MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Big Tech and research funding:

A bibliometric approach

Emil Ahmadov

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

BIG TECH AND RESEARCH FUNDING:

A BIBLIOMETRIC APPROACH

by

Emil Ahmadov

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor / Co Supervisor: Bruno Damásio

Co Supervisor*:* Sandro Mendonça

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Emil Ahmadov

Lisbon, 23rd of November 2022

DEDICATION

Dedicated to my father, Mustafa Ahmadov, who himself is a genuis and one of the most caring people one can ever see. He has done enormous things for me to be at the place I am right now. I have always felt his unconditional support throughout my studies and he has been the one to teach me to never give up in life.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Bruno Damasio and Sandro Mendonça for their support for developing my thesis. Also with their support, a version of this thesis was sent to Eu-SPRI conference (https://euspri2022.nl/) and was accepted to be presented.

I am also grateful to my parents, Mustafa Ahmadov and Fargana Ahmadova for supporting me during these years, even though being far away, to my brother, Yashar, for always being there whenever I needed him and for his guidance. I would also like to thank my sisters, Sakina and Farida for their support. Last, but not least to the friends for their support and inspiration during the process. Furthermore, I would like to thank my amazing girlfriend, Cristina, who always provided her support throughout this process.

ABSTRACT

Technology companies have radically transformed our daily life in the recent years with help of the wide usage of internet. While transforming our lives, these companies also have grown up even bigger in the recent times and have become more powerful not only financially, but also in terms of computing power and data. Although there have been lots of research done on the influence of large digital economy players (Big Tech) in different fields, the academic influence of these companies is little understood. By drawing on 130,000 academic papers for which there is evidence of support by the Big Tech, the present work applies bibliometric approaches (on the metadata) and text mining techniques (on the contents) to shed a light on the outcomes of this relationship. In particular, we take into consideration research funding (direct strategies) and conference sponsorships (indirect strategies) to empirically explore this relatively unexplored side of Big Tech's influence in contemporary society. While developing the analysis a key limitation was the scarcity of prior work exploring the connections between digital platforms and the scientific enterprise. There are several results that come to light from such a perspective, one of these findings is that among the research supported by Big Tech companies, there is big gap between the number of outcomes with the content about the technical perspectives (like machine learning or artificial intelligence) than the content about reflexive (say ethical or environmental) dimensions of innovation, ladder being very small. These findings may stimulate further inquiries into identifying the possible risks, if any, are generated from the direct and indirect financial support by corporate informational giants to academia. The causes and consequences of this non-market activity by companies with big market power may require further attention and research in this field.

KEYWORDS

Big Tech; Digital platforms; Research funding; Science policy; Regulation

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This thesis is concerned about the ways in which the science system owes to Big Tech, the large digital companies that rose to global dominance in a number of increasingly interrelated internet-based areas of business (software, search, social-media, content-streaming, etc.). We focus on research support, and on a particular set of possible types of material assistance: conference sponsorships (indirect support) and funding (direct support). Bibliometric techniques are used to identify the main corporate actors that are becoming central in providing assistance, institutions that are benefiting, the scientific fields involved, etc. Text mining techniques are deployed to determine the main topics of interest and in getting deeper insights regarding the content and context of the papers and conferences being targeted and encouraged.

Research is an organized activity that requires resources. To have access for these resources, some conditions must be met, for instance, research agents must abide by a set of written rules and unwritten norms that are traditionally non-market based. The allocation of resources to the scientific enterprise is non-trivial and constitutes an aspect of the research system that warrants examination. A reassessment becomes urgent when the context dramatically changes, in particular if new cohort of very powerful intangibles-intensive private players becomes interested in the academic knowledge-production activities.

The internet became a clearly recognized all-purpose informational infrastructure at around the year 2000 (see Freeman and Louçã, 2001). In the ensuing years the information and communication technology (ICT) sector went through number of structural changes. A number of leading companies of the new ICT paradigm found themselves hard-pressed to adapt their operations and products (e.g. Apple, Microsoft) as the “back-office” mode of operation became cloud-based and the “front-office” interface of convenience became mobile-driven. The window of opportunity opened by the “web 2.0” (user-derived raw materials and final goods) made room for the rise of a new cohort of operations based on structurally novel technologies and services (e.g. Amazon, Google, Facebook, Netflix). By the 2010 these companies were losing their counter-culture “Silicon Valley” allure and in the next years begun being referred to by shorthands such as “Big Tech”, “tech giants”, “digital platforms”, and a number of reshuffling acronyms (FAANG, GAFA, etc.) (see Oremus, 2017; Wu, 2018). At the beginning of the 2020s, these companies that controlled entire expansive “ecosystems” together had a combined value of over 8.5 trillion US dollars, which is higher than the GDP of all the countries in the world except US and China (O'Neill, 2022).

While these entities were accumulating economic power, effects were also being experienced in several other non-market dimensions as years went by. One of the examples of these is their impact on social values such as privacy (Zubboff, 2019), how their normal activities were able to shape the formation public discourse (US Senate, 2019; Wall Street Journal, Oct 8, 2019) and their lobbying investment interfered with the political process (Cowgill et al., 2021; Lévêque, 2011; Marty, 2021). In addition to forms of influence projection already known, this thesis inquires the footprint of these platforms in the scientific domain. There could be a number of metrics for this, a classic one being in-house R&D spending: for the year 2020, Amazon topped the world ranking, followed by Alphabet, while Apple, Microsoft and Facebook coming up in the top 10; this represents a step jump from 2010-2012 when the best performing company was Microsoft in the 29th place (Dernis et al., 2015; Irwin-hunt, 2020). In what concerns the outside research that Big Tech are able to induce there is a greater gap in the available knowledge. Given the financial might of these companies, the magnitude of the phenomenon is expected to be non-trivial. In other words, Big Tech money is possibly a major element at play in the contemporary scientific scene, mostly likely in the US for certain frontier fields.

Sponsors and funders can shape research in different ways, for example, sometimes by eliciting work in topics that would otherwise remain underexplored or other times, by guiding the development direction by selective supporting communities or agendas (see, e.g., Geuna, 1999; Lyall et al., 2013). Traditionally, the allocation of expenditure to research groups and research projects by public agency is a costly endeavour and many countries have installed research evaluation programmes, reflecting increasing demands for greater accountability (Geuna and Martin, 2003). However, the nexus of socio-economic and political-institutional circumstances underpinning research support defines the "social contract for science" and should be examined in historical context (Martin, 2003). Likewise, the incentives and organizational structures that comprise the sponsorship and funding environment evolve, which invite a heightened awareness of the potentialities of intended and unintended consequences (both good and bad) (David, 2004). For instance, corporate patronage of research by big business have poured big money into academia has been proven to lead to perverse outcomes that have proved salient in a variety of industries, from consumer goods to industrial core inputs (Oreskes and Conway, 2014). As the global driving forces shift, new questions can be raised about current chances and their possible implications for science, truth and the common good.

Here we seek answers in 134.748 papers Bibliometric data that were sourced from Web of Science (WoS) and connected to the Scopus database for data completeness. The corpus consists of two kinds of papers: those that are outcomes of conferences sponsored by Google, Amazon, Microsoft, Apple, Facebook, or Netflix (114480 papers), and those that acknowledge funding of these companies (20268 papers). These companies were selected because they are the ones with the biggest power in their fields. These companies have enormous amount of not only financial resources, but also extensive computing power and data (Rikap and Lundvall, 2022). Once filtered from false positives and cleared out of duplications, the database includes research areas, publication dates, affiliations, and countries as some of the fields considered. In addition to bibliometrics, the content (abstracts and titles) of the papers were further studied with recourse to text mining techniques. Text materials were then cleaned from the stopwords and recurring wordage, and the content investigated.

We find that Big Tech conference sponsorships became expressive after the year 2000 and acknowledgements to Big Tech support in published papers conspicuously surface after 2005. Big Tech support is greatly associated to technologies related to data science (e.g. machine learning, artificial intelligence, augmented reality), while underplaying critical topics related to the social sciences and the humanities (e.g. ethics, data privacy, environmental challenges).

In subsequent parts of the thesis will firstly address the necessary background and define the main concepts that are used in this analysis. Secondly, the procedure regarding data collection, cleaning, preparation, and analysis will be detailed. Subsequently, we will present the findings of the results and conclusions that can be drawn from these results.

2. RESEARCH, SUPPORT, AND FUNDING

2.1. Resource allocation to research activities

Time, skills, software and equipment, assistants and administrative staff, technical services, conference fees and travel costs are some of expenses must be met in order to conduct research, maintain teams, and sustain institutions. Although one of the main classic financial sources for science is the State, there have been initiatives to increase the cooperation between the academia and industry. Efforts to increase research cooperation have been widely supported by many governments from time to time. It is believed that higher cooperation between these two will result in several beneficial outcomes such as increase of resources for academic research, faster and accelerated knowledge transfer, increased competitiveness, and eventually leading to economic development (Bahrens and Grey, 2001).

There are two types of main sources of to support research activities: money that flows to the academic institution (restricted funds) and money that flows to universities for specific purposes (unrestricted funds) (Kenway et al., 2007). While first one relies on the state or government funding, alumni donations, or financial activities that institution generates income from, such as publishing, sales of assets etc., the second type of support comes from the external funders to the universities for specific purposes. These purposes can range from funder to funder which might mean there can less or more "strings attached".

Most of the academic organizations require researchers to give acknowledgement of the funders. For instance, according to IEEE's webpage, any support that was received (financial, in-kind support from funding bodies, sponsors, industrial partners, labs, and other collaborators) in writing of the academic paper must be acknowledged (Information for authors, 2017). Research support is thus one of the crucial points in the governance of research.

Resources allocated to research is thus an activity that is supposed to leave back some traces. Combining the standard bibliographic sources for papers that benefited from Big Tech support we come up with Figure 1. We searched for papers published in proceedings of conferences that acknowledged at least one of Big Tech as conference sponsor (we call this "indirect support") and for papers that themselves acknowledge some sort of indebtedness to Big Tech (we call this "direct support"). Since many name variations of supporters exist in many thousands of papers and because many can be duplicate (that is, a given conference or paper having received support from more than one entity at the same time), the returns are noisy. To get at an illustrative approximation of the phenomenon we applied a simple technique of truncation and consolidation: we took the raw data from the Top 50 entities and generated Top 20 tables for which we rank consolidated entities (say, Microsoft and Microsoft Research were considered to be same entity). We find i) the "big five" digital corporations to be in both rankings, ii) State-based funding institutions are not very collaborative with Big Tech, and iii) there are more data points for indirect than for direct support, as expected (Figure 1).

Sponsor	Count
Microsoft	75096
IEEE	55322
Google	46817
IBM	32808
Facebook	18338
Assoc Comp Machinery	16847
NSF	13103
Baidu	12309
Amazon	11407
Intel	10890
Huawei	10644
Qualcomm	7863
Adobe	6375
Springer	5910
Samsung	5762
ACM SIGCHI	5579
Nuance	5411
SAP	4725
KUKA	4561
Apple	4233

Funder	Count
Google	5026
Microsoft	4015
Amazon	877
Intel Corporation	618
Facebook Inc	474
NSF	373
Alfred P. Sloan Foundation	285
NVIDIA	228
IBM	222
Samsung	183
DARPA	179
NSERC	167
CNPq	158
Adobe	157
Huawei Technologies	157
CAPES	146
Vmware	138
Qualcomm	131
Cisco	119
Oracle	100

Source: An approximation to the Top 20 donor entities beside Big Tech, on the basis of WoS

Figure 1 – Papers benefiting from indirect and direct Big Tech support

2.2. Corporations investing in science

History has been witness of the influence by corporates on the academic research. Situations have been documented of too close proximity leading to conflicts of interest and even to falsification of evidence and manipulation the research outcomes.

One famous example of this is the “Big Tobacco” companies manipulating the truth about the tobacco consumption to maintain their business. Starting from the 1950s, tobacco companies hid the evidence that smoking is harmful and addictiveness of nicotine. In 1970s these companies spent enormous amount of money on the campaigns which aimed to deflect the science to expose the links between smoking and cancer. Big Tobacco also distributed briefing papers that promoted the ideas for the favor of these companies. It was in 1980s that Big Tobacco found its way to the science society and hired “independent” scientists to reverse popular and scientific opinions about the harms of smoking (WHO, 2019).

Additionally, there is “Big Oil”. Scandals related to giant oil companies have been reported in the past, where these companies have been associated to the denial or underplaying of global warming (Sullivan, 2020). Considering that global warming is one of the important environmental issues in the world today, which could have been tackled in the early stages, vested interests have waged campaigns, hid evidence, and steered stakeholders. And research has been caught in this maneuvering.

Marty Hoffert, who was a worker of Exxon, shared his predictions with his managers which was showing the possible results of fossil fuels. However, company made public statements contradicting the results of this research (Keane, 2020). Additionally, it was identified that companies undermined the climate science by directly funding think tanks and contrarian scientists (Jacques et al., 2008).

In addition to Big Oil and Big Tobacco, “Big Pharma” has also been known to affect the outcomes of scientific work. Although “sponsorship bias” was recognized for more than a century ago, it was in 1980 that empirical research about the topic was conducted in a systematic way (Jefferson, 2020). In ensuing research, it was found that higher rates of harms were associated with lower likelihood in the reports published, meaning, significant differences were found between protocols submitted and subsequent published reports.

Considering the above examples of Big Tobacco, Big Oil and Big Pharma, the attention on Big Tech is found lacking. There is a need for a closer look to the results of these companies pouring money into academia, borrowing their assets to researchers (data, computing power), and entwining them in their agendas. This becomes even more important knowing that, as mentioned above, these companies have the unmatched potential to impact every corner of the global society.

It is important to investigate how financial support of Big Tech companies affects research. For this purpose, we can start by recognizing the weight of these companies in scientific conferences and appreciating how research outputs refer to their help.

2.3. CORPORATE RESEARCH SUPPORT (IN THE NEW ERA)

The lead corporate players of the Information Revolution are essentially a new cohort, that is, they did not branch out of organizations forged during the industrial era (Louçã and Mendonça, 2002; Mendonça, 2006;). Some were early day protagonists of the year of the microcomputer (such as Apple and Microsoft) while others were born intangible as the Internet was already the new infrastructure (such as Google or Facebook), while a few cases exist of entrants in between these two waves (Amazon, which started as book e-tailer specialist).

Appearing before and after the year 2000, these incumbents and insurgents all grew to become giants in the global economy in the 2000s as “web 2.0” gained traction, i.e. a repositioning of the internet economy through a combination of cloud infrastructure (high fix costs) and user-generated content (network effects) that favored large integrated corporate ecosystems that would become known as “Digital Platforms”. These new actors have been shaping the nature of the “network society” ever since (Costa et al., 2019).

3. METHODOLOGY AND DATA

3.1. RESEARCH DESIGN

This thesis analyses data on scientific papers published up until recently that were somehow supported by Big Tech companies. We combine text mining and bibliometric techniques in order to gain perspective on the interaction between informational wealth and science system. Support to research was considered from two perspectives: conference sponsorships (indirect support) and paper support (direct support).

One the one hand, what is conference sponsorship indicating? International scientific meetings have grown into large events that are complex and expensive to organize. Key procedures involve planning, finding a venue, arranging accommodation for participants, establishing registration systems, setting up a communication strategy, inviting key-note speakers, etc. (Rivlin, 1995). Fundraising and attracting sponsors thus become leading duties for science event organizers, since registration fees alone are not enough income to put it in place. This relationship gives a special place to sponsors, which can become promoters and partners. While there is not much research on the motivations of sponsors, we will assume they have something to gain: they may wish a certain set of topics are researched, they may be interested in creating brand awareness, they may be interested in establishing or strengthening a reputation as science-friendly entities (“giving back to society”), creating favor with prospective new employees (like young talent who may be presenting papers, etc.)

On the other hand, what are paper acknowledgements indicating? Direct support for specific research or researchers is a much more targeted and selective type of support. This support may be in terms of funding or in-kind (access to data, use of computational power, etc.) and is directly beneficial to the teams involved and the research institutions around them (Morillo, 2016). Overall, evidence can be gleaned from available sources by exploring under-exploited aspects on the empirical material (see, e.g. Costa and Mendonça, 2019).

3.2. DEFINING BIG TECH

One can have a hard time to define which companies should be considered as “Big Tech”. In this thesis we have focused on the companies that are not only have revolutionized their industries, but also have huge power in terms of finance, computational power and data. Recent developments in the fields such as machine learning, artificial intelligence and etc. has brought very big importance to the data and use cases of data. However, these developments have also brought lots of doubts with itself, such as data privacy and ethical sides of the algorithms that are used by the companies (Abdalla and Abdalla, 2021) . These doubts also sometimes have been sparkled by the scandals around these companies. In addition to the enormous computing power and data, these companies also have very big financial power (Rikap and Lundvall, 2022). This money has been influencing a lot of sides of our world, which science world is no exception. Considering these, we have included Microsoft, Facebook, Amazon, Google, and Apple, which are for most of the people undoubtedly in Big Tech. In addition, we have also included Netflix, which has recently become an integral part of our lives and entertainment sector.

