CRISP-ML (Q):

Um framework para abordar problemas com ML

Foto: https://www.nationalmuseum.af.mil/Upcoming/Photos/igphoto/2000526757/

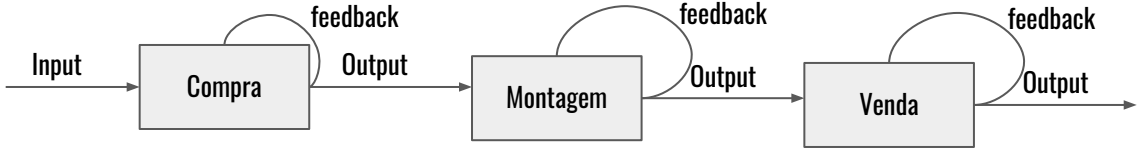
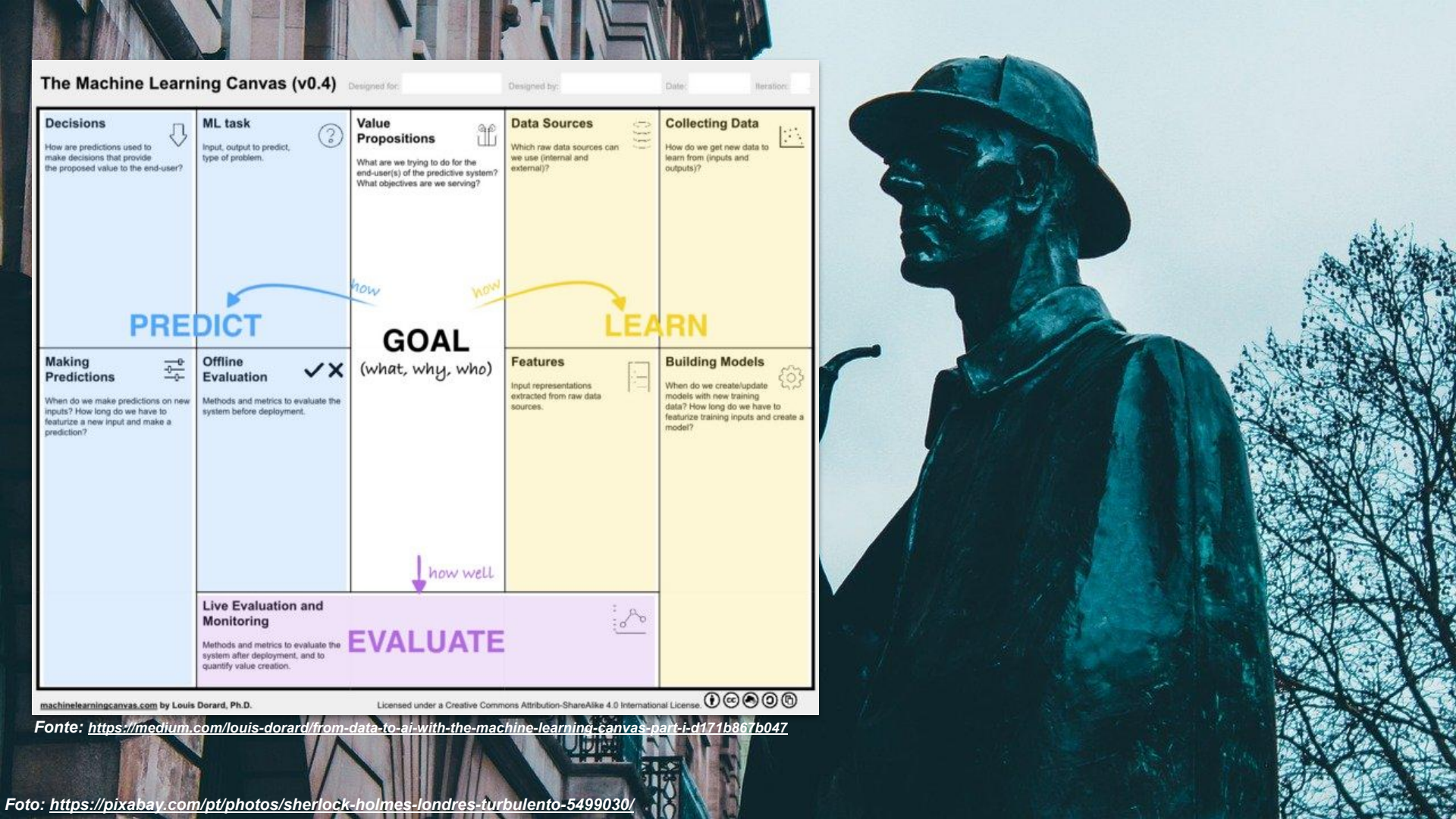


