Introduction to Data Exploration

Data Processing vs. Querying vs. Exploration

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Explain modalities for data exploration

Exploratory Search

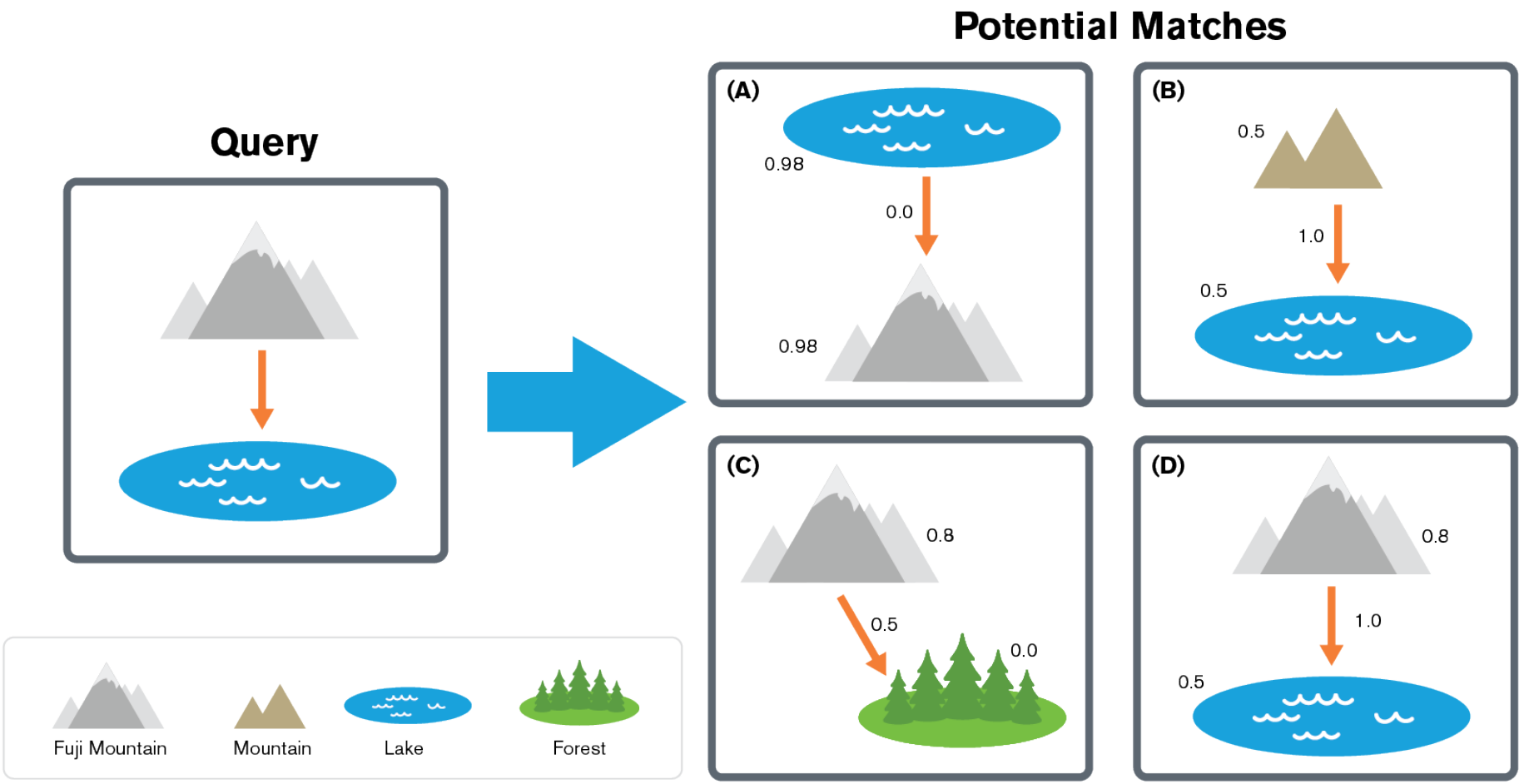
Acquiring new knowledge and revealing new facts

- Analysis (identify common patterns or outliers)
- Comparison (quantify similarity/differences)
- Aggregation (create groups, clusters)
- Transformation (use a more convenient representation
- Visualization

Exploratory Querying

- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Query by Example / Similarity Search



Ranked retrieval

- When not all sub-goals need to be satisfied (i.e., partial matches are allowed) each and every data element in the database is a potential match
- Hence the query results need to be ranked according to some objective or subjective criteria