3.3. BIBLIOMETRICS AND TEXT MINING

Combining classic and novel scientometric techniques can unveil the trends and turns of new technologies and business models (see, e.g., Mendonça et al., 2022). This thesis deploys approximation techniques, bibliometric tools, and text mining techniques on the details of the papers that are supported by Big Tech. Since this is relatively unexplored field, simple approximations can also be beneficial for identifying some hidden patterns and open the way for further research. In addition, by using bibliometrics, we aim to identify main patterns in these papers and with the help of text mining we want to achieve to extract content patterns from documents (Tan et al., 2000).

As mentioned in section 2, we used a simple method of truncation and consolidation to arrive at an approximation of the phenomena that serves as an illustration. Raw data from the Top 50 entities was taken and entities that have different variations of their name were standardized and added together. For example, Microsoft has also other name variations such as “Microsoft Inc” or “Microsoft Research”, to avoid this breakdown, we changed both names to “Microsoft”. After this consolidation, we created Top 20 out of this data. Alternative to this method, would be identifying all the different variations of the entity names in the whole dataset and standardizing the names and get overall ranking. However, this would be very time consuming and sometimes it might not be possible to identify all the different versions of names of entities.

Bibliometric methodology utilizes the quantitative methods on bibliometric data of the scientific papers such as citation information (Broadus, 1987). Although, bibliometrics can be perceived as recently emerging methodology, first discussion about the bibliometrics have started in the 1950s (Wallin, 2005). However, the proliferation of bibliometric is relatively recent thanks to the emergence of scientific databases such as Web of Science and Scopus, which have been also utilized in this thesis. These databases have made the large bibliometric data accessible in a structured format. With the help of these databases, number of publications that have used bibliometrics have emerged in recent years with an average of 1021 publications in the last decade . One can see the utilization of bibliometrics in various fields such as human resources, marketing and management. Two features that make this methodology highly usable are (1) ability to work with large volumes of scientific data, and (2) producing high research impact (Donthu, 2021).

Other frequently utilized methods are meta-analysis and systematic literature reviews. Meta-analysis estimates “the overall strength and direction of effects or relationships,” and “the across-study variance in the distribution of effect-size estimates and the factors that explain this variance”. Although this methodology also has the ability to handle the huge amounts of literature, it mainly focuses on condensing the empirical evidence by examining the direction and strength of effects and relationships between different variables. It is used in the cases that focuses on summarizing the results without engaging with the content of the research papers. On the other hand, bibliometrics engages with the content and tries to analyze in different social and structural relationships among the numerous actors involved in research such as funders, countries, and authors. In contrast with meta-analysis and bibliometrics systematic literature review cannot handle large datasets because it is done manually and thus requires very narrow scope. Which in our case could not be utilized because of the enormous number of papers directly or indirectly supported by Big Tech thus creating a very large dataset.

Method	End goal	Use cases	Analysis
Bibliometric analysis	Summarizing the big amount of bibliometric data to find out insights about the intellectual structure and emerging trends	• Broad scope • Large Dataset	• Quantitativ e • Qualitative
Meta-analysis	Summarize the empirical evidence of relationship between variables	• Focus is summarizing instead of engaging with content • Broad scope • Large Dataset	• Quantitativ e
Systematic literature review	Summarize and synthesize the findings of existing literature on a specific topic	• Small scope • Small Dataset	• Qualitative

Figure 2 – Comparison of methodologies (Donthu, 2021)

Text mining is also a helpful methodology for extraction of patterns that might not be visible at first sight and interesting knowledge from unstructured documents that are in text format. Although it is an extension of data mining, it is not as simple as data mining because it deals with mostly unstructured form of data. At this point, it is important to compare text mining with Natural Language Processing (NLP) methodology. Although they both analyze the text data, there are certain differences between these two methodologies. NLP helps to understand the actions and emotions behind the languages by analyzing semantics and grammatical structures. In contrast, text mining does not consider the semantics and grammar. While end goal of NLP is to extract grammatical structures and sentiment from the language, text mining focuses mostly on quantifying the text data in a structured format and deals with patterns and frequency of words.

To reiterate, two perspectives will be used for investigating such effects: (1) examining the data from the research papers that are indirectly supported through the conference sponsorships and (2) the papers that are directly supported by these companies through direct funding or in-kind support.

It is worth noting that computer sciences, robotics, and electronics/digital fields seem to constitute a major user of proceedings for purposes of publications (Bar-Ilan, 2010) and also that the support sources reflected in the acknowledgements can be used to characterize the supporting entities (Meija and Kajikawa, 2018).

3.4. DATA COLLECTION

Bibliometric data of the scientific papers that are related to Big Tech were collected from the database of Web of Science (WOS). All the record that contains any of “Microsoft”, “Facebook”, “Google”, “Apple”, “Netflix” and “Amazon” in any field related to conference information (conference title, conference sponsor, conference date, conference location) of the record were collected and downloaded. Since there is a restriction on the number of records that can be exported at a time from the website, each batch of records were stored in different file. All the data fields that are related to the papers was exported, resulting in 114.480 records for sponsorships and 20.268 records for funding organizations.

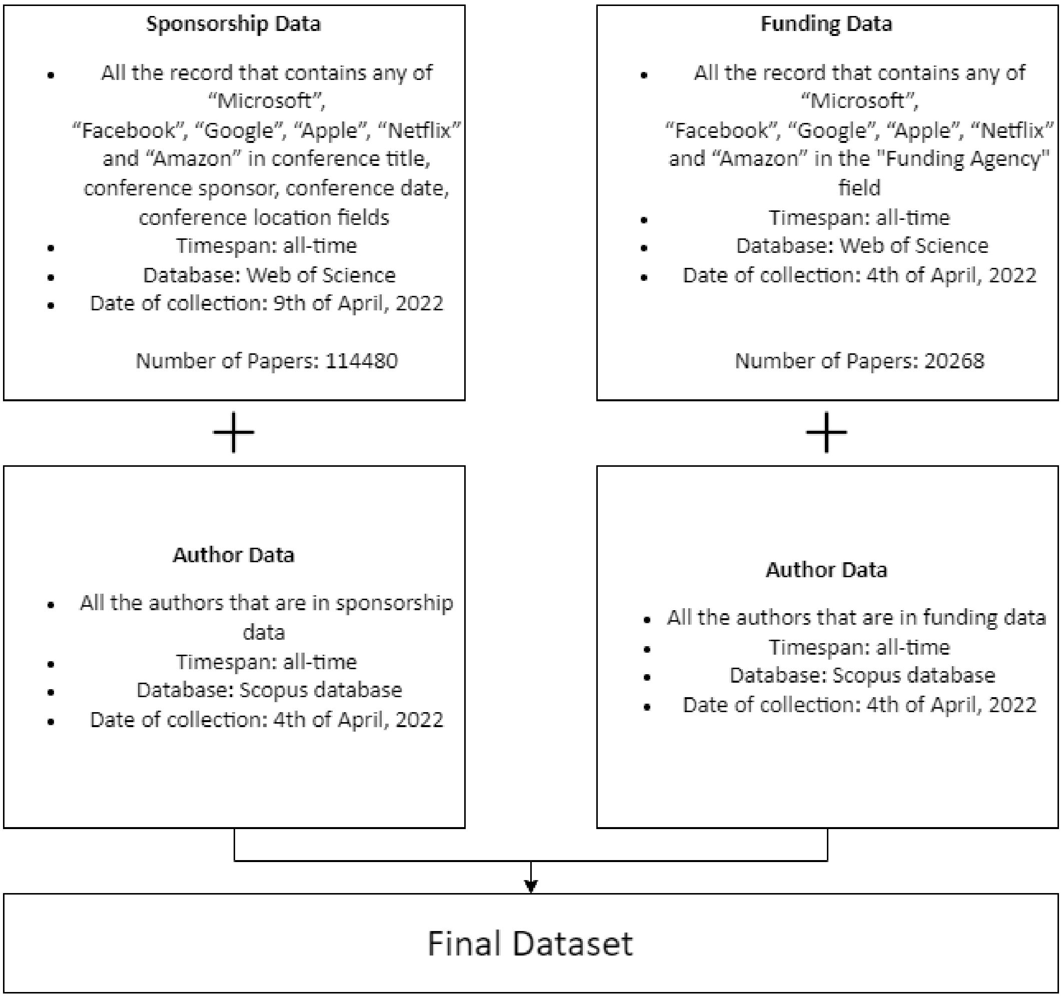


Figure 3 – Data collection procedure. Plus sign means complementation, meaning, those two databases were used in a complementary way to each other.

For author related data such as author nationality, their research areas, citation count and affiliate organization, Scopus database was used. Data was collected from the database through Scopus API by utilizing Pybliometrics package for Python (Rose and Kitchin, 2019). Web of Science data was connected to Scopus author data by ORCIDs and Scopus IDs of the authors. To achieve this goal, first ORCIDs of the authors of the papers were extracted from WOS data and then Scopus IDs of these authors were retrieved through Scopus API. Next, these Scopus IDs were stored in a file for shortening the data retrieval times in the next times. Subsequently, these identifiers were used for retrieving the data from the database. This retrieved data was also saved in a file for eliminating the need for using API in the next times since it is costly in terms of time to retrieve all these data each time.

3.5. Data Preparation

After records of the scientific papers were collected in different files, they were imported to Python and all these files were merged into one data frame. To maintain the consistency of the data, search was done to identify the duplicate records. As a result, 135 duplicate records were identified and discarded from the “sponsorships” data. No duplicates were found in “funding” data. To start with, dataset was filtered to only contain the papers that were published before 2022, since year was not completed at the data retrieval date. According to the purpose of the analysis, it was mandatory that data contains only the records that are related to the Big Tech companies. To maintain this, false positives, i.e., papers that are not related to these companies but appear there only because they contain the string with company name in the related fields (for example a paper with a funding organization name of “Apple Pickers Foundation”), were identified firstly.

In case of sponsorship data this was done by listing all different sponsor names that contain Big Tech company names in it and replacing unrelated sponsor names with “Other” (Appendix A). For funding dataset, different approach was deployed after the listing of all different variations of the funding organizations names, since in funding organizations there were much more false positives than the sponsors. Instead of replacing all the different false positives one by one, most frequent combinations

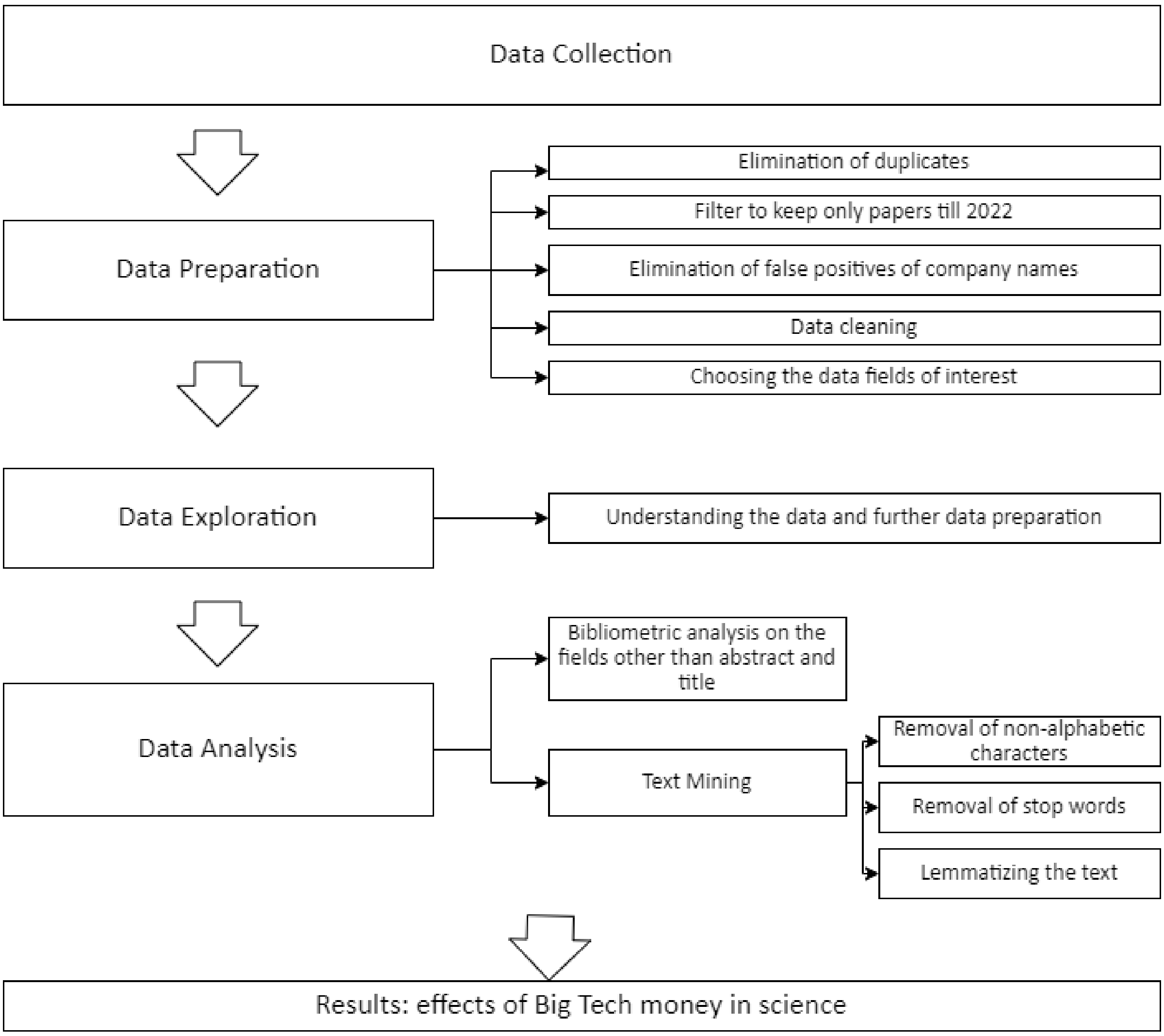


Figure 4 – Flowchart of the data processing and analysis

of the words in funding organizations' names that causes to false positives were used (combinations that were used for elimination, and exact sponsor names that were replaced, can be

found in Appendix B). For example: if the name contains both "apple" and "peach" or "apple" and "berry", these names were replaced by the string "Other". After this replacement of strings, data was filtered based on the fields "Funding Orgs" and "Conference Sponsor" which are respectively containing information about the funding organizations and sponsors of the conferences to include Big Tech companies in these fields. This filter resulted in 18396 and 113295 scientific paper records for each analysis. In total there were 68 data fields for each record, however data quality problems were spotted in some of these data fields in the exploratory data analysis process, one being the missing data problem. To tackle this problem, data fields that contained higher than 80% of missing values were discarded from the analysis. In addition, also some other fields were discarded from the data since those fields were not significant for the purpose of the analysis. This process resulted in the simplification of the analysis and data. After this simplification process, exploratory data analysis was conducted to understand the data better.

Data understanding also identified more room for data cleaning and more data engineering. Some of the columns were discarded from the analysis since there were other fields which already contained the information, and this was causing for data replication. For example, "Authors" field was discarded, because "Author Full Names" already contains that information. Furthermore, it was identified that some of the fields contained more than one value separated by delimiters. "Author Keywords", "Conference Sponsors", "Research Areas" and "Keywords Plus" are some of the fields that contained more than one value with delimiters. These records were cleaned from delimiters and were stored separately. There have been also identified records with different conference names but, referring to same conference. This was caused by typos or different written versions of the conference names. For identifying those cases, conferences with the same conference date were gathered in one list. After identification, they were standardized by keeping the longer or correct version of the conference names. In addition, to visualize the evolution of the number of papers throughout the time, years were extracted from date data. One of the examples for this is the extracting of conference year from conference date field. However, problems were identified in some of the date data such as years showing as "5007", in which case were replaced with the logical year values such as "2007". Additionally, there have been other kinds of data quality problems noticed that were related to the country names of the conference locations and author's addresses. Some of the country names were not given as it is in the official name or were given in the form of state name (mostly in USA). Those cases were identified and replaced with the official name of the country. Moreover, countries inside United Kingdom were indicated separately such as "England", "Scotland", "Wales" and "Northern Ireland". Those country names were converted to "UK" for more accurate analysis. Also, it was noticed that some of the country names were given as "Electr Network". After checking with the websites of these conferences, it was identified that these conferences were held online. Records were corrected accordingly.

3.6. Analysis

This thesis utilizes bibliometric techniques, co-word analysis and text mining techniques in combination to identify the main patterns in the outcomes of Big Tech financial support in academia.

Bibliometric techniques such as performance analysis were used to identify the mostly cited papers and authors.

After the data understanding, data cleaning and data preparation, data analysis was conducted. First part of data analysis included summarizing and quantifying the data. This part included all data fields other than article title and abstract. Distributions of the occurrences in each data field was analysed in this part of the analysis. Other than these, co-occurrence matrix was created from conference sponsors and funding agencies fields to identify the collaboration between Big Tech. To achieve this matrix, binary lists were created for each of Big Tech company indicating if the company name appear in the corresponding column. Subsequently these lists were stored in one data frame and distinct combination of these binary values and number of records were counted. This was done to identify the companies that appeared together most when providing financial support.

In the second part main objective was to analyze the contents of texts of the article titles and abstracts to identify the patterns and understanding the contents of the papers better. In this stage, firstly, article titles were divided into time periods according to their publication years. While doing this division, titles from articles that are published before 1991, between 1991 and 2001, 2001-2011, 2011-2021 were grouped together. Reason for grouping the articles before 1991 together was that the number of articles published in that period were relatively smaller when compared to other time periods. Next, all non-alphabetical characters were replaced with space in these titles and stop-words were cleaned from the texts. Stop words are the most frequently used words that does not add much value to the text. Stop word list that was used in the text cleaning is provided in Appendix C. Next, words were lemmatized just to keep the vocabulary form of the words. Afterwards, all unigrams, bigrams and trigrams were identified and stored in the corresponding decade lists.

