1 - Data Management basics

Data Management

Michele Mastroianni
Michele.mastroianni@unicampania.it
mmastroianni@unisa.it

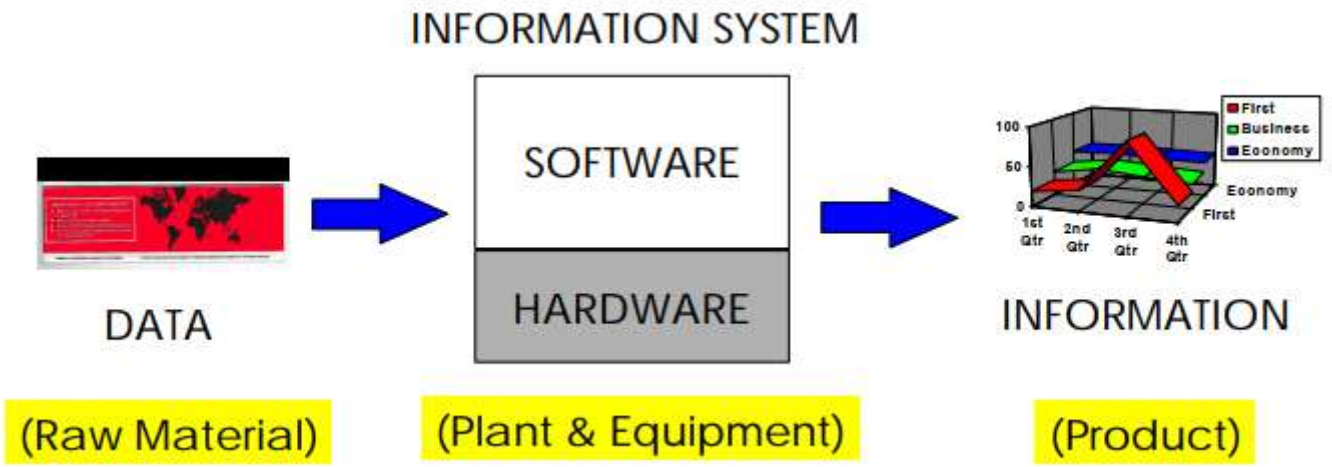
Università degli Studi della Campania *Luigi Vanvitelli*

Dipartimento di Matematica e Fisica

Data and Information

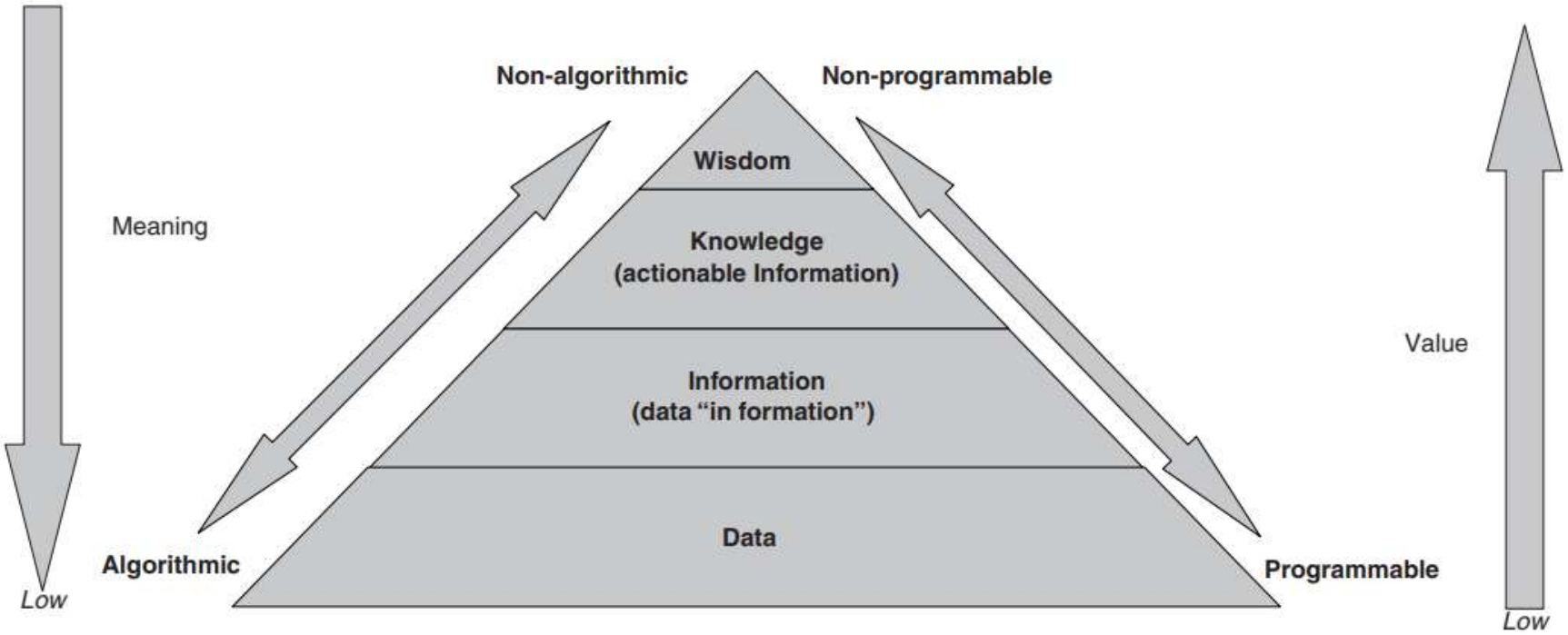
Manufacturing analogy:

