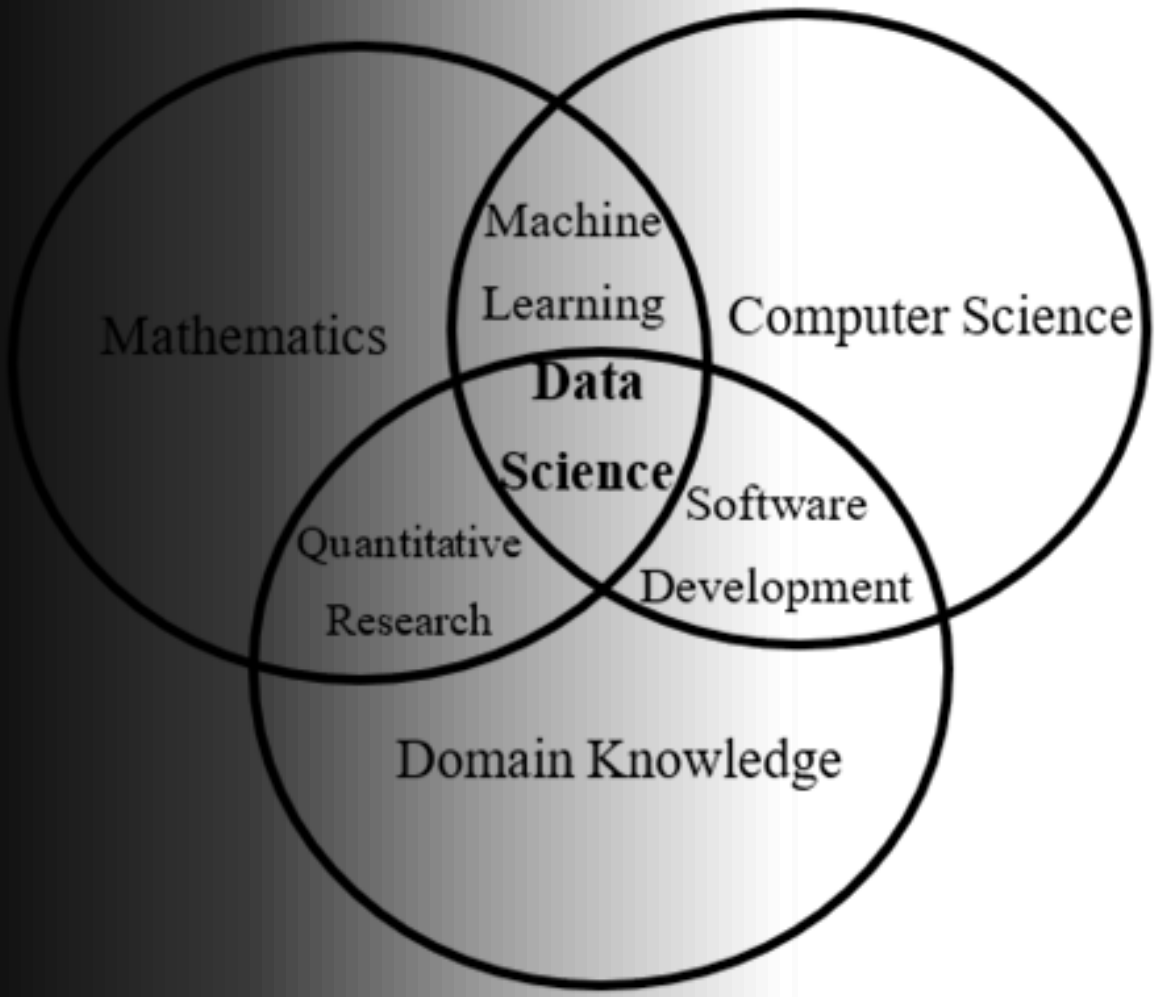
DATA SCIENCE PROJECT DEVELOPMENT

Carlos J. Costa, ISEG

Lisbon School of Economics & Management
Universidade de Lisboa

Context

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



Aparicio et al.(2019).

PROJECT MANAGEMENT

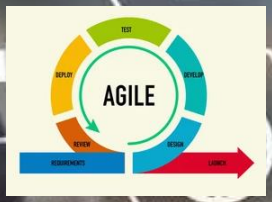
Project Management Institute.

