SESAR LAB

UNIVERSITÀ
DEGLI STUDI
DI MILANO

OVERVIEW ON BI METHODS

Paolo Ceravolo

paolo.ceravolo@unimi.it

Dipartimento di Informatica - Sede di Crema

Thursday 14.00 - 15.00

SESAR LAB

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Outline of the course

1. Definition of BPI
2. **Overview on BI methods**
3. Focus on Process Mining algorithms
4. Requisites for a well founded BI project
5. Exercise: review BPI works

BI methods

As said, BI implies to integrate different methods

1. Data Provisioning
2. Data Modelling and Analysis
3. Process Modelling and Analysis
4. Data Reporting and Knowledge Acquisition

BI methods

As said, BI implies the integration of different methods

1. Data Provisioning

   Data Extraction, Online Analytical Processing, Data Integration, Data Preparation

2. Data Modelling and Analysis

   Data Mining, Text Mining, Social Network Analysis, Organisational Mining, Decision Mining

3. Process Modelling and Analysis

   Process Model, Resource Planning, Process Mining

4. Data Reporting and Knowledge Acquisition

   Data Distribution, Data Visualisation, Decision Making

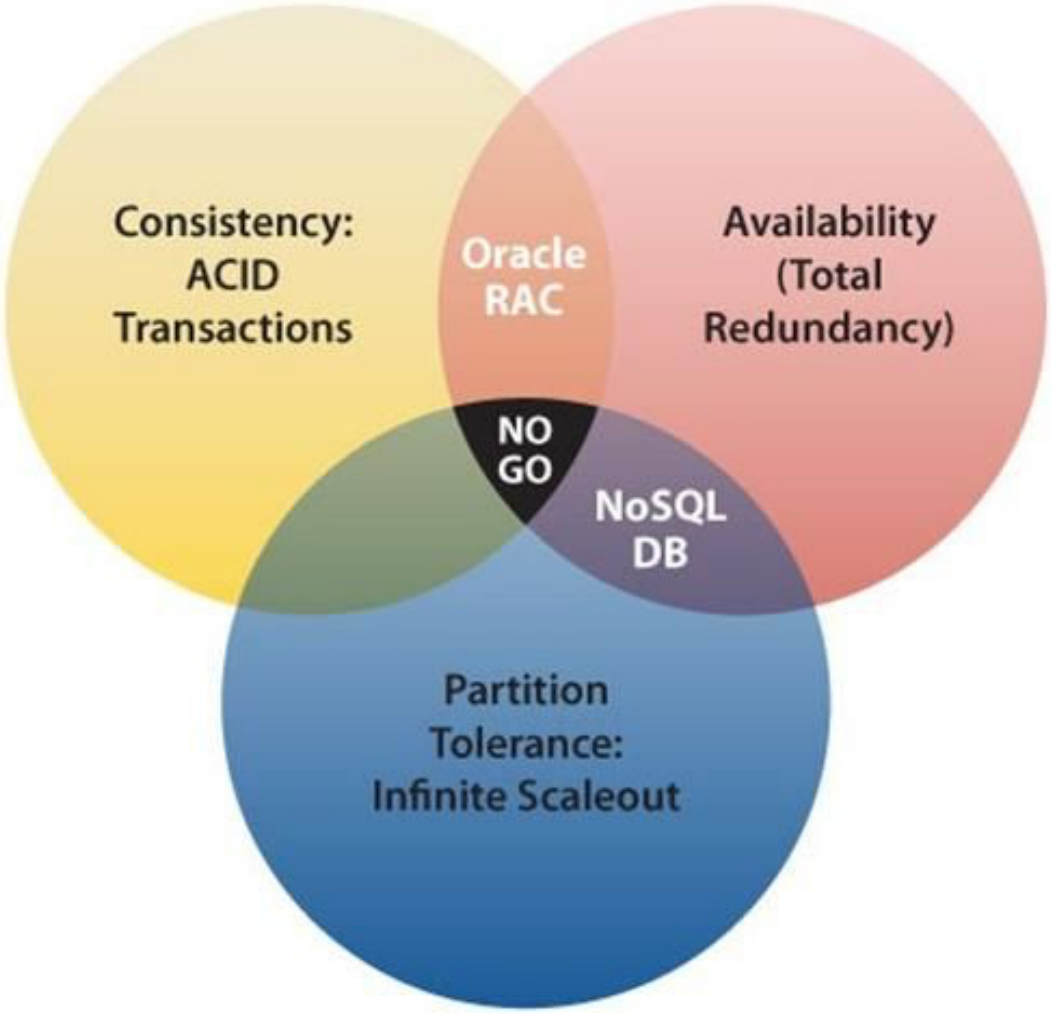
Data Provisioning

- The first challenge for DP is the variety of data sources:
  - serialisation (storage format)
  - structure (flat, hierarchical, document, process oriented)
  - conceptual model (semantics, granularity)
- The range of data variability in an operational system may exceed the expectations of designers

Data Provisioning

- This set is usually referred as *Extraction-Transformation-Load* (ETL) process
  - Staging step transforms data to a common model
  - Extraction may be offline (snapshot) or online (updates)

