KU LEUVEN

Active Preference Learning for Ranking Patterns

Vladimir Dzyuba
Matthijs van Leeuwen, Siegfried Nijssen, Luc De Raedt

KU Leuven, Belgium

ICTAI 2013 – November 6, Herndon, VA, USA

Outline

1. Introduction: Case for interactive pattern mining
2. Algorithm: Interactive learning of subjective quality measures
3. Experimental evaluation
4. Take-away messages

Outline

- Techniques for exploratory data analysis
- Patterns are concise descriptions of *interesting* (???) regularities in data

Let the domain expert determine what is interesting, interactively and with minimal effort

Subgroups are regions in data with substantial deviation in a property of interest

Description attributes		Target
A_1	A_2	A_T
False	3	−
False	4	+
True	10	+
True	12	+
True	11	+
True	7	−
False	3	−

Description: $A_1 = True$

Quality φ:

- $Sensitivity = \frac{|G^+|}{|\mathcal{D}^+|} = \frac{3}{4}$
- $Specificity = 1 - \frac{|G^-|}{|\mathcal{D}^-|} = \frac{2}{3}$
- ...

Objective quality measures do not capture *subjective* interestingness

A_1	A_2	A_T
$False$	3	$-$
$False$	4	$+$
$True$	10	$+$
$True$	12	$+$
$True$	11	$+$
$True$	7	$-$
$False$	3	$-$

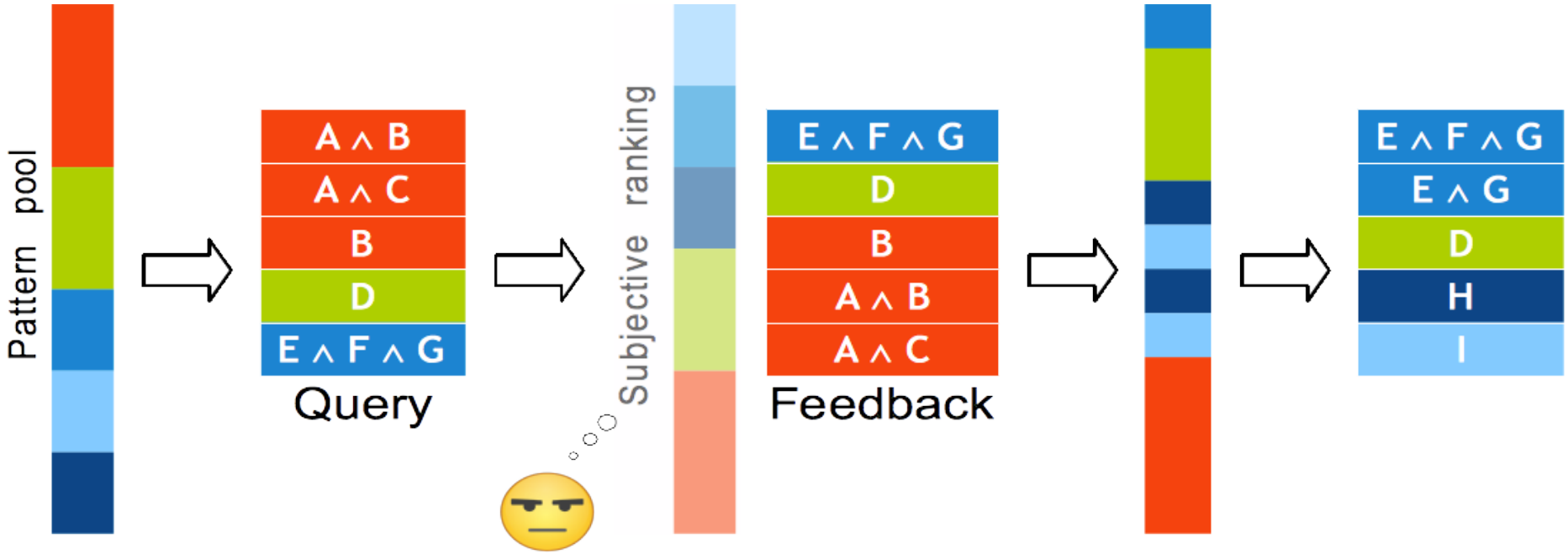
Subgroup		*Sensitivity*
$A_1 = True$	3+ 1−	$3/4$
$A_2 > 5$	3+ 1−	$3/4$

