

Data Driven Project



Wize



DML

Team



Maryam Ramezani

CTO & Co-founder @Wize Analytics
| PhD candidate @DML & Lecturer
at Sharif University of Technology
maryam.ramezani@sharif.edu



Amin Kashiri

Data Scientist @Wize Analytics |
Sharif University of Technology
kashiri.amin@gmail.com

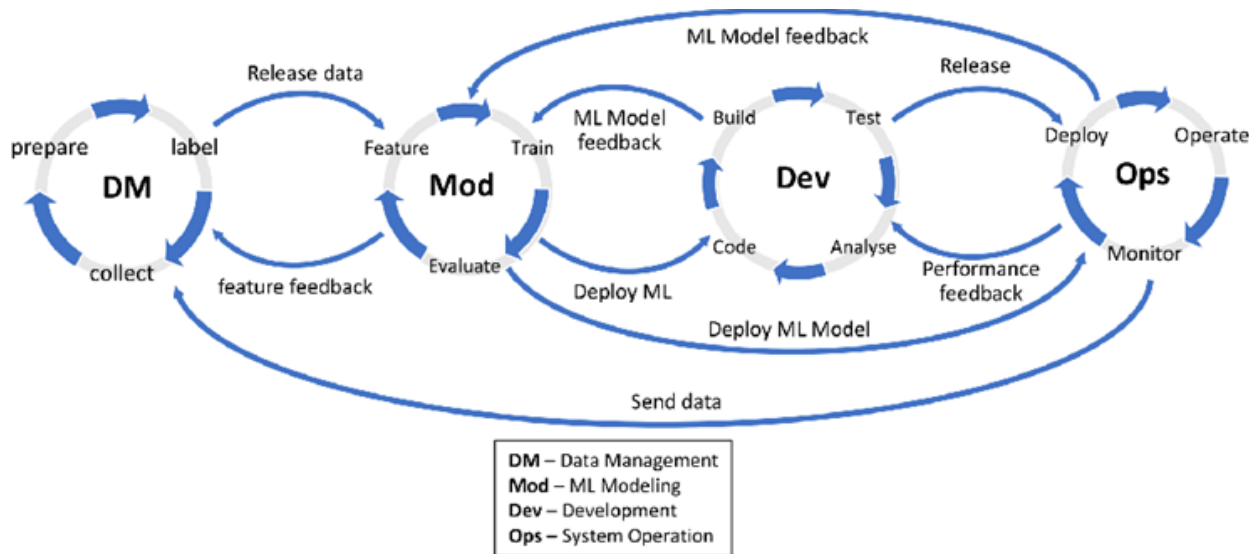
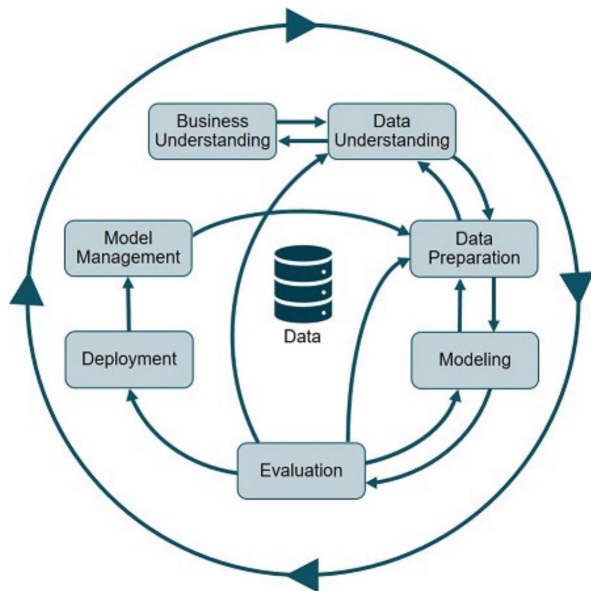


Gita Sarafranz

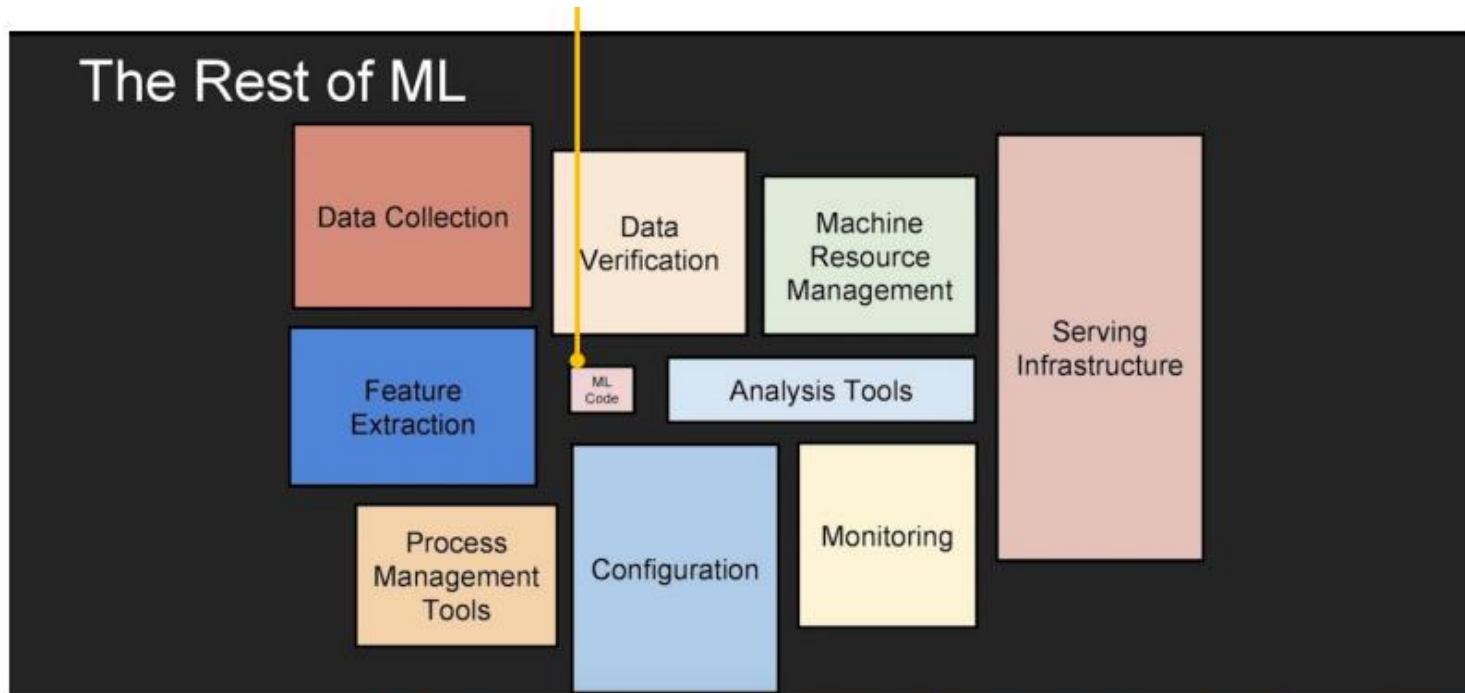
Data Scientist @Wize Analytics |
PhD Student @DML |
Sharif University of Technology
gita.sarafranz@gmail.com

Process Model: CRISP

- The iterative nature of CRISP-DM—something it has in common with the Agile philosophy—makes it a good way to think about data science projects.