Foto: https://www.nationalmuseum.af.mil/Upcoming/Photos/igphoto/2000526757/

Foto: https://pixabay.com/pt/photos/sherlock-holmes-londres-turbulento-5499030/



Fonte: https://medium.com/louis-dorard/from-data-to-ai-with-the-machine-learning-canvas-part-i-d171b867b047

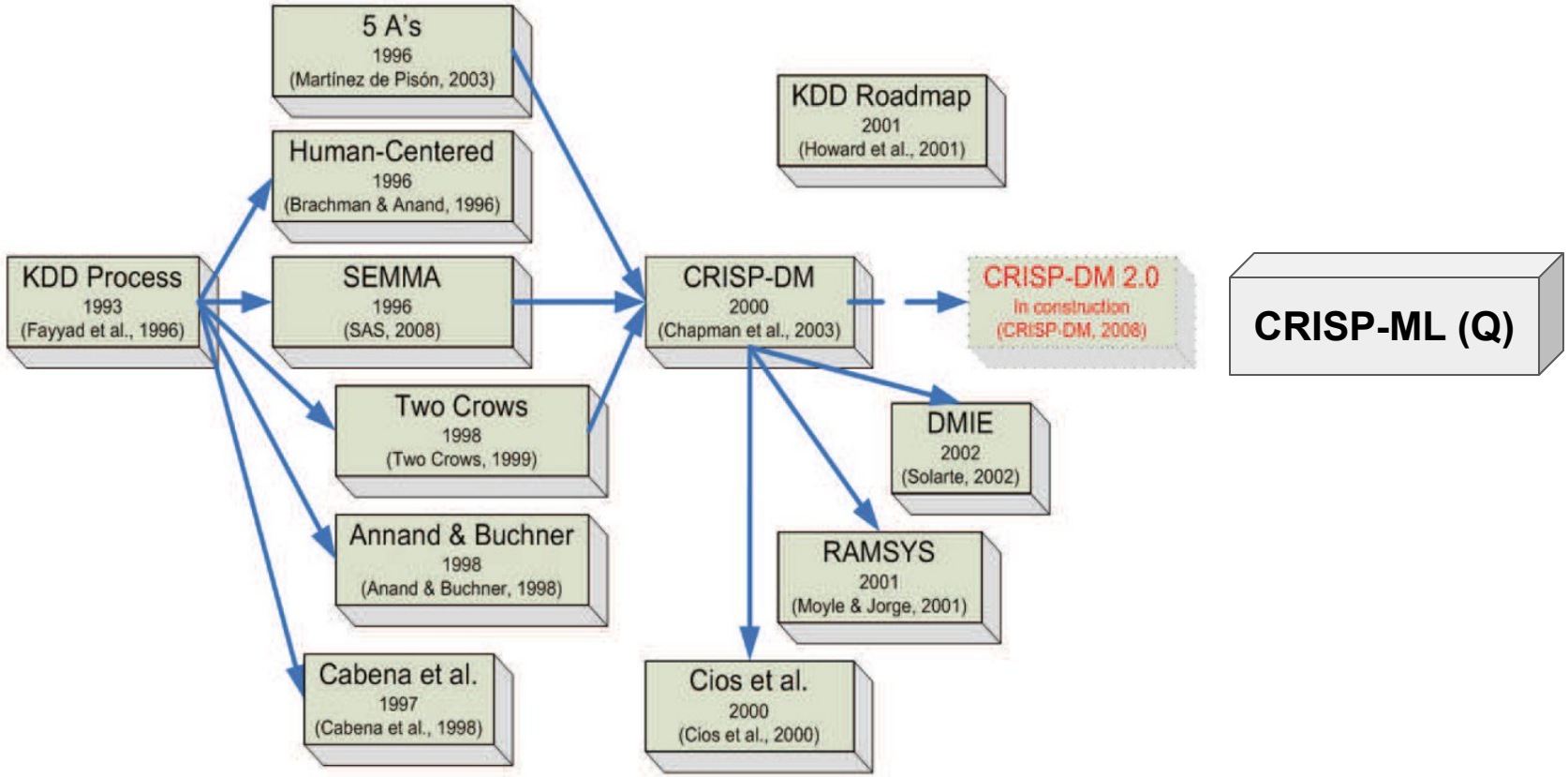
Foto: https://pixabay.com/pt/photos/sherlock-holmes-londres-turbulento-5499030/

