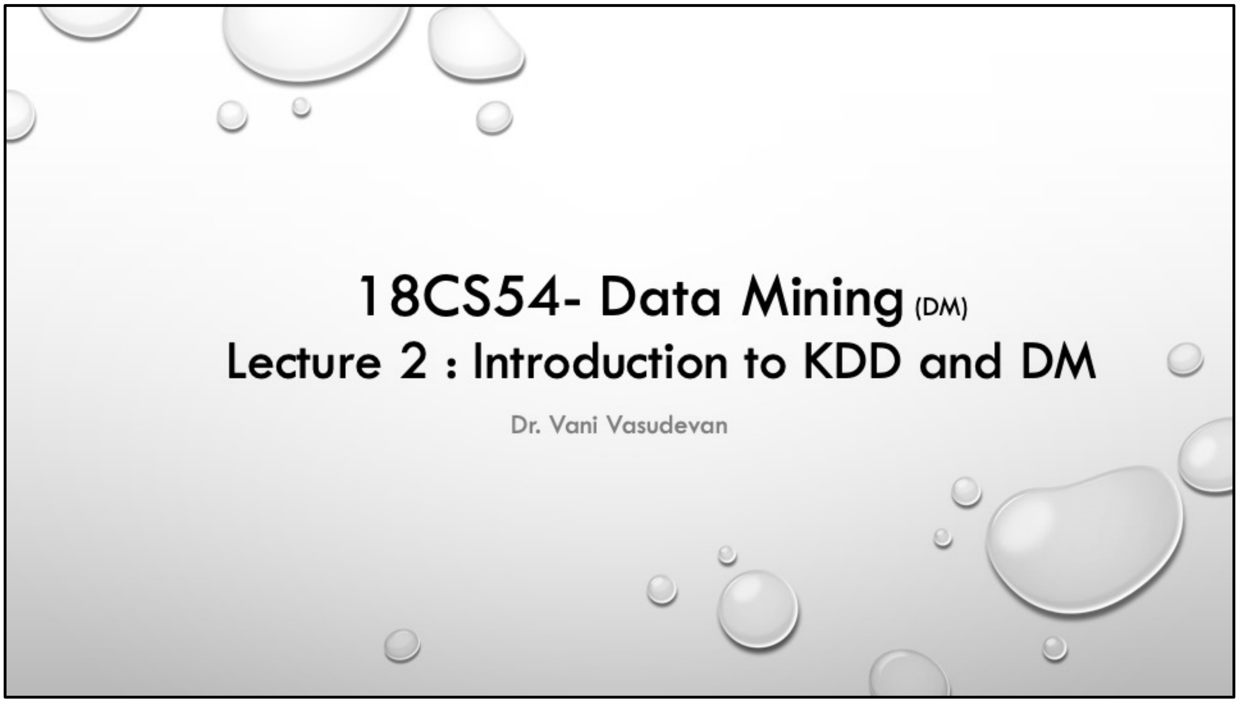
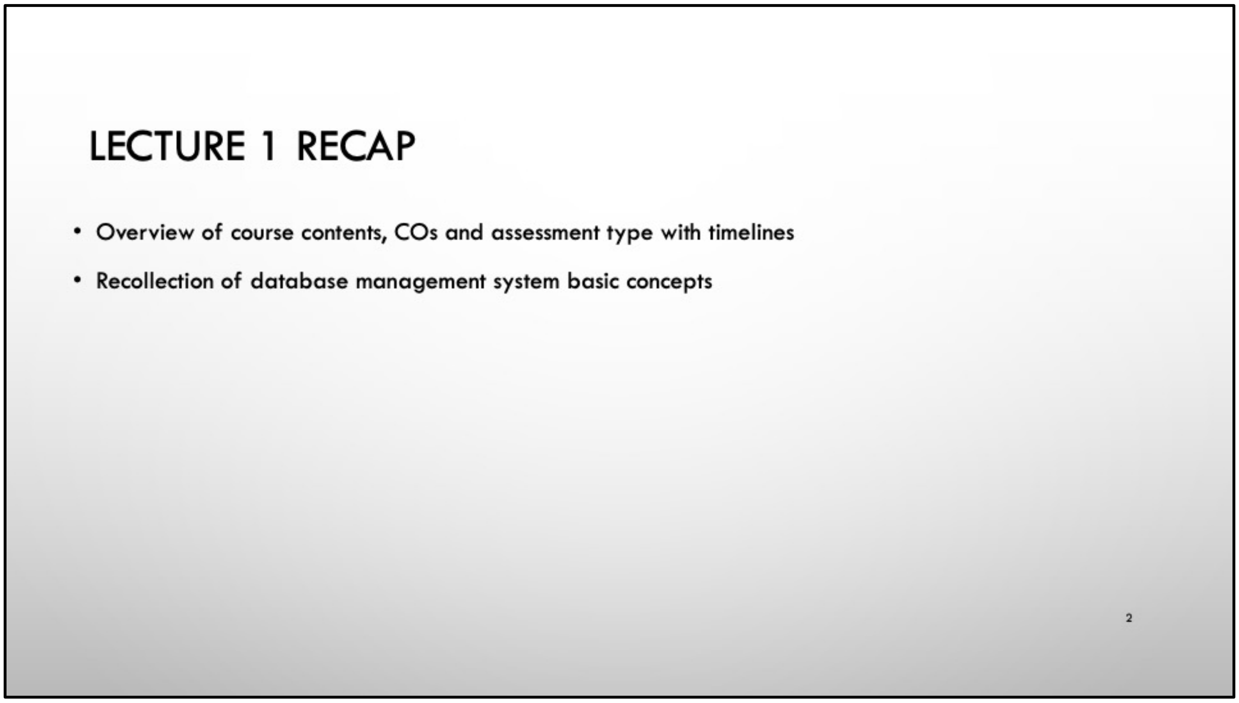


18CS54- Data Mining (DM)

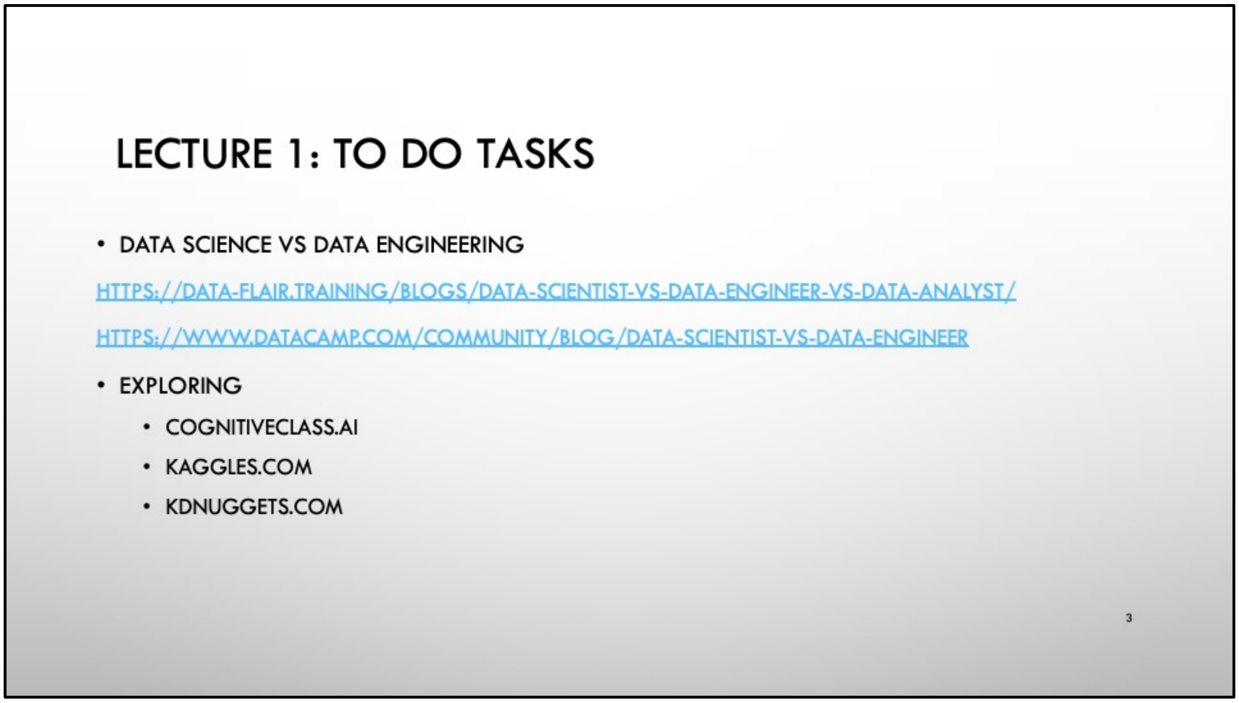
Lecture 2 : Introduction to KDD and DM

Dr. Vani Vasudevan



LECTURE 1 RECAP

- Overview of course contents, COs and assessment type with timelines
- Recollection of database management system basic concepts



