1

WHAT IS DATA SCIENCE?

What is data science?

- Definitions of data science
- Our definition of data science
- Importance of data science
- Who are data scientists?
- An nice ambiguity: science with data or science of data ?
- Research vs practice
- Data science = data curation + data analytics

Some definitions of data science

"…work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information …

Data scientists play active roles in … four related areas: data architecture, data acquisition, data analysis, and data archiving … Data Science is an applied activity and data scientists serve the needs and solve the problems of data users."

— Jeffrey Stanton, *Introduction to Data Science*, 2013

•

"… the science of planning for, acquisition, management, analysis of, inference, and discovery from data"

—"Final Report from StatSNSF [Support for the *Statistical Sciences* at NSF] subcommittee"
Iain Johnstone, Fred Roberts, Co-Chairs July 18, 2014.

Other terms related to data science

Other terms related to data science are:

- *informatics*, *data mining*, *knowledge discovery*, *eScience*, *cyberinfrastructure*, *data analytics*, *data curation*, etc.
- As well as, of course, the colloquial phrase “big data”.

Our definition of data science

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*

The importance of data science today

Several things have combined to create this revolution:

1. vast quantities of digital data from a wide variety of sources and often arriving in real time
2. stunning advances new analytical strategies and computational methods
3. easy to use powerful analysis tools
4. high-speed global communications
5. access to distributed high performance computing from almost anywhere in the world
6. a large community of experts in the new strategies for data management and analysis

The importance of data science today

Those things create exciting new opportunities for major advances in many industries, professions, and disciplines. . .

. . .including medicine, health, engineering, defense, safety, agriculture, business, the arts, community services, government services, and more

Data science — the with/of ambiguity

Existing definitions of data science reflect an ambiguity in the phrase “data science”.

Specifically: Is data science to be understood as the science *of* data, or as science *with* data?

Wikipedia has defined data science as:

1) “. . . the study of the generalizable extraction of knowledge from data.” — Wikipedia c. 2013-14
2) “. . . the extraction of knowledge from data” — Wikipedia c. 2014-15

The literal difference is clear:

In 1) doing data science is doing research on computational methods, that is: developing new tools and techniques, studying those tools and techniques, and studying their applications.

In 2) doing data science is *using* data science methods to do research in *other* fields (genetics, medicine, engineering, marketing, etc.).

Our definition deliberately avoids the issue. [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

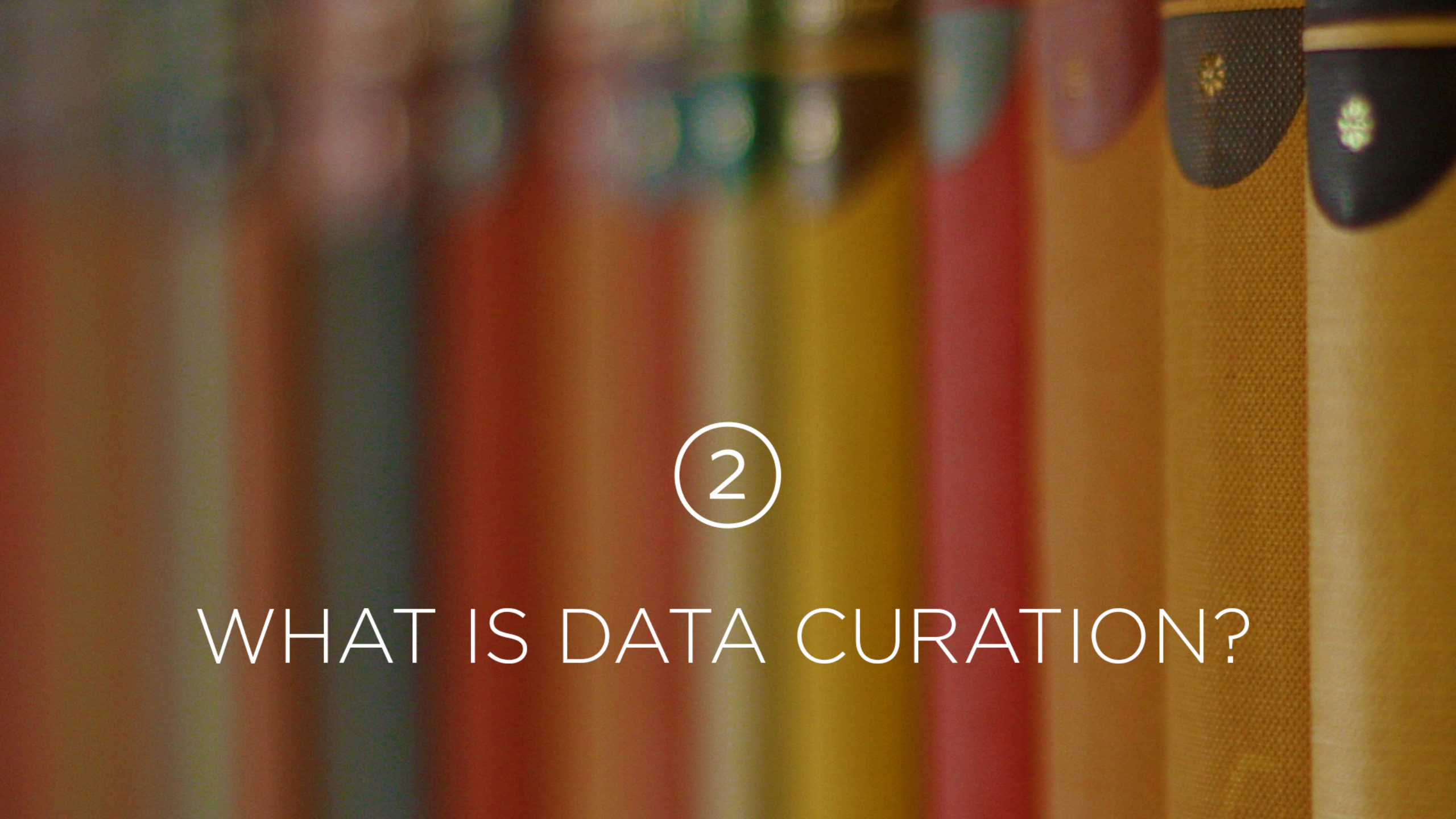
Data science has two components:

