Rensselaer

why not change the world?®

Example of Data Science and Data and information acquisition (curation) and metadata/ provenance - management

Ahmed Eleish

Data Science – ITWS/CSCI/ERTH 4350/6350

Week 2, September 5th, 2024

Admin info (keep/ print this slide)

- Class: ITEC/CSCI/ERTH-4350/6350
- **Hours: 11:00 am - 1:50pm ET on Thursdays**
- **Location: Lally room 104**
- **Instructor:** Ahmed Eleish
- **Instructor contact:** eleisa2@rpi.edu
- **Instructor office hours:** Tue from 12:30 PM - 1:30 PM ET/Wednesday from 12:00 PM - 1:00 PM ET or by appointment/email
- **Instructor office location:** Amos Eaton 134
- **TA:** Benita Chinemerem **-** chineb@rpi.edu
- **TA office hours:** Tuesday 12-2pm / Friday 12-1pm
- **TA office:** Lally 205
- Web site: https://tw.rpi.edu/classes/data-science-2024

Course Assignments

- Assignment 1 (available 09/05 – due 09/19) – 10% (written)
- Assignment 2 (available 09/12 – due 09/26) – 15% (written) / 5% (presentation)
- Assignment 3 (available 10/03 – due 10/17) – 20% (written)
- Assignment 4 (available 10/24 – due TBA) – 25% (written) / 5% presentation/poster
- Assignment 5 (available 11/07 – due 12/01) – 10% (written)

Reading Assignments

- Changing Science: Chris Anderson: [1]
- Rise of the Data Scientist [2]
- Where to draw the line? [3]
- Career of the Future [4]
- What is Data Science (I) [5]
- Data Science vs. Data Analytics [6]
- Data Scientist: The Hottest Job You've Never Heard Of [7]
- What Is a Data Scientist? [8]
- Data Scientist - sexiest job of the 21st Century ? [9]
- Big Data [10]
- A Very Short History of Data Science [11]
- 7 Phases of Data-Life-Cycle [12]

