Association Rule Mining

Session 01

Brock Tibert

BU

Boston University Questrom School of Business

Outline for Today

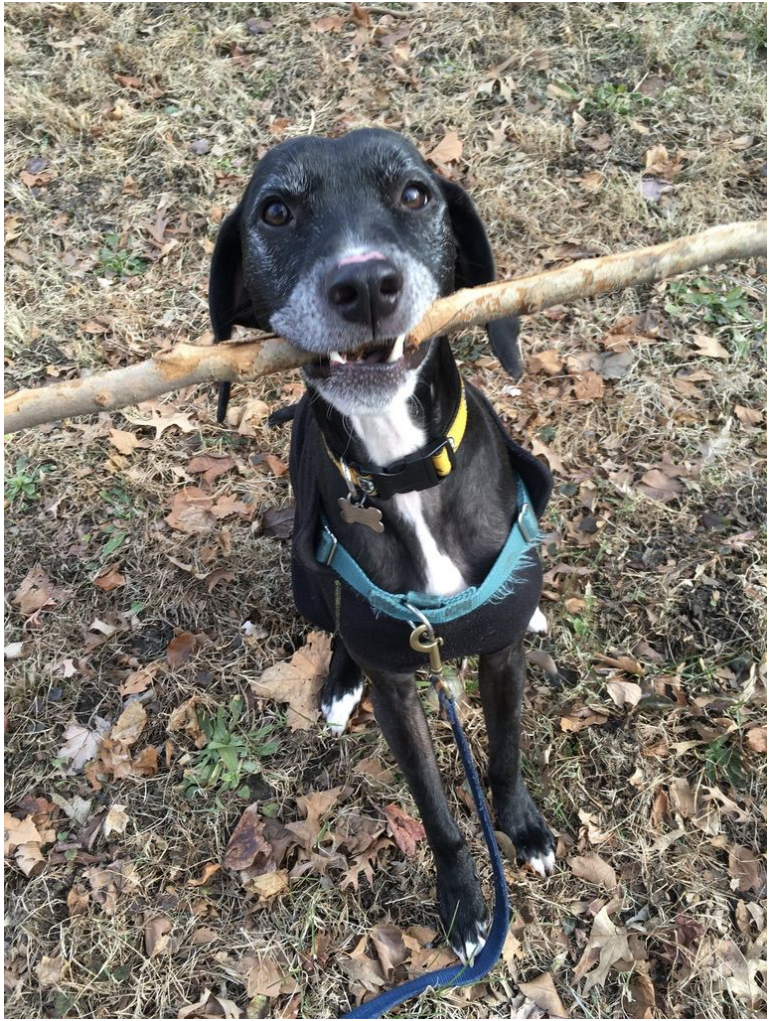
- Discuss the mechanics and delivery of the course
- Base `python` versus External Packages
- What we will cover and learn this semester
- Topic 1: Association Rules for Pattern Discovery
  - Method/Approach
  - Evaluation
  - Applications
  - Hands on lab

I will use the terms Association Rules and Market Basket Analysis interchangeably

