

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 intricate networks of thin, light-colored lines forming a mesh or web-like structure. Scattered throughout these areas are numerous small, colored dots in shades of green, blue, and orange. In the center, a large, white, angular shape resembling a stylized 'V' or a folded piece of paper serves as a backdrop for the title. To the left of this central shape, there is a smaller, rectangular inset image showing a cluster of orange and red dots on a light background, with a horizontal bar chart or heatmap overlaid on it. The overall aesthetic is technical and data-driven.

Lecture 7. Sequential Pattern Mining

Lecture 7. Sequential Pattern Mining

- ❑ Sequential Pattern and Sequential Pattern Mining
- ❑ Sequential Pattern Mining Algorithms
 - ❑ **GSP** (Generalized Sequential Patterns)
 - ❑ Vertical Format-Based Mining: **SPADE**
 - ❑ Pattern-Growth Methods: **PrefixSpan**
- ❑ Mining Closed Sequential Patterns: **CloSpan**
- ❑ Constrain-Based Sequential Pattern Mining



++

Session 1. Sequential Pattern and Sequential Pattern Mining

Sequence Databases & Sequential Patterns

- ❑ Sequential pattern mining has broad applications
 - ❑ Customer shopping sequences
 - ❑ Purchase a laptop first, then a digital camera, and then a smartphone, within 6 months
 - ❑ Medical treatments, natural disasters (e.g., earthquakes), science & engineering processes, stocks and markets, ...
 - ❑ Weblog click streams, calling patterns, ...
 - ❑ Software engineering: Program execution sequences, ...
 - ❑ Biological sequences: DNA, protein, ...
- ❑ Transaction DB, sequence DB vs. time-series DB
- ❑ Gapped vs. non-gapped sequential patterns
 - ❑ shopping, clicking streams vs. biological sequences


Sequential Pattern and Sequential Pattern Mining

- Sequential pattern mining: Given a set of sequences, find the **complete set of frequent subsequences** (i.e., satisfying the min_sup threshold)

A sequence database

SID	Sequence
10	<a(<u>ab</u> c)(a <u>c</u>)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(<u>ab</u>)(df) <u>c</u> b>
40	<eg(af)cbc>

A sequence: < (ef) (ab) (df) c b >



- An element may contain a set of *items* (also called *events*)
- Items within an element are unordered and we list them alphabetically

<a(bc)dc> is a subsequence of <a(abc)(ac)d(cf)>

- Given support threshold min_sup = 2, <(ab)c> is a sequential pattern

Sequential Pattern Mining Algorithms

- ❑ Algorithm requirement: Efficient, scalable, finding complete set, incorporating various kinds of user-specific constraints
- ❑ The Apriori property still holds: If a subsequence s_1 is infrequent, none of s_1 's super-sequences can be frequent
- ❑ Representative algorithms
 - ❑ **GSP** (Generalized Sequential Patterns): Srikant & Agrawal @ EDBT'96)
 - ❑ Vertical format-based mining: **SPADE** (Zaki@Machine Learning'00)
 - ❑ Pattern-growth methods: **PrefixSpan** (Pei, et al. @ICDE'01)
- ❑ Mining closed sequential patterns: **CloSpan** (Yan, et al. @SDM'03)
- ❑ Constraint-based sequential pattern mining