

# Doing Science : from Start to Finish (day 4)

Dr. Félix E. Rivera-Mariani

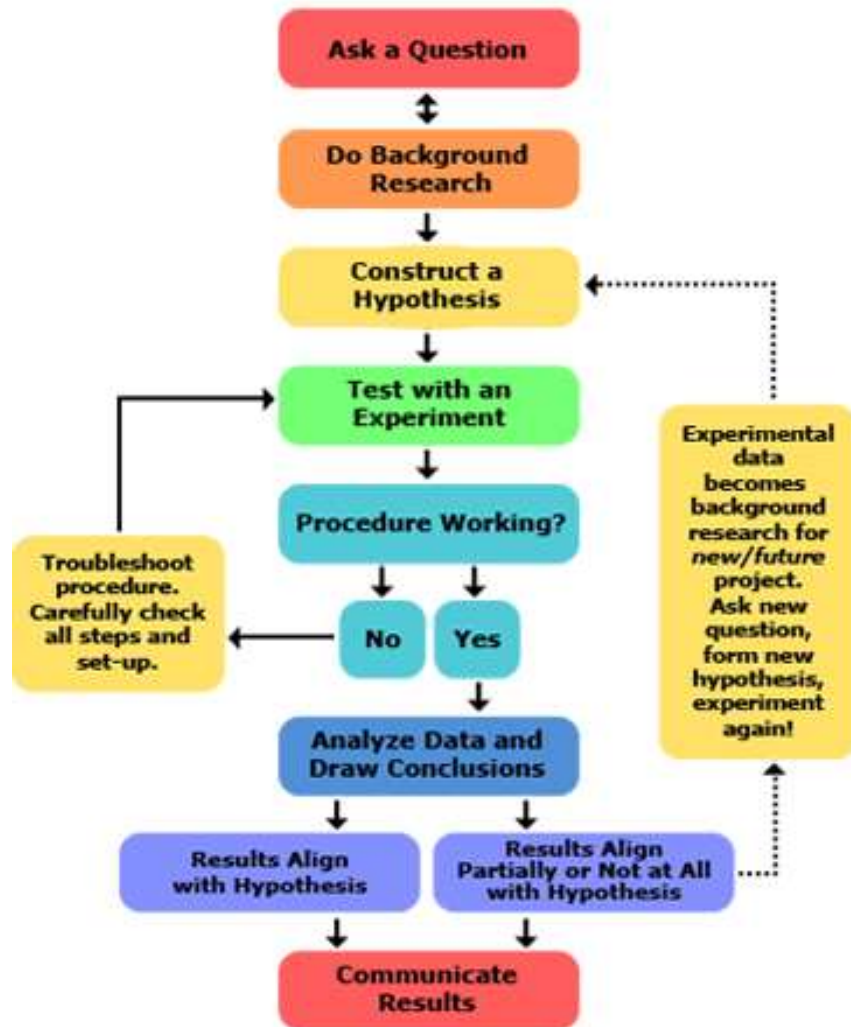
<https://github.com/friveramariani/DoingScienceWorkshops>

# Today's goals

---

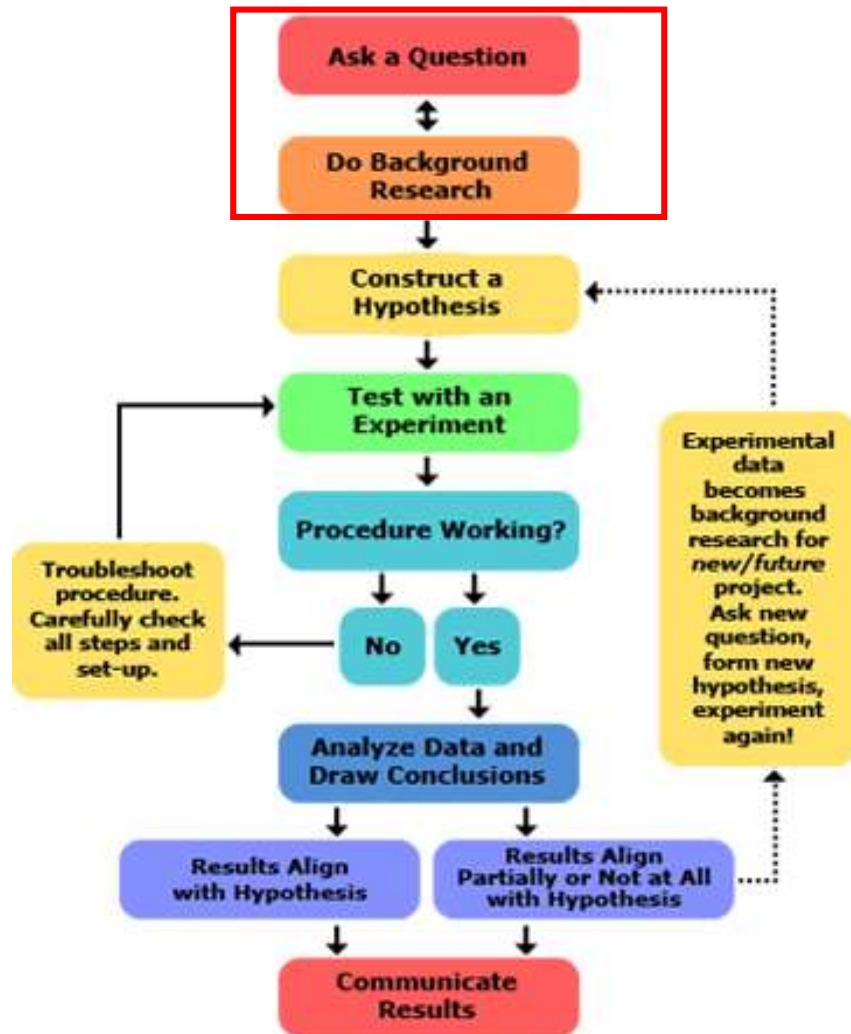
- Differentiate between good and poor structures in questions, hypothesis, and purpose
- Designing questions and hypotheses from existing data
- Implement questions related to the epicycle of data analysis
- Familiarize with the type of data and format for properly gathering data

# The Scientific Method: Review



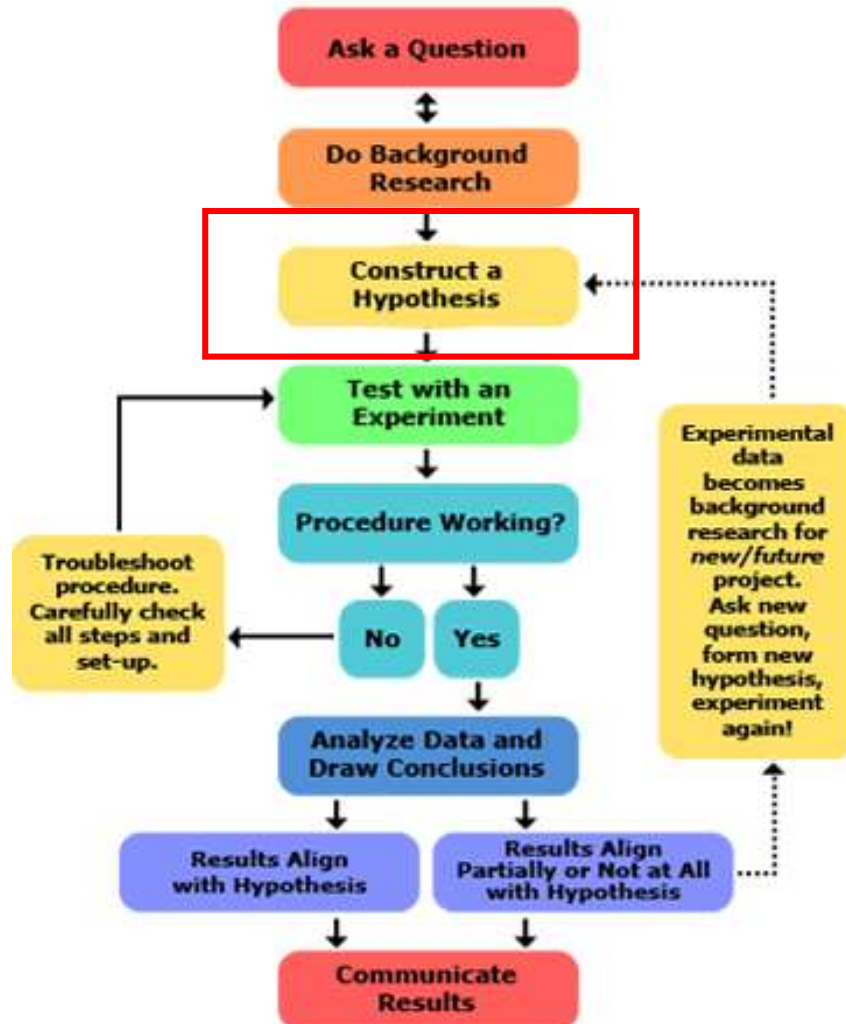
- Two directions between question and background research
- Epicycles with the experiment
- Epicycle between results and hypothesis

