Week 1

MA5821- Advanced Statistical Methods for Data Scientists

Kazi Arif Hossain
kazi.hossain@jcu.edu.au

SUBJECT OVERVIEW

- Key Dates:
  - Census date 12th September, 2019
  - Last date to withdraw without academic penalty 19th September, 2019
- Assessment Items:
  - Assessment 1: weekly quizzes (30%)
  - Assessment 2: weekly workbook exercises (30%)
  - Assessment 3: capstone project (40%)
  - Self-learning practical workbook submissions (non-graded)

O-WEEK Q&A HIGHLIGHT

- Referential materials:
  - Statistics1: Introduction to ANOVA, Regression, and Logistic Regression
  - SAS Visual Analytics 7.3 - Getting started with exploration and reporting
  - Exploring Data with SAS Visual Analytics
- Capstone Datasets: chose from TUN SAS Data Dictionaries

WEEK 1 TO-DO LIST

- Go through week 1 materials, 4 topics
- Go through week exercise
- Complete and submit week 1 self-learning practical (due 09/Sep/2019)
- Complete week 1 quiz (60 Mins, due 08/Sep/2019)

TOPICS FOR WEEK 1

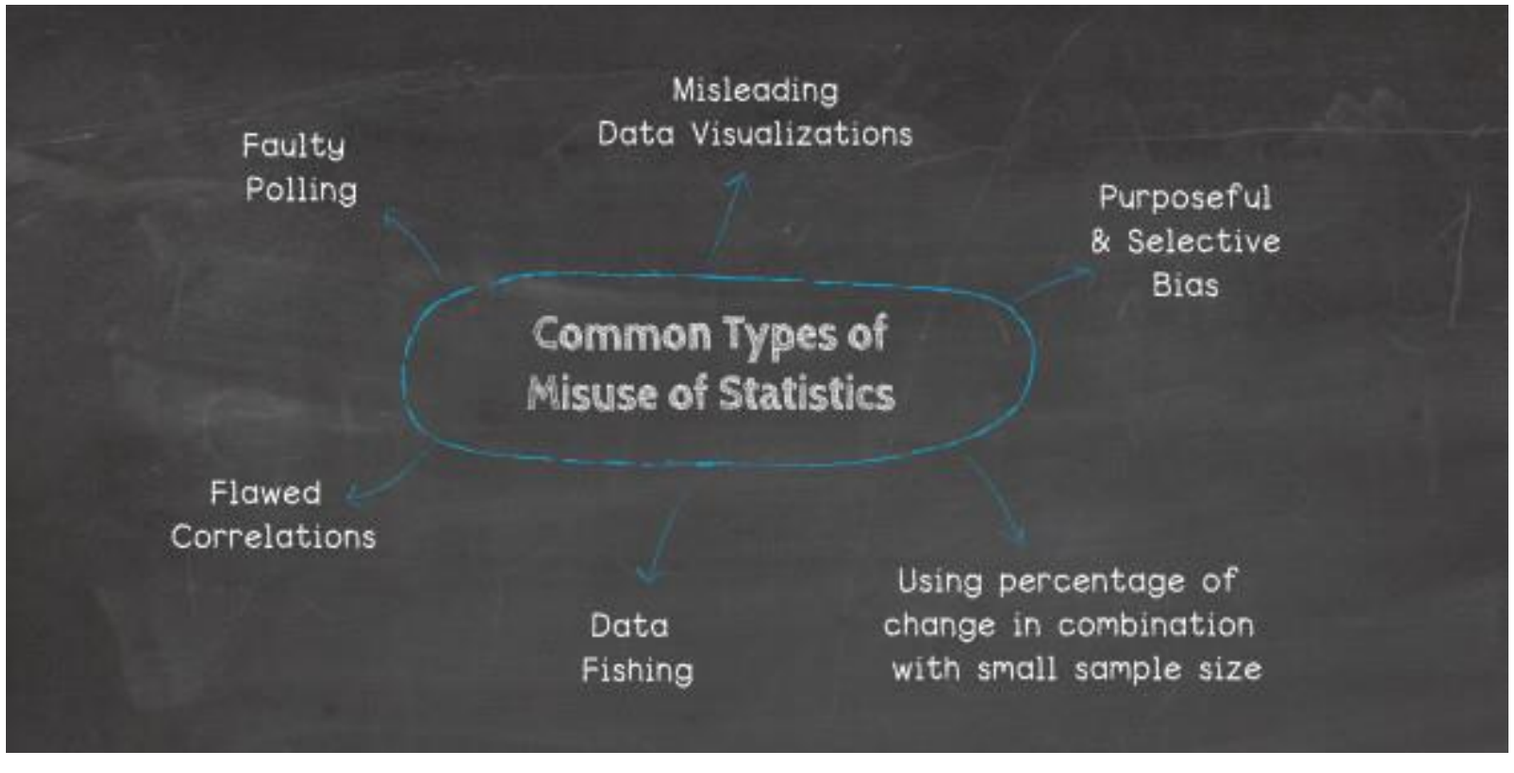
- Topic 1: Why are analytics and statistics important?
- Topic 2: Introduction to core concepts in advanced statistical modelling
- Topic 3: Advanced Probability
- Topic 4: Frequentist vs Bayesian statistics

