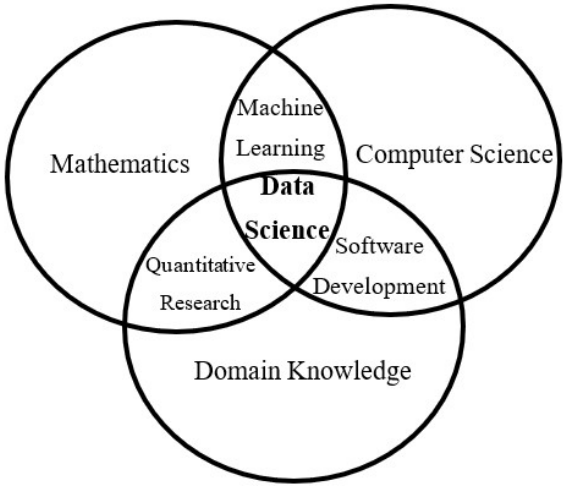
POST-DS: A METHODOLOGY TO BOOST DATA SCIENCE IN THE CONTEXT OF COVID 19

Carlos J. Costa, ISEG

João Tiago Aparicio, IST

Context

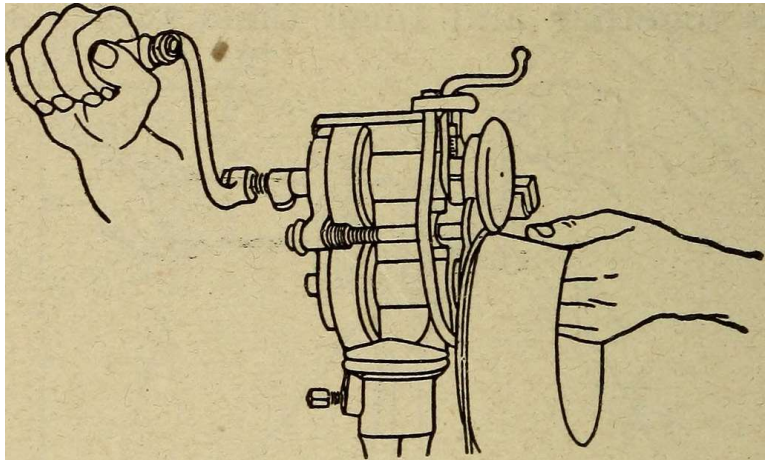
- Data Science includes techniques developed in some traditional fields like artificial intelligence, statistics or machine learning.



Aparicio et al. (2019).

Context

- It is essential creating a methodology that may contribute to the improvement of the knowledge creation outputs.

