This week

Follow this schedule

- Assessment-A Q&A

Do Not jump ahead

1. Intro to this week
2. Association Rule Analysis – Notes
3. Demo 1
4. Python Library – How to install
5. Demo 2 – Explanation -> Exercise to complete

Follow me

Step by Step

- Assessment-B - Discussion

Do Not jump ahead

TU 257 – Fundamentals of Data Science

Data Analytics

L9 – Association Rule Analysis

Brendan Tierney

Agenda

- Application Areas
- Market Basket Analysis
- A Simple Example
- A Big Search Space Problem
- Frequent Item Sets
- Apriori Algorithm
- How the Apriori Algorithms Example
- Evaluating & Filtering Rules (Support, Confidence & Lift)
- Data Privacy Issues

Machine Learning

- the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

Data Mining/Machine Learning

Supervised ML

Classification

Regression

Naïve Bayes
Logistic Regression
Decision Trees
Gradient Boosting
Random Forest
Support Vector Machine
K-nearest neighbors
Neural Networks
Regression
AutoML

UnSupervised ML

Clustering

Kmeans
Hierarchical Clustering
DBSCAN Clustering

Association Rules

Apriori
Fp-growth

Machine Learning

- the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

Data Mining/Machine Learning

Supervised ML

Classification

Regression

Naïve Bayes
Logistic Regression
Decision Trees
Gradient Boosting
Random Forest
Support Vector Machine
K-nearest neighbors
Neural Networks
Regression
AutoML

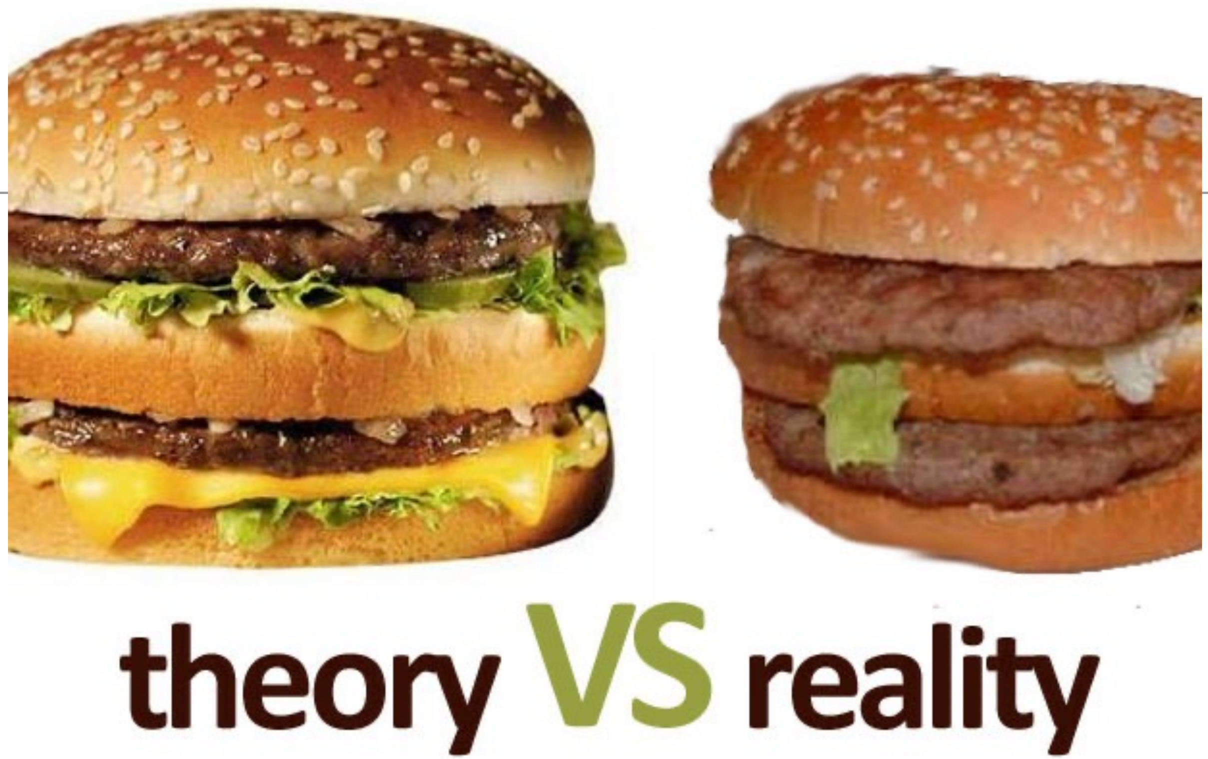
UnSupervised ML

Clustering

Kmeans
Hierarchical Clustering
DBSCAN Clustering

Association Rules

Apriori
Fp-growth



- Most Data Analytics etc can be done in a few lines of code

- Don’t worry about the Theory, we might touch upon some of it, but it isn’t necessary to know in-depth

- You’ll never have to write an algorithm from scratch

Domain Knowledge

- Built up over time
- Experience
- Can take many years or decades
- In-depth knowledge of what is happening and why
- Can be based on Gut Feelings

Domain Knowledge

Domain Knowledge

- Built up over time
- Experience
- Can take many years or decades
- In-depth knowledge of what is happening and why
- Can be based on Gut Feelings

vs

- But what about the DATA
- What is the DATA telling you

Exercise

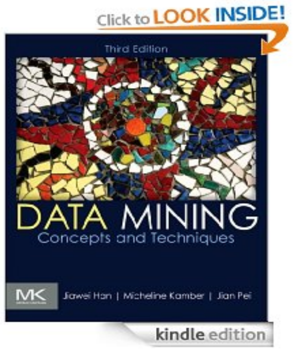
- Think about 2 items you regularly buy in your local shop, supermarket, convenience store, etc?
  - What are they?
  - What can you tell about their placement in the shop?
  - Why are they placed like that?

amazon.co.uk Brendan's Amazon | Today's Deals | Gift Cards | Sell | Help

Shop by Department ▾ Search All ▾ data mining

Kindle Store Buy A Kindle Kindle Books Kindle Singles Kindle Daily Deals Newsstand Accessories Refurbished Kindle Discussions Manage Your Kindle Kindle Support

Start reading *Data Mining: Concepts and Techniques: Concepts and Techni...* on your Kindle **in under a minute**. Don't have a Kindle? Get your Kindl



Share your own customer images

Data Mining: Concepts and Techniques: Concepts and Techniques (The Morgan Kaufmann Series in Data Man

Jiawei Han (Author), Micheline Kamber (Author), Jian Pei (Author)

(1 customer review)

Print List Price: ~~£45.99~~
Kindle Price: £28.42 includes VAT* & free wireless delivery via ***Amazon Whispernet***
You Save: £17.57 (38%)

* Unlike print books, digital books are subject to VAT.

