ScholarSuccess

An End-to-End Pipeline for Conference Acceptance rate Prediction

A Project for 2110403 Introduction to Data Science and Data Engineering

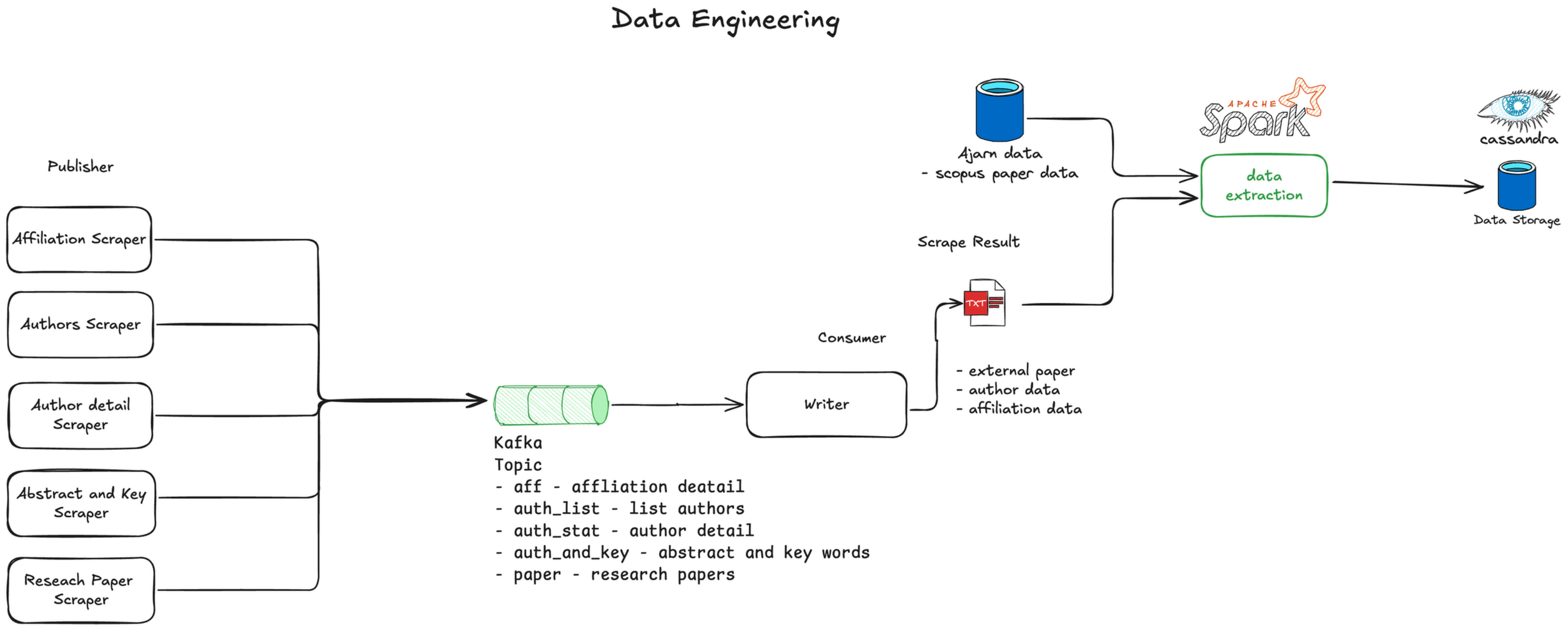
CΣDT

WHY CONFERENCE PROCEEDING ?

- Search for an important and efficient way to make others aware of a researcher’s work and have it visible faster when an academic conducts a literature review

- help make researchers aware of new insights and industry innovations, while citing research underway on a given subject

DATA ENGINEERING



DATA COLLECTION

- Data Provided (2018-2023)


```
{"abstracts-retrieval-response": {
    "item": {…
    },
    "affiliation": […
    ],
    "coredata": {…
    },
    "idxterms": {"mainterm": […
    ]},
    "language": {"@xml:lang": "eng"},
    "authkeywords": null,
    "subject-areas": {"subject-area": […
    ]},
    "authors": {"author": […
    ]}
}}
```


20216 records

- Data Scraped
  - Research Papers
  - Authors List of Research Papers
  - Author Details
  - Abstract, Reference, Keywords of Research Papers
  - Affiliation Details

Scopus®

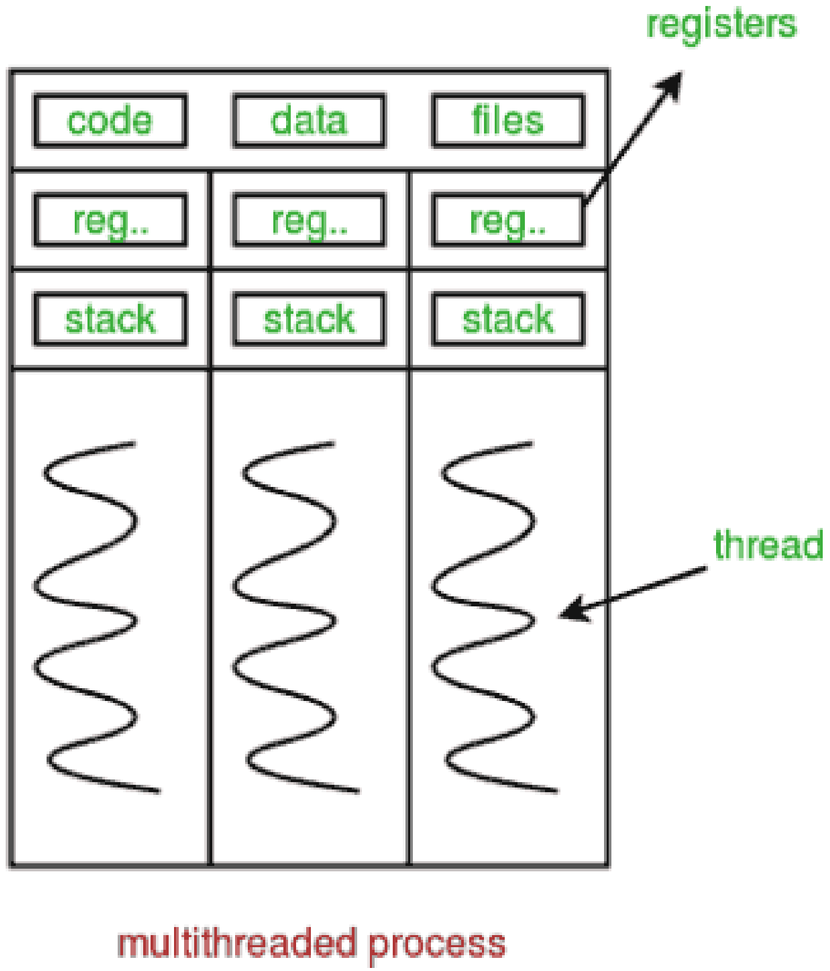
Elsevier Developer Portal

DATA COLLECTION

- Scraping Strategy

```
import concurrent.futures

with concurrent.futures.ThreadPoolExecutor(max_workers=self.max_workers) as executor:
  author_results = list(filter(None, executor.map(self.scrape_author_data, authors)))
  .
  .
  .
```

```
from concurrent.futures import ThreadPoolExecutor, as_completed

with ThreadPoolExecutor(max_workers=self.max_workers) as executor:
            progress_bar = tqdm.tqdm(
                total=len(self.data),
                desc="Fetching Abstracts",
                unit="paper"
            )

            futures = {
                executor.submit(self._process_paper, paper): paper
                for paper in self.data
            }

            for future in as_completed(futures):
                result = future.result()
                .
```



