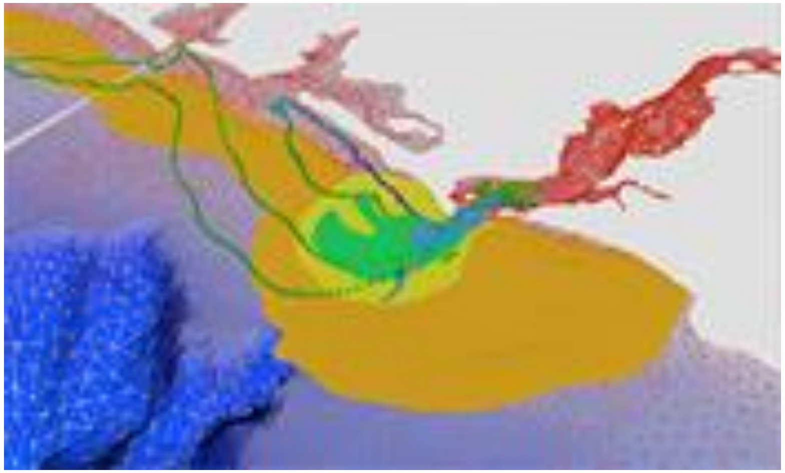
Searching for Data Near Here:
Ranked Similarity Search over Scientific Big Data



CMOP: “Virtual Columbia River”

Veronika Megler
PhD Candidate
Portland State University

Advised by: David Maier

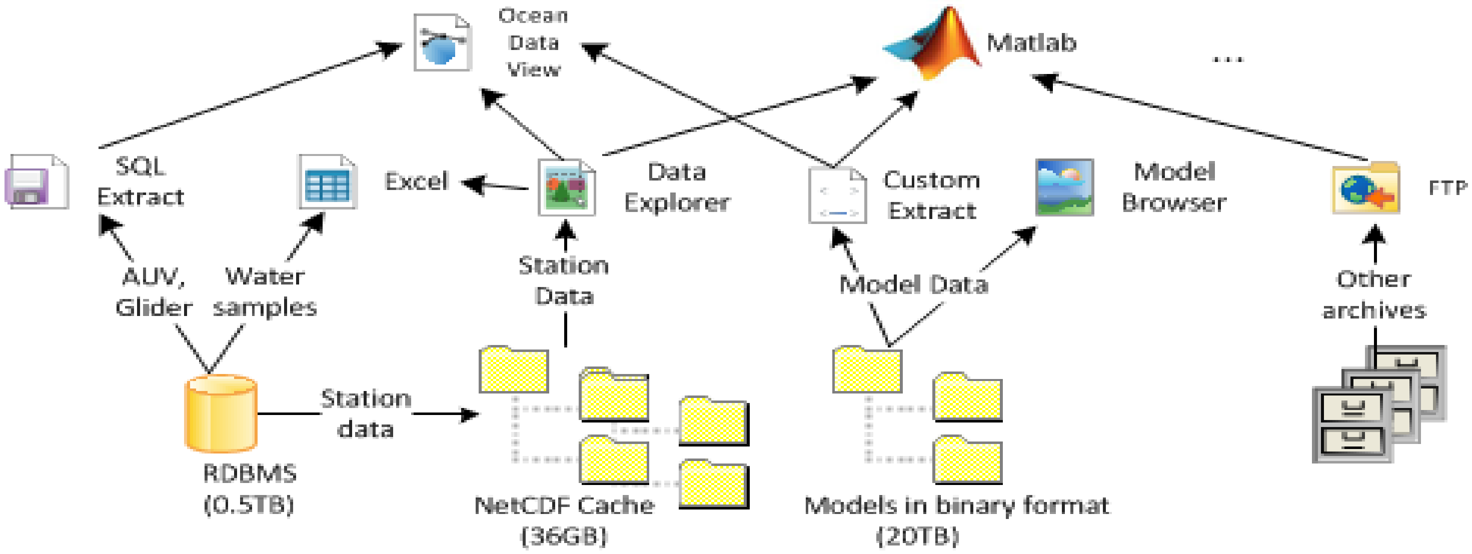
With thanks to the scientists at Center for Coastal Margin Observation and Prediction (CMOP). This work is supported by NSF award OCE-0424602.

Motivation

Scientists have difficulty finding data relevant to their research questions

- Current approaches time-consuming and error-prone
- Example information need:

  “observations collected near [some lat,long] in mid-2010, with temperature between 5-10C”



Heterogeneity of Data Formats and Data Access Tools in One Scientific Archive

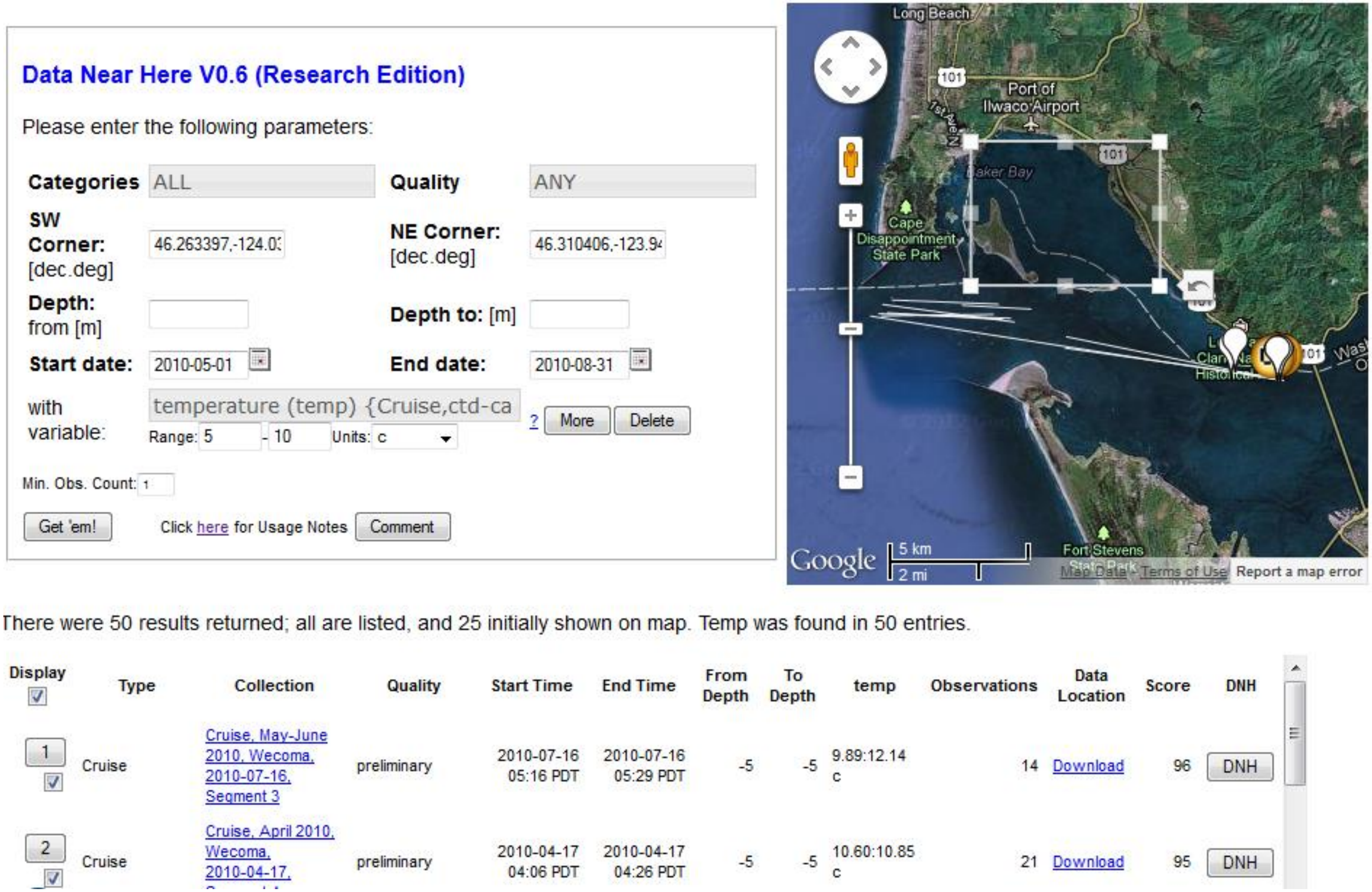
Motivation: Current Approaches to Finding Data

- “Data access” approaches
  - Search via menu selections, portals
  - Each selection individually reviewed (Does not scale)

- Individual visualization of large datasets
  - Does not scale

- Text-based search of metadata
  - Results depend on quality of metadata provided
    Metadata provision still primarily manual
  - Many scientific search criteria are numeric

Our Approach



Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

Categories	ALL	**Quality**	ANY
SW Corner: [dec.deg]	46.263397,-124.0	**NE Corner:** [dec.deg]	46.310406,-123.9
Depth: from [m]		**Depth to:** [m]	
Start date:	2010-05-01	**End date:**	2010-08-31
with variable:	temperature (temp) {Cruise,ctd-ca Range: 5 - 10 Units: c	? More Delete	

Min. Obs. Count: 1

Get 'em!   Click here for Usage Notes   Comment

There were 50 results returned; all are listed, and 25 initially shown on map. Temp was found in 50 entries.

Display ☑	Type	Collection	Quality	Start Time	End Time	From Depth	To Depth	temp	Observations	Data Location	Score	DNH
1 ☑	Cruise	Cruise, May-June 2010, Wecoma, 2010-07-16, Segment 3	preliminary	2010-07-16 05:16 PDT	2010-07-16 05:29 PDT	-5	-5	9.89:12.14 c	14	Download	96	DNH
2 ☑	Cruise	Cruise, April 2010, Wecoma, 2010-04-17,	preliminary	2010-04-17 04:06 PDT	2010-04-17 04:26 PDT	-5	-5	10.60:10.85 c	21	Download	95	DNH

- Apply Information Retrieval techniques to scientific data

My Research

- “The principal contribution … is to define a new problem”[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

- Provided evidence of utility
  - Two user studies
  - “Defined a baseline ranking function against which future developments can be compared” [1]

- (In progress) Evaluate scalability

1. Comment from anonymous reviewer

My Research

- “The principal contribution … is to define a new problem”[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

- Provided evidence of utility
  - Two user studies
  - “Defined a baseline ranking function against which future developments can be compared” [1]

- (In progress) Evaluate scalability

[1] Comment from anonymous reviewer

Applying Information Retrieval Techniques

- Definition:

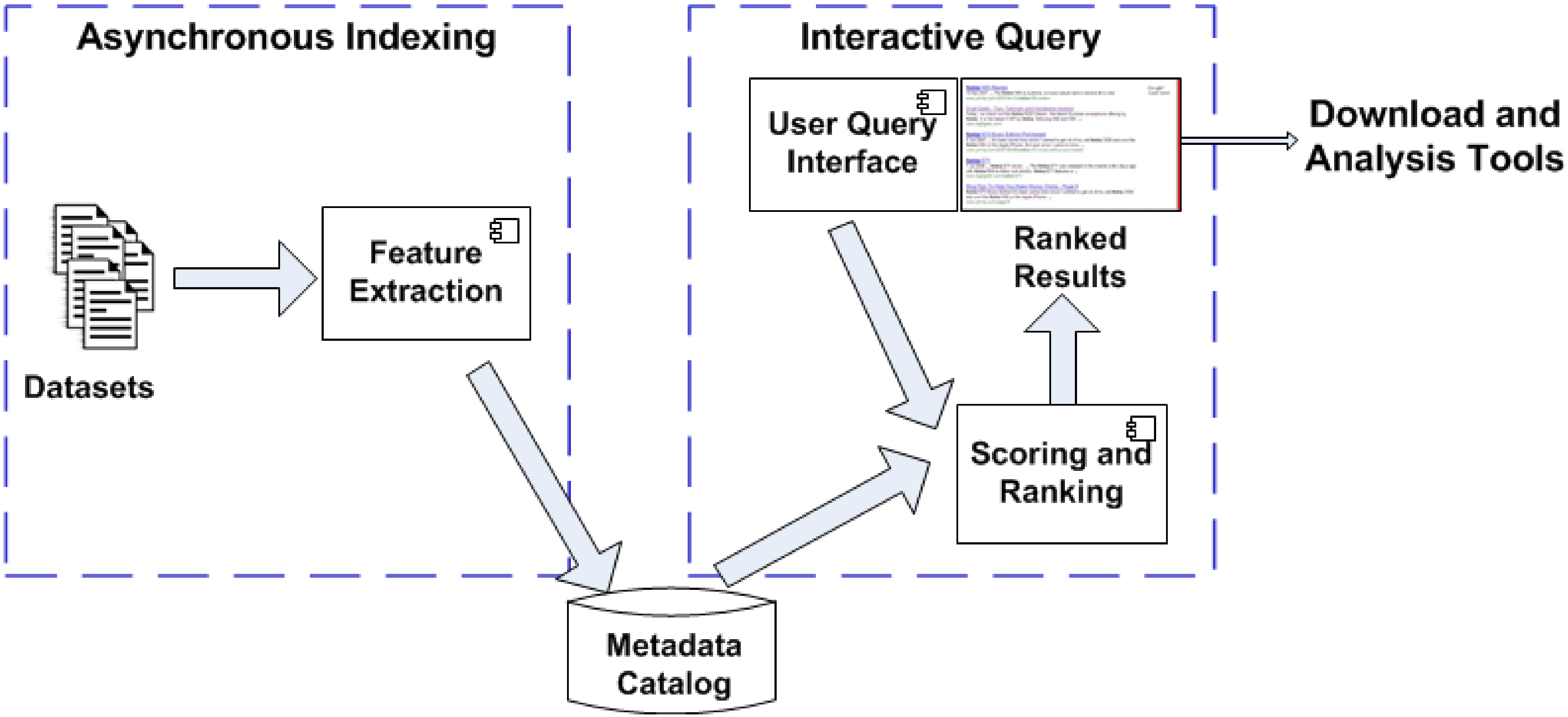
  A dataset is *relevant* if the scientist perceives that it contains data relevant to the scientist's information need.[2]

- Two major approaches to retrieving relevant items:
  - *Boolean retrieval*: only exact matches are returned
  - *Ranked retrieval*:
    - Each item given a *score*: item's relevance to the query
    - Result list *ranked:* from highest to lowest score

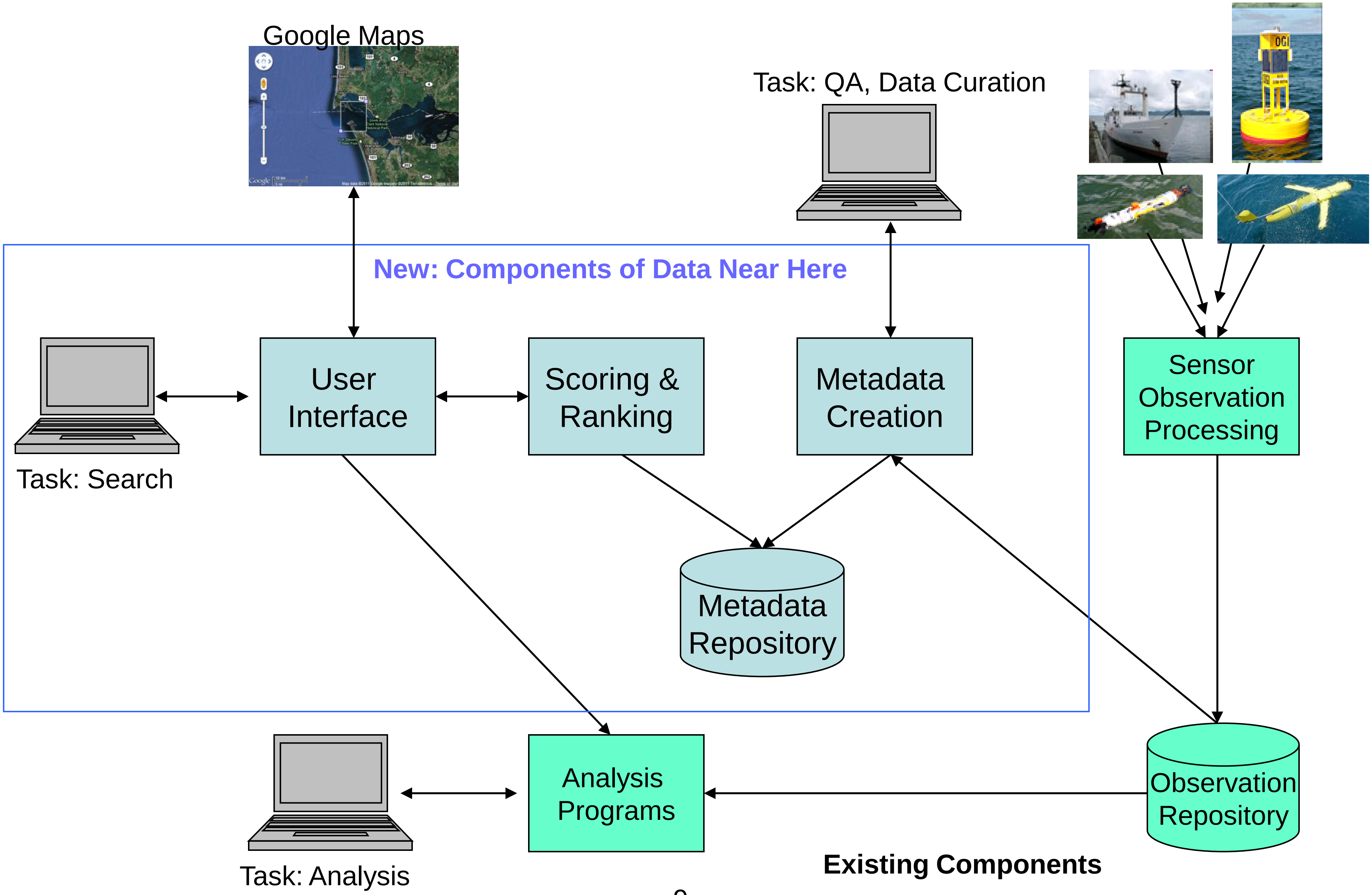
- To apply ranked IR techniques we need:
  1. a method for extracting features from datasets
  2. to express a scientific information need as a set of query conditions
  3. a similarity measure to compare query conditions to the extracted features

[2] Adapted from Manning, C.D. et al.: Introduction to Information Retrieval. Cambridge University Press (2008).

