Data Warehouse and Data Mining

Dhruv Gupta

10-January-2022

NTNU | Norwegian University of Science and Technology

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Data and Data Mining
   - Data Modeling
   - Databases and Data Warehouses
   - Association Rules
   - Classification
   - Clustering
   - Summary

Course Team — Instructors & TA Leaders

Instructor
Dhruv Gupta
(*dhruv.gupta@ntnu.no*)

Teaching Assistant Leader
David Baumgartner
(*david.baumgartner@ntnu.no*)

Teaching Assistant Leader
Shiva Shadrooh
(*shiva.shadrooh@ntnu.no*)

Course Team

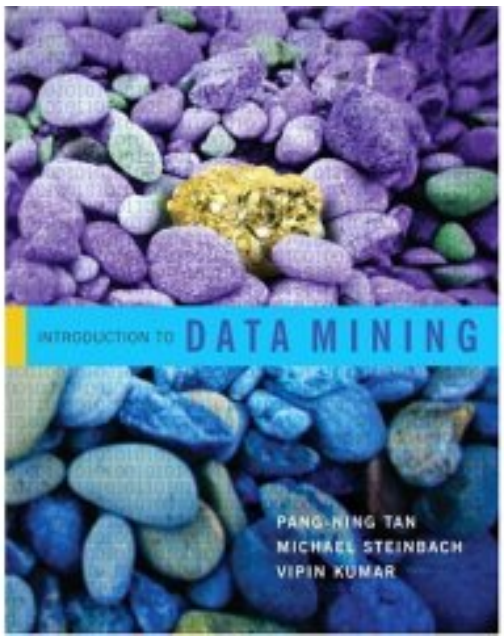
- Course ID: TDT4300 Data Warehouse and Data Mining
- Course Instructor: Dhruv Gupta
- Teaching Assistants:
  - David Baumgartner (*david.baumgartner@ntnu.no*)
  - Shiva Shadrooh (*shiva.shadrooh@ntnu.no*)
  - Mattias Ness
  - Peder Smith
- Course Email / Assistant Email: tdt4300-undass@idi.ntnu.no
- Lecture Timings:
  - Mondays (1515-1700 HRS)
  - Wednesdays (1615-1800 HRS)
- Piazza is the main platform for discussions.
- Blackboard is the main platform for course-related information.

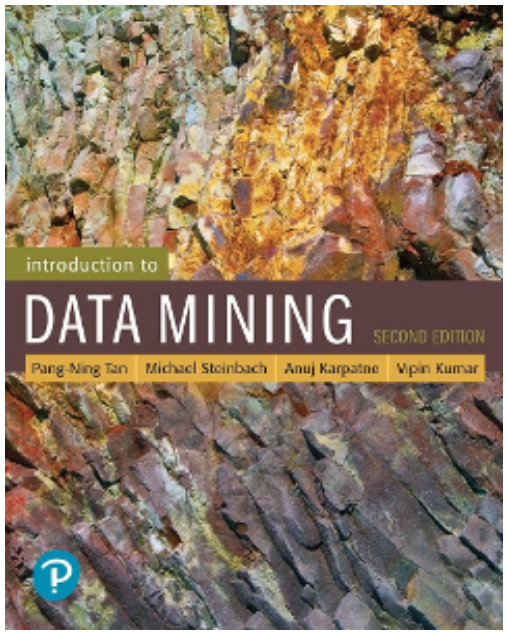
Course Prerequisites

- The subject material is at an advanced level.
- Recommended prerequisites:
  - TDT4145 Data Modeling and Database Systems or equivalent
  - Programming experience through ITGK,
    programming GK, and
    AlgDat (or equivalent).
- However, this course is not too much programming-oriented.

Course References

- Many available books on the course topic.
- Main course book for the majority of the topics[1,2].
  Tan, Steinbach, Karpatne, and Kumar,
  Introduction to Data Mining (1st or 2nd Edition).
- We will cover topics of data warehouses from other books.





[1] Image Credit: https://www-users.cs.umn.edu/~kumar001/dmbook/firsted.php

[2] Image Credit: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Guest Lectures

- There will be guest lecture from BearingPoint.
- Intention is to supplement the course knowledge from industry by daily practitioners.
- The guest lecture will be in digital format.
- Announcement at an upcoming date.

Assignments

- 5 mandatory exercises, where 4 out of 5 must be approved.
- Either individual or two-person groups.
  Assignments are doable by one person.
- The exercises will be published approximately two weeks before the submission deadline.
- In this subject, we will (in contrast to many other subjects) expect that most of the exercise is done and correct, otherwise one will be asked to submit again.
- Plagiarism will lead to fail grade on the assignment.
- Tentative exercise plan:

Deadline (week)	Theme
03.February.2022 (05)	Modeling of data warehouses and OLAP operations
17.February.2022 (07)	Association rules
03.March.2022 (09)	Clustering
17.March.2022 (11)	Classification
31.March.2022 (13)	Miscellaneous

Examination

- Examination date: 24-May-2022 at 1500 HRS.
- Exam format: School exam.
- Old exam assignments: Posted on Blackboard.
  There is also a solution sketch for most questions.

Course Contents

- **Academic Content**
  - The course deals with methods and theory for developing data warehouses and performing data analysis using data mining:
    1. Data quality and techniques for preprocessing data
    2. Modeling and design of data warehouses
    3. Algorithms for classification, clustering, and association rule detection
    4. Practical use of data analysis software
- **Learning Objectives**
  - The course will provide students with knowledge and practical experience with methods and theory for developing data warehouses and performing data analysis using data mining.

Lecture Plan (Tentative)

Week No./ Day	Topic	Book (Chap. No.)
2 / Monday	Introduction to Data Mining	Tan (1)
2 / Wednesday	Data Warehouse and multidimensional data	Han (4)
3 / Monday	Data Warehouse	Han (4)
3 / Wednesday	Data Warehouse / Indexing and Data and Types of Attributes	Han (4) and Tan (2)
4 / Monday	Data Quality and Preprocessing	Tan (2)
4 / Wednesday	Objectives for Equality and Inequality of Rules and Association Rules	Tan (2) and Tan (6)
5 / Monday	Association Rules	Tan (6)
5 / Wednesday	Association Rules and Clustering	Han (6.2.4 and 5) and Tan (8)
6 / Monday	K-means	Tan (8)
6 / Wednesday	HAC, DBScan, and Cluster Validation	Tan (8)
7 / Monday	Classification	Tan (4)
7 / Wednesday	Classification	Tan (4)
8 / Monday	Web Use Mining	Liu (12)
8 / Wednesday	Guest Lecture from BearingPoint	

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Data and Data Mining
   - Data Modeling
   - Databases and Data Warehouses
   - Association Rules
   - Classification
   - Clustering
   - Summary

Data

- "Data is the new Oil" — Clive Humby[3]
-  "People send more than 500 million Tweets per day."[4]
-  "Amazon had over 162 million unique mobile browser visitors in December 2018 ..."[5]
-  'At the maximum approximately 67 billion webpages were indexed by Google'[6]

---

[3] https://medium.com/project-2030/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294

[4] https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

[5] https://www.businessinsider.com/mobile-app-users-amazon-2019-3

[6] https://www.worldwidewebsize.com/

"Data is the New Oil"

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."