Top-K Search

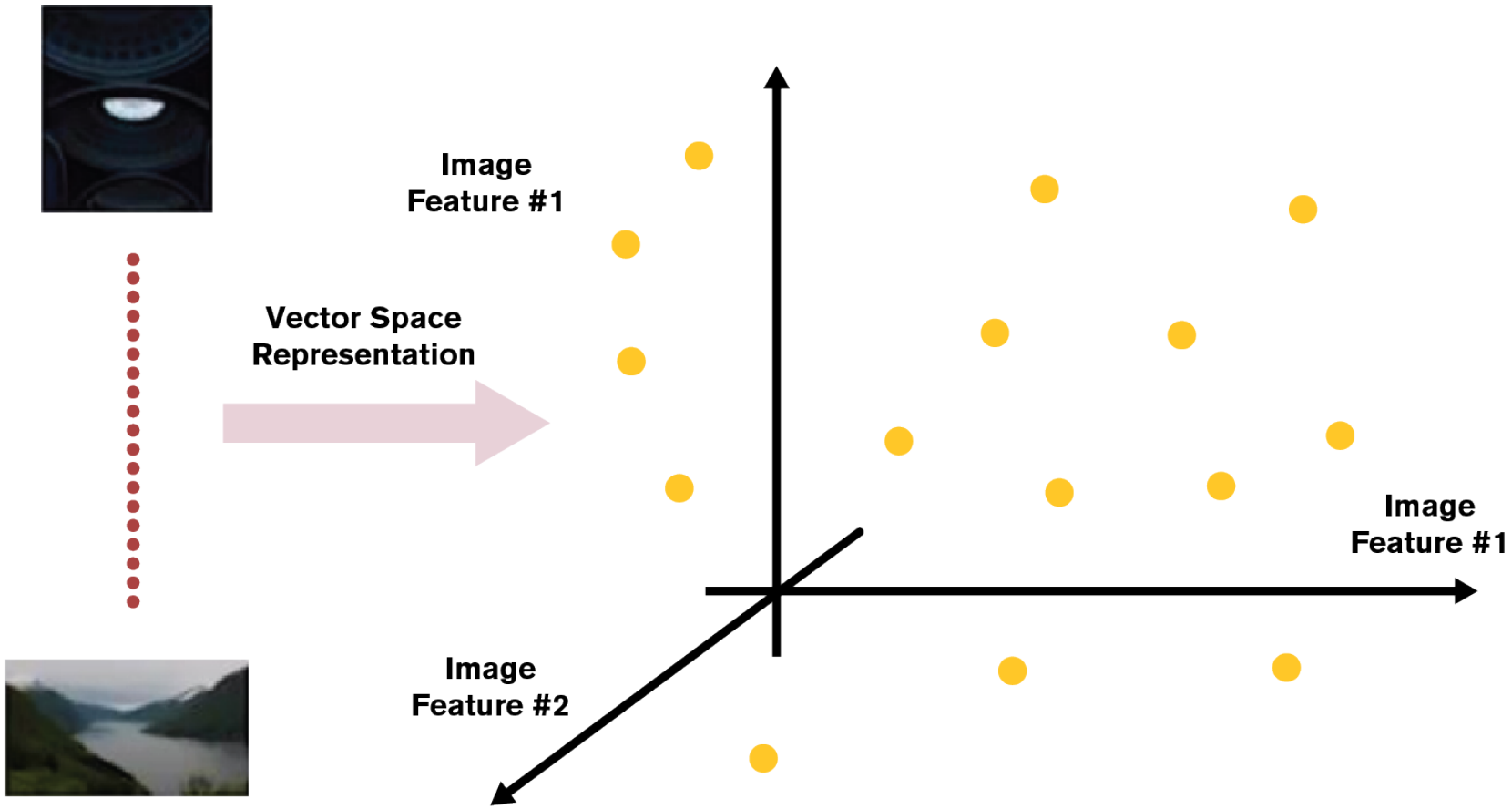
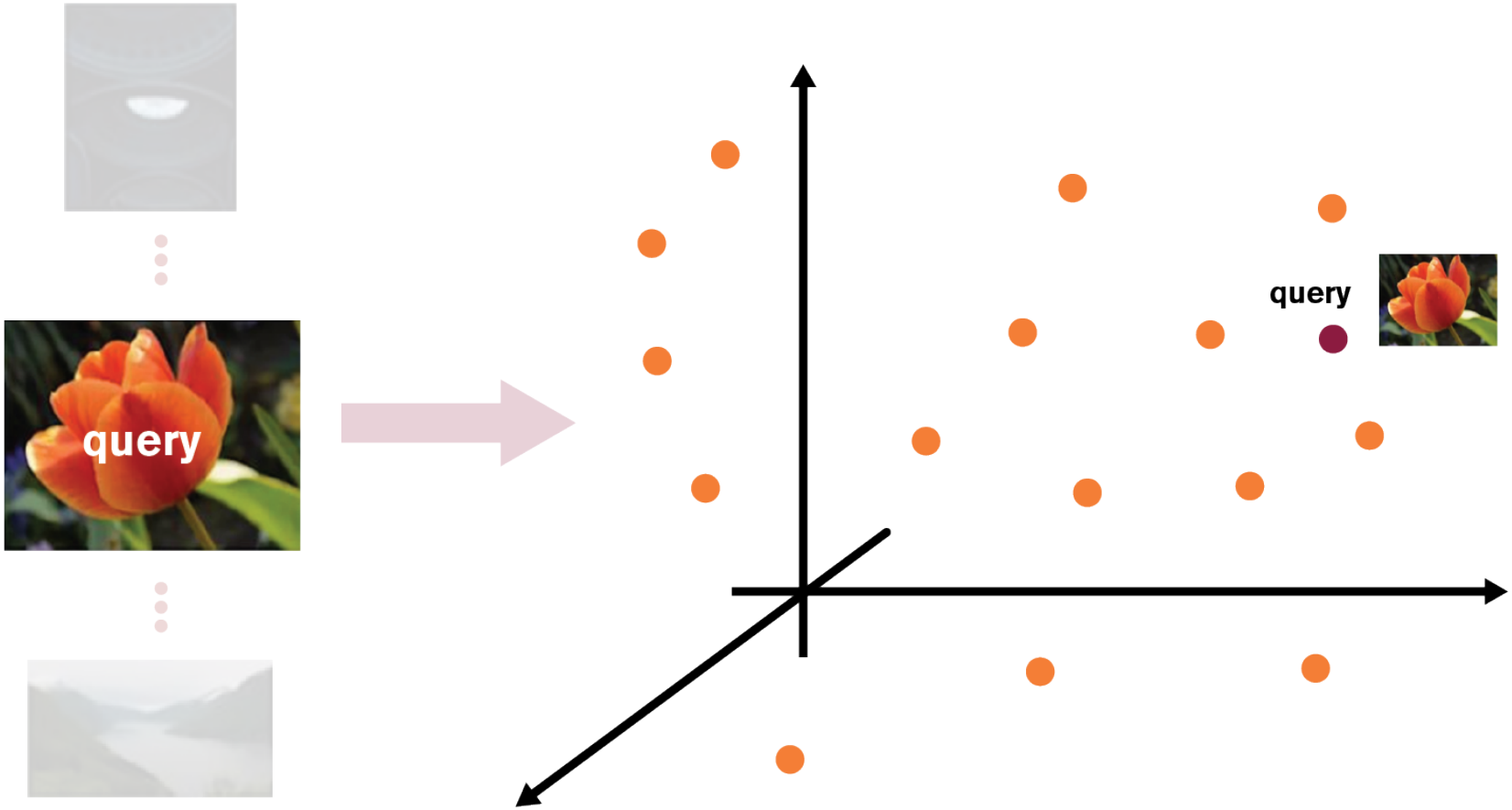
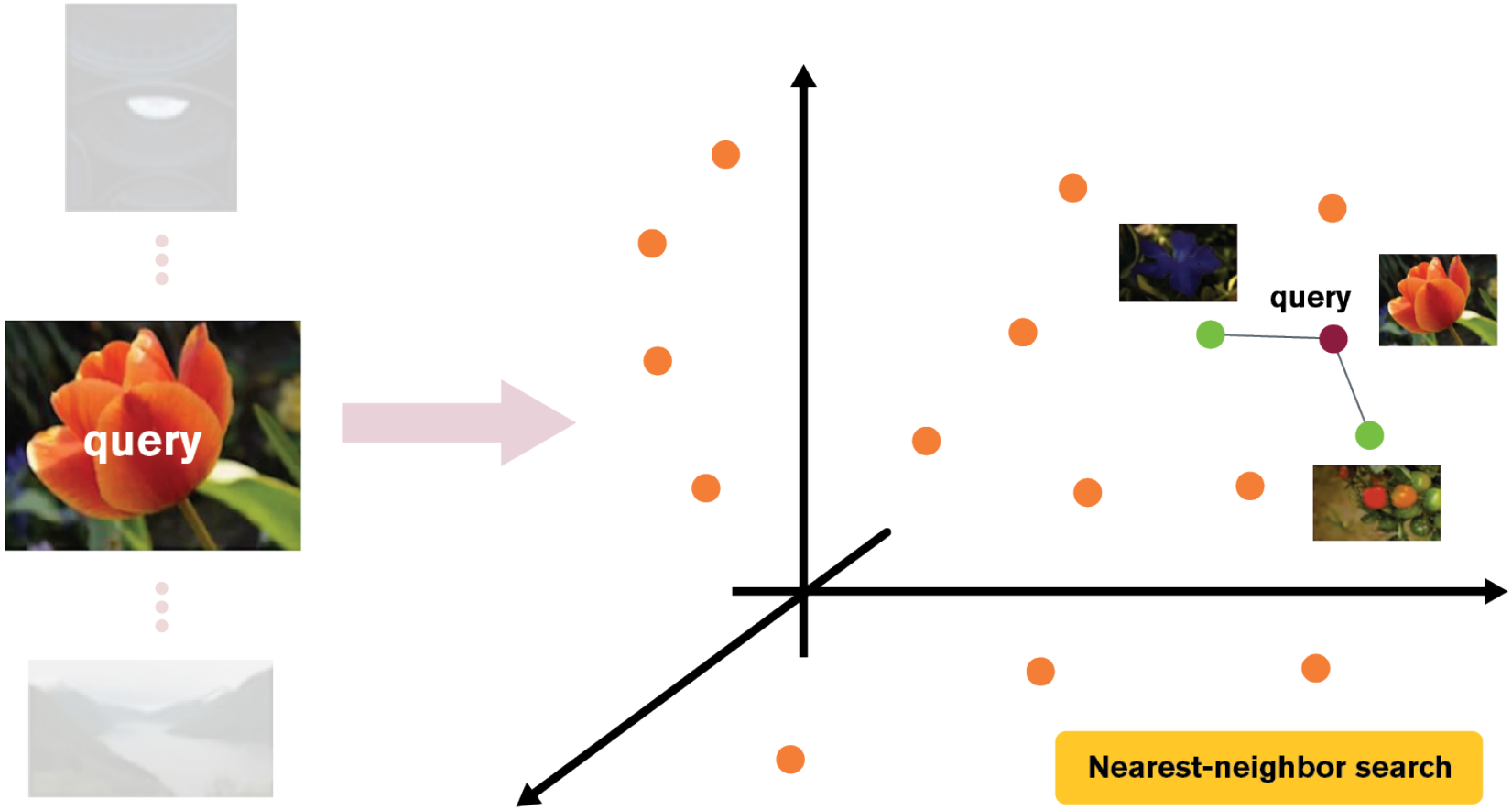


