Rensselaer

why not change the world?®

# Data Management, Data and information acquisition (curation, preservation), Provenance

**Ahmed Eleish**

**Data Science – ITWS/CSCI/ERTH 4350/6350**

**Week 2, September 9th, 2026**

Tetherless World Constellation
Rensselaer Polytechnic Institute

# Admin information

- Class: ITWS-4350, ITWS-6350, CSCI-4350, CSCI-6350, ERTH-4350, ERTH-6350
- Hours:

  11:00am ET - 1:50pm ET on Wednesdays – Lally 104
- Instructor: Ahmed Eleish
  - Email: eleisa2@rpi.edu
  - Webex: https://rensselaer.webex.com/meet/eleisa2
  - Instructor office hours: Wednesdays from 02:00 PM - 4:00 PM ET or by appointment/email
  - Instructor office location: Lally 315 / Amos Eaton 134
- TA: Raman (Ram) Varadharaj
- Web site: https://tw.rpi.edu/classes/data-science-2026
- LMS (http://lms.rpi.edu/)

# Readings from last week

- Required:

What is Data Science (I) [4]

- Choose 1 from:

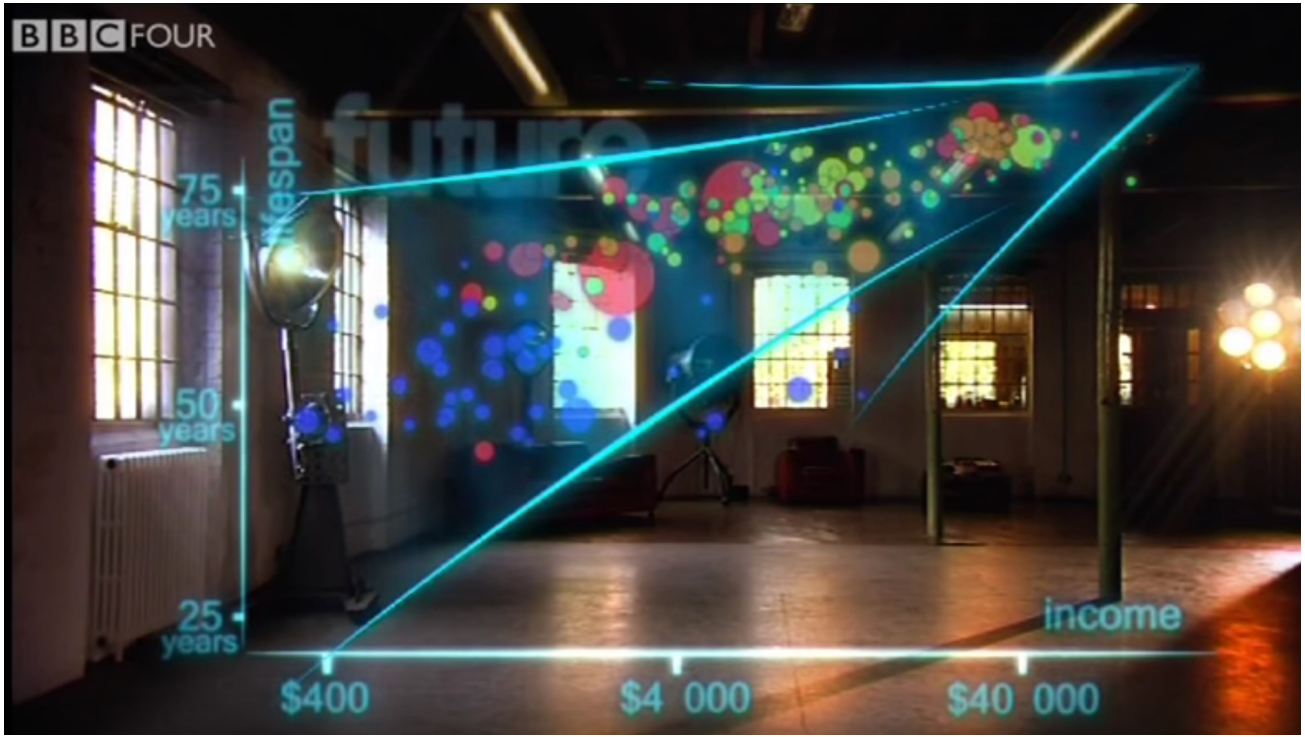
Changing Science: Chris Anderson: [1]
Rise of the Data Scientist [2]
Where to draw the line? [3]
Data Science vs. Data Analytics [5]
What Is a Data Scientist? [6]
Big Data [7]
A Very Short History of Data Science [8]
7 Phases of Data-Life-Cycle [9]