— Clive Humby[7]

[7] https://medium.com/project-2030/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294

"Data is the New Oil"

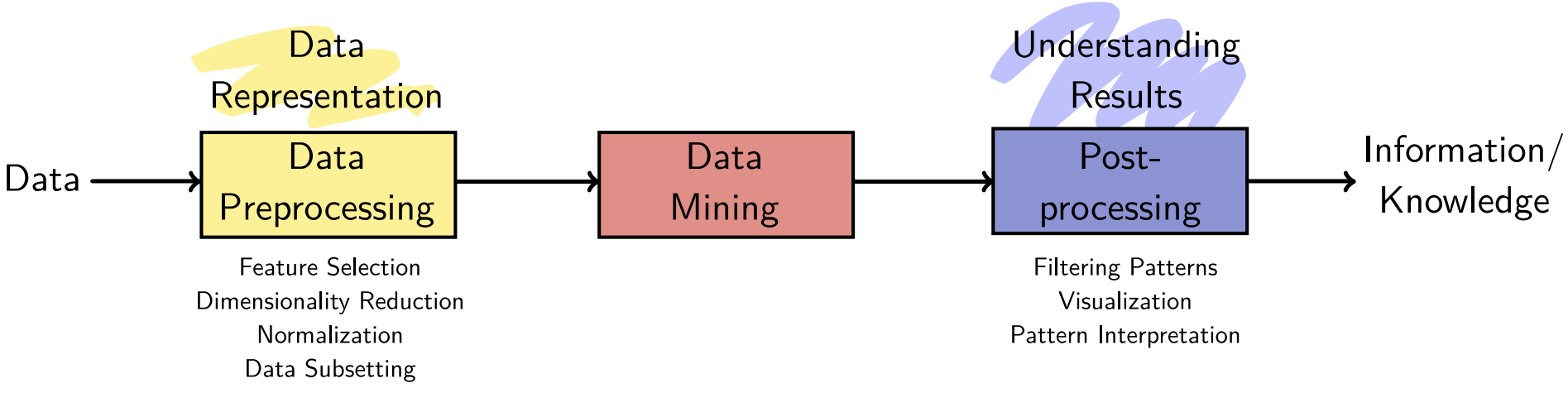
- Data collection is growing at a rapid pace as every aspect of our lives are digitized.
- Example of everyday data collection:
  - Shopping transactions
  - Search Engine queries
  - Smart home appliances
- However, the capability to analyze this "Big Data" is limited. Why?

Data Mining

- Key challenge that arises when analyzing massive amounts of data: scalability.
- Traditional statistical techniques when applied to massive amounts of data (think Terabytes) will often not converge to results quickly.
- Data mining thus leverages advances in modern computational capabilities to scale such statistical techniques to provide answers quickly.

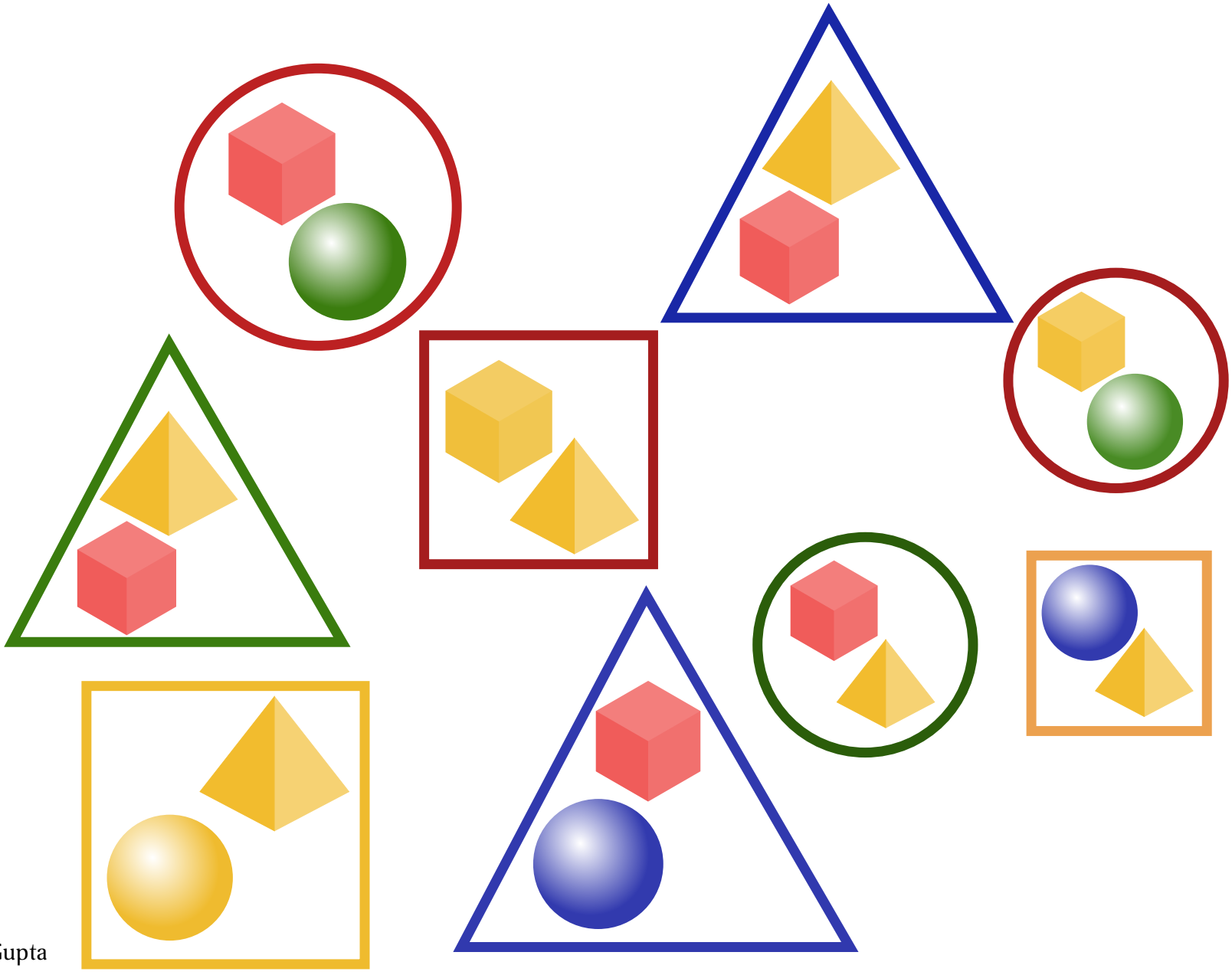
Data Mining

- Objective: Discover interesting insights from Big Data using computational methods. This can be done in a user-driver manner (e.g., analytical queries) or automatically (e.g., application of machine learning algorithms).
- This Knowledge Discovery process can be summarized as follows[8]:

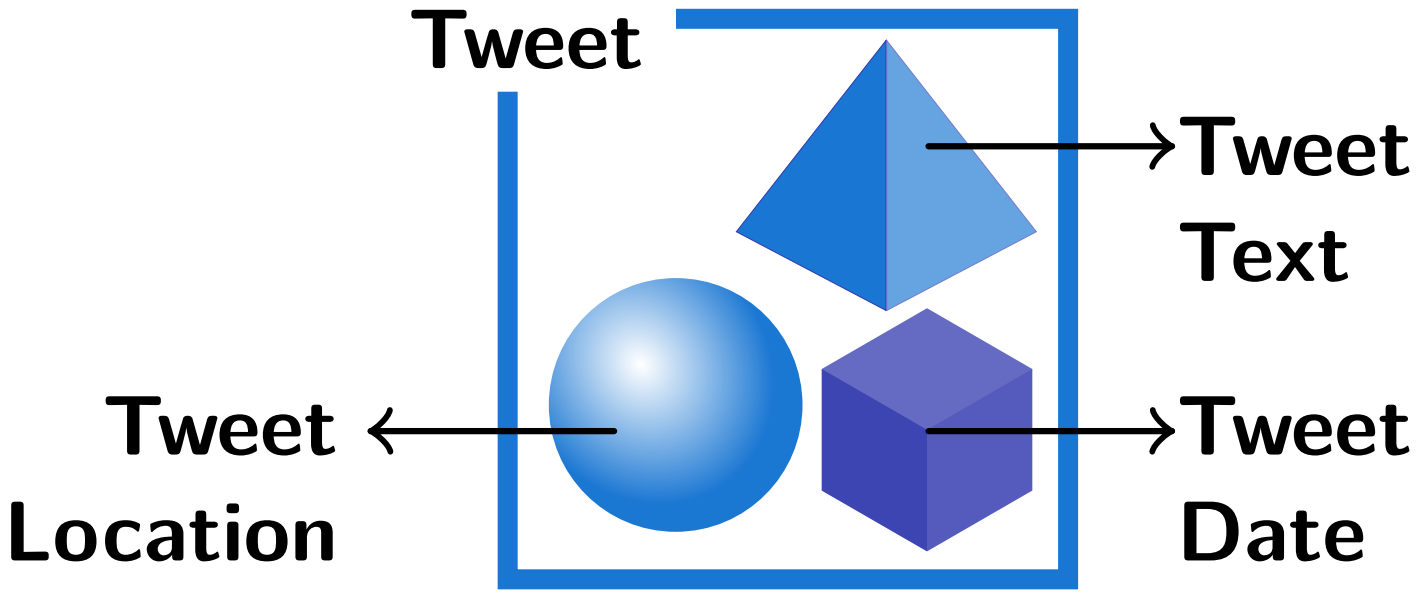


[8] Tan et al. Introduction to Data Mining (1st Edition). 2006.

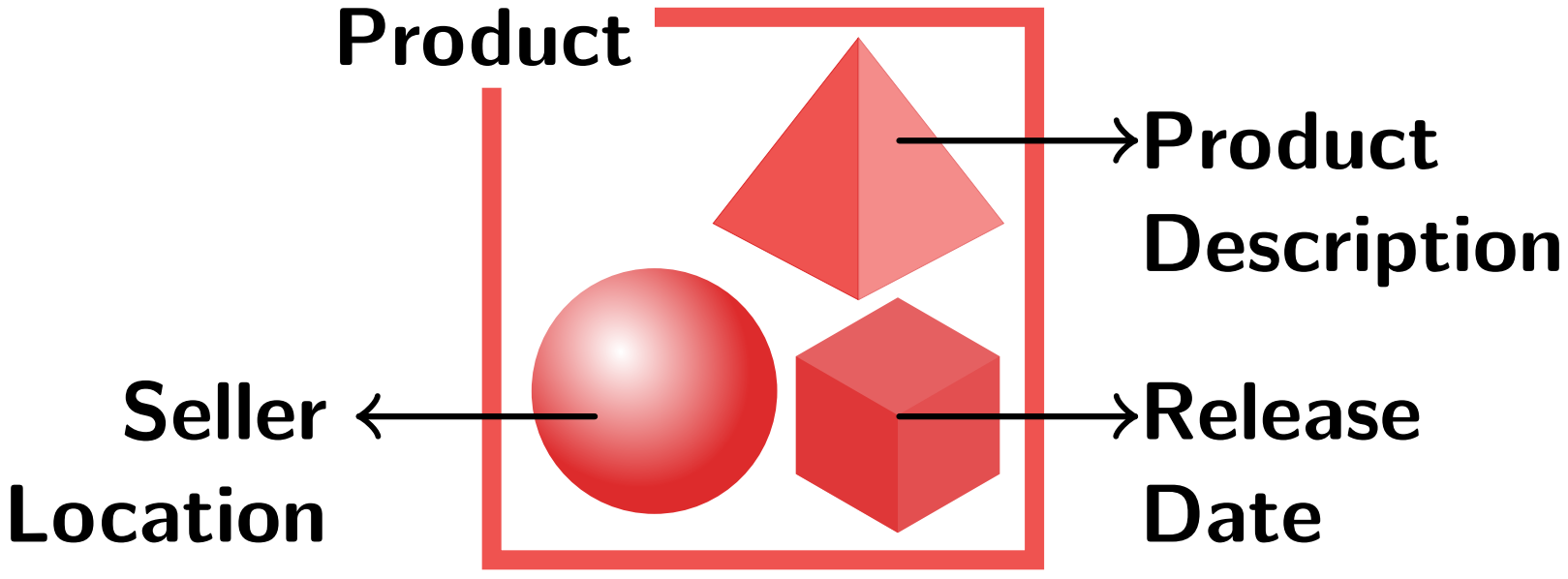
Data — Observations Example (Abstract)