IR Architecture Adapted to Scientific Data Search



System Components



Google Maps

Task: QA, Data Curation

New: Components of Data Near Here

User Interface

Scoring & Ranking

Metadata Creation

Sensor Observation Processing

Task: Search

Metadata Repository

Analysis Programs

Observation Repository

Task: Analysis

Existing Components

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

My Research

- “The principal contribution … is to define a new problem”[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

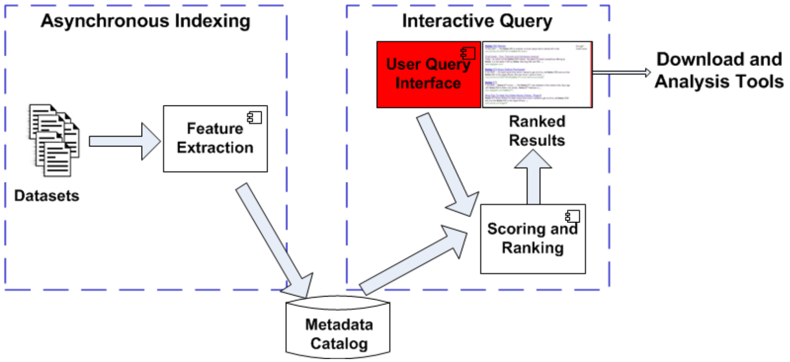
- Provided evidence of utility
  - Two user studies
  - “Defined a baseline ranking function against which future developments can be compared” [1]

- (In progress) Evaluate scalability

[1] Comment from anonymous reviewer

Prototype: “Data Near Here” (DNH)

- Implemented at CMOP
- Search with interactive response times
  - > 1B observations
- Datasets represented by summaries
- Explore, plot or download results



Asynchronous Indexing

Datasets

Feature Extraction

Metadata Catalog

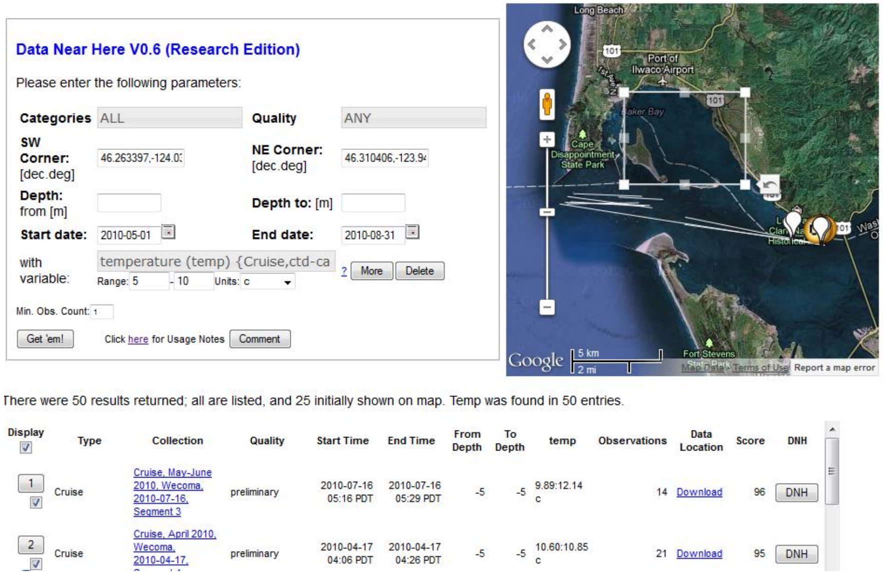
Interactive Query

User Query Interface

Ranked Results

Scoring and Ranking

Download and Analysis Tools



Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

Categories	ALL	**Quality**	ANY
SW Corner: [dec.deg]	46.263397,-124.03	**NE Corner:** [dec.deg]	46.310406,-123.94
Depth: from [m]		**Depth to:** [m]	
Start date:	2010-05-01	**End date:**	2010-08-31
with variable:	temperature (temp) {Cruise,ctd-ca Range: 5 - 10 Units: c		

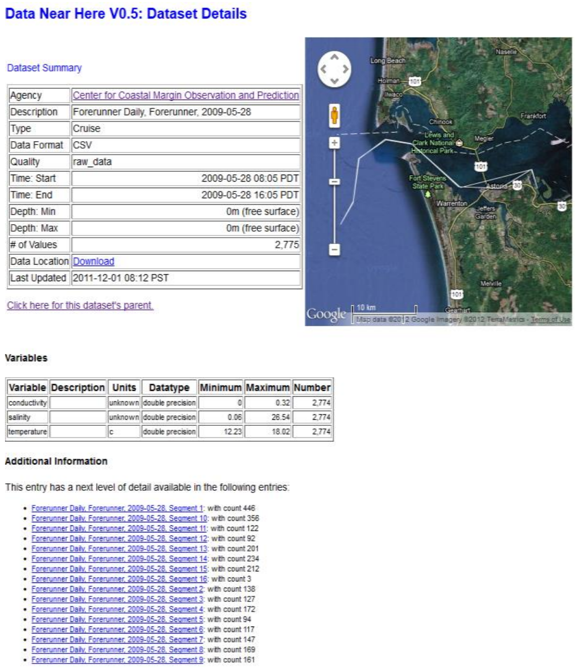
Min. Obs. Count: 1

Click here for Usage Notes

There were 50 results returned; all are listed, and 25 initially shown on map. Temp was found in 50 entries.

Display ☑	Type	Collection	Quality	Start Time	End Time	From Depth	To Depth	temp	Observations	Data Location	Score	DNH
1 ☑	Cruise	Cruise, May-June 2010, Wecoma, 2010-07-16, Segment 3	preliminary	2010-07-16 05:16 PDT	2010-07-16 05:29 PDT	-5	-5	9.89:12.14 c	14	Download	96	DNH
2 ☑	Cruise	Cruise, April 2010, Wecoma, 2010-04-17,	preliminary	2010-04-17 04:06 PDT	2010-04-17 04:26 PDT	-5	-5	10.60:10.85 c	21	Download	95	DNH

User Interface: Search-and-Results Screen



Data Near Here V0.5: Dataset Details

Dataset Summary

Agency	Center for Coastal Margin Observation and Prediction
Description	Forerunner Daily, Forerunner, 2009-05-28
Type	Cruise
Data Format	CSV
Quality	raw_data
Time: Start	2009-05-28 08:05 PDT
Time: End	2009-05-28 16:05 PDT
Depth: Min	0m (free surface)
Depth: Max	0m (free surface)
# of Values	2,775
Data Location	Download
Last Updated	2011-12-01 08:12 PST

Click here for this dataset's parent.

Variables

Variable	Description	Units	Datatype	Minimum	Maximum	Number
conductivity		unknown	double precision	0	0.32	2,774
salinity		unknown	double precision	0.06	26.54	2,774
temperature		c	double precision	12.23	18.02	2,774

Additional Information

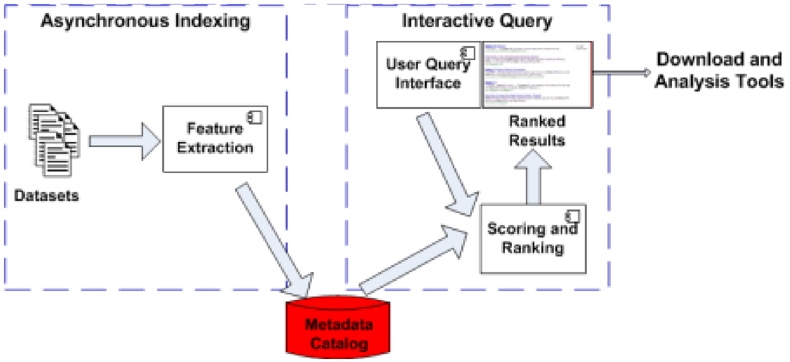
This entry has a next level of detail available in the following entries:

- Forerunner Daily, Forerunner, 2009-05-28, Segment 1: with count 446
- Forerunner Daily, Forerunner, 2009-05-28, Segment 10: with count 356
- Forerunner Daily, Forerunner, 2009-05-28, Segment 11: with count 122
- Forerunner Daily, Forerunner, 2009-05-28, Segment 12: with count 92
- Forerunner Daily, Forerunner, 2009-05-28, Segment 13: with count 201
- Forerunner Daily, Forerunner, 2009-05-28, Segment 14: with count 234
- Forerunner Daily, Forerunner, 2009-05-28, Segment 15: with count 212
- Forerunner Daily, Forerunner, 2009-05-28, Segment 16: with count 3
- Forerunner Daily, Forerunner, 2009-05-28, Segment 2: with count 138
- Forerunner Daily, Forerunner, 2009-05-28, Segment 3: with count 127
- Forerunner Daily, Forerunner, 2009-05-28, Segment 4: with count 172
- Forerunner Daily, Forerunner, 2009-05-28, Segment 5: with count 94
- Forerunner Daily, Forerunner, 2009-05-28, Segment 6: with count 117
- Forerunner Daily, Forerunner, 2009-05-28, Segment 7: with count 147
- Forerunner Daily, Forerunner, 2009-05-28, Segment 8: with count 169
- Forerunner Daily, Forerunner, 2009-05-28, Segment 9: with count 161

User Interface: Dataset Details

Prototype: Feature Extraction

- Features extracted during one-time scan of each dataset
  - Build a “dataset summary”
  - A feature may be: a column and its data range; or, global metadata



- Multiple types of data handled:
  - Single location, single time

    Water samples, “casts”
  - Single location, multi-year

    “Fixed stations”
  - Mobile devices (3D, 4D)

    Cruises, AUV, glider

- Data from other archives added

  No modifications to summary required

- “Available in test/dev”:

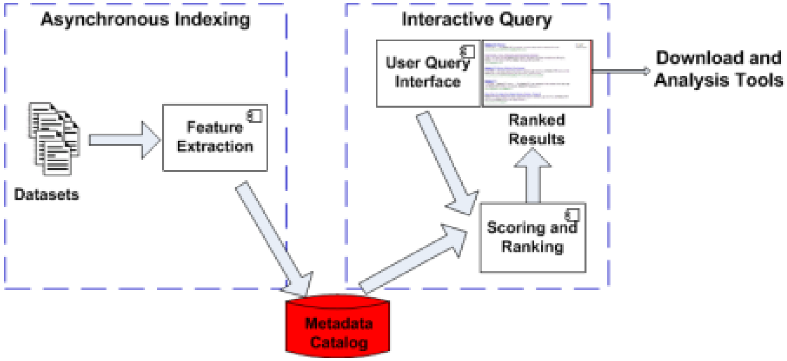
  Satellite, model data [dense grids]

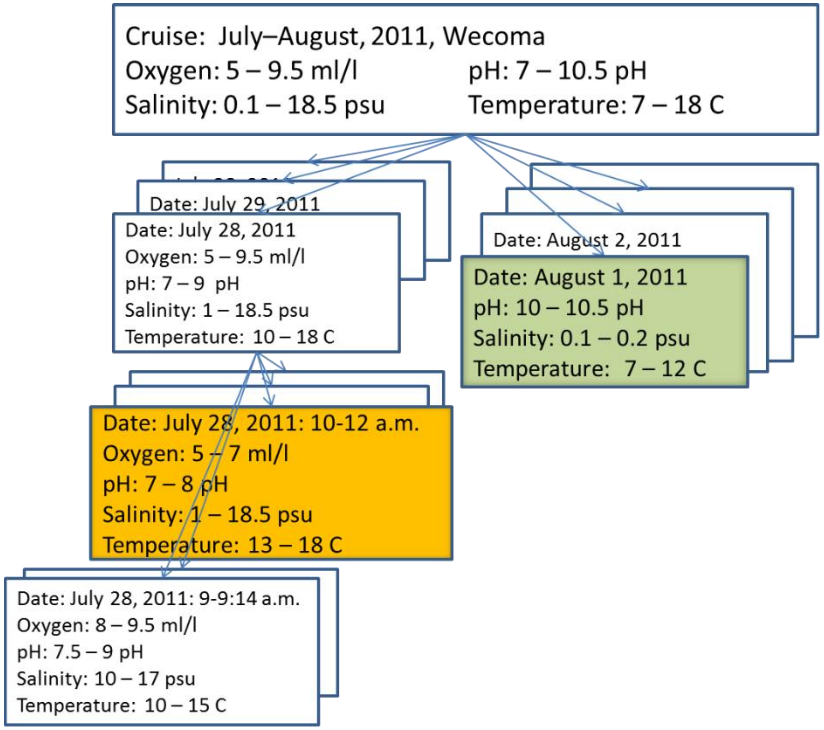
Dataset id:	**saturn01.ctd.201005**
Description:	Saturn-01 Profiler, May 2010
Quality:	Verified
Times [start .. end]:	2010-05-14 .. 2010-05-31
Geometry (location):	Point(-123.87,46.23)
Elevations, datum:	-13 .. 2.5 [m], NGVD27
# Observations:	247,377
Data Location:	http://...
Data Format:	NetCDF
Variables [units] (values):	Salinity [psu] (0 .. 29.6) Temperature [C] (8.2 .. 14.6) Time [secs since epoch] (1,273,869,578 .. 1,275,378,800)

Example “Dataset Summary”

Prototype: Adaptive Metadata Hierarchy

- Multiple granularities of data via unbalanced hierarchy of summaries
- Curator makes decision(s) once per kind of data/dataset





Cruise: July–August, 2011, Wecoma
Oxygen: 5 – 9.5 ml/l pH: 7 – 10.5 pH
Salinity: 0.1 – 18.5 psu Temperature: 7 – 18 C

