COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Data Collection

Dr SHI Lei

Durham University

learning lab

Problem Framing

Data Preparation

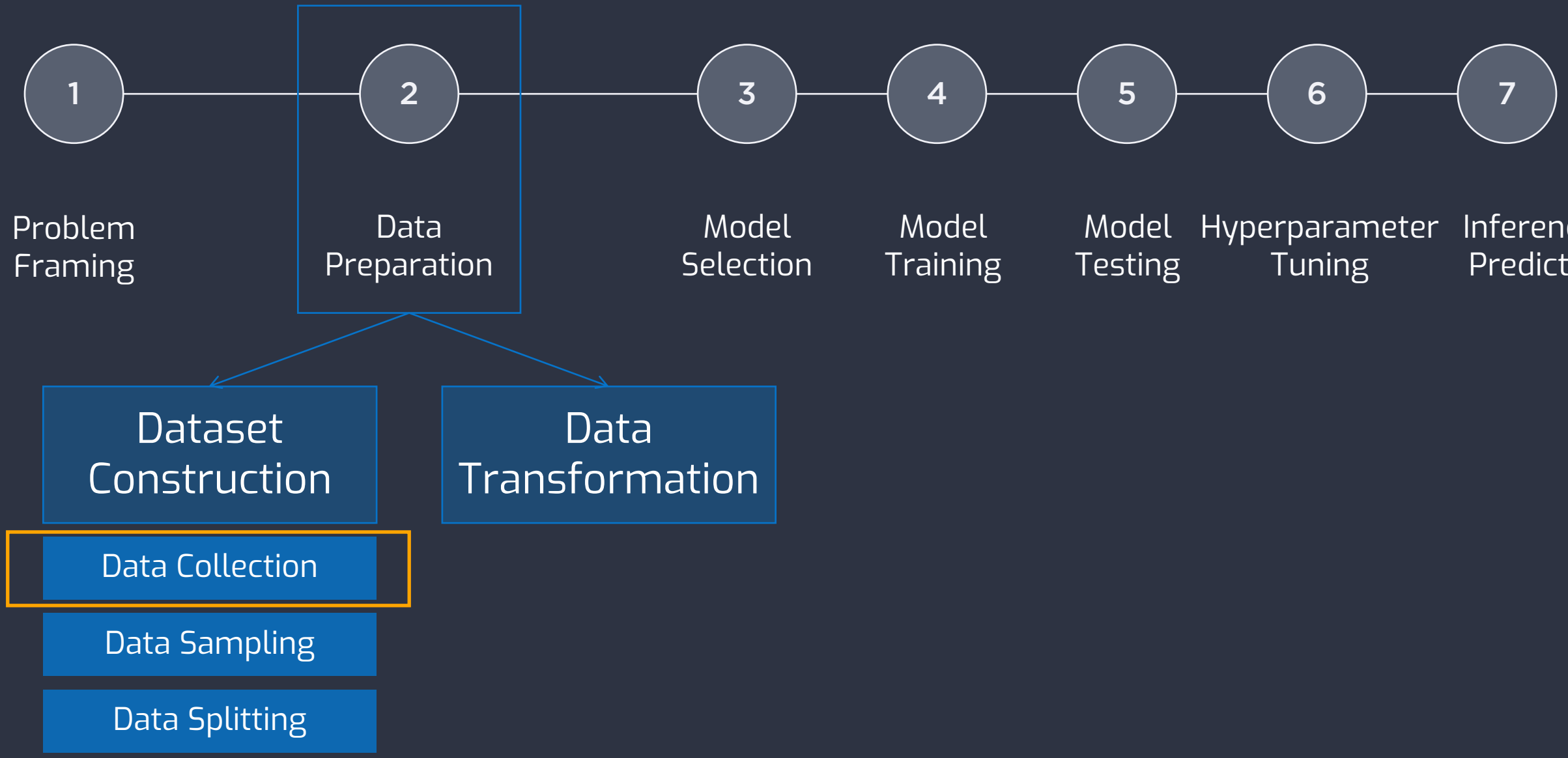
Model Selection

Model Training

Model Testing

Hyperparameter Tuning

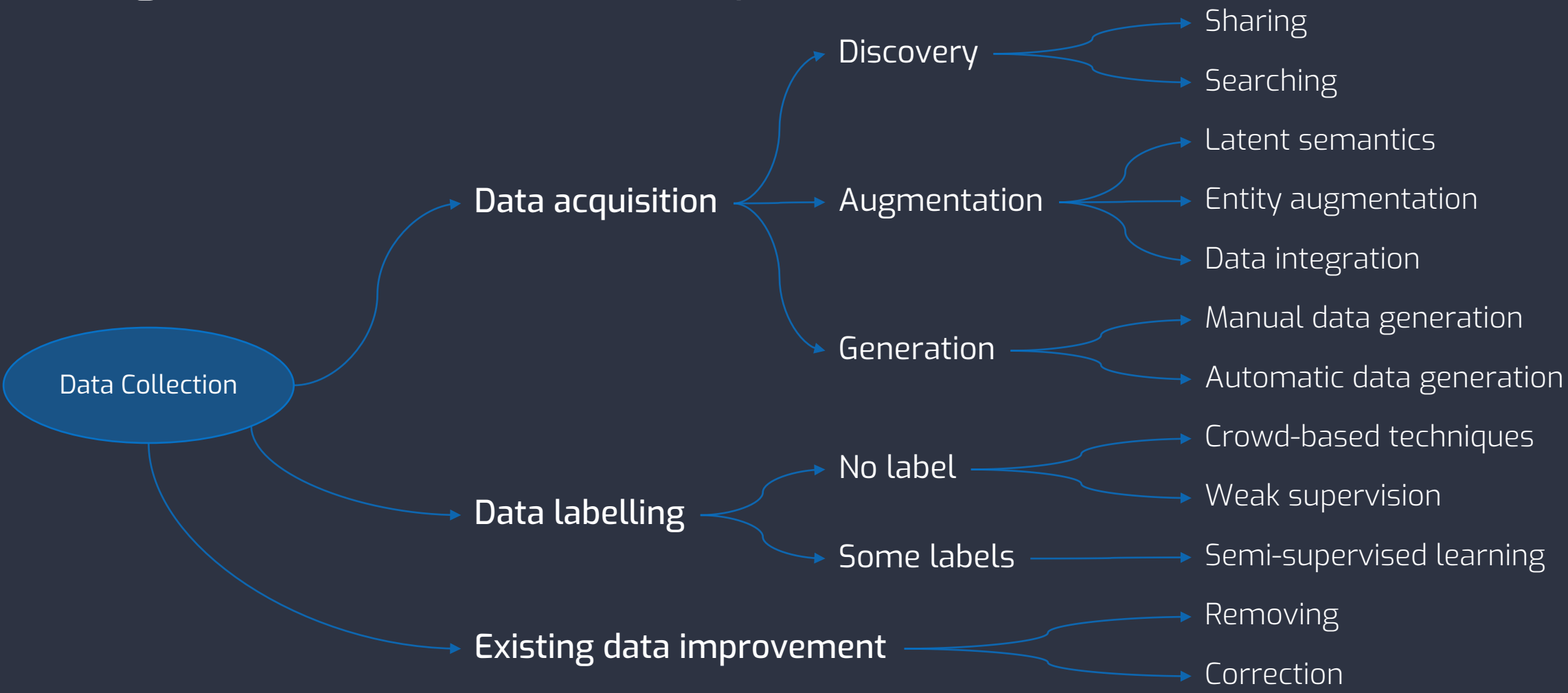
Inference / Prediction


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Transformation
Data Collection
Data Sampling
Data Splitting