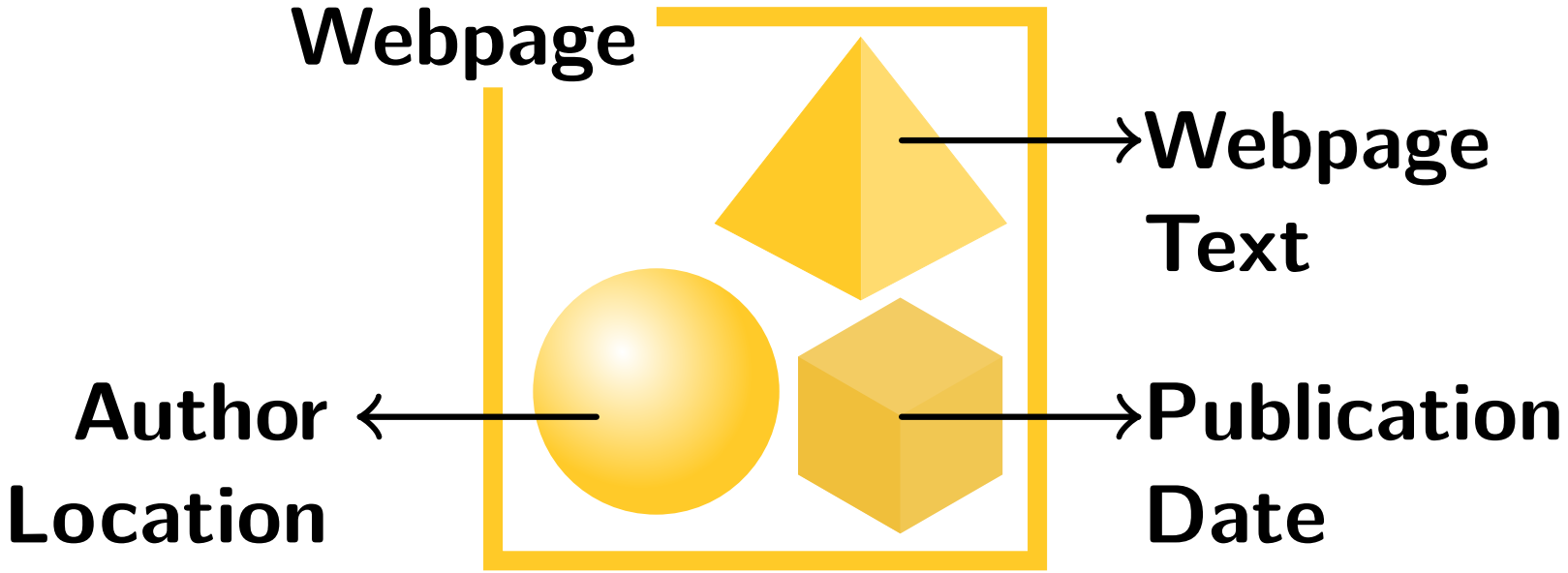
Data — Tweets Example



Data — Amazon Products Example

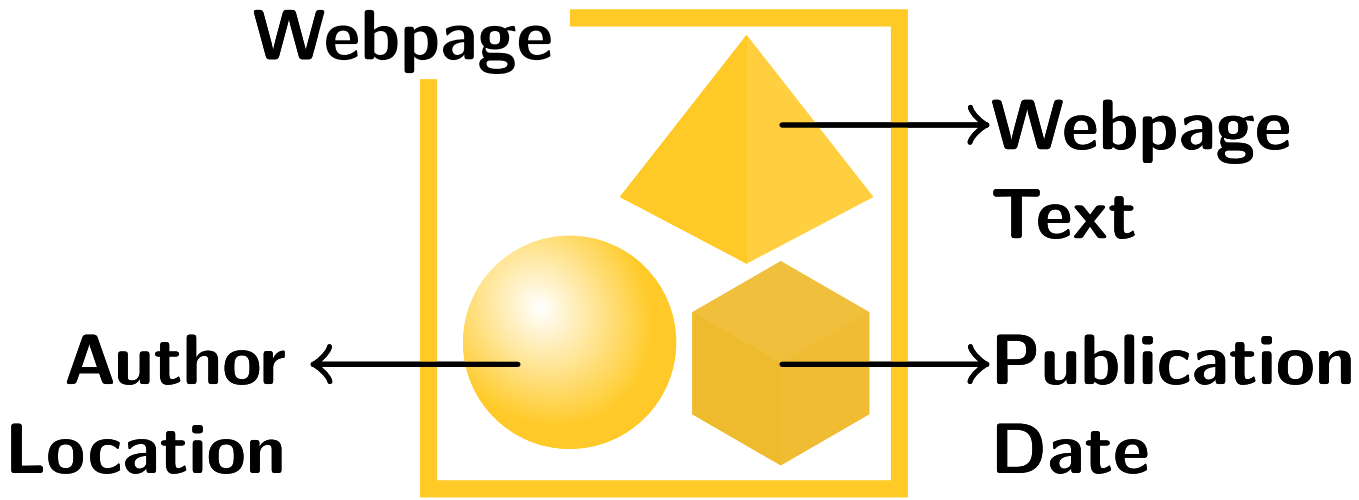


Data — Google Webpages Example



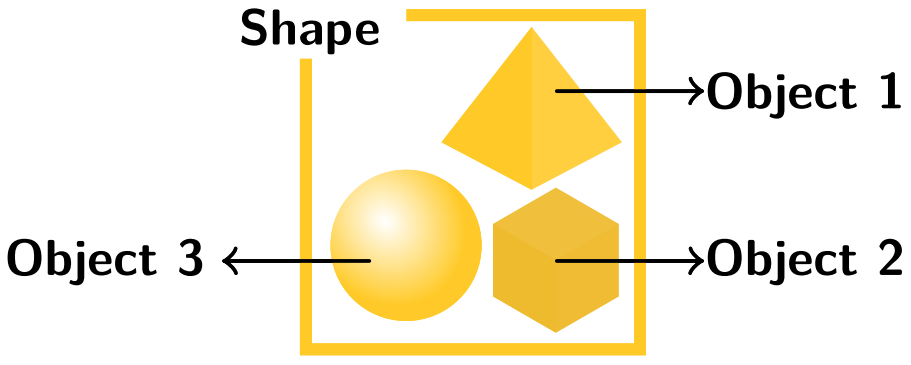
Data — Modeling and Storage Example

```
Class Webpage{
 String text; // Webpage Text
 Date   pubDate; // Publication Date
 Geo    location; // Publisher Location
}
```

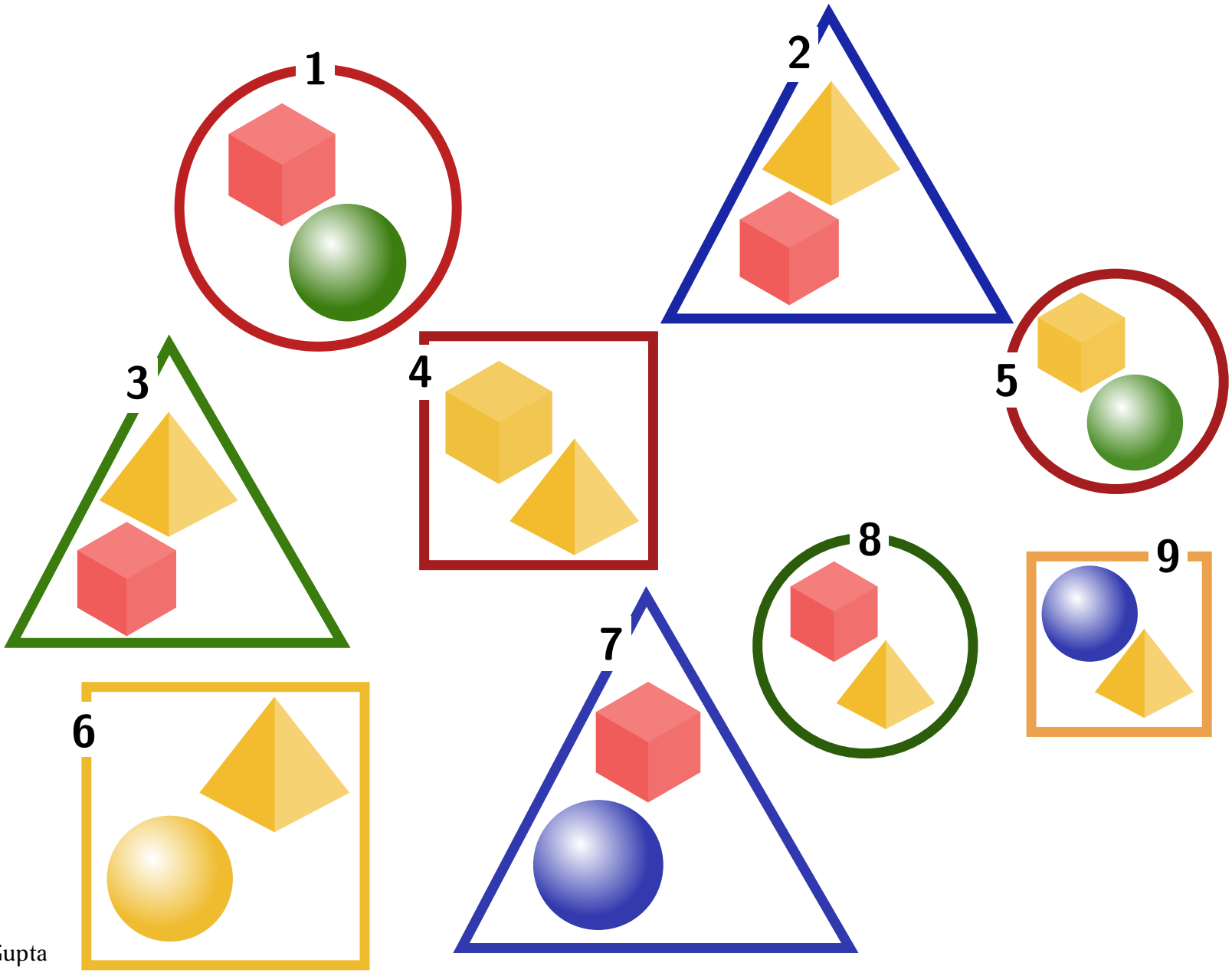


Data — Modeling and Storage (Abstract)

```
Class Data {
 String shape;
 Color  shapeColor;
 String[] objects;
 Color[] objectColors;
}
```



