

# KE5106 – DATA WAREHOUSING FOR BUSINESS ANALYTICS

---

*Singapore Research and Development –  
Analysis of funds allocation & outcomes*

---

Submitted by:

Abu Matthew Thoppan - A0178303H

Balaji Nataraj – A0178294N

Chokkalingam Shanmugasiva – A0178230J

Viknesh Kumar Balakrishnan – A0178304E

Yesupatham Kenneth Rithvik – A0178448M

# Table of Contents

|   |           |
|---|-----------|
| KE5106 – DATA WAREHOUSING FOR BUSINESS ANALYTICS..... | 0         |
| <b>Introduction .....</b>                             | <b>2</b>  |
| <b>Business Understanding.....</b>                    | <b>2</b>  |
| <b>Business Goals .....</b>                           | <b>2</b>  |
| <b>Questions.....</b>                                 | <b>3</b>  |
| <b>Metrics.....</b>                                   | <b>3</b>  |
| <b>Target Audience .....</b>                          | <b>3</b>  |
| <b>Insights &amp; Decisions .....</b>                 | <b>3</b>  |
| <b>System Architecture.....</b>                       | <b>4</b>  |
| <b>Data Collection.....</b>                           | <b>5</b>  |
| <b>Technical Approach.....</b>                        | <b>7</b>  |
| <b>Extract-Transfer-Load.....</b>                     | <b>7</b>  |
| <b>Database Design .....</b>                          | <b>8</b>  |
| <b>Dashboard .....</b>                                | <b>10</b> |
| <b>Researchers .....</b>                              | <b>11</b> |
| <b>Expenditure .....</b>                              | <b>14</b> |
| <b>Research Outcome Metrics.....</b>                  | <b>17</b> |
| <b>Conclusion.....</b>                                | <b>19</b> |

## Introduction

Singapore is at the forefront of research and innovation. The recent Global Innovation Index 2018 ranks Singapore 5<sup>th</sup> in the world in terms of innovation and research. Research, innovation and enterprise are cornerstones of Singapore's national strategy to develop a knowledge-based innovation-driven economy and society. As part of this strategy, the government will be sustaining its commitment to research, innovation and enterprise, and will invest \$19 billion for the RIE2020 Plan over 2016 to 2020. The RIE2020 plan will be executed by the National Research Foundation (NRF) which will be source of funding and grants for the researchers and research institutes across Singapore. The funds will capitalise on the progress made in public sector research and help build industry R&D capabilities and innovative enterprises. It therefore becomes necessary to ensure proper management of the funds allocated towards research. Through proper planning and effective implementation, Singapore can secure its future in the forefront of R&D internationally.

## Business Understanding

Since research is a field in which the desired outcome is often not obtained, and success comes only after multiple failure, fund estimation and allocation is a problem. Usually funds are allocated towards research with the concept that even the few successes will contribute vastly to improving the society. But due to constraints on budget, it is difficult to blindly invest into research hoping for success. It is important to plan the investments and measure the success obtained to make best use of the funds. The main investments that go towards research are the monetary expenditure and the human resources. Research outcomes are difficult to measure accurately but measures like the number of patents, research papers and revenue obtained due to the research act as suitable indicators of success. It is therefore the responsibility of NRF to manage the inputs and maximize outcomes of research.

## Business Goals

Based on the business understanding, we can identify the following business goals:

1. Analyse the current state of research in Singapore
2. Identify areas to improve fund allocation and human resources to maximize research outcomes

## Questions

Answering the following questions will help achieve the business goal:

1. What is expenditure on R&D over years?
2. What is the strength of human resources in R&D over years?
3. Which areas have obtained most successful outcomes (patents, revenue)?
4. Can efficiency of inputs to outcome be improved? In which sectors?
5. How will efficiency change in the future?
6. How can fund allocation be modified to obtain better outcomes?

## Metrics

1. Headcount of human resources working in R&D
2. Expenditure on research
3. Number of patents applied and awarded
4. Licensing and sales revenue obtained through research
5. Measures of efficiency:
  - a. Patent conversion ratio:  $\text{Patents awarded} / \text{Patents Applied}$
  - b. Gross profit per patent:  $(\text{Revenue} - \text{Expenditure}) / \text{Patents awarded}$
  - c. Patents per 100 researchers:  $(\text{Patents awarded} / \text{No. of researchers}) * 100$

## Target Audience

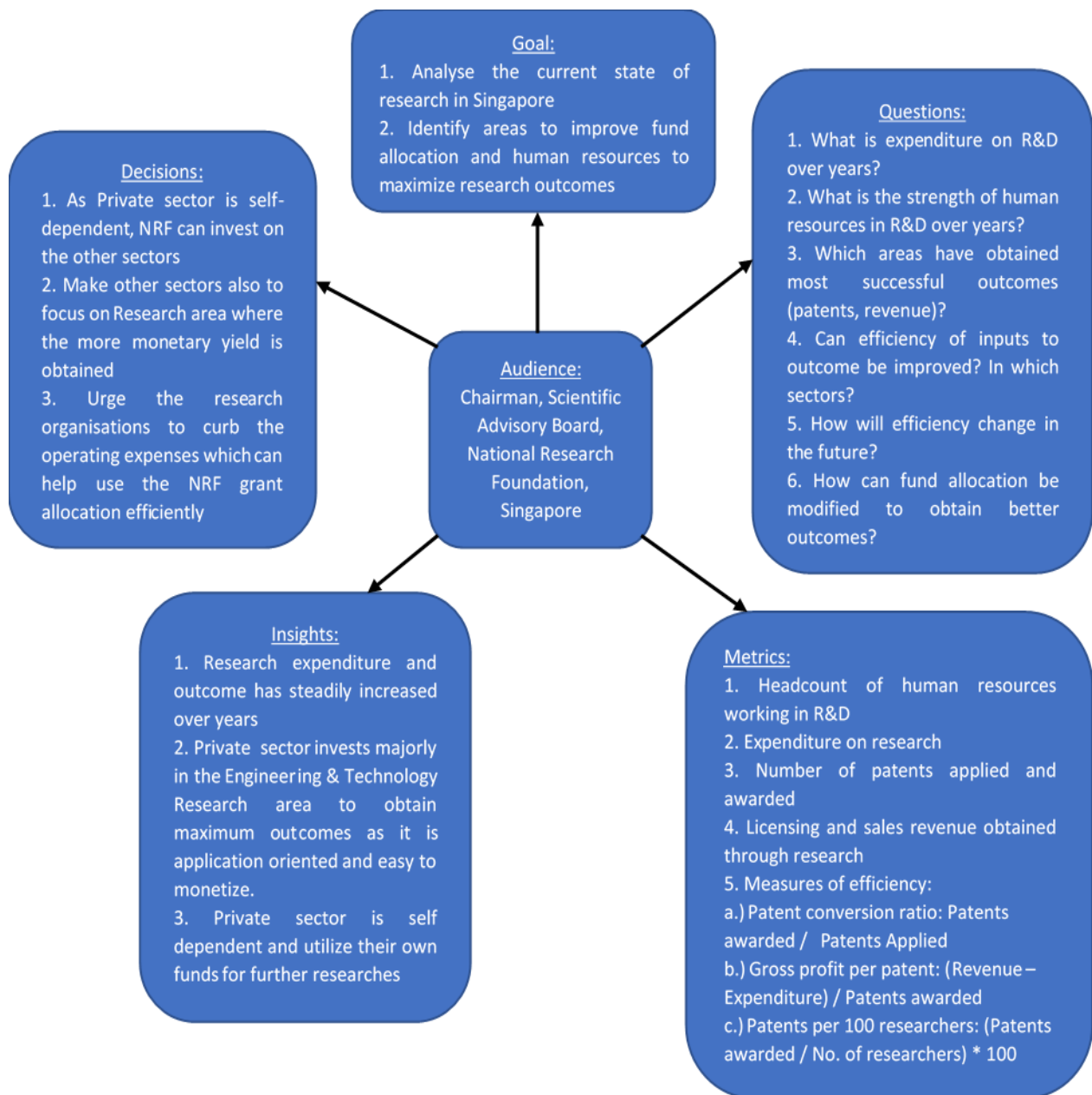
The NRF has a Scientific Advisory Board (SAB) that consists of a multi-disciplinary international board that convenes annually to advise on NRF's policies and programmes. One of its roles includes assisting and advising NRF on the management of R&D, including the allocation of funding and the assessment of research outcomes. We target the Prof Sir Richard Friend, Chairman of SAB as the main audience for this project.

