

ANALYTICS & SOLUTIONS INTERNSHIP SUMMER 2025: CAPSTONE PROJECT

Luke Finkelstein

August 7, 2025



PERSONAL BACKGROUND

Education

Rising Senior @ Dickinson College
Double Major: Data Analytics & Mathematics

Activities Involved In

Captain of Men's Soccer, Admissions,
Fellow, and Tour Guide

Technical Skills

Python, SQL, R, Tableau, Machine Learning

Relevant Coursework

- Intro to Data Science
- Intro to Computing
- Probability & Statistics I & II
- Data Systems
- Statistical Machine Learning

Experience

- Data & Finance Intern @ Miller Environmental Group
- Technical Staff Data Analyst Intern @ Birmingham Legion, USL Championship
- Independent Machine Learning Projects (Waze, Salifort Motors)
- Web Scraping & Data Analysis Project (GoodReads)
- Coursera Google Advanced Data Analytics Certificate



ANALYTICS & SOLUTIONS INTERNSHIP STRUCTURE

- Program Managers: Dr. Anncy Thomas & Kimberly Velez
- 10-week internship (June–August 2025)
- Rotational experience: exposure to various teams across Northwell
- Mix of hands-on technical work, team meetings, lectures, and independent learning
- Mentorship from industry professionals and cross-functional teams

Week 1: Orientation & Healthcare Informatics

Week 2: Data Science

Week 3: Healthcare Analytics + Business Intelligence

Week 4: Software Engineering

Week 5: IT Project Management

Weeks 6-10: Capstone Project

CAPSTONE PROJECT: AUTOMATED EXTRACTION & CLASSIFICATION PIPELINE USING LLMS

Luke Finkelstein & Kevin Coppa

INTRODUCTION



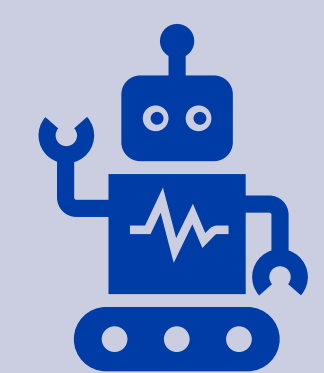
Project Set Up



Problem: Most research data is uncited in published papers (~86%, *Scientometrics*)



Business Value: Researchers don't receive credit & data isn't reliable/traceable



Host: Kaggle, an online community for machine learning & data science



Goal: Automate citation identification & classification

Challenges

- Unstructured data format
- Citations not provided
- Varying citation formats
(required extensive EDA)
- Ambiguity & context dependence
- Pipeline scale (optimization & speed)

What does it look like?

- Open Source Scientific Publications (biomedical)
- PDF Format
- 500+ training articles
- 30 testing articles

Citations:

1) Digital Object Identifiers (DOIs)

<https://doi.org/10.17882/49388>

2) Accession Numbers from databases (GEO, SRA, etc.)

GSE144193

RESEARCH ARTICLE

10.1002/2017JC013030

Key Points:

- The b_{bp} -to- $Chl a$ relationship varies along the water column, as well as with seasons and oceanic regions
- The b_{bp} -to- $Chl a$ ratio is a valuable biogeochemical proxy for assessing the nature of the particulate assemblage and revealing photoacclimation processes
- The BGC-Argo float network yields an unprecedented amount of quality data for studying biogeochemical processes at a global scale and along the vertical dimension

Supporting Information:

- Supporting Information S1
- Supporting Information S2
- Supporting Information S3
- Table S1
- Table S2
- Table S3

Correspondence to:

M. Barbieux,
barbieux@obs-vlfr.fr

Citation:

Barbieux, M., Uitz, J., Bricaud, A., Organelli, E., Poteau, A., Schmechtig, C., ... Claustre, H. (2018). Assessing the variability in the relationship between the particulate backscattering coefficient and the chlorophyll a concentration from a global Biogeochemical-Argo database. *Journal of Geophysical Research: Oceans*, 123, 1229–1250. <https://doi.org/10.1002/2017JC013030>

Received 26 APR 2017








Accepted 7 DEC 2017

Accepted article online 28 DEC 2017

Published online 15 FEB 2018

© 2017. American Geophysical Union.
All Rights Reserved.

