

# Data science for decision making

Prof. Amos DAVID

*DICEN-IDF Laboratory, Paris, France*

*Lorraine University, Nancy, France*

# Outline

- The Big Data Pipeline
- Competitive intelligence process
- Data-driven decision pipeline
- Issues, Approach of solution, Illustration

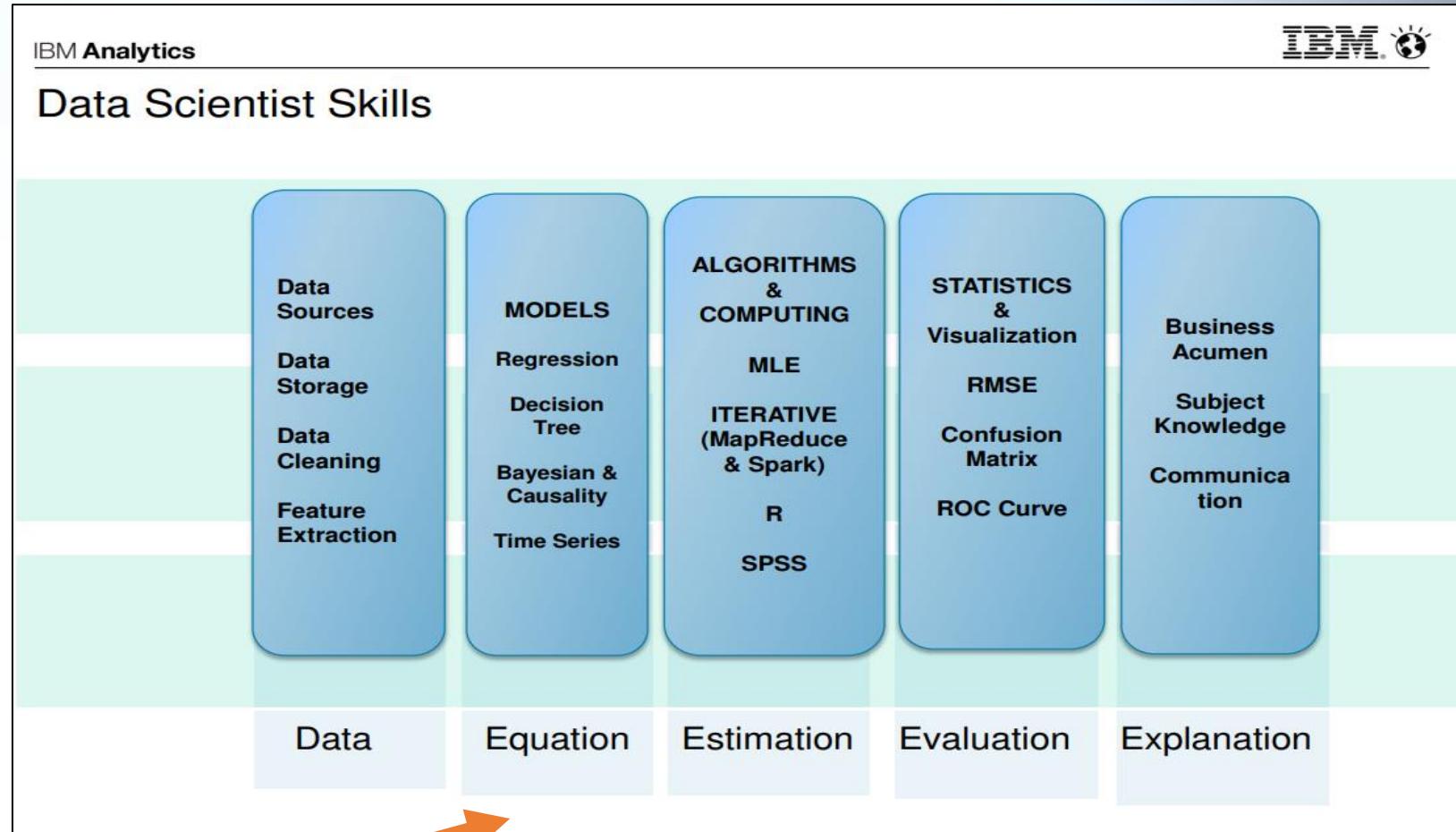
# Aim of the paper

- Show the importance of **decisional problem specification**, to enhance the results obtained using data science as decision support system

