UNIVERSITAS GADJAH MADA

CRISP-ML(Q): Cross-Industry Standard Process for the Development of Machine Learning Applications with Quality Assurance

Dr.Eng. Igi Ardiyanto

PROFILE

Dr. IGI ARDIYANTO

Deep Learning • Computer Vision • Automation
Robotics • Embedded Systems

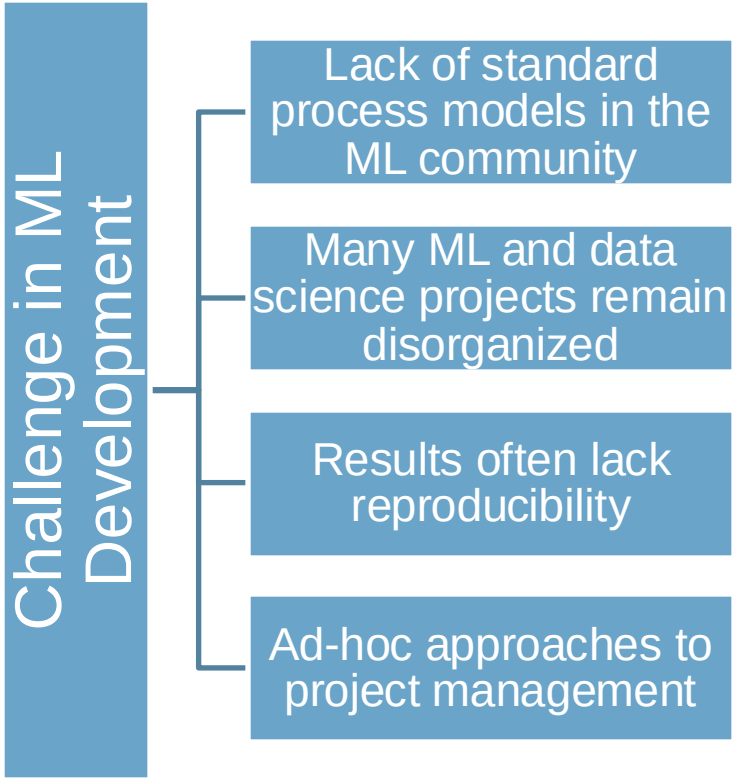
CONTACT
Dept. of Electrical Engineering and Information Tech.
Faculty of Engineering, Universitas Gadjah Mada
Jl. Grafika No.2 Bulaksumur
Yogyakarta, Indonesia 55281
E-mail : igi@ugm.ac.id

More Information ??
http://te.ugm.ac.id/~igi

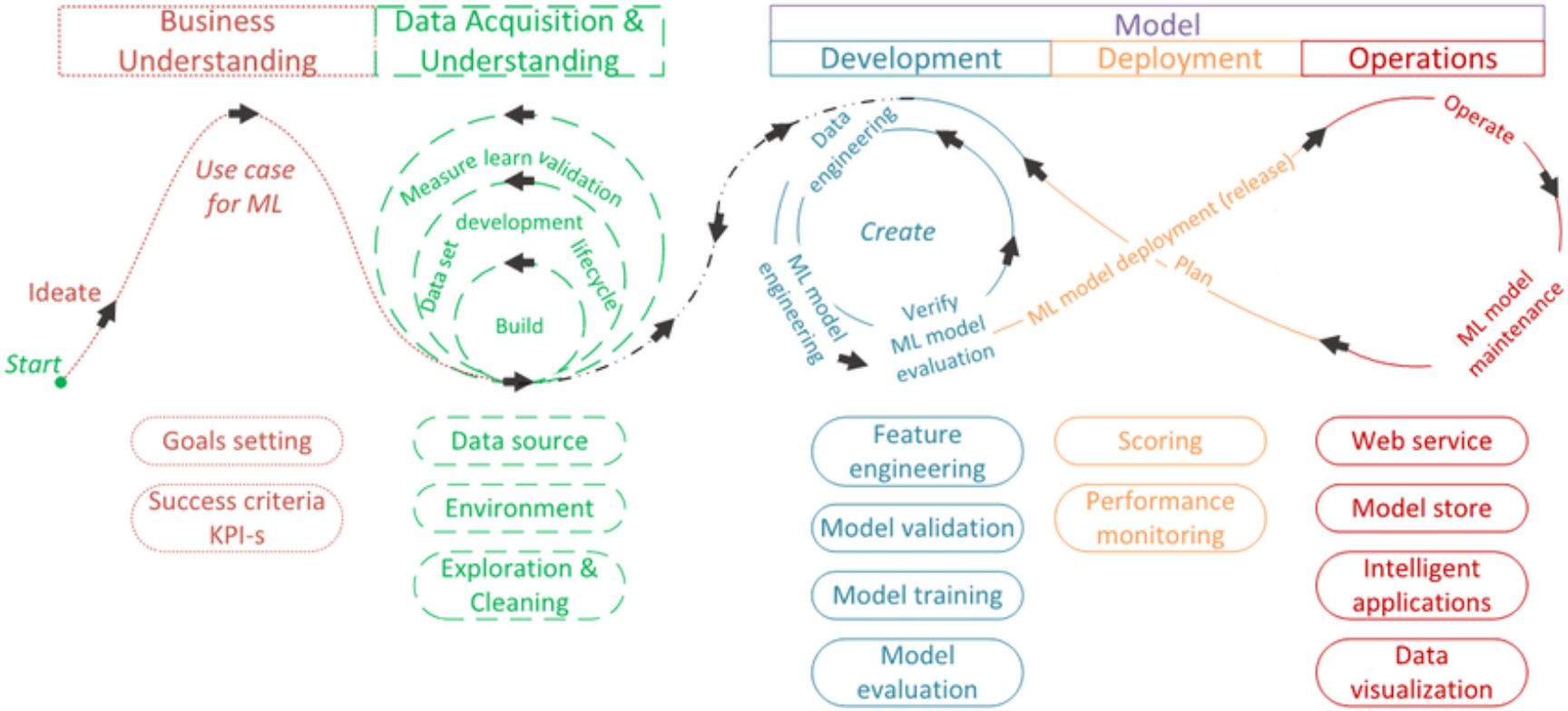
Academic :

- ❑2012 – 2015 Doctoral Degree, Toyohashi University of Technology, Japan
- ❑2010 – 2012 Master Degree, Toyohashi University of Technology, Japan

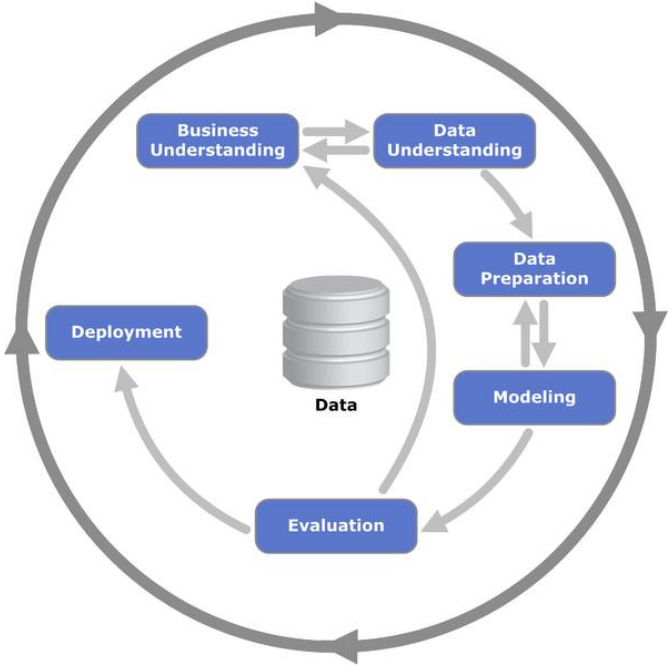
Challenge in ML Development

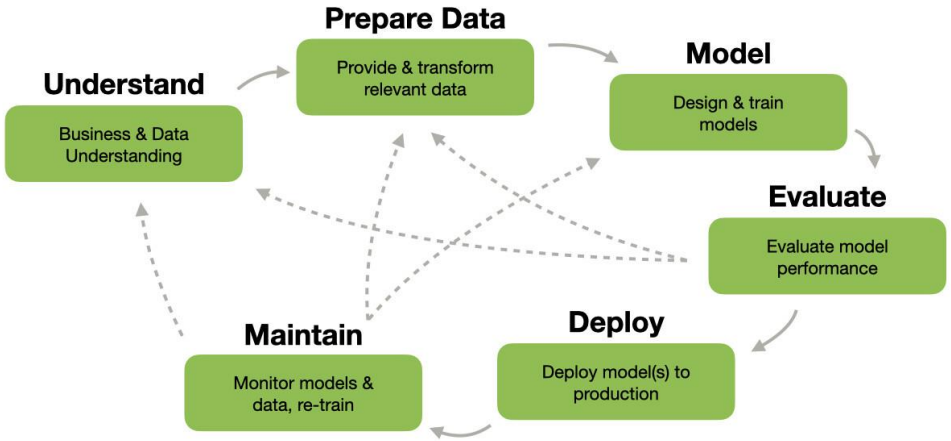


What is CRISP-ML(Q)



CRISP-DM vs CRISP-ML(Q)

