Finding Haystacks with Needles: Ranked Search for Data Using Geospatial and Temporal Characteristics

V.M. Megler
David Maier
Portland State University

Acknowledgements: This work is supported by NSF award OCE-0424602. We thank the staff of the Center for Coastal Margin Observation and Prediction for their support. We also thank the students and professionals who were willing to take part in the user study.

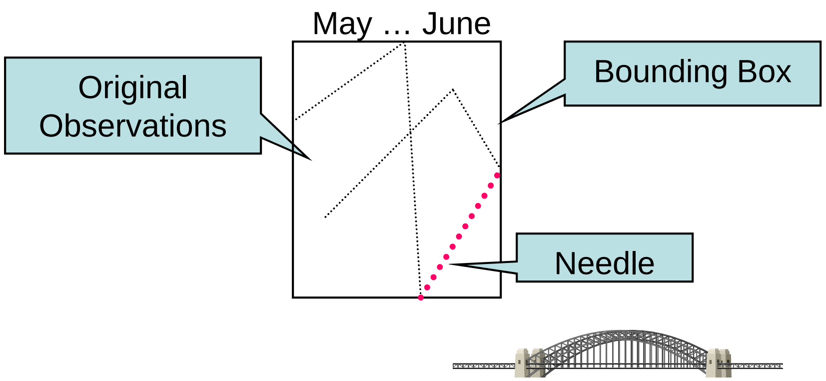
Haystacks

- Many environmental sensors deployed in last decade
- Each sensor collects environmental observations
  - Sometimes many per second
- Each observation has:
  - a time;
  - a location;
  - observed variables
- Observational data stored in many formats, many datasets

Needles

- Scientists at CMOP name "finding data relevant to their research" as one of their biggest problems[2]

- Example query:
  - "Any observations near the Astoria bridge in June 2009"

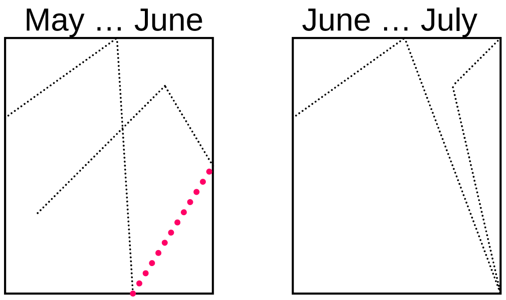


2. Center for Coastal Margin Observation and Prediction RIG Meeting, July 15 2010

Problem: Finding Haystacks that Contain Needles

- Problem: Which datasets contain relevant data?
  - Many scientific datasets have no metadata
  - Many scientific datasets not indexed

- Potential solution: extract simple dataset bounds, perform Boolean search
  - But: many false positives



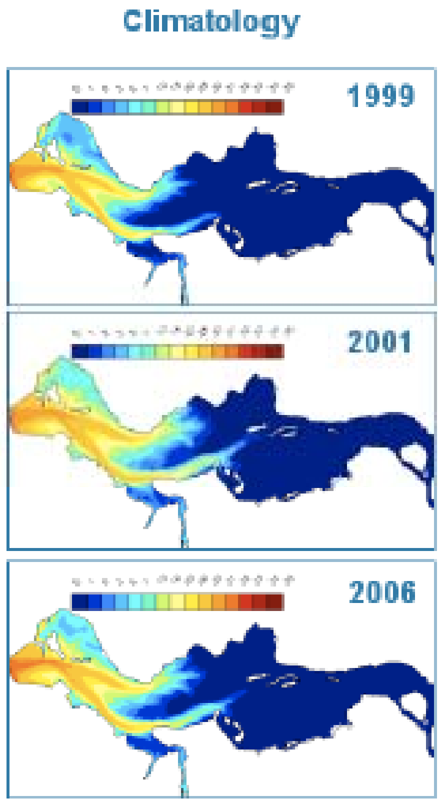
Our Approach:

1. Create hierarchical metadata to represent dataset contents
2. Query over metadata
3. Rank query results

Current Approaches / Related Work (1)

- Search via data visualization
  - Given a specific dataset and data ranges, display the (large amount of) data
  - Most common approach so far

- But: How does the scientist identify relevant datasets and ranges for visualization?



