Dataspaces: A New Abstraction for Data Management

Mike Franklin, Alon Halevy,
David Maier, Jennifer Widom

Today’s Agenda

- Why databases are great.
- What problems people really have
  - Why databases are not great.
- Data integration and sharing:
  - Nice, but doesn’t address all the problem.
- Dataspaces:
  - Initial concepts, a note on politics
  - Research challenges

Databases Are Great

- Very clean abstraction for data management.
- High-level querying with efficient query processing.
- Strong guarantees. Your data will survive anything.
- *Put your data in the database*, and your worries will go away.

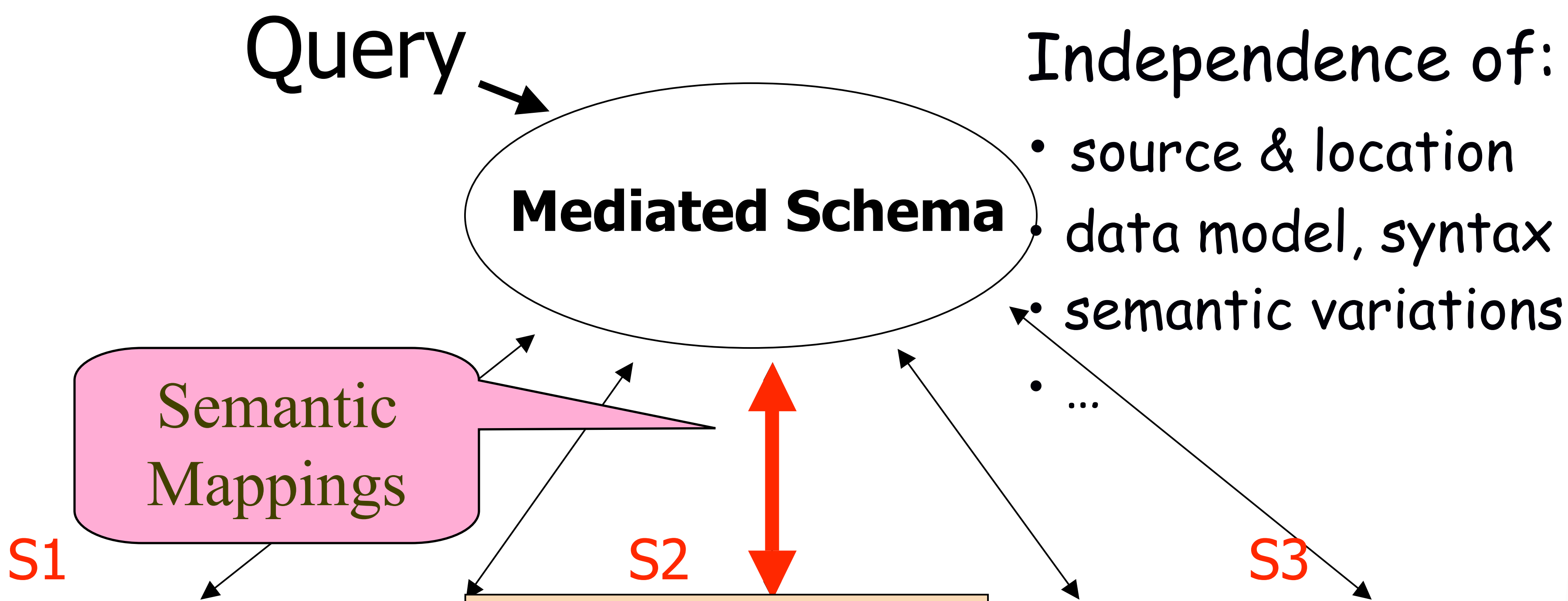
Today’s DM Challenges

- A set of inter-related data sources:
  - The enterprise
  - Large science projects
  - Government agencies
  - The battlefield
  - The desktop (and its extensions)
  - A library
  - The ‘smart’ home
- We’ve heard this before. What’s new?

A Quick History of Data Integration

- Until late 90’s:
  - Integration by warehousing
  - Integration by custom code
- Late 90’s (boom years):
  - Virtual data integration (data stays at the source, queried on the fly)
  - Nimble, Cohera and others.
  - EII (Enterprise Information Integration): new buzzword. Still buzzing now too.