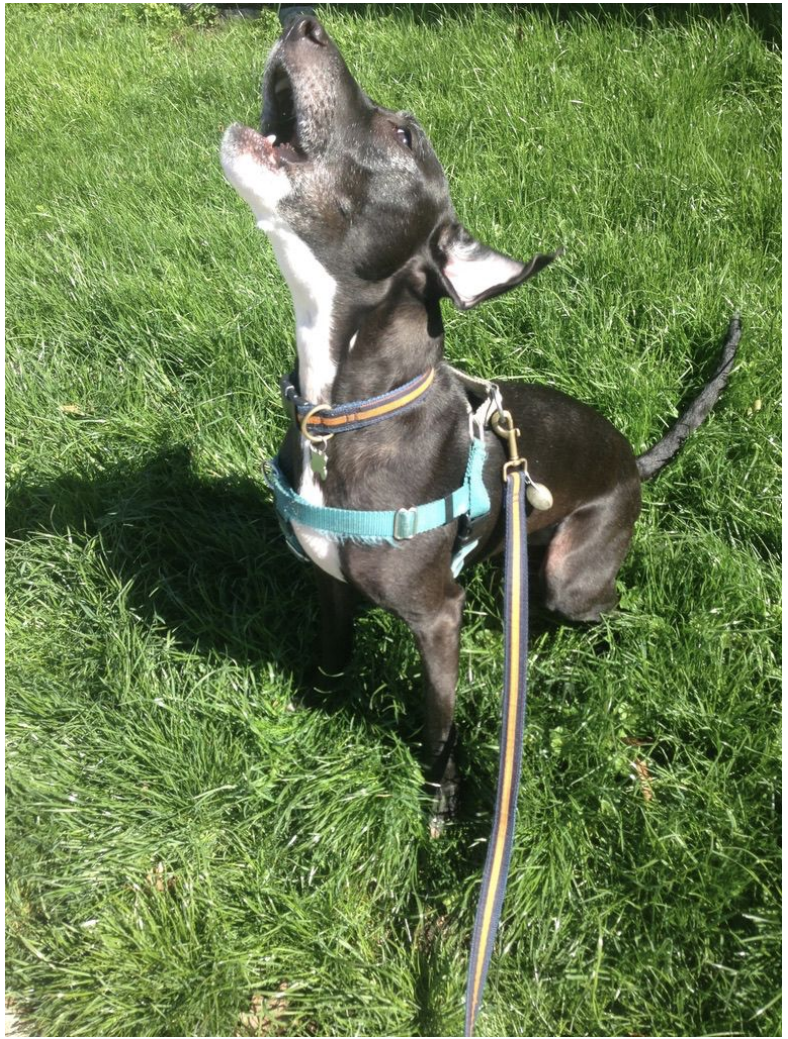
About Me

- 17 years applying “data science” in a variety of industries
- Prior to joining the IS Faculty in Questrom I ran a small edtech firm
- Interests are:
  - applied data science
  - graph analytics
  - product management
  - Sports analytics
  - applications of text mining (Conversational agents)

My Dog Bodhi (Most of the time)



My Dog Bodhi (*sometimes*)



Course Delivery Overview

Details on setup will be provided later this week

Our class sessions will be recorded for review on demand. The recordings ***should*** be posted automatically after the recorded session has finished processing. These can be found on QuestromTools.

- The majority of the resources will be found in the class repo
  - I will post our lecture notes there, and add content as we go through the semester, so pull changes often!
  - Discussions

- We will build on prior classes and use Google Cloud tools
  - Big Query to query our datasets
  - I encourage you to use this course to continue to practice your cloud skills!

- Whatever environment works for you is 100% ok with me
- I will *mostly* write code in VS Code, but may hop into Colab for certain modules

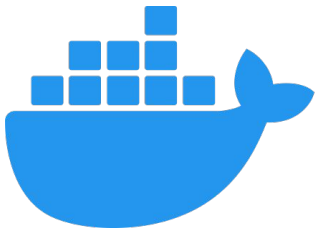
Notebook environments are great for exploration, but the non-linear flow can be a source of bugs/error

Deploying our work as data scientists (more often than not) requires that we have structure in our projects, which is why I encourage you to get comfortable outside of notebooks*