$A_2 > 5$ is *subjectively* more interesting than $A_1 = True$

User must be able to communicate such information to the algorithm

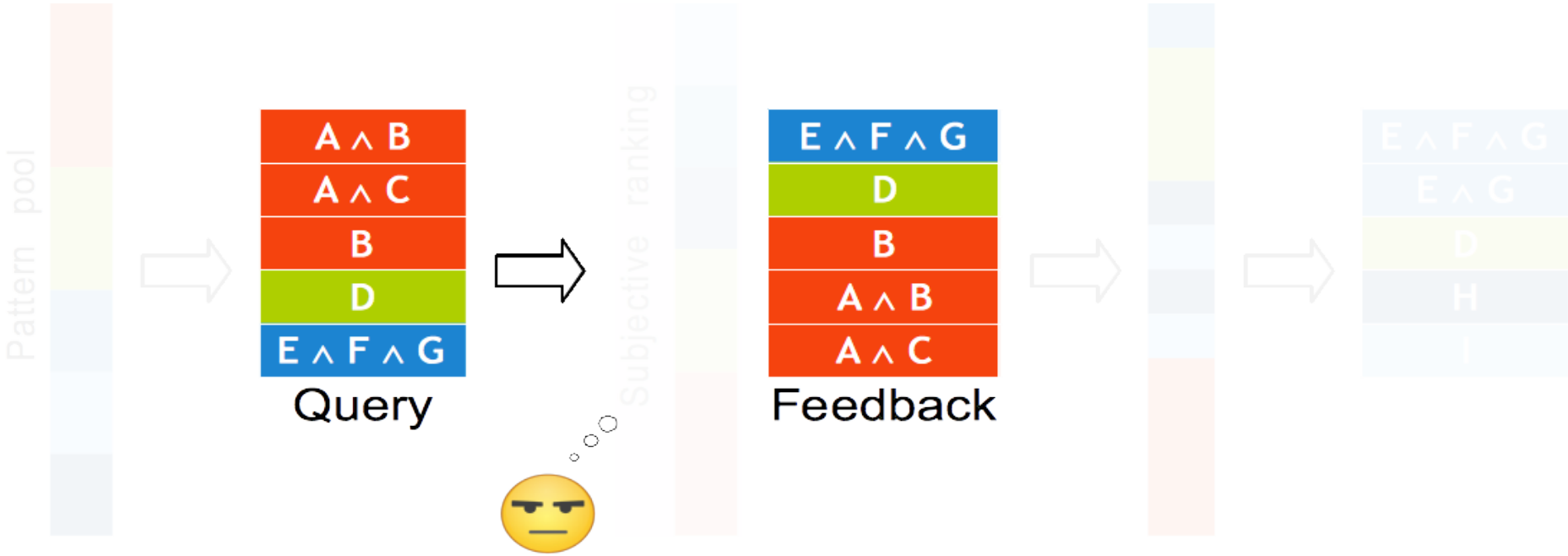
Core assumption:

The user has an (implicit) subjective ranking of all possible patterns

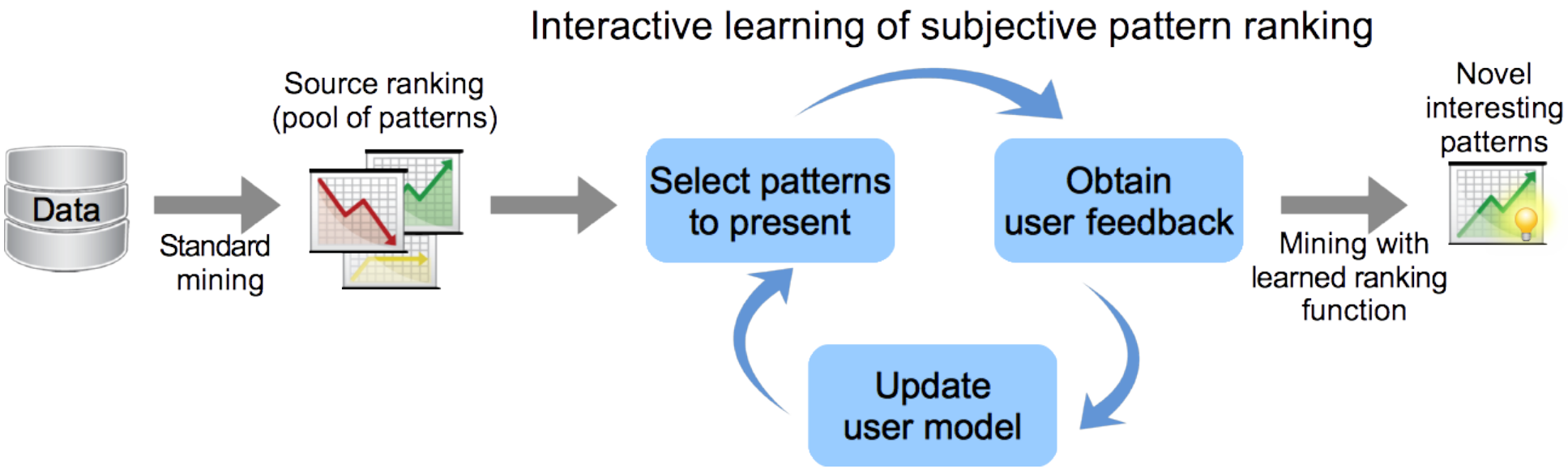


Proposed solution: User perspective

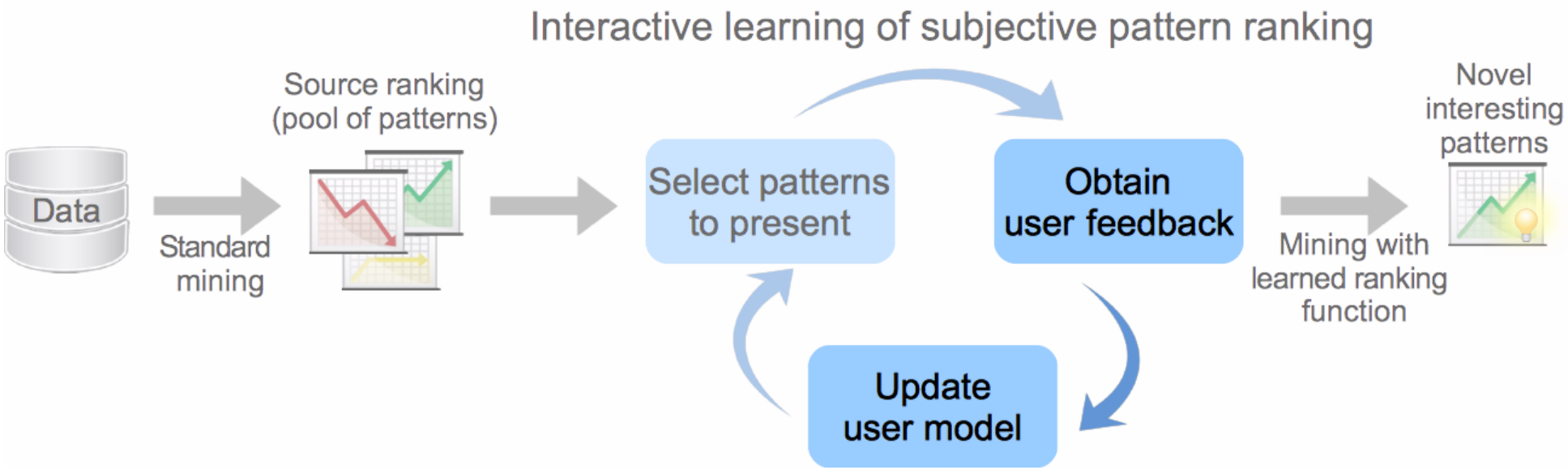
The user only evaluates small sets of patterns