## Insights & Decisions

The following insights and decisions can be obtained:

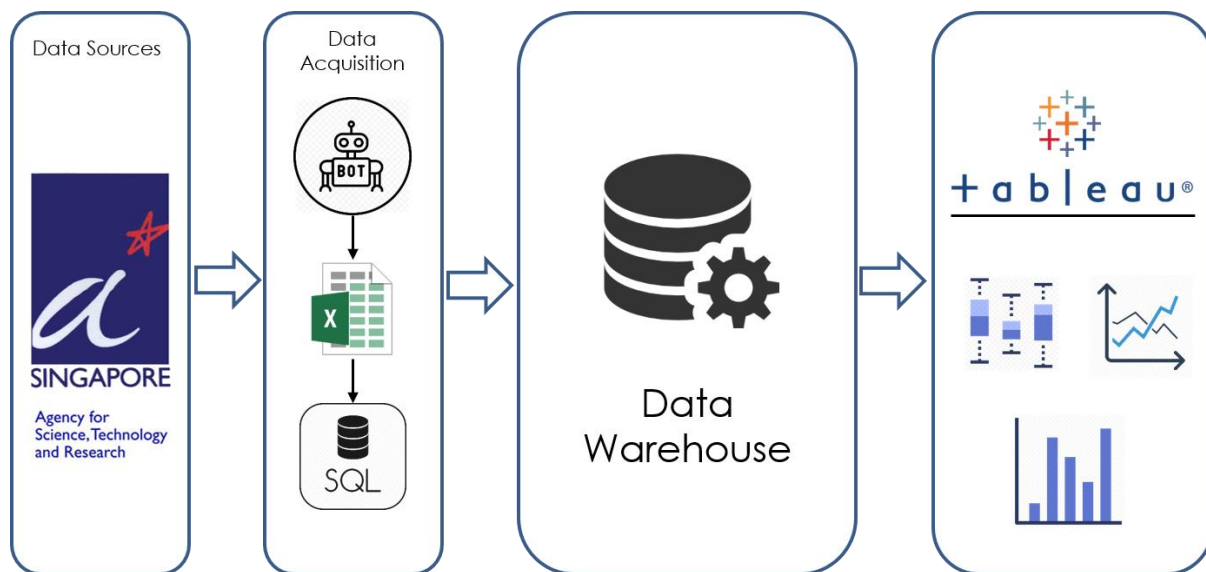
1. Research expenditure and outcome has steadily increased over years
2. Private sector manages to obtain maximum outcomes as most of their research is in engineering and technology which is application oriented and easy to monetise
3. Private sector mostly depends on its own funds; therefore, the government can afford to reduce funds towards the private sector and redistribute across other sectors

Government, private and higher education sectors have improved across years, but public research institutes haven't shown much improvement in terms of efficiency



## System Architecture

The architecture of our system is described visually below. We use A-Star reports as our data source, later we use tools to convert these reports to a spreadsheet form. We modify the granularity of these tables to match each and load them to a data-base in a normalized way. We later create data-warehouse out of this database and supply data to our visualization tool, Tableau.



## Data Collection

For analysing the spend of funds on research and development, and the resulting output generated by them, we need an adequate and reliable data source to work upon. After searching the web, we have decided to use The National Survey of Research and Development published by A-Star annually [here](#).

This report contains data about the investment of funds in research and development by the different sectors, area in which research was done, amount of resources used, the intellectual property gained through said research, revenue generated by the research and also describes the diversity of the people carrying out this research.

We decided to work on data from the last seven years available on the website that is from 2010 to 2016. This can give us a good trend analysis of how the funding dynamics have changed over the years while also helping us forecast and prepare for subsequent years to come.

The tables that we used for this analysis were: -

| Table                       | Description   |
|-----------------------------|---|
| R&D Manpower                | Gives the no. of people categorised by their qualification level and the sector that they are currently working in.   |
| R&D Manpower by Nationality | Gives the no. of people categorised by their qualification level and the sector that they are currently working in, while also indicating whether they are Singapore citizens/PR's or not |

|  |   |
|--|---|
| R&D Manpower by Age Group  | Gives the no. of people categorised by their qualification level and the sector that they are currently working in, while also categorizing their age group range   |
| R&D Manpower by Gender   | Gives the no. of people categorised by their qualification level and the sector that they are currently working in, while also the gender of the researchers.   |
| R&D Expenditure by Type of Costs   | Gives the expenditure in million dollars(\$\$), spent by the different sectors and the type of cost incurred  |
| R&D Expenditure by Source of Funding   | Gives the expenditure in million dollars(\$\$), spent by the different sectors and the source from which the fund was obtained  |
| Patenting Indicators   | Gives the breakup of the patents applied, awarded and owned by the different sectors.   |
| Revenue Indicators   | Gives the type of revenue earned by the different sectors   |
| Researchers by Field of Science & Technology   | Gives the researchers categorised by the sector they are working in and their educational qualification, while also the field of science and technology they are working in.  |
| Private Sector Researchers by Enterprise Ownership/Size and Field of Science & Technology                  | Gives the researchers categorised by the enterprise ownership/size of the organization that they are working in and their educational qualification, while also indicating the field of science and technology they are working in. |
| R&D Expenditure by Type of R&D and Field of Science & Technology   | Gives the expenditure categorized by the sector, field of science and technology and also the type of research work that the money is being spent upon.   |
| Private Sector R&D Expenditure by Enterprise Ownership/Size, Type of R&D and Field of Science & Technology | Gives the expenditure categorised by the enterprise ownership/size of the organization, field of science and technology and the type of research work that the money is being spent upon.   |

Table 1. Tables that were used for analysis

The tables that were got from this report are not in a format that can be used readily for analysis. We first convert them into a spreadsheet format using an online tool found here, that uses OCR to convert pdf files to a spreadsheet format. The granularity of the data was later changed to be uniform across all tables to ensure uniform merges and comparisons while doing analysis. This

data was later loaded into a data-base management system and normalized to carry out further analysis.

