

# ISA 401: Business Intelligence & Data Visualization

## 23: A Short Introduction to Exploratory Data Mining

Fadel M. Megahed, PhD

Associate Professor  
Department of Information Systems and Analytics  
Farmer School of Business  
Miami University

Twitter: [FadelMegahed](#)

GitHub: [fmegahed](#)

Email: [fmegahed@miamioh.edu](mailto:fmegahed@miamioh.edu)

Office Hours: [Automated Scheduler for Virtual Office Hours](#)



Spring 2022

# A Recap of What we Learned Last Week


- Define a “business report” & its main functions
- Understand the importance of the right KPIs
- Automate traditional business reports
- Dashboards as real-time business reporting tools

# Course Objectives Covered so Far

[Y]ou will be re-introduced to **how data should be explored** ... Instead, the focus is on understanding the underlying methodology and mindset of **how data should be approached, handled, explored, and incorporated back into the domain of interest.** ... You are expected to:

- ✓ Be capable of extracting, transforming and loading (ETL) data using multiple platforms (e.g.  & Tableau).
- ✓ Write basic  scripts to preprocess and clean the data.
- ✓ Explore the data using visualization approaches that are based on sound human factors (i.e. account for human cognition and perception of data).
- ✗ Understand how data mining and other analytical tools can capitalize on the insights generated from the data viz process.
- ✓ Create interactive dashboards that can be used for business decision making, reporting and/or performance management.
- ✗ Be able to apply the skills from this class in your future career.

# Learning Objectives for Today's Class

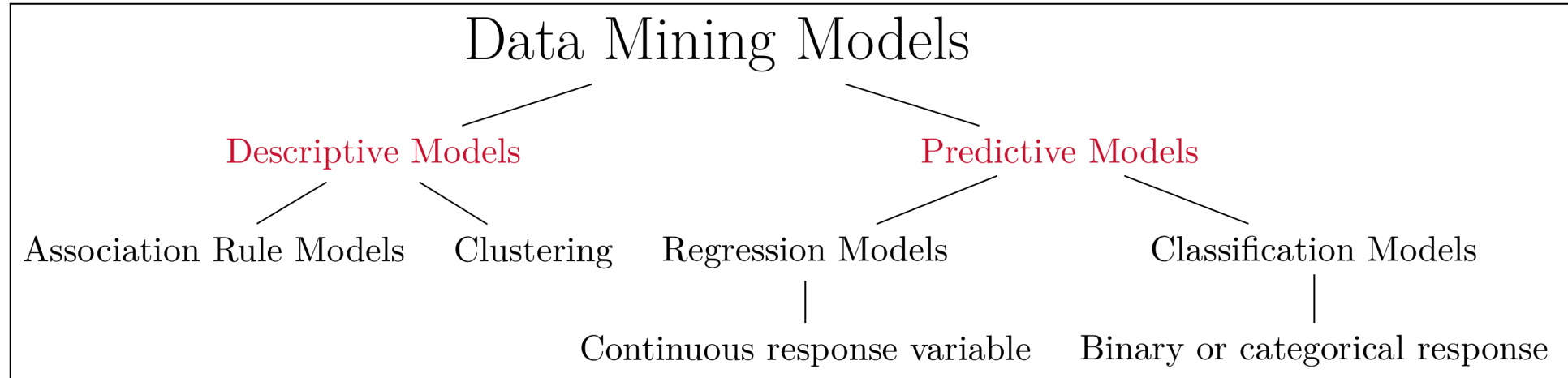
- Describe the goals & functions of data mining
- Understand the statistical limits on data mining
- Describe the data mining process
- What is “frequent itemsets” & the application of this concept
- Explain how and why “association rules” are constructed
- Use  to populate both concepts

# An Overview of Data Mining

# What is Data Mining?

- The most common definition of data mining is the discovery of models from data.
- Discovery of **patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
- Subsidiary Issues:
  - **Data cleansing:** detection of bogus data
  - **Data visualization:** something better than MBs of output
  - **Warehousing** of data (for retrieval)

# A Simplistic View of Data Mining Models



A simplistic summary of data mining models. Note that, in ISA 401, we will only briefly cover descriptive/exploratory data mining models