MISLEADING STATISTICS

Misleading statistics are simply the misusage – purposeful or not – of a numerical data. The results provide a misleading information to the receiver, who then believes something wrong if he or she does not notice the error or does not have the full data picture.

73.6% Of All Statistics Are Made Up

33.7% of scientists surveyed admitted to questionable research practices, including modifying results to improve outcomes, subjective data interpretation, withholding analytical details and dropping observations because of gut feelings…. Scientists!



PROBABILITY

$$\text{Key Idea: } P(A) = \frac{\text{number times event A occus}}{\text{number of all events}}$$

Rules: $0 \leq P(A) \leq 1$, $P(\acute{A}) = 1 - P(A)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cap B) = P(A) \times P(B)$ *(A and B are independent events)*

Example:

Throwing Dice: When a single die is thrown, there are **six** possible outcomes: **1, 2, 3, 4, 5, 6**.

The probability of any **one** of them is $\frac{1}{6}$

CONDITIONAL PROBABILITY

Conditional probability can be thought of as a means of **adjusting** probability in light of **new** information.

We express this mathematically as **P(X|Y)** and can be read as the **conditional** probability of **X** given that **Y** has occurred, which can be thought of as an **adjusted** version of the probability of **X** in light of the **additional** information that **Y** has occurred or is occurring.

The vertical line **|** in the formula immediately after **X** designates conditional probability.

The formula can be expressed as,

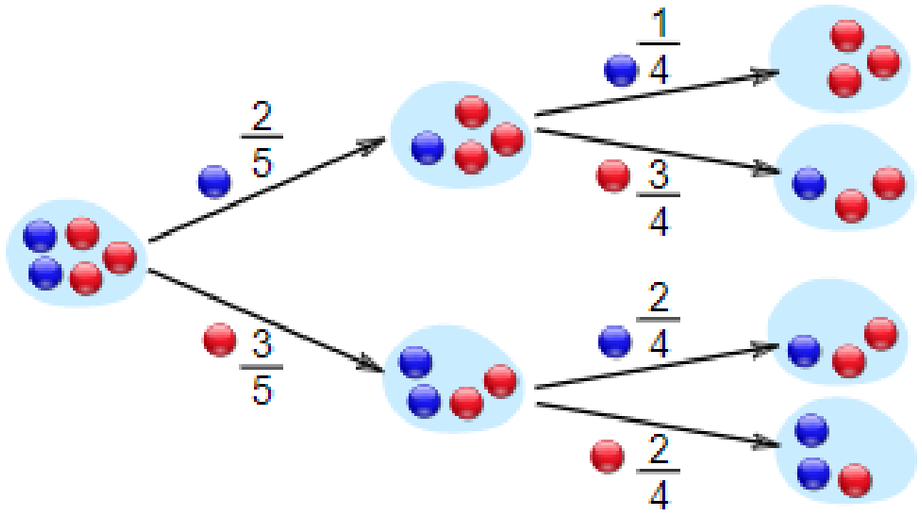
$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

provided that P(Y) is not zero.

CONDITIONAL PROBABILITY

Example: 2 blue and 3 red marbles are in a bag. What are the chances of getting a blue marble?

The chance is **2 in 5 (But after taking one out** the chances change!) Let's see how



"What are the chances of drawing 2 blue marbles?" *(with replacement: Independent)*

P(A) =2/5, P (B)=2/5

P(A ∩ B)= P(A) × P(B)= 2/5 × 3/5 = **6/25**

"What are the chances of drawing 2 blue marbles?"*(without replacement: Dependent)*

Event A is "get a Blue Marble **first": P(A) =2/5**

Event B is "get a Blue Marble **second**" *(condition: got a blue one **first**)* : **P(B|A) = 1/4**

P(A ∩ B)= P(A) × **P(B|A)**= 2/5 × 1/4 = **1/10**

ODDS

Probability is often stated in the form of **odds**. The concept of odds has two forms, namely, the odds in **favour** of event **A** and the odds **against** event **A**.

