# ML Solutions for Publication Growth

Dawid Kubkowski

Supervised by:
Dr Michał Burdukiewicz
Dr Grzegorz Jagiella

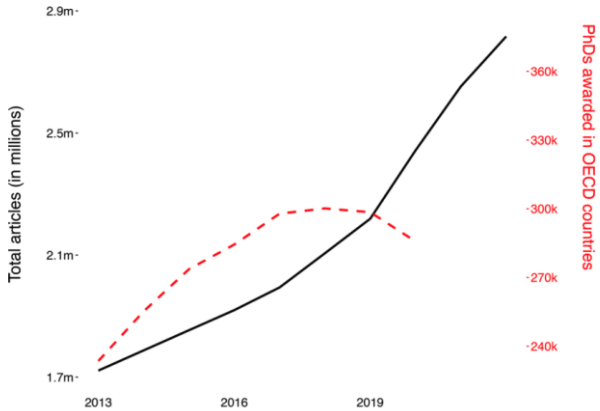
April 15, 2025

# Introduction

- The exponential rise in scientific publications has made manual review impractical.
- Researchers struggle to identify relevant studies amidst the overwhelming volume of articles.
- Our solution: An ML-driven approach to automatically assess and prioritize research articles based on relevance.

# Publication Deluge: The Growing Challenge

The number of scientific publications is increasing rapidly.




Source: Hanson et al., 2024, *The Strain on Scientific Publishing*, *Quantitative Science Studies*, vol. 5, no. 4, pp. 823–843, DOI:10.1162/qss$_{a0}$0327, licensed under CC BY-NC-SA 4.0.

# Limitations of Manual Curation

Traditional methods struggle to keep pace:

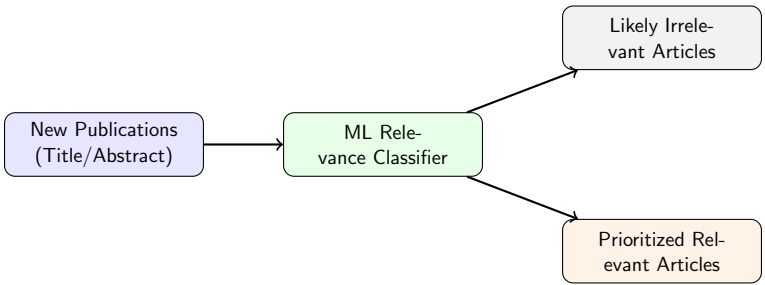
- **Time-Intensive:** Screening thousands of articles manually is extremely slow.
- **Costly:** Requires significant human resources and expertise.
- **Inconsistent:** Subjectivity and reviewer fatigue can lead to errors and biases.

A scalable, automated solution is needed.

# Our Approach: AI-Powered Prioritization

We propose using Machine Learning (ML) to assist researchers:

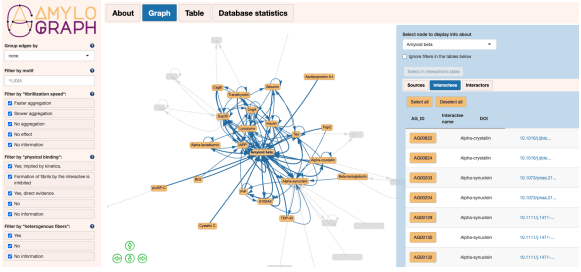
- **Goal:** Develop an automated system to classify research articles based on relevance using titles and abstracts.
- **Benefit:** Prioritize potentially relevant papers, significantly reducing manual screening effort.
- **Method:** Train classification models on curated datasets to predict article relevance.



# Amyloids & AmyloGraph

To demonstrate our approach, we focus on the challenging domain of amyloid research:

- We focus on amyloids—proteins involved in neurodegenerative disorders.
- Understanding amyloid interactions is crucial, as they can contribute to disease onset.
- A major challenge: Different experimental techniques highlight different aspects of these interactions.
- To standardize this knowledge, we developed an ontology for amyloid interactions and created **AmyloGraph**, the first dedicated database for these interactions.

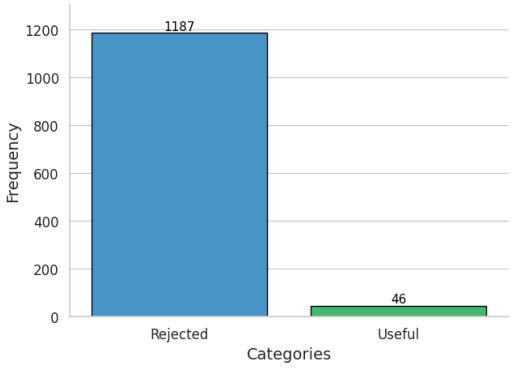


# Data Curation: Building the Training Set for Amyloids

High-quality labeled data is essential for training the ML model.

**Curation Process:**

- Manually reviewed articles based on title/abstract.
- Labeled articles as **Relevant** (Accepted) or **Non-Relevant** (Rejected).
- **Criteria for Relevance:** Must report experimental data on antibody-amyloid interactions affecting amyloid formation.
- **Common Rejection Reasons:**
  - Review articles, preprints, non-English
  - Missing experimental data
  - Irrelevant scope (e.g., in silico only, wrong protein type)




Distribution of Relevant vs. Non-Relevant articles.

Note the class imbalance.

# Model Pipeline: Preprocessing

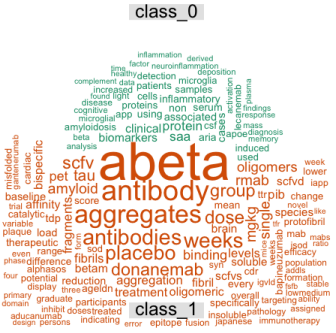
- We are developing an ML model to classify research articles based on title and abstract.
- **Preprocessing:**
  - Text cleaning and creating a document-term matrix (DTM), which counts how often each word appears in a document.

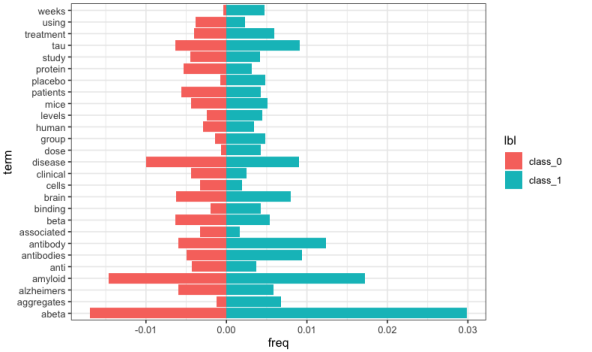
```
corpus <- create_corpus(c("This is first example.",
                "This is another example",
                "This is another another example"))

words = find_words(corpus, frequency = 1)
create_document_term_matrix(corpus, words)
```

```
##   example first another
## 1       1     1       0
## 2       1     0       1
## 3       1     0       2
```

# Exploratory Data Analysis

Visualization of Common Words in the Text



# Model Pipeline: Classification

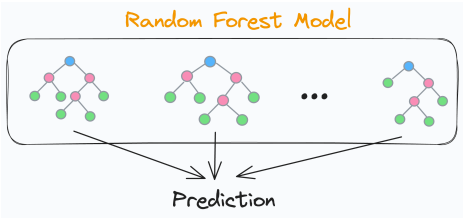
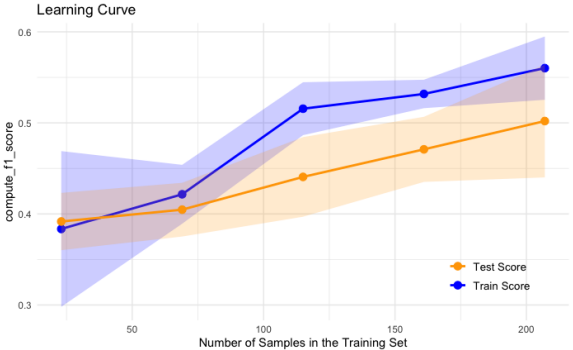



Figure source: blog.dailydoseofds.com

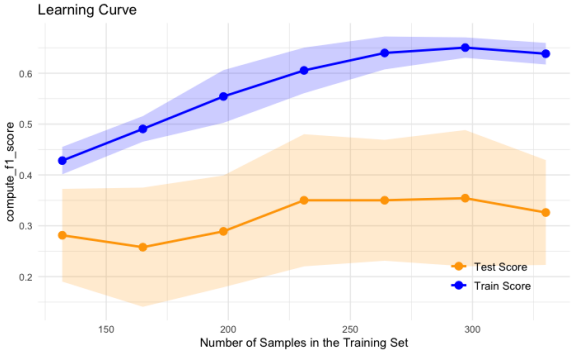
- **Classification:**
  - Binary classification model (relevant vs. non-relevant) trained on curated labels.
- **Evaluation Metrics:**
  - Accuracy, precision, recall, F1-score.

# Learning Curves

Do we benefit from more data? Learning curves show model performance vs. training set size.




Dataset: Peptide




Dataset: Amyloids

# Hyperparameter Grid Search

Fine-tuning the model for better generalization.

- We systematically explored different Random Forest settings (`mtry`, `nodesize`, `ntree`, etc.) using Grid Search with Cross-Validation.
- **Goal:** Find the hyperparameter combination that yields the best performance on unseen data (test set).
- Performance was assessed using mean training score, cross-validation score, and test score.

| | mtry | nodesize | ntree | class_val | min_freq | threshold | mean_train_score | cv_score | test_score |
|---|---|---|---|---|---|---|---|---|---|
| 12223 | 0.03 | 5 | 100 | 0.4 | 5 | 0.18 | 0.9374350 | 0.5109524 | 0.3636364 |
| 11416 | 0.03 | 1 | 100 | 0.5 | 7 | 0.14 | 0.9966102 | 0.5032756 | 0.4571429 |
| 16480 | 0.07 | 10 | 100 | 0.5 | 7 | 0.22 | 0.8524413 | 0.5003868 | 0.3589744 |
| 5918 | sqrt | 1 | 100 | 0.3 | 7 | 0.12 | 1.0000000 | 0.4929365 | 0.2580645 |
| 14434 | 0.05 | 10 | 100 | 0.3 | 5 | 0.24 | 0.8761981 | 0.4851732 | 0.2500000 |
| 6818 | sqrt | 10 | 100 | 0.1 | 7 | 0.18 | 0.9075274 | 0.4841270 | 0.2400000 |
| 16481 | 0.07 | 10 | 100 | 0.5 | 7 | 0.24 | 0.8983229 | 0.4790476 | 0.4516129 |
| 13071 | 0.05 | 1 | 100 | 0.3 | 7 | 0.12 | 0.9731007 | 0.4785931 | 0.3243243 |

# Acknowledgements

We gratefully acknowledge the support for this research:

- **Institution:**
  Bioinformatics and Multiomics Analysis Laboratory,
  Clinical Research Centre,
  Medical University of Białystok.

- **Funding Source:**
  National Science Center, Poland, via the SONATA 19 grant.
  Project No: DEC-2023/51/D/NZ7/02847.

- **Project Title:**
  *"Taming aggregation with AmyloGraphem 2.0: database and predictive model of amyloid self-organization of modulators"*.