ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Data Visualisation Process Book

by

Selmane Kechkar Ayoub

Bilguun Chinzorig

Olagoke Lukman Olabisi

submitted

in the
Data Visualization Course
Course Code: COM-480

Lecturer: Dr. Benzil Kirell
Teaching Assistant: Volodymyr Miz

December 2017

The purpose of visualization is insight, not pictures.

Ben Shneiderman

Abstract

This processbook describes the visualization project on the patent data set[1] . Generally, our aim is in two-folds. First, to analyze the data set to obtain insight and extract information. Second, building on step one, we build the right visualization idiom to encode the information in an interactive and engaging way. To achieve these aims, we start by carrying out statistical analysis to understand the semantics of the data set and the corresponding metadata. Next, we generate any relevant derived attribution that can help summarize the data. Lastly, we iteratively apply the right visualization libraries that we have seen in class, to test repeatedly and fine-tune different visual representation concurrently. The result of this technique-driven approach is a visualization that provides the correct data abstraction and has broad awareness of the information space; while also being interactive and providing the right information depth.

Contents

Abstract **ii**

1 Introduction: Project Overview and Design Considerations **1**

1.1 Overview And Target 1

1.1.1 Motivation 2

1.1.1.1 What Are We trying to Show? 3

1.1.1.2 Related Inspiration? 3

1.2 Design Consideration 3

1.2.1 The Big Picture: The What-why-Analysis loop 5

1.2.1.1 Domain Situation 5

1.2.1.2 Task Abstraction And Visual Encoding: The Why-What-how loop 5

1.2.1.3 Visual Encoding: How? 6

1.2.1.4 Algorithm 6

1.3 Conclusion 7

2 Exploratory Data Analysis **8**

2.1 Introduction 8

2.1.1 First Look into the Data set 8

2.1.2 Exploratory analysis 9

2.1.3 Map Data Preparation 11

2.2 Conclusion 12

3 Patent Impact **13**

3.1 Introduction 13

3.2 Defining Academic Impact 13

3.3 Defining Economic Impact 14

3.4 Relationship between economic and academic impact 14

4 Geospatial Map **16**

4.1 Introduction 16

4.1.1 Data set Preparation 16

4.2 Conclusion 18

5 Orogonality Ranking **19**

5.1 Overview And Target 19

5.1.1 Cleaning Pipeline 19

5.1.2 Look and Feel of Visualisation Idiom 20
5.2 Peer Assessment 22
5.3 Conclusion 23

Bibliography **24**

Chapter 1

Introduction: Project Overview and Design Considerations

1.1 Overview And Target

In general the aim is to come up with interactive, engaging, user-friendly and fun visual idioms after careful exploratory data analysis on the patent data set.It is obvious that given the size of the data set, the normal human mind would be unable to fathom any meaning in the data set. However, by enlisting the power of computation, we can augment the human perception system. Thus our target is to help users make meaning of the data set. This does not mean that we sacrifice depth for simplicity but rather we try to optimize on both.

In summary, we intend to embed any derived information in a geo-spatial context (maps). From this abstraction we can get some geospatially relevant information which might, at first, not be obvious to the un-aided human eye. In addition, We intend to go further to discover to derive the impact of patents on fortune company revenue (interesting!). Lastly, we were curious how the originality score of patents for different countries for best performing countries - in terns of patents count.

1.1.1 Motivation

The motivation for this project stems from the fact that different countries record different patent number for each year. We are thus motivated by the following thought provoking questions:

First from the analysis point of view we will focus on:

1. Is there a trend for the number of patents applications during the years?
2. Which countries have the most assigned patents?
3. What are the most popular technology fields for patents assigned in the last 5 years?
4. What's the most frequent technology field by inventor's country?
5. What percentage of patents belongs to private and what to organizations?
6. Are the citations inside each patent and the citations to a patent increasing with passing of the years?
7. How the number of citations relates to the category of the patent?

From the visualization perspective we ask the following questions:

1. Data types appropriate for the representation of results from analysis (we will consider using a tree , Network or geometry or a combination of geometry and networks.
2. Possible actions for interacting with the Viz: Search, Query or Consume, Explore, Navigate.
3. visualization targets: Trends, Outliers, Features, Correlation etc
4. Spatial information of patents (spatial data e.g country locations)
5. Iteratively do what-why-how analysis on each visualization idioms proposals and optimize for efficiency

1.1.1.1 What Are We trying to Show?

In general, we want the to show case a visualization that answers some of the questions that motivated the project - enumerated above. However, we aim to do this in an interactive and easy-to-use way. In achieving this we want each user to be able to understand and gain meaningful insight on the data set instantly .

We were able to answer 90 percent of the above questions that motivated us. Proposals on ways to improve this visualization will be discussed in conclusion section.

1.1.1.2 Related Inspiration?

The motivation for our visualization implementation comes from various online sources, Stackoverflow and the D3JS git repo. In particular for the visualization on patent impact we took insporation from [2]. In addition for the Timeline ranking of countries - presented later - we took inspiration from [3].

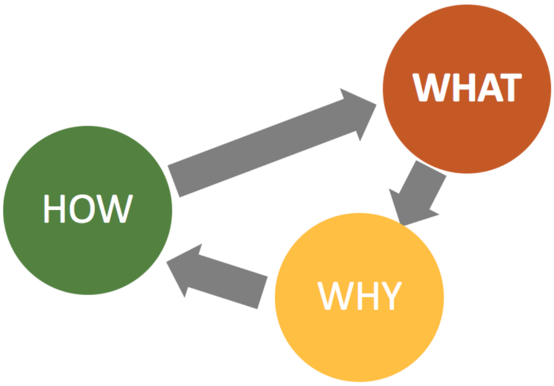
For the map graph our main reference is D3JS and StackOverflow.