Example of visualization approach
[Howe et al. 2009]

Current Approaches / Related Work (2)

- Metadata search
  - Text search of manually-added metadata
    - E.g. “Salinity, Columbia River”
  - Boolean search on time and location (rare)
    - Some advanced geoportals provide spatial tests:
      - E.g. dataset *intersects* or *completely contains* query area
- But:
  - Boolean search: No matches: no results (1)
  - Search results not ranked (2)

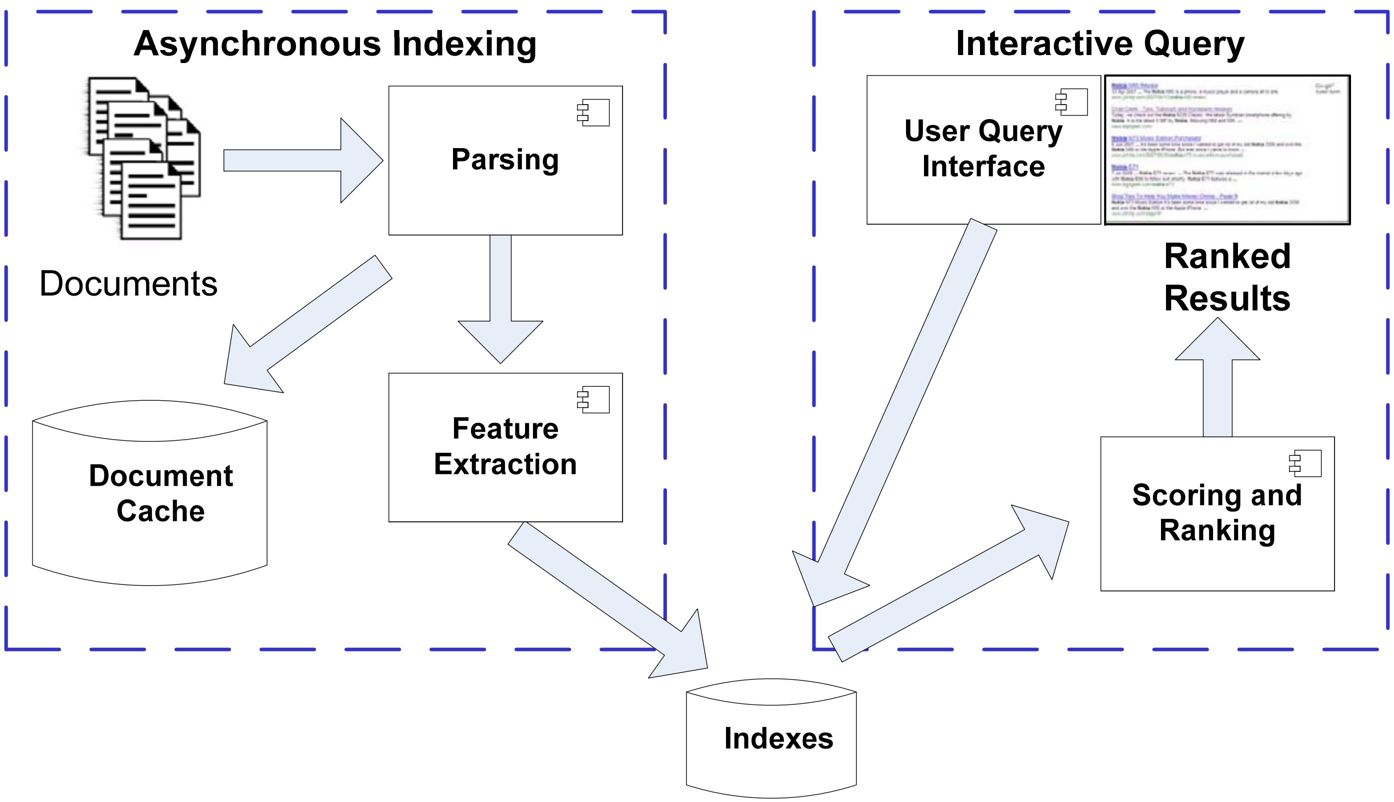
(1)

(2)

There were 3840 results returned

Current Approaches / Related Work (3)

- In Information Retrieval:
  - Ranked retrieval of unstructured text documents



- But text retrieval techniques not suited to searching the contents of scientific datasets

Research Questions

- ? How can we rank datasets?
  - ? Does the ranking approach resonate with users?
- ? What features should we extract from scientific datasets …
- ? … that would allow us to perform real-time search over the extracted features?

