AI in Practice

Muhammad Ghifary, PhD
Head of Artificial Intelligence

June 2020

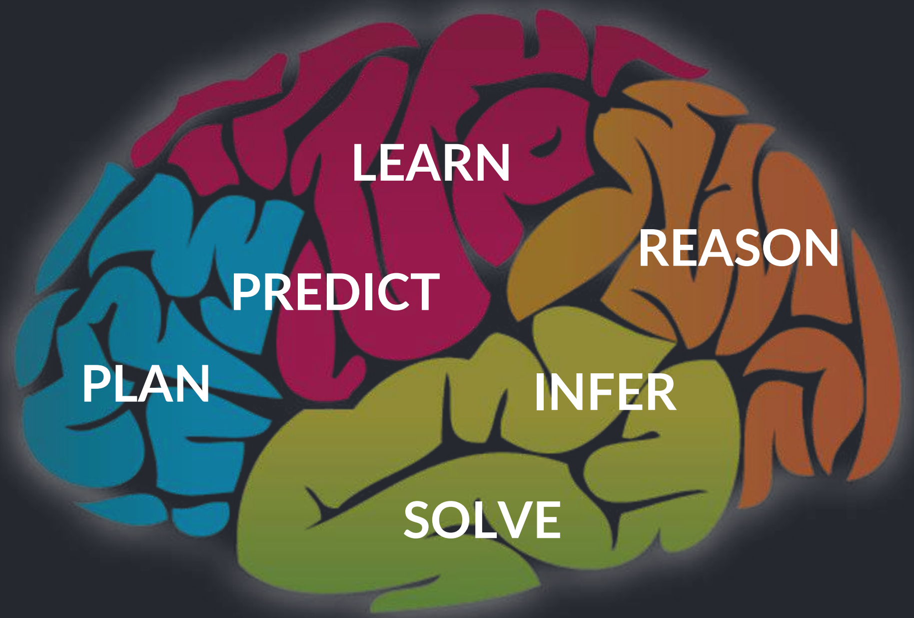
Overview

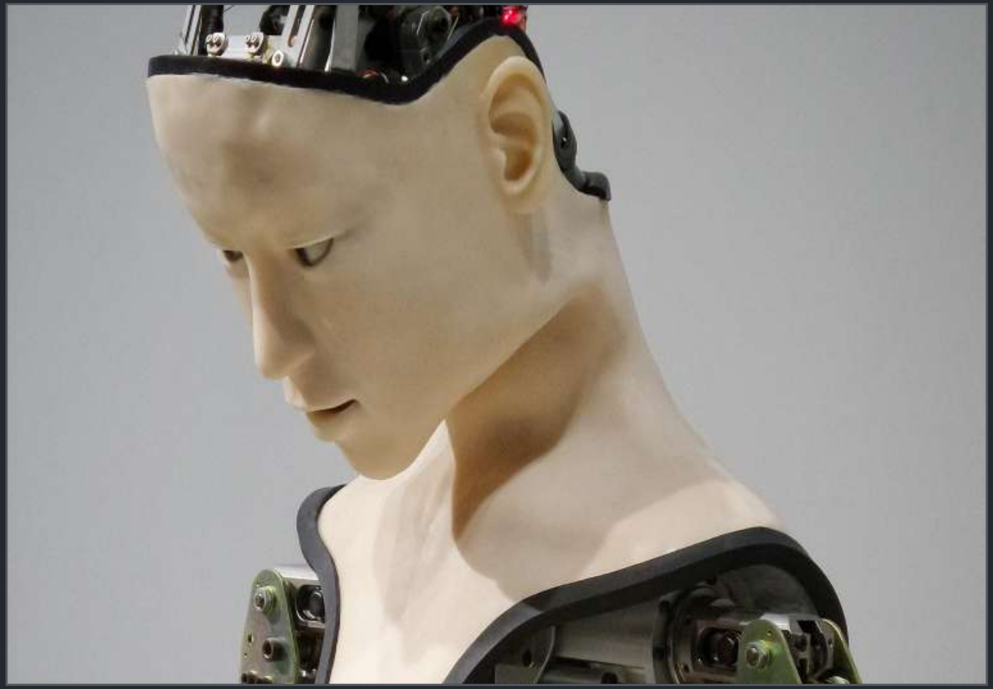
1. What is AI?
2. AI/ML in Practice
3. AI in Bukalapak
4. AI Organizations and Skill Set

What is AI?

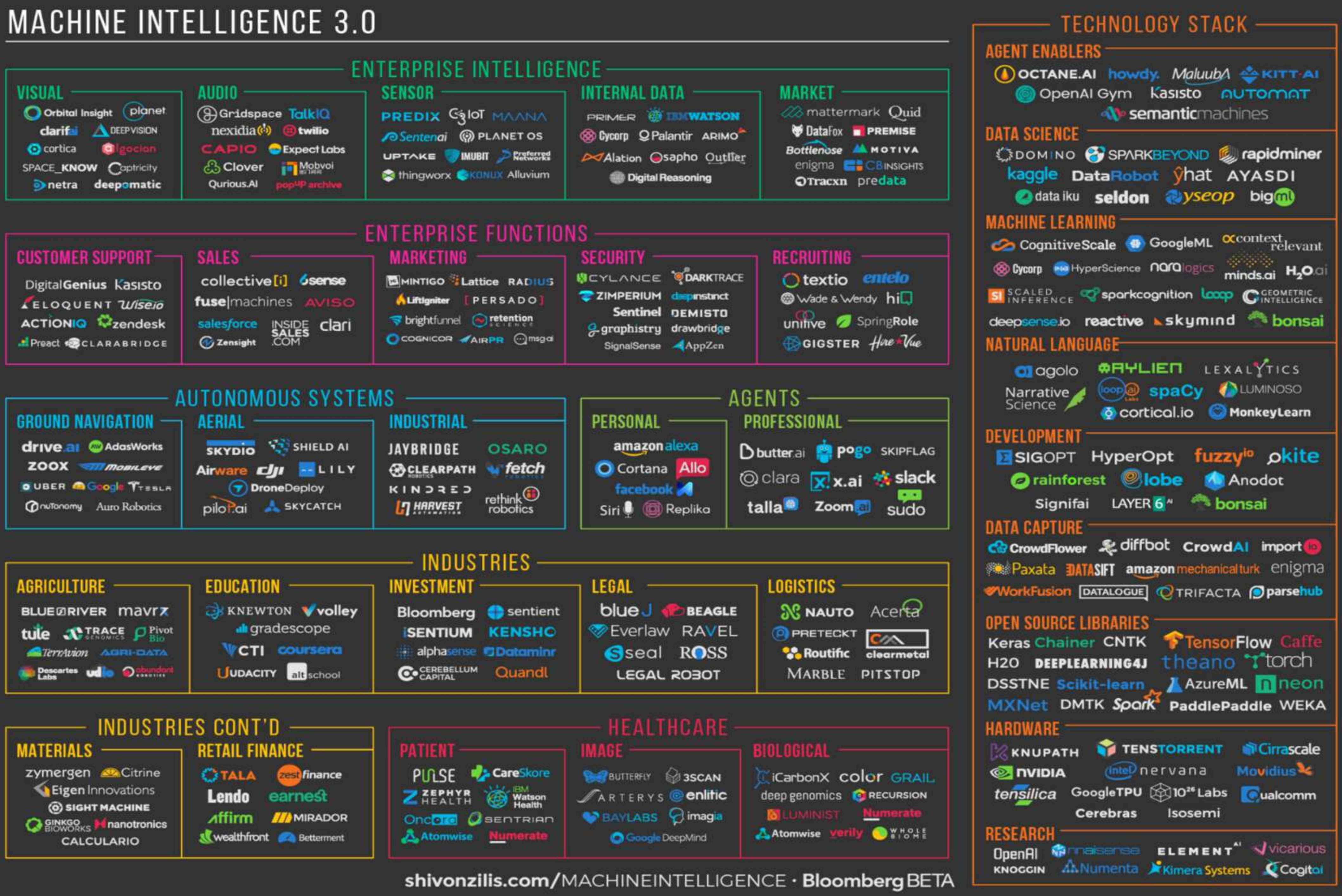
Artificial Intelligence (AI)

The creation of machines that mimic human intelligence



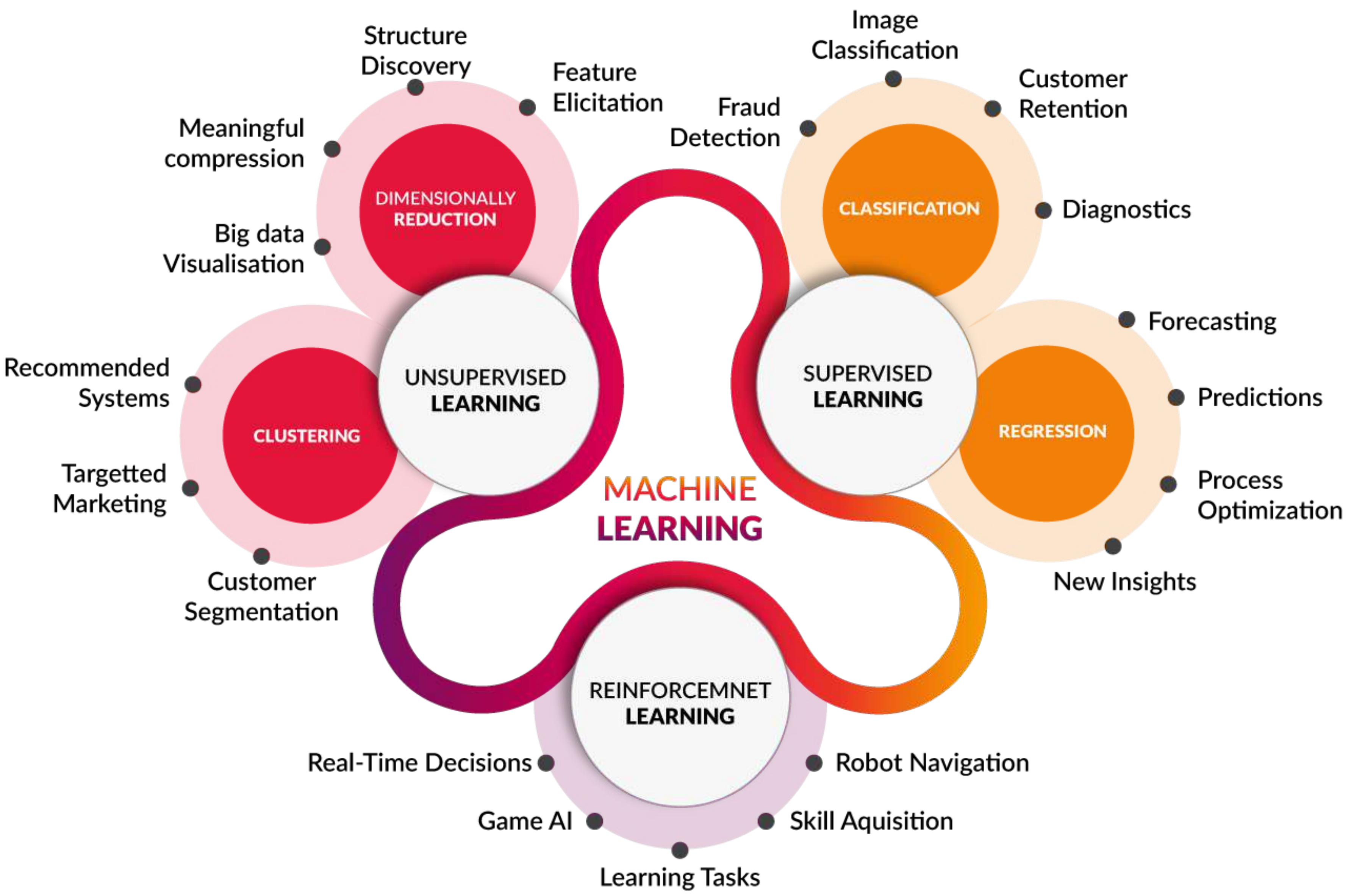


AI



Bukalapak

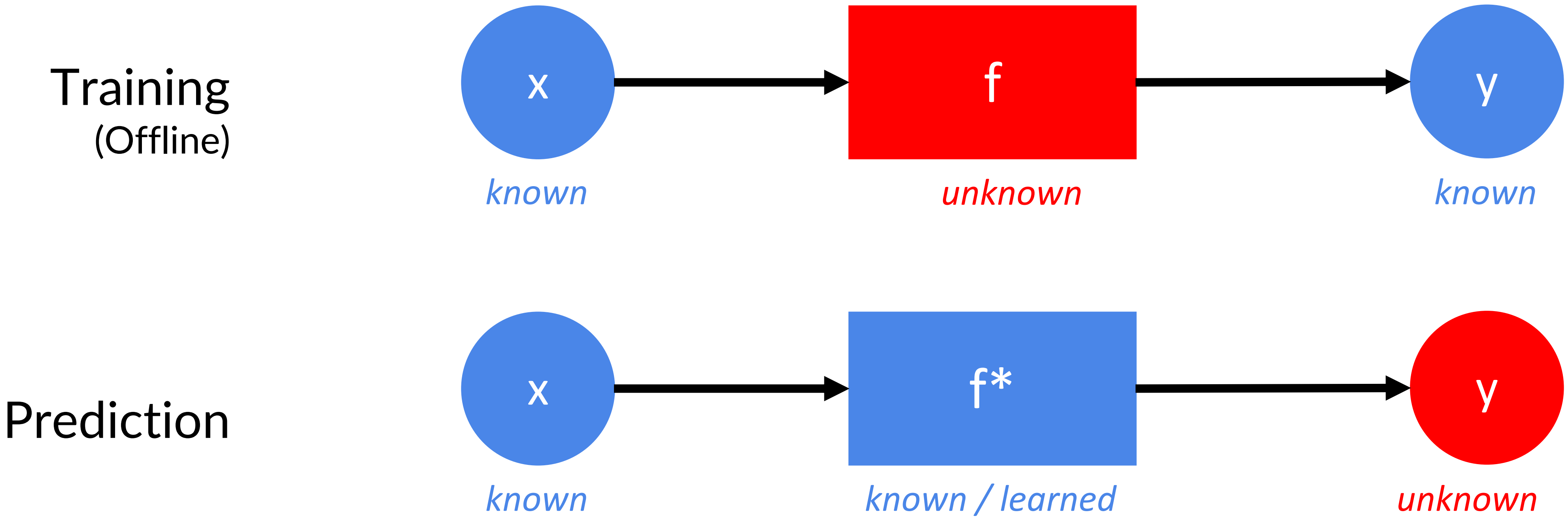
Machine Learning - programming intelligence into computers, through learning from data.



Source: https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

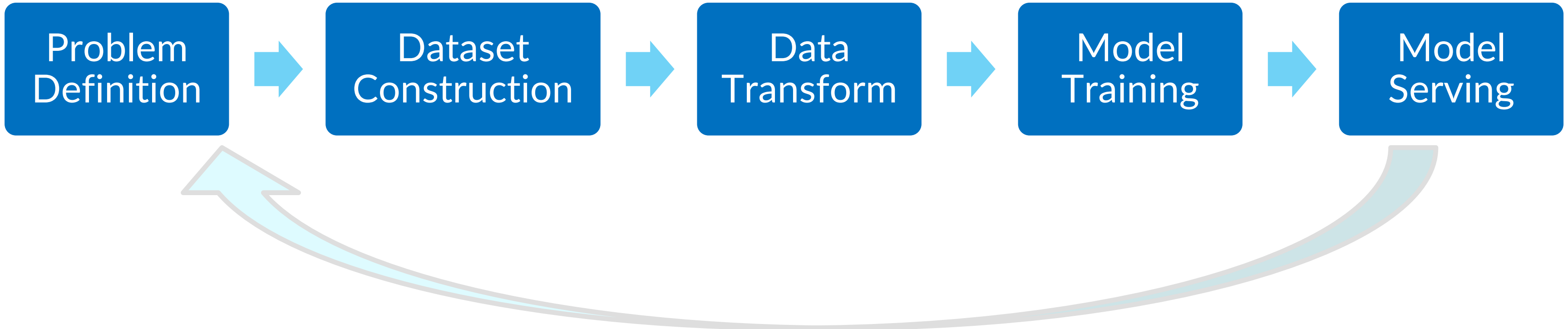
Supervised Learning Template

Two Phases



AI/ML in Practice

Empirical Science



1. Set the research goal
2. Make a hypothesis
3. Collect the data
4. Build a model and test your hypothesis
5. Analyze your results
6. Reach a conclusion
7. Refine hypothesis and repeat