DATA COLLECTION

- Data Scraped
  - Research Papers
  - Authors List of Research Papers

42565 records

```
headers = {
  'X-ELS-APIKey': self.api_key,
  'Accept': 'application/json'
}
```

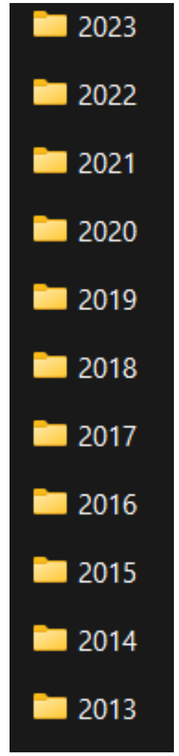
```
params = {
  "query": queries[query_option],
  "start": start_index,
  "count": fetch_count,
  "field": ...
   }
```

```
"https://api.elsevier.com/content/search/scopus"
response = requests.get(base_url, headers=headers, params=params)
```

```
response = self.session.get(url, headers=headers, timeout=10)
```

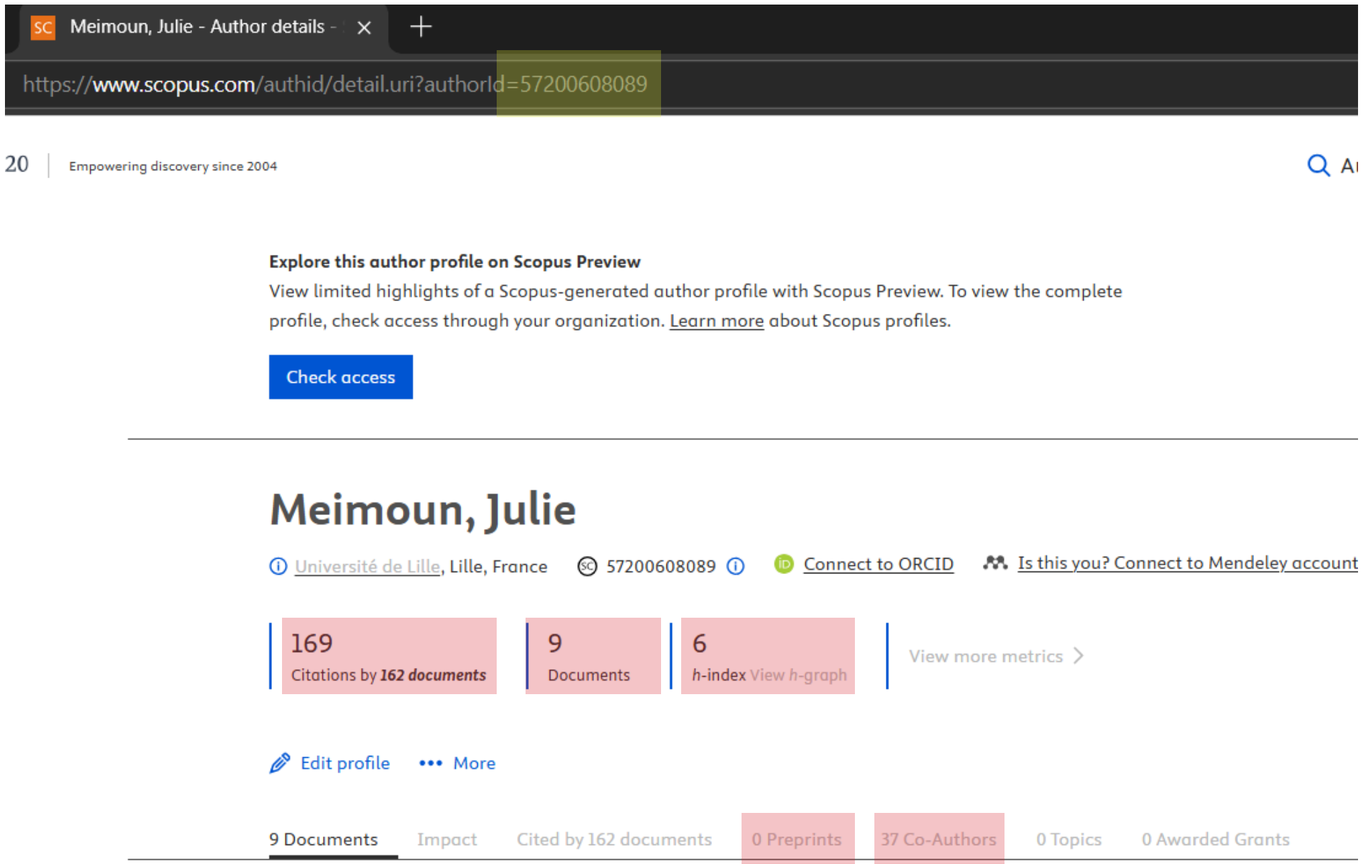
Scopus API


```
{
    "@_fa": "true",
    "prism:url": "https://api.elsevier.com/content/abstract/scopus_id/84891140835",
    "dc:identifier": "SCOPUS_ID:84891140835",
    "eid": "2-s2.0-84891140835",
    "dc:title": "Vice chairman's message",
    "dc:creator": "Gulati S.",
    "prism:publicationName": "2013 International Conference on Control, Computing, C
    "prism:isbn": [ ···
    ],
    "prism:pageRange": null,
    "prism:coverDate": "2013-12-31",
    "prism:coverDisplayDate": "2013",
    "prism:doi": "10.1109/ICCCCM.2013.6648897",
    "citedby-count": "0",
    "prism:aggregationType": "Conference Proceeding",
    "subtype": "ed",
    "subtypeDescription": "Editorial",
    "openaccess": "0",
    "openaccessFlag": false,
    "author": [ ···
    ],
    "source-id": "21100276776"
},
```

