Introduction to Clustering in Apache Beam

Jasper Van den Bossche
ML6

BEAM SUMMIT
NYC 2023

Agenda

- What is clustering?
  - Online vs offline clustering
  - What are the applications?
- How does clustering in Apache Beam work
  - High level overview of the transform
- Example pipeline

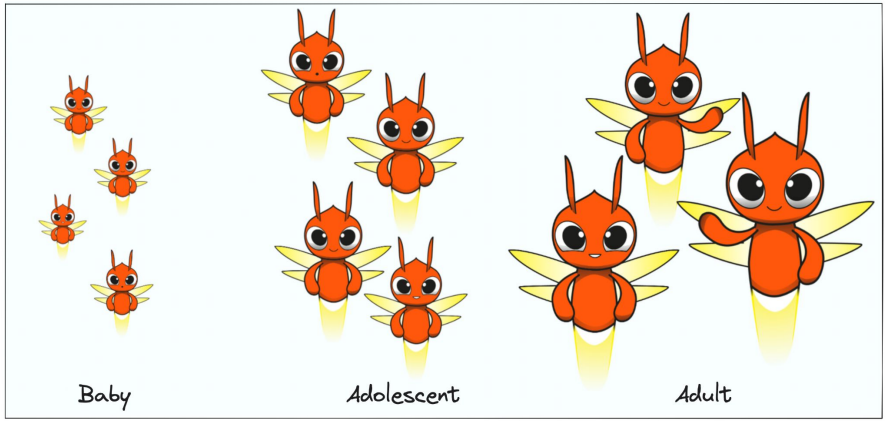
What is clustering?

What is clustering?

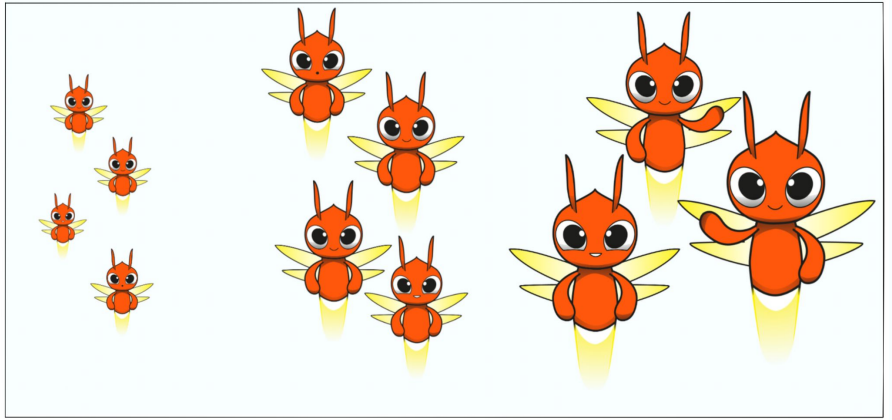
What is clustering?

Clustering is an ***unsupervised*** technique used to ***group similar data points*** together based on their ***characteristics*** or ***patterns***.

What is Unsupervised Training?

