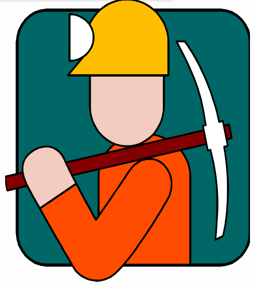
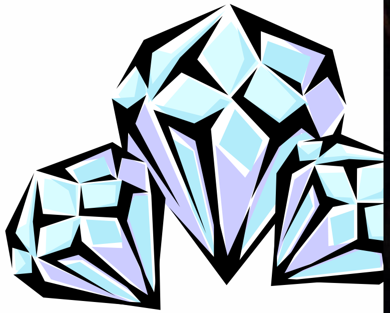
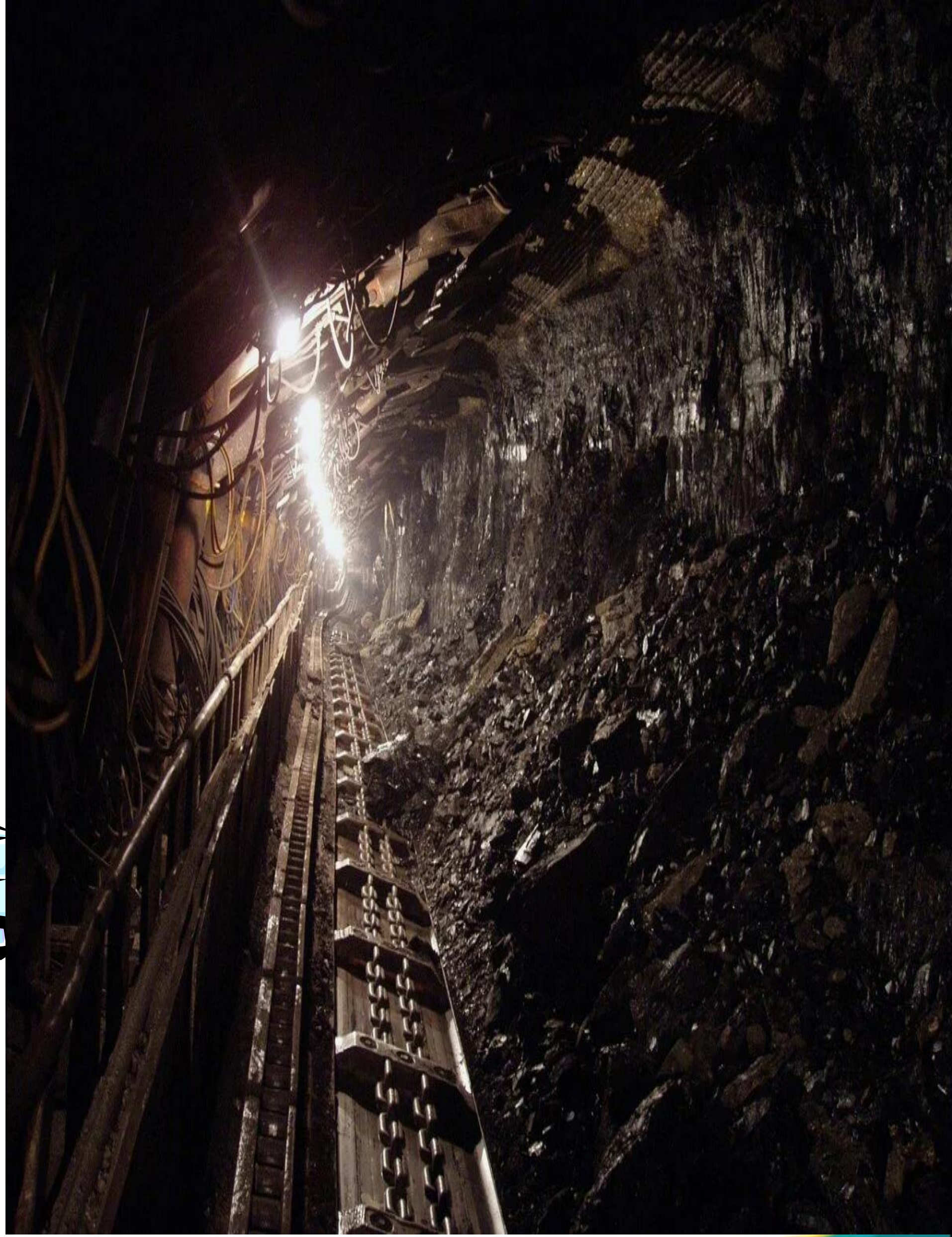