# Data Mining is Hard

Data mining is hard since it has the following issues:

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation

**Note that I have intentionally not included fitting/training a model since this is relatively easy if you understand the data, engineered/captured the important predictors, and have the data in the "correct" shape/quality.**



# Association Rules

Data	Top 5 Rules	Scatter Plot of all Rules	Graph-based Plot of Top 5 Rules
------	-------------	---------------------------	---------------------------------

```
## transactions as itemMatrix in sparse format with
## 9835 rows (elements/itemsets/transactions) and
## 169 columns (items) and a density of 0.02609146
##
## most frequent items:
##      whole milk other vegetables      rolls/buns      soda
##      2513      1903      1809      1715
##      yogurt      (Other)
##      1372      34055
##
## element (itemset/transaction) length distribution:
## sizes
##      1      2      3      4      5      6      7      8      9      10      11      12      13      14      15      16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117  78  77  55  46
##      17      18      19      20      21      22      23      24      26      27      28      29      32
##      29      14      14      9      11      4      6      1      1      1      1      3      1
##
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      1.000  2.000  3.000  4.409  6.000 32.000
##
```

# Association Rules

Data	Top 5 Rules	Scatter Plot of all Rules	Graph-based Plot of Top 5 Rules
------	-------------	---------------------------	---------------------------------

```
##      lhs                                rhs      support
## [1] {Instant food products, soda}      => {hamburger meat} 0.001220132
## [2] {soda, popcorn}                    => {salty snack}    0.001220132
## [3] {flour, baking powder}             => {sugar}          0.001016777
## [4] {ham, processed cheese}            => {white bread}    0.001931876
## [5] {whole milk, Instant food products} => {hamburger meat} 0.001525165
##      confidence coverage    lift    count
## [1] 0.6315789  0.001931876 18.99565 12
## [2] 0.6315789  0.001931876 16.69779 12
## [3] 0.5555556  0.001830198 16.40807 10
## [4] 0.6333333  0.003050330 15.04549 19
## [5] 0.5000000  0.003050330 15.03823 15
```

# Association Rules

Data

Top 5 Rules

Scatter Plot of all Rules

Graph-based Plot of Top 5 Rules

# Association Rules

Data

Top 5 Rules

Scatter Plot of all Rules

Graph-based Plot of Top 5 Rules

04:00

# Clustering of Traffic Volume on I-85

Data

Calendar Plot of Clustered Data

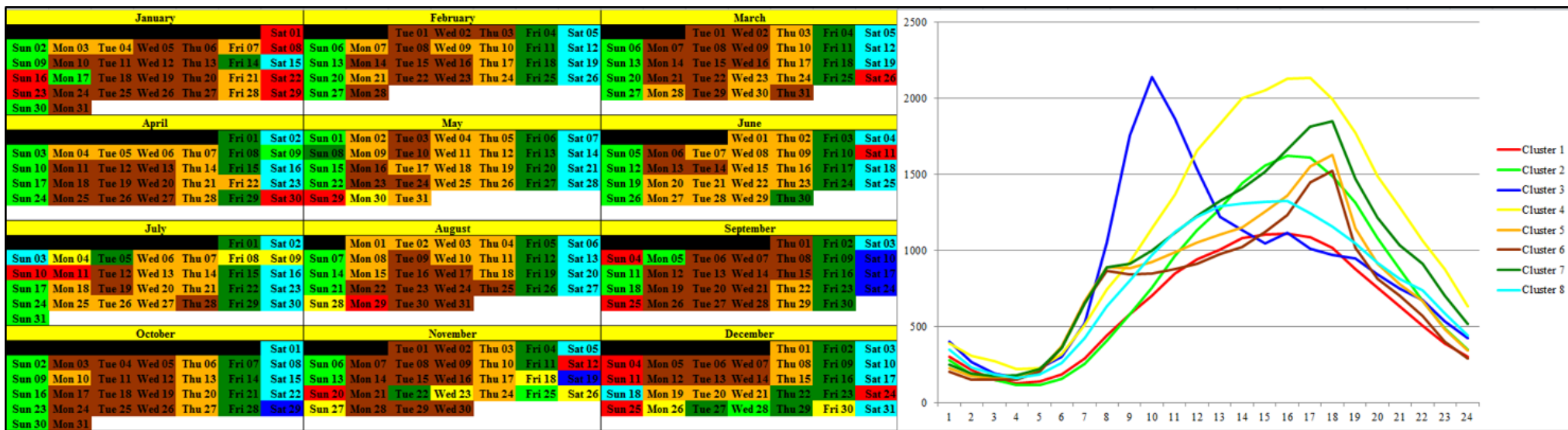
Insights from Chart?