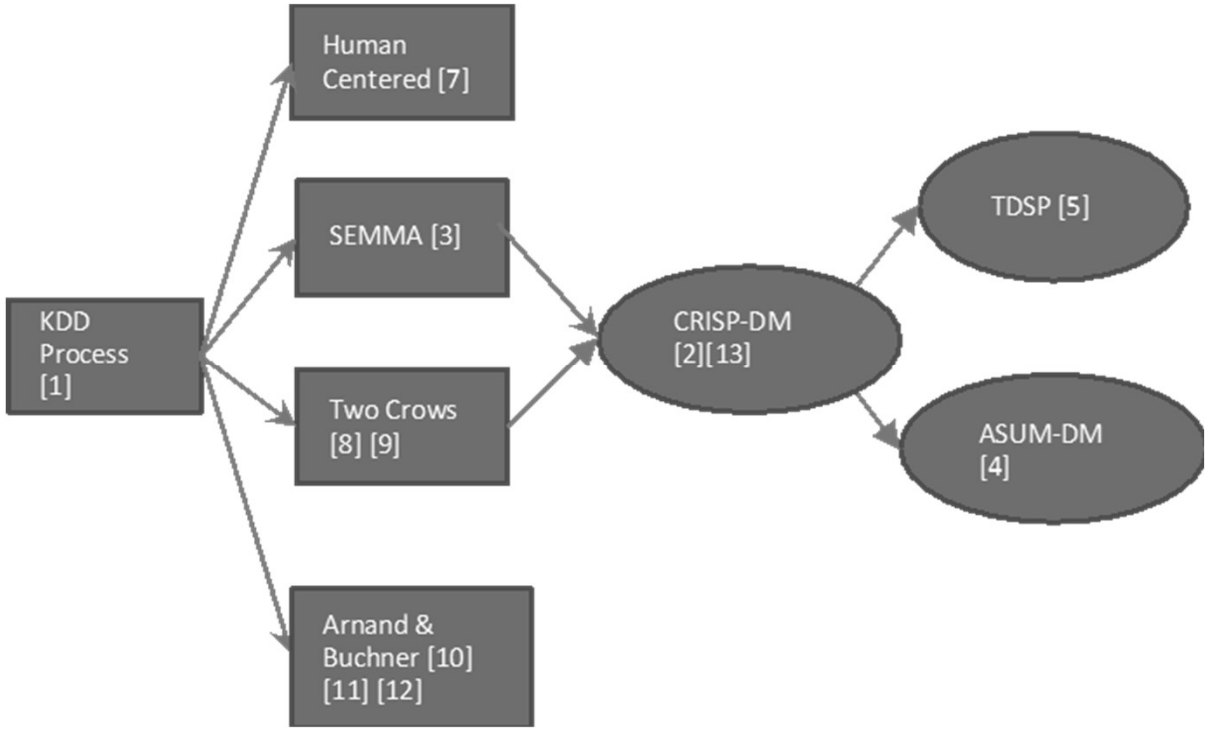


Literature Review

- It is in this context crucial to identify possible approaches.
- Identification of main limitations