For example: **A** is the event of winning a race.

Odds in favour of **A** = $\frac{P(A)}{1-P(A)}$

Odds against A = $\frac{1-P(A)}{P(A)}$)

Dividend= 1+ *odds against* =1 + $\frac{1-P(A)}{P(A)}$= $\frac{1}{P(A)}$

How to find probability from dividend?

P(A)= $\frac{1}{Dividend}$ = $\frac{1}{5.5}$ = $\frac{2}{11}$

Bookmakers quote **odds** as odds **against** winning. A horse quoted at the fixed odds of **9** to **2** (often written as the ratio **9/2**) is expected to lose **9** and win just **2** out of every **11** races.

In Australia betting agencies often quote dividends for a **$1** bet rather than quoting odds against winning

For example, suppose that the odds against are quoted as **9** to **2**, then the winning dividend on a $1 stake is given by:

1 + $\frac{9}{2}$ = $\frac{1}{2/11}$ = $5.5

ODDS

Let's look at a betting **example**. Consider the betting market on an AFL match that was played in a recent season between St Kilda and Essendon. In the following table the dividend for this match is given for a $1 stake on a win before the match began.

Game	Team	Dividend for $1 stake on a win	Odds against
1	St Kilda Essendon	$2.90 $1.40	1.9 to 1 (or 19 to 10) 0.4 to 1 (or 4 to 10)

St Kilda would pay a dividend of $2.90 for a **$1** bet on a win. Thus, for every **$1** successfully bet on a St Kilda win, you would win **$1.90** (dividend amount −amountbet)

So a St Kilda win was quoted as odds against of 1.9 to 1 or 19 to 10. So St Kilda was expected to lose 19 out of every 29 games.

On the other hand Essendon would pay a dividend of **$1.40** for a **$1** bet on a win. An Essendon $1 bet would therefore pay, on an Essendon win, **$0.40** (dividend amount −amountbet).

So Essendon is expected to lose just 4 out of every 14 games.

FREQUENTIST VS BAYESIAN

I have misplaced my phone somewhere in the home. I can use the phone locator on the base of the instrument to locate the phone and when I press the phone locator the phone starts beeping.

Problem: Which area of my home should I search?

Frequentist Reasoning

I can hear the phone beeping. I also have a mental model which helps me identify the area from which the sound is coming. Therefore, upon hearing the beep, I infer the area of my home I must search to locate the phone.

Bayesian Reasoning

I can hear the phone beeping. Now, apart from a mental model which helps me identify the area from which the sound is coming from, I also know the locations where I have misplaced the phone in the past. So, I combine my inferences using the beeps and my prior information about the locations I have misplaced the phone in the past to identify an area I must search to locate the phone.

BAYESIAN THEOREM

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

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

$$P(B \cap A) = P(B|A) \times P(A) = P(A \cap B)$$

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

BAYESIAN THEOREM EXAMPLE

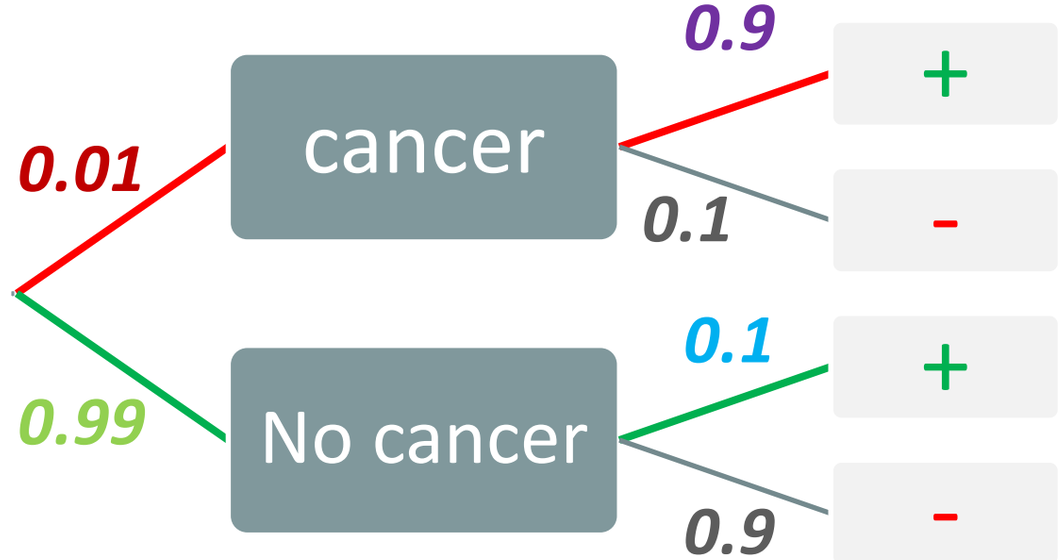
Approximately **1%** of women aged 40-50 have breast cancer. A woman with breast cancer has a **90%** chance of a positive test from a mammogram, while a woman without has a **10%** chance of a false positive result. What is the probability a woman has breast cancer given that she just had a positive test?

