Data Visualization and Insight CA3

**—————————————————————————————————————————-**

**Student Name - Mitul Srivastava**

**Student ID - C00313606**

**Programme title - MS Data Science**

**Programme code - CW\_KCDAR\_M**

**Module title - Data Visualization and Insight**

**Lecturer’s name - Agnes Maciocha**

**Submission date - 20 December, 2024**

**Word count - 1574**

**—————————————————————————————————————————-**

### **Research Question :**

**How have hardware, data, and investment trends shaped the landscape of Artificial General Intelligence (AGI) research over the past decade?**

#### **What is AGI?**

AGI meaning Artificial General Intelligence, which is the something we are aiming for in Artificial Intelligence is something that it can reason, learn and adapt like humans and apply the skills gained through that to many tasks that are not limited to a domain. In contrast to narrow AI, which shines in individual applications, AGI aims for an overarching understanding similar to how a human thinks.

#### **Why is it important to research about AGI (Significance)?**

Here are some key points highlighting the potential impact of AGI on society and technology:

* Unprecedented problem-solving capabilities: AGI can solve issues such as climate change, diseases, and the problems of resource management more efficiently than humans, which may result to solutions for some of the world’s biggest problems.
* Scientific research acceleration: With the help of AGI, there is the possibility of accelerating the process of scientific advancement and discovery in all areas, which may translate to faster development in areas such as medicine, physics, and space exploration.
* Enhanced decision-making: With data-based recommendations, AGI has the potential of transforming policy making and business planning thus offering better ways of governance and economic systems.
* Redefining human-machine interaction: AGI can also transform the current forms of communication between man and machine into more friendly and complex ones, possibly influenced by the way we engage with technology.

**—————————————————————————————————————————-**

**Importing the libraries required for the analysis.**

library(dplyr)  
library(tidyverse)  
#new libraries introduced  
library(rvest) #Used for web scraping the data on data generation.  
library(httr) #Used in conjunction with rvest for making HTTP requests.  
library(stringr) #Used for string manipulation in the web scraping section.

**—————————————————————————————————————————**

### **Hardware Power Analysis**

**The GPU\_benchmarks\_v7 contains the data for GPU and CPU performance over years.**

GPU\_performance\_df <- read.csv("C:/Users/Mitul/Desktop/Study/R class/CA 3/GPU\_benchmarks\_v7.csv")

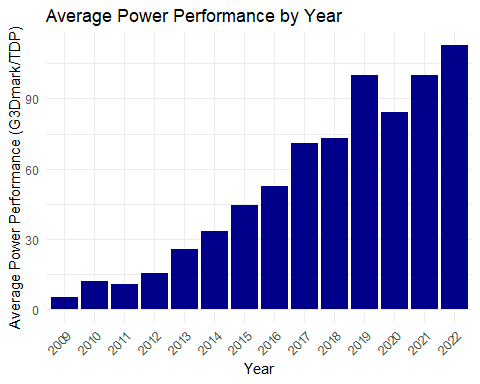
**Analysis of GPU data:**

1. Adding new column year derived from testdate.
2. Grouping by year and summarizing using mean value of powerPerformance.
3. Filtering NA from year.

GPU\_performance\_df\_summary <- GPU\_performance\_df %>%  
 mutate(year = as.integer(substr(testDate, 1, 4))) %>%  
 group\_by(year) %>%  
 summarise(avg\_powerPerformance = mean(powerPerformance, na.rm = TRUE)) %>%  
 filter(!is.na(year))

**Visualization of GPU data:**

ggplot(GPU\_performance\_df\_summary, aes(x = factor(year), y = avg\_powerPerformance)) +  
 geom\_bar(stat = "identity", fill = "darkblue") +  
 geom\_smooth(method = "lm", color = "red", size = 1) +  
 labs(title = "Average Power Performance by Year",  
 x = "Year",  
 y = "Average Power Performance (G3Dmark/TDP)") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#### Insight - From 2013 to 2022, 337 percent rise in PowerPerformance of GPU going from 25.7 to 112.5 G3d/TDP.

