Data mining

COSC430—Advanced Databases

David Eyers

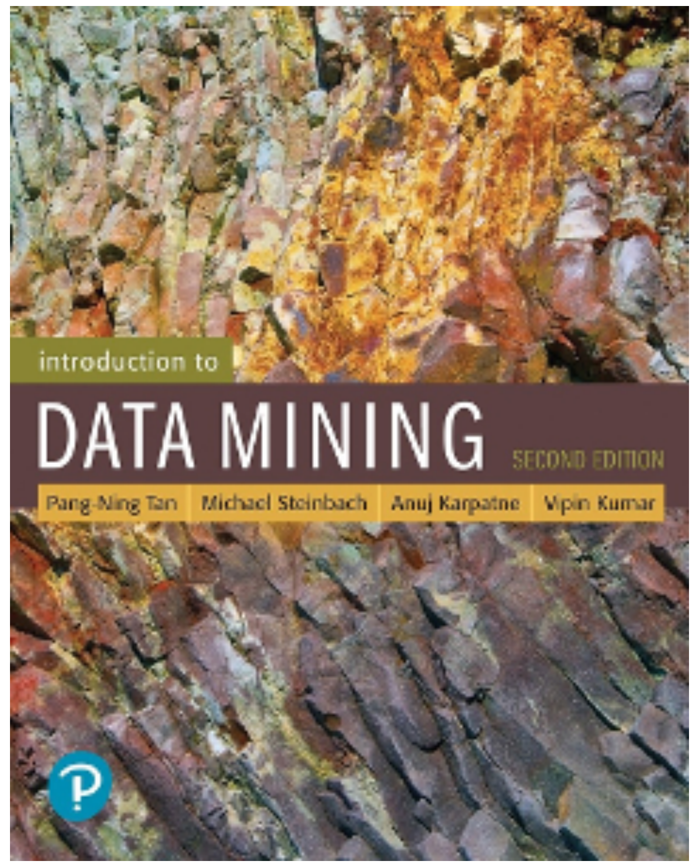
Lecture notes derived from Haibo’s, in turn derived from:
Jiawei Han, https://wiki.illinois.edu//wiki/display/cs412/2.+Course+Syllabus+and+Schedule
Vipin Kumar, https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Learning objectives

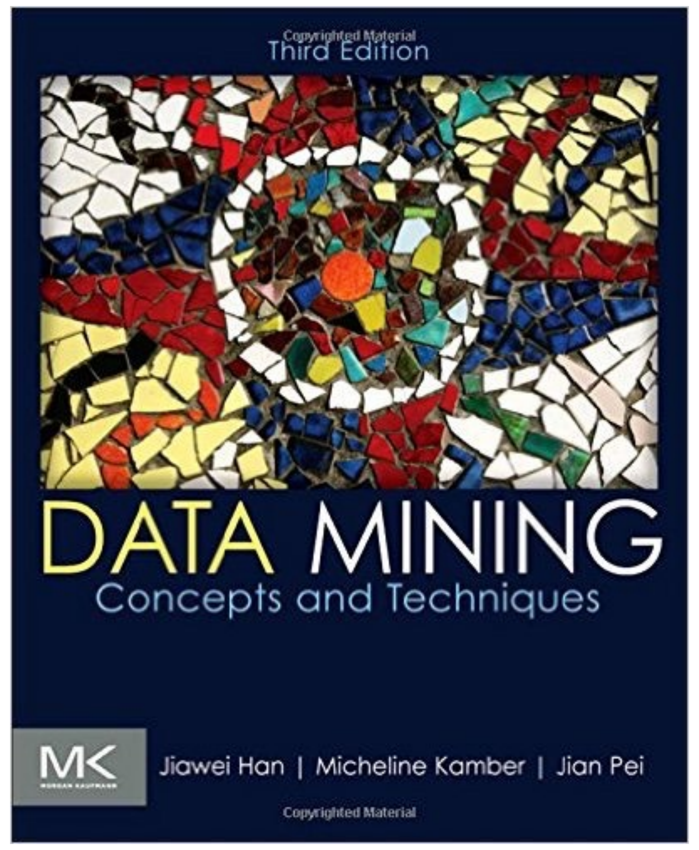
- You should:
  - understand what **data mining** is, and why we need it
  - understand the process of **knowledge discovery**
  - be able to explain what **frequent itemsets** are, and how to mine them
  - be able to distinguish and explain the difference between **classification** and **cluster** analysis
- Exploring scientific research—
  - the research paper that introduces the **Apriori** algorithm

Recommended text books on data mining

- Introduction to Data Mining (2nd Ed.)
  - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar
  - https://www.amazon.com/Introduction-Mining-Whats-Computer-Science/dp/0133128903



- Data Mining: Concepts and Techniques (3rd Ed.)
  - Jiawei Han, Micheline Kamber, and Jian Pei
  - https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1



What is data mining?

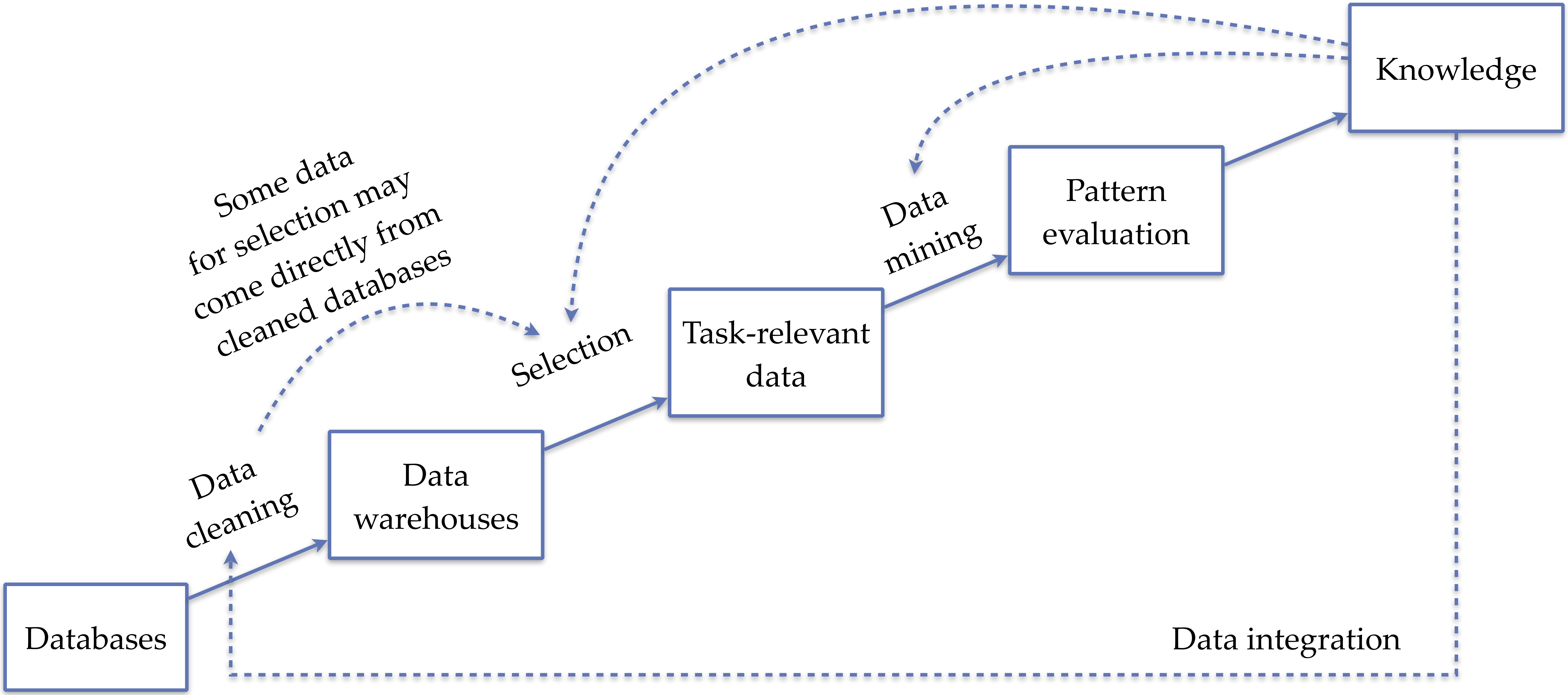
- Data mining is knowledge discovery from data
  - For example, detection of interesting patterns
- The type of knowledge discovered should be:
  - non-trivial
  - implicit
  - previously unknown
  - potentially useful

