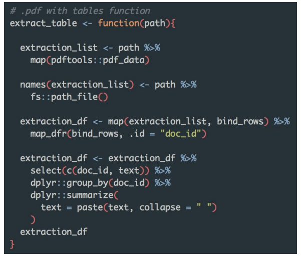
***Project VentureWell:***
*Data Science Corps Wrangle,*
*Analyze, Visualize*
*Spring & Fall 2020*

**BACKGROUND:** To maximize creativity, innovation incubator VentureWell allows applicants to submit proposals to their E-Teams Grant Program in an open-ended, narrative format. However, without consistency in structure or filetype, systematically searching and analyzing the thousands of proposals manually is challenging. So, our challenge was to extract proposal text data and wrangle it into a dataframe with tagged keyword counts.

## METHODS

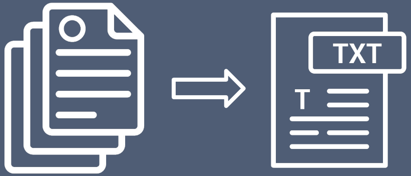
1. Used Scrum agile framework to stay organized
2. Communicated weekly with VentureWell staff
3. Explored text extraction packages, regular expressions, and SQL concepts
4. Leveraged technology (i.e. GitHub, Zoom, Slack) to continue work remotely
5. Wrote functions to extract text from for text-based and scanned .pdf, .doc, .and docx files
6. Combined functions into master script that produces dataframe upon execution
7. Created script to count occurence of keywords
8. Tested scripts on representative sample of proposal documents

## RESULTS



```r
# .pdf with tables function
extract_table <- function(path){

  extraction_list <- path %>%
    map(pdftools::pdf_data)

  names(extraction_list) <- path %>%
    fs::path_file()

  extraction_df <- map(extraction_list, bind_rows) %>%
    map_dfr(bind_rows, .id = "doc_id")

  extraction_df <- extraction_df %>%
    select(c(doc_id, text)) %>%
    dplyr::group_by(doc_id) %>%
    dplyr::summarize(
      text = paste(text, collapse = " ")
    )
  extraction_df
}
```

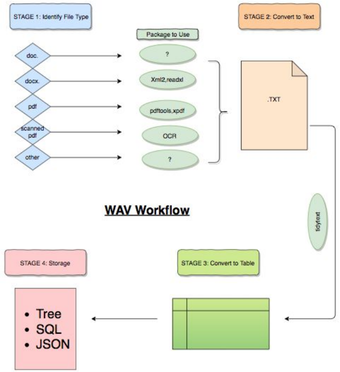
Sample of .pdf extraction function

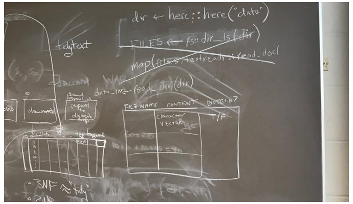


Manually searching and analyzing thousands of proposals is inefficient. We created a **function** to produce a data frame of **extracted text** from **documents** and to **tag keywords**.

**Take a picture** to learn more about DSC-WAV, our multi-institution, NSF-funded data science workforce development project!

https://dsc-wav.github.io/www/projects.html





PI Ben Baumer, Joyce Huang, Sunni Raleigh, Emma Scott, Rachel Yan, Annabel Yim