Image
Feature #1
Vector Space
Representation
Image
Feature #1
Image
Feature #2

“Find K=2 most similar images”



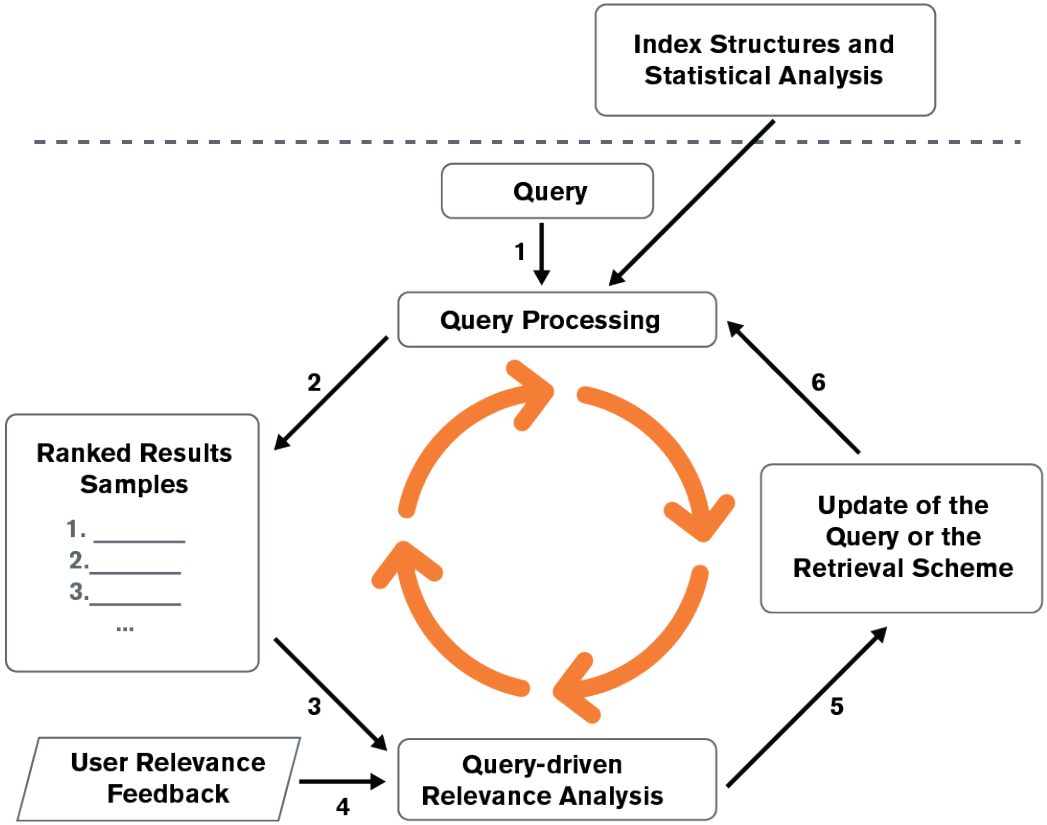
“Find K=2 most similar images”



