Inspire...Educate...Transform.






Inspire…Educate…Transform.

Methods and Algorithms in Machine Learning

Rules and Rule Induction

Dr. Sridhar Pappu

Executive VP – Academics, INSOFE

January 28, 2017

And the awards go to…

- **#10: CART (34 votes)**
- **#7: Naive Bayes (45)**
- **#7: kNN (45)**
- **#7: AdaBoost (45)**
- **#6: PageRank (46)**
- **#5: EM (Expectation Maximization) (48)**
- **#4: Apriori (52)**
- **#3: SVM (58)**
- **#2: K-Means (60)**
- **#1: C4.5 (61)**

An *if-then* Rule

- If (x) and (y) and (z), then A

  If (Attendance <= 70%) *and* (Cumulative Grade < 50%) *and* (Feedback on Sridhar Pappu < Excellent), *then* (No Certificate)

- x, y, z: Antecedent or Precondition
- A: Consequent or Conclusion

- Length of a rule: Number of antecedents

GOODNESS METRICS TO FILTER RULES – THE STATISTICAL / DATA SCIENTIST PERSPECTIVE

"if CCAvg is medium, then loan = accept"

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

3 in 13. 23% of the data is covered by this rule.

This is called **SUPPORT** or **COVERAGE**.

Support is the % of cases in the data that contain both X and Y. *Recall Joint Probability P(X and Y).*

* Adapted from Universal Bank data on predicting loan purchase likelihood of existing bank customers

"if CCAvg is medium, then loan = accept"

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

Of the three occasions LHS is present, RHS too is present.

This rule has 100% **CONFIDENCE** or **ACCURACY**.

Confidence is the % of cases containing X that also contain Y.

Recall Conditional Probability P(Y|X).

CSE 7405c

“if loan = accept, then CCAvg is medium”

ID	Age	Income	Family	CCAvg	Personal Loan
1	Young	Low	4	Low	0
2	Old	Low	3	Low	0
3	Middle	Low	1	Low	0
4	Middle	Medium	1	Low	0
5	Middle	Low	4	Low	0
6	Middle	Low	4	Low	0
10	Middle	High	1	High	1
17	Middle	Medium	4	Medium	1
19	Old	High	2	High	1
30	Middle	Medium	1	Medium	1
39	Old	Medium	3	Medium	1
43	Young	Medium	4	Low	1
48	Middle	High	4	Low	1

What are SUPPORT and CONFIDENCE now?

Support remains same at 3/13 (23%), but Confidence dips to 3/7 (43%).

CSE 7405c

Findings

- Support remains same if rule is switched
- Confidence changes

Recall P(Y and X) = P(X and Y) but $P(Y|X) \neq P(X|Y)$

CSE 7405c

Let us look at another case

- A transaction table contains 100 records. We want to find rules with Support greater than 20% and Confidence higher than 80%.

CSE 7405c

- Let us say X is present in 25 transactions out of 100 transactions/records. If Y is an obvious class (e.g., does not have cancer or something like that), it is present in, say, all the records. What are Support and Confidence?

  - Support = 25/100 = 25% (both items are present in a total of 25% of transactions)
  - Confidence = 25/25 =100% (if X is present then Y is always present).

Confidence is Not Adequate

- This seems to have a very high confidence.
- But, in reality, Y is present whether or not X is present. So, there is no REAL relationship between the two.

- Clearly, we need another metric. That metric is called **LIFT**.

CSE 7405c

LIFT as a Goodness Metric

- We divide the confidence of Y with the probability of Y.

- The *lift* of a rule $X \Rightarrow Y$ is defined as:
  - LIFT = P(Y|X) / P(Y) = $\frac{P(X \text{ and } Y)}{P(X)P(Y)}$

$$= \frac{\text{Joint Probability of } X \text{ and } Y}{(\text{Marginal Probability of } X) * (\text{Marginal Probability of } Y)}.$$

  - That is, LIFT is the ratio of the confidence to the % of cases containing Y.

CSE 7405c

LIFT as a Goodness Metric – Case 1

- In our example, as Y is there in all transactions, P(Y)=1.
- Since Confidence = 1, **LIFT =1**.

CSE 7405c

LIFT as a Goodness Metric – Case 2

- Total transactions 100. Y occurs in 20 where X also occurred (recall X occurred 25 times) and does not occur elsewhere. What are Support, Confidence and Lift now?
  - Support of X->Y: 20/100 = 20% (where X and Y occur).
  - Confidence of X->Y: 20/25 = 80%.
  - P(Y) = 20/100 = 20%.
  - **Lift** = 0.8/0.2 (confidence/probability of Y) **= 4**

CSE 7405c

LIFT as a Goodness Metric – Case 3

- Y occurs in 20 records where X occurs and 70 transactions where X does not occur. What are Support, Confidence and Lift now?
  - Support of X->Y: 20/100 = 20% (where X and Y occur).
  - Confidence of X->Y: 20/25 = 80%.
  - P(Y) = 90/100 = 90%.
  - **Lift** = 0.8/0.9 (confidence/probability of Y) **= 0.89**