*The most powerful GPU available in the market is GeForce RTX 4090 GPU which can do more than 1300 TOPS, or trillions of operations per second. In comparison Copilot+ PC lineup by Microsoft, which includes neural processing units (NPUs) able to perform upwards of 40 TOPS.*

**—————————————————————————————————————————-**

### **Amount of Data in world**

**Amount of data produced in the world taken from *“explodingtopics”* blog:**

url <- "https://explodingtopics.com/blog/data-generated-per-day"  
webpage <- read\_html(url)  
  
# Extracting the table  
table\_data <- webpage %>%  
 html\_nodes("table") %>%  
 .[[2]] %>%  
 html\_table()

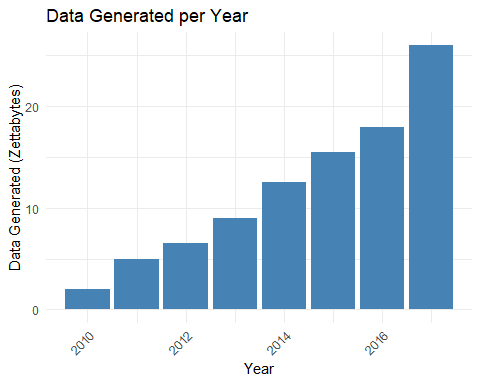
**Analysis of world data :**

1. Selecting the Year and Data\_Generated column from the table.
2. Adding columns Year and Data\_Generated after converting them to numeric.
3. Filtering NA values if any.

WorldData <- table\_data %>%  
 select(Year = 1, Data\_Generated = 2) %>%  
 mutate(Year = as.numeric(Year),  
 Data\_Generated = as.numeric(gsub(" zettabytes", "", Data\_Generated))) %>%  
 filter(!is.na(Year) & !is.na(Data\_Generated))  
  
# The data has been read from source and is been written in a csv to provide as backup.  
write.csv(WorldData,"WorldData.csv")

**WorldData Visualization bar plot:**

# Create the bar chart  
ggplot(WorldData, aes(x = Year, y = Data\_Generated)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 labs(title = "Data Generated per Year",  
 x = "Year",  
 y = "Data Generated (Zettabytes)") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#### Insight - From 2010 to 2017, 1200 percent rise in world data going from 2 zb to 26 zb.

*1 billion TB in 1 ZB. The 90 percent of data which exist in the world today was created in the last two years from 2021-2023*

**—————————————————————————————————————————-**

### **AI funding received over years**

**Importing Private investment data from [ourworldindata.org].**

ai\_investment\_data <- read.csv("C:/Users/Mitul/Desktop/Study/R class/CA 3/private-investment-in-artificial-intelligence.csv")

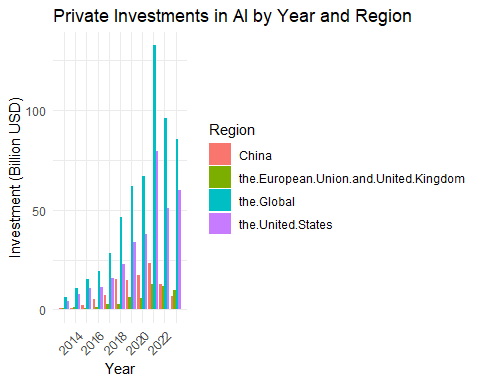
**Analysis of Investment data:**

1. Pivoting the columns to change the data shape from wide to long.
2. Two new columns created **Region** having places and **Investment** having amounts invested.
3. Removing “Private investment in AI in” from Region names.

investment\_df <- pivot\_longer(  
 ai\_investment\_data,  
 cols = starts\_with("Private.investment"),  
 names\_to = "Region",  
 values\_to = "Investment"  
)  
  
# Clean up region names for better readability  
investment\_df$Region <- gsub("Private.investment.in.AI.in.", "", investment\_df$Region)

**Visualization of Investment data:**

# Create the bar chart  
ggplot(investment\_df, aes(x = Year, y = (Investment/1000000000), fill = Region)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(  
 title = "Private Investments in AI by Year and Region",  
 x = "Year",  
 y = "Investment (Billion USD)",  
 fill = "Region"  
 ) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 scale\_y\_continuous(labels = scales::comma)



#### Insight - From 2013 to 2021, 1814 percent rise in investment going from 4.2 billion USD to 80.4 billion USD.

**—————————————————————————————————————————**

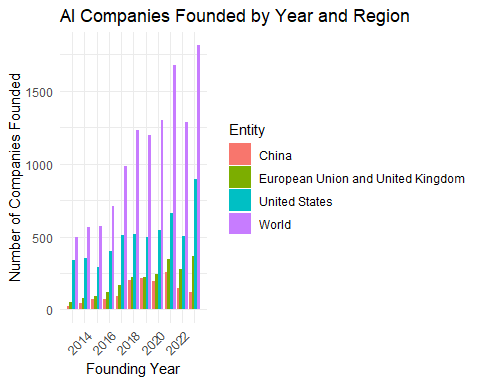
### **AI companies founded in the last decades**

**Importing the data for number of companies founded related to AI from [ourworldindata.org].**

Companies\_year\_df <- read.csv("https://ourworldindata.org/grapher/newly-funded-artificial-intelligence-companies.csv?v=1")  
  
# The data has been read from source and is been written in a csv to provide as backup.  
write.csv(Companies\_year\_df, "companies\_year\_data.csv", row.names = FALSE)

**Visualization of companies data:**

ggplot(Companies\_year\_df, aes(x = Year, y = Newly.founded.AI.companies, fill = Entity)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(  
 x = "Founding Year",  
 y = "Number of Companies Founded",  
 title = "AI Companies Founded by Year and Region"  
 ) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



#### Insight - From 2013 to 2023, 266 percent rise in number of startup going from 495 to 1812.

**—————————————————————————————————————————-**

### **Research towards AGI**

**Webscraping data to get research papers published on the topic of “AGI”.**

base\_url <- "https://arxiv.org/search/?query=AGI&searchtype=all&start="  
  
# Initialize an empty data frame  
Research\_df <- data.frame(Title = character(), SubmissionYear = character(), stringsAsFactors = FALSE)  
  
# Looping through pages  
for (page in seq(0, 450, by = 50)){   
 url <- paste0(base\_url, page)  
   
 # Fetching webpage content  
 response <- GET(url, user\_agent("Mozilla/5.0"))  
 webpage <- read\_html(content(response, as = "text"))  
   
 # Extracting the titles  
 titles <- webpage %>% html\_nodes(".title.is-5.mathjax") %>% html\_text(trim = TRUE)  
   
 # Extracting raw submission dates  
 raw\_submission\_dates <- webpage %>%   
 html\_nodes("p.is-size-7") %>%   
 html\_text(trim = TRUE)  
   
 # Filtering valid submission dates  
 valid\_submission\_dates <- raw\_submission\_dates[str\_detect(raw\_submission\_dates, "Submitted")]  
   
 # Extracting the year from submission dates  
 cleaned\_submission\_years <- str\_extract(valid\_submission\_dates, "\\d{4}")  
   
 # Aligning lengths of titles and submission years  
 min\_length <- min(length(titles), length(cleaned\_submission\_years))  
 titles <- titles[1:min\_length]  
 cleaned\_submission\_years <- cleaned\_submission\_years[1:min\_length]  
   
 # Create a data frame for the current page  
 page\_data <- data.frame(Title = titles, SubmissionYear = cleaned\_submission\_years, stringsAsFactors = FALSE)  
   
 # Append to the main data frame  
 Research\_df <- bind\_rows(Research\_df, page\_data)  
   
 # Print progress  
 cat("Scraped page starting at:", page, "with", nrow(page\_data), "records.\n")  
}

## Scraped page starting at: 0 with 50 records.  
## Scraped page starting at: 50 with 50 records.  
## Scraped page starting at: 100 with 50 records.  
## Scraped page starting at: 150 with 50 records.  
## Scraped page starting at: 200 with 50 records.  
## Scraped page starting at: 250 with 50 records.  
## Scraped page starting at: 300 with 50 records.  
## Scraped page starting at: 350 with 50 records.  
## Scraped page starting at: 400 with 50 records.  
## Scraped page starting at: 450 with 45 records.

