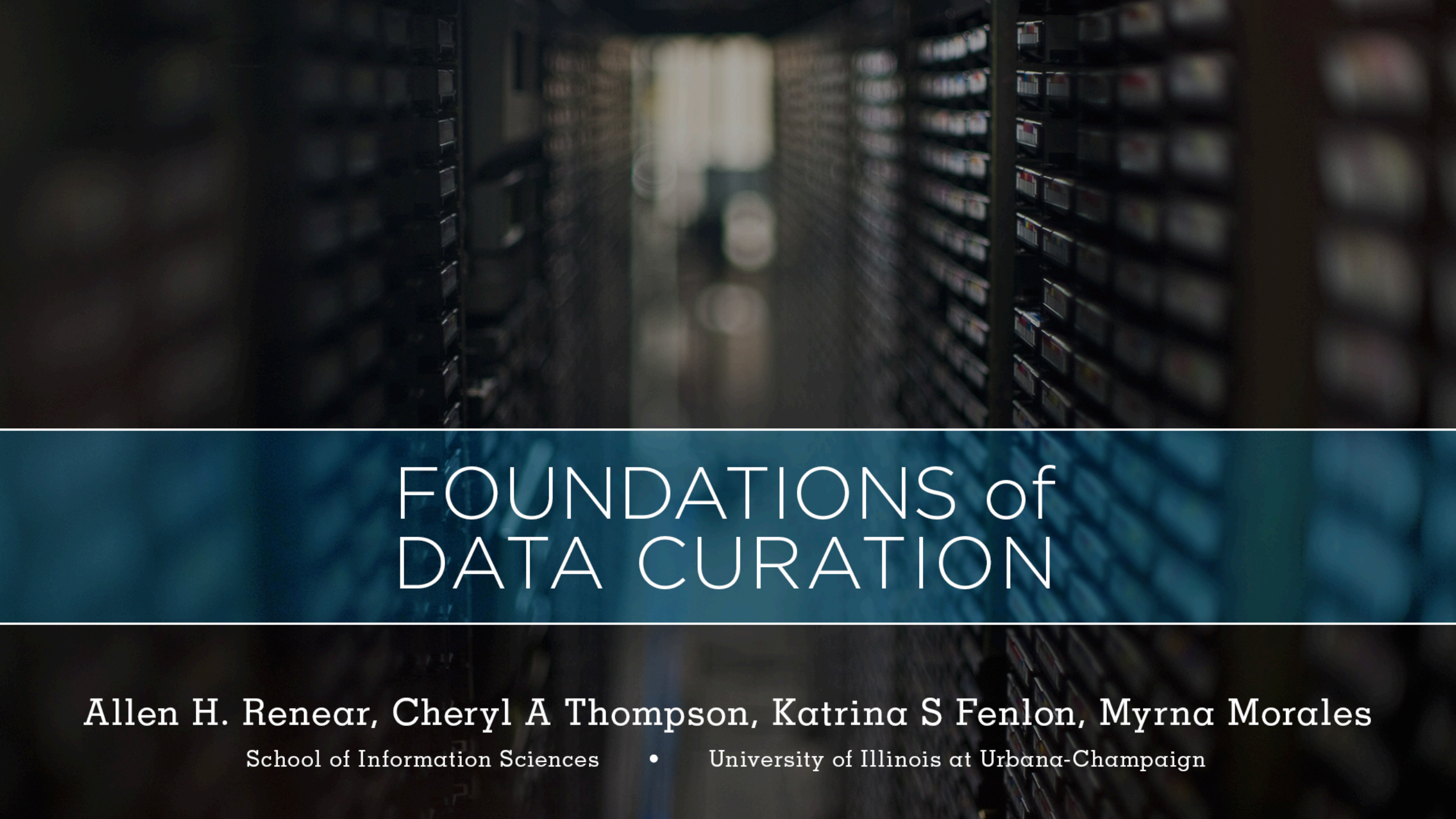


FOUNDATIONS of DATA CURATION

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

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

Workflow and Provenance

(anything profound, and the cool slides, is from Bertram Ludäscher. Everything else is from Renear

V2. Provenance

What is provenance

Why is provenance important?

Kinds of provenance

Origin of the term “provenance”

General:

The place of origin or earliest known history of something (OED)

In the humanities:

A record of ownership of a work of art or an antique, used as a guide to authenticity or quality (OED)

Computational provenance:

Origin and processing history of an artifact – usually: data (products), figures, ...
sometimes: workflow (and script) evolution (DataOne)

. . . the sources of information, including entities and processes, involved in producing or delivering an artifact (W3C)

Computational provenance

The heart of computational provenance:

What data was used?

What calculations were performed?

and also: “What in the world exactly happened just now?!”

Why is provenance important?

Access to provenance information supports

- understanding
- reliability
- reproducibility
- trust
- attribution and credit
- discovery and reuse of data, tools, and algorithms

Moreover. . .

The consequences of inadequate provenance information can be profound.

Not only failures of understanding and scientific reputation,
and the abstract loss of reliability and trust.

But also:

Failure of transportation or power systems; medical systems, instruments, and therapies; failures of heavy machine operations, etc.

Civil and criminal penalties for use of data or methods that violated property rights or legal restrictions, or should have been known to be flawed

Levels of provenance (Ludaescher)

Black-box

Little is explicitly and verifiably known about what data and methods are being used.

White-box

A mathematically exact representation of data and algorithms used is available.

