

Data Discovery Using Digests

Dr. Peter Mork

703-983-1465 • pmork@mitre.org

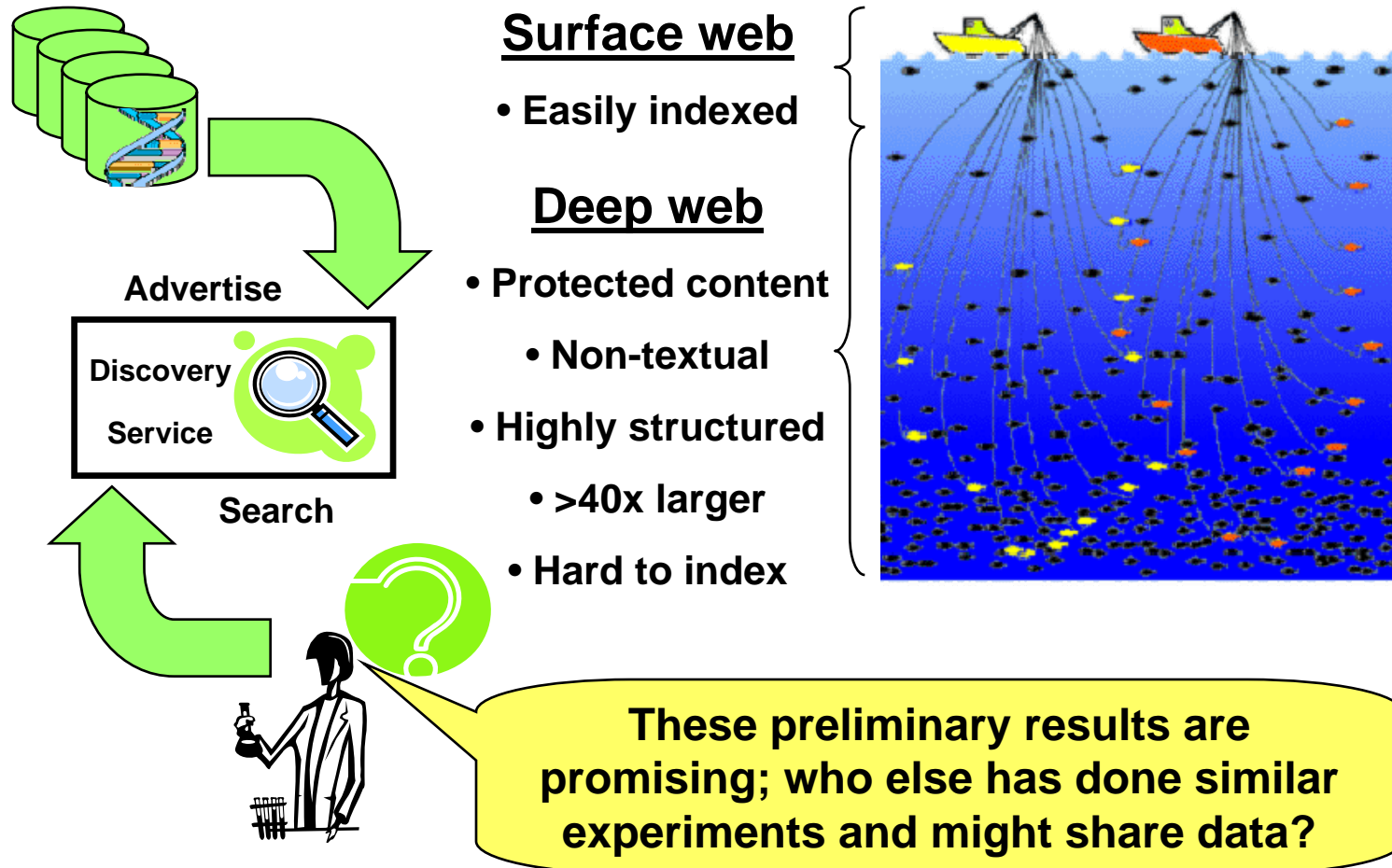
MITRE Sponsored Research



Problem

- Potential data consumers need ways to **search** for useful data resources
- Potential data providers need ways to **advertise** existence of data resources
- How can a **discovery** service:
 - Maximize precision and recall?
 - Maximize data privacy?
 - Minimize response time?