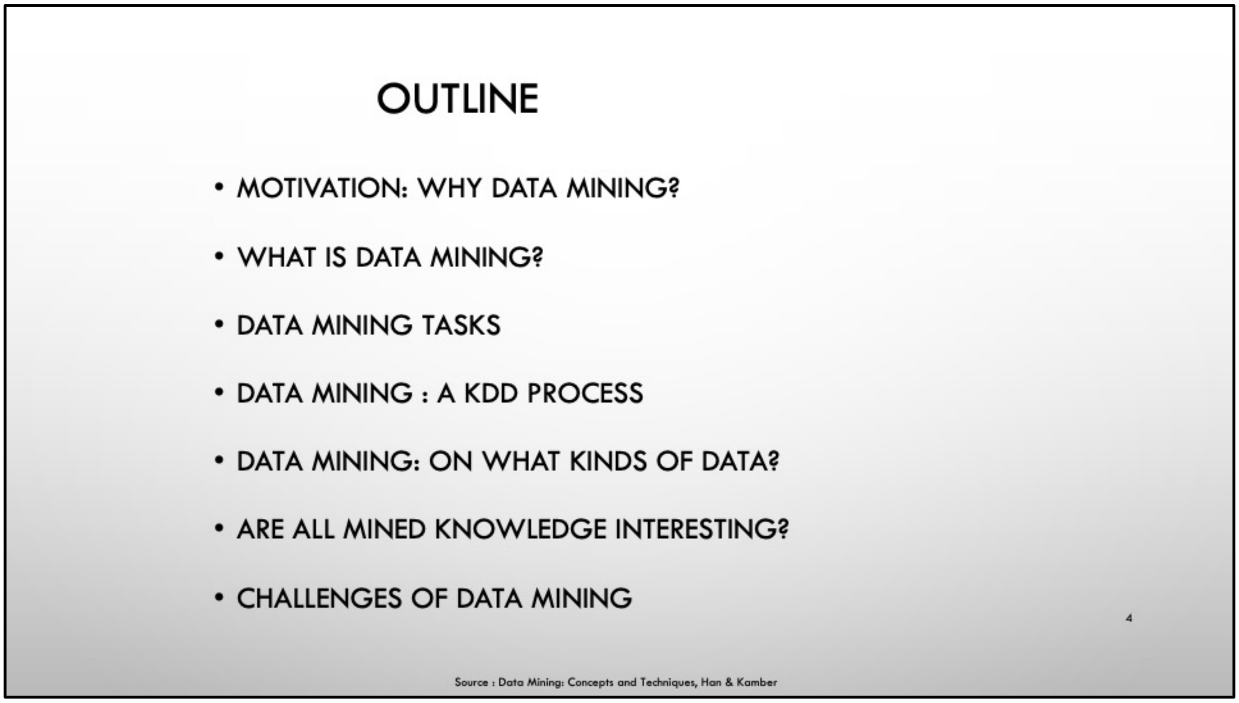
LECTURE 1: TO DO TASKS

- DATA SCIENCE VS DATA ENGINEERING

HTTPS://DATA-FLAIR.TRAINING/BLOGS/DATA-SCIENTIST-VS-DATA-ENGINEER-VS-DATA-ANALYST/

HTTPS://WWW.DATACAMP.COM/COMMUNITY/BLOG/DATA-SCIENTIST-VS-DATA-ENGINEER

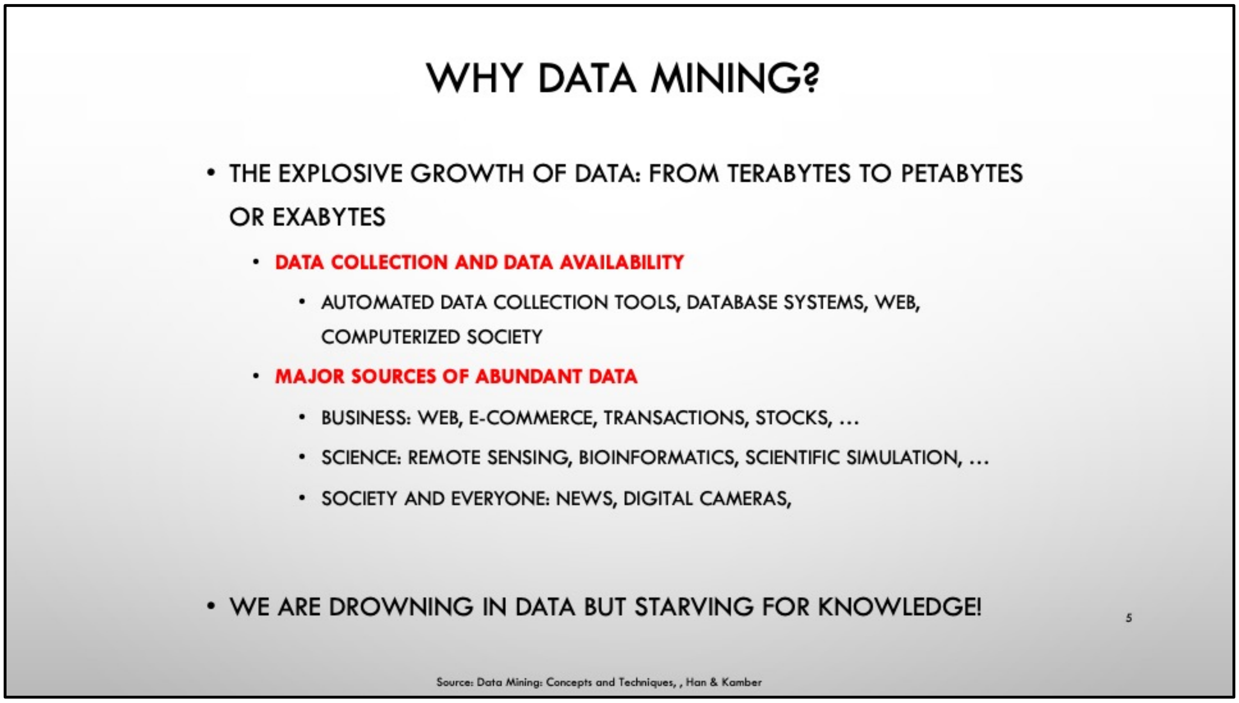
- EXPLORING
  - COGNITIVECLASS.AI
  - KAGGLES.COM
  - KDNUGGETS.COM



OUTLINE

- MOTIVATION: WHY DATA MINING?
- WHAT IS DATA MINING?
- DATA MINING TASKS
- DATA MINING : A KDD PROCESS
- DATA MINING: ON WHAT KINDS OF DATA?
- ARE ALL MINED KNOWLEDGE INTERESTING?
- CHALLENGES OF DATA MINING

Source : Data Mining: Concepts and Techniques, Han & Kamber

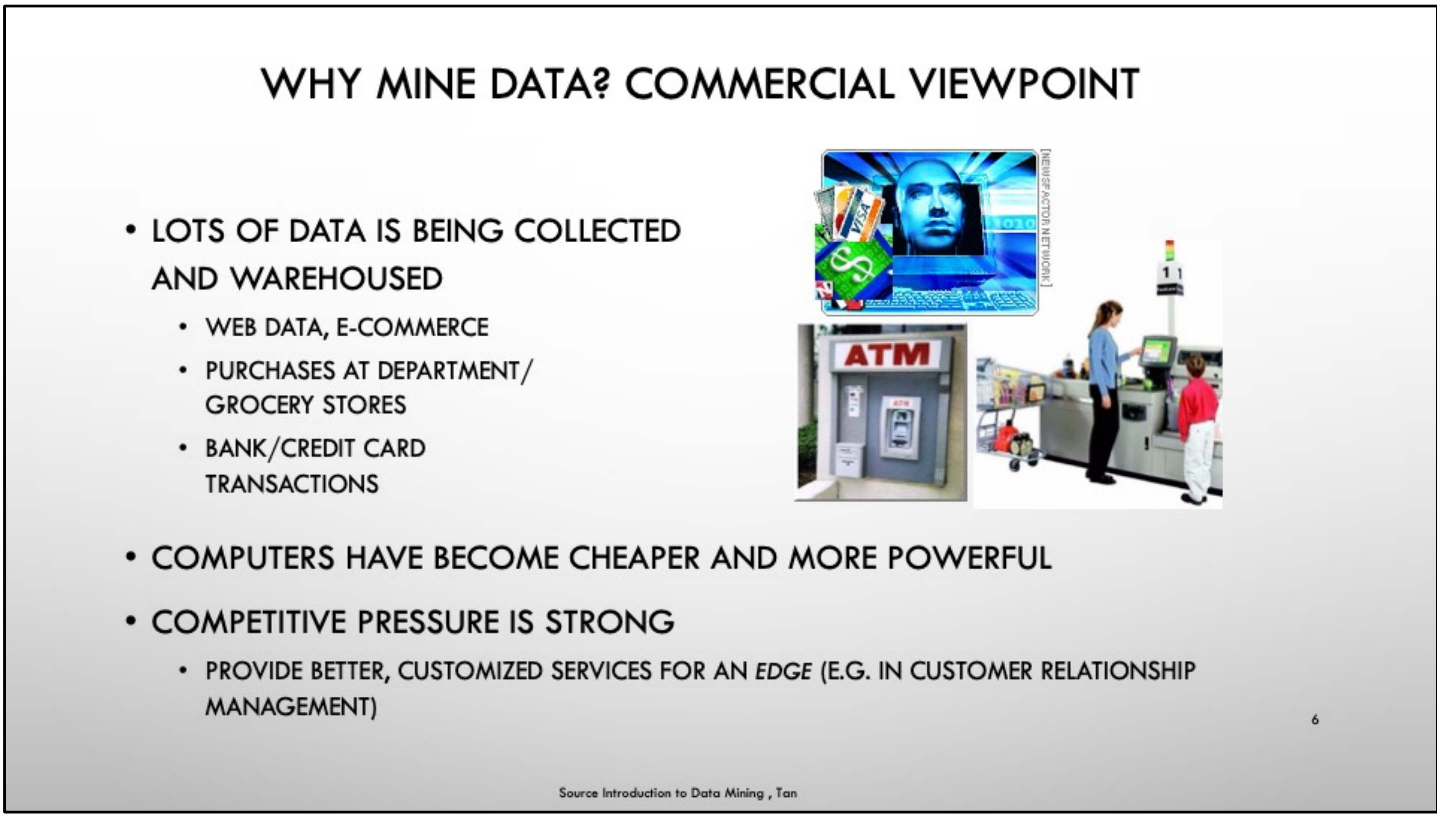


WHY DATA MINING?

- THE EXPLOSIVE GROWTH OF DATA: FROM TERABYTES TO PETABYTES OR EXABYTES
  - **DATA COLLECTION AND DATA AVAILABILITY**
    - AUTOMATED DATA COLLECTION TOOLS, DATABASE SYSTEMS, WEB, COMPUTERIZED SOCIETY
  - **MAJOR SOURCES OF ABUNDANT DATA**
    - BUSINESS: WEB, E-COMMERCE, TRANSACTIONS, STOCKS, …
    - SCIENCE: REMOTE SENSING, BIOINFORMATICS, SCIENTIFIC SIMULATION, …
    - SOCIETY AND EVERYONE: NEWS, DIGITAL CAMERAS,

- WE ARE DROWNING IN DATA BUT STARVING FOR KNOWLEDGE!