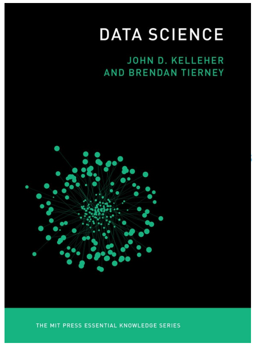
- Length: 626 pages (Contains Real Page Numbers)
- Don't have a Kindle? Get your Kindle here or start reading now with a free Kindle Reading App.

Formats	Amazon Price	New from	Used from
Kindle Edition	£28.42	--	--
Hardcover	£38.41	£21.17	£32.31

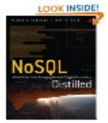
Trade-In Store

Amazon.co.uk Trade-In Store
Did you know you can trade in your unwanted old books for an Amazon.co.uk Gift Card? Plus, until February 23, 2014, get an extra £5 Promotior
Books Trade-In Store for more details. Learn more.

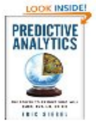
Customers Who Bought This Item Also Bought



Data Mining Techniques: For Marketing, Sales, and Customer Relationship ...
Gordon S. Linoff
Kindle Edition
£20.39



NoSQL Distilled: A Brief Guide to the Emerging ...
› Pramod J. Sadalage
(7)
Kindle Edition
£12.99



Predictive Analytics: The Power to Predict Who ...
› Eric Siegel
(9)
Kindle Edition
£11.45

Seven Databases in Seven Weeks: A Guide to ...
› Eric Redmond
(3)
Kindle Edition
£18.81

Python for Data Analysis: Data Wrangling with ...
› Wes McKinney
(5)
Kindle Edition
£12.67

The Data Warehouse Tool The Complete ...
› Ralph Kimball
(14)
Kindle Edition
£27.81


TESCO
CLUBCARD
Every little helps
634000 0200000000000
MRS J SMITH


£2.00
CLUBCARD
£1.50
In-store • Online •
£1.50


TESCO
25 extra points
when you spend £1 or more on confectionery
25 extra points
when you buy Mr Kipling lemon slices or angel slices
KIPLING
25 extra points
when you spend £1 or more on fresh savoury snack items
50p off
when you buy any Spam
SPAM
25 extra points
when you buy any wine
MumsMonkey2012


DON'T FORGET TO USE YOUR COUPONS
VALID FOR A LIMITED TIME ONLY!
TESCO | Every little helps
CP72217
25p off
AINSLEY HARRIOTT
Use your Clubcard with this coupon
Valid until 13/03/11
CP72403
£1 off
when you spend £5 or more on celebration products
Includes cards, wrap, party bags, balloons & table decorations
TESCO Every little helps
Use your Clubcard with this coupon
Valid until 13/03/11
CP72351
Finest
Use your Clubcard with this coupon
Valid until 13/03/11
CP72413
£2.50 off
when you spend £10 or more on any kids' clothing
Includes back to school
TESCO Every little helps
Use your Clubcard with this coupon
Valid until 13/03/11
CP72668
25 extra points
when you spend £1 or more on condiments
Use your Clubcard with this coupon
Valid until 13/03/11
CP72125
50 bonus Airmiles
when you exchange a minimum of £15 in Clubcard Vouchers for Airmiles in one transaction at tesco.com/rewards.
Exchange by 31/03/11.
Visit www.airmiles.co.uk/rewards for more information
See overleaf for Terms & Conditions
AIR MILES
Make your money fly
Valid until 31/03/11

What Is Association Mining?

- Association rule mining:
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories

Frequent Pattern: A pattern (set of items, sequence, etc.) that occurs frequently in a database

Examples of Application Areas

- Beer and Nappies Example. (google this story)

- Others include
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis
  - Web log (click stream) analysis, DNA sequence analysis, etc

Another Example

"A bank's marketing department is interested in examining associations between various retail banking services used by customers. They would like to determine both typical and atypical service combinations"

- The **BANK** data set has over 32,000 rows coming from 8,000 customers. Each row of the data set represents a customer-service combination. Therefore, a single customer can have multiple rows in the data set, each row representing one of the products he or she owns. The median number of products per customer is three

Name	Model Role	Measurement Level	Description
ACCOUNT	ID	Nominal	Account Number
SERVICE	Target	Nominal	Type of Service
VISIT	Sequence	Ordinal	Order of Product Purchase

Case Study

- The 13 products are represented in the data set as follows:
  - ATM automated teller machine debit card
  - AUTO automobile installment loan
  - CCRD credit card
  - CD certificate of deposit
  - CKCRD check/debit card
  - CKING checking account
  - HMEQLC home equity line of credit
  - IRA individual retirement account
  - MMDA money market deposit account
  - MTG mortgage
  - PLOAN personal/consumer instalment loan
  - SVG saving account
  - TRUST personal trust account

Case Study (cont…)

- Rules generated by analysis
  - ATM automated teller machine debit card
  - AUTO automobile installment loan
  - CCRD credit card
  - CD certificate of deposit
  - CKCRD check/debit card
  - CKING checking account
  - HMEQLC home equity line of credit
  - IRA individual retirement account
  - MMDA money market deposit account
  - MTG mortgage
  - PLOAN personal/consumer installment loan
  - SVG saving account
  - TRUST personal trust account

