MLPR PROJECT

INDIAN JOB RECOMMENDATION SYSTEM

AMOL HARSH
ARYAMAN KHANDELWAL
RISHI VIJAYWARGIYA

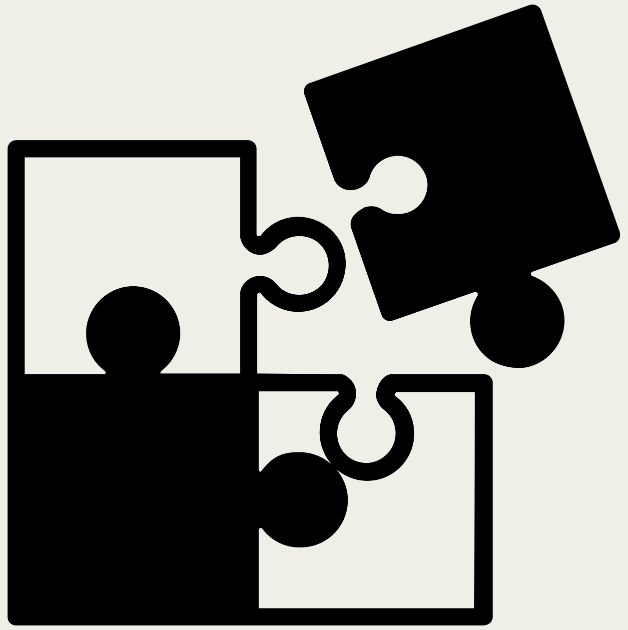
DESIGNED AND MADE IN PLAKSHA

AGENDA

- Problem Statement
- Literature Review
- Dataset and Features Preprocessing
- Future Methodologies

PROBLEM STATEMENT :

Developing a machine learning-based career recommendation system for Plaksha University students to provide personalized, accurate career path suggestions post-graduation.



PROBLEM STATEMENT : STAKEHOLDER CHALLENGES

NO GRANULARITY IN JOB PROFILING

Graduating individuals often struggle to choose a job profile that matches their academic and industrial experiences, making the transition from college to career a challenging decision

3RD YEAR-4TH YEAR-UG STUDENTS-TLP STUDENTS:

Facing uncertainty in making informed career decisions due to the rapidly changing job market trends in India, resulting in difficulty in matching their skillsets with industry demands.

PLAKSHA CAREER DEVELOPMENT CELL:

Despite having access to all student resumes, the process of matching students with relevant job roles remains manual and resource-intensive, leading to uncertainty about which types of companies to bring on campus.

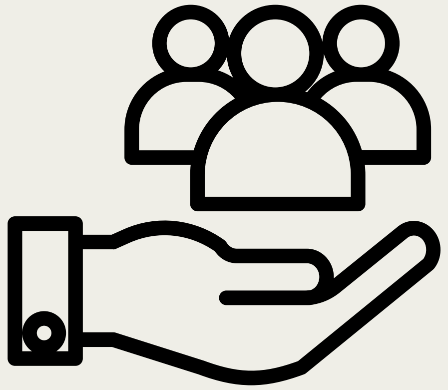
POTENTIAL APPLICATION AND IMPACT

PERSONALISED STUDENT CAREER GUIDANCE:

It can provide personalised career guidance to students at different academic levels (UG, TLF) and across majors, helping them make informed decisions about job roles

RESOURCE OPTIMIZATION FOR PLAKSHA CAREER DEVELOPMENT CELL:

For career development cell, it can automate the process of matching students with job roles, reducing the manual workload and optimizing resource allocation.



LITERATURE SURVEY

LITERATURE SURVEY-

Survey on Job Recommendation Systems using Machine Learning

Raj Thali
Department of Information Technology
Pillai College of Engineering
New Panvel,India
thaliraj1@gmail.com

Suyog Mayekar
Department of Information Technology
Pillai College of Engineering
New Panvel,India
suyog.mayekar12@gmail.com

Shubham More
Department of Information Technology
Pillai College of Engineering
New Panvel,India
ssmore2312@gmail.com

Sanjana Barhate
Department of Information Technology
Pillai College of Engineering
New Panvel,India
sanju1234.barhate@gmail.com

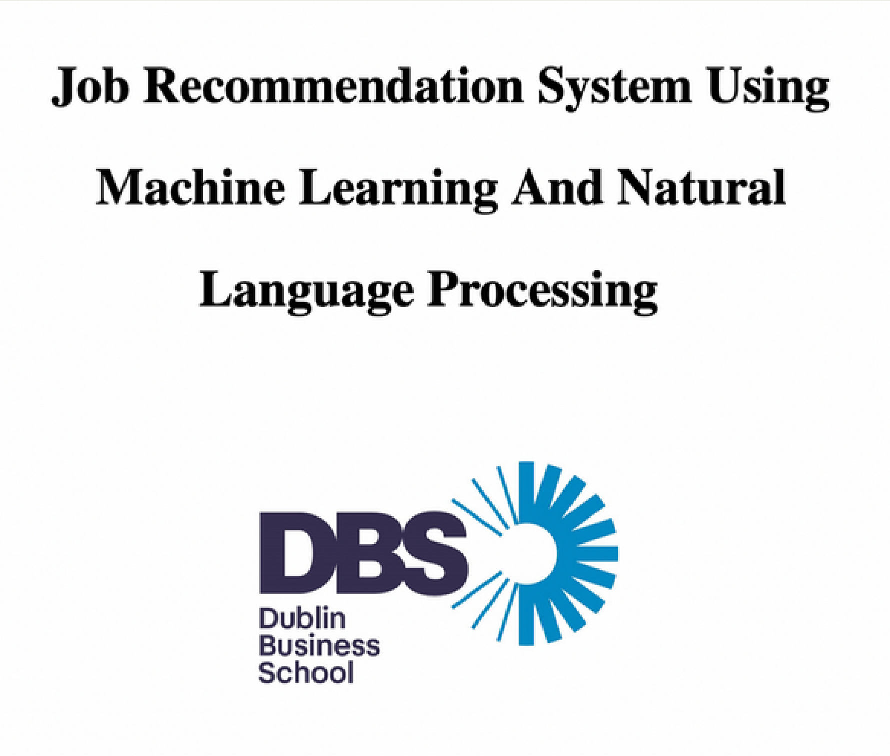
Sangeetha Selvan
Department of Computer Engineering
Pillai College of Engineering
New Panvel,India
sangeethas@mes.ac.in

Collaborative filtering, commonly used in recommendation systems, focuses on using the preferences and behaviors of similar users to suggest jobs. It emphasizes the 'community' aspect of recommendations.

A modern approach in job recommender systems is not just to match jobs but to recommend skills that users might need to learn to improve their employability.

https://ieeexplore.ieee.org/document/10100122

LITERATURE SURVEY- COMMON COLLECTION PROCEDURE


Job Recommendation System Using Machine Learning And Natural Language Processing

DBS
Dublin Business School

Open Source

Stack Overflow
Job Listing Surveys

Web Scraping

Linkedin/
Glassdoor

https://esource.dbs.ie/bitstream/handle/10788/4254/msc_jeevankrishna_2020.pdf?sequence=1&isAllowed=y

LITERATURE SURVEY-

CITES:
RESEARCH PAPER: RESEARCH GATE
'PERSONALISED RETRIEVAL FOR ONLINE RECRUITMENT SERVICES'

Once communities of related users are constructed the recommendation process can then proceed in a way that is analogous to the memory-based approach, except that instead of selecting k neighbours for the target profile, we select the members of the target profile's community. Of course, the immediate benefit of this cluster-based approach is that it is possible to identify larger groups of users that are related to the target user and thus provide a richer recommendation base.

$$Quality(j,P) = \frac{|\{p \in P : p \text{ contains } j\}|}{|P|}$$

Definition 3: Quality {where j is a job and P is a community of profiles}

evaluation of two versions of the Adaptive Collaborative Filtering (ACF) algorithm for personalised job recommendations. The evaluation was carried out manually by selecting ten target users from different virtual communities and producing two recommendation lists containing ten jobs each.

Memory-based collaborative filtering is probably the simplest form of the general collaborative filtering approach. Users are related on the basis of a direct similarity between their profiles, for example, by measuring the degree of overlap between their profile items, or by measuring the correlation coefficient between their grading lists [2, 12, 13]. This leads to a lazy (in the machine learning sense) form of collaborative filtering whereby the target user is used to select the k nearest profiles. Currently CASPER uses a simple overlap metric (Definition. 1) to determine profile similarity.

$$Overlap(t,p) = \frac{|Items(t) \cap Items(p)|}{|Items(t) \cup Items(p)|} \qquad Quality(j,t,P) = \sum_{\forall pi : j \in pi} Overlap(t,p_i)$$

Definition 1: Overlap {where: t and p are profiles (t being the target profile) and j is a job}

Definition 2: Quality {where: t and p are profiles (t being the target profile) and j is a job}

The two ACF versions evaluated were Memory (ACF-NN) and Cluster (ACF-Cluster). The grading of the recommendations was based on how similar the recommended jobs were to the existing jobs in each target user profile.