CSE 7405c

Interpreting Lift

- If lift = 1, then X and Y are independent (case 1)
- If lift > 1, then X and Y are positively correlated (case 2)
- If lift < 1, then X are Y are negatively correlated (case 3)

CSE 7405c

APPLICATION OF THE METRICS TO FILTER RULES – THE BUSINESS PERSPECTIVE

CSE 7405c

Defining Minimum Support

- Rules with a certain minimum support alone are important. How do we know that?
  - Domain expertise
  - But, this is subjective

CSE 7405c

Identifying Trivial Rules

- Ask the business user

- A short rule with high support and confidence

$$\frac{Support+Confidence}{Length}$$

- For example, a customer who buys bread also buys butter

Assessing Actionability

- If (the mother is B positive) and (smoked during pregnancy) and (the kid is eating a lot of carbohydrates), then the kid is likely to get asthma

- All three attributes have different actionabilities

CSE 7405c

Analysis of Attributes for Universal Bank to Make Decisions

- **Non-actionable**: Acts of God (weather), external factors (price of gold, rupee value, etc.)
- **Actionable**: Age, experience, income, family, education
- **Actionable and changeable**: Mortgage, mortgage status, average credit card spending and other statistics, usage of other accounts (cc, cd, online & securities), infoReq (information requested by Phone or Email)

CSE 7405c

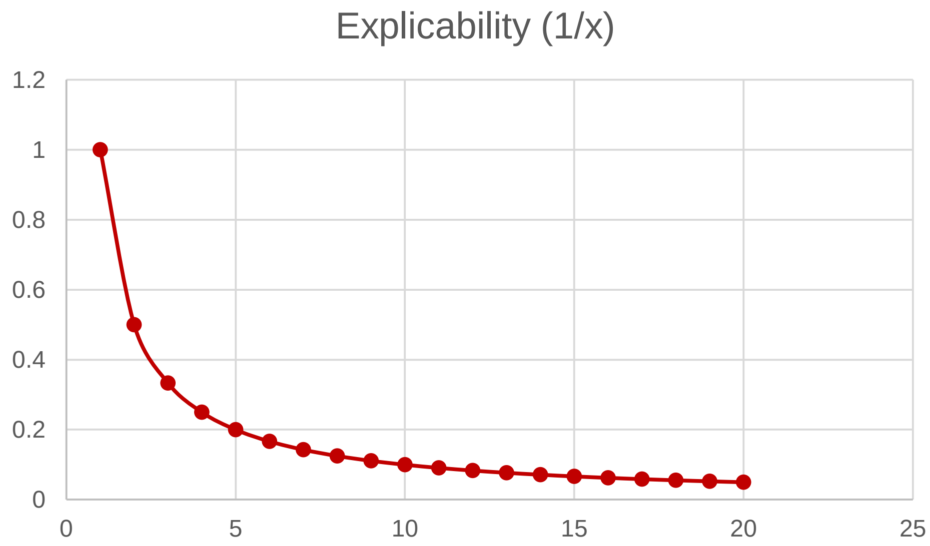
Actionability of a Rule

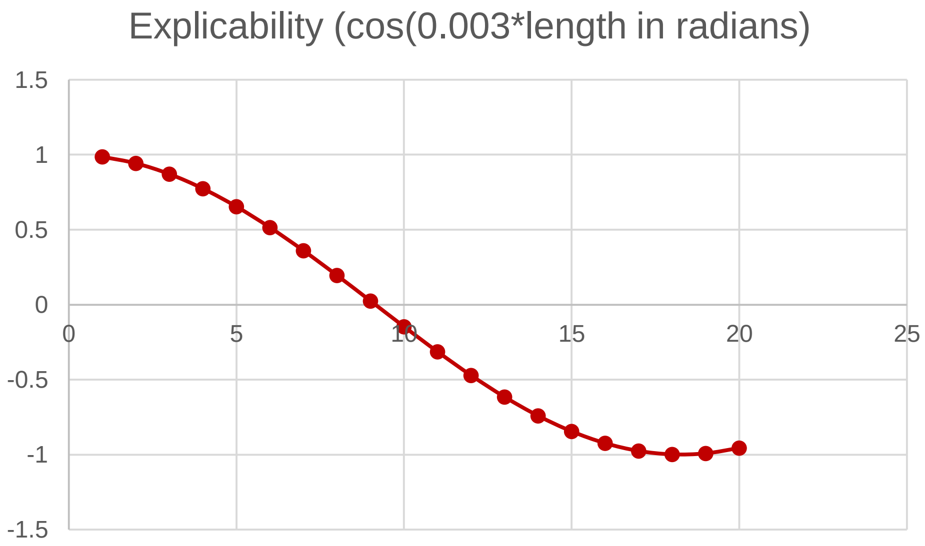
- $Actionability = \dfrac{\sum Actionability\ of\ antecedents}{Total\ number\ of\ attributes\ in\ the\ antecedent}$

- If we take the numerator alone, a long rule and short rule with same actionability come out as equals

Explicability

- More precedents, less explicable





CSE 7405c

Cost of a Rule

- Cost of a rule is the sum of the costs of collection of each attribute

CSE 7405c

Trying out What-If scenarios

- **Generalizing:** removing an attribute

- What does it do to support?
- What does it do to confidence?

