



Data Discovery using Digests

Peter Mork

703-983-1465

pmork@mitre.org

MSR

Problem

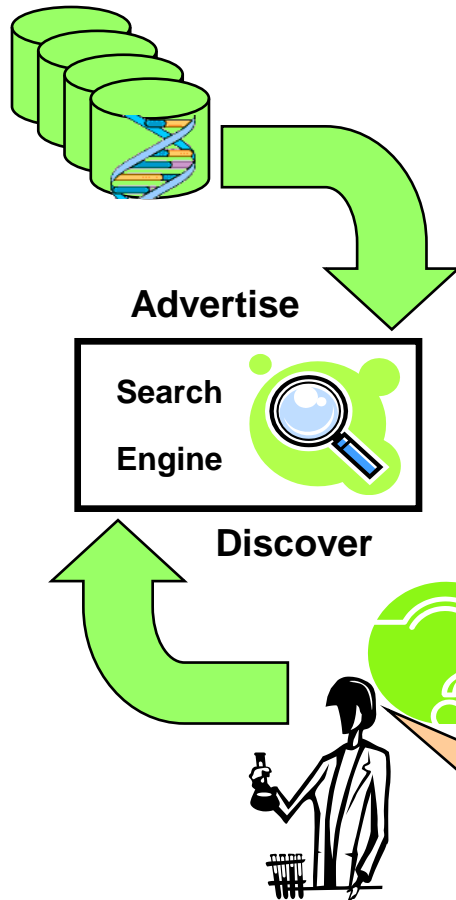


There is currently no efficient way to sift through **large numbers of **structured** data resources with **sensitive** content to identify relevant data.**

Existing efforts:

- Assume unstructured hyperlinked text (e.g., Google), or**
- Rely on manually generated data catalogs (e.g., using Department of Defense Metadata Specification)**

Background

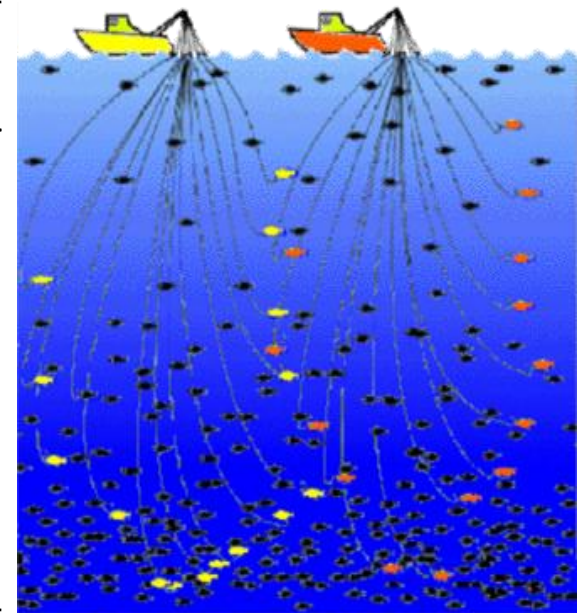


Surface web

- Keyword queries common

Deep web

- Protected content
- Non-textual
- Highly structured
- >500x larger
- Keyword queries often not applicable



How many patients with Alzheimer's are there within 50 miles of me between the age of 55 and 65?