Virtual Data Integration



Independence of:

- source & location
- data model, syntax
- semantic variations
- …

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	…	…

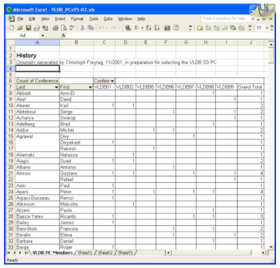
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

. .

```
<cd>    <title> The best of … </title>
             <artist> Carreras
          </artist>
             <artist> Pavarotti
          </artist>
             <artist> Domingo
          </artist>
             <price> 19.95      </price>
        </cd>
```

. . .

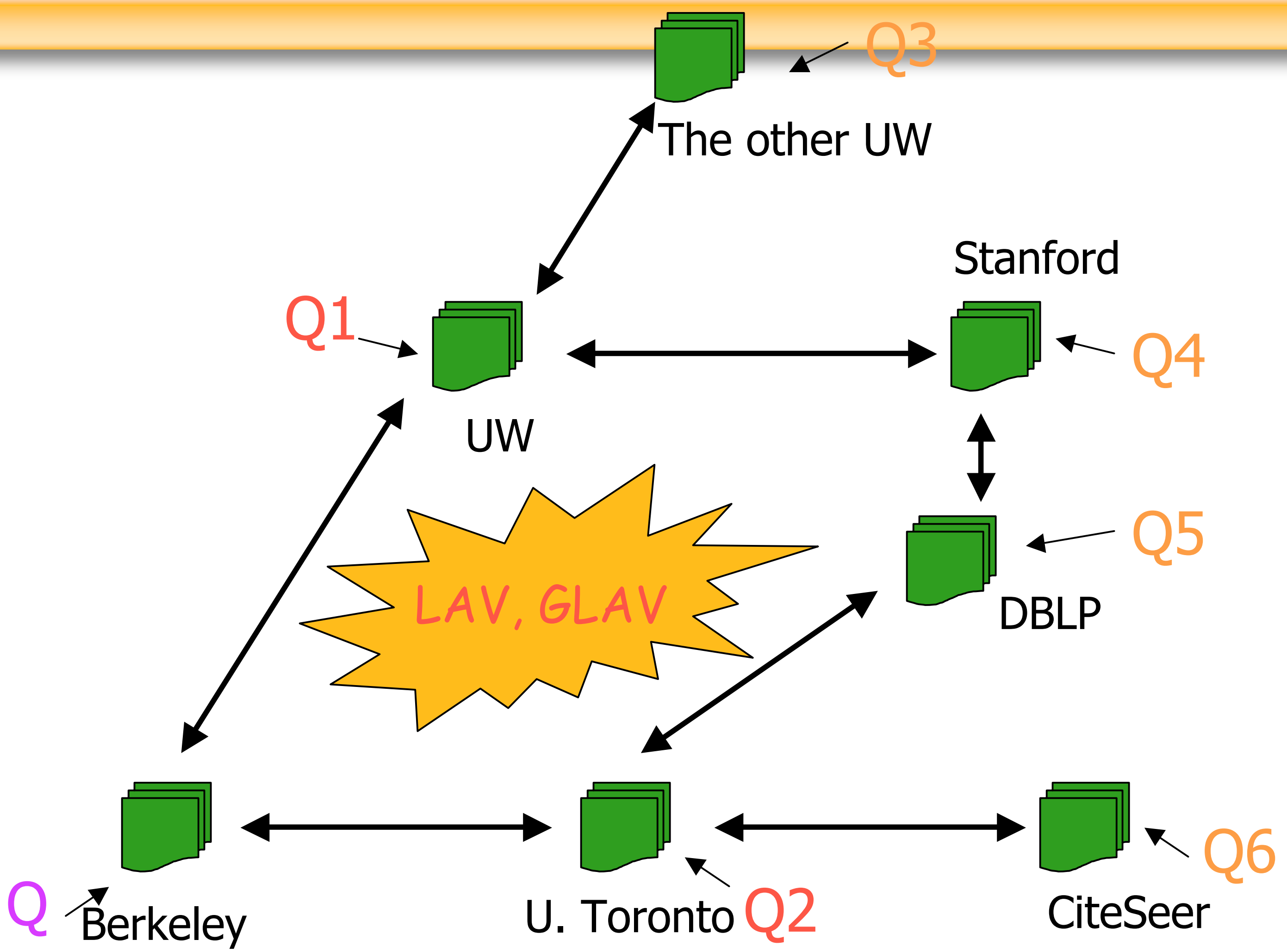

Microsoft Excel - VLDB_PCs95-02.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
A4
VLDB PC Members
Sheet1
Sheet2
Sheet3
Ready