DATA COLLECTION

- Data Scraped
  - Author Details



Browser Scraping

76182 records

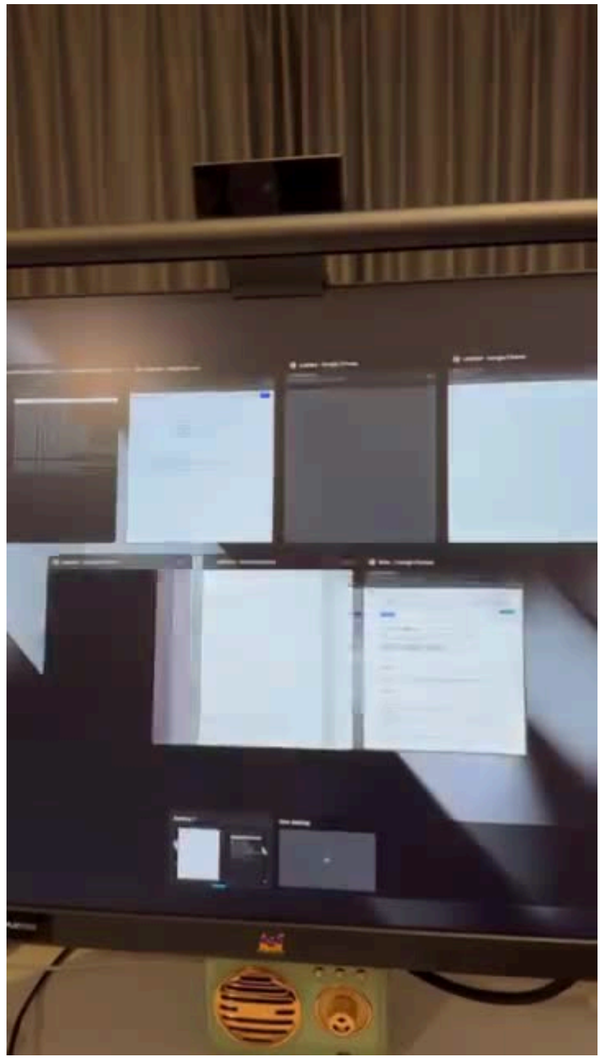

```
{
  "auid": "22980355000",
  "author_stat": {
    "citations": 25617,
    "documents": 481,
    "h_index": 80,
    "co_authors_count": 11143,
    "cited_by_count": 10306,
    "preprints_count": 228
  }
},
```