Rule description

MAP	Rule
RULE1	CKING & CCRD ==> CKCRD
RULE2	CKCRD ==> CKING & CCRD
RULE3	CKCRD ==> CCRD
RULE4	CKING & CKCRD ==> CCRD
RULE5	CCRD ==> CKCRD
RULE6	CCRD ==> CKING & CKCRD
RULE7	HMEQLC ==> CKING & CCRD
RULE8	CKING & CCRD ==> HMEQLC
RULE9	HMEQLC ==> CCRD
RULE10	HMEQLC & CKING ==> CCRD
RULE11	CCRD ==> HMEQLC
RULE12	CCRD ==> HMEQLC & CKING
RULE13	SVG & HMEQLC ==> CKING & ATM
RULE14	CKING & ATM ==> SVG & HMEQLC
RULE15	HMEQLC ==> SVG & CKING & ATM
RULE16	SVG & CKING & ATM ==> HMEQLC
RULE17	SVG & ATM ==> HMEQLC
RULE18	SVG & ATM ==> HMEQLC & CKING
RULE19	HMEQLC ==> SVG & ATM
RULE20	HMEQLC & CKING ==> SVG & ATM
RULE21	HMEQLC ==> CKING & ATM
RULE22	CKING & ATM ==> HMEQLC
RULE23	SVG & HMEQLC ==> ATM
RULE24	SVG & HMEQLC & CKING ==> ATM
RULE25	ATM ==> SVG & HMEQLC
RULE26	ATM ==> SVG & HMEQLC & CKING
RULE27	CD & ATM ==> SVG & CKING
RULE28	HMEQLC ==> ATM

What are the most interesting findings ?

Case Study (cont…)

- The most interesting findings from the analysis included:
  - The strongest rule is checking, and credit card implies check card.
    - This is not surprising given that many check cards include credit card logos
  - It appears that customers with auto loans typically have checking and savings accounts (and are ATM users),
  - but do not utilize other services (at least with sufficient support and confidence to be included in the presented analysis)
- Patterns in the data are discovered. Your job is to put a meaning on the patterns.
  - This isn't always possible

Association Rule: Basic Concepts

- Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)

- Find: all rules that correlate the presence of one set of items with that of another set of items
    - E.g., *98% of people who purchase tires and auto accessories also get automotive services done*

- Applications
    - *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales)
    - *Home Electronics* (What other products should the store stocks up?)
    - Attached mailing in direct marketing
    - Detecting “ping-pong”ing of patients, faulty “collisions”

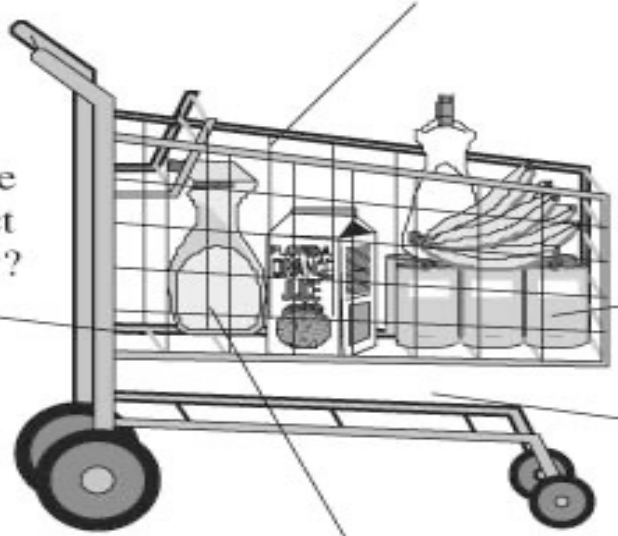
Market Basket Analysis

- **Market basket analysis** is a typical example of frequent itemset mining
- Customers buying habits are divined by finding associations between different items that customers place in their “shopping baskets”
- This information can be used to develop marketing strategies
- One basket tells you about what one customer purchased at one time
- A loyalty cards make it possible to tie together purchases by a single customer (or household) over time

In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.

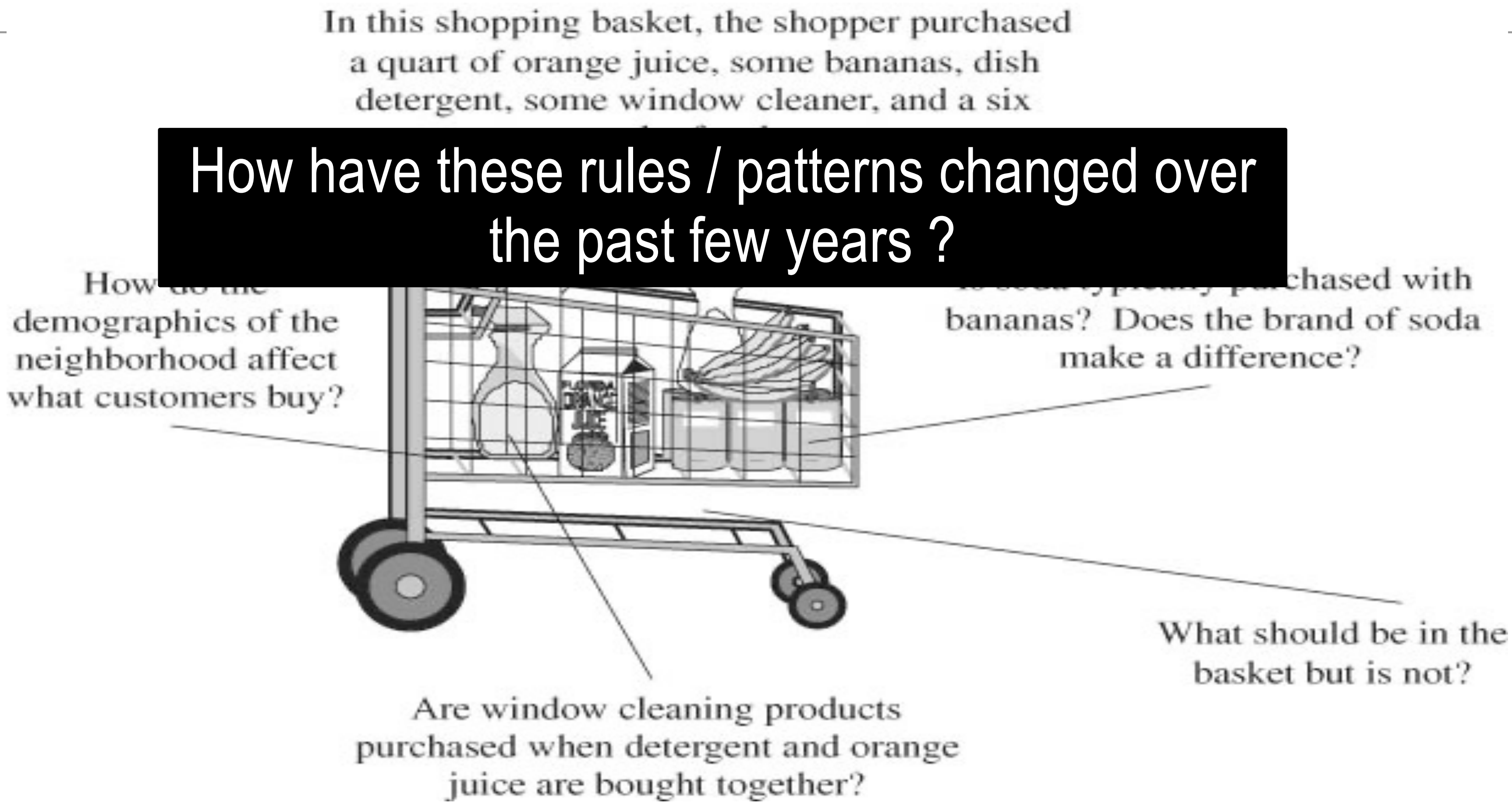
How do the demographics of the neighborhood affect what customers buy?

