KU LEUVEN

Interactive Discovery of Interesting Subgroup Sets

Vladimir Dzyuba, Matthijs van Leeuwen

Department of Computer Science,
KU Leuven, Belgium

IDA 2013 – October 17, London, UK

Outline

1. Introduction: Case study & Encountered issues
2. IDSD: Interactive Subgroup Discovery algorithm
3. Experimental evaluation
   - Emulated feedback
   - User study
4. Take-away messages

Outline

We are interested in what makes a team good at scoring
Can we discover it from data?

Player 1	Statistic 1	Scoring
Out	1	Low
Out	2	$High$
In	30	$High$
In	40	$High$
In	50	$High$
In	60	Low
Out	700	Low

We can use *Top-k Subgroup Discovery*:

- "Mining" descriptions of regions in the data with substantial deviation in a property of interest

Pl.1	Stat.1	Scoring
Out	1	Low
Out	2	$High$
In	30	$High$
In	40	$High$
In	50	$High$
In	60	Low
Out	700	Low

We can use *Top-k Subgroup Discovery*:

- "Mining" descriptions of regions in the data with substantial deviation in a property of interest

Description attributes		**Target**
Pl.1	Stat.1	Scoring
Out	1	Low
Out	2	$High$
In	30	$High$
In	40	$High$
In	50	$High$
In	60	Low
Out	700	Low

- Description
  $Pl.1 = In \wedge Stat.1 > 10$

We can use *Top-k Subgroup Discovery*:

- "Mining" descriptions of regions in the data with substantial deviation in a property of interest

Description attributes		**Target**
Pl.1	Stat.1	Scoring
Out	1	Low
Out	2	$High$
In	30	$High$
In	40	$High$
In	50	$High$
In	60	Low
Out	700	Low

- Description
  $Pl.1 = In \wedge Stat.1 > 10$
- Cover $G \Rightarrow |G| = 4$

We can use *Top-k Subgroup Discovery*:

- "Mining" descriptions of regions in the data with substantial deviation in a property of interest

Description attributes		**Target**
Pl.1	Stat.1	Scoring
Out	1	Low
Out	2	$High$
In	30	$High$
In	40	$High$
In	50	$High$
In	60	Low
Out	700	Low

- Description
  $Pl.1 = In \wedge Stat.1 > 10$
- Cover $G \Rightarrow |G| = 4$
- Quality $\varphi(G) \propto 4 \cdot \left(\frac{3}{4} - \frac{4}{7}\right)$

Description	Size	Quality
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge thabeet = Out$	219	0.0692
$opp\_def\_reb = Low \wedge opponent \neq ATL$	222	0.0689
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge ajohnson = Out$	222	0.0689
$opp\_def\_reb = Low \wedge thabeet = Out \wedge opponent \neq PHI$	225	0.0685
$opp\_def\_reb = Low \wedge opponent \neq PHI$	228	0.0682

Top 5 subgroups discovered by DSSD with default settings

Top subgroups are slight variations of one theme:

- Only 231 out of 923 game segments in the dataset are covered

Description	Size	Quality
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge thabeet = Out$	219	0.0692
$opp\_def\_reb = Low \wedge opponent \neq ATL$	222	0.0689
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge ajohnson = Out$	222	0.0689
$opp\_def\_reb = Low \wedge thabeet = Out \wedge opponent \neq PHI$	225	0.0685
$opp\_def\_reb = Low \wedge opponent \neq PHI$	228	0.0682

Top 5 subgroups discovered by DSSD with default settings

Trivial facts that are already known by experts

- Subgroups with high *objective* quality

Description	Size	Quality
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge thabeet = Out$	219	0.0692
$opp\_def\_reb = Low \wedge opponent \neq ATL$	222	0.0689
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge ajohnson = Out$	222	0.0689
$opp\_def\_reb = Low \wedge thabeet = Out \wedge opponent \neq PHI$	225	0.0685
$opp\_def\_reb = Low \wedge opponent \neq PHI$	228	0.0682

Top 5 subgroups discovered by DSSD with default settings

Not interesting or actionable descriptions:

- Useless for analysis and decision making

Description	Size	Quality
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge thabeet = Out$	219	0.0692
$opp\_def\_reb = Low \wedge opponent \neq ATL$	222	0.0689
$opp\_def\_reb = Low \wedge opponent \neq ATL \wedge ajohnson = Out$	222	0.0689
$opp\_def\_reb = Low \wedge thabeet = Out \wedge opponent \neq PHI$	225	0.0685
$opp\_def\_reb = Low \wedge opponent \neq PHI$	228	0.0682

Top 5 subgroups discovered by DSSD with default settings

Observations:

- Interestingness is inherently *subjective*
- Issues cannot be solved effectively *after search*
- Algorithm is not easy to tune for *domain experts*

Let the user directly influence subgroup search:

1. Introduce *within-the-search* feedback mechanism
2. Make it easy-to-use for domain experts
3. Allow them communicate with algorithm in domain terms (e.g. $Player\ 1$ instead of φ)

Note

Here we strongly focus on the goal of *eliminating undesired results*

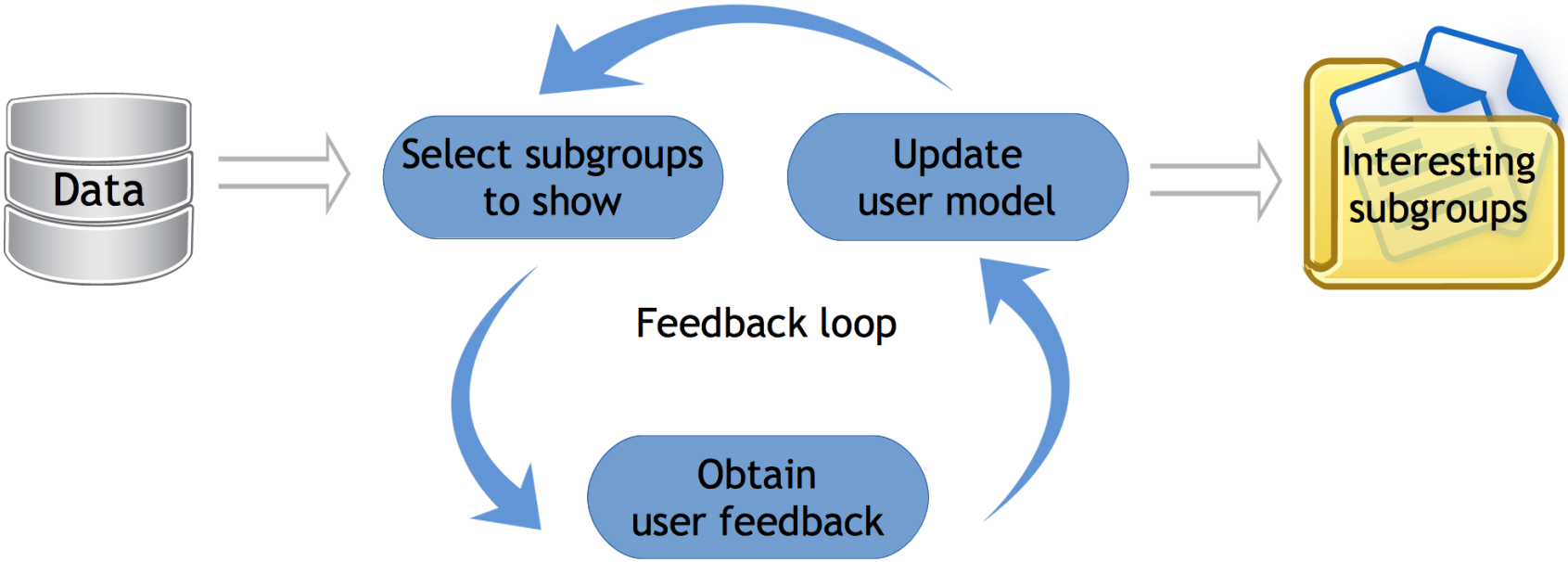
Domain expert – a basketball journalist – used augmented algorithm
Results were *subjectively* more interesting

