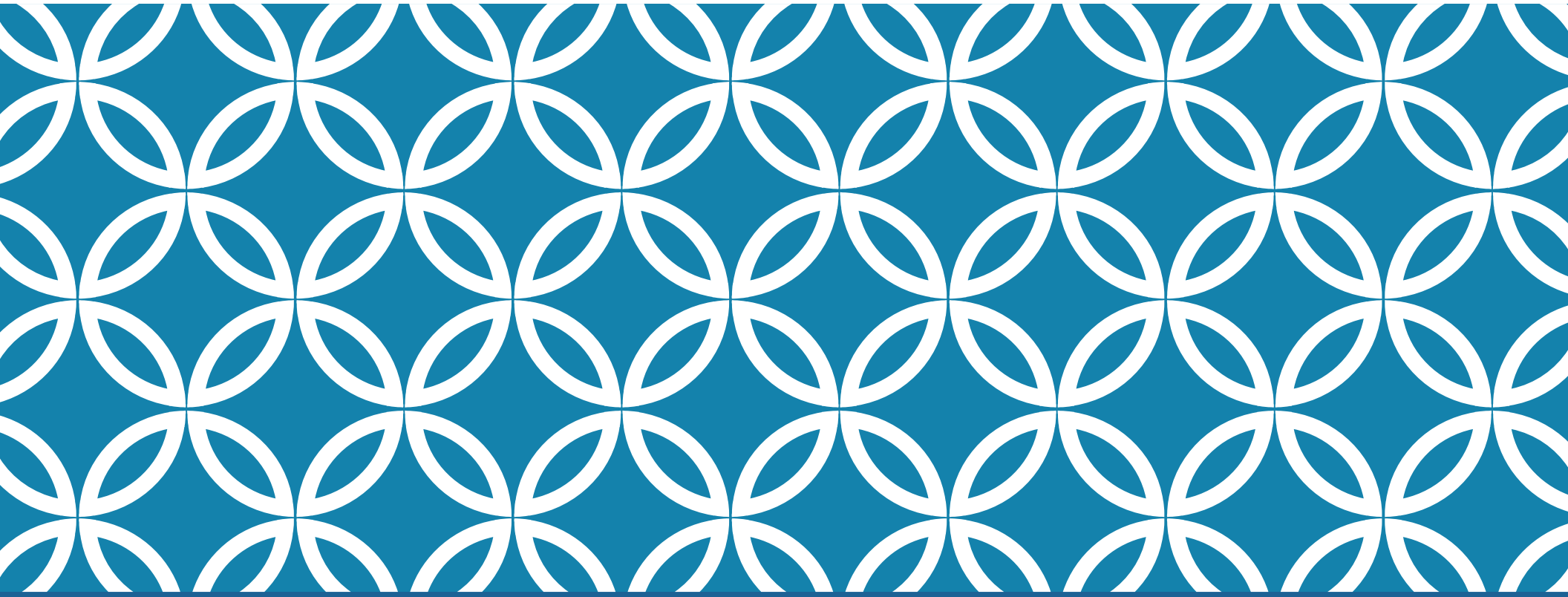


LETS RECAPITULATE!!!

DATA WAREHOUSING AND DATA MINING

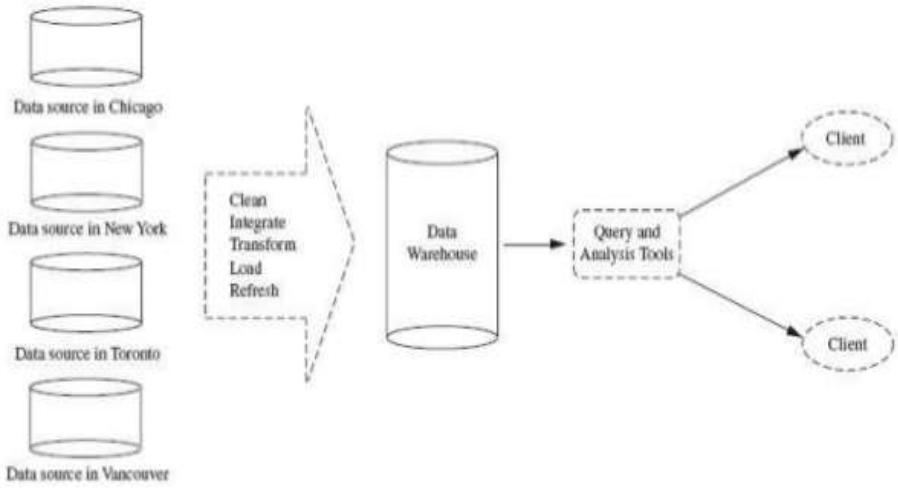
What Is Data Mining??

- Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.

- Alternative names:
  - Knowledge discovery(mining) in databases (KDD),
  - knowledge extraction,
  - data/pattern analysis,
  - data archeology,
  - business intelligence, etc.

Data Warehouses

- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



Data Warehouses (2)

- Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization

A typical DM System Architecture (2)

On What Kinds of Data?

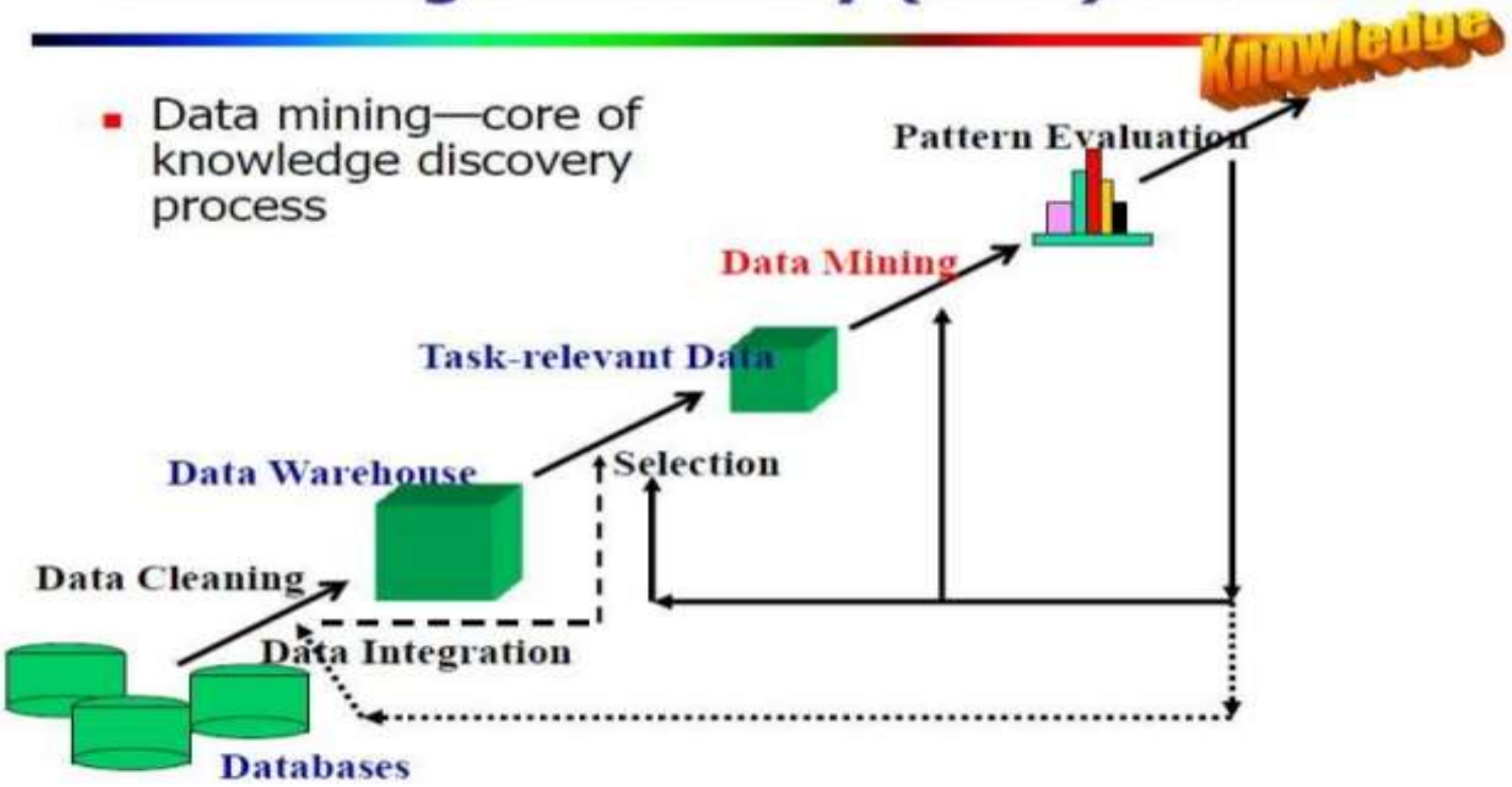
- **Database-oriented data sets and applications**
  - Relational database, data warehouse, transactional database
- **Advanced data sets and advanced applications**
  - Object-Relational Databases
  - Temporal Databases
  - Spatial Databases and Spatiotemporal Databases
  - Text databases and Multimedia databases
  - Heterogeneous Databases
  - Data Streams
  - The World-Wide Web etc.

