DATA MINING AND KNOWLEDGE DISCOVERY

2019
3/13

Dr. Jiawei Chang

WHY DATA MINING

- With the explosive growth of data
  - Web, transactions, sensor networks, bioinformatics, news, digital cameras, …
- Raw data are not so meaningful…
- **We are drowning in data, but starving for knowledge!**

WHAT IS DATA MINING

- Knowledge Discovery from Data

- Extraction of previously unknown and potentially useful patterns from huge amount of data

WHAT KINDS OF DATA

- Various Data
  - Texts
  - Photos
  - Videos
  - Time-series data

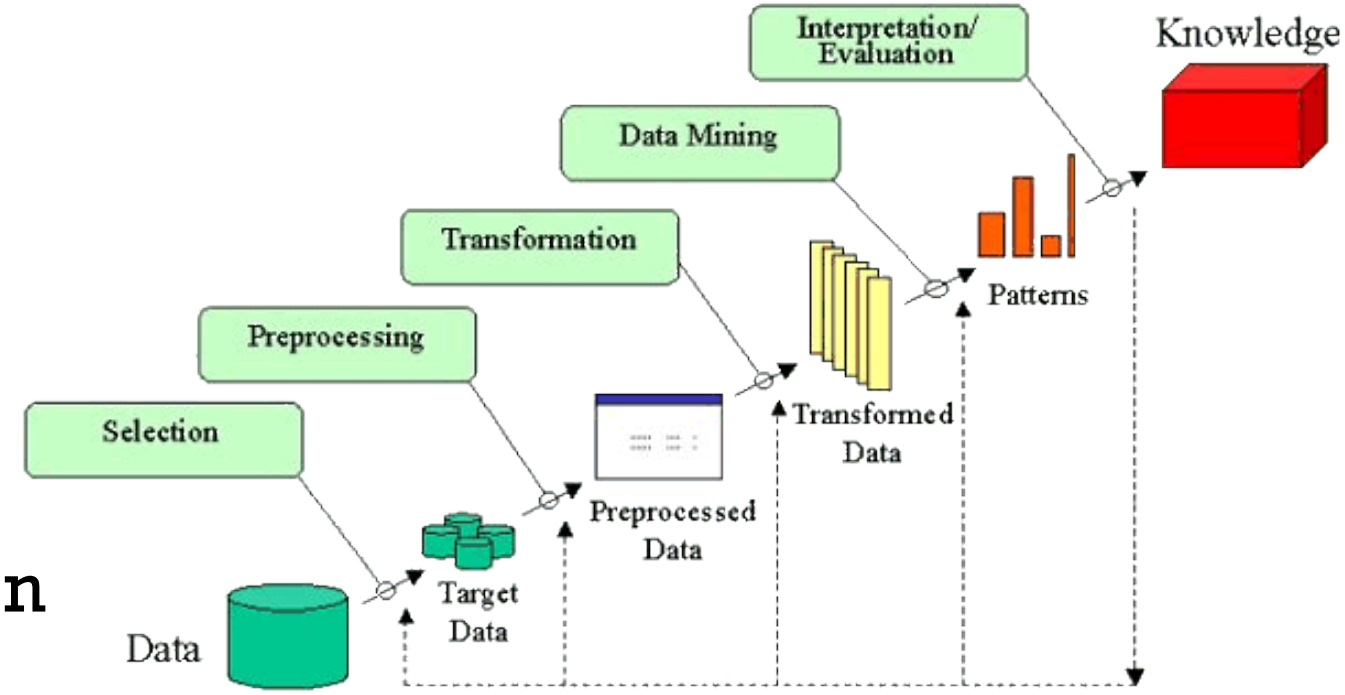
- The Preprocessing Methods to Different Data Types
  - Text: remove stop-words, count co-occurrence
  - Image: resize images, remove noise, do segmentation
  - IoT: sensor graph, spatial or temporal data

APPLICATIONS OF DATA MINING

- Marketing
- Recommender systems
- Web page analysis
- Biological and medical data analysis
- Internet of things
- Other dedicated knowledge discovery

KNOWLEDGE DISCOVERY PROCESS

1. Source Selection
2. Data Preprocessing
3. Data Transformation
4. Data Mining
5. Evaluation and Interpretation



SOURCE SELECTION

- **Source selection** is the process of finding related or useful data to target problem or application.
- **Data collection** is the process of gathering information on targeted variables in a systematic fashion.

