showing: **<u>option space</u>**)
- **Sweeping line** normal to vector **w**
- Sweeps from top-corner (1,1) towards origin
- Order an option is met ↔ order in ranking!
  - E.g. top-2 = { $\mathbf{r_1}$, $\mathbf{r_2}$ }
- At current position: ∀ option above (below) the line, higher (lower) score than $\mathbf{r_2}$



Relationship to Convex Hull

- **Convex Hull**: The smallest convex polytope that includes a set of points (options)
- Fact: The top-1 option for **any** query vector is on the hull!
  - [Dantzig63]: LP text



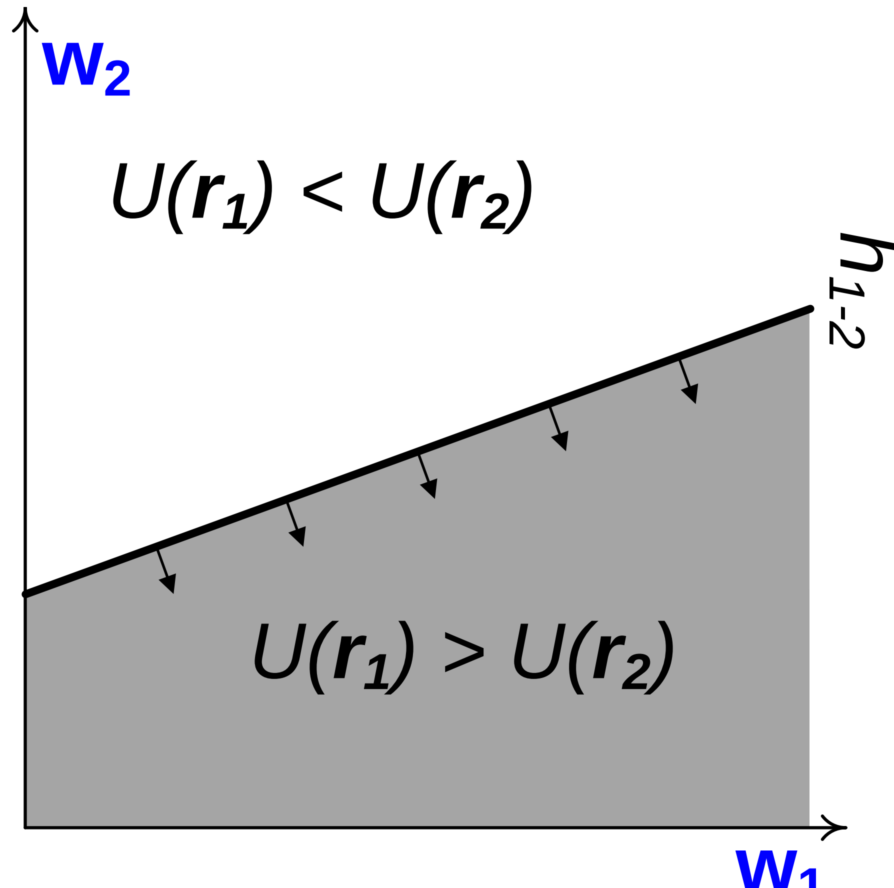
Utility order and equivalent half-space

- $U(\mathbf{r_1}) = U(\mathbf{r_2}) \leftrightarrow$

a **hyper-plane** in pref. domain

- $U(\mathbf{r_1}) > U(\mathbf{r_2}) \leftrightarrow$

a **half-space** in pref. domain

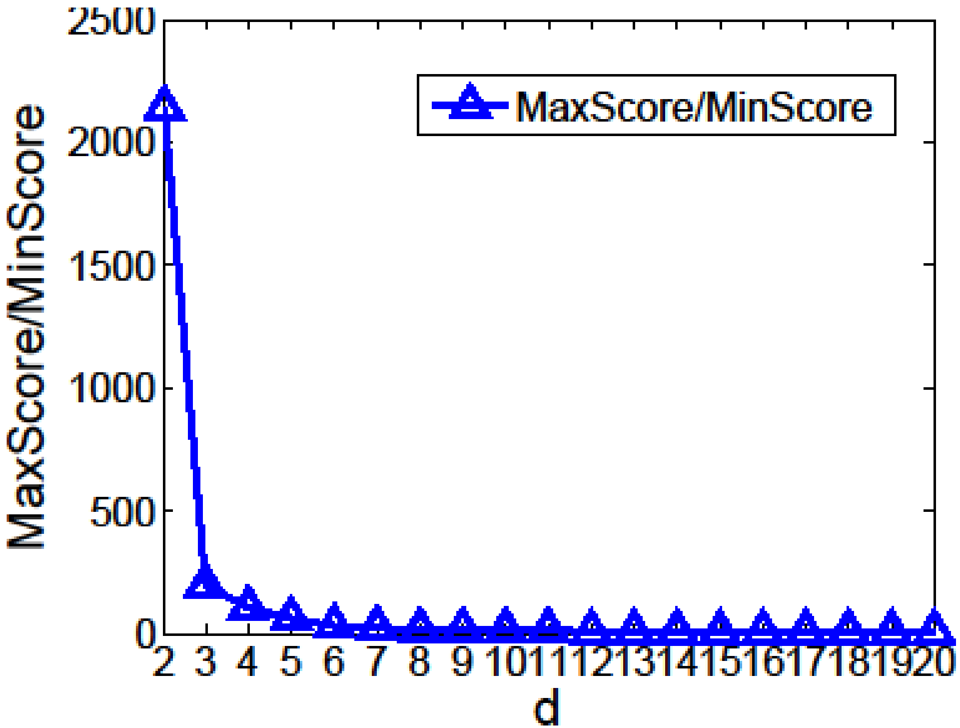


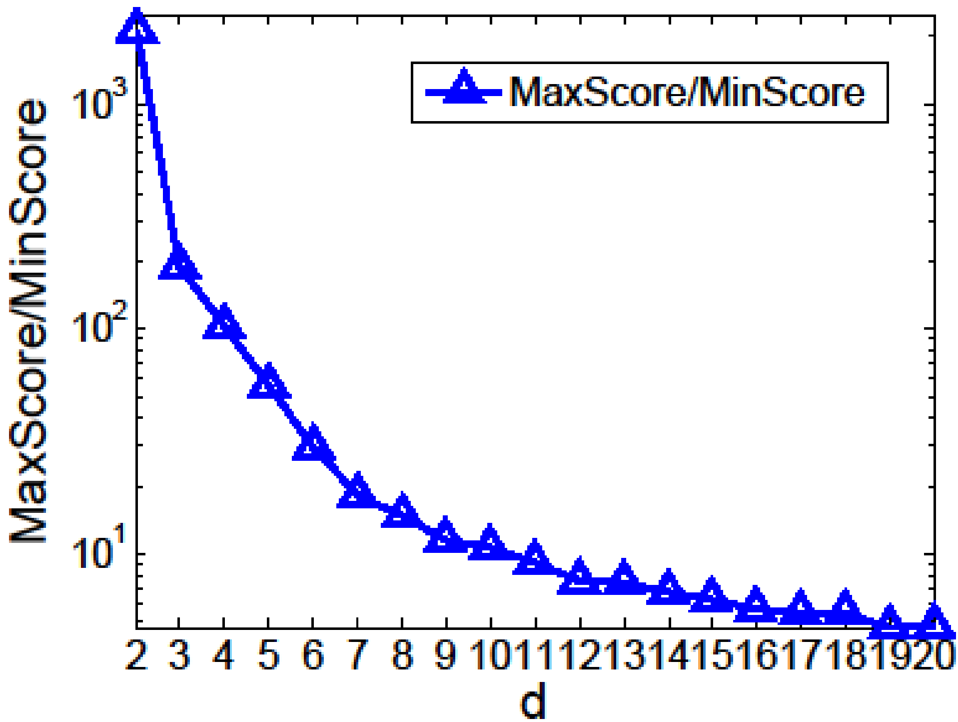
Top-k in High-D?

- Unless the data are very sparse or overly correlated, top-k is meaningless in more than 5-6 dimensions!
- As d grows, the **highest score** across the dataset approaches the **lowest score**!
- I.e. ranking by score no longer offers distinguishability $\leftrightarrow$ looses its usefulness