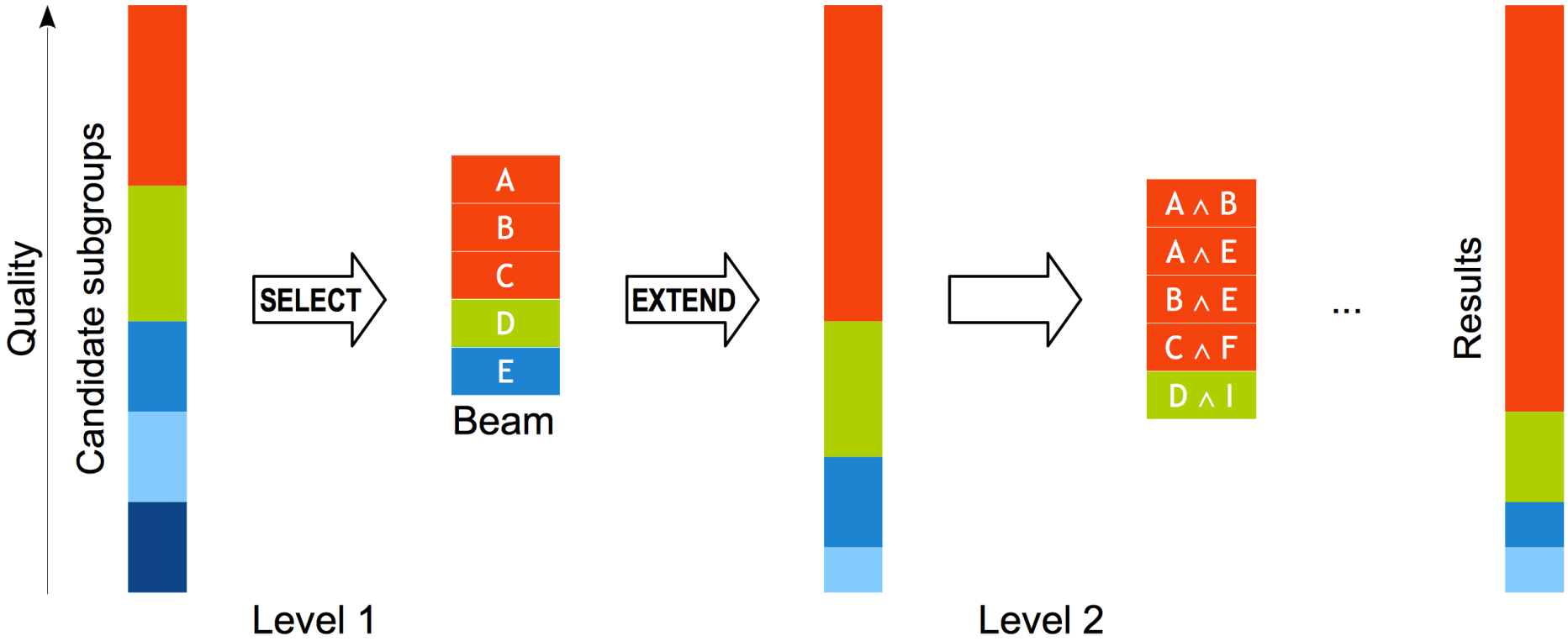
Description	Size	Quality
$crawford = Out \wedge matthews = In$	96	0.0382
$hickson = In$	186	0.0219
$crawford = Out \wedge hickson = In$	328	0.0211
$matthews = In \wedge hickson = In$	290	0.0163
$matthews = In \wedge pace < 88.518$	303	0.0221

Top 5 subgroups discovered by interactive algorithm

Outline

Key idea: alternate between mining and user feedback




Quality
Candidate subgroups
SELECT
A
B
C
D
E
Beam
EXTEND
A ∧ B
A ∧ E
B ∧ E
C ∧ F
D ∧ I
...
Results
Level 1
Level 2

Simple “binary" feedback:

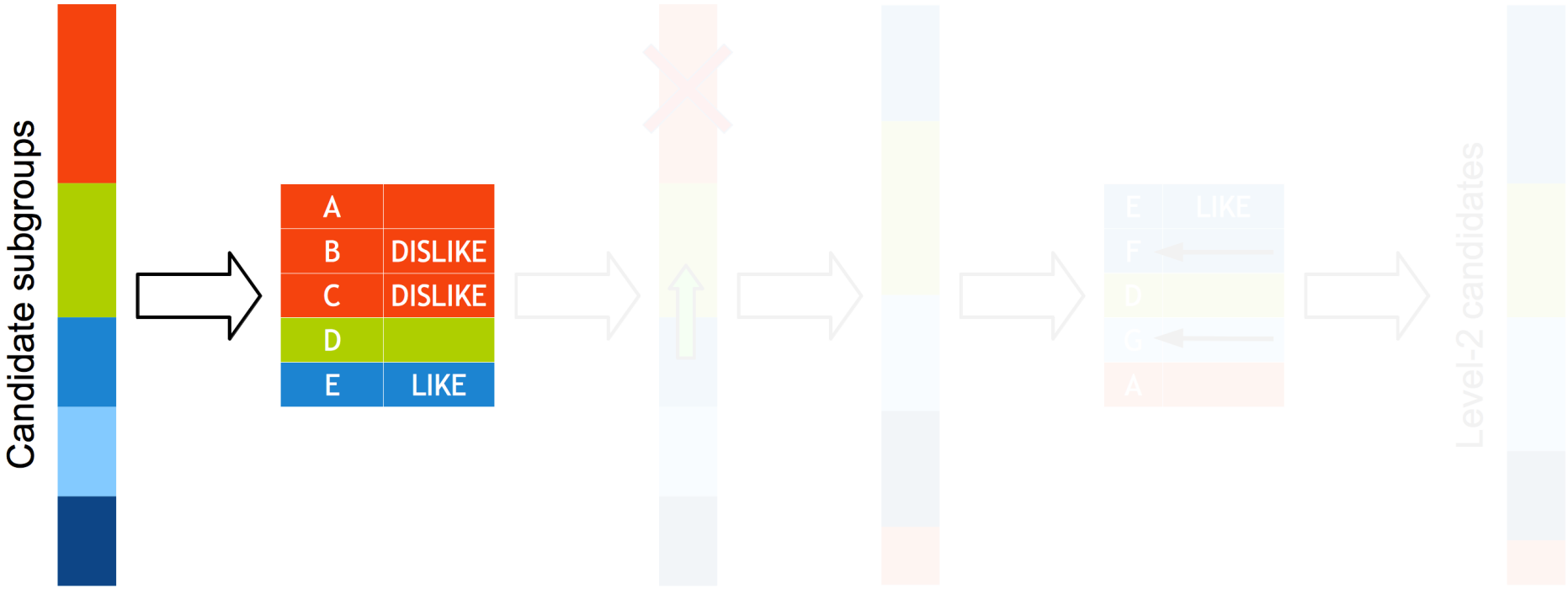
-  LIKE, i.e. positive evaluation/"subgroup is interesting"
-  DISLIKE, i.e. negative evaluation
- Abstaining from feedback is possible