Learning Objectives

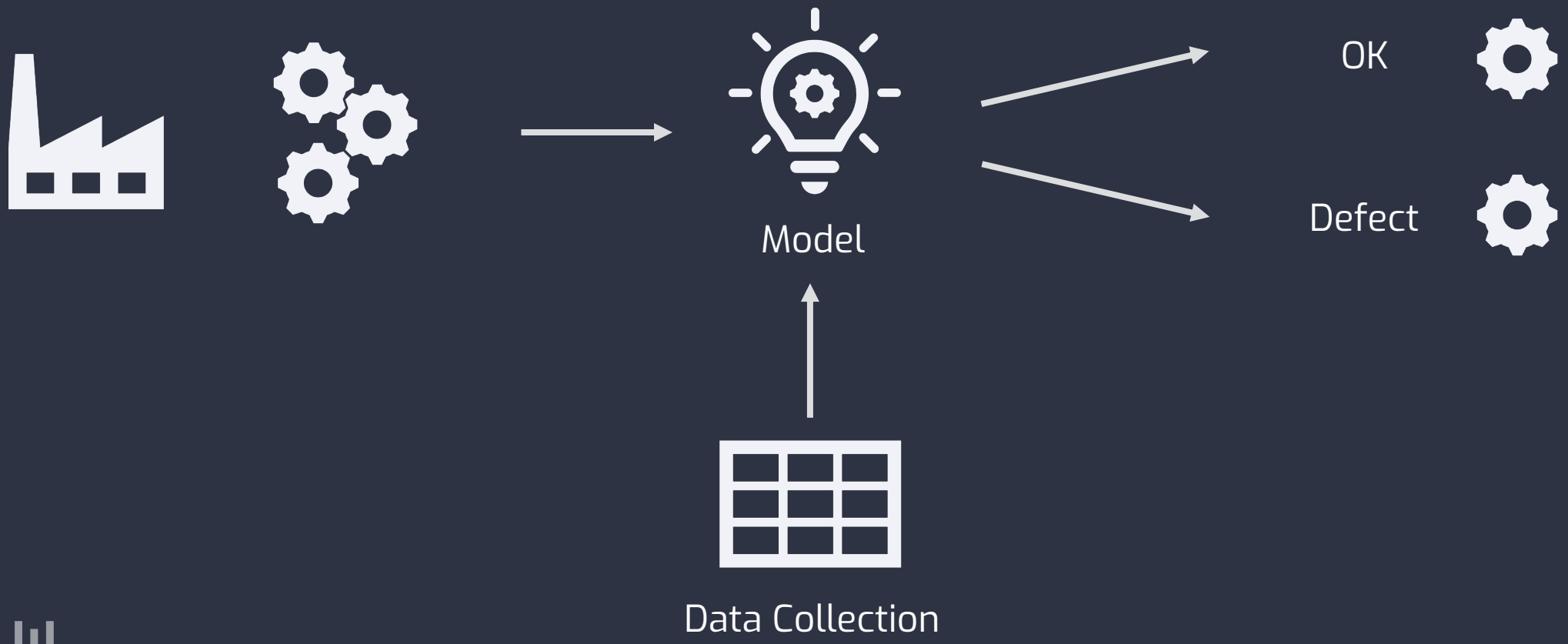
- Understand the process of Data Collection
- Understand different techniques for Data Collection