## Technical Approach

### Extract-Transfer-Load

From the spreadsheet files generated from the pdf files we prepared the data to make it loadable to the data warehouse. We use MSSQL Server, which is a very scalable and robust database management system, to maintain our data warehouse. Even though we have tools like SSIS which can handle the ETL process efficiently and in ease with MSSQL Server, we used Microsoft excel to prepare the data. We chose to do so since there were many anomalies in the raw data converted by the OCR, which required manual attention.

The original data files provide the following data,

- Research & Development Manpower headcounts by sector, field of science & technology, nationality, age and gender per year
- Research & Development expenses by types of costs, field of science & technology and sources of funding per sector per year
- Research & Development Revenue from licensing & sales per sector per year
- Patents filed and awarded per sector per year

Since the original data is raw enough to load into a database, we had to do some pre-processing to change the structure of the data to fit it to a relational database. From the original data, we identified the key entities and normalised the data by splitting it in to reference and value tables as following.

### Reference Tables

1. Age Groups
2. Cost Types
3. Fields of Science & Technology
4. Funding Sources
5. Gender
6. Nationality
7. Researcher Types
8. Sectors

### Value Tables

1. Expenses by field




2. Head Counts
3. Revenue and patents by sector


We used the Data Import feature in the Microsoft SQL Server Management Studio to import the data into the database.

## Database Design


### AgeGroups

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | AgeGroup    | nvarchar(255) | <input type="checkbox"/> |


### CostTypes

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | CostType    | nvarchar(255) | <input type="checkbox"/> |


### Fields

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | Field       | nvarchar(255) | <input type="checkbox"/> |


### FundingSources

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | Source      | nvarchar(255) | <input type="checkbox"/> |

### Gender

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | Gender      | nvarchar(255) | <input type="checkbox"/> |

### Nationality

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
|  | Id          | int           | <input type="checkbox"/> |
|   | Nationality | nvarchar(255) | <input type="checkbox"/> |

## **ResearchTypes**

|   | Column Name        | Data Type     | Allow Nulls              |
|---|--------------------|---------------|--------------------------|
| 🔑 | Id                 | int           | <input type="checkbox"/> |
|   | ResearcherCategory | nvarchar(255) | <input type="checkbox"/> |

## **Sectors**

|   | Column Name | Data Type     | Allow Nulls              |
|---|-------------|---------------|--------------------------|
| 🔑 | Id          | int           | <input type="checkbox"/> |
|   | Sector      | nvarchar(255) | <input type="checkbox"/> |

## **ExpensesByField**

|   | Column Name     | Data Type | Allow Nulls                         |
|---|-----------------|-----------|-------------------------------------|
| 🔑 | RowNumber       | int       | <input type="checkbox"/>            |
|   | Year            | int       | <input type="checkbox"/>            |
|   | SectorId        | int       | <input checked="" type="checkbox"/> |
|   | CostTypeId      | int       | <input checked="" type="checkbox"/> |
|   | FundingSourceId | int       | <input checked="" type="checkbox"/> |
|   | FieldId         | int       | <input checked="" type="checkbox"/> |
|   | Expenditure     | float     | <input type="checkbox"/>            |

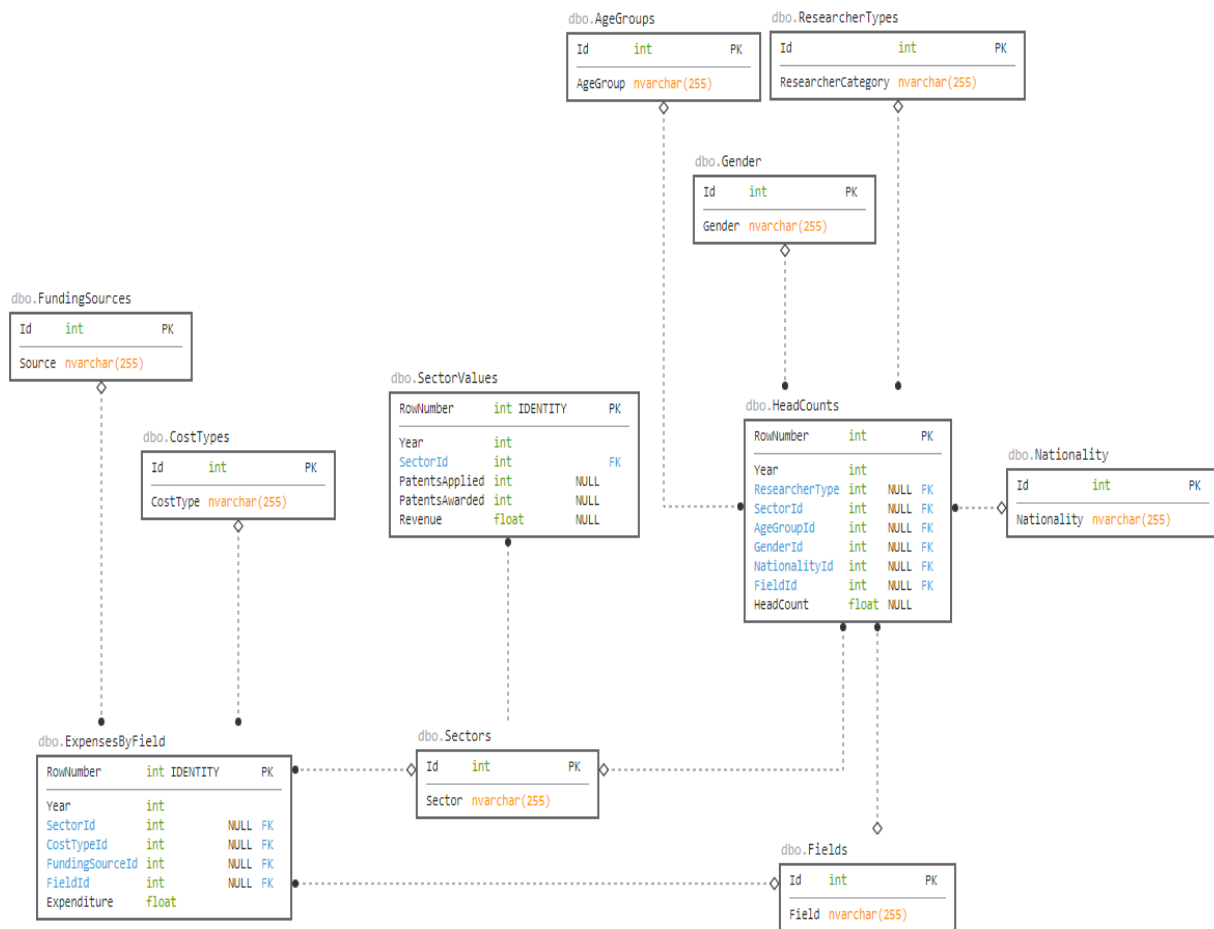
## **HeadCounts**

|   | Column Name    | Data Type | Allow Nulls                         |
|---|----------------|-----------|-------------------------------------|
| 🔑 | RowNumber      | int       | <input type="checkbox"/>            |
|   | Year           | int       | <input type="checkbox"/>            |
|   | ResearcherType | int       | <input checked="" type="checkbox"/> |
|   | SectorId       | int       | <input checked="" type="checkbox"/> |
|   | AgeGroupId     | int       | <input checked="" type="checkbox"/> |
|   | GenderId       | int       | <input checked="" type="checkbox"/> |
|   | NationalityId  | int       | <input checked="" type="checkbox"/> |
|   | FieldId        | int       | <input checked="" type="checkbox"/> |
|   | HeadCount      | float     | <input type="checkbox"/>            |

