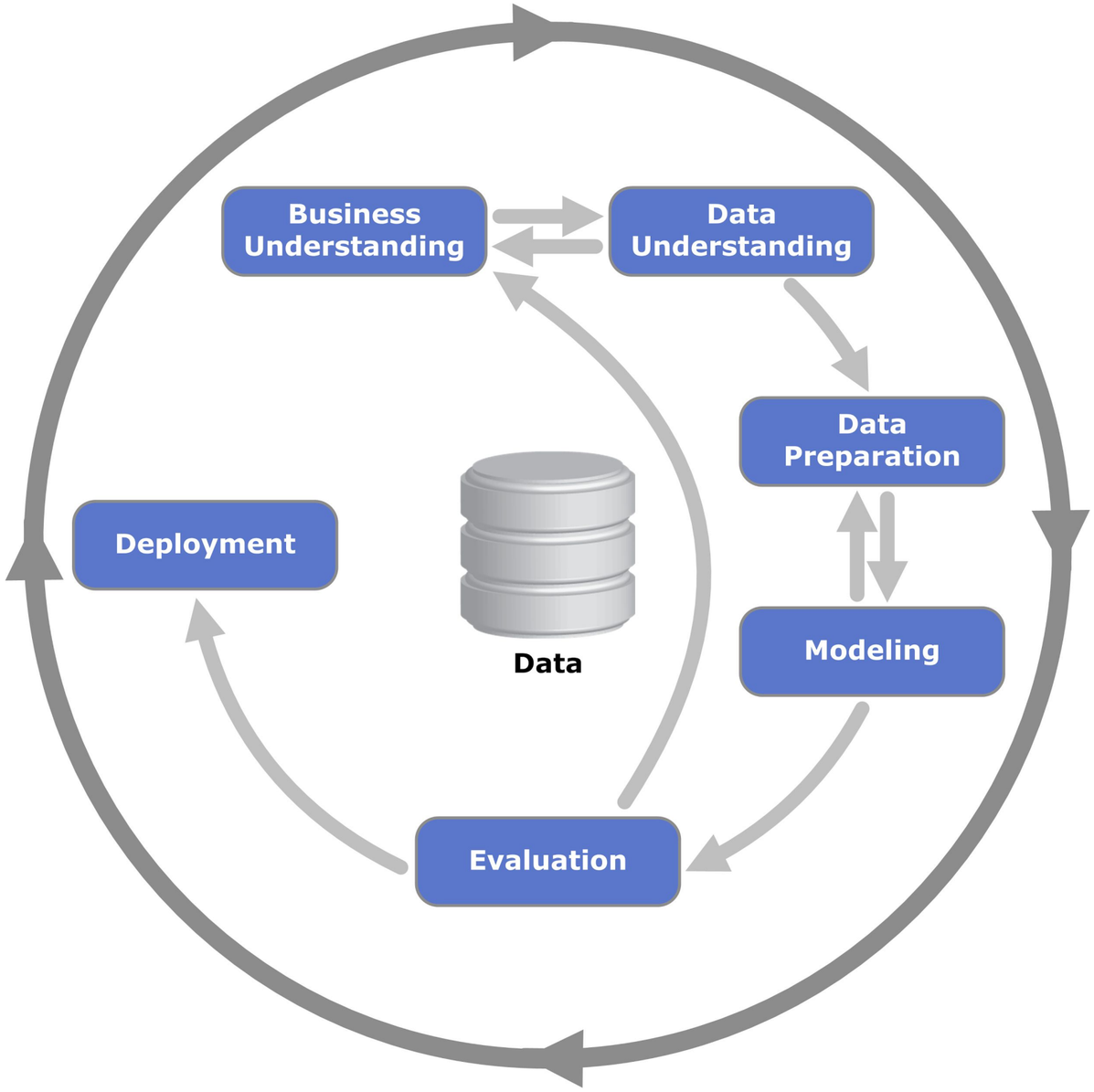
CDSF17 and CDSF18

Recap VI

DS Academy

Data Science Process

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Project definition

- Mapping the Problem
  - Do we have a problem?
  - What is the problem?
    - define the key business question you hope to solve
  - What is the impact?
  - How do you solve this today?
  - What data is available?