Date: July 29, 2011

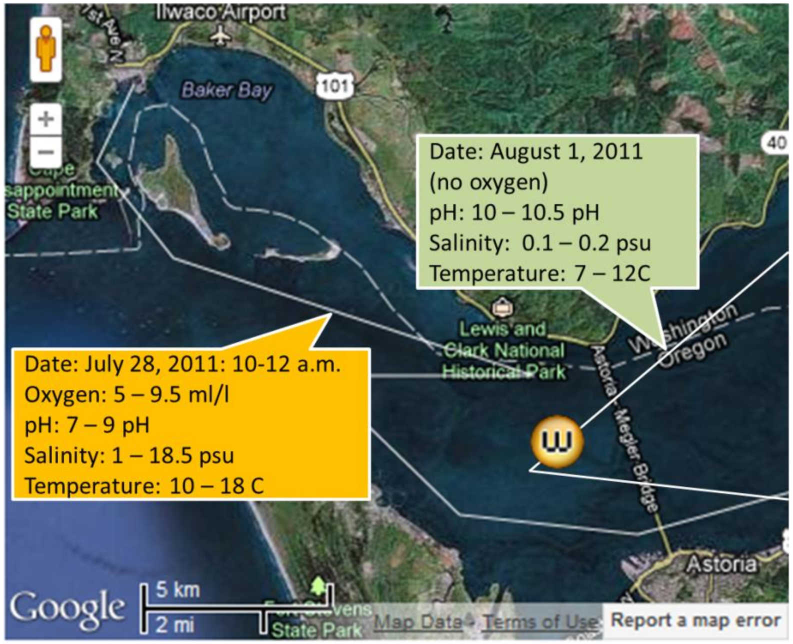
Date: July 28, 2011
Oxygen: 5 – 9.5 ml/l
pH: 7 – 9 pH
Salinity: 1 – 18.5 psu
Temperature: 10 – 18 C

Date: August 2, 2011

Date: August 1, 2011
pH: 10 – 10.5 pH
Salinity: 0.1 – 0.2 psu
Temperature: 7 – 12 C

Date: July 28, 2011: 10-12 a.m.
Oxygen: 5 – 7 ml/l
pH: 7 – 8 pH
Salinity: 1 – 18.5 psu
Temperature: 13 – 18 C

Date: July 28, 2011: 9-9:14 a.m.
Oxygen: 8 – 9.5 ml/l
pH: 7.5 – 9 pH
Salinity: 10 – 17 psu
Temperature: 10 – 15 C

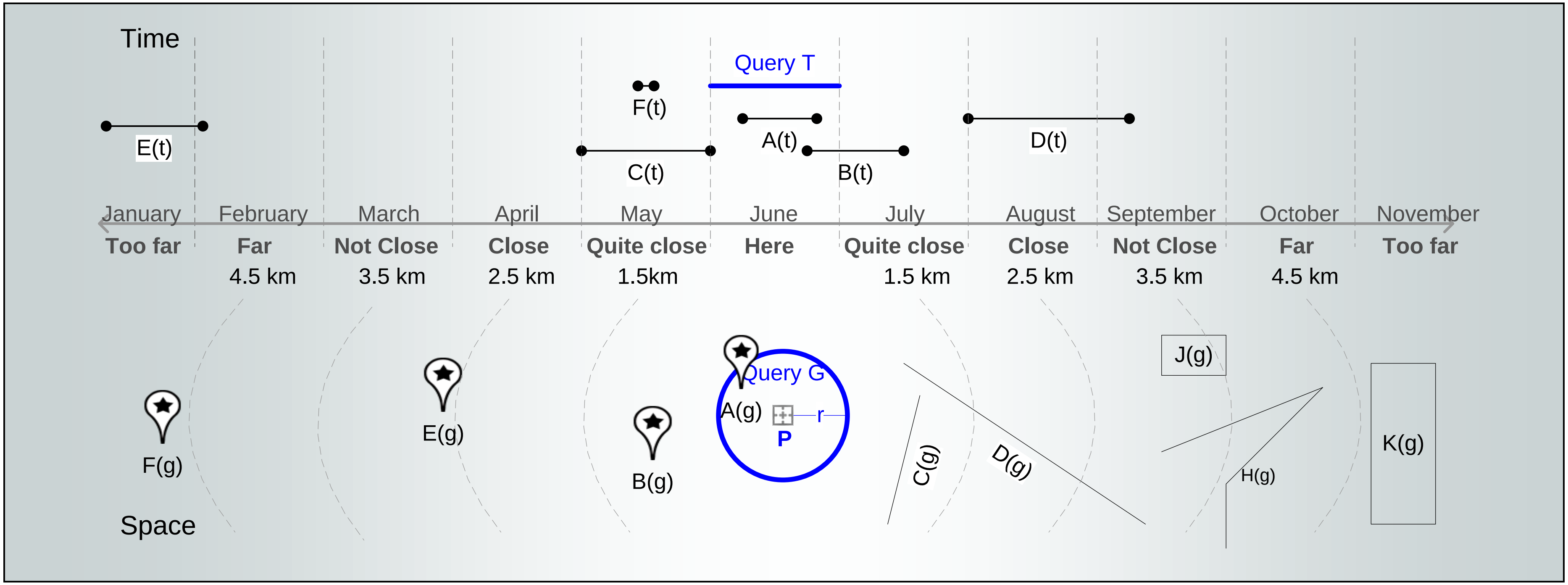


Date: August 1, 2011
(no oxygen)
pH: 10 – 10.5 pH
Salinity: 0.1 – 0.2 psu
Temperature: 7 – 12C

Date: July 28, 2011: 10-12 a.m.
Oxygen: 5 – 9.5 ml/l
pH: 7 – 9 pH
Salinity: 1 – 18.5 psu
Temperature: 10 – 18 C

Space-Time Ranking: Mental Model

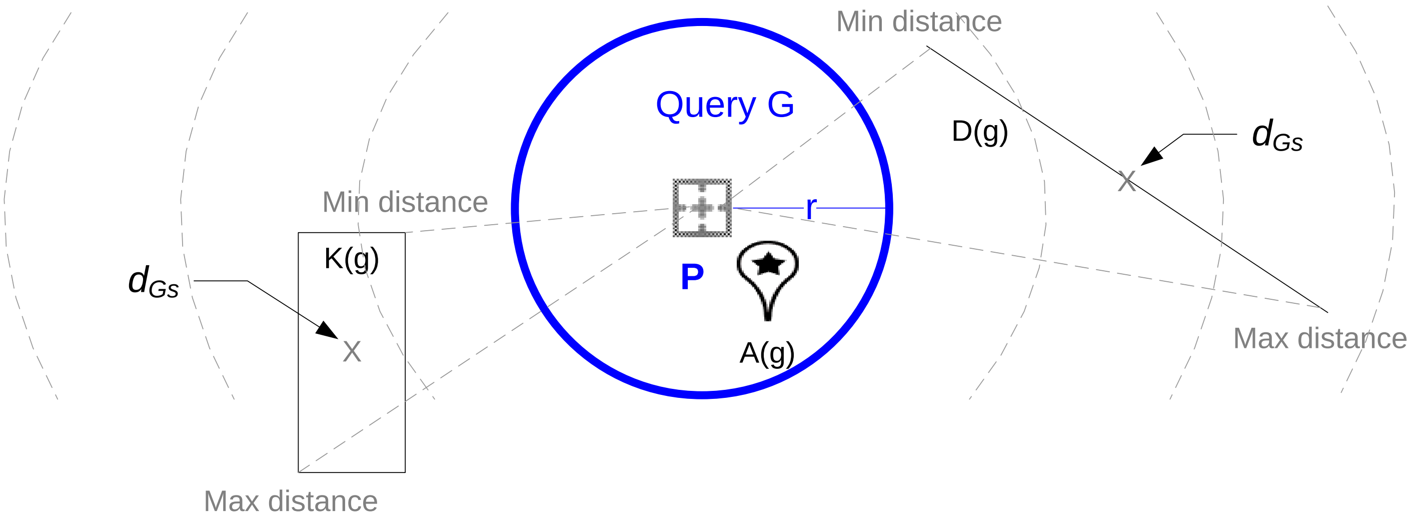
- Example Query: “Observations within ½ km of point ‘P’, in June 2009”
- Each dataset A, B, … represented by its time extent A(t), B(t), … and its geospatial extent A(g), B(g), …



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

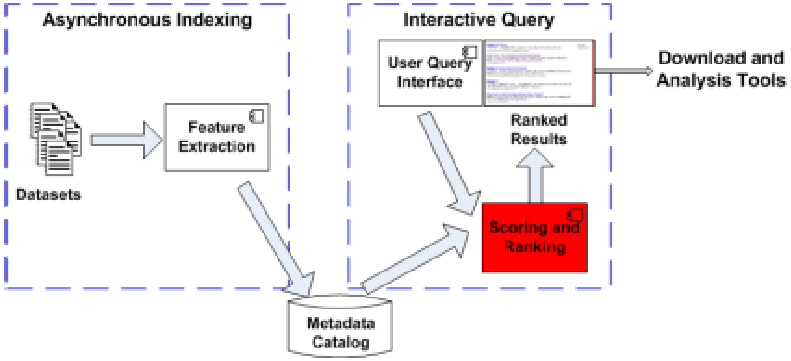
Scoring Datasets (1)

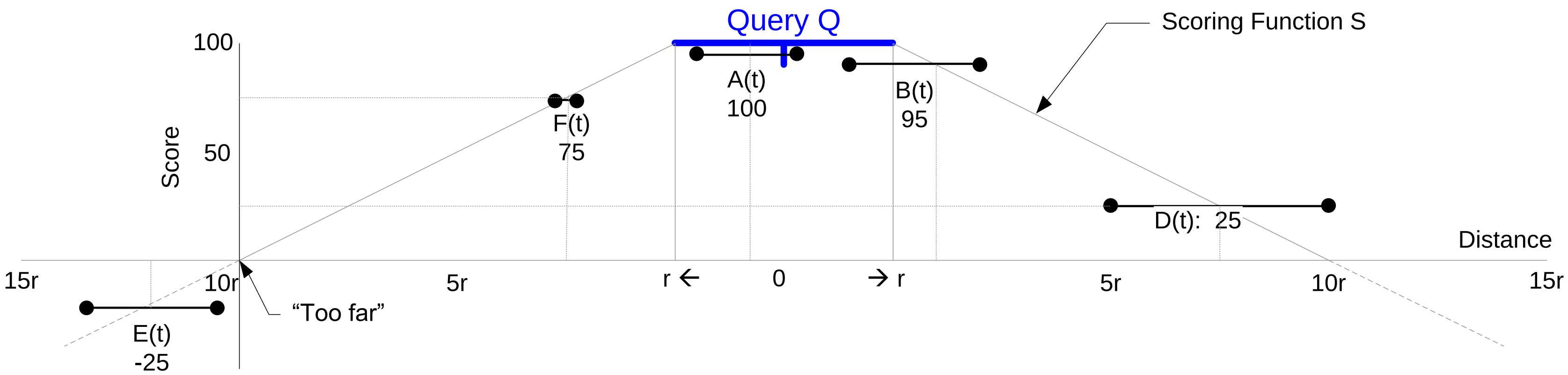
- Score each dataset using formulae that quantify the model
- Given a geospatial query *G*, calculate spatial-relevance score d_{Gs} for dataset *d*
- Spatial relevance is approximated by:
  - ½ (min distance + max distance) / radius
  - Apply scoring function to the result



Prototype: Scoring Datasets

- Simple distance-based formula
- Each variable's "distance" converted to "unit-less" measure
  - Distance: number of query radii from query term
  - Adjusted for overlap with query term





- Scoring performed per query term
- Relative importance of query terms defined by range

My Research

- "The principal contribution … is to define a new problem"[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

- Provided evidence of utility
  - Two user studies
  - "Defined a baseline ranking function against which future developments can be compared" [1]

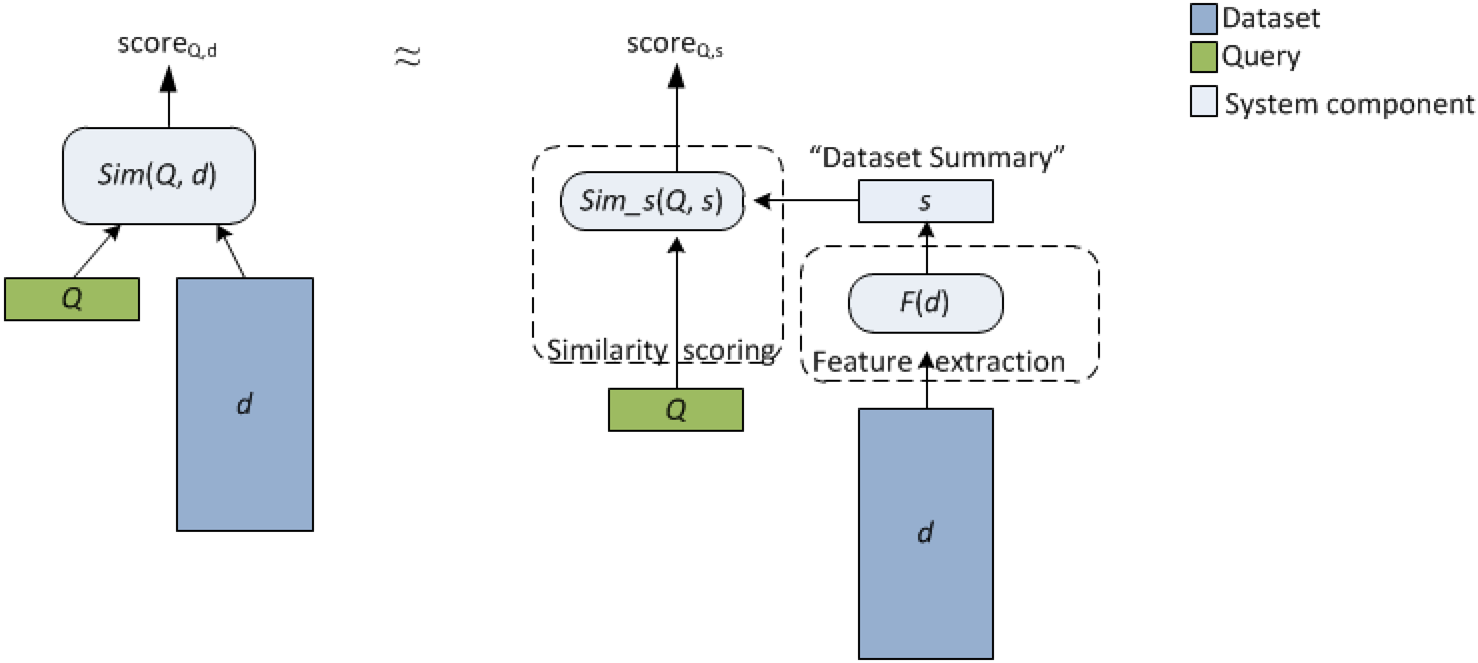
- (In progress) Evaluate scalability

[1] Comment from anonymous reviewer

Model (1)

- Requirements:
  - Assess similarity of query to dataset
  - Allow scaling independent of dataset size
  - Provide multiple data granularities: “most useful meaning-bearing unit”