Applications of Data Mining

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process

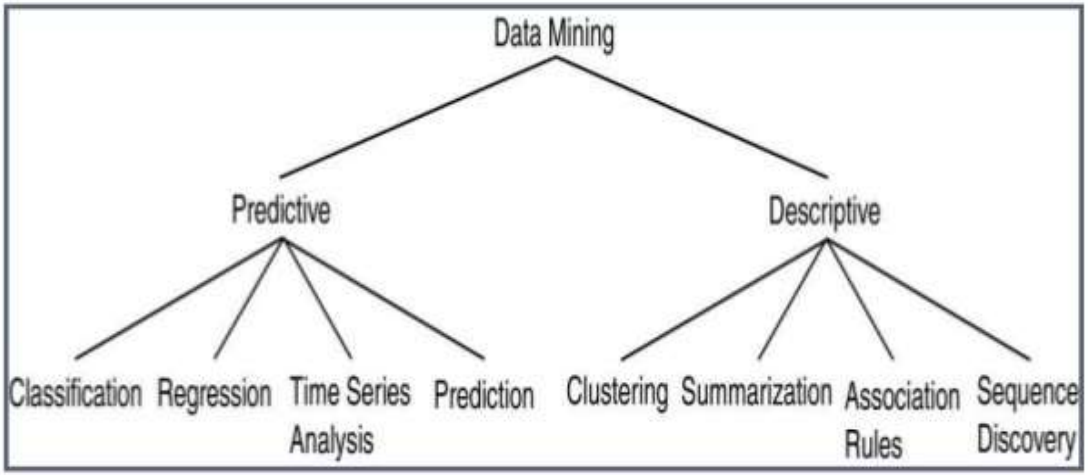


Data mining functionalities-

1. **Descriptive Data mining Tasks:** Describe general properties of existing data.

2. **Predictive Data mining Tasks:** Attempt to do predictions based on inference on available data.

Data Mining Functionalities/Tasks



Classification

Classification is a process of predicting class label for unseen new data based on the data tuples with known class labels

Examples:

- Predict whether a new customer buy a Computer in the store ?
- Predict the loan applicant status as safe or risky

Classification(Cont'd)

Classification is a 2 step process

Step-1: **Construction or training of a model/ classifier** using training data tuples with known class labels

Step-2: **Testing the accuracy of a classifier** using the test data tuples for which the class label is already known

