



# Rensselaer

**why not change the world?®**

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

**Ahmed Eleish**

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

**Week 1, September 3<sup>rd</sup>, 2025**

Tetherless World Constellation  
Rensselaer Polytechnic Institute



# Admin info

- Class: ITEC/CSCI/ERTH-4350/6350
- Hours: **11:00 am - 1:50pm ET on Wednesdays**
- Location: **West Hall Auditorium**
- Instructor: Ahmed Eleish
- Instructor contact: [eleisa2@rpi.edu](mailto:eleisa2@rpi.edu)
- Instructor office hours: Wednesday from 2-4 PM ET or by appointment/email
- Instructor office location: Amos Eaton 134
- TA: Iris :
- TA office hours: TBA
- TA office: TBA
- Web site: <https://tw.rpi.edu/classes/data-science-2025>
- LMS (<http://lms.rpi.edu/>)



# Contents

- Learning objectives
- What is expected
- Assignments
- What skills are needed
- Outline of the course
- Definitions / Current Challenges
- 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.
- For both to know tools, and requirements to properly handle data and information.
- Learn and evaluate the full life-cycle of data and relevant methods, technologies and best practices.



# Learning Objectives

- Through class lectures, readings, written and oral presentation assignments and projects, students should:
- Develop and demonstrate skill in data collection and management.
- Develop data models and generate metadata.
- Demonstrate knowledge application of data and metadata standards.
- Demonstrate skill in data science tool use and evaluation of the results.
- Demonstration the application of the data life-cycle principles.
- 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
- 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
  - Additional questions or tasks in assignments
- 4000 Level are welcome to complete these higher requirements to extra grade



# Assessment and Assignments

- Written assignments with specific percentage of grade allocation provided with each assignment.
- Projects with written reports and group oral presentations with 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.
- 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 (1 week).
- Summaries should be short! (max 3 paragraphs)



# 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.



# 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 (**zero grade for that assignment**). 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, 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. **If you have any question concerning this policy before submitting an assignment, please ask for clarification.**



# 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 😊



# Current Syllabus/Schedule

- Week 1 (Sep. 02): History of Data and Information, Data, Information, Knowledge Concepts and State-of-the-Art
- Week 2 (Sept. 09): Data and information acquisition (curation, preservation) and metadata - management
- Week 3 (Sept 16): Data formats, metadata standards, conventions, reading and writing data and information
- Week 4 (Sept. 23): Module 2 & 3 Review, Knowledge Graphs, Data Analysis I
- Week 5 (Sept. 30): Data Analysis II
- 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 I
- Week 9 (Oct. 28): Data Mining II - Project teams and introducing project instructions.
- Week 10 (Nov. 04): Data Quality, Uncertainty and Bias, Final Project Preparation – Start project meetings with team
- Week 11 (Nov. 11): Scientific Workflows and Data Stewardship
- Week 12 (Nov. 18): Webs of Data
- Week 13 (Nov. 25): Guest Lecture
- Week 14 (Dec. 02): No classes: continue project and assignment work
- Week 15 (Dec. 09): Final project presentations, written reports due Dec 12



# Introductions

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



# Questions so far?



# So, what are we talking about?



<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.



- But merely using data isn't really 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.