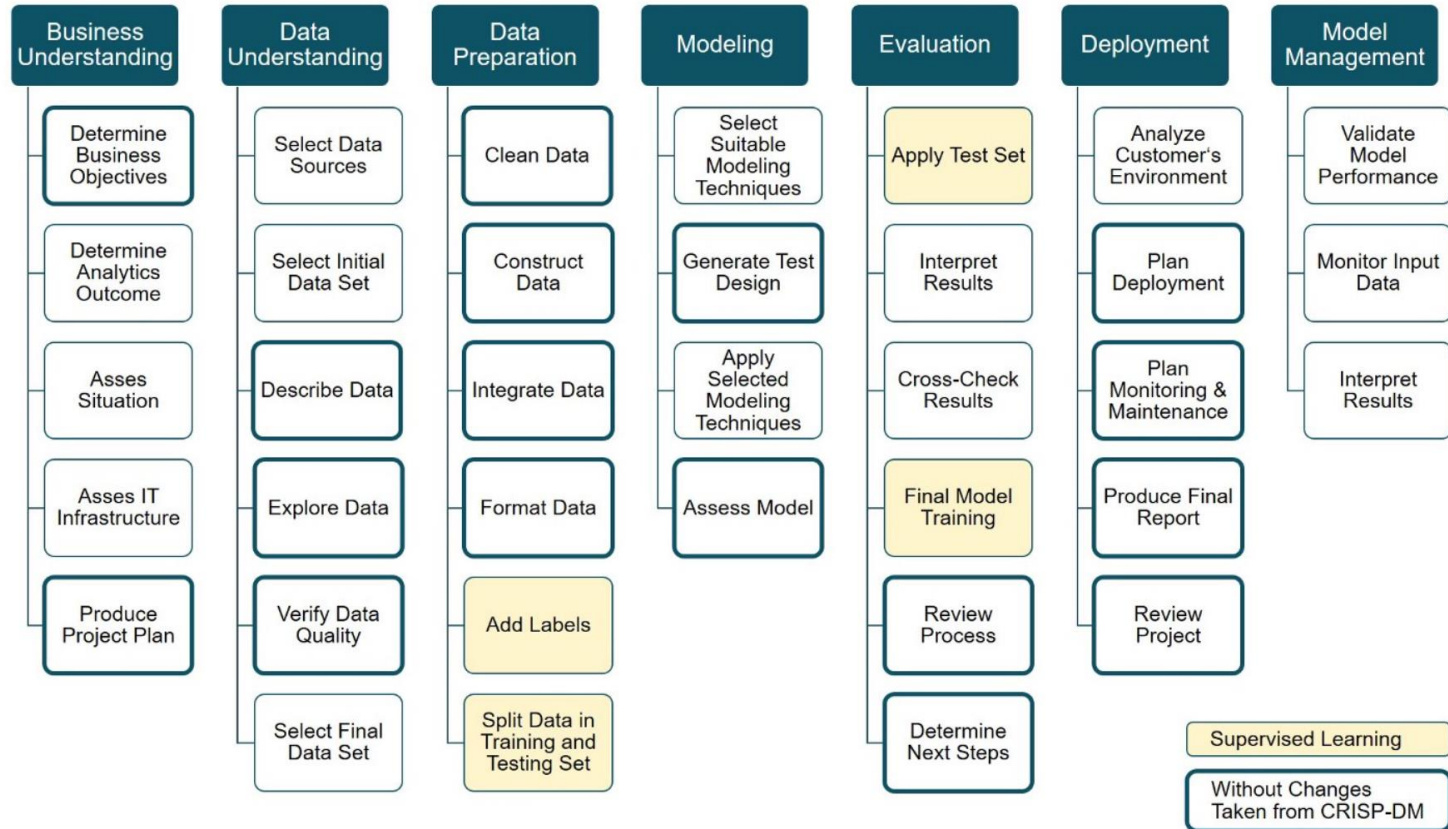


Machine Learning Systems

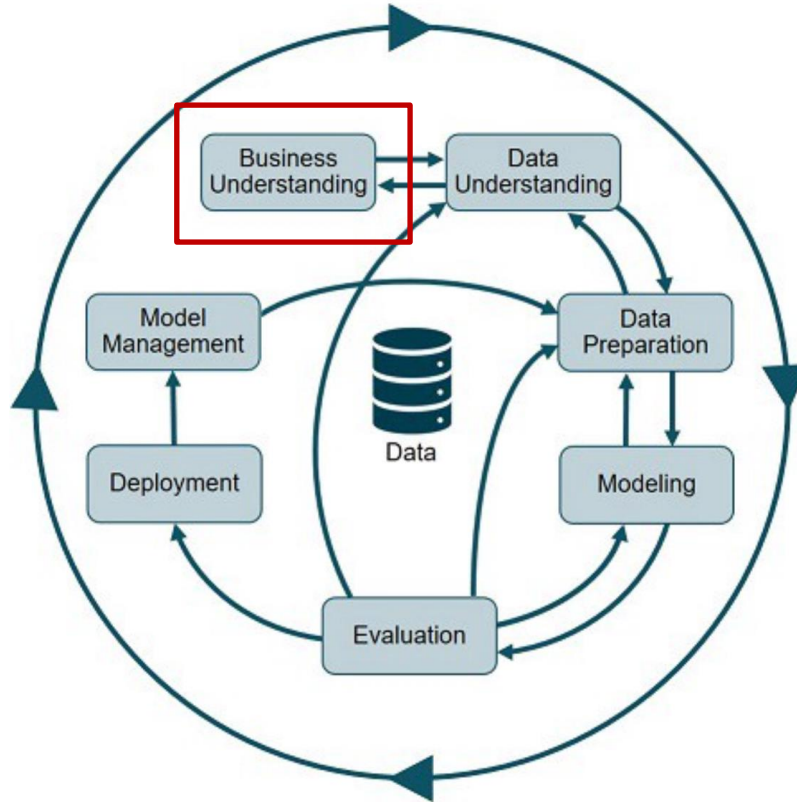


-- D. Sculley, et. al., Hidden Technical Debt in Machine Learning Systems, NIPS 2015¹

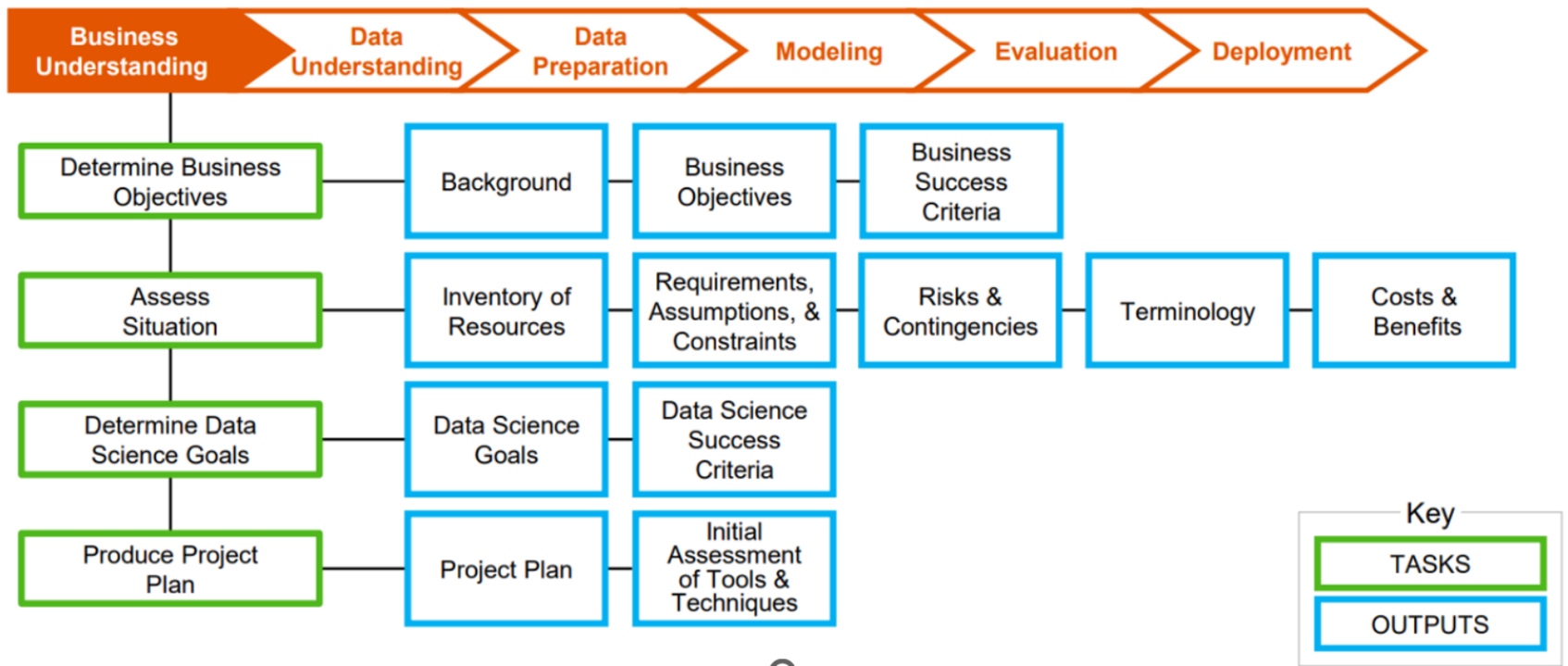
Process Model: CRISP



Process Model: CRISP



CRISP: Business Understanding



CRISP: Business Understanding

- Align on what strategic business problem you are aiming to solve using ML
- What is the credible business use case?
- What is the value you are creating?
- What stakeholders do you need to involve in the problem?
- What are the KPIs for the success of your model?
- How does it align with the business strategy?

