Introduction to Data Exploration

Data Processing vs. Querying vs. Exploration

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Explain data warehouses and exploratory querying

Data Processing vs. Querying vs. Exploration

Data processing

- user knows what s/he wants
- user have a function/procedure /workflow to compute what s/he wants

Querying

- user knows what s/he wants
- user can describe what s/he wants

Navigation

- user knows what s/he wants
- user does not know how to describe/locate what s/he wants

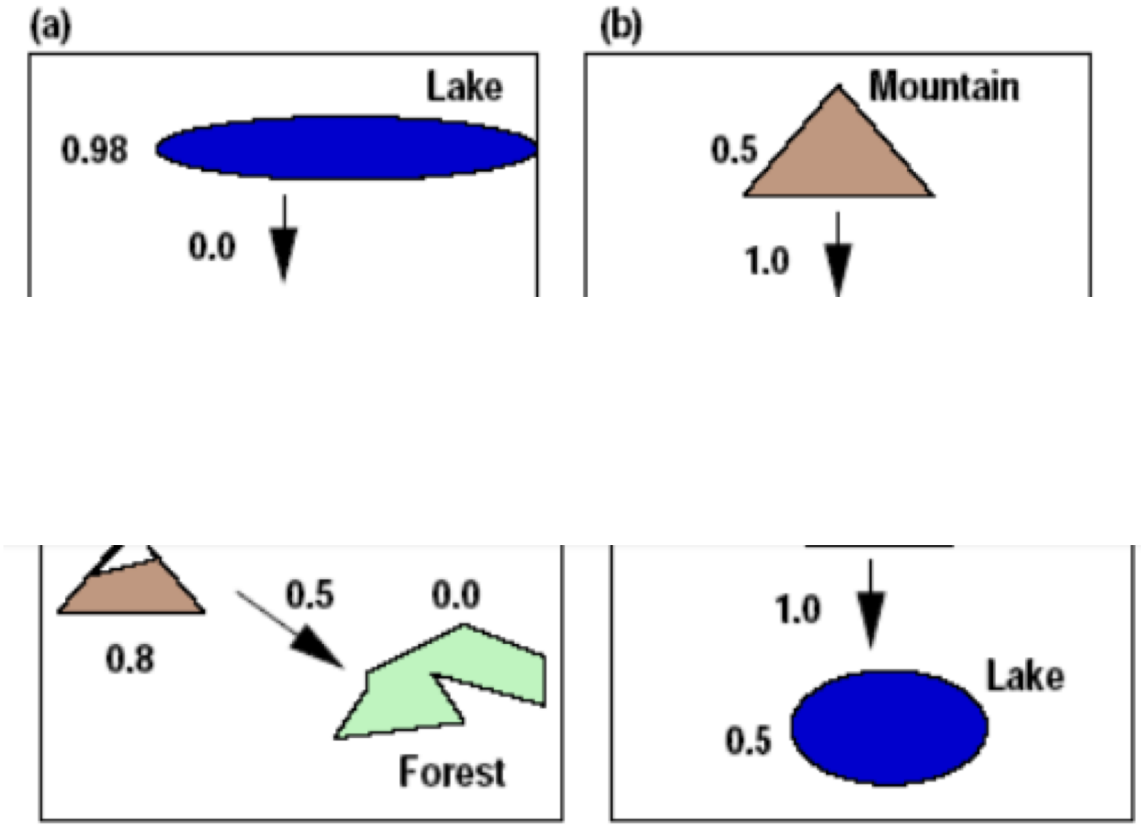
Exploration

- user does not precisely know what s/he wants
- user wants to get an idea about the available data

Query by Example / Similarity Search

- When not all sub-goals need to be satisfied each and every data element in the database is a potential match

- query results need to be ranked according to some objective or subjective criteria



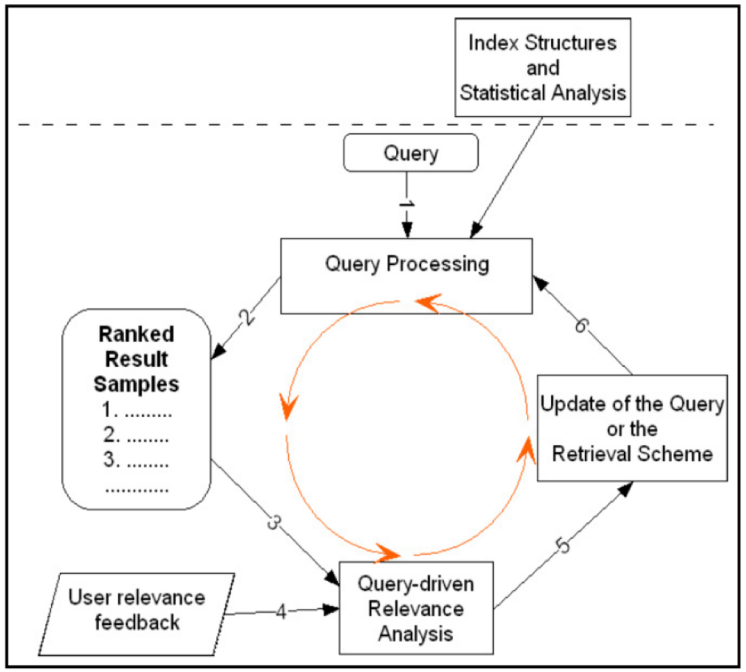
Relevance Feedback

Sematic gap/subjectivity:

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

Computational challenges:

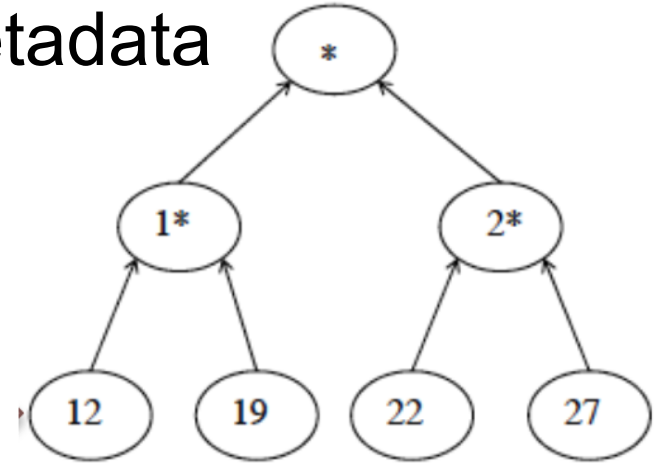
- When the query is not precisely known in advance, the number of candidate matches explodes



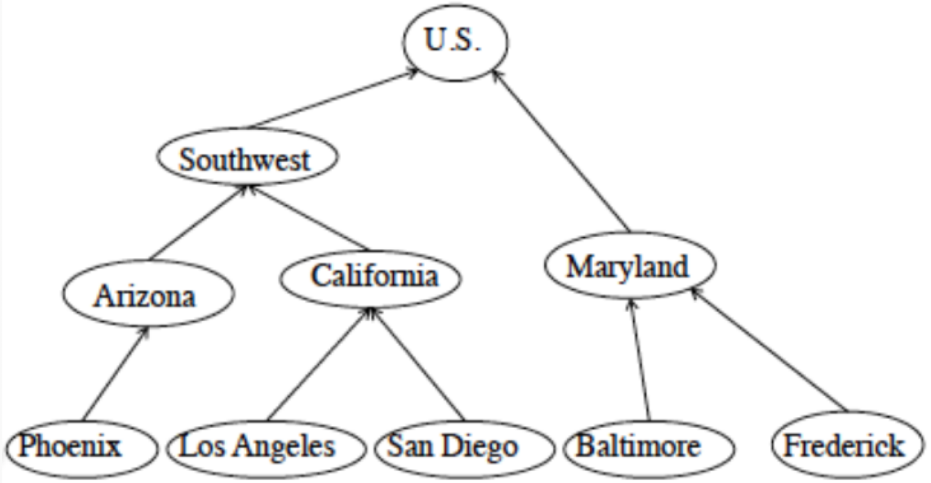
Exploratory Querying

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

Age Metadata



Location Metadata

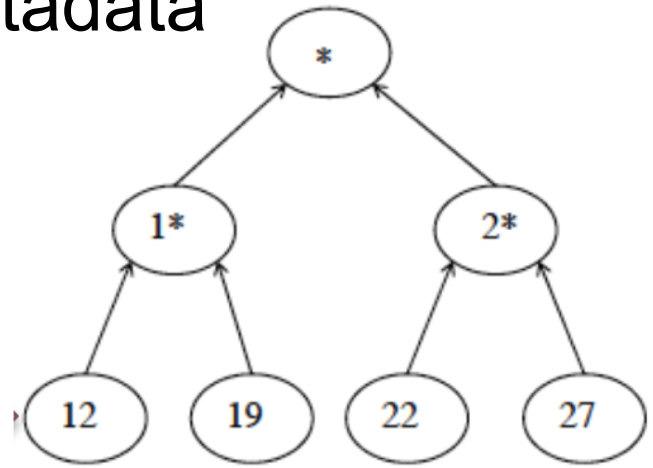


Data Table

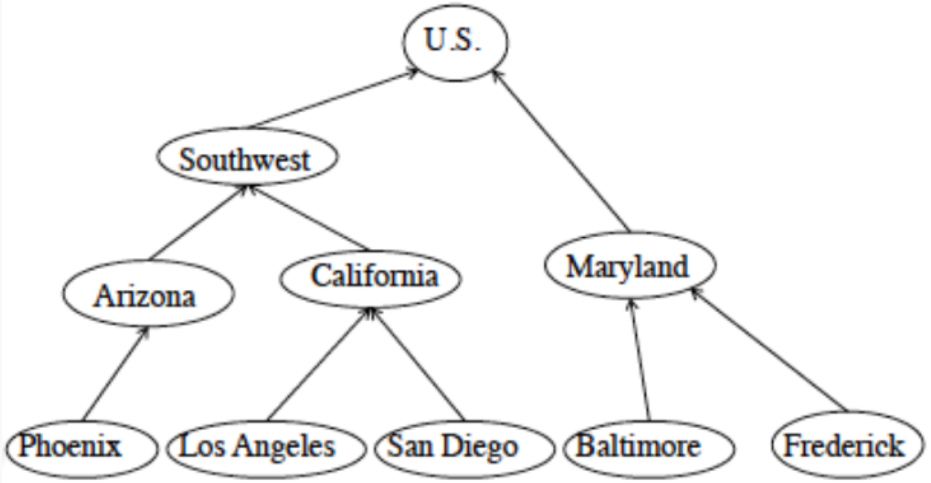
NAME	AGE	LOCATION
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Roll-Up on Age

