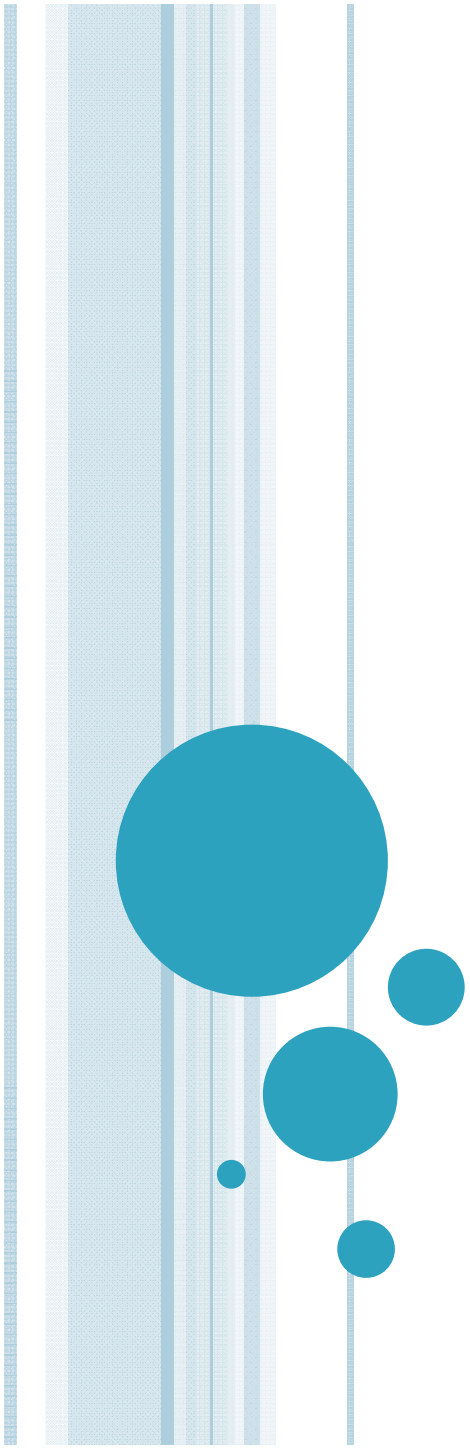


Sistemi per la Ricerca di Dati Multimediali

Claudio Lucchese e Salvatore Orlando

Terzo Workshop di Dipartimento
Dipartimento di Informatica
Università degli studi di Venezia

Ringrazio le seguenti persone, coinvolte nel progetto SAPIR

Raffaele Perego
Fausto Rabitti

CONSIGLIO NAZIONALE DELLE RICERCHE

Paolo Bolettieri
Fabrizio Falchi
Tommaso Piccioli
Matteo Mordacchini

Non solo la Terra è precaria.
RicAt
Ricercatori Atipici
CNR Pisa

WHAT AND, MORE IMPORTANTLY, *WHY* ??!

- Multi-Media Objects:
  - text, web pages *(Google, Yahoo!, …)*
  - images *(Flickr bought by Yahoo!, …)*
  - video *(YouTube bought by Google, …)*
  - audio
  - etc. ...
- Why we are interested in MM objects ?
  - 1 image every 10 documents.
  - Increasing multimedia content on the web.
  - Yet, search in audio-visual content is limited to associated text and metadata annotations.

OK, BUT GOOGLE ALREADY DOES THE JOB …

- We mean to ***search by content !***

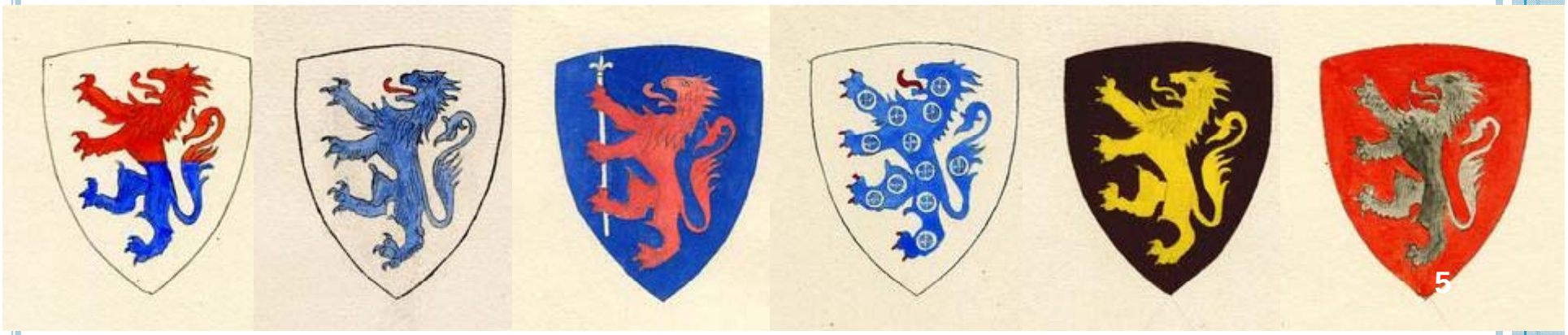
- Google does text-based (annotations) search
  - Do we have enough text ?
  - How is the quality of such text ?
  - Is it correlated with the actual content ?

- Look at this:
  - http://www.nmis.isti.cnr.it/khi/
  - ISTI-CNR (Pisa) and Max Planck Kunsthistorische Institute project on Florentine Coat of Arms

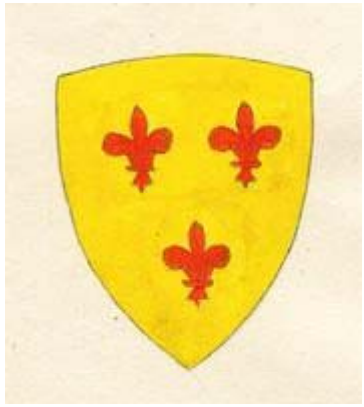
CONTENT-BASED SEARCH

Shape Selection

Results

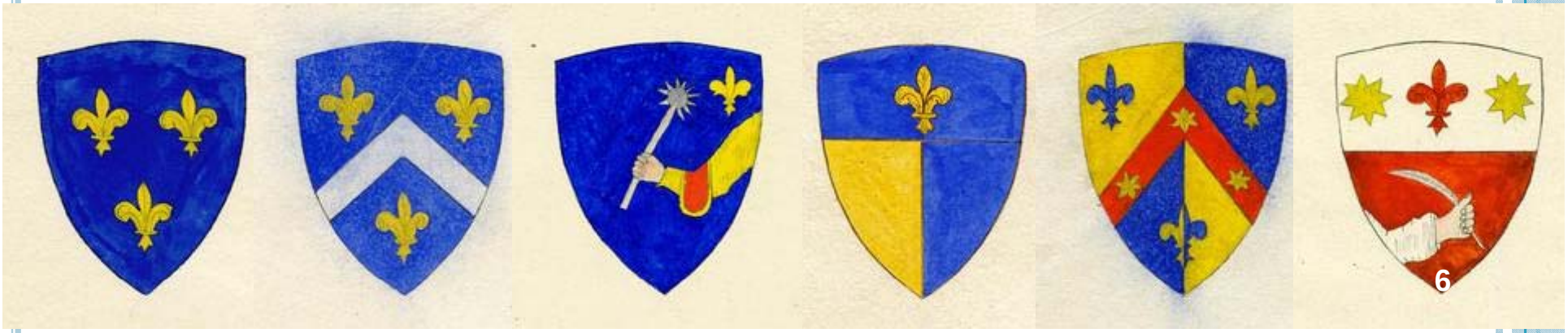


CONTENT-BASED SEARCH



Shape Selection

Results



WHAT IS THE ADDED VALUE ?