A high level research landscape of data collection for machine learning

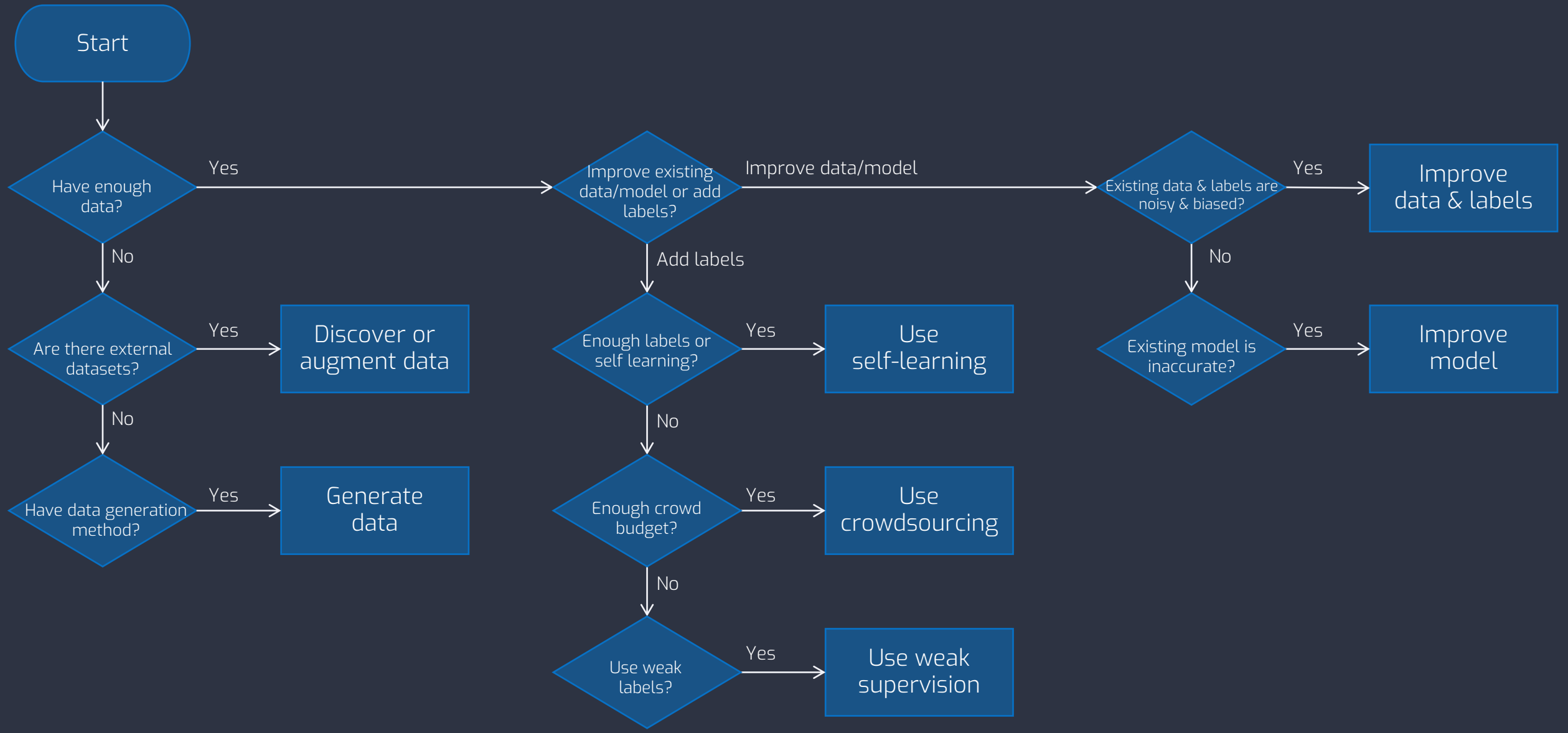


Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

EXAMPLE. Faulty Detection in Smart Factory

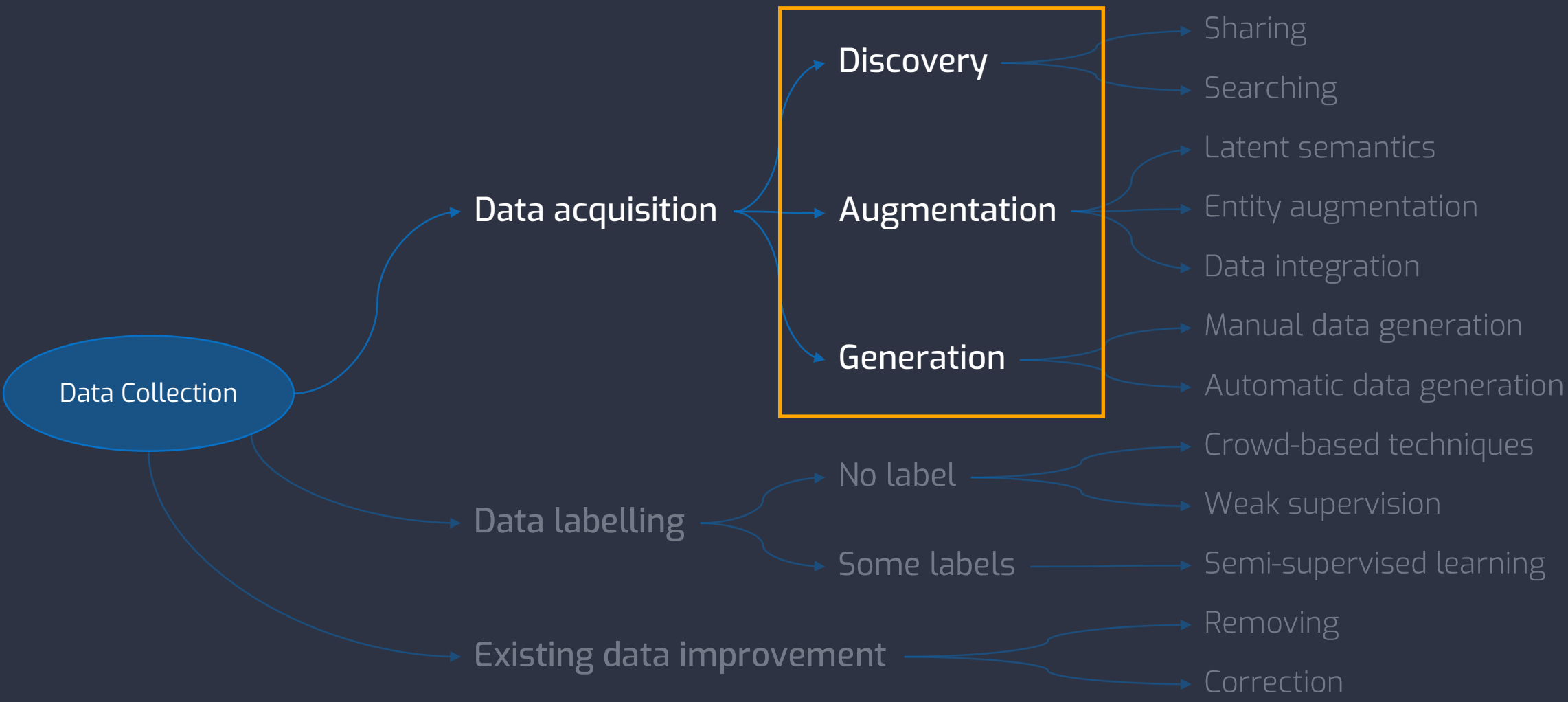


A decision flow chart for Data Collection

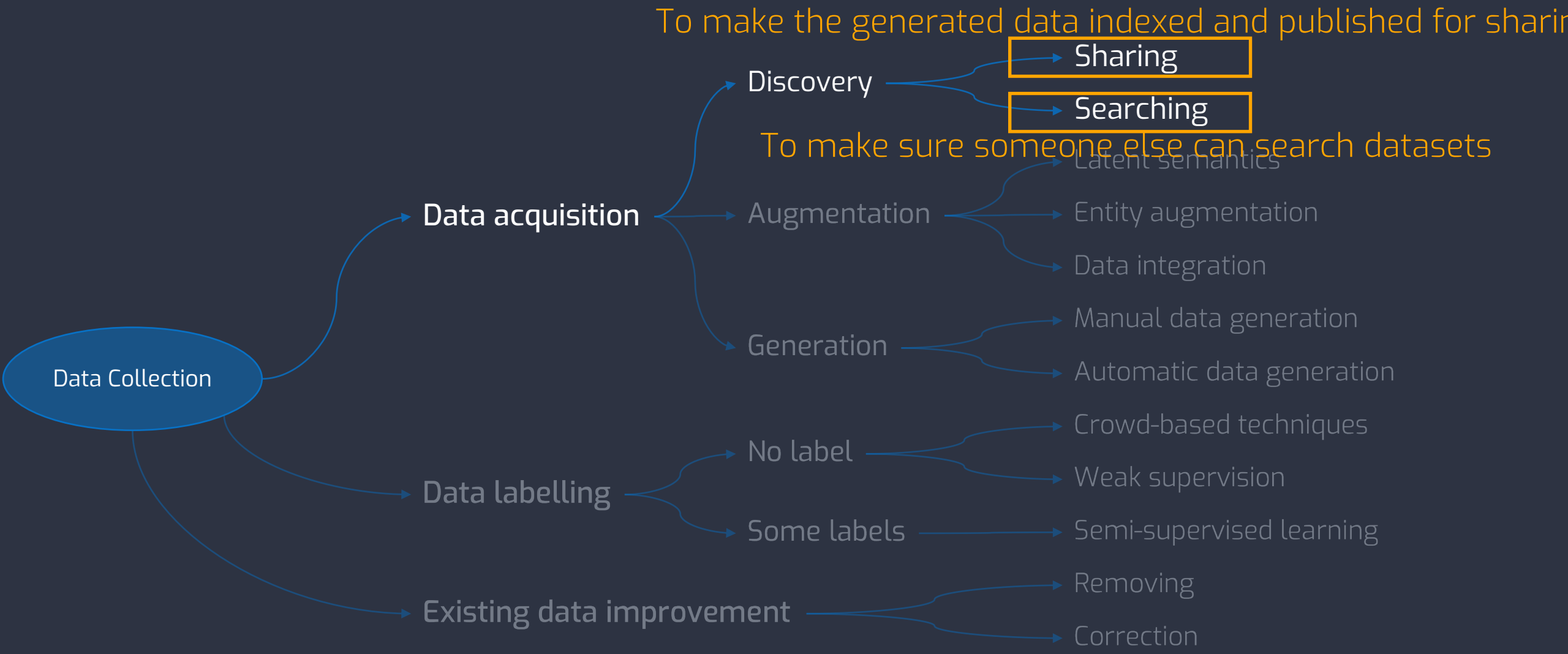


Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

To find the dataset that machine learning model can be trained on.



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Data Sharing / Data Searching

Challenge	many systems are not built with the intention of sharing datasets.
	use a post-hoc approach where metadata is generated after the datasets are created without the help of the dataset owners.

Collaborative Systems	are designed to make data sharing easy.

Collaborative Analysis

collaboratively analysing different versions of datasets using version control systems; an individual or team can perform machine learning tasks on their own version of a dataset and later merge with other versions.

