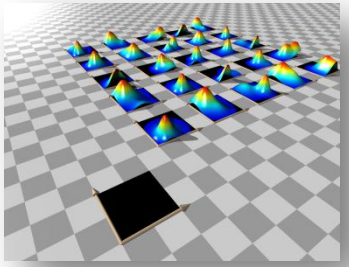
Summarizing data collections by their spatial, temporal, textual, and image footprint: Techniques for source selection and beyond

Andreas Henrich and Daniel Blank
andreas.henrich@uni-bamberg.de
University of Bamberg
Media Informatics Group

Leipzig, 15.11.2012

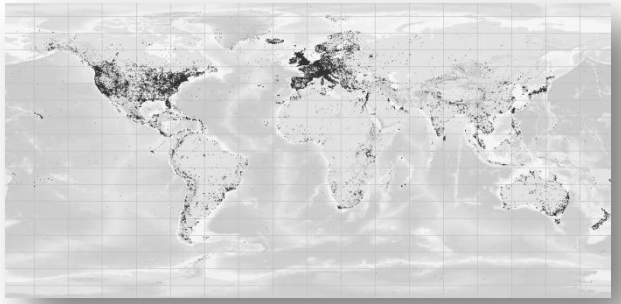
Agenda



1. Motivation and Introduction
   (1) Distributed indexing and search
   (2) Media types in use
   (3) Scenario: source selection for image retrieval
2. Source Selection: Research Goals & Work in our Group
   (1) Efficient source selection with HFS and UFS
   (2) Visualizing the source selection process
   (3) Applicability in other application fields (geographic IR, access methods)
3. Conclusion and Outlook

1. Motivation and Introduction

- Dramatic increase in data volumes in many areas:
  - in the world wide web
  - on private devices
  - in companies
  - …

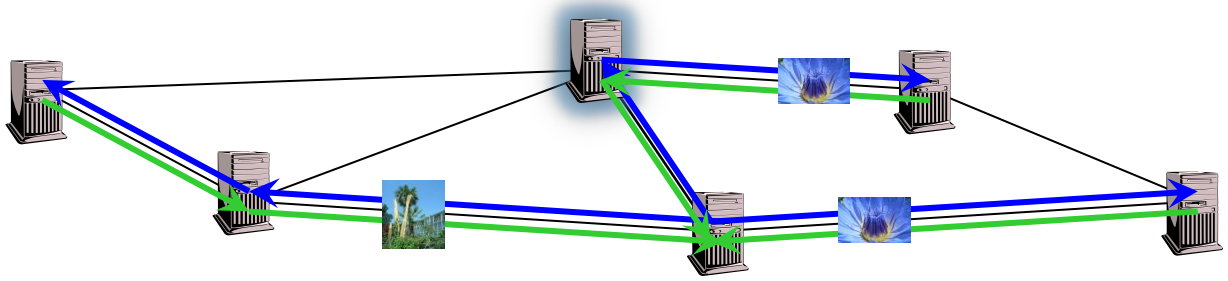


⇒ adequate indexing and search techniques needed

⇒ distributed indexing and search

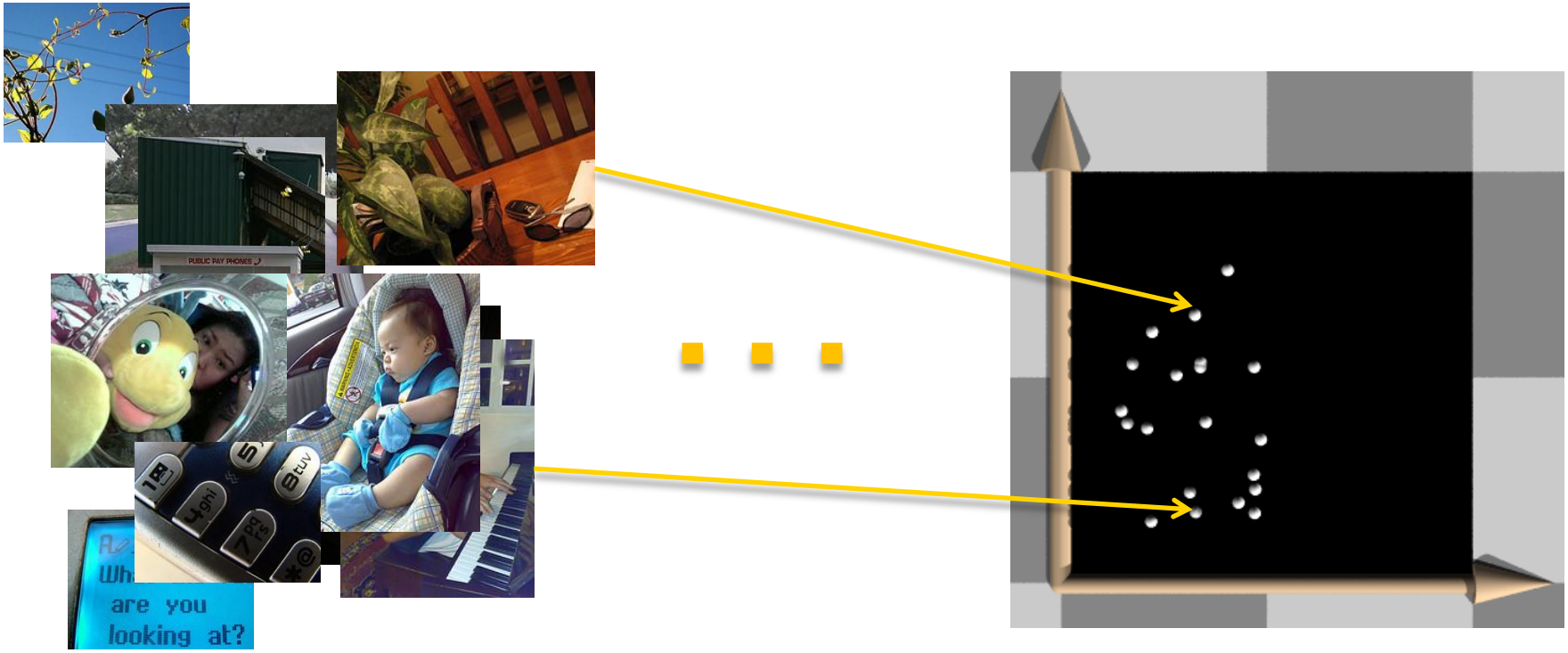
Distributed indexing and search

- global vs. local indexing:
  - global indexing: “one” global index is distributed
    - inserts and updates → high network load
    - no locality of data (autonomy) → replication
  - local Indexing: many local indexes & query routing based on local data summaries
    - query processing with logarithmic cost hard to guarantee
    - hybrid approaches: caching, replication, data specialization, …

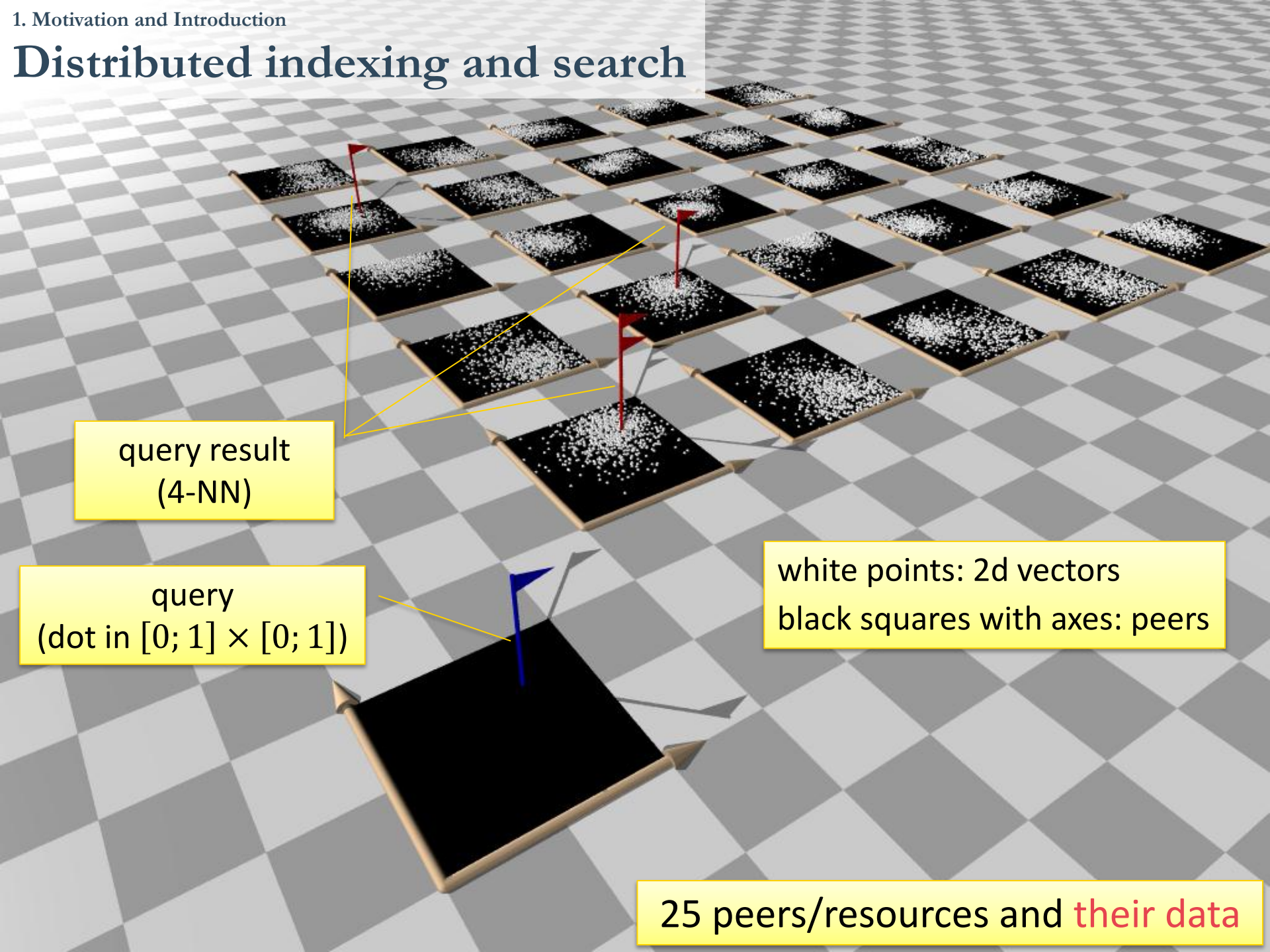


Distributed indexing and search

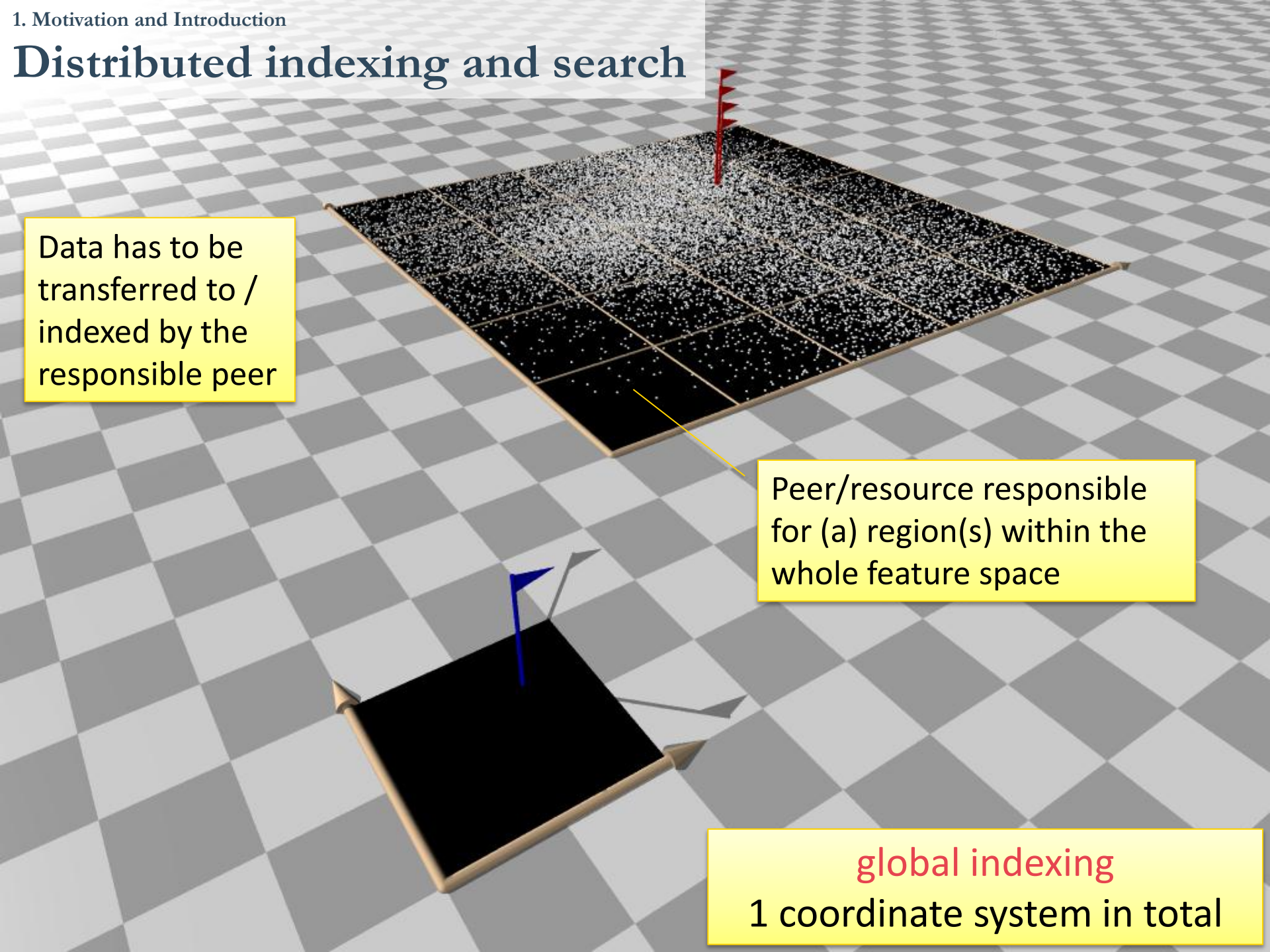
Visualization: features of 1 object/image ⇔ 1 point in 2d



Distributed indexing and search

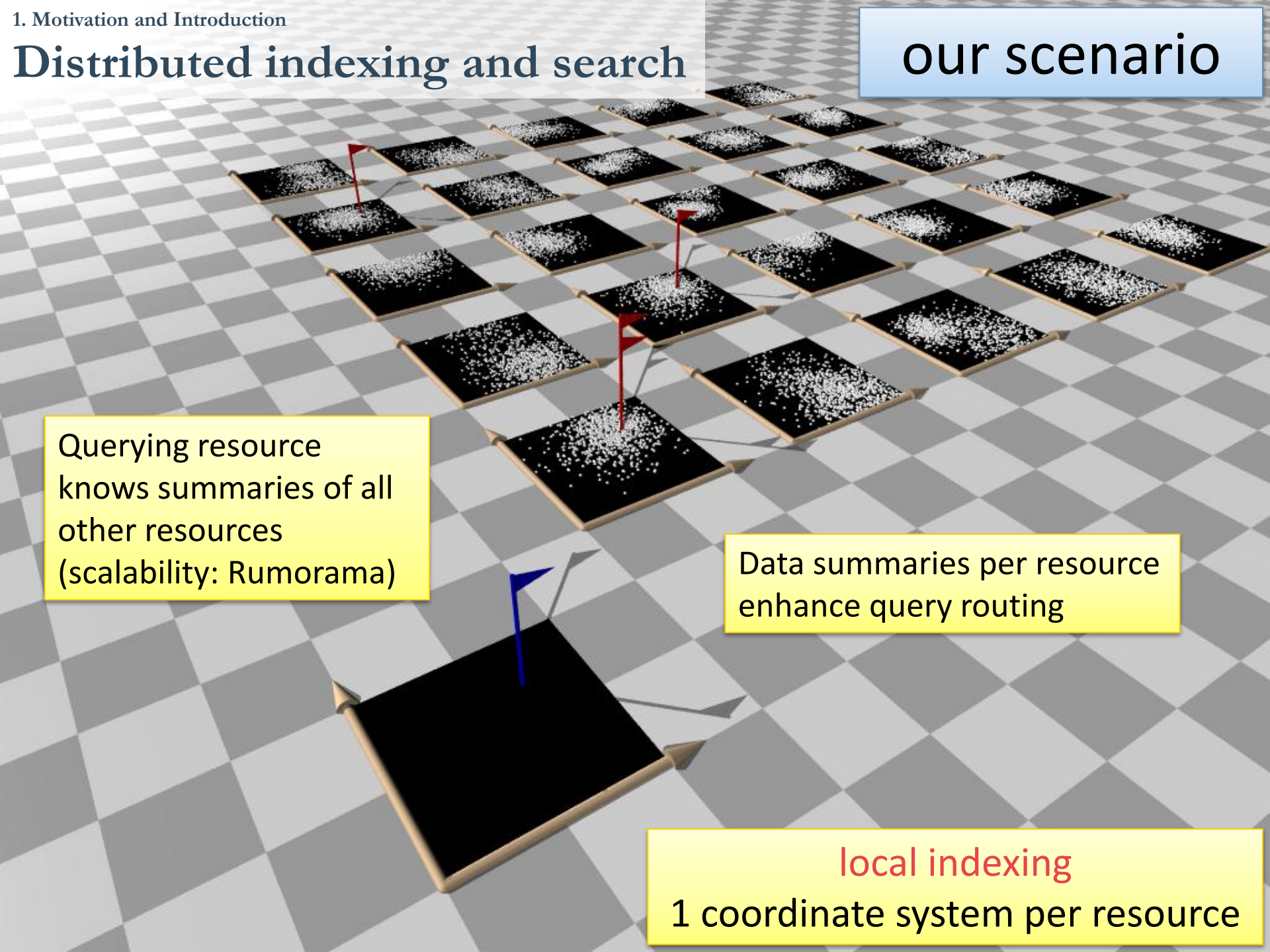


Distributed indexing and search



Distributed indexing and search

our scenario



Querying resource knows summaries of all other resources (scalability: Rumorama)

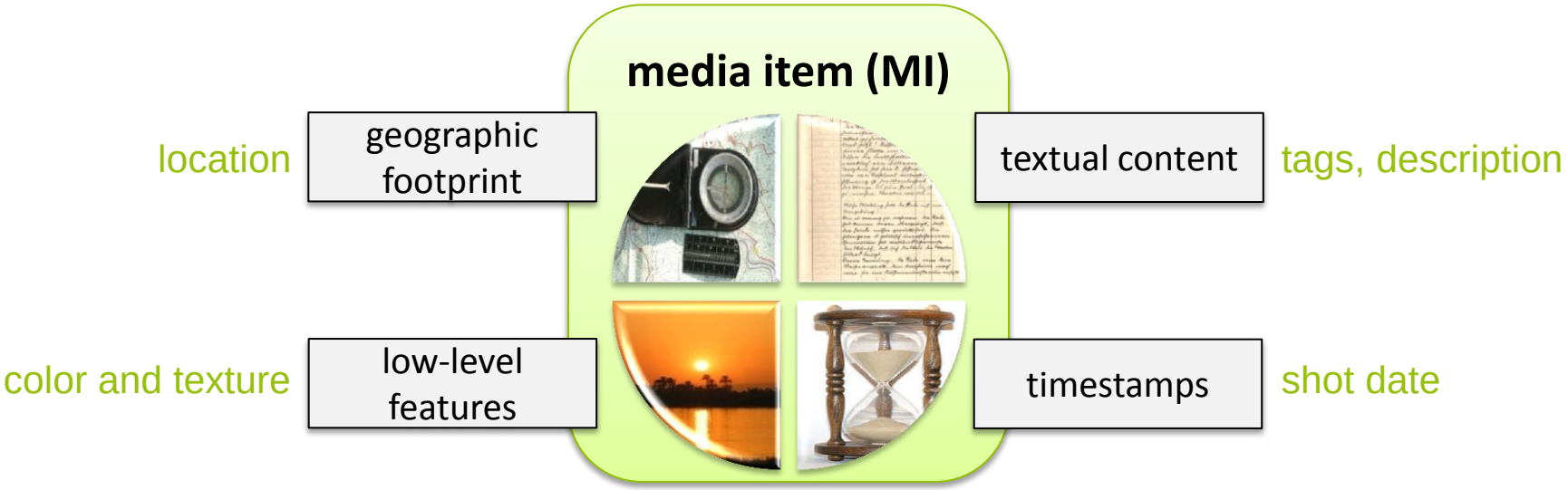
Data summaries per resource enhance query routing

local indexing

1 coordinate system per resource

Media types in use

Administration of distributed media items:

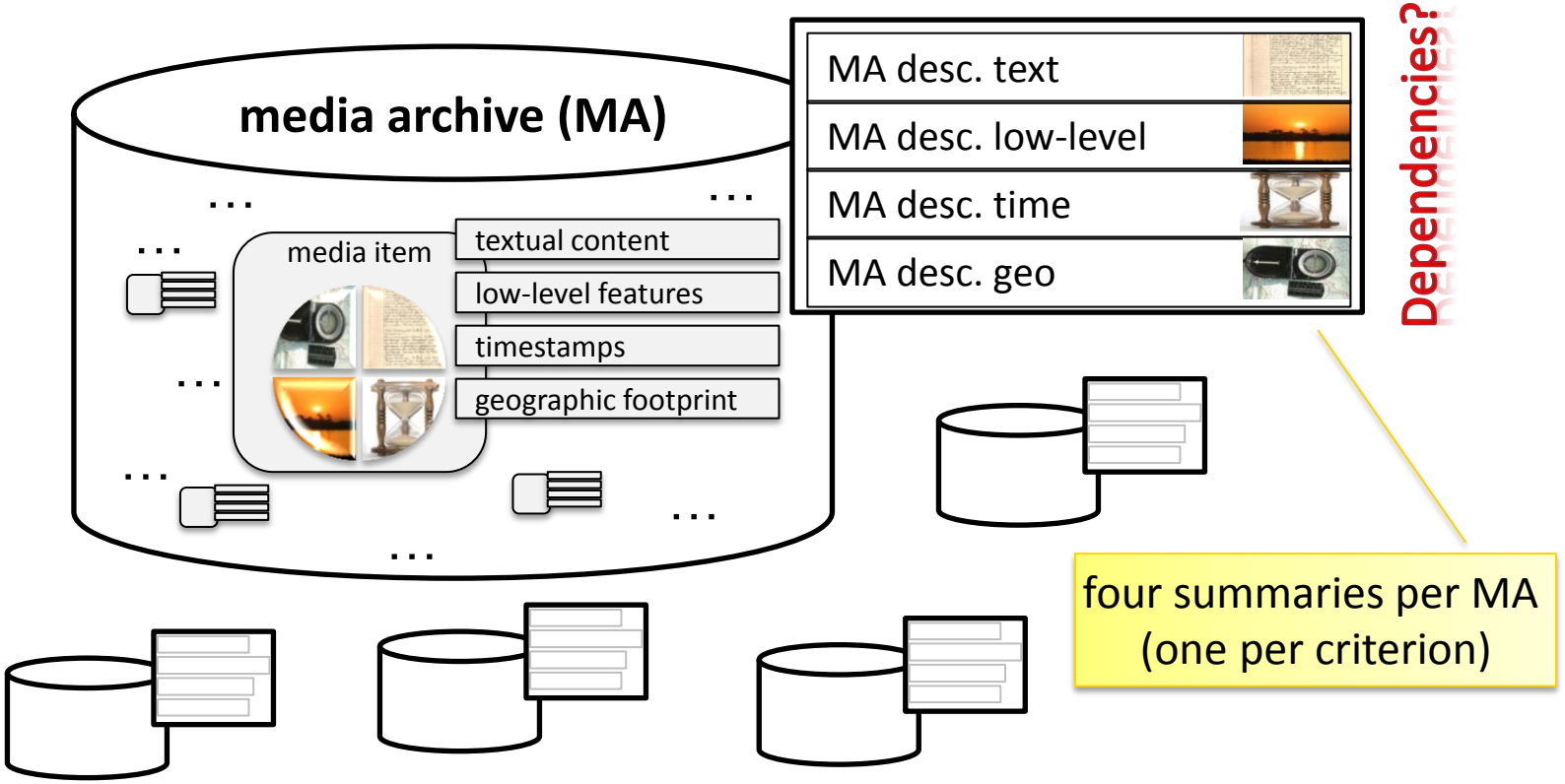


Content-based image retrieval (CBIR) just an example scenario.

Goal: Provide resource description and selection techniques | for general metric spaces | and apply them elsewhere (e.g. geo-context)

Media types in use

Resource description and selection: Identify promising MAs according to a query based on a set of known data summaries.