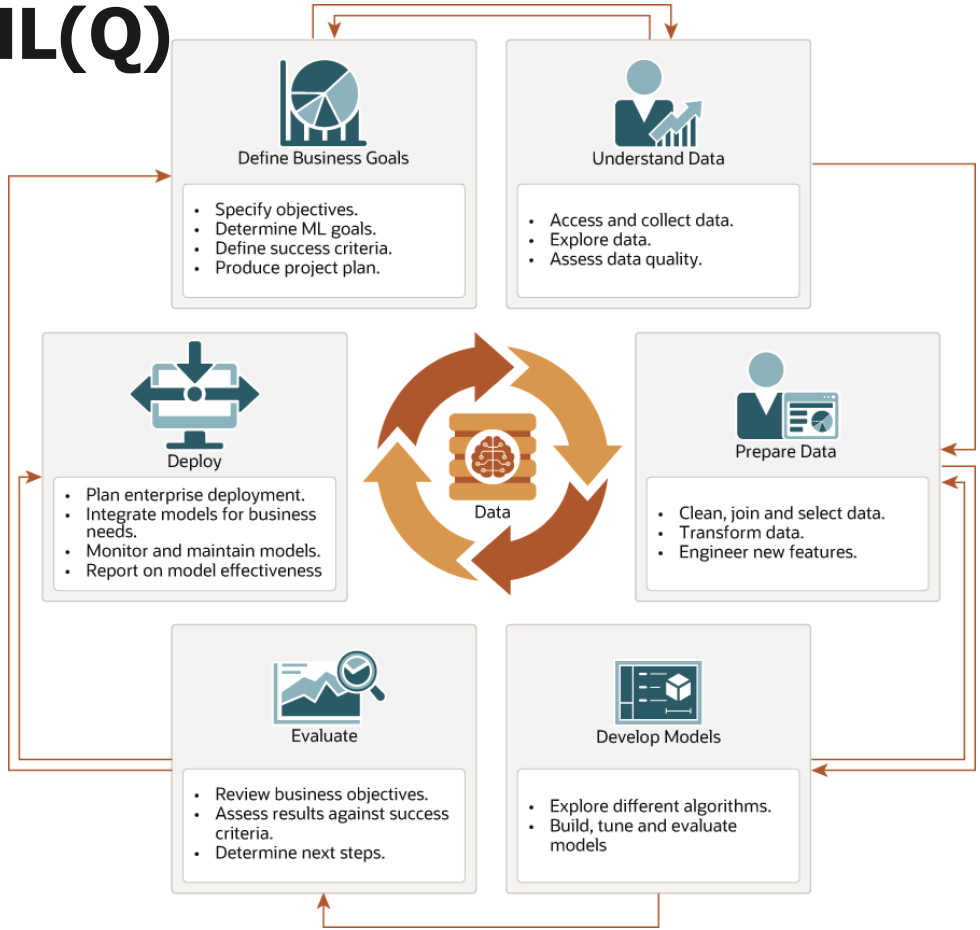




CRISP-DM vs CRISP-ML(Q)

CRISP-ML(Q)	CRISP-DM	Amershi *et al.* [28]	Breck *et al.* [29]	
Business & Data Understanding	Business Understanding	Requirements	-	
	Data Understanding	Collection	Data	Infrastructure
Data Preparation	Data Preparation	Cleaning		
		Labeling		
		Feature Engineering		
Modeling	Modeling	Training	Model	
Evaluation	Evaluation	Evaluation	-	
Deployment	Deployment	Deployment	-	
Monitoring & Maintenance	-	Monitoring	Monitoring	

CRISP-ML(Q)


Define Business Goals
Specify objectives.
Determine ML goals.
Define success criteria.
Produce project plan.
Understand Data
Access and collect data.
Explore data.
Assess data quality.
Deploy
Plan enterprise deployment.
Integrate models for business needs.
Monitor and maintain models.
Report on model effectiveness
Data
Prepare Data
Clean, join and select data.
Transform data.
Engineer new features.
Evaluate
Review business objectives.
Assess results against success criteria.
Determine next steps.
Develop Models
Explore different algorithms.
Build, tune and evaluate models

UNIVERSITAS GADJAH MADA

Phase 1:

BUSINESS UNDERSTANDING

Business Understanding: Types of Analytics

4 Types of Data Analytics

Descriptive Analytics

Make better predictions about future events based on historical data.

1

Diagnostic Analytics

Understand why something happened by analyzing data from previous events.

2

Predictive Analytics

Predict what will happen in the future by analyzing data from previous events.

3

Prescriptive Analytics

Predict & recommend how best to achieve those results by analyzing data from previous events.

4

Business Understanding

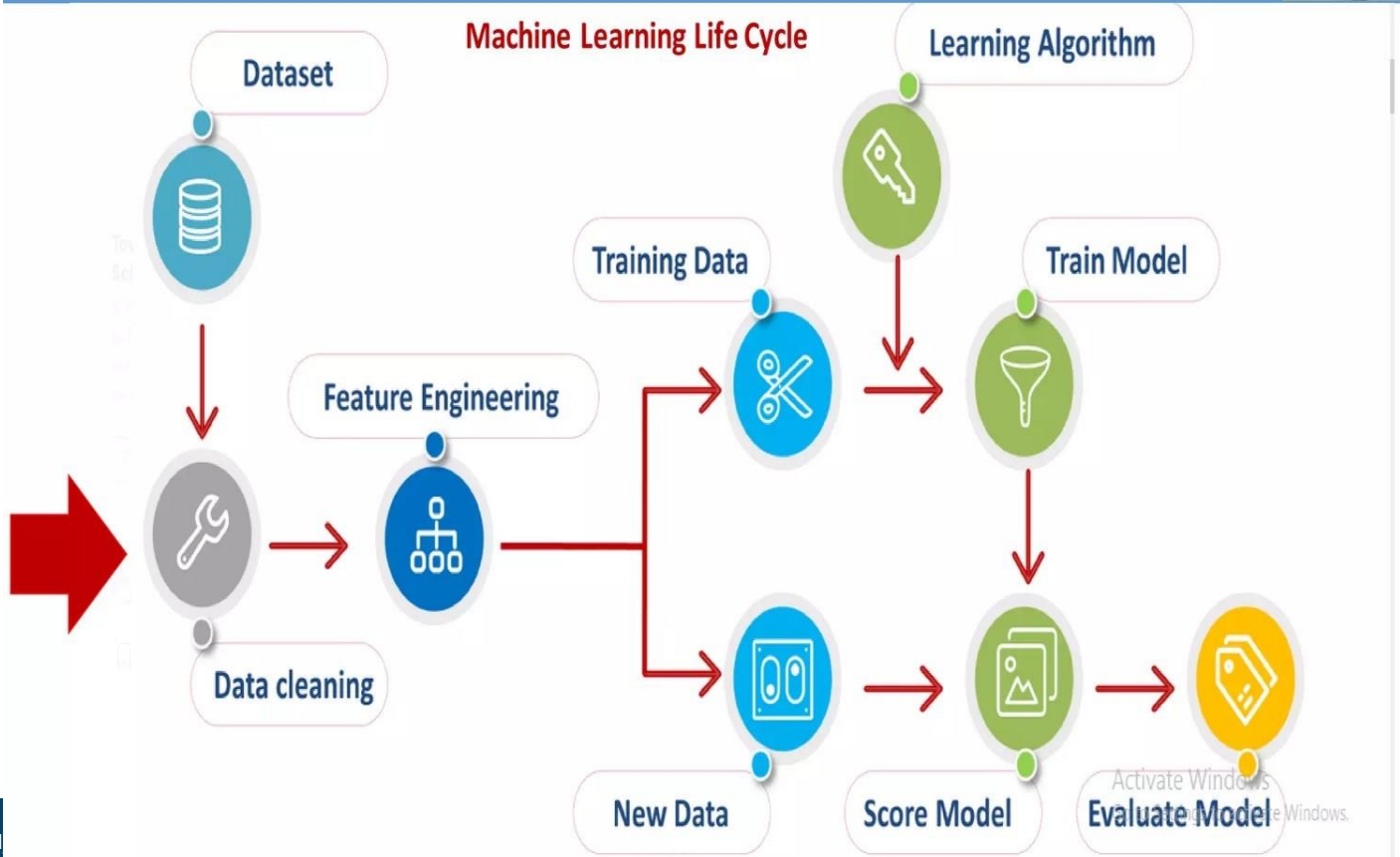
1.1 Define the problem	1.2 Define success	1.3 Identify factors / levers
• Identify the target user • Write the problem statement • Why it matters • Identify the current state and gap(s) in current state • Frame in terms of analytic approach	• State and quantify expected impact • Identify constraints • Translate impact and constraints into metrics – both business and technical • Define success targets for metrics	• Gather domain expertise • Brainstorm possible factors that may influence problem statement

UNIVERSITAS GADJAH MADA

Phase 2 & 3:

DATA UNDERSTANDING & PREPARATION

Where is Data Cleaning used?



Data Pre-processing

- Data preprocessing is an **important** step in ML
- The phrase "**garbage in, garbage out**" is particularly applicable to data mining and machine learning projects.
- It involves transforming raw **data** into an **understandable format**.
- Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
- **Data preprocessing** is a proven method of resolving such issues

Why Data Pre-processing?

- Data in the real world is "dirty"
  - incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

Why Data Pre-processing?

- A manager at ***All Electronics*** and have been charged with analyzing the company's data with respect to the sales at a branch.
- He carefully inspect the **company's database** and data warehouse, identifying dimensions to be included, such as *item*, *price*, *units sold,* and *session* .
- He notice that several of the attributes for various tuples have **no recorded value**. For analysis, he would like to include information.
- In other words, the data he wish to analyze by **machine learning techniques** is incomplete, noisy and inconsistent.

Why Data Pre-processing?

Item	Price	Unit Sold	Session
TV	7200	44	All
Fan	480	27	Summer
Tube light	54	30	All
AC	27000	38	
Fridge		40	Summer
Switches	58	35	
2 mm Wire	520		All
Backup Light	790	48	Winter
Fan Regulator	83	50	All
Bulb	87	37	Rainy Session

What do you mean by data Pre-processing

- It is **cleaning** and explorating data for analysis
- Prepping **data for modeling**
- Modeling in Python requires **numerical** input

- Data preprocessing is a technique that involves transforming raw data into an **understandable format**.
- Data preprocessing is a proven method of **resolving such issues**.