Why is data mining useful?

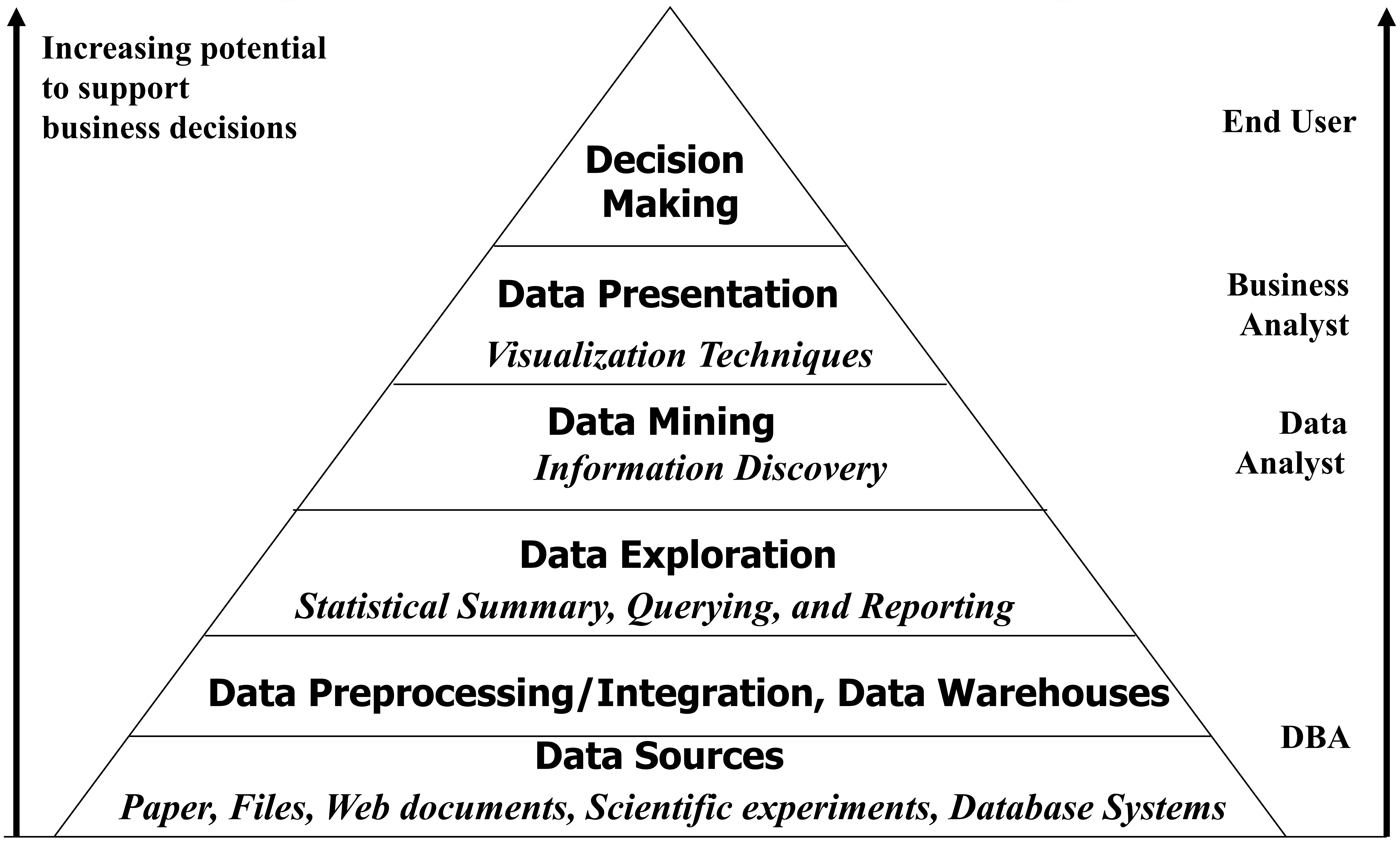
- Particularly today, there is an explosive growth of data
  - … but much of the raw information is not useful knowledge

- Data mining overlaps with many other terms:
  - knowledge discovery (mining) in databases (KDD);
  - knowledge extraction;
  - data/pattern analysis;
  - data archeology;
  - information harvesting;
  - business intelligence; …

KDD process: a view from databases



Data mining in business intelligence



KDD Process: a view from ML and stats

Data pre-processing	Data mining	Post-processing
Data integration	Pattern discovery	Pattern evaluation
Normalisation	Classification	Pattern selection
Feature selection	Clustering	Pattern interpretation
Dimension reduction	Outlier analysis	Pattern visualisation

Data preprocessing

- Data cleaning
  - Handle missing data; smooth noisy data; identify or remove outliers; and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes or files
- Data reduction
  - Dimensionality reduction; numerousity reduction (a representation that's smaller, e.g., linear regression/sampling); data compression
- Data information and discretisation (i.e., putting data into bins)
  - Normalisation (e.g. rescale min/max)
  - Concept hierarchy generation (e.g., bin address into city, then country)

Pattern discovery

- What are patterns?
  - A set of items, sub-sequences, or sub-structures that occur frequently together (or strongly correlated) in a data set
- Pattern discovery:
  - Uncovering patterns from massive data sets
- Motivating examples:
  - What products are often purchased together?
  - What are the subsequent purchases after buying an iPad?
  - What code segments are likely to contain copy-paste bugs?
- Broad applications
  - Cross-marketing, web log analysis, biological sequence analysis, etc.