Age Metadata



Location Metadata

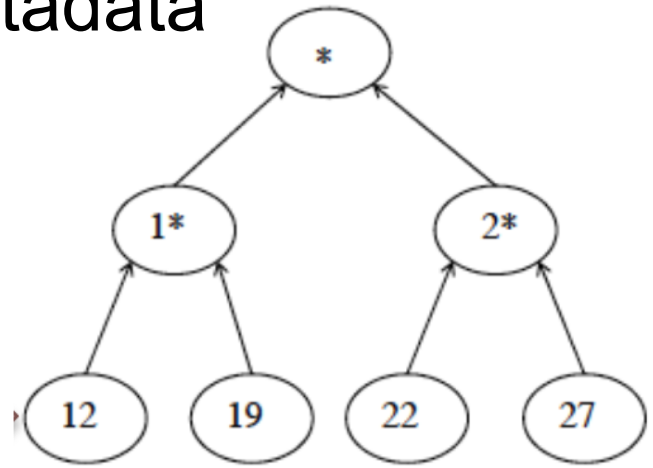


Data Table

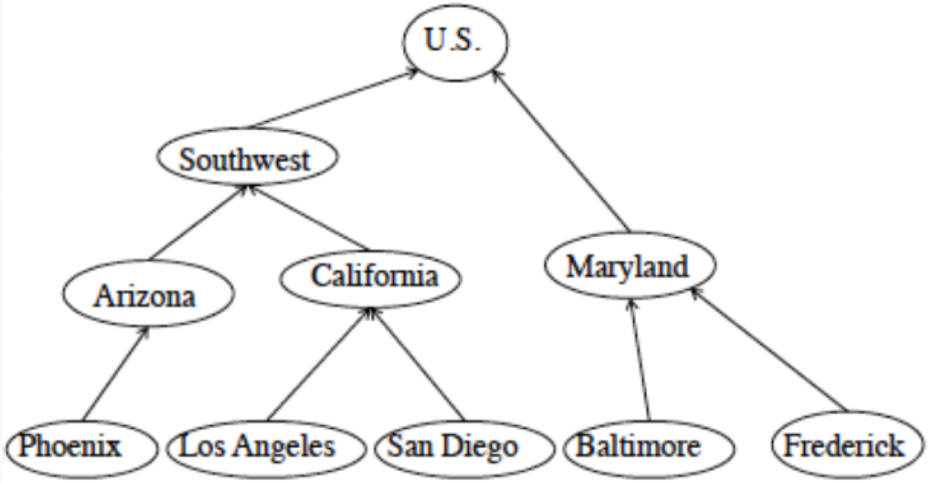
NAME	AGE	LOCATION
John	1*	Phoenix
Sharon	1*	Los Angeles
Mary	1*	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Roll-Up on Location

Age Metadata



Location Metadata

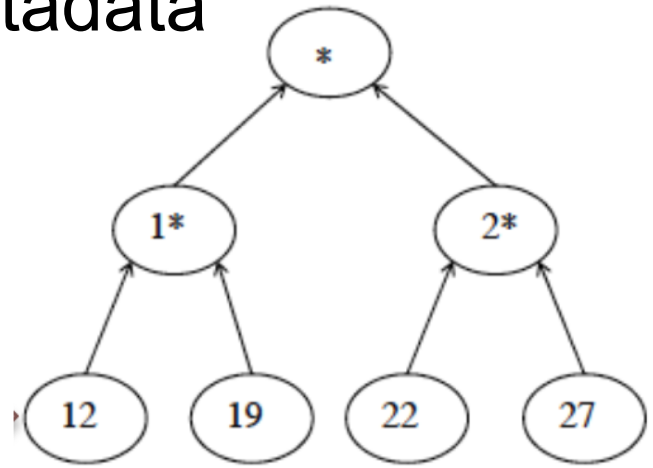


Data Table

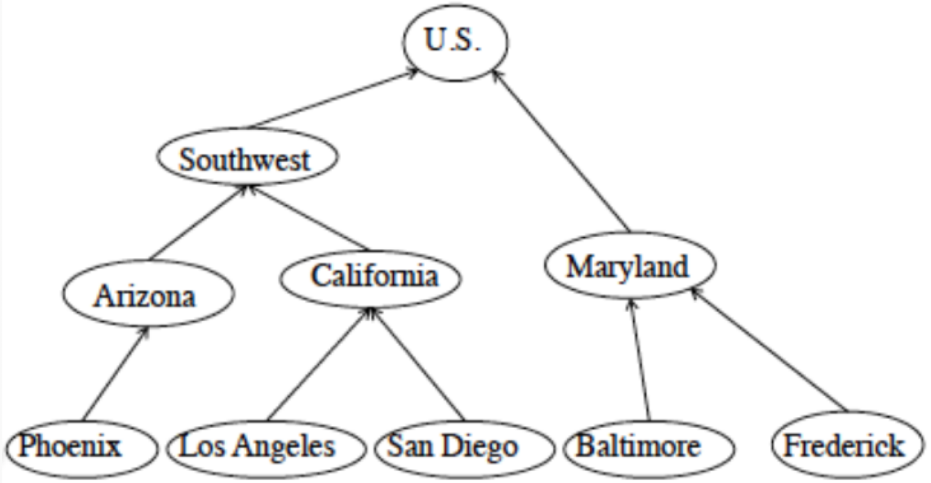
NAME	AGE	LOCATION
John	1*	Arizona
Sharon	1*	California
Mary	1*	California
Peter	2*	Maryland
James	2*	Maryland
Alice	2*	Maryland

Drill Down on Age (1*)

Age Metadata



Location Metadata

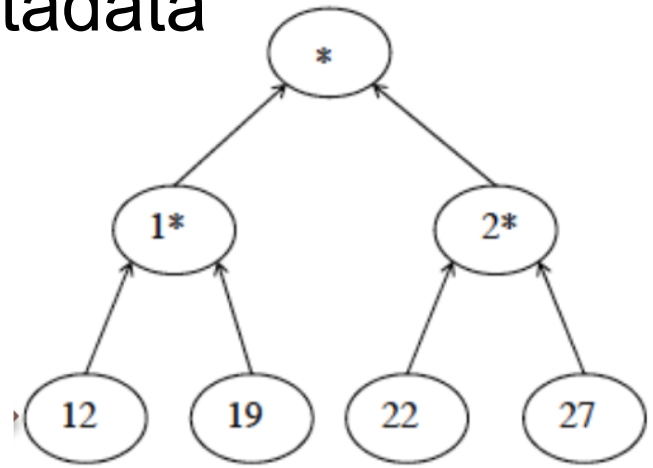


Data Table

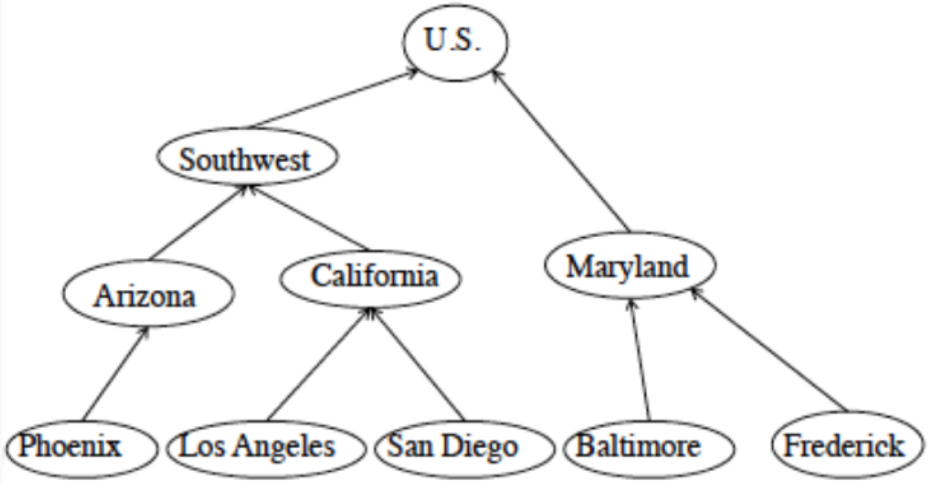
NAME	AGE	LOCATION
John	12	Arizona
Sharon	19	California
Mary	19	California
Peter	2*	Maryland
James	2*	Maryland
Alice	2*	Maryland

Roll-Up on Location(Arizona)

Age Metadata



Location Metadata

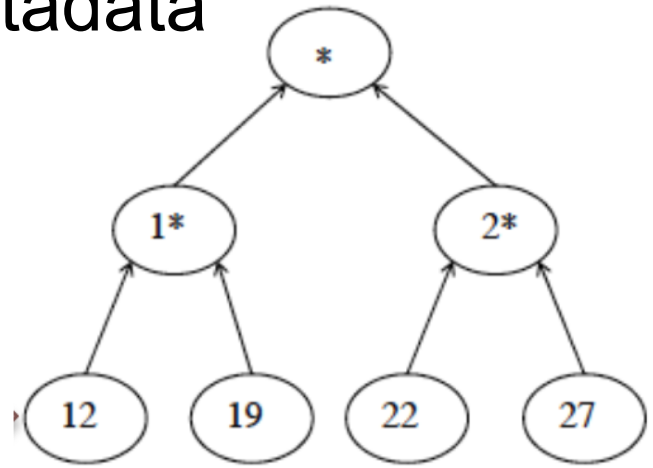


Data Table

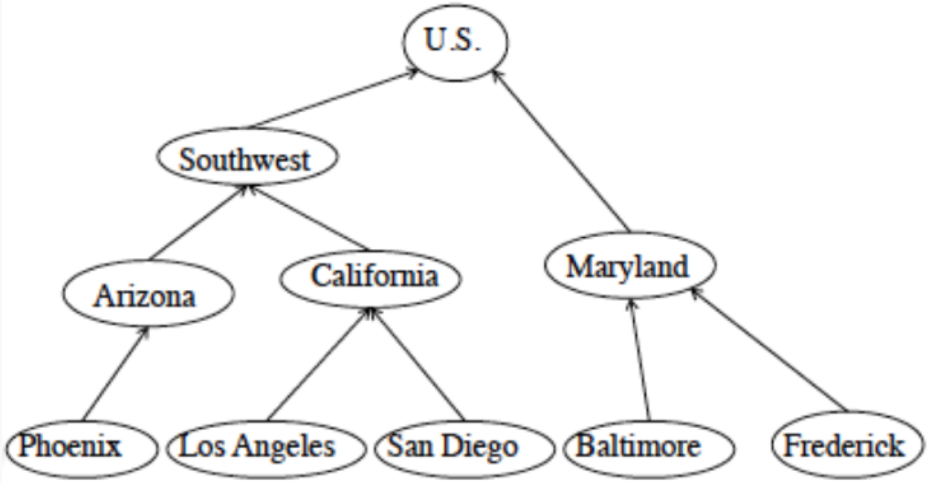
NAME	AGE	LOCATION
John	12	Southwest
Sharon	19	California
Mary	19	California
Peter	2*	Maryland
James	2*	Maryland
Alice	2*	Maryland

Summarize (target row = 2)