Each target user received a cumulative grading score across the 10 recommended jobs from each ACF technique, and each grading score was normalized by dividing by the maximum cumulative grade of 30.

The experimental study is based on the user profiles generated from server logs between 2/6/98 and 22/9/98. These logs contained a total of 233,011 job accesses from 5132 different users. These profiles spanned a total of 8248 unique jobs with an average profile size of approximately 14 jobs and nearly 3000 profiles containing less than 10 jobs – and indication of CASPER's extremely sparse profile space.

LITERATURE SURVEY-

D. *Similarity Method Dealing with Text*

In student job hunting system, student resume information and job descriptions are stored in the form of text in the database. To compare the similarity between two pieces of information, we represent each piece of information as space vector and use cosine similarity distance calculation.

For example, job description is expressed as a vector like this: (job name, location, job type, field, category name). It is represent by $\vec{J}=(j_1,j_2,j_3,j_4,j_5)$; student resume is expressed as a vector like this: (college, major, degree, home place, gender). It is represent by $\vec{S}=(s_1,s_2,s_3,s_4,s_5,s_6)$.

The similarity between two jobs or two students can be calculated by the formula (10) and (11):

$$sim(J_1,J_2)=\cos(\theta_j) \tag{10}$$

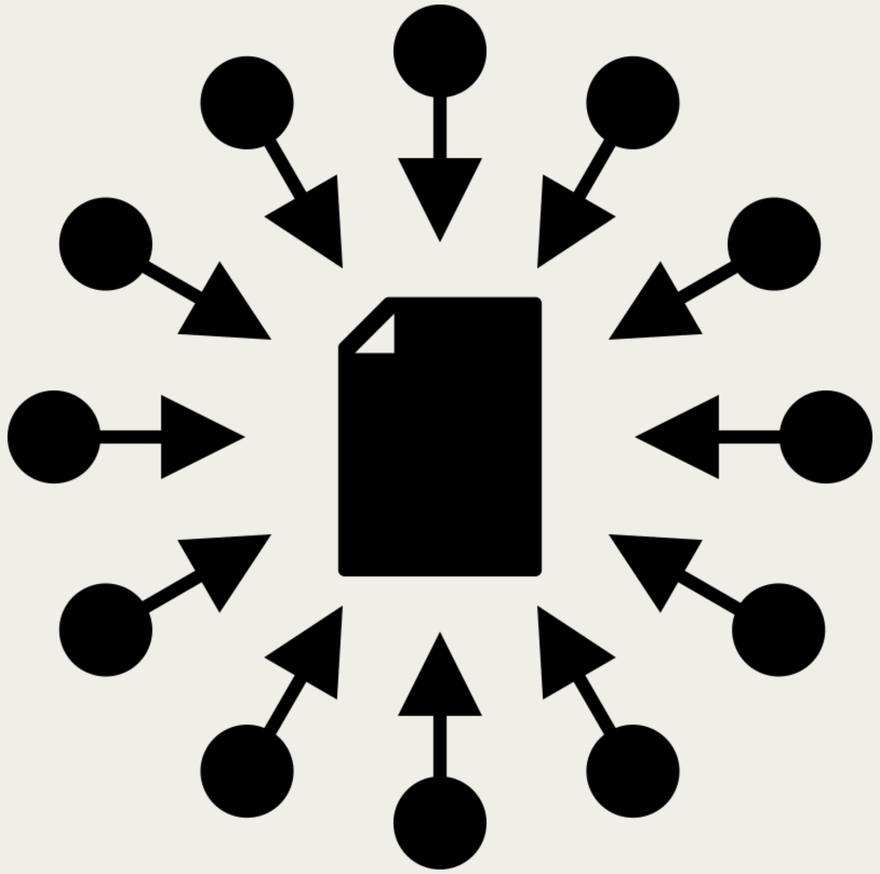
$$sim(S_1,S_2)=\cos(\theta_u) \tag{11}$$

Job descriptions and student resumes are converted into vector format. Each attribute of a job description or a resume is represented as a component in its respective vector.

Cosine similarity is used to calculate the similarity between two vectors. It measures the cosine of the angle(θ) between two vectors in a multidimensional space.

By representing the resumes and job descriptions as vectors, the system can compute how closely a student's qualifications (resume vector) match the requirements of a job (job vector).

Dataset and Features Preprocessing



Data Collection → Data Cleaning → Data Wrangling → ML Methodology → ML Deployment

DATA COLLECTION

Source

- Data was collected using **Selenium** and **Beautiful Soup.**
- Data was also collected **manually** by us, due to the change in rendering structure of LinkedIn on web.
- **Data Augmentation** was also done which will be explained in the future slide.

About Data

- 1200 data points collected*
- 800 real data points
  - 500 using Selenium
  - 300 manually
- 400 synthetic data points
- 100 features in the data

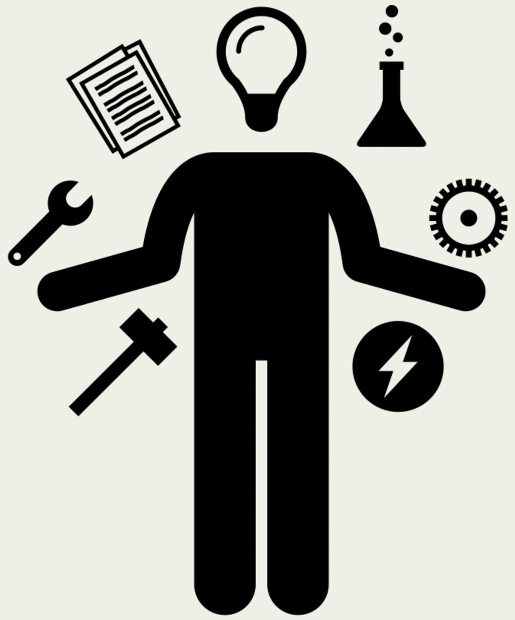
Ethical Concerns

Ethical concerns were addressed by ensuring **Anonymity.**

Data Collection → Data Cleaning → Data Wrangling → ML Methodology → ML Deployment

DATA COLLECTION

Experiences



Skills + Projects + Certifications

Education

Data Collection → Data Cleaning → Data Wrangling → ML Methodology → ML Deployment

DATA COMPOSITION

About Section

- Professional Summary:
- Career Objectives or Goals:
- Key Achievements and Skills:
- Relevant Keywords:

Crucial for understanding the individual's involvement in the industry and classify him based on characteristics in different and in correlation with different samples

Qualification

- Level of qualification
- Degree Procured

Qualification helps segregating the colleges as an early filter, classifying the possible colleges based on alikeness

Certificates

- Name of professional certificates from online /offline platforms.
- Keywords based on frameworks

Required for segmenting the individuals based on professional capacity and the level of expertise and proficiency.

DATA AUGMENTATION - SYNTHETIC VALUES

BACK TRANSLATING

The simultaneous conversion between english to specific language and back to engage

LANGCHAIN GENERATION

Generating artificial text based on contextual information from established dataset consisting of information.

SYNONYMS GENERATION

Replacing the words in the data with their synonyms adding variability to the data.

Data Collection → Data Cleaning → Data Wrangling → ML Methodology → ML Deployment