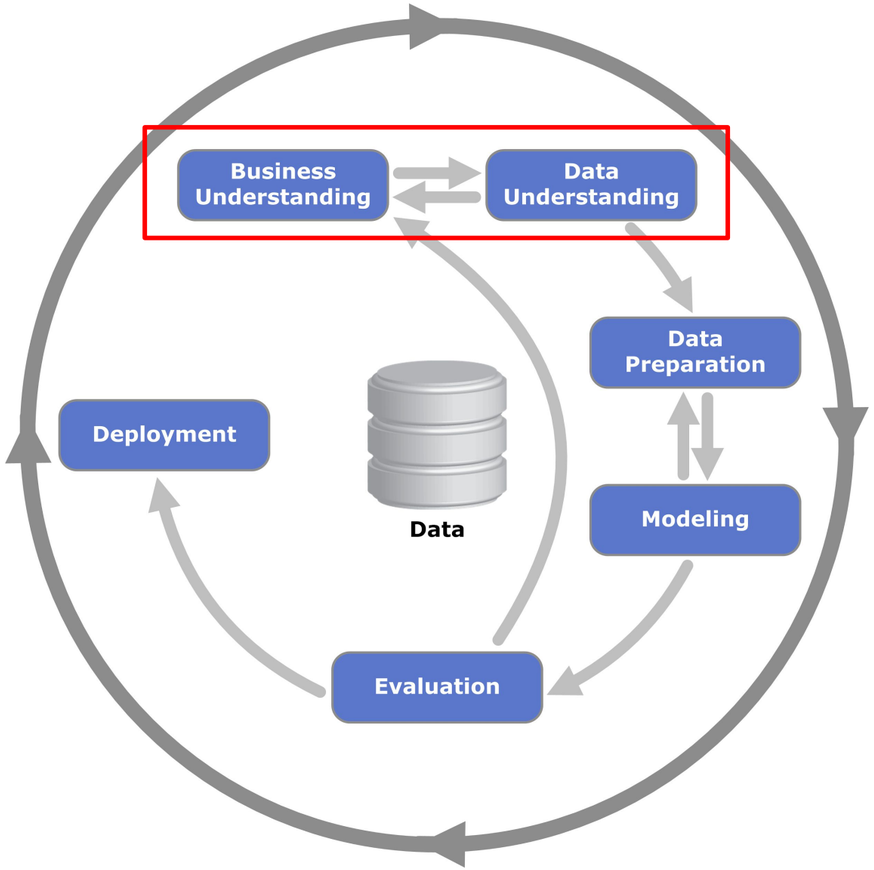
Project definition

- Mapping the Problem
  - Do we have a problem?
  - What is the problem?
    - define the key business question you hope to solve
  - What is the impact?
  - How do you solve this today?
  - What data is available?

- Example: The churn problem.
  - What is the definition of churn for you?
  - “When a customer is likely to churn” is very different from “Why are my customers churning?”
  - How much does a churn cost?
  - Is there any retention action for a possible churn?
    - How does it cost to retain a client?
    - How much time you spend on doing that?

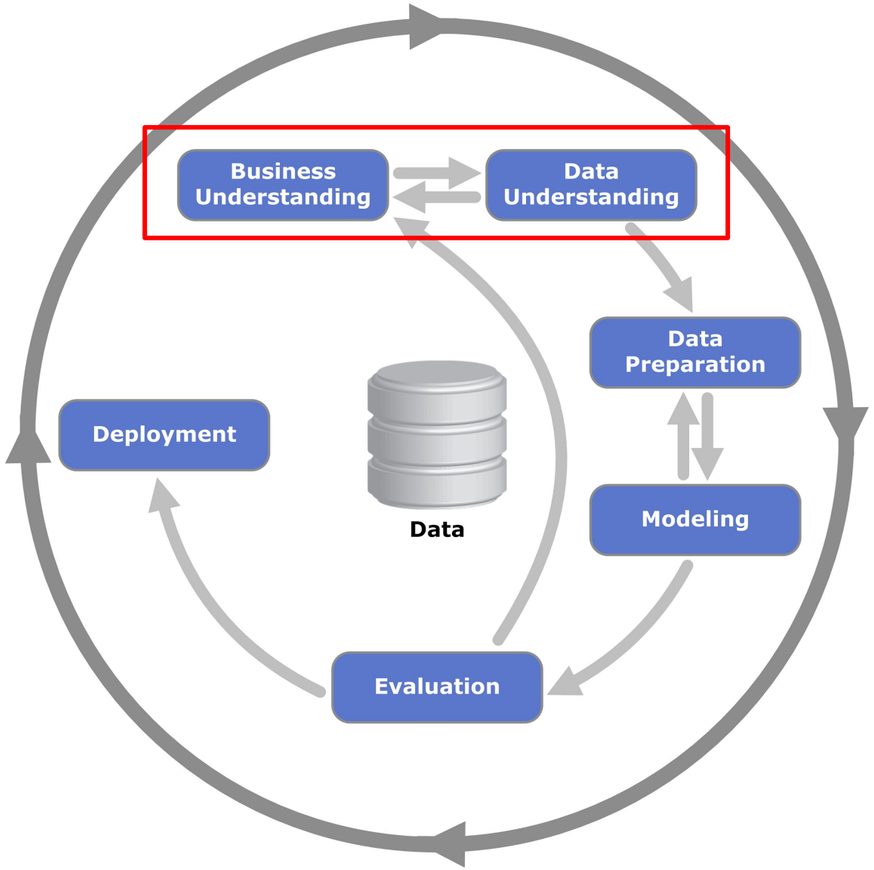
Exploratory Data Analysis (EDA)