Data Provisioning

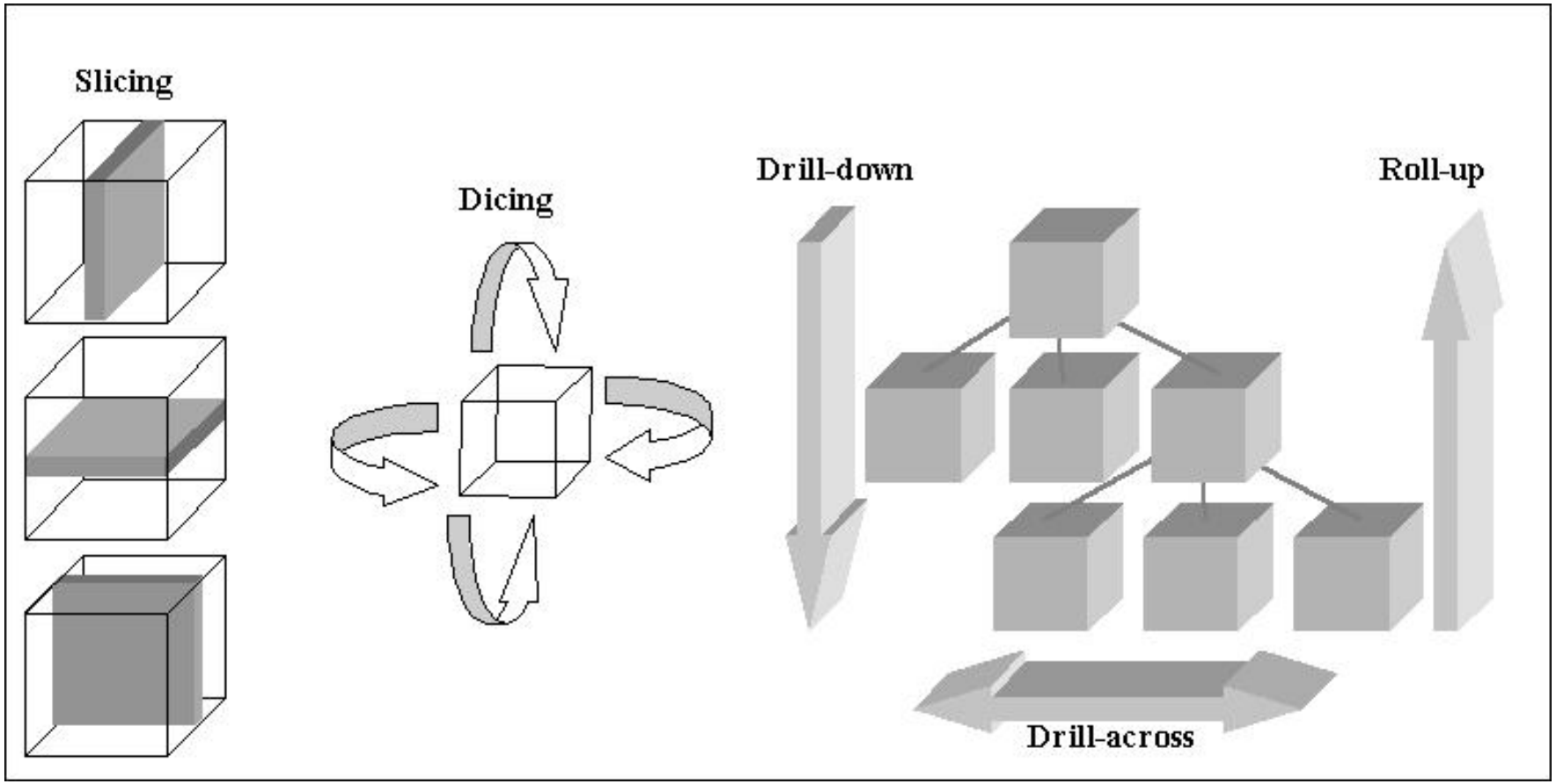


Data Provisioning

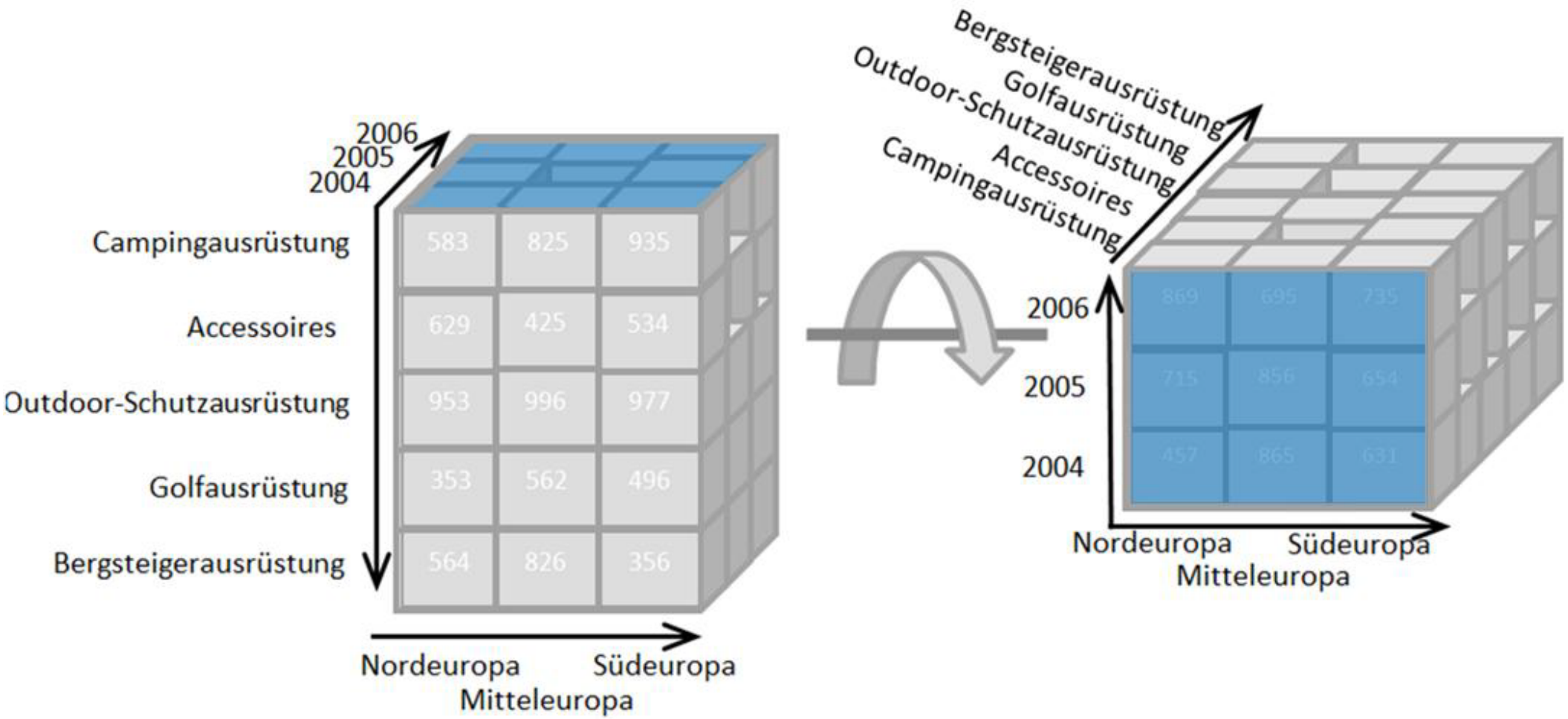
- Online analytical processing (OLAP)
  - Query multidimensional array of data
  - In database theory, an OLAP cube is an abstract representation of a projection of an RDBMS relation
    - Roll-up
    - Drill-down
    - Slice and dice
    - Pivot (rotate)

Data Provisioning

- Online analytical processing (OLAP)



Data Provisioning



Data Modelling and Analysis

- Select a combination of analytics to analyse data
- Each technique is specific to a data model and data type (a restrict set of)
- Typically data must be transformed to an analytical data format

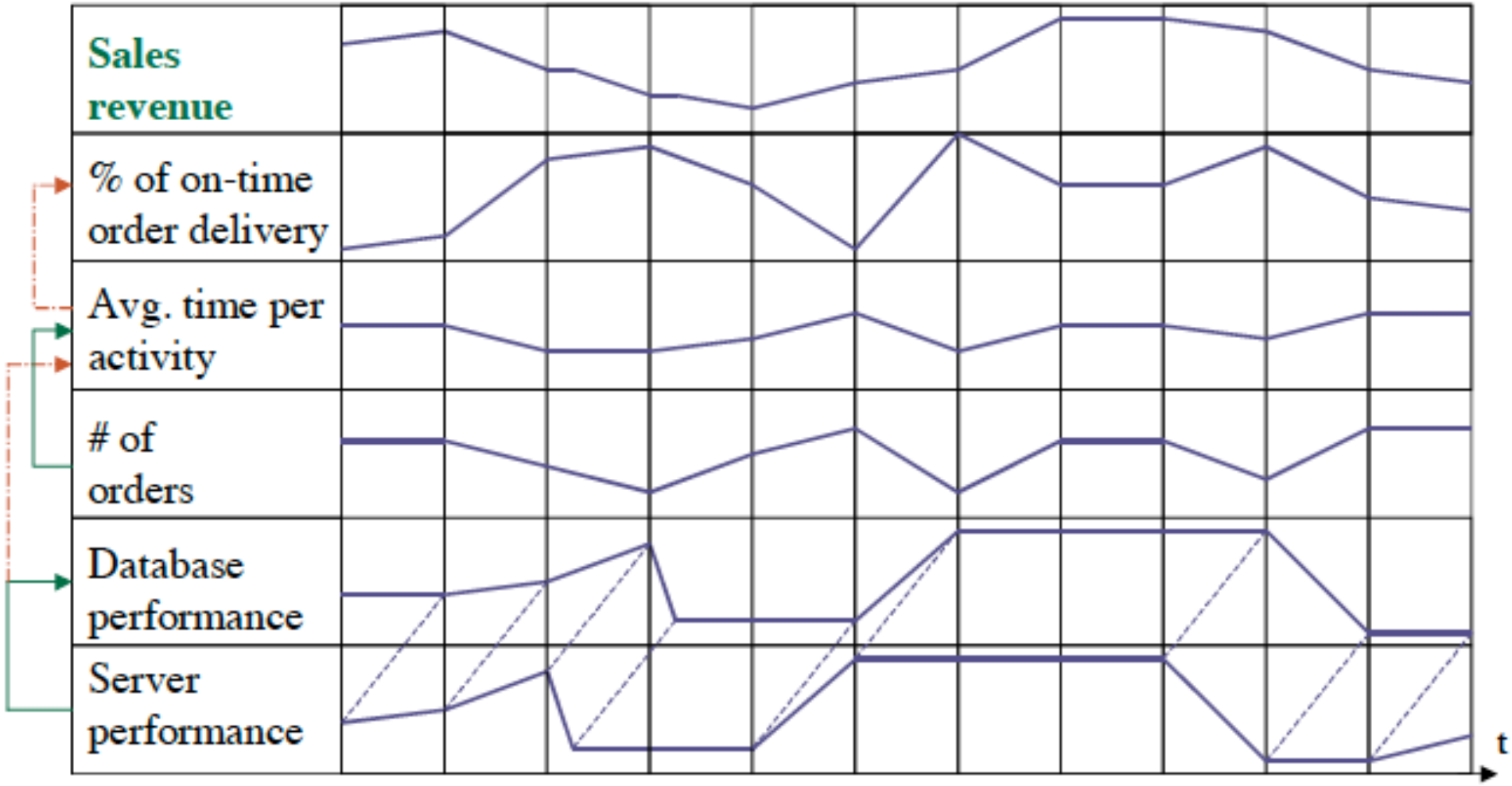
Data Modelling and Analysis

- A variety of Data Mining algorithm exists
  - Supervised - Non Supervised
  - Categorical Data - Continous Data
  - Streaming - Batch data
- We will focus on temporal data:
  - Time series
  - Sequence Mining
  - Episode Mining

Temporal Data

- A time sequence is defined as a sequence of time-stamped data with attributes values resulting from some measurement-observation
- A time series is a *time sequence* with equidistant predefined observation times (often a restrict set of events are observed over the evolution of their occurrence in a continuous time interval)
- A state sequence is a *time sequence* where the state variable *S* attains only a finite number of possible values. If the observed time are not known of minor importance we denote it a a *chain*: and ordered sequence of observations over the state variable