Media types in use → resource descriptions


text → (counting) Bloom filters, Topic models, etc.


time → histogram, clustering, our approach feasible


geo → discussed at the end of the talk


low-level → focus in the remainder of the talk (CBIR)

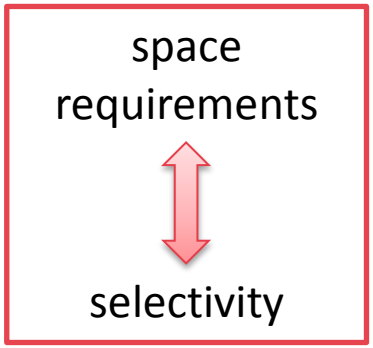
Search by multiple criteria: e.g. *'town hall Leipzig' at sunset*

- merging of several criteria-specific resource rankings
- need to model correlations in the resource descriptions
- future work: not covered so far

Scenario: source selection for image retrieval

Callan [Cal00] defines distributed information retrieval (IR) based on three problems and tasks:

1. resource description
   adequate descriptions/summaries
2. resource selection
   adequate selection mechanisms
3. result merging
   merging of resource-specific search results



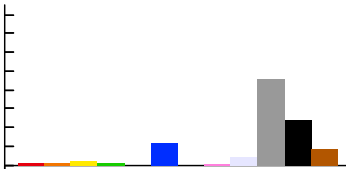
Often in literature (and also in this talk):

- resource selection = resource description & selection;
- result merging trivial

Scenario: source selection for image retrieval

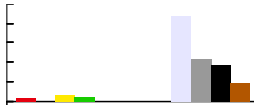
exemplarily

- Image retrieval: often text search & tags
- Only text search not sufficient:
  - Bolettieri et al. [BEF+09]: >100 Mio. Flickr images
    - 30%: no comments and no tags → searchable?
    - on average: 0.5 comments and 5.0 tags
  - tag spam, homonyms/synonyms, selectivity (‘phone’), …
- Benefits of content-based image retrieval (CBIR):
  - automatically extracted image properties
  - features: color (histograms), texture, salient points, …



Scenario: source selection for CBIR

given: query object q
database O



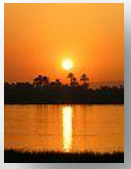
wanted: 'similar' images w.r.t. query; criterion: $d(q,o)$ with $o \in O$

- important query types:
  - range queries (search radius r):
    $range(q,r) = \{o \in O | d(q,o) \leq r\}$
  - k-nearest-neighbor queries (desired #hits: k):
    $kNN(q,k) = K$ with $\forall o \in K, o' \in O \setminus K: d(q,o) \leq d(q,o')$ and $|K| = k$

Scenario: source selection for CBIR

resource A

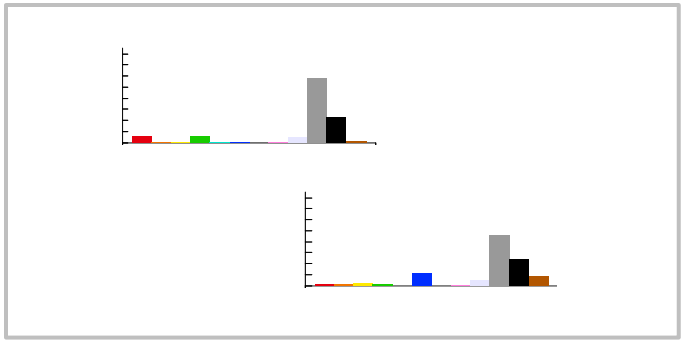

resource B

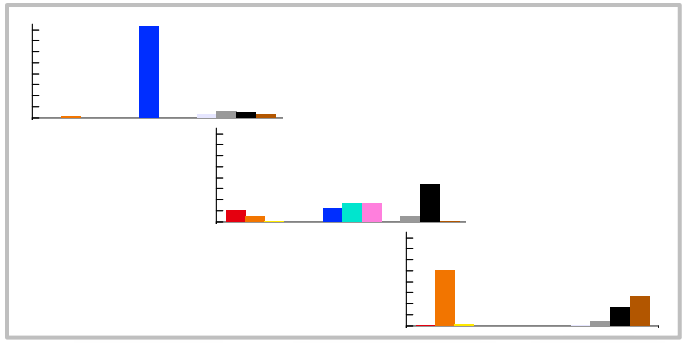

resource C

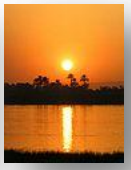

Scenario: source selection for CBIR

resource A

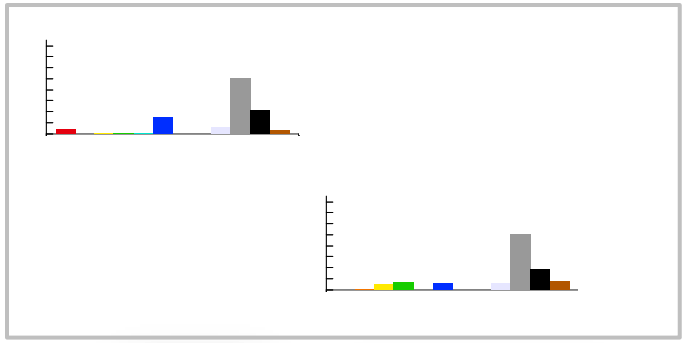


resource B

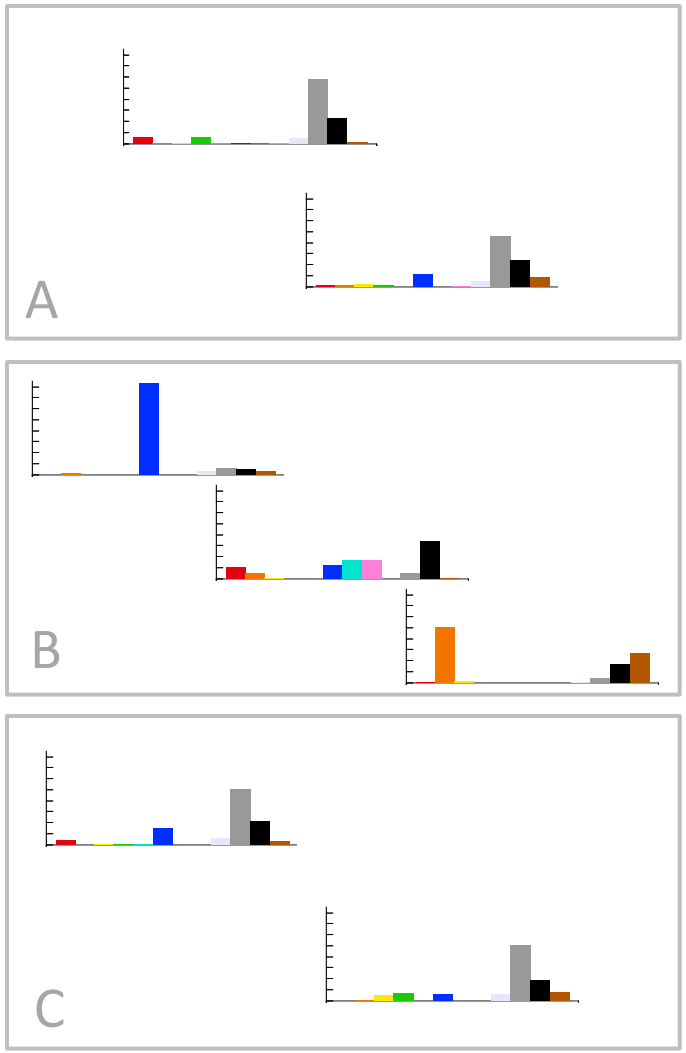




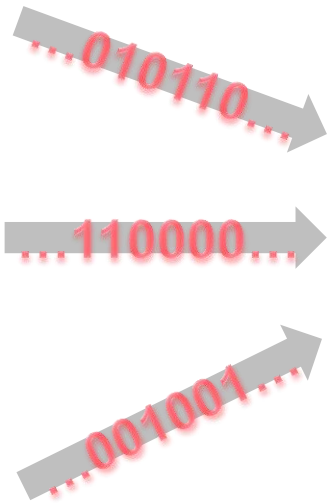
resource C



Scenario: source selection for CBIR



resource description



similarity query

resource selection

1. C
2. A
3. ~~B~~

2. Source Selection: Research Goals and Work in our Group

❶ Design resource selection techniques which improve earlier work w.r.t. efficiency:

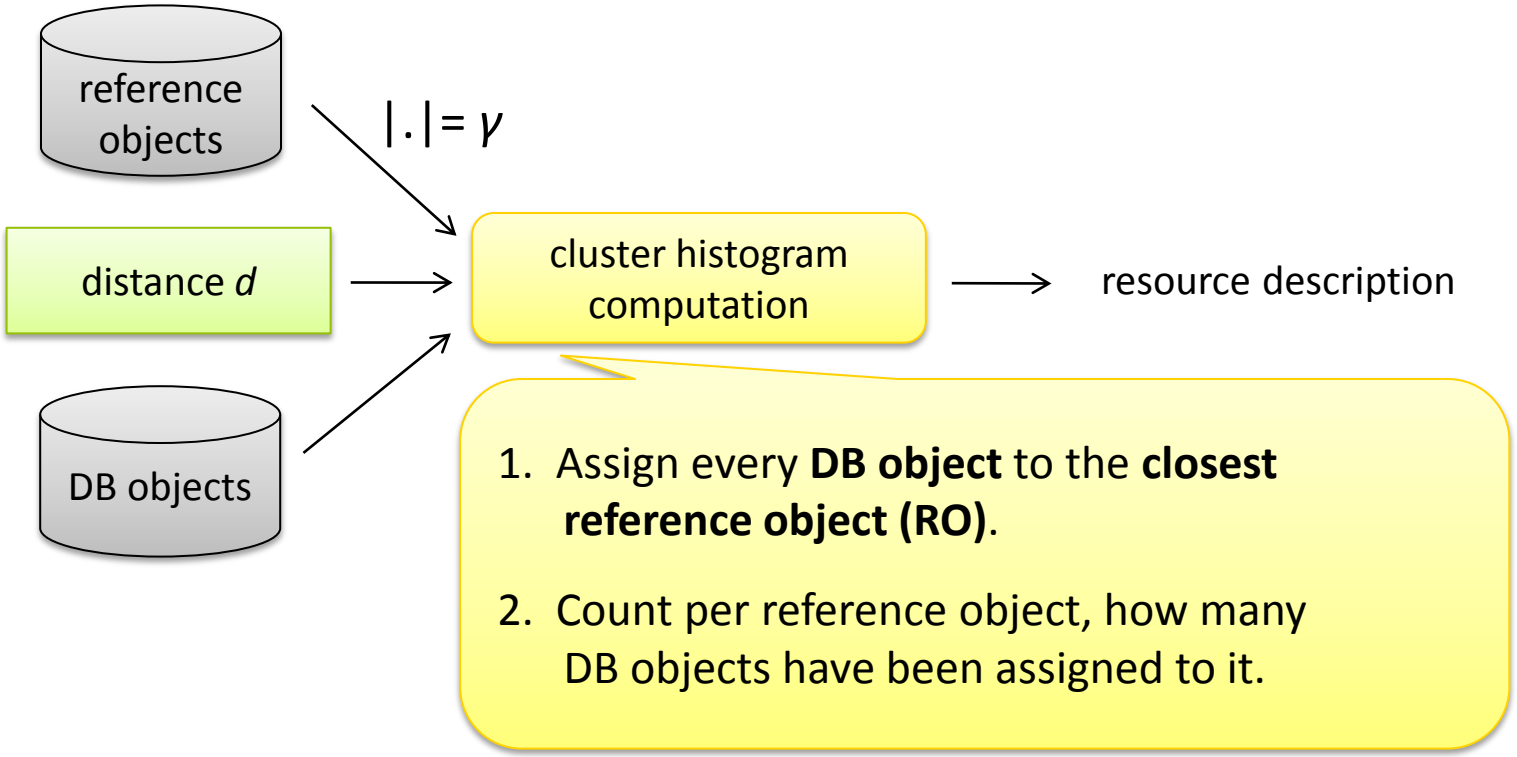
Space efficiency and selectivity of the resource descriptions

❷ Visualize the source selection process

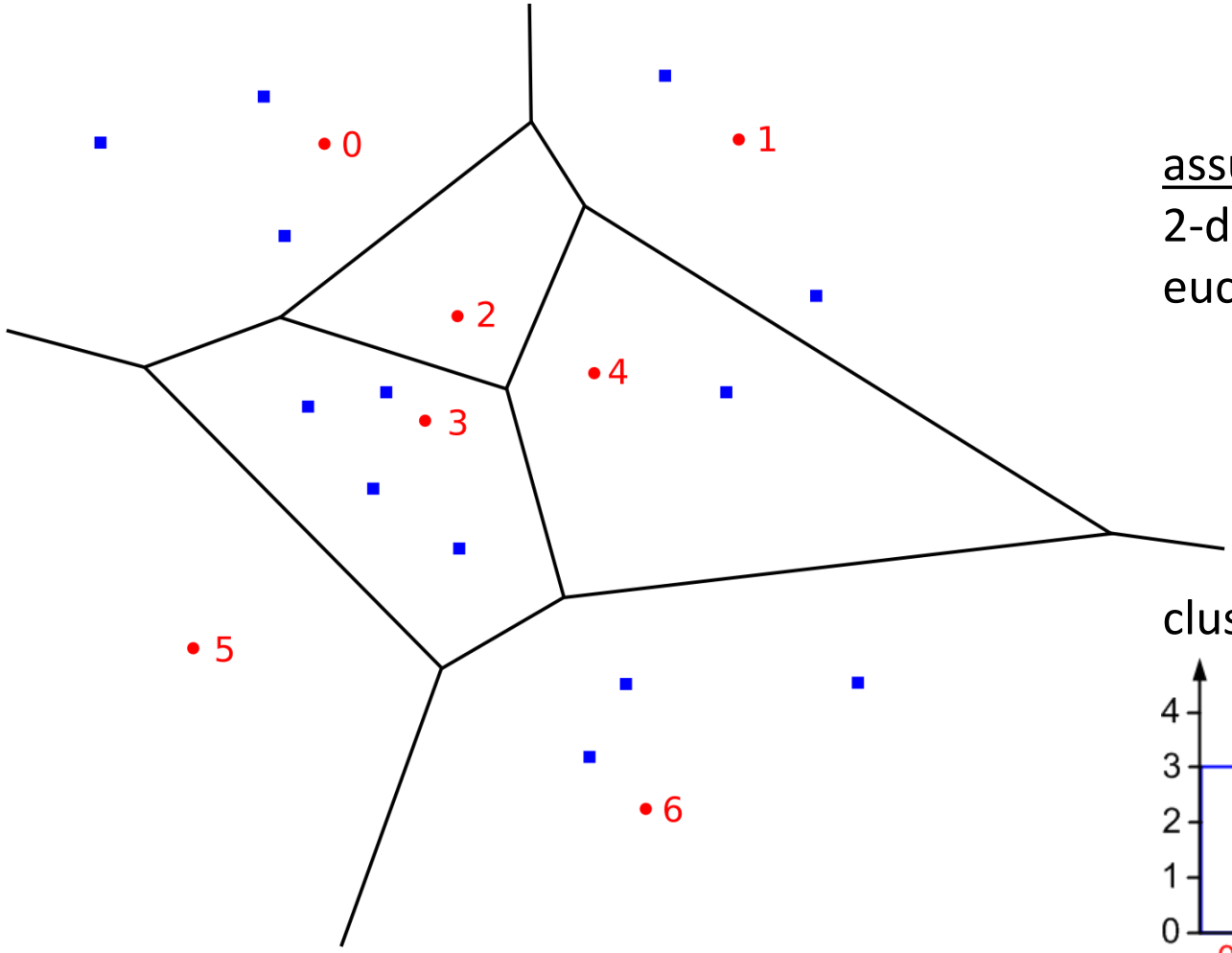
❸ Show usefulness of the techniques in other scenarios (apart from distributed IR: metric access methods, geo IR, …)

❶ Efficient source selection with HFS and UFS

Baseline: [MEH05b]

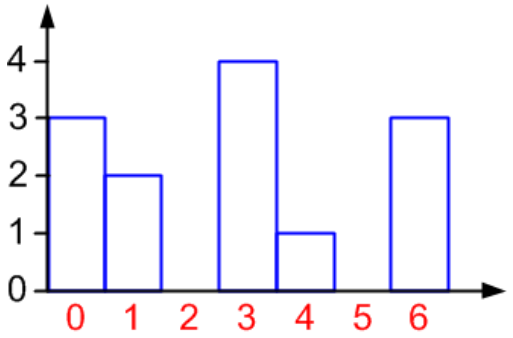


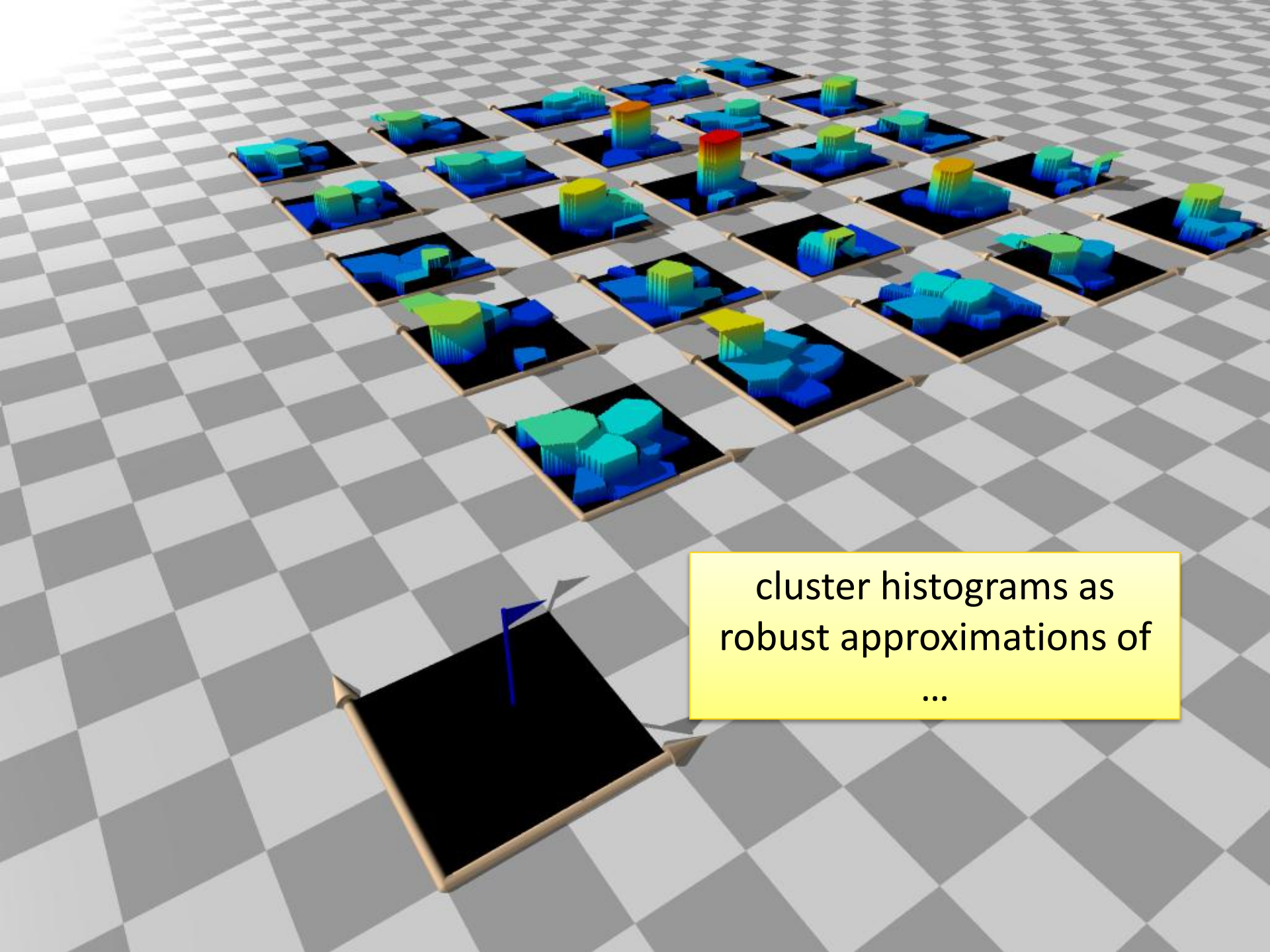
❶ Efficient source selection with HFS and UFS

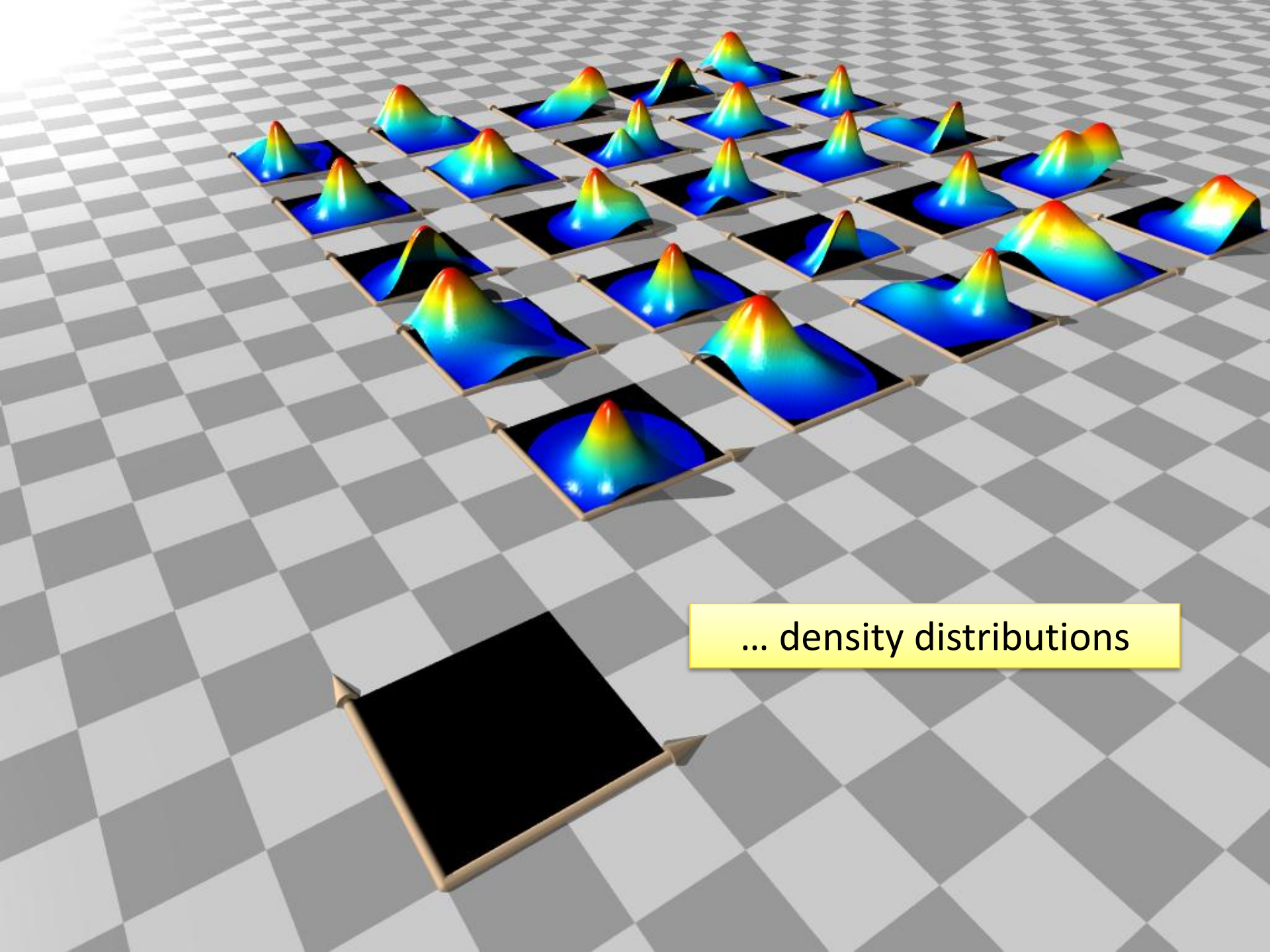


assumption:
2-dim. feature space
euclidean distance

cluster histogram:

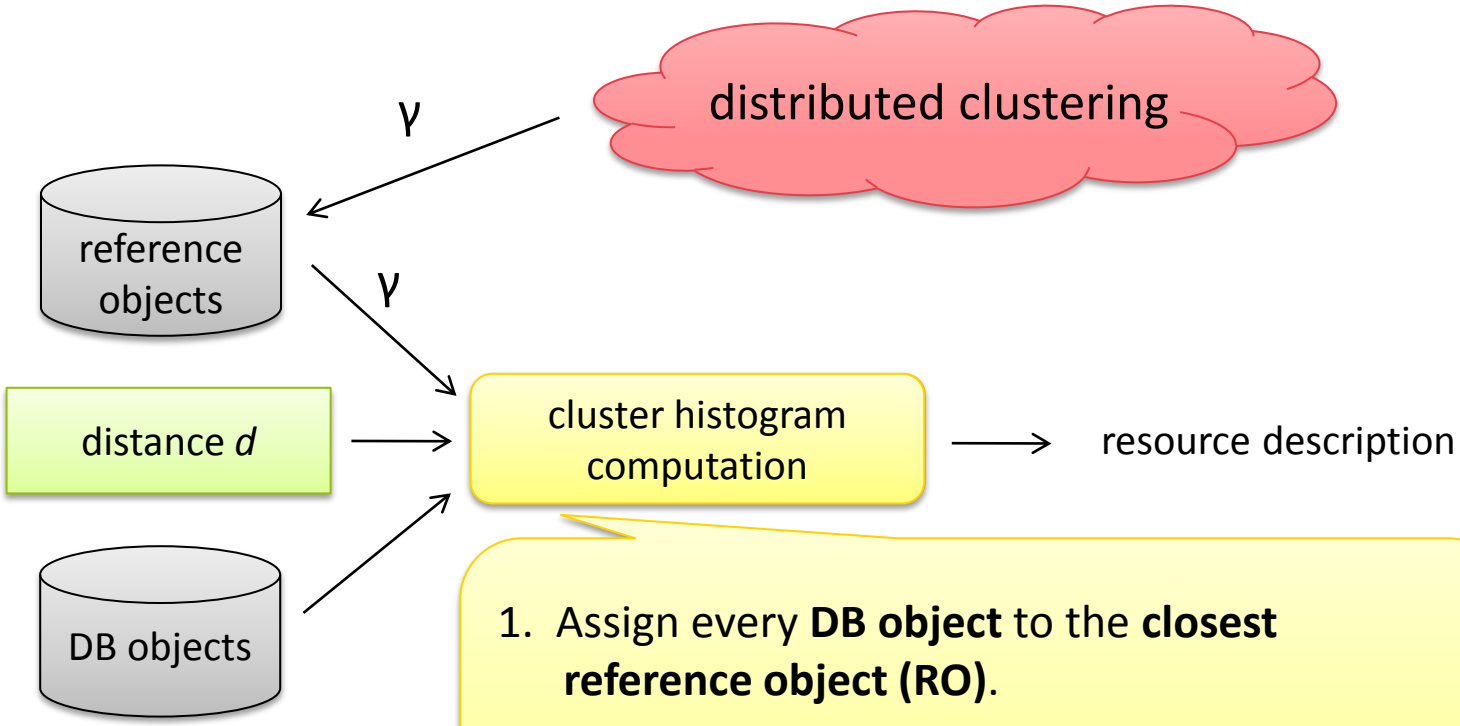



cluster histograms as
robust approximations of
…


… density distributions

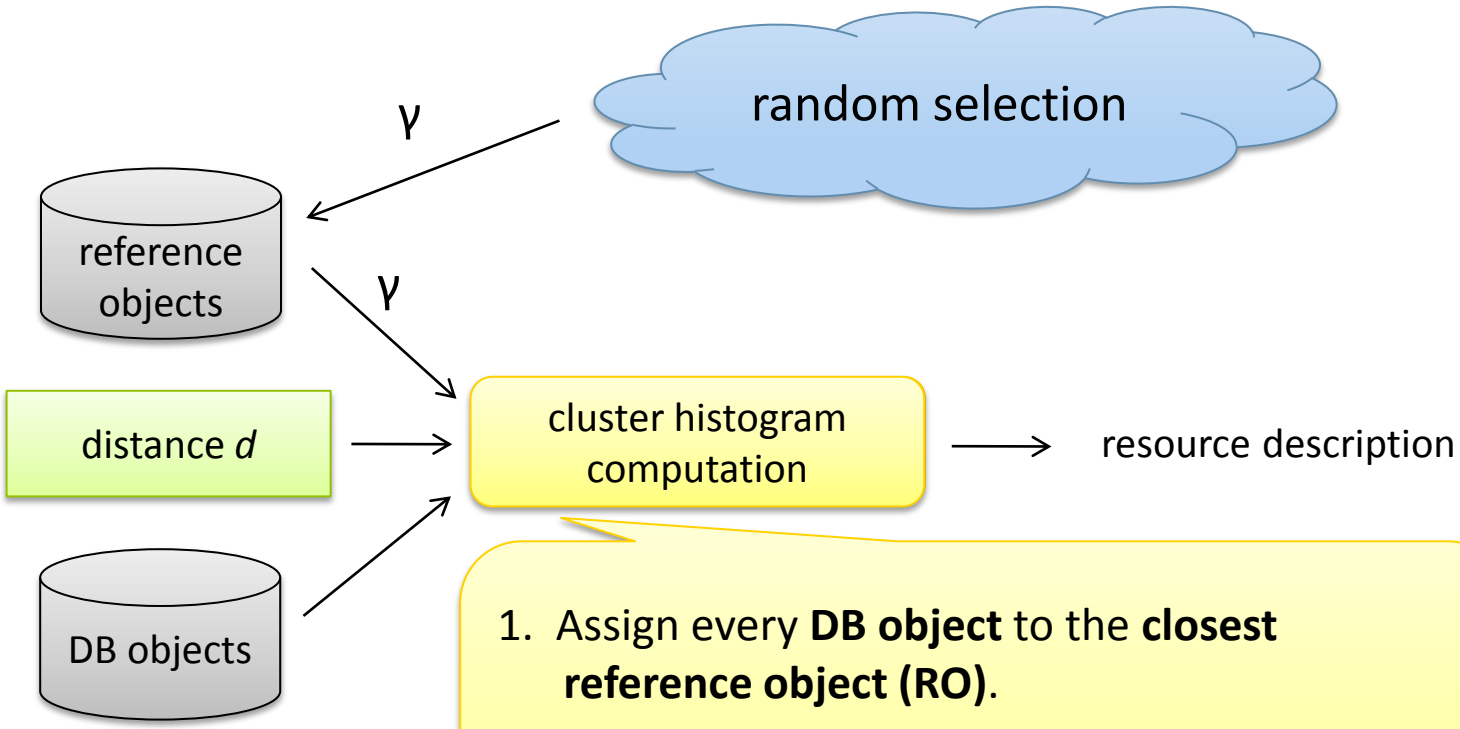
❶ Efficient source selection with HFS and UFS

[MEH05b]



❶ Efficient source selection with HFS and UFS

[EMH+06]



❶ Efficient source selection with HFS and UFS

Source Selection as in [EMH+06]: version 1 for small γ

1. compute the reference object c_j closest to q w.r.t. $d(q, c_j)$
2. ranking of peers

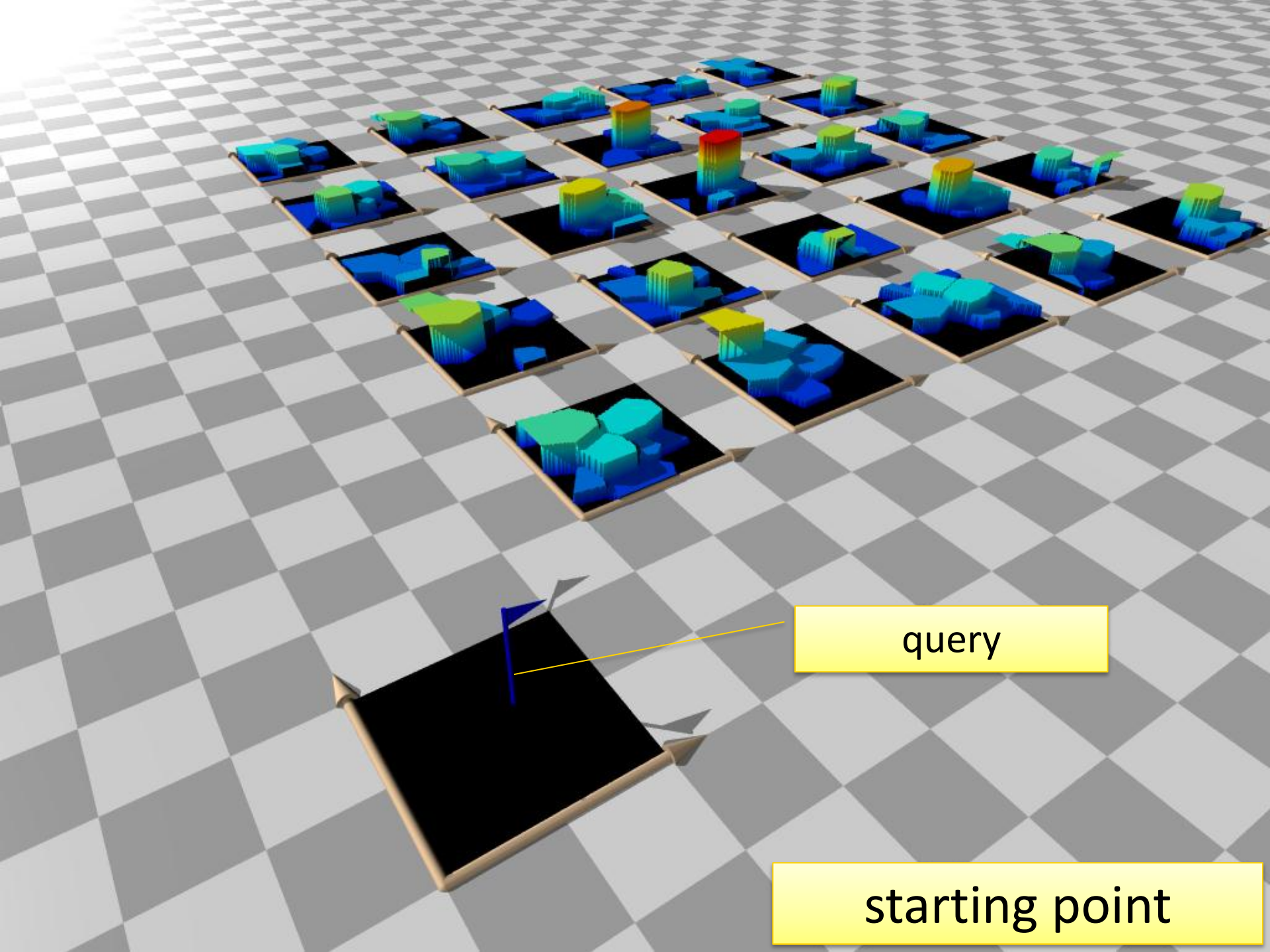
   peer p_a has more DB objects in cluster c_j than p_b

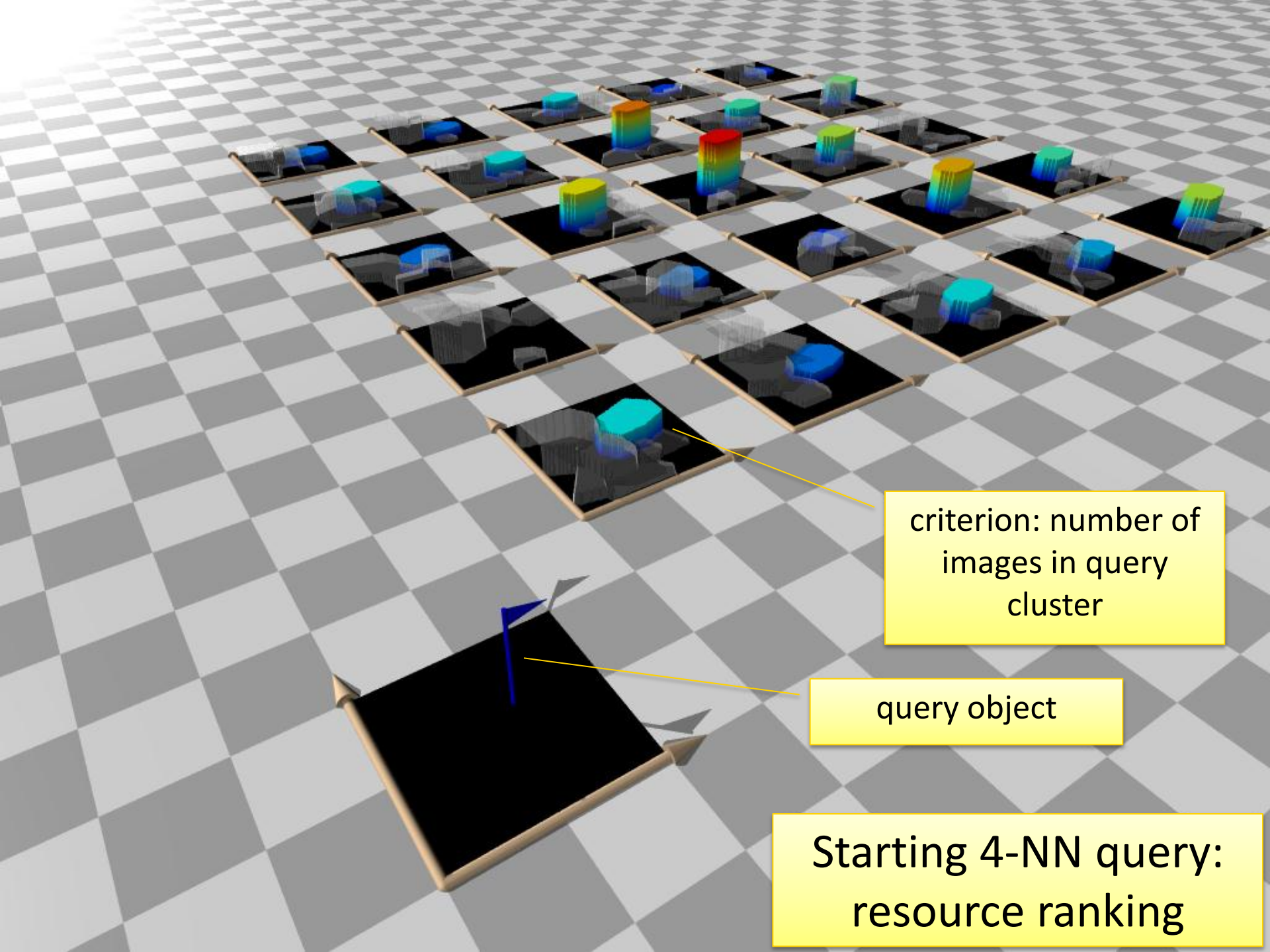
   $\Rightarrow$ p_a is contacted before p_b

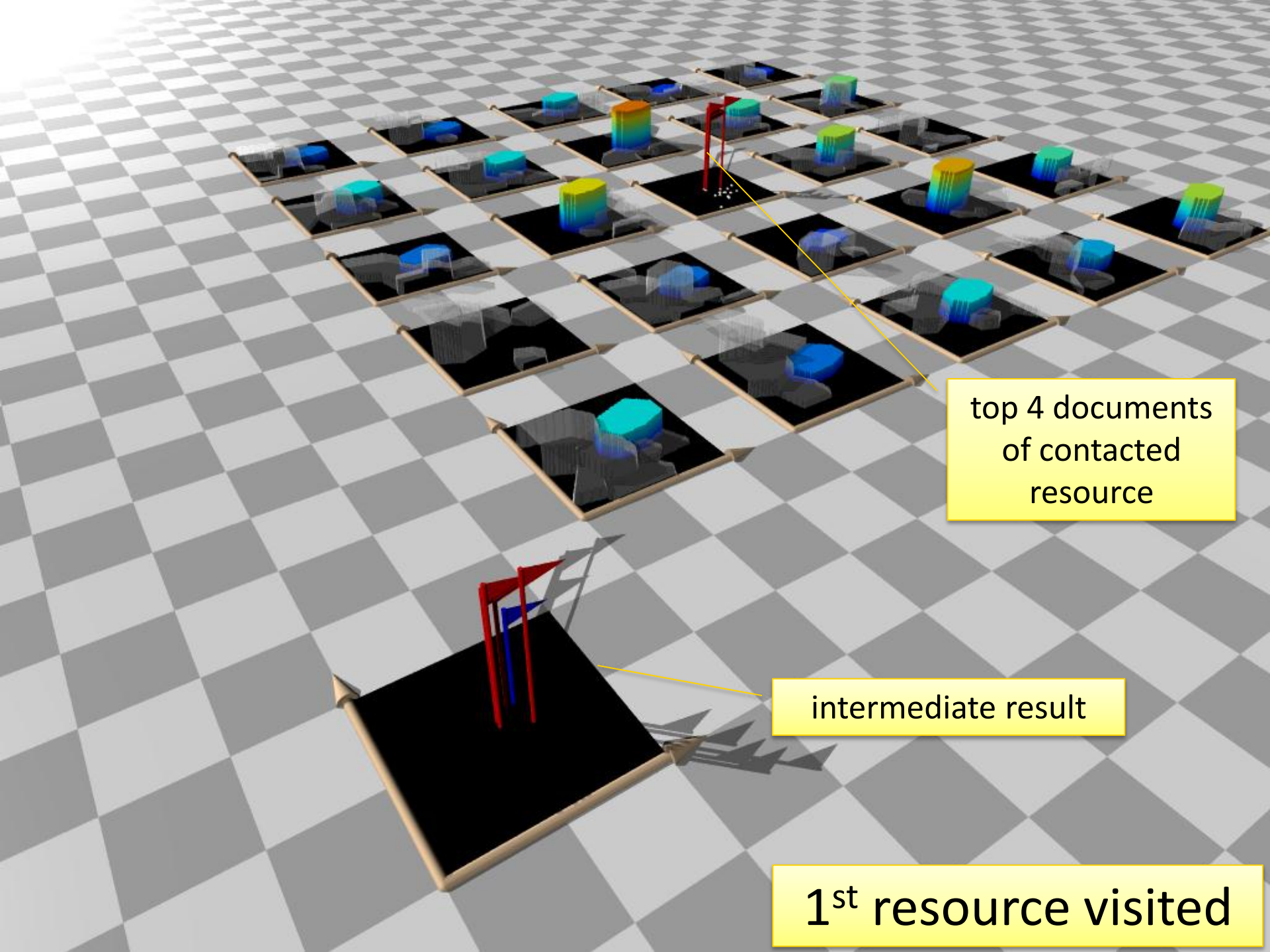
   peer p_b has more DB objects in cluster c_j than p_a

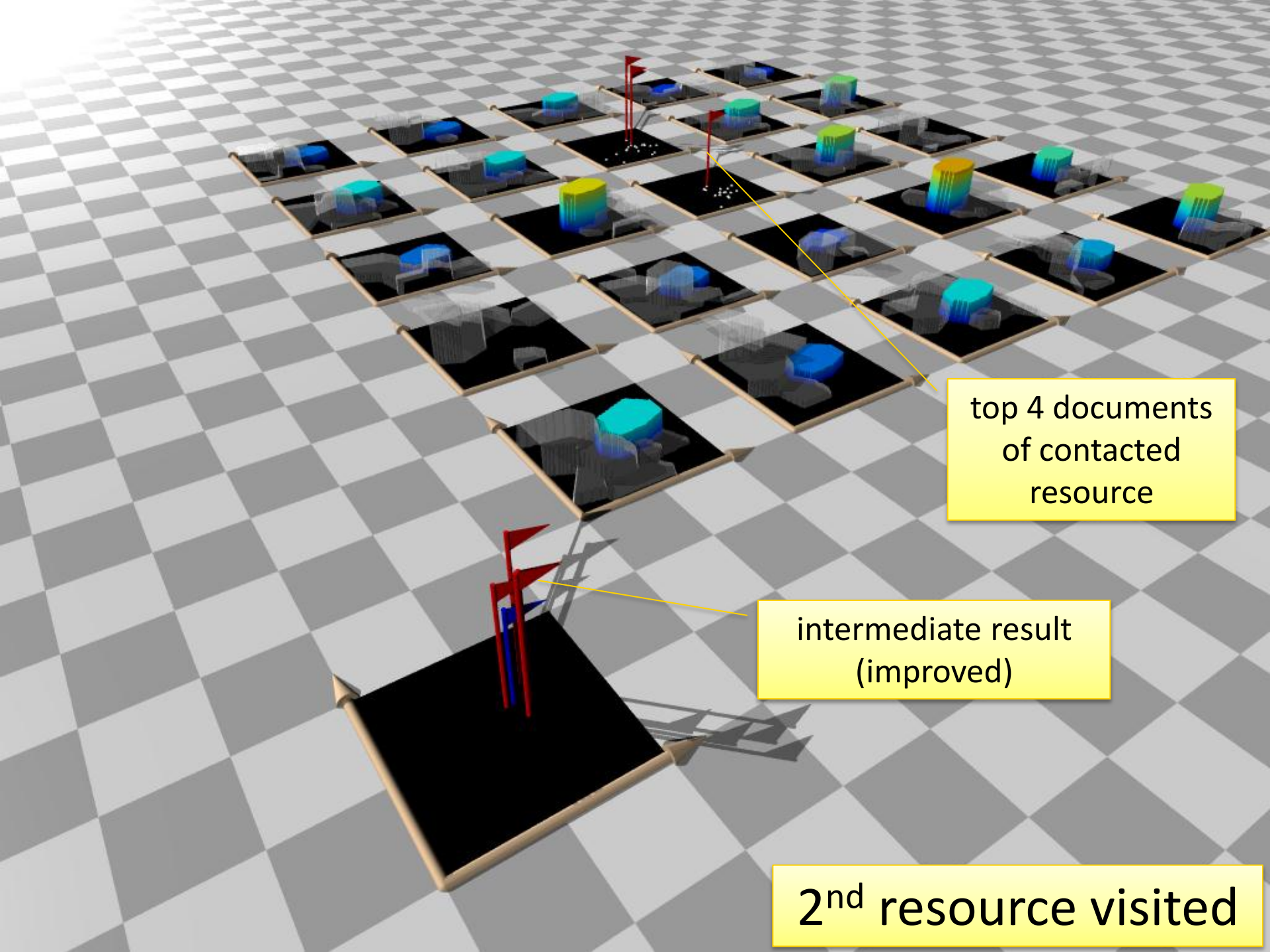
   $\Rightarrow$ p_b is contacted before p_a

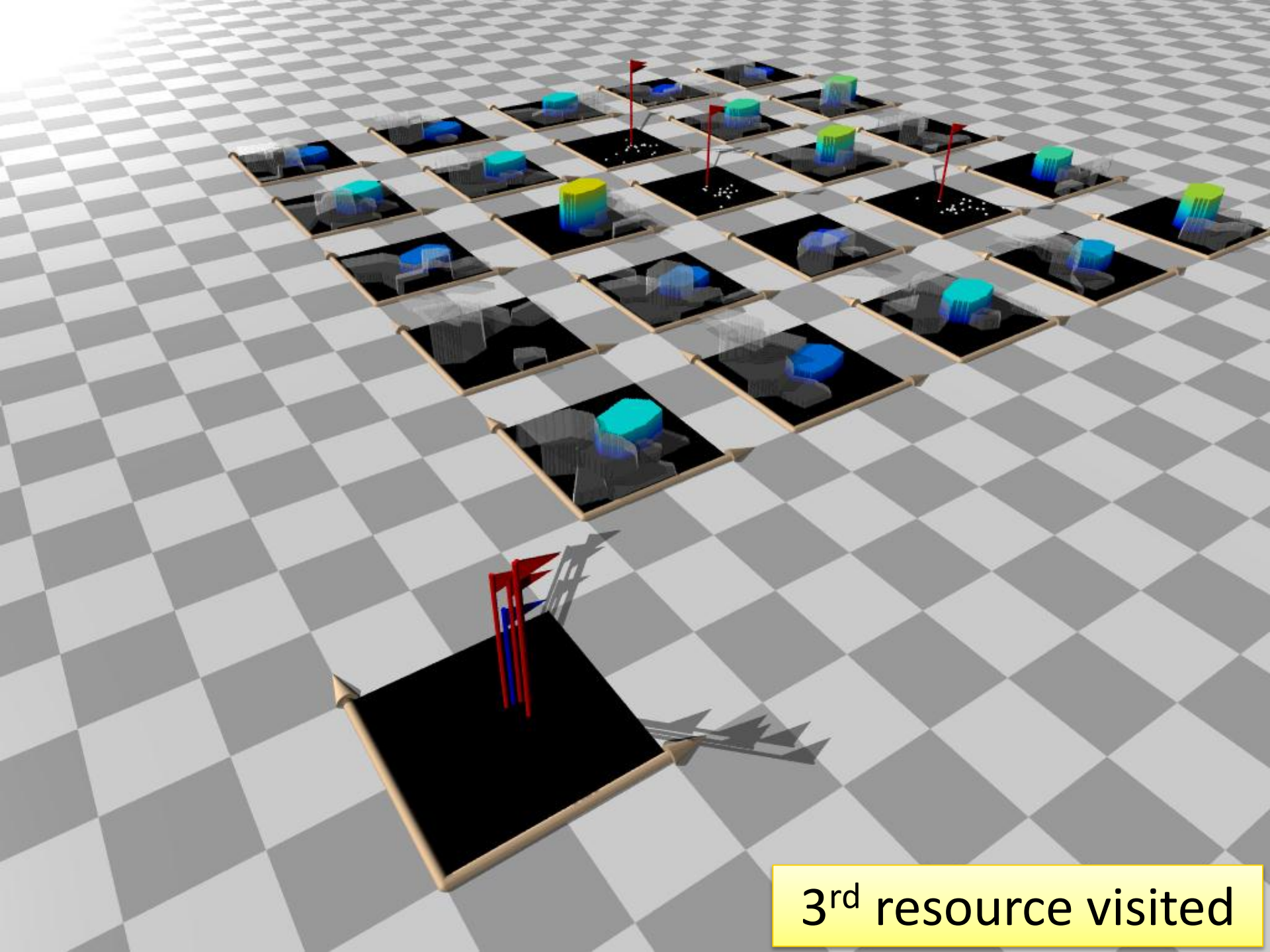
   otherwise (same number of objects in c_j) a random decision is made

