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

Ahmed Eleish

Data Science ITWS/CSCI/ERTH-4350/6350

Week 1, August 29th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

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: TBA**
- **TA office:** Lally 205
- Web site: https://tw.rpi.edu/classes/data-science-2024
- LMS (http://lms.rpi.edu/)

Contents

- About this course
- Learning objectives
- What is expected
- 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

Assessment and Assignments

- Written assignments with specific percentage of grade allocation provided with each assignment
- Individual oral presentations with specific percentage of grade allocation provided
- Group presentationsdepending on class size
- Participation in class(not to exceed 10% of total, start with 10% and lose % by not participating)
- Late submission policy: first time with valid reason no penalty, otherwise 20% of score deducted each late day

Assessment and Assignments

- **This is a Reading Heavy Course.**
- Reading assignments
- Are given almost every week
- Most are background and informational
- Some are key to completing assignments
- Some are relevant to the current week's class (i.e. follow up reading)
- Others are relevant to following week's class (i.e. pre- reading)
- Will be discussed in class and participation in these discussions is taken into account

At the beginning of each class, we will discuss the assigned reading assignments. (come prepared!)

- You will progress from individual work to group work

Assignments

- 5 Assignments
- 1 Group Project (Data Science Research Project Research Paper Group Work: Due at the end of the semester).
- You will given guidance in finding data for your group project.
- At the end of the semester, you and your group members will submit your final project paper (report) and present your project results as poster. Paper and Poster templates will be available on LMS.

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 understandand 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, practical sessions, 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

What is expected

- Attend class, complete assignments
- Participate (esp. reading discussion)
- 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

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 (Aug. 29): History of Data and Information, Data, Information, Knowledge Concepts and State-of-the-Art
- Week 2 (Sept. 05): Data and information acquisition (curation, preservation) and metadata - management
- Week 3 (Sept 12): Data formats, metadata standards, conventions, reading and writing data and information
- Week 4 (Sept. 19): Module 2 & 3 Review, Data Analysis I
- Week 5 (Sept. 26): Class exercise - collecting data - individual
- Week 6 (Oct. 03): Class Presentations: present your data, Part of Assignment 2
- Week 7 (Oct. 10): Class Presentations: present your data, Part of Assignment 2
- Week 8 (Oct. 17): Academic basis for Data Science, Data Models, Schema, Markup Languages, group project, working with someone else's data
- Week 9 (Oct. 24): Introduction to Data Mining for Data Science
- Week 10 (Oct. 31): Data Analysis II and Class exercise
- Week 11 (Nov. 07): Data Workflow Management, Preservation, and Data Stewardship
- Week 12 (Nov. 14): Data Quality, Uncertainty and Bias, Final Project Preparation – Project work discussion with the instructor
- Week 13 (Nov. 21): Webs of Data and Data on the Web, the Deep Web, Data Discovery, Data Integration, Data Citation
- Week 14 (Nov. 28): No classes: Thanksgiving break – continue project and assignment work
- Week 15 (Dec. 05): Final project work discussion with the instructor – Group One-on-Ones
- Week 16 (Dec. TBA): Final Project Report Submission and Presentations

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 and can be the basis of graphs, images of a set of variables.
- **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

