



Rensselaer

why not change the world?®

Definitions, History of Data (and Information), Data Science, Data Lifecycle, Current Challenges

Ahmed Eleish

Data Science ITWS/CSCI/ERTH-4350/6350

Week 1, September 2nd, 2026

Tetherless World Constellation
Rensselaer Polytechnic Institute



Course webpage:

tw.rpi.edu/classes/data-science-2026



Admin information

- Class: ITWS-4350, ITWS-6350, CSCI-4350, CSCI-6350, EARTH-4350, EARTH-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: Rei Uchida
- TA office hours: TBA
- TA office: TBA
- Web site: <https://tw.rpi.edu/classes/data-science-2026>
- LMS (<http://lms.rpi.edu/>)



Contents

- Learning objectives
 - What is expected
 - Assignments
 - What skills are needed
-
- Outline of the course
 - Definitions
 - History of data and information
 - What is data science; who is a data scientist
 - Data life cycle



Learning Objectives

- Instruct future scientists how to sustainably generate / collect and use data for their research as well as for others.
- Instruct future technologists how to understand and support essential data and information needs of a wide variety of producers and consumers.
- To develop students' knowledge of the full life-cycle of data and relevant frameworks, technologies, standards, and best practices.
- To train students in the methods, applications, and tools necessary to properly handle data and information.
- To introduce students to methods for data analysis, data visualization, and the development and evaluation of predictive models.



Learning Outcomes

- Through class lectures, readings, written reports and oral presentation assignments and projects, students should:
- Develop and demonstrate skill in data collection and management.
- Develop data models and generate metadata.
- Demonstrate experience in the application of data and metadata standards.
- Demonstrate skill in data science tool use and evaluation of the results.
- Demonstrate knowledge of the principles of the data life-cycle.
- Become proficient in data and information product generation.
- **(6000 level)** Demonstrate knowledge of conventions and best practices in data stewardship.



What is expected

- Attend class, complete assignments
- Submit summaries of reading assignments
- Ask questions, provide answers (if you have them)
- Work both individually and in a group
- Work constructively in group and class sessions



4000 Level / 6000 Level

- 6000 Level students are assessed at:
 - Higher level of demonstration of knowledge
 - Additional questions or tasks in assignments
- 4000 Level are welcome to complete these higher requirements for extra credit



Assessment and Assignments

- Written assignments with the specific percentage of grade allocation provided with each assignment.
- Projects with written reports and group oral presentations with the specific percentage of grade allocation provided.
- Reading assignments (10% of total)
- Late submission policy: first time with valid reason no penalty, otherwise 20% of score deducted each late day



Reading Assignments

- Reading assignments are given almost every week
- At the end of the lecture I will share links to the readings, 1-2 will be mandatory and you will select another 1-2, the specifics will be in the lecture slides, then posted on the course webpage + LMS.
- Most are background and informational, some are key to completing assignments
- You are required to submit **short** summaries of the assigned readings by the following class session (1 week).
- Summaries should be short! (max 3 paragraphs per reading)



Assignments

- 5 assignments
- 2 group projects (3 week data collection exercise, 6 week data analysis project)
- You will be given guidance in finding data for your data analysis project.
- At the end of each project, you and your group members will submit a paper (report) and present your project results.



Current assignment structure (no final exam)

- Assignment 1 Data Collection Exercise (Proposals) (09/09 – due 09/25) – 10% (written)
- Assignment 2 Data Collection Exercise (09/25 – 10/14) – 15% (written) / 5% (presentation)
- Assignment 3 Data Analysis Project (10/14 – 12/11) – 25% (written) / 5% presentation
- Assignment 4 Data Archival (10/28 – 11/18) – 20% (written)
- Assignment 5 (11/18 – 12/09) – 10% (written)
- Reading Summaries (10%)



Skills needed

- Literacy with computers and applications that can handle data
- Basic knowledge of data structures, computer programming
- Ability to access internet and retrieve / acquire data
- Presentation of assignments / results
- Working alone and collaborating with others
- Problem solving / patience 😊