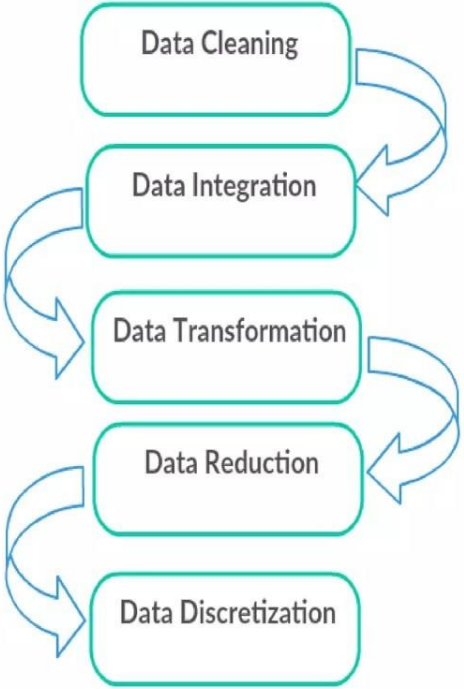
Data Understanding : Relevance of data

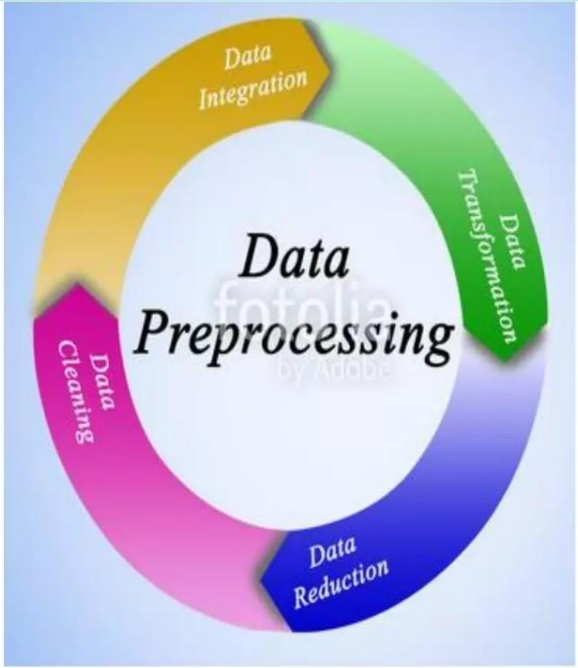
- What data is **available** for the task?
- Is this **data** relevant?
- Is **additional** relevant data available?
- How much **historical data** is available?

Data Understanding: Quantity of data

- Number of instances (records, objects)
- *Rule of thumb: 5,000 or more desired*
- if less, results are less reliable; use special methods (boosting, …)
- Number of attributes (fields)
- *Rule of thumb: for each attribute, 10 or more instances*
- If more fields, use feature reduction and selection
- if very unbalanced, use sampling

Data Pre-processing Steps





Data Pre-processing Steps

Data Cleaning

- **Importance**

Data cleaning is the **number one problem** during working with large data

Data Cleaning Tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration

Data Cleaning: Missing Data

- **Data is not always available**

E.g., while admission filling form by student at the time of admission, he might be don't known local guardian contact number.

- **Missing data may be due to**

✓equipment malfunction

✓inconsistent with other recorded data and thus deleted

✓data not entered due to misunderstanding

✓certain data may not be considered important at the time of entry

✓no register history or changes of the data

✓expansion of data schema

How to Handle Missing Data?

- Ignore the tuple (loss of information)
- Fill in missing values manually: tedious, infeasible?
- Fill in it automatically with

✓a global constant : e.g., unknown, a new class?!

✓**Imputation**: Use the attribute **mean** to fill in the missing value,

✓ Use the most **probable value** to fill in the missing value.

Noisy Data

- Noise: random **error** or variance in a measured variable
- Incorrect attribute values may due to
  - ✓ faulty data collection instruments
  - ✓ data entry problems
  - ✓ data transmission problems
  - ✓ technology limitation
  - ✓ inconsistency in naming convention
- Other data problems which requires data cleaning
  - ✓ duplicate records
  - ✓ incomplete data
  - ✓ inconsistent data

How to handle noisy data?

- **Binning method:**

✓first **sort data** and **partition** into (equi-depth) bins

✓then one can smooth by bin **means**, smooth by bin **median**, smooth by bin **boundaries**, etc.

- **Combined computer and human inspection**

✓detect suspicious values and check by human

Data Integration

Data integration:

Its **combines data** from multiple sources

- **Schema integration**

Integrate metadata from different sources

Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#

- **Detecting and resolving data value conflicts**
- for the same real world entity, attribute values from different sources are different, e.g., different scales, metric vs. British units
- **Removing duplicates and redundant data**

Data Transformation

Data Transformation

- Smoothing: **remove noise** from data
- **Normalization**: scaled to fall within a small, specified range
- Attribute/feature construction

✓ New attributes constructed from the given ones

- **Aggregation**: summarization

✓ Integrate data from different sources (tables)

Data Reduction

- **Data is too big to work with**

✓ Too many instances

✓ too many features (attributes)

✓**Data Reduction**

✓ Obtain a reduced representation of the data set that is much smaller in **volume** but yet produce the same (or almost the same) **analytical** results (easily said but difficult to do)

- **Data reduction strategies**

✓ Dimensionality reduction – remove unimportant attributes

✓ Aggregation and clustering -

✓ Remove redundant or close associated ones

✓ Sampling

Data Discretization

- Discretization is a process that transforms **quantitative** data into **qualitative** data.
- It significantly improve the quality of **discovering knowledge**.
- It **reduces** the running time of various machine learning tasks such as association rule discovery, classification, clustering and prediction.
- It reduce the number of values for a given continuous attribute by dividing the range of the attribute into **intervals**.
- Interval labels can then be used to **replace** actual data values

Data Discretization

Discretization - Example

- **Example: discretizing the "Humidity" attribute using 3 bins.**

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Low = 60-69
Normal = 70-79
High = 80+

Humidity
High
High
Normal
High
High
Normal
Low
High
Normal
High
Normal
High
Normal
High

UNIVERSITAS GADJAH MADA

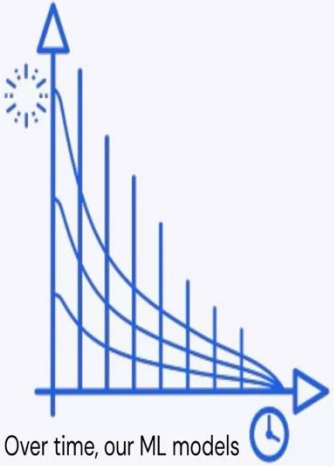
Phase 3 & 4:

MODEL & EVALUATION

Model Drift

Machine Learning Model Drift

The fundamental assumption in any machine learning model is that the data and logic used actually mimics the real-world.



Over time, our ML models make worse predictions. Also called "decay".

Don’t Get Too Caught Up In Terminology

ML Drift, Model Drift, Model Decay, Prediction Drift = Your predictions are getting worse

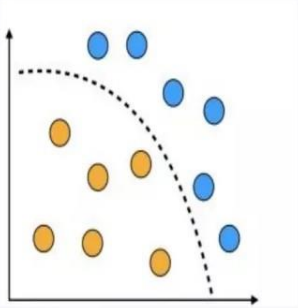
Experience, types, causes, and indicators of drift are sometimes used together, overlap, and don’t have direct mappings to each other.

Multiple types of drift can happen at the same time.

How We Experience ML Drift

Not really drift but can appear to be

Key Types of Drift

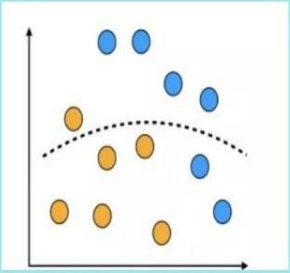


Training data with decision boundary

$P(Y|X)$
Probability of y output given x input

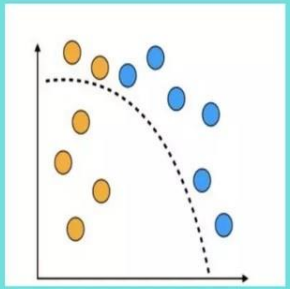
Concept Drift

$P(Y|X)$ Changes



- Reality/behavioral change
- Relationships change, not the input

Data Drift*

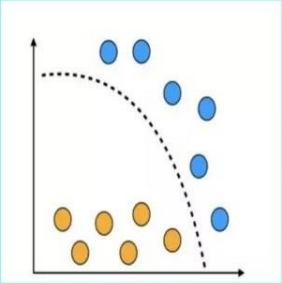


Label Drift

- Output data shifts
- $P(Y)$ Changes

Feature Drift

- Input data shifts
- $P(X)$ Changes



Virtual Drift

Data changes but boundary still works

Drift Examples for a Loan Application Model

Concept Drift

An income level that was earlier considered creditworthy is now considered riskier.

Label Drift

A larger proportion of credit-worthy applications start showing up.

Feature Drift

Incomes of most applicants increase or decrease. Or you suddenly get more application from one region.

Triggers of ML Model Drift

Real Data Distribution Change

- Label or feature distribution changes e.g. product launch in a new market
- Concept can change. e.g. a competitor launching a new service

May require a new model

Data Integrity Issues

- Correct data enters at source but faulty data engineering. E.g. debt-to-income values & age values are swapped in the input.
- Incorrect data enters at source. E.g., due to a front-end issue, a website form accepts leaving a field blank.