History

Originally generated by Christoph Freytag, 11/2001, in preparation for selecting the VLDB '03 PC

Count of Conference		Confere							
Last	First	VLDB01	VLDB02	VLDB95	VLDB96	VLDB97	VLDB98	VLDB99	Grand Total
Abbadi	Amr-El						1		1
Abel	David							1	1
Aberer	Karl	1	1						2
Abiteboul	Serge				1			1	2
Acharya	Swarup							1	1
Adelberg	Brad						1		1
Adiba	Michel			1	1				2
Agrawal	Divy					1			1
	Divyakant	1							1
	Rakesh			1					1
Ailamaki	Natassa		1						1
Alagic	Suad	1	1						2
Albano	Antonio				1				1
Alonso	Gustavo	1	1			1		1	4
	Rafael							1	1
Aoki	Paul	1							1
Apers	Peter	1		1		1	1		4
Arpaci-Dusseau	Remzi	1							1
Atkinson	Malcolm		1						1
Atzeni	Paolo						1		1
Baeza-Yates	Ricardo	1				1	1		3
Bailey	James	1							1
Bancilhon	Francois				1		1		2
Baralis	Elena				1			1	2
Barbara	Daniel						1		1
Barga	Roger	1							1

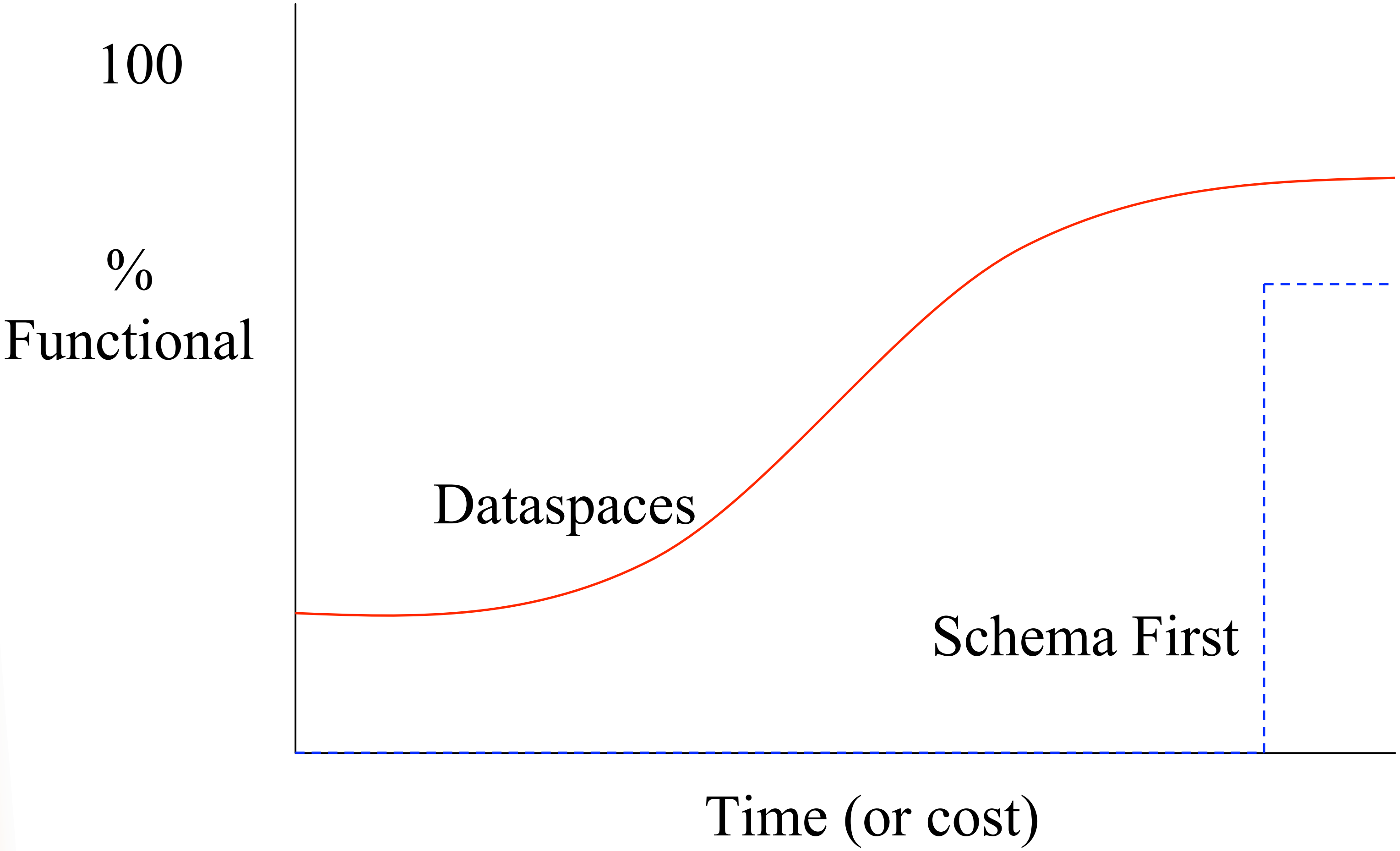
Peer Data Management Systems



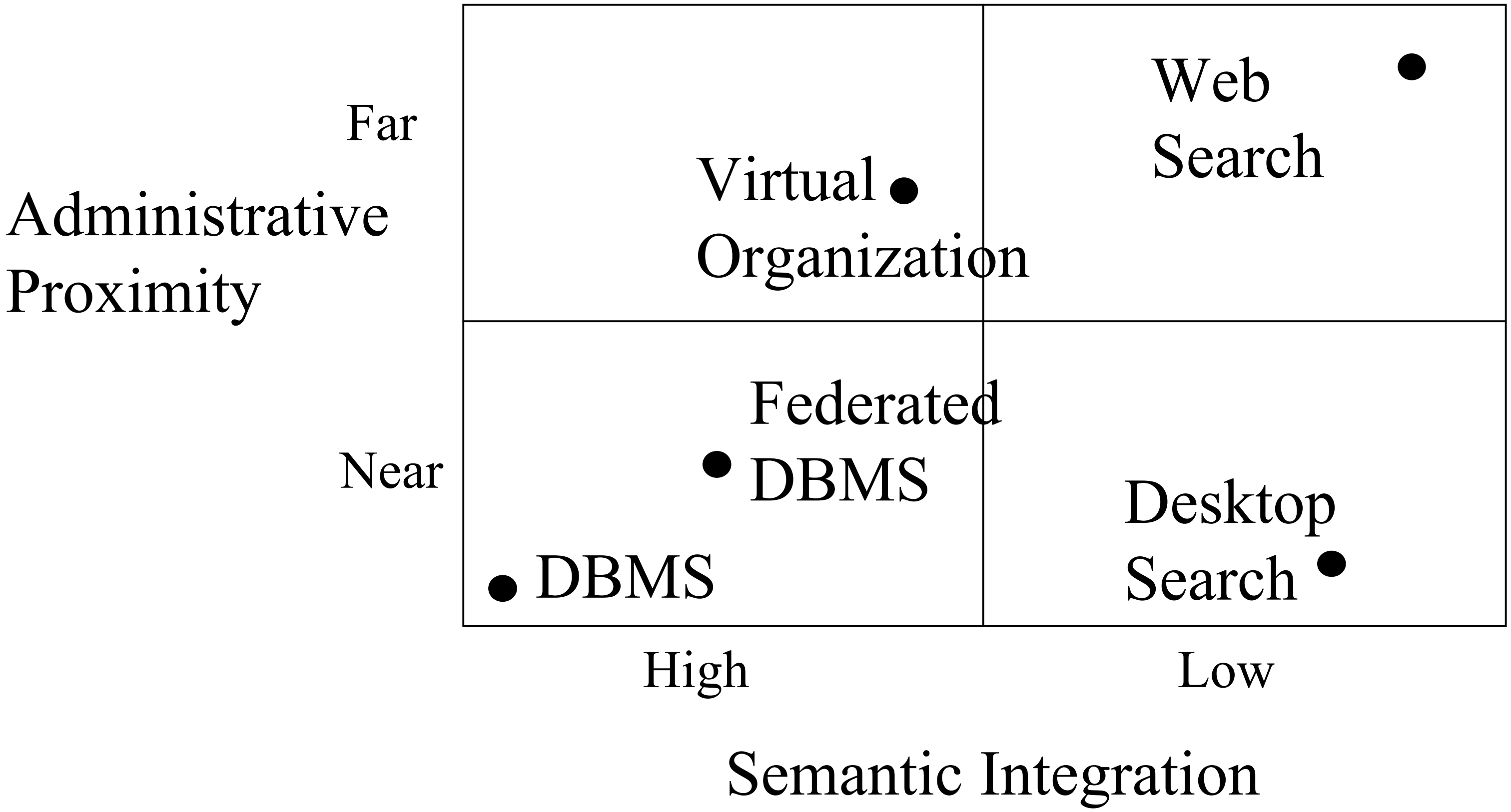
DI: Nice but Limited

- Still thinking about it like DB people.
- You can only manage data if it is:
  - Explicitly put in the database (or some source)
  - Fully mapped to the mediated schema.
- Upfront cost is too high:
  - Benefits not always clear at the outset.

