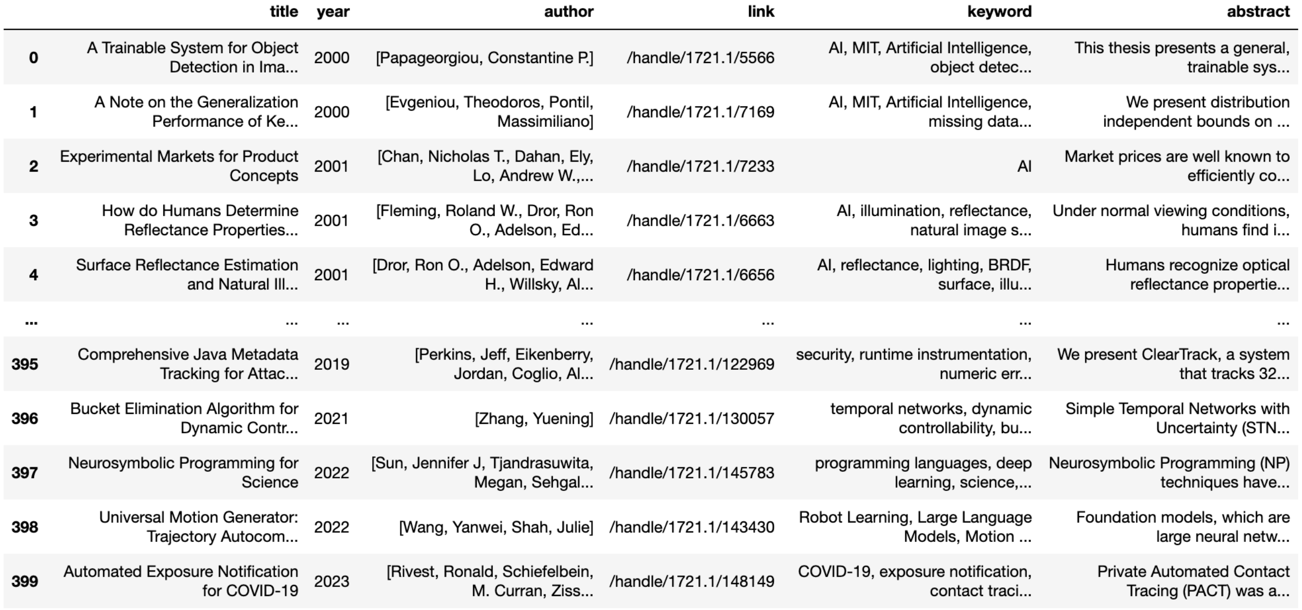
RA Task – XiaoyuOuyang

Part 1: Compile data

The objective of the following section is to generate a representative sample of academic papers emanating from the MIT Computer Science and Artificial Intelligence Lab. Specifically, papers published after 2000 are considered for inclusion. A web scraping methodology is employed to gather pertinent information from each paper, such as the title, year of publication, authors, keywords, and abstracts. This procedure entails extracting data from every paper on each webpage and subsequently navigating through the ensuing pages to retrieve additional papers. To ensure that the final dataset encapsulates a span from 2000 to 2023, a random subset of 400 papers is chosen from the complete dataset, which is then exported in CSV format for further analysis. Below are some sample rows of the output data frame:

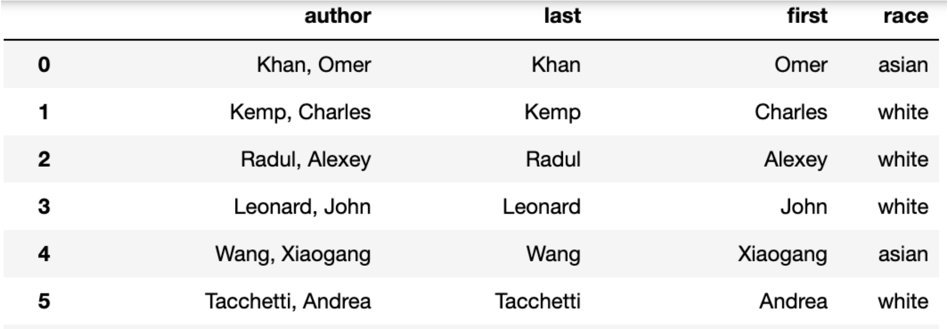


	title	year	author	link	keyword	abstract
0	A Trainable System for Object Detection in Ima...	2000	[Papageorgiou, Constantine P.]	/handle/1721.1/5566	AI, MIT, Artificial Intelligence, object detec...	This thesis presents a general, trainable sys...
1	A Note on the Generalization Performance of Ke...	2000	[Evgeniou, Theodoros, Pontil, Massimiliano]	/handle/1721.1/7169	AI, MIT, Artificial Intelligence, missing data...	We present distribution independent bounds on ...
2	Experimental Markets for Product Concepts	2001	[Chan, Nicholas T., Dahan, Ely, Lo, Andrew W.,...	/handle/1721.1/7233	AI	Market prices are well known to efficiently co...
3	How do Humans Determine Reflectance Properties...	2001	[Fleming, Roland W., Dror, Ron O., Adelson, Ed...	/handle/1721.1/6663	AI, illumination, reflectance, natural image s...	Under normal viewing conditions, humans find i...
4	Surface Reflectance Estimation and Natural Ill...	2001	[Dror, Ron O., Adelson, Edward H., Willsky, Al...	/handle/1721.1/6656	AI, reflectance, lighting, BRDF, surface, illu...	Humans recognize optical reflectance propertie...
...	...	...	...	...	...	...
395	Comprehensive Java Metadata Tracking for Attac...	2019	[Perkins, Jeff, Eikenberry, Jordan, Coglio, Al...	/handle/1721.1/122969	security, runtime instrumentation, numeric err...	We present ClearTrack, a system that tracks 32...
396	Bucket Elimination Algorithm for Dynamic Contr...	2021	[Zhang, Yuening]	/handle/1721.1/130057	temporal networks, dynamic controllability, bu...	Simple Temporal Networks with Uncertainty (STN...
397	Neurosymbolic Programming for Science	2022	[Sun, Jennifer J, Tjandrasuwita, Megan, Sehgal...	/handle/1721.1/145783	programming languages, deep learning, science,...	Neurosymbolic Programming (NP) techniques have...
398	Universal Motion Generator: Trajectory Autocom...	2022	[Wang, Yanwei, Shah, Julie]	/handle/1721.1/143430	Robot Learning, Large Language Models, Motion ...	Foundation models, which are large neural netw...
399	Automated Exposure Notification for COVID-19	2023	[Rivest, Ronald, Schiefelbein, M. Curran, Ziss...	/handle/1721.1/148149	COVID-19, exposure notification, contact traci...	Private Automated Contact Tracing (PACT) was a...

Table 1. data frame of papers' information

Part 2: Categorize Ethnicity of Authors

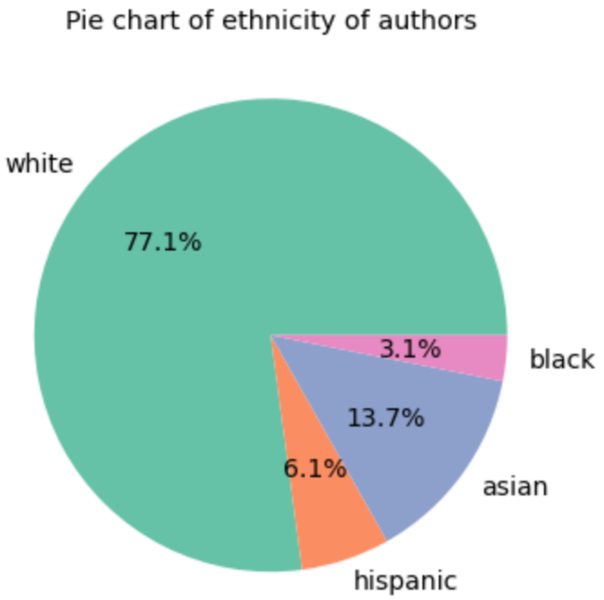
To determine the ethnicity of the authors, the full name of each author was first divided into a first name and a last name. Authors with incomplete names were excluded from further analysis. Due to the lack of reliable information sources regarding the nationality or ethnicity of the authors, a Python package known as "ethnicolr" was utilized to predict the authors' most probable ethnicities. The "pred_fl_reg_name" function within the package, which employs the full-name FL model, was specifically utilized to provide mean, standard error, lower and upper bounds of the confidence interval for races including White, Black, Hispanic, Asian. And the most probable race of the individual is the one with the highest mean. The first few rows of the data frame are shown below.