- Behaviour very similar to nearest neighbor query, known to suffer from the dimensionality curse

Top-k in High-D?

- IND data
- …of fixed cardinality n = 100K
- …we vary data dimensionality





*m*IR problem

- Tang, Mouratidis, Han: "*On m-Impact Regions and Standing Top-k Influence Problems*". SIGMOD'21
- *m-Impact Regions Problem (mIR):* Given an option set *D*, a user set *W*, and a positive integer *m*, the *mIR* problem is to compute the maximal region *R* in option space, inside which any (existing or hypothetical) option ***r*** is in the top-k set of at least *m* users

*m*IR example

Option set: hotels

Attributes (dimensions): Value, Service

User set includes 4 users

Hotel	Value	Service
$\boldsymbol{p}_1$	0.57	0.24
$\boldsymbol{p}_2$	0.81	0.42
$\boldsymbol{p}_3$	0.93	0.91
$\boldsymbol{p}_4$	0.18	0.43
$\boldsymbol{p}_5$	0.34	0.75

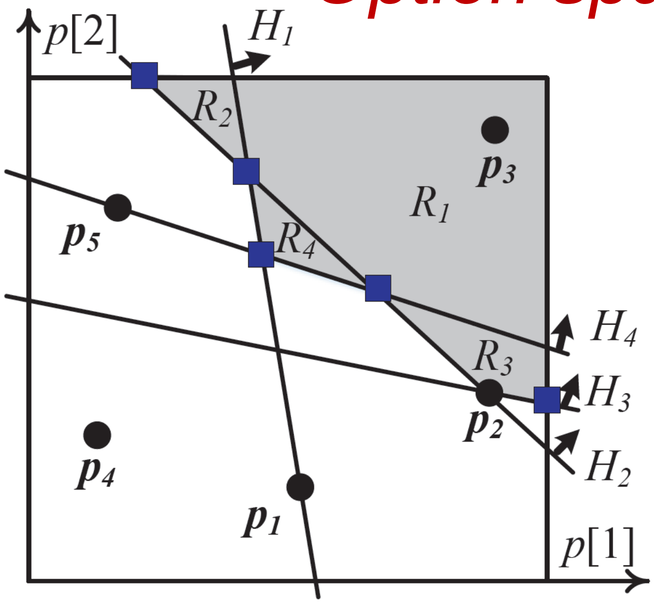
User	$w[1]$	$w[2]$	k	Top-k result
$\boldsymbol{w}_1$	0.76	0.24	3	$\{\boldsymbol{p}_3, \boldsymbol{p}_2, \boldsymbol{p}_1\}$
$\boldsymbol{w}_2$	0.49	0.51	2	$\{\boldsymbol{p}_3, \boldsymbol{p}_2\}$
$\boldsymbol{w}_3$	0.09	0.91	3	$\{\boldsymbol{p}_3, \boldsymbol{p}_5, \boldsymbol{p}_2\}$
$\boldsymbol{w}_4$	0.28	0.72	2	$\{\boldsymbol{p}_3, \boldsymbol{p}_5\}$