Review from last week

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

Story telling with Data

- https://youtu.be/jbkSRLYSojo

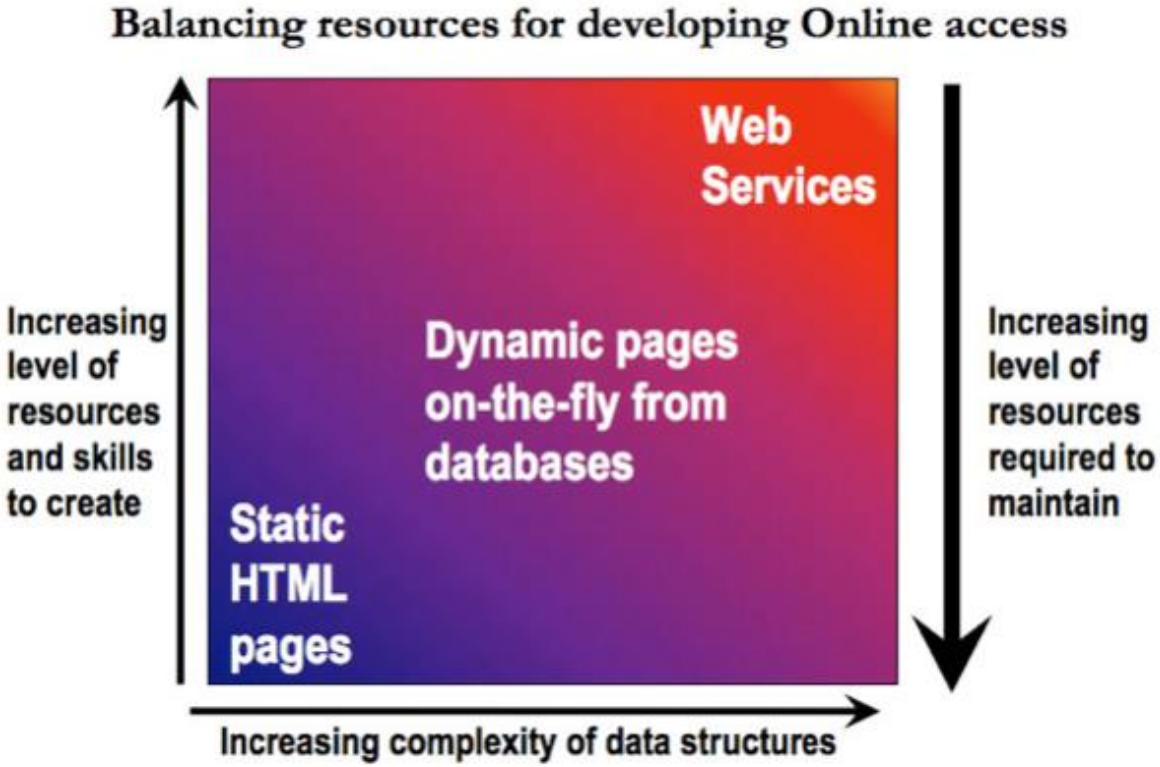
Components of Data Science

- Data life cycle – acquisition, curation and preservation
- Data management and 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
- Presenting findings and conclusions

Shifting the Burden from the User to the Provider



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

Data pipelines: we have problems

- Data is coming in faster, in greater volumes and forms and outstripping our ability to perform adequate quality control.

- Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision.

- We often fail to capture, represent and propagate manually generated information that need to go with the data flows.

- Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects.

- The task of event determination and feature classification is onerous and we don't do it until after we get the data.

- And now much of the data is on the Internet/Web (good or bad?)

Provenance

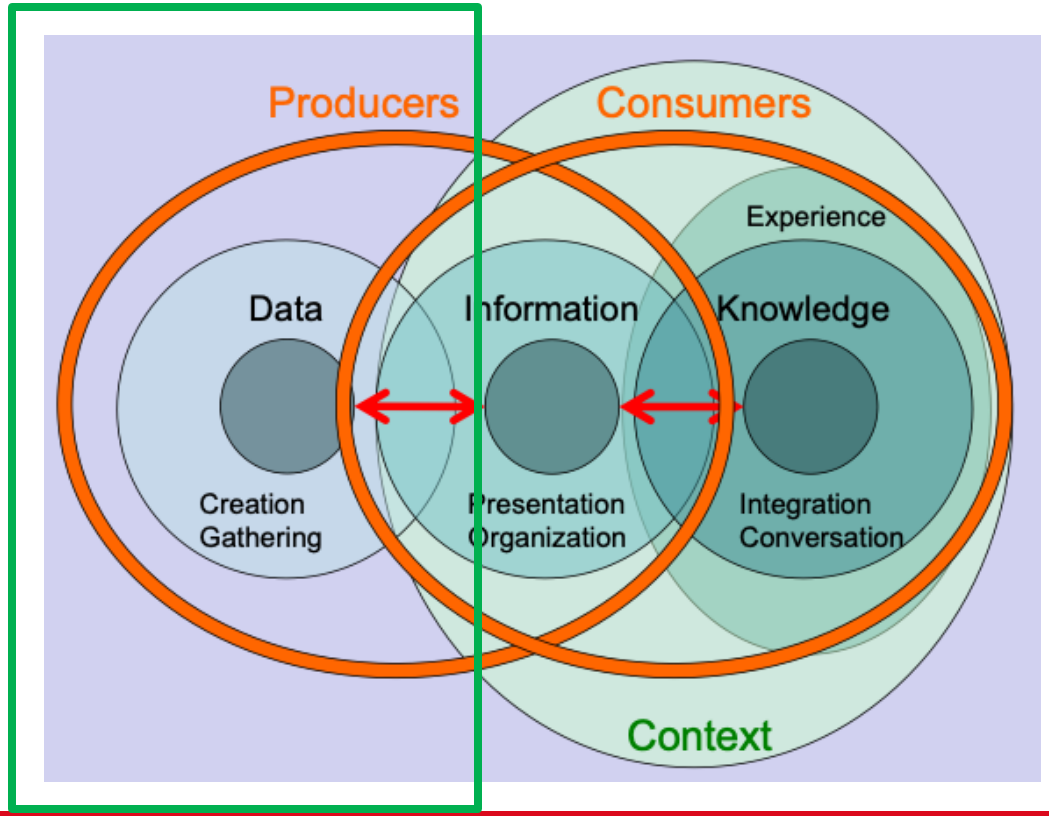
• Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility

• More on this later in the semester...

