Tech Funding Distribution

Pitchaya Manopchantaroj

Background: Tech Company Funding Since 2020

The past decade has witnessed an unprecedented surge in technological innovation, driven by advancements in various technologies such as artificial intelligence, blockchain, cloud computing, etc. As startups play a crucial role in shaping this digital transformation, we can use data on venture capital and private investments to provide insight into funding trends and how it has fuel innovation worldwide. It includes information on companies from different regions, spanning multiple industry. The dataset captures key funding details, including the amount raised, the investment stage (e.g., Seed, Series A), and the date of funding.

Motivation and Research Focus

A key objective of this analysis is to uncover patterns in investment behavior. This can be seeing whether certain regions are emerging as new innovation hubs or specific technology sectors are experiencing surges in funding. Understanding these trends can offer insights into broader market dynamics, including the evolving priorities of investors and the changing technological landscape. A satisfactory analysis would reveal not only regional disparities in funding but also shifts in industry focus, highlighting where capital is flowing and which innovations are gaining momentum.

Tech Funding Data

The dataset contains records of funding events for technology startups, capturing key details about each company and its investment history.

The main variables include:

Company: The name of the funded startup

Region: The geographic location where the company is based

Vertical: The industry or sector the company operates in

Funding Amount (USD): The total funding received in U.S. dollars

Funding Stage: The stage of investment

Funding Date: The date when the funding was recorded

```
techfund_raw <- read_csv('tech_fundings.csv')
techfund <- techfund_raw |>
  # Convert funding to numeric, properly handling commas
  mutate(`Funding Amount (USD)` = ifelse(is.na(`Funding Amount (USD)`),
                                        0, `Funding Amount (USD)`)) |>
  mutate(`Funding Amount (USD)` = suppressWarnings(as.numeric(gsub(
                                        ",", "", `Funding Amount (USD)`))))|>
  # Handle missing values in Region
  mutate(Region = ifelse(is.na(Region), "Unknown", Region))

# Group smaller regions
techfund <- techfund |>
  mutate(RegionClean = fct_lump_n(Region, 20, other_level = "Other Regions"))

# Table of first 5 entries to get an idea of what the dataset looks like
techfund |>
  head(5) |>
  select(-1, -3, -ncol(techfund)) |>
  kable(caption = "First 5 Entries of Tech Fundings",
        align = rep("c", ncol(select(techfund, -1, -3, -ncol(techfund)))))
```

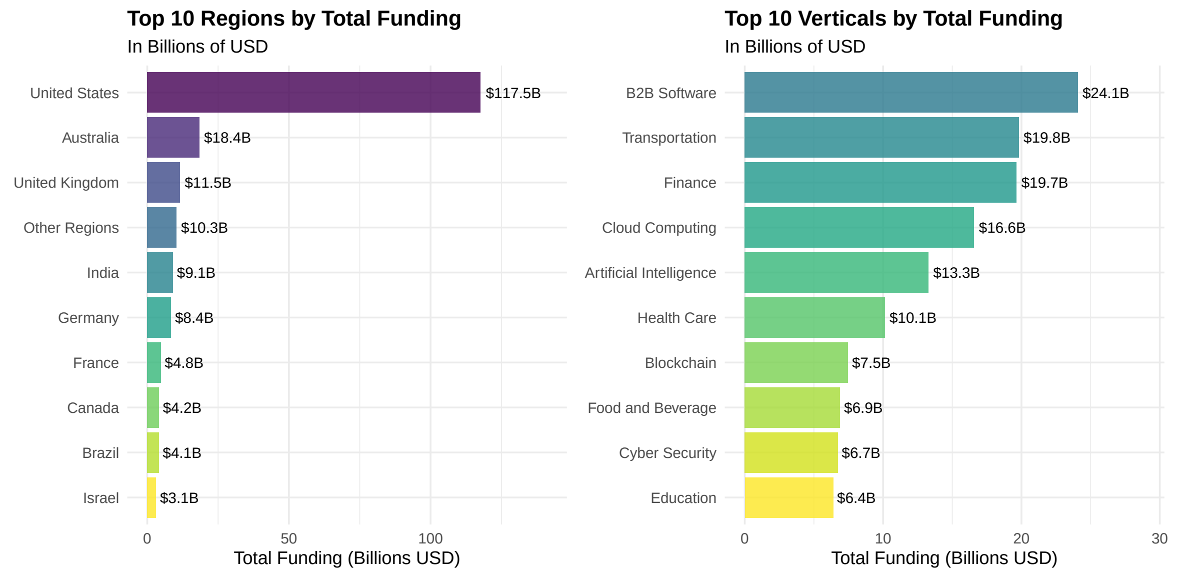
Table 1: First 5 Entries of Tech Fundings

