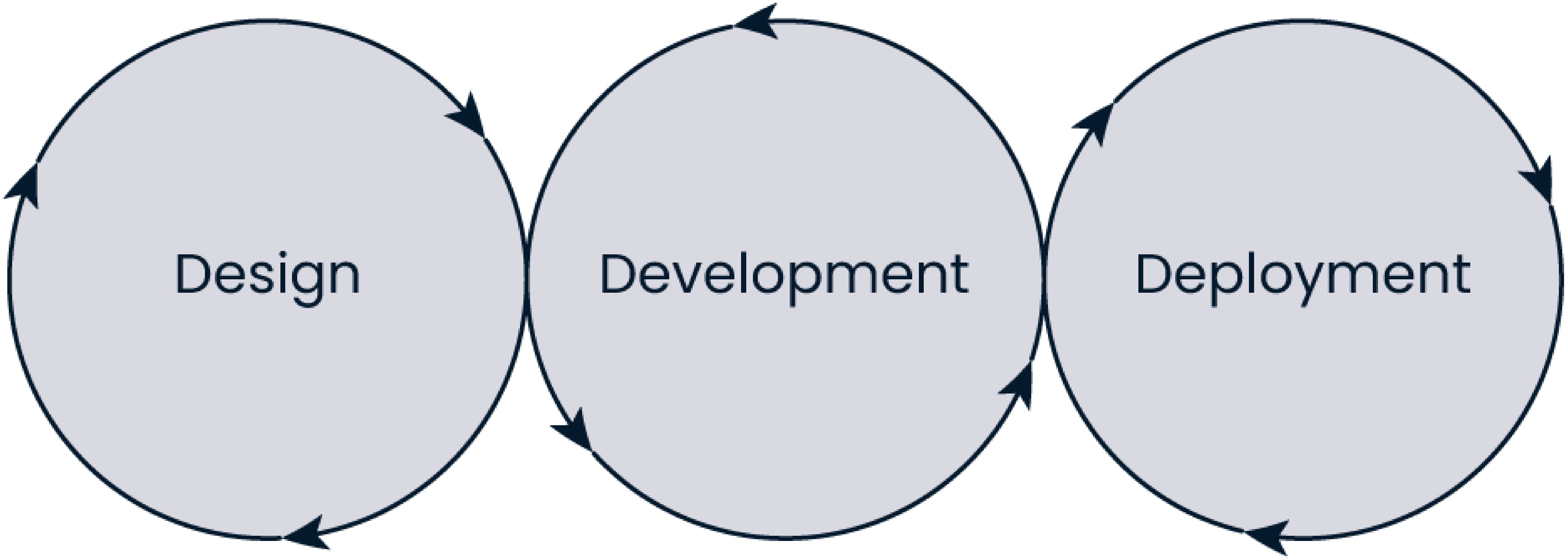
MLOps design

MLOPS CONCEPTS

Folkert Stijnman
ML Engineer

Machine learning design



- Added value
- Business requirements
- Key metrics
- Data processing

- Feature engineering
- Experiment tracking
- Model training & evaluation

- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring & retraining

Added value

Added value estimation

- Predict whether a customer will churn

Added value estimation

- Predict whether a customer will churn

100K customers

$10 per month

Added value estimation

- Predict whether a customer will churn

100K customers

$10 per month

80% accuracy predicting churn

1,000 customers churn

50% decrease of churn

Added value estimation

- Predict whether a customer will churn

100K customers

$10 per month

80% accuracy predicting churn

1,000 customers churn

50% decrease of churn

1,000 customers x 80% x 50% = 400 customers p/m

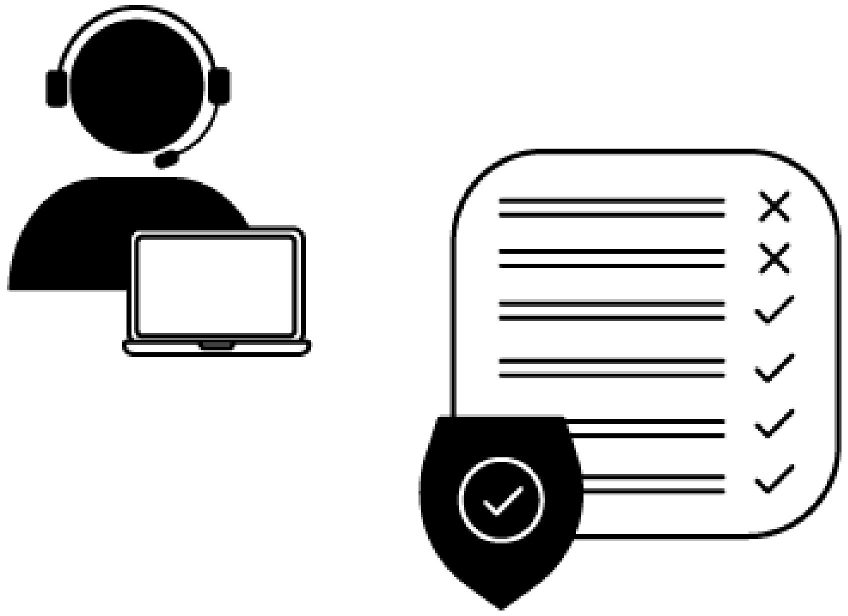
400 x discounted subscription $8 = $3200 per month