	author	last	first	race
0	Khan, Omer	Khan	Omer	asian
1	Kemp, Charles	Kemp	Charles	white
2	Radul, Alexey	Radul	Alexey	white
3	Leonard, John	Leonard	John	white
4	Wang, Xiaogang	Wang	Xiaogang	asian
5	Tacchetti, Andrea	Tacchetti	Andrea	white

Table 2. data frame of authors' names and ethnicity

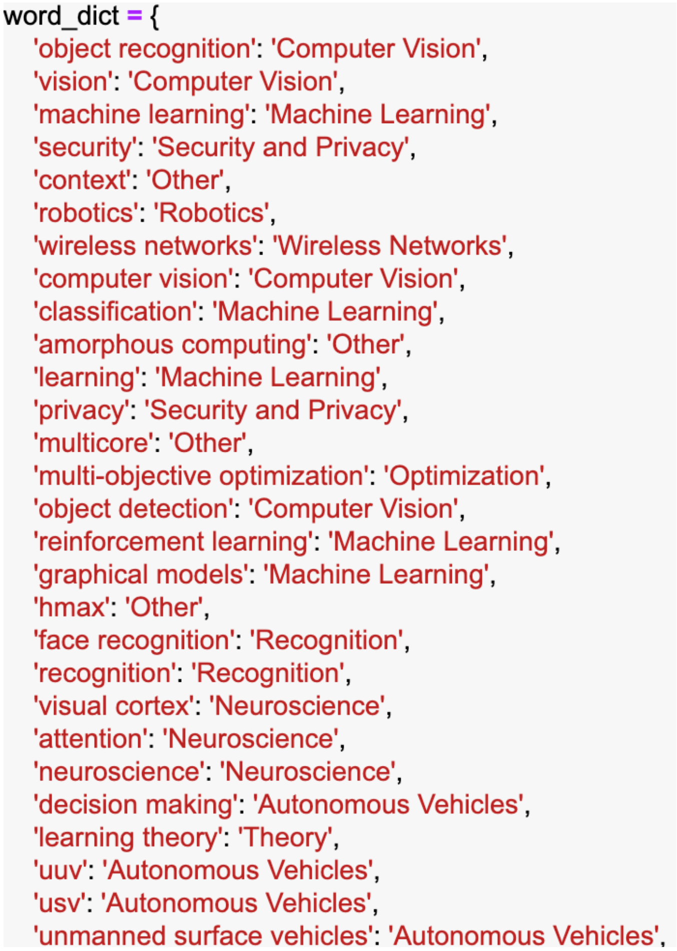
Ultimately, the pie chart Graph 1 was created to depict the proportion of authors from different ethnic groups among the 400 papers in the dataset. According to the graph, a significant majority of authors associated with papers which are published after 2000 and procured from MIT's Computer Science and Artificial Intelligence Lab identify as White, comprising approximately 77% of the total. Conversely, individuals identifying as Asian constitute 13.7% of the authors, while Hispanic and Black authors represent only 6.1% and 3.1%, respectively. These results suggest a disparity in racial representation within the MIT CS&AI Lab's academic publications, with White authors being overrepresented compared to other racial groups.