# Important observation

Currently, the emphasis in data science is not on the integration of the aspect of ***decisional problem*** but rather on the **extraction of insights** from **existing dataset**

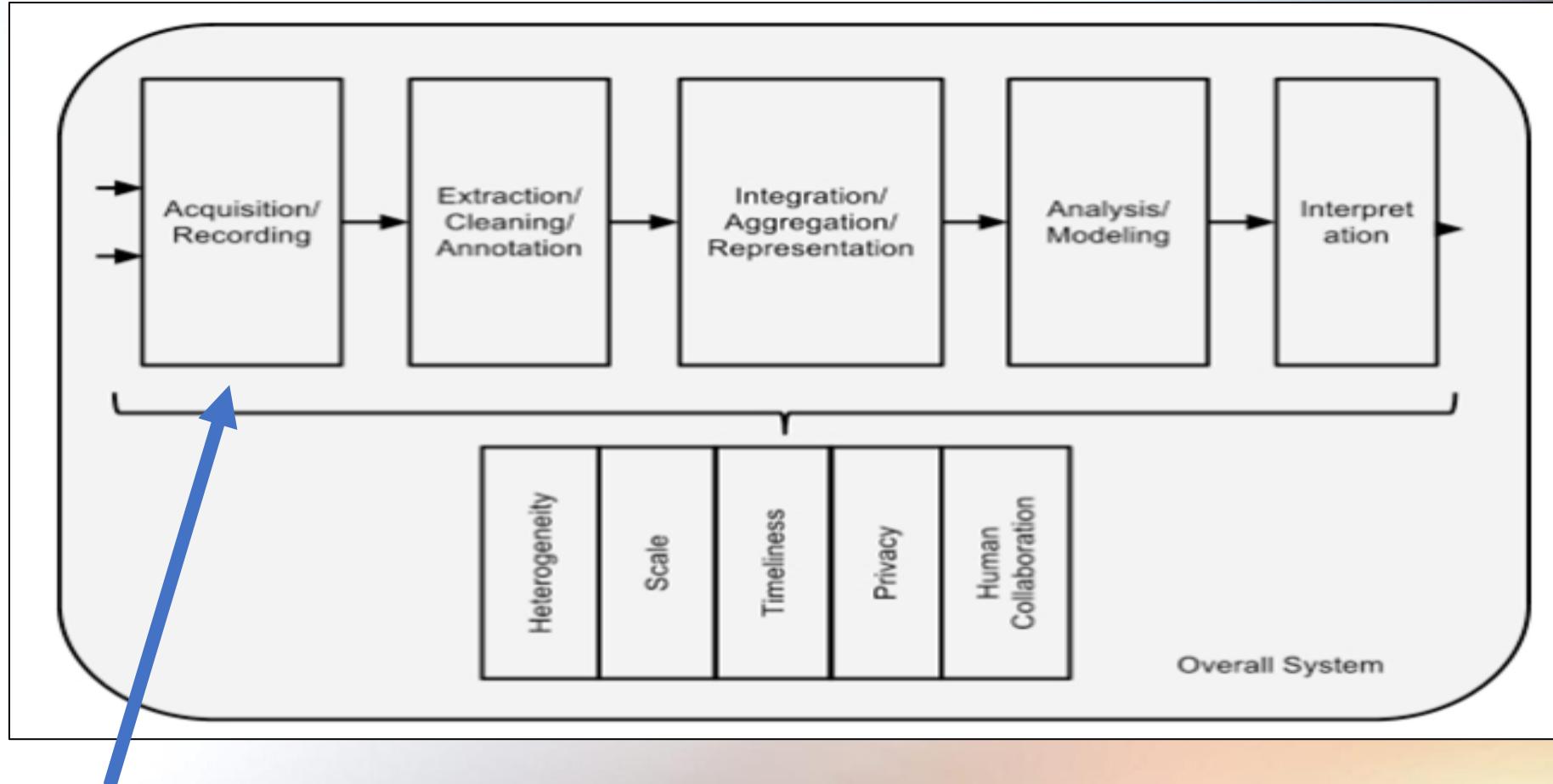
# Data Scientist Skills



No explicit reference to decisional problem

# The Big Data pipeline

*The Big Data pipeline [3]*



No explicit reference to decisional problem

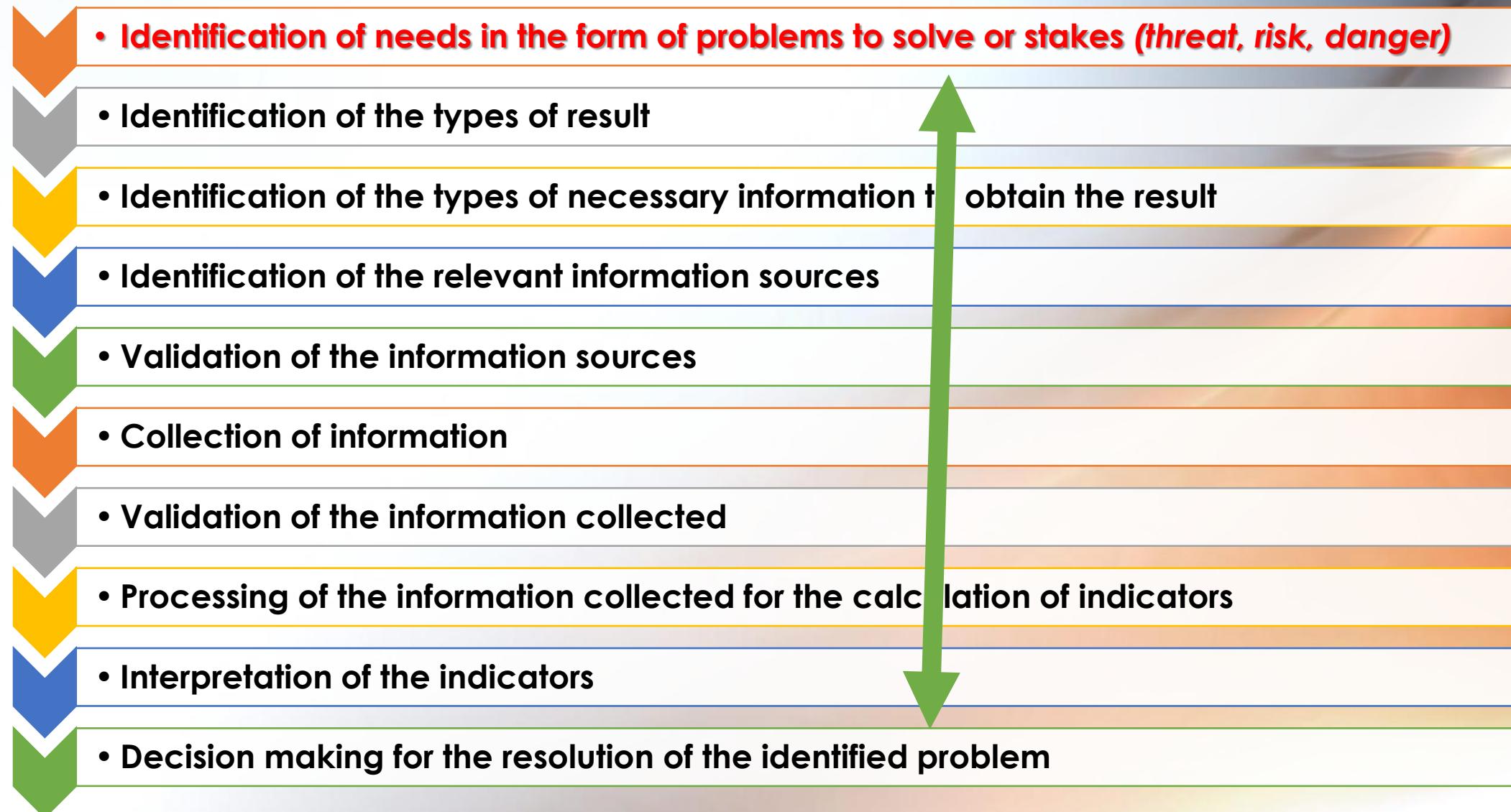