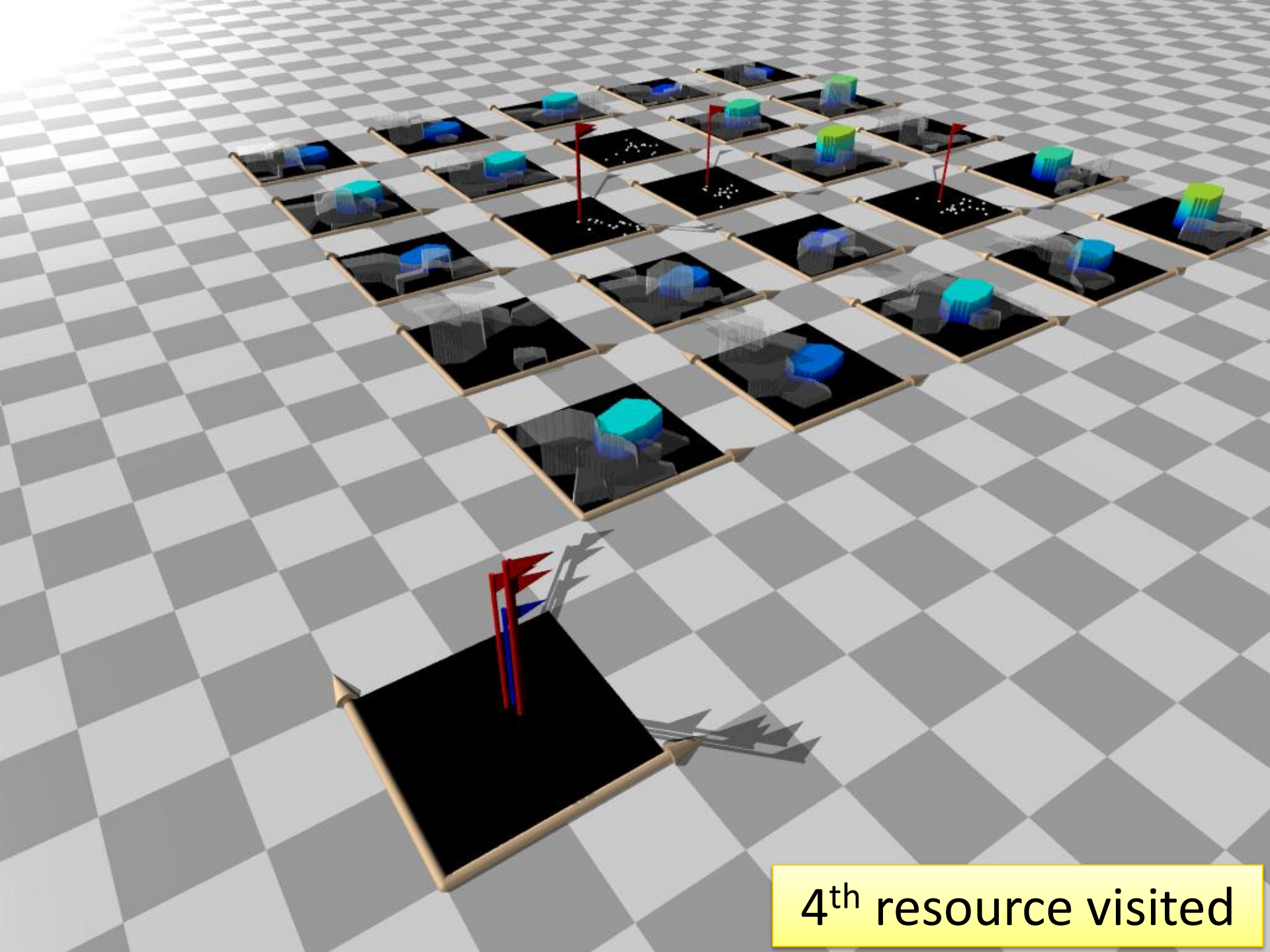

query
starting point

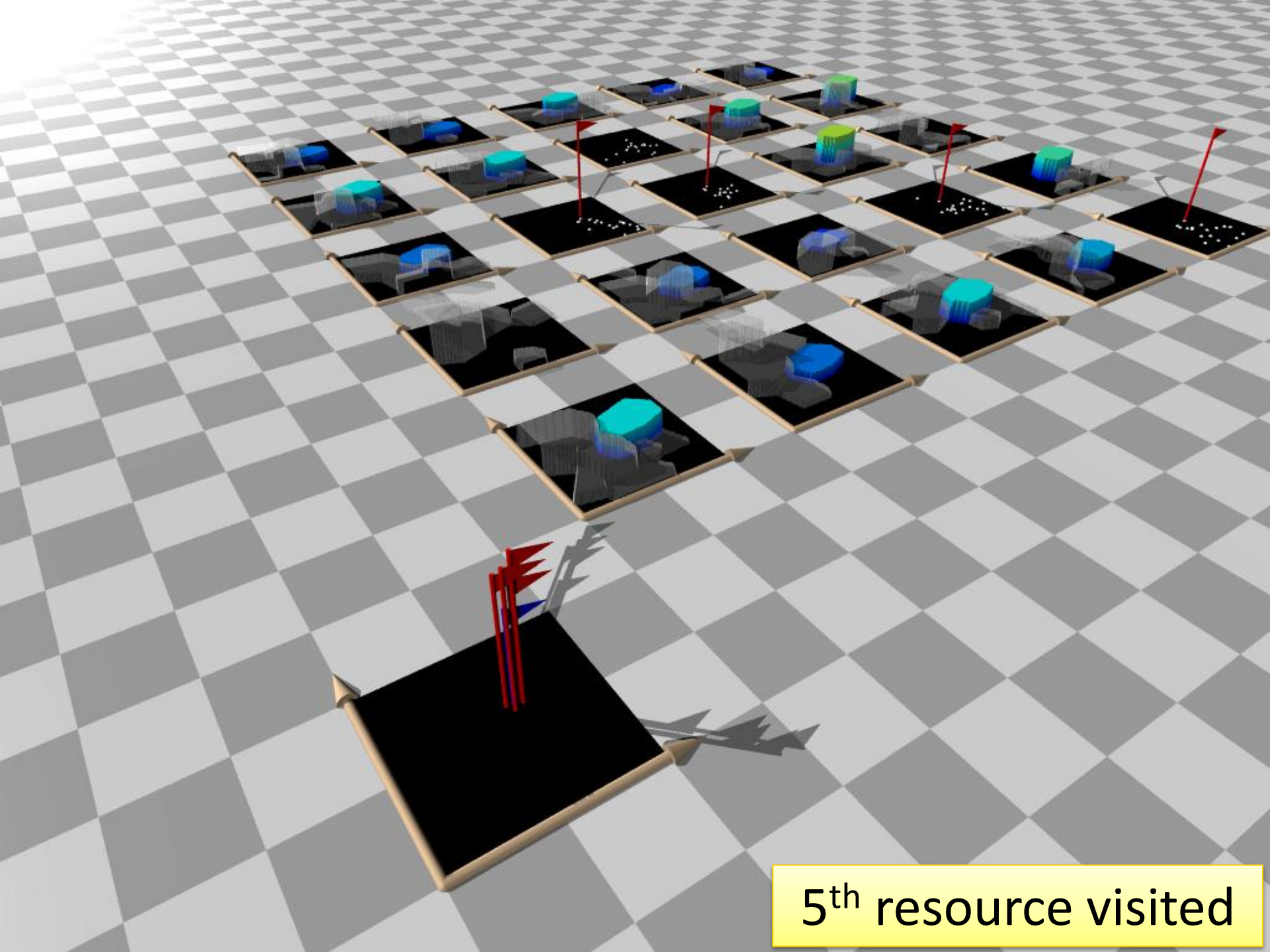

criterion: number of images in query cluster
query object
Starting 4-NN query: resource ranking


top 4 documents of contacted resource
intermediate result
1st resource visited


top 4 documents of contacted resource
intermediate result (improved)
2nd resource visited

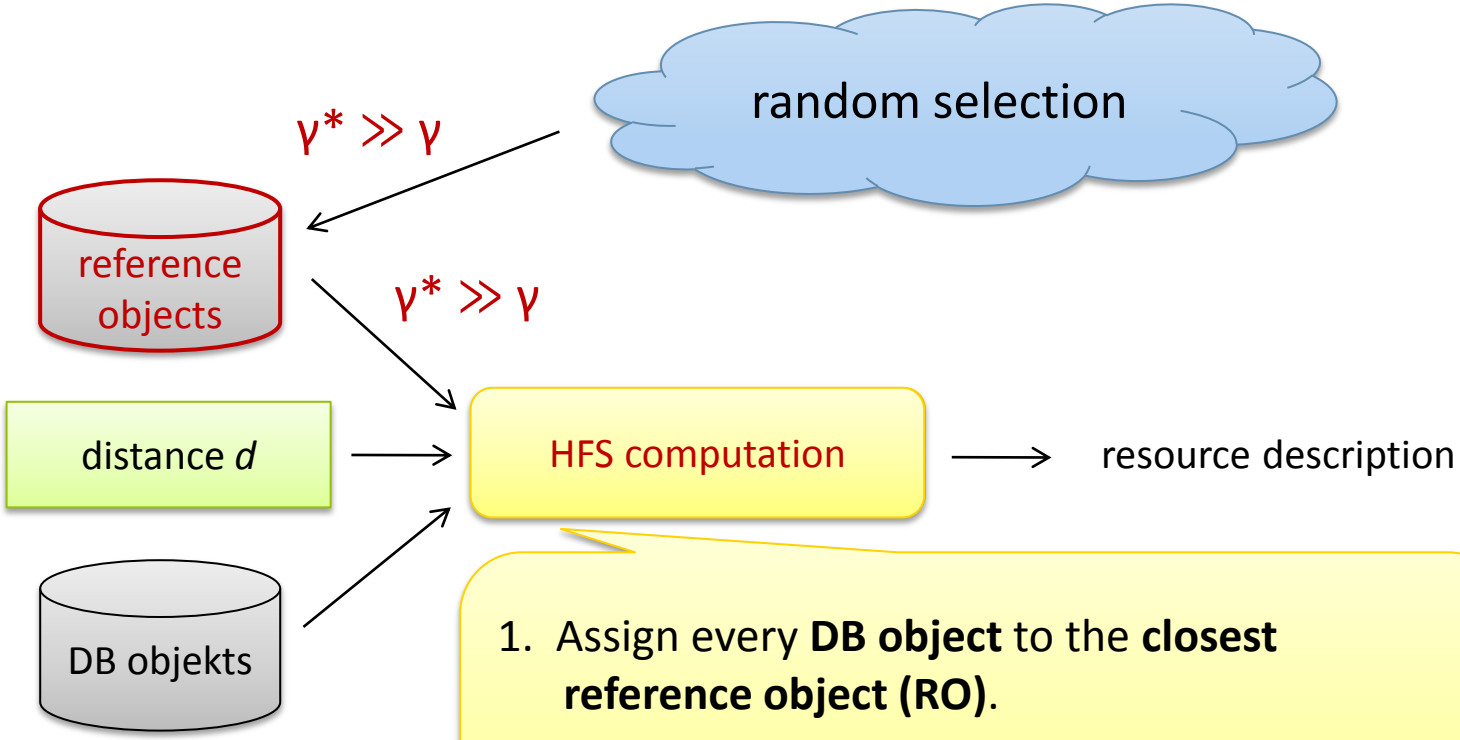

3rd resource visited

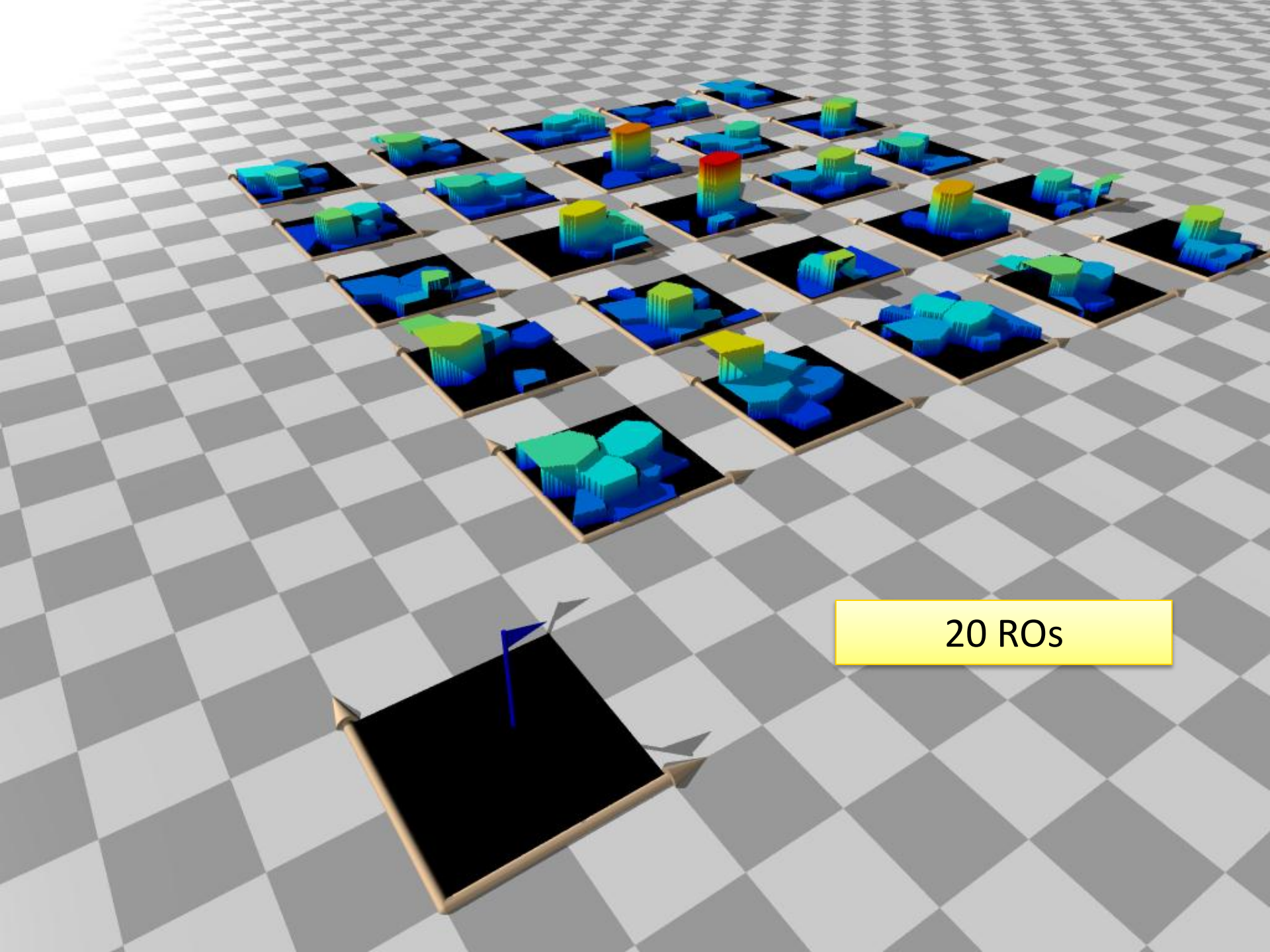

4th resource visited

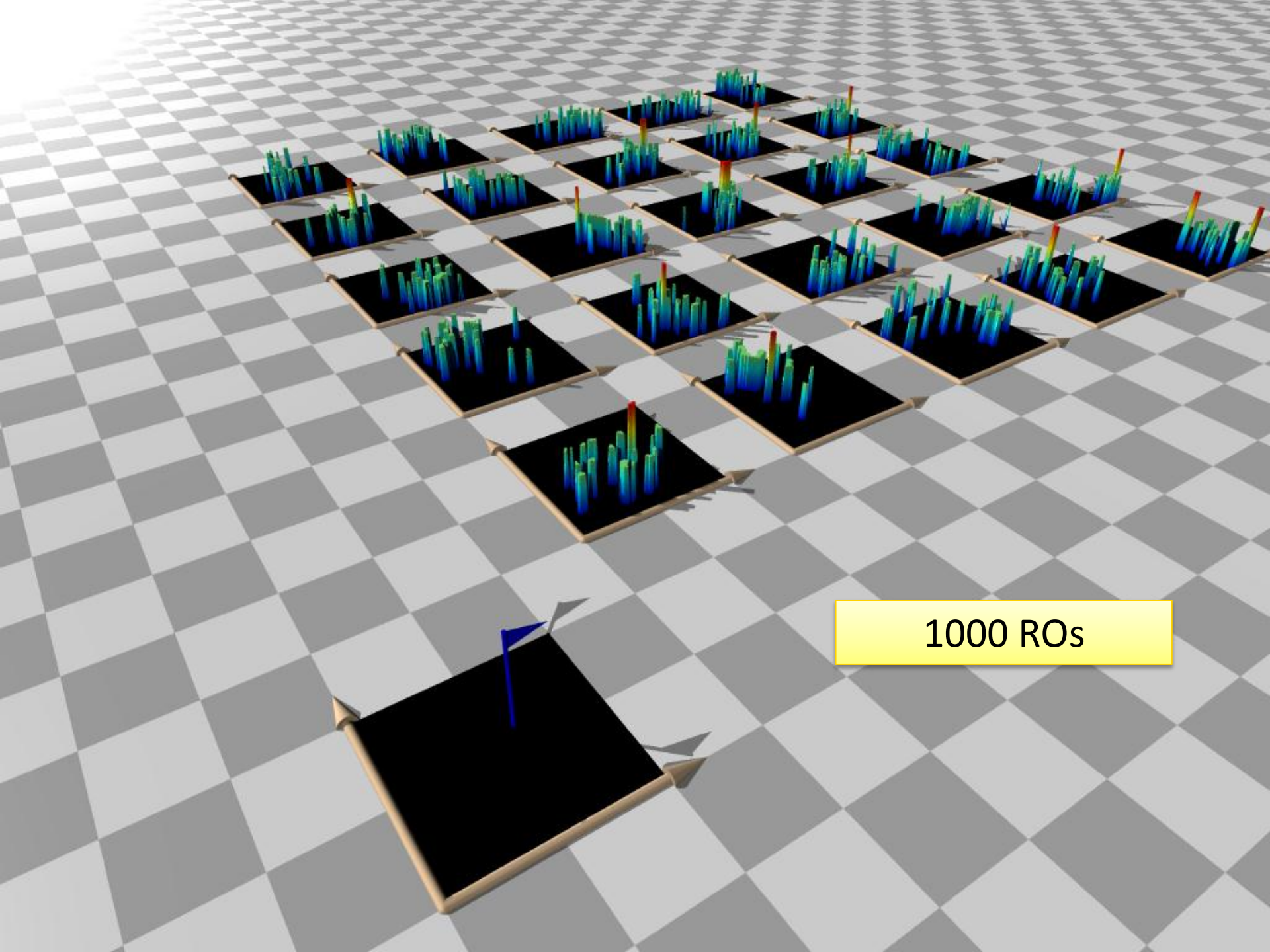

5th resource visited

❶ Efficient source selection with HFS and UFS

[BEM+07]




20 ROs


1000 ROs

❶ Efficient source selection with HFS and UFS

Source Selection as in [EMH+06]: version 2 for higher values of γ

1. compute list L

   sort reference objects c_j ascending w.r.t. $d(q, c_j)$

2. ranking of peers

   Iterate over the list L (from beginning to end):

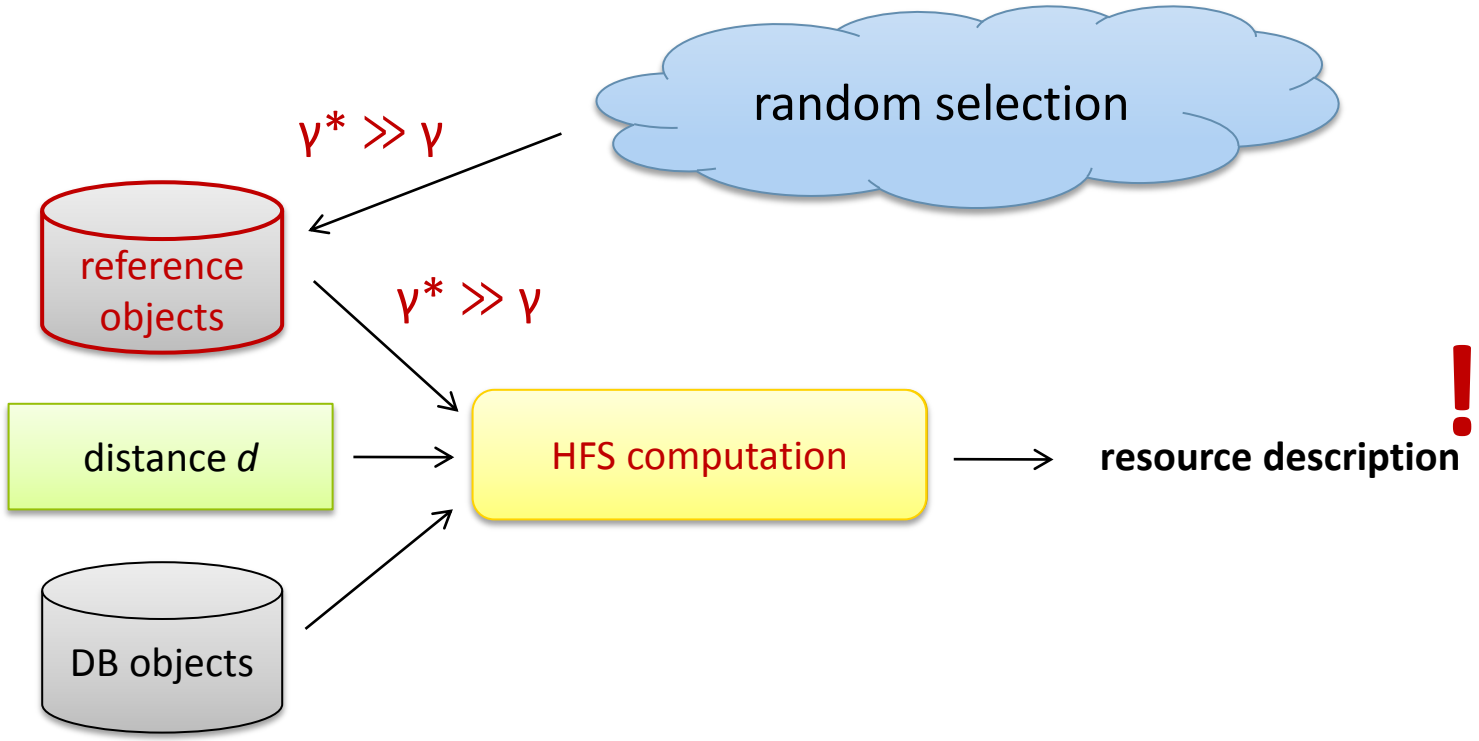
   peer p_a has more DB objects in current cluster than p_b
   $\Rightarrow p_a$ is contacted before p_b

   peer p_b has more DB objects in current cluster than p_a
   $\Rightarrow p_b$ is contacted before p_a

   otherwise (same number of objects in all clusters) random decision is made

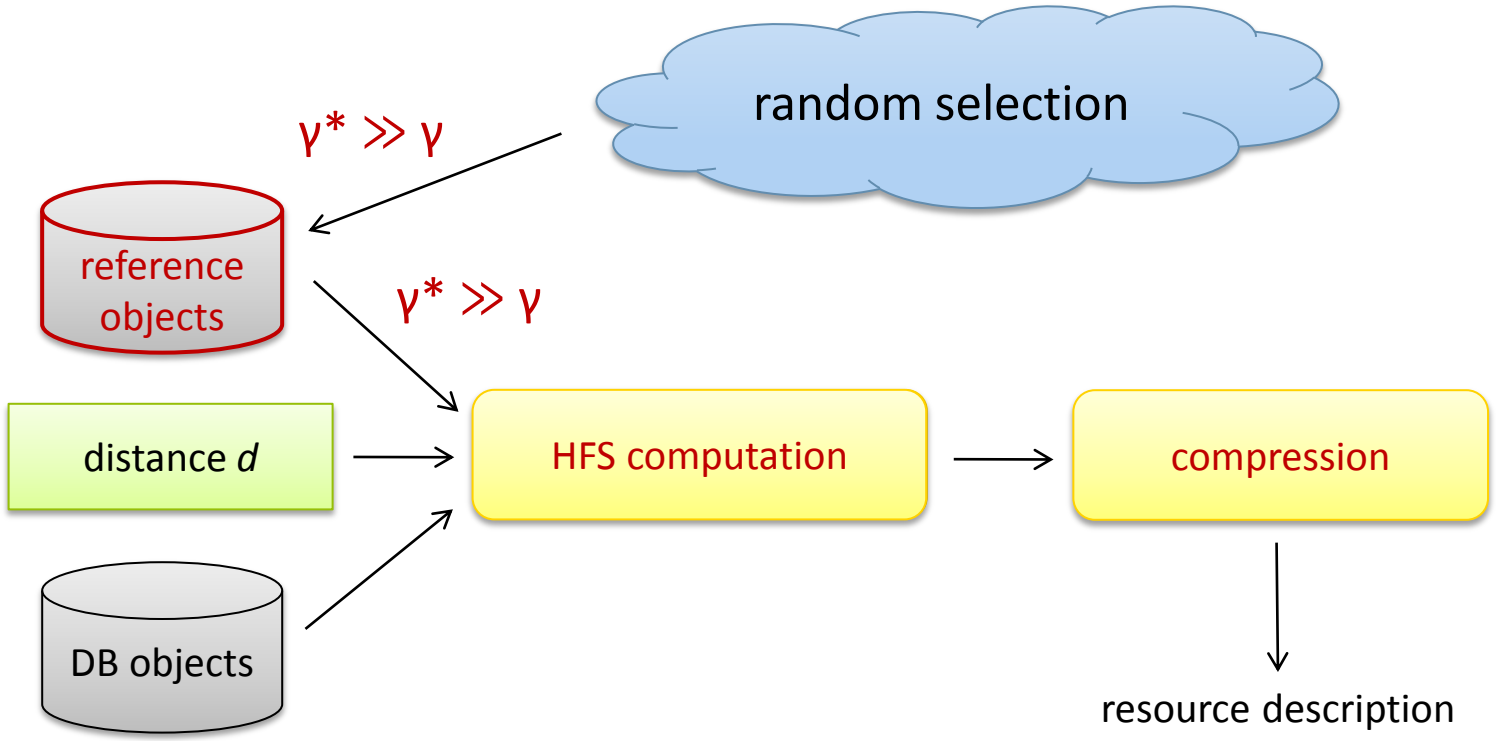
❶ Efficient source selection with HFS and UFS

