TEAM: ABRACA-DATA

HONGHUI WANG (301378426) CHHAVI VERMA (301371249) SHRAY KHANNA (301367221)



AGENDA

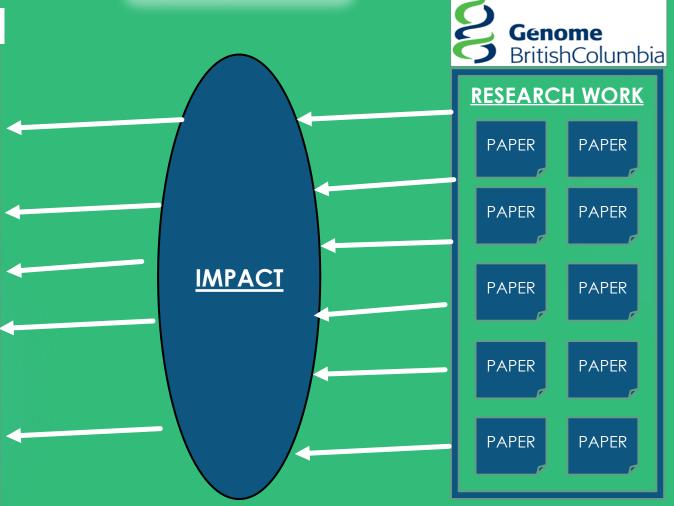
- > GOAL and MOTIVATION
- > SCHEDULE
- > PROGRESS
- > CHALLENGES
- > NEXT STEPS

OVERVIEW

GOAL

DOWNSTREAM DOCUMENTS



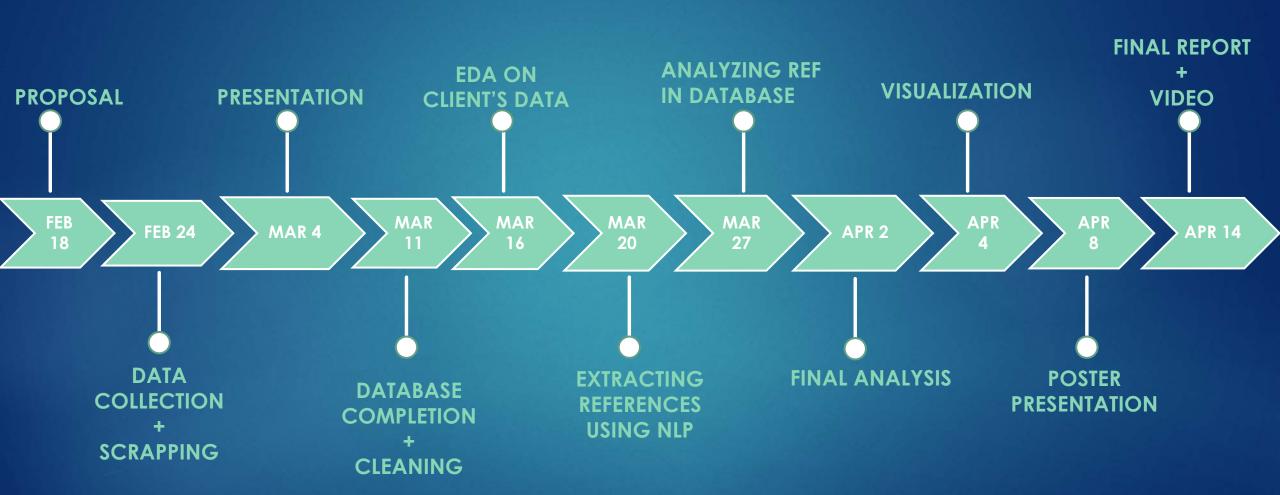




MOTIVATION

- ☐ Checking the impact of research publication in Healthcare for future implementation.
- ☐ Make healthcare research more impactful and beneficial to people.
- ☐ Help researchers understand the real world impacts of healthcare publications.

SCHEDULE



PROGRESS

DATA ACQUISITION

❖ DATA FROM 5 DIFFERENT SOURCES SHOWN BELOW:











- * MANUALLY DOWNLOADING UNSTRUCTURED DATA.
- *** USING TOOLS LEARNED TO SCRAP THE DOCUMENTS FROM THE GIVEN WEBSITES.**

```
import wellih
■ downloadPDFs_prj
- m downloadPDFs
 in pycache
 - m spiders
                           class downloadPDFs(scrapy.Spider):
   · m _pycache_
                              name = "downloadPDFs"

    _init_.py

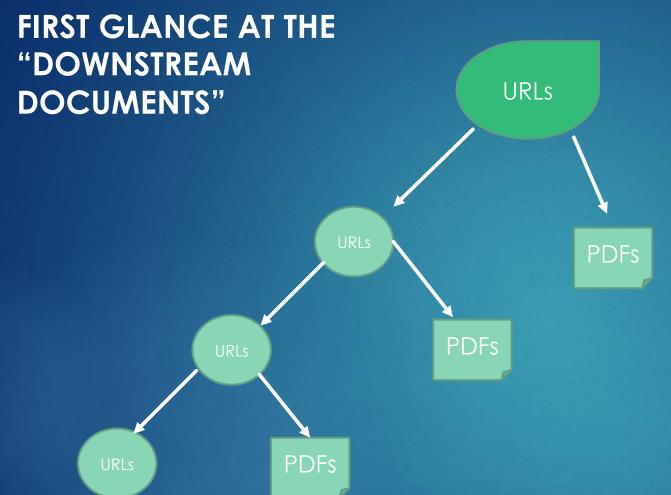
                               allowed domains = ["www.bccancer.bc.ca"]
  R_init_.py
  R items.pv
                                  "http://www.bccancer.bc.ca/health-professionals",
  ₽ middlewares.py

    □ pipelines.py

                                   #"http://www.bccancer.bc.ca/health-professionals/clinical-resources/chemotherapy-protocols/genitourinar
  - m pdfs
 → m about-site
                               def parse(self, response):
  books ...
                                   for href in response.css('a::attr(href)').extract():
 - m cancer-managemer
                                      if href.endswith('.pdf'):
   - m Documents
                                           yield response.follow(href, self.pdf save)
    2015 Follow-up
                                       else:
  - m centre-for-the-nort
                                         vield response.follow(href, self.parse)
   - m Documents
                                      # for *test* yield response.follow(href, self.test)
     III CN%20Guidebo
     © GEPsmallndfro
                              #def test(self, response):
   centre-for-the-soul
                              # print("Harry Wang"+response.url)
   - m Documents
     a Confidentiality. 25
     B CSI-April-suppo 26
     © CSI-February-st. 27
                                path = 'pdfs' + urllib.parse.urlparse(response.url).path
     IB CSI-March-supc 30
                                   self.logger.info(f'Saving PDF {path}')
     @ CSI-May-suppor
                                   dirnm = os.path.dirname(path)
     ≅ Kelowna_Relax
                                   if not os.path.exists(dirnm):
     © Launhter%20Y
                                     os.makedirs(dirnm)
     III MBSR%20Post
                                   with open(path, 'wb') as f
```

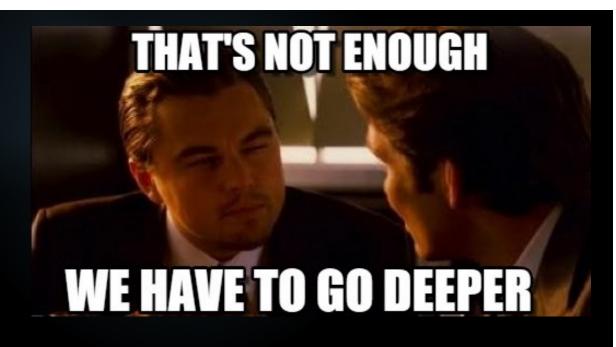
```
hon-your members and sept from
hon-your members and from
hon-your members an
```