## **SectorValues**

|   | Column Name    | Data Type | Allow Nulls                         |
|---|----------------|-----------|-------------------------------------|
| 🔑 | RowNumber      | int       | <input type="checkbox"/>            |
|   | Year           | int       | <input type="checkbox"/>            |
|   | SectorId       | int       | <input type="checkbox"/>            |
|   | PatentsApplied | int       | <input checked="" type="checkbox"/> |
|   | PatentsAwarded | int       | <input checked="" type="checkbox"/> |
|   | Revenue        | float     | <input checked="" type="checkbox"/> |

The Entity-Relationship diagram for the data warehouse is as following.

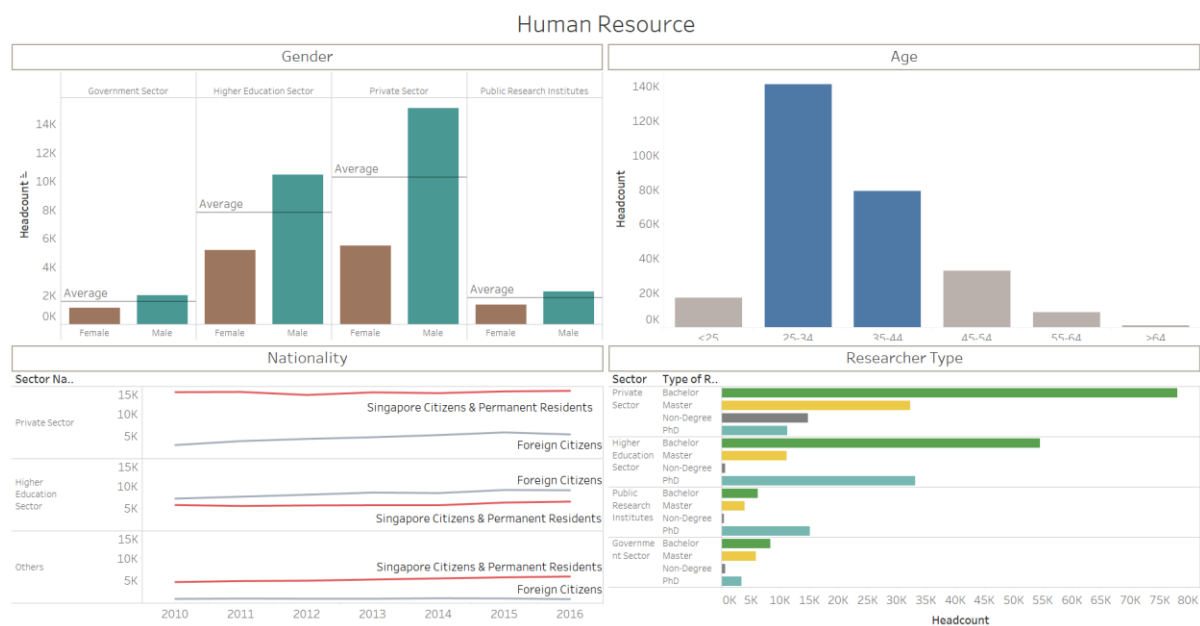


## Dashboard

In total, we have 3 dashboards that depict the utilization of human resources, expenditure incurred to carry out the researches and finally the metrics that can be helpful in evaluating the performance of the research works. To create the dashboard, we have utilized various dimensional data ranging from the year 2010 to 2016. This can help management personnel or director by providing several insights influenced by various factors to take appropriate business decisions at any moment of time. Below are the dashboard details:

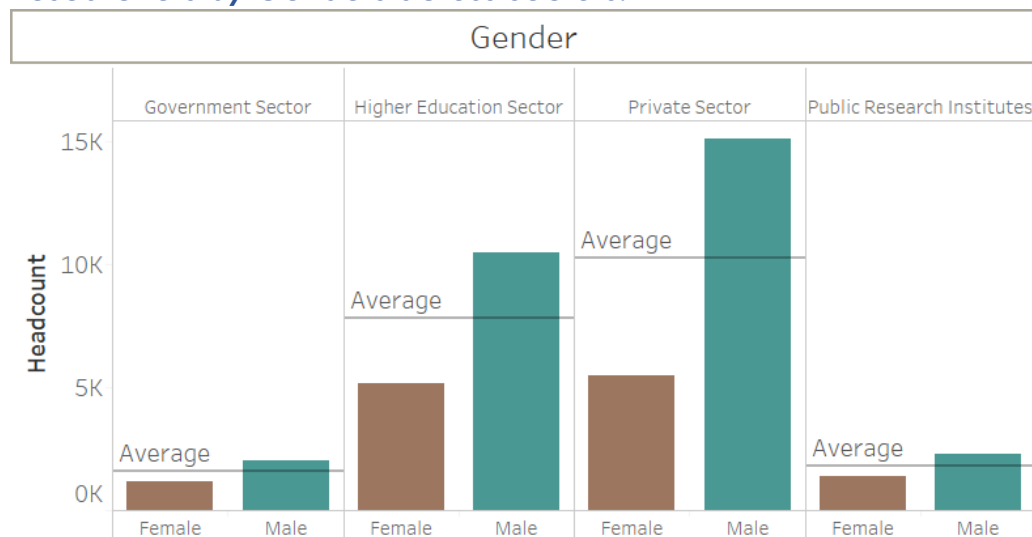
1. Dashboard-1: Researchers
2. Dashboard-2: Expenditure
3. Dashboard-3: Research Outcome Metrics

## Researchers



This dashboard visually shows all the current statistics related to the researchers and how they are distributed across various groups as per the data from the Singapore government.