# The Scientific Method: *The question*



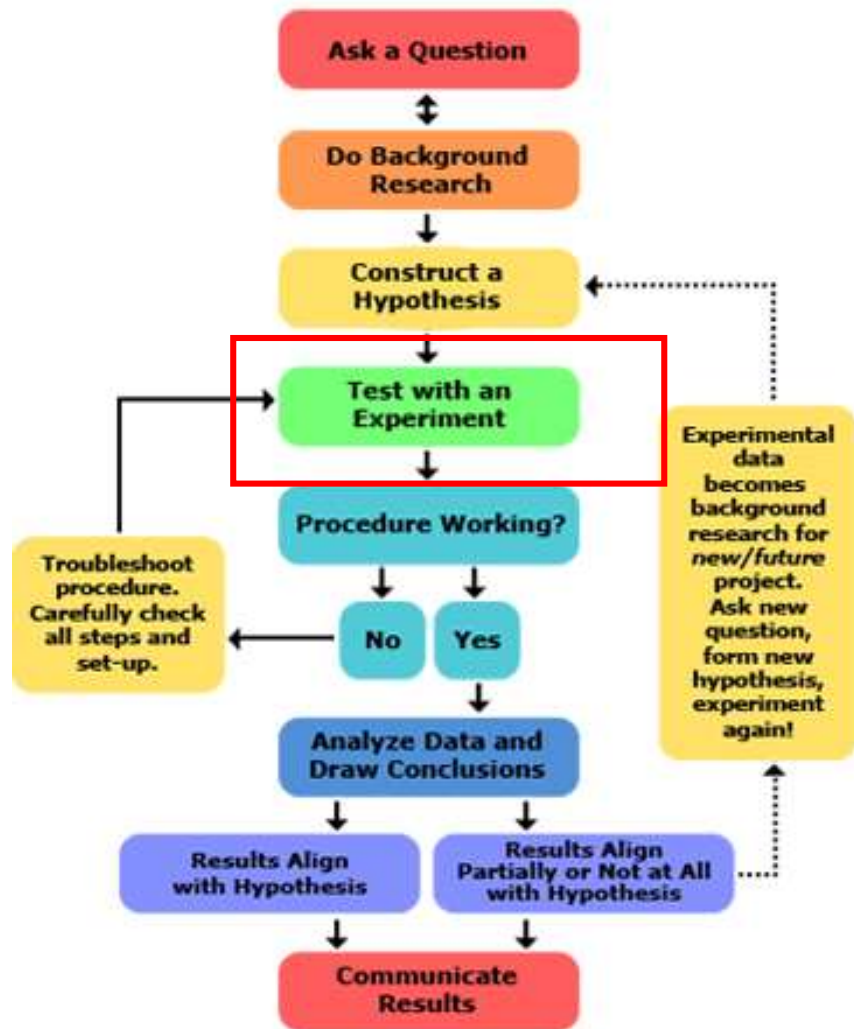
- What are the most frequent words in the students expectations' about the workshop "*Doing Science: from Start to Finish*"?

# The Scientific Method: *The Hypothesis*



The students expectations' of the Doing Science workshop will contain words that are in the title of the workshop.

# The Scientific Method



Name	Date modified	Type	Size
Students_Expectations1.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations2.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations3.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations4.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations5.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations6.txt	10/10/2016 6:06 PM	Text Document	1 KB
Students_Expectations7.txt	10/10/2016 6:06 PM	Text Document	1 KB
Students_Expectations8.txt	10/17/2016 1:41 PM	Text Document	1 KB
Students_Expectations9.txt	10/17/2016 1:42 PM	Text Document	1 KB
Students_Expectations10.txt	10/17/2016 1:42 PM	Text Document	1 KB
Students_Expectations11.txt	10/24/2016 7:00 A...	Text Document	1 KB
Students_Expectations12.txt	10/24/2016 10:01 ...	Text Document	1 KB

```
# Import libraries
# Import the most frequent words using the students_expectations
# Import libraries
library(ta)
library(SnowballC)
library(ggplot2)
library(WordCloud)
library(cluster)
library(fpac)

# Load directory
std_exp <- "C:/Users/Felix/Dropbox/Scientific/Doing_Science_Workshops/1_TextMining/students_expectations"

# Create corpus
std_exp_corpus <- Corpus(DirSource(std_exp))

# Convert files
inspect(std_exp_corpus)

# Create list for file names
# Remove punctuation from file names
std_exp_names <- ta_map(std_exp_corpus, removePunctuation)

# Remove numbers from file names
std_exp_names <- ta_map(std_exp_names, removeNumbers)

# All lowercase
std_exp_lower <- ta_map(std_exp_names, tolower)

# Remove common words in corpus and English words
std_exp_lower <- ta_map(std_exp_lower, removeWords, stopwords("english"))

# Remove white spaces
std_exp_names <- ta_map(std_exp_lower, stripWhiteSpace)

# Map file names to file content
std_exp_plain <- ta_map(std_exp_names, PlainTextDocument)

# Create document matrix
std_exp_mtx <- DocumentMatrix(std_exp_plain)

# Convert document matrix to TermDocumentMatrix
std_exp_mtx <- TermDocumentMatrix(std_exp_mtx)
```

```
# Exploratory analysis
# Create, order, and examine frequencies
freq_std_exp <- colSums(as.matrix(std_exp_mtx))
ord_freq_std_exp <- order(freq_std_exp)
freq_std_exp[head(ord_freq_std_exp)]

# Create, order, and examine frequencies of word sparse terms, Spanish terms
dms_std_nospr <- removeSparseTerms(std_exp_mtx, 0.50)
freq_std_nospr <- colSums(as.matrix(dms_std_nospr))
ord_freq_std_nospr <- order(freq_std_nospr)
freq_std_nospr[head(ord_freq_std_nospr)]

# Sort frequencies of non-sparse terms for Spanish from most to least frequent
std_freq_nospr_sort <- sort(colSums(as.matrix(dms_std_nospr)), decreasing=TRUE)

# Plot frequencies
# Create data frame
wf_df_std <- data.frame(word=names(std_freq_nospr_sort), freq=std_freq_nospr_sort)

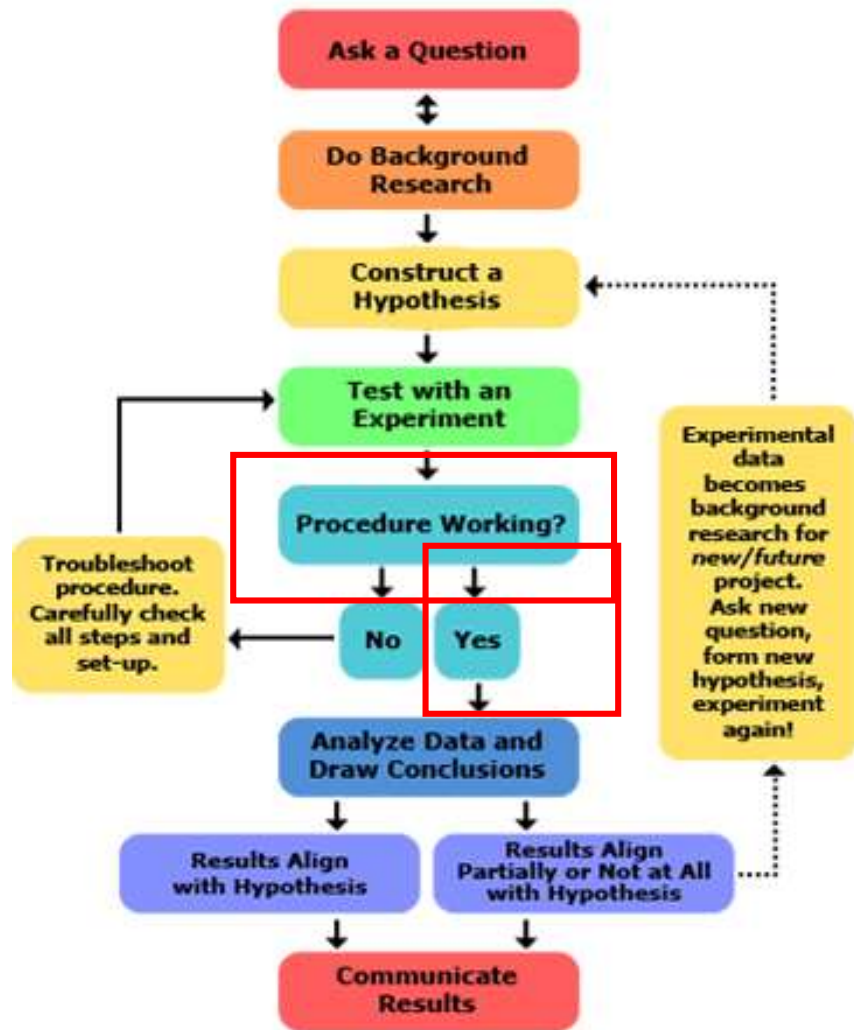
# Plot frequencies
plot_std_exp <- ggplot(subset(wf_df_std, freq>1), aes(word, freq)) + geom_bar(stat="identity")
plot_std_exp <- plot_std_exp + theme(axis.text.x=element_text(angle=45, adjust=1))

print(plot_std_exp)

# Word cloud
wordcloud(names(freq_std_nospr), freq_std_nospr, min.freq = 3,
          colors=brewer.pal(6, "Dark2"))
```



