Help! They dumped a dataset on our doorstep…

… and want us to predict something

By Richard Berendsen

devCON

is powered by

luminis
Conversing worlds

My world is:

Nou moe?! - Guust Flater

All our worlds › www.luminis.eu

About me

Richard Berendsen

- Search and data engineer at Luminis
- BSc & MSc Artificial Intelligence
- PhD Information retrieval (aka search)

@rwberendsen

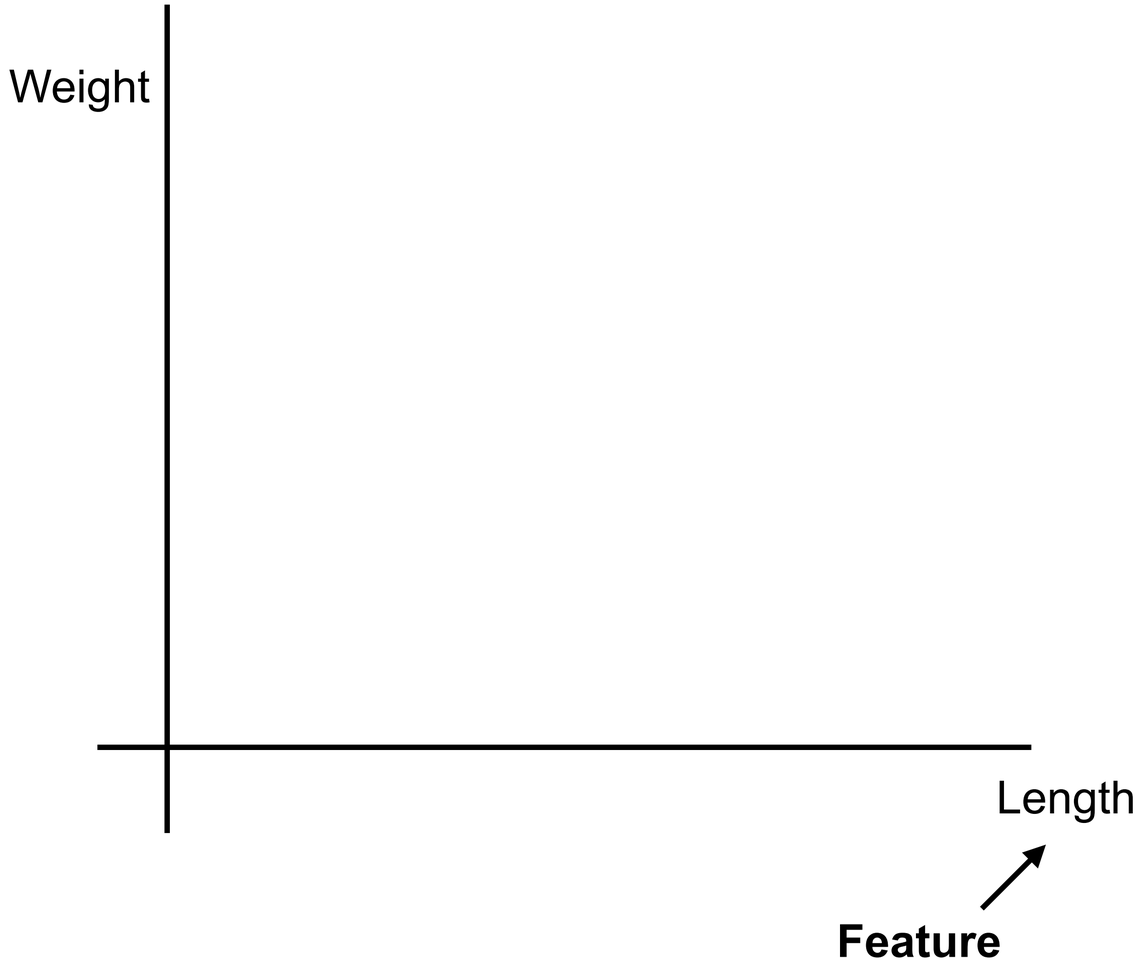
luminis
Conversing worlds

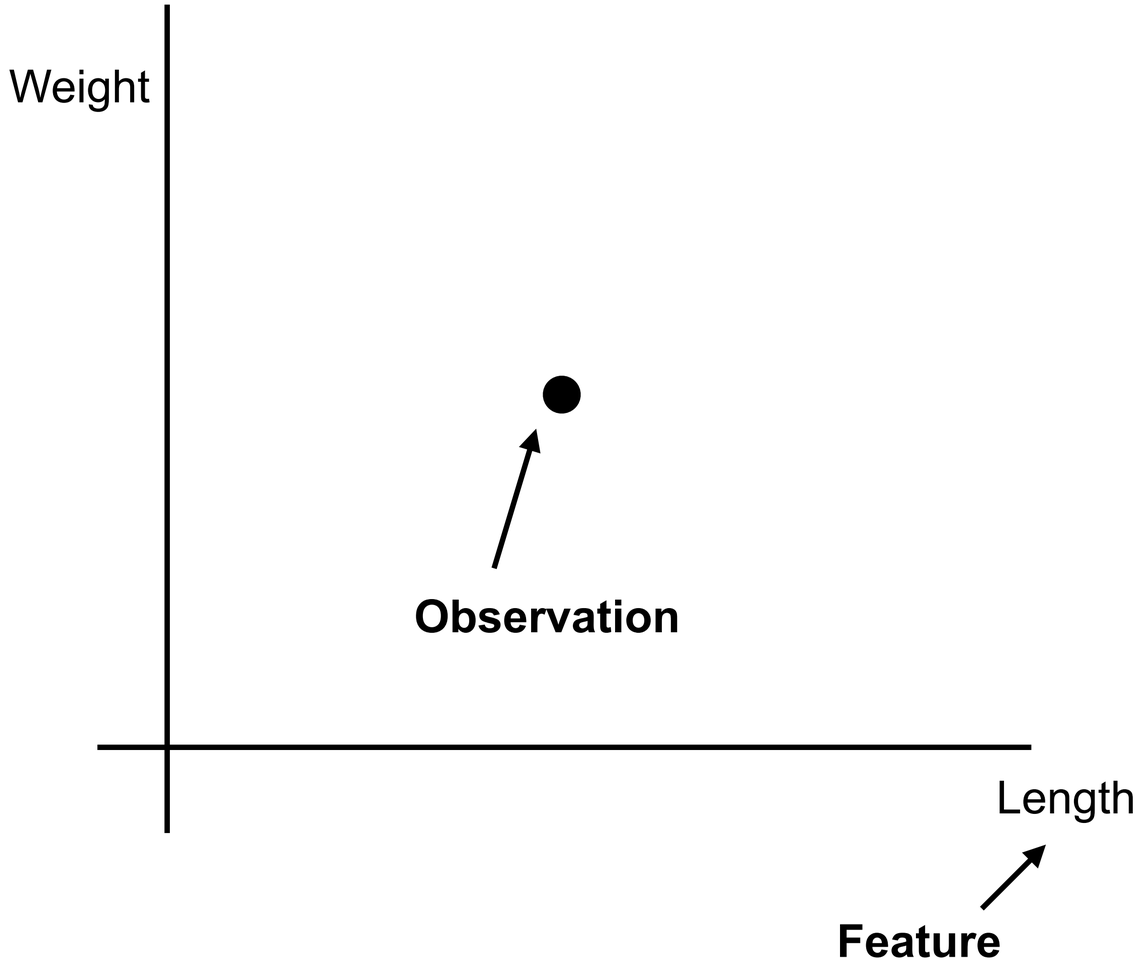
Emphasis

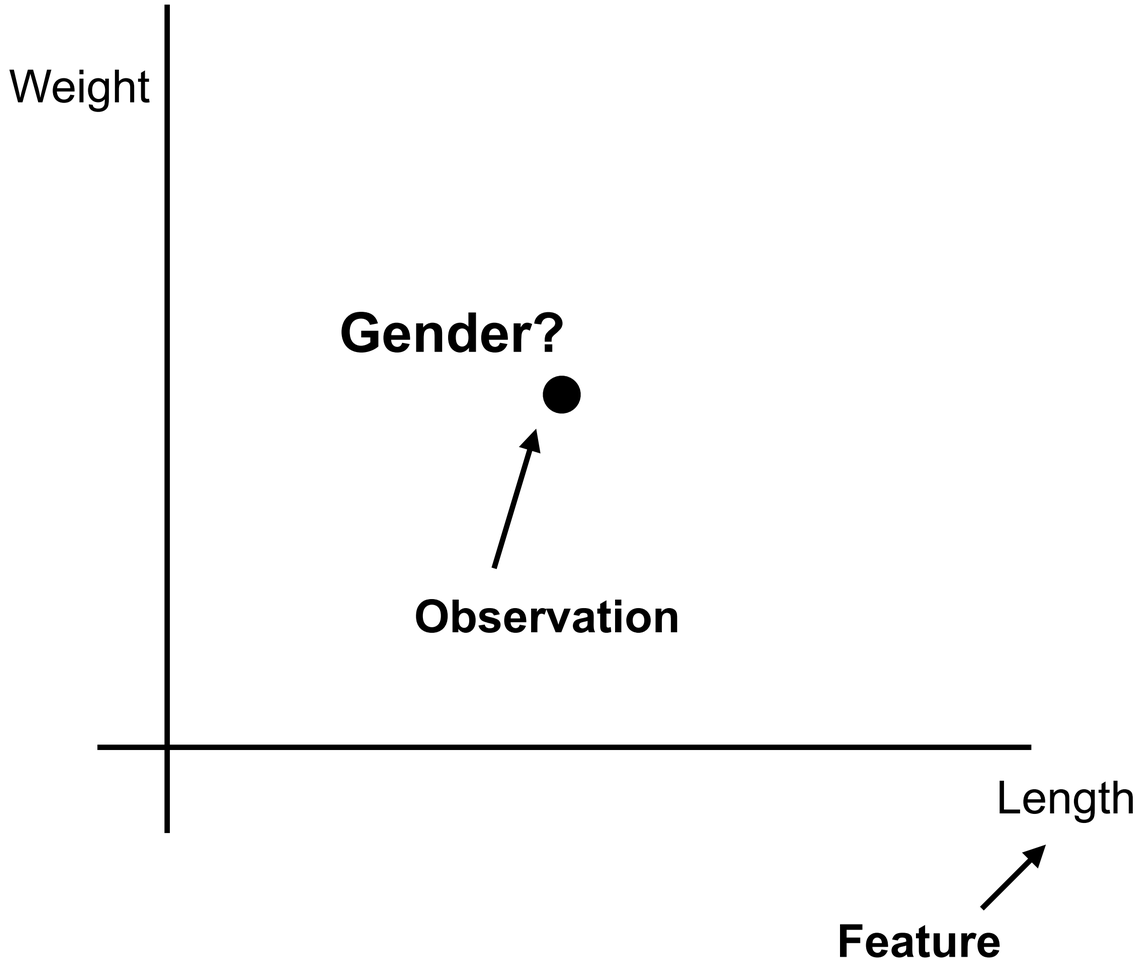
- Data science process
- Choices
- Questions
- Step by step demo

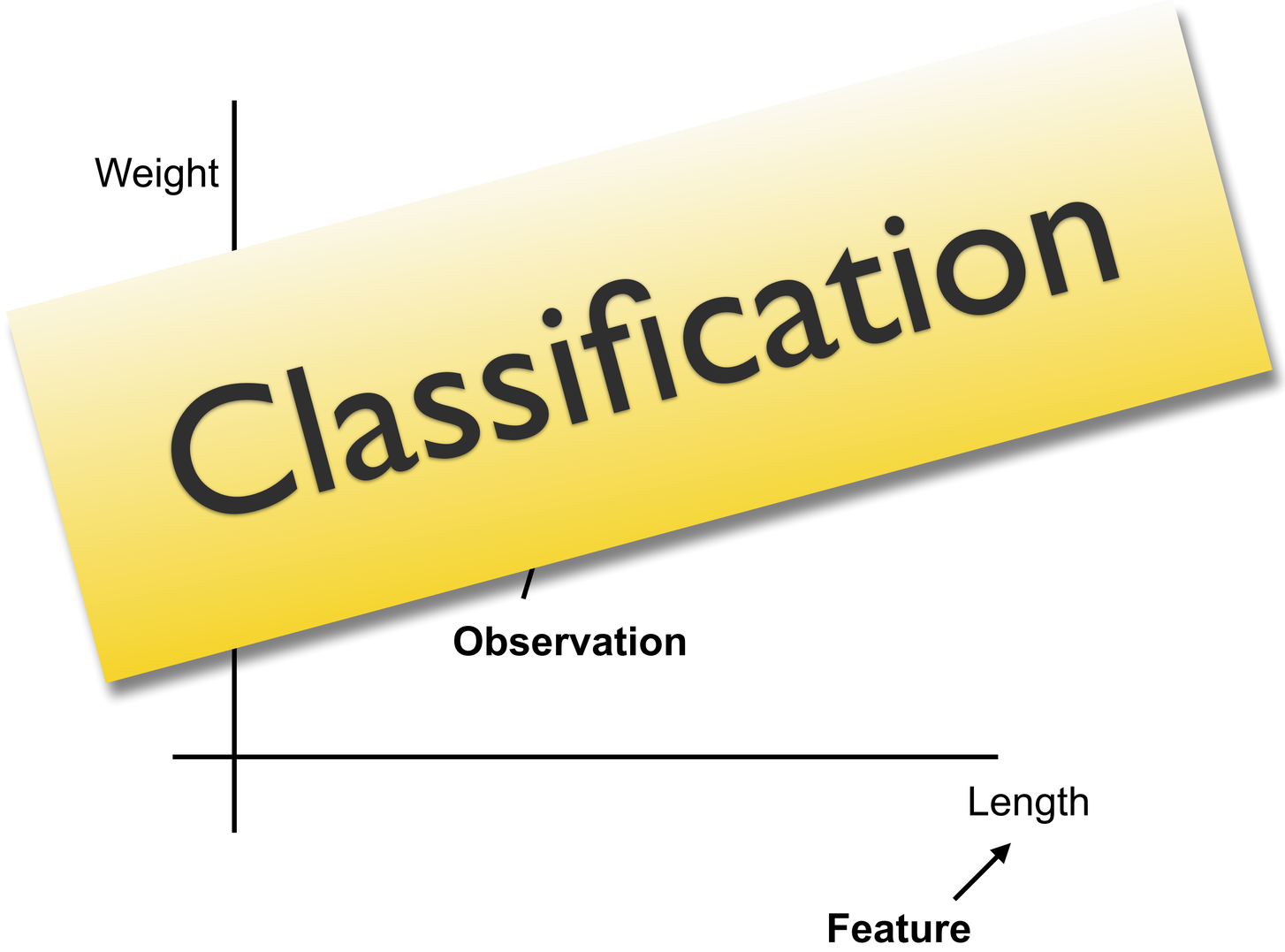
But first…

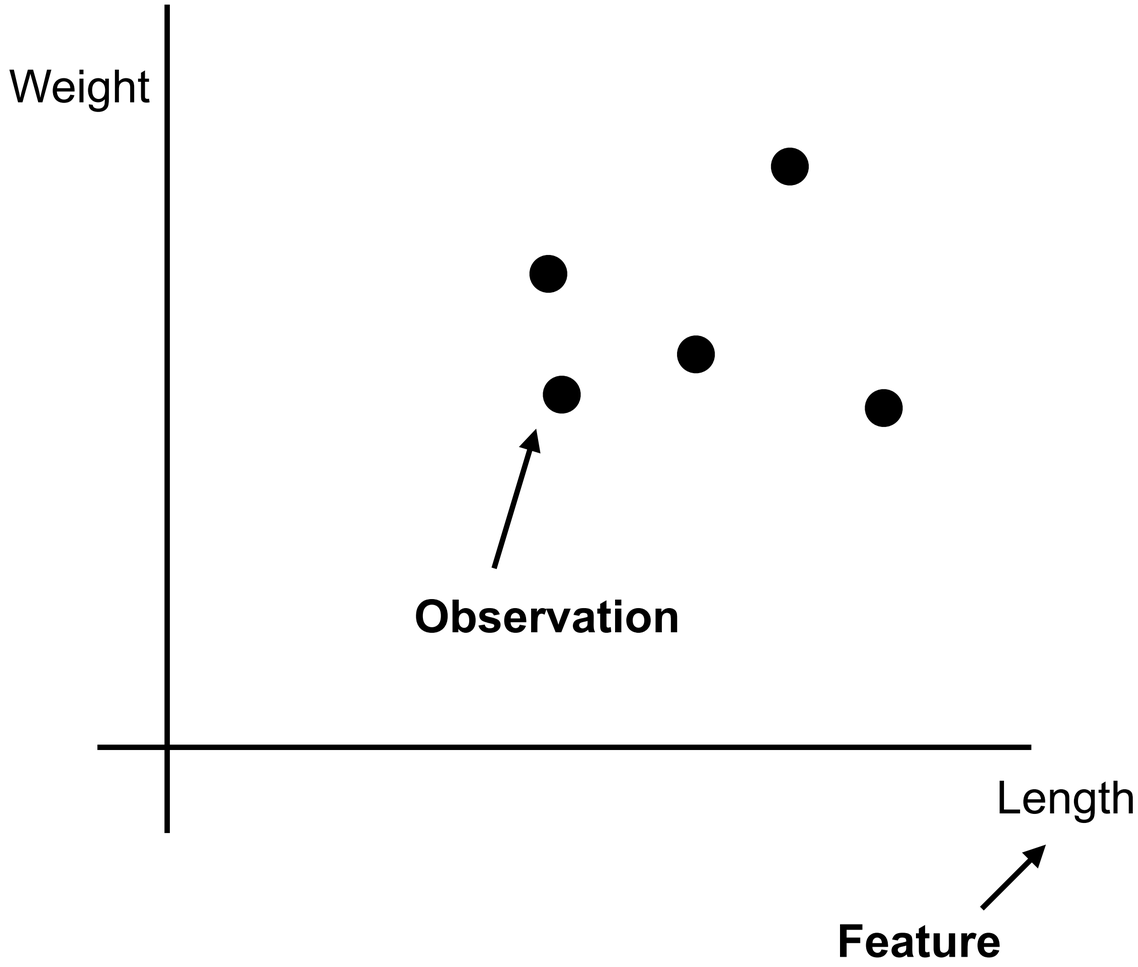
- A little data science intro
- A quick look at the stack