Basic concepts: k-itemset; abs/rel. support

Tid	Items bought
10	beer, nuts, nappies
20	beer, coffee, nappies
30	beer, nappies, eggs
40	nuts, eggs, milk
50	nuts, coffee, nappies, eggs, milk

- **Itemset**: a set of one or more items
- **k-itemset**: $X=\{x_1,\ldots x_k\}$
  - e.g. {beer, nuts, nappies} is 3-itemset
- **sup{X}**: absolute support of X
  - Frequency or number of occurrences of itemset X
    - e.g., sup{beer} = 3, sup{beer,eggs} = 1
- **s{X}**: relative support of X
  - Fraction of items that contain X
    - e.g., s{beer} = 3/5=60%; s{beer,eggs}=1/5=20%

Frequent itemsets (patterns)

Tid	Items bought
10	beer, nuts, nappies
20	beer, coffee, nappies
30	beer, nappies, eggs
40	nuts, eggs, milk
50	nuts, coffee, nappies, eggs, milk

- An itemset (or a pattern) X is frequent if the support of X is no less than a **minsup** threshold σ
- For given dataset with σ=50%
  - All frequent 1-itemsets:
    - {beer}:3/5=60%; {nuts}:3/5=60%; {nappies}:4/5=80%; {eggs}:3/5=60%
  - All frequent 2-itemsets:
    - {beer, nappies}:3/5=60%
  - All frequent 3-itemsets? There are none

From frequent itemsets to association rules

Tid	Items bought
10	beer, nuts, nappies
20	beer, coffee, nappies
30	beer, nappies, eggs
40	nuts, eggs, milk
50	nuts, coffee, nappies, eggs, milk

- Rules more useful than itemsets alone
  - e.g. mining the "rule", nappies→beer
    - i.e., buying nappies implies will also buy beer
- How strong is this rule?
  - Look at **support** (**s**) and **confidence** (**c**)
  - Measuring association rules X→Y (s,c) for itemsets X and Y
  - Support s: probability item will contain X∪Y (i.e., union of both itemsets)
    - s{nappies, beer} = 3/5 = 60%
  - Confidence c: conditional probability Tid containing X also contains Y
    - c = sup(X∪Y) / sup(X)
    - e.g., c{nappies,beer} = sup{nappies,beer}/sup{nappies} = 3/4 = 75%

Mining frequent itemsets & assoc. rules

Tid	Items bought
10	beer, nuts, nappies
20	beer, coffee, nappies
30	beer, nappies, eggs
40	nuts, eggs, milk
50	nuts, coffee, nappies, eggs, milk

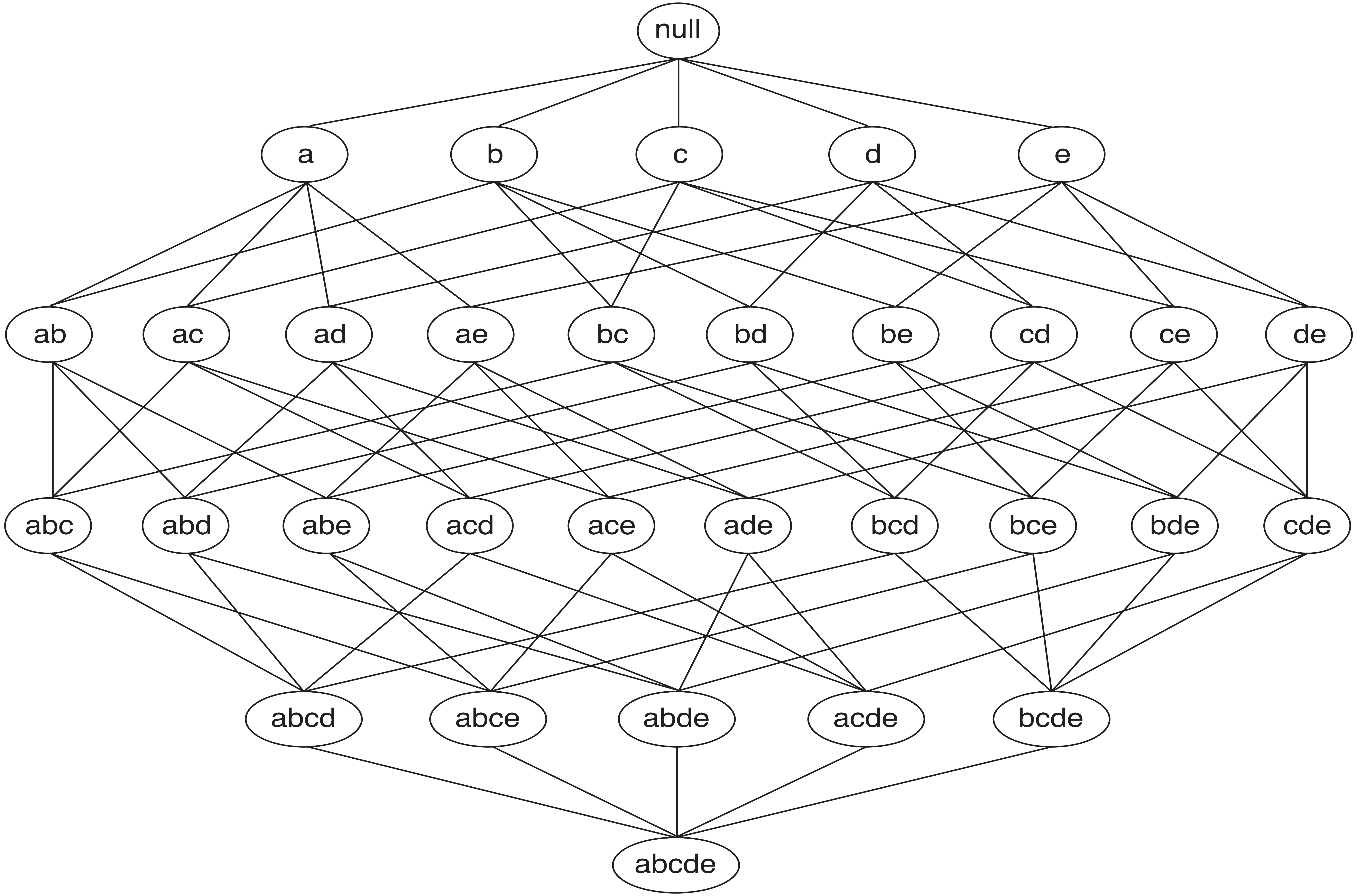
- Association rule mining; find all rules:
  - Given two thresholds: minsup, minconf
  - $X \rightarrow Y$ (s,c) s>=minsup, c>=minconf
- For example, let minsup=50%
  - freq. 1-itemsets: beer:3; nuts:3; nappies:4; eggs:3
  - freq. 2-itemsets: {beer,nappies}:3
- Then let minconf=50%
  - beer→nappies (60%,100%)
  - nappies→beer (60%, 75%)
- Observations:
  - Mining association rules and mining frequent patterns are close problems
  - Scalable methods needed to mine large datasets

Challenge: too many frequent patterns!