Mike’s First Figure



Mike’s Second Figure



Far

Administrative Proximity

Near

Virtual ● Organization

Web Search ●

● Federated DBMS

● DBMS

Desktop Search ●

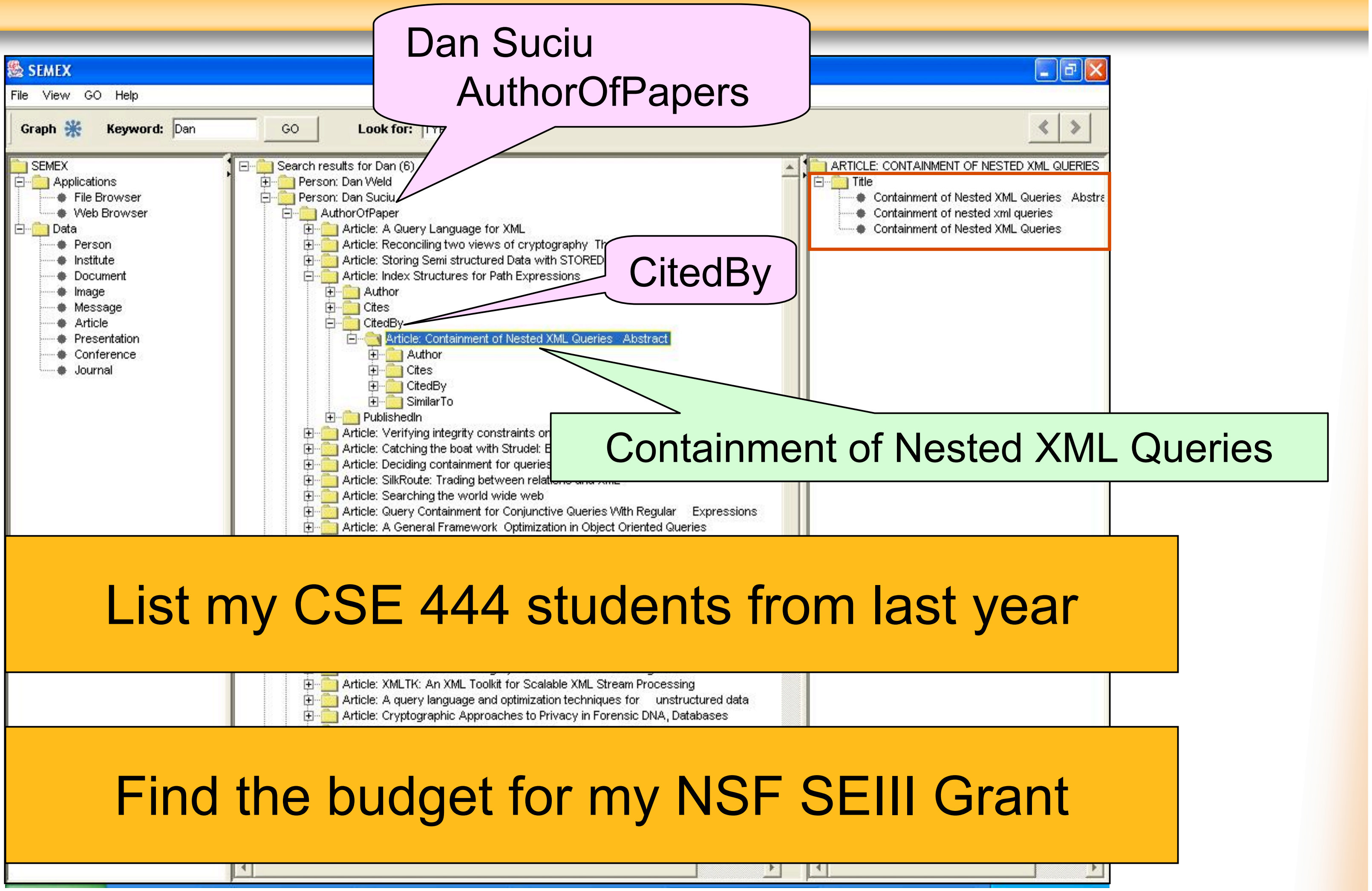
High

Low

Semantic Integration

Bernstein's Story

The Desktop



```
SEMEX
  Applications
    File Browser
    Web Browser
  Data
    Person
    Institute
    Document
    Image
    Message
    Article
    Presentation
    Conference
    Journal
```

```
Search results for Dan (6)
  Person: Dan Weld
  Person: Dan Suciu
    AuthorOfPaper
      Article: A Query Language for XML
      Article: Reconciling two views of cryptography  Th
      Article: Storing Semi structured Data with STORED
      Article: Index Structures for Path Expressions
        Author
        Cites
        CitedBy
          Article: Containment of Nested XML Queries  Abstract
            Author
            Cites
            CitedBy
            SimilarTo
        PublishedIn
      Article: Verifying integrity constraints or
      Article: Catching the boat with Strudel: E
      Article: Deciding containment for queries
      Article: SilkRoute: Trading between relat
      Article: Searching the world wide web
      Article: Query Containment for Conjunctive Queries With Regular  Expressions
      Article: A General Framework  Optimization in Object Oriented Queries
      Article: XMLTK: An XML Toolkit for Scalable XML Stream Processing
      Article: A query language and optimization techniques for  unstructured data
      Article: Cryptographic Approaches to Privacy in Forensic DNA, Databases
```

```
ARTICLE: CONTAINMENT OF NESTED XML QUERIES
  Title
    Containment of Nested XML Queries  Abstra
    Containment of nested xml queries
    Containment of Nested XML Queries
```