IPMA®

PRINCE2®

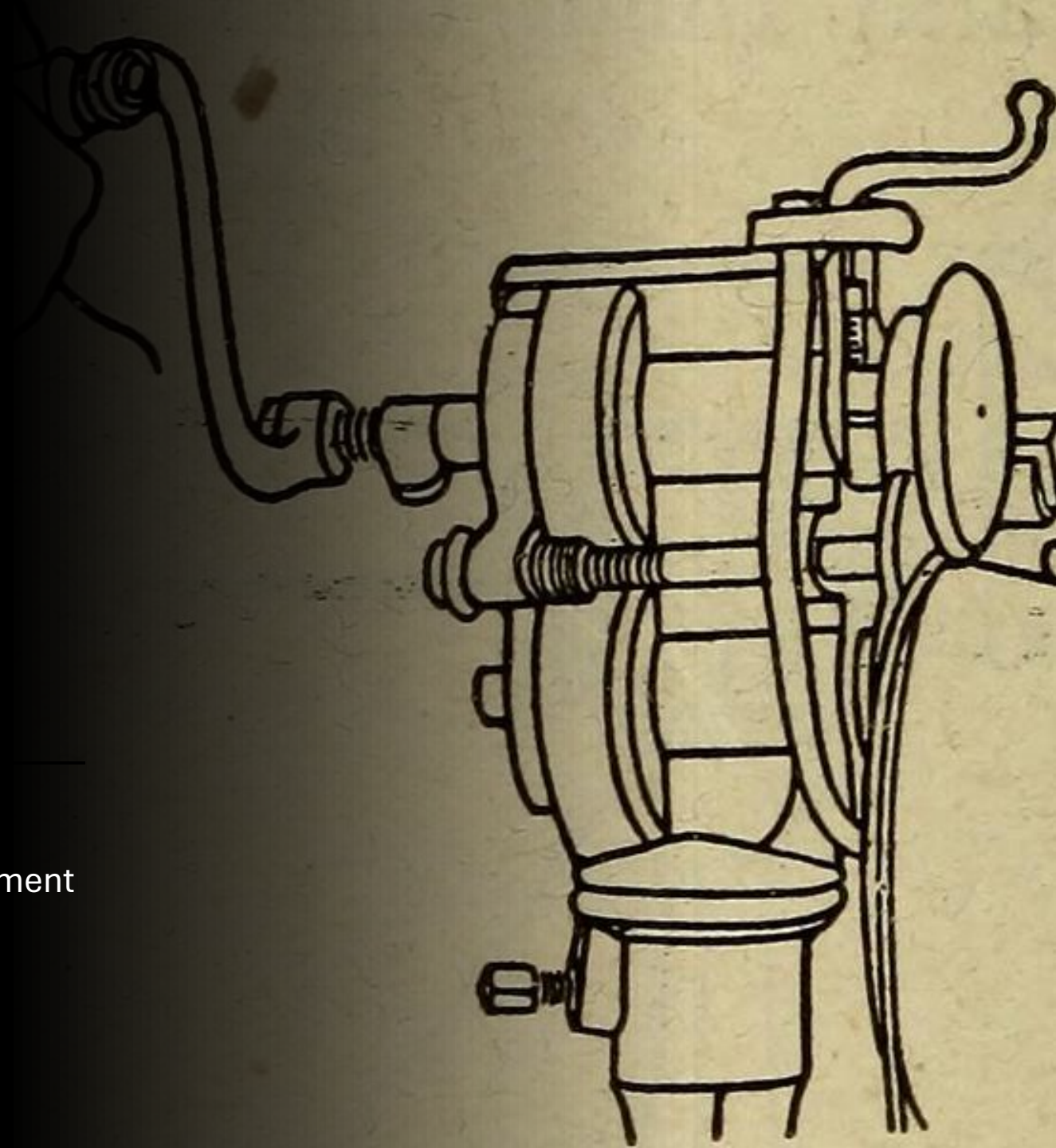
AXELOS GLOBAL BEST PRACTICE

Scrum Alliance CSPO CERTIFIED



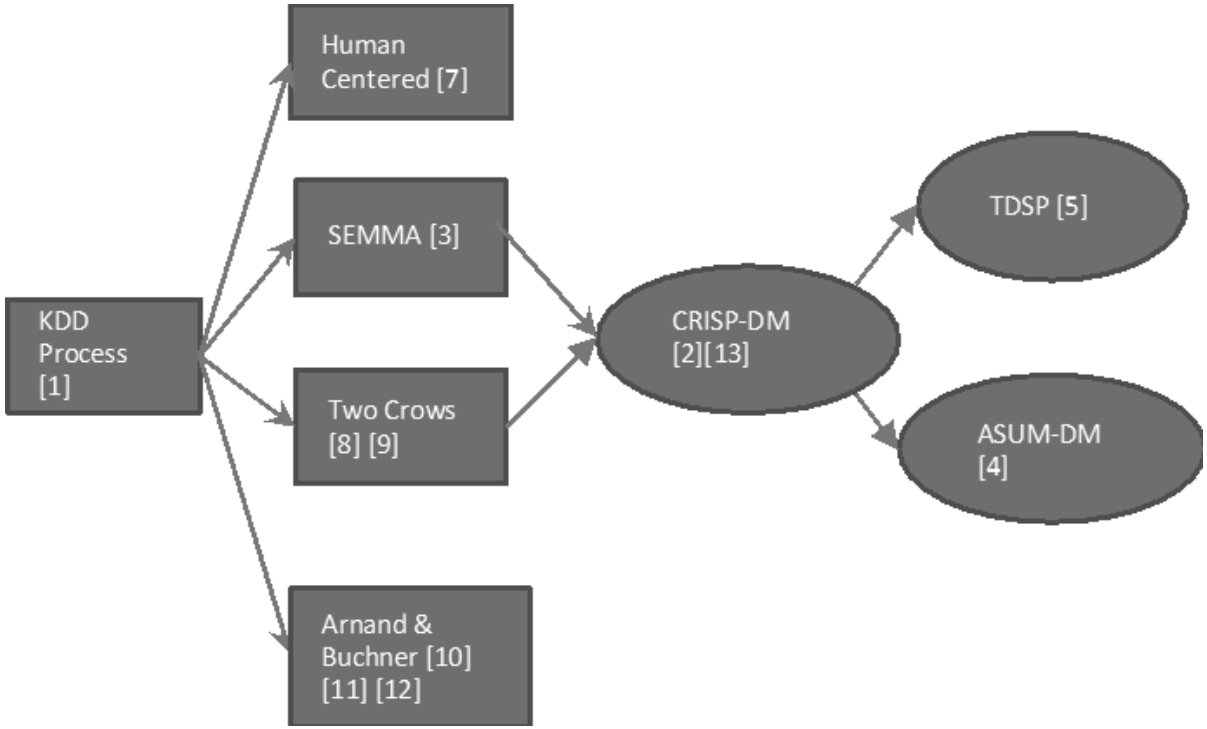
Context

methodology that may contribute to the improvement of the knowledge creation outputs.

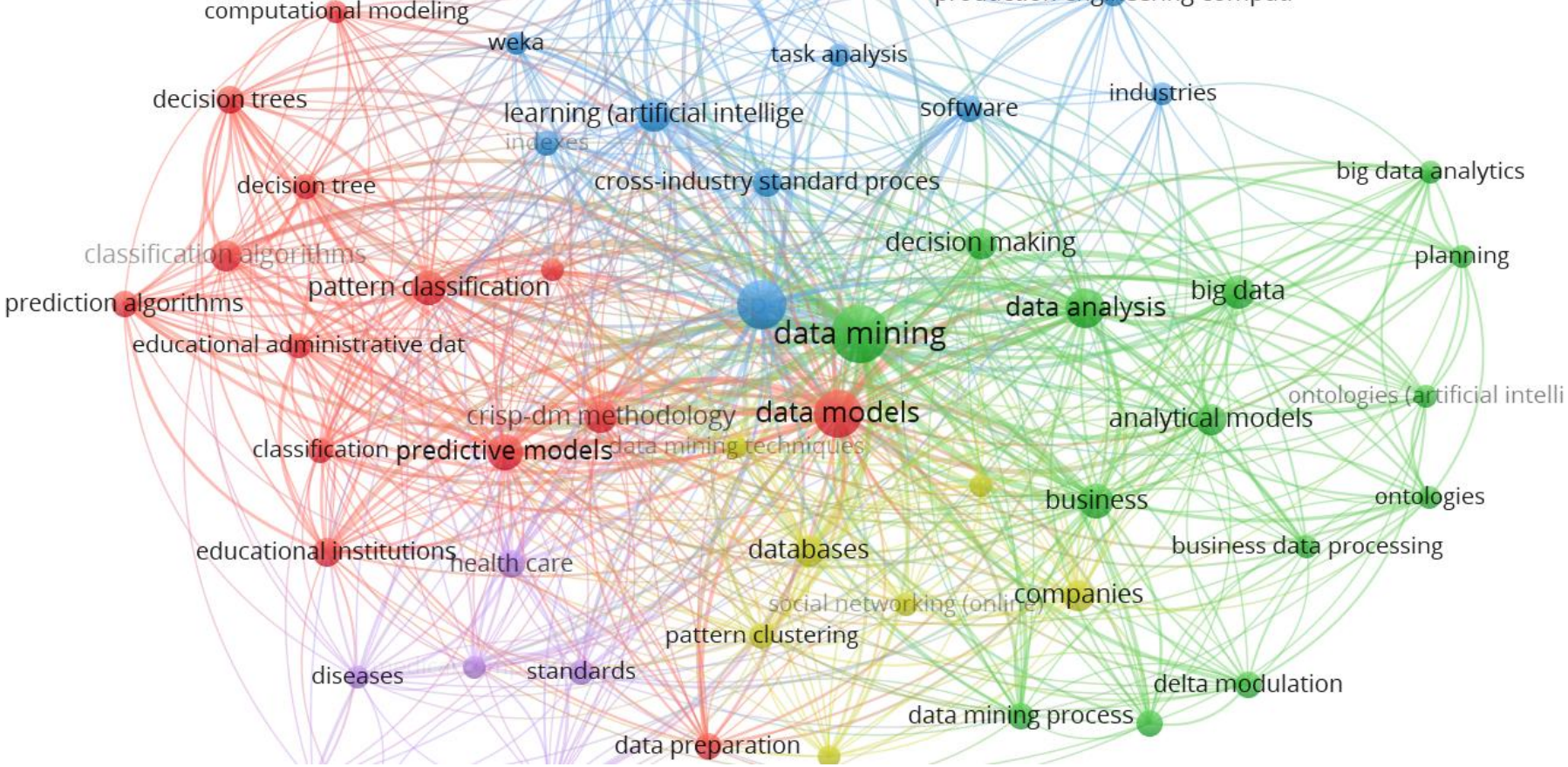


Related Work

- Process



Costa & Aparicio (2020)