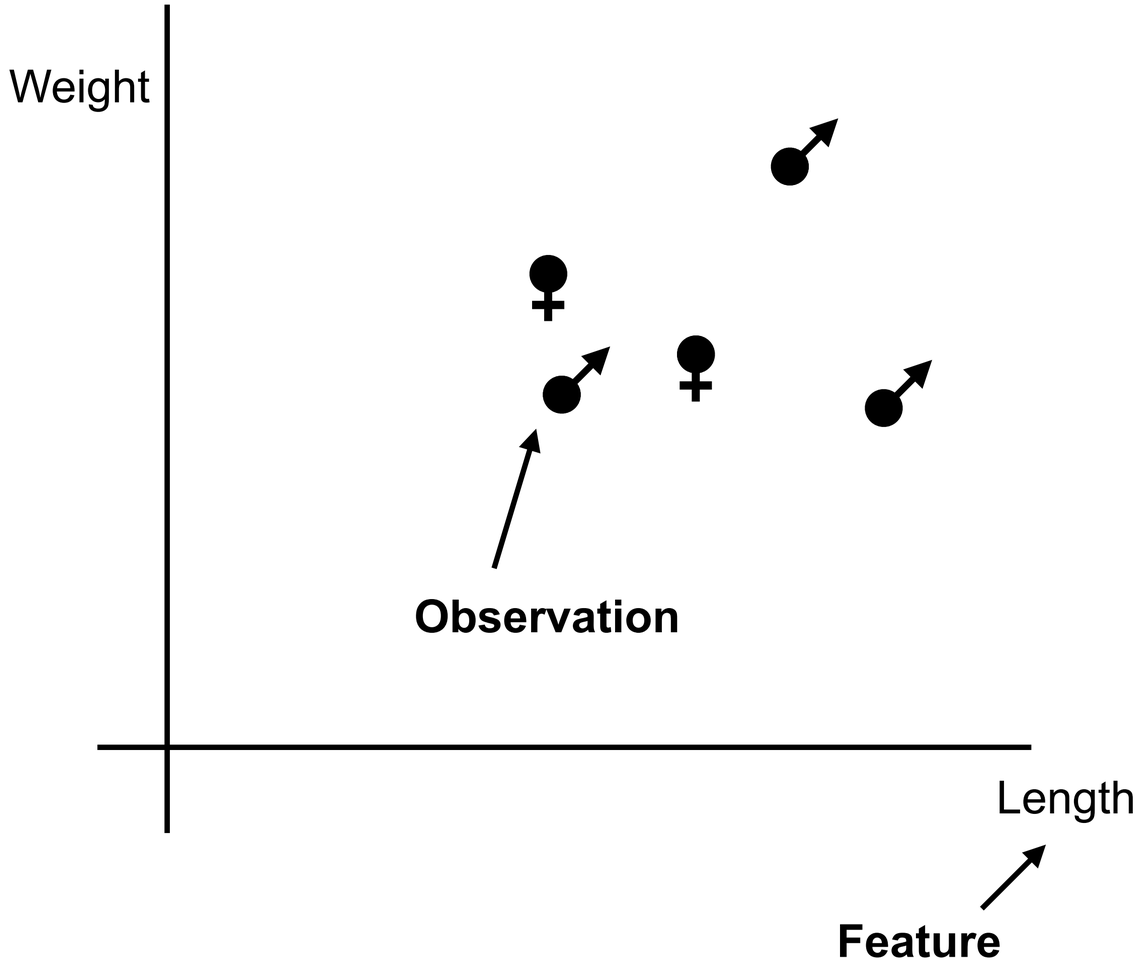

Weight
Length
Feature

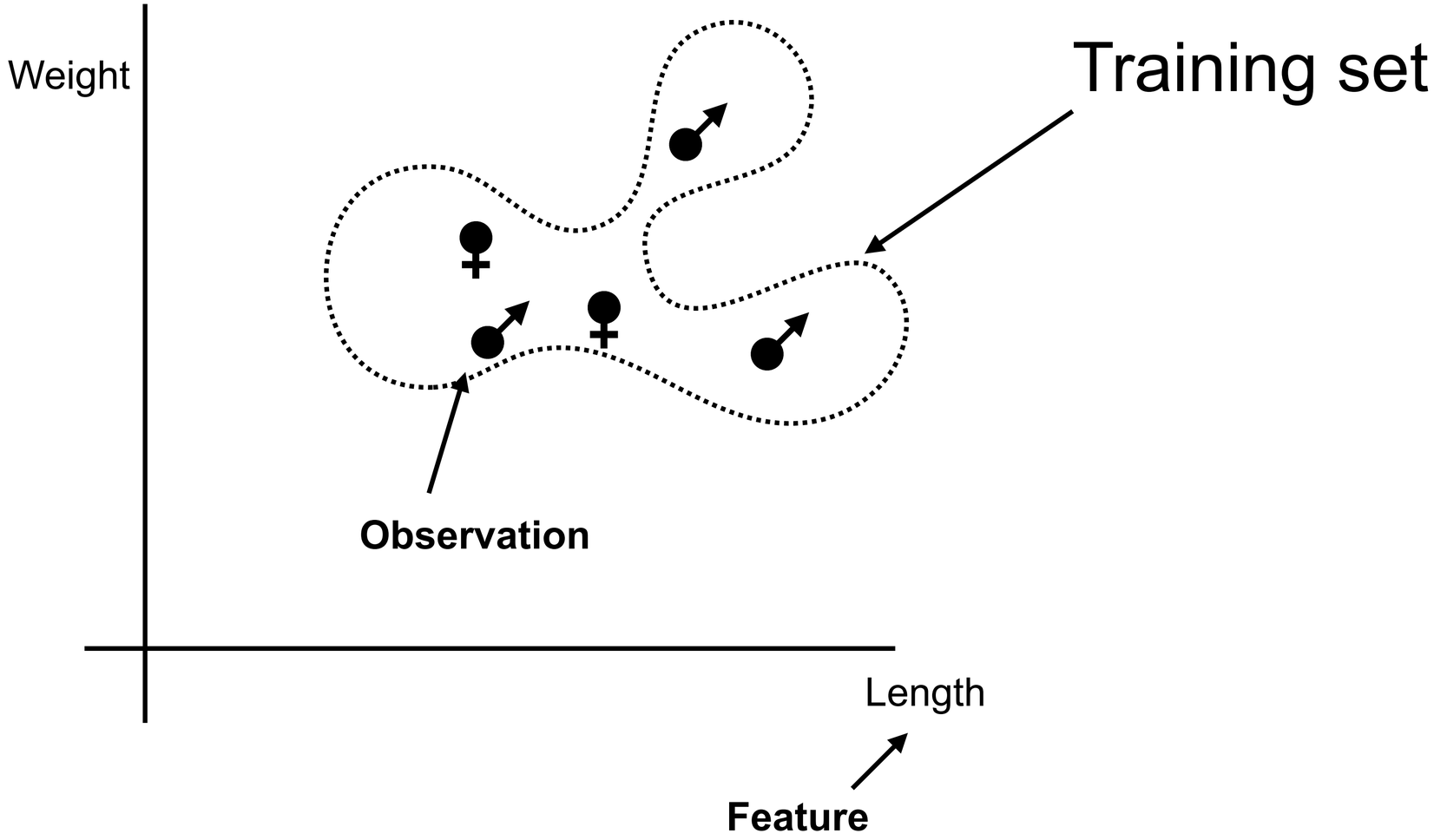

Weight
Observation
Length
Feature

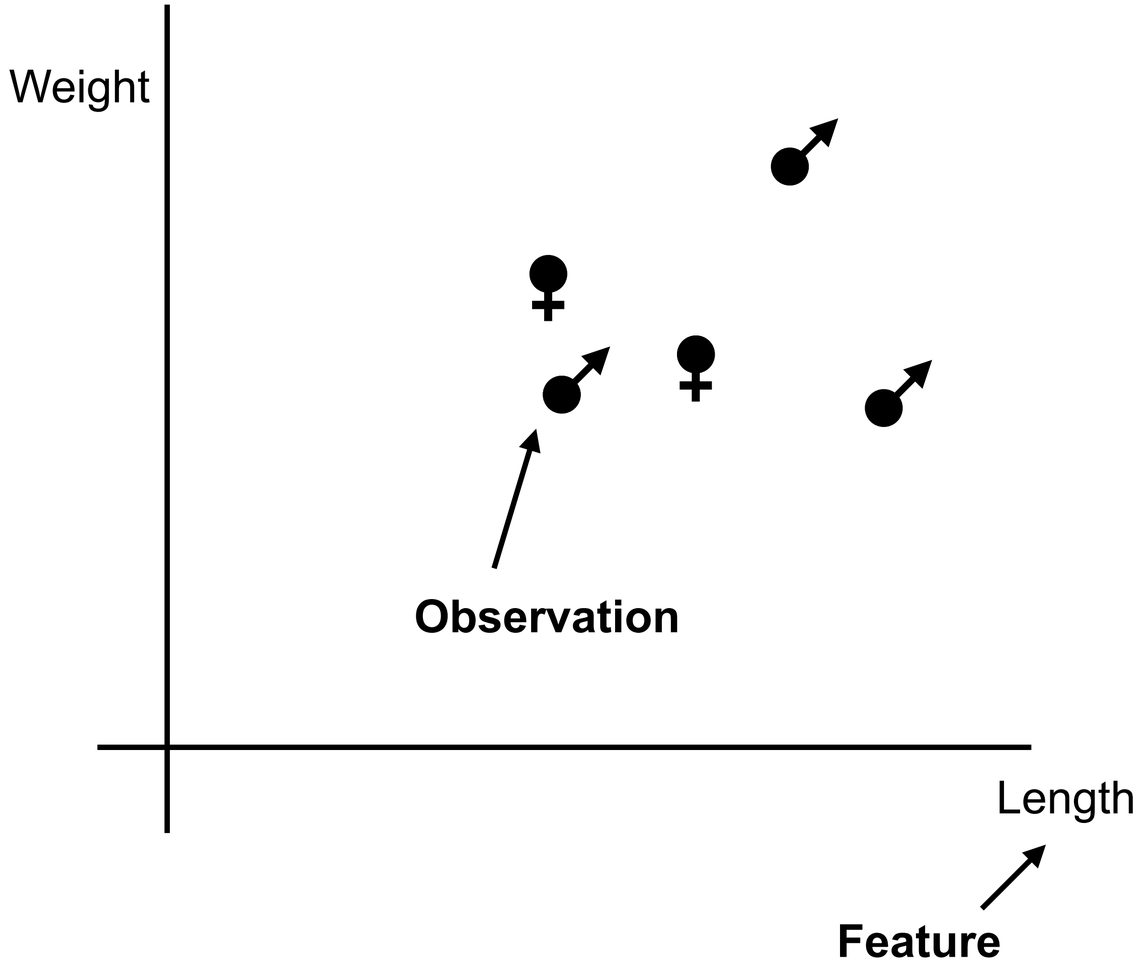

Weight
Gender?
Observation
Length
Feature

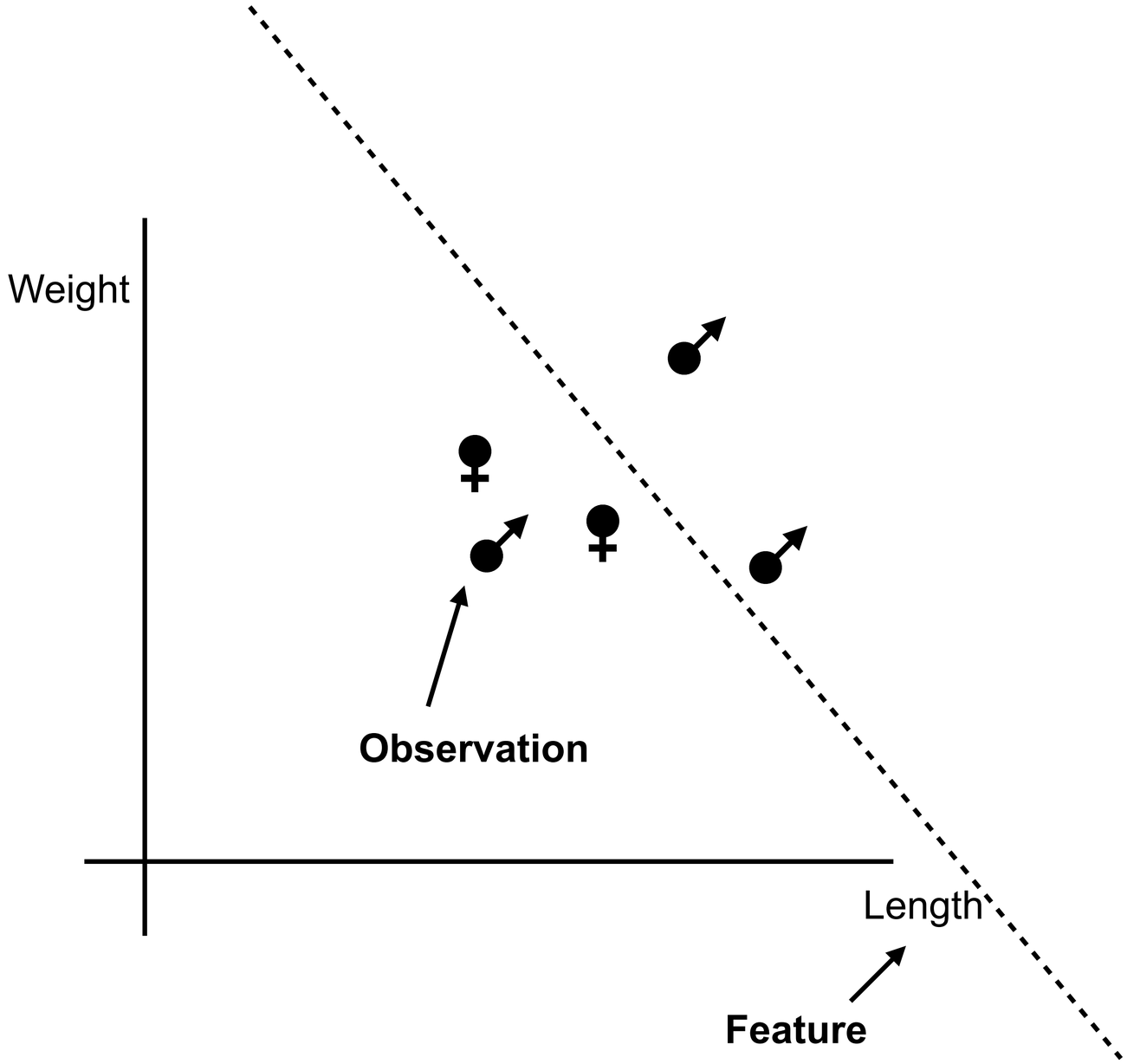

Classification
Weight
Observation
Length
Feature

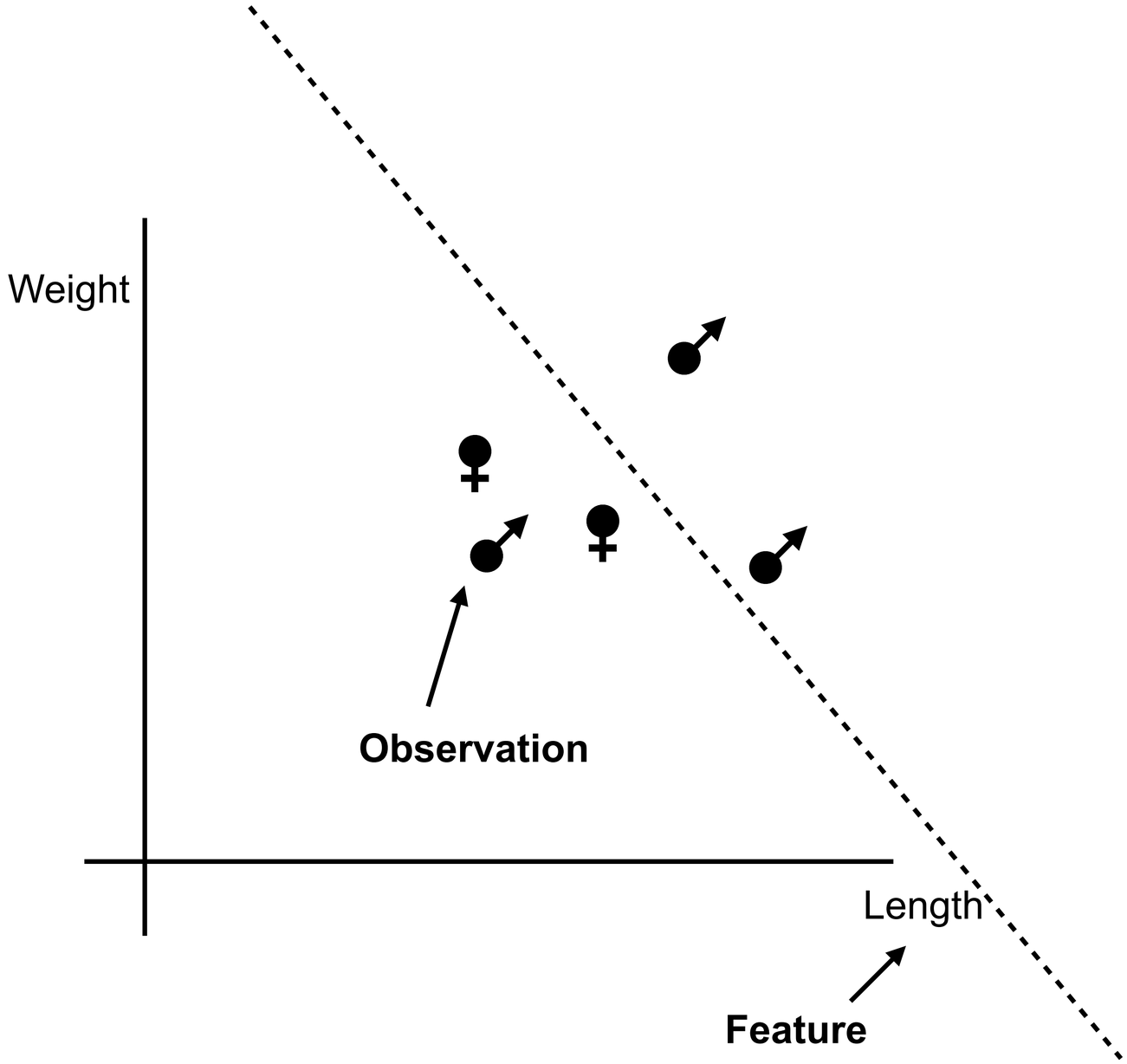

Weight
Observation
Length
Feature

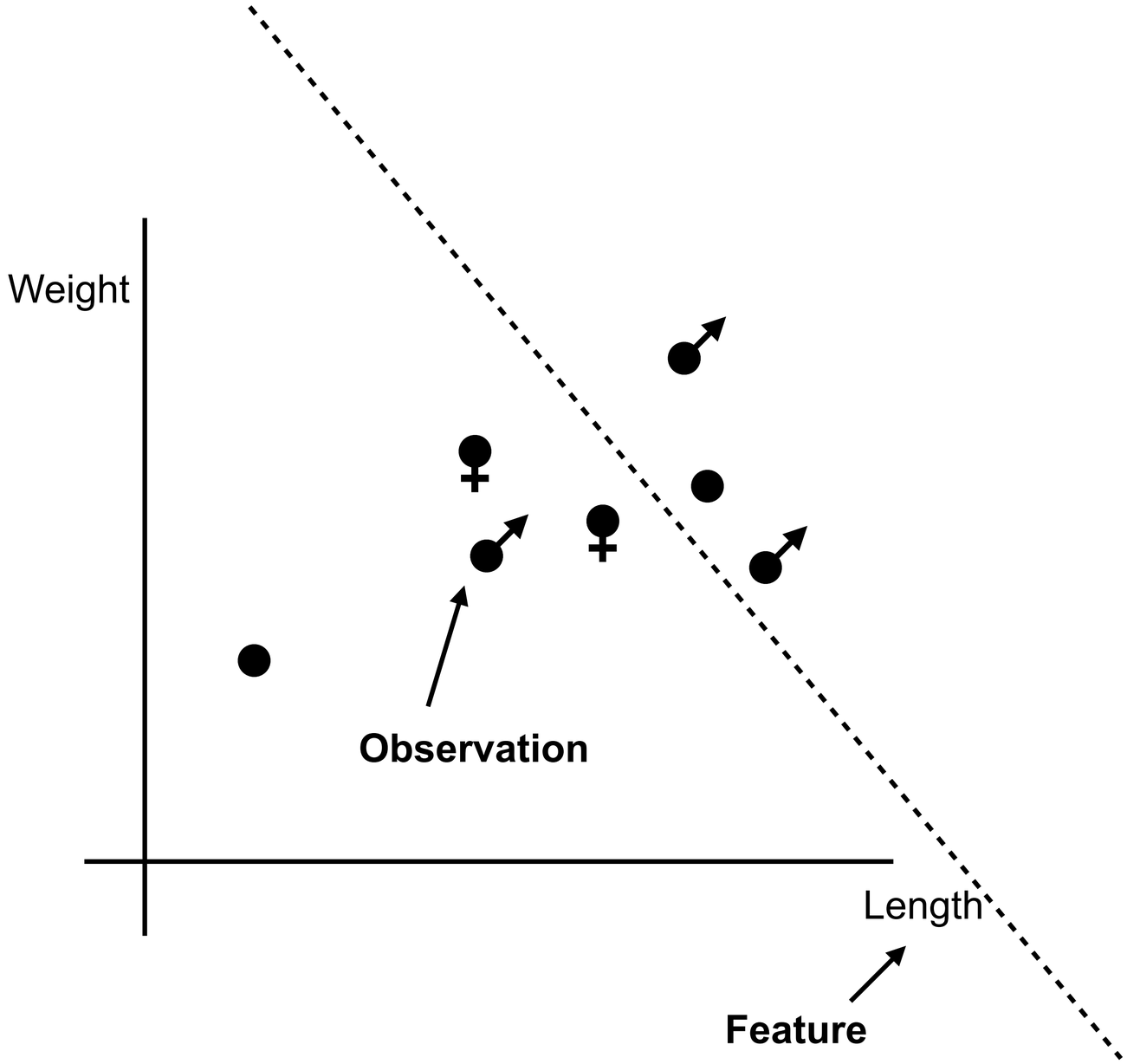

Weight
Observation
Length
Feature

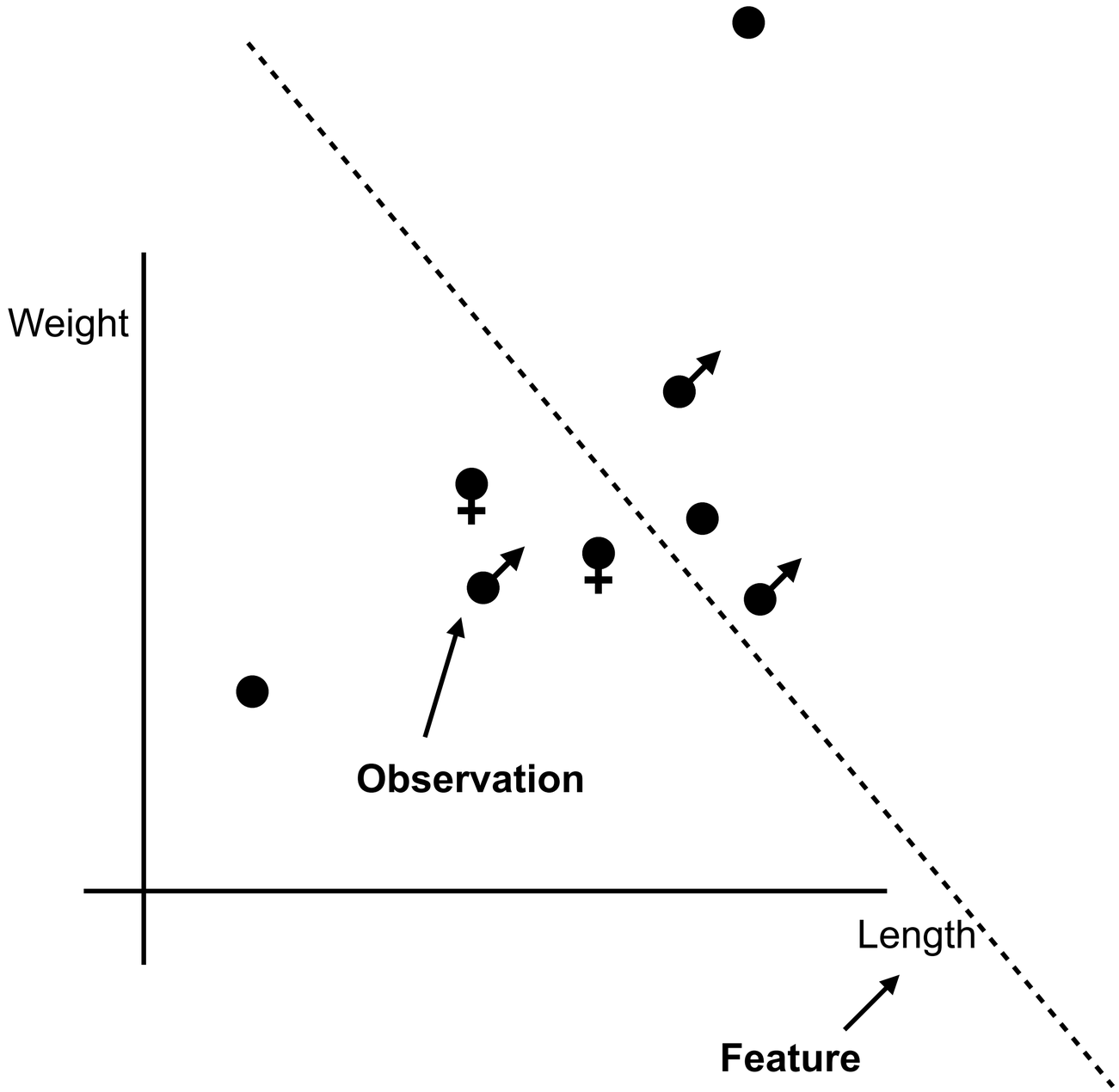

Weight
Training set
Observation
Length
Feature

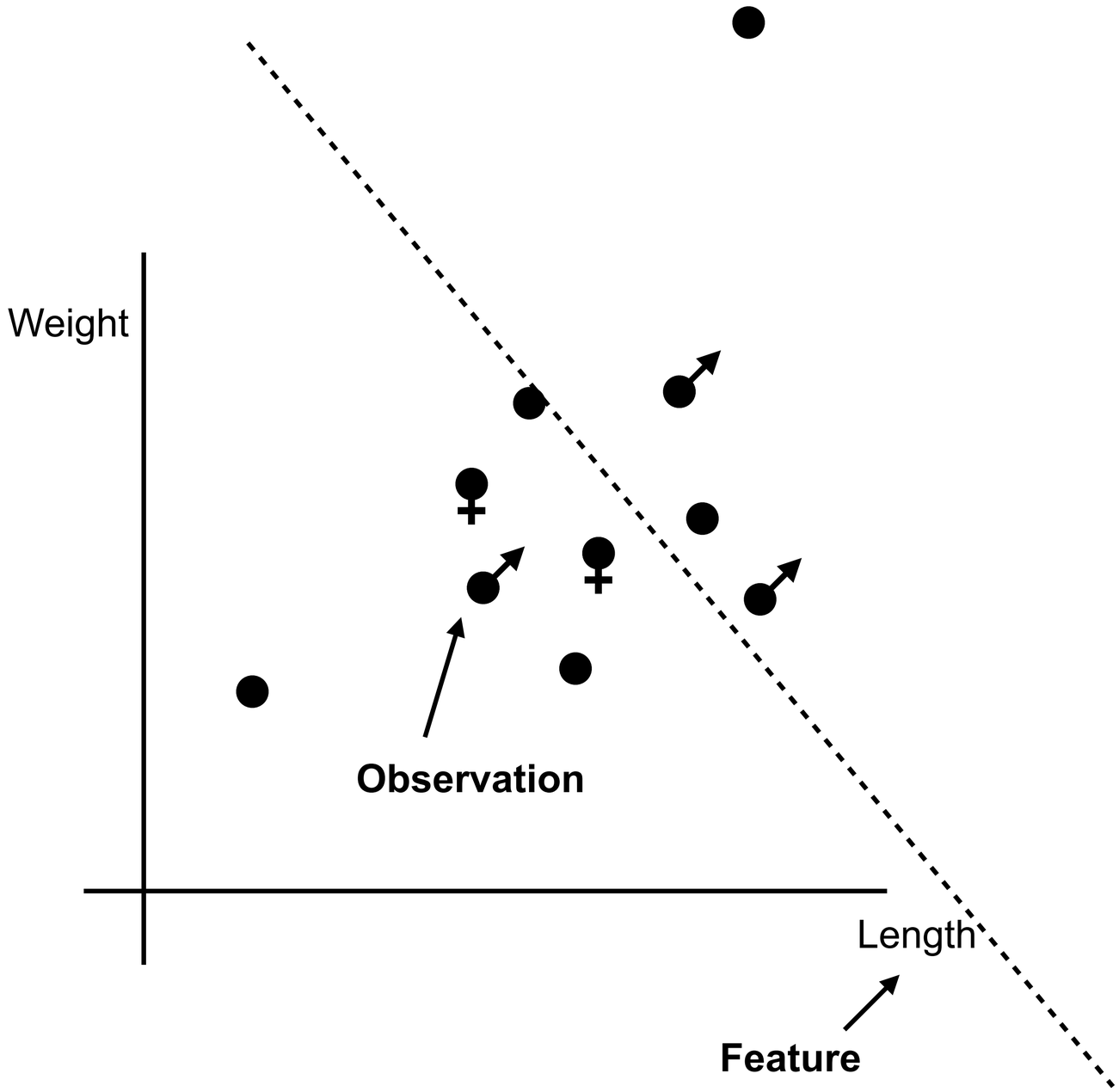

Weight
Observation
Length