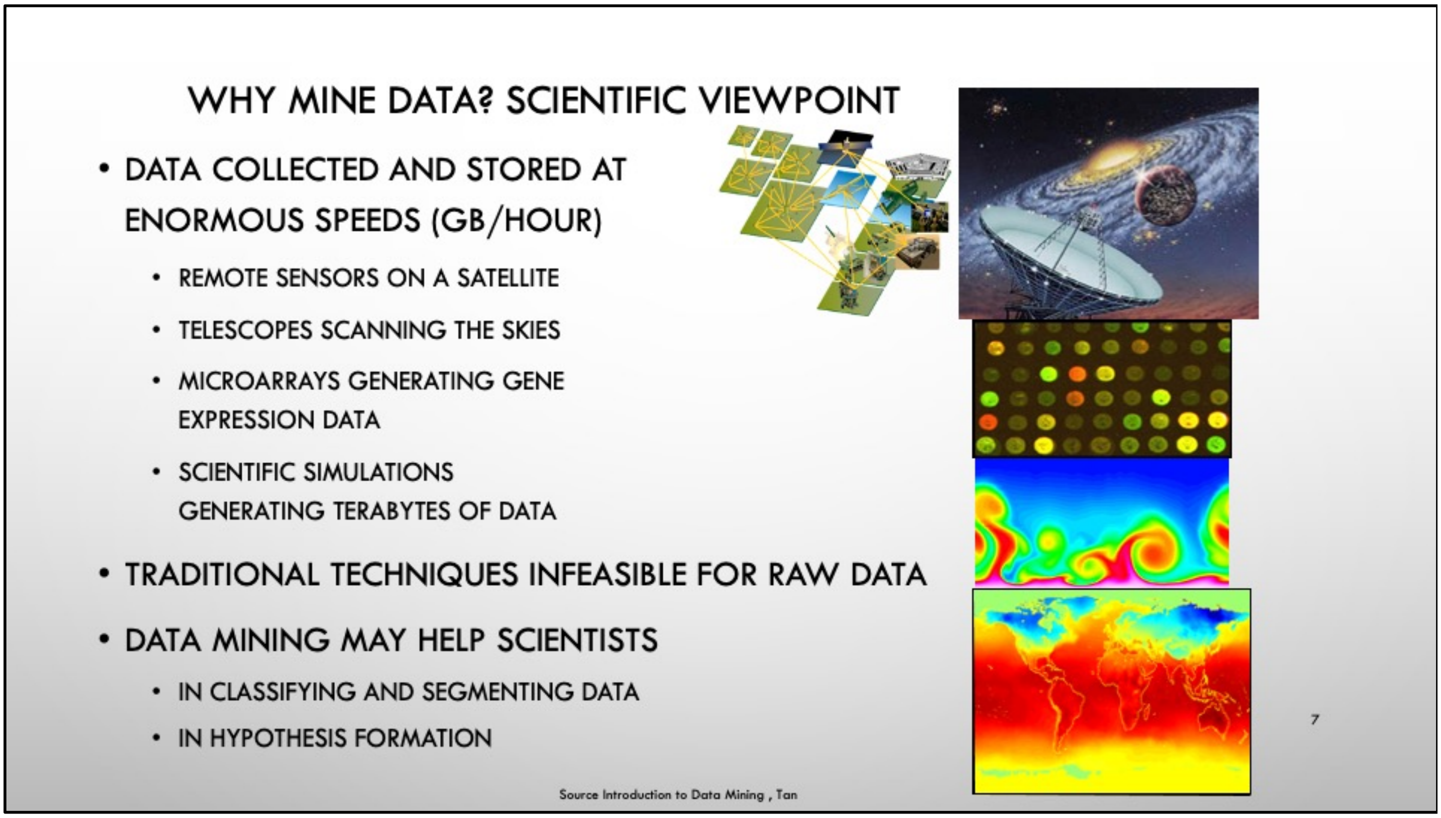
Source: Data Mining: Concepts and Techniques, , Han & Kamber



WHY MINE DATA? COMMERCIAL VIEWPOINT

- LOTS OF DATA IS BEING COLLECTED AND WAREHOUSED
  - WEB DATA, E-COMMERCE
  - PURCHASES AT DEPARTMENT/ GROCERY STORES
  - BANK/CREDIT CARD TRANSACTIONS
- COMPUTERS HAVE BECOME CHEAPER AND MORE POWERFUL
- COMPETITIVE PRESSURE IS STRONG
  - PROVIDE BETTER, CUSTOMIZED SERVICES FOR AN *EDGE* (E.G. IN CUSTOMER RELATIONSHIP MANAGEMENT)

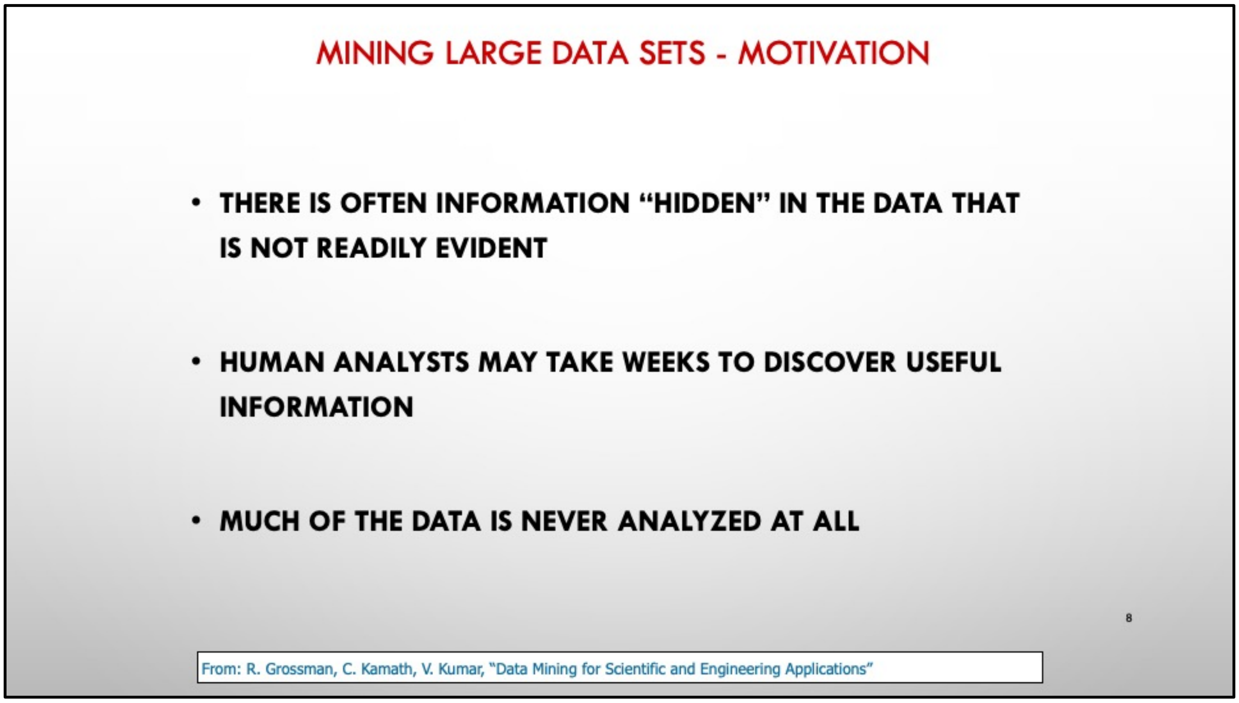
Source Introduction to Data Mining , Tan



WHY MINE DATA? SCIENTIFIC VIEWPOINT

- DATA COLLECTED AND STORED AT ENORMOUS SPEEDS (GB/HOUR)
  - REMOTE SENSORS ON A SATELLITE
  - TELESCOPES SCANNING THE SKIES
  - MICROARRAYS GENERATING GENE EXPRESSION DATA
  - SCIENTIFIC SIMULATIONS GENERATING TERABYTES OF DATA
- TRADITIONAL TECHNIQUES INFEASIBLE FOR RAW DATA
- DATA MINING MAY HELP SCIENTISTS
  - IN CLASSIFYING AND SEGMENTING DATA
  - IN HYPOTHESIS FORMATION

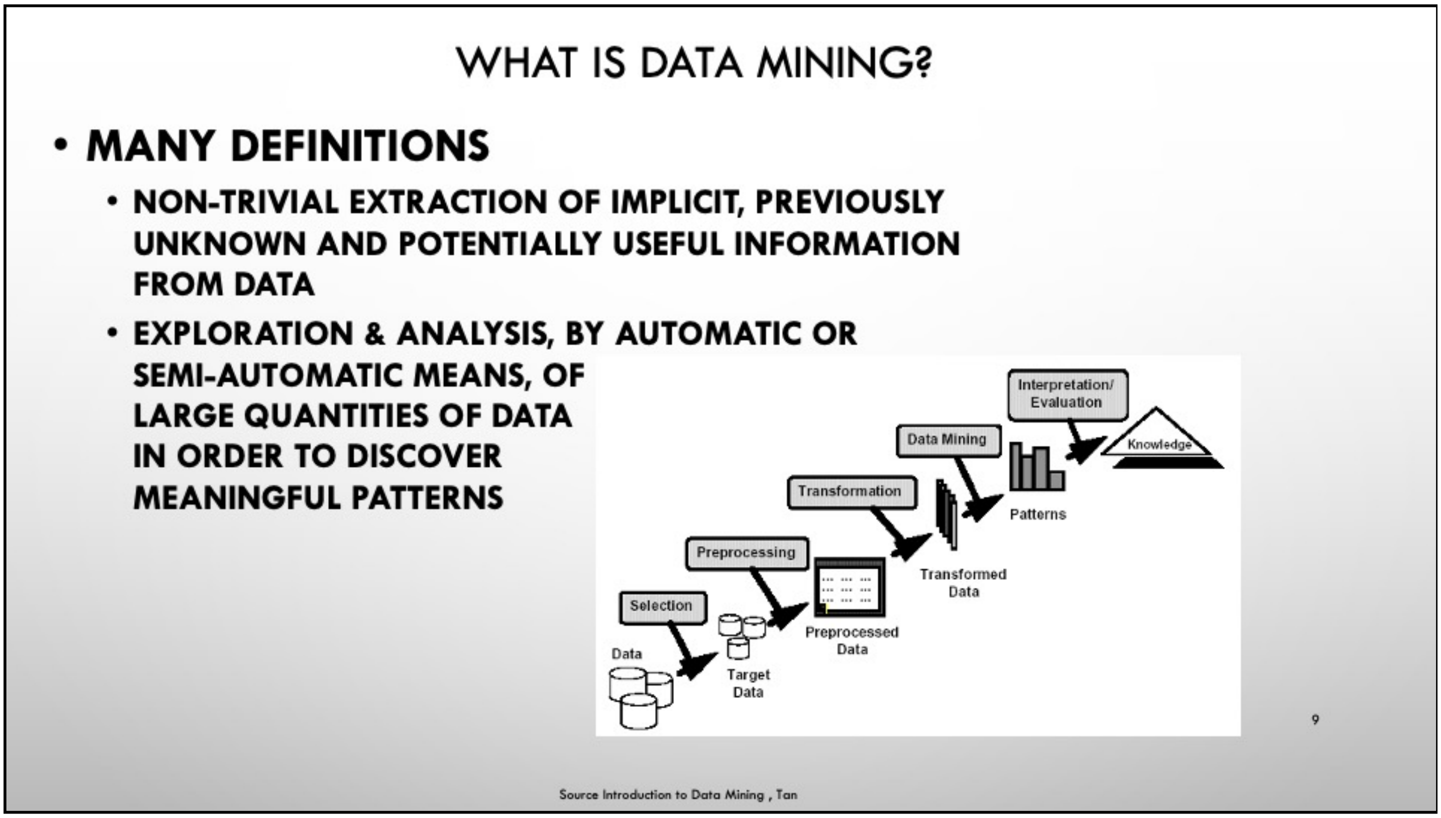
Source Introduction to Data Mining , Tan