- Content-based search for ***disambiguation***:
  - What about searching for “sapphire” on Flickr ?

CONTENT-BASED, I.E. SIMILARITY SEARCH

- Objects are ***"unknown"***

  distances between objects is ***"known"***
- ***Metric Space assumption:***
  - symmetry
  - identity
  - triangle inequality
- Distance functions inducing a metric space:
  - Minkowski distances, edit distance, jaccard distance...
- Typical queries: ***Range*** or ***kNN***
- Applications:

Photo Search

SIMILARITY SEARCH

- Obje
- Metr
  - sy
  - id
  - tri
- Dista
  - M
- Typi
- Applic

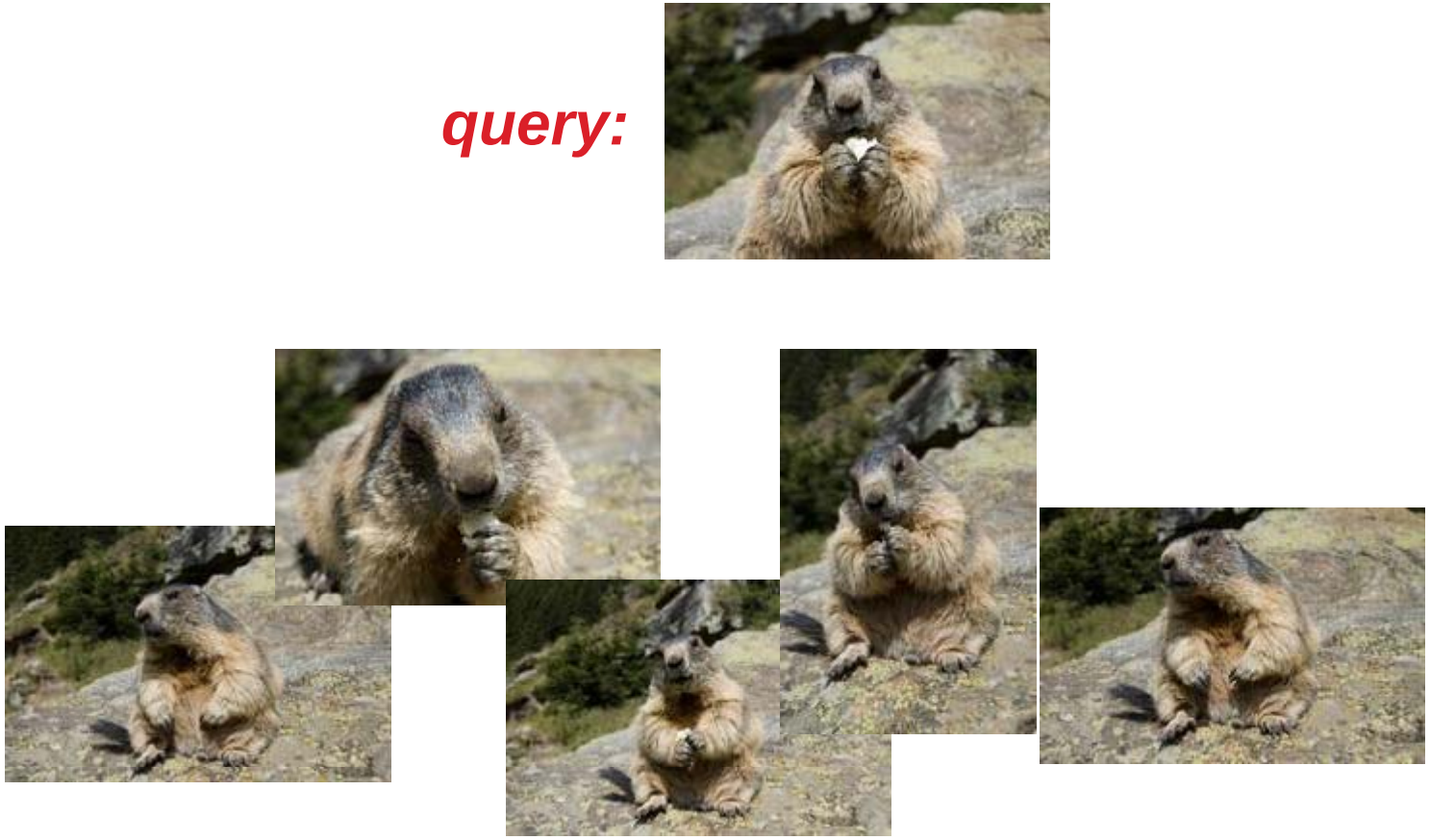


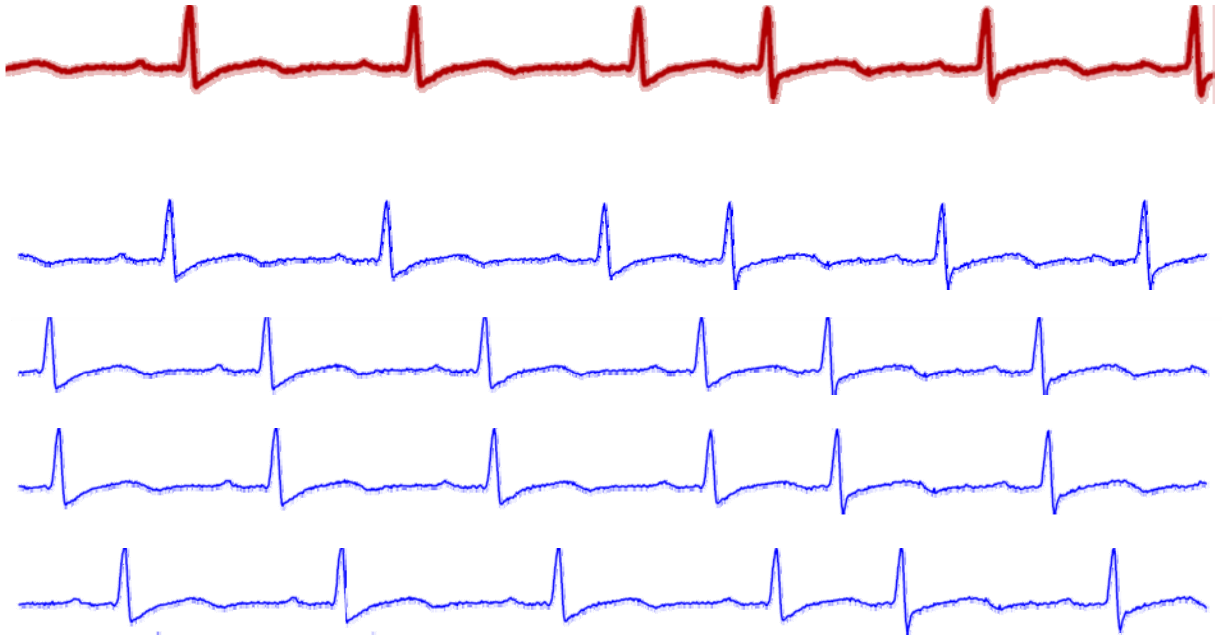
Photo Search

Medical Data Search

SIMILARITY SEAR

- Obje
- Me
  -
  -
  -
- Dis
  -
- Typ
- Ap

query:



SIMILARITY SEAR