Business Understanding: Objectives and outcomes

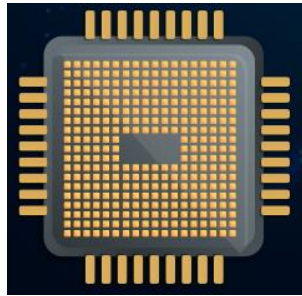
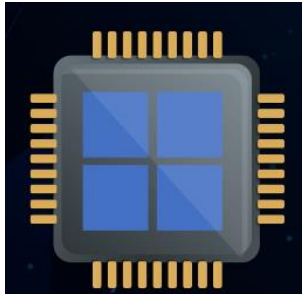
- Understanding the business goals: minimize later risk by clarifying problems, goals, and resources.
- Success criteria for the data project (the evaluation is performed with respect to)
- Task List
 - Start gathering background information about the current business situation.
 - Document specific business objectives decided upon by key decision makers.
 - Agree upon criteria used to determine data mining success from a business perspective.

Business Understanding: Assessing the Situation

Now that you have a clearly defined goal, it's time to make an assessment of where you are right now. This step involves asking questions such as:

- What source of data are available for analysis?
- Do you have the personnel needed to complete the project?
- What are the biggest risk factors involved?
- Do you have a contingency plan for each risk?

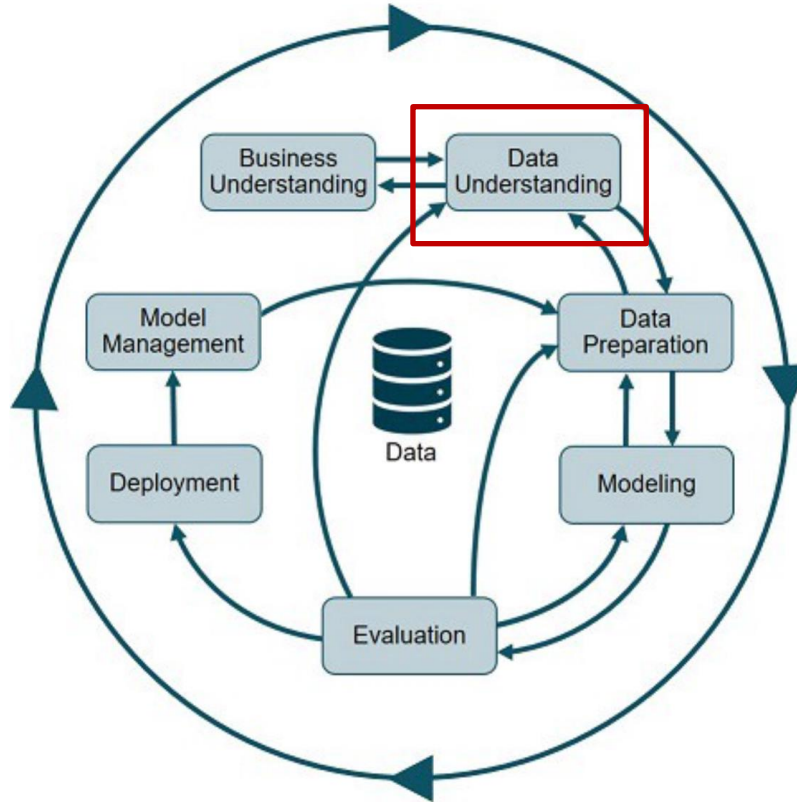
Business Understanding: Assess IT Infrastructure



Business Understanding: Produce Project Plan

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data mining consultant, some database analyst time	Data problems, technology problems
Modeling	2 weeks	Data mining consultant, some database analyst time	Technology problems, inability to find adequate model
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data mining consultant, some database analyst time	Economic change, inability to implement results

Process Model: CRISP



Data Understanding: Overview

- Understand what data do you have
- Where to get the data
- What is in your data
- If your data is of quality

Data Understanding: Collect Initial Data

- *WHAT* data to acquire
- *WHERE* to acquire
- *HOW* to acquire
- *HOW MUCH* data to acquire



What data can help solving the problem?

Data is not always complete or enough, but the problem exists!

Different data is needed for different parts of the problem. In a vehicle plate decoding problem, we need data for plate detection, and data for plate decoding.

→ Data collection report

Data Understanding: Collect Initial Data

- *WHAT* data to acquire
 - *WHERE* to acquire
 - *HOW* to acquire
 - *HOW MUCH* data to acquire
- Data is not already prepared!
We may need:
- Generating (labeling data)
 - Writing scripts (even for labeling data!)
 - Crawling
 - Developing data collection framework

→ Data collection report

Data Understanding: **Collect Initial Data**

- **WHAT** data to acquire
- **WHERE** to acquire
- **HOW** to acquire
- **HOW MUCH** data to acquire

→ Data has different forms and sources:

- Streams (GPS or IoT sensors)
- Images
- Documents
- Databases
- Objects structured in folders (Often hierarchical)
- APIs

Data needs to be stored

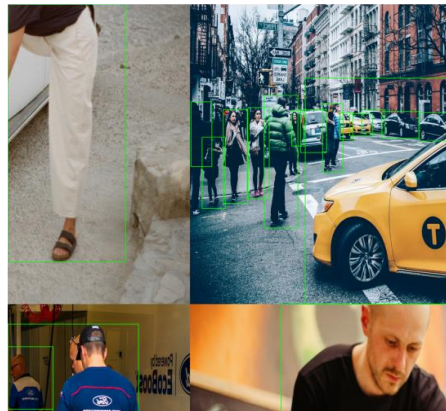
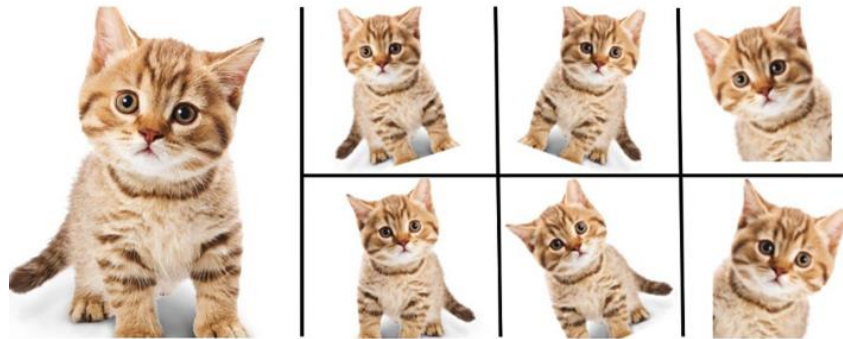
→ Data collection report

Data Understanding: Collect Initial Data

- *WHAT* data to acquire
- *WHERE* to acquire
- *HOW* to acquire
- *HOW MUCH* data to acquire



Sometimes we may need techniques like Data Augmentation!



Data Understanding: Describe Data

- Properties of data
- Format of the data
- Quantity of data
- The records and fields in each table or datasets

→ Data Description Report

Data Understanding: Explore Data

- Querying
 - Visualization
 - Reporting
 - Summary report
- We investigate statistical properties (max, min, mean, correlation between variables and ...).
 - We visualize distributions, trend, classes and Sometimes even simply looking at the data and chart is beneficial.
 - We may find problems like class imbalance, biased data and We should use this information we designing our model.
 - In this step we may need going back to earlier steps. Maybe this data does not answer our problem. We can't just ignore problems or leave solving the problem.

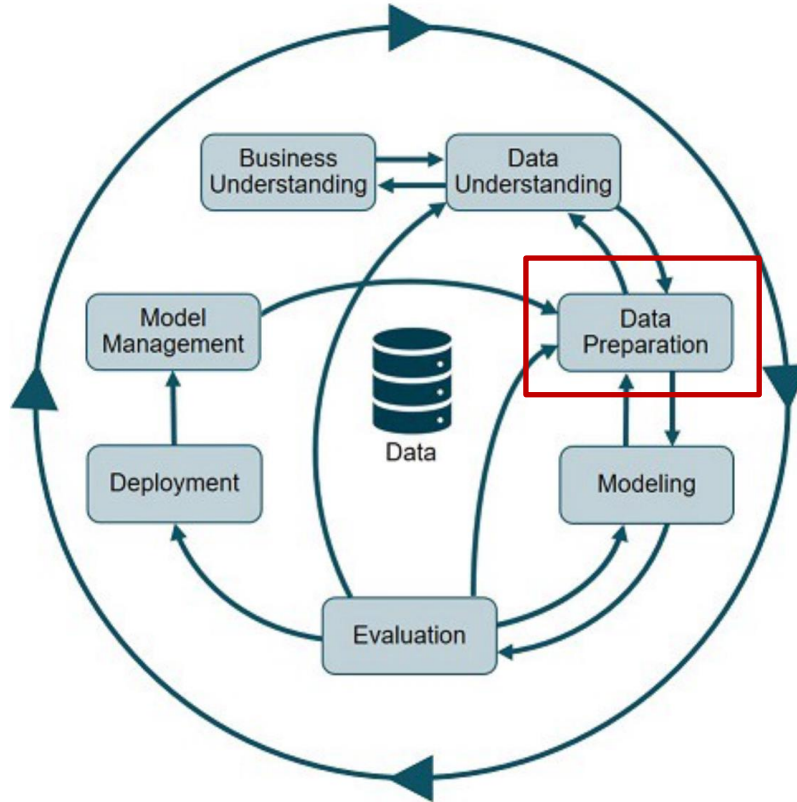
→ Data Exploration Report

Data Understanding: Verify Data Quality

- Is Data complete?
- Does data have errors and missing values?
- How much?
- What can be done?