Deciding on ML

01 Start simply and clearly

02 Define an ideal outcome

03 Set success / failure metrics

04 Design appropriate model outputs

05 **Start with heuristics / rules**

Occam’s Razor

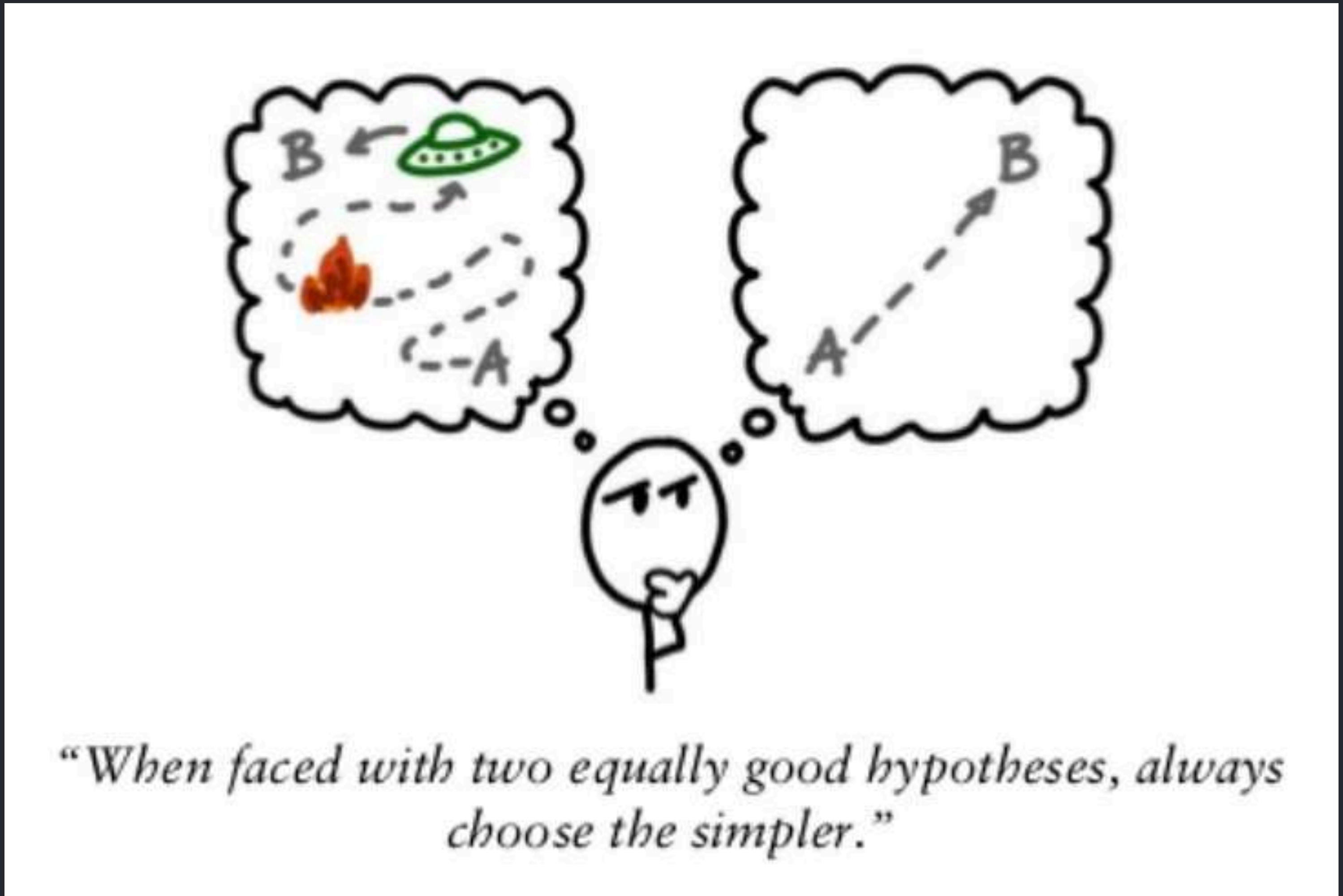
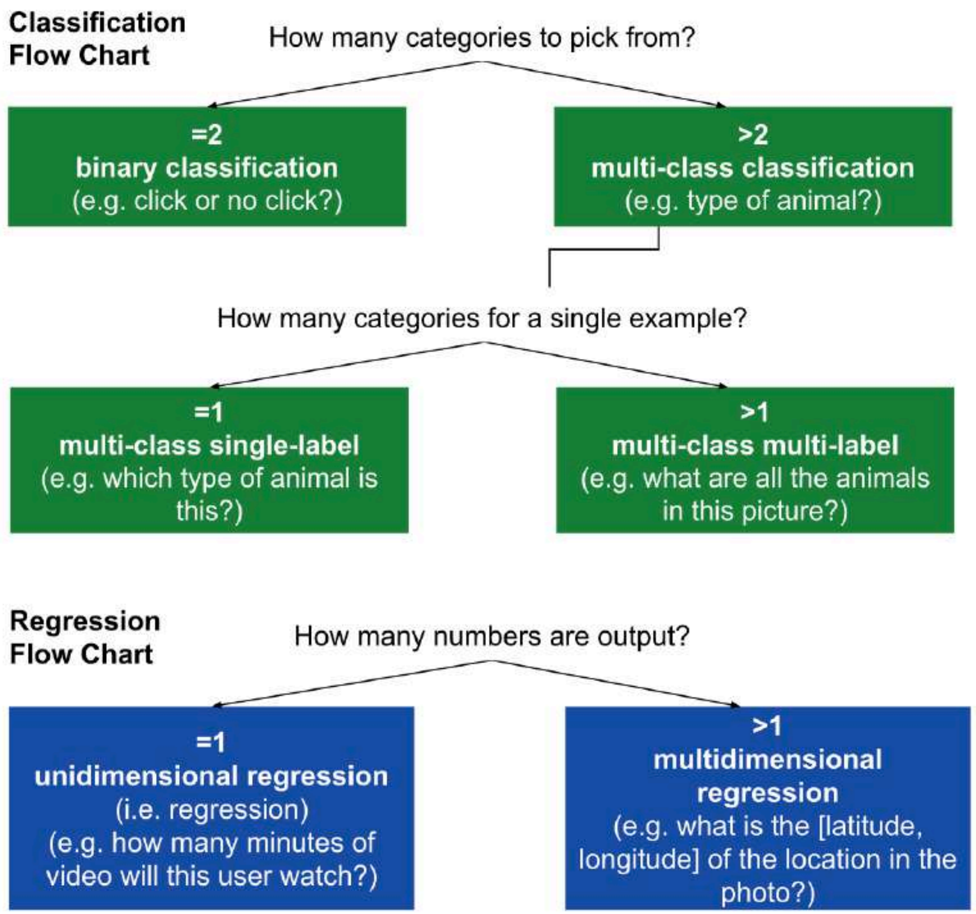


Image source: ClubStreetPost.com

Problem Definition

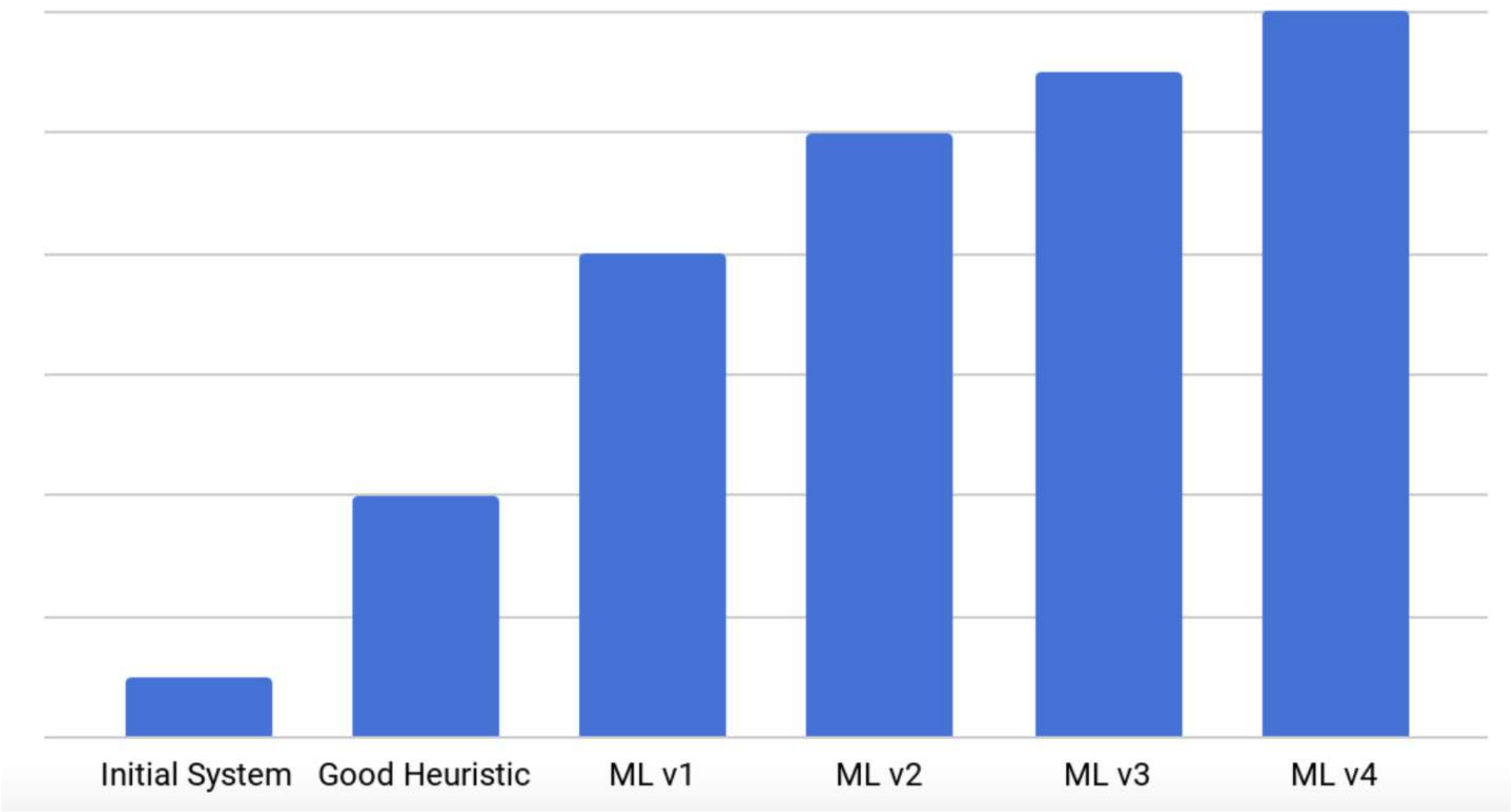
ML problem is best framed as:

- Binary classification
- Multi-class single-label classification
- Multi-class multi-label classification
- Uni-dimensional regression
- Multi-dimensional regression
- Clustering (unsupervised)
- Other (translation, parsing, bounding box, etc)



Source: https://developers.google.com/machine-learning/problem-framing/formulate

Biggest gain in ML is first launch



Source: https://developers.google.com/machine-learning/problem-framing/formulate

AI

Dataset Construction

01 Collect the raw data

02 Identify feature and label sources

03 Select a sampling strategy

04 Split the data

Bukalapak

AI

Data Transform

01 Normalization

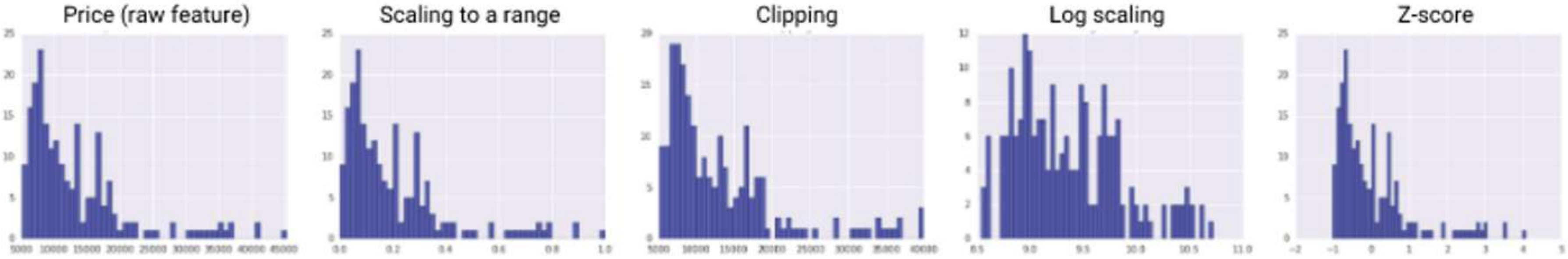
02 Bucketing

03 Categories to numbers

Bukalapak

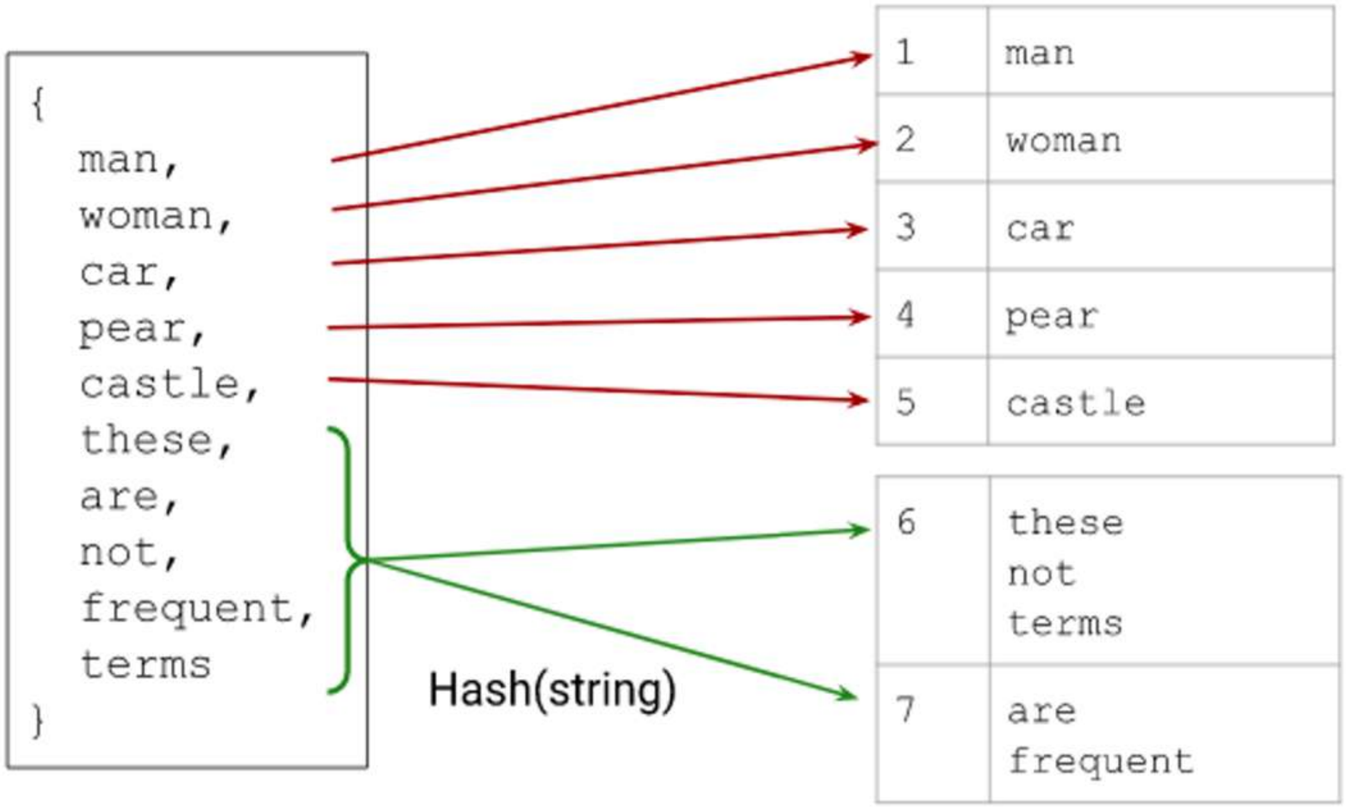
Normalization

Normalization Technique	Formula	When to Use
Linear Scaling	$x' = (x - x_{min})/(x_{max} - x_{min})$	When the feature is more-or-less uniformly distributed across a fixed range.
Clipping	if x > max, then x' = max. if x < min, then x' = min	When the feature contains some extreme outliers.
Log Scaling	x' = log(x)	When the feature conforms to the power law.
Z-score	x' = (x - μ) / σ	When the feature distribution does not contain extreme outliers.

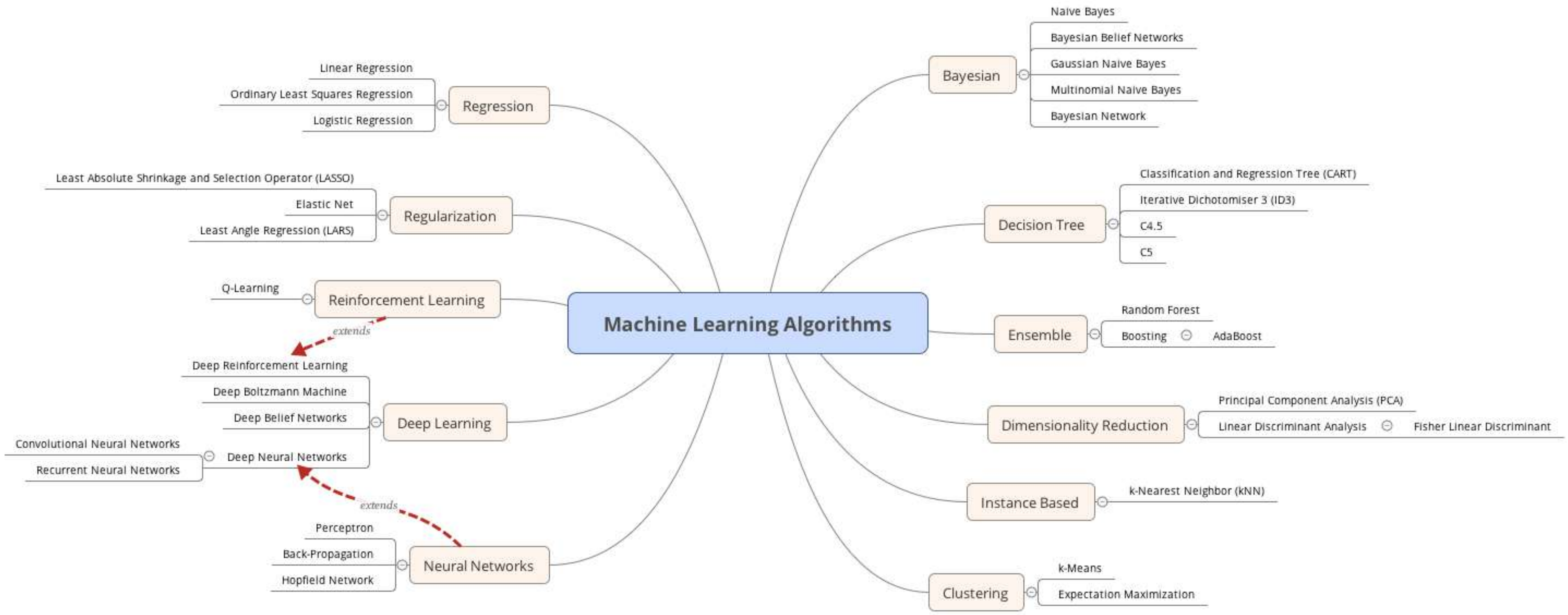


Categories to numbers

- Vocabulary
- Hashing

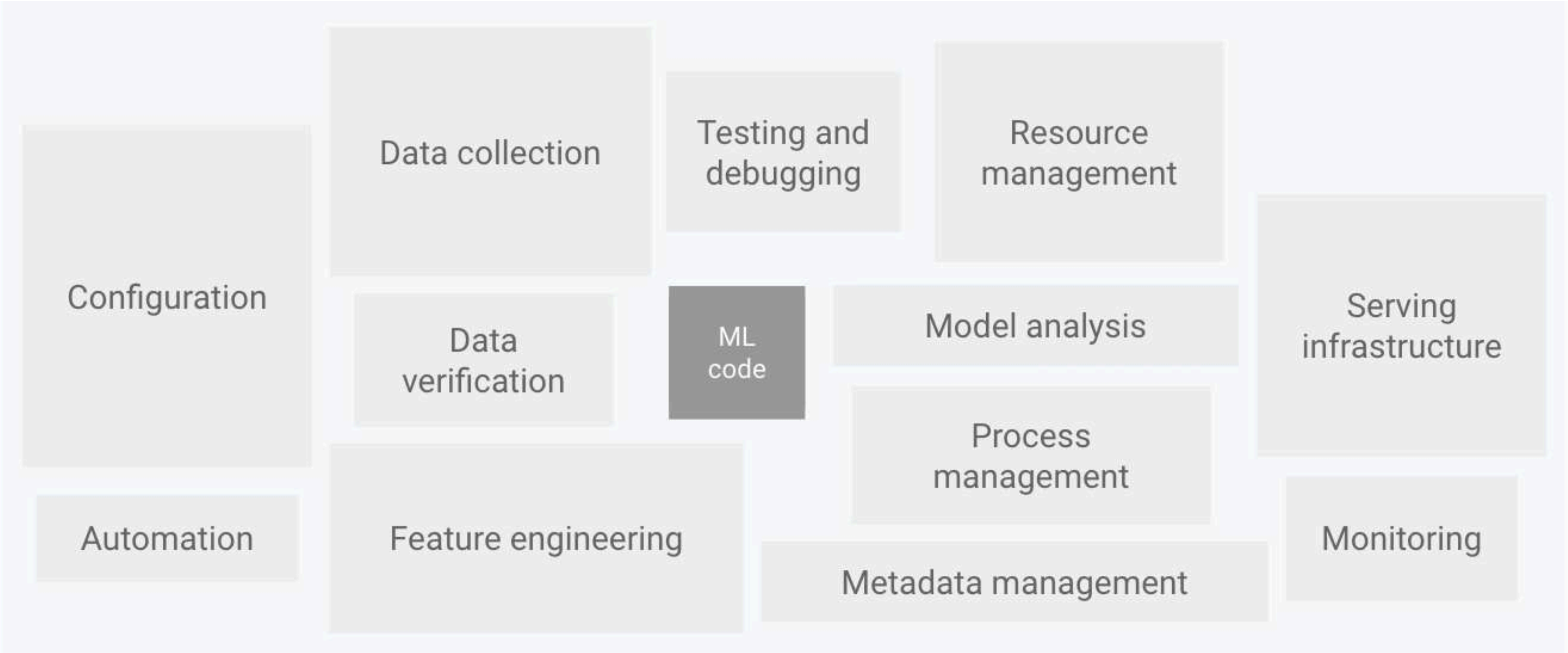


Model Training



Source: http://web.cs.ucla.edu/~shi.feng/Machine_Learning.html

Model Inference: ML in Production



[Sculley et al. NIPS 2015] ”Hidden Technical Debt in Machine Learning Systems

DevOps vs MLOps

DevOps

- Continuous Integration (CI)
- Continuous Delivery (CD)

