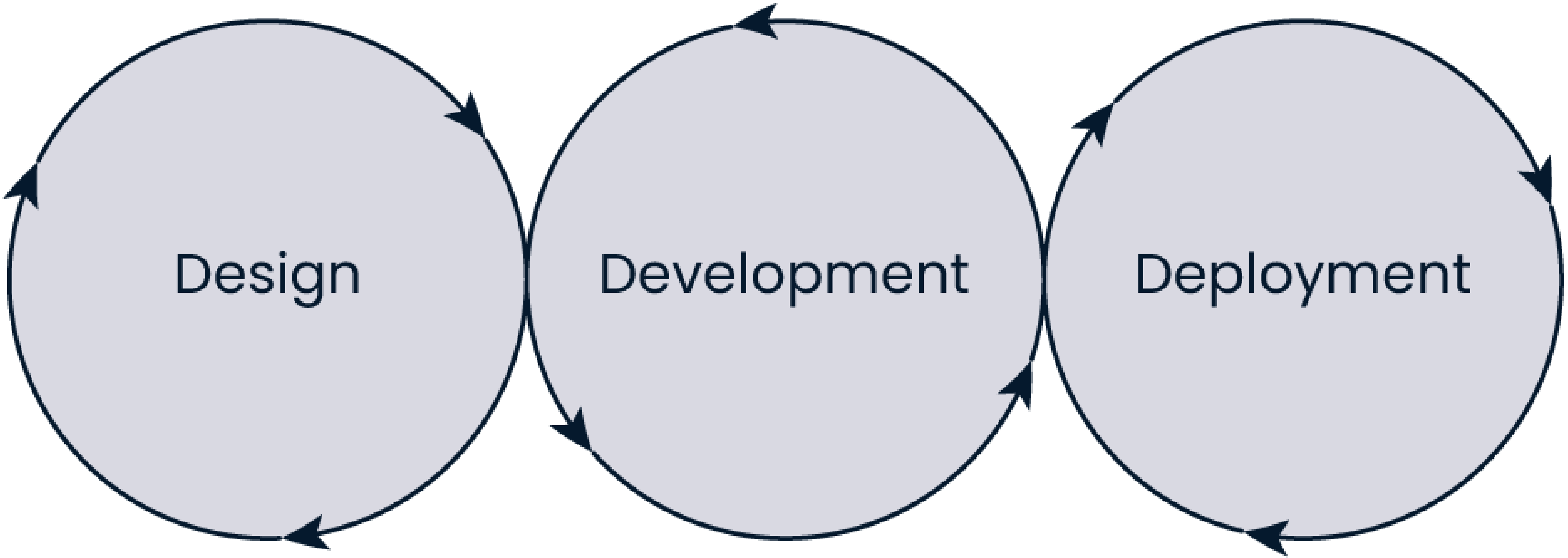
MLOps design

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

- Estimate the expected value
- ML is experimental and uncertain
- Aids in resource allocation, prioritization, and setting expectations

Business requirements

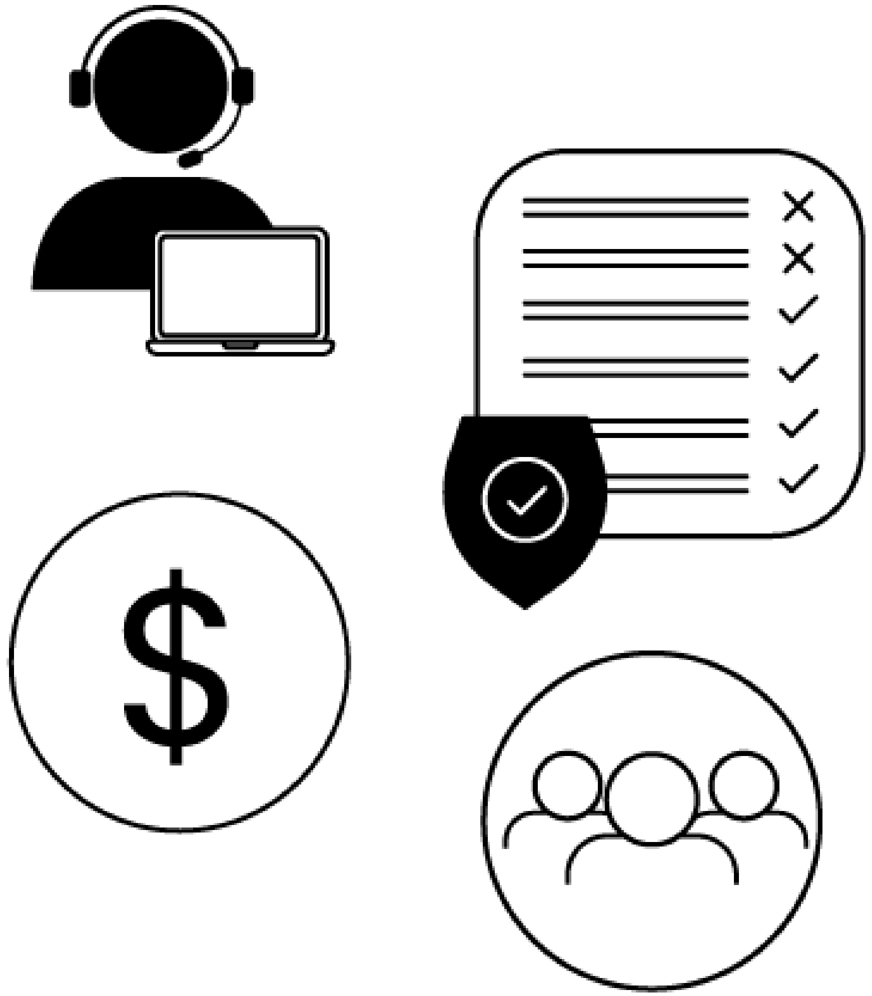
- End user
  - Speed
  - Accuracy
  - Transparency