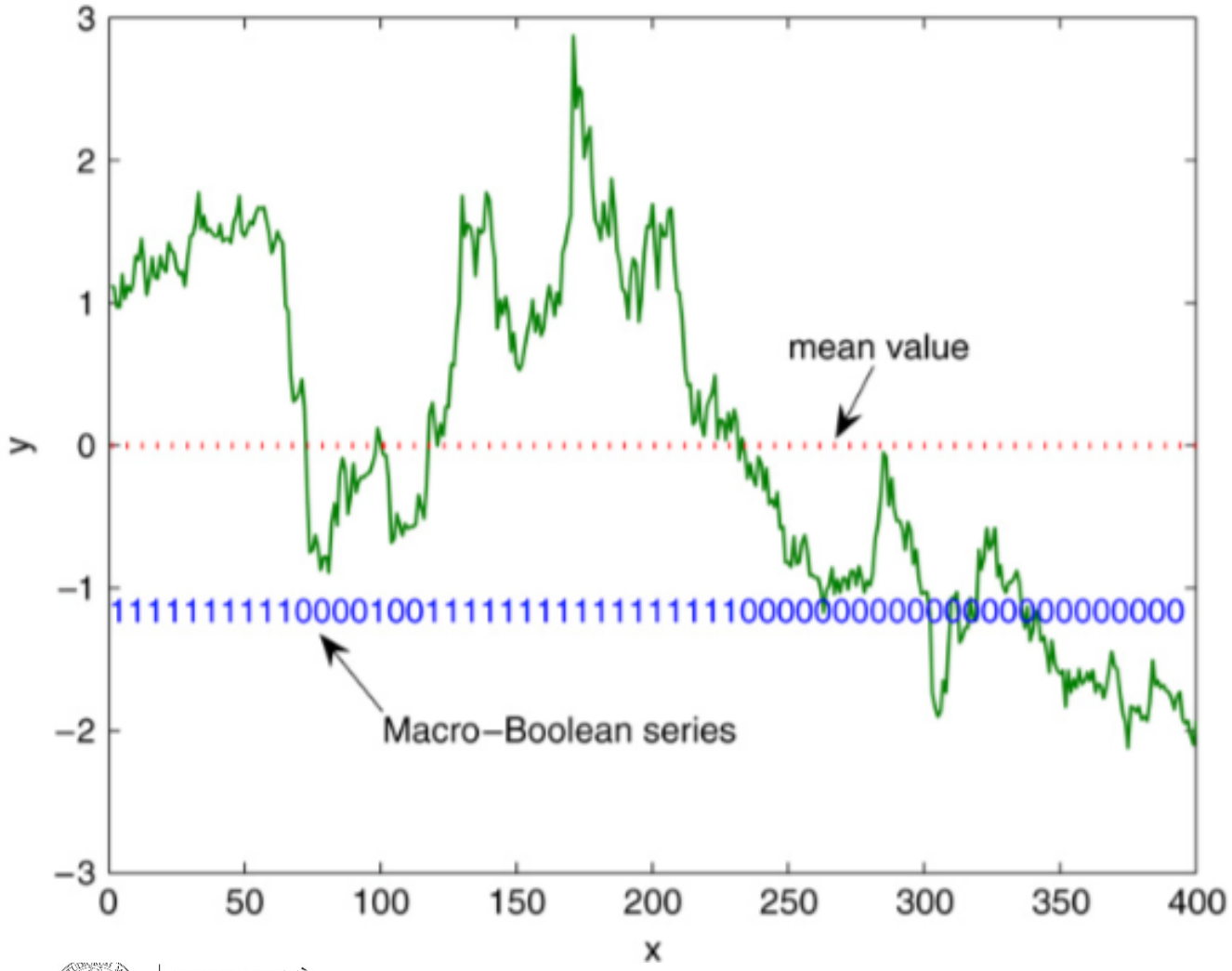
Time Series



Detecting time correlations in time-series data streams, Sayal, Mehmet, 2004.

