Week 7
Looking for sentiment words, regular expressions
Chapter 8

Dr. Dave Campbell

Using csv files

Week 2 lab:

2 Loading a csv file and dealing with data types

Download the file "pokemon_2018.csv" data file from `https://www.kaggle.com/alopez247/pokemon` or from canvas.

You may need to use the setwd and getwd commands at this time.

```
poke = read.csv(file = "pokemon_2019.csv",
    header = TRUE, sep = ",")
head(poke)
```

```
##   Number       Name Type_1 Type_2 Total HP Attack
## 1      1  Bulbasaur  Grass Poison   318 45     49
## 2      2    Ivysaur  Grass Poison   405 60     62
## 3      3   Venusaur  Grass Poison   525 80     82
## 4      4 Charmander   Fire          309 39     52
## 5      5 Charmeleon   Fire          405 58     64
## 6      6  Charizard   Fire Flying   534 78     84
```

Midterm (value 16%)

If you do better on the final we will shrink this to 10%.

Go to lab, practice on lab machines, do extra problems, be organized, have common code snippets ready

Finding or replacing text

grep(pattern, x,…) #returns a vector of the indices of the elements

gregexpr(pattern, text,…) #gregexpr returns a list of the same length as text each element of which is of the same form as the return value for regexpr, except that the starting positions of every (disjoint) match are given.

gsub(pattern, replacement, x,…) #gsub replaces all occurrences

grep:
input a vector or list
output indices of elements with the pattern

```
(ExampleText = c("1 - small thing to do","2 - 2 big things that we're doing","Some Small things that were done 4 free - 333", "four Things that we've done"))

grep("that", ExampleText) #which elements have that

grep("small", ExampleText) #which elements have small (but not Small)

grep("small", ExampleText, ignore.case = TRUE) #which elements have small or Small or smAll,…

grep("small", ExampleText, ignore.case = TRUE,value=TRUE) #which elements contains small, or Small, or smALl,..
```

grepl:
input a vector or list
output TRUE/FALSE for pattern presence in element

```
(ExampleText = c("1 - small thing to do","2 - 2 big things that we're doing","Some Small things that were done 4 free - 333", "four Things that we've done"))

grepl("that", ExampleText) #which elements have that

grepl("small", ExampleText) #which elements have small (but not Small)

grepl("small", ExampleText, ignore.case = TRUE) #which elements have small or Small or smAll,…
```

substr: subset a string of text
input is a string of text
output is the elements from **start** to **stop.**

(Example1 = "1 - small thing to do")

grep just says if “mall” exists (1) or not (0) in Example1

grep("mall",Example1) #element in which it exists (there is only one element)

grep("mdall",Example1) #can’t be found

#Extract the letters between positions **start** and **stop** (inclusive)

substr(Example1,**start**=6,**stop**=15)

gregexpr:
input a vector or list of text
output list showing text start position and pattern length

```
(ExampleText = c("1 - small thing to do","2 - 2 big things that we're doing","Some Small things that were done 4 free - 333", "four Things that we've done"))



gregexpr("thing", ExampleText)

gregexpr("small", ExampleText)

gregexpr("small", ExampleText, ignore.case = TRUE)
```

gsub:
replace a pattern within a string
input a list, vector, or string

```
(ExampleText = c("1 - small thing to do","2 - 2 big things that we're doing","Some Small things that were done 4 free - 333", "four Things that we've done"))
```

```
gsub("thing", "stuff", ExampleText) #replace “thing” with “stuff”

gsub("thing", "stuff", ExampleText,ignore.case = TRUE) # replace “thing”,”Thing”, “ThInG”,… with “stuff”
```

Finding General Pieces:

R uses "\" to say "let's do something fancy", follow it with:

\w = word characters (groups of letters)

\W = no word characters

\s = space characters

\S = no space characters

\d = digits

\D = no digits

\b = word edge

\B = no word edge

\< = word beginning

\> = word end

Finding General Pieces:

To use these you may need to use the argument: perl=TRUE

\\w = word characters (groups of letters)

\\W = no word characters

\\s = space characters

\\S = no space characters

\\d = digits

\\D = no digits

\\b = word edge

\\B = no word edge

\\< = non-word end

\\> = word end

Finding General Pieces:

To use these you may need to use the argument: perl=TRUE

```
gsub("\\w", "*", ExampleText)
#word characters

gsub("\\W", "*", ExampleText)  #No
word characters

gsub("\\s", "*", ExampleText)
#space characters

gsub("\\S", "*", ExampleText)
#non space characters

gsub("\\d", "*", ExampleText)  #
digits
```

```
gsub("\\D", "*", ExampleText)
#Non  digits

gsub("\\b", "*", ExampleText,perl
=TRUE)  # word bound

gsub("\\B", "*", ExampleText,perl
=TRUE)  #not word bounds

gsub("\\>", "*", ExampleText) #
word end

gsub("\\<", "*", ExampleText) #
NON-word end
```

Another way of finding General Pieces:

```
gsub("[[:digit:]]", "*", ExampleText)    #numbers 0-9

gsub("[[:lower:]]", "*", ExampleText)  #lower case a-z

gsub("[[:upper:]]", "*", ExampleText)  #UPPER CASE A-Z

gsub("[[:alpha:]]", "*", ExampleText) #letters

gsub("[[:alnum:]]", "*", ExampleText)  #numbers and letters

gsub("[[:punct:]]", "*", ExampleText)  # punctuation marks

gsub("[[:graph:]]", "*", ExampleText)  # numbers, letters, and punctuation

gsub("[[:space:]]", "*", ExampleText) #just spaces

gsub("[[:blank:]]", "*", ExampleText)  # spaces and tabs

gsub("[[:print:]]", "*", ExampleText) # all printable characters
```

https://www.rstudio.com/resources/cheatsheets/

regular expressions:

https://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

Composites & finding very specific pieces

gsub("(s|S)mall", "Bigly", ExampleText) # replace **small or Small**

gsub("we[[:punct:]]", "*", ExampleText) #replace **we've and we're but not were**

gsub("\\B", "*", ExampleText,perl =TRUE)   # replace **non-word end punctuation word end**

gsub("\\b[[:alpha:]]{4}\\b", "*", ExampleText,perl =TRUE) # replace four letter words

gsub("[[:digit:]]*[[:alpha:]]+[[:punct:]]", "*", ExampleText,perl =TRUE) #

gsub("\\b[[:alpha:]]{4,5}\\b", "*", ExampleText,perl =TRUE) # replace four or five letter words

Table “National Parks and National Park Reserves” from https://en.wikipedia.org/wiki/List_of_National_Parks_of_Canada

NationalParks = c(" Aulavik Northwest Territories 12,200 km2 (4,710 sq mi) 1992
Auyuittuq Pangnirtung Fiord S 2 2001-07-15.jpg Nunavut 19,089 km2 (7,370 sq mi) 2001
Banff Moraine Lake 17092005.jpg Alberta 6,641 km2 (2,564 sq mi) 1885
Bruce Peninsula CyprusLake - Bruce Peninsula.jpg Ontario 154 km2 (59 sq mi) 1987
Cape Breton Highlands NS CapeBretonHighlands1 tango7174.jpg Nova Scotia 949 km2 (366 sq mi) 1936
Elk Island Bison Elk Island.jpg Alberta 194 km2 (75 sq mi) 1913
Forillon Forillon National Park of Canada 1.jpg Quebec 244 km2 (94 sq mi) 1970
Fundy Fundy NP New Brunswick 1.jpg New Brunswick 206 km2 (80 sq mi) 1948
Georgian Bay Islands BeausoleilIslandCedarSprings2004.jpg Ontario 14 km2 (5 sq mi) 1929
Glacier Glacier np canada.JPG British Columbia1,349 km2 (521 sq mi) 1886
Grasslands Saskatchewan - Grasslands National Park 02.JPG Saskatchewan 907 km2 (350 sq mi) 1981
Gros Morne NLW GrosMorne4 tango7174.jpg Newfoundland and Labrador 1,805 km2 (697 sq mi) 1973
Gulf Islands (Reserve) Gulfislfromair.jpg British Columbia36 km2 (14 sq mi) 2003
Gwaii Haanas[A] (Reserve) Haida Heritage Centre.jpg British Columbia1,495 km2 (577 sq mi) 1988
Ivvavik[B] Canada--yukon--ivvavik-np--spe 3021.jpg Yukon 10,168 km2 (3,926 sq mi) 1984
Jasper Fryatt Valley top.jpg Alberta 10,878 km2 (4,200 sq mi) 1907
Kejimkujik Kejimkujik NP Nova Scotia 3.jpg Nova Scotia 404 km2 (156 sq mi) 1968
Kluane[C] (two units: a Park and a Reserve) Donjek Valley.jpgYukon 22,013 km2 (8,499 sq mi) 1976 (Reserve) 1993 (Park)
Kootenay Kootenay National Park.jpg British Columbia1,406 km2 (543 sq mi) 1920
Kouchibouguac Kouchibouguac.JPG New Brunswick 239 km2 (92 sq mi) 1969
La Mauricie Ile aux pins.jpg Quebec 536 km2 (207 sq mi) 1970
Mingan Archipelago (Reserve) Monolithes de L'Archipel de Mingan.jpg Quebec 151 km2 (58 sq mi) 1984
Mount Revelstoke Revelstoke from Mount Revelstoke.jpg British Columbia 260 km2 (100 sq mi) 1914
Naats'ihch'oh[4] (Reserve) Howard's Pass Yukon Territory 1.jpg Northwest Territories 4,850 km2 (1,873 sq mi) 2014
Nahanni (Reserve) Nahanni - VirginiaFalls.jpg Northwest Territories 30,000 km2 (11,583 sq mi) 1976
Pacific Rim (Reserve) Longbeach prnp.jpg British Columbia511 km2 (197 sq mi) 1970
Point Pelee Point Pelee looking south.jpg Ontario 15 km2 (6 sq mi) 1918
Prince Albert Prince Albert National Park.jpg Saskatchewan 3,874 km2 (1,496 sq mi) 1927
Prince Edward Island Peicoast.jpg Prince Edward Island 22 km2 (8 sq mi) 1937
Pukaskwa HorseshoeBayPukaskwaPark23.jpg Ontario 1,878 km2 (725 sq mi) 1978
Qausuittuq Peary caribou - looking west towards Evan's Bay.jpg Nunavut 11,000 km2 (4,247 sq mi) 2015
Quttinirpaaq[E] Tanquary Fiord 16 1997-08-05.jpg Nunavut 37,775 km2 (14,585 sq mi) 2001
Riding Mountain[F] Bison herd - Lake Audy - Riding Mountain National Park.JPG Manitoba2,973 km2 (1,148 sq mi) 1933
Rouge Little Rouge River Lookout.jpg Ontario 36 km2 (14 sq mi) 2015
Sable Island (Reserve) SableHorses.jpg Nova Scotia 34 km2 (13 sq mi) 2013
Sirmilik Sirmilik Glacier 2 1997-08-06.jpg Nunavut 22,200 km2 (8,571 sq mi) 2001
Terra Nova NLC TerraNova3 tango7174.jpg Newfoundland and Labrador 400 km2 (154 sq mi) 1957
Thousand Islands Thousand Islands 2.JPG Ontario 24 km2 (9 sq mi) 1904
Torngat Mountains Nachvak Fjord Labrador 2008.JPG Newfoundland and Labrador 9,700 km2 (3,745 sq mi) 2008
Tuktut Nogait Hornaday River.jpg Northwest Territories 16,340 km2 (6,309 sq mi) 1996
Ukkusiksalik Eisbär 1996-07-23.jpg Nunavut 20,885 km2 (8,064 sq mi) 2003
Vuntut Vontut National Park.jpg Yukon 4,345 km2 (1,678 sq mi) 1995
Wapusk Bärenmutter & Junges 3 2004-11-17.jpg Manitoba 11,475 km2 (4,431 sq mi) 1996
Waterton Lakes[G] Upper Waterton Lake.JPG Alberta 505 km2 (195 sq mi) 1895
Wood Buffalo Wood-Buffalo-NP Gros Beak Lake 2 98-07-02.jpg Alberta Northwest Territories 44,807 km2 (17,300 sq mi) 1922
Yoho YohoNP-Takakkaw IMG 1372-800x533byBMK.jpg British Columbia 1,313 km2 (507 sq mi) 1886")

How to split into Rows and Columns?

FewRows = c(" Aulavik Northwest Territories 12,200 km2 (4,710 sq mi) 1992

  Auyuittuq Pangnirtung Fiord S 2 2001-07-15.jpg Nunavut 19,089 km2 (7,370 sq mi) 2001

  Banff Moraine Lake 17092005.jpg Alberta 6,641 km2 (2,564 sq mi) 1885

  Bruce Peninsula CyprusLake - Bruce Peninsula.jpg Ontario 154 km2 (59 sq mi) 1987

  Cape Breton Highlands NS CapeBretonHighlands1 tango7174.jpg Nova Scotia 949 km2 (366 sq mi) 1936

  Elk Island Bison Elk Island.jpg Alberta 194 km2 (75 sq mi) 1913")

How to split into Rows and Columns?

FewRows = c(" Aulavik Northwest Territories 12,200 km2 (4,710 sq mi) 1992

  Auyuittuq Pangnirtung Fiord S 2 2001-07-15.jpg Nunavut 19,089 km2 (7,370 sq mi) 2001

  Banff Moraine Lake 17092005.jpg Alberta 6,641 km2 (2,564 sq mi) 1885

  Bruce Peninsula CyprusLake - Bruce Peninsula.jpg Ontario 154 km2 (59 sq mi) 1987

  Cape Breton Highlands NS CapeBretonHighlands1 tango7174.jpg Nova Scotia 949 km2 (366 sq mi) 1936

  Elk Island Bison Elk Island.jpg Alberta 194 km2 (75 sq mi) 1913")

RowSplits = strsplit(FewRows,"\n")

One Row

Row1 = c(" Aulavik   Northwest Territories
12,200 km2 (4,710 sq mi)   1992")

One Row

Row1 = c(" Aulavik   Northwest Territories
12,200 km2 (4,710 sq mi)   1992")

OneRow = strsplit(Row1,"\t")

Make a data frame by splitting into rows and columns

See Stat 341 for other R options (i.e. using apply, lapply, parLapply,…)

(RowSplits = strsplit(FewRows,"\n|\t")) #Split at line breaks “\n” OR tabs “\t”:

splitup = unlist(strsplit(FewRows,"\n|\t")) #Make a data.frame:

Make a data frame by splitting into rows and columns

```
(RowSplits = strsplit(FewRows,"\n|\t")) #Split at line breaks “\n” OR tabs “\t”:

splitup = unlist(strsplit(FewRows,"\n|\t")) #Make a data.frame


N = length(splitup)

Name        = splitup[seq(1,N,by=5)]

Photo       = splitup[seq(2,N,by=5)]

Location    = splitup[seq(3,N,by=5)]

Area        = splitup[seq(4,N,by=5)]

Established = splitup[seq(5,N,by=5)]

(Parks = cbind(Name,Location,Area,Established))
```

Dealing with Area

Area:

Area = "12,200 km2 (4,710 sq mi)"

Split into 2 columns: km2 & sq mi

Finding special characters requires escaping from the regular way of using them.

gsub("(", "*", Area) # error

“\\(” escapes and then looks for the (

gsub("\\(", "*", Area)

“.” means any character except a line break

“*” means **≥0** matches

gsub("\\(.*\\)", "*", Area)

How do we get rid of bracket stuff AND punctuation?

How do we get vectors of sqmi and km2?

Making a data frame from web table copied data

```
(km2 = gsub("[[:punct:]]|km2.*\\)", "", Area) )

(km2 = as.numeric(gsub("[[:punct:]]|km2.*\\)", "",
Area) ))

(sqmi = as.numeric(gsub(".*\\(|[[:alpha:]]|[[:punct:]]", "",
Area) ))
```

```
(Parks =
cbind(Name,Location,km2,sqmi,Establis
hed))
```

NationalParks = c(" Aulavik Northwest Territories 12,200 km2 (4,710 sq mi) 1992
Auyuittuq Pangnirtung Fiord S 2 2001-07-15.jpg Nunavut 19,089 km2 (7,370 sq mi) 2001
Banff Moraine Lake 17092005.jpg Alberta 6,641 km2 (2,564 sq mi) 1885
Bruce Peninsula CyprusLake - Bruce Peninsula.jpg Ontario 154 km2 (59 sq mi) 1987
Cape Breton Highlands NS CapeBretonHighlands1 tango7174.jpg Nova Scotia 949 km2 (366 sq mi) 1936
Elk Island Bison Elk Island.jpg Alberta 194 km2 (75 sq mi) 1913
Forillon Forillon National Park of Canada 1.jpg Quebec 244 km2 (94 sq mi) 1970
Fundy Fundy NP New Brunswick 1.jpg New Brunswick 206 km2 (80 sq mi) 1948
Georgian Bay Islands BeausoleilIslandCedarSprings2004.jpg Ontario 14 km2 (5 sq mi) 1929
Glacier Glacier np canada.JPG British Columbia1,349 km2 (521 sq mi) 1886
Grasslands Saskatchewan - Grasslands National Park 02.JPG Saskatchewan 907 km2 (350 sq mi) 1981
Gros Morne NLW GrosMorne4 tango7174.jpg Newfoundland and Labrador 1,805 km2 (697 sq mi) 1973
Gulf Islands (Reserve) Gulfislfromair.jpg British Columbia36 km2 (14 sq mi) 2003
Gwaii Haanas[A] (Reserve) Haida Heritage Centre.jpg British Columbia1,495 km2 (577 sq mi) 1988
Ivvavik[B] Canada--yukon--ivvavik-np--spe 3021.jpg Yukon 10,168 km2 (3,926 sq mi) 1984
Jasper Fryatt Valley top.jpg Alberta 10,878 km2 (4,200 sq mi) 1907
Kejimkujik Kejimkujik NP Nova Scotia 3.jpg Nova Scotia 404 km2 (156 sq mi) 1968
Kluane[C] (two units: a Park and a Reserve) Donjek Valley.jpgYukon 22,013 km2 (8,499 sq mi) 1976
(Reserve) 1993 (Park)
Kootenay Kootenay National Park.jpg British Columbia1,406 km2 (543 sq mi) 1920
Kouchibouguac Kouchibouguac.JPG New Brunswick 239 km2 (92 sq mi) 1969
La Mauricie Ile aux pins.jpg Quebec 536 km2 (207 sq mi) 1970
Mingan Archipelago (Reserve) Monolithes de L'Archipel de Mingan.jpg Quebec 151 km2 (58 sq mi) 1984
Mount Revelstoke Revelstoke from Mount Revelstoke.jpg British Columbia 260 km2 (100 sq mi) 1914
Naats'ihch'oh[4] (Reserve) Howard's Pass Yukon Territory 1.jpg Northwest Territories 4,850 km2 (1,873 sq mi) 2014
Nahanni (Reserve) Nahanni - VirginiaFalls.jpg Northwest Territories 30,000 km2 (11,583 sq mi) 1976
Pacific Rim (Reserve) Longbeach prnp.jpg British Columbia511 km2 (197 sq mi) 1970
Point Pelee Point Pelee looking south.jpg Ontario 15 km2 (6 sq mi) 1918
Prince Albert Prince Albert National Park.jpg Saskatchewan 3,874 km2 (1,496 sq mi) 1927
Prince Edward Island Peicoast.jpg Prince Edward Island 22 km2 (8 sq mi) 1937
Pukaskwa HorseshoeBayPukaskwaPark23.jpg Ontario 1,878 km2 (725 sq mi) 1978
Qausuittuq Peary caribou - looking west towards Evan's Bay.jpg Nunavut 11,000 km2 (4,247 sq mi) 2015
Quttinirpaaq[E] Tanquary Fiord 16 1997-08-05.jpg Nunavut 37,775 km2 (14,585 sq mi) 2001
Riding Mountain[F] Bison herd - Lake Audy - Riding Mountain National Park.JPG Manitoba 2,973 km2 (1,148 sq mi) 1933
Rouge Little Rouge River Lookout.jpg Ontario 36 km2 (14 sq mi) 2015
Sable Island (Reserve) SableHorses.jpg Nova Scotia 34 km2 (13 sq mi) 2013
Sirmilik Sirmilik Glacier 2 1997-08-06.jpg Nunavut 22,200 km2 (8,571 sq mi) 2001
Terra Nova NLC TerraNova3 tango7174.jpg Newfoundland and Labrador 400 km2 (154 sq mi) 1957
Thousand Islands Thousand Islands 2.JPG Ontario 24 km2 (9 sq mi) 1904
Torngat Mountains Nachvak Fjord Labrador 2008.JPG Newfoundland and Labrador 9,700 km2 (3,745 sq mi) 2008
Tuktut Nogait Hornaday River.jpg Northwest Territories 16,340 km2 (6,309 sq mi) 1996
Ukkusiksalik Eisbär 1996-07-23.jpg Nunavut 20,885 km2 (8,064 sq mi) 2003
Vuntut Vontut National Park.jpg Yukon 4,345 km2 (1,678 sq mi) 1995
Wapusk Bärenmutter & Junges 3 2004-11-17.jpg Manitoba 11,475 km2 (4,431 sq mi) 1996
Waterton Lakes[G] Upper Waterton Lake.JPG Alberta 505 km2 (195 sq mi) 1895
Wood Buffalo Wood-Buffalo-NP Gros Beak Lake 2 98-07-02.jpg Alberta
Northwest Territories 44,807 km2 (17,300 sq mi) 1922
Yoho YohoNP-Takakkaw IMG 1372-800x533byBMK.jpg British Columbia 1,313 km2 (507 sq mi) 1886")

