MMA/MMAI 869
Machine Learning and AI

Association Rule Learning

Stephen Thomas

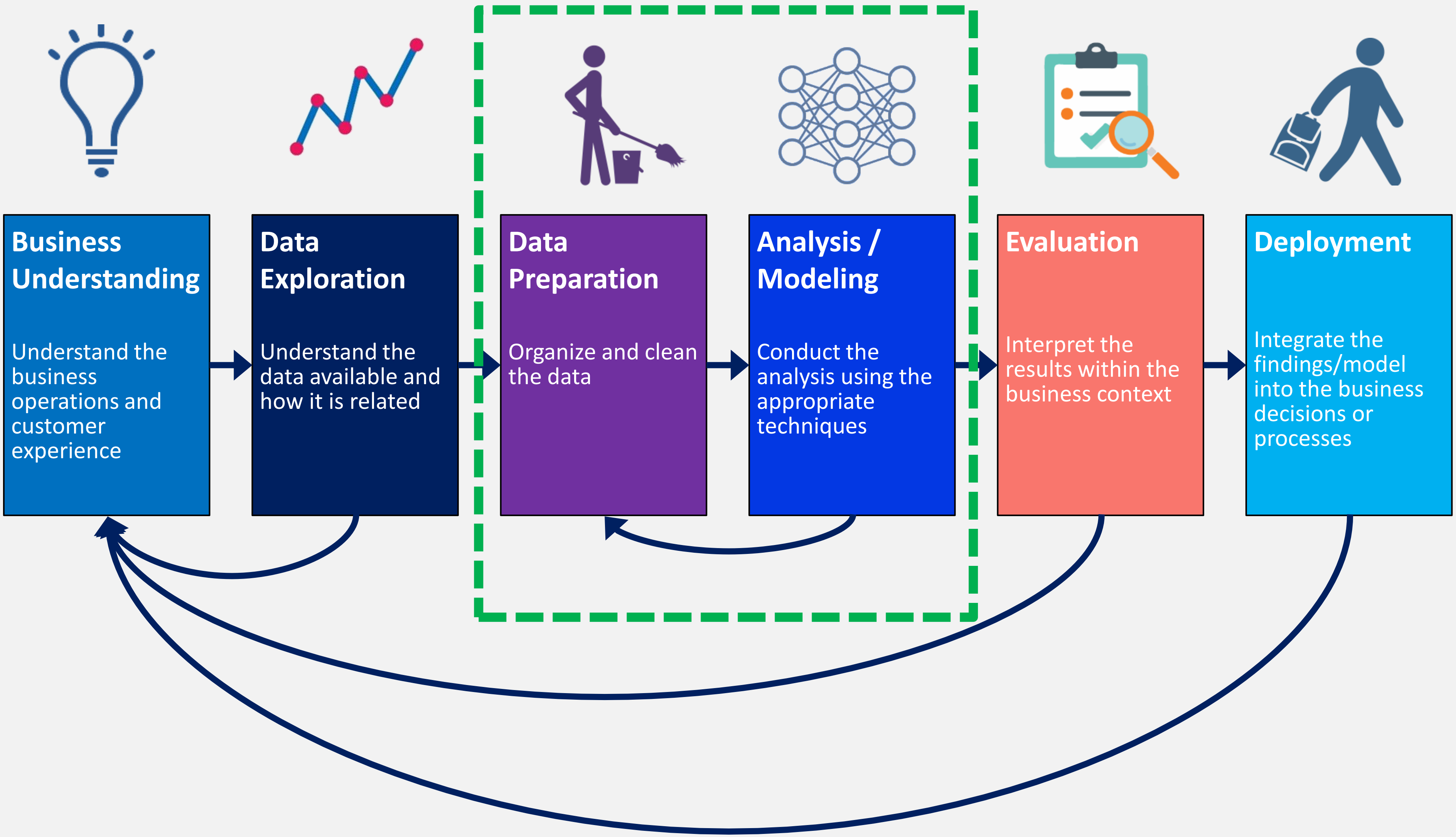
Updated: Nov 7, 2022

Smith School of Business | Queen's University

Outline

- What is the Market Basket Model?
- What are Association Rules and why do we care?
- How do algorithms find them?
- How to use interestingness measures?

The Analytics Process: CRISP-DM



More Detail

Data Preparation

Organize and clean the data

Feature Engineering

Normalization
Discretization
Coding
Temporal, text, image

Feature Selection

Filter
Wrapper

Analysis / Modeling

Conduct the analysis using the appropriate techniques

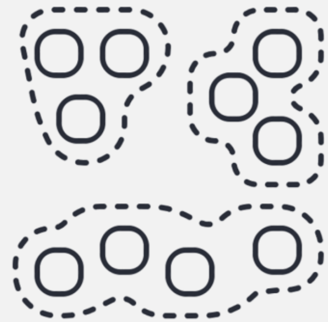
Model Training

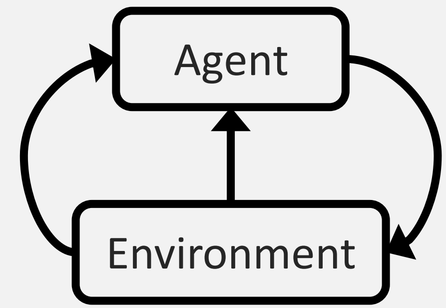
Algorithms
Ensembles
Parameter tuning

Model Selection/ Evaluation

Cross validation
A/B Testing

Three Types of Machine Learning





	Supervised	Unsupervised	Reinforcement
What	Predict something in the future	Find relationships	Learn through trial and error
How	Algorithm builds model from past data	Algorithms finds patterns in data	Algorithm takes actions, gets rewards
Data	Labeled	Unlabeled	None
Tasks/ Algorithms	• Classification – Decision Tree, SVM, Naïve Bayes • Regression – Linear, Polynomial, Lasso • Recommenders – Collaborative filtering, matrix decomposition	• Clustering – K-Means, DBSCAN, Hierarchical • Association rules – Apriori, Eclat, FP-Growth • Dimensionality Reduction – PCA, NMF, LDA, GDA, t-SNE	• Q-learning • SARSA • Deep Q Network

OVERVIEW

Association Rule Learning

Association Rule Learning

noun

- Discovering interesting **relationships** ("rules") in data

Training Phase

ID	Items
1	{citrus fruits, bread, margarine}
2	{tropical fruit, yogurt, cream cheese}
3	{vegetables, whole milk, diapers}
4	{whole milk, butter, yogurt, rice}
5	{rolls/buns, gin, tonic, lime}
…	…

unlabeled training data

association rules algorithm

Apriori, FP-Growth

Who	What	When		What	Confidence	Lift
Men	Diapers	Friday	→	Beer	40%	2.5x
-	Sausage	-	→	Mustard	31%	1.9x
Women	Candy, Pickles	-	→	Pregnancy Test	37%	40.4x
-	Lobsters	-	→	Butter	80%	6.6x

Market Basket Model

- Theory that if you buy certain groups of items, you are more (or less) likely to buy another group of items
  - People who buy Kale are more likely to buy ….
  - People who buy baby food are more likely to buy …

Terms

- ***Item***: product
  - E.g., cheese, milk, cereal
- ***Itemset***: set of items
  - E.g., {beer, diapers, milk} is a 3-itemset
- ***Basket***: an itemset that someone buys
  - E.g., Steve’s basket is {cereal, cheese, apples}
- ***Association rule***: Expression of the form X → Y
  - X and Y are disjoint itemsets
  - E.g., {Dough, Cheese, Pizza sauce} → {Pepperoni}
  - E.g., {Cheeseburger, Fries} →{Milkshake}
  - E.g., {Sausage} → {Mustard}

Exercise

Which of the following describes the rule that people who buy diapers are more likely to buy beer?

- {Diapers, Beer}
- {Diapers, Beer } -> {Beer}
- {Beers} -> {Diaper}
- {Diapers} -> {Beer}

Example: Groceries

Transaction	Items
1	{citrus fruits, semi-finished bread, margarine, ready soups}
2	{tropical fruit, yogurt, cream cheese}
3	{vegetables, whole milk, diapers}
4	{whole milk, butter, yogurt, rice}
5	{rolls/buns, gin, tonic, lime}
6	{vegetables, milk, rolls/buns, bottled beer, liquor}
7	{beef, pickles, butter sausage}
8	{salsa, chips, avocado, salt, garlic, beer, tortilla, sausage, milk}
	…

X	→	Y	Support	Confidence	Lift
{Beer}	→	{Diapers}	5%	20%	1.0
{Gin, Lime}	→	{Tonic}	0.1%	1%	0.3
{Jalapeno, Salsa}	→	{Tortilla}	3%	79%	1.5
{Sausage}	→	{Mustard}	2%	85%	1.9
{Candy, Pickles}	→	{Pregnancy Test}	2%	55%	1.4
{Lobsters}	→	{Butter}	1%	90%	1.6

Example: Medical

Patient	Symptoms
1	{chills, abrasion, blurred vision, anxiety}
2	{swelling, chest pain, alexia, blisters}
3	{otorrhea, heartburn, rash, and chest pain}
4	{back pain, apnea, cough, itching}
5	{hearing loss, fatigue}
6	{fever, dizziness, discharge}
7	{double vision, itching, chills}
8	{nasal discharge, drymouth, muscle weakness}
	…

X	→	Y	Support	Confidence	Lift
{fever, dry cough, tiredness}	→	{covid-19}	…	…	…
{rash, fever, chills}	→	{headache}	…	…	…
{weight loss, blood loss}	→	{fever}	…	…	…
{chest pain, shortness of breath}	→	{neck pain}	…	…	…