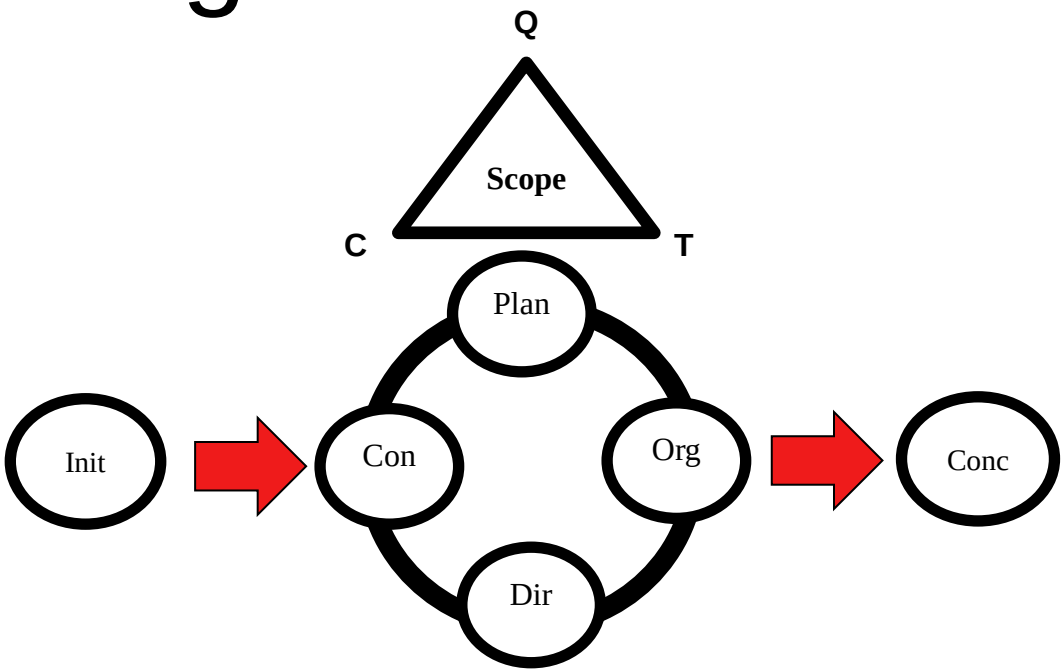
Related Work

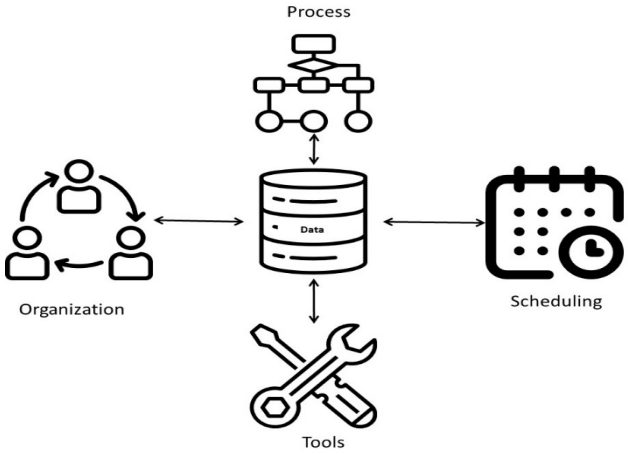
Costa & Aparicio (2021)