4. BIG TECH'S INDIRECT SUPPORT: CONFERENCE SPONSORSHIPS

Between 1988 and 2021, we detect around 1300 conferences held with the sponsorships of Big Tech companies, reaching to highest number of conferences (103) in 2019 and the lowest number of conferences (0) in 1994. First boom in the number of conferences can be observed in early 2000's with the use of internet becoming more widespread and second boom in early 2010's. Drastic decrease in the number of conferences held in 2020 and 2021 can be also observed with the effect of global pandemic (Figure 2). These conferences were held in several cities in almost 70 countries around the world and have produced around 113295 academic papers in total. Besides the locations of the conferences, geography of authors that were involved in the research also captures numerous countries.

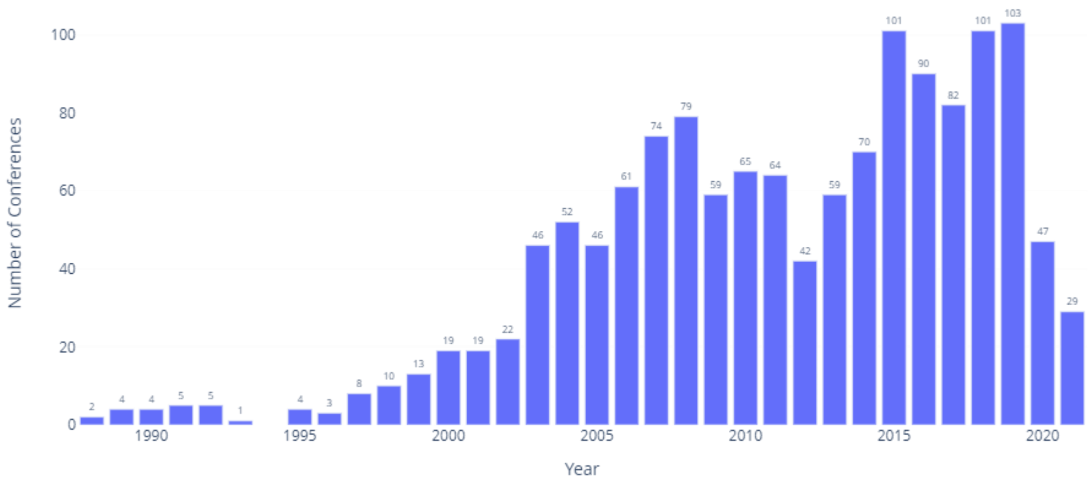


Figure 5 – Number of Big Tech sponsored conferences per year

To understand the publications related to conferences, we first need to understand who main sponsors of these conferences were. Microsoft is referred in 960 of these conferences as sponsor, making them 1st among the Big Tech, where Google is the second top sponsor with 487 conferences sponsored by them and Amazon, Apple, Netflix share the next places respectively being 4th, 5th, and 6th. In terms of number of outputs from these conferences, Microsoft is the top sponsor of these papers with indirectly supporting around 80 thousand of papers, which corresponds to almost 70% of all papers published, whereas Google is referred in around 50 thousand papers as conference sponsor and is the second biggest sponsor of these papers among the Big Tech (Table 1, Figure 3). Facebook comes in third place among Big Tech companies with 21 thousand publications related to the conferences sponsored by them. Fourth, among the big tech companies, is Amazon with 18 thousand publications related. Least sponsorship is coming from Netflix with around one thousand papers relating to the conferences sponsored by the company and Apple takes fifth place with around 7 thousand papers.

It is also important to observe the dynamics of each company's sponsorship in terms of timeline. When number of papers per each year was analysed, it was seen that Apple is the earliest sponsor that was mentioned in the papers as conference sponsor among the Big Tech, starting from 1989. However, number of papers are not very consistent for Apple, meaning, there are drastic changes from year to year and some years there are no papers mentioning the company, such as between 2010 and 2015.

For example, there are 900 papers in company's name for the year of 2003 and in the next year there are only 63 papers. Highest number of papers were published in 2018 with 950 papers. Second earliest mentioned company is Microsoft, with mentions in 183 papers in 1990. Number of papers follows steady increase between 1995 and 2008, with exceptions being only in 2000, 2006, and 2007. After 2008 there is a drastic decrease in the number of papers and in the next 3 years number of papers fluctuate between 4077 and 3242. From 2013 to 2015 number of papers steadily increase and reaches the global maximum in 2015 with 5606. After 2015, number of papers steadily decrease, reaching to 2265 in 2020. Although third earliest mentioned company is Amazon with 62 papers in 2000, there are no papers coming from the conferences supported by the company until 2013. While number of papers are close in 2013 and 2014, there is more than 3 times increase in 2015, corresponding to 2267 papers. In the next two years number of papers are almost on same level with 2252 and 2025 papers correspondingly. Interestingly, there is another drastic increase in 2018, reaching to 3841 papers and the global maximum. Even though there is drastic decrease in 2020 with the effect of pandemic, in 2021 company reached to almost same level as 2017 with 1909 papers. Fourth earliest mentioned company is Google in 2003 with 84 papers referring to the company as conference sponsor. Number of papers increases between 2005 and 2015 only exceptions being in 2009 and 2012, reaching the global maximum in 2015 with 6439 papers. There is a drastic decrease in 2016, and number of papers stayed almost on the same level till 2019, where second highest number is observed with 5824 papers. As in the other cases, with the pandemic, there is a drastic decrease in the number of papers. However, in 2021 huge increase can be observed with 3728 papers. Relatively younger companies, Facebook and Netflix share the next two places respectively, with the papers first referring to the companies as conference sponsors in 2012 and 2017. There is a steady increase in the papers mentioning Facebook, between 2012 and 2015 reaching to 2897 papers in 2015. Although there is a decrease in 2016, number of papers steadily increases till 2019 and reaching the global maximum of 3950 papers. Effect of the pandemic can be also observed in Facebook data. For Netflix, steady increase continued till 2019 with 722 papers and in 2020 pandemic effect can be observed (Appendix H).

Table 1 – Number of papers from conferences sponsored by Big Tech

Sponsor	**Count**
Microsoft	80342
Google	50835
Facebook	21453
Amazon	17784
Apple	7464
Netflix	1561

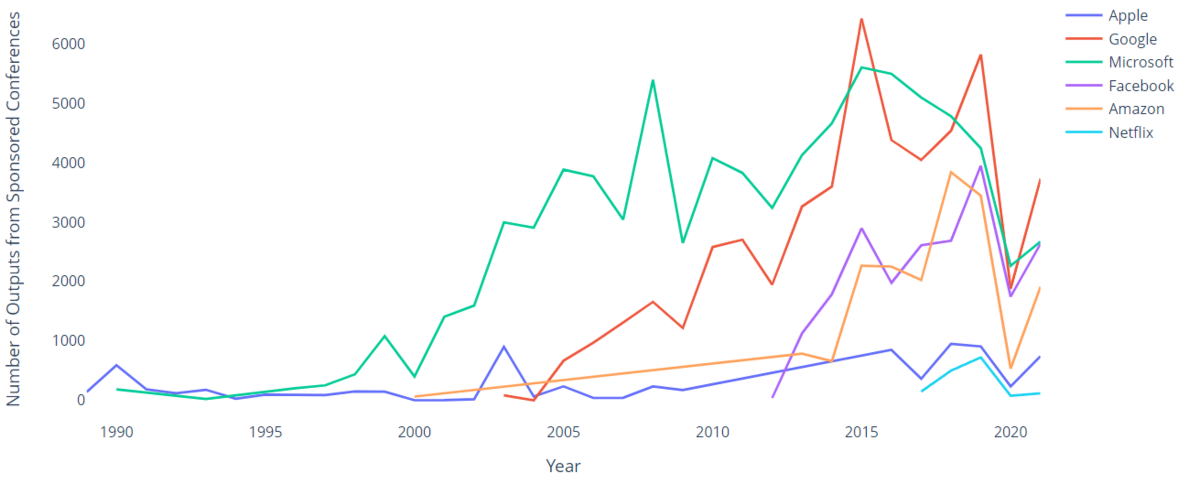


Figure 6 – Number of papers from conferences supported by Big Tech, over time

While analysing the sponsorship counts for Big Tech it is also important to pay attention to the overlapping sponsorships between these companies, which might suggest the mutual interest of the companies. Co-occurrence matrix shows that most overlap occurs between Google and Microsoft among these companies, with 13468 papers referring to only these companies as conference sponsors. Following three places are formed with the addition of the rest of the Big Tech to these two companies in order of Facebook, Amazon, and Apple. In contrast, lowest number of overlapping occurs between Google, Amazon, and Apple with only 17 papers in common. However, when only bilateral relations (sponsored only by two of Big Tech companies) are considered, second largest overlap is between Microsoft and Facebook with 2769 papers, whereas third largest overlap occurs between Google and Facebook with 1623 papers. Next biggest overlap for both of Microsoft and Google are same: Amazon. Nevertheless, Microsoft has more papers in common with Apple than Netflix, whereas it is the opposite for Google. As one can expect, for all the rest of Big Tech, biggest 2 overlaps are with Microsoft and Google respectively, except for Amazon, meaning Amazon has more overlaps with Google than Microsoft. Another interesting point is that Apple has no other bilateral overlaps with the rest of the companies. In addition, Facebook has more papers in common with Netflix, than with Amazon. Full co-occurrence matrix can be found in the Appendix D.

In terms of the types of publications, these papers consist of conference proceedings, series, books, and journals. Almost 103.000 of these papers are conference proceedings as can be expected, and around 7.000 are described as series publication type. The rest is shared by books and journals. However, in terms of the document type, almost all these documents are proceedings papers. In terms of publishers, IEEE and Springer are the top publishers with number of 47 thousand and 26 thousand of papers respectively. Third largest publisher is Association for Computing Machinery with around 21 thousand of papers. The analysis done on the publisher cities shows that New York is the city with the biggest number of papers published with around 55 thousand papers, which is almost half of all the papers, followed by Berlin with almost 19% of papers. Third biggest city for the number of publisher cities is Los Alamitos, California.

Third interesting point is to understand how many papers these conferences generated per year. First of these papers was published in 1988, which is the year that Big Tech started sponsoring conferences. In first two years of sponsorship number of papers published stays around same level and we can see a little peak in 1990 with the technological developments. However, in the following four years fluctuation can be seen with the global minimum in 1994. From 1995 till 1999 and from 2000 till 2005 there is steady increase in the number of publications related to the conferences with the highest percent change being in 1995 compared to previous year. Although there are fluctuations in following years, we can see increasing trend and reach to global maximum in 2019. In 2020 with the effect of pandemic, drastic decrease in the number of papers is observed (Figure 4). Nevertheless, two years preceding the pandemic are showing high numbers.

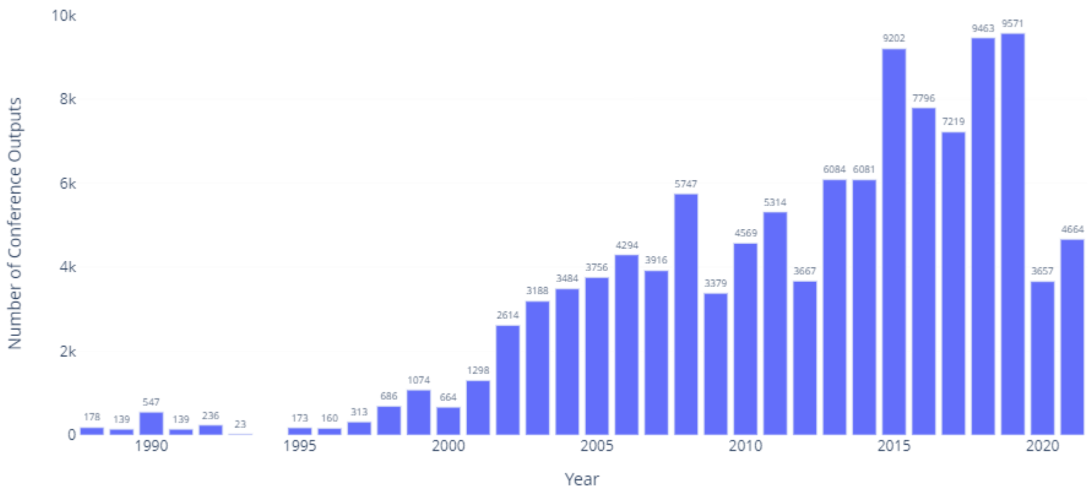


Figure 7 – Number of papers published in the conferences supported by Big Tech per year

Next important topic of question is about the locations of these conferences held. United States of America dominates the locations with around 30 percent of all papers that are published from the conferences sponsored by the tech giants, where China and Canada are second and third respectively with the percentage of 7 and 5 (Figure 5). The only European country that makes it to top five is Italy. In terms of continents, there are 1 from Australia, 2 countries from both North America and South America, 6 countries from Asia and 9 from Europe. This is an interesting fact that Europe accounts for the almost half of the locations. In addition, effects of pandemic can be seen in the location of the conferences, with 6.2 percent of the papers coming from conferences that were held online.

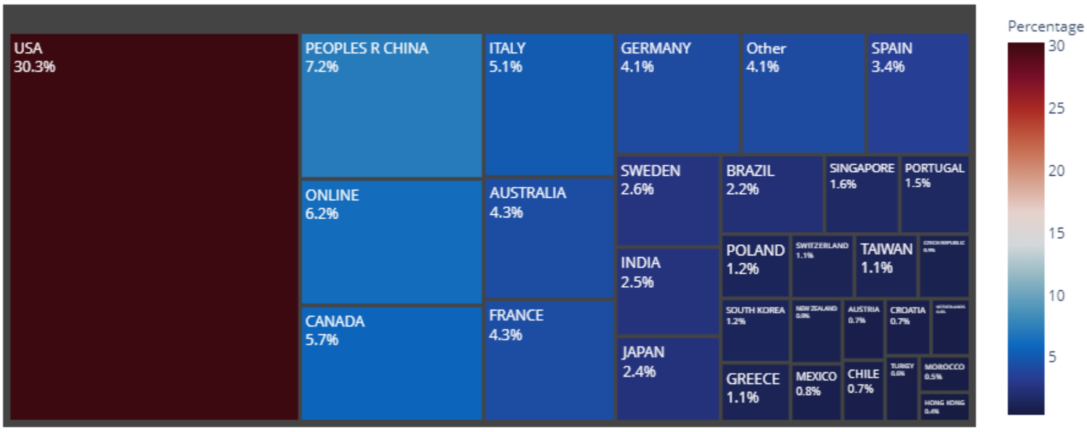


Figure 8 – Locations that the conferences were held sponsored by Big Tech

Even though conferences are held in different countries and different locations, almost all the languages of these papers are in English, followed by Portuguese with 168 papers and 54 in French. Spanish, Chinese and Czech are other languages that papers have been published in, with 2, 2, 1 paper respectively.

Each paper is assigned into different categories depending on the topics by Web of Science. To have more information about the topics of papers, these categories were analysed. Results show that main part of the research is related with technology and engineering while humanitarian subjects are in minority (Table 2). Category with the most papers is "Computer Science, Theory & Methods" with around 51 thousand records. Four of top 5 categories are related to Computer Science and one to Electrical and Electronic Engineering. Some of the categories that are not related to Computer Science, are Linguistics and Acoustics. When top 30 categories are analysed, at 25th place we can see Social Issues category.

Table 2 – Top 10 of the WoS categories of indirectly supported papers

Category	Count
Computer Science, Theory & Methods	51158
Engineeering, Electrical & Electronic	34712
Computer Science, Artificial Intelligence	34218
Computer Science, Information Systems	27787
Computer Science, Software Engineering	20236
Computer Science, Interdisciplinary Applications	12556
Telecommunications	11577
Computer Science, Hardware & Architecture	8680
Robotics	7534
Imaging Science & Photographic Technology	5925

In terms of the subjects of authors involved in these papers, Computer Science is again holding the first place with 759 authors associated with the subject. As a general view, most of the subjects of these authors are related with computer science, software, engineering, and artificial intelligence.

While most of the subjects in top 20 are related to engineering and computer sciences very little part of these authors are related to Human-Computer Interaction. To be more precise, only 539 of 10516 authors are related to this field, which corresponds to 5.1% of the top 20 fields. Remaining fields in top 20 are Computer Networks and Communications, Information Systems, Hardware and Architecture, Computer Vision and Pattern Recognition, Computer Graphics and Computer Aided Design.

To understand the content of these papers, another important field to look at, is the Author Keywords and Keywords Plus. When we analyse the most common keywords, we can see that "machine learning" is the top keyword used by authors to describe their papers. This goes in hand to hand with the recent developments in the technology and data science field and shows the huge interest of the research world in machine learning. Second place is held by "deep learning" which is again one of the topics on surge. Especially considering that these big tech companies are recently trying to improve their products in the field of virtual assistant such as Siri, Alexa, and Google Assistant, it is not a surprise that "speech recognition holds the 4th place in mostly used keywords. Some of the highest mentioned are related to the field of data science such as "neural networks", "deep learning" etc. However, there are also interesting keywords that are catching one's attention. These keywords are "social media", "security", "privacy" and "crowdsourcing" (Table 3). Analysis of additional keywords (Keywords Plus) supports the ideas introduced in the keywords. These are also mainly concentrated on algorithms, models, systems, networks, etc.