# The Scientific Method



Name	Date modified	Type	Size
Students_Expectations1.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations2.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations3.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations4.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations5.txt	10/10/2016 6:05 PM	Text Document	1 KB
Students_Expectations6.txt	10/10/2016 6:06 PM	Text Document	1 KB
Students_Expectations7.txt	10/10/2016 6:06 PM	Text Document	1 KB
Students_Expectations8.txt	10/17/2016 1:41 PM	Text Document	1 KB
Students_Expectations9.txt	10/17/2016 1:42 PM	Text Document	1 KB
Students_Expectations10.txt	10/17/2016 1:42 PM	Text Document	1 KB
Students_Expectations11.txt	10/24/2016 7:00 A...	Text Document	1 KB
Students_Expectations12.txt	10/24/2016 10:01 ...	Text Document	1 KB

```
## ---- Setup ----
## If you are the most frequent words among the students' expectations

# ---- Text file ----
# Load packages
library(tse)
library(SkimStats)
library(ggplot2)
library(wordcloud)
library(cluster)
library(fsc)

# Load directory
std_exp <- "~/Users/Felix/Dropbox/Scientific/Doing_Science_Workshops/1_TextMining/students_expectations"

# Create corpus
std_exp_corpus <- Corpus(DirSource(std_exp))

# Document files
inspect(std_exp_corpus)

# ---- Clean for plotting ----
# Remove punctuation from each file
std_exp_nospr <- tm_map(std_exp_corpus, removePunctuation)
# Remove numbers from each file
std_exp_nospr <- tm_map(std_exp_nospr, removeNumbers)
# All documents
std_exp_incor <- tm_map(std_exp_nospr, toLower)

# Remove common words to simplify the original words
std_exp_re_incor <- tm_map(std_exp_incor, removeWords, stopwords("english"))

# Remove words known
std_exp_re_incor <- tm_map(std_exp_re_incor, stripWhitespace)

# Load some state terms
std_exp_plain <- tm_map(std_exp_re_incor, PlainTextDocument)

# Create document matrix
std_exp_mtx <- DocumentMatrix(std_exp_plain)

# Create word-document matrix
std_exp_mtx_d <- TermDocumentMatrix(std_exp_plain)

## ---- Exploratory analysis ----
# Create, order, and examine frequencies
freq_stdexp <- colSums(as.matrix(std_exp_mtx_t))
ord_freq_stdexp <- order(freq_stdexp)
freq_stdexp[head(ord_freq_stdexp)]

# Create, order, and examine frequencies of word sparse terms, spanish terms
dms_stdnospr <- removeSparseTerms(std_exp_mtx, 0.50)
freq_std_nospr <- colSums(as.matrix(dms_stdnospr))
ord_freq_std_nospr <- order(freq_std_nospr)
freq_std_nospr[head(ord_freq_std_nospr)]

# Sort frequencies of word sparse terms for spanish from most to least frequent
std_freq_nospr_sort <- sort(colSums(as.matrix(dms_stdnospr)), decreasing=TRUE)

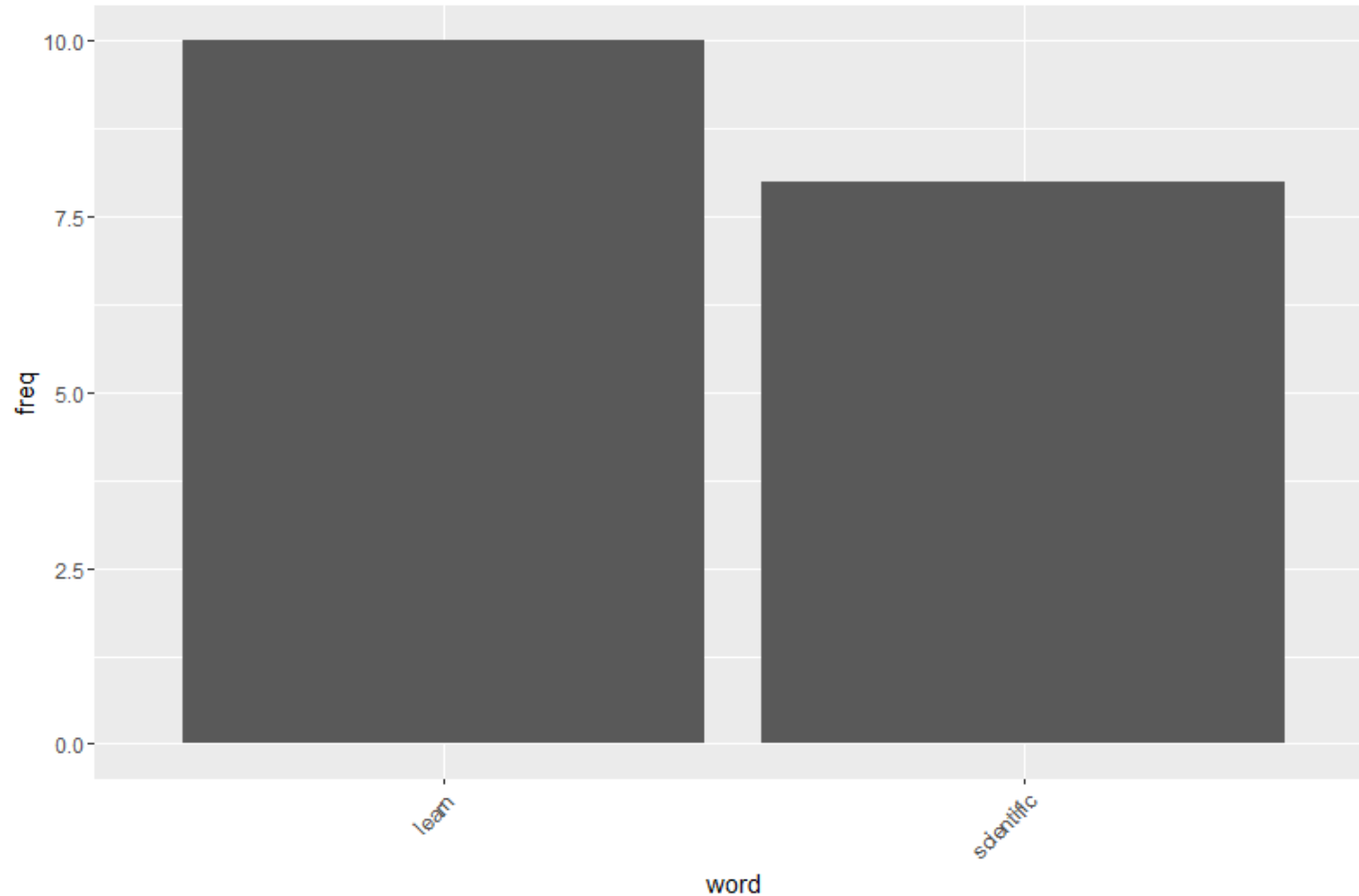
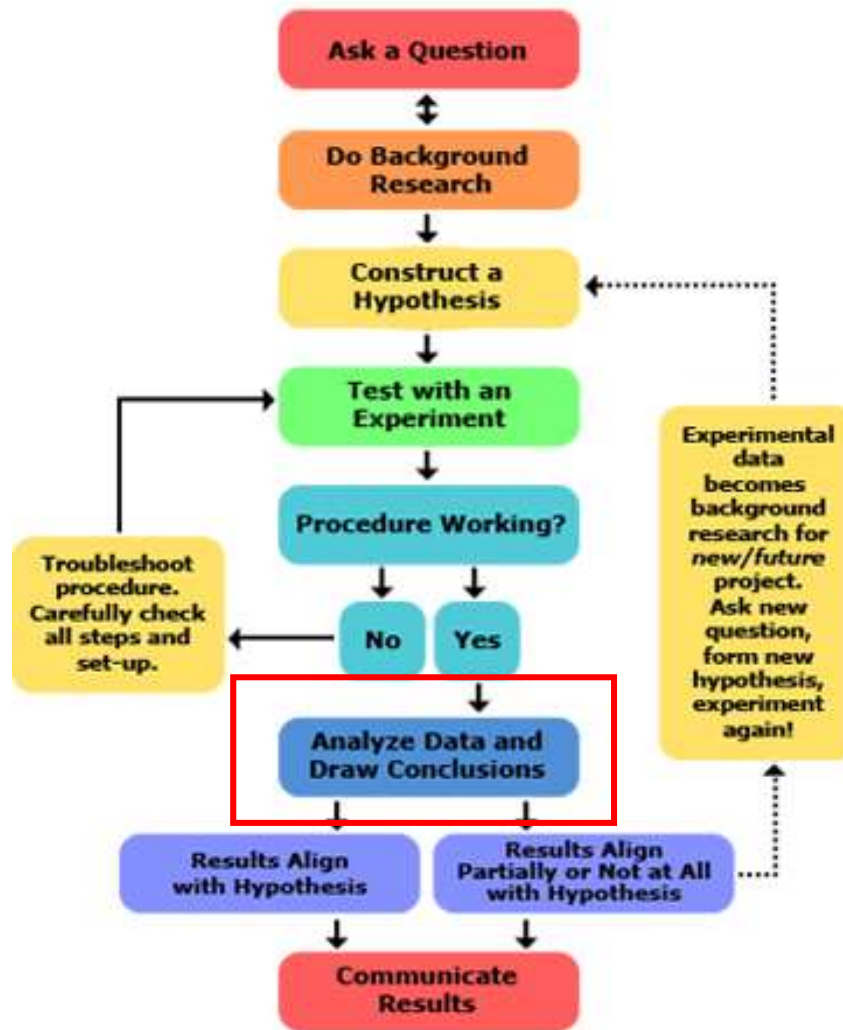
# ---- Plot frequencies ----
wf_df_std <- data.frame(word=names(std_freq_nospr_sort), freq=std_freq_nospr_sort)

plot_std_exp <- ggplot(subset(wf_df_std, freq>1), aes(word, freq)) + geom_bar(stat="identity")
plot_std_exp <- plot_std_exp + theme(axis.text.x=element_text(angle=45, adjust=1))

print(plot_std_exp)

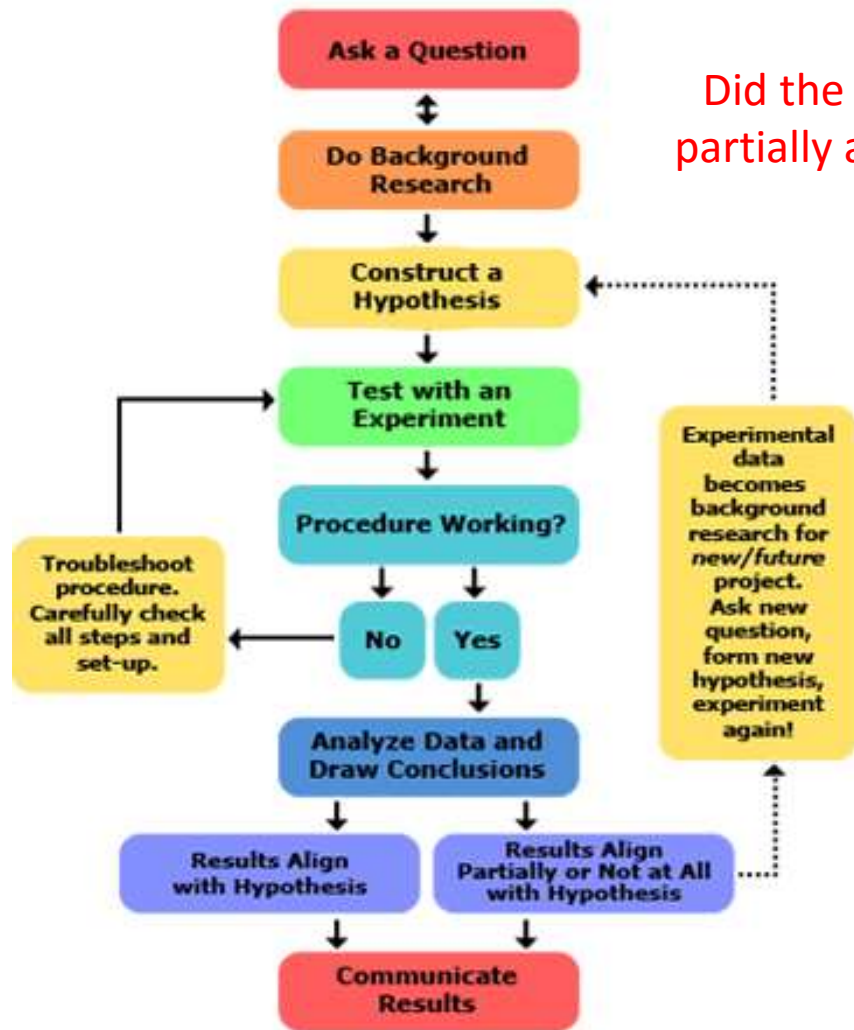
# ---- Word cloud ----
wordcloud(names(freq_std_nospr), freq_std_nospr, min.freq = 3,
          colors=brewer.pal(6, "Dark2"))
```