# Clustering of Traffic Volume on I-85

Data

Calendar Plot of Clustered Data

Insights from Chart?



# Clustering of Traffic Volume on I-85

Data

Calendar Plot of Clustered Data

Insights from Chart?

**Based on the previous tab, what are 2-3 main insights you have learned about the traffic volume in Montgomery, AL?** Write them down below

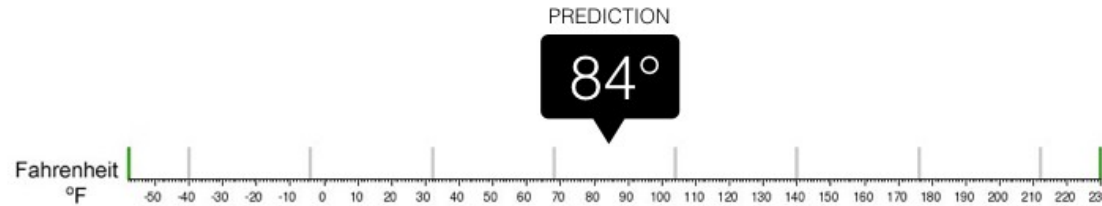
Edit me and insert your solution here

# Regression vs Classification



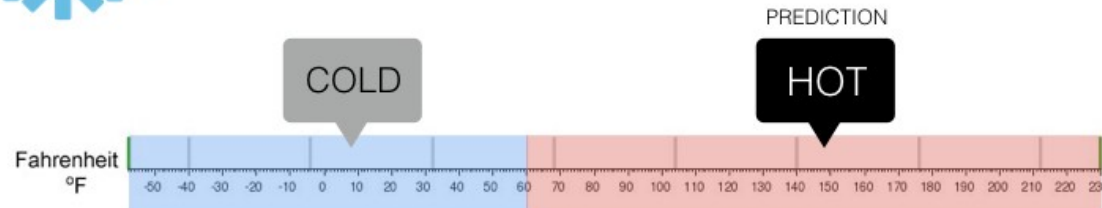
## Regression

What is the temperature going to be tomorrow?



## Classification

Will it be Cold or Hot tomorrow?



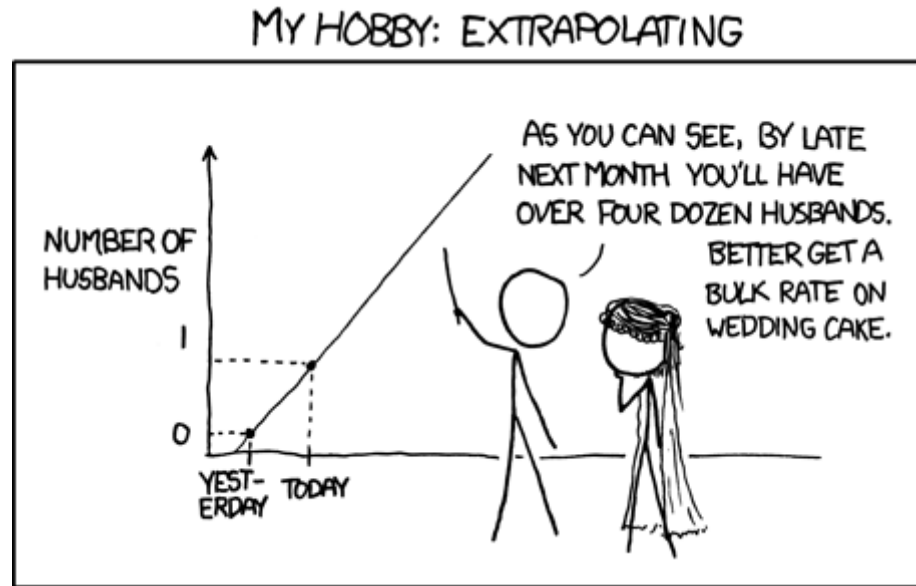


# An Overview of Common Data Mining Models

# Limits on Data Mining

# Meaningfulness of Answers from DM Models

- A big risk when data mining is that you will discover patterns that are meaningless.
- **Bonferroni's Principle:** (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find.