Table 3 – Top 10 of the keywords of indirectly supported papers

Keyword	Count
machine learning	857
deep learning	703
virtual reality	479
speech recognition	473
security	444
privacy	390
social meadia	380
neural networks	334
design	328
crowdsourcing	320

Considering that article titles are one sentence summaries of the papers, analysis of these titles also presents interesting insights about the content of the papers in addition to the insights from keywords. Looking at the mostly occurring trigrams and bigrams is helpful in this case. As mentioned in Section 4, titles were divided to decades to not only capture the contents, but also to capture how the contents have been changing throughout the decades. It is not surprising that most of the bigrams in Top 10, for last decade, include the terminology relating to Data Science, such as "neural networks", "deep learning", "machine learning" and "speech recognition". Although timespan between 2001 and 2011 shows also how some of these terms related to Data Science were becoming more popular, there were also terms related to the surge of internet and sensors such as wireless networks, web services and wireless sensors. Bigrams from the period before 2001 are also providing interesting insights about the first years of the Big Tech companies, such as "educational software", "software development" and "information technology". Trigrams are also following the similar pattern as bigrams in terms of giving more weight to machine learning, artificial intelligence terms in more recent decades and more weight on simulation models in earlier decades. Top 10 lists of trigrams and bigrams can be found in Appendix E.

Another interesting point is related to the authors that are mostly publishing papers from the conferences sponsored by Big Tech. First place in the number of papers belongs to Yang Liu with 110 publications (Figure 6). He is an Assistant Professor in University of California, and he states his goal as "building responsible machine learning tools with humans in the loop" and "developing fair and accountable machine learning treatments to better serve our society" according to his website. Wei Wang and Wen Gao are the second and third placed respectively with around 103 and 90 publications.

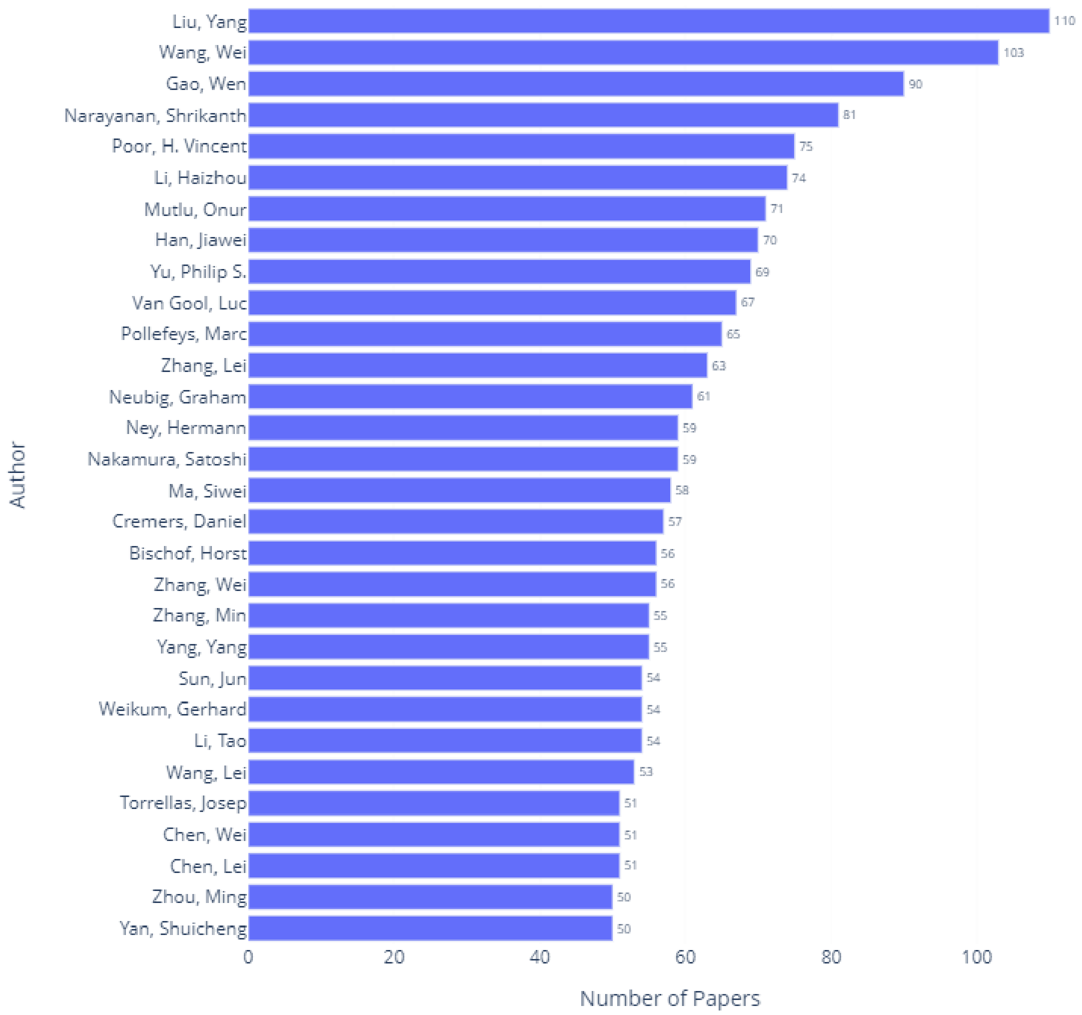


Figure 9 – Number of publications from the sponsored conferences per author (top 30)

One of the important things for understanding the geography of the research is the author addresses. In terms of the countries, 2495 of the authors are living in United States of America. Italy and Spain are sharing second and third places with 1973 and 1759 authors respectively. 6 out of 10 top countries for number of authors from Europe (Figure 7). Portugal is one of the countries catching attention at 8th place by passing countries like India and China.

Figure 10 – Number of authors with at least one paper from the conferences sponsored by Big Tech, per country

Although number of authors is an important metric to see for analysis, it might be biased since each country has different number of populations. Countries with higher number of populations might be advantageous. To better understand the geography of research and preserve the unbiasedness, analysis was done based on number of authors per person. Population data is taken accounted for 2020. This analysis shows that highest number of authors per person is in Portugal. While Luxembourg and Finland are the countries in second and third places, Ireland and Australia are closing top 5.

5. BIG TECH'S DIRECT SUPPORT: PAPERS ACKNOWLEDGING BIG TECH FUNDING

There have been 18396 academic papers published between the years of 1988 and 2021 that are funded by at least one of the big tech companies. Till 2006, number of the papers were changing between zero and four with the highest being in 1993. However, after 2006 there is a steady increase till 2020, pandemic year, only exception being in 2013 in which number of papers stayed almost at the same level as previous year (Figure 8). Highest percentage change in the number of publications is in 2006 with the boom and second-highest percentage change after 2006 is seen in 2009 with 88% of increase. Global maximum of the number of publications is observed in 2019 with 2351 papers.

Another important part is to see how each company funding importance changes from year to year. Data shows that Apple is the first among the Big Tech companies to directly fund a paper, first one being in 1988. However, number of papers are not very consistent till 2006, meaning, there have been no papers published in this 16-year period, only exceptions being in 1993 and 1997 with 4 and 1 papers respectively. There are also fluctuations in the number of papers between the period of 2006 and 2014, with highest number of papers being 10 (in 2010) and lowest number of papers being 3 (in 2007 and 2009). However, after 2014, there is a steady increase in the number of papers funded. It is interesting to see that pandemic did not affect the number of the papers funded by company and in the proceeding year highest number of papers were funded with 45 papers.

Second earliest mentioned company is Microsoft, with mentions in only one paper in 2003. Although first paper is in 2003, number of funded papers start to take off from 2006 and follows steady increase till 2011, reaching 538 papers in that year. In next 4 years, number of funded papers stays almost constant with 520,508,530 and 546 papers respectively. After 2015, there is a steady increase till 2018, which is also the global maximum with 882 papers acknowledging the funding by company. In following 3 years steady decrease can be observed.

Next two places are shared between Amazon and Google, both companies funding first paper in 2006 with 1 and 5 pages respectively. Although both companies funded their first papers in 2006, Amazon did not fund any paper in the next year. However, after 2008 number of funded papers steadily increased till 2019 reaching the global maximum in 2019, with 389 papers, only exception being in 2015, where decrease can be observed. In following 2 years number of funded papers by Amazon is almost steady with 383 and 385 papers. In Google's case, steady increase continued till 2016, reaching to 658 papers in that year. Although there is a big leap (almost two times increase, with 1185 papers) between 2016 and 2017, in next 3 years number of papers also grew steadily till 2020 and reached the global maximum in 2019 with 1240 funded papers. In next two years, 2020 and 2021, decrease can be observed having 1066 and 1018 papers in these years.

Fourth and fifth places are also shared by two companies, Facebook, and Netflix, both starting funding research in 2010. Number of papers funded by Netflix shows some fluctuations, meaning, some years there are no funded papers and there is no steady increase or decrease. However, in general view between 2010 and 2021 some increase can be observed. Most number of papers funded by Netflix was in 2021 with 9 papers. In Facebook's case, steady increase can be observed throughout the last 11 years, only exception being in 2020 with a minor decrease. As Netflix, Facebook also reached the highest number of funded papers was reached in 2021 with 272 papers (Appendix I).

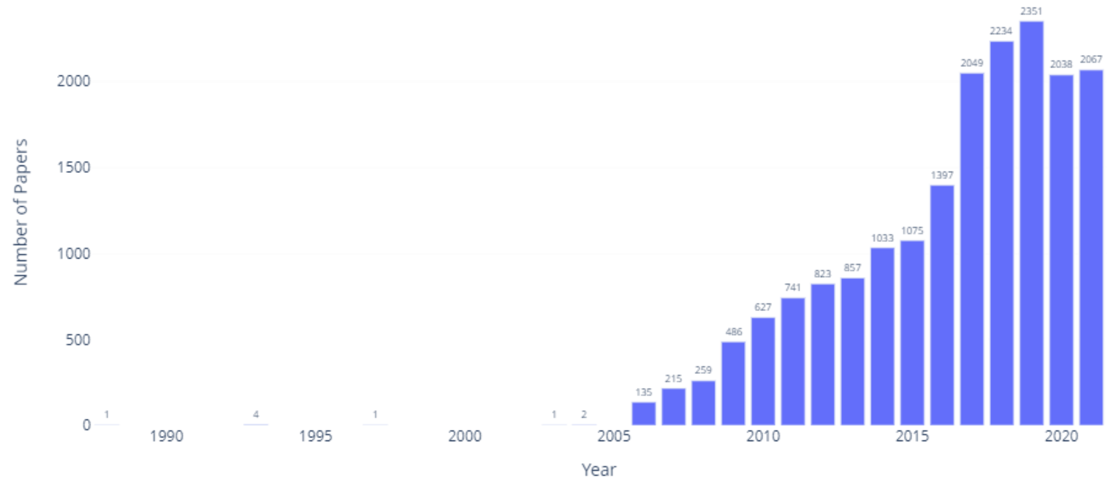


Figure 11 – Number of papers acknowledging direct Big Tech support over years

When we analyse number of papers funded by Big Tech, we can see that Google holds the first place with number of 8585 which corresponds to almost 50% of all papers produced. Second place is held by Microsoft with 8493 papers, however there is drastic difference between 2nd and third places, which is held by Amazon with 1987 papers funded (Table 4). Fourth and fifth places are shared by Facebook and Apple respectively.

Table 4 – Number of papers funded by each of Big Tech

Company	Count
Google	8585
Microsoft	8493
Amazon	1987
Facebook	1119
Apple	239
Netflix	26

Another interesting point while analysing direct support from Big Tech is about the overlapping between these companies. Co-occurrence matrix shows that the biggest overlap occurs between Google and Microsoft with 493 papers referring to only these companies as conference sponsors among Big Tech. Second largest overlap between two of Big Tech companies is between Google and Amazon with 198 papers. Furthermore, overlap between Google and Facebook is third largest

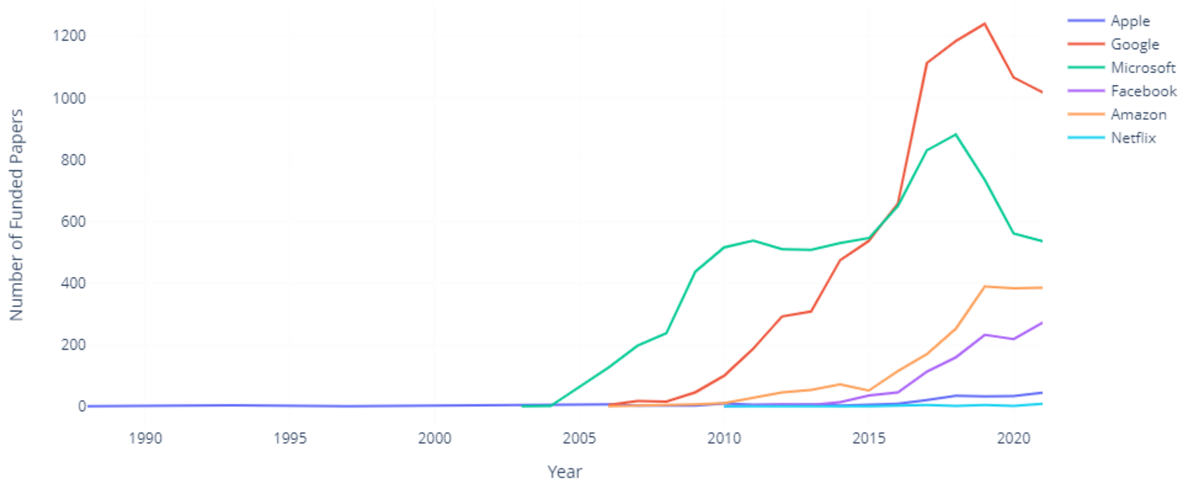


Figure 12 – Number of papers funded by each of Big Tech over time

overlap with 9 papers less than the latter one, 189 papers. When only bilateral (funded only by two of Big Tech companies) overlaps considered, Google is first company for all other Big Tech with only one exception being Netflix, which has only one bilateral overlap, with Facebook. Second biggest bilateral overlap for all the companies occurs with Amazon, except Apple. For Apple second biggest overlap occurs with Microsoft. Another interesting point is that Facebook and Apple have no funded papers in common, meaning including only these two companies as funding companies among Big Tech. Full co-occurrence matrix can be found in Appendix F.

Big tech has funded different publication types of research throughout the time such as journals, conference proceedings, series and even some books. Among these publication types, these companies mostly invested in journals, with 10108 papers, which corresponds to 55% of all the papers. Second highest number of publications relates to conference proceedings, which is not very surprising when considered the number of conferences sponsored by these companies. Third place is series with 16 papers. In terms of publishers, IEEE is the biggest publisher of these papers with the number of 4550 papers in total. Second and third places are held by Association for Computing Machinery (ACM) and Springer with 3264 and 1733 papers accordingly. Other notable publishers are Elsevier and Wiley. Analysing the publisher cities shows that New York is the central city for these publishers with around 30% of the papers related to it. Another city from United States of America, Los Alamitos is the second largest publisher city with 1218 papers in its name. Third place is also held by another US city, Piscataway, with 914 papers. Two European cities, London and Amsterdam are sharing the fourth and fifth places respectively. While top cities mostly consist of the cities from the US, some of the notable publisher cities in Europe are Oxford, Berlin, and Dordrecht. Meanwhile there are no other city from the continents other than Northern America and Europe in top 20.

Another important point is to see the countries of the authors producing research funded by Facebook, Amazon, Apple, Microsoft, Netflix and Google. There have been 6101 authors from 103 countries, whom research was funded by the Big Tech. Most of these countries are situated in Europe (38 countries), while 31 of them are in Asia. Rest of the countries are shared between Africa, South

America, North America, and Oceania. When compared by the number of authors, unsurprisingly, USA holds the first place by far with 1572 authors involved in the research from the country, which corresponds to almost 26 percent of all authors (Figure 10). Great Britain comes second with 649 authors, and the third is Brazil with 399 authors. Furthermore, six out of ten top countries are from Europe, which are Great Britain, Italy, Spain, Germany, France, and Netherlands. Rest of the top ten countries are Australia and China. Surprisingly, China is holding the 8th place. Even though geography of authors spans various number of countries and continents, produced research is in English, except only one paper being in Spanish.

Figure 13 – Number of authors with at least one paper funded by Big Tech, per country

To have a better view of the distribution of number of authors across the world and to eliminate the bias caused by number of populations, number of authors were also analyzed by the number of populations. To be more precise, number of authors were divided by the population to find the number of authors per person. When analyzed from this perspective, Saint Kitts and Nevis holds the first place, followed by Australia and Denmark. Portugal holds fourth place followed by Ireland in the fifth place. Last five place is held by Nigeria, Sudan, Ethiopia, Vietnam, and Philippines.

Besides the geography of the authorship, another important point is to understand the content of the papers produced with the funding of big tech. As mentioned before, categories that are supplied by Web of Science is a good and precise way for understanding outputs of these papers. Similar to the papers in Indirect Support section, very big part of these papers is related to Computer Science. Category with the greatest number of academic papers is "Computer Science, Theory & Methods" with 4655 papers and second highest category is "Computer Science, Artificial Intelligence". While 3rd highest category is "Engineering, Electrical and Electronic", all other categories in top 5 are also related to Computer Science: Information Systems and Software Engineering. While most of the top categories are related to engineering and computer science, category of "Ecology" is at the 24th place with 250 papers related to it.