1. Data is the raw material
2. Software and hardware are the plant and equipment
3. Information is the end product that is delivered to the customer



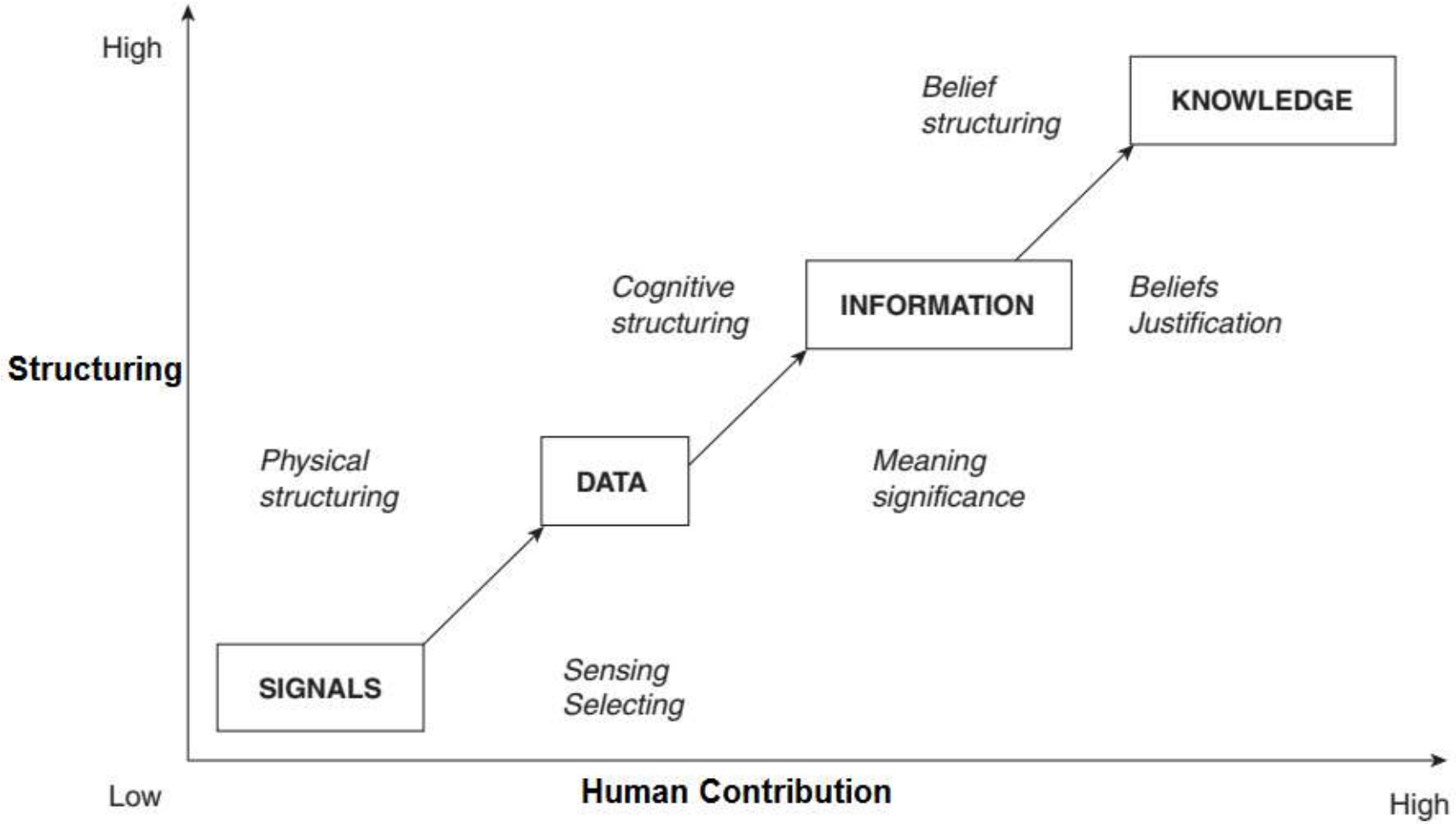
The DIKW Hyerarchy

The data–information–knowledge–wisdom (DIKW) hierarchy is a model for representing structural and/or functional relationships between Data, Information, Knowledge, and Wisdom