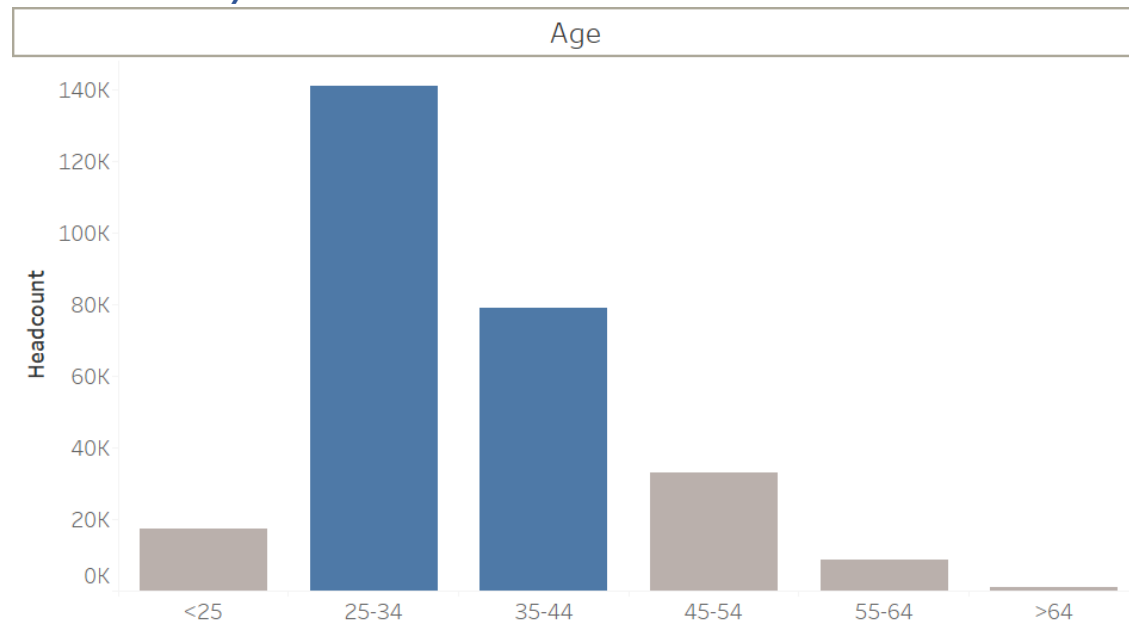
### Researchers by Genders across Sectors:



Here we did a comparison of researchers based on the gender and it reveals that number of female researchers are very low in comparison with the male researchers overall. As we all know, research involves intense amount of human effort which may have created disinterest to female researchers but still we need more concrete data to conclude on that. We have used bar graph here to show the level of changes in the number of researchers in

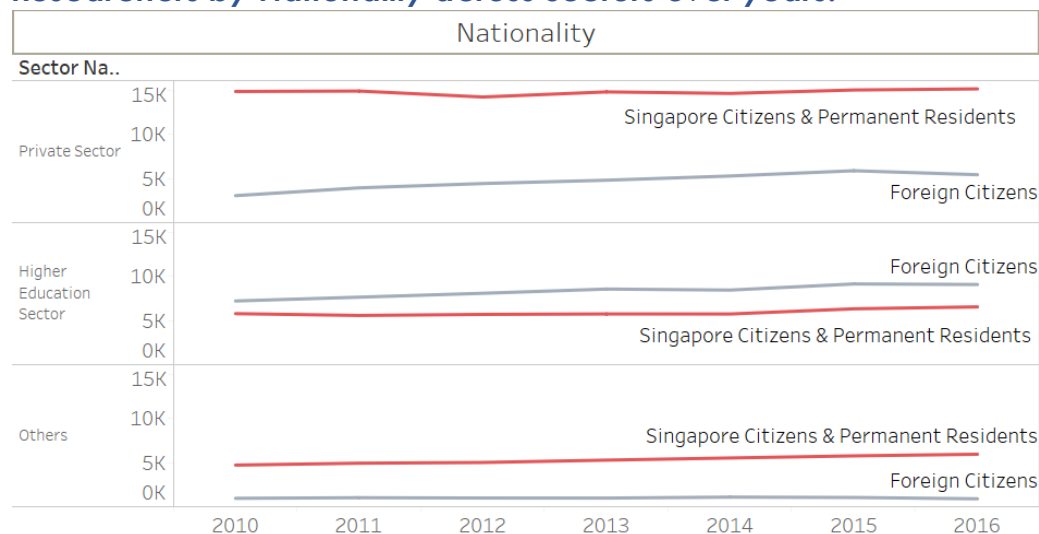
each sector i.e. Government Sector, Private Sector, Public research institutes and Higher Education institutes.

### Researchers by Genders across Sectors:



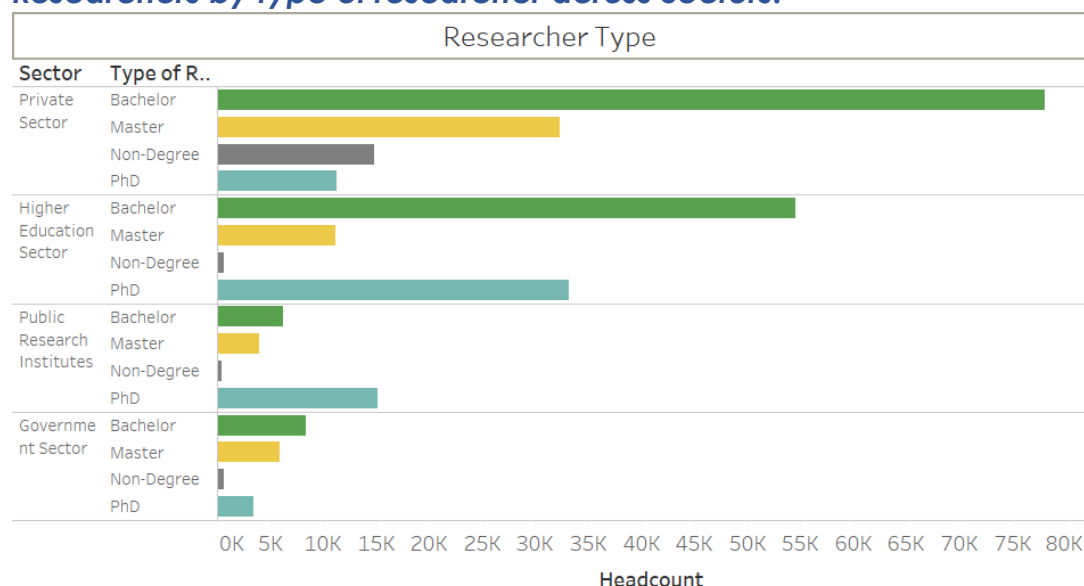
As like gender, we had the split up of researchers count for different age groups which represent that most of the researchers are from mid age group i.e. 25 to 44. But this graph would not be constant and would change every year as the volume distribution of one age group would propagate to other age group to make the graph as skewed.