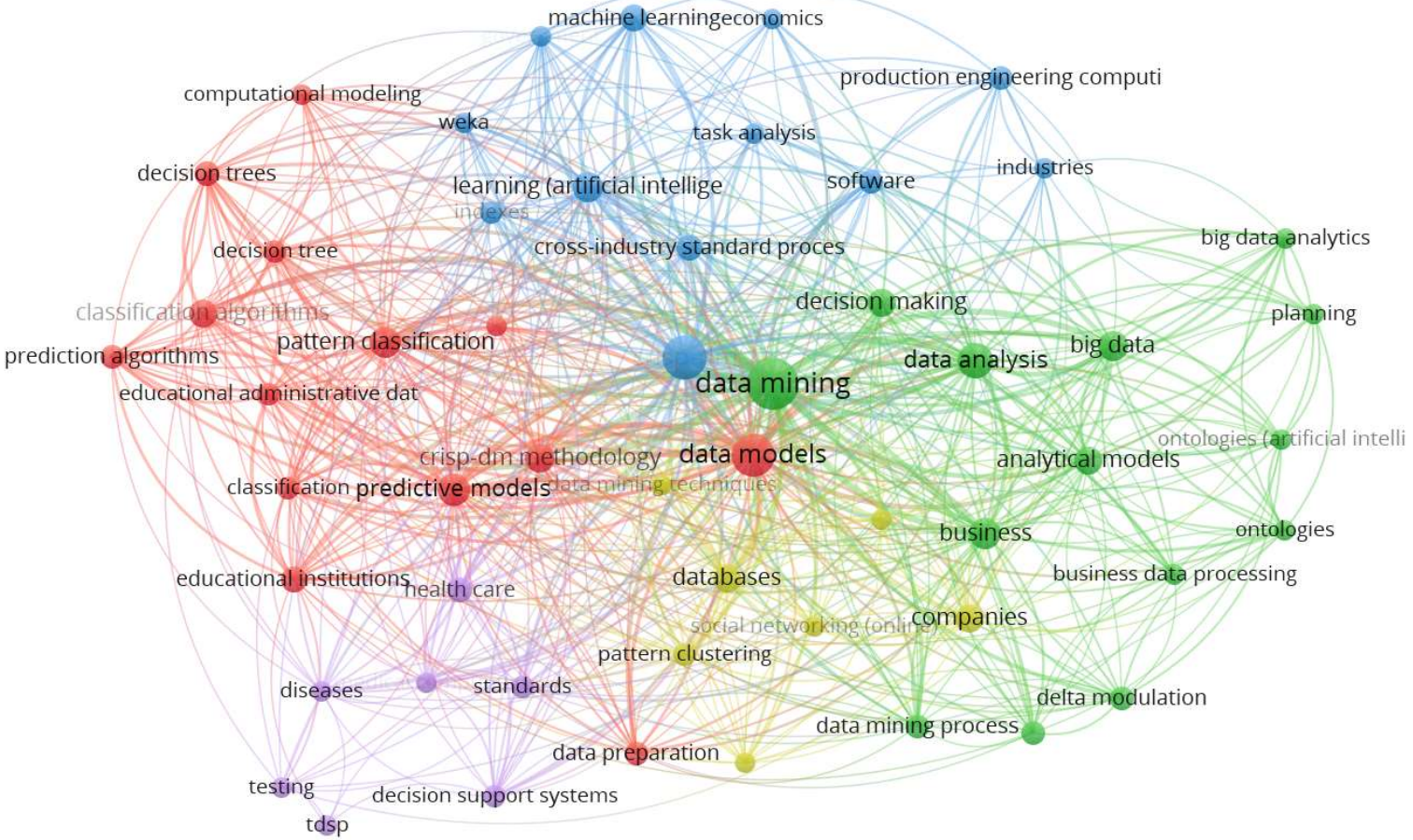
Literature Review

- Process



Costa & Aparicio (2020)

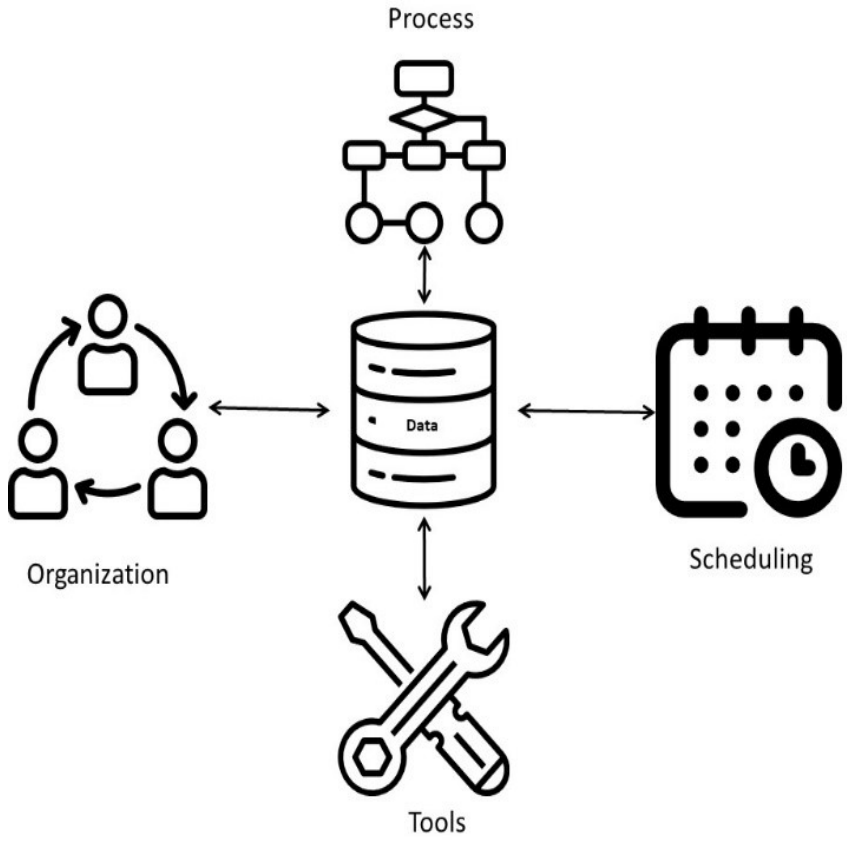
Literature Review



Literature Review

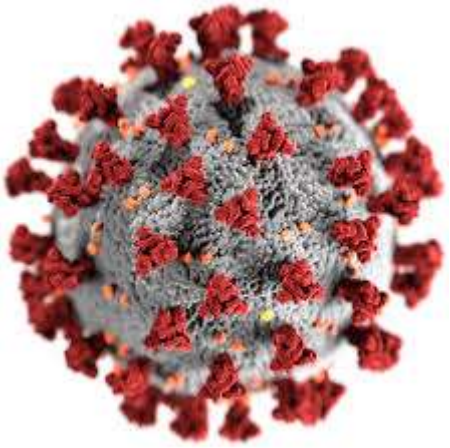
- Summarizing, the approaches related to data mining, machine learning and data science may be interrelated.
- CRISP-DM is one of the most used and the one that inspired many other approaches.
- Nevertheless, other features may be added to this approach:
  - Organization
  - Scheduling
  - Tools