Integrate feedback into beam search:

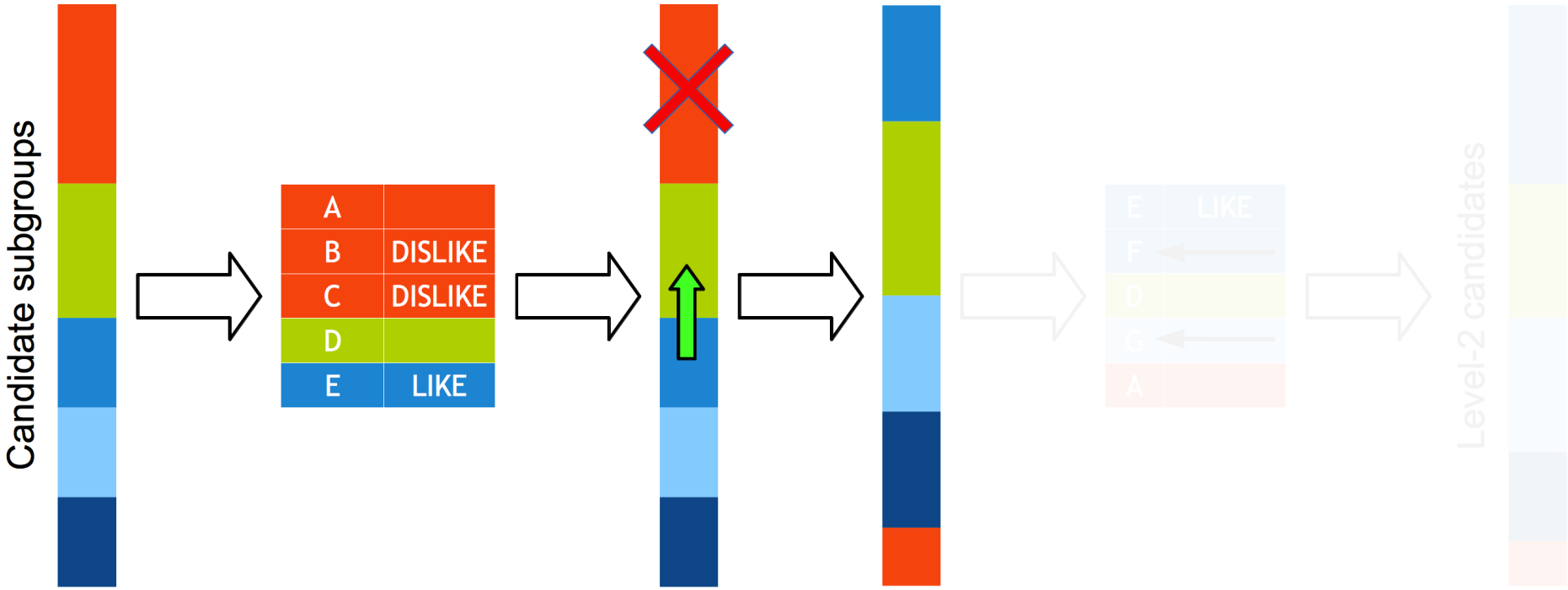
- User provides feedback on each beam
- Subgroup similarity measure is used to generalise this feedback