Problemas de ML devem ser abordados por **processos investigativos iterativos**.

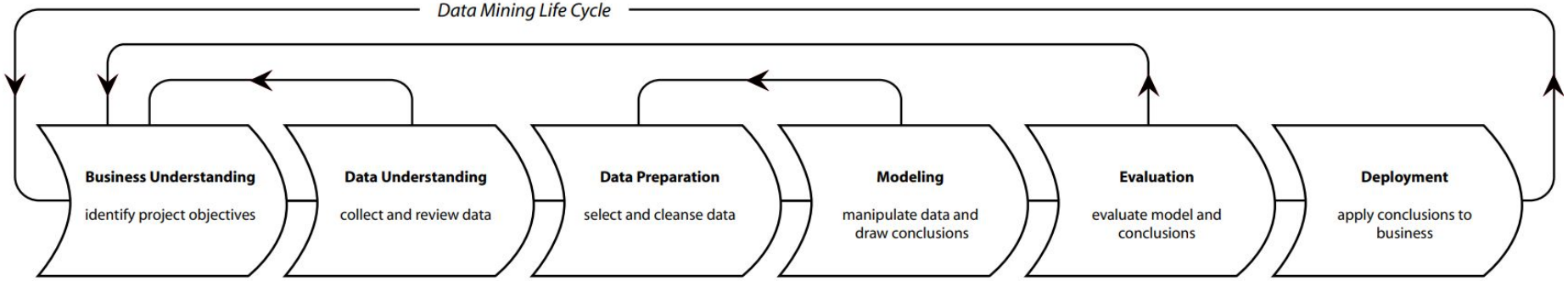
Foto: https://www.nationalmuseum.af.mil/Upcoming/Photos/igphoto/2000526757/

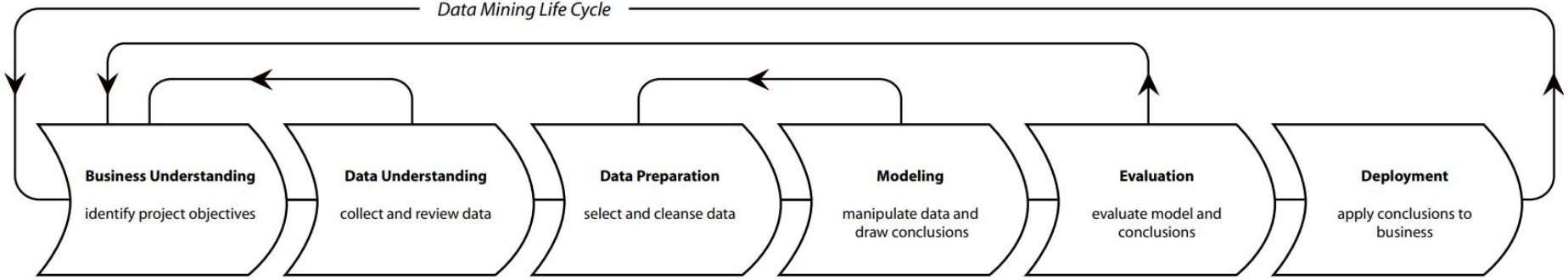
Foto: https://pixabay.com/pt/photos/sherlock-holmes-londres-turbulento-5499030/

Existem diversos frameworks para problemas com dados



Fonte: https://medium.com/data-hackers/project-data-science-pds-39d5a78e058a


Data Mining Life Cycle
Business Understanding
identify project objectives
Data Understanding
collect and review data
Data Preparation
select and cleanse data
Modeling
manipulate data and draw conclusions
Evaluation
evaluate model and conclusions
Deployment
apply conclusions to business