Graph 1. pie chart of ethnicity of author

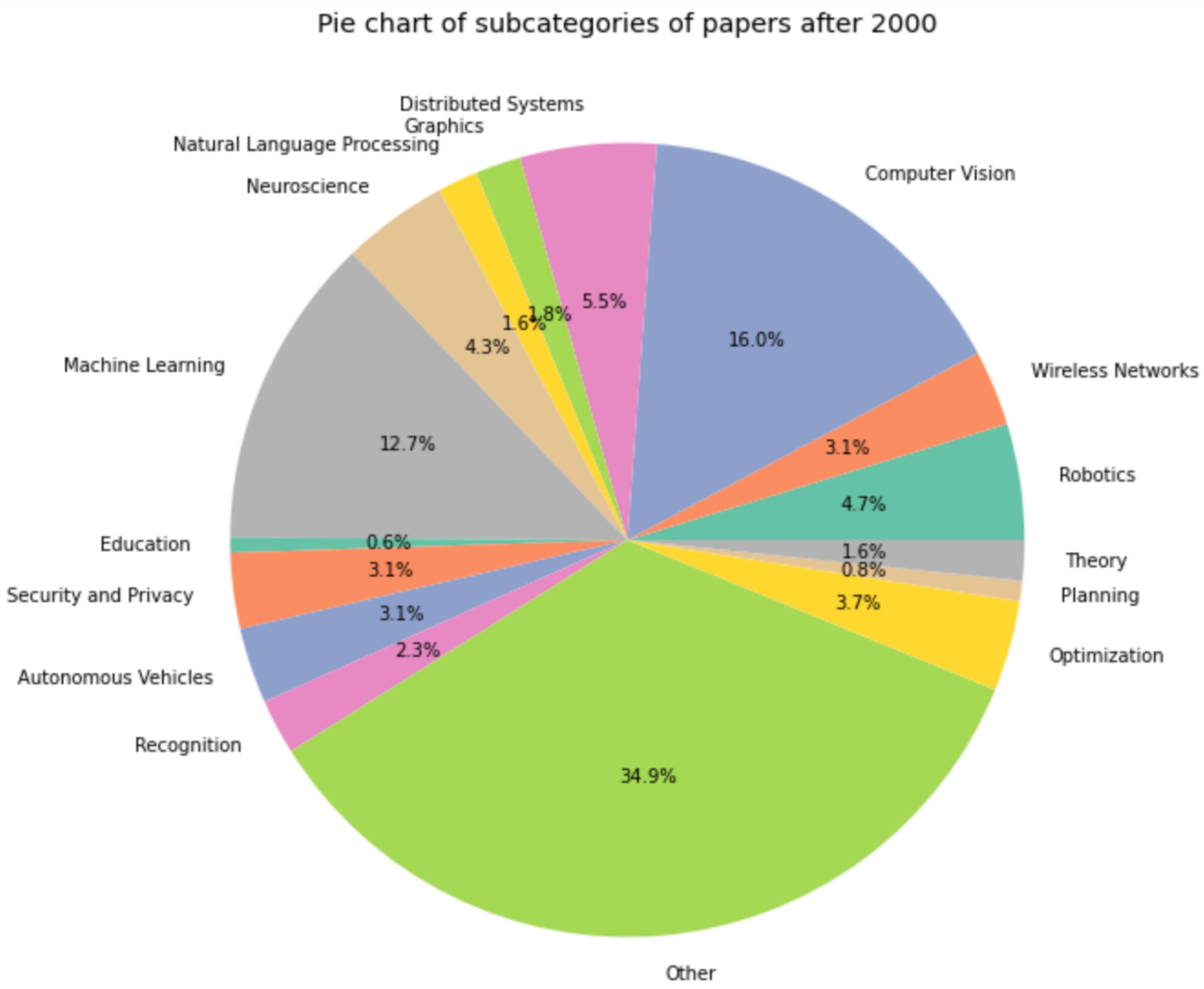
<u>Part 3: Categorize Research Topics</u>

To categorize the research topics of these papers, I used the keywords provided in the webpage of each article as the primary source. Specifically, I selected 200 most common keywords among all keywords of papers in the sample and defined 16 subcategories of AI including: 'Robotics', 'Wireless Networks', 'Computer Vision', 'Distributed Systems', 'Graphics', 'Natural Language Processing', 'Neuroscience', 'Machine Learning', 'Education', 'Security and Privacy', 'Autonomous Vehicles', 'Recognition', 'Other', 'Optimization', 'Planning', and 'Theory' (in alphabetical order). Then I utilized ChatGPT to assign each keyword to one of 16 subcategories of AI with manual adjustment and created a dictionary to store the mapping. Some key-value pairs are shown below.



```
word_dict = {
    'object recognition': 'Computer Vision',
    'vision': 'Computer Vision',
    'machine learning': 'Machine Learning',
    'security': 'Security and Privacy',
    'context': 'Other',
    'robotics': 'Robotics',
    'wireless networks': 'Wireless Networks',
    'computer vision': 'Computer Vision',
    'classification': 'Machine Learning',
    'amorphous computing': 'Other',
    'learning': 'Machine Learning',
    'privacy': 'Security and Privacy',
    'multicore': 'Other',
    'multi-objective optimization': 'Optimization',
    'object detection': 'Computer Vision',
    'reinforcement learning': 'Machine Learning',
    'graphical models': 'Machine Learning',
    'hmax': 'Other',
    'face recognition': 'Recognition',
    'recognition': 'Recognition',
    'visual cortex': 'Neuroscience',
    'attention': 'Neuroscience',
    'neuroscience': 'Neuroscience',
    'decision making': 'Autonomous Vehicles',
    'learning theory': 'Theory',
    'uuv': 'Autonomous Vehicles',
    'usv': 'Autonomous Vehicles',
    'unmanned surface vehicles': 'Autonomous Vehicles',
```

Ultimately, a pie chart Graph 2 is created to show the percentages of subcategories of all papers after 2000 as shown below. The most popular research subcategories of AI since 2000 are computer vision (16.0%) and general machine learning (12.7%) with all other categories lower than 10%.