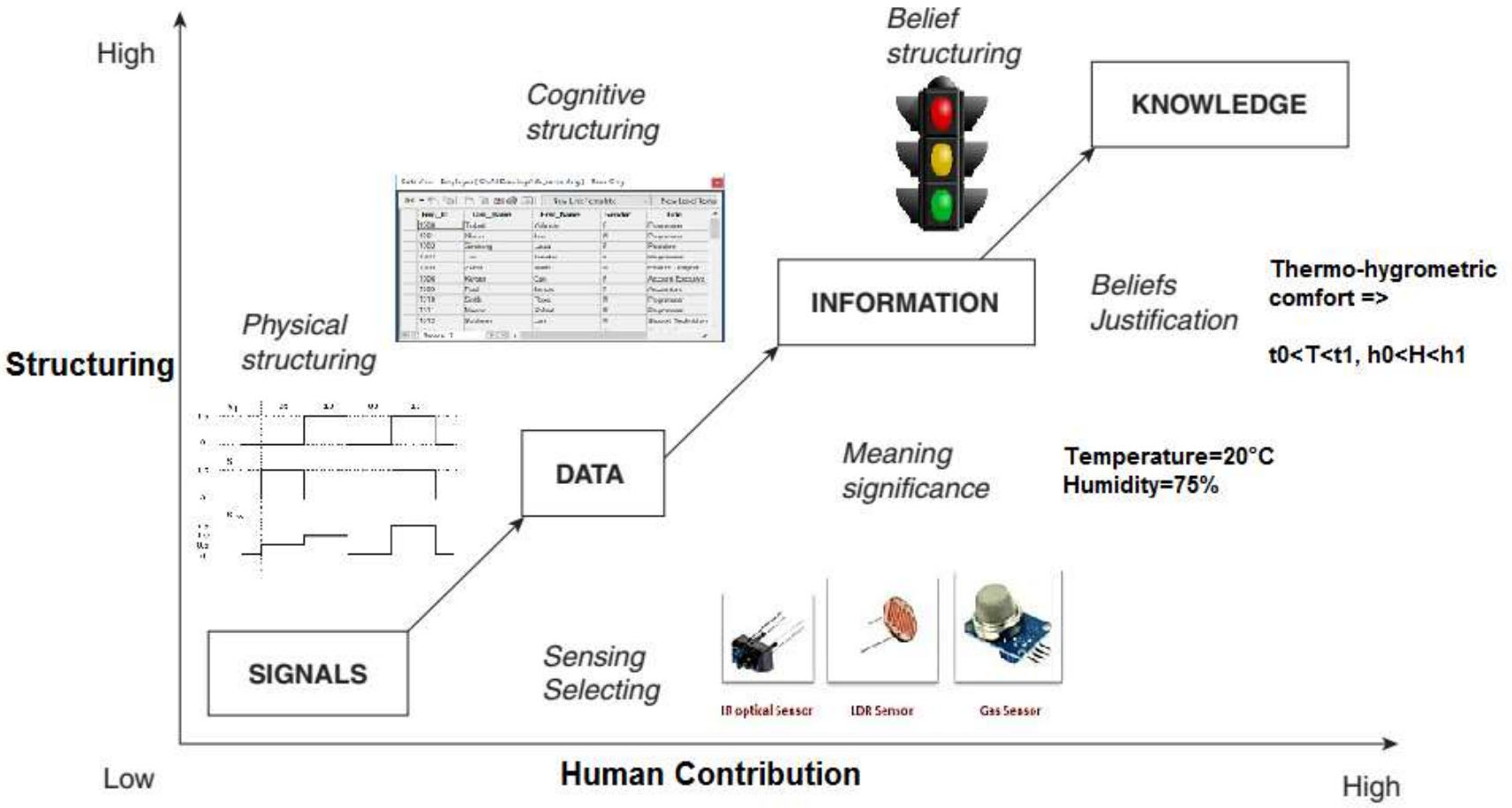
Università
degli Studi
della Campania
Luigi Vanvitelli

The transformation processes between signals, data, information and knowledge



Università
degli Studi
della Campania
Luigi Vanvitelli

Example: Environmental conditioning



Università
degli Studi
della Campania
Luigi Vanvitelli

Information needs

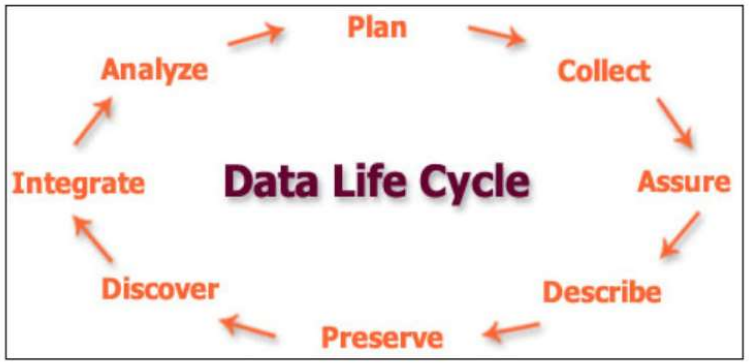
1. What specific issue is to be addressed?
2. Who are the main decision-making and influencers?
3. How do these groups currently use informations?
4. What constraints do they work under?
5. Do a policy exists within decisions are being made?
6. What specific information is required to help implement/evaluate policy?
7. How, when and to whom should this information be delivered?