Business requirements

- End user
  - Speed
  - Accuracy
  - Transparency

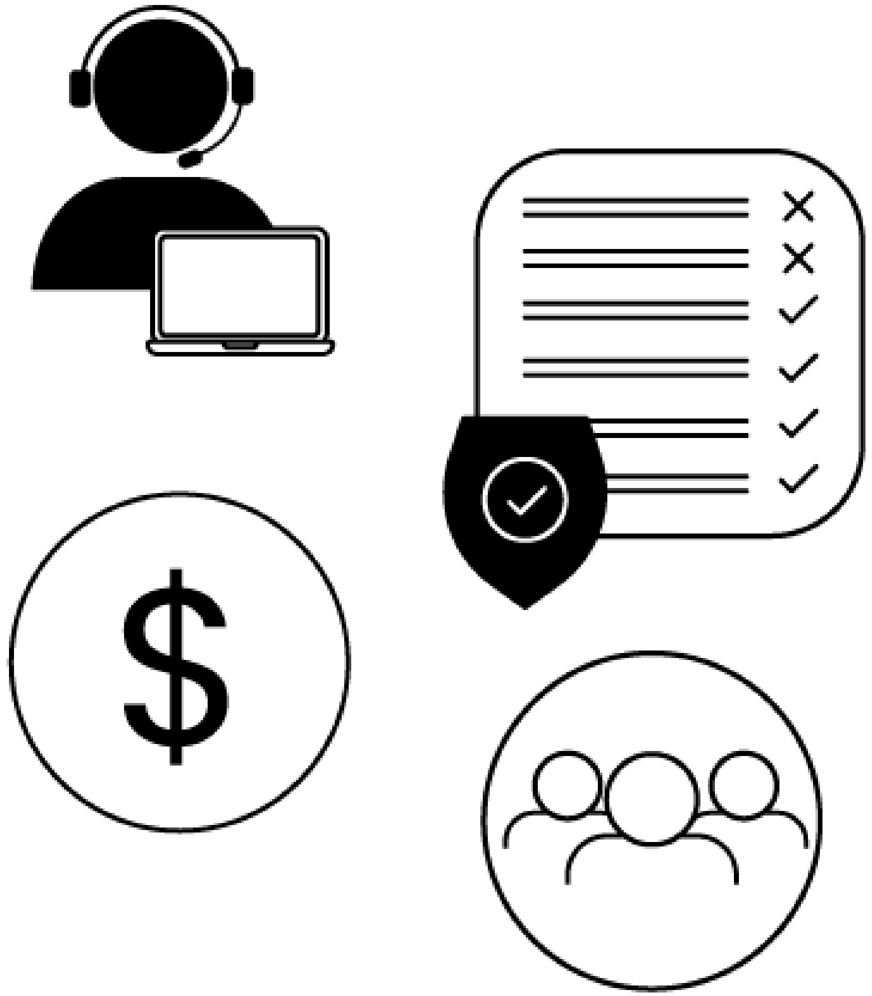
Business requirements

- End user
  - Speed
  - Accuracy
  - Transparency
- Compliance and regulations



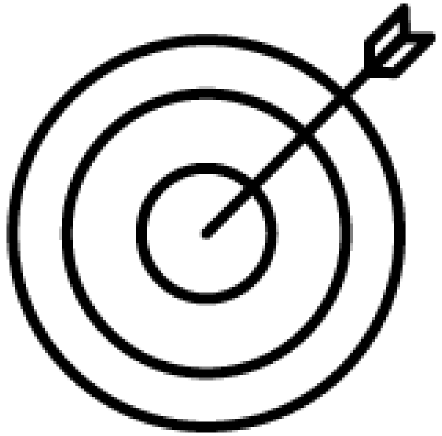
Business requirements

- End user
    - Speed
    - Accuracy
    - Transparency
- Compliance and regulations
- Budget
- Team size



