

The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements. In the upper and lower portions, there are dense networks of thin, light-colored lines forming a web-like structure. Scattered throughout these networks are small, colored dots in shades of green, blue, and yellow. On the left side, there is a vertical strip containing a grid of small, light-colored plus signs. In the center, a large, white, angular shape resembling a stylized 'V' or a folded piece of paper points downwards, serving as a backdrop for the title. The overall aesthetic is technical and data-driven.

What Is Cluster Analysis?

What Is Cluster Analysis?

- ❑ What is a cluster?

- ❑ A cluster is a collection of data objects which are
 - ❑ Similar (or related) to one another within the same group (i.e., cluster)
 - ❑ Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)

- ❑ Cluster analysis (or *clustering*, *data segmentation*, ...)

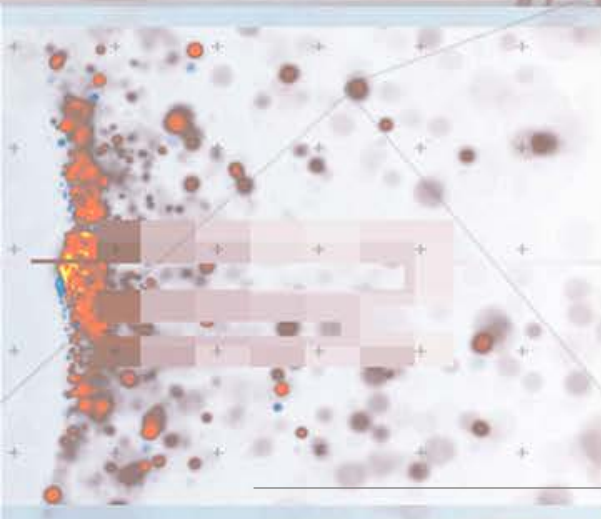
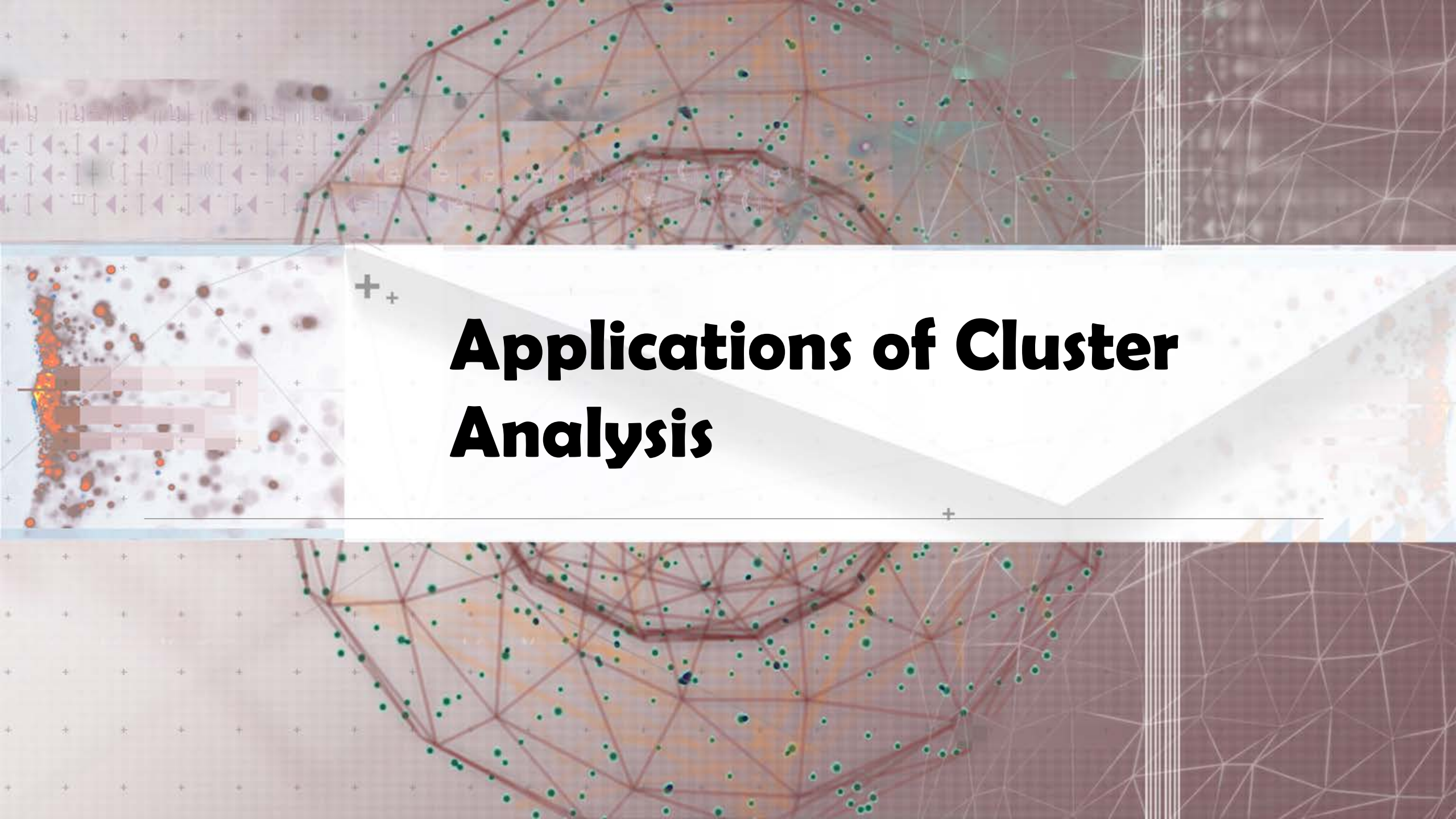
- ❑ Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible

- ❑ Cluster analysis is **unsupervised learning** (i.e., no predefined classes)

- ❑ This contrasts with *classification* (i.e., *supervised learning*)

- ❑ Typical ways to use/apply cluster analysis

- ❑ As a stand-alone tool to get insight into data distribution, or
- ❑ As a preprocessing (or intermediate) step for other algorithms



++

Applications of Cluster Analysis

+

Cluster Analysis: Applications

- ❑ A key intermediate step for other data mining tasks
 - ❑ Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
 - ❑ Outlier detection: Outliers—those “far away” from any cluster
- ❑ Data summarization, compression, and reduction
 - ❑ Ex. Image processing: Vector quantization
- ❑ Collaborative filtering, recommendation systems, or customer segmentation
 - ❑ Find like-minded users or similar products
- ❑ Dynamic trend detection
 - ❑ Clustering stream data and detecting trends and patterns
- ❑ Multimedia data analysis, biological data analysis and social network analysis
 - ❑ Ex. Clustering images or video/audio clips, gene/protein sequences, etc.

The background features a complex geometric pattern of thin, light-colored lines forming a network of triangles and polygons. Overlaid on this are several semi-transparent elements: a horizontal band of purple and blue wavy patterns at the top, a vertical band of orange and red wavy patterns on the left, and a large white banner with a grey border that contains the title. Small grey plus signs are scattered throughout the background.

Requirements and Challenges

Considerations for Cluster Analysis

❑ Partitioning criteria

- ❑ Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)

❑ Separation of clusters

- ❑ Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)

❑ Similarity measure

- ❑ Distance-based (e.g., Euclidean, road network, vector) vs. connectivity-based (e.g., density or contiguity)

❑ Clustering space

- ❑ Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Requirements and Challenges

□ Quality

- Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
- Discovery of clusters with arbitrary shape
- Ability to deal with noisy data

□ Scalability

- Clustering all the data instead of only on samples
- High dimensionality
- Incremental or stream clustering and insensitivity to input order

□ Constraint-based clustering

- User-given preferences or constraints; domain knowledge; user queries

□ Interpretability and usability

The background is a complex collage of abstract elements. It features a grid of small grey plus signs, a network of red lines connecting green dots, and a vertical strip of orange and red dots. A large, light grey, angular shape is positioned behind the text.