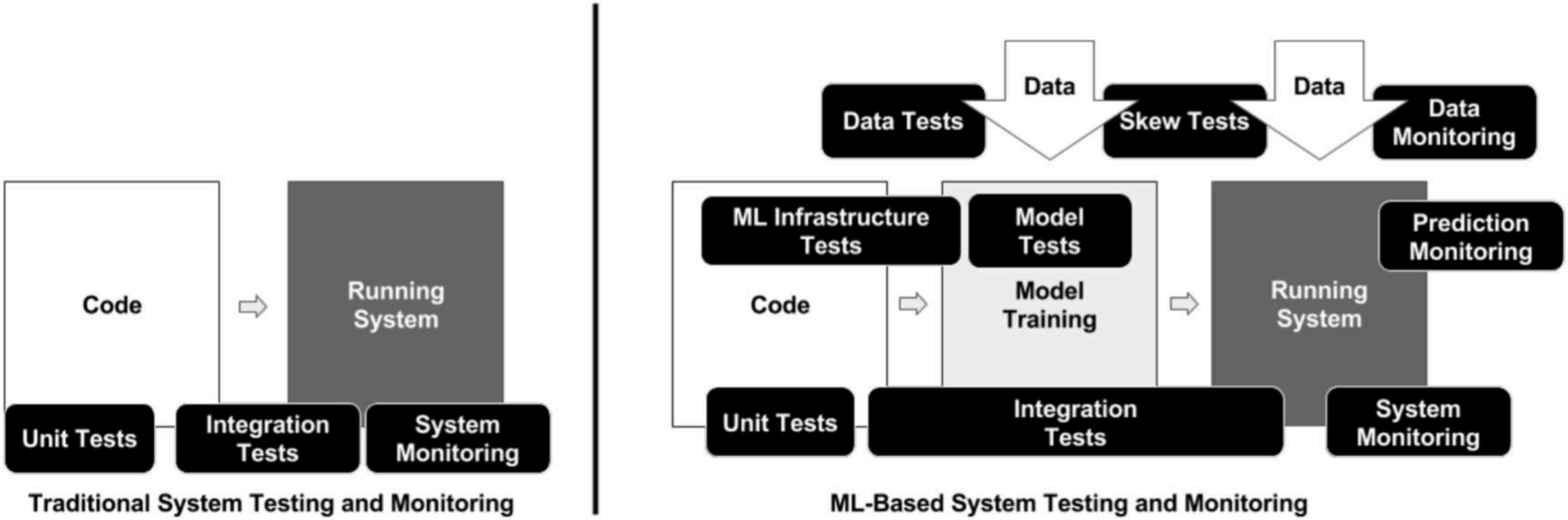
Asset:
Code + Components

MLOps

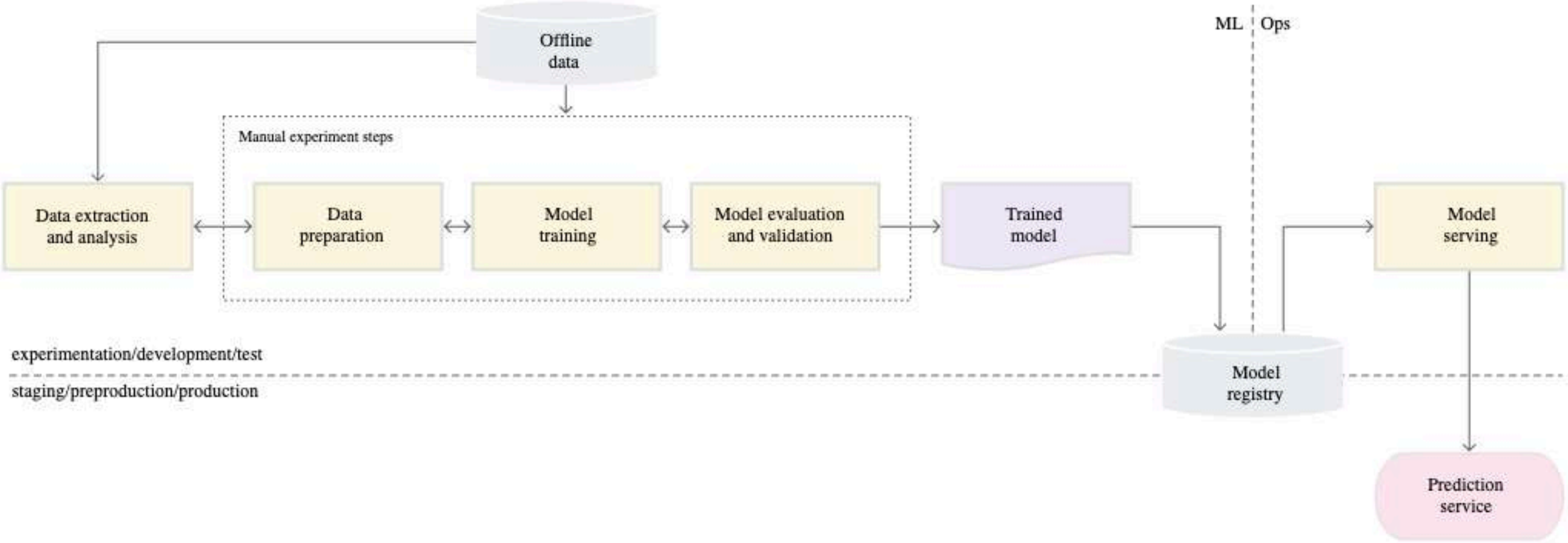
- Continuous Integration (CI)
- Continuous Delivery (CD)
- Continuous Training (CT)

Asset:
DevOps + Data + Model

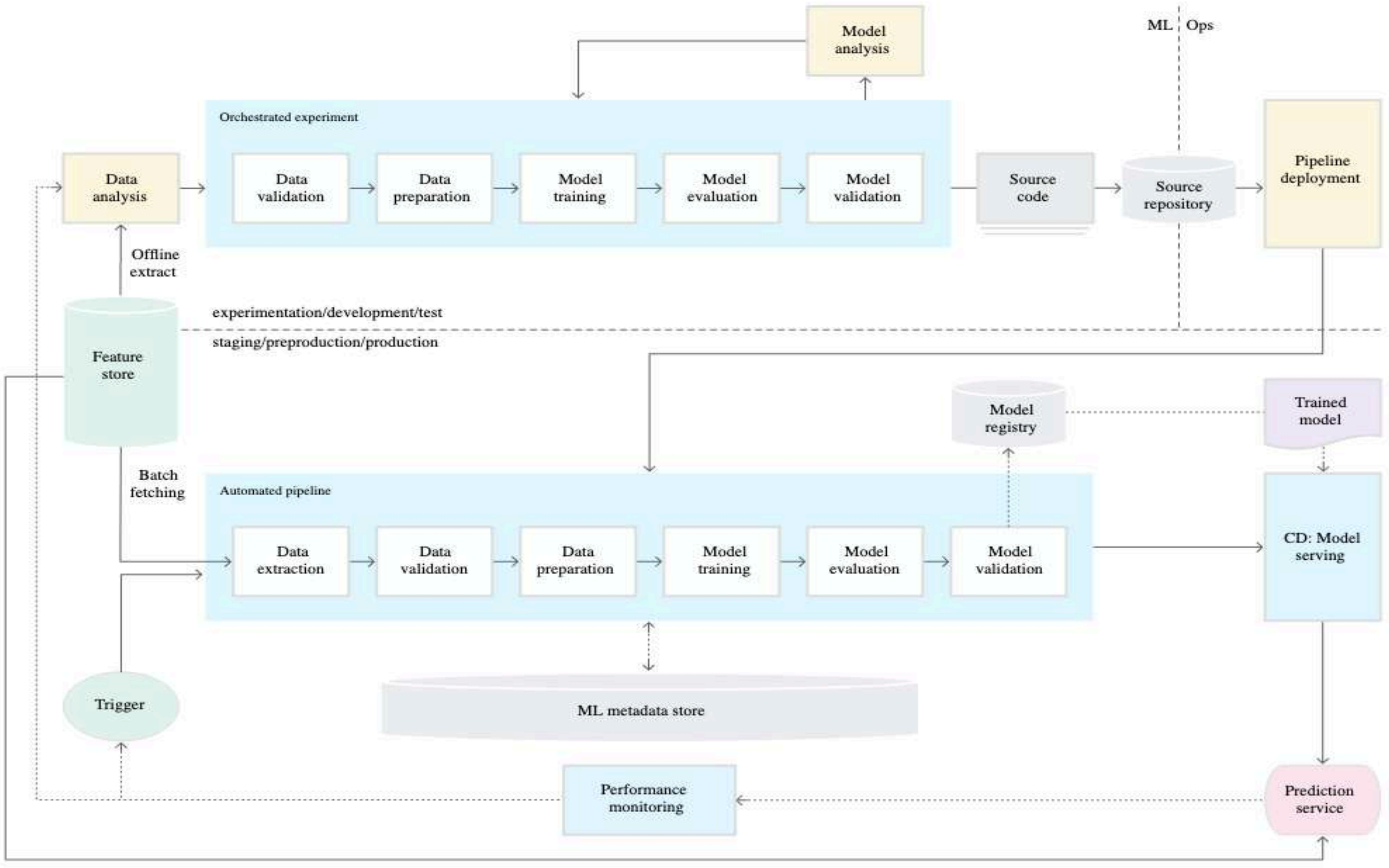
DevOps vs MLOps (2)



MLOps: Manual process



MLOps: Pipeline automation



AI in Bukalapak

One of the largest e-commerces in Southeast Asia

1,9M

Numbers of kiosk and agents combined

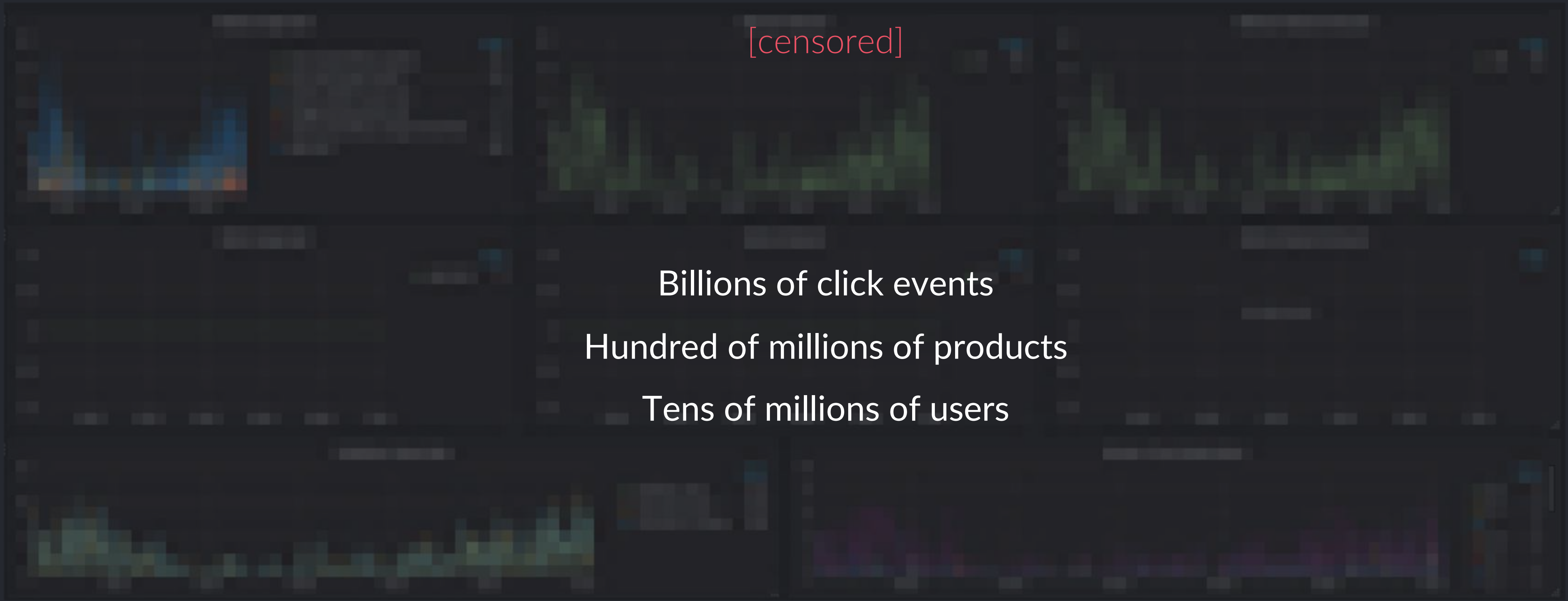
+4M

Engaged more than 4 million sellers

+50M

More than 50 million active users

PBs of data!



[censored]

Billions of click events

Hundred of millions of products

Tens of millions of users

Recommender System

Definition

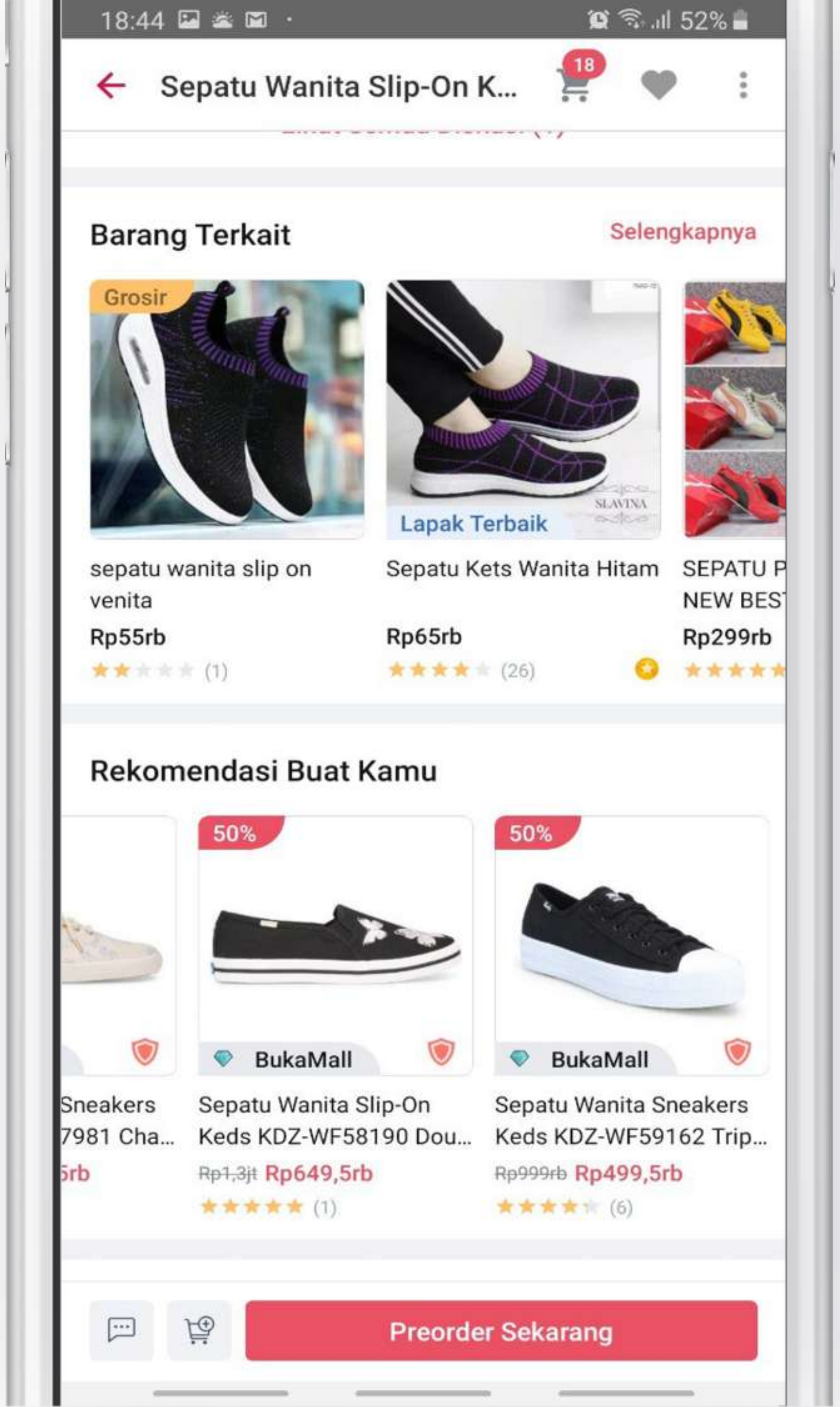
Recommender System

A computerized systems that suggest goods and service by predicting user’s preference and ratings.

Recommender systems in e-commerce identify a similarity in the preferences or tastes of one consumer and others (e.g. goods purchased, products viewed); and make recommendations for new purchases drawn from the set of other goods bought or viewed by each of the like-minded consumers.

Scope

Recommendation may be combined with Personalization



Service

Recommendation

Used in

Recommendation on PDP, Cart, and Homepage

Providing relevant recommendation based on personalization, popular products, and sellers to users. Recommendation is in Product Detail Page, Cart, and Homepage

Total Impact (since 2017)