- A long pattern has a combinatorial number of sub-patterns
- How many frequent itemsets does the following contain?
  - T_1: $\{a_1, \ldots, a_{50}\}$; T_2:$\{a_1, \ldots, a_{100}\}$
  - Let's have a try if we assume (absolute) minsup = 1
    - 1-itemsets: $\{a_1\}$:2, $\{a_2\}$:2, …, $\{a_{50}\}$:2, $\{a_{51}\}$:1, …, $\{a_{100}\}$:1
    - 2-itemsets: $\{a_1,a_2\}$:2, …, $\{a_1,a_{50}\}$:2, $\{a_1,a_{51}\}$:1, …, $\{a_{99},a_{100}\}$:1,
    - …
    - 99-itemsets: $\{a_1,a_2,\ldots,a_{99}\}$:1, …, $\{a_2,a_3,\ldots,a_{100}\}$:1
    - 100-itemsets: $\{a_1,a_2,\ldots,a_{100}\}$:1
- The total number of frequent itemsets: (i.e. a really large number!)

$$\binom{100}{1} + \binom{100}{2} + \binom{100}{3} + \cdots + \binom{100}{100} = 2^{100} - 1$$

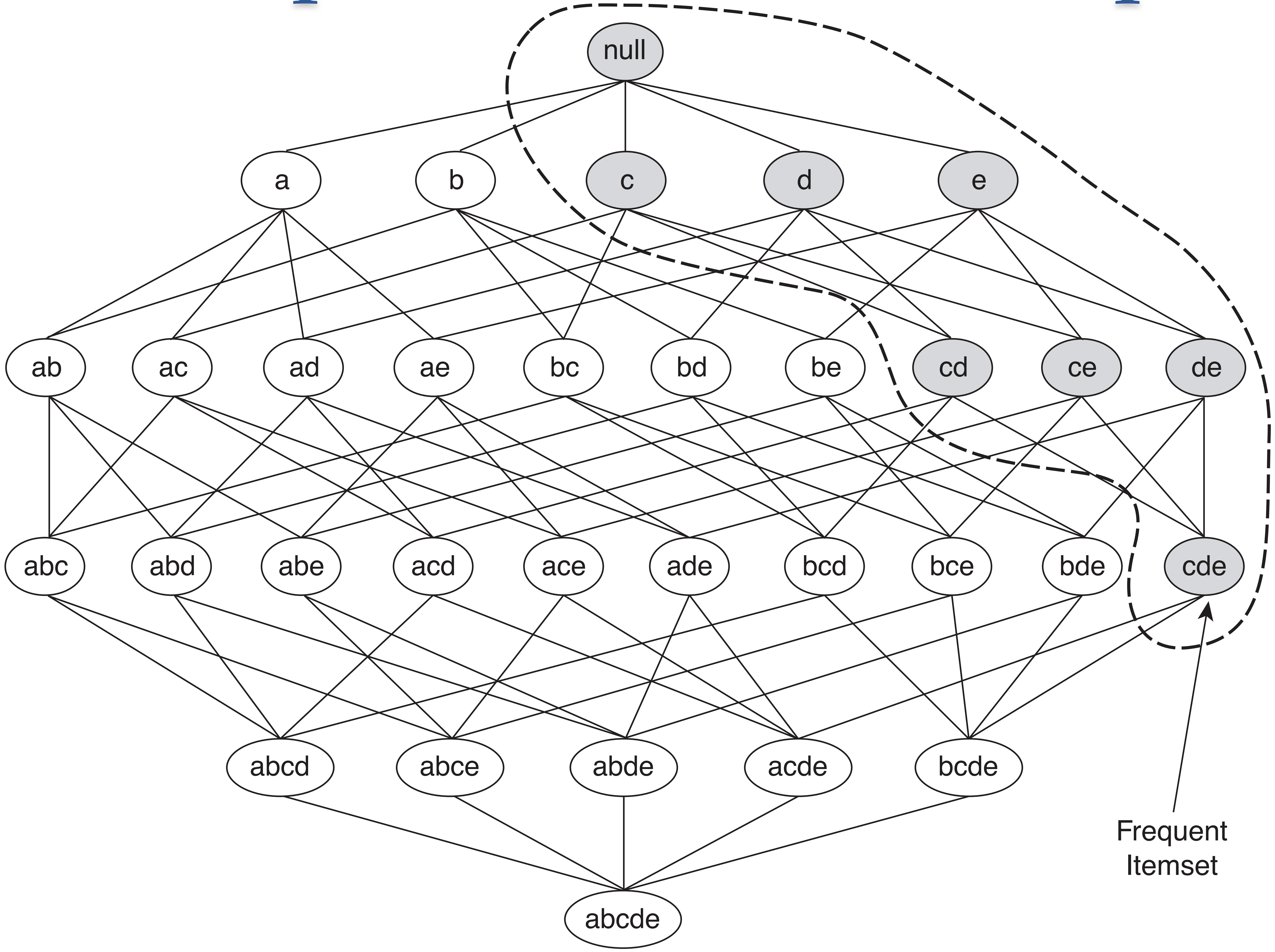
Challenge: too many frequent patterns!