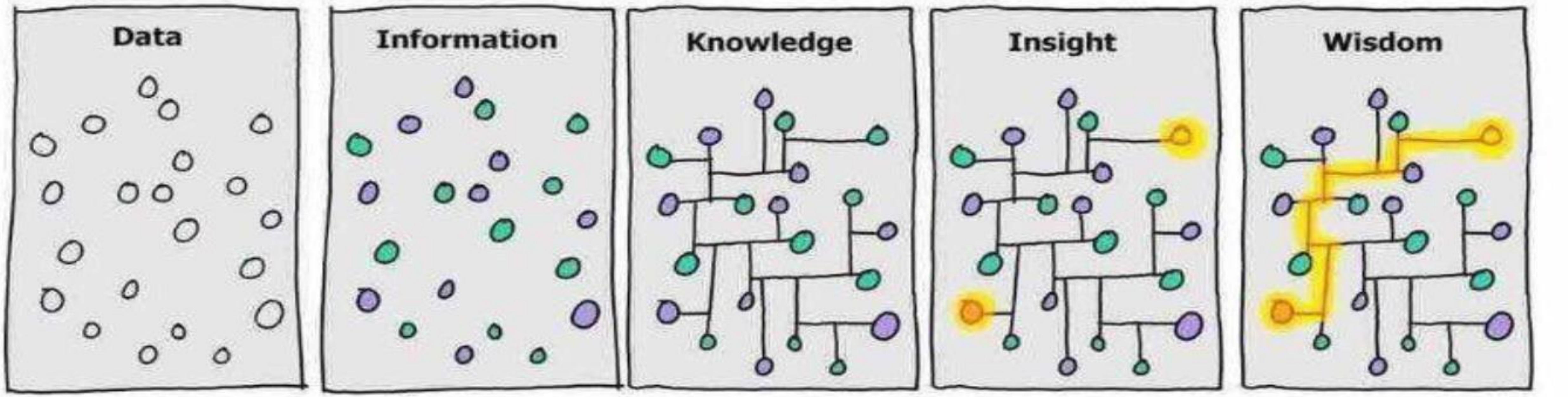
Chapter 1: Introduction to

Data Mining



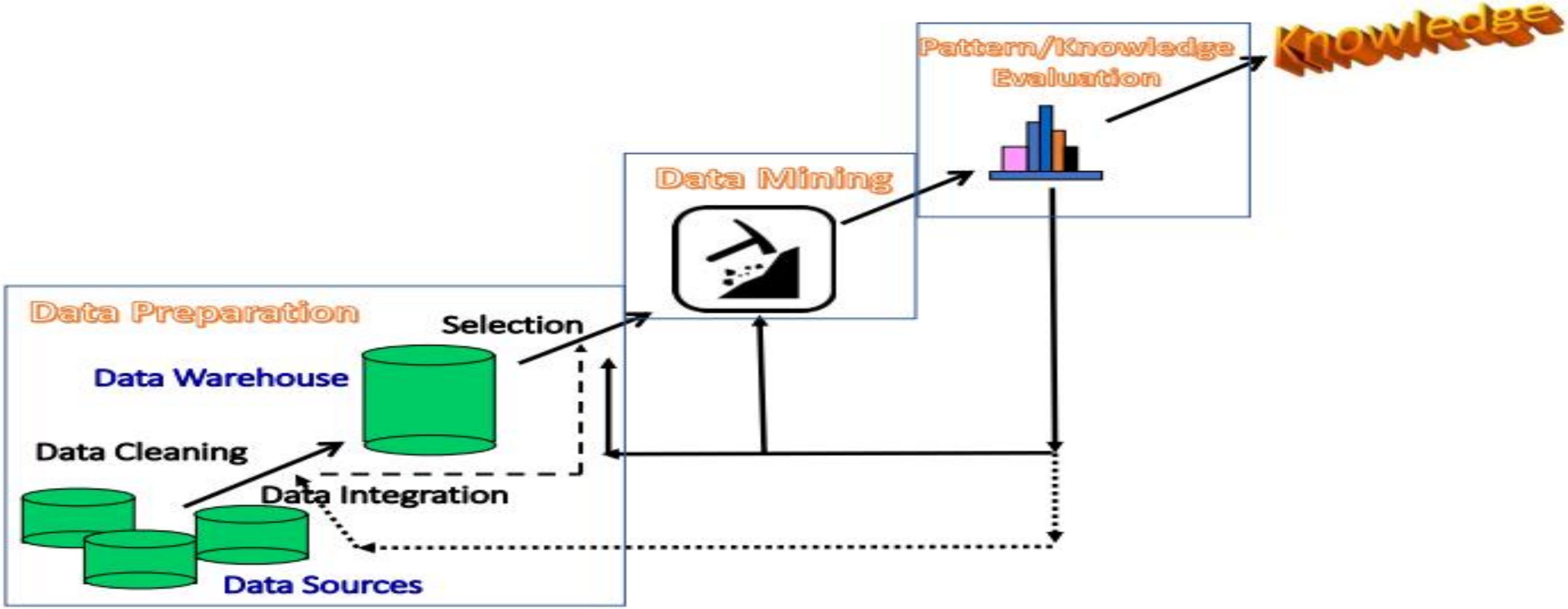


DIKIW MODEL



PIPELINE

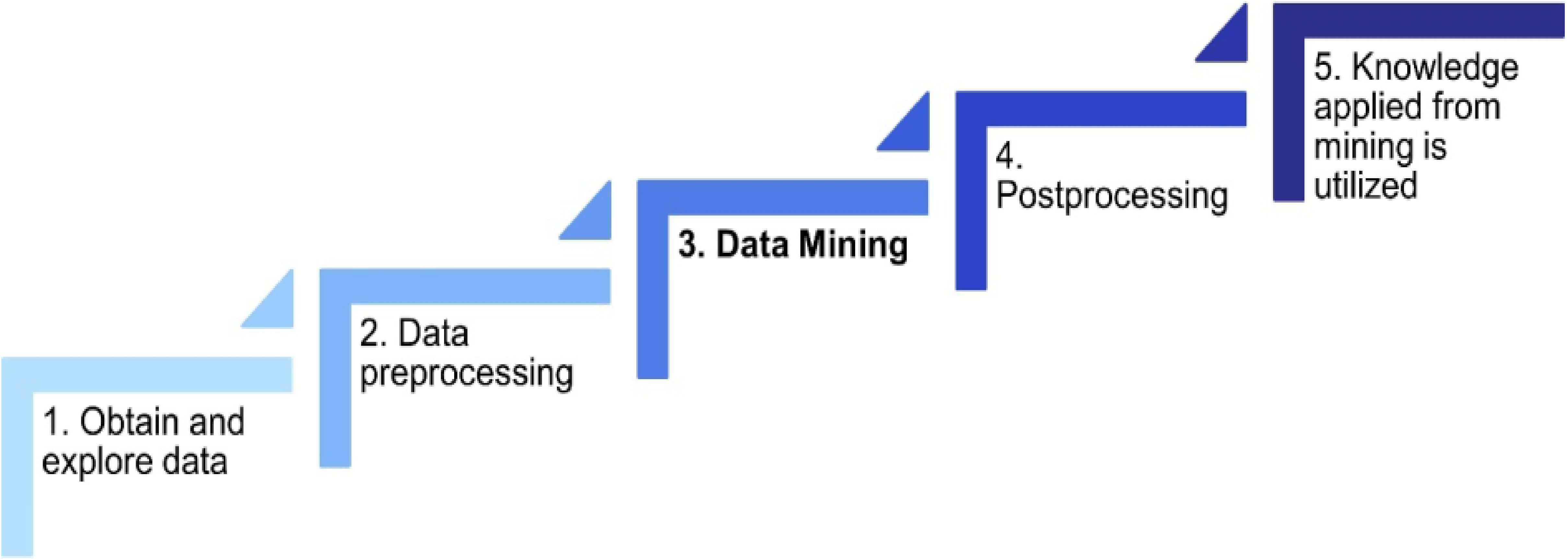
Data contains value and knowledge



Jiawei Han
Jian Pei
Hanghang Tong

PIPELINE

CRISP-DM (Cross-Industry Standard Process for Data Mining)



1. Obtain and explore data

2. Data preprocessing

3. Data Mining

4. Postprocessing

5. Knowledge applied from mining is utilized

What is data mining?

- After years of data mining there is still no unique answer to this question.

- A tentative definition:

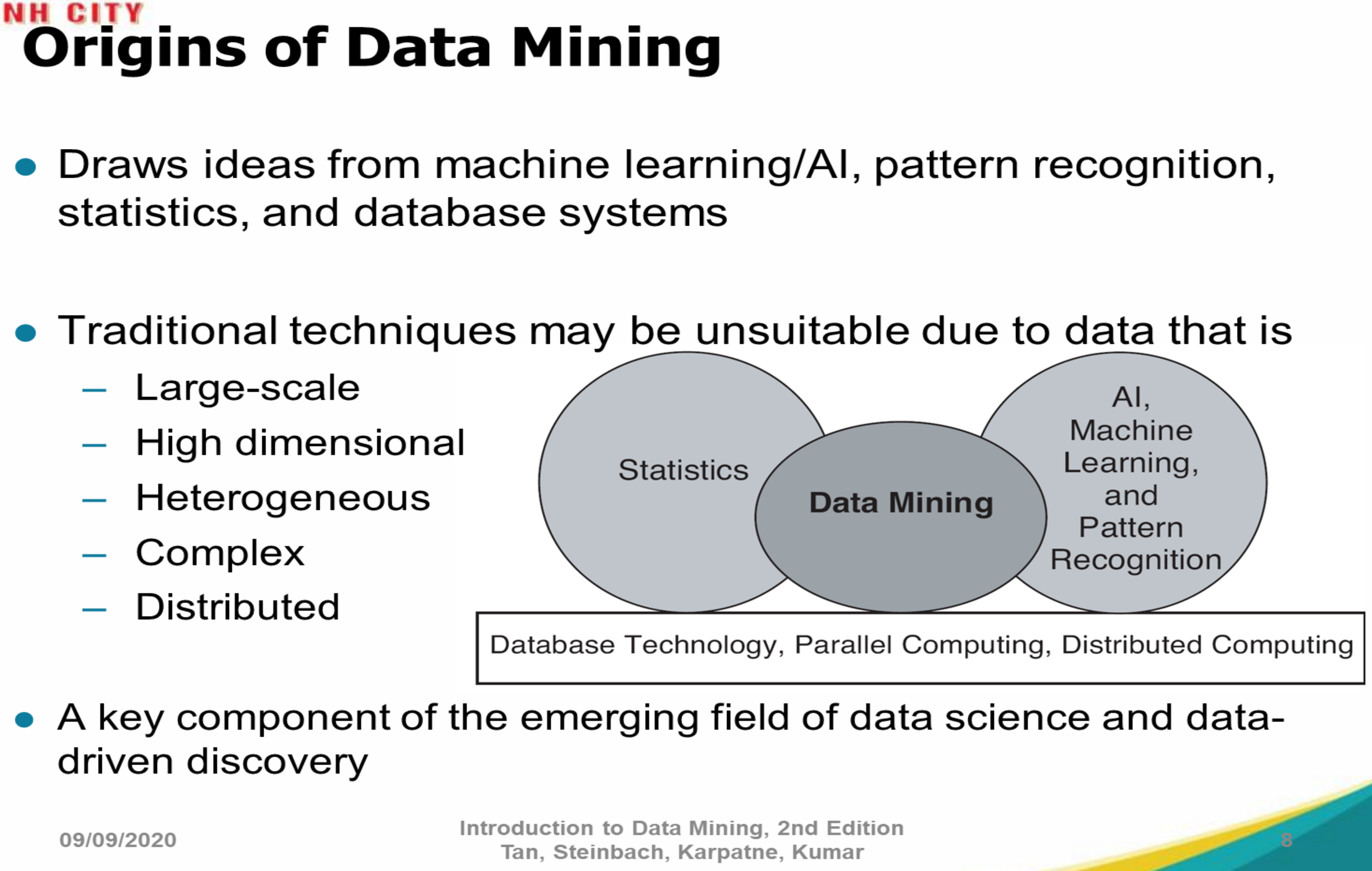
Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.

Why do we need data mining?

- Really, really huge amounts of raw data!!
  - In the digital age, TB of data is generated by the second.
    - Web, Wikipedia, Mobile devices, Digital photographs and videos, Facebook, Twitter, Instagram, Transactions, sensor data, behavioral data, scientific measurements, wearable computing
  - New ways of generating data are constantly created.
  - Cheap storage has made possible to maintain this data
- Need to analyze the data to extract knowledge

Why do we need data mining?