Company	Region	Vertical	Funding Amount (USD)	Funding Stage	Funding Date
Internxt	Spain	Blockchain	278940	Seed	Jan-20
Dockflow	Belgium	Logistics	292244	Seed	Jan-20
api.video	France	Developer APIs	300000	Seed	Jan-20
Buck.ai	United States	Artificial Intelligence	300000	Seed	Jan-20
Prodsight	United Kingdom	Artificial Intelligence	529013	Seed	Jan-20

Exploratory Analysis

With this dataset, we will begin by using graphical and descriptive techniques to provide a broad view of the global startup ecosystem, allowing us to analyze how funding is distributed across regions and how investor focus has shifted over time. To begin our analysis, we will conduct a univariate exploration of the dataset, generating visualizations to better understand its distribution and key characteristics.

```r
# Aggregate funding data by region
region_funding <- techfund |>
  filter(!is.na(Region), !is.na(`Funding Amount (USD)`)) |>
  group_by(RegionClean) |>
  summarize(
    TotalFunding = sum(`Funding Amount (USD)`, na.rm = TRUE),
    AvgFunding = mean(`Funding Amount (USD)`, na.rm = TRUE),
    MedianFunding = median(`Funding Amount (USD)`, na.rm = TRUE),
    NumCompanies = n_distinct(Company),
    NumDeals = n()
  ) |>
  arrange(desc(TotalFunding))

# Top regions plot
regions_plot <- region_funding |>
  head(10) |>
  ggplot(aes(x = reorder(RegionClean, TotalFunding), y = TotalFunding / 1e9))+
  geom_bar(stat = "identity", fill = viridis(10, alpha = 0.8)) +
  geom_text(aes(label = sprintf("$%.1fB", TotalFunding / 1e9)),
            hjust = -0.1, size = 3) +
  coord_flip() +
  expand_limits(y = max(region_funding$TotalFunding) * 1.2 / 1e9) +
  labs(title = "Top 10 Regions by Total Funding",
       subtitle = "In Billions of USD",
       x = NULL,
       y = "Total Funding (Billions USD)") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"))

# Aggregate funding data by vertical
vertical_funding <- techfund |>
  filter(!is.na(Vertical), !is.na(`Funding Amount (USD)`)) |>
  group_by(Vertical) |>
  summarize(
    TotalFunding = sum(`Funding Amount (USD)`, na.rm = TRUE),
    AvgFunding = mean(`Funding Amount (USD)`, na.rm = TRUE),
    MedianFunding = median(`Funding Amount (USD)`, na.rm = TRUE),
    NumCompanies = n_distinct(Company),
    NumDeals = n()
  ) |>
  arrange(desc(TotalFunding))

# Top verticals plot
verticals_plot <- vertical_funding |>
  head(10) |>
  ggplot(aes(x = reorder(Vertical, TotalFunding), y = TotalFunding / 1e9)) +
  geom_bar(stat = "identity", fill = viridis(10, alpha = 0.8, begin = 0.4)) +
  geom_text(aes(label = sprintf("$%.1fB", TotalFunding / 1e9)),
            hjust = -0.1, size = 3) +
  coord_flip() +
  expand_limits(y = max(vertical_funding$TotalFunding) * 1.2 / 1e9) +
  labs(title = "Top 10 Verticals by Total Funding",
       subtitle = "In Billions of USD",
       x = NULL,
       y = "Total Funding (Billions USD)") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"))

print(regions_plot+verticals_plot)
```



Regional Funding Distribution

The disparity between the US and other regions is striking - the US has received approximately 6.4× more funding than Australia, the second-highest funded region. The US remains the dominant hub for tech funding, suggesting continued centralization of innovation capital despite global digital transformation. Besides the US, we see the total funding of the other 9 countries to be in a similar range from one another. Countries like India and Brazil appear in the top 10, indicating growing investment in emerging markets as well. Something else that is interesting to consider is the fact that I grouped every country outside the top 10 regions as `Other regions` and those combined are still behind countries like the UK and Australia.