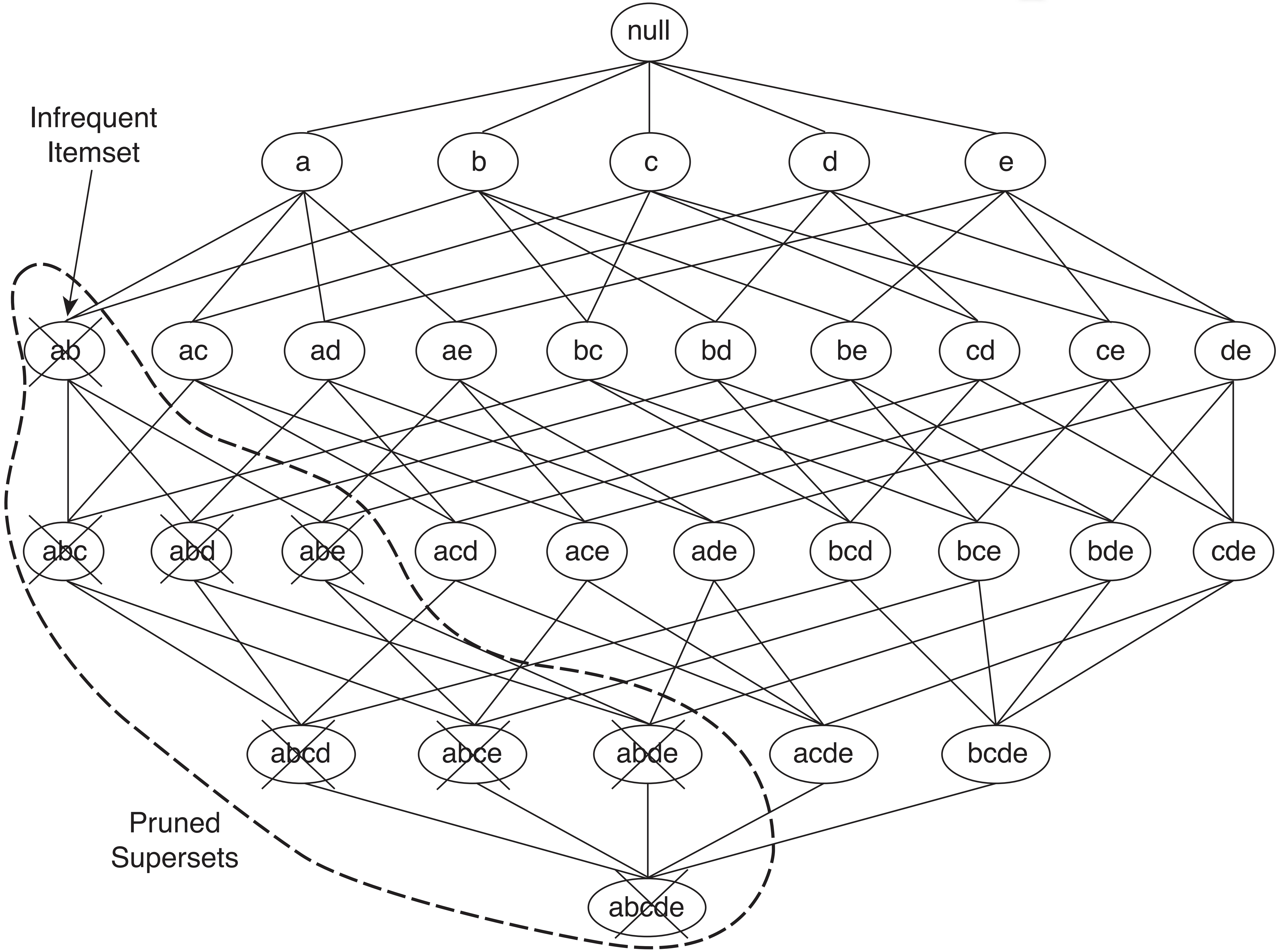
Efficient pattern mining methods

- The **downward closure** (also called "Apriori") property
  - if {beer,nappies,nuts} is frequent, so is {beer,nappies}
  - i.e., any subset of a frequent itemset must be frequent
- **Apriori pruning principle**
  - if any subset of an itemset S is infrequent, then there is no chance for S to be frequent
- Major approaches:
  - Level-wise, join-based approach: Apriori (Agrawal & Srikant, 1994)
  - Frequent pattern projection & grown: FPgrowth (Han, et al., 2000)
  - Vertical data format approach: Eclat (Zaki, et al., 1997)

Subsets of frequent items are frequent



Freq. itemsets can't have infrequent subsets

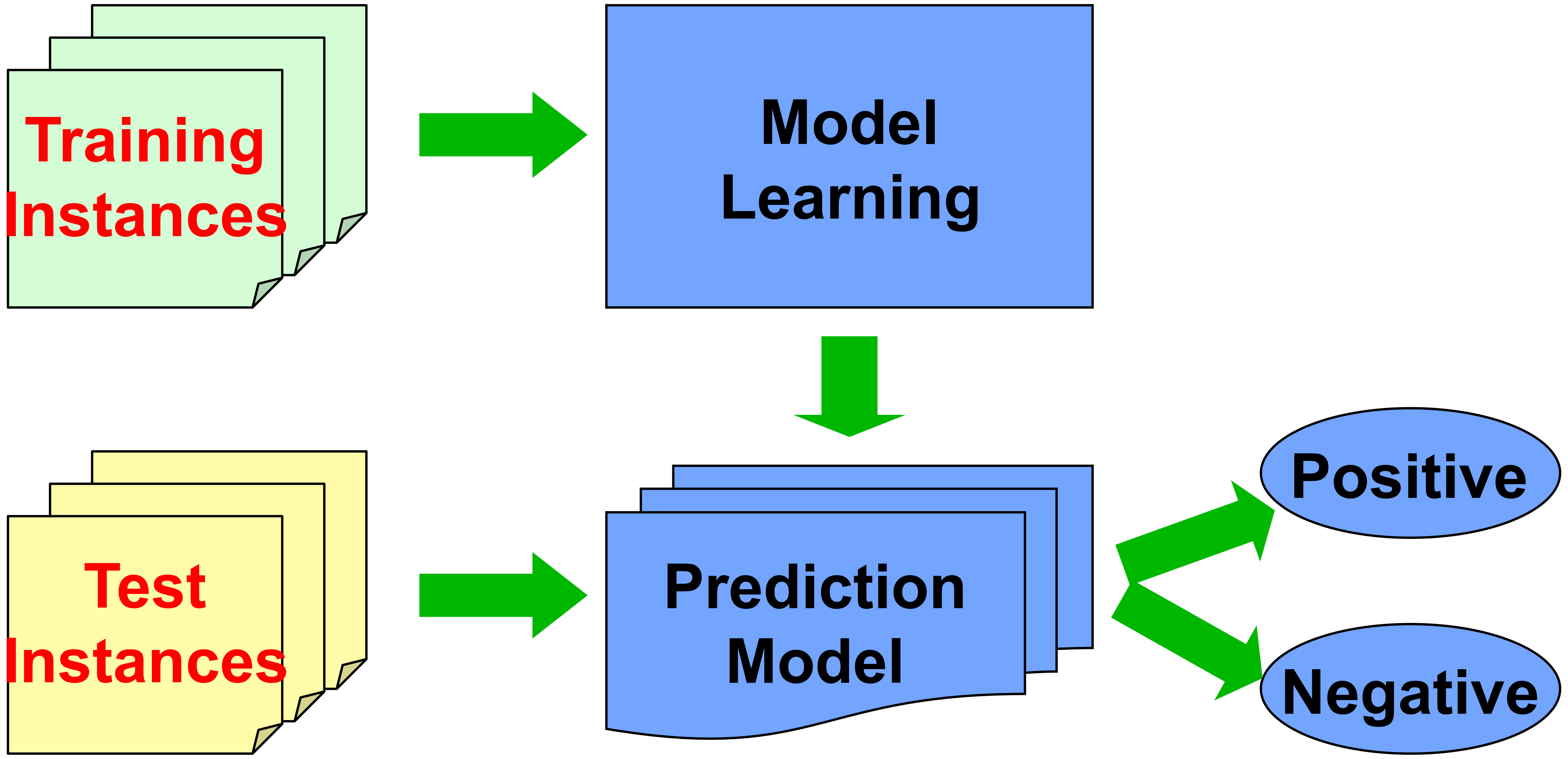


Classification

- Supervised learning
  - Training data (observations, measurements) accompanied by labels indicating the classes to which they belong
  - New data is classified using models built from training set