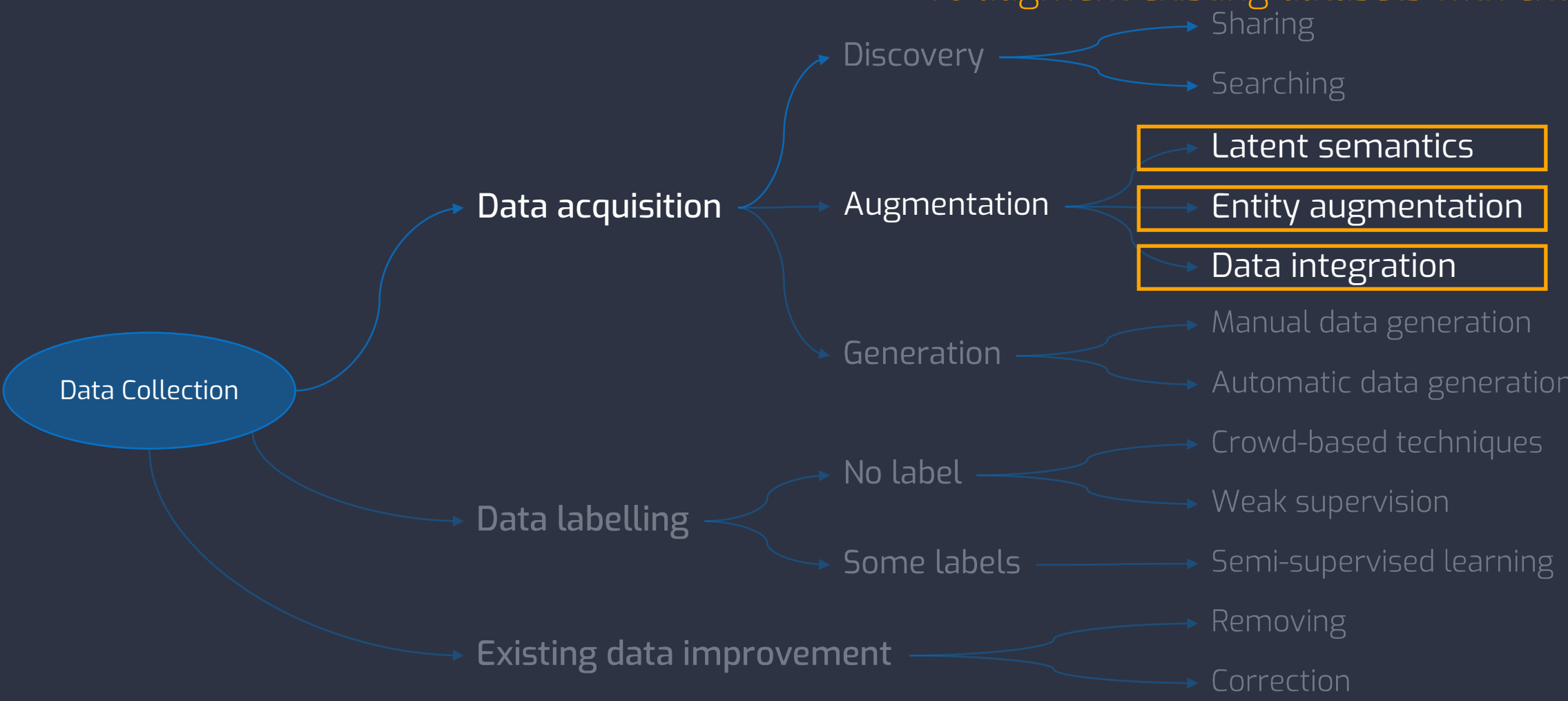
Internet

others can access to them via Web search. Some datasets are on the Internet and free of use with certain copy rights, and others are in the so-called data marketplaces where anyone can buy and sell their own datasets.

Data Sharing / Data Searching

Challenge	lots of datasets generated internally – difficult for others to discover.
	Data Lake systems, a method to search datasets so that others don't have to make redundant efforts to re-generate the datasets for their own ML projects.
Google Data Search	Catalogues metadata of datasets from various storage systems.
Data Civilizer	Builds linkage graph of data.

To augment existing datasets with external data.



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Latent semantics

- can be used to extract, represent and generate the contextual usage meaning of words or text, so that the "original" data is augmented in order to be better used in machine learning.