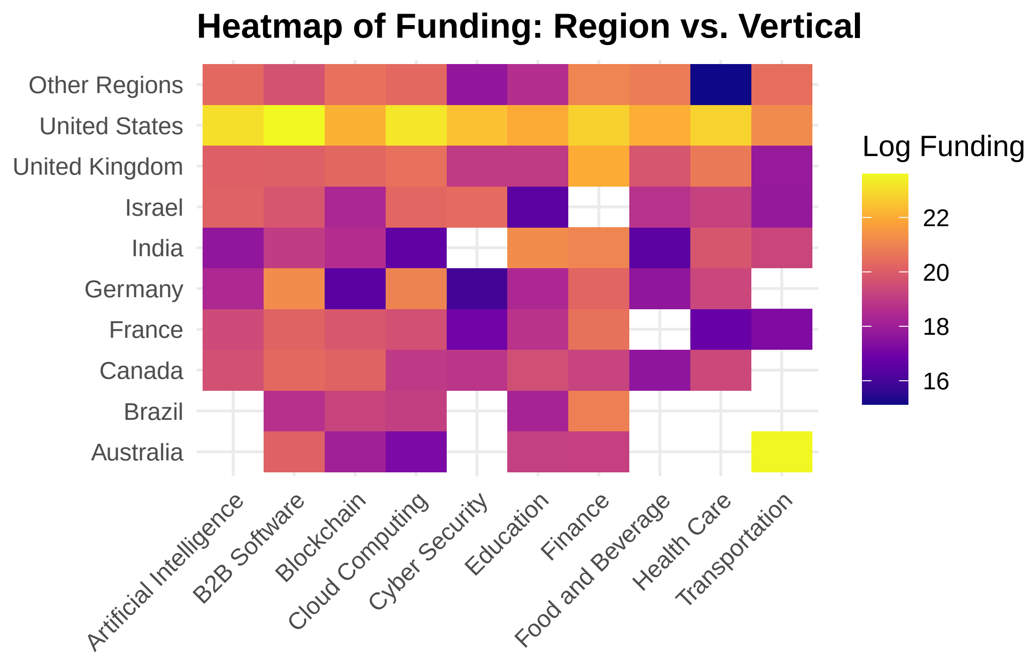
Vertical Funding Distribution

The funding distribution across verticals shows greater balance than regional distribution, with less pronounced differences between top categories. B2B Software, transportation, and finance lead in total funding, underscoring the importance of enterprise solutions and traditional infrastructure. Future-focused technologies like AI, cloud computing, and blockchain attract substantial investment, while essential sectors like food and education lag behind. This investment pattern reveals a key market dynamic: investors favor emerging technologies with higher growth potential and disruptive capabilities over staple sectors serving established markets with consistent demand, despite the universal necessity of these basic services.

Heatmap of Funding: Region vs. Vertical

```r
# Heatmap of Region and Vertical
region_vertical_funding <- techfund |>
  filter(!is.na(Region), !is.na(Vertical), !is.na(`Funding Amount (USD)`)) |>
  group_by(RegionClean, Vertical) |>
  summarize(
    TotalFunding = sum(`Funding Amount (USD)`, na.rm = TRUE),
    NumDeals = n()
  ) |>
  ungroup()

# Get top 10 regions and top 10 verticals for the heatmap
top_regions <- region_funding$RegionClean |> head(10)
top_verticals <- vertical_funding$Vertical |> head(10)

# Filter data for heatmap
heatmap_data <- region_vertical_funding |>
```

```
  filter(RegionClean %in% top_regions, Vertical %in% top_verticals)

# Create heatmap for region vs. vertical
heatmap_plot <- heatmap_data |>
  ggplot(aes(x = Vertical, y = RegionClean, fill = log1p(TotalFunding))) +
  geom_tile() +
  scale_fill_viridis_c(name = "Log Funding", option = "plasma") +
  labs(title = "Heatmap of Funding: Region vs. Vertical",
       x = NULL, y = NULL) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold"))

print(heatmap_plot)
```