Related Work

- 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

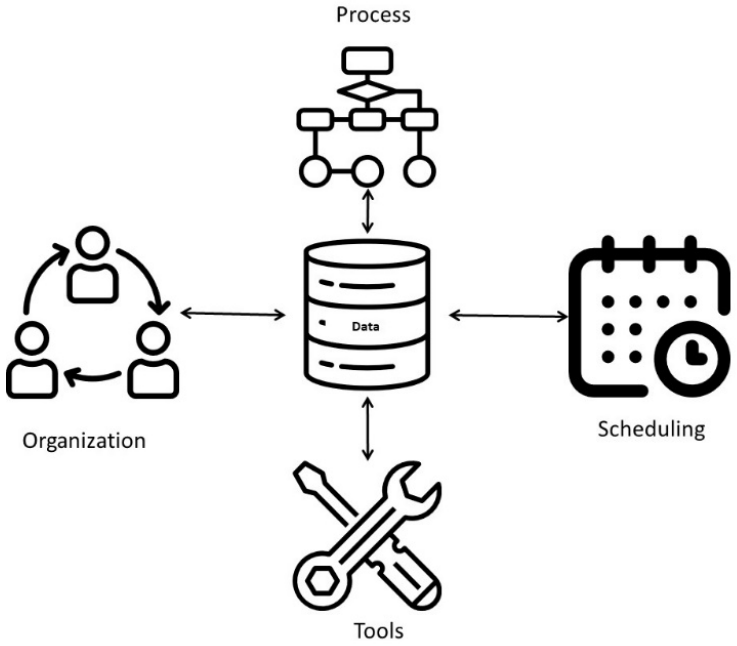
Proposing a Model



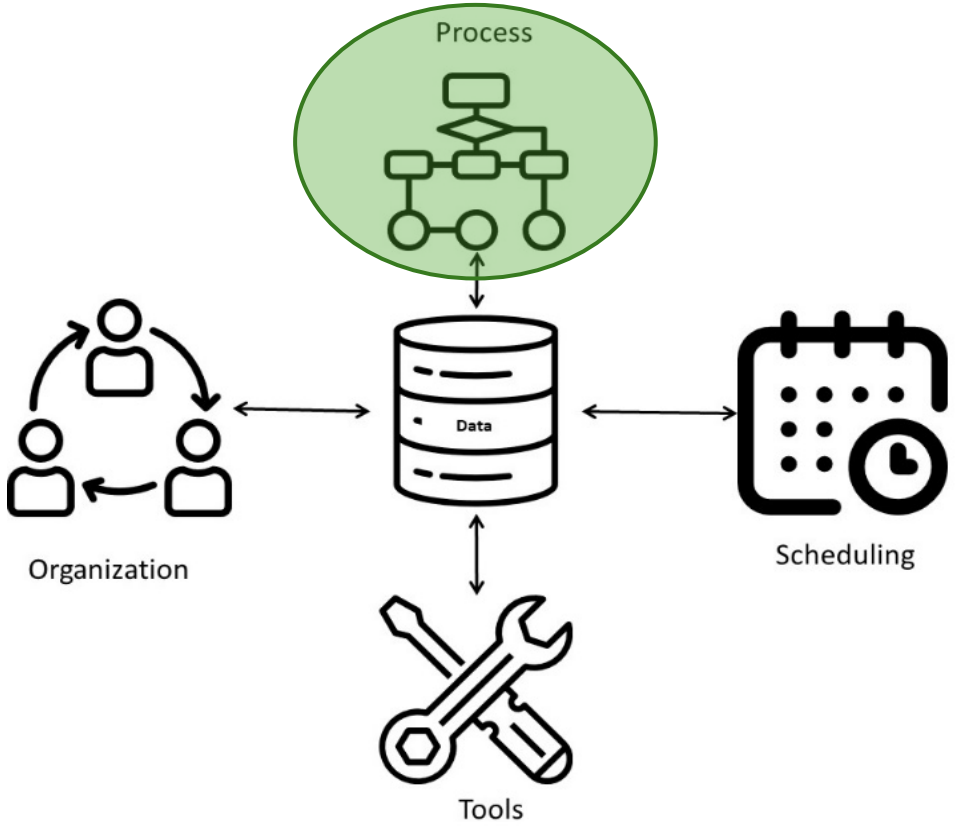


Proposing a Model

- POST-DS



Process



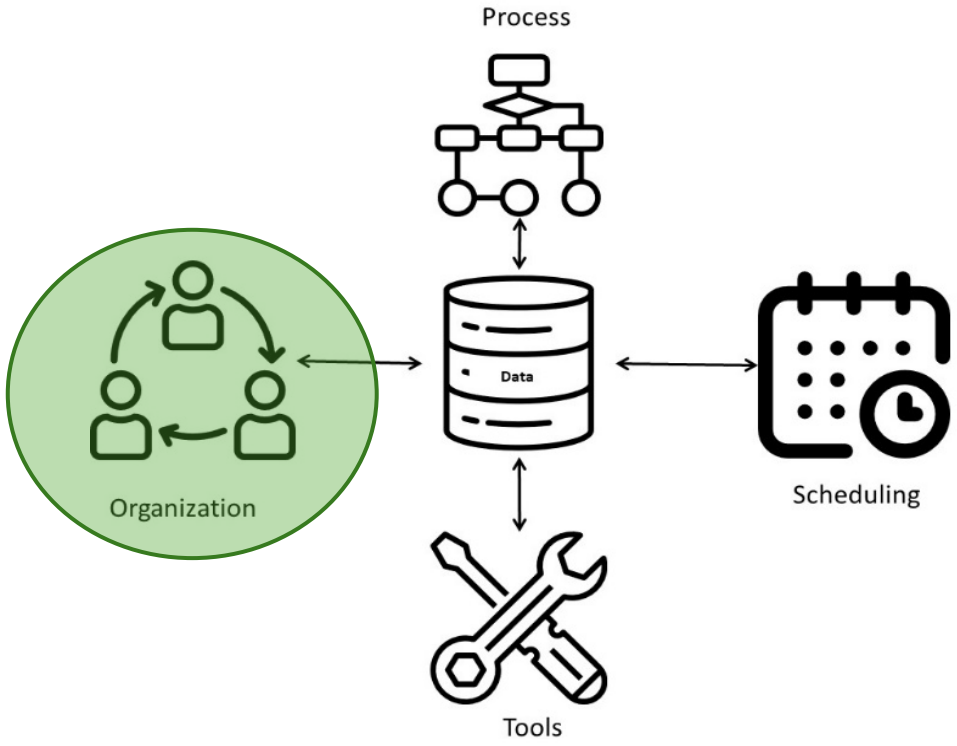
Process

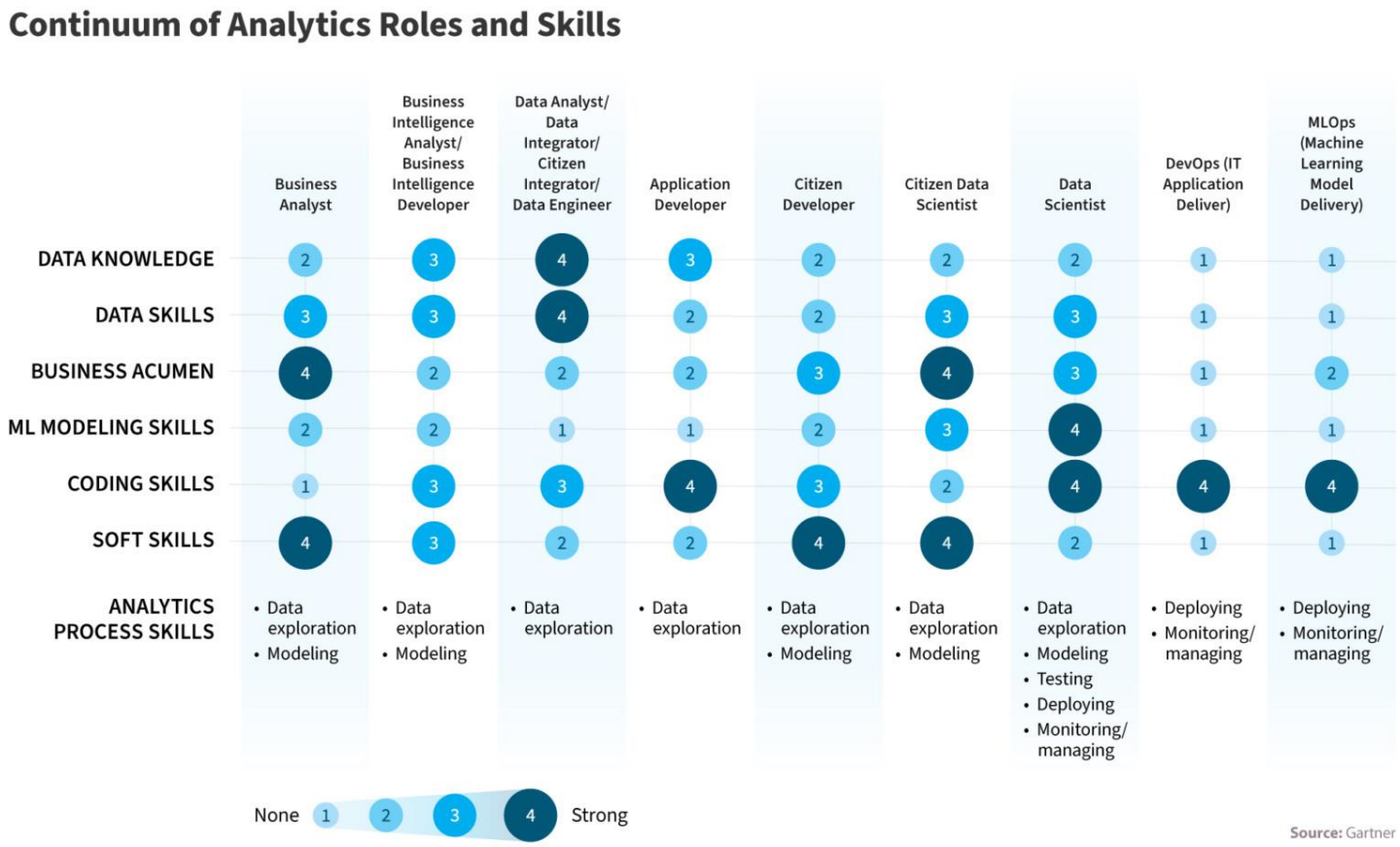
- Process
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modelling
  - Evaluation
  - Deployment

Process

		BA	DE	DS	WD	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	Tools and Resource
	Business Understanding																				
.1.	Define Business Objectives																				
.2.	Identify ethical values and privacy	A/R				L															meeting
.3.	Assess Situation	A/R				L															meeting
.4.	Define Data Science Goals	A/R				L															meeting
.5.	Produce Project Plan	A/R	R	R		L															WBS, GANTT
	Data Understanding																				
.1.	Collect Initial Data		A/R			H															open data, scraping,
.2.	Describe Data		A/R			L															use Jupyter/python/Pandas
.3.	Explore Data		A/R			M															use Jupyter/python/Pandas
.4.	Verify Data Quality			A/R		H															use Jupyter/python/Pandas
	Data Preparation			A/R																	
.1.	Select Data			A/R		M															Meeting
.2.	Clean Data			A/R		M															use Jupyter/python/Pandas
.3.	Construct Data			A/R		M															use Jupyter/python/Pandas
.4.	Integrate Data			A/R		H															use Jupyter/python/Pandas
.4.	Format Data			A/R		H															use Jupyter/python/Pandas
	Modeling																				
.1.	Select Modeling Techniques			A/R		H															MIT flowchart
.2.	Generate Test Design			A/R		H															use Jupyter/python/Pandas
.3.	Build Model			A/R		M															use Jupyter/python/Pandas
.4.	Assess Model			A/R		H															use Jupyter/python/Pandas
	Evaluation																				
.1.	Evaluate Results, icnluding ethical	A/R		R		H															use Jupyter/python/Pandas
.2.	Review Process	A/R				L															meeting
.3.	Determine Next Steps	A/R				L															meeting
	Deployment																				
.1.	Plan Deployment	A		R	R	H															PowerBI or Flash
.2.	Plan Monitoring and Maintenance	A				M															meeting
.3.	Produce Final Report	A/R	R	R	R	M															PowerBI or Flash
.4.	Review Project	A/R		R		M															meeting