1.2 Design Consideration

In this section we provide a discussion of the project from the perspective of the what-why-how fig framework 1.1. This framework will be embedded in an analytical validation approach fig 1.2 that encapsulates the entire visualization pipeline. The framework serves as a reference approach with which the project was built. Each of the visualization that has been realised will be analysed in the context of this framework or approach.

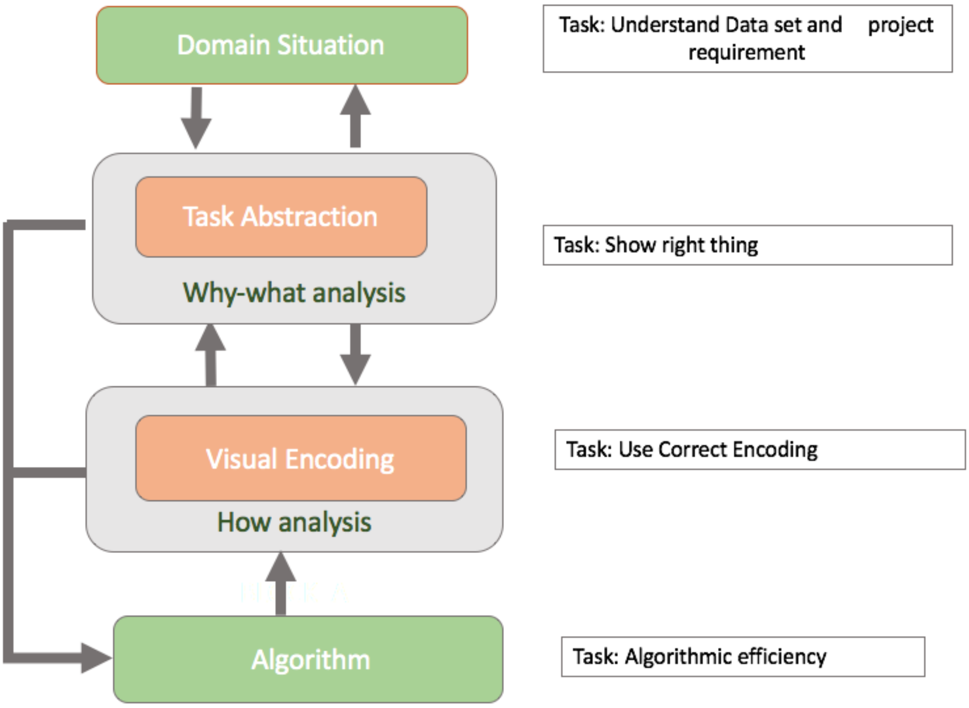
While we study related works our aim is to leverage the theoretical underpinnings from class to filter the various related works in a nice scientific framework. Then eventually we can come up with a solution that respects those dos and dont discussed in class in a scientific framework. This approach enable us to analyse not only the algorithm but the visual output.

Figure 1.1: WHAT-WHY-HOW FRAMEWORK



[Adapted from [4]]

Figure 1.2: ANALYTICAL VALIDATION FRAMEWORK



[Adapted from [4]. This is used]

1.2.1 The Big Picture: The What-why-Analysis loop

1.2.1.1 Domain Situation

The first line of approach for arriving at a suitable visualization starts with understanding the project requirement and the database section. Similarly, the data set consists of patents granted in the US. It was important that we understand the definition of columns of the data set. It was observed that there was the country of origin of inventor of patent, the patent ID and the originality of the patent. From this, it was possible to generate a derived attribution that holds the number of patents. In addition we could abstract the problem further to generate the correlation between number of patents and the company revenues. Furthermore, we could create another abstraction where we can rank countries by the originality of patents score. This completes the Domain situation. Further details will be provided in the appropriate section

1.2.1.2 Task Abstraction And Visual Encoding: The Why-What-how loop

Based on the domain situation discussed above , following proposals visualization idioms [4] were proposed:

1. Proposal 1: Use Geometry data type in addition with various marks and channels to generate the right spatial information about the data. This is essentially a chloropleth: but more interactive. Use Network data set type to encode which country cite more from the other

2. Proposal 2 : Using geometrical element such as lines (of varying area, shape , color , hue and saturation) for the realization of the relationship between citations and revenues of companies. The company dta

3. Proposal 3: to provide a time-line ranking of countries (with highest patent count) based on the yearly originality score of their patents. Thus can be achieved using appropriate line channel to follow a the path , that connects the country rank for each year based on originality score.

The target of the last proposal is for enjoyment, presentation of facts and summary of the information. [4]. It also afford users the ability to compare the countries based on originality of patents.

The second proposal targets deriving new attributions(company revenue based on patents score) and show impact of patent on company growth and research.

The first proposal displays the spatial relationship between number of patents, and the citation dependencies (how often does one patent cites the other.)

1.2.1.3 Visual Encoding: How?

The present here the visual encoding for each

1. Proposal 1: Uses Map, Change,navigate,filter and select vis idioms to achieve the task abstraction discussed above for proposal 1.
2. Proposal 2: Navigate, Embed, Change and hue vis idioms for the realization of task abstraction defined for proposal 2.
3. Proposal 3: Shape, hue, Filter and select operation to achieve the task abstraction discussed above for proposal 1.

Hence we will refer to proposal one as *map*, proposal 2 as *patent impact* and proposal 3 as *Originality Ranking*. The reasons for these names will be obvious in later chapters.