Detecting Issues

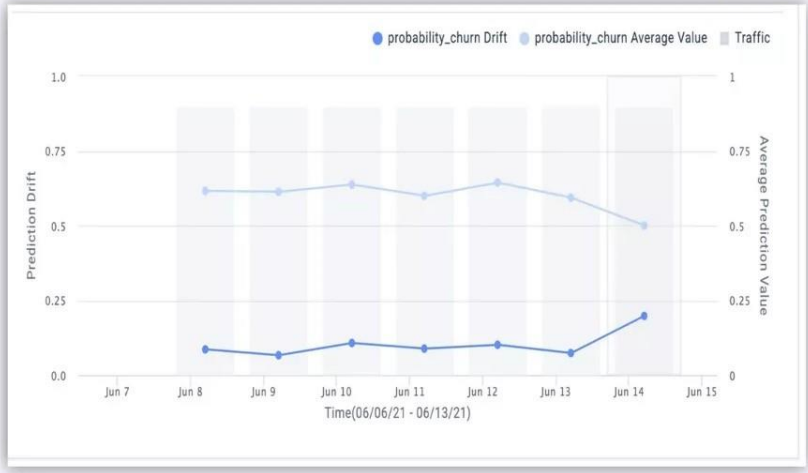
Performance Monitoring & Supervised Learning

Data Drift Monitoring & Unsupervised Learning

Data Integrity Monitoring

Performance Monitoring & Supervised Learning

Works well if you have ground truth/labels!!



Monitor performance metrics

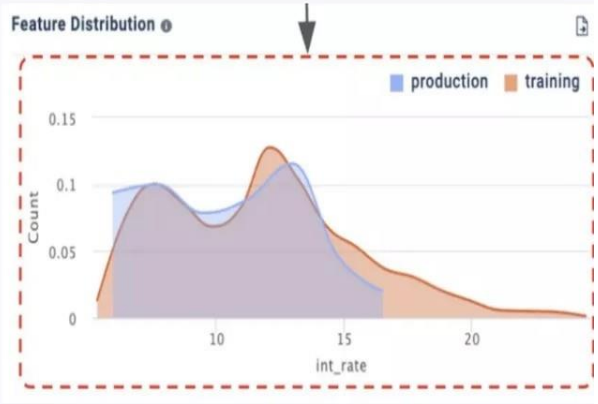
- Statistical measures
- Accuracy, precision, FPR, AUC etc.

Supervised learning methods – Ref "A Survey of Concept Drift Adaptation"

- Sequential analysis (SPRT – CUMSUM & PH) – tune alarms on false positives
- Statistical process control (SPC) – rate of change
- Monitoring 2 distributions (ADWIN) – more precise, more overhead

Data Drift Monitoring & Unsupervised Learning

Monitor Statistical Distributions Metrics



- **Population Stability Index (PSI)** compares current scoring variable to the predicted probability from training data
- **Kullback–Leibler (KL)** divergence measures the differences of one probability distribution to a reference probability distribution
- **Jensen-Shannon (JS)** based on KL, measuring the similarity between two probability distributions - but is notably symmetric and finite.
- **Kolmogorov-Smirnov test (or KS test)** quantifies a distance between the distribution of the sample and the cumulative distribution of the reference (non-parametric)

Data Drift Monitoring & Unsupervised Learning

Un / Semi-supervised Learning

An overview of unsupervised drift detection methods

- Can be more accurate
- Online methods look at each instance (batch methods are more efficient)
- Most are global oriented
- May miss gradual drift / sensitivity issues
- Not intuitive for explaining

Method	Drift type	Application
NN-DVI	Local	Independent
FAAD	Local/global	Sequence anomaly
UDetect	Global	Independent
SQSI-IS	Global	Independent
MD3	Global	Independent
MD3-EGM	Global	Independent
PDetect	Global	Adversarial attacks
DDAL	Global	Independent
IKS-bdd	Global	Independent
CD-TDS	Local/global	Transactional data
OMV-PHT	Global	Independent
NM-DDM	Global	Independent
Plover	Global	Independent
SAND	Global	Independent
DSDD	Global	Independent
DbDDA	Global	Independent

Data Integrity & Outlier Monitoring

Data errors slowly degrade performance and can look like real drift

- Missing values
- Range and Type mismatch
- Schema mismatch

Changes in the business (new cataloged products, revoked pricing, etc.)

Broken data pipeline due to bugs or API updates

Getting to the Root Cause

Attribute to drift in features

Account for feature importance

Analyze affected traffic

Fix It!

Retrain
new data and/or
relabel old data

Adapt/Augment
model behavior, weighting,
biz logic, data collections

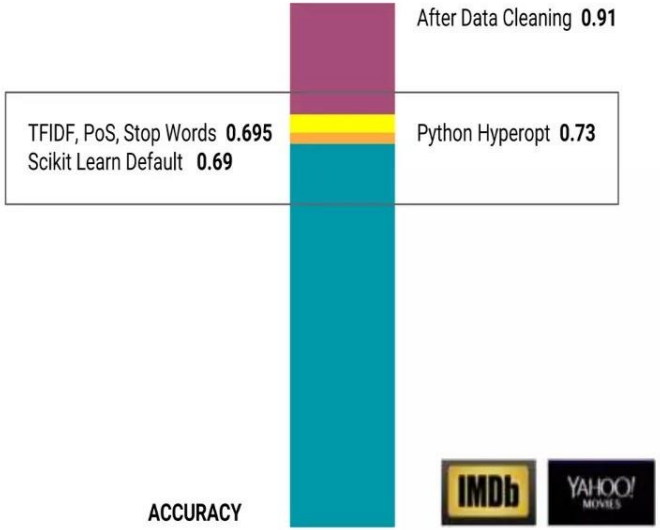
Model Mgmt
Archive/schedule model,
ensemble balancing

UNIVERSITAS GADJAH MADA

Phase 5 & 6

QUALITY ASSURANCE FOR ML

Why Quality Data Matters?

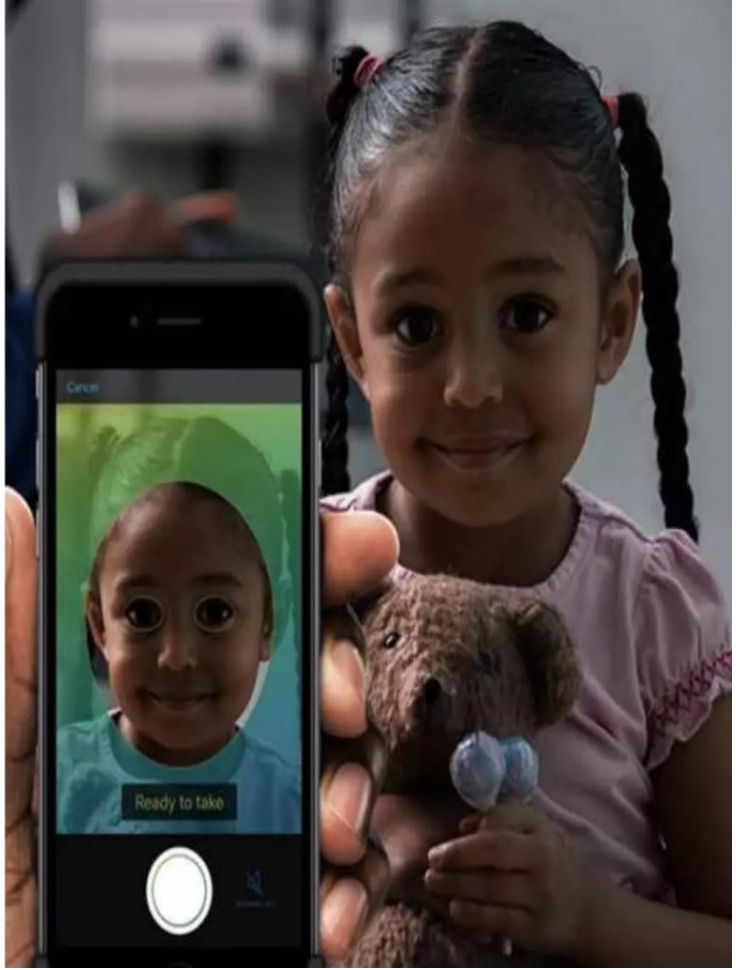


Sigmod2016
Sanjay Krishnan (UC Berkeley)
And Jiannan Wang (Simon Fraser U.)
https://sigmod2016.org/sigmod_tutorial1.shtml

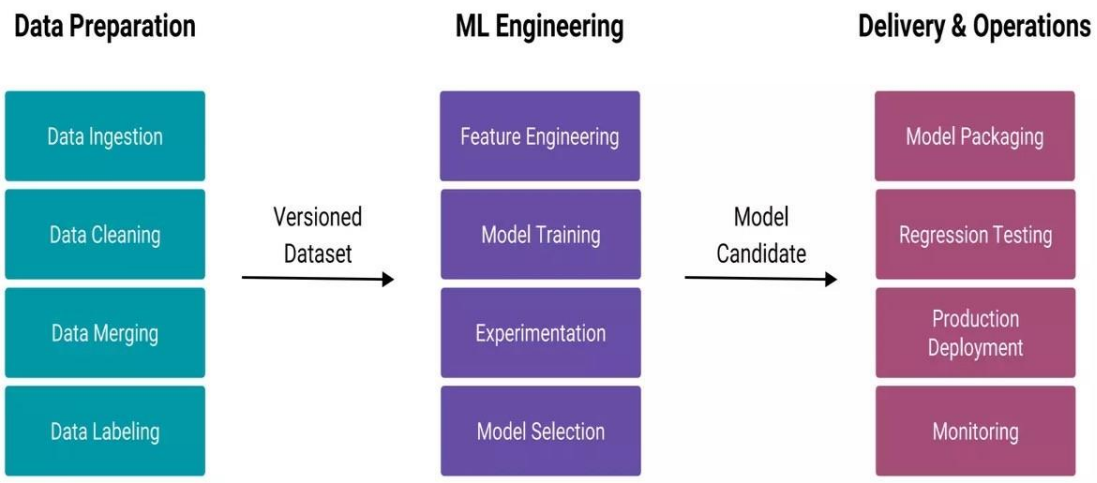
GoCheck Kids Case Study

End-to-end deep learning image classification models to detect child gaze, strabismus, crescent, and dark iris/pupil population.