Background



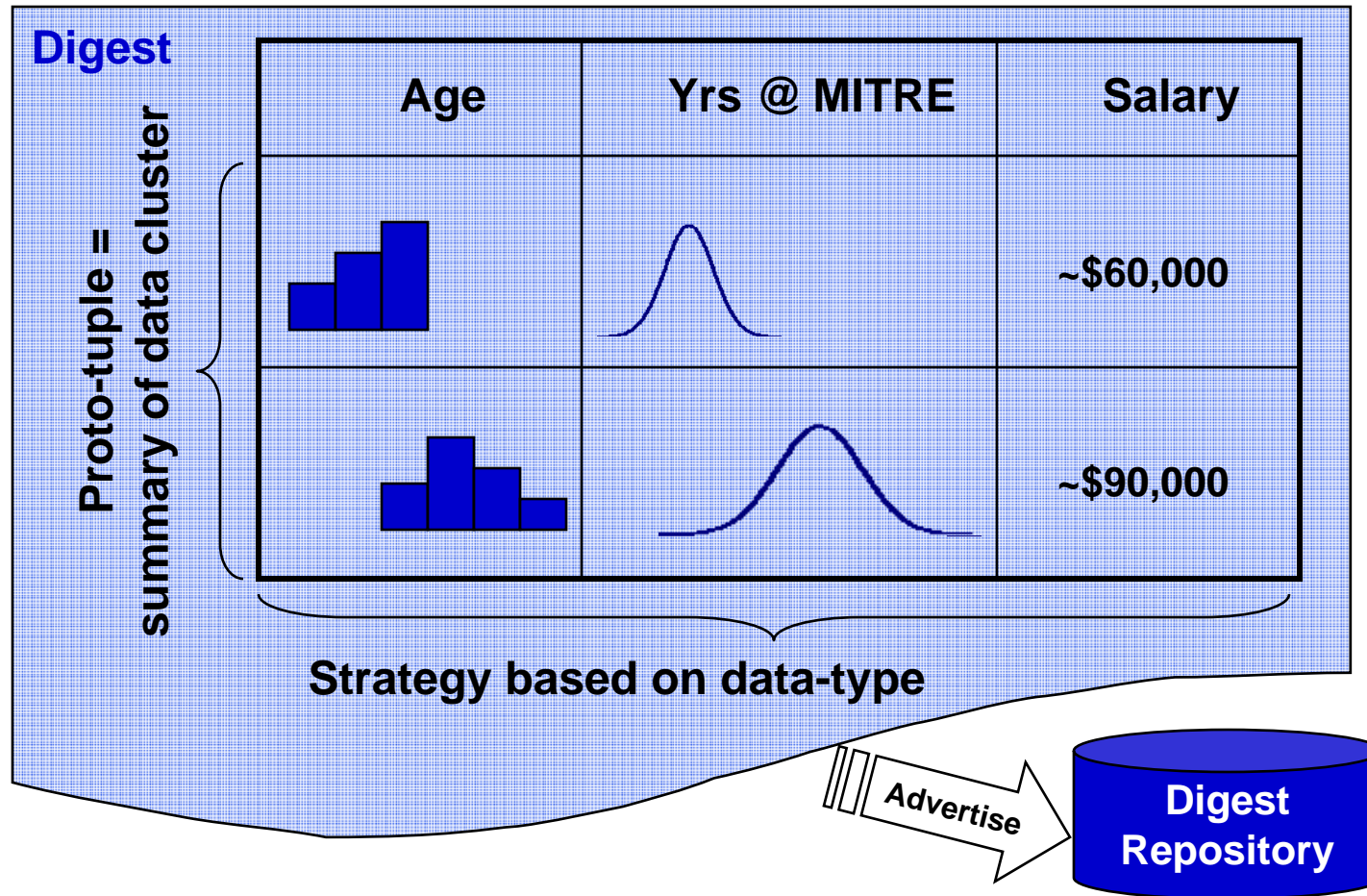
Objective

- **Adapt research on data summarization and data privacy to problem of data discovery**
- **Develop tools to help:**
 - **Data providers advertise their data**
 - **Data consumers search for data**
- **Publish research results and share findings across MITRE**
- **Transition tools to sponsors to facilitate increased data sharing**