Focus for the rest of this module

- Preparing for data collection
- Managing data
- Data and metadata formats
- Data life-cycle : acquisition
  - Modes of collecting
  - Examples
  - Information as data
  - Bias, provenance
- Curation

Data-Information-Knowledge Ecosystem



MIT DDI Alliance Life Cycle

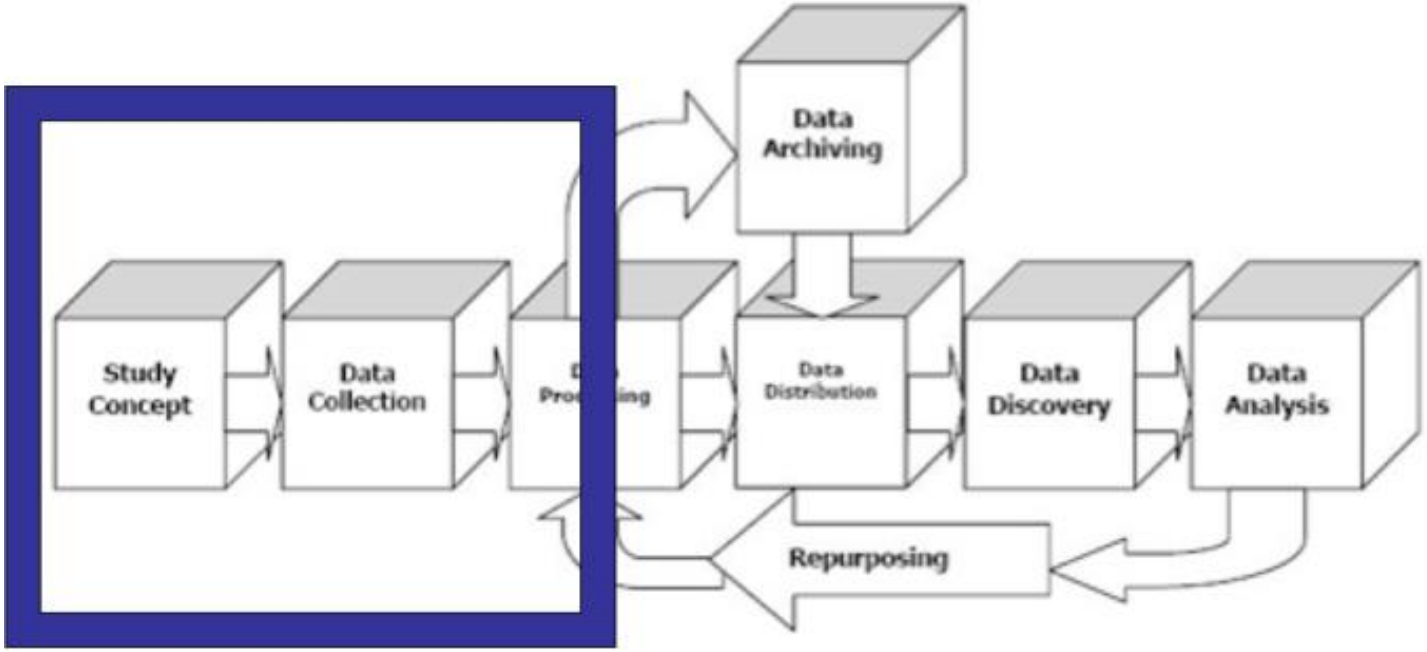


Image Resource: MIT Data Documentation Initiative

Modes of collecting data, information

- Observation
- Measurement
- Generation

- Driven by
  - Questions
  - Research idea
  - Exploration

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.

Management

- Creation of logical collections

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

- Physical data handling

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

Management

• Interoperability support
– Normally the data does not reside in the same place, or various data collection (like catalogues) should be put together in the same logical collection.

• Security support
– Data access authorization and change verification. This is the basis of trusting your data.

• Data ownership
– Define who is responsible for data quality and meaning

Management

- Metadata collection, management and access
  – Metadata are data about data

- Persistence
  – data need not change if the underlying technology changes

- Knowledge and information discovery
  – Ability to identify useful relations and information inside the data collection

Management

- Data distribution and publication

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

Logical Collections

- Identifying naming conventions and organization
- Aligning cataloguing and naming to facilitate search, access, use
- Provision of contextual information
- Related to metadata – why?

Physical Data Handling

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

Interoperability Support

- Bit/byte and platform/wire neutral encodings
- Programming or application interface access
- Data structure and vocabulary (metadata) conventions and standards

- Definition of interoperability?

– Smallest number of things to agree on so that you do not need to agree on anything else

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

- 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

- 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

• 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?

Discovery

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

• How would you provide discovery of collections, versus files, versus ‘bits’?