Proposed solution: Interactive learning of quality measures 6/16



Outline

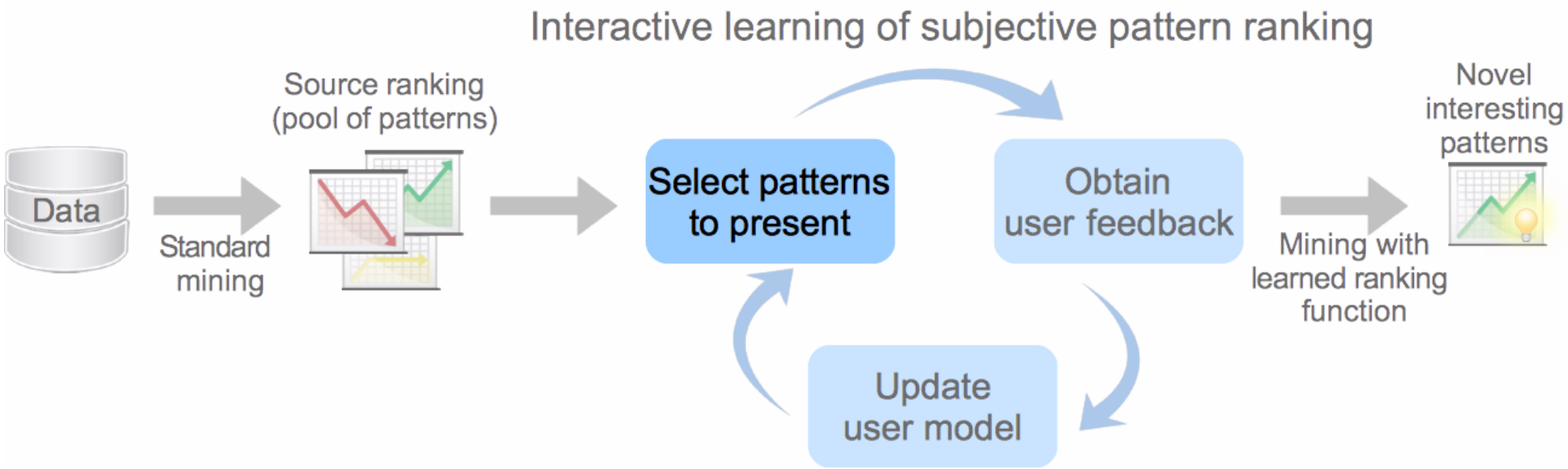


Learn a ranking function from feedback $\Rightarrow$ *Preference learning*

Ordered feedback:

- $\{G_1, G_2, G_3\} \Rightarrow G_3 \succ G_1 \succ G_2$
- To learn ranking functions we use an off-the-shelf preference learning algorithm, RANKING SVM with linear kernels

	Subgroup features			
G_3	1	1	1	1
G_1	1	1	0	0
G_2	1	1	0	1
Training examples	$\{G_3 \succ G_1,\ G_3 \succ G_2,\ G_1 \succ G_2\}$			
Learned weights	0.0000	0.0000	0.0050	−0.0025



Minimize required user effort $\Rightarrow$ *Active learning*

Pool-based batch active learning

Select subgroup sets of (small) fixed size that are then ranked by the user

Goal:
Minimize user effort required to learn accurate ranking functions

$$\mathit{Effort}(G_1 \succ G_2 \succ G_3) = |\{G_1 \succ G_2,\ G_1 \succ G_3,\ G_2 \succ G_3\}| = 3$$

Challenge:
Balance pattern space exploration with exploitation of available feedback

Greedy heuristics are used for query selection

IR-inspired heuristics:

- Aim at getting the top of the ranking right
- Balance (current approximation of) quality and query diversity

Uncertainty-based heuristics:

- Exploit properties of RANKING SVM
- Focus on the most uncertain pairs

Outline

Feedback emulation is required for extensive evaluation

Target "subjective" ranking is defined by an objective measure that is not known to the algorithm

We use χ^2 (*Chi Squared*) as target measure:

$$\chi^2(G) = \sum_{c \in \{-,+\}} \frac{(|G| \cdot (|G^c| - |\mathcal{D}^c|))^2}{|G| \cdot |\mathcal{D}^c|} + \frac{(|G| \cdot (|G^c| - |\mathcal{D}^c|))^2}{(|\mathcal{D}| - |G|) \cdot |\mathcal{D}^c|}$$

