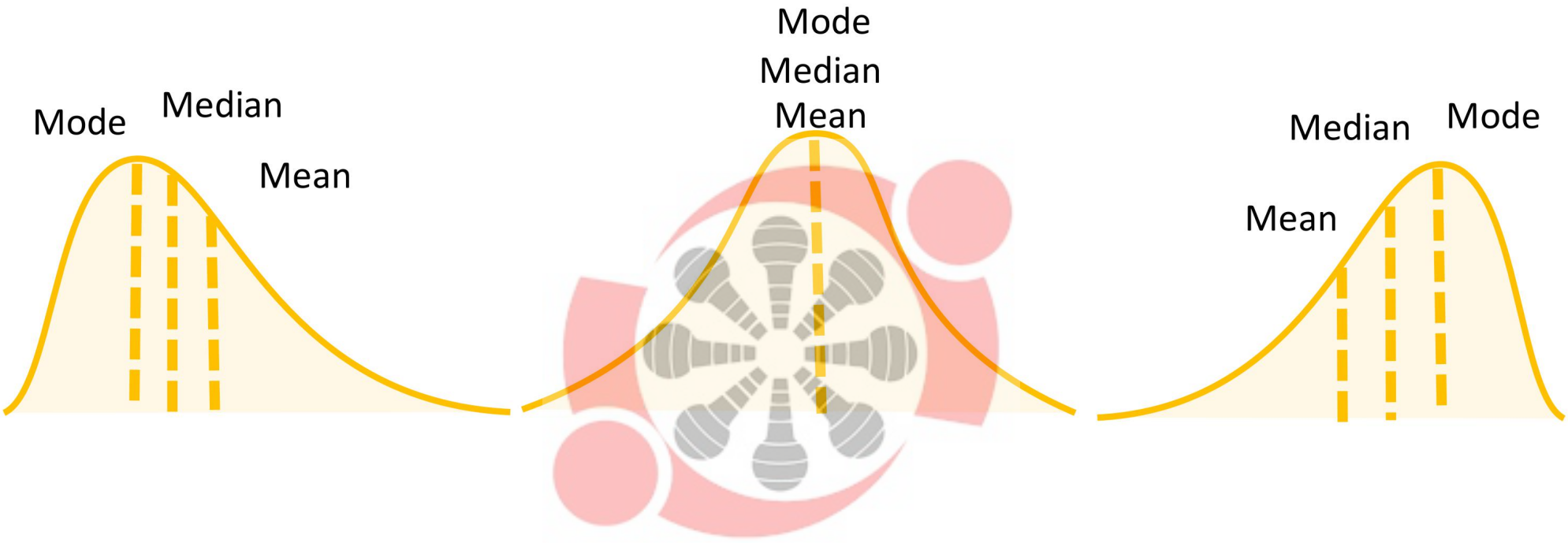
The Central Tendencies Review

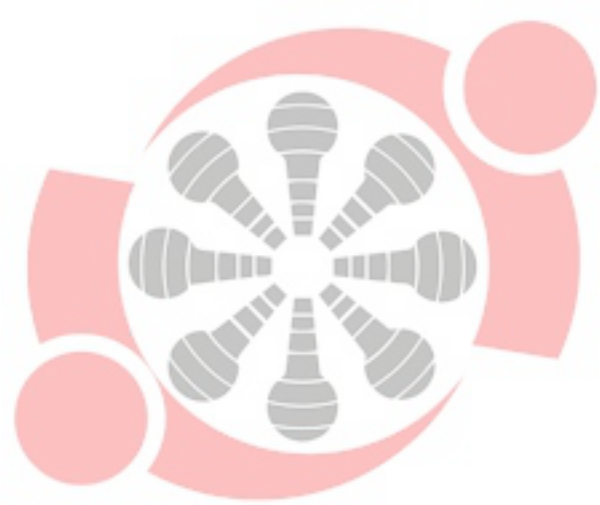
The Central Tendencies



The Central Tendencies

For the dataset, $23, 14, 17, 28, 15, 20$ the median is

- 18.5
- 17
- 15
- 20



The Central Tendencies

The spread of the data in a dataset could be studied using

_____

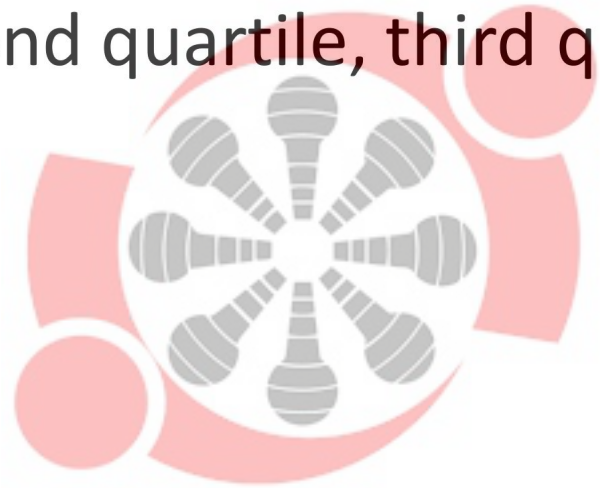
- Interquartile range
- Variance
- Standard Deviation
- Range (max-min)
- All of the above

The Central Tendencies

Given the numbers are **168, 183, 158, 184, 200, 164**,

Find:

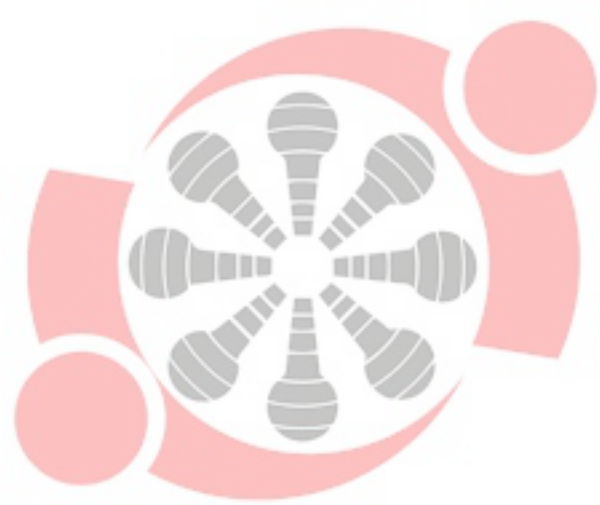
1. first quartile, second quartile, third quartile
2. IQR
3. Any Outliers



The Central Tendencies

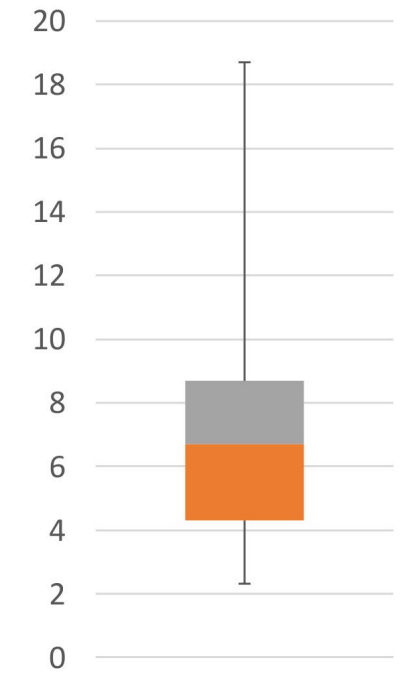
Which of the following plot is used to analyze interquartile range

- Scatterplot
- Histogram
- Lineplot
- Boxplot
- All of the above



The Central Tendencies

What term would best describe the shape of the given boxplot?