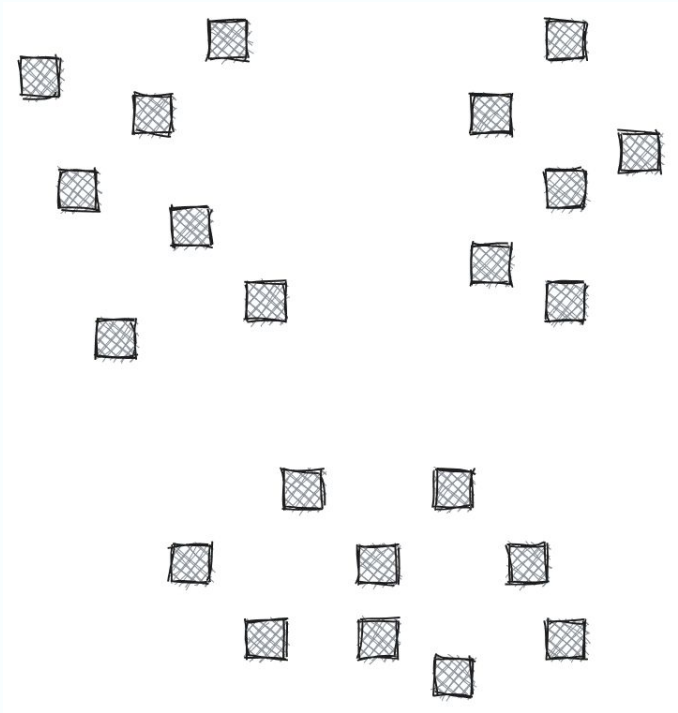


Supervised

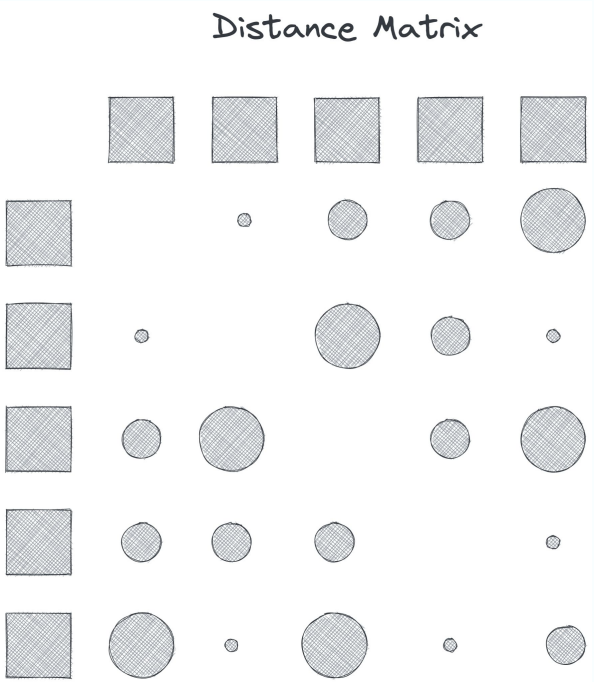


Unsupervised

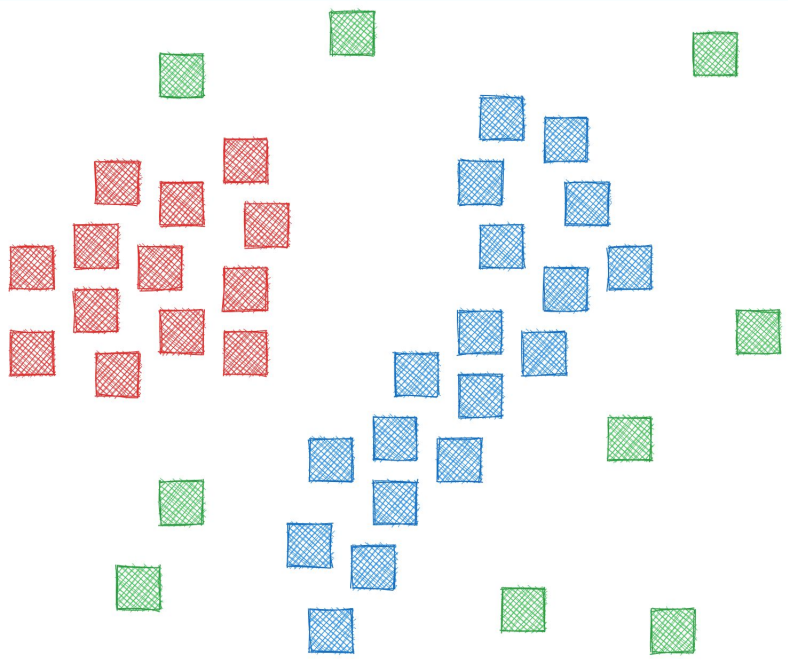
How are datapoints grouped together?



How are datapoints grouped together?