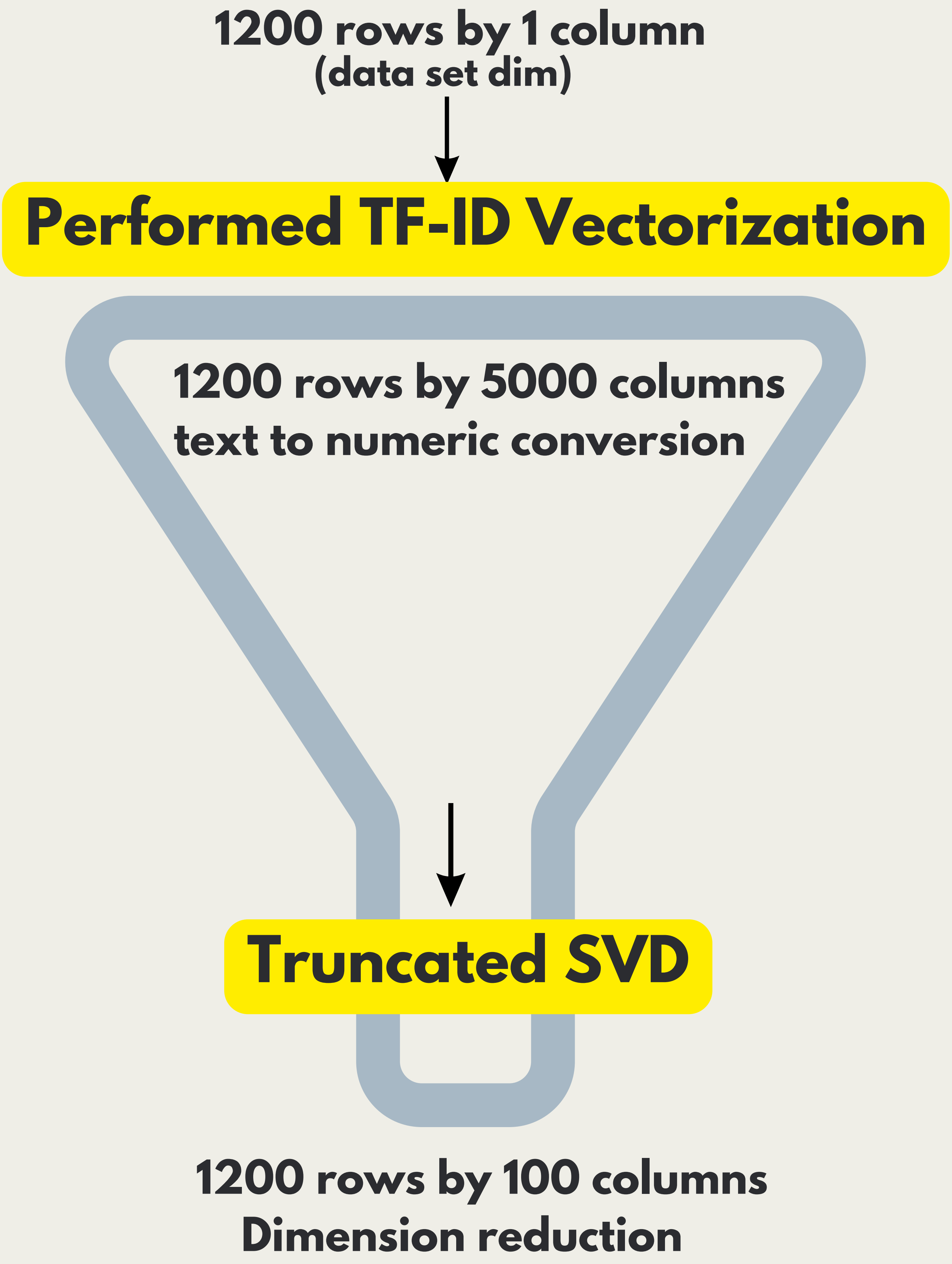
DATA COMPOSITION - ASSEMBLY

	A
1	combined_text
2	
3	I am a Senior Data Analyst within McKinsey's Growth, Marketing & Sales Practice, specializing in (RGM) Revenue Growth Management Solution. My role involves harnessing data, advanced analytics, and technology to guide clients in making informed decisions and
4	I have 2 years of experience as a Data Analyst, excelling in both independent and teamwork environments. My expertise includes SQL, ETL Tool, and Data Visualization, Data Mining using Python packages like Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn for var
5	nan Data AnalystData AnalystDTDC - India · Full-timeDTDC - India · Full-timeMar 2022 - Present · 1 yr 9 mosMar 2022 - Present · 1 yr 9 mos-Managing data warehousing,reporting and. Data analysis,requirement gathering.-Using sql and big query for data analysis. -G
6	#I hold a Bachelor of Technology in Information Technology and a Master of Technology in Distributed & Mobile Computing. My passion lies in data analysis, and I possess an analytical mindset for solving real-world problems. Currently, I am an experienced data a
7	I am a Masters in Data Analytics graduate from National College of Ireland with Business Intelligence and Data Analytics expertise. I help convert raw unstructured data into meaningful insights and patterns that directly translate into business growth and developm
8	I am a data analyst who loves automating the processes. • Have Working experience with Data Analytics, Outreach, Marketing, and Management. • Have worked on individual projects as well as ,with Teams, and as a Team Leader. • Aim to work on projects that make
9	IT Professional with in-depth knowledge in the working of computers and its technologies with a client and customer oriented attitude looking to join a challenging position where I can add value to the bottom line of the Company. Senior Data AnalystSenior Data
10	I am a post graduate of Enterprise Business Analytics from National University of Singapore. Currently, I am working in Election Commission of India as a senior data analyst. My work involves making analytical and statistical reports for the Commission and analysin
11	As a data analyst, I specialize in using data to drive business decisions and improve performance. My technical skills include proficiency in SQL, Excel, Python and data visualization tools such as Power BI. I have a strong understanding of statistical analysis and Explo
12	nan Mercedes-Benz Research and Development IndiaMercedes-Benz Research and Development India1 yr 11 mos1 yr 11 mosBengaluru, Karnataka, IndiaBengaluru, Karnataka, IndiaData AnalystData AnalystFull-timeFull-timeApr 2023 - Present · 8 mosApr 2023 · Pr
13	I am working as a Lead Data Analyst at Imarc Services Private Limited and responsible for carrying out various analytical operations contributing to fulfil the business requirement. I have strong Analytical and Documentation skills which in turn contributes to help
14	Experienced Data Analyst, 4+ experience of experience in Business and Analytics. Hands on experience on Python, R, Machine Learning,Tableau, and Advance Excel. Worked with Data-driven business solution ,coupling theoretical data science techniques with real-
15	A management student turned Data Analyst. Always open to learning new technology or an emerging existing one! My current interests outside of work lie in exploring and learning about blockchain and cryptocurrencies (like everyone else's!) Data AnalystData Ana
16	I pursued my B Tech in Computer Science from National Institute of Information Technology (NIIT University). I am a Data Science enthusiast and continuous learner. I have a keen interest in the field of Machine Learning. I give high productivity while working unde
17	Database: MS SQL Server, HiveProgramming Languages: Python, RBI Tools: Tableau, Power BILibraries: Pandas, Scikit, Seaborn, MatplotLibAlgorithms : Random Forest, XGBoost, Clustering and other fundamental models Data AnalystData AnalystAmerican Express
18	Previously, as a data analyst at Google, I created and maintained complex reporting dashboards, identified and resolved data discrepancies, and provided real-time insights into region-wise abuse alerts. I worked with the Business Strategy and Operations team with
19	Currently working as a Senior Data Analyst for Automation Coding Process in BuddiHealth(Formerly known as Claritrics India) ChennaiDevelopment and analysis of Computer-Assisted Coding process for (CPT and ICD10) Radiology & Surgery coding.CODINGKnowle
20	Data analyst with a curious mind and a passion for uncovering insights hidden within vast amounts of data. With 7 years of experience working in the field, I've honed my skills in Power BI, SQL, MS Excel, Power Query, ETL, and love putting them to use in solving co
21	Self-driven data analyst with a passion to create business impact, guide data into business insights.Result-oriented individual with strong analytical thinking and ability to clearly communicate, seekingproduct and business analyst opportunities. Data AnalystData
22	Welcome to my LinkedIn profile! I am a Data Analyst with expertise in Core Banking Operations and a focus on delivering data-driven solutions. I am currently working with Tata Consultancy Services (TCS) as a vendor for State Bank of India.In my role, I am immerse
23	I'm a Senior Associate Engineer at Caterpillar on a data analytics team focused on Drive train controls validation and machinery health. As an analyst, my primary role is to provide actionable insights for the data (typically high frequency time series data) provided a
24	Data Analyst with of experience of 5 years in data field . Currently Working In FMCG industry as Data Analyst. Sharing insight from raw data after transform into insightful and meaningful data. Utilizing these insights by business to take decision for sales growth. I giv
25	Experienced Data Analyst with a demonstrated history of working in the marketing and advertising industry. Skilled in Market Research, Microsoft Excel, Data Analysis, Data Visualization and Tableau. Digitas IndiaDigitas India4 yrs 11 mos4 yrs 11 mosSenior Associ
26	[illegible]

DATA PREPROCESSING- FREQUENCY MEASUREMENT

- **Data lemmitization/Stemming - reducing uncontextual words.**
- **Keyword detection related to a parameter(job role)**
- **Removing the words with frequency(f) < threshold and combining similar contextual words in a base word**
- **Tools used**
  - re (Regular Expressions)
  - string
  - nltk (Natural Language Toolkit)

Hyper-parameter Tuning



```
0        hi guy ai research scientist blended experienc...
1        hi guy ai research scientist blended experienc...
2        ai applied research scientist ai product manag...
3        research scientist specialize field artificial...
4        machine learning engineer demonstrated history...
                                 ...
```

After TF-ID Vectorization

	2019	ab	aba	abap	abaqus	abb	abdm	ability	abin	able	...	zenly	zeppelin	zero	zest	zonal	zookeeper	zscaler	zw	zx
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.08429	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.08429	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

```
In [114]:  1 tfidf_df.columns.shape
Out[114]: (5000,)
```

After Dimension Reduction

	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93
0	0.333657	-0.070946	0.085722	0.003010	-0.030875	-0.045888	0.002208	0.022203	-0.016190	0.096740	...	-0.017455	0.027243	-0.012284	-0.005156
1	0.333657	-0.070946	0.085722	0.003010	-0.030875	-0.045888	0.002208	0.022203	-0.016190	0.096740	...	-0.017455	0.027243	-0.012284	-0.005156
2	0.167763	-0.019305	-0.044584	-0.022066	0.297777	-0.093103	0.114402	-0.043691	0.012949	-0.110555	...	-0.040186	0.042883	-0.014071	0.013443
3	0.269379	-0.028949	0.021850	-0.072194	0.224431	-0.084479	0.102876	-0.045232	0.007290	0.022560	...	-0.019210	0.046007	0.007454	0.045438
4	0.326130	-0.002581	0.015993	-0.046596	-0.021838	-0.066116	-0.046237	-0.108558	0.249998	0.043570	...	0.019542	-0.005403	-0.005002	-0.094342
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