# DIKW model

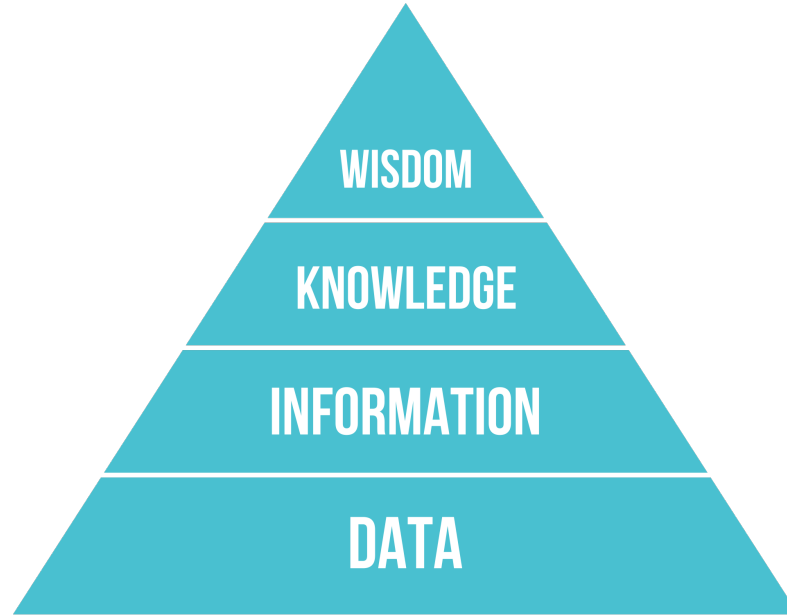


Image Source: [https://en.wikipedia.org/wiki/DIKW\\_pyramid](https://en.wikipedia.org/wiki/DIKW_pyramid)

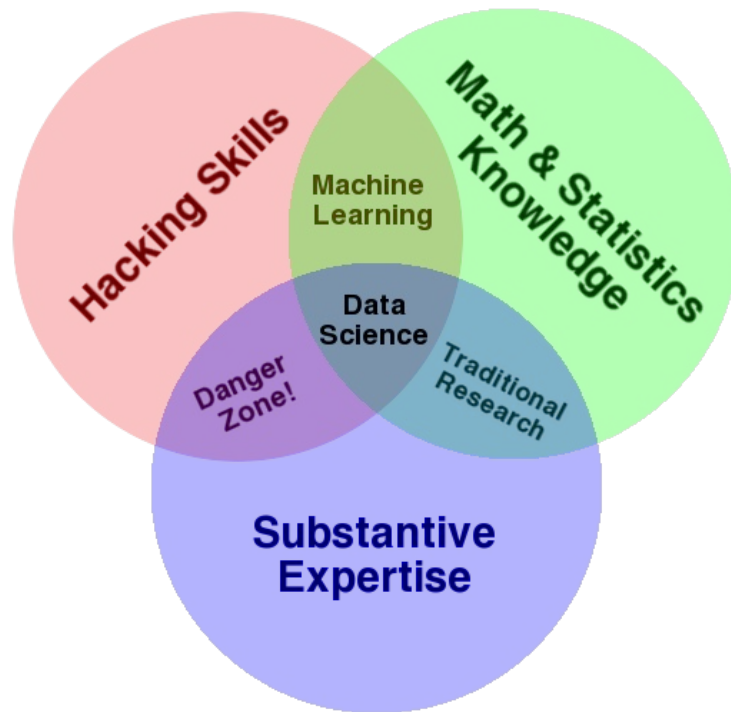
# What is Data Science?

- **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](https://en.wikipedia.org/wiki/Data_science)



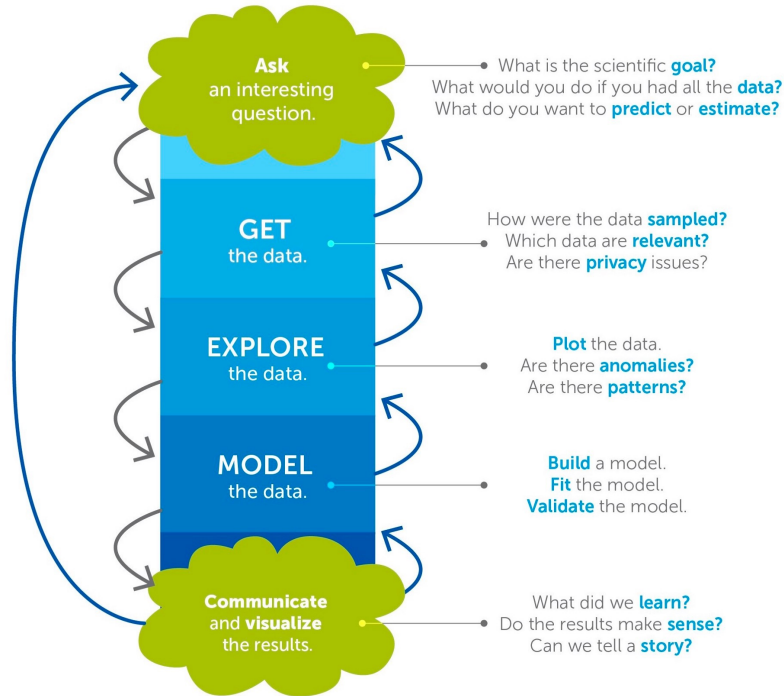
# Data Science



<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



Rensselaer

Tetherless World Constellation



# 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 someone else's data to come up with the Kepler's laws of planetary motion.