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 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 (dc.x)
  - – Encoding/wrapper standards – 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

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

Example 1

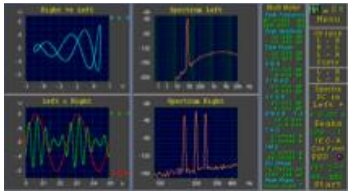
• “the record of the time when the CDTA bus 87 arrives at the bus stop on 15th street under the RPI walk over bridge. The data collection need is being driven by the desire to have a **more precise idea of the time when the bus will arrive at that bus stop** in the hopes that it will be closer to reality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

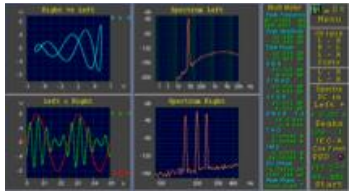
Example 2



• ‘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, ...

Example 3

• ‘The goal of my data collection exercise was to observe and generate historical stock price data of large financial firms within a specified time frame of the years 2007 to 2009. This objective was primarily driven by general questions and exploration purposes – in particular, a question I wanted to have answered was **how severe the ramifications of the economic crisis were on major financial firms**.’

Example 3

• ‘The goal of my data collection exercise was to observe and generate historical stock price data of large financial firms within a specified time frame of the years 2007 to 2009. This objective was primarily driven by general questions and exploration purposes – in particular, a question I wanted to have answered was **how severe the ramifications of the economic crisis were on major financial firms**.’

• Lessons
– Irregularities in data due to company changes (buy-out, bankrupt), no metadata
– had to create it all, quality was very high, choice of sampling turned out to be crucial, ...

Example 4

• I performed a survey among a sample set of people to
• determine how many prefer carbonated drinks (like Coke) to fruit juice.

The goal of this data collection exercise was to determine **which option is more popular and if any health-related issues occur due to the consumption of these drinks**. The data collection need was primarily driven by the question - whether consumption of caffeine, soda and excessive sugar present in these drinks actually cause health problems like obesity, cholesterol, dental decay etc. The mode of data collection was by observation.

Example 4

• I performed a survey among a sample set of people to
• determine how many prefer carbonated drinks (like Coke) to fruit juice. The goal of this data collection exercise was to determine **which option is more popular and if any health related issues occur due to the consumption of these drinks**. The data collection need was primarily driven by the question - whether consumption of caffeine, soda and excessive sugar present in these drinks actually cause health problems like obesity, cholesterol, dental decay etc. The mode of data collection was by observation.

• Lessons

– The measurement unit for the amount of drink consumed daily was notf ixed before starting the data collection exercise. During the data collection process, some gave me the amount in ml whereas some in ounces and some others in number of glasses. Later, I had to convert those units to the standard unit that I was using – ml.

– Some people were reluctant to disclose health related issues and I had to guarantee them anonymity. This solved the problem to a great extent

Example 5

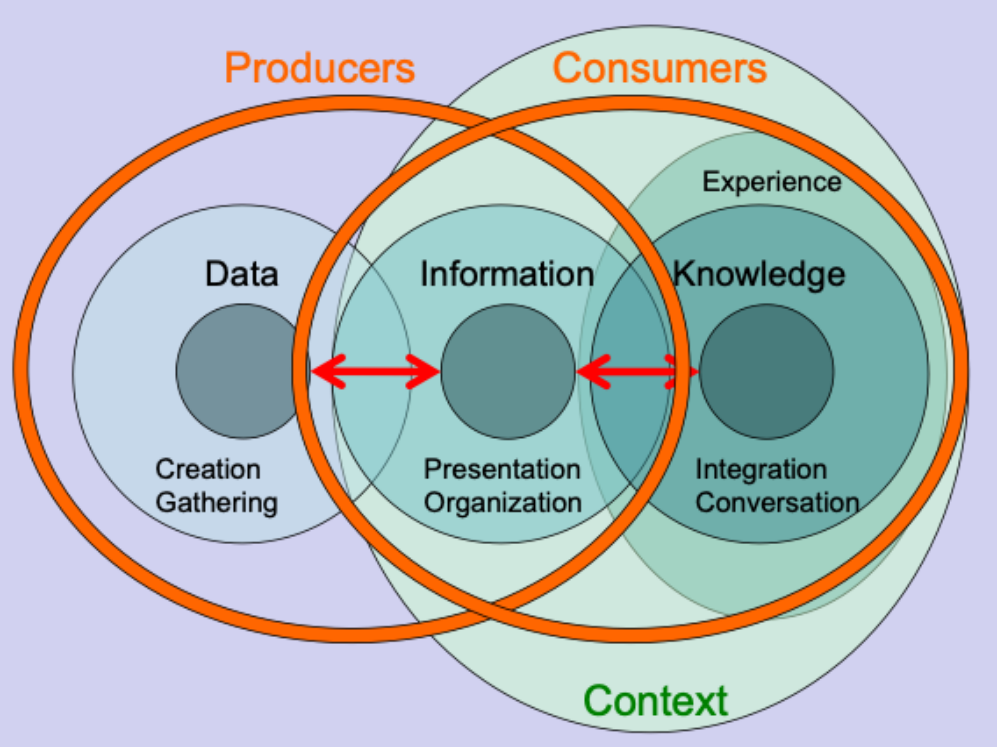
• The goal of my data collection exercise is to generate the average rainfall levels during the hurricane seasons from the year 2005-2010. Along with this data, we would generate data on the average height of the Hudson River during the hurricane season months. The data collection need is being driven by questions such as how likely will it be for the Hudson River to cause major flood damage during the hurricane season to the cities surrounding it .