>Rp200B per month

additional income for sellers

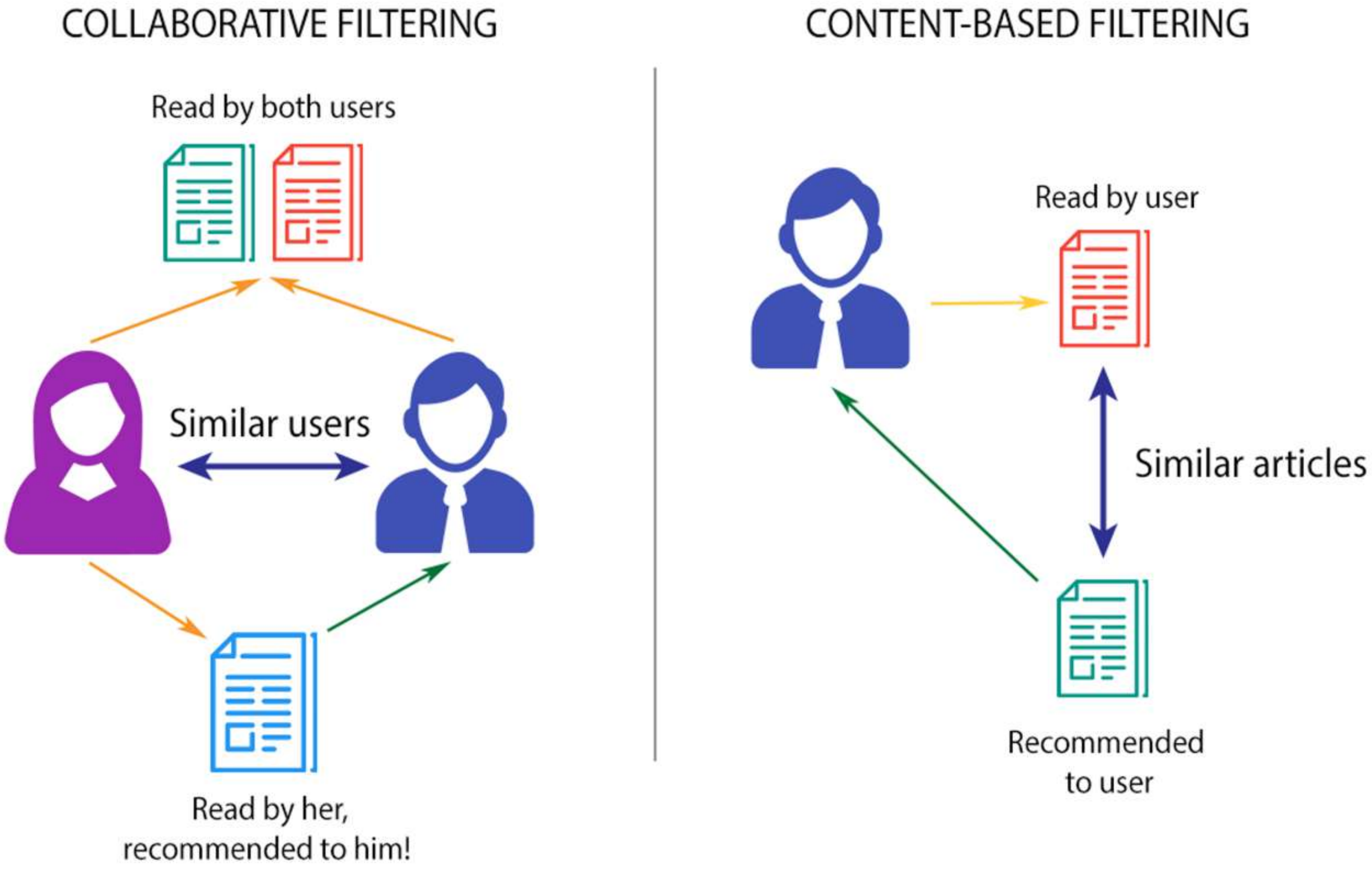
>45% of traffics to product detail

end up clicking our recommendation

>15% of traffics to product detail

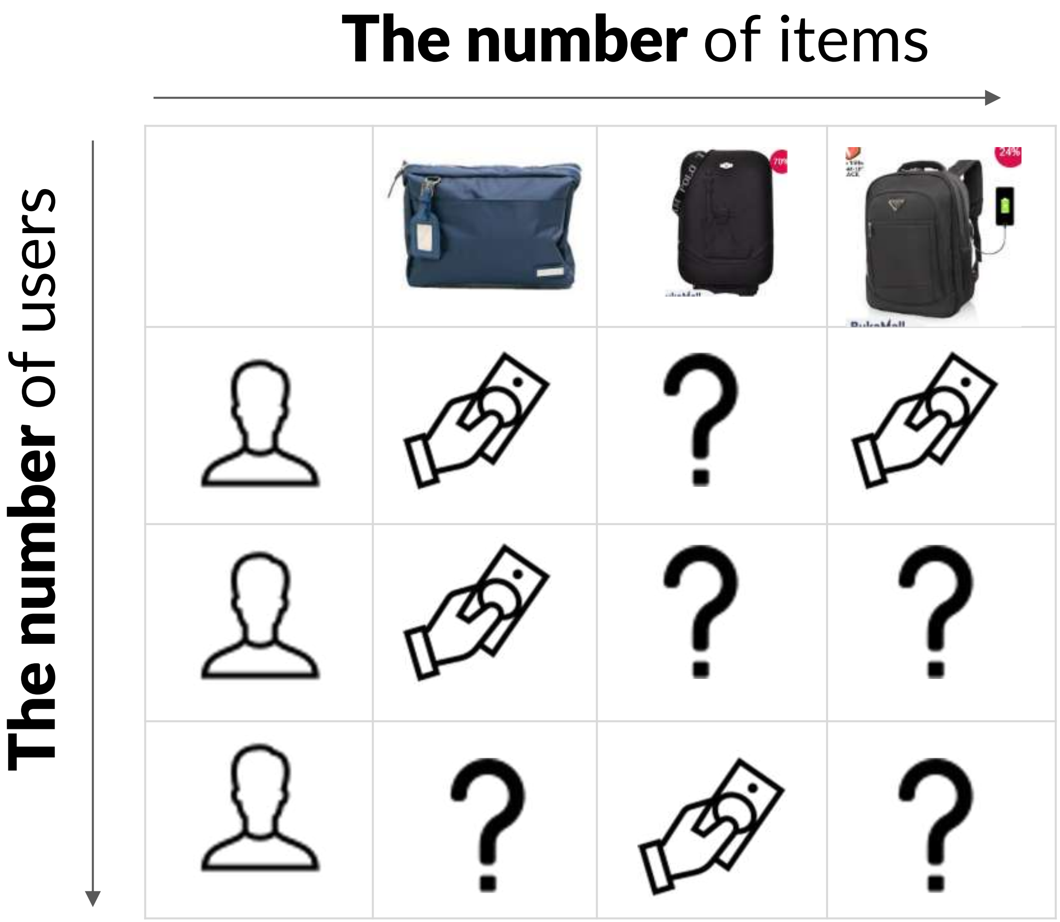
is accounted from our recommendation

RecSys Model



Source: https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-1-knn-item-based-collaborative-filtering-637969614ea

Collaborative Filtering



MATRIX FACTORIZATION

Classical approach in collaborative filtering

The task is to complete the matrix and predict the user preference

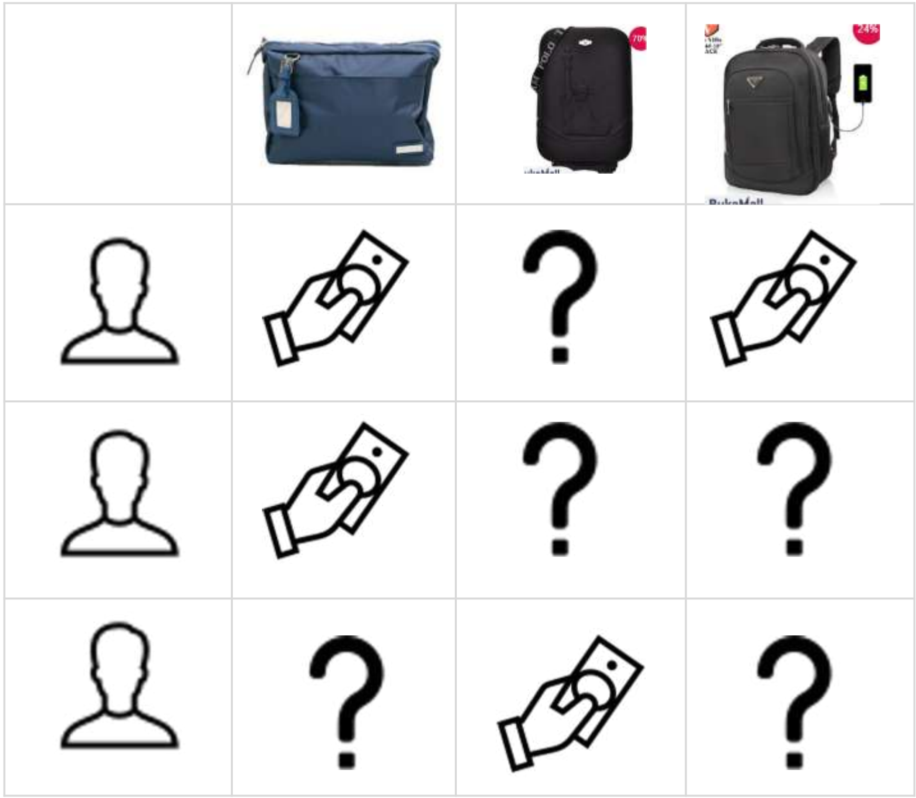
Represent the large user item matrix into multiplication of two low-rank matrices (e.g. **Singular Value Decomposition**) :

- User Factors (**k-size vector** for each user)
- Item Factors (**k-size vector** for each item)

Matrix Factorization
(challenges)

millions of items!

millions of users!



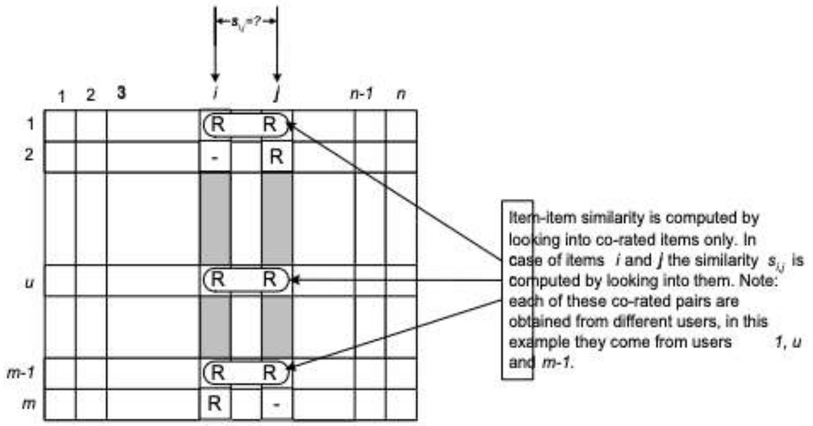
Data sparsity becomes an **issue**

Too many **zeros** in the matrix

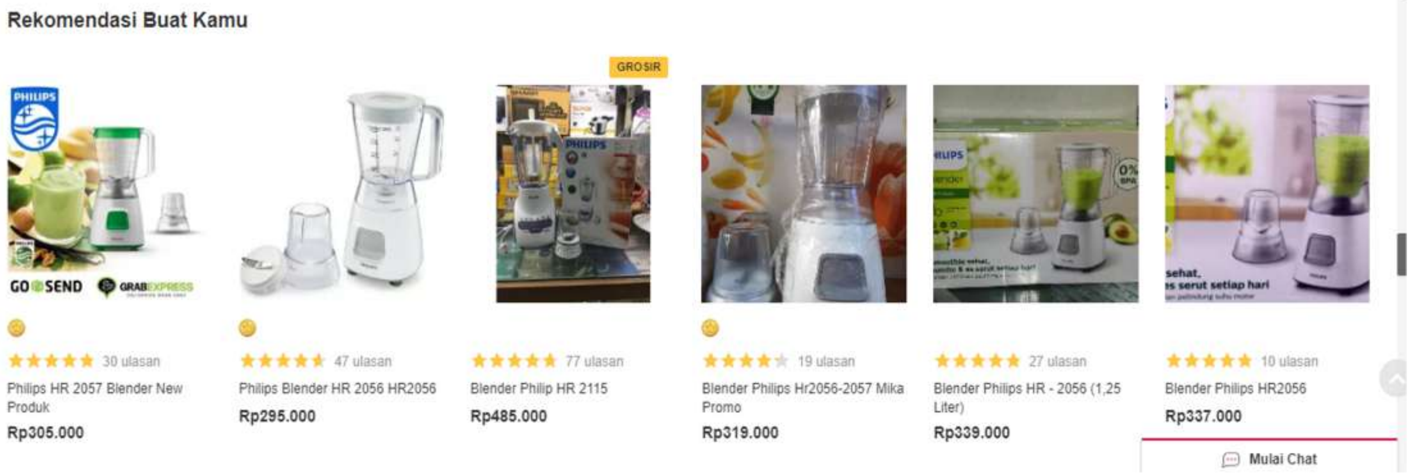
Similar products sold from different seller ⇒ Does it **translate** to different user preference?

Similar Item Recommendation

“Mitigating the scalability issue”



[Sarwar et al. 2001] Item-based Collaborative Filtering Recommendation Algorithms



For Buyer

Streamline the customer browsing journey

If we can provide similar items well, the user doesn’t need to **go back and forth** between pages

For Sellers

Increase **exposure** to various products

Help our sellers to **increase their income**

E-commerce Data

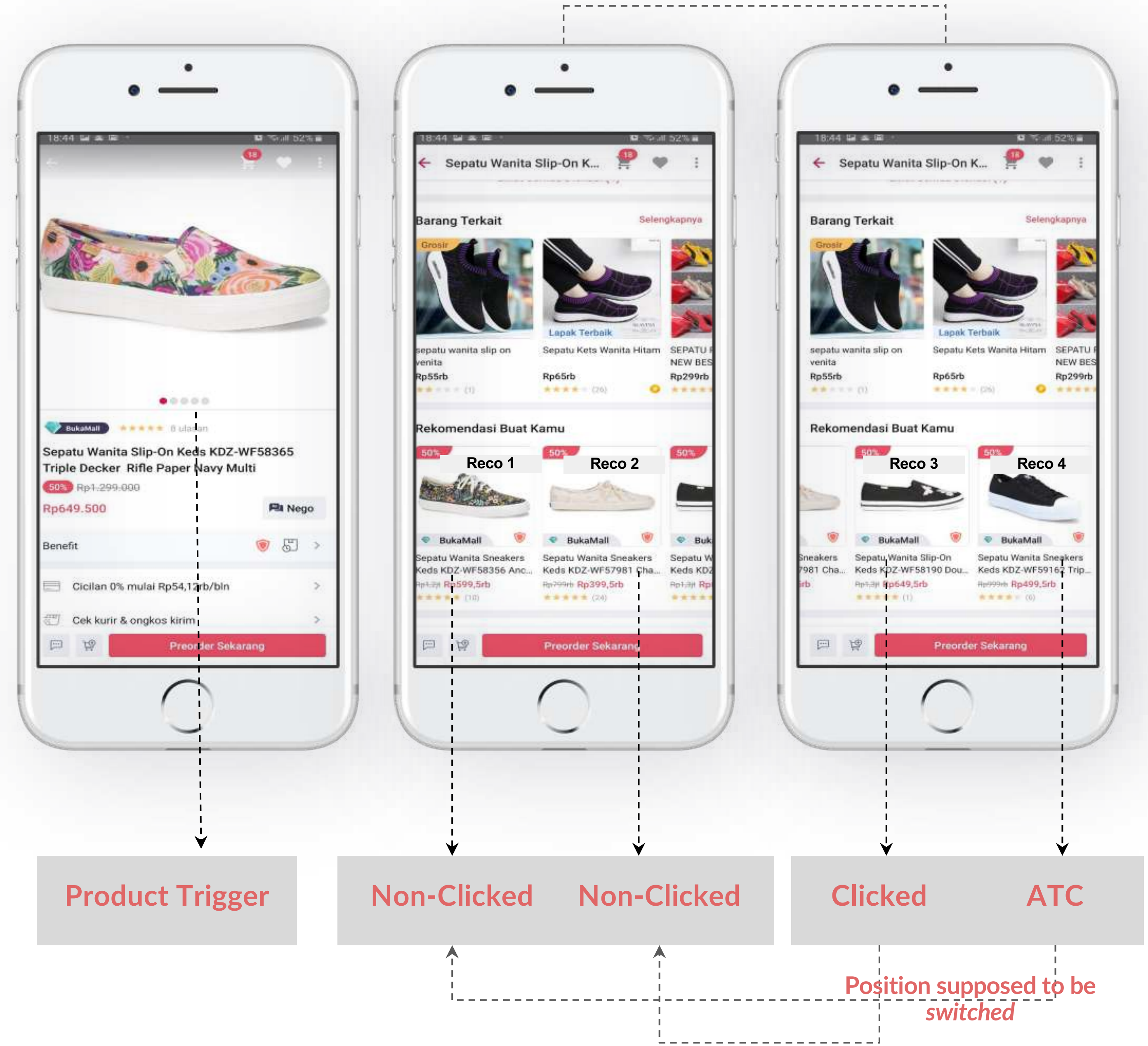
(User Feedback)

User click indicates some level of interest. Clicks are abundant. But **noisy!**

Okay, I'm interested in buying this stuff! The big question is does the user eventually buy it?

I definitely like it. I have paid for the product!

I **love** this item!



Brovman, et al. 2016. Optimizing Similar Item Recommendations in a Semi-structured Marketplace to Maximize Conversion. In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16).

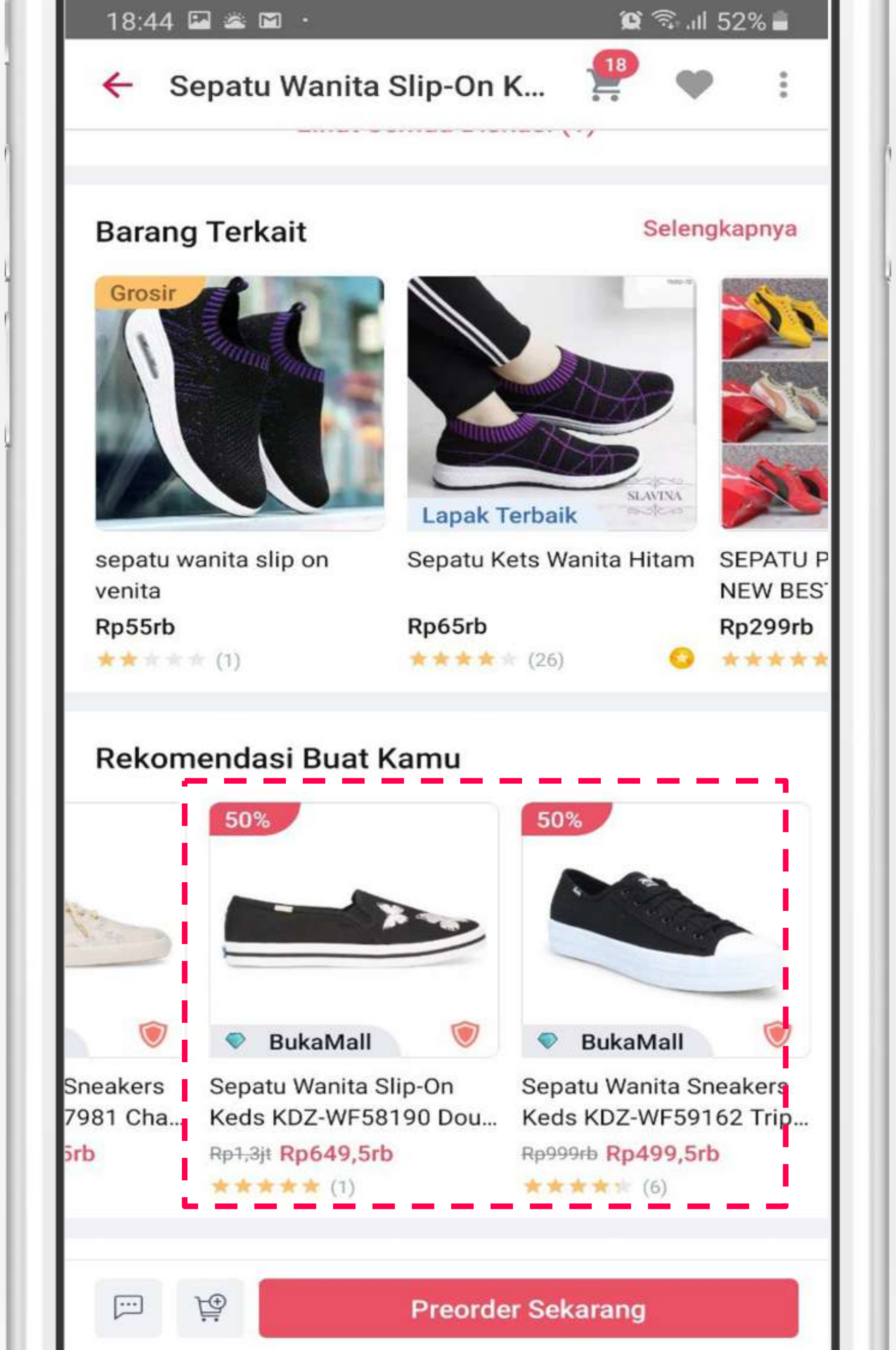
Enhance re-ranking through Learning-to-Rank

Comparison Features

1. Title Similarity
2. Price Ratio
3. Category
4. etc.

Item Quality Features

1. Product Rating
2. Seller Feedbacks
3. Revenue
4. etc.



[L. Evalina, et al. ICACSIS 2019] "Toward Improving Similar Item Recommendation for a C2C Marketplace"