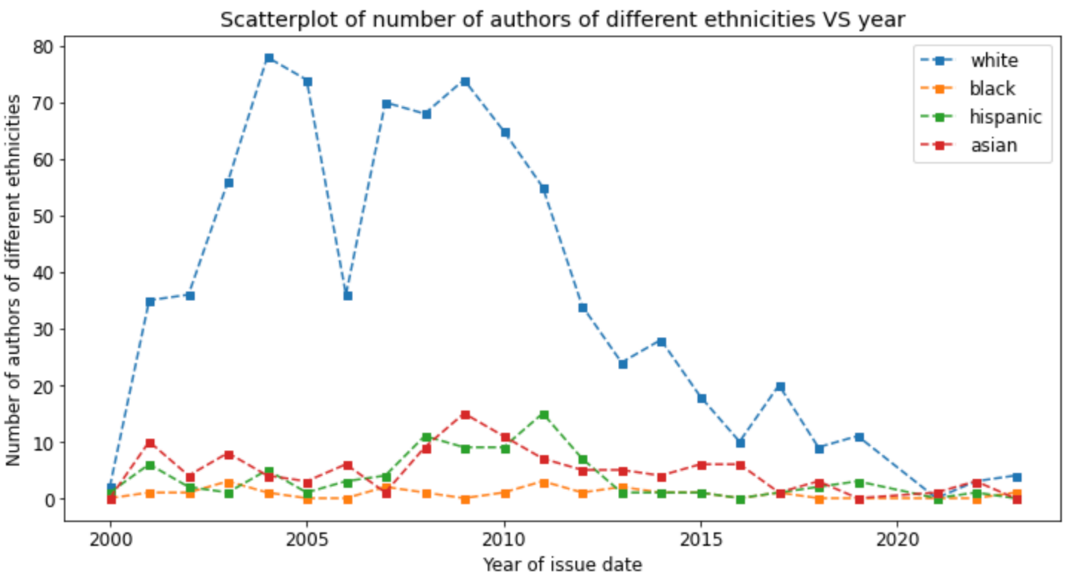


Graph 2. Pie chart of subcategories of papers after 2000

Part 4: Identify Patterns Over Time

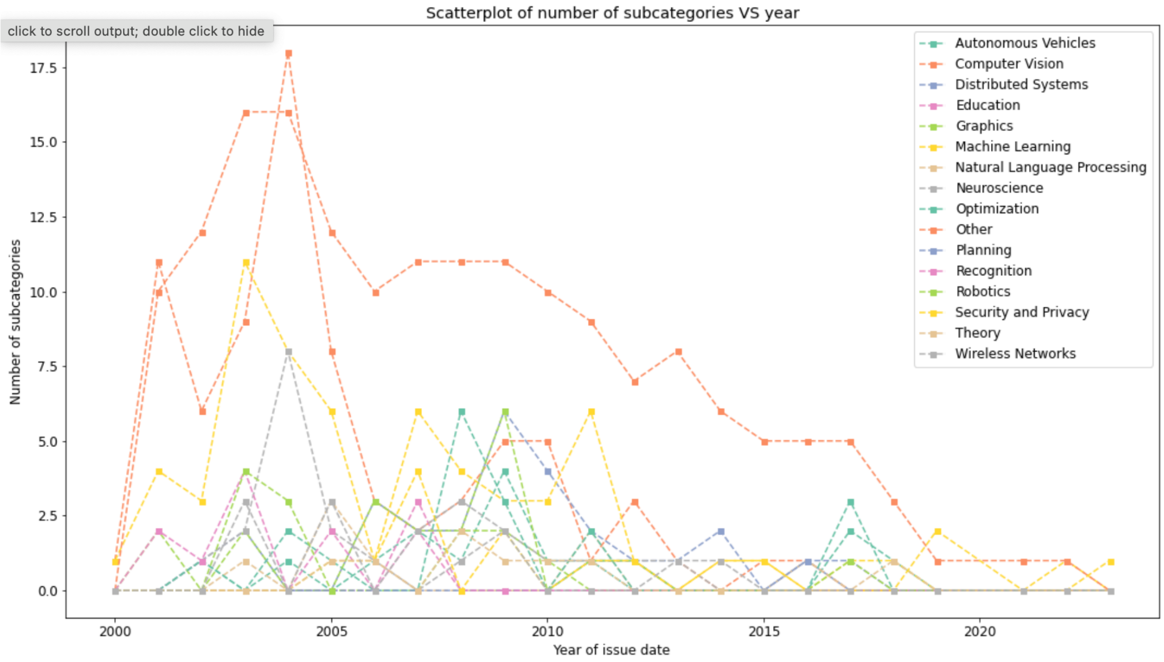
- Ethnicity

From the scatterplot in Graph 3, it is observed that in the period 2005-2006 and 2008-2010, authors of all ethnicities have published many papers, while the number of papers declines over time after 2010. This general pattern applies to White, Hispanic, and Asian authors but not to Black authors who published approximately constant and small number of papers through 2000 to 2023.



Graph 3