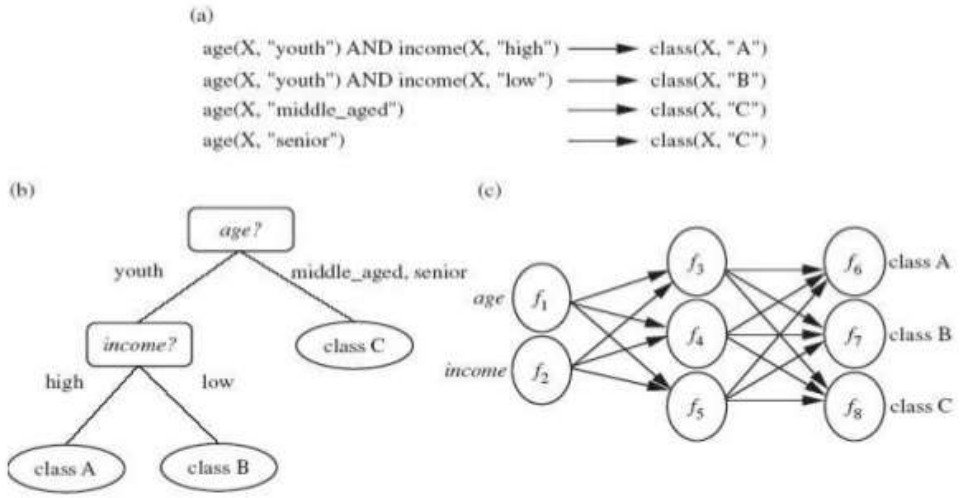
Classification Algorithms

- Decision tree-Induction(Tree like flowchat)
- Back propagation (Neural Network)
- Bayesian Classification (statistical probability)
- Rule-based classification (If..Else rules)

Classification is called as **supervised-Learning**

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?



An Example

(from *Pattern Classification by* Duda & Hart & Stork – Second Edition, 2001)

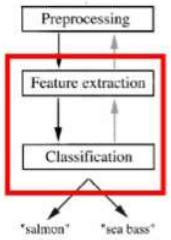
- A fish-packing plant wants to automate the process of sorting incoming fish according to species
- As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

An Example (continued)

Features (to distinguish):

Length
Lightness
Width
Position of mouth

An Example (continued)



- **Preprocessing:** Images of different fishes are isolated from one another and from background;
- **Feature extraction:** The information of a single fish is then sent to a feature extractor, that measure certain "features" or "properties";
- **Classification:** The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

An Example (continued)

- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature: (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier *learns* itself from the training examples, how to distinguish Salmon from Bass based on the *fish_length*

Classification Example 2

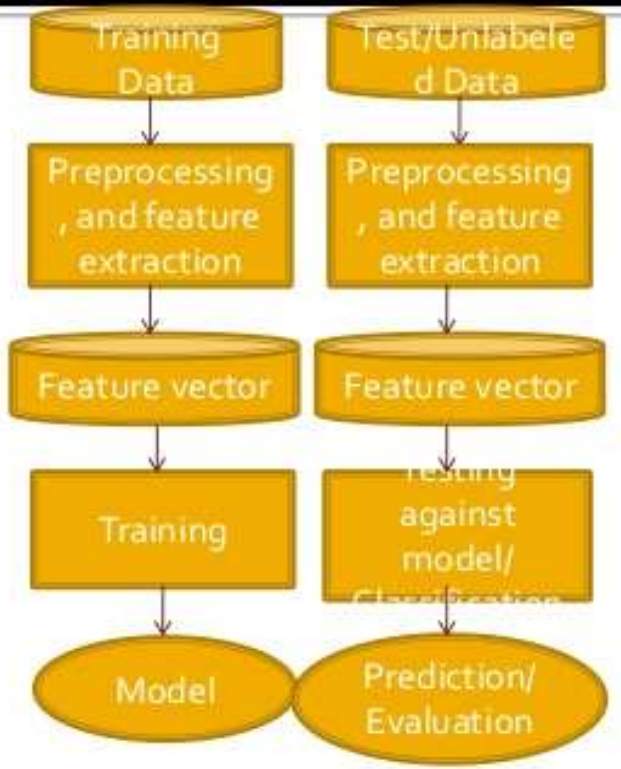
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

categorical categorical continuous class

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

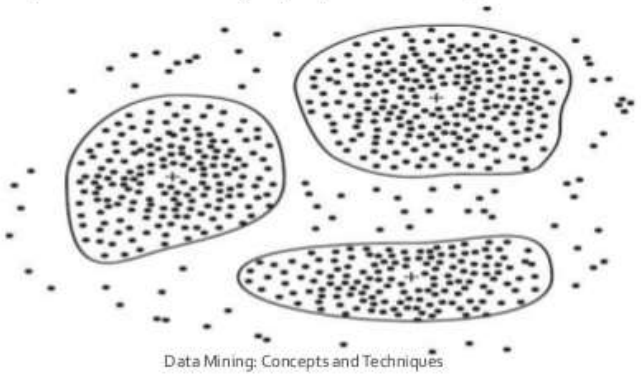
An Example (continued)

- So the overall classification process goes like this →



Data Mining Functionalities (2)

- Cluster Analysis
  - Class label is unknown: group data to form new classes
  - Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*
    - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



Data Mining Functionalities (2)

- Outlier Analysis
  - Data that do no comply with the general behavior or model.
  - Outliers are usually discarded as noise or exceptions.
  - Useful for fraud detection.
    - E.g. Detect purchases of extremely large amounts
- Evolution Analysis
  - Describes and models regularities or trends for objects whose behavior changes over time.
    - E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

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.