- Symmetric
- Skewed with right tail
- Skewed with left tail
- Normal

The Central Tendencies

A sample of 1000 Hyderabad households is selected and several variables are recorded. Which of the following statements is correct?

- Socioeconomic status (recorded as "low income", "middle income", or "high income") is nominal level data

- The number of people living in a household is a discrete variable

- The primary language spoken in the household is ordinal level data (recorded as "Kannada", "Tamil", etc)

The Central Tendencies

We studied Quartiles in depth and mentioned Deciles and Percentiles in passing. However, just as Quartiles divide data into 4 equal parts, Deciles divide it into 10 equal parts and Percentiles into 100 equal parts.

Given the above, find the 25^{th}, 50^{th}, 75^{th} and the 90^{th} percentiles for the top 16 global marketing sectors for advertising spending for a recent year according to *Advertising Age. Also, find Q2 and* IQR. Data in next slide.

Sector	Ad spending (in $ million)
Automotive	22195
Personal Care	19526
Entertainment and Media	9538
Food	7793
Drugs	7707
Electronic	4023
Soft Drinks	3916
Retail	3576
Restaurants	3571
Cleaners	3553
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137
Toys	699

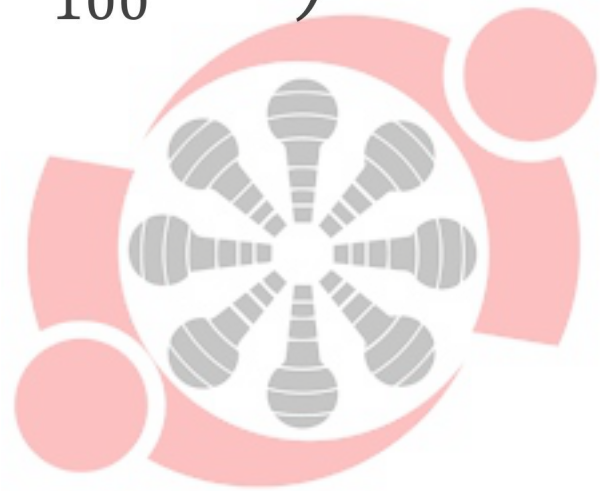
The Central Tendencies

N = 16

90^{th} percentile = $\left(p * \frac{n-1}{100} + 1\right)$

$= \frac{90(16-1)}{100} + 1$

= 14.5



Probability Basics

Understanding Probability

Consider the following statements. How do you interpret "probability" in each one of those? And how is it computed?

- Coin Toss – Probability of Head is ½
- Weather – Probability of thunderstorm tomorrow is 25%
- Cricket– India has only a 80% chance to win world cup

Probability vs Statistics

- Probability – Predict the likelihood of a future event
- Statistics – Analyze the past events

Questions addressed -

- Probability – What will happen in a given ideal world?
- Statistics – How ideal is the world?

Probability - Applications

- Gaming industry – Establish charges and payoffs
- Manufacturing/Aerospace – Prevent major breakdowns
- Business – Deciding on a business proposal based on probability of success vs cost
- Risk Evaluation – Scenario analysis

Assigning Probabilities

Classical Method – *A priori or Theoretical*

Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ even\ occurs}{total\ possible\ \#\ of\ outcomes}$$

Example: Tossing of a fair die

Computing A priori Probability

Find the probability of pulling a yellow marble from a bag of 3 yellow, 2 red, 3 green and 1 blue marbles

$$P(yellow) = \frac{No\ of\ yellow\ marbles}{Total\ number\ of\ marbles}$$

$$= \frac{3}{9}$$

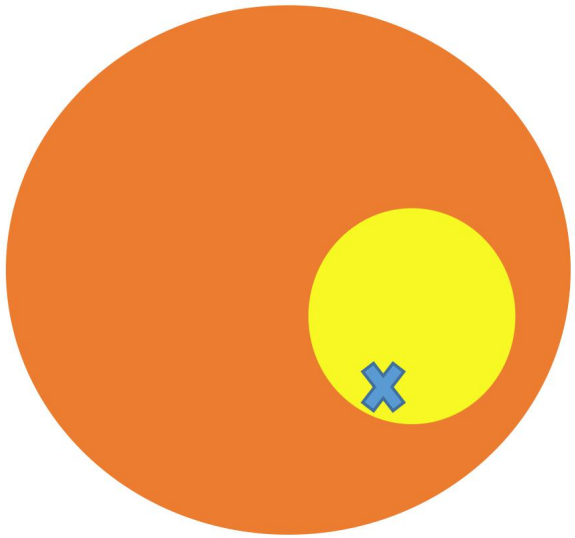
Computing probability

There are two concentric circles, The circumference of a circle is 36π. Contained in that circle is a smaller circle with an area of 16π. A point is selected at random from inside the larger circle. What is probability that the point also in the same circle.

$$\text{Area of smaller circle} = 16\,\pi$$

$$\text{Area of larger circle} = \pi * \left(\frac{36\pi}{2\pi}\right)^2$$

$$= 324\,\pi$$



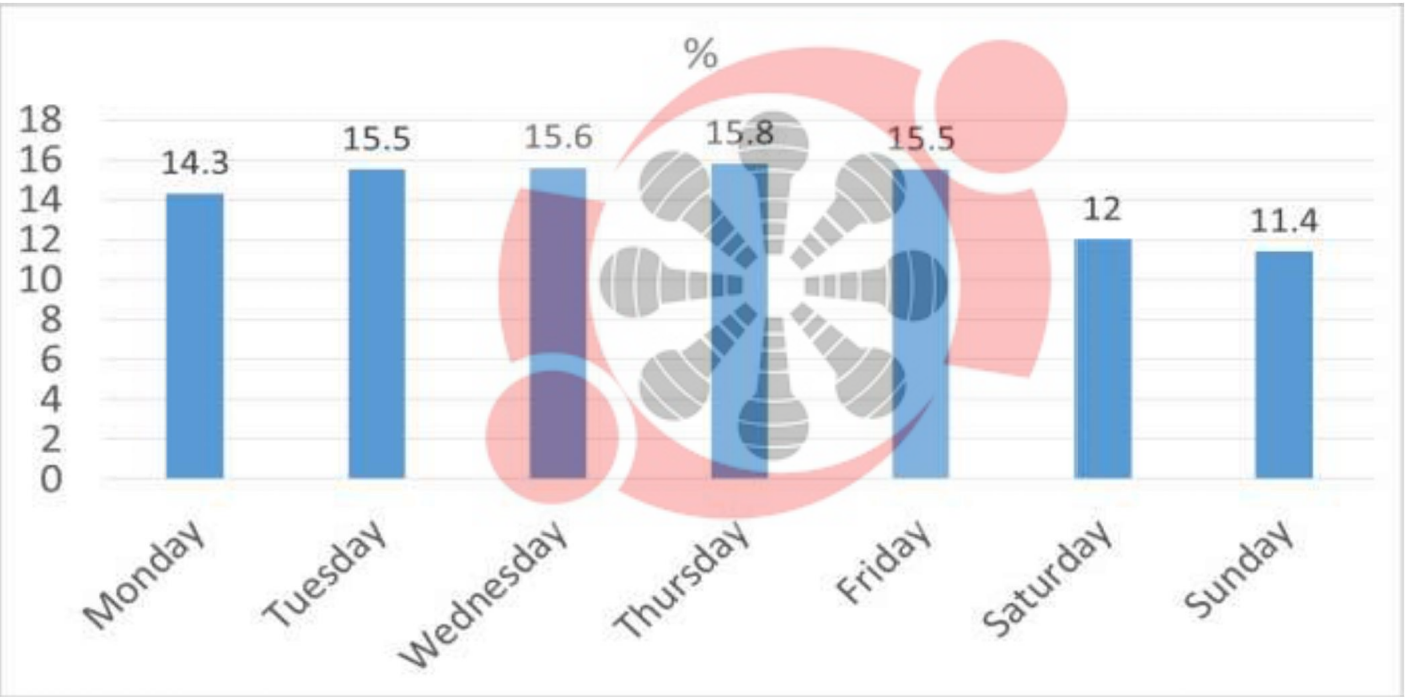
$$P(\text{point in small circle}) = \frac{Area\ of\ Large\ circle}{Area\ of\ small\ circle}$$