# The Scientific Method: *The Results*



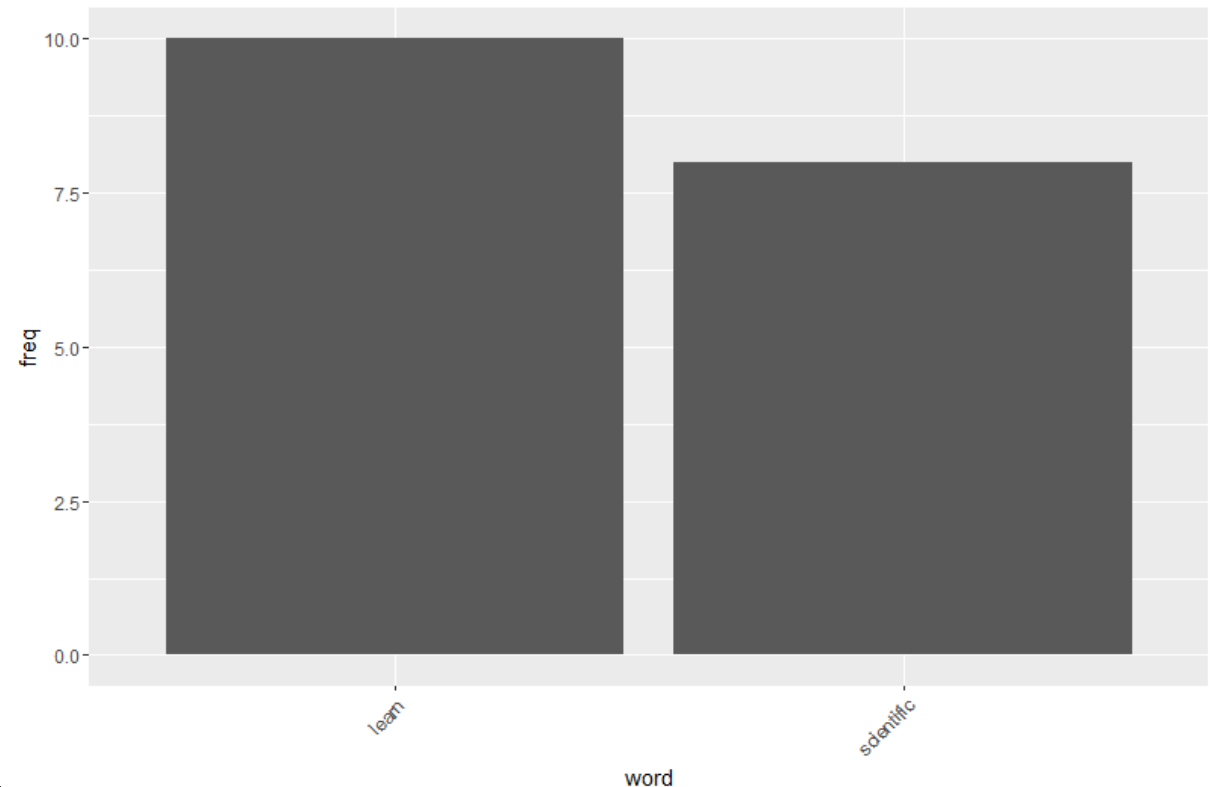


# The Scientific Method: *Align or partially align?*

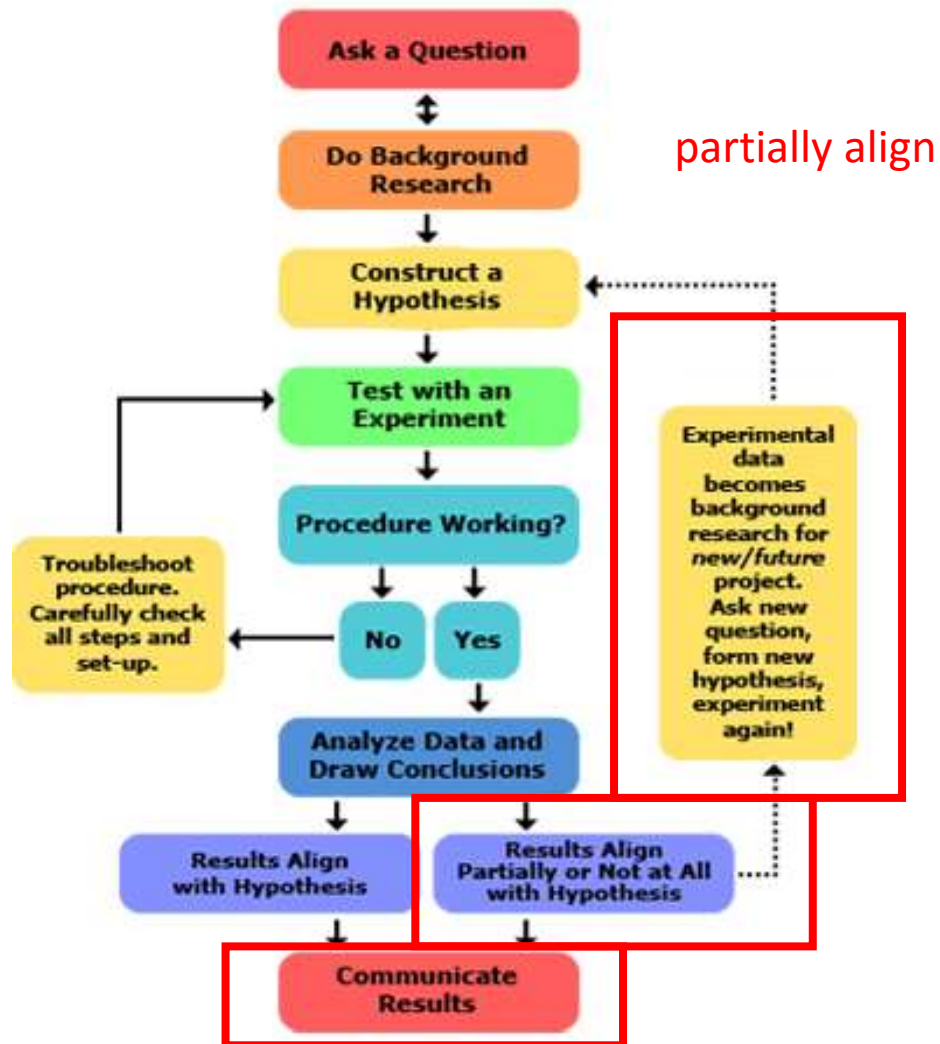


Did the results align,  
partially align, or not at  
all?