# Rhines Paradox: An Example of Overzealous DM?

- Joseph Rhine was a parapsychologist in the 1950s who hypothesized that some people had **Extra-Sensory Perception**.
- He devised an experiment where subjects were asked to guess 10 hidden cards **red** or **blue**.
- He discovered that almost 1 in 1000 had ESP they were able to get all 10 right!
- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- **What did he conclude?**
  - He concluded that you should not tell people they have ESP; it causes them to lose it.
  - **Why is this an incorrect conclusion?**

# Ethical Issues with Data Mining

# In the News: AI Implementation Scandals

FROM POLITICO

## Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.



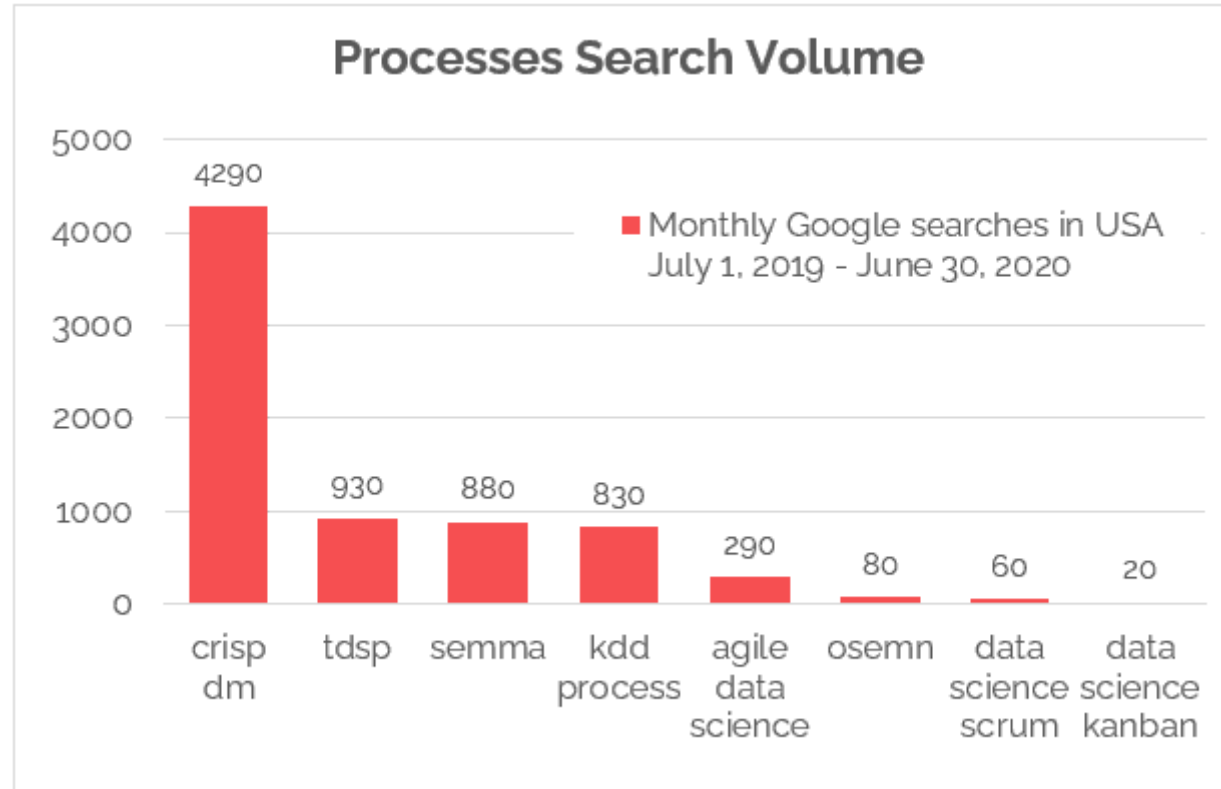
As the world turns to AI to automate their systems, the Dutch scandal shows how devastating they can be | Dean Mouhtaropoulos/Getty Images