- What data is available
  - Structured data?
  - How it aligns to the business problem
  - Who is the person that knows this data?



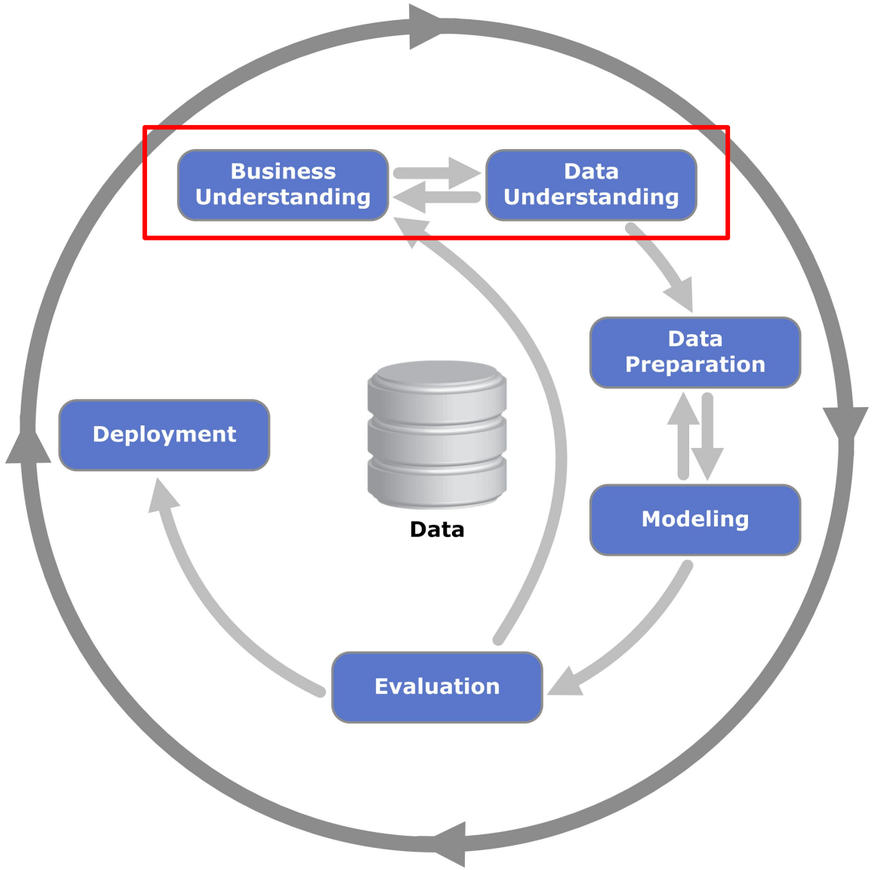
Exploratory Data Analysis (EDA)

- How much data is available?
  - The more, the better?
    - You need have sufficient data
  - Was there any change in the system, e.g., migration?



