

Common Ground: Agile Data Sharing

Ken Smith

703-983-6115

kps@mitre.org

CEM MSR

Problem

- **Data sharing is often prohibitively expensive in time and labor**
- **Yet, many high-value data sharing activities are urgent or funding-constrained**
- **How can we “lower the activation energy” for data sharing**

Background

Producing exhaustive agreement
disallows vital sharing!

“Emergency Preparedness & Response”

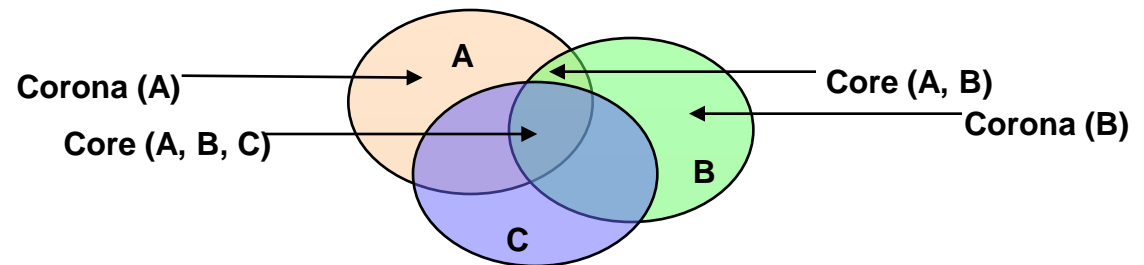
“Bad Guys Without Borders”

State law enf.
database coverage
stops here

*But criminals
don't!*



*But sharing can
start sooner ...*



Objectives



- **Rapidly identify a semantically important “core” set of attributes from the data models of sharing partners, thus**
 - Reducing the problem size
 - Focusing limited integration resources
 - Building trust and momentum by “getting in the game” sooner
- **Evaluate the sensitivity of each partner’s core instance values**

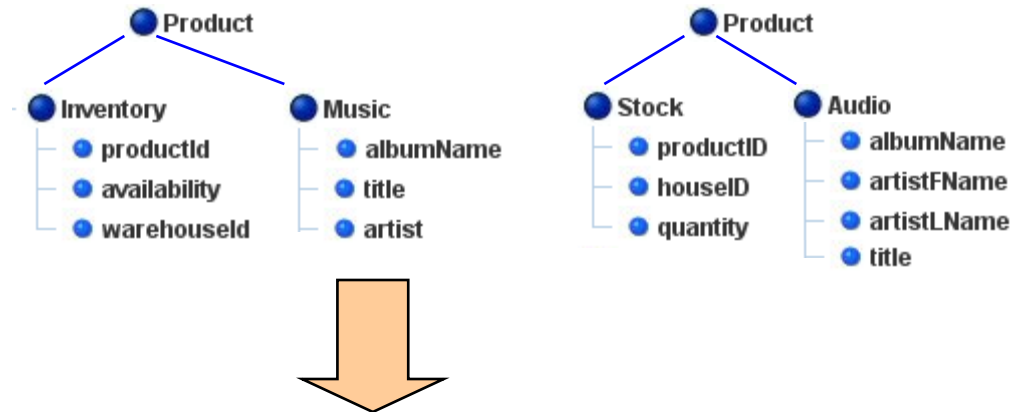
Activities

- **Toward a core data model**
 - **Enhance MITRE’s existing Harmony schema-matching tool for “industrial strength” matching problems**
 - **Develop a schema repository to manage data models and share knowledge about them (e.g., core model and known correspondences)**
 - **Develop schema clustering techniques to quickly suggest likely groups of common attributes**
- **Investigate *data sensitivity metrics* to assist data releasers once a sharable core is identified**