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.



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.

3

OBJECTIVES, ACTIVITIES, METHODS

The data curation objective

Data curation is concerned with all aspects of the management of data

in order to efficiently and reliably support the analysis of data, and enable reuse over time.

Areas of curatorial activities

Collection	Support the collection and acquisition of data
Organization	Employ an appropriate data model and use appropriate standards
Storage	Support reliable and effective storage
Preservation	Ensure that data will be understandable and useable in the future
Discoverability	Support the ability to search for and locate relevant data
Access	Support the ability to retrieve and distribute data
Workflow	Support the ability to systematize data workflows
Identification	Support the ability to identify, authenticate, and validate data
Integration	Support integration of data from different sources using different data models
Reformatting	Support reformatting for use by different tools or to match new format standards
Reproducibility	Support ability to reproduce results, ensuring scientific validity and reliability
Sharing	Support sharing data between researchers, teams, and institutions
Communication	Support representation, publishing, and visualizations that provide insight
Provenance	Support .identifying what inputs, processes, and calculations are responsible for data values
Modification	Support management of corrections and updates
Compliance	Ensure compliance to legal, regulatory, and local policy requirements
Security	Ensure that data is secure from tampering or inappropriate access and distribution

A Closer Look

Now let’s take a look at each of the areas. . .

Collection

Support the collection and acquisition of data

Includes support for, e.g., coordination of instrument calibration, protocols, procedures, collection area division, interview transcription, etc.

Of particular importance: recording information (as metadata) related to collection activity so that all relevant aspects of context are available later to support full understanding, authentication, and provenance.

Organization

Employ an appropriate data model and use appropriate standards

Determine an appropriate data model and schema

Use abstraction and indirection to manage data

Identify and use any relevant standards for both syntax and semantics

Of particular importance:

- Document schema attributes (including specifying datatypes and constraints).
- Document all changes to schemas.
- Maintain metadata for schema changes.

Storage

Support reliable and effective storage

Select storage strategies that proved the right mix of reliability, security and access

Preservation

Ensure that data will be understandable and useable in the future

Maintain a documented preservation strategy.

This includes not just bit sequence preservation and syntax documentation, but also the documentation of semantics for data elements and the generation and preservation of all metadata needed to ensure that the data is useable and understandable, and can be authenticated and audited for provenance.

Execute that strategy with discipline, documenting all actions taken.

Discoverability

Support the ability to search for and locate relevant data

Develop metadata to support searching for and finding relevant data in relevant formats.

Support searching that provide relevance ranking and recommends related datasets.

Access

Support the ability to retrieve and distribute data

Maintain systems, tools, and metadata that support the efficient and reliable retrieval and distribution of data.

Add metadata describing file formats

Where appropriate control access appropriately and maintain data on distribution and access.

Workflow

Support the ability to systematize work with data

The processing of data should be carried out a well-designed modular system of transformations.

The role of each module should be documented

The execution of a workflow should be documented as well.

To the greatest extent possible documentation should be generated automatically and should itself be both machine readable and executable.

Specifically: well-maintained scripts should be developed and used to document as well as execute data transformations.

Identification

Support the ability to identify, authenticate, and validate data

Identifier systems must be carefully designed.

Attention must be given to *what* (conceptually) is being identified and to the *method* of identification.

