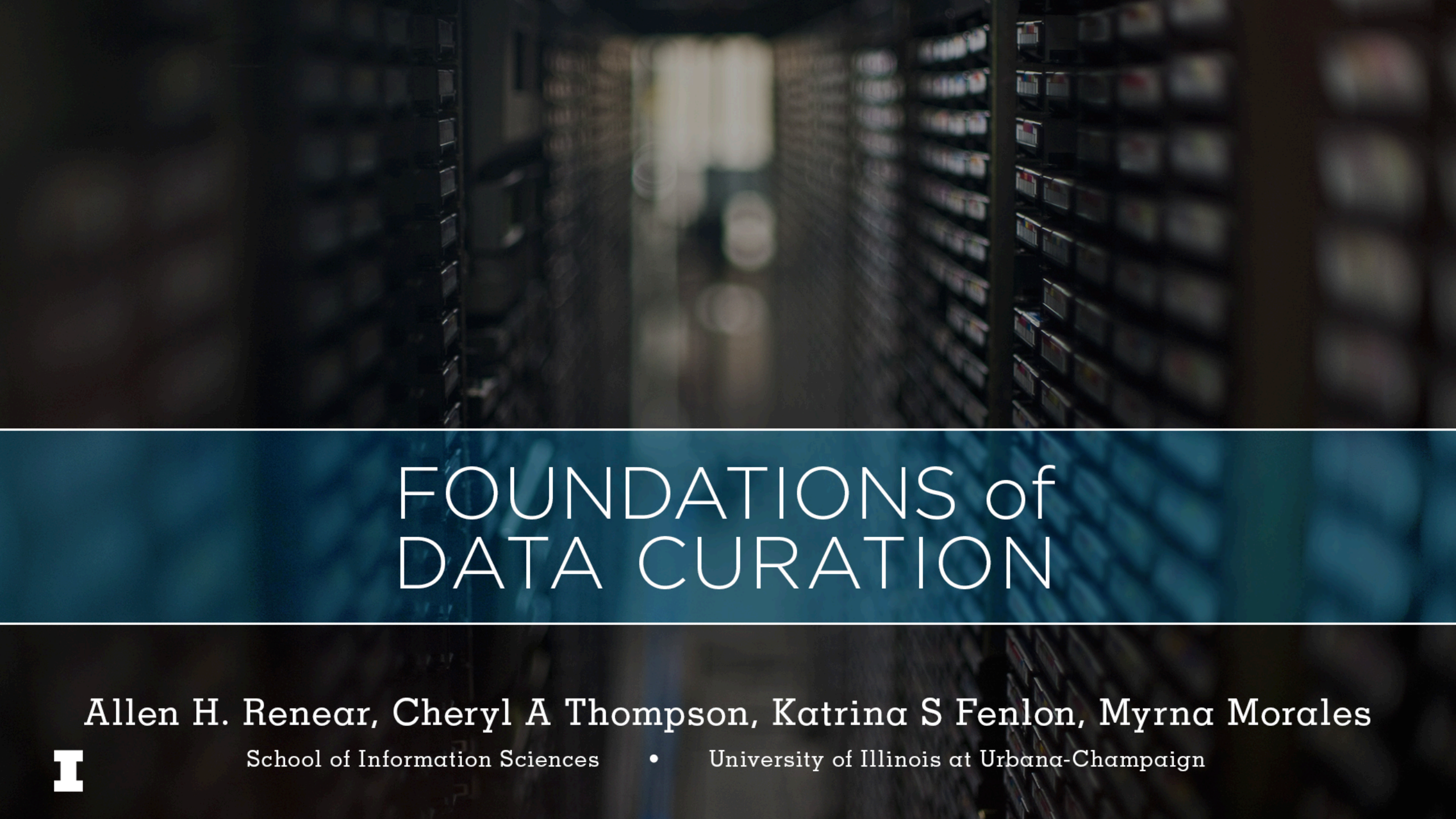