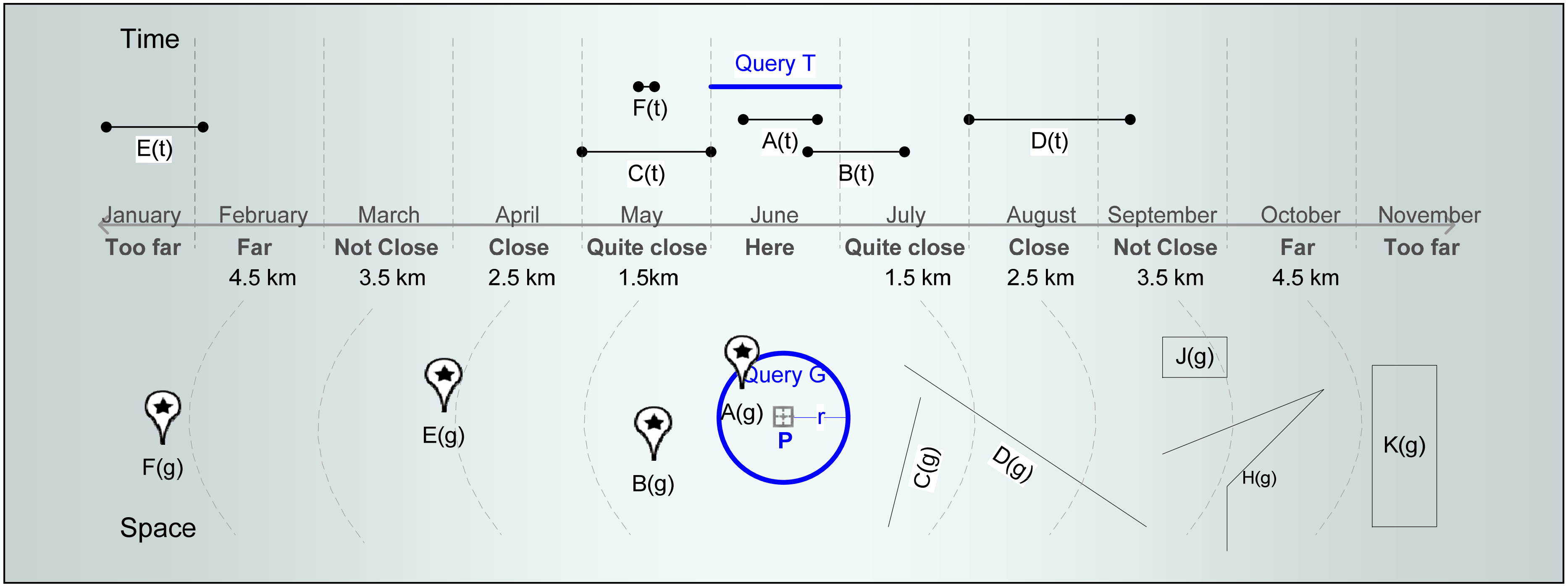
Spatial and temporal features selected for initial case study

Research Contributions

- Proposed a mental model of how scientists perceive dataset similarity for space and time characteristics
  - Tested mental model in a user study
- Developed hierarchical metadata to represent dataset contents
  - Extracting features at multiple granularities
- Developed a prototype query engine with real-time response

Space-Time Ranking: Mental Model

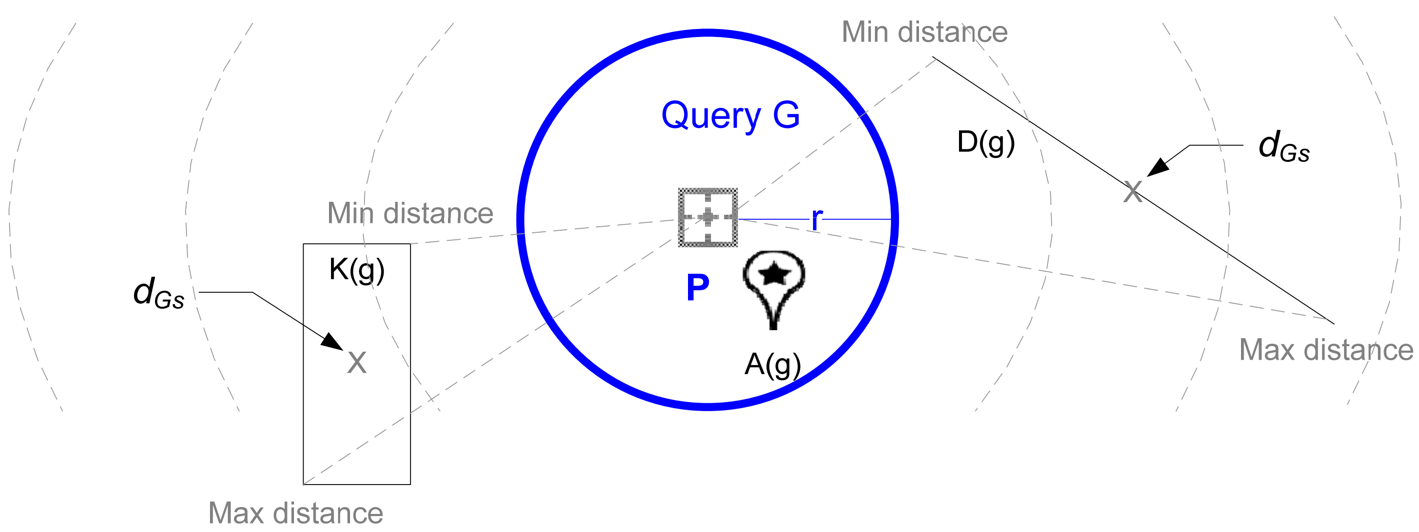
- Example Query: "Observations within ½ km of point 'P', in June 2009"
- Each dataset A, B, … represented by its time extent A(t), B(t), … and its geospatial extent A(g), B(g), …