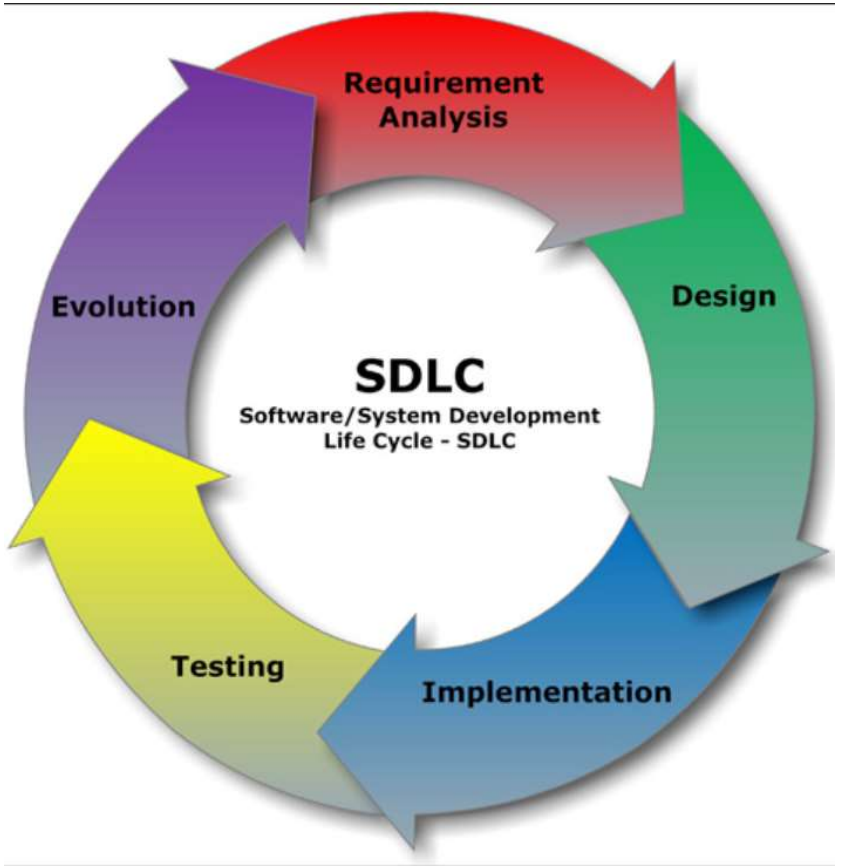
The Data Life Cycle (DLM)

Università
degli Studi
della Campania
Luigi Vanvitelli

This is only an example of a possible DLC

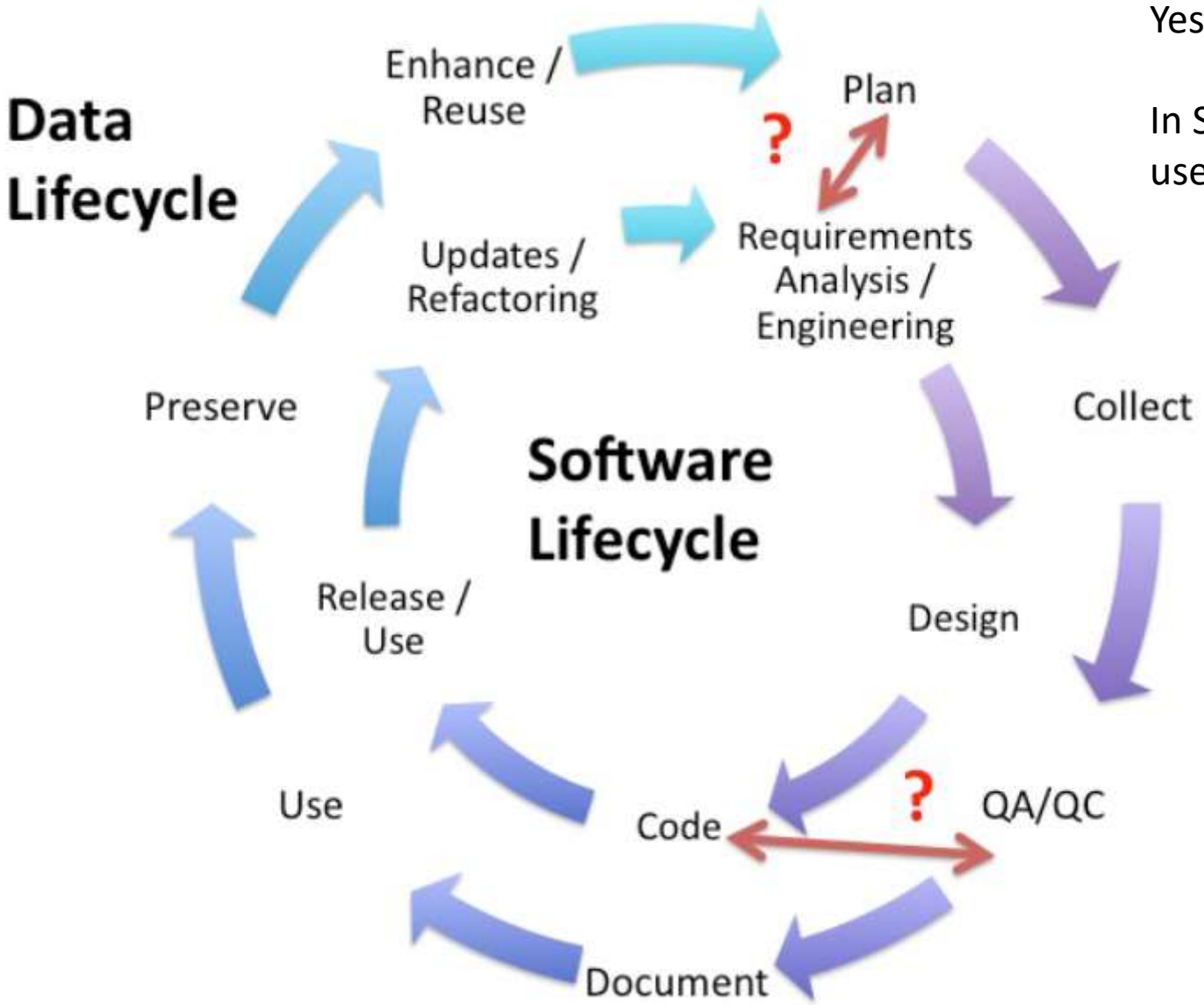
Data Lifecycle and Software Lifecycle Management