Dan Suciu
AuthorOfPapers

CitedBy

Containment of Nested XML Queries

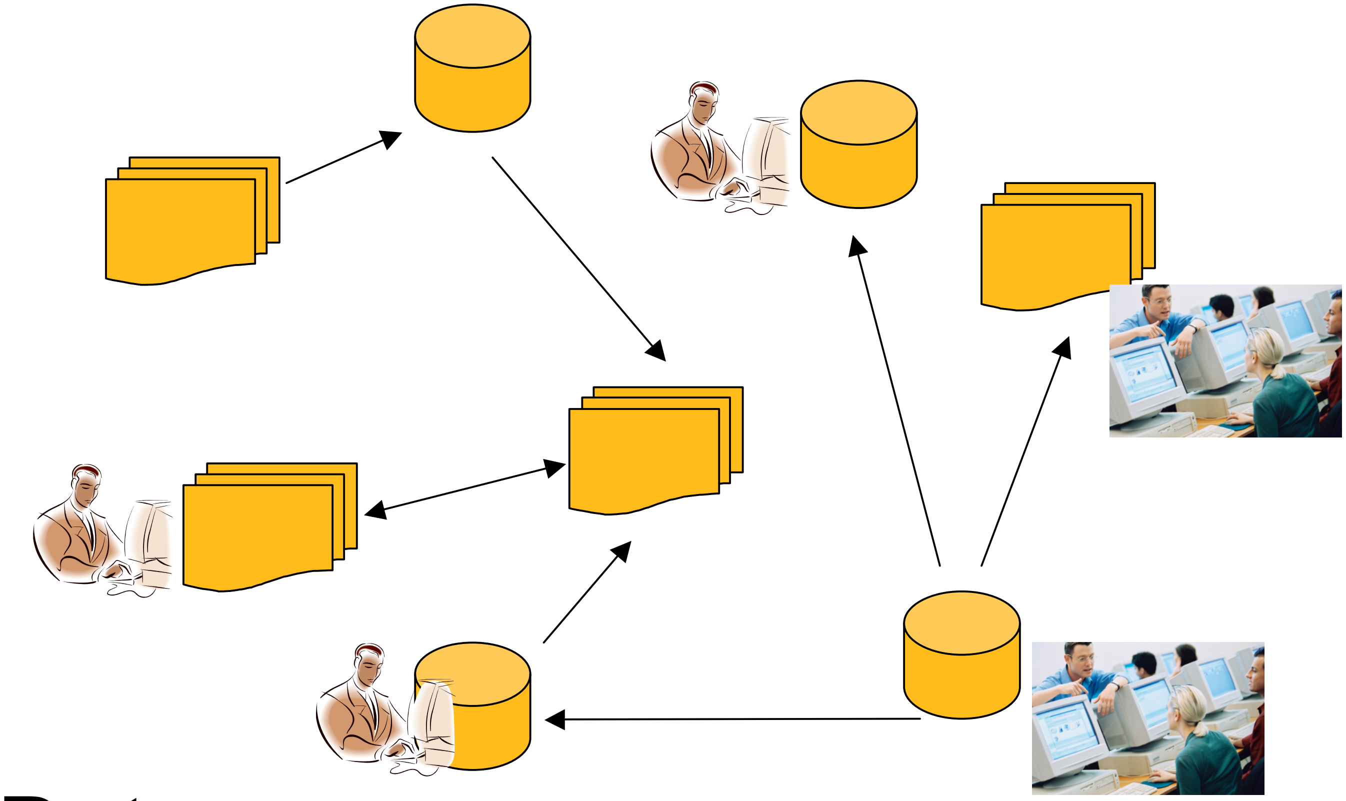
List my CSE 444 students from last year

Find the budget for my NSF SEIII Grant

(Big) Science

Find the experiments run an hour before the SIGMOD deadline.
What *were* we thinking?

Alon’s First Figure



A Dataspace

Participants: Examples

- Structured databases (relational, XML)
- Files of various applications
- Code collections
- Web services, software packages
- Sensors

- Different query capabilities
- Some updateable, others not
- Some more structured than others
- May stream