Business requirements

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

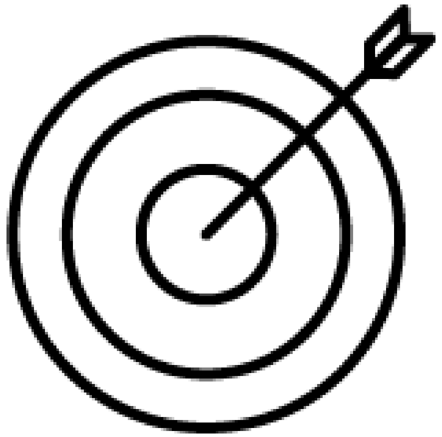
Business requirements

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



Key metrics

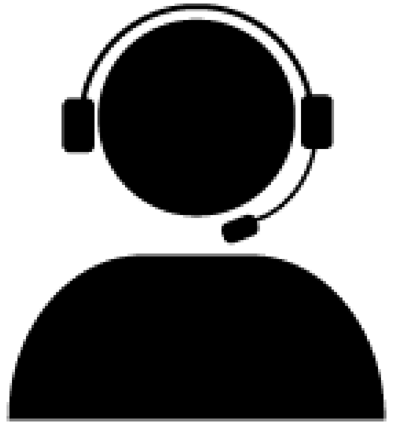
Data
scientist



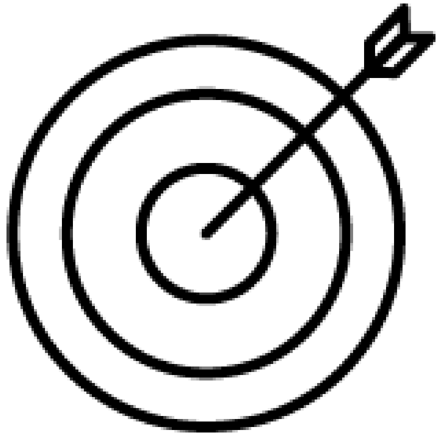
Accuracy

Key metrics

Data
scientist



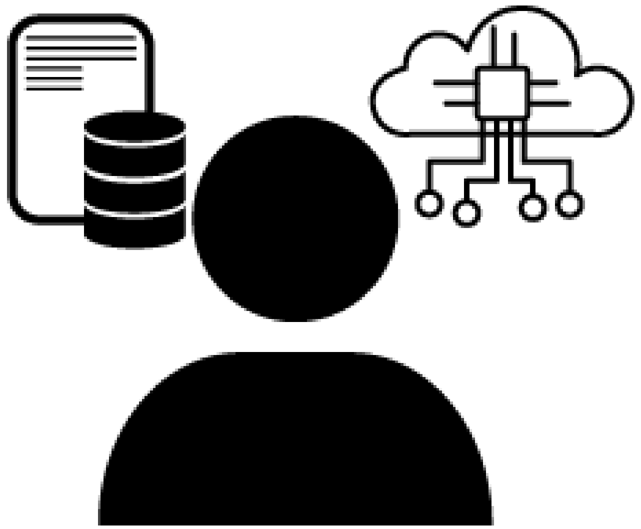
Subject matter
expert



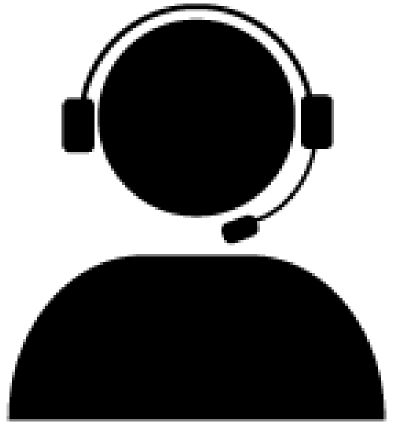
Accuracy

Customer happiness