Example: Store Locations

Customer	Stores Visited
1	{McDonald's, Indigo, Cineplex, Shoppers, Costco, Sport Chek, LCBO, The Keg}
2	{Rexall, Home Depot, Metro, McDonald's, Taco Bell, Sobey's}
3	{Loblaws, Tiffany, Chipotle, Taco Bell, Petro Canada, Tom's Fish, Lowe's}
4	{Home Hardware, Burger King, The Keg, McDonald's, Rexall, Cineplex, Metro}
5	{Pandora, Bed Bath and Beyond, Adidas, Nike Shop, Wendy's}
6	{Old Navy, Greenhouse Juices, PrettyLittleThing}
7	{McDonald's, Carter's, Bonnie Togs, The Mansion, Red House, LCBO}
8	{SkipTheDishes, Costco, Sobey's, Safeway, Walmart, Beer Store}
	…

X	→	Y	Support	Confidence	Lift
{Lululemon, Yogashop}	→	{Greenhouse Juices}	…	…	…
{Le BBQ Shop, BBQing.com}	→	{Bob's Butcher}	…	…	…
{Tiffany, Porsche}	→	{Chipotle}	…	…	…
{Hollister, Forever 21, H&M}	→	{Cineplex}	…	…	…

Example: Web Analytics

User	Websites visited
1	{msn.com, google.com, espn.com, reddit.com}
2	{fox.com, Tumblr.com, thechive.com, facebook.com, twitter.com, tmall.com}
3	{Instagram.com, facebook.com, twitter.com, tiktok.com, naver.com}
4	{youtube.com, lululemon.com, ebay.com, zoom.com}
5	{stackoverflow.com, google.com, youtube.com}
6	{lichess.com, chess24.com, youtube.com}
7	{msn.com, reddit.com, lichess.com}
8	{wikipedia.com, google.com, facebook.com, Instagram.com}
	…

X	→	Y	Support	Confidence	Lift
{reddit.com, tumblr.com}	→	{thechive.com}	…	…	…
{lichess.com}	→	{chess24.com}	…	…	…
{stackoverflow.com}	→	{udemy.com}	…	…	…
{linkedin.com, bloomberg.com}	→	{forbes.com}	…	…	…

Business Application: Price Bundling

{Samsung Suitcase} → {Samsung Handbag}

Business Application: Assortment

{Barbie dolls} → {Skittles}

Business Application: Consumer Profiles

{broccoli, kale} → {rolled oats}

{broccoli, kale} → {hemp hearts}

…

→

Business Application: Cross Selling

{cheeseburger} → {milkshake}

ALGORITHMS

Main Algorithms

- **Apriori**: The original algorithm
  - Basically, just counting
  - Apriori principle: if {A, B} is frequent, then {A} and {B} must both be frequent.
    - Thus, if {A} is not frequent, then {A, B} cannot be frequent
- **FP Growth**: Newer, faster algorithm
  - Divide-and-conquer allows it to scale to huge datasets

Apriori

https://stream.queensu.ca/Watch/Aa47Fqo9

Apriori Example

- Suppose we want all itemsets that occur at least 2 times

TID	Baskets
1	{A, C, D}
2	{B, C, E}
3	{A, B, C, E}
4	{B, E}

1st scan

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

prune

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

generate

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

Itemset	Support
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

prune

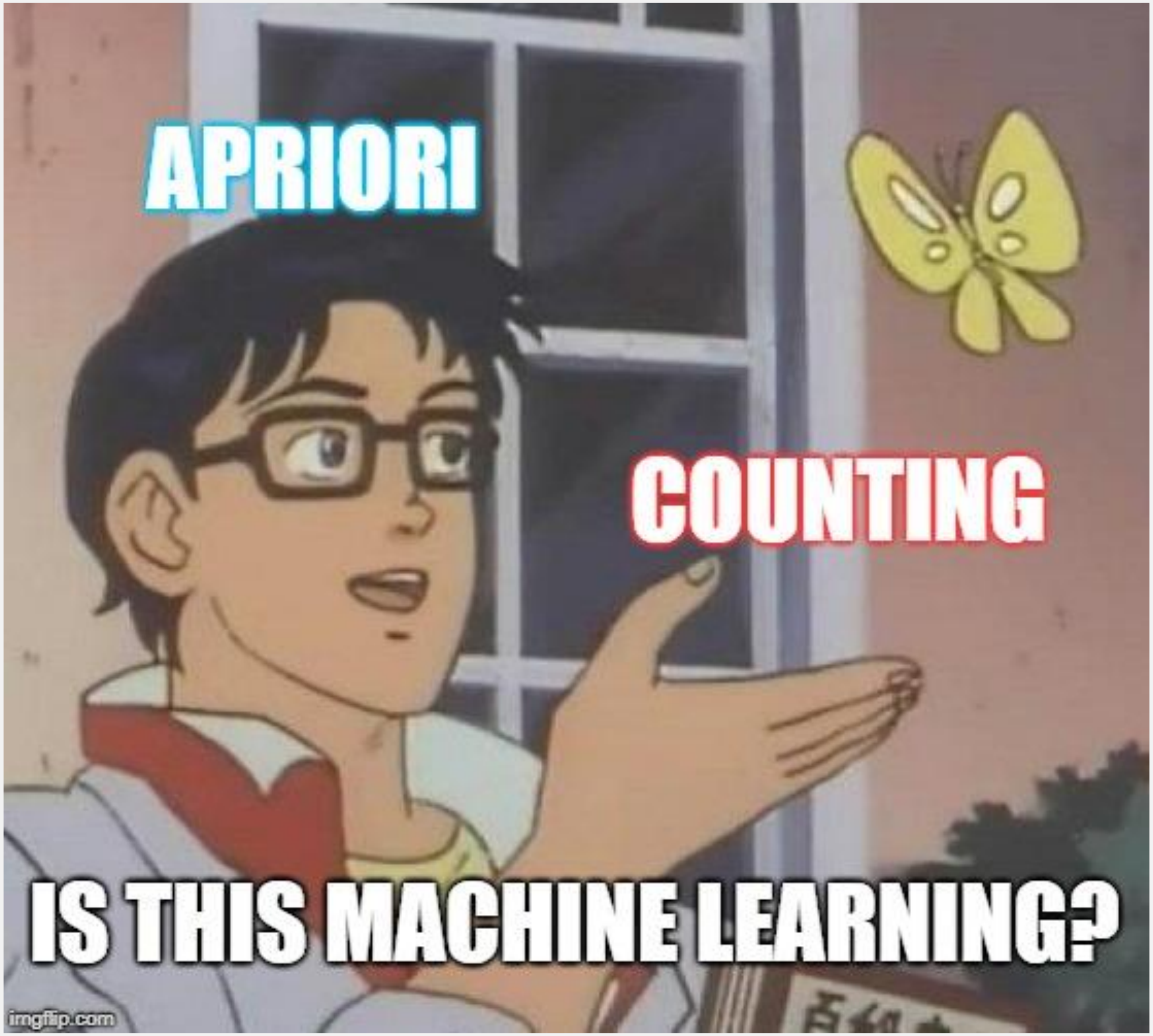
Itemset	Support
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

generate

Itemset
{B, C, E}

3rd scan

Itemset	Support
{B, C, E}	2


APRIORI
COUNTING
IS THIS MACHINE LEARNING?
imgflip.com

INTERESTINGNESS MEASURES

Interestingness Measures

- Usually, there *a lot* of rules discovered from a dataset
  - Many are uninteresting or redundant
- ***Interestingness measures*** can be used to prune/rank the rules
  - Support, Confidence, Lift

Consider the rule {**Sausage**} → {**Mustard**}

	Definition:	**Low Value means:**	**High Value means:**
Support 0% - 100%	How many baskets contain Sausage and Mustard?	**Few** baskets contain Sausage and Mustard	**Many** baskets contain Sausage and Mustard
Confidence 0% - 100%	How likely are Sausage buyers to also buy Mustard?	Sausage buyers ***unlikely*** to also buy Mustard	Sausage buyers ***likely*** to also buy Mustard
Lift 0 - ∞	Do Sausage buyers buy Mustard more often than average buyers?	Sausage buyers buy Mustard **less than average**	Sausage buyers buy Mustard **more than average**

Support

- $S(X \rightarrow Y) = \dfrac{\#\text{ transactions with X and Y}}{\#\text{ transactions}}$

- Percentage of transactions/baskets containing **X** and **Y**
- Order doesn't matter

TID	Bread	Milk	Coke	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	1	0	1	1
5	0	1	1	0	1

Rule: X → Y	S(X → Y)
{Milk} → {Coke}	3/5 = 60%
{Diaper, Milk} → {Beer}	??
{Coke, Milk} → {Bread}	??

Confidence

- $C(X \rightarrow Y) = \frac{S(X \,\&\, Y)}{S(X)}$

- How likely are **X** buyers to also buy **Y**?
- Order does matter

TID	Bread	Milk	Coke	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	1	0	1	1
5	0	1	1	0	1

Rule: X → Y	C(X → Y)
{Milk} → {Coke}	3/4 = 75%
{Diaper, Milk} → {Beer}	??
{Coke, Milk} → {Bread}	??

Exercise

Calculate the following:

- S({Beer} → {Coke}) =
- C({Pepsi} → {Juice}) =
- S({Beer, Coke} → {Milk}) =

TID	Beer	Coke	Pepsi	Milk	Juice
1	1	0	1	0	0
2	0	1	0	1	0
3	1	1	0	0	1
4	0	0	1	1	0
5	1	1	0	1	0
6	0	0	0	1	1
7	0	0	1	0	1
8	1	1	0	1	1

Drawback of Confidence

- {Tea} → {Coffee}
- C({Tea} → {Coffee}) = 0.75
  - Pretty high, right?
- S({Coffee}) = 0.90
  - 90% of people buy coffee
  - But only 75% of tea buyers do
  - Tea buyers are ***less*** likely to buy coffee, despite high confidence
- Reason for pitfall: confidence does not take into account the support of the RHS

Lift

- $L(X \rightarrow Y) = \dfrac{C(X \rightarrow Y)}{S(Y)}$
- Do **X** buyers buy **Y** more often than average buyers?
- Order matters
- Like confidence, but takes into account support of the RHS

TID	Bread	Milk	Coke	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	1	0	1	1
5	0	1	1	0	1

Rule: X → Y	C(X → Y)	S(Y)	L(X → Y)
{Milk} → {Coke}	75%	60%	75/60 = 1.25
{Diaper, Milk} → {Beer}	?	?	?
{Coke, Milk} → {Bread}	?	?	?