- "The data is the computer"
  - Large amounts of data can be more powerful than complex algorithms and models
    - Google has solved many Natural Language Processing problems, simply by looking at the data
    - Example: misspellings, synonyms
  - Data is power!
    - Today, the collected data is one of the biggest assets of an online company
      - Query logs of Google, The friendship and updates of Facebook, Tweets and follows of Twitter, Amazon transactions
  - Data for the people:
    - Using data from the people activity we can improve their individual lives but also the overall society life.
  - We need a way to harness the collective intelligence
- From Data mining to Data Science



Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed

Statistics

Data Mining

AI, Machine Learning, and Pattern Recognition

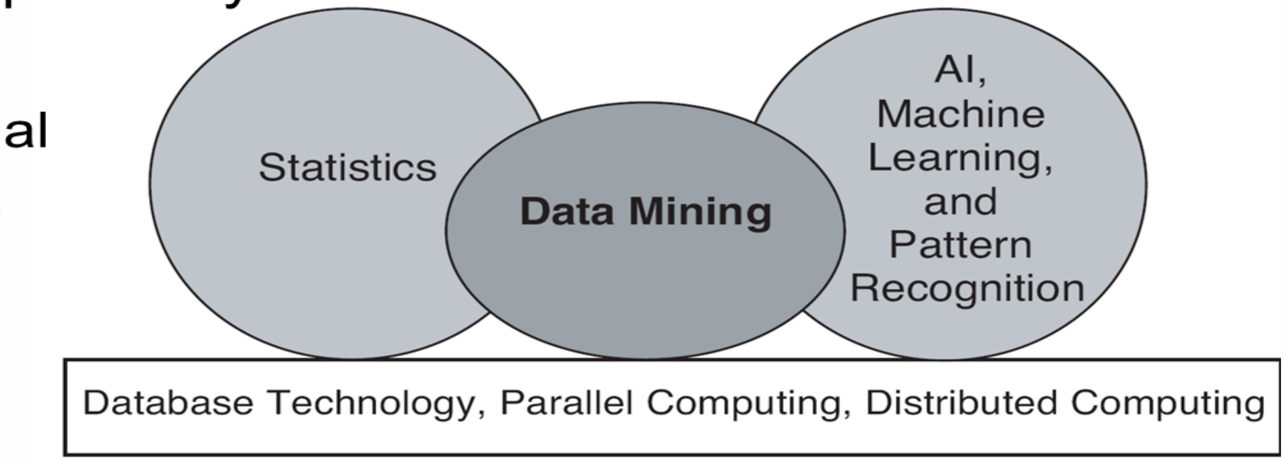
Database Technology, Parallel Computing, Distributed Computing

- A key component of the emerging field of data science and data-driven discovery

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



Statistics

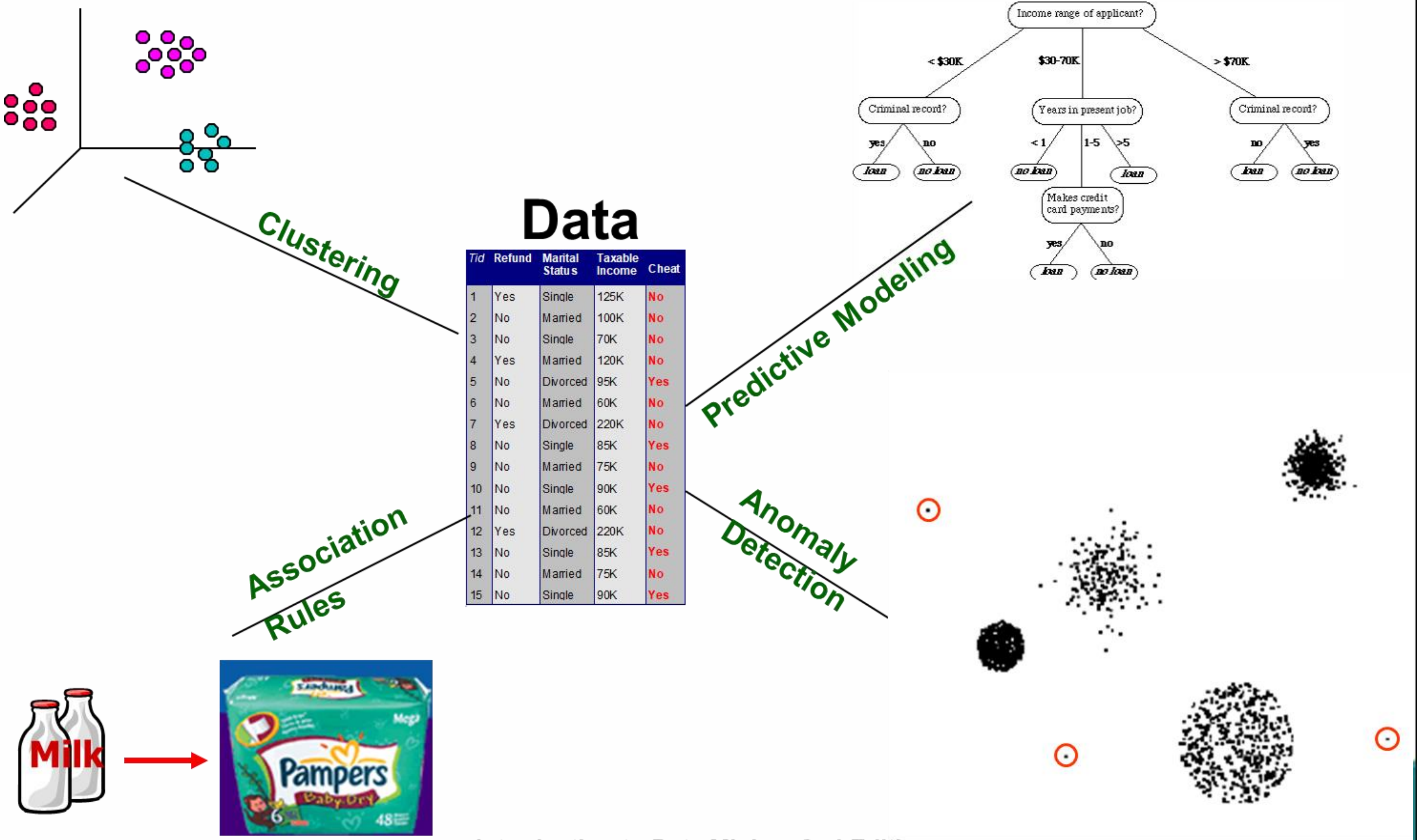
Data Mining

AI, Machine Learning, and Pattern Recognition

Database Technology, Parallel Computing, Distributed Computing

- A key component of the emerging field of data science and data-driven discovery

Data Mining Tasks ...



Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

09/09/2020

Introduction to Data Mining, 2nd Edition
Tan, Steinbach, Karpatne, Kumar

10

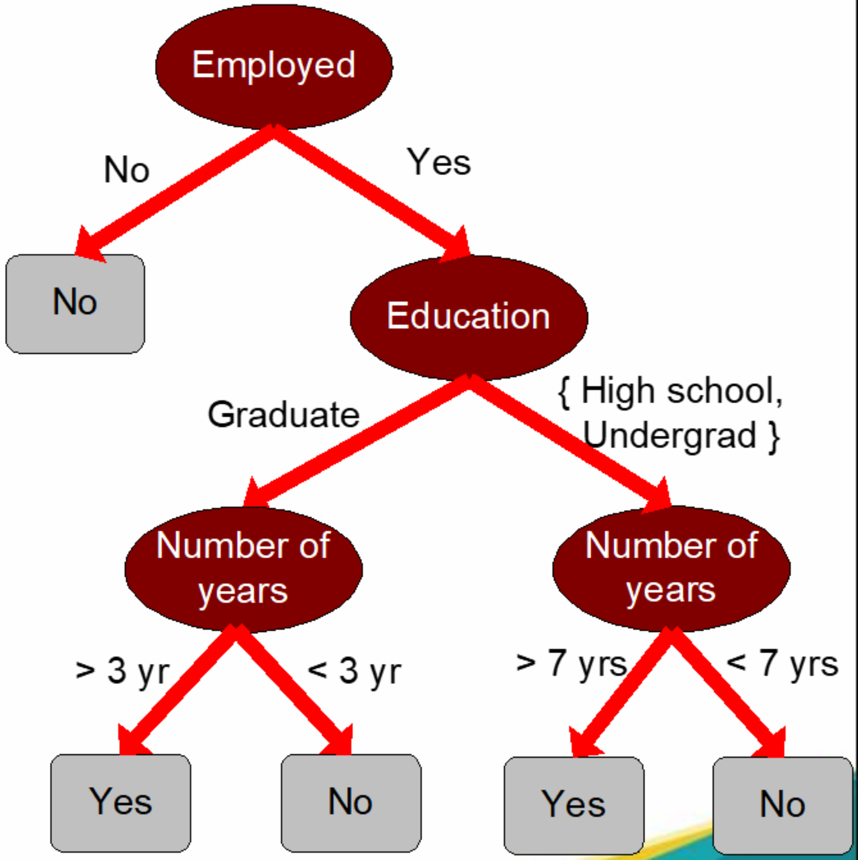
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness



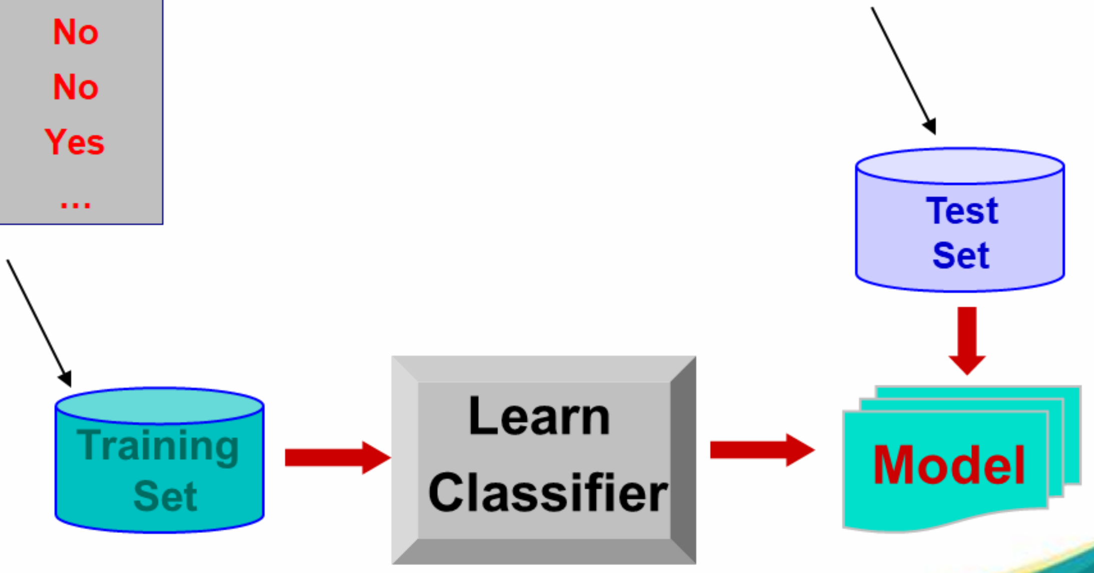
09/09/2020

Introduction to Data Mining, 2nd Edition
Tan, Steinbach, Karpatne, Kumar

Classification Example

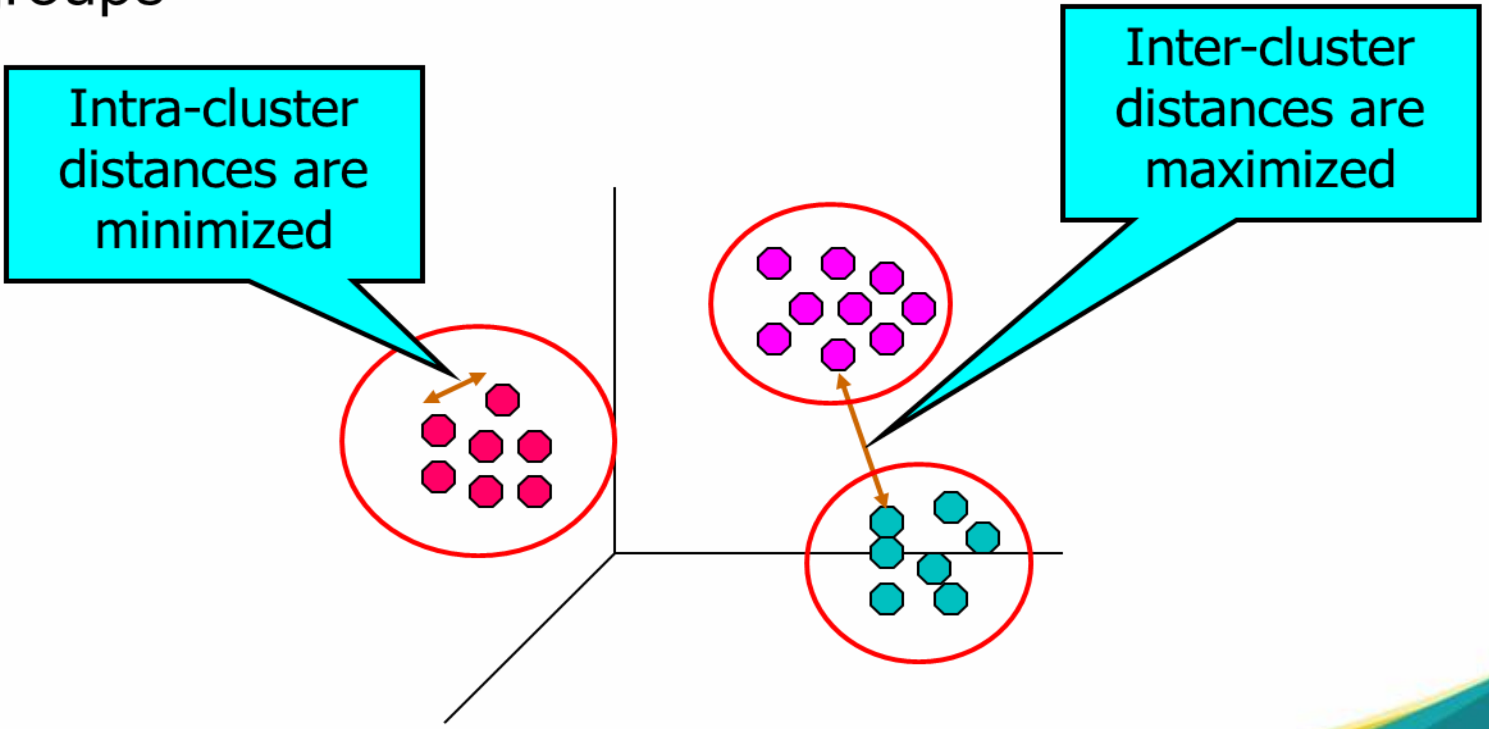
	categorical	categorical	quantitative	class
Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



