

The background features a complex network graph with red lines connecting green and blue nodes. On the left, there is a smaller inset showing a cluster of orange and red nodes. The overall aesthetic is technical and data-driven.

# **Pattern Mining Applications: Mining Quality Phrases from Text Data**

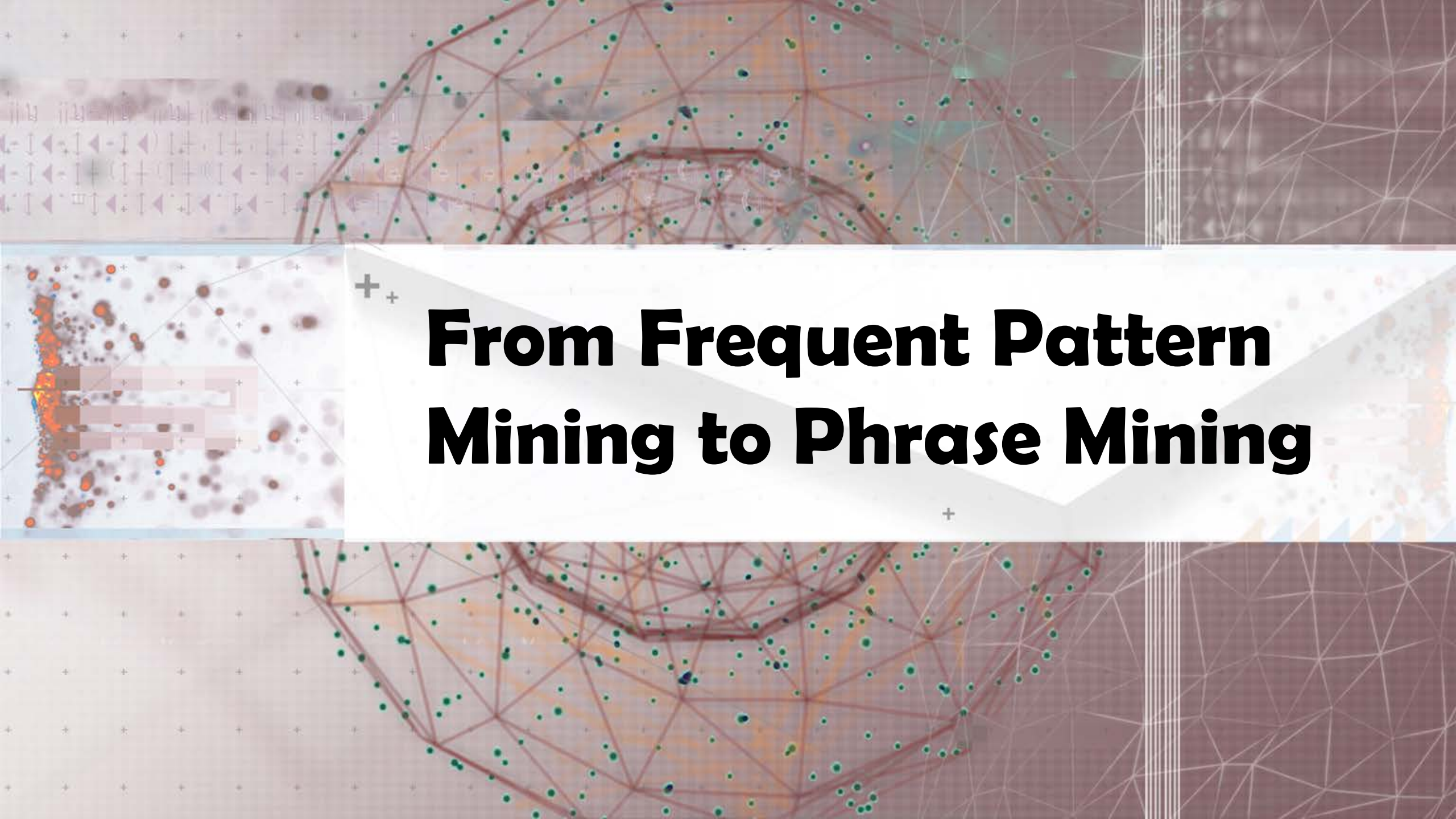
# Pattern Mining Applications: Mining Quality Phrases from Text Data

---

- ❑ From Frequent Pattern Mining to Phrase Mining
- ❑ Previous Phrase Mining Methods
- ❑ ToPMine: Phrase Mining without Training Data
- ❑ SegPhrase: Phrase Mining with Tiny Training Sets

Thanks to Ahmed El-Kishky@UIUC, Jialu Liu@UIUC, Jingbo Shang@UIUC, Xiang Ren@UIUC, Chi Wang@MSR and Marina Danilevsky@IBM for their contributions



The background features a complex network of thin, light-colored lines forming a web-like structure. Overlaid on this are various data visualizations: a grid of small plus signs in the top-left, a series of purple arrows pointing left in the top-center, a dense cluster of orange and red dots with a heatmap overlay in the bottom-left, and a large network of green dots connected by red lines in the center and bottom-right. The overall color palette is muted, with earthy tones and soft pastels.

# **From Frequent Pattern Mining to Phrase Mining**

# Why Phrase Mining?

---

- ❑ Unigrams vs. phrases
  - ❑ **Unigrams** (single words) are often *ambiguous*
    - ❑ Example: “United”: United States? United Airline? United Parcel Service?
  - ❑ **Phrase**: A natural, meaningful, *unambiguous* semantic unit
    - ❑ Example: “United States” vs. “United Airline”
- ❑ Mining semantically meaningful phrases
  - ❑ Transform text data from *word granularity* to *phrase granularity*
  - ❑ Enhance the power and efficiency at manipulating unstructured data

# From Frequent Pattern Mining to Phrase Mining

---

- General principle
  - Exploit information redundancy and data-driven criteria to determine phrase boundaries and salience
- Methodology: Exploring three ideas
  - Frequent pattern mining and colocation analysis
  - Phrasal segmentation
  - Quality phrase assessment
- Recent developments of phrase mining methods
  - ToPMine: Mining quality phrase without training (A. El-Kishky, et al., 2015)
  - SegPhrase: Mining quality phrase with tiny training sets (J. Liu, et al., 2015)