It is possible to create a relationship Between DLC and software Developement lifecycle?

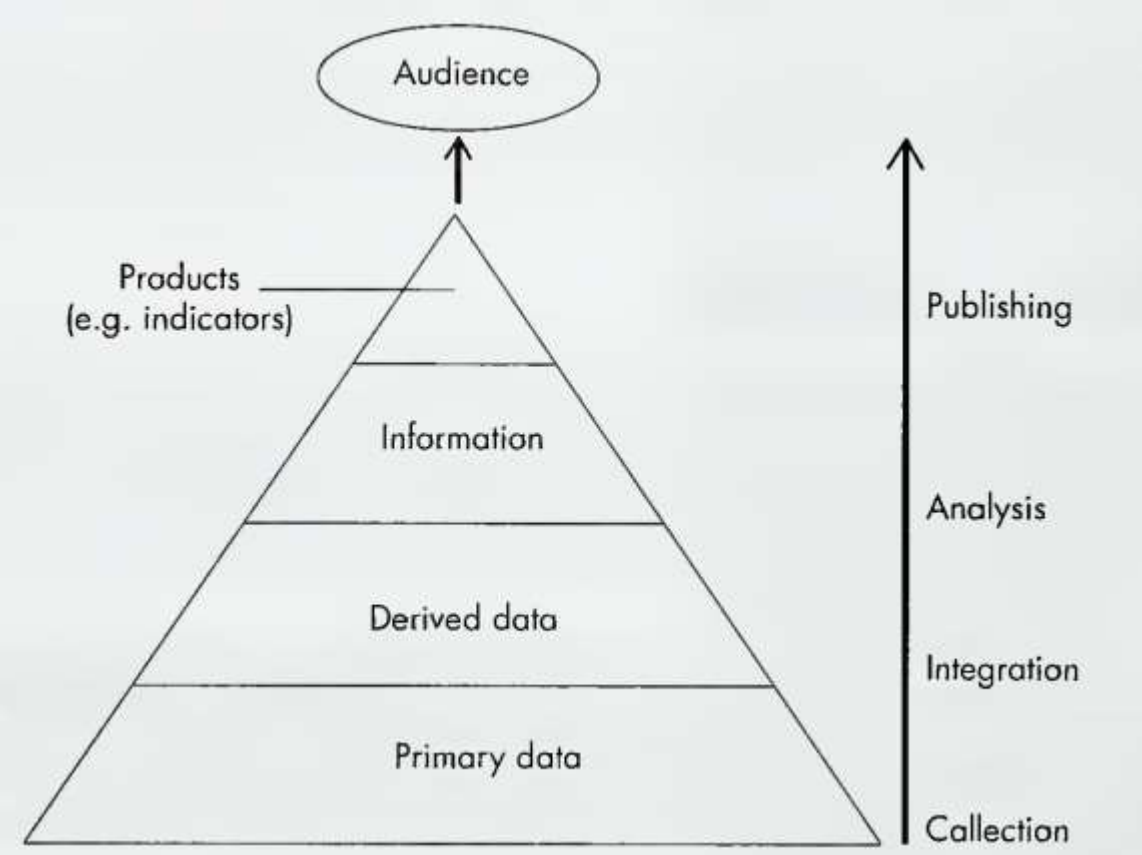
Data Lifecycle and Software Lifecycle Management



