Data Mining &

Knowledge Discovery

(Knowledge Acquisition,

Machine Learning)

Extraction of hidden,

previously unknown,

and potentially useful

high-level information

from low-level data

Machine Learning

Michalski (1986): constructing or modifying representations of what is being experienced

Simon (1983): denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time

Minsky (1986): making useful changes in the workings of our minds

[	Size	Ink-color	Body-color	Attitude	]
	small	blue	blue	positive	
	large	blue	black	negative	
	large	red	blue	positive	
	small	black	blue	negative	
	large	black	blue	negative	
	large	blue	blue	positive	
	large	black	black	negative	
	small	blue	black	negative	

2, 2, 2

(Ink-color, blue) & (Body-color, blue) -> (Attitude, positive)

1, 1, 1

(Ink-color, red) -> (Attitude, positive)

1, 3, 3

(Body-color, black) -> (Attitude, negative)

1, 3, 3

(Ink-color, black) -> (Attitude, negative)

[Working_capital/Total_assets
Retained_earnings/Total_assets
Earnings_before_interest_and_taxes/Total_assets
Market_value_equity/Book_value_of_total_debt
Sales/Total_assets
Prediction]

Working_capital/Total_assets	Retained_earnings/Total_assets	Earnings_before_interest_and_taxes/Total_assets	Market_value_equity/Book_value_of_total_debt	Sales/Total_assets	Prediction
36.7	-62.8	-89.5	54.1	1.7	bankruptcy
24.0	3.3	-3.5	20.9	1.1	bankruptcy
-61.6	-120.8	-103.2	24.7	2.5	bankruptcy
-1.0	-18.1	-28.8	36.2	1.1	bankruptcy
18.9	-3.8	-50.6	26.4	0.9	bankruptcy
-57.2	-61.2	-56.2	11.0	1.7	bankruptcy
3.0	-20.3	-17.4	8.0	1.0	bankruptcy
-5.1	-194.5	-25.8	6.5	0.5	bankruptcy
17.9	20.8	-4.3	22.6	1.0	bankruptcy
5.4	-106.1	-22.9	23.8	1.5	bankruptcy
23.0	-39.4	-35.7	69.1	1.2	bankruptcy
-67.6	-164.1	-17.7	8.7	1.3	bankruptcy
-185.1	-308.9	-65.8	35.7	0.8	bankruptcy
13.5	7.2	-22.6	96.1	2.0	bankruptcy
-5.7	-118.3	-34.2	21.7	1.5	bankruptcy
72.4	-185.9	-280.0	12.5	6.7	bankruptcy
17.0	-34.6	-19.4	35.5	3.4	bankruptcy
-31.2	-27.9	6.3	7.0	1.3	bankruptcy
14.1	-48.2	6.8	16.6	1.6	bankruptcy
-60.6	-49.2	-17.2	7.2	0.3	bankruptcy
26.2	-19.2	-36.7	90.4	0.8	bankruptcy
7.0	-18.1	-6.5	16.5	0.9	bankruptcy
-53.1	-98.0	-20.8	26.6	1.7	bankruptcy
-17.2	-129.0	-14.2	267.9	1.3	bankruptcy
32.7	-4.0	-15.8	177.4	2.1	bankruptcy
26.7	-8.7	-36.3	32.5	2.8	bankruptcy
-7.7	-59.2	-12.8	21.3	2.1	bankruptcy
18.0	-13.1	-17.6	14.6	0.9	bankruptcy
2.0	-38.0	1.6	7.7	1.2	bankruptcy
-35.3	-57.9	0.7	13.7	0.8	bankruptcy
5.1	-8.8	-9.1	100.9	0.9	bankruptcy
0.0	-64.7	-4.0	0.7	0.1	bankruptcy
25.2	-11.4	4.8	7.0	0.9	bankruptcy
35.2	43.0	16.4	99.1	1.3	survival
38.8	47.0	16.0	126.5	1.9	survival
14.0	-3.3	4.0	91.7	2.7	survival

2, 29, 29
(Retained_earnings/Total_assets, -308.9..8.5) &
(Sales/Total_assets, 0.1..2.7) -> (Prediction,
bankruptcy)

1, 26, 26
(Market_value_equity/Book_value_of_total_debt,
0.7..53.4) -> (Prediction, bankruptcy)