Assessing the Variability in the Relationship Between the Particulate Backscattering Coefficient and the Chlorophyll a Concentration From a Global Biogeochemical-Argo Database

Marie Barbieux¹ , Julia Uitz¹, Annick Bricaud¹, Emanuele Organelli^{1,2} , Antoine Poteau¹ , Catherine Schmechtig³ , Bernard Gentili¹, Grigor Obolensky⁴, Edouard Leymarie¹ , Christophe Penker¹, Fabrizio D'Ortenzio¹ , and Hervé Claustre¹ 

¹Sorbonne Universités, UPMC Univ Paris 06, CNRS, Observatoire Océanologique de Villefranche, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France, ²Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, United Kingdom, ³OSU Ecce Terra, UMS 3455, CNRS and Université Pierre et Marie Curie, Paris 6, Paris, France, ⁴ERIC Euro-Argo, 29280 Plouzané, France

Abstract Characterizing phytoplankton distribution and dynamics in the world's open oceans requires in situ observations over a broad range of space and time scales. In addition to temperature/salinity measurements, Biogeochemical-Argo (BGC-Argo) profiling floats are capable of autonomously observing at high-frequency bio-optical properties such as the chlorophyll fluorescence, a proxy of the chlorophyll a concentration ($Chl a$), the particulate backscattering coefficient (b_{bp}), a proxy of the stock of particulate organic carbon, and the light available for photosynthesis. We analyzed an unprecedented BGC-Argo database of more than 8,500 multivariable profiles collected in various oceanic conditions, from subpolar waters to subtropical gyres. Our objective is to refine previously established $Chl a$ versus b_{bp} relationships and gain insights into the sources of vertical, seasonal, and regional variability in this relationship. Despite some regional, seasonal and vertical variations, a general covariation occurs at a global scale. We distinguish two main contrasted situations: (1) concomitant changes in $Chl a$ and b_{bp} that correspond to actual variations in phytoplankton biomass, e.g., in subpolar regimes; (2) a decoupling between the two variables attributed to photoacclimation or changes in the relative abundance of nonalgal particles, e.g., in subtropical regimes. The variability in the b_{bp} : $Chl a$ ratio in the surface layer appears to be essentially influenced by the type of particles and by photoacclimation processes. The large BGC-Argo database helps identifying the spatial and temporal scales at which this ratio is predominantly driven by one or the other of these two factors.

1. Introduction

Our ability to observe the dynamics of phytoplankton biomass and associated carbon fluxes on relevant space and time scales considerably limits our understanding and prediction skills of the biogeochemical role of phytoplankton in the carbon biological pump (Honjo et al., 2014; Legendre et al., 2015; Volk & Hoffert, 1985). For example, in situ measurements of primary production and phytoplankton carbon biomass are particularly challenging and remain scarce, although novel promising techniques have been recently proposed (Graff et al., 2012, 2015; Riser & Johnson, 2008). To overcome space-time coverage sampling limitations, bio-optical oceanographers have implemented optical sensors on a variety of in situ or remote platforms, from research vessels and moorings to ocean color satellites, gliders, and profiling floats, each with specific complementary space-time observation scales (Claustre et al., 2010; Dickey, 2003). Such platforms enable to monitor bio-optical properties that serve as proxies for major biogeochemical variables. Those include the concentration of chlorophyll a ($Chl a$) and the particulate backscattering coefficient at 700 nm (hereafter referred simply as b_{bp}). The chlorophyll a concentration is the most commonly used proxy for the phytoplankton carbon biomass (Cullen, 1982; Siegel et al., 2013), although it is well known that the ratio of $Chl a$ to carbon shows large fluctuations driven by a variety of factors such as phytoplankton physiology (Álvarez et al., 2016; Geider, 1993; Staehr et al., 2002) or community composition (Geider et al., 1997; Halsey & Jones, 2015; MacIntyre et al., 2002). In the absence of mineral particles (i.e., in most open ocean waters), b_{bp} generally covaries with, and is therefore used as a proxy of, the stock of particulate organic carbon (POC; Bishop,

GLOSSARY

LLM: Large Language Models (GPT, Gemini, Copilot, etc.)