(a) Option set and User set

m = 3

→

Option space



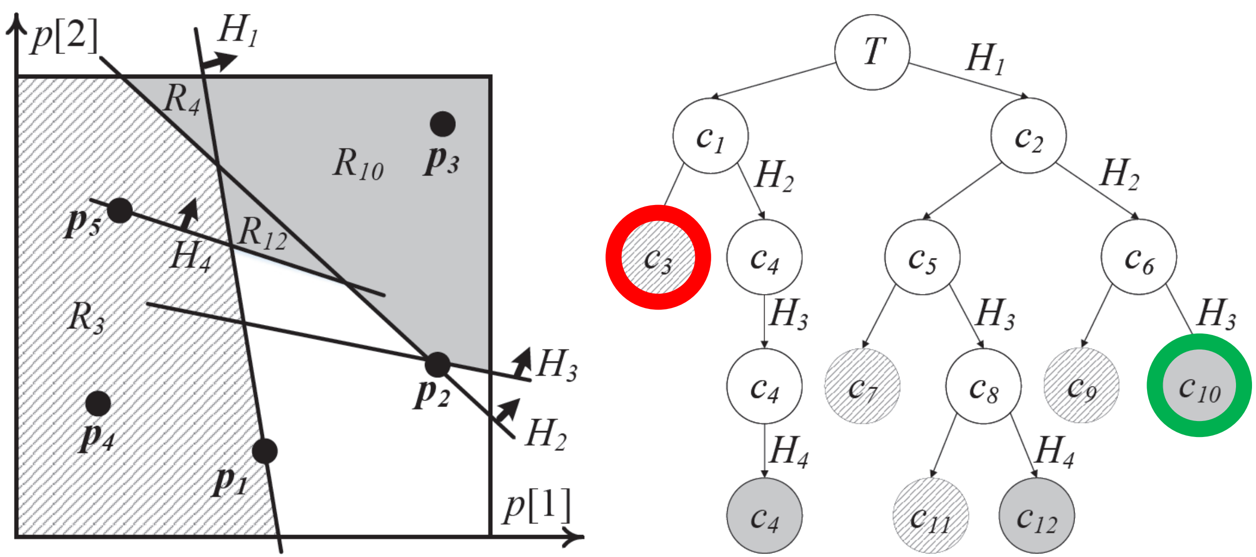
(b) *m*IR result (shown shaded)

Algorithmic basis for *m*IR

- Let c_i be the top-k-th score for user $\mathbf{w_i}$ in *D*
- $\boldsymbol{r}$ is in top-k set of $\mathbf{w_i} \Leftrightarrow U_{\mathbf{wi}}(\mathbf{r}) \geq c_i$
- …which is a half-space in the pref. space, called the *impact half-space* of $\mathbf{w_i}$
- Basic idea:
  - produce the impact half-space for each user
  - partition the pref. space by these half-spaces
  - report the partitions (*cells*) included in ≥ *m* impact half-spaces
  - complexity……… $O(|W|^d)$

Algorithmic basis for *m*IR

- Insert half-spaces one by one into a ***cell tree***
- Early reporting and pruning possible
- Still too slow



