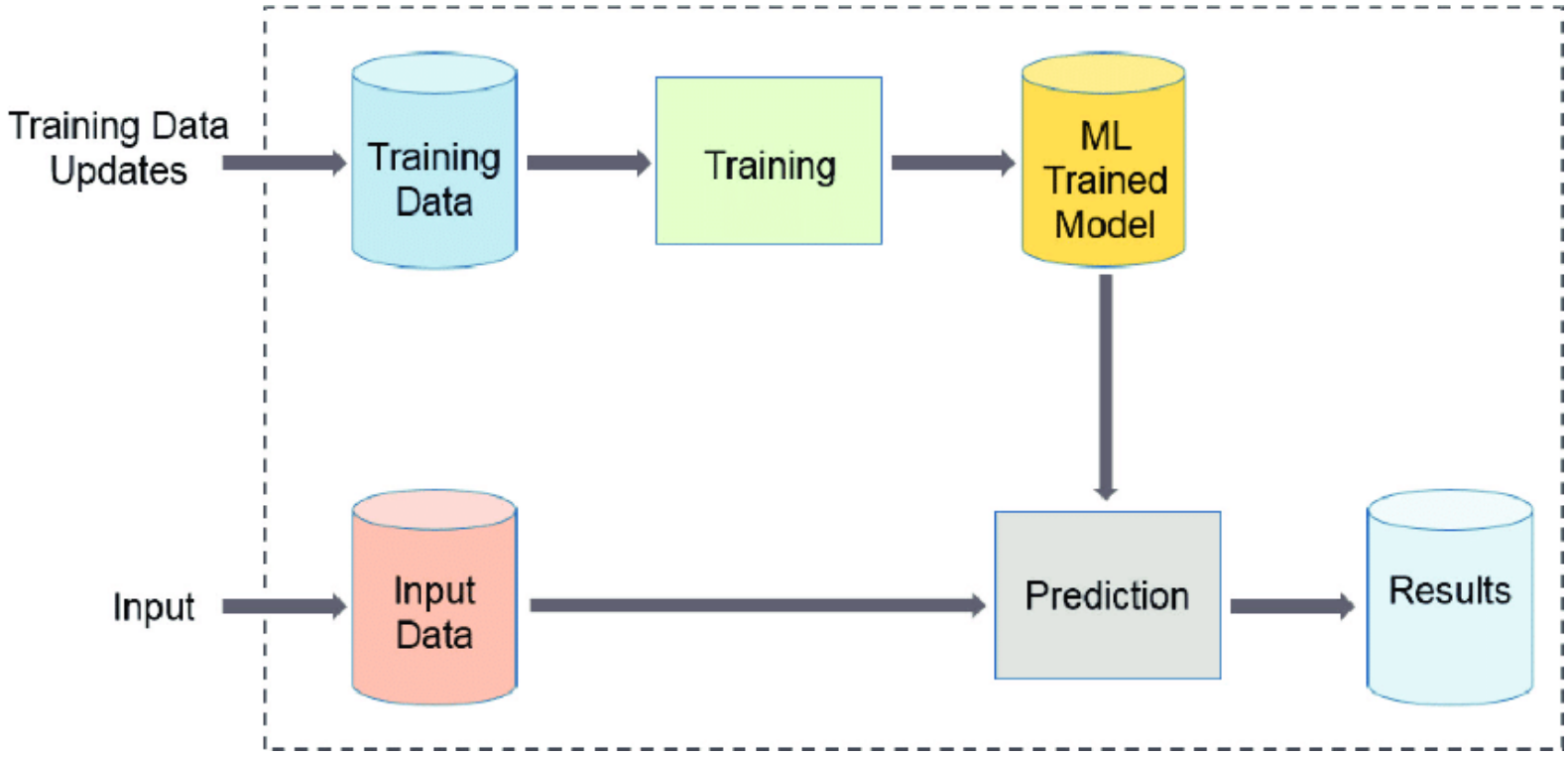
Machine Learning and Artificial Intelligence