2, 32, 32
(Market_value_equity/Book_value_of_total_debt,
53.4..771.7) & (Retained_earnings
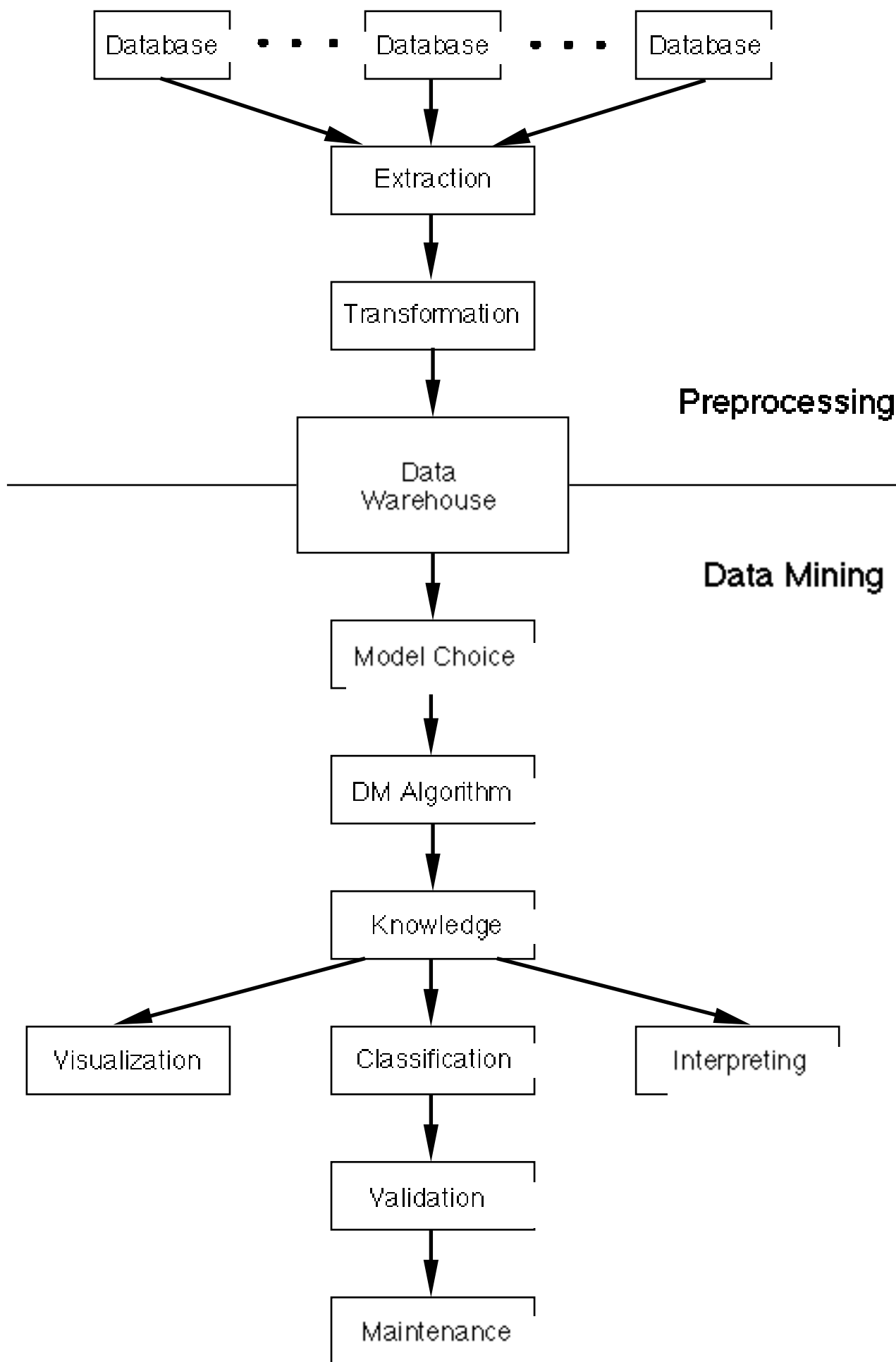/Total_assets, 8.5..68.6) -> (Prediction, survival)

2, 4, 4
(Sales/Total_assets, 2.7..6.7) &
(Market_value_equity/Book_value_of_total_debt,
53.4..771.7) -> (Prediction, survival)