Feature

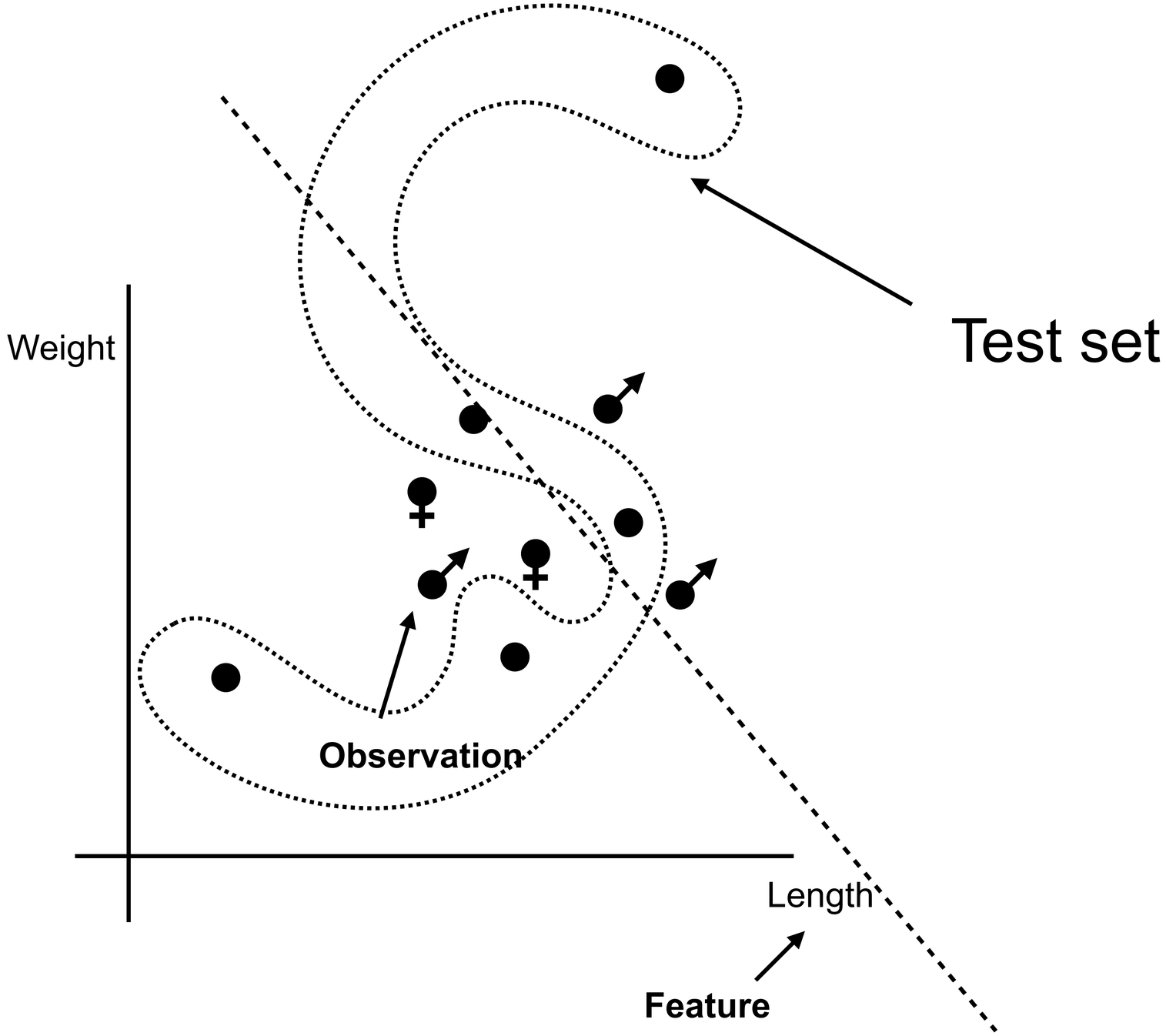

Weight
Observation
Length
Feature

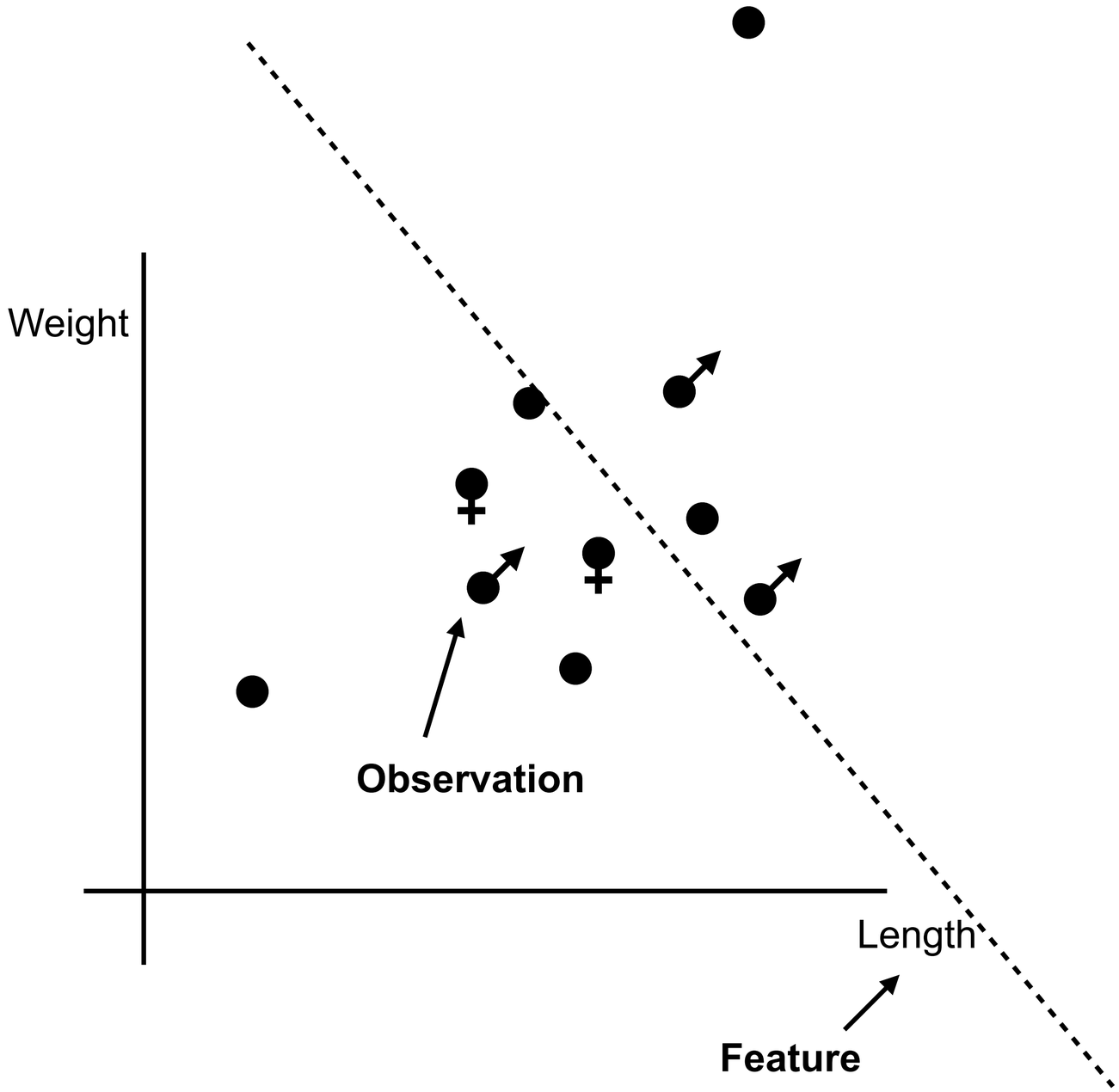

Weight
Observation
Length
Feature


Weight
Observation
Length
Feature


Weight
Observation
Length
Feature


Weight
Observation
Length
Feature

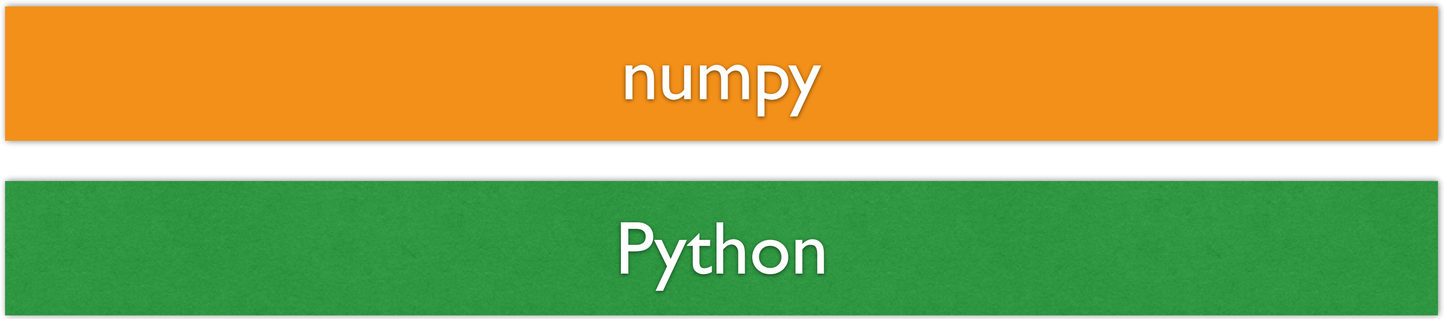

Test set
Weight
Observation
Length
Feature


Weight
Observation
Length
Feature

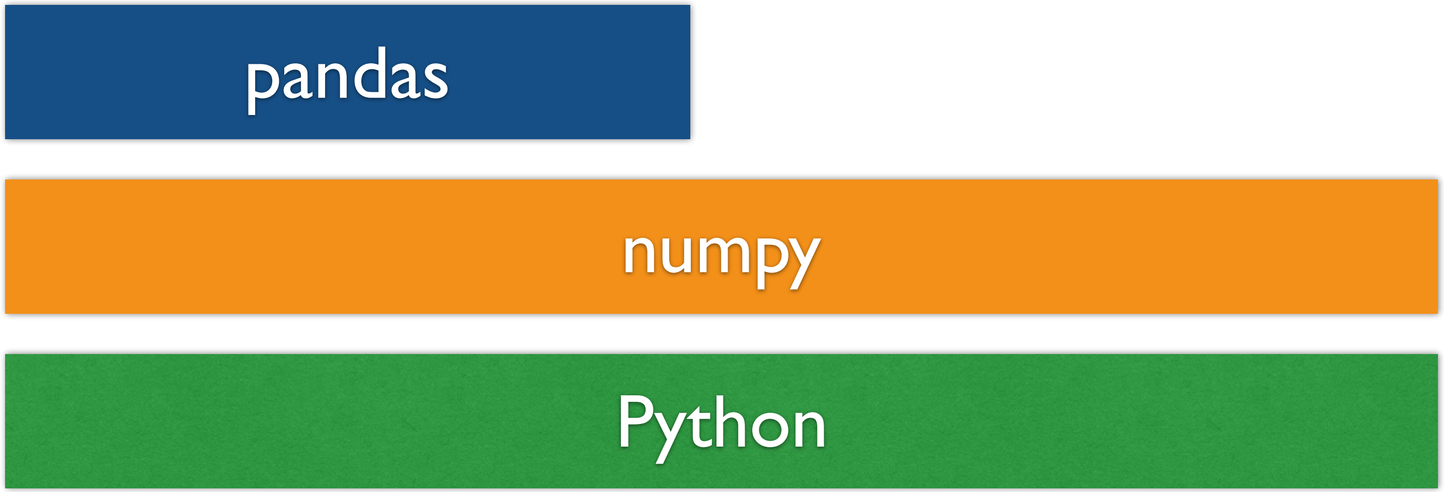
The SciPy Stack (parts we use)

Python

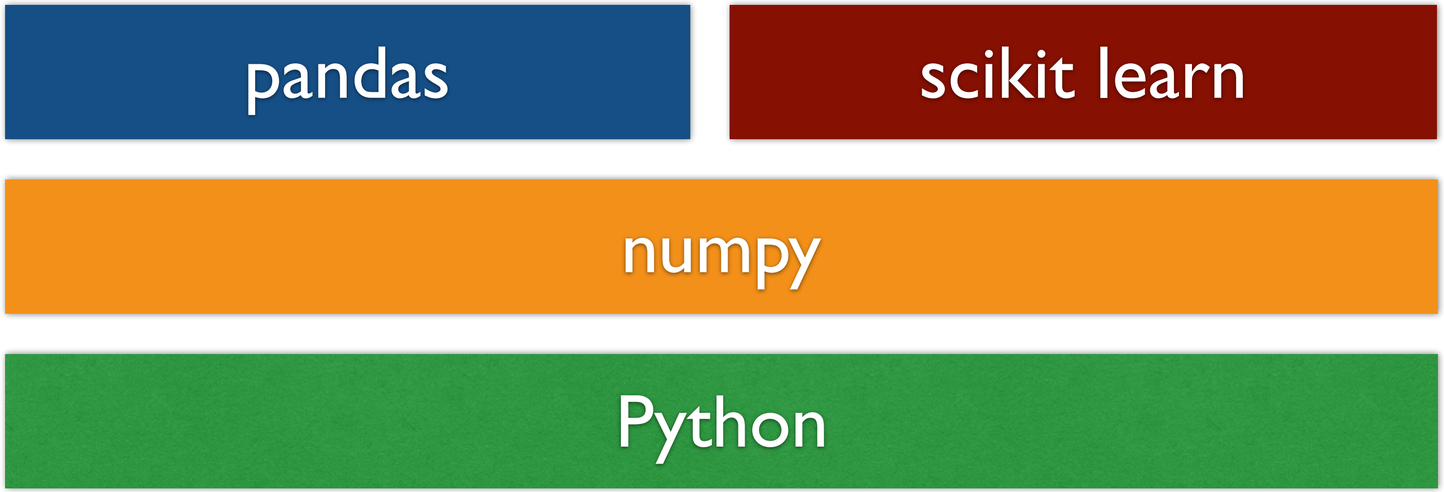
The SciPy Stack



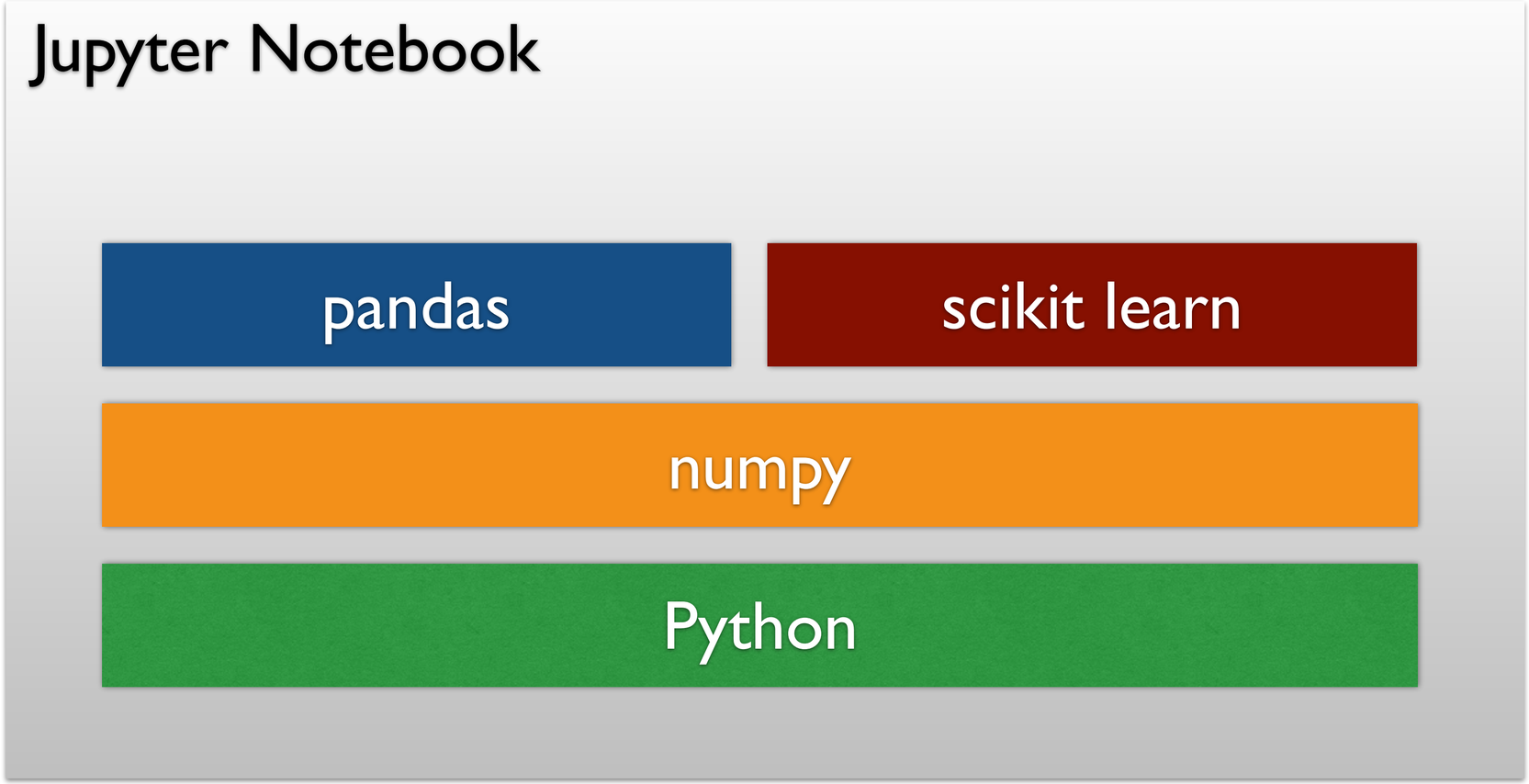
The SciPy Stack



The SciPy Stack



The SciPy Stack



Installation requirements

- What I did, anyway.
- Install the latest Anaconda Python 3 distribution: https://www.continuum.io
- Install pandas: `conda install pandas`
- Install Jupyter: `conda install jupyter`
- You are now ready to go, just `jupyter notebook`

The data science process

Data science process

- Business model
- Data cleaning
- Exploration
- Experimental setup
- Feature engineering
- Model fitting (with decision trees)

Business model

Business model

- Understand it!
- What to learn / predict?
- Use case?
- Data?

devCON

Our aim

- Predict who will order again in the next period
- Predict how much they will spend