let A = "the woman has breast cancer" and B = "a positive test“ P(A|B)=??

$$P(\text{cancer}|+)=\frac{P(+|\text{cancer})\,P(\text{cancer})}{P(+)}$$

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

$$P(A|B)=\frac{P(B|A)\times P(A)}{P(B|A)\times P(A)+P(B|\sim A)\times P(\sim A)}$$



$$\frac{(0.9)\times(0.01)}{(0.9)\times(0.01)+(0.99)\times(0.1)}=\frac{9}{108}=\mathbf{8.3\ \%}$$

BIRTHDAY PROBABILITY PROBLEM

If there are 23 people in a room, what is the probability that at least two of them have the same birthday??

P(A) = the probability that **at least** two people in the room have the same birthday

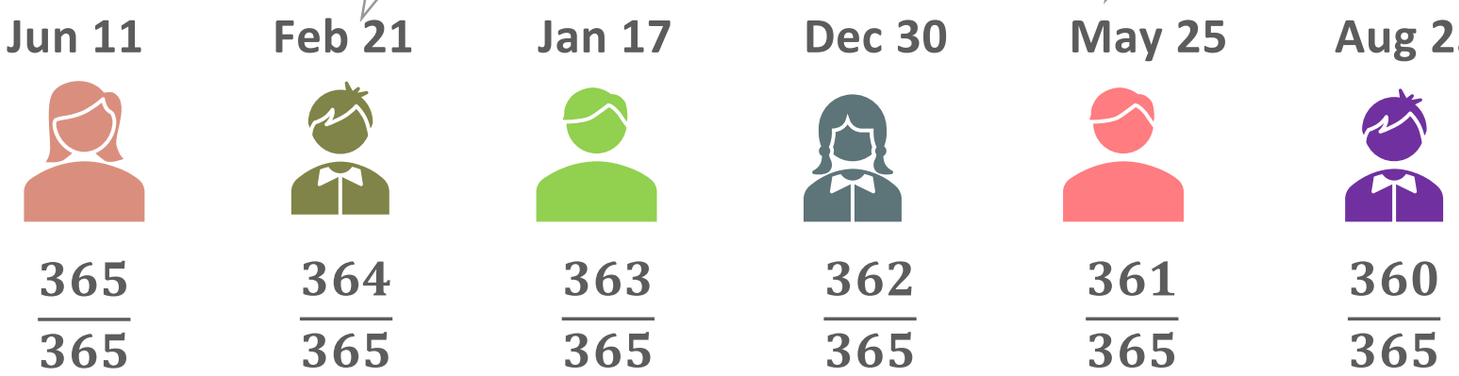
P(A')= the probability that **no two** people in the room have the same birthday.

$$P(A) = 1 - P(A')$$

The event that all **23** people have **different** birthdays is the **same** as the event that **person 2** does not have the same birthday as **person 1**, and that **person 3** does not have the same birthday as either **person 1 or person 2**, and so on, and finally that person **23** does not have the same birthday as any **of persons 1 through 22**

the probability of Event **2** is **364/365**, as person 2 may have any birthday other than the birthday of **person 1**.

the probability of Event **5** is **361/365**, as person 5 may have any birthday other than the birthday of **person 1,2,3 and 4**.