- Subcategories

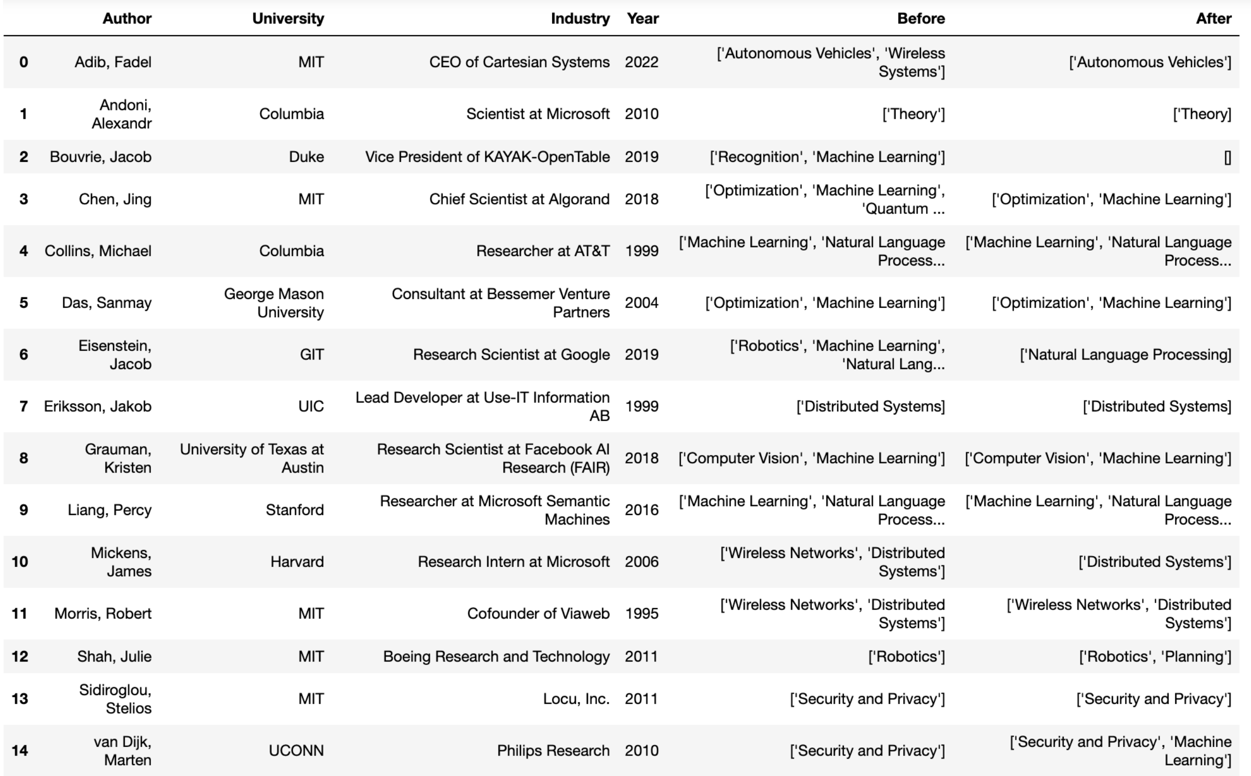


Graph 4

Part 5. Career Trajectories

This section involved a sample of 50 authors who were randomly selected. From this initial group, 30 authors were further selected as they were found to be current faculty members of US top ranked universities, based on their LinkedIn pages and online CVs. Then I researched their career paths and relevant information such as the universities where they were employed, their professional positions outside of academia, and the year they began working in industry were recorded in an Excel spreadsheet.