1.2.1.4 Algorithm

For this we will use D3JS and JavaScript as the backbone for algorithmic implementation. As each vis presented in this work presents its own challenges, a discussion of relevant algorithmic challenges and the proposed solutions will be left till later chapters.

We will conclude that this is of course an iterative process as illustrated in the diagram above. We do not consider each of the abstractions in isolation.

1.3 Conclusion

In this chapter a general overview of what the motivation, target and framework within which we will work. In the next chapter we discuss the exploratory data analysis.

Chapter 2

Exploratory Data Analysis

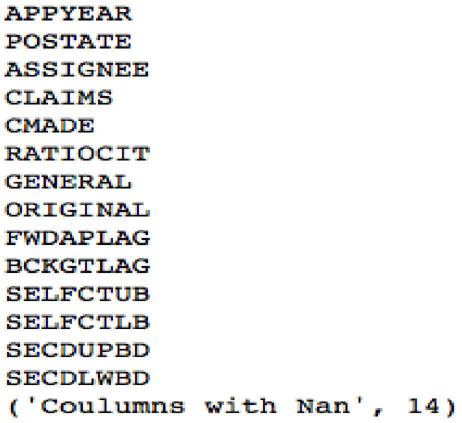
2.1 Introduction

Evidently, there is need to carry out exploratory data analysis on the data set in order to gain any meaningful insight. This is a very crucial step that we decide to discuss it separately in a chapter. It goes without saying that if the analysis is not correct, then the visualization which is the depends on the analysis, will mislead users

2.1.1 First Look into the Data set

The size of the data set is $(2923922 * 23$, which is is fairly big. First, we want to know the columns that have Nans. Next, we analyze the number of columns in the dataset: In particular it is very difficult to get the definition of the column names. We provide this below: The code and definition of the column_codes, types and meaning are provided below: The columns we worked with the most are Patent , Grant year, Measure of originality and Country. With this we want to achieve simplicity and answer the questions that were posed in the first chapter. Other datasets were also added as needed. In addition we also used the citing and cited dataset (also from [1]). This shows a cited patent and the corresponding of citing patent. For completion, the first few columns of the data set is provided below: The citing and cited data set would be used to plot the links between the country of the cited patent and the contry of the citing patents.

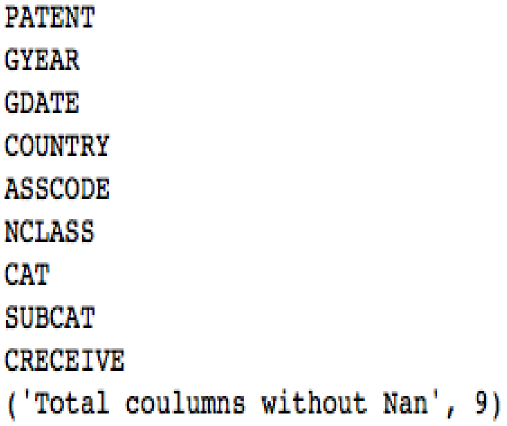
Figure 2.1: Column With Nans



```
APPYEAR
POSTATE
ASSIGNEE
CLAIMS
CMADE
RATIOCIT
GENERAL
ORIGINAL
FWDAPLAG
BCKGTLAG
SELFCTUB
SELFCTLB
SECDUPBD
SECDLWBD
('Coulumns with Nan', 14)
```

[From Our Ipython Notebook on Github]

Figure 2.2: Columns Without NAn



```
PATENT
GYEAR
GDATE
COUNTRY
ASSCODE
NCLASS
CAT
SUBCAT
CRECEIVE
('Total coulumns without Nan', 9)
```

[From Our Ipython Notebook on Github]

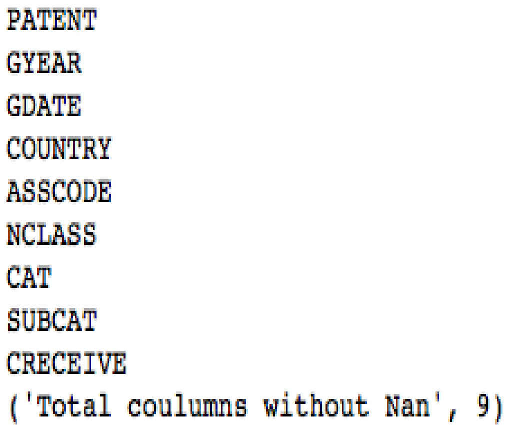
In addition , we also used the fortune data set from `http://archive.fortune.com/magazines/fortune/fortune500_archive/full/`. For completeness we also show the first few columns of the data sets:

2.1.2 Exploratory analysis

We answer the various questions presented in question one. But in particular we will present those questions that influence our visualisation decision:

Since the data is very large we first set out to answer the question on which countries have the most patent. The result is presented in the graph below:

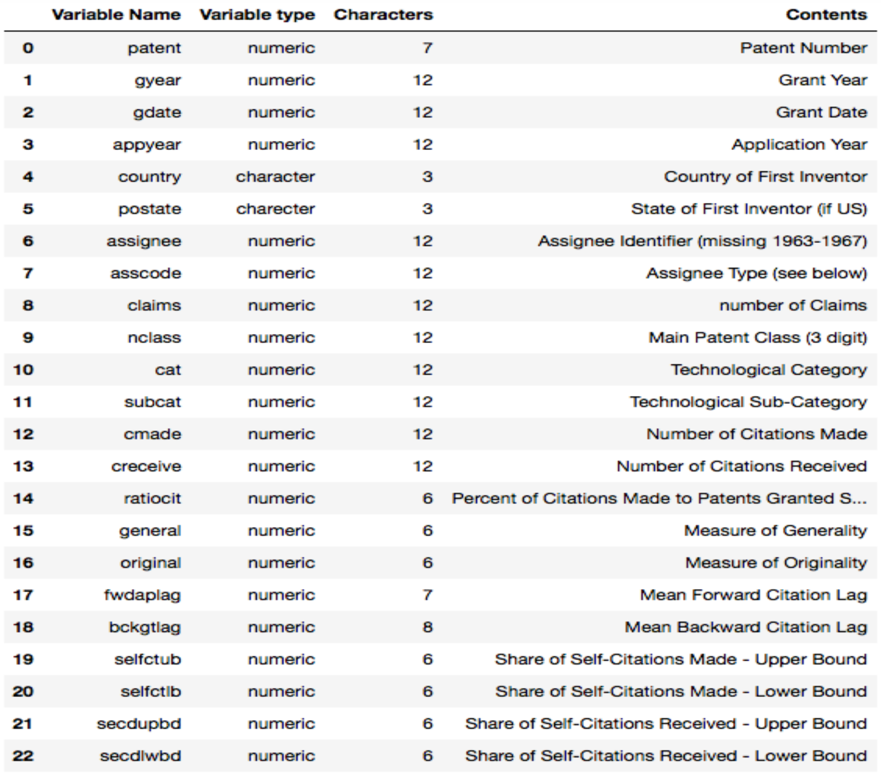
Figure 2.3: Columns With NAn



```
PATENT
GYEAR
GDATE
COUNTRY
ASSCODE
NCLASS
CAT
SUBCAT
CRECEIVE
('Total coulumns without Nan', 9)
```

[From Our Ipython Notebook on Github]

Figure 2.4: Column codes and their definition



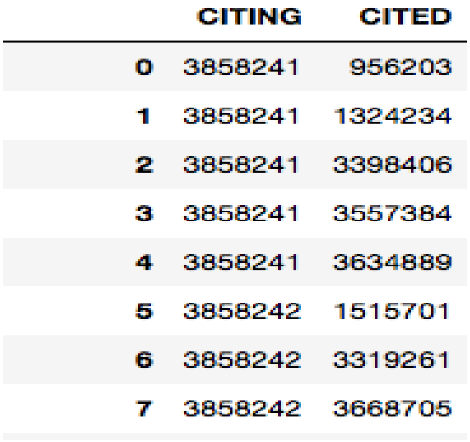
	Variable Name	Variable type	Characters	Contents
0	patent	numeric	7	Patent Number
1	gyear	numeric	12	Grant Year
2	gdate	numeric	12	Grant Date
3	appyear	numeric	12	Application Year
4	country	character	3	Country of First Inventor
5	postate	charecter	3	State of First Inventor (if US)
6	assignee	numeric	12	Assignee Identifier (missing 1963-1967)
7	asscode	numeric	12	Assignee Type (see below)
8	claims	numeric	12	number of Claims
9	nclass	numeric	12	Main Patent Class (3 digit)
10	cat	numeric	12	Technological Category
11	subcat	numeric	12	Technological Sub-Category
12	cmade	numeric	12	Number of Citations Made
13	creceive	numeric	12	Number of Citations Received
14	ratiocit	numeric	6	Percent of Citations Made to Patents Granted S...
15	general	numeric	6	Measure of Generality
16	original	numeric	6	Measure of Originality
17	fwdaplag	numeric	7	Mean Forward Citation Lag
18	bckgtlag	numeric	8	Mean Backward Citation Lag
19	selfctub	numeric	6	Share of Self-Citations Made - Upper Bound
20	selfctlb	numeric	6	Share of Self-Citations Made - Lower Bound
21	secdupbd	numeric	6	Share of Self-Citations Received - Upper Bound
22	secdlwbd	numeric	6	Share of Self-Citations Received - Lower Bound

[From Our Ipython Notebook on Github]

This information will be used in the originality ranking vis idiom of countries that fall into this category. Clearly the US leads the table.

In addition, we also wanted to know which feature of the patent data set correlate the most with the economy. The result of such analysis is provided below:

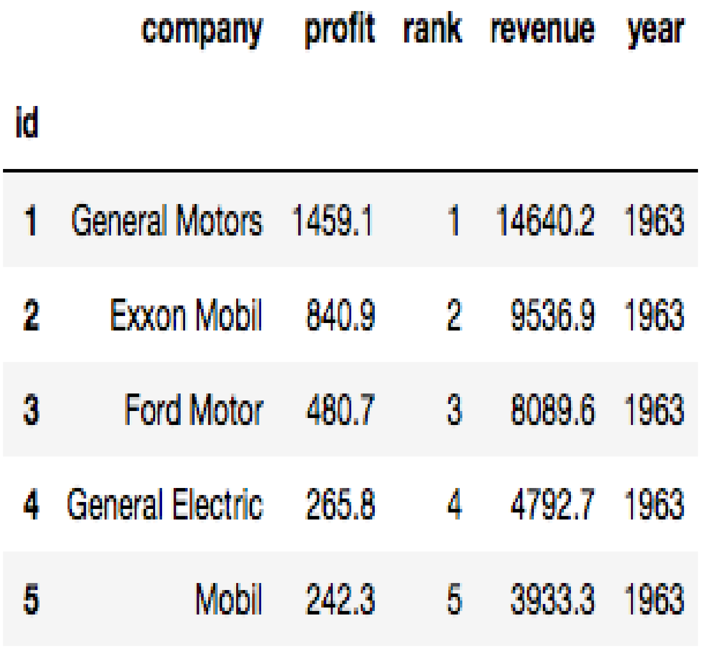
Figure 2.5: citing and cited Data set



	CITING	CITED
0	3858241	956203
1	3858241	1324234
2	3858241	3398406
3	3858241	3557384
4	3858241	3634889
5	3858242	1515701
6	3858242	3319261
7	3858242	3668705