Is soda typically purchased with bananas? Does the brand of soda make a difference?



Are window cleaning products purchased when detergent and orange juice are bought together?

What should be in the basket but is not?



How have these rules / patterns changed over the past few years ?

- **more than just the contents of shopping carts**

- It is also about what customers do not purchase, and why.
- If customers purchase baking powder, but no flour, what are they baking?
- If customers purchase a mobile phone, but no case, are you missing an opportunity?
- Are they a drug dealer?

- It is also about key drivers of purchases; for example, the gourmet mustard that seems to lie on a shelf collecting dust until a customer buys that particular brand of special gourmet mustard in a shopping excursion that includes hundreds of dollars' worth of other products. Would eliminating the mustard (to replace it with a better-selling item) threaten the entire customer relationship?

Data has a Value

In the USA

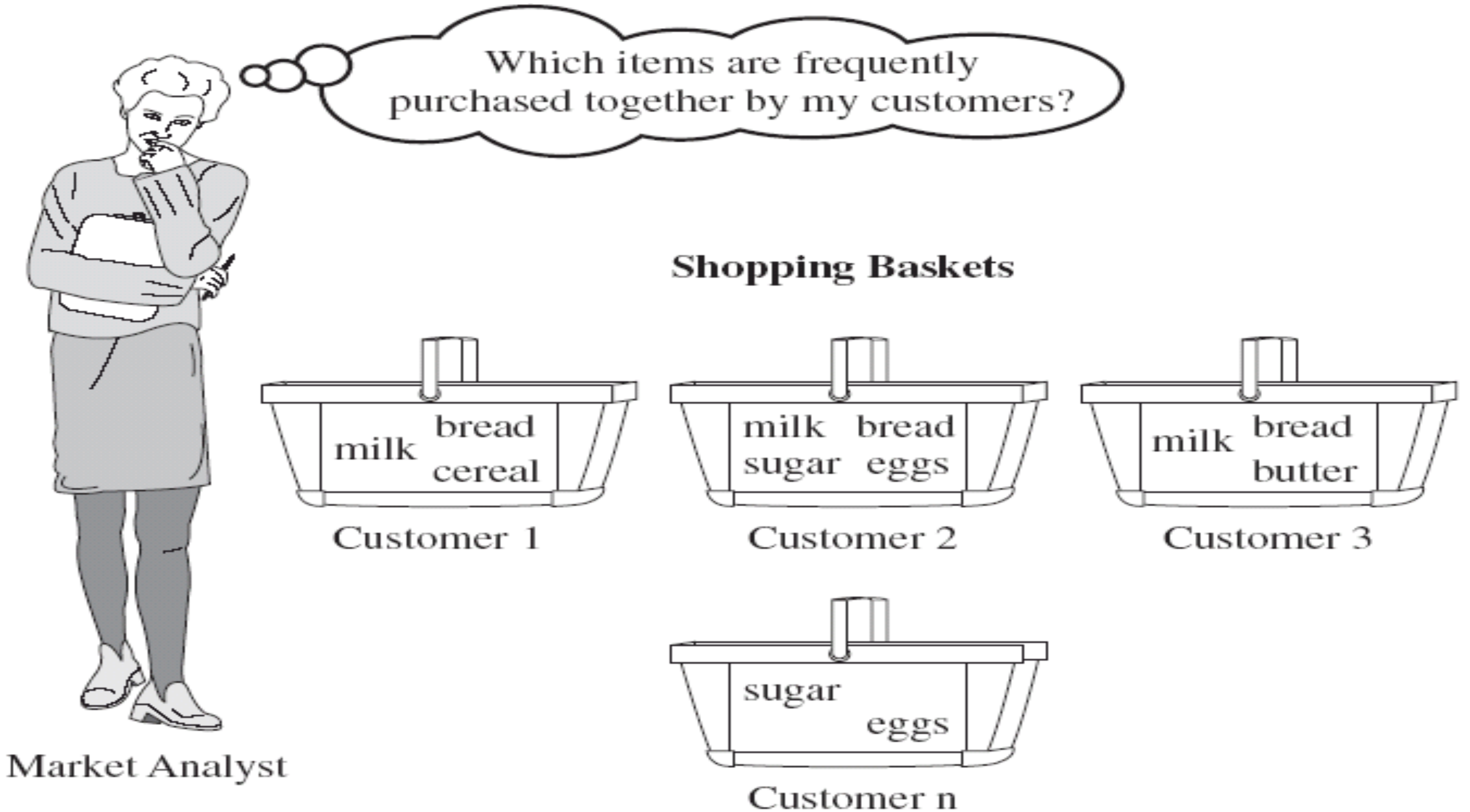
Some retails stores are making more money out of your data

(using it, selling it, etc.)

