COMP9033 DATA ANALYTICS

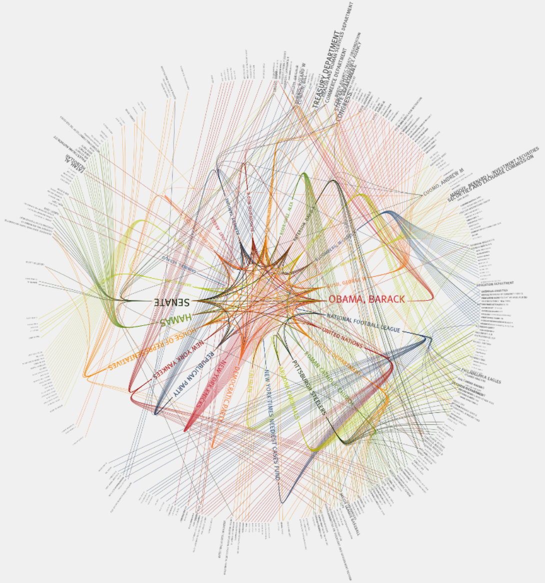
10/12

ASSOCIATION RULE MINING

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2018.04.17



Credit: Jer Thorp

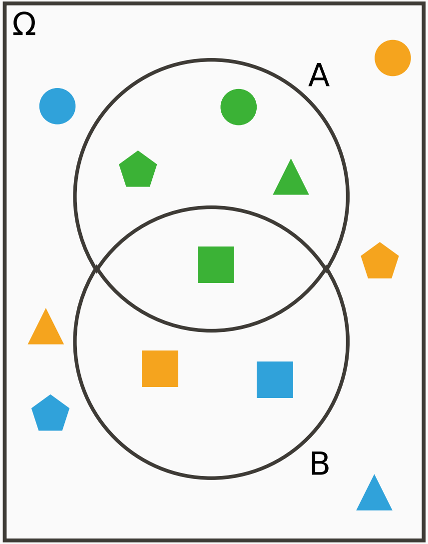
Overview

1. Clustering:
    - What it is.
    - How it is used.
    - Real world examples.
    - Flavours / variants.
2. Combinatorial clustering:
    - Brute force.
    - K-means.
    - K-medoids.
    - Performance considerations.
    - How to choose K.
3. Hierarchical clustering:
    - Divisive vs. agglomerative.
    - Dendrograms.
    - Determining similarity.

1. Sets and probability:
    - Set notation.
    - Set cardinality.
    - Probability.
    - Conditional probability.
    - Examples.
2. Association rule mining:
    - What it is.
    - Terminology.
    - Support, confidence and lift.
    - Brute force mining.
    - Efficient rule mining.

Sets and probability

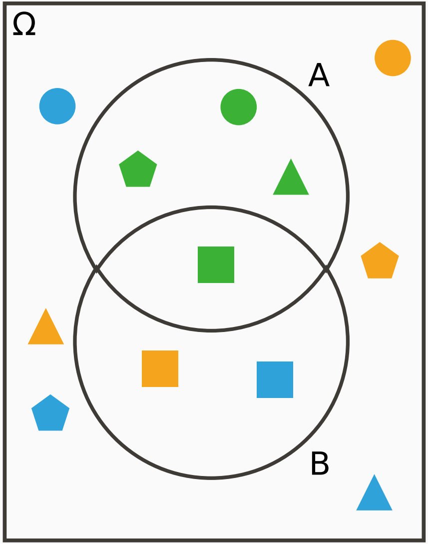
- A *set* is a collection of distinct objects.
- Consider the set diagram to the right:
    - Ω is the set of *all* items.
    - $\emptyset$ is the set of *no* items (*i.e.* the null set).
    - A is the subset of green items ($A \subseteq \Omega$).
    - B is the subset of square items ($B \subseteq \Omega$).
    - $A \cap B$ is the set of green *and* square items.
    - $A \cup B$ is the set of green *or* square items.
- Ω is referred to as the *sample space*.
- $A \cap B$ is referred to as the *intersection* of A and B.
- $A \cup B$ is referred to as the *union* of A and B.