Lab 07 – ML Pipelines

20/04/2021

The Machine Learning workflow



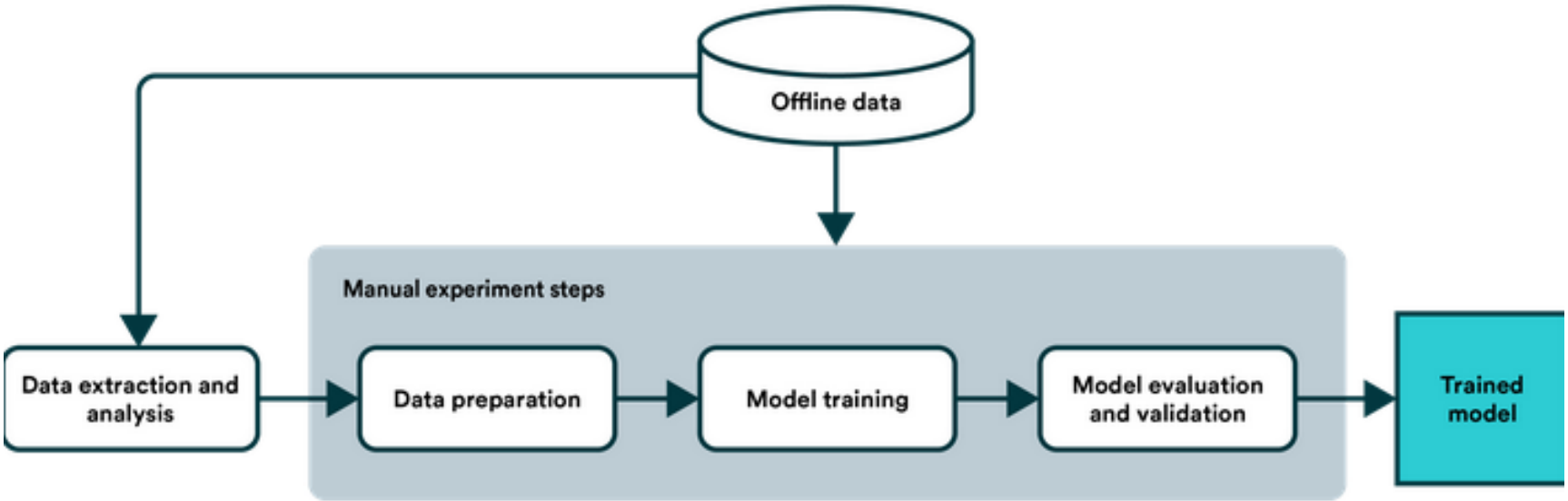
Manual workflow

- Typical problems tend to be specific and related to a single business problem, e.g: recognise the logo.

- Teams tend to start with a manual workflow, where no real infrastructure exists.