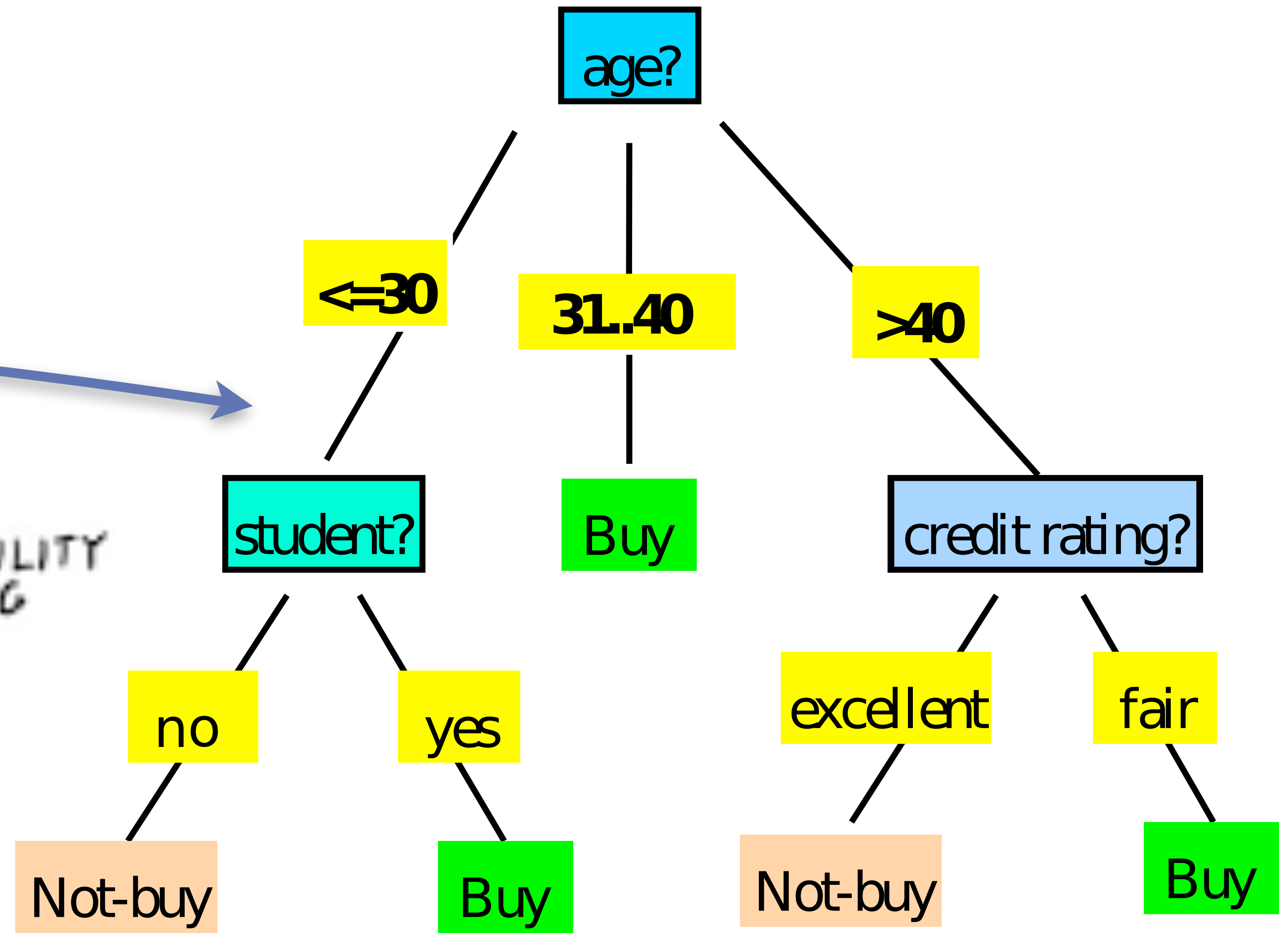
Training Data with class label:

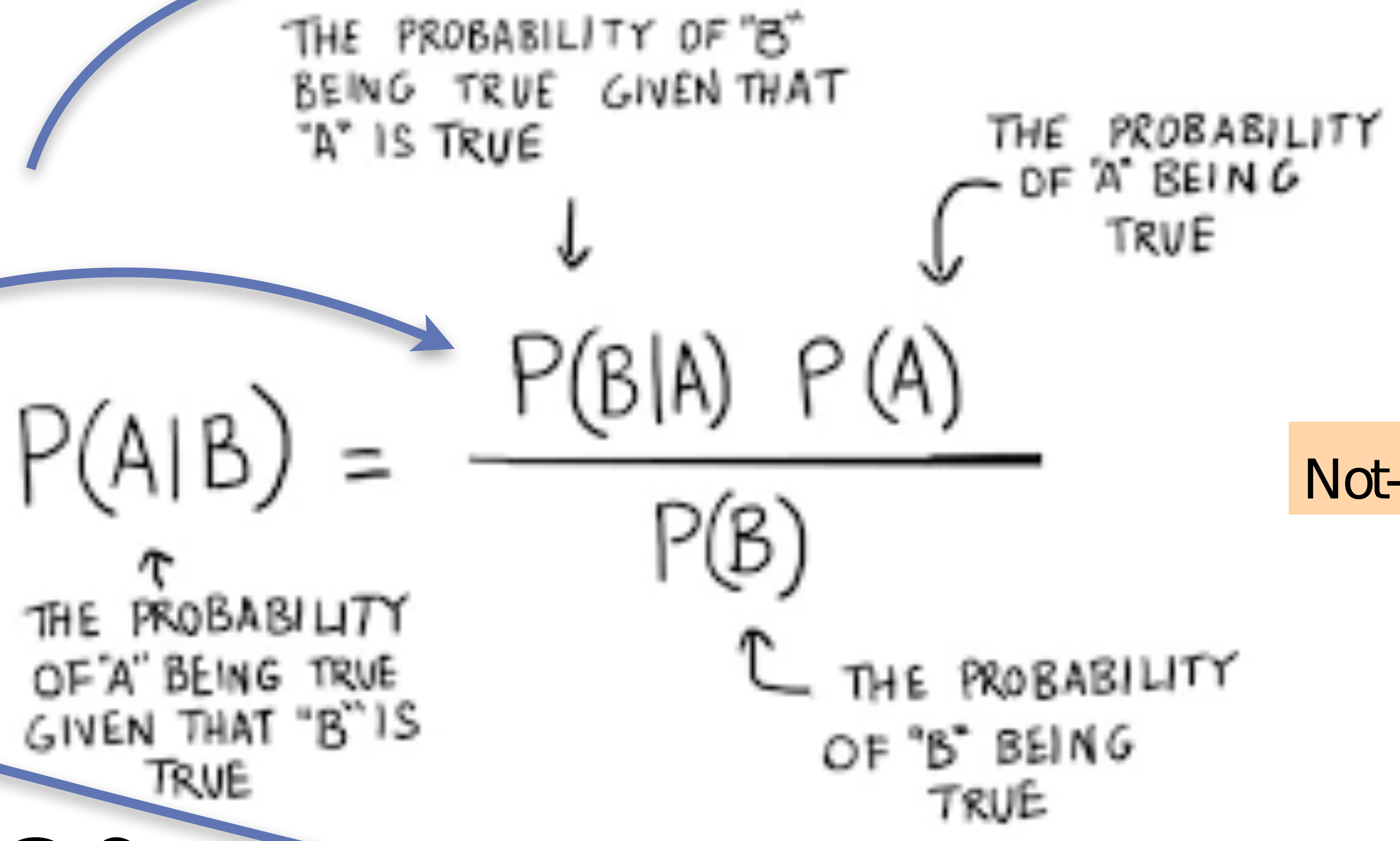
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