BY MELISSA HEIKKILÄ

March 29, 2022 | 6:14 pm

# The Data Mining Process

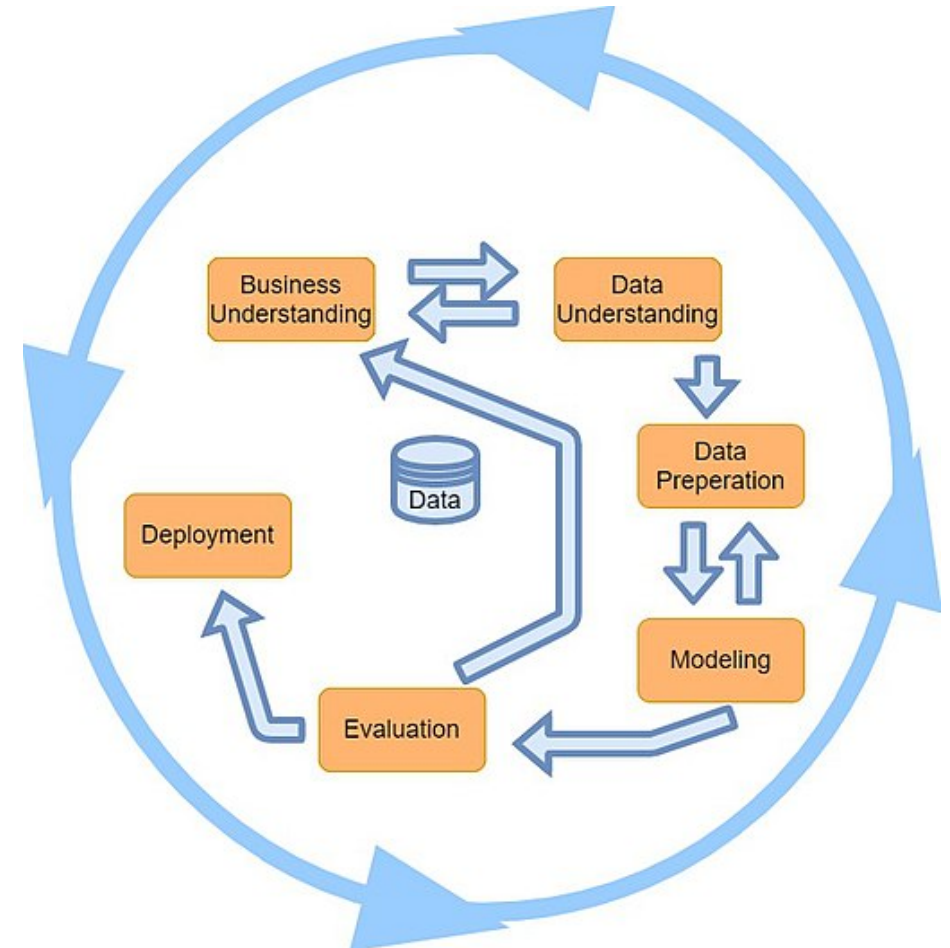
# Frameworks for Data Mining Projects





# The CRISP-DM Process

- You are expected to read the **original CRISP-DM paper**
- Each step has several substeps
- **Most of the project time is typically spent in steps 1-3**



# Frequent Itemsets, Market Basket Analysis and Association Rule Mining

# Association Rule Discovery

## Supermarket shelf management – Market-basket model:

- **Goal:** Identify items that are bought together by sufficiently many customers
- **Approach:** Process the sales data collected with barcode scanners to find dependencies among items
- **A classic rule:**
  - If someone buys diaper and milk, then he/she is likely to buy beer
  - Don't be surprised if you find six-packs next to diapers!

# The Market-Basket Model

- A large set of **items**
  - e.g., things sold in a supermarket
- A large set of **baskets**
- Each basket is a **small subset of items**
  - e.g., the things one customer buys on one day
- Want to discover **association rules**
  - People who bought {x,y,z} tend to buy {v,w}
  - Amazon!