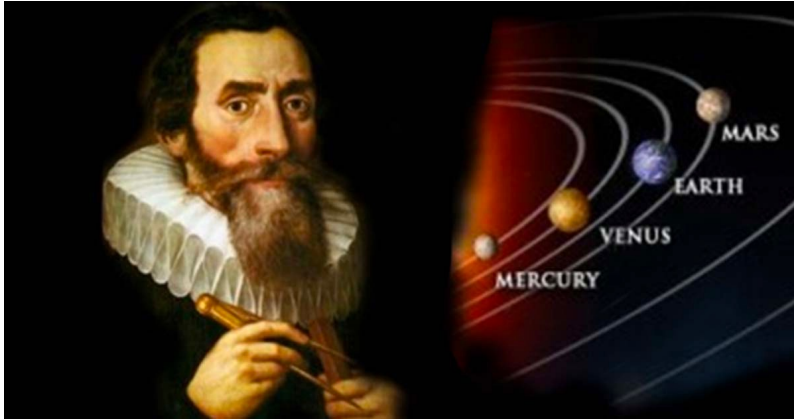


Image Credit: <https://www.ancient-origins.net/history-famous-people/how-did-skeptical-astrology-johannes-kepler-contribute-our-view-cosmos-007944>



# History

- Observation of our natural world – Kepler's laws of planetary motion
  - Aurora and sunspots (Chinese historians documented this phenomena)



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



# How do we 'get' data?

## 1. Computation

e.g. GPA, credit score, age

## 2. Observation

e.g. outcome of an experiment, athletic performance

## 3. Measurement

e.g. thermometer, camera, any sensor



# 42?



# 42?



# 42?

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

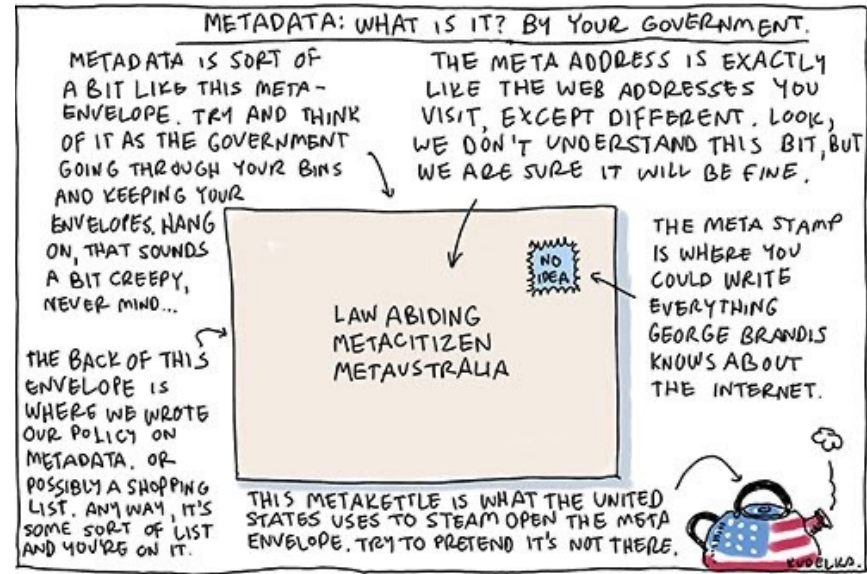
Douglas Adams  
The Hitchhiker's Guide to the Galaxy

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



# Definitions ctd.

- Metadata – data about data



- Data documentation – integrated collection of information and metadata intended to support all aspects of data (find, access, use...)