Example Measures

You usually want to find rules that are high in all three

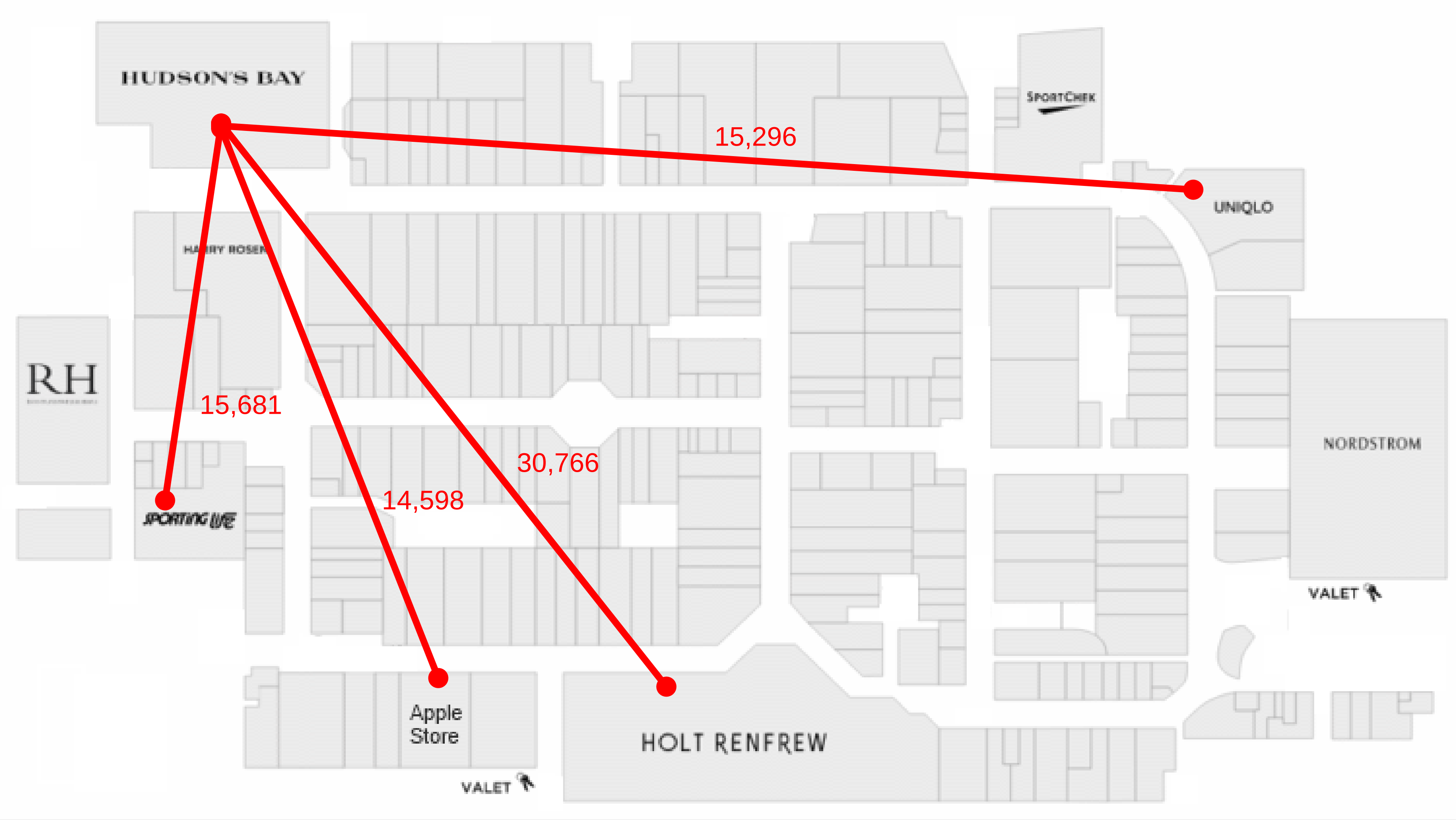
LHS		RHS	Support	Confidence	Lift
{Canned Beer}	→	{Milk}	5%	20%	1.0
{Canned Beer}	→	{Berries}	0.1%	1%	0.3
{Canned Beer}	→	{Chips}	0.3%	79%	1.5
{Sausage}	→	{Mustard}	3%	85%	1.9
{Sausage}	→	{Ketchup}	2%	55%	1.1
{Sausage}	→	{Milk}	1%	30%	0.8
{Sausage, Chips}	→	{Canned Beer}	1%	90%	1.6

Possible Actions

- Suppose X → Y is a "good" rule
  - High support, confidence, lift, etc.
- Possible actions:
  - Put X and Y close together
  - Package X with Y
  - Package X and Y with a poorly selling item
  - Give discount on only one of X and Y
  - Increase the price of X and lower the price of Y
  - Advertise only one of X and Y
  - …

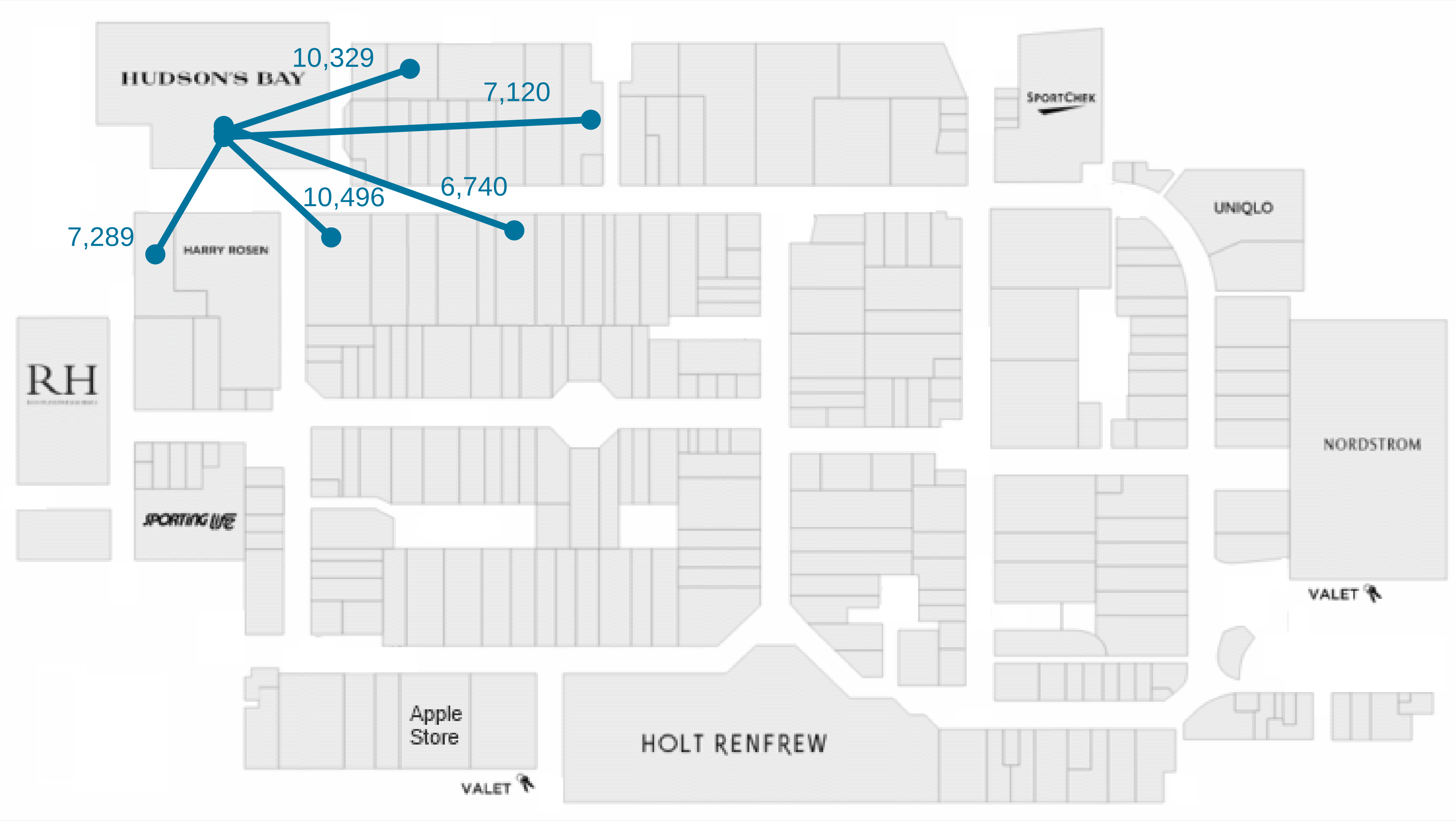
Example: Yorkdale Mall

- Built rules from 1.8M customer journeys
- Wanted rules that showed a ***long customer journey***
- Red: Rules containing HB and LHS is greater than 200 meters away