Proposing a Model

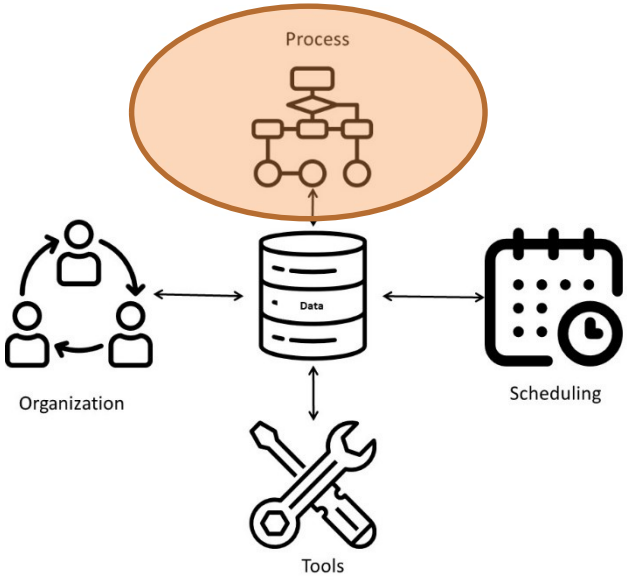


Using the model

- Context of Covid19
- Using Data From several sources
- Need of many skills



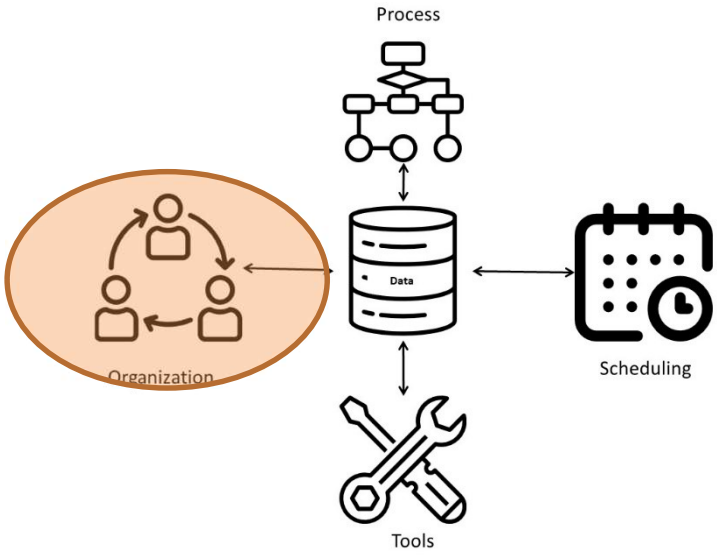
Process



Process: CRISP-DM

- Process
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modelling/Chart
  - Evaluation
  - Deployment