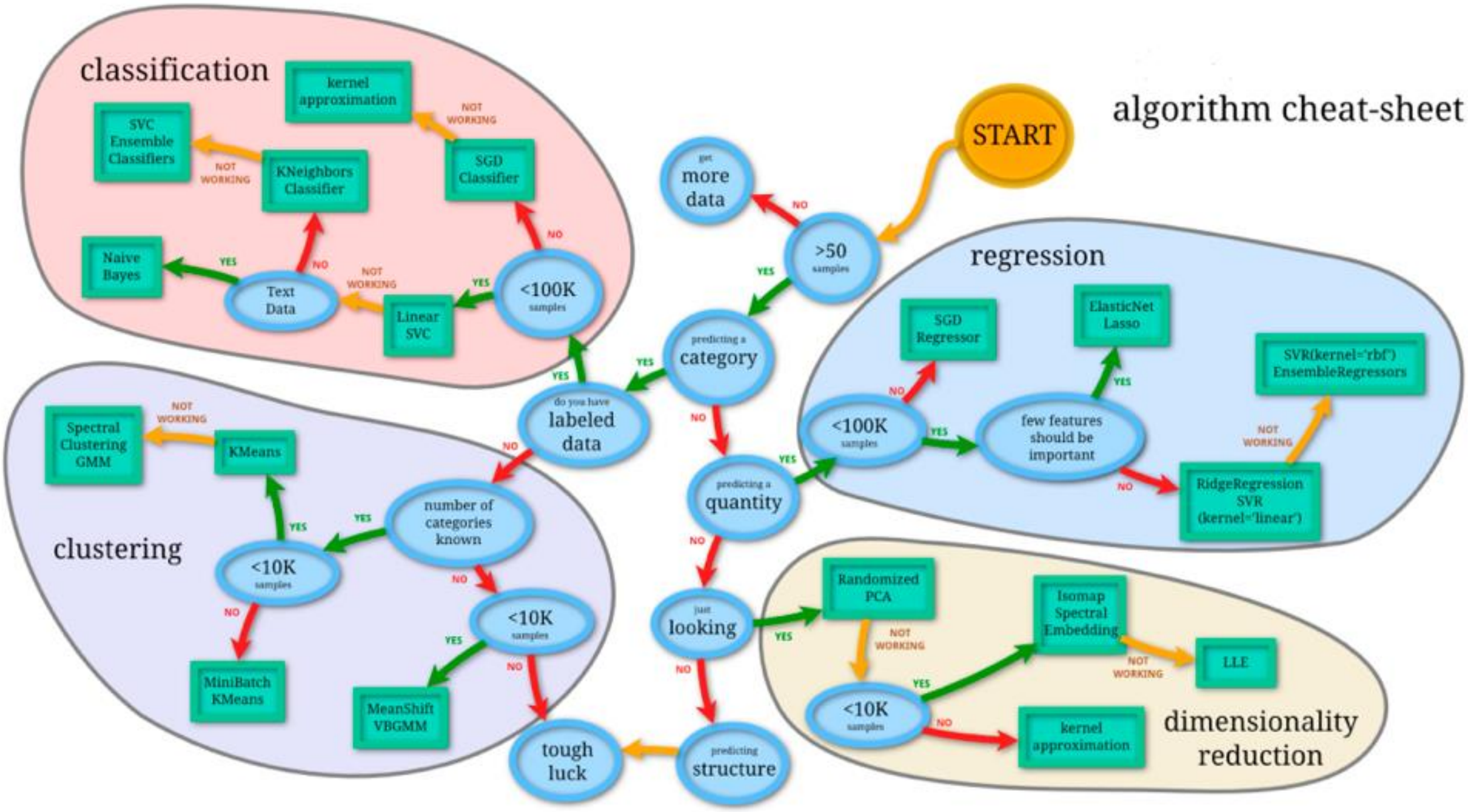
$$P(\acute{A}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \frac{361}{365} \times \ldots \frac{343}{365} = 0.492703$$

$P(A) = 1 - 0.492703 = 0.507297$ (**50.7297%**)

MACHINE LEARNING SYNOPSIS

Purpose	Problem Space	ML Technique
Anomaly Detection	more features, aggressive boundary	One-class SVM
	less features, fast training	PCA-based anomaly detection
Prediction	Linear model, fast training	Linear regression
	Linear model, small dataset	Bayesian linear regression
	Accuracy, long training time	Neural network regression
	Accuracy, fast training	Decision forest regression
	Predict event counts	Poisson regression
	Accuracy, fast training, large memory	Boosted decision tree regression
Discovering structure	Clustering	K-means
Classification (two class, multi-class)	Fast training, linear model	Logistic regression
	Accuracy, long training time	Neural network
	Accuracy, fast training	Decision forest, Decision jungle
	More features	Deep SVM
Recommendation	What you may also like	Association rules, matchbox
Text Analytics	NER, Sentiment Analysis	Rule based, SVM
Computer Vision	Image recognition	CNN, OpenCV Library

STATISTICAL MODELS



WEEK 1 EXERCISE

- SAS Visual Analytics Set up
- Task 3 Demonstration
  - Date format issue