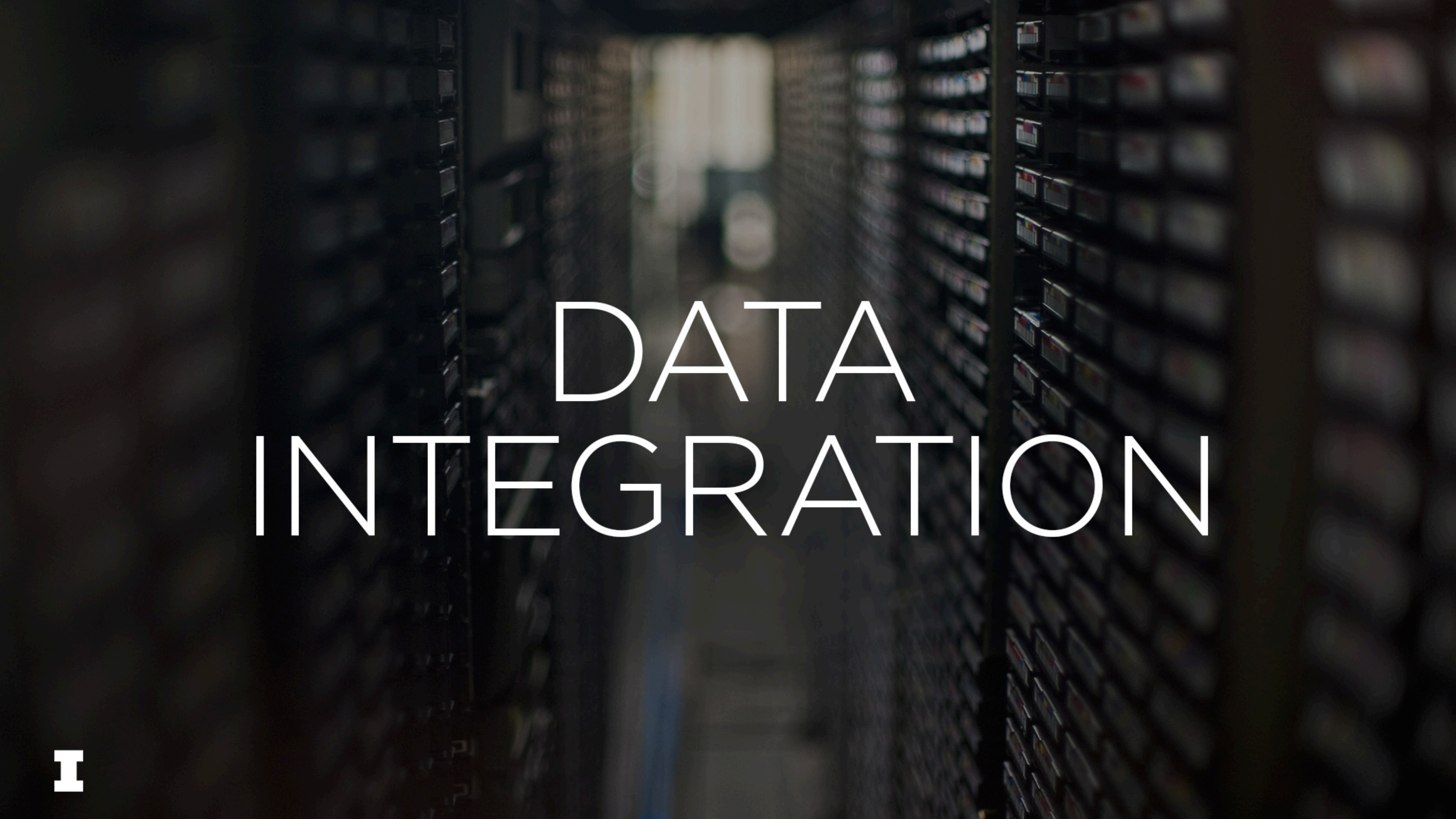
FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign



DATA INTEGRATION

①

DATA CLEANING, DATA INTEGRATION

Data cleaning, Data integration

- Data cleaning
- Data integration
- Heterogeneity
- Federation vs derived combinations