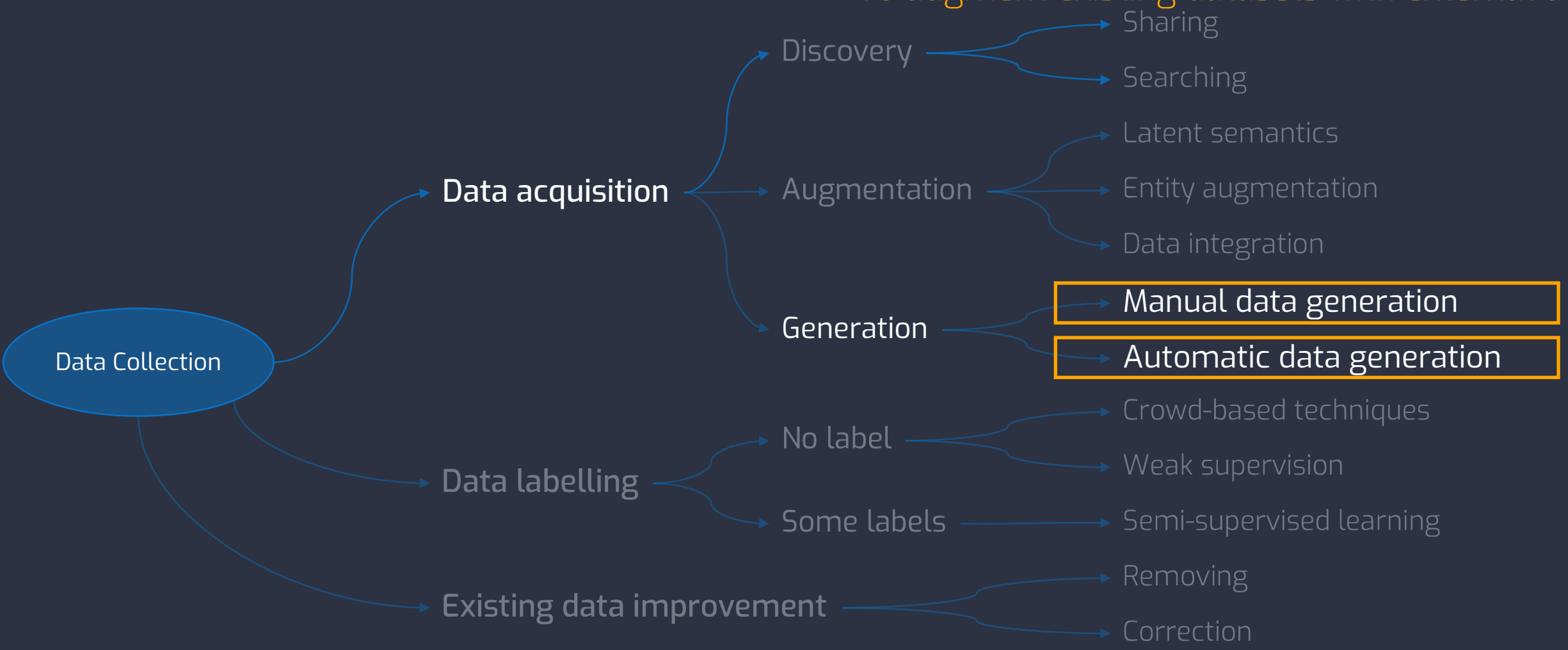
Entity augmentation

- To fille in the dataset by gathering more information. For example, we could fill in missing values of attributes in some or all of the entities by matching multiple tables using schema matching.

Data integration

- is particularly used when we want to extend existing dataset with other acquired ones. For example, we could join multiple tables using foreign keys, in order to have a single file, i.e. a single dataset to train on.

To augment existing datasets with external data.



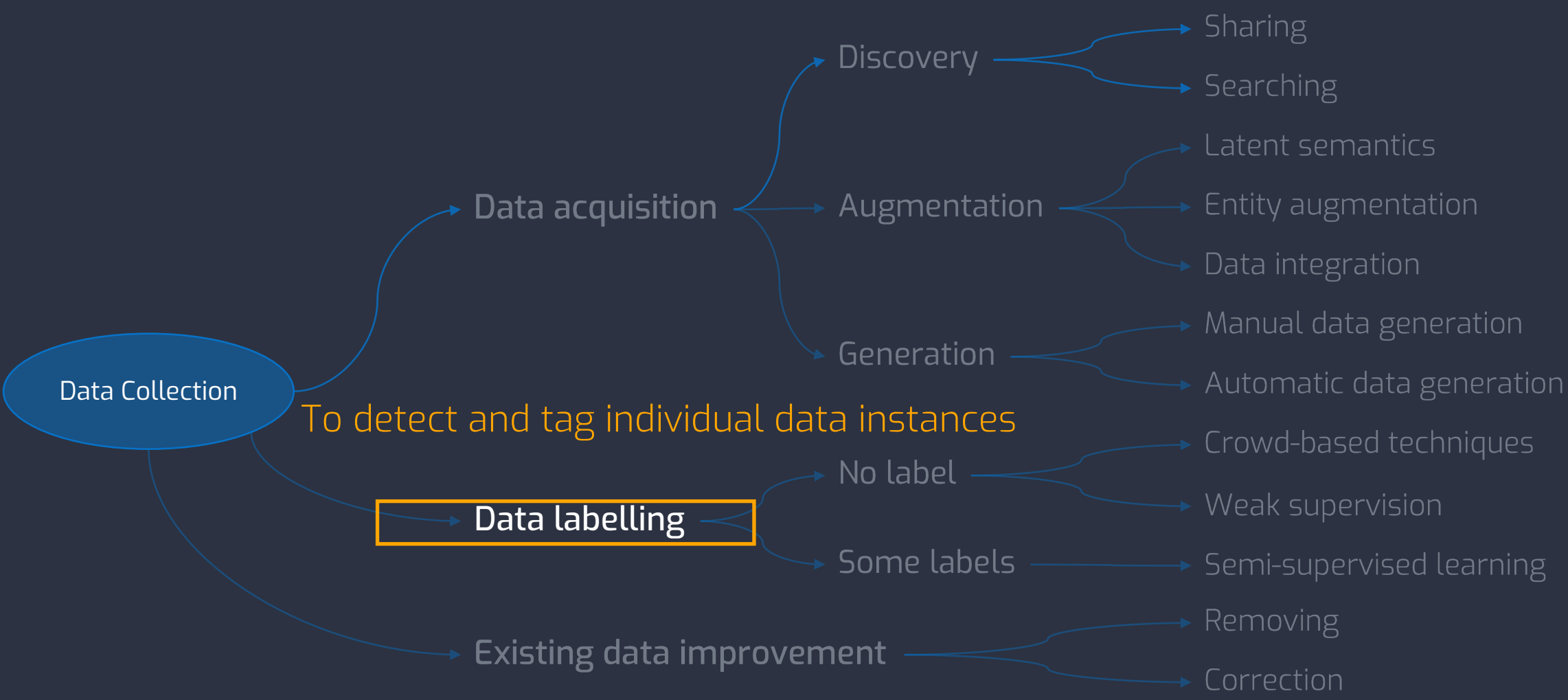
Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Manual data generation

