Machine Learning Zoomcamp

Session #1.4

CRISP-DM ML Process

DataTalks.Club

Session #1.4: Plan

- CRISP-DM — methodology for organizing ML projects

Session #1.4: Plan

- CRISP-DM — methodology for organizing ML projects
- From problem understanding to deployment

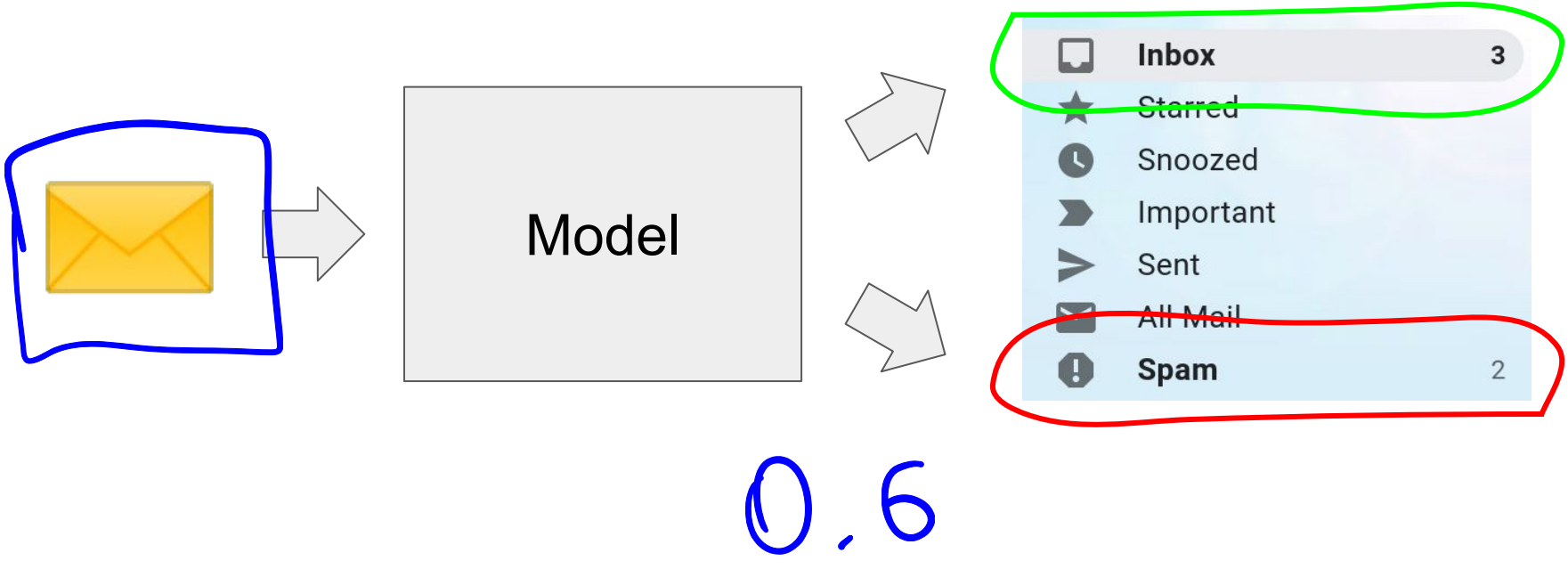
Session #1.4: Plan

- CRISP-DM
- From problem understanding to deployment
- Spam detection example

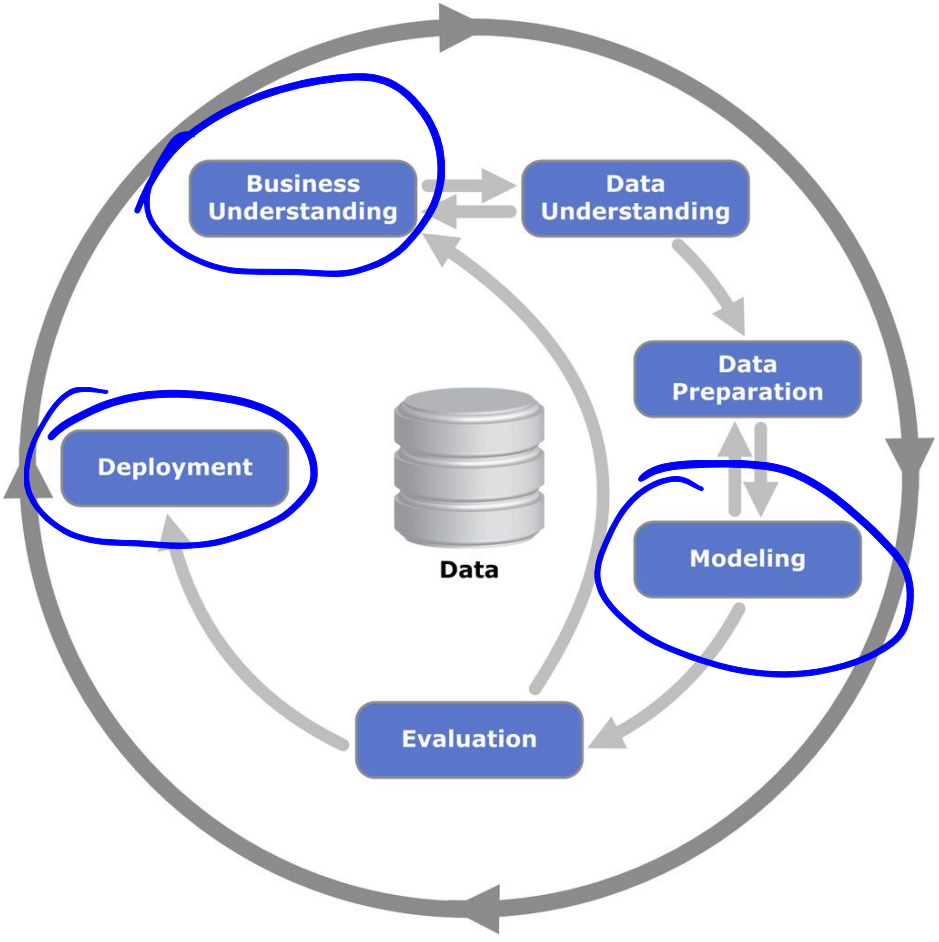
ML Projects