DATA COLLECTION

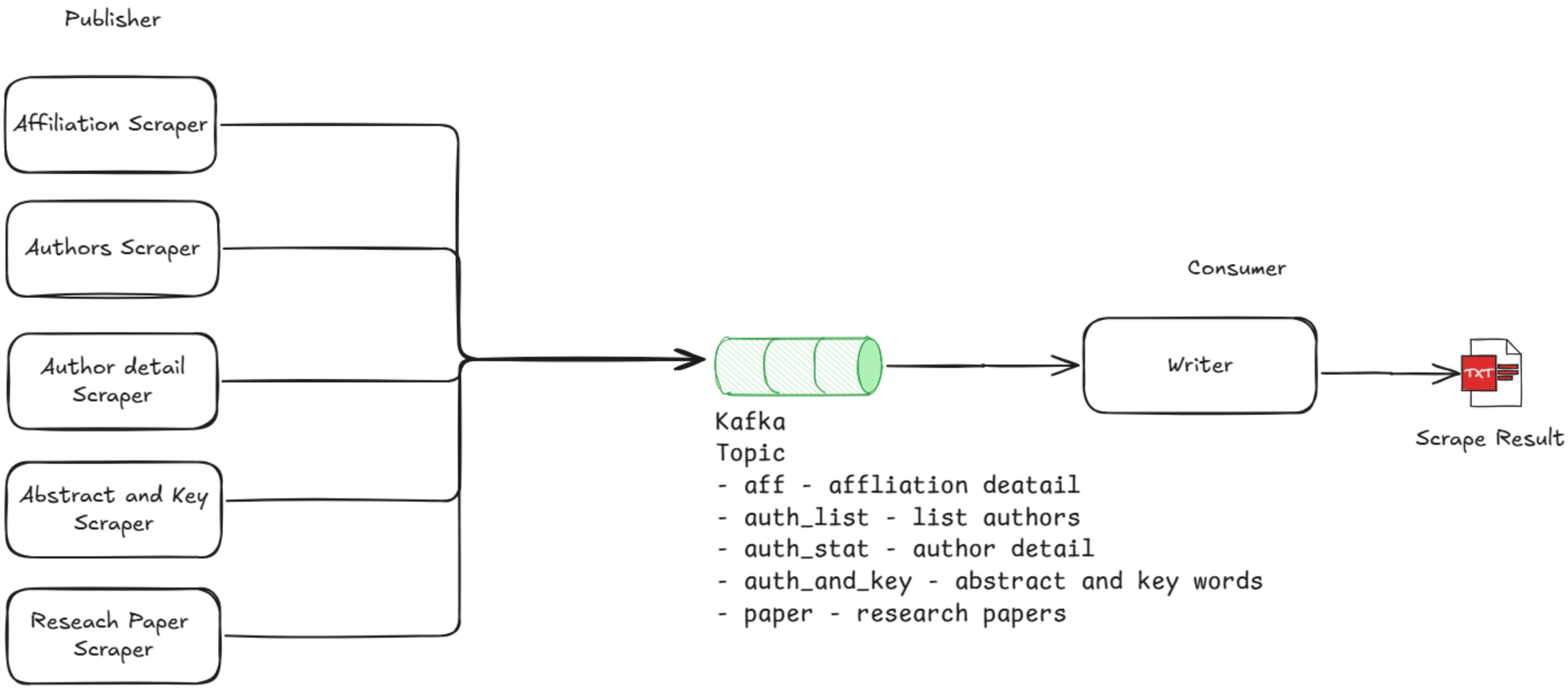
- Data Scraped
  - Author Details

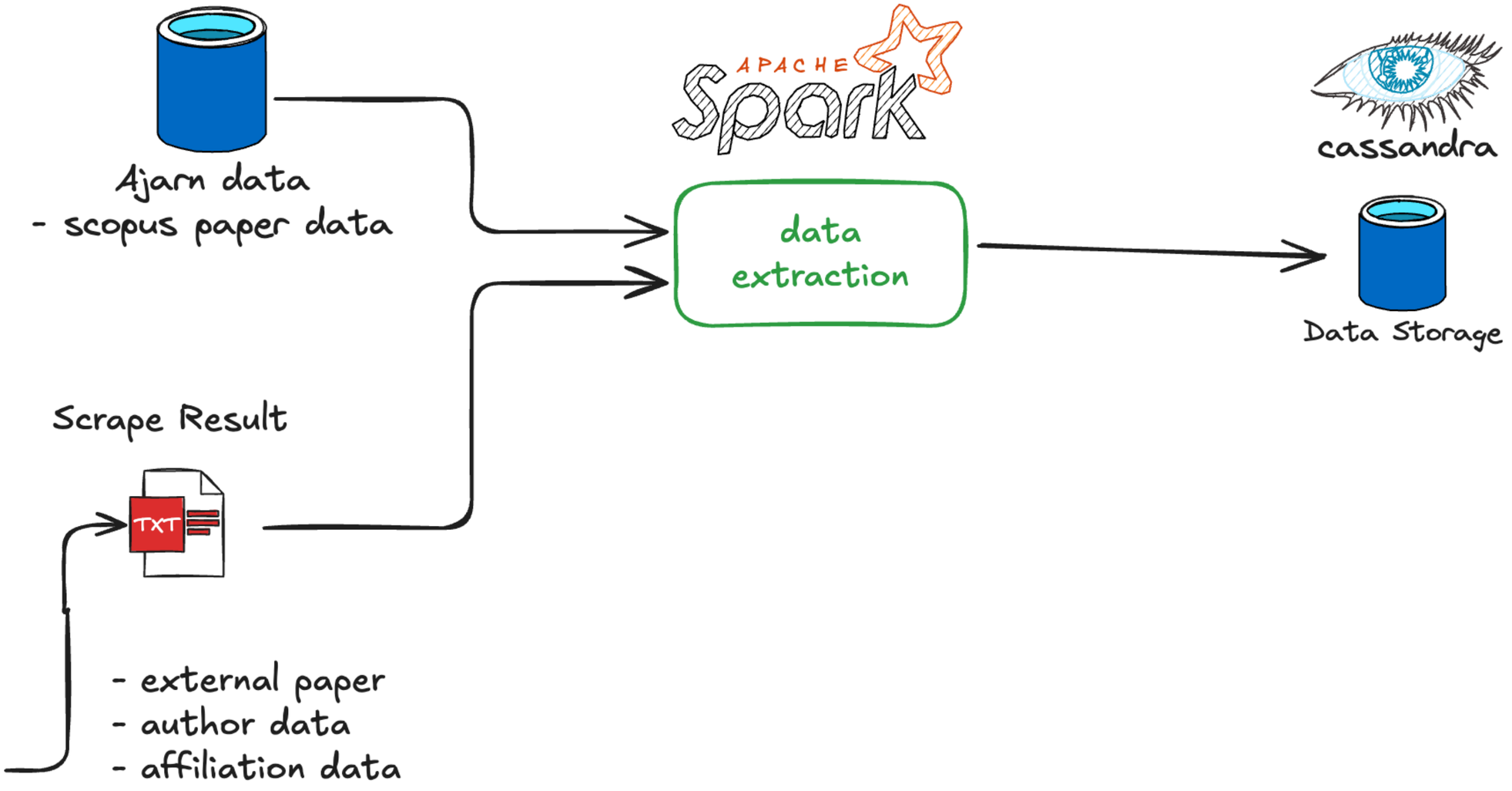
```
author_id = author.get('@auid')

driver = self._create_webdriver()

url = f'https://www.scopus.com/authid/detail.uri?authorId={author_id}'
driver.get(url)

return self._extract_metrics(driver, author_id)
```

```
element = driver.find_element(By.CSS_SELECTOR, f"{selector} > span")
```



DATA INGESTION




Ajarn data
- scopus paper data
APACHE
Spark
cassandra
data
extraction
Data Storage
Scrape Result
TXT
- external paper
- author data
- affiliation data

```
session.execute("""
    CREATE TABLE IF NOT EXISTS scopus_data.records (
        doi text PRIMARY KEY,
        title text,
        abstract text,
        document_type text,
        source_type text,
        publication_date text,
        source_title text,
        publisher text,
        author_keywords text,
        subject_code text,
        subject_name text,
        subject_abbrev text,
        author_given_name text,
        author_surname text,
        author_url text,
        author_affiliation text,
        author_details text,
        funding_details text,
        ref_count int,
        open_access boolean,
        affiliation text,
        language text,
        cited_by int,
        status_state text,
        delivered_date text,
        subject_area text
    )
""")
```

```
session.execute("""
            CREATE TABLE IF NOT EXISTS scopus_data.affiliations (
                affiliation_id text PRIMARY KEY,
                preferred_name text,
                name_variants list<text>,
                documents_count int,
                authors_count int,
                country text,
                city text,
                state text,
                street_address text,
                postal_code text,
                contact_url text,
                hierarchy_ids list<text>
            )
        """)
```

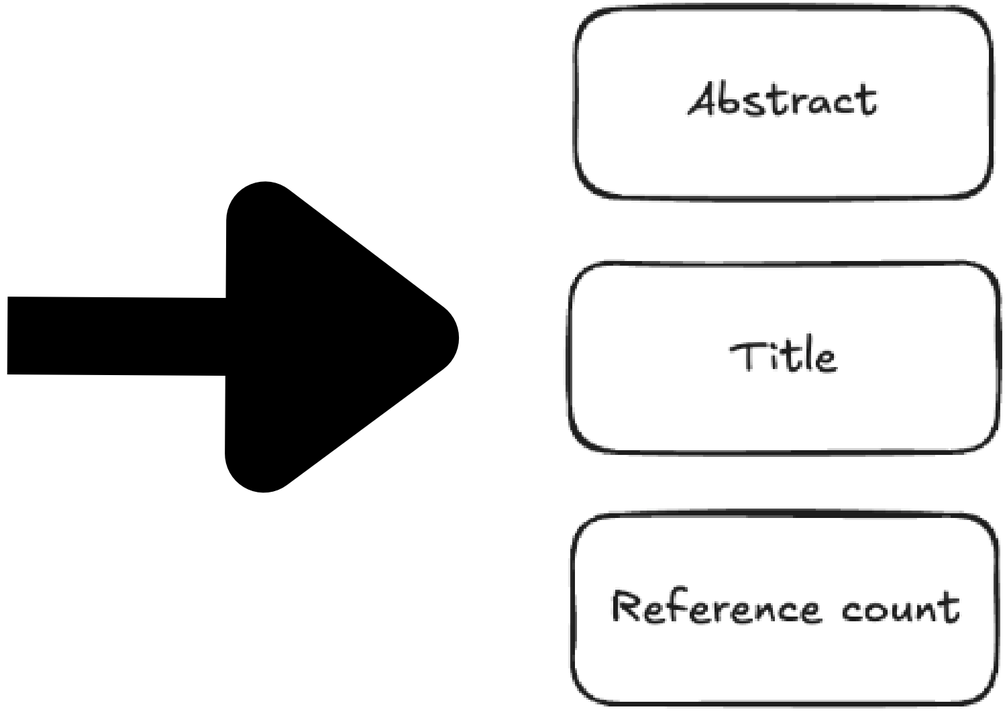
```
session.execute("""
            CREATE TABLE IF NOT EXISTS scopus_data.author_metrics (
                author_id text PRIMARY KEY,
                citations bigint,
                documents int,
                h_index int,
                co_authors_count int,
                cited_by_count bigint,
                preprints_count int
            )
        """)
```

- records 62781
- author_metrics: 76182
- affiliations: 7625

Data Processing

```
session.execute("""
    CREATE TABLE IF NOT EXISTS scopus_data.records (
        doi text PRIMARY KEY,
        title text,
        abstract text,
        document_type text,
        source_type text,
        publication_date text,
        source_title text,
        publisher text,
        author_keywords text,
        subject_code text,
        subject_name text,
        subject_abbrev text,
        author_given_name text,
        author_surname text,
        author_url text,
        author_affiliation text,
        author_details text,
        funding_details text,
        ref_count int,
        open_access boolean,
        affiliation text,
        language text,
        cited_by int,
        status_state text,
        delivered_date text,
        subject_area text
    )
""")
```

