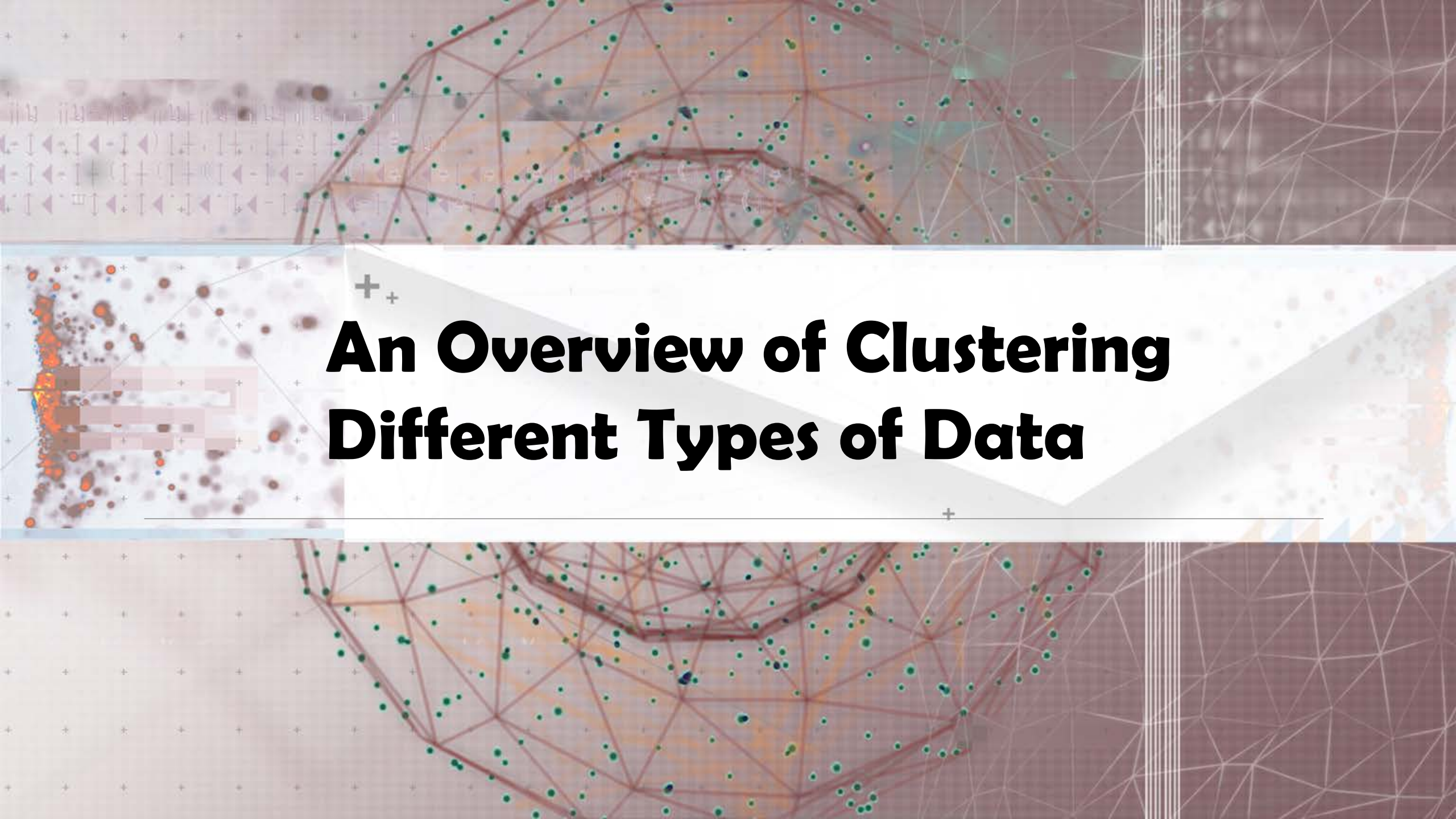

An Overview of Clustering Different Types of Data

Clustering Different Types of Data (I)

- **Numerical data**
  - Most earliest clustering algorithms were designed for numerical data
- **Categorical data** (including binary data)
  - Discrete data, no natural order (e.g., sex, race, zip-code, and market-basket)
- **Text data**: Popular in social media, Web, and social networks
  - Features: High-dimensional, sparse, value corresponding to word frequencies
  - Methods: Combination of k-means and agglomerative; topic modeling; co-clustering
- **Multimedia data**: Image, audio, video (e.g., on Flickr, YouTube)
  - Multi-modal (often combined with text data)
  - Contextual: Containing both behavioral and contextual attributes
    - Images: Position of a pixel represents its context, value represents its behavior
    - Video and music data: Temporal ordering of records represents its meaning

Clustering Different Types of Data (II)

- **Time-series data**: Sensor data, stock markets, temporal tracking, forecasting, etc.
  - Data are temporally dependent
  - Time: contextual attribute; data value: behavioral attribute
  - Correlation-based online analysis (e.g., online clustering of stock to find stock tickers)
  - Shape-based offline analysis (e.g., cluster ECG based on overall shapes)
- **Sequence data**: Weblogs, biological sequences, system command sequences
  - Contextual attribute: Placement (rather than time)
  - Similarity functions: Hamming distance, edit distance, longest common subsequence
  - Sequence clustering: Suffix tree; generative model (e.g., Hidden Markov Model)
- **Stream data**:
  - Real-time, evolution and concept drift, single pass algorithm
  - Create efficient intermediate representation, e.g., micro-clustering

Clustering Different Types of Data (III)

- **Graphs and homogeneous networks**
  - Every kind of data can be represented as a graph with similarity values as edges
  - Methods: Generative models; combinatorial algorithms (graph cuts); spectral methods; non-negative matrix factorization methods
- **Heterogeneous networks**
  - A network consists of multiple typed nodes and edges (e.g., bibliographical data)
  - Clustering different typed nodes/links together (e.g., NetClus)
- **Uncertain data**: Noise, approximate values, multiple possible values
  - Incorporation of probabilistic information will improve the quality of clustering
- **Big data**: Model systems may store and process very big data (e.g., weblogs)
  - Ex. Google's MapReduce framework
    - Use *Map* function to distribute the computation across different machines
    - Use R*educe* function to aggregate results obtained from the Map step