Using the dataframe containing information on 15 professors, it was found that none of them had published papers in MIT's Computer Science and Artificial Intelligence Lab both before and after their period of employment outside of academia from the pool of papers after 2000. To further explore this issue, the authors' works were searched on the Google Scholar website in order to obtain a larger sample size. Subsequently, the papers were categorized based mainly on their titles and abstracts, and relevant information was recorded in the dataframe presented below.



	Author	University	Industry	Year	Before	After
0	Adib, Fadel	MIT	CEO of Cartesian Systems	2022	['Autonomous Vehicles', 'Wireless Systems']	['Autonomous Vehicles']
1	Andoni, Alexandr	Columbia	Scientist at Microsoft	2010	['Theory']	['Theory']
2	Bouvrie, Jacob	Duke	Vice President of KAYAK-OpenTable	2019	['Recognition', 'Machine Learning']	[]
3	Chen, Jing	MIT	Chief Scientist at Algorand	2018	['Optimization', 'Machine Learning', 'Quantum ...	['Optimization', 'Machine Learning']
4	Collins, Michael	Columbia	Researcher at AT&T	1999	['Machine Learning', 'Natural Language Process...	['Machine Learning', 'Natural Language Process...
5	Das, Sanmay	George Mason University	Consultant at Bessemer Venture Partners	2004	['Optimization', 'Machine Learning']	['Optimization', 'Machine Learning']
6	Eisenstein, Jacob	GIT	Research Scientist at Google	2019	['Robotics', 'Machine Learning', 'Natural Lang...	['Natural Language Processing]
7	Eriksson, Jakob	UIC	Lead Developer at Use-IT Information AB	1999	['Distributed Systems]	['Distributed Systems]
8	Grauman, Kristen	University of Texas at Austin	Research Scientist at Facebook AI Research (FAIR)	2018	['Computer Vision', 'Machine Learning']	['Computer Vision', 'Machine Learning']
9	Liang, Percy	Stanford	Researcher at Microsoft Semantic Machines	2016	['Machine Learning', 'Natural Language Process...	['Machine Learning', 'Natural Language Process...
10	Mickens, James	Harvard	Research Intern at Microsoft	2006	['Wireless Networks', 'Distributed Systems']	['Distributed Systems']
11	Morris, Robert	MIT	Cofounder of Viaweb	1995	['Wireless Networks', 'Distributed Systems']	['Wireless Networks', 'Distributed Systems']
12	Shah, Julie	MIT	Boeing Research and Technology	2011	['Robotics']	['Robotics', 'Planning']
13	Sidiroglou, Stelios	MIT	Locu, Inc.	2011	['Security and Privacy']	['Security and Privacy']
14	van Dijk, Marten	UCONN	Philips Research	2010	['Security and Privacy']	['Security and Privacy', 'Machine Learning']

Based on the presented data frame above, it is apparent that the research topics of all professors remain substantially relevant both before and after their involvement in activities outside of academia. The data indicates that 4 professors retained their research interests exactly as they were before, while 5 professors have reduced some of their previous research interests, and 2 have broadened their research scope. Therefore, it can be inferred that a majority of the professors maintain their focus on their established area of expertise, irrespective of their engagement in non-academic pursuits, while some choose to narrow their research focus or explore new directions over time.

It is important to acknowledge that certain limitations still persist, which could benefit from further refinement. Notably, the sample size of authors being limited to 15 may compromise the generalizability of our findings. Therefore, the inclusion of a greater number of papers from diverse data sources, such as ResearchGate, may provide a more comprehensive and representative dataset. Moreover, the process of categorizing research topics solely based on paper titles or keywords could be enhanced to ensure greater efficiency and accuracy. The implementation of machine learning techniques to analyze abstracts of papers is a promising avenue for improving the precision of research topic categorization.