4V

Volume Scale of data. Amount of data.

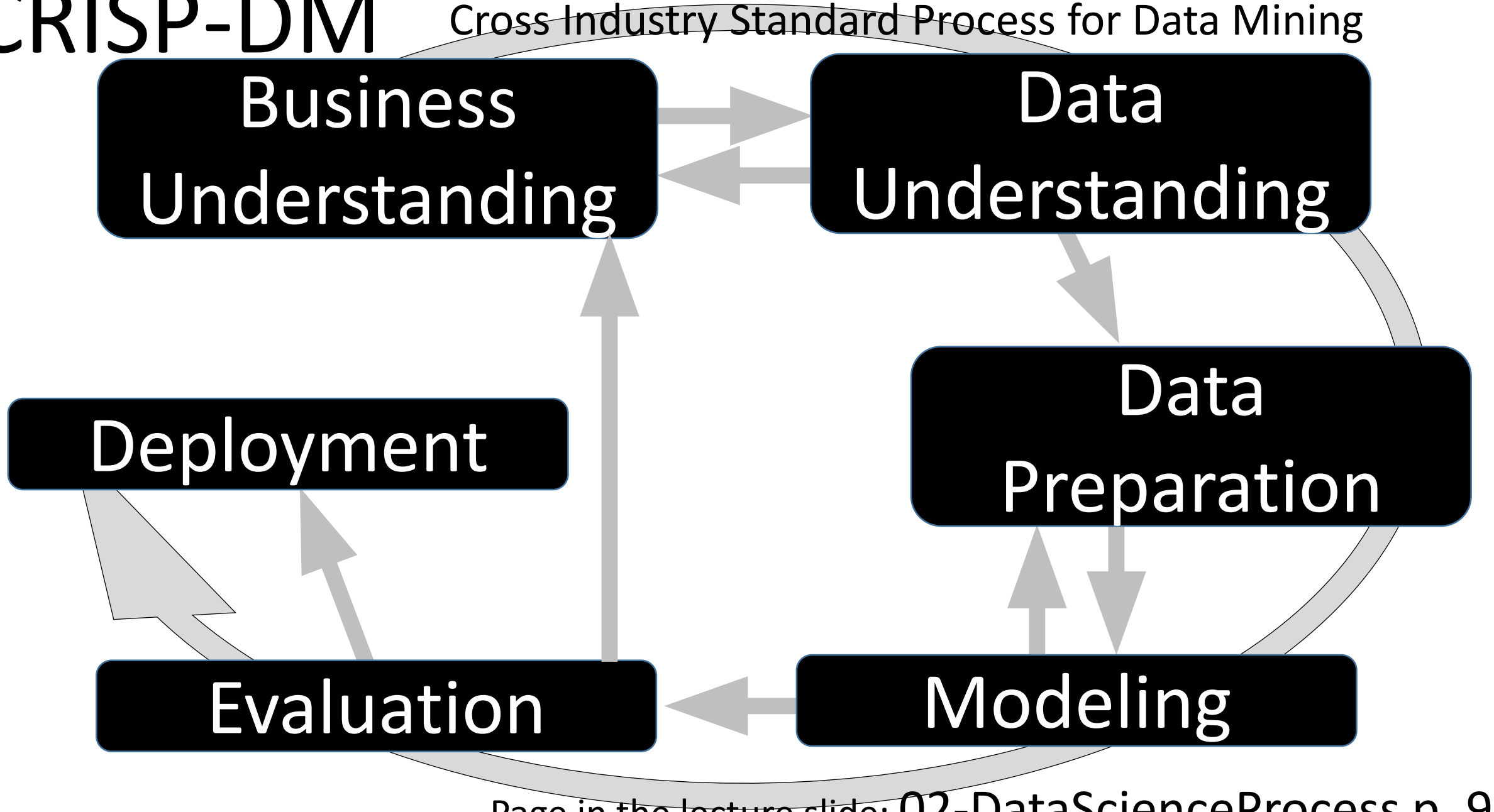
Velocity Speed of data.
Speed of processing and generation of the data.

Variety Diversity of data.
Number of types of data.

Veracity Certainty of data.

CRISP-DM

Cross Industry Standard Process for Data Mining



Data Preparation

Z-Score

Z-Score(X) = (X – mean(X)) / SD(X)

SD(X): Standard Deviation of the field values

Outliers

Extreme values that lie near the limits of the data or go against the trend. Identifying them is important since they might be errors in data entry.