Age Metadata



Location Metadata

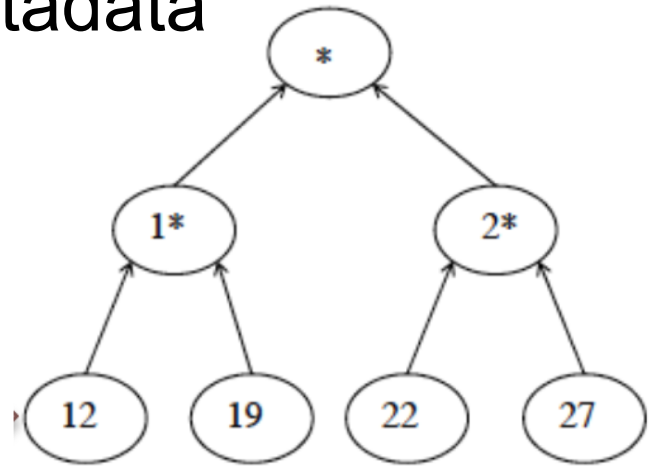


Summarized Data Table

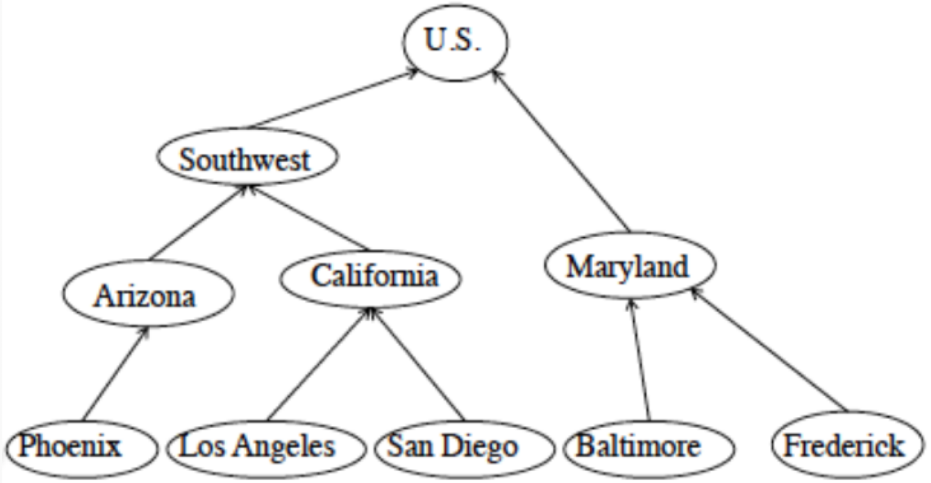
NAME	AGE	LOCATION	Aggregate (Count)
-	1*	Southwest	3
-	2*	Maryland	3

Summarize (target row = 2)

Age Metadata



Location Metadata

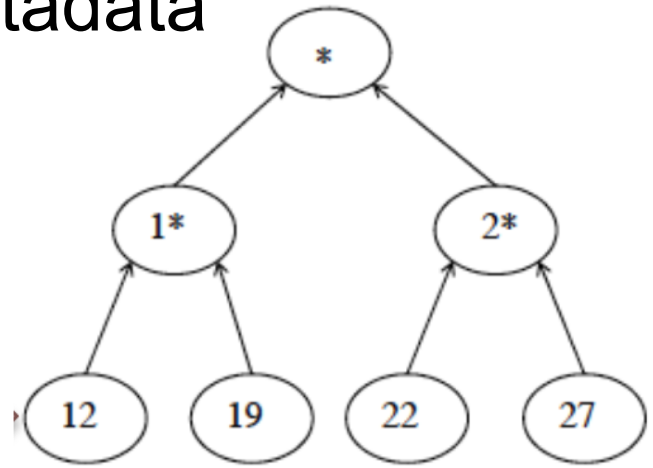


Summarized Data Table

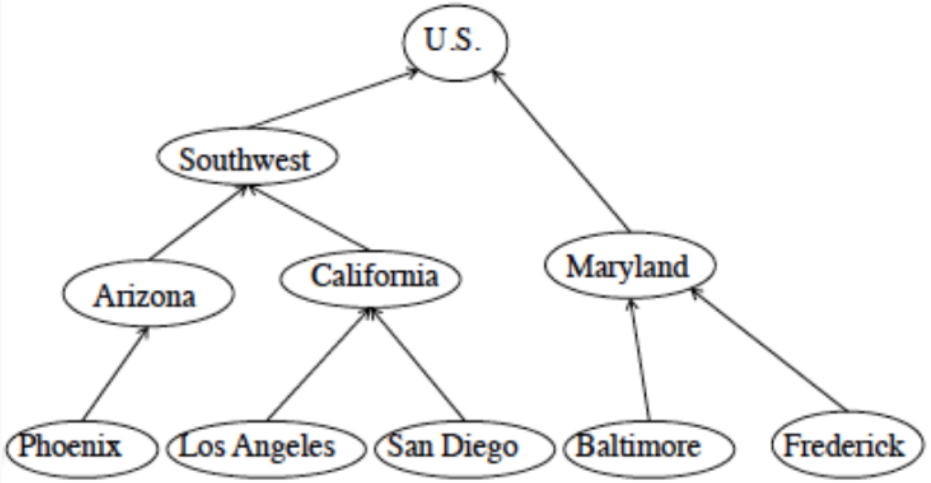
NAME	AGE	LOCATION	Aggregate (Count)
-	19	Southwest	3
-	27	Maryland	3

Summarize (target row = 2)

Age Metadata



Location Metadata

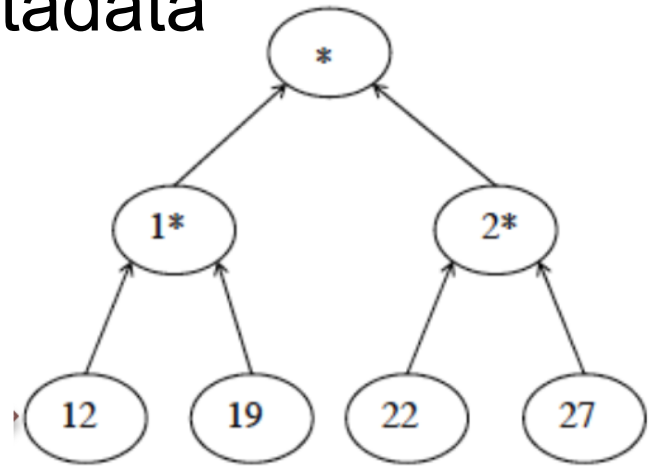


Summarized Data Table

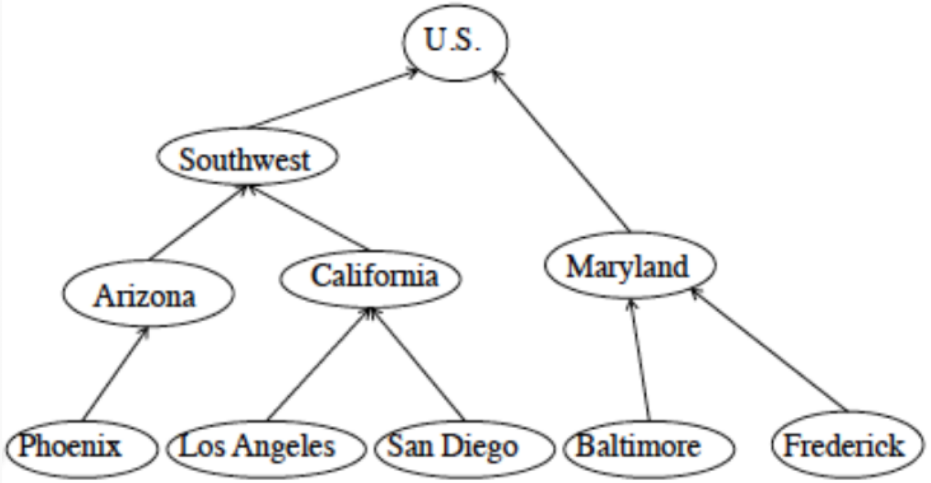
NAME	AGE	LOCATION	Aggregate (Count)
-	27	Maryland	3

Summarize (target row = 2)