Yes, but carefully

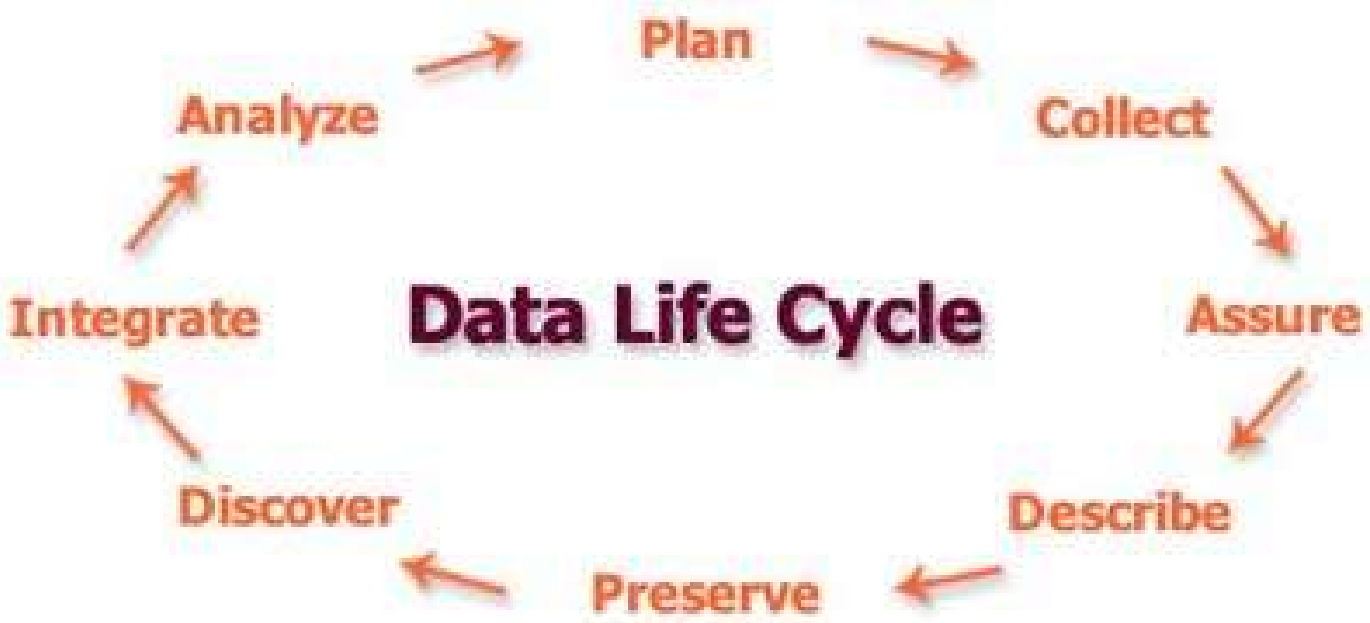
In Scientific SW may be useful

A first example of DLC: Information Pyramid



- **Collection:** consist in receiving the raw data of various natures.
- **Integration:** is to set rules and policies to integrate the distributed data because the methods of collection are different.

DataOne Data Lifecycle

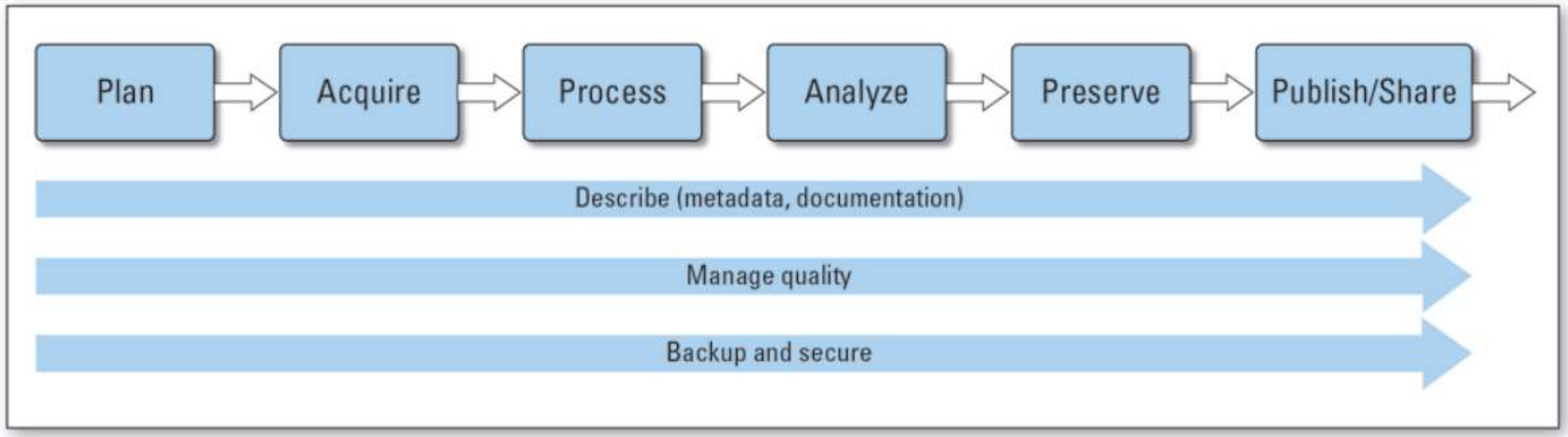


DataOne DLC

- **Plan**: In this phase, a description of the data and how these data will be managed and made accessible throughout their lifecycle.
- **Collect**: observations are made by hand or with sensors or other tools and data is placed under digital format
- **Assure**: controls and inspections are carried out to ensure the quality of the data
- **Describe**: metadata standards are used accurately and correctly to describe the data.
- **Preserve**: Data is stored in a suitable long-term archive.
- **Integrate**: data from different sources are combined to have a homogeneous set of data.
- **Analyze**: data is exploited and analyzed to draw conclusions and interpretations of decision support