* papermill is the exception

BigQuery

tableau

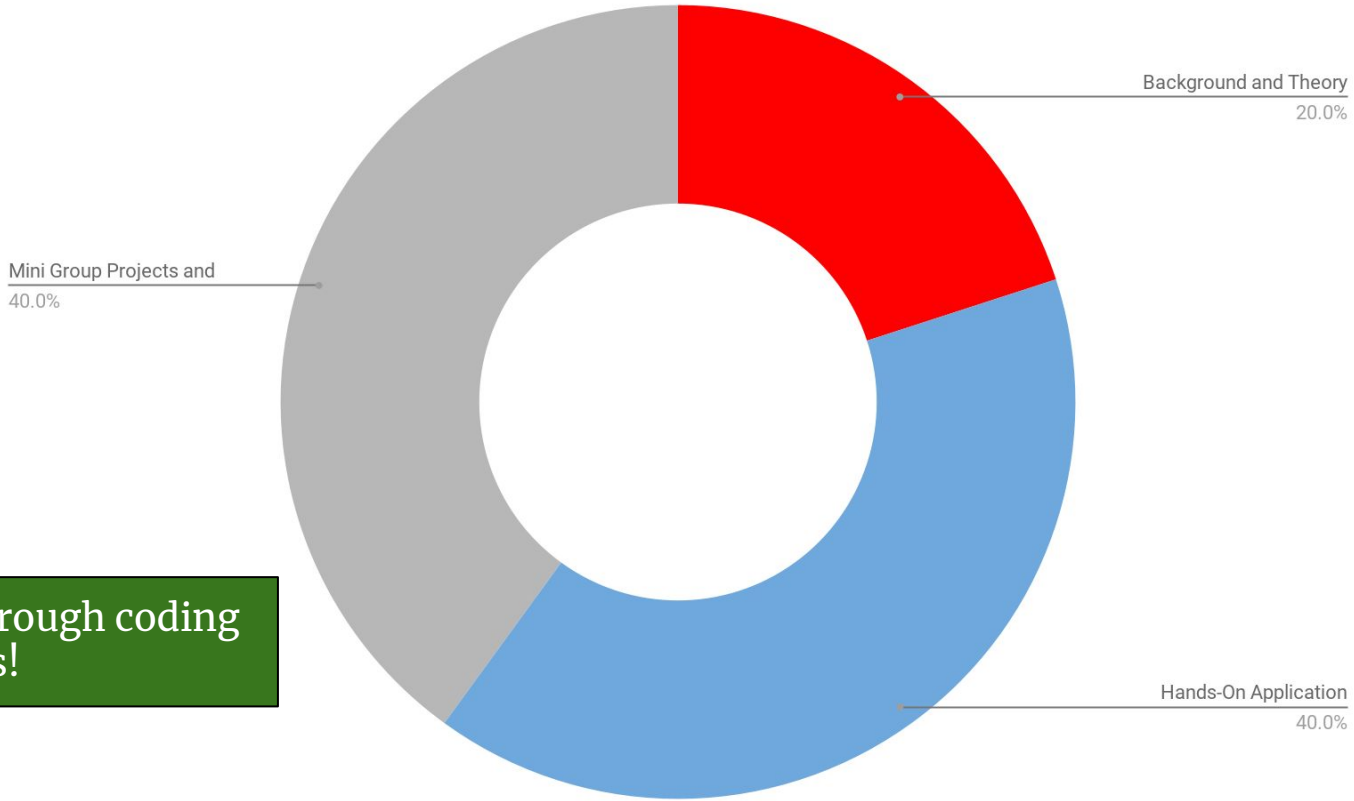


docker

Streamlit

FastAPI

Composition of the Course



Learning through coding up solutions!

Course Delivery

- Discuss the method and the mechanics/understanding of how it works
- Explore the tools available to us in the `python` ecosystem
- Build mental models of how to include each technique into a larger analytics project, with a focus on reporting technical results to a nontechnical audience
- In-class Exercises / Case Studies / Competitions to be completed as teams
- Yes we will write code, but it's about coding up the model **AND** interpreting/presenting the findings

Overview

Class preparation makes these advanced topics, especially when we are writing code, easier in the long run. I do leave it to you, though, to determine what works best for your learning style

- DataCamp review and practice
- I will record supplemental videos
- Github Repository
  - Discussions (you get credit!) and peer-to-peer support
  - Cheat Sheets and resources to supplemental readings
- Supplemental books listed in the syllabus
- You will leave this class with the ability to build both write code to solve analytical challenges, and discuss your findings with a variety of audiences
- It's about understanding the methods, applying them, and telling a story of the results. Why does it matter? Are we introducing risk/uncertainty into the business?

Beyond base python

Moving Forward

- I am certain by now you are increasingly comfortable with python and tools like `pandas`
- We will be using a large number of external packages this course
- **Please use the Discussion forums on Github for help with installation and setup of these packages, though a number do come pre-configured on Colab**
- There are many, many ways to attack some of the problems that we are doing, and some may be handled differently in DataCamp. That's ok!
    - **Exposure to what is possible and a variety of ways to solve the problem is important!**

What we will be covering in BA820

Unsupervised Machine Learning

- Association Rules (~ Reco Engines)
- Clustering Segmentation Methods
- Dimensionality Reduction/Embeddings

Text Analytics

- Text Pre-processing
- Sentiment Analysis
- Text Classification and Clustering
- NLP

- Unsupervised Machine Learning helps identify structure and patterns in our datasets, which also can be used to generate features downstream supervised ML tasks.

- The majority of the data we generate on a daily basis is unstructured. We will learn how the foundation of Natural Language Processing and Understanding (NLP/NLU) and see how both UML and SML techniques still apply to text!

Market Basket Analysis Overview

Market Baskets

A classic application of market basket analysis addresses this question:

Which items are likely to be purchased together?