Age Metadata



Location Metadata

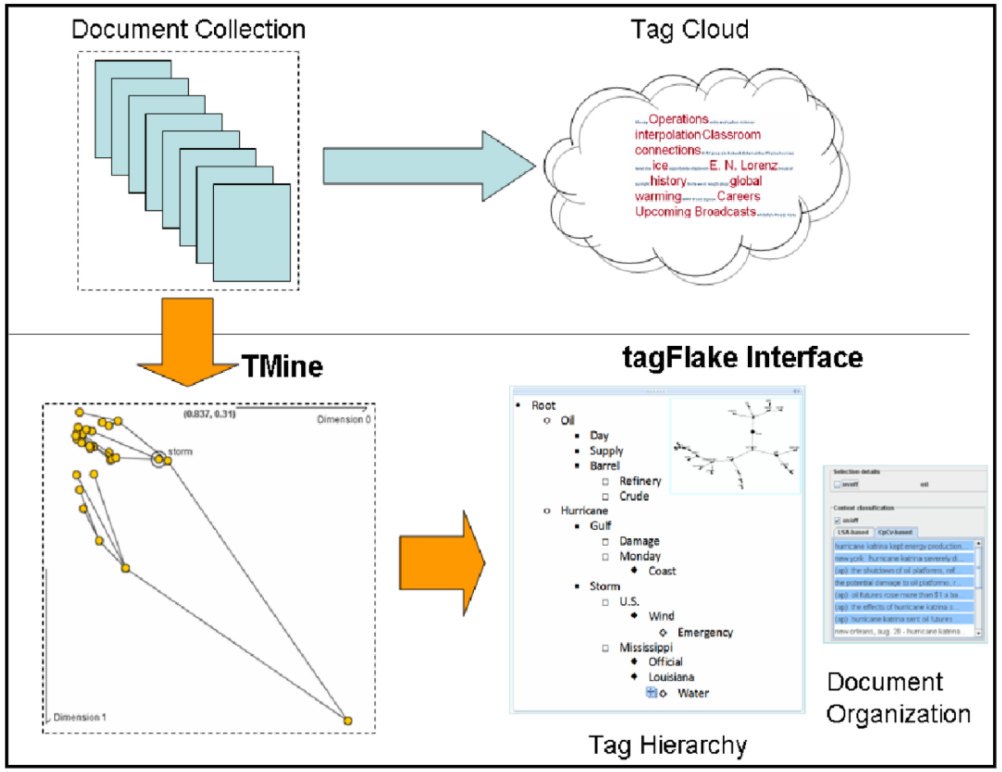


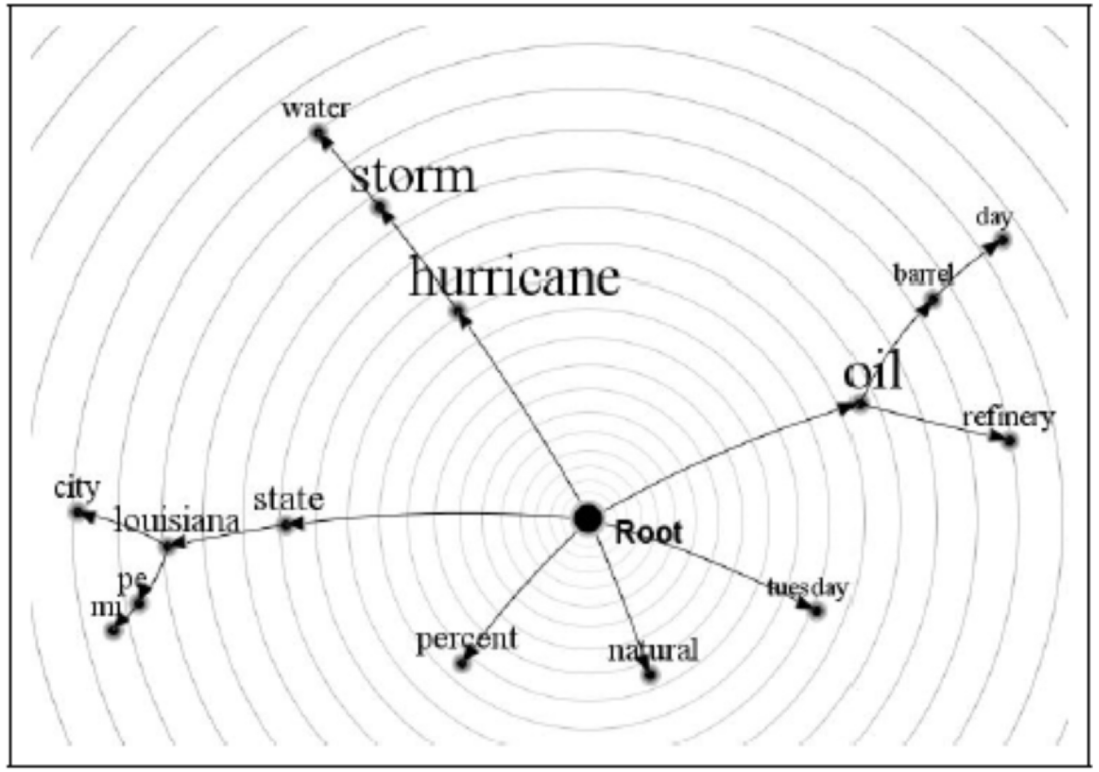
Summarized Data Table

NAME	AGE	LOCATION	Count
-	-	U.S	5
Sharon	19	Los Angeles	1

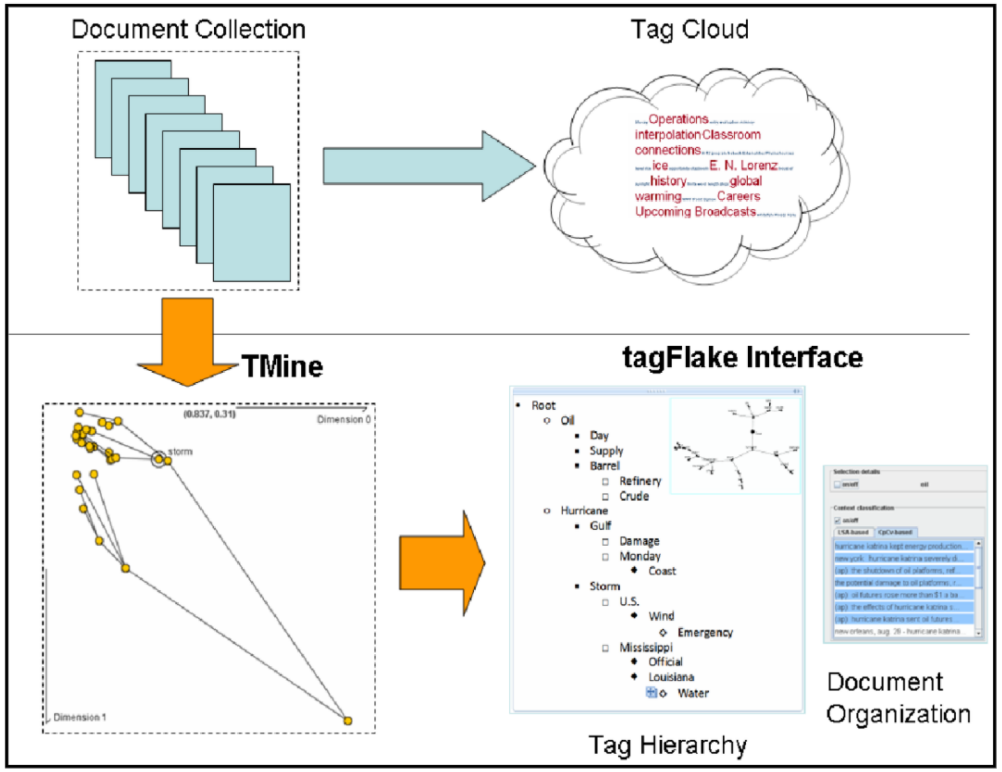
Sketch: Word Cloud