- The *indicator function*, $I_X(x_i)$, indicates whether the data point x_i is a member of the set X, *i.e.*

$$I_X(x_i) = \begin{cases} 1 & \text{if } x_i \in X, \\ 0 & \text{if } x_i \notin X. \end{cases} \tag{10.1}$$

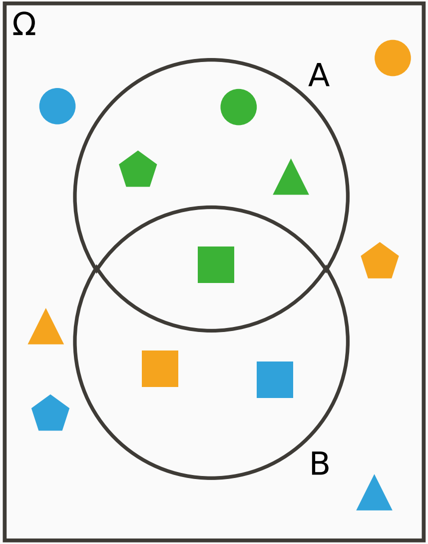
- For instance, in the diagram to the right:
  - $I_A(x_i) = 1$ when x_i corresponds to a green object and is zero otherwise.
  - $I_B(x_i) = 1$ when x_i corresponds to a square object and is zero otherwise.



- We can use the indicator function to describe the size of a set:

$$|A| = \sum_{x_i \in \Omega} I_A(x_i), \tag{10.2}$$

where $|A|$ is number of elements in the set A, also known as its *cardinality*.



- For instance, in the diagram to the right:

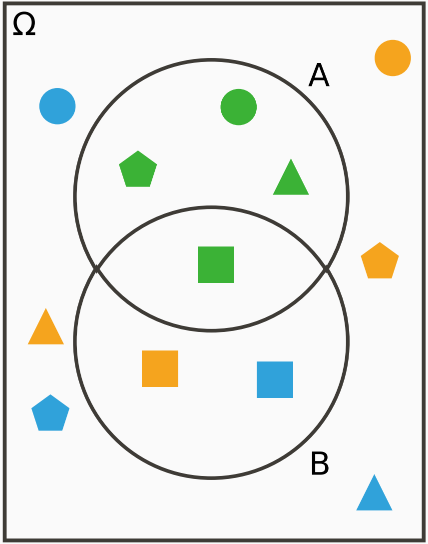
$$|\Omega| = 12,$$

$$|A| = 4,$$

$$|B| = 3,$$

$$|A \cap B| = 1,$$

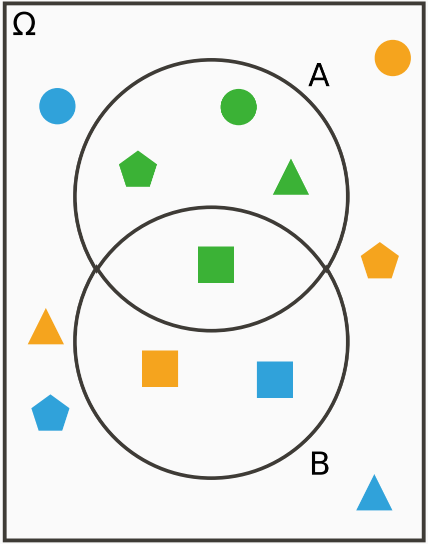
$$|A \cup B| = 6.$$



- In the case where each outcome in the sample space is equally likely to occur, we can state that

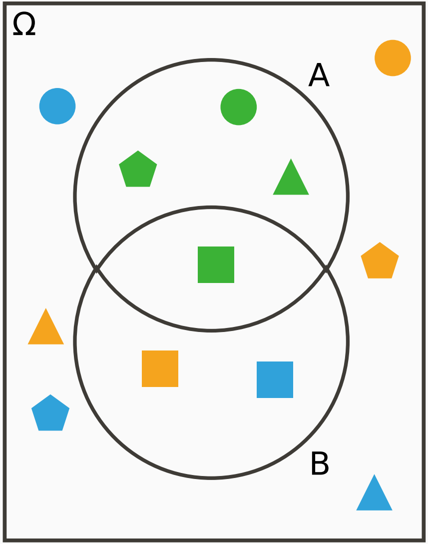
$$P(X) = \frac{|X|}{|\Omega|}, \tag{10.3}$$

where $P(X)$ is the probability of the outcome X.



