Lab 1 - What is Data Science

YourNameGoesHere

For class Tuesday, February 8th

Labs

Labs (what we will call class activities whether or not they involve using software) will help you practice with data science concepts, coding, and communicating. Labs will usually start with a synopsis of the activity, so you can understand the purpose of the tasks that follow. Labs will often involve group work.

Today's lab is really a set of discussions that you will have in small groups with major ideas being shared with the class when we come back together as a group. The lab will help you get to know some classmates and work through some thoughts on what data science is, even if you haven't gotten through the first assigned reading.

You can follow along with this lab in R, if you have everything set up for it already (without Git is fine for today) or simply keep notes on your own paper following along with the pdf. If working in R, you'll still want the .pdf to see the images.

Getting Started

Today's groups were randomly generated. Introduce yourselves to each other by sharing your preferred name, at least one academic interest, and at least one extracurricular activity. Take a few minutes for this - the goal is to start getting to know your classmates.

Next, we'll tackle some discussion questions. We'll come back together as a class as groups finish Part I, and hopefully again after Part II.

Part I

Answer the following questions with your group, discussing your ideas. Is there a main idea you'd be willing to share with the class for each question?

What words come to mind when you hear the phrase `data science`?

What does `data science` mean to you?

Some descriptions of data science are visual. (References at end of the lab.)

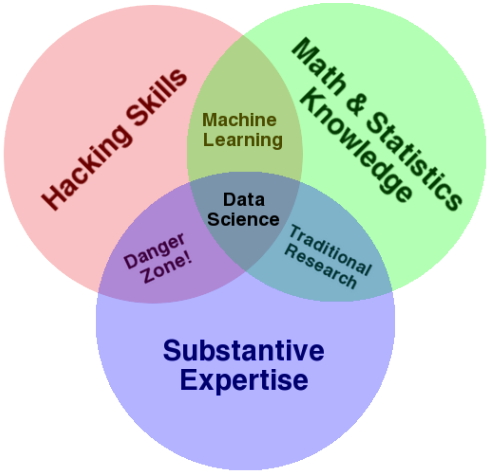


Figure 1: Conway: Data Science Venn Diagram.

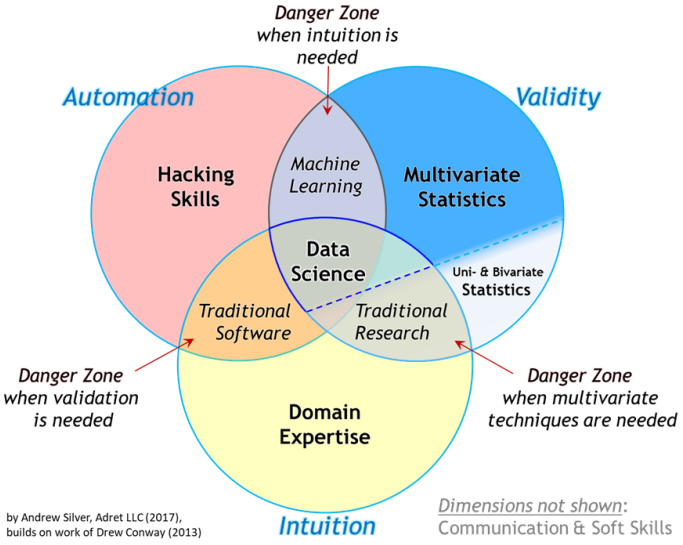


Figure 2: Towards Data Science: Data Science Venn Diagram.

Do either of these graphics match the group consensus about what `data science` is? Discuss what you see.

The second graphic is based on the first. Do you think the changes help in explaining what data science is?

Where is the science in `data science`?

Our course is project-focused to allow us to practice following the data analysis cycle. This is also referred to as the data science life cycle.

How would you describe this cycle?

Some descriptions of the cycle are visual. (References at end of the lab.)

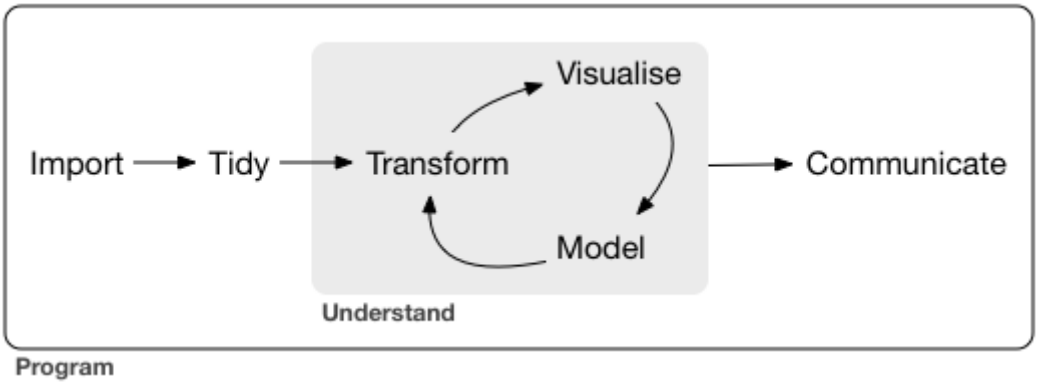


Figure 3: R for Data Science: Cycle.

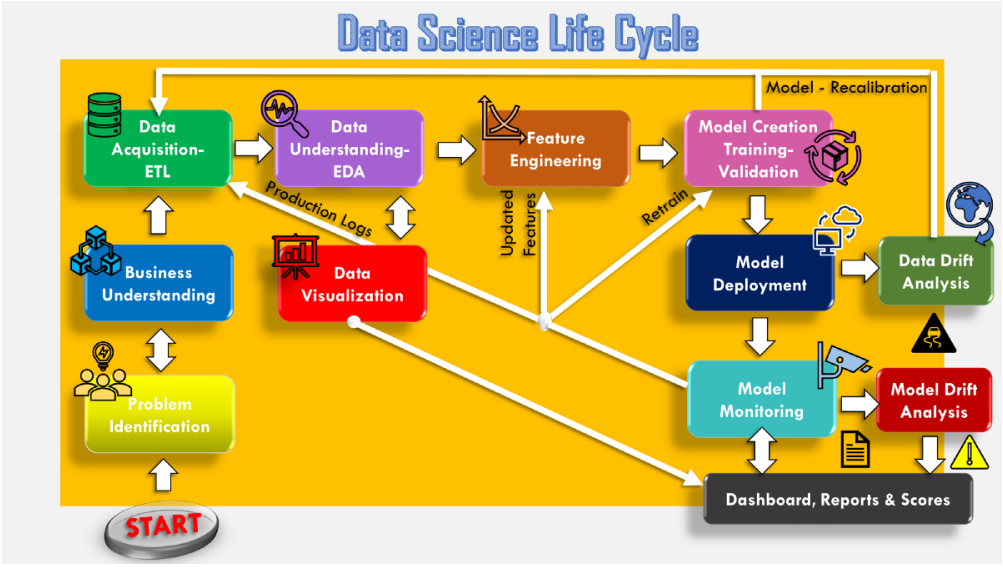


Figure 4: Towards Data Science: Cycle.

Is there anything surprising to you about either depiction of the cycle? Anything you would add or remove? Which depiction do you prefer?

What role does ethics play in `data science`?

Part II

You can continue on into Part II after completing Part I. I expect we will come back together as a class to discuss some ideas from Part I.

Everyone in the group should pick an application area of interest to them. Have everyone do a google search for "data science applicationarea" where you put your application area in. For example, I like learning about archaeology, so I searched for "data science archaeology".

What kind of results do you find with your google search? Share with your group.

Next, we're going to get a sense of what data science is by looking at a some code and a visual it generates. (Visual and code based on an activity by Prof. Bailey.) We'll address a few related questions on the way.

We will be working with the voting history of countries in the United Nations General Assembly. There are R packages that allow access to many different datasets. For this visual, we will be using data from the **unvotes** package. Additionally, we will make use of the **tidyverse** and **lubridate** packages for the analysis.

What is your understanding of what an R package is?

The **unvotes** package provides three datasets we can work with: `un_roll_calls`, `un_roll_call_issues`, and `un_votes`. Each of these datasets contains a variable called `rcid`, the roll call id, which can be used as a unique identifier to join the three datasets together. We will learn details of this code later, for now, just follow along.

The `un_votes` dataset provides information on the voting history of the United Nations General Assembly. It contains one row for each country-vote pair.

```
head(un_votes, 4)
```

```
## # A tibble: 4 x 4
##    rcid country       country_code vote
##   <dbl> <chr>         <chr>        <fct>
## 1     3 United States US           yes
## 2     3 Canada        CA           no
## 3     3 Cuba          CU           yes
## 4     3 Haiti         HT           yes
```

The `un_roll_calls` dataset contains information on each roll call vote of the United Nations General Assembly.

```
head(un_roll_calls, 4)
```

```
## # A tibble: 4 x 9
##    rcid session importantvote date       unres   amend  para short    descr
##   <int>   <dbl>         <int> <date>     <chr>   <int> <int> <chr>    <chr>
## 1     3       1             0 1946-01-01 R/1/66      1     0 AMENDME~ "TO ADOPT~
## 2     4       1             0 1946-01-02 R/1/79      0     0 SECURIT~ "TO ADOPT~
## 3     5       1             0 1946-01-04 R/1/98      0     0 VOTING ~ "TO ADOPT~
## 4     6       1             0 1946-01-04 R/1/107     0     0 DECLARA~ "TO ADOPT~
```

The `un_roll_call_issues` dataset contains (topic) classifications of roll call votes of the United Nations General Assembly. Many votes had no topic, and some have more than one.

```
head(un_roll_call_issues, 4)
```

```
## # A tibble: 4 x 3
##    rcid short_name issue
##   <int> <chr>      <fct>
## 1    77 me         Palestinian conflict
## 2  9001 me         Palestinian conflict
## 3  9002 me         Palestinian conflict
## 4  9003 me         Palestinian conflict
```

Data prep

In order to do our analysis, we first need to combine our three datasets into one.

```
unvotes <- un_votes %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid")
```

What do you think one case/observation in our final unvotes data set represents? A country? A UN-session?

```
head(unvotes)
```

```
## # A tibble: 6 x 14
##    rcid country  country_code vote  session importantvote date       unres amend
##   <dbl> <chr>    <chr>        <fct>   <dbl>         <int> <date>     <chr> <int>
## 1     6 United ~ US           no          1             0 1946-01-04 R/1/~     0
## 2     6 Canada   CA           no          1             0 1946-01-04 R/1/~     0
## 3     6 Cuba     CU           yes         1             0 1946-01-04 R/1/~     0
## 4     6 Dominic~ DO           abst~       1             0 1946-01-04 R/1/~     0
## 5     6 Mexico   MX           yes         1             0 1946-01-04 R/1/~     0
## 6     6 Guatema~ GT           no          1             0 1946-01-04 R/1/~     0
## # ... with 5 more variables: para <int>, short <chr>, descr <chr>,
## #   short_name <chr>, issue <fct>
```

We started from the `un_votes` data set and basically, just added on additional information to each row. So, the cases/observations are the same as `un_votes` - they are country-vote pairs.

Visualization

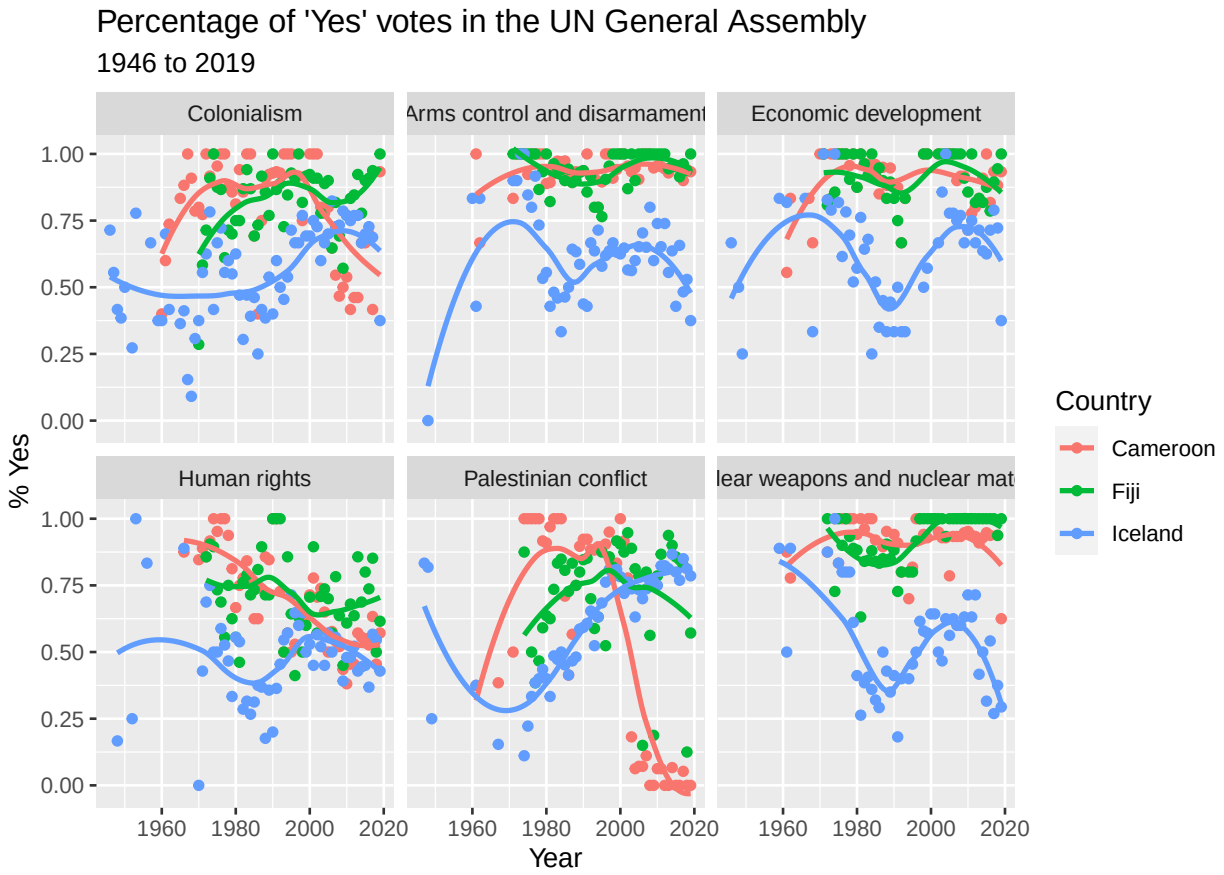
Now we can create a visualization that displays how the voting record changed over time across countries on six broader issues. We will pick 3 countries at random (I used R to select these.)

Take a moment to look through the R code on the next page. What functions are new to you?

We'll be focusing on making visuals and wrangling for the next 2 weeks.

What functions seem understandable to you? That is, for what functions do you get a sense of what they are doing?

```
unvotes %>%
  filter(country %in% c("Cameroon",
                        "Fiji",
                        "Iceland")) %>%
  group_by(country, year = year(date), issue) %>%
  summarize(votes = n(),
            percent_yes = mean(vote == "yes")) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year,
                       y = percent_yes,
                       color = country)) +
    geom_point() +
    geom_smooth(method = "loess", se = FALSE) +
    facet_wrap(~ issue) +
    labs(title = "Percentage of 'Yes' votes in the UN General Assembly",
         subtitle = "1946 to 2019",
         y = "% Yes",
         x = "Year",
         color = "Country")
```



Consider the final visual. How would you summarize what you've found?

Reflecting on all your discussions and the code/visual example, brainstorm skills you might want to develop or improve as part of our course work. Share with your group.

References

1. Drew Conway - Data Science Venn Diagram - http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
2. Towards Data Science - Data Science Venn Diagram - Andrew Silver - https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40
3. R for Data Science - Wickham and Grolemund - Data Science Cycle - https://r4ds.had.co.nz/introduction.html
4. Towards Data Science - Data Science Cycle - Sivakar Sivarajah - https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc
5. David Robinson (2017). unvotes: United Nations General Assembly Voting Data. R package version 0.2.0. https://CRAN.R-project.org/package=unvotes.
6. Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of International Organization, edited by Bob Reinalda (published May 27, 2013).
7. Much of the analysis has been modeled on the examples presented in the unvotes package vignette.