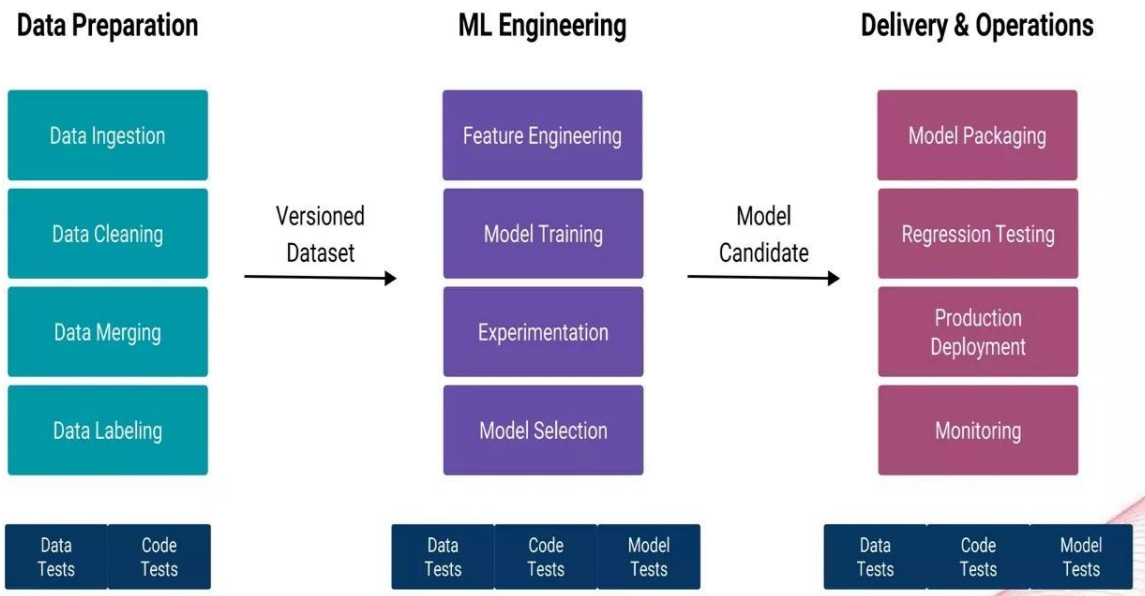
	Before	**After Data QA**
Precision	32%	**40%**
Recall	89%	**91%**
FPR	19%	**17%**
PR AUC	57%	**76%**



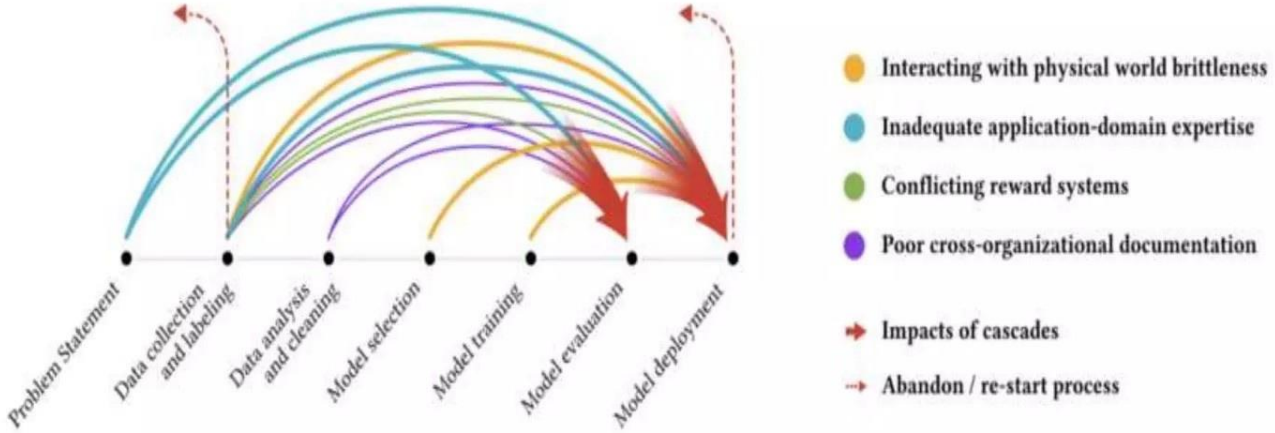
Machine Learning Lifecycle



All Stages of ML Lifecycle Require QA



Error Cascades



How: Practical Challenges and Solutions

Common Challenge #1:

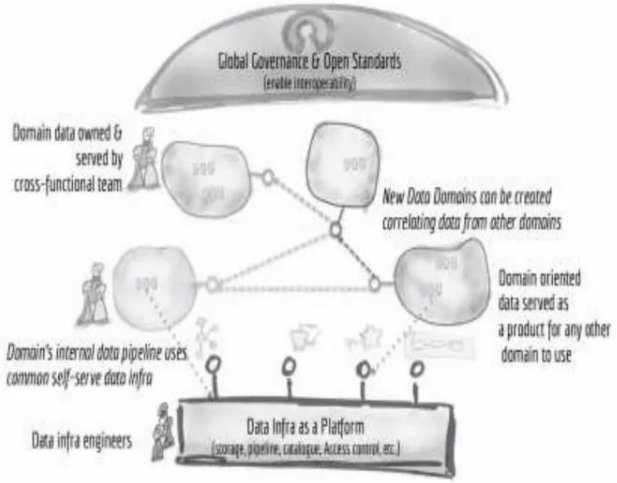
How to find & access the data I trust?

1. Data is **scattered** across multiple data sources and technologies: RDMS, DWH, Data Lakes, Blobs
2. Data **ownership** is not clear
3. Data **requirements and SLAs** are not clear
4. **Metadata** is not discoverable
5. **As a result, all investments into Data and ML are killed by data access and discoverability issues**

Solution: Migrate to Data Mesh

Data Mesh is in the convergence of Distributed Domain-Driven Architecture, Self-Serve Platform Design, and Product Thinking with Data

- Brings data closer to Domain Context
- Introduces the concept of Data as a Product and all appropriate data contracts
- Sorts out data ownership issues



Invest into Global Data Catalog

The solution to answer questions like:

- Does this data exist? Where is it?
- What is the source of truth of the data?
- Who and/or which team is the owner?
- Who are the users of the data?
- Are there existing assets I can reuse?
- Can I trust this data?

Amundsen

Data Hub

Apache Atlas

* There are no established leaders

* Commercial vendors are not listed

Common Challenge #2:

How to get started with QA for Data and ML?

1. What exactly to test?
2. Who should test (Traditional QA, Data Engs, ML Engs, Analysts)?
3. What tools to use?
4. **As a result, low productivity of ML Engineers having to deal with data quality issues.**

Data: What to Test

Default **data quality** checks:

- Duplicates
- Missing values
- Syntax errors
- Format errors
- Semantic errors
- Integrity

Data: What to Test

Default **data quality** checks:

- Duplicates
- Missing values
- Syntax errors
- Format errors
- Semantic errors
- Integrity checks

Advanced **unsupervised** methods:

- Distribution tests
- KS, Chi-squared tests
- Outlier detection with AutoML
- Auto Constraints suggestion
- Data Profiling for Complex Dependencies

Unsupervised Constraints Generation

Use cases:

- existing data with poor documentation or schema
- rapidly evolving data
- rich structure
- starting from scratch

1. Compute data profiles/summaries

2. Generate checks on:

- types
- completeness
- ranges
- uniqueness
- distributions

Extensible:

- e.g., conventions on column naming

3. Evaluate on holdout subset

4. Review and add to test suites

Data Testing: Available Tools

- Deequ
- GreatExpectations
- Tensorflow Data Validation
- dbt

* Commercial vendors are not listed

- expect_column_values_to_not_be_null
- expect_column_values_to_match_regex
- expect_column_values_to_be_unique
- expect_column_values_to_match_strftime_format
- expect_table_row_count_to_be_between
- expect_column_median_to_be_between

Model Testing

Model Testing: Analyzing Input and Output Datasets

	date	tempe rature	weather	event	day of week	month	sold	sold prediction
Training set (January - October)	01.01	-1°C	clear		Tue	Jan	x	
	02.01	-5°C	rain	festival	Wed	Jan	254	
	03.01	0°C	snow		Thu	Jan	189	
	...	...	...	...	...	...	...	
Test set (November - December)	28.12	3°C	snow		Mon	Dec	165	150
	29.12	3°C	clear	football	Tue	Dec	203	210
	30.12	6°C	drizzle		Wed	Dec	240	243

Model Testing: Datasets Are Test Suites with Test Cases

- Golden UAT datasets
- Security datasets
- Production traffic replay
- Regression datasets
- Datasets for bias
- Datasets for edge cases

Model Testing: Bias

Bias is considered to be a disproportionate inclination or prejudice for or against an **idea** or **thing**.