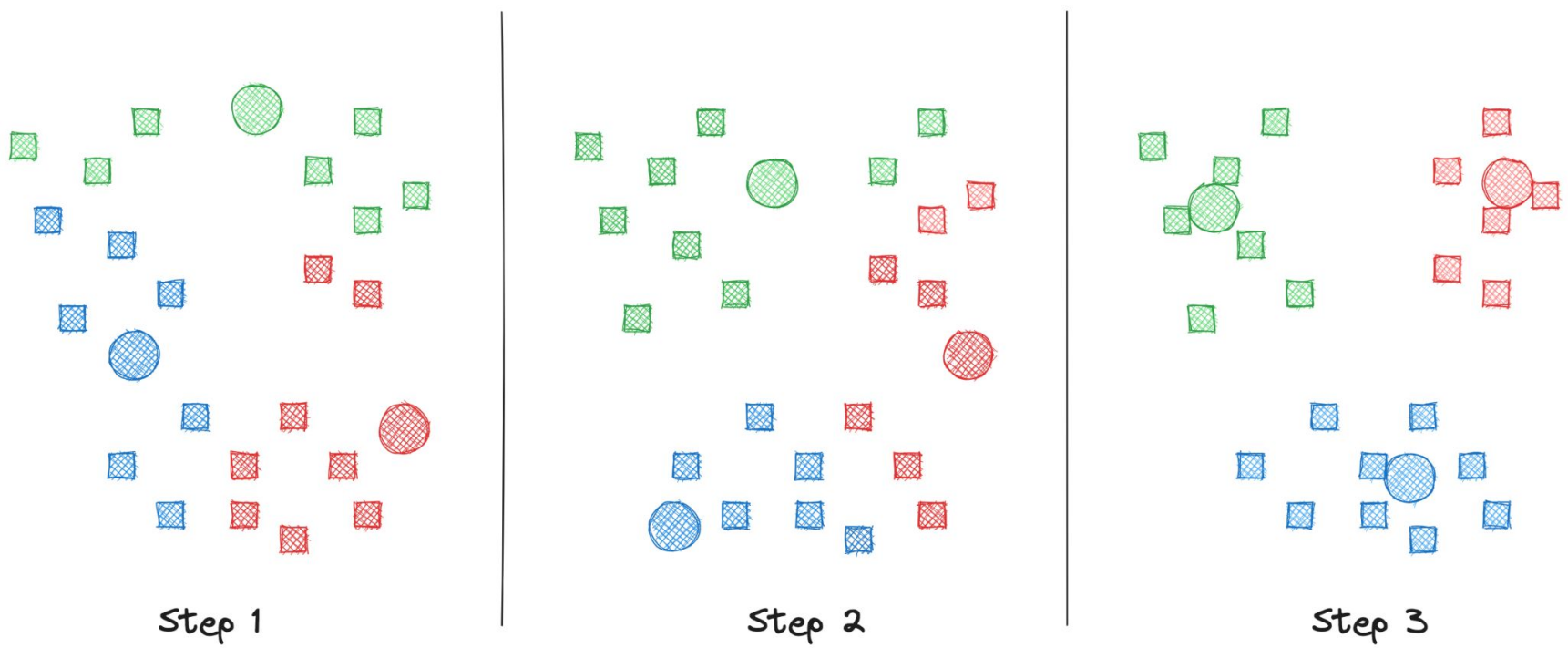


Spectral Clustering



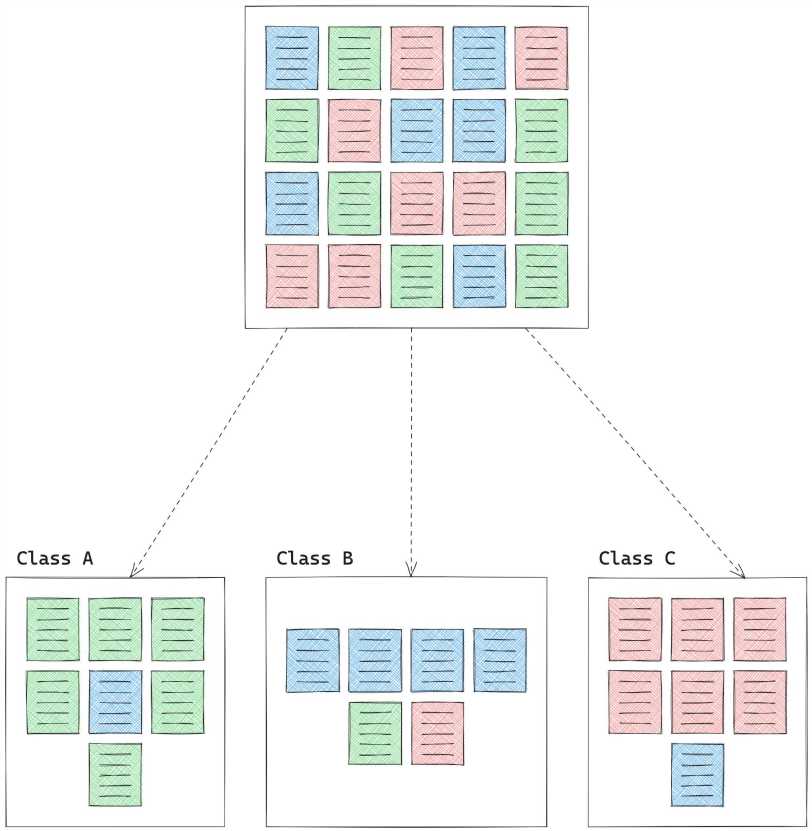
DBSCAN

K-means clustering

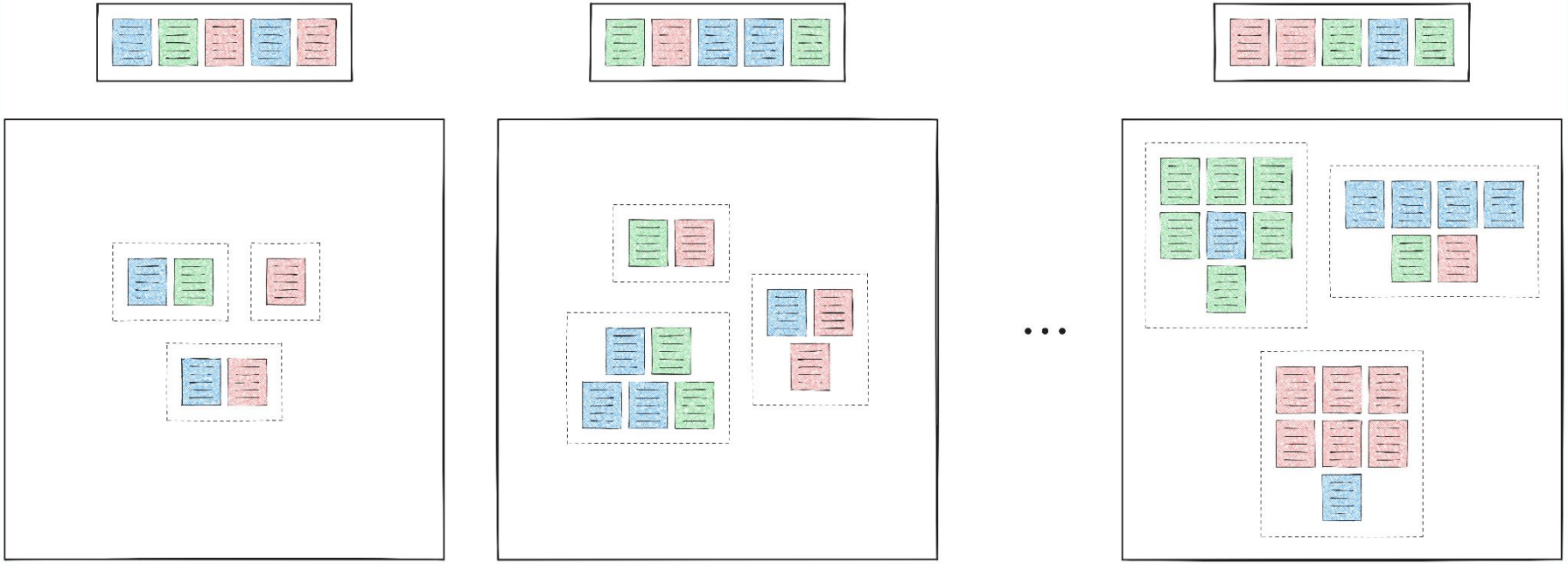


Online vs offline clustering

Offline Clustering



Online Clustering