MINING LARGE DATA SETS - MOTIVATION

- THERE IS OFTEN INFORMATION “HIDDEN” IN THE DATA THAT IS NOT READILY EVIDENT
- HUMAN ANALYSTS MAY TAKE WEEKS TO DISCOVER USEFUL INFORMATION
- MUCH OF THE DATA IS NEVER ANALYZED AT ALL

From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

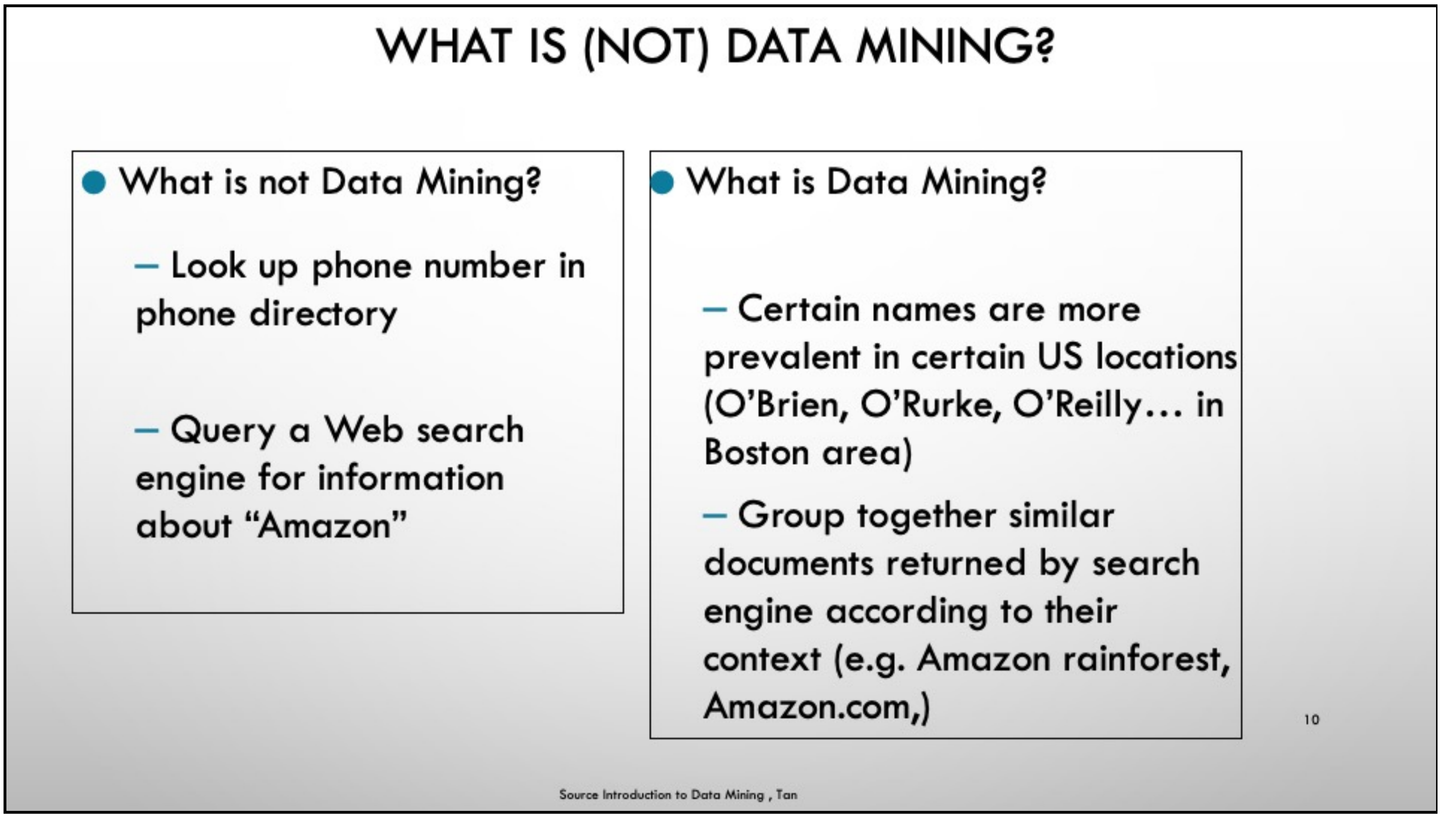


WHAT IS DATA MINING?

- **MANY DEFINITIONS**
  - **NON-TRIVIAL EXTRACTION OF IMPLICIT, PREVIOUSLY UNKNOWN AND POTENTIALLY USEFUL INFORMATION FROM DATA**
  - **EXPLORATION & ANALYSIS, BY AUTOMATIC OR SEMI-AUTOMATIC MEANS, OF LARGE QUANTITIES OF DATA IN ORDER TO DISCOVER MEANINGFUL PATTERNS**

Source Introduction to Data Mining , Tan

What Is Data Mining?
Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation, such as predicting whether a newly arrived customer will spend more than $100 at a department store - Tan