# Examples

- Rock sample:

- Data: weight, composition, shape, size
- Information: distribution of rock types in area
- Knowledge: evidence of geologic activity
- Metadata: location and time of sample collection
- Documentation: published lab report ...

- Weather

- Data: wind speed and direction, temperature, ..
- Information: weather map with contours and features
- Knowledge: high pressure system, stable weather
- Metadata: type of radar, sensor, model





# Definitions ctd.

- Data Management: Process of arranging for discovery, access and use of data, information and all related elements.
  - 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 elements)

- **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 it is available for discovery and re-use in the future (<https://www.dcc.ac.uk/about/digital-curation> )
  - **Preservation:** Process of retaining usability of data in some source form for intended and unintended use
- 
- **Stewardship:** Process of maintaining integrity for acquisition, curation and/or preservation



# 7 Phases of Data Life Cycle

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

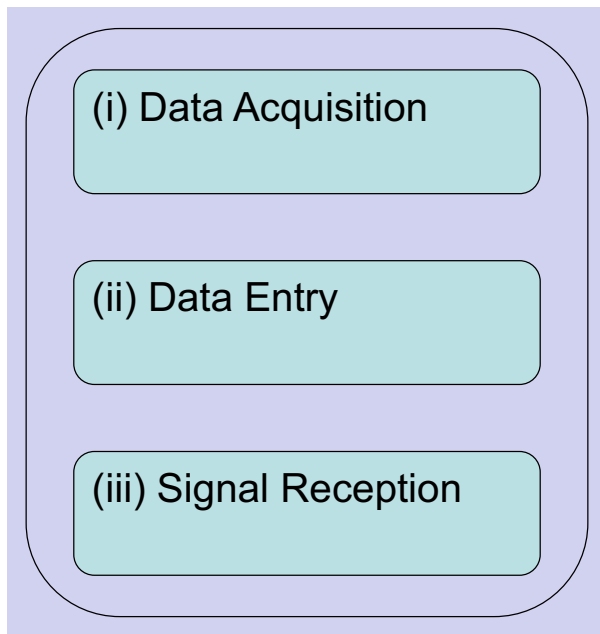
Article on Data Life Cycle:

<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 get 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, typically important in control system, but becoming more important for information systems within the Internet of Things (IoT).



## 2. Data Maintenance

- Once the data has been captured, it usually encounters data maintenance, this can be defined as:
  - Supplying the data to 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 require 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 organizational 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.
  - Is it legal to use the data in the way business people want?





# 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:

- The data archive is simply a place where data is stored, but no maintenance, active usage or publication occurs.

(copying of data to an environment where it is stored, in case it is needed again)



## 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 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.



# 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 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 on the Internet/Web (good or bad?)