Key metrics

Data
scientist

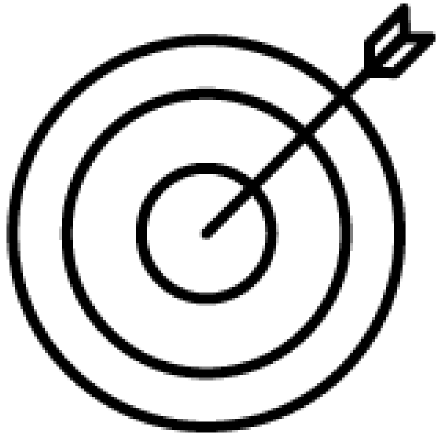


Accuracy

Key metrics

Data
scientist

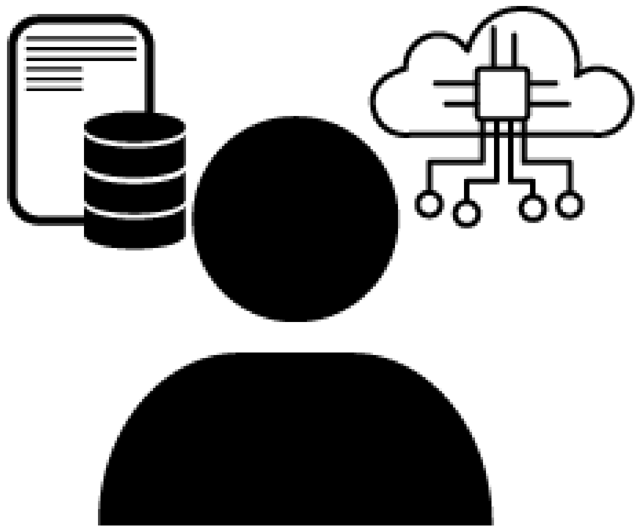
Subject matter
expert



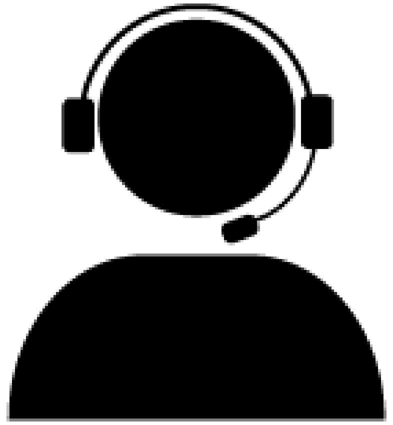
Accuracy

Customer happiness

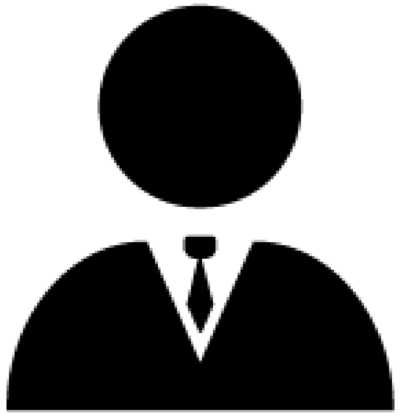
Key metrics



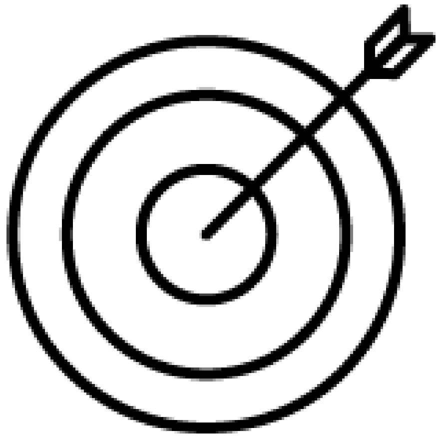
Data scientist



Subject matter expert

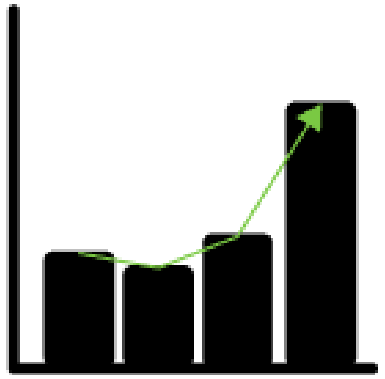


Business stakeholder



Accuracy

Customer happiness



Generated revenue

Let's practice!

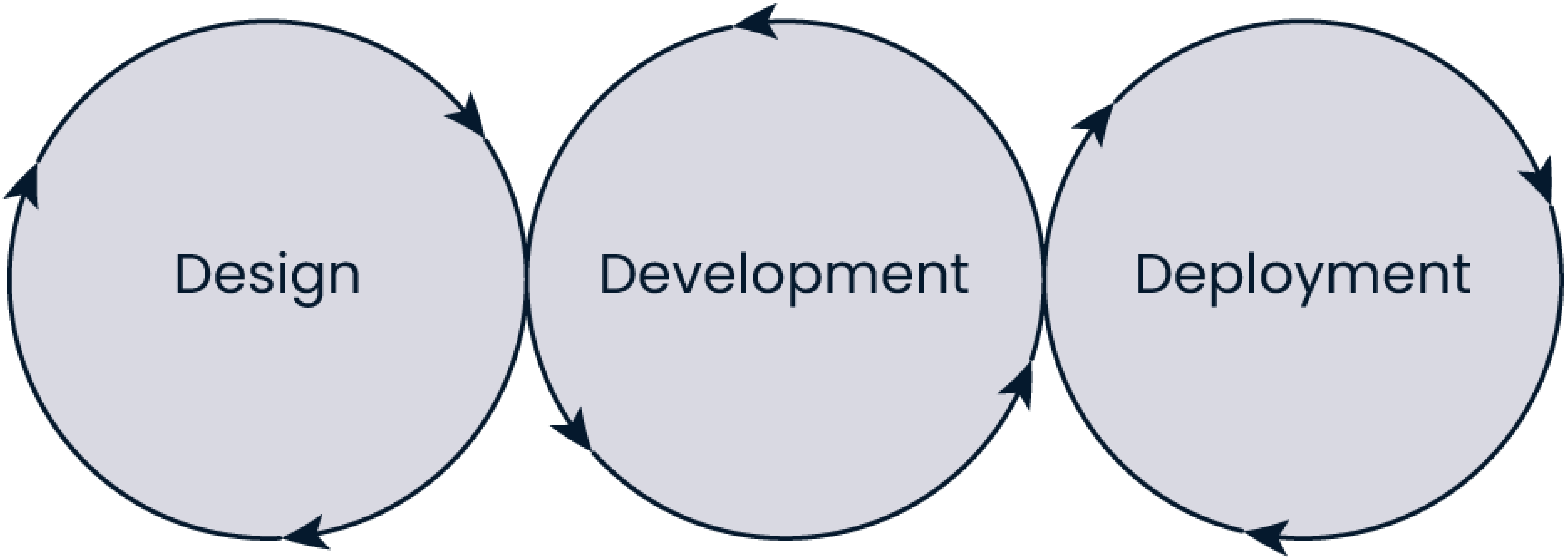
MLOPS CONCEPTS

Data quality and ingestion

MLOPS CONCEPTS

Folkert Stijnman
ML Engineer

Data quality and ingestion



- Added value
- Business requirements
- Key metrics
- Data processing

- Feature engineering
- Experiment tracking
- Model training & evaluation

- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring & retraining

What is data quality?

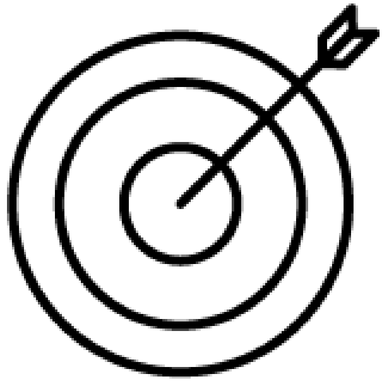
- "*Data quality refers to both the characteristics associated with high-quality data and the processes used to measure or improve the quality of data.*" - DAMA Dictionary of Data Management
- The core of the machine learning model
- Poor data quality impacts the model

Data quality dimensions

- Accuracy
- Completeness
- Consistency
- Timeliness

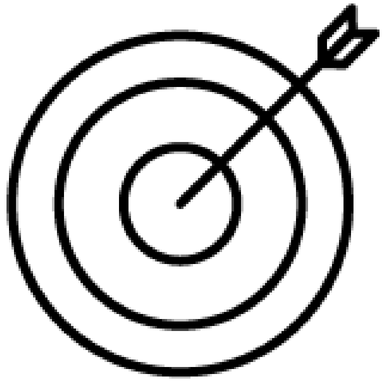
Data quality dimensions

- **Accuracy:** representation of reality
- **Completeness**
- **Consistency**
- **Timeliness**



