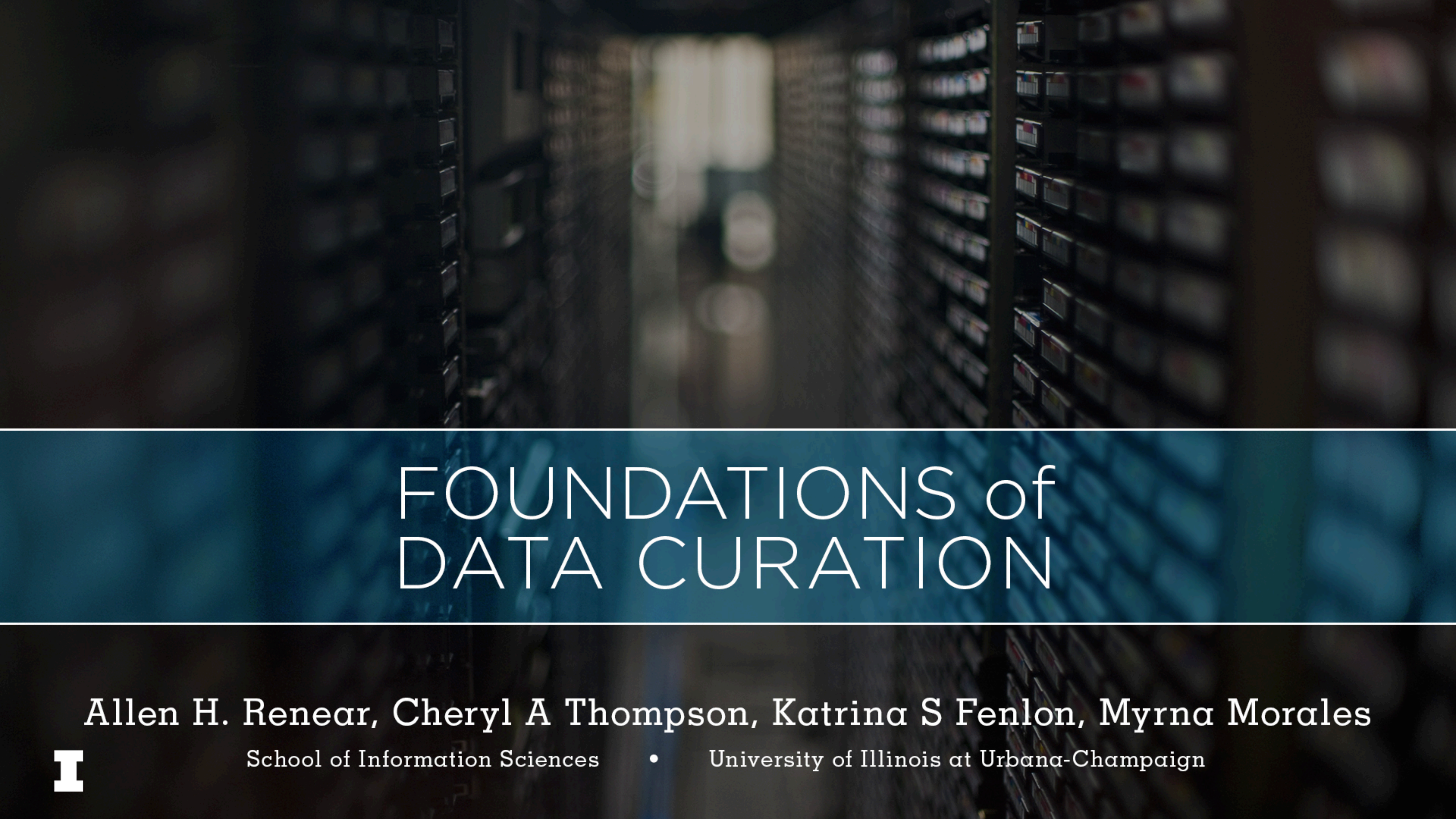