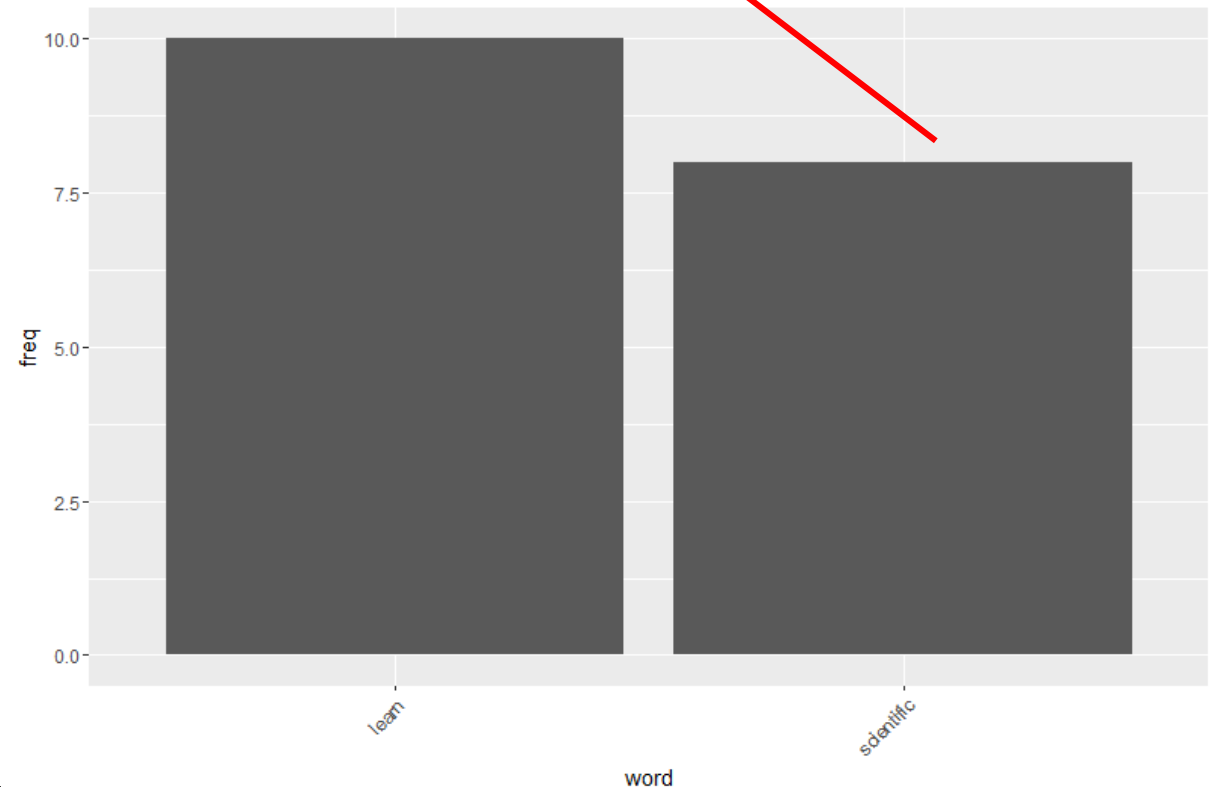
**Hypothesis:** The students expectations' of the Doing Science workshop will contain words that are in the title of the workshop.



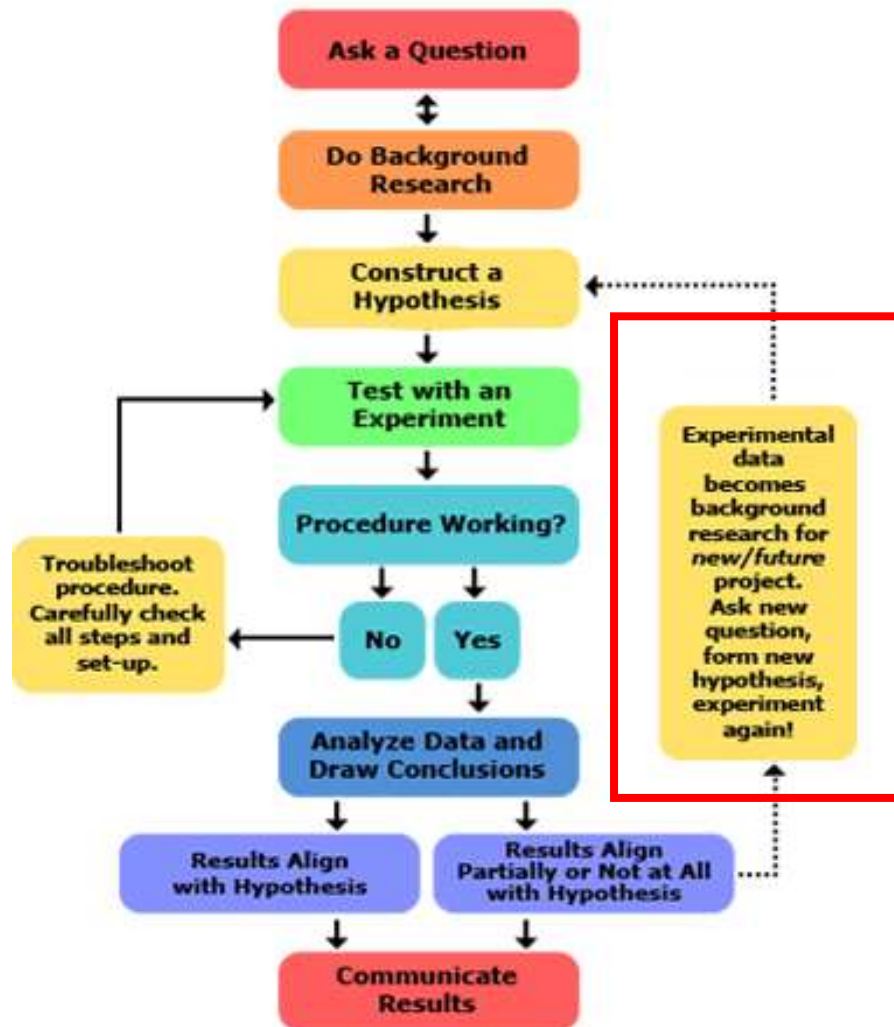
# The Scientific Method: *partially align*



Hypothesis: The students expectations' of the Doing Science workshop will contain words that are in the title of the workshop.