→ Data Quality Report

Process Model: CRISP



Data Preparation: Overview

Data preparation phase covers all activities to construct the final dataset from the initial raw data in order to prepare the data for further processing.

Benefits

- Identify and fix errors before feeding them to the model
- High quality data
- Make better business decisions
- Make model interpret and produce good results

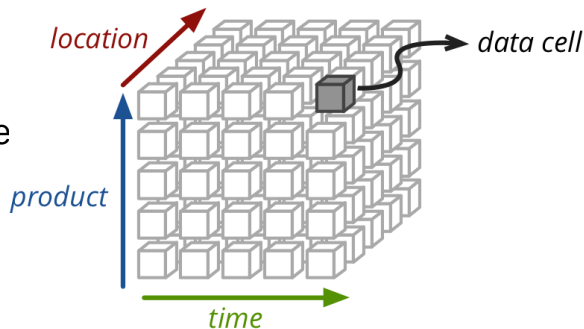
Data Preparation: **Select Data**

- Select data that is relevant to business goal
- List of data that need to be excluded, along with its reasons
- Store data that is needed

Data Preparation: Integrate Data

- Joining datasets, and data stored in different forms
- Edit metadata, to handle inconsistencies like types, naming ...
- Usually SQL knowledge is important in this step
- Designing Data Warehouses

Sometimes most of the business problems are solved in this step



Data Preparation: Clean Data

How to address the data quality problems
that were reported during Data Quality
Verification step of Data Understanding

- Handle missing values and features
- Handle duplicate or redundant data
- Handle outliers and noises
- Handle inconsistent and conflicting data

Steps:

- Screening
- Diagnosing
 - Missing data
 - Error
 - True extreme
- Treatment
 - Do nothing
 - Delete
 - Correct

Data Preparation: Clean Data

Two examples:

- Duplicate data
- Missing Value

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX		Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX		Mid
6	CA	76600	High
7	NY	85000	High
8	CA		Entry
9	CT	45000	Entry

Missing values

	A	B	C	D
1	Last Name	Sales	Country	Quarter
2	Smith	\$16,753.00	UK	Qtr 3
3	Johnson	\$14,808.00	USA	Qtr 4
4	Williams	\$10,644.00	UK	Qtr 2
5	Jones	\$1,390.00	USA	Qtr 3
6	Brown	\$4,865.00	USA	Qtr 4
7	Smith	\$16,753.00	UK	Qtr 3
8	Williams	\$12,438.00	UK	Qtr 1
9	Johnson	\$9,339.00	UK	Qtr 2
10	Smith	\$18,919.00	USA	Qtr 3
11	Jones	\$9,213.00	USA	Qtr 4
12	Jones	\$7,433.00	UK	Qtr 1
13	Smith	\$16,753.00	UK	Qtr 3
14	Brown	\$3,255.00	USA	Qtr 2
15	Williams	\$14,867.00	USA	Qtr 3
16	Williams	\$19,302.00	UK	Qtr 4
17	Smith	\$9,698.00	USA	Qtr 1
18				

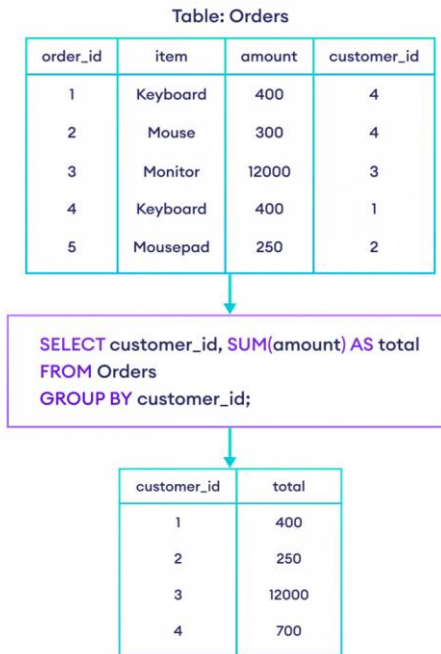
Note: Cleaning data is needed in every step of the procedure

- Measurement
- Collecting
- Storing
- Integrating
- Transforming

Cleaning is not a one-time step

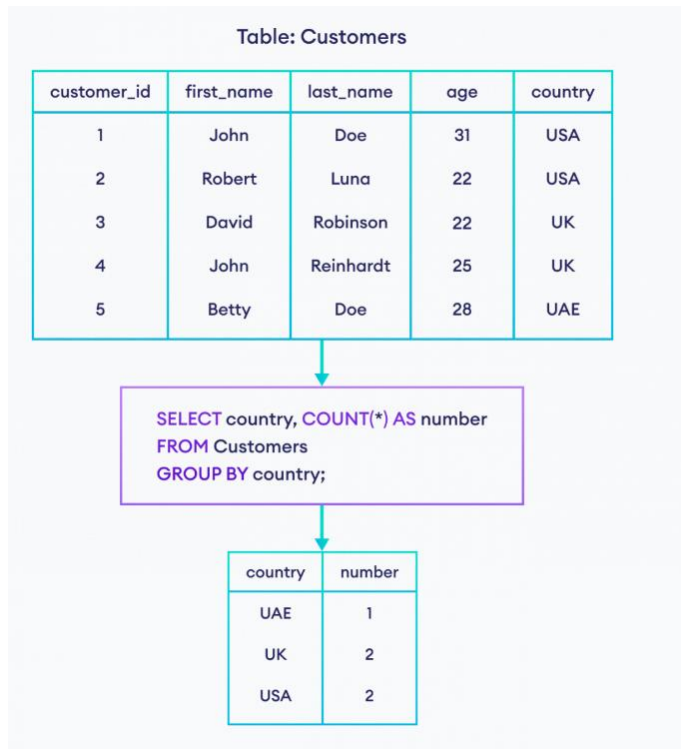
Data Preparation: Transform/Format Data

- Normalization / Scaling
- Aggregation
 - City > State > Country
- Generalization
- Feature Construction
 - Add features derived from existing ones



Data Preparation: Reduce Data

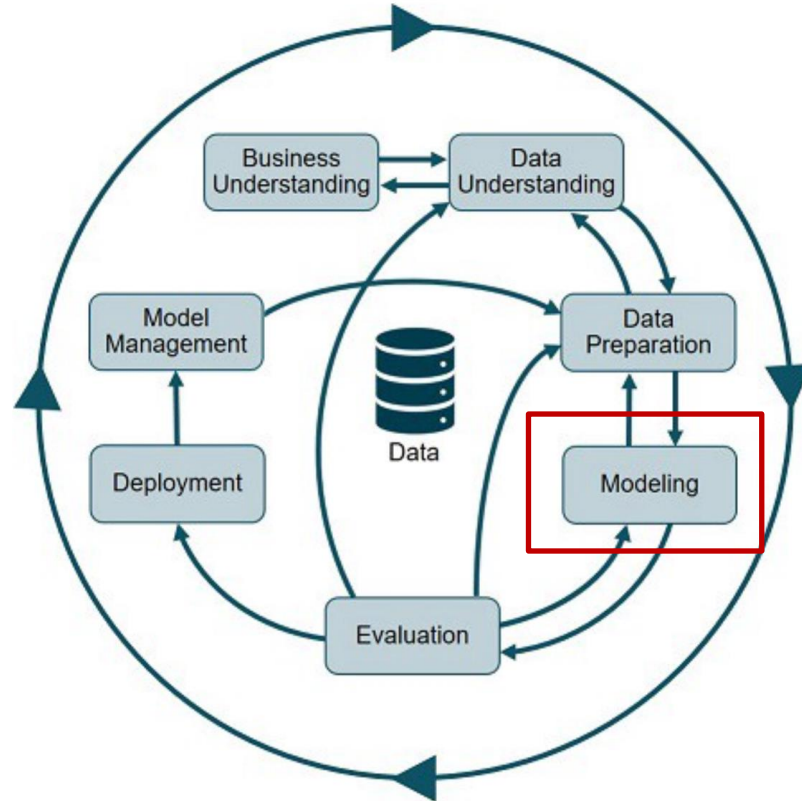
- Dimensionality Reduction and Data Compression
 - Feature Selection, PCA
- Aggregation and Generalization (like last step)