- Understand the problem
- Collect the data
- Train the model
- Use it

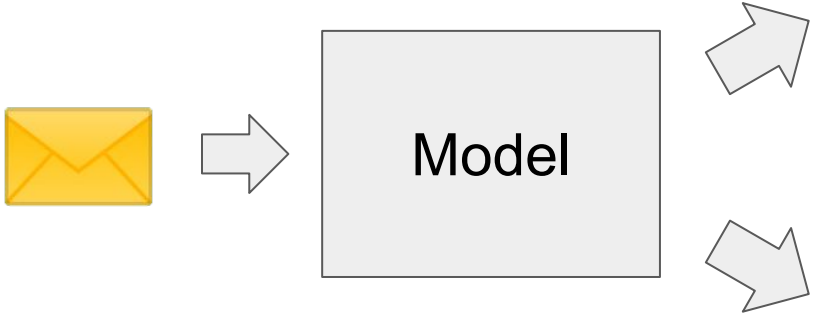
Spam detection

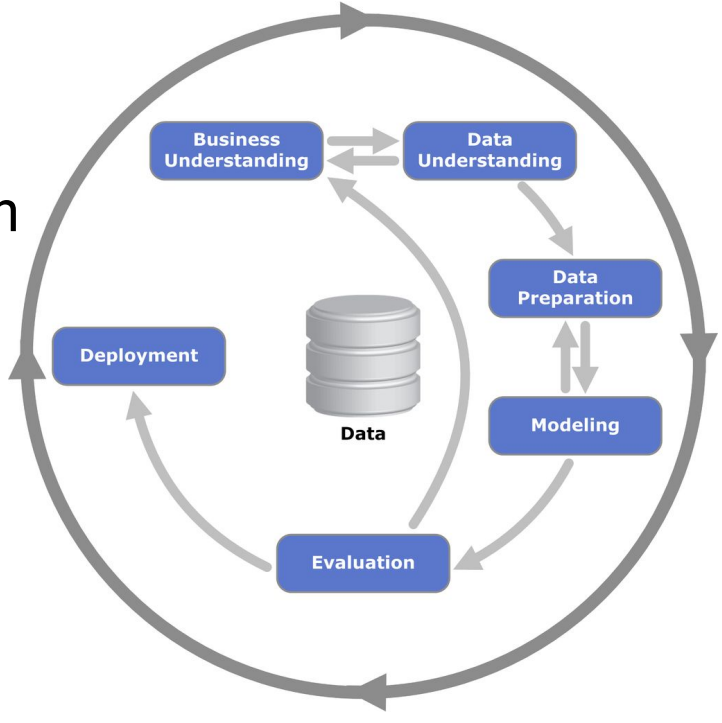


CRISP-DM

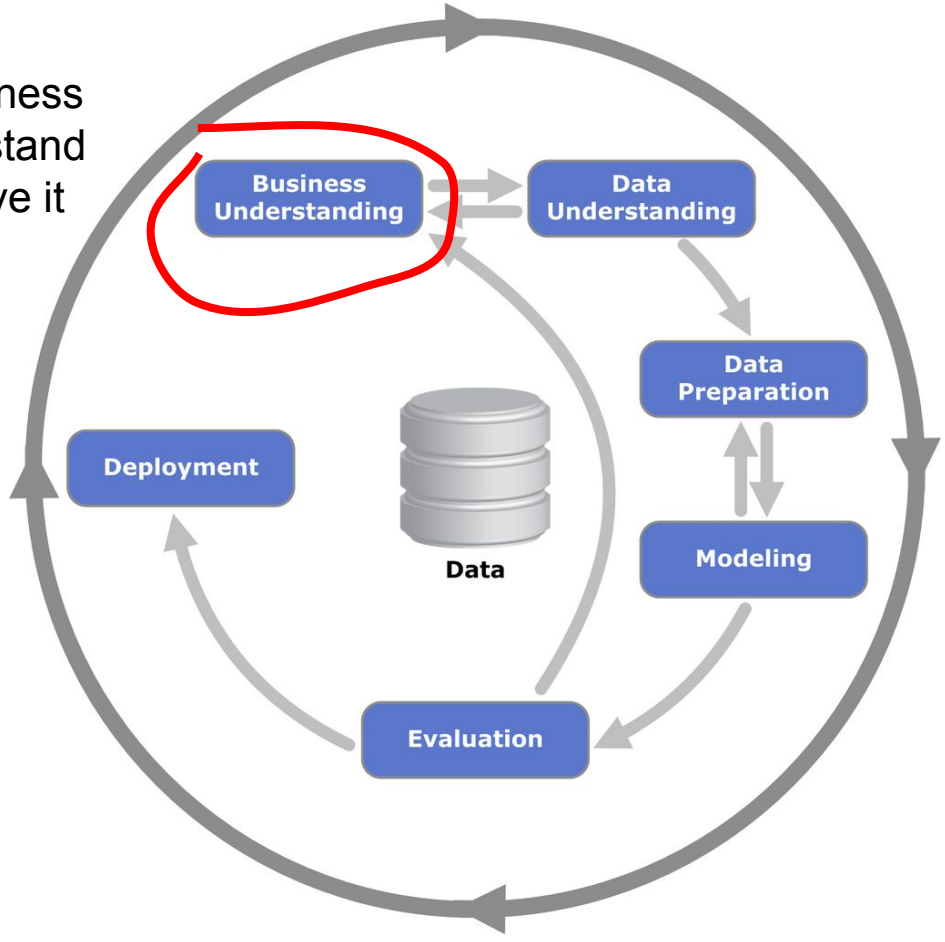


Picture: CRISP-DM


Model
spam
not spam

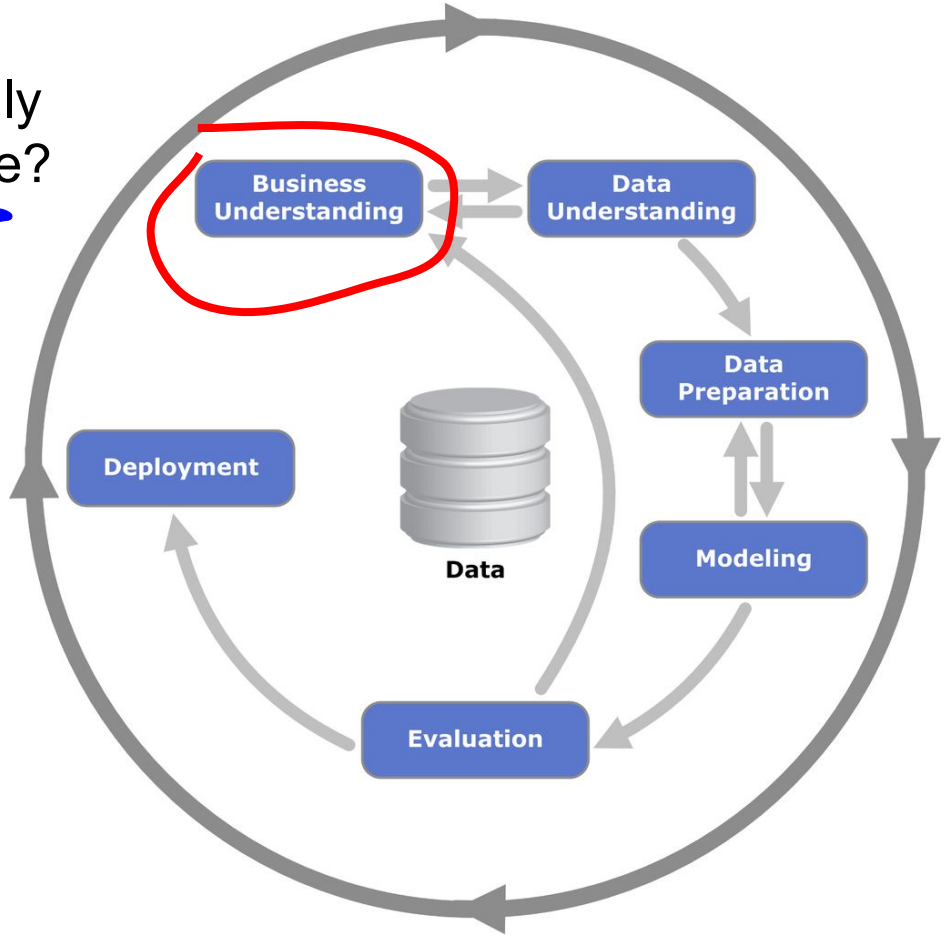

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Identify the business problem, understand how we can solve it



Picture: CRISP-DM

Do we actually need ML here?