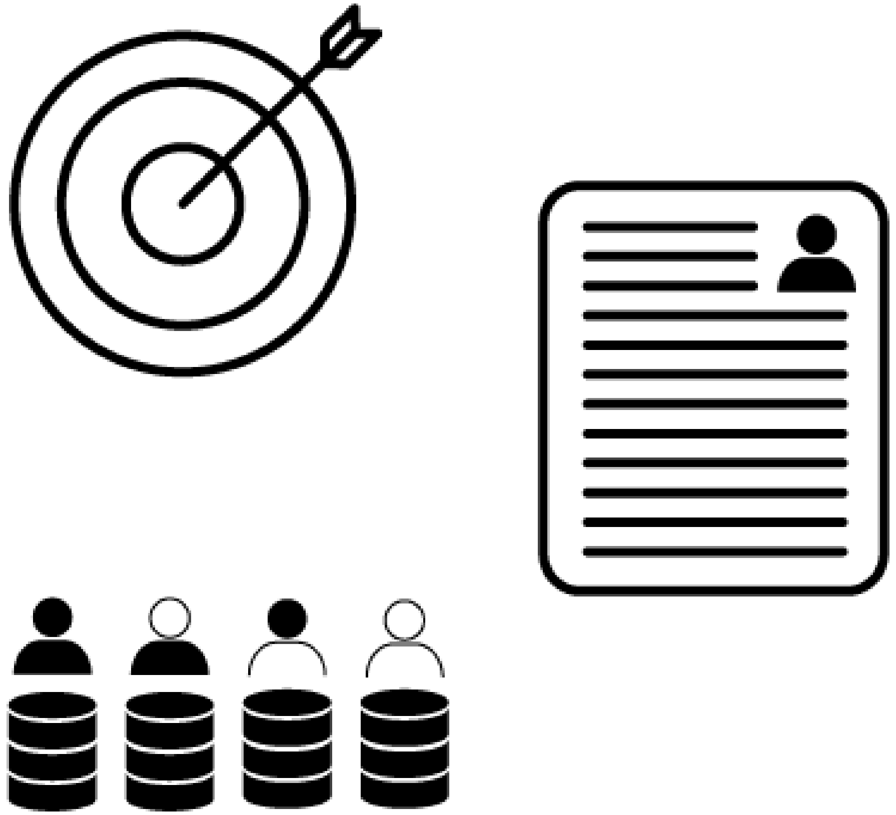
Data quality dimensions

- **Accuracy:** representation of reality
- **Completeness:** thorough description
- **Consistency**
- **Timeliness**



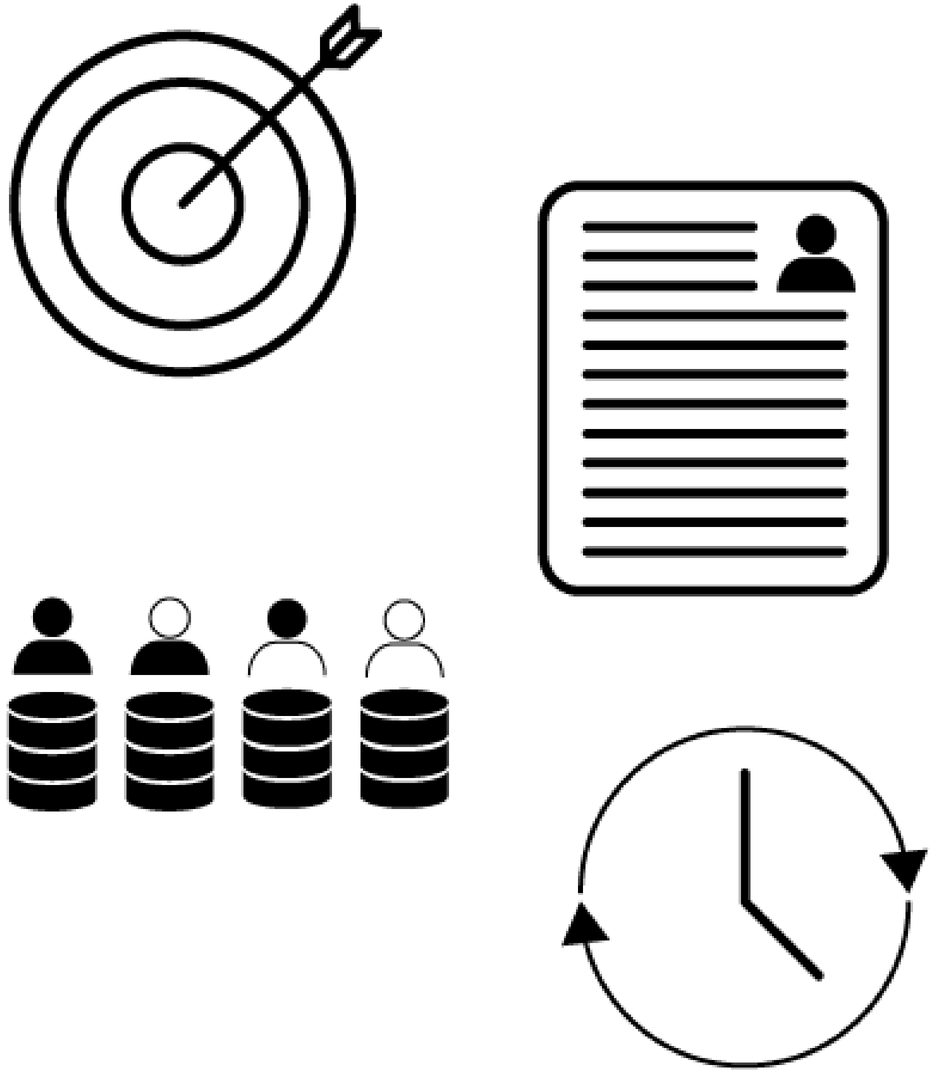
Data quality dimensions

- **Accuracy:** representation of reality
- **Completeness:** thorough description
- **Consistency:** similar definitions
- **Timeliness**