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	–
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	GRILL
TOOL	TONGS
PLACE	OUTSIDE

COOKING

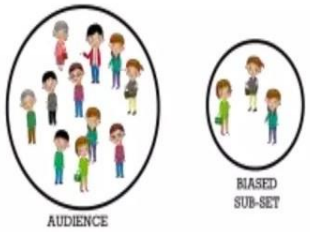
ROLE	VALUE
AGENT	WOMAN
FOOD	VEGETABLES
HEAT	STOVE
TOOL	TONGS
PLACE	KITCHEN

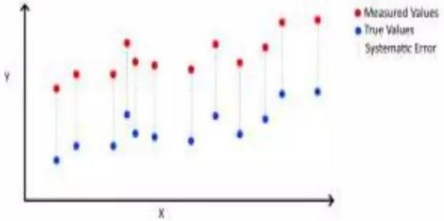
COOKING

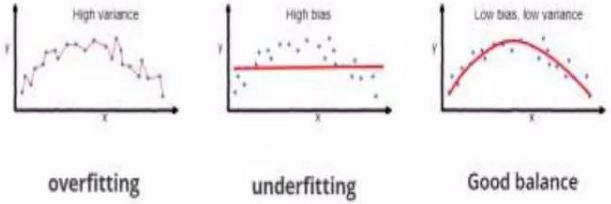
ROLE	VALUE
AGENT	MAN
FOOD	–
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

10+ Bias Types

- **Selection Bias** – The selection of data in such a way that the sample is not representative of the population
- **The Framing Effect** – Annotation questions that are constructed with a particular slant
- **Systematic Bias** – Consistent and repeatable error.
- **Outlier Data, Missing Values, Filtering Data**
- Bias / Variance Trade off
- Personal Perception Bias







SECURITY IN ML

FAILURE MODES IN MACHINE LEARNING

- **Intentional failures** where the failure is caused by an active adversary attempting to subvert the system to attain their goals – either to misclassify the result, infer private training data, or to steal the underlying algorithm

- **Unintentional failures** where the failure is because an ML system produces a formally correct but completely unsafe outcome [1]

SOME NECESSARY CONCEPTS

"C.I.A. TRIAD"

- Confidentiality
- Integrity
- Availability

BLACKBOX/WHITEBOX ATTACKS

- Knowledge and clearance about the ML system

FAILURE MODES IN MACHINE LEARNING

Intended Failures	
#	**Attack**
1	Perturbation attack
2	Poisoning attack
3	Model Inversion
4	Membership Inference
5	Model Stealing
6	*Reprogramming ML system*
7	*Adversarial Example in Physical Domain*
8	*Malicious ML provider recovering training data*
9	Attacking the ML supply chain
10	Backdoor ML
11	Exploit Software Dependencies

Unintended Failures	
#	**Attack**
12	*Reward Hacking*
13	*Side Effects*
14	Distributional shifts
15	Natural Adversarial Examples
16	*Common Corruption*
17	*Incomplete Testing*

Out of scope (Sorry...)
Will be only discussed (No Demo)...
Present in the demo! (Yesssss...)

INTENTIONAL ATTACKS

PERTURBATION ATTACK

In perturbation style attacks, the attacker stealthily modifies the query to get a desired response [1]

COUNTERMEASURES

- Inputs monitoring: Histograms, Anomaly detection, confidence bands
- Coherence check in model deployment
- NLP: Word distribution, TF-IDF score shift
- Computer Vision: Color density histogram, Human-in-the-Loop

INTENTIONAL ATTACKS

EXPLOIT SOFTWARE DEPENDENCIES

In this attack, the attacker does not manipulate the algorithms. Instead, exploits the software vulnerabilities such as a buffer overflow within dependencies of the system.[1]

COUNTERMEASURES

- Dependency management tools can help (Snyk, Poetry, Bandit, Safety, requires.io)
- Regular security scanning
- If it's possible, follow security reports
- Avoid Wrappers and pre-built environments
- Use less dependencies as possible (*i.e. avoid the 5 tonnes Sandwich*)

UNINTENDED FAILURES

DISTRIBUTIONAL SHIFTS

The system is tested in one kind of environment, but is unable to adapt to changes in other kinds of environment [1]

COUNTERMEASURES

- Monitoring Trinity: Inputs, Model Outputs, KPIs
- Input monitoring: Histograms, Accepted Variational Bands
- Keep-up with exogenous aspects (*i.e.* respect Causality)
- Adaptive Shift Alerting

UNINTENDED FAILURES

NATURAL ADVERSARIAL EXAMPLES

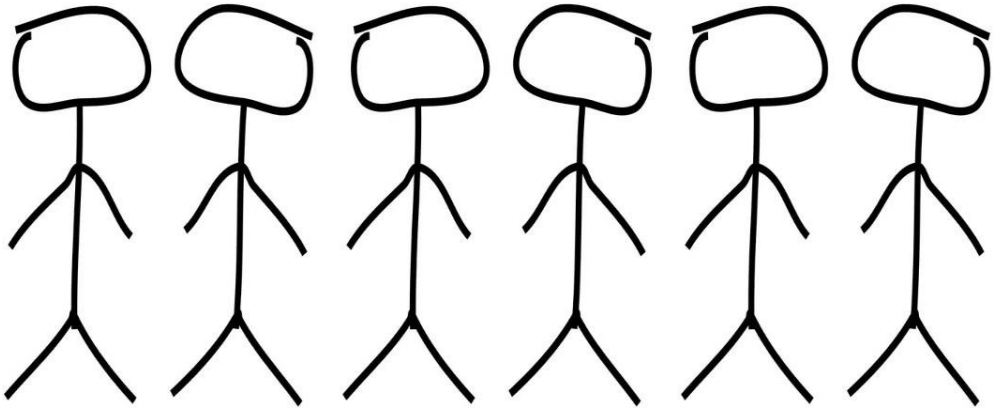
The system misrecognizes an input that was found using hard negative mining [1][2]

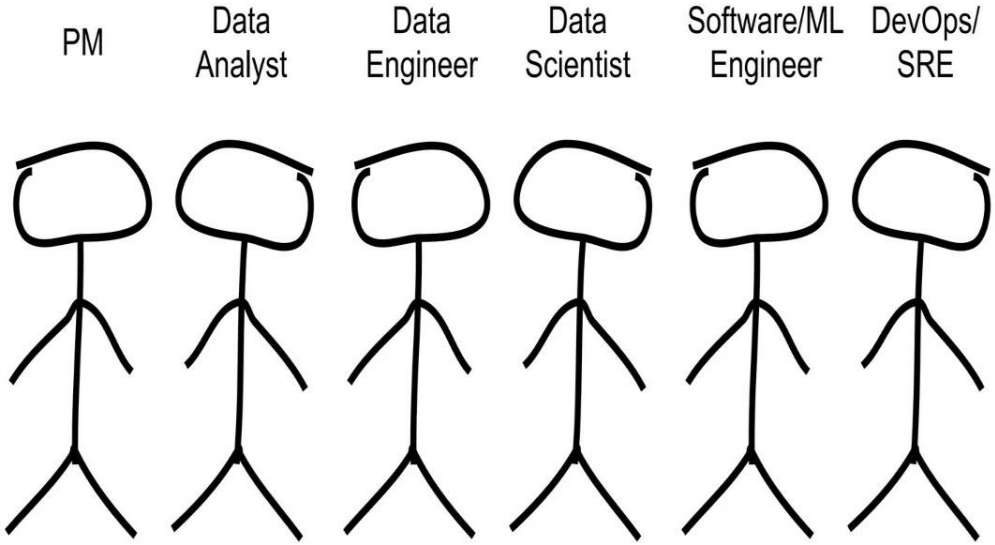
COUNTERMEASURES

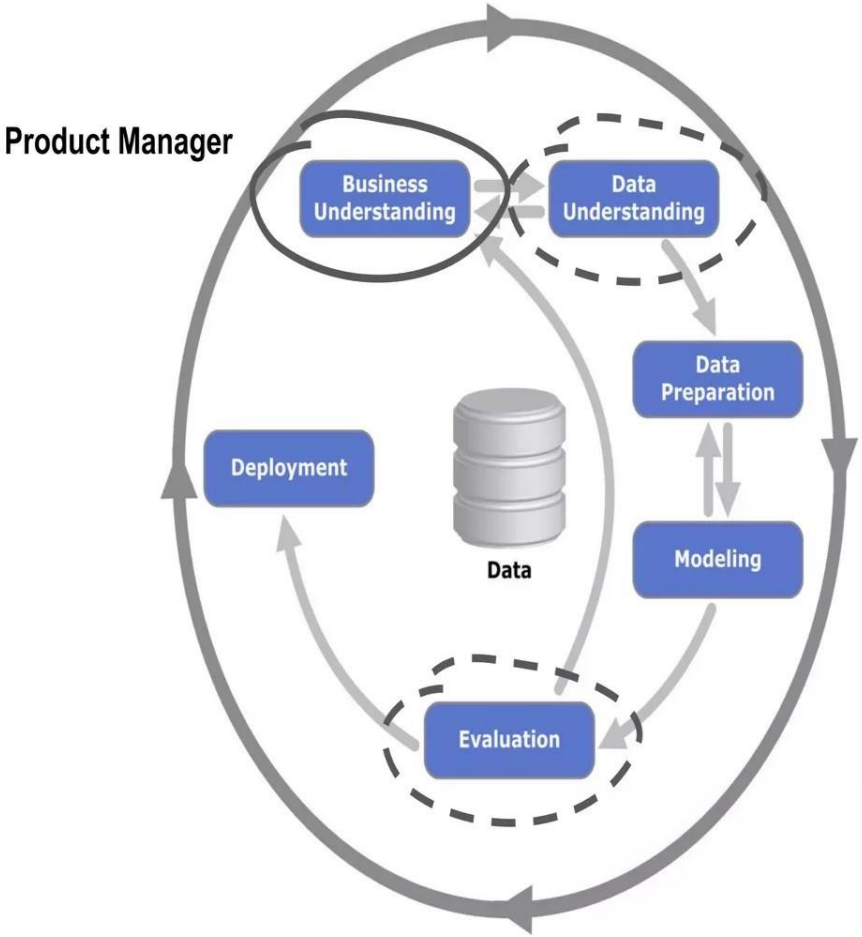
- Extra layers of checking
- Diversity in labeled data
- Good Data Augmentation strategy
- If it's possible, check the model behavior against corner cases to understand the limits
- Error analysis