Data — Representation



Data — Representation

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

Databases — Transactional Processing

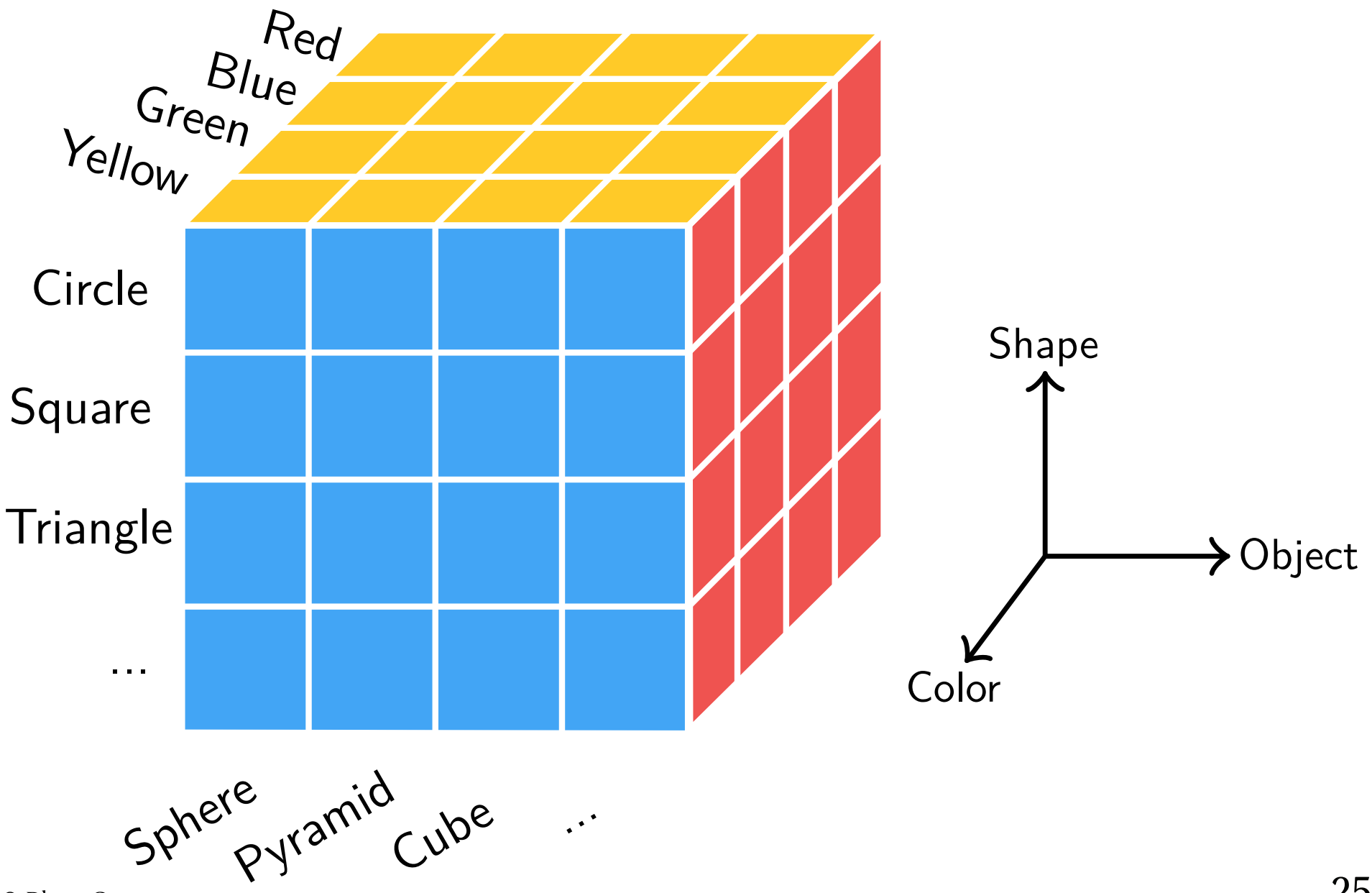
- Databases are great at storing large amounts of data for querying and selecting a subset of the data.
- In short, they are good for Online Transactional Processing (OLTP).
- However, the query language (SQL) becomes too cumbersome to perform analytical tasks.

Databases — Analytical Processing

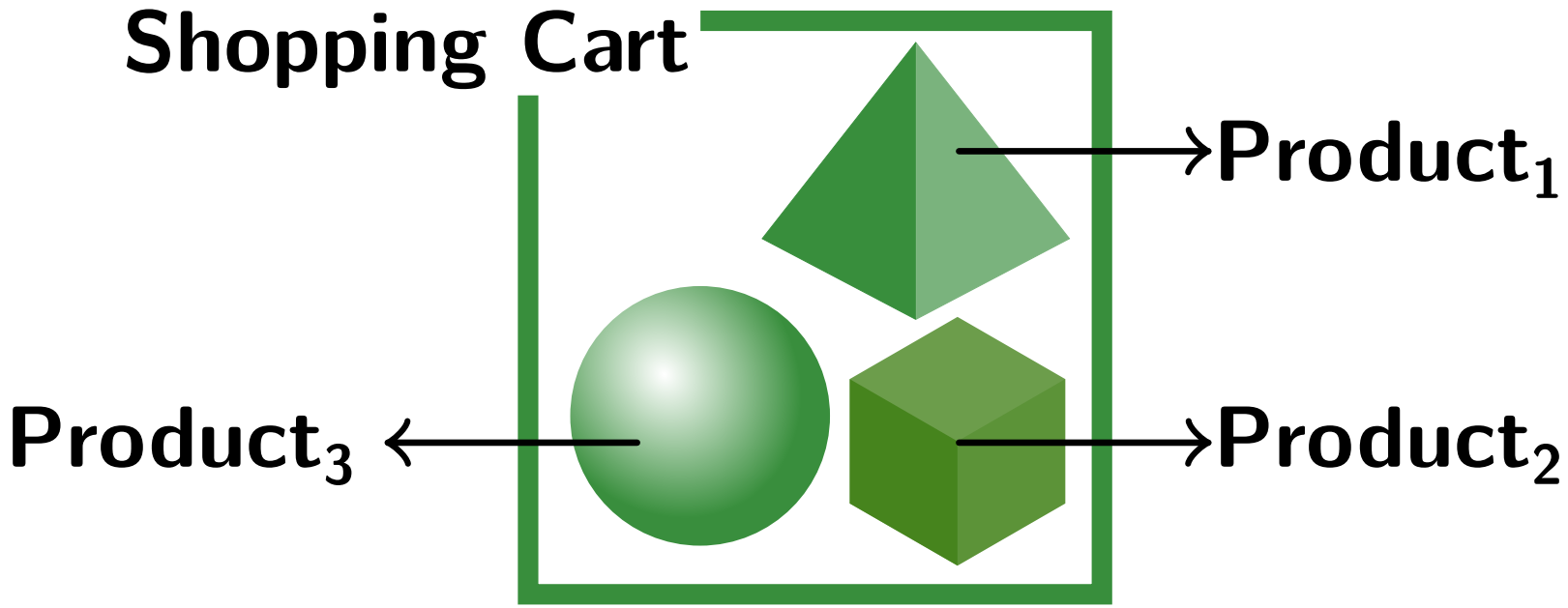
- How many Circles are there that contain a "Red Pyramid"?
- How many objects are there that have a color "Yellow"?
- Count the number of Squares and Circles that contain either a "Yellow Cube" or "Red Sphere"?