### Researchers by Nationality across Sectors over years:



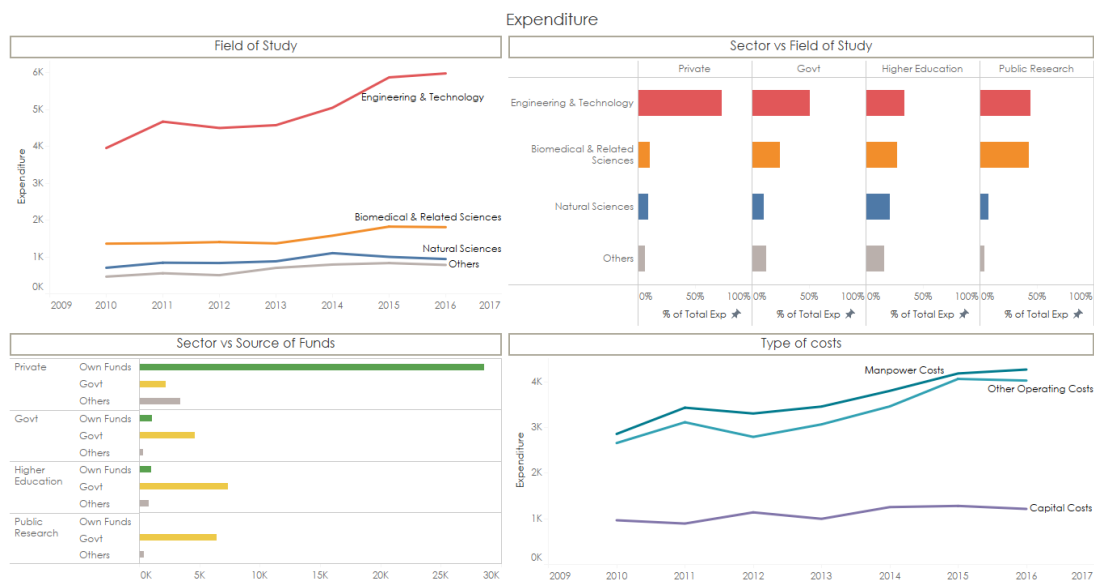
As far as nationality is concerned, we obtained the data which had 3 categories namely Singapore Citizens, Singapore Permanent Residents and Foreign Citizens. Here, we had grouped the sectors: Government and Public Research Institutes as Others. As Public research institute consists only A\* which inturn is also an government-aided organisation and also to get a proper visualization. In general, we would think that the country's own nationals would outnumber the foreign nationals. This is true in most cases but only in higher education sector, we can spot an unusual scenario where most of the researches were undertaken by foreign nationals. This could be also due the popularity of the educational universities in Singapore (eg: NUS, NTU etc.) which attracts students from various part of world to pursue their higher education here and also the the facility and standard provided for the students to perform their research. We have used line graph to show the changes over the years and differentiated the nationality using colours.

### Researchers by type of researcher across Sectors:



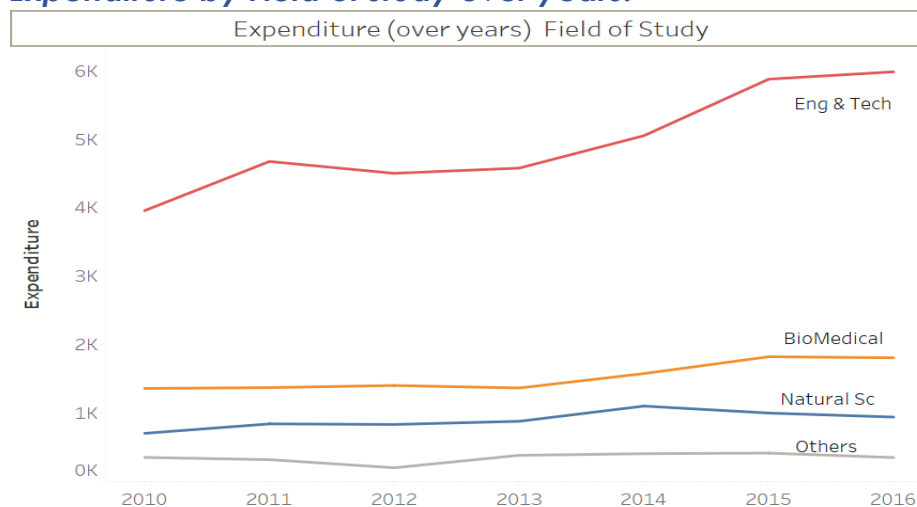
As far as research and development in Singapore, most of the contributions comes from Private sector organisations which is also evident by the number of researchers which is far ahead than the other sectors (also showed in the gender bar graph). Here this bar chart shows the different types of researchers (based on the qualification) involved in the various sectors. General myth is that researchers are basically PhD's. Yes, it is true to some extend in the higher education and public research institute sectors. Whereas across this island nation, most of the researches are undertaken by the bachelors in most of the sectors. Only PhD's are preferred the most in Public research institutes as they focus to improve the quality of life of the people in the country.

## Expenditure



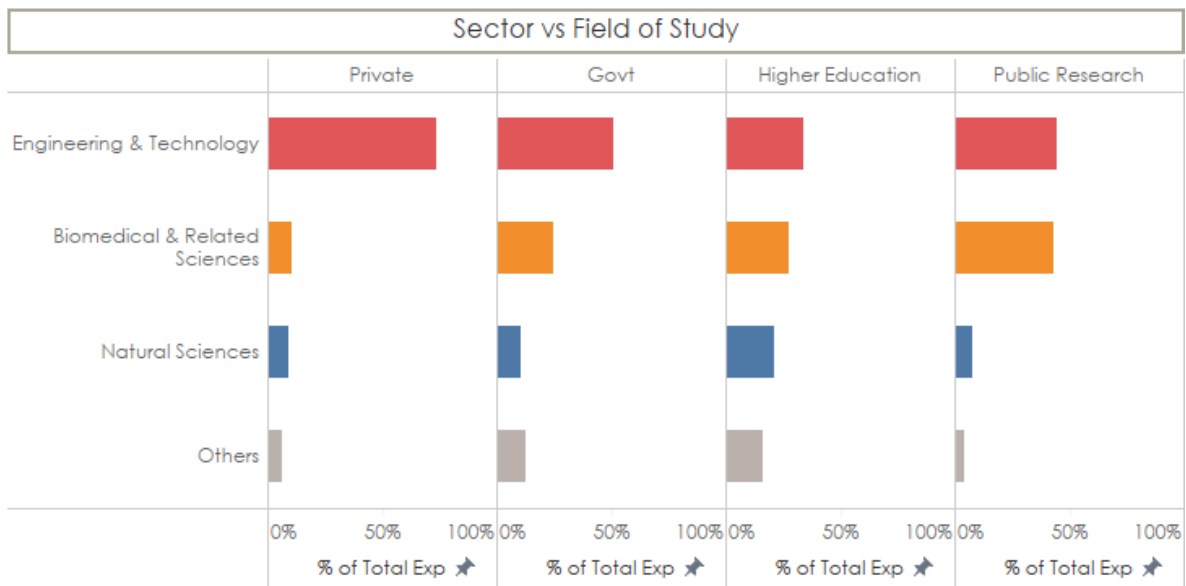
This dashboard visually shows all the expenditure related to the researchers and their research works across various groups as per the data from the Singapore government.