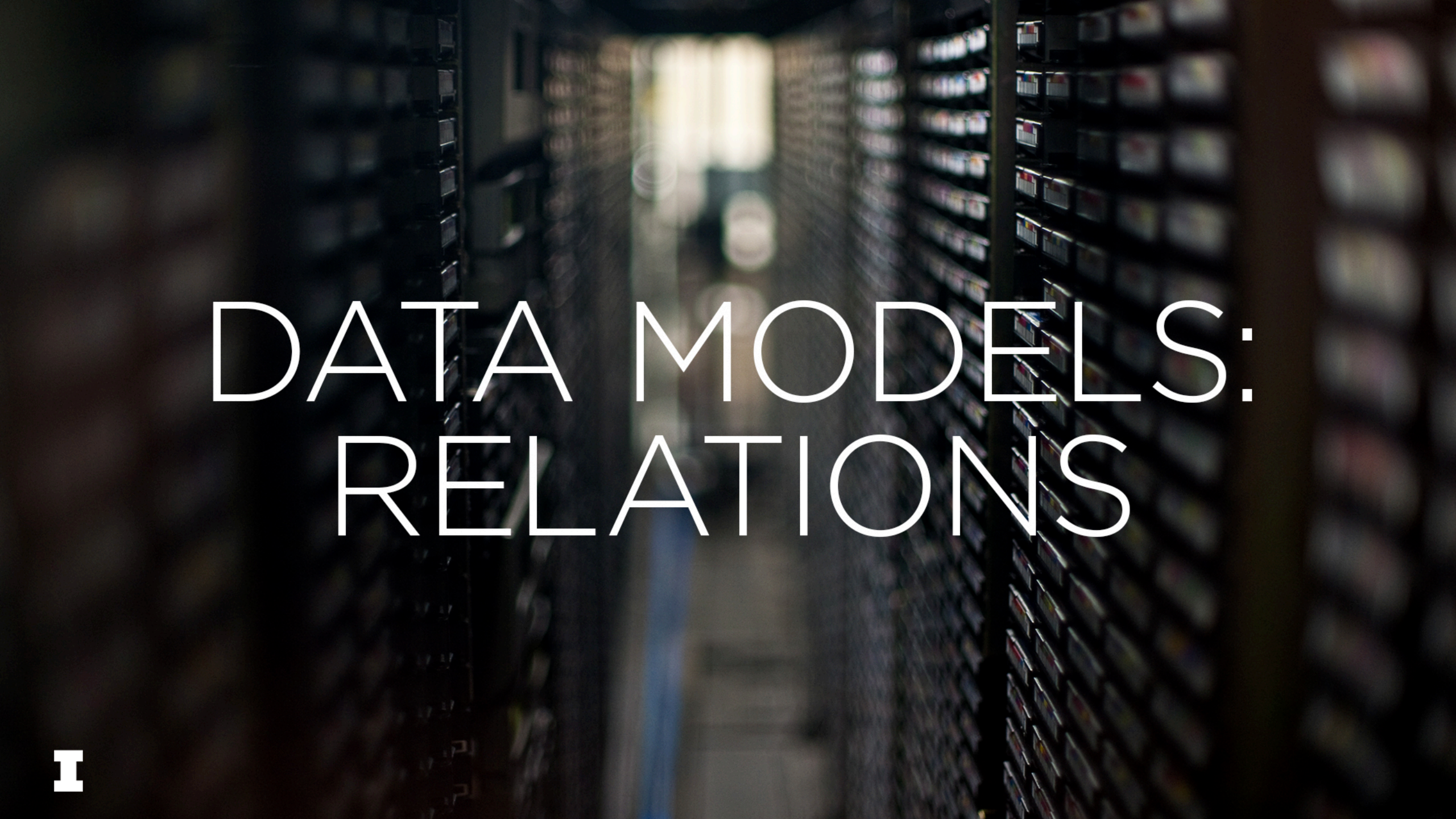