Because the value is identified as the extreme value that stay away from the other numbers. The value “###” is too small/big compared to the other data.

Characteristics of SAN

Storage Attached Network

- Provides direct access from multiple computers at the block level.
- Access Control and Translation from file-level to block-level operations must take place on the client node.

Example Veritas Cluster File System and DataPlow Nasan File System

Characteristics of NAS

Network Attached Storage

- Fault tolerance and high availability by data replication of one sort or another
- Fast disk-access time and small amount of CPU-processing time over distributed structure

Example Veritas Cluster File System and DataPlow Nasan File System

NAS vs. SAN

NAS

- Shared storage over **shared** network
- File System
- Easier management

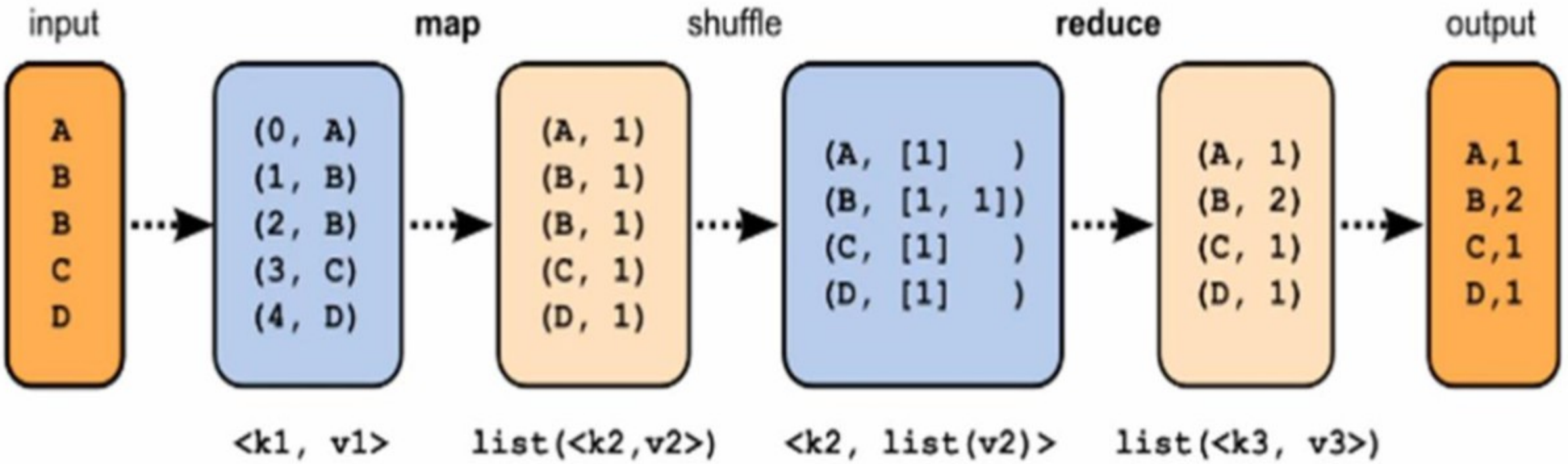
SAN

- Shared storage over **dedicated** network
- Raw storage
- Fast, but costly

Design goal of distributed file system

- Access transparency
- Location transparency
- Concurrency transparency
- Failure transparency
- Replication transparency
- Migration transparency
- Scalability
- Heterogeneity

Map Reduce


input
map
shuffle
reduce
output
A
B
B
C
D
(0, A)
(1, B)
(2, B)
(3, C)
(4, D)
(A, 1)
(B, 1)
(B, 1)
(C, 1)
(D, 1)
(A, [1])
(B, [1, 1])
(C, [1])
(D, [1])
(A, 1)
(B, 2)
(C, 1)
(D, 1)
A,1
B,2
C,1
D,1
<k1, v1>
list(<k2,v2>)
<k2, list(v2)>
list(<k3, v3>)

Shuffle process

1. Split Creation
2. Map
3. Spill
4. Merge
5. Copy
6. Sort
7. Reduce

Word count program (1) // Mapperの実装

```
public static class Map extends
        Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritableone = new IntWritable(1);
        private Text word = new Text();
        @Overrideprotected
        void map(LongWritablekey, Text value, Context context)
                throws IOException, InterruptedException{
                String line = value.toString();
                StringTokenizertokenizer = new StringTokenizer(line);
                while (tokenizer.hasMoreTokens()) {
                        word.set(tokenizer.nextToken());
                        context.write(word, one);
                }
        }
```