Organization



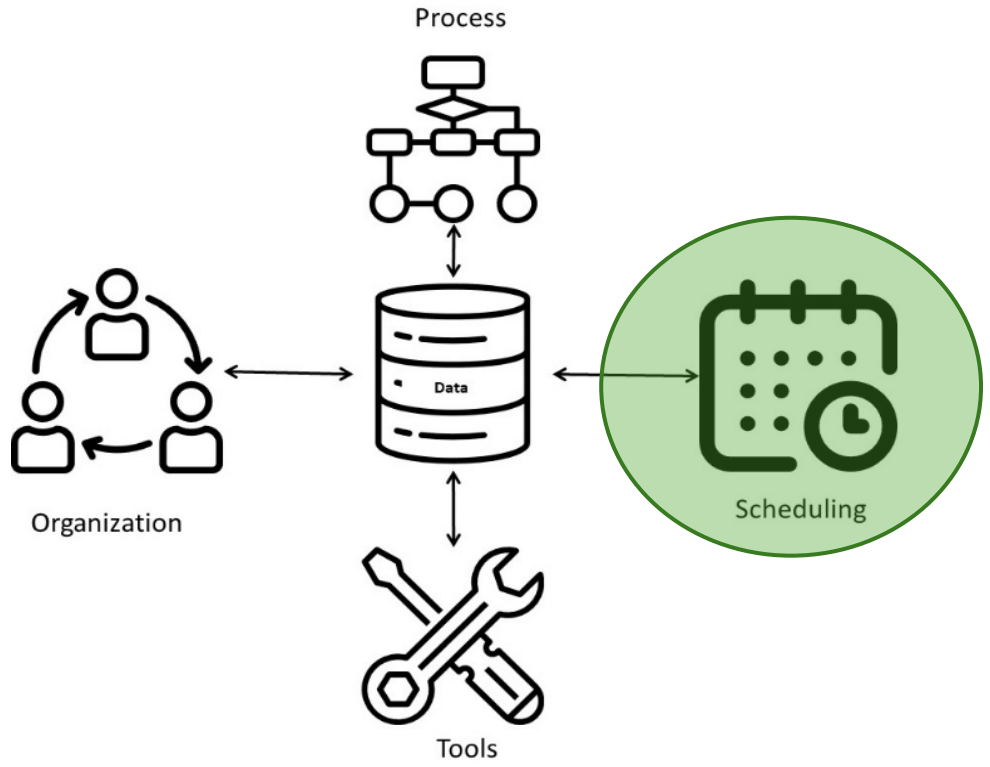


Organization

Organization

		A	DE	DS	WL	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	Tools and Resource
1	**Business Understanding**																				
1.1.	Define Business Objectives																				
1.2.	Identify ethical values and privacy	/R				L															meeting
1.3.	Assess Situation	/R				L															meeting
1.4.	Define Data Science Goals	/R				L															meeting
1.5.	Produce Project Plan	/R	R	R		L															WBS, GANTT
2	**Data Understanding**																				
2.1.	Collect Initial Data		A/R			H															open data, scraping,
2.2.	Describe Data		A/R			L															use Jupyter/python/Pandas
2.3.	Explore Data		A/R			M															use Jupyter/python/Pandas
2.4.	Verify Data Quality			A/R		H															use Jupyter/python/Pandas
3	**Data Preparation**			A/R																	
3.1.	Select Data			A/R		M															Meeting
3.2.	Clean Data			A/R		M															use Jupyter/python/Pandas
3.3.	Construct Data			A/R		M															use Jupyter/python/Pandas
3.4.	Integrate Data			A/R		H															use Jupyter/python/Pandas
3.4.	Format Data			A/R		H															use Jupyter/python/Pandas
4	**Modeling**																				
4.1.	Select Modeling Techniques			A/R		H															MIT flowchart
4.2.	Generate Test Design			A/R		H															use Jupyter/python/Pandas
4.3.	Build Model			A/R		M															use Jupyter/python/Pandas
4.4.	Assess Model			A/R		H															use Jupyter/python/Pandas
5	**Evaluation**																				
5.1.	Evaluate Results, icnluding ethical	/R		R		H															use Jupyter/python/Pandas
5.2.	Review Process	/R				L															meeting
5.3.	Determine Next Steps	/R				L															meeting
6	**Deployment**																				
6.1.	Plan Deployment			R	R	H															PowerBI or Flash
6.2.	Plan Monitoring and Maintenance					M															meeting
6.3.	Produce Final Report	/R	R	R	R	M															PowerBI or Flash
6.4.	Review Project	/R		R		M															meeting