## Input:

Basket #	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

## Output: Discovered Rules

{Milk} --> {Coke}  
{Diaper, Milk} --> {Beer}

# Definitions: Support & Support Threshold

- **Simplest question:** Find sets of items that appear together “frequently” in baskets
- **Support for itemset  $I$ :** Number of baskets containing all items in  $I$ 
  - Often expressed as a fraction of the total number of baskets
- Given a **support threshold  $s$** , then sets of items that appear in at least  $s$  baskets are called frequent itemsets

**Input:**

Basket #	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

**Support of {Beer, Bread}: = 2**

# Non-graded Activity: Frequent Itemsets

Activity	Your Solution
----------	---------------

**Items** = {Milk, Coke, Pepsi, Beer, Juice}

**With a support threshold of 3 baskets, find all frequent itemsets based on these 8 baskets:**

- $B_1 = \{\text{Milk, Coke, Beer}\}$
- $B_2 = \{\text{Milk, Pepsi, Juice}\}$
- $B_3 = \{\text{Milk, Beer}\}$
- $B_4 = \{\text{Coke, Juice}\}$
- $B_5 = \{\text{Milk, Pepsi, Beer}\}$
- $B_6 = \{\text{Milk, Coke, Beer, Juice}\}$
- $B_7 = \{\text{Coke, Beer, Juice}\}$
- $B_8 = \{\text{Coke, Beer}\}$

04:00

# Non-graded Activity: Frequent Itemsets

Activity

Your Solution

**Identify all frequent singletons, doubles, triples, etc.**

Edit me and insert your solution here

# Association Rules

- **Association Rules:** If-then rules about the contents of baskets
- $\{i_1, i_2, \dots, i_k\} \rightarrow j$  means: "if a basket contains all of  $i_1, \dots, i_k$  then it is likely to contain  $j$ "
- **In practice there are many rules, want to find significant/interesting ones!**
- **Confidence** of this association rule is the probability of  $j$  given  $I = \{i_1, \dots, i_k\}$

$$conf(I \rightarrow j) = P(j \mid I) = \frac{support(I \cap j)}{support(I)}$$

- **Not all high-confidence rules are interesting**
  - The rule  $\mathbf{X} \rightarrow \mathbf{milk}$  may have high confidence for many itemsets  $\mathbf{X}$ , because **milk** is just purchased very often (independent of  $\mathbf{X}$ ) and the confidence will be high
- **Lift** of an association rule  $I \rightarrow J$  is the ratio between its confidence and the fraction of baskets containing  $j$ :  
$$lift(I \rightarrow j) = \frac{conf(I \rightarrow j)}{Pr(j)}$$



# Non-Graded Activity: Confidence and Lift

Activity	Your Solution
<ul style="list-style-type: none"><li>• <math>B_1 = \{\text{Milk, Coke, Beer}\}</math></li><li>• <math>B_3 = \{\text{Milk, Beer}\}</math></li><li>• <math>B_5 = \{\text{Milk, Pepsi, Beer}\}</math></li><li>• <math>B_6 = \{\text{Coke, Beer, Juice}\}</math></li></ul>	<ul style="list-style-type: none"><li><math>B_2 = \{\text{Milk, Pepsi, Juice}\}</math></li><li><math>B_4 = \{\text{Coke, Juice}\}</math></li><li><math>B_6 = \{\text{Milk, Coke, Beer, Juice}\}</math></li><li><math>B_8 = \{\text{Coke, Beer}\}</math></li></ul>