Statistics	Data Mining
• Mathematical rigor	• Adventurous attitude
• Sampling	• Entire population
• Model	• Predictive power
• Computation is secondary	• Algorithms are the most important
• Slow	• Quick
• Lack of explanation	• Explanation in rules, trees,...

```
┌──────────┐
│   Raw    │
│ training │
│   data   │
└────┬─────┘
     │
     ▼
 ╭─────────────╮
(  Preprocessing )
 ╰──────┬──────╯              ┌──────────┐
        │                     │   Raw    │
        ▼                     │  unseen  │
   ┌──────────┐               │   data   │
   │ Refined  │               └────┬─────┘
   │ training │                    │
   │   data   │                    │
   └────┬─────┘                    ▼
        │                    ╭─────────────╮
        ▼                   (  Preprocessing )
  ╭──────────╮               ╰──────┬──────╯
 (  Induction  )                    │
  ╰─────┬─────╯                     ▼
        │                      ┌──────────┐
        ▼                      │ Refined  │
   ┌──────────┐                │  unseen  │
   │Knowledge │                │   data   │
   └────┬─────┘                └────┬─────┘
        │                           │
        └────────┐         ┌────────┘
                 ▼         ▼
            ╭───────────────────╮
           (    Classification    )
            ╰─────────┬─────────╯
                      │
                      ▼
                ┌──────────┐
                │Classified│
                │  unseen  │
                │   data   │
                └──────────┘
```
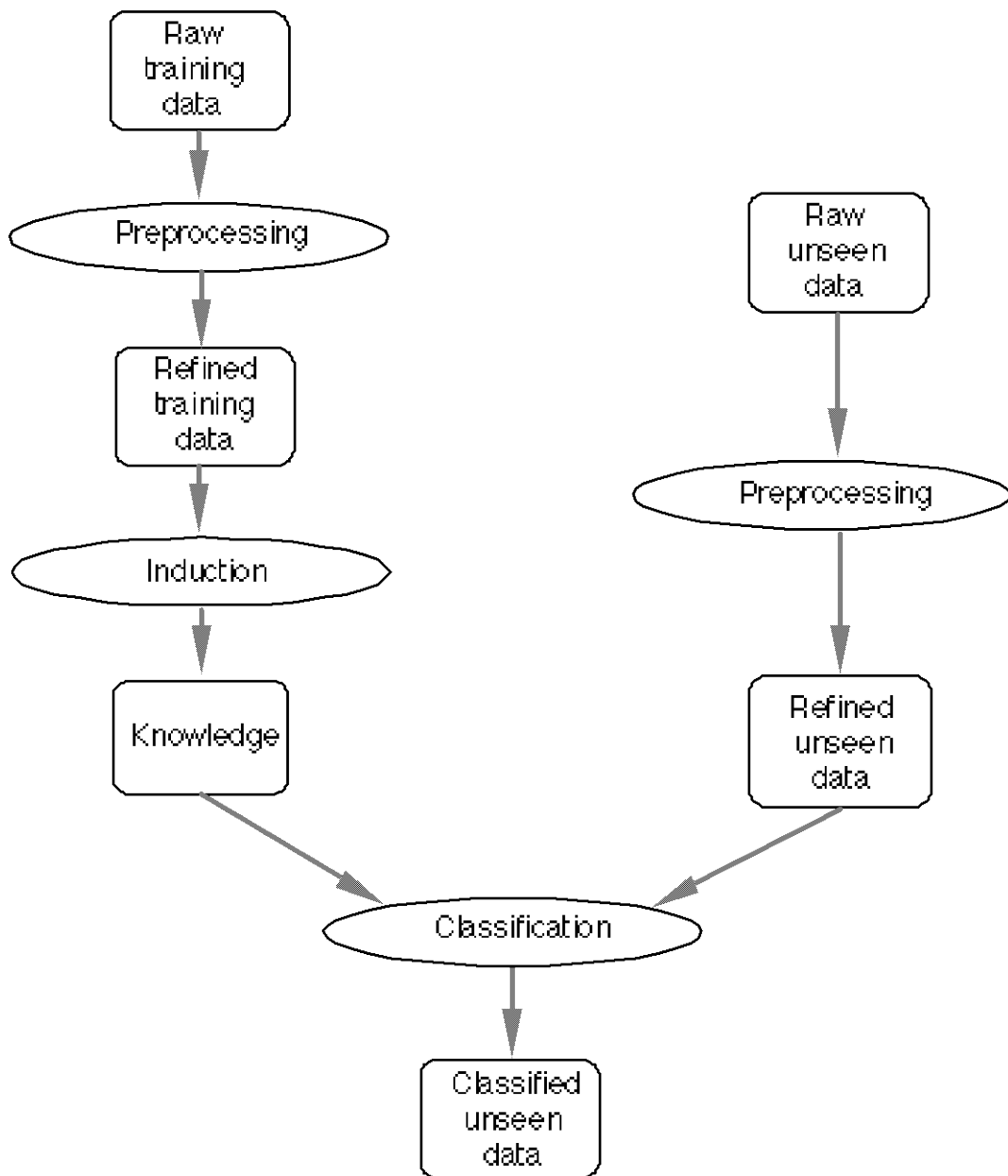
Type of knowledge acquired

production rules

decision trees

taxonomies

parameters in algebraic expressions

formal grammars

graphs

logic expressions

programs

Data Mining

Statistical Methods

Machine Learning

Case-Based Reasoning

Text Mining

Time Series

Machine Learning

Similarity Based Learning (Empirical Learning),

Explanation-Based Learning

Computational Learning Theory

Genetic Algorithms

Neural Nets

Machine Learning

Incremental way of learning

Constructive Induction

Ensembles of classifiers

Large Training Sets

Imbalanced Data Sets