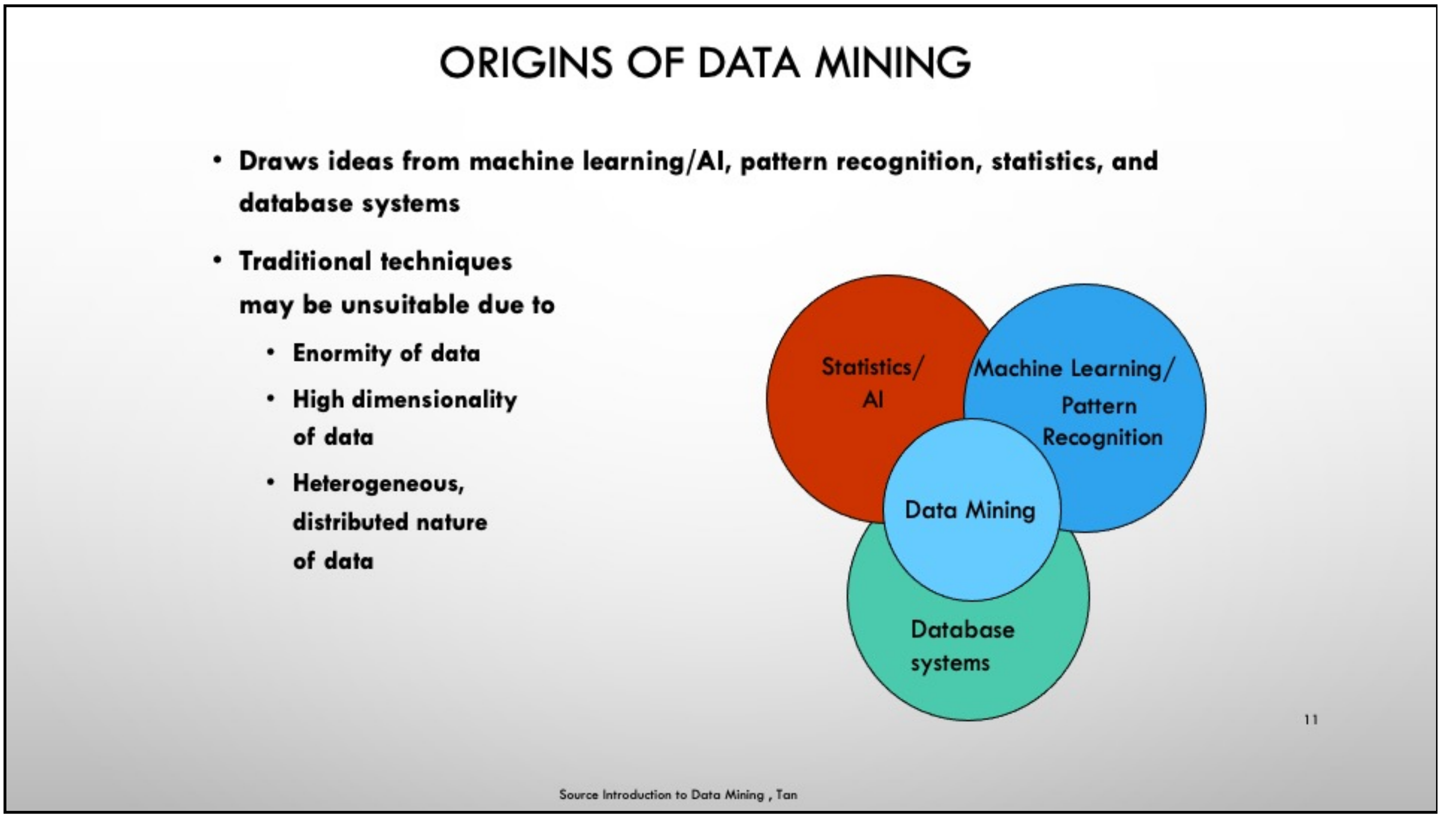


WHAT IS (NOT) DATA MINING?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

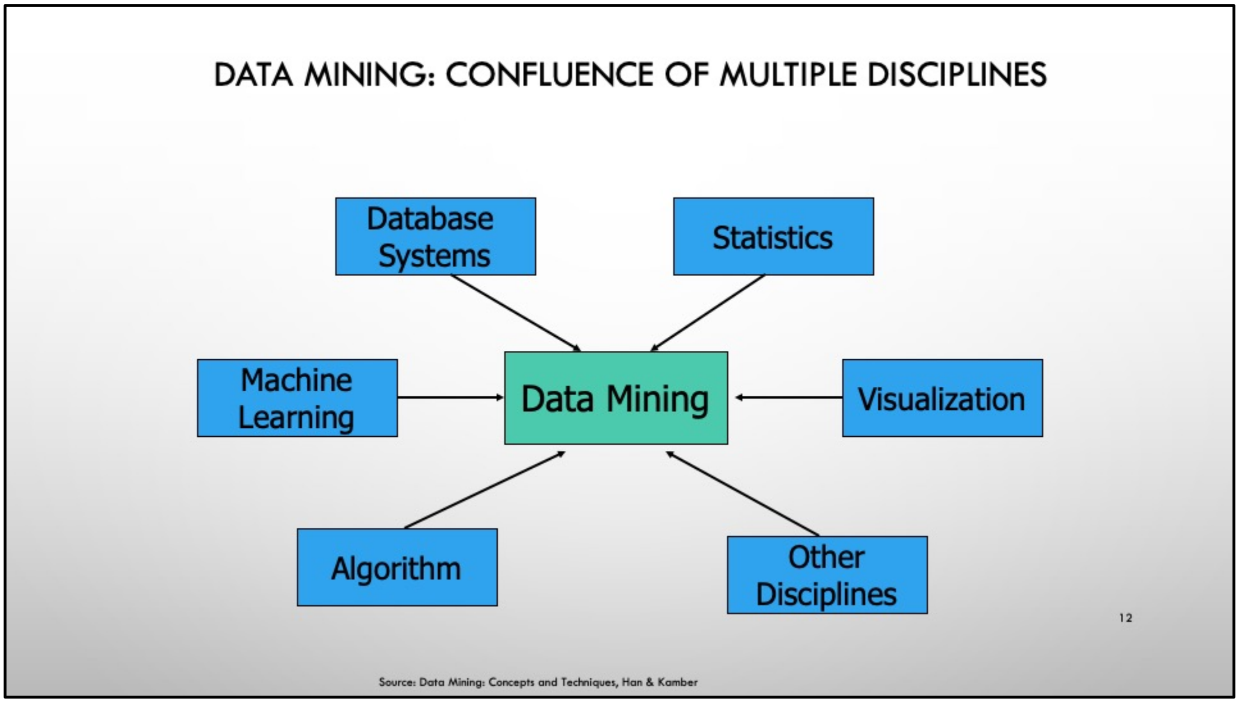
Source Introduction to Data Mining , Tan



ORIGINS OF DATA MINING

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**

Source Introduction to Data Mining , Tan



DATA MINING: CONFLUENCE OF MULTIPLE DISCIPLINES

Database Systems

Statistics

Machine Learning

Data Mining

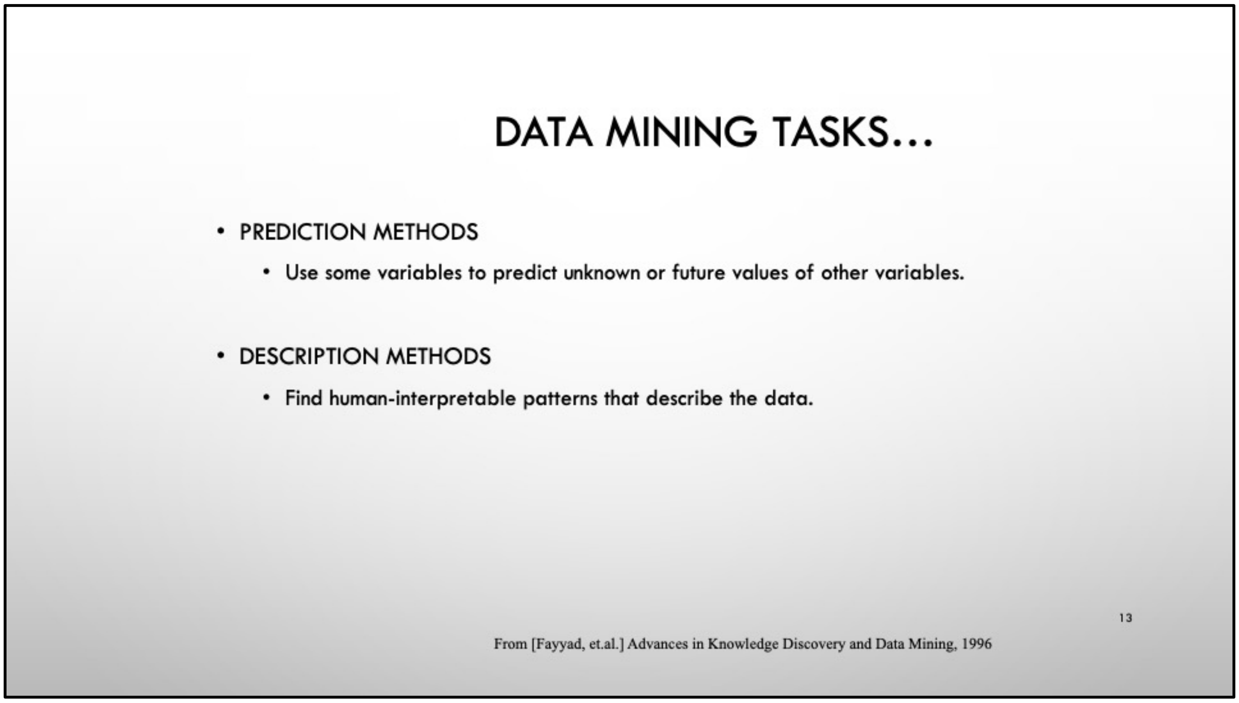
Visualization

Algorithm

Other Disciplines

12

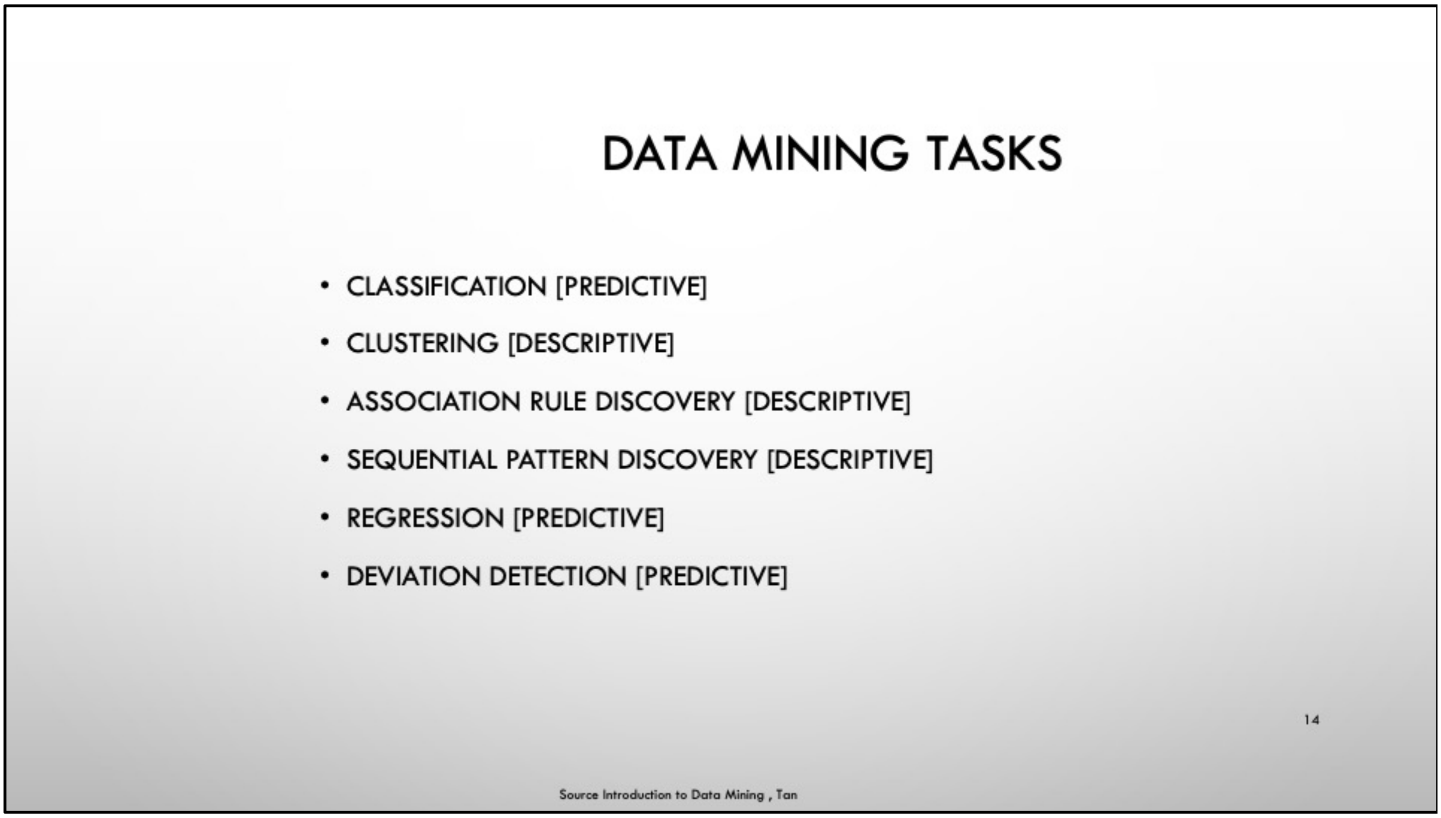
Source: Data Mining: Concepts and Techniques, Han & Kamber