Data quality dimensions

- **Accuracy:** representation of reality
- **Completeness:** thorough description
- **Consistency:** similar definitions
- **Timeliness:** availability of data

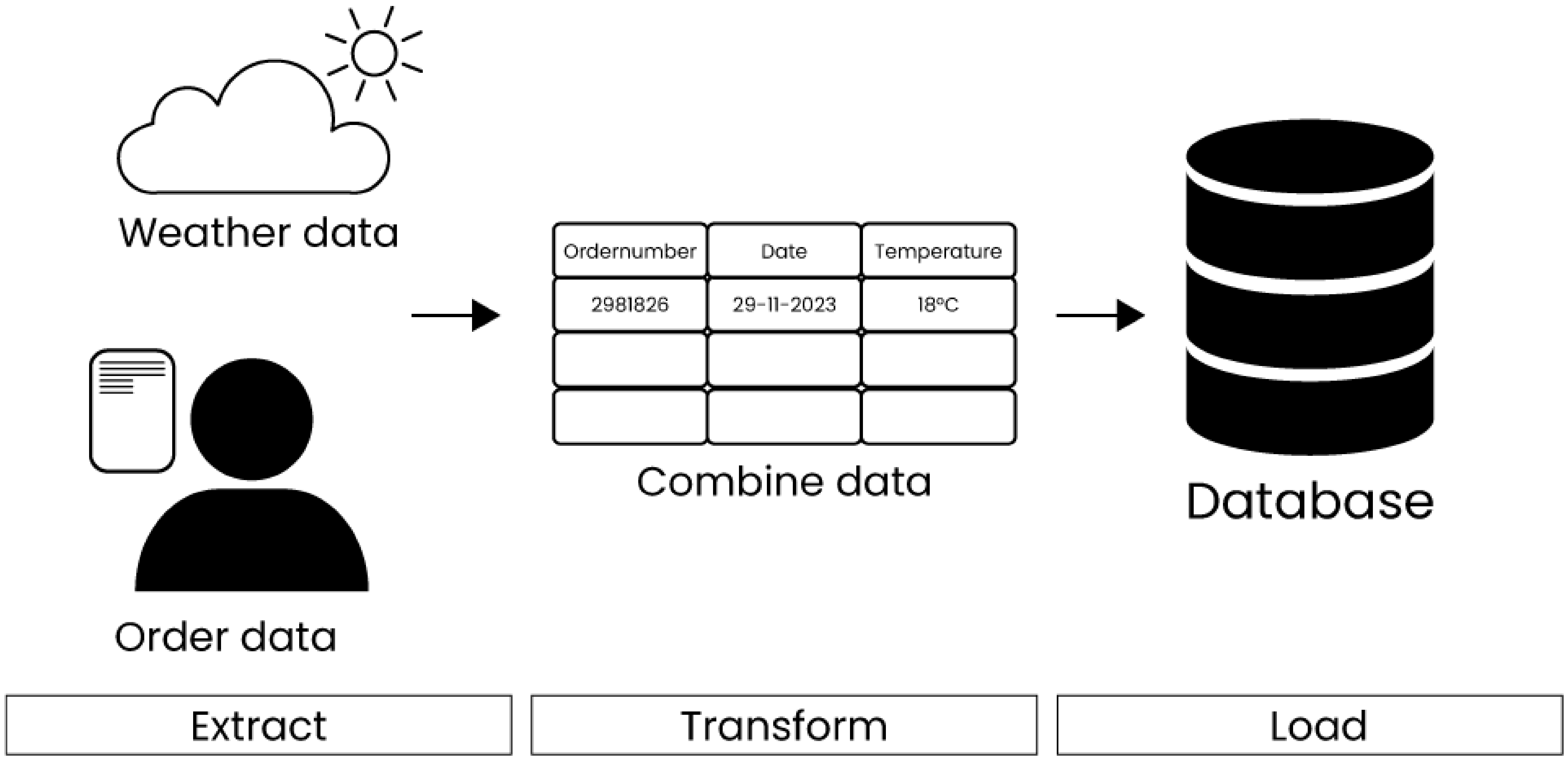


Data quality dimensions example

Dimension	Example question to answer	Example of dimension quality
Accuracy	Does our data correctly describe the customer?	The customer's age in the data is 18, but is actually 32.
Completeness	Is there any customer data missing?	For 80% of the customers, we don't have a last name.
Consistency	Is the definition of the customer synchronized throughout the company?	The customer is stated as active in one database but not active in another.
Timeliness	When is the customer ordering data available?	The customer orders are synchronized at the end of the day but are not available in real-time.

Low data quality is not the end of the project!

Data ingestion



Let's practice!