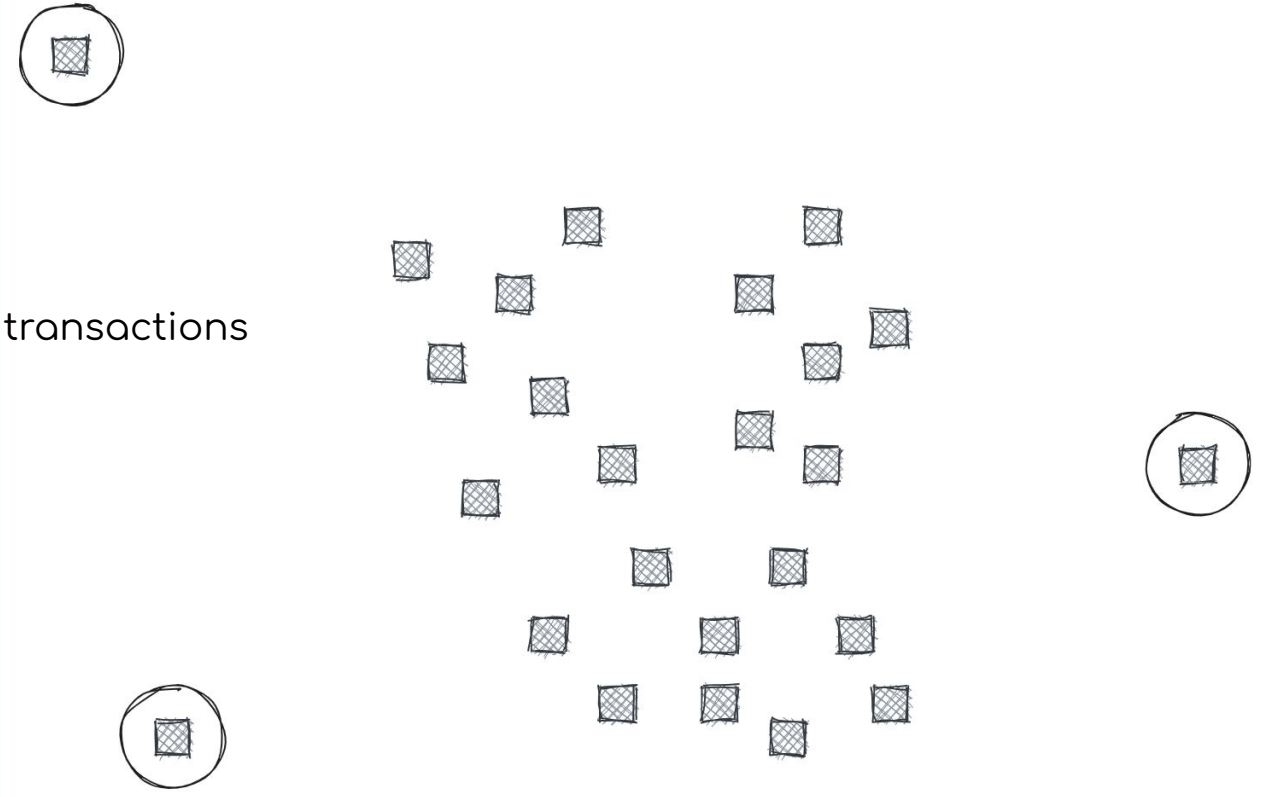


What are the applications of clustering?

What are the applications of clustering?

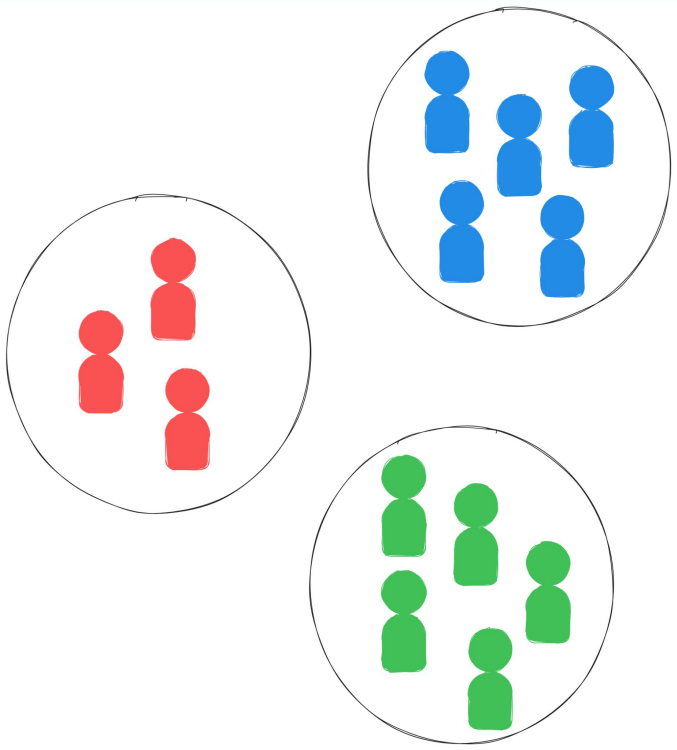
Anomaly detection

- Detect fraudulent transactions
- Detect diseases
- Quality control
- Spam filters