CSE 7405c

Trying out What-If scenarios

- **Specifizing:** Adding an attribute

- What does it do to support?
- What does it do to confidence?

CSE 7405c

PROPOSITIONAL RULE INDUCTION

CSE 7405c

PRISM Algorithm

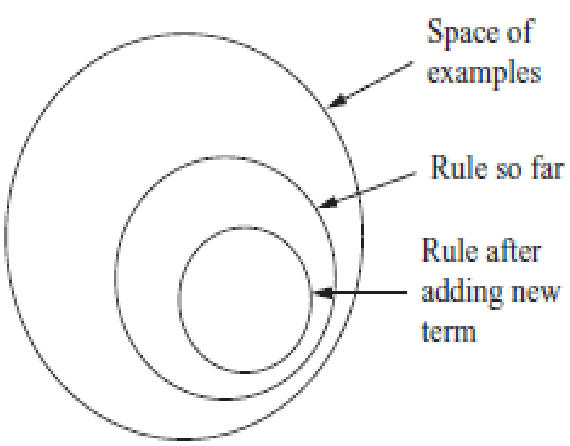
- Operate by adding tests to the rule that is under construction, always striving to create a rule with maximum accuracy/confidence.

- Involves finding an attribute to split on. Then chooses an attribute–value pair to maximize the probability of the desired classification.

- Suppose the new rule covers a total of t instances, of which p are positive examples of the class and t–p are in other classes—that is, they are errors made by the rule.

- Then choose the new term to maximize the ratio p/t.

- Stop when $p/t = 1$ OR the set of instances/examples cannot be split further.



PRISM Algorithm

For each class *C*
    Initialize to the set of all examples ***E***
    While ***E*** contains examples in class *C*
        Create a rule ***R*** with an empty left-hand side that predicts class *C*
        Until ***R*** is 100% accurate (or there are no more attributes to use) do:
            For each attribute ***A*** not in ***R***, and each value ***v***
                Consider adding the condition (attribute-value pair) ***A=v***
                to the left hand side of ***R***
                Sellect ***A*** and ***v*** to maximize the accuracy and covering of the
                attribute-value pair
            Add ***A=v*** to ***R***
        Remove the examples covered by ***R*** from ***E***

CSE 7405c

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

If ?, then recommendation = hard

For the unknown term ?, we have nine choices:

Attribute	Value	p/t
Age	young	2/8
Age	pre-presbyopic	1/8
Age	presbyopic	1/8
Spectacle prescription	myope	3/12
Spectacle prescription	hypermetrope	1/12
Astigmatism	no	0/12
Astigmatism	yes	4/12
Tear production rate	reduced	0/12
Tear production rate	normal	4/12

For each class *C*
  Initialize to the set of all examples *E*
  While *E* contains examples in class *C*
    Create a rule *R* with an empty left-hand side that predicts class *C*
    Until *R* is 100% accurate (or there are no more attributes to use) do:
      For each attribute *A* not in *R*, and each value *v*
        Consider adding the condition (attribute-value pair) *A*=*v* to the left hand side of *R*
        Sellect *A* and *v* to maximize the accuracy and covering of the attribute-value pair
      Add *A*=*v* to *R*
    Remove the examples covered by *R* from *E*

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

astigmatism = yes

```
For each class C
    Initialize to the set of all examples E
    While E contains examples in class C
        Create a rule R with an empty left-hand side that predicts class C
        Until R is 100% accurate (or there are no more attributes to use) do:
            For each attribute A not in R, and each value v
                Consider adding the condition (attribute-value pair) A=v
                to the left hand side of R
                Sellect A and v to maximize the accuracy and covering of the
                attribute-value pair
            Add A=v to R
        Remove the examples covered by R from E
```

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

CSE 7405c

If astigmatism = yes and ?, then recommendation = hard

Attribute	Value	p/t
Age	young	2/4
Age	pre-presbyopic	1/4
Age	presbyopic	1/4
Spectacle prescription	myope	3/6
Spectacle prescription	hypermetrope	1/6
Tear production rate	reduced	0/6
Tear production rate	normal	4/6

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

CSE 7405c

tear production rate = normal

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	normal	none

CSE 7405c

If astigmatism = yes and tear production rate = normal and ?, then recommendation = hard

Attribute	Value	p/t
Age	young	2/2
Age	pre-presbyopic	1/2
Age	presbyopic	1/2
Spectacle prescription	myope	3/3
Spectacle prescription	hypermetrope	1/3

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	yes	normal	none

If astigmatism = yes and tear production rate = normal and spectacle prescription = myope, then recommendation = hard

- Only covers three out of the four hard recommendations.
- Delete these three from the set of instances
- Start again, looking for another rule of the form:

If ?, then recommendation = hard

CSE 7405c

Other Rules from the Contact Lens Dataset

```
IF    TearProduction = reduced
THEN  ContactLenses = none        [#soft=0 #hard=0 #none=12]

IF    TearProduction = normal
  AND Astigmatism = no
THEN  ContactLenses = soft        [#soft=5 #hard=0 #none=1]

IF    TearProduction = normal
  AND Astigmatism = yes
  AND SpectaclePrescription = myope
THEN  ContactLenses = hard        [#soft=0 #hard=3 #none=0]

IF    TearProduction = normal
  AND Astigmatism = yes
  AND SpectaclePrescription = hypermetrope
THEN  ContactLenses = none        [#soft=0 #hard=1 #none=2]
```

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Propositional Rule Induction

- Produces compact/understandable knowledge
- Can find every possible pattern and hence overfit
- They are slow to induce
- It is a separate-and-conquer process

Advances in Rule Induction

- Rule induction is greedy search
- Beam search: Instead of one, pick *n* candidates at every stage
- Other methods are available too

Unsupervised rule induction

ASSOCIATION RULES – AFFINITY ANALYSIS / MARKET BASKET ANALYSIS

- Popularized by the 1993 paper* by Agrawal *et al.* on finding regularities based on POS transactions in supermarkets.

- Market basket analysis **doesn't refer** to a single technique.

- Useful for cross-selling, up-selling, influencing sales promotions, loyalty programs, store layouts, discount plans, intrusion detection, bioinformatics, and many more applications.

**Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.*

CAN WE REALLY GET INSIGHTS FROM MARKET BASKETS

CSE 7405c

CSE 7405c

405c

It is not over yet

- Most likely he/she is a vegetarian!
- He/she has been exposed to some overseas culture (how many Indians eat pickles, not the Indian pickles!)

Market Basket Analysis

- Provides insight into **which products** tend to be **purchased together** and which are most amenable to **promotion**.

- The findings were that men between 30- 40 years in age, shopping between 5PM and 7PM on Fridays, who purchased diapers, were most likely to also have beer in their carts.

Market Basket Analysis

- Suppose the POS system has the following data:
  - Total transactions = 600,000
  - Transactions containing diapers = 7,500 (1.25%)
  - Transactions containing beer – 60,000 (10%)
  - Transactions containing both beer and diapers = 6,000 (1%)
- Assuming (null hypothesis) that beer and diaper purchases are independent (no association), and knowing that 10% of all transactions contain beer, 10% of the transactions containing diapers should be EXPECTED to contain beer.

Market Basket Analysis

- 10% of 7500 = 750. However, 6000 transactions containing diapers contained beer, which is an 8-fold **increase over the expected value**. So, the LIFT is 8.

$$\text{Recall, } Lift = \frac{P(X \text{ and } Y)}{P(X)P(Y)} = \frac{0.01}{0.0125 * 0.1} = 8$$

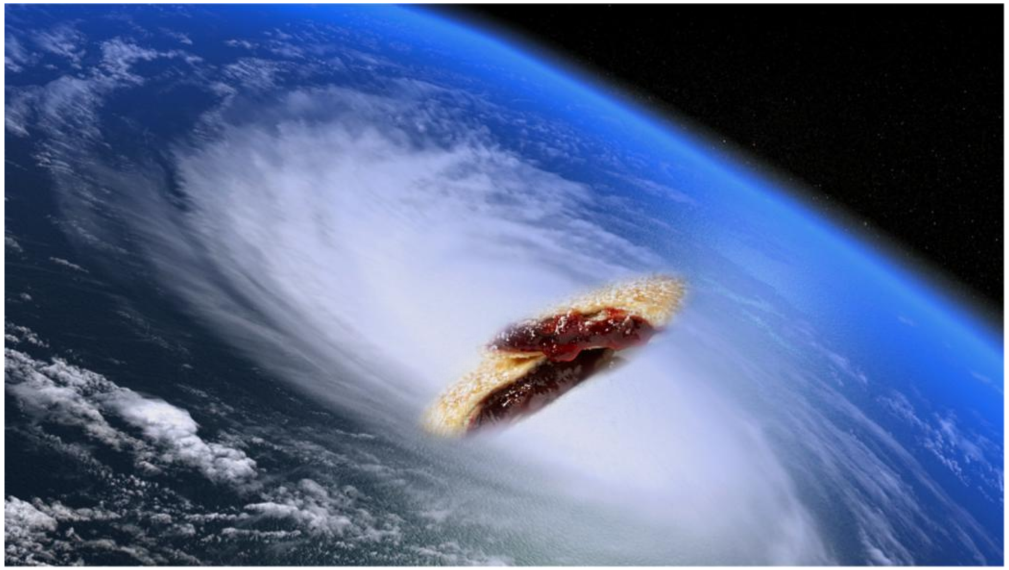
CSE 7405c

Market Basket can give Rules that are

- Seemingly interesting, actually not
  - People buying conference calling facility also buy call forwarding service
- Trivial
  - People who buy shoes also buy socks
- Inexplicable
  - People who buy shirts also buy milk
- Actionable
  - People who bought Dove soap also bought Barbie doll. What should the business do?
  - Target store case

Market Basket can lead to interesting discoveries

Wal-Mart in Florida found in 2004 that **strawberry pop tart** sales before a hurricane had a lift of 7 over normal shopping days.