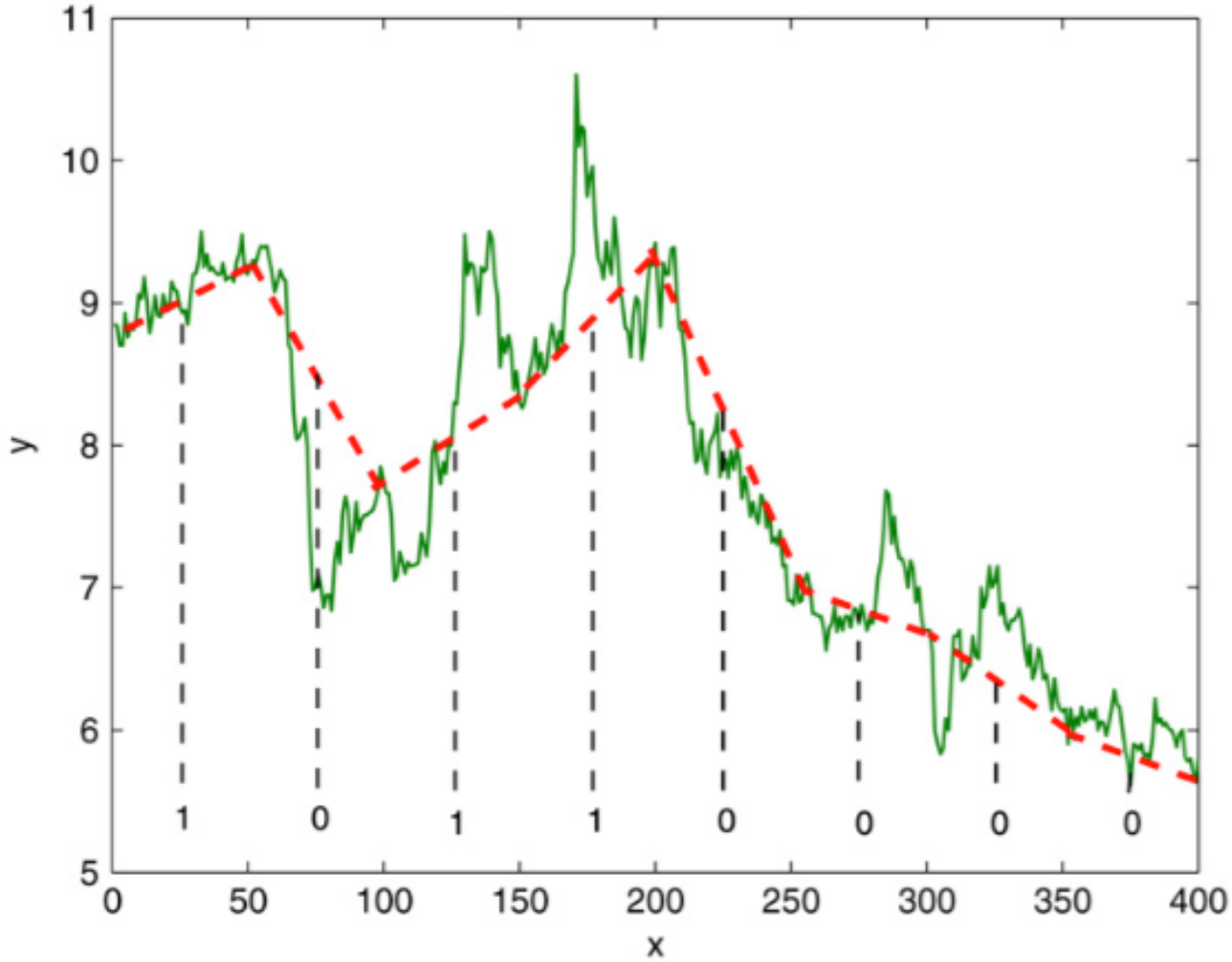
Time Series

T. Zhang et al. / Computers and Mathematics with Applications 57 (2009) 937–948

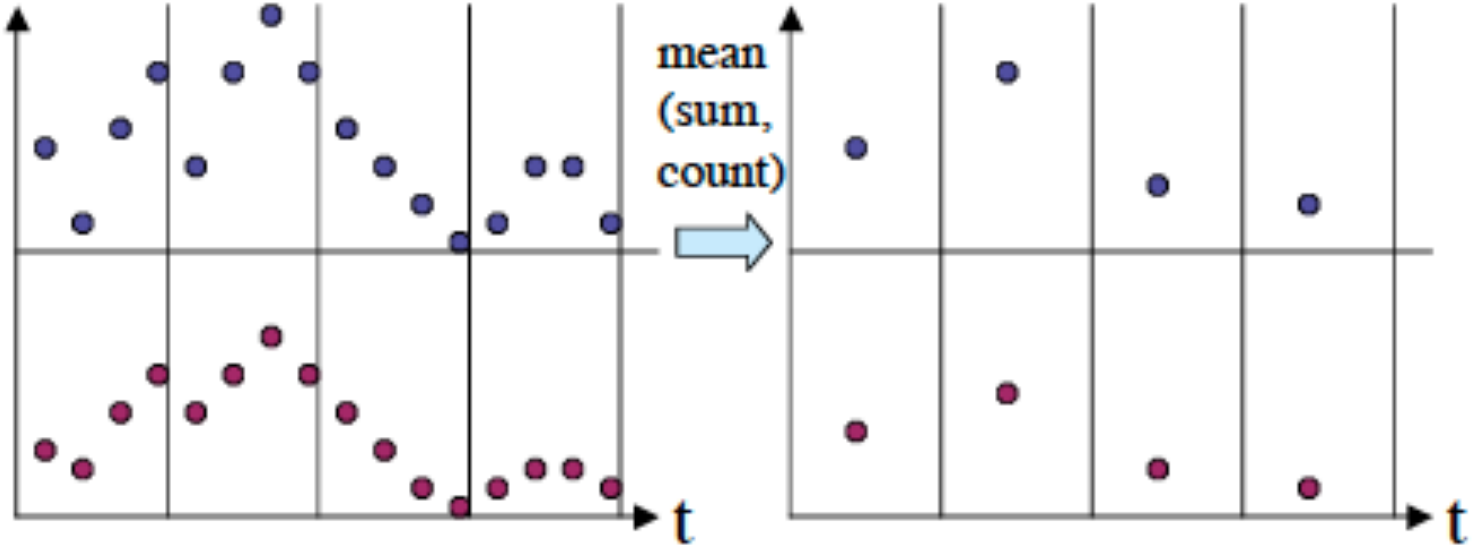


Time Series

T. Zhang et al. / Computers and Mathematics with Applications 57 (2009) 937–948

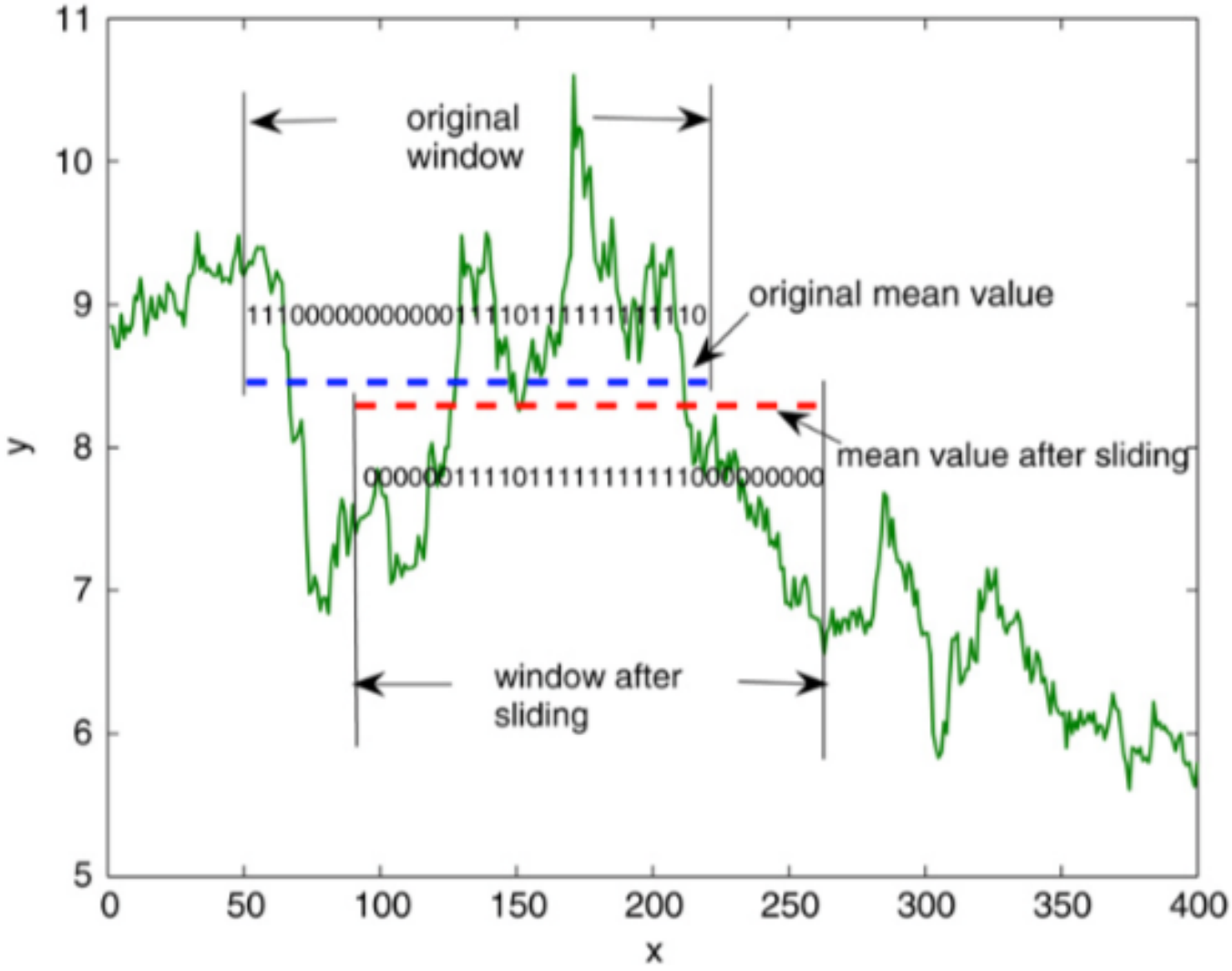


Time Series



Time Series

T. Zhang et al. / Computers and Mathematics with Applications 57 (2009) 937–948



Sequence Mining

TID	Items
1	{Pane,Latte}
2	{Pane,Pannolini,Birra,Uova}
3	{Latte,Pannolini,Birra,Cola}
4	{Pane,Latte;pannolini,Birra}
5	{Pane,Latte,Pannolini,Cola}