# Concepts from last week

- Data
- Information
- Knowledge
- Metadata/documentation
- Data life-cycle

# Story telling with Data

- https://youtu.be/jbkSRLYSojo



# Components of Data Science

- Data life cycle – acquisition, curation and preservation
- Data management and data/information products
- Forms of analysis, errors, and uncertainty
- Technical tools and standards

# Functions of Data Science

- Definition of hypotheses, guiding questions
- Finding and integrating datasets
- Planning and executing analyses and investigations
- Presenting results and visualizations
- Communicating findings and conclusions

# Reminder

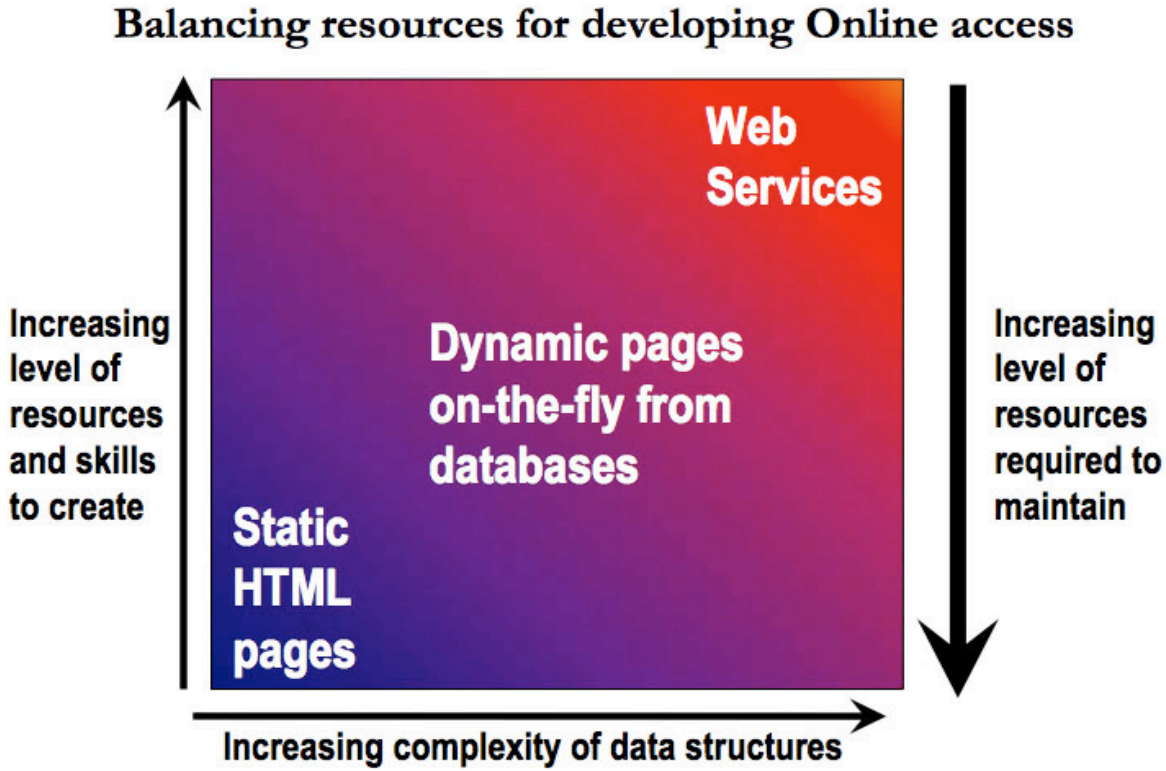
• Science data (and information) challenges are being identified as increasingly common.