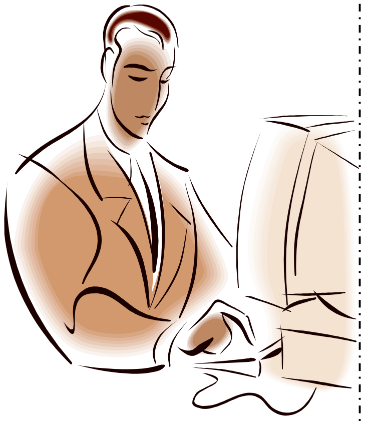
Relationships: Examples

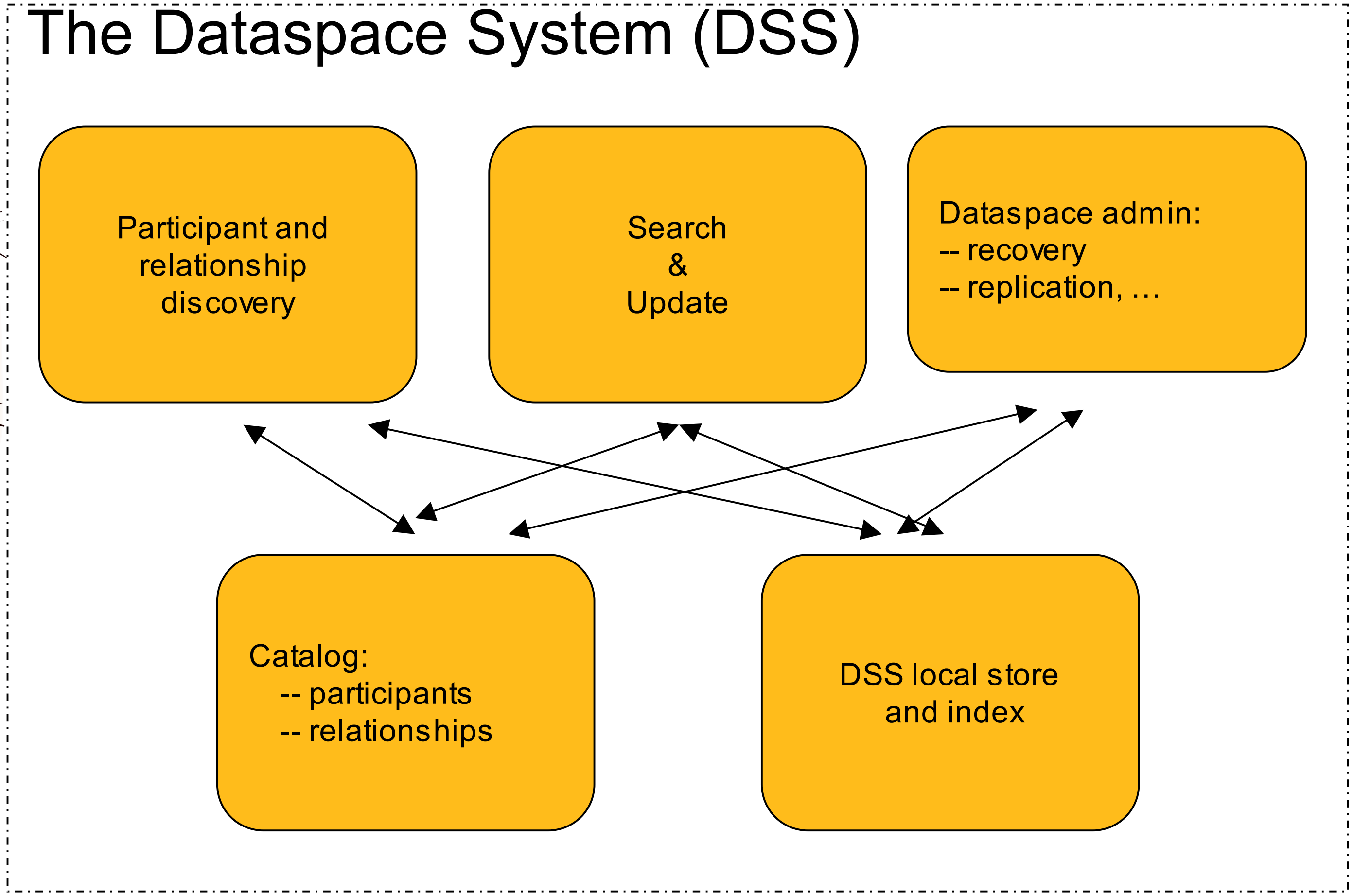
- Full schema mappings
  - E.g., views of each other, replicas
- *A* was manually created from *B* and *C*
- *A* is a snapshot of *B* on a certain date
- *A* and *B* reflect the same underlying physical entity (but are different)
- *A* was sent to me at the same time as *B*.

Dataspace Services

- Search & query: on data, schema, meta-anything.
  - Query lineage, hypothetical queries, …
- Mining.
- Set up workflows.
- Monitoring for special events.
- Soft constraints, recovery, consistency, …

Alon’s Second Figure





The Dataspace System (DSS)

Participant and relationship discovery

Search & Update

Dataspace admin:
-- recovery
-- replication, …

Catalog:
-- participants
-- relationships

DSS local store and index

A Note on Politics

- RDBMS have been a great identity
  - But has it served its purpose?
  - We’ve moved on, but the external perception hasn’t.
  - Too much alcohol served at CIDR.
- Dataspaces could be a new identity
  - 80% of our work is already on it anyway
  - Some exciting new problems (next)
  - *“Because that’s the size of the problem”*

Challenges: Search/Query

- What does search mean over a heterogeneous collection? Ranking?
- Answer queries despite schema heterogeneity and with no mappings.
- Support spectrum of search to query