[From Our Ipython Notebook on Github]

Figure 2.6: Fortune 500 data



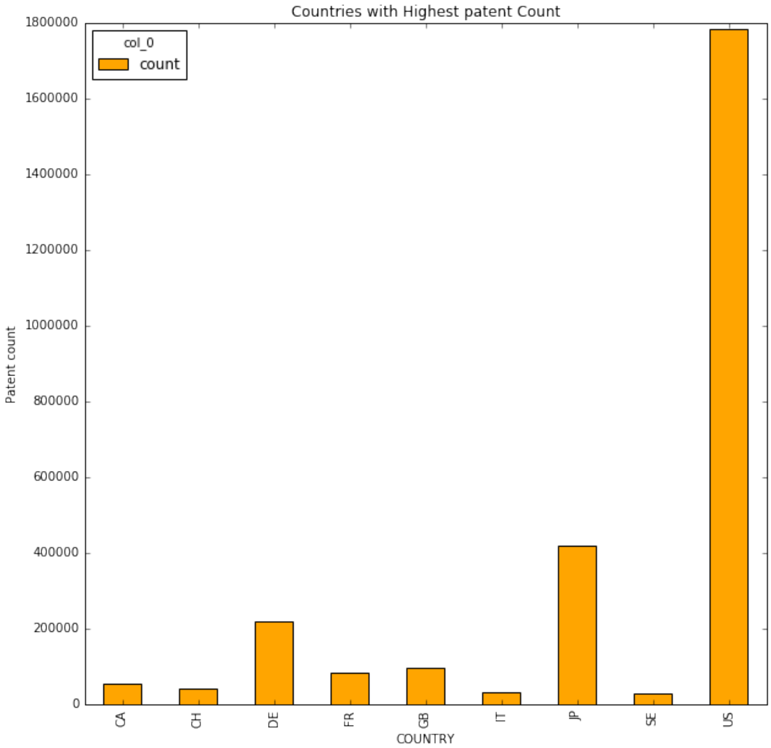
	company	profit	rank	revenue	year
id					
1	General Motors	1459.1	1	14640.2	1963
2	Exxon Mobil	840.9	2	9536.9	1963
3	Ford Motor	480.7	3	8089.6	1963
4	General Electric	265.8	4	4792.7	1963
5	Mobil	242.3	5	3933.3	1963

[From Our Ipython Notebook on Github]

2.1.3 Map Data Preparation

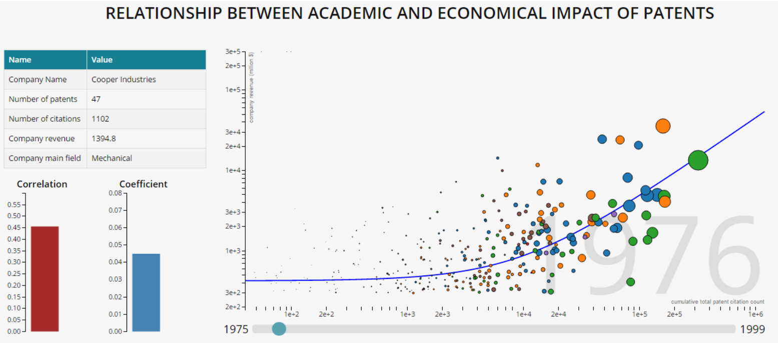
The Map vis idiom was a bit chanllenging. We gad to experiment with various available data sets onine data sets for the map visualisation. When we had the map set up. We had to join the data set to the to map data set. Finally, we genrated the links on the map that connects cited country to the citing country. This was donr using an inner merge on the patent data set and citing_cited data set discussed above. From this merge

Figure 2.7: Countres with Highest Patent count



[Image from Git Repo,]

Figure 2.8: Feature set and Economy Impact


RELATIONSHIP BETWEEN ACADEMIC AND ECONOMICAL IMPACT OF PATENTS

Name	Value
Company Name	Cooper Industries
Number of patents	47
Number of citations	1102
Company revenue	1394.8
Company main field	Mechanical

[Image from online source]

we get the country code and use this as basis for the links on the map.

2.2 Conclusion

In this section we provided description of the data set cleaning pipeline that is relevant to the visualization pipeline. Further relevant details can be found on our git hub page.

Chapter 3

Patent Impact

3.1 Introduction

The number of patents and their quality is important index of country's innovation and economic growth. However, according to our data set, the total number of patents granted by only US Patent and Trademark Office (USPTO) from 1963 to 1999 is almost 3 million. In other words, over 200 patents were granted per day, and it's increasing over time. Such exponential growth raises important questions like "How can one measure patent's quality", "How to distinguish high quality patents" and "Is it possible to predict, if whether patent is good or bad". By answering to those questions, both researchers and business companies can benefit and save huge amount of assets and time.

In this section, we will try to measure both economic and academic impacts of each patents and try to find if there is any correlation or interesting relationship between them. In order to do that, we have used NBER patents data set and fortune rankings list from 1963 to 1999

3.2 Defining Academic Impact

According to the `https://www.nap.edu/read/5976/chapter/7`, the number of patents issued and the technical and scientific literature citations on the patents can be used to develop quantitative measures of innovative output and science-technology linkages. Thus, to measure academic impact of each patent, first, we have used number of citations it has received using citations data set. Even though our citation data set contains only information related from 1975 to 1999, there is over 16 million citation relationships between patents.

