DATA MINING

Data Mining

- The extraction of useful information from data
- The automated extraction of hidden predictive information from (large) databases
- Business, huge data bases, customer data, mine the data
  - Also Medical, Genetic, Astronomy, etc.
- Data often unlabeled – unsupervised clustering, etc.
- Focuses on learning approaches which scale to massive amounts of data
  - and potentially to a large number of features
  - sometimes requires simpler algorithms with lower big-O complexities (and which are more intelligible)

Data Mining Applications

- Often seeks to give businesses a competitive advantage

- Which customers should they target
  - For advertising – more focused campaign
  - Customers they most/least want to keep
  - Most favorable business decisions

- Associations
  - Which products should/should not be on the same shelf
  - Which products should be advertised together
  - Which products should be bundled

- Information Brokers
  - Make transaction information available to others who are seeking advantages

Data Mining

- Basically, a particular niche of machine learning applications
  - Focused on business and other large data problems
  - Focused on problems with huge amounts of data which needs to be manipulated in order to make effective inferences
  - "Mine" for "gems" of actionable information

Data Mining Popularity

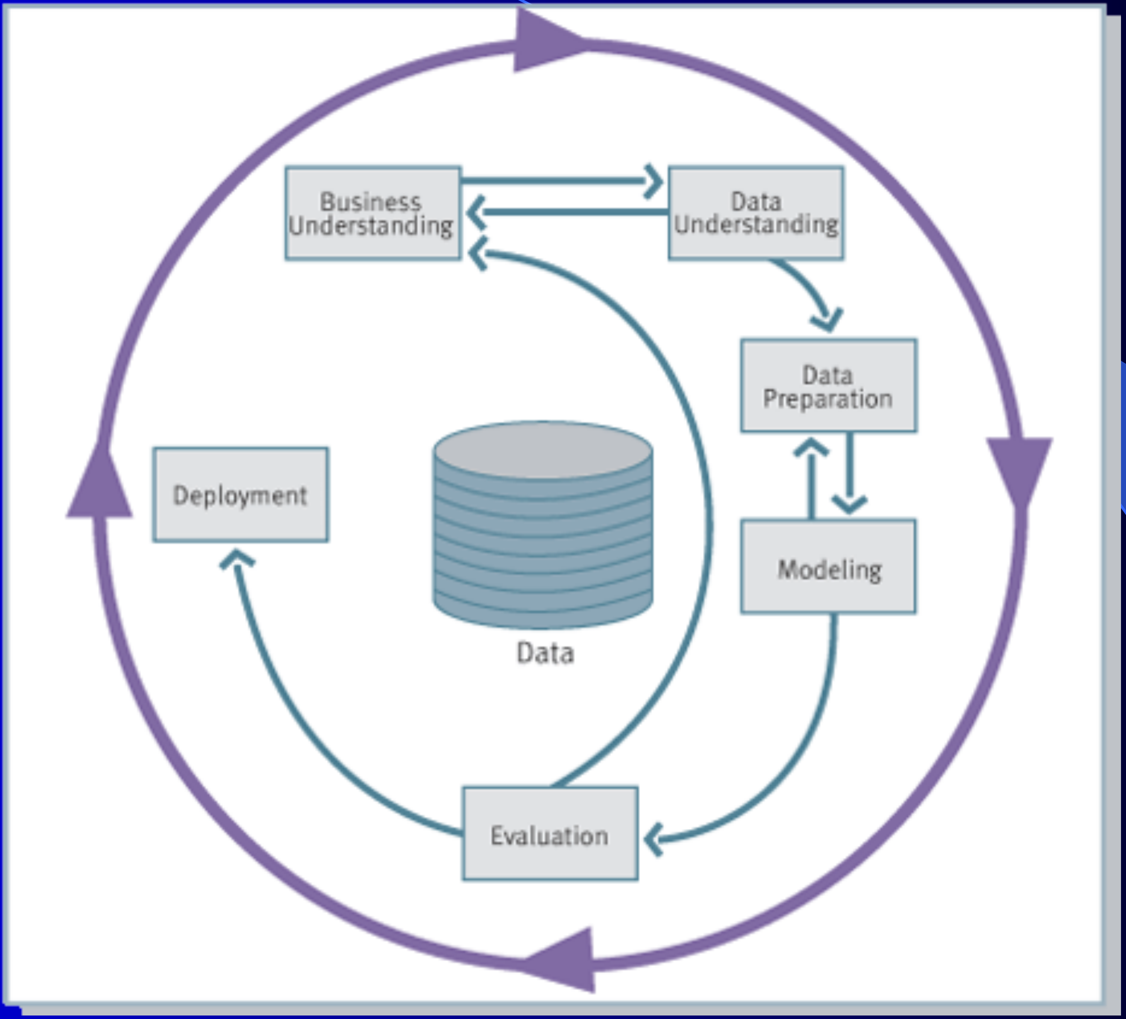
- Recent Data Mining explosion based on:
- Data available – Transactions recorded in data warehouses
  - From these warehouses specific databases for the goal task can be created
- Algorithms available – Machine Learning and Statistics
  - Including special purpose Data Mining software products to make it easier for people to work through the entire data mining cycle
- Computing power available
- Competitiveness of modern business – need an edge