Sematic gap/subjectivity

Relevance feedback

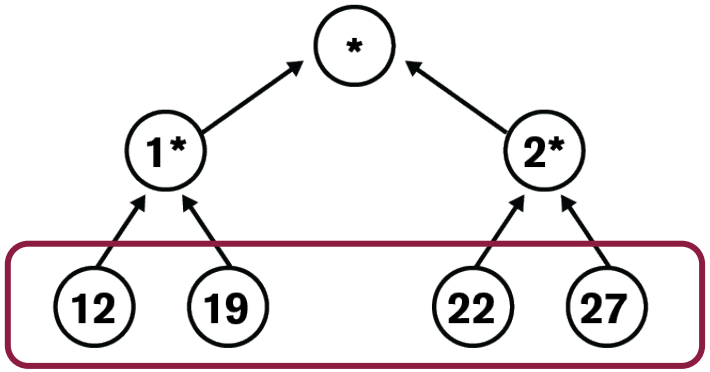
- In a request to identify images “similar” to an example, the visual features of the images are relevant for the user’s query must be inferred from feedback to identify the most relevant images.



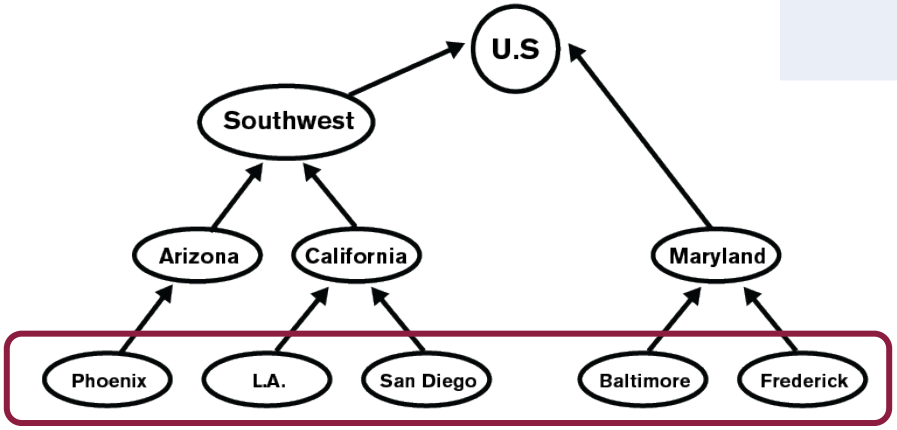
Exploratory Querying

- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Age Metadata



Location Metadata

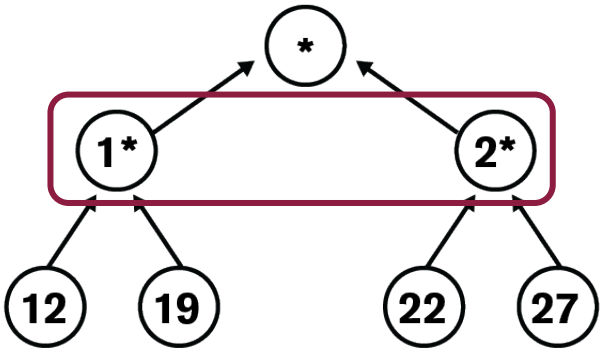


Data Table

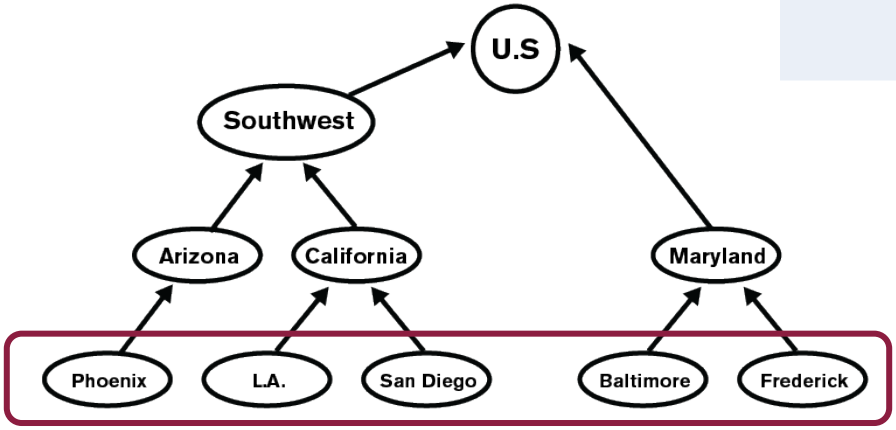
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Roll-Up on Age

Age Metadata



Location Metadata

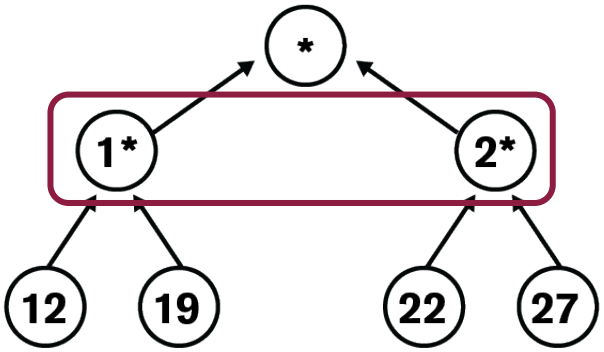


Data Table

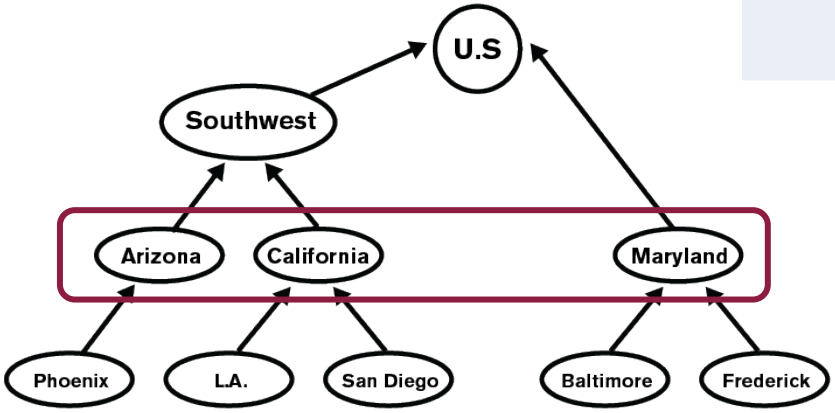
Name	Age	Location
John	1*	Phoenix
Sharon	1*	Los Angeles
Mary	1*	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Roll-Up on Location