• Data (and information) science now accompanies theory, observation/experiment and simulation as a means of doing science.

• Scientists and technologist are not well prepared to cope with 21st century data management and use of tools.

• Making data available is now a responsibility not a privilege.

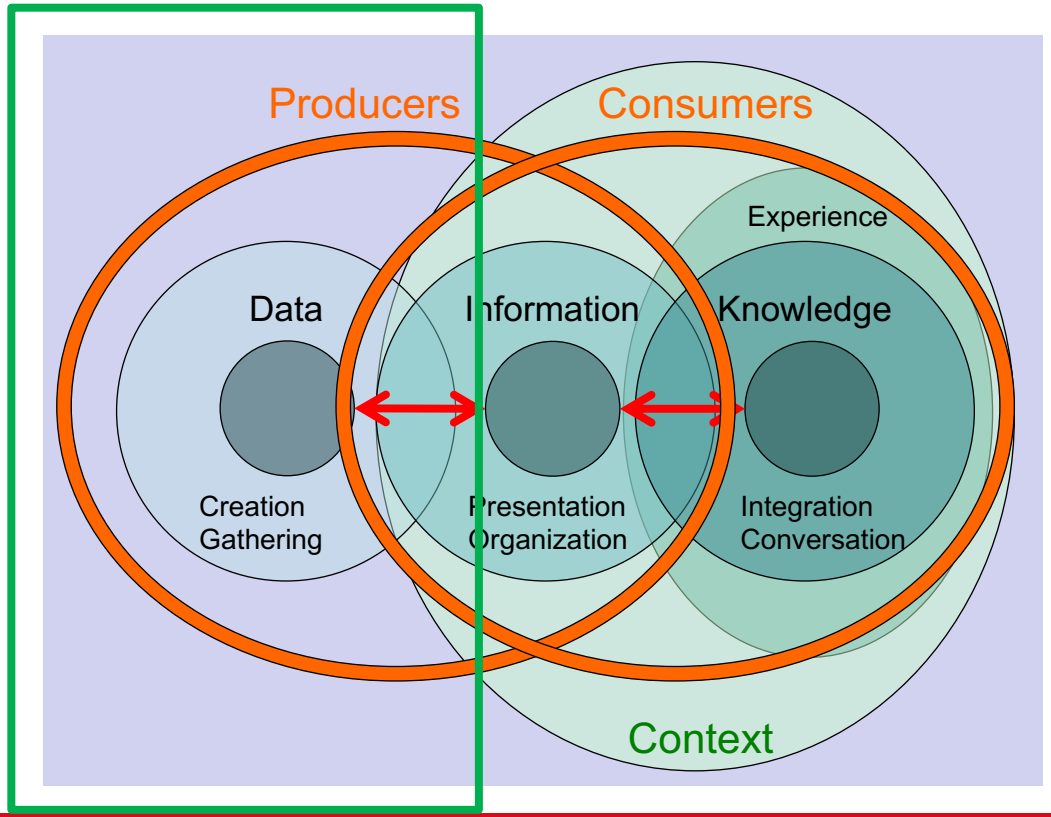
# Shifting the Burden from the User to the Provider