A Multi-Dimensional Categorization

Cluster Analysis: A Multi-Dimensional Categorization

❑ Technique-Centered

- ❑ Distance-based methods
- ❑ Density-based and grid-based methods
- ❑ Probabilistic and generative models
- ❑ Leveraging dimensionality reduction methods
- ❑ High-dimensional clustering
- ❑ Scalable techniques for cluster analysis

❑ Data Type-Centered

- ❑ Clustering numerical data, categorical data, text data, multimedia data, time-series data, sequences, stream data, networked data, uncertain data

❑ Additional Insight-Centered

- ❑ Visual insights, semi-supervised, ensemble-based, validation-based

The background is a collage of various data visualization techniques. It includes a network graph with red lines and green nodes, a scatter plot with orange and blue points, a heatmap with a color gradient from blue to red, and a grid of small plus signs. The text is centered over a white, angular shape.

An Overview of Typical Clustering Methodologies

Typical Clustering Methodologies (I)

□ Distance-based methods

- Partitioning algorithms: K-Means, K-Medians, K-Medoids
- Hierarchical algorithms: Agglomerative vs. divisive methods

□ Density-based and grid-based methods

- Density-based: Data space is explored at a high-level of granularity and then post-processing to put together dense regions into an arbitrary shape
- Grid-based: Individual regions of the data space are formed into a grid-like structure

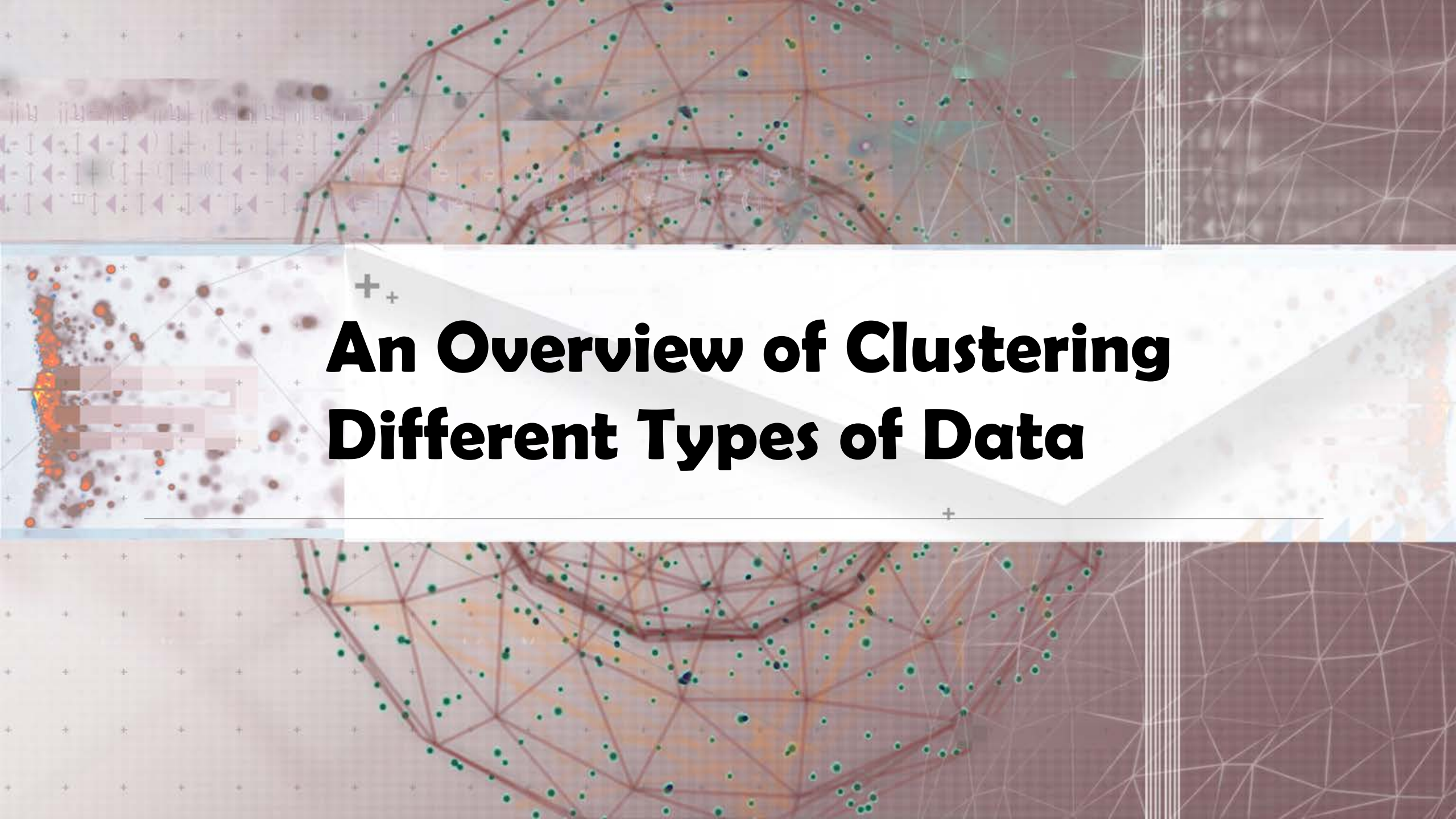
□ Probabilistic and generative models: Modeling data from a generative process