Photo Search

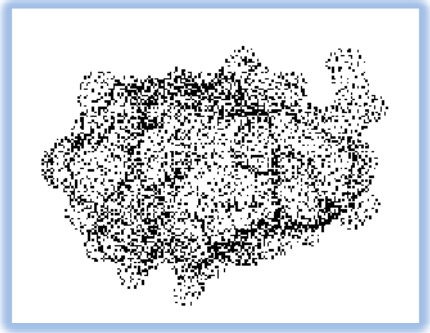
Medical Data Search

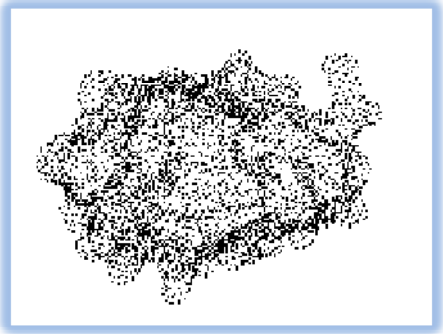
3D Shape Search

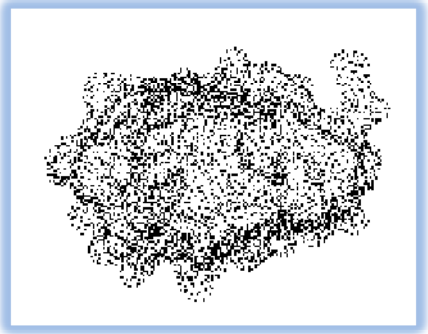
- Objects are *"unknow*

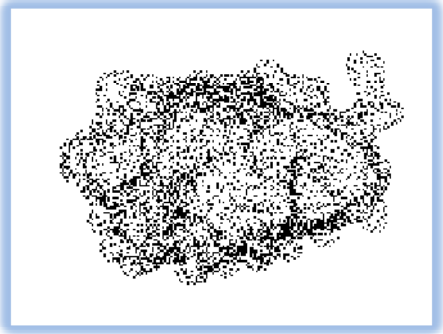
distances be

query:
(haemoglobin molecule)









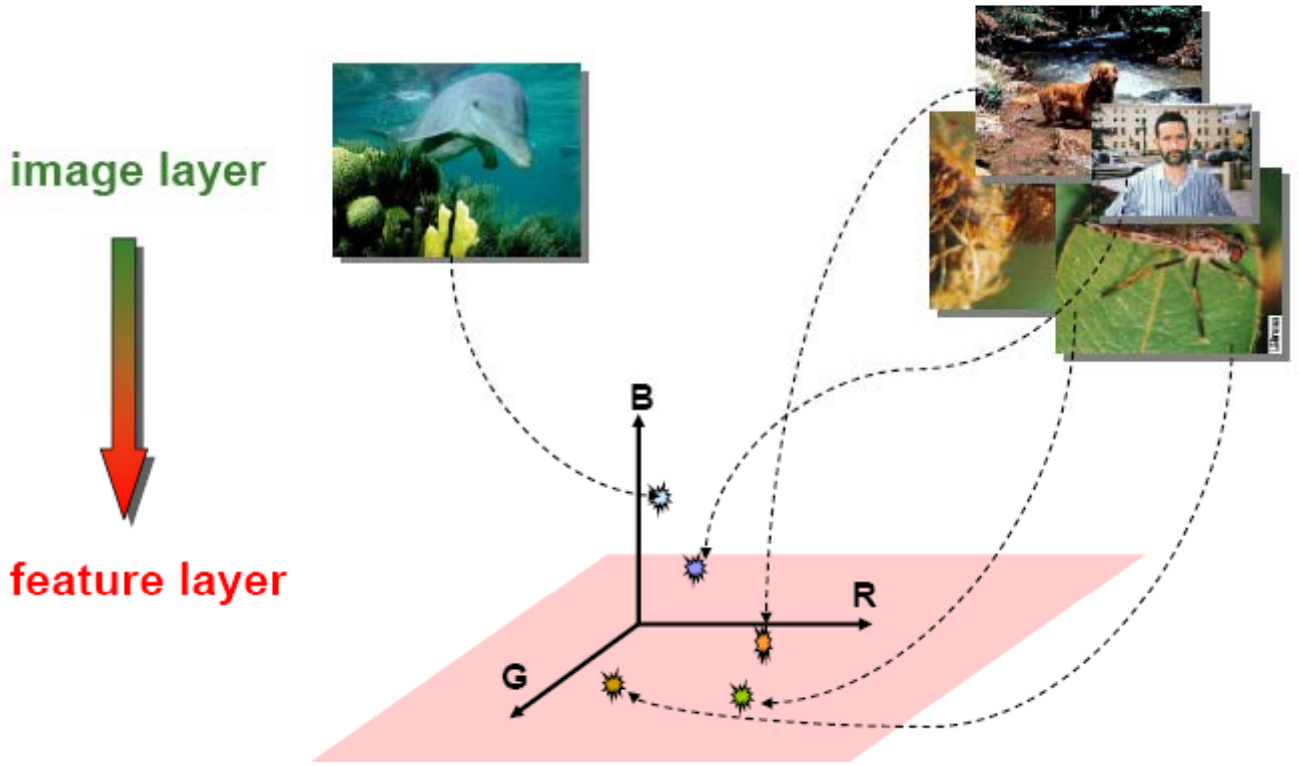
SIMILARITY SEARCH

- Objects are ***"unknown"***

  distances between objects is ***"known"***
- ***Metric Space*** assumption:
  - symmetry
  - identity
  - triangle inequality
- Distance functions inducing a metric space:
  - Minkowski distances, edit distance, jaccard distance...
- Typical queries: ***Range*** or ***kNN***
- Applications:
  - photos, 3D shapes, medical images ***but also***
  - ***text, dna, graphs, etc. etc.***