$$= 16\pi/324\pi$$

Assigning Probabilities

What is the probability of a baby being born on a Wednesday?

A-priori probability = $\frac{1}{7} = 14.3\%$



Data from "Risks of Stillbirth and Early Neonatal Death by Day of Week", by Zhong-Cheng Luo, Shiliang Liu, Russell Wilkins, and Michael S. Kramer, for the Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. Data of 3,239,972 births in Canada between 1985 and 1998. The reported percentages do not add up to 100% due to rounding.

Assigning Probabilities

Empirical Method – *A posteriori or Frequentist*

Probability can be determined post conducting a thought experiment.

$$P(E) = \frac{\#\ of\ times\ an\ event\ occured}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occured}$$

Example: Tossing of a weighted die…well!, even a fair die.

The larger the number of experiments, the better the approximation.

This is the most used method in statistical inference.

Assigning Probabilities

Subjective Method

Based on feelings, insights, knowledge, etc. of a person.

What is the probability of India winning the upcoming World cup 2019?

Probability - Terminology

Sample Space – Set of all possible outcomes, denoted S.

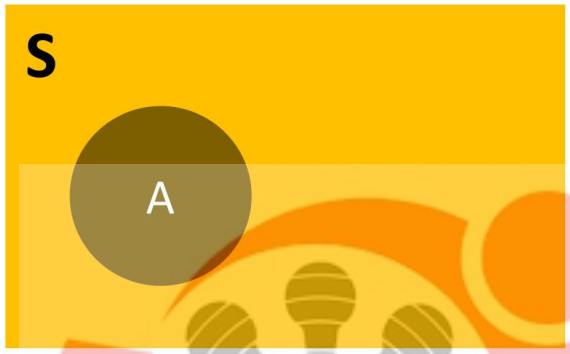
Example:

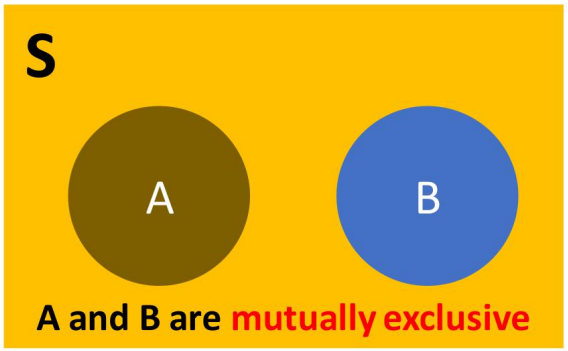
After 2 coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

Event – A subset of the sample space.

An Event of interest might be - HH

Probability - Rules





$P(s) = 1$

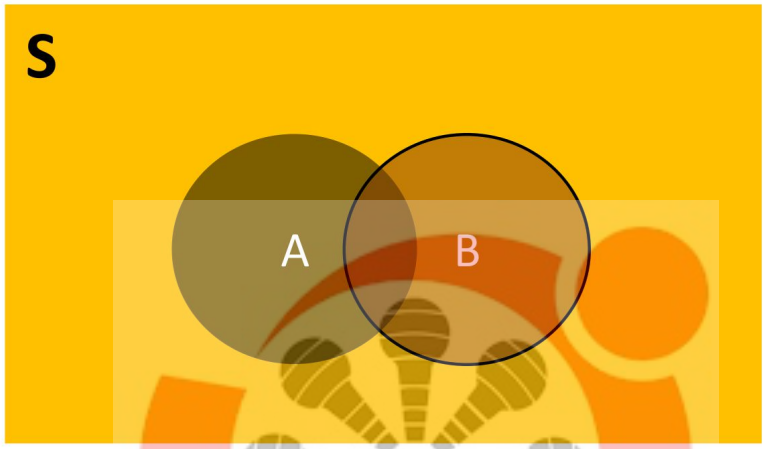
$0 \leq P(A) \leq 1$

$P(A \; or \; B) = P(A) + P(B)$

Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

Mutually Exclusive – If event A happens, event B cannot.

Probabilities Rules



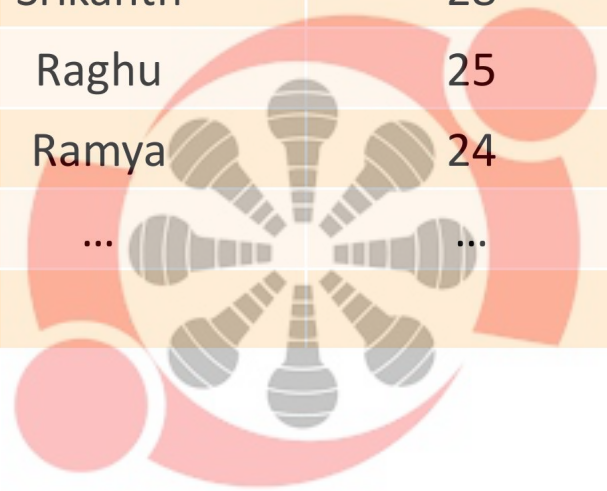
$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

Example

- Event A – Customers who default on loans
- Event B – Customers who are High Net Worth Individuals

Probability - Types

Customer-Id	Customer Name	Age	Default
846596	Srikanth	28	Yes
846597	Raghu	25	No
846598	Ramya	24	No
…	…	…	…



Probability - Types

Contingency table summarizing 2 variables, *Loan Default* and *Age:*

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	5252	13684	130	19066
	Yes	1793	2426	60	4279
	Total	7045	16110	190	23345