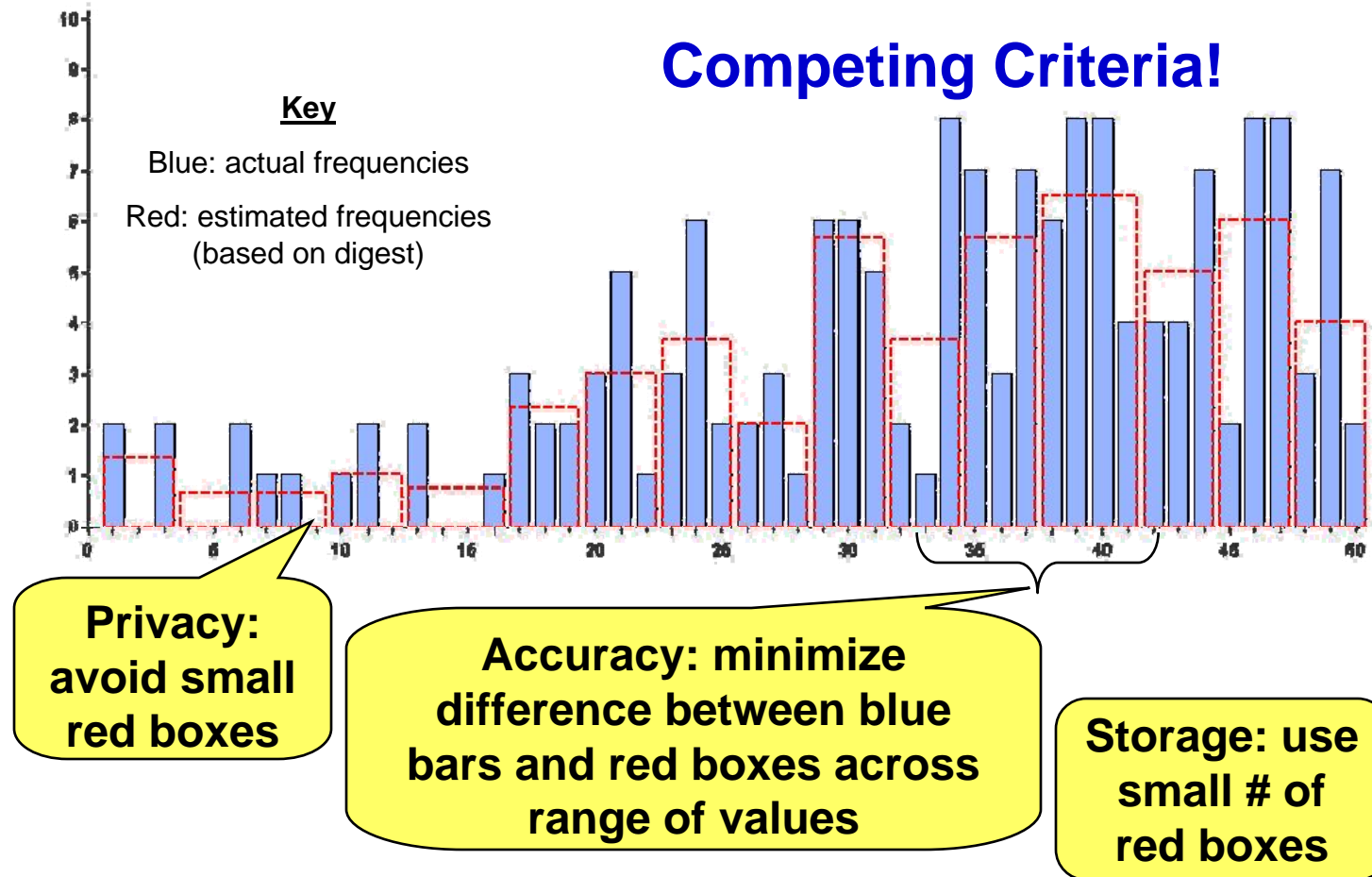
Activities

- Developing techniques for creating succinct database summaries (**digests**) that allow:
 - Accurate aggregate queries
 - Without exposing individual records
- Creating prototype of digest **repository** to support searching for data resources
- Developing language and tools for expressing and enforcing **privacy** constraints

Highlight



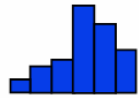
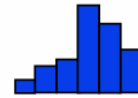
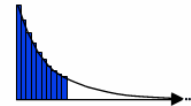

Demonstration



Impacts

- **Increases visibility of structured data resources buried behind access controls**
- **Extends search capabilities provided by metadata registries**
- **Addresses open academic research problem: how to balance sensitivity and accuracy**
- **Leverages MITRE's expertise in bridging database theory and practice**
- **Applicable to range of MITRE sponsors**

Future Plans

Attribute Type:	numeric	categorical	string	text
Characteristics:	continuous	discrete, enumerated	lexical tokens, zipfian	language
Examples:	lat/long event time age temperature frequency	country alert status state Fr/So/Jr/Sr diagnosis	person name place name book title menu item part name	description physicians notes comments news story
Summarization Strategies:	 histogram	 histogram	 zipfian	 bag of words
	Current Focus	Ongoing Research		Incorporate Prior Work