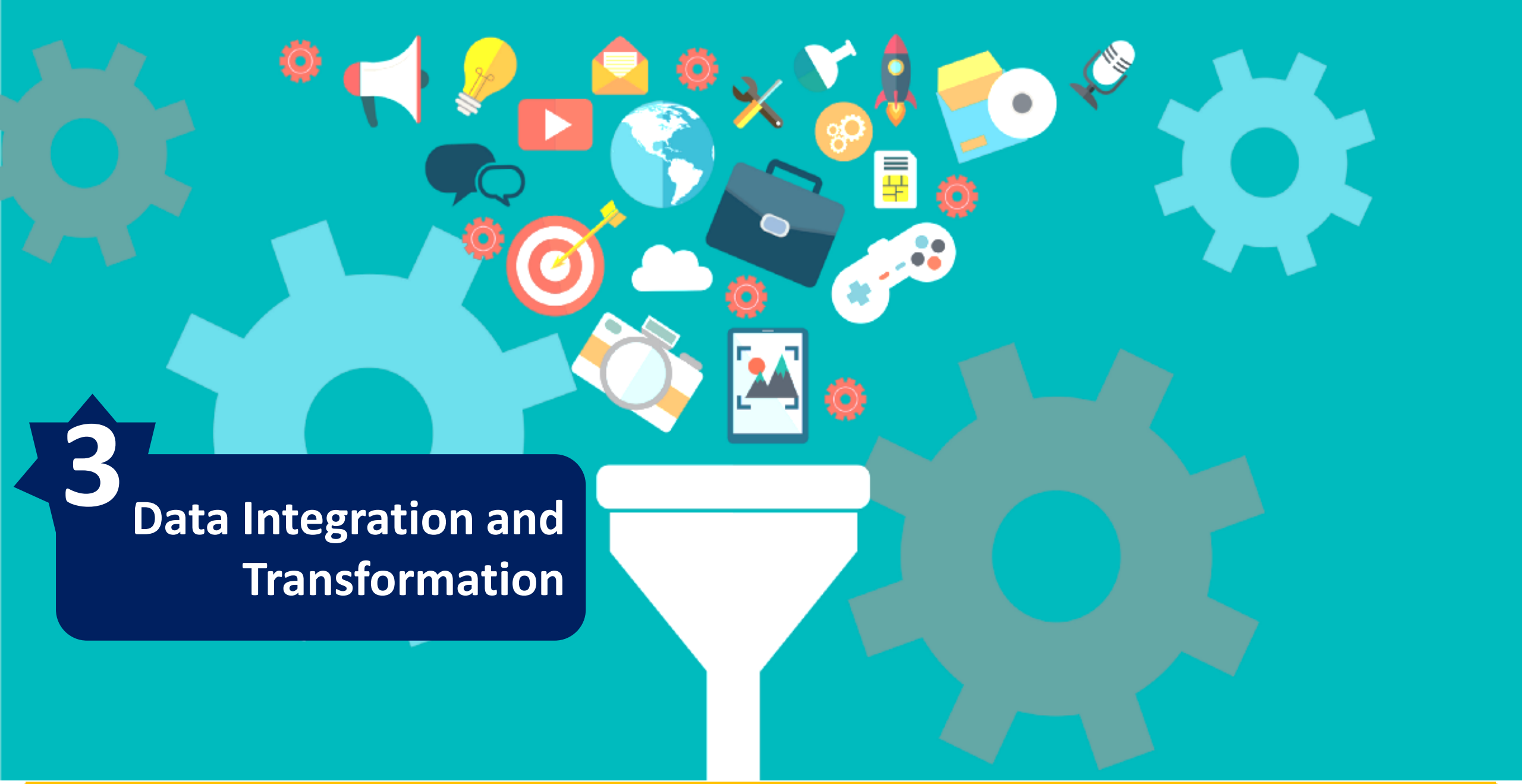
DATA MINING PRE-PROCESSING

Outline

1. Why data preprocessing?
2. Data cleaning
3. Data integration and transformation
4. Data reduction
5. Discretization and concept hierarchy generation
6. Summary



3 Data Integration and Transformation

Data Integration

Data integration:

- combines data from multiple sources into a coherent store

Schema integration

- integrate metadata from different sources
- Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#

Detecting and resolving data value conflicts

- for the same real world entity, attribute values from different sources are different
- possible reasons: different representations, different scales, e.g., metric vs. British units, different currency