[BEM+07]



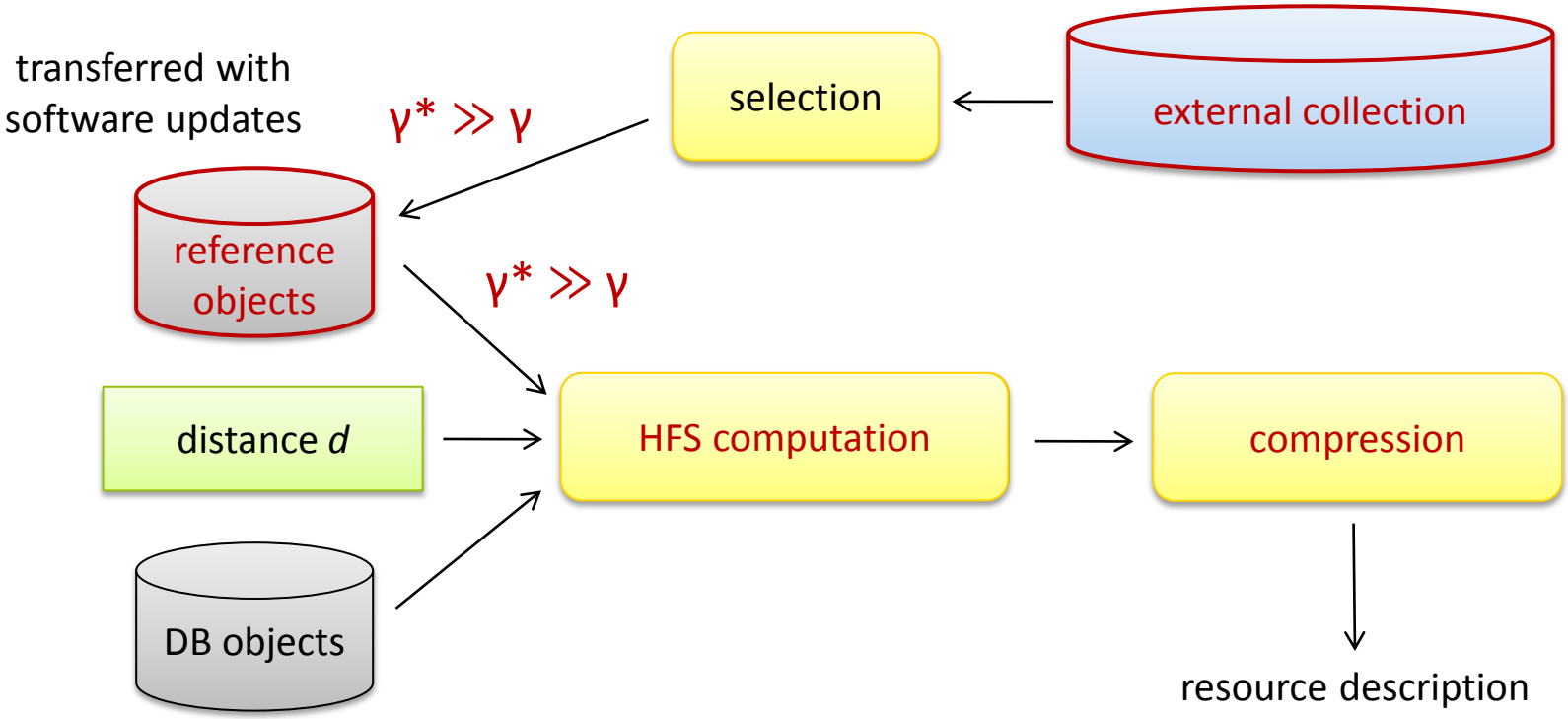
❶ Efficient source selection with HFS and UFS

[BEM+07]



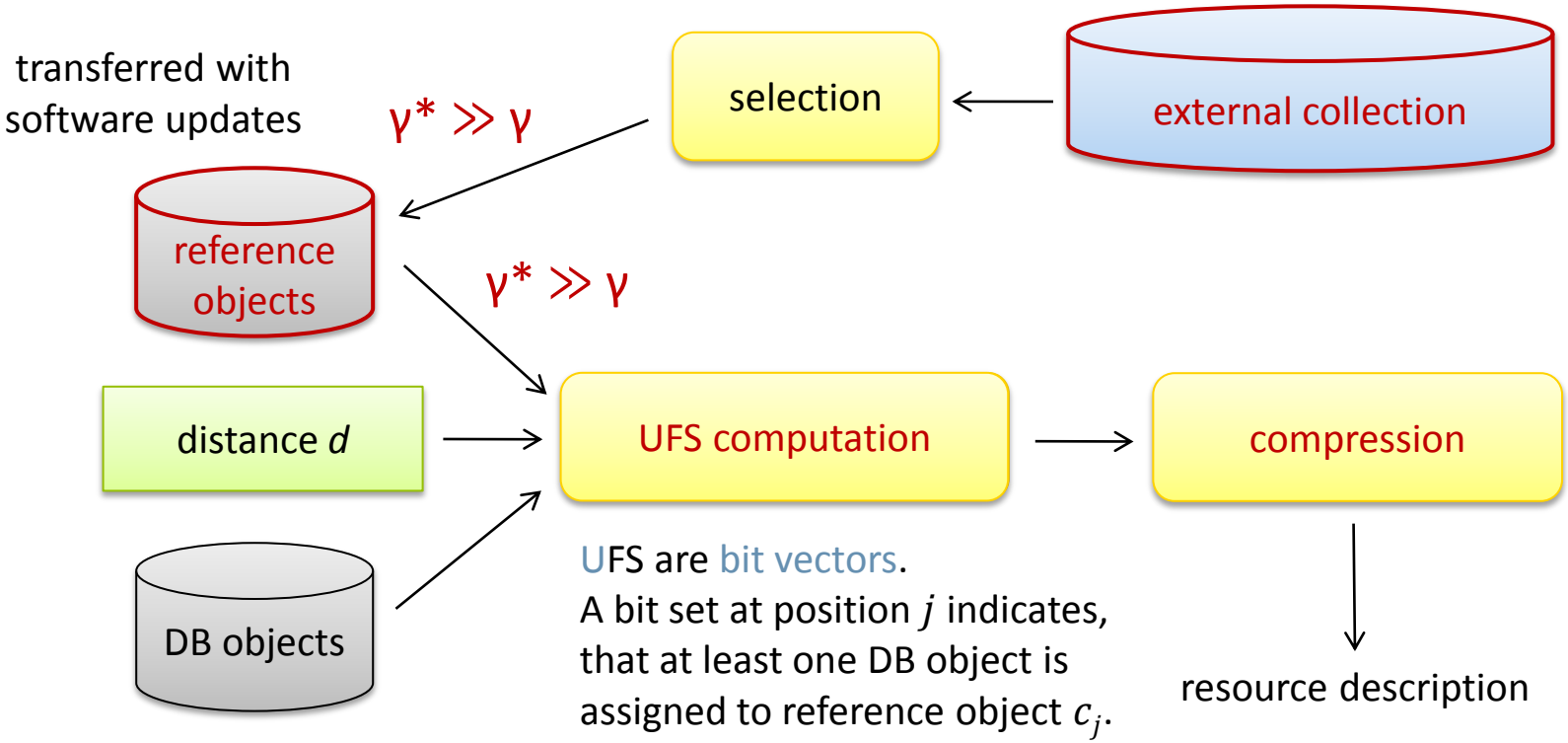
❶ Efficient source selection with HFS and UFS

[BEM+07]



❶ Efficient source selection with HFS and UFS

[BH10a, BH10b]



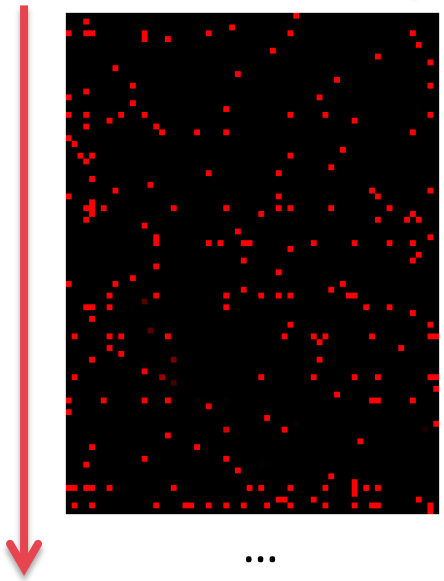
❷ Visualizing the source selection process

Another view on the source selection process with UFS:

source selection as an $\gamma \times n$ image.

Set of ROs: $C = \{c_j | 1 \leq j \leq \gamma\}$,
(1 RO c_j with cluster $[c_j]$ per column)

Set of peers:
$P = \{p_i | 1 \leq i \leq n\}$,
(1 summary $s^i[j]$ per row)



...

$s^i[j] = 0$
p_i has NO docs in $[c_j]$

$s^i[j] > 0$
p_i has ≥ 1 doc(s) in $[c_j]$

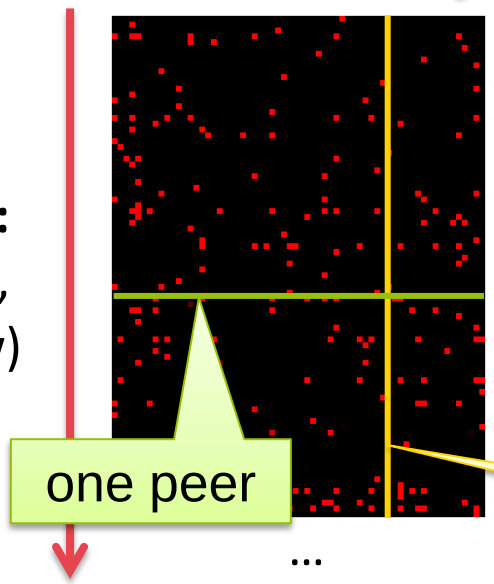
❷ Visualizing the source selection process

Another view on the source selection process with UFS:

source selection as an $\gamma \times n$ image.

Set of ROs: $C = \{c_j | 1 \leq j \leq \gamma\}$,
(1 RO c_j with cluster $[c_j]$ per column)

Set of peers:
$P = \{p_i | 1 \leq i \leq n\}$,
(1 summary $s^i[j]$ per row)

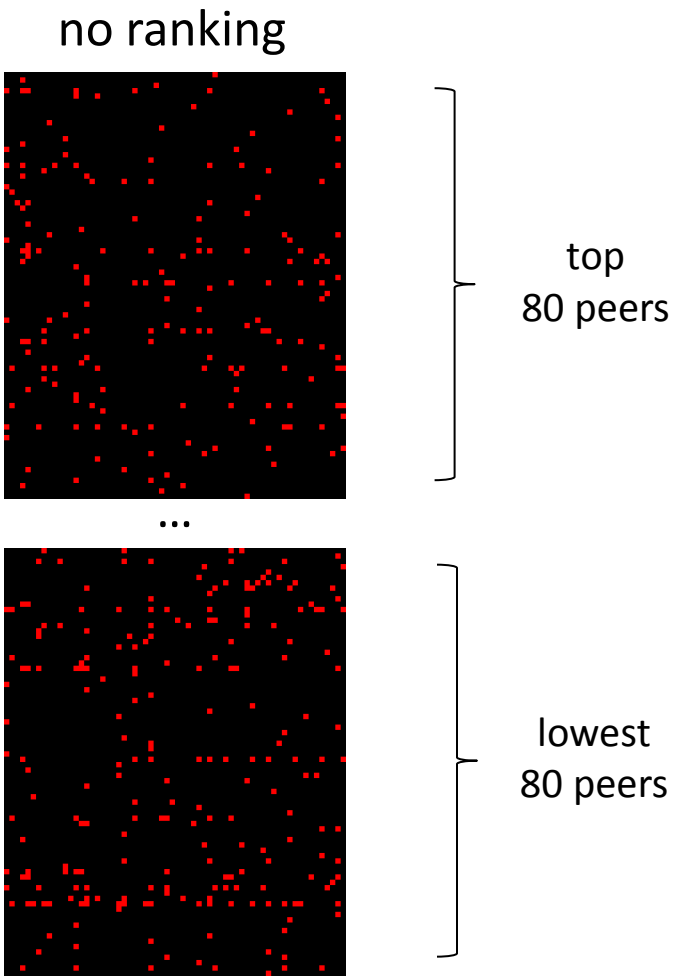


$s^i[j] = 0$
p_i has NO docs in $[c_j]$

$s^i[j] > 0$
p_i has ≥ 1 doc(s) in $[c_j]$

one cluster / reference object

❷ Visualizing the source selection process

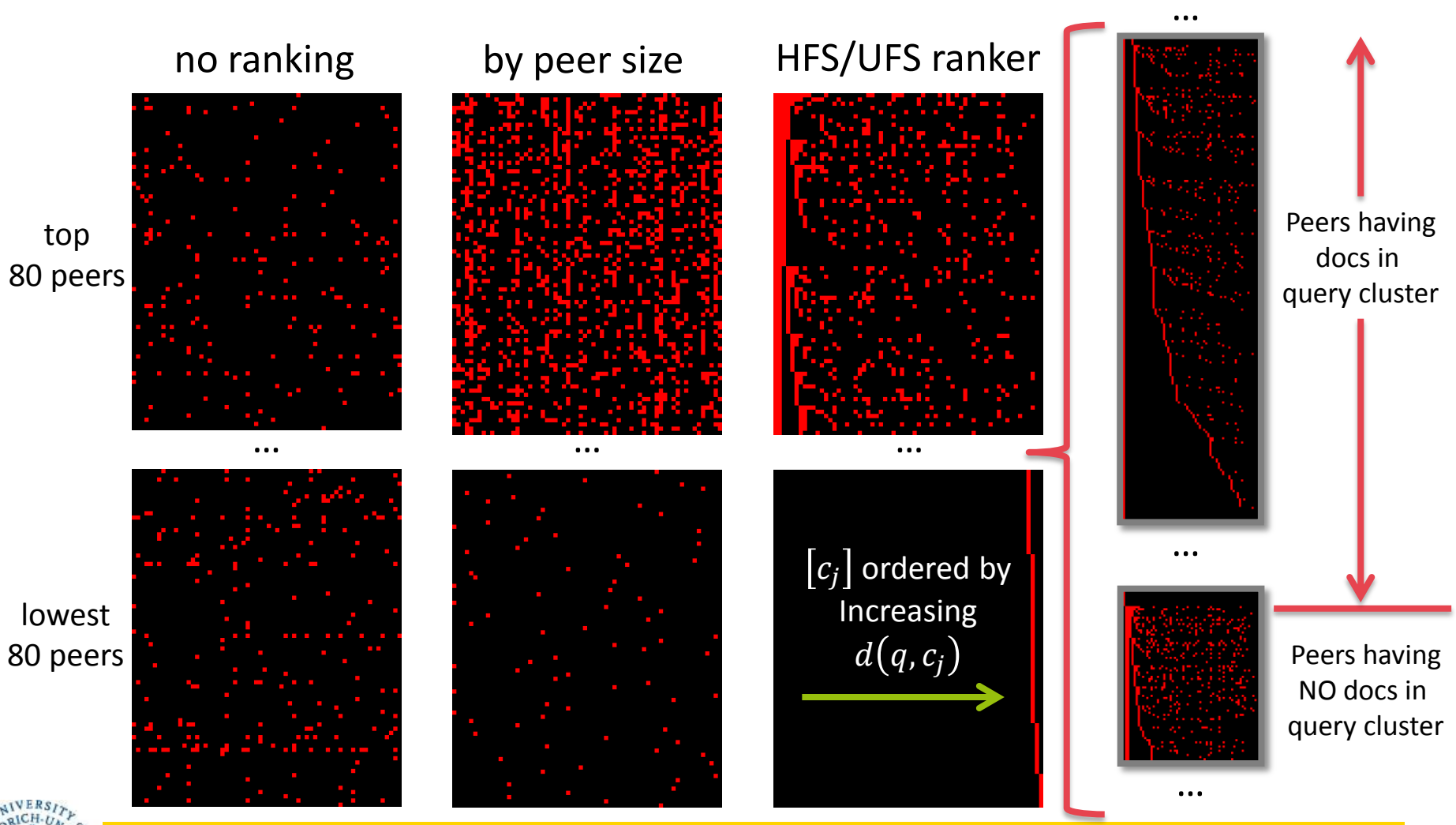


MIRFLICKR 25k collection:
http://press.liacs.nl/mirflickr/

24.258 images assigned to 9.862 peers based on the Flickr-UserID

CEDD descriptor + Hellinger distance

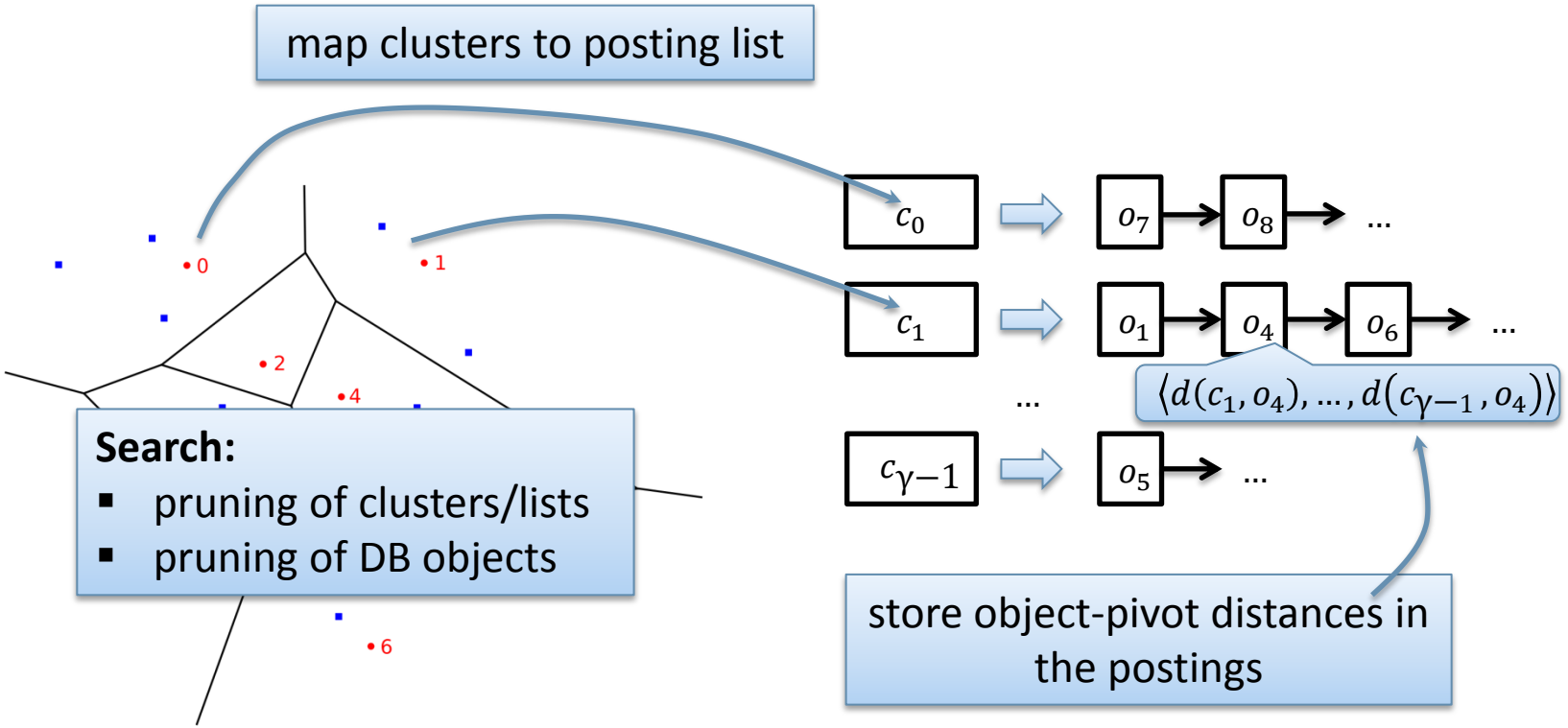
❷ Visualizing the source selection process



❸ Applicability in other application fields

exact search

a) Design of a MAM based on inverted files:



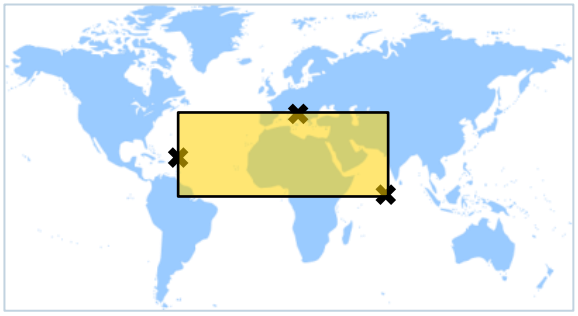
❸ Applicability in other application fields

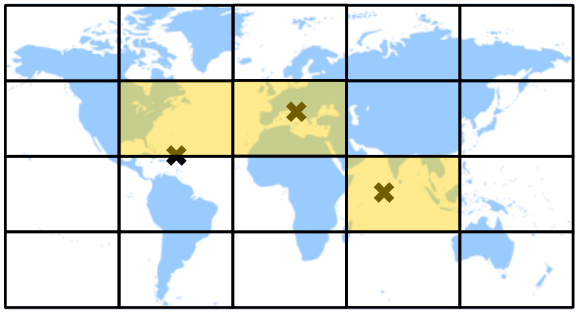
exact search

b) Apply HFS/UFS and extensions in the context of geographic data and compare it with:

geometric approaches

space partitioning approaches



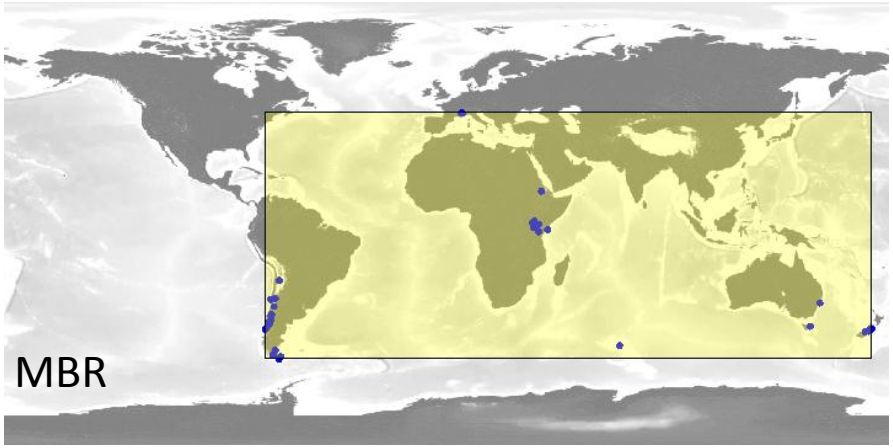


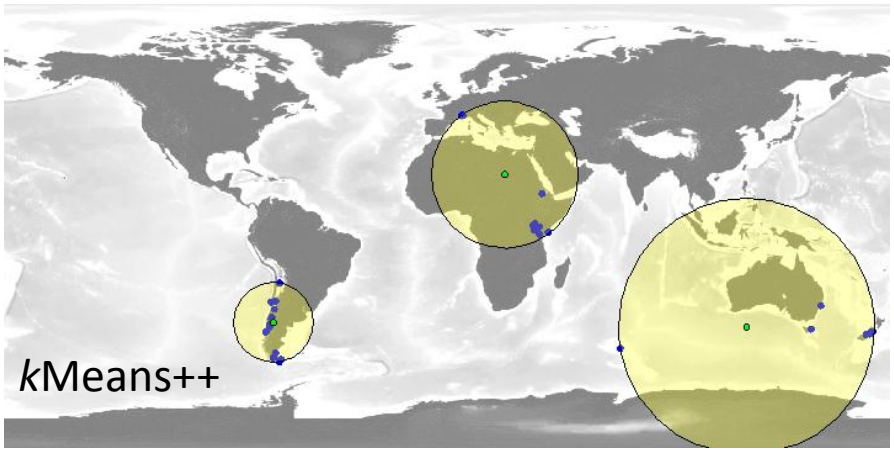
✖ = peer data

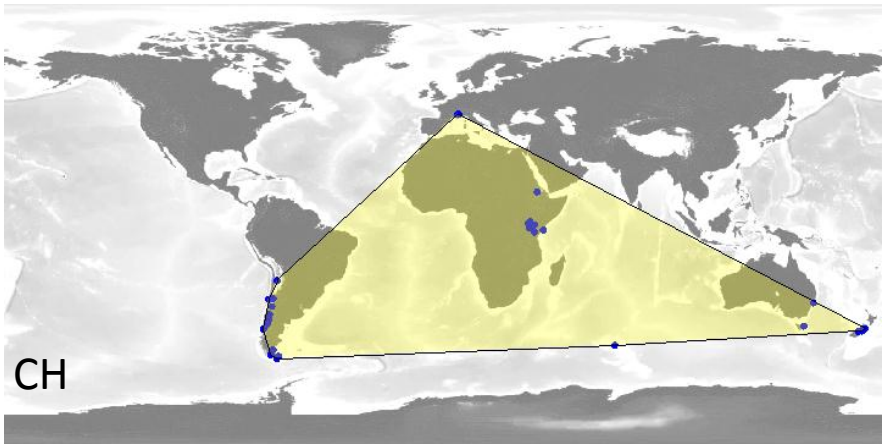
→ also: hybrid combinations!