Word count program (2) // Reducerの実装

```
public static class Reduce extends
        Reducer<Text, IntWritable, Text, IntWritable> {
        private IntWritable value = new IntWritable(0);
        @Overrideprotected void reduce(Text key,
                Iterable<IntWritable> values, Context context)
                throws IOException, InterruptedException {
                int sum = 0;
                for (IntWritable value : values)
                        sum += value.get();
                value.set(sum);
                context.write(key, value);
        }
}
```

Motivation of Apache Spark

- Difficulty of programming directly in Hadoop MapReduce
- Performance bottlenecks, or batch not fitting use cases
- Better support iterative jobs typical for machine learning

Two Main Abstractions

- RDD – Resilient Distributed Dataset
- DAG – Direct Acyclic Graph

RDD

Resilient Distributed Dataset

Collection of data items split into partitions and stored in memory on worker nodes of the cluster

RDD is the main and only tool for data manipulation in spark

- Transformations
- Actions

Driver

Spark Cluster

- Entry point of the Spark Shell (Scala, Python, R)
- The place where SparkContext is created
- Translates RDD into the execution graph
- Splits graph into stages
- Schedules task and controls their execution
- Stores metadata about all the RDDs and their partitions
- Brings up Spark WebUI with job informaiton

Executor

Spark Cluster

- Stores the data in cache in JVM heap or on HDDs
- Reads data from external sources
- Writes data to external sources
- Performs all the data processing

Two winters

- In 1959, there was receptive fields of single neurons in the cat's striate cortex appeared.
- In 1962, there was receptive fields, binocular interaction and functional architecture in the cat's visual cortex.

Problem issue and solution

We understand the text based on the previous words + the current word, but Neural network cannot process in such way.

Recurrent Neural Network(RNN) can process a sequence of data by applying a recurrence formula in the every steps.

Data Mining Definition

Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data Mining – Difference to AI

The difference is in its purpose.

Data mining: to discover the unknown characteristics of the data.

Machine Learning: to predict something by learning the known training data.

Machine Learning:

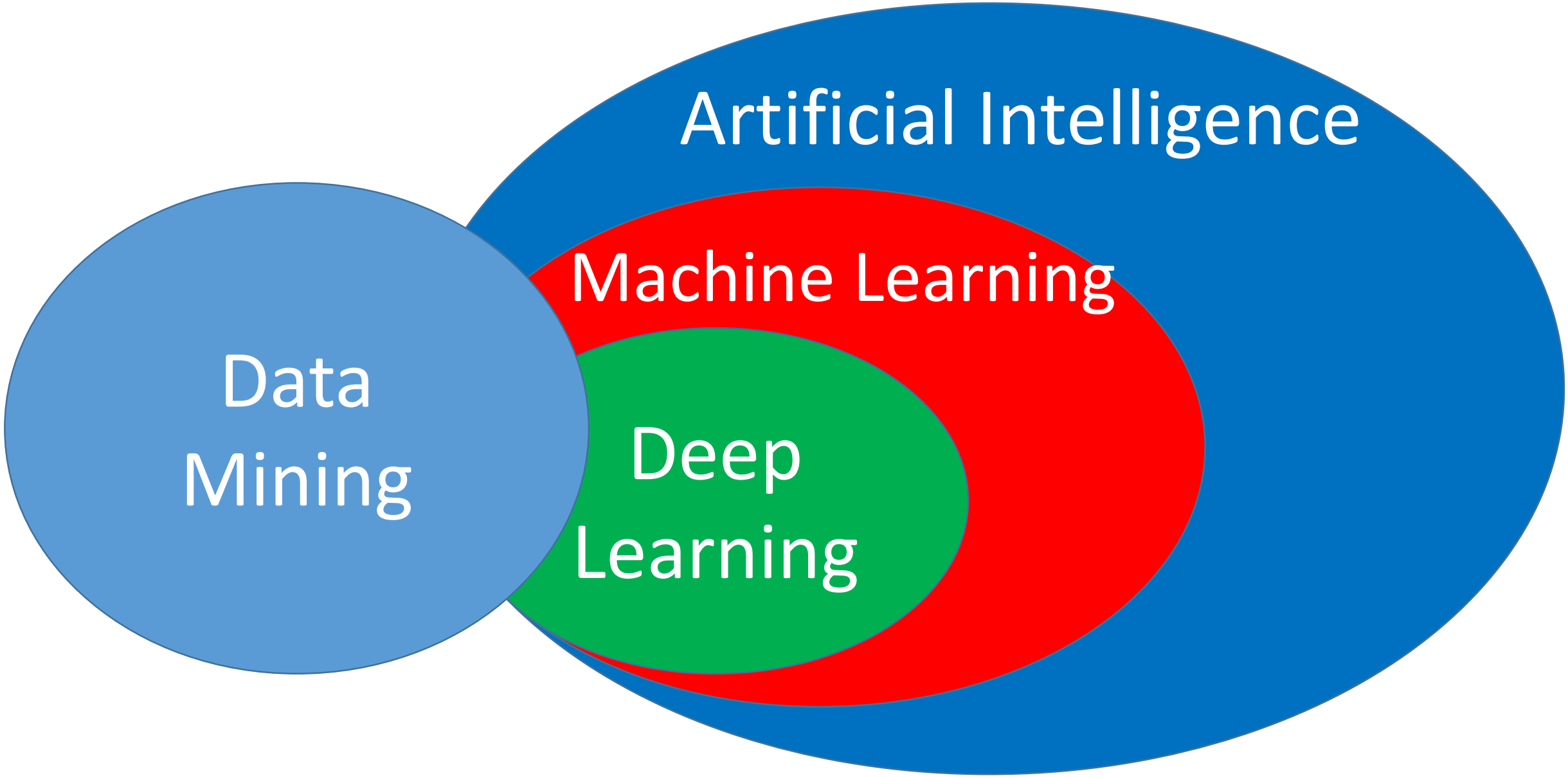
Evaluated by whether it can regenerate the known knowledges