Secondly, since only the number of citations each patent received isn't good method to evaluate academic impact. Because, the patent which influenced another patent which has received huge amount of citations, might have bigger impact, but number of citations alone can't represent such cases. So, we have measured second layer of citations, which is the number of citations of each citing patents received for given patent. Thus we also considered them.

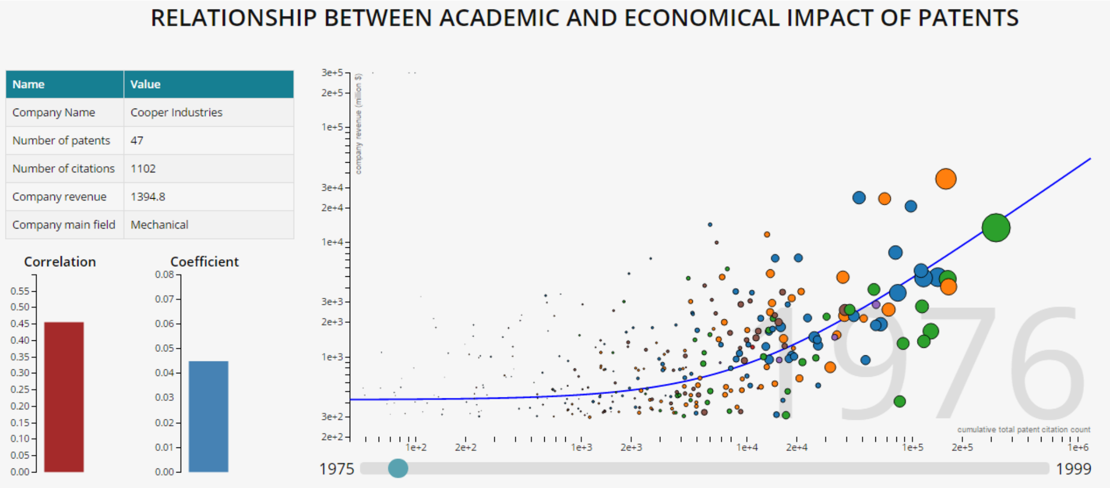
3.3 Defining Economic Impact

The data set provided by NBER doesn't have any information related to economy except the company names of each patents are assigned to. Thus, in this section, we have scraped Fortune magazine's rankings of top 500 companies in US over 1963 to 1999, with their revenues and profits in million `http://archive.fortune.com/magazines/fortune/fortune500_archive/full/` By combining these 2 data sets, we can get patents information of each company owns and their economic measurements (profit and revenue) for given year.

3.4 Relationship between economic and academic impact

Once we have successfully extracted each patent's impact, we tried to visualize and analyze their relationship. Our visualization consists of main graph plot and supplementary table and bar charts.

In the main graph plot, each bubble represents one company that has mentioned in given year's fortune list and their size is for number of patents they own and color is for dominant category of patents. In X axis, we have number of citations to every patents owned by each company, while in Y axis, we have selected company's revenue. The blue line indicates the best fit of linear regression model. You can see parameters value at the table



RELATIONSHIP BETWEEN ACADEMIC AND ECONOMICAL IMPACT OF PATENTS

Name	Value
Company Name	Cooper Industries
Number of patents	47
Number of citations	1102
Company revenue	1394.8
Company main field	Mechanical

[The relationship between academic and economic impact of patents]

In order to analyze relationship, we have trained linear regression model for our data set and extracted coefficient of the model and correlation between revenue and number of citations. These two variables will tell us, how much do academic impacts affects economical ones and how accurate our model is.

In general the correlation is pretty high and showing clear evidence that ACADEMIC impact indeed influence on ECONOMY. But interestingly, this graph might be inferring some historical events. For example around early 1980, there was global economical recession happened and it was ended around 1984. And also, through 1989 to 1991, due to inconvenient government policy economical recession happened. However the 1990 to 2000 is considered the golden years of US economy and in 1994 the number of jobs created were peaked. `https://en.wikipedia.org/wiki/1990s_United_States_boom`,`https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States`. Even though it is tempting to say that economic and academic relationship is somehow related to nations economical state and policy, further research and expertise are needed.

Chapter 4

Geospatial Map

4.1 Introduction

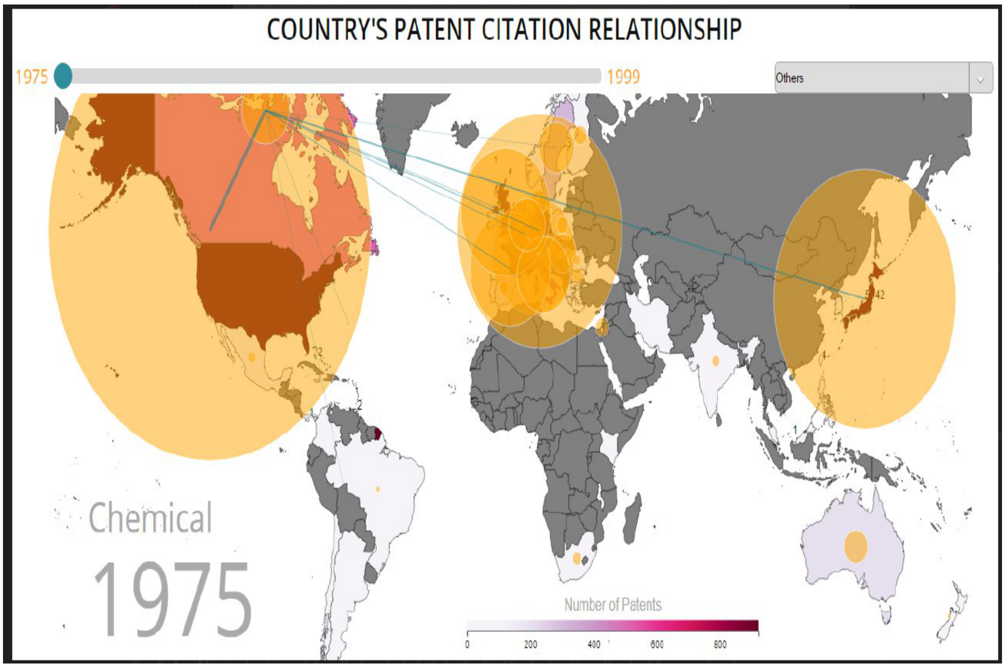
The aim of the map is two fold; reflecting the number of patents generated for each country around the world during a period from 1975 to 1999 for each category of patent. For a given category of patent selection, the opacity of the colours and and the size of the bubbles corresponds to the patent count for the country: the denser the color and the bigger the bubble size, higher the number of patent for the country. On the other hand, the map also enables us to show the patent citation dependencies between countries, this is reflected by the outgoing edges form the patent issuer country to the the patent citing countries, the magnitude for the edge tells us how much the cited patent countries is cited from the other citing patent countries.

4.1.1 Data set Preparation

We used the patent data set. We filter to extract only the patent ID, Country, Grant year and the category of patent. This information is merged with the spatial coordinate data for the base map.

This geometry information together with the right combination of hue and saturation and area channel in the form of bubble adds the relevant information about the patent count per country.

Figure 4.1: map

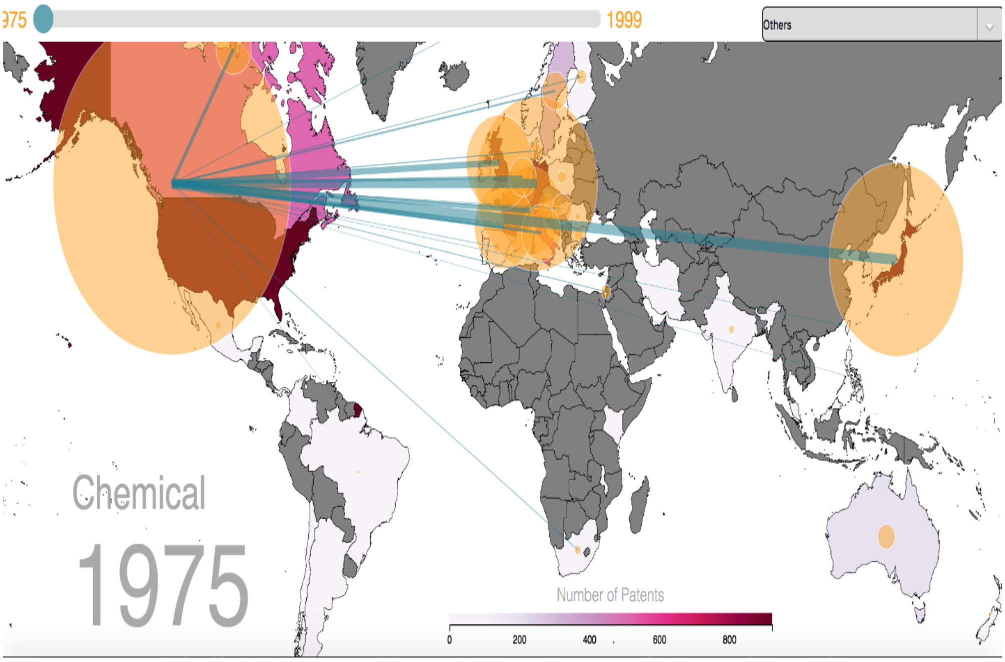


[Image from Git Repo,]

From the visualization we can select the category and click on the bubbles to generate the link between the cited country and citing country

Moving the slider enables one to see the changes in bubble size or patent count over the years for the category selected. In addtion from the image below we notice that both USA and Japan has a huge number of patent , the former patents are increasingly cited from europe countries and Japan

Figure 4.2: map



[Image5 from Git Repo,]

4.2 Conclusion

The map visualisation is the most challenging visualisation. But it was worth the challenge.

Chapter 5

Orogonality Ranking

5.1 Overview And Target

In this last chapter, we document the last visualization idiom. First, we discuss the data cleaning process pertinent to the visualization. Then a little implementation details will be analysed and lastly, a screen short of how the visualization works will be provided.

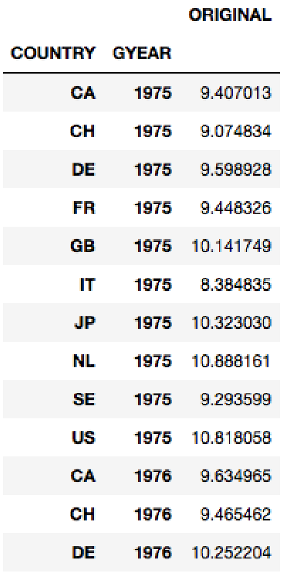
5.1.1 Cleaning Pipeline

For this visualisation we only need the top 10 countries with hoghest visualisation. Next, we compute the average originality score for each country for each year. Lastly for each year considered we place the country with the highest score at the top of the ranking. This is Domain Situation level for this visualization.