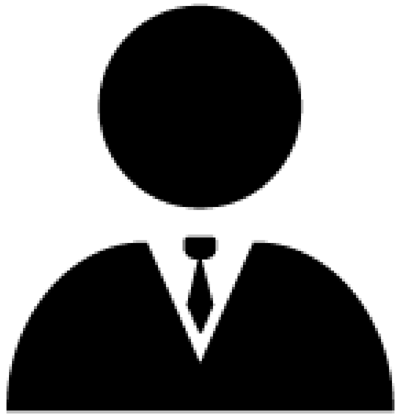
Key metrics



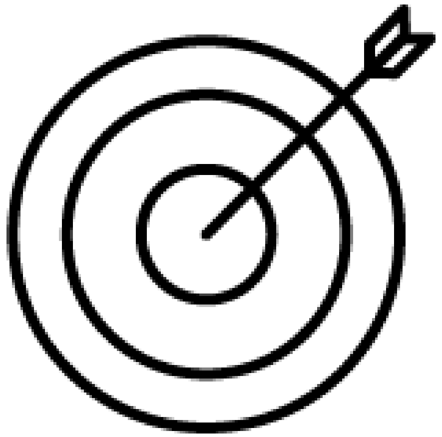
Data
scientist



Subject matter
expert

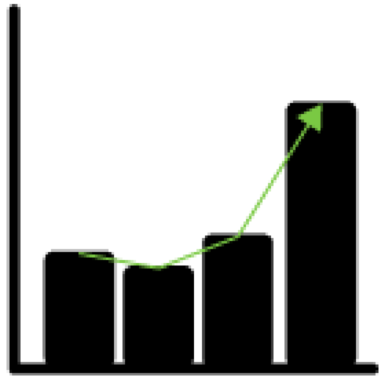


Business
stakeholder



Accuracy

Customer happiness



Generated revenue

Let's practice!

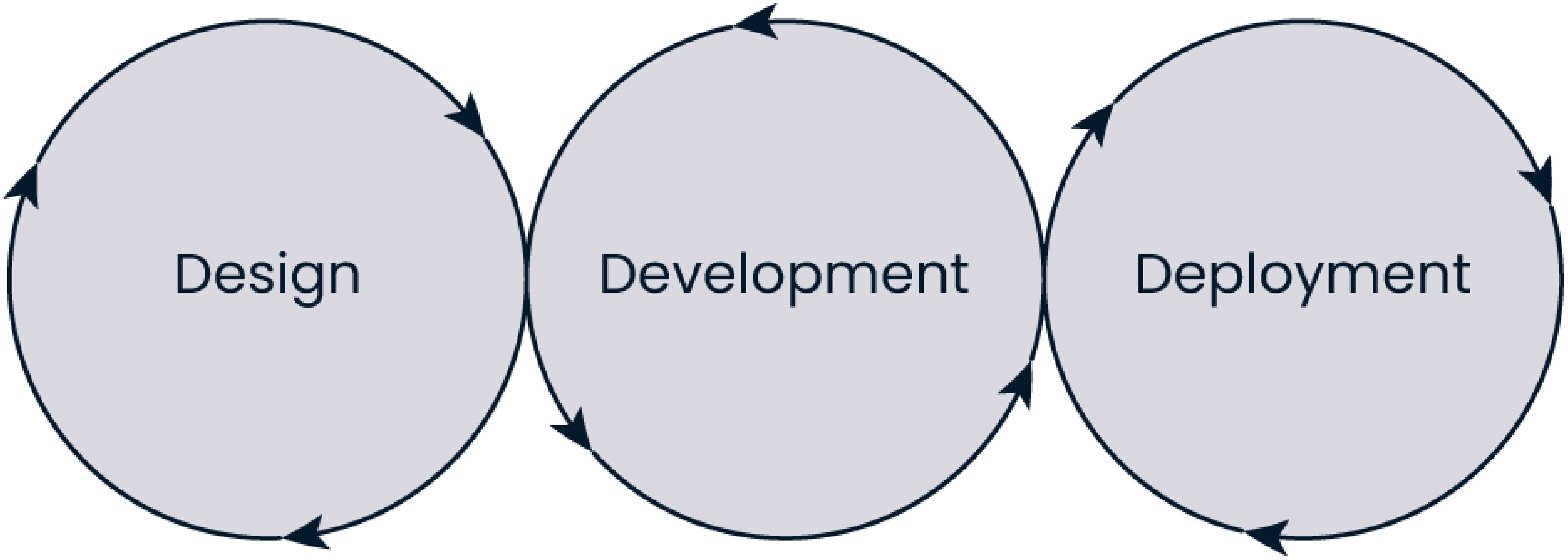
MLOPS CONCEPTS

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 is a measure of how well data serves its intended purpose
- Evaluated through various dimensions
- Quality of ML model depends on data