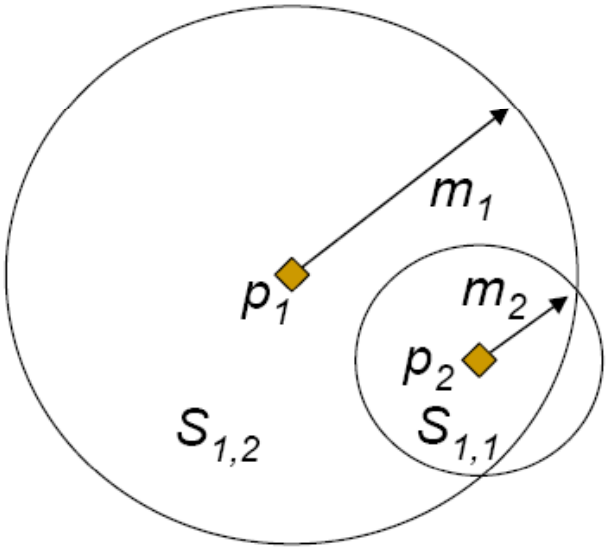
FEATURE-BASED APPROACH

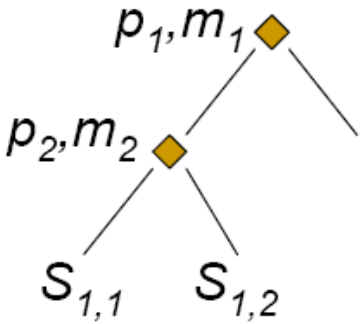
- From the ***object space*** to the ***feature space***