- Relative "weight" of space to time given by the "range" of each query term

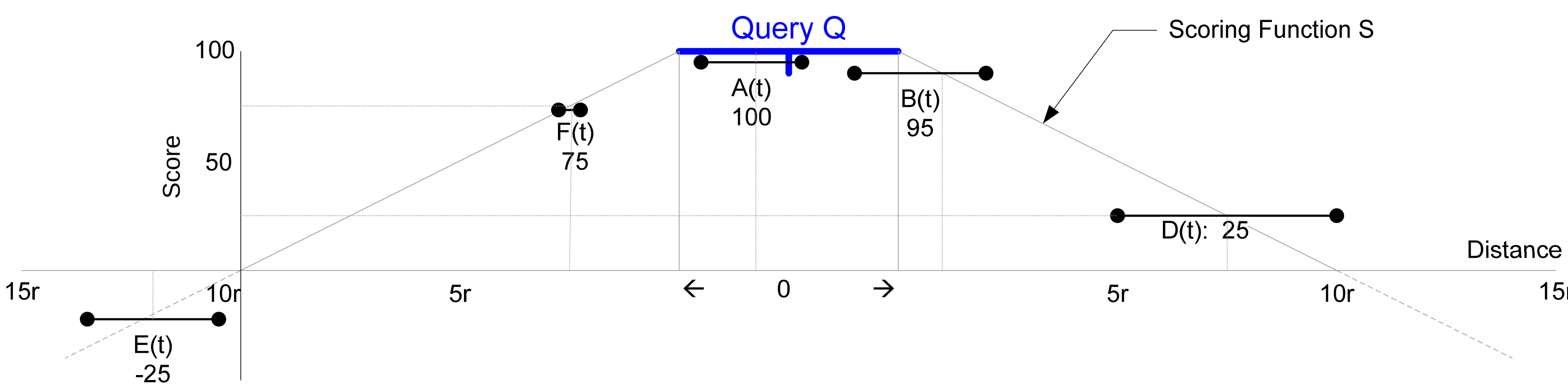
Scoring Datasets (1)

- Score each dataset using formulae that quantify the model
- Given a geospatial query *G*, calculate spatial-relevance score d_{Gs} for dataset *d*
- Spatial relevance is approximated by:
  - ½ (min distance + max distance) / radius
  - Apply scoring function to the result



Scoring Datasets (2)

- Given a time query *T*, calculate a time-relevance score d_{Ts} for dataset *d*



- Calculated scores can range from 100 for an exact match to query terms to negative numbers for datasets “too far” from query

Ranking Datasets

- Overall relevance score d_{score} for each dataset d is composed using the geospatial and temporal scores:

$$d_{score} = (d_{Gs} + d_{Ts}) / 2$$

- Datasets are then ranked by decreasing relevance score.