Data cleaning

Data frame

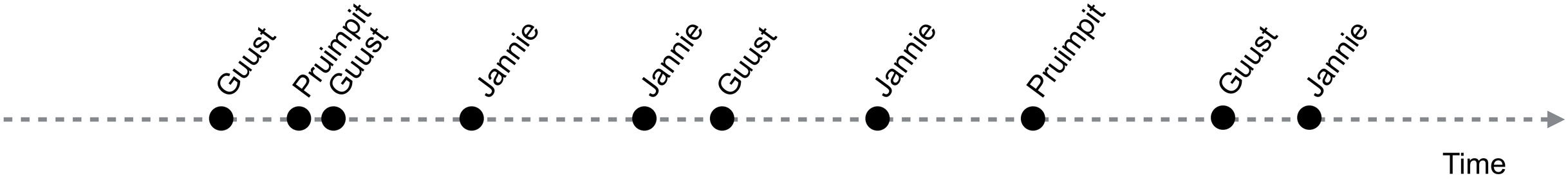
Index	Email	…	Paid amount	Refunded amount
Order 1				
Order 2				
Order 3				
Order 4				

devCON

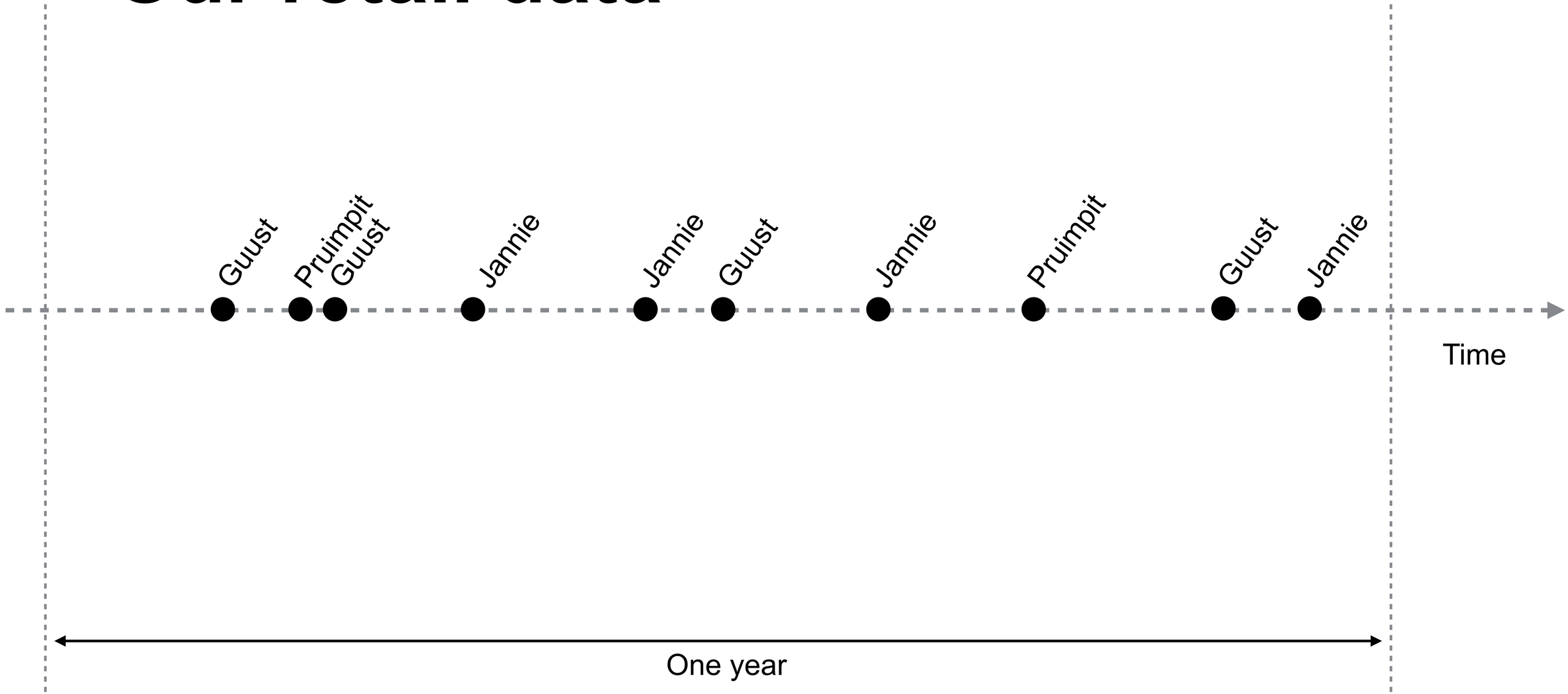
Exploration

Our retail data

Our retail data

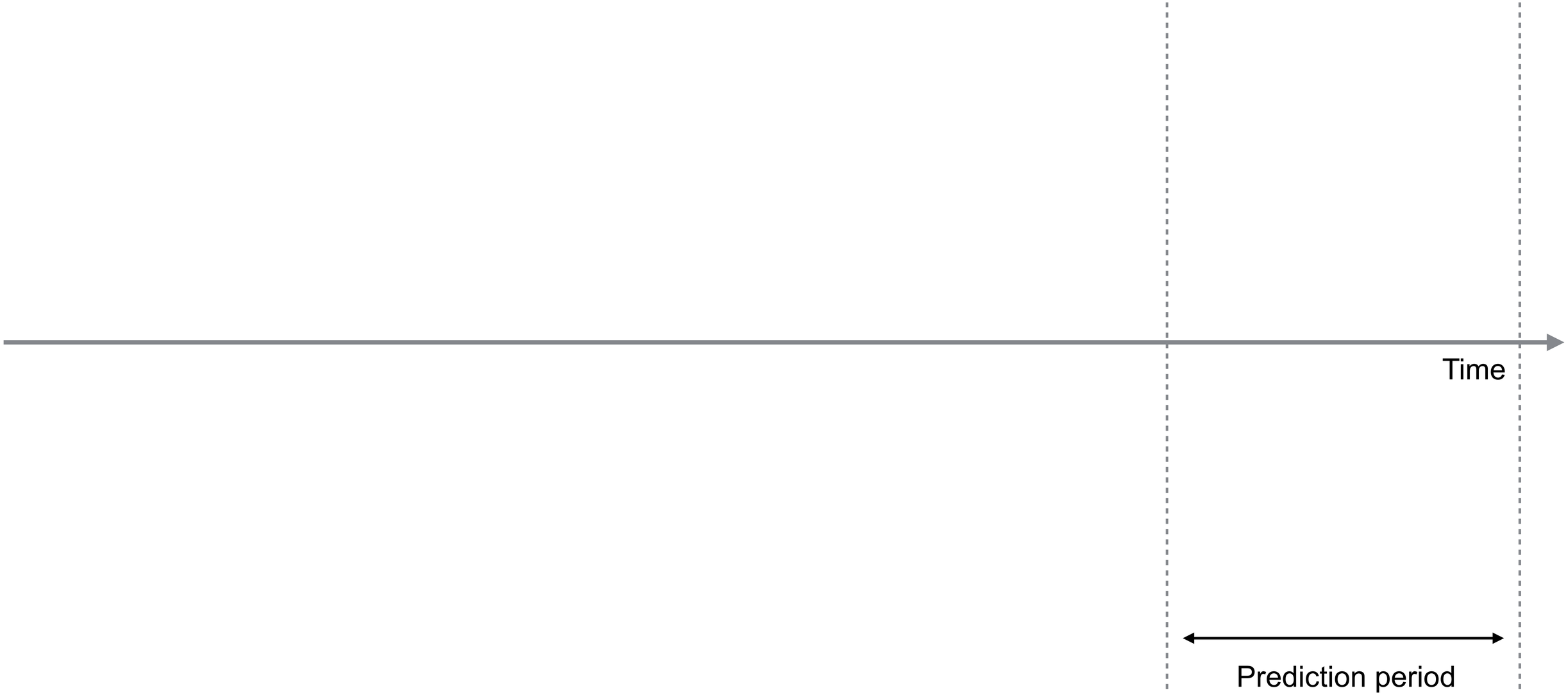


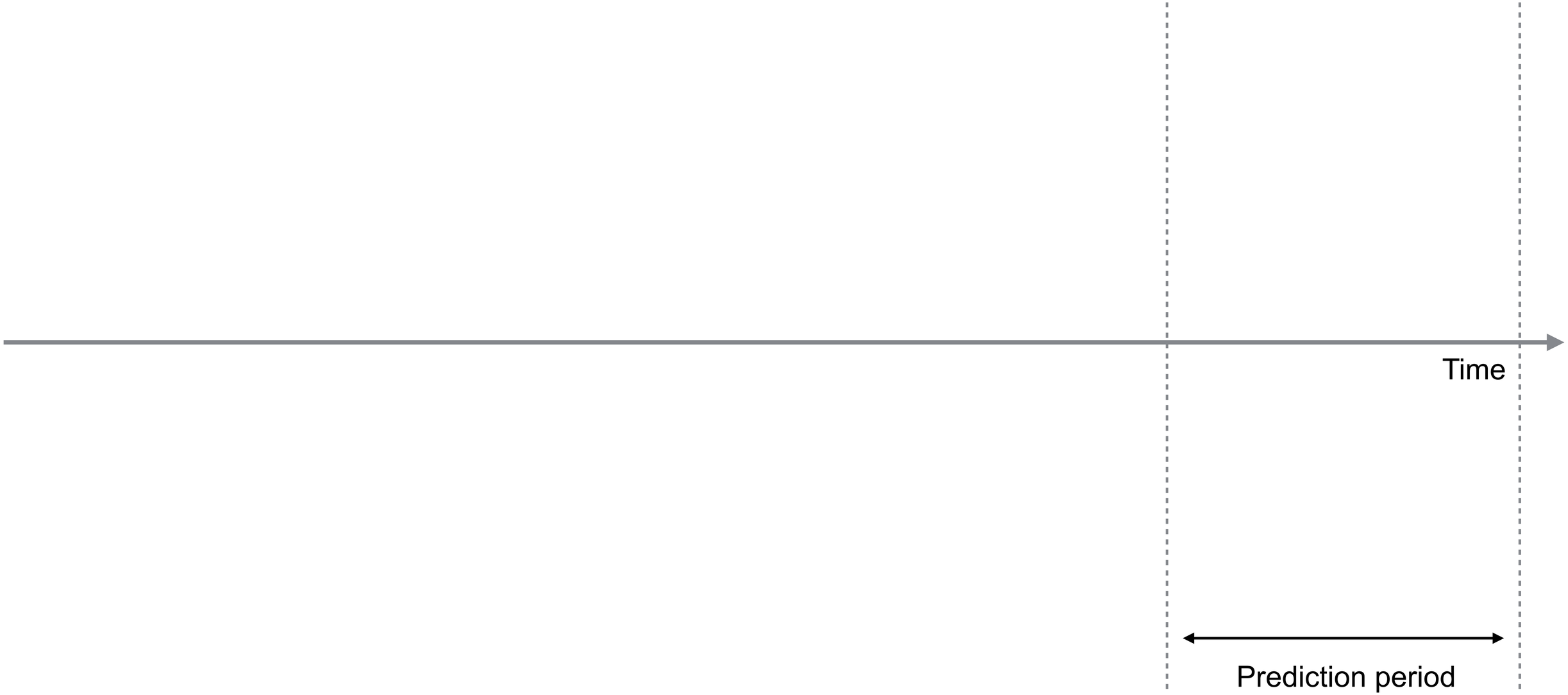
Our retail data

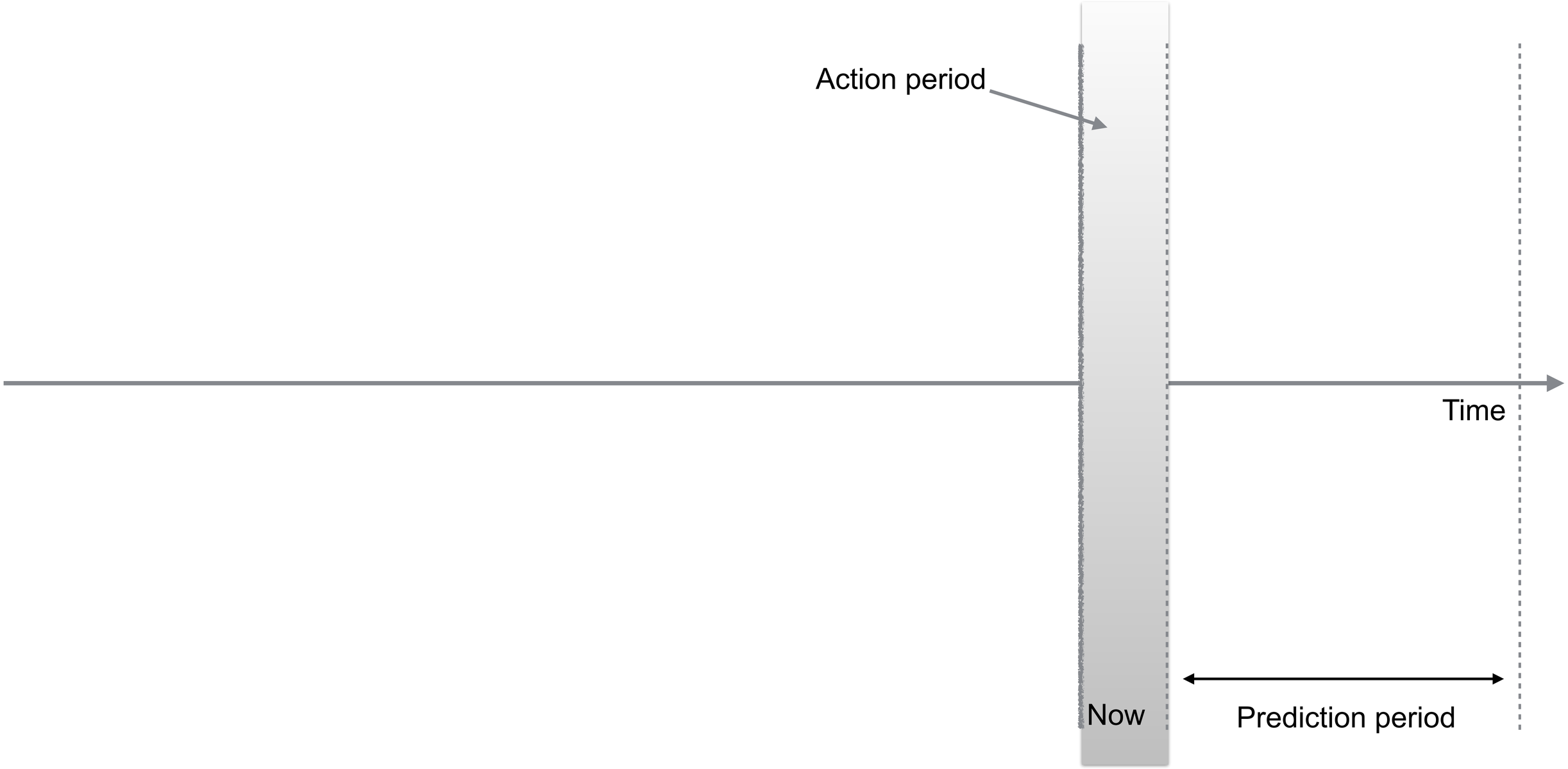


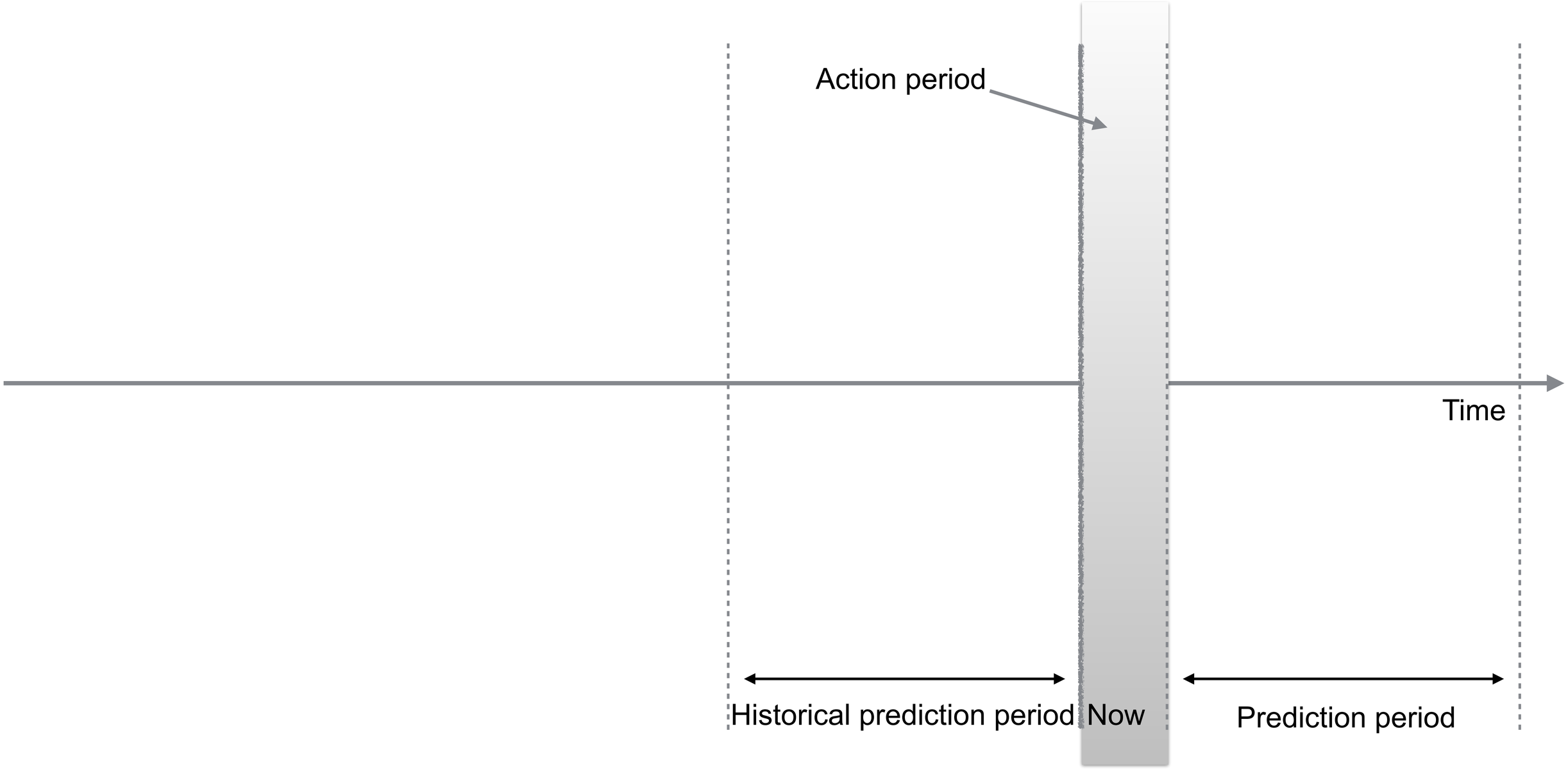
devCON

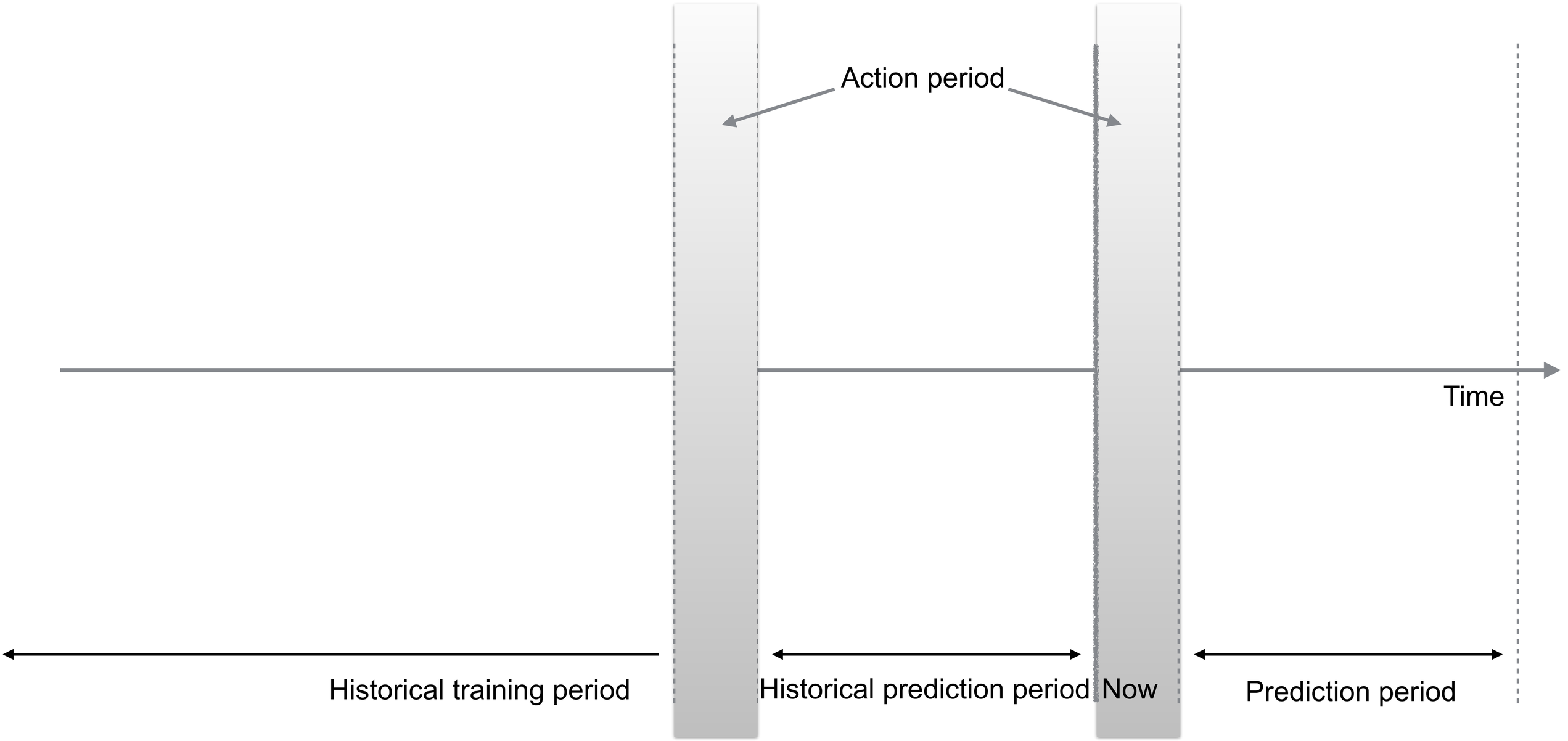
Experimental setup

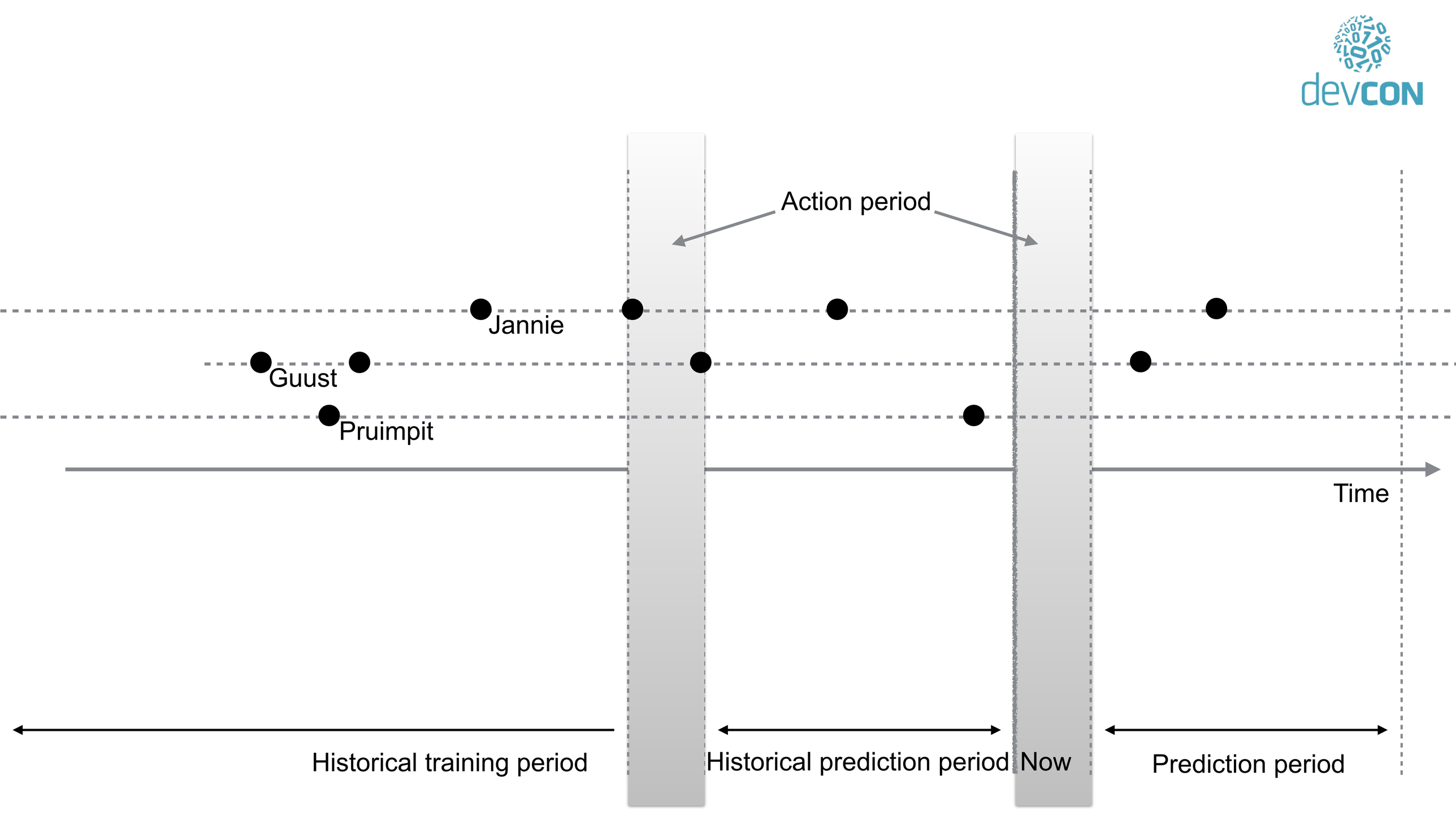

Time
Prediction period

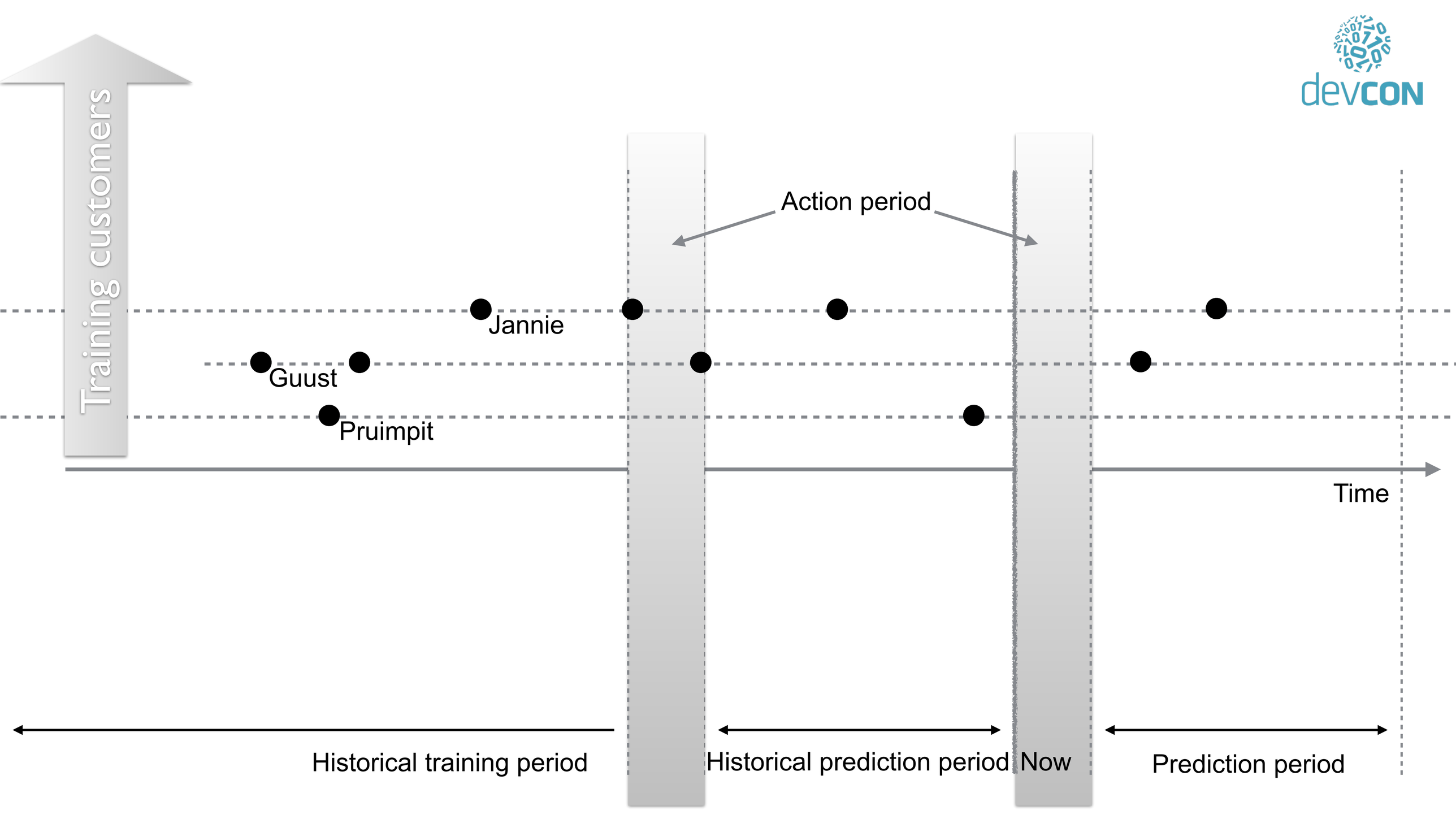

Time
Prediction period

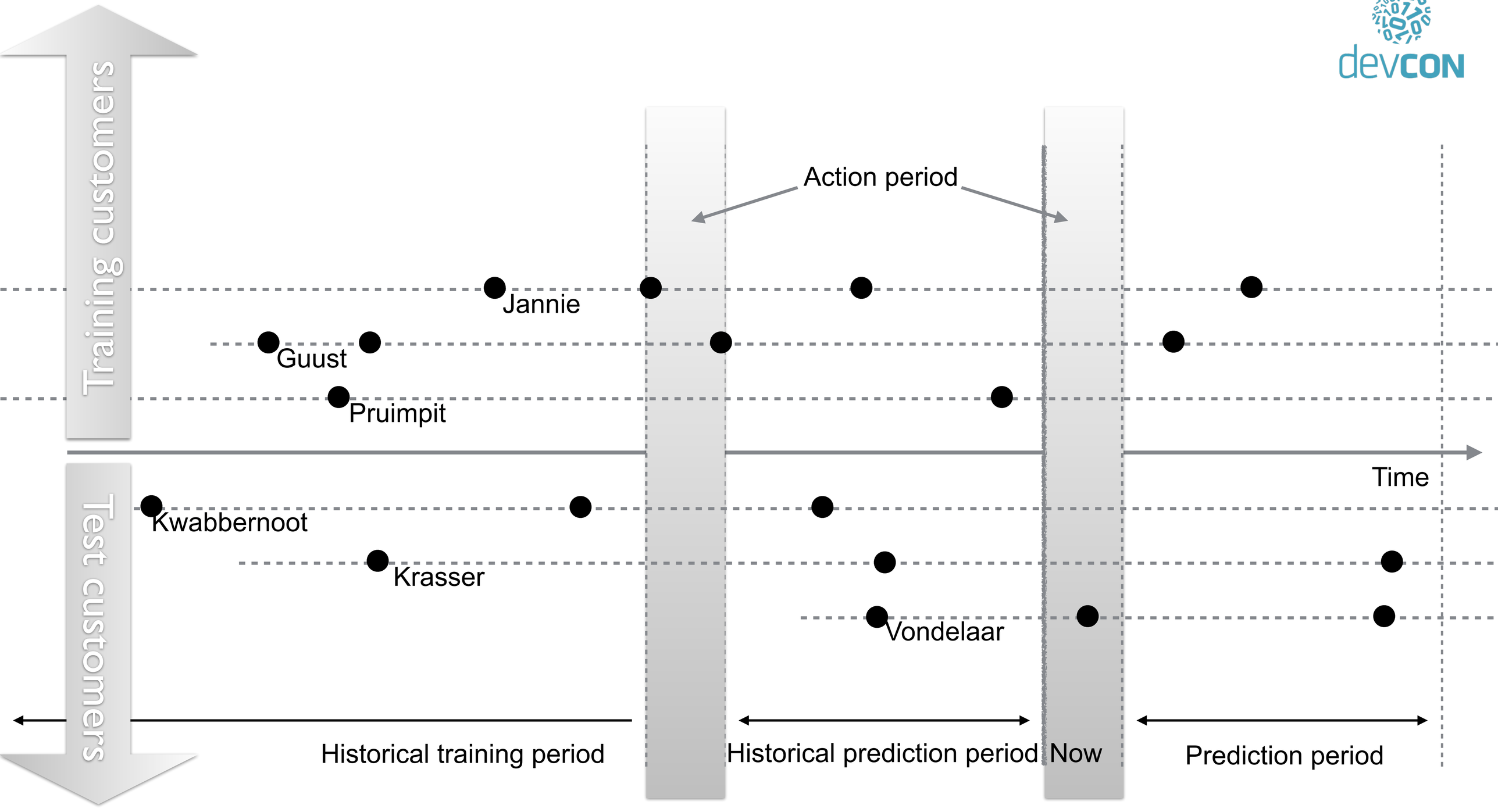

Action period
Time
Now
Prediction period

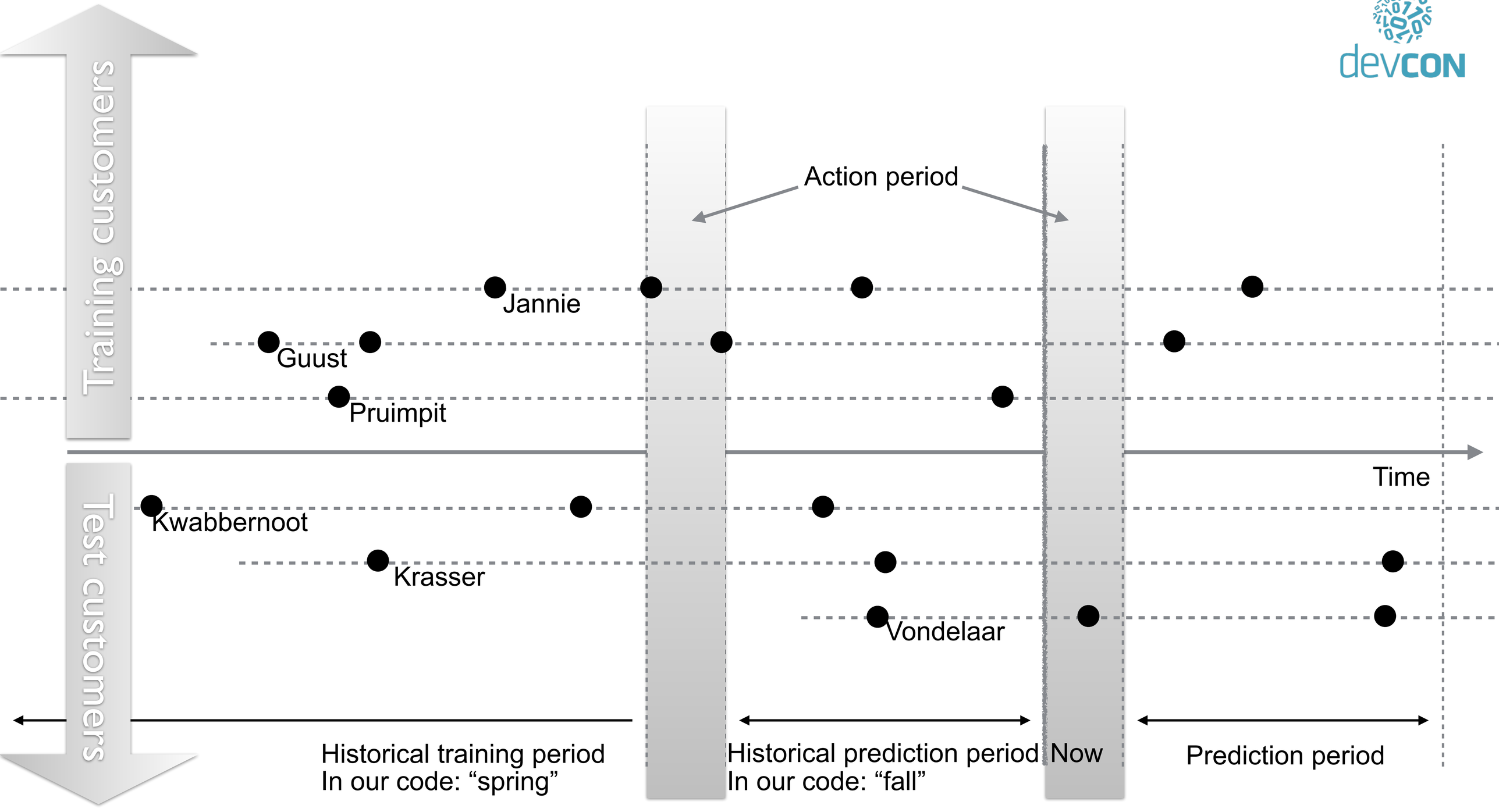

Action period
Time
Historical prediction period
Now
Prediction period