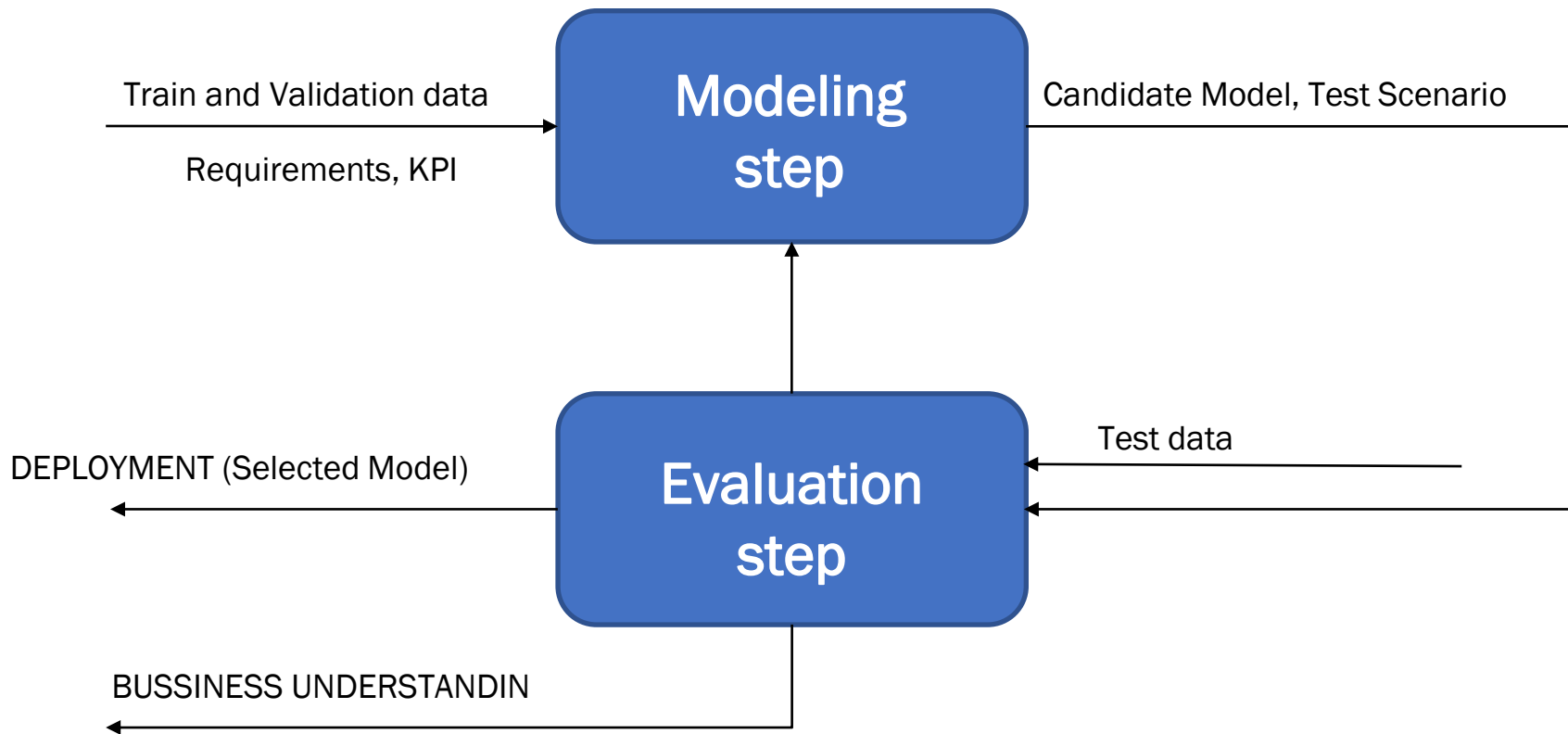
Data Preparation: Split Data

- Split data to train, test and validation set
 - train: Fitting and optimizing model parameters
 - validation: Hyperparameters tuning and avoid overfitting
 - test: Evaluate the model
- Balance data distribution between sets (or sometimes the opposite)

Process Model: CRISP



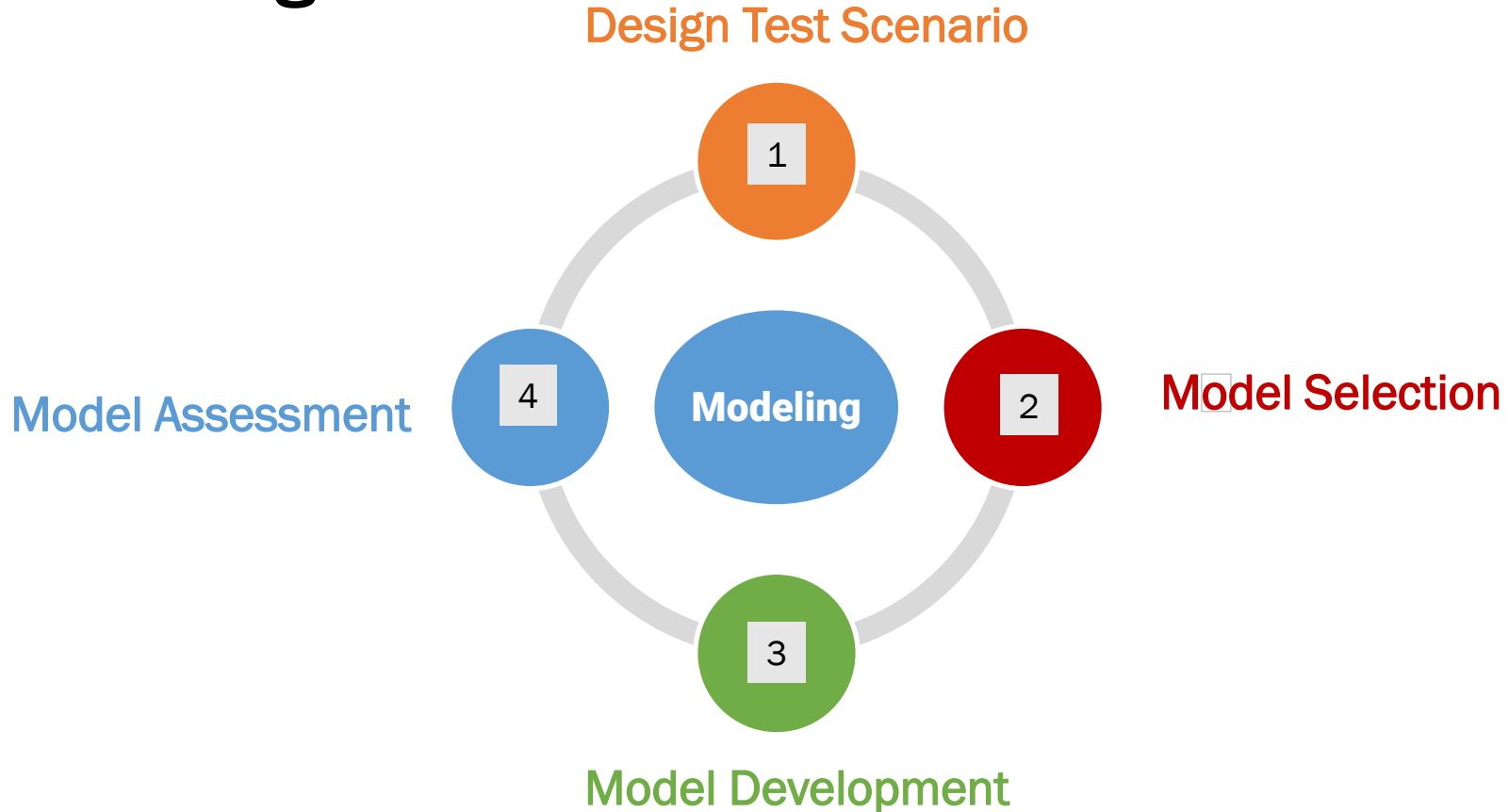
Modeling and Evaluation



Modeling

- This is the actual Data Science step
- **Modelling** includes **selecting**, **configuring** and **testing** various algorithms, as well as deciding on their sequence, which creates a model.
- In the modeling phase we would develop our machine learning model/product to answer the business question.