The USGS DLC

Defined by the U.S. Geological Service (USGS)



The phases of USGS DLC

- **Plan**: resources and planned outputs for each stage of the cycle. Output: data management plan
- **Acquire**: represents the activities by which new or existing data is collected, generated, or considered and evaluated for reuse. It involves the collection or addition of data banks.
- **Process**: refers to actions or measurements taken on the data to verify, organize, transform, integrate and extract data in an appropriate output form for future use.
- **Analyze**: actions and methods performed on data that help to describe the facts, detect trends, develop explanations and test assumptions. This includes assurance of quality data, statistical analysis of data, modeling and interpretation of test results.

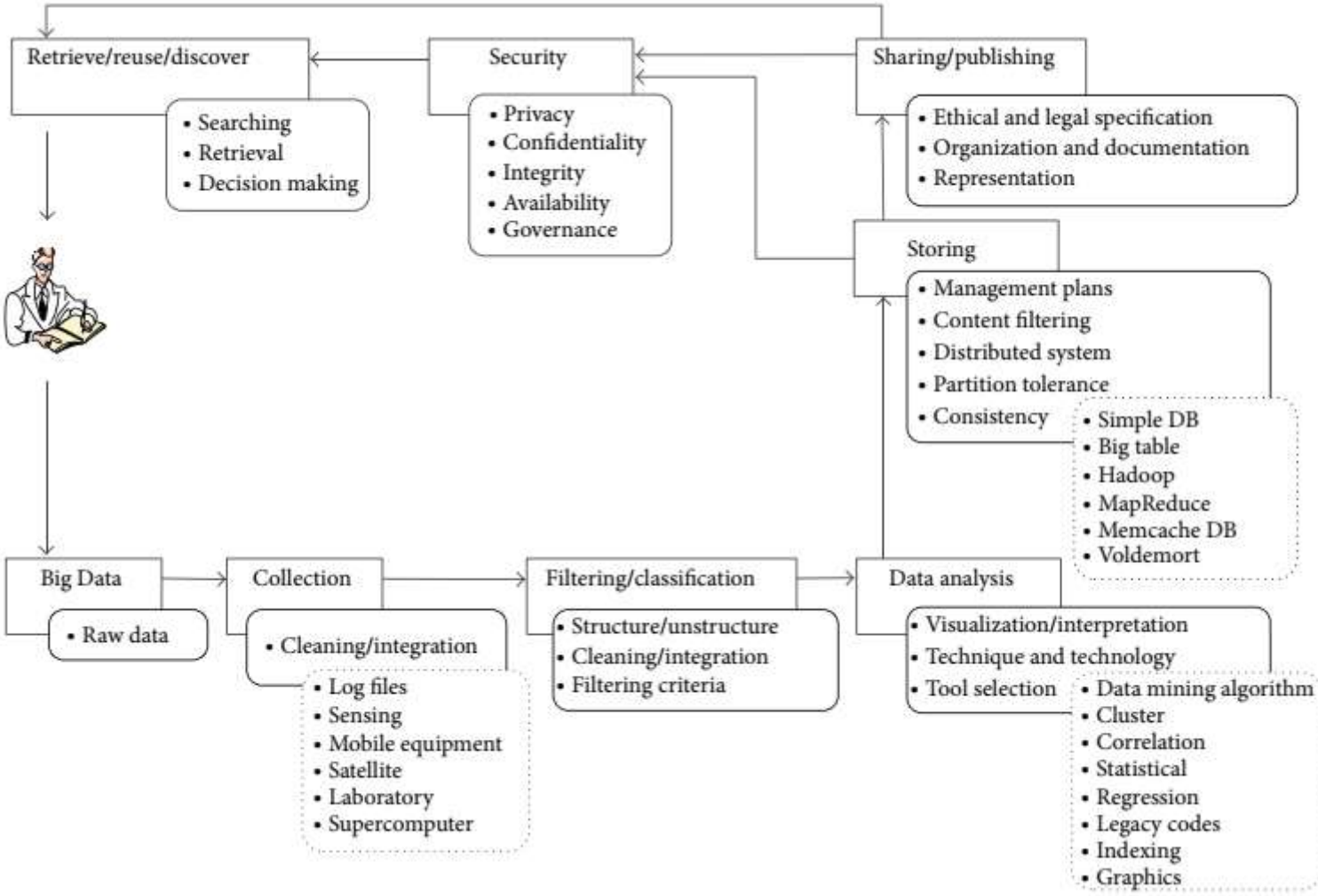
The phases of USGS DLC (cont'd)

- **Preserve**: involves actions and procedures for storing data for a certain period of time and/or setting up side data for future use, and includes archiving data and/or submitting data to a reference data.