Market Basket can lead to interesting discoveries

A major electronics store used association rule mining to find that customers who bought VHS players/recorders tended to return 3-4 months later to buy camcorders. They use discount coupons to successfully engage such customers.

3-4 months

CSE 7405c

It is not just for retail and baskets

- Unusual combinations of **insurance claims** can be a sign of fraud and can spark further investigation.

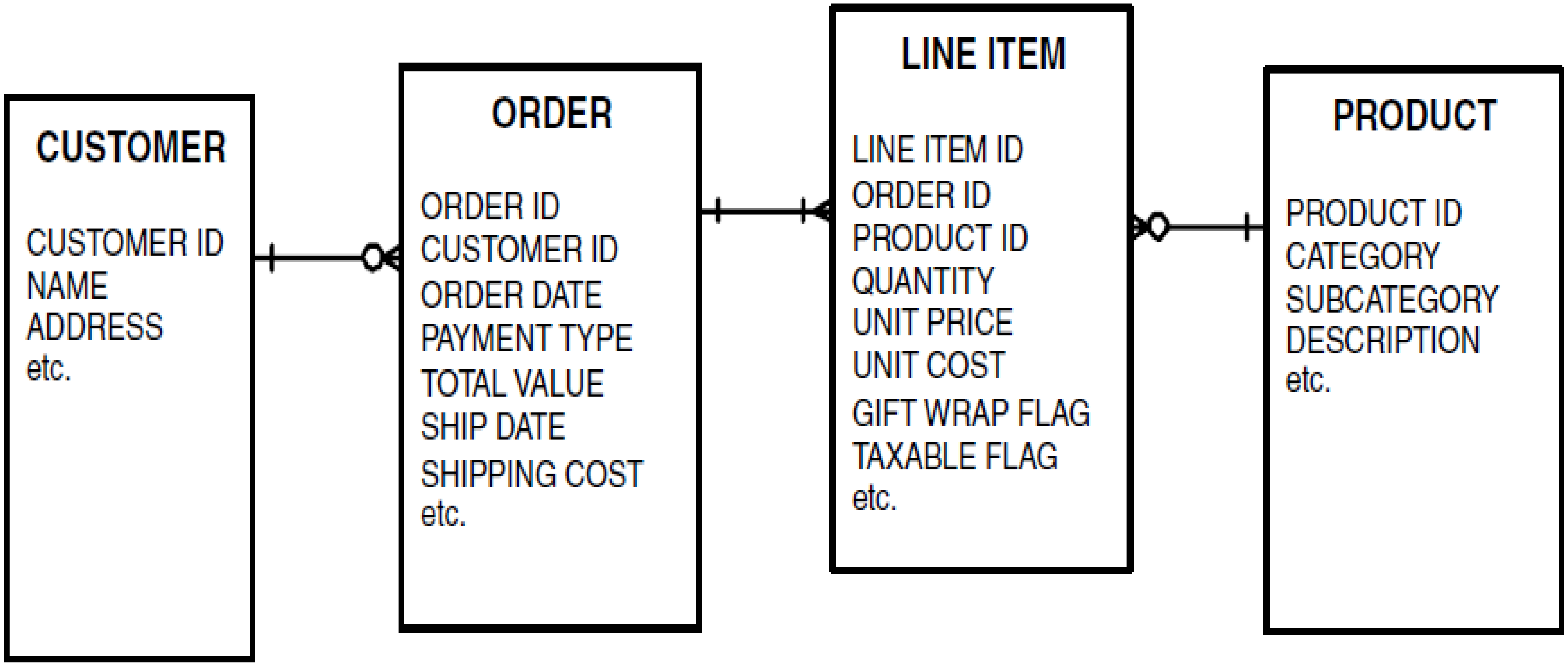
- Medical patient histories can give indications of likely complications based on certain **combinations** of **treatments**.

CSE 7405c

- Supervised or Unsupervised?

CSE 7405c

Let us hand work some rules



CSE 7405c

Some straightforward Market Basket insights

- What is the average number of orders per customer?
- What is the most common item found in a one-item order?
- What is the average number of unique items per order?
- What is the average number of items per order?

The Process – Step 1: Transform the data

ORDER ID	LINE ITEM ID	B
ORDER ID	LINE ITEM ID	C

	PRODUCT A	PRODUCT B	PRODUCT C	PRODUCT D
ORDER ID	0	1	1	0

CSE 7405c

The Process – Step 2: Co-occurrence table

	Product A	Product B	Product C	Product D
Product A				
Product B				
Product C				
Product D				

CSE 7405c

Line item table

ID	Order ID	Product ID	Quantity
1	1	1	2
2	1	2	1
3	2	3	3
4	2	1	2
5	2	4	1
6	3	1	2
7	3	5	3
8	4	1	1
9	4	5	1
10	4	2	2
11	5	2	2
12	5	4	3

Product table

ID	Product
1	Orange juice
2	Soda
3	Milk
4	Window cleaner
5	Detergent

Order ID	Products
1	Orange juice, Soda
2	Milk, orange juice, window cleaner
3	Orange juice, detergent
4	Orange juice, detergent, soda
5	Window cleaner, soda

Co-occurrence Table

Product	OJ	Window Cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	2
Window cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	2	0	0	1	2

Order ID	Products
1	Orange juice, Soda
2	Milk, orange juice, window cleaner
3	Orange juice, detergent
4	Orange juice, detergent, soda
5	Window cleaner, soda

Insights

Product	OJ	Window Cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	2
Window cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	2	0	0	1	2

- Orange juice and soda OR Orange juice and detergent are more likely to be purchased together than any other two items.
- Detergent is never purchased with window cleaner or milk.
- Milk is never purchased with soda or detergent.

Important considerations in building rules

- Choosing the right set of items
- The co-occurrence tables can be huge
  - Overcoming the practical limits imposed by thousands or tens of thousands of items

CSE 7405c

Different purposes require us to go to different degrees of depth of products

Supermarket owner may be fine if rules are found at

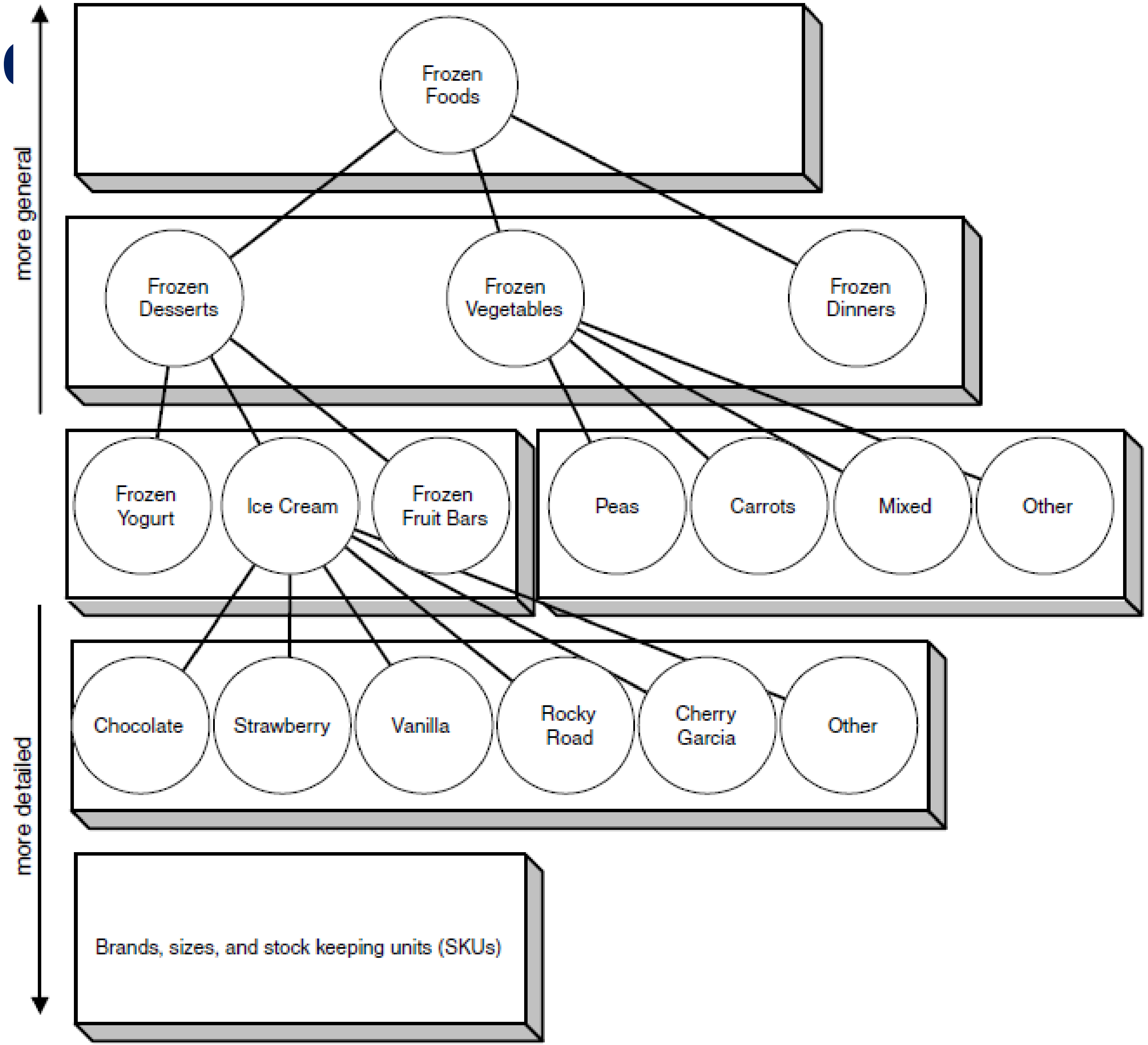
CUSTOMER	PIZZA	MILK	SUGAR	APPLES	COFFEE
1	✓				
2		✓	✓		
3	✓			✓	✓
4		✓			✓
5	✓		✓	✓	✓

The maker of pizza in the above table may be interested in

CUSTOMER	EXTRA CHEESE	ONIONS	PEPPERS	MUSHROOMS	OLIVES
1	✓	✓			✓
2			✓		
3	✓	✓		✓	
4		✓			✓
5	✓		✓	✓	✓

CSE 7405c

Use Hierarchies to determine when to stop item sel



CSE 7405c

Use Hierarchies to determine when to stop item selection

- Product hierarchies are built in coding SKUs
- # of combinations increases rapidly as items increase suggesting using items from higher levels
- On the other hand, more specific items provide more actionable rules
- A solution to the above conflict could be to first use more general items and then when honing on specifics, use only the subset of data containing them

Best use of Market Basket analysis

- Best results obtained when items occur in roughly the same number of transactions to prevent common items from dominating the rules
- Use hierarchies to roll up rare items to more general items, and leave more common items as is
- Data quality is extremely important for Association Rules

APRIORI ALGORITHM

Association Rules

- There are a large number of association rules algorithms.
- They all use different strategies and data structures.
- We will study the Apriori algorithm and its variants as they are the widely used.

CSE 7405c

Other Algorithms

- Eclat algorithm
- FP-growth algorithm
- AprioriDP
- Context Based Association Rule Mining Algorithm
- Node-set-based algorithm
- GUHA procedure ASSOC
- OPUS search

Two-Step Process

- Find all itemsets that have **minimum support**. These are called ***frequent itemsets***.

- Use ***frequent itemsets*** to generate **rules**.

Apriori Property

- The **key idea** behind the algorithm is called the **apriori** property or **downward closure** property.

- Downward closure property: Any **subset** of a frequent itemset is **also** a frequent itemset.

CSE 7405c

Closed

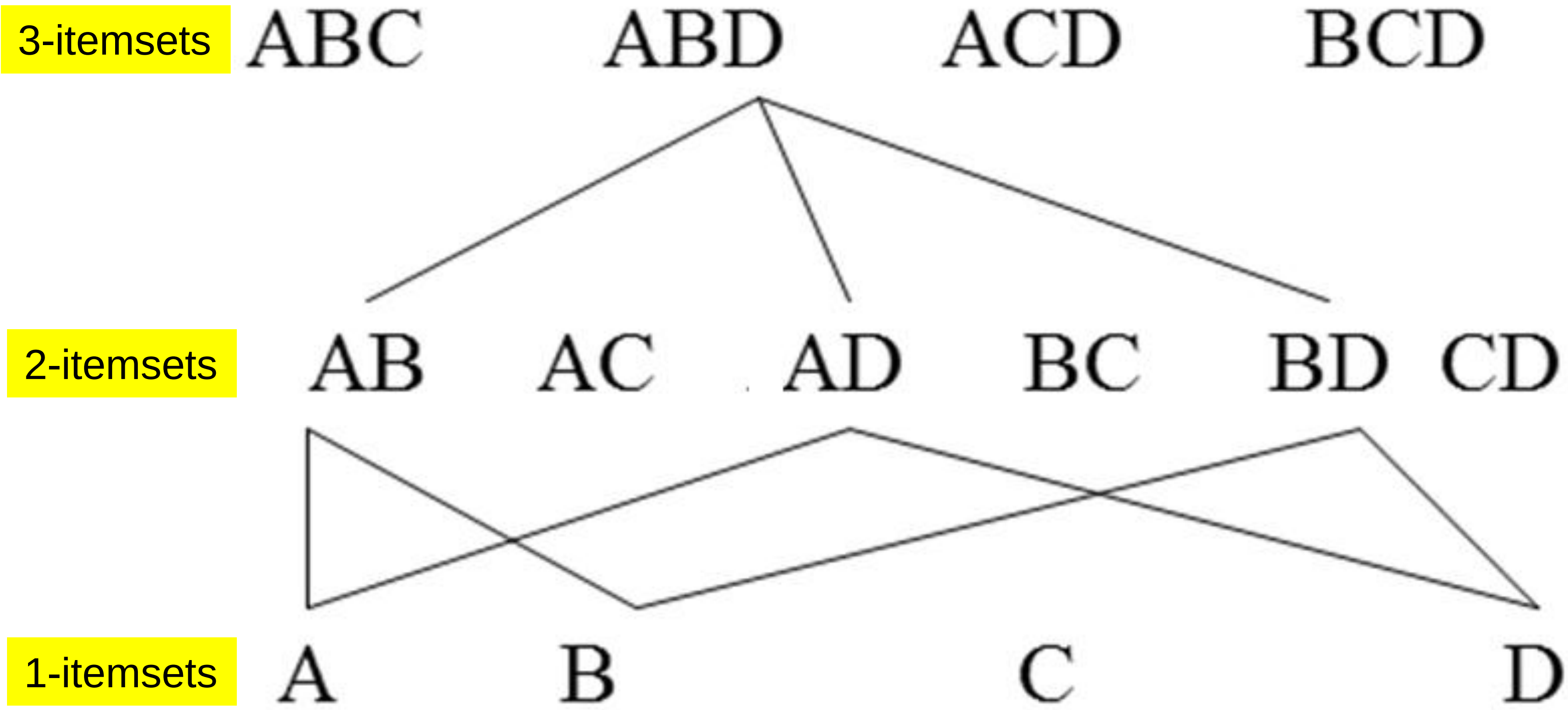
- A **set** is said to be closed under an **operation**, if the **operation** produces **another number** of the **set**.