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

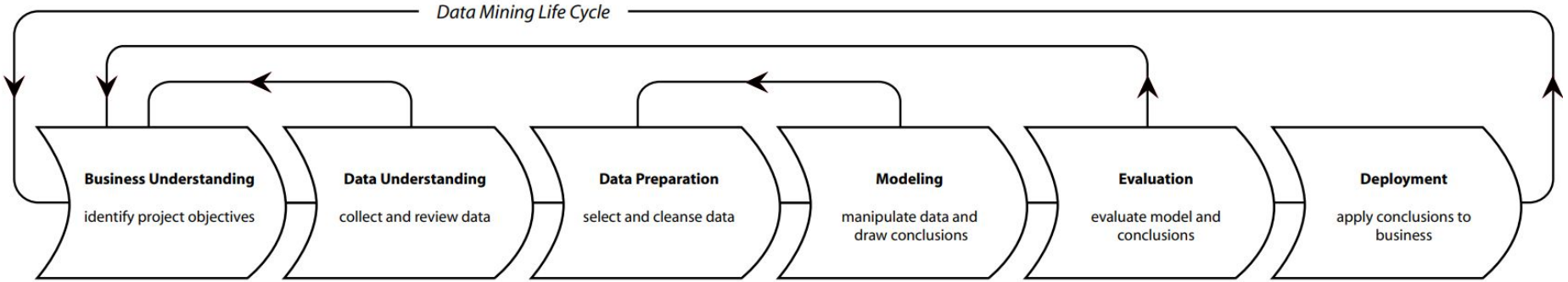
Review Project
Experience Documentation
(Log and Report Process)

Generic Tasks
Specialized Tasks
(Process Instances)

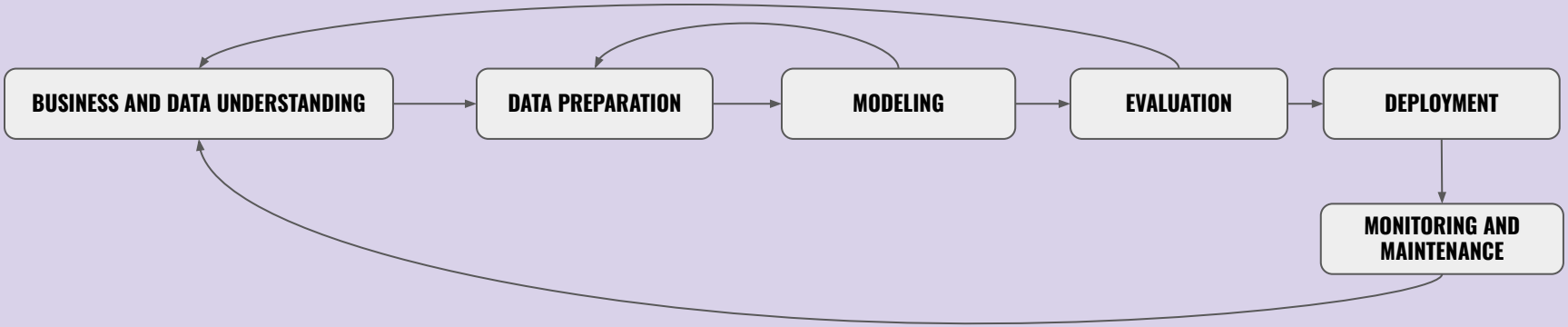
a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

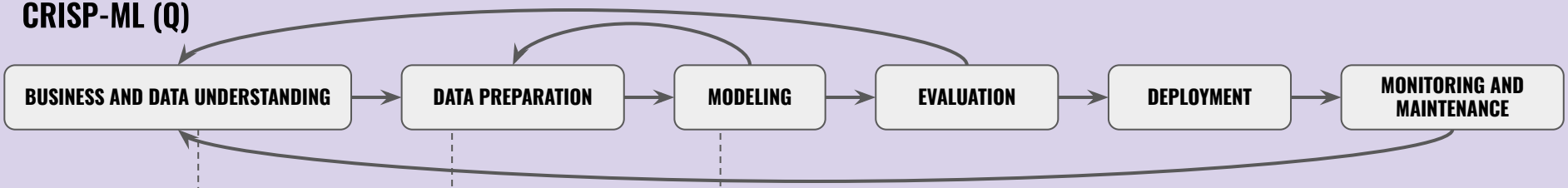
CRISP-DM



CRISP-ML (Q)