Highlight

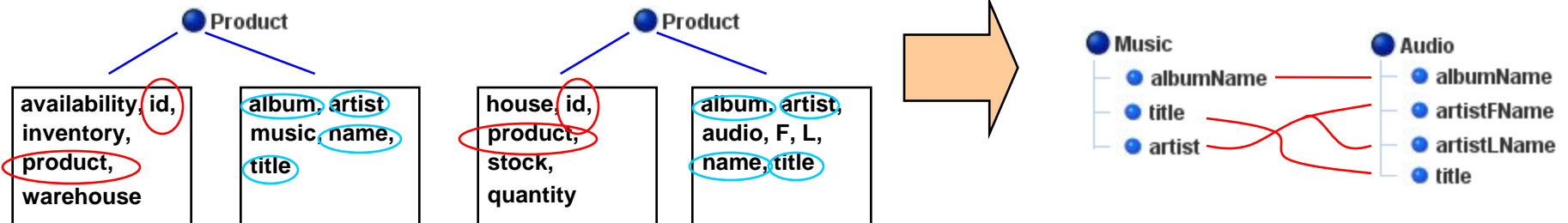
We want to match two similar, yet very large schemas; however their complexity makes conventional approaches too expensive

Fragment Matching



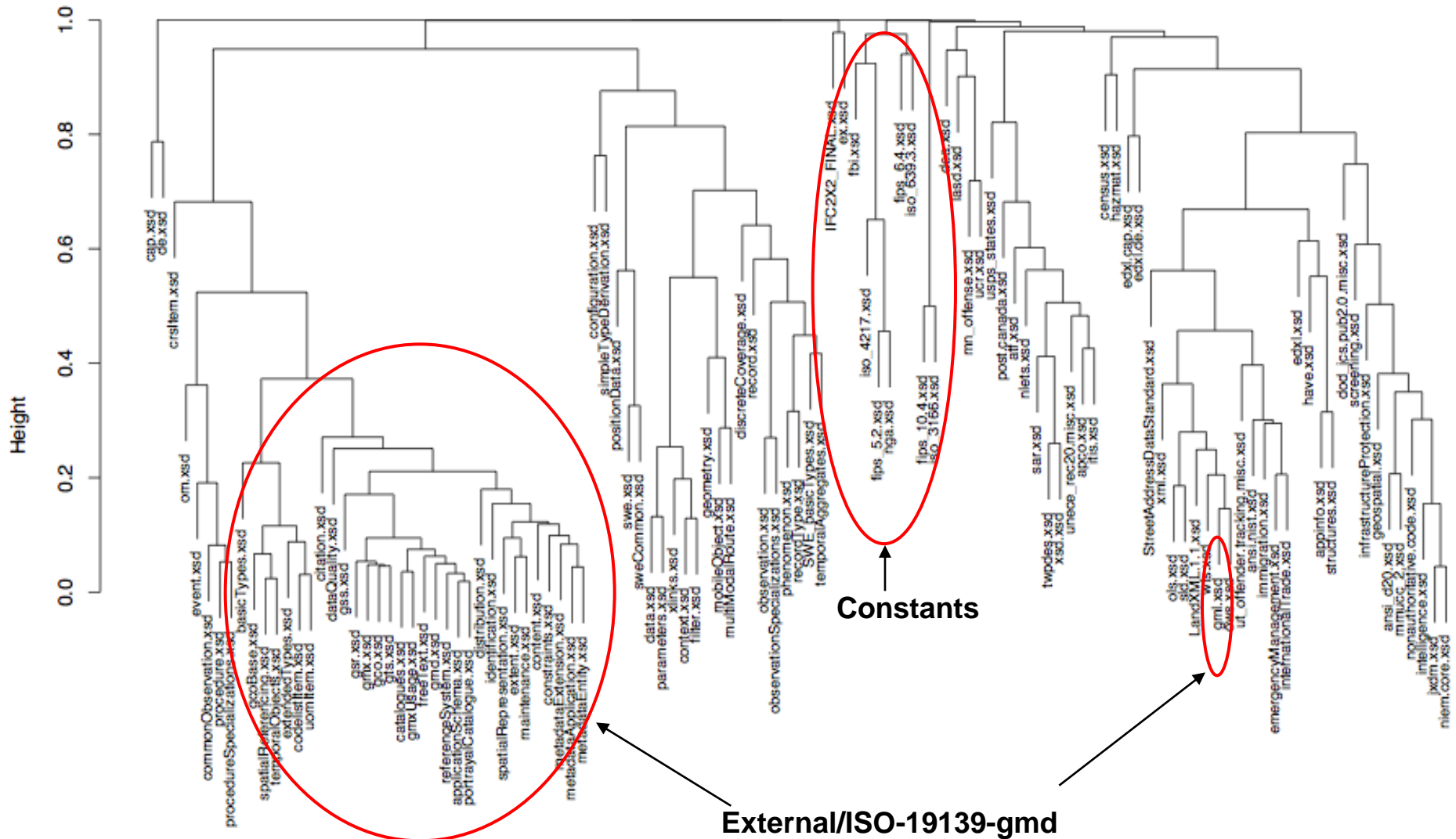
First, compute approximate matches using document similarity metrics (e.g., cosine distance)