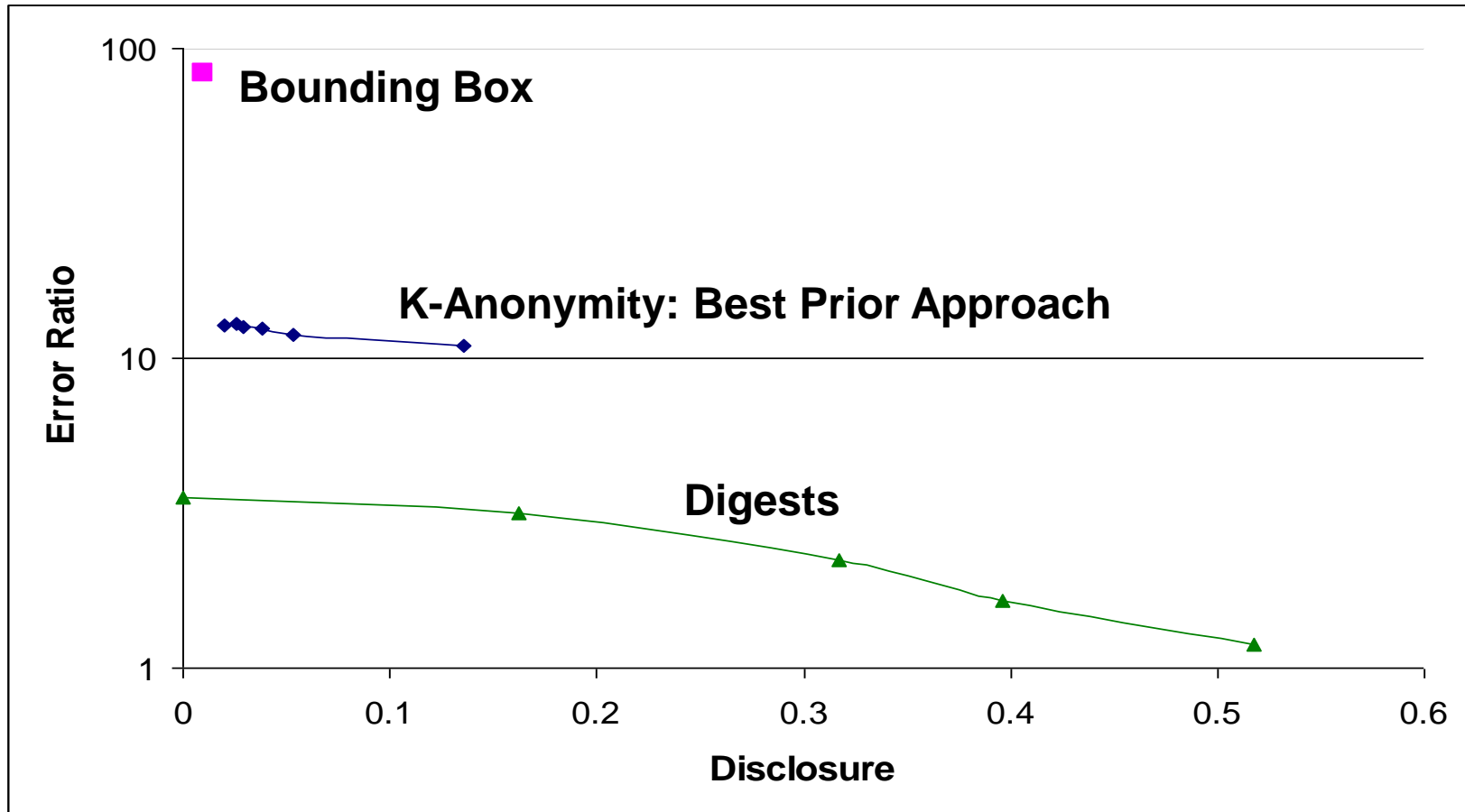
Objective

Develop techniques to **automatically** summarize structured data to allow more accurate **discovery** while minimizing **disclosure** of sensitive information

- Adapt research on data summarization and data privacy
- Develop tools to help data providers and consumers
- Publish research results and share findings across MITRE
- Demonstrate scope of problem to vendors
- Transition techniques to sponsors