Besides Web of Science categories, Author Keywords and Keyword Plus fields are also providing good insights about the contents of the research produced. When keywords are analyzed, it can be seen that Data Science related terms are holding first and second places, which are “machine learning” and “deep learning” with 319 and 303 number of papers related to the terms. In addition, “cloud computing”, “algorithms” and “performance” are following these two terms. Other than the terms related to technical subjects, some of the top interesting keywords are “social media”, “crowdsourcing”, “security” and “privacy”. However, there are big difference between the papers with technical terms related to data science and the terms related to security or privacy side of these innovations. Term “privacy” is at 10th place with 85 papers and “security” at 7th place with 101 papers related to it (Table 5). Analysis of the Keywords Plus also shows the concentration of the companies more on the technical side rather than more humanitarian side. Some of the highest used additional keywords are “model”, “algorithm”, “classification”, “recognition” etc.

Table 5 – Top 10 keywords

Keyword	Count
machine learning	319
deep learning	303
cloud computing	230
algorithms	203
performance	172
design	148
security	101
social media	89
crowdsourcing	88
privacy	85

To have a closer look of the content of the papers identification of the mostly occurring trigrams and bigrams is helpful. As mentioned before, titles were divided to decades to not only capture the contents, but also to capture how the contents have been changing throughout the decades. It is again not of a surprise that majority of the bigrams in top 10, for past decade, includes the phrases related to machine learning and artificial intelligence, such as “neural networks”, “deep learning”, “machine learning” and “speech recognition”. However, it is interesting to see that no bigrams related to these fields are appearing in the top 10 of previous decade. Period spanning the years between 2001 and 2011 shows the terms related to the surge of internet and sensors such as wireless networks, web services and wireless sensors and social networks. Trigrams are also following the similar pattern as bigrams in terms of giving more weight to machine learning, artificial intelligence terms in most recent decade. For example, convolutional neural networks, deep neural networks and recurrent neural network are one of the mostly used trigrams. Top 10 lists of trigrams and bigrams can be found in Appendix G.

Next interesting point of analysis is about the authors that are producing research in a collaboration with Big Tech companies.

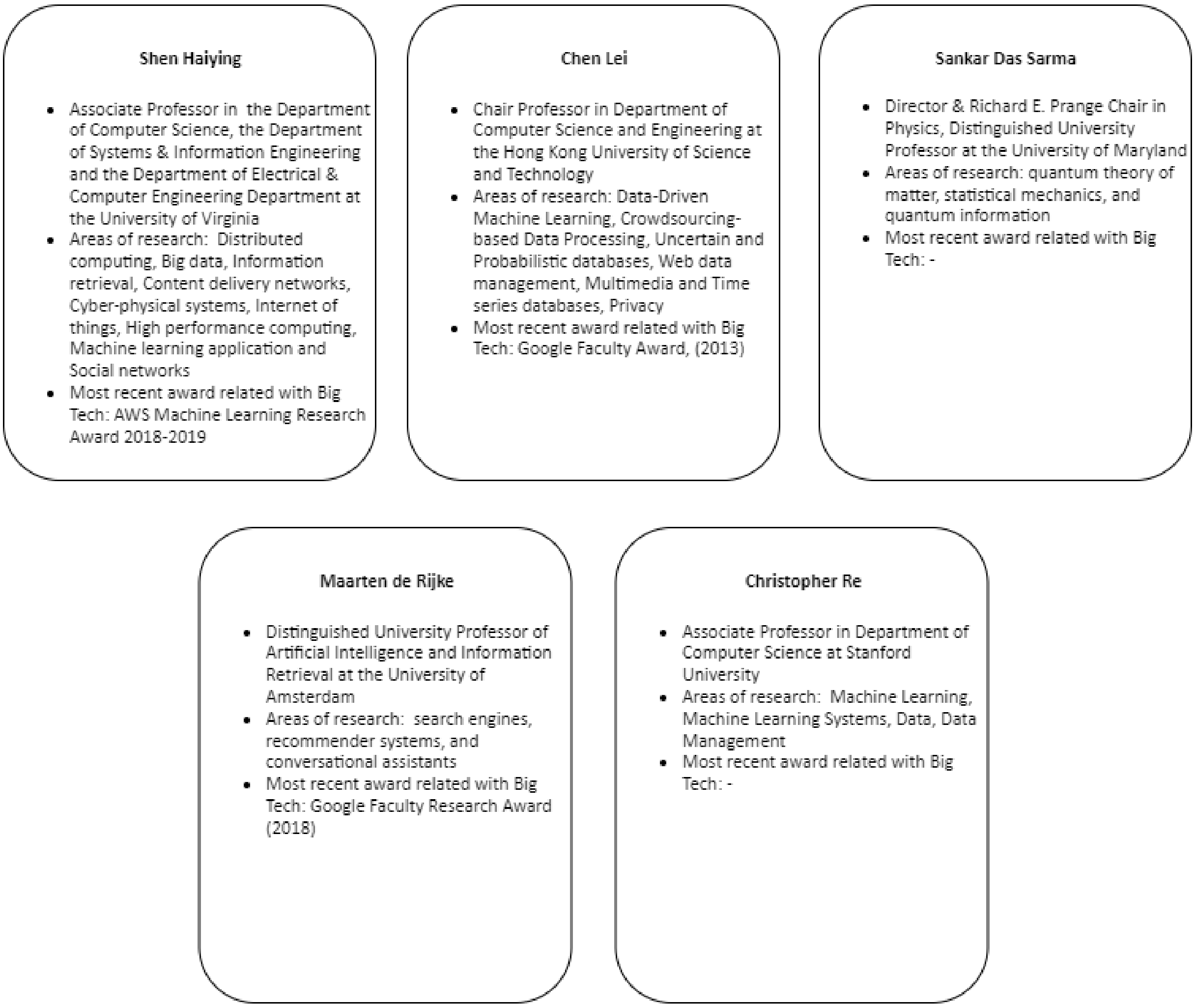


Figure 14 – Overview of the top 5 authors (in terms of numbers of paper funded by Big Tech)

Professor Shen Haiying is the top author in terms of taking part in the research funded with number of 159 papers (Figure 11). According to her web page, her main areas research are Distributed Systems and Networks, Cloud/Edge Computing, Distributed Machine Learning, Machine Learning Applications, Cyber-Physical Systems, and Smart City/Building. Second author with the highest number of papers is Chen Lei. He is affiliated to Hong Kong University of Science and Technology. Sankar Das Sarma holds third place with 90 papers in his name.

Furthermore, when subject areas of authors are analyzed, it can be seen that "Biochemistry, Genetics and Molecular Biology (all)" is the top subject area with 137 authors. "Computer Science Applications" is in the second place just with five less authors corresponding to 132 authors. "Multidisciplinary", "Theoretical Computer Science", and "Electrical and Electronic Engineering" complete the top five subject areas with 128, 104 and 98 authors respectively (Table 6). While the rest of the top 20 subject categories mostly correspond to the areas related to Computer Science and Engineering, some of the interesting subject areas in the top 20 are “Molecular Biology”, “Human-Computer Interaction”, and “Medicine”.

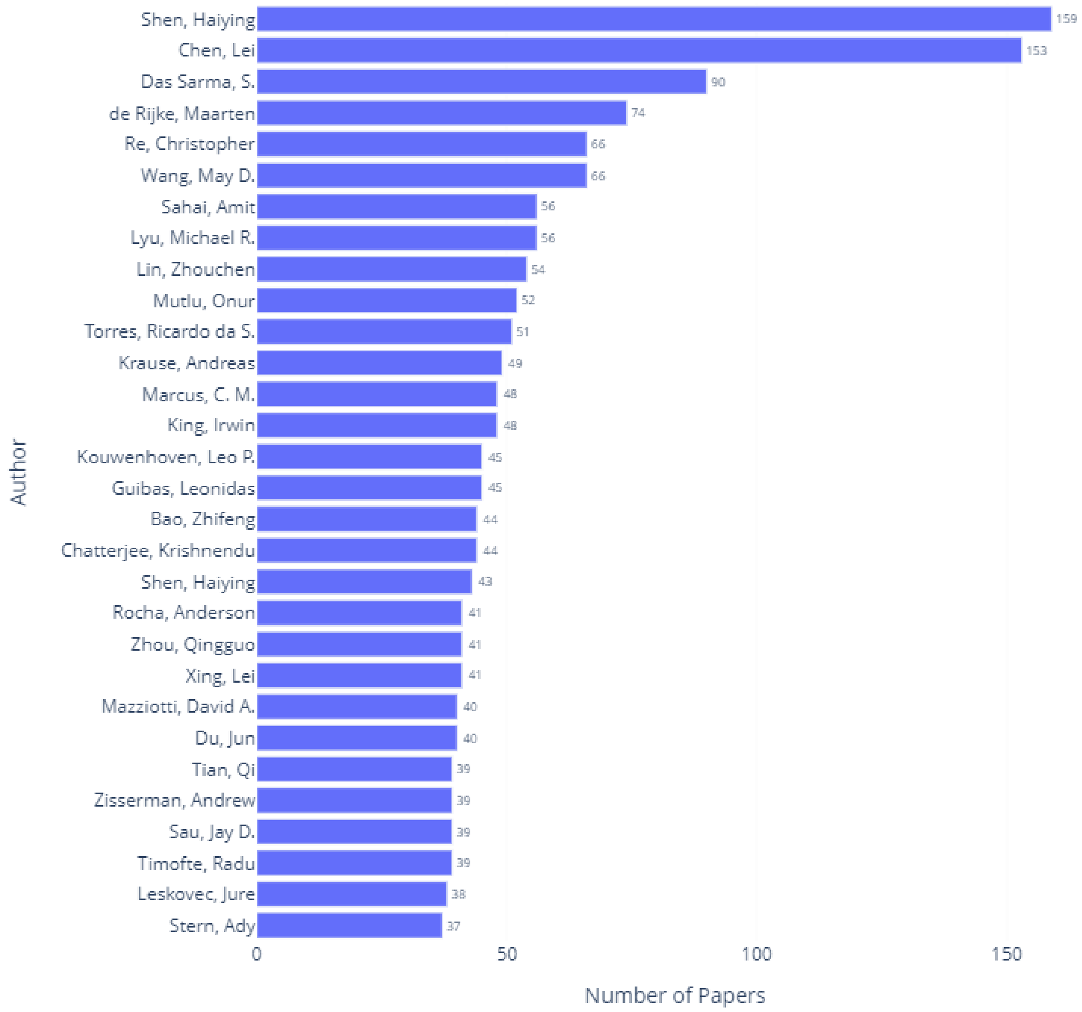


Figure 15 – Number of papers funded by Big Tech per author (top 30)

Table 6 – Scientific fields of authors that are directly funded

Subject	Country
Biochemistry, Genetics and Molecular Biology	137
Computer Science Applications	132
Multidisciplinary	128
Theoretical Computer Science	104
Electrical and Electronic Engineering	98
Artificial Intelligence	96
Software	96
Computer Science (all)	94
Engineering (all)	92

Information Systems	89
Medicine (all)	86
Human-Computer Interaction	86
Hardware and Architecture	82
Computer Vision and Pattern Recognition	81
Computer Networks and Communications	81
Physics and Astronomy (all)	80
Applied Mathematics	78
Computational Theory and Mathematics	77
Molecular Biology	76
Control and Systems Engineering	74

6. DISCUSSION OF RESULTS

In this section results from sections 5 and 6 are discussed by drawing parallels between indirect and direct support. Differentiating points between the two will be also discussed to identify the similarities and differences between outcomes of different types of support. If we think that companies most of the times directly support the papers that they consider as essential things for their company and conference support as a bit more marketing technique and improving company's image in the eye of stakeholders, results can be of interest: a) what public knowledge environments they choose sponsor and promote (indirect support), and b) what research actors they want to see working given topics (direct support).

In terms of timespan of both types of financial support, it is observed that Big Tech started to sponsor the conferences earlier than directly funding the research. Direct funding is more notable after 2005 whereas the boom for conference sponsorships started from late 1990s, early 2000s. Another point that should be noted is the effect of pandemic to these support types. Namely, drastic decrease is observed in the number of papers that are outcomes of the sponsored conferences whereas in direct funding, decrease is not as big as the first one. This can be explained by the cancellations of conferences during pandemic.

Another relevant point is which company gives higher importance to direct funding or indirect funding. When we look at the rankings of Big Tech companies in both types of support, some differences can be noted. While top sponsor for conferences is Microsoft, Google is the top funder in direct funding. Amazon and Facebook also change places, to be more precise, Amazon is third biggest direct funder, while Facebook is the third largest sponsor in conferences. However, Apple and Netflix stay in the same rank in both types of support. In terms of bilateral co-occurrences, Google and Microsoft are the mostly co-occurring companies in both types of support. One interesting difference is that Amazon is the largest co-occurring company with Microsoft and Google in directly supported papers, while in indirect support Facebook takes the second place. Also, in indirect funding, Facebook and Amazon appear together less than they do in directly funded papers.

As expected, type of publications mainly consists of conference proceedings for indirect support. While there are some conference proceedings in direct funding too, mostly they consist of the journals. Biggest publishers of both types of papers are IEEE, Springer, and ACM. Also, in terms of publisher cities, in indirect support there is one European city, Berlin, in top 3, while all of the publishers in top 3 of directly funded papers are from USA. Furthermore, language of conference papers has more variation than funded papers. Namely, funded papers are strictly written in English. This shows the preference of companies in English language as the language of science.

Another comparison can be drawn about the contents of the two types of supports. When WoS categories of each type is compared, it can be seen that both types mostly concentrate on technical sides of innovations than the consequences of these innovations for society and our world. That said, in both lists of top 30 for categories, social issues and environmental issues take 25th place. Interesting difference here is that while direct funding focuses on environmental issues, indirect support focuses more on social issues. In terms of timeline, environmental issues start appearing after 2008, making a drastic increase in 2014 and reaching to highest number in 2020. On the other hand, social issues in indirect support starts from one year earlier than environmental issues, namely in 2007. While number of papers categorized as social issues reaches highest point in 2009, there is a drastic decrease in the next year and there are no papers related to category afterwards, only exception being in 2016 with 100 papers published that was categorized as social issues. Same thing applies to the fields of authors that produce scientific papers through these supports. Most of the authors are on the fields that are related computer science. One interesting distinction between the author fields is that direct support also involves authors on biochemistry, medicine, physics, and astronomy besides computer sciences, while indirect support involves mostly the authors in the field of computer science. This shows that while in direct support environmental issues are discussed by the authors that are in fields of biochemistry, physics and astronomy, social issues in indirect support are discussed by mostly authors that are related to computer science.

Comparison of the keywords suggests one more interesting distinction between two types of support. While security and privacy are more recurrent for the papers indirectly supported, it is less recurrent for directly supported papers. While these terms start to appear in indirectly supported papers in early 2000s and reaches the highest point in 2019, in directly funded papers these terms start appearing at 2006 and seeing the peak point at 2017 (security) and 2018 (privacy). As it was mentioned in first paragraph, one can see this difference as how Big Tech wants to be perceived vs what they want to really focus on. That said, Big Tech wants to be perceived as they are working on improving security and privacy, in real side they concentrate on technical sides of the innovations. One similar point about the keywords is that machine learning, neural networks and deep learning are the mostly used keywords in both types of supports. However, there are some keywords that are related to the computer science field but does not appear in direct support as frequent as they appear in indirect support. These are speech recognition and virtual reality. Instead, “algorithms” keyword appears more frequent in directly supported papers. The list of most frequent trigrams from the titles of both types of support also supports the idea that Big Tech mainly focuses on the technical side of the innovations more than the ethical sides.

7. CONCLUSIONS

This work goes into a set of paramount influencing-wielding organizations in the global informational economy. It provides a perspective on science governance by focusing on rise of Big Tech in the realm of research support. The institutions of science and the practice of research has in the past shown vulnerabilities from Big Business, whether from food, drugs, or energy sectors. A precautionary concern suggests itself regarding the most recent wave of corporate might, one so unique that is shaping the entire mode of production of the contemporary economy as a whole. Accusations of "having too much power" were bipartisan in the US Congress when the chiefs of platform industries were summed regarding concerns that they were "squelching competition, creativity and innovation." (Financial Times, 39 July 2020, p.2)

It has been noticed that the largest technology groups have been "pouring money" into the knowledge and policy ecosystem: their donations for think-tanks in the US more than doubled from 2017-2018 to 2019-2020 (perhaps reaching close to $3mn) and place them with Oil & Gas as top donors (Financial Times, 2 February 2022, p.6). Tech companies have building presence for years and became focused on influencing elites. Now the time has come to know about the Big Tech and Modern Science as bedfellows. With valuation over 1tn dollars, non-corporates cannot match the resources Big Tech can put in the hands of collaborators. How will this chance the rate and direction of research, the narrowing and instrumentalization of agendas is still unknown.

In this thesis we focus on conference sponsorships and funding acknowledgements as indicators of Big Tech support to research activities. We find that Big Tech are paramount players: their presence is substantively felt from the mid-2000s in terms of indirect support and from the mid-2010s in terms of direct support. Computer science topics is the main focus of this support, rather than reflexive themes (like the socio-political impacts of platformisation, the psychological effects of social media on children, or the environmental cost of large database exploitation) although security and private challenges has appeared as of late.