Data Mining Process Model

- You will use much of this process in your group project

1. Identify and define the task (e.g. business problem)
2. Gather and Prepare the Data
   - Build Data Base for the task
   - Select/Transform/Derive features
   - Analyze and Clean the Data, remove outliers, etc.
3. Build and Evaluate the Model(s) – Using training and test data
4. Deploy the Model(s) and Evaluate business related Results
   - Data visualization tools
5. Iterate through this process to gain continual improvements both initially and during life of task
   - Improve/adjust features and/or machine learning approach

Data Mining Process Model - Cycle



Monitor, Evaluate, and update deployment

Data Science and Big Data

- Interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data
  - Machine Learning
  - Statistics/Math
  - CS/Database/Algorithms
  - Visualization
  - Parallel Processing
  - Etc.

- Increasing demand in industry!

- New DS emphasis in BYU CS began Fall 2019

Data Warehouses

- Companies have large data warehouses of transactions
  - Records of sales at a store
  - On-line shopping
  - Credit card usage
  - Phone calls made and received
  - Visits and navigation of web sites, etc…
- Many/Most things recorded these days and there is potential information that can be mined to gain business improvements
  - For better customer service/support and/or profits

Association Analysis – Link Analysis

- Used to discover relationships in large databases
- Relationships represented as *association rules*
  - Unsupervised learning, any data set
- One example is *market basket analysis* which seeks to understand more about what items are bought together
  - This can then lead to improved approaches for advertising, product placement, etc.
  - Example Association Rule: {Cereal} ⇒ {Milk}