Modeling



Modeling: Design Test Scenarios

- Design your modeling test design by splitting the data into training, test, validation sets, or cross-validation. Justify the reason for how you design the test.
- Test scenarios based on:
 - Metrics
 - KPI
 - Resources
 - Requirements
 - ...

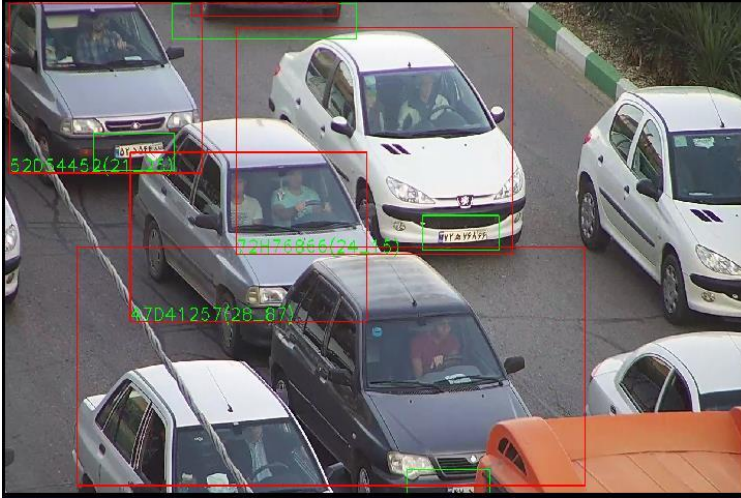

Modeling: Design Test Scenarios



تاریخ : 1398/02/04 دوربین : 202

ساعت : 10 : 22 : 03 توضیحات : تست

۵۷ ۸۴۳ د ۲۳

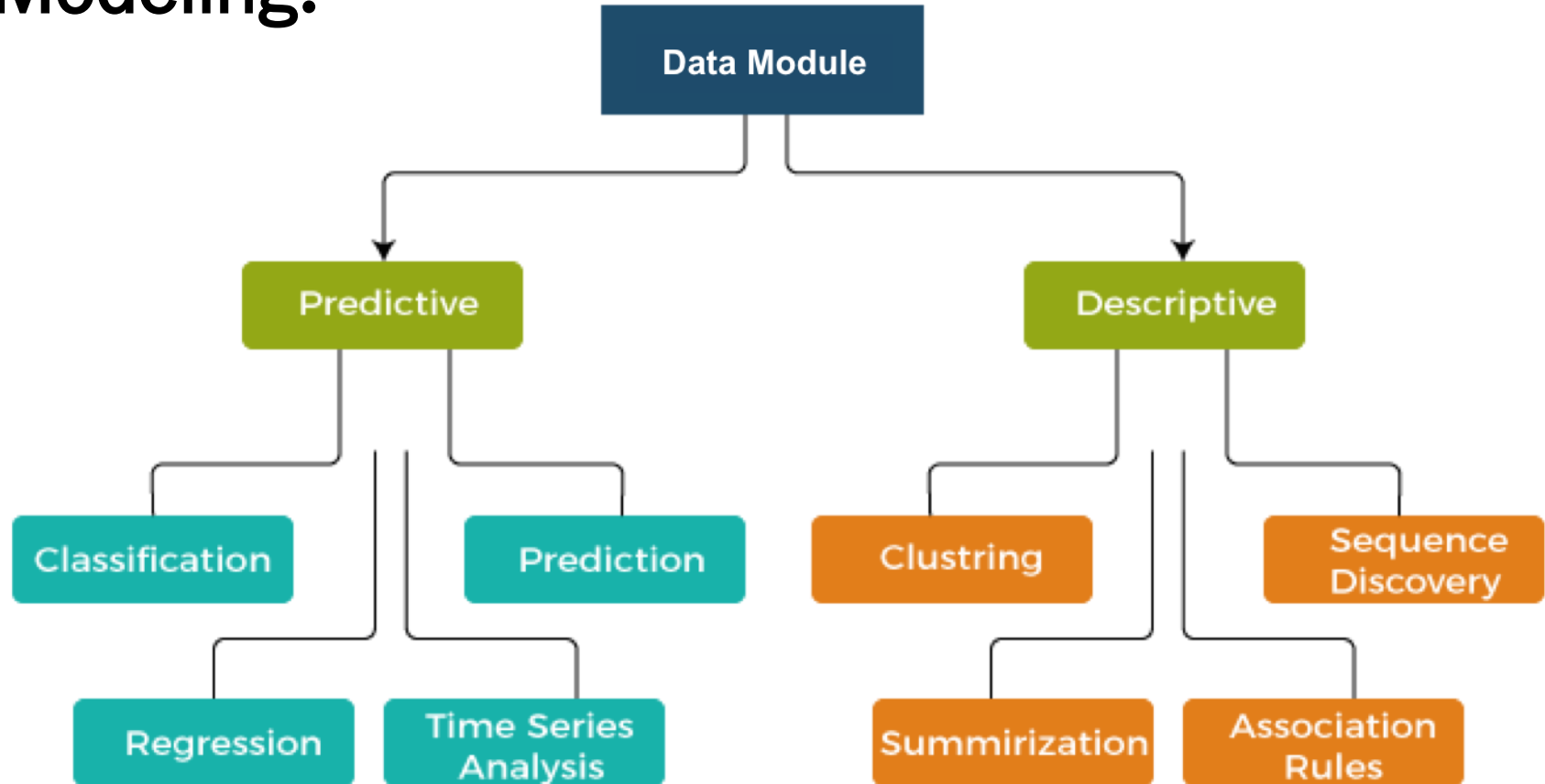



۵۷ ۸۴۳ د ۲۳	۵۲ ۸۴۳ د ۲۳	۵۷ ۲۲۱ د ۱۵	۵۲ ۵۴۲ د ۵۲	۶۶ ۷۶۸ ه ۷۲	

Modeling: Model Selection

- Using which machine learning algorithms and frameworks to try. You might want to experiment with many models. It would be desirable if you could explain why you select a certain algorithm.
- Customize the selected algorithms
- Decide on the sequence of algorithms and aggregation
- Prior knowledge

Modeling:



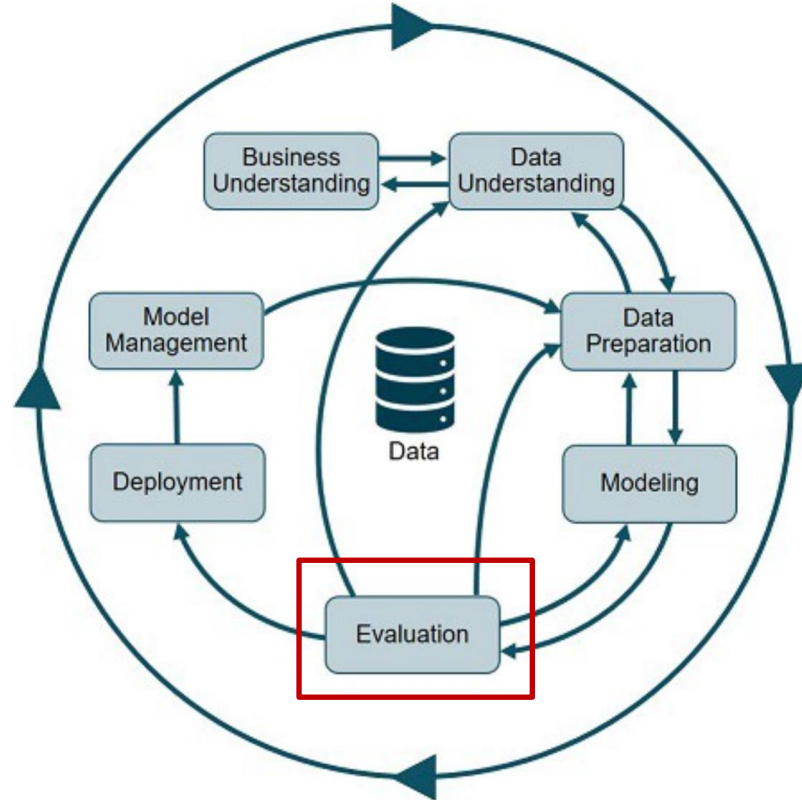
Modeling: Model Development

- Fit your model using the data you have prepared.
- Manage your resources well here — it might take a long time, depend on your data and your experiment design.
- The development should consider how the business question and industrial scene would be as well, such as “Would the model I develop is possible in the business,” “Is the resources I need to develop this model is costly?”

