Data Transformation: Key Methods

Data transformation is a crucial step in data preprocessing. It involves converting data from one format or structure into another, making it more suitable for analysis and modeling. Here's a summary of the key methods, drawing from the document:

1. **Smoothing:**
    - **Purpose:** To remove noise (random variation or errors) from the data.
    - **How it works:** Uses algorithms to highlight important patterns and suppress irrelevant fluctuations.
    - **Example:** Applying a moving average to a time series of stock prices to smooth out daily fluctuations and reveal the overall trend. If you have daily sales data with lots of ups and downs, smoothing might involve calculating a 7-day rolling average to see the weekly trend more clearly.
2. **Aggregation:**
    - **Purpose:** To summarize data, often from multiple sources, into a more concise form.
    - **How it works:** Combines data, often using summary statistics (like sum, average, count).
    - **Example:** Combining daily sales data from multiple stores into monthly sales totals for each region. Or, taking individual customer transaction records and aggregating them to calculate the total spending per customer.
3. **Discretization:**
    - **Purpose:** To convert continuous attributes (those with a range of numeric values) into discrete attributes (with a limited set of categories or intervals).
    - **How it works:** Divides the range of a continuous attribute into intervals and assigns a label to each interval.
    - **Example:** Converting a person's age (e.g., 25, 38, 62) into age groups (e.g., "Young" (0-20), "Middle-aged" (21-50), "Senior" (51+)). Or, transforming exam scores (0-100) into letter grades (A, B, C, D, F).
4. **Attribute Construction (Feature Construction):**
    - **Purpose:** To create new attributes (features) from existing ones.
    - **How it works:** Combines or transforms existing attributes to create new ones that might be more informative for analysis.
    - **Example:** From "Date of Birth" and "Current Date," you can create a new attribute "Age." Or, from "Length" and "Width" of a rectangle, you can create a new attribute "Area." In the document's example, creating a new attribute "Joined in 2019" based on employee data.
5. **Generalization:**

- **Purpose:** To replace low-level (detailed) data with higher-level concepts.
- **How it works:** Uses concept hierarchies to move from specific values to more general categories.
- **Example:** Replacing specific street addresses with city names, or city names with country names. Generalizing "Red Delicious Apple" to "Apple" to "Fruit."

6. **Normalization:**
    - **Purpose:** To scale data to fall within a specific range. This is essential when attributes have vastly different ranges, which can skew the results of some data mining algorithms.
    - **How it works:** Applies mathematical transformations to bring all attribute values into a common range (e.g., 0 to 1, or -1 to 1).
    - **Example:** Scaling exam scores (originally 0-100) and GPA (originally 0.0-4.0) to both be between 0 and 1 so that one attribute doesn't dominate the other in a clustering algorithm. Common methods include:
        - **Min-Max Normalization:** Scales data to a range (often 0 to 1).
        - **Z-score Normalization:** Scales data to have a mean of 0 and a standard deviation of 1.

In summary, data transformation is a critical set of techniques to prepare data and the methods that are listed and defined in the document are:

1. Smoothing
2. Aggregation
3. Discretization
4. Attribute Construction
5. Generalization
6. Normalization.