Example: Yorkdale Mall

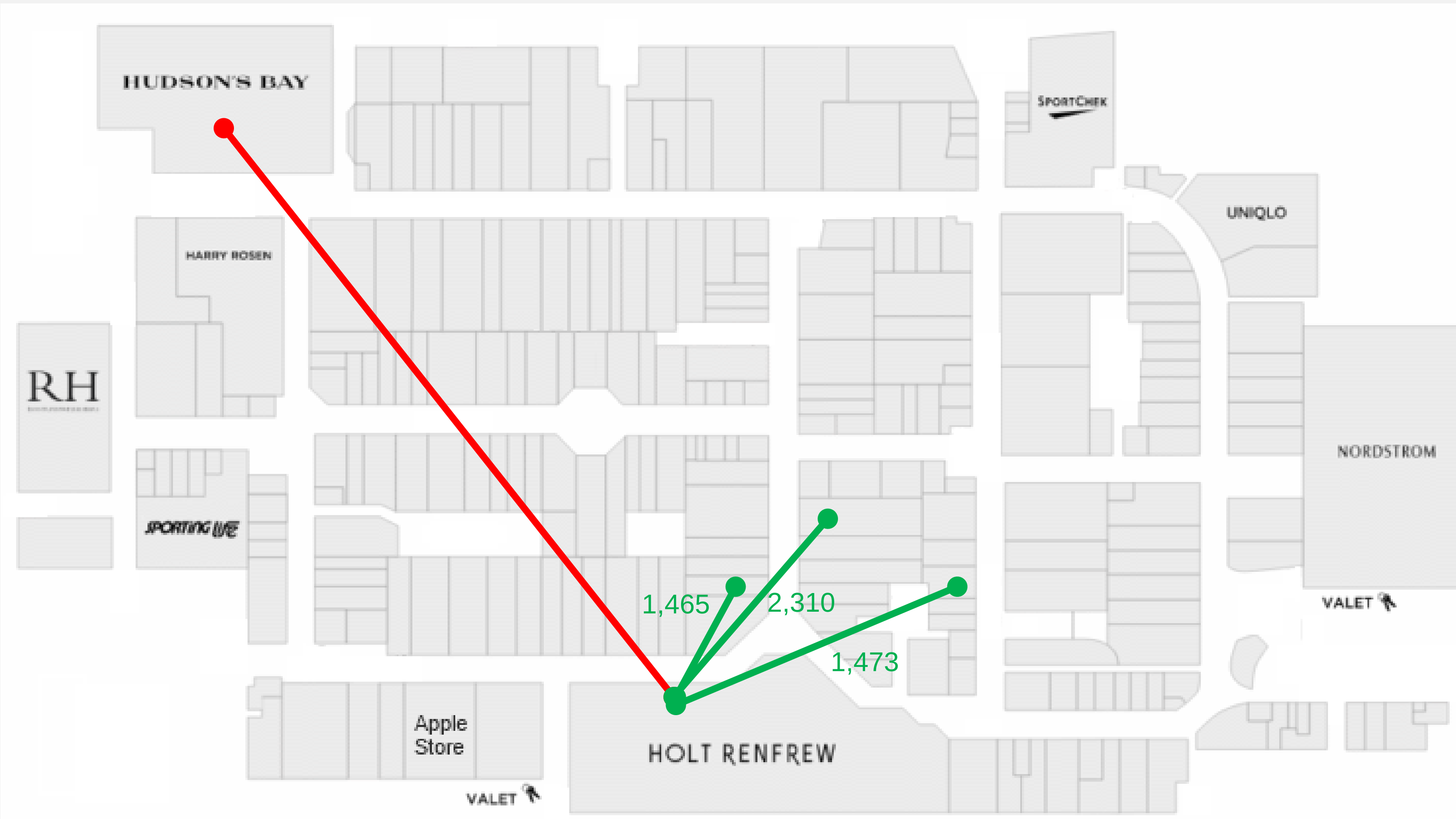
- Wanted rules that showed very high confidence
- Blue: Rules containing HB and have confidence > 10%



Source: Team Bay, MMA 2019W

Example: Yorkdale Mall

- Wanted rules with ***high spend***
- Green: Rules with HB and HR on LHS, and RHS contains "luxury" stores with > 30% conf



Source: Team Bay, MMA 2019W

ASSOCIATION RULES IN PYTHON, R

mlxtend

Python package for association rule learning (and more)

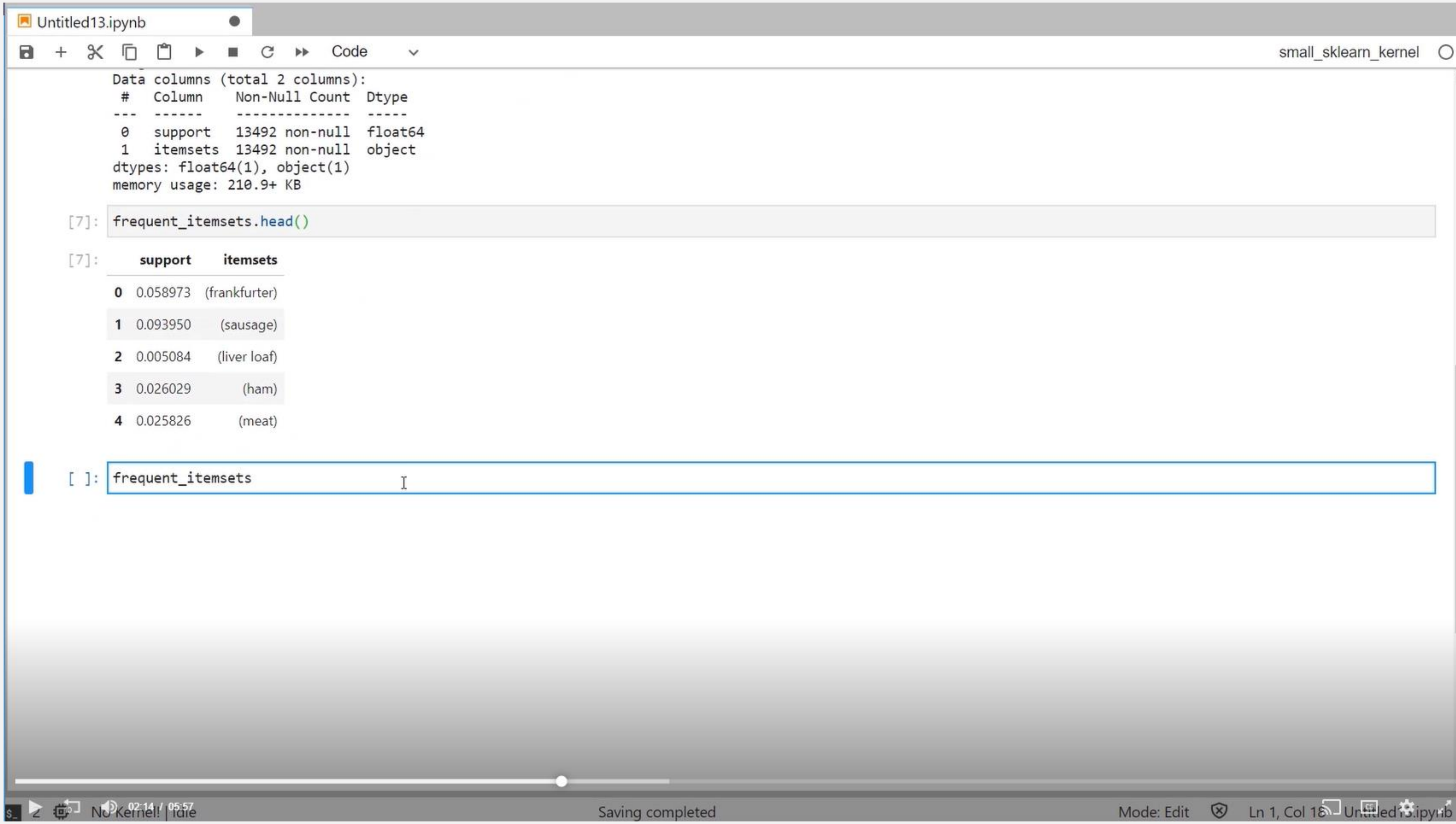
```
from mlxtend.frequent_patterns import apriori

%time frequent_itemsets = apriori(df, min_support=0.001, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
```

```
Wall time: 16.6 s
```

```
frequent_itemsets.head(10)
```

	support	itemsets	length
0	0.058973	(frankfurter)	1
1	0.093950	(sausage)	1
2	0.005084	(liver loaf)	1
3	0.026029	(ham)	1
4	0.025826	(meat)	1
5	0.006507	(finished products)	1
6	0.002237	(organic sausage)	1
7	0.042908	(chicken)	1
8	0.008134	(turkey)	1
9	0.057651	(pork)	1



https://stream.queensu.ca/Watch/Mj57JnNf

arules

- Popular R package for association rule learning
- arules.Rmd

```r
library(arules)
data("Groceries")

# Build the rules
rules <- apriori(Groceries, parameter = list(supp = 0.01, conf = 0.3, target = "rules"))

# List the top 40 rules, ordered by lift
inspect(head(rules, n=40, by = "lift"))

# Calculate two more interestingness measures: hyperConfidence, and conviction
quality(rules) <- cbind(quality(rules),
                  hyperConfidence = interestMeasure(rules, measure = "hyperConfidence",
                                                    transactions = Groceries),
                  conviction = interestMeasure(rules, measure="conviction", transactions = Groceries))

# List the top 40 rules, ordered by lift
inspect(head(rules, n=40, by = "conviction"))
```

17:46 (Top Level) R Script

SUMMARY

Summary

- **Market basket model**
  - People who buy one thing, often buy another thing
- **Association Rules**
  - Rules of the form X→Y
- **Algorithms**
  - Apriori: original, slower
  - FP Growth: newer, faster
- **Interestingness Measures**
  - Support, confidence, lift, …
- **Business Applications**
  - Product bundling, cross selling, shelf placement, …
- Other topics (not covered today)
  - How other algorithms work under the hood
  - How to visualize rules
  - Subjective interestingness measures (surprise, utility, novelty, …)
  - Data preprocessing (continuous values, time stamps, etc.)
  - Sequence Mining