- If product A and product B often go together, then placing a more **expensive alternative** to B near the display for A can create an **upsell opportunity.**
- If product A and B are often purchased together, putting **one item on sale** at a time can drive purchases for both.

Hardware Store Example

A hardware store has 25 shopping aisles. *To help customers find product easily*, which products should be placed near one another?

- Key-cutting near paint or near door hardware?
- Lawn ornaments near garden or near indoor decorative ornaments?
  - Could be harder to rely on intuition to figure out

Association Rules can be used as simple recommendation engine but is more commonly thought as a pattern recognition technique to surface "if this, than that" type of rules.

What it is:

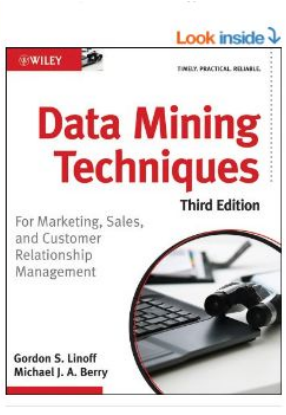
- Find patterns in transactional data in order to determine the co-occurence, or association, between items.

What it does:

- Given a known set of items (LHS), what other items (RHS) tend to co-occur with the LHS

What it means:

- We can leverage the RHS items and treat these as recommendations given the LHS.



Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management Kindle Edition

by Gordon S. Linoff (Author), Michael J. A. Berry (Author)

4.0 out of 5 stars 31 customer reviews

See all 9 formats and editions

Kindle	Paperback	Unknown Binding
$26.39	$30.80	from $64.68
Read with Our Free App	67 Used from $11.66 56 New from $21.39	3 Used from $64.68 1 New from $140.67

The leading introductory book on data mining, fully updated and revised!

When Berry and Linoff wrote the first edition of *Data Mining Techniques* in the late 1990s, data mining was just starting to move out of the lab and into the office and has since grown to become an indispensable tool of modern business. This new edition—more than 50% new and revised— is a significant update from the previous one, and shows you how to harness the newest data mining

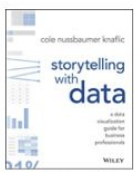
Read more

Customers who bought this item also bought

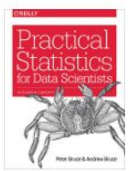
Page 1 of 11

How to Measure Anything: Finding the Value of Intangibles in Business
› Douglas W. Hubbard
4.5 out of 5 stars 156
Kindle Edition
$26.39

R for Data Science: Import, Tidy, Transform, Visualize, and Model Data
› Hadley Wickham
4.5 out of 5 stars 99
Kindle Edition
$19.49

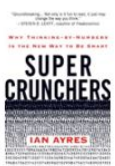


Storytelling with Data: A Data Visualization Guide for Business Professionals
› Cole Nussbaumer...
4.5 out of 5 stars 273
#1 Best Seller in Library Management

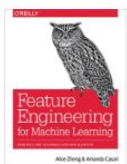


Practical Statistics for Data Scientists: 50 Essential Concepts
› Peter Bruce
4.0 out of 5 stars 44
Kindle Edition
$12.79

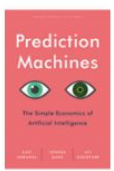
Data Science for Business: What You Need to Know about Data Mining and...
› Foster Provost
4.5 out of 5 stars 186
Kindle Edition
$20.48



Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart
› Ian Ayres
4.0 out of 5 stars 157
Kindle Edition
$1.99



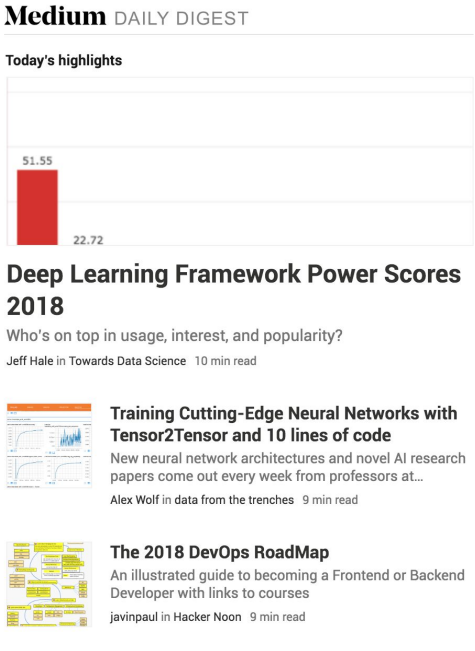
Feature Engineering for Machine Learning: Principles and...
› Alice Zheng
4.5 out of 5 stars 6
Kindle Edition
$29.99