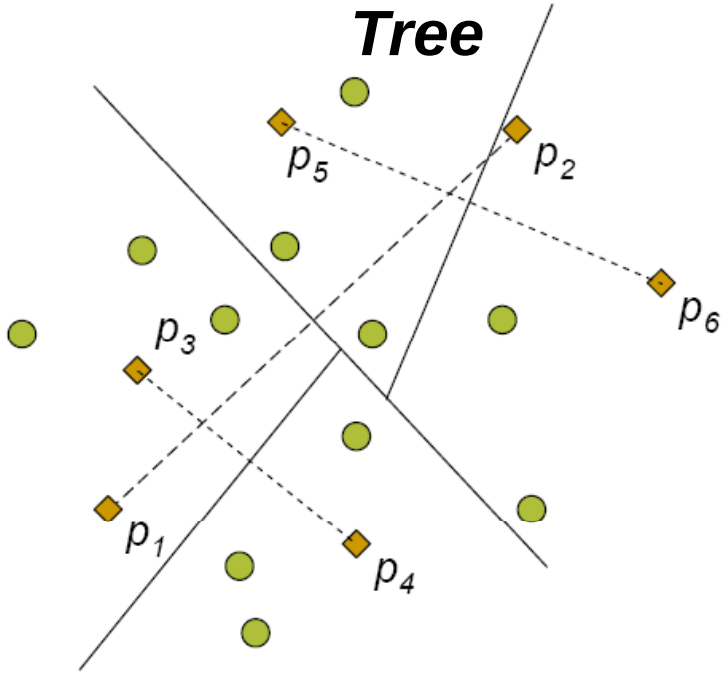
DATA STRUCTURES

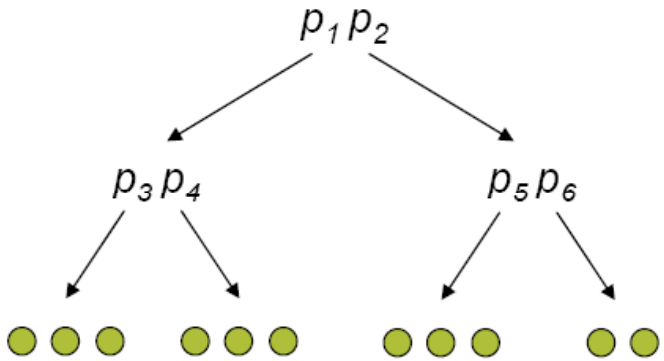
Vantage Point Tree



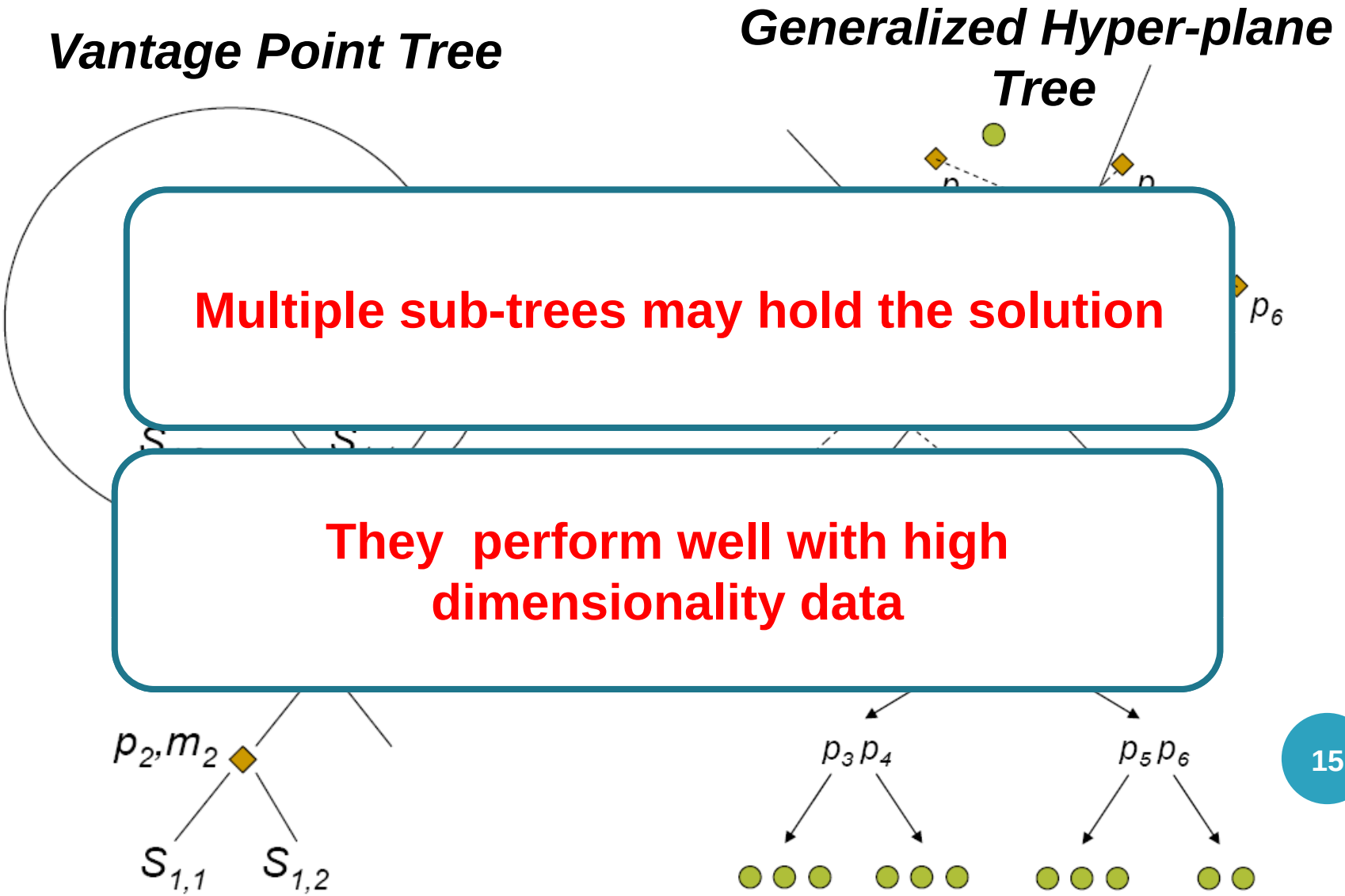


Generalized Hyper-plane Tree





DATA STRUCTURES

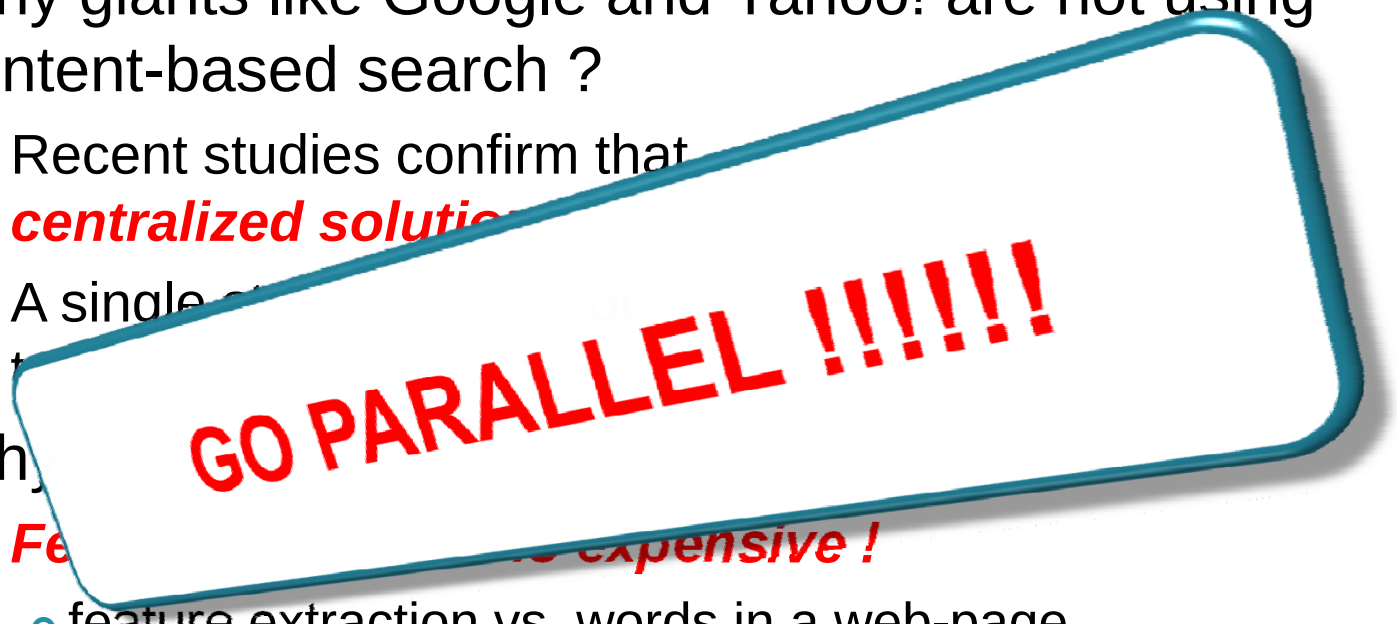


I HAVE NEVER SEEN ANYTHING LIKE THAT IN THE WEB !!!

- Why giants like Google and Yahoo! are not using content-based search ?
  - Recent studies confirm that ***centralized solutions are not scalable !***
  - A single standard PC would need about 12 years to process a collection of 100 million images.
- Why ?
  - ***Feature extraction is expensive !***
    - feature extraction vs. words in a web-page
  - ***Searching is expensive !***
    - similarity search vs. boolean search

I HAVE NEVER SEEN ANYTHING LIKE THAT IN THE WEB !!!

- Why giants like Google and Yahoo! are not using content-based search ?
  - Recent studies confirm that ***centralized solutio***
  - A single
- Wh
  - ***Fe*** ***expensive !***
    - feature extraction vs. words in a web-page
  - ***Searching is expensive !***
    - similarity search vs. boolean search