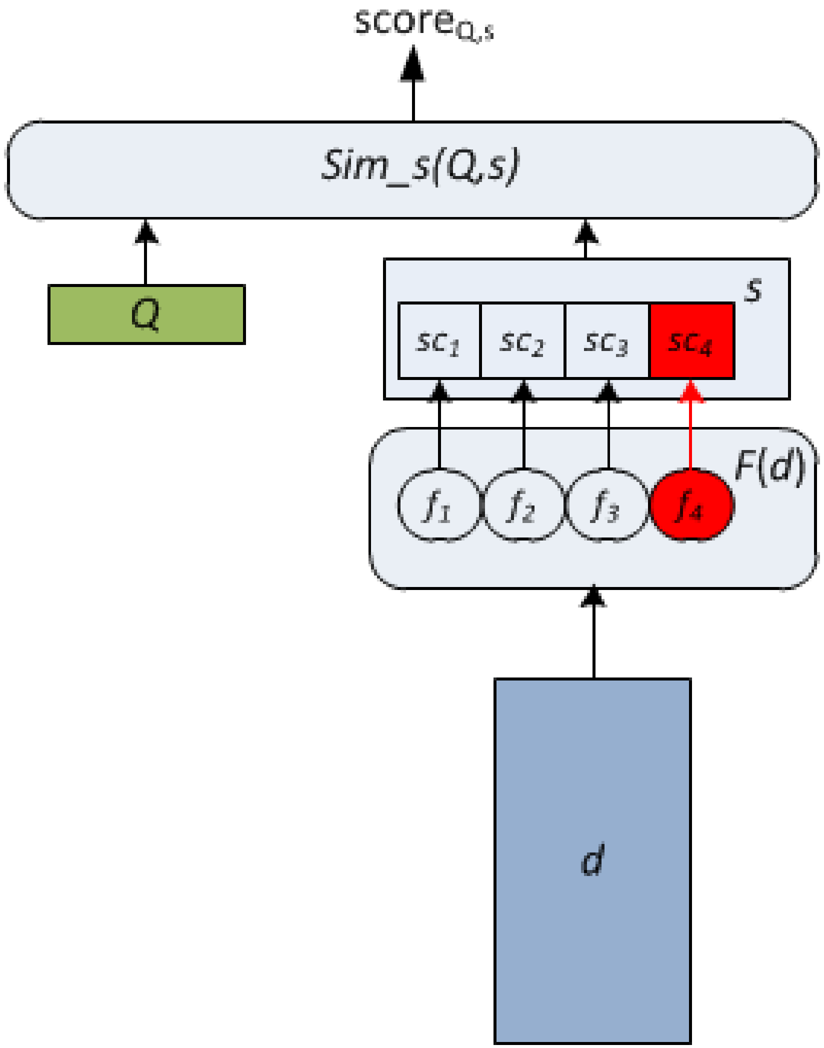
- Approach:
  - De-couple feature extraction from similarity scoring
  - Identify lightweight $Sim\_s(Q,s)$ where: $Sim\_s(Q,s) \approx Sim(Q,d)$

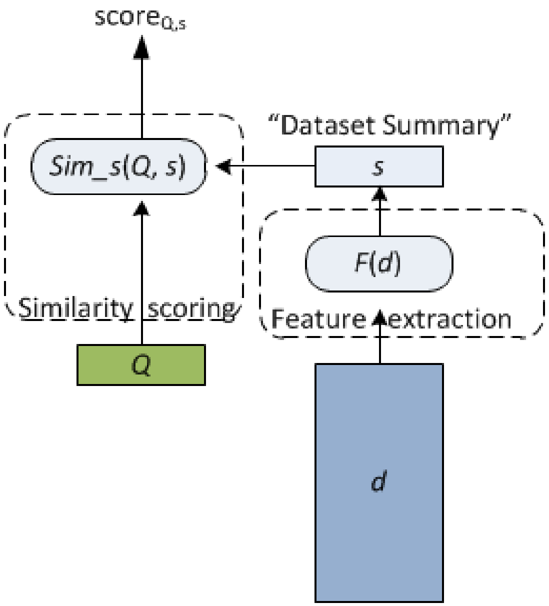


Model (2): Feature Extraction

For each dataset *d* in an archive *D:*

- Input: dataset *d*
- Processing: perform componentized extractions (f_1 .. f_n)
- Return: summary *s*

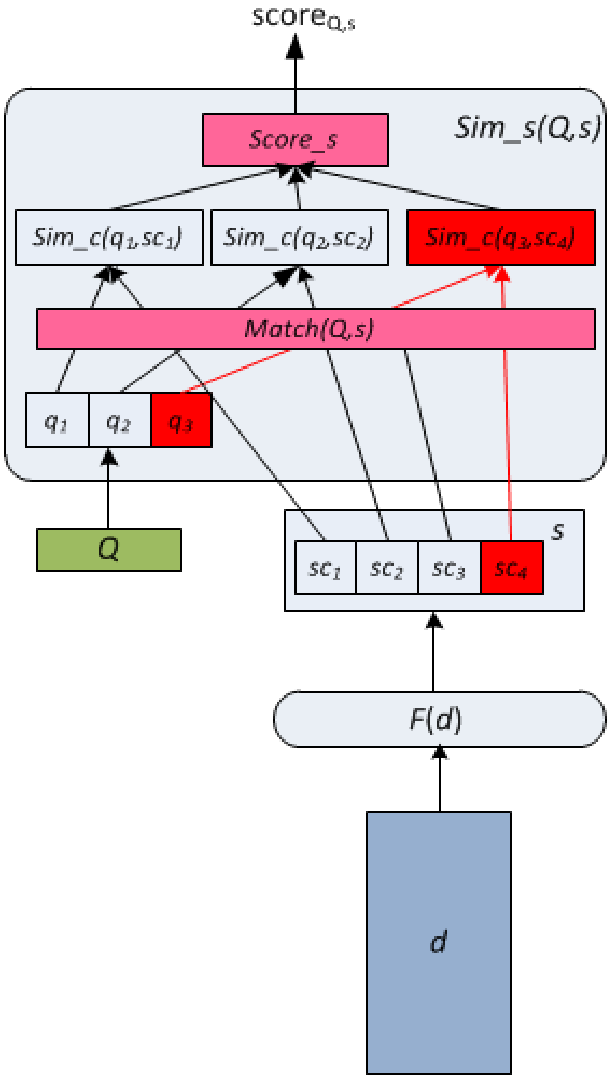


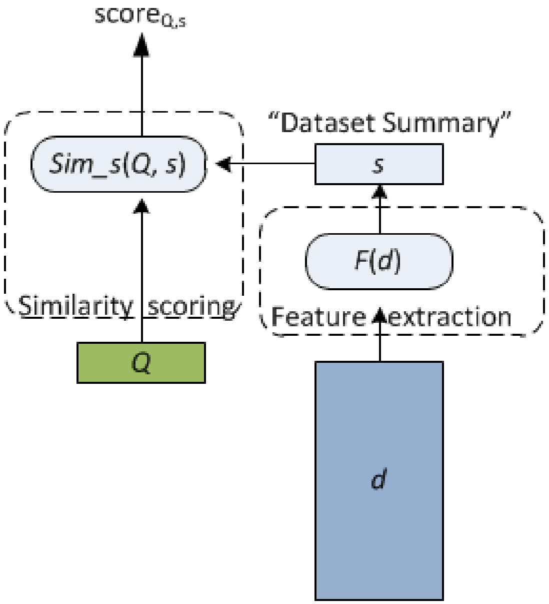


Model (3): Similarity Scoring

- Inputs: query *Q*, set of summaries *S*
- Processing:
  - For each dataset summary *s* in *S*:
    1. *Match:* Pair each query term with a summary feature

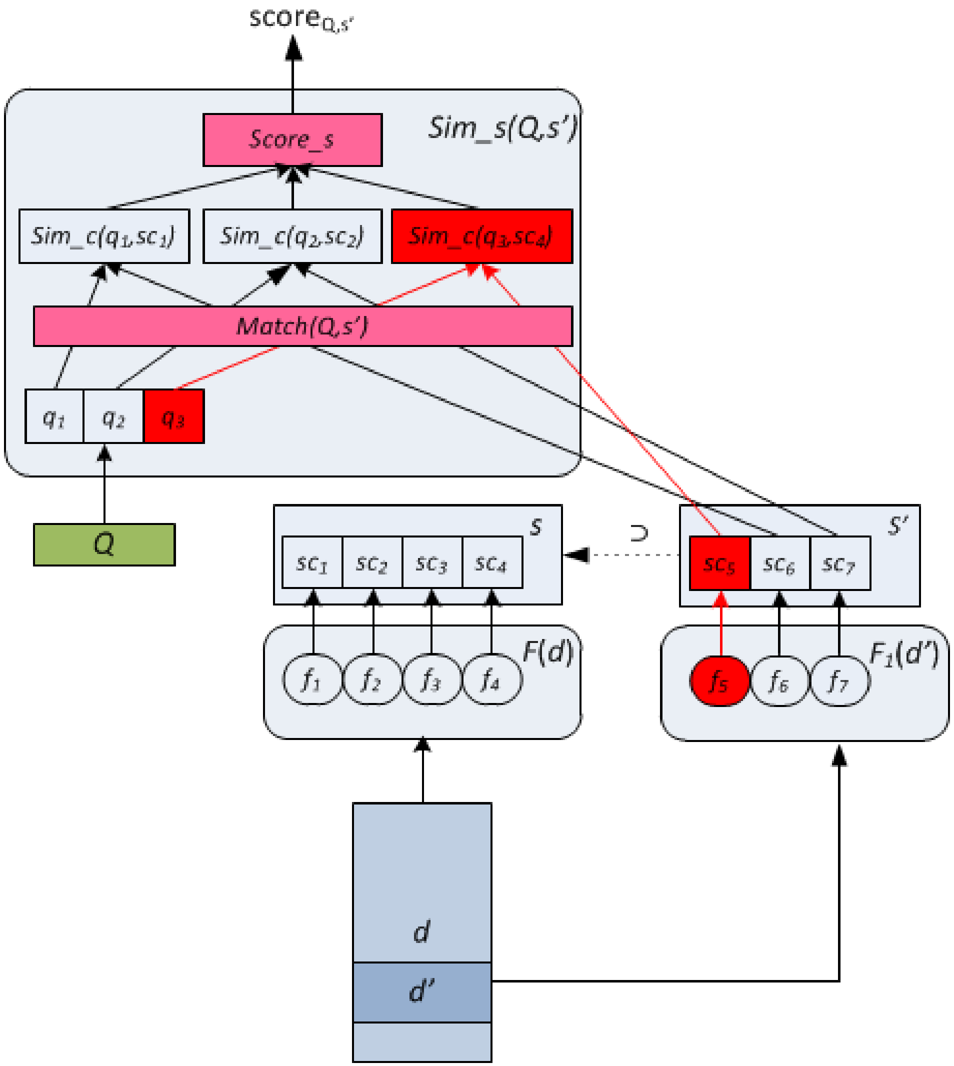
       Sim_c: Calculate similarity between each (query term, feature) pair
    2. *Score_s*: Combine into final score
- Return: *k* top-scoring summaries





Model (4): Summaries in Adaptive Hierarchies

- Purpose: Provide access to data at multiple granularities
- Feature extraction:
  - Create multiple summaries for same data
  - Maintain subset/superset relationships
- Similarity scoring:
  - Return top-scoring summaries from any level of hierarchy

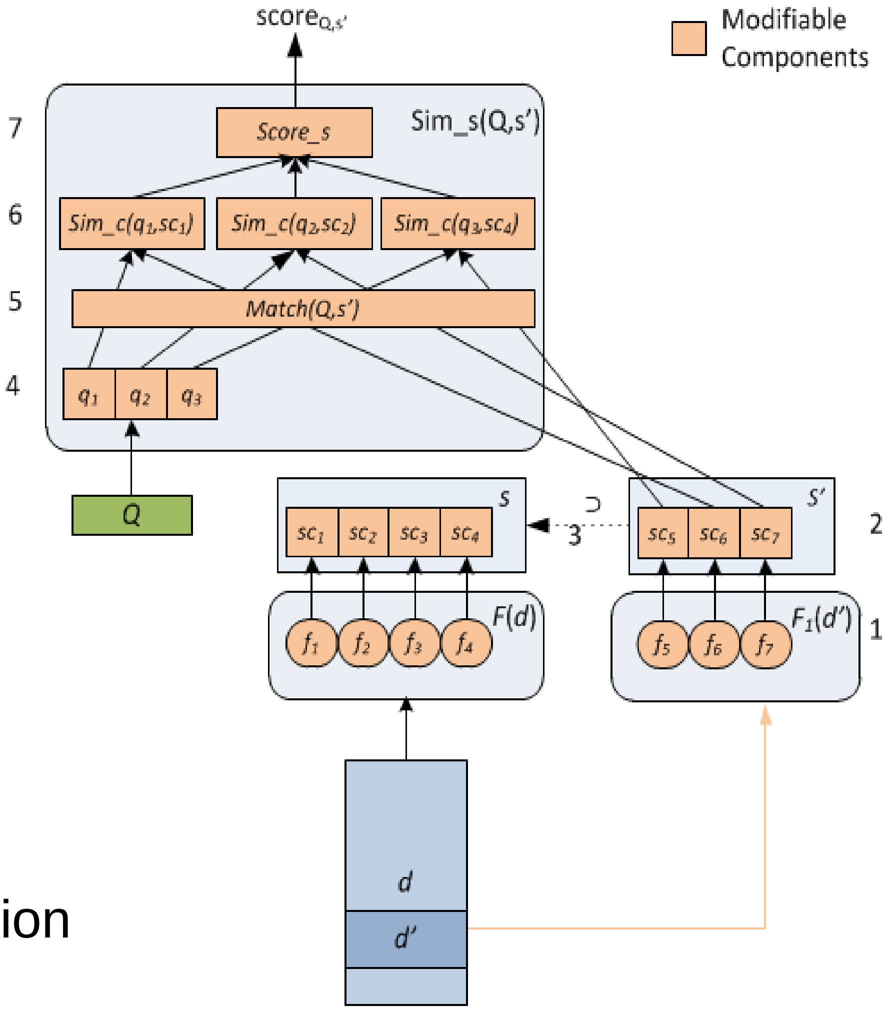


Model (5): Pluggable Components

- Allows individual modification of:
  1. Dataset summarization approaches
  2. Summary contents
  3. Hierarchical partitioning
  4. Form of the query terms
  5. Matching approaches
  6. Similarity functions
  7. Score combining

- Some component dependencies exist

- Supports componentized implementation architecture



My Research

- “The principal contribution … is to define a new problem”[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

- Provided evidence of utility
  - Two user studies
  - “Defined a baseline ranking function against which future developments can be compared” [1]

- (In progress) Evaluate scalability

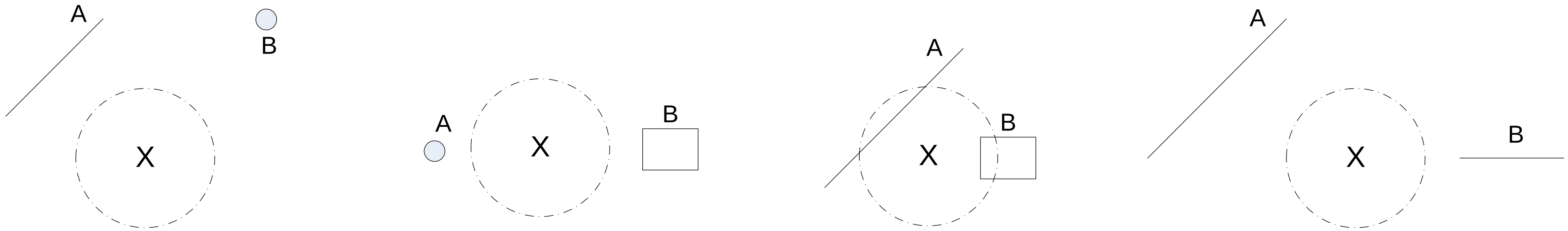
[1] Comment from anonymous reviewer

Utility: User Study 1

- Premise: Candidate similarity function resembles “human perceptions”

- Populations: Two populations, each *n*=20
  - “Scientists” (domain experts)
  - “Non-scientists” (non-domain experts)

- Findings:
  - Similarity function adequately reflects respondent’s assessments
  - Respondents related to “dataset summary” concept
  - Space, time, and space-and-time comparisons resonated with respondents



Example “spatial comparison” questions from User Study 1

Utility: User Study 2

- Premise:
  - Similarity measure extends to variable search
  - Implementation effective for dataset search