- For example, let's say we put all the shapes from the diagram opposite into a bag and drew one at random.
- The probability of drawing a green square would be

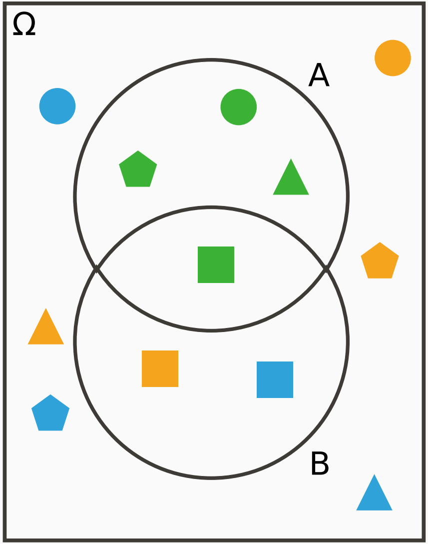
$$P(A \cap B) = \frac{|A \cap B|}{|\Omega|}$$
$$= \frac{1}{12}.$$



- In the case where each outcome in the sample space is equally likely to occur, we can also state that

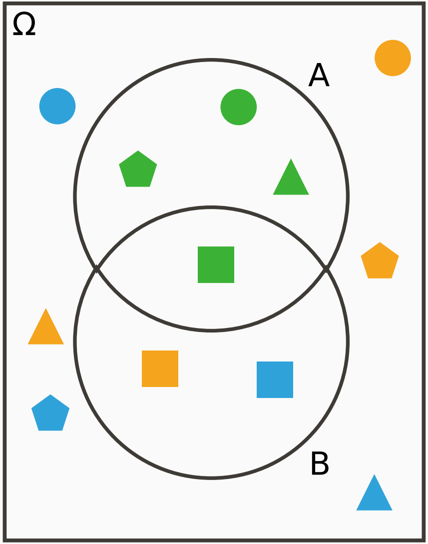
$$P(Y|X) = \frac{|X \cap Y|}{|X|}, \tag{10.4}$$

where $P(Y|X)$ represents the probability of Y occurring, given that X has already occurred.



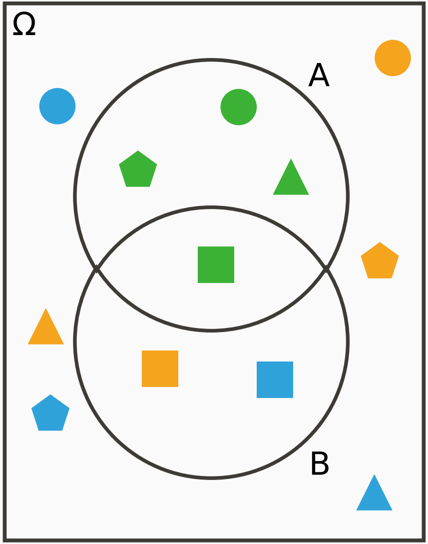
- For example, let's say we returned the green square from the previous example to our bag of shapes and drew another at random.
- On drawing the new shape, we see that it is green.
- At this point, the probability of it also being square is given by

$$P(B|A) = \frac{|A \cap B|}{|A|}$$
$$= \frac{1}{4}.$$



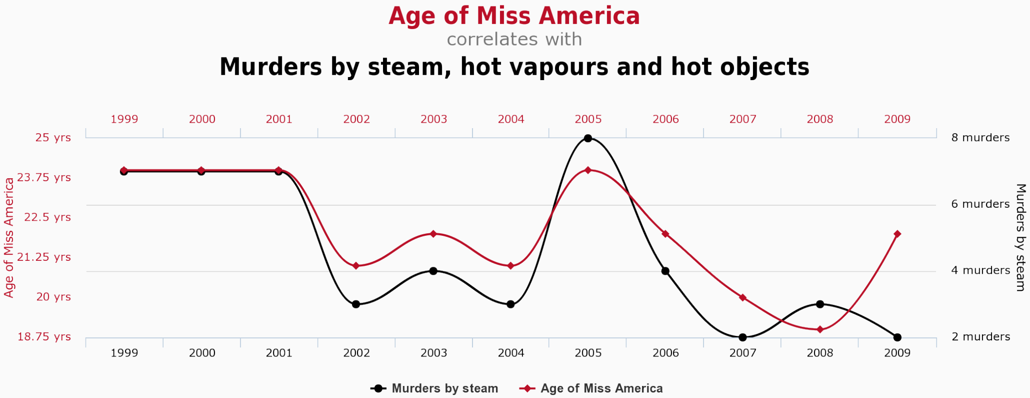
- In general, $P(Y|X)$ is not equal to $P(X|Y)$.
- For instance, we can compute the probability that, if we picked a square item, it would be green, as

$$P(A|B) = \frac{|A \cap B|}{|B|}$$
$$= \frac{1}{3}.$$



Association rule mining

- Association rule mining is an *unsupervised* machine learning technique that can be used to find hidden relationships in *categorical* data.
- Associations are typically formed by examining the co-occurrence of items or events: when *X* occurs, *usually Y* also occurs, *e.g.*
  - When a customer buys tea, they usually also buy biscuits.
  - When a user clicks on link A, they usually also click on link B.
  - When a server fails, monitoring software usually detects anomalies on its failover.
- Associations that occur frequently may be of interest as they could indicate a previously unknown relationship.
- However, associations can also occur by random chance!