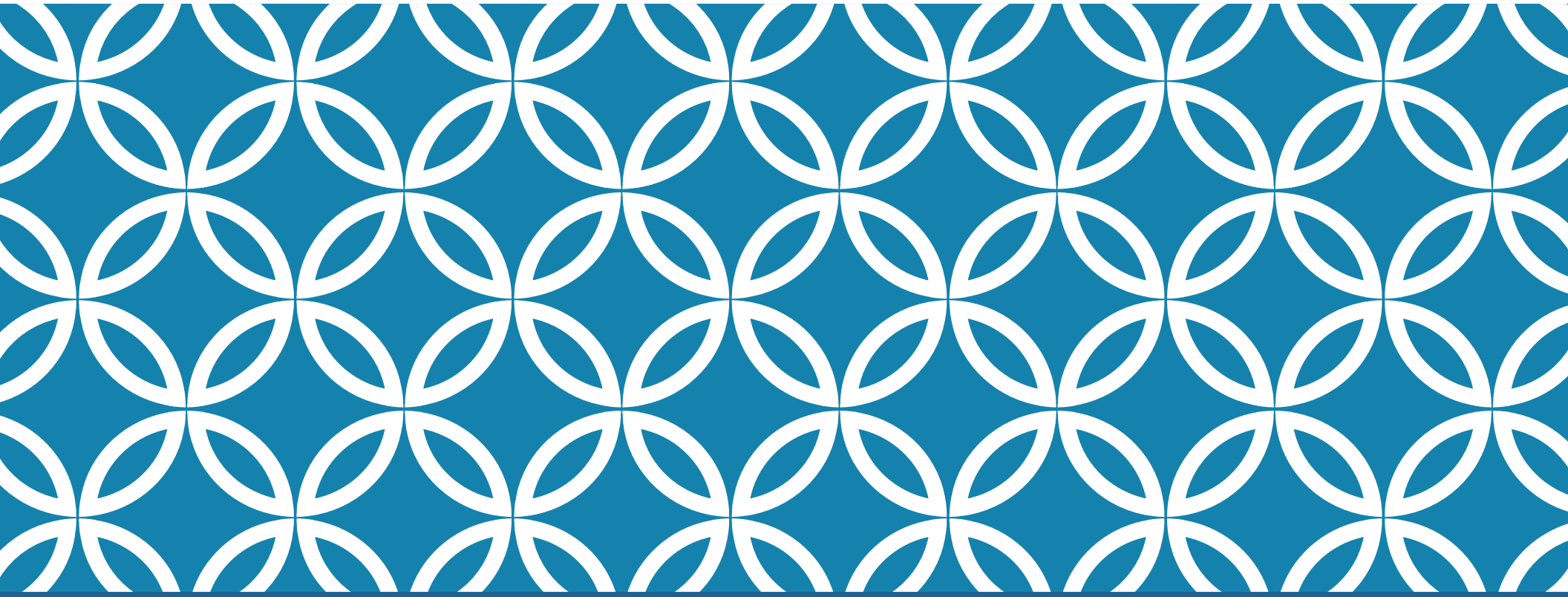
Association Rule Discovery: Definition

- 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 Rules
  - Implication: $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$;
  - Support of AR (s) $X \rightarrow Y$:
    - Percentage of transactions that contain
- $X \cup Y$
  - Probability that a transaction contains $X \cup Y$.
-
  - Confidence of AR (a) $X \rightarrow Y$:
    - Ratio of number of transactions that contain $X \cup Y$ to the number that contain X
    - Conditional probability that a transaction having X also contains *Y*.



THANKS !!!