DATA MINING TASKS…

- PREDICTION METHODS
  - Use some variables to predict unknown or future values of other variables.
- DESCRIPTION METHODS
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996



DATA MINING TASKS

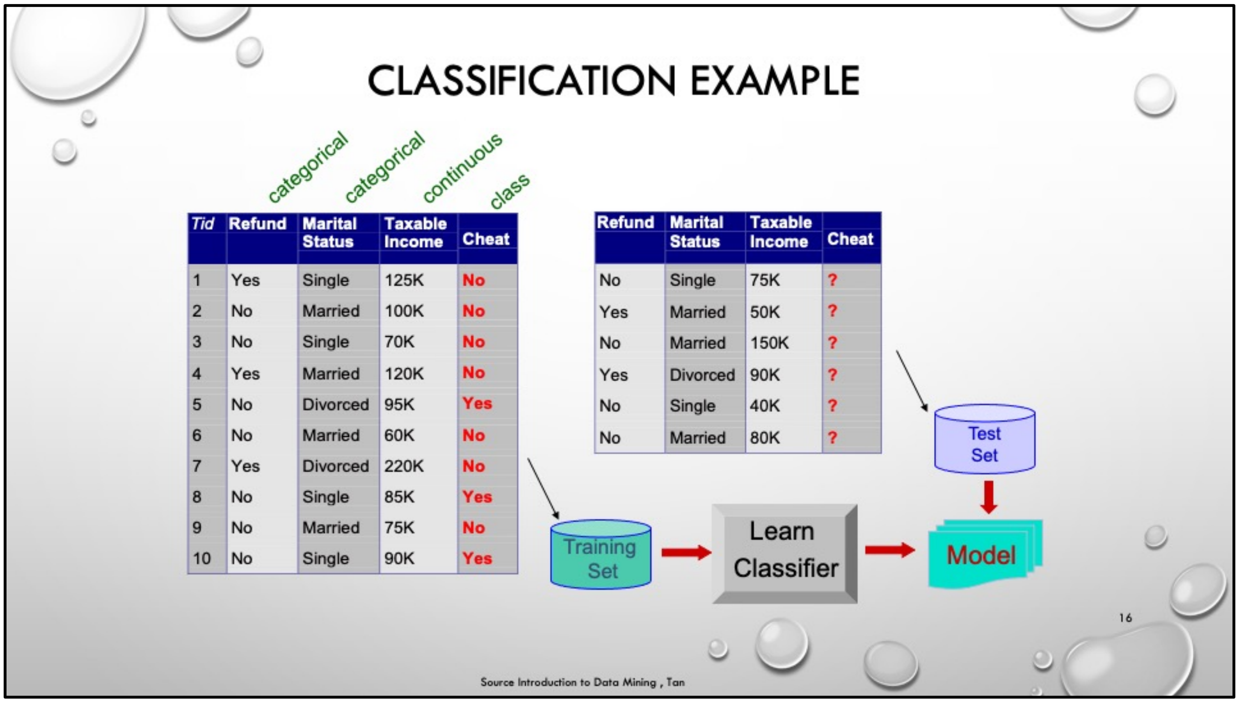
- CLASSIFICATION [PREDICTIVE]
- CLUSTERING [DESCRIPTIVE]
- ASSOCIATION RULE DISCOVERY [DESCRIPTIVE]
- SEQUENTIAL PATTERN DISCOVERY [DESCRIPTIVE]
- REGRESSION [PREDICTIVE]
- DEVIATION DETECTION [PREDICTIVE]

Source Introduction to Data Mining , Tan

CLASSIFICATION: DEFINITION

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Source Introduction to Data Mining , Tan



CLASSIFICATION EXAMPLE

categorical categorical continuous class

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**

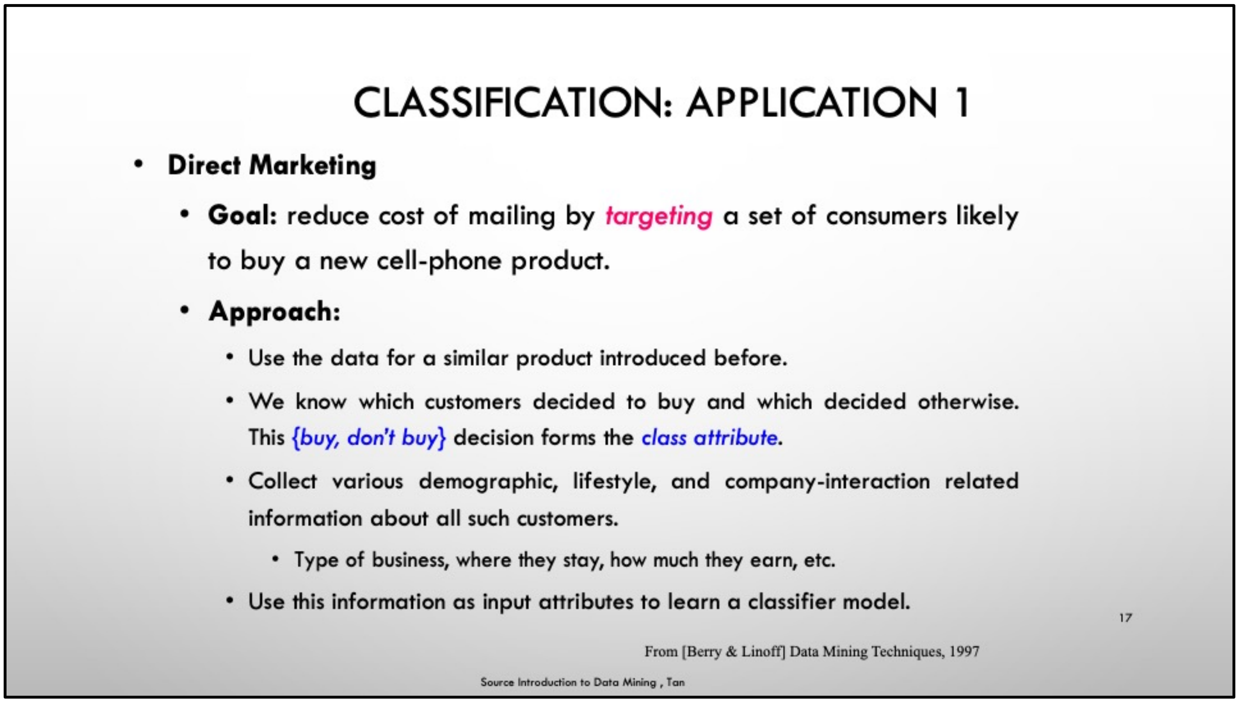
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**

Test Set

Training Set → Learn Classifier → Model

16

Source Introduction to Data Mining , Tan



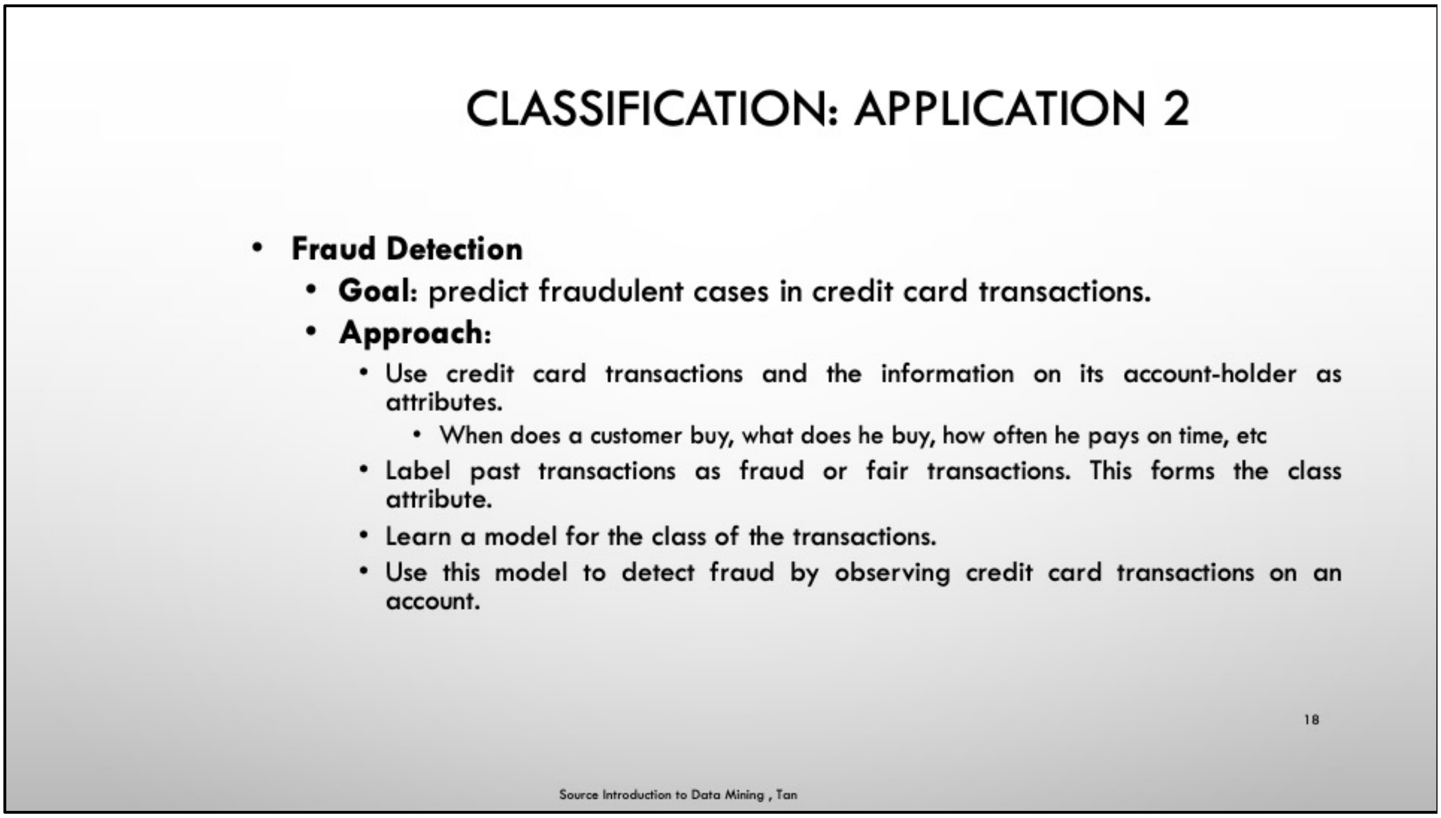
CLASSIFICATION: APPLICATION 1

- **Direct Marketing**
  - **Goal:** reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - **Approach:**
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