than they are through their normal day-to-day business

Be careful what you agree too when signing up to “Special Offers”

Quick Exercise – What are the most commonly bought combination of products



Association Rules: The Problem of Lots of Data

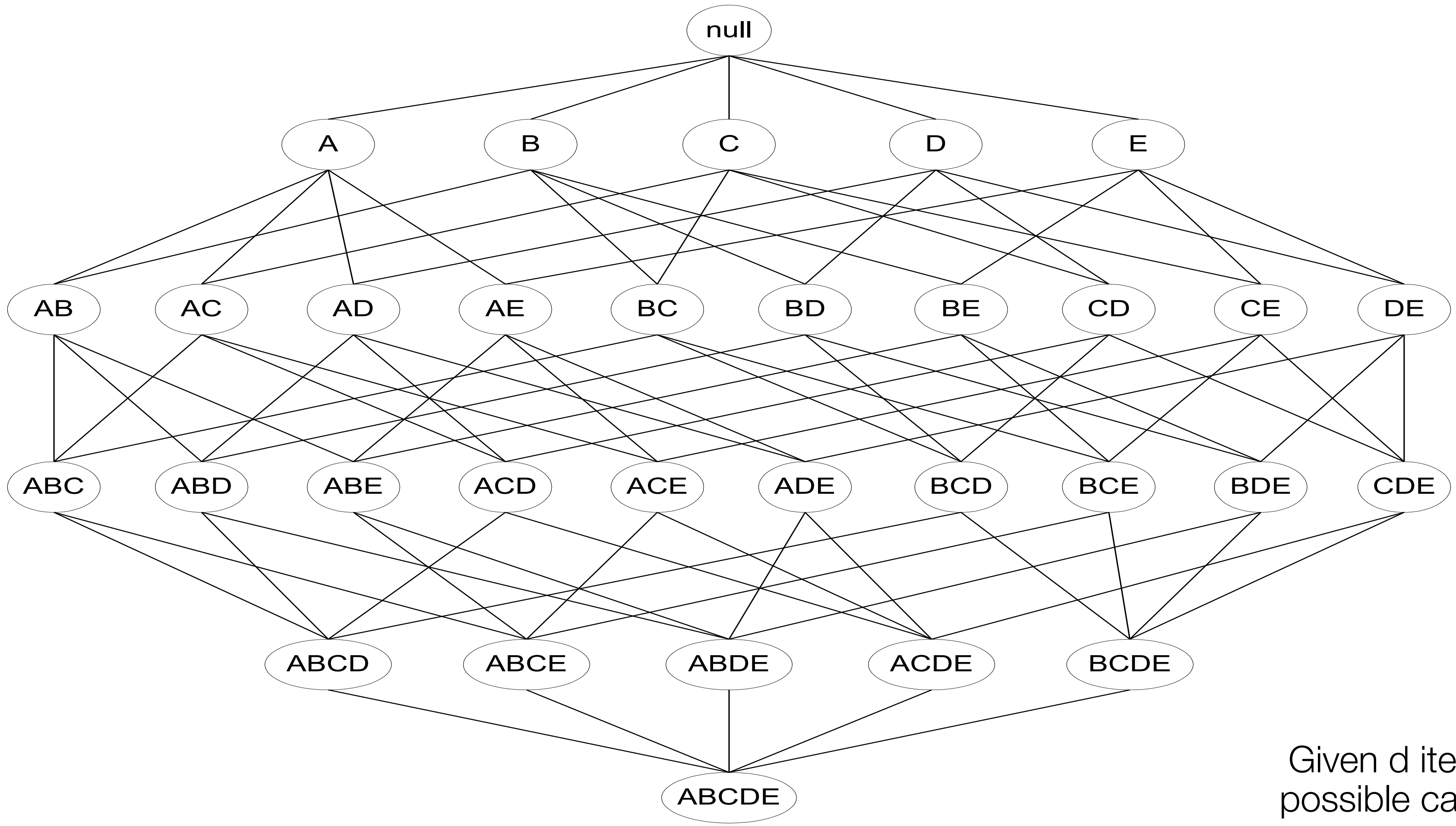
- Fast Food Restaurant…could have 100 items on its menu
  - How many combinations are there with 3 different menu items?
  - 161,700 !
- Supermarket…10,000 or more unique items
  - 50 million 2-item combinations
  - 100 billion 3-item combinations
- Use of product hierarchies (groupings) helps address this common issue
- Finally, know that the number of transactions in a given time-period could also be huge (hence expensive to analyse)

Try writing some SQL queries to find frequently items sets for these situations!

Itemsets & Frequent Itemsets

- An **itemset** is a set of items
- A $\boldsymbol{k}$**-itemset** is an itemset that contains k items
- The **occurrence frequency of an itemset** is the number of transactions that contain the itemset
  - This is also known more simply as the **frequency**, **support count** or **count**
- An itemset is said to be **frequent** if the support count satisfies a **minimum support count threshold**
- The set of frequent itemsets is denoted L_k

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Association Rule Mining

- In general association rule mining can be reduced to the following two steps:
  1. Find all frequent itemsets
     - Each itemset will occur at least as frequently as as a minimum support count
  2. Generate strong association rules from the frequent itemsets
     - These rules will satisfy minimum support and confidence measures

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!