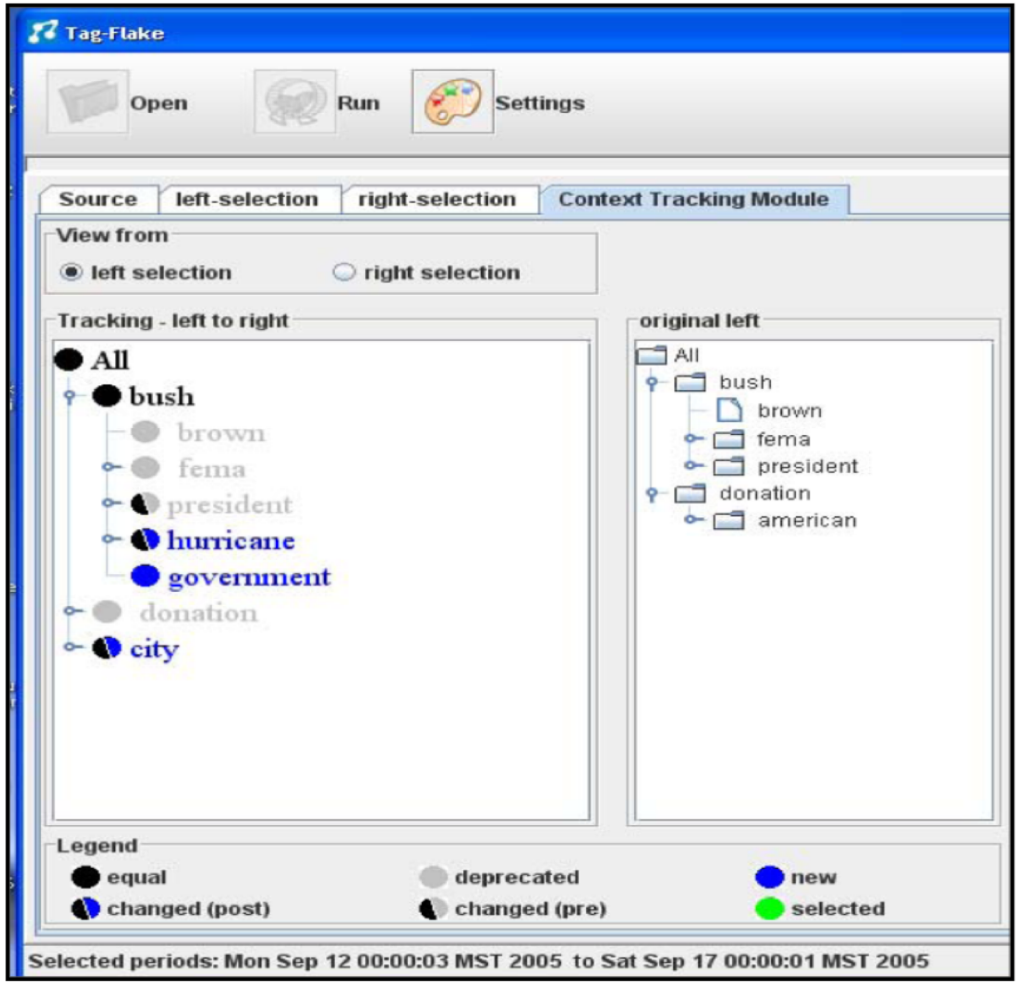
Extract Latent Structure



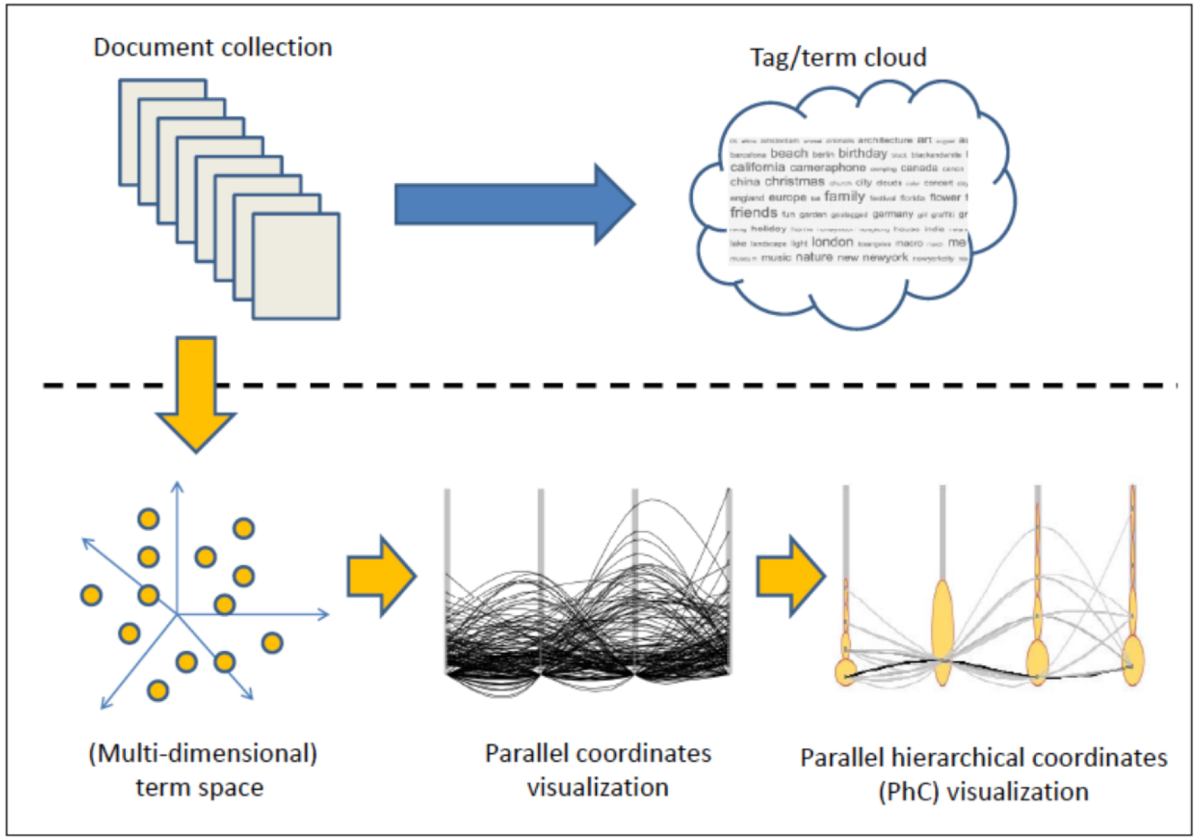


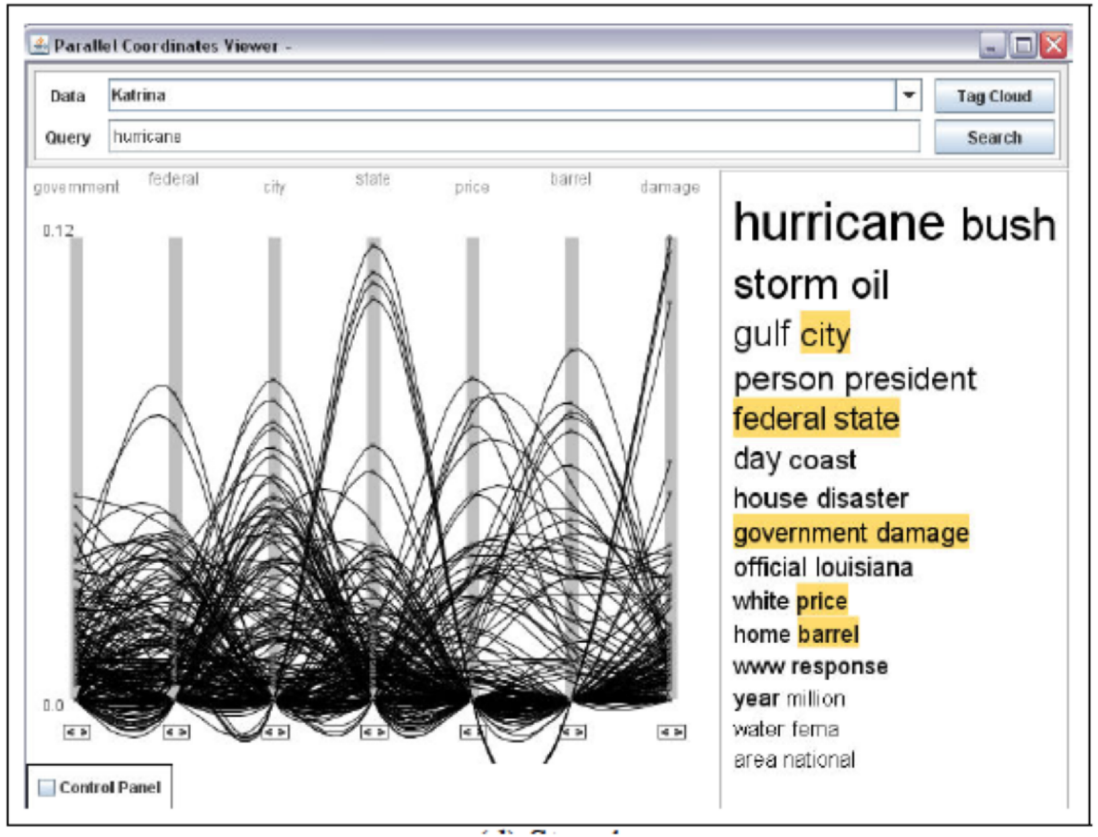
Support Navigation



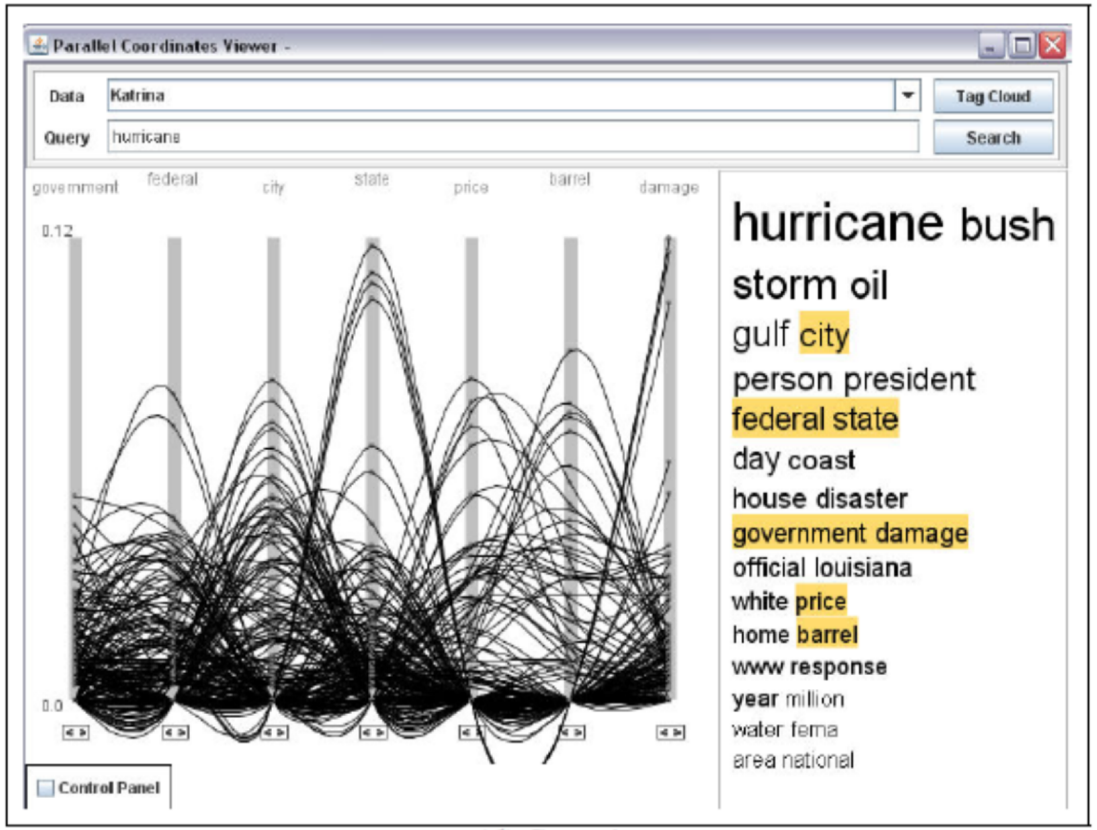


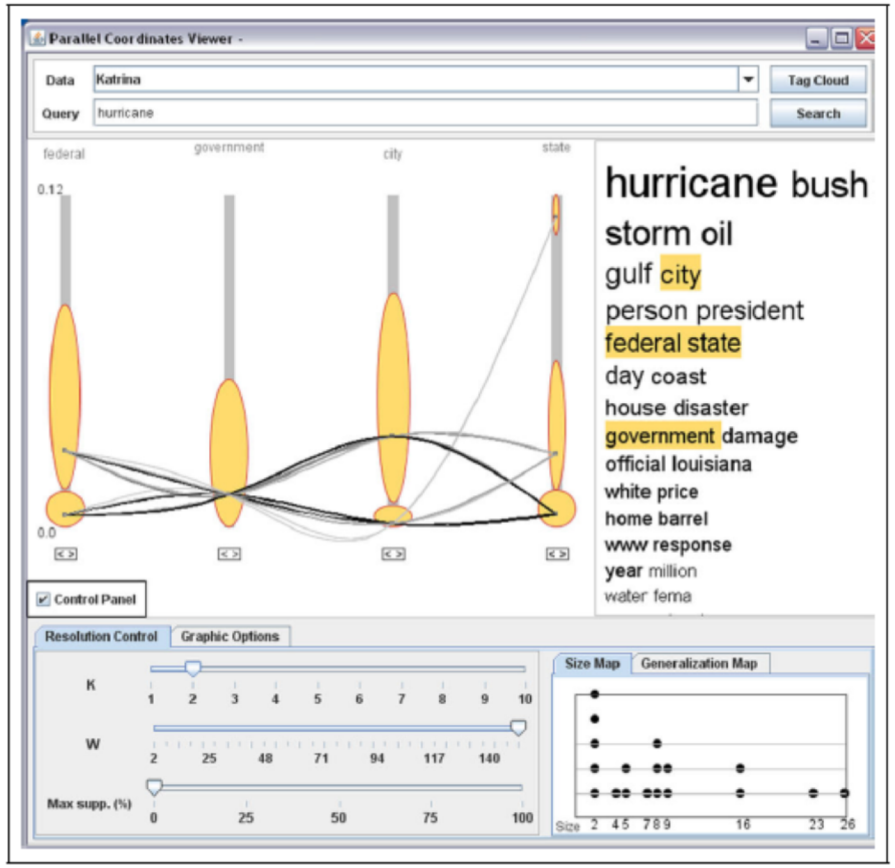
Extract Latent Structure [Alt#2]