Classification Result

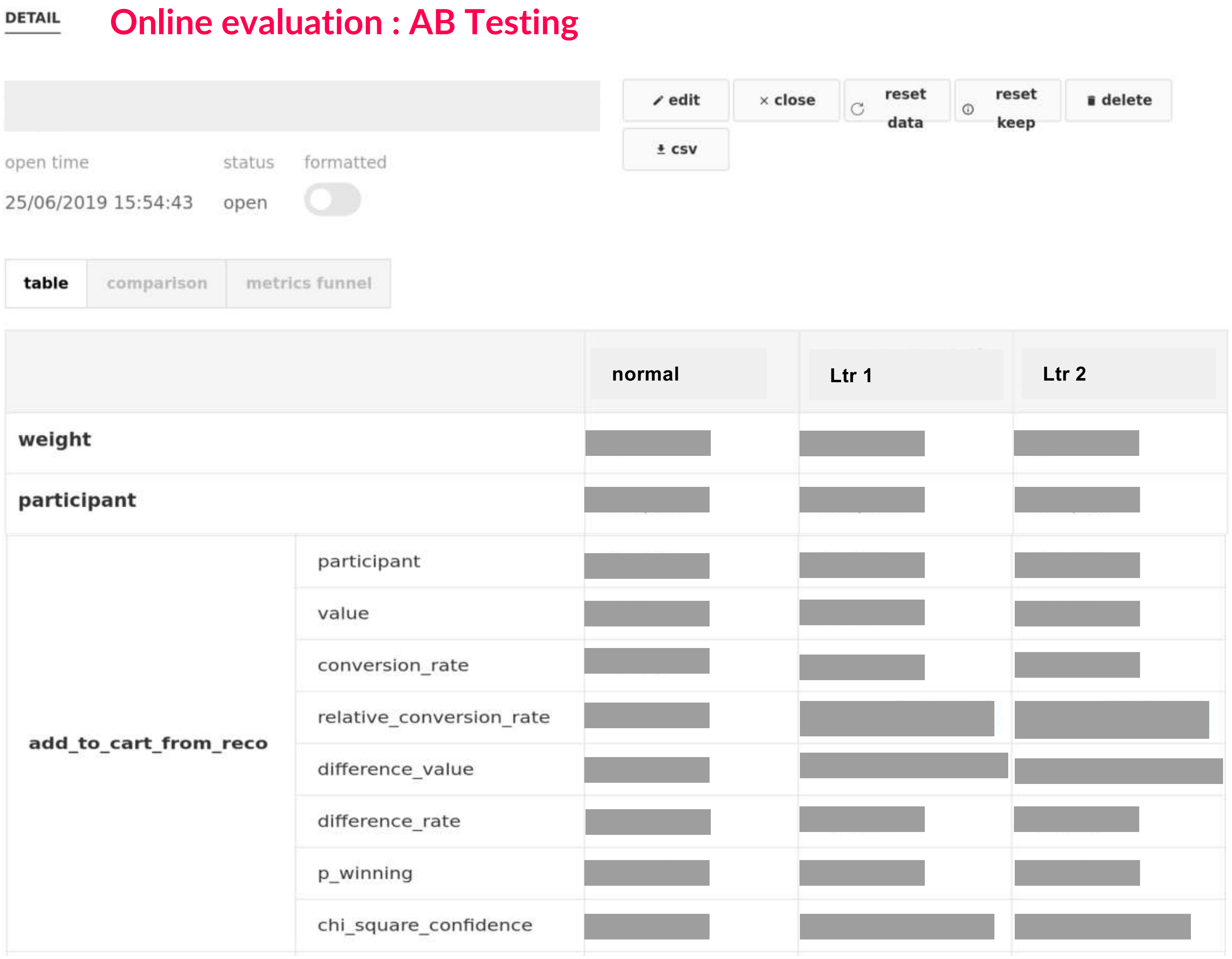
We use Logistic Regression, Random Forest, XGB. However, ***LogReg*** came up with the best ***performance***.

Metrics	Baseline	LTR
MAP	28.43%	31.35%

Notes :

1. A/B test shows ***positive*** result for ***paid*** and ***atc*** conversion.
2. Rank aware metrics are ***correlated*** with the ***online testing.***

Online evaluation : AB Testing



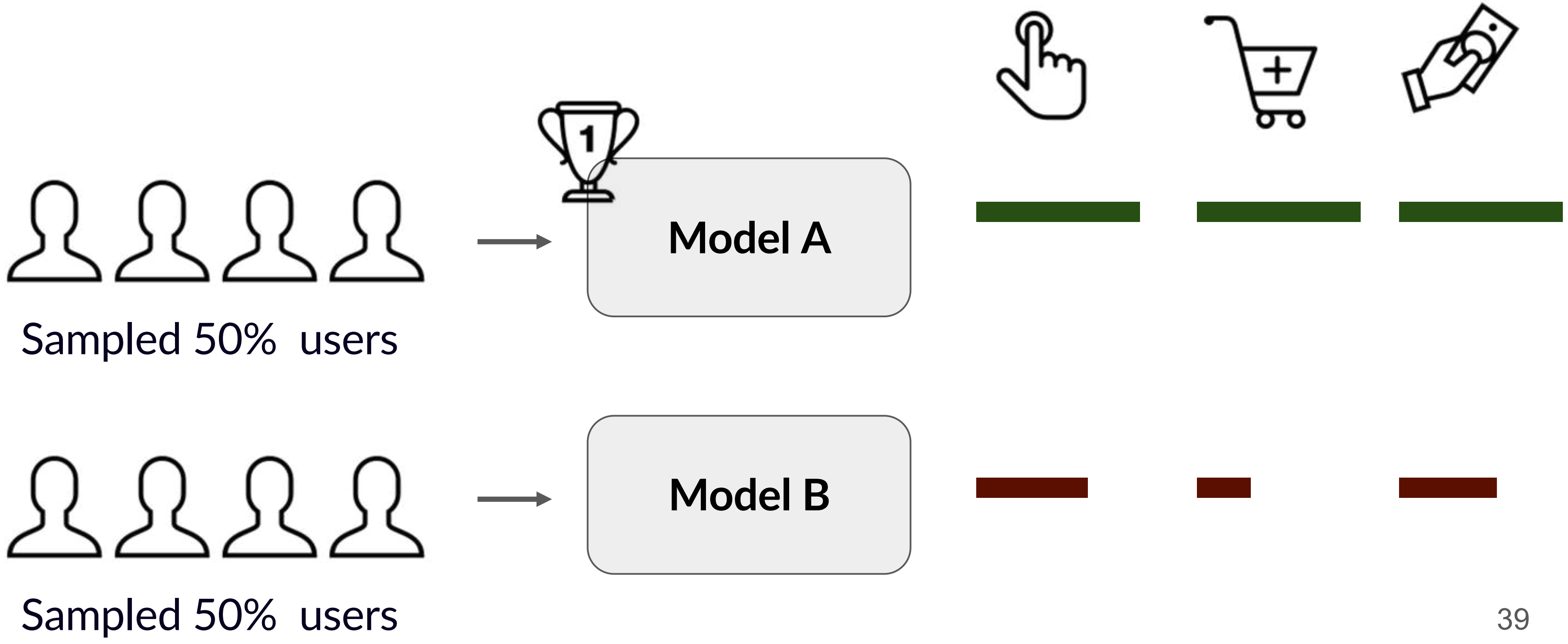
Evaluation

OFFLINE EVALUATION

Accuracy, Diversity, Qualitative Check

ONLINE EVALUATION

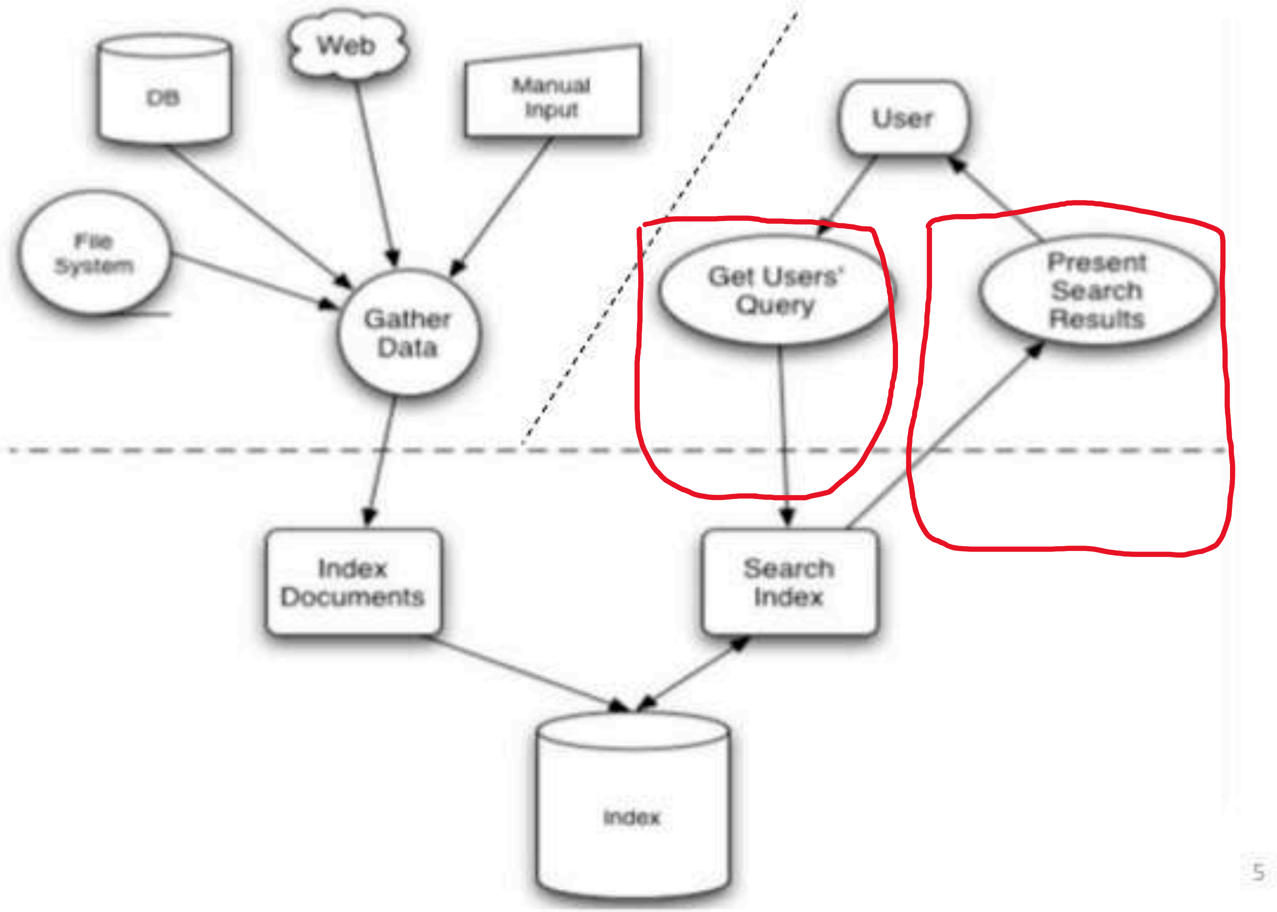
A/B Testing



Deployment

Search

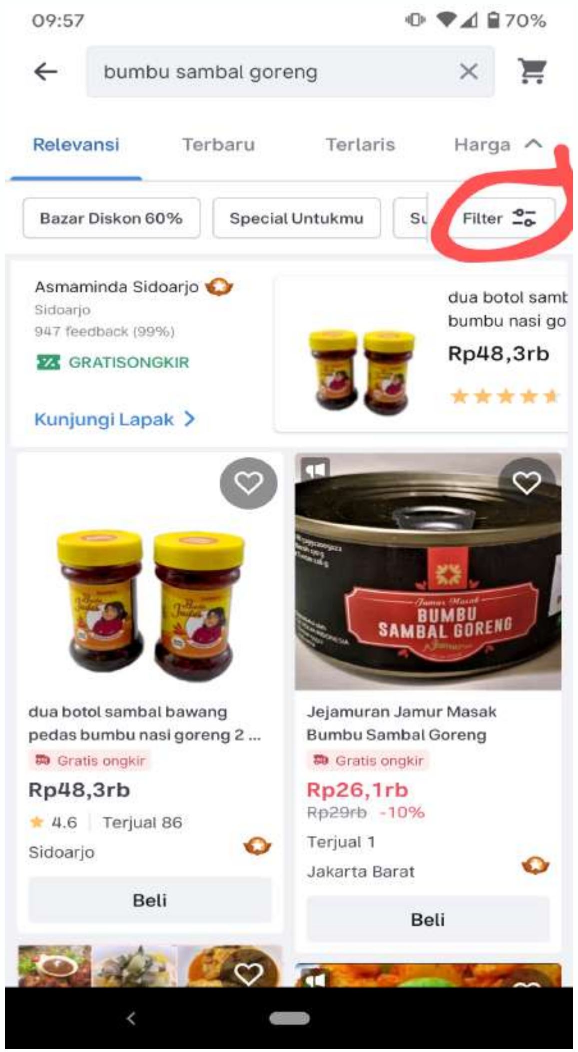
Search Engine Architecture

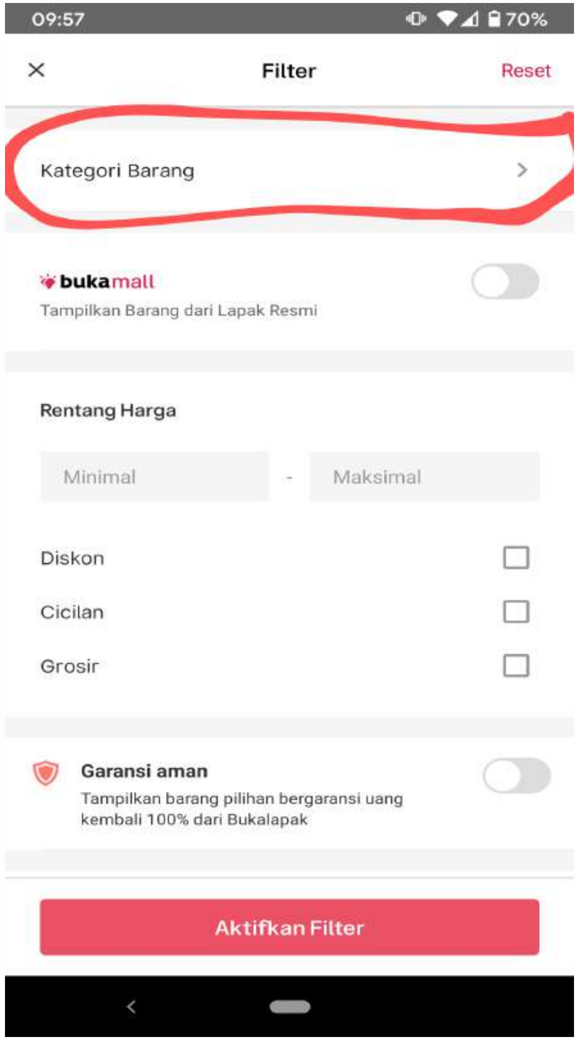


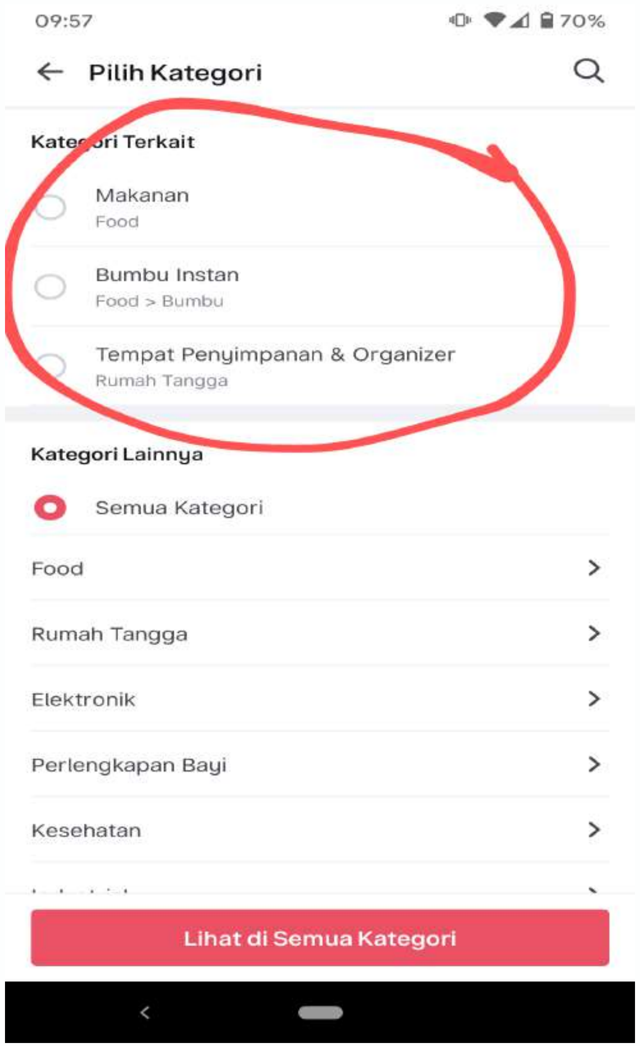
Source: https://www.slideshare.net/Zhiguang/intro-to-elasticsearch-63475620

Query to Category Mapping

Useful for search result filtering by category







Query Typo Corrections

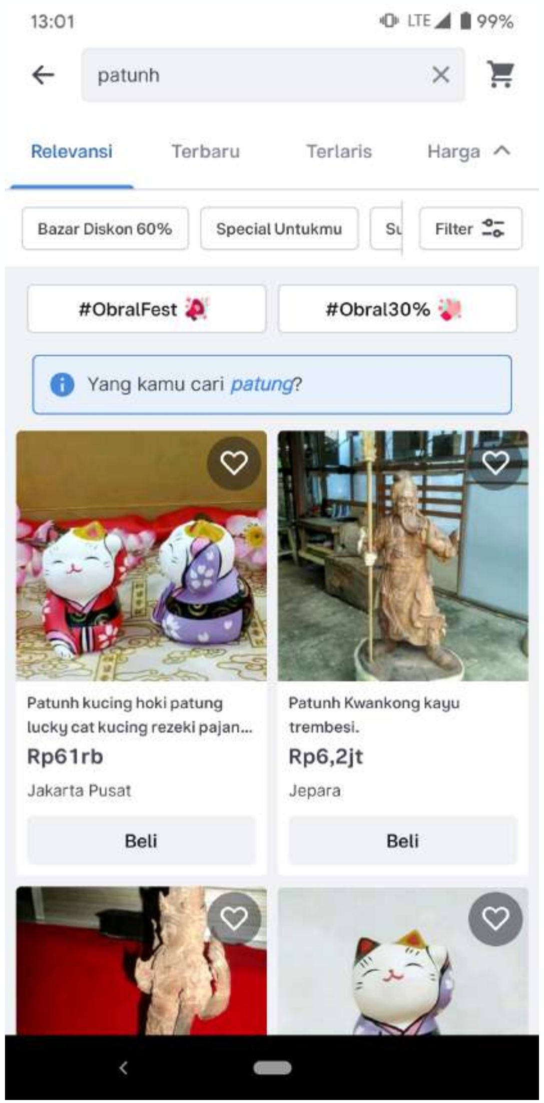
Bi-gram Language Model

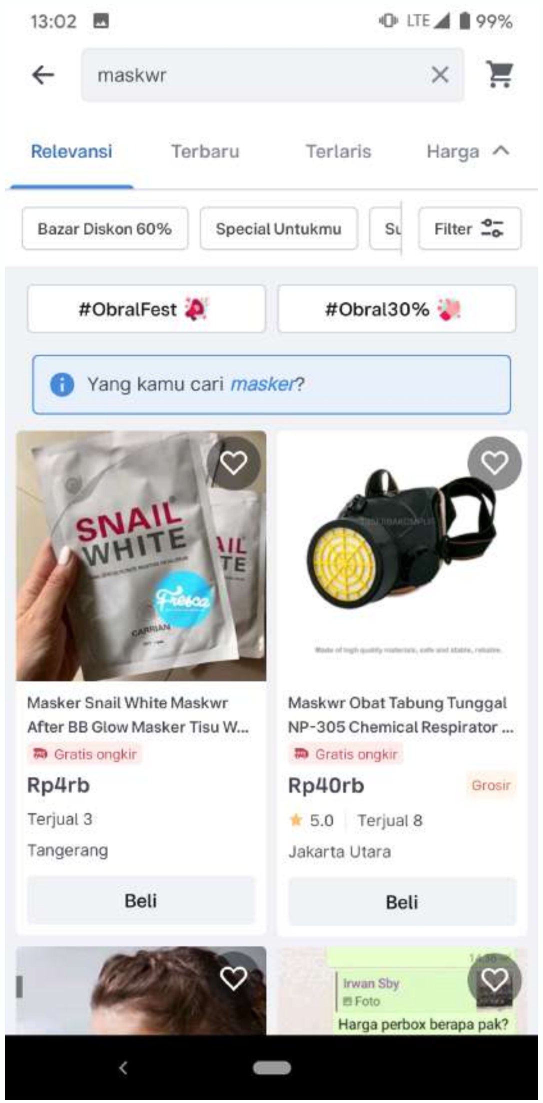
- Frequency count-based: easy to implement and productionize
- Fast inference