Modeling: Model Assessments

- Set your success technical metrics and choose the best model(s) viable for solving the business question. Try to explain why you decide on certain metrics and why not the other.

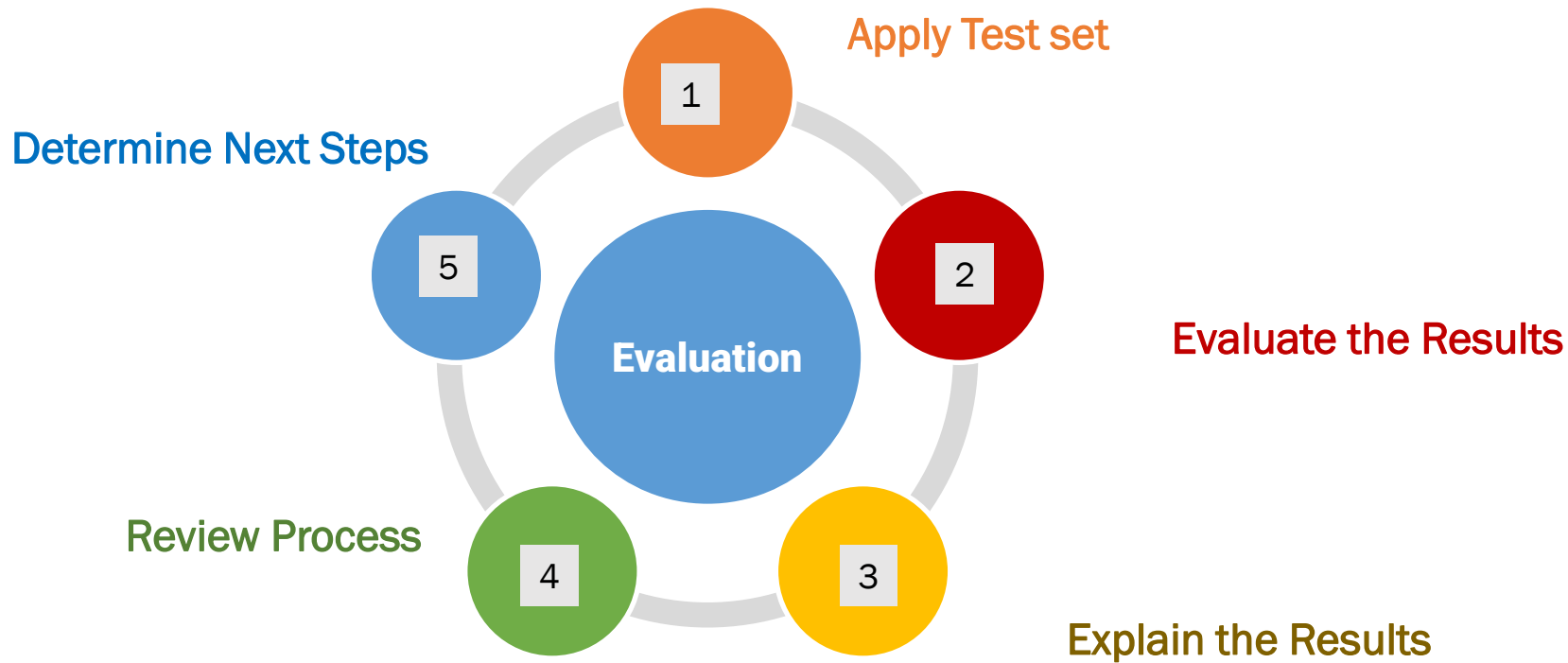
Process Model: CRISP



Evaluating

The Evaluation phase is different from the Modeling technical evaluation. This phase **evaluates the model concerning the business indicator and what to do next.**

Evaluating



Evaluating: Apply Test sets, Evaluate and Explain the Results

- Apply test sets base on selected test scenario
- Would the business success criteria be met using your model?
- which model(s) would you choose?
- This is the phase when you should explain how your model would help the business. Explain it as realistic as possible, and don't use too much technical jargon that people outside of the data world could understand.

Evaluating: Review Process

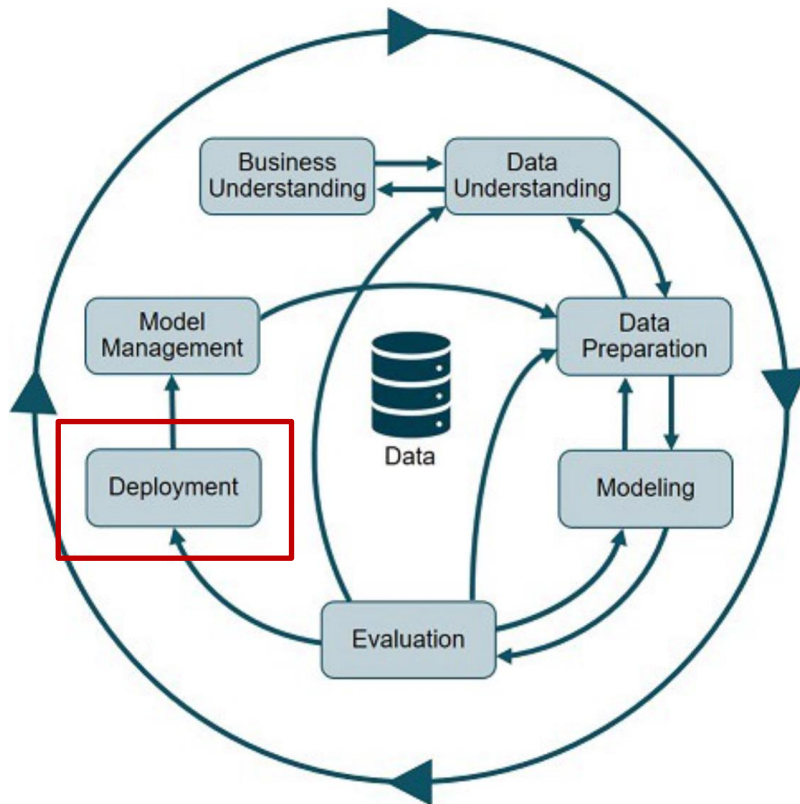
- Review your work process.
- Was anything missing?
- Need more time?
- Were all phases executed?
- Try to Summarize your findings and correct anything if required.

Evaluating: Determine Next Steps

Based on the previous three tasks, determine whether to proceed to:

- Deployment
- Iterate further
- Initiate new projects

Process Model: CRISP



Deployment: Plan Deployment

- The first step is to summarize your results--both models and findings.
 - This helps you determine which models can be integrated within your database systems and which findings should be presented to your colleagues.
- For each deployable model, create a step-by-step plan for deployment and integration with your systems.
 - Note any technical details such as database requirements for model output. For each conclusive finding, create a plan to disseminate this information to strategy makers.
- Consider how the deployment will be monitored.
 - How will you decide when the model is no longer applicable?
- Identify any deployment problems and plan for contingencies.
 - For example, decision makers may want more information on modeling results and may require that you provide further technical details.