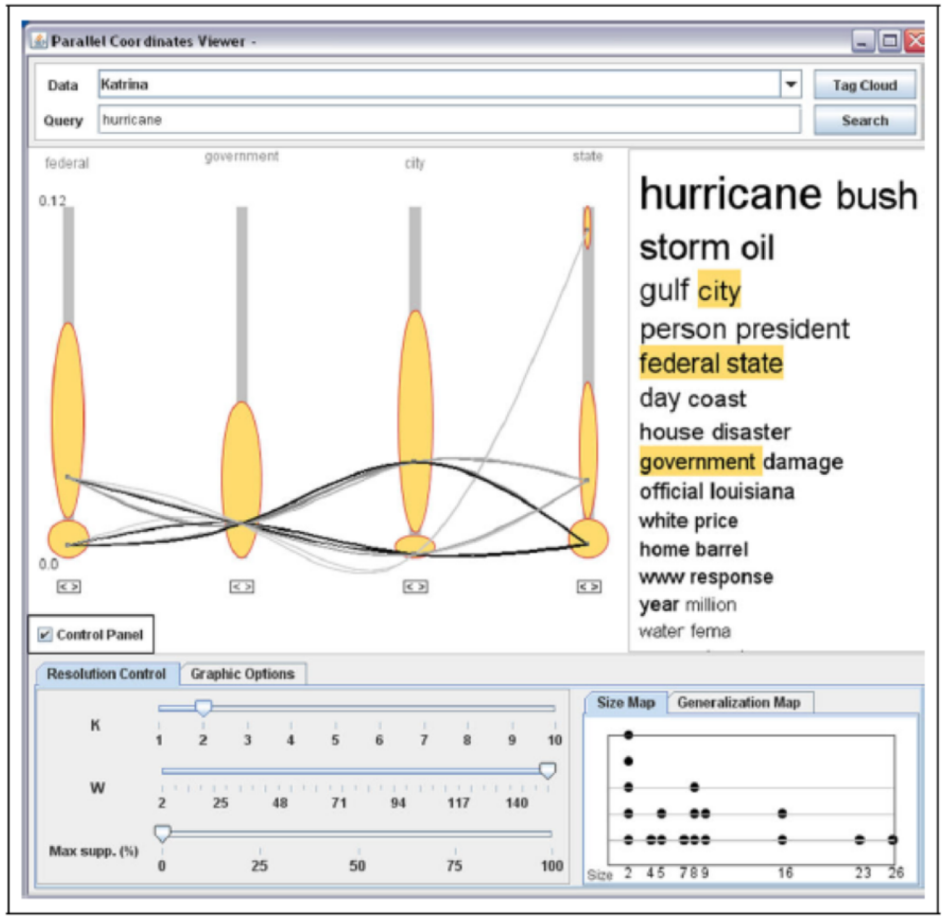


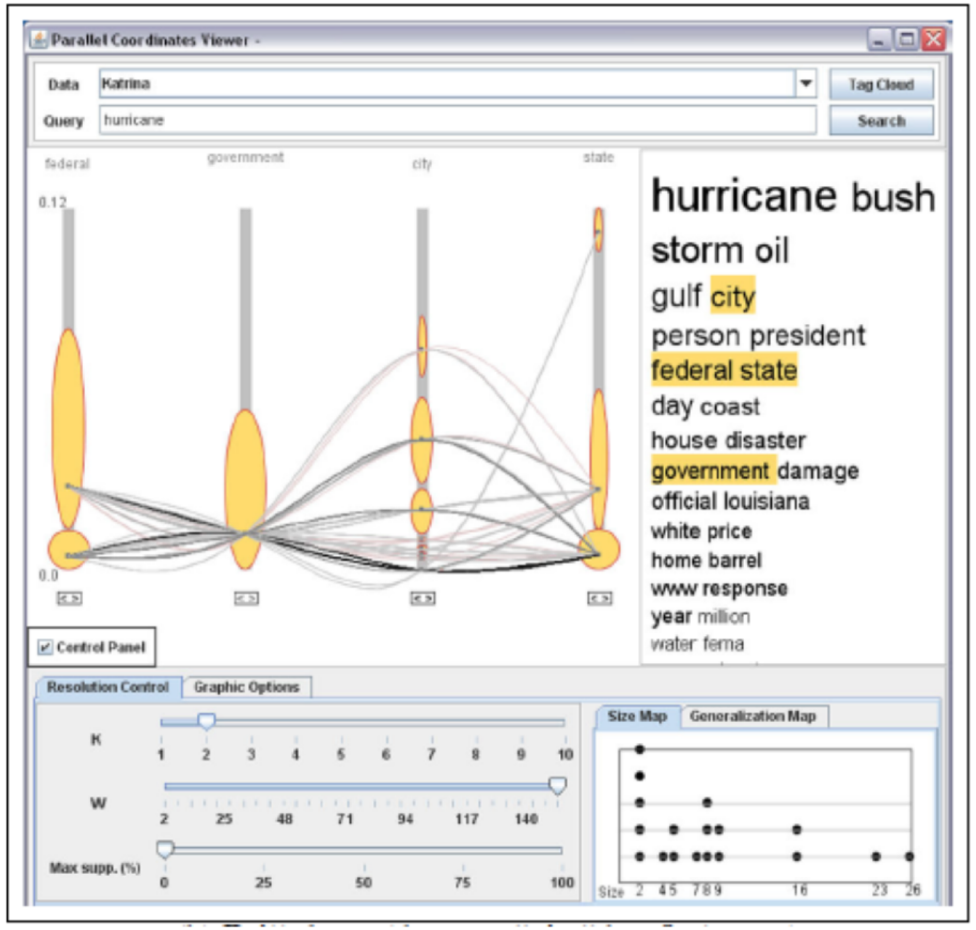
Support Exploration



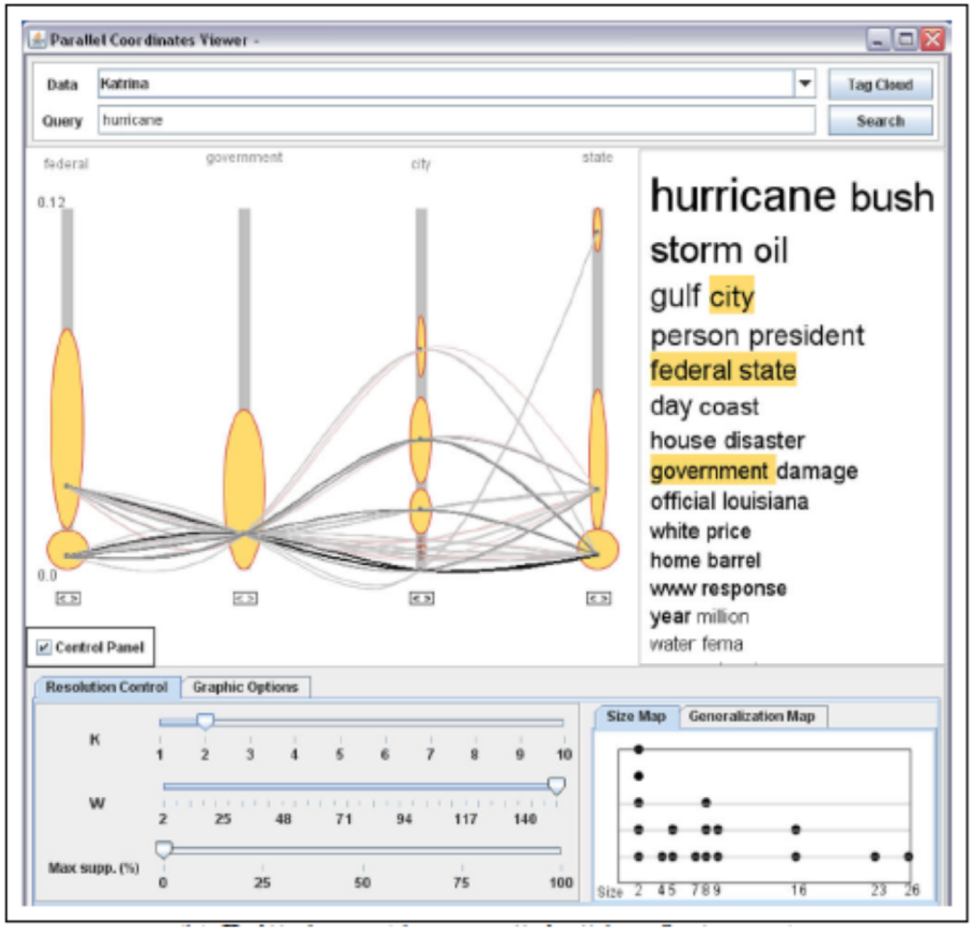


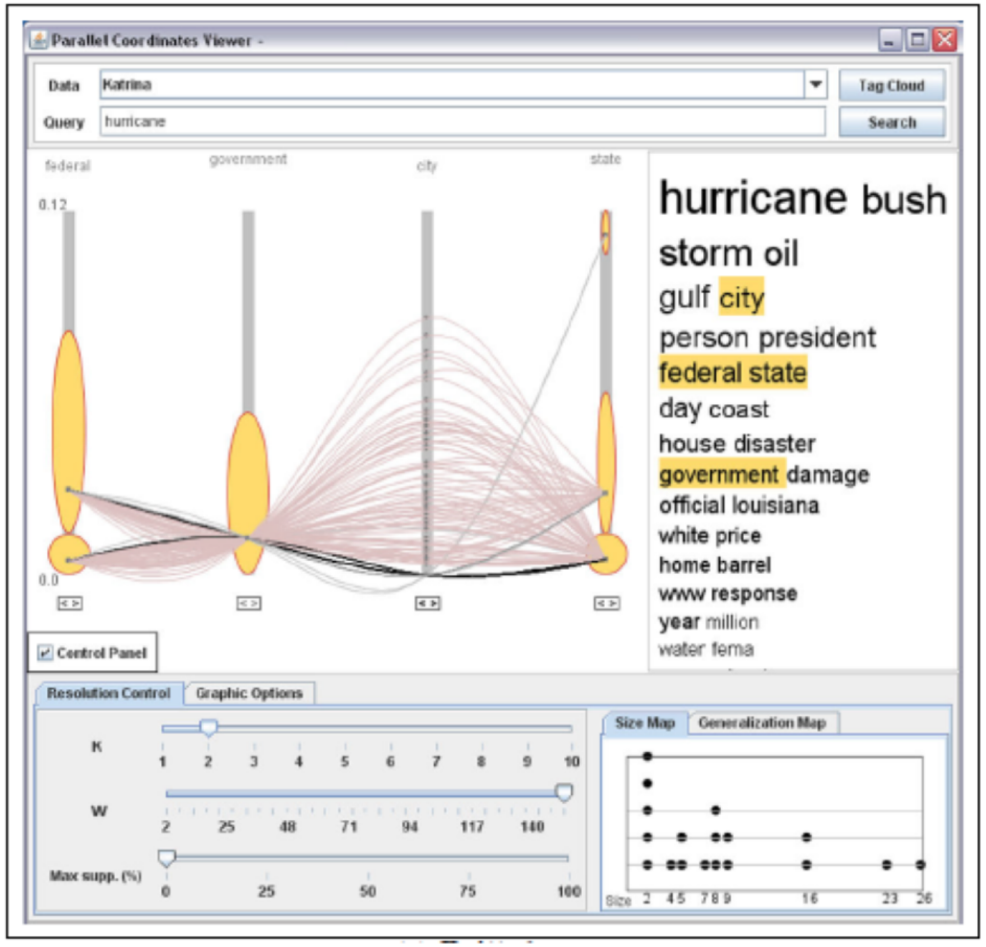
Support Exploration: Drill-Down



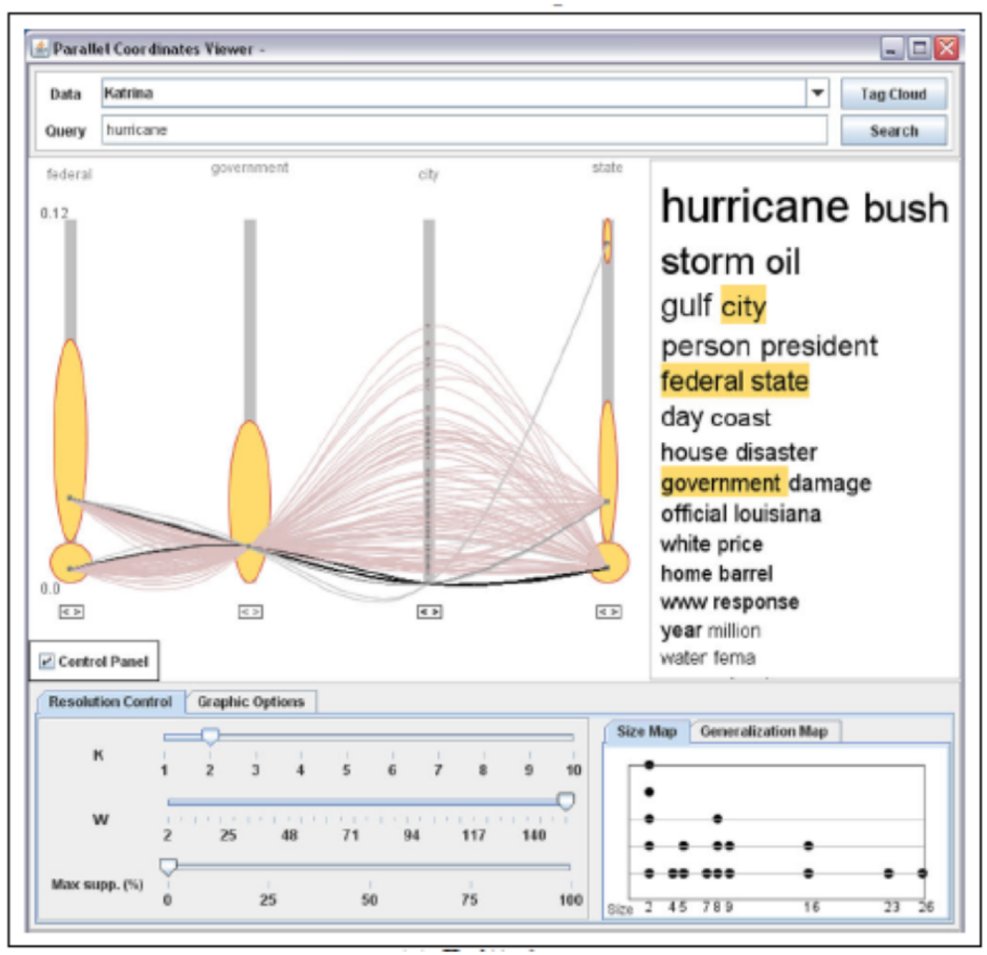


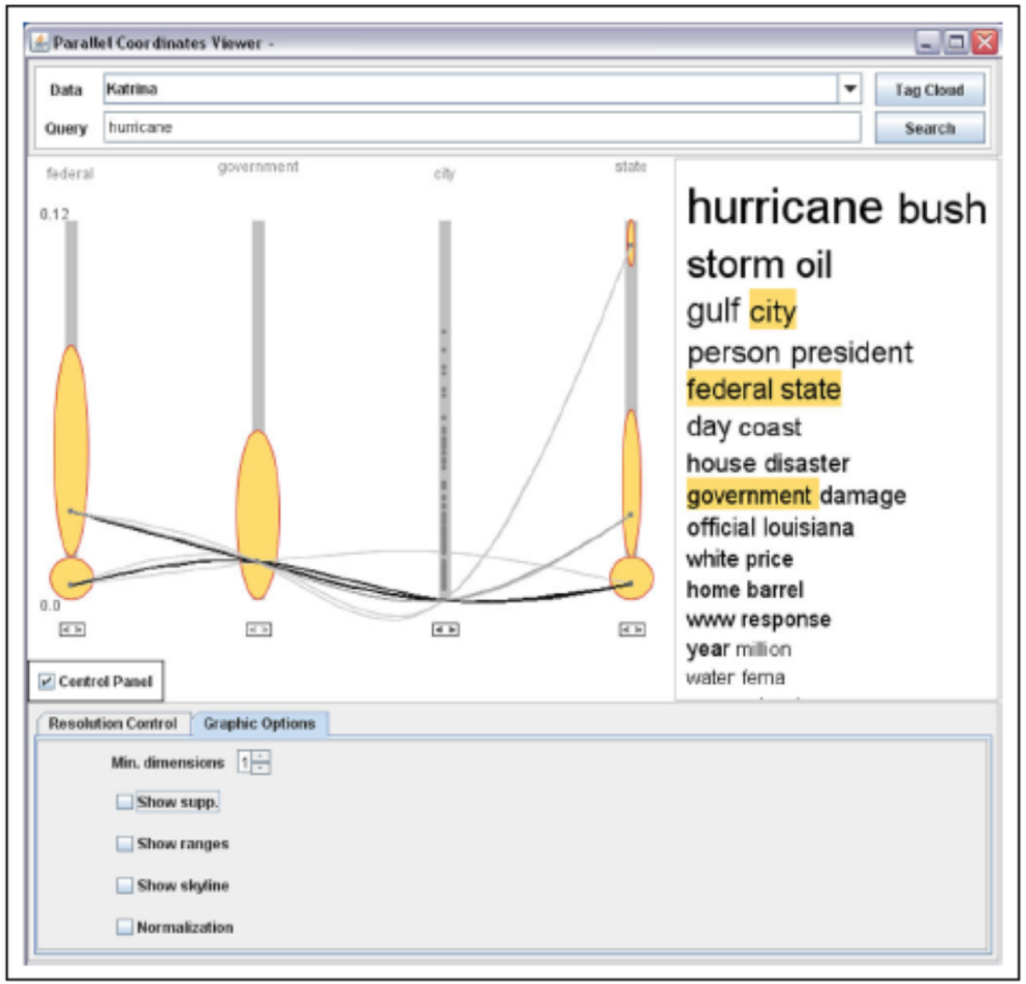
Support Exploration: Drill-Down