CSE 7405c

Natural Numbers (Whole, Non-negative)

- Closed under
  - Addition (e.g., 3+5 gives 8)
  - Multiplication (e.g., 3*5 gives 15)
- Not closed under
  - Subtraction (e.g., 3-5 gives a negative number)
  - Division (e.g., 3/5 gives a fraction)

Downward closure



CSE 7405c

- Suppose {A,B} is frequent. Since each occurrence of A, B includes both A and B, then both A and B must also be frequent
- Similar argument for larger itemsets
- So, if a *k*-itemset is frequent all its subsets (*k-1*, *k-2* itemsets) are also frequent

The Apriori Algorithm — Example

Database D

Minsup = 0.5

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

CSE 7405c

Finalizing Rules from Apriori

Calculate confidence for each of the finalized *k*-itemsets, formulate rules and finalize those meeting the minimum confidence required

CSE 7405c

- Example itemset

  {Milk, Diaper, Beer}

- Rules (Calculate Support and Confidence for each of the below)

  s = 0.4, c = 0.67

  {Milk, Diaper} → {Beer} s = 0.4, c = 1.00

  {Milk, Beer} → {Diaper} s = 0.4, c = 0.67

  {Diaper, Beer} → {Milk} s = 0.4, c = 0.67

  {Beer} → {Milk, Diaper} s = 0.4, c = 0.50

  {Diaper} → {Milk, Beer} s = 0.4, c = 0.50

  {Milk} → {Diaper, Beer}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Observation:** Rules originating from the same itemset have identical support but can have different confidence *(once again recall joint and*

Apriori Recap

Definition

- An expression of the form X → Y is a rule, where X and Y form the itemset
- X is the rule's antecedent and Y is the rule's consequent

Example: $\{Milk, Diaper\} \Longrightarrow Beer$

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Rule Evaluation Metrics

- Support (s)
  - Fraction of transactions that contain both X and

$$s = \frac{\sum(Milk, Diaper, Beer)}{|T|} = \frac{2}{5} = 0.4$$

- Confidence (c)
  - Measures how often Y appears in transactions that contain X

$$c = \frac{\sum(Milk, Diaper, Beer)}{\sum(Milk, Diaper)} = \frac{2}{3} = 0.67$$

CSE 7405c

Titanic Survivors - R

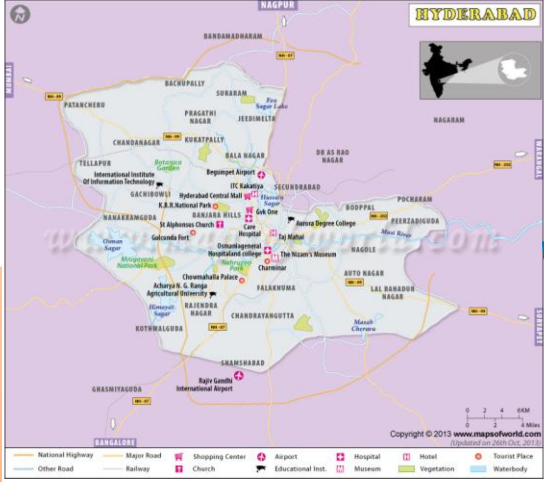
CSE 7405c

Titanic Survivors - R

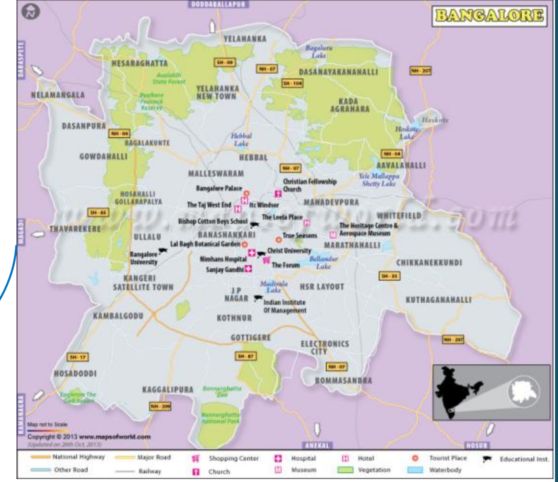
Age/gender	Class/crew	Number aboard	Number saved	Number lost	Percentage saved	Percentage lost
Children	First Class	6	5	1	83%	17%
	Second Class	24	24	0	100%	0%
	Third Class	79	27	52	34%	66%
Women	First Class	144	140	4	97%	3%
	Second Class	93	80	13	86%	14%
	Third Class	165	76	89	46%	54%
	Crew	23	20	3	87%	13%
Men	First Class	175	57	118	33%	67%
	Second Class	168	14	154	8%	92%
	Third Class	462	75	387	16%	84%
	Crew	885	192	693	22%	78%
Total		**2224**	**710**	**1514**	**32%**	**68%**

Source: https://en.wikipedia.org/wiki/RMS_Titanic
Last accessed: January 27, 2016

CSE 7405c



Inspire...Educate...Transform.



HYDERABAD
Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU
Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.