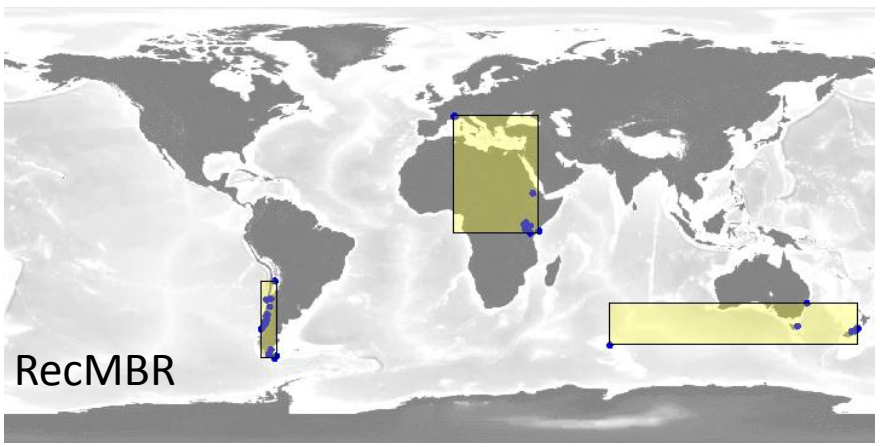
❸ Applicability in other application fields

geometric approaches









● peer data

❸ Applicability in other application fields

Exemplary ranking for MBR summaries, other rankers for geometric resource descriptions work likewise!

Given:

Query q and two peers: p_a with MBR_a and p_b with MBR_b

Peer Ranking:

minimum distance to MBR

reciprocal size of area covered by MBR

$$contains(MBR_a, q) \wedge \neg contains(MBR_b, q) \Rightarrow p_a \succ p_b$$

$$contains(MBR_b, q) \wedge \neg contains(MBR_a, q) \Rightarrow p_b \succ p_a$$

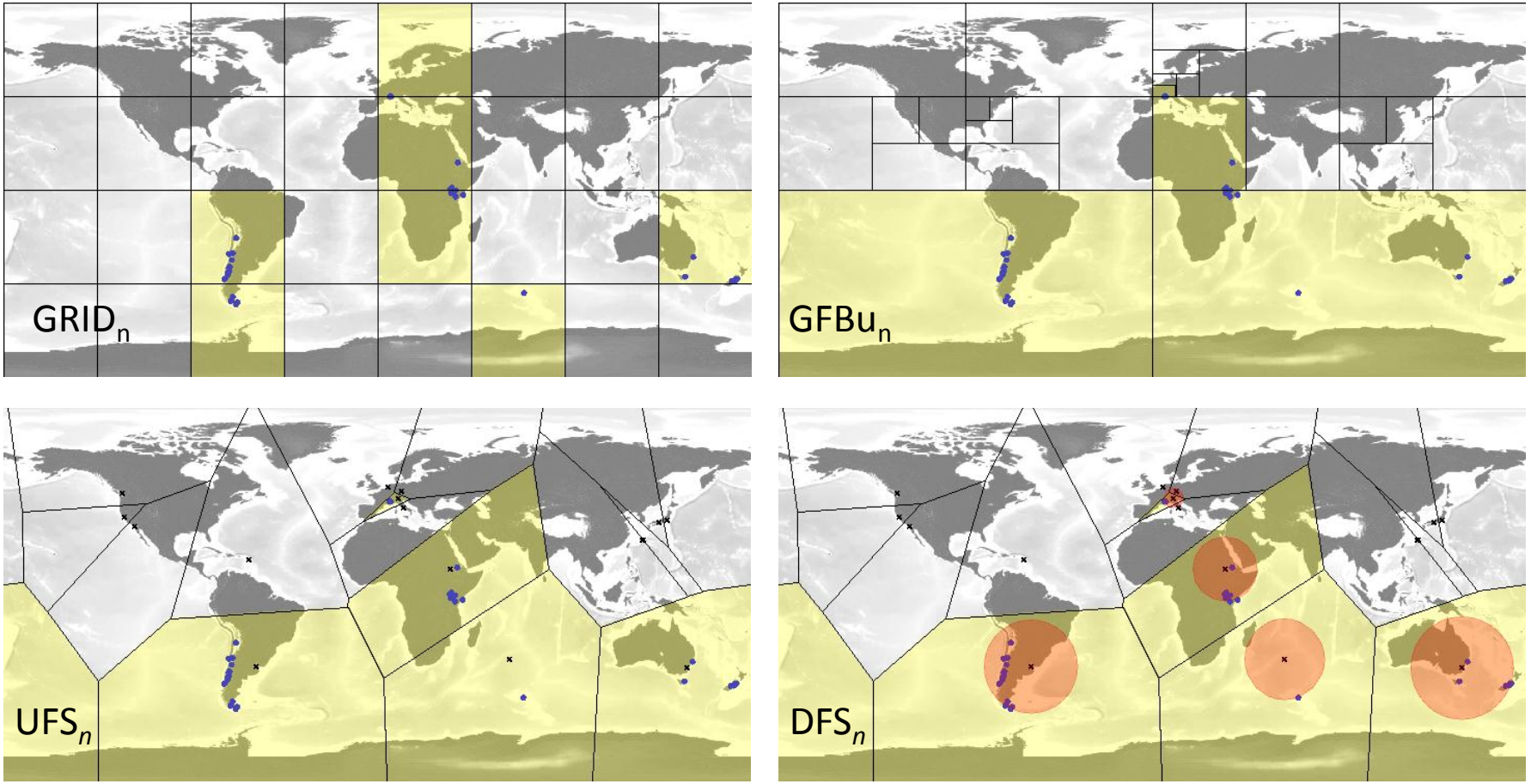
$$contains(MBR_a, q) \wedge contains(MBR_b, q) \Rightarrow recMBRSize(MBR_a, MBR_b)$$

$$\neg contains(MBR_a, q) \wedge \neg contains(MBR_b, q) \Rightarrow distToMBR(q, MBR_a, MBR_b)$$

*k*Means++ and RecMAR: For each peer a queue of shapes is created (ordered by minimum distance to query point), queue based consideration of shapes.

❸ Applicability in other application fields

space partitioning approaches



●peer data ×reference object

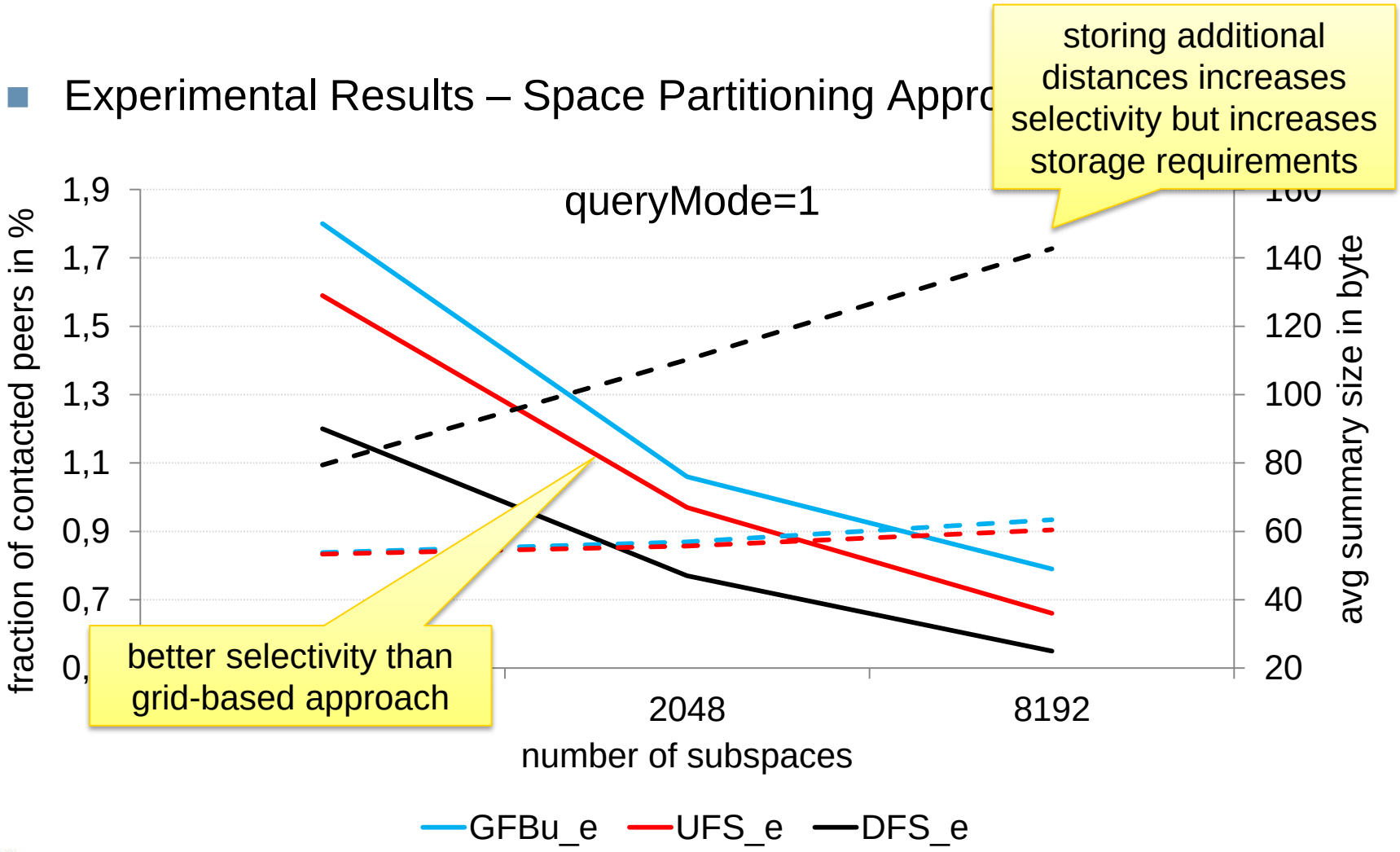
❸ Applicability in other application fields

1. Create list *L* of subspaces/clusters sorted by minimum distance of subspace to query object (cf. HFS/UFS as before)
2. j = 1;
3. Choose subspace $[c_j]$, positioned *j*-th in list *L* ($1 \leq j \leq \gamma$):
   - p_a administers documents in $[c_j]$, p_b does not $\Rightarrow p_a \succ p_b$
   - p_a and p_b both do (not) administer documents in $[c_j]$
     - *j*++; GOTO 3;

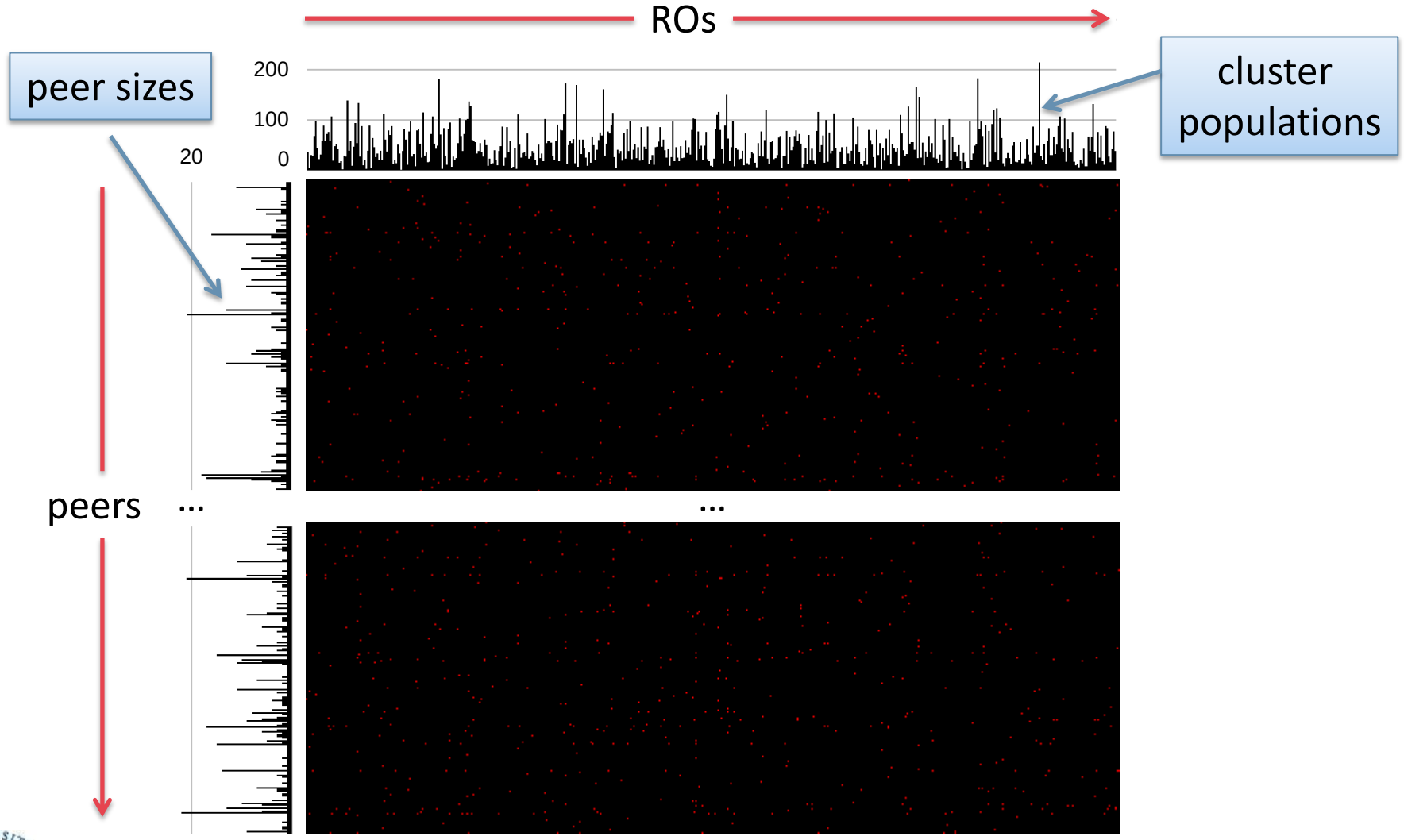
Grid: considers rings of neighboring cells instead of single subspaces
UFS/DFS: list ordered by min. distance of ROs to query object

❸ Applicability in other application fields

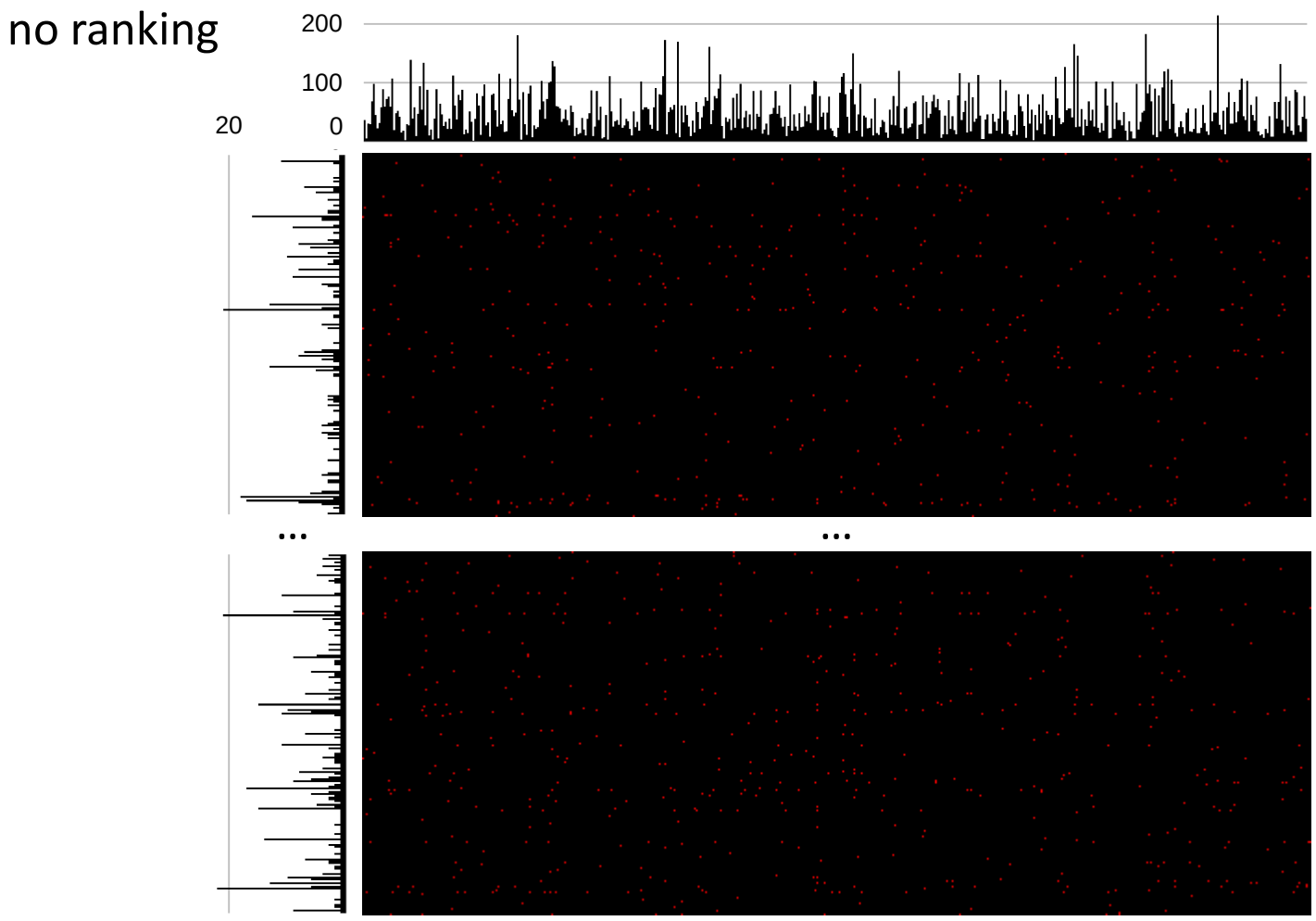
- Experimental Results – Space Partitioning Appro



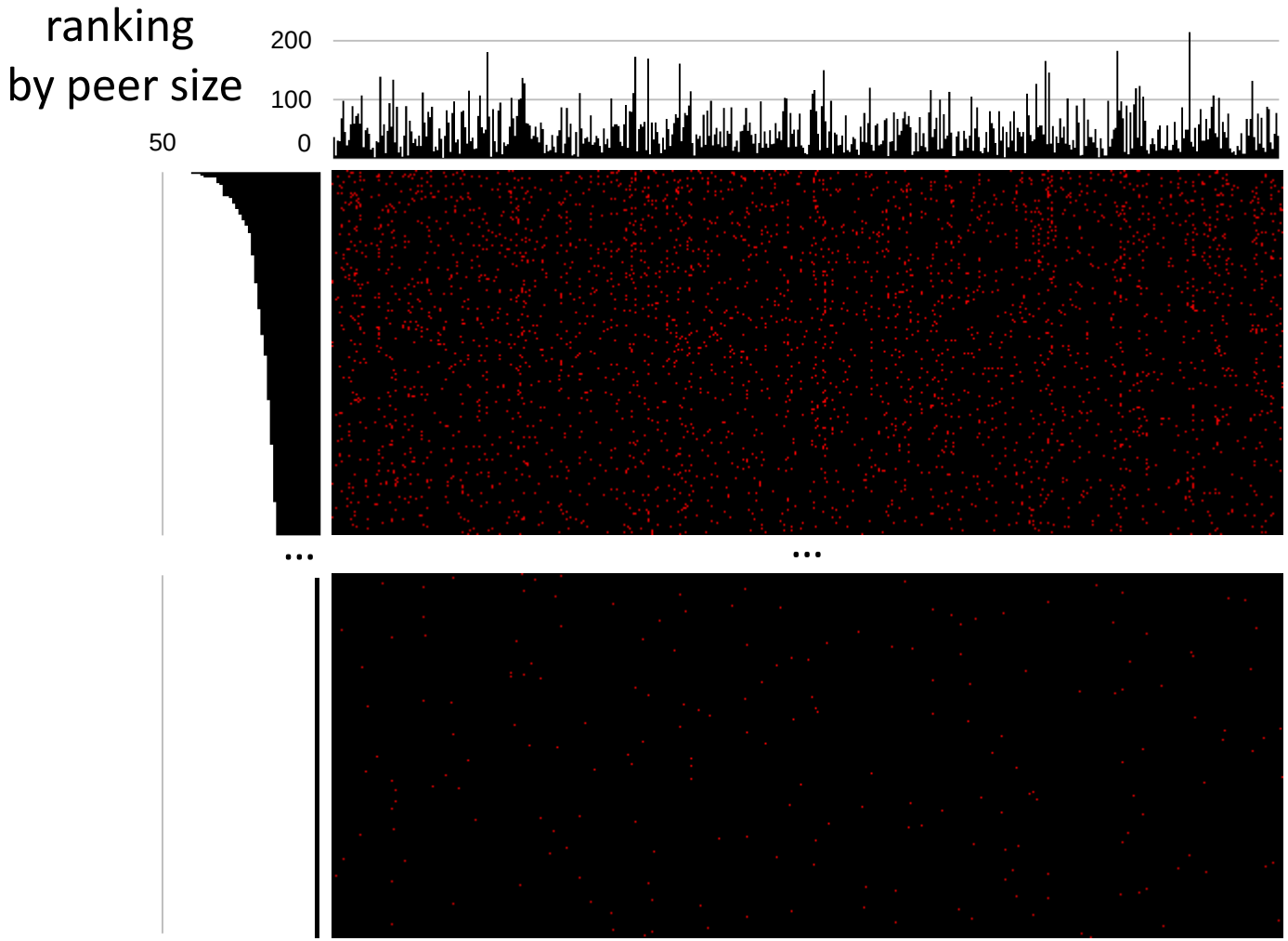
3. Conclusion and Outlook: 'Source selection through visual analytics'