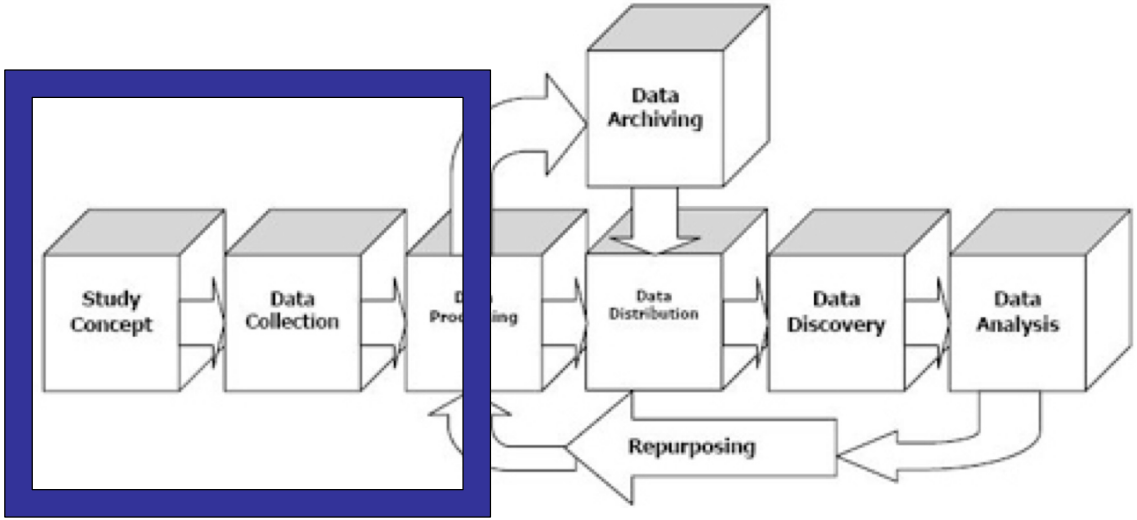
# Focus for the rest of this module

- Data management
- Data and metadata formats
- Data life-cycle
    - Acquisition & curation
    - Bias
    - Provenance
- Example data collection exercises

# Data-Information-Knowledge Ecosystem



# DDI Alliance Life Cycle




https://en.wikipedia.org/wiki/Data_Documentation_Initiative

Image Resource: Data Documentation Initiative

# Modes of collecting data, information

- Observation
- Measurement
- Generation

- Driven by
  - Questions
  - Research idea
  - Exploration

# Data Management

- Creation of logical collections
- Physical data handling
- Interoperability support
- Security support
- Data ownership
- Metadata collection, management, and access
- Persistence
- Knowledge and information discovery
- Data distribution and publication

# Creation of logical collections

– The primary goal of a data management system is to abstract physical data into logical collections. The resulting view of the data is a uniform homogeneous library collection.

# Logical Collections

- Identifying naming conventions and organization scheme.
- Catalogue, index to facilitate search, access, use.
- Provide contextual information.
- Related to metadata – why?

# Physical data handling

– This layer maps between the physical data and the logical data views. Here you find items like data replication, backup, caching, etc.

# Physical Data Handling

- Where does the data come from?
- How is it transferred into a physical form?
- Backup, archiving, and caching...
- Data formats
- Naming conventions

# Interoperability support

– Users may need to use data from different sources in the same application OR use the same data may be used in different applications.. How can that be made possible?

# Interoperability Support

- Using platform/OS neutral encodings, protocols, etc.
- Providing programming or application interface access
- Using data structure and metadata conventions

– "*Smallest number of things to agree on so that you do not need to agree on anything else…" Peter Fox*

# Security support

– Controlling data access, authorization and change verification. This provides a basis for trust your data.

# Security

- What mechanisms exist for securing data?
- Who performs this task?
- Change and versioning (yes, the data may change), who does this, how?
- Who has access?
- How are access methods controlled, audited?
- Who and what – authentication and authorization?
- Encryption and data integrity.

# Data ownership

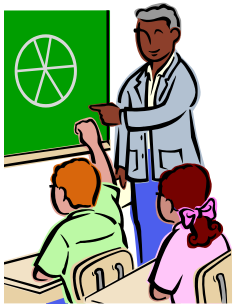
– Define who is responsible for data quality and meaning.

# Data Ownership

- Rights and policies – definition and enforcement.
- Limitations on access and use.
- Requirements for acknowledgement and use.
- Who and how is quality defined and ensured?
- Who may ownership migrate too?
- How to address replication?
- How to address revised/ derivative products?

# Metadata collection, management and access

– Metadata are data about data!



# Metadata

- Know what conventions, standards, best practices exist.
- Use them – can be hard, use tools.
- Understand costs of incomplete and inconsistent metadata.
- Understand the line between metadata and data and when it is blurred.
- Know where and how to manage metadata and where to store it (and where not to).
- Metadata CAN be added later in many cases.