Probability - Types

Convert it into probabilities

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability - Types

Marginal Probability

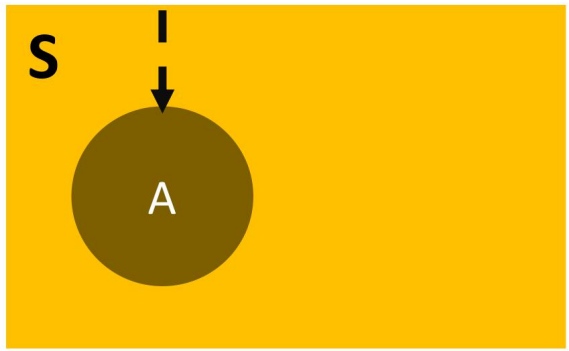
		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Marginal Probability

Probability describing a single attribute

P(Middle) = 0.690

P(old) = 0.008



Probability - Types

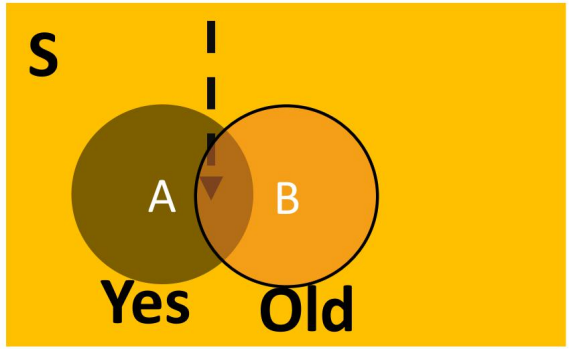
Joint Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Joint Probability

Probability describing a combination of attribute

P(Yes and old) = 0.003



Probability - Types

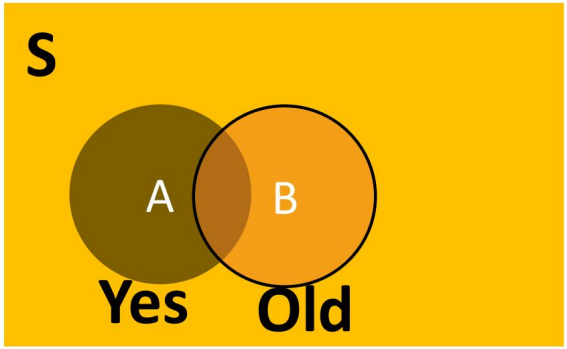
Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(Yes or old) = P(Yes) + P(old) – P(Yes and old)

= 0.184 + 0.008 – 0.003

= 0.189

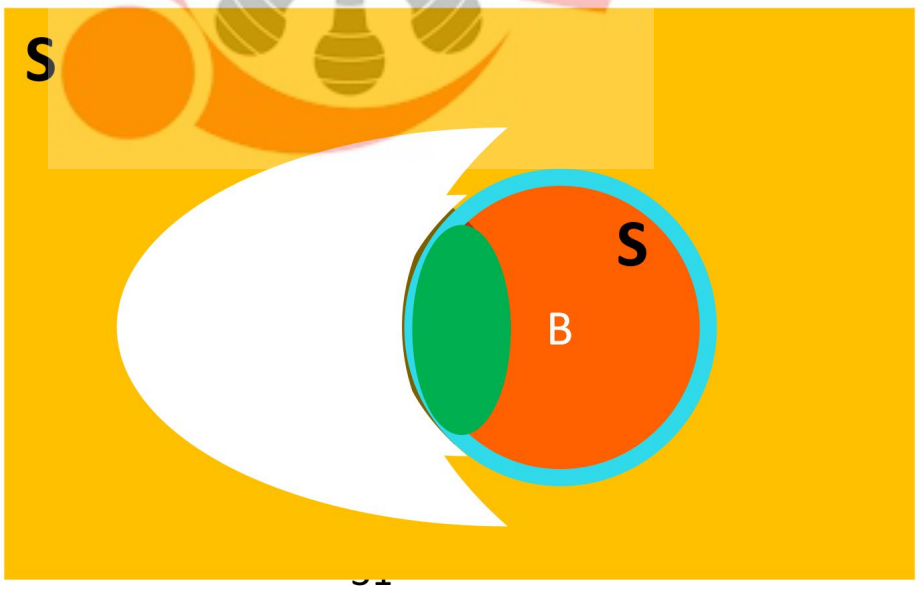


Probability - Types

Conditional Probability

- Probability of *A occurring* ***given that*** B has occurred.
- The sample space is restricted to a single row or column.
- This makes rest of the sample space irrelevant.

Probability, i.e., $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$



Probability - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

What is the probability that a person will not default on the loan payment **given she is middle-aged?**

Probability, i.e., $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

P(No | Middle-Aged) = 0.586/0.690 = 0.85

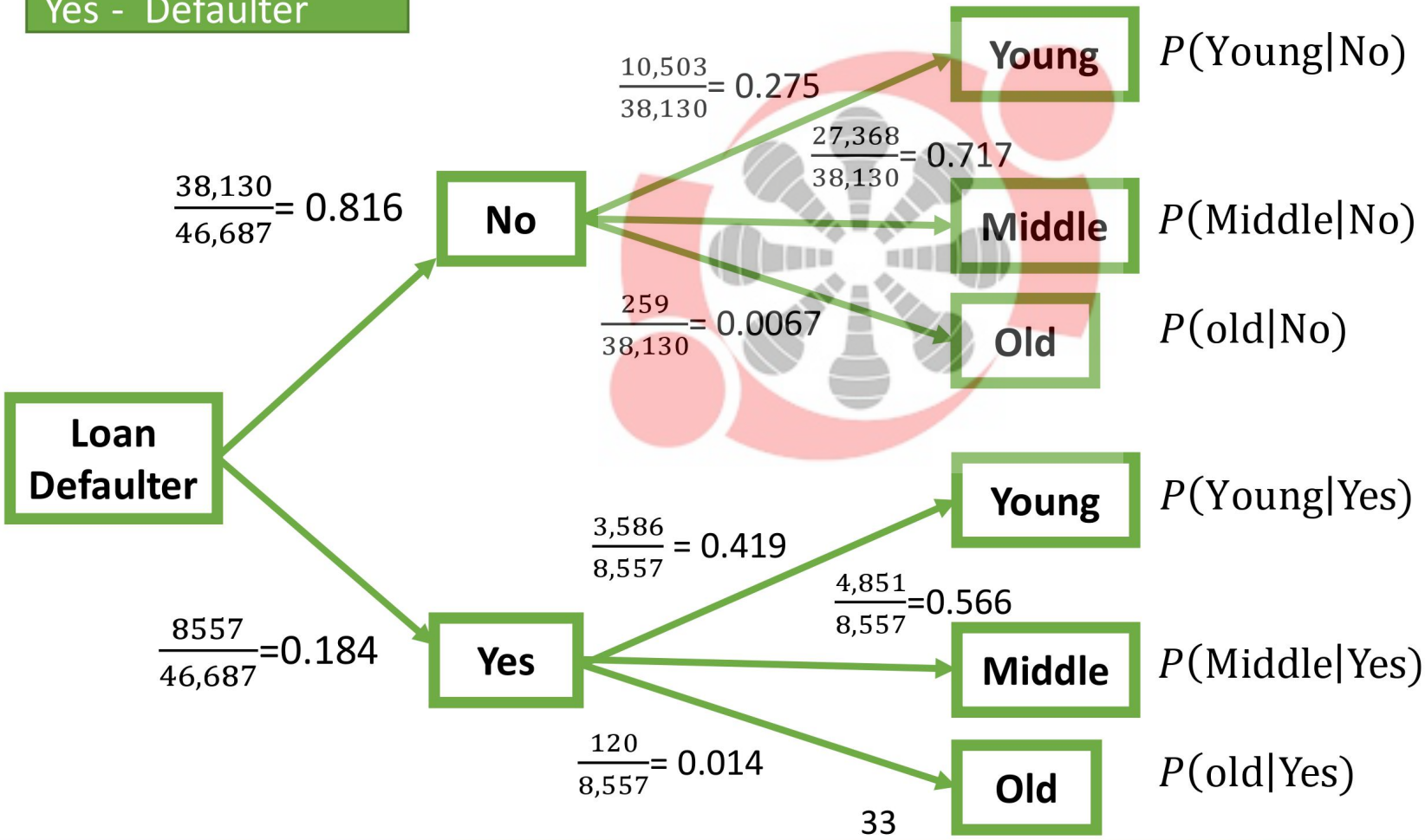
Note that this is the ratio of **Joint Probability to Marginal**