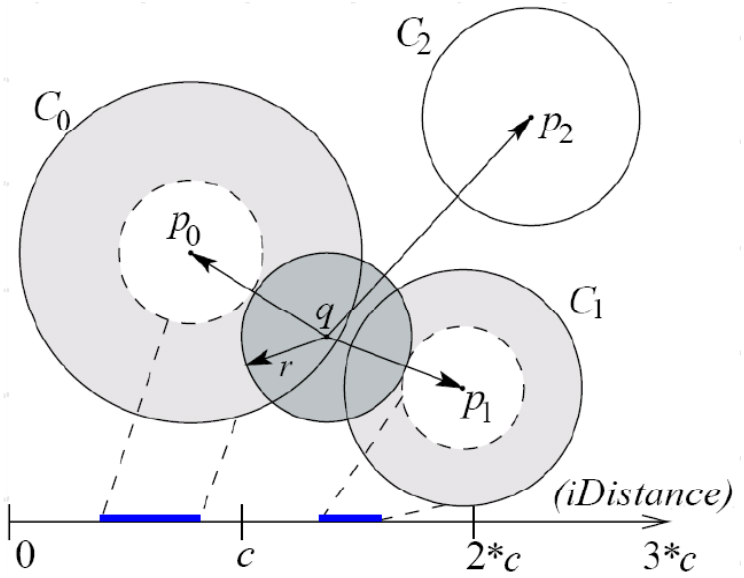


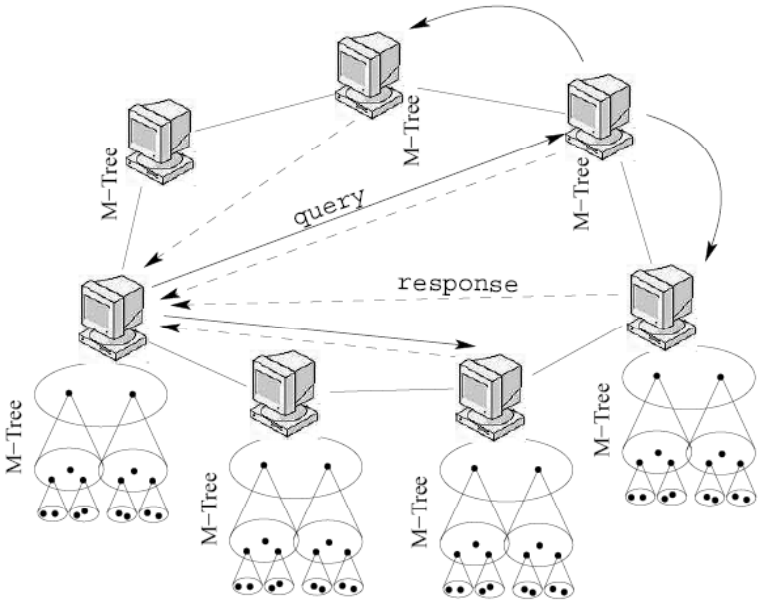
PARALLEL FEATURE EXTRACTION

- ***We crawled Flickr*** to gather Image Ids
- A set of clients were deployed on the EDEE Grid:
  - Retrieve 1000 imaged Ids
  - Download Image
  - Extract MPEG-7 features
  - Parse Flicker Photo-page to obtain additional metadata
  - Send metadata to our centralized repository
- We have about ***50 millions images with features***
  - it seems that Flickr as at least 1 billion images
  - Yahoo! images has at least 2 billion images
  - and … they grow exponentially !
- Our metadata collection is called ***CoPhIR***
  - It is largely the largest publicly available collection

PARALLEL SEARCH

- The search space is mapped into a linear interval
- This is mapped onto a distributed network thanks to some ***DHT-based*** algorithm
- Each node of the network holds an M-Tree





Our Proposal

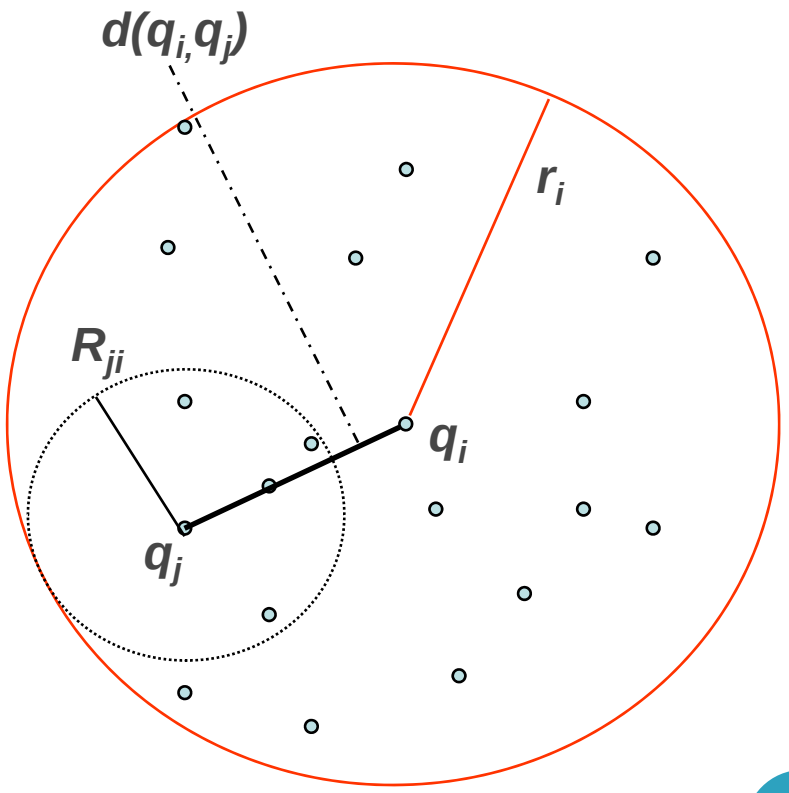
- ***Metric cache***:
  - Cache is widely used in traditional search engines
- ***Trivial***:
  - Store results previous queries
  - Return results when submitted query is stored
- ***Less trivial ...***
  - Store results of previous queries
  - ***Best-effort*** query answering
  - (***with guarantees***)
  - Use past queries to ***optimize database queries***.

A CACHE WITH ONE ENTRY ...

- q_i is the query in cache with its ***k*-NN**.
- r_i is the "radius of the query".
- q_j is a new query.
- If $d(q_i,q_j)<r_i$ then the cached objects at distance

$$R_{ji} = r_i - d(q_i,q_j) \text{ from } q_j$$

are *the top-k' results of the new query*.