Age Metadata



Location Metadata

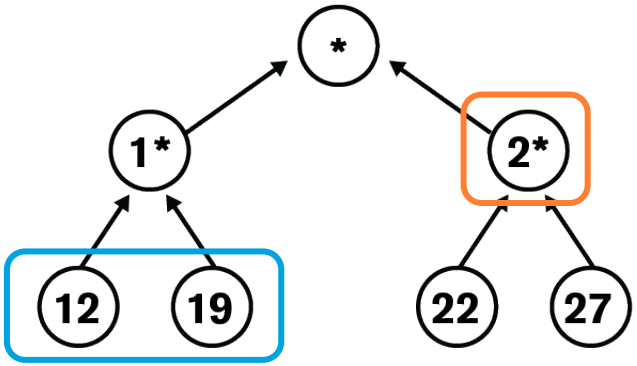


Data Table

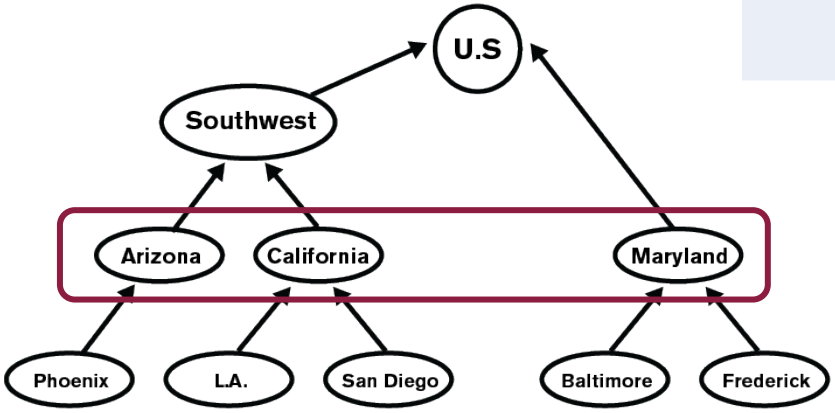
Name	Age	Location
John	1*	Phoenix
Sharon	1*	Los Angeles
Mary	1*	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Drill Down on Age (1*)

Age Metadata



Location Metadata

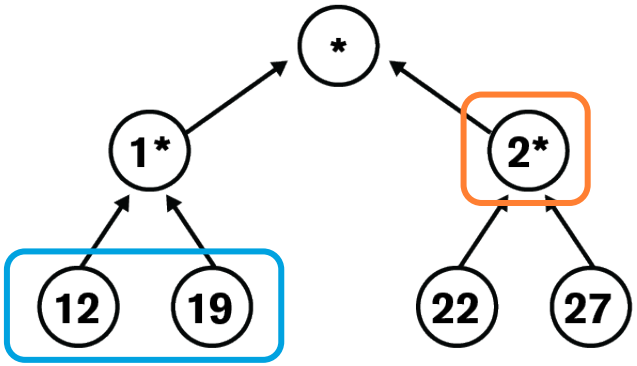


Data Table

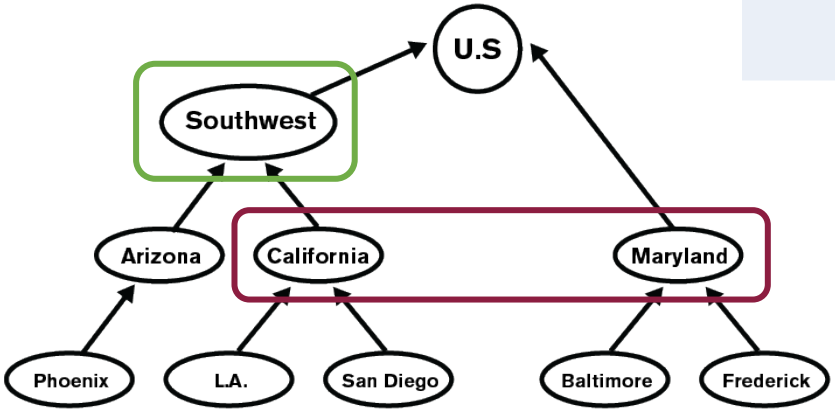
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Roll-Up on Location (Arizona)

Age Metadata



Location Metadata



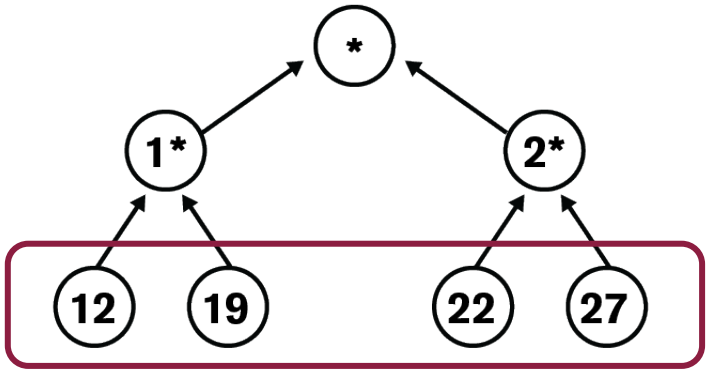
Data Table

Name	Age	Location
John	12	Southwest
Sharon	19	California
Mary	19	California
Peter	2*	Maryland
James	2*	Maryland
Alice	2*	Maryland