The Apriori Algorithm

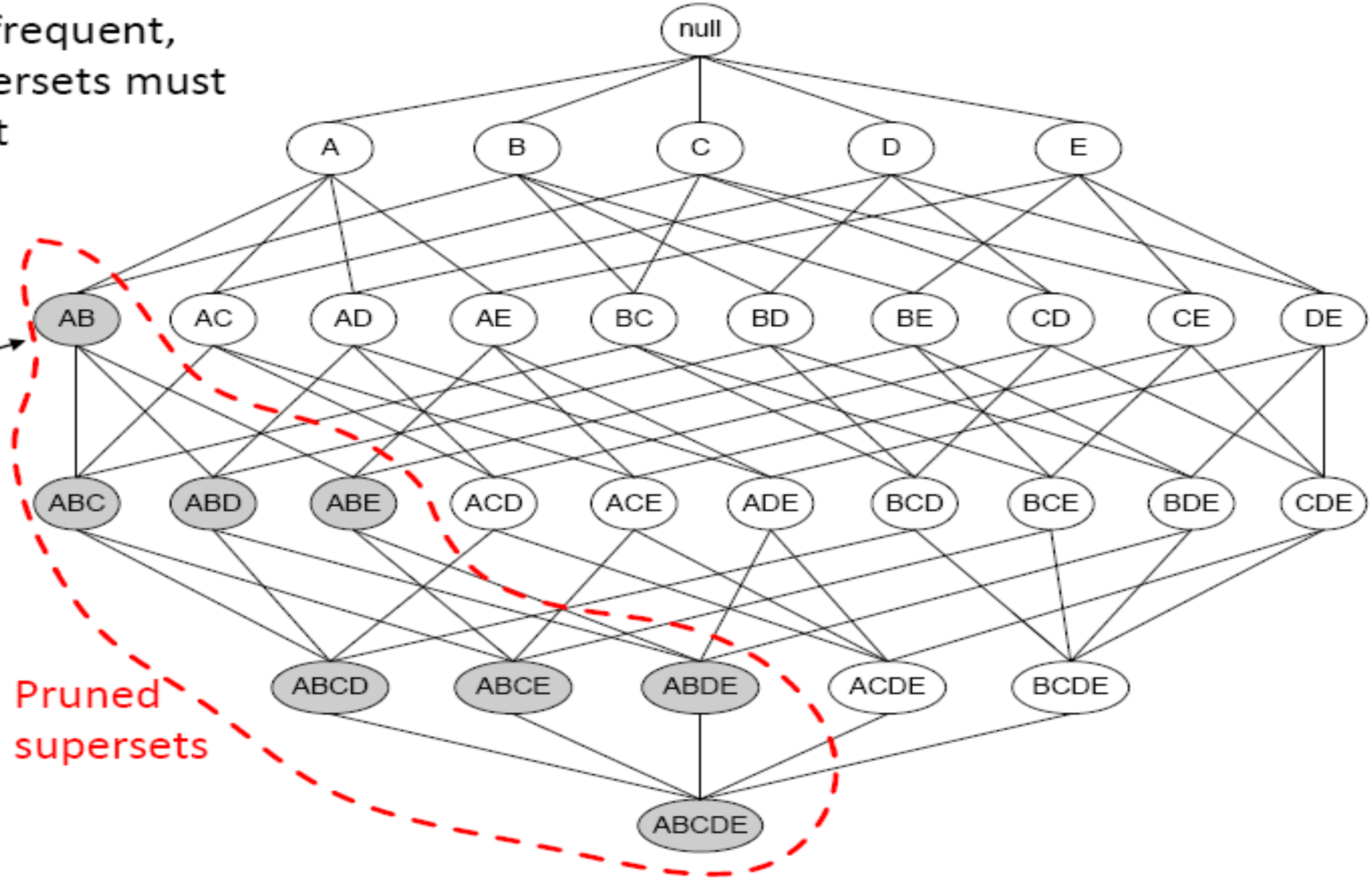
- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code:

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=$\emptyset$; k++) **do begin**
C_{k+1} = candidates generated from L_k;
for each transaction t in database do
increment the count of all candidates in C_{k+1} that are contained in t
L_{k+1} = candidates in C_{k+1} with min_support
end
return $\cup_k L_k$;

Apriori principle for pruning candidates

If an itemset is infrequent, then all of its supersets must also be infrequent



Found to be Infrequent

Generating Association Rules

- Once all frequent itemsets have been found, the association rules can be generated

- Strong association rules from a frequent itemset are generated by calculating the confidence in each possible rule arising from that itemset and testing it against a minimum confidence threshold

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

Another Example

TID	List of item_IDs
T100	Beer, Crisps, Milk
T200	Crisps, Bread
T300	Crisps, Nappies
T400	Beer, Crisps, Bread
T500	Beer, Nappies
T600	Crisps, Nappies
T700	Beer, Nappies
T800	Beer, Crisps, Nappies, Milk
T900	Beer, Crisps, Nappies

ID	Item
I1	Beer
I2	Crisps
I3	Nappies
I4	Bread
I5	Milk

Scan D for count of each candidate →

C_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare candidate support count with minimum support count →

L_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Generate C_2 candidates from L_1 →

C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

Scan D for count of each candidate →

C_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Compare candidate support count with minimum support count →

L_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Generate C_3 candidates from L_2 →

C_3

Itemset
{I1, I2, I3}
{I1, I2, I5}

Scan D for count of each candidate →

C_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate support count with minimum support count →

L_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Association Rule Support & Confidence

- We say that an association rule $A \Rightarrow B$ holds in the transaction set D with **support**, S, and **confidence**, C

- The **support** of the association rule is given as the percentage of transactions in D that contain both A and B (or $A \cup B$)
  - So, the support can be considered the probability $P(A \cup B)$

- The **confidence** of the association rule is given as the percentage of transactions in D containing A that also contain B

- So, the confidence can be considered the conditional probability $P(B \mid A)$

- Association rules that satisfy minimum support and confidence values are said to be **strong**

Support & Confidence Again

- Support and confidence values can be calculated as follows:

$$support(A \Rightarrow B) = P(A \cup B)$$

$$= \frac{support\_count(A \cup B)}{count(\ )}$$

$$confidence(A \Rightarrow B) = P(B \mid A)$$

$$= \frac{support(A \cup B)}{support(A)}$$

$$= \frac{support\_count(A \cup B)}{support\_count(A)}$$

Rule Measures: Support and Confidence

- Find all the rules X & $Y \Rightarrow Z$ with minimum confidence and support
  - support, *s*, probability that a transaction contains $\{X \cup Y \cup Z\}$
  - confidence, *c*, conditional probability that a transaction having $\{X \cup Y\}$ also contains Z

Transaction ID	Items Bought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

Let minimum support 50%, and minimum confidence 50%, we have

$A \Rightarrow C$ (50%, 66.6%)
$C \Rightarrow A$ (50%, 100%)

Rule Measures: Support and Confidence

- Find all the rules *X & Y* $\Rightarrow$ *Z* with minimum confidence and support
  - support, *s*, probability that a transaction contains {X ∪ Y ∪ Z}
  - confidence, *c,* conditional probability that a transaction having {X ∪ Y} also contains *Z*