- Two-part user study:
  1. Qualitative assessment of query experience

     Likert scale
  2. Quantitative assessment of relevance

     Respondents rate relevance of individual datasets returned by prototype

- Population: 13 CMOP scientists

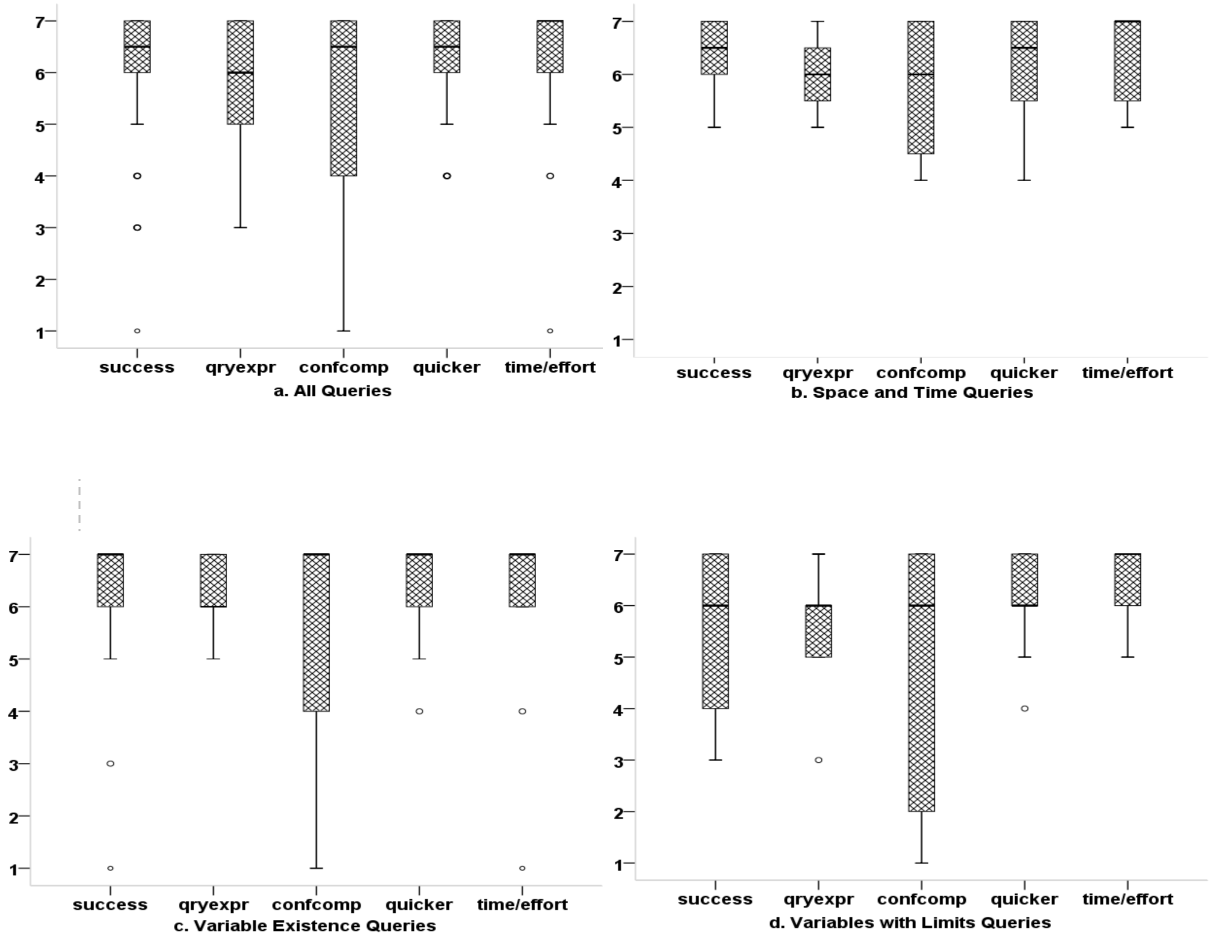
- Information needs and queries provided by respondents

User Study 2: Qualitative Assessment

- Finding: DNH receives high scores on all subjective assessments
  - 7-step Likert scale (1:poor, 7:excellent)
  - Best scores on variable existence; poorest on variables with limits

1. How successful was this search in helping with your information need? [success]
2. How well does this style of query allow you to express your information need? [qryexpr]
3. How confident are you in the completeness of search results? [confcomp]
4. Was using this tool quicker than finding the most relevant results by other means? [quicker]
5. How valuable are the search results versus time expended? [time/effort]

Study Questions



a. All Queries

b. Space and Time Queries

c. Variable Existence Queries

d. Variables with Limits Queries

Example: How Alternative Rankers are Evaluated

- Wanted: a relevance measure that simulates users' rankings

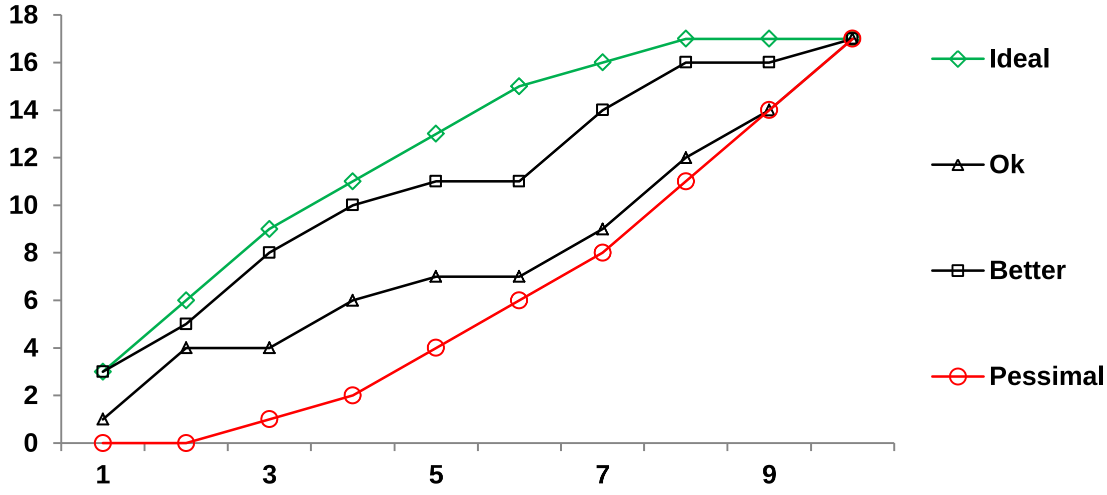
  Most-relevant items near top; least-relevant near bottom

- Focus on accuracy in the top few items returned
  - "Discount" rankings of items further down the list
  - Discounted Cumulative Gain (DCG)

    commonly-used evaluation measure

Order of Rankings			
Ideal	Better	Ok	Pessimal
3	3	1	0
3	2	3	0
3	3	0	1
2	2	2	1
2	1	1	2
2	0	0	2
1	3	2	2
1	2	3	3
0	0	2	3
0	1	3	3

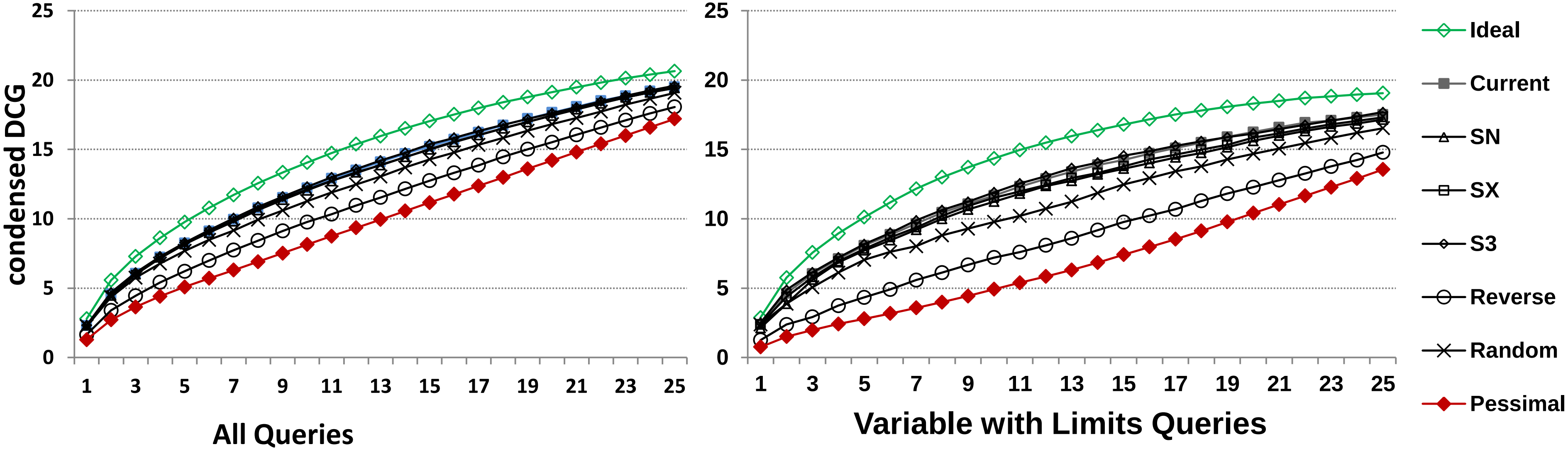
Example Rankings (3=high, 0=not relevant)



Discounted Cumulative Gain (DCG) of Example Rankings

User Study 2: Quantitative Assessment (1)

- Finding: existing ranking method performs well, compared to ideal
  - 2 different comparisons used (condensed DCG and Average RBP)
  - Alternative rankings studied not significantly better
  - Random, pessimal and reverse lines show potential for “worse”



DNH Rankings: condensed Discounted Cumulative Gain

My Research

- “The principal contribution … is to define a new problem”[1]

- Defined a new approach
  - Apply Information Retrieval (IR) techniques: ranked search
  - Use adaptive, hierarchical metadata

- Developed prototype

  In production use by CMOP scientists

- Defined formal model & componentized architecture

- Provided evidence of utility
  - Two user studies
  - “Defined a baseline ranking function against which future developments can be compared” [1]

- (In progress) Evaluate scalability

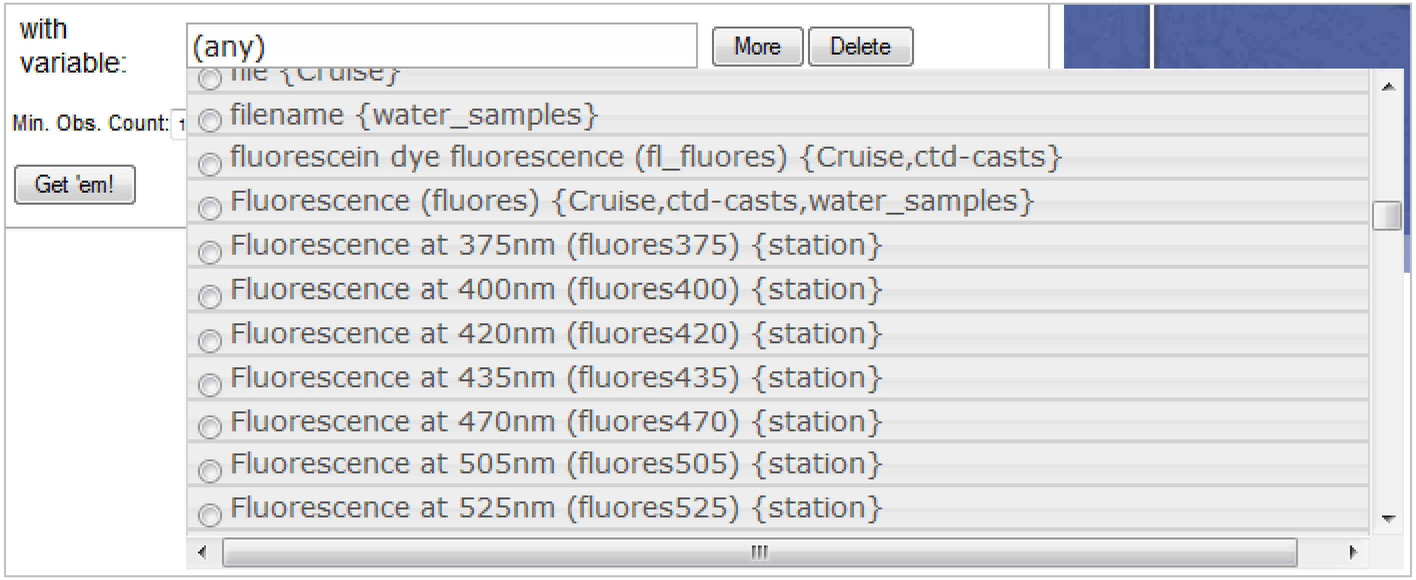
[1] Comment from anonymous reviewer

The Metadata Mess

- Working assumption: each named column in a (publicly available) dataset represents a valid variable

- Result: Ever increasing number of variables (over 300 at CMOP)

- Problem:
  - Hard for searchers to navigate, locate desired variable
  - Not what the archive wants to expose – “metadata mess” ← our focus



with variable: (any)

Min. Obs. Count: 1

○ file {Cruise}
○ filename {water_samples}
○ fluorescein dye fluorescence (fl_fluores) {Cruise,ctd-casts}
○ Fluorescence (fluores) {Cruise,ctd-casts,water_samples}
○ Fluorescence at 375nm (fluores375) {station}
○ Fluorescence at 400nm (fluores400) {station}
○ Fluorescence at 420nm (fluores420) {station}
○ Fluorescence at 435nm (fluores435) {station}
○ Fluorescence at 470nm (fluores470) {station}
○ Fluorescence at 505nm (fluores505) {station}
○ Fluorescence at 525nm (fluores525) {station}

Figure: Variable List as Exposed in Search Tool

Characterizing the Metadata Mess

- Archive curator’s goal: to present the metadata he wishes he had

- Sources of the mess:
  - Poor, unenforced or multiple naming standards
  - Data from multiple or external sources or systems
  - Changes in systems, standards and personnel over time
  - Many researchers, from different fields
  - Changing research foci

- Can’t we repair the archive?
  - Datasets must be modified or regenerated – not practical
  - May require changing code, systems – expensive, limited payoff
  - Names may be set by vendors or external data providers
  - Time-consuming, error-prone – and problems recur
  - Change is constant

The Metadata Mess (2)

- Alternative approach: compensate for the mess
- How?
  - Reduce semantic diversity

    Perfection not needed

  - Provide transformation layer from “what is” to “what should be”

Categories of Semantic Diversity

Category	**Example**
Minor variations and misspellings	*air_temperature*, *air_temperatrue*, *airtemp*
Synonyms	*C*, *degC*, *Centigrade*
Abbreviations	*MWHLA*
Excess variables	Quality assurance variables: *qa_level*
Ambiguous usages	*temp*: *temporary* or *temperature*?
Source-context naming variations	*temperature* may mean *air_temperature* or *water_temperature*, depending on source context
Concepts at multiple levels of detail	*Fluorescence*, vs. *fluores375*, *fluores400*

Semantic Diversity: Overall Approach

- Principles:
  - No one approach sufficient
  - All approaches must be:
    - Simple
    - Robust
    - Tolerant of continued growth and ambiguity
  - “Refunds and exchanges available”
    - Provide defaults
    - Improve results via overrides, modifications, adjustments
    - Be non-destructive: re-doable metadata processing

- “Semi-curated” model
  - Curator performs some work for each new type of data indexed
  - Curator can review, adjust and override currently-used defaults and prior decisions

Reducing Variable-Name Diversity: Possible Approaches

Category	**Example**	**Desired Result**	**Possible Technical Approach**
Minor variations and misspellings	*air_temperature*, *air_temperatrue*, *airtemp*	Make them the same	Translate current to desired name
Synonyms	*C*, *degC*, *Centigrade*	Make them the same	Translate current to desired name
Abbreviations	*MWHLA*	Use full/canonical variable name	Translate current to desired name
Excess variables	Quality assurance variables: *qa_level*	Exclude from search Show in detailed dataset views	Mark variables Exclude from search
Ambiguous usages	*temp*: *temporary* or *temperature?*	Identify and expose variables. Allow curator to: • clarify where possible • hide variable • leave as is	Provide interface to specify options
Source-context naming variations	*Temperature*: *air_temperature* or *water_temperature* depending on source context	Specify context of variable Make context accessible to user	Link to multiple taxonomies
Concepts at multiple levels of detail	*Fluorescence*, vs. *fluores375*, *fluores400*	Collapse or expose as needed	Allow variables to be grouped Support hierarchical menus

Patent, Papers, Presentations

Patent filed:

- US Patent Application Number 13/175,611, “A Search Tool that Utilizes Numerical Scientific Metadata Matched Against User-Entered Parameters”, Megler and Maier, filed June 2011.