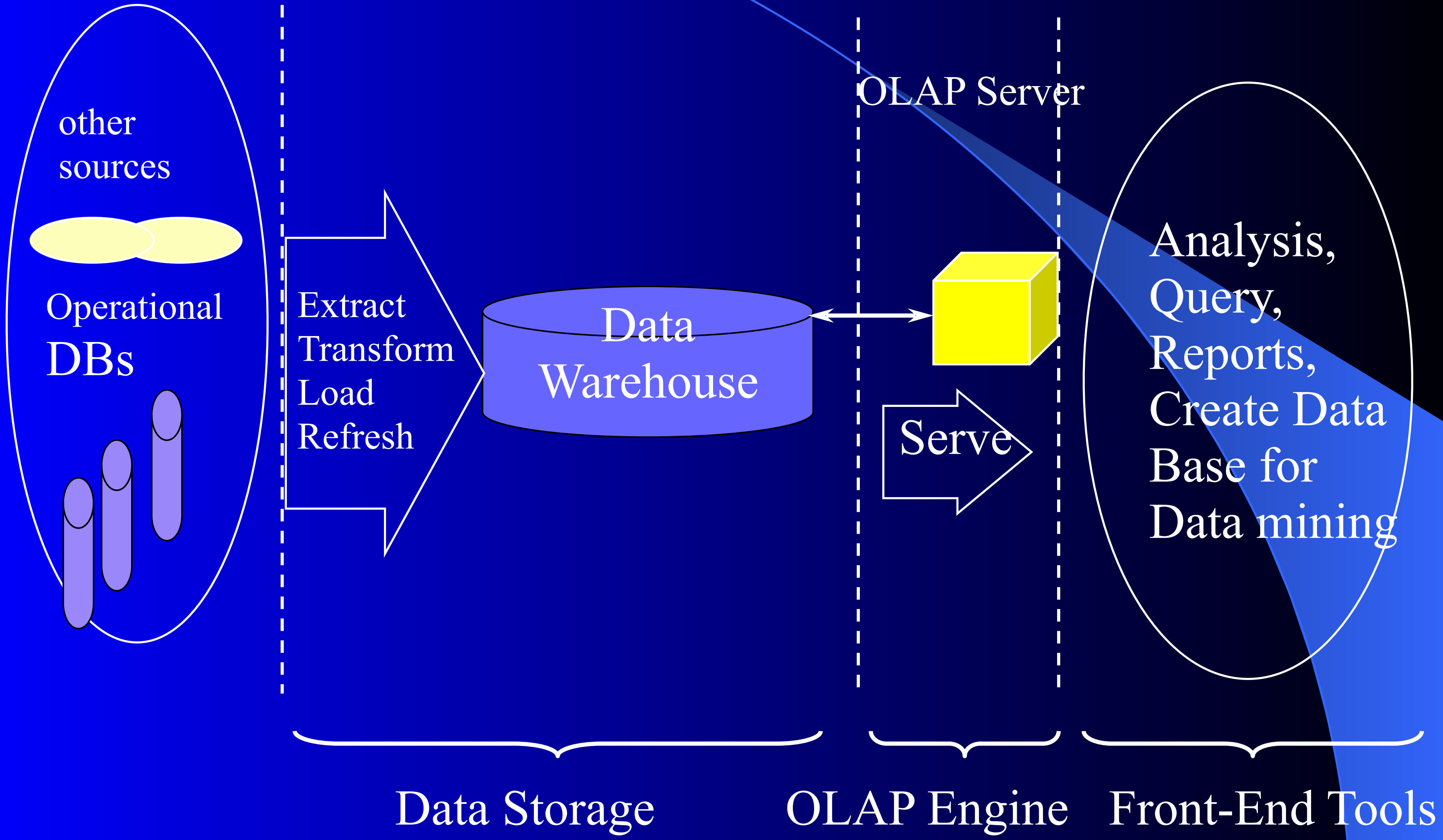
Transaction ID and Info	**Items Bought**
1 who, when, etc.	{Ice cream, milk, eggs, cereal}
2	{Ice cream}
3	{milk, cereal, sugar}
4	{eggs, yogurt, sugar}
5	{Ice cream, milk, cereal}

Association Discovery

- Association rules are not causal, show correlations
- *k*-item set is a subset of the possible items – {Milk, Eggs} is a 2-item set
- Which item sets does transaction 3 contain
- Association Analysis/Discovery seeks to find frequent item sets

TID	Items Bought
1	{Ice cream, milk, eggs, cereal}
2	{Ice cream}
3	{milk, cereal, sugar}
4	{eggs, yogurt, sugar}
5	{Ice cream, milk, cereal}

The Big Picture: DBs, DWH, OLAP & DM



Summary

- Association Analysis useful in many real world tasks
  - Not a classification approach, but a way to understand relationships in data and use this knowledge to advantage
- Also standard classification and other approaches
- Data Mining continues to grow as a field
  - Data and features issues
    - Gathering, selection and transformation, preparation, cleaning, storing
  - Data visualization and understanding
  - Outlier detection and handling
  - Time series prediction
  - Web mining
  - etc.

Group Projects

- Review timing and expectations
  - Proposal - due Friday May 30th
  - Project
    - Gathering, cleaning, transforming the data can be the most critical part of the project, so get that going ASAP!
    - Plenty of time to try some different ML models and some iterations on your features and/or ML models to get improvements.
  - Final report

- Questions?