Handling Redundant Data in Data Integration

Redundant data occur often when integrating multiple DBs

- The same attribute may have different names in different databases
- One attribute may be a “derived” attribute in another table, e.g., annual revenue

Redundant data may be able to be detected by correlational analysis

$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$$

Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

Smoothing: remove noise from data (binning, clustering, regression)

Aggregation: summarization, data cube construction

Generalization: concept hierarchy climbing

Normalization: scaled to fall within a small, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Attribute/feature construction

- New attributes constructed from the given ones

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

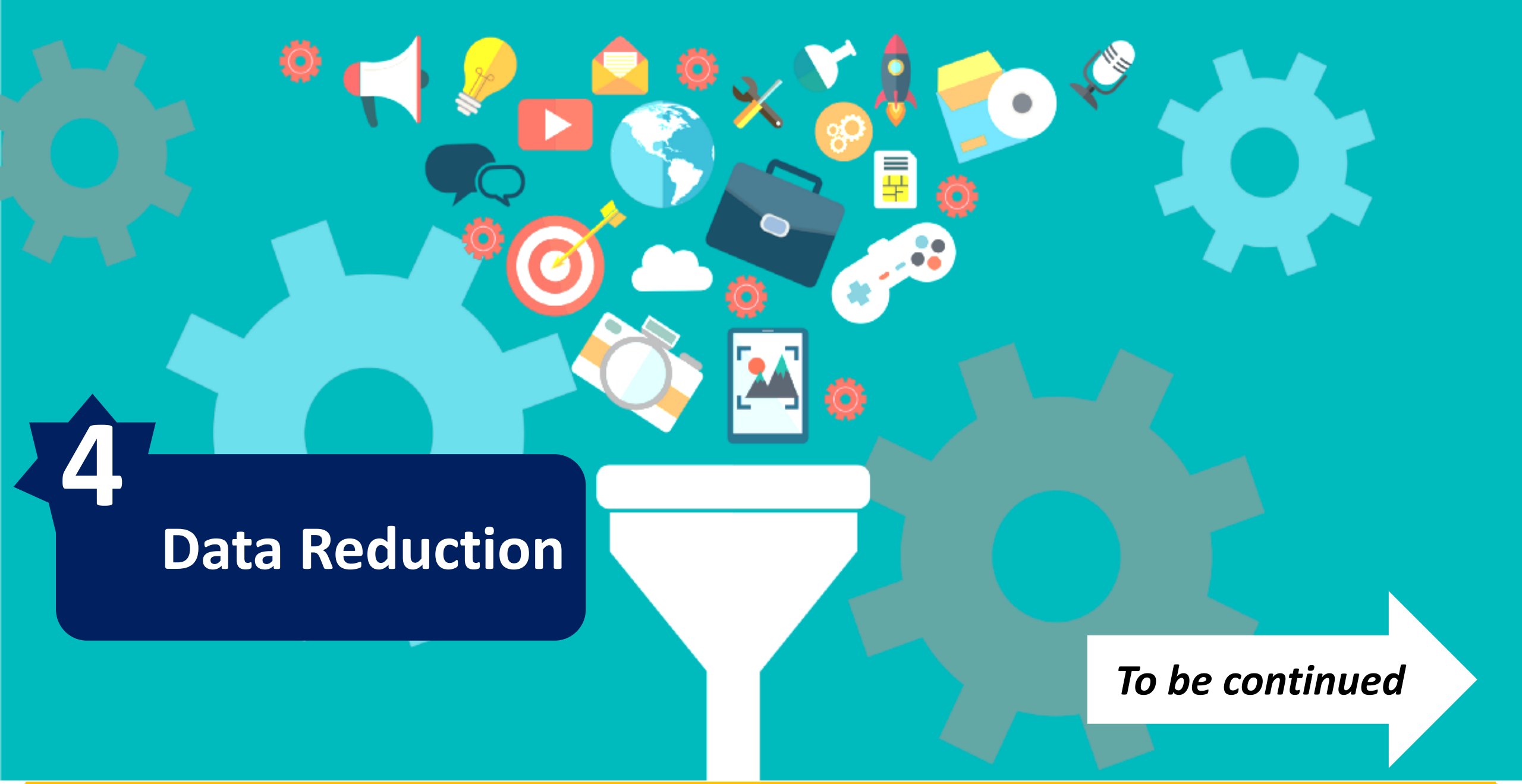
- z-score normalization

$$v' = \frac{v - mean_A}{stdev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that $Max(|v'|) < 1$



4 Data Reduction

To be continued