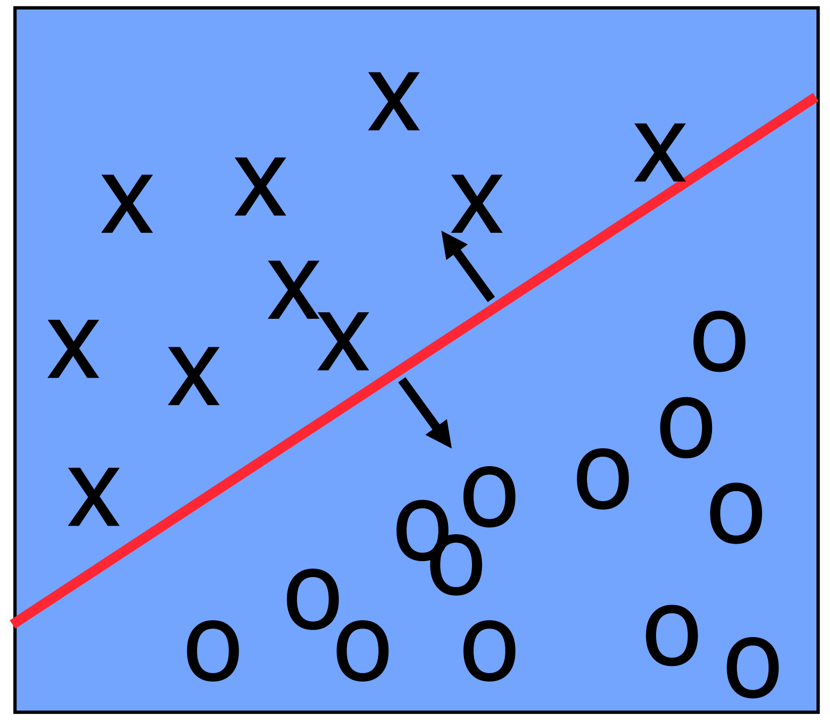


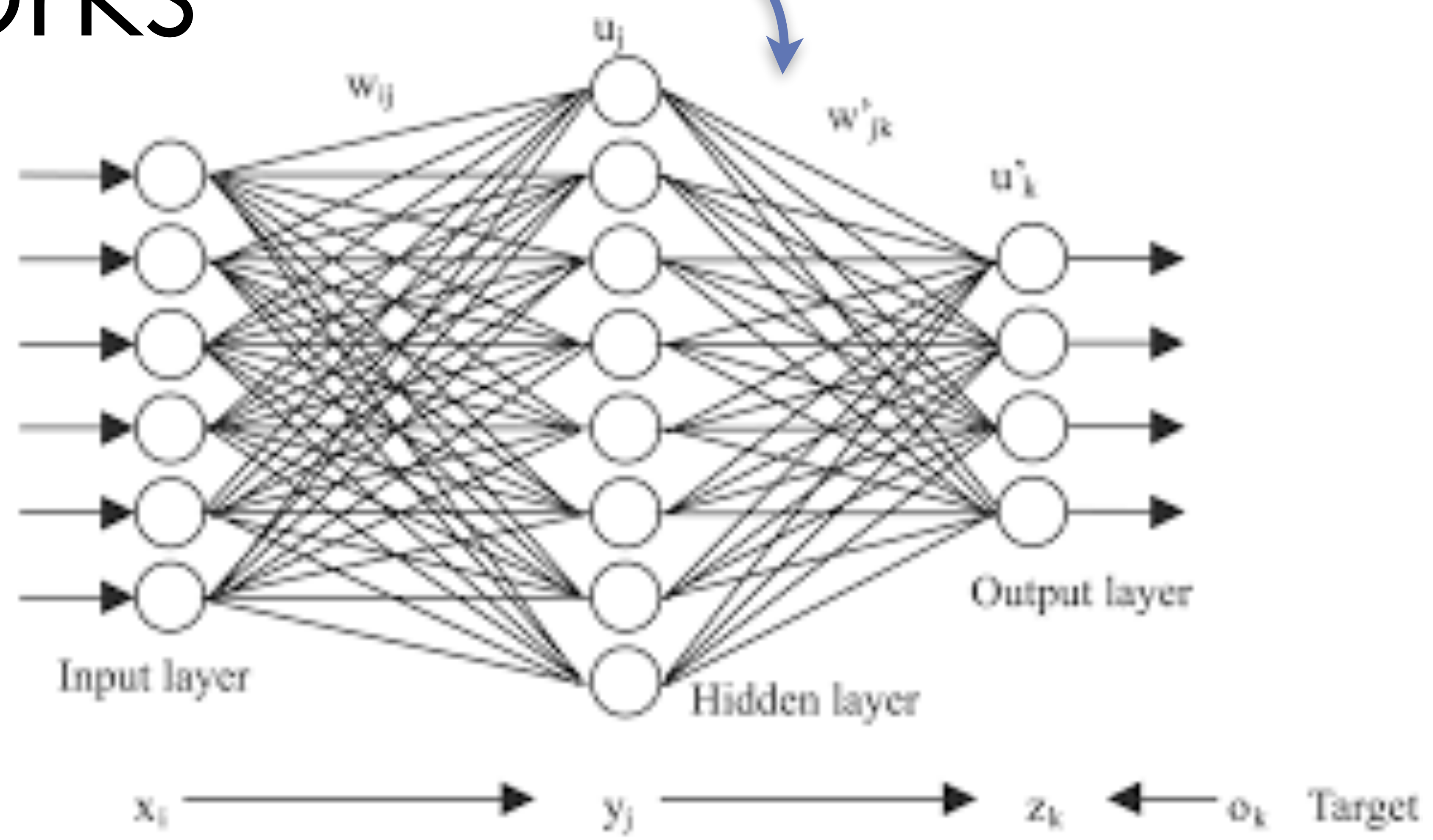
Popular classification methods

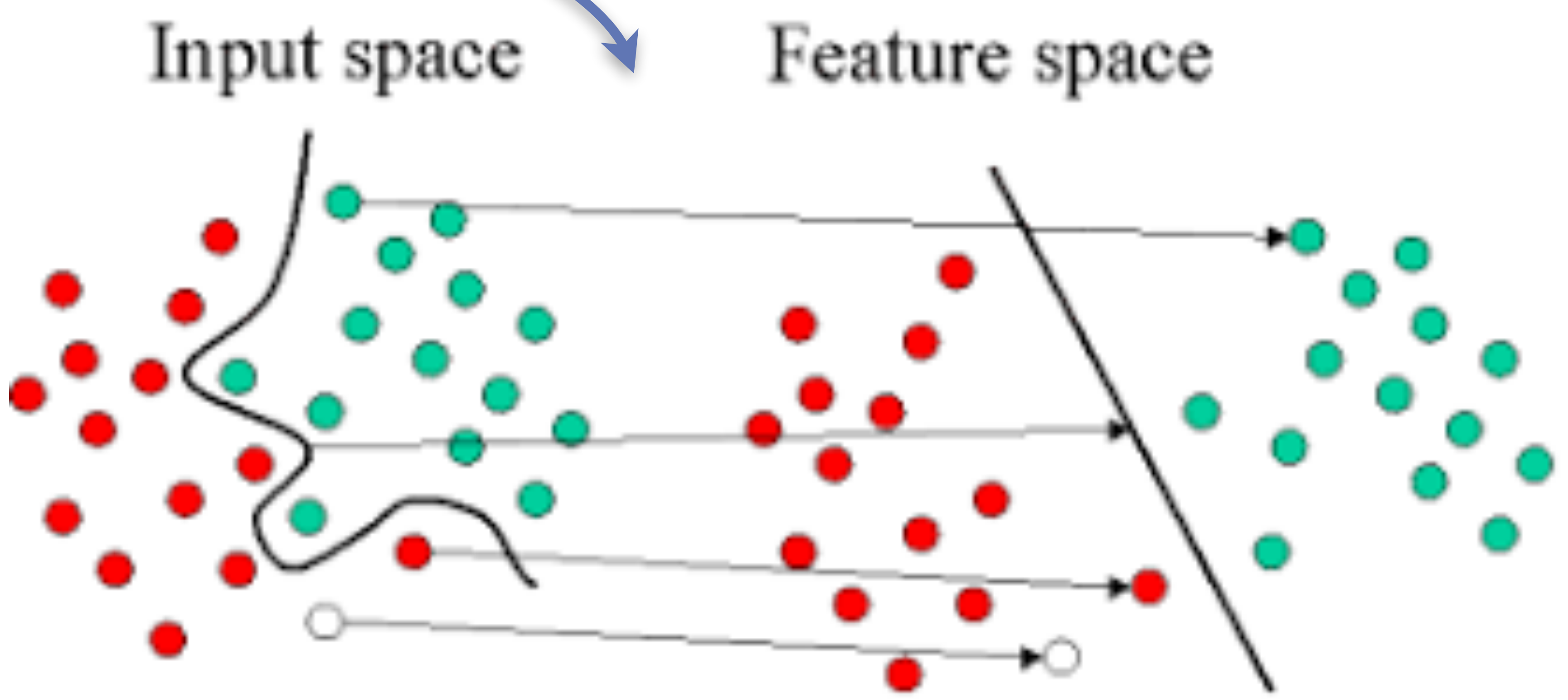
- Decision tree induction
- Bayes classification
- Linear regression
- Support vector machines
- Neural networks
- …