DATA PREPROCESSING

- **Data cleaning** is the process of identifying incomplete or incorrect parts of the data and then replacing, modifying, or deleting the dirty data.
- **Data integration** involves combining different data sources. Data integration appears with increasing frequency as the volume (that is, big data) and the need to share existing data explodes.

VIEWTABLE: Work.Sample

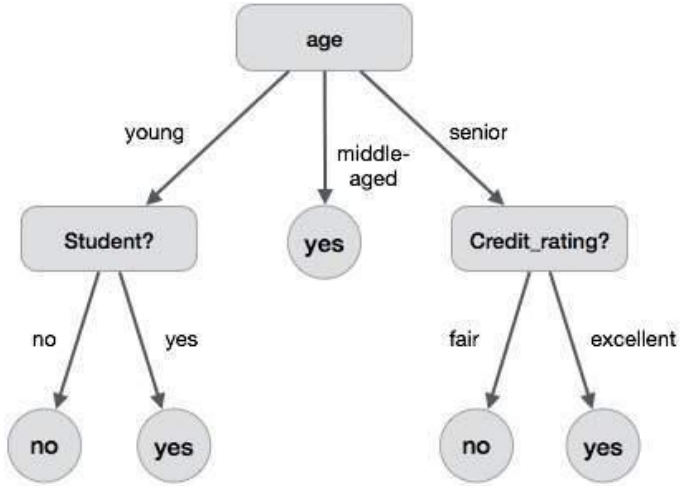
	StudentID	Gender	DOB	Race	Ethnicity	Class	Weight	Height	Enrollment_Date	State_Residency
1	5	1	08/15/1991	2	1	1	226	70	08/15/2012	In state
2	9	1	11/01/1991	3	1	1	144	71	08/15/2012	
3	35	1	10/29/1990	1	.	1	.	.	08/15/2012	Out of state
4	70	2	04/06/1994	1	2	1	175	63	08/15/2012	In state
5	44	1	01/31/1991	1	2	2	170	77	.	In state
6	51	1	.	1	1	2	177	71	08/15/2011	Out of state
7	85	2	09/26/1991	.	2	2	141	.	.	Out of state
8	19	1	05/25/1991	1	2	3	184	74	.	In state
9	40	1	10/29/1990	1	2	3	170	67	08/15/2010	In state
10	43	1	02/03/1990	2	2	3	.	.	08/15/2010	Out of state
11	24	1	09/04/1993	1	2	4	167	73	08/15/2007	In state
12	39	1	08/12/1993	3	2	4	150	73	08/15/2006	Out of state
13	45	1	03/09/1994	1	2	4	161	71	08/15/2007	In state
14	79	2	02/16/1992	1	2	4	143	62	08/15/2008	In state
15	89	.	09/11/1993	1	2	4	128	64	08/15/2009	Out of state

Missing numeric values are a period.

Missing character values are blank.