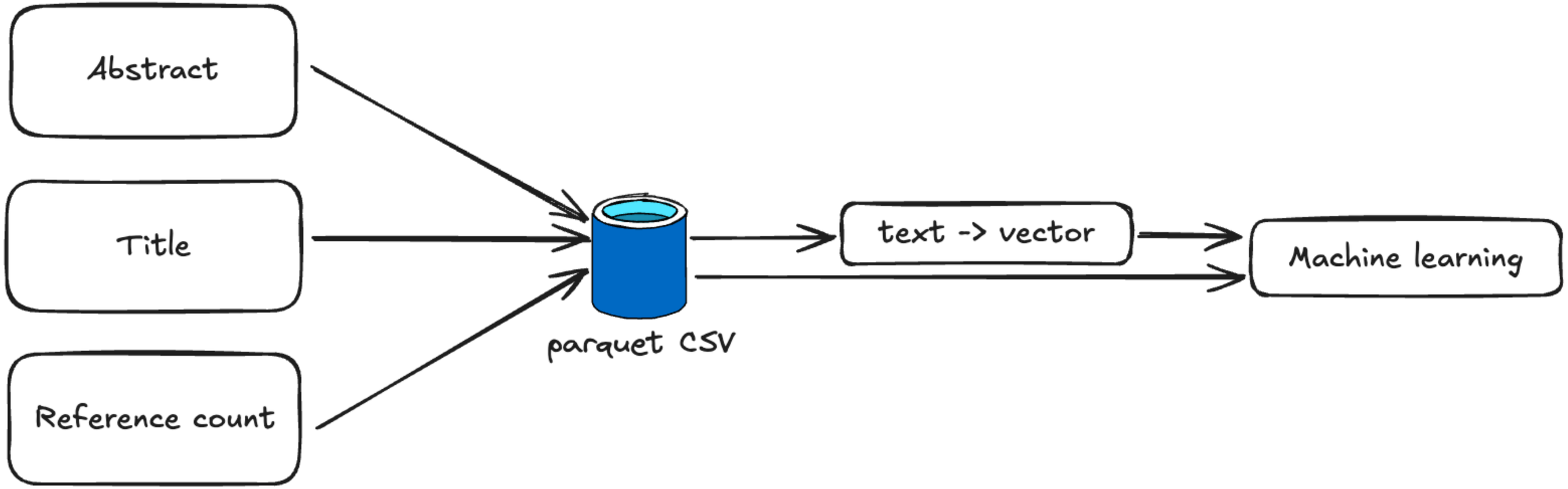
- records 62781



Data Processing

features selections

- reduce features following feature important model
- for more explainable and faster runtime



Data Processing

text feature extraction (bag of words)

- **count how frequency that words appear in datas**
- **normalize by total of that words in documents**

$$\texttt{tf}(t, d)$$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

x

$$\texttt{idf}(t, D)$$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

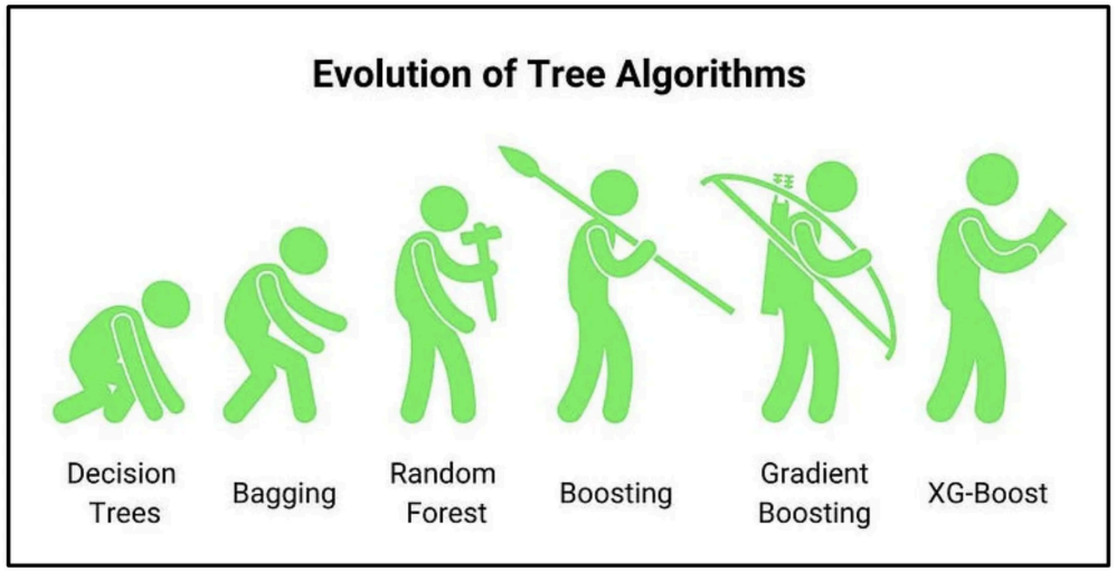
- TF-IDF: Multiply TF and IDF scores, use to rank importance of words within documents
  - Most important word for each document is highlighted

$$\texttt{tfidf}(t, d, D) = \texttt{tf}(t, d) \cdot \texttt{idf}(t, D)$$

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	**0.100**	0.0417	0
4	0	0.0209	**0.100**	**0.100**	**0.100**	0	0.0417	0