# Persistence

– Making data available beyond the scope of their initial use.

# Persistence

- Where will you put your data so that someone else (e.g. one of your class members) can access it?
- What happens after the class, the semester, after you graduate?
- What other factors are there to consider?

# Data discovery

– Ability to find and access data/datasets with minimal knowledge about their contents.

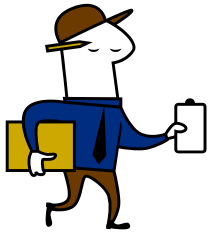
# Discovery

• If you choose so (see ownership and security), how does someone find your data?

• How would you provide for discovery of variables (in the data), versus collections/files?

# Data distribution and publication (dissemination)

– Mechanism to make the interested parties aware of changes and additions to the collections.



# Dissemination

- Who should do this?
- How and what needs to be put in place?
- How to advertise?
- How to inform about updates?
- How to track use, significance?

# Data Encodings, Formats – preview

- ASCII, UTF-8, ISO 8859-1
- Self-describing formats
- Table-driven
- Markup languages and other web-based
- Databases
- Graphs
- Unstructured

# Metadata formats


- ASCII, UTF-8, ISO 8859-1
- Table-driven
- Markup languages and other web-based
- Database, graphs, ...
- Unstructured
- Look familiar? Yes, same as data!

- Some metadata standards
  - Dublin Core (DCMI)
  - METS (METS: Metadata Encoding and Transmission Standard)
  - ISO in general, e.g. ISO/IEC 11179
  - Geospatial, ISO 19115-2, FGDC
  - Time, ISO 8601, xsd:datetime

# Considerations

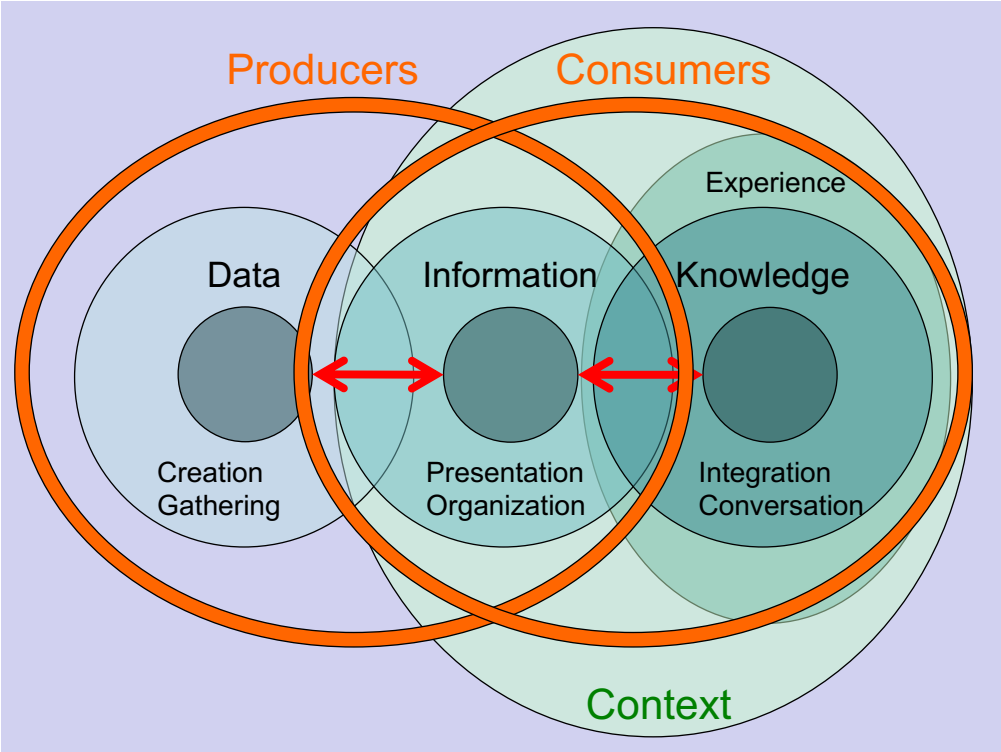
Information has content, context and structure. The notion of "unstructured" really means information that is "unmanaged" with conventional technologies (i.e., metadata, markup and databases).