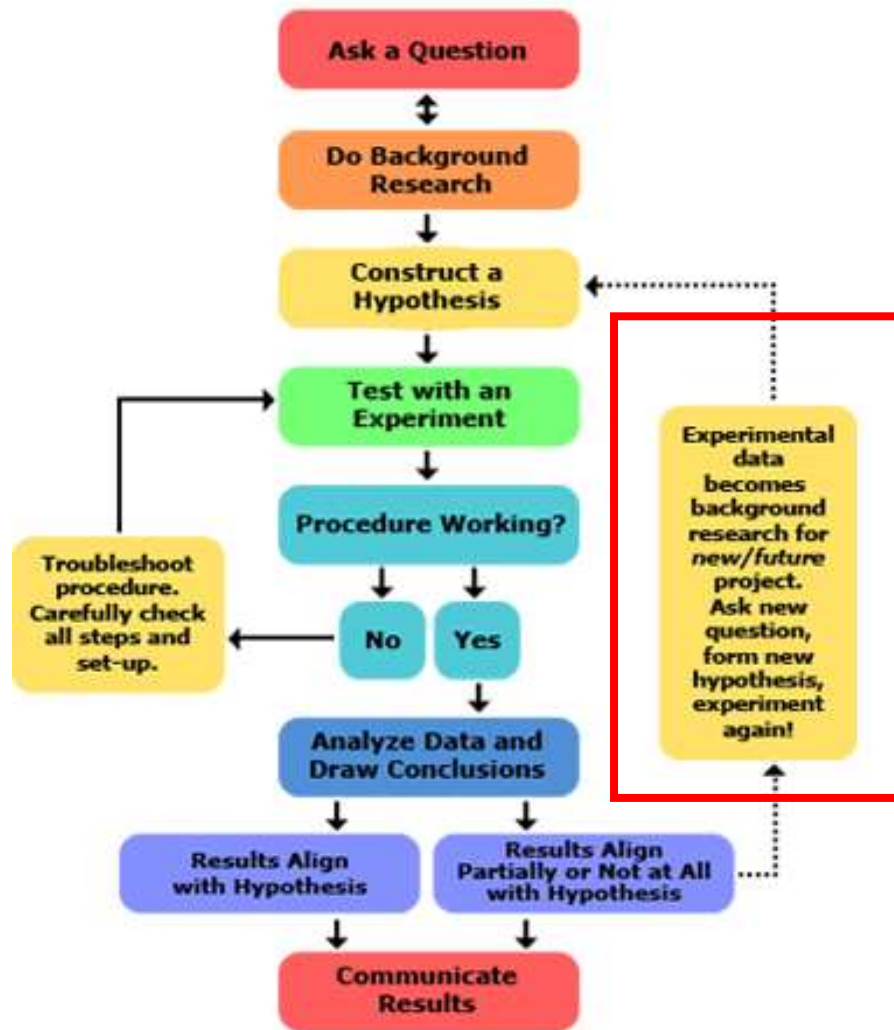


# The Scientific Method: *new question(s)*



If each of you sent a hypothesis, and then resubmitted a new hypothesis, *what could be a question that may have originated from our experiment about student's expectations?*

# The Scientific Method: *good questions?*



- a) Will the students' hypothesis 1 and hypothesis 2 align with the results of the students expectations' results?
- b) Will students hypotheses (hypothesis 1 vs hypothesis 2) share commo words?
- c) Will students' hypothesis 1 have higher number of letters than hypothesis 2?
- d) Will students' hypothesis 2 be better than hypothesis 1?

# The Scientific Method: *good questions?*

---

Meet the criteria:

1. Specific
2. Testable

~~a) Will the students' hypothesis 1 and hypothesis 2 align with the results of the students expectations' results?~~

b) Will students hypotheses (hypothesis 1 vs hypothesis 2) share commo words?

c) Will students' hypothesis 1 have higher number of letters than hypothesis 2?

~~d) Will students' hypothesis 2 be better than hypothesis 1?~~

# The Scientific Method: *hypothesis from the question*

---

Using the same words used already in these two questions, design a hypothesis for each question

- b) Will students hypotheses (hypothesis 1 vs hypothesis 2) share common words?
- c) Will students' hypothesis 1 have higher number of letters than hypothesis 2?



# The Scientific Method: *hypothesis from the question*

---

Using the same words used already in these two questions, design a hypothesis for each question

- b) Students' hypothesis 1 vs hypothesis 2 will share common words.
- c) Students' hypothesis 1 will have higher number of letters than hypothesis 2.

Students' hypothesis 1 vs hypothesis 2 will share common words.

---

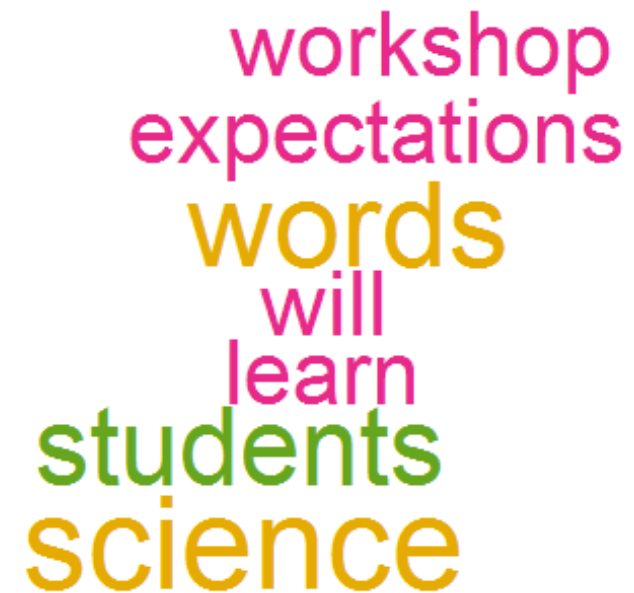
### Common words in hypothesis 1



A word cloud for hypothesis 1. The word 'students' is the largest and most prominent, colored orange. Above it, the word 'words' is smaller and colored blue. To the left of 'students', the word 'learn' is written vertically in green. To the right of 'students', the word 'workshop' is written vertically in green.

words  
students  
learn  
workshop

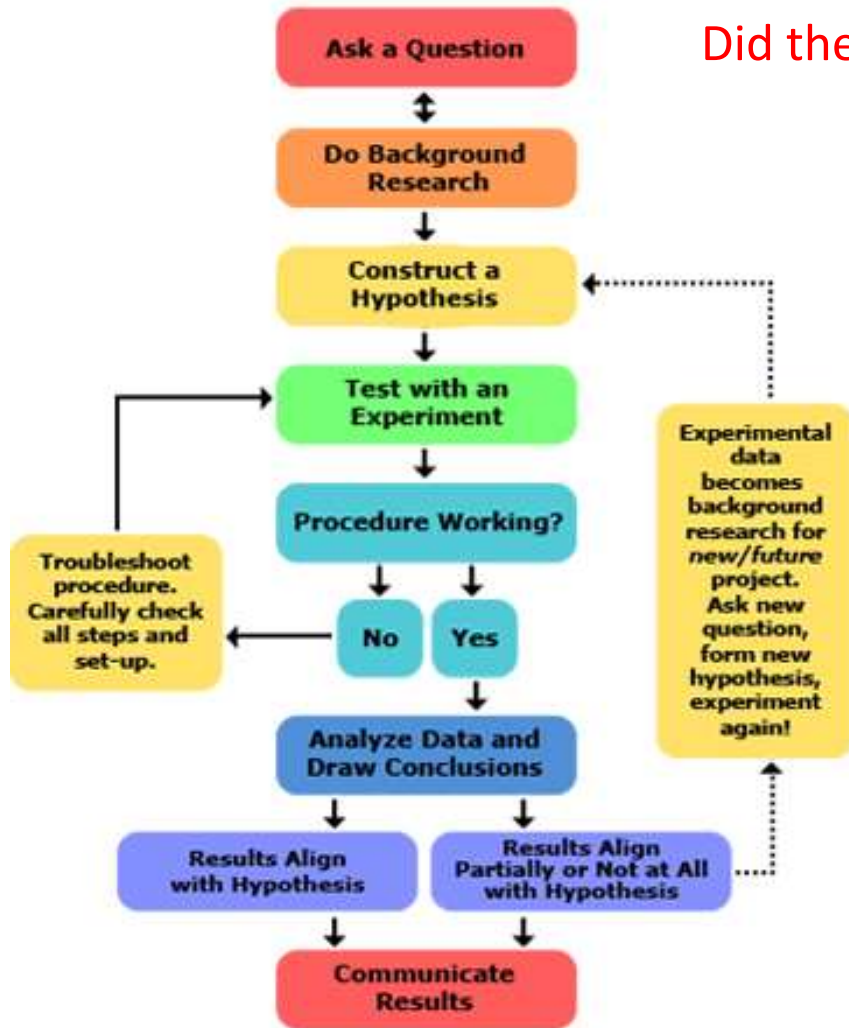
### Common words in hypothesis 2



A word cloud for hypothesis 2. The word 'workshop' is the largest and most prominent, colored pink. Below it, the word 'expectations' is also in pink. Below 'expectations', the word 'words' is in orange. Below 'words', the word 'will' is in pink. Below 'will', the word 'learn' is in pink. Below 'learn', the word 'students' is in green. At the bottom, the word 'science' is in orange.

workshop  
expectations  
words  
will  
learn  
students  
science

Students' hypothesis 1 vs hypothesis 2 will share common words.



Did the results align, partially align, or not at all with the hypothesis?