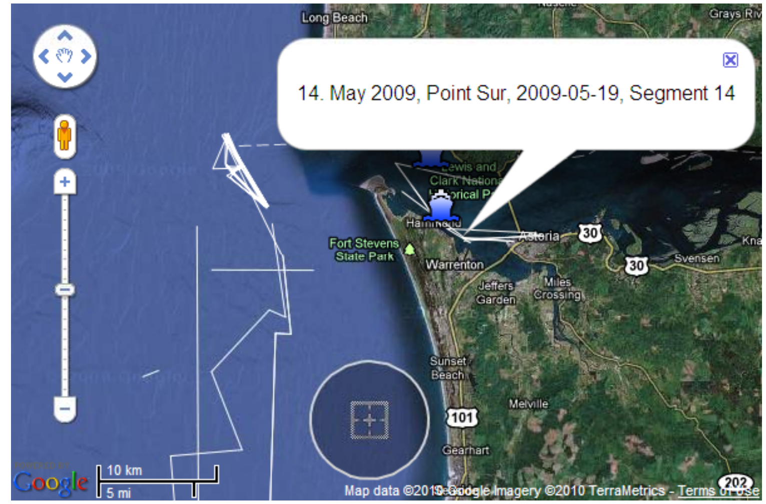
Ranking

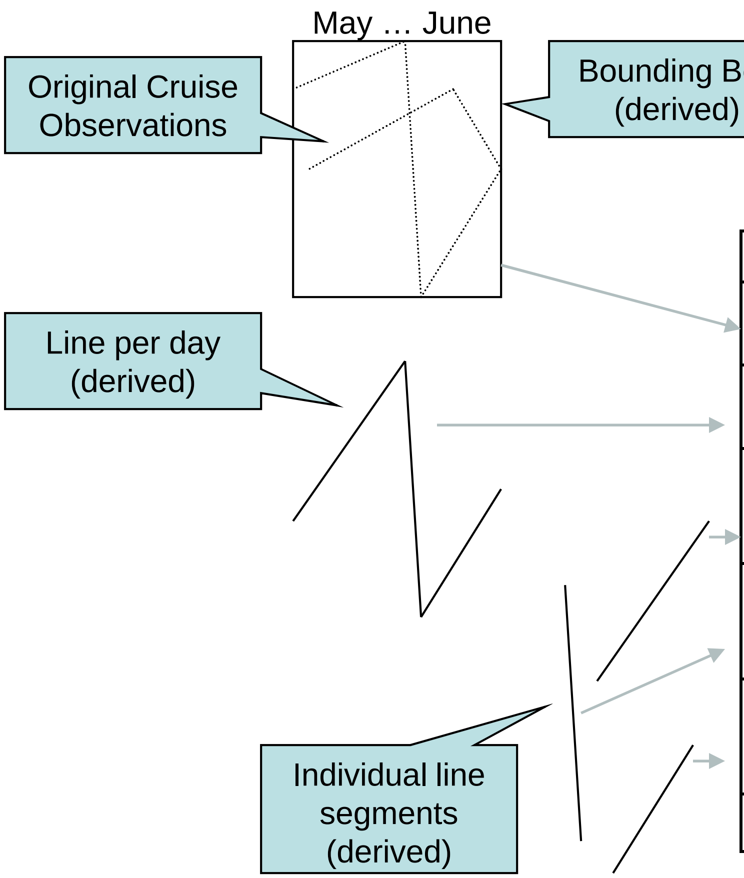
- Tested relevance ranking with a user study:
  - Proposed relevance measure appears to approximate user expectations
  - Relevance-measure "tuning" may further improve match with user expectations
    - "Closest edge" has more weight than "centroid" or "farthest edge"

- Scoring/ranking approach assumes appropriate indexes over which to operate
  - Query terms should relate to indexed features
  - Features represent metadata used to describe dataset content

Creating Metadata: Extracting Features for Space and Time

- Transform observations into features
  - Extract at multiple granularities
  - Model features as “footprints”
  - E.g.: 1 million observations over 3 weeks





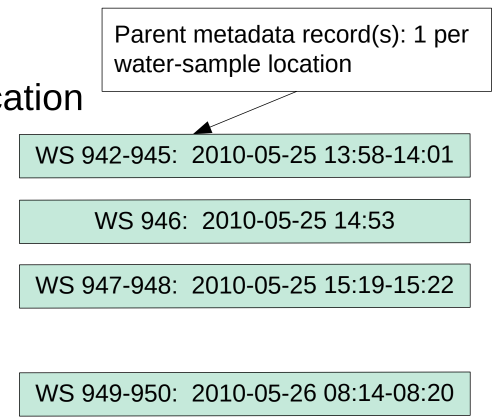
DNH Metadata Table

	Geometry	**Mintime**	**Maxtime**	**Parent**
May 2009, Point Sur	Polygon [bounding box]	5/19/2009	6/10/2009	<null>
May 2009, Point Sur, 2009-05-19	Polyline(p1, p2, p3, p4)	5/19/2009, 00:00	5/19/2009, 23:59	May 2009, Point Sur
May 2009, Point Sur, 2009-05-19, Segment 1	Line(p1, p2)	5/19/2009, 00:00	5/19/2009, 06:14	May 2009, Point Sur, 2009-05-19
May 2009, Point Sur, 2009-05-19, Segment 2	Line(p2, p3)	5/19/2009, 06:15	5/19/2009, 14:23	May 2009, Point Sur, 2009-05-19
May 2009, Point Sur, 2009-05-19, Segment 3	Line(p3, p4)	5/19/2009, 14:24	5/19/2009, 15:01	May 2009, Point Sur, 2009-05-19
….				

Metadata: Adaptive Hierarchy

Water samples:

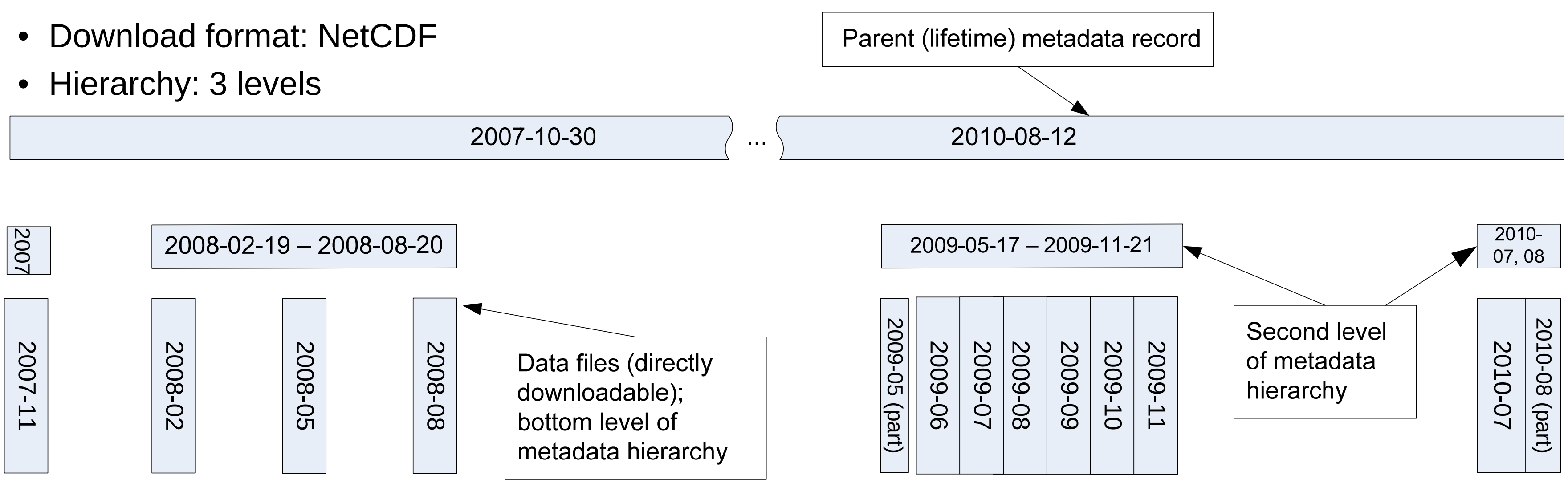
- 1-3 observations per location
- Time: minutes
- Download format: CSV
- Hierarchy: 1 level



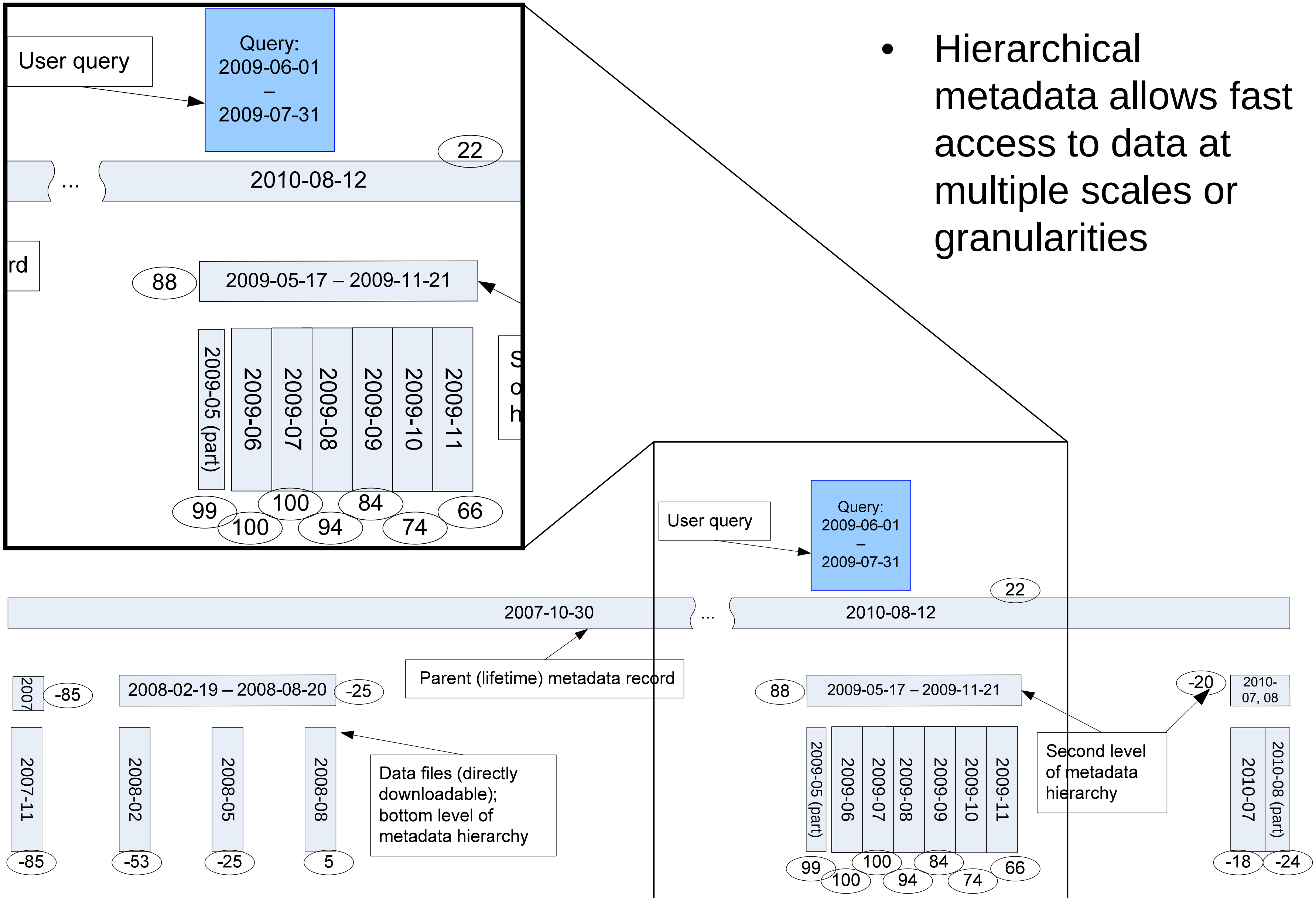
- Multiple depths of hierarchy are accommodated simultaneously
- Curation decision(s) made once per kind of data/dataset