6 UCI datasets

3 *source rankings* per dataset: *Sensitivity*, *Specificity*, and *Coverage*

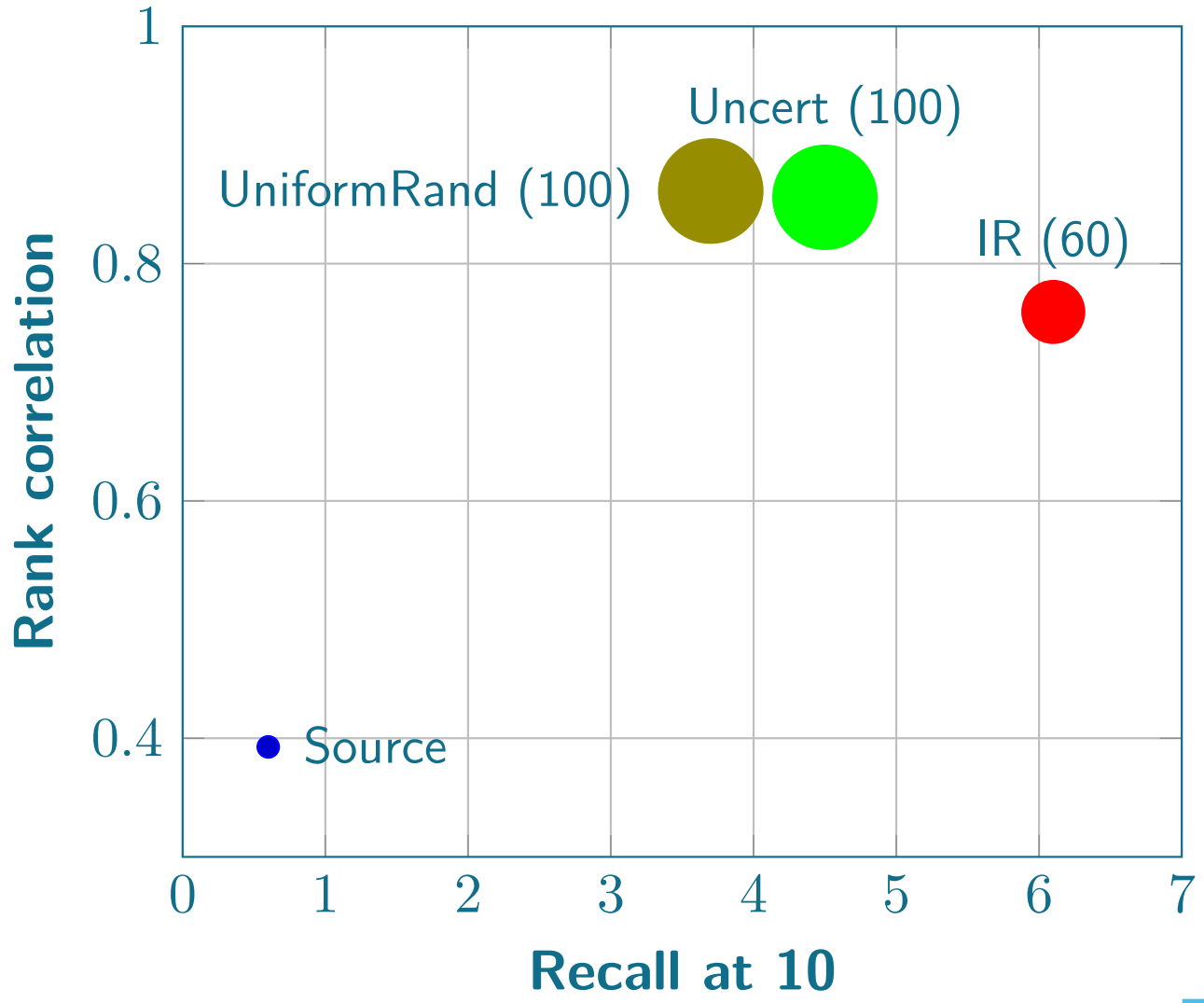
- Substantially different from the target ranking, by intention