# The stage of Acquisition/Recording

- “The problems start right away during data acquisition, when the data tsunami requires us to **make decisions**, currently in an ad hoc manner, about **what data to keep and what to discard**, and how to store what we keep reliably with the right metadata”

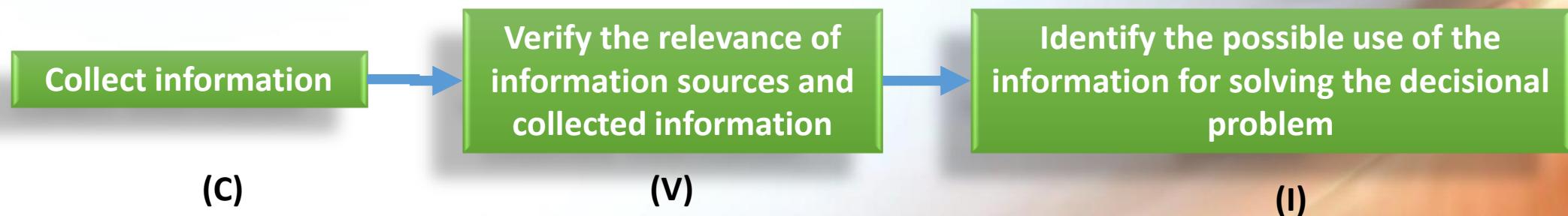
# Economic Intelligence process

- Identification of needs in the form of problems to solve or stakes (*threat, risk, danger*)
- Identification of the types of result
- Identification of the types of necessary information to obtain the result
- Identification of the relevant information sources
- Validation of the information sources
- Collection of information
- Validation of the information collected
- Processing of the information collected for the calculation of **indicators**
- Interpretation of the indicators
- Decision making for the resolution of the identified problem