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Source: https://web.stanford.edu/~jurafsky/slp3/3.pdf





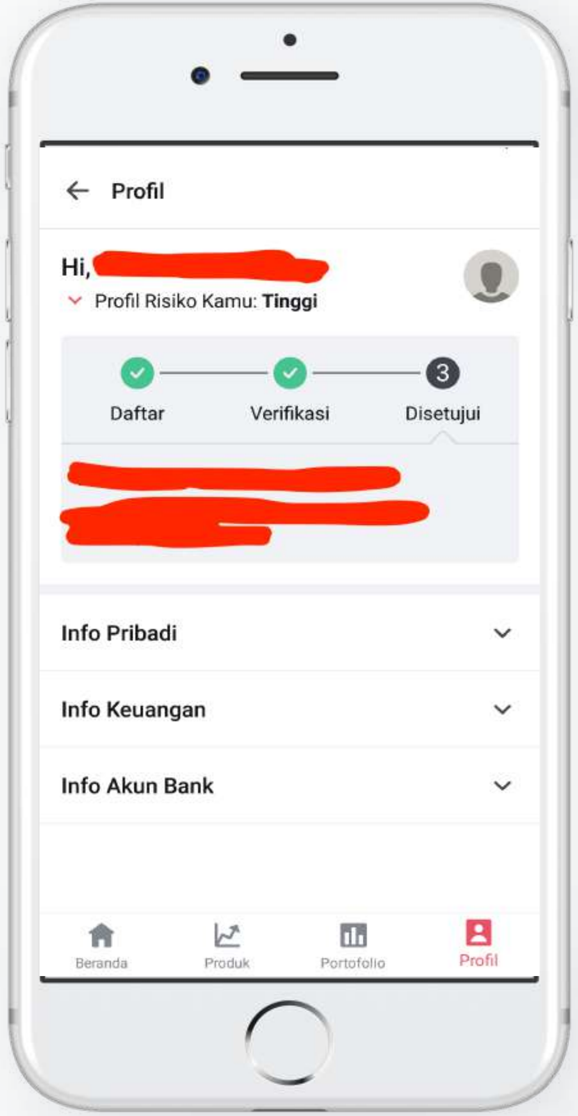
Investment Advisory

Risk Scoring

Predict the user’s risk profile of from the meta-data / attributes before filling the questionnaire

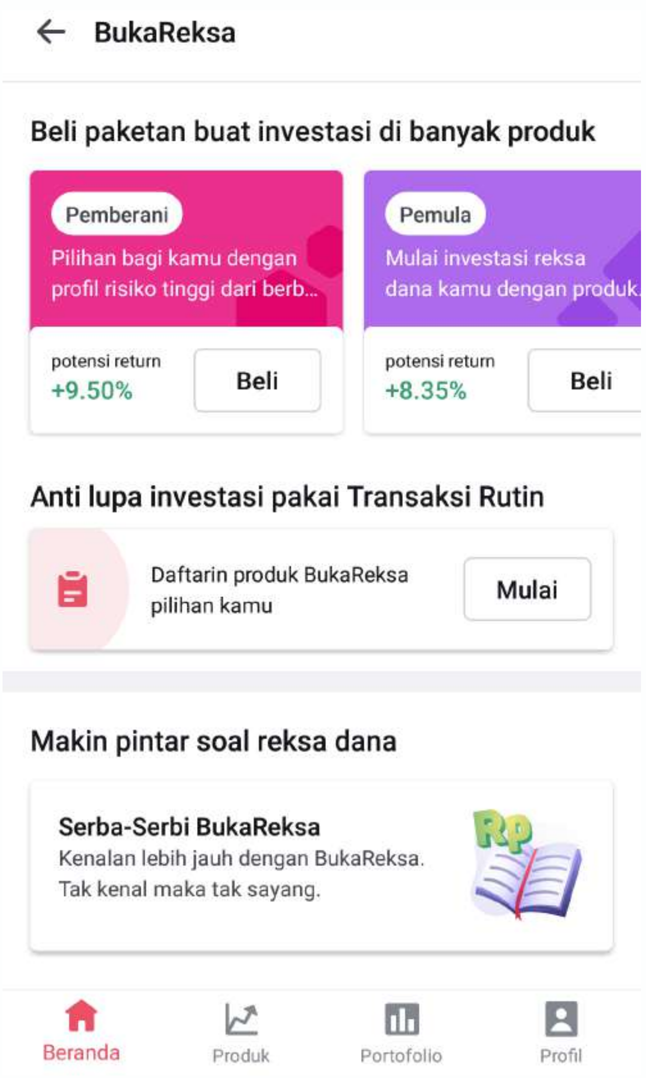
Source: https://analyticsindiamag.com/a-step-by-step-to-creating-credit-scoring-model-from-scratch/

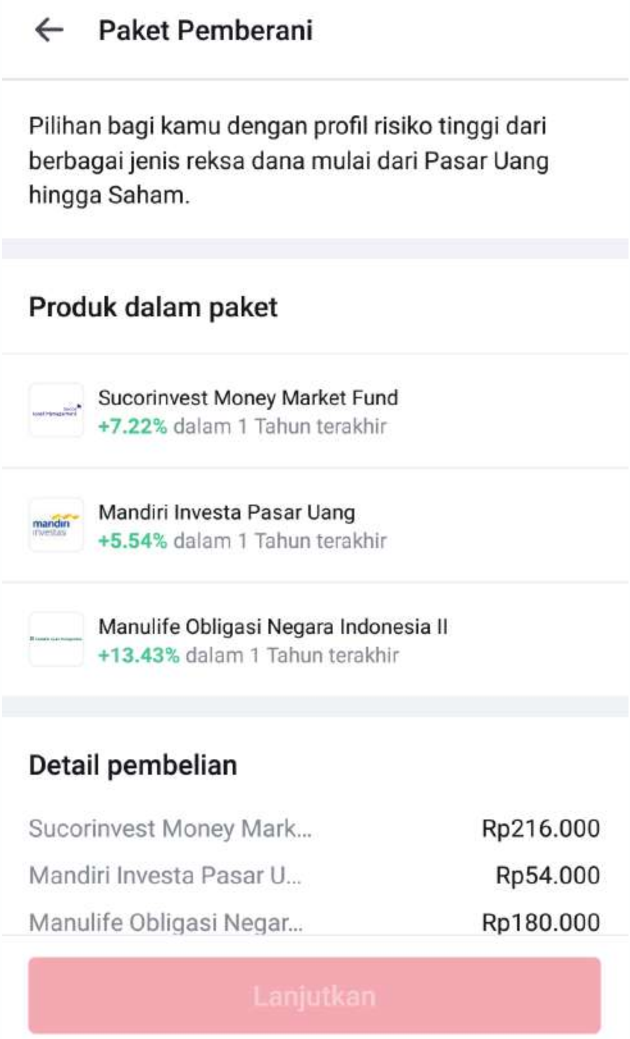
BukaReksa



ReksaDana Portfolio Selection

Provide ReksaDana packages that maximize return and minimize risk according to the user's risk profile.





Forbidden Product Filtering

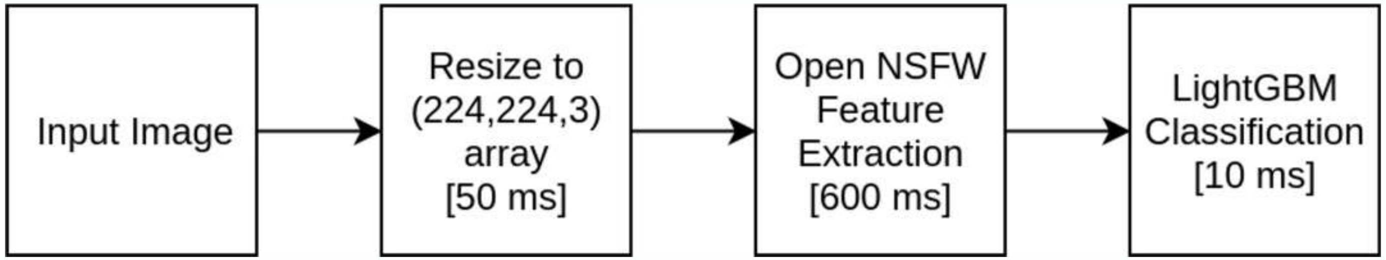
Forbidden Product Filtering

Usages:

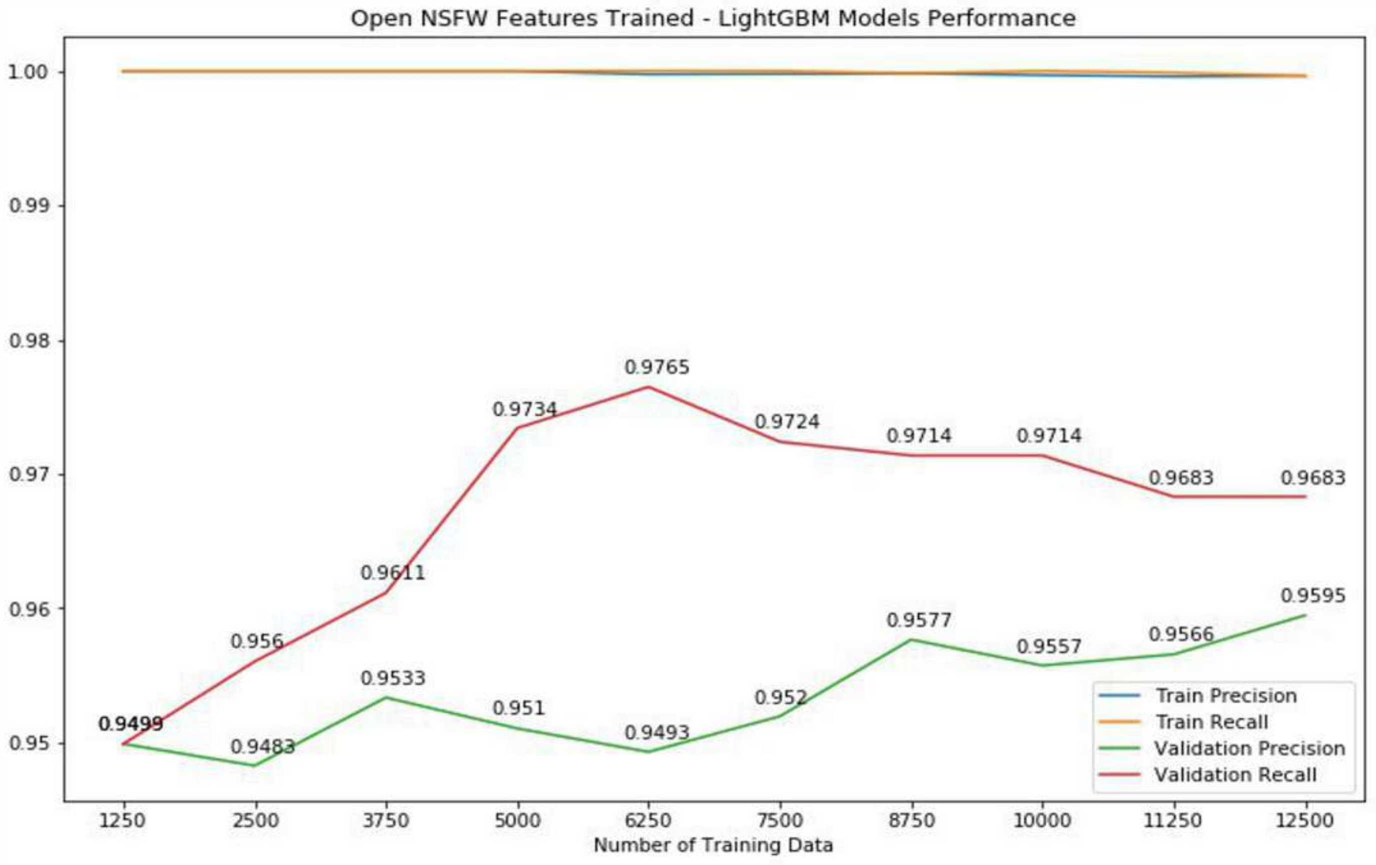
- Automatic filtering of products before being ingested by Ad Campaign
- Help Ops team to take down the products from marketplace

Handling Sexy Contents

- Deep learning-based feature extraction
- Transfer learning from Yahoo Open NSFW model -- source data are not available
- Human labelling for training is super important



Handling Sexy Contents (cont'd)



Seems good, right? But ...

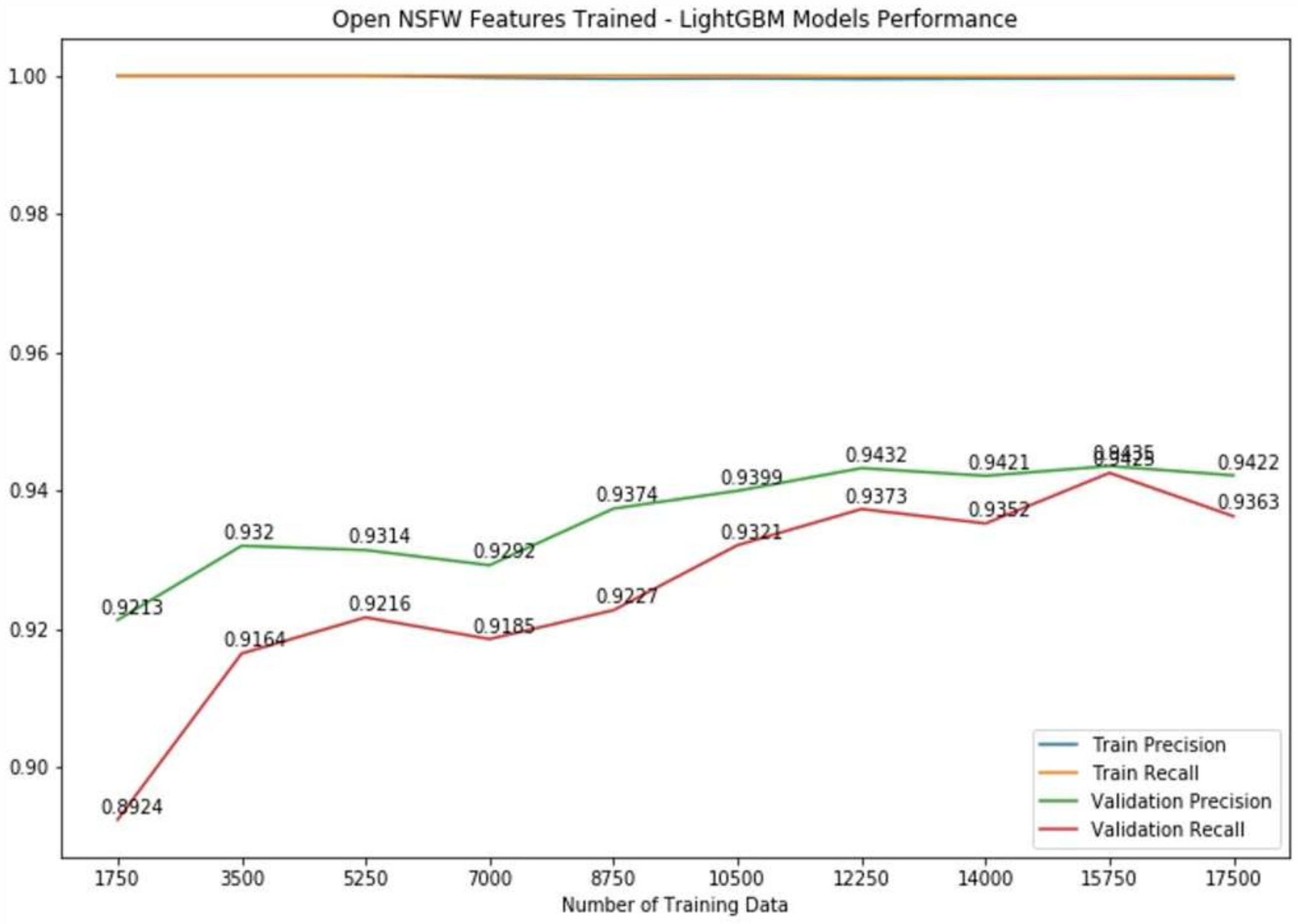
Handling Sexy Contents (cont’d)

← 97% NSFA

→ 96% NSFA

Not exactly what we wanted

Handling Sexy Contents (cont'd)



Seeing only the numbers, it seems to be worse. But, ...

Handling Sexy Contents (cont’d)

← 24% NSFA

→ 28% NSFA

Though it’s not perfect, this is closer to what we wanted ☺

AI Organizations & Skill Set

The materials are mostly taken from https://workera.ai/candidates/report

AI Organizations

Data Science

To make scientific decisions, help businesses run more effectively

Machine Learning

To automate tasks, decrease operational costs, scale a product

Data Engineering

Provide the necessary data to achieve the modeling or business analysis task.

Modeling

Prototyping models to exploit patterns found in data to predict outcomes, identify business risks and opportunities.

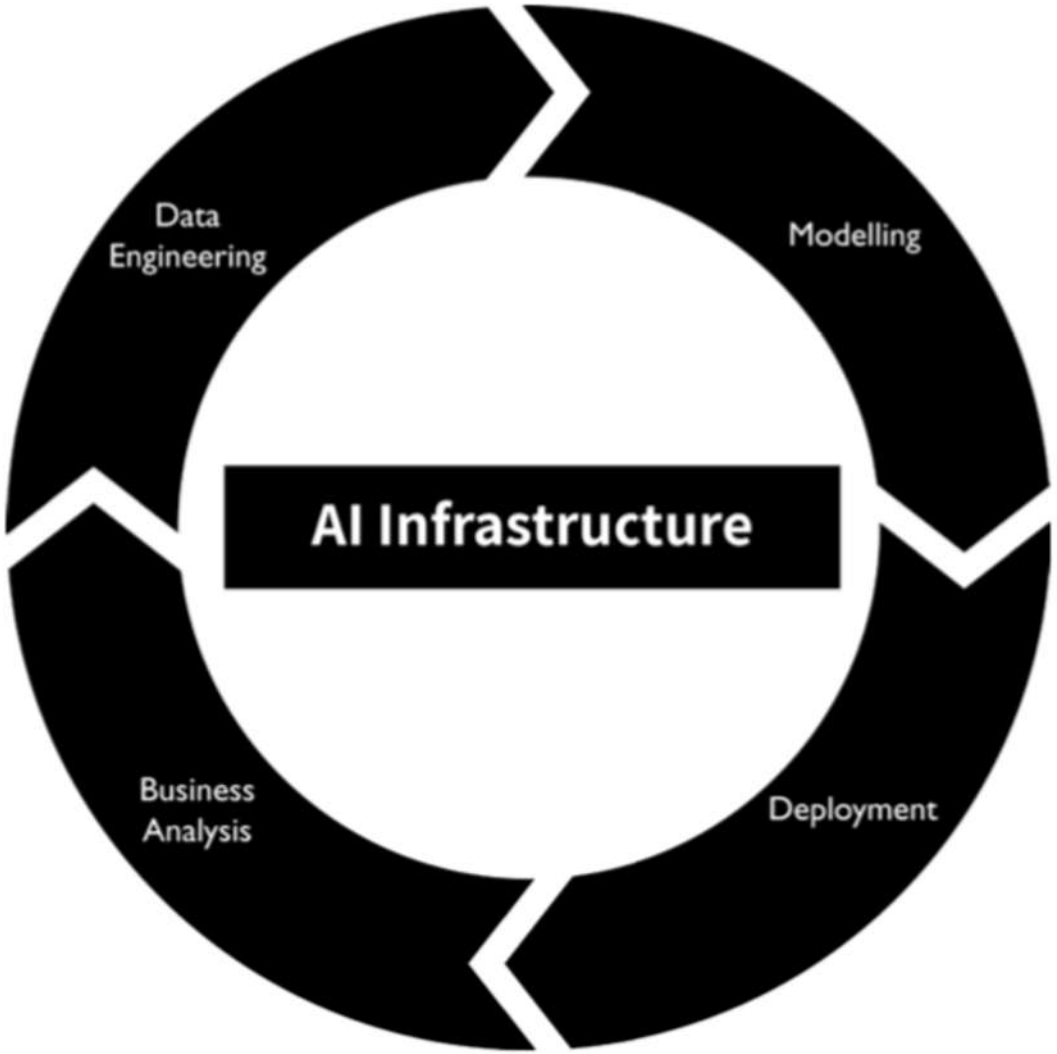
Deployment

All activities that make a model available for use, requiring the ability to write production code.