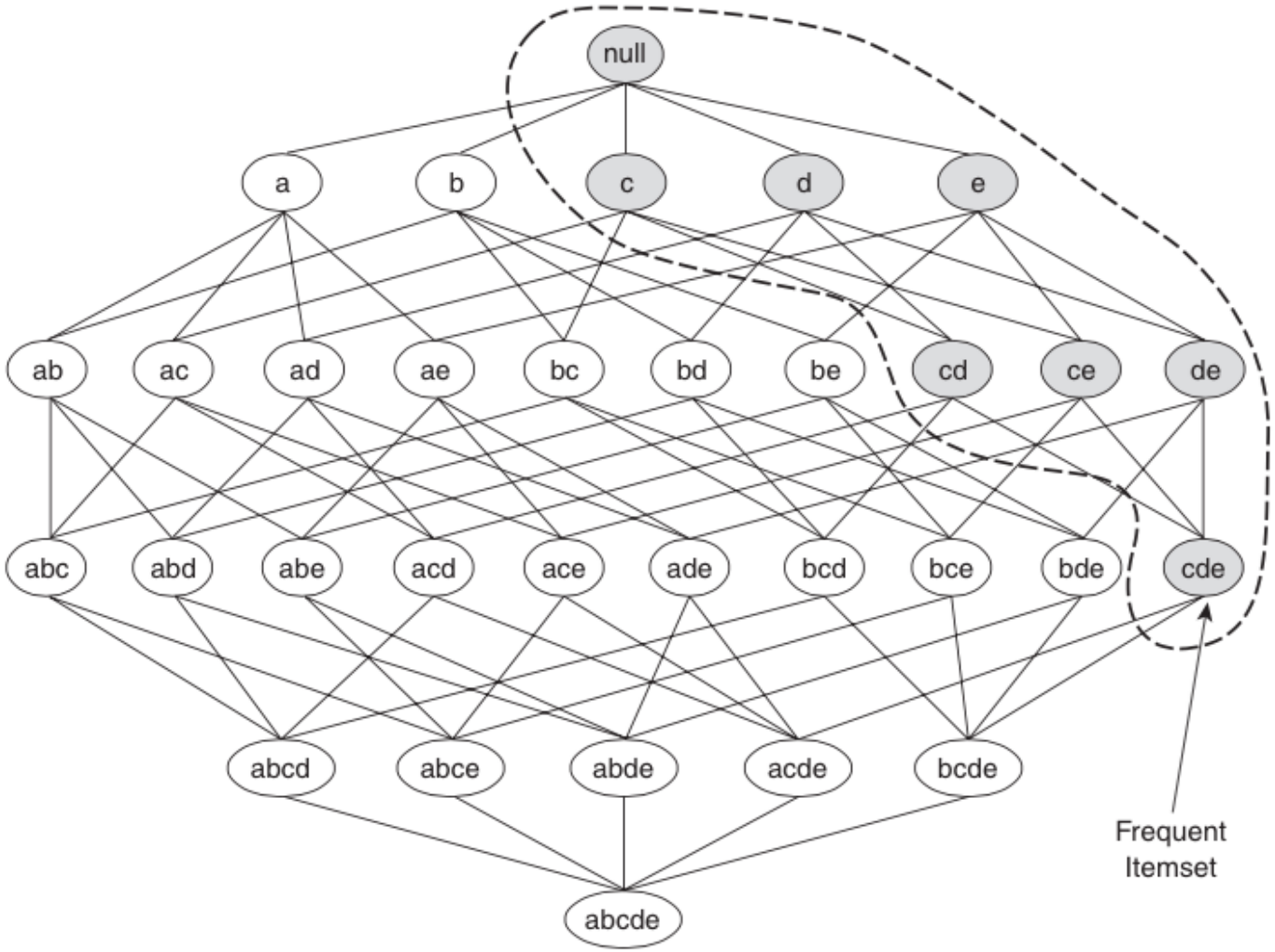
Support of an item set $\sigma(X)$

$\sigma(X) = \sum t_i | X \subseteq t_i, t_i \in T$

Sequence Mining

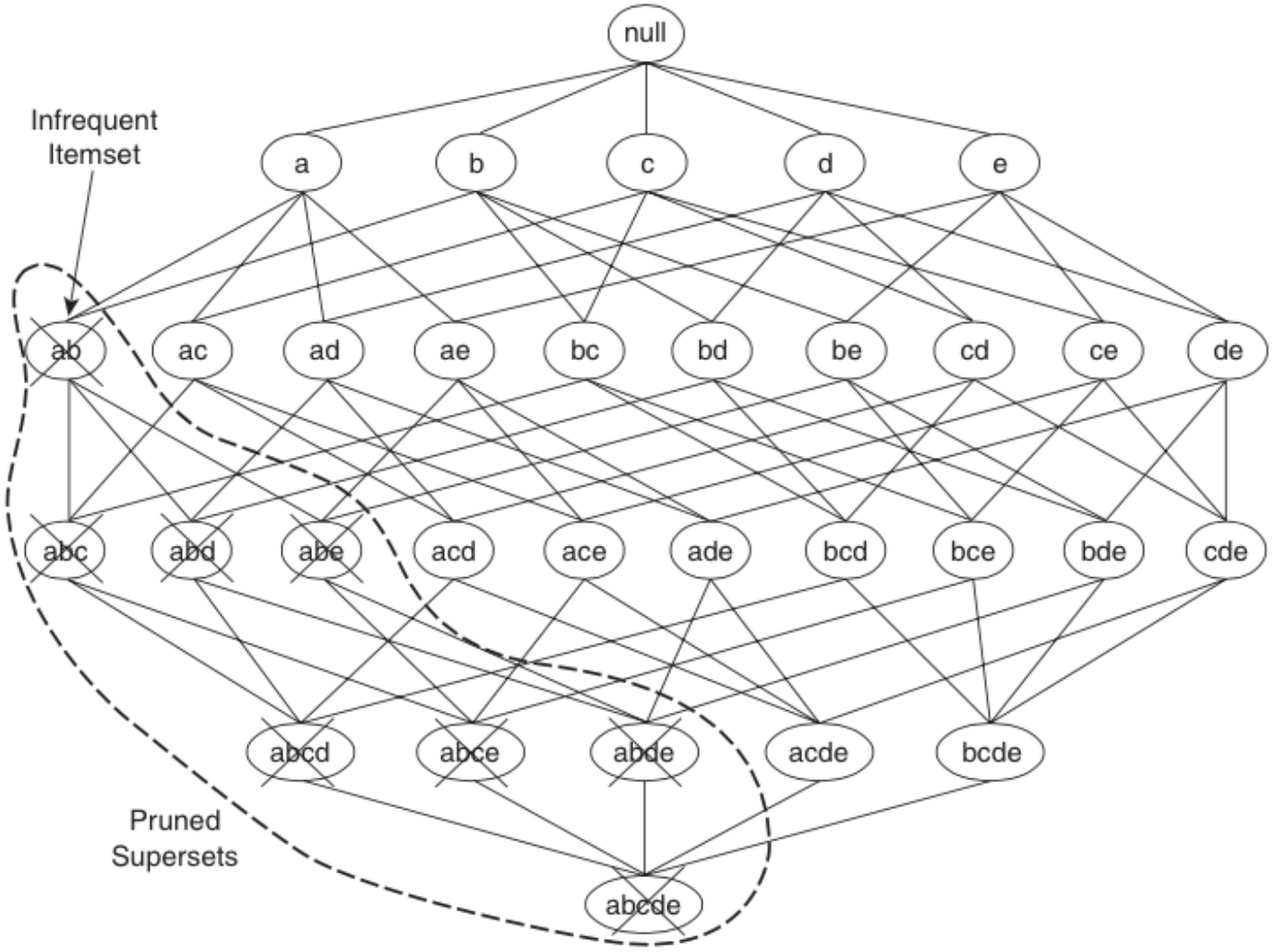
- A priori Algorithm
  - The Apriori algorithm addresses the generation phase of the frequent itemset for successive approximations, since the itemset with a single element
  - The number of iterations is 1 + k_{max}, where k_{max} is the maximum cardinality of a itemset frequent. This algorithm uses a generate-and-test approach

Sequence Mining



Suppose that {c, d, e} is a frequent itemset. Consequently, according to the Priori principle all subsets of {c, d, e} must be frequent, that is, each transaction that contains {c, d, e} also contains} {c, d}, {c, e}, {d, e}, {c}, {d}, and {e}

Sequence Mining

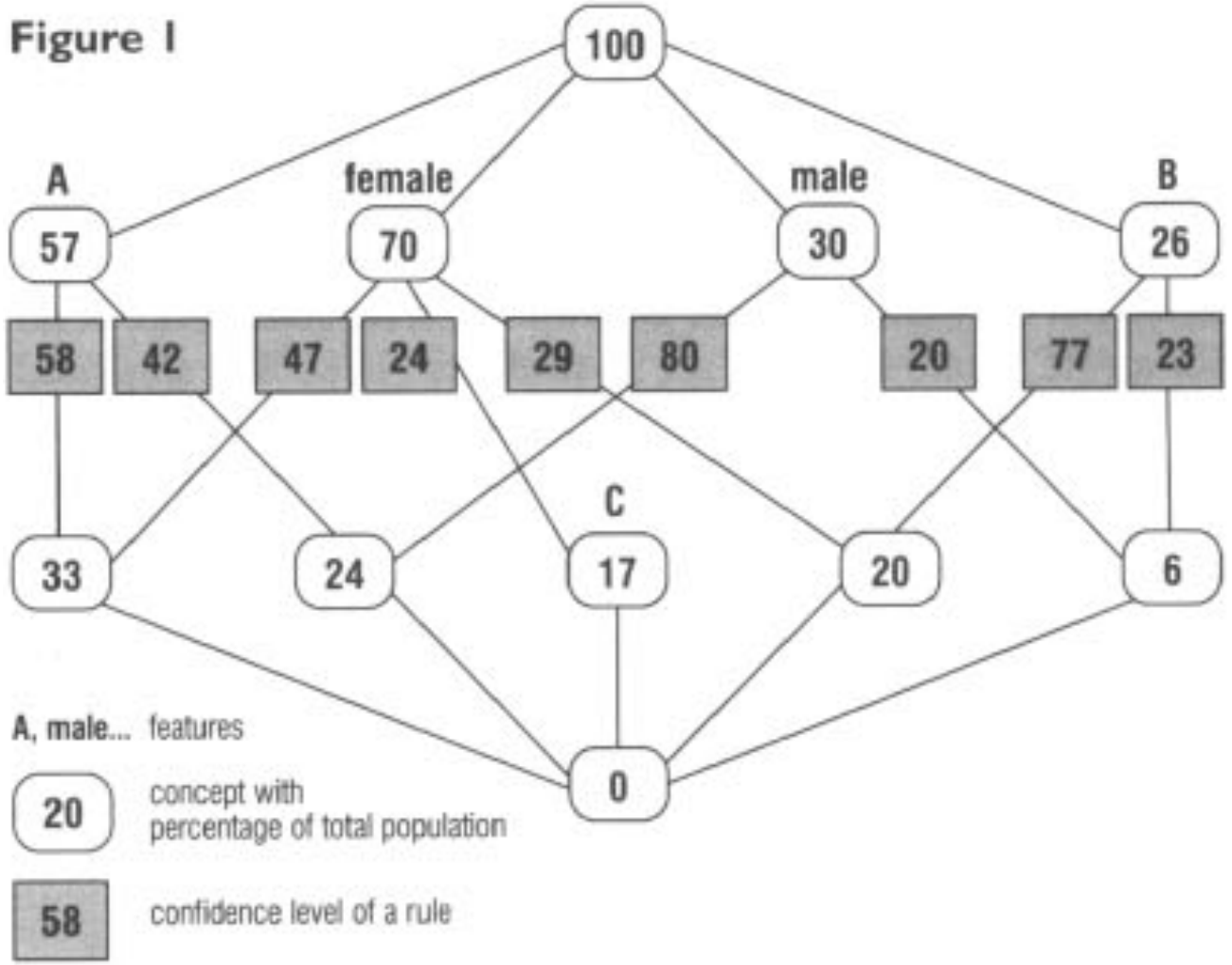


On the contrary, if we assume that {a, b} is an infrequent itemset, all itemset that contain it should not be frequent. This implies that all the itemsets containing {a, b} can be eliminated