Example 5

• The goal of my data collection exercise is to generate the average rainfall levels during the hurricane seasons from the year 2005-2010. Along with this data, we would generate data on the average height of the Hudson River during the hurricane season months. The data collection need is being driven by questions such as how likely will it be for the Hudson River to cause major flood damage during the hurricane season to the cities surrounding it .

• Lesson(s): some historical data is very hard to find. Places where average height is measured do not provide optimal comparison with where rainfall is measured.

Data-Information-Knowledge Ecosystem



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

Bias

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

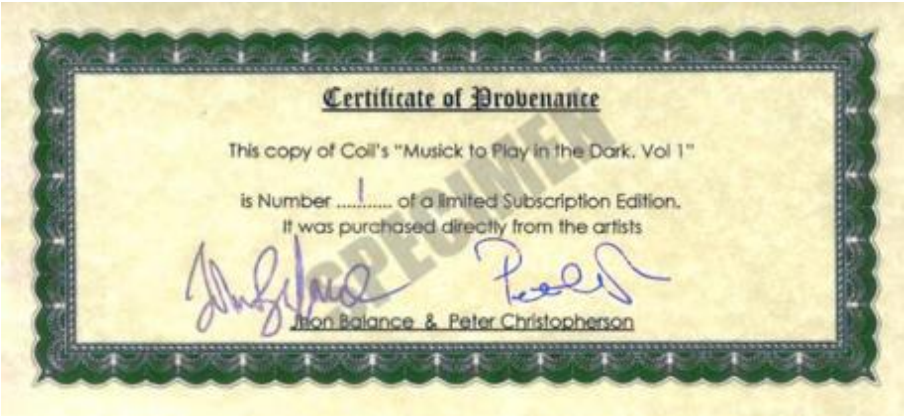
• A partiality that prevents objective consideration of an issue or situation [syn: prejudice, preconception]

• For acquisition – sampling bias is your enemy

Provenance*

• Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility

• Internal?
• External?
• Mode?