17

From [Berry & Linoff] Data Mining Techniques, 1997

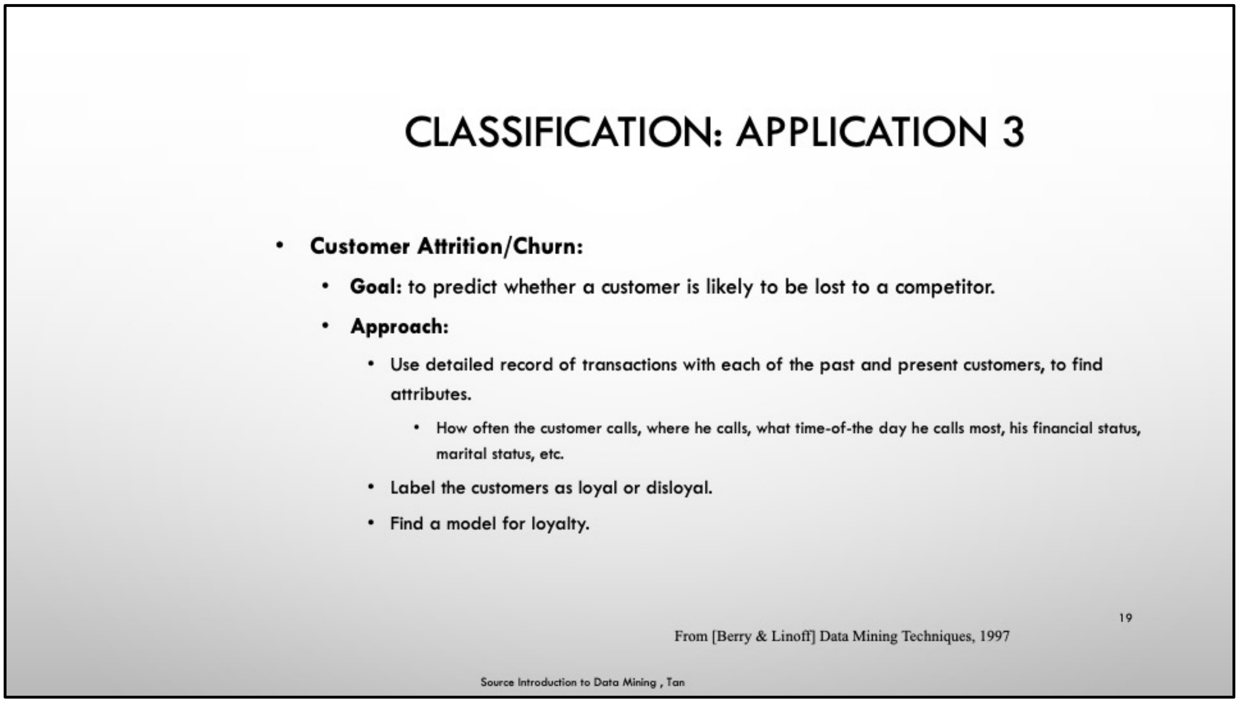
Source Introduction to Data Mining , Tan



CLASSIFICATION: APPLICATION 2

- **Fraud Detection**
  - **Goal**: predict fraudulent cases in credit card transactions.
  - **Approach**:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Source Introduction to Data Mining , Tan

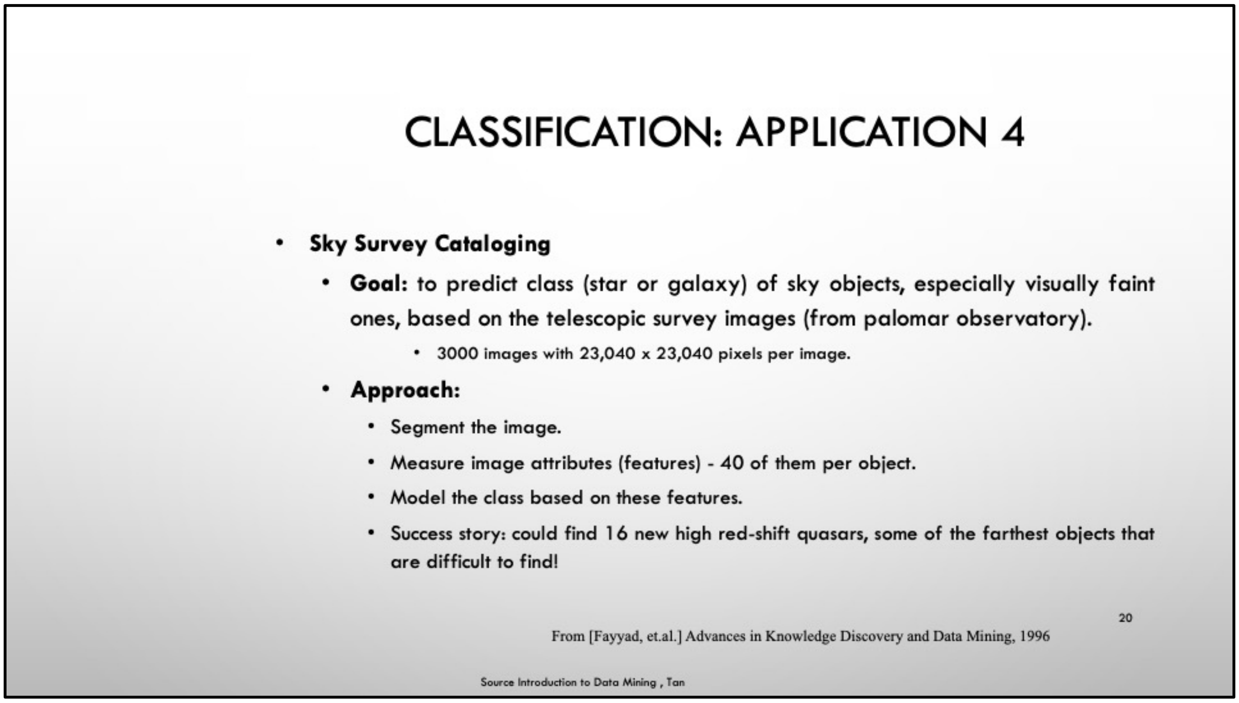


CLASSIFICATION: APPLICATION 3

- **Customer Attrition/Churn:**
  - **Goal:** to predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Source Introduction to Data Mining , Tan

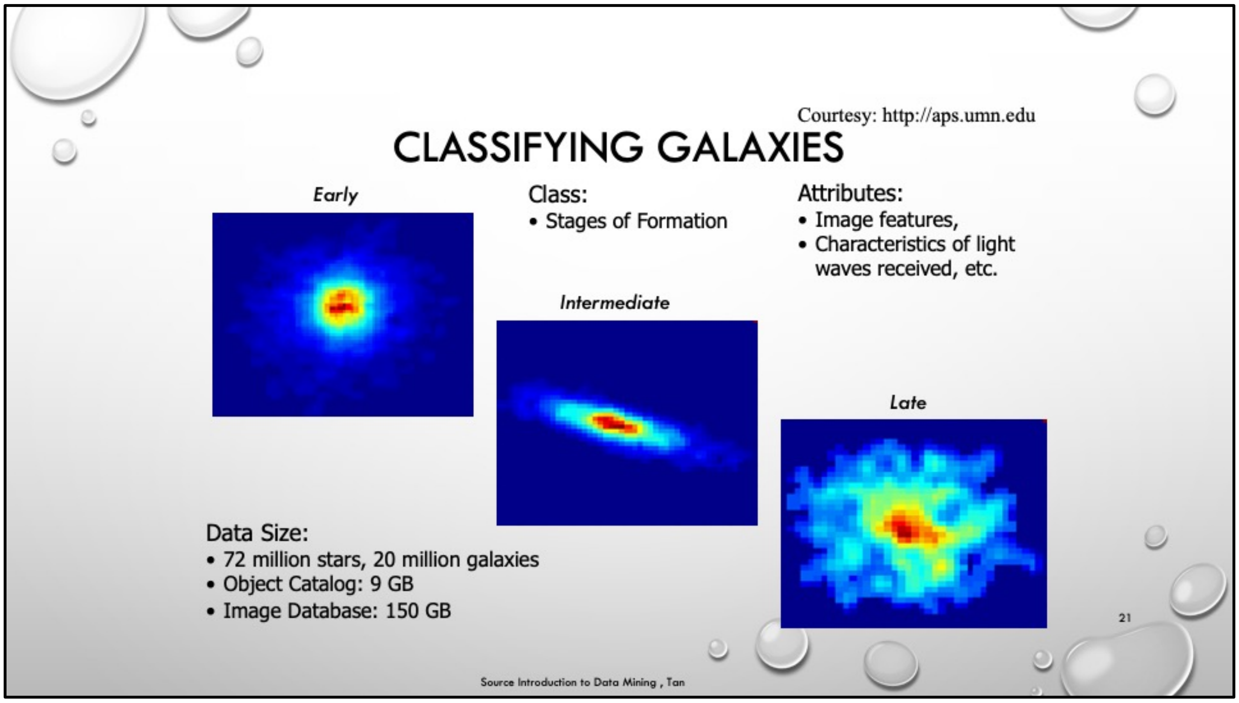


CLASSIFICATION: APPLICATION 4

- **Sky Survey Cataloging**
  - **Goal:** to predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from palomar observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success story: could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Source Introduction to Data Mining , Tan