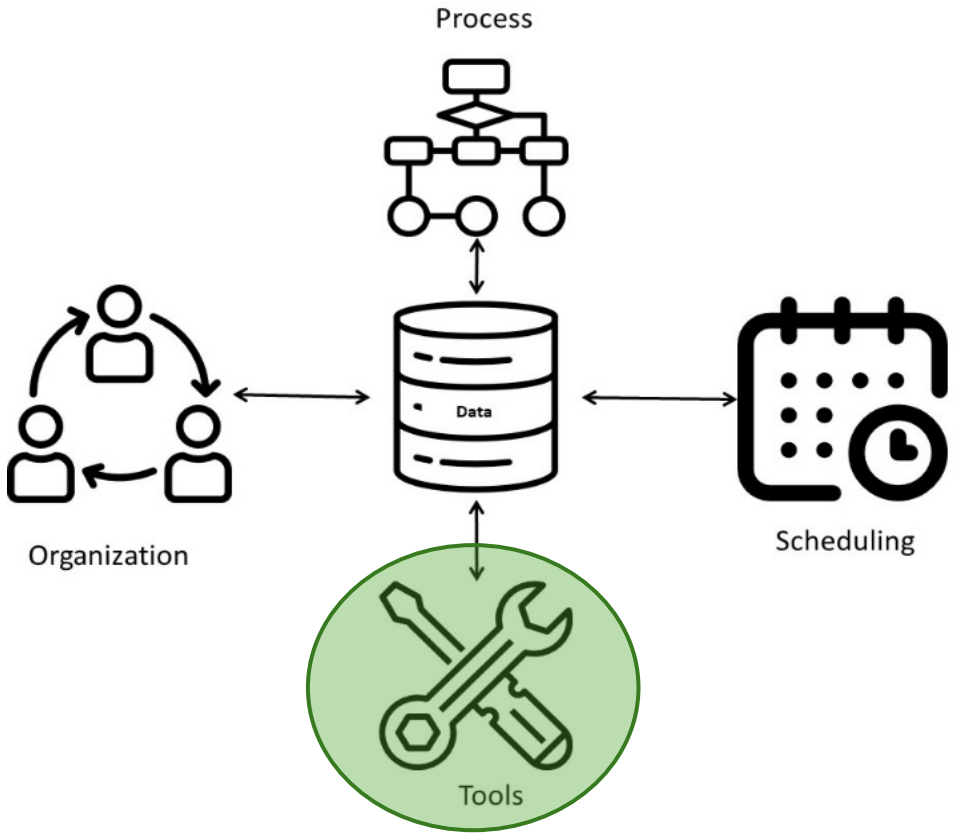
Scheduling



Scheduling

		BA	DE	DS	WD	Risk	w	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	14	Tools and Resource
1	**Business Understanding**																				
1.1.	Define Business Objectives																				
1.2.	Identify ethical values and privacy	A/R																			eeting
1.3.	Assess Situation	A/R																			eeting
1.4.	Define Data Science Goals	A/R																			eeting
1.5.	Produce Project Plan	A/R	R	R																	VBS, GANTT
2	**Data Understanding**																				
2.1.	Collect Initial Data		A/R																		pen data, scraping,
2.2.	Describe Data		A/R																		se Jupyter/python/Pandas
2.3.	Explore Data		A/R																		se Jupyter/python/Pandas
2.4.	Verify Data Quality			A/R																	se Jupyter/python/Pandas
3	**Data Preparation**			A/R																	
3.1.	Select Data			A/R																	leeting
3.2.	Clean Data			A/R																	se Jupyter/python/Pandas
3.3.	Construct Data			A/R																	se Jupyter/python/Pandas
3.4.	Integrate Data			A/R																	se Jupyter/python/Pandas
3.4.	Format Data			A/R																	se Jupyter/python/Pandas
4	**Modeling**																				
4.1.	Select Modeling Techniques	I		A/R																	IIT flowchart
4.2.	Generate Test Design	I		A/R																	se Jupyter/python/Pandas
4.3.	Build Model	I		A/R																	se Jupyter/python/Pandas
4.4.	Assess Model	I		A/R																	se Jupyter/python/Pandas
5	**Evaluation**																				
5.1.	Evaluate Results, icnluding ethical	A/R		R																	se Jupyter/python/Pandas
5.2.	Review Process	A/R																			eeting
5.3.	Determine Next Steps	A/R																			eeting
6	**Deployment**																				
6.1.	Plan Deployment	A		R	R																owerBI or Flash
6.2.	Plan Monitoring and Maintenance	A																			eeting
6.3.	Produce Final Report	A/R	R	R	R	M															PowerBI or Flash
6.4.	Review Project	A/R		R		M															meeting

Tools



Tools

		BA	DE	DS	WD	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	ools and Resource
1	**Business Understanding**																				
1.1.	Define Business Objectives																				
1.2.	Identify ethical values and privacy	A/R				L															meeting
1.3.	Assess Situation	A/R				L															meeting
1.4.	Define Data Science Goals	A/R				L															meeting
1.5.	Produce Project Plan	A/R	R	R		L															WBS, GANTT
2	**Data Understanding**																				
2.1.	Collect Initial Data		A/R			H															open data, scraping,
2.2.	Describe Data		A/R			L															use Jupyter/python/Pandas
2.3.	Explore Data		A/R			M															use Jupyter/python/Pandas
2.4.	Verify Data Quality			A/R		H															use Jupyter/python/Pandas
3	**Data Preparation**			A/R																	
3.1.	Select Data			A/R		M															Meeting
3.2.	Clean Data			A/R		M															use Jupyter/python/Pandas
3.3.	Construct Data			A/R		M															use Jupyter/python/Pandas
3.4.	Integrate Data			A/R		H															use Jupyter/python/Pandas
3.4.	Format Data			A/R		H															use Jupyter/python/Pandas
4	**Modeling**																				
4.1.	Select Modeling Techniques	I		A/R		H															MIT flowchart
4.2.	Generate Test Design	I		A/R		H															use Jupyter/python/Pandas
4.3.	Build Model	I		A/R		M															use Jupyter/python/Pandas
4.4.	Assess Model	I		A/R		H															use Jupyter/python/Pandas
5	**Evaluation**																				
5.1.	Evaluate Results, icnluding ethical	A/R		R		H															use Jupyter/python/Pandas
5.2.	Review Process	A/R				L															meeting
5.3.	Determine Next Steps	A/R				L															meeting
6	**Deployment**																				
6.1.	Plan Deployment	A		R	R	H															PowerBI or Flash
6.2.	Plan Monitoring and Maintenance	A				M															meeting
6.3.	Produce Final Report	A/R	R	R	R	M															PowerBI or Flash
6.4.	Review Project	A/R		R		M															eeting

Tools

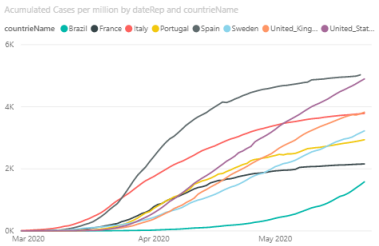
- Charting approaches
- Modeling concepts: supervised algorithms and non supervised algorithms
- Techniques
- Programming Languages

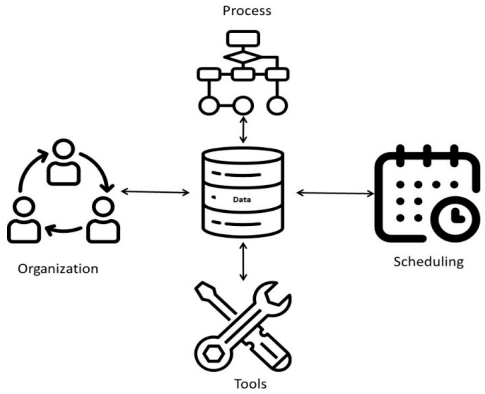
POST-DS

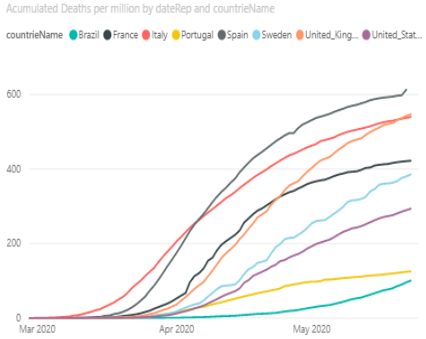
		BA	DE	DS	WD	Risk	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	Tools and Resource
1	**Business Understanding**																				
1.1.	Define Business Objectives																				
1.2.	Identify ethical values and privacy	A/R				L															meeting
1.3.	Assess Situation	A/R				L															meeting
1.4.	Define Data Science Goals	A/R				L															meeting
1.5.	Produce Project Plan	A/R	R	R		L															WBS, GANTT
2	**Data Understanding**																				
2.1.	Collect Initial Data		A/R			H															open data, scraping,
2.2.	Describe Data		A/R			L															use Jupyter/python/Pandas
2.3.	Explore Data		A/R			M															use Jupyter/python/Pandas
2.4.	Verify Data Quality			A/R		H															use Jupyter/python/Pandas
3	**Data Preparation**			A/R																	
3.1.	Select Data			A/R		M															Meeting
3.2.	Clean Data			A/R		M															use Jupyter/python/Pandas
3.3.	Construct Data			A/R		M															use Jupyter/python/Pandas
3.4.	Integrate Data			A/R		H															use Jupyter/python/Pandas
3.4.	Format Data			A/R		H															use Jupyter/python/Pandas
4	**Modeling**																				
4.1.	Select Modeling Techniques	I		A/R		H															MIT flowchart
4.2.	Generate Test Design	I		A/R		H															use Jupyter/python/Pandas
4.3.	Build Model	I		A/R		M															use Jupyter/python/Pandas
4.4.	Assess Model	I		A/R		H															use Jupyter/python/Pandas
5	**Evaluation**																				
5.1.	Evaluate Results, icnluding ethical	A/R		R		H															use Jupyter/python/Pandas
5.2.	Review Process	A/R				L															meeting
5.3.	Determine Next Steps	A/R				L															meeting
6	**Deployment**																				
6.1.	Plan Deployment	A		R	R	H															PowerBI or Flash
6.2.	Plan Monitoring and Maintenance	A				M															meeting
6.3.	Produce Final Report	A/R	R	R	R	M															PowerBI or Flash
6.4.	Review Project	A/R		R		M															meeting