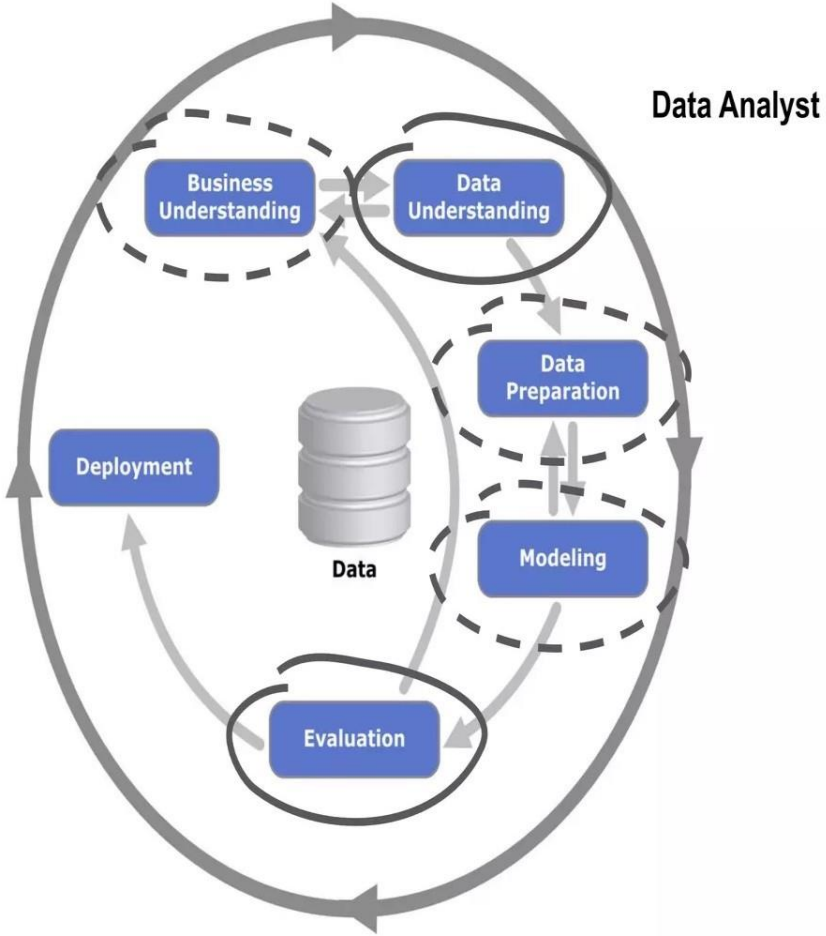
BUILDING TEAM FOR ML

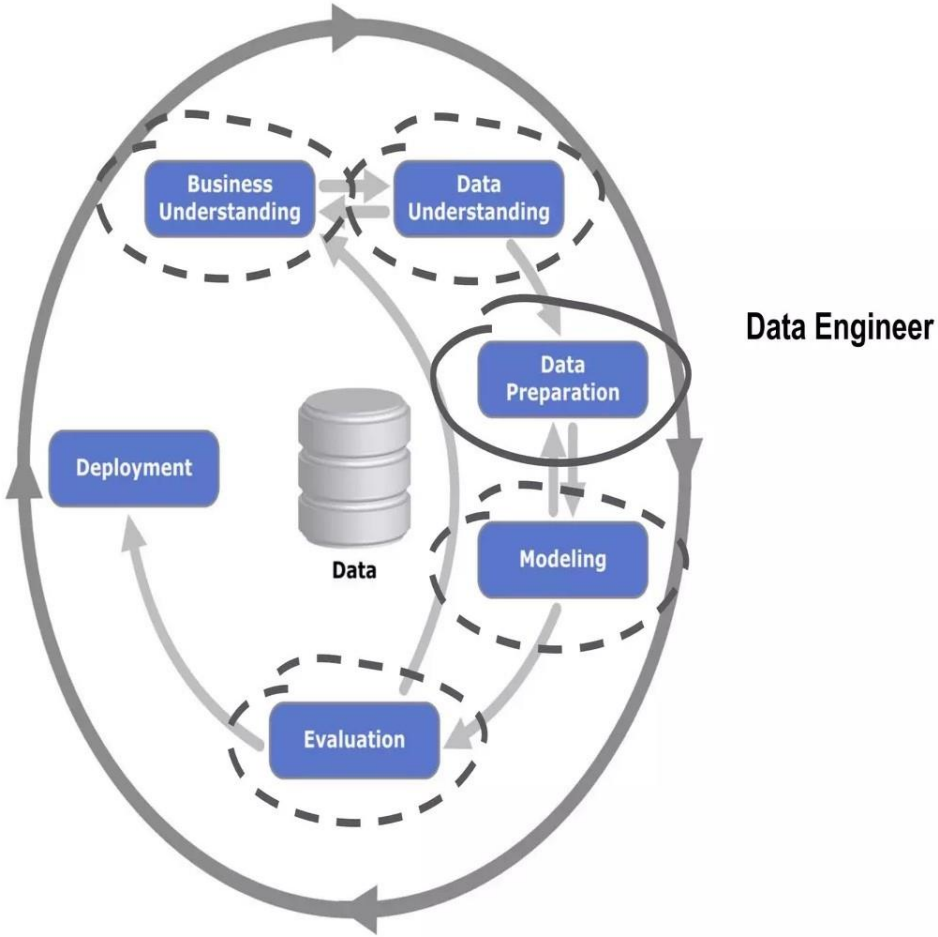
We need a team for that!