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

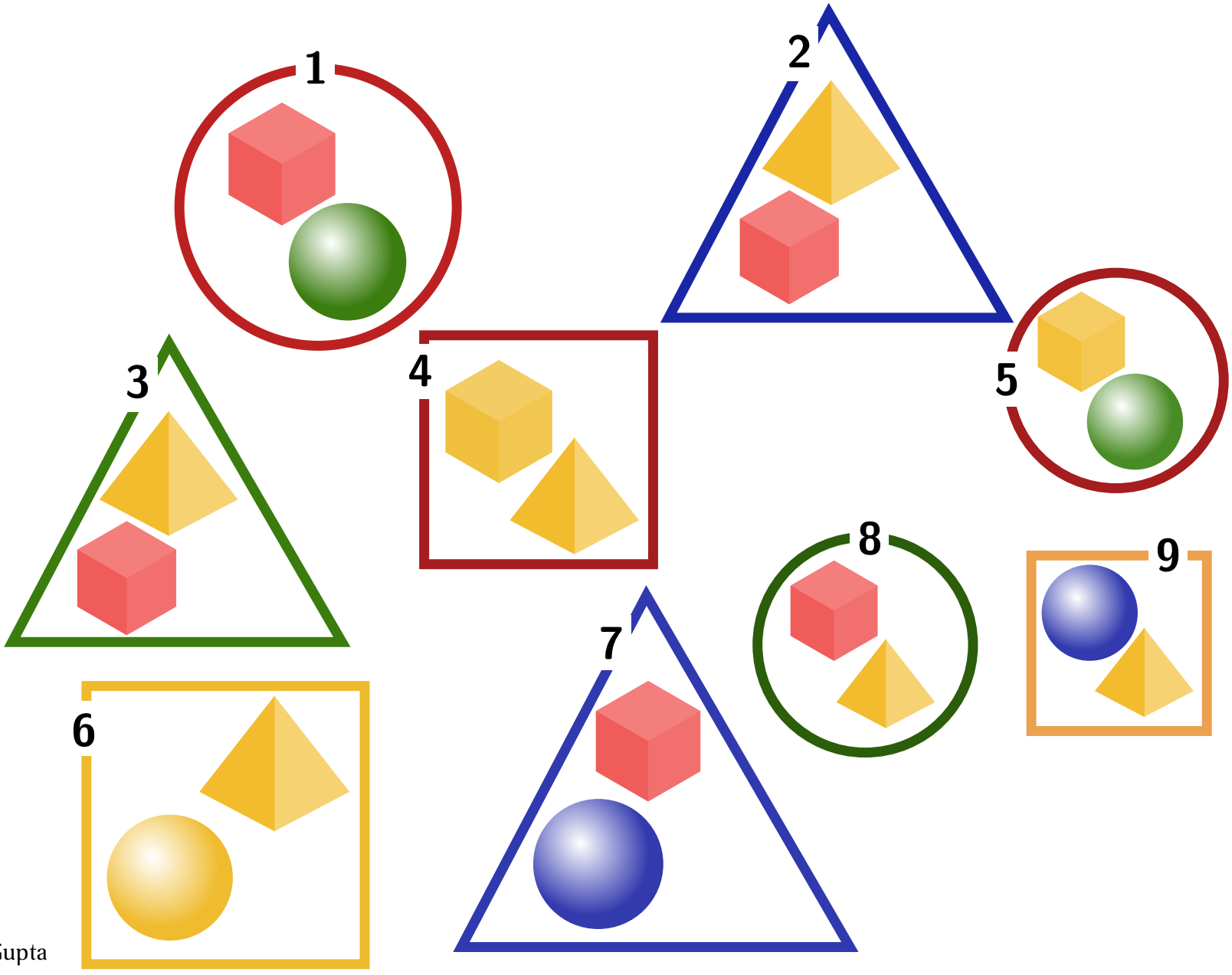
Data Warehouses — Online Analytical Processing



Association Rules — Shopping Cart



Association Rules — Transactions

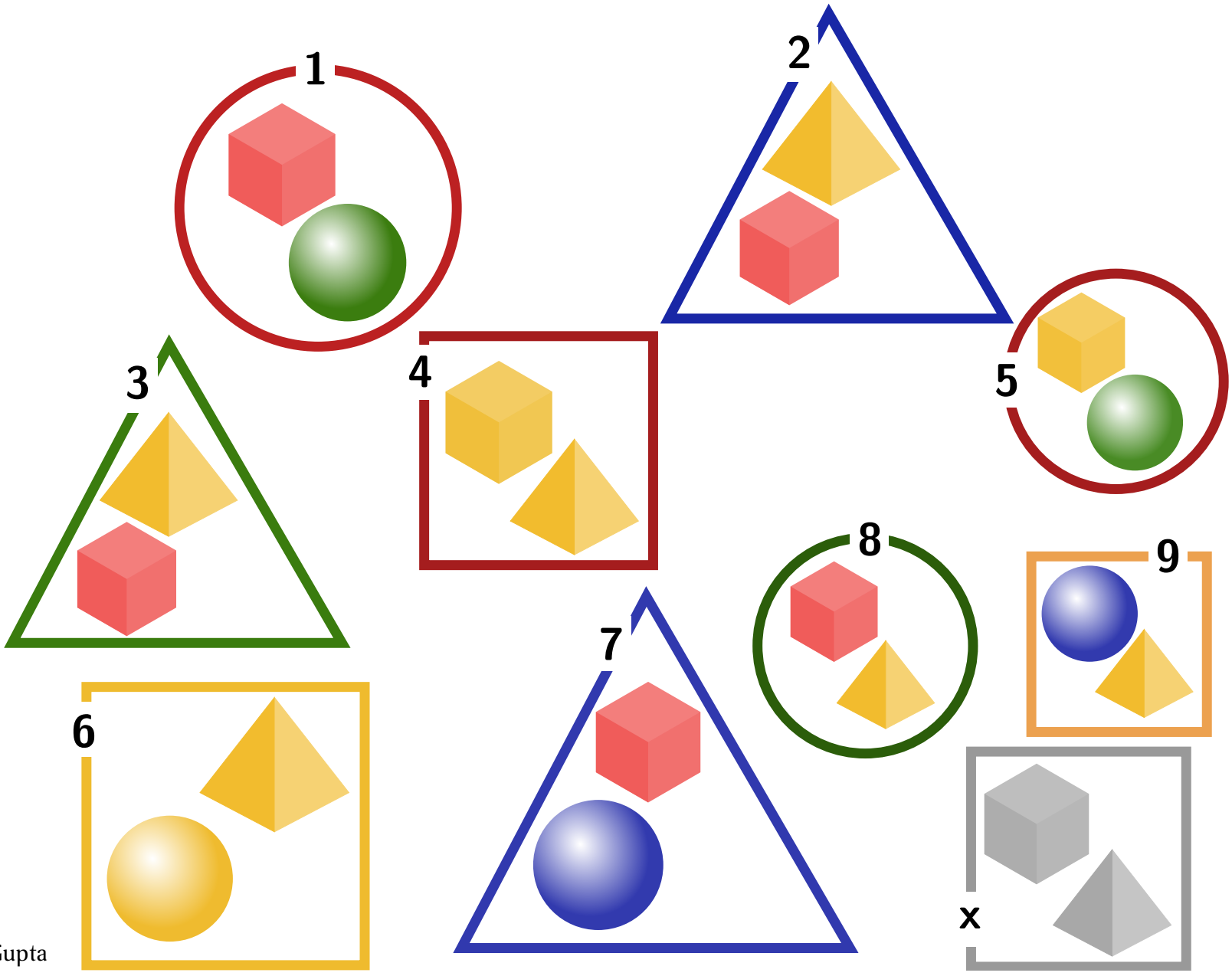


Association Rules

- Consider each row in the table below as a shopping transaction.
- Identify sets of products that are commonly bought together.
- Identify such sets for recommendation or placement of products in physical stores.
- For example, in a grocery store bread and eggs will be physically placed together as customer frequently buy them together.