Academic Integrity

- Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these. In this class, all assignments that are turned in for a grade must represent the student's own work. In cases where help was received, or teamwork was allowed, a notation on the assignment should indicate your collaboration.
- Submission of any assignment that is in violation of this policy will result in a penalty. If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty of full **loss of grade** for the work in violation, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: **warning, probation, suspension, expulsion**, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities.
- Second violation will result in **failure** of the course.
- **If you have any question concerning this policy **before** submitting an assignment, please ask for clarification.**



Current Schedule

- Week 1 (Sep. 02): Definitions, History of Data (and Information), Data Science, Data Lifecycle
- Week 2 (Sept. 09): Data Management, Data and information acquisition (curation, preservation), Provenance
- Week 3 (Sept 16): Data Formats, Metadata Standards, Conventions, Reading and Writing Data and Information
- Week 4 (Sept. 23): Knowledge Graphs, Webs of Data
- Week 5 (Sept. 30): Data Analysis
- Week 6 (Oct. 07): Class Presentations: present your data, Part of Assignment 2
- Week 7 (Oct. 14): Class Presentations: present your data, Part of Assignment 2
- Week 8 (Oct. 21): Data Mining
- Week 9 (Oct. 28): Data in Research, Scientific Workflows and Data Stewardship
- Week 10 (Nov. 04): Responsible Data Science
- Week 11 (Nov. 11): Project Check in
- Week 12 (Nov. 18): Guest Lecture
- Week 13 (Nov. 25): No classes (Thanksgiving break)
- Week 14 (Dec. 02): Final project presentations, written reports due Dec 11
- Week 15 (Dec. 09): Final project presentations, written reports due Dec 11



Questions so far?





Peter Arthur Fox, Ph.D.
1959 - 2021

Introductions

- Who you are, background?
- Why you are here? What you expect to learn?
- Interests/Hobbies?



Data who?



<http://images2.fanpop.com/image/photos/9400000/Lt-Commander-Data-star-trek-the-next-generation-9406565-1694-2560.jpg>



Definitions (at least for this course)

- **Data** - are encodings that represent the qualitative or quantitative attributes of a variable or set of variables.
- **Data** (plural of "datum", which is seldom used) - are typically the results of measurements, computations, or observations.
- **Data** - are *often* viewed as the lowest level of abstraction from which information and knowledge are derived.
- **Information** - data put into context



DIKW model

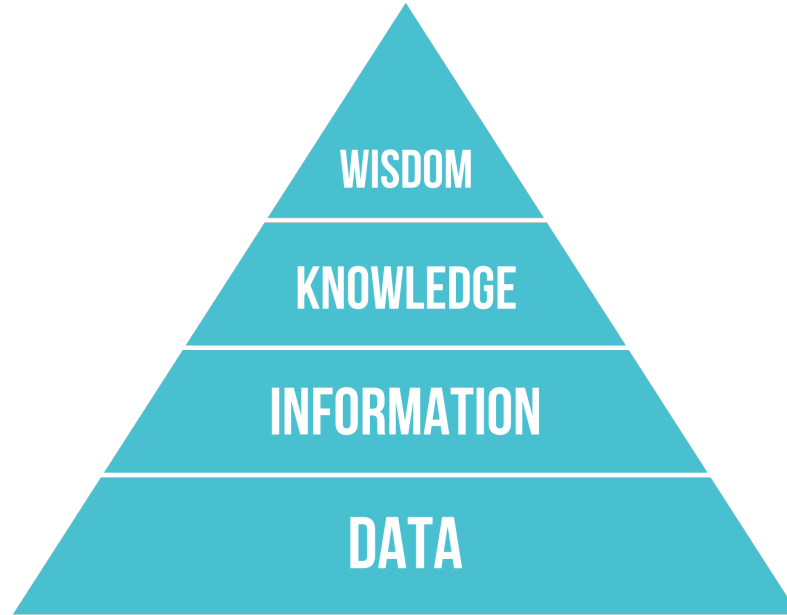


Image Source: https://en.wikipedia.org/wiki/DIKW_pyramid

What is data science?