Related entities (such as the data abstractly and the same data represented in different formats) must be both precisely distinguished and precisely related.

Version control for format changes, corrections, etc. must be implemented.

Authentication (the data is in fact the data it claims to be) and validation (the schema constraints, syntax and semantics, are met) are both fundamental.

Integration

Support integration of data from different sources using different data models

Both variations in syntax and data element semantics must be accommodated if data from multiple sources is to be combined and related to solve real world problems.

Use schema alignment and cross-walking techniques to integrate data

Document integration strategies in detail so that any conflation, data loss, etc. is noted.

Reformatting

Support reformatting for use by different tools or to match new format standards

Data must frequently be reformatted in order to support new tools, new versions of existing tools, or to meet new format standards..

Reformatting must be documented and any changes in semantics or meaning must be identified.

Reproducibility

Support ability to reproduce results, ensuring scientific validity and reliability

Data curation for reproducibility includes documenting not only data collection and management, but also documenting processing and analysis.

Sharing

Support sharing data between researchers, teams, and institutions

There are many obstacles to data sharing, ranging from formats, to lack of documentation, to concerns about misuse or misunderstanding.

Data curation must address these, typically with policies, documentation, metadata, and interoperable systems.

Communication

Support representation, publishing, and visualizations that provide insight.

To be useful data must be presented in forms that provide insight (such as scientific visualizations) and integrated clearly and efficiently into the full life-cycle of scientific work, which includes scientific publishing. Related communication issues are relevant to other data curation activities: in entertainment, documentation, services, etc. Here data curation overlaps with interface design.

Provenance

Support identifying what inputs, calculations, and actions are responsible for data values

When one data set (or view) is derived from another, reliable use and understanding requires that the inputs, calculations, and actions responsible for data values can be identified.

Modification

Support management of corrections and updates

Data must be updated and corrected.

This must be supported and managed so that errors are not introduced but so that the changes overtime can be tracked and audited.

Compliance

Ensure compliance to legal, regulatory, and local policy requirements

The issues here range from intellectual property rights to regulations regarding the privacy of medical, financial, and personal information.

Security

Ensure that data is secure from tampering or inappropriate access and distribution

This will involve methods for controlling access and determining user identity and privileges, as well as data identity, authentication and validation.

Methods of curatorial action

Analysis: To determine needs, and develop relevant data models and *metadata*, and reformat, correct, or update data.

Documentation: To record essential information (typically via *metadata)*

System design and implementation: To support all data curatorial activities
To support the generation and use of data documentation and processing documentation

Policy: To specify objectives, procedures, practices, and formats.

Process: To ensure success and efficiency by managing the development of appropriate organizational units and roles, providing training, advocating for change, and managing curatorial activities.

4

ORGANIZATIONS, CONFERENCES, LITERATURE

Organizations

Research Data Alliance* (https://www.rd-alliance.org/)

International Digital Curation Centre * (http://www.dcc.ac.uk/)

Association for Information Science and Technology (https://www.asist.org/)

National Digital Stewardship Alliance (http://ndsa.org/about/)

*Be sure to browse.

Selected conferences (general)

Research Data Alliance* **	https://www.rd-alliance.org/
International Provenance and Annotation Workshop Series*	http://www.ipaw.info/
ISWC The International Semantic Web Conference*	https://iswc2017.semanticweb.org/
International Conference on Digital Preservation*	https://ipres-conference.org/
Dublin Core Metadata Initiative*	http://dcevents.dublincore.org/IntConf/dc-2017
CODATA*	http://www.codata.org/
Balisage: The Markup Conference*	https://www.balisage.net/
Digital Preservation	http://ndsa.org/meetings/
IDCC	http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc/
Association for Information Science and Technology	https://www.asist.org/events/annual-meeting/
iConference :	http://ischools.org/the-iconference/
Research Data Access and Preservation Summit	https://www.asist.org/events/rdap-summit/

**Be sure to browse*

Some specialized data curation-related conferences

IDigbio (Biodiversity data) https://www.idigbio.org/

ESIP (Earth sciences data) http://meetings.esipfed.org/

eScience (data-intensive science) http://escience-2016.idies.jhu.edu/

JCDL (Digital libraries) http://www.jcdl.org/

IASSIST (Social science data) http://www.iassistdata.org/

DH (Humanities) https://www.adho.org/

Literature — journals

International Journal of Digital Curation	http://www.ijdc.net
CODATA Data Science Journal	https://datascience.codata.org/
A list of specialized data journals:	http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList

Workforce

Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century
National Science Board, 2005

https://www.nsf.gov/pubs/2005/nsb0540/

Preparing the Workforce for Digital Curation,
National Research Council of the National Academies, National Academies Press, 2015

https://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation

[Electronic download is free]