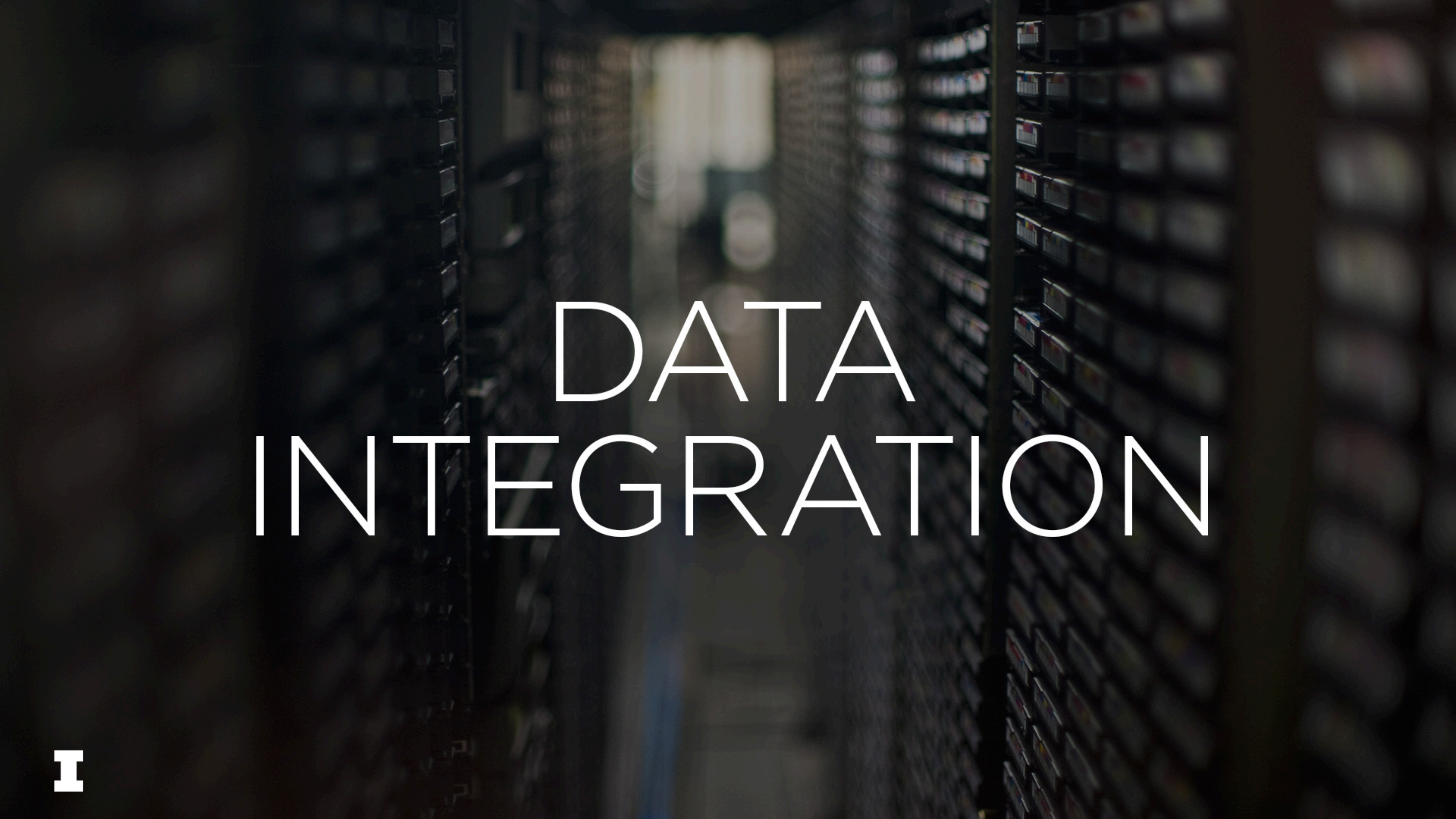