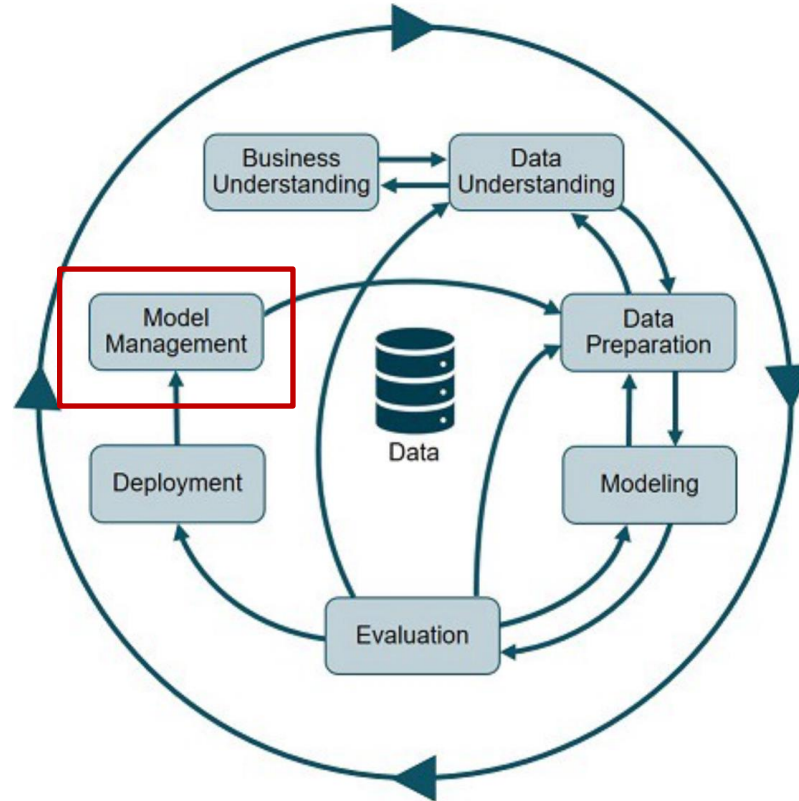
Deployment: Monitoring and Maintenance

- For each model or finding, which factors or influences (such as market value or seasonal variation) need to be tracked?
- How can the validity and accuracy of each model be measured and monitored?
- How will you determine when a model has "expired"? Give specifics on accuracy thresholds or expected changes in data, etc.
- What will occur when a model expires? Can you simply rebuild the model with newer data or make slight adjustments? Or will changes be pervasive enough as to require a new data mining project?
- Can this model be used for similar business issues once it has expired? This is where good documentation becomes critical for assessing the business purpose for each data mining project.

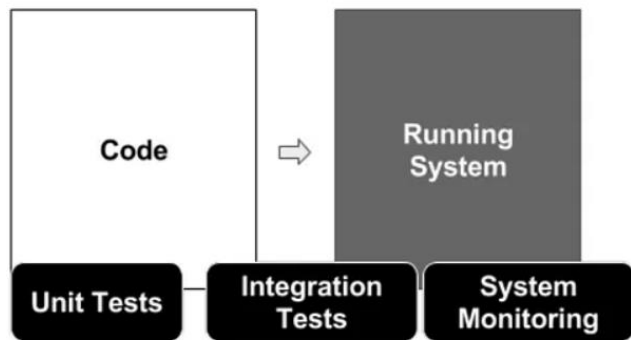
Deployment: Producing a Final Report

- A thorough description of the original business problem
- The process used to conduct the project
- Costs of the project
- Notes on any deviations from the original project plan
- A summary results, both models and findings
- An overview of the proposed plan for deployment
- Recommendations for further data work, including interesting leads discovered during exploration and modeling

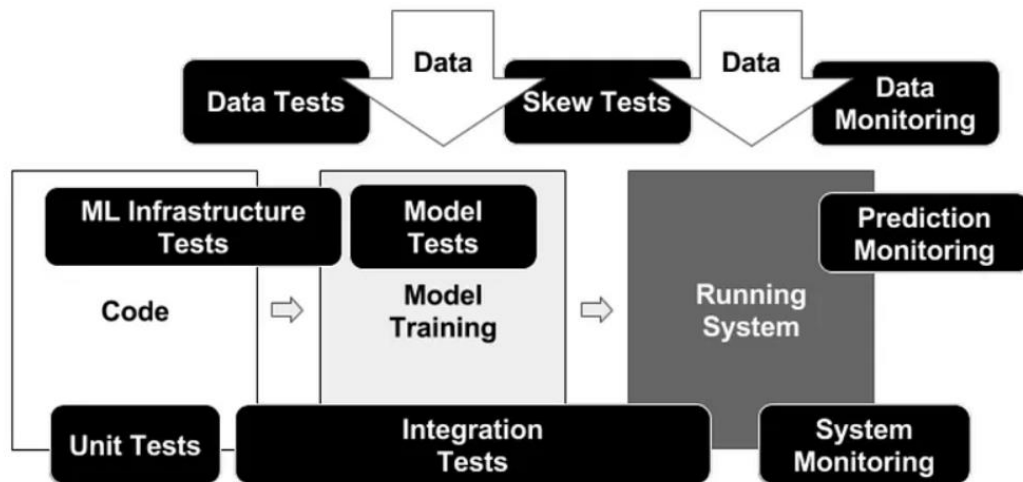
Process Model: CRISP



Model Management:



Traditional System Testing and Monitoring



ML-Based System Testing and Monitoring

Model Management: Concept Drifts

- ML estimates " $f(x) = y$ "
 - What if the relationship between "x" & "y" changes over time?
 - What if "f" does not capture certain relationships?
- In general, impossible to predict
 - Continuously monitor and update model

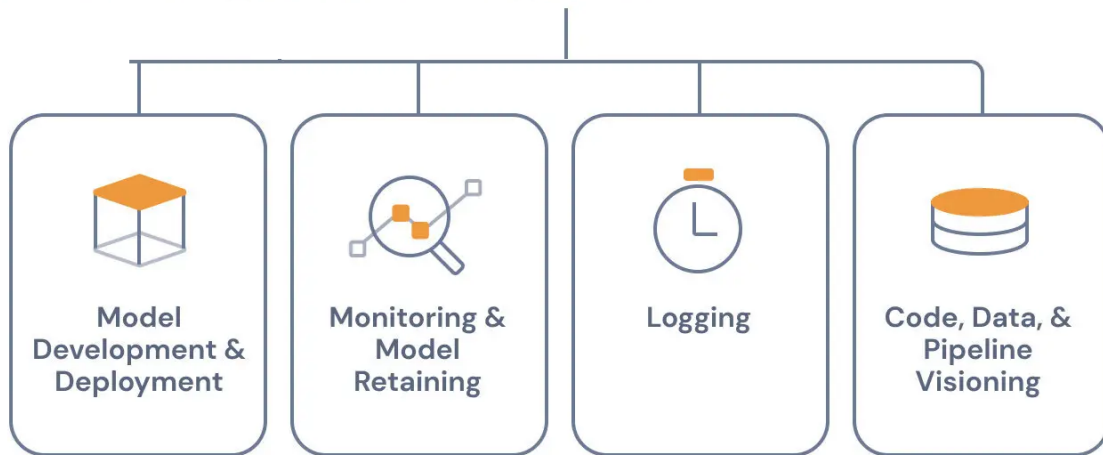
Model Management: Measurements in AI-Enabled Systems

- Organizational objectives
 - e.g., revenues, growth, lives saved, societal benefits
 - Often not directly measurable from system output; slow indicators
- Leading indicators
 - Customer sentiment: Do they like the product?
 - Customer engagement: How often do they use the product?
 - But can be misleading (more daily active users => higher profits?)
- User outcomes
 - Does the system achieve what it promises to users?
- Model Properties
 - Accuracy of predictions, error rates
 - Performance (e.g., prediction time)
 - Cost: Training time, amount of data required

Model Management: Monitoring

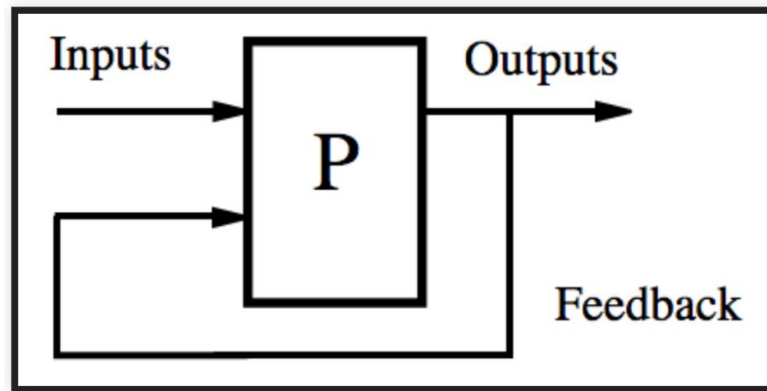
- Controlling models
- Performance
- Scalability

Machine Learning Model Management



Model Management: Feedback Loops

- Every system is deployed as part of an environment
- Output influences the environment
 - In turn, affects input back to the system
 - Over time, may lead to undesirable (and difficult to reverse) outcome
 - Higher risks if initial data set & model is biased



Important Notes

- Searching skill is very import
- A technician is not a data scientist !!!!
- Algebra/ Statistics are very important
- The wheel should not be reinvented
- Small ideas, big changes
- Scrum (prototype)

Important Notes

- Model generalization is too important (data, model)
- Interpretation and explanation is a key factor
- Your first data science project iteration did not need to be the perfect one. Learning from the mistake is part of the process and what the company wants to see as well.
- In a real-world working environment, we don't try to achieve perfection. What we want is a “good enough”. If you feel you haven't achieved that “99% Accuracy” model, it is fine. What is important is you able to explain the process.

Questions?