Picture: CRISP-DM

Business understanding

- Our users complain about spam
- Analyze to what extent it's a problem
- Will Machine Learning help?
- If not: propose an alternative solution

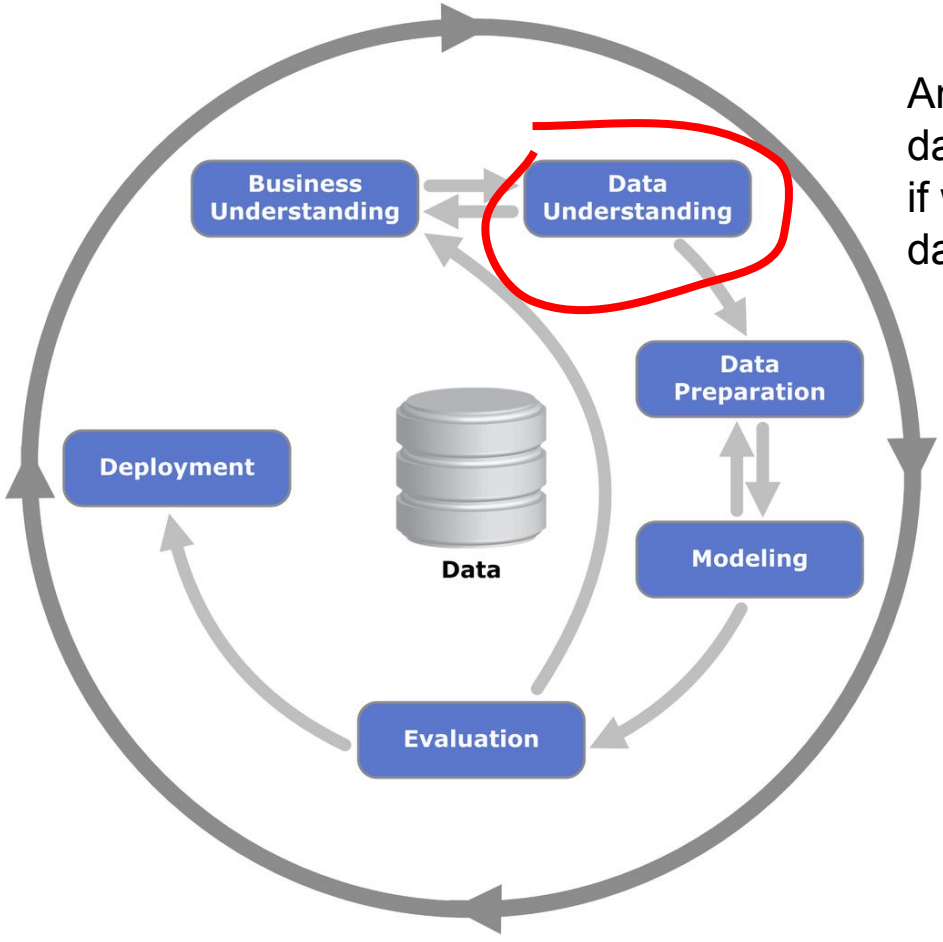
Business understanding

Define the goal:

- Reduce the amount of spam messages, or
- Reduce the amount of complaints about spam

The goal has to be measurable

- Reduce the amount of spam by 50%



Analyze available data sources, decide if we need to get more data

Picture: CRISP-DM

Data understanding

Identify the data sources

- We have a report spam button
- Is the data behind this button good enough?
- Is it reliable?
- Do we track it correctly?
- Is the dataset large enough?
- Do we need to get more data?

Data understanding

Identify the data sources

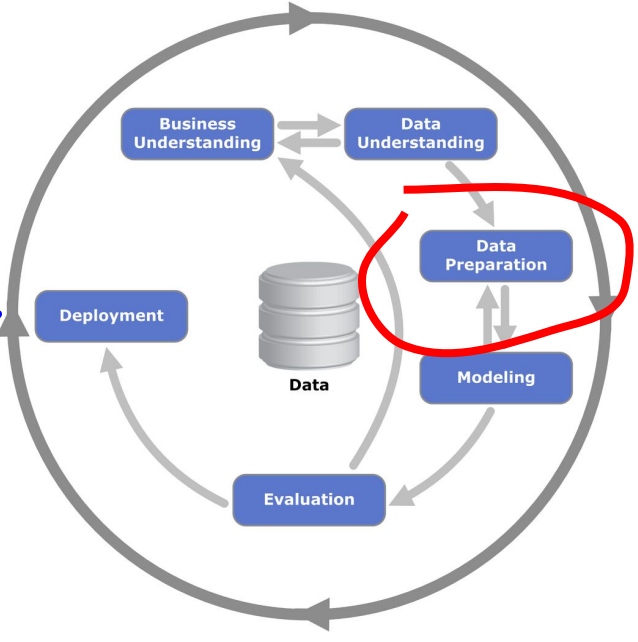
- It may influence the goal
- We may go back to the previous step and adjust it