User feedback processing

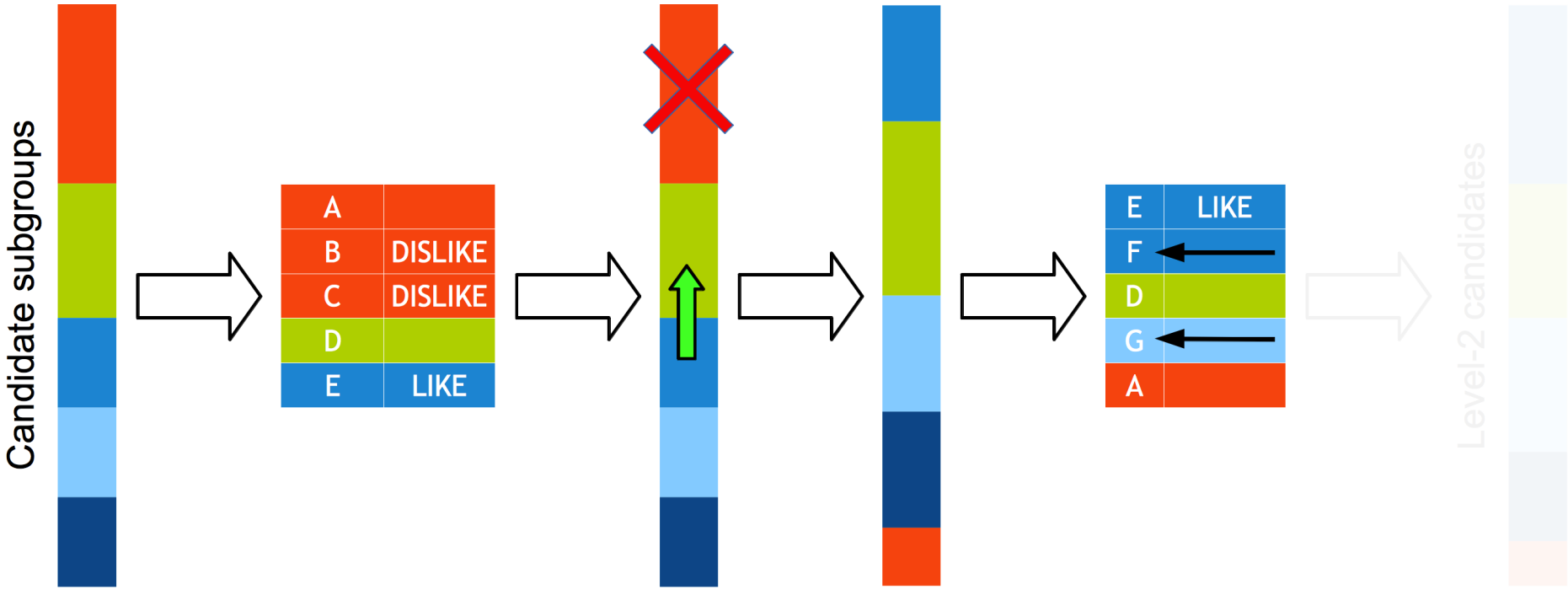
User provides feedback on the selected beam



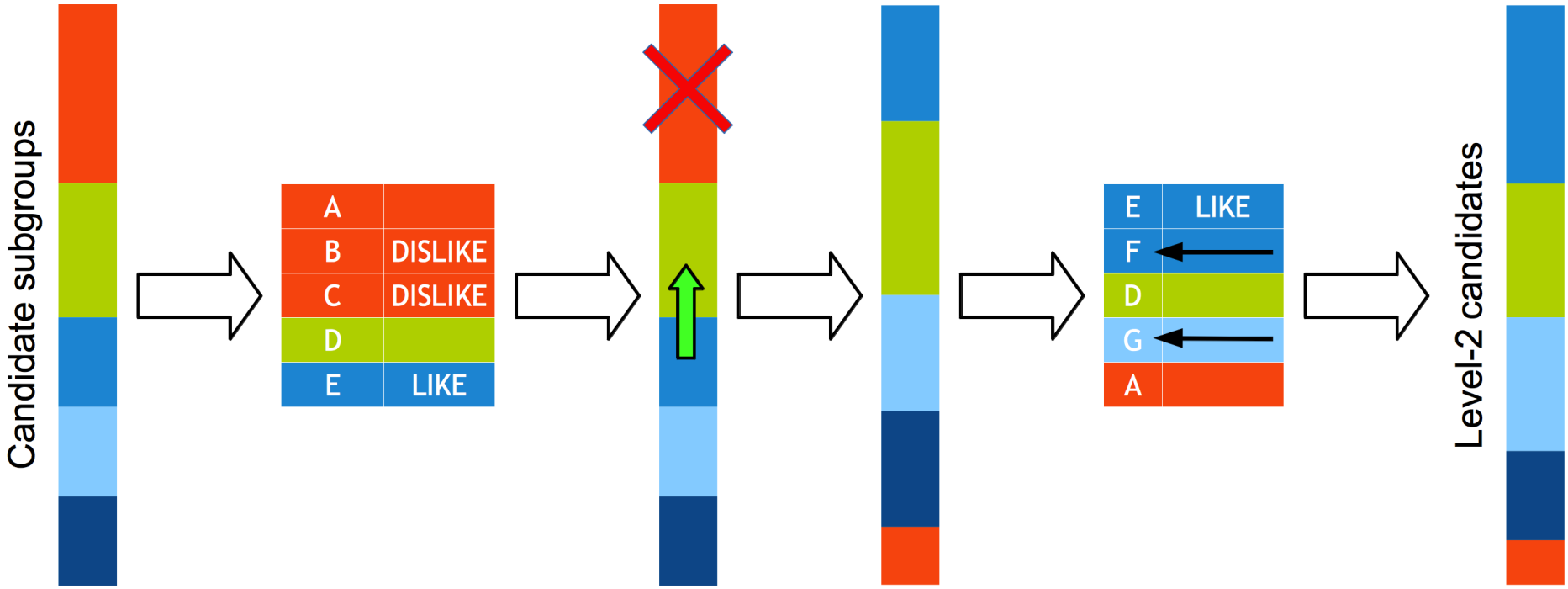
Candidate subgroups are re-ranked



Disliked subgroups are replaced



Level-2 candidates are generated



Outline

Outline

Motivated by basketball case study:

- Top subgroups correspond to well-known facts
- They should be avoided in order to obtain interesting results

For each dataset:

1. Mine subgroups using DSSD
2. Assign few top non-overlapping subgroups to *negative background knowledge* BK
3. Dislike subgroup during search if similarity σ with any element of BK is greater than threshold β

Note

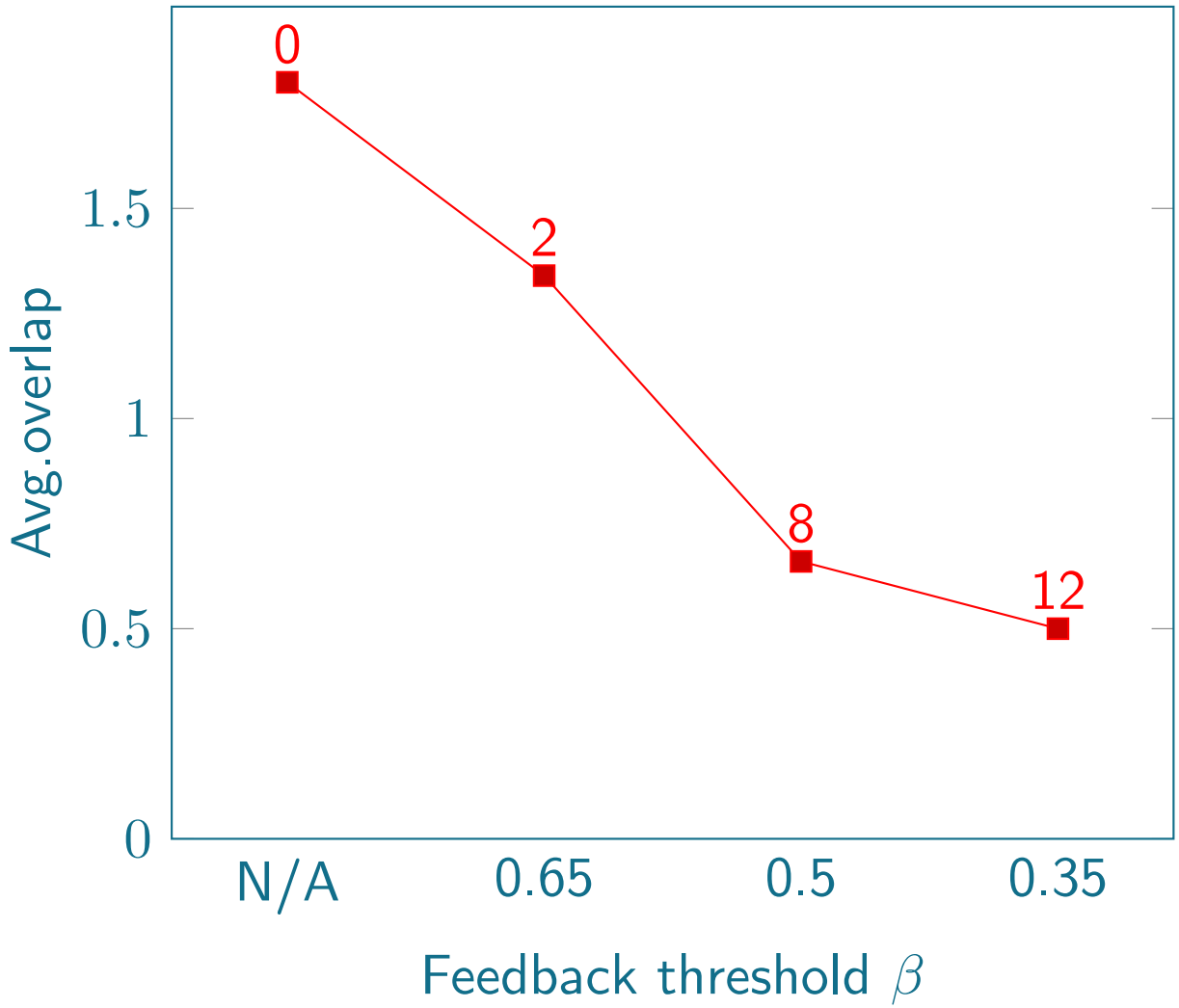
Only negative feedback is emulated

- UCI dataset *german credit*
- 2 subgroups in BK
- Description similarity σ_{desc}, e.g. $\sigma(A \wedge B; A \wedge C) = \frac{1}{2}$