MACHINE LEARNING MODEL

EMSEMBLE METHODS

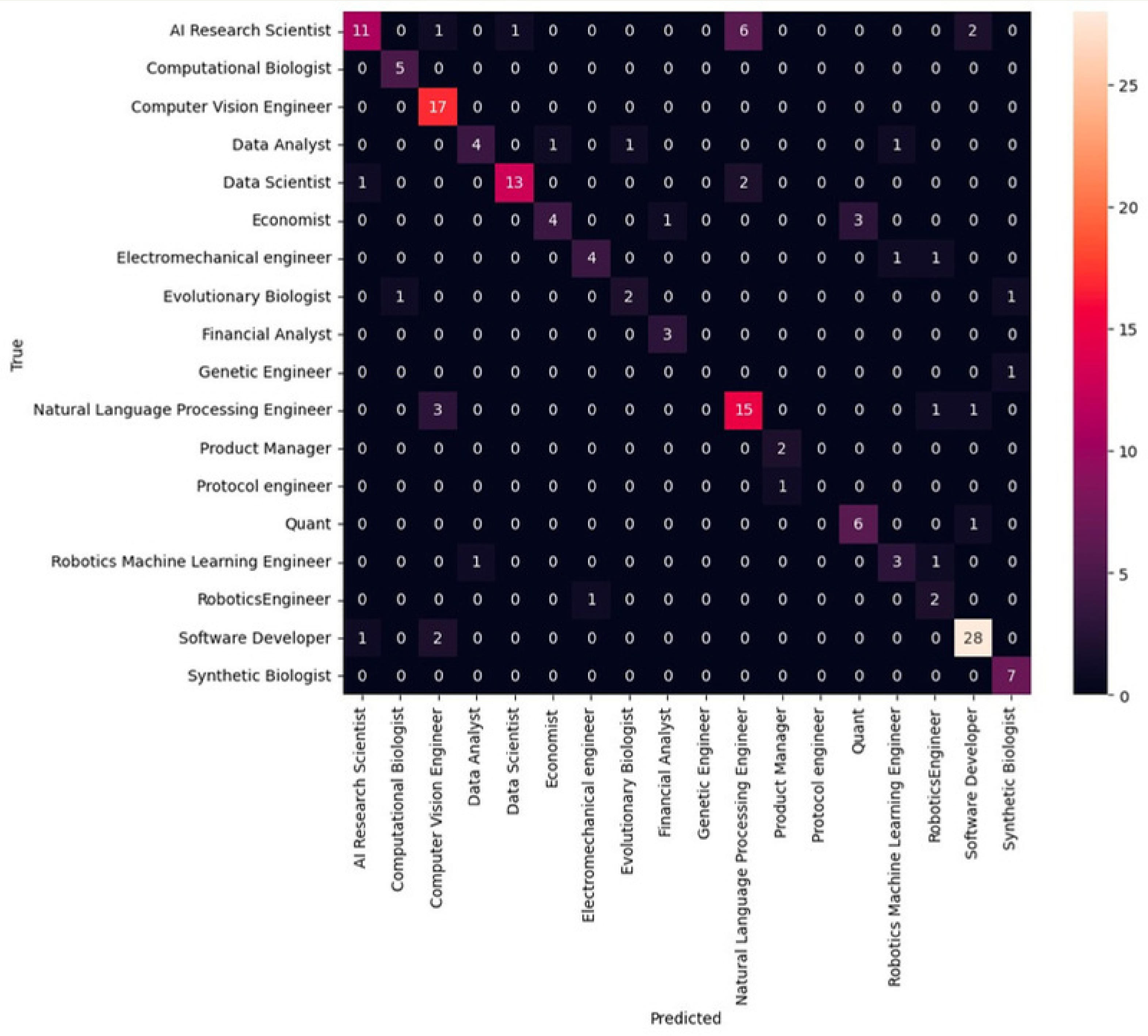
CLUSTERING ALGORITHM

Data Collection → Data Cleaning → Data Wrangling → **ML Methodology** → ML Deployment

ML MODEL - ENSEMBLE METHOD(XGBOOST)

	precision	recall	f1-score	support
AI Research Scientist	0.87	0.62	0.72	21
Computational Biologist	1.00	0.80	0.89	5
Computer Vision Engineer	0.84	0.94	0.89	17
Data Analyst	0.83	0.71	0.77	7
Data Scientist	0.76	0.81	0.79	16
Economist	0.88	0.88	0.88	8
Electromechanical engineer	0.80	0.67	0.73	6
Evolutionary Biologist	1.00	0.75	0.86	4
Financial Analyst	0.75	1.00	0.86	3
Genetic Engineer	1.00	1.00	1.00	1
Natural Language Processing Engineer	0.67	0.70	0.68	20
Product Manager	0.67	1.00	0.80	2
Protocol engineer	0.00	0.00	0.00	1
Quant	0.86	0.86	0.86	7
Robotics Machine Learning Engineer	0.43	0.60	0.50	5
RoboticsEngineer	0.75	1.00	0.86	3
Software Developer	1.00	0.94	0.97	31
Synthetic Biologist	0.78	1.00	0.88	7
accuracy			0.81	164
macro avg	0.77	0.79	0.77	164
weighted avg	0.83	0.81	0.81	164

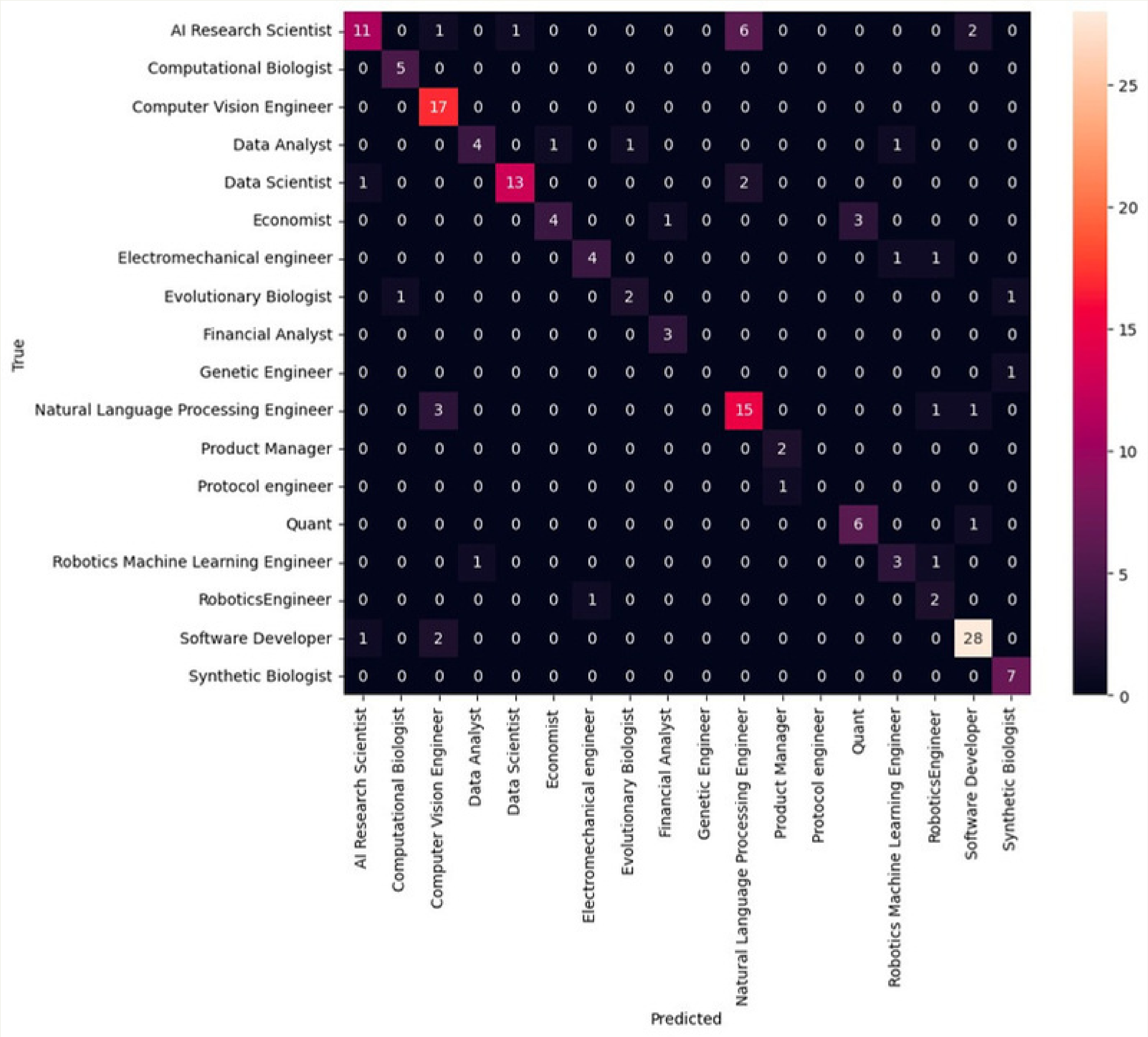
Accuracy: 81.10%



ML MODEL - EMSEMBLE METHOD(XGBOOST)

Why did we not use it?