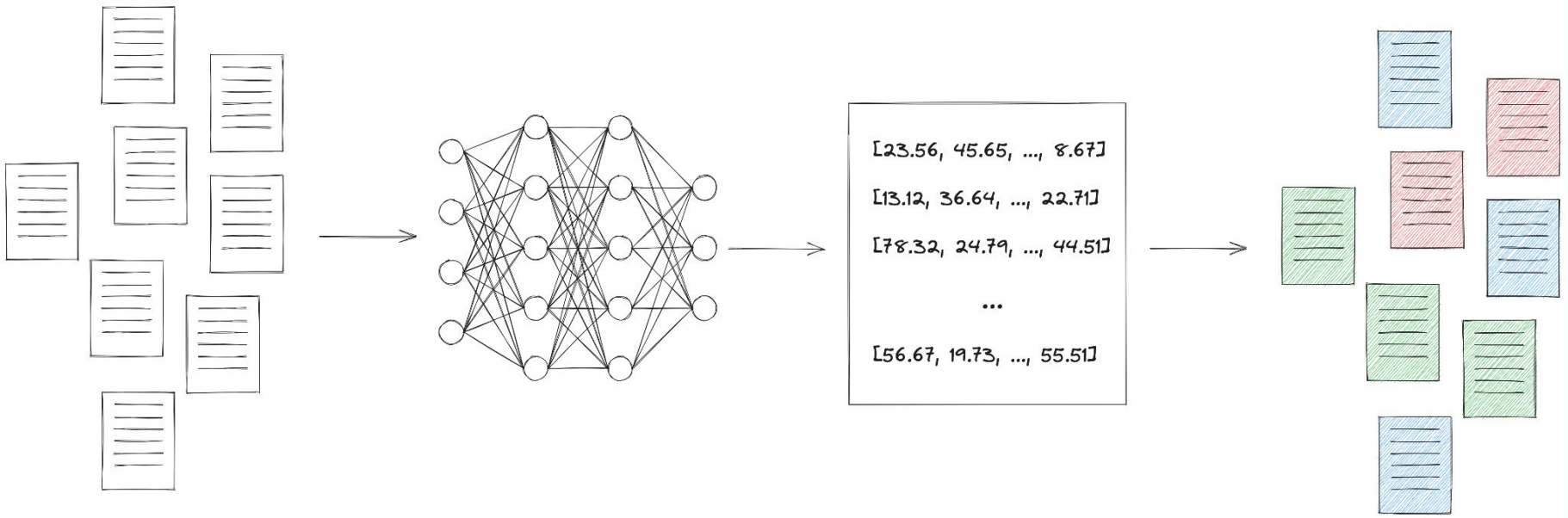


Personalisation

- Personalised ads
- Movie/music recommendations



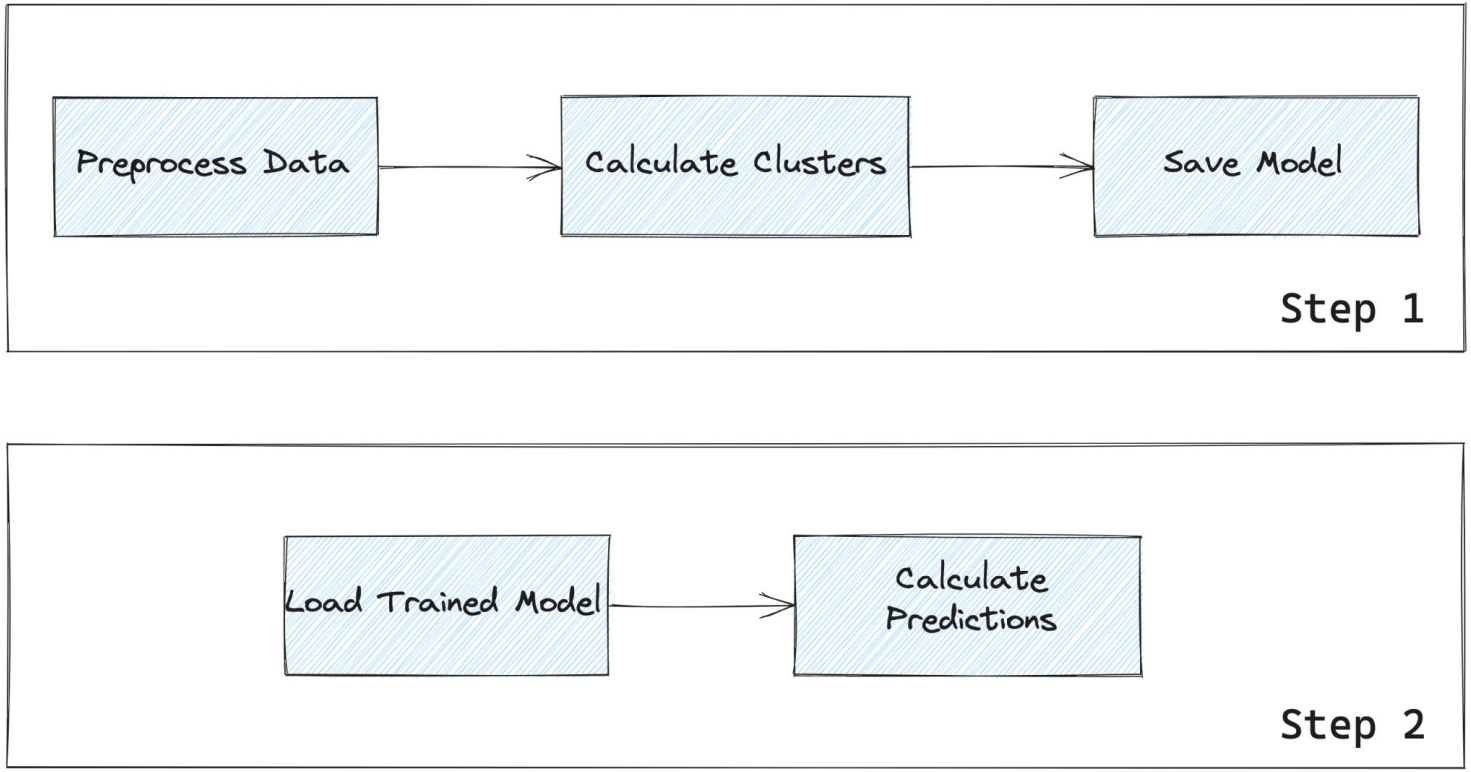
Grouping documents



1. Use a language model to calculate embeddings