- Everyone* uses data daily, individually, in business, academia, science..
- ..but merely using data isn't what we mean by “data science.”
- A data application acquires its value from the data and creates more data as a result.
- It's not just an application with data; it's a data product.
- Data science enables the creation of data products.

* pretty much



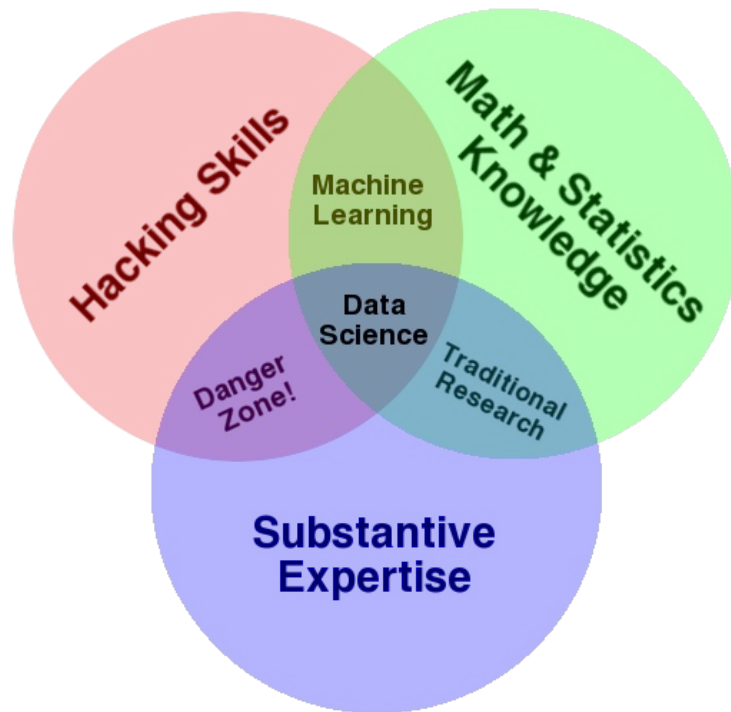
What is data science?

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data.
- Data science is multifaceted and can be described as a science, a research paradigm, a research method, a discipline, a workflow, and a profession.
- To study phenomena through data using computation, statistical analysis, data mining, predictive modeling, and visualization.

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



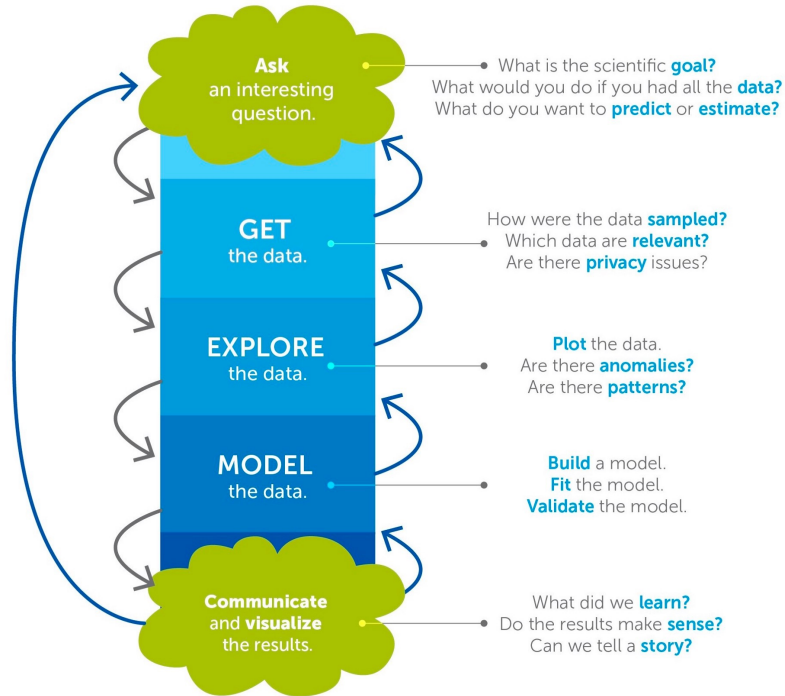
Data Science



<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>



The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.