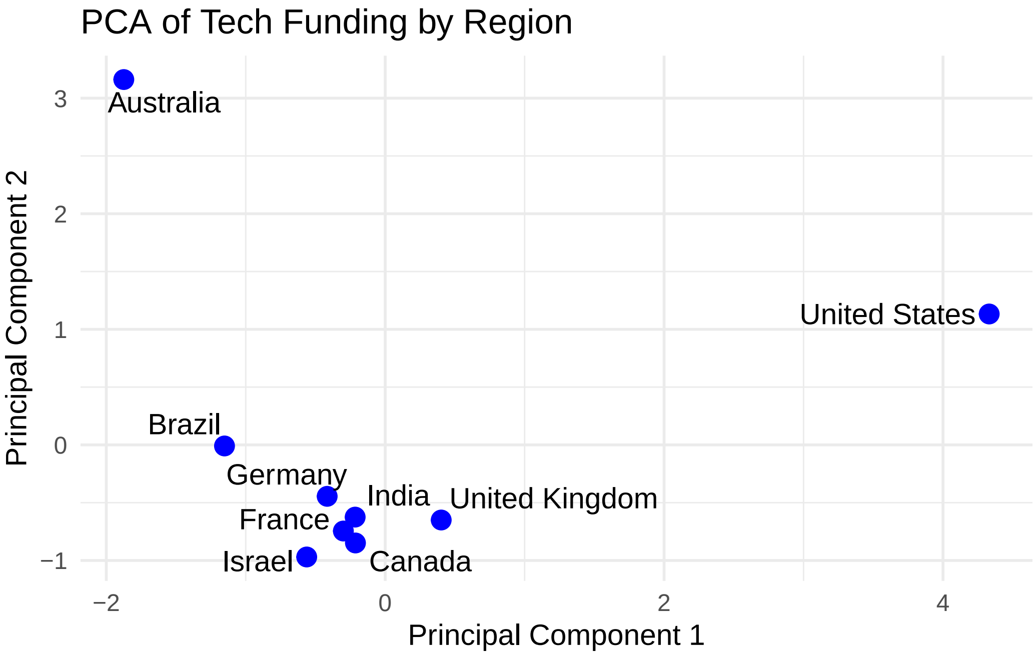
Investment Concentrations Based on Heat Map

The heatmap reveals which regions are specializing in specific verticals. For example, the US shows consistently high funding (yellow/light orange) across most verticals, confirming its role as a dominant innovation hub across multiple sectors. On the other hand, Australia shows particularly strong funding in Transportation (brightest yellow), suggesting it may be emerging as a hub for transportation technology, beating the US in funding despite its smaller overall funding. Other countries like the UK shows strength in Education, India demonstrates notable activity in Finance, and Germany focuses on B2B Software, indicating regional specialization rather than uniform global investment patterns. The white spaces in the heatmap identify sectors where funding is notably absent, potentially indicating untapped market opportunities. Despite seeing from the previous bar graphs how much countries are investing in future technology, it is surprising to see that there is a lack of investments in cybersecurity.

Principal Component Analysis: Regional Investment Patterns in Tech Funding

```r
# Aggregate data for PCA: Using mean, variance, and count of funding
pca_data <- techfund |>
  filter(Region %in% top_regions) |>
  group_by(Region) |>
  summarize(
    mean_funding = mean(`Funding Amount (USD)`, na.rm = TRUE),
    sd_funding = sd(`Funding Amount (USD)`, na.rm = TRUE),
    total_funding = sum(`Funding Amount (USD)`, na.rm = TRUE),
    num_investments = n(),
    num_verticals = n_distinct(Vertical)
  ) |>
  filter(!is.na(mean_funding), !is.na(sd_funding)) |>
  column_to_rownames("Region") |>
  as.matrix()

# Perform PCA on the data
pca_result <- prcomp(pca_data, scale. = TRUE)

# Convert PCA results to dataframe
pca_df <- as.data.frame(pca_result$x[, 1:2])
pca_df$Region <- rownames(pca_df)

# Plot PCA results
pca_plot <- ggplot(pca_df, aes(x = PC1, y = PC2, label = Region)) +
  geom_point(size = 3, color = "blue") +
  geom_text_repel(max.overlaps = 10) +
  labs(title = "PCA of Tech Funding by Region",
       x = "Principal Component 1",
       y = "Principal Component 2") +
  theme_minimal()
```

```
print(pca_plot)
```



PCA Results

Based on the PCA results, we can visualize how regions cluster according to their funding characteristics (mean amount, variability, total volume, investment count, and sector diversity), revealing which tech hubs exhibit similar investment behaviors and which stand as outliers in the global funding landscape.