Data Mining:

Important is whether it can discover the unknown knowledges

Data Mining is using the method of machine learning in some parts.

Machine learning is also using the data mining method for unsupervised learning.


Artificial Intelligence
Machine Learning
Data Mining
Deep Learning

Predictive and Descriptive

Descriptive

Identifies what happened in the past by analyzing stored data

- Tracking assignment and assessment **grades**
- Comparing **pre-test and post test** assessments
- Analyzing course **completion rates** by learner or by course

Predictive

Describes what can happen in the future by analyzing past data

- **Credit score** when we apply to the credit card used in finalcial service.
- Analysis of market price in **FX/Share**, etc.
- Prediction of the **sales of this year** based on the sales in last year.

Classification and Clustering

Classification

Classify the data into one of numerous already defined definite classes.

Involved in **Supervised learning**

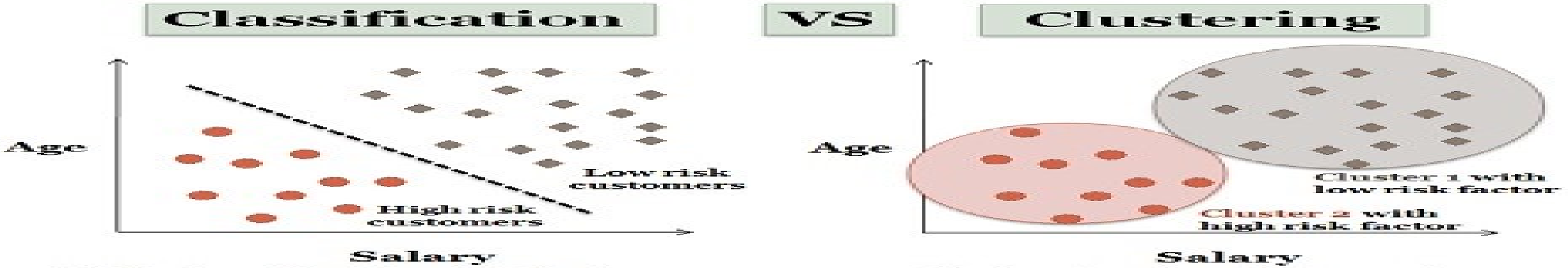
Provided training sample

Clustering

Organize a group of data into classes and clusters where the items has simil.

Involved in **Unupervised learning**

NOT Provided training sample



Risk classification for the loan payees on the basis of customer salary

3 Methods of classification

Gini Index (10-1 p. 34)

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

Entropy (INFO) (10-1 p. 42)

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

Misclassification error

$$Error(t) = 1 - \max_i P(i \mid t)$$