APPENDIX

FP Growth Example: Pass One

- Read through once to find support for each item
- Discard infrequent items
- Re-order items in each basket from highest to lowest support

TID	Items
1	A, B
2	D, B, C
3	C, E, A, D
4	E, A, D
5	B, A, C
6	A, D, B, C
7	A, F
8	C, A, B
9	B, D, A
10	B, E, C

1st scan →

Itemset	Support
{A}	8
{B}	7
{C}	6
{D}	5
{E}	3
{F}	1

prune →

Itemset	Support
{A}	8
{B}	7
{C}	6
{D}	5
{E}	3

reorder →

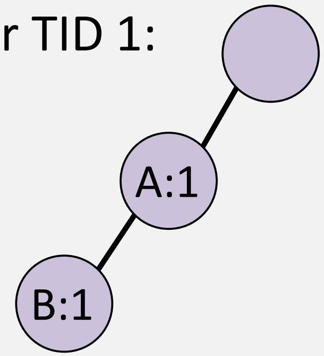
TID	Items
1	A, B
2	B, C, D
3	A, C, D, E
4	A, D, E
5	A, B, C
6	A, B, C, D
7	A
8	A, B, C
9	A, B, D
10	B, C, E

FP Growth Example: Pass Two

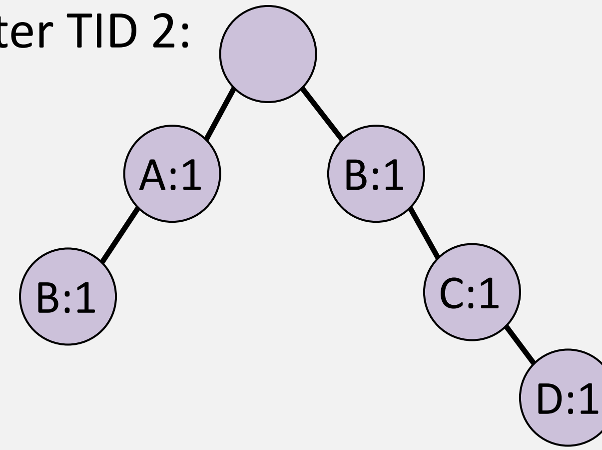
For each basket, add a path to the tree

TID	Items
1	A, B
2	B, C, D
3	A, C, D, E
4	A, D, E
5	A, B, C
6	A, B, C, D
7	A
8	A, B, C
9	A, B, D
10	B, C, E

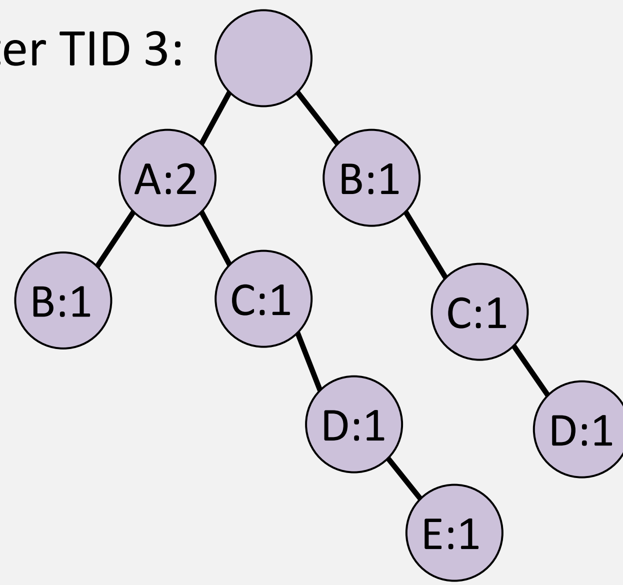
After TID 1:



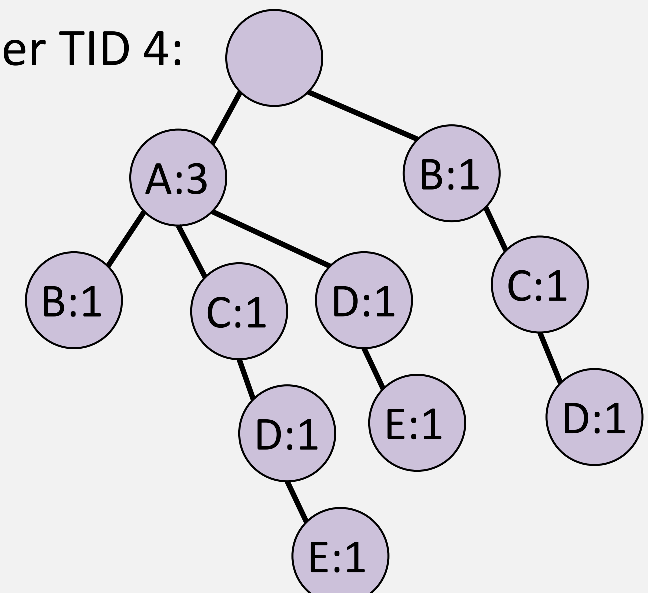
After TID 2:



After TID 3:



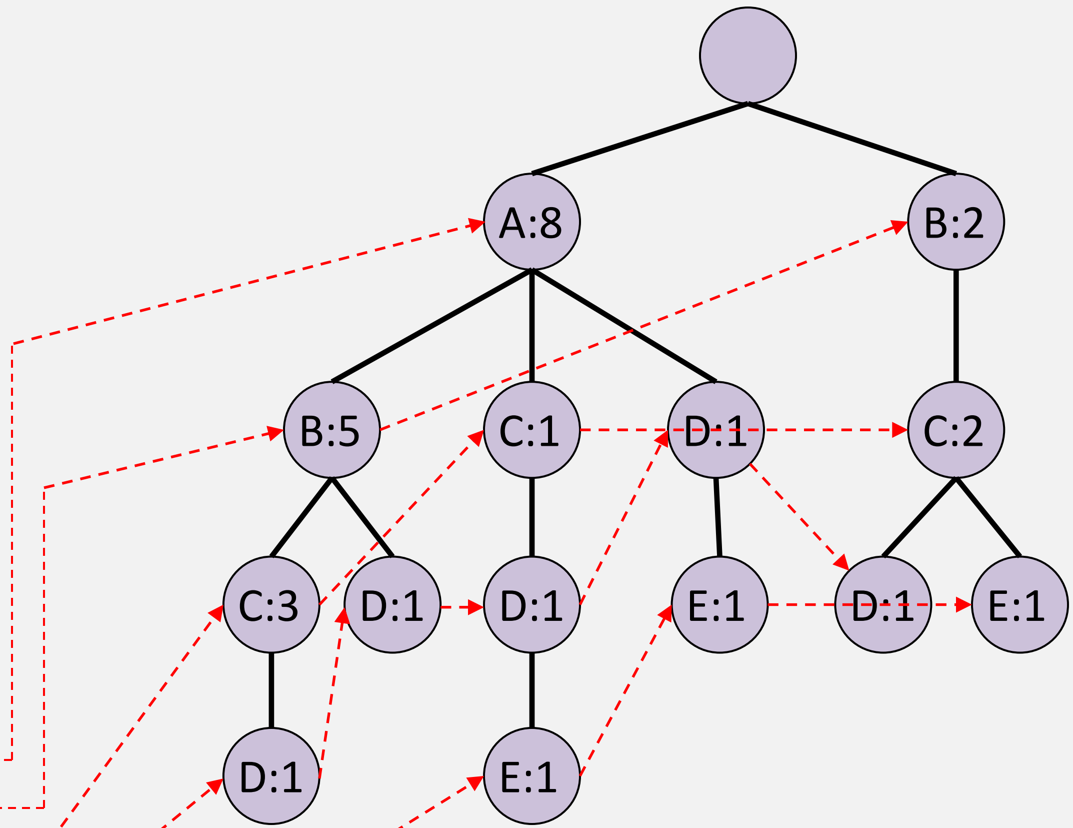
After TID 4:



FP Growth Example: Final Tree

- Final tree for all 10 transactions
- Add pointers to assist in frequent itemset generation
- Generate frequent itemsets from a bottom-up traversal of tree (details omitted here)