Data quality dimensions

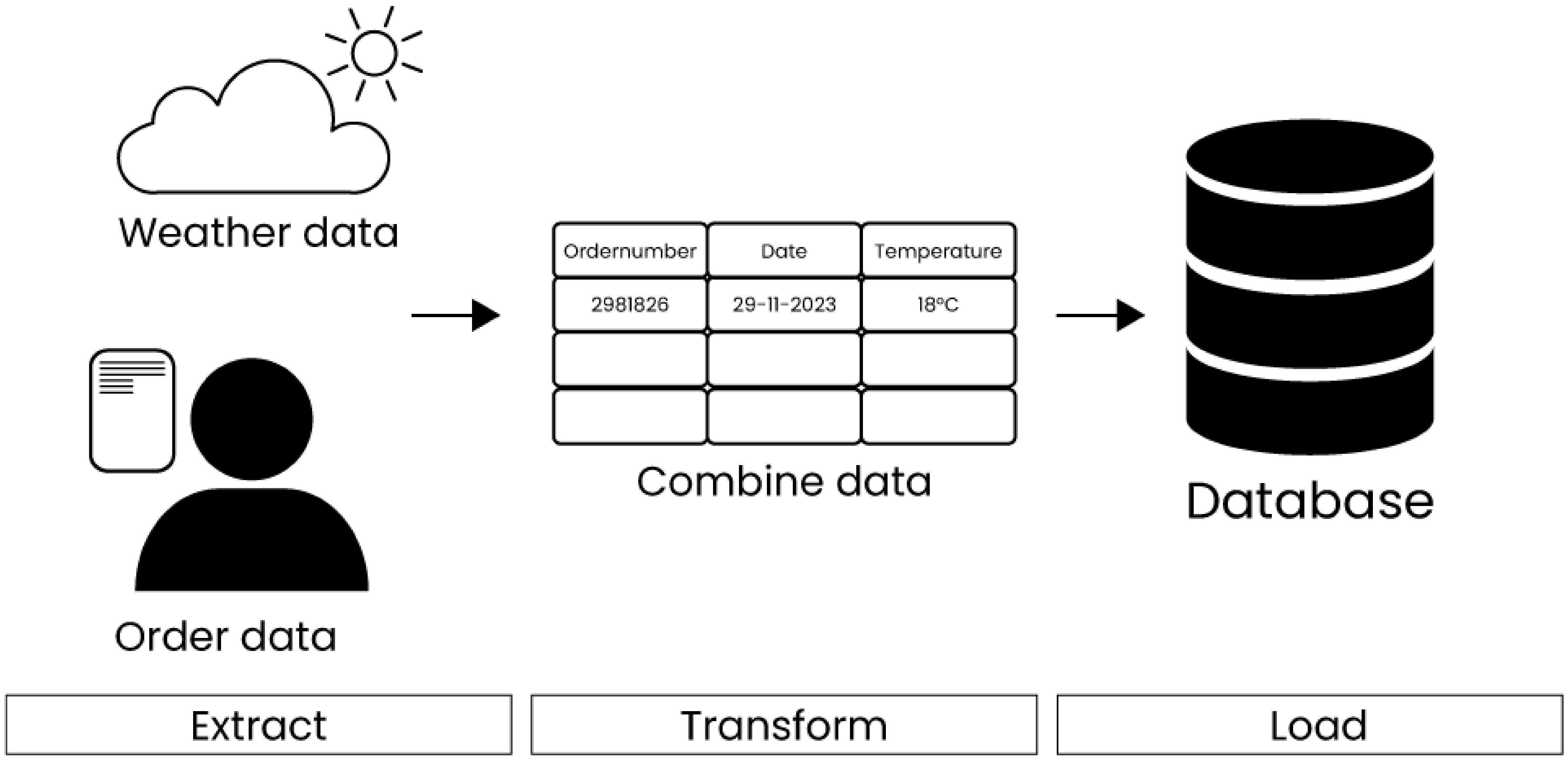
- Accuracy
- Completeness
- Consistency
- Timeliness

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!

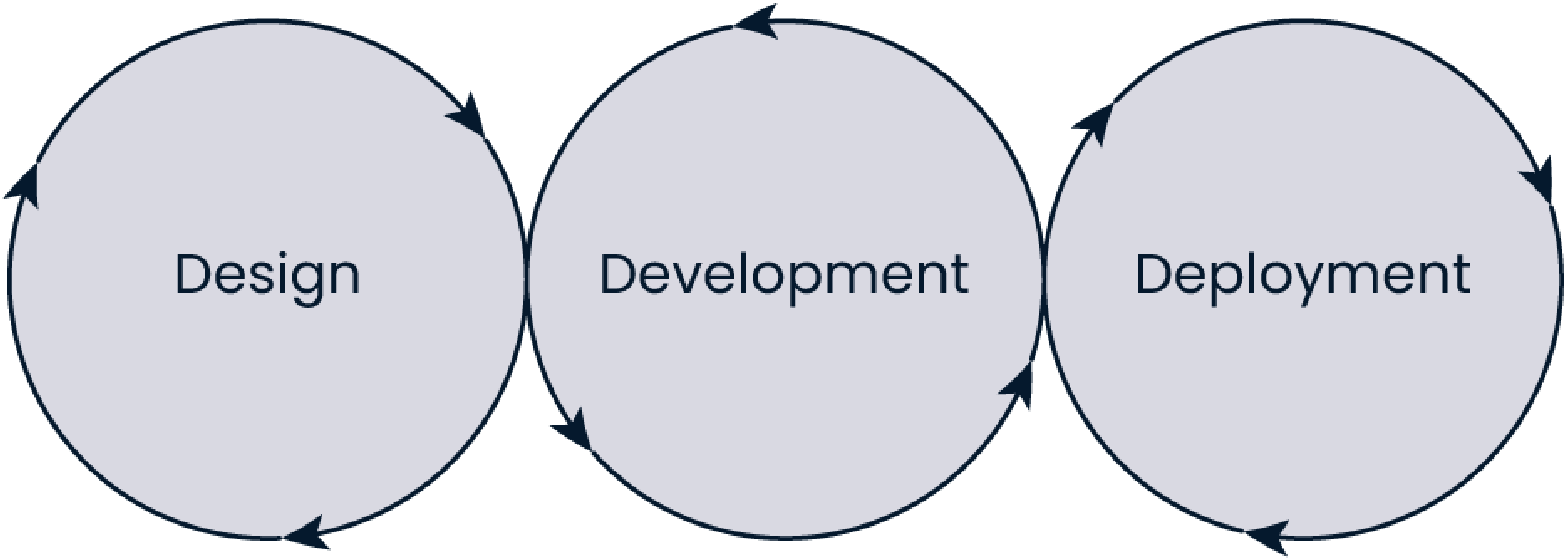
MLOPS CONCEPTS

Feature engineering

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
- We can use raw data, but also create our own