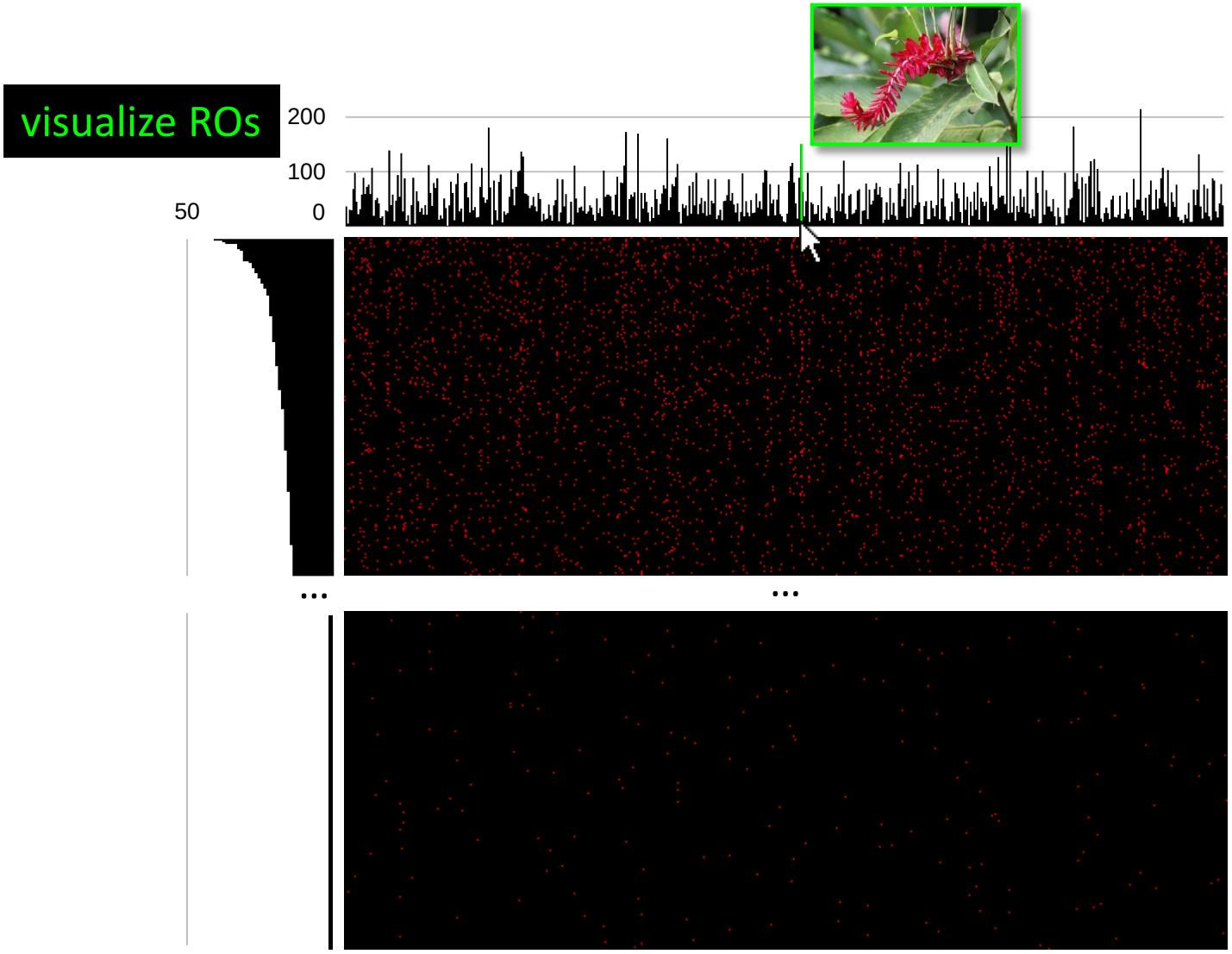
3. Conclusion and Outlook: 'Source selection through visual analytics'



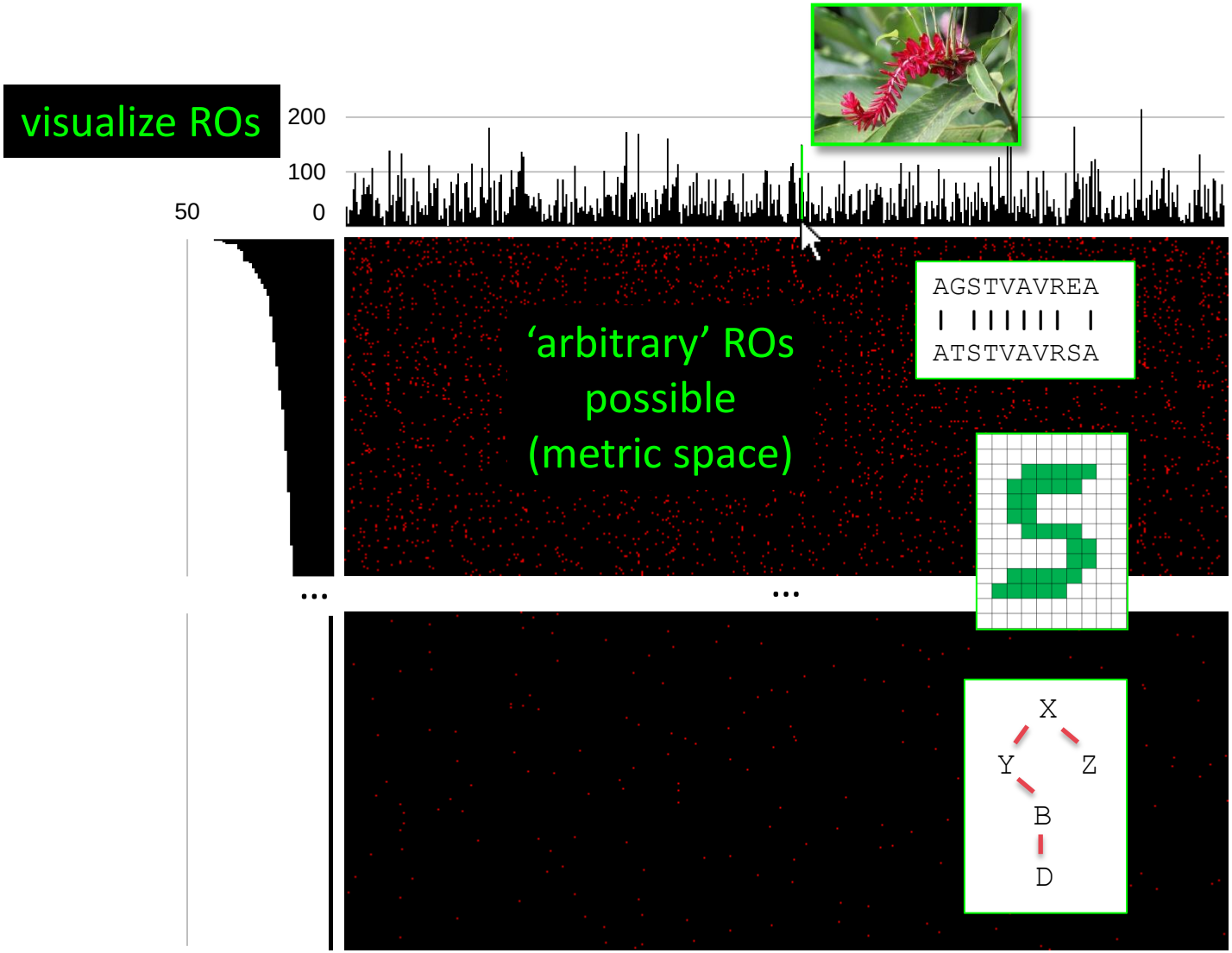
3. Conclusion and Outlook: 'Source selection through visual analytics'



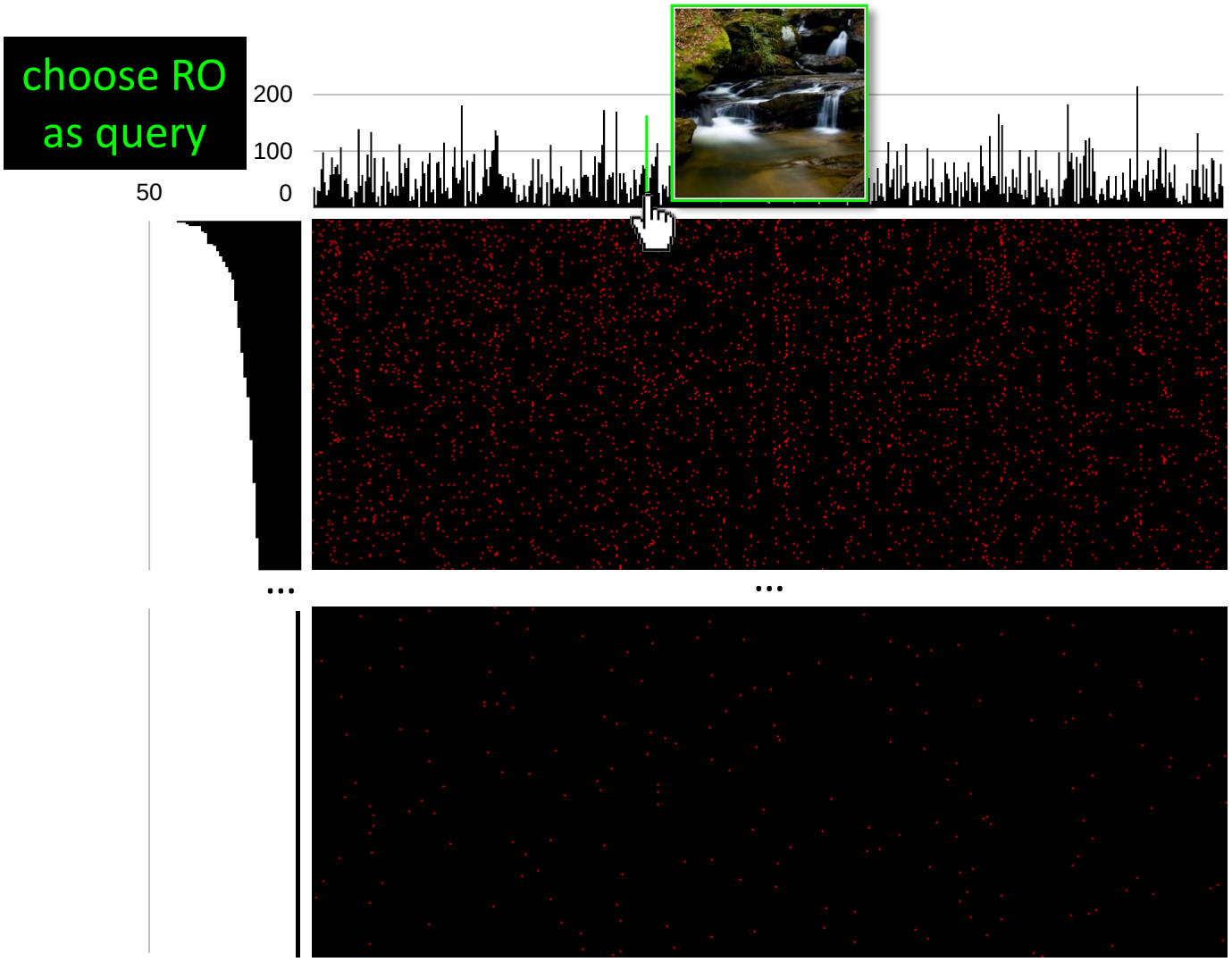
3. Conclusion and Outlook: 'Source selection through visual analytics'



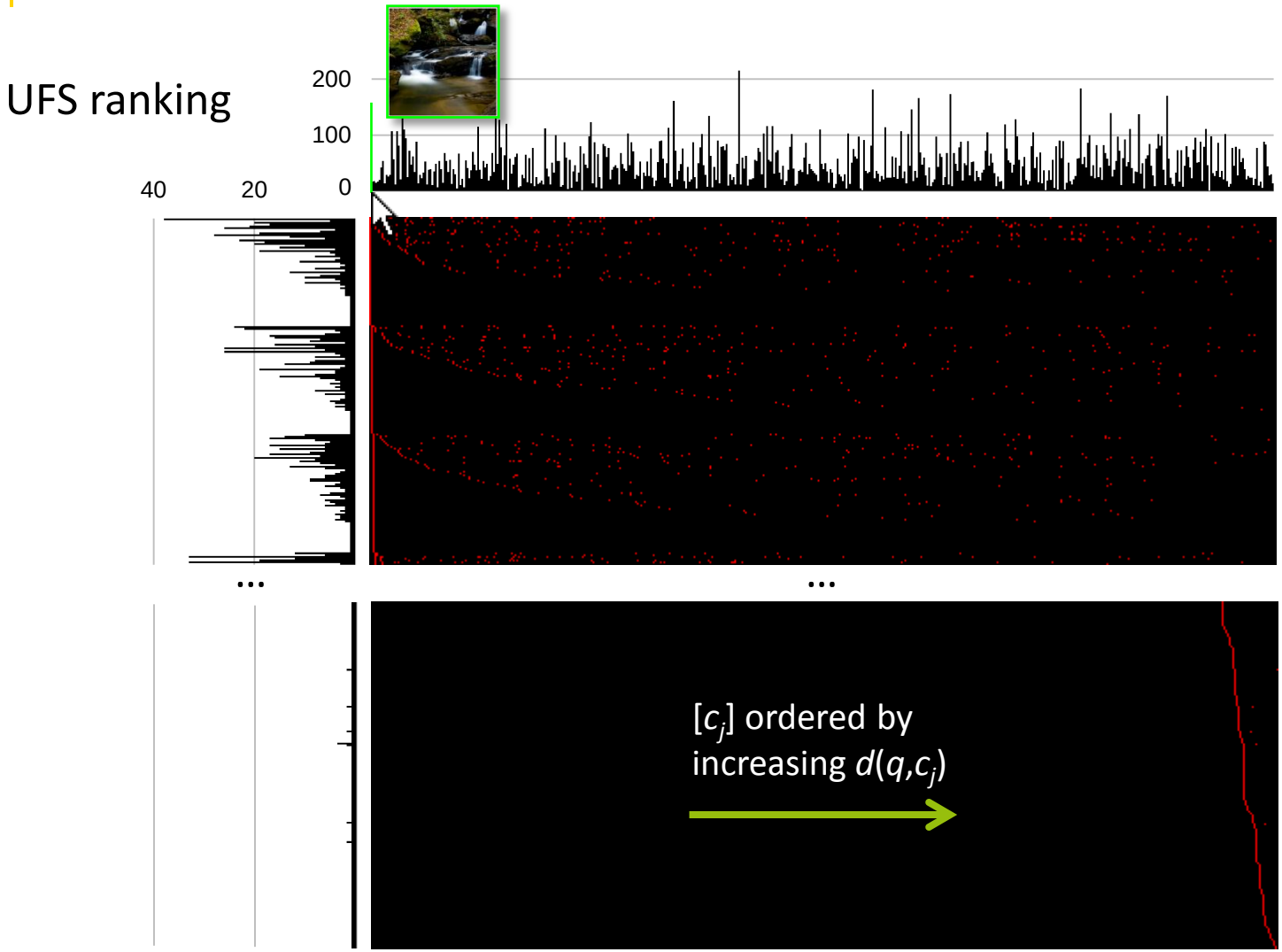
3. Conclusion and Outlook: 'Source selection through visual analytics'



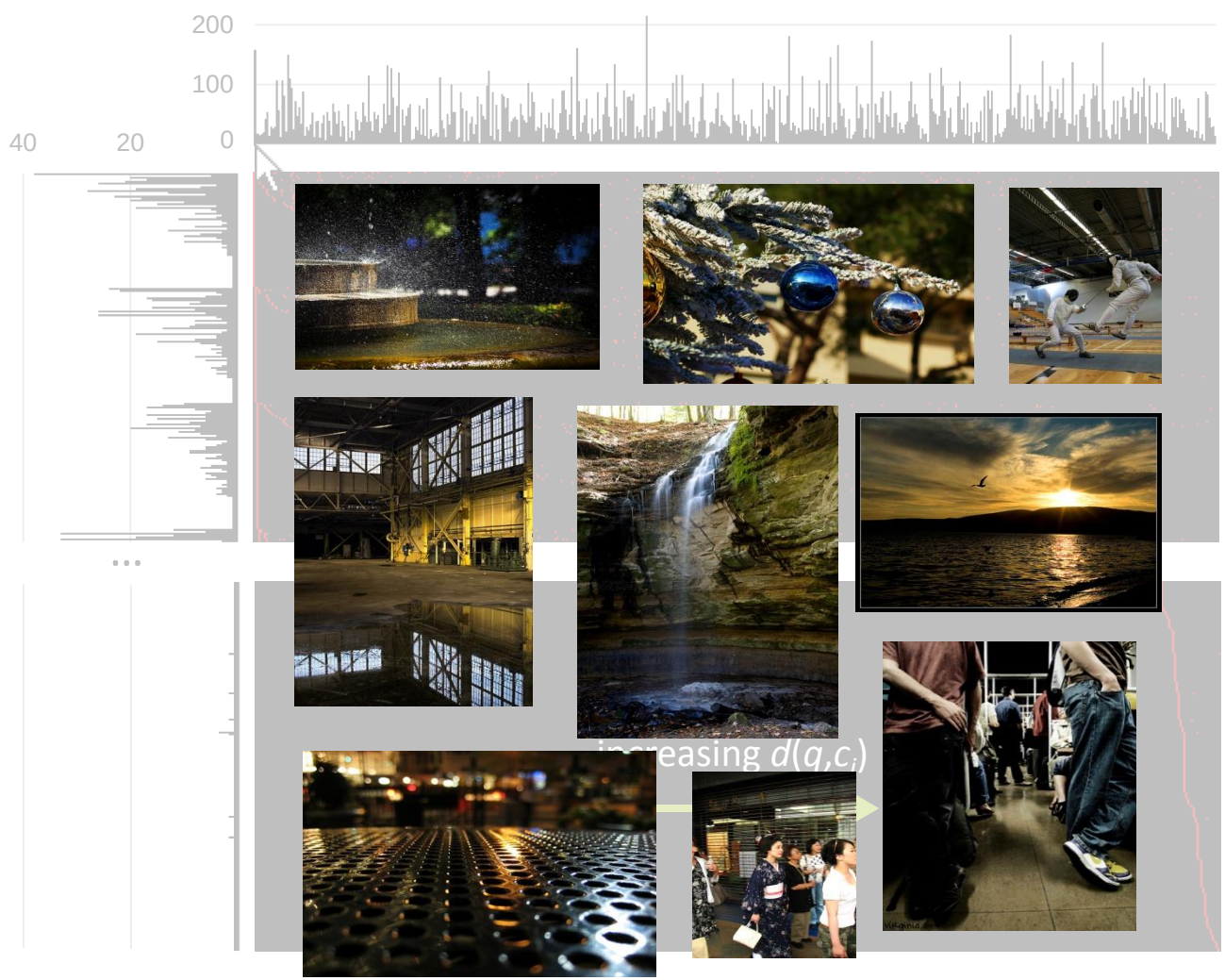
3. Conclusion and Outlook: 'Source selection through visual analytics'



3. Conclusion and Outlook: 'Source selection through visual analytics'



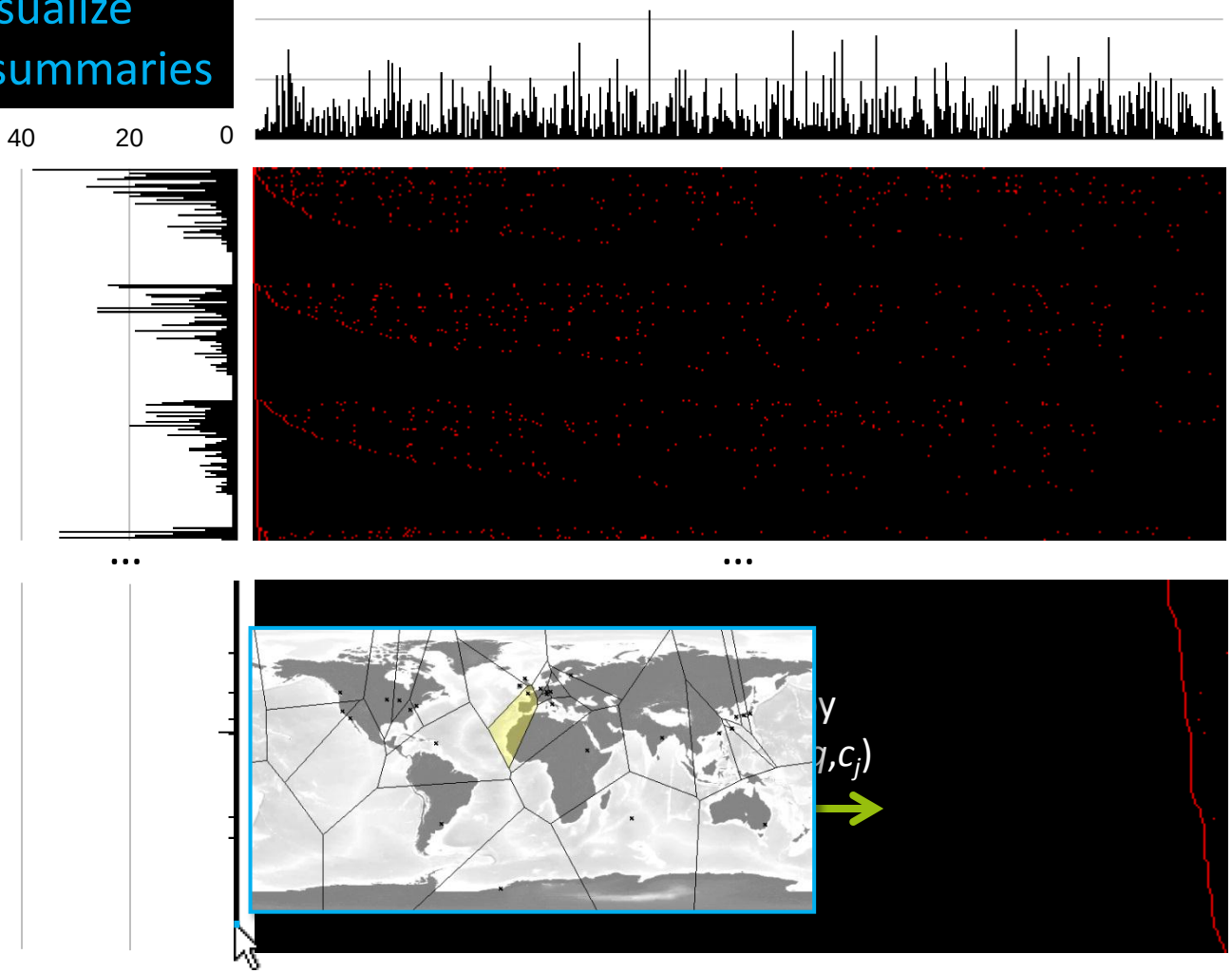
3. Conclusion and Outlook: 'Source selection through visual analytics'