2. Group together points in the embedding space close to each other

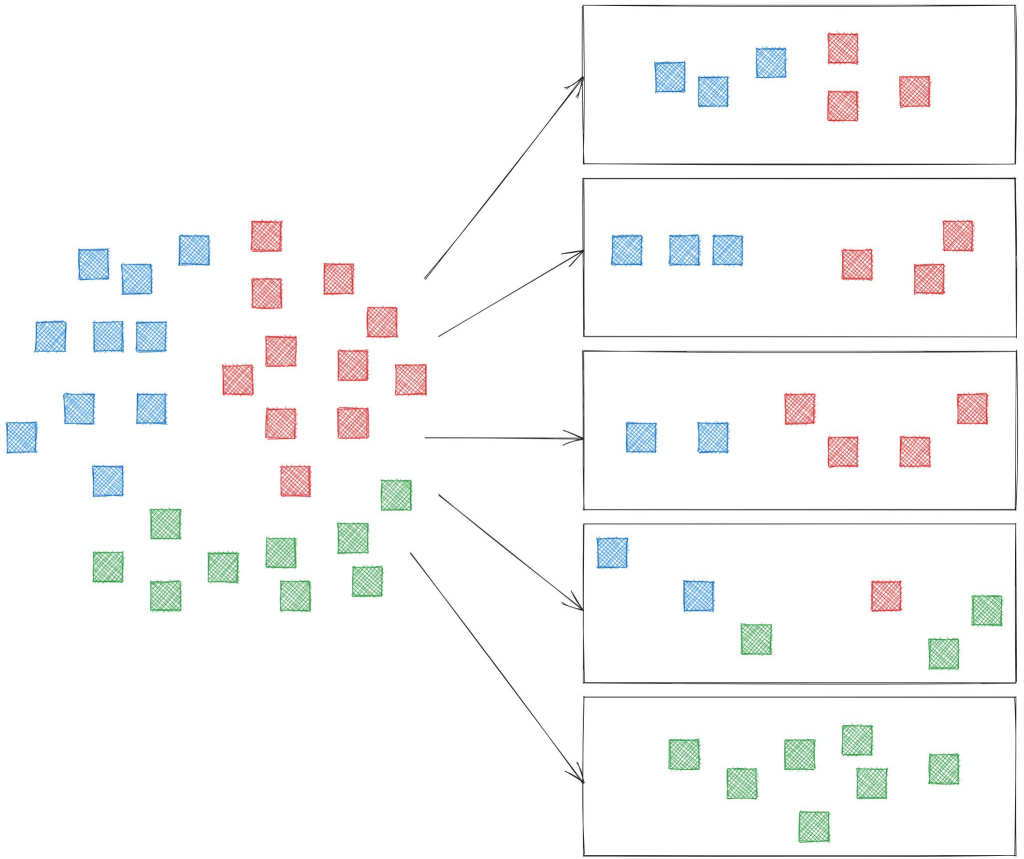
How does clustering in Apache Beam work

A High Level Look Behind the Scenes

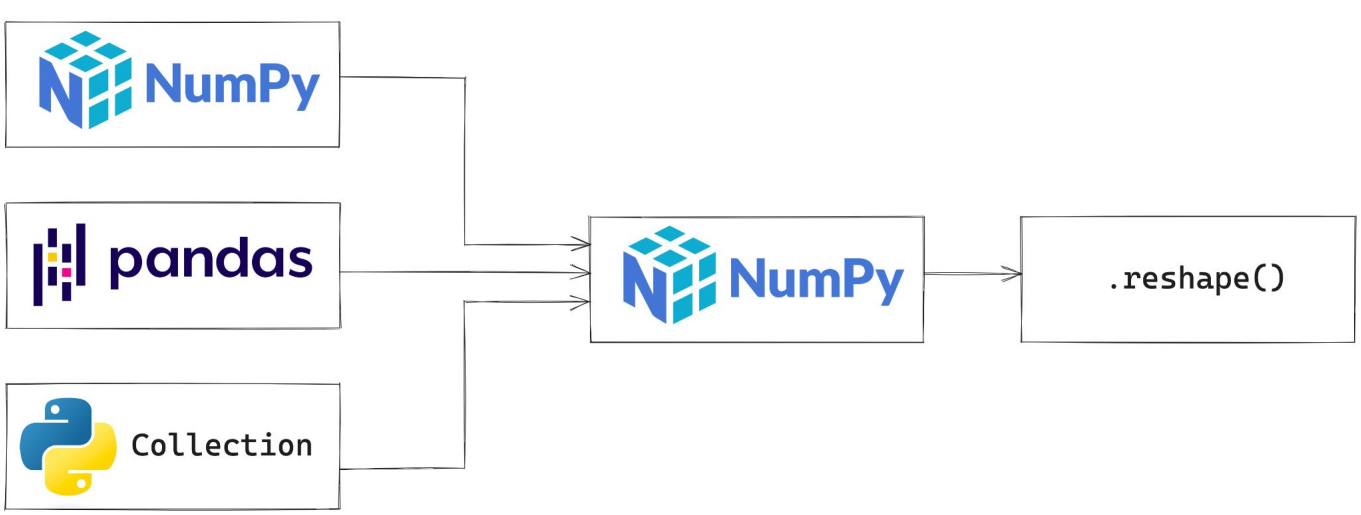


Preprocessing

1. Create Batches of Datapoints