### Expenditure by Field of study over years:



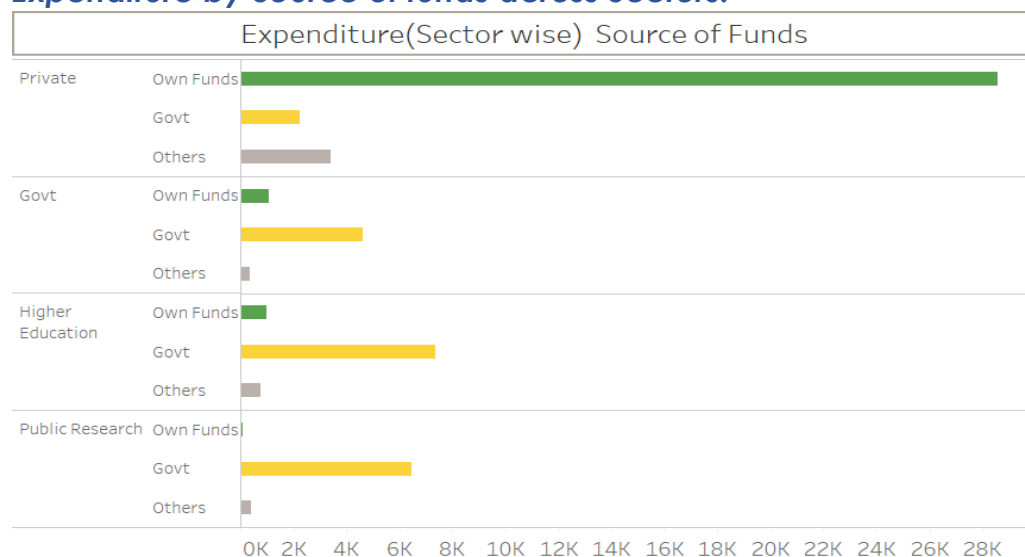
We have used line graph to depict the research expenditure over the years (2010 to 2016). As per the data source, we had 6 research areas namely: Engineering & Technology, Biomedical, Natural Science, Agricultural & Food Sciences and other areas. Singapore invests more in the Engineering & Technology area when compared to other research areas. We have displayed only the top 3 research areas in terms of expenditure and grouped the other research areas into one.

### Expenditure by Field of Study across Sectors:



In the previous dashboard, we know that most of the research work was contributed by the private sector which was evident by the number of researchers and the expenditure spent on research work. To have a deeper look to understand which research area were invested by each sector, we used horizontal bar graph. Every sector invests heavily on Engineering & Technology research area, but a major callout is that, 73% expenditure of private sector's money is spent on this area whereas other sectors proportionally invest in other research areas. One possible reason could be that private sector is driven towards profit whereas others are not.

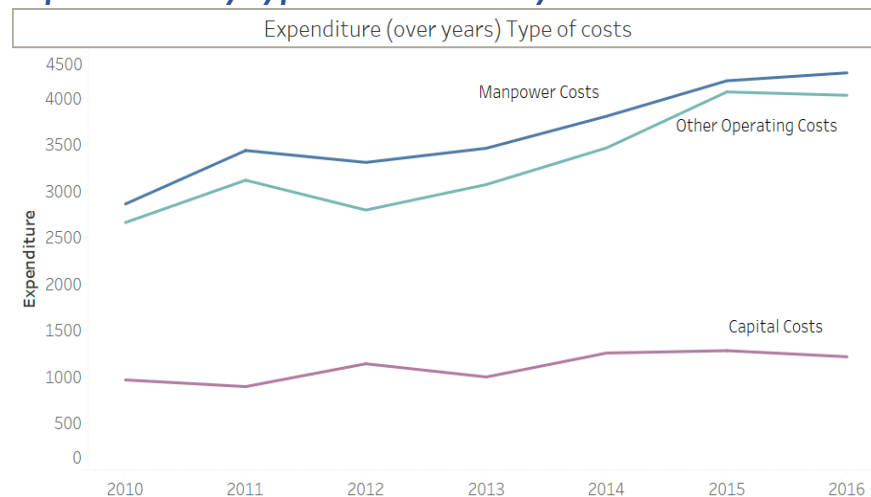
### Expenditure by Source of funds across Sectors:





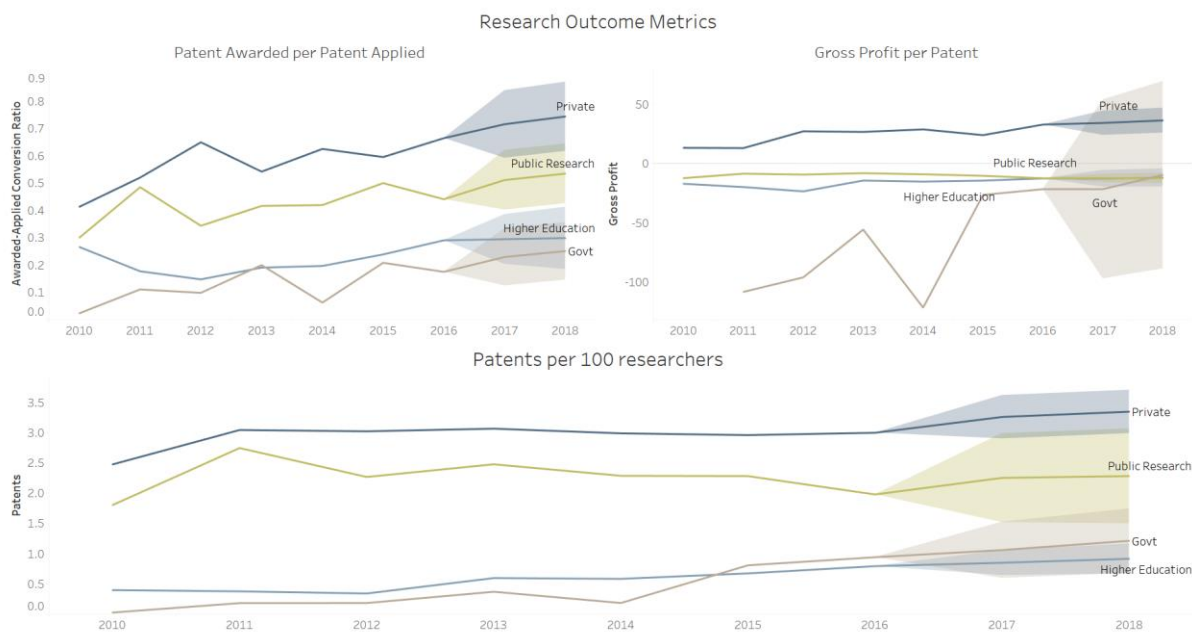
Here again, Horizontal bar graph is used to display the source of funds for the expenditure across sectors and private sectors are self-dependent as they mostly utilize their own funds to spend further. Whereas other sectors are funded majorly through government (Ex: NRF) and by this visualization, government can curb the fund allocation towards private sectors and can focus more on other sectors.