Transaction ID	Items Bought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

Let minimum support 50%, and minimum confidence 50%, we have

$A \Rightarrow C$ (50%, 66.6%)
$C \Rightarrow A$ (50%, 100%)

Rule Measures: Support and Confidence

- Find all the rules *X & Y* ⇒ *Z* with minimum confidence and support
  - support, *s*, probability that a transaction contains {X ⍰ Y ⍰ Z}
  - confidence, *c*, conditional probability that a transaction having {X ⍰ Y} also contains *Z*

Transaction ID	Items Bought
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

Let minimum support 50%, and minimum confidence 50%, we have

$A \Rightarrow C$ (50%, 66.6%)
$C \Rightarrow A$ (50%, 100%)

Mining Association Rules: An Example

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

$$support(A \Rightarrow C) = \frac{support\_count(\{A\} \cup \{C\})}{count()}$$

$$= 50\%$$

$$confidence(A \Rightarrow C) = \frac{support\_count(\{A\} \cup \{C\})}{support\_count(\{A\})}$$

$$= 66.7\%$$

The Apriori principle:

Any subset of a frequent itemset must be frequent

Mining Association Rules: An Example

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Frequent pattern	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

$$support(C \Rightarrow A) = \frac{support\_count(\{C\} \cup \{A\})}{count()}$$

$$= 50\%$$

$$confidence(C \Rightarrow A) = \frac{support\_count(\{C\} \cup \{A\})}{support\_count(\{C\})}$$

$$= 100\%$$

Support

- The rule X ⇒ Y holds with supports if *s*% of transactions in D contain X∪Y.

TID	Items	Support = Occurence / Total Support
1	ABC	
2	ABD	Total Support = 5
3	BC	Support {AB} = 2 / 5 = 40%
4	AC	Support {BC} = 3 / 5 = 60%
5	BCD	Support {ABC} = 1 / 5 = 20%

- Rules that have a *S* greater than a user-specified support is said to have minimum support.

- **Support:** Support of a rule is a measure of how frequently the items involved in it occur together. Using probability notation: support (A implies B) = P(A, B).

Confidence

- The rule X⇒Y holds with confidence *c* if c% of the transactions in D that contain X also contain Y.

TID	Items
1	ABC
2	ABD
3	BC
4	AC
5	BCD

Given $X \Rightarrow Y$

Confidence = Occurrence {Y} / Occurrence {X}

Confidence $\{A \Rightarrow B\} = 2 / 3 = 66\%$

Confidence $\{B \Rightarrow C\} = 3 / 4 = 75\%$

Confidence $\{AB \Rightarrow C\} = 1 / 2 = 50\%$

- Rules that have a *C* greater than a user-specified confidence is said to have minimum confidence.

- **Confidence:** Confidence of a rule is the conditional probability of B given A. Using probability notation: confidence (A implies B) = P (B given A).

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

Lift

- Lift
  - Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support. It provides information about the improvement, the increase in probability of the consequent given the antecedent. Lift is defined as follows.
    - (Rule Support) /(Support(Antecedent) * Support(Consequent))
    - This can also be defined as the confidence of the combination of items divided by the support of the consequent.
- Example
  - Convenience store customers who buy orange juice also buy milk with a 75% confidence.
  - The combination of milk and orange juice has a support of 30%.

  - This at first sounds like an excellent rule, and in most cases, it would be. It has high confidence and high support. However, what if convenience store customers in general buy milk 90% of the time? In that case, orange juice customers are actually less likely to buy milk than customers in general.

  - in our milk example, assuming that 40% of the customers buy orange juice, the improvement would be:
  - 30% / (40% * 90%) = 0.83 – an improvement of less than 1.
  - Any rule with an improvement of less than 1 does not indicate a real cross-selling opportunity, no matter how high its support and confidence, because it actually offers less ability to predict a purchase than does random chance.

  - If lift > 1, then items are positively correlated
  - lift < 1, then negatively correlated
  - lift = 1, then are independent

Sequence Databases and Sequential Pattern Analysis

- Frequent patterns vs. (frequent) sequential patterns

- Applications of sequential pattern mining
  - Customer shopping sequences:
    - First buy computer, then a Printer, and then Head Set, within 3 months.
  - Medical treatment
  - Natural disasters (e.g., earthquakes)
  - Science & engineering processes,
  - Stocks Markets
  - Telephone calling patterns,
  - Weblog click streams
  - DNA sequences and gene structures
  - …

But

What about privacy issues?

Does this matter?

Forbes

TECH | 2/16/2012 @ 11:02AM | 2,337,246 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

334 comments, 173 called-out + Comment Now + Follow Comments

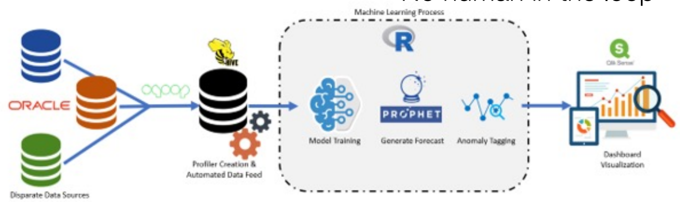
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