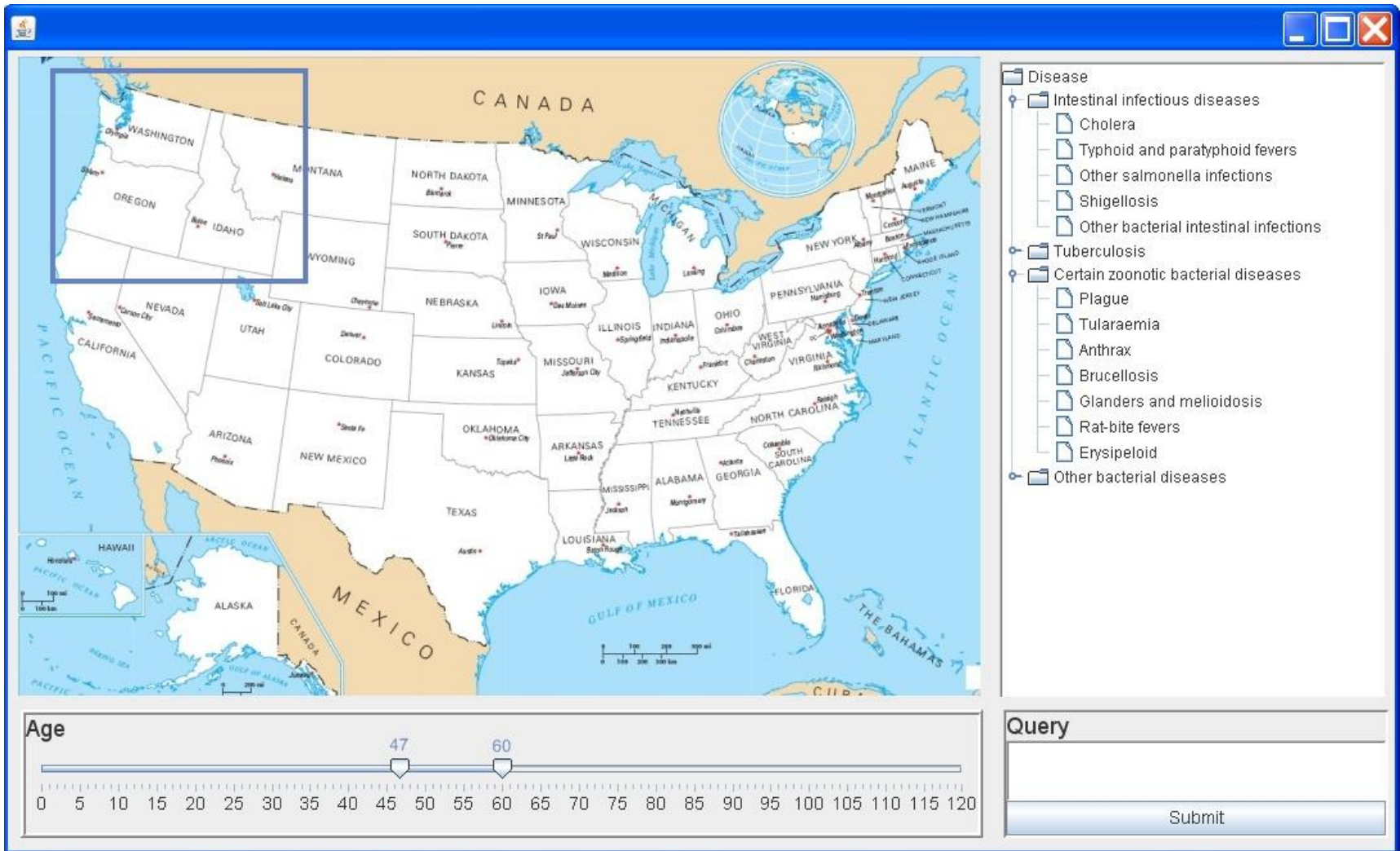
- **Datatype-specific summarization strategies for:**
 - Numeric data (e.g., patient age)
 - Geospatial data (e.g., patient address)
 - Hierarchically coded data (e.g., diagnosis)
- **Quantifying tradeoff between**
 - Disclosure: advertising
 - Error: discovery
- **Use case tailored to medical records**

Highlight: Error vs. Disclosure



$$\frac{\Pr(e_a = v | e_q \wedge D) - \Pr(e_a = v | e_q)}{1 - \Pr(e_a = v | e_q)}$$

Demonstration: Medical Records



The screenshot shows a web-based application interface for medical records. The main component is a map of the United States with a blue rectangular box highlighting the Pacific Northwest region, specifically Washington, Oregon, and Idaho. To the right of the map is a sidebar menu titled "Disease" with a tree structure of categories and sub-items:

- Disease
 - Intestinal infectious diseases
 - Cholera
 - Typhoid and paratyphoid fevers
 - Other salmonella infections
 - Shigellosis
 - Other bacterial intestinal infections
 - Tuberculosis
 - Certain zoonotic bacterial diseases
 - Plague
 - Tularaemia
 - Anthrax
 - Brucellosis
 - Glanders and melioidosis
 - Rat-bite fevers
 - Erysipeloid
 - Other bacterial diseases

Below the map is an "Age" slider control. The slider is labeled "Age" and has a scale from 0 to 120 in increments of 5. Two diamond-shaped markers are positioned on the slider at the values 47 and 60. To the right of the slider is a "Query" input field, which is currently empty. Below the input field is a "Submit" button.

Impacts

- **National Center for Research Resources (NIH) project to track “monkey census”:**
 - Sensitive data (due to PETA threats)
 - “Who has three 5-10 yr. old female Macaque available?”
- **“In a recent proof of concept, MITRE presented the concept of the hierarchy rollup to **Informatica** ... to help **bridge the gap**, Informatica has made significant strides in the latest release of our software ...”**
- **Co-organizing **first international workshop on resource discovery****
- **Co-author of: “Biological Metadata Management and Resource Discovery” in Encyclopedia of Database Systems, Liu and Özsu (eds.)**

Future Plans (2nd half of FY'08)