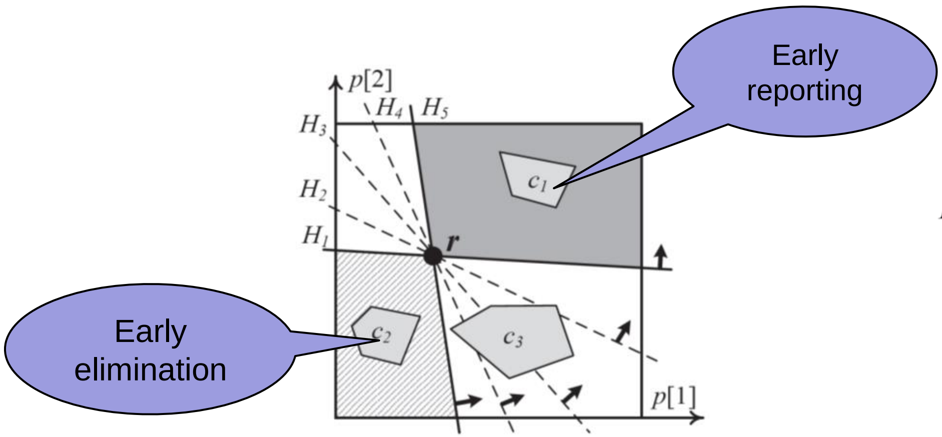
(a) Halfspace arrangement

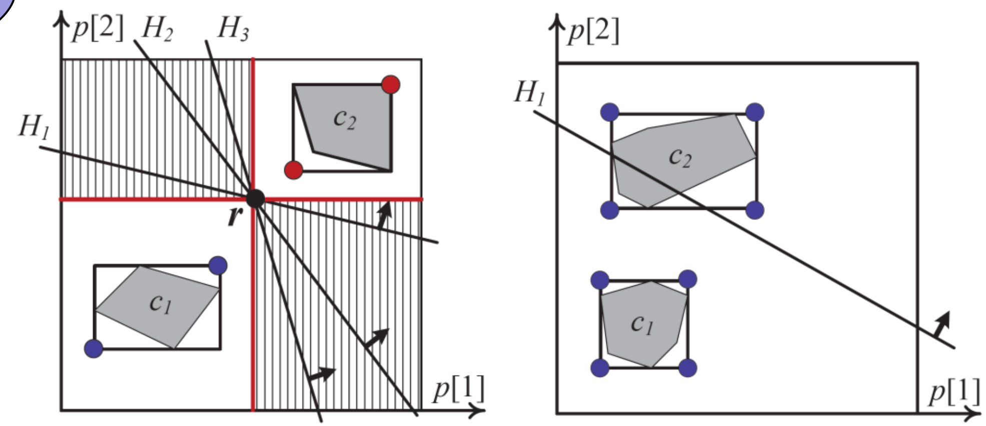
(b) Binary tree representation

Early reporting

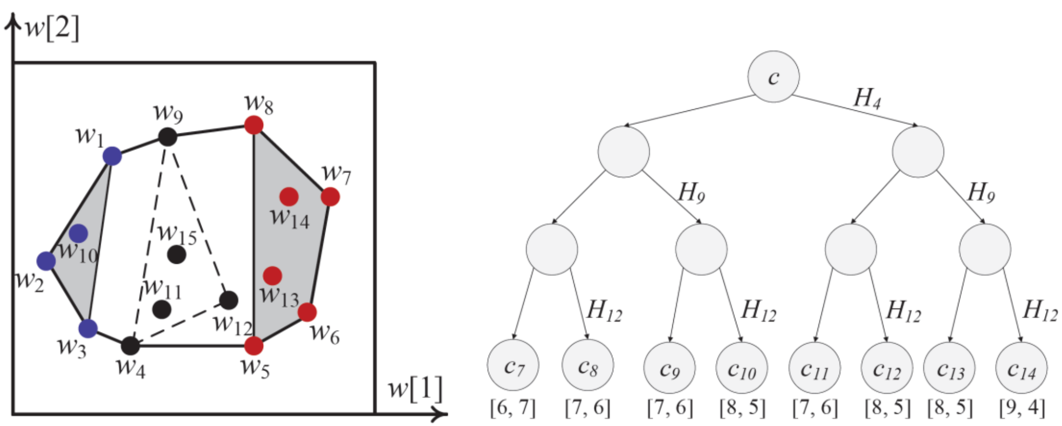
Early elimination

Snapshots of our methodology





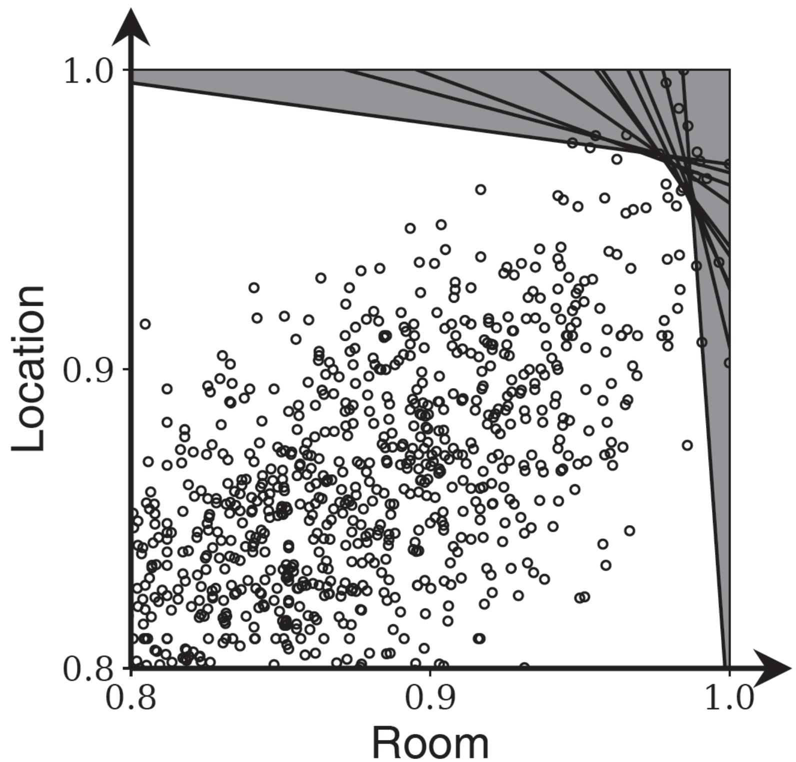
(a) Group testing w.r.t. cells (b) Classifying user groups



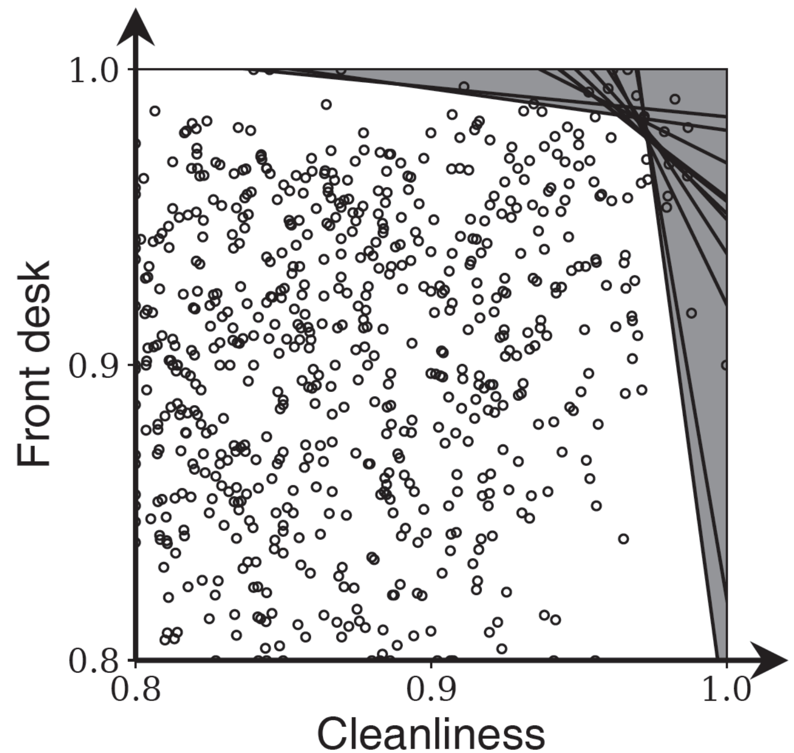
(a) G in weight space (b) Arrangement (sub-)tree of *c*

Case study

- TripAdvisor data (137,563 users and 1,850 hotels)
- $d = 2$, $k = 10$, $m = 0.5 \cdot |U|$