A number of policy and regulatory implications suggest themselves. For instance, Big Tech could be compelled to conform to strengthened disclosure rules. Moreover, beneficiaries must declare their business ties. Research actors should find ways to compete for other types of support and civil society scientific bodies should avoid Big Tech dependence.

REFERENCES

Abdalla, M., & Abdalla, M. (2021). The Grey Hoodie Project: Big Tobacco, big tech, and the threat on academic integrity. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3461702.3462563

Bar-Ilan, J. (2010). Web of science with the conference proceedings citation indexes: The case of computer science. Scientometrics, 83(3), 809–824. https://doi.org/10.1007/s11192-009-0145-4

Behrens, T. R., & Gray, D. O. (2001). Unintended consequences of cooperative research: Impact of industry sponsorship on climate for academic freedom and other graduate student outcome. Research Policy, 30(2), 179–199. https://doi.org/10.1016/s0048-7333(99)00112-2

Broadus, R. N. (1987). Toward a definition of "bibliometrics." *Scientometrics*, *12*(5-6), 373–379. https://doi.org/10.1007/bf02016680

Broomfield, CO: One Earth Future Foundation. http://acuns.org/wp-content/uploads/2013/11/gg-weiss.pdf

Castellacci, F., Grodal, S., Mendonca, S. & Wibe, M. (2005). Advances and challenges in innovation studies. Journal of Economic Issues, 39, 91-121.

Castells, M. (1996). The Information Age: Economy, Society, and Culture. Vol. 1: The Rise of the Network Society, London: John Wiley and Sons.

Confraria, H., Ferreira, V.H. and M.M. Godinho (2021). Emerging 21st Century technologies: Is Europe still falling behind?. REM Working Paper 0188-2021.

Conway, E. M., & Oreskes, N. (2014). Merchants of doubt. Bloomsbury Publishing.

Costa, C. M., & Mendonça, S. (2019). Knowledge-intensive consumer services. understanding KICS in the innovative global health-care sector. *Research Policy*, *48*(4), 968–982. https://doi.org/10.1016/j.respol.2018.10.026

Costa, C. M., Quintanilha, T. L., & Mendonça, S. (2019). Manuel Castells and informationalism 1. In *Management, Organizations and Contemporary Social Theory* (1st ed.). essay, Routledge.

Cowgill, B., Prat, A. and Valletti, T. (2021). Political power and market power. Working Paper, June. https://arxiv.org/pdf/2106.13612.pdf

David, P. A. (2004). Understanding the emergence of 'open science' institutions: Functionalist economics in historical context. Industrial and Corporate Change, 13(4), 571–589. https://doi.org/10.1093/icc/dth023

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Dutton, W.H. (ed.) (2013). The Oxford Handbook of Internet Studies. Oxford: Oxford University Press.

Epstein, D., Boden, R., & Kenway, J. (2007). Winning and managing research funding (academic's support kit). SAGE Publications.

Freeman, C. & Louçã, F. (2001). As Time Goes By: From the Industrial Revolutions to the Information Revolution. Oxford: Oxford University Press.

Gambardella, A., Heaton, S. Novelli, E. & Teece, D.J. (2021). Profiting from enabling technologies?. Strategy Science, 6(1), 75-90.

Geroski, P. (2003). The Evolution of New Markets. Oxford: Oxford University Press.

Geuna, A. (1999), The Economics of Knowledge Production: Funding and the Structure of University Research, Cheltenham: Edward Elgar.

Geuna, A., & Martin, B. R. (2003). University Research Evaluation and funding: Na international comparison. Minerva, 41(4), 277–304. https://doi.org/10.1023/b:mine.0000005155.70870.bd

Glänzel, W., Moed, H.F., Schmoch, U. & Thelwall, M. (eds) (2019), Springer Handbook of Science and Technology Indicators. Berlin: Springer.

Irwin-hunt, A. (2020). Global outlook R&D. FDI Intelligence, April/May, pp. 24-27.

Information for authors. (2017). IEEE Transactions on Aerospace and Electronic Systems, 53(2), 1065–1066. https://doi.org/10.1109/taes.2017.2696918

Jacques, P. J., Dunlap, R. E., & Freeman, M. (2008). The organisation of denial: Conservative think tanks and environmental scepticism. Environmental Politics, 17(3), 349–385. https://doi.org/10.1080/09644010802055576

Jefferson, T. (2020), "Sponsorship bias in clinical trials: growing menace or dawning realisation?", Journal of the Royal Society of Medicine, Vol. 113, No. 4, 148-157.

Keane, P. (2020, September 20). How the oil industry made us doubt climate change. BBC. Retrieved May 15, 2022, from https://www.bbc.com/news/stories-53640382.

Klinger, J., Mateos-Garcia, J.C., & Stathoulopoulos, K. (2020), A narrowing of AI research?, Available at SSRN 3698698.

Lévêque, F. (2021). Les Entreprises Hyperpuissantes: Géants et Titans, La Fin du Modèle Global?. Paris: Odile Jacob.

Louçã, F. & Mendonça, S. (2002). Steady change: The 200 largest US manufacturing firms throughout the 20th century. Industrial and Corporate Change, 11 (4), 817-845.

Lyall, C., Bruce, A., Marsden, W., & Meagher, L. (2013). The role of funding agencies in creating interdisciplinary knowledge. Science and Public Policy, 40(1), 62–71. https://doi.org/10.1093/scipol/scs121

Mansell, R., Avgerou, C., Silverstone, R. & Quah, D. (2007). The Oxford Handbook of Information and Communication Technologies. Oxford: Oxford University Press.

Mansell, R. & Steinmueller, W.E. (2020). Advanced Introduction to Platform Economics. Cheltenham: Edward Elgar.

Mansell, R. (2021). Adjusting to the digital: Societal outcomes and consequences. Research Policy, 50 (9), 1-10.

Martin, B. R. (2003). The Changing Social Contract for Science and the evolution of the University. Science and Innovation, 7–29. https://doi.org/10.4337/9781781950241.00011

Martin, B. (2019). The future of science policy and innovation studies: Some challenges and the factors underlying them. In Simon, D., Kuhlmann, S., Stamm J. & Canzler, W. (eds), Handbook on Science and Public Policy. Chelteham: Edward Elgar, 523-542.

Marty, F. (2021). Hyperpowerful companies: Giants and Titans, the end of the global model, François Lévêque. Concurrences, N° 3-2021, Art. N° 101235.

Mejia, C., & Kajikawa, Y. (2017). Using acknowledgement data to characterize funding organizations by the types of research sponsored: The case of robotics research. Scientometrics, 114(3), 883–904. https://doi.org/10.1007/s11192-017-2617-2

Mendonça, S. (2006). The revolution within: ICT and the shifting knowledge base of the world's largest companies. *Economics of Innovation and New Technology*, *15*(8), 777–799. https://doi.org/10.1080/10438590500510442

Mendonça S. (2012). Trademarks as a telecommunications indicator for industrial analysis and policy. In Hadjiantonis A.M., & Stiller, B. (eds), Telecommunication Economics. Lecture Notes in Computer Science, Vol. 7216. Heidelberg: Springer, 33-41.

Mendonça, S., Pereira, T.S. & Godinho, M.M. (2004). Trademarks as an indicator of innovation and industrial change. Research Policy, 33 (9), 1385-404.

Mendonça, S., Schmoch, U. & Neuhäusle, P. (2019). Interplay of patents and trademarks as tools in economic competition. In Glänzel, W., Moed, H.F., Schmoch, U. & Thelwall, M. (eds), Springer Handbook of Science and Technology Indicators, Berlin: Springer, 1023-35.

Mendonça, S., Damásio, B., de Freitas, L.C., Oliveira, L., Cichy, M., and Nicita, A. (2022). The rise of 5G technologies and systems: A quantitative analysis of knowledge production. Telecommunications Policy, forthcoming.

Morillo, F. (2016). Public–private interactions reflected through the funding acknowledgements. Scientometrics, 108(3), 1193–1204. https://doi.org/10.1007/s11192-016-2032-0

Nelson, R.R., Dosi, G., Helfat, C.E. & Winter, S.G. (2018). Modern Evolutionary Economics: An Overview. Cambridge University Press.

O'Neill, A. (2022, January 4). Gross domestic product (GDP) ranking by country 2020. Statista. Retrieved May 14, 2022, from https://www.statista.com/statistics/268173/countries-with-the-largest-gross-domestic-product-gdp/

Oremus, W. (2017, November 17). The rise of the phrase "big tech" should worry Silicon Valley. Slate Magazine. Retrieved May 14, 2022, from https://slate.com/technology/2017/11/how-silicon-valley-became-big-tech.html

Paredes, A., Mendonça, J., Bação, F., & Damásio, B. (2022). Does R&D tax credit impact firm behaviour? Micro evidence for Portugal. Research Evaluation, forthcoming.

Philippon T., (2020), The Great Reversal: How America Gave Up on Free Markets, Belknap Press.

Righi, R., Samoili, S., Cobo, M.L., Baillet, M.V., Cardona, M., Prato, G. (2020). The AI techno-economic complex System: Worldwide landscape, thematic subdomains and technological collaborations. Telecommunications Policy, 44 (6), 101943.

Rikap, C., & Lundvall, B. Å. B. (2022). Big tech, knowledge predation and the implications for development. Innovation and Development. https://doi.org/10.1080/2157930X.2020.1855825

Rivlin, G. (1995). Guide to organizing an international scientific conference. https://doi.org/10.1159/isbn.978-3-318-04737-0

Rose, M. E., & Kitchin, J. R. (2019). Pybliometrics: Scriptable Bibliometrics using a python interface to scopus. SoftwareX, 10, 100263. https://doi.org/10.1016/j.softx.2019.100263

Sullivan, L. (2020, September 11). New Plastics Economy Archives - Wrong Kind of Green. Wrong Kind of Green RSS. Retrieved May 15, 2022, from https://www.wrongkindofgreen.org/tag/new-plastics-economy/

Taalbi, J. (2020). Evolution and structure of technological systems - An innovation output network. Research Policy, 49 (8), 104010.

Tan, A., Mui, H., & Terrace, K. (2000). Text Mining: The state of the art and the challenges.

Tunzelmann, N. (2003). Historical coevolution of governance and technology in the industrial revolutions. Structural Change and Economic Dynamics, 14 (4), 365-384.

US Senate (2019). Report of the Select Committee on Intelligence United States Senate Covering the Period January 3, 2017 to January 3, 2019 (March 28, 2019). Washington, D.C.: US Senate.

Wallin, J. A. (2005). Bibliometric methods: Pitfalls and possibilities. Basic and Clinical Pharmacology and Toxicology, 97(5), 261–275.

Wedel, J. (2014). Unaccountable: How Elite Power Brokers Corrupt Our Finances, Freedom, and Security. New York: Pegasus Books.

Wedel, J., Hussain, N., and Dolan, D.A. (2017). Political Rigging: A primer on political capture and influence in the 21st century, Oxfam America

Weiss, Thomas G., Conor D. Seyle, and Kelsey Coolidge. 2013. “The Rise of Non-State Actors in Global Governance: Opportunities and Limitations.” One Earth Future Discussion Paper.

Wells, G., McMillan, R., & Volz, D. (2019). Senate Faults Google, Other Tech Giants for Role in Russian Election Meddling. The Wall Street Journal. https://doi.org/https://www.wsj.com/articles/senate-report-seeks-to-curb-possible-russian-influence-in-2020-election-11570556521

World Health Organization (WHO). (2019). Tobacco industry: decades of deception and duplicity. Retrieved May 14, 2022, from https://applications.emro.who.int/docs/FS-TFI-198-2019-EN.pdf?ua=1

Wu, T. (2017). The Attention Merchants: The Epic Scramble To Get Inside Our Heads. New York: Knopf.

Wu, T. (2018). The curse of bigness: Antitrust in the new gilded age. Columbia Global Reports.

Zuboof, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs.

APPENDIX

A: Sponsor names that were discarded from the sponsorship papers

‘Green Apple Data Ctr’

‘NEW ZEALAND APPLE & PEAR MKT BOARD’

‘Cafe Amazon’

‘Inst Nacl Pesquisas Amazonia’

‘Minist Environm Water Resources & Legal Amazon’

‘Univ Amazonas’

‘Daegu Kyungbook Apple Growers Agr Cooperat Union’

‘Daegu Kyungbook Apple Grower Agr’

B:

Combinations with Apple:

'fruit','food','genome','agricultural','agriculture','farm','peach','flavor','growers','picker','tree','breed',’ orchard','china','agro','juice','plant','agricultual', 'law experience', 'medical corporation', 'pear', 'berry', 'loess', '(anr)', 'rural', 'michigan', 'horticult', 'linnean', '(ardp)', 'pools', 'products research and education council', 'thurgood', 'adelphi', 'beatrice apple', 'baosteel', 'association of apple producers', 'california apple commission', 'apple industry', 'arthur boller', 'ny apple research and development program', 'fondazione cassa di risparmio', 'fondo unico progetti', 'idaho apple commission', 'dwarf apple', 'apple cultivars', 'modern apple industry technology', 'apple products', 'mr apple', 'ny apple rd', 'ny state apple', 'new york apple research', 'new york state apple research', 'jude apple', 'apple in postharvest', 'pennsylvania apple program', 'pennsylvania apple marketing board', 'shaanxi', 'apple proliferation', 'liaoning', 'south tyrolean apple consortium', 'south african apple', 'apple association', 'apple slices', 'apple rootstock', 'apple commission', 'yangling', 'apple pests', 'apple as a model for genomic information exchange', 'david j. apple', 'laboratory of apple resources innovation and genetic improvement', 'fertilizer', 'early apple wilt disease', 'virginia apple research program', 'wine', 'william s. apple', 'national apple'

Combinations with Amazon:

'brazil', 'amazonas', 'amazon region', 'amazonia', 'western amazon', 'amazon basin conservation association (acca)', 'amazon conservation', 'amazon cooperation', 'amazonica', 'amazon forest', 'amazon fund', 'amazon initiative', 'amazon malaria initiative', 'amazon research foundation', 'amazon river', 'amazon state research foundation', 'fapeam', 'amazon state', 'amazonandes', 'federal university of amazon', 'amazonica', 'peruvian amazon', 'amazon biodiversity program', 'capes', 'climamazon', 'environmental changes in the amazon', 'eastern amazon', 'amazon biota', 'amazon fish', 'amazon basin', 'iatecam', 'asica', 'fapespa', 'fapes', 'university of amazon', 'gordon and betty', 'forest', 'wild', 'adapta', 'ipeam', 'iesam', 'itegam', 'icaa', 'usaid', 'amazon oil', 'land use in the amazon', 'margaret mee', 'national amazon research institute', 'national science foundation', 'natural environment research council', 'proamazon', 'ufam', 'satellite monitoring', 'state of the amazon', 'university of the amazon', 'universidad', 'lower amazon', 'in the amazon', 'green amazon', 'wood', 'oils of the amazon'

C: Stop Words used for text cleaning

make,btw,a,as,able,about,above,according,accordingly,across,actually,after,afterwards,again,against,aint,aint,all,allow,allows,almost,alone,along,already,also,although,always,am,among,amongst,an,and,another,any,anybody,anyhow,anyone,anything,anyway,anyways,anywhere,apart,appear,appreciate,appropriate,are,arent,arent,around,as,aside,ask,asking,associated,at,available,away,awfully,b,be,became,because,become,becomes,becoming,been,before,beforehand,behind,being,believe,below,beside,besides,best,better,between,beyond,both,brief,but,by,c,ca,cmon,cmon,cs,came,can,cant,cant,cannot,cant,cause,causes,certain,certainly,changes,clearly,co,com,come,comes,concerning,consequently,consider,considering,contain,containing,contains,corresponding,could,couldnt,couldnt,course,currently,d,definitely,described,despite,did,didnt,didnt,different,do,does,doesnt,doesnt,doing,don,dont,dont,done,down,downwards,during,due,e,each,edu,eg,eight,either,else,elsewhere,enough,entirely,especially,essentially,et,etc,even,ever,every,everybody,everyone,everything,everywhere,ex,exactly,example,except,f,far,few,feel,fifth,finally,first,five,followed,following,follows,for,former,formerly,forth,four,from,further,furthermore,g,get,gets,getting,given,gives,go,goes,going,gone,got,gotten,greetings,h,had,hadnt,hadnt,happens,hardly,has,hasnt,hasnt,have,havent,havent,having,he,hes,hes,hello,help,hence,her,here,heres,heres,hereafter,hereby,herein,hereupon,hers,herself,hi,him,himself,his,hither,hopefully,how,howbeit,however,i,id,id,ill,ill,im,ive,ive,ie,if,ignored,immediate,in,inasmuch,inc,indeed,indicate,indicated,indicates,inner,insofar,instead,into,inward,is,isnt,isnt,it,itd,itd,itll,itll,its,its,its,itself,j,just,k,keep,keeps,kept,know,knows,known,l,last,lately,later,latter,latterly,least,less,lest,let,lets,lets,likely,little,literally,look,looking,looks,ltd,m,mainly,many,may,maybe,me,mean,meanwhile,merely,might,more,moreover,most,mostly,much,must,my,myself,n,name,namely,nd,near,nearly,necessary,need,needs,neither,never,nevertheless,new,next,nine,no,nobody,non,none,noone,nor,normally,not,nothing,novel,now,nowhere,o,obviously,of,off,often,oh,ok,okay,old,on,once,one,ones,only,onto,or,other,others,otherwise,ought,our,ours,ourselves,out,outside,over,overall,own,p,particular,particularly,per,perhaps,placed,please,plus,possible,presumably,probably,provides,q,que,quite,qv,r,rather,rd,re,really,reasonably,regarding,regardless,regards,relatively,respectively,right,s,said,same,saw,say,saying,says,second,secondly,see,seeing,seem,seemed,seeming,seems,seen,self,selves,sensible,sent,serious,seriously,seven,several,shall,she,should,shouldnt,since,six,so,some,somebody,somehow,someone,something,sometime,sometimes,somewhat,somewhere,soon,sorry,specified,specify,specifying,still,sub,such,sup,sure,t,ts,take,taken,tell,tends,th,than,thank,thanks,thanx,that,thats,thats,the,their,theirs,them,themselves,then,thence,there,theres,theres,thereafter,thereby,therefore,therein,theres,thereupon,these,they,theyd,theyd,theyll,theyll,theyre,theyre,theyve,theyve,think,third,this,thorough,thoroughly,those,though,three,through,throughout,thru,thus,to,together,too,took,toward,towards,tried,tries,truly,try,trying,twice,two,u,un,under,unfortunately,unless,unlikely,until,unto,up,upon,us,use,used,useful,uses,using,usually,uucp,v,value,various,very,via,viz,vs,w,want,wants,was,wasnt,wasnt,way,we,wed,wed,well,weve,well,were,weve,welcome,well,went,were,werent,werent,what,whats,whats,whatever,when,whence,whenever,where,wheres,wheres,whereafter,whereas,whereby,wherein,whereupon,wherever,whether,which,while,whither,who,whos,whos,whoever,whole,whom,whose,why,will,willing,wish,with,within,without,wo,wont,wont,wonder,would,would,wouldnt,wouldnt,x,y,yes,yet,you,youd,youd,youll,youre,youve,youll,youre,youve,your,yours,yourself,yourselves,z,zero