ID	Object_1	Object_2
1	Cube	Sphere
2	Pyramid	Cube
3	Pyramid	Cube
4	Cube	Pyramid
5	Cube	Sphere
6	Pyramid	Sphere
7	Cube	Sphere
8	Cube	Pyramid
9	Sphere	Pyramid

Classification — Toy Example

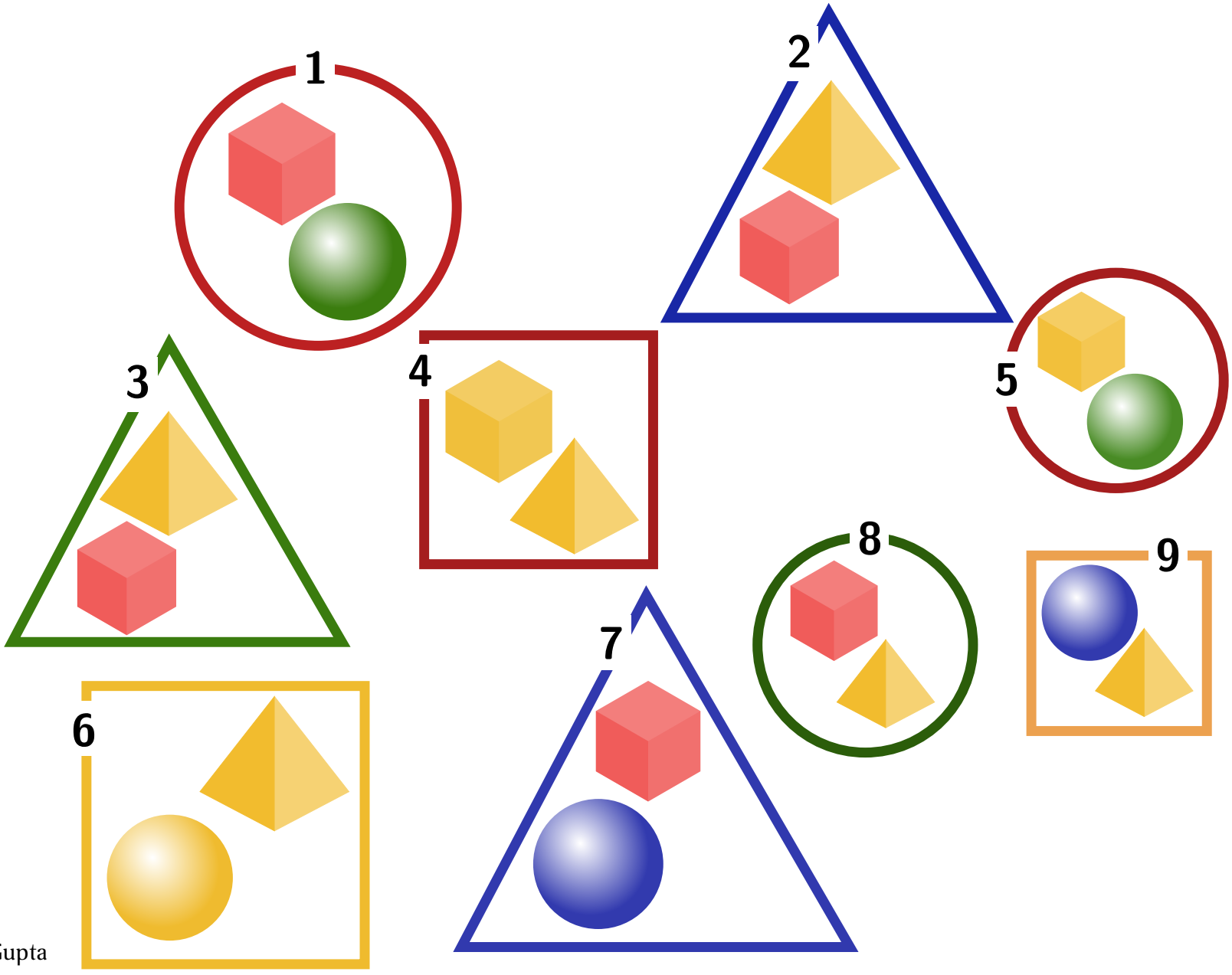


Classification — Supervised Learning

- Consider the example observations from earlier.
- Given this set of data can we learn something, so that when a new observation is made (e.g., "x") we can predict its other characteristics?

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow
x	Square	?	Cube	?	Pyramid	?