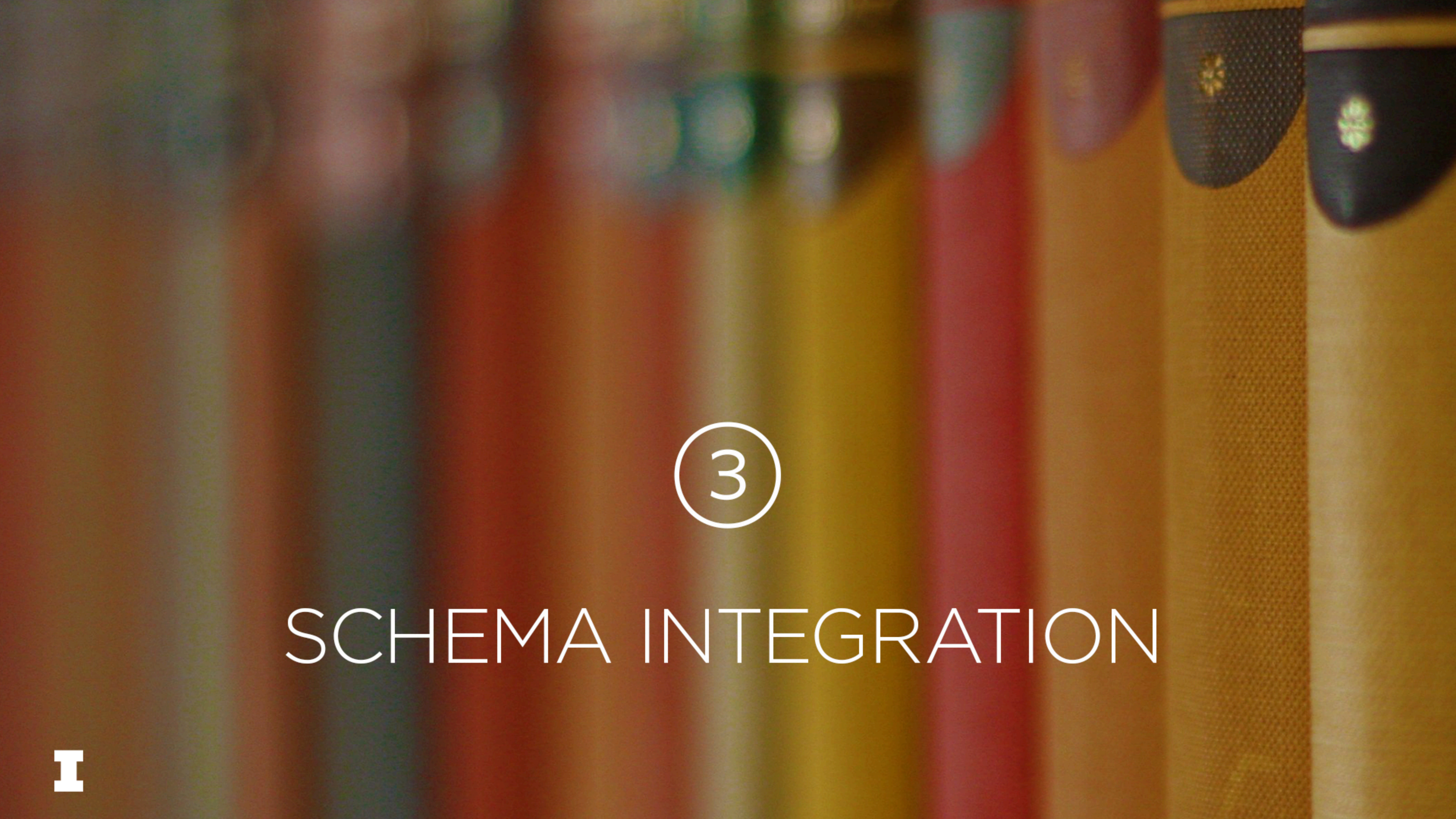
FOUNDATIONS of DATA CURATION

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

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



DATA INTEGRATION



③

SCHEMA INTEGRATION

Schema Integration

Definition

Representational challenges

Semantic challenges

- Homonyms, synonyms, overlaps, missing relationships, generality . . .

Definition: Schema integration

Schema Integration

> *"the process of merging several conceptual schemas into a global conceptual schema that represents all the requirements of the application"*. (Batini et al)

- If the approach is federation a federating schema is created and mapped to individual schemas.
- If the approach is a derived combination dataset a single schema is created for the derived dataset.

Schema integration usually takes place at the conceptual level (as indicated above).

Or at the logical level (e.g., relational schemas or XML schemas).

Or even across levels (e.g. an XML schema and an ontology);
(particularly when the distinction between logical and conceptual level is unclear).

Two kinds of schema integration problems

Representational (aka *structural*)

Different choices about modeling constructs and integrity constraints

Semantic

Synonyms, homonyms, missing relationships, different but related concepts, etc

In what follows we focus on naming problems.