FOUNDATIONS of DATA CURATION

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

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



DATA MODELS: RELATIONS


2
THE PROBLEM

What's The Problem?

- The situation (circa 1960)
  - Data is stored in radically different ways
  - Interaction with data is immediately and directly via storage methods
  - Explicit and formal conceptualization of data *as data* is rare
    (and typically only in human memory)
- Why is this a problem?
  - Huge operational inefficiencies
  - Lack of functionality
  - Lack of data independence

What's the Problem?

Imagine a large company with many divisions

Left to itself each division will probably:

(i) conceptualize their domain in different ways

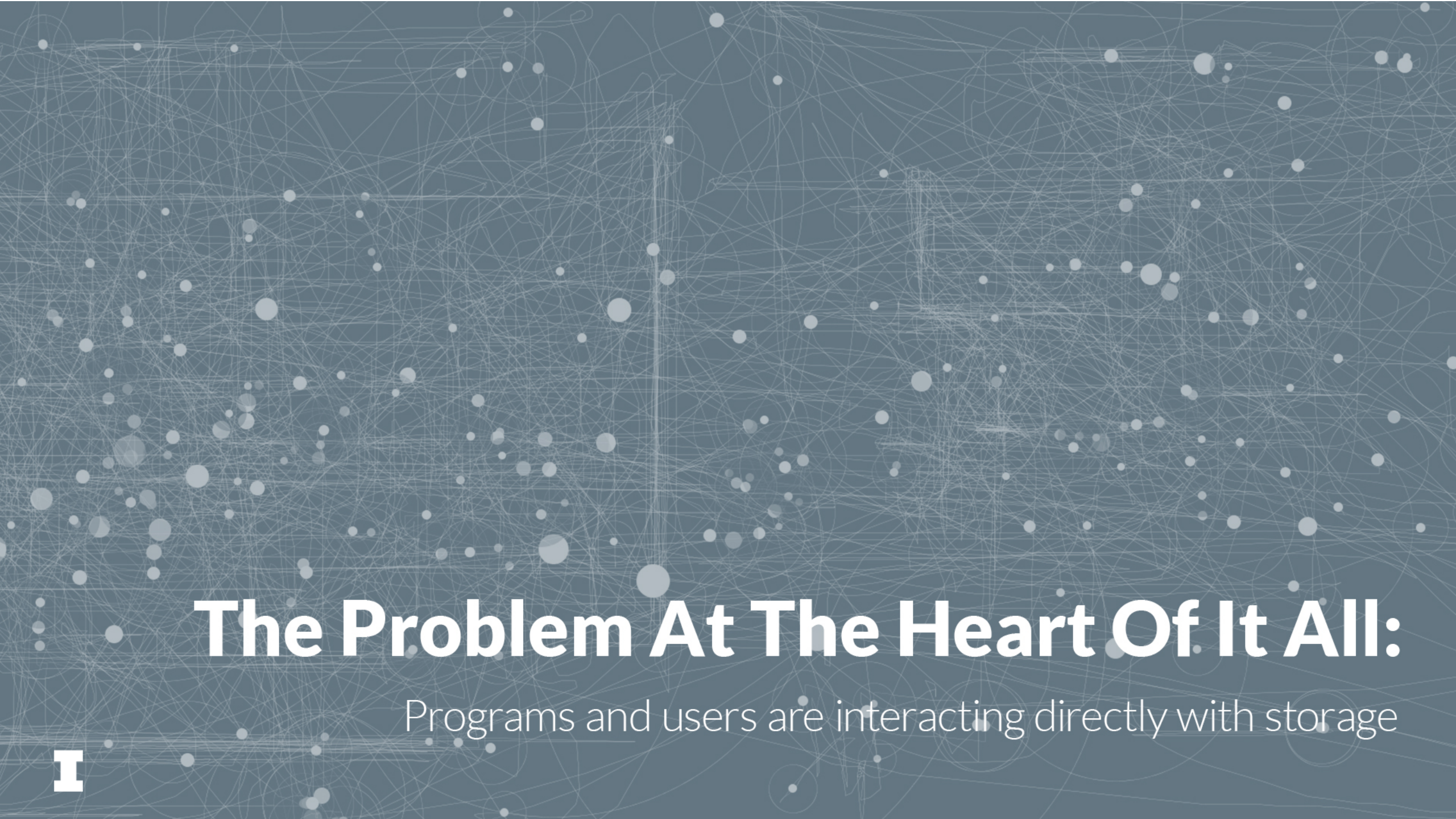
(ii) develop different methods of representation and storage