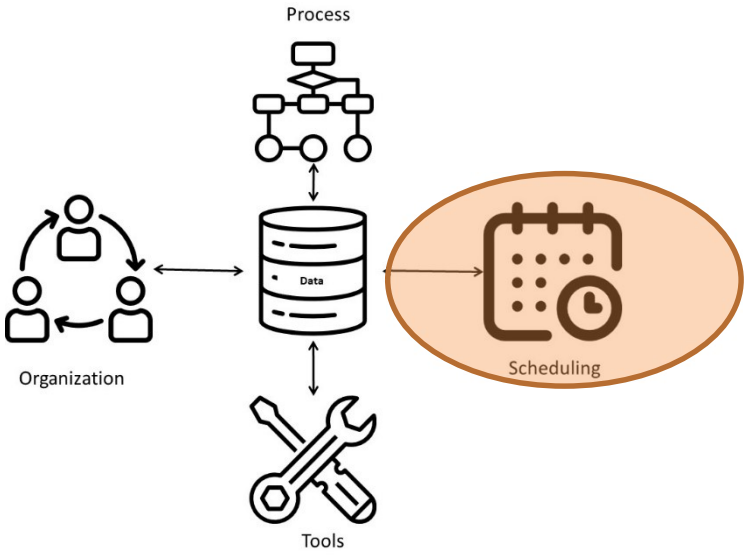
Organization



Organization: RACI

		BA	DE	DS	WD	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w
1	**Business Understanding**																	
1.1.	Define Business Objectives																	
1.2.	Identify ethical values and privacy	A/R				L												
1.3.	Assess Situation	A/R				L												
1.4.	Define Data Science Goals	A/R				L												
1.5.	Produce Project Plan	A/R	R	R		L												
2	**Data Understanding**																	
2.1.	Collect Initial Data		A/R			H												
2.2.	Describe Data		A/R			L												
2.3.	Explore Data		A/R			M												
2.4.	Verify Data Quality			A/R		H												
3	**Data Preparation**			A/R														
3.1.	Select Data			A/R		M												
3.2.	Clean Data			A/R		M												
3.3.	Construct Data			A/R		M												
3.4.	Integrate Data			A/R		H												
3.4.	Format Data			A/R		H												
4	**Modeling**																	
4.1.	Select Modeling Techniques	I		A/R		H												
4.2.	Generate Test Design	I		A/R		H												
4.3.	Build Model	I		A/R		M												
4.4.	Assess Model	I		A/R		H												

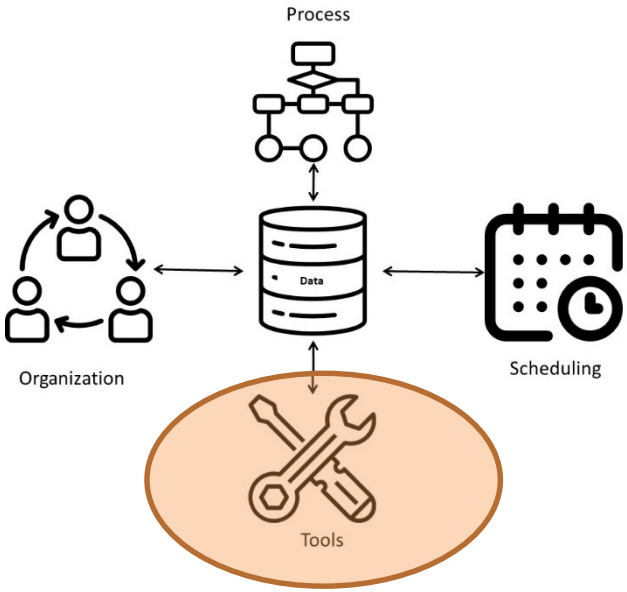
Scheduling



Scheduling: GANTT

		BA	DE	DS	WD	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w
1	**Business Understanding**																	
1.1.	Define Business Objectives																	
1.2.	Identify ethical values and privacy	A/R				L												
1.3.	Assess Situation	A/R				L												
1.4.	Define Data Science Goals	A/R				L												
1.5.	Produce Project Plan	A/R	R	R		L												
2	**Data Understanding**																	
2.1.	Collect Initial Data		A/R			H												
2.2.	Describe Data		A/R			L												
2.3.	Explore Data		A/R			M												
2.4.	Verify Data Quality			A/R		H												
3	**Data Preparation**			A/R														
3.1.	Select Data			A/R		M												
3.2.	Clean Data			A/R		M												
3.3.	Construct Data			A/R		M												
3.4.	Integrate Data			A/R		H												
3.4.	Format Data			A/R		H												
4	**Modeling**																	
4.1.	Select Modeling Techniques	I		A/R		H												
4.2.	Generate Test Design	I		A/R		H												
4.3.	Build Model	I		A/R		M												
4.4.	Assess Model	I		A/R		H												