- Assume a specific form of the generative model (e.g., mixture of Gaussians)
- Model parameters are estimated with the Expectation-Maximization (EM) algorithm (using the available dataset, for a maximum likelihood fit)
- Then estimate the generative probability of the underlying data points

Typical Clustering Methodologies (II)

▣ High-dimensional clustering

- ▣ Subspace clustering: Find clusters on various subspaces
 - ▣ Bottom-up, top-down, correlation-based methods vs. δ -cluster methods
- ▣ Dimensionality reduction: A vertical form (i.e., columns) of clustering
 - ▣ Columns are clustered; may cluster rows and columns together (co-clustering)
- ▣ Probabilistic latent semantic indexing (PLSI) then LDA: Topic modeling of text data
 - ▣ A cluster (i.e., topic) is associated with a set of words (i.e., dimensions) and a set of documents (i.e., rows) simultaneously
- ▣ Nonnegative matrix factorization (NMF) (as one kind of co-clustering)
 - ▣ A nonnegative matrix A (e.g., word frequencies in documents) can be approximately factorized two non-negative low rank matrices U and V
- ▣ Spectral clustering: Use the *spectrum* of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions

The background is a collage of various data visualization techniques. It includes a network graph with red lines and green nodes, a scatter plot with orange and blue points, a heatmap with a color gradient from blue to red, and a grid of small plus signs. The text is centered over a white, angular shape.

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 *Reduce* function to aggregate results obtained from the Map step

The background features a complex network of red lines connecting green and blue dots, suggesting a data graph or social network. On the left, there is a smaller inset showing a heatmap with orange and red clusters. The title text is centered on a white, angular geometric shape.

An Overview of User Insights and Clustering

User Insights and Interactions in Clustering

- **Visual insights:** One picture is worth a thousand words
 - Human eyes: High-speed processor linking with a rich knowledge-base
 - A human can provide intuitive insights; HD-eye: visualizing HD clusters
- **Semi-supervised insights:** Passing user's insights or intention to system
 - User-seeding: A user provides a number of labeled examples, approximately representing categories of interest
- **Multi-view and ensemble-based insights**
 - Multi-view clustering: Multiple clusterings represent different perspectives
 - Multiple clustering results can be ensembled to provide a more robust solution
- **Validation-based insights:** Evaluation of the quality of clusters generated
 - May use case studies, specific measures, or pre-existing labels

Recommended Readings

❑ Major Reference Books on Cluster Analysis

- ❑ Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011 (Chapters 10 & 11)
- ❑ Charu Aggarwal and Chandran K. Reddy (eds.). Data Clustering: Algorithms and Applications. CRC Press, 2014
- ❑ Mohammed J. Zaki and Wagner Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014

❑ Reference paper for this lecture

- ❑ Charu Aggarwal. An Introduction to Clustering Analysis. *in* Aggarwal and Reddy (eds.). Data Clustering: Algorithms and Applications (Chapter 1). CRC Press, 2014