Databases, metadata and markup provide control for managing digital information, but they are not convenient. This is why less than 20% of the available digital records are managed with these technologies (i.e., conventional technologies are not scalable).

Search engines are extremely convenient, but provide limited control for managing digital information (i.e., long lists of ranked results conceal relationships within and between digital records). The search engine problem is that accessing more information does not equal more knowledge.

We already have effectively infinite and instantaneous access to digital information. The challenge is no longer access, but being able to objectively integrate information based on user-defined criteria independent of scale to discover knowledge.

# Data-Information-Knowledge Ecosystem



# Acquisition

- Learn / read what you can about the developer of the means of acquisition
  - Even if it is you (the observer)
  - Beware of **bias**!!!
- Document things
  - See notes from Class 1
- Have a checklist (see Management) and review it often
- Be mindful of who or what comes after your step in the data pipeline

# Modes of collecting data, information

- Observation
- Measurement
- Generation

- Driven by
  - Questions
  - Research idea
  - Exploration

# Information as a basis for data

• Don’t over think this... Data extracted from an information source, e.g. a web page, an image, a table

• If information is data in context (for human use) then there is data behind the information, e.g. name, address, for a web page form, measure of intensity of light for an image, numerical values for a table

**•But data can also be acquired from information with a different context, e.g. the number of people in an image that are wearing green**

# Curation (partial)

- Consider the organization and presentation of your data
- Document what has been (and has not been) done
- Consider and address the provenance of the data to date, put yourself in the place of the next person
- Be as technology-neutral as possible
- Look to add information and metainformation

# Bias

- To incline to one side; to give a particular direction to; to influence; to prejudice; to prepossess. [1913 Webster]

- A partiality that prevents objective consideration of an issue or situation [syn: prejudice, preconception]

- For acquisition – sampling bias is your enemy

# Sources of Bias

• Sampling/Selection bias: where some members of a population are over or under-represented in the sample.. example

* **Sampling** and **selection** bias are often used synonymously and sometimes **sample selection** bias

• Confirmation bias: when a researcher looks for patterns to confirm their preconceived opinions..

• Publishing bias: where researchers are more likely to publish positive or interesting findings rather than negative results which may be as important

# Provenance*

• Origin or source from which something (data) originates, intention for use, for whom/what was it generated/manufactured/produced, manner of creation, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility.

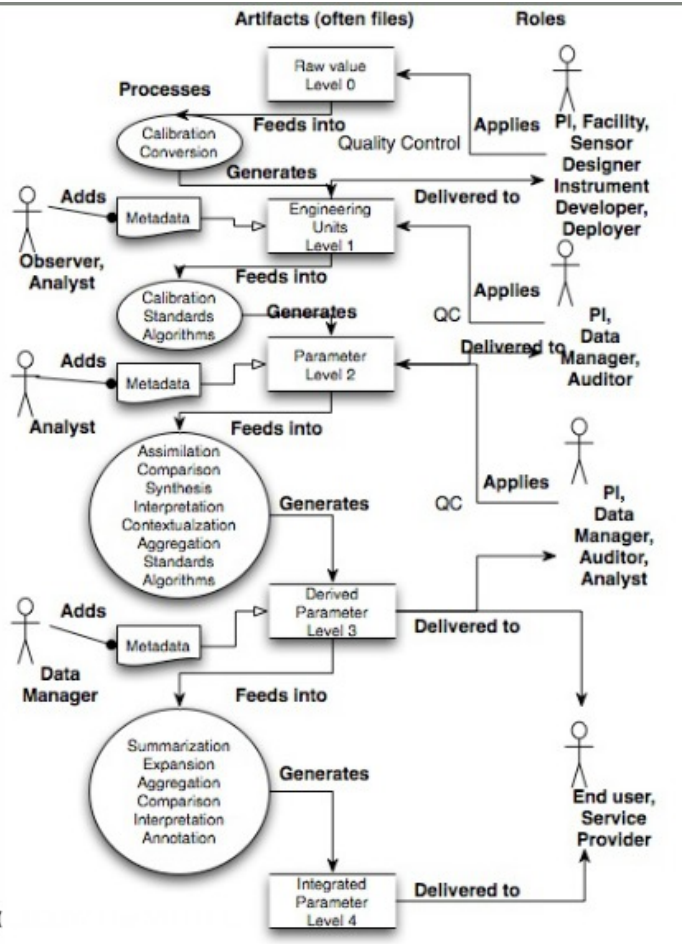
• Internal?
• External?
• Mode?