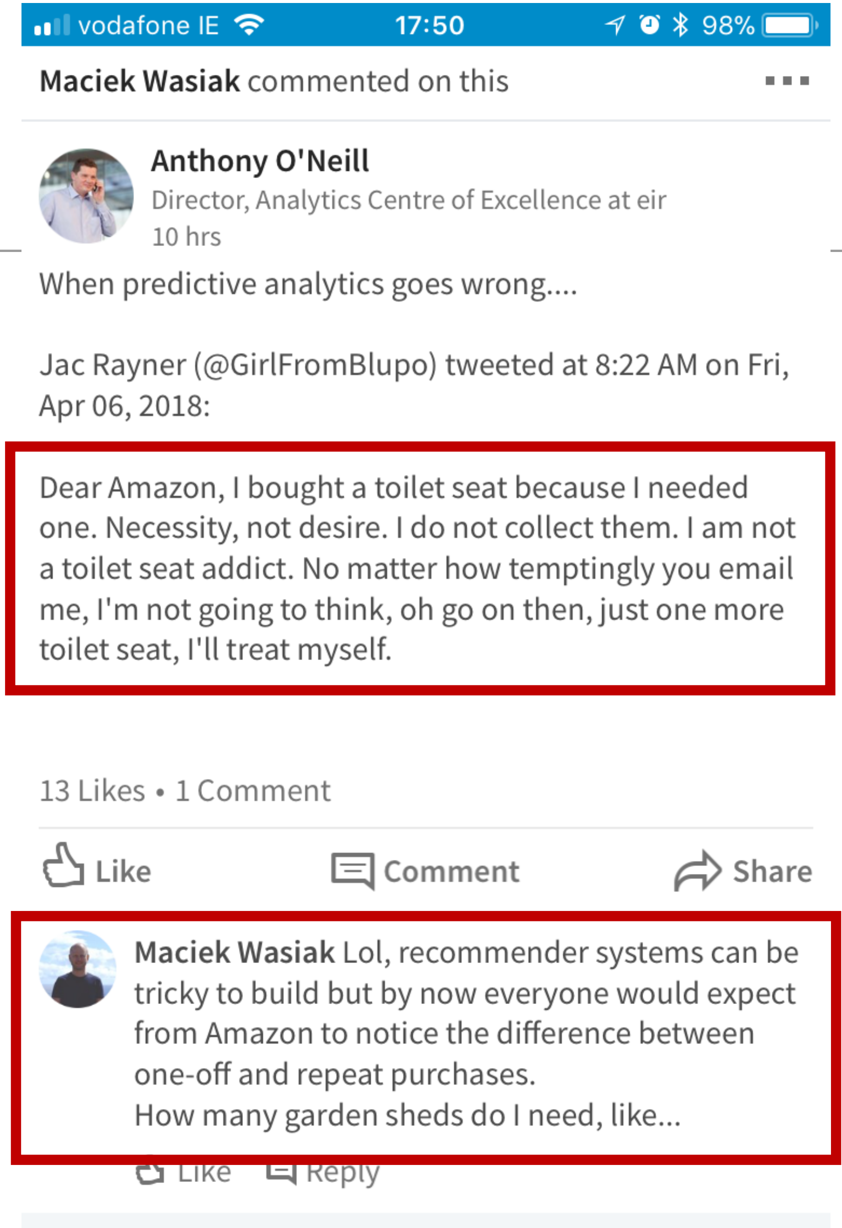
Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

Automated Pattern Discovery

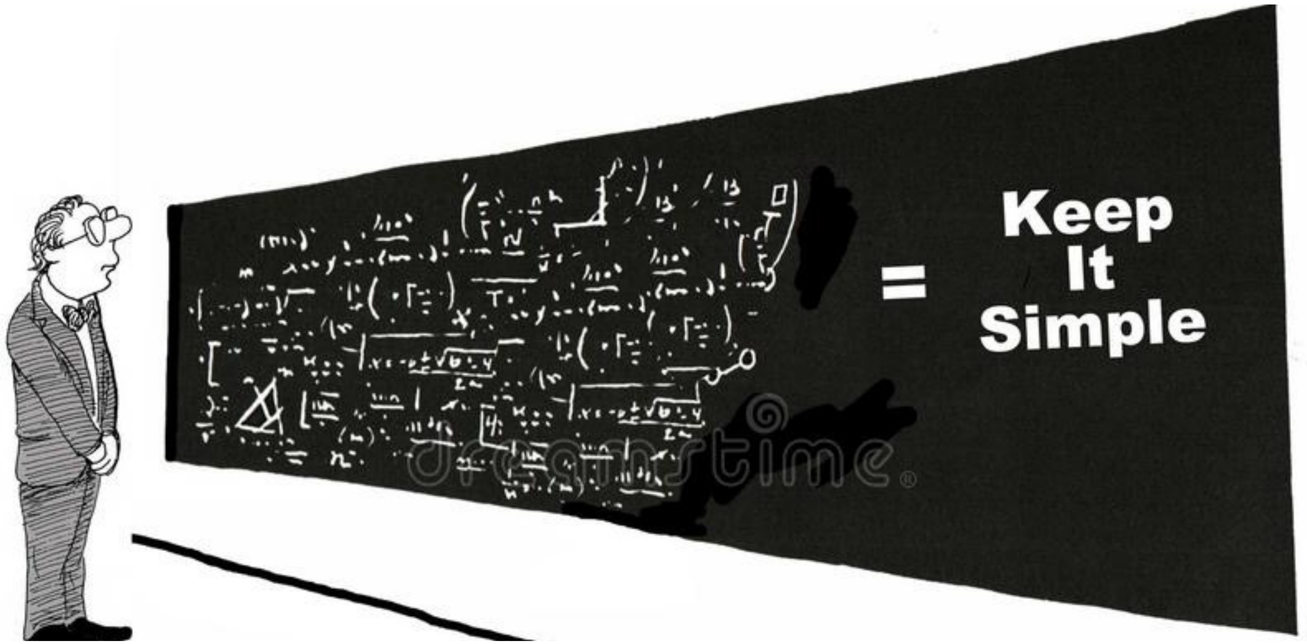
No human in the loop



https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=7385e4a16668


vodafone IE
17:50
98%
Maciek Wasiak commented on this
Anthony O'Neill
Director, Analytics Centre of Excellence at eir
10 hrs
When predictive analytics goes wrong....
Jac Rayner (@GirlFromBlupo) tweeted at 8:22 AM on Fri, Apr 06, 2018:
Dear Amazon, I bought a toilet seat because I needed one. Necessity, not desire. I do not collect them. I am not a toilet seat addict. No matter how temptingly you email me, I'm not going to think, oh go on then, just one more toilet seat, I'll treat myself.
13 Likes • 1 Comment
Like
Comment
Share
Maciek Wasiak Lol, recommender systems can be tricky to build but by now everyone would expect from Amazon to notice the difference between one-off and repeat purchases.
How many garden sheds do I need, like...
Like
Reply


= too complex


= Keep It Simple


EASY
HARD


EASY
HARD

Time for an Example

Any Questions ?

What Now/Next ?