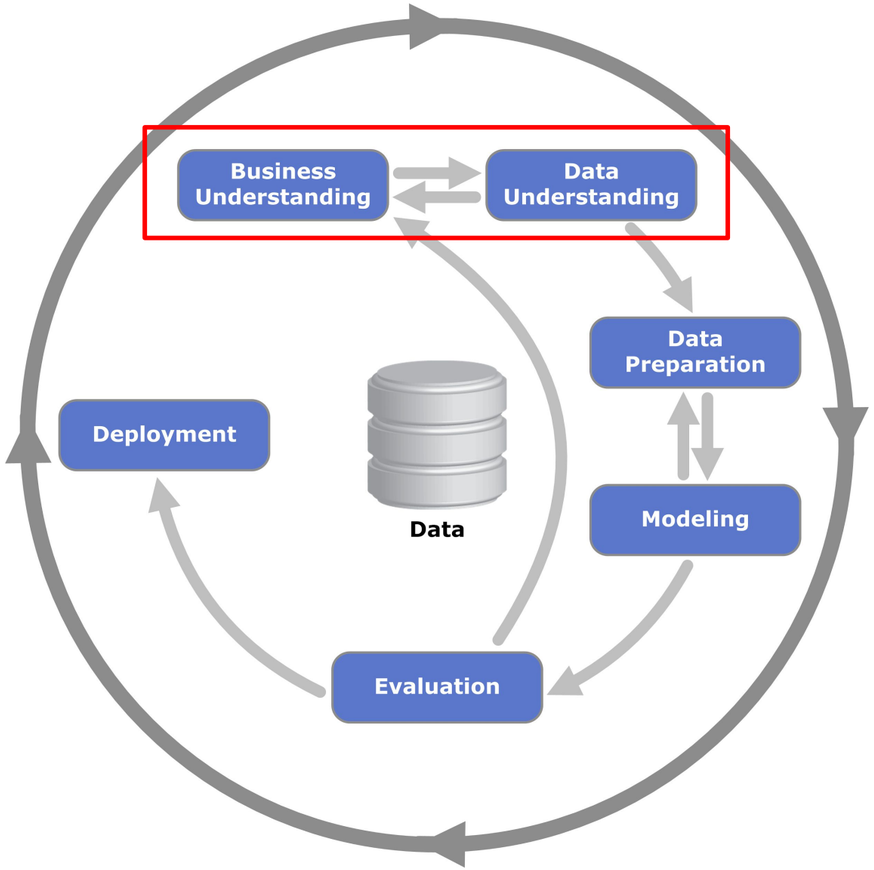
Exploratory Data Analysis (EDA)

- What format will the data be in and where does it reside?
  - What are the datatypes?
    - CSV, JSON, PARQUET?
  - Data Base, API, DUMP?
    - How can the data be accessed?
  - How do the multiple data sources get joined?



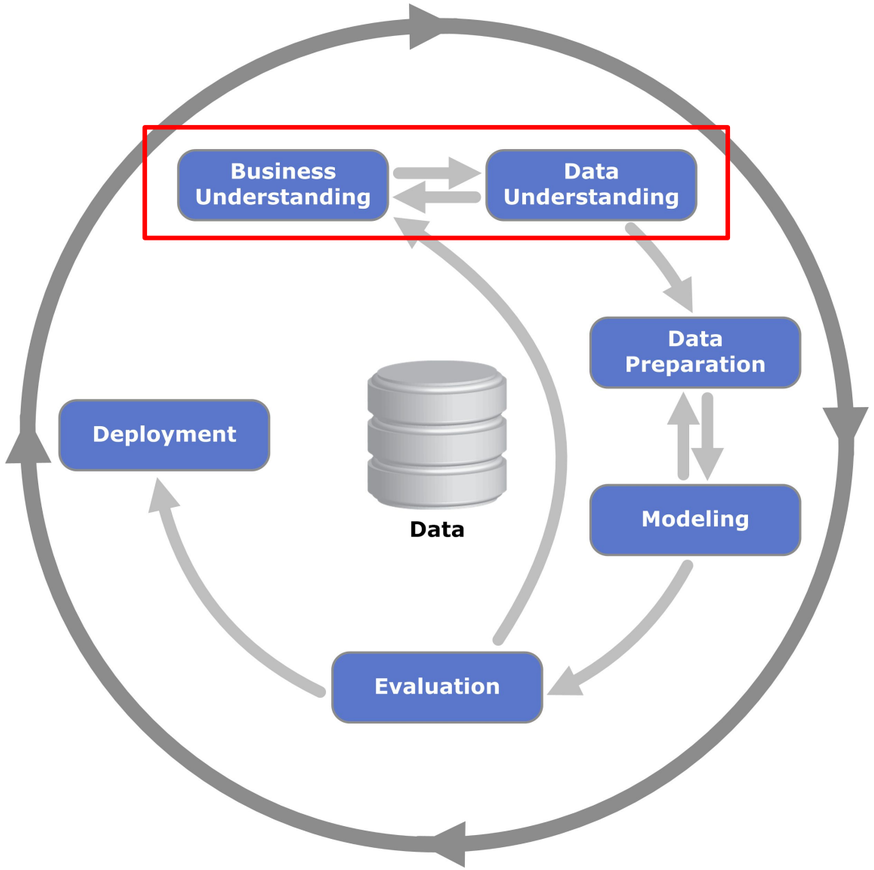
Exploratory Data Analysis (EDA)

- Do you have access to the target?
  - How can I build the target?
  - Validated if your data is balanced
  - Do we have metrics related to the target?
  - Correlation between target and variables.