(a) Room-location product space



(b) Cleanliness-front desk product space

Marrying top-k with skyline

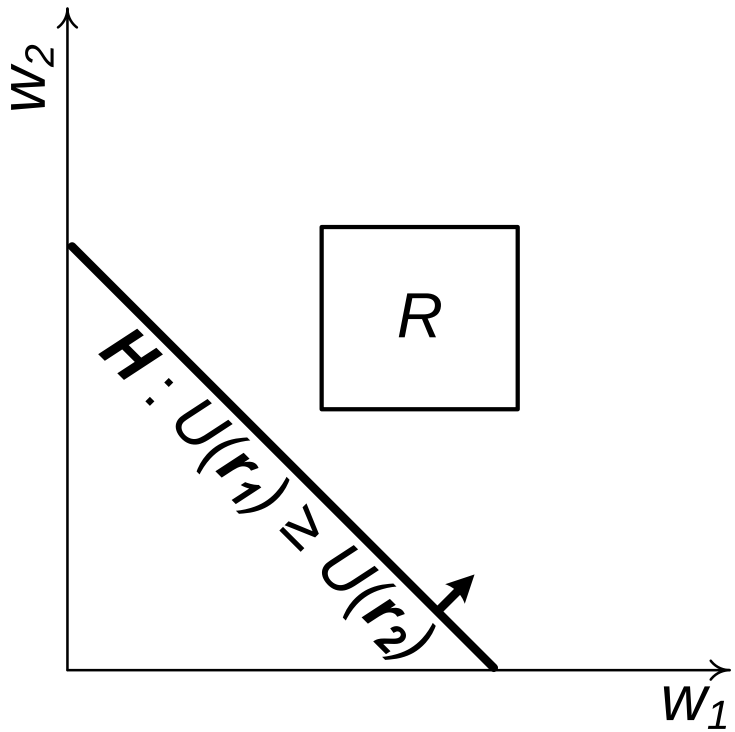
- Mouratidis, Li, Tang: "Marrying Top-k with Skyline Queries: Relaxing the Preference Input while Producing Output of Controllable Size". SIGMOD'21
- Skyline: not personalized, no output-size control
- Top-k: whether mined or user-input, **w** is only an **estimate** $\Rightarrow$ too rigid ranking
- Strong requirements:
  - Personalized
  - Output-size specified – (**OSS**)
  - Flexible preference specification

Previous operators

Operator	Personalized	OSS	Flexible Input
Skyline/k-Skyband	✘	✘	✔
Top-k	✔	✔	✘
OSS skylines	✘	✔	✔
Regret-minimizing sets	✘	✔	✔
Fixed-region techniques	✔	✘	✔
Proposed (ORD and ORU)	✔	✔	✔

Fixed-region (appr. 1): r-skyband

- Consider opts. $\mathbf{r_1}$, $\mathbf{r_2}$ and a region R in pref. domain
- $\forall \mathbf{w}$ in R, $U(\mathbf{r_1}) > U(\mathbf{r_2})$: $\mathbf{r_1}$ **r-dominates** $\mathbf{r_2}$
- **r-skyband**: options r-dominated by <k others

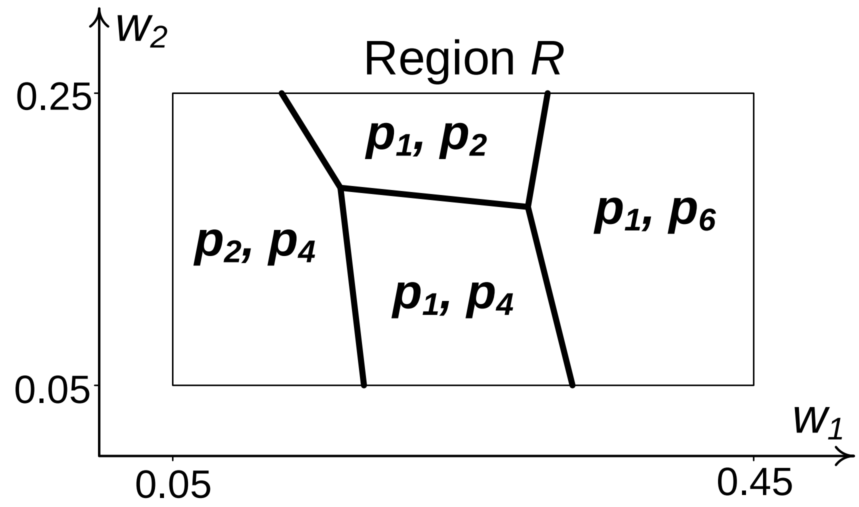


Fixed-region (appr. 2): Uncertain top-k

- Given: region *R* in pref. space
- **UTK**: report all possible top-k opts. when **w** $\in$ *R*

Hotel	Svc.	Cln.	Loc.
p_1	8.3	9.1	7.2
p_2	2.4	9.6	8.6
p_3	5.4	1.6	4.1
p_4	2.6	6.9	9.4
p_5	7.3	3.1	2.4
p_6	7.9	6.4	6.6
p_7	8.6	7.1	4.3

Dataset

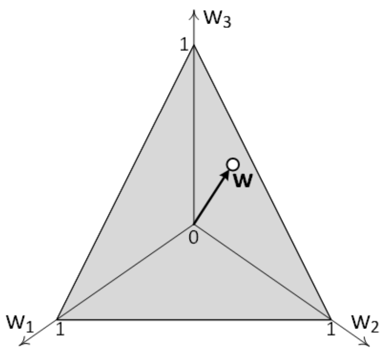


UTK output for $k = 2$
(in preference space)

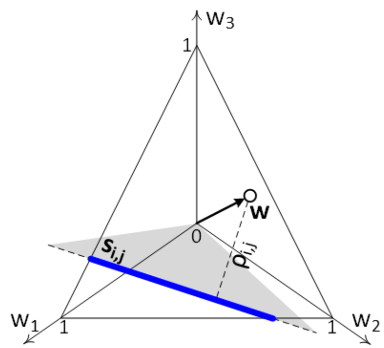
Problem definition: ORD & ORU

- Input: vector (seed) **w**, value *k*, desired output size *m*
- *ρ*-dominance: a record *ρ*-dominates another if it has higher utility $\forall$ pref. vector within radius *ρ* from **w**
- **ORD**: report the options that are *ρ*-dominated by fewer than k others, *for the minimum ρ that produces m records in the output*
- Stopping radius *ρ* unknown to the algo. in advance
- The user and application are both transparent to *ρ*
- **ORU**: report the options that are in top-k result for at least one pref. vector within distance *ρ* from **w**, *for the minimum ρ that produces m records in the output*