Fixed Stations:

- 1 location
- Time: months-decades
- Observations: millions
- Download format: NetCDF
- Hierarchy: 3 levels

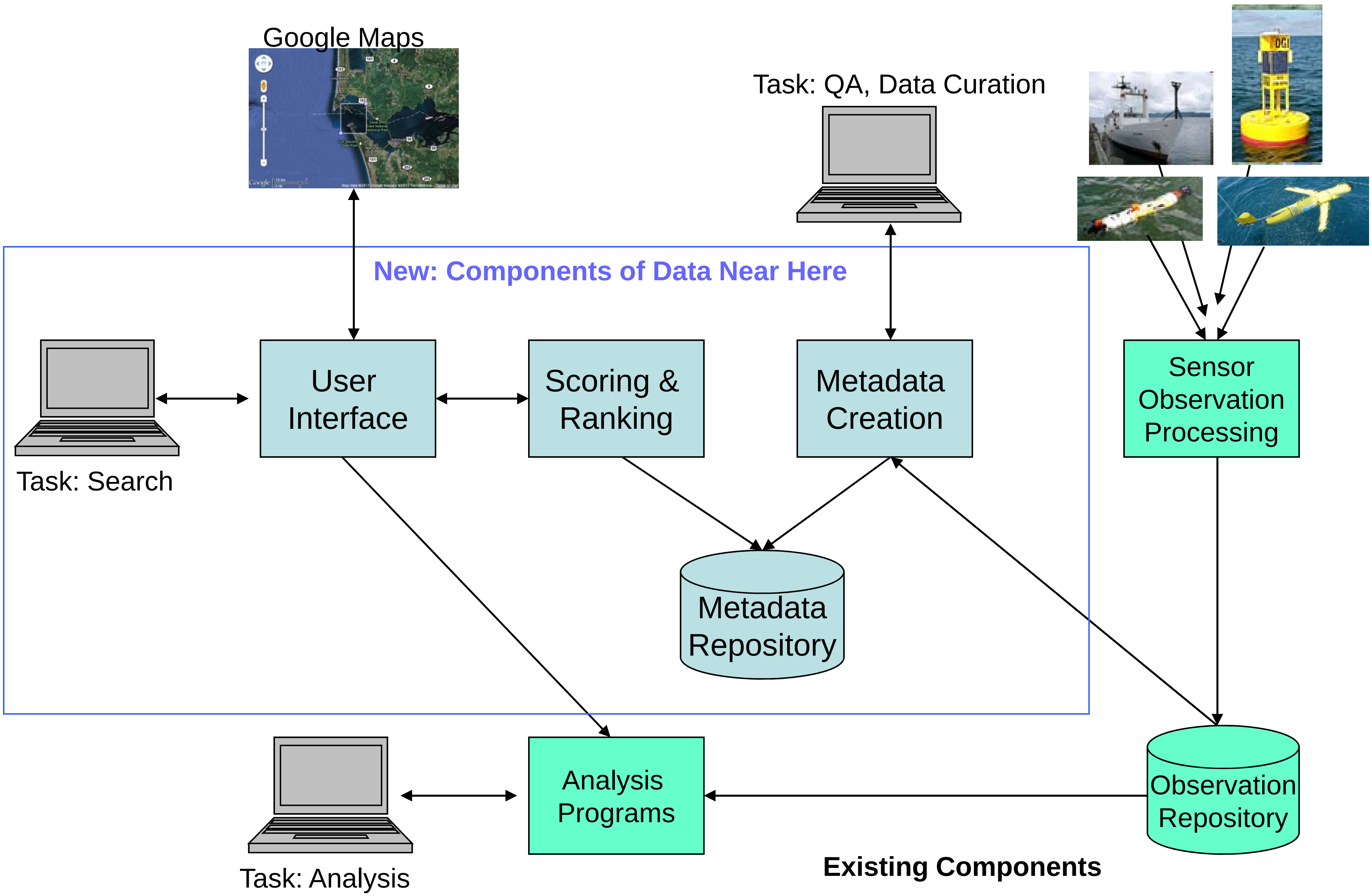


Scoring using Hierarchical Metadata

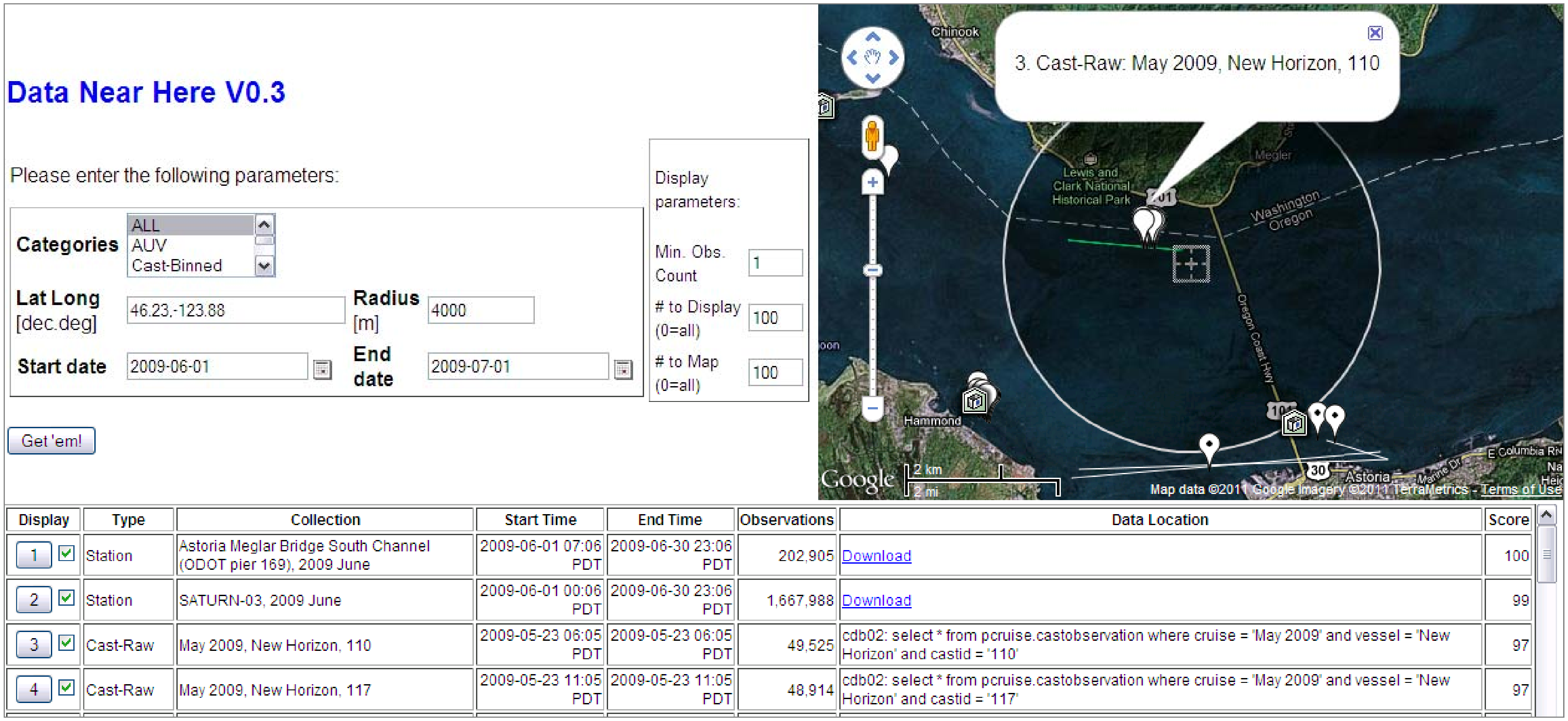


- Hierarchical metadata allows fast access to data at multiple scales or granularities

System Components



The Prototype: “Data Near Here”



- ✓ Extracted metadata for ¼ billion observations → 15,500 metadata records
- ✓ Developed an interactive user interface: Demo
  - ✓ Accepts spatial and temporal query terms
  - ✓ Ranks datasets by decreasing score
  - ✓ Provides real-time response

Conclusion

Our research demonstrates methods for:

- ✓ Ranking scientific datasets in response to a spatio-temporal query
- ✓ Automatically extracting hierarchical metadata from scientific datasets …
- ✓ … and searching over the extracted features
- ✓ Providing real-time response times for queries over ¼ billion observations in a multi-terabyte data repository

Current Research

- Evaluation of metadata scalability
- Add elevation / depth: 4-dimensional search
  - 2+1+1 versus 3+1
- Add additional search criteria:
  - Observational variables
  - … “with oxygen below 3 mg/liter, where Myrionecta Rubra are present”