D: Cooccurrence matrix for conference sponsorships

	Microsoft	Google	Facebook	Amazon	Apple	Netflix	Count
0	True	False	False	False	False	False	47337
1	False	True	False	False	False	False	18198
2	True	True	False	False	False	False	13468
3	True	True	True	False	False	False	4871
4	False	False	False	True	False	False	4589
5	True	True	True	True	False	False	4409
6	False	False	False	False	True	False	3361
7	True	True	True	True	True	False	3069
8	True	False	True	False	False	False	2769
9	False	False	True	False	False	False	2590
10	True	True	False	True	False	False	1908
11	False	True	True	False	False	False	1623
12	False	True	False	True	False	False	984
13	False	True	True	True	False	False	717
14	True	True	True	True	False	True	676
15	True	False	False	True	False	False	533
16	True	False	False	False	True	False	279
17	True	False	False	False	False	True	267
18	False	True	False	True	False	True	238
19	True	False	True	True	False	False	218
20	True	True	True	False	True	False	170
21	False	False	True	True	True	False	142
22	True	True	False	False	True	False	140
23	False	False	False	False	False	True	138
24	True	False	False	True	True	False	123
25	True	True	False	True	True	False	105
26	False	True	False	False	False	True	99
27	False	True	True	False	False	True	85
28	False	False	True	False	False	True	58
29	False	True	False	False	True	False	58
30	False	False	True	True	False	False	56
31	False	True	False	True	True	False	17

E: Trigrams and Bigrams by decade, from paper titles coming from sponsored conferences

	1991	2001	2011	2021
0	((large, magellanic, cloud), 14)	((ad, hoc, network), 13)	((wireless, sensor, network), 140)	((deep, neural, network), 323)
1	((star, magellanic, cloud), 7)	((monte, carlo, simulation), 9)	((ad, hoc, network), 107)	((convolutional, neural, network), 320)
2	((dynamic, recrystallization, hot), 3)	((mobile, ad, hoc), 7)	((support, vector, machine), 75)	((recurrent, neural, network), 188)
3	((small, magellanic, cloud), 3)	((support, collaborative, learn), 6)	((hide, markov, model), 41)	((neural, machine, translation), 160)
4	((magellanic, cloud, planetarynebulae), 3)	((world, wide, web), 6)	((mobile, ad, hoc), 40)	((deep, reinforcement, learn), 114)
5	((starclusters, magellanic, cloud), 3)	((learn, bayesian, network), 6)	((markov, random, field), 36)	((wireless, sensor, network), 113)
6	((hiiregions, magellanic, cloud), 3)	((support, vector, machine), 6)	((wireless, mesh, network), 31)	((name, entity, recognition), 68)
7	((chemical, evolution, magellanic), 3)	((quantum, monte, carlo), 5)	((artificial, neural, network), 28)	((generative, adversarial, network), 68)
8	((evolution, magellanic, cloud), 3)	((decision, support, systems), 4)	((automatic, speech, recognition), 27)	((deep, convolutional, neural), 58)
9	((simplify, analysis, pwm), 2)	((componentbased, software, engineer), 4)	((spacetime, block, cod), 26)	((online, social, network), 58)

Bigrams

	1991	2001	2011	2021
0	((magellanic, cloud), 72)	((bayesian, network), 32)	((sensor, network), 314)	((neural, network), 1361)
1	((educational, software), 15)	((software, engineer), 29)	((neural, network), 220)	((deep, learn), 451)
2	((large, magellanic), 14)	((monte, carlo), 29)	((case, study), 205)	((social, network), 393)
3	((dynamic, recrystallization), 12)	((position, summary), 28)	((web, service), 200)	((reinforcement, learn), 366)
4	((recrystallization, texture), 10)	((case, study), 22)	((wireless, network), 166)	((case, study), 363)
5	((star, formation), 8)	((atm, network), 19)	((information, retrieval), 166)	((machine, learn), 354)
6	((expert, system), 7)	((ad, hoc), 18)	((wireless, sensor), 152)	((deep, neural), 344)
7	((information, technology), 7)	((model, check), 17)	((ad, hoc), 150)	((social, media), 329)
8	((static, recrystallization), 7)	((software, development), 16)	((semantic, web), 130)	((convolutional, neural), 326)
9	((star, magellanic), 7)	((data, mine), 15)	((speech, recognition), 123)	((speech, recognition), 320)

F: Cooccurrence matrix of the companies in directly funded papers

	Microsoft	Google	Facebook	Amazon	Apple	Netflix	Count
0	True	False	False	False	False	False	7578
1	False	True	False	False	False	False	7138
2	False	False	False	True	False	False	1519
3	False	False	True	False	False	False	686
4	True	True	False	False	False	False	493
5	False	True	False	True	False	False	198
6	False	True	True	False	False	False	189
7	False	False	False	False	True	False	166
8	True	True	True	False	False	False	73
9	True	False	False	True	False	False	71
10	False	False	True	True	False	False	48
11	True	True	False	True	False	False	44
12	False	True	False	False	True	False	42
13	True	True	True	True	False	False	34
14	False	True	True	True	False	False	31
15	True	False	True	False	False	False	29
16	False	False	False	False	False	True	25
17	True	True	True	True	True	False	10
18	True	False	False	False	True	False	4
19	False	True	True	False	True	False	3
20	False	True	False	True	True	False	3
21	True	True	False	True	True	False	3
22	False	False	False	True	True	False	3
23	True	False	True	True	False	False	2
24	True	False	False	True	True	False	1
25	True	True	False	False	True	False	1
26	False	False	True	False	False	True	1
27	False	True	True	True	True	False	1

G: Trigrams and Bigrams of titles of directly funded papers

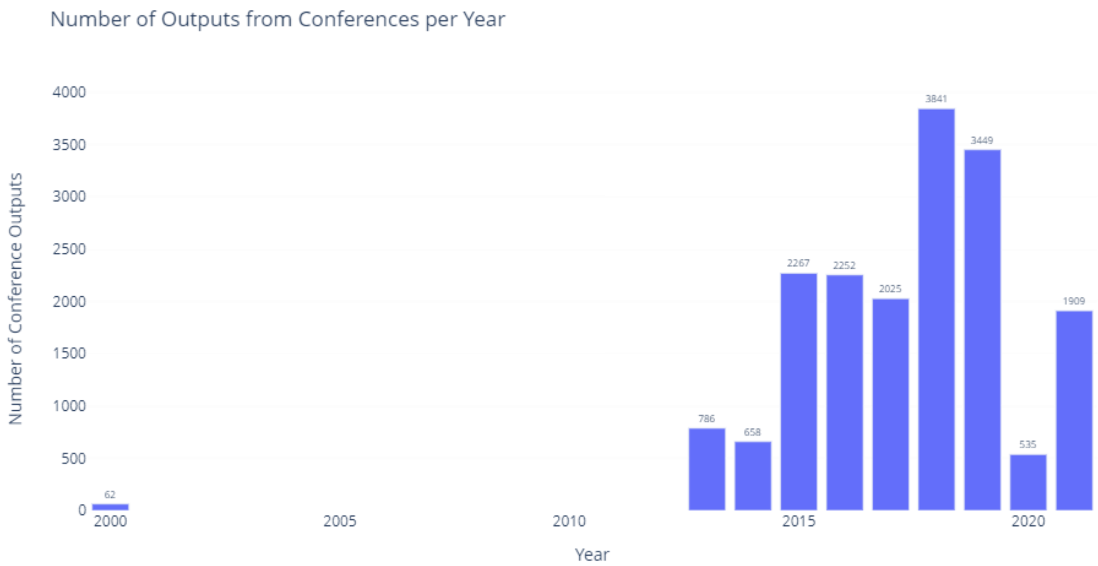
	2001	2011	2021
0	((transform, representation, spectra), 2)	((wireless, sensor, network), 15)	((convolutional, neural, network), 80)
1	((representation, spectra, acoustic), 2)	((fractional, quantum, hall), 10)	((deep, neural, network), 80)
2	((spectra, acoustic, speech), 2)	((twoelectron, reduce, density), 9)	((deep, reinforcement, learn), 30)
3	((acoustic, speech, segment), 2)	((quantum, hall, state), 8)	((neural, machine, translation), 26)
4	((schedule, techniques, reduce), 1)	((antihermitian, contract, schrodinger), 8)	((generative, adversarial, network), 23)
5	((techniques, reduce, processor), 1)	((contract, schrodinger, equation), 8)	((recurrent, neural, network), 23)
6	((reduce, processor, energy), 1)	((quantum, hall, effect), 7)	((graph, neural, network), 18)
7	((processor, energy, macos), 1)	((ad, hoc, network), 7)	((wireless, sensor, network), 17)
8	((prosthesisware, class, software), 1)	((reduce, density, matrices), 6)	((neural, network, base), 16)
9	((class, software, support), 1)	((reduce, density, matrix), 5)	((online, social, network), 16)

Bigrams

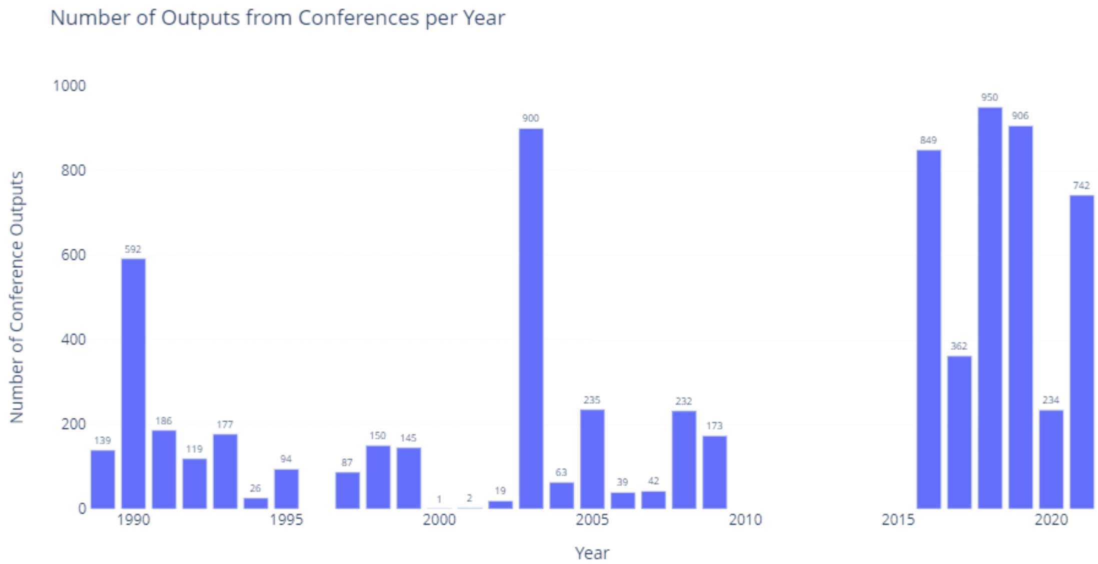
	2001	2011	2021
0	((transform, representation), 2)	((sensor, network), 33)	((neural, network), 349)
1	((representation, spectra), 2)	((quantum, hall), 24)	((deep, learn), 197)
2	((spectra, acoustic), 2)	((social, network), 22)	((machine, learn), 170)
3	((acoustic, speech), 2)	((wireless, sensor), 17)	((reinforcement, learn), 113)
4	((speech, segment), 2)	((image, retrieval), 14)	((social, network), 91)
5	((schedule, techniques), 1)	((web, service), 13)	((deep, neural), 88)
6	((techniques, reduce), 1)	((network, cod), 11)	((case, study), 85)
7	((reduce, processor), 1)	((case, study), 11)	((convolutional, neural), 82)
8	((processor, energy), 1)	((antihermitian, contract), 11)	((big, data), 74)
9	((energy, macos), 1)	((reduce, density), 11)	((social, media), 74)

H: Evolution of number of papers from conference sponsorships by years, per company