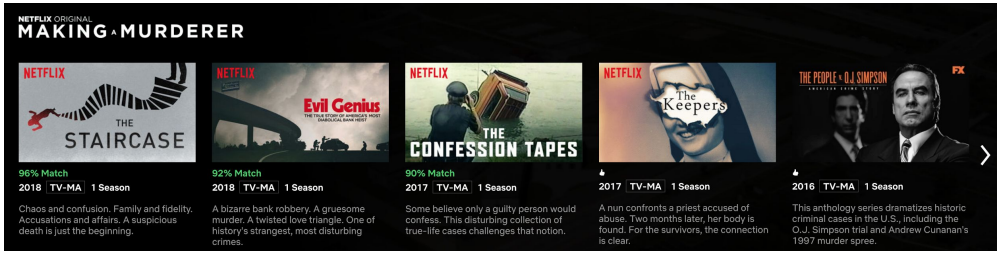


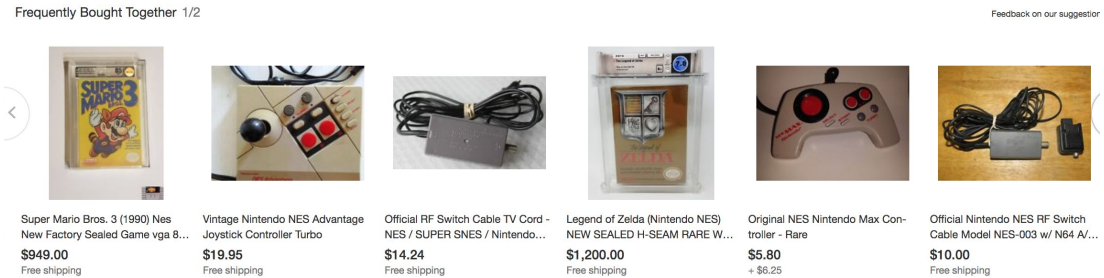
Prediction Machines: The Simple Economics of Artificial Intelligence
Ajay Agrawal
4.5 out of 5 stars 42
Kindle Edition
$16.19

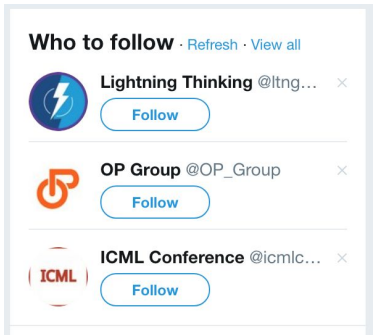
Practical Time Series Forecasting with R: A Hands-On Guide [2nd...
› Galit Shmueli
4.0 out of 5 stars 12
Kindle Edition
$9.99

Other Examples?









The Algorithm

This approach is popular because it's fairly simple to frame for non-technical decision makers.

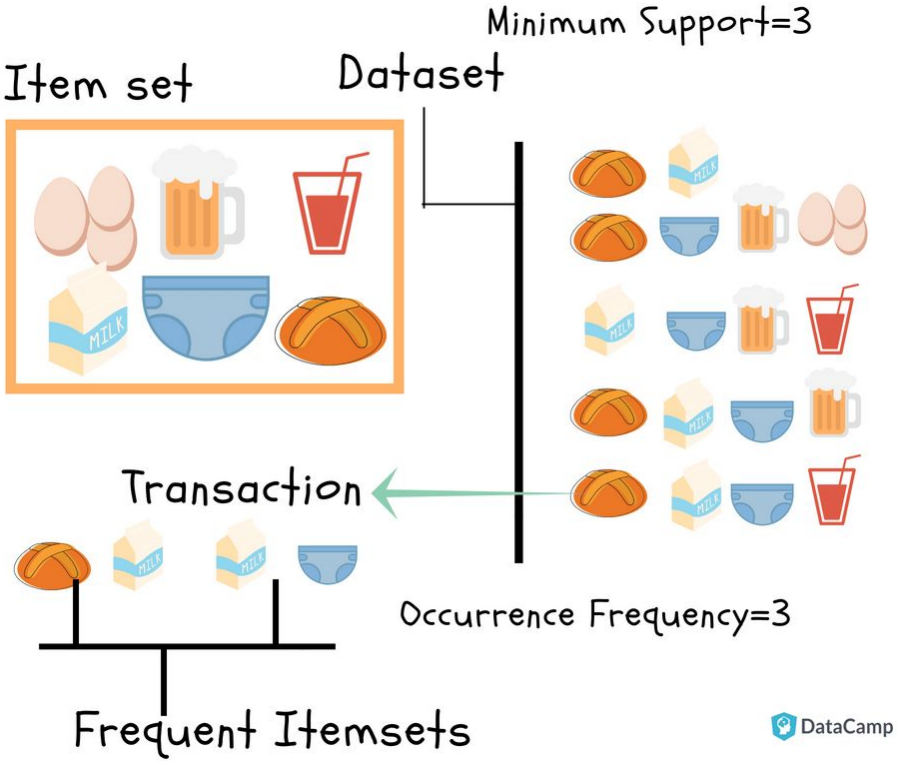
We will be using the Apriori algorithm. This is found within the frequent patterns module in the mlxtend package.