Credit: Tyler Vigen

- Association rules are representations of relationships between items or groups of items.
- Typically, they are written in the form $X \Rightarrow Y$, where $X = \{x_1, x_2, \ldots, x_n\}$ represents the *antecedent* (*i.e.* source) item set and $Y = \{y_1, y_2, \ldots, y_m\}$ represents the *consequent* (*i.e.* target) item set, *e.g.*
    - $\{\text{Star Wars}\} \Rightarrow \{\text{Indiana Jones}\}$
    - $\{\text{tea}, \text{milk}\} \Rightarrow \{\text{biscuits}\}$
    - $\{\text{California}, \text{Hawaii}\} \Rightarrow \{\text{Pacific}, \text{USA}\}$
- One thing to note is that the antecedent items and the consequent items must be distinct sets, *i.e.* there cannot be any shared items ($X \cap Y = \emptyset$).

- The terminology in association rule mining differs slightly from that in general machine learning:
  - Instead of features, we have *items*.
  - Instead of samples, we have *transactions*.
- A transaction may involve zero, one or many items.
- Generally, we want to discover relationships between items, across transactions.

items $\longrightarrow$

transactions $\downarrow$

	A	B	C	D
1	1	0	1	0
2	0	1	1	1
3	1	0	0	1
4	1	1	1	1

- Generally, the occurrence of an item in a transaction is represented as in the table opposite:
    - The presence of an item is represented as a 1.
    - The absence of an item is represented as a 0.
- Note that representing the presence of items only means we lose information about the *quantity* of items!

items $\longrightarrow$

transactions $\downarrow$

	A	B	C	D
1	1	0	1	0
2	0	1	1	1
3	1	0	0	1
4	1	1	1	1

- The *transaction count* of an item set is defined as the number of transactions that the item set appears in.
- More formally, we can write this as

$$\text{count}(X) = |\{t_i \mid X \subseteq t_i, t_i \in T\}|, \tag{10.5}$$

  where t_i corresponds to the set of items present in the i^{th} transaction and $T = \{t_1, t_2, \ldots, t_n\}$ is the set of all transactions.
- The right hand side of Equation 10.5 can be read as "*the number of elements in the set of transactions that include the item set X as a subset*".

TRANSACTION	BREAD	MILK	DIAPERS	BEER	EGGS	BUTTER
1	1	0	0	0	1	0
2	0	0	1	1	0	1
3	0	1	1	1	1	0
4	1	1	1	1	0	1
5	1	1	1	0	1	0

Q. Given the data above, compute the transaction count of the item set $\{\text{milk}, \text{diapers}\}$.

A. Using Equation 10.5, we can write

$$\begin{aligned}\text{count}(\{\text{milk}, \text{diapers}\}) &= |\{3, 4, 5\}| \\ &= 3.\end{aligned}$$

- The *support* of an association rule is a measure of how often its relationship occurs in the set of transactions.
- In the set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, the support of the rule $X \Rightarrow Y$ is given by

$$\begin{aligned} \text{support}(X \Rightarrow Y) &= \frac{\text{count}(X \cup Y)}{|T|} \\ &= \frac{\text{count}(X \cup Y)}{n}, \end{aligned} \tag{10.6}$$

i.e. the support of $X \Rightarrow Y$ is the ratio of the number of transactions that include the items in both X *and* Y to the total number of transactions.

- Using the definition in Equation 10.3, we can infer that support is an approximate measure of the probability that X and Y occur together in transactions *in general*.
- If the transactions are a representative sample of the true relationship between X and Y, then this approximation becomes more accurate as the number of transactions grows larger.
- If the support of a rule is very low, then we can infer that the probability of observing the relationship it describes is small:
    - Infrequently occurring relationships may not be “interesting”.
    - Very infrequently occurring relationships may be the result of chance.