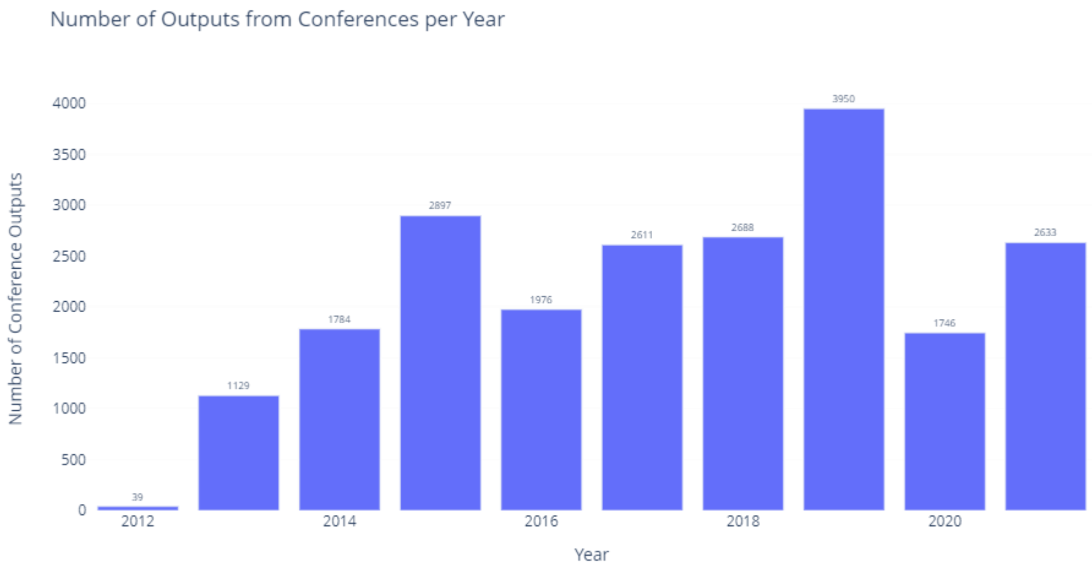
Amazon



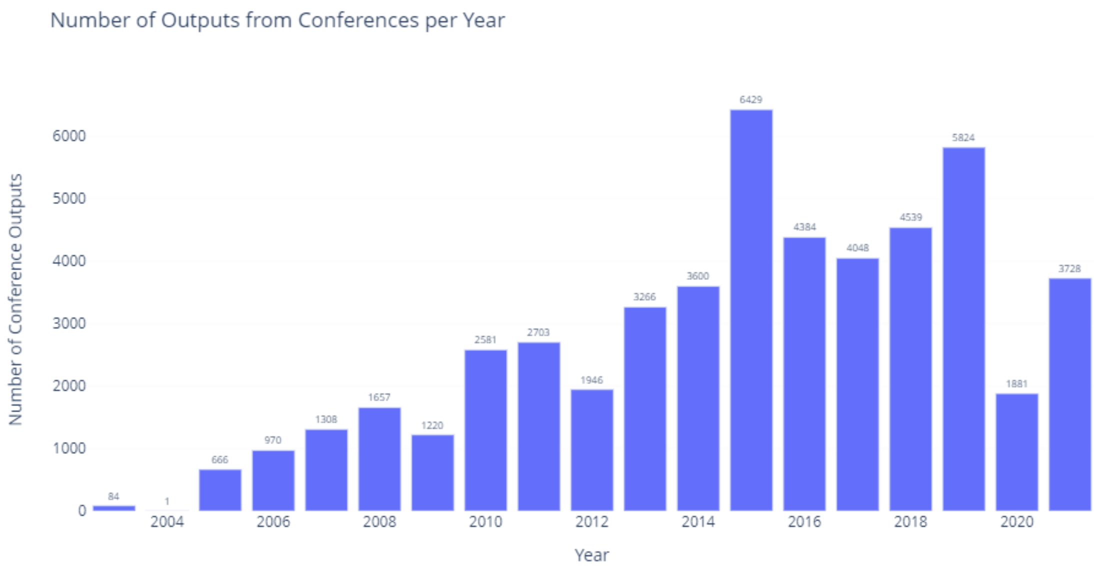
Apple



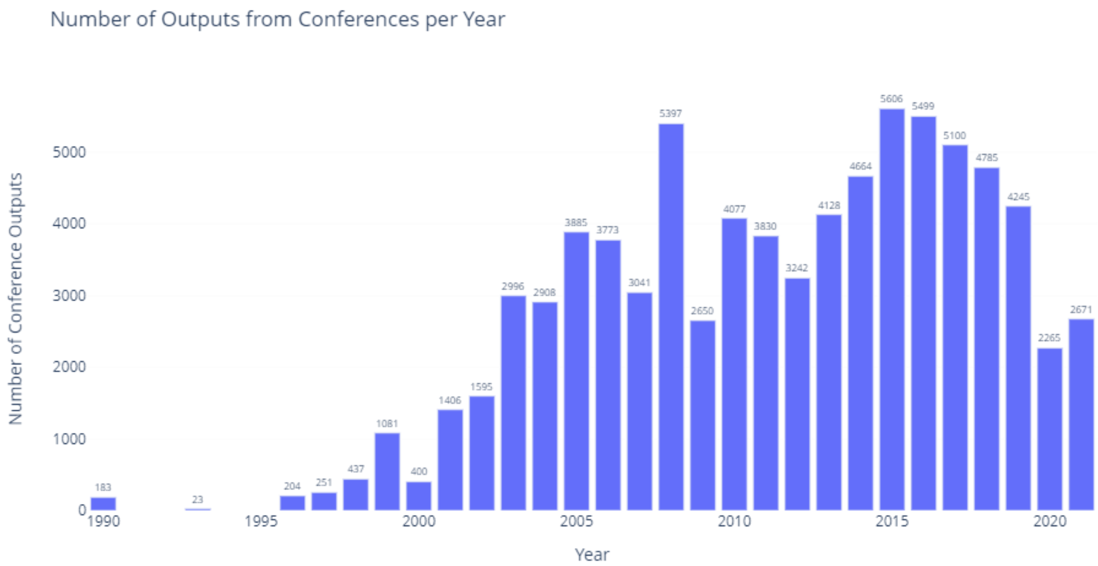
Facebook



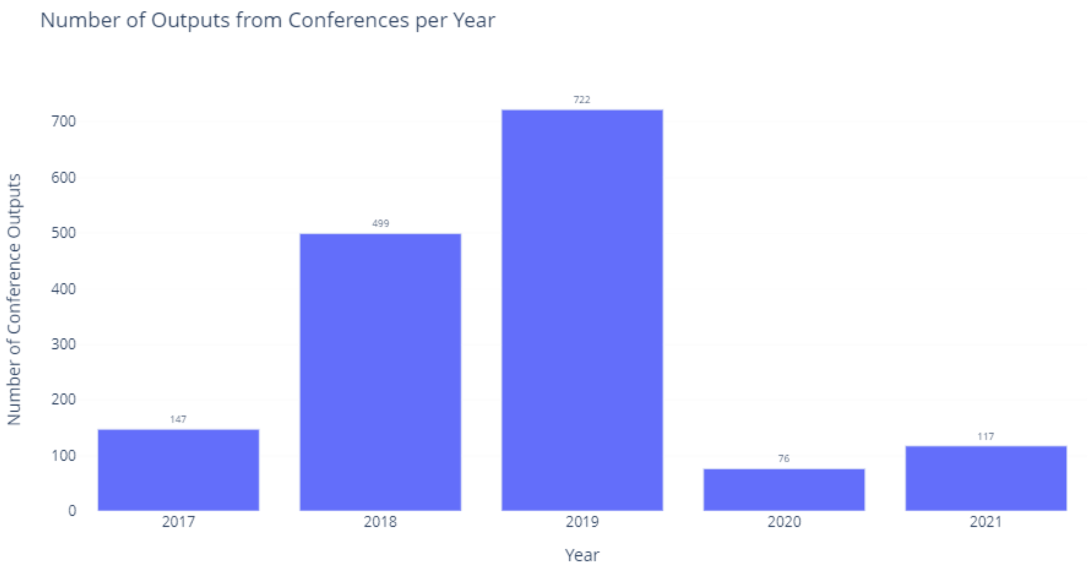
Google



Microsoft

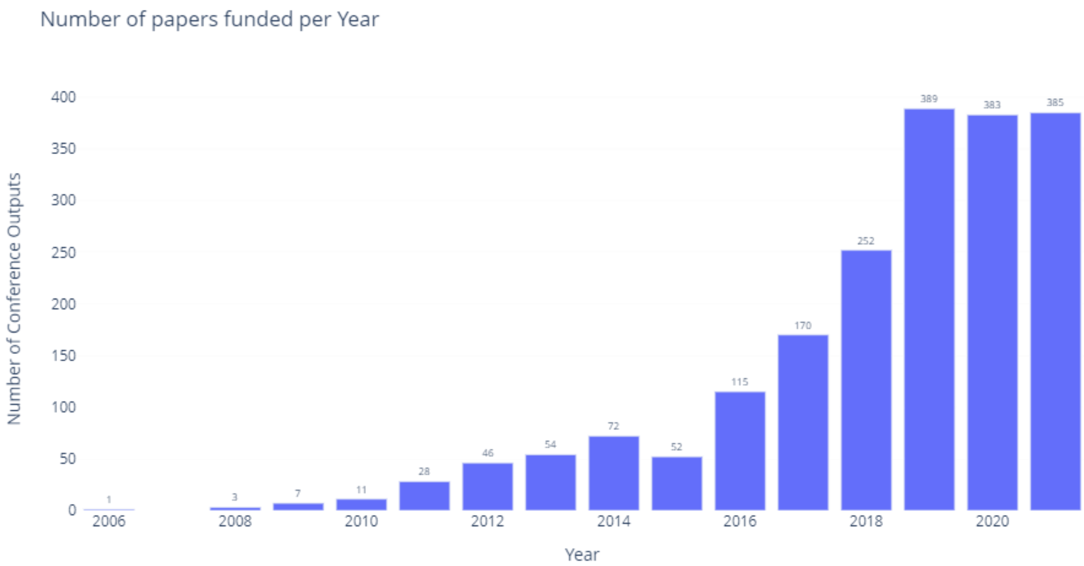


Netflix

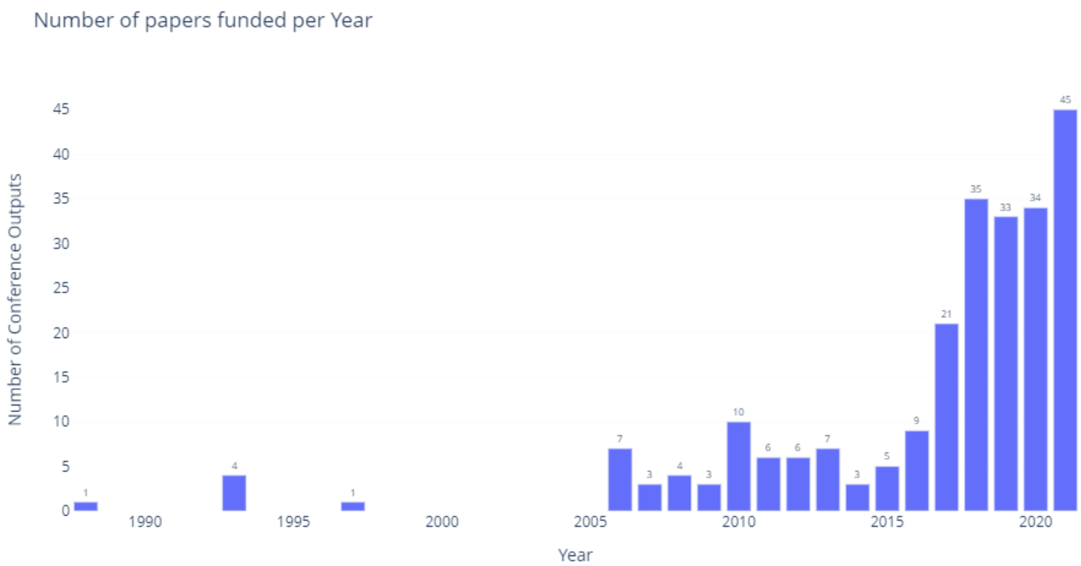


I: Evolution of number of papers directly funded by years, per company

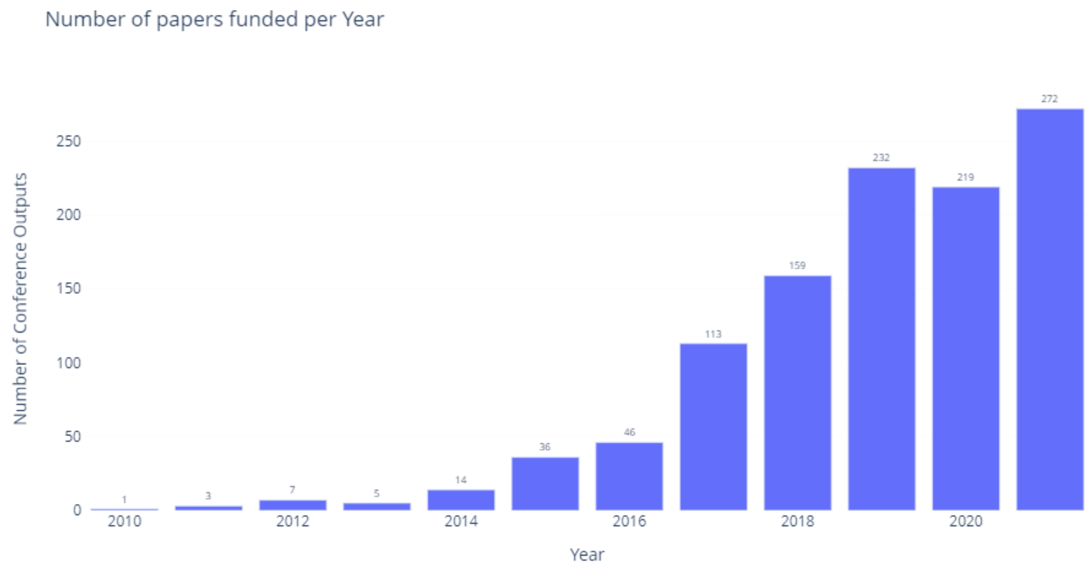
Amazon



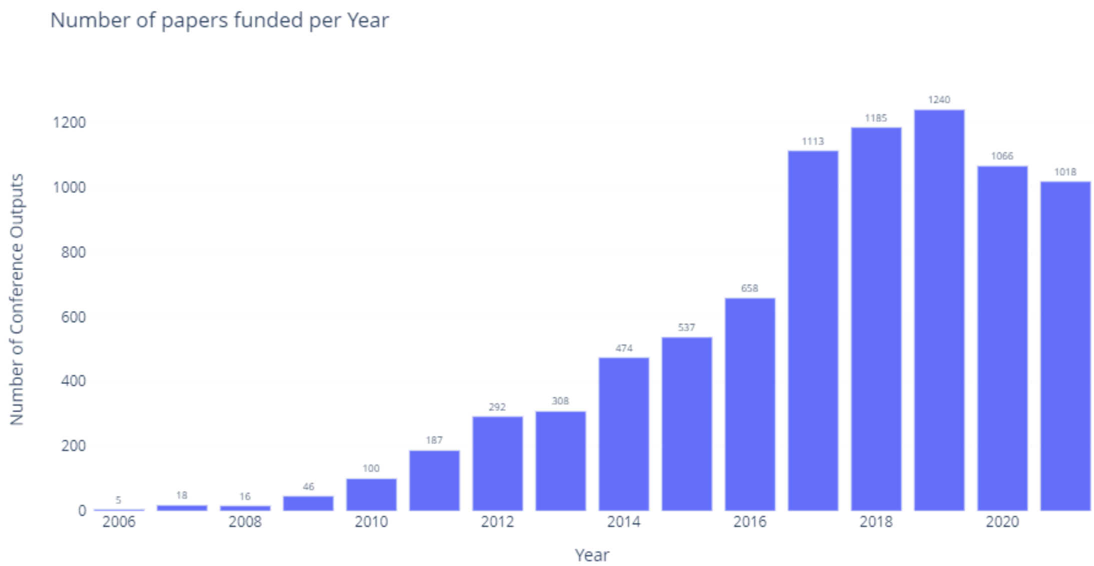
Apple



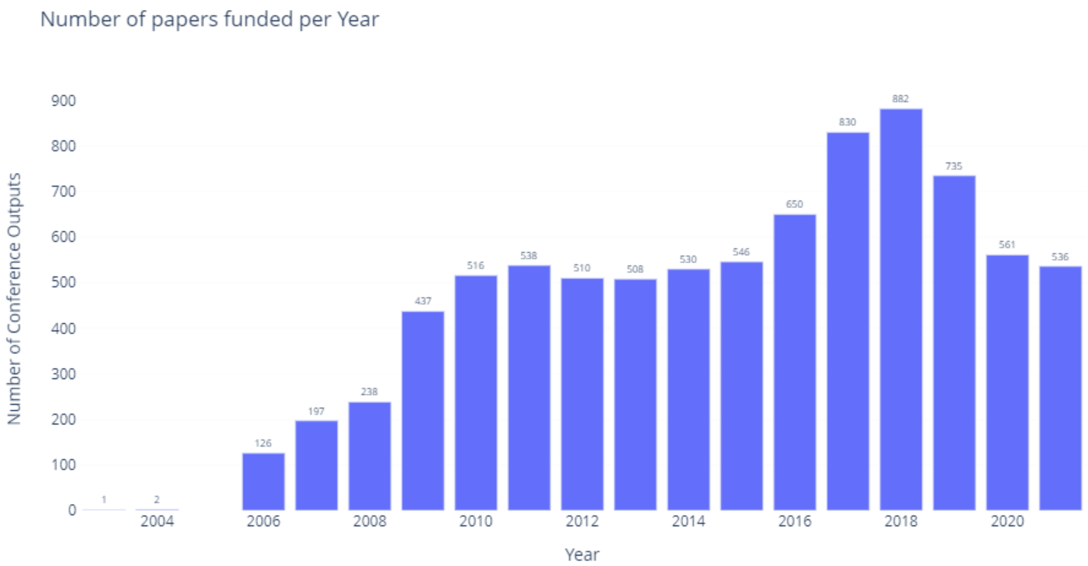
Facebook



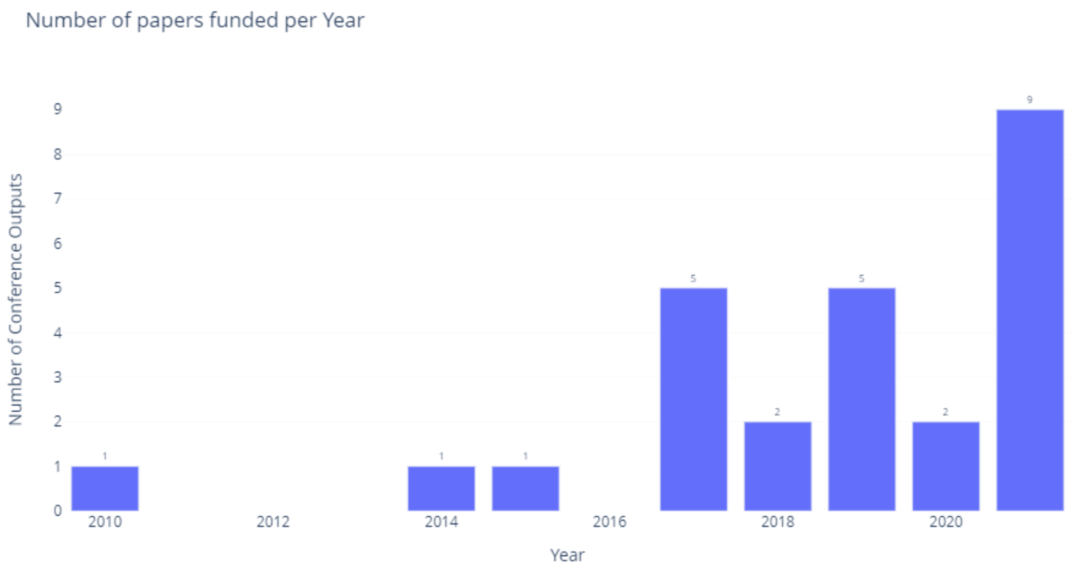
Google



Microsoft


Number of papers funded per Year
Number of Conference Outputs
Year
0
100
200
300
400
500
600
700
800
900
2004
2006
2008
2010
2012
2014
2016
2018
2020
1
2
126
197
238
437
516
538
510
508
530
546
650
830
882
735
561
536

Netflix


Number of papers funded per Year
Number of Conference Outputs
Year
0
1
2
3
4
5
6
7
8
9
2010
2012
2014
2016
2018
2020
1
1
1
5
2
5
2
9

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa