# Agenda

| Start | End | Item |
|---|---|---|
| | | Core Concepts in Data Mining |
| | | Break |
| | | More R learning & EDA |
| | | |
| | | |

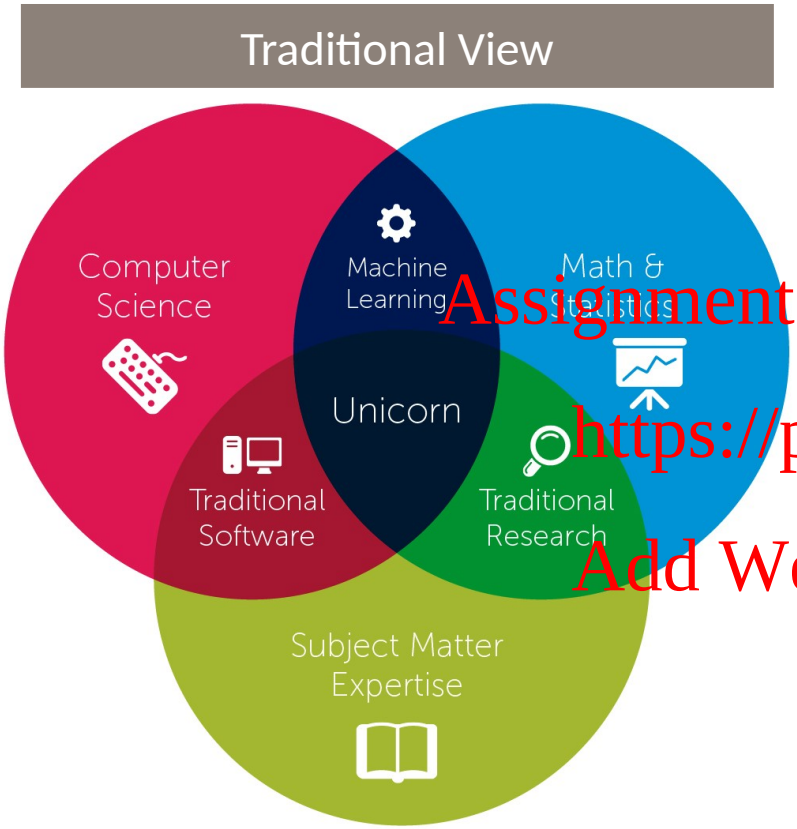
# Data Mining in this course

- **Business Analytics** – analyzing historical business data with basic math, SME rules, tallies, tables, summary statistics etc
- **Business Intelligence** – what has happened or is happening that can help current business decisions, often done with visuals, powerpoint, dashboards i.e. tableau
- **Data Mining** – includes machine learning (ML) & data science; applies more sophisticated methods to understand and predict business outcomes.



In this class we explore basic analytics, some business intelligence and ML methods in an effort to bring quantitative judgment to bear on business decisions.

# Data Mining & Science is almost always missing business acumen.








## Data Science

The study of information with the goal of extracting meaningful insights and creating actionable recommendations.

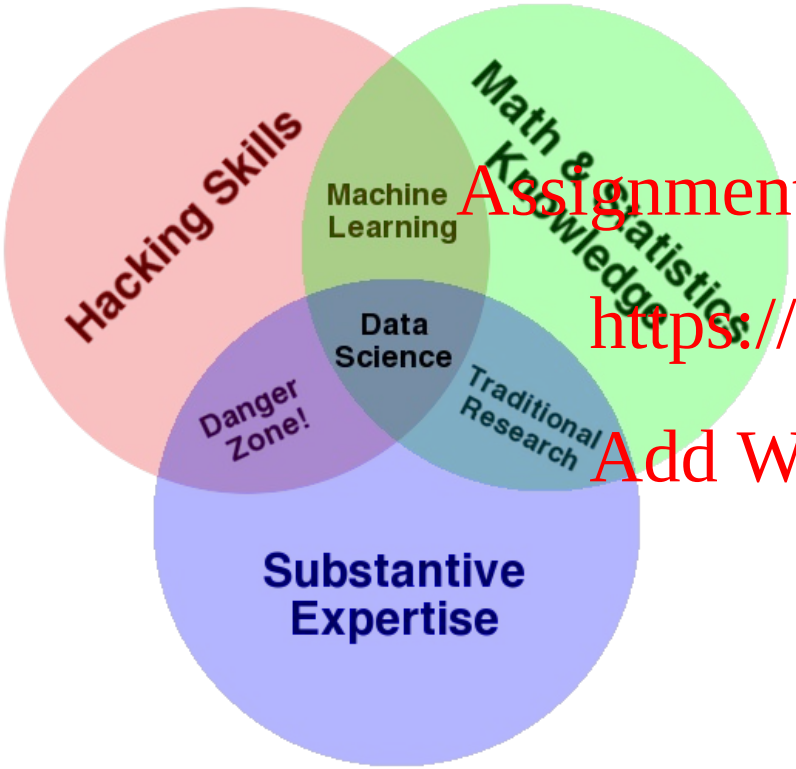
**often does not require "big data" or extremely exotic approaches to have a business impact*

## Machine Learning

An outgrowth of artificial intelligence, machine learning is the set of tools, methodologies and techniques allowing a computer to "learn" about a specific situations represented with data.

# Expertise is not confined to math or CS…but learning business implications.

Another Popular View




## Why hacking skills?

- Data science takes creativity
  - Art & Science
- A sprinkle of obsessive behavior to explain the data phenomenon

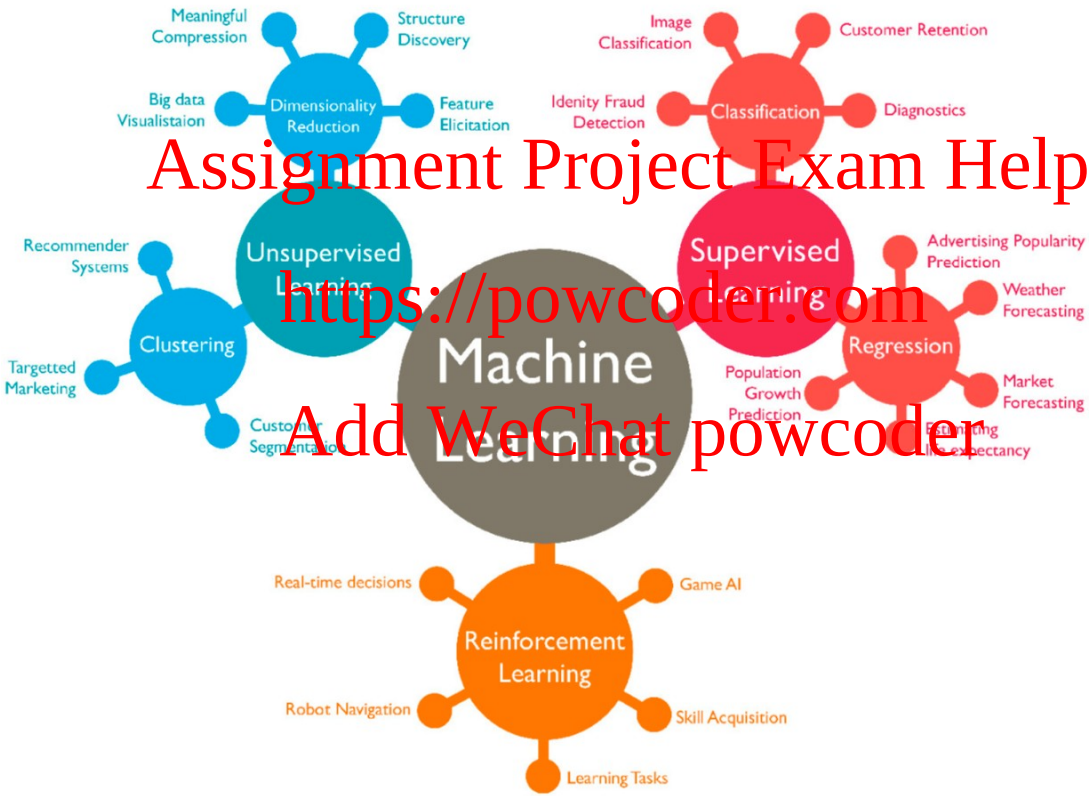


## My $0.02

Both diagrams have expertise yet it is often overlooked. Many data scientists are technically sound but lack business acumen or substantive expertise.

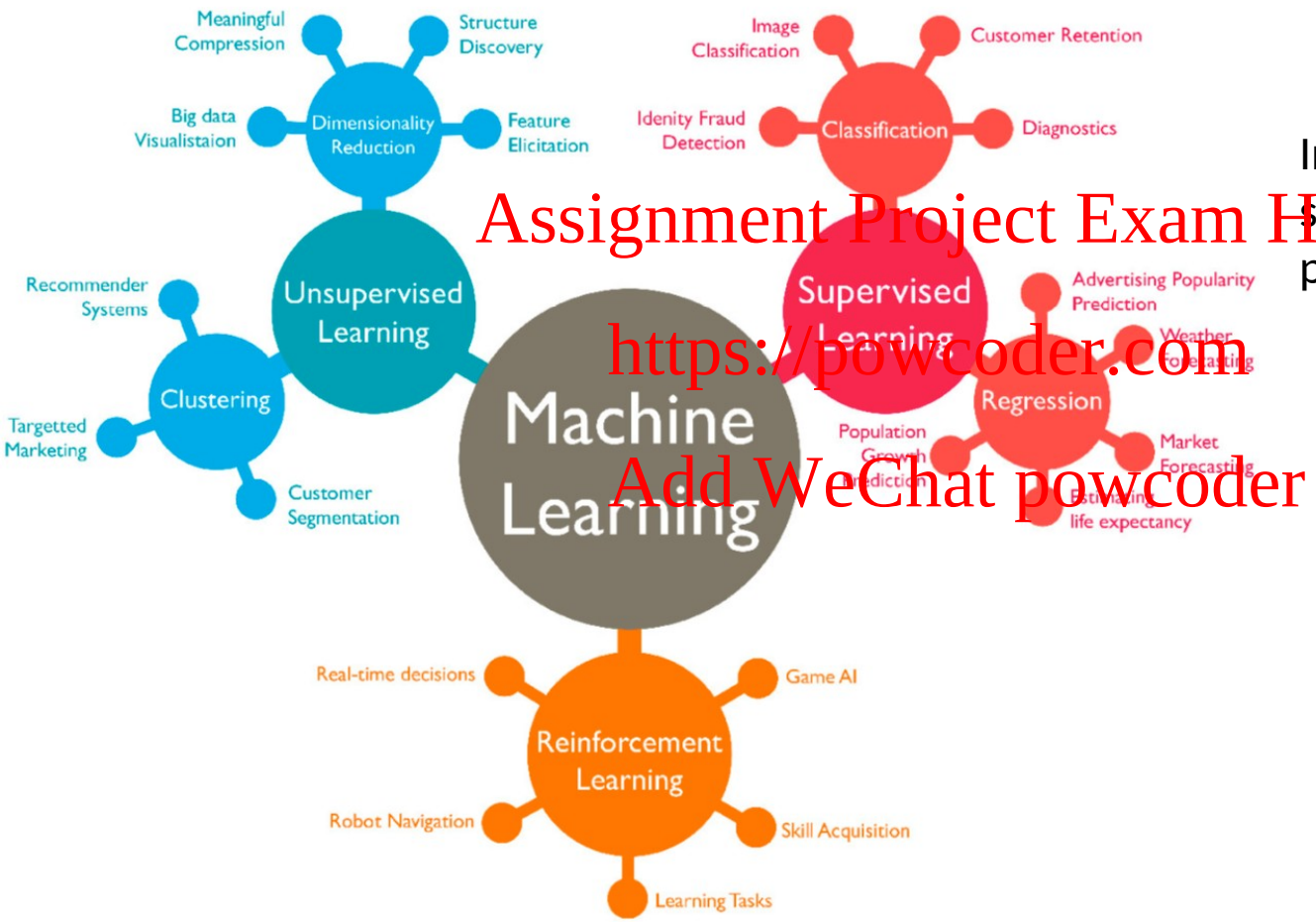
# What is Machine Learning anyway?

Machine Learning (or Statistical Learning) refers to a huge set of tools for understanding data.



# What is Machine Learning anyway?

Machine Learning (or Statistical Learning) refers to a huge set of tools for understanding data.



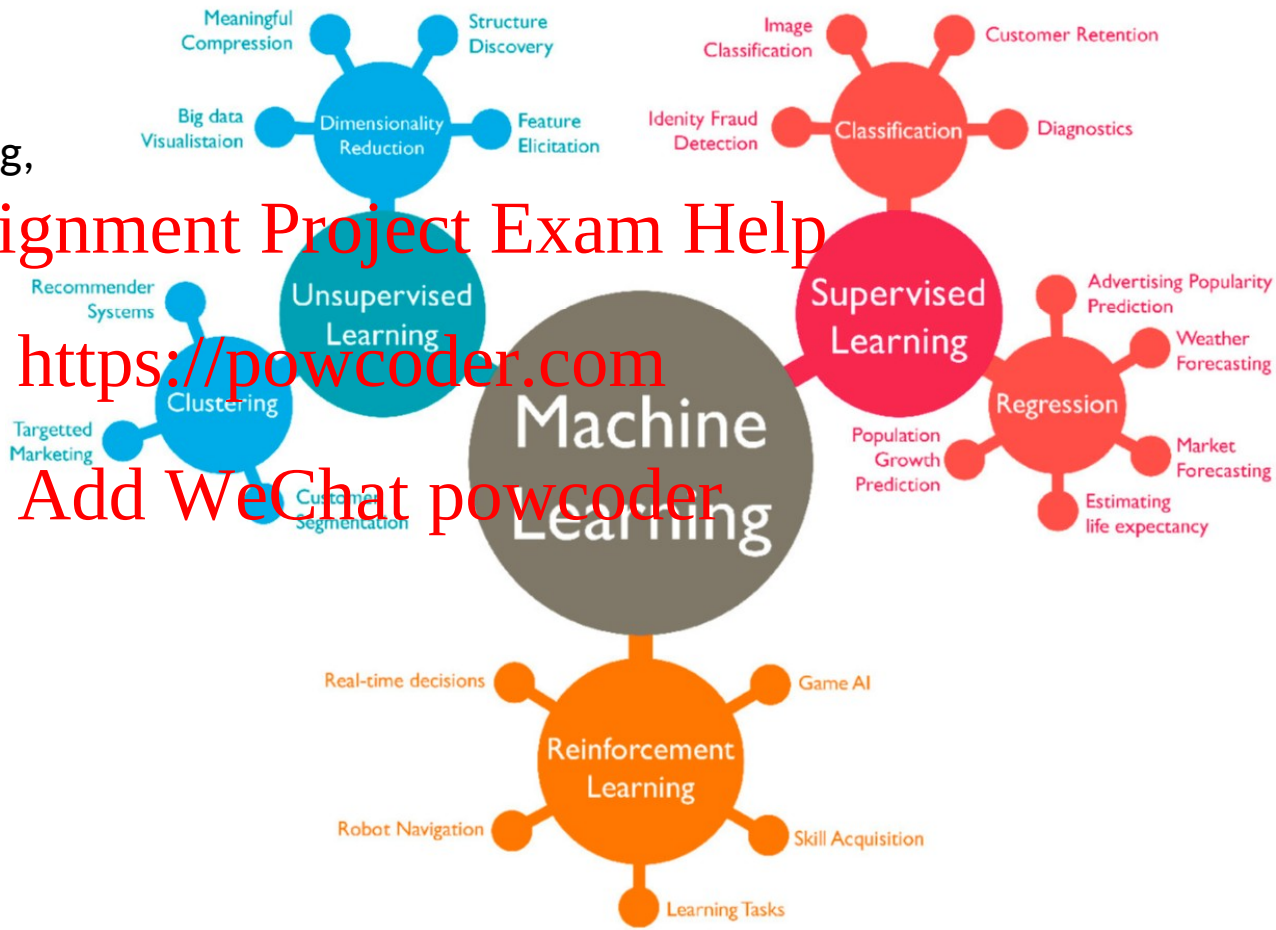



In Supervised Learning, a statistical model is built to predict a labeled output.

# What is Machine Learning anyway?

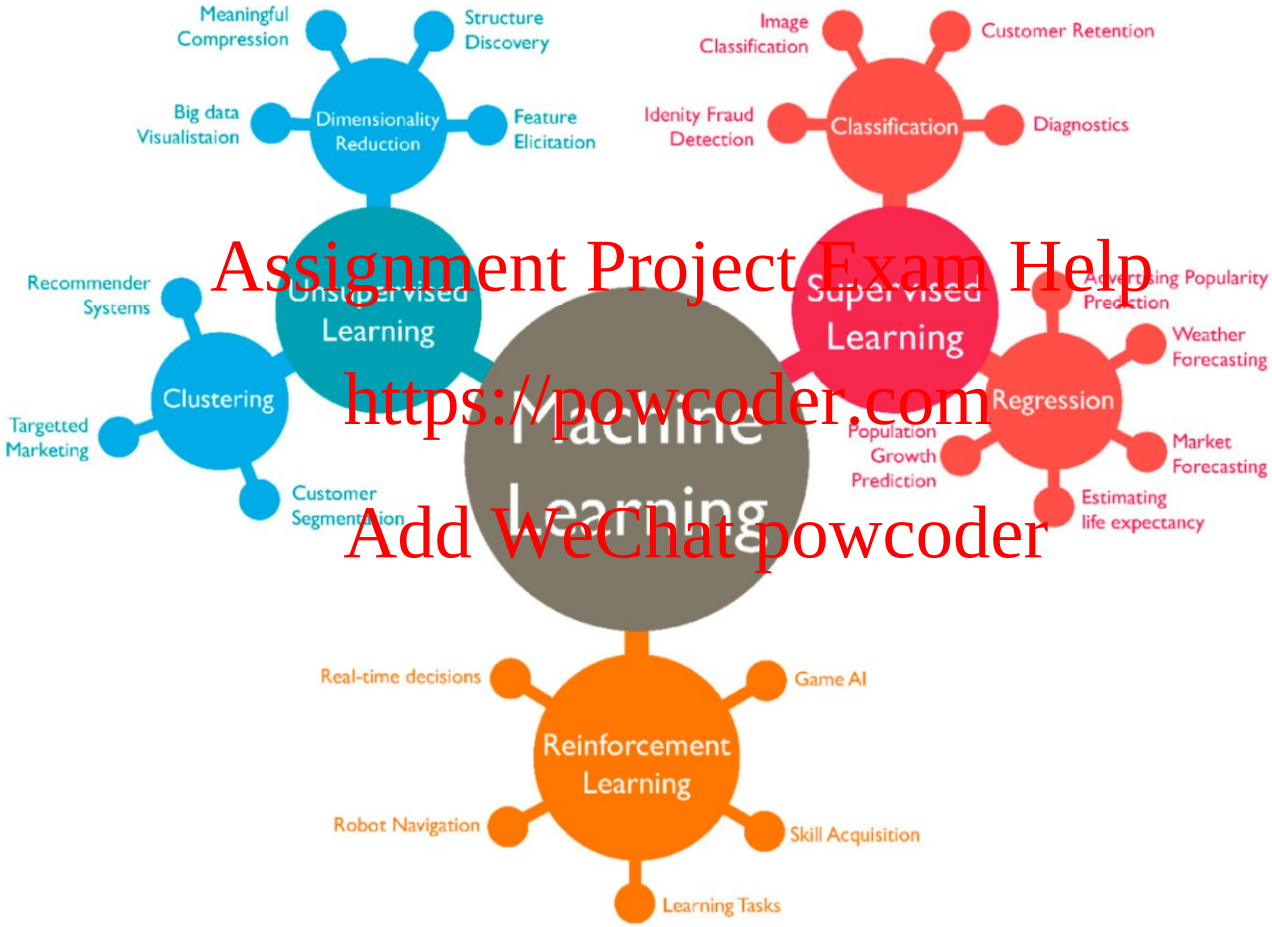
Machine Learning (or Statistical Learning) refers to a huge set of tools for understanding data.

With Unsupervised Learning, there are inputs but no labeled response, you are looking for patterns and cohorts in the data.



# What is Machine Learning anyway?

Machine Learning (or Statistical Learning) refers to a huge set of tools for understanding data.






In Reinforcement Learning, a model learns how to behave in a defined environment by performing actions, seeing the results, going back to iterate on the actions etc.

# Diagnosing & Defining a data mining project

## Questions to Ask:

- Is this a data mining problem? If so, what data would be helpful?
- What is the current state?
- What are the possible outcomes of the business scenario?
- How will the outcome of the data mining project be used?
- What is success for this project?

# Pitfalls

**Without asking these questions your efforts will:**

- Have scope-creep or never end!
- Be difficult to define success
- Be difficult to implement or have a limited impact

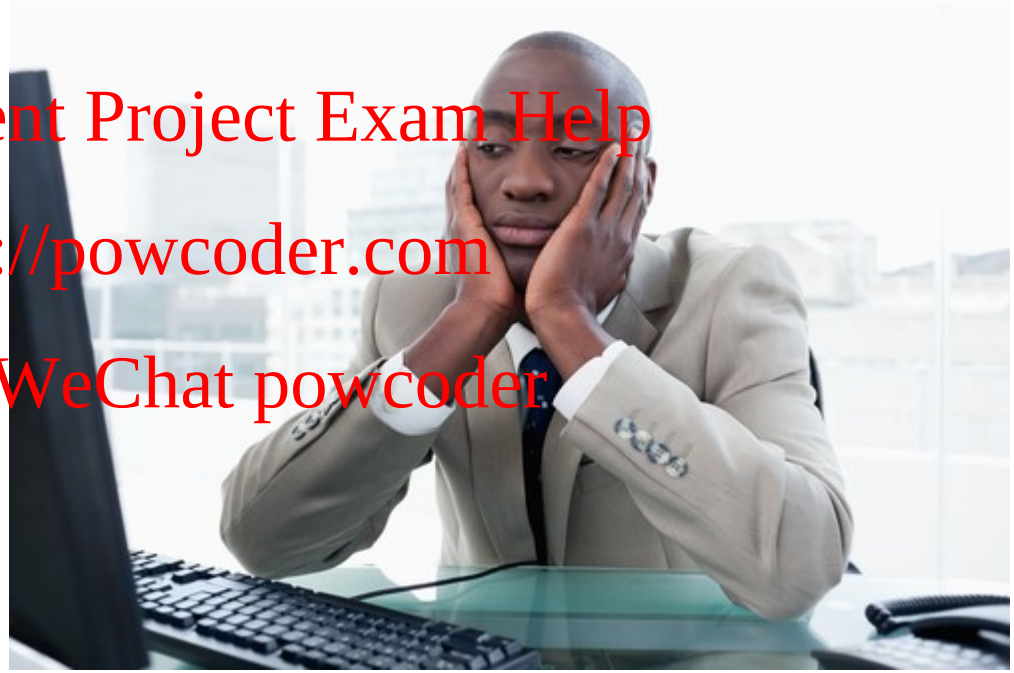


# Let’s do this for real…

Meet Dale

- Runs the analytics group at Busy-ness Corp a large conglomerate that makes, distributes and services corn dogs.

- He looks miserable because senior leaders make his job harder than needed.

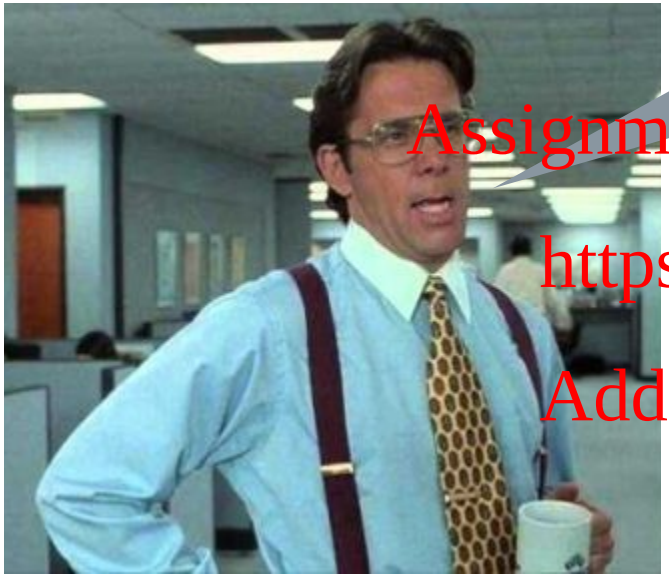




Let’s help Dale add some structure to his data mining projects.

# Let’s do this for real

Hey Dale, its me…the boss.
I read in the WSJ that everyone should be using blockchain. Should we?





Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?

# Let’s do this for real

Your Daleness, I work in marketing and want to look intelligent (hence the glasses). I want to do a mailing to prospective corndog eaters. Can you identify how many postcards we should send & ROI?



Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?

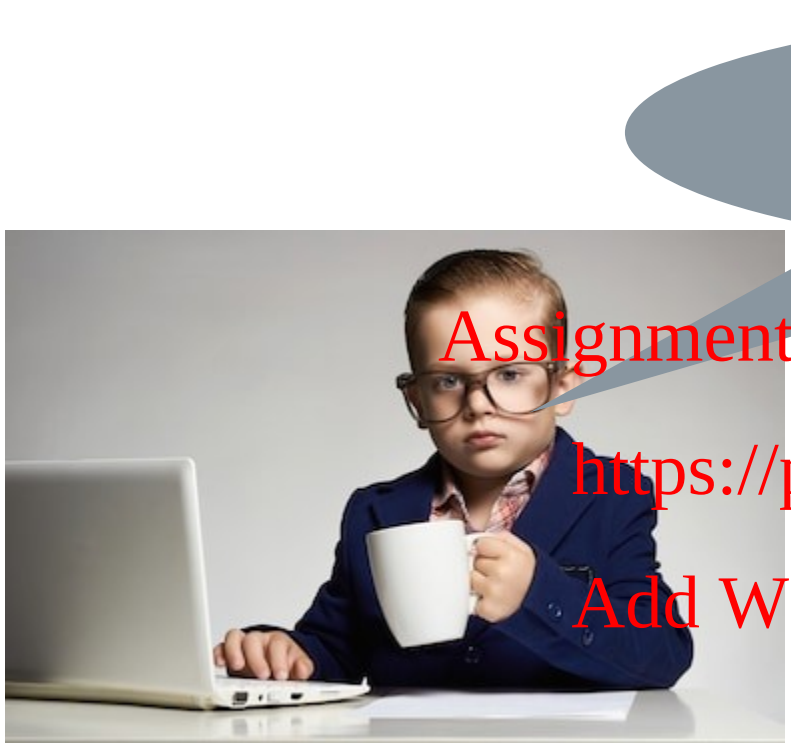
# Let’s do this for real

Dale-areno! I think we are getting a lot more calls than usual about defective corn dogs. Can you look into whether or not that’s true?



Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?

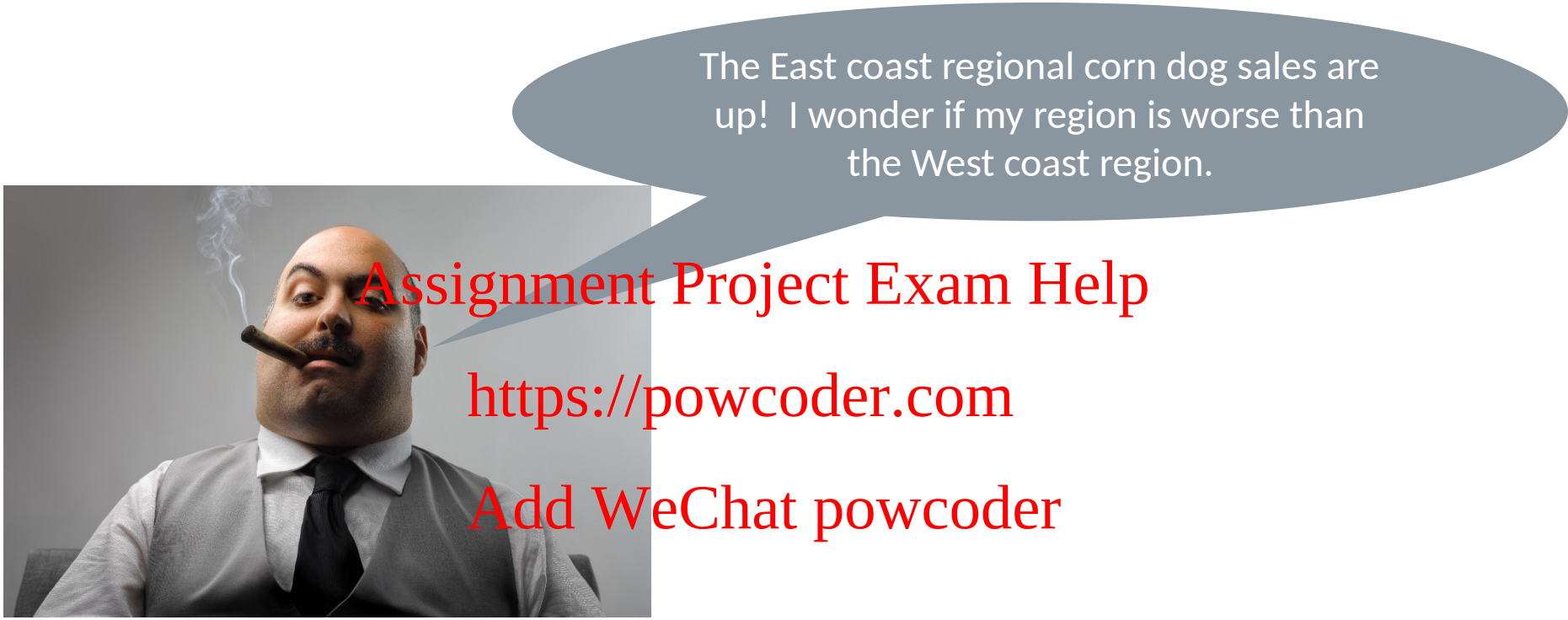
# Let’s do this for real






Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?

# Let’s do this for real






Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?



HARVARD UNIVERSITY

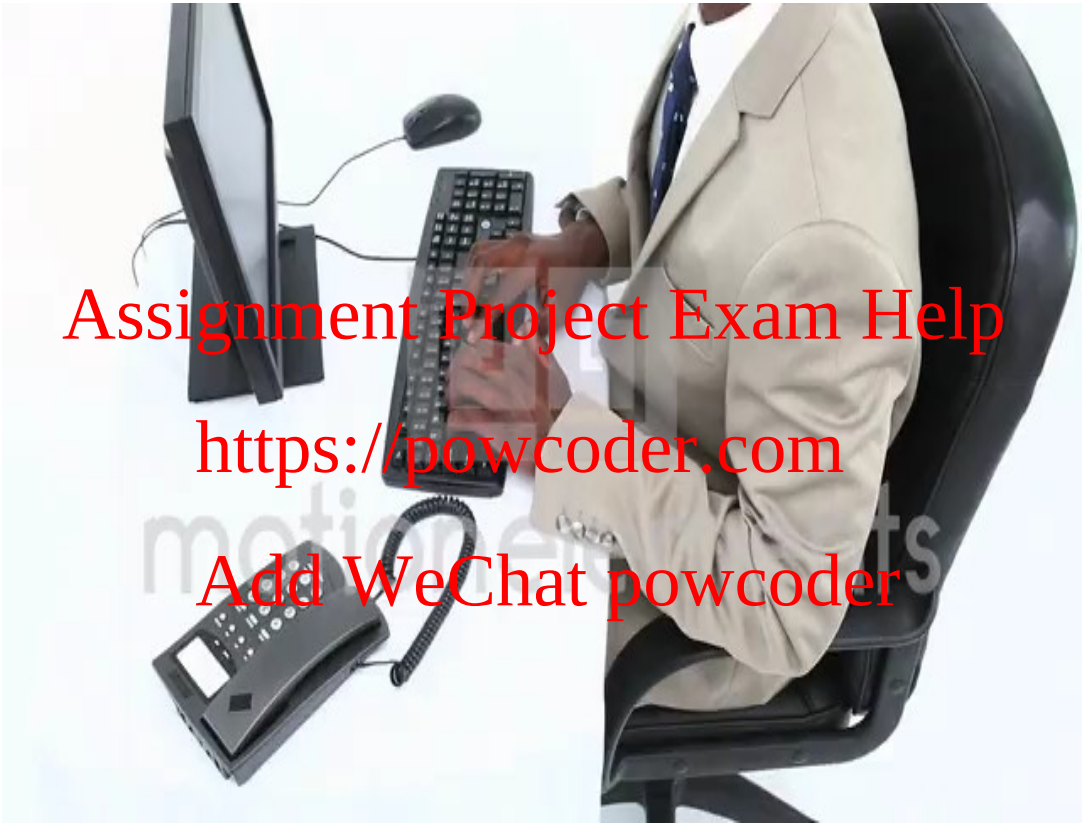
# Let’s do this for real

Despite that idiot in marketing selling defective dogs, we need a new warehouse in the East region. Which zip code should we build it in to support a strong workforce?



Is this a data mining problem?
What is the current state?
What are the possible outcomes of the business scenario?
How will the outcome of the data mining project be used?

# Let’s call Dale to tell him what we learned.





https://www.motionelements.com/stock-video-3902092-successful-black-businessman

# Types of Problems

- Retrospective
- Descriptive
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System

# Types of Problems

- **Retrospective**
- Descriptive
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System



How did we do last quarter?

*Usually point in time and standalone information not summary.*

# Types of Problems

- Retrospective
- **Descriptive**
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System



What is the average number of donuts served each morning?

*Retrospective but can be summary and/or in comparison to other data.*

# Types of Problems

- Retrospective
- Descriptive
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

Will the next person that calls accept the credit card offer?

The outcome is 1 = yes, 0=no, they will accept.

# Supervised Learning

- Goal: Predict a single “target” or “outcome” variable
- Training data, where target value is known
- Score to data where value is not known
- Methods: Classification and Prediction

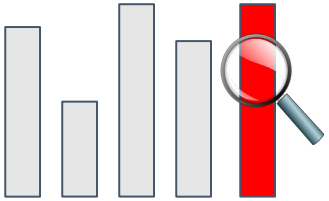
# Supervised Learning

## Inferring a function from labeled data.

*"Learn from telling", "Look at my data and I will tell you what to predict"*



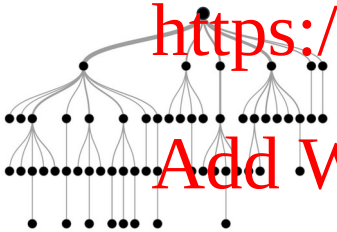
**Data Setup**



*Flat "Excel" file. Each row is a record or observation. Each column is an attribute of the record.*

***<u>One column is the outcome, y or target attribute.</u>***

**Method**



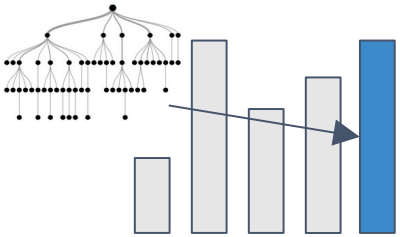
*Modeling e.g. K-NN, linear regression, decision tree, random forest etc.*

**Business Examples**

**Marketing**-Will a customer buy yes or no? How much will a customer spend?

**Operations**- Will an applicant default? When will a machine break?

**Application**



*Use the model to make predictions for the target label on the new data.*



HARVARD UNIVERSITY

# Types of Problems

- Retrospective
- Descriptive
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

Will the next person that calls accept the credit card offer?

The outcome is 1 = yes, 0=no, they will accept.

# Supervised: Classification

- Goal: Predict categorical target (outcome) variable
- Examples: Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy…
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Target variable is often binary (yes/no)

# Types of Problems

- Retrospective
- Descriptive
- Data Science
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

Will the next patron order a wine, beer, or cocktail?

The outcome is one of three classes, wine, beer, or cocktail.

# Types of Problems

- Retrospective
- Descriptive
- **Data Science**
  - **Predictive**
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - **Continuous**
  - Forecasting
  - Unsupervised Learning
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

How many ice cream cones will we sell on an 85 degree, Saturday?
(the outcome is a continuous 0 to some number of cones)

# Supervised: Prediction

- Goal: Predict numerical target (outcome) variable
- Examples: sales, revenue, performance
- As in classification:
- Each row is a case (customer, tax return, applicant)
- Each column is a variable
- Taken together, classification and prediction constitute "predictive analytics"

# Types of Problems

- Retrospective
- Descriptive
- **Data Science**
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - **Forecasting**
  - Unsupervised Learning
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

How much corn meal will we need for our corn dogs this month?
(there is an outcome, and the data is time related)

# Types of Problems

- Retrospective
- Descriptive
- **Data Science**
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - **Unsupervised Learning**
  - Associative System

Ask yourself is there an outcome variable that we want to predict.

Can our customer data base be grouped in some meaningful way?
(there is no clear outcome to predict, we can observe and explore the clusters within the customer db)

# Unsupervised Learning

- Goal: Segment data into meaningful segments; detect patterns



- There is no target (outcome) variable to predict or classify



- Methods: Association rules, data reduction & exploration, visualization

# Unsupervised Learning

## Trying to find hidden structure in unlabelled data.

*"Learn from observing", "Look at my data and tell me about it"*



### Data Setup

*Flat "Excel" file. Each row is a record or observation. Each column is an attribute of the record.*

### Method

*Clustering e.g. K-Means, Hierarchical Clustering etc*

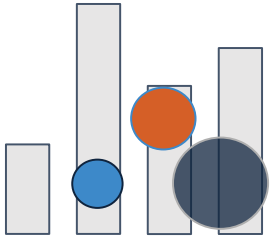
### Business Examples

**Marketing**-Find customer segments for specific marketing campaigns.

**Operations**-Identify locations for cell towers based on population density and area characteristics.

**Text Analysis**- Topic modeling of articles

### Application



*In new data find the customers/observations that most likely are part of a particular cluster.*

# Types of Problems

- Retrospective
- Descriptive
- **Data Science**
  - Predictive
    - Supervised Learning
      - Classification
        - Binary
        - Multi-Class
      - Continuous
  - Forecasting
  - Unsupervised Learning
  - **Associative System**

Ask yourself is there an outcome variable that we want to predict.

What should we offer customers that bought the corn dog?

(There is not really a distinct outcome, only observed data similar to the unsupervised example. Should we offer additional dogs, condiments, orange soda, red wine, steak etc. Among all choices, how are items associated based on purchase history?)

# Unsupervised: Association Rules

- Goal: Produce rules that define “what goes with what”
- Example: “If X was purchased, Y was also purchased”
- Rows are transactions
- Used in recommender systems – “Our records show you bought X, you may also like Y”
- Also called “affinity analysis”

# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | Will the Celtics (basketball team) win the game? |

# Your turn…

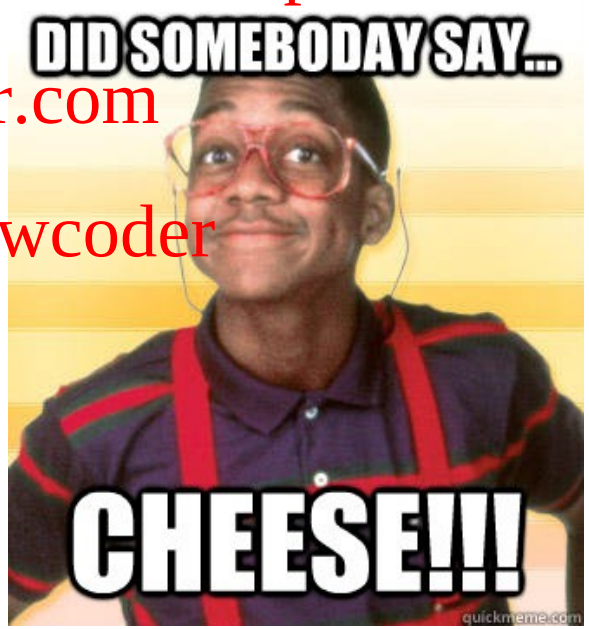
| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | How did same stores sales change year over year? |



# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | What should we place next to the cheese in the grocery cooler? |

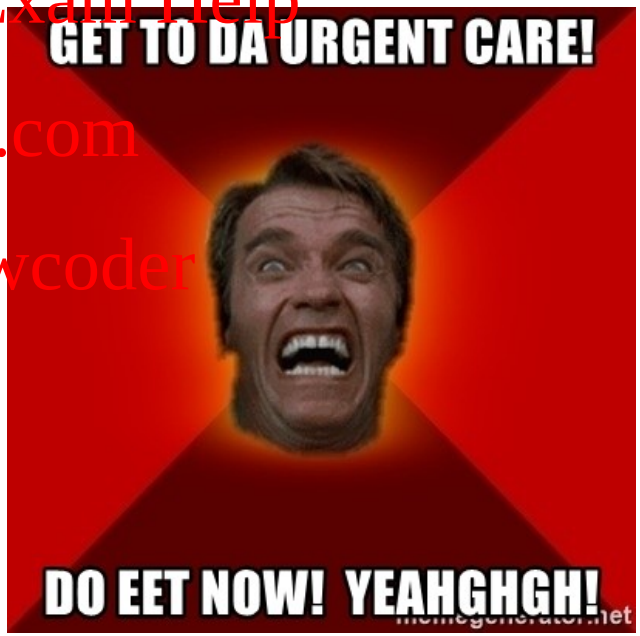




# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | How many patients should we expect in the urgent care tomorrow? |





# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | How many wickets will the Chennai SuperKings make? |

# Your turn…

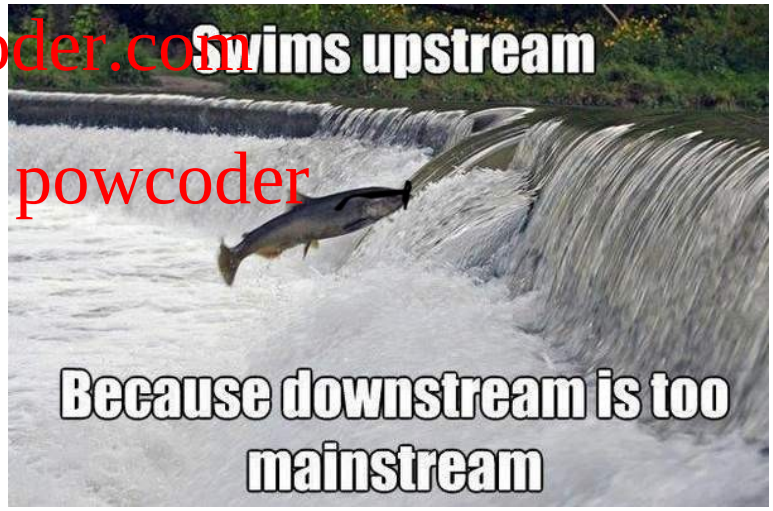
| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | What are our customer personas and how are they similar by account attribute? |



# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | How many fish did each vessel catch yesterday? |





# Your turn…

| Methods | Scenarios |
| --- | --- |
| • Retrospective<br>• Descriptive<br>• Data Science<br>  • Predictive<br>  • Supervised Learning<br>  • Classification<br>  • Binary<br>  • Multi-Class<br>  • Continuous<br>  • Forecasting<br>  • Unsupervised Learning<br>  • Associative System | Will the Celtics win the game?<br>How many wickets will the Chennai SuperKings make?<br>What should we place next to the cheese in the grocery cooler?<br>How did same stores sales change year over year?<br>How many patients should we expect in the urgent care tomorrow?<br>What are our customer personas and how are they similar by account attribute?<br>How many fish did each vessel catch yesterday? |

# Quiz!

Is there a target variable in unsupervised learning?



True or False? Data Science is defined as the study of information with the goal of extracting meaningful insights and creating actionable recommendations.



Data scientists need what type of skills?



What is an example of a supervised learning business question?
Name three data attributes would you need for that example.

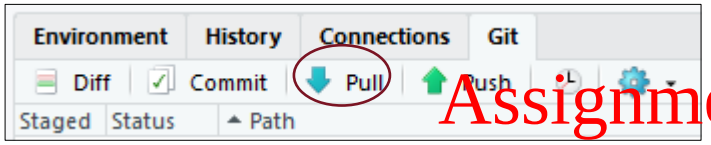
# Agenda

| Start | End | Item |
|---|---|---|
| | | Core Concepts in Data Mining |
| | | Break |
| | | More R learning & EDA |
| | | |
| | | |

# Perform a Git Pull to get the scripts & data

https://github.com/kwartler/Harvard_DataMining_Business_Student

If you have git software, when do a "git pull" in Rstudio.

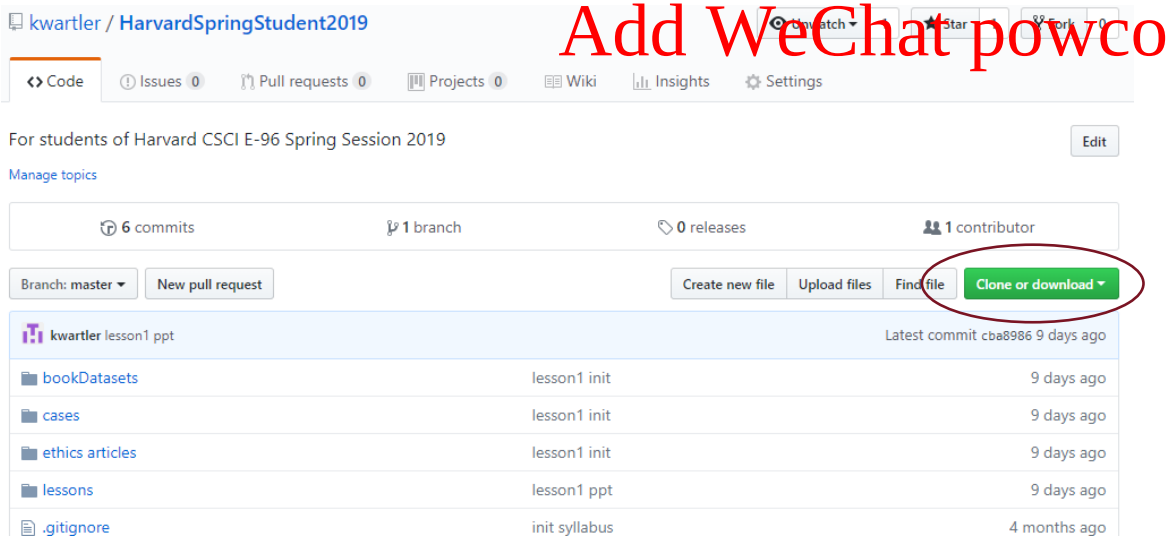





Alternatively you can download a zip of the repo on github.com but this can be cumbersome with file updates.










HARVARD UNIVERSITY

# Let's Practice!

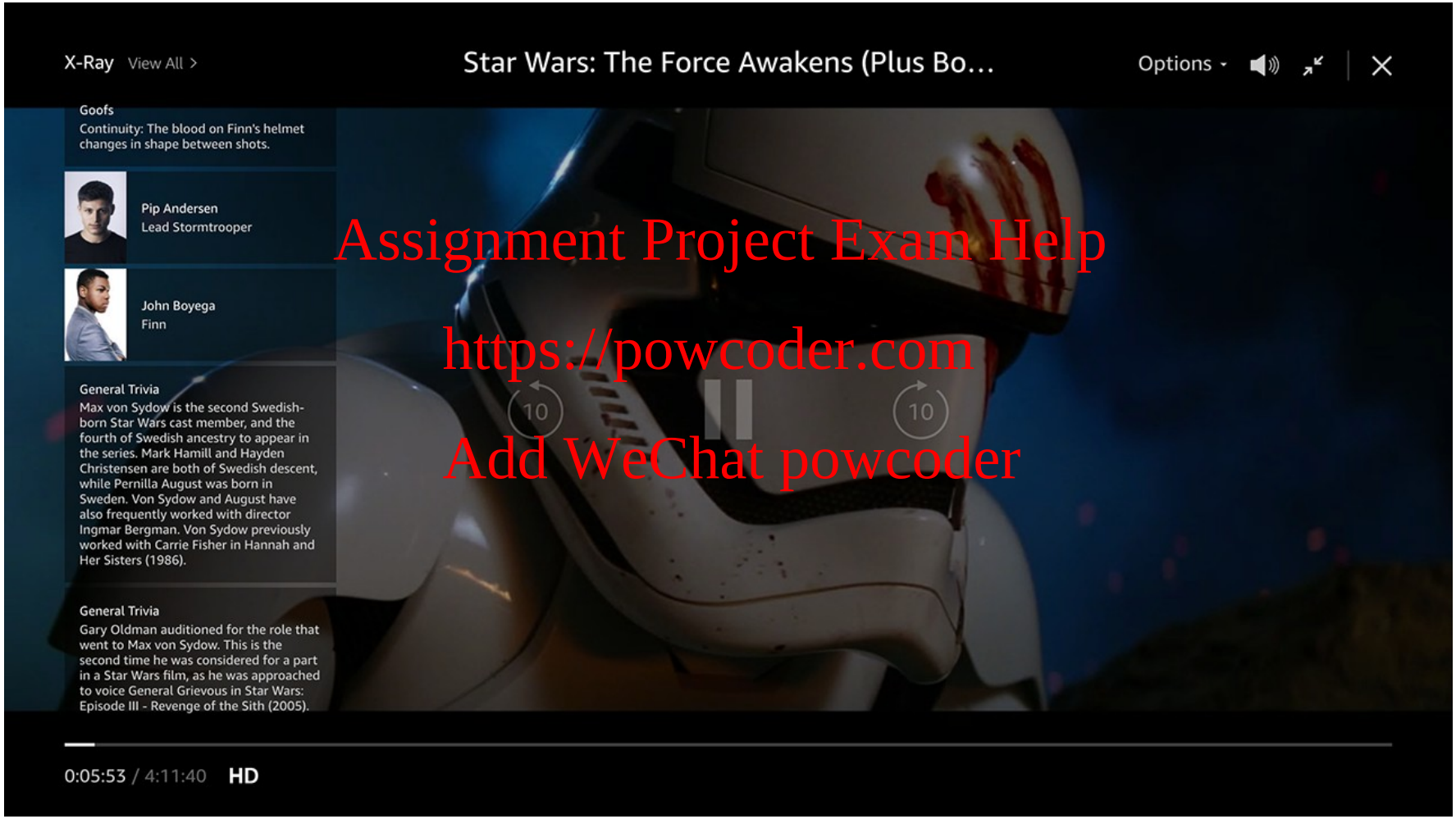
## Open:
## A_Basic_Test_Drive.R

- Simple Operators ie "+"
- Define Variables using "<-"
- Review objects and types
- Use paste()
- Find help with "?"
- Create a data.frame()
  - Add a column
  - Navigate the DF
  - write.csv()
  - read.csv()
- scatterplot()
- tables()
- barplot()
- saving a visual as a .jpeg
- "IF" Conditional loops
- "FOR" loops



Open the first script, get familiar with the basic R operations by execution.

# Do or Do Not There is No Try…



# Let’s explore Amazon’s x-ray feature

## Character Background

| | A | B | C | D |
|---|---|---|---|---|
| 1 | char.name | char.story | char.url | |
| 2 | General Hux | Ruthless commander in power struggle with Kylo Ren for the First Order leadership and being exceeded only by Snoke. | http://ia.media-imdb.com/images/M/MV5B | |
| 3 | Poe Dameron | Poe Dameron is portrayed by Oscar Isaac in Star Wars: Episode VII The Force Awakens. Isaac's casting in the film was | http://ia.media-imdb.com/images/M/MV5B | |
| 4 | Maz Kanata | A thousand-year old female pirate and past acquaintance to Han Solo. Around thirty years after the Battle of Endor, Maz I | http://ia.media-imdb.com/images/M/MV5B | |
| 5 | Unkar Plutt | Crolute junk dealer on Jakku. He is very stingy with food ration payments to Rey, until he sees BB-8 with her and offers h | http://ia.media-imdb.com/images/I/81cVbp | |
| 6 | Finn | Finn is a former storm trooper (FN-2187) who befriends Poe Dameron, Rey, Han Solo, Chewbacca, and General Organ | http://ia.media-imdb.com/images/M/MV5B | |
| 7 | Snap Wexley | NA | http://ia.media-imdb.com/images/M/MV5B | |
| 8 | Captain Phasma | Legion Commander who reports to General Hux. She wears special armor that can change shape and purpose based u | http://ia.media-imdb.com/images/M/MV5B | |



## Official Scenes



| | A | B | C |
|---|---|---|---|
| 1 | defined.scenes | start | end |
| 2 | Studio Logo | 0 | 9 |
| 3 | Star Wars Crawl | 9 | 102 |
| 4 | The First Order raids a village; Jakku | 102 | 573 |
| 5 | Poe is held captive by the First Order; Star Destroyer | 573 | 643 |
| 6 | Rey raids an aging Imperial Star Destroyer; Jakku | 643 | 1004 |



## Character Appearances

| | | | |
|---|---|---|---|
| 1 | character | start | end |
| 2 | BB-8 Performed By | 157 | 573 |
| 3 | Lor San Tekka | 171 | 573 |
| 4 | Poe Dameron | 171 | 573 |
| 5 | Jakku Villager | 236 | 573 |
| 6 | Finn | 336 | 573 |

# But first, what is a ggplot?

The first layer is to define a ggplot, with screenTime as the data. The aesthetics (aes) define that information should be colored by each character. However, there is no other information at this point.

```
ggplot(screenTime, aes(colour=screenTime$character))
```



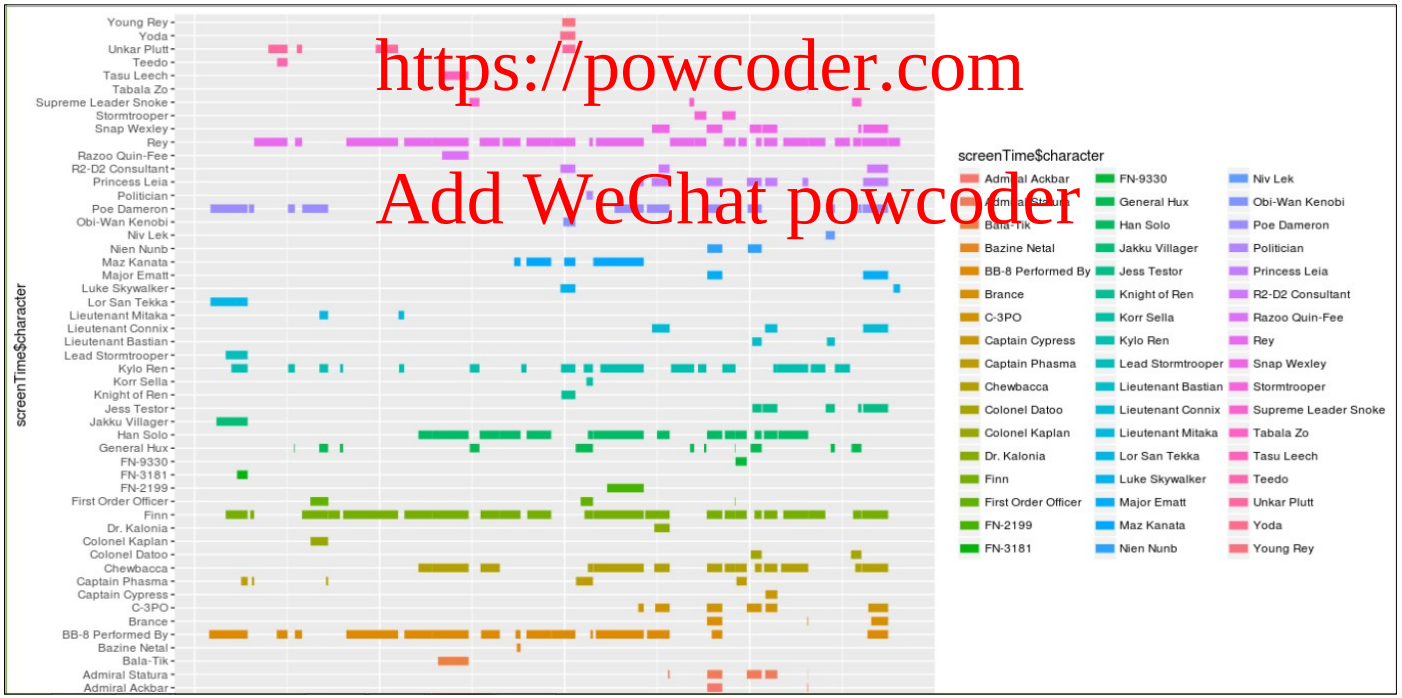
Ggplot is a “grammar of graphics” package. It works by adding layers with an “+” to construct a visual.

# Understanding ggplot…

The second layer adds a line segment for each character and defines the size of each.

```
ggplot(screenTime, aes(colour=screenTime$character)) +
geom_segment(aes(x=screenTime$start, xend=screenTime$end,y=screenTime$character,
yend=screenTime$character),size=3)
```





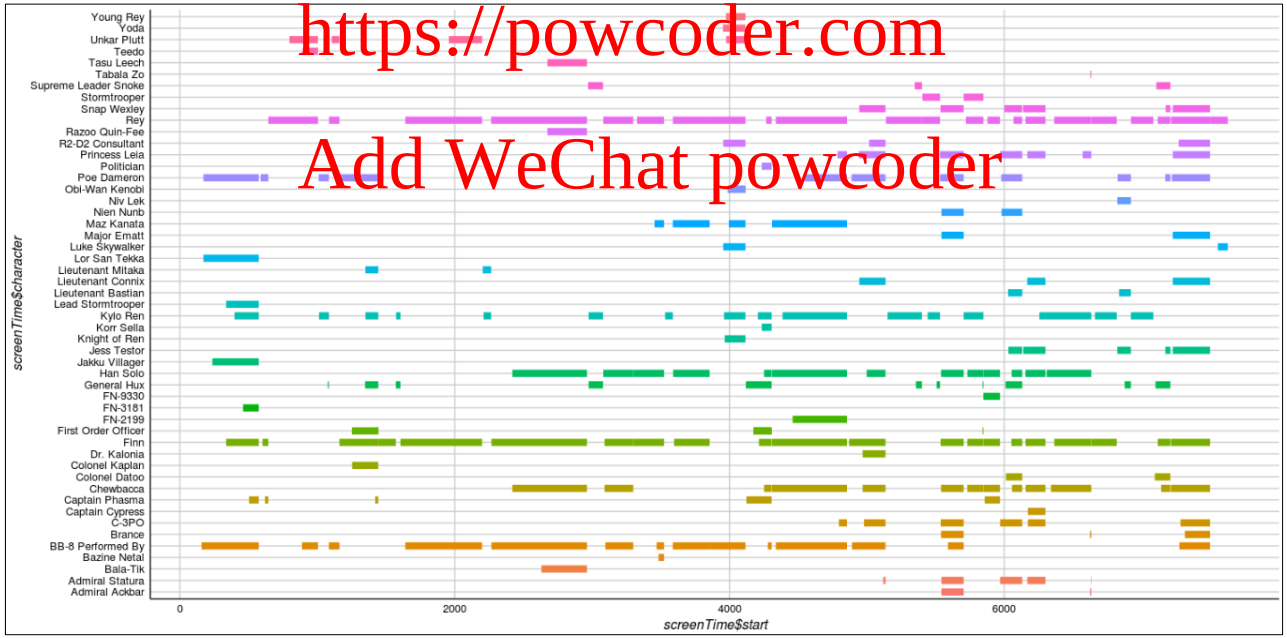
# Understanding ggplot…

The third layer changes the background, axis & colors. The fourth layer removes the legend which is redundant in this context.

```
ggplot(screenTime, aes(colour=screenTime$character)) +
geom_segment(aes(x=screenTime$start, xend=screenTime$end,y=screenTime$character,

yend=screenTime$character),size=3) + theme_gdocs() + theme(legend.position="none")
```





# Let's Practice!

## Open: B_Functions_EDA_Viz.R

- read.csv
- dim()
- table()
- indexing
- subset()
- sample()
- as.matrix()
- barplot()
- ggplot()
- Bokeh::figure()



Open the second script, get familiar with libraries, reading data, functions applied to objects & making visuals.

# Let’s Practice on geospatial data!

Open:

C_geospatial.R



- read.csv
- ggplot()
- Google maps with ggmap
- leaflet()



Open the third file, explore geospatial information.

# Common R Object Types - Vectors

Objects in R can be various forms and even made to be "custom" types.

## Numeric/Integer

| |
|---|
| 1 |
| 10 |
| 12 |
| 3.47 |
| 82 |

```
c(1,10,12,3.47)
```

## Factors (Distinct Classes)

| |
|---|
| MALE |
| FEMALE |
| FEMALE |

Unordered

```
as.factor(c('MALE','FEMALE','FEMALE'))
```

```
[1] MALE FEMALE FEMALE
Levels: FEMALE MALE
```

Cardinality 2

| |
|---|
| High |
| Med |
| Low |

Ordinal

```
as.factor(c('High','Med','Low'))
```

```
[1] high med  low
Levels: high low med
```

Cardinality 3

## Boolean

| |
|---|
| TRUE |
| TRUE |
| FALSE |
| TRUE |
| FALSE |

```
c(T, T, F, T, F)
c(TRUE,TRUE, FALSE, TRUE,FALSE)
c(T,TRUE, F, TRUE,FALSE)
```

## STRING (just text)

```
c('MALE','FEMALE','FEMALE')
```

```
[1] "MALE" "FEMALE" "FEMALE"
```



**In R, a vector can be numeric, Boolean (T/F), factors, or contain strings.**

# More Complex Common R Object Types - Matrix

Matrices are 2 dimensional data (rows/columns). Each column must be the same type.

| RowID | breaks | wool | tension |
|---|---|---|---|
| 1 | 26 | A | L |
| 2 | 30 | A | L |
| 3 | 54 | A | L |
| 4 | 25 | A | L |
| 5 | 70 | A | L |
| 6 | 52 | A | L |
| 7 | 51 | A | L |
| 8 | 26 | A | L |
| 9 | 67 | A | L |
| 10 | 18 | A | M |

```
> as.matrix(warpbreaks[1:10,])
   breaks wool tension
1  "26"   "A"  "L"
2  "30"   "A"  "L"
3  "54"   "A"  "L"
4  "25"   "A"  "L"
5  "70"   "A"  "L"
6  "52"   "A"  "L"
7  "51"   "A"  "L"
8  "26"   "A"  "L"
9  "67"   "A"  "L"
10 "18"   "A"  "M"
```

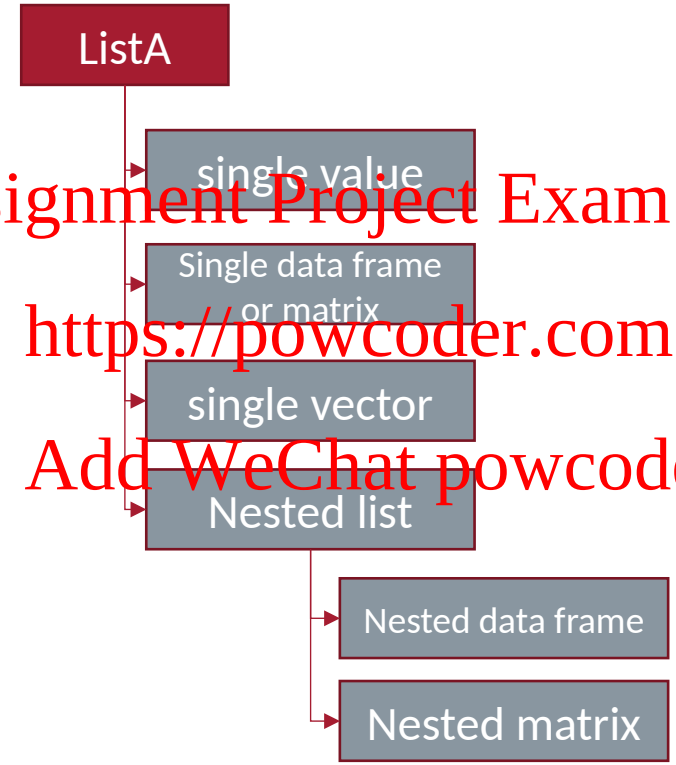
All strings



Matrices are organized into rows and columns. In R, the row names are not actually a vector of the matrix but are an attribute of the matrix. In excel you would need a standalone vector to capture that information.

# More Complex Common R Object Types – List

Lists are multi-dimensional objects that can contain different data types of different lengths.






Lists are useful for data organization but can be complex and difficult to navigate to get specific information.

# More Complex Common R Object Types – Data Frame

**Data Frames are like 2 dimensional data objects but can have mixed data types.**

*A data frame is actually a named list but with equal length elements. Being a list lets it contain mixed data types.*

| RowID | breaks | wool | tension |
|---|---|---|---|
| 1 | 26 | A | L |
| 2 | 30 | A | L |
| 3 | 54 | A | L |
| 4 | 25 | A | L |
| 5 | 70 | A | L |
| 6 | 52 | A | L |
| 7 | 51 | A | L |
| 8 | 26 | A | L |
| 9 | 67 | A | L |
| 10 | 18 | A | M |

```
> warpbreaks[1:10,]
   breaks wool tension
1      26    A       L
2      30    A       L
3      54    A       L
4      25    A       L
5      70    A       L
6      52    A       L
7      51    A       L
8      26    A       L
9      67    A       L
10     18    A       M
```

Integers | Factor | Factor



Data frames are used often because they can hold different types of vectors, but can be switched back and forth with as.matrix() and as.data.frame(). **Remember that the vector classes could change!!**



HARVARD UNIVERSITY

# Data Structure for Analysis & Modeling

Often the 1st Column is a unique identifier but the identifier could also be a row attribute (not actually a vector)



| name | mfr | type | calories | protein | fat | sodium | fiber | carbo | sugars | potass | vitamins | shelf | weight | cups | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100%_Bran | N | C | 70 | 4 | 1 | 130 | 10 | 5 | 6 | 280 | 25 | 3 | 1 | 0.33 | 68.40297 |
| 100%_Natural_Bran | Q | C | 120 | 3 | 5 | 15 | 2 | 8 | 8 | 135 | 0 | 3 | 1 | 1 | 33.98368 |
| All-Bran | K | C | 70 | 4 | 1 | 260 | 9 | 7 | 5 | 320 | 25 | 3 | 1 | 0.33 | 59.42551 |
| All-Bran_with_Extra_Fiber | K | C | 50 | 4 | 0 | 140 | 14 | 8 | 0 | 330 | 25 | 3 | 1 | 0.5 | 93.70491 |
| Almond_Delight | R | C | 110 | 2 | 2 | 200 | 1 | 14 | 8 | | 25 | 3 | 1 | 0.75 | 34.38484 |
| Apple_Cinnamon_Cheerios | G | C | 110 | 2 | 2 | 180 | 1.5 | 10.5 | 10 | 70 | 25 | 1 | 1 | 0.75 | 29.50954 |
| Apple_Jacks | K | C | 110 | 2 | 0 | 125 | 1 | 11 | 14 | 30 | 25 | 2 | 1 | 1 | 33.17409 |
| Basic_4 | G | C | 130 | 3 | 2 | 210 | 2 | 18 | 8 | 100 | 25 | 3 | 1.33 | 0.75 | 37.03856 |
| Bran_Chex | R | C | 90 | 2 | 1 | 200 | 4 | 15 | 6 | 125 | 25 | 1 | 1 | 0.67 | 49.12025 |
| Bran_Flakes | P | C | 90 | 3 | 0 | 210 | 5 | 13 | 5 | 190 | 25 | 3 | 1 | 0.67 | 53.31381 |
| Cap'n'Crunch | Q | C | 120 | 1 | 2 | 220 | 0 | 12 | 12 | 35 | 25 | 2 | 1 | 0.75 | 18.04285 |

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

# Data Structure for Analysis & Modeling

Informative features are usually independent & do not lend information to other rows (auto-correlation). Can be called informative columns, independent variables, or features.
*Remember in a DF, these can be mixed with decimals, integers, factors, strings, T/F.*



| name | mfr | type | calories | protein | fat | sodium | fiber | carbo | sugars | potass | vitamins | shelf | weight | cups | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100%_Bran | N | C | 70 | 4 | 1 | 130 | 10 | 5 | 6 | 280 | 25 | 3 | 1 | 0.33 | 68.40297 |
| 100%_Natural_Bran | Q | C | 120 | 3 | 5 | 15 | 2 | 8 | 8 | 135 | 0 | 3 | 1 | 1 | 33.98368 |
| All-Bran | K | C | 70 | 4 | 1 | 260 | 9 | 7 | 5 | 320 | 25 | 3 | 1 | 0.33 | 59.42551 |
| All-Bran_with_Extra_Fiber | K | C | 50 | 4 | 0 | 140 | 14 | 8 | 0 | 330 | 25 | 3 | 1 | 0.5 | 93.70491 |
| Almond_Delight | R | C | 110 | 2 | 2 | 200 | 1 | 14 | 8 | | 25 | 3 | 1 | 0.75 | 34.38484 |
| Apple_Cinnamon_Cheerios | G | C | 110 | 2 | 2 | 180 | 1.5 | 10.5 | 10 | 70 | 25 | 1 | 1 | 0.75 | 29.50954 |
| Apple_Jacks | K | C | 110 | 2 | 0 | 125 | 1 | 11 | 14 | 30 | 25 | 2 | 1 | 1 | 33.17409 |
| Basic_4 | G | C | 130 | 3 | 2 | 210 | 2 | 18 | 8 | 100 | 25 | 3 | 1.33 | 0.75 | 37.03856 |
| Bran_Chex | R | C | 90 | 2 | 1 | 200 | 4 | 15 | 6 | 125 | 25 | 1 | 1 | 0.67 | 49.12025 |
| Bran_Flakes | P | C | 90 | 3 | 0 | 210 | 5 | 13 | 5 | 190 | 25 | 3 | 1 | 0.67 | 53.31381 |
| Cap'n'Crunch | Q | C | 120 | 1 | 2 | 220 | 0 | 12 | 12 | 35 | 25 | 2 | 1 | 0.75 | 18.04285 |



Generally we will use data frames to avoid complexity but you will be exposed to other data types.

# Data Structure for Analysis & Modeling

If we are doing supervised learning, there is a dependent variable.

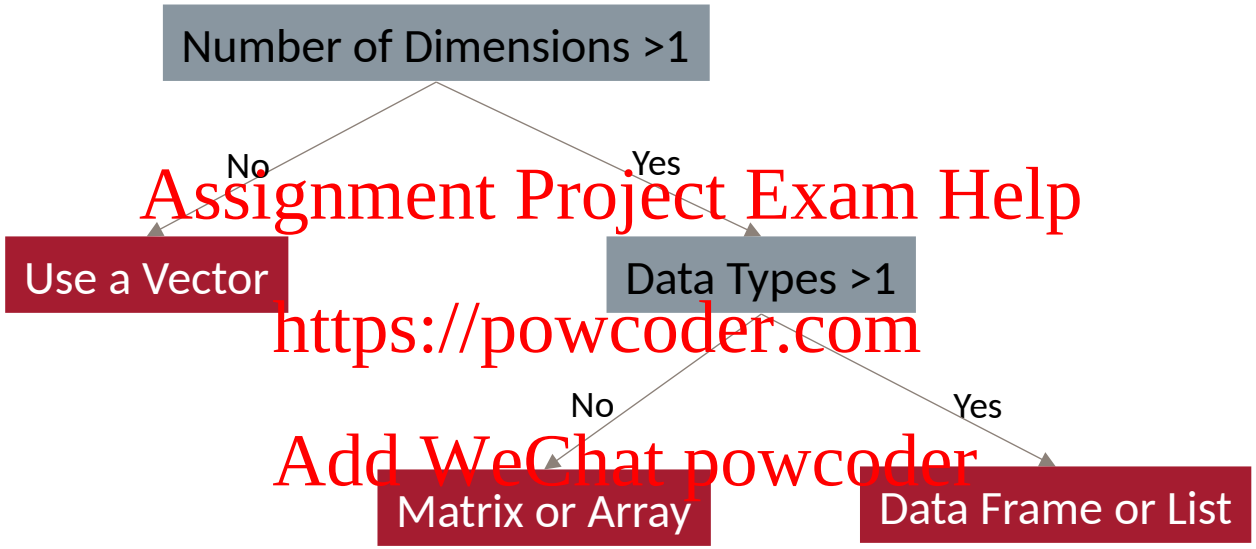
This is the outcome and is “dependent” on the informative columns. An analysis with this vector can be binary, classification, or predictive.



| name | mfr | type | calories | protein | fat | sodium | fiber | carbo | sugars | potass | vitamins | shelf | weight | cups | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100%_Bran | N | C | 70 | 4 | 1 | 130 | 10 | 5 | 6 | 280 | 25 | 3 | 1 | 0.33 | 68.40297 |
| 100%_Natural_Bran | Q | C | 120 | 3 | 5 | 15 | 2 | 8 | 8 | 135 | 0 | 3 | 1 | 1 | 33.98368 |
| All-Bran | K | C | 70 | 4 | 1 | 260 | 9 | 7 | 5 | 320 | 25 | 3 | 1 | 0.33 | 59.42551 |
| All-Bran_with_Extra_Fiber | K | C | 50 | 4 | 0 | 140 | 14 | 8 | 0 | 330 | 25 | 3 | 1 | 0.5 | 93.70491 |
| Almond_Delight | R | C | 110 | 2 | 2 | 200 | 1 | 14 | 8 | | 25 | 3 | 1 | 0.75 | 34.38484 |
| Apple_Cinnamon_Cheerios | G | C | 110 | 2 | 2 | 180 | 1.5 | 10.5 | 10 | 70 | 25 | 1 | 1 | 0.75 | 29.50954 |
| Apple_Jacks | K | C | 110 | 2 | 0 | 125 | 1 | 11 | 14 | 30 | 25 | 2 | 1 | 1 | 33.17409 |
| Basic_4 | G | C | 130 | 3 | 2 | 210 | 2 | 18 | 8 | 100 | 25 | 3 | 1.33 | 0.75 | 37.03856 |
| Bran_Chex | R | C | 90 | 2 | 1 | 200 | 4 | 15 | 6 | 125 | 25 | 1 | 1 | 0.67 | 49.12025 |
| Bran_Flakes | P | C | 90 | 3 | 0 | 210 | 5 | 13 | 5 | 190 | 25 | 3 | 1 | 0.67 | 53.31381 |
| Cap'n'Crunch | Q | C | 120 | 1 | 2 | 220 | 0 | 12 | 12 | 35 | 25 | 2 | 1 | 0.75 | 18.04285 |

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

# When should you use a specific data type?






Most analyses start with a data frame, and change classes as needed.

# Let’s Practice!

## Open D_R objects.R:

- c() to combine values into a vector
- as.matrix() to create a matrix object
- data.frame()
- as.list()
  - List elements by index
  - List elements by name



Skip depending on time!

# Data Exploration (EDA)

- Data sets are typically large, complex & messy
- Need to review the data to help refine the task
- Use techniques of Reduction and Visualization

# Exploring Data: Sampling to Save Time

- Data mining typically deals with huge databases
- For piloting/prototyping, algorithms and models are typically applied to a sample from a database, to produce statistically-valid results
- Once you develop and select a final model, you use it to "score" (predict values or classes for) the observations in the larger database

# Rare Event Over-Sampling

- Often the event of interest is rare
- Examples: response to mailing, fraud in taxes, …
- Sampling may yield too few "interesting" cases to effectively train a model
- A popular solution: oversample the rare cases to obtain a more balanced training set
- Later, need to adjust results for the oversampling



What are some cases where you think over sampling rare cases makes sense?

# Sampling & Oversampling

**TABLE 2.4** **SAMPLING IN R**

code for sampling and over/under-sampling

```
# random sample of 5 observations
s <- sample(row.names(housing.df), 5)
housing.df[s,]

# oversample houses with over 10 rooms
s <- sample(row.names(housing.df), 5, prob = ifelse(housing.df$ROOMS>10, 0.9, 0.01))
housing.df[s,]
```



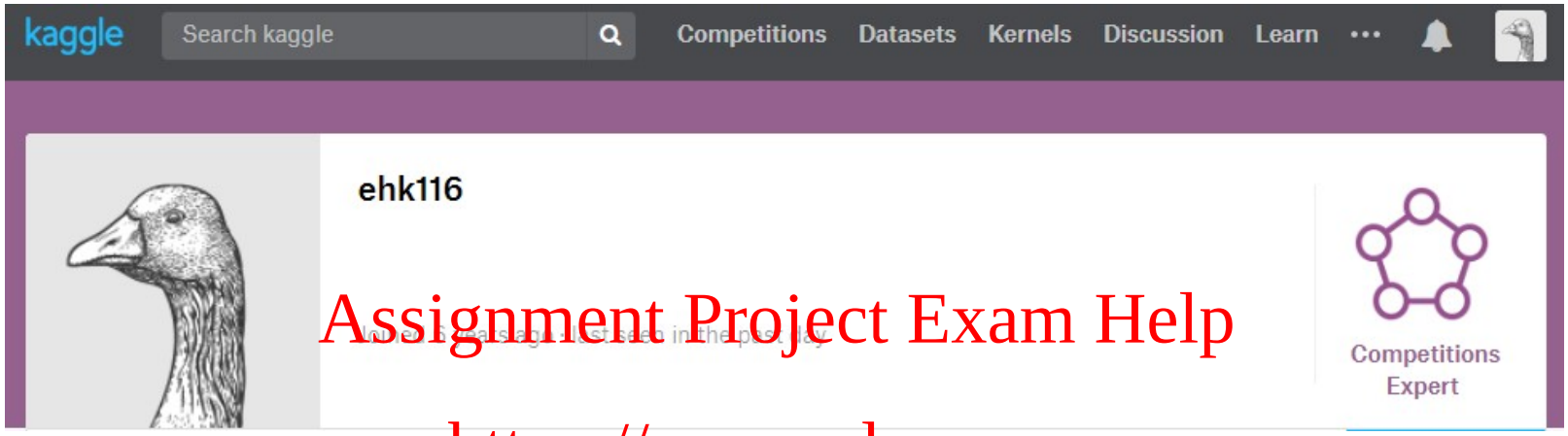
Create an index of random numbers from 1 to the number of rows.

`idx <- sample(`a vector to choose from, the number to choose`)`

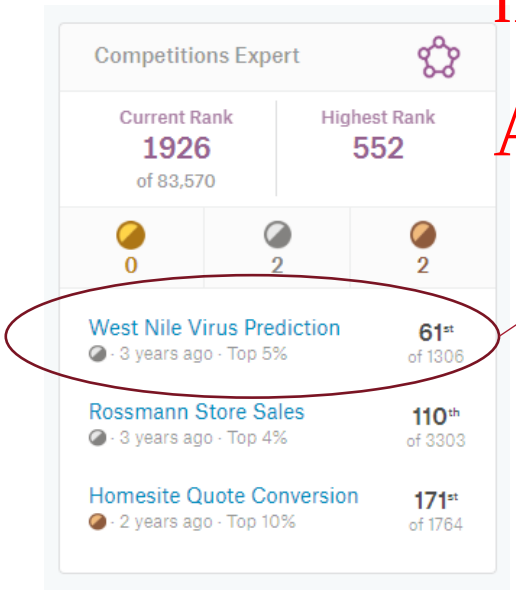
Use the index of randomly chosen numbers to select rows

`dataFrame[ idx, ]`

# What's the value of good EDA?






Asked to predict the presence of West Nile Virus in Chicago mosquitos traps.



HARVARD UNIVERSITY

# EDA let me realize a flaw!

HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH

**West Nile Virus Prediction**

Predict West Nile virus in mosquitos across the city of Chicago

$40,000 · 1,306 teams · 3 years ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Late Submission**



| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | ▼1 | Syowen | | | | |
| 59 | ▼3 | Let's find Mosquito | | 0.81303 | | |
| 60 | ▲35 | JustQ | | 0.81265 | 56 | 3y |
| 61 | ▲19 | ehk116 | | 0.81196 | 25 | 3y |
| 62 | ▼8 | H2O.ai | | 0.81171 | 42 | 3y |
| 63 | ▼3 | Artem | | 0.81109 | 28 | 3y |



Simple EDA by year would show that West Nile was 2x in 2012

After fitting an algorithm, I merely doubled predictions if they were within 2012 for the test set. Not great DS but an easy way to move up the leaderboard.



HARVARD UNIVERSITY

# Let's Practice

## Open E_EDA work.R:

- Lots of basic R options
  - str()
  - dim()
  - class()
  - head()
  - nlevels()
  - summary()
  - cor()
  - unique()
  - mean()
  - colSums()
  - is.na()

- Specific packages make life easier
  - library(DataExplorer)
    - plot_str()
    - plot_missing()
    - plot_histogram()
    - plot_density()
    - plot_scatterplot()
  - library(radiant.data)



On this script you will fill in the object, vector and information into the code scaffold. Then spend 5-10min exploring the data with radiant.data

# Housekeeping , Reading & Homework

Now that the cohort has a level foundation of R knowledge, the real fun begins…applications in a real business scenario!

- Homework…check the syllabus



- Groups will be assigned so start working on Case2 in code or with Radiant.Data!



- Read Chapter 3



- Ask questions publicly on Piazza