TRANSACTION	BREAD	MILK	DIAPERS	BEER	EGGS	BUTTER
1	1	0	0	0	1	0
2	0	0	1	1	0	1
3	0	1	1	1	1	0
4	1	1	1	1	0	1
5	1	1	1	0	1	0

Q. Given the transaction history data above, compute the support of the rule $\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}$.

A. We can compute support using Equation 10.6, *i.e.*

$$\begin{aligned} \text{support}(\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}) &= \frac{\text{count}(\{\text{milk}, \text{diapers}, \text{beer}\})}{n} \\ &= \frac{2}{5}. \end{aligned}$$

- The *confidence* of an association rule is a measure of its strength.
- The confidence of the rule $X \Rightarrow Y$, is given by

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{count}(X \cup Y)}{\text{count}(X)}, \tag{10.7}$$

i.e. the confidence of $X \Rightarrow Y$ is the ratio of the number of transactions that include the items in both X *and* Y to the number of transactions that include the items in X *only*.

- Using the definition in Equation 10.4, we can infer that the confidence of the rule $X \Rightarrow Y$ is an approximate measure of the probability of observing Y in a transaction given that we have already observed X.
- As with support, the approximation becomes more accurate as the number of transactions grows larger (assuming that the transactions represent the relationship between X and Y well).
- If the confidence of a rule is high, then we can infer that the probability of observing Y after observing X is high.
- However, this *not* does imply that X causes Y — correlation does not imply causation!

TRANSACTION	BREAD	MILK	DIAPERS	BEER	EGGS	BUTTER
1	1	0	0	0	1	0
2	0	0	1	1	0	1
3	0	1	1	1	1	0
4	1	1	1	1	0	1
5	1	1	1	0	1	0

Q. Given the transaction history data above, compute the support, confidence and lift of the rule $\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}$.

A. We can compute confidence using Equation 10.7, *i.e.*

$$\begin{aligned}\text{confidence}(\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}) &= \frac{\text{count}(\{\text{milk}, \text{diapers}, \text{beer}\})}{\text{count}(\{\text{milk}, \text{diapers}\})} \\ &= \frac{2}{3}.\end{aligned}$$

- The *lift* of an association rule is a measure of its strength relative to the frequency of the consequent item set.
- The lift of the rule $X \Rightarrow Y$, is given by

$$\mathrm{lift}(X \Rightarrow Y) = n \times \frac{\mathrm{count}(X \cup Y)}{\mathrm{count}(X) \times \mathrm{count}(Y)}, \tag{10.8}$$

  where $n = |T|$, *i.e.* the total number of transactions.
- Lift is a similar measure to confidence, but takes the transaction counts of both X and Y into account.

- Using the definition in Equation 10.3, we can infer that the lift of the rule $X \Rightarrow Y$ is an approximate measure of the relative probability that X and Y occur together to the probability that X and Y occur independently of one together.
- As before, the approximation grows more accurate as the number of transactions increases.
- If $\text{lift}(X \Rightarrow Y) = 1$, then we can infer that X and Y occur independently of one another
- If $\text{lift}(X \Rightarrow Y) > 1$, then we can infer X and Y are dependent on one another: transactions with X are more likely to have Y and vice-versa.
- If $\text{lift}(X \Rightarrow Y) < 1$, then we can infer X and Y are anti-dependent on one another: transactions with X are less likely to have Y and vice-versa.

TRANSACTION	BREAD	MILK	DIAPERS	BEER	EGGS	BUTTER
1	1	0	0	0	1	0
2	0	0	1	1	0	1
3	0	1	1	1	1	0
4	1	1	1	1	0	1
5	1	1	1	0	1	0

Q. Given the transaction history data above, compute the support, confidence and lift of the rule $\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}$.

A. We can compute lift using Equation 10.8, *i.e.*

$$\begin{aligned}\text{lift}(\{\text{milk}, \text{diapers}\} \Rightarrow \{\text{beer}\}) &= n \times \frac{\text{count}(\{\text{milk}, \text{diapers}, \text{beer}\})}{\text{count}(\{\text{milk}, \text{diapers}\}) \times \text{count}(\{\text{beer}\})} \\ &= 5 \times \frac{2}{3 \times 3} = \frac{10}{9}.\end{aligned}$$

- Association rules can be mined in a brute force manner.
- However, the total number of possible rules that must be evaluated is given by

$$R = 3^k - 2^{k+1} + 1, \tag{10.9}$$

where k is the number of item categories.