- Not optimal Accuracy.
- Centric & Bias Classification.
- Does not cover all nuances of a profile.




```
Electromechanical engineer: 99.20%
RoboticsEngineer: 0.23%
Robotics Machine Learning Engineer: 0.18%
```

We need something that suggests and not dictates!

CUSTOM MULTI-CLASS CENTROID CLASSIFIER

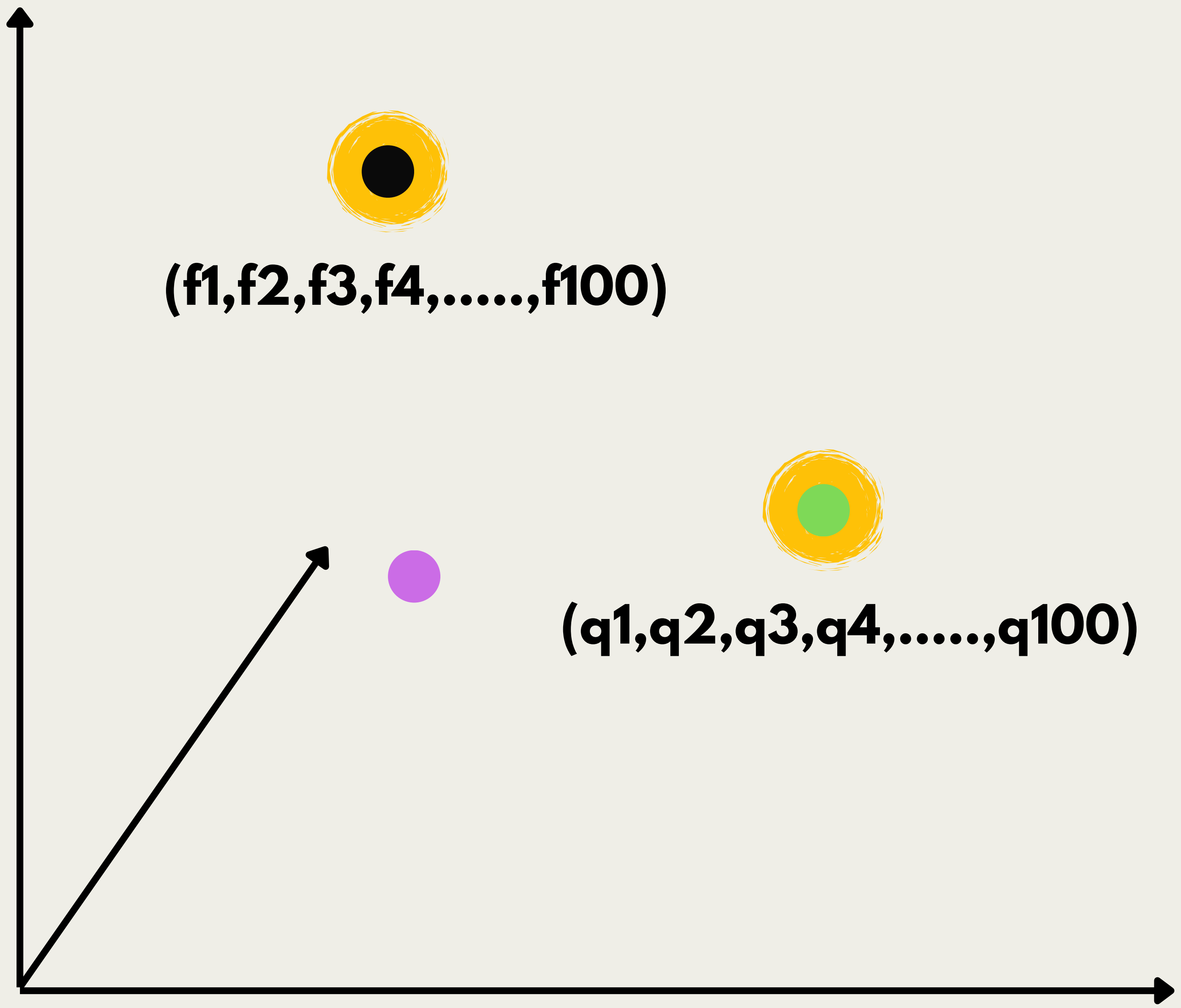
Tailored Career Guidance and Efficient Multi-Class Categorization: Utilizes cosine similarity for personalized recommendations and incorporates clustering for effective classification.

f1 f2--f100

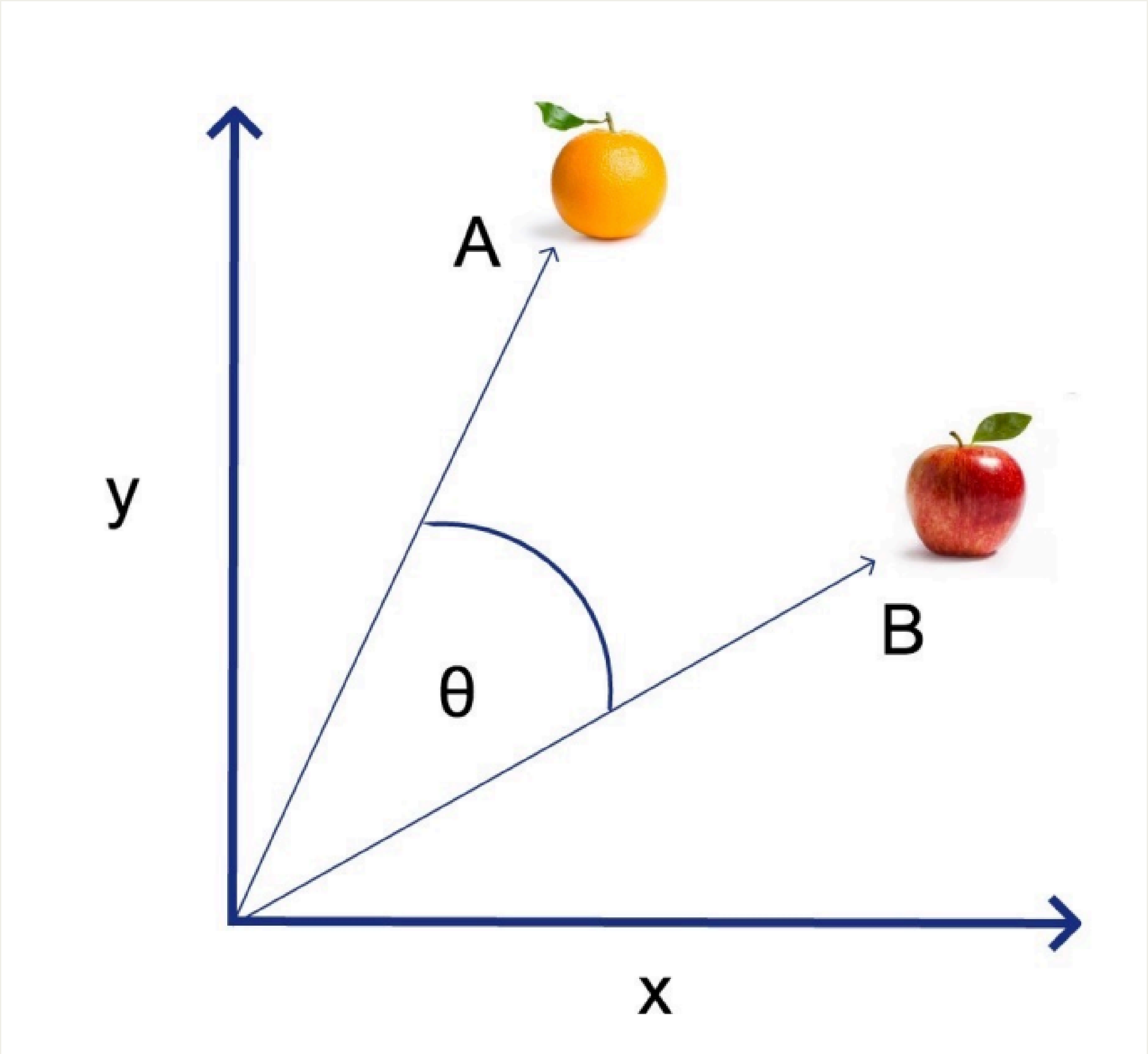
	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93
0	0.333657	-0.070946	0.085722	0.003010	-0.030875	-0.045888	0.002208	0.022203	-0.016190	0.096740	...	-0.017455	0.027243	-0.012284	-0.005156
1	0.333657	-0.070946	0.085722	0.003010	-0.030875	-0.045888	0.002208	0.022203	-0.016190	0.096740	...	-0.017455	0.027243	-0.012284	-0.005156
2	0.167763	-0.019305	-0.044584	-0.022066	0.297777	-0.093103	0.114402	-0.043691	0.012949	-0.110555	...	-0.040186	0.042883	-0.014071	0.013443
3	0.269379	-0.028949	0.021850	-0.072194	0.224431	-0.084479	0.102876	-0.045232	0.007290	0.022560	...	-0.019210	0.046007	0.007454	0.045438
4	0.326130	-0.002581	0.015993	-0.046596	-0.021838	-0.066116	-0.046237	-0.108558	0.249998	0.043570	...	0.019542	-0.005403	-0.005002	-0.094342
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Job profiles: Data Scientist

HOW DOES IT WORK?