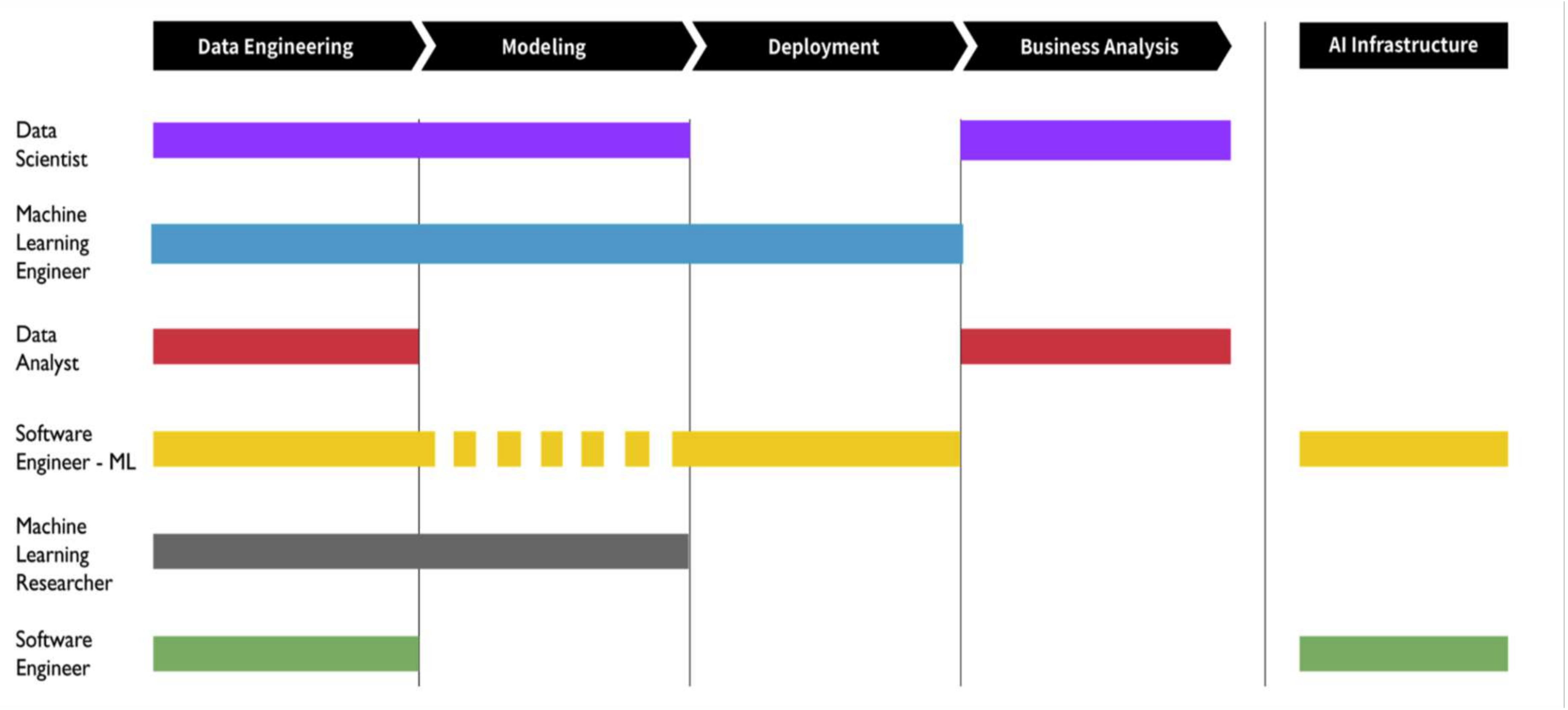
Business Analysis

Analytics, business activities related to communicating with clients and colleagues, thought leadership, and marketing.