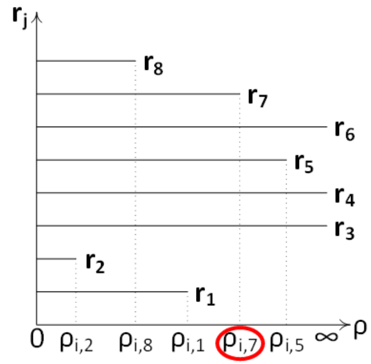
Snapshots of our methodology



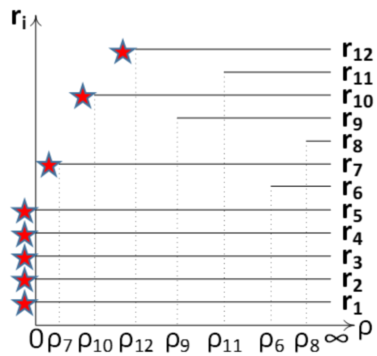
Pref. domain



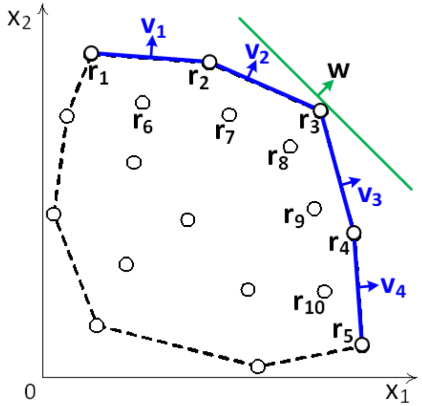
ρ-dominance



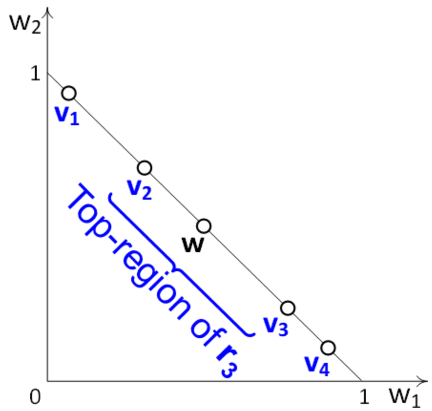
Inflection radius



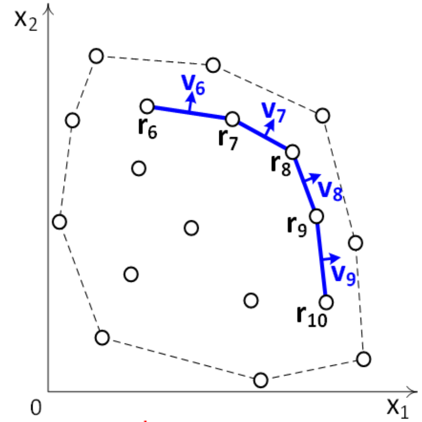
ρ-skyband vs. ρ



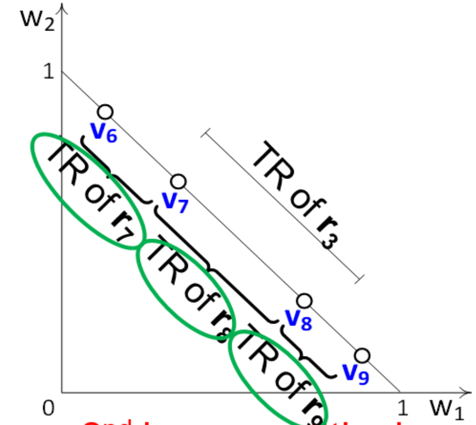
Upper hull & facet norms



1st layer partitioning



2nd hull layer

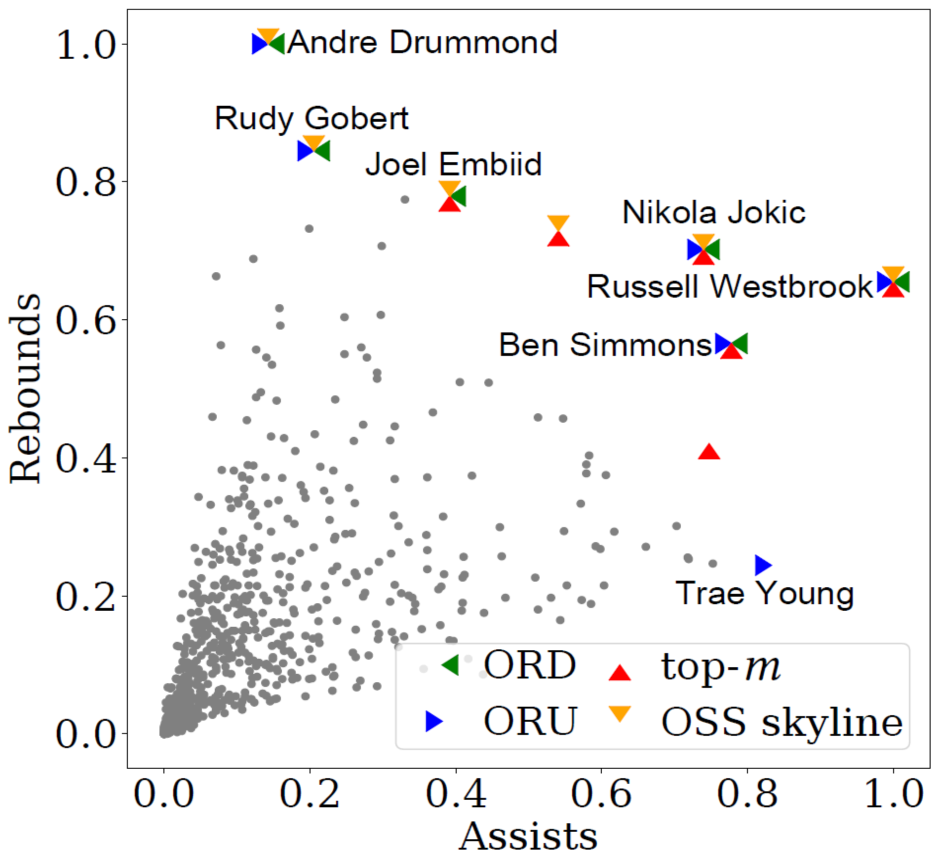


2nd layer partitioning

Case study

- NBA 2018-19 statistics (k = 2, m = 6)

w = (0.49, 0.51)