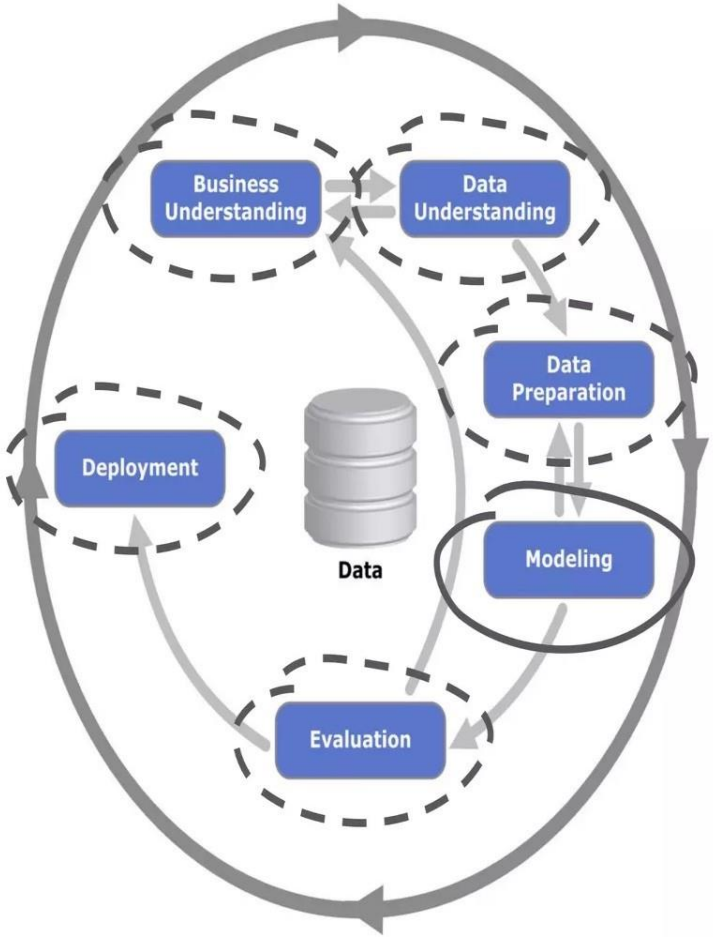



PM
Data Analyst
Data Engineer
Data Scientist
Software/ML Engineer
DevOps/ SRE

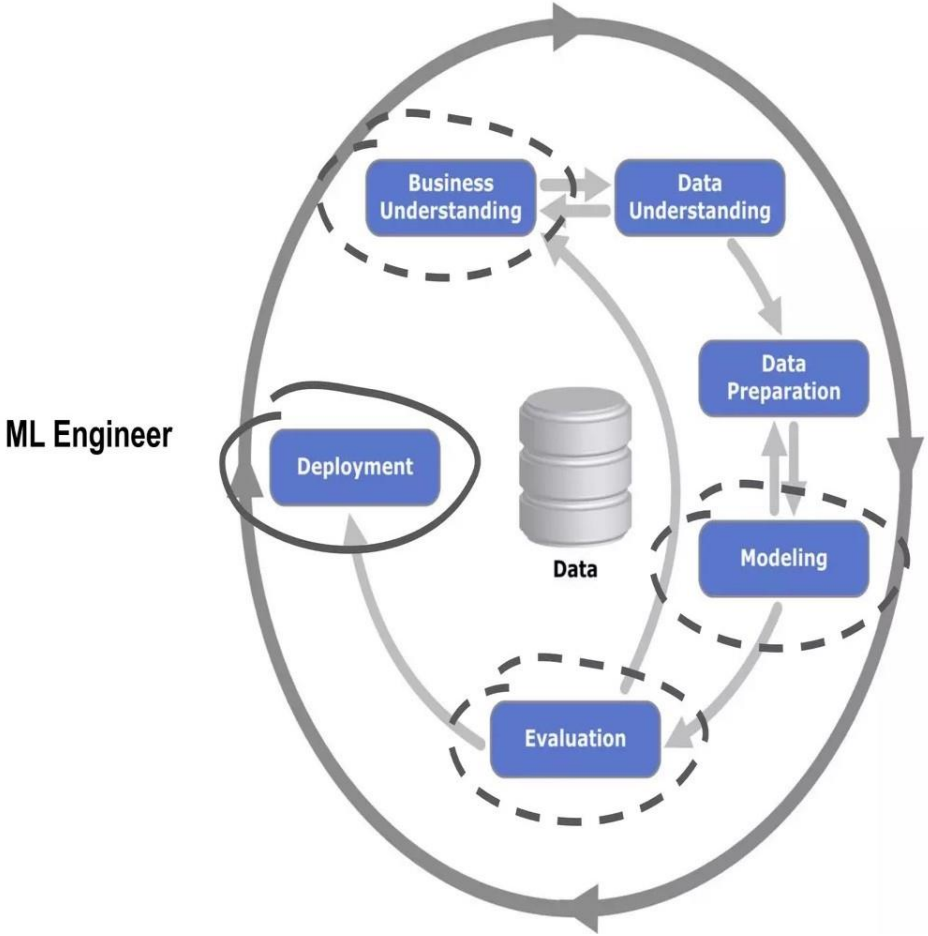

Product Manager
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

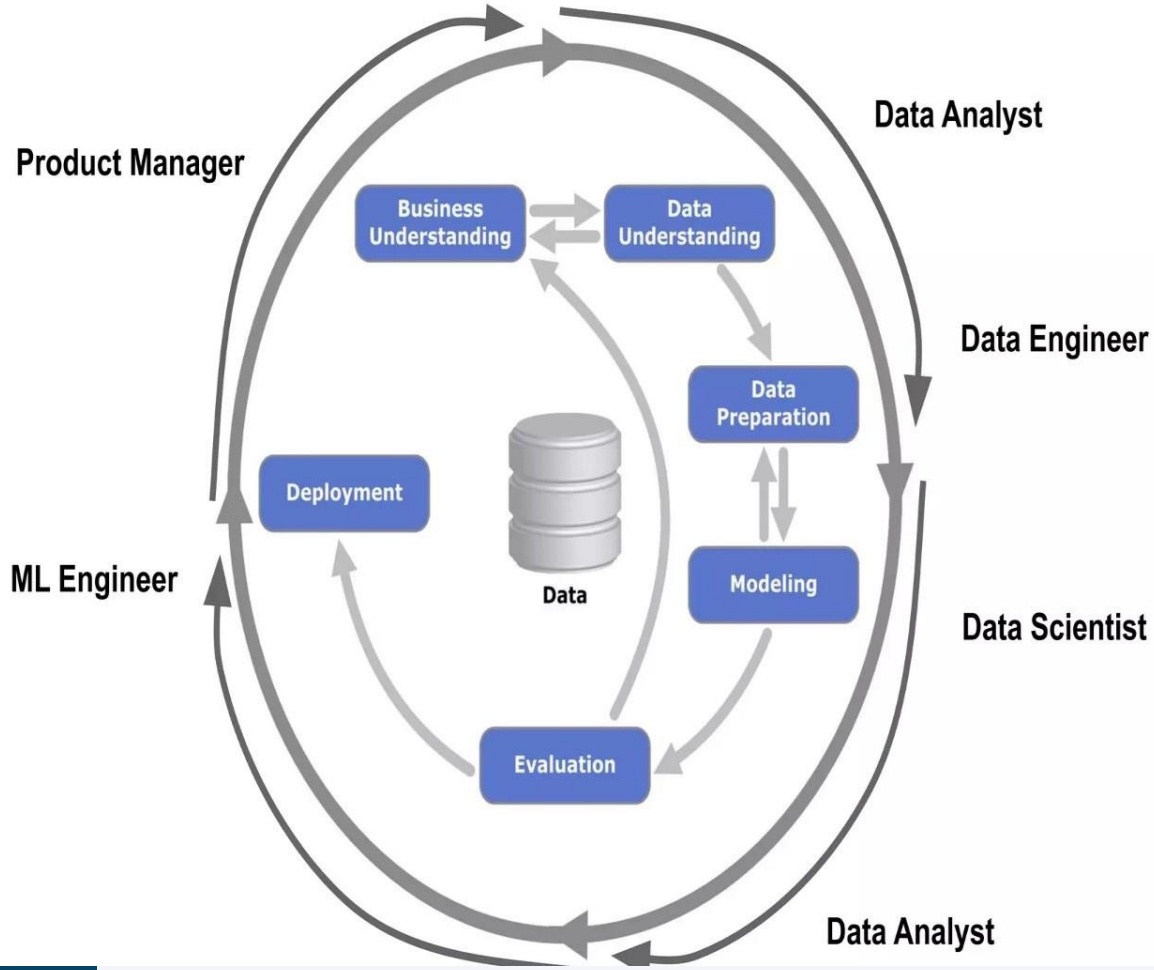

Data Analyst
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

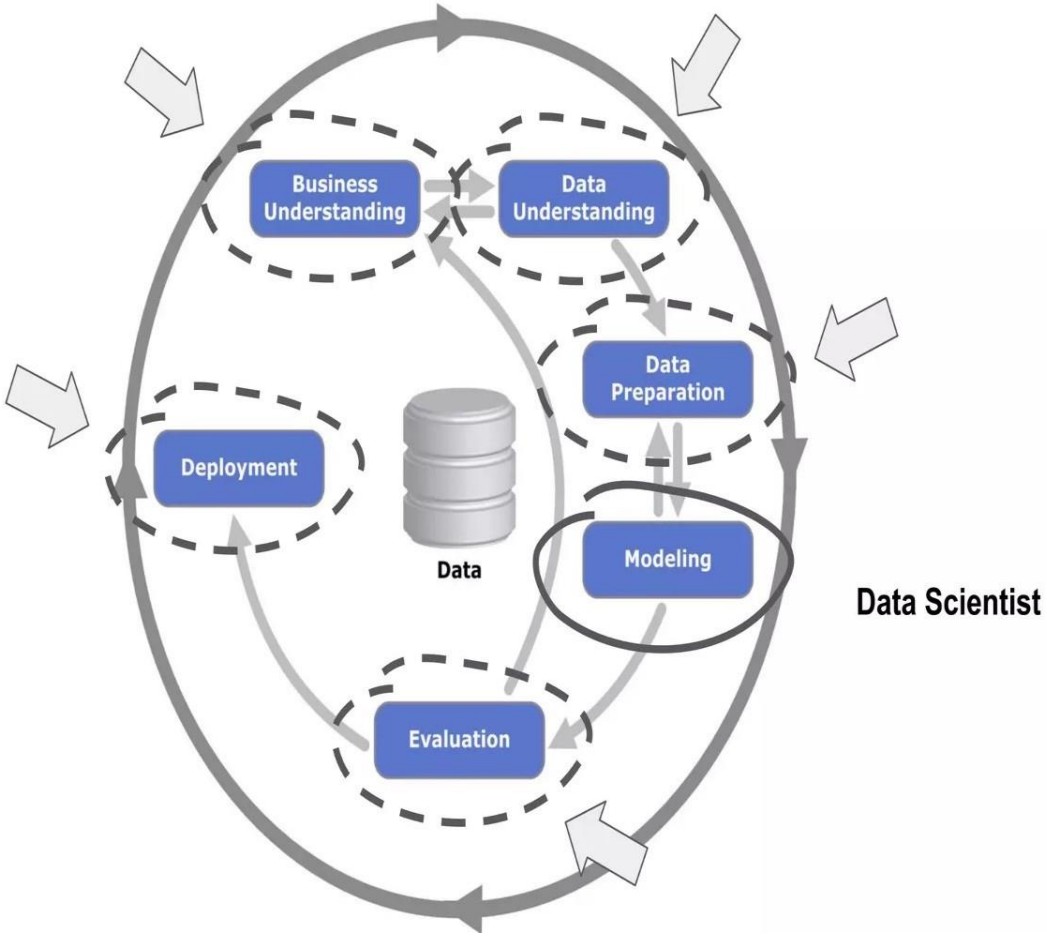

Business Understanding
Data Understanding
Data Preparation
Data Engineer
Deployment
Data
Modeling
Evaluation



Data Scientist


ML Engineer
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data


Product Manager
Data Analyst
Business Understanding
Data Understanding
Data Engineer
Data Preparation
Deployment
Data
ML Engineer
Modeling
Data Scientist
Evaluation
Data Analyst


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data
Data Scientist

80/20 rule

- Break down the role into core areas and skills
- Order the skills by importance
- Pick the most important ones
- Practice practice practice

Product management

- Strategy
- UX & Design
- Communication
- Planning
- Evaluation

- Requirement gathering
- Prioritization
- Stakeholder management

Data engineering

- SQL Databases
- NoSQL Databases
- Stream processing
- Batch processing
- Data pipelines

- MySQL
- AWS (S3, Kinesis)
- Spark
- Airflow

Backend engineering

- SQL & NoSQL Databases
- CS fundamentals
- Languages and frameworks
- Web services
- Best practices

- Python
- Docker
- Tests
- Clean code

DevOps

- Infrastructure
- Automation
- Monitoring
- Reliability

- AWS
- Kubernetes
- Terraform

Thank you

Terima Kasih

Matur nuwun