AI Project Development Task Lifecycle

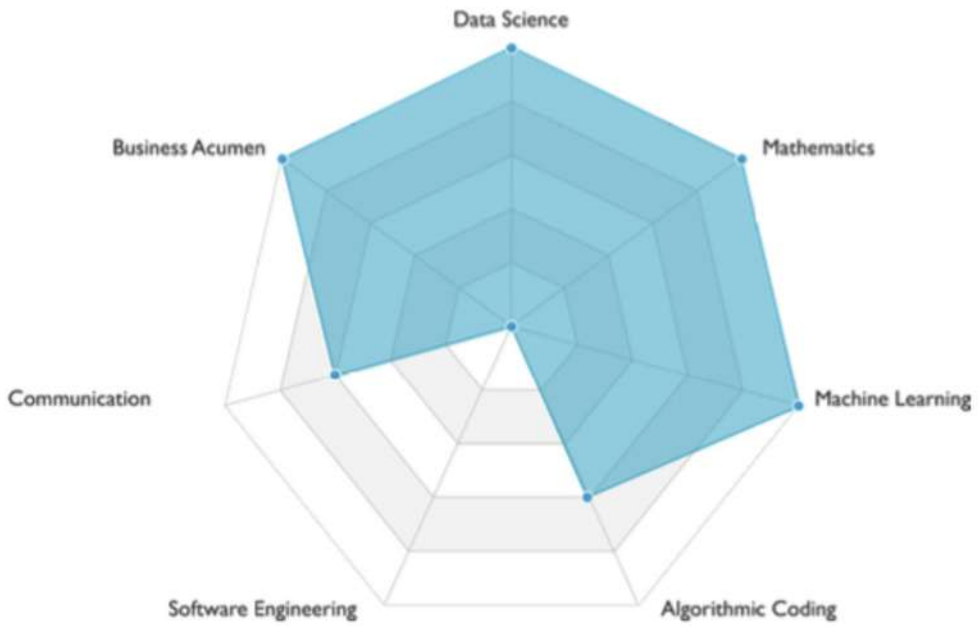


6 Roles of an AI Team

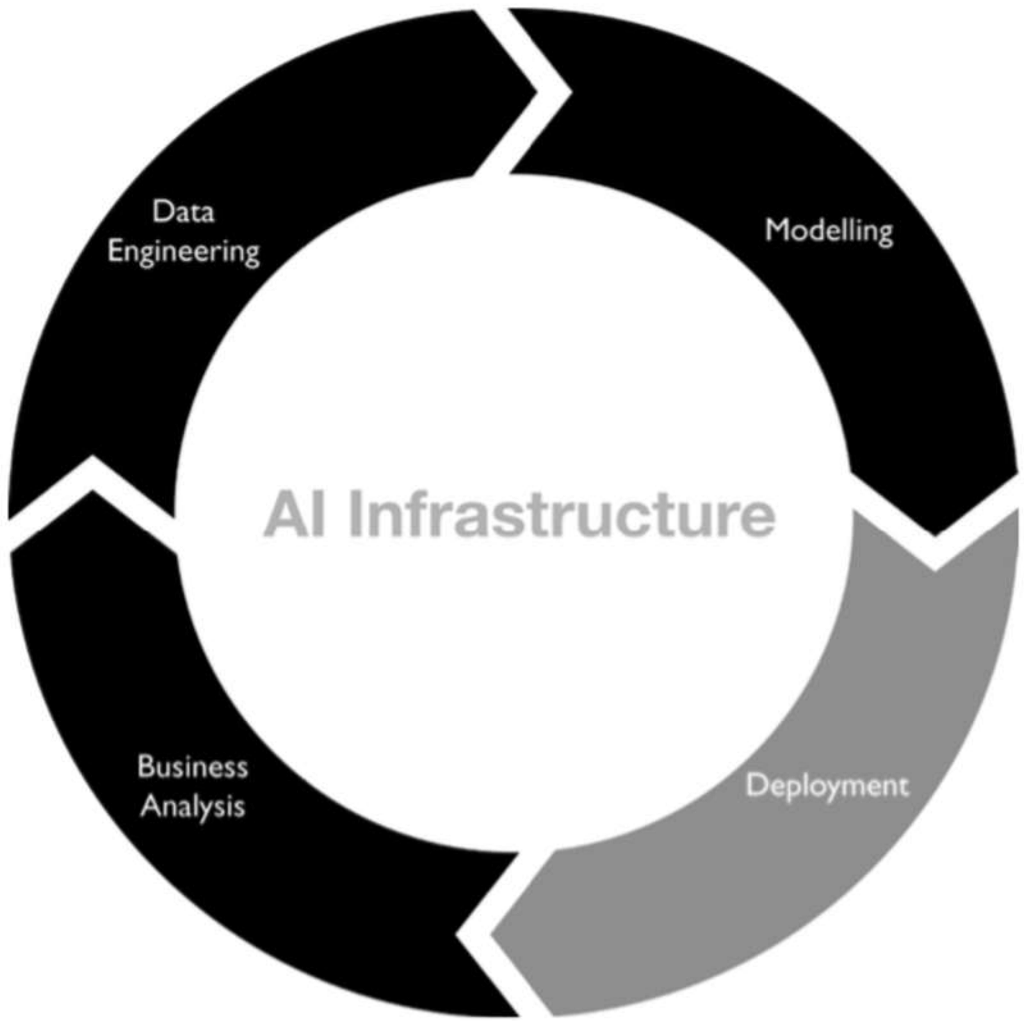


Data Scientist

SKILL PROFILE

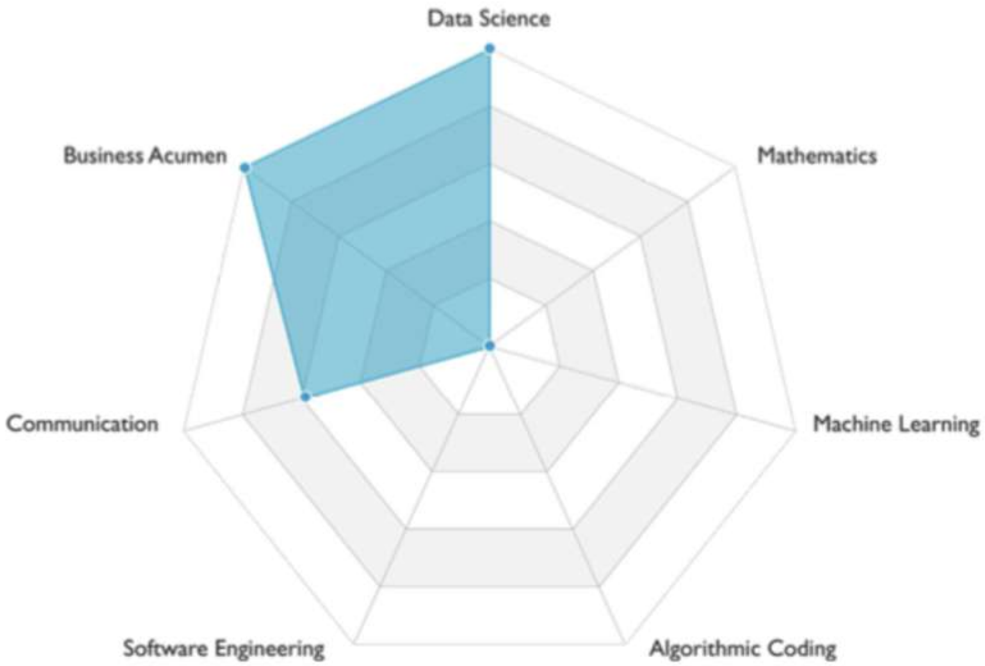


TASKS

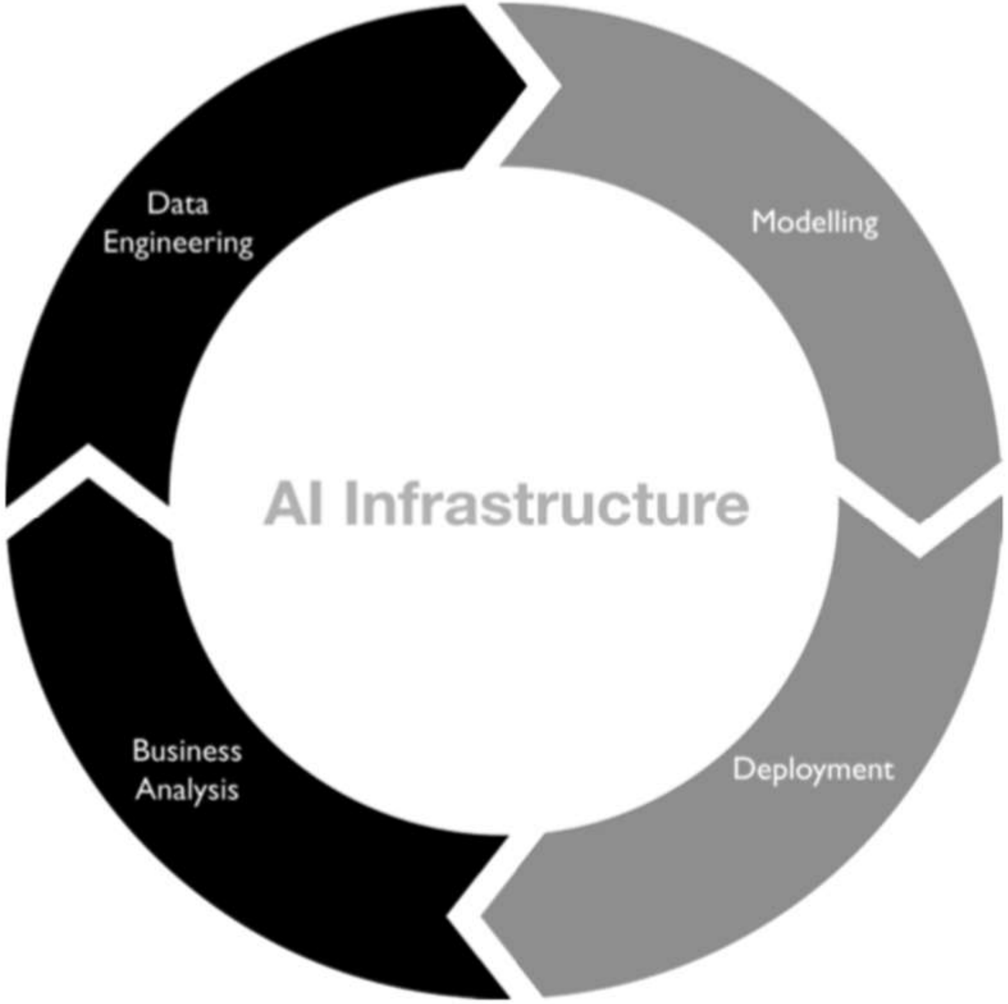


Data Analyst

SKILL PROFILE

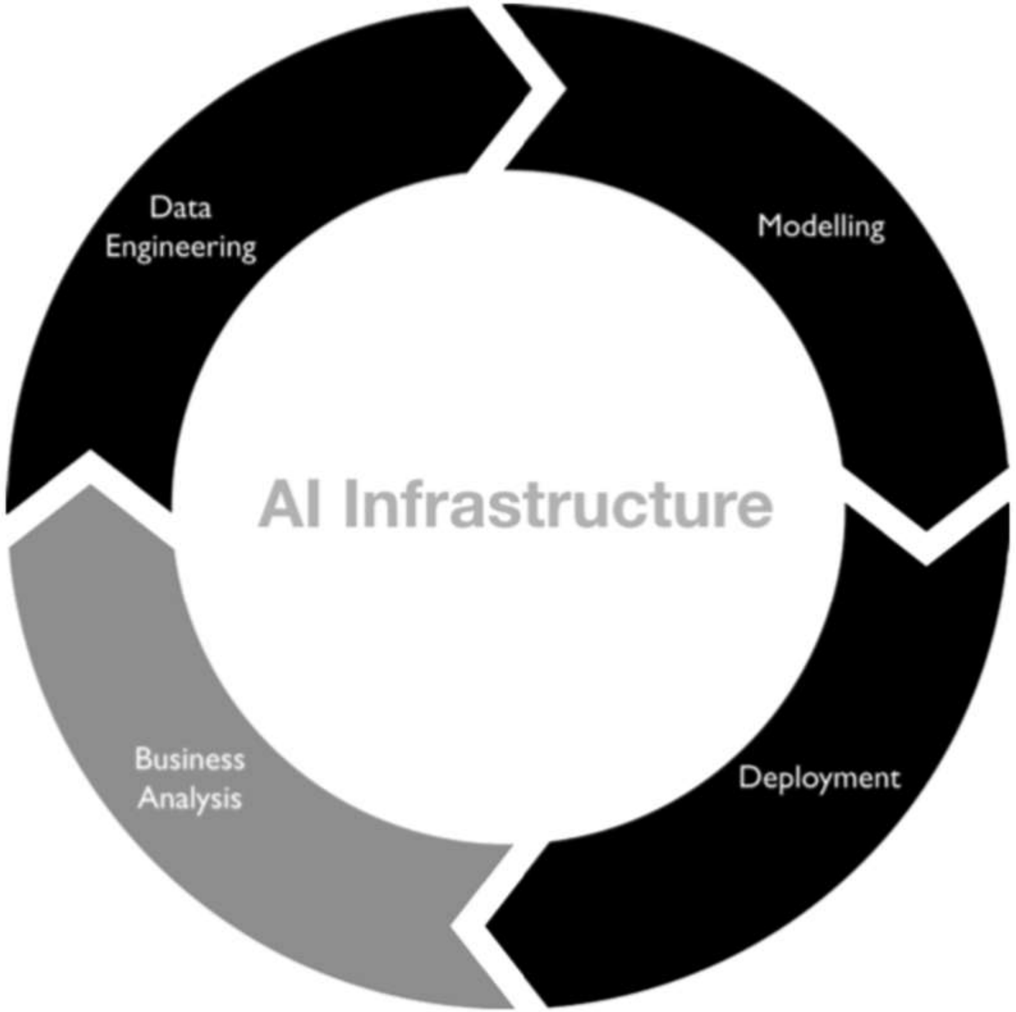


TASKS

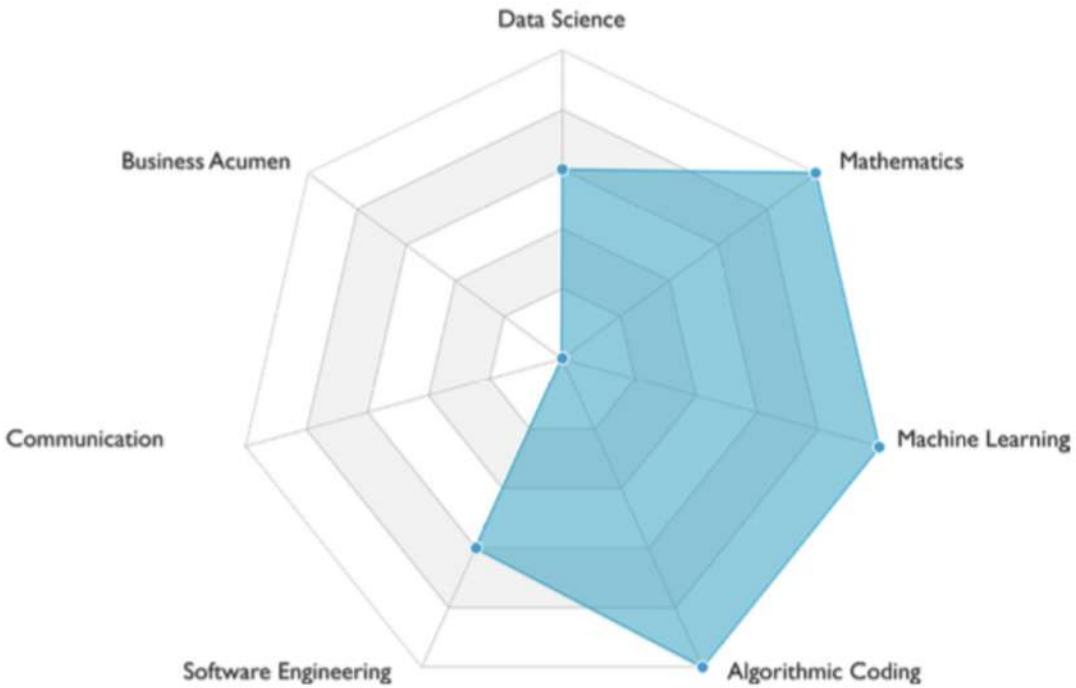


Machine Learning Engineer

TASKS

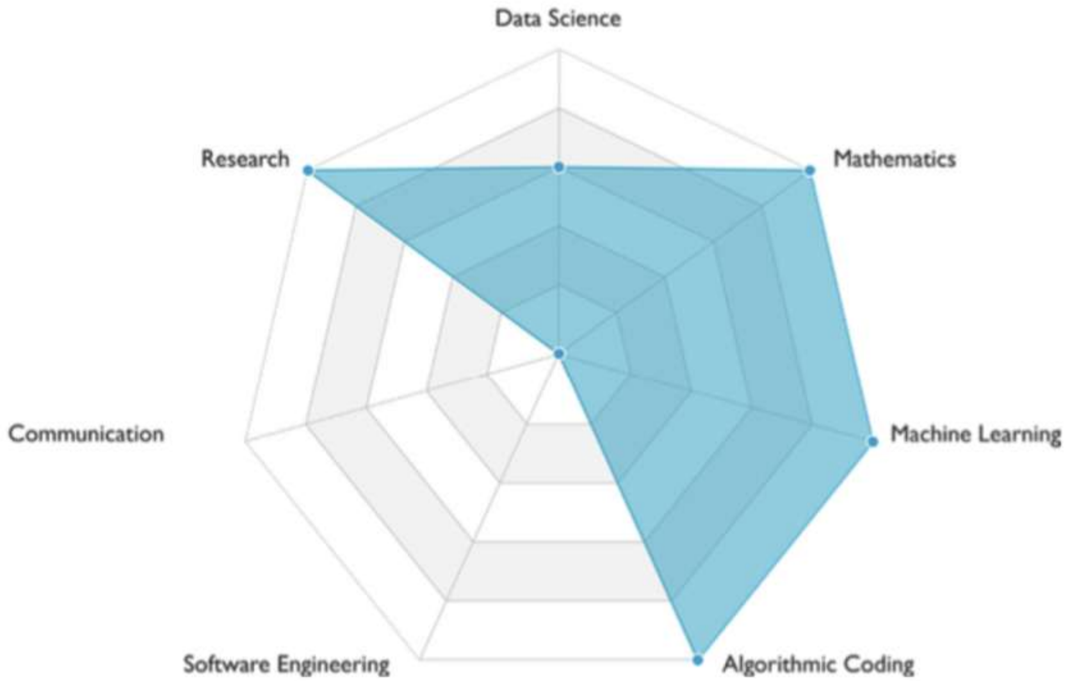


SKILL PROFILE

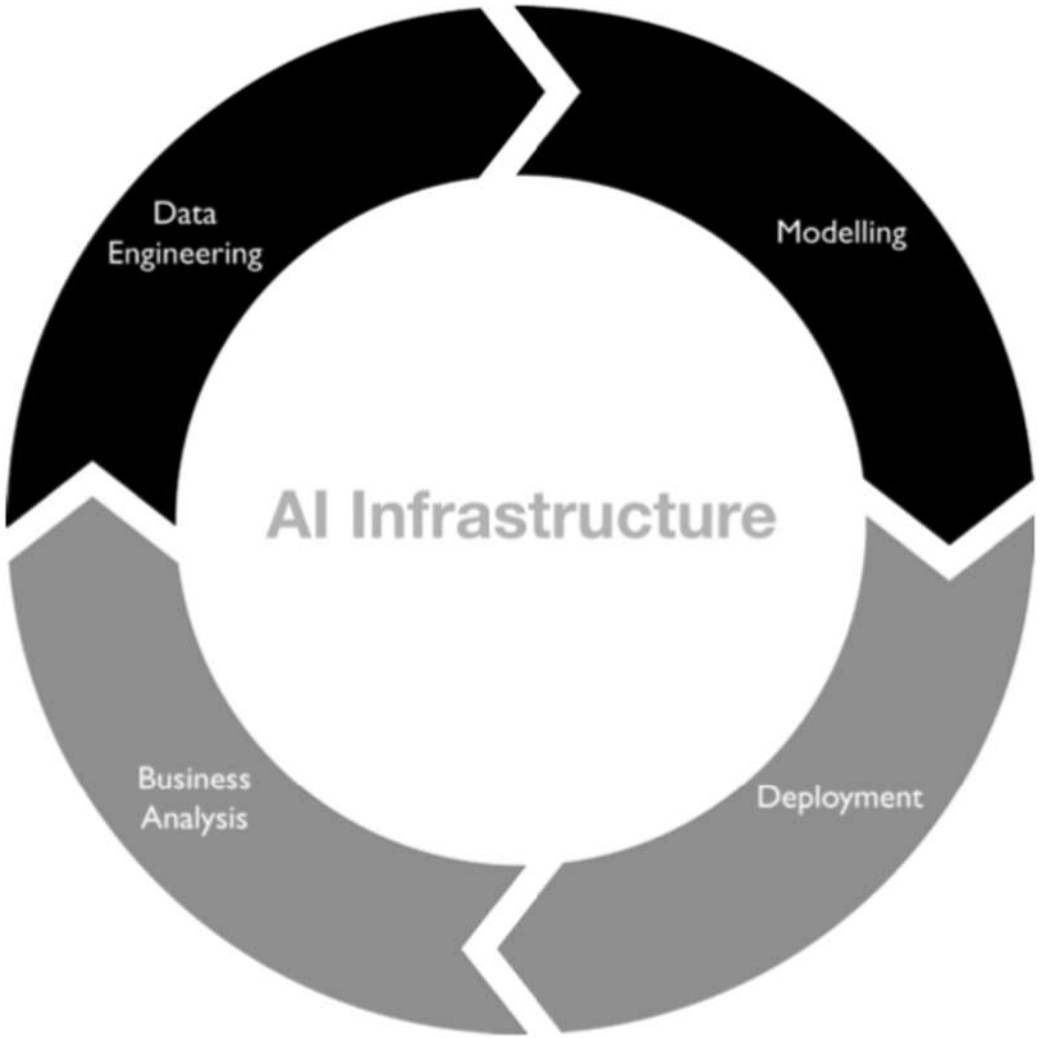


Machine Learning Researcher

SKILL PROFILE

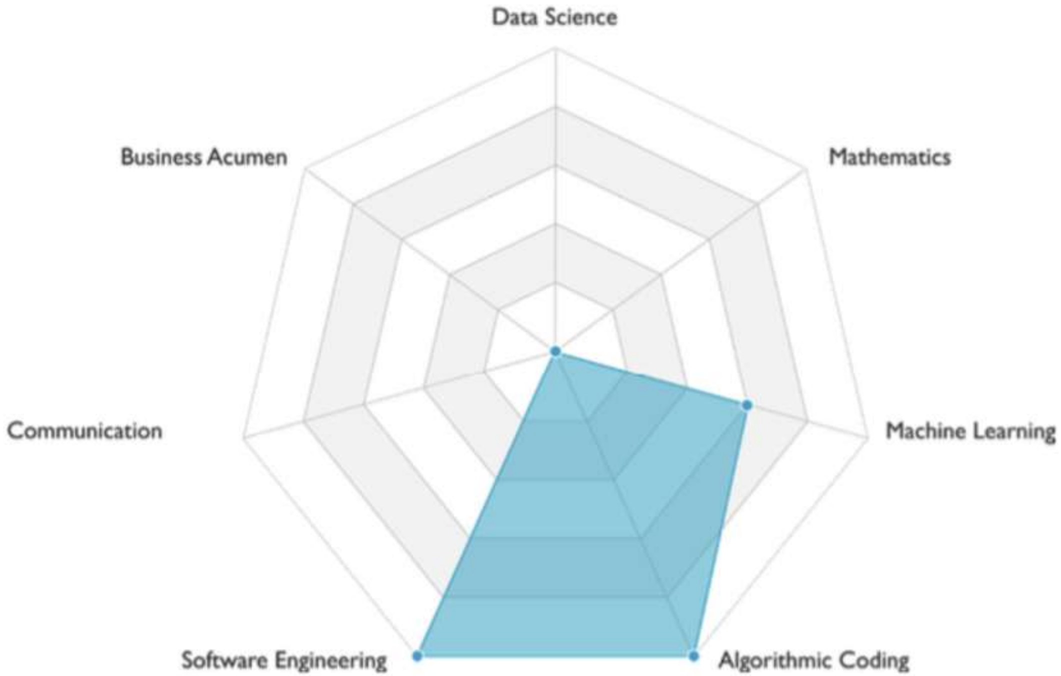


TASKS

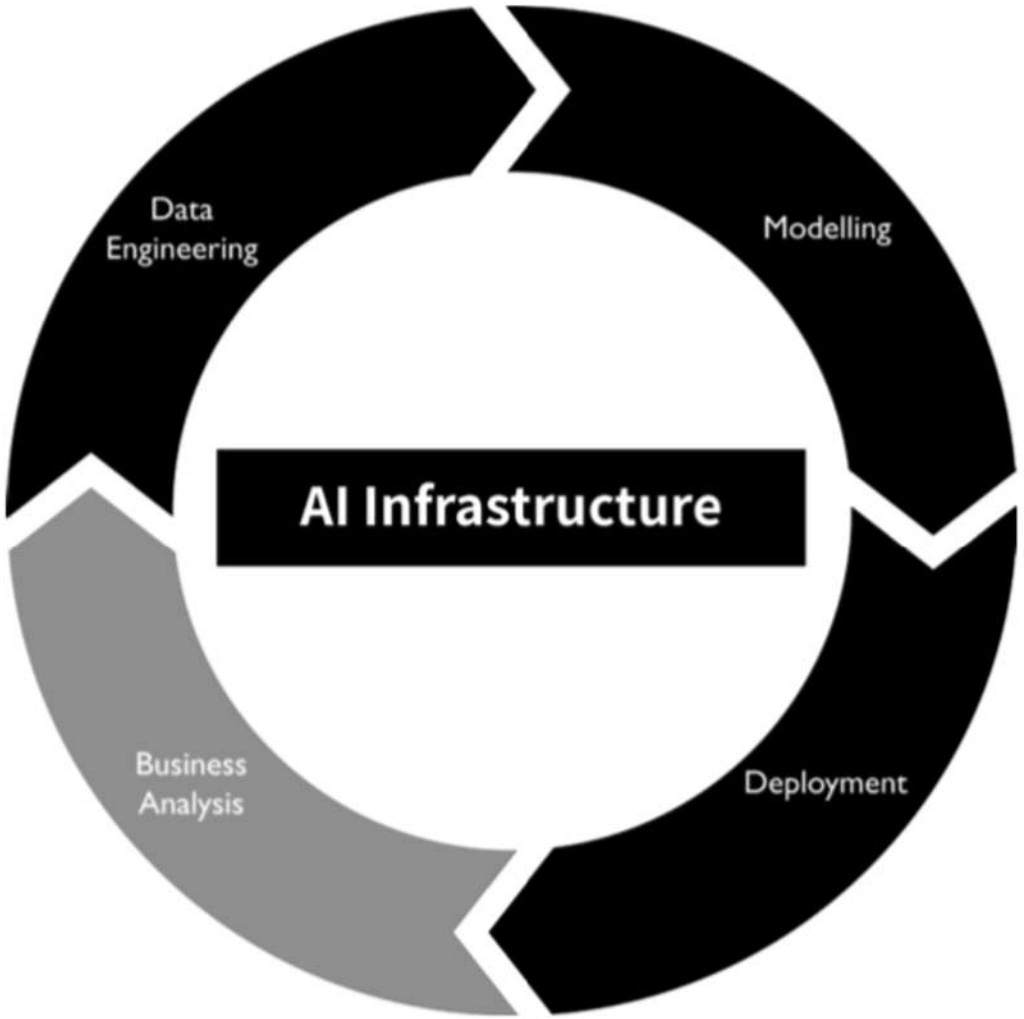


Software Engineer - Machine Learning

SKILL PROFILE

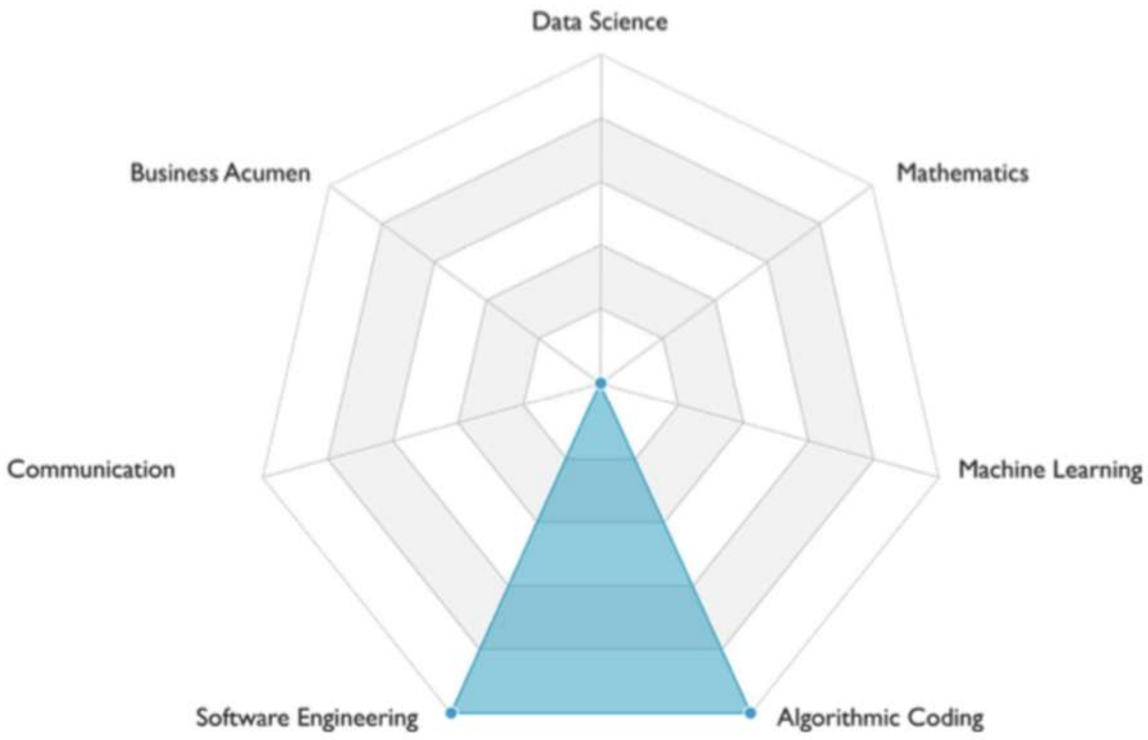


TASKS

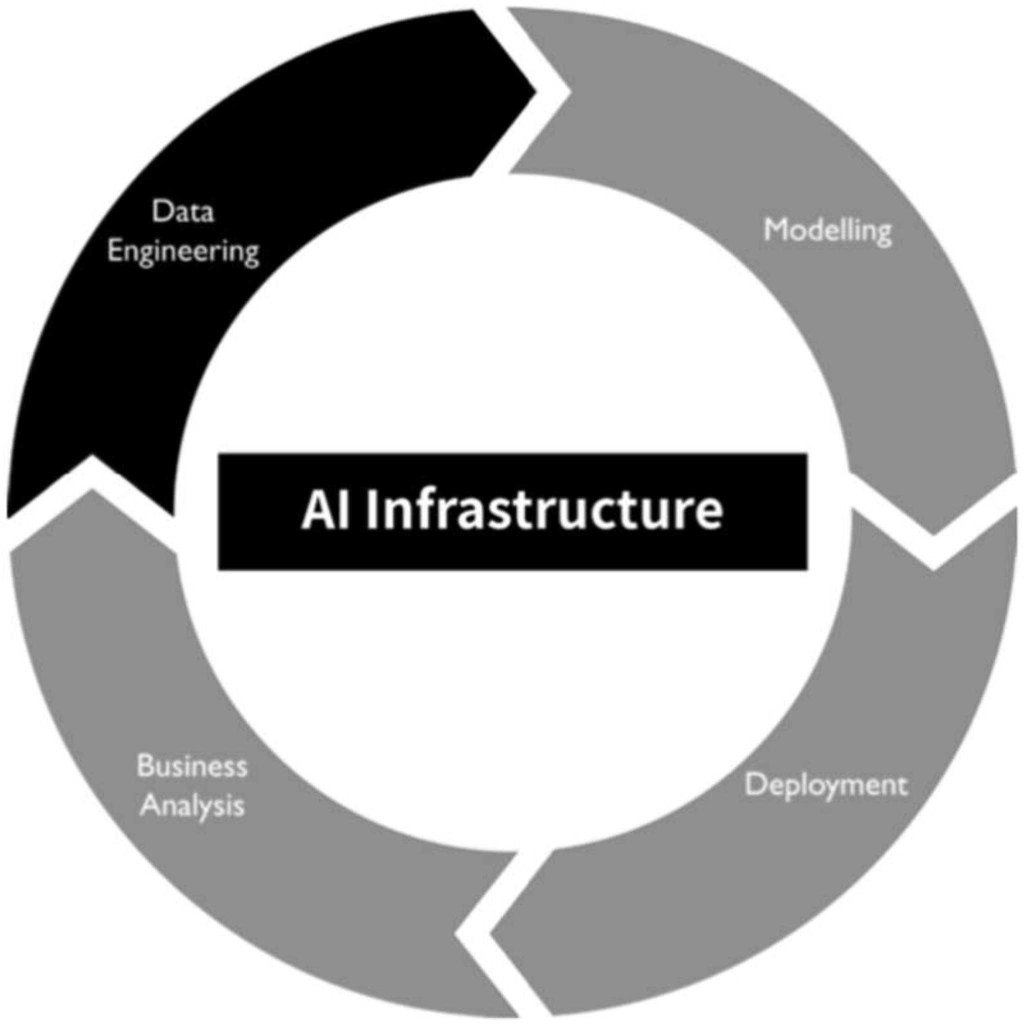


Software Engineer

SKILL PROFILE



TASKS



Thank you

AI Applications in Industry

Muhammad Ghifary, PhD

Head of Artificial Intelligence

muhammad.ghifary@bukalapak.com

Bukalapak