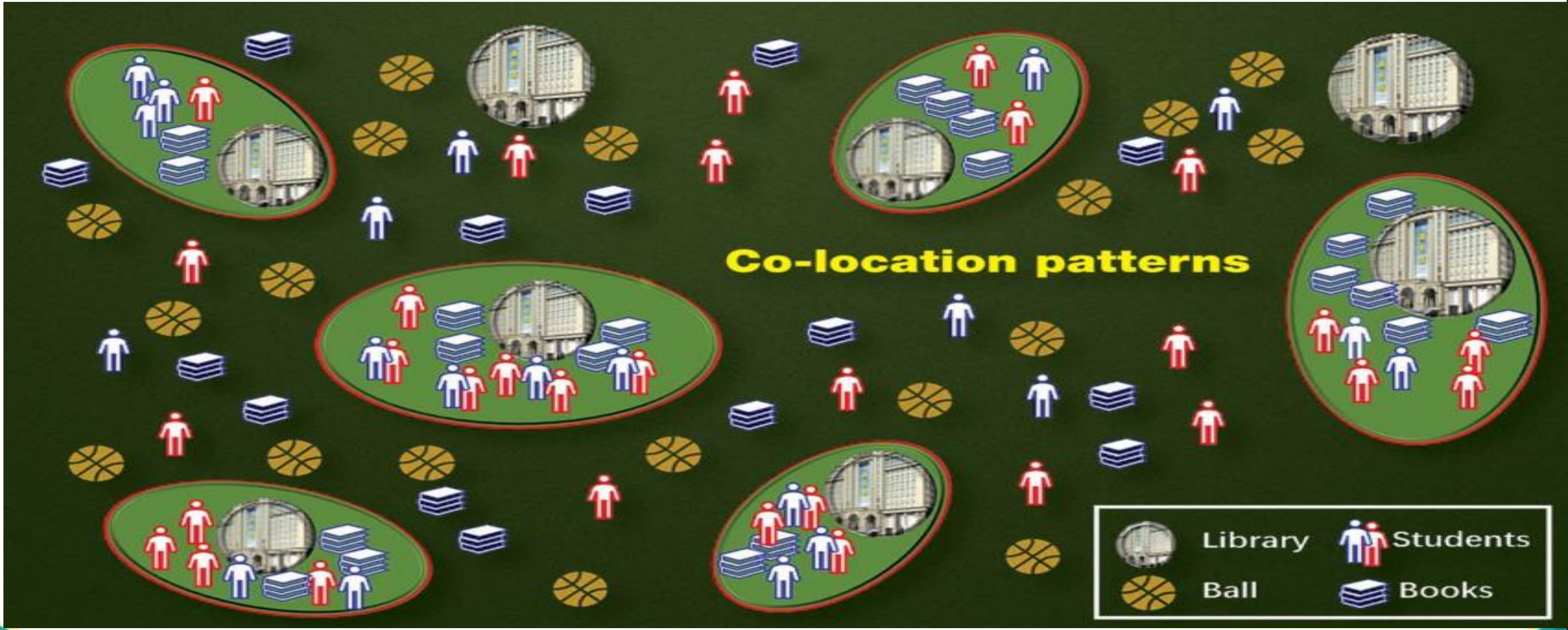
Intra-cluster distances are minimized

Inter-cluster distances are maximized

09/09/2020

Introduction to Data Mining, 2nd Edition
Tan, Steinbach, Karpatne, Kumar

Clustering: Application 1



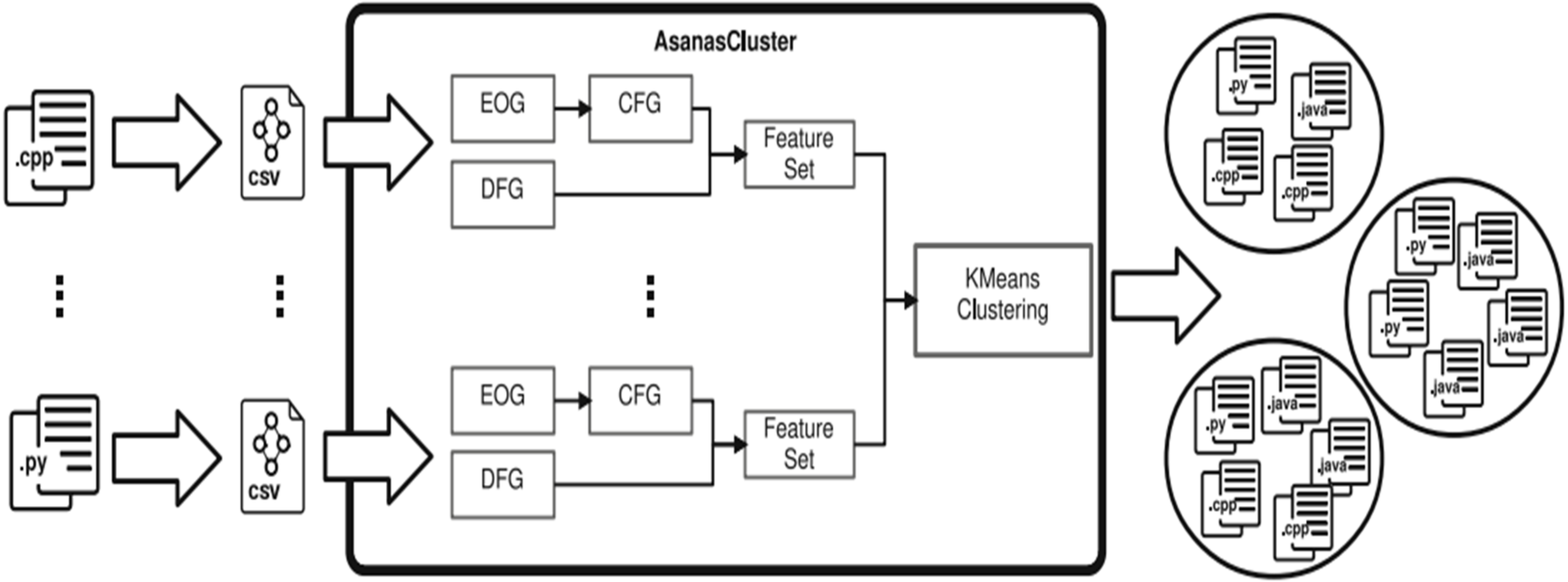
Clustering: Application 2

- Document Clustering:
  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset

scott@onlinemailbox.net
Cindy Olson
marty.sunde@enron.com
Gordon Savage
Michael Brown
tom.briggs@enro
Phillip K. Allen
Mike Smith
Larry
mark.metts@enron.com
jordan.mintz@enron.com
Jeffrey T. Hodge
Chris Longgov
Steve Hall
evan.hughes@
John Lavorato
Mark Frevert
Steven J. Kean
vance.meyer@enron.com
gail.brownfeld@enron.co
kevin.garland@enron.com
mark.koenig@enron.com
herrick@enron.com
miyung.b
john.sherriff@enron.com
elizabeth.tilney@enron.com
steven.kean@enron.com
Wall
Steve Montovano
Tom
muller@enron.com
richard.causey@enron.com
john.shelk@enron.com
brian.stanley@enron.com
bryan.seyfried@enron.com
janel.guerrero@enron.com