Action period
Time
Historical training period
Historical prediction period
Now
Prediction period


devCON
Action period
Jannie
Guust
Pruimpit
Time
Historical training period
Historical prediction period
Now
Prediction period


devCON
Training customers
Action period
Jannie
Guust
Pruimpit
Time
Historical training period
Historical prediction period
Now
Prediction period


devCON
Training customers
Action period
Jannie
Guust
Pruimpit
Time
Test customers
Kwabbernoot
Krasser
Vondelaar
Historical training period
Historical prediction period
Now
Prediction period


Training customers
Action period
Jannie
Guust
Pruimpit
Time
Test customers
Kwabbernoot
Krasser
Vondelaar
Historical training period
In our code: “spring”
Historical prediction period
In our code: “fall”
Now
Prediction period

Evaluation

- How many customers that came back did we predict correctly?
- How many that left?
- How close was our prediction of what they spent?

Data frame layout

Index	Independent 1	Independent 2	Returning Customer?	Revenue
Customer 1				
Customer 2				
Customer 3				
Customer 4				

Feature engineering

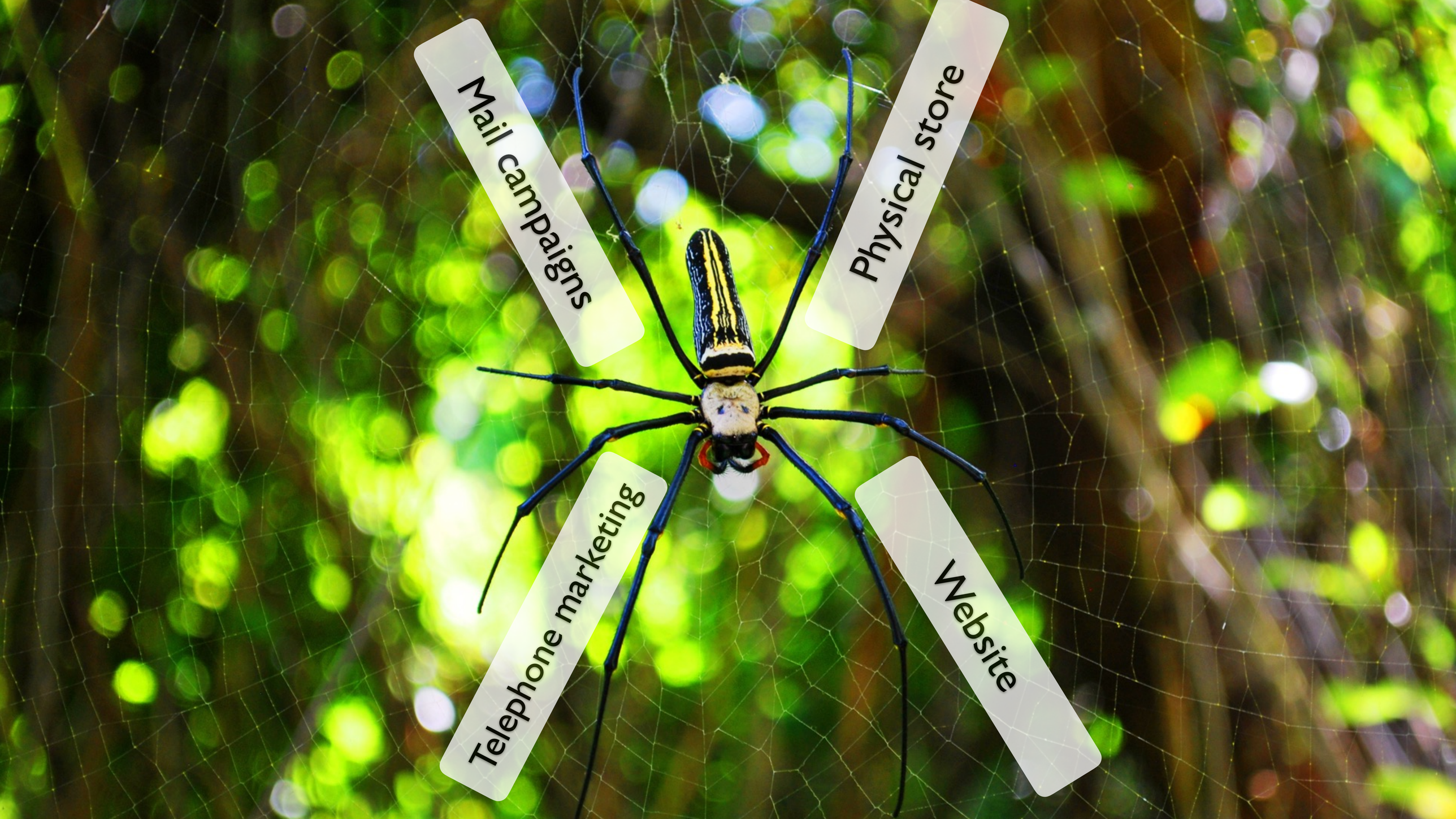
Customer journey

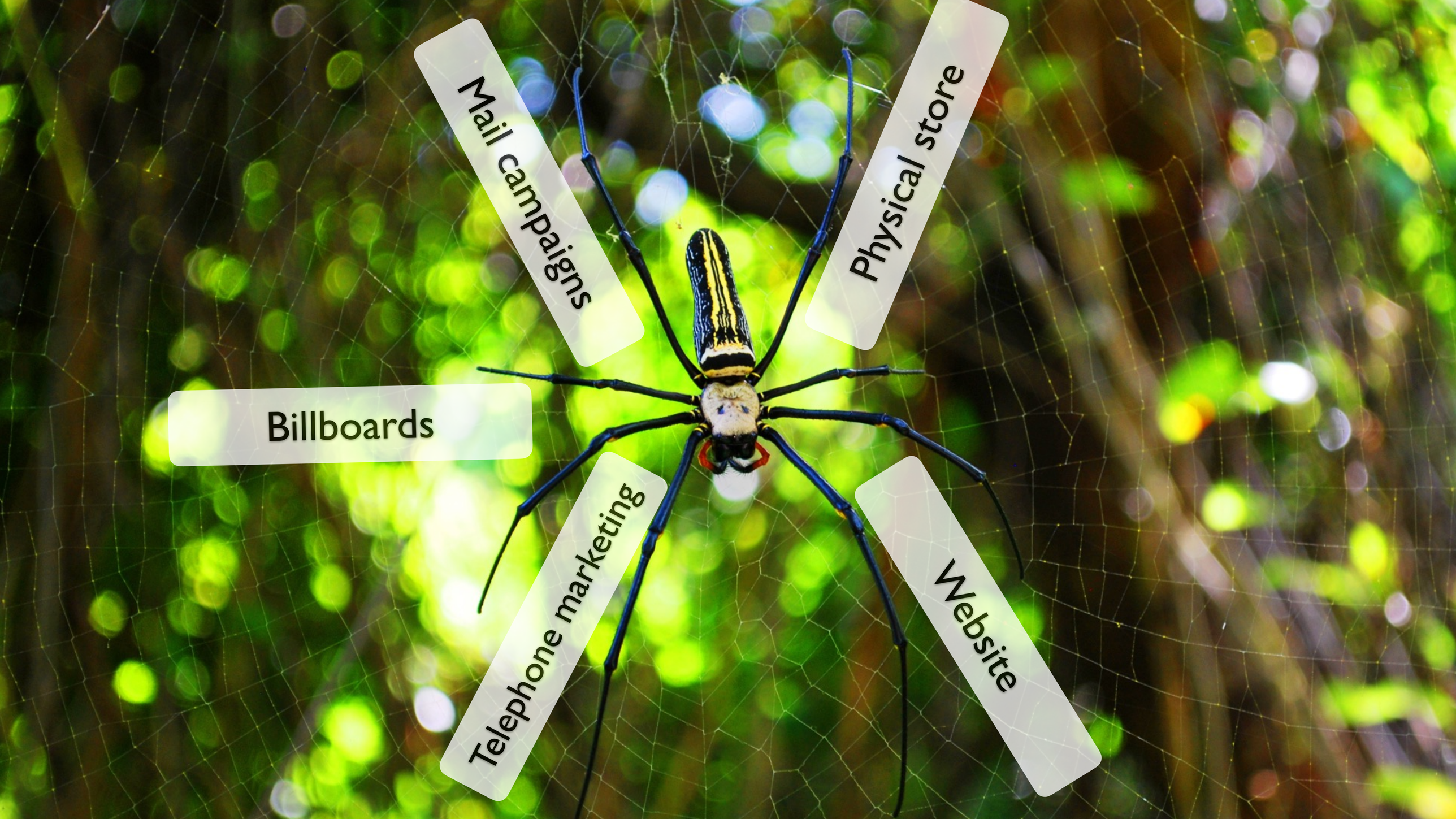
- So let’s take a look at a typical one

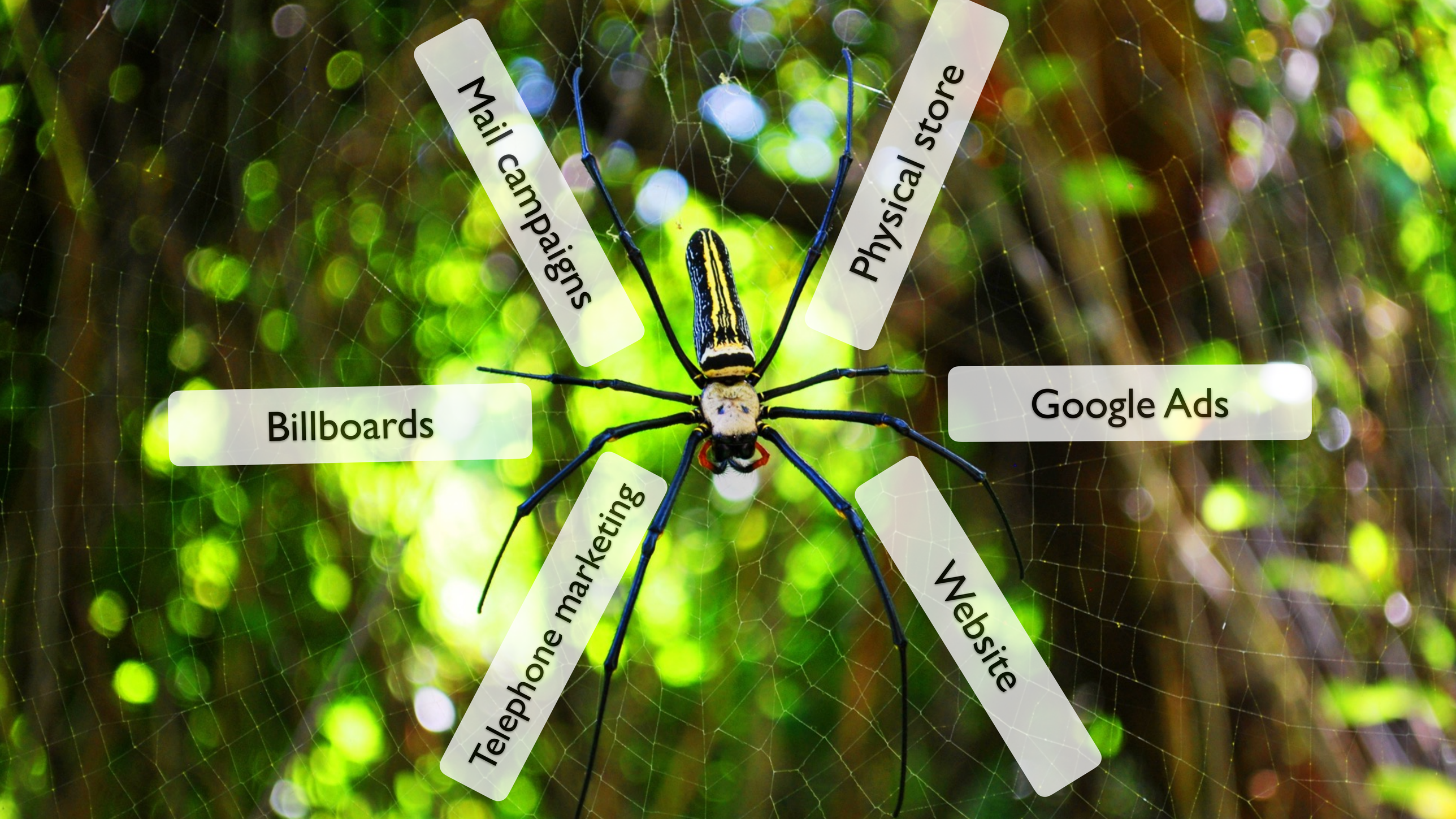

Physical store


Physical store
Website


Mail campaigns
Physical store
Website


Mail campaigns
Physical store
Telephone marketing
Website


Mail campaigns
Physical store
Billboards
Telephone marketing
Website


Mail campaigns
Physical store
Billboards
Google Ads
Telephone marketing
Website

Customer buying behaviour

- How many orders
- How much per order
- How much time between orders
- Paying behaviour