For simplicity , a Network data set type will be used with a line mark or channel. This each line connect the ranking of the countries considered across a timeline. Approproate hue, color and saturation were added so that countries can be distinguishe form each other. Overall the target is for the user to enjoy the visualisation while having a fun interaction with the interface in other to gain useful insight. In addition this was meant to be a summary of facts that showcase originality among countries with most patent.

Below we show the output of top columns of the data set used in this visualization:

Figure 5.1: Originality ranking data set



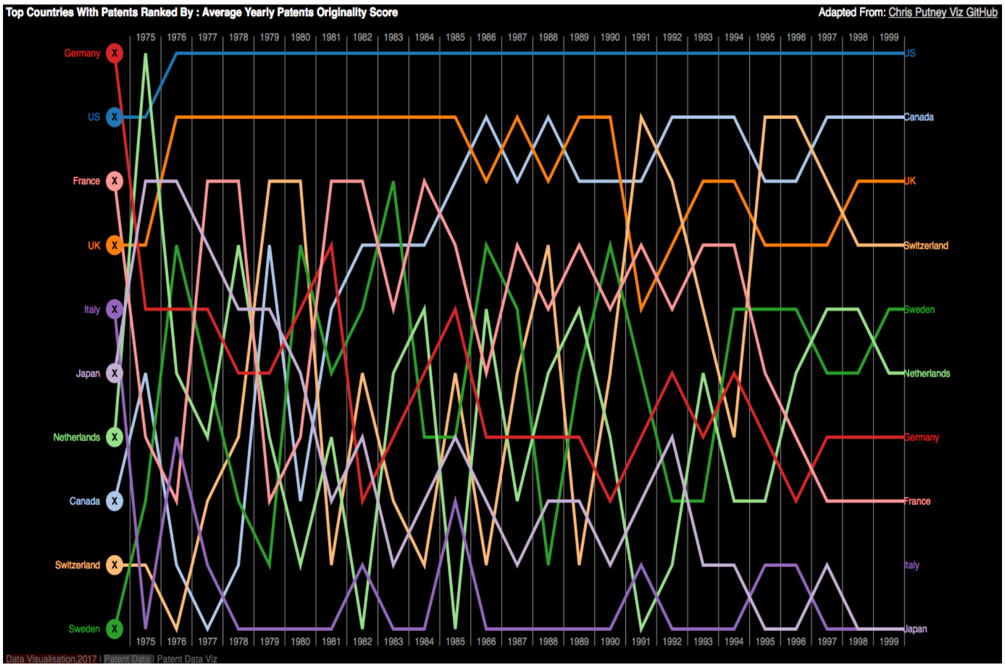
		ORIGINAL
COUNTRY	GYEAR	
CA	1975	9.407013
CH	1975	9.074834
DE	1975	9.598928
FR	1975	9.448326
GB	1975	10.141749
IT	1975	8.384835
JP	1975	10.323030
NL	1975	10.888161
SE	1975	9.293599
US	1975	10.818058
CA	1976	9.634965
CH	1976	9.465462
DE	1976	10.252204

[Image from Git Repo, This values were converted into integer that corresponds to the country originality ranking for each year - the is a value in the range 1 to 10]

5.1.2 Look and Feel of Visualisation Idiom

On launching the originality ranking vis idiom we get the following:

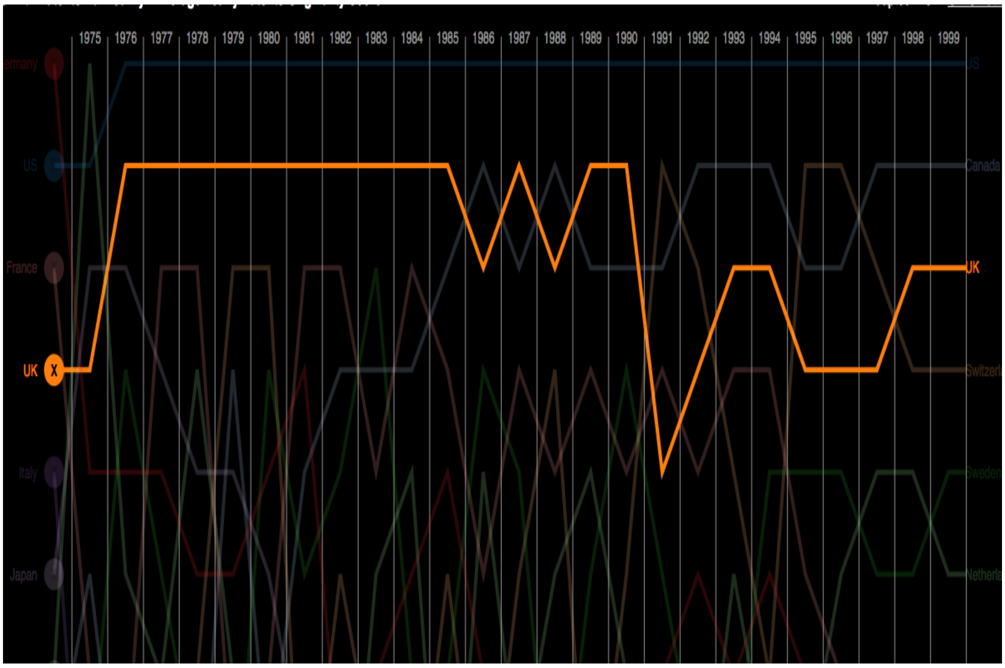
Figure 5.2: Countres with Highest Patent count



[Image from Git Repo,]

The visualisation is optimised for optimal interactive display. When one selects any country of one's choice one gets the following:

Figure 5.3: Originality rankinf Illustration