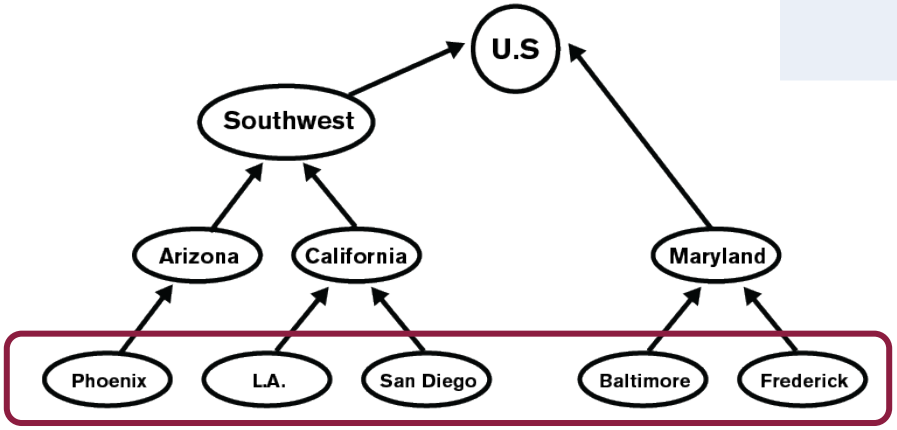
Exploratory Querying

- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Age Metadata



Location Metadata

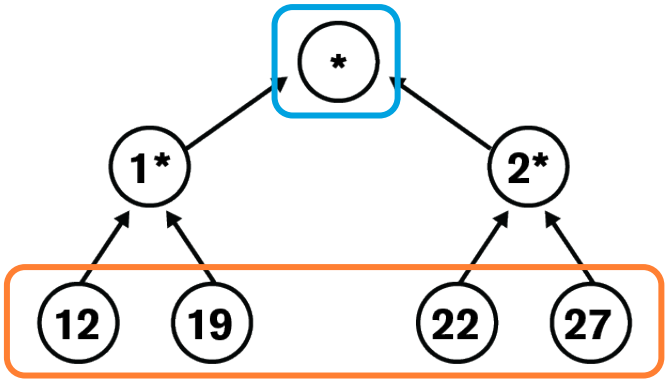


Data Table

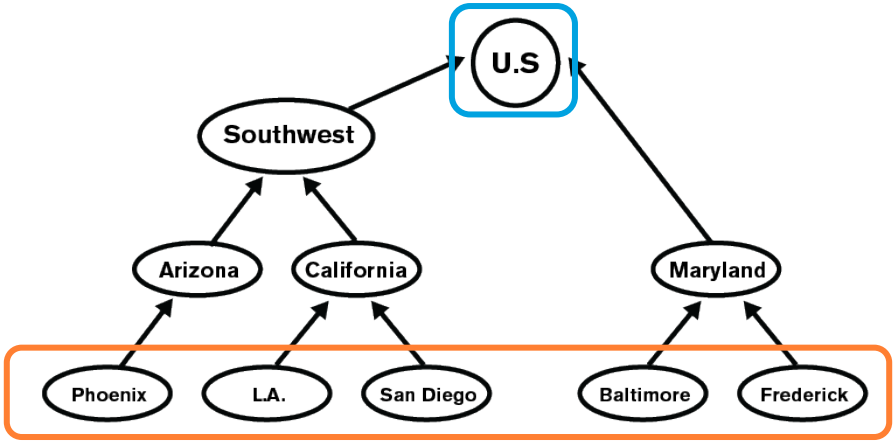
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Summarization (target # rows = 2)

- Age Metadata



- Location Metadata



Summarized Data Table

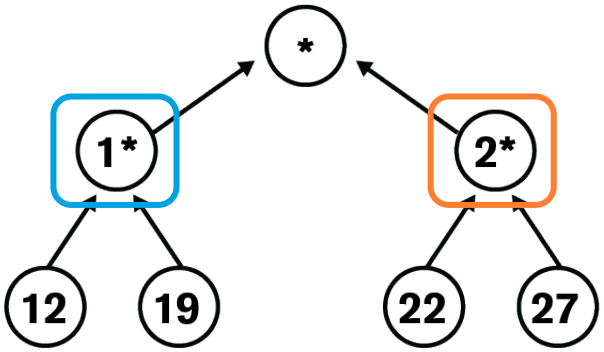
Name	Age	Location	Aggregate (count)
-	1*	Southwest	3
-	2*	Maryland	3

Exploratory Querying

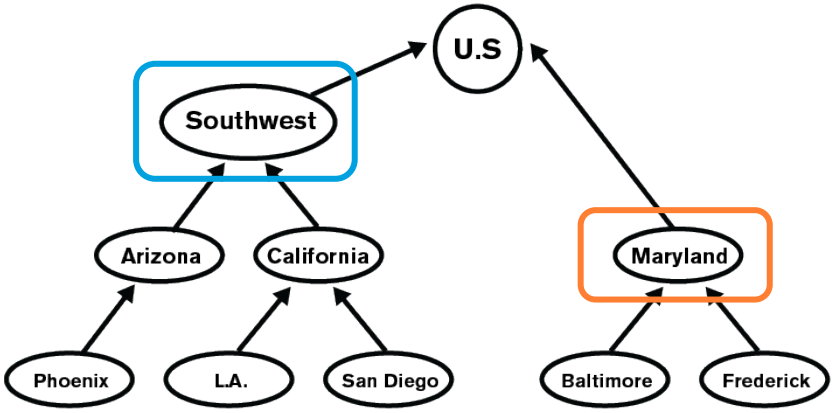
- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Alternative Summarization (target # rows = 2)