Prompt Engineering: designing instructions passed to LLM to get better results

Zero-Shot Prompting: Giving LLM instructions with no examples

Few-Shot Prompting: Instructions + examples

Parsing: Extracting & Analyzing info from a file

Natural Language Processing (NLP): Field of AI focused on interaction between human & machine languages

Named Entity Recognition (NER): Technique to locate + classify parts of text as certain categories (names, locations, etc.)

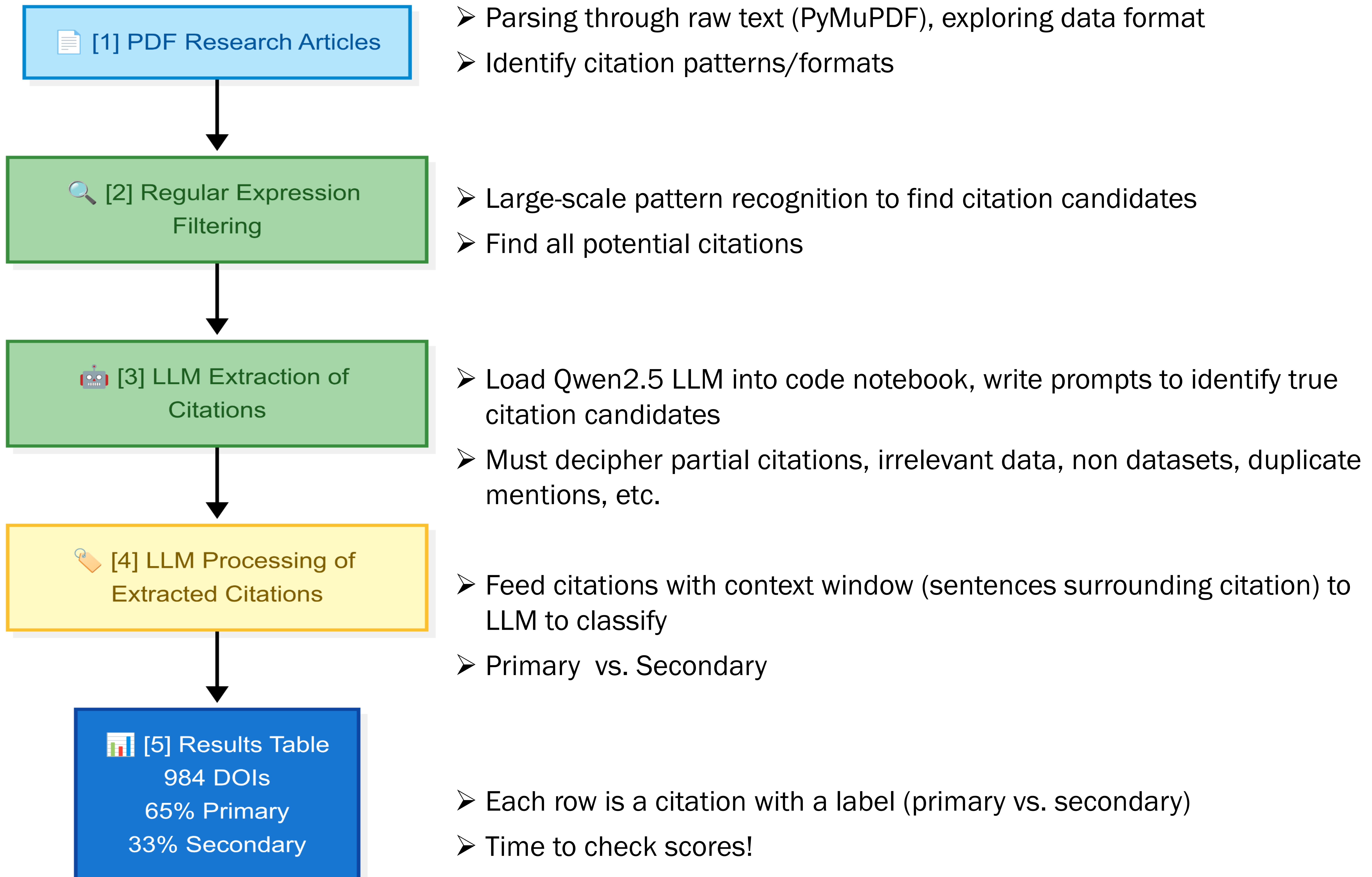
Classification: Assigning labels to data (emails, news, images, etc.)