Transform the data so it can be put into a ML algorithm

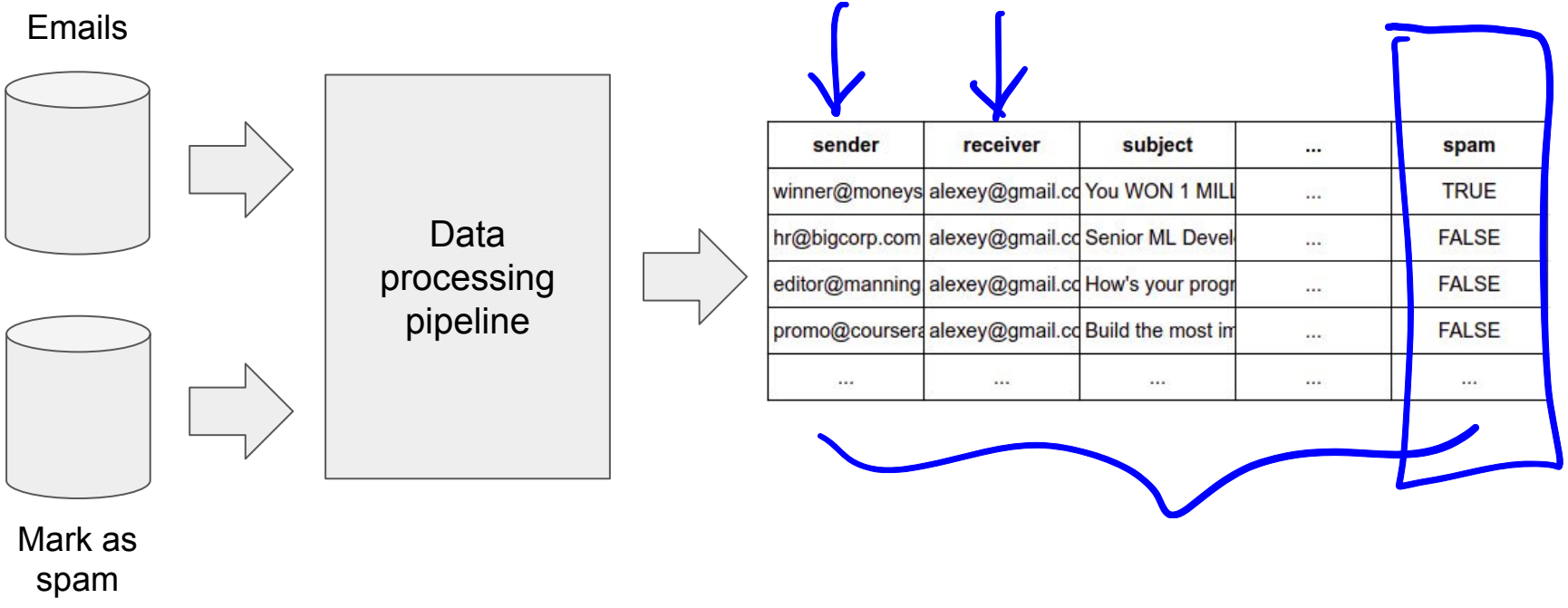
Picture: CRISP-DM

Data preparation

- Clean the data
- Build the pipelines
- Convert into tabular form



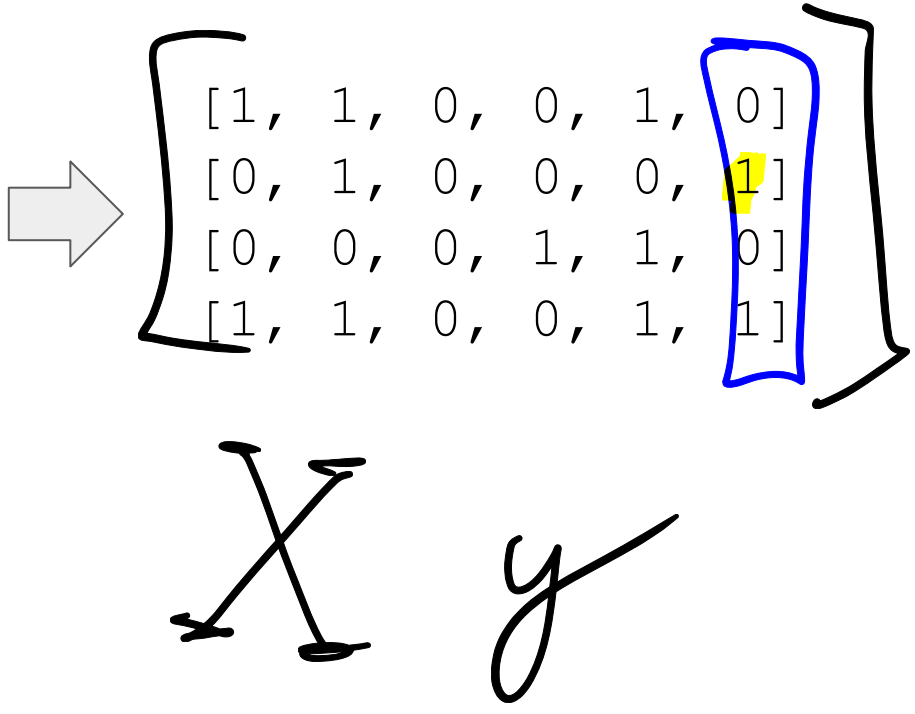
Data preparation

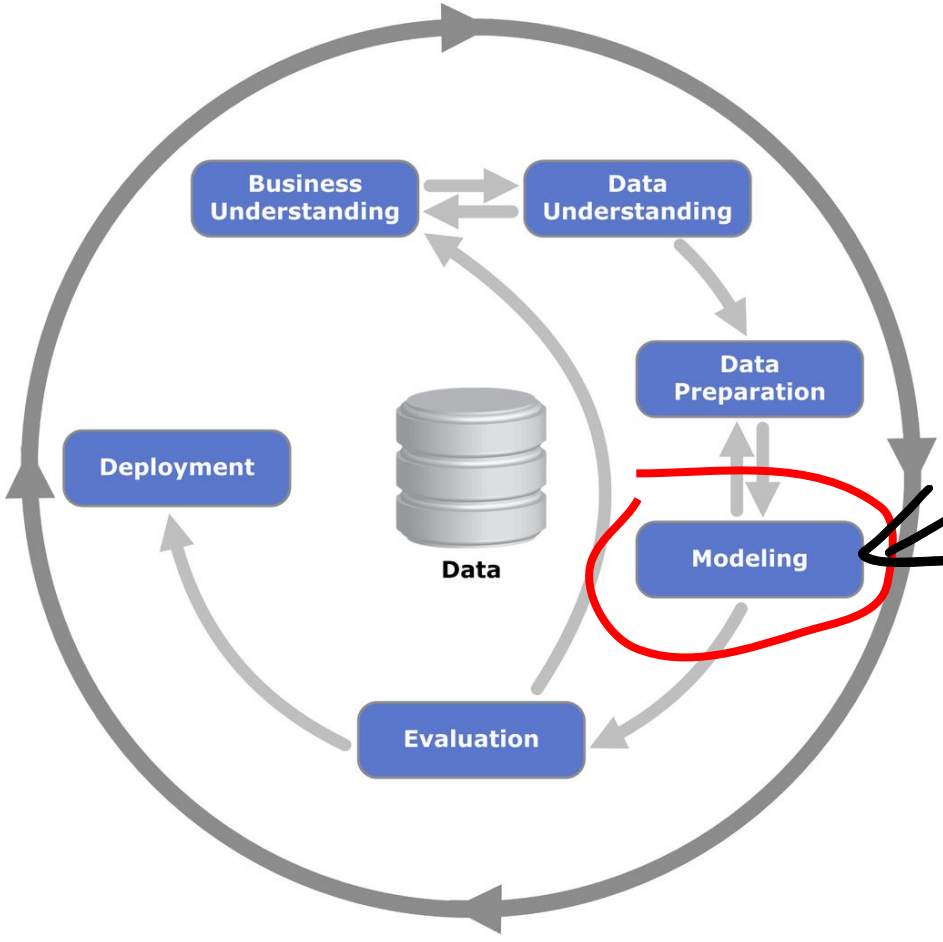


Subject: You won 1 MILLION!
From: winner@moneys.com

Congratulations! You've won $1,000,000!
In order to access the money, deposit $100 to XXXXXX

Yours sincerely,
Moneyball





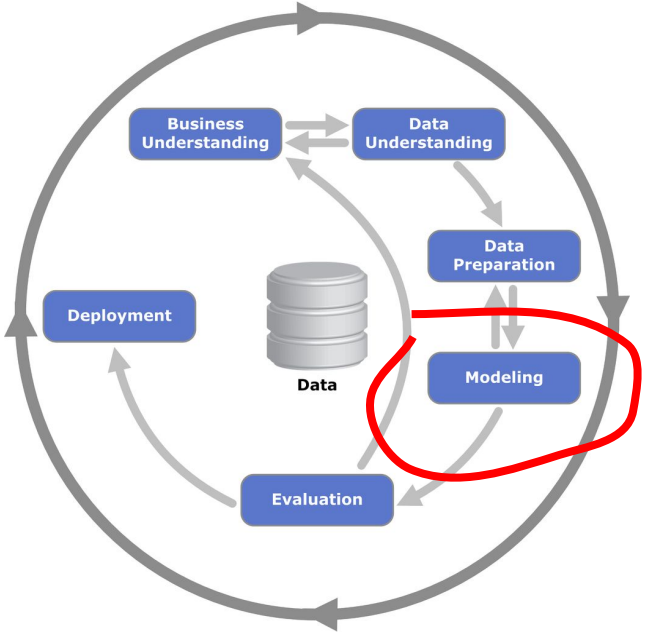
Training the models: the actual Machine Learning happens here