Age Metadata



Location Metadata

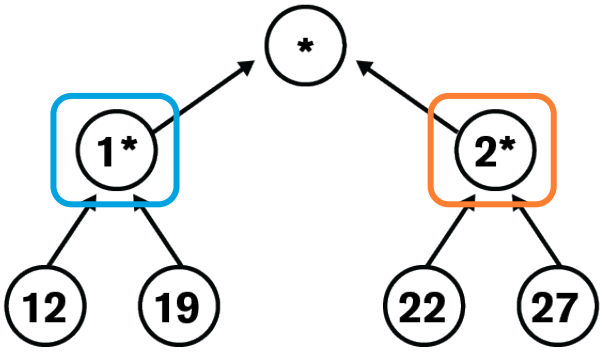


Summarized Data Table

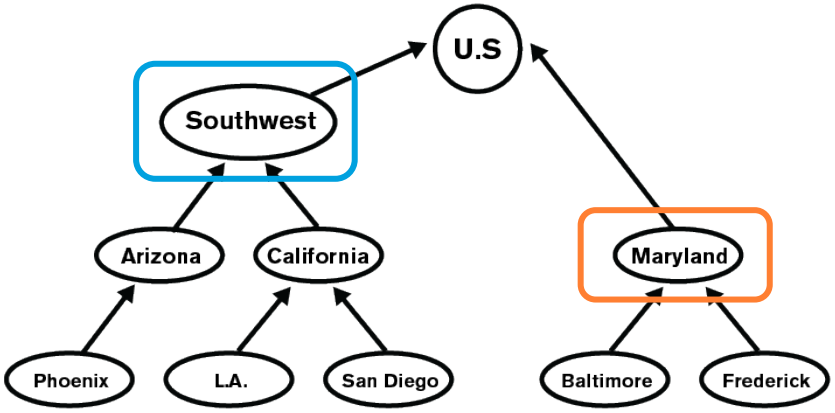
Name	Age	Location	Aggregate (count)
-	1*	Southwest	3
-	2*	Maryland	3

Summarization + Aggregation
(target # rows = 2; max(age))

- Age Metadata



- Location Metadata

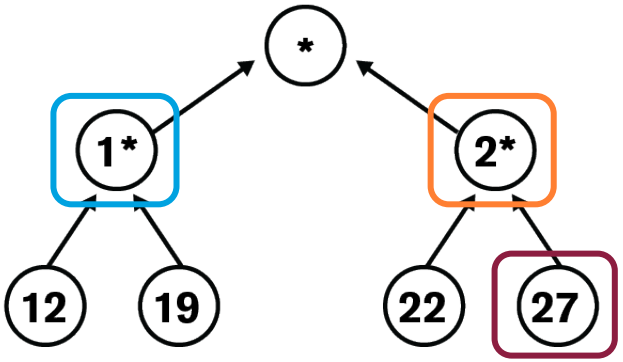


Summarized/Aggregated Data Table

Name	Max(AGE)	Location	Aggregate (count)
-	19	Southwest	3
-	27	Maryland	3

Summarization + Iceberg
(target # rows = 2; max(AGE)>20)

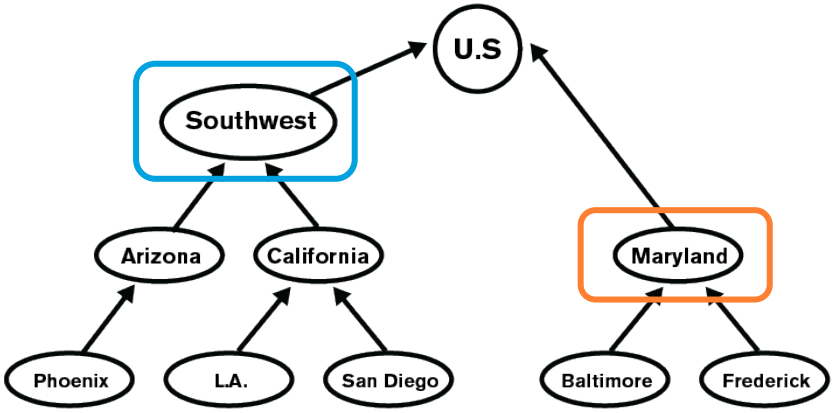
- Age Metadata



Summarized/Aggregated Data Table

Name	Age	Location	Aggregate (count)
-	27	Maryland	3

- Location Metadata



Exploratory Querying

- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Skylines

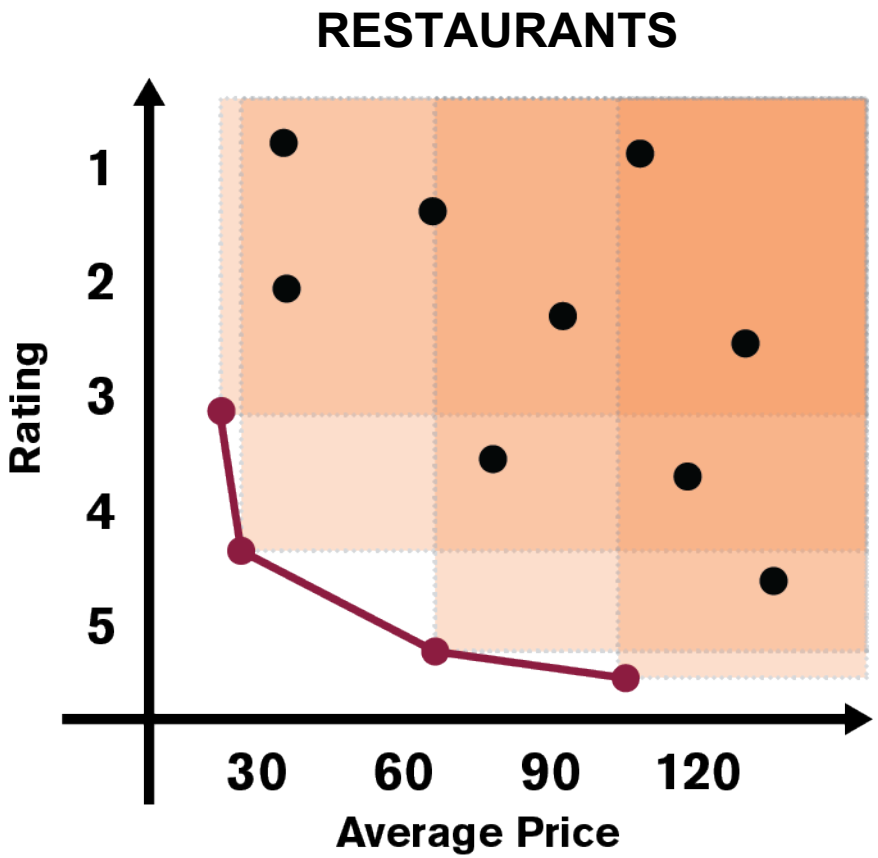
Question

- The higher the rating, the better
- The cheaper the price the better

Which restaurants would you consider?