Courtesy: http://aps.umn.edu

CLASSIFYING GALAXIES

Early

Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate

Late

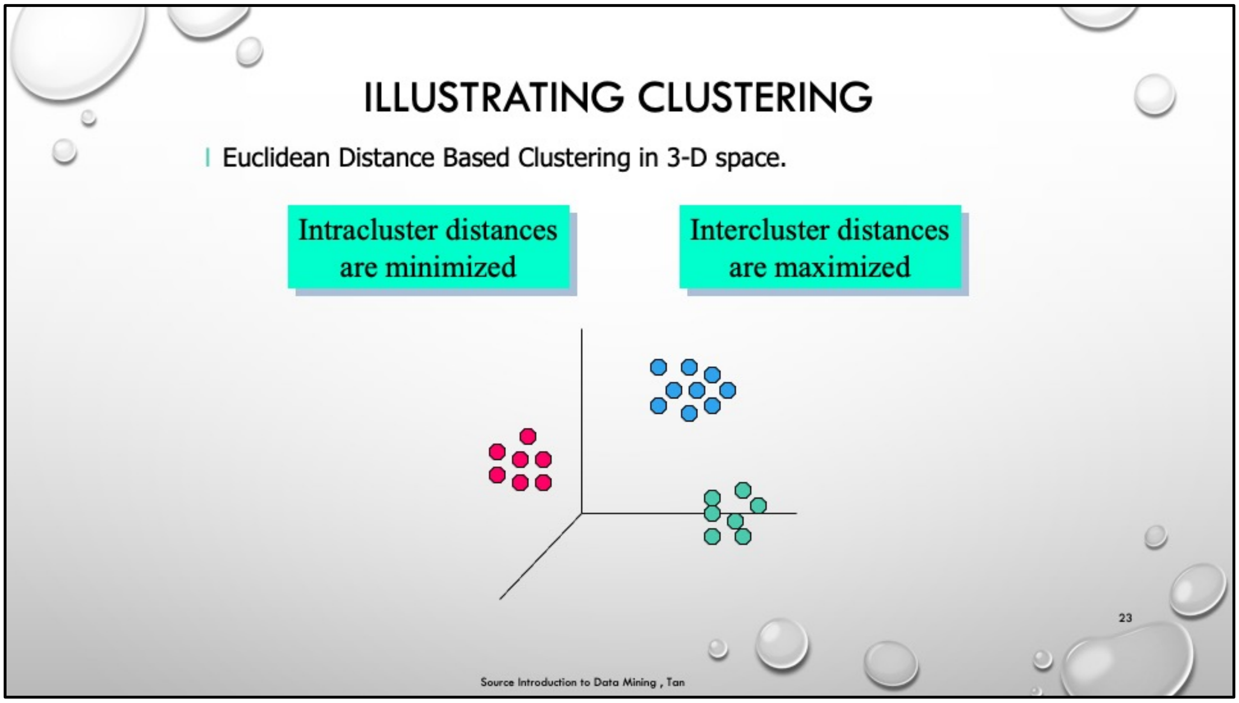
Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Source Introduction to Data Mining , Tan

CLUSTERING DEFINITION

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in **one cluster are more similar** to one another.
  - Data points in **separate clusters are less similar** to one another.
- **Similarity measures:**
  - Euclidean distance if attributes are continuous.
  - Other problem-specific measures.

Source Introduction to Data Mining , Tan



ILLUSTRATING CLUSTERING

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized

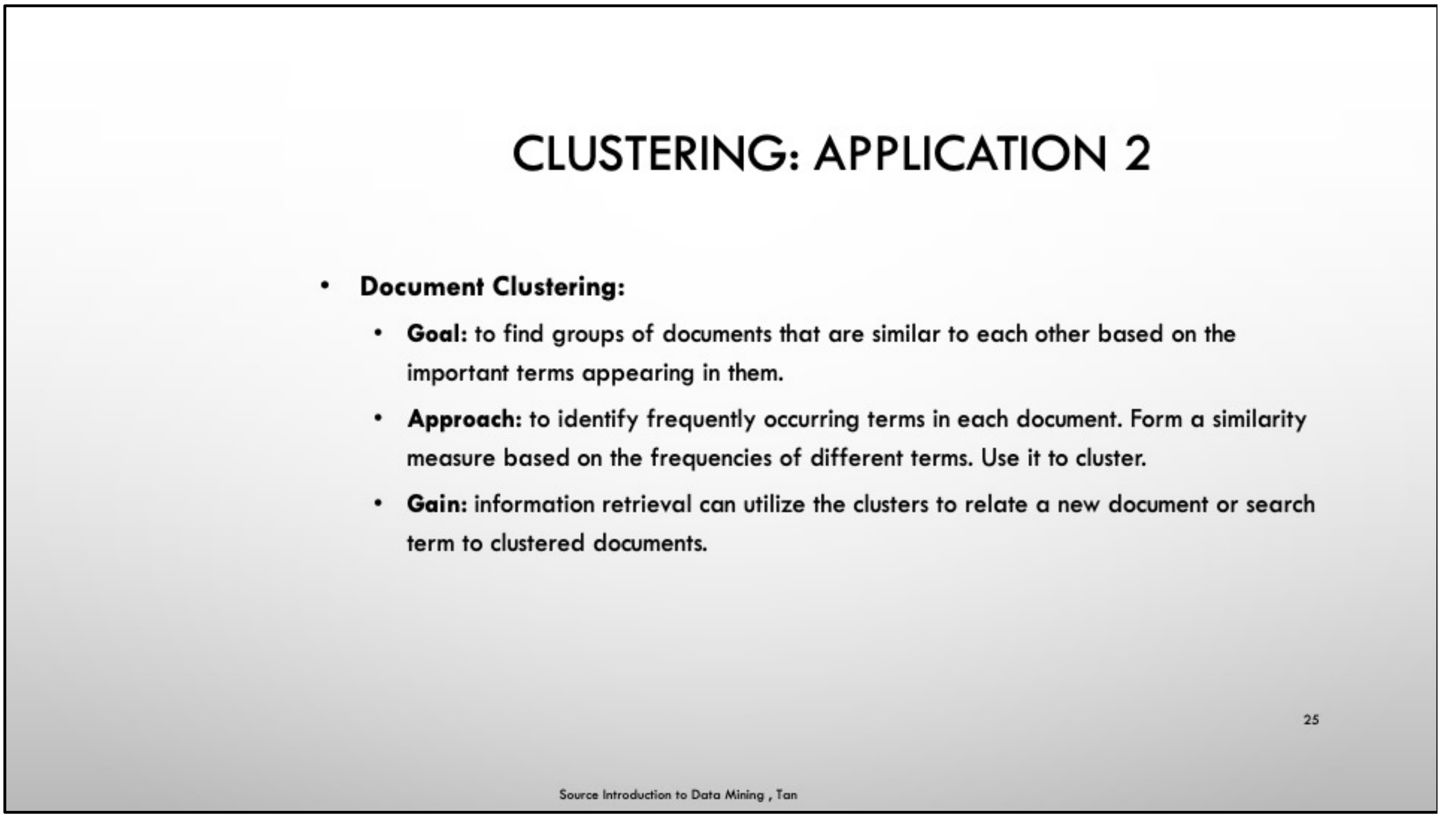
23

Source Introduction to Data Mining , Tan

CLUSTERING: APPLICATION 1

- **Market Segmentation:**
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. Those from different clusters.

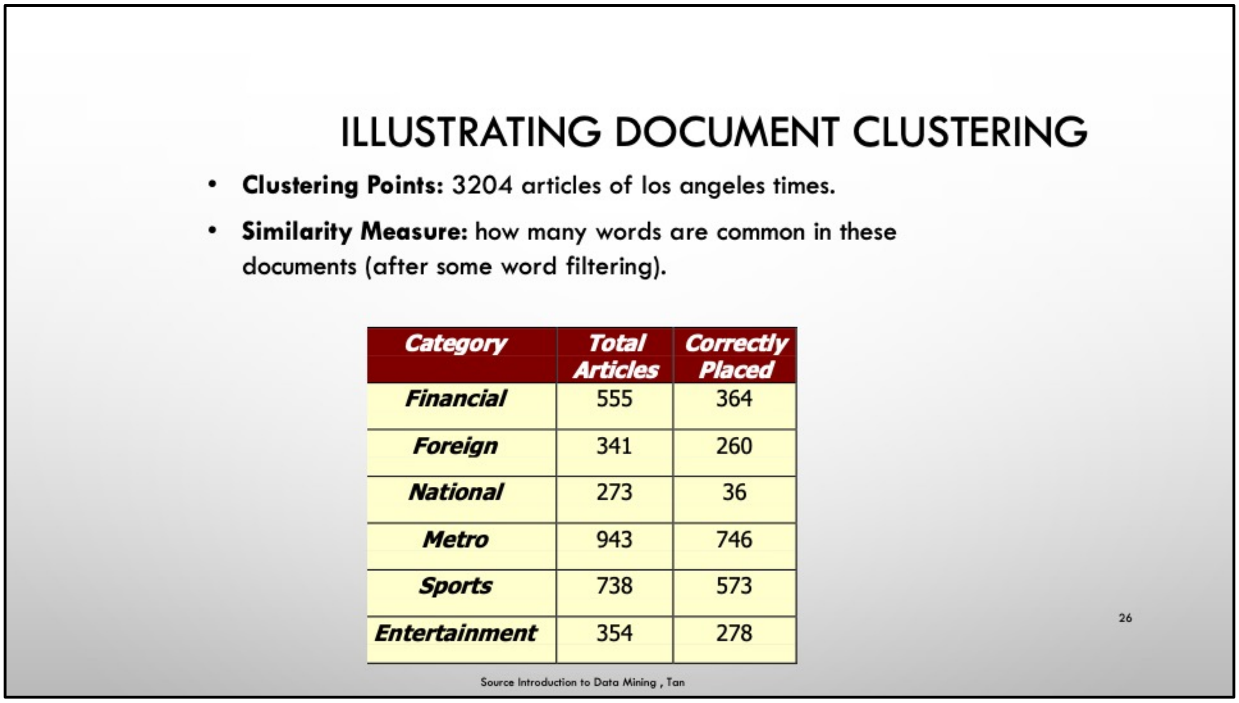
Source Introduction to Data Mining , Tan



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.
  - **Gain:** information retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Source Introduction to Data Mining , Tan



ILLUSTRATING DOCUMENT CLUSTERING

- **Clustering Points:** 3204 articles of los angeles times.
- **Similarity Measure:** how many words are common in these documents (after some word filtering).

Category	***Total Articles***	***Correctly Placed***
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Source Introduction to Data Mining , Tan