CHALLENGES

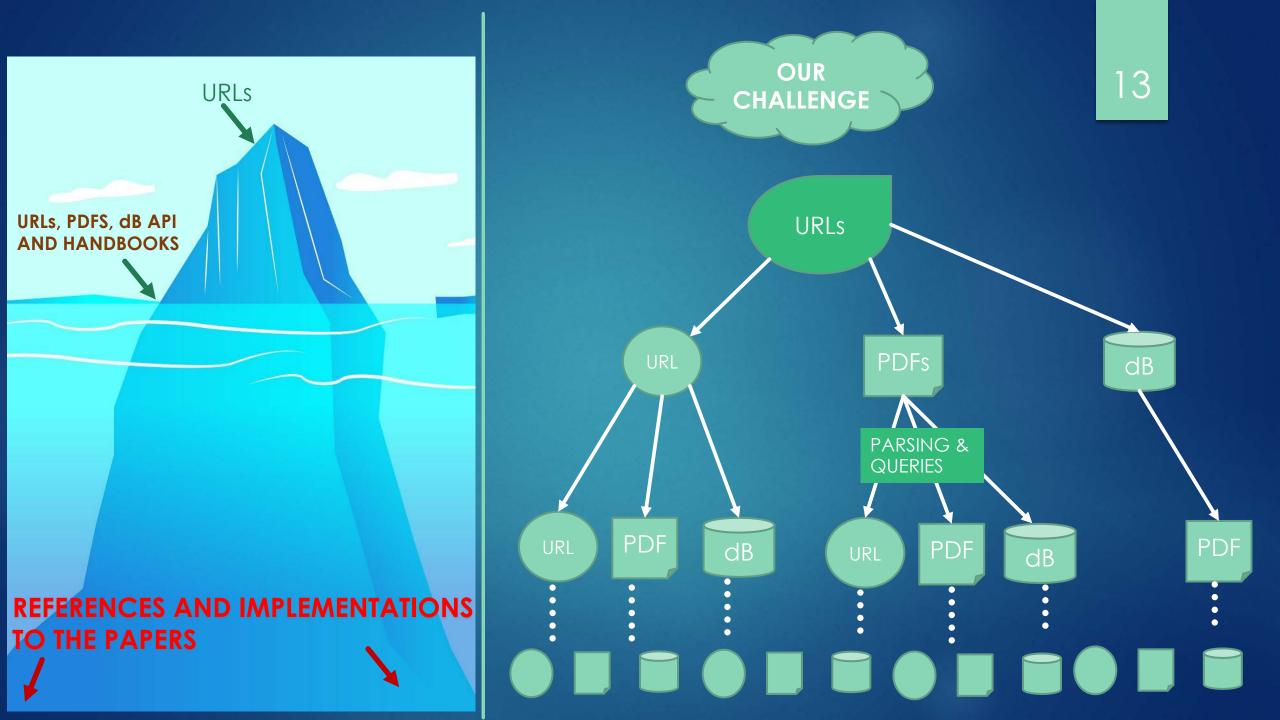


IS IT THIS SIMPLE TO SOLVE???



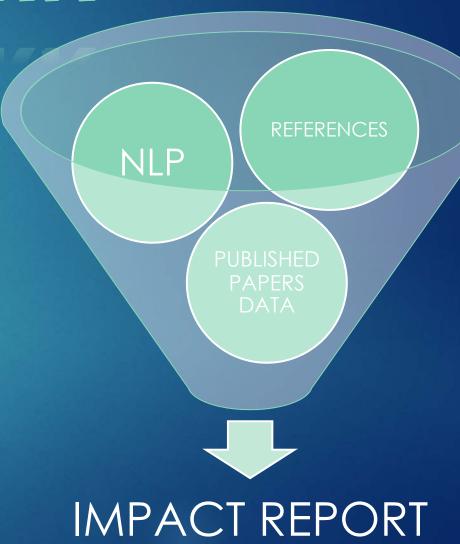




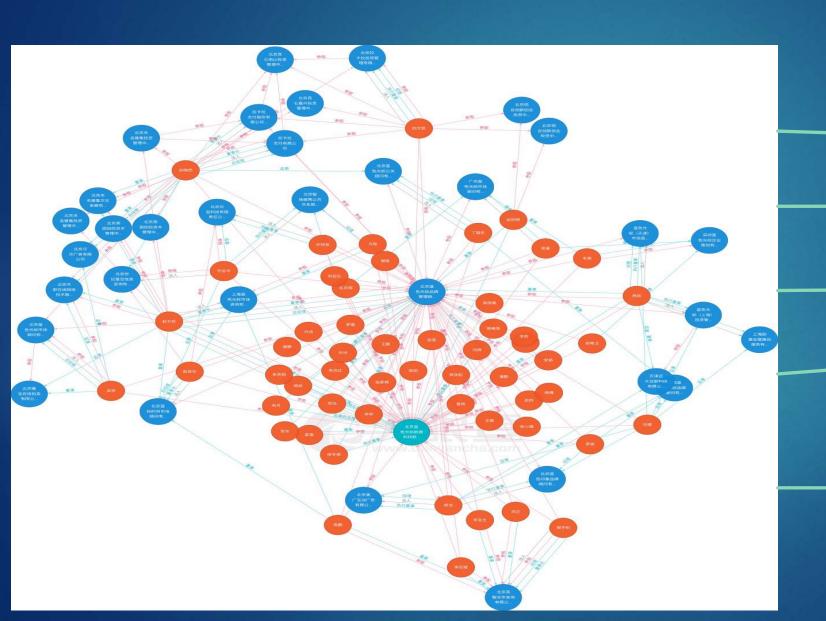


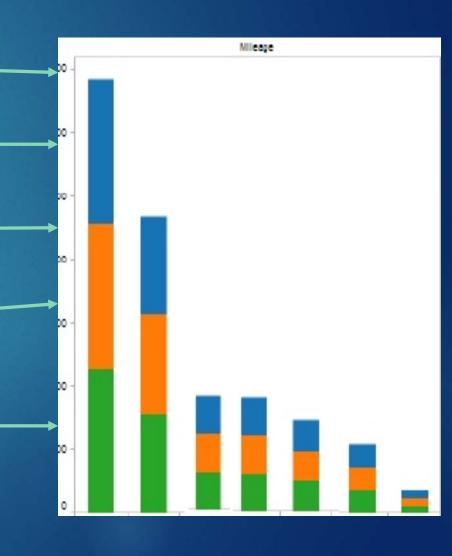
FUTURE WORK

- ❖ EXTRACTING REFERENCES TO BUILD THE DATABASE
- ❖ BUILD LABELLED TRAINING DATA FROM PAPERS WHO HAVE DOIS FOR NLP MODEL.
- ❖ USING NLP TO GET CORRECT KEYWORDS
- ❖ EDA ON GENOME BC's DATA
- ❖ GETTING IMPACTS FROM EACH PUBLICATION USING ANALYSIS
- ❖ VISUALIZATION USING TABLEAU
- ❖ AUTOMATE SCRAPING PROCESS TO TEST



OUR GOAL





QUESTIONS?