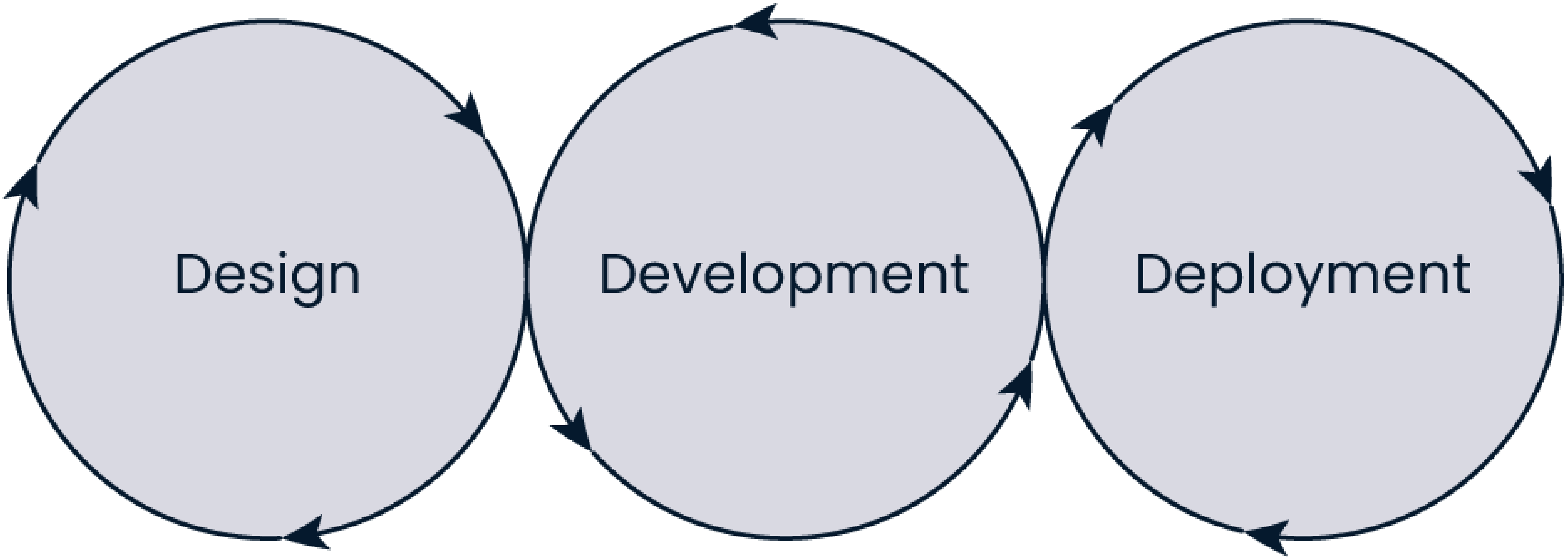
MLOPS CONCEPTS

Feature engineering and the feature store

MLOPS CONCEPTS

Folkert Stijnman
ML Engineer

Feature engineering



- Added value
- Business requirements
- Key metrics
- Data processing

- Feature engineering
- Experiment tracking
- Model training & evaluation

- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring & retraining

Feature engineering

... is the process of selecting, manipulating, and transforming raw data into features.

A feature is a variable, such as the column in a table

Customer data

Customer ID	Number of orders	Total expenditure
0	4	$1982
1	2	$8545
2	8	$102
...	...	...

Customer data

Customer ID	Number of orders	Total expenditure
0	4	$1982
1	2	$8545
2	8	$102
...	...	...

→

Average expenditure
$495.50
$4272.50
$12.75
...

Feature engineering weigh-off

More features can

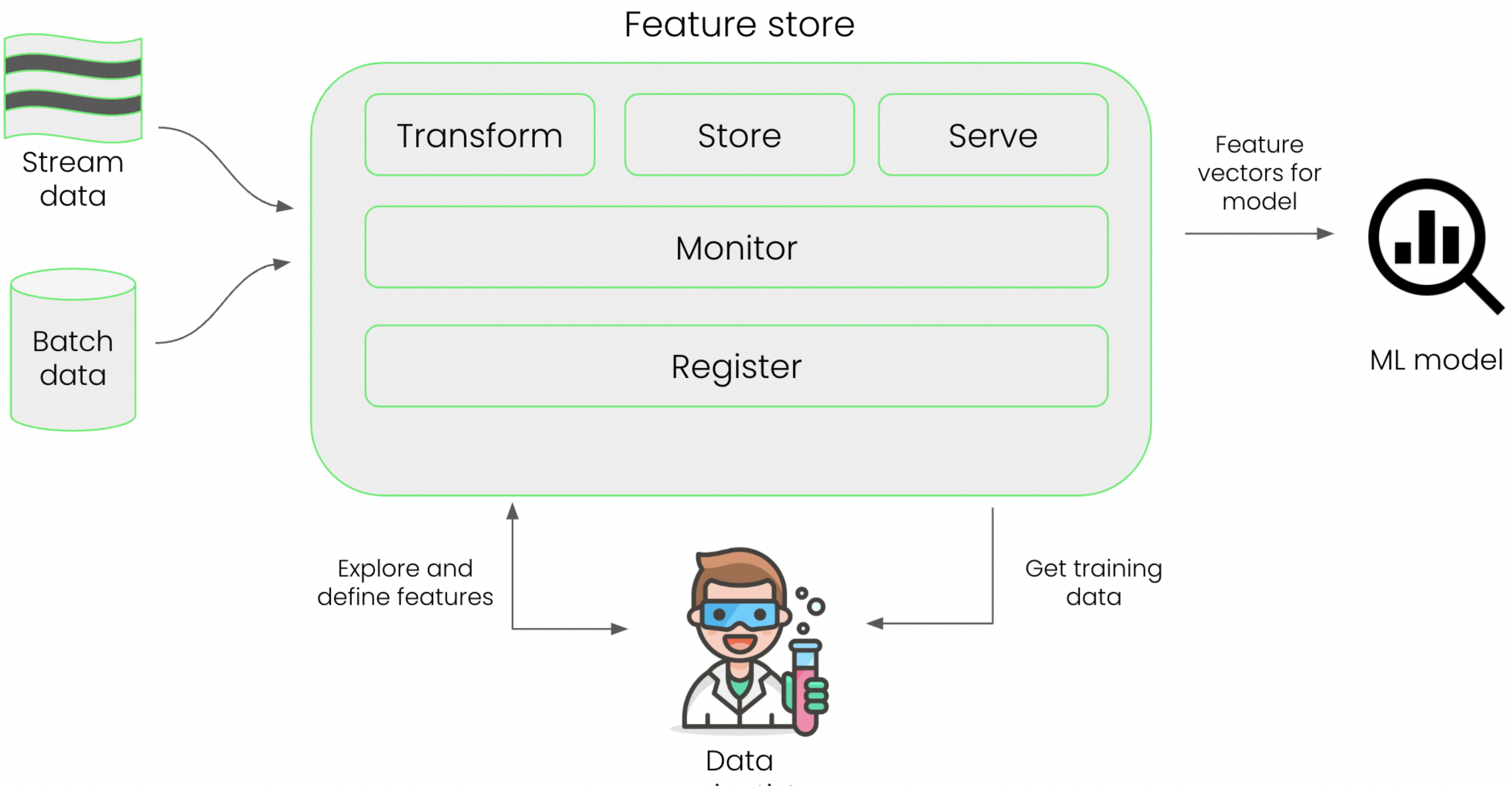
- produce a very accurate model
- achieve more stability
- be more expensive due to additional pre-processing steps
- require more maintenance
- lead to noise, or over-engineering

