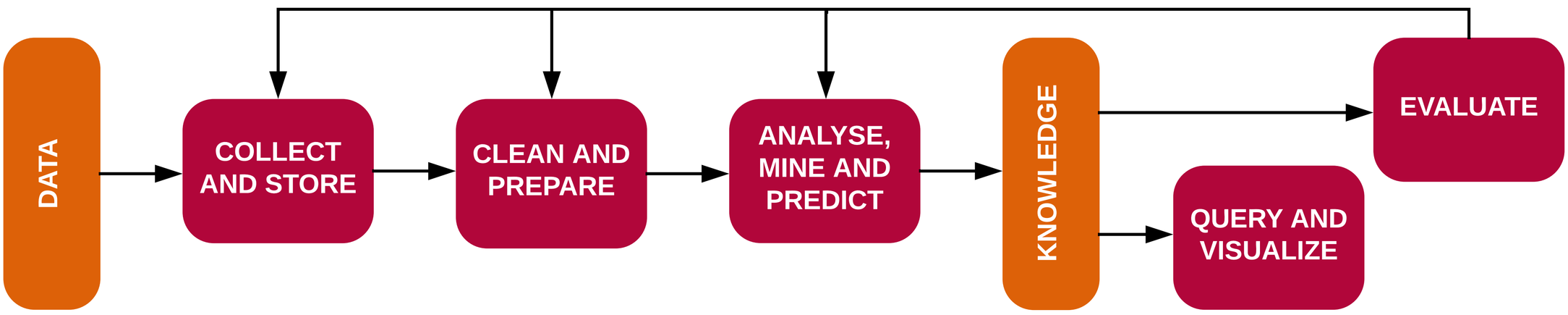
Data Engineering in der Praxis
An Overview

Abstract
This poster shows an overview of technologies and puts them in context of the data engineering process. The overview focuses mainly on batch-processing technologies and does not show the stream-processing technologies introduced in the lectures.

The process

In the diagram below you can see the coarse process of data engineering. Many of the lectures* showed similar processes to this one or specialized in parts of it.



Technologies

Below the process different tools and techniques are shown that were introduced in some of the lectures*. Their ordering represents in what part of the process these tools and techniques are typically useful but also in which order you would typically use them (e.g. oozie is not cleaning or preparing data, but it is used before the next step of the process). Of course, this overview of technologies is not complete. Furthermore, each step of the process blends together with the other steps and each technology has a wide variety of use cases and could be used in other parts of the process as well.

- outlier filtering
- data enrichment
- duplicate detection
- m-fold cross-validation
- AB-Tests
- capture feedback
- precision and recall
- root-mean squared error

Michael Vaichenker
Bachelor

* Eating news from the web - lambda architecture for scalable text analytics in a news stream
Data mining & predictive maintenance for energy efficient coal mining
SAP Leonardo Machine Learning - Turning insights into action
Amazon Redshift Spectrum
Data engineering for job recommendations
Data engineering in the newsroom