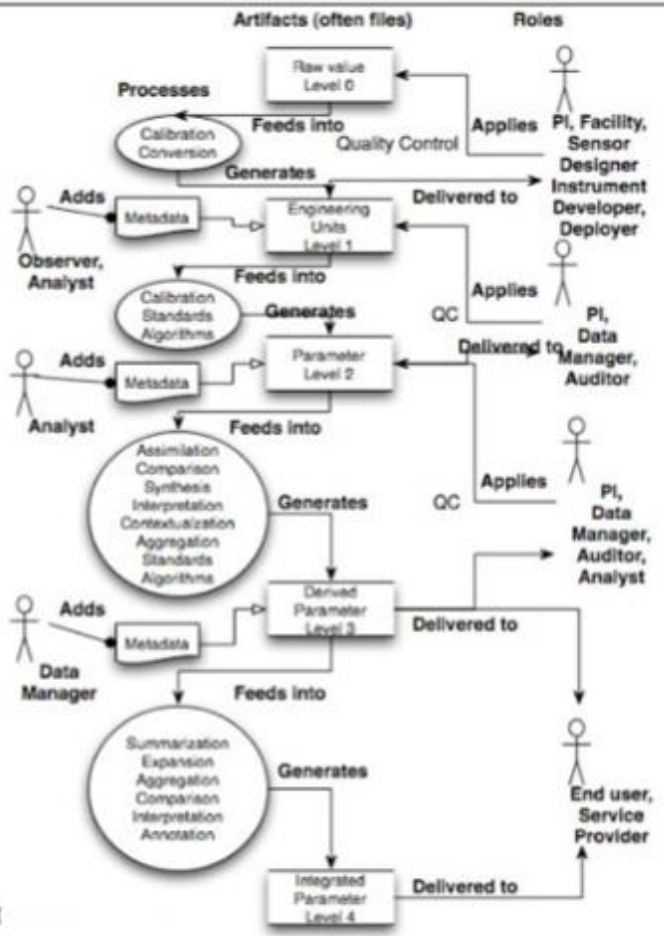


Image/Photo Credit: http://www.brainwashed.com/coil/images/others/provenance.jpg

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?

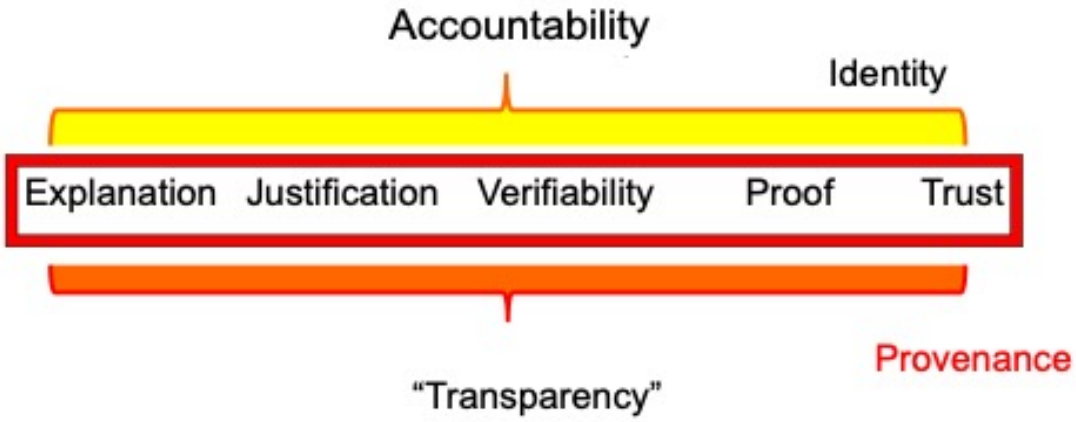


At the least

- 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?

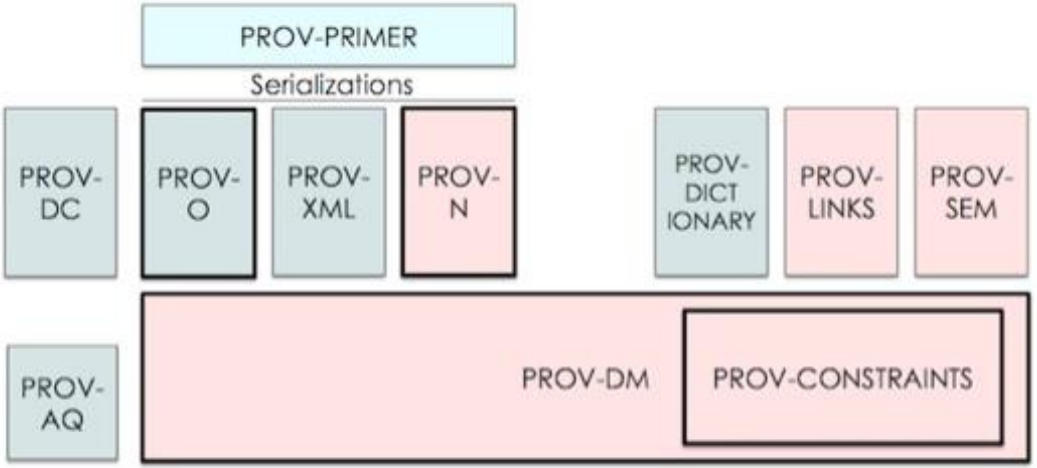
It is an entire ecosystem

- 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....

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

Assignment 1

• Propose two data collection exercises by writing a data management plan and 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) – see LMS for the Assignment 1 link.

• **Note: Assignment 1 is due on Sept 19th, 2024, 11AM EST**.

What is next

- Reading – see web page/ Module 2 on course web page (Data Management, Provenance)

- Next (Data formats, metadata standards, conventions, reading and writing data and information)