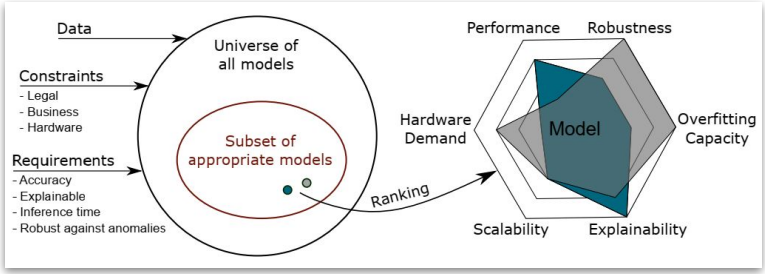
CRISP-ML (Q)



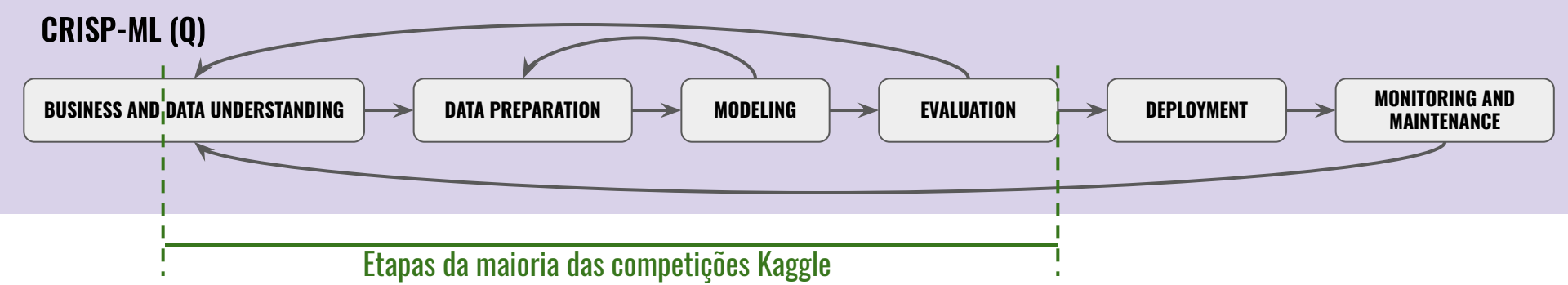
Understanding
- escopo do desenvolvimento
- critério de sucesso
- relatório com a verificação da qualidade dos dados.
- para o modelo: requisitos e condições de contorno do problema

Data preparation
- um *dataframe* para alimentar o modelo.
- usualmente, as etapas de preparação são:
  - select data;
  - clean data;
  - construct data (feature engineering e data augmentation);
  - standardize data (normalização, unidades).

Modeling
- Model selection (revisão da literatura, definição das métricas de avaliação + qualidade do modelo)
- Garantir reprodutibilidade



Performance	Performance metric on unseen data e.g. accuracy, AUROC, F1-score, mean square error (MSE), mean absolute error (MAE) etc.
Robustness	Resiliency of the model to inconsistent inputs e.g. adversarial attacks, out-of-distribution samples, anomalies and distribution shifts and to failures in the underlying execution environment e.g. sensor, actuators and computational platform.
Scalability	The property of the model to scale to high data volume during the training and re-training in the production system. Complexity analysis on the execution time and hardware demand dependent on the number of samples and feature dimension.
Explainability	Models could be either directly explainable or given by post-hoc explanations. The decisions of explainable models could be inspected manually and could increase the user acceptance. In addition, uncertainty and confidence estimates provide guidance on indecisive decisions.
Model Complexity	Models with large capacities overfit easily on small data sets. Assure that the capacity of your model suits the complexity of your data and use proper regularization.
Resource Demand	The model has to be deployed on hardware and is restricted by its memory. In addition, the inference time has to be considered dependent on the application.


CRISP-ML (Q)
BUSINESS AND DATA UNDERSTANDING
DATA PREPARATION
MODELING
EVALUATION
DEPLOYMENT
MONITORING AND MAINTENANCE
Etapas da maioria das competições Kaggle