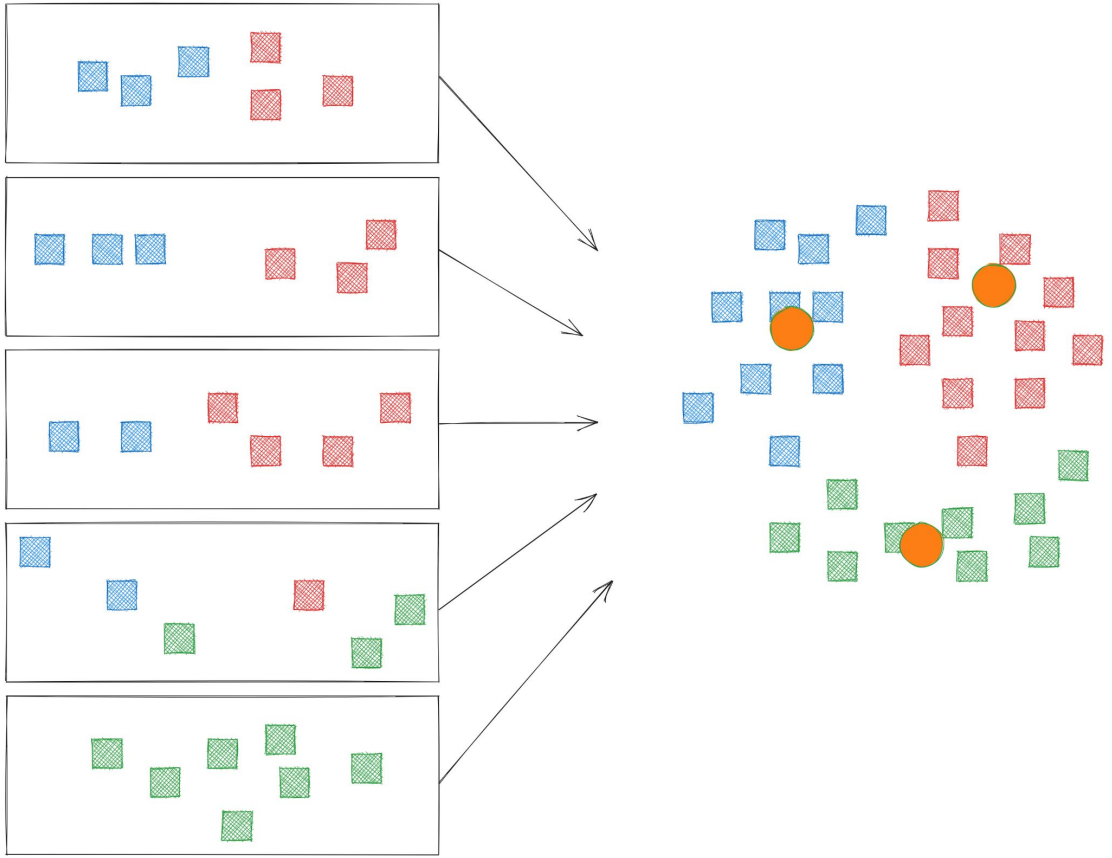
Preprocessing



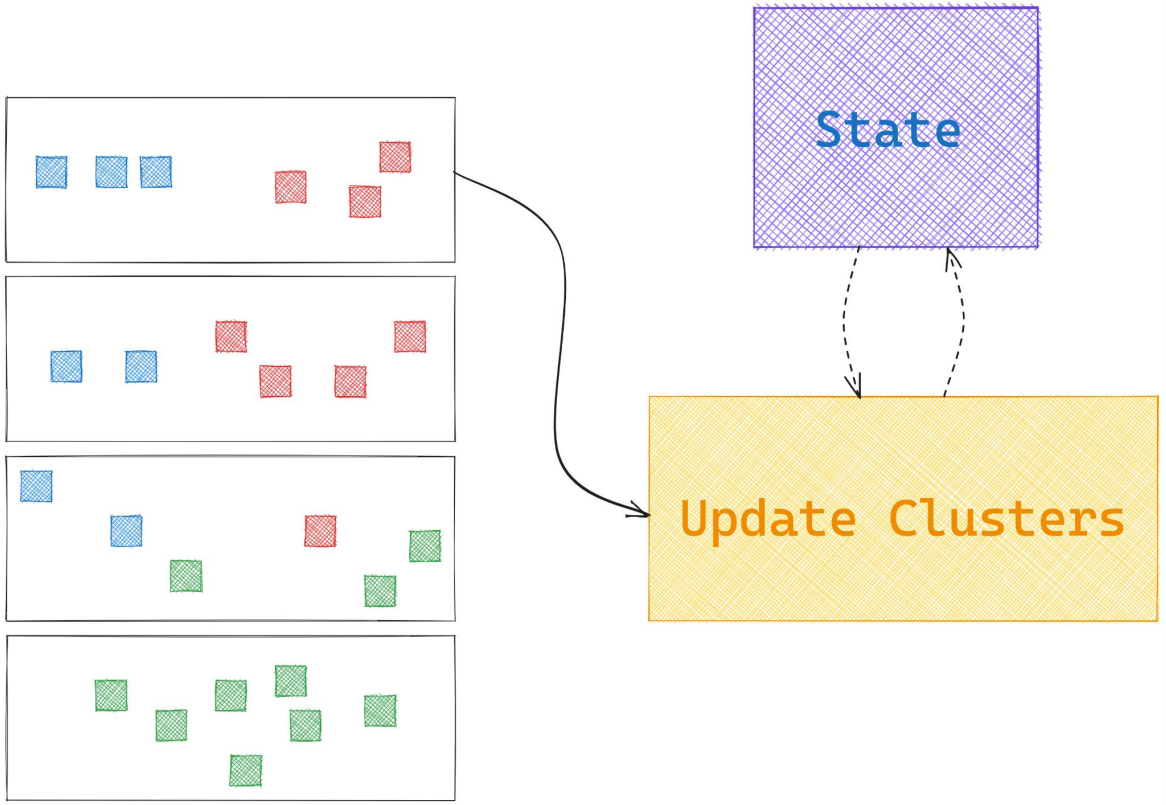
2. Convert to Numpy and Reshape

Calculate Cluster Centers

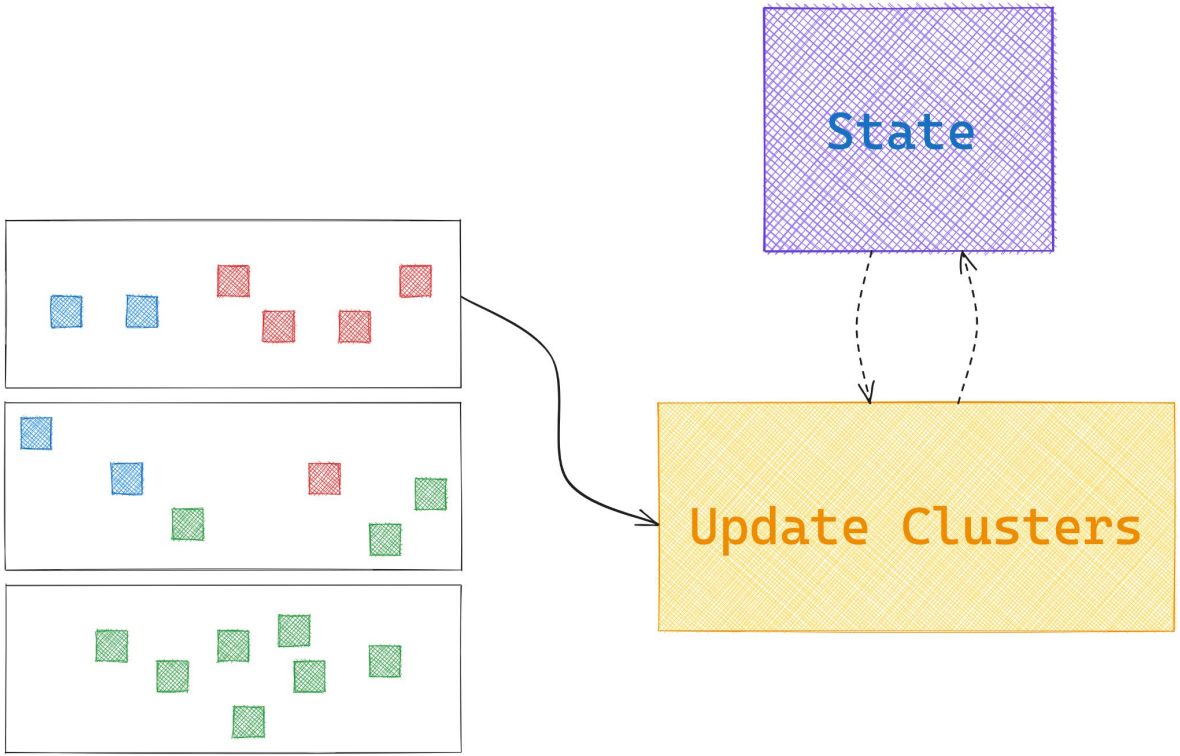
Process batch by batch
to calculate cluster centers