# Save the results to a CSV file  
write.csv(Research\_df, "arxiv\_AGI\_data\_year.csv", row.names = FALSE)

**—————————————————————————————————————————-**

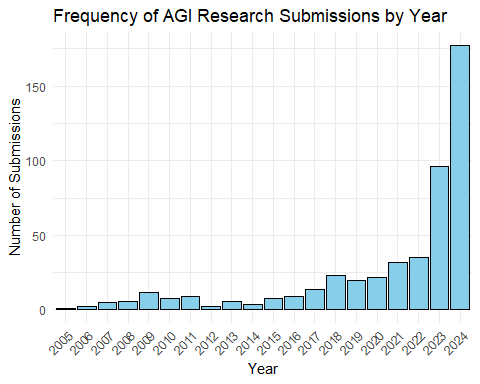
**Analysis of research data:**

1. Filtering where years are not NA.
2. Grouping by SubmissionYear.
3. Calculating the frequency of papaers per year.

freq\_research <- Research\_df %>%  
 filter(!is.na(SubmissionYear)) %>%  
 group\_by(SubmissionYear) %>%  
 summarise(count = n())

**Visualization of reseach data:**

Research\_plot <- ggplot(freq\_research, aes(x = SubmissionYear, y = count)) +  
 geom\_bar(stat = "identity", fill = "skyblue", color = "black") +  
 labs(  
 title = "Frequency of AGI Research Submissions by Year",  
 x = "Year",  
 y = "Number of Submissions"  
 ) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
print(Research\_plot)



#### Insight - From 2013 to 2023, 1500 percent rise in number of research papers going from 6 to 96.

**—————————————————————————————————————————-**

### **Future Prospects:**

1. Advances in Computational Power Hardware like GPUs, TPUs, and neuromorphic chips (e.g., NVIDIA H100) are enabling more powerful AGI training, with quantum computing on the horizon.
2. Rising Investments and Collaborations Tech giants like OpenAI and DeepMind, alongside governments, are heavily investing in AGI research, emphasizing performance and safety.
3. Improved Learning Paradigms New methods like reinforcement learning (e.g., AlphaGo) and unsupervised learning (e.g., GPT models) are driving AGI’s ability to generalize tasks efficiently.

**—————————————————————————————————————————**

### **Conclusion :**

The analysis offers a powerful overview of the unfolding trajectories of Artificial General Intelligence (AGI) with the hardware, data generation and capital inputs in our holistic environment. It highlights the narrative of technological innovation through the explorations and visualizations of trends and patterns in research. This study highlights the transformative potential of AGI while particularly discussing the challenges and future of this field. Overall, the analysis highlights AGI’s crucial role in determining the future of both technology and humanity.

**—————————————————————————————————————————**

### **References :**

* <https://arxiv.org/search/?query=AGI&searchtype=all&source=header>
* <https://ourworldindata.org/grapher/private-investment-in-artificial-intelligence>
* <https://explodingtopics.com/blog/data-generated-per-day>
* <https://www.kaggle.com/datasets/alanjo/gpu-benchmarks>
* <https://www.perplexity.ai/>

### **Video link**

<https://setuo365-my.sharepoint.com/:v:/g/personal/c00313606_setu_ie/EXSrkoZnkrJPn4yTbLPin8wBJMmeOaYynImTYyAu4ztv-Q?e=a7T0zB&nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbFZpZXciOiJTaGFyZURpYWxvZy1MaW5rIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXcifX0%3D>