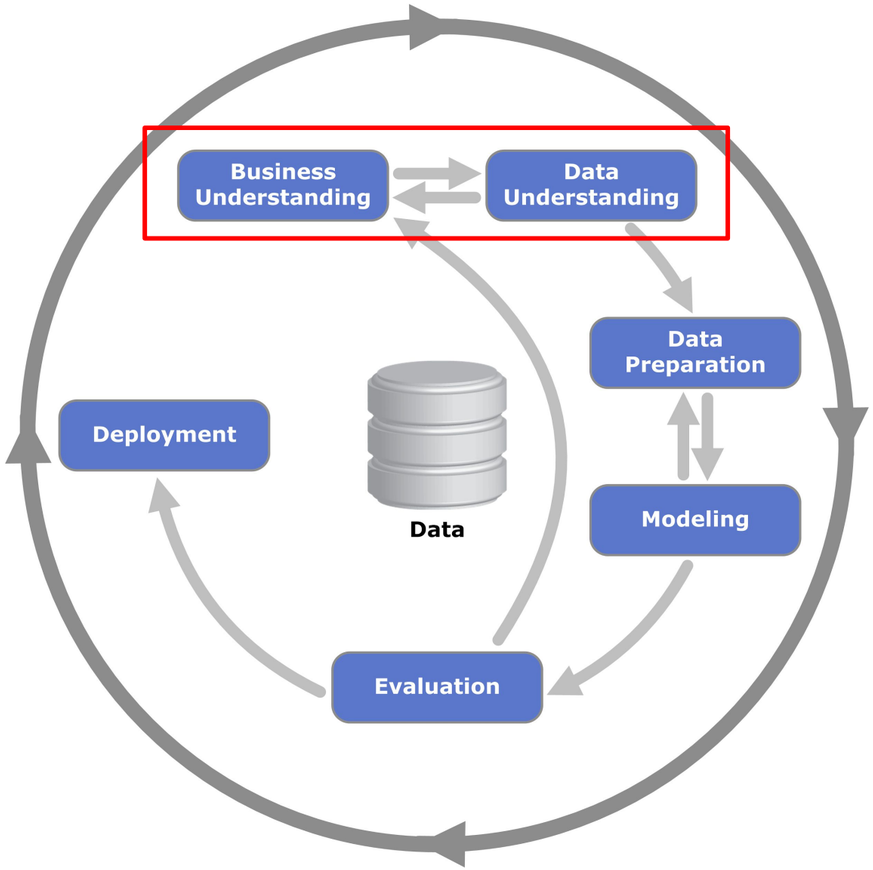
Exploratory Data Analysis (EDA)

- Which fields are most important?
  - What important metrics are reported using this data?
  - Check for Outliers
  - Check for missing values
  - Is there bias in your data
  - Create visualizations