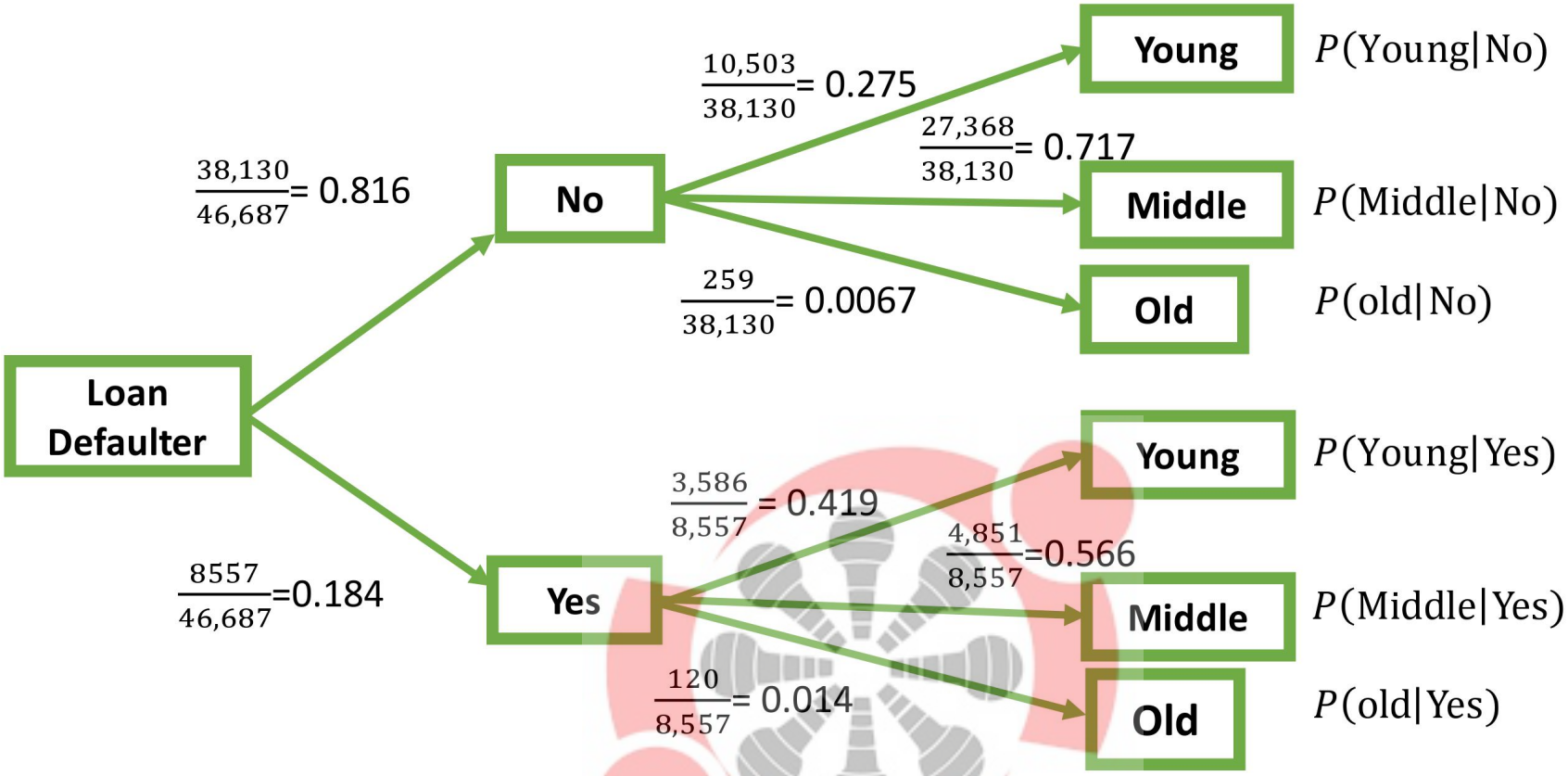
P(Middle-Aged | No) = $\frac{0.586}{0.816}$ = 0.72 (Order Matters)

		Age			
		Young	Middle-aged	Old	Total
Loan Defaults	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Age			
Young	Middle-aged	Old	Total
0.225	0.586	0.005	0.816
0.077	0.104	0.003	0.184
0.302	0.690	0.008	1.000