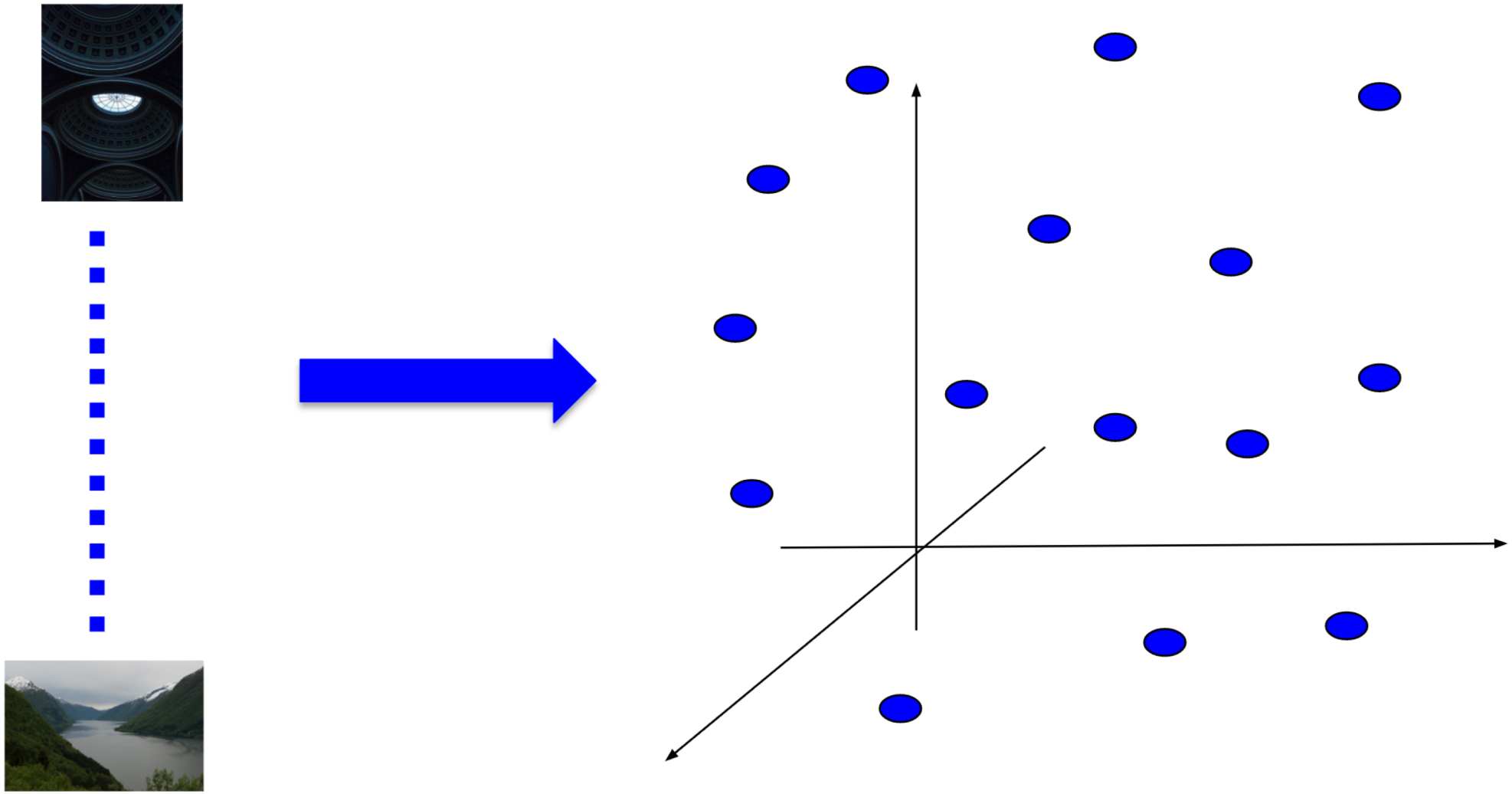


Support Exploration: Eliminate Outliers

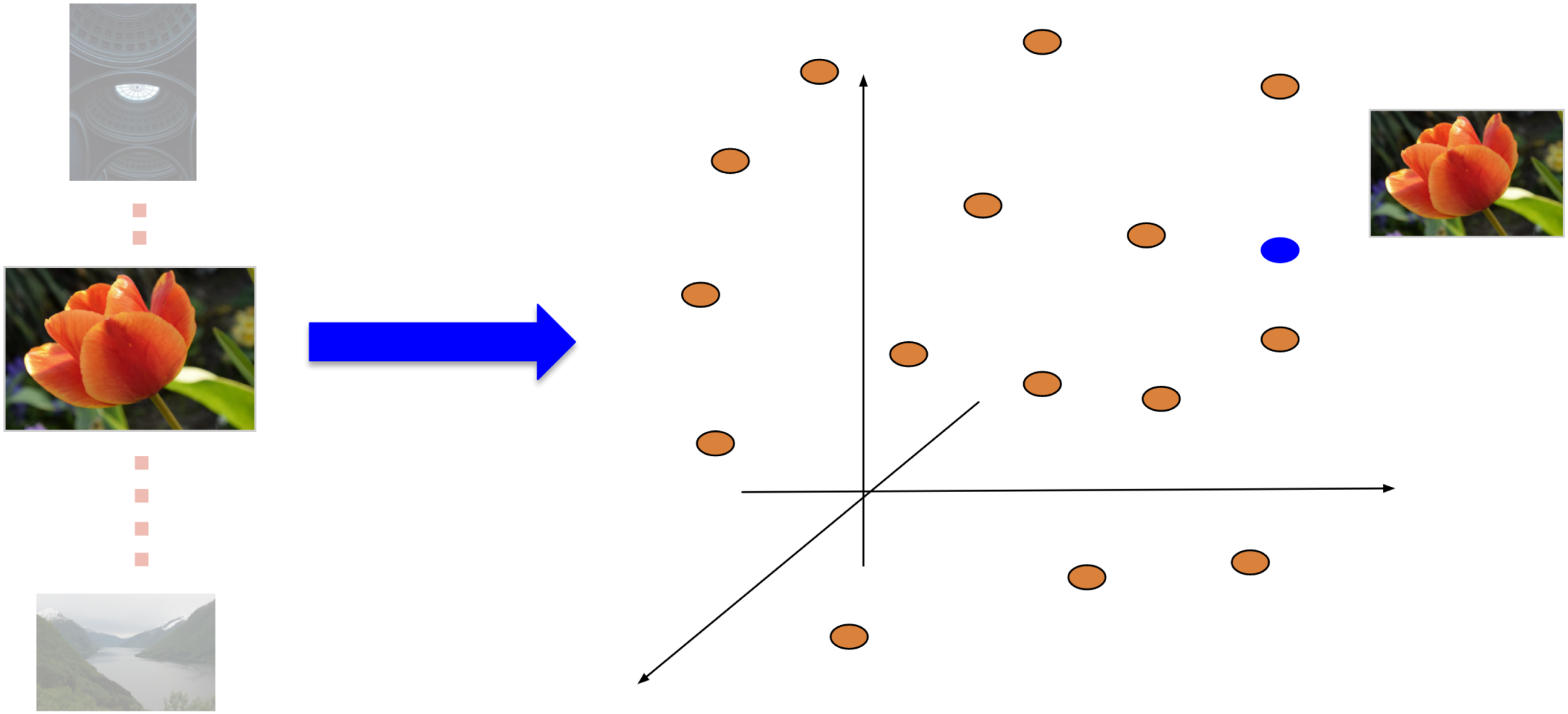




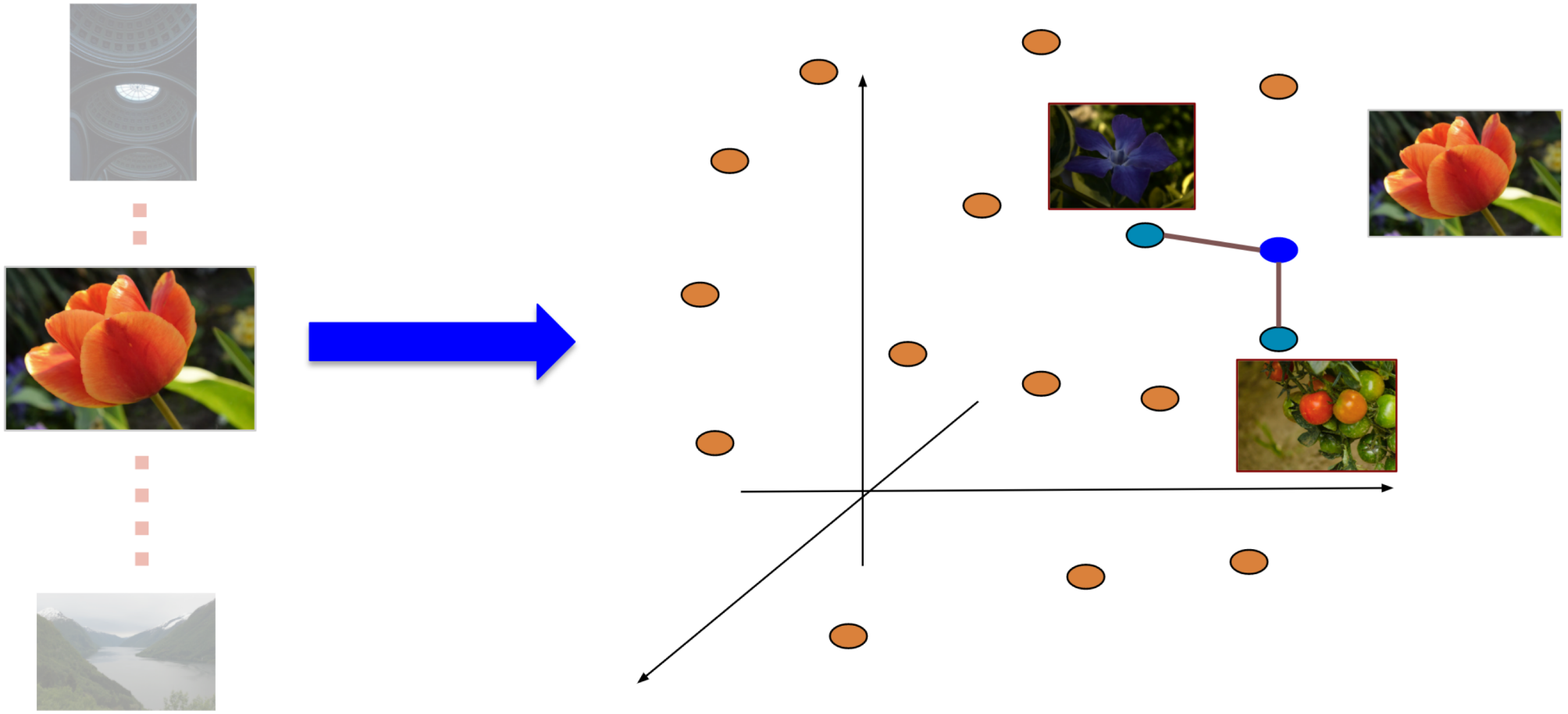
Top-K Search



Top-K Search

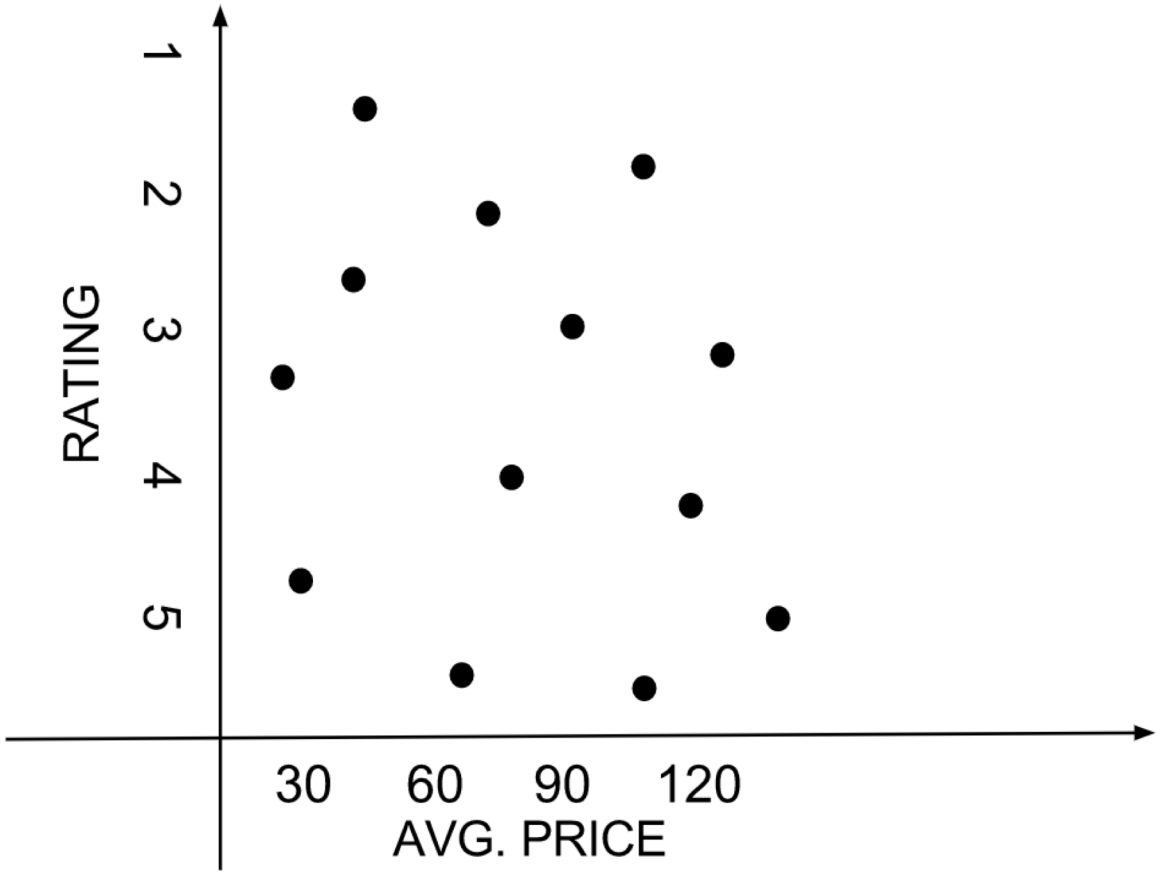


Nearest Neighbor Search



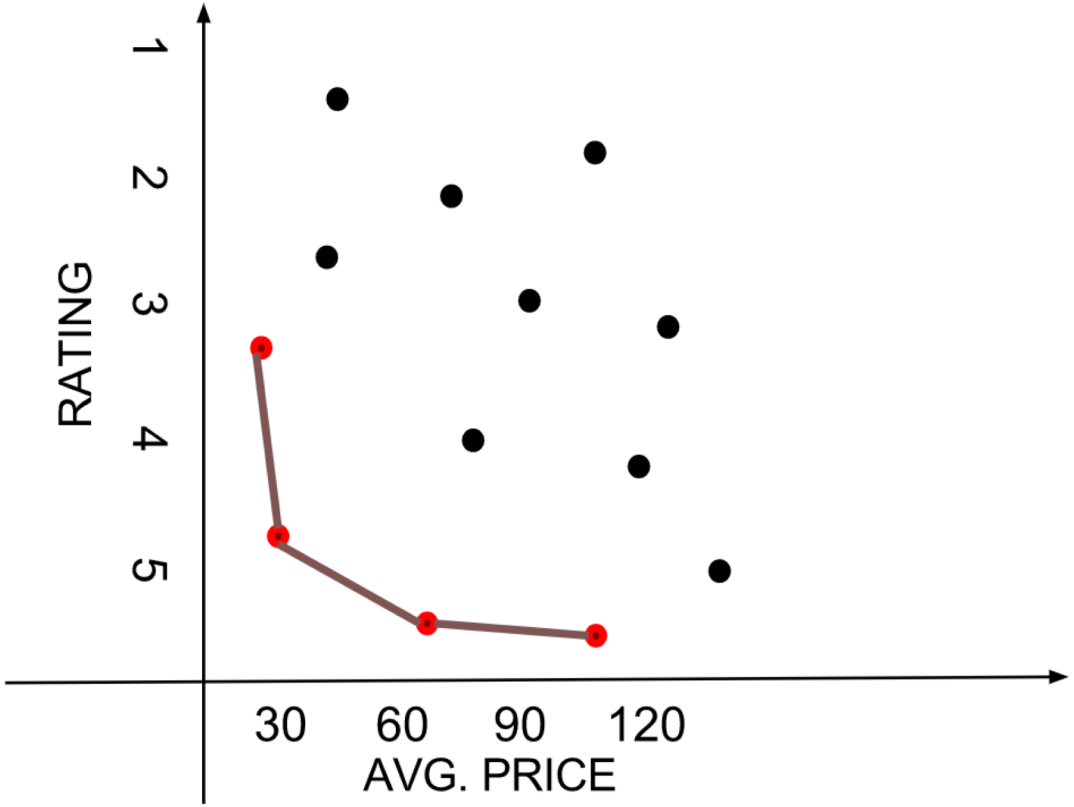
Skylines