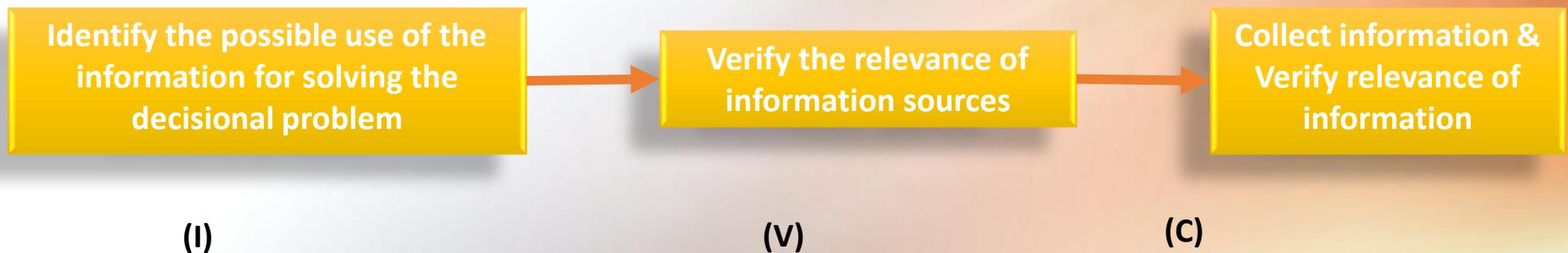
# Economic Intelligence process



### *Current framework*



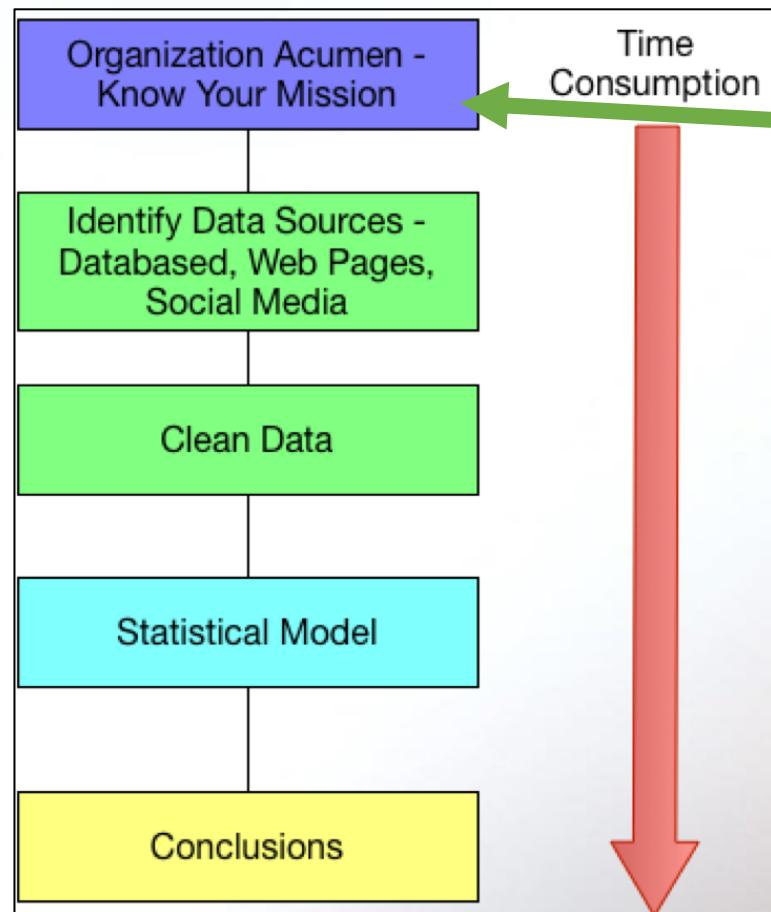
### *Proposed framework*



# Illustration

- The International Forum on the Information Society proclaims that **scientists and research from developing countries are not visible**. The forum considers this situation unacceptable and wants to improve it. This is a decisional problem identified by the World Summit on the information Society

# A timeline for making data-driven decisions



**Acumen:**  
Ability to make good judgements and take quick decisions

# Data science approach

- Get the “correct” dataset
- Use data science to, *extract patterns and facts from that dataset, and utilize those facts to make inferences*

# **Correct data set**

- Bibliographic references from bibliographic databases

# Issues

- In the current approach in data science, this is the most normal solution – relevant information sources, volume big enough for inference making, etc.
- The problem here is to know if this decisional problem is at the origin of the creation of the bibliographic references database. This brings us to the question of WHY we create bibliographic reference databases.

# Approach of solution

- How to represent a decisional problem
- Approaches in the field of competitive intelligence, such as the work of Najoua Bouaka [5] and Philippe Kislin [6]
  - Specification of decisional problem and transforming decisional problem into information retrieval problem

# The concept of STAKE

- {Object, Signal, Hypotheses}
  - **The object** (information needed to solve the problem)
  - **The signal** (the level of knowledge of the decision maker on the needed information)
  - **The hypotheses** (what the decision-maker thinks he stands to gain or lose

# From stake to Information retrieval problem

- The concept of indicators
  - Attributes / Features in data science
- We know that the country indicated in a paper is generally the country of the institution in which paper was written
- There is no cross analysis that can effectively produce the correlation between the country of origin of the authors and the country indicated as the country of publication because the attribute “country of origin” is not used for representing a bibliographic reference.
- In fact, some authors from developing countries may prefer indicating a foreign country because of the view that a publication from a foreign country is better rated.

# Conclusion

- It is obvious that if the decisional-problem is not well stated before data collection, there is high chance of not having the right data set no matter how well the analysis technique. This is particularly the case for attributes that are not represented in the dataset.



Thanks for your attention