Centroid of each job profile

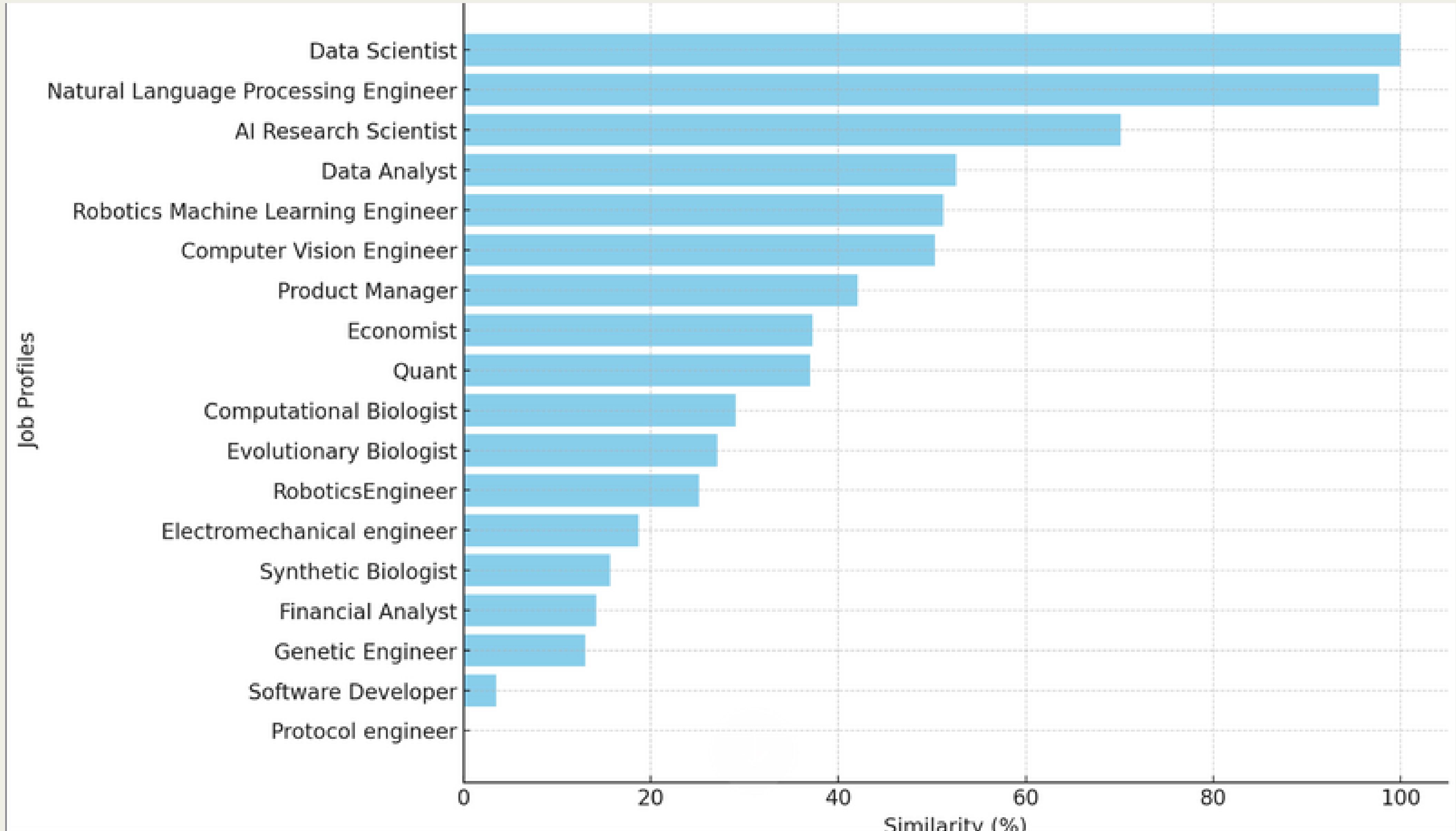


Cosine Similarity to calculate the similarity between two items

```
Raw Cosine Distances:  {0: 0.4853091955097788, 1: 0.83979722331453, 2: 0.44636496589278274, 3: 0.8740463259779488,
4: 0.743878150390753, 5: 0.9177837351738369, 6: 0.8025860259966566, 7: 0.8943448173065744, 8: 0.9654885670577621,
9: 0.9327317519512219, 10: 0.6467594122252713, 11: 0.6331316847689132, 12: 0.8470847119434224, 13: 0.91090047926235
81, 14: 0.6658305984258175, 15: 0.8038857619251599, 16: 0.9135085496740252, 17: 0.908351094010676}
```

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

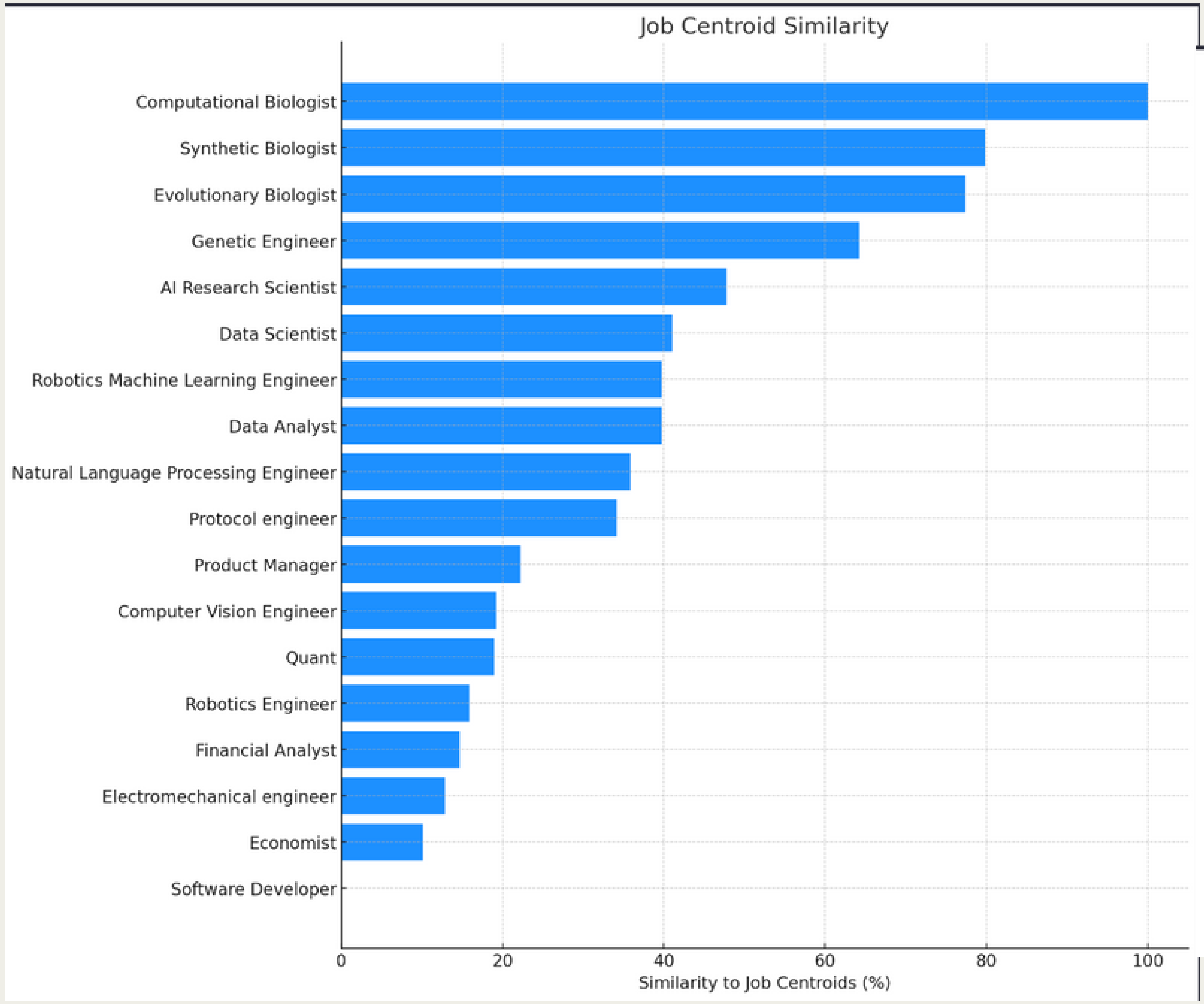
```
True Label: Data Scientist
Similarity Label: Data Scientist, Similarity: 100.00%
Similarity Label: Natural Language Processing Engineer, Similarity: 97.67%
Similarity Label: AI Research Scientist, Similarity: 70.13%
Similarity Label: Data Analyst, Similarity: 52.58%
```