Papers:

- “Are Datasets Like Documents?” (submitted), V.M. Megler, David Maier.
- “Data Near Here: Bringing Relevant Data Closer to Scientists” (in press), V.M. Megler, David Maier, *Computing in Science and Engineering*, 2013
- “Taming the Metadata Mess”, V.M. Megler, *Workshop for Ph.D. Students at ICDE*, 2013
- “When Big Data Leads to Lost Data” (Best Paper Award), V.M. Megler, David Maier, *PIKM 2012: 5th Workshop for Ph.D. Students at CIKM*, 2012
- “Navigating Oceans of Data”, David Maier, V.M. Megler, António M. Baptista, Alex Jaramillo, Charles Seaton, Paul J. Turner, in *Scientific and Statistical Database Management*, 2012, vol. 7338, pp. 1–19.
- “Finding Haystacks with Needles: Ranked Search for Data Using Geospatial and Temporal Characteristics”, Megler, V.M. & Maier, D. *Scientific and Statistical Database Management*, 2011, vol. 6809.

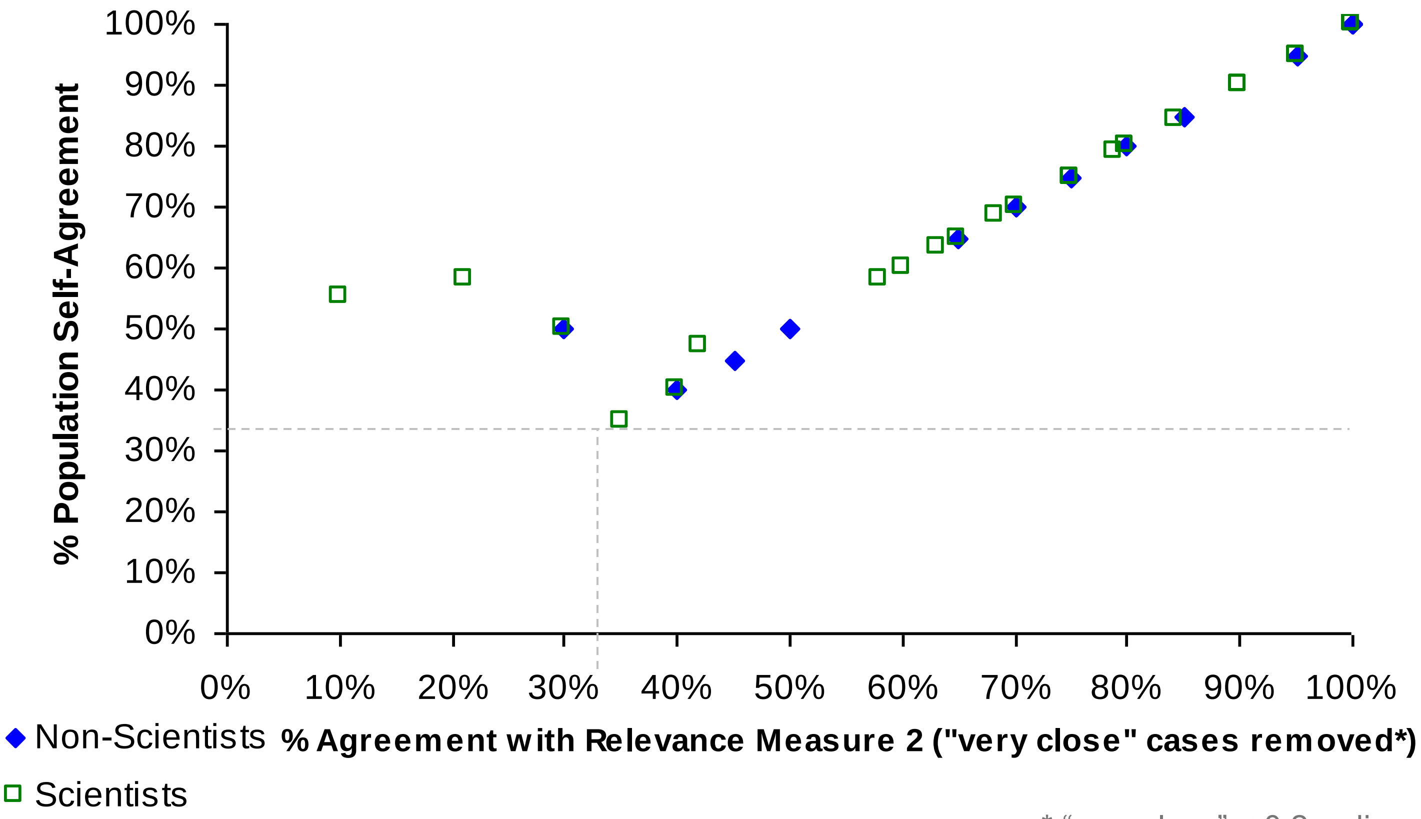
Conference & External Presentations:

- Presentation to National Science Foundation STC review committee, June 2012.
- “Needles in Haystacks: Finding Observational Data with Geospatial and Temporal Characteristics (Take 2)”, Veronika Megler and David Maier, Association of American Geographers Annual Conference (AAG), Seattle, Washington, April 2011.
- “Needles in Haystacks: Finding Observational Data with Geospatial and Temporal Characteristics”, Veronika Megler and David Maier, GIS In Action Conference, URISA, Portland, March 2011.

Backup

User Study 1: Sample Finding #1

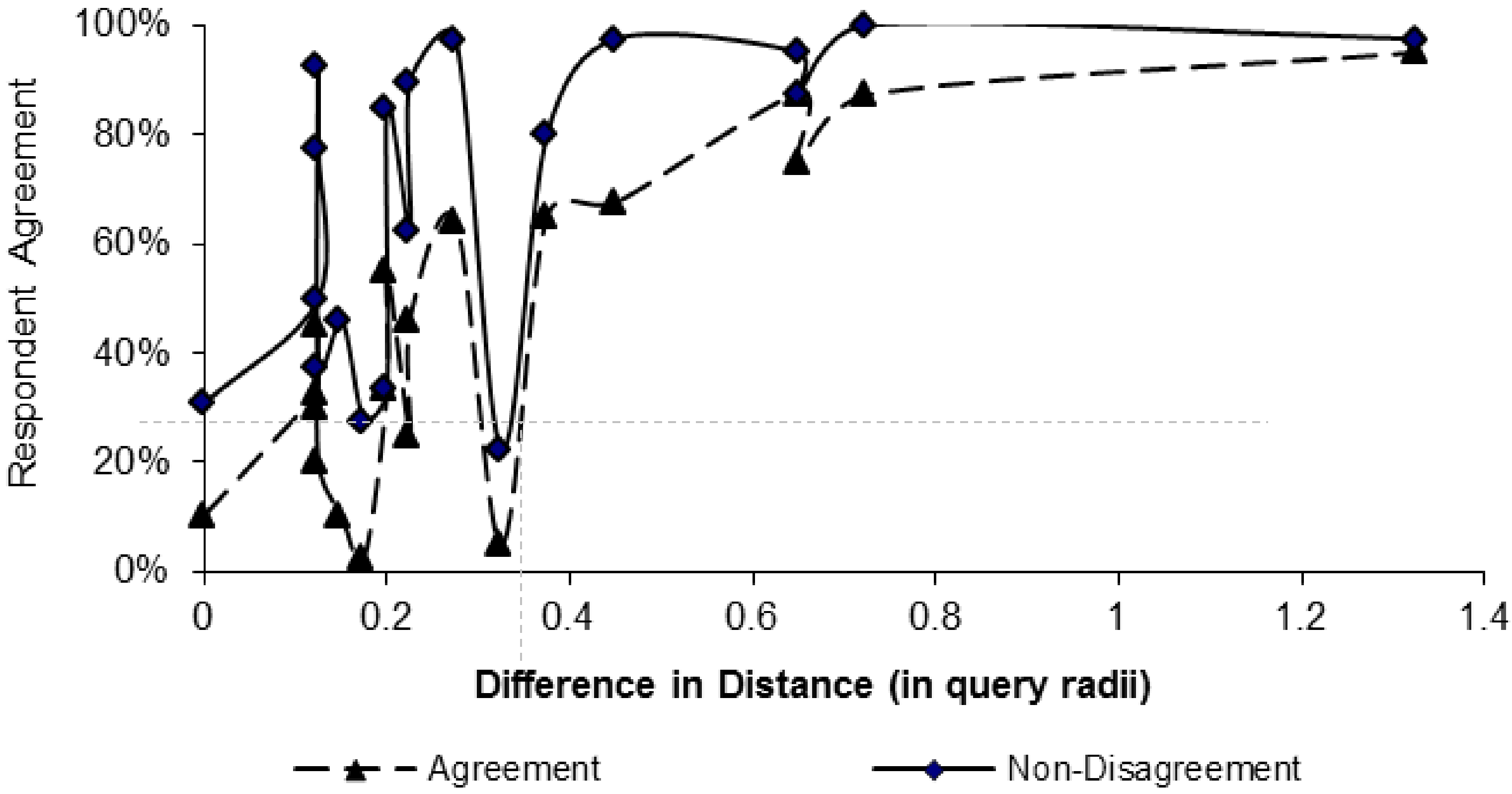
- Finding: Ordinal responses are independent of:
  - Type of question (time only, space only, time and space combined)
  - Shape (point, line, polyline, polygon)



* “very close” < 0.2 radius difference in distance

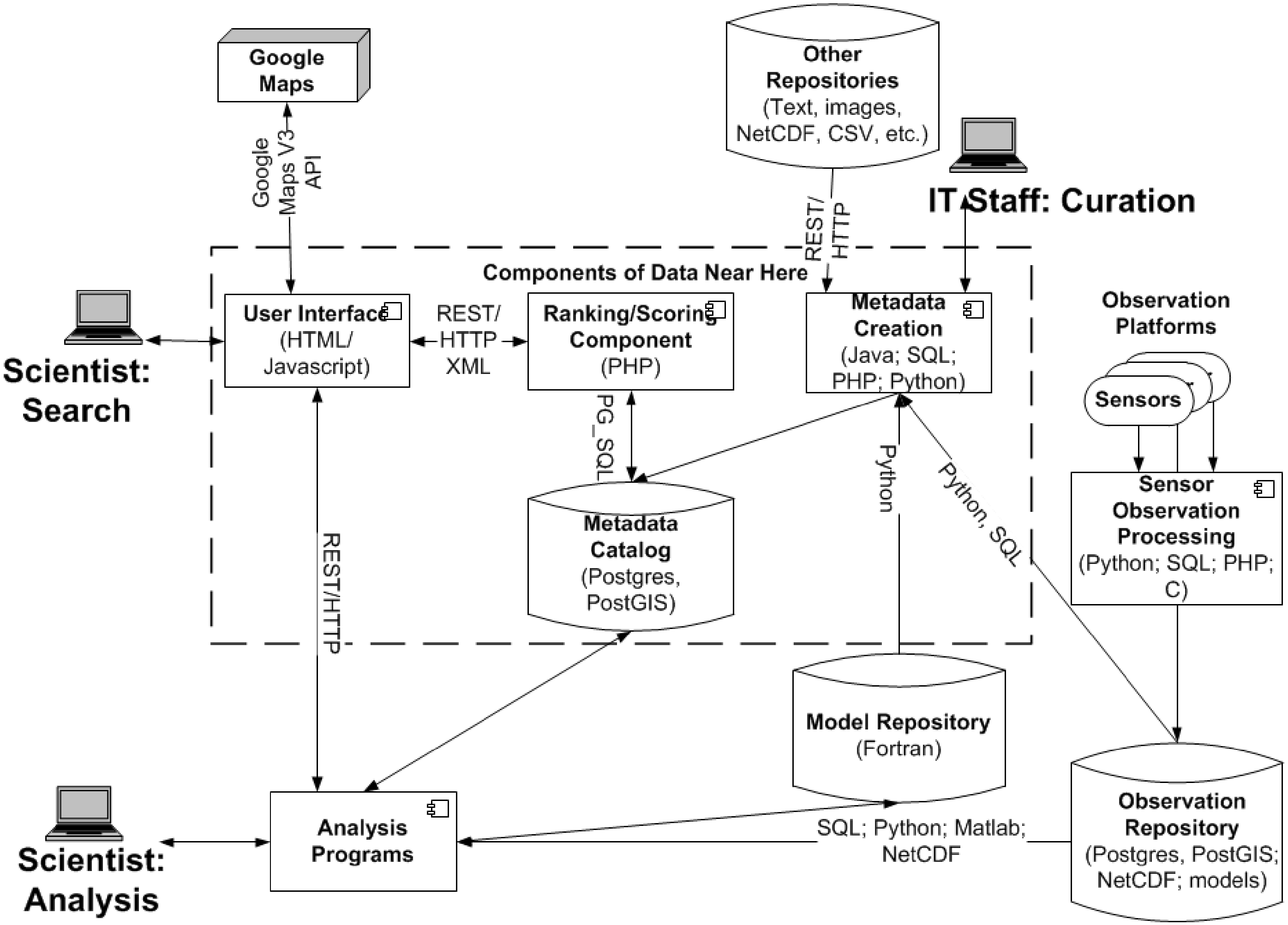
User Study 1: Sample Finding #2

- Finding: As differences between distance to two objects decreases, the assessment of which one is closer becomes more variable

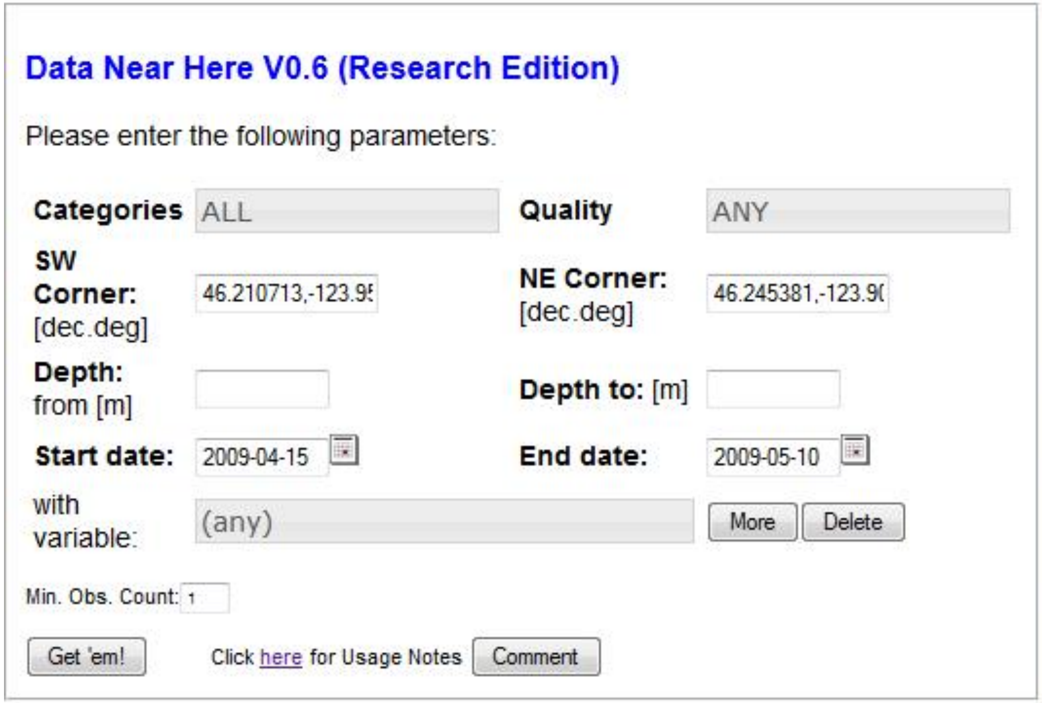


Prototype Implementation: “Data Near Here”

- Data Near Here components designed to “add” to existing environment
- Implementation technologies chosen based on CMOP standards