- A large number of rules can be formed from even just a small number of item categories, *e.g.*
  - For ten item categories, $R = 57002$.
  - For one hundred item categories, $R \approx 5.15 \times 10^{47}$.
- Consequently, for large numbers of items, a brute force approach is generally not feasible.

- Several efficient algorithms exist for mining association rules, *without* resorting to heuristics, *e.g*. Apriori, FP-Growth, Eclat.
- Generally, efficient rule mining recognises that we're only interested in rules with high support *and* high confidence.
- Consequently, if we determine that a rule has low support, we don't need to evaluate it any further and can discard it.
- This prompts a two stage search:
    1. Frequent item set generation: generate a set of candidate item sets that all meet some predefined support threshold.
    2. Association rule extraction: generate candidate rules and filter them by eliminating those that do not meet a predefined confidence threshold.

- To generate the frequent item sets, the we can take advantage of a set property called *downward closure*.
- The key idea is that, if a rule describes a frequent relationship (*e.g.* $A \Rightarrow \{B, C\}$), then all relationships that are *subsets* of that relationship (*e.g.* $A \Rightarrow B$, $A \Rightarrow C$) must also be frequent.
- Conversely, if a rule describes an infrequent relationship (*e.g.* $C \Rightarrow D$), then all relationships that are *supersets* of that relationship (*e.g.* $C \Rightarrow \{D, E\}$, $C \Rightarrow \{D, F\}$) must also be infrequent.
- Consequently, if we determine that a given item set is infrequent, we can discard all of the rules that depend on supersets of that item set.

- We can apply downward closure in an iterative way to generate the frequent item sets:
    1. The supports of all single item sets are computed, *e.g.* $\{A\}, \{B\}, \{C\}, \{D\}$.
    2. Sets that do not meet a predefined support threshold are discarded, *e.g.* $\{D\}$.
    3. The supports of all two item sets that can be formed from the undiscarded single item sets are computed, *e.g.* $\{A, B\}, \{A, C\}, \{B, C\}$.
    4. Again, sets that do not meet a predefined support threshold are discarded.
    5. Steps 3 and 4 are repeated for increasing set sizes until every possible set size has been evaluated.
- When the process has completed, we are left with a collection of item sets of varying sizes, each of which has exceeded the minimum support threshold we have defined.

- Once we have gathered our frequent item sets, we can then mine them for association rules:
    1. Select a frequent item set, *e.g.* $\{A, B, C\}$.
    2. Partition the frequent item set into every possible combination of rules, *e.g.* $A \Rightarrow B$, $A \Rightarrow C$, $A \Rightarrow \{B, C\}$, $B \Rightarrow A$, $B \Rightarrow C$, $B \Rightarrow \{A, C\}$, $C \Rightarrow A$, $C \Rightarrow B$, $C \Rightarrow \{A, B\}$.
    3. Compute the confidence for each rule and discard those that do not meet a predefined threshold.
    4. Repeat Steps 1-3 for all frequent item sets found previously.
- Once the process finishes, we will have a collection of rules whose support and confidence exceed our predefined thresholds.

2.22 / DRAWBACKS OF SUPPORT FILTERING

- Using a large support threshold helps to reduce the number of item sets / rules we must assess, but can lead us to eliminate rare, valid, and potentially interesting, relationships.
- By eliminating rules with low support, we avoid “relationships” that may have occurred simply by chance.
- However, if an item is rare, then by definition it will have low support, and so filtering will eliminate it, even if the relationship is valid.
- This can lead us to eliminate some interesting relationships accidentally.

Summary

X.1 / SUMMARY

- Association rule mining:
  - Used to find things that occur together frequently.
  - Quality measures: support, confidence and lift.
  - Computational complexity: brute force vs. efficient rule mining.
  - Drawbacks: spurious correlation, quantities are ignored, support filtering.
- Lab work:
  - Mine association rules for grocery transaction data.
  - Compute support.
  - Compute confidence.
- Next lecture: big data systems!

1. Hastie et al. *The Elements of Statistical Learning: Data mining, Inference and Prediction.* 2nd edition, February 2009. (stanford.io/2i1T6fN)
2. Tan et al. *Introduction to Data Mining.* Pearson, 2006. (bit.ly/1Avz7VR)
3. Ullman et al. *Mining of Massive Data Sets.* Cambridge University Press, 2014. (stanford.io/1qtgAYh)