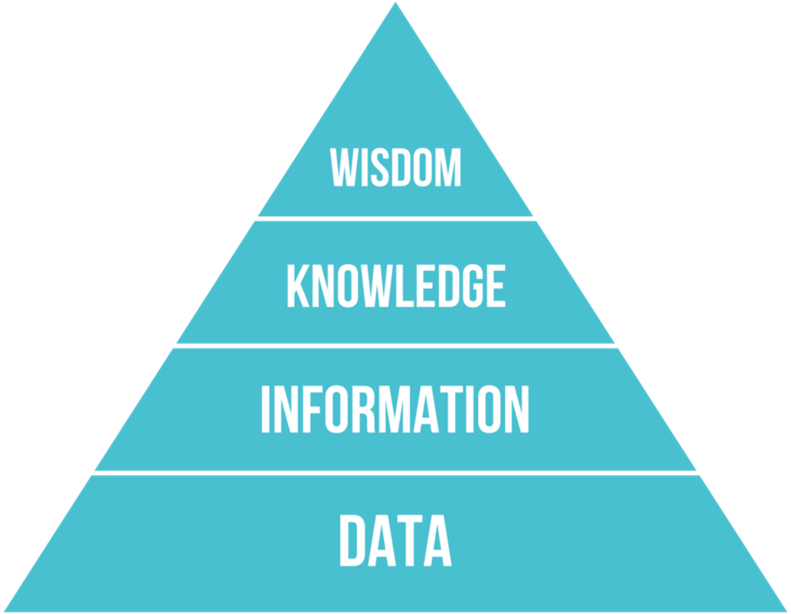


Image Source: https://en.wikipedia.org/wiki/DIKW_pyramid

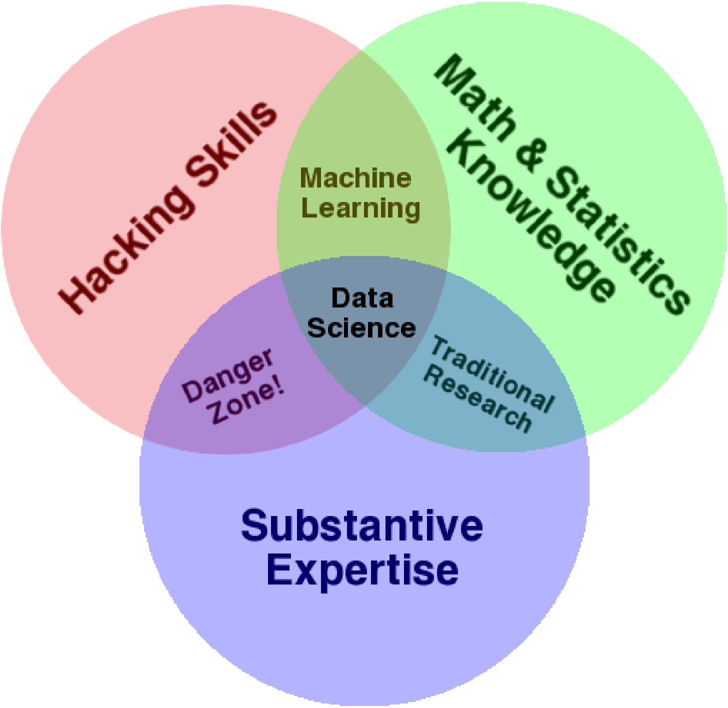
What is Data Science?

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data
- To gain insights into data through computation, statistics, and visualization

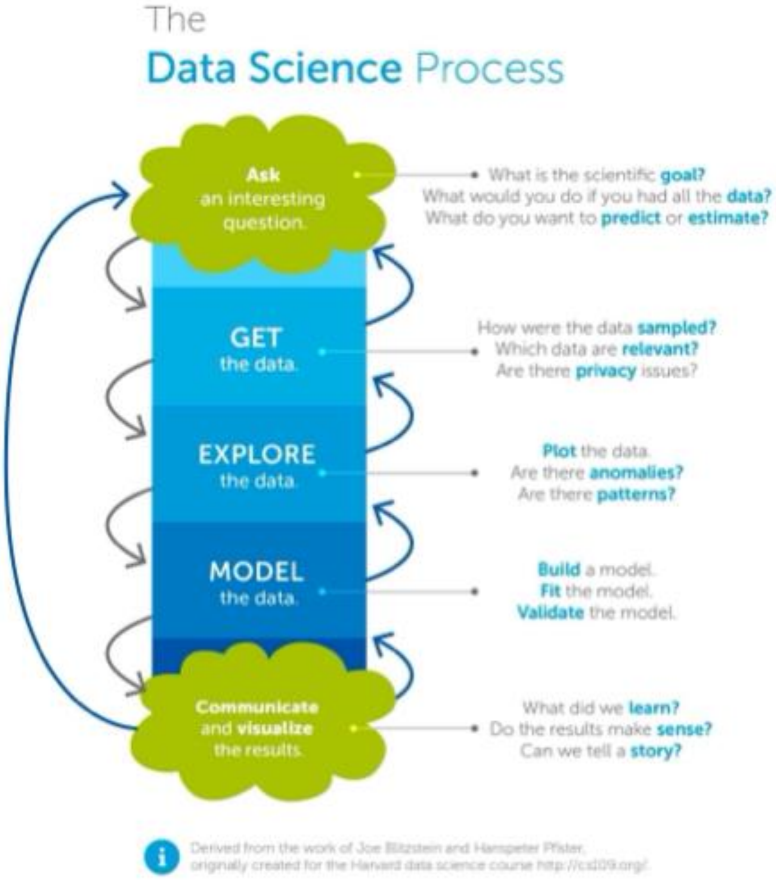
Go over: Week 1 Reading Materials: “What is Data Science” in the course webpage: https://tw.rpi.edu/classes/data-science-2024