Google Maps
Google Maps V3 API
Other Repositories (Text, images, NetCDF, CSV, etc.)
REST/ HTTP
IT Staff: Curation
Components of Data Near Here
Scientist: Search
User Interface (HTML/ Javascript)
REST/ HTTP XML
Ranking/Scoring Component (PHP)
Metadata Creation (Java; SQL; PHP; Python)
Observation Platforms
Sensors
PG_SQL
Python
Python, SQL
Metadata Catalog (Postgres, PostGIS)
REST/HTTP
Sensor Observation Processing (Python; SQL; PHP; C)
Model Repository (Fortran)
Scientist: Analysis
Analysis Programs
SQL; Python; Matlab; NetCDF
Observation Repository (Postgres, PostGIS; NetCDF; models)

Prototype: Default Page



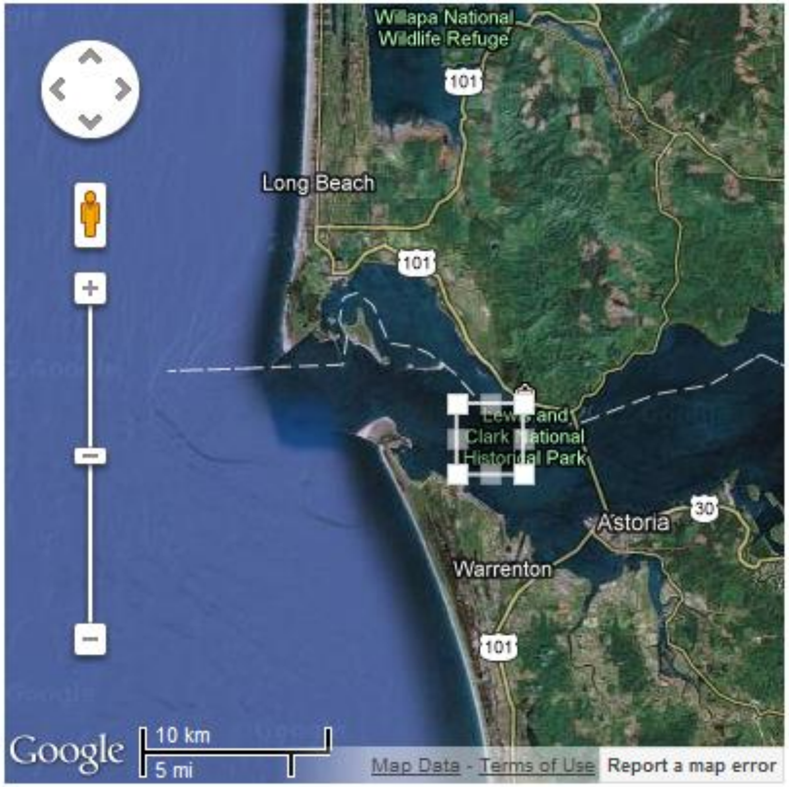
Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

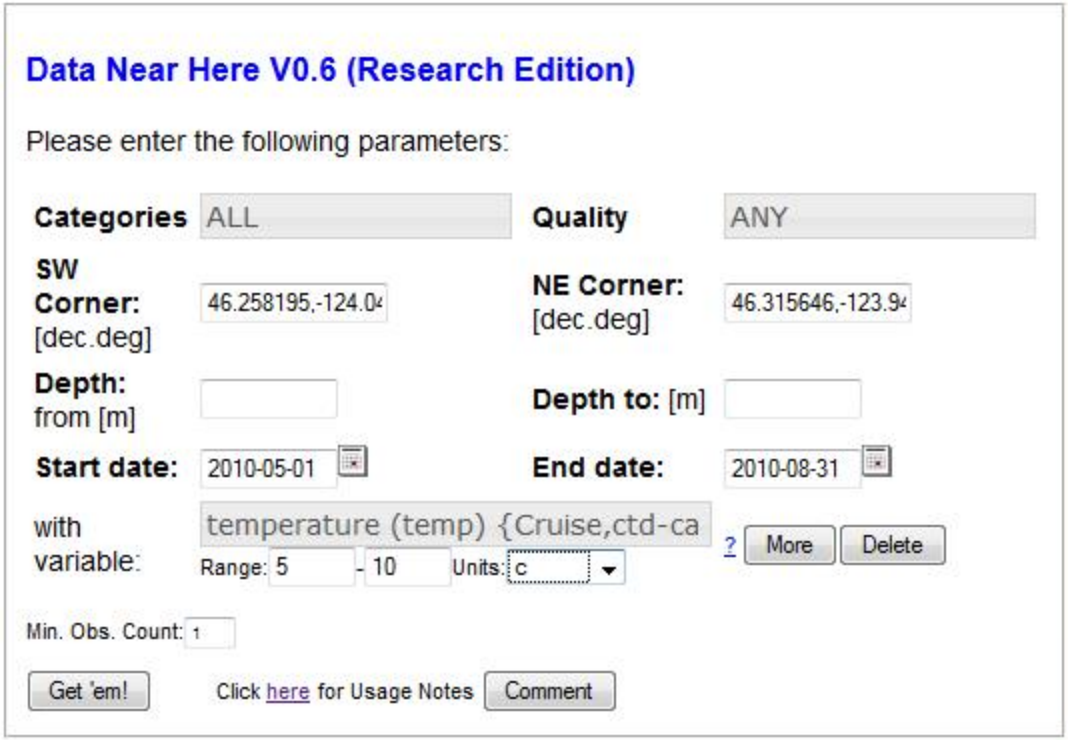
Categories	ALL	**Quality**	ANY
SW Corner: [dec.deg]	46.210713,-123.95	**NE Corner:** [dec.deg]	46.245381,-123.90
Depth: from [m]		**Depth to:** [m]	
Start date:	2009-04-15	**End date:**	2009-05-10
with variable:	(any)	More	Delete

Min. Obs. Count: 1

Get 'em!   Click here for Usage Notes   Comment



Prototype: Enter Query



Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

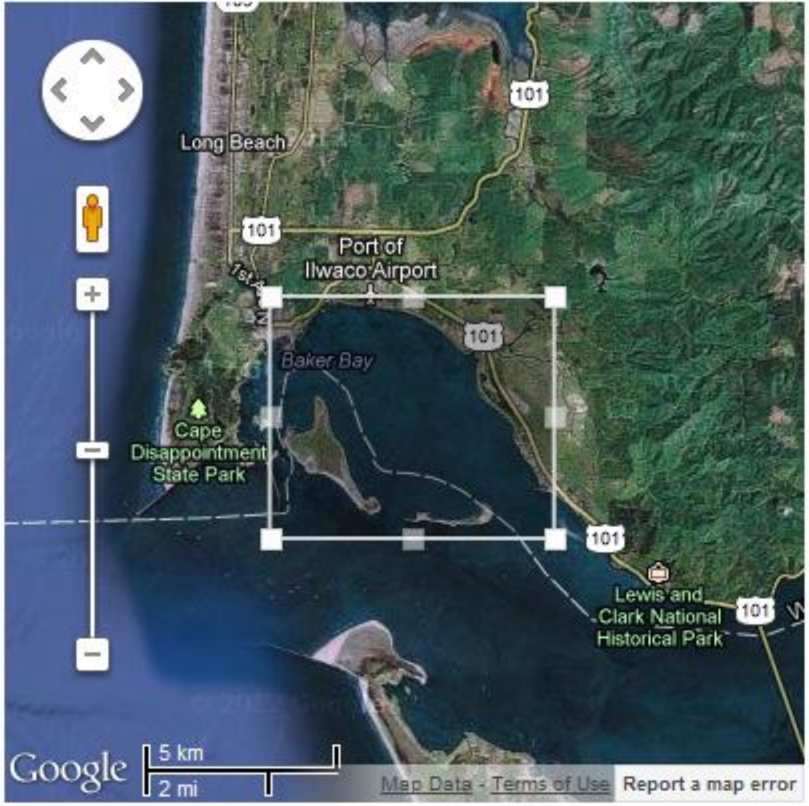
Categories	ALL	**Quality**	ANY
SW Corner: [dec.deg]	46.258195,-124.04	**NE Corner:** [dec.deg]	46.315646,-123.94
Depth: from [m]		**Depth to:** [m]	
Start date:	2010-05-01	**End date:**	2010-08-31

with variable: temperature (temp) {Cruise,ctd-ca ? More Delete

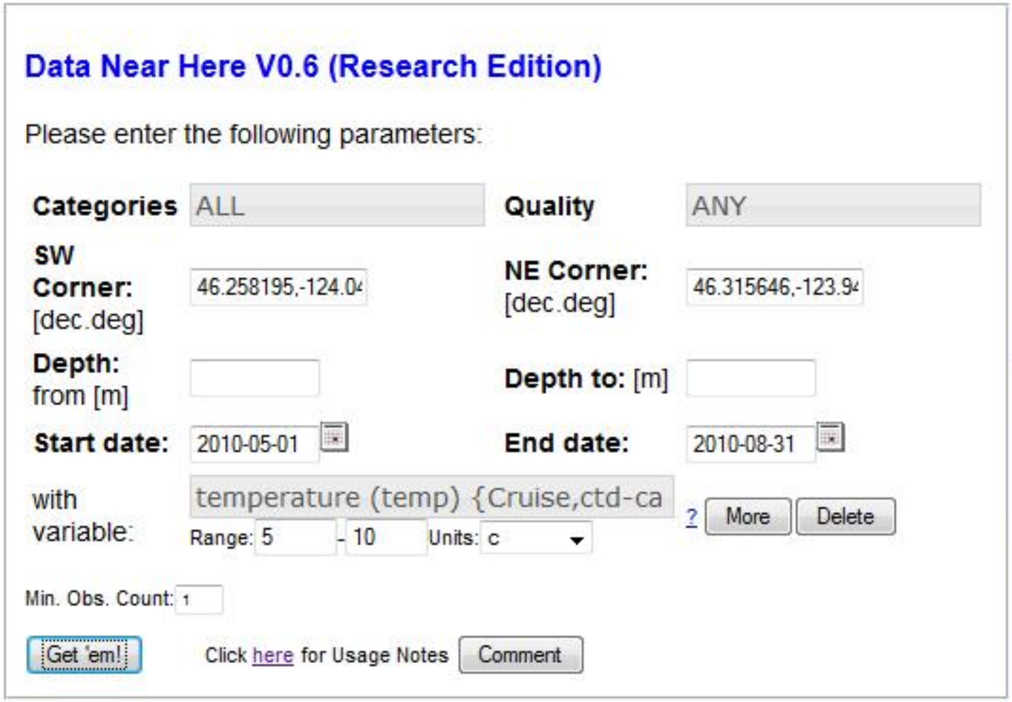
Range: 5 - 10 Units: c

Min. Obs. Count: 1

Get 'em! Click here for Usage Notes Comment



Prototype: Query Results



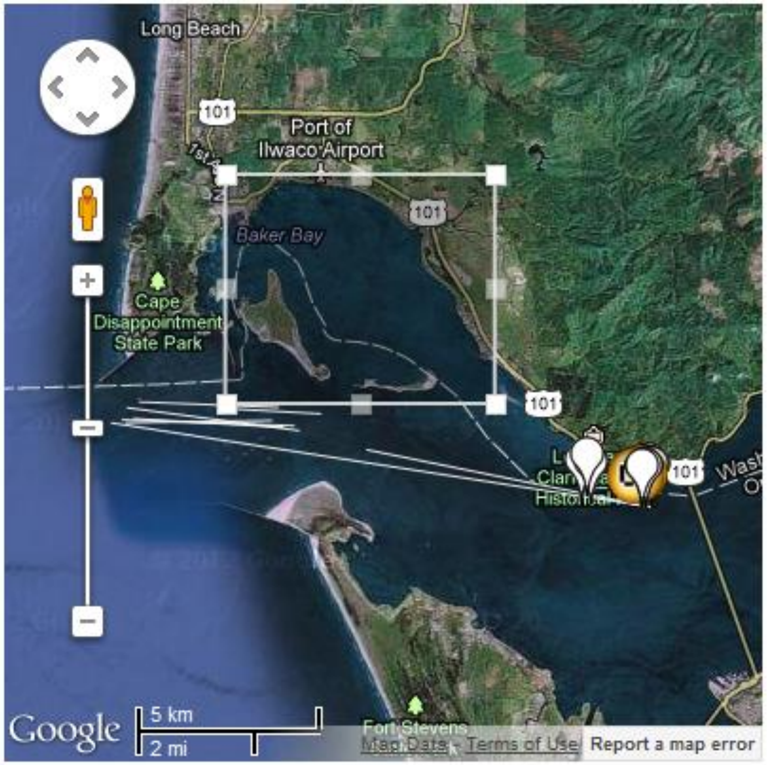
Data Near Here V0.6 (Research Edition)

Please enter the following parameters:

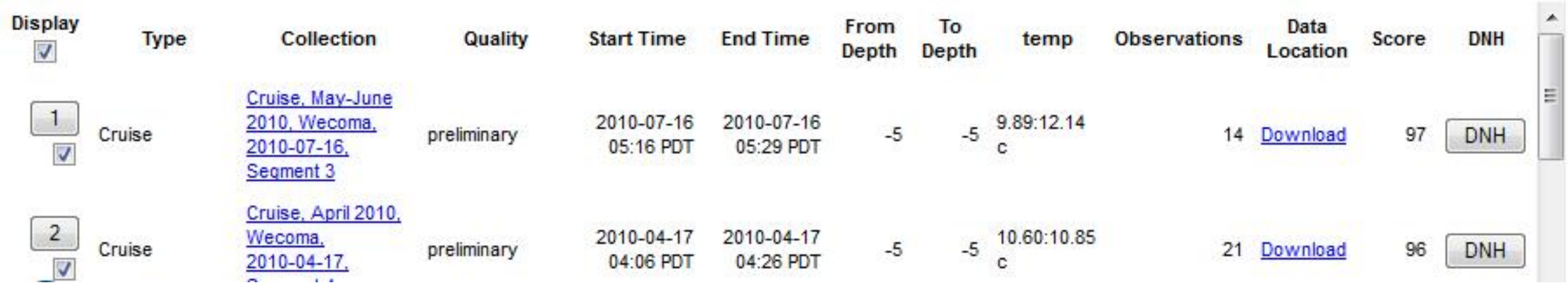
Categories	ALL	**Quality**	ANY
SW Corner: [dec.deg]	46.258195,-124.0	**NE Corner:** [dec.deg]	46.315646,-123.9
Depth: from [m]		**Depth to:** [m]	
Start date:	2010-05-01	**End date:**	2010-08-31
with variable:	temperature (temp) {Cruise,ctd-ca Range: 5 - 10 Units: c		

Min. Obs. Count: 1

Click here for Usage Notes

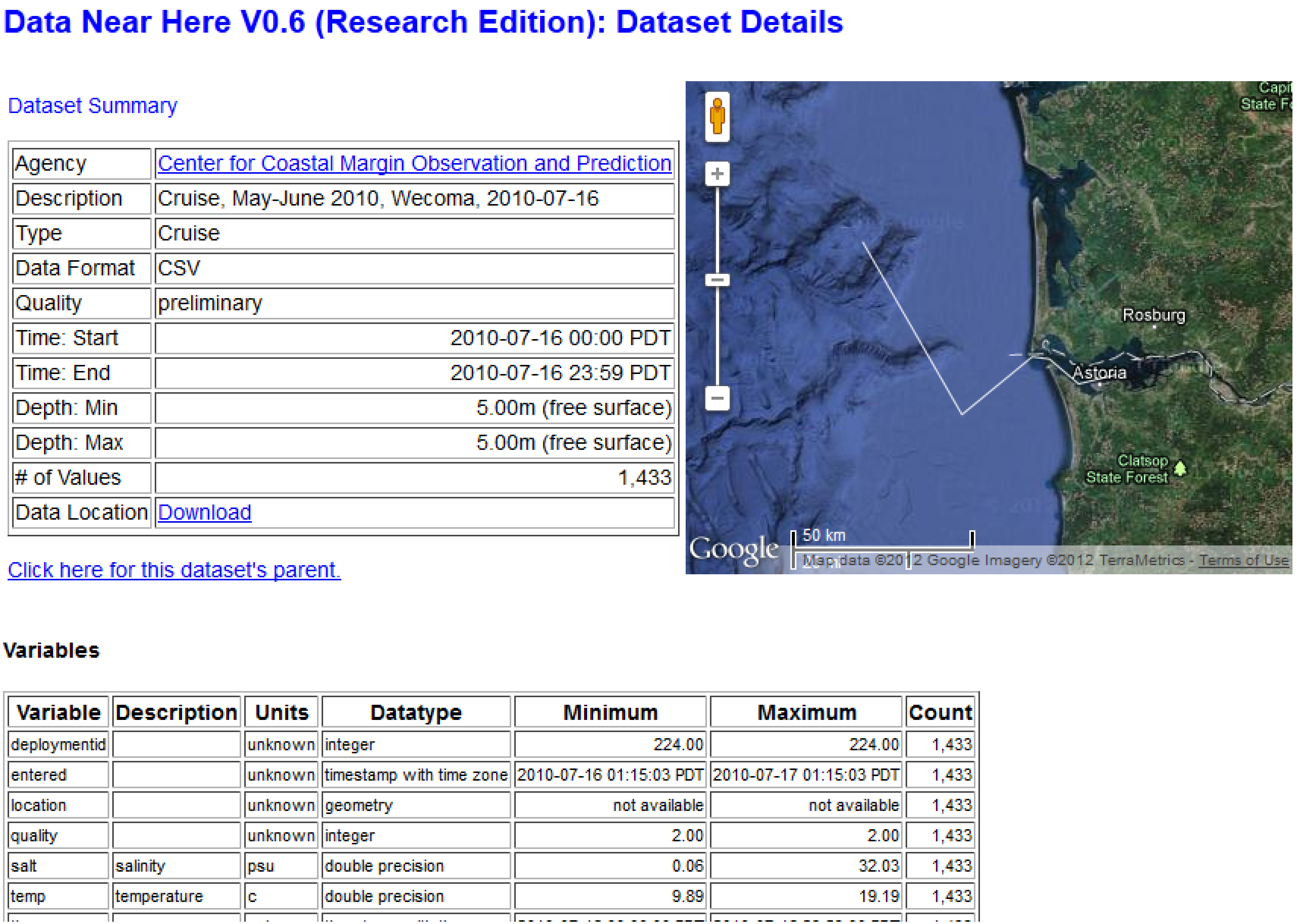


There were 50 results returned; all are listed, and 25 initially shown on map. Temp was found in 50 entries.