Then, perform complex, expensive matches only on similar fragments



Demonstration

Schema Clustering Dendrogram: 115 NIEM Schemas; 19 Seconds



Impacts



- **Help MITRE’s sponsors who must share high value data under pressing time and money constraints by:**
 - **Developing tools for matching, managing, and visualizing the “semantic topology” of a set of data models**
 - **Providing a deeper understanding of sensitivity (e.g., privacy) evaluation in structured data sets**
 - **Influencing research and vendor activities in these areas**

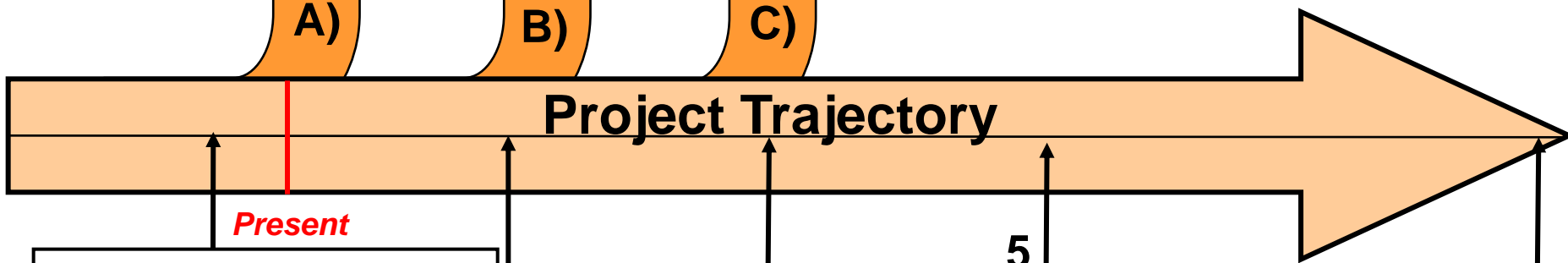
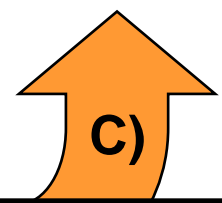
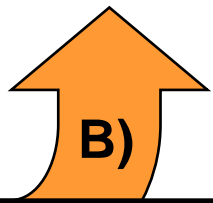
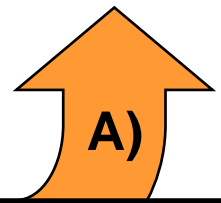
Future Plans

Develop a general and open schema repository system for data fusion

Schema clustering research

Early transition opportunities

Opportunities



Present

1 Groundwork: restart and solidify Harmony 1.0; reengineer it to scale to non-trivial binary matches

2 Develop new matching strategies (e.g., fragment match) suited to large-scale matching tasks

3 Determine top priority for data sensitivity research

Planned Tasks

4 Rapidly perform an N-way match on N non-trivial schemas

5 Turn the result of an N-way match into the core data model of N non-trivial schemas

6 Transition tools and published methodologies to

- Rapidly compute the core data model of N non-trivial schemas
- Rapidly evaluate the sensitivity of pre-shared data