Regular Expression: a sequence of characters defining a pattern for finding/matching text (ex: ^cat)

PIPELINE

Why Qwen2.5?

- Strong performance with classification
- Fast processing for large scale project
- Open Source



SCORING METRICS

Recall: Proportion of true positive citations correctly identified
(important during **extraction** phase)

Precision: Proportion of positive predictions of citations actually correct
(important during **classification** phase)

F1 Score: Harmonic mean of Precision & Recall
Main Metric for evaluating this project

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

ITERATIONS OF PIPELINE & RESULTS

V1

1. PDF Research Articles
2. Regular Expression Filtering & Citation Extraction
 - Strict pattern criteria (low recall)
3. LLM Processing of Extracted Citations
 - Model: Qwen2.5-0.5B-instruct
4. Results Table

V2

1. PDF Research Articles
2. Regular Expression Filtering
 - Broader criteria (boosted **recall**)
3. **LLM Extraction of Citations**
 - Model: Qwen2.5-0.5B-instruct
 - More true mentions (boosted **precision**)
4. LLM Processing of Extracted Citations
 - Model: Qwen2.5-0.5B-instruct
5. Results Table

V3

1. PDF Research Articles
2. Regular Expression Filtering
3. LLM Etraction of Citations
 - **Model: Qwen2.5-3B-instruct**
 - Improved prompts → better **recall**
4. LLM Processing of Extracted Citations
 - **Model: Qwen2.5-3B-instruct** (stronger model, boosted **precision** for classification)
5. Results Table

ITERATIONS OF PIPELINE & RESULTS

V1

- 1. PDF Research Articles
- 2. Regular Expression Filtering & Citation Extraction
 - Strict pattern criteria (low recall)
- 3. LLM Processing of Extracted Citations
 - Model: Qwen2.5-0.5B-instruct
- 4. Results Table

V2

- 1. PDF Research Articles
- 2. Regular Expression Filtering
 - Broader criteria (boosted **recall**)
- 3. **LLM Extraction of Citations**
 - Model: Qwen2.5-0.5B-instruct
 - More true mentions (boosted **precision**)
- 4. LLM Processing of Extracted Citations
 - Model: Qwen2.5-0.5B-instruct
- 5. Results Table

V3

- 1. PDF Research Articles
- 2. Regular Expression Filtering
- 3. LLM Etraction of Citations
 - **Model: Qwen2.5-3B-instruct**
 - Improved prompts → better **recall**
- 4. LLM Processing of Extracted Citations
 - **Model: Qwen2.5-3B-instruct** (stronger model, boosted **precision** for classification)
- 5. Results Table

	V1	V2	V3	Difference
Precision	0.05	0.14	0.25	+ 0.20
Recall	0.05	0.13	0.35	+ 0.30
Accuracy	0.10	0.41	0.64	+ 0.54
F1	0.05	0.11	0.229	+ 0.224

LEARNING GOALS ACHIEVED



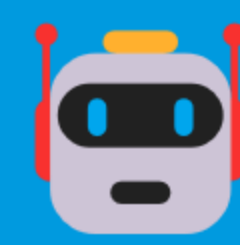
PDF Text parsing in Python (**PyMuPDF**, **fitz**)