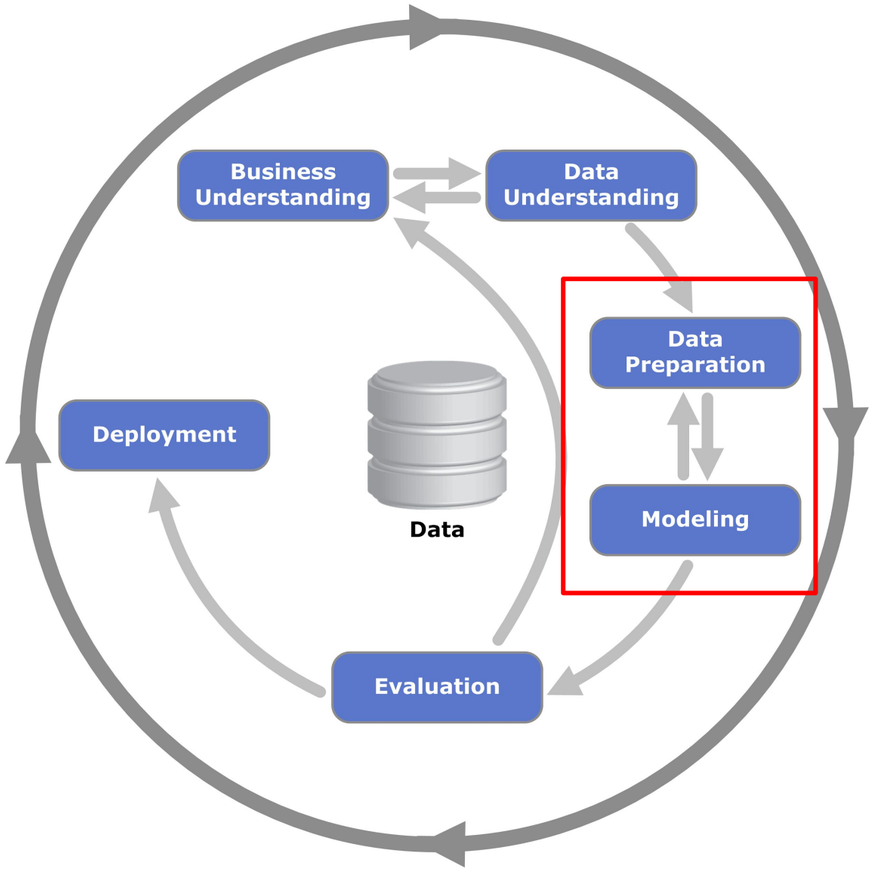
Exploratory Data Analysis (EDA)

- Real data is messy
- Document your findings.



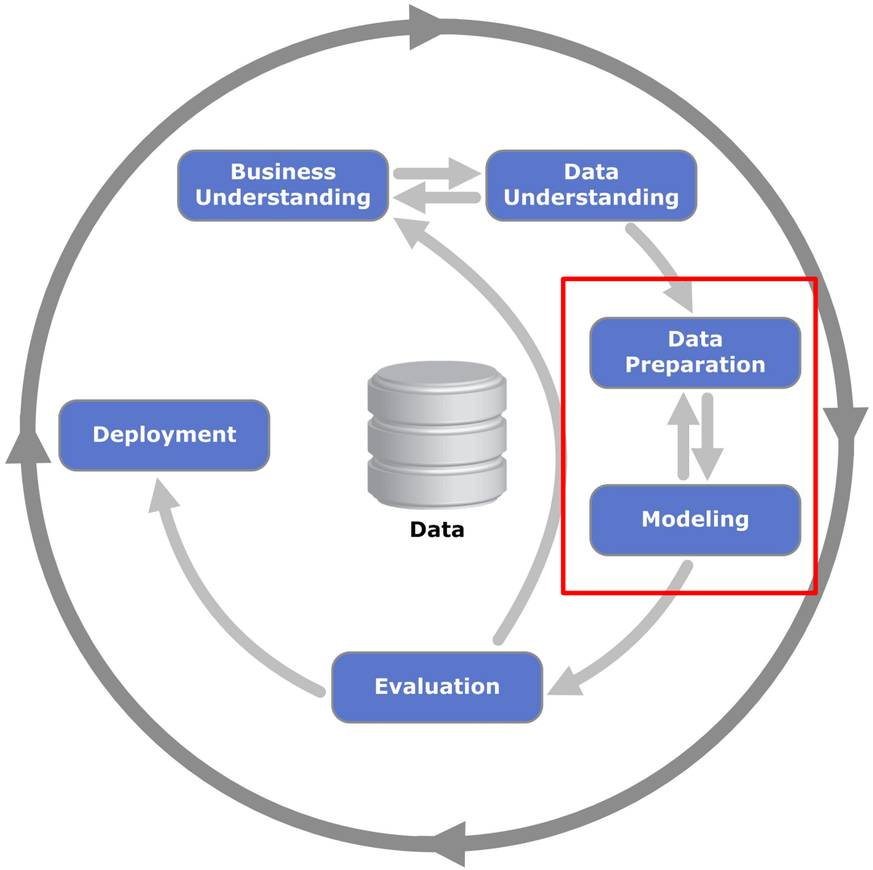
Data Prep. and Modeling

- Keep a set of data out for testing!!
- Create your validation strategy
  - Always left a part of the date out for testing.
  - Remember, you cannot compare apple with oranges.
    - Your validation set should be always the same for different models.
    - The experiments must be as similar as possible.