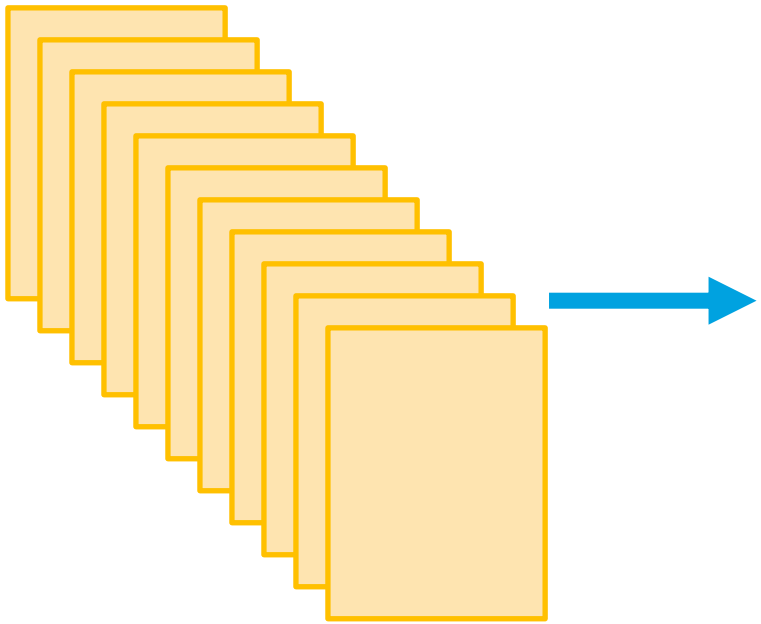
Skylines

- **Objects in the "skyline" are not dominated by any other objects in the database**
  - Also known as the Maximum Vector Problem [Kung 75]
  - Coined as "Skylines" in [Börzsönyi01]



Data sketches – example: tag clouds

Document Collection

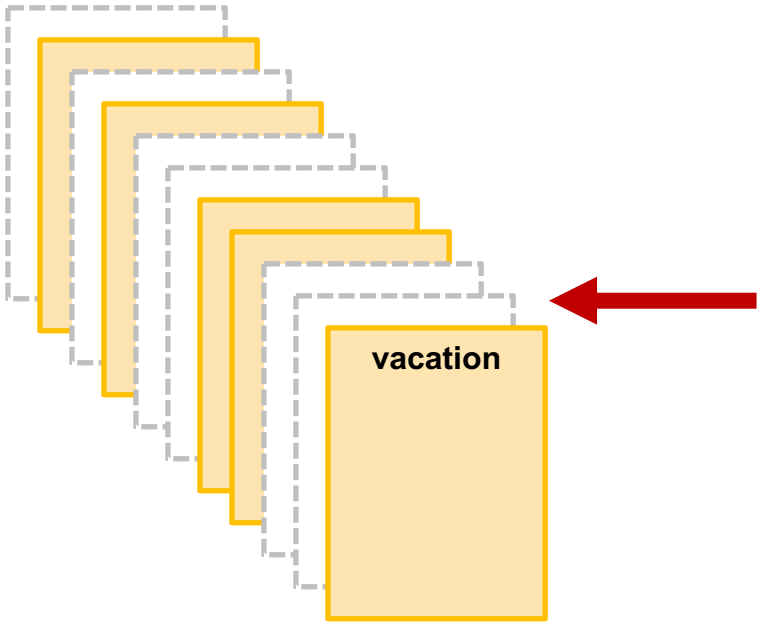


Tag/term cloud

amsterdam animal animals april architecture art australia baby barcelona beach berlin bird birthday black blackandwhite blue boston building bw california cameraphone camping canada canon car cat cats chicago china christmas church city clouds color concert day dc dog dogs england europe family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon hongkong house india ireland italy japan july kids lake landscape light live london losangeles macro march may me mexico moblog mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo portrait red river roadtrip rock rome san sanfrancisco school scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo

Data sketches – example: tag clouds

Document Collection

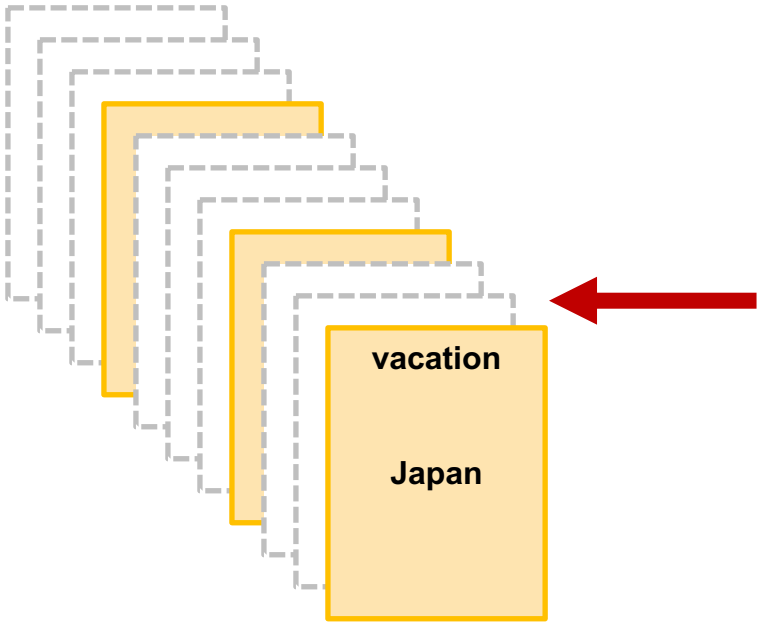


Tag/term cloud

amsterdam animal animals april architecture art australia baby barcelona beach berlin bird birthday black blackandwhite blue boston building bw california cameraphone camping canada canon car cat cats chicago china christmas church city clouds color concert day dc dog dogs england europe family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon hongkong house india ireland italy japan july kids lake landscape light live london losangeles macro march may me mexico moblog mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo portrait red river roadtrip rock rome san sanfrancisco school scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo

Data sketches – example: tag clouds

Document Collection



Tag/term cloud