1. Identify all items (and subsequently, itemsets) that meet certain thresholds, such as a minimum number of times an item must appear across all of the transactions
2. Generate rules from the Frequent itemsets. These rules must also meet certain criteria such as how many times they appear in the transactions.

The apriori algorithm needs to perform full scans of the database to evaluate candidate itemsets and rules. This can be computationally intensive on larger datasets and the total number of items to consider.

Terminology

- `Items`: The set of objects or items available to be purchased, or viewed, or streamed.
- `Transaction`: A trip to the store, Netflix viewing history, your most recent Spotify songs streamed. A transaction contains one or more items.
- `Rule`: Can be considered an if *this* then *that*.
  - If purchase bread and eggs, then also purchase milk.
  - {Bread, Eggs} => {Milk}
- `LHS`, or `antecedent`: Left-hand side of the rule
  - The known set of objects. {Bread, Eggs}
- `RHS`, or `consequent`: The items that are associated, or co-occur with the `LHS`
  - Above, this would be {Milk}.
- `(Frequent) Itemset`: A collection of one or more items.



A Visual Reference

- The dataset of transactions is on the right
- Items = eggs, milk, bread, beer, diapers, cola
- Itemsets are a collection of k-items (usually 2 or more)
- Frequent Itemsets are itemsets that are common across the transactions and meet our thresholds for support and confidence.

You also see two new terms; Support and Frequency. As you can imagine, Association rules are built on top of statistics and we uses these metrics to guide our decision making to find interesting rules. We will come back to the evaluation in a moment.

Let’s see this in action

Market Basket Analysis: Support

Support (A=> B) =

$$\frac{\text{transactions containing } \{A,B\}}{\text{all transactions}}$$

Interpretation:
For all of the transactions in the dataset, how frequently does the itemset {A,B} occur

Market Basket Analysis: Confidence

Confidence (A=> B) =

$$\frac{\text{transactions containing } \{A,B\} = \text{Support } \{A,B\}}{\text{transactions containing } A = \text{Support } \{A\}}$$

Interpretation:
For **all** of the transactions containing LHS **A**, how confident are we that the rule/transaction also includes RHS **B** is true.

Market Basket Analysis: Lift

Lift (A=> B):

$$\frac{\text{transactions containing } \{A,B\} = \text{Support } \{A,B\}}{\text{Support } \{A\} * \text{Support } \{B\}}$$

Interpretation:

- Often considered a quality measure for a rule. Lift > 1 is how much better we are predicting {B} than just knowing support {B}. Higher values indicate more effective rules in predicting RHS
- Lift = 1 suggests independence between {A} and {B}
- Lift is very similar correlation (general measure of association between two itemsets). It doesn't change if you swap the LHS / RHS

Calculate the Metrics

Rule	Support	Confidence	Lift
A => D	2/5 =.4	2/3	(2/5)/ (3/5)*(3/5) = 1.11
C => A			
A => C			
B & C => D			