Image/Photo Credit: http://www.brainwashed.com/coil/images/others/provenance.jpg

# When thinking of provenance

- Who?
- What?
- Where?
- Why?
- When?
- How?

- Provenance in this data pipeline
- Provenance is metadata in *context*
- What context?
  - Who you are?
  - What you are asking?
  - What you will use the answer for?

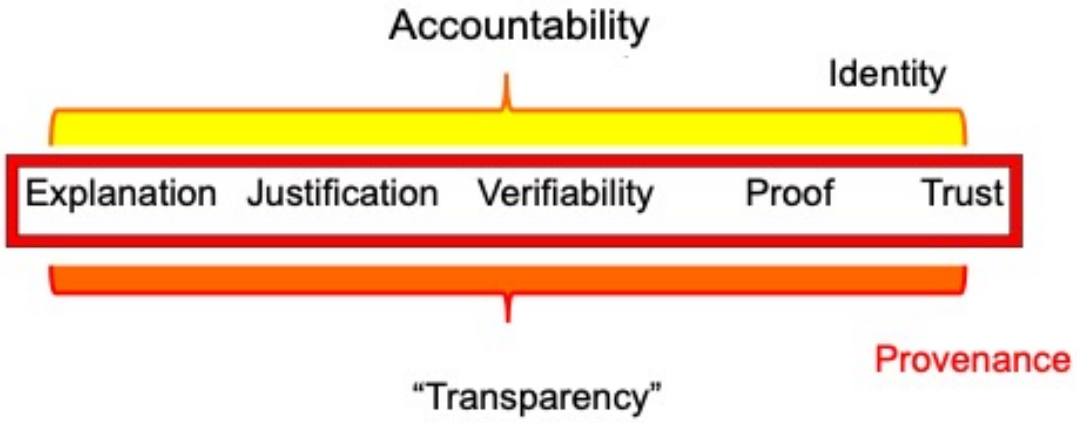


# At the least

- Keyword-value pair (contextual form of metadata – more on this next week)
  – Obs_start_time="Thur 15 Sep 2022 19:22:30 EST" is this okay?
  – Observer="John Doe" – is this okay?
  – You get the idea?

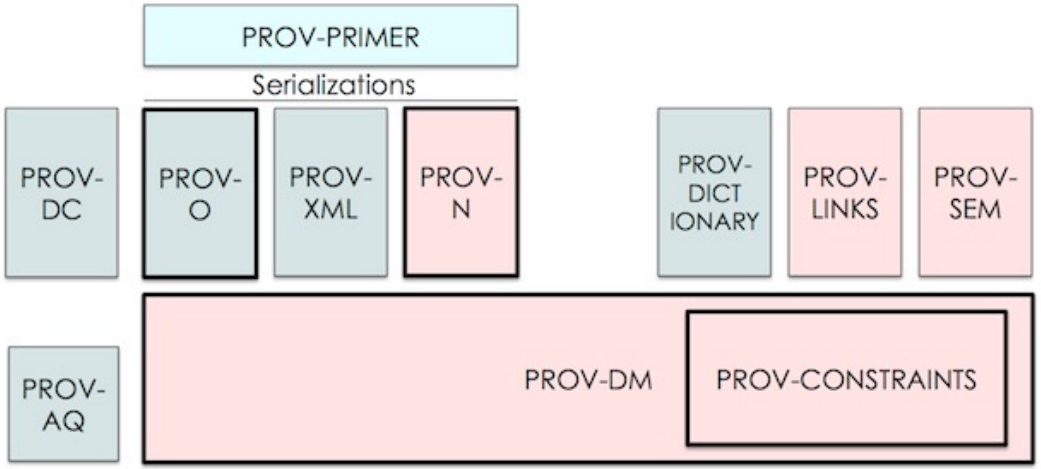
# It is an entire ecosystem

- The elements that make up provenance are often scattered
- But these are what enable data scientists to explore/confirm/deny their data science investigations