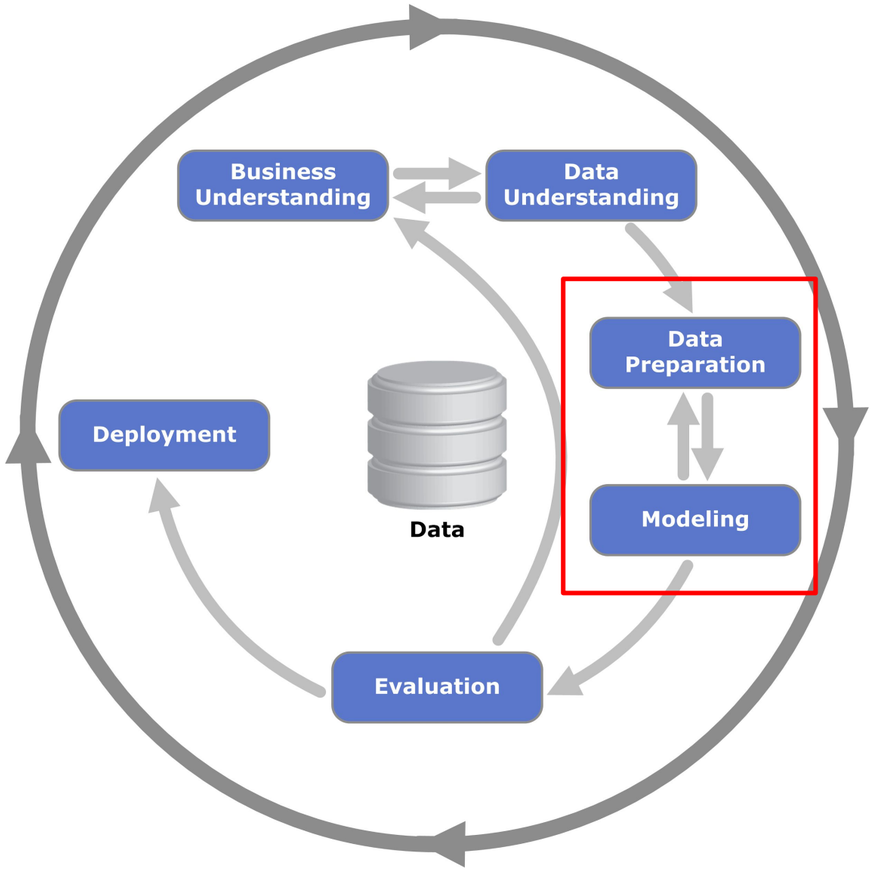
Data Prep. and Modeling

- Create your Baseline
  - The simplest model you can think
  - If there exists a process solving this, this could be the baseline.
  - Our goal is to beat the baseline.