  - Given keywords, identify what db may be relevant.
- No single data model, not even mediated.

Challenges: Lineage and Uncertainty

- When everything is fluffy, life is uncertain.
- Need to model:
  - Uncertainty and lineage *and* the relationship between them.
  - Hypothetical queries.
  - Different types of uncertainty:
    - Is it in the data?
    - Is it a result of approximate integration and translations?

Indexing a Dataspace

- Build a heterogeneous index on *everything.*
- Think: Google desktop, but with clever indexing of (semi)-structured sources.
- Resolve multiple references to objects in the dataspace.
- Materialize some of the data for faster access.

Dataspace Discovery

- What do I have in my enterprise??
- Tasks:
  - Find the sources and classify them.
  - Suggest mappings between sources.
  - Suggest which sources may be related.
  - Maintain this over time.
  - Create associations between data items.

Consistency and Recovery

- Mike?

Reuse, Reuse and Reuse

- Reuse any human effort related to a dataspace.
- First example:
  - Reuse schema mappings
  - E.g., everyclassified.com includes 4500 mappings. Reuse was key.
- Next steps:
  - Reuse other human annotations
  - Reuse for more removed tasks.

Summary

Dataspaces -- because:

- That’s the size of the problem
- The field needs funding
- There is a ton of exciting stuff to do