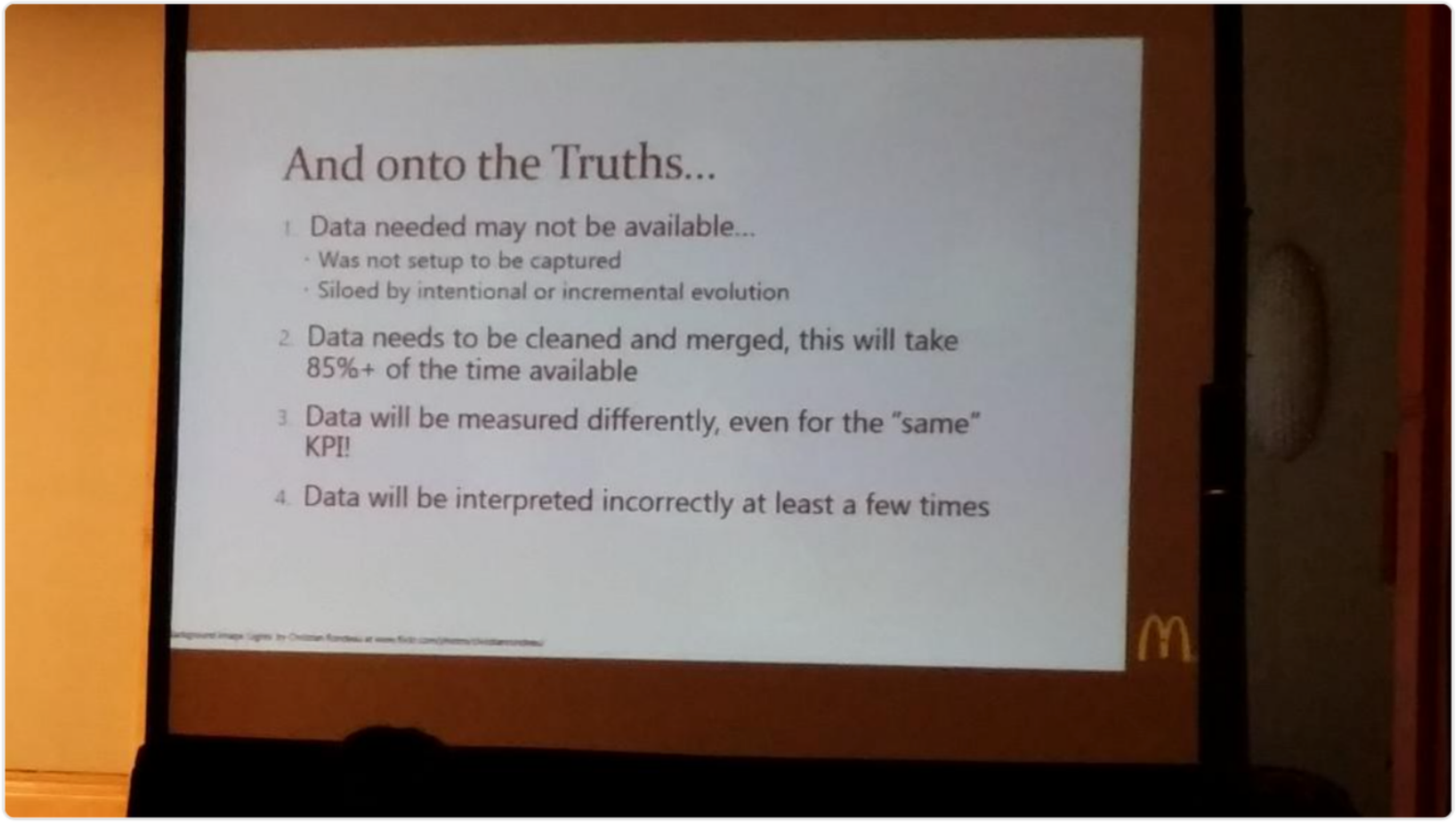
```
splitup = unlist(strsplit(NationalParks,"\n|\t")) #Make a
data.frame


N = length(splitup)

Name        = splitup[seq(1,N,by=5)]

Photo       = splitup[seq(2,N,by=5)]

Location    = splitup[seq(3,N,by=5)]

Area        = splitup[seq(4,N,by=5)]

Established  = splitup[seq(5,N,by=5)]

(Parks = cbind(Name,Location,Area,Established))
```

https://twitter.com/CIPSToronto/status/831610306111614977

CIPSToronto
@CIPSToronto

Follow

Diep of @McD_Canada tells #BigDataCA that 85% of time spent on data cleansing merging



RETWEET
1

1:05 PM - 14 Feb 2017 from Toronto, Ontario