Source: 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

Ask an interesting question.
- What is the scientific **goal?**
- What would you do if you had all the **data?**
- What do you want to **predict** or **estimate?**

GET the data.
- How were the data **sampled?**
- Which data are **relevant?**
- Are there **privacy** issues?

EXPLORE the data.
- **Plot** the data.
- Are there **anomalies?**
- Are there **patterns?**

MODEL the data.
- **Build** a model.
- **Fit** the model.
- **Validate** the model.

Communicate and **visualize** the results.
- What did we **learn?**
- Do the results make **sense?**
- Can we tell a **story?**

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.”
~ Josh Blumenstock

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

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?

Through observations and measurements Analog versus digital ‘data’

- Examples:

– Thermometer, watch, sensors ...

– We keep a digital exhaust of data from our day-to-day activities.

42?

42?

Colorbox

42?

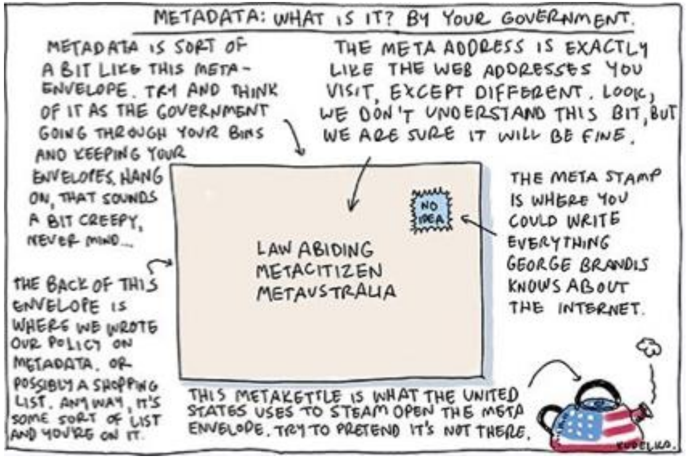
“The answer to the great question…of Life, the Universe and Everything…is…forty-two.”

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: WHAT IS IT? BY YOUR GOVERNMENT.
METADATA IS SORT OF A BIT LIKE THIS META-ENVELOPE. TRY AND THINK OF IT AS THE GOVERNMENT GOING THROUGH YOUR BINS AND KEEPING YOUR ENVELOPES. HANG ON, THAT SOUNDS A BIT CREEPY, NEVER MIND...
THE META ADDRESS IS EXACTLY LIKE THE WEB ADDRESSES YOU VISIT, EXCEPT DIFFERENT. LOOK, WE DON'T UNDERSTAND THIS BIT, BUT WE ARE SURE IT WILL BE FINE.
NO IDEA
THE META STAMP IS WHERE YOU COULD WRITE EVERYTHING GEORGE BRANDIS KNOWS ABOUT THE INTERNET.
LAW ABIDING
METACITIZEN
METAUSTRALIA
THE BACK OF THIS ENVELOPE IS WHERE WE WROTE OUR POLICY ON METADATA. OR POSSIBLY A SHOPPING LIST. ANYWAY, IT'S SOME SORT OF LIST AND YOU'RE ON IT.
THIS METAKETTLE IS WHAT THE UNITED STATES USES TO STEAM OPEN THE META ENVELOPE. TRY TO PRETEND IT'S NOT THERE.
KUDELKA.

