

# Data Preparation: Data Types

Mining Massive Datasets

Prof. Carlos Castillo

Topic 03



uncurated  
raw data

data engineer

data  
scientist

# Main Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 2) + [slides by Lijun Zhang](#)
- Introduction to Data Mining 2<sup>nd</sup> edition (2019) by Tan et al. (Chapter 2)
- Data Mining Concepts and Techniques, 3<sup>rd</sup> edition (2011) by Han et al. (Chapter 3)

“凡事豫（预）则立，不豫（预）则废”——《礼记·中庸》

Success depends upon previous preparation, and without such preparation there is sure to be failure – Confucius

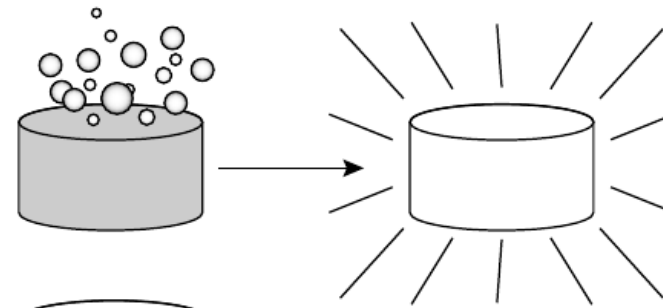
# Typical datasets

- Records / Matrices
- Documents
- Transactions
- Graphs
- Temporal / Sequences
- Spatial

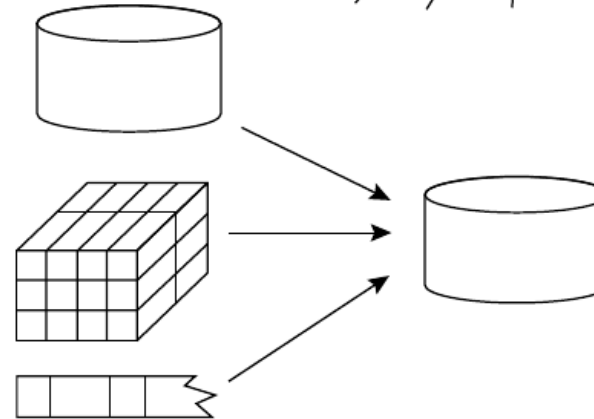
# Data preparation

- Feature Extraction and Portability
  - Extract relevant elements for our analysis
  - Convert heterogeneous data types
- Data Cleaning
  - Deal with missing, erroneous, and inconsistent data
- Data Integration
  - Bring different data sources into a common framework
- Data Reduction, Selection, and Transformation
  - Done for both efficiency and effectiveness

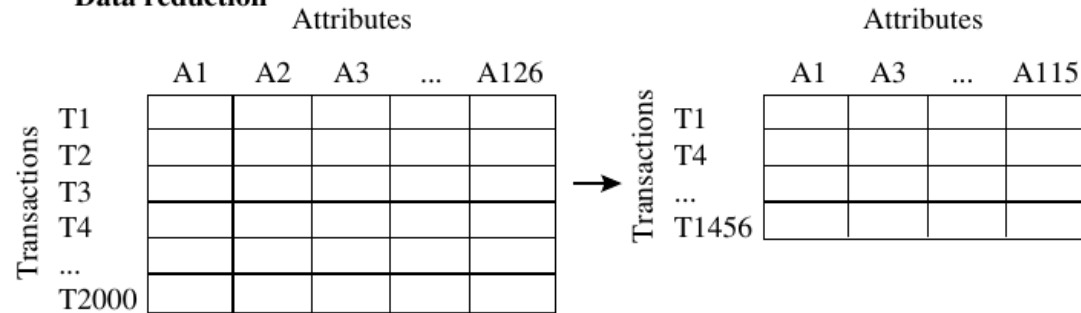
**Data cleaning**



**Data integration**



**Data reduction**



**Data transformation**

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

# Feature extraction examples

Domain	Raw Data	Features
Sensor	Low-level signals	Wavelet or Fourier transforms
Image	Pixels	Color histograms Visual words
Web logs	Text strings	IP address Action
Network traffic	Characteristics of the network packets	Number of bytes transferred Network protocol
Document data	Text strings	Bag-of-words Entity extraction

This is both a skill and an art that the analyst develops over the years.



# Data type conversions

# Data type conversions

- Data is often **heterogeneous**
  - A demographic data set may contain both numeric and mixed attributes
- Possible solution
  - Designing an algorithm for an **arbitrary combination** of data types
    - Time-consuming and sometimes impractical
- **Converting** between various data types
  - Using off-the-shelf tools for processing

# Data type conversions (cont.)

Some ways of converting between data types

Source data type	Destination data type	Methods
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent Semantic Analysis ( <i>LSA</i> )
Time series	Discrete sequence	Symbolic Aggregate Approximation ( <i>SAX</i> )
Time series	Numeric multidimensional	Discrete Wavelet Transform ( <i>DWT</i> ), Discrete Fourier Transform ( <i>DFT</i> )
Discrete sequence	Numeric multidimensional	
Spatial	Numeric multidimensional	2-d <i>DWT</i>
Graphs	Numeric multidimensional	Multidimensional Scaling ( <i>MDS</i> ), spectral
Any type	Graphs	Similarity graph (restricted applicability)