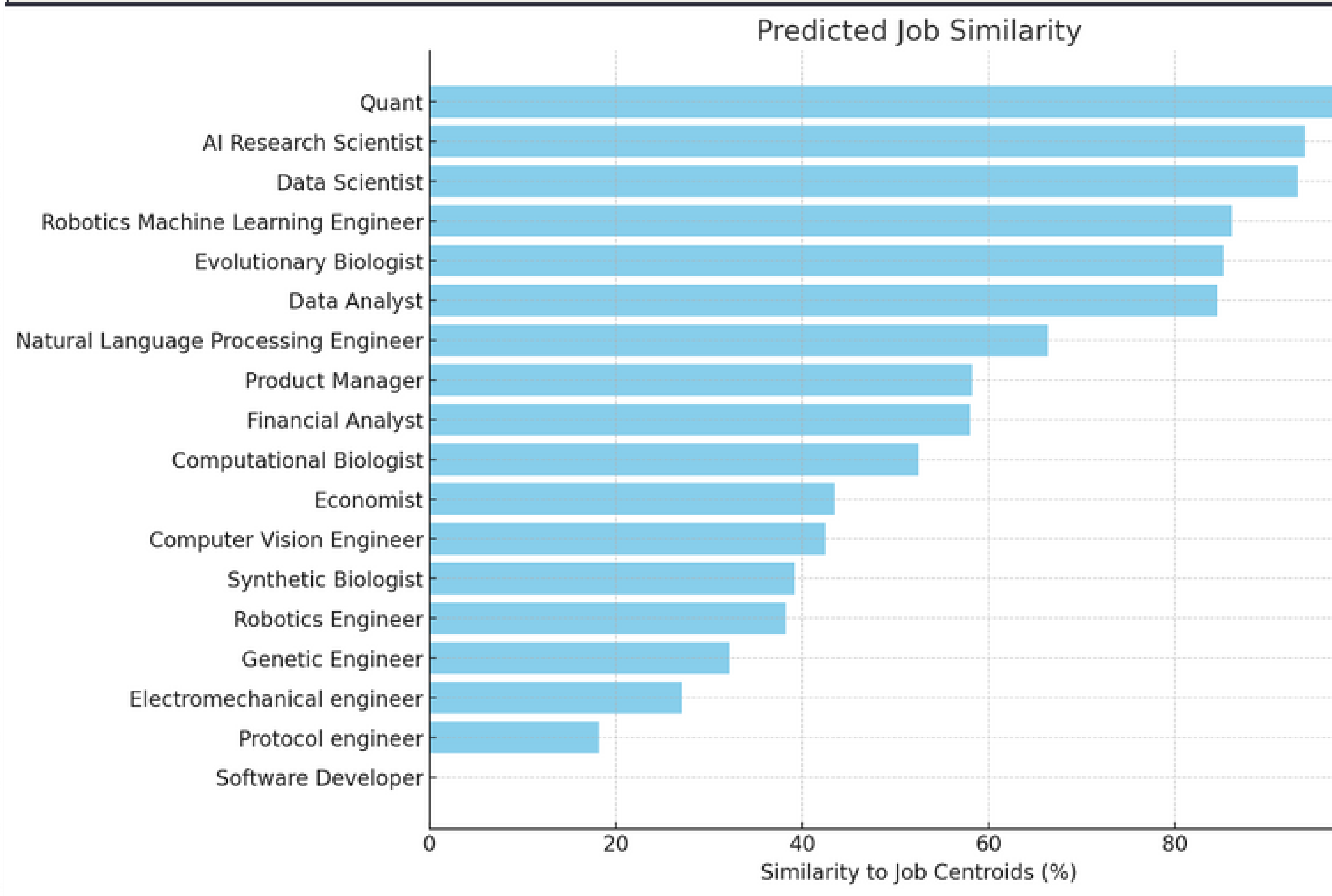


Model Accuracy:

```
Top-3 Accuracy: 95.1219512195122 %
```

RESULTS ON REAL WORLD RESUMES!





BSE

DSEB

Data Collection → Data Cleaning → Data Wrangling → ML Methodology → ML Deployment

MODEL - METHODOLOGIES(FILTERING)

USER PROFILE VECTOR:

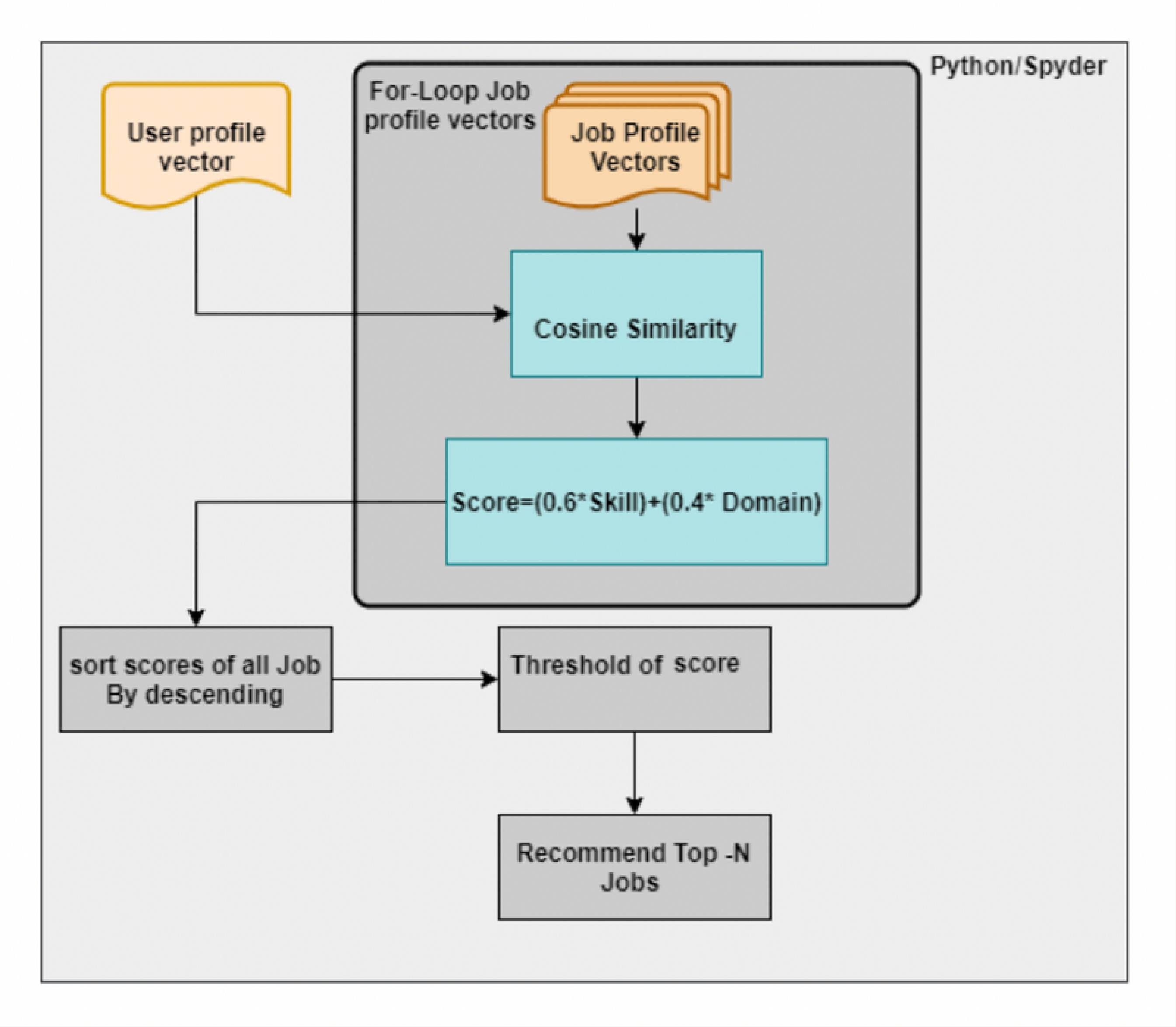
DOMAIN KNOWLEDGE, AND OTHER ATTRIBUTES OF A USER. IT'S ESSENTIALLY A NUMERICAL REPRESENTATION OF A USER'S QUALIFICATIONS AND PREFERENCES.