What if the number of ML projects increases?

Feature A	Feature B	Feature C
0,298	92,5	1
0,721	24,0	0
0,980	56,8	0
...	...	...

Feature A	Feature B	...	Feature Z
0,298	92,5	...	2
0,721	24,0	...	8
0,980	56,8	...	5
...	...	...	...

The feature store



When to use a feature store?

- Computational cost of computing or transforming features
- Amount of projects and thus ML models for the same features

Let's practice!

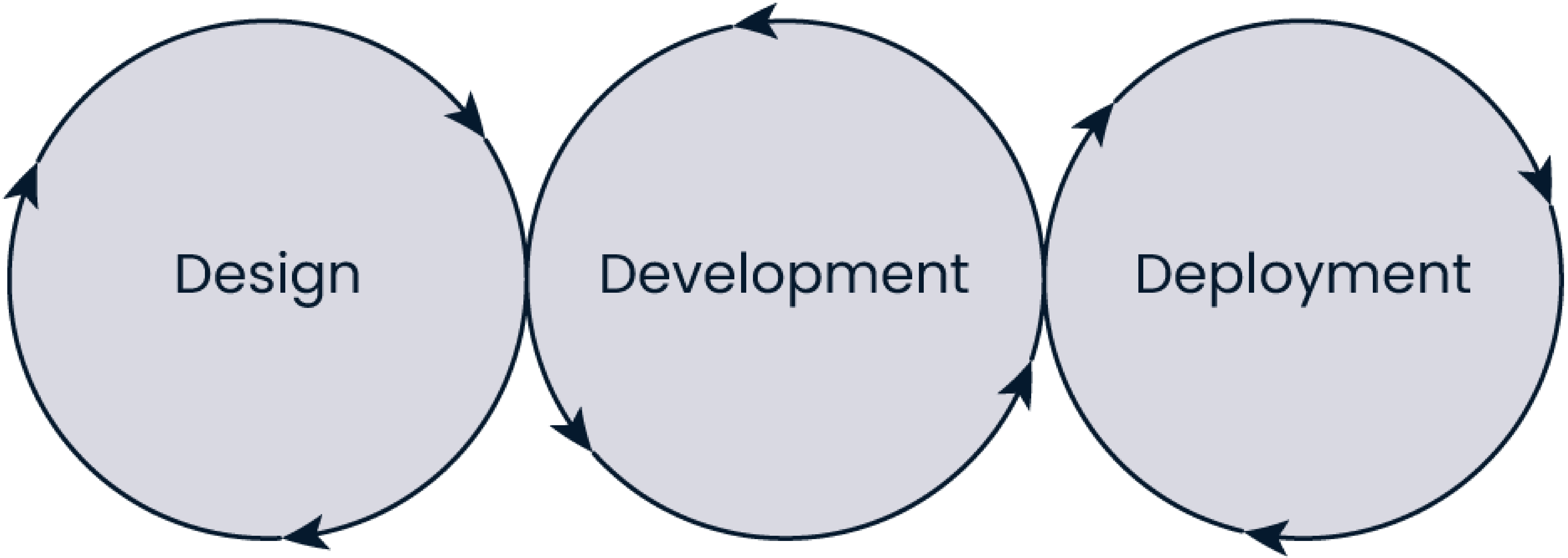
MLOPS CONCEPTS

Experiment tracking

MLOPS CONCEPTS

Folkert Stijnman
ML Engineer

The machine learning experiment



- Added value
- Business requirements
- Key metrics
- Data processing

- Feature engineering
- Experiment tracking
- Model training & evaluation

- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring & retraining

Why is experiment tracking important?

In each experiment, the following factors can be configured:

Machine learning models

Model hyperparameters

Versions of data

Execution scripts

Environment configurations

- The amount of different configurations can become huge
- Each experiment can have a different outcome

Using experiment tracking in the ML lifecycle

Experiment tracking can help to:

- Compare and evaluate experiments
- Reproduce results from earlier experiments
- Collaborate on experiments with developers and stakeholders
- Report on results to stakeholders

How to track experiments?

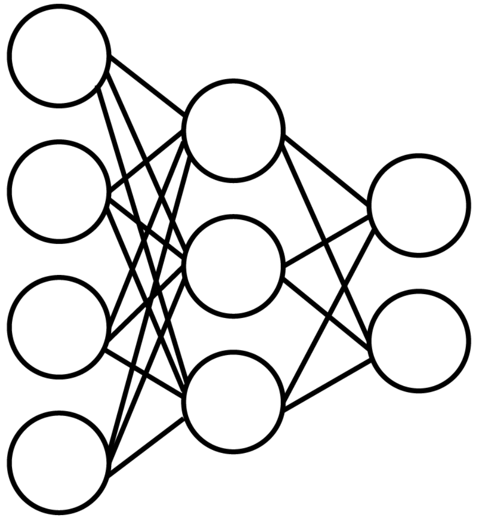
Tool	Pro	Con
Spreadsheet	Straightforward, easy to use	Require a lot of manual work
Proprietary platform	Custom solution specific for our process	Require time and effort
Experiment tracking tool	Specifically designed for experiments	Can be expensive