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

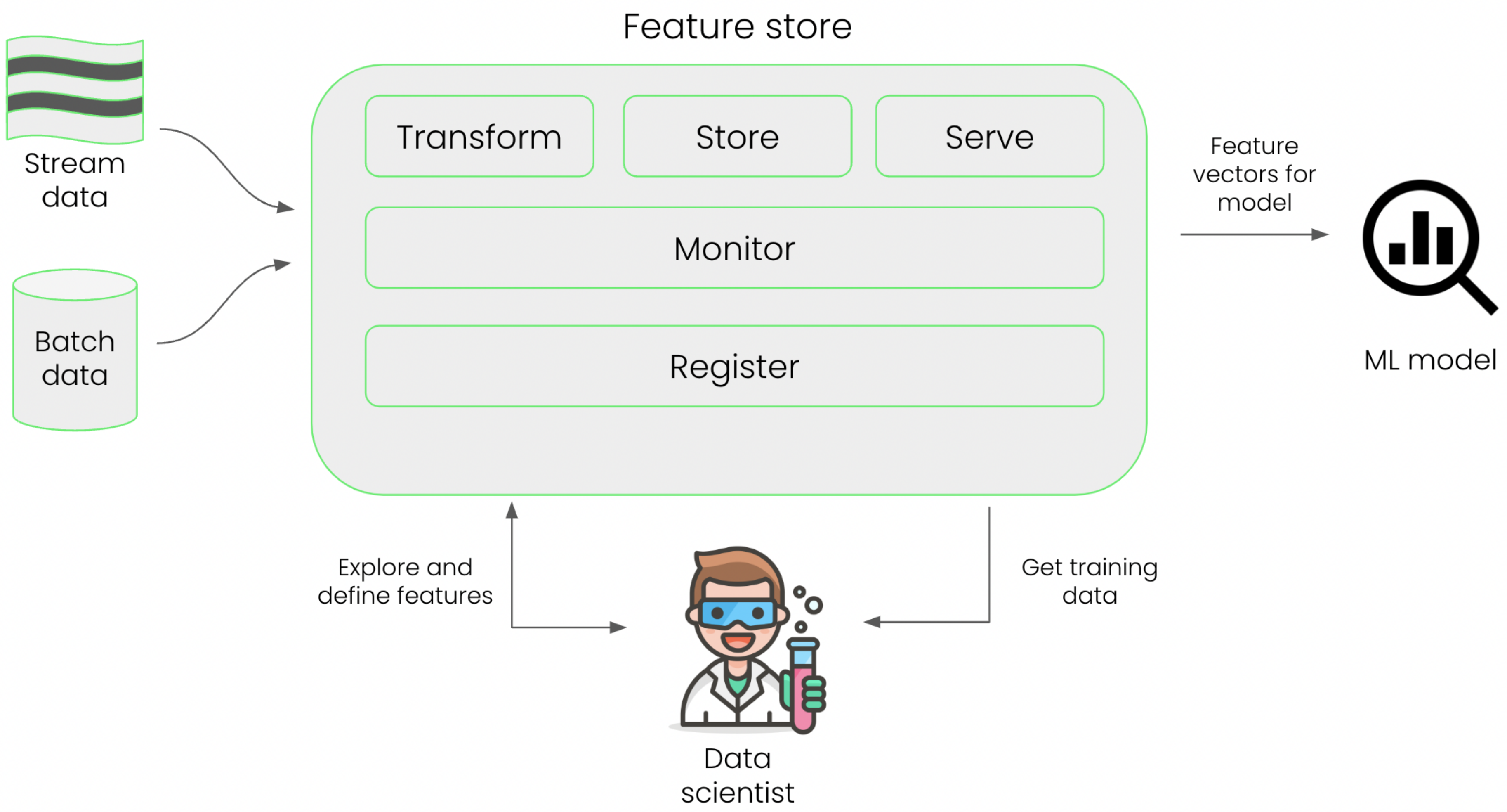
- Goal is to enhance model performance
- Tools and techniques help to process, select, and maintain features:
  - Feature selection
  - Feature store
  - Data version control

Feature selection

- Domain-specific knowledge
- Correlation
- Feature importances
- Other methods: univariate selection, Principal Component Analysis (PCA), Recursive Feature Elimination (RFE)