- **Publish/Share**: prepare and publish, or disseminate, data with good quality to the public and other organizations. We need to make sure that the data is
- shared, but with controls to protect the data ownership and pre-decision and data integrity.

- **Describe** (metadata, documentation): throughout the cycle, documents must be updated to reflect measurements taken on the data.

The Khan et al (2014) “Hindawi” model



Hindawi DLC phases

- **Big Data**: allows to enhance the raw data collected by researchers and organizations. The data are transformed from its initial state and are stored in a state of added value.

- **Collection**: large amounts of data are created and others come from sensors, mobile equipment, satellites, laboratories, supercomputers, research forms, messages on internet forums and microblog messages.

- **Filtering/classification**: the data collected are of low density and high value. This phase allows the classification of the data in structured/ unstructured, as well as a filtering according to certain criteria.

- **Data analysis**: allows an organization to process abundant information that can affect the company and accurately predict future observations. Storing: Storing, managing and determining large amounts of data.

Hindawi DLC phases (cont'd)

Sharing/publishing: Enables the public, governments, academics, researchers, scientific partners, federal agencies and other stakeholders to benefit from the information being processed.

Security: describes data security and roles in data management to protect legitimate privacy, confidentiality and intellectual property.

Retrieve/reuse/discover: Data recovery ensures data quality, adding value and retaining data by reusing existing data to discover new and valuable information.

Redacting a DLC

There are many good practices in DLC definition

Researchers may develop a specific DLC for their purposes, even using part of existing DLCs

I suggest to read El Arass et al. (2017) [7]

They introduce a “reference DLC” to compare several DLC models

The “reference” phases are: Planning, Creation/Reception, Integration, Filtering, Anonymity, Enrichment, Analysis, Visualization, Storage, Destruction and Archiving

Lifecycle score according to the retained phases

	Information lifecycle	Hindawi	DataONE	USGS	Big Data	IBM	DDI	CIGREF	CRUD	Enterprise	Pyramid	PII
Plan	0	0	1	1	0	0	1	0	0	0	0	0
Create/Receive	1	1	1	1	1	1	1	1	1	1	1	1
Integration	1	1	1	1	0	0	0	1	0	0	1	0
Filtering	0	1	0	1	1	0	0	0	0	0	0	0
Anonymity	0	1	0	0	0	1	0	0	0	0	0	1
Enrichment	0	0	0	0	1	0	0	0	0	0	0	0
Analyze	0	1	1	1	1	1	1	1	1	0	1	0
Visualization	0	1	0	0	1	0	0	0	0	0	0	0
Storage	1	1	0	1	1	1	0	1	1	1	0	0
Destruction	1	0	0	0	0	0	0	0	1	1	0	1
Archiving	1	1	1	1	0	1	1	0	1	1	0	0
Total	5	8	5	7	6	5	4	4	5	4	3	3

References

1. World Conservation Monitoring Centre. 1996. "**Guide to Information Management in the Context of the Convention on Biological Diversity**". United Nations Environment Programme, ISBN: 92-807-1591-5, Nairobi, Kenya. http://www.mekonginfo.org/assets/midocs/0003032-utilities-communications-guide-to-information-management-in-the-context-of-the-convention-on-biological-diversity.pdf
2. Rowley, J. (2007). **The wisdom hierarchy: representations of the DIKW hierarchy**. Journal of information science, 33(2), 163-180. https://journals.sagepub.com/doi/abs/10.1177/0165551506070706?casa_token=kZHC0hnp354AAAAA:Om5KYFRjQ7Yl0BHaOWYu_Iazeb24ezb631_kja5Rc0C-P7-_HwH0tE2jA1Bb_vQ2KBw72GDJf3fl
3. Lenhardt, W C et al 2014 "**Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science**". Journal of Open Research Software, 2(1): e15, pp. 1-4, DOI: http://dx.doi.org/10.5334/jors.ax
4. Allard, S. (2012). DataONE: Facilitating eScience through collaboration. Journal of eScience Librarianship, 1(1), 3. https://pdfs.semanticscholar.org/91e6/472248cef044b7720b21353451fb7779895f.pdf
5. FAUNDEEN, John L., et al. **The United States geological survey science data lifecycle model**. US Department of the Interior, US Geological Survey, 2013. https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf
6. KHAN, Nawsher, et al. **Big data: survey, technologies, opportunities, and challenges**. The scientific world journal, 2014, 2014. https://downloads.hindawi.com/journals/tswj/2014/712826.pdf
7. El Arass, M., Tikito, I., & Souissi, N. (2017, April). "**Data lifecycles analysis: towards intelligent cycle**". In 2017 Intelligent Systems and Computer Vision (ISCV) (pp. 1-8). IEEE. https://ieeexplore.ieee.org/document/8054938

The transformation processes between signals, data, information and knowledge

Università
degli Studi
della Campania
Luigi Vanvitelli