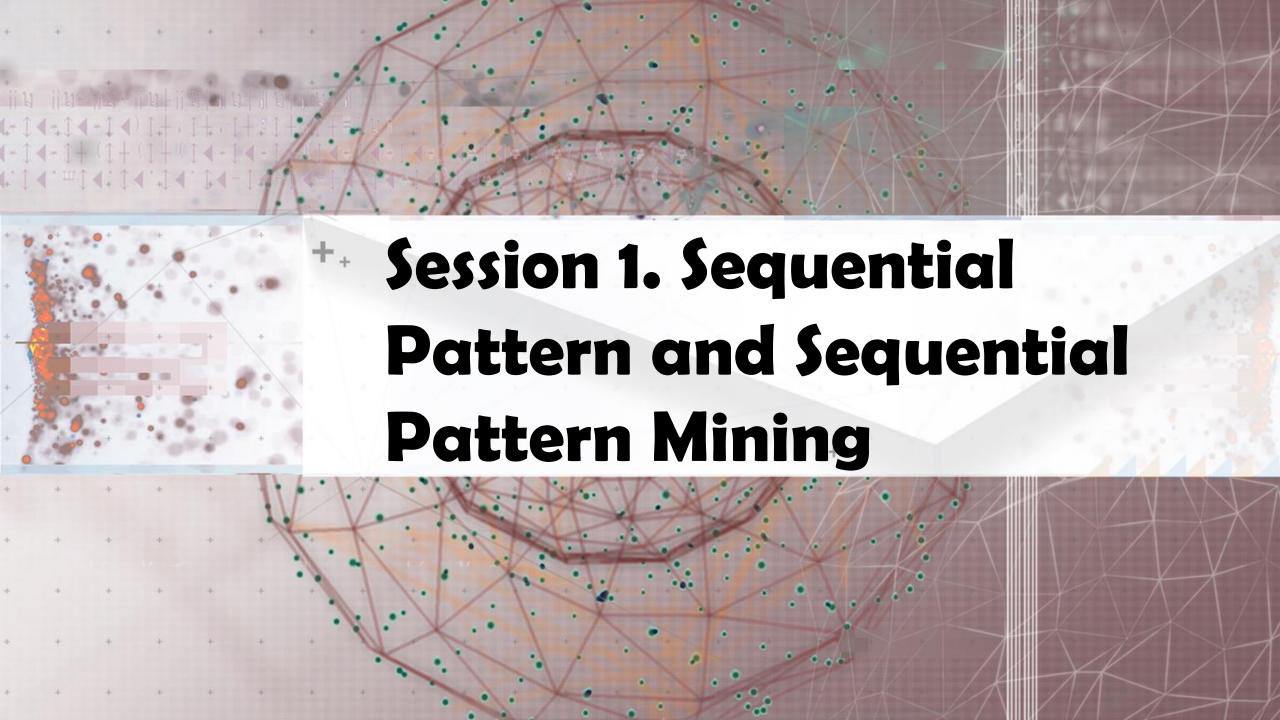


# Lecture 7. Sequential Pattern Mining

- 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



## Sequence Databases & Sequential Patterns

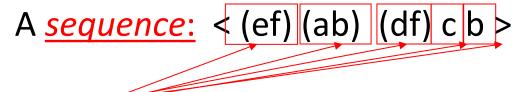
- Sequential pattern mining has broad applications
  - Customer shopping sequences
    - Purchase a laptop first, then a digital camera, and then a smartphone,
      within 6 months
  - Medical treatments, natural disasters (e.g., earthquakes), science & engineering processes, stocks and markets, ...
  - 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 <u>sequence database</u>

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



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

 $<a(bc)dc>is a <u>subsequence</u> of <math><\underline{a(abc)(ac)\underline{d(cf)}}>$ 

Given <u>support threshold</u> min\_sup = 2, <(ab)c> is a <u>sequential pattern</u>

## 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 Leanining'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