[1] https://www.datacamp.com/tutorial/tutorial-datails-on-correlation

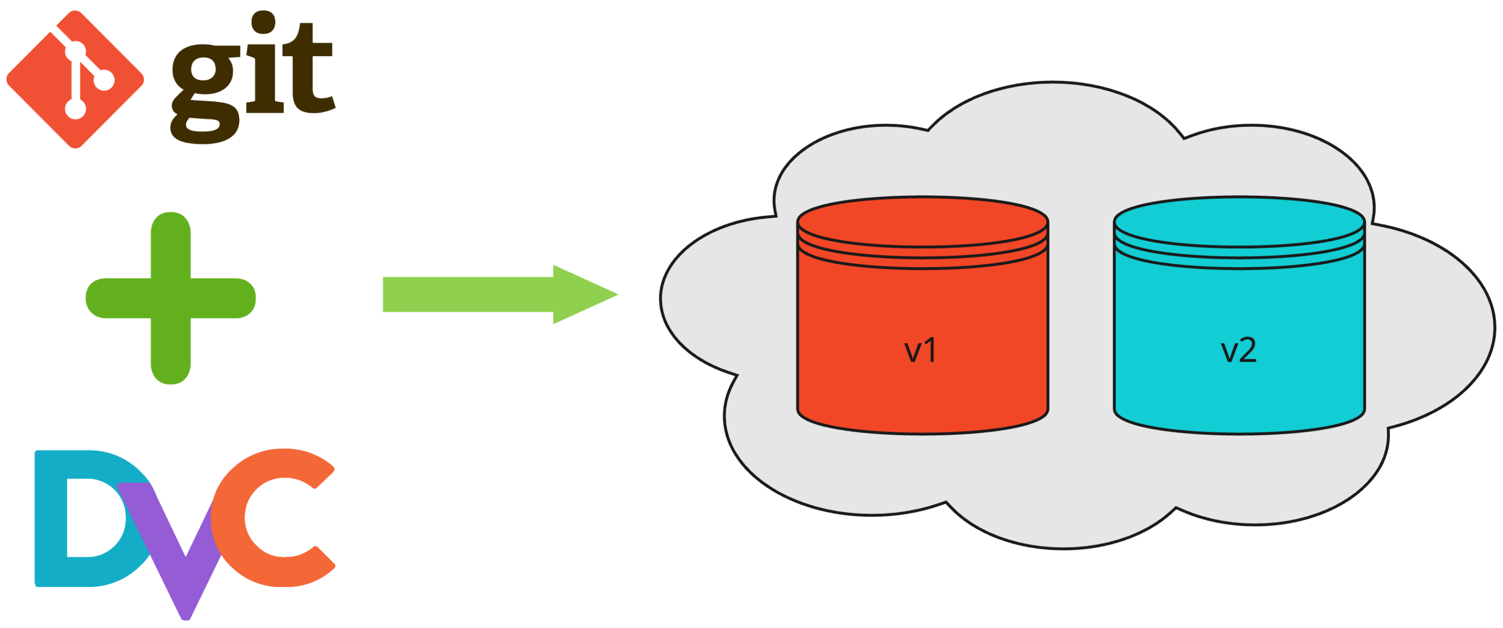
The feature store



Only relevant for large teams working on multiple projects that use the same features

Data version control

- Tracking dataset changes
- Maintaining consistency throughout the development lifecycle



[1] https://www.datacamp.com/courses/cicd-for-machine-learning

Let's practice!