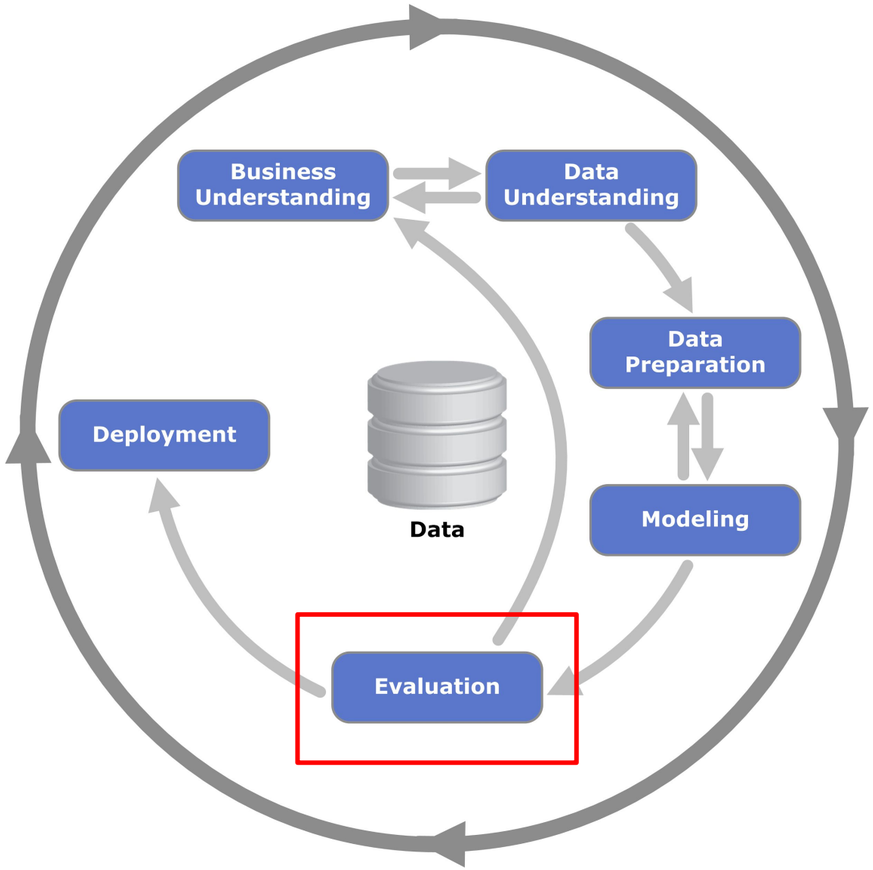
Data Prep. and Modeling

- Research and development.
  - What algorithm to use?
  - Is there any market solution that is doing the same?
    - What do they use?
  - Find papers about the subject.
  - Find the best framework to use



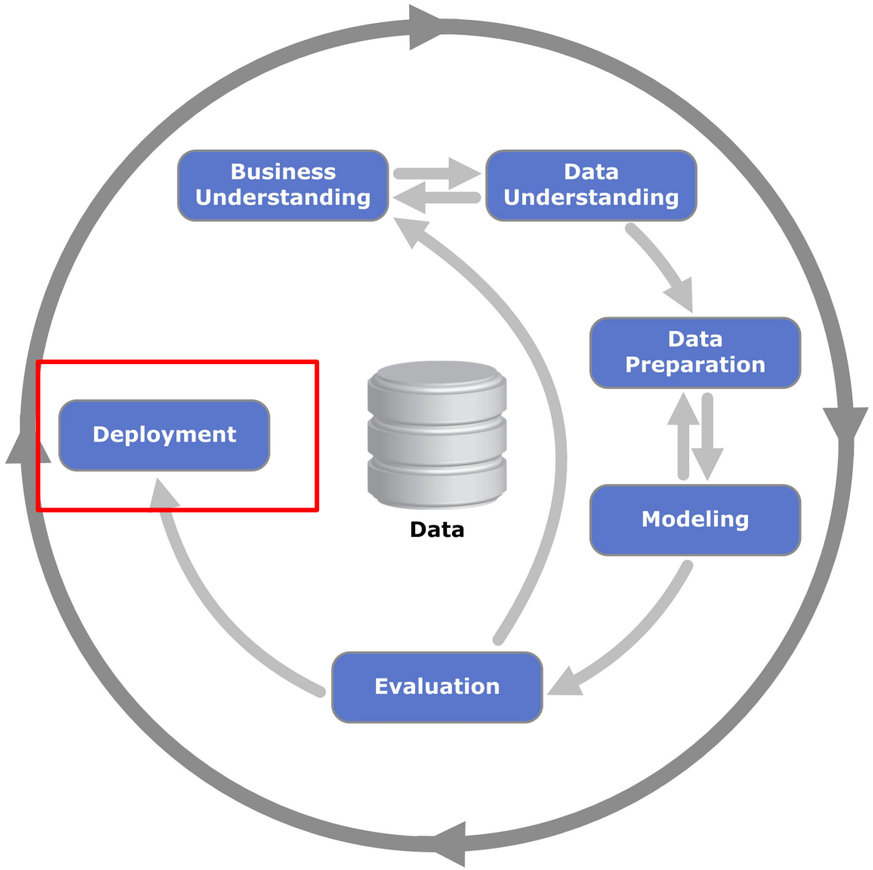
Evaluation

- At the beginning, create your model and add features little by little
  - Check the importance of the features
  - Does it make sense what you are seeing?
- If the model is too good, something is wrong!
  - check for leaks
- Check where you are getting it wrong
- Validate, validate and validate.



Deployment

- Are we ready?
  - How good is my model in my test data?
  - How will we validate the results when in production
  - Will my code scale?
  - How often will I retrain the model?
  - How will I serve the model?
  - Do I have any technical dept?
- Monitoring after deploy.
  - How good is the model in production?
  - How often do I need to revisit it?



Master in Carol

Master in Carol

- How to use Carol to build and deploy your ML app.
- TOTVS university will release the videos in the next days
- People that participate in Carol Data Science Foundation have priority to use the platform this year

Useful Resources

Useful Resources

Technical Debt:
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

THANK YOU

totvs.com

company/totvs

@totvs

fluig.com

Tecnologia + Conhecimento são nosso DNA.

O sucesso do cliente é o nosso sucesso.

Valorizamos gente boa que é boa gente.

#SOMOSTOTVERS