Sequence Mining

TABLE 1			
	Male	Female	Total
A	24%	33%	57%
B	6%	20%	26%
C	0%	17%	17%
	30%	70%	100%

Sequence Mining



Sequence Mining

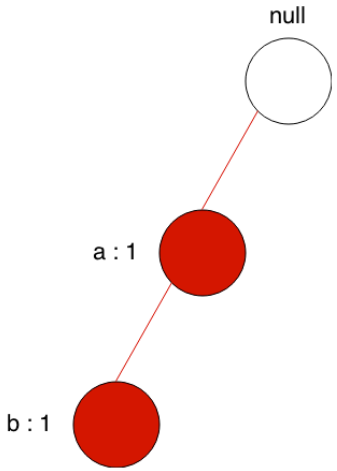
- The FP-Growth Algorithm,
  - An efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree).

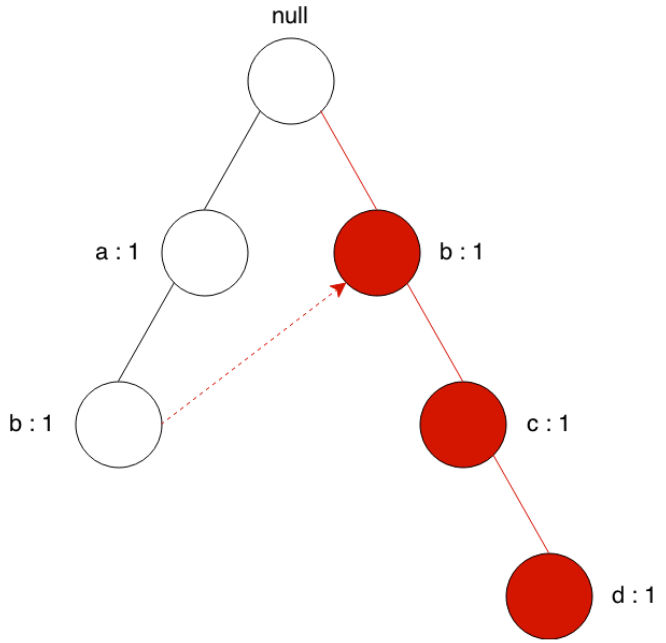
Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (SIGMOD '00). ACM, New York, NY, USA, 1-12. DOI=http://dx.doi.org/10.1145/342009.335372