ORD/ORU report distinct results from past approaches (and from each other)

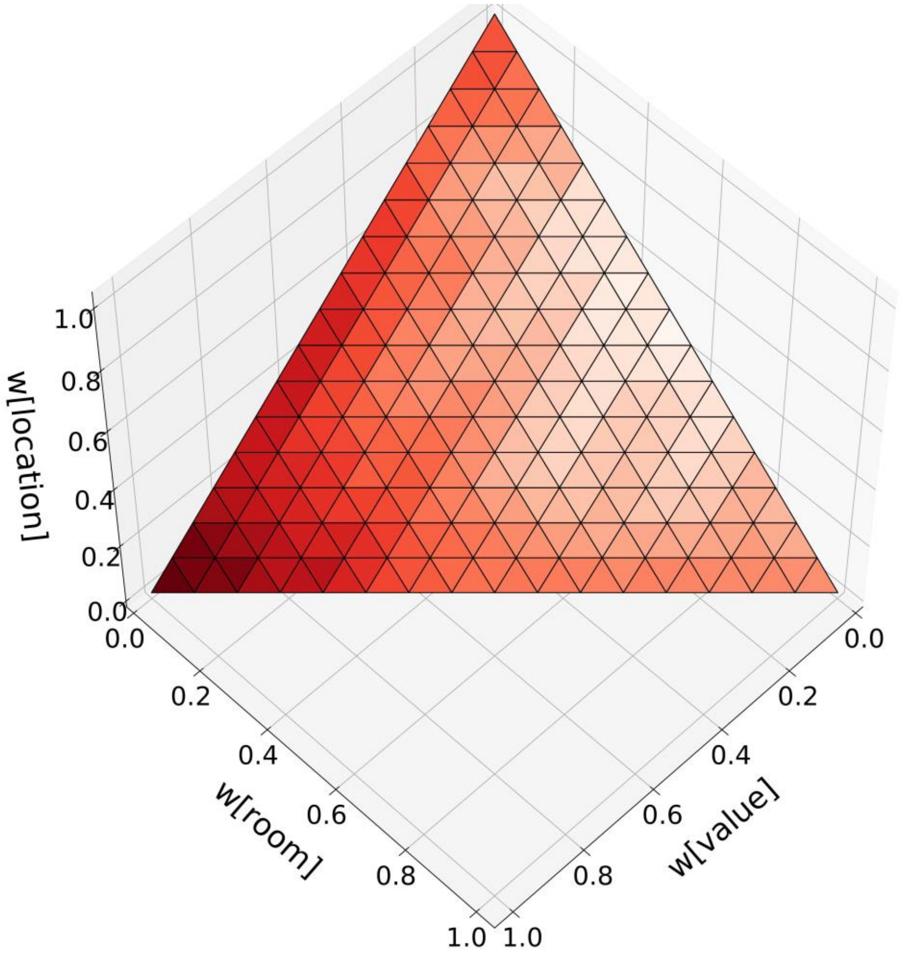
ORD/ORU report records that are particularly strong for alternative, very similar preferences to the seed **w**

Quantifying Dataset Competitiveness

- Mouratidis, Li, Tang: "*Quantifying the Competitiveness of a Dataset in Relation to General Preferences*". VLDBJ, to appear
- Change of focus… to the dataset itself
- Objective: assess its competitiveness w.r.t. different possible preferences
- We define measures of competitiveness, and represent them in the form of a heat-map in the pref. space

Case study (TA)

- TripAdvisor 1,850 hotels
- d = 3 (loc/n, room, value)
- Pref. space: simplex
- Partition into cells
- Focus on the **fringe** of *D*:
  - Use r-skyband
- Utility-based measure
  - MaxMin_k
  - for-granted utility that any of the possible top-k hotels would have for any preference in the cell