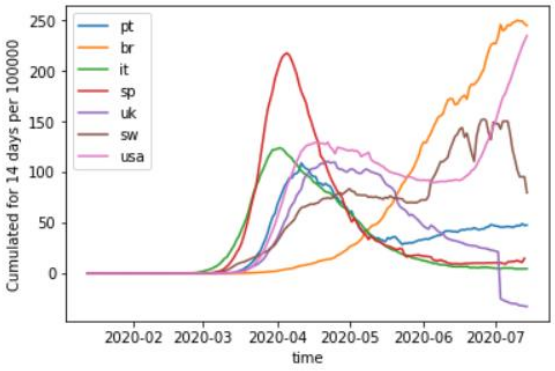
Use Cases

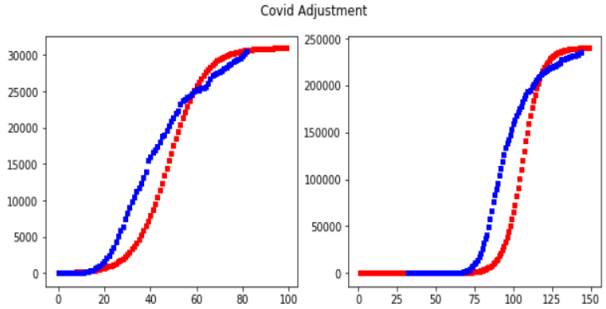
- Data Science and Business Analytics
- Covid
- Financial market
- Software Development Business
- Academic context
- Professional work