Grey-box

Identification of datasets and high level processes is available

Prospective vs Retrospective provenance (Ludaescher)

Prospective [aka “Compile-time” provenance, aka “workflow land”]

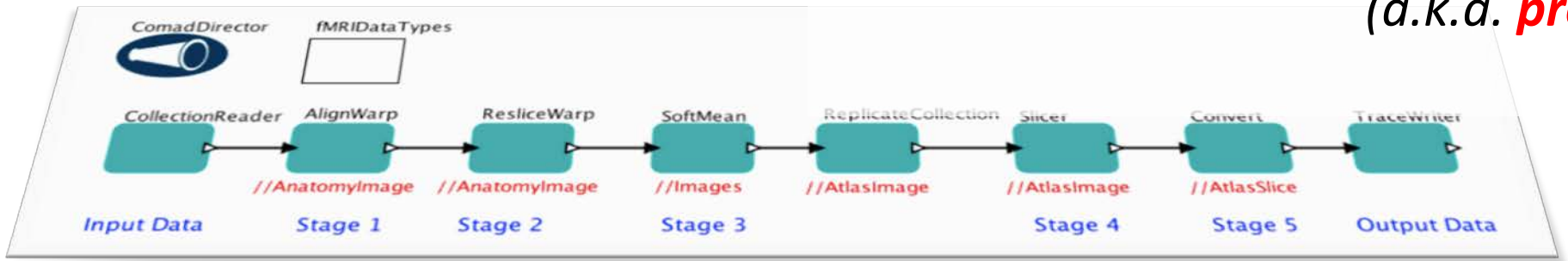
a correct specification the workflow scenario

Retrospective [aka “Run-time” provenance, aka “trace land”]

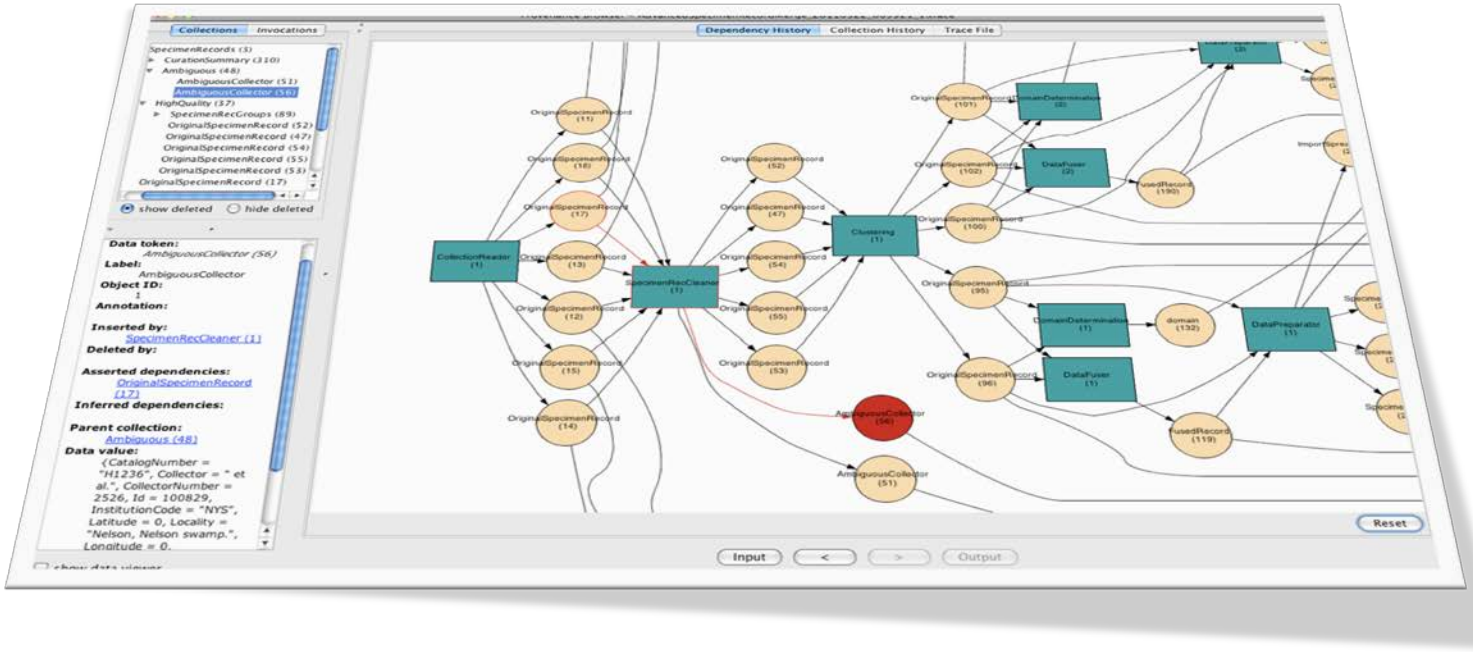
generated data on the execution of the workflow scenario

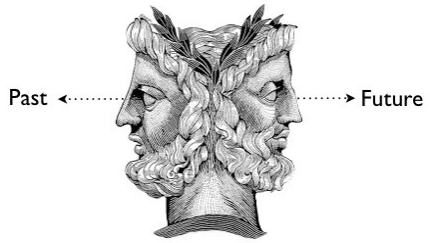
The Workflow⇔ Provenance Connection … (Bertram Ludäscher)

Workflow Modeling & Design
(a.k.a. ***prospective*** ***provenance***
"Workflow-land"*)*



Runtime Provenance
(a.k.a. ***traces, logs,***
retrospective ***provenance,***
"Trace-land"*)*





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.