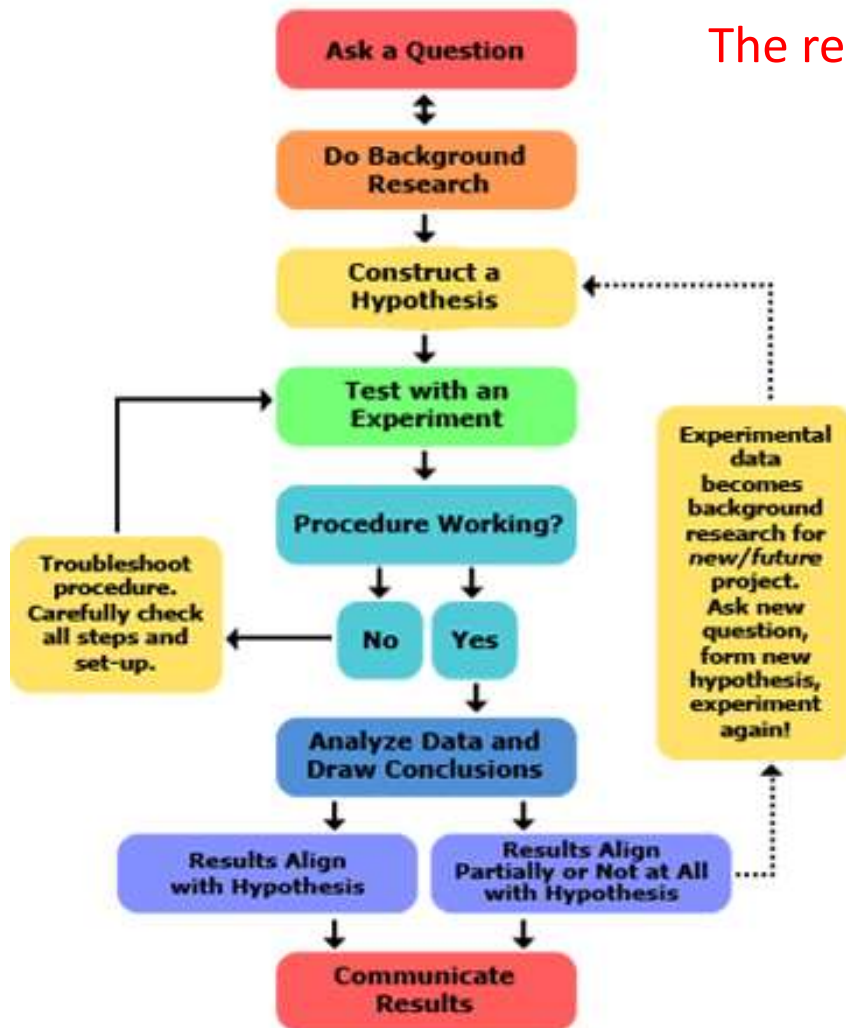
Words in hypothesis 1

words  
students  
learn  
workshop

Words in hypothesis 2

workshop  
expectations  
words  
will  
learn  
students  
science

Students' hypothesis 1 vs hypothesis 2 will share common words.



The results align, and there was an additional finding: students have more common words in common in hypothesis 2

Words in hypothesis 1

words  
students  
learn  
workshop

Words in hypothesis 2

workshop  
expectations  
words  
will  
learn  
students  
science

Students' hypothesis 1 will have higher number of letters than hypothesis 2.

---

Students' hypothesis 1 will have higher number of letters than hypothesis 2.

- Is this format correct for data analysis?

hyp1	hyp2
685	155
216	183
137	92
226	147
101	124
127	124
76	140
117	143
	78
	120



Students' hypothesis 1 will have higher number of letters than hypothesis 2.

- Is this format correct for data analysis?

No

hyp1	hyp2
685	155
216	183
137	92
226	147
101	124
127	124
76	140
117	143
	78
	120

Students' hypothesis 1 will have higher number of letters than hypothesis 2.

- What is the dependent variable?
- What is the independent variable?

hyp1	hyp2
685	155
216	183
137	92
226	147
101	124
127	124
76	140
117	143
	78
	120

Students' hypothesis 1 will have higher number of letters than hypothesis 2.

- What is the dependent variable?
  - Number of letters
- What is the independent variable?
  - Hypotheses

hyp1	hyp2
685	155
216	183
137	92
226	147
101	124
127	124
76	140
117	143
	78
	120

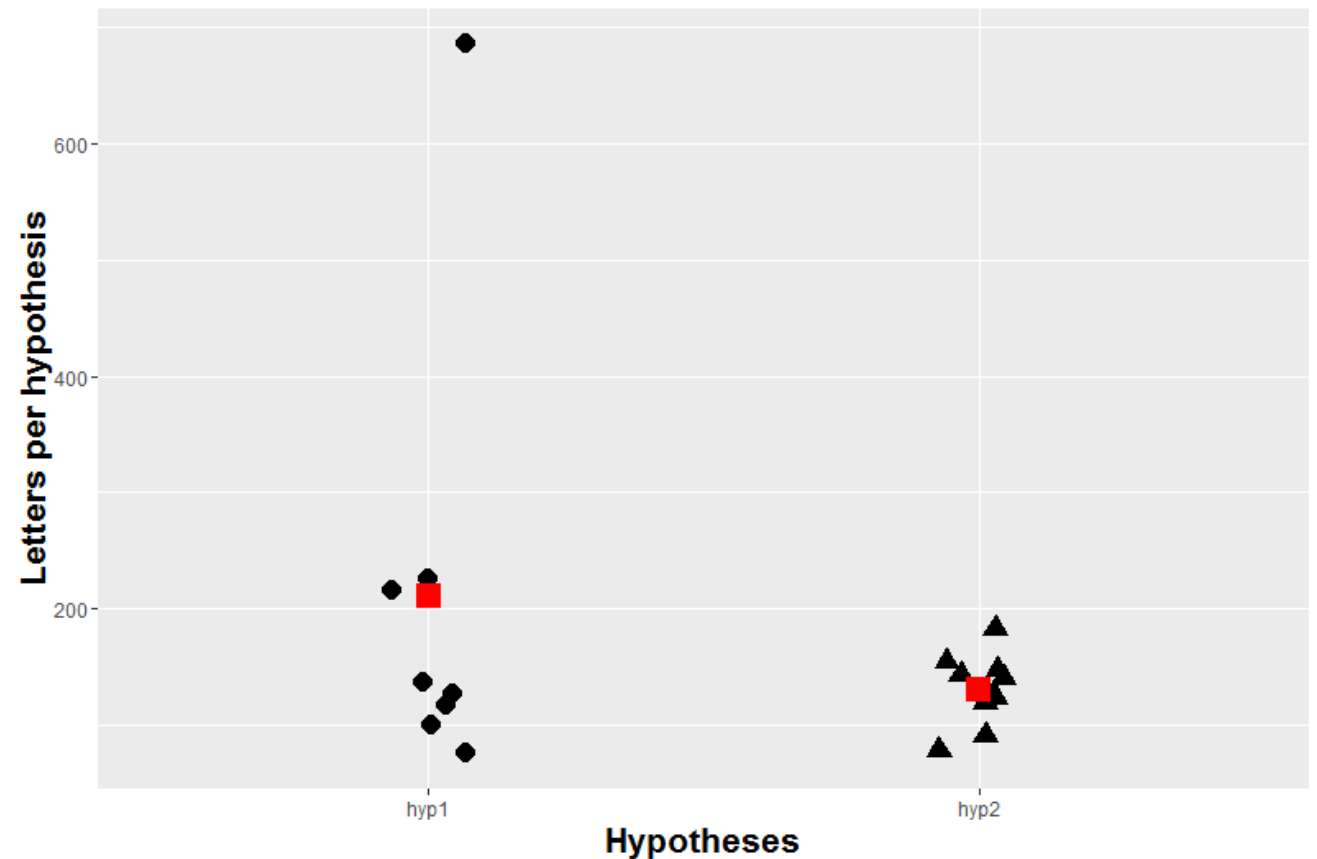
Students' hypothesis 1 will have higher number of letters than hypothesis 2.

- Variables = titles of columns
- Rows = observations for each variable

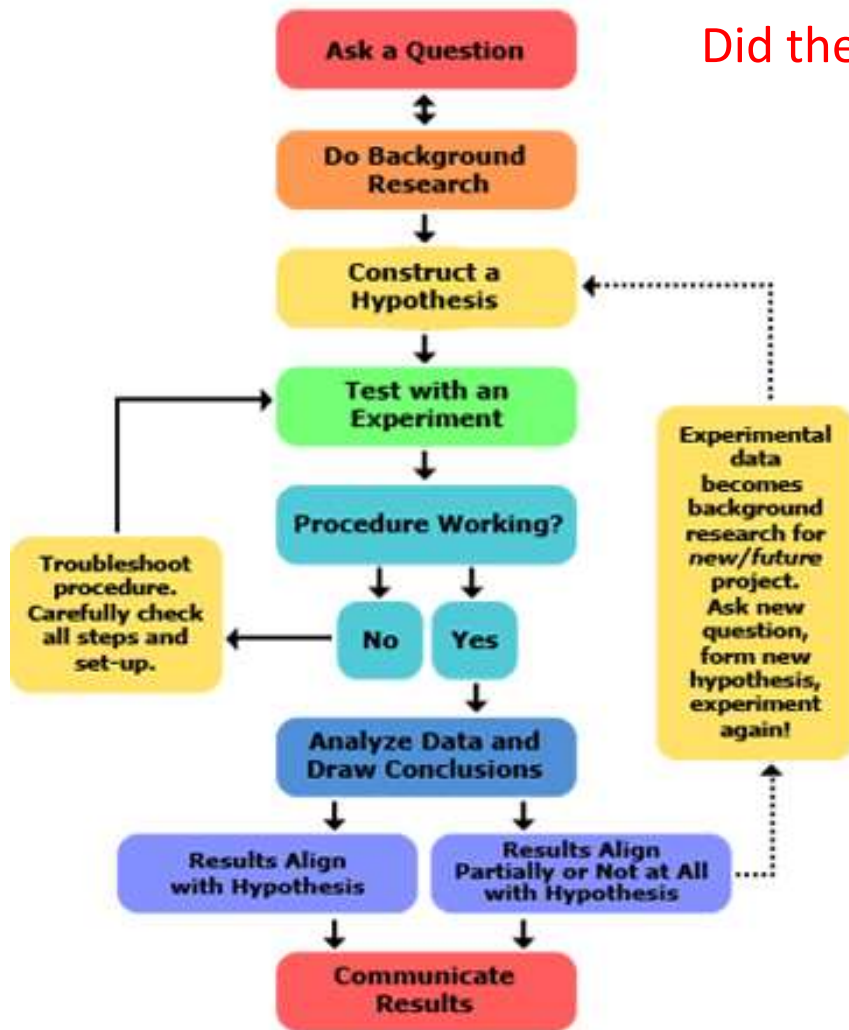
hypothesis	letters
hyp1	685
hyp1	216
hyp1	137
hyp1	226
hyp1	101
hyp1	127
hyp1	76
hyp1	117
hyp2	155
hyp2	183
hyp2	92
hyp2	147
hyp2	124
hyp2	124
hyp2	140
hyp2	143
hyp2	78
hyp2	120

Students' hypothesis 1 will have higher number of letters than hypothesis 2.