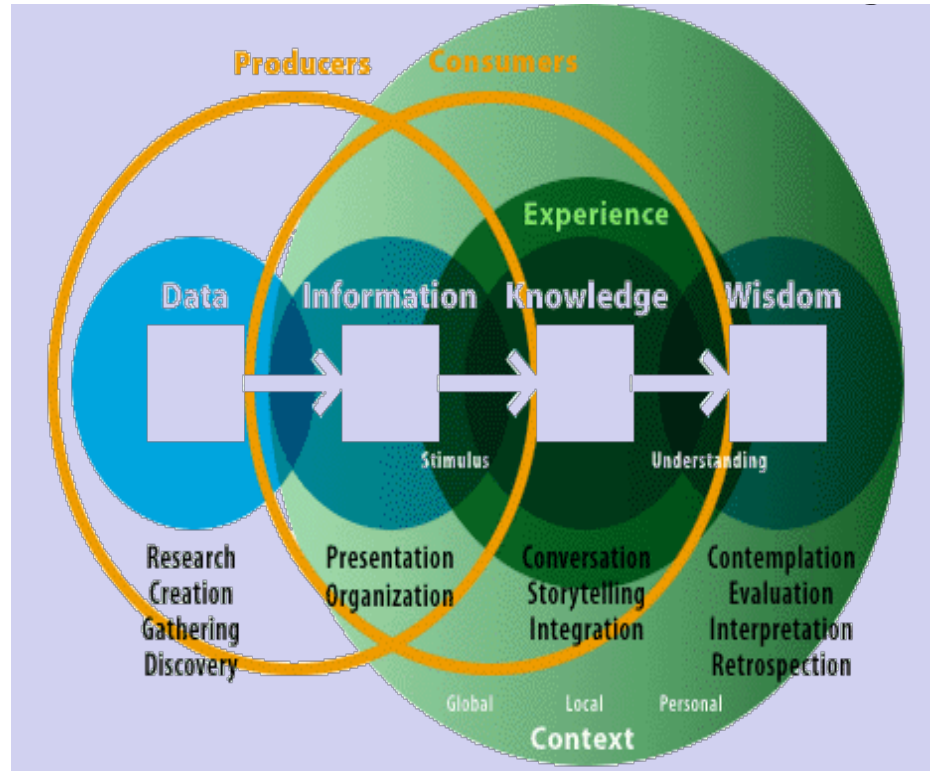


# Provenance

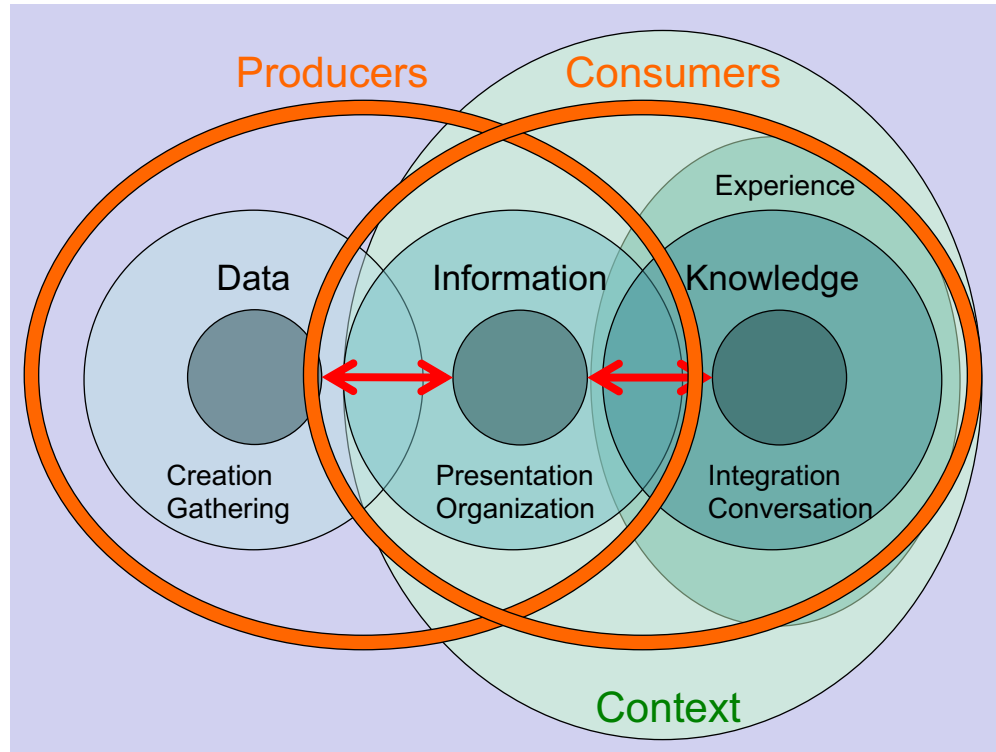
- Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility
- “Metadata in a given context”



# An OLD view of data, information and knowledge



# Data-Information-Knowledge Ecosystem



We will see more on this later...



# Summary

- 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 technologists are not well prepared to cope with 21<sup>st</sup> century data management and use of tools.
- Making data available is now a responsibility not a privilege.



# 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 by email ([eleisa2@rpi.edu](mailto:eleisa2@rpi.edu))



Next class: Sept. 5<sup>th</sup> – Data and information acquisition (curation) and metadata/ provenance – management.



# Thanks!