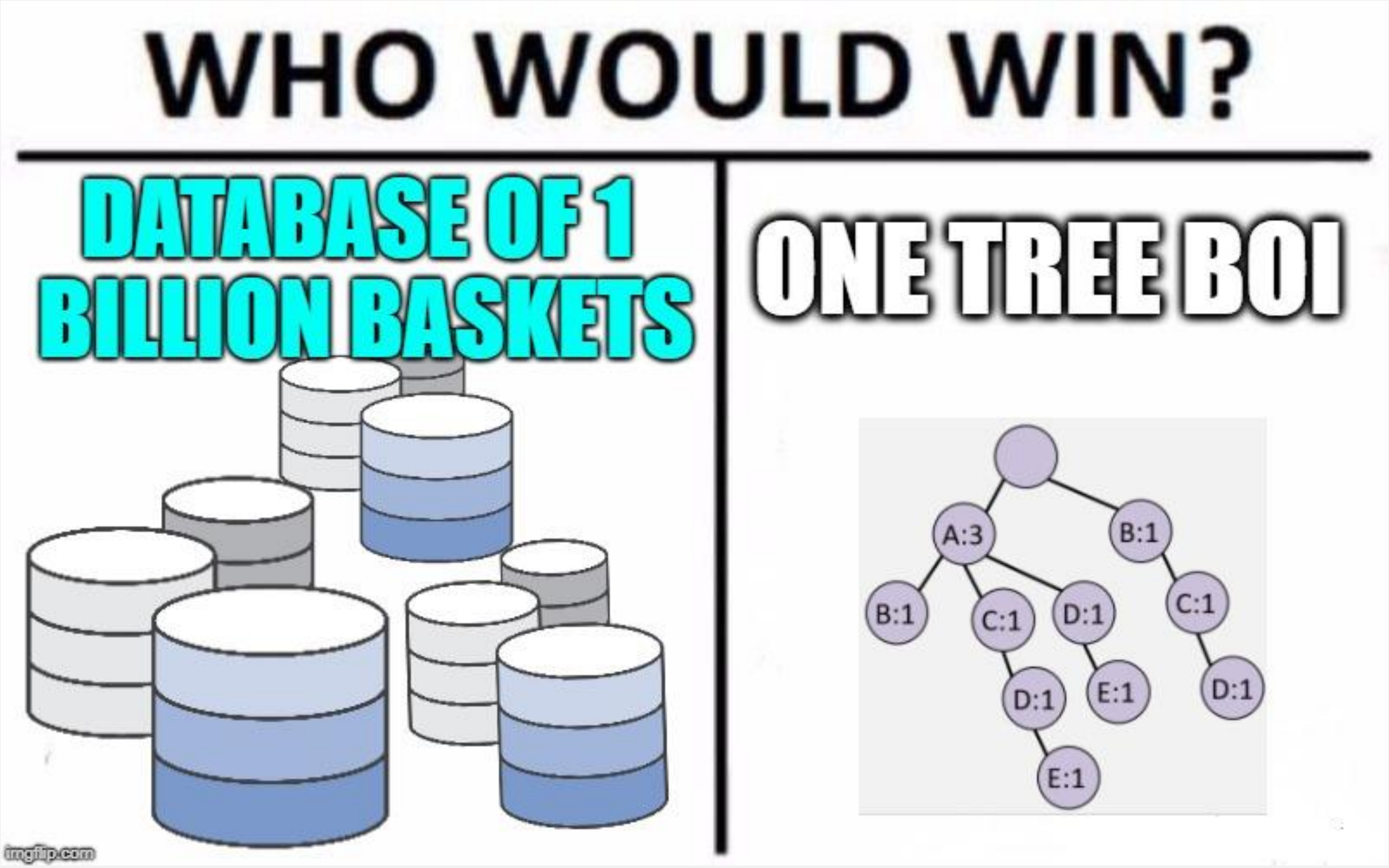
TID	Items
1	A, B
2	B, C, D
3	A, C, D, E
4	A, D, E
5	A, B, C
6	A, B, C, D
7	A
8	A, B, C
9	A, B, D
10	B, C, E



Item	Pointer
A	
B	
C	
D	
E	

Header table

FPGrowth is Scalable




HADOOP
APRIORI
FPGROWTH
imgflip.com

Uncle Steve's Algorithm Comparison

Algorithm	Summary	Pros	Cons
Apriori	*Makes N passes over the data to find increasingly-large itemsets*	✓ Simple ✓ Implemented everywhere	× Slow × Lots of passes over data × Uses lots of memory
FP Growth	*Builds an FP Tree and extracts itemsets from the tree*	✓ Only 2 passes over data ✓ Compresses data ✓ Can be parallelized ✓ Much faster than Apriori	× Not as available (yet)

Subjective Interestingness Measures

With the help of a domain expert, rules can be selected or removed based on certain criteria:

Concise	• Contains relatively few items
Diverse	• Items different significantly from each other
Surprising	• Contracts existing knowledge or expectations
Actionable	• How difficult it is it implement the rules
Utility	• Helps reach a goal (e.g., items with high margins)

Generating Rules

Once all frequent itemsets are discovered (from any algorithm), association rules are generated by:

- Taking all possible subsets s of each frequent itemset I
  - E.g., I = {A, B, C}
    - {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, {A, B, C}
- Considering all rules of the form s → {I – s}
  - {A} → {C}
  - {A, B} → {C}
  - {B} → {A}
  - {B} → {C}
  - {B} → {A, C}
  - etc.
- Evaluate each rule against the minimum confidence threshold.

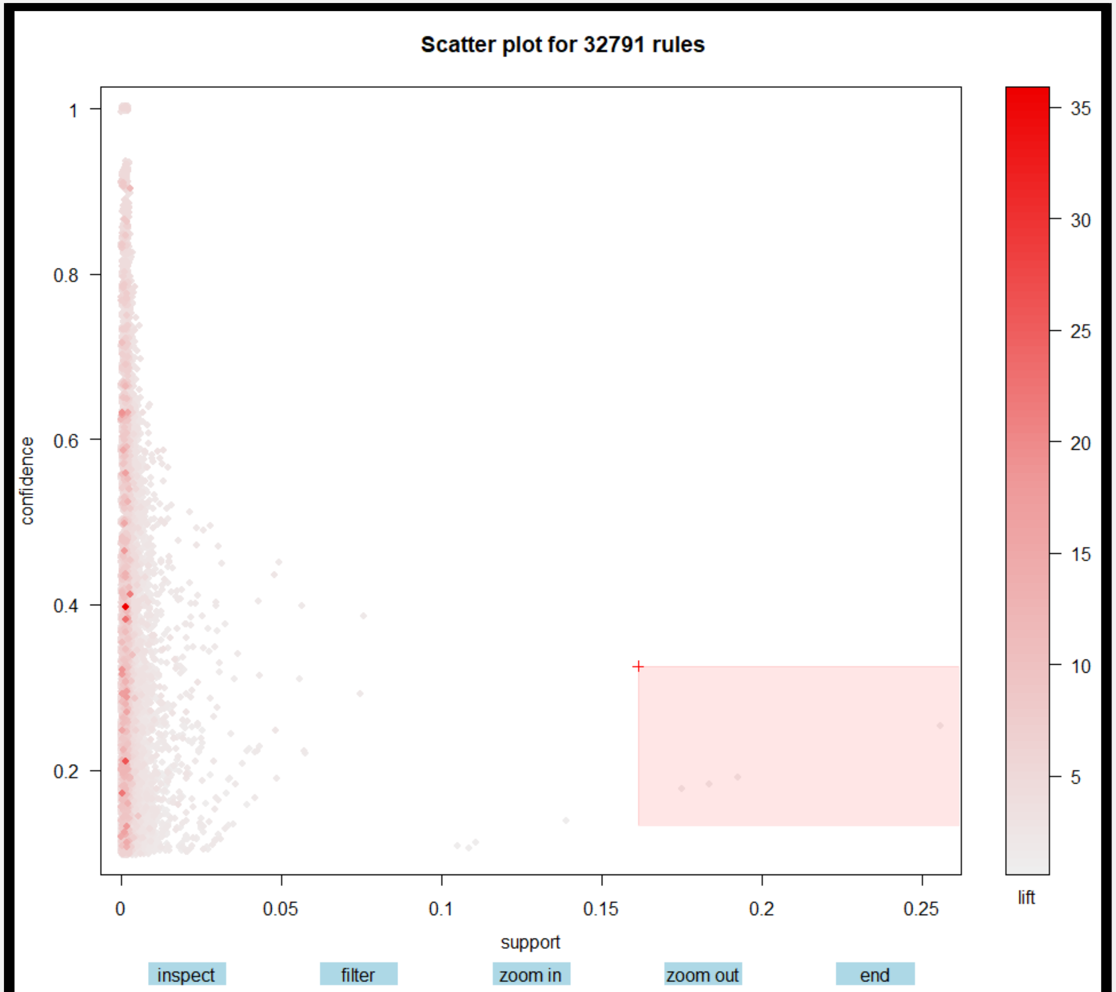
VISUALIZATION OF ASSOCIATION RULES

Visualizations of Rules

- Visualizations of rules can help you explore and understand
- Main R package: arulesViz

Interactive Scatter Plots

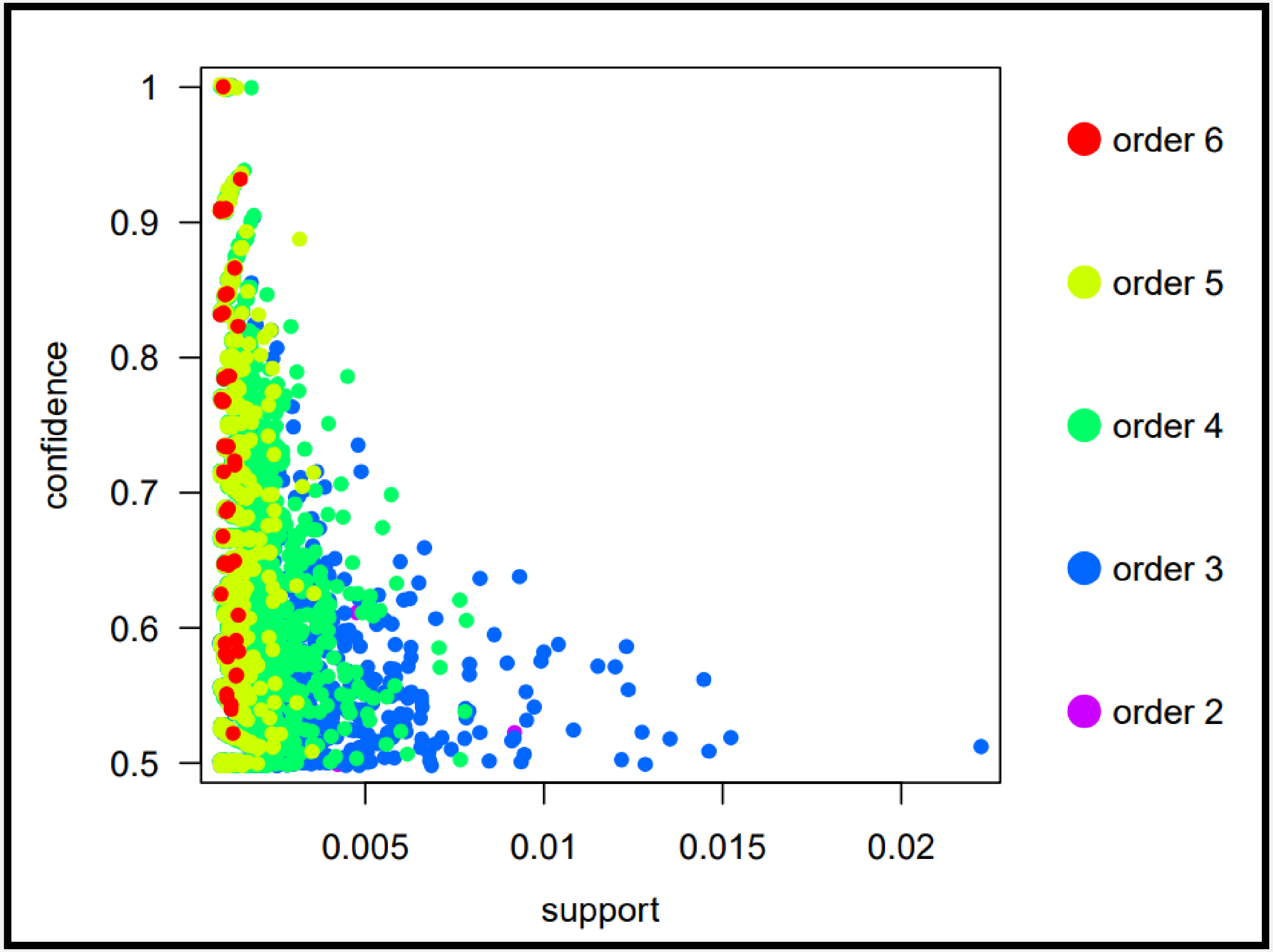
```
plot(rules, interactive = T)
```




```
Number of rules selected: 4
    lhs    rhs                  support   confidence lift conviction order
[1] {}  => {rolls/buns}         0.1839349 0.1839349  1    1          1
[2] {}  => {soda}               0.1743772 0.1743772  1    1          1
[3] {}  => {other vegetables}   0.1934926 0.1934926  1    1          1
[4] {}  => {whole milk}         0.2555160 0.2555160  1    1          1
```