Display ☑	Type	Collection	Quality	Start Time	End Time	From Depth	To Depth	temp	Observations	Data Location	Score	DNH
1 ☑	Cruise	Cruise, May-June 2010, Wecoma, 2010-07-16, Segment 3	preliminary	2010-07-16 05:16 PDT	2010-07-16 05:29 PDT	-5	-5	9.89:12.14 c	14	Download	97	DNH
2 ☑	Cruise	Cruise, April 2010, Wecoma, 2010-04-17,	preliminary	2010-04-17 04:06 PDT	2010-04-17 04:26 PDT	-5	-5	10.60:10.85 c	21	Download	96	DNH

Prototype: Dataset Details Page



Data Near Here V0.6 (Research Edition): Dataset Details

Dataset Summary

Agency	Center for Coastal Margin Observation and Prediction
Description	Cruise, May-June 2010, Wecoma, 2010-07-16
Type	Cruise
Data Format	CSV
Quality	preliminary
Time: Start	2010-07-16 00:00 PDT
Time: End	2010-07-16 23:59 PDT
Depth: Min	5.00m (free surface)
Depth: Max	5.00m (free surface)
# of Values	1,433
Data Location	Download

Click here for this dataset's parent.

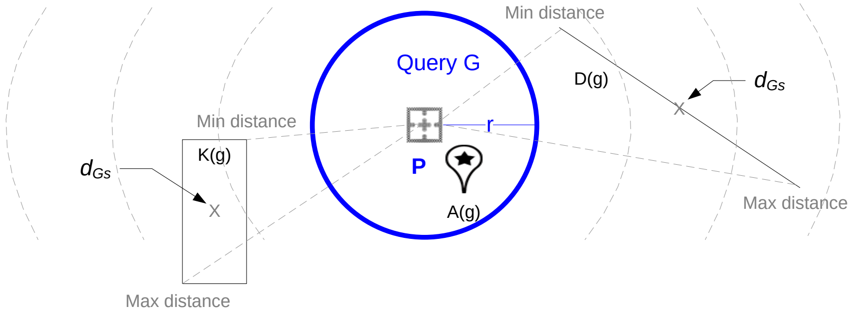
Capi
State F
Rosburg
Astoria
Clatsop
State Forest
Google
50 km
20 mi
Map data ©2012 Google Imagery ©2012 TerraMetrics - Terms of Use

Variables

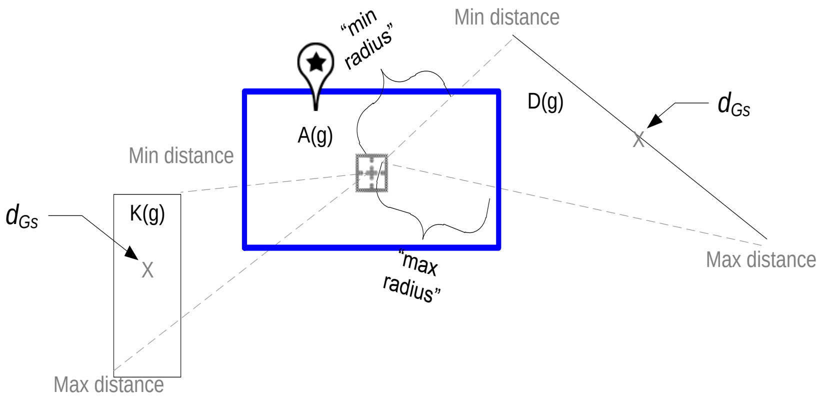
Variable	Description	Units	Datatype	Minimum	Maximum	Count
deploymentid		unknown	integer	224.00	224.00	1,433
entered		unknown	timestamp with time zone	2010-07-16 01:15:03 PDT	2010-07-17 01:15:03 PDT	1,433
location		unknown	geometry	not available	not available	1,433
quality		unknown	integer	2.00	2.00	1,433
salt	salinity	psu	double precision	0.06	32.03	1,433
temp	temperature	c	double precision	9.89	19.19	1,433

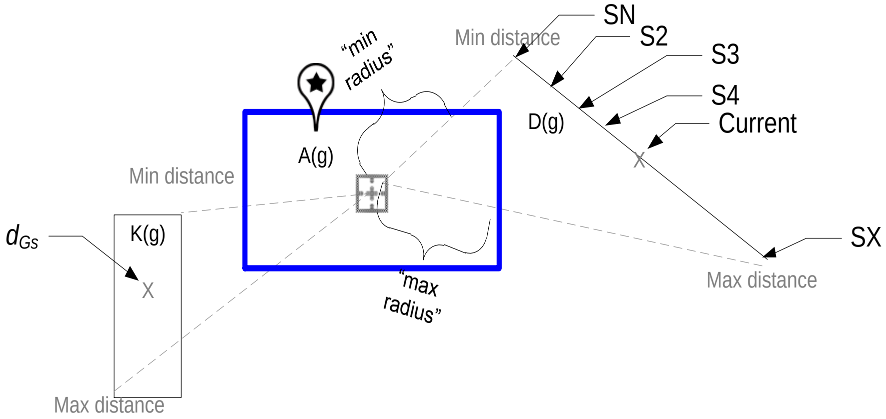
Prototype: Scoring Datasets

- “Current”: Spatial distance is approximated by:
  - ½ ((min distance)/radius + (max distance)/ radius)
  - Apply scoring function to the result



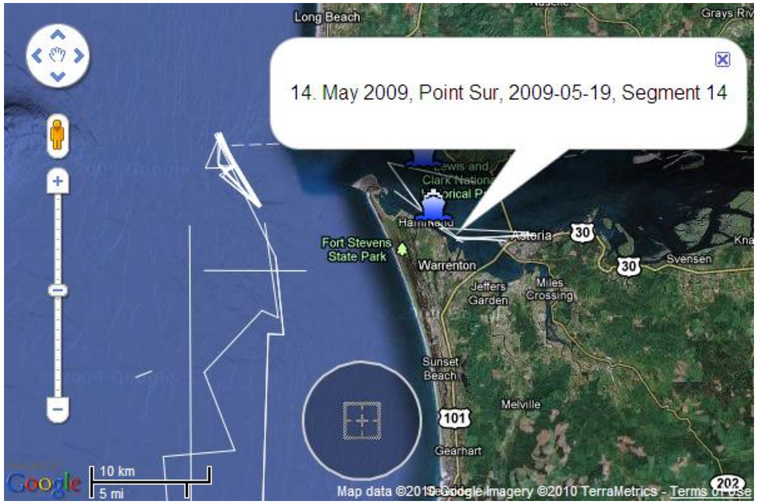
- Alternate rankings vary weighting of min and max

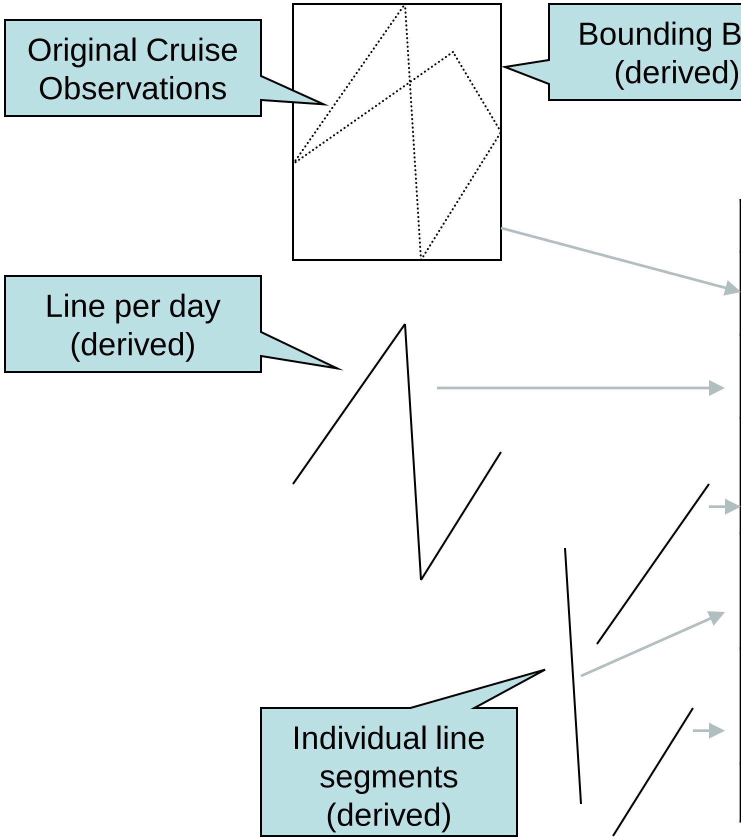




Prototype: Creating Metadata: Space

- A complex, multi-week cruise track; >1 million observations
  - Process: Extract bounding box, polylines, lines
  - Result: a small set of metadata records





Original Cruise Observations

Bounding Box (derived)

Line per day (derived)

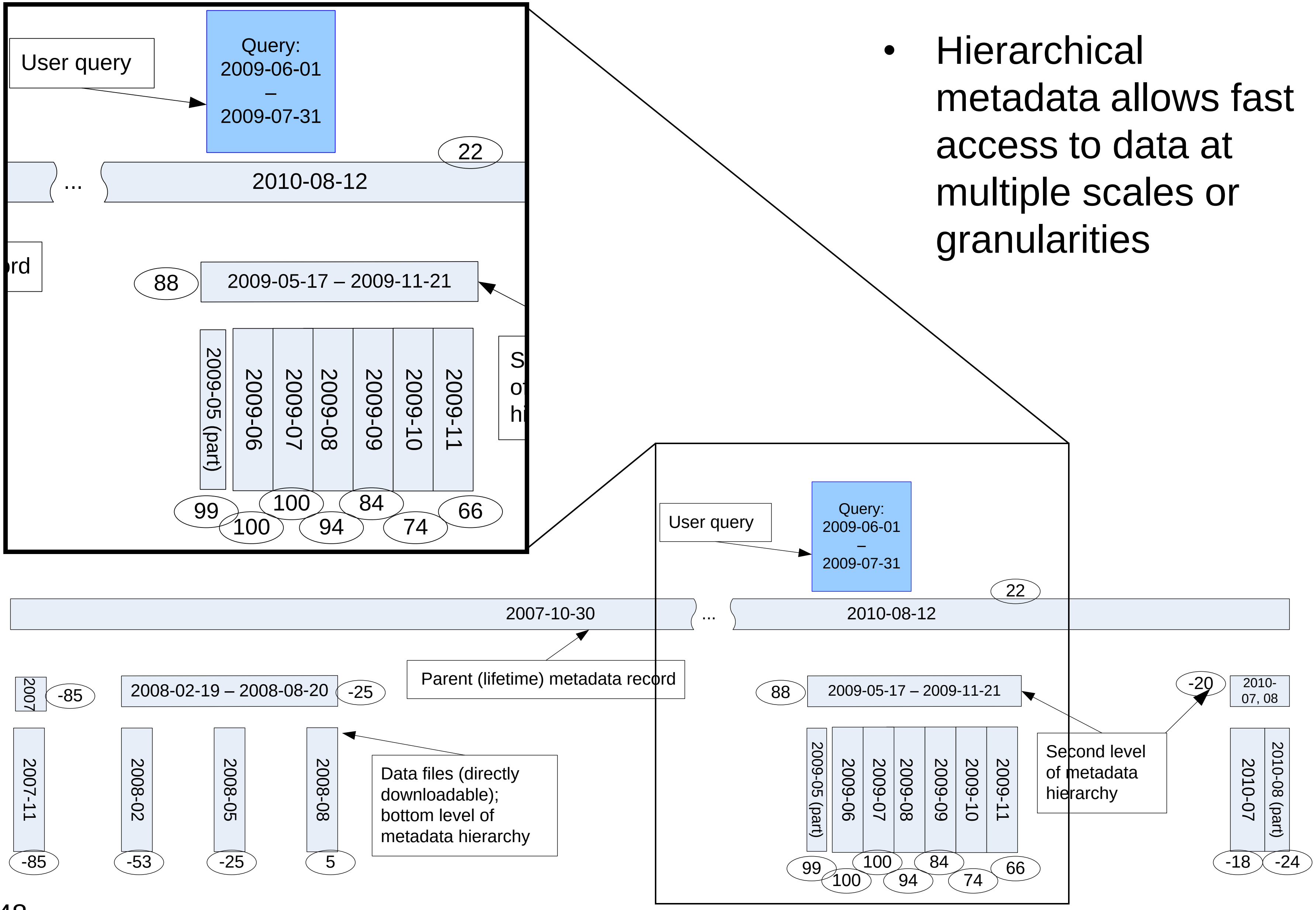
Individual line segments (derived)

DNH Metadata Table

	Geometry	Min. Time	Max. Time	Parent
May 2009, Point Sur	Polygon [bounding box]	5/13/2009	5/25/2009	<null>
May 2009, Point Sur, 2009-05-19	Line(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
….				

Prototype: Scoring using Hierarchical Metadata

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


User query
Query: 2009-06-01 – 2009-07-31
22
...
2010-08-12
rd
88
2009-05-17 – 2009-11-21
2009-05 (part)
2009-06
2009-07
2009-08
2009-09
2009-10
2009-11
S
of
hi
99
100
100
84
94
74
66
User query
Query: 2009-06-01 – 2009-07-31
22
2007-10-30
...
2010-08-12
Parent (lifetime) metadata record
2007
-85
2008-02-19 – 2008-08-20
-25
88
2009-05-17 – 2009-11-21
-20
2010-07, 08
2007-11
2008-02
2008-05
2008-08
Data files (directly downloadable); bottom level of metadata hierarchy
2009-05 (part)
2009-06
2009-07
2009-08
2009-09
2009-10
2009-11
Second level of metadata hierarchy
2010-07
2010-08 (part)
-85
-53
-25
5
99
100
100
94
84
74
66
-18
-24