JOB PROFILE VECTORS:

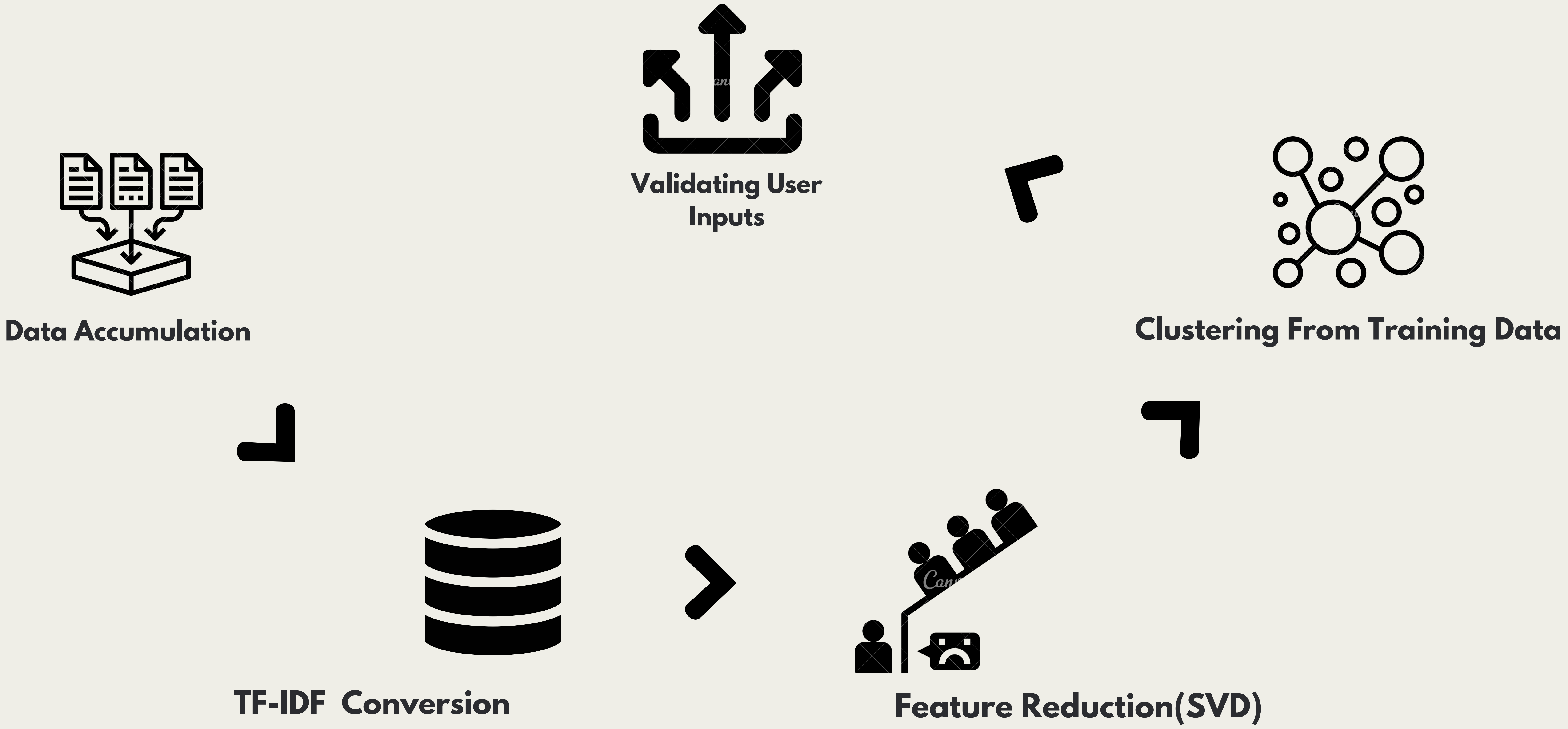
THESE ARE NUMERICAL REPRESENTATIONS OF VARIOUS JOB PROFILES. EACH VECTOR WILL LIKELY CONTAIN INFORMATION ABOUT THE REQUIRED SKILLS, DOMAIN KNOWLEDGE, AND OTHER RELEVANT CRITERIA FOR A PARTICULAR JOB.

COSINE SIMILARITY:

A MATHEMATICAL MEASURE IS USED TO DETERMINE HOW SIMILAR THE USER PROFILE VECTOR IS TO EACH JOB PROFILE VECTOR. CLOSENESS TO SIMILARITY REFERS TO RELATIBILITY TO A JOB PROFILE



*MCC - PIPELINE



POSSIBLE CHALLENGES

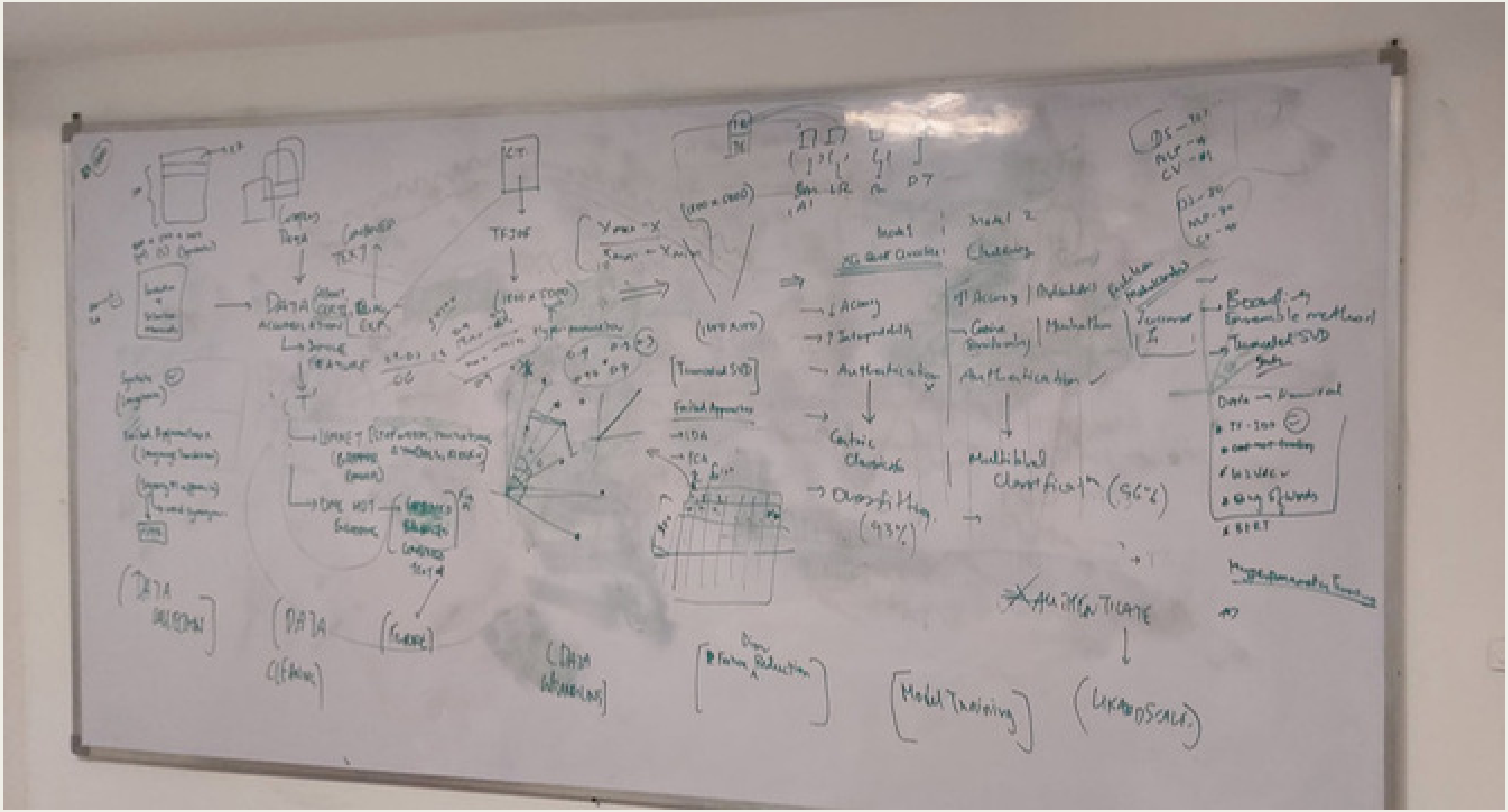
- **Deployment at Plaksha**
  - Integration with CDC: Implement the system within CDC's framework for data-driven job role and company insights.
  - Resume Processing: Use the model to analyse student resumes for matching with potential job roles.

- **Steps for Deployment**
  - Data Collection and Privacy: Ensure ethical collection and secure storage of resumes while respecting data privacy.
  - Integration with Educational Systems: Seamlessly integrate the system with existing university platforms.

Challenges in Scaling Up:

  - Diverse Profiles Handling: Accurately accommodate a wide range of student academic backgrounds and interests.
  - Customisation and Flexibility: Ensure the system's adaptability to individual needs and various industries.

BEHIND THE SCENES



Thank you!