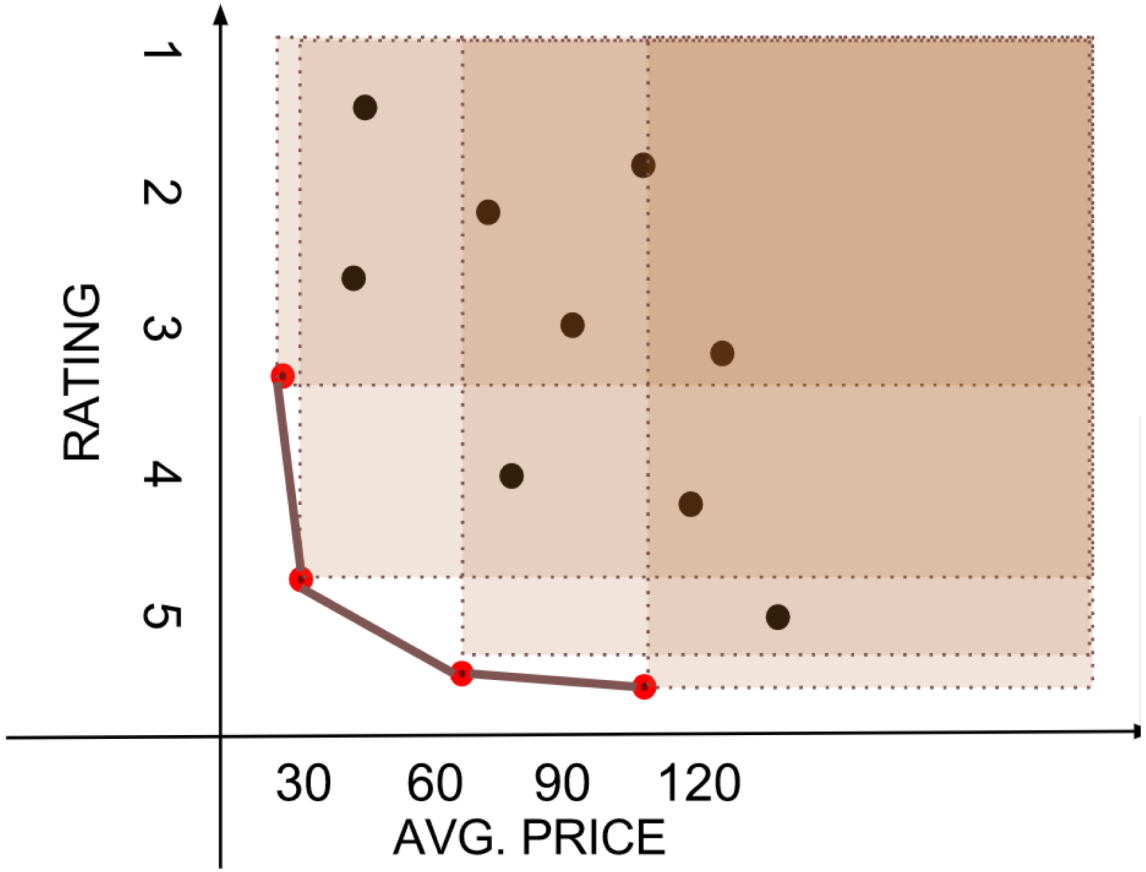
- Known as the Maximum Vector Problem [Kung75]
- Coined as “Skyline” in [Börzsönyi01]



Skylines

- The higher the rating, the better
- The cheaper the price the better
- Which restaurants to consider?



Skylines

- Objects in the “skyline” are not dominated by any other objects in the database