Two-key Plot

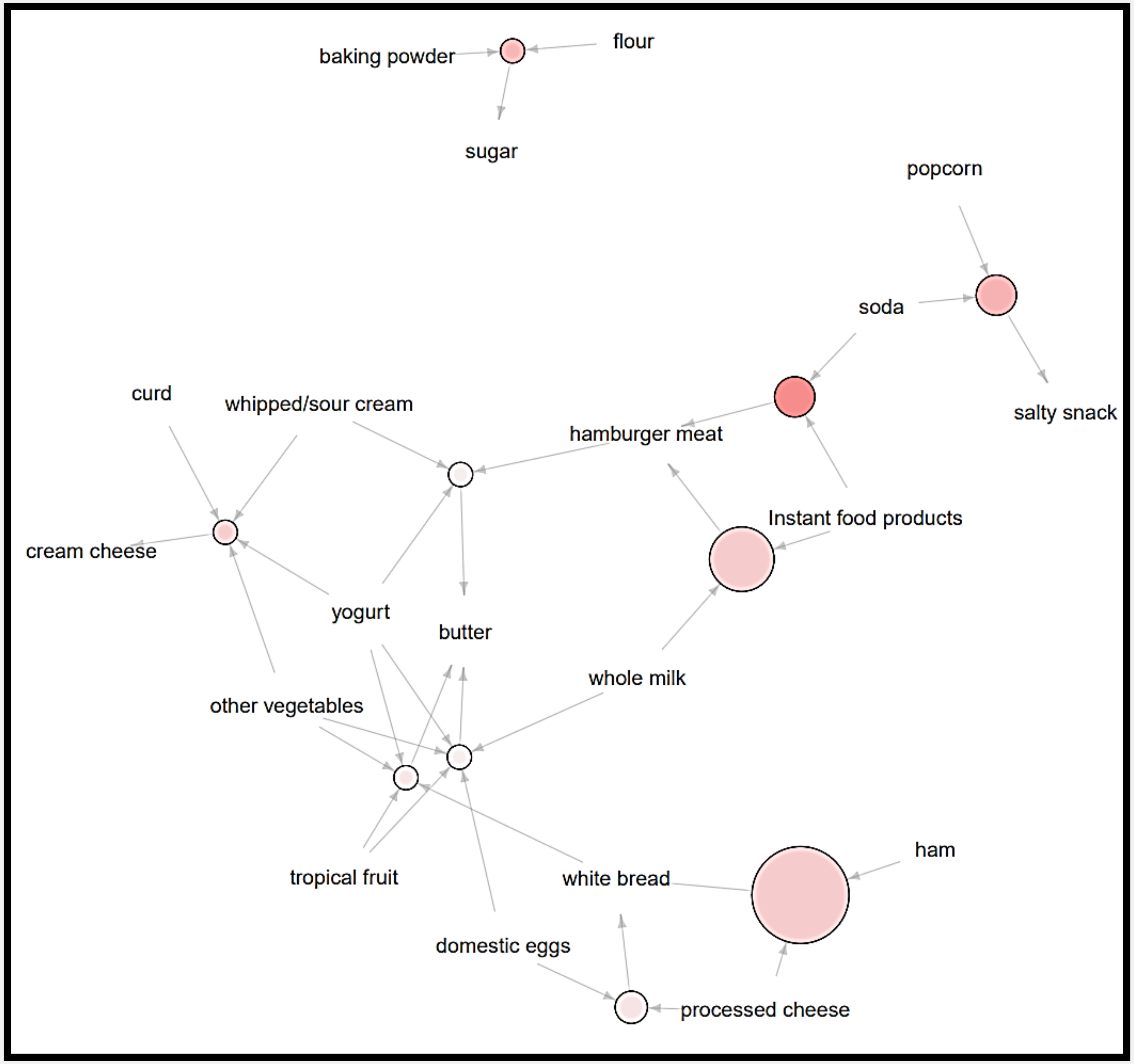
```
plot(rules, method = "two-key plot")
```



- ***Order*** is the number of items contained in the rule
- In this plot, it is clear that order and support have inverse relationship

Graph-based Visualization

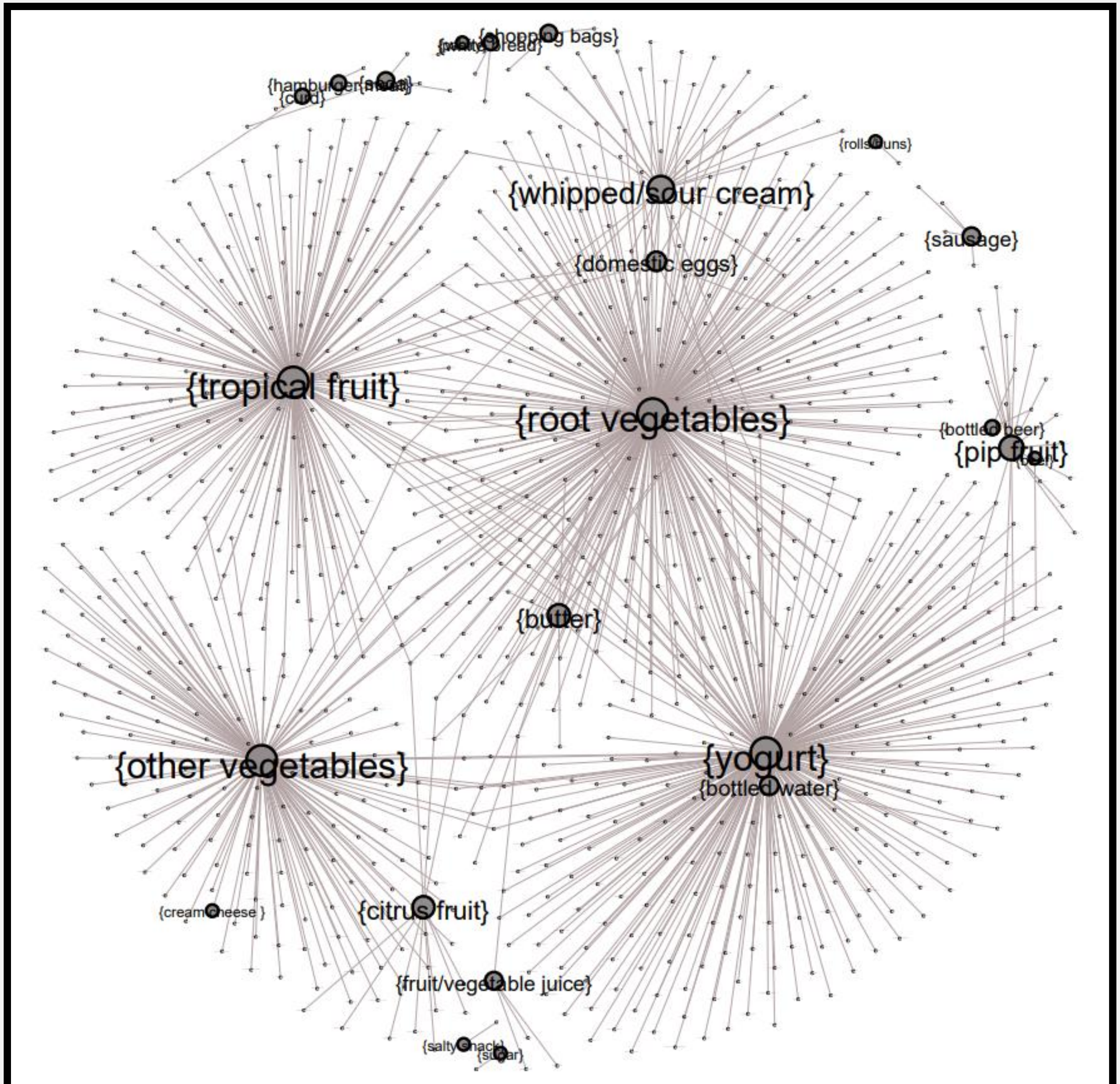
```
plot(rules, method = "graph")
```



- Circles are rules
- Graph useful for showing rules that share items
- Size: support
- Color: lift