Advanced Regular Expression patterns for identification (**re**)



Data cleaning, deduplication, and pipeline automation



NLP with Large Language Models (**transformers**, **torch**)



Prompt engineering (**transformers**)



Model evaluation: Accuracy, Precision, Recall, F1



Debugging & workflow best practices (**tqdm**)



Iterative development & improvement



Real-world data science challenges

FUTURE WORK & BROADER VALUE

Potential Improvements

- Experiment with few-shot prompting examples
- Experiment with different models
- Larger, Diverse Training Set
- Allocate more time to Pattern Recognition
- Train my own models on research paper data for improved extraction & classification (to boost recall + precision)

Similar Applications

- Common NLP tasks like NER, Classification of large texts
- Insights from studies in PDF format, eliminates repetitive tasks
- Pipeline to analyze/classify unstructured data
 - Ex: summarizing patient notes using LLMs

AUTOMATED EXTRACTION OF DATA CITATIONS FROM BIOMEDICAL RESEARCH PAPERS

LUKE FINKIELSTEIN & KEVIN COPPA



Project Workflow

➤ BACKGROUND (or INTRODUCTION)

- Analyzed a dataset of scientific research PDFs—primarily in biomedical fields—provided as part of a Kaggle competition.
- Premise was to create a model capable of extracting and classifying the data citations present in the research papers
- Opportunity originated with a Kaggle competition focused on benchmarking automated approaches for extracting dataset citations from research papers. (Make Data Count)

➤ OPPORTUNITY (or CHALLENGE)

- The challenge: Identify and classify data citations within thousands of full-text biomedical articles using scalable automated methods.
- Importance: Manual extraction from scientific literature is impractical at scale, only solution is automation
- 2-part project:
 - 1) Named Entity Recognition
 - 2) Classification

➤ GOALS (or AIMS or OBJECTIVES)


- Develop an NLP pipeline to locate, extract and classify dataset citations within the context of research papers
- Primary metric of interest: F1 score on official Kaggle test data (no concrete passing threshold, just relative leaderboard performance).
- Produce a reproducible, modular workflow suitable for competitive and real-world analytics settings

 [1] PDF Research Articles

 [2] Regular Expression Filtering

 [3] LLM Extraction of Citations

 [4] LLM Processing of Extracted Citations

 [5] Results Table
984 DOIs
65% Primary
33% Secondary

INTERVENTION (or SOLUTION)

- Parsed biomedical research PDFs into context-rich windows, then used regex and a large language model (LLM) to extract and classify valid data citation DOIs.
- Applied robust, automated post-processing: filtered self/preprint DOIs, deduplicated citations, and labeled each as Primary, Secondary, or Irrelevant data.
- Implemented the entire workflow in Python with Pandas, regex, torch, and HuggingFace transformers.

OUTCOMES

- F1 score: 0.229, demonstrating measurable performance in dataset citation extraction & classification
- Improvement of + 0.224 from baseline
- Significant improvement in prediction accuracy and reliability across pipeline iterations
- Validated the effectiveness of combining rule-based filters with advanced language models.
- Efficient pipeline delivery actionable insights from unstructured data
- Framework can be transferred to similar real-world tasks involving entity recognition and classification across industries.

REFLECTIONS

- Accurate data citation extraction is harder than expected—context and prompt wording are critical
- LLMs enable fast automation, but rules and post-processing are still essential for precision.
- Surprised by frequent misclassification of repository DOIs; classification remains an open challenge.
- Learned the importance of iterative testing and validation on real-world data.
- Refining simple, downstream filters often yields bigger impact than tuning the extraction model itself.

WRAP UP

Broad exposure of Northwell teams within Analytics & Solutions

Opportunity to work with many industry-standard tools, like...

Kaggle, Google Colab, Smartsheet, JIRA, AI Hub, VS Code, AMIA Learning Center, SQL, Tableau

Deeper understanding of possible work initiatives within Data