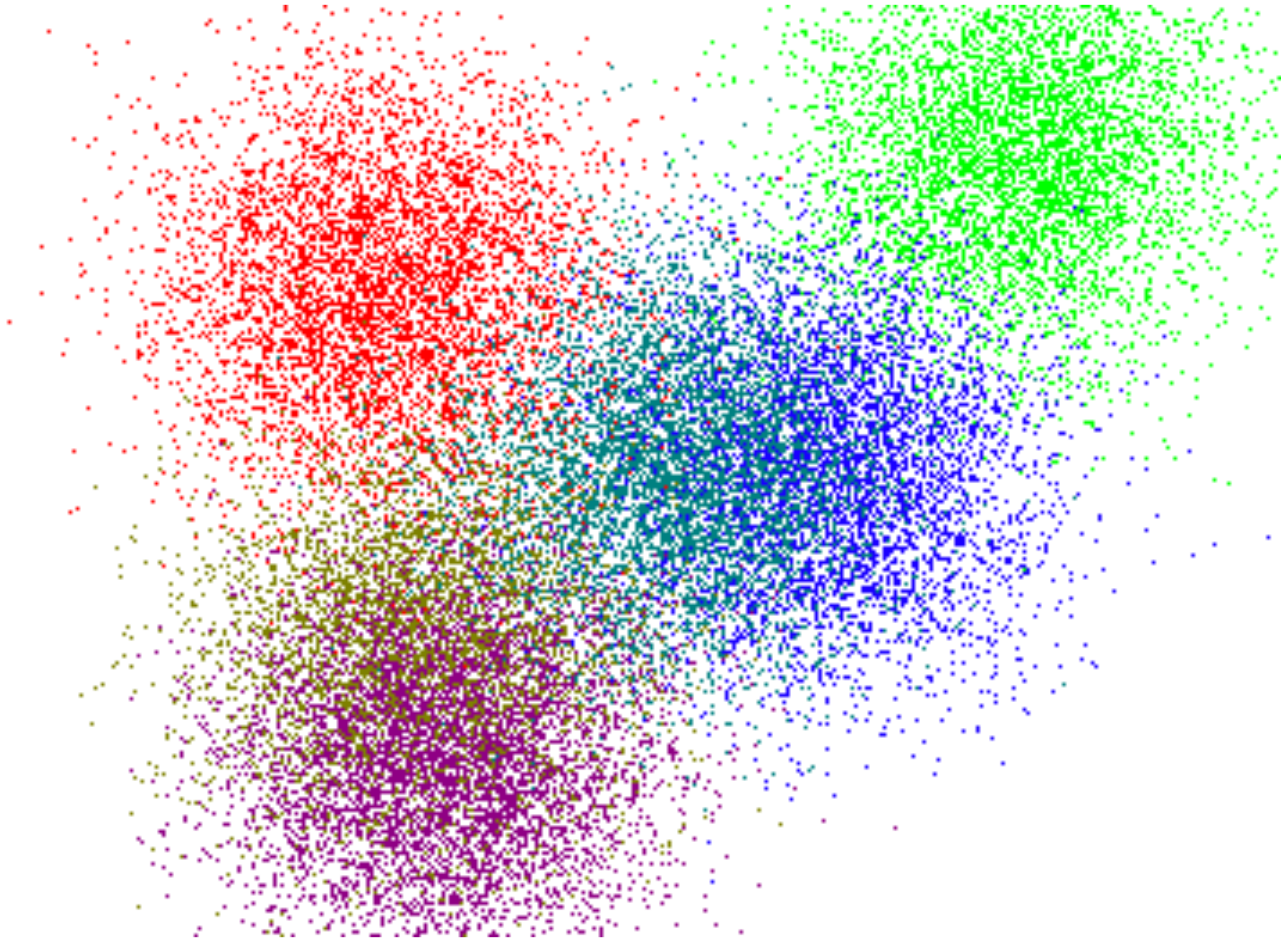






Cluster Analysis

- Unsupervised learning (i.e., no predefined classes)
  - Given a set of data points, partition them into a set of groups (i.e., clusters)
  - High intra-class similarity and low inter-class similarity



Partitioning concepts

- Partitioning
  - Discovering groupings in data by optimising an objective function and iteratively improving the quality of partitions
- K-partitioning
  - Partitioning a dataset D of n objects into a set of K clusters so that an objective function is optimised
  - A typical objective function is sum of squared errors (SSE)

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_{i \in C_k}} \| x_i - c_k \|^2$$

K-means clustering

- K-means (MacQueen 1967, Lloyd 1957, 1982)
  - Each cluster is represented by the centre of the cluster
- K-means clustering algorithm—
  - Select k points as initial centroids
  - Repeat until convergence criterion is satisfied:
    - Form k clusters by assigning each point to its closest centroid
    - Re-compute the centroids of each cluster
- Different kinds of measures can be used
  - Manhattan distance L^1 norm; Euclidean distance L^2 norm; …

Variations of k-means

- There are many variants of the k-means method:
  - Choosing better initial centroid estimates
    - K-means++; Intelligent K-means; genetic K-means
  - Choosing different representative prototypes for the clusters
    - K-medoids; K-medians; K-modes
  - Applying feature transformation techniques
    - Weighted K-means, Kernel K-means

Summary

- Data mining and its applications
- KDD from different views
- Mining frequent itemsets and association rules
- Classification methods
- Cluster analysis methods

References

- Association mining:

  - Rakesh Agrawal, Ramakrishnan Srikant. (1994). Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the 20th International Conference on Very Large Data Bases, VLDB.
  - Jiawei Han, Jian Pei, Yiwen Yin. (2000). Mining frequent patterns without candidate generation, Proceedings of the ACM SIGMOD international conference on Management of data.
  - Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). Parallel algorithms for discovery of association rules. Data mining and knowledge discovery, 1(4), 343–373.

- K-means and variants:

  - MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, California. https://projecteuclid.org/euclid.bsmsp/1200512992
  - Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory. 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.
  - David Arthur, Sergei Vassilvitskii, K-means++: the advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007.
  - H.S. Park , C.H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 36, (2) (2009), 3336–3341.
  - Gereon Frahling, Christian Sohler, A fast k-means implementation using coresets, Proceedings of the twenty-second annual symposium on Computational geometry, 2006.
  - Qinghao Hu, Jiaxiang Wu, Lu Bai, Yifan Zhang, Jian Cheng, Fast K-means for Large Scale Clustering, Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017.

- Also recall the textbooks included on slide 3