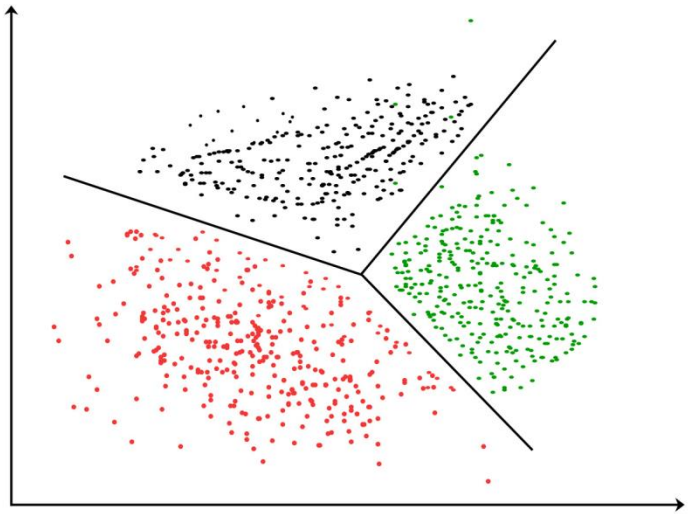
DATA TRANSFORMATION

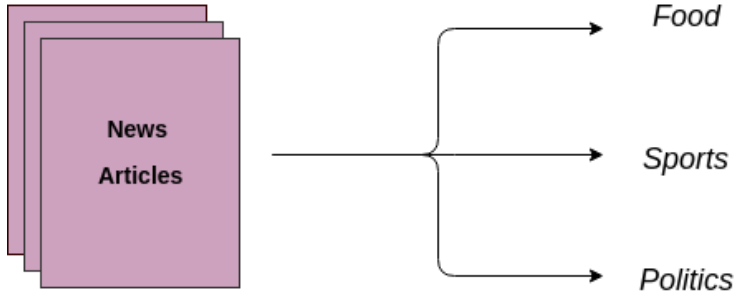
- **Data warehousing**
  - A data warehouse is a repository of data collected from multiple data sources and is intended to be used as a whole under the same unified schema.
- **Feature selection**
  - The primary objective is the determination of appropriate data type and source that allow investigators to answer research questions.

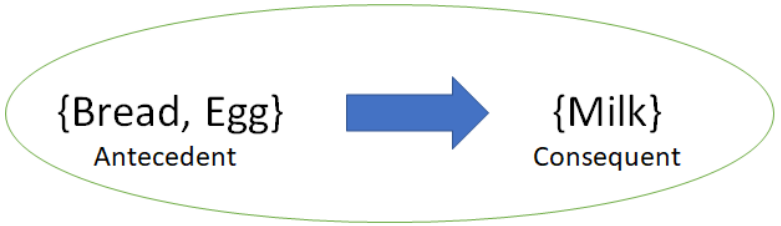


DATA MINING

1. **Clustering**
2. **Classification**
3. **Association Rule Learning**



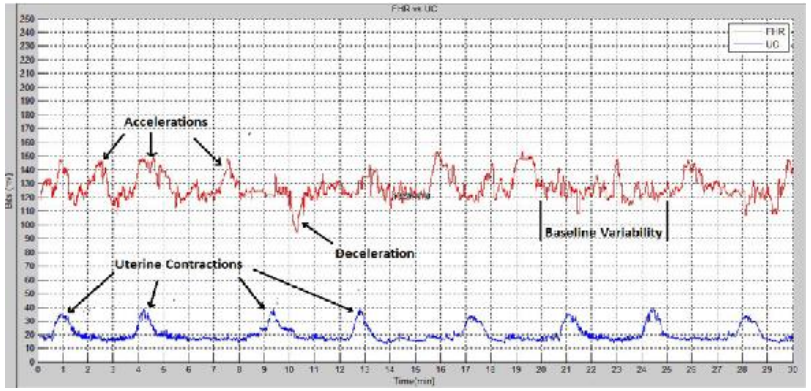




Itemset = {Bread, Egg, Milk}

EVALUATION

- **Pattern evaluation**: strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge discovery** is the crucial step in which clever techniques are applied to extract patterns potentially useful.



INTERPRETATION

- **Information presentation** means the discovered knowledge is visually represented to the user. This essential step uses visualization skills to help users understand and interpret the data mining results.
- **Decision making** means the user make a decision depending on the discovered knowledge.

13

THANK YOU