Usage behaviour

- Complaints
- Newsletter opt-in / opt-out

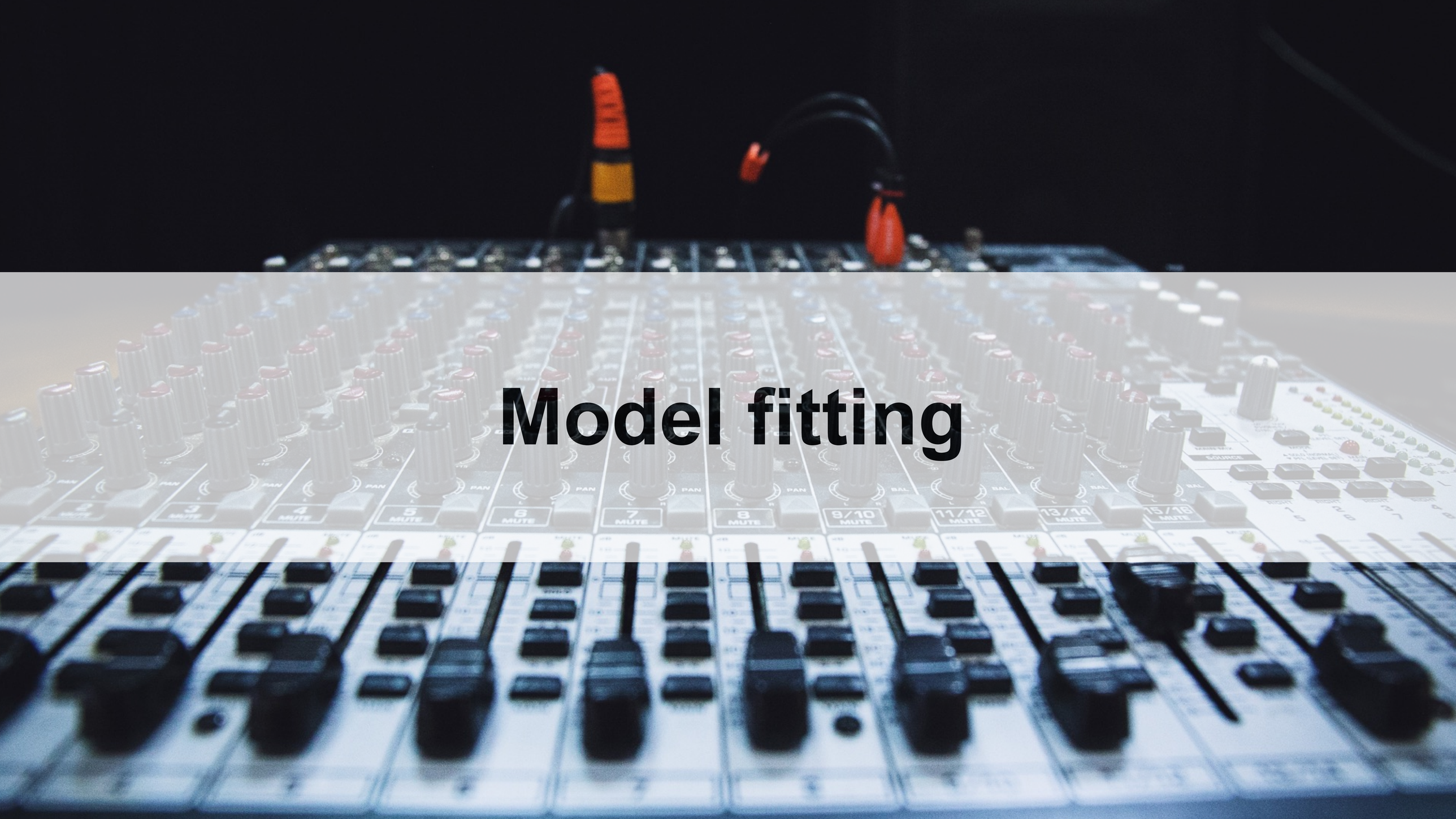
Sharing behaviour

- Social media
- Member get member

Discarding behaviour

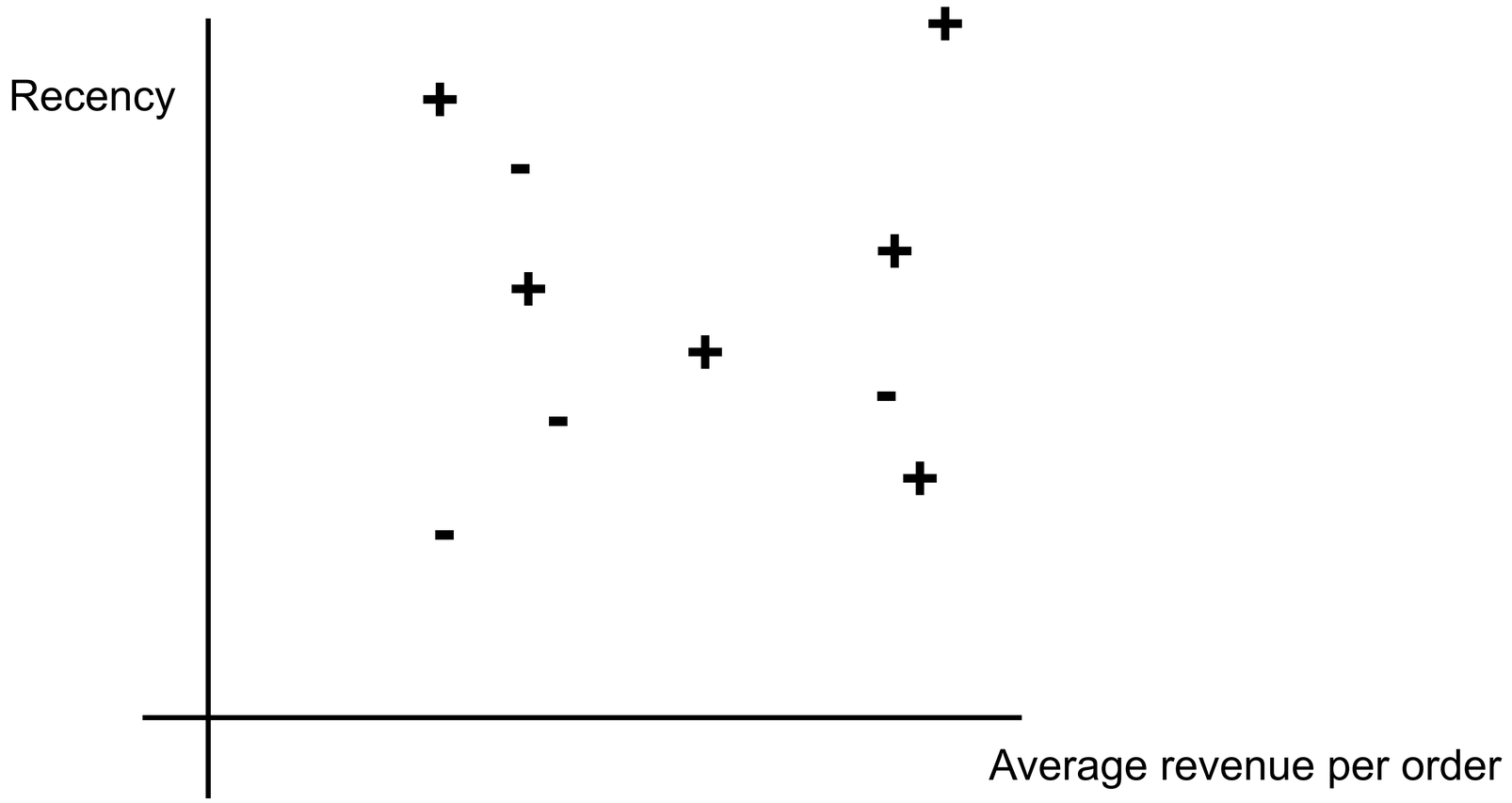
- Behaviour observed around ending of customer relationship

devCON

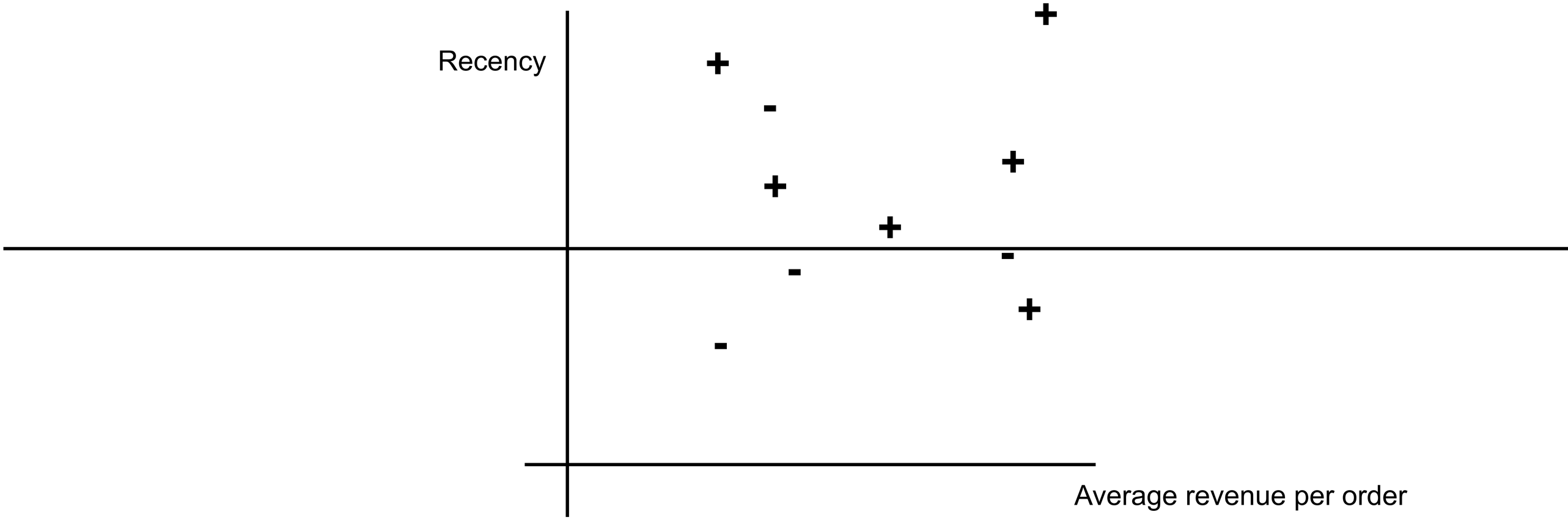


Model fitting

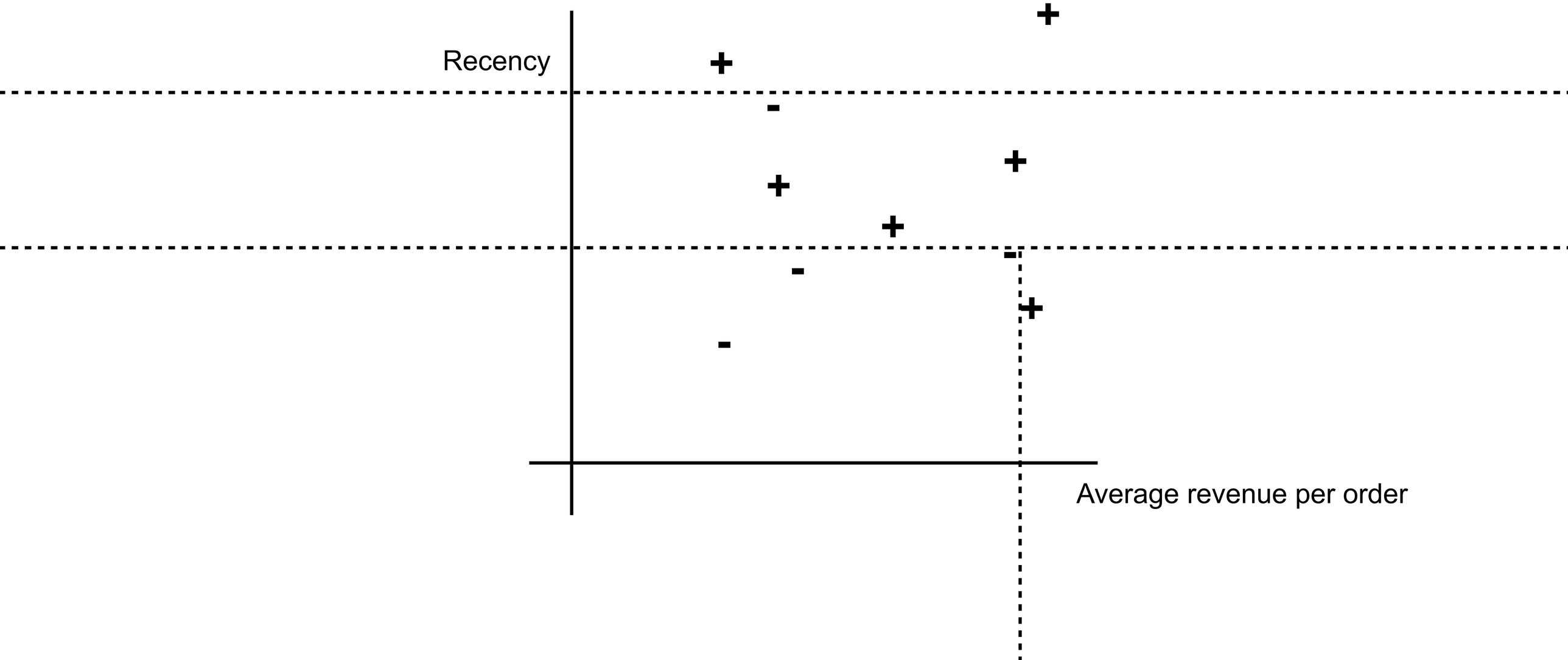
Decision trees



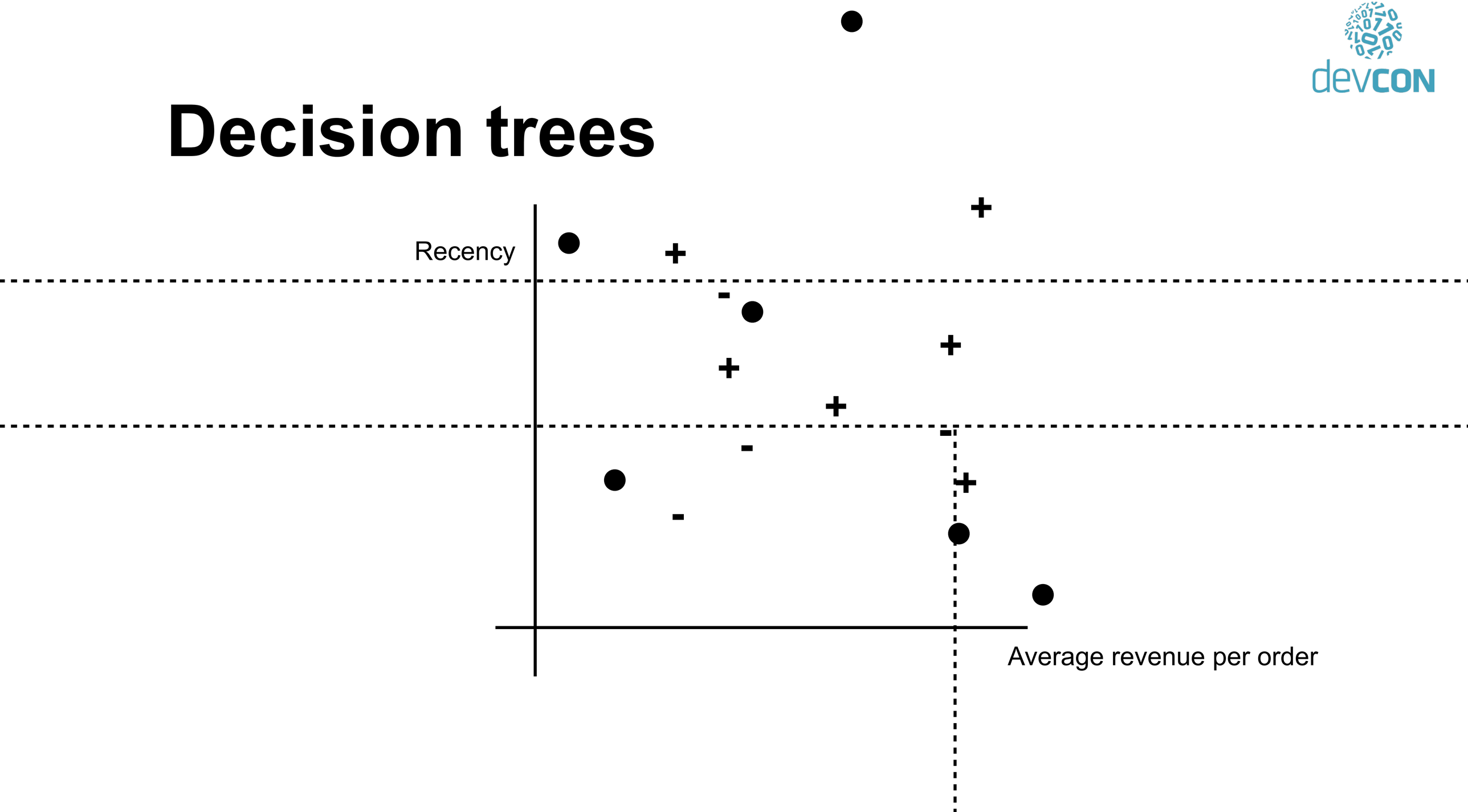
Decision trees



Decision trees



Decision trees



devCON

Take away points

- Understand business model
- Careful experimental setup
- Interactive data science:
  - Fun
  - Productive

Thank you

Any Questions?

devCON

is powered by

luminis
Conversing worlds

My world is:

Nou moe?! - Guust Flater

All our worlds › www.luminis.eu

Pearson correlation

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{s(x)s(y)}.$$

Precision

- tp / fp + tp

Recall

- tp / tp + fn

Fβ-score

- (1 + β^2) (precision x recall) / (β^2 precision + recall)

Gini impurity in node m

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Summing over class labels k

Mean square error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

R^2: coefficient of determination

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2}$$