Sequence Mining

Dataset

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

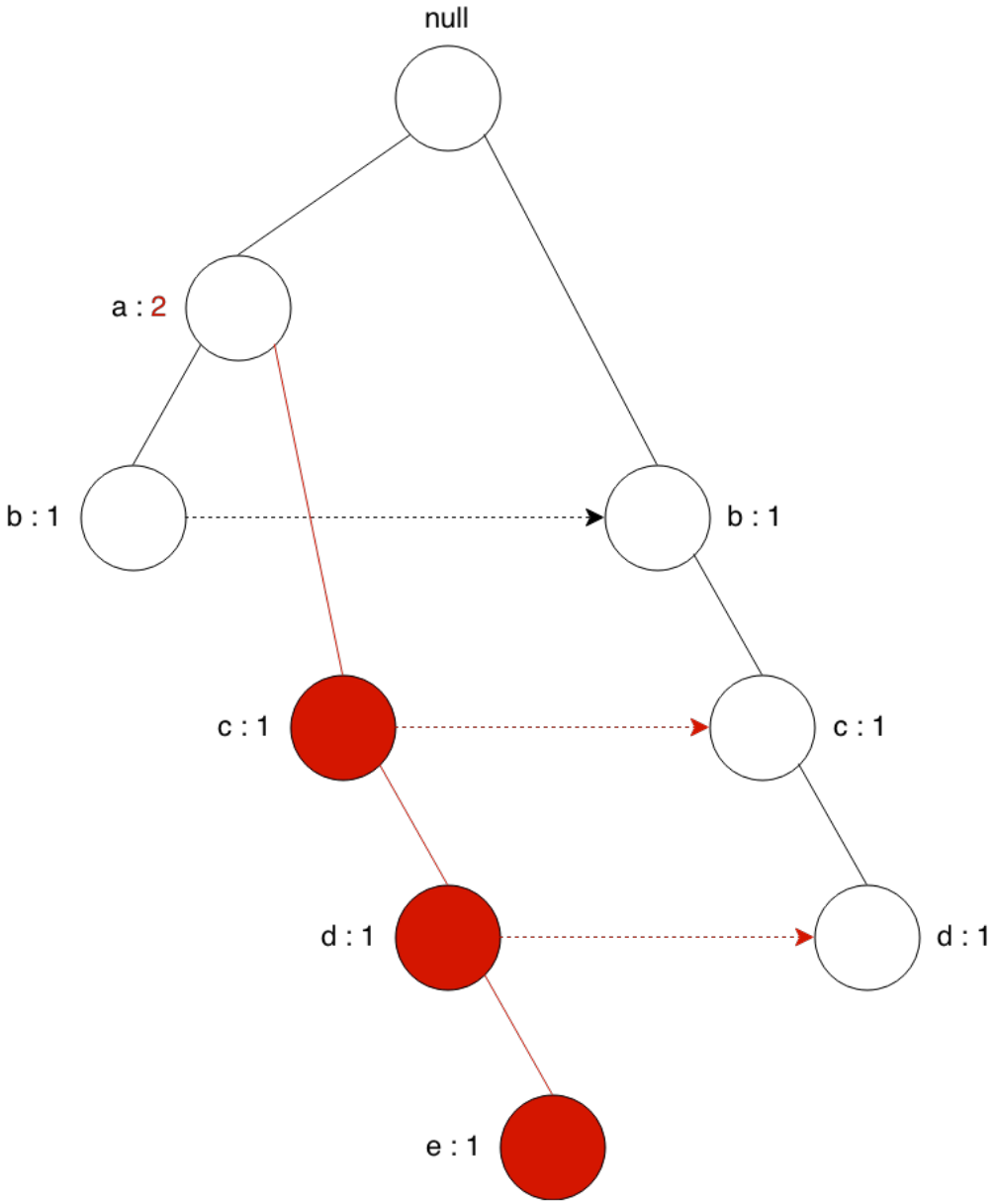




Sequence Mining

Dataset

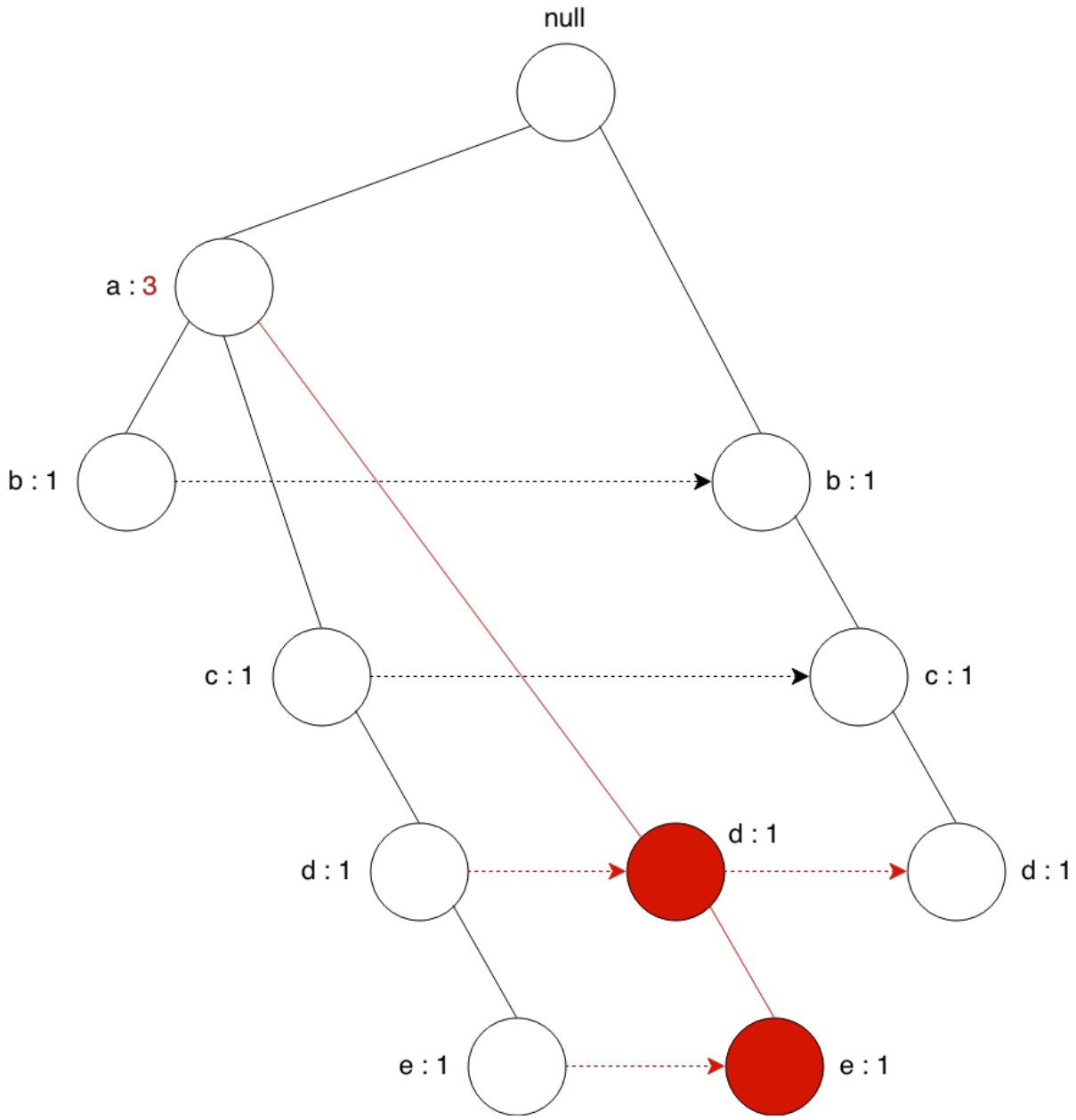
TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}



Sequence Mining

Dataset

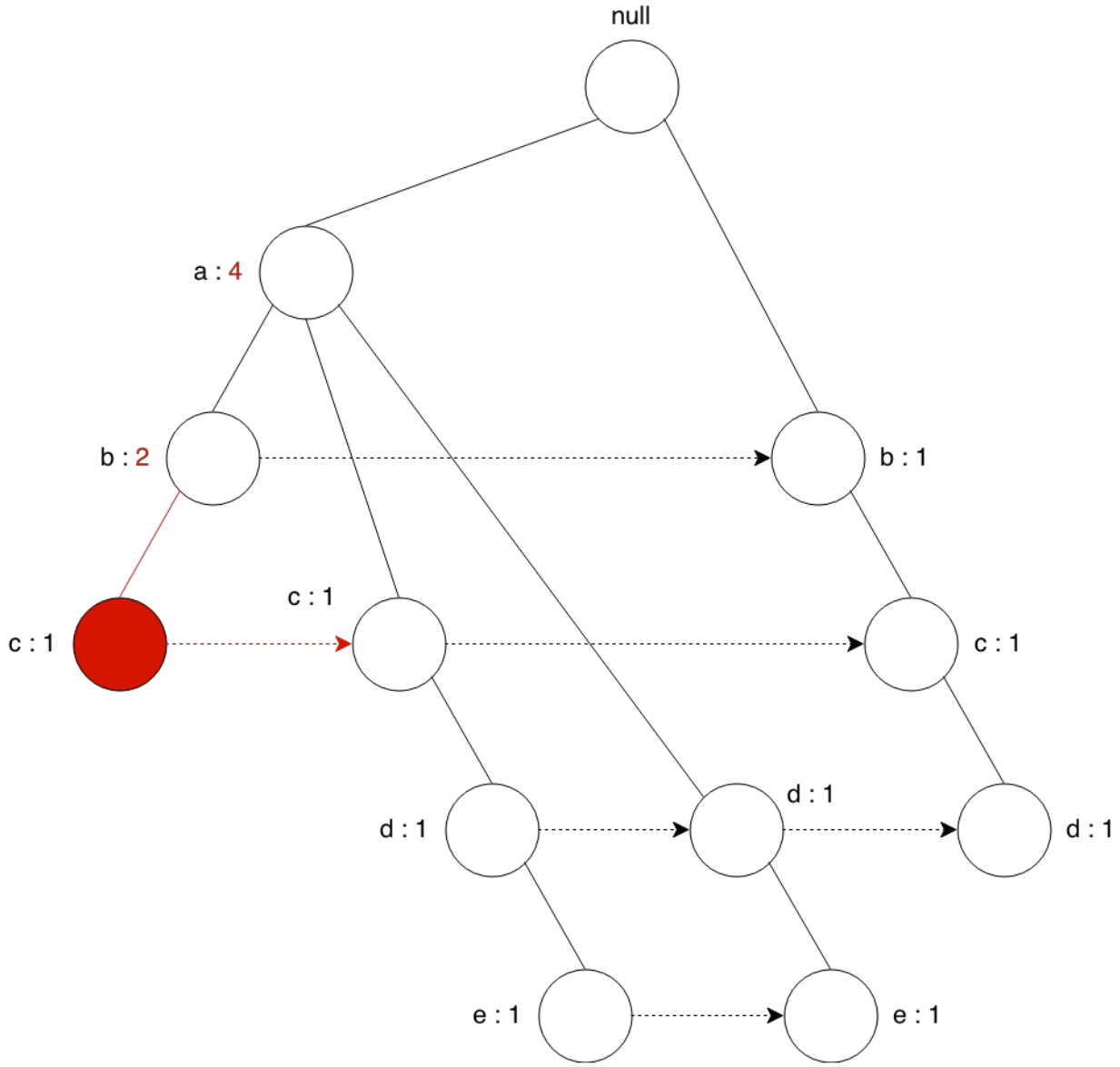
TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}



Sequence Mining

Dataset

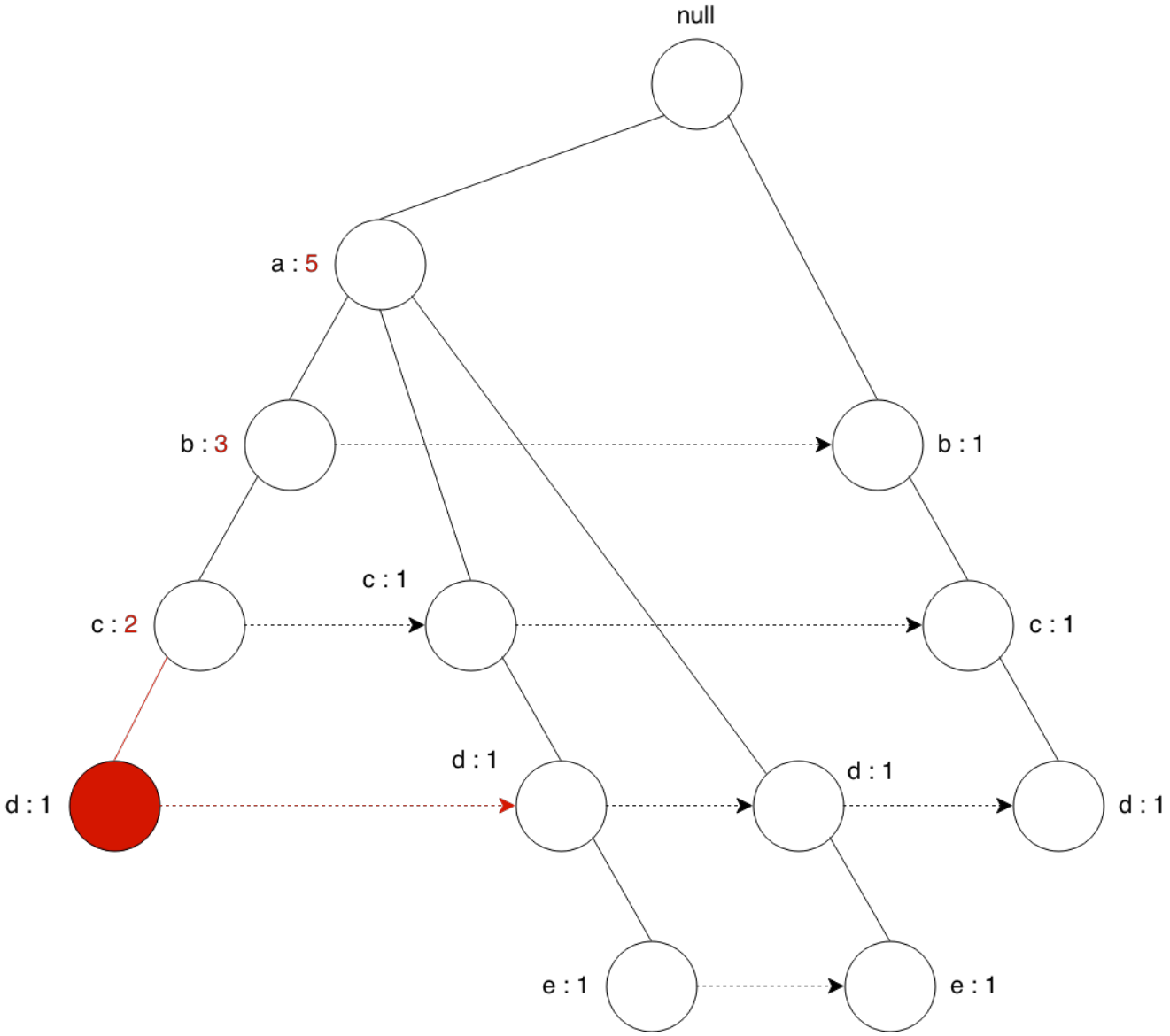
TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}