Clustering is a stateful transform

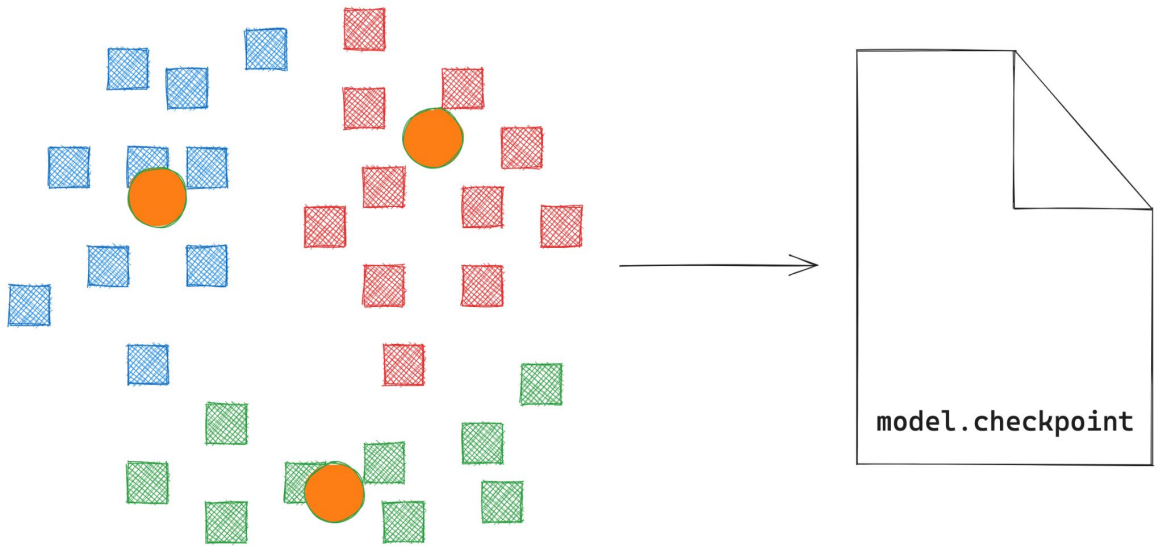


Clustering is a stateful transform



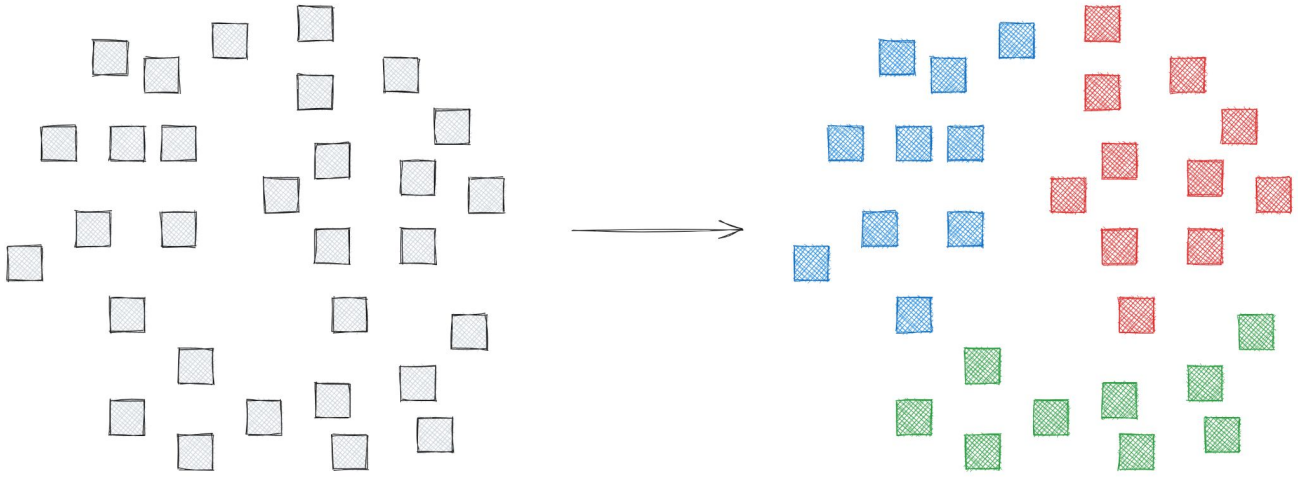
Save Model

Save the trained model to persistent storage



Assign Labels

Assign all datapoints a label using the trained model



Let’s look at an example!

Example: Clustering California Houses

Group similar houses based on location and income of the owner

longitude	latitude	income
-122.23	37.83	52.000
-122.28	37.81	152.000
-122.17	37.82	48.000
-122.26	37.79	56.000
-122.23	37.84	72.000

Preparing Data

```
# 1. Calculate clustering centers and save model to persistent storage
model = (
    housing_features
    | "Train clustering model" >> OnlineClustering(
        OnlineKMeans,
        n_clusters=6,
        batch_size=256,
        cluster_args={},
        checkpoints_path=known_args.checkpoints_path))
```

Training the Clustering Model

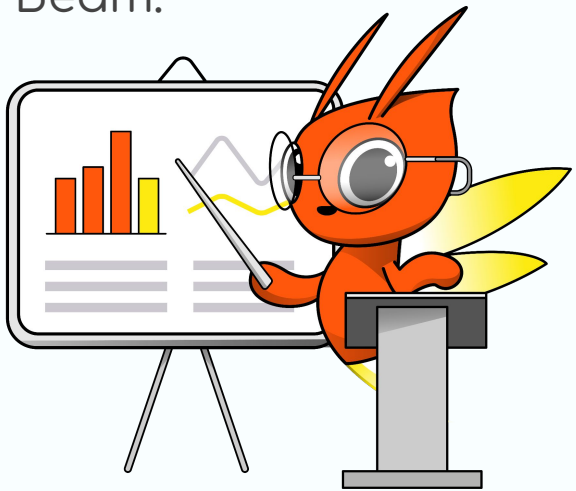
```
# 2. Calculate labels for all records in the dataset
# using the trained clustering model using in memory model
_ = (
    housing_features
    | "RunInference" >> AssignClusterLabelsInMemoryModel(
        model=pvalue.AsSingleton(model),
        model_id="kmeans",
        n_clusters=6,
        batch_size=512)
    | beam.Map(print))
```

Calculating Predictions

```
pipeline = test_pipeline
if not test_pipeline:
  pipeline = beam.Pipeline(options=pipeline_options)

data = pipeline | read_csv(known_args.input)

features = ['longitude', 'latitude', 'median_income']

housing_features = to_pcollection(data[features])
```

Summary

- Clustering is a technique to group similar datapoints based on their characteristics
- Many applications ranging from anomaly detection to document grouping
- Clustering is a twofold transform in Apache Beam:
  - Data preprocessing and model training
  - Assigning cluster labels to datapoints



Jasper Van den Bossche

QUESTIONS?