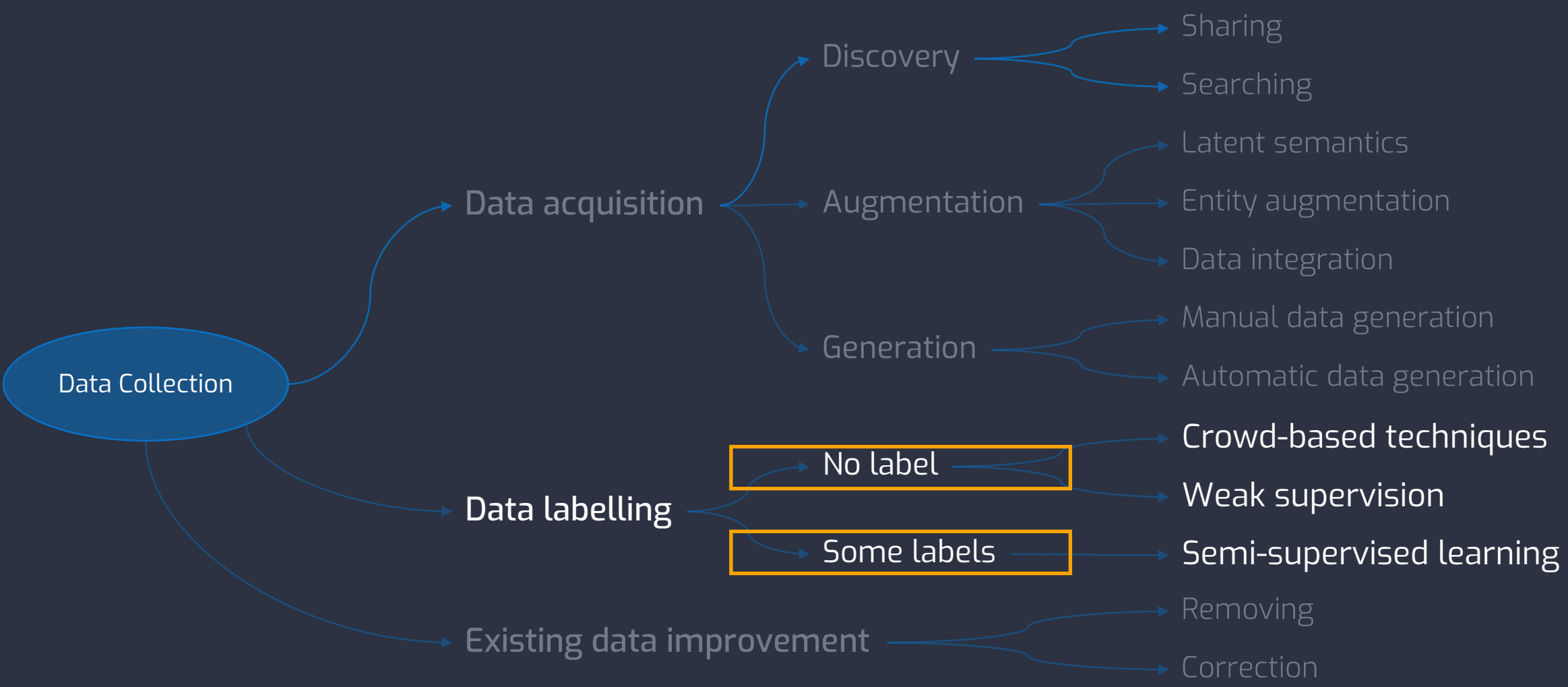
- Crowdsourcing is usually used in this scenario.
- Amazon Mechanical Turk, for business (known as Requesters) to hire remotely located "crowdworkers" to perform discrete on-demand tasks that computers are currently unable to do.
- Employers post Human Intelligence Tasks (HITs). "Crowdworkers" (known as Turkers) search and perform existing tasks to get compensations.

Automatic data generation

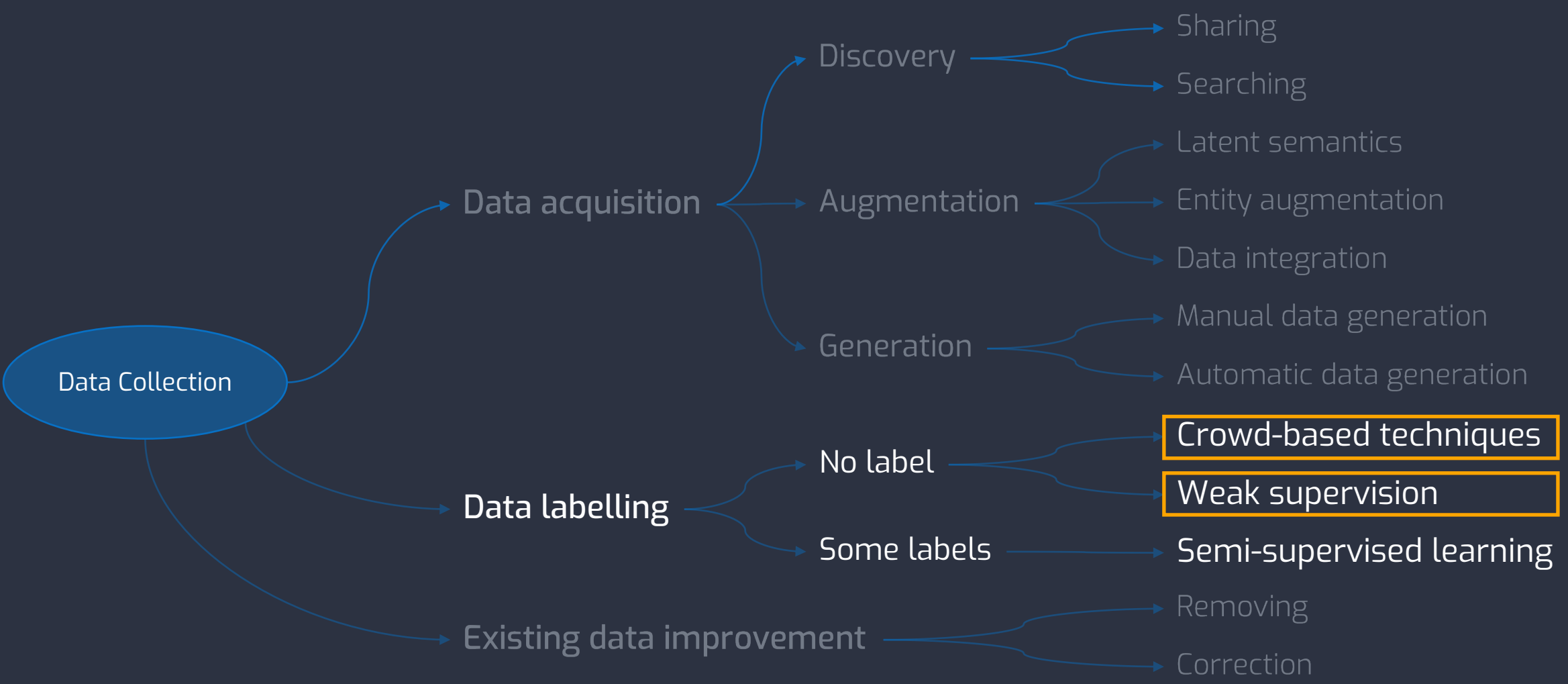
- To generate a repository of data synthetically.
- Benefits from low cost and flexibility.
- A simple method is to use probability distribution and generate a sample from that distribution.
- We could use the methods provided by scikit-learn, which create random regression, classification, or clustering problem dataset with added noise.



