Foundations of Data Curation Review!!

Data Curation
What is it?

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*

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.

Areas of curatorial activities

Collection:	Support the collection and acquisition of data (and documentation of that, *throughout list)*
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
Sharing:	Support sharing data between researchers, teams, and institutions.
Communication:	Support representation, publishing, and visualizations that provide insight
Provenance:	Support identifying what inputs 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

Methods of curatorial action

Five categories stand out as particularly important:

Analysis

To determine needs, 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.