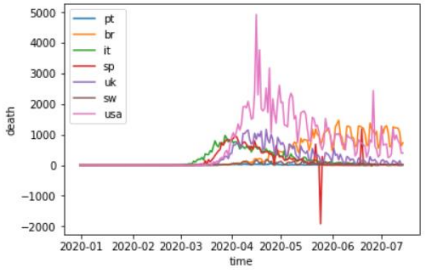


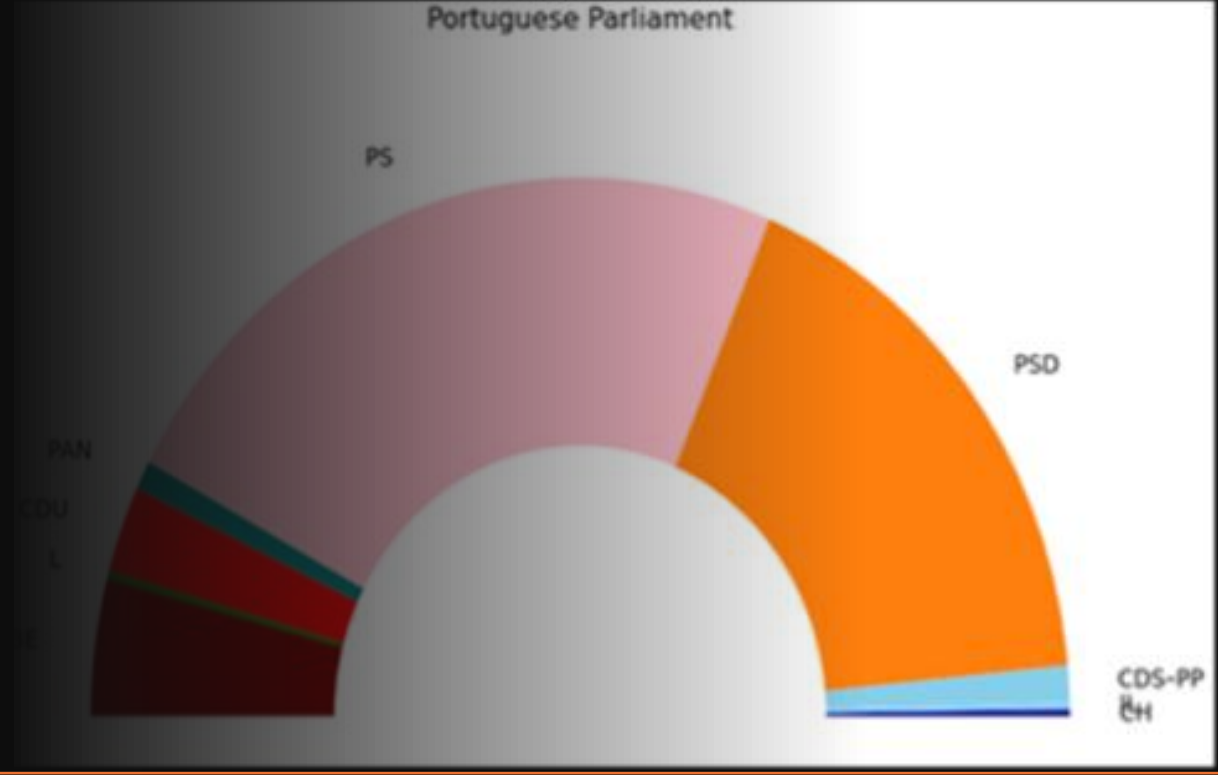










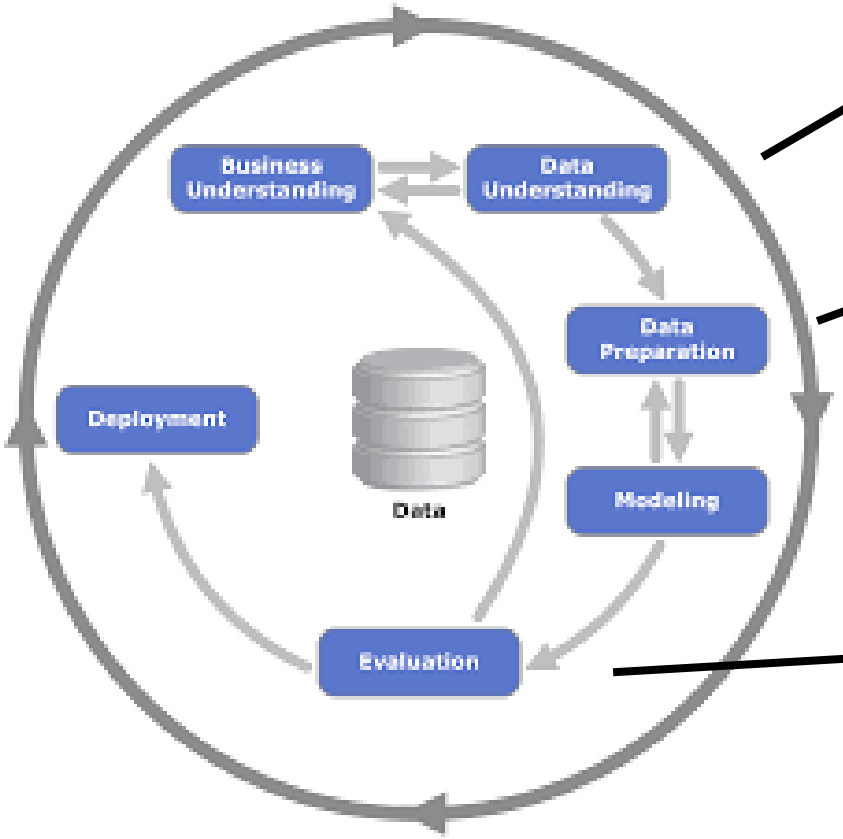


Use Cases

Emotion analysis of Portuguese Political Parties Communication

Use Cases

- Process



Collect Tweets

Collect Programs

Clean, translate

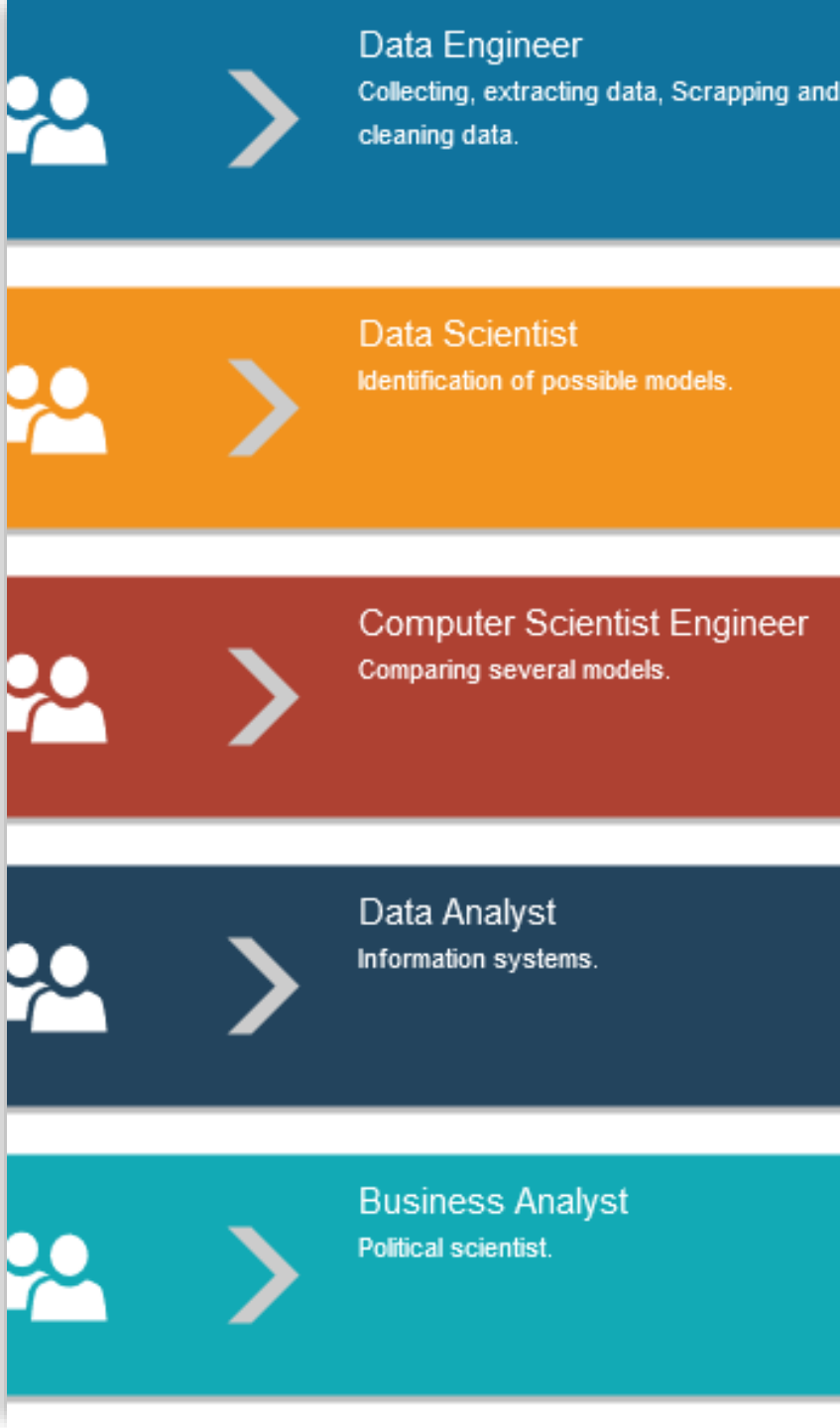
Regression

Similarity analysis

Political science analysis

Use Cases

- Process
- Organization
- Scheduling
- Tools



Use Cases

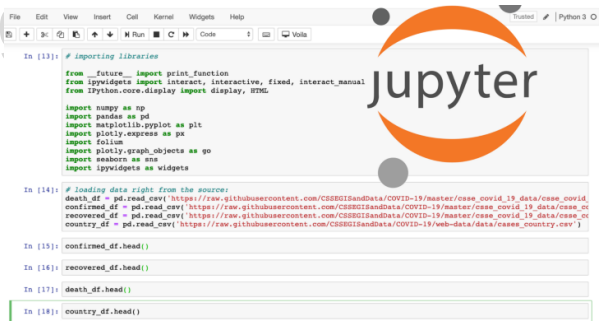
- Process
- Organization
- Scheduling
- Tools

Use Cases

- Process
- Organization
- Scheduling
- Tools

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.4423	0.415	-3.473	0.001	-2.256	-0.628
Favorites	0.2296	0.001	349.796	0.000	0.228	0.231
neg	2.0056	0.685	2.930	0.003	0.664	3.347
neu	0.3320	0.435	0.762	0.446	-0.522	1.186
pos	-0.4631	0.594	-0.780	0.436	-1.627	0.701
lenTex	-0.0038	0.001	-4.718	0.000	-0.005	-0.002
Hashtags	0.2503	0.049	5.119	0.000	0.154	0.346
Mentions	0.3500	0.086	4.069	0.000	0.181	0.519

Omnibus:	47844.549	Durbin-Watson:	1.638
Prob(Omnibus):	0.000	Jarque-Bera (JB):	738581430.277
Skew:	15.841	Prob(JB):	0.00
Kurtosis:	865.992	Cond. No.	2.59e+03



Test Data

	R2	MAE	MSE
OLS	0.764957	3.444953	104.684972
Ridge	0.764956	3.444938	104.685139
Lasso	0.765343	3.421643	104.512759
BayesianRidge	0.764878	3.438733	104.720064
Polynomial Regression	0.717263	2.814491	125.927009
Neural Network (MLP)	0.746654	2.942042	112.836870

Conclusions

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

References

- Aparicio, J.T, Sequeira, J & Costa, J. (2021) Emotion analysis of Portuguese Political Parties Communication over the covid-19 Pandemic in 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6, doi: 10.23919/CISTI52073.2021.9476557.

- 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. DOI:10.23919/CISTI.2019.8760820

- 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. Doi:10.23919/CISTI49556.2020.9140932

- Costa, C, Aparicio, M &Aparicio, J.T. (2021). Sentiment Analysis of Portuguese Political Parties Communication. In The 39th ACM International Conference on Design of Communication (SIGDOC '21). Association for Computing Machinery, New York, NY, USA, 63–69. DOI:10.1145/3472714.34736241.

- Costa CJ, Aparicio JT.(2021) A Methodology to Boost Data Science in the Context of COVID-19. Advances in Parallel & Distributed Processing, and Applications. Published online 2021:65-75. doi:10.1007/978-3-030-69984-0_7