Your Turn

Calculate the Metrics

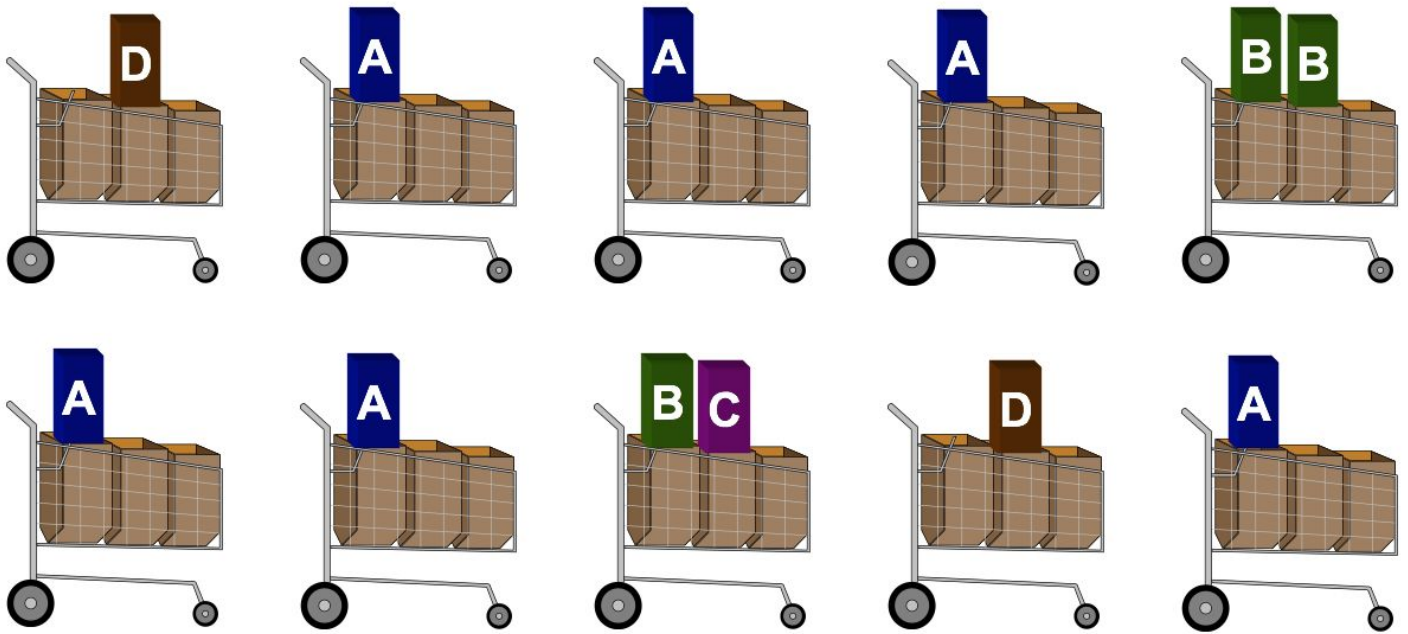
Rule	Support	Confidence	Lift
A => D	2/5 =.4	2/3	(2/5)/ (3/5)*(3/5) = 1.11
C => A	2 / 5	2 / 4	.4 / (4/5)*(3/5) = .833
A => C	2 / 5	2 / 3	.4 / (3/5)*(4/5)=.833
B & C => D	1 / 5	1 / 3	.2 / (3/5)*(3/5) = .555

Thought Experiment:

Stepping back, as an analyst trying to solve a business problem, how can we combine Support / Confidence / Lift in relationship the rules?

Does the data support Market Basket Analysis

We need data with collection of items, if minimal, association rules are probably not a good technique



Some Parting Notes

- Rules need to be actionable (not *very* rare, but this is also debatable)
- Rules are not causation, but help us find patterns of co-occurrence
- Rules can be symmetric, so it often helps to frame the problem in the context of the LHS or RHS
  - I tend to frame relative to the RHS
  - To sell more of the RHS, find the itemsets in the LHS to help build strategies for upsell or cross-promotion

But what does a dataset look like?

The transactional datasets can vary in how they are stored

trans_id	item
1	b
1	c
2	c
2	a
3	c

Single format

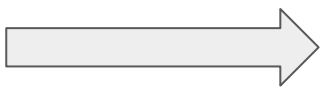
trans_id	items
1	b,c
2	c,a
3	c

Basket format

Luckily, python/pandas make it really easy to modify the origin source data to fit the libraries expected format

The mlxtend format

trans_id	item
1	b
1	c
2	c
2	a
3	c



	a	b	c
1	False	True	True
2	True	False	True
3	False	False	True

What do you notice about the dataset on the right?

Let’s write some python!