(iii) mix processing instructions with data when convenient

There will typically be

- no clear separation between storage methods and the intrinsic structure of the information being represented
- no general abstract understanding of the nature of the information being represented [other than what is in the memories of staff and programmers]

Flickr | ArnoldReinhold



The Problem At The Heart Of It All:

Programs and users are interacting directly with storage

Storage representations

Variable-length fields, delimited

```
W54825,Moby Dick,1851
W85246,The Scarlett Letter,1860
W55427,Fanshawe,1828
```

```
W54825    Moby Dick    1851
W85246    The Scarlett Letter        1860
W55427    Fanshawe     1828
```

Indexed byte offsets

00000000	(WorkID)
00000010	(Title)
00000020	(Year)

Fixed-length fields

…counting from the left: 0, 6, 25

W	5	8	4	2	5	M	o	b	y		D	i	c	k											1	8	5	1
W	8	5	2	4	6	T	h	e		S	c	a	r	l	e	t	t		L	e	t	t	e	r	1	8	6	0
W	5	5	4	2	7	F	a	n	s	h	a	w	e												1	8	2	8

…or from the right: 28, 22, 3

The problems this causes (1)

With no formally defined general approach application development is arduous.

- Many different unique access subroutines must be developed, tested, and maintained.
- Tools developed for different divisions are not interoperable.
- There cannot be a sustainable 3rd party industry of common applications and tools.
- Specialized applications (searching, analysis, etc.) must be custom developed, cannot reference high level constructs
- All tools must be be modified often as storage formats change.

The problems this causes (2)

There are more problems caused by a lack of a general conception of data:

- Workflow and transformation cannot be usefully tracked, audited, or logged
- Documentation cannot exploit a conceptual understanding of the data
- Documentation must be updated frequently for changes in storage format or processing changes.
- Data validation and quality assurance is difficult as standard tools for syntax checking, typing, constraint management, etc., cannot be used.
- Schemas, if they exist, focus only on storage and do not help us with general data management.

As a result. . .

This result is that systems and practices are:

- Inefficient
- Error-prone
- Untrustworthy
- Difficult to document
- Difficult to repurpose and reuse
- Difficult to preserve for future use
- Dependent on memory and workplace practices
- Dependent on custom tools and applications

Data Independence

One significant consequence of this chaos is a failure of data independence.

This failure comes in two varieties:

Type 1: If the storage method changes, then the end user programs accessing the data will fail to perform as expected.

Type 2: If new kinds of data need to be represented, then again end user programs may fail or give the wrong result.

First, keep these variations in mind

Variable-length fields, delimited

```
W54825,Moby Dick,1851
W85246,The Scarlett Letter,1860
W55427,Fanshawe,1828
```

```
W54825      Moby Dick     1851
W85246      The Scarlett Letter          1860
W55427      Fanshawe      1828
```

Indexed byte offsets

00000000	(WorkID)
00000010	(Title)
00000020	(Year)

Fixed-length fields

…counting from the left: 0, 6, 25

W	5	8	4	2	5	M	o	b	y		D	i	c	k										1	8	5	1
W	8	5	2	4	6	T	h	e		S	c	a	r	l	e	t	t		L	e	t	t	e	r	1	8	6	0
W	5	5	4	2	7	F	a	n	s	h	a	w	e												1	8	2	8

…or from the right: 28, 22, 3

Lack of data independence (type 1)

If the physical storage method changes, then the end user programs accessing the data will fail to perform as expected.

For instance, if the storage method switches from a fixed field approach to a delimited field approach then the access programs (and other tools) will return the wrong results.

Lack of data independence (type 2)

If new kinds of data need to be represented then, again, end user programs may fail or give the wrong result.

For instance, if a new attribute is accommodated by adding a delimited field to the right side of a record, then any program or other tool that has been identifying fields by counting delimiters right to left will probably return the wrong result.

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.