Here are the notable findings:

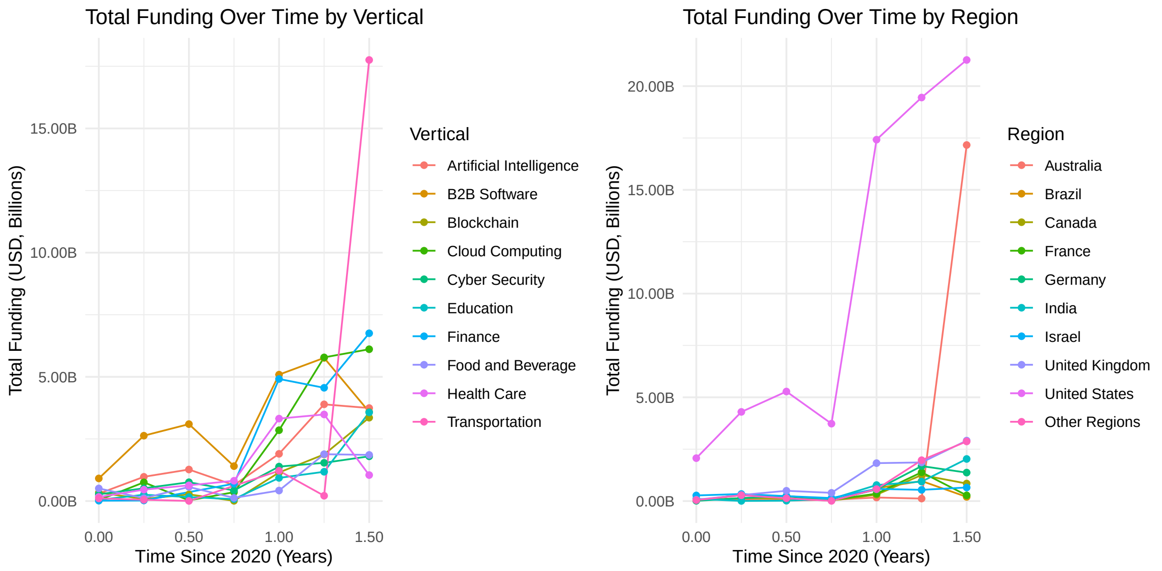
1. Distinct Investment Profiles: The PCA plot shows clear separation between regions, particularly with the United States and Australia positioned distinctly from other countries, suggesting they have unique investment profiles.
2. United States Dominance: The US appears far to the right on PC1, indicating it leads in total funding volume and diversity of investments across multiple sectors.
3. Australia's Unique Position: Australia's high position on PC2 while being relatively low on PC1 suggests it has diversified investments across verticals despite lower overall funding compared to the US.

4. Clustered Emerging Markets: Brazil, Germany, India, France, Israel, and Canada cluster together, indicating similar investment patterns with moderate diversity and funding.
5. UK's Transition Position: The UK falls between the main cluster and the US, suggesting it's transitioning toward a more mature funding ecosystem but is still far from the US.