Clustering source code from automated assessment of programming assignments



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

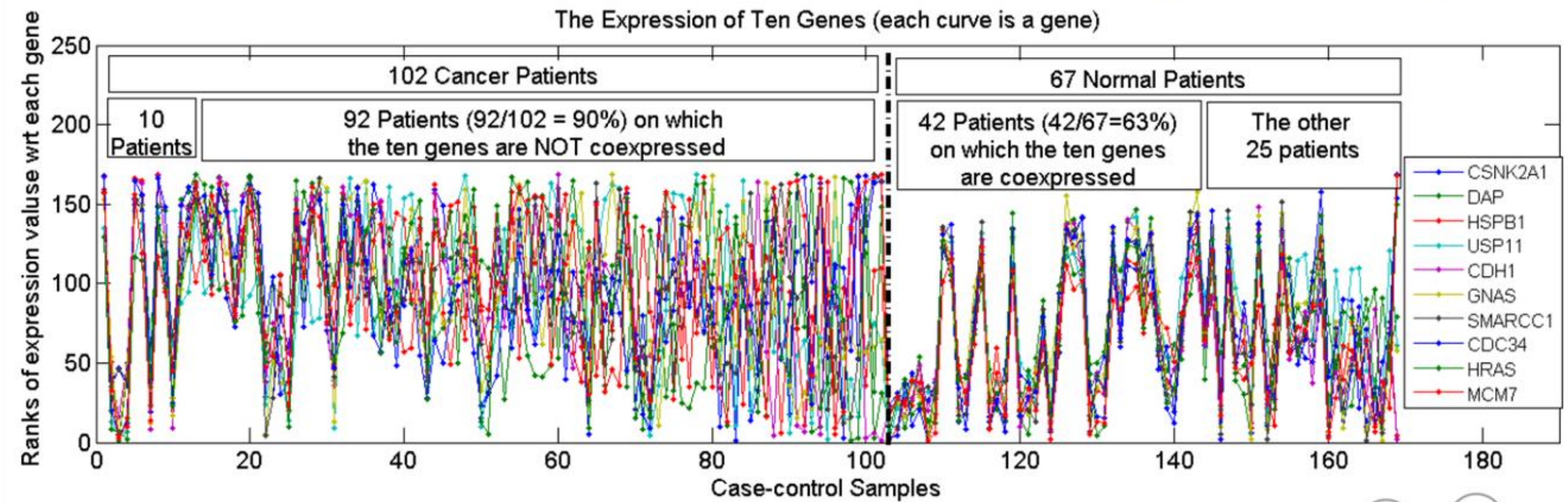
{Milk} --> {Coke}

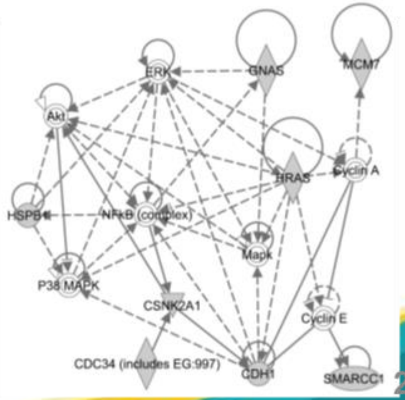
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]

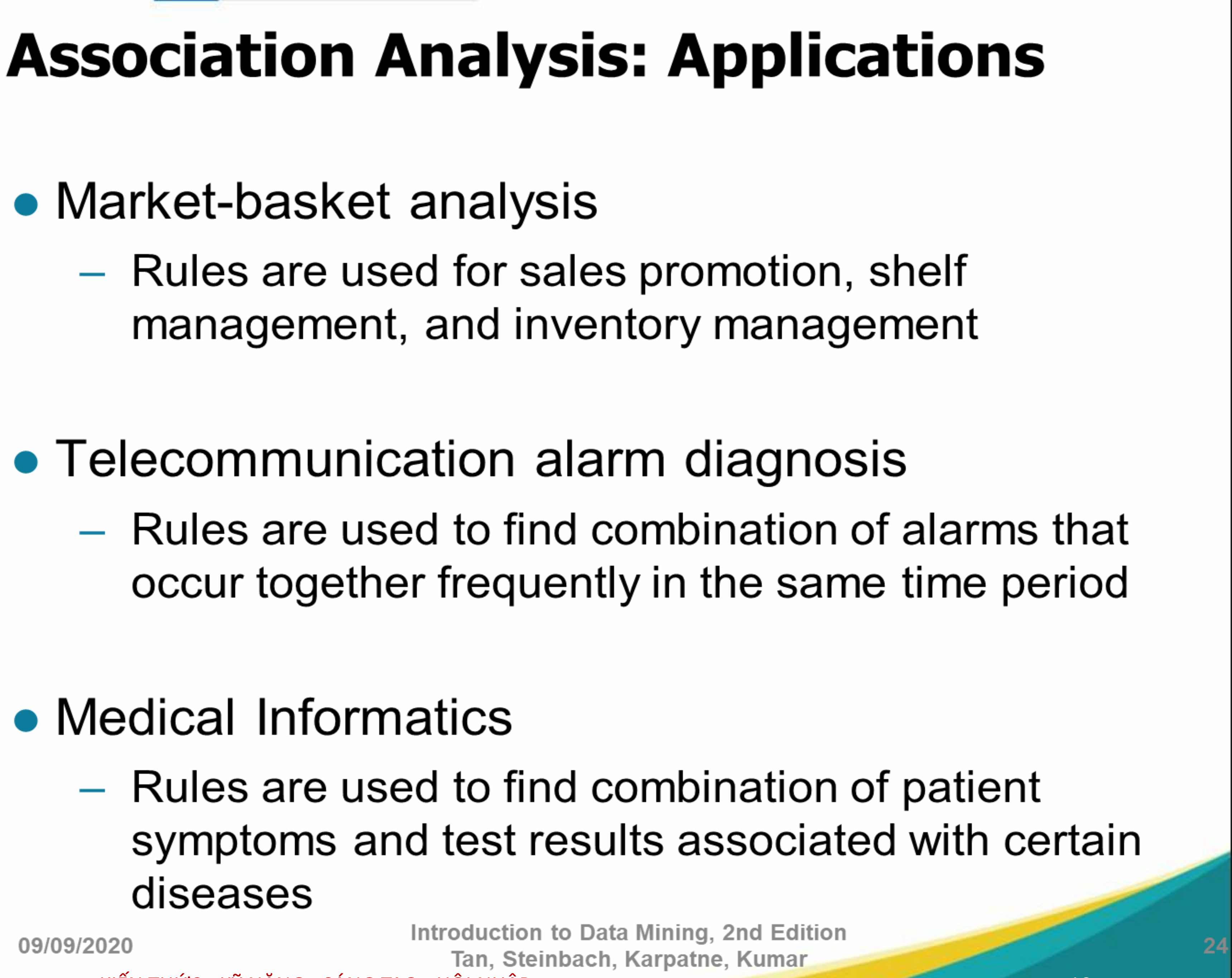




Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer

P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

[Fang et al PSB 2010]



Association Analysis: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.