Simple naming problems: *synonyms* and *homonyms*

For example:

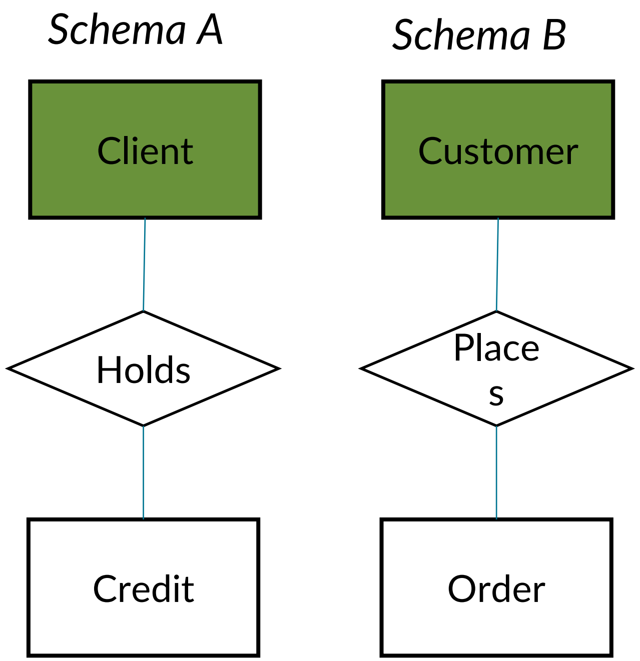
1) Two schemas use different names (F and G) for the same properties
2) Two schemas use the same name (F) for different properties

Once it is determined that 1) or 2) is the case the remedy, either in a meta-schema for federation, or the new comprehensive schema for derived combination is obvious.

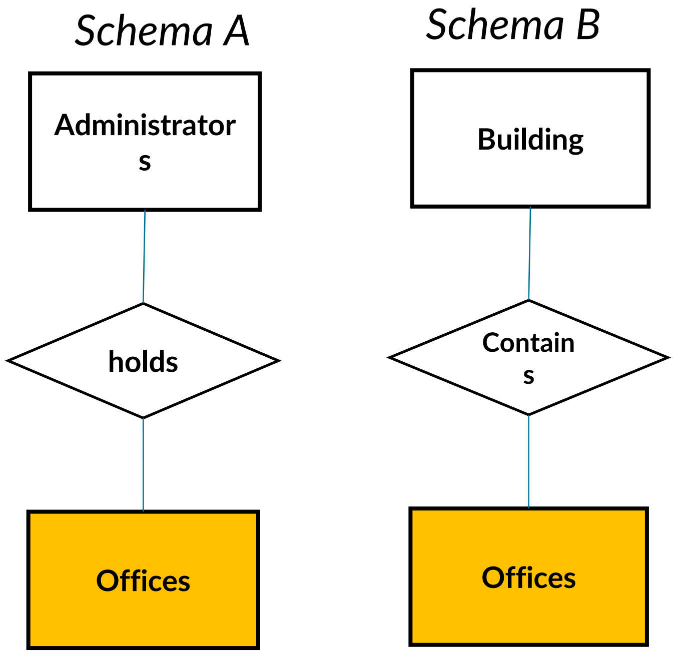
- What is hard is determining that there is in fact a synonym/homonym problem.
  - This might involve examining documentation, interviewing schema designers and data collectors, etc.
  - And, of course examining the distribution of values in the dataset instances themselves
  - [In all cases using background knowledge of the domain is essential.]
- But none of these are decisive.

Synonyms, Homonyms

Synonyms



Homonyms



Identifying “client” and “customer” as the same entity type in different schemas is called “schema matching”.

Not so simple naming problem: conceptual overlaps

A more challenging problem is when there is a conceptual overlap.

For instance:

- Schema A has an attribute *name* interpreted as applying to nicknames and legal names, but not to aliases
- Schema B has an attribute *name* interpreted as applying to legal names and aliases, but not nicknames

Like other earlier problems overlap may be difficult to discover.

But it is also difficult to remedy without information loss.