- Results of experiment tracking are stored in a *metadata store*

A machine learning experiment

Experiment 1

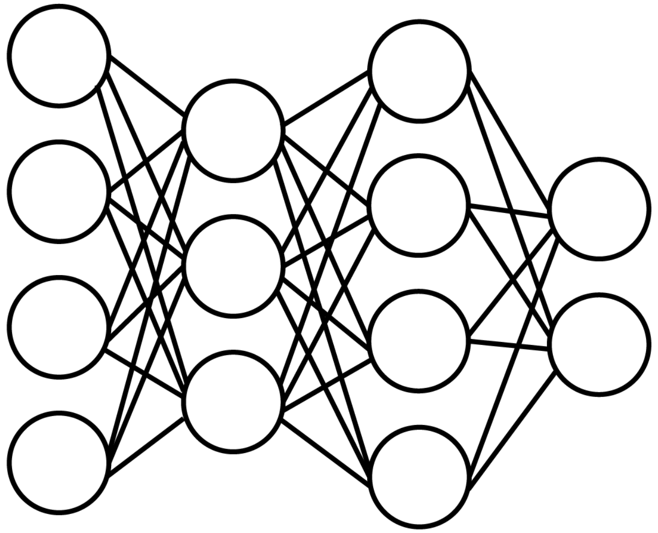
1.000 images of cats & dogs



A neural network with 1 hidden layer

Experiment 2

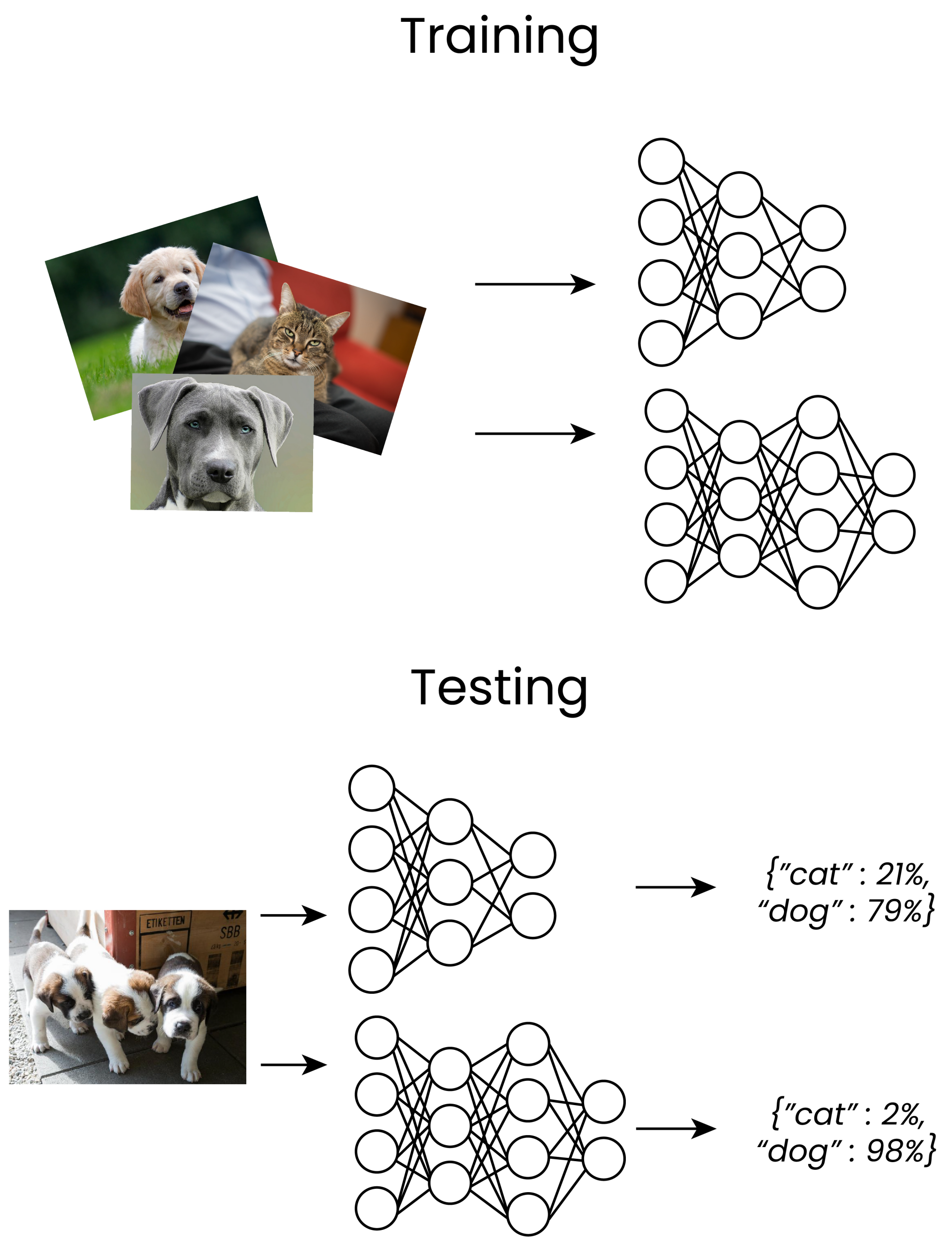
2.000 images of cats & dogs
(including puppies & kittens)



A neural network with 2 hidden layers

The experiment process

1. Formulate a hypothesis: *"We expect that..."*
2. Gather images and labels
3. Define experiments, e.g., types of models, hyperparameters, datasets
4. Setup experiment tracking
5. Train the machine learning model(s)
6. Test the models on a hold-out test set
7. Register the most suitable model
8. Visualize and report back to team and stakeholders, and determine next steps



Let's practice!

MLOPS CONCEPTS