Image Credit: Joe Bilizsten, Hanspeter Pfister / Harvard Data Science



How Data tells a story

- Walmart growth visualization:
 - <http://projects.flowingdata.com/walmart/>
- Target growth visualization:
 - <http://projects.flowingdata.com/target/>



Who is a Data Scientist?

- “A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

~ Joshua Blumenstock

- “Data Scientist = statistician + programmer + coach + storyteller + artist” ~ Shlomo Argamon



An Early Data Scientist...

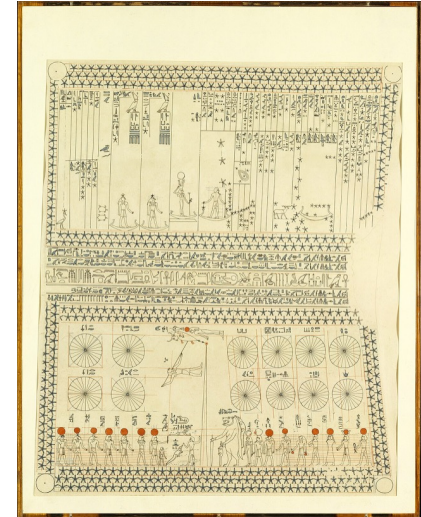
- **Johannes Kepler** did not make direct observations of the planets, he analyzed Tycho Brahe's data to come up with the Kepler's laws of planetary motion, which he published from 1608-1621



Image Credit: <https://www.ancient-origins.net/history-famous-people/how-did-skeptical-astrology-johannes-kepler-contribute-our-view-cosmos-007944>

Data throughout history

- Observation of our natural world – Kepler's laws of planetary motion
- Aurora and sunspots (documented by Chinese historians)
- Celestial diagrams of constellations (recorded by ancient Egyptians)



<https://academic.oup.com/pasj/article-abstract/69/4/65/4004641?redirectedFrom=fulltext>

https://en.wikipedia.org/wiki/Astronomical_ceiling_of_Senenmut%27s_Tomb

Modes of collecting data, information

- Through observations, measurements, and computations...

Examples:

Observation: outcome of an experiment, status of a patient, condition of wildlife

Measurement: thermometer, watch, electrical/mechanical sensor

Generation: credit score, course grade, social media post

Driven by:

- Questions, research idea, exploration ...



68?



42?

“The answer to the great question...of
Life, the Universe and
Everything...is...forty-two.”

Douglas Adams

The Hitchhiker's Guide to the Galaxy

<https://www.scientificamerican.com/article/for-math-fans-a-hitchhikers-guide-to-the-number-42/>



Definitions ctd.

- **Metadata** – data about data
- Metadata are often automatically generated and stored along with the original data, e.g. creating a new file records its type, data of creation, size, etc.
- Metadata may be added manually and stored separately from the data.