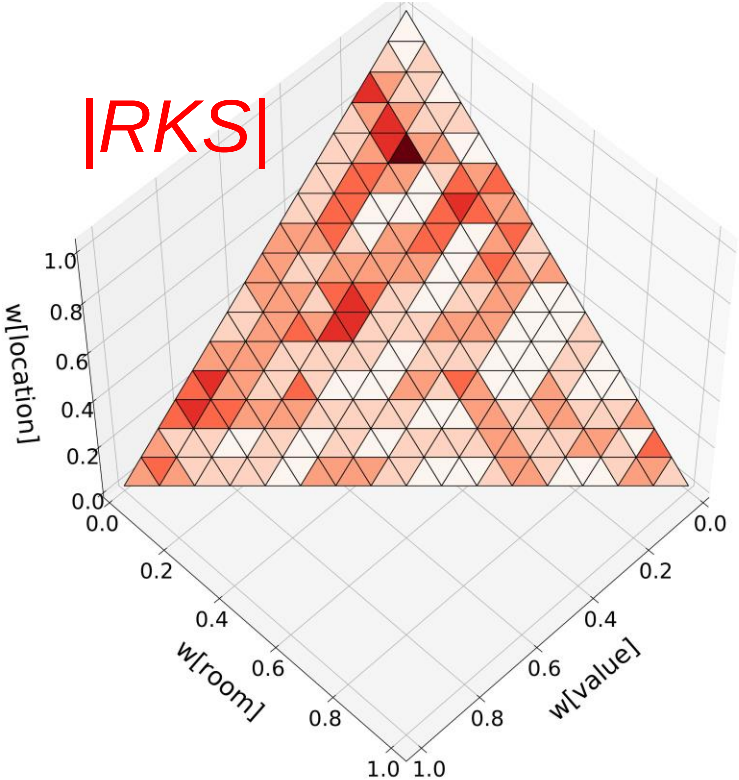
Utility-based heat-map

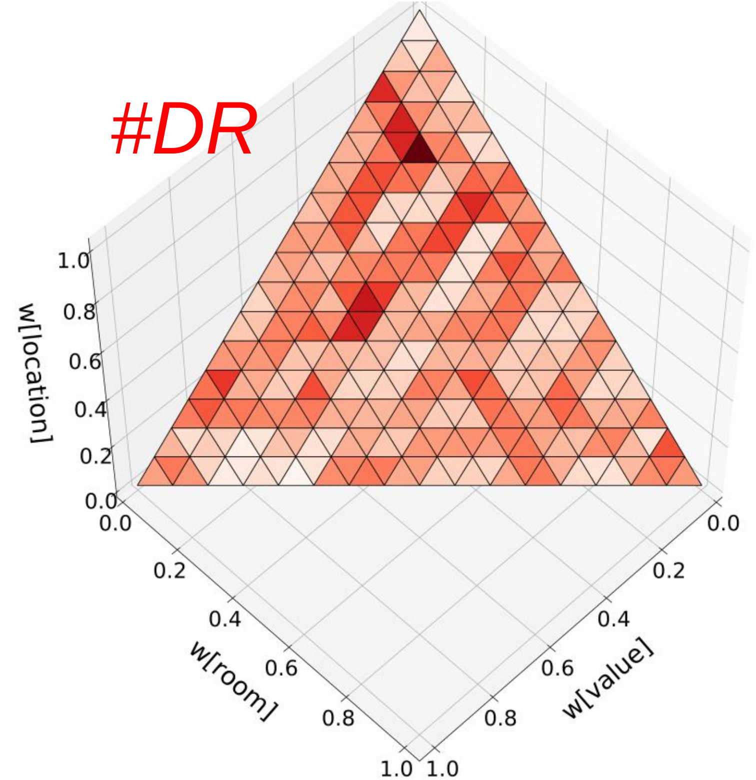
Applications

- **Market Analysis**: hottest cells is where the market's strength lies
  - i.e., the hotel market caters best for users who prioritize value over room quality and location.
- **Business Development**: hottest cells indicate market saturation
  - e.g., coldest cells may indicate sweet spots for a new hotel
- **Identifying outstanding options in the market**
- **$\mathbf{MaxMin_k}$ can speed up top-k computation**
- First two applications benefit when the distribution (or a sample) of user preferences is known

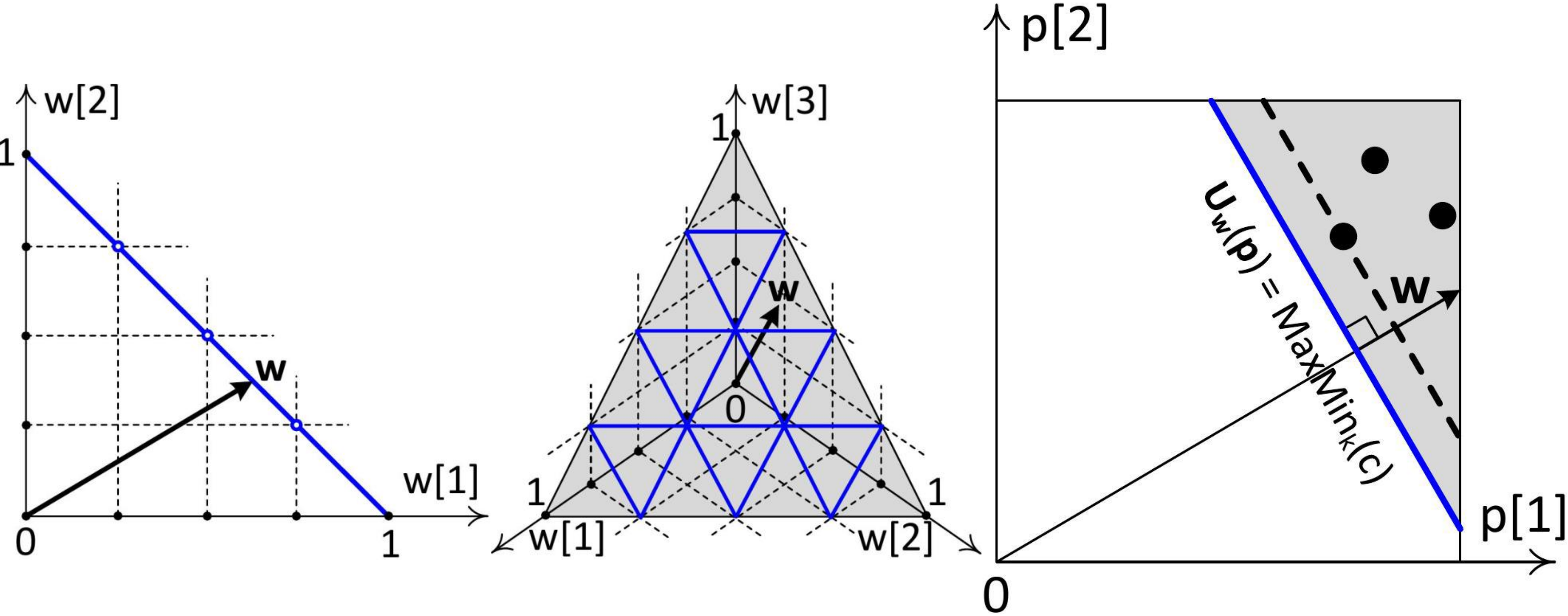
Competitiveness measures

- **Type I** (utility-based): how satisfied the different user types with the products available in *D*
- **Type II** (competition-based): how steep the competition among alternative products





Snapshots of our methodology



Lemma 2 *In general dimensionality d, there are* $\binom{2^h-1+d}{d} - \binom{2^h}{d}$ *nodes at depth h of the simplex pyramid.*

Conclusion

- We have overviewed the topic of **multicriteria/preference querying** and its many relationships to spatial indexing/querying
- We looked deeper into 3 specific examples (problem definitions)
- Overall, we saw that a skillset typical to SIGSPATIAL attendees may apply to an exciting, non-spatial domain

Thank you!