- In this paradigm and first stage, **the model is the product.**

Manual workflow

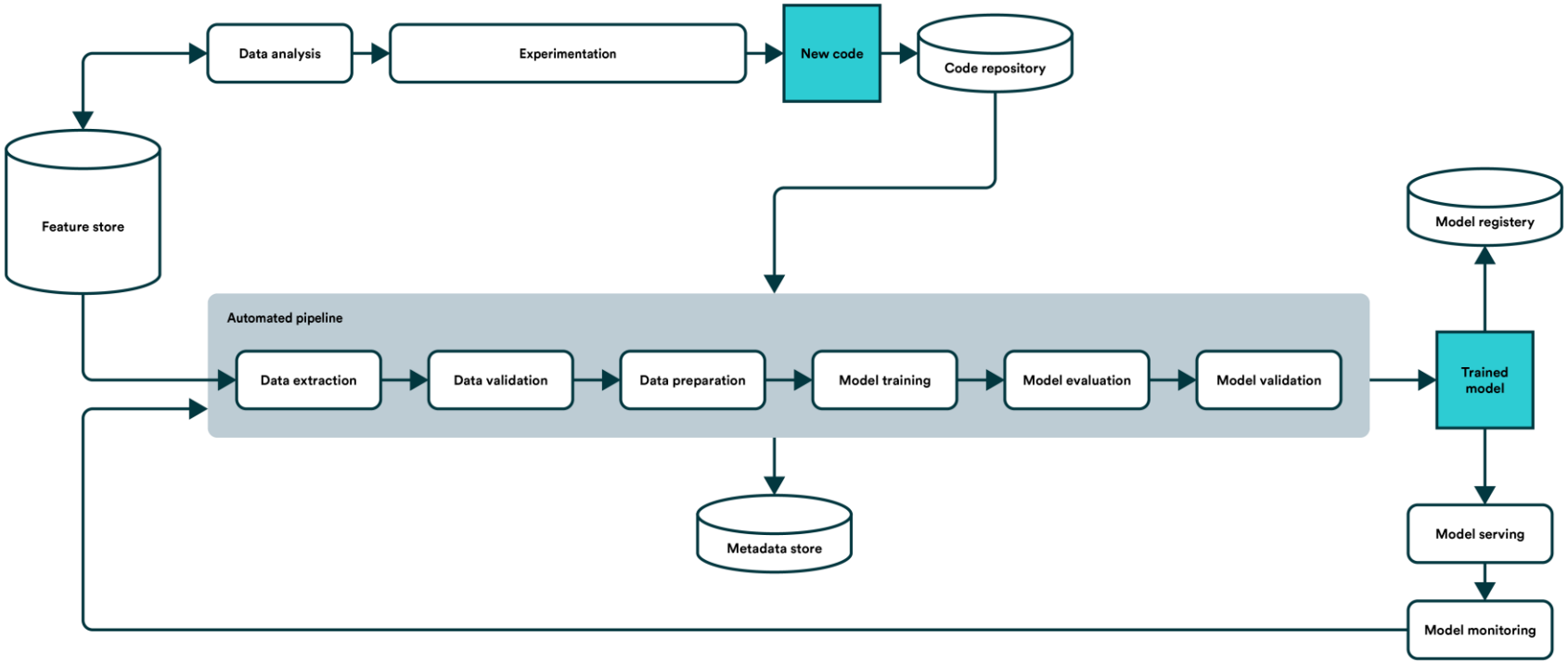


Automated workflow

- Once the problem is clear and the team has a solution, the problem shifts towards a way to keep updating the model in production.

- The product is not the model anymore, but the whole process, aka the **pipeline**.

Automated workflow



Chaining components

- Once we have known solutions and re-usable components, we can chain them together and form a sequential pipeline, without needing to manually adjust the single components.

- The components can be the following:
  - Data validation
  - Data cleanup
  - Model training
  - Model evaluation

Practical example: PCA + K-Means

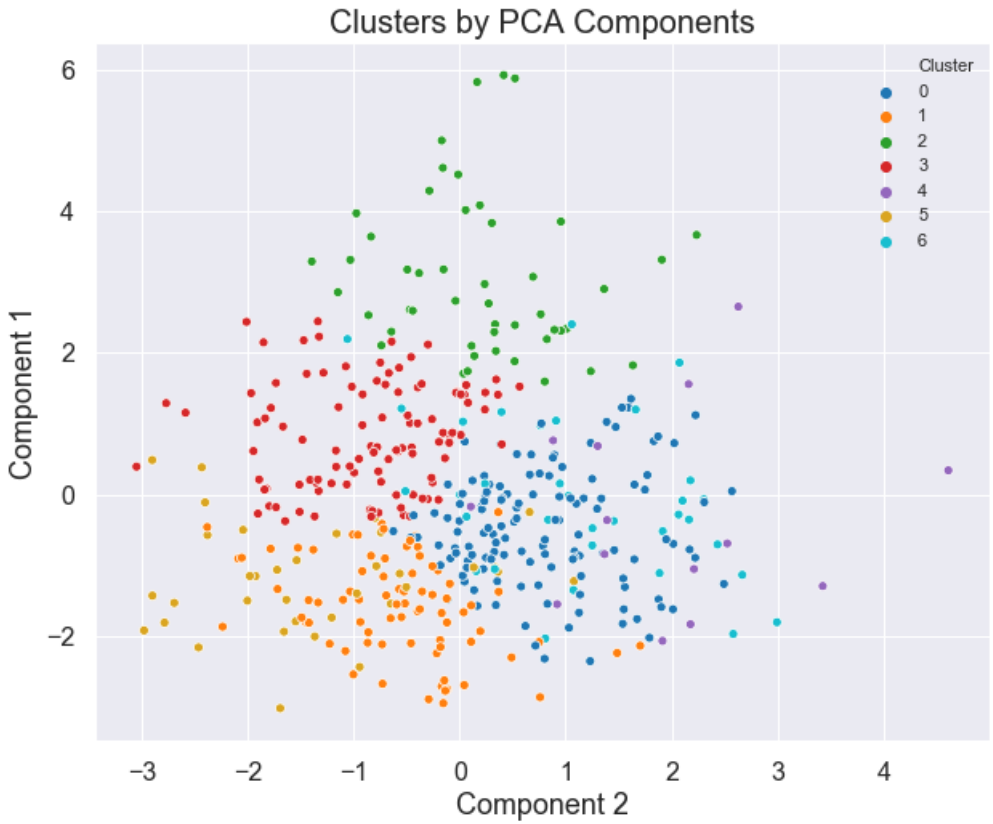
- A well-known combination, stemming from the work of Ding et al. (2004) «***K*-means clustering via principal component analysis**».

- We can first reduce the dimensionality of the data, obtaining in this way a feature space where the clustering result can be understandable (remember orthogonal dimensions!!!).

PCA + K-Means

- We want to create a sequential process which:
  1. Pre-processes the data.
  2. Applies PCA to the pre-processed data (Nr. Of components?).
  3. Applies K-Means clustering to the reduced data (K?).
  4. Displays the results on the reduced data.
  5. Displays a general per cluster analysis on the real data.

PCA + K-Means


Clusters by PCA Components
Cluster
0
1
2
3
4
5
6
Component 1
6
4
2
0
−2
Component 2
−3
−2
−1
0
1
2
3
4

Sklearn links

- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html#sklearn.preprocessing.MinMaxScaler
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

Exercises