Data Cleaning

Data cleaning is a general term used colloquially to describe preparing data for analysis. When a single schema is involved the phrase typically suggests dealing with:

- duplicate records
- values that are missing, out of bounds, or inconsistent
- data typing errors or inconsistencies
- data entry errors
- attribute interpretation errors
- variation in null handling
- misapplication of standards
- inadequate normalization
- missing or inadequate schemas
- missing relationships
- failures of referential integrity or other constraints
- failure to match schema (e.g. as identified by a formal grammar or XML parser)

Data cleaning is profoundly important – without it data cannot be used reliably, or at all.
It also is a major industry expense and consume much staff time.

When multiple schemas (or disparate instances) are involved the data preparation task is *data integration*

Data Integration

Data integration:

“. . . combining data residing in difference sources and providing users with a unified view. . .”

[Lenzerini 2002]

Data model relationships

Entities, Relationships

Conceptual models, UML or ER models, ontologies

Schemas: ER, UML . . .

Schemas: RDFS, OWL . . .

Conceptual Level

Logical Level

Relations

e.g., Relational databases

Schemas: column and key descriptions

Trees

e.g., XML Documents

Schemas: grammars (e.g. DTDs),

Triples

e.g., RDF triple stores

Schemas: serialization descriptions.

Physical Level [or: Storage]

[files, records, delimiters, data structures, indexes, etc.]

Why is it important?

Real world problems are profoundly interdisciplinary, solving them requires integrating diverse data from multiple sources

e.g., an effective response to an impending natural disaster can require understanding how many people will be affected, hospital location and capacity, and transportation routes, and so on.

Many different disparate databases will need to be accessed (demographic, meteorological, geographical)

And, most importantly, the data elements will need to be related: concentrations of people connected to transportation routes, the storm path, hospital capacity, etc.

If this cannot happen or cannot happen reliably, or efficiently,
much valuable data will be useless, opportunities lost, and problems unaddressed.

Why is it hard?

When datasets are developed by different communities and for specific purposes; integrating them with other datasets is often not anticipated.

(and accommodation would be hard in any case)

These datasets often use different data models, schemas, and encodings that are very hard to related to each other, even when describing the same real world feature.

But unless common concepts can be found to connect data across datasets, and to either standardize or refactor related data elements, integration is impossible.

The obstacle to data integration is therefore: *heterogeneity*.

Kinds of heterogeneity

*Relatively easy
**Often difficult
***Usually very difficult

Encoding heterogeneity *

Different mappings from bitstreams into bytes, characters, numbers, or other logical units

Syntax heterogeneity *

Different data description languages for the same model type: e.g. RDF/XML vs N3

Model heterogeneity **

Different model type; e.g., relations vs entities/relationships

Representational heterogeneity **

Different modeling choices within a model type; e.g. relationships vs entities.

Semantic heterogeneity ***

Different conceptualization of similar domain features

Processing heterogeneity **

e.g. different maintenance and update regimes

Policy heterogeneity **

e.g. different privacy and security rules, varying ownership and licensing, etc.

Adapted from Bertram Ludäscher, Kai Lin, Shawn Bowers, Efrat Jaeger-Frank, Boyan Brodaric, Chaitan Baru, "Managing scientific data: From data integration to scientific workflows", *Geoinformatics: Data to Knowledge, 2006*; and Amit Sheth. "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics". In M. Goodchild, M. Egenhofer, R. Fegeas,and C. Kottman, editors, *Interoperating Geographic Information Systems* Kluwer, 1998.

Two general approaches to integration: federation vs derivation

Federation

For relevance: Standardized metadata attached to each dataset can be used to determine its relevance, indicating spatial and temporal location and general nature of content; this facilitates discovery.

For queries: Views on and queries against multiple databases are supported by mappings to a mediating meta-schema.

Derivation

A single dataset is derived from multiple sources and governed by a single schema.
[compare Extract, Transform, and Load (ETL) data warehouses]

In either case heterogeneity remains a huge challenge

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.