Picture: CRISP-DM

Modeling

Training a model:

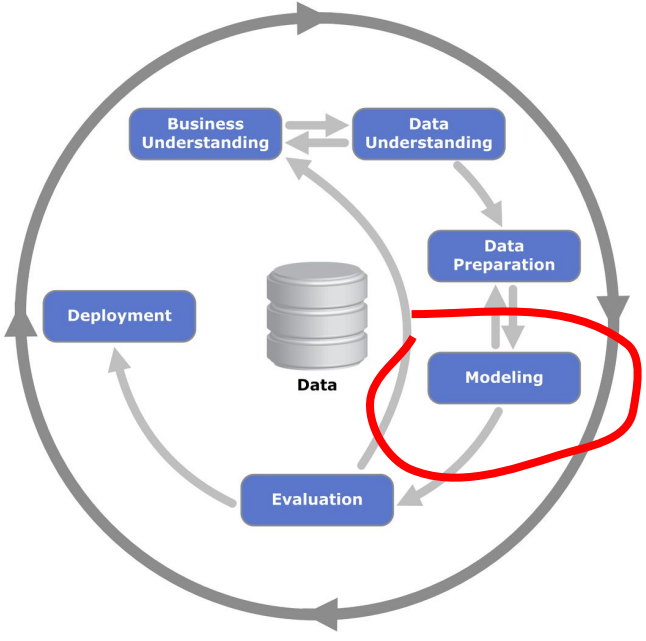
- Try different models
- Select the best one



Modeling

Which model to choose?

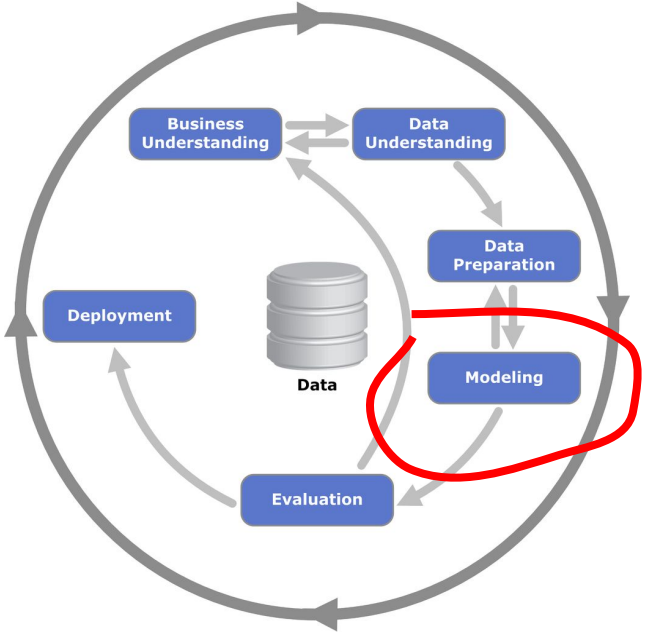
- Logistic regression
- Decision tree
- Neural network
- Or many others

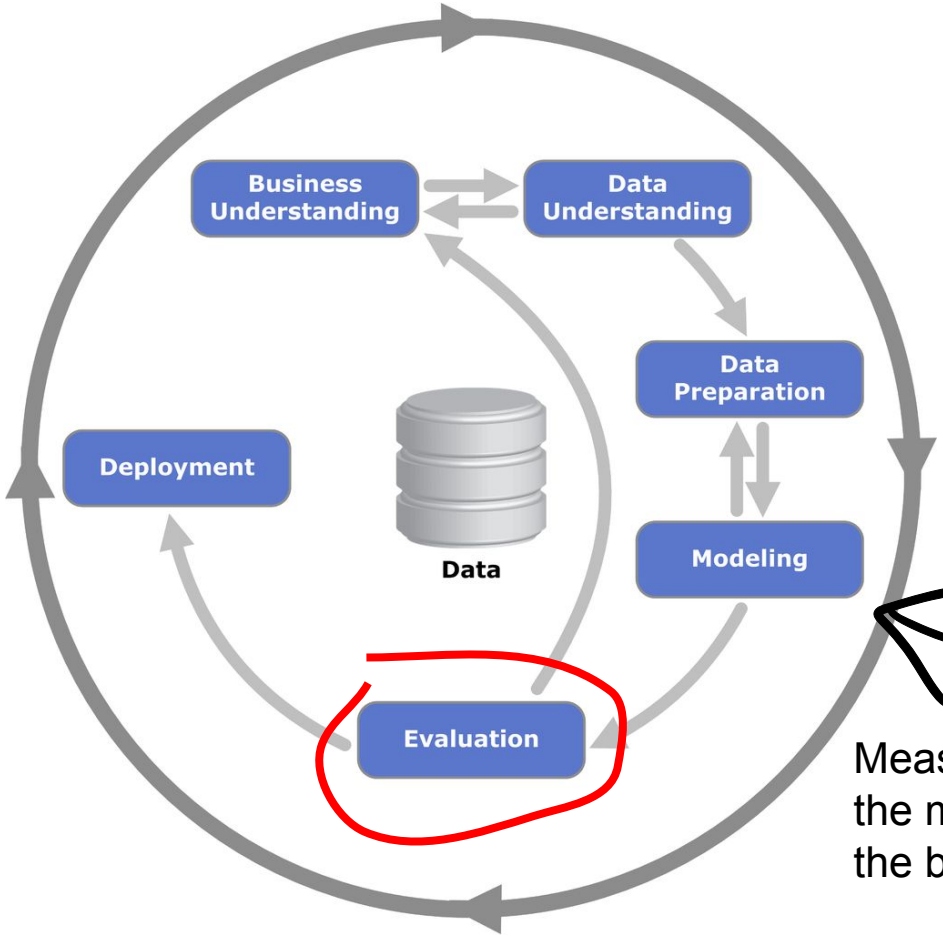


Modeling

Sometimes, we may go back to data preparation:

- Add new features
- Fix data issues





Measure how well
the model solves
the business problem

Picture: CRISP-DM

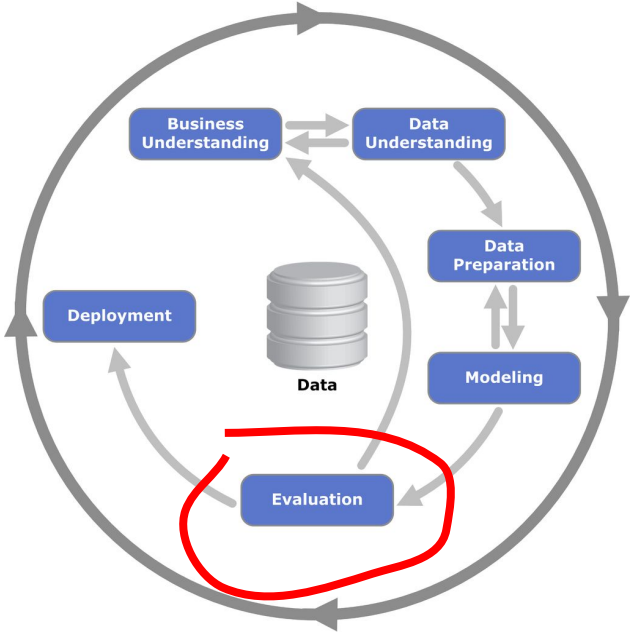
Evaluation

Is the model good enough?