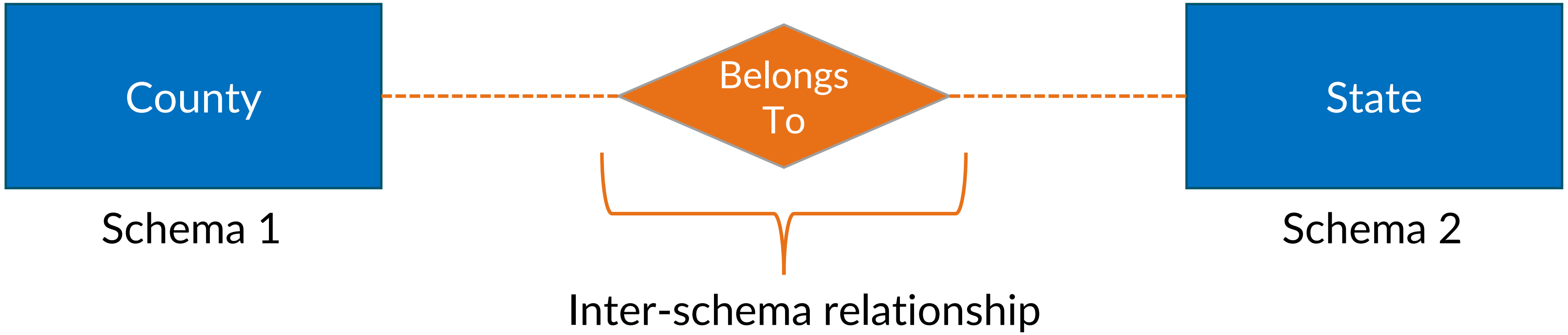
What are the options?

Inter-schema relationships

Suppose one schema has the entity type *county*, but not the entity type *state*, another has the entity type *state* but not the entity type *county*.

This limits the information we can extract from these datasets.

Adding *belongs_to* to the integrating schema and a mapping of counties to states will allow instance data in the first to be connected to states as well as counties.



NB: These relationships are not represented in either schema; they must be inferred.

Adapted from Batini, C., Lenzerini, M., and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. ACM Computing Surveys 18(4). doi>10.1145/27633.27634

Generality variation

Schema A has an entity type *vehicle* but no way to indicate kinds of vehicles.

Schema B has an entity type for *motorcycle* but no entity type for *vehicle.*

[Motorcycle is, or course, a proper subset, and so subclass, of *vehicle]*

A simple solution would be to generalize *motorcycle* to *vehicle*, but this would lose information.

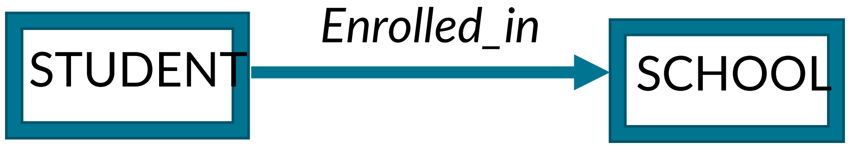
Retaining both *vehicle* and *motorcycle* with *motorcycle* as a subclass of *vehicle* retains all the information we have, similarly a *kind* attribute with motorcycle value could be used as well.

[No information is lost, but in both cases an unevenness in instance descriptions results.]

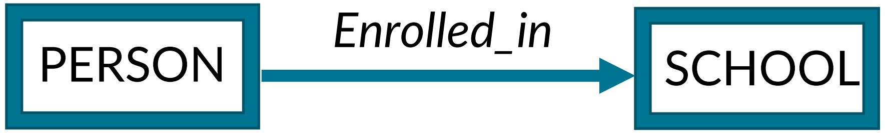
As always the basic strategy is the same whether for federations or derived combinations.

Entities, relationships, subclasses

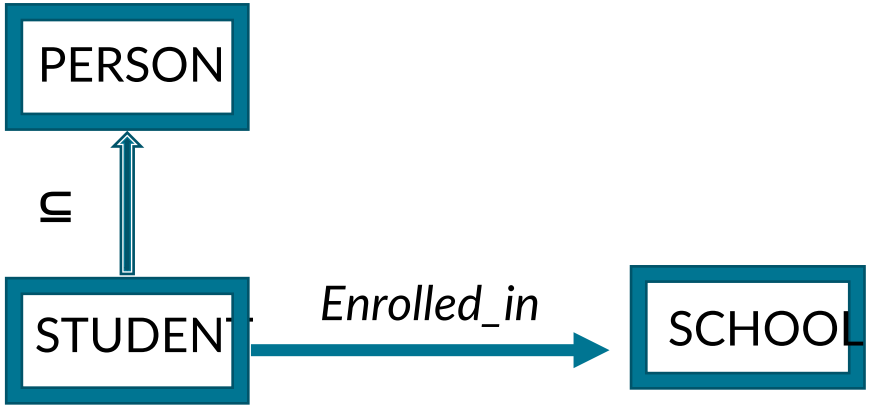
Schema A



Schema B



Schema C



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.