Clustering — Toy Example

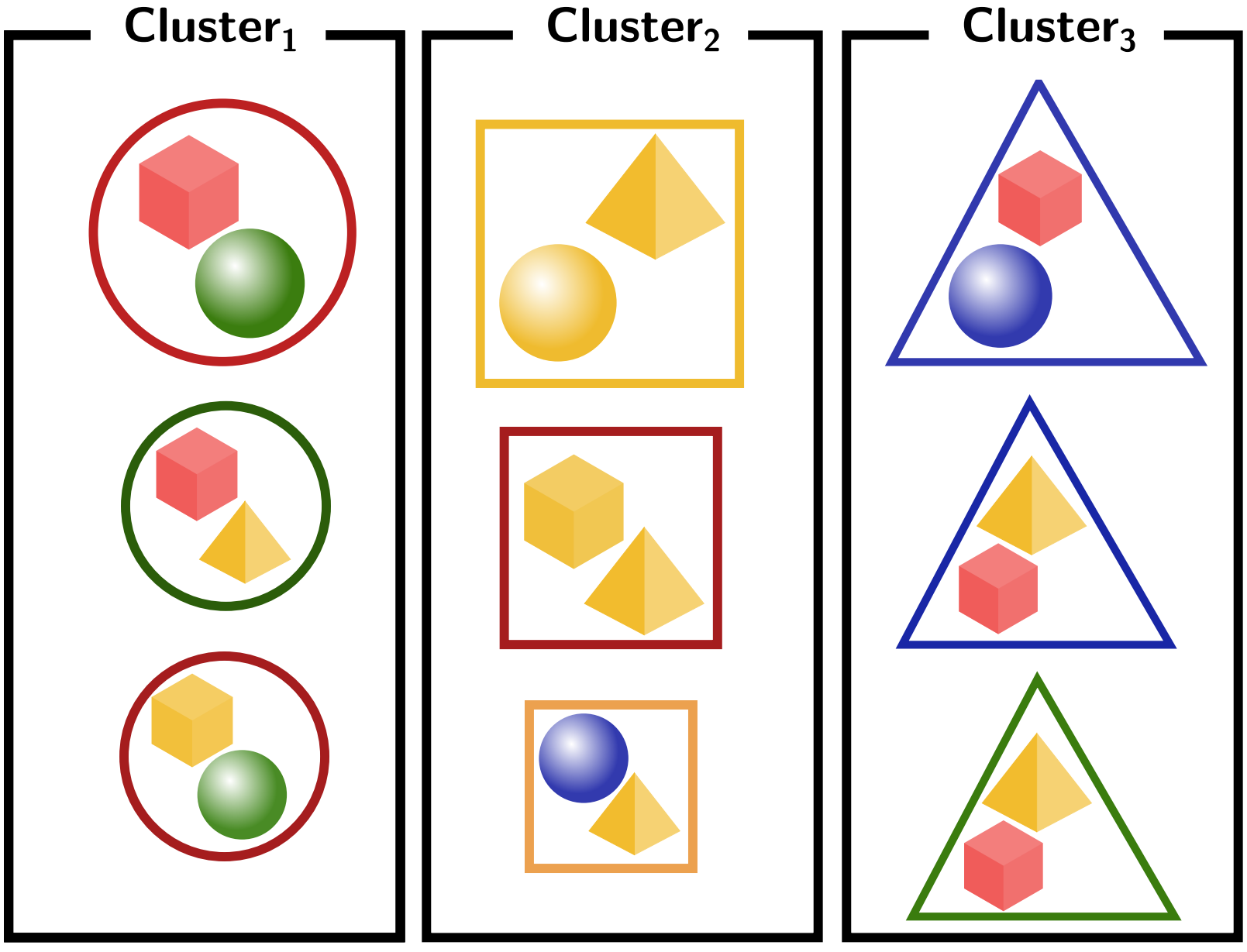


Clustering — Unsupervised Learning

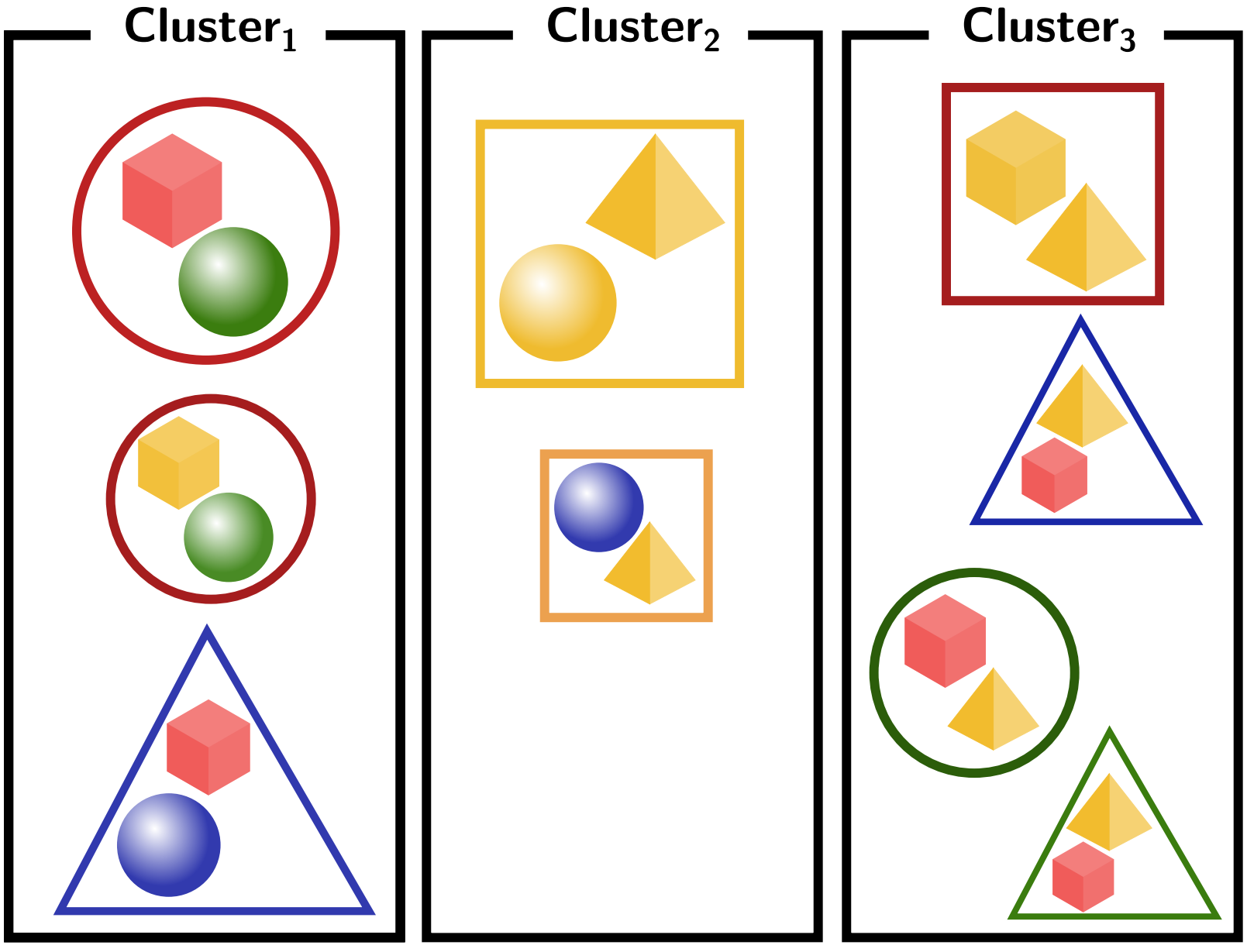
- Consider the example observations from earlier.
- Given this set of data, how can we organize this data based on its attributes so that related observations are placed together and unrelated observations are excluded?

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

Example Clustering — Based on Shapes

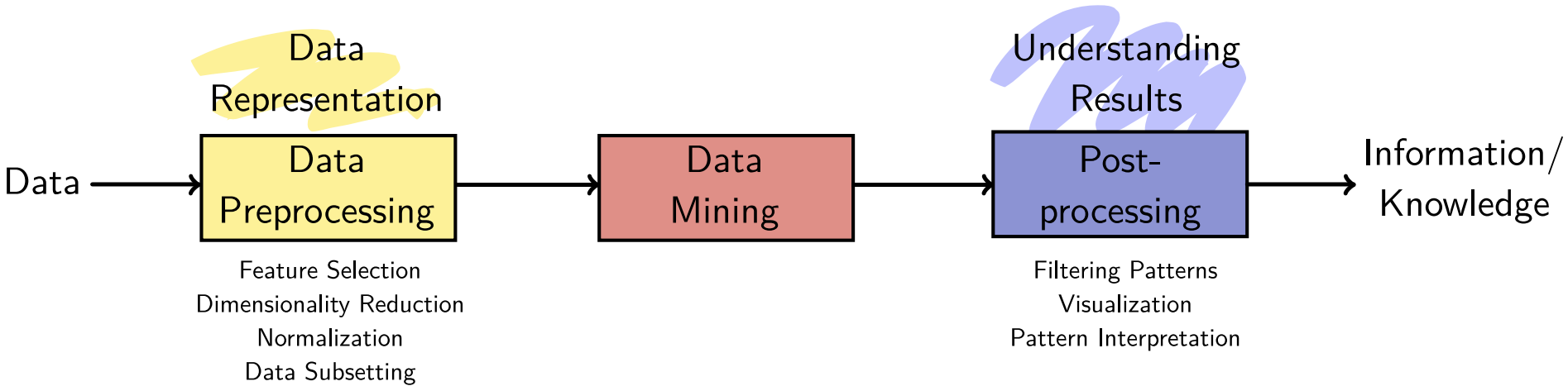


Example Clustering — Based on Objects



Summary

- Data is growing at a rapid pace.
  However, our capability to analyze it is limited.
- Key challenge in analyzing massive amounts of data is scalability.
- Data management systems offer the capability to organize the data effectively for efficient data analysis.
- Objective of data mining is to discover insights into Big Data either through user-driven exploration or automatically through algorithms.
- This Knowledge Discovery process can be summarized as follows[9]:



[9] Tan et al. Introduction to Data Mining (1st Edition). 2006.