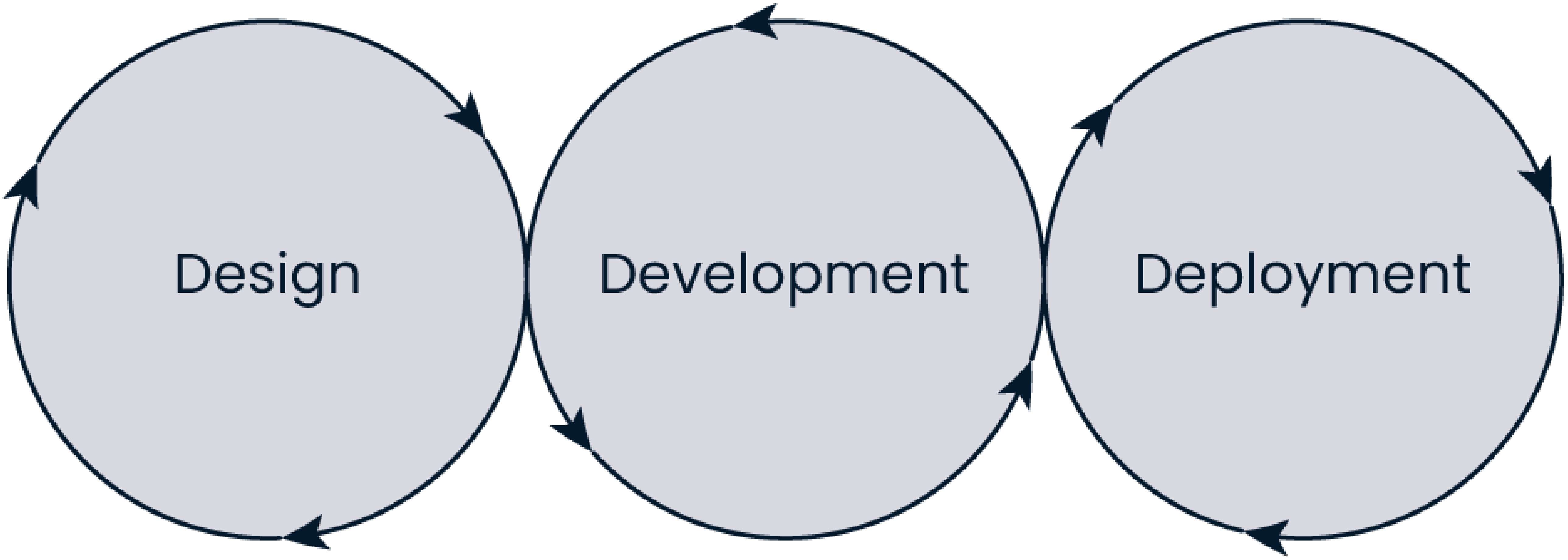
MLOPS CONCEPTS

Experiment tracking

MLOPS CONCEPTS

Folkert Stijnman
ML Engineer

The machine learning experiment



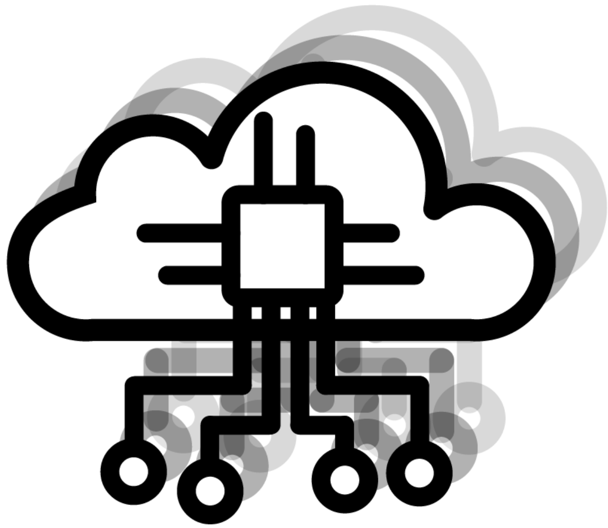
- Added value
- Business requirements
- Key metrics
- Data processing

- Feature engineering
- Experiment tracking
- Model training & evaluation

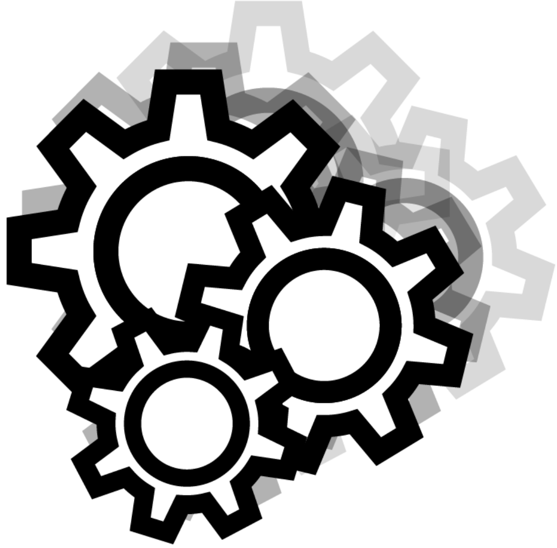
- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring & retraining

Why is experiment tracking important?

In each experiment, the following factors can be configured:

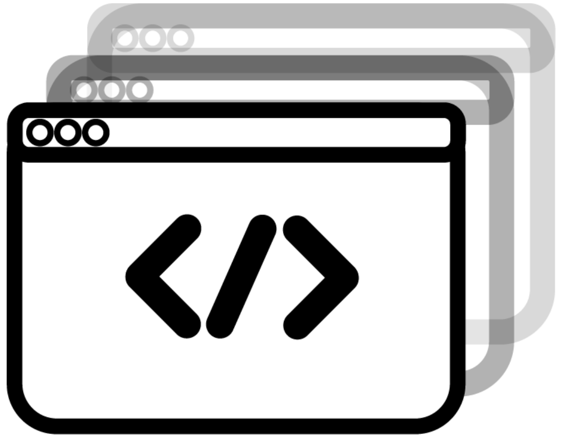


Machine learning models

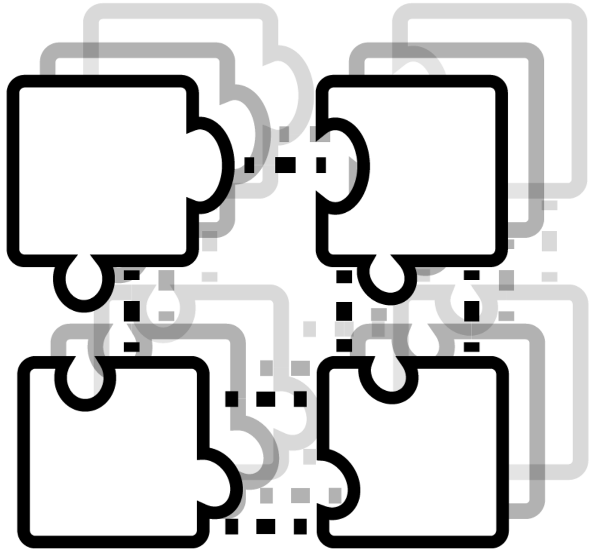


Model hyperparameters

Versions of data



Execution scripts



Environment configurations

Using experiment tracking in the ML lifecycle

Experiment tracking can help to:

- Compare results
- Reproduce past experiments
- Collaborate with developers and stakeholders
- Report on results to stakeholders

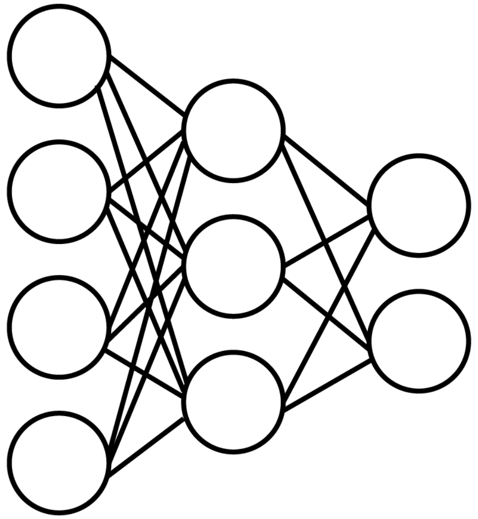
How to track experiments?