# Provenance standards

- International standards:
  - ISO lineage (see reading)
  - PROV (reading)




- Old ones...
  - Open Provenance Model
  - Proof Markup Language

**More on this in during later lectures....**

# Assignment 1

- Propose **two** data collection exercises by writing a data management plan, perform a survey of data formats, metadata and application support for data management suitable for the data you intend to collect in two weeks (10% of grade)

Assignment 1 instructions

- **Note: Assignment 1 is due on Sept 25th, 2026, 10PM ET**

# Example 1

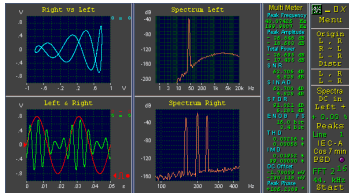
- “the record of the time when the CDTA bus 87 arrives at the bus stop on 15th street under the RPI walk over bridge. The data collection need is being driven by the desire to have a **more precise idea of the time when the bus will arrive at that bus stop** in the hopes that it will be closer to reality than the official CDTA schedule for bus 87.”

# Example 1

• “the record of the time when the CDTA bus 87 arrives at the bus stop on 15th street under the RPI walk over bridge. The data collection need is being driven by the desire to have a **more precise idea of the time when the bus will arrive at that bus stop** in the hopes that it will be closer to reality than the official CDTA schedule for bus 87.”

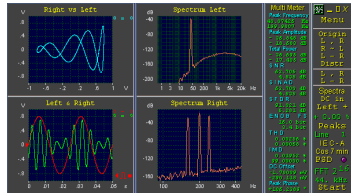
• Lessons:
– Other buses, hard to see the bus, calibrated time source, unanticipated metadata, better to have prepared tables for recording, ...

# Example 2




- ‘The goal of the data collection was to **explore the relative intensity of the wavelengths in a white-light source through a colored plastic film**. By measuring this we can find properties of this colored plastic film.’
- ‘We used a special tool called a spectrometer to measure the relative intensity of this light. It’s connected to a computer and records all values by using a software program that interacts with the spectrometer.’

# Example 2



• ‘The goal of the data collection was to **explore the relative intensity of the wavelengths in a white-light source through a colored plastic film**. By measuring this we can find properties of this colored plastic film.’
•‘We used a special tool called a spectrometer to measure the relative intensity of this light. It’s connected to a computer and records all values by using a software program that interacts with the spectrometer.’

• Lessons
– Noise from external light, inexperience with the software, needed to get help from experienced users, more metadata than expected, software used different logical organization, ...

# Metadata/Provenance

- Required:

DMP Tool (about) [14]

- Choose 2 from:

Data Lineage [10]
Earth Science Information Partners Data Management Workshop: [11]
Earth Science Information Partners: [12]
Data Management Plan - Univ. Minnesota [13]
Machine Actionable Plans (MAP) Case Studies [15]
Metadata and Provenance Management [16]
Provenance Management in Astronomy (case study) [17]
Web Data Provenance for QA [18]
W3 PROV Primer [19]
W3 PROV Data Model [20]

- Submit summaries for 3 readings by email (eleisa2@rpi.edu)

# A Note on the Readings

- They are meant to enlighten not burden. If there is a section you do not understand, do not go down rabbit holes. Mention that in the summary and ask me to elaborate in class/office hours.

- Some readings contain links with relevant information. Where possible, follow the links and include that content in the summary, especially where the original reading item only contains summary information and links.

# Next class: Sept. 16th – Data formats, metadata standards, conventions, reading and writing data and information

# Thanks!