- Have we reached the goal?
- Do our metrics improve?

Goal: Reduce the amount of spam by ~~50%~~ 30

- Have we reduced it? By how much?
- (Evaluate on the test group)



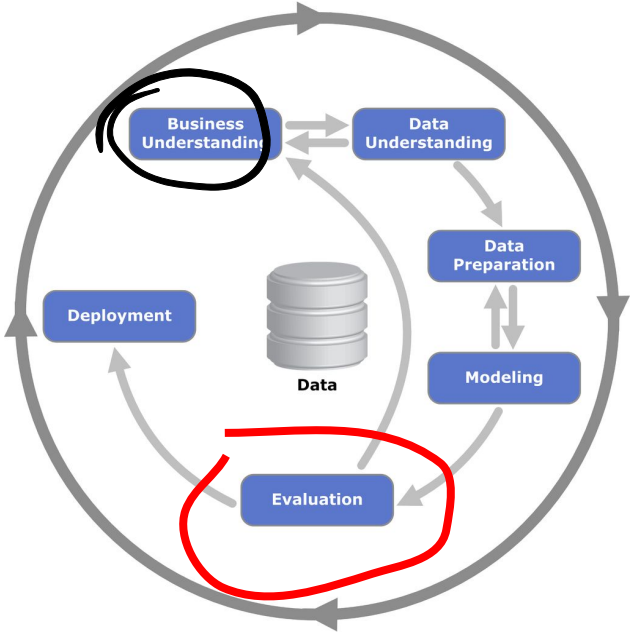
Evaluation

Do a retrospective:

- Was the goal achievable?
- Did we solve/measure the right thing?

After that, we may decide to:

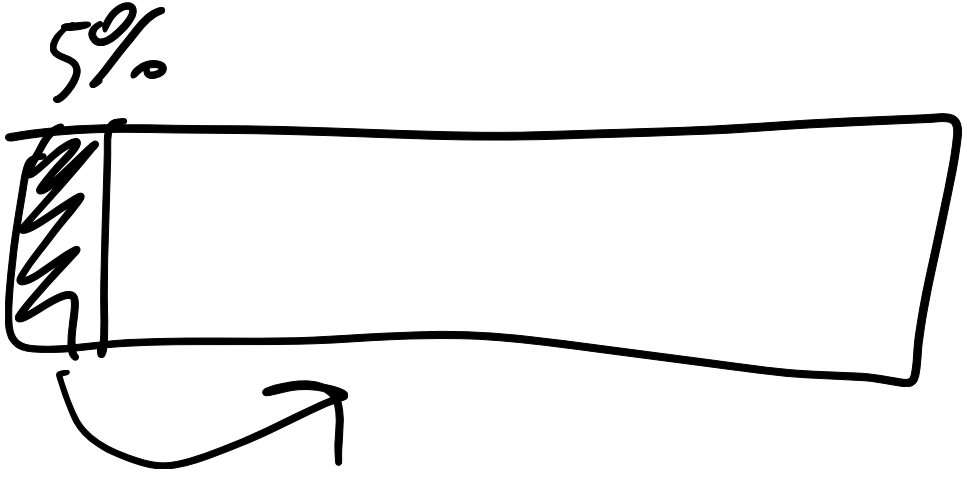
- Go back and adjust the goal
- Roll the model to more users/all users
- Stop working on the project

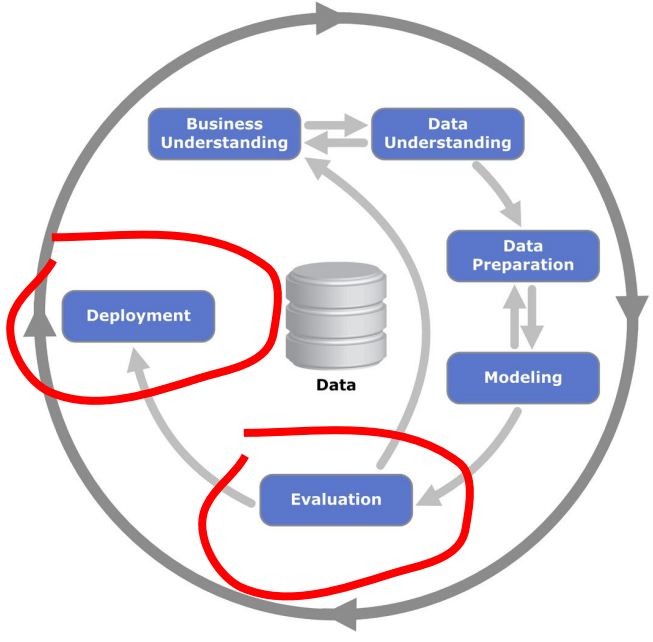


Evaluation + Deployment

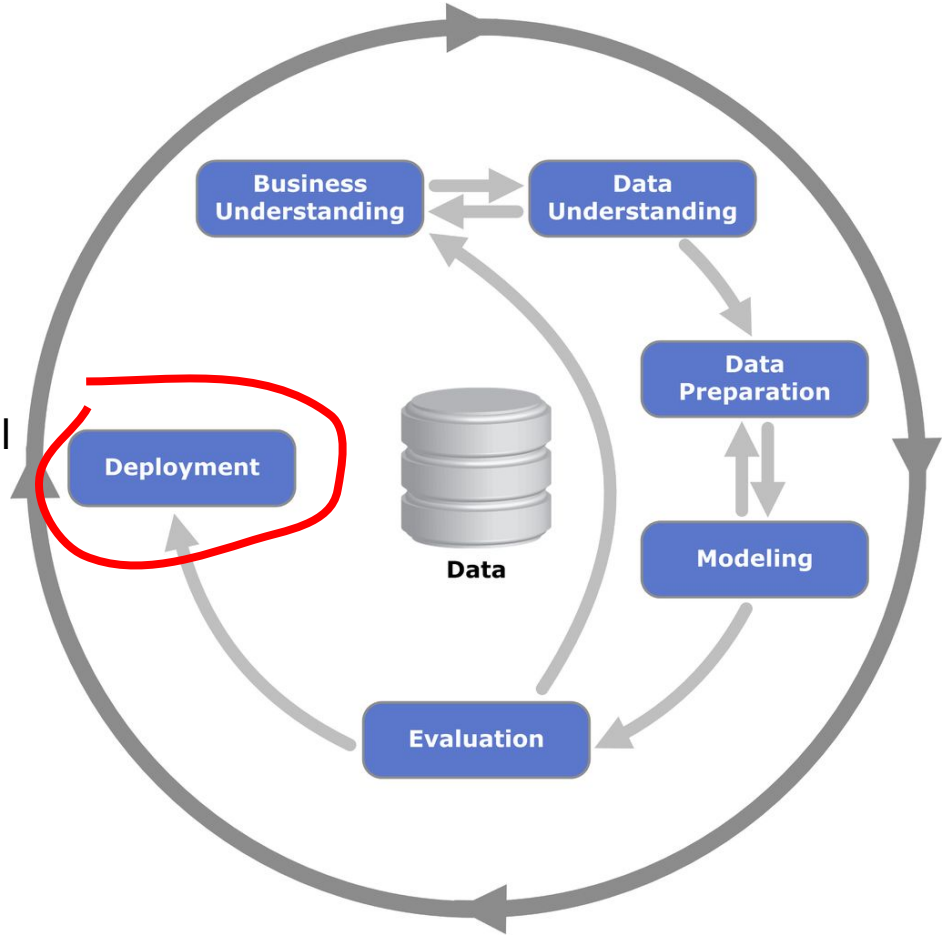
Often happens together:

- Online evaluation: evaluation of live users
- It means: deploy the model, evaluate it





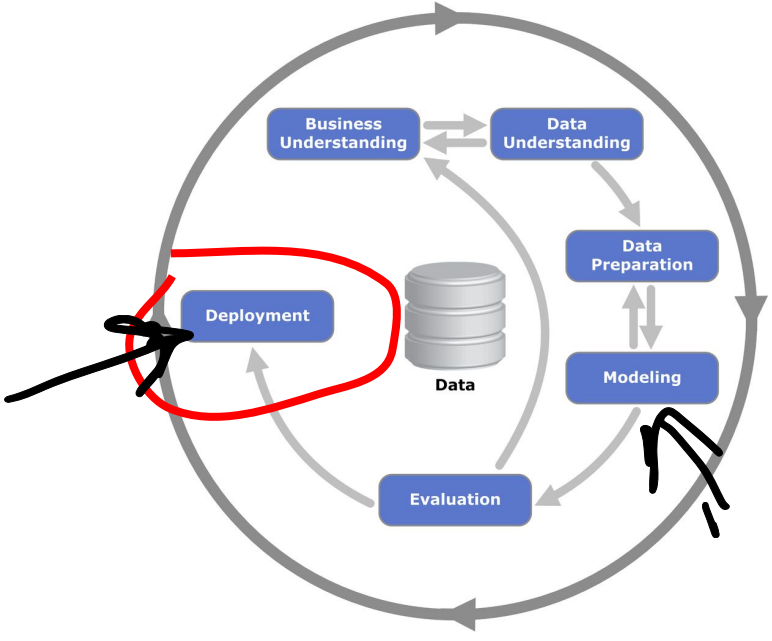
Deploy the model to production



Picture: CRISP-DM

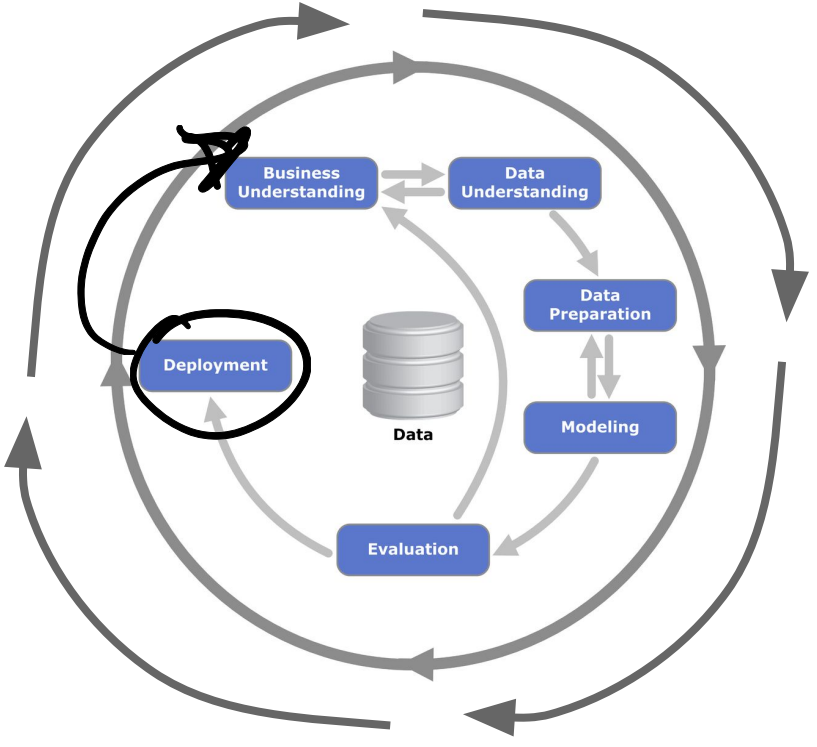
Deployment

- Roll the model to all users
- Proper monitoring
- Ensuring the quality and maintainability



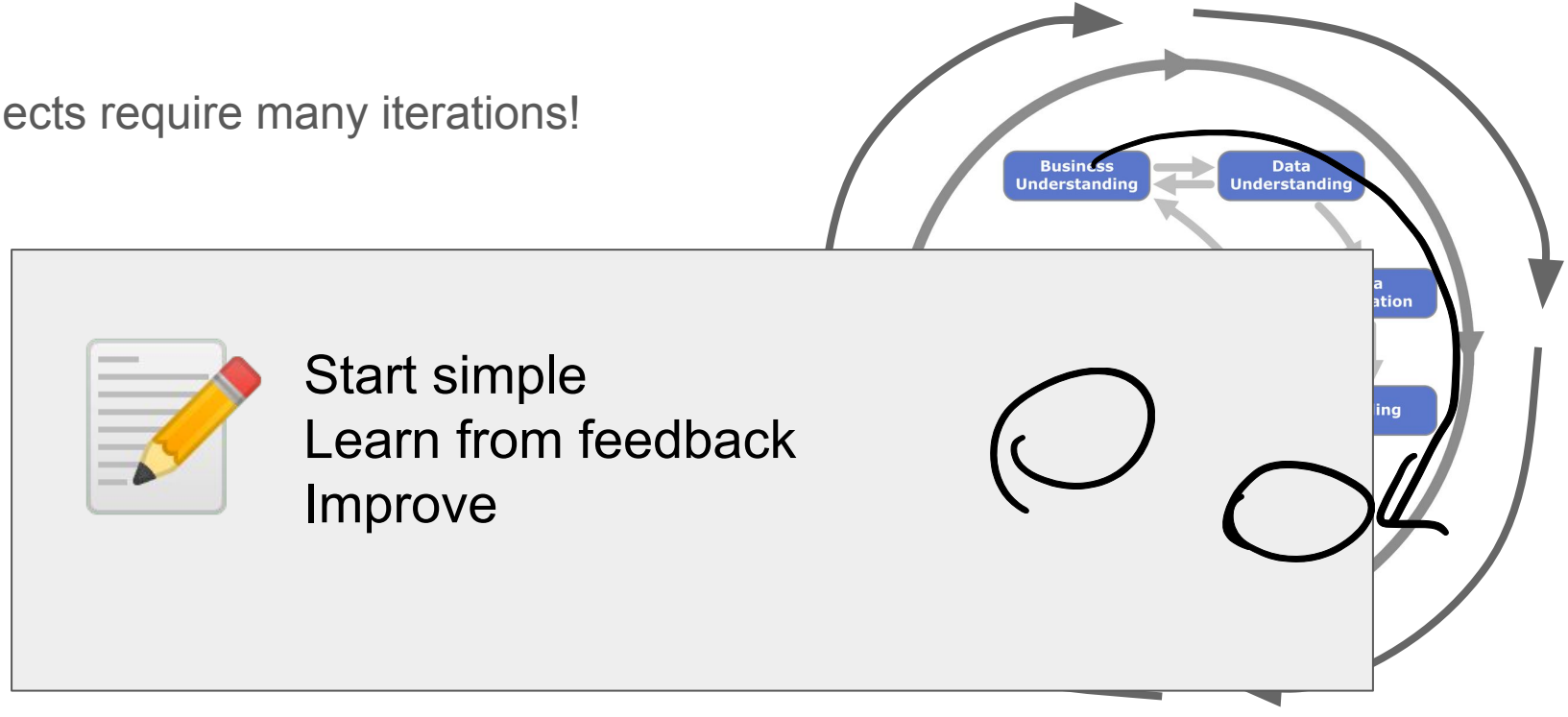
Iterate!