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	Requires getting familiar with the tool

A machine learning experiment

Experiment 1

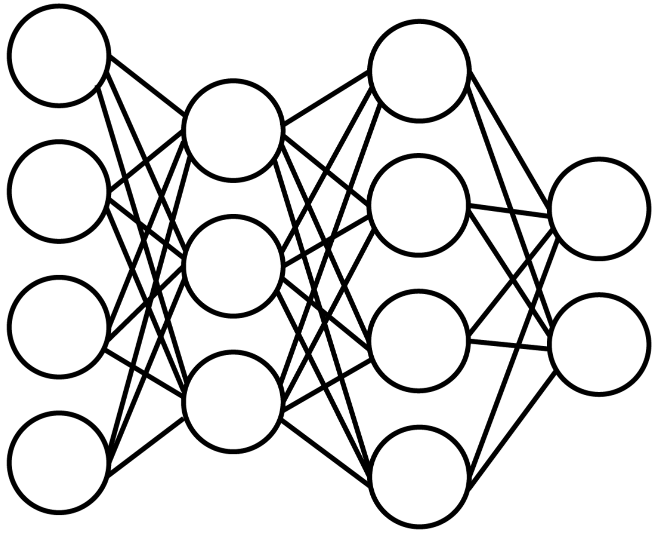
1.000 images of cats & dogs



A neural network with 1 hidden layer

Experiment 2

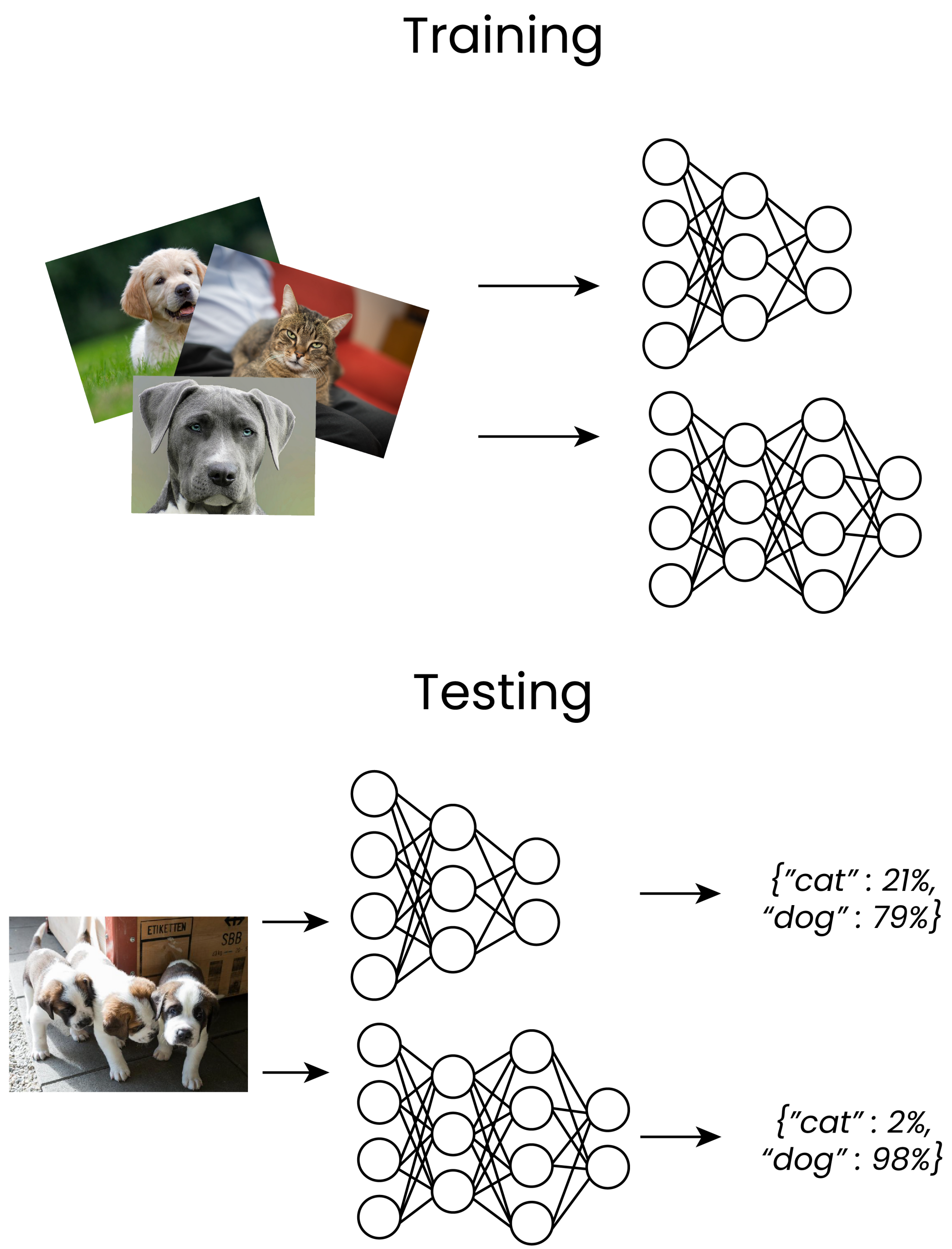
2.000 images of cats & dogs (including puppies & kittens)



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