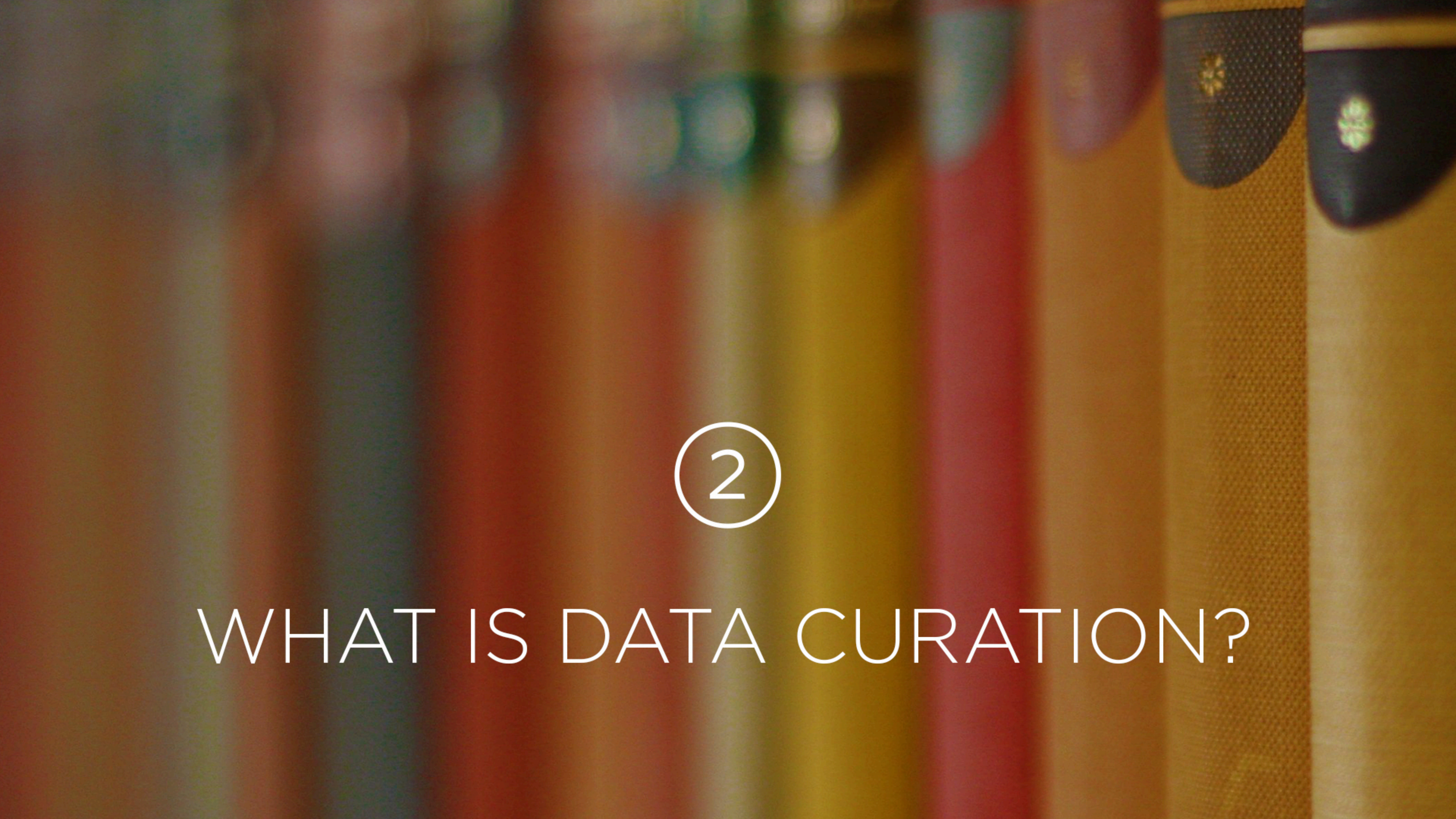

②

WHAT IS DATA CURATION?

What is data curation?

- Definitions of data curation
- Science vs practice of data curation
- Data analytics values / Data curation values
- Importance of data curation
- Relative size ($$, employment) of data curation vs data analytics

Our definition of data science (again)

Data science is concerned with all aspects of the **creation, management, analysis,** and **communication** of data focusing particularly on the application of computational methods to digital data

The data science objective: *extracting useful knowledge from data*

Data science = Data Curation + Data Analytics

Data science has two components:

Data curation: Ensuring that data can be efficiently and reliably found and used

Data analytics: Employing specific techniques to extract knowledge from data

Data curation is concerned primarily with the *management of data* in order to better support *the analysis of data**

It includes among many other things: acquisition and collection, modeling, workflow, provenance, validity and integrity, metadata, preservation, integration, retrieval, re-use, policy, standards, identifiers, format conversions, processing levels, supporting reproducibility, etc.

[*] but the boundary is not a sharp one.

Data curation: the Illinois definition

From Wikipedia:

“According to the University of Illinois [School of Information Sciences],

"Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.”

“Data Curation”, Wikipedia. Retrieved May 1 2017.

Original source: “An Educational Program on Data Curation”. *ALA Science & Technology Section Conference*. Cragin, Melissa; Heidorn, P. Bryan; Palmer, Carole L.; Smith, Linda C. (2007).

Science of... *vs* Practice of…

The *science* of data curation:

research and development on new methods of data management and use; draws on mathematical and engineering methods, but also on methods from social science, law, economics, and other disciplines.

The *practice* of data curation:

the use and adaptation of data management methods to meet user needs and support data analytics

Once again, not a sharp distinction, but a real one nonetheless [the same distinction can be made for data analytics]

Data science values

Data analytics values:

Extraction should be novel, fast, precise, accurate.

Data curation values:

Data should be efficient and reliable: findable, useable, legal

(thereby supporting novelty, speed, precision, accuracy.)

Importance of data curation (1)

Where real world interdisciplinary challenges are concerned, curatorial problems are acute:

Large amounts of rapidly changing data, often heterogeneous in nature and developed by different scientific communities, must be found, retrieved, authenticated, reformatted, integrated with other data and managed for effective use, and demonstrably reliable even after processing and preparation

Importance of data curation (2)

Supporting analysis, discovery, and use is an enormous challenge.

. . . it involves the complex management of large-scale data storage and preservation, creation of metadata and tools for retrieval and context documentation, preparation of computationally accessible documentation of provenance and workflow, conducting reliable format conversions to support new tools and applications, the management of identifiers and validity checks that accommodate format changes, the integration of related data elements from substantially different data sources, and more. . . .

Importance of data curation (3)

Without successful data curation successful data analysis is not possible, it would be prohibitively expensive and and dangerously unreliable.

Data curation is the larger part of data science

Not only is data curation essential for reliable efficient analysis, but most of the cost associated with using data is, by far, in curation, not analysis, and most of the workforce needs are, also by far, in curation, not analysis.

Ask any data manager in industry will tell you, it is curatorial work where they make the largest investment, of money, staff, time, and effort.