### **Expenditure by type of costs over years:**



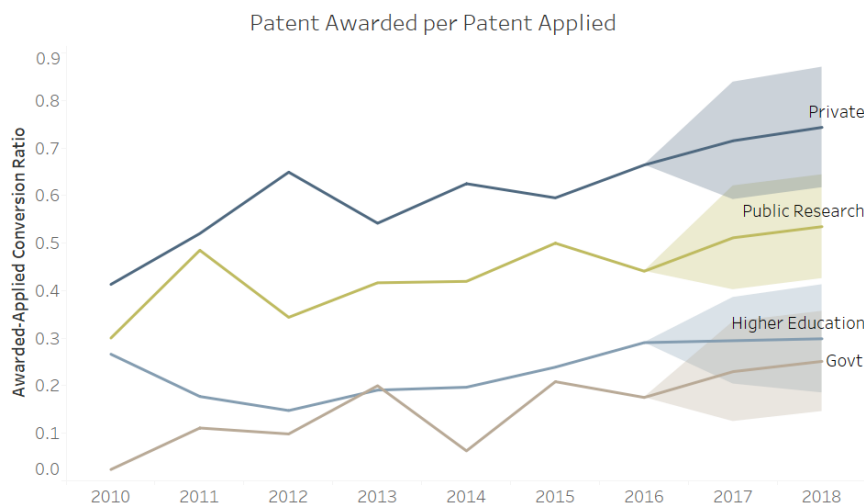
Research expenditure consists of 3 major costs: Manpower costs, Capital costs and other operating costs. Manpower costs is the amount spent on the human resources or the researchers. Capital cost is the amount spent on setting up a research lab or institute or workplace. This line graph explains that the expenditure can go up if the number of researchers/manpower are increased and only way to spend efficiently is by reducing the Other operating costs which also has a heavy impact on the total expenditure of the organisations.

## Research Outcome Metrics



This dashboard shows the various measures of efficiency in obtaining the research outcomes. This can be used to identify the change in efficiency in different sectors over years.

### Patent conversion ratio across sectors over years:

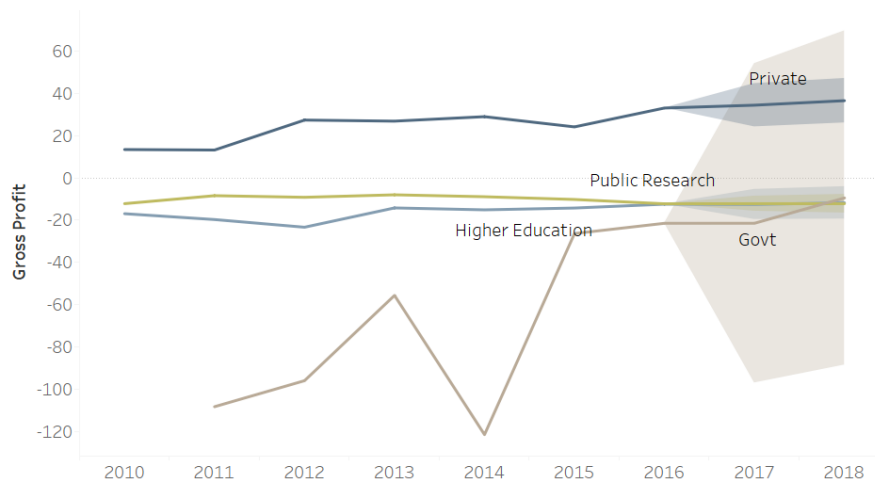


The patent conversion ratio shows the ratio of number of patents awarded to number of patents applied. This measures efficiency of the work done as the work done on unawarded patent is poor utilisation of resources and time. A line chart is used to show the change in the conversion ratio over years and different colours are used to differentiate between sectors. It is seen that all sectors have shown gradual improvement in conversion ratio over years and private sector has the best ratio. The chart also includes a forecast of the

ratio for the next two years which predicts that the sectors will continue to improve.

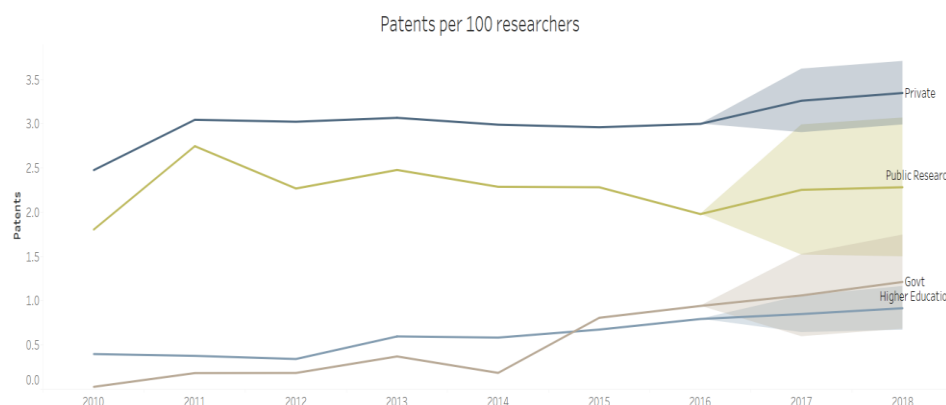
### Gross profit per patent across sectors over years:

Gross Profit per Patent



Gross profit per patent shows the effective gross profit, i.e. total revenue on licensing and sales of patents minus the total expenditure on research, for each patent in a sector over years. This measures the monetary impact of working on each patent. A line chart shows the measure over years and different colours indicate sectors. It is seen that only private sector shows a positive profit per patent which means that for each patent produced, the private sector earns revenue. But the other sector shows negative profitability. While private, public and higher education sectors have remained static over years, the government sector has had varying profits with the most recent years pushing its profits close to public and higher education sector. However, due to the high variability of the government sector, its value cannot be predicted accurately for the future years.

### Patents per 100 researchers across Sectors over years:



Patents per 100 researchers shows the effective numbers of patents contributed towards by every 100 researchers in every sector across years. This gives a measure of human resource required to produce a patent. A line chart shows the measure over years and colours differentiate between the sectors. The private, government and higher education sectors have shown considerable improvement in this, especially the government sector having overtaken the higher education sector. The public research institutes however have seen a decline in this measure. The forecast for every sector shows improvement in future years.

## Conclusion

From the analysis, it is seen that private sector is the best in terms of all efficiency metrics. One of the reasons is that the private sector is self-sufficient with 86% of own funding. All the other sectors depend heavily on government funding. Private sector also invests heavily in the field of engineering and technology, almost 73% of the expenditure. Other sectors proportionally invest in other research areas as well. This contributes to the high performance of private sector as application research in the engineering field is relatively more fruitful than other sectors. Another factor is that significant part of the research in the private sector is carried out by undergraduates and non-degree (interns), but the other sectors depend heavily on post-graduates and PhDs, who are paid more.

In terms of efficiency, we looked at patent conversion ratio, gross profit per patent and patent per 100 researchers. Private sector leads in all measures of efficiency and government sector has shown significant improvement over years. Only private sector shows positive profitability which can be attributed to its application-oriented research in the field of engineering and technology.

Finally, some of the key decisions that NRF needs to take are:

1. As Private sector is mostly self-dependent, NRF can choose to invest in the other sectors
2. Create motive for other sectors also to focus on research areas where high monetary yield is obtained by providing more funds in those research areas
3. Urge the research organisations to reduce the unnecessary operating expenses which can help use the NRF grant allocation efficiently