Model Selections

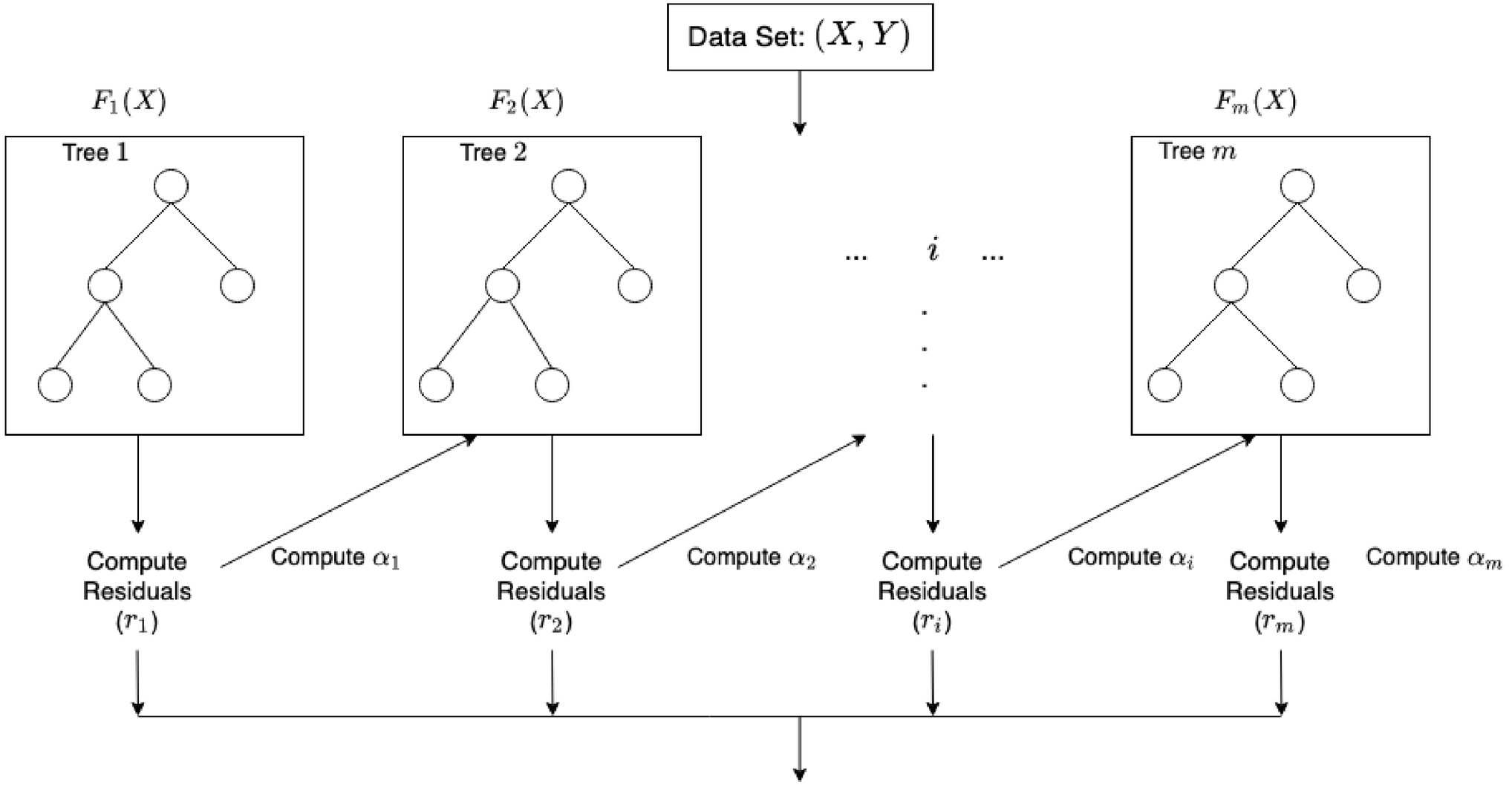
- **classification -> probability**
- **tree based models**
  - **XGBoost**
  - CatBoost
  - LightGBM
- **hyperpameters**



AutoGluon

explainable model

XGBoost

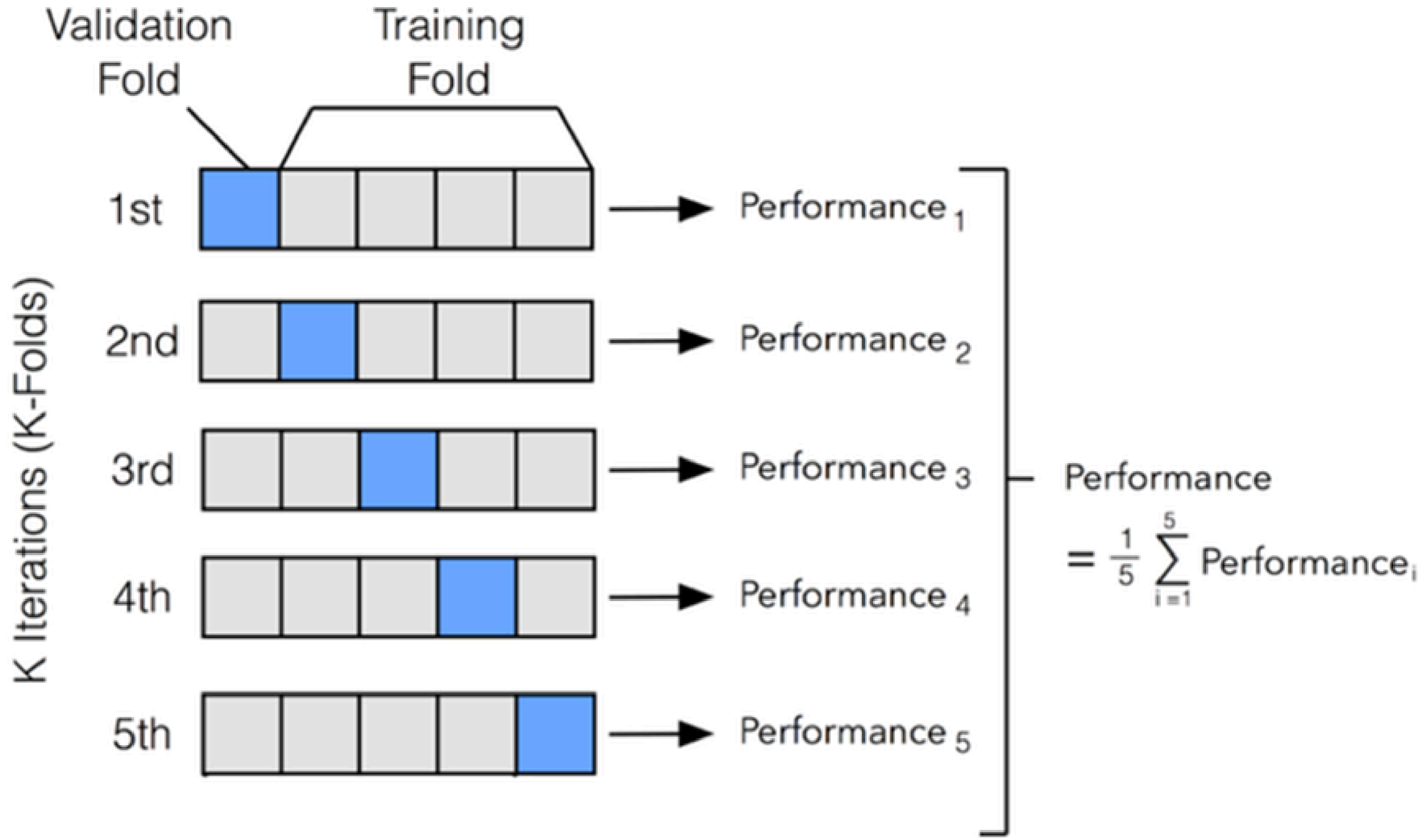


$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$$

where α_i, and r_i are the regularization parameters and residuals computed with the i^{th} tree respectfully, and h_i