Graph-based Visualization

```
saveAsGraph(head(rules, n = 1000, by = "lift"), file = "rules.graphml")
```



- Rules can be imported into graph tools like Gephi
- Interactivity is then enabled

APPLYING RULES

Example: Price Bundling

- Want to boost sales at a local Korean restaurant
  - Lunch special, loyalty program, combo menu, ...?
- Easiest option is to offer a combo menu with existing food items
- But which items, and how much of discount to offer?
- Method:
  - Run association rules to discover popular combinations
  - Run simulation to determine level of discount

Source: Ryan Park (MMA), *Association Rule and Bundle Pricing*

Example: Price Bundling

- Ran association rules, and looked for rules with a min confidence of 30%
- Management liked the second rule in particular

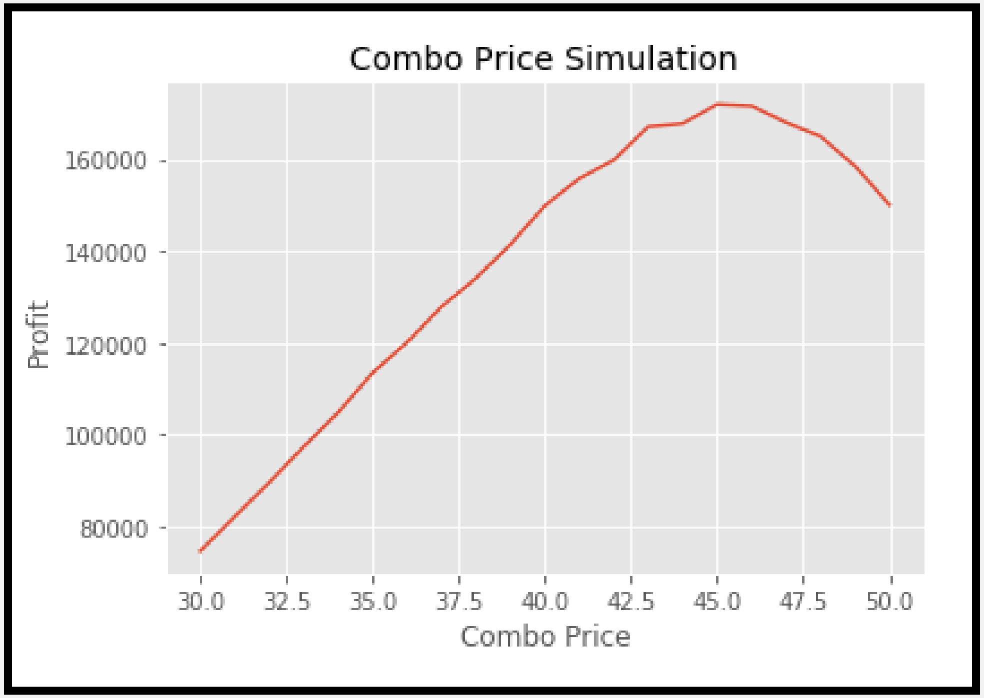
LHS		RHS	Confidence	Lift
{Jap Chea}	→	{LA Galbi}	32%	2.2
{Seafood Pancake}	→	{Gam Ja Tang}	30%	1.8
{Jap Chea}	→	{Seafood Pancake}	31%	2.4

Seafood Pancake

Gam Ja Tang

Example: Price Bundling

- How much to charge?
- Assume:
  - Seafood Pancake = $19.99 (profit $8), Gam Ja Tang = $29.99 (profit $12)
  - A customer's reservation price is normally distributed around current price, with standard deviation of $5
  - A customer would purchase the combo if it costs the same or less their reservation price of the two individual items
- Run simulation of 10,000 customers



Optimal price is $45 (10% discount)

ADVANCED TOPICS

Continuous Variables

- What if data has continuous variables as well?

Age	Work	Education	Race	Sex	Hrs.	Native	Income
39	State-gov	Bachelors	White	Male	40	United-States	<=50K
50	Self-emp-not-inc	Bachelors	White	Male	13	United-States	<=50K
38	Private	HS-grad	White	Male	40	United-States	<=50K
53	Private	11th	Black	Male	40	United-States	<=50K
28	Private	Bachelors	Black	Female	40	Cuba	<=50K
37	Private	Masters	White	Female	40	United-States	<=50K
49	Private	9th	Black	Female	16	Jamaica	<=50K
52	Self-emp-not-inc	HS-grad	White	Male	45	United-States	>50K
31	Private	Masters	White	Female	50	United-States	>50K
42	Private	Bachelors	White	Male	40	United-States	>50K
37	Private	Some-college	Black	Male	80	United-States	>50K
30	State-gov	Bachelors	Asian-Pac-Islander	Male	40	India	>50K
23	Private	Bachelors	White	Female	30	United-States	<=50K
32	Private	Assoc-acdm	Black	Male	50	United-States	<=50K
40	Private	Assoc-voc	Asian-Pac-Islander	Male	40	?	>50K
34	Private	7th-8th	Amer-Indian-Eskimo	Male	45	Mexico	<=50K
25	Self-emp-not-inc	HS-grad	White	Male	35	United-States	<=50K
32	Private	HS-grad	White	Male	40	United-States	<=50K
38	Private	11th	White	Male	50	United-States	<=50K
43	Self-emp-not-inc	Masters	White	Female	45	United-States	>50K
40	Private	Doctorate	White	Male	60	United-States	>50K

Continuous Variables

- What if data has continuous variables as well?
- Solution: transform variables into buckets

Age	Work	Education	Race	Sex	Hrs.	Native	Income
30-39	State-gov	Bachelors	White	Male	40-60	United-States	<=50K
50-59	Self-emp-not-inc	Bachelors	White	Male	10-20	United-States	<=50K
30-39	Private	HS-grad	White	Male	40-60	United-States	<=50K
50-59	Private	11th	Black	Male	40-60	United-States	<=50K
20-29	Private	Bachelors	Black	Female	40-60	Cuba	<=50K
30-39	Private	Masters	White	Female	40-60	United-States	<=50K
40-49	Private	9th	Black	Female	10-20	Jamaica	<=50K
50-59	Self-emp-not-inc	HS-grad	White	Male	40-60	United-States	>50K
30-39	Private	Masters	White	Female	40-60	United-States	>50K
40-49	Private	Bachelors	White	Male	40-60	United-States	>50K
30-39	Private	Some-college	Black	Male	80-100	United-States	>50K
30-39	State-gov	Bachelors	Asian-Pac-Islander	Male	40-60	India	>50K
20-29	Private	Bachelors	White	Female	30-40	United-States	<=50K
30-39	Private	Assoc-acdm	Black	Male	40-60	United-States	<=50K
40-49	Private	Assoc-voc	Asian-Pac-Islander	Male	40-60	?	>50K
30-39	Private	7th-8th	Amer-Indian-Eskimo	Male	40-60	Mexico	<=50K
20-29	Self-emp-not-inc	HS-grad	White	Male	30-40	United-States	<=50K
30-39	Private	HS-grad	White	Male	40-60	United-States	<=50K
30-39	Private	11th	White	Male	40-60	United-States	<=50K
40-49	Self-emp-not-inc	Masters	White	Female	40-60	United-States	>50K
40-49	Private	Doctorate	White	Male	40-60	United-States	>50K

Time-stamped Data

- Can we take advantage of time-stamped data?

TID	Date	Bread	Milk	Coke	Beer	Diaper
1	2018-01-12	1	1	1	0	0
2	2018-01-12	1	0	0	1	0
3	2018-01-13	0	1	1	1	1
4	2018-01-15	1	1	0	1	1
5	2018-01-17	0	1	1	0	1

- Yes!
- Group data by period (day, month, or year) and run association rule algorithm separately on each group.
  – Some tools will even track rules' measures over time.

Differential MBA

- Can you include other data, like customer demographics?

TID	Date	StoreID	Customer Sex	Customer Age	Bread	Milk	Coke	Beer	Diaper
1	2018-01-12	33	M	20-29	1	1	1	0	0
2	2018-01-12	34	M	20-29	1	0	0	1	0
3	2018-01-13	35	F	40-49	0	1	1	1	1
4	2018-01-15	33	F	30-39	1	1	0	1	1
5	2018-01-17	65	F	40-49	0	1	1	0	1

- Yes!
- Group data by attribute (StoreID, Sex, Age) and run association rule algorithm separately on each group.
- Example: if we observe a rule holds in one store, but not in any other, then there must be something about that store.

Itemset Support

Support for itemset I*:* Percentage of transactions containing I

TID	Bread	Milk	Coke	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	1	0	1	1
5	0	1	1	0	1

S({Beer}) = 3/5

S({Bread}) = 3/5

S({Beer, Bread}) = 2/5

Frequent Itemset

Frequent itemset: itemset with support at least s

– You chose a support threshold s

TID	Bread	Milk	Coke	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	1	0	1	1
5	0	1	1	0	1

S({Beer}) = 3/5

S({Bread}) = 3/5

S({Beer, Bread}) = 2/5

If support threshold is 50%, then which itemsets are frequent?

Choosing Minimum Support Threshold

- Algorithms need you to define a minimum support threshold
- How to set appropriately?
  - If too high, you could miss rules involving interesting rare items (e.g., expensive jewelry)
  - If too low, algorithms takes too longer to run, and number of rules is too big
  - Answer depends on your goals and allotted time
- Trial and error is often used

History: Diapers and Beer