Total Funding Over Time

```r
# Convert "MMM-YY" format to proper date (changing it to be by quarter)
techfund <- techfund |>
  mutate(
    parsed_date = as.Date(paste0("01-", `Funding Date`), format = "%d-%b-%y"),
    funding_year = year(parsed_date),
    funding_quarter = quarter(parsed_date),
    time_since_2020 = funding_year - 2020 + (funding_quarter - 1) / 4
  ) |>
  filter(!is.na(parsed_date))  # Remove invalid dates

# Filter techfund to only include top regions and verticals
techfund_top <- techfund |>
  filter(RegionClean %in% top_regions, Vertical %in% top_verticals)

# Find the Vertical with the Most Funding Over Time
vertical_trends <- techfund_top |>
  group_by(time_since_2020, Vertical) |>
  summarise(total_funding = sum(`Funding Amount (USD)`, na.rm = TRUE) / 1e9,
            .groups = 'drop')

vertical_time <- ggplot(vertical_trends, aes(x = time_since_2020,
                                             y = total_funding,
                                             color = Vertical)) +
  geom_line(linewidth = 0.5) +
  geom_point() +
  scale_x_continuous(labels = scales::label_number(accuracy = 0.01)) +
  scale_y_continuous(labels = label_number(suffix = "B", accuracy = 0.01)) +
  labs(title = "Total Funding Over Time by Vertical",
       x = "Time Since 2020 (Years)",
       y = "Total Funding (USD, Billions)",
       color = "Vertical") +
  theme_minimal()
```

```
# Find the Country/Region with the Most Growth in Funding
region_trends <- techfund_top |>
  group_by(time_since_2020, RegionClean) |>
  summarise(total_funding = sum(`Funding Amount (USD)`, na.rm = TRUE) / 1e9,
            .groups = 'drop')

region_time <- ggplot(region_trends, aes(x = time_since_2020,
                                         y = total_funding,
                                         color = RegionClean)) +
  geom_line(linewidth = 0.5) +
  geom_point() +
  scale_x_continuous(labels = scales::label_number(accuracy = 0.01)) +
  scale_y_continuous(labels = label_number(suffix = "B", accuracy = 0.01)) +
  labs(title = "Total Funding Over Time by Region",
       x = "Time Since 2020 (Years)",
       y = "Total Funding (USD, Billions)",
       color = "Region") +
  theme_minimal()

print(vertical_time + region_time)
```



Analysis of Total Funding Over Time by Vertical/Region

The line graphs reveal evolving funding dynamics since 2020. AI, Finance, and Cloud Computing demonstrate consistent, substantial growth, indicating strong investor confidence. In contrast, Cybersecurity and Education show fluctuating trends, possibly reflecting market uncertainty. Transportation experienced a big surge in funding during 2021, particularly in Australia, suggesting regional leadership in this sector. Healthcare showed steady growth but unexpectedly declined in 2021 despite the ongoing COVID-19 pandemic. On the regional front, the United States maintains investment leadership, while Australia, the UK, and India show increasing momentum, pointing toward a gradual diversification of global tech investments.

Summary of Findings

The analysis of tech funding distribution highlights key trends in investment patterns. The United States dominates the global technology investment landscape, far outpacing other regions, though emerging markets like Brazil and Israel are gaining traction. Established hubs such as Australia, the UK, and India continue to attract significant capital. In terms of sectors, B2B Software, Transportation, and Finance lead in funding, reflecting strong investor confidence in infrastructure and enterprise solutions. The substantial investments in artificial intelligence, blockchain, and cloud computing demonstrate a focus on disruptive technologies, while relatively lower funding for cybersecurity, education, and food & beverage suggests these sectors may be undervalued despite their potential long-term significance.

Despite these insights, the analysis has certain limitations. The dataset may not fully capture all funding events, leading to potential gaps in regional or sector representation. Additionally, the focus on total funding amounts overlooks nuances such as investment returns, startup success rates, and long-term sustainability. External factors like government policies, economic fluctuations, and geopolitical risks also influence funding distribution, especially since the data was tracked during 2020-2021 which was during the COVID-19 pandemic. These constraints highlight the need for a more comprehensive approach that integrates qualitative factors alongside financial trends.

Future research should investigate strategies for emerging markets to attract greater investment capital and examine whether currently underfunded sectors like cybersecurity and education will gain investor attention. The limited funding in cybersecurity presents a particular paradox given the accelerating digitalization across industries and the growing need for robust consumer protection measures. Further examination of regulatory frameworks, shifts in investor sentiment, and technological breakthroughs would provide valuable insights into the forces driving funding allocation patterns and could help predict future investment trends.

Citation

“Bansal, Shivam. ‘Tech Company Fundings Data Exploration.’ Kaggle, 2021, https://www.kaggle.com/code/shivamb/tech-company-fundings-data-exploration.”