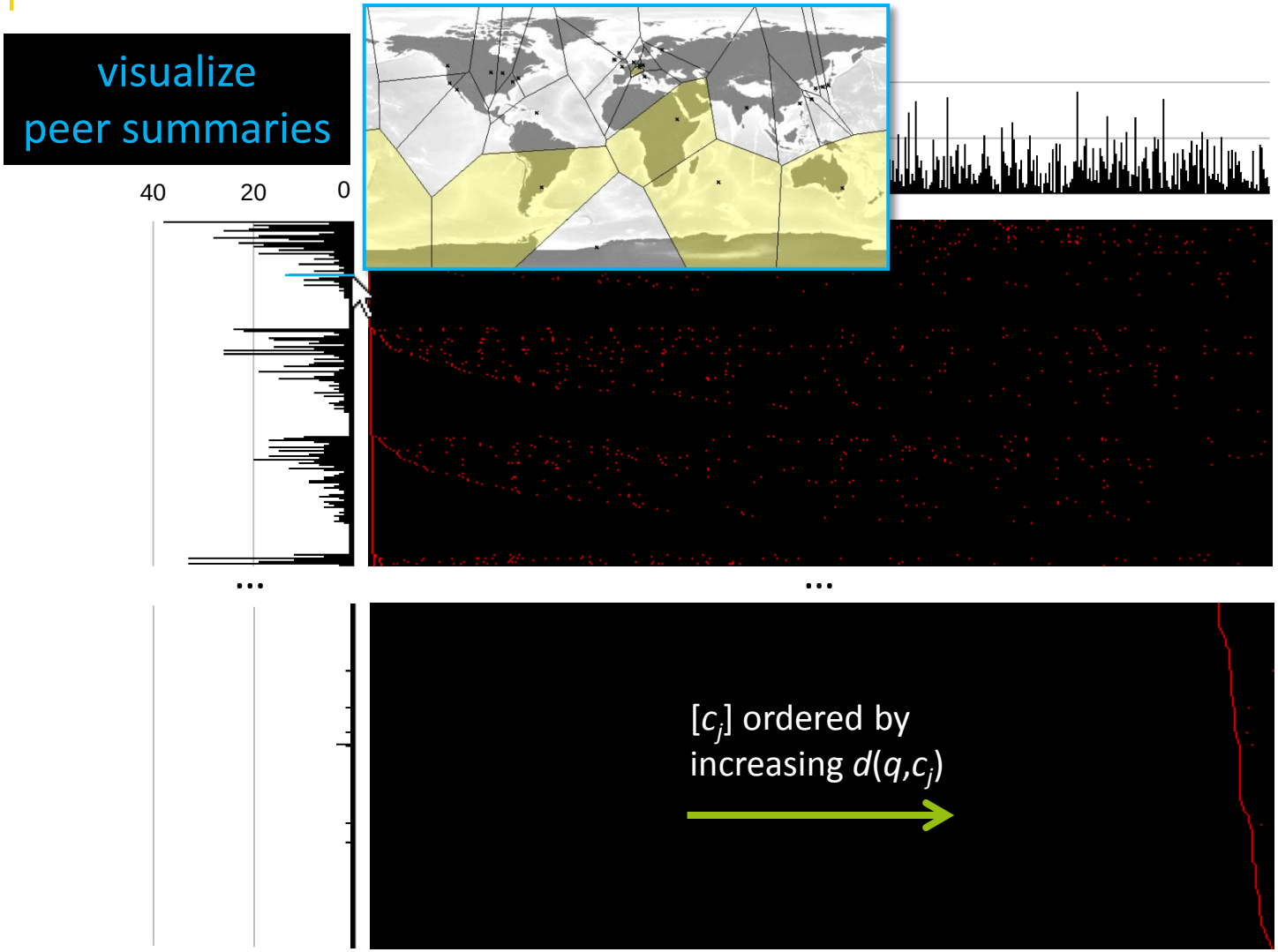
9 NNs

3. Conclusion and Outlook: 'Source selection through visual analytics'

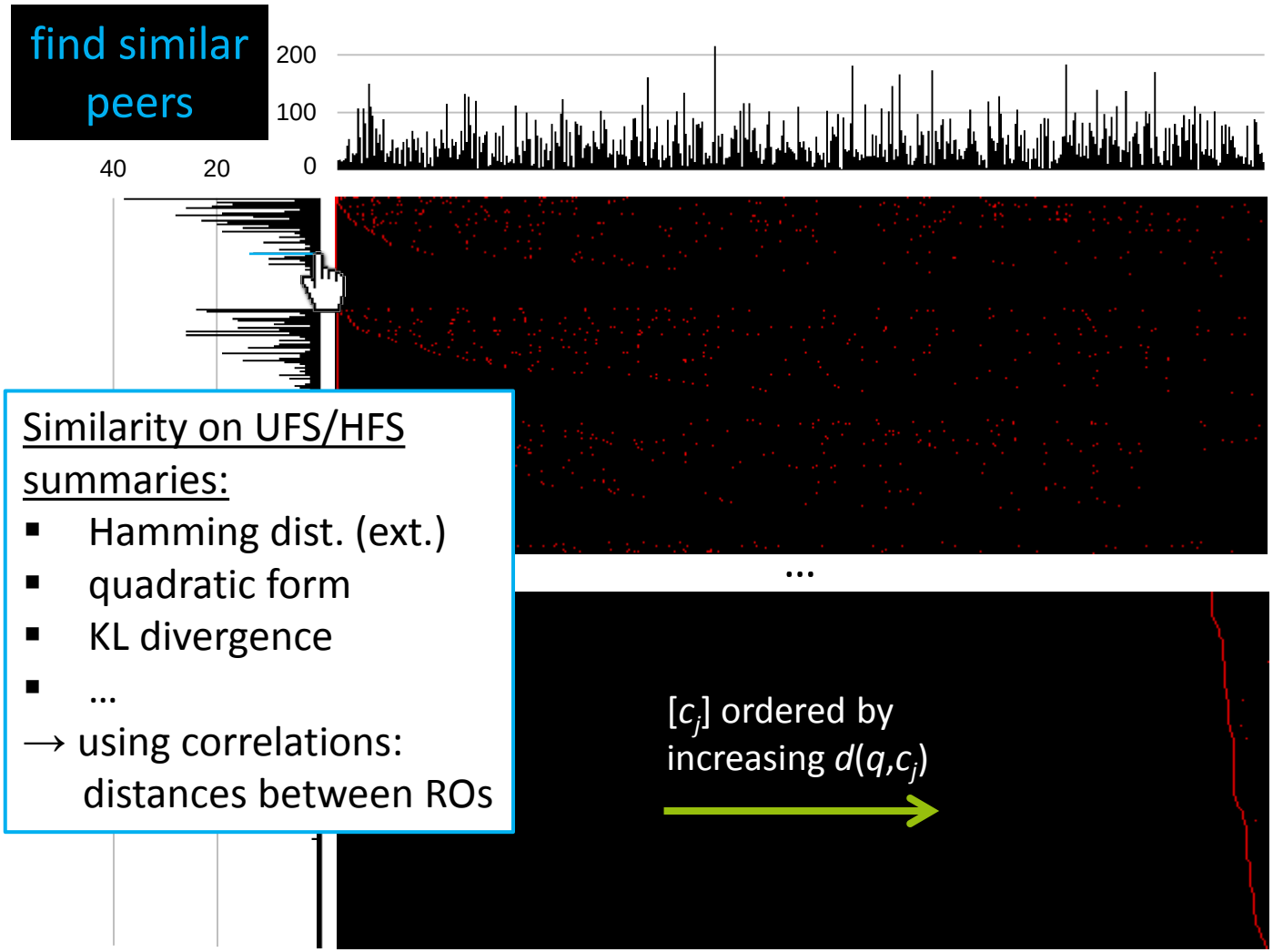
visualize peer summaries



3. Conclusion and Outlook: 'Source selection through visual analytics'



3. Conclusion and Outlook: 'Source selection through visual analytics'

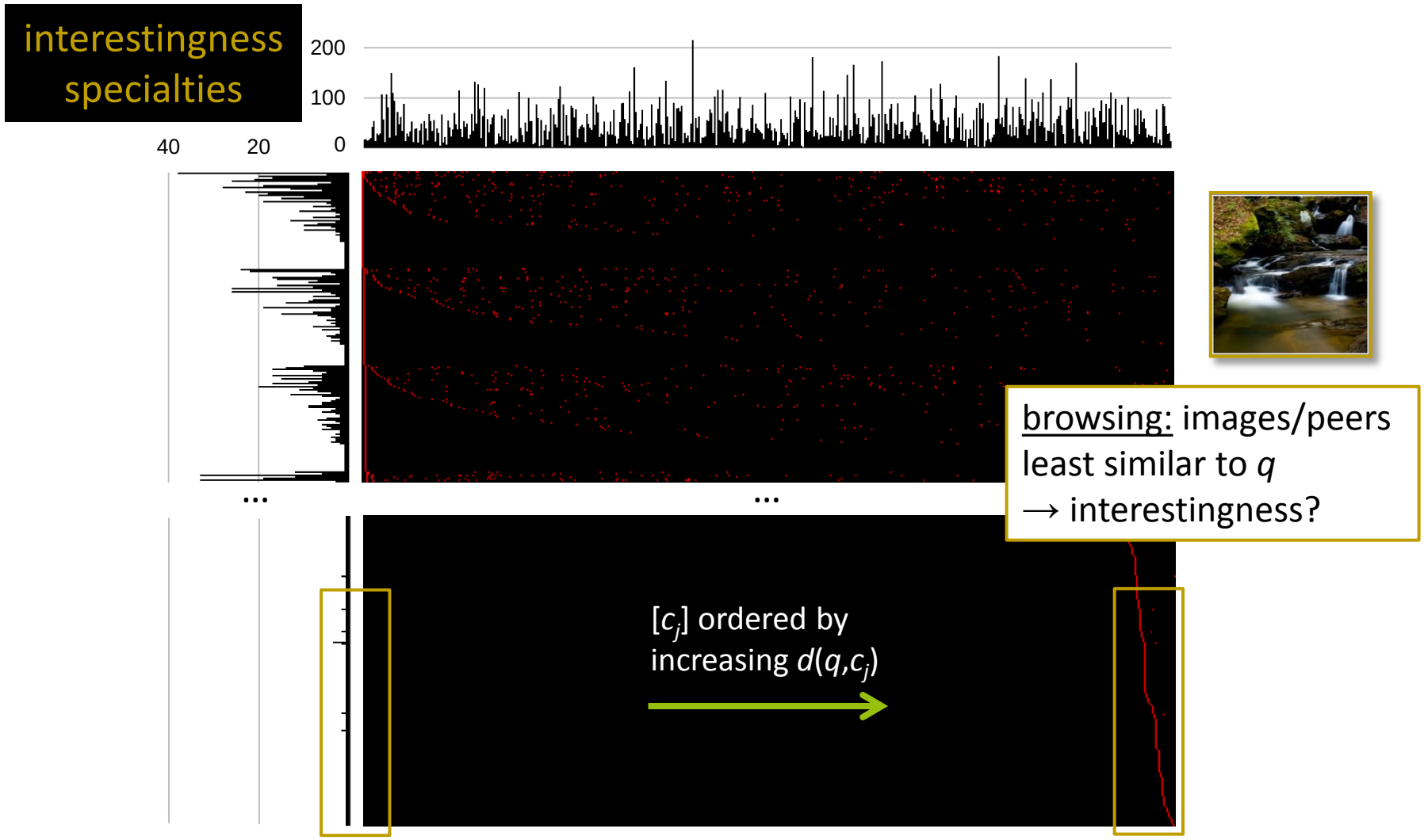


Similarity on UFS/HFS summaries:

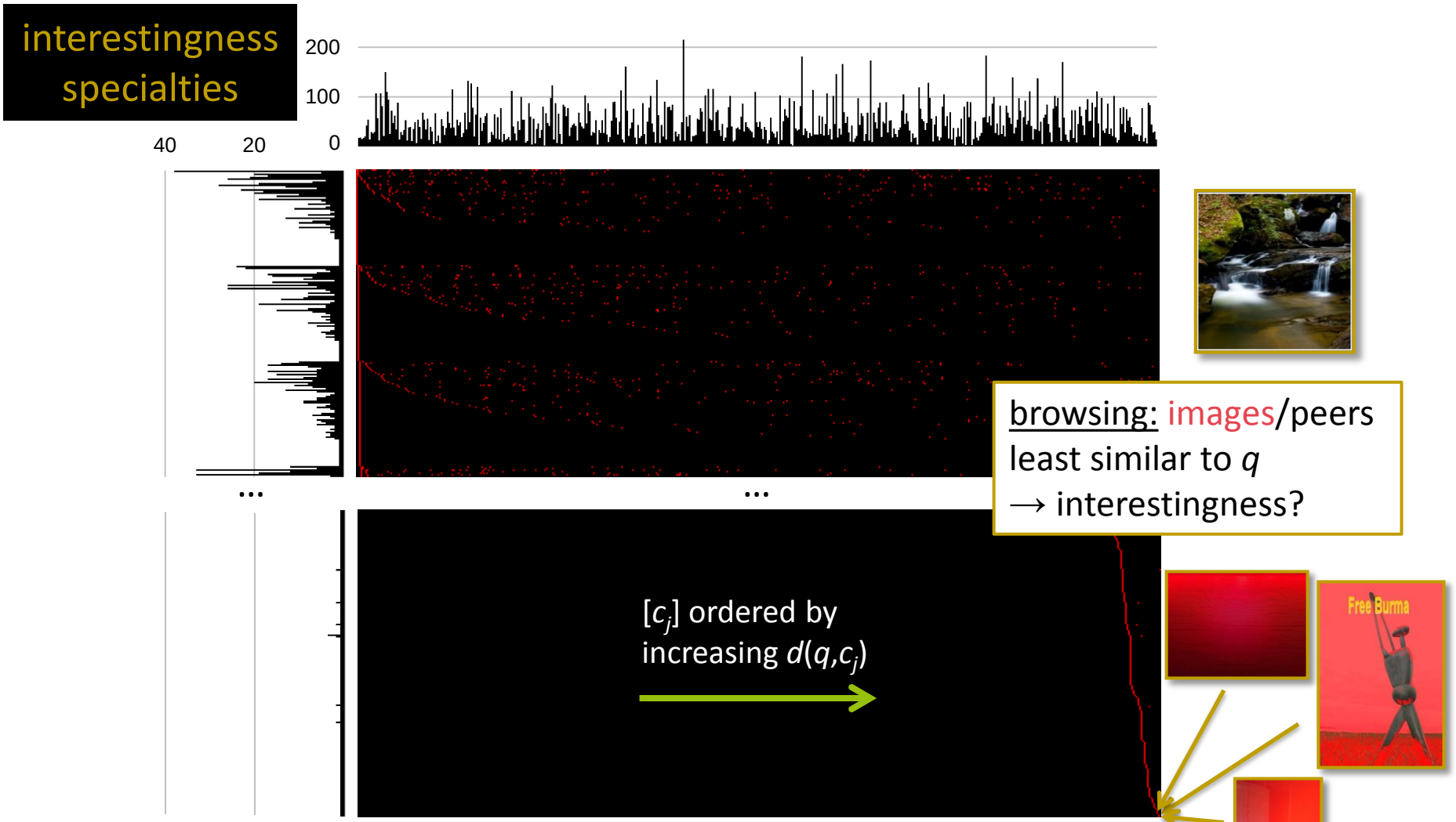
- Hamming dist. (ext.)
- quadratic form
- KL divergence
- …

→ using correlations: distances between ROs

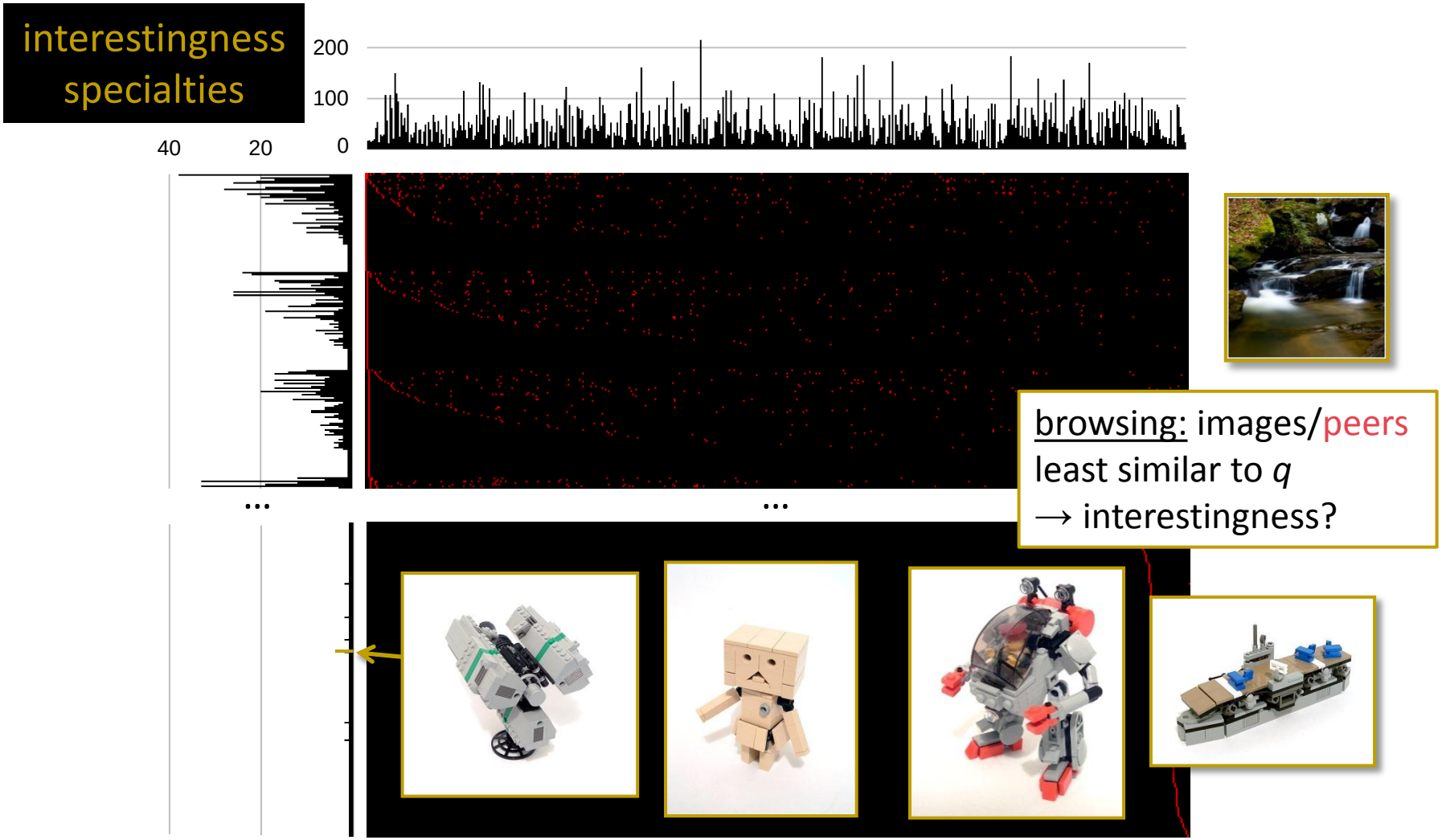
3. Conclusion and Outlook: 'Source selection through visual analytics'



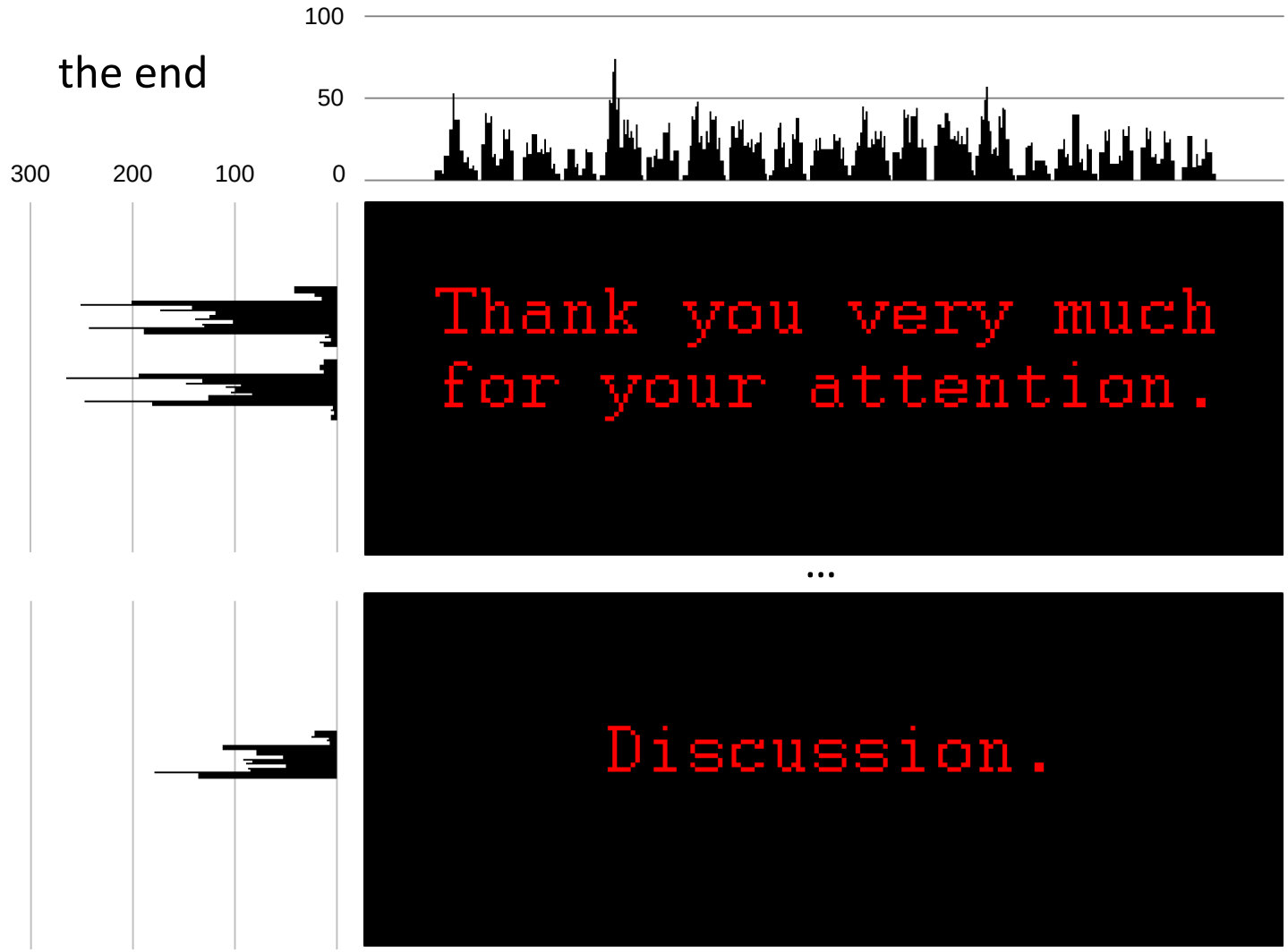
3. Conclusion and Outlook: 'Source selection through visual analytics'



3. Conclusion and Outlook: 'Source selection through visual analytics'



3. Conclusion and Outlook: 'Source selection through visual analytics'



Literaturverzeichnis (1/2)

[BDP04] S. Berretti, A. Del Bimbo, and P. Pala. Merging results for distributed content based image retrieval. Multimedia Tools Appl., 24(3):215–232, 2004.

[BEF+09] P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti. CoPhIR: a Test Collection for Content-Based Image Retrieval. CoRR, abs/0905.4627v2, http://arxiv.org/abs/0905.4627v2 (last visit: 2.11.2011), 2009.

[BEM+07] D. Blank, S. El Allali, W. Müller, and A. Henrich. Sample-based creation of peer summaries for efficient similarity search in scalable peer-to-peer networks. In Intl. SIGMM Workshop on Multimedia Information Retrieval, pages 143–152, Augsburg, Germany, 2007. ACM.

[BH10a] D. Blank and A. Henrich. Binary histograms for resource selection in peer-to-peer media retrieval. In Proc. of LWA Workshop Lernen, Wissen, Adaptivität, pages 183–190, Kassel, Germany, 2010.

[BH10b] D. Blank and A. Henrich. Description and selection of media archives for geographic nearest neighbor queries in P2P networks. In Proc. of the Information Access for Personal Media Archives Workshop, pages 22–29, http://doras.dcu.ie/15373/ (last visit: 2.11.2011), 2010.

[BH12] D. Blank and A. Henrich. Inverted file-based indexing for efficient multimedia information retrieval in metric spaces. In Proc. of the 27th ACM Intl. Symp. on Applied Computing. Riva del Garda, Italy, 2012 (to appear).

Literaturverzeichnis (2/2)

[Cal00] J. Callan. Distributed information retrieval. In W. B. Croft, editor, Advances in Information Retrieval, pages 127–150. Kluwer Academic Publishers, 2000.

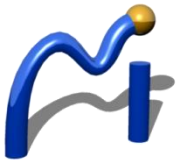
[EMH+06] M. Eisenhardt, W. Müller, A. Henrich, D. Blank, and S. El Allali. Clustering-based source selection for efficient image retrieval in peer-to-peer networks. In Proc. of the 8th Intl. Symp. on Multimedia, pages 823–830, San Diego, CA, USA, 2006. IEEE.

[MEH05a] W. Müller, M. Eisenhardt, and A. Henrich. Scalable summary based retrieval in P2P networks. In Proc. of the 14th Intl. Conf. on Information and Knowledge Management, pages 586–593, Bremen, Germany, 2005. ACM.

[MEH05b] W. Müller, M. Eisenhardt, and A. Henrich. Fast retrieval of high-dimensional feature vectors in P2P networks using compact peer data summaries. Multimedia Systems, 10(6):464–474, 2005.

[MKL+03] D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu. Peer-to-peer computing. Technical Report HPL-2002-57 (R.1), HP Laboratories, Palo Alto, CA, USA, July 2003.

[ZAD+05] P. Zezula, G. Amato, V. Dohnal, and M. Batko. Similarity Search: The Metric Space Approach. Springer New York, Inc., Secaucus, NJ, USA, 2005.



Bildquellen des Vortrags

Seite 3: [BEF+09]

Seite 5: http://commons.wikimedia.org/wiki/File:Bamberg-Rathaus1-Asio.JPG

Seite 6: http://commons.wikimedia.org/wiki/File:Dreieck.svg?uselang=de

Seite 7 und 12: http://commons.wikimedia.org/wiki/File:Bamberg-Rathaus3-Asio.JPG

Seite 10 und 11:

http://commons.wikimedia.org/wiki/File:Bamberg-Rathaus1-Asio.JPG
http://commons.wikimedia.org/wiki/File:Bamberg_Klein-Venedig_I.jpg
http://commons.wikimedia.org/wiki/File:Marienberg_wuerzburg.jpg
http://commons.wikimedia.org/wiki/File:Marienkapelle.JPG
http://commons.wikimedia.org/wiki/File:Zugspitze_Sonnenuntergang_Westseite.JPG
http://commons.wikimedia.org/wiki/File:Sonnenuntergang_001.jpg
Das pinke Sonnenuntergangsbild ist derzeit nicht mehr bei Wikimedia verfügbar.

Seite 17: http://commons.wikimedia.org/wiki/File:Voronoi_diagram.svg