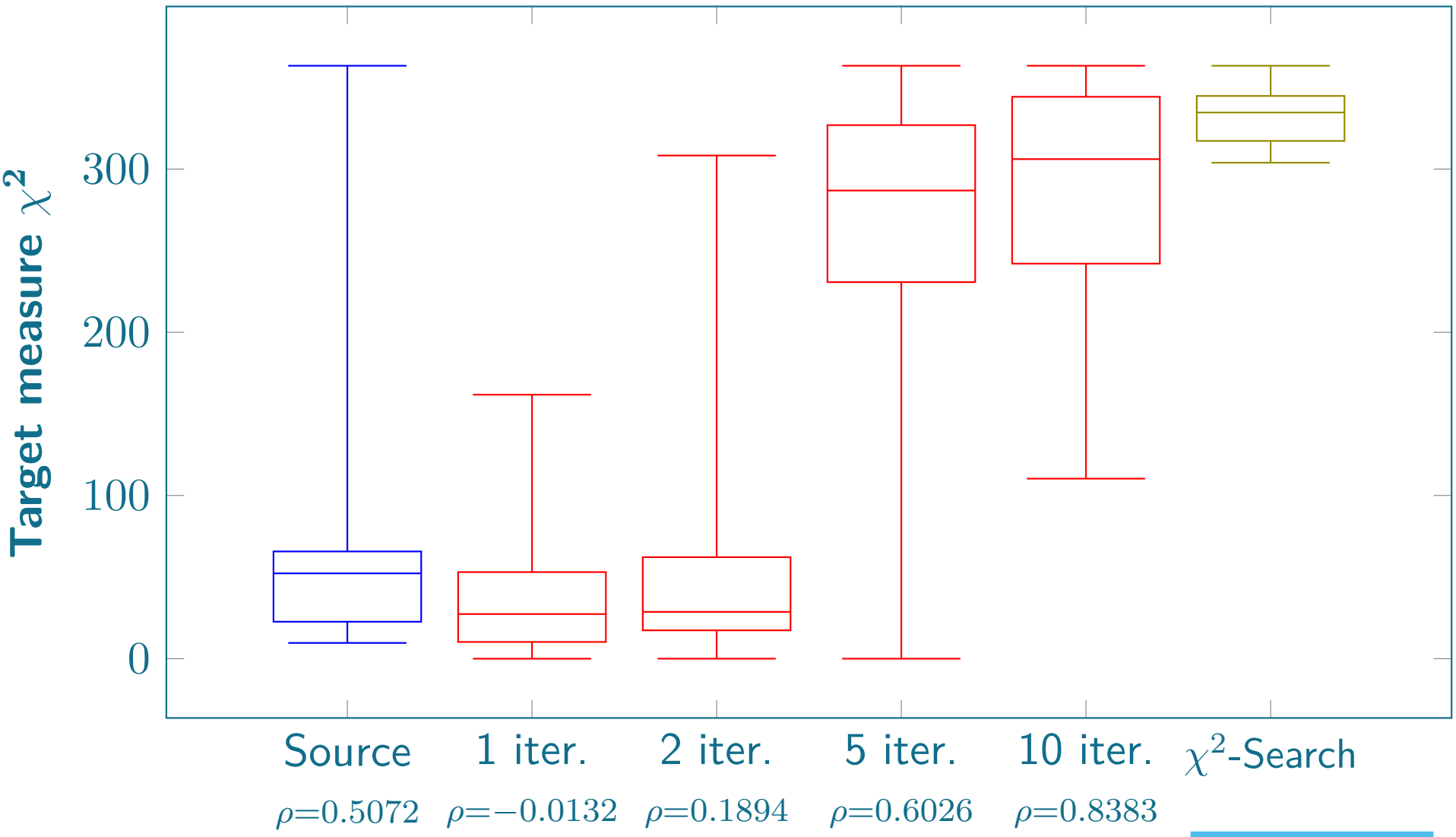
5 subgroups in query, 10 feedback iterations (maximal effort is 100 pairs)

Performance measures:

- *Recall at 10* – Getting top of the ranking right
- *Rank correlation* – Getting the entire ranking right

Comparing active learning heuristics




Target measure χ^2
300
200
100
0
Source
1 iter.
2 iter.
5 iter.
10 iter.
χ^2-Search
ρ=0.5072
ρ=−0.0132
ρ=0.1894
ρ=0.6026
ρ=0.8383

Outline

Active learning of subjective preferences over patterns is an important building block for interactive pattern mining systems

- Preference learning captures subjective quality of subgroups
- Active learning reduces required user effort
- Using learned ranking functions in search allows discovering novel interesting subgroups
- Approach can be easily extended to other types of patterns

Note

Principled evaluation is challenging

Thank you for your attention!

May I answer any questions?