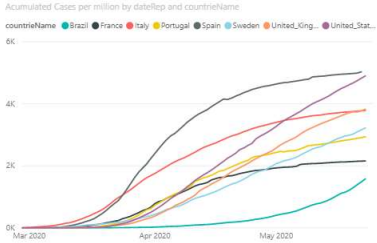
Tools

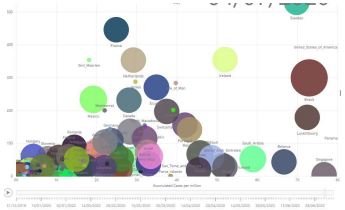


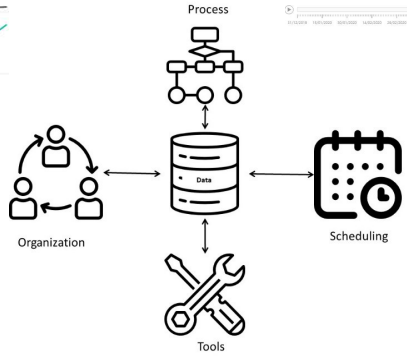
Tools

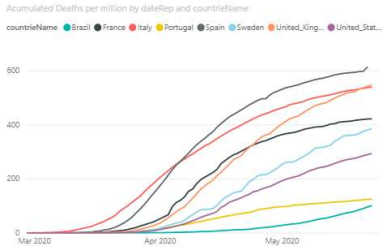
- Charting approaches
- Modeling concepts
- Techniques
- Programming Languages

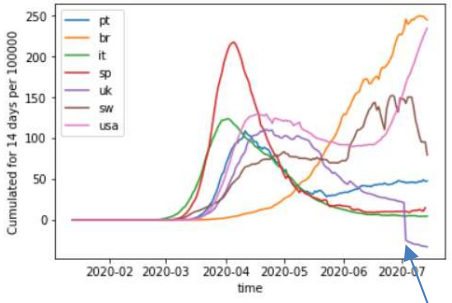
Results



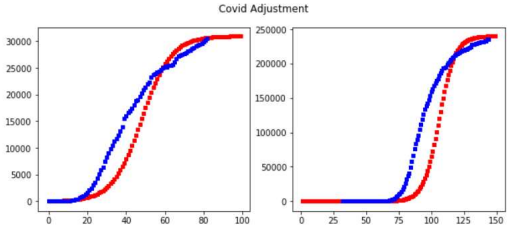


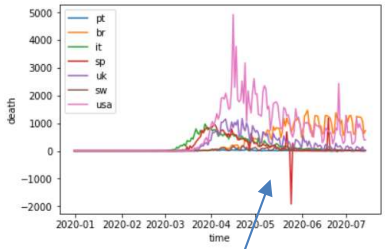

Process
Organization
Data
Scheduling
Tools




pt
br
it
sp
uk
sw
usa
Cumulated for 14 days per 100000
2020-02 2020-03 2020-04 2020-05 2020-06 2020-07
time

?


Covid Adjustment


pt
br
it
sp
uk
sw
usa
death
2020-01 2020-02 2020-03 2020-04 2020-05 2020-06 2020-07
time

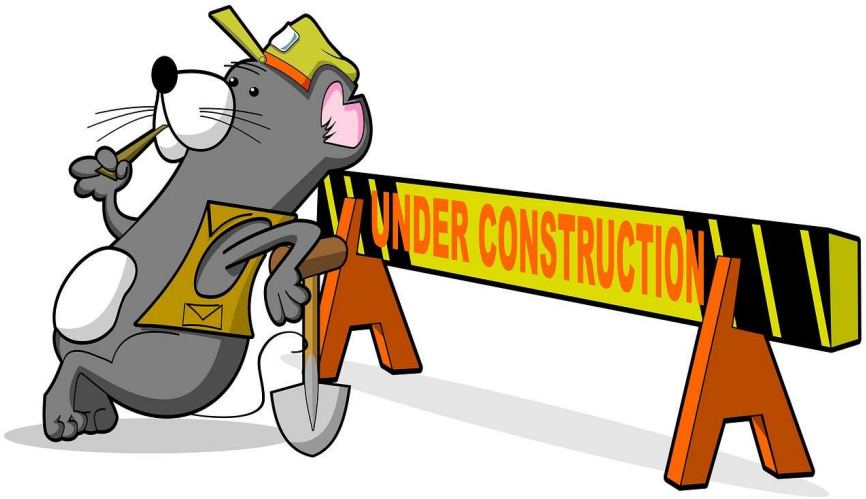
?

Conclusions

- Adequate Approach
- Many roles and people with different backgrounds
- Improve organization contribution
- Improve scheduling
- Allows results vs. expectations adjustment
- Main limitation: Bureaucracy

Future Work

- Risk analysis
- Budgeting



References

- Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019). Data Science and AI: trends analysis. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE.
- Costa, C. J. & Aparicio, J.T. (2020). POST-DS: A Methodology to Boost Data Science . In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE.


Thank you