**For the association rule:**  $\{\text{Milk, Beer}\} \rightarrow \text{Coke}$ , compute both its confidence and lift.

# Non-Graded Activity: Confidence and Lift

Activity

Your Solution

**Computing the confidence and lift for the association rule** {Milk, Beer} → Coke

Edit me and insert your solution here

# Finding Association Rules

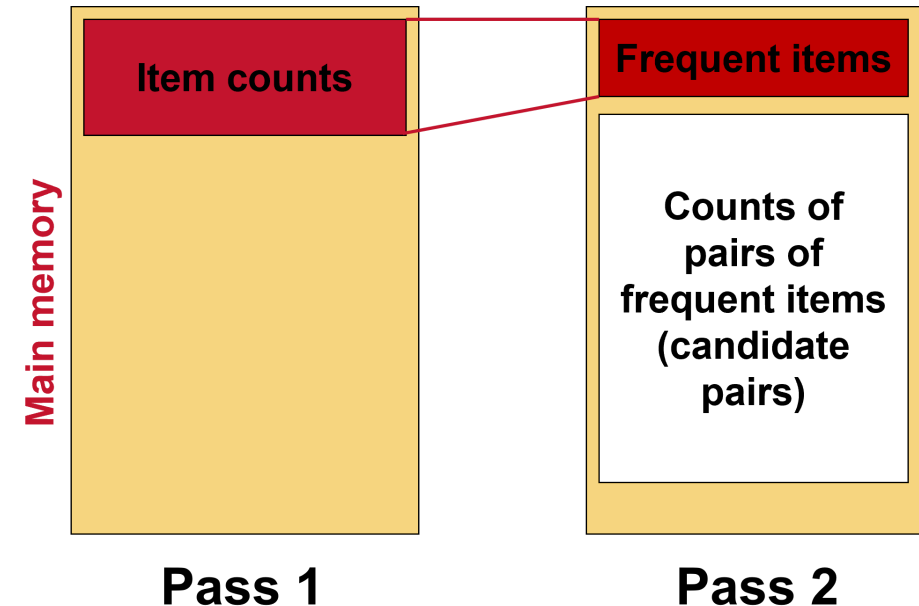
- **Problem:** Find all association rules with support  $\geq s$  and confidence  $\geq c$ 
  - **Note:** Support of an association rule is the support of the set of items on the left side
- **Hard part:** Finding the frequent itemsets!
  - If  $\{i_1, i_2, \dots, i_k\} \rightarrow j$  has high support and confidence, then:
  - both  $\{i_1, i_2, \dots, i_k\}$  and both  $\{i_1, i_2, \dots, i_k, j\}$  will be “frequent”

# Naïve Approach to Counting Frequent Itemsets

- Naïve approach to finding frequent pairs
- **Read file once, counting in main memory the occurrences of each pair:**
  - From each basket of  $n$  items, generate its  $\frac{n(n-1)}{2}$  pairs by two nested loops
- Fails if  $(\text{\#items})^2$  exceeds main memory
  - Remember: #items can be 100K (Wal-Mart) or 10B (Web pages)
  - Suppose  $10^5$  items, counts are 4-byte integers
  - Number of pairs of items:  $\frac{10^5(10^5-1)}{2} = 5 * 10^9$
  - Therefore,  $2 * 10^{10}$  (20 gigabytes) of memory needed

# A-Priori Algorithm

- **Pass 1:** Read baskets and count in main memory the occurrences of each **individual item**
  - Requires only memory proportional to #items
- **Items that appear  $\geq s$  times are the frequent items**
- **Pass 2:** Read baskets again and count in main memory **only those pairs where both elements are frequent (from Pass 1)**



# Using to Mine Association Rules

In class, we will go through this R code, explaining: (a) what each function is doing, and (b) the outputs from each step.

```
if(require(pacman)==FALSE) install.packages('pacman')
pacman::p_load(arules, tidyverse)

data('Groceries') # note its class

summary(Groceries)

itemFrequency(Groceries) # returns frequency in alphabetic order
itemFrequency(Groceries) %>% sort(decreasing = T)

itemFrequencyPlot(Groceries, support = 0.1)
itemFrequencyPlot(Groceries, topN = 20)


# mine association rules with a certain min support and confidence
grocery_rules = apriori(
  Groceries, parameter = list(
    support = 0.01, confidence = 0.5, minlen = 2, maxlen = 5) )

summary(grocery_rules)
inspect(grocery_rules)

sort(grocery_rules, by = 'lift', decreasing = T)[1:3] %>% inspect()
```

# Recap

# Summary of Main Points

- Describe the goals & functions of data mining
- Understand the statistical limits on data mining
- Describe the data mining process
- What is “frequent itemsets” & the application of this concept
- Explain how and why “association rules” are constructed
- Use  to populate both concepts