Numerical  Categorical

# Numerical to categorical: discretization

- Divide the range for the numerical variable into  $\Phi$  different ranges

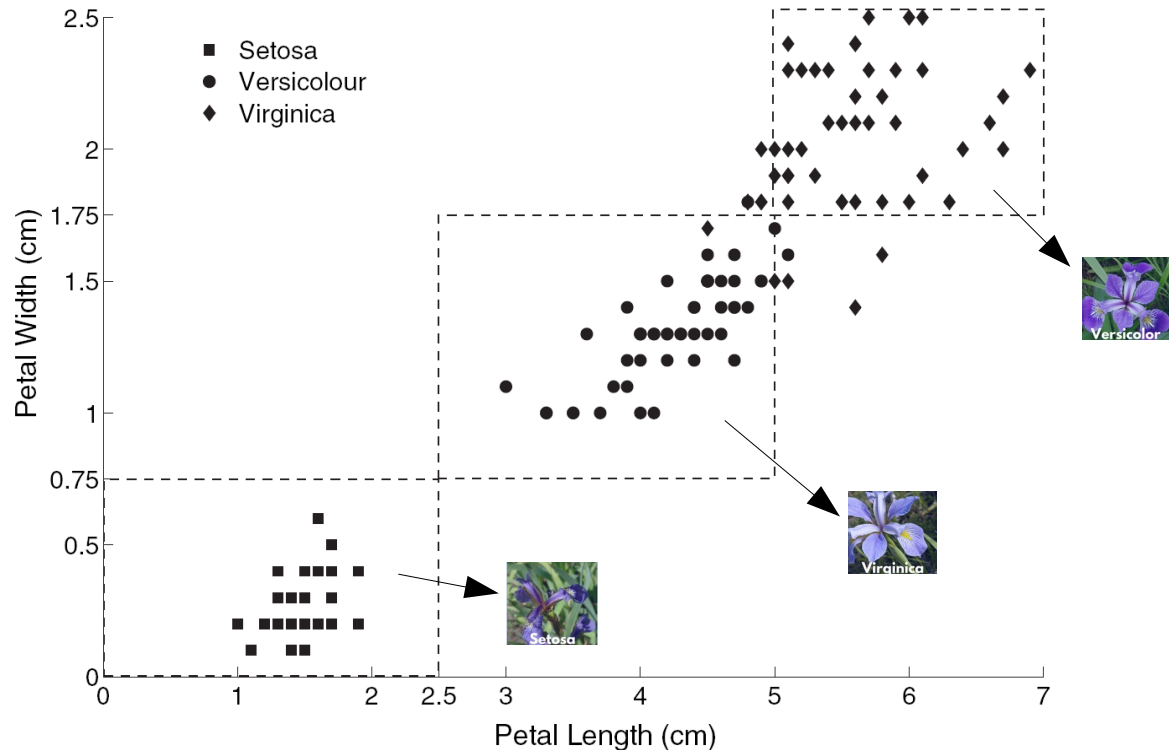


# Numerical to categorical: discretization (cont.)



- **Equi-width** ranges ( $l_i - r_i$  is constant)
- **Equi-log** ranges ( $\log r_i - \log l_i$  is constant)
- **Equi-depth** ranges (num. items in  $[l_i, r_i]$  constant)

# Example discretization in IRIS dataset



Continuous variables are converted to three possible values per feature: small, medium, large

Answer in  
Google Spreadsheet  
(Link in Aula Global)

# Exercise

- Given this database
- Create two categorical (ordinal) attributes
  - Salary bin (equi width) with salary binned into equi-width categories
  - Salary bin (equi depth) with salary binned into equi-depth categories
- Values: low, medium, high

Person	Salary
a	34,000
b	49,000
c	53,000
d	54,000
e	32,000
f	44,000
g	41,000
h	37,000
i	48,000



# Categorical to numerical: binarization (one-hot encoding)

- One categorical value with K categories  
⇒ indicator vector with K binary variables

User name	gender
alice	Female
bob	Male
cara	Female
...	...



alice	bob	cara	Female	Male	...
1	0	0	1	0	...
0	1	0	0	1	...
0	0	1	1	0	...
...	...	...	...	...	...

# Series and sequences

# Time series to discrete sequence

- Symbolic aggregate approximation (SAX)
  - Window-based averaging
    - Evaluate the average value in each window
  - Value-based discretization
    - Discretize the average value by equi-depth intervals
- How to ensure equi-depth without seeing the entire series?
  - Assume certain distribution, such as Gaussian
  - Estimate the distribution

# Time series to numeric data

- Discrete Wavelet Transform (DWT)
- Discrete Fourier transform (DFT)

(Seen elsewhere, e.g., signal processing)

# Discrete sequence to numeric

- Discrete sequence to a set of (binary) time series
  - ACACACTGTGACTG (4 Symbols)
  - 10101000001000 (A)
  - 01010100000100 (C)
  - 00000010100010 (T)
  - 00000001010001 (G)
- Map each of these time series into a multidimensional vector
- Features from the different series are combined

Graphs ↔ Numerical

# Convert any data type to a graph

- Determine distance  $d(u, v)$  between **all pairs** of elements  $(u, v)$
- All elements with  $d(u, v) \leq \theta$  are **connected**

# Graphs to numerical

- Graph embeddings
  - Each node is converted into a point in a low-dimensional space
  - Nearby nodes in the low-dimensional space are connected by short paths in the graph

We might see more on this on the *spectral graph clustering* topic, possibly (if we get to that topic)



# Summary

# Things to remember

- Converting across data types

# Exercises for TT03-TT05

- Exercises 3.7 of Data Mining Concepts and Techniques, 3<sup>rd</sup> edition (2011) by Han et al.
- Exercises 2.6 of Introduction to Data Mining, Second Edition (2019) by Tan et al.
  - Mostly the first exercises, say 1-6