– Metainformation – information about information

- 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 – images of the rock as collected
  - Knowledge – evidence of geologic activity
  - Metadata – location and time of 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, use of 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

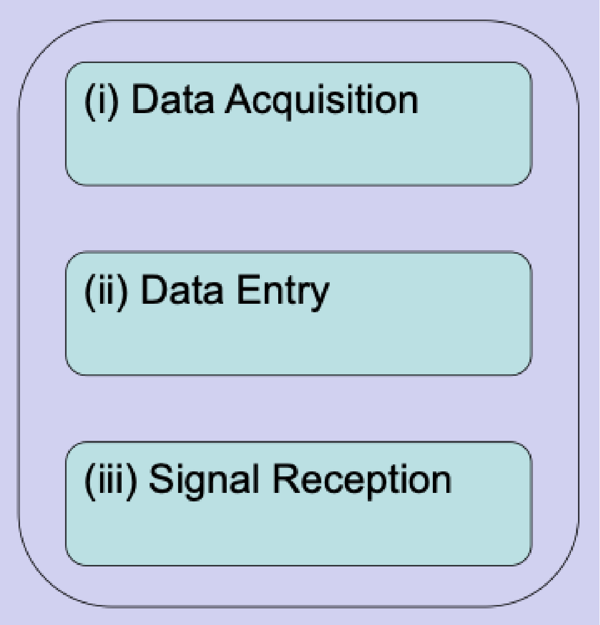
1. Data Capture
2. Data Maintenance
3. Data Synthesis
4. Data Usage
5. Data Publication
6. Data Archival
7. Data Purging

Read the 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:
  - The supplying of the data to point at which Data Synthesis and Data 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 - Enrichment

3. Data Synthesis

- This is relatively new and perhaps still not very common phase in Data Life Cycle.
- The creation of data values via inductive logic, using other data as input.
- Inductive logic require some kind of expert knowledge, judgement and/or as a part of the logic.
  - Example: The way credit score is created.
- It is the arena of analytics that uses modeling

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 archival is simply a place where data is store, 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.

- Data governance challenge is this phase is 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 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**

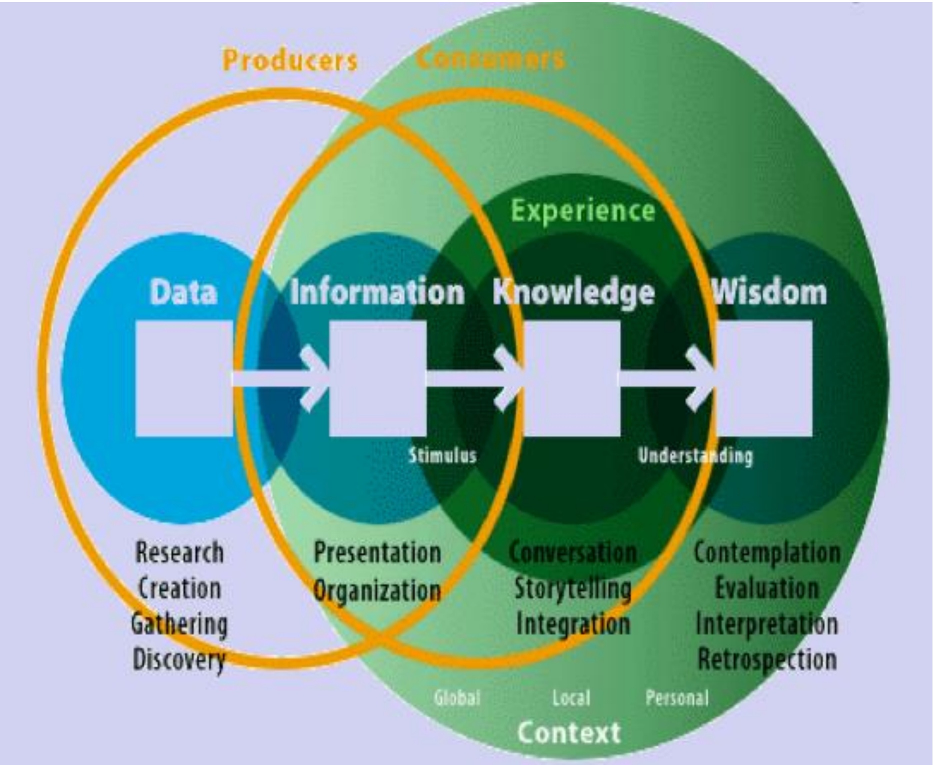
- And now much of the data is on the Internet/Web (good or bad?)