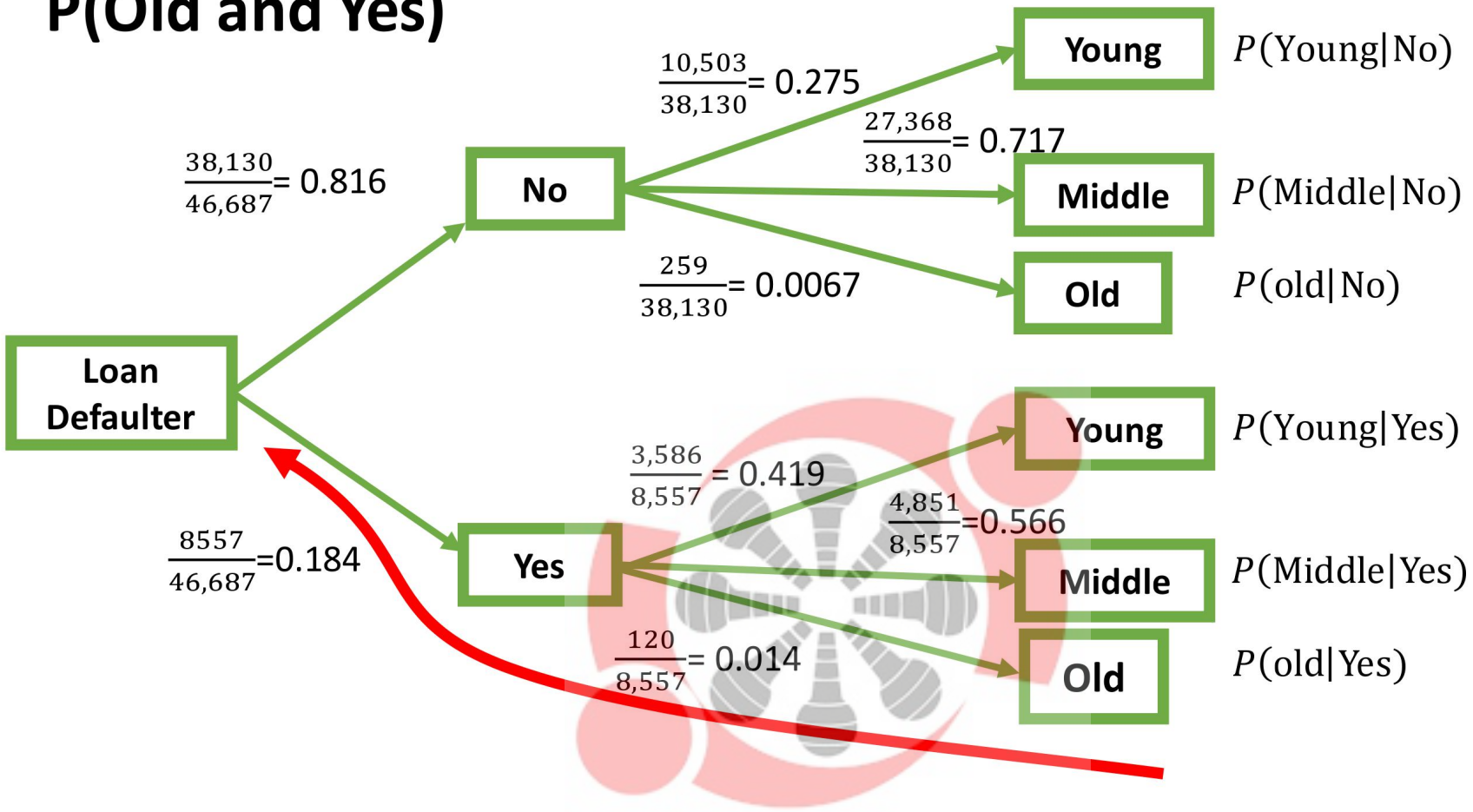
No – Non-defaulter
Yes - Defaulter





- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

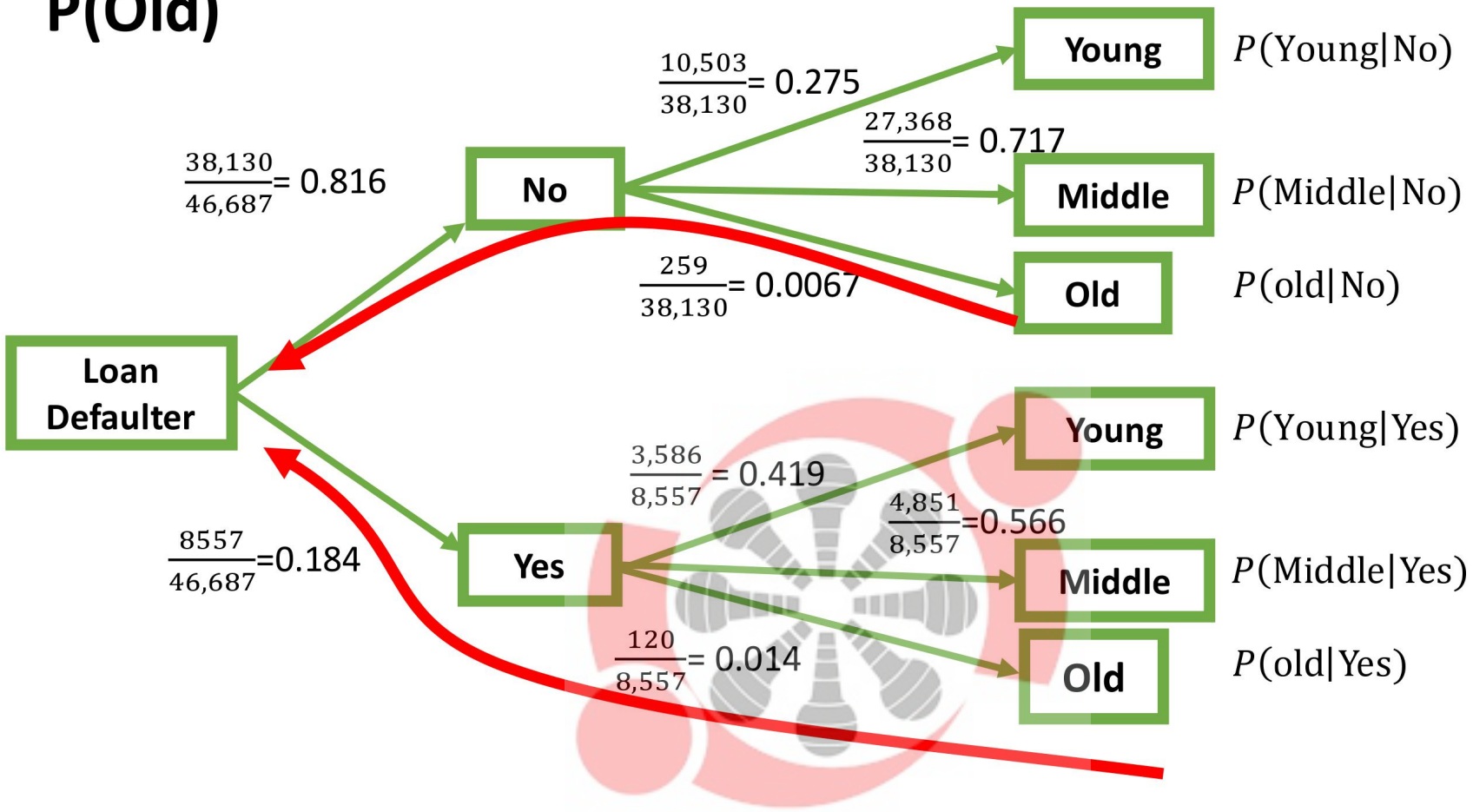
P(Old and Yes)