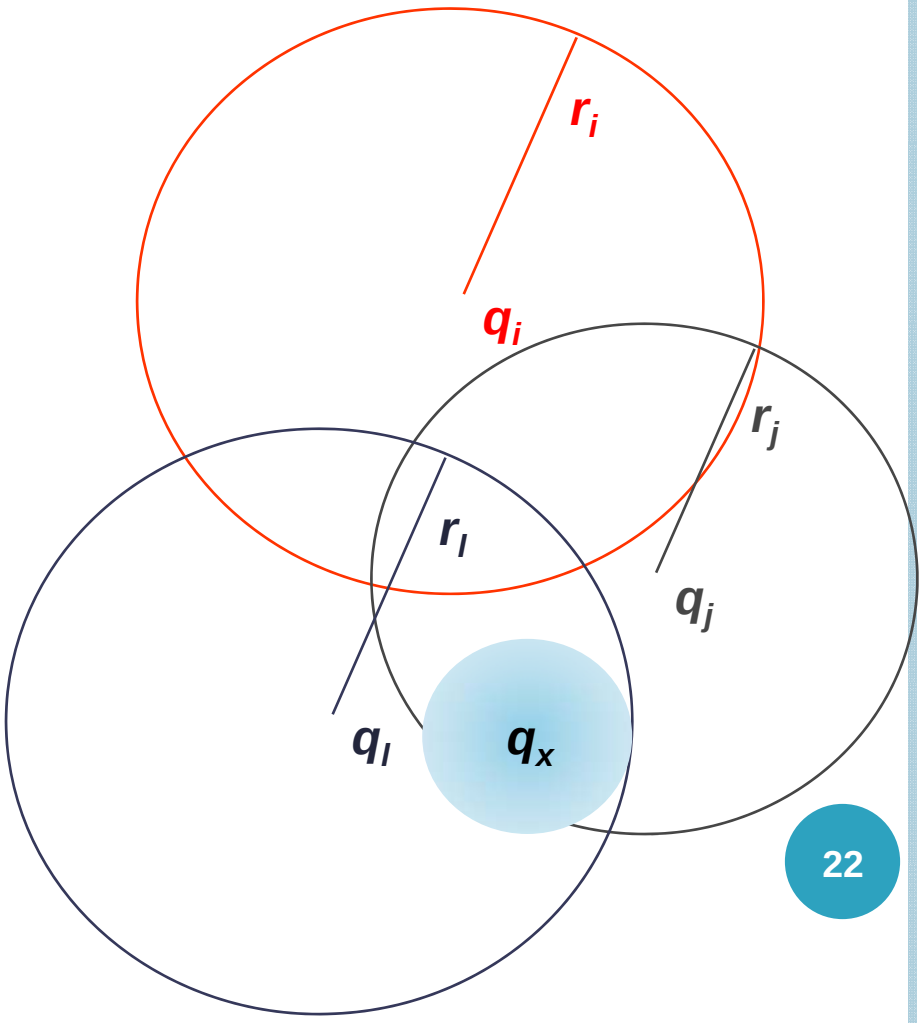


A Cache With Many Entries

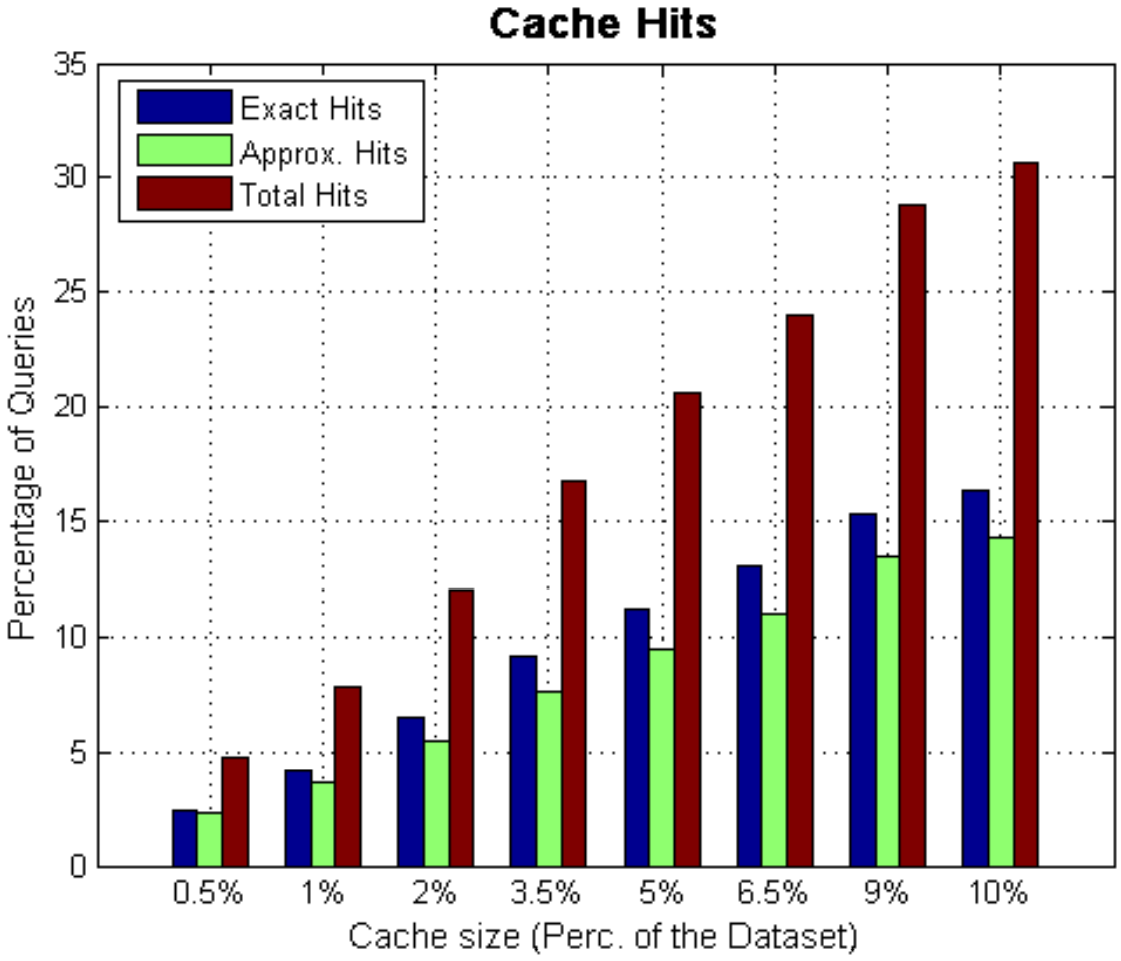
- Given the new query q_x find the *largest* R_{xi} corresponding to the cached query q_i.

- The cached objects within distance R_{xi} from q_x are the most similar to the query.

