# Data Collection

——Background——

"Data Mining: Methods and Applications"

1

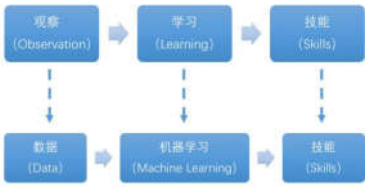
## Contents


- **Background**
- **Data Acquisition**
- **Data Labeling**
- **Improvement of Existing Data and Models**
- **How to Decide which Data Collection Techniques to Use When**
- **Interesting Future Research Challenge**


2

## Background

- **The reasons for data collection**
  - New applications(data mining, machine learning) do not necessarily have enough labeled data.
    - Traditional Machine Translation and Object Detection: massive amounts of data
    - New Applications: manual labeling(expensive and domain expertise )
  - Unlike traditional machine learning, deep learning techniques automatically generate features, which saves feature engineering costs, but in return may require larger amounts of labeled data.

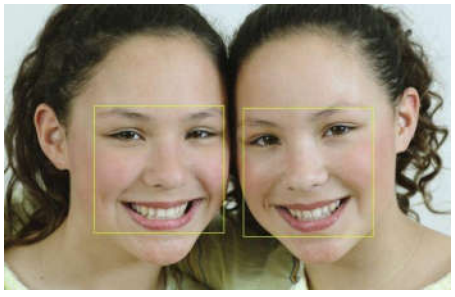




3

## Background

- **Related Applications**
  - Machine Learning(ML), Natural Language Processing(NLP) and Computer Vision(CV)
    - End to end (端到端) machine learning applications: collecting, cleaning, analyzing, visualizing, and feature engineering
  - Data Management



4

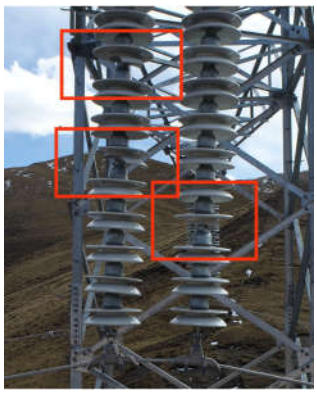
## Background

- **There is a pressing need of accurate and scalable data collection techniques in the era of Big data.**
  - Share and search new datasets: data acquisition techniques can be used to discover, augment, or generate datasets.
  - Once the datasets are available, various data labeling techniques can be used to label the individual examples.
  - Instead of labeling new datasets, it may be better to improve existing data or train on top of trained models.

5

## Background

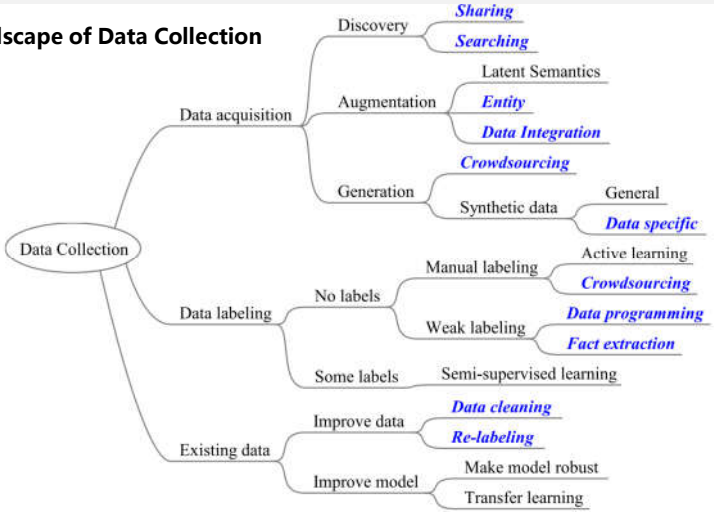
- **Labeling Data**



6

## Background

- **Research Landscape of Data Collection**



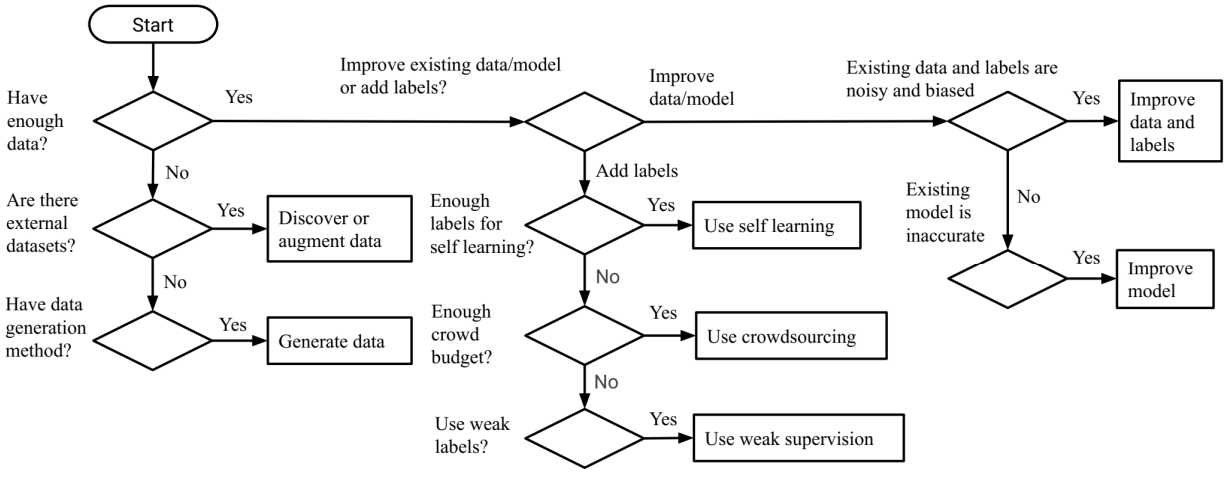

**Research Landscape**:

The topics that are at least partially contributed by the data management community are highlighted using blue italic text.

7

## Background

- **A general decision flow chart of the data collection techniques**



## Background

- **Data Acquisition :** data discovery, data augmentation and data generation
- **Data Labeling :** utilizing existing labels, using crowdsourcing techniques, and using weak supervision
- **Improving Existing Data or Models**
- **Method Integration :** Put all techniques together

9

## Contents

- **Background**
- **Data Acquisition**
- **Data Labeling**
- **Improvement of Existing Data and Models**
- **How to Decide which Data Collection Techniques to Use When**
- **Interesting Future Research Challenge**

10

# Thanks !