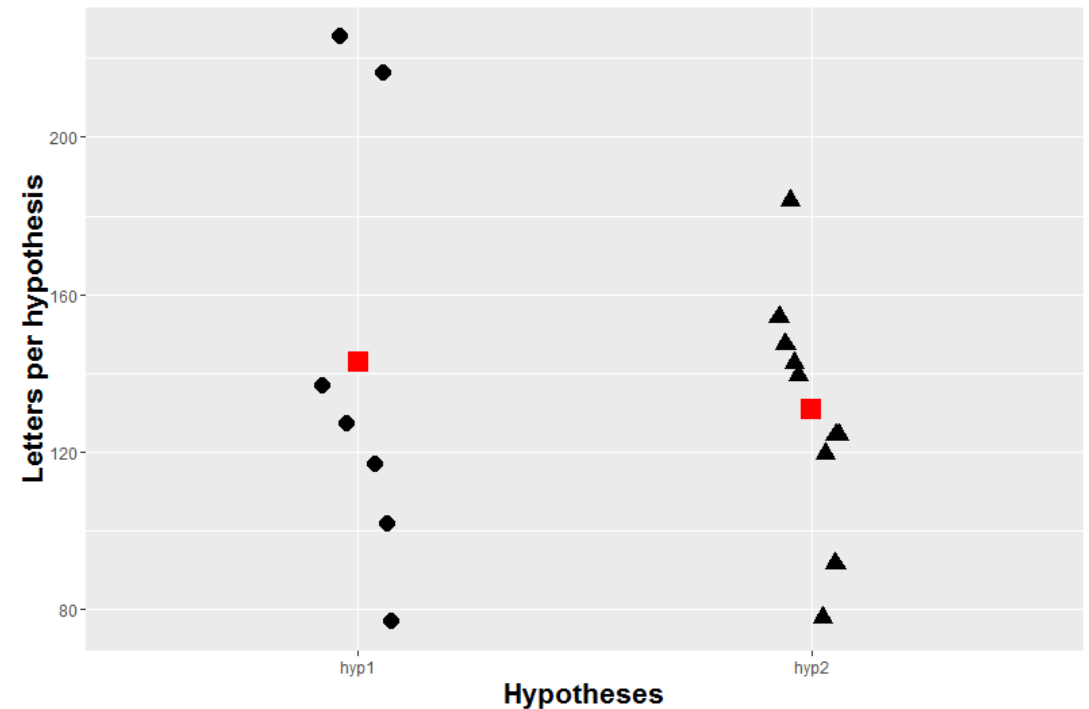
- There is an outlier that may prevent us from correctly interpreting the results



Students' hypothesis 1 will have higher number of letters than hypothesis 2.



Did the results align, partially align, or not at all with the hypothesis?





# How many variable are present in this table?

name	Quiz1	Quiz2	Quiz3	Quiz4	Quiz5	Quiz6	Topic	Outline	First Draft	Final Draft	Reading Assignments	Attendance	Bonus	Exam1	Exam2	Exam3	Grade
E	0	4	6	1.8			5	5	10		30	30	19	78.3	64.4		79.20%
K	9	5	7	6			5	5	10		30	30	21	73	91.8		91.50%
D	7	11	10	8.8			5	5	10		29	29.5	12.5	85.8	108.4		100.60%
J	3	5	6	6			5	5	10		30	30	16.5	85.8	86.8		90.30%
L	9.5	0	5	2			0	5	0		28	29	19	100.6	79.6		86.80%
B	10	7	8	5.2			5	5	10		30	30	21.5	75.7	83		90.80%
U	2	3	7	6			5	5	10		29.5	30	15	85.8	69.4		83.70%
J	12	8	9	7.5			5	5	10		30	30	13	82.1	75.5		89.70%
K	10	5	9	8.3			5	5	0		30	30	25.5	82.1	81.3		91%
Y	0	6	3	10.7			5	5	10		29.5	29.5	13.5	78.3	70.5		81.60%
M	8	5	5	7.8			5	5	10		29.5	29.5	18.5	96	89.5		96.50%
C	10	12	12	12			5	5	10		30	30	19.5	93.2	94.3		104.10%
B	0	7	0	4.7			5	5	10		29.5	29.5	11	82.1	80.4		82.60%
J	5	8	7	2.8			5	5	10		29.5	29	14.5	71	70.6		80.40%
D	8	5	6	3			5	5	10		30	29.5	19.5	78.4	79.5		87.20%
C	5.5	3	3	7.2			5	0	10		30	29.5	18.5	67.3	54.6		73%
J	5.5	7	9	10.7			5	5	10		30	30	17	96.9	103.8		103.10%
M	11.5	12	8	1.7			5	0	10		29	29.5	17	75.7	83.6		88.40%
D	8	5	7	1.7			5	5	10		30	30	23.5	82.1	74.4		88%
D	9.5	5	0 EX				5	5	10		30	29.5	18	78.4	77.6		86.50%
R	EX	4	5 EX				5	5	10		29	29.5	10	75.6	47.7		73.60%
D	11.5	6	8	9.6			5	5	10		30	30	18	85.8	72.2		91%
D	5	6	8	8.7			5	5	10		30	29.5	11.5	78.4	91		90%
E	9	6	12	9.2			5	5	10		30	30	12.5	111.7	87.2		102.40%
Daniel Ol	8	9	0	9			5	5	10		30	29.5	20	89.5	86.8		94.30%

# How many variable are present in this table?

1	2						3	4	5		6	7	8	9			10
name	Quiz1	Quiz2	Quiz3	Quiz4	Quiz5	Quiz6	Topic	Outline	First Draft	Final Draft	Reading Assignments	Attendance	Bonus	Exam1	Exam2	Exam3	Grade
E. N.	0	4	6	1.8			5	5	10		30	30	19	78.3	64.4		79.20%
K. V.	9	5	7	6			5	5	10		30	30	21	73	91.8		91.50%
D. t	7	11	10	8.8			5	5	10		29	29.5	12.5	85.8	108.4		100.60%
J. e	3	5	6	6			5	5	10		30	30	16.5	85.8	86.8		90.30%
L. z	9.5	0	5	2			0	5	0		28	29	19	100.6	79.6		86.80%
B. i	10	7	8	5.2			5	5	10		30	30	21.5	75.7	83		90.80%
U. e	2	3	7	6			5	5	10		29.5	30	15	85.8	69.4		83.70%
J. f	12	8	9	7.5			5	5	10		30	30	13	82.1	75.5		89.70%
K. a	10	5	9	8.3			5	5	0		30	30	25.5	82.1	81.3		91%
Y. a	0	6	3	10.7			5	5	10		29.5	29.5	13.5	78.3	70.5		81.60%
M. g	8	5	5	7.8			5	5	10		29.5	29.5	18.5	96	89.5		96.50%
C. F	10	12	12	12			5	5	10		30	30	19.5	93.2	94.3		104.10%
B. e	0	7	0	4.7			5	5	10		29.5	29.5	11	82.1	80.4		82.60%
J. n	5	8	7	2.8			5	5	10		29.5	29	14.5	71	70.6		80.40%
D. g	8	5	6	3			5	5	10		30	29.5	19.5	78.4	79.5		87.20%
C. e	5.5	3	3	7.2			5	0	10		30	29.5	18.5	67.3	54.6		73%
J. e	5.5	7	9	10.7			5	5	10		30	30	17	96.9	103.8		103.10%
M. l	11.5	12	8	1.7			5	0	10		29	29.5	17	75.7	83.6		88.40%
D. v	8	5	7	1.7			5	5	10		30	30	23.5	82.1	74.4		88%
D. i	9.5	5	0 EX				5	5	10		30	29.5	18	78.4	77.6		86.50%
R. a EX		4	5 EX				5	5	10		29	29.5	10	75.6	47.7		73.60%
D. l	11.5	6	8	9.6			5	5	10		30	30	18	85.8	72.2		91%
D. v	5	6	8	8.7			5	5	10		30	29.5	11.5	78.4	91		90%
E. r	9	6	12	9.2			5	5	10		30	30	12.5	111.7	87.2		102.40%
Daniel Ol	8	9	0	9			5	5	10		30	29.5	20	89.5	86.8		94.30%