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Crowd-based techniques / Weak supervision

Active learning

- Focuses only on the most “interesting” instances.
- Assigned to expert labelling workers.
- Could bring bias to the instances and so the learning algorithm cannot be reused.
- Techniques: uncertainty sampling, query-by-committee, decision theoretic approaches…

Crowdsourcing

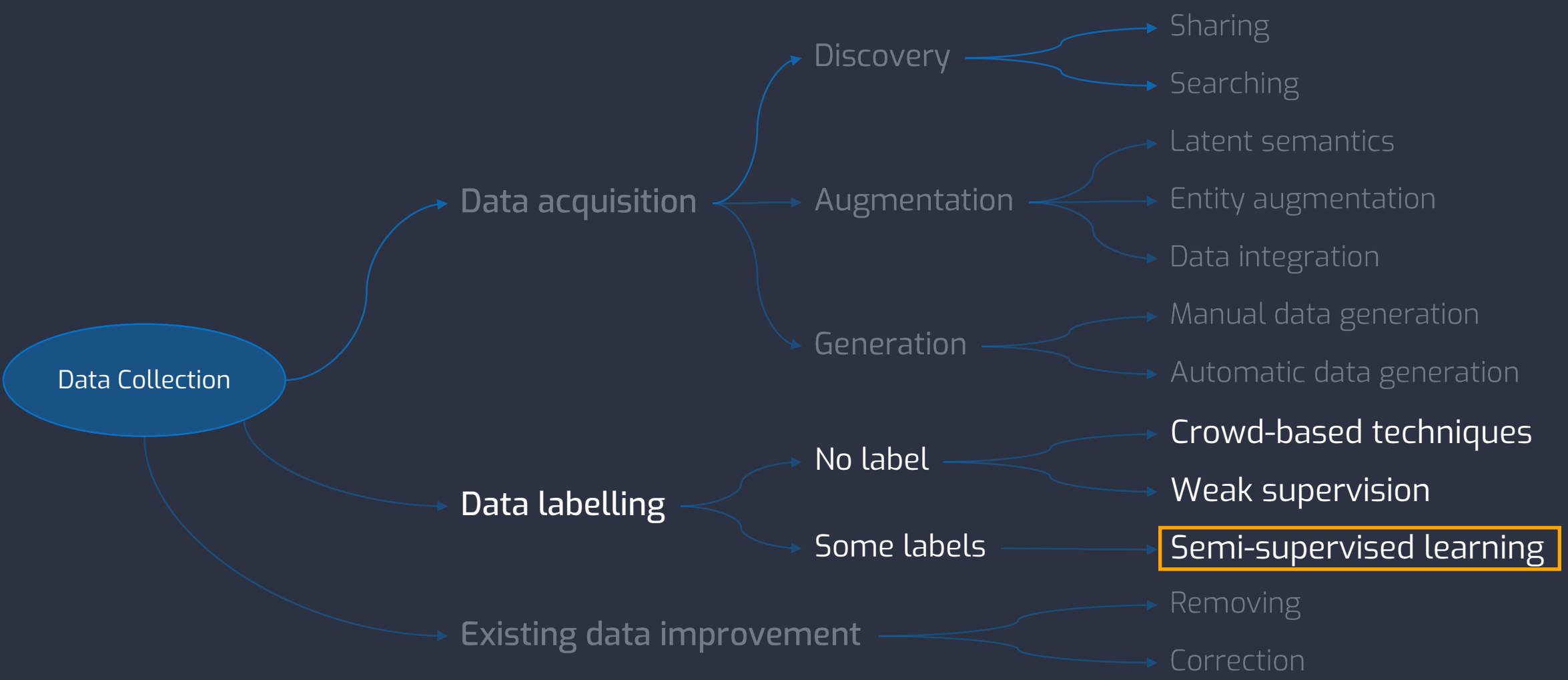
- Focuses on assigning task to many works.
- Not necessarily to be assigned to experts.
- Interaction between workers could reduce mistakes and biases.
- Techniques: user interaction, quality control, scalability, regression…

Data Programming

- Can be used to generate large amounts of labels.
- The basic idea is to express weak supervision strategies or domain heuristics as user-defined labelling functions that collectively generate a large but potentially overlapping set of labels.
- Can increase accuracy and usability of the generated labels.

Fact Extraction

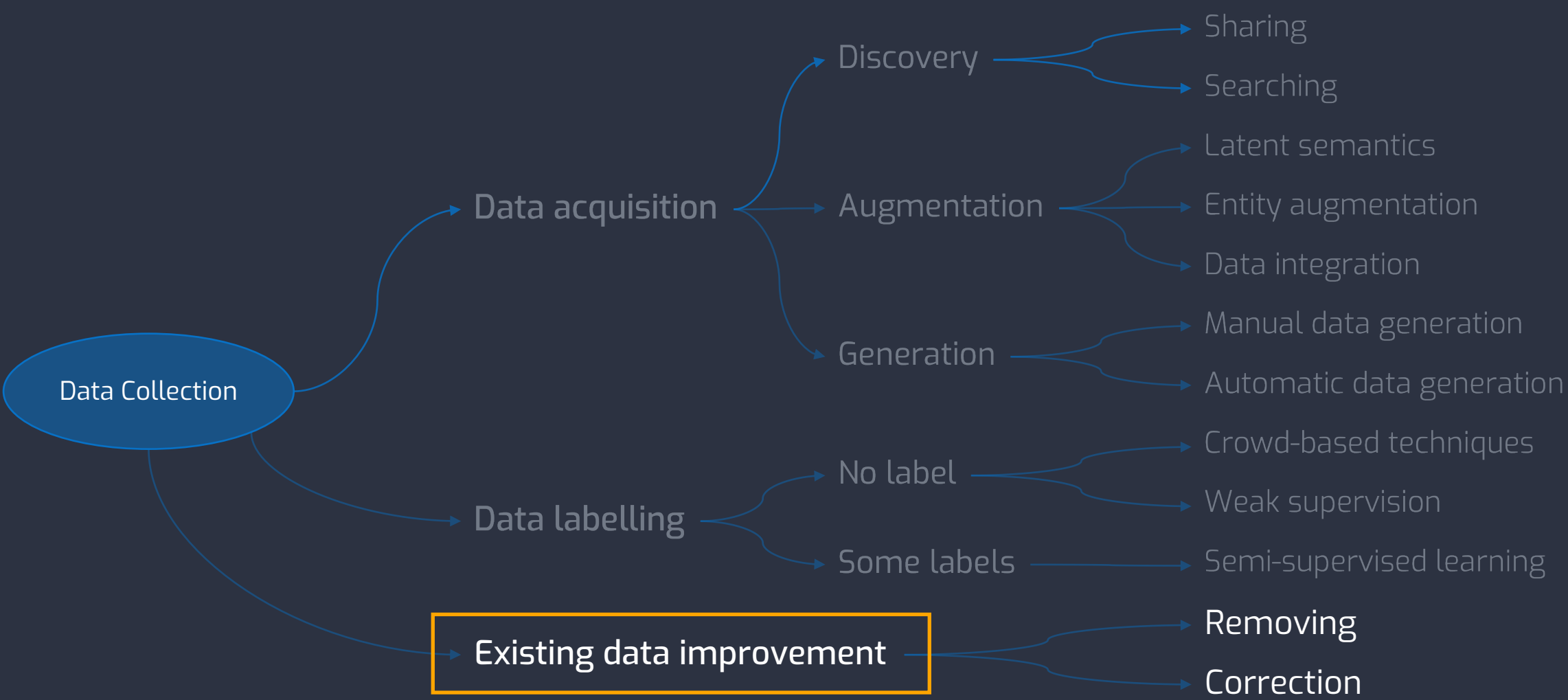
- Facts contained in knowledge bases are extracted from various sources including the Internet.
- Facts can be considered as positively labelled instances and can be used as seed labels for distant supervision when generating weak labels.
- Fact extraction roots from the broader topic of information extraction which aims to extract structured data from the Internet.