ML projects require many iterations!



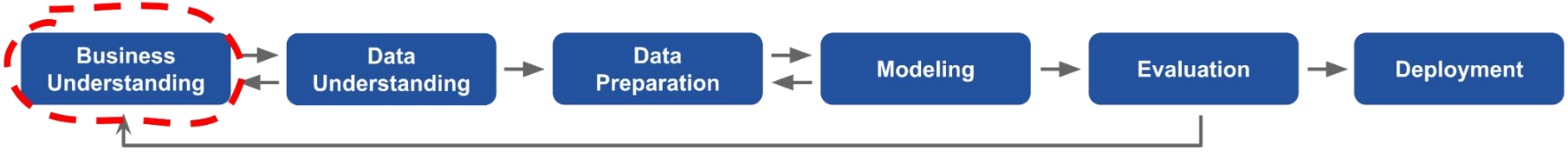
Iterate!

ML projects require many iterations!



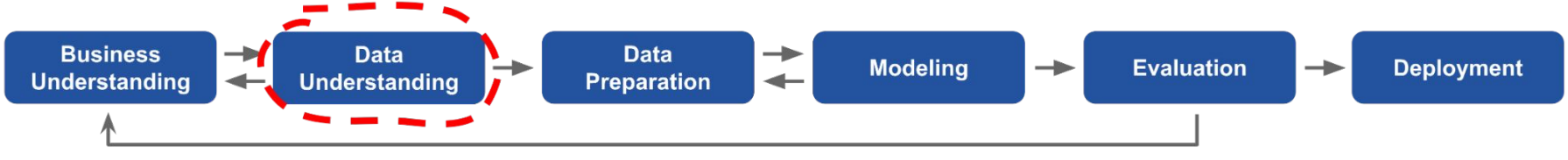
Summary

- Business understanding: define a measurable goal. Ask: do we need ML?



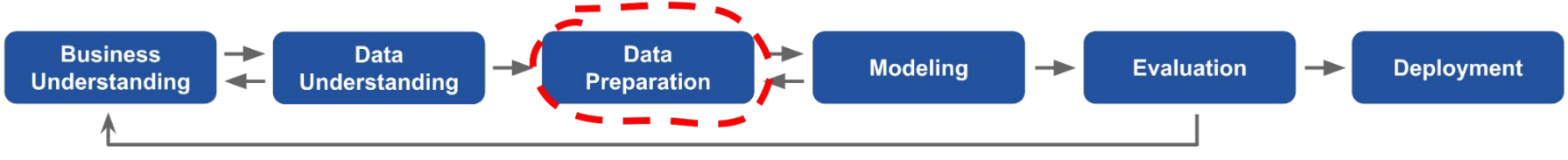
Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?



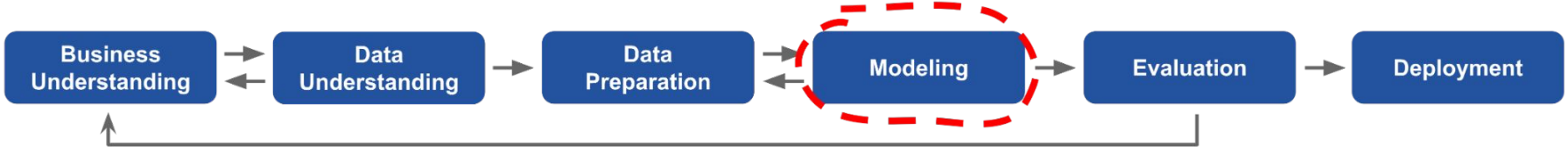
Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?
- Data preparation: transform data into a table, so we can put it into ML



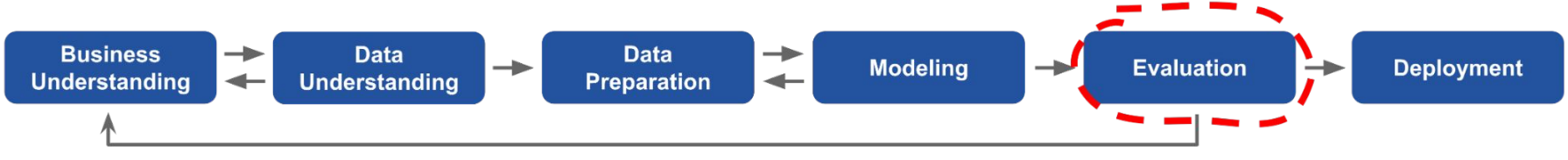
Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?
- Data preparation: transform data into a table, so we can put it into ML
- Modelling: to select the best model, use the validation set



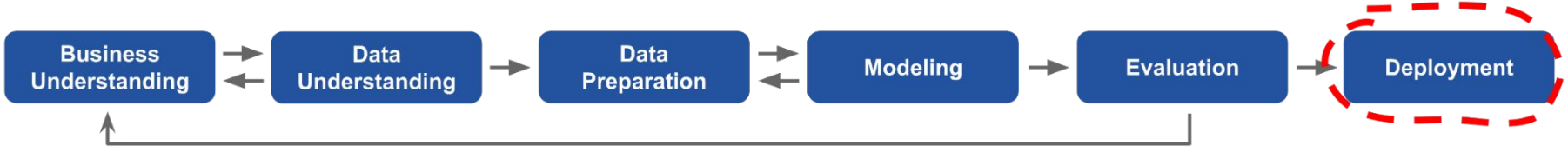
Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?
- Data preparation: transform data into a table, so we can put it into ML
- Modelling: to select the best model, use the validation set
- Evaluation: validate that the goal is reached



Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?
- Data preparation: transform data into a table, so we can put it into ML
- Modelling: to select the best model, use the validation set
- Evaluation: validate that the goal is reached
- Deployment: roll out to production to all the users



Summary

- Business understanding: define a measurable goal. Ask: do we need ML?
- Data understanding: do we have the data? Is it good?
- Data preparation: transform data into a table, so we can put it into ML
- Modelling: to select the best model, use the validation set
- Evaluation: validate that the goal is reached
- Deployment: roll out to production to all the users
- Iterate: start simple, learn from the feedback, improve

Next

The modelling step of CRISP-DM