Provenance

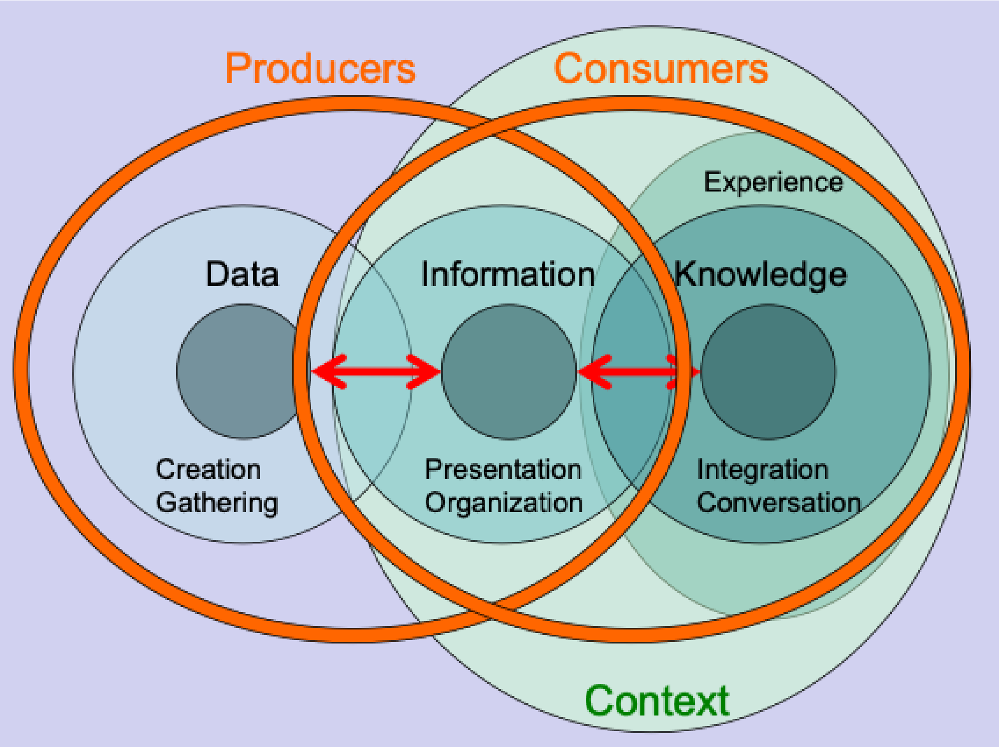
- Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility

- “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 21st century data management and use of tools

- Making data available is now a responsibility not a privilege

Rise of the Data Scientist

- http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/
- https://www.bloomberg.com/professional/blog/7-phases-of-a- data-life-cycle/
- http://drewconway.com/zia/2013/3/27/where-to-draw-the-line-on-data-science
- http://radar.oreilly.com/2010/06/what-is-data-science.html
- http://www.wired.com/magazine/2010/06/ff_sergeys_search/all/1
- https://blogs.msdn.microsoft.com/escience/2009/10/16/the- fourth-paradigm-data-intensive-scientific-discovery-book- released/
- https://hbr.org/2022/02/the-new-rules-of-data-privacy

- And other reading material listed on course webpage for this week (week 1 reading)...

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

Thanks!