Sequence Mining

Dataset

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

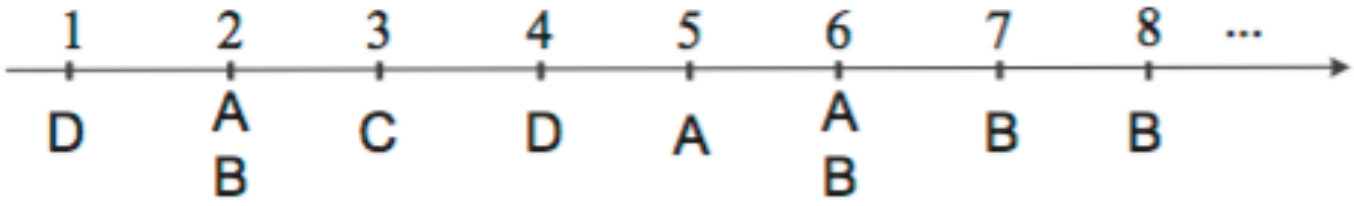


Episode Mining

- Similar to sequence mining, episode mining aims at finding frequent sequences but there are two basic differences:
  - Input Data: *sequence mining* grounds on transaction data; *episode mining* assume a continuos stream of events
  - Pattern Structure: s*equence mining* aims at finding maximum sequences of frequent item sets; *episode mining* aims at finding partially ordered sequences of events occurring together

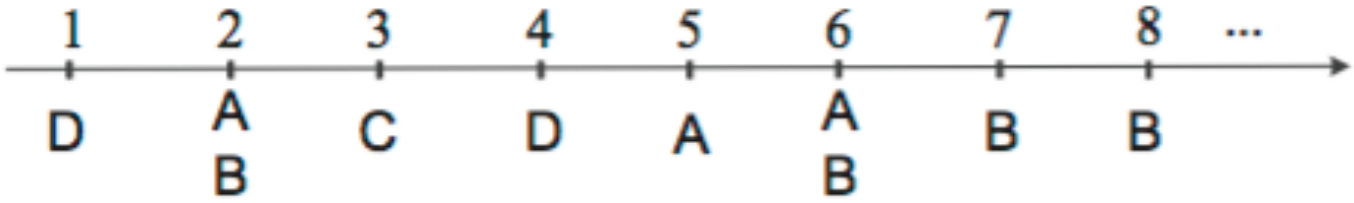
Episode Mining

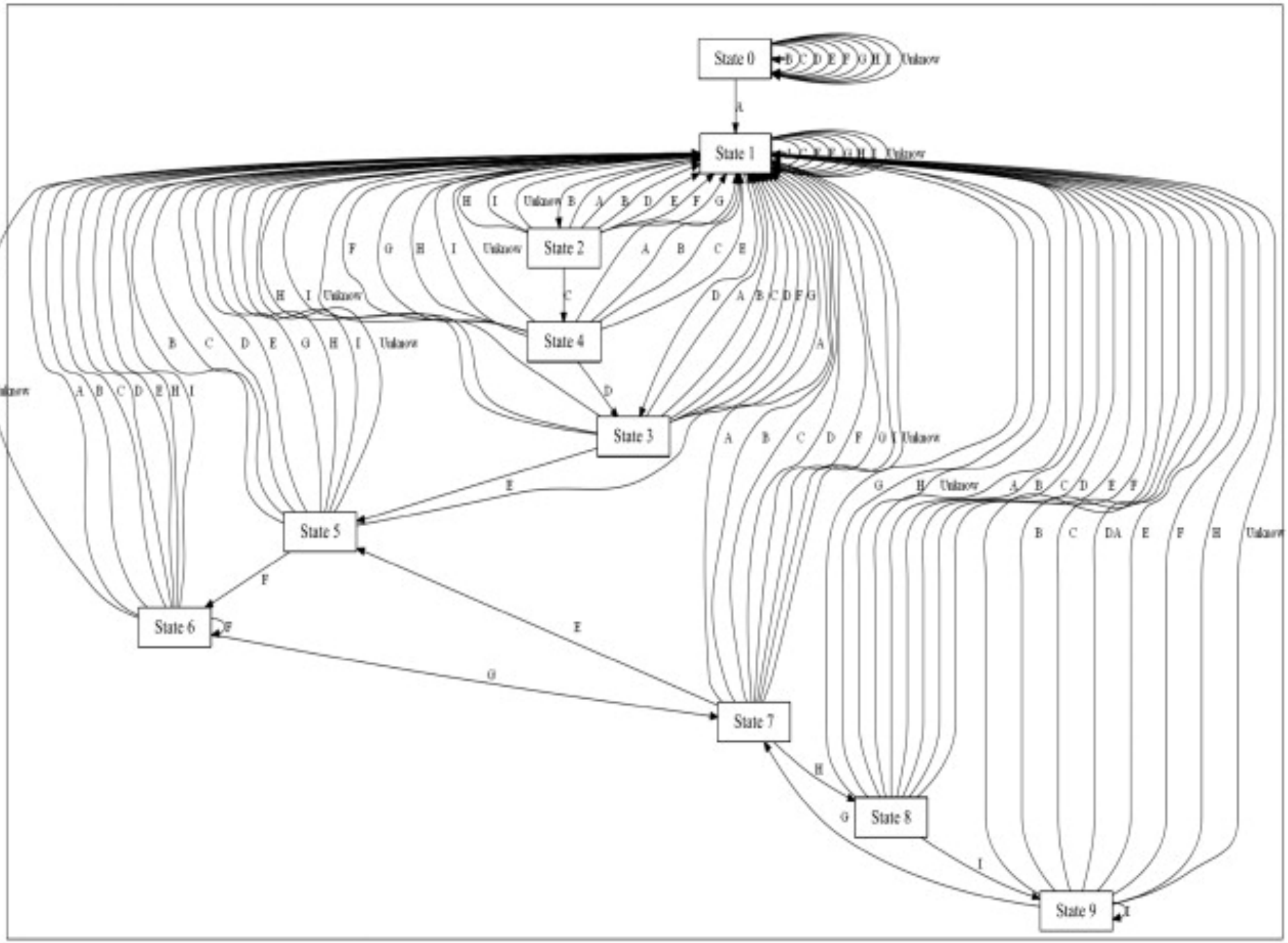
- Basically in *episode mining* the patterns detected are more structured including serial and parallel events and focusing on the partial order instead of the exact sequence



Episode Mining

- Basically in *episode mining* the patterns detected are more structured including serial and parallel events and focusing on the partial order instead of the exact sequence
  - A different notion of sub sequence is considered: $\beta = D \rightarrow C$ is a sub-episode of $D \rightarrow A \rightarrow C$, but $\beta' = A \rightarrow D$ is not.




State 0
State 1
State 2
State 3
State 4
State 5
State 6
State 7
State 8
State 9
A
B
C
D
E
F
G
H
I
Unknow

Data Reporting Knowledge Acquisition

- To increase user acceptance of a BI system, it can be advisable to consult business users at an early stage of the BIS design, for example at the requirements gathering phase
  - Acceptability
  - Usability
  - Incentive Schema

Data Reporting Knowledge Acquisition