- Overlap of descriptions with subgroups in BK
- Objective quality of discovered subgroups
  - Strength of dependency
- Redundancy of result sets

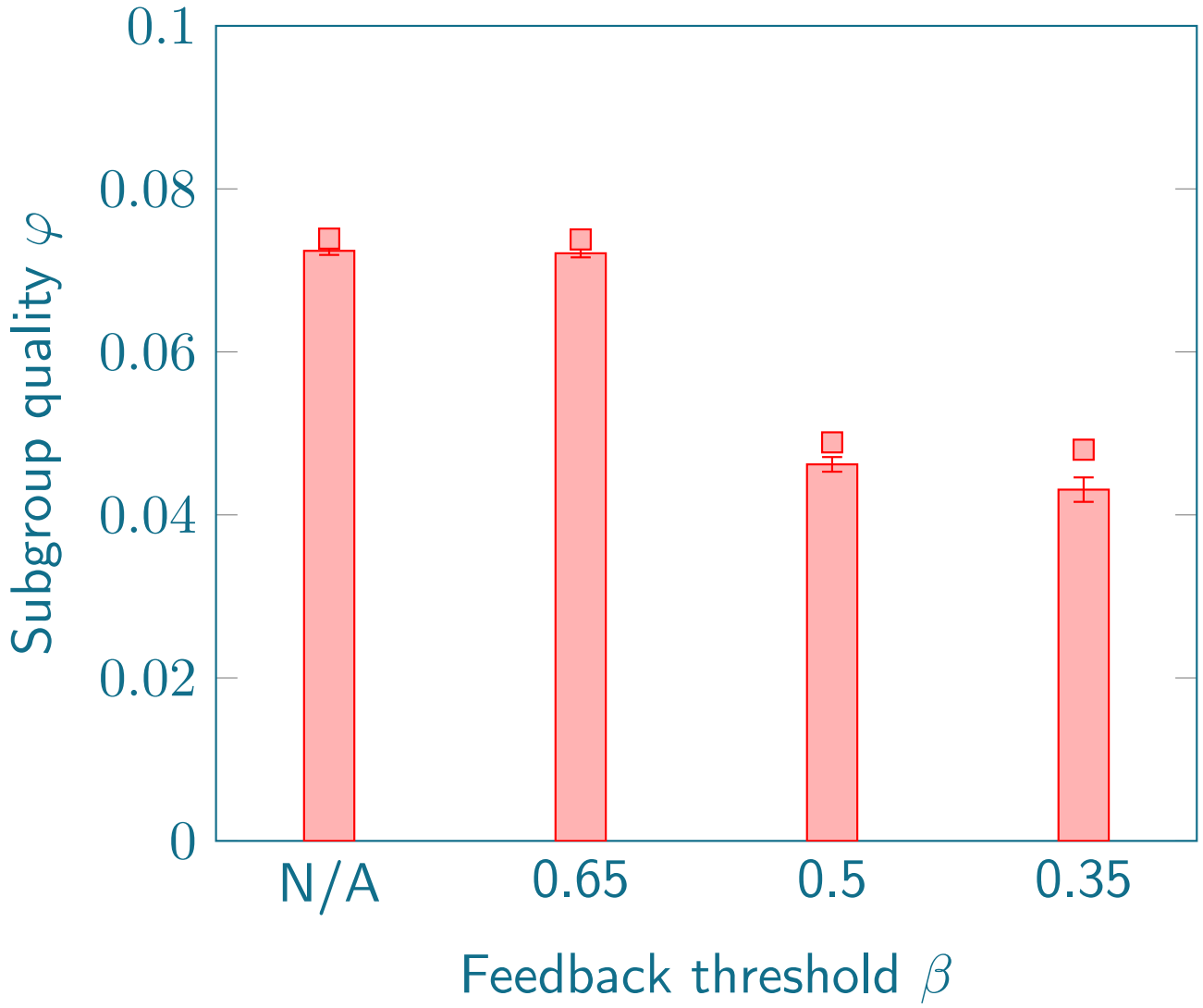
Characteristic example: Overlap with BK

Few dislikes (12 at most) allow eliminating undesired conditions from results



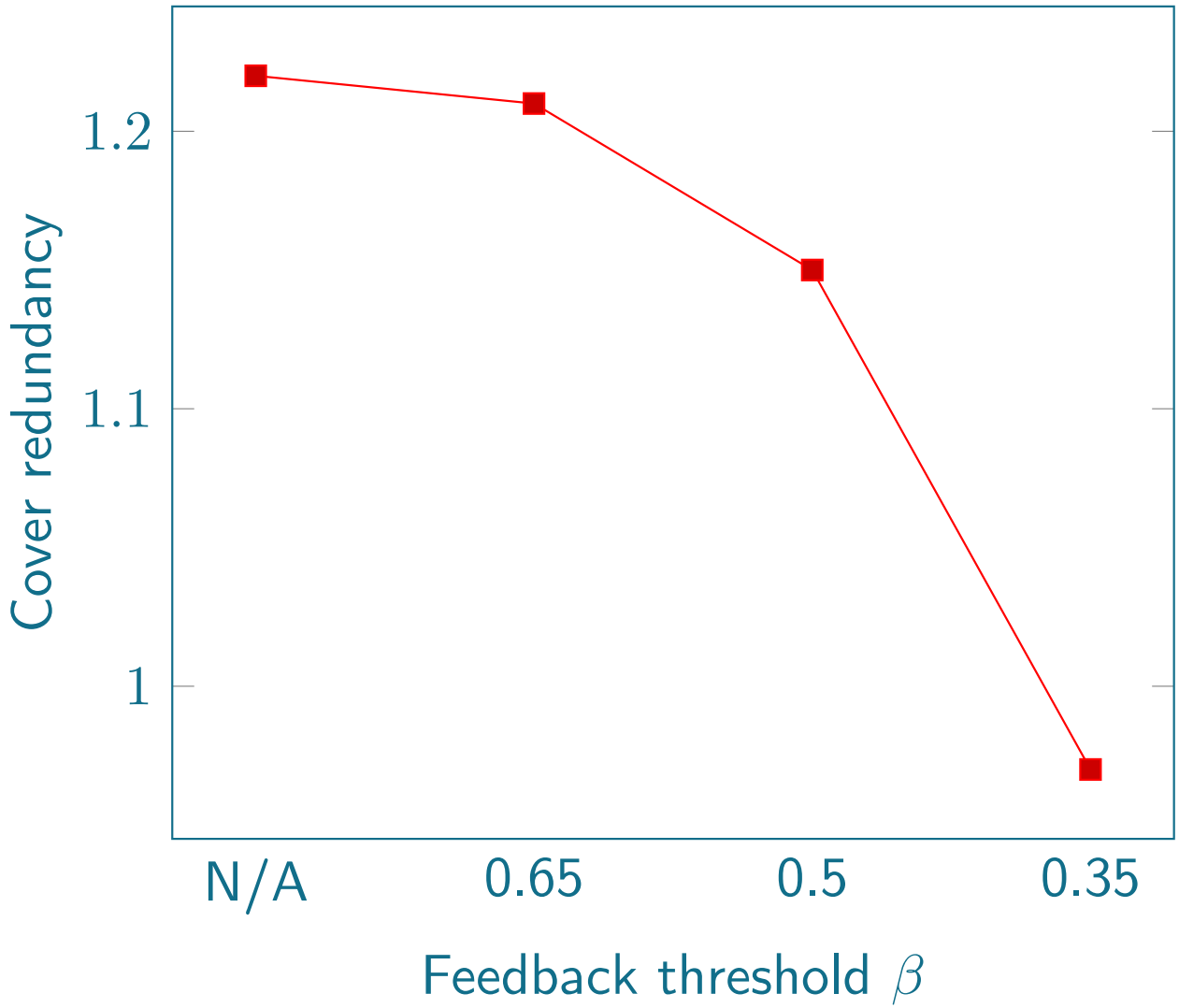
Characteristic example: Subgroup quality

Objective quality of discovered subgroups remains relatively high



Characteristic example: Redundancy

Redundancy slightly decreases with more feedback



Outline

Domain expert – basketball journalist:

- 7 liked and 11 disliked subgroups
- Results were *subjectively interesting*

Description	Size	Quality
$crawford = Out \wedge matthews = In$	96	0.0382
$hickson = In$	186	0.0219
$crawford = Out \wedge hickson = In$	328	0.0211
$matthews = In \wedge hickson = In$	290	0.0163
$matthews = In \wedge pace < 88.518$	303	0.0221

Remark

Not all analysis sessions were as successful

Assumption

User inspects subgroups one by one:

1. To provide feedback
   $E_{beam} = \textit{Size of all beams} + \textit{Number of disliked subgroups}$
2. To find interesting subgroups in result set
   $E_{results} = \textit{Lowest rank of interesting subgroup in result set}$

Hence, to obtain results above:

- $E_{DSSD} = E_{results} = 1049$
- $E_{IDSD} = E_{results} + E_{beam} = 5 + (3 \times 10 + 11) = 46$

Outline

Active user involvement in the knowledge discovery process is essential for obtaining genuinely interesting results

Even proof-of-concept implementation of within-the-search interactivity:

- Eliminates undesired results effectively
- Reduces required effort
- Makes algorithm easier-to-use for domain experts

Challenges

- Estimating *interestingness* and *effort*
- Designing principled evaluation methods

Thank you for your attention!

May I answer any questions?