training method

- **split datasets to k-fold then train k models**

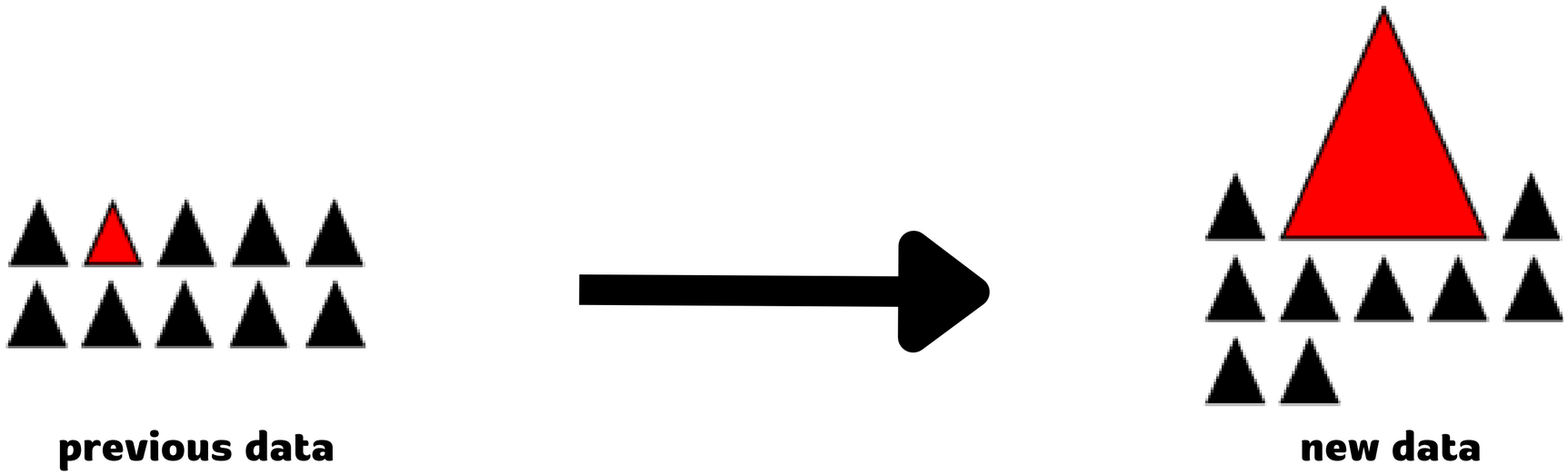


bag of words

not good enough

dealing with imbalance

- **weight loss following distribution of class**
- have true class ~**7 %**



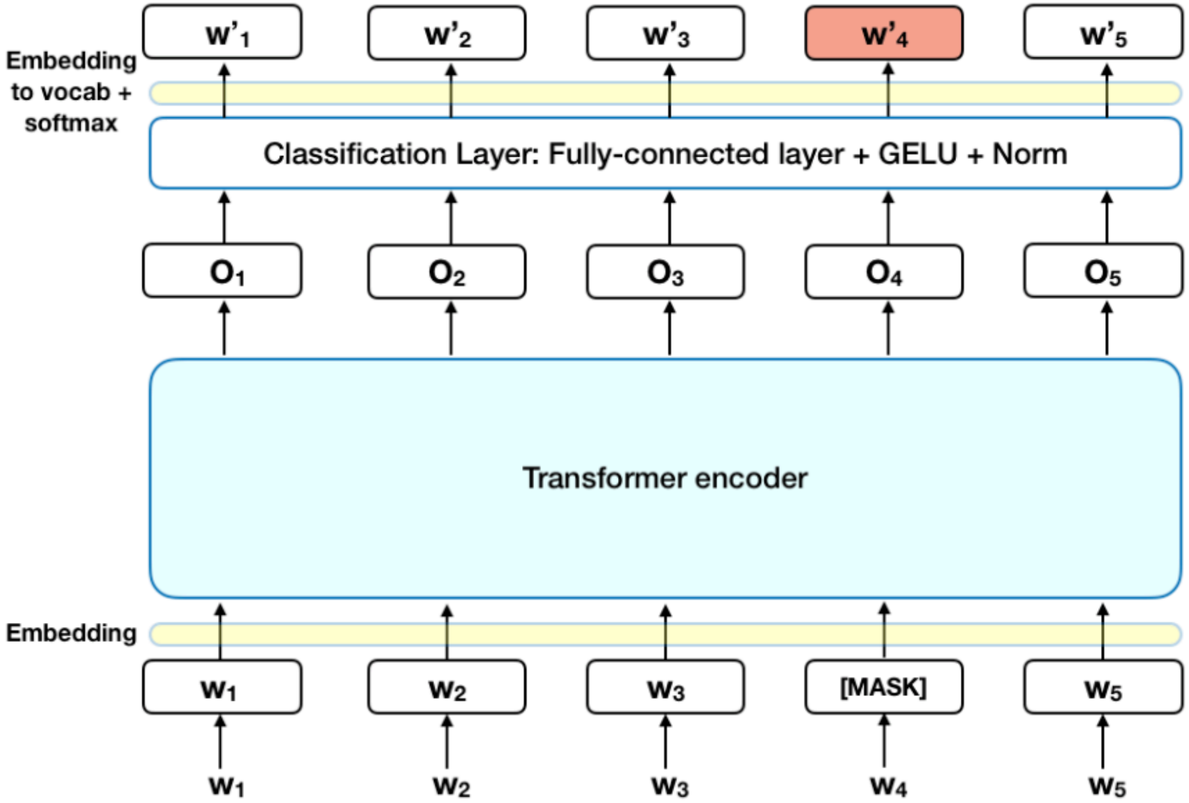
previous data

new data

Data Processing

text feature extraction (Deep learning)

- **use Bert that train with paper data (SciBert)**
- **have special token for papers**
- **512 + 512 tokens**



allenai/scibert_scivocab_uncased

SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy **Kyle Lo** **Arman Cohan**
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

model results

Overall Model Accuracy

94.3%

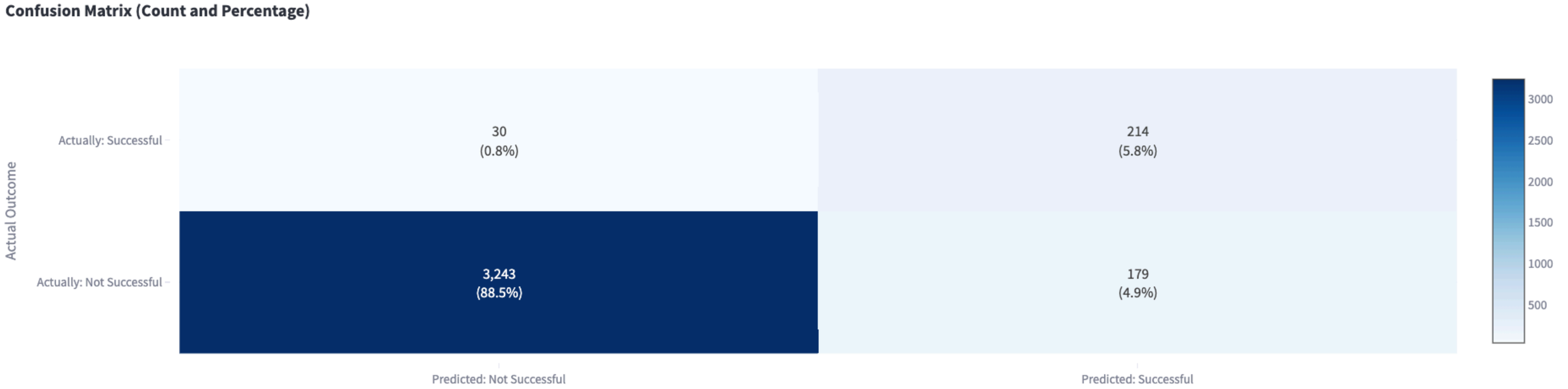
model results

Recall:

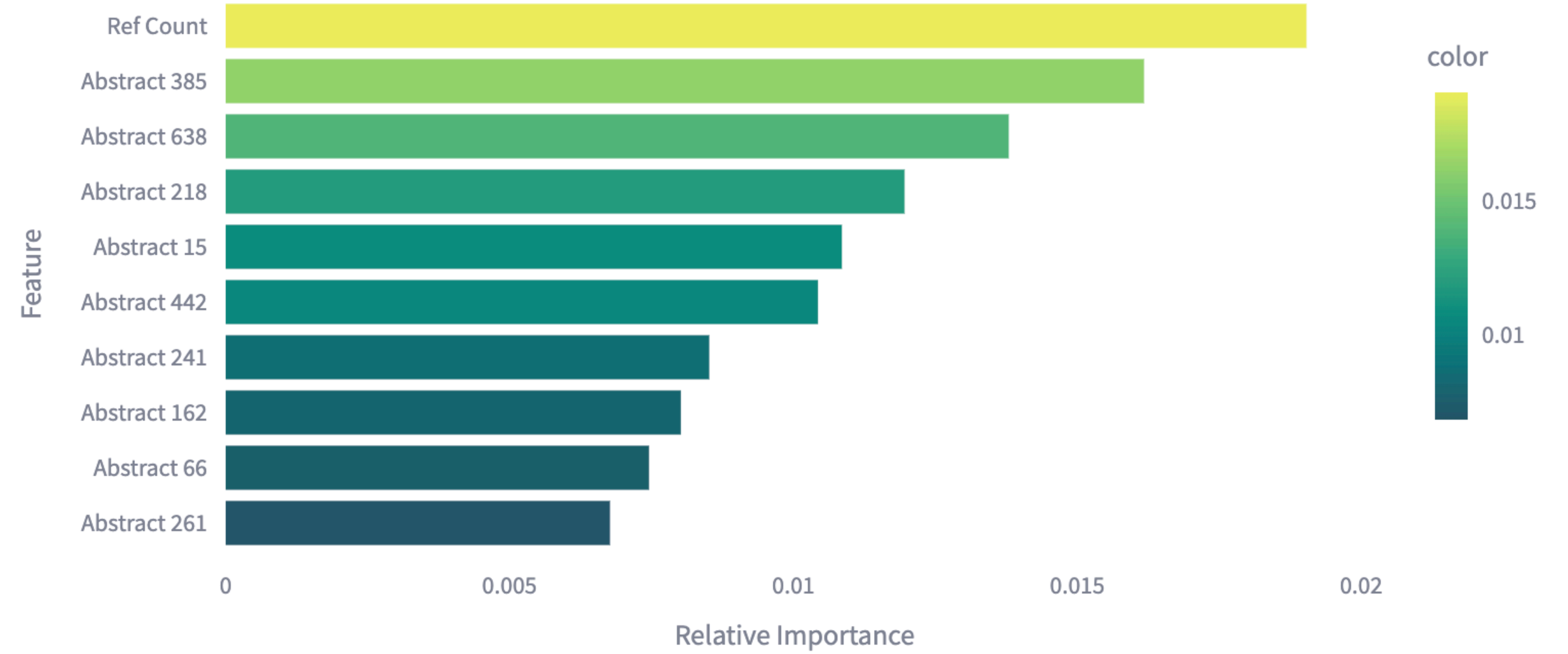
- Correctly identifies **87.7%** of actually successful scholars

Dataset Distribution:

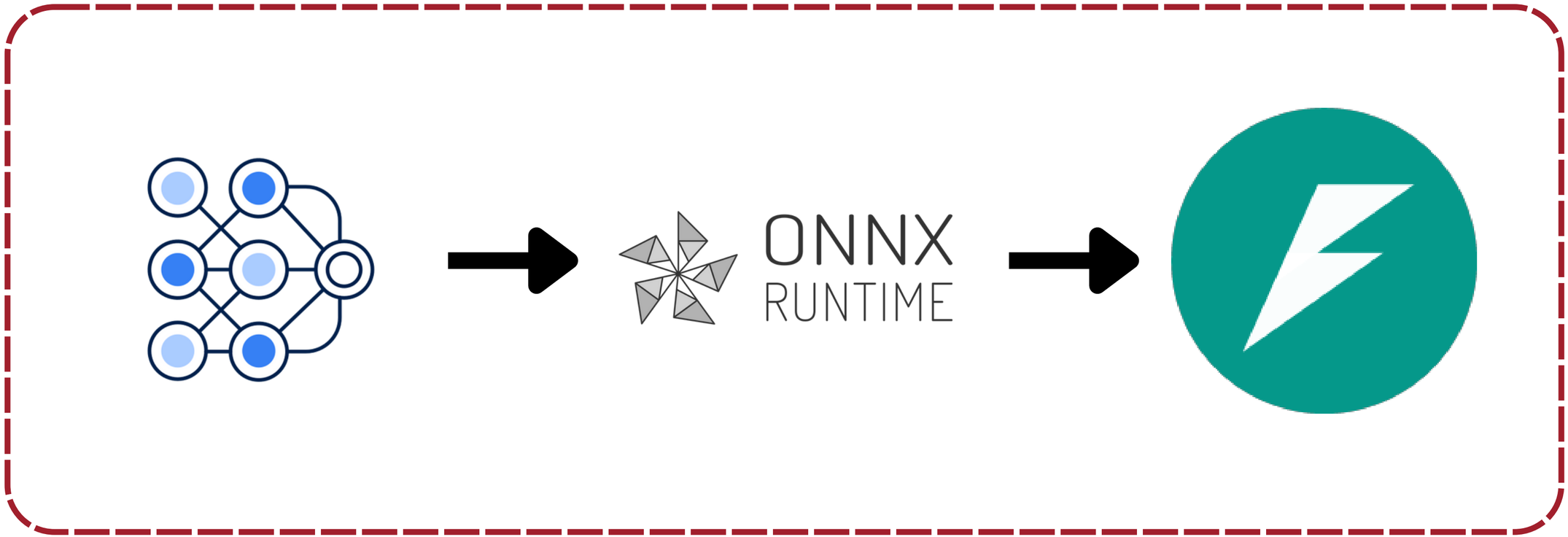
- Total scholars: 3666
- Successful: 244 **(6.7%)**
- Non-successful: 3422



Top 10 Most Important Features for Prediction



deployment



DATA VISUALIZATOIN

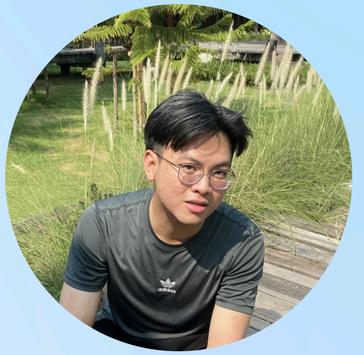
Team Credential

SIRAWIT CHANABURANASAK

Software Developer

SADIT WONGPRAYON

Machine Learning engineer



PAPAWIN TANGCHITPORN

Software developer

SIRASIT TANAJIRAWAT

Project manager

SAKDIPAT SUKHANESKUL

Software developer