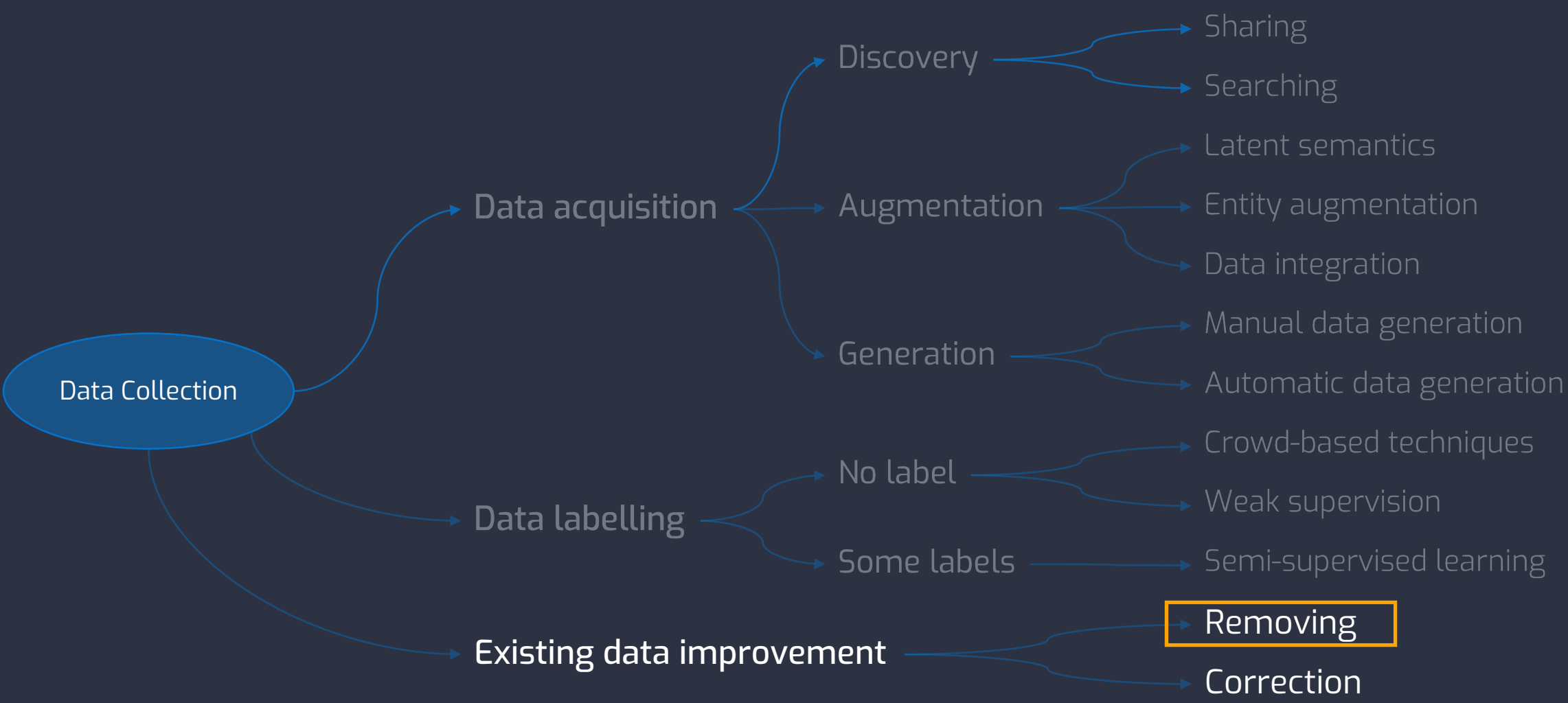
Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Semi-supervised learning

- Incorporates supervised learning and unsupervised learning to label large amount of data with only a small labelled dataset.
- The idea is to use supervised learning models trained on the small labelled dataset to predict labels for unlabelled data.
- Could be time and cost efficiency, because we only need a small amount of manually labelled training data.
- However, the accuracy strictly depends upon that small labelled dataset. If the quality of that small labelled dataset is not good, then the generated labels may also be less accurate.



To improving existing data instead of seeking new data



Roh, et al. A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective

Removing / Correction

Unwanted instances

- "Unwanted" includes "redundant" and "irrelevant".
- **Redundant** instances usually appear during data collection, e.g., combining datasets, scraping data.
- **Irrelevant** instances are what are not useful for a specific task, e.g. when training a model to predict house prices in England, those instances from Scotland are not useful thus should be removed.

Outliers

- Outliers may cause issues, especially for those models that are sensitive to outliers e.g. linear regression models.
- We must only remove outliers if they are "proven guilty". We should not remove an outlier just because it is "such a small number", because it could be informative for our model.

Removing / Correction

Structural errors

- Need to deal with typos or inconsistent English spelling (British vs American English), inconsistent capitalisation, or abbreviation.
- This is extremely important especially in case of categorical features, e.g., “ML” and “Machine Learning” should be a single category.

Missing data

- We may decide to drop the instances with missing values, because most machine learning algorithms do not accept missing values.
- Need to be careful when removing missing data – we cannot just simply drop them or impute the missing values based on other instances.

Takeaway Points

- For data collection, there are three options to consider: data acquisition, data labelling and improving existing data.
- When developing a system, it is important to think about how we want our data to be managed for analysis.
- New data can come from augmentation or through automatic/manual generation.
- Sometimes improving existing data could be considered over seeking new datasets.