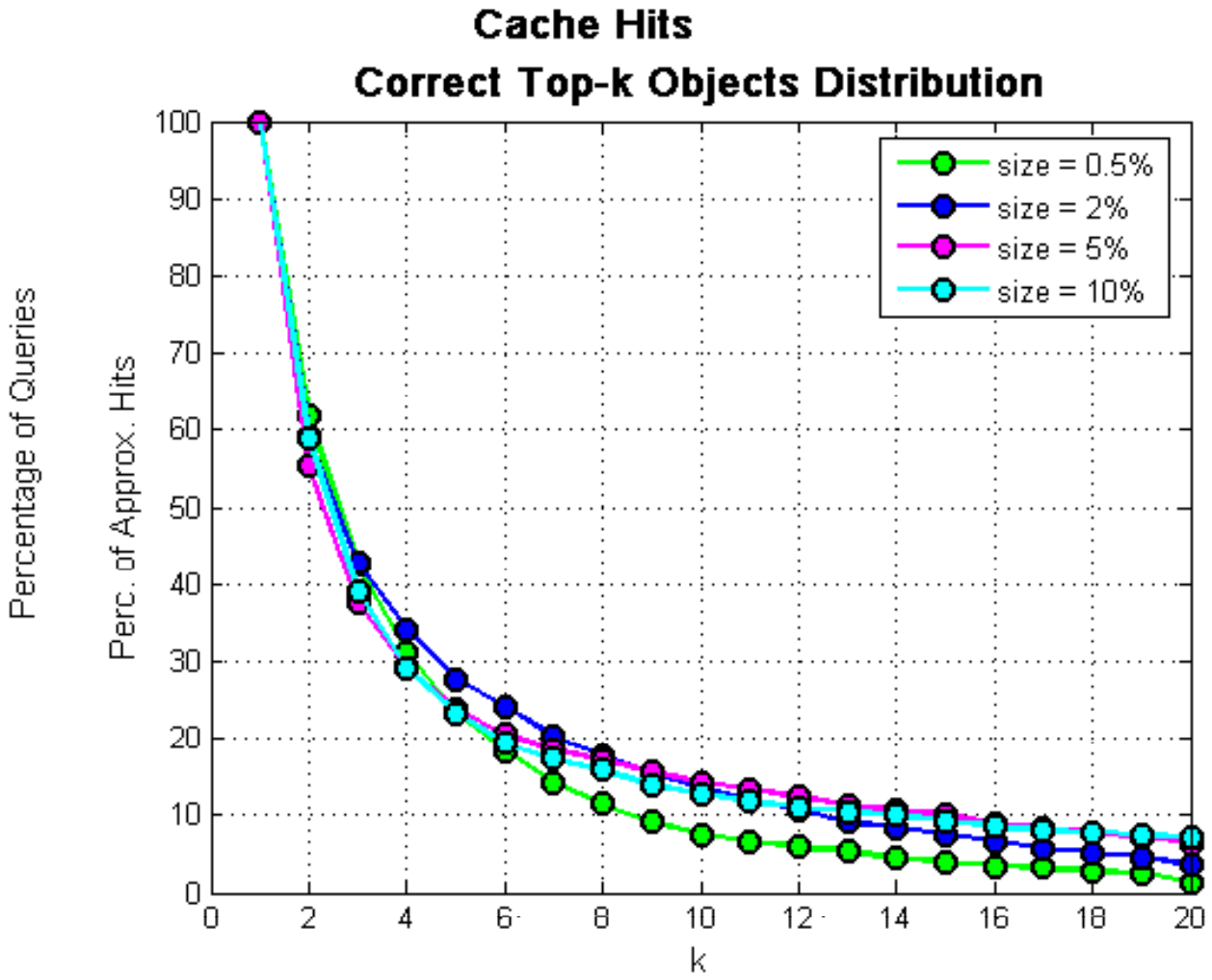
- Additionally, one could use other objects in the cache to provide an approximate answer.



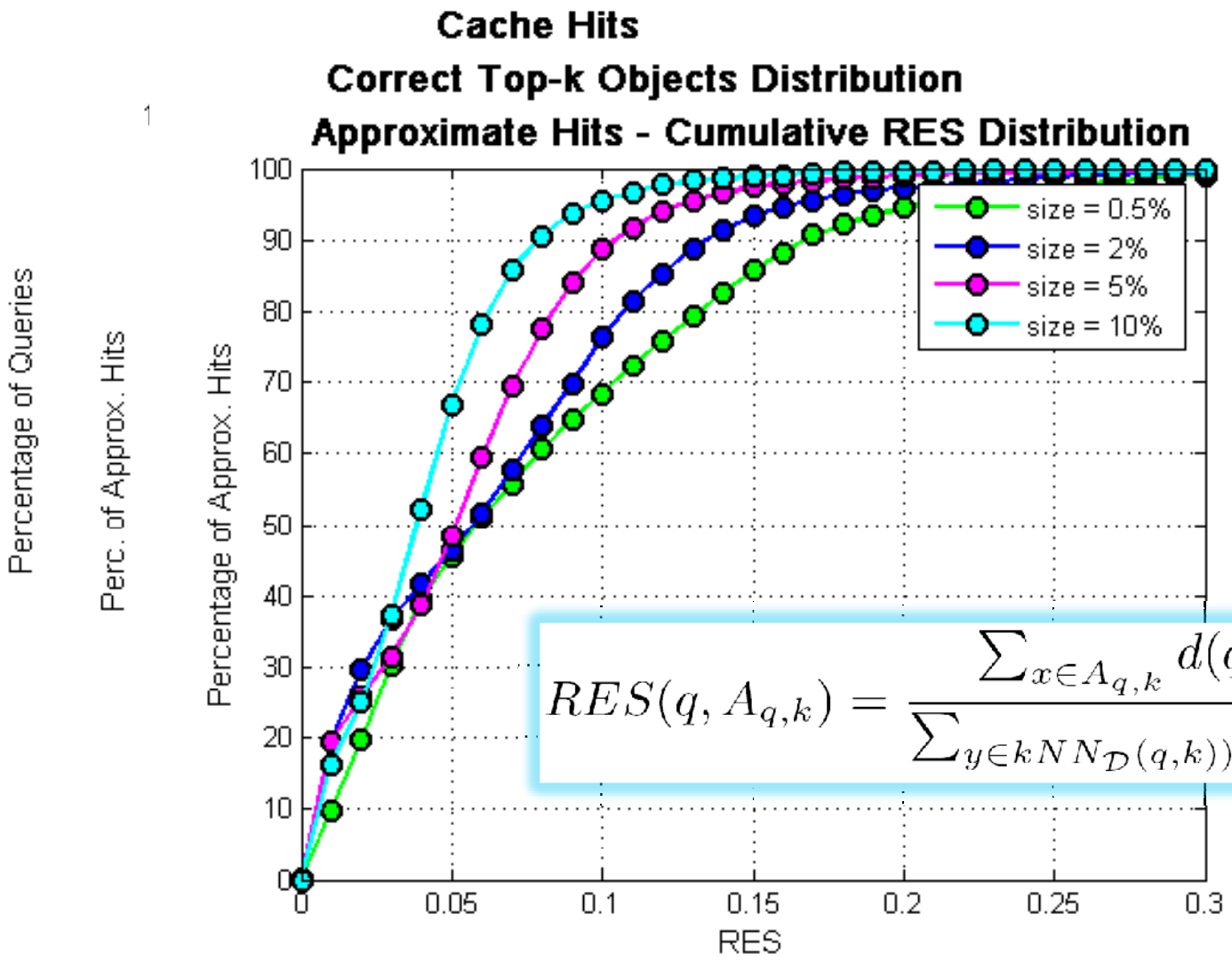
SOME RESULTS ...



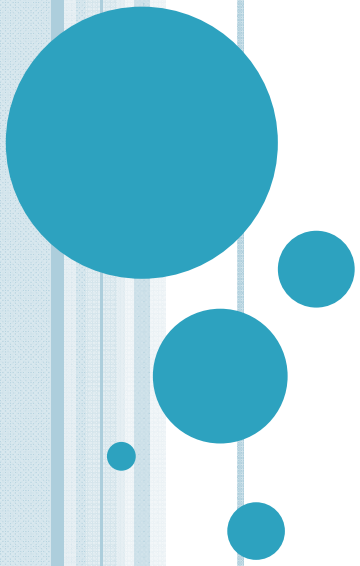
SOME RESULTS ...



SOME RESULTS ...



$$RES(q, A_{q,k}) = \frac{\sum_{x \in A_{q,k}} d(q,x)}{\sum_{y \in kNN_{\mathcal{D}}(q,k))} d(q,y)} - 1$$



The End.

... Grazie ! =)