$$P(\text{Old}\,|\,\text{Yes}) = \frac{P(\text{Old and Yes})}{P(\text{Yes})}$$

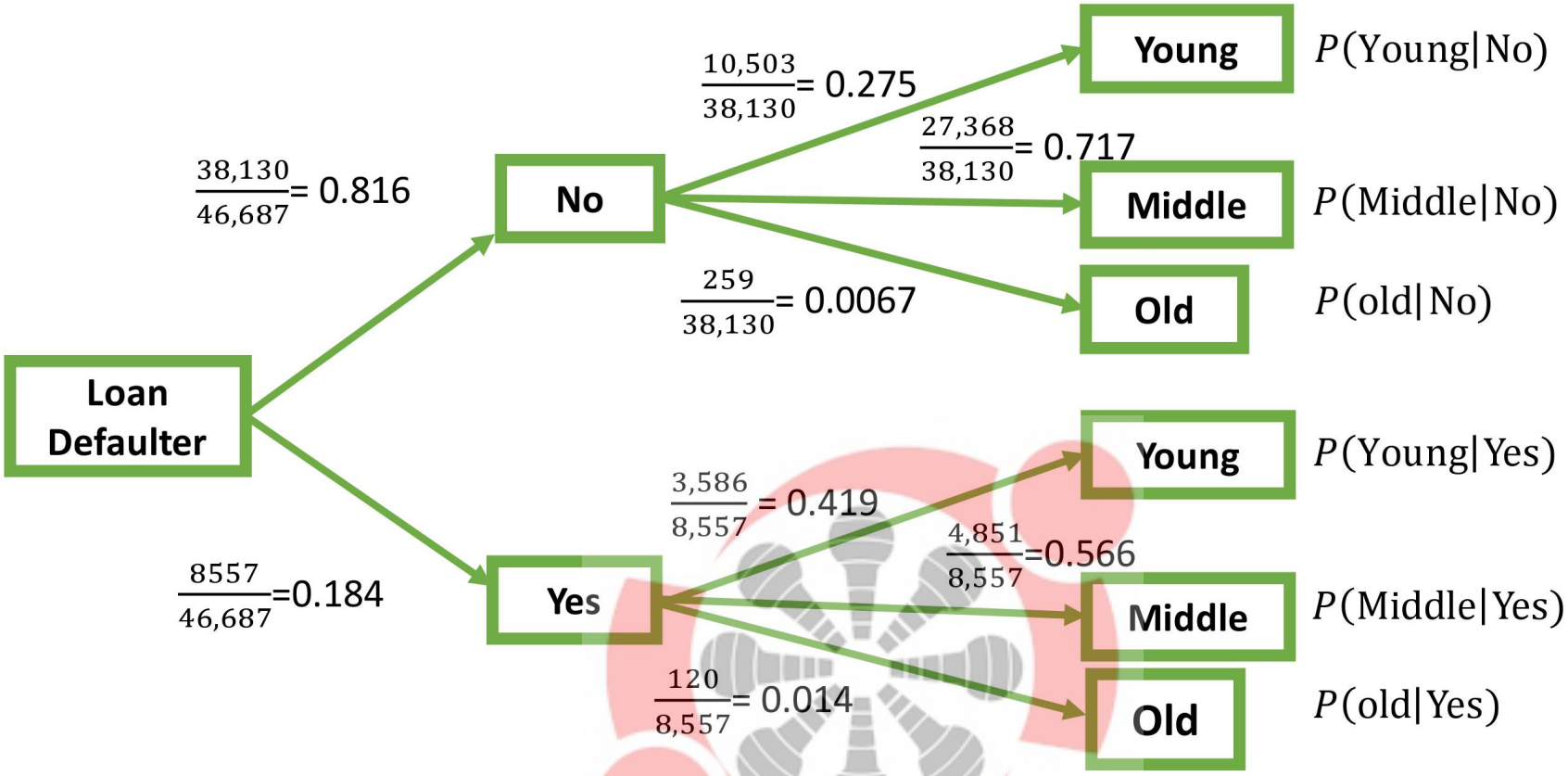
$$P(\text{Old and Yes}) = P(\text{Old}\,|\,\text{Yes}) * P(\text{Yes}) = 0.014 * 0.184$$

P(Old)



$$P(\text{Old}) = P(\text{Old and Yes}) + P(\text{Old and No})$$

$$P(Old) = 0.014 * 0.184 + 0.0067 * 0.816$$



- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(B) * P(A|B)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \Rightarrow P(A \text{ and } B) = P(A) * P(B|A)$$

Equating, we get

$$P(B) * P(A|B) = P(A) * P(B|A)$$

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Probability - Types

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

In loan defaulters older people make up only 1.4%. Now the probability that someone defaults on a loan is 0.184, Find the probability default on loan knowing that he is old person. Older people make up only 0.8%.

Ans:

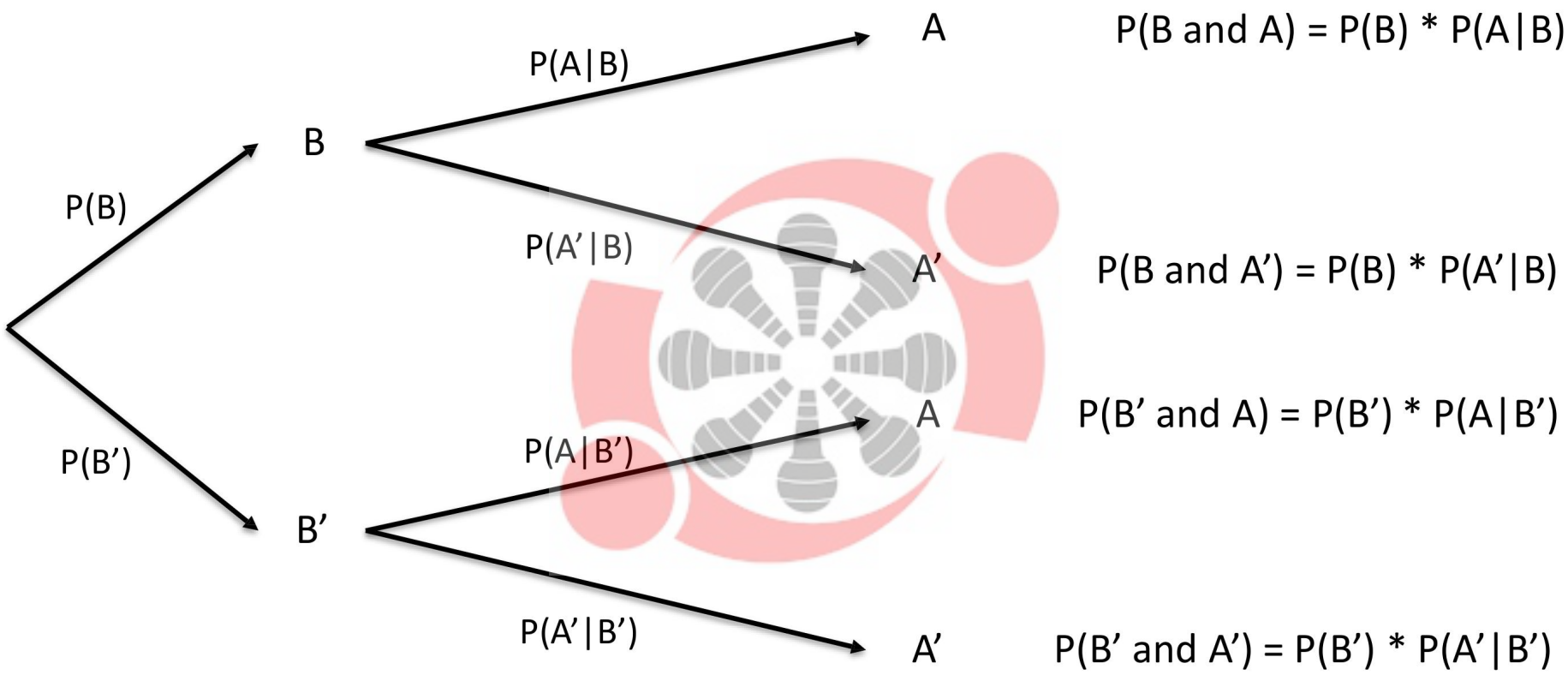
$P(\text{Old}|\text{Yes}) = 0.014$

$P(Yes) = 0.184$

$P(Old) = 0.008$

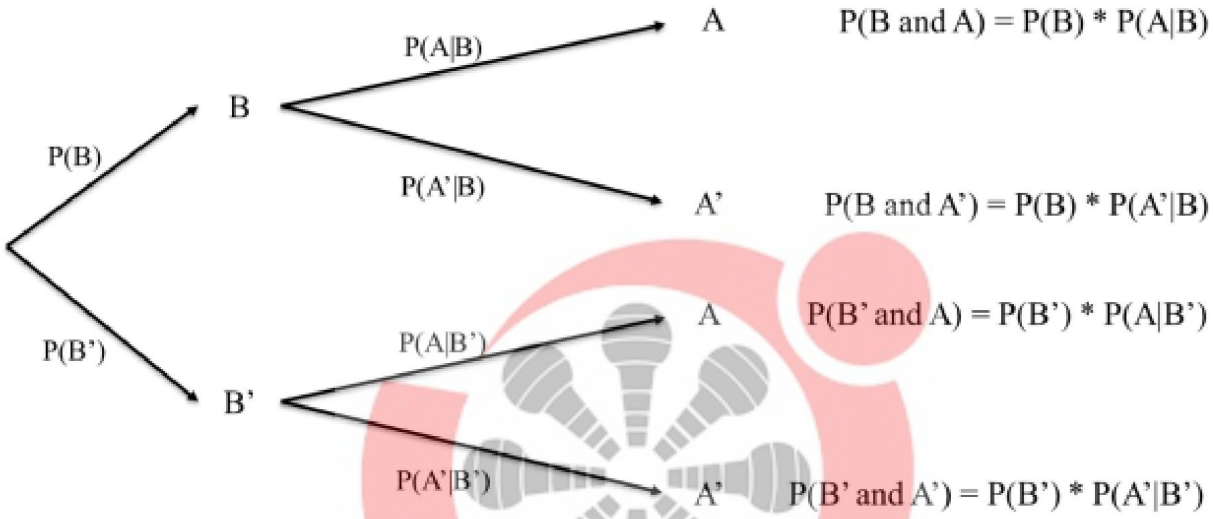
$$P(\text{Yes}|\text{Old}) = \frac{P(\text{Yes}) * P(\text{Old}|\text{Yes})}{P(\text{Old})} = \frac{0.184 * 0.014}{0.008} = 0.32$$

Generalized Probability Tree



State each probability in English; note B' means "not B".

Conditional Probability → Bayes Theorem



$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A|B) * P(B) + P(A|not\,B\,) * P(not\,B)}$$

Note B' means "not B"

Bayes' Theorem ⇒Spam filtering

Apache SpamAssassin™

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 30% of the mails marked as spam, i.e., P(Free | Spam) = 0.30. Assuming 1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Draw the probability tree diagram

Bayes' Theorem

P(Spam) = 0.50

P(Free | Spam) = 0.30 (*aka* Prior Probability)

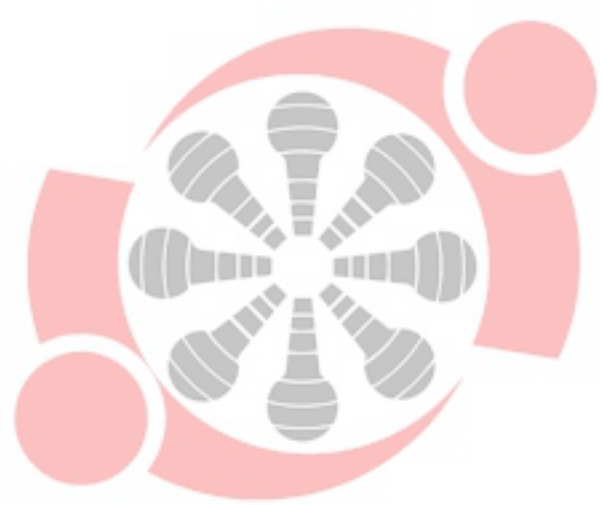
P(Free | No spam) = 0.01

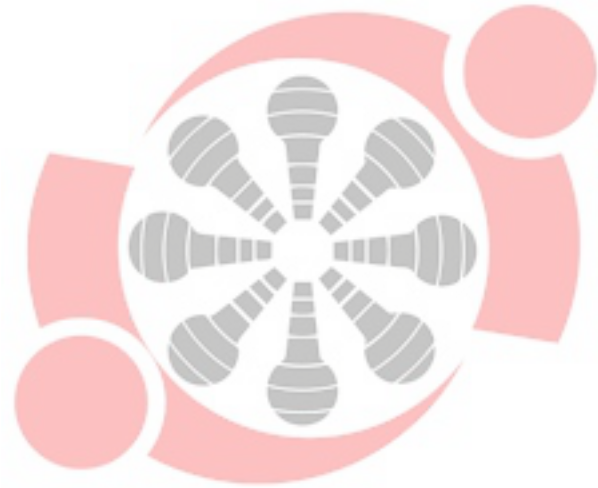
P(Spam | Free) = ? (*aka* Posterior or Revised Probability)

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.3}{0.3 * 0.5 + 0.01 * 0.5} = \frac{0.15}{0.155} = 0.9677$$

This helps the spam filter automatically classify the messages as spam.





How Good is Your Classification

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	1904	1052	2956
	Negative	334	6050	6384
Total		2071	7102	9340

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Spam filtering		Predicted		
		Positive	Negative	Total
Actual	Positive	1904	1052	2956
	Negative	334	6050	6384
Total		2071	7102	9340

$$Recall\ (sensitivity) = \frac{1904}{2956} = 0.644$$

$$\text{Precision} = \frac{1904}{2071} = 0.851$$

$$Accuracy = \frac{1904 + 6050}{1904 + 1052 + 334 + 6050} = \frac{7954}{9340} = 0.852$$

$$Spcificity = \frac{6050}{6384} = 0.948$$

$$F_1 = 2 * \frac{Precision * Recall}{Preceision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = 0.733$$

Which measure(s) is/are more important?

Confusion Matrix

Cancer		Predicted		
		Positive	Negative	Total
Actual	Positive	1904	1052	2956
	Negative	334	6050	6384
Total		2071	7102	9340

$$Recall\ (sensitivity) = \frac{1904}{2956} = 0.644$$

$$\text{Precision} = \frac{1904}{2071} = 0.851$$

$$Accuracy = \frac{1904 + 6050}{1904 + 1052 + 334 + 6050} = \frac{7954}{9340} = 0.852$$

$$Spcificity = \frac{6050}{6384} = 0.948$$

$$F_1 = 2 * \frac{Precision * Recall}{Preceision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = 0.733$$

Which measure(s) is/are more important?

Reference

- Headfirst Book