DIKW model

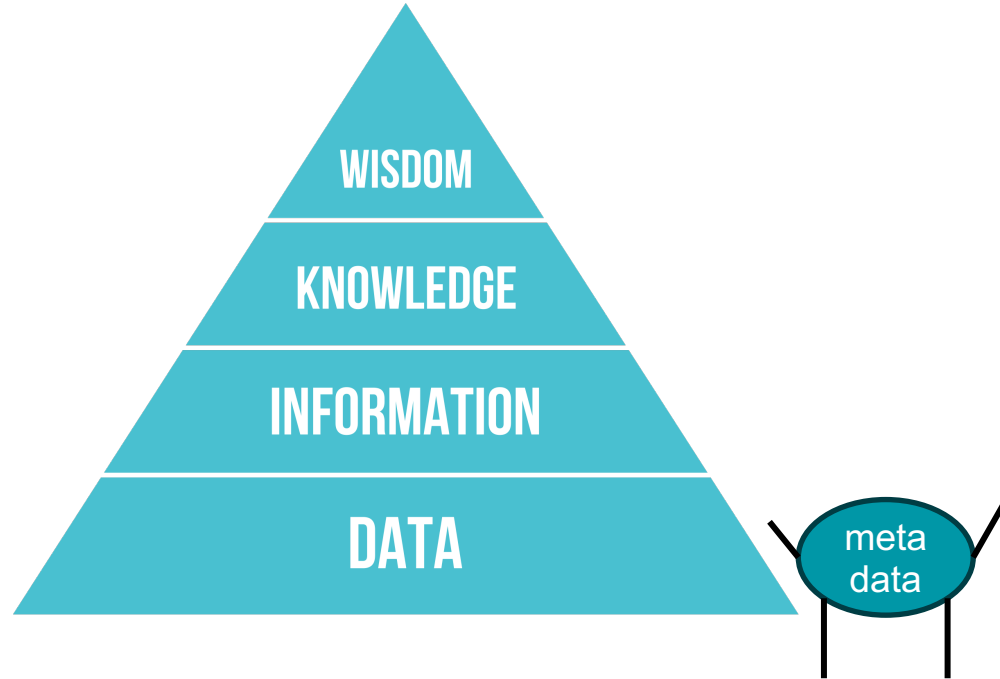


Image Source: https://en.wikipedia.org/wiki/DIKW_pyramid

Examples

- Rock sample:
 - Data: weight, composition, shape, size
 - Metadata: location and time of sample collection
 - Information: distribution of rock types in area
 - Knowledge: evidence of geologic activity
- Weather
 - Data: wind speed and direction, temperature, ..
 - Metadata: type of radar, sensor, model
 - Information: weather map with contours and features
 - Knowledge: high pressure system, stable weather



Definitions ctd.

- **Data Management:** Process of arranging for discovery, access and use of data, information and all related elements (metadata, documentation, etc.)
 - Also oversees or effects control of processes for acquisition, curation, preservation and stewardship.
 - This involves fiscal and intellectual responsibility.



Definitions ctd.

- **Data life-cycle** (3 main components)
 - **Acquisition**: Process of recording or generating a concrete artefact from the concept.
 - **Curation**: The activity of managing the use of data from its point of creation to ensure its availability for discovery and re-use in the future (<https://www.dcc.ac.uk/about/digital-curation>).
 - **Preservation**: The process of retaining usability of data in some form for intended and unintended use
- **Stewardship**: Process of maintaining data integrity for acquisition, curation and/ or preservation



Data Life Cycle Management - 7 Phases

1. Data Capture
2. Data Maintenance
3. Data Synthesis
4. Data Usage
5. Data Publication
6. Data Archival
7. Data Purging

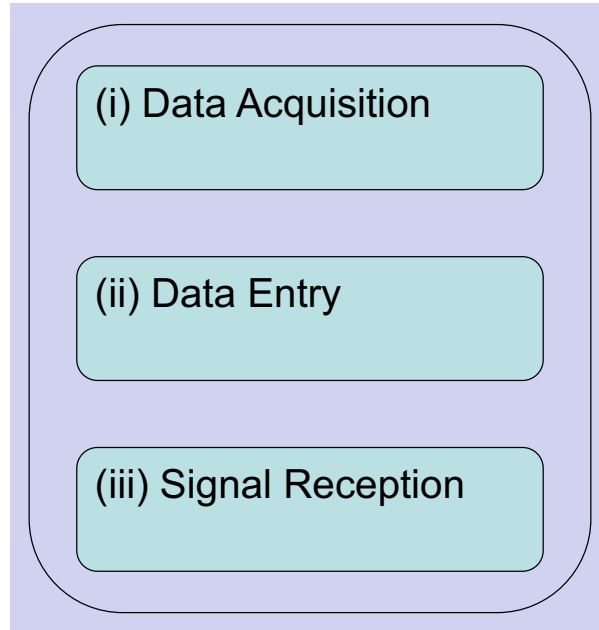
Article on Data Life Cycle:

<https://www.bloomberg.com/professional/blog/7-phases-of-a-data-life-cycle/>



1.Data Capture

The act of creating data values that do not exist and have never existed within the organization.



1. Data Capture

(i) Data Acquisition : The ingestion of already existing data that has been produced by an outside organization.

(ii) Data Entry : The creation of new data values for the organization by human operator or devices that generate data for the organization.

(iii) Signal Reception : The capture of data created by devices, sensors, etc. typically important in control systems, but becoming more important for information systems.



2. Data Maintenance

- Once the data have been captured, data maintenance is usually required, this can be defined as:
 - Supplying the data to a point at which synthesis and usage occur.
- What Data Maintenance is about:
 - Processing the data without deriving any value from it for the enterprise.
- Data Maintenance involves tasks such as:
 - Movement
 - Integration
 - Cleansing



3. Data Synthesis

- The creation of data values via inductive logic, using other data as input.
- Inductive logic requires some kind of expert knowledge, judgement and/or opinion as a part of the logic.

e.g. The way credit score is created.



4. Data Usage:

- The application of data as information to tasks that the organization needs to run and manage itself.
- This would normally be outside of the data life cycle, however, data is becoming more central many organizations. For instance, data may be itself be a product or service that an enterprise offers.
- Data usage has special data governance challenges.
 - Questions of legality and ethics may need to be addressed.



5. Data Publication:

- The sending of data to a location outside of the organization.
- Once the data has been sent outside the organization, it is de facto, impossible to recall it.
- Data values that are wrong can not be corrected as they are beyond the reach of the organization.
- Data breaches also fall under the category of data publication. `



6. Data Archival:

- Copying data to an environment for long term storage.
- The data archive is simply a place where data are stored, but no maintenance, active usage or publication occurs.



7. Data Purging:

- The removal of every copy of data item from the organization.
- Ideally this will be done from an archive.
- A data governance challenge of this phase of the data life cycle is proving that the purge has actually been done properly.



The nature of the challenge

- To do data science today
 - You may play many roles in different parts of the life cycle, or all of them.
 - You may not get all the metadata or information you need even if you get the data.
 - You will need skills that you were not taught.
- To work with data scientists today
 - You might need lots of technical experience.
 - You will need new skills in addressing the changing use of data and information.



The nature of the challenge

- 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 needs to go with the data.
- Each time we develop a **new instrument**, we develop a new data ingest procedure and collect different metadata and organize them differently.
- And now much of the data is publicly available on the Internet/Web.

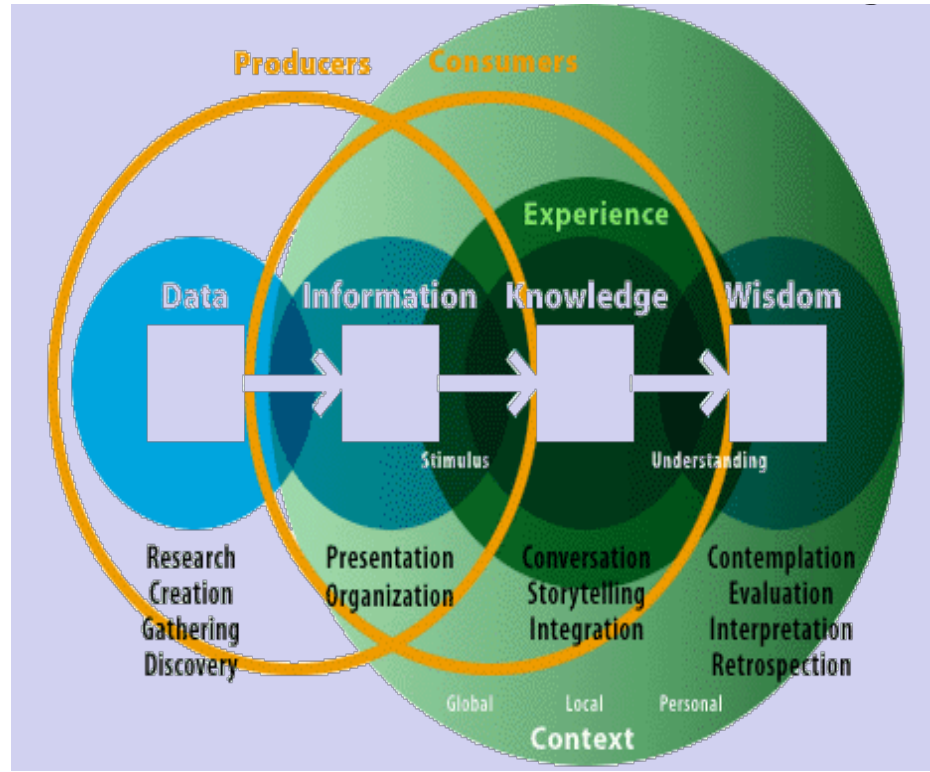


Provenance

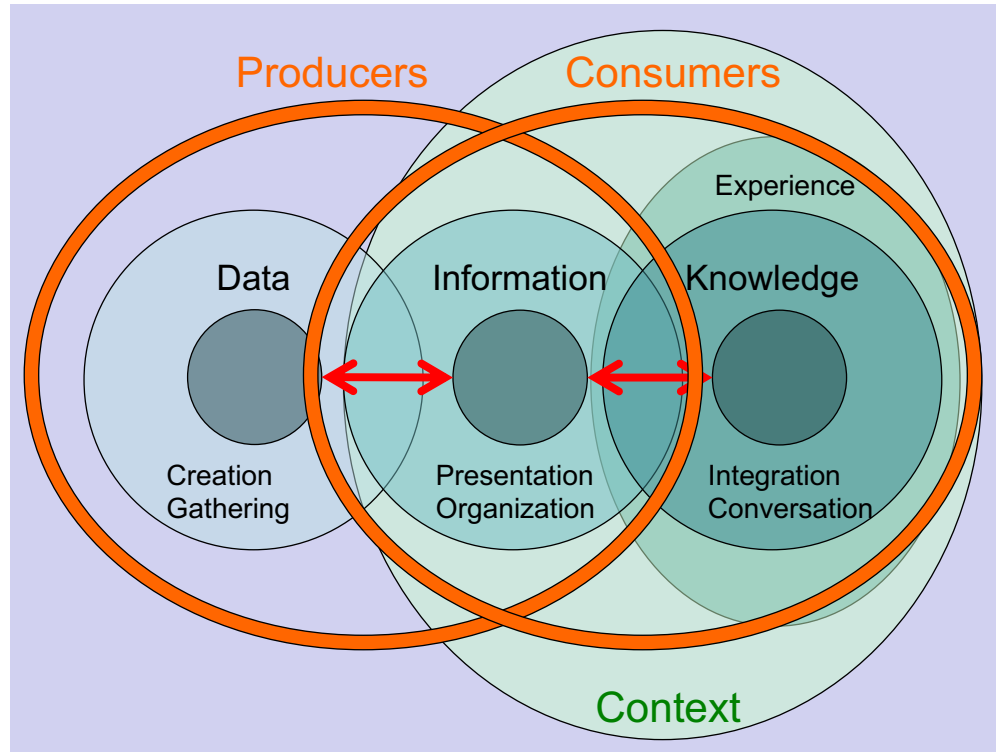
- Origin or source from which something comes, intention for use, who/what it was generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility ..
- “Metadata in a given context”



An OLD view of data, information and knowledge



Data-Information-Knowledge Ecosystem



We will see more on this later...

Summary

- What is data science? Who's a data scientist?
- Data (& metadata), data life cycle, data-information-knowledge..
- Scientists and technologists are not well prepared to cope with 21st century data management and use of tools.



Rise of the Data Scientist

- 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](#)[\[9\]](#)

- Submit summaries for 2 readings to LMS



Project Teams

- Assignments 1, 2, & 3 are group work/projects.
- Teams should have 2-3 members.
- Please form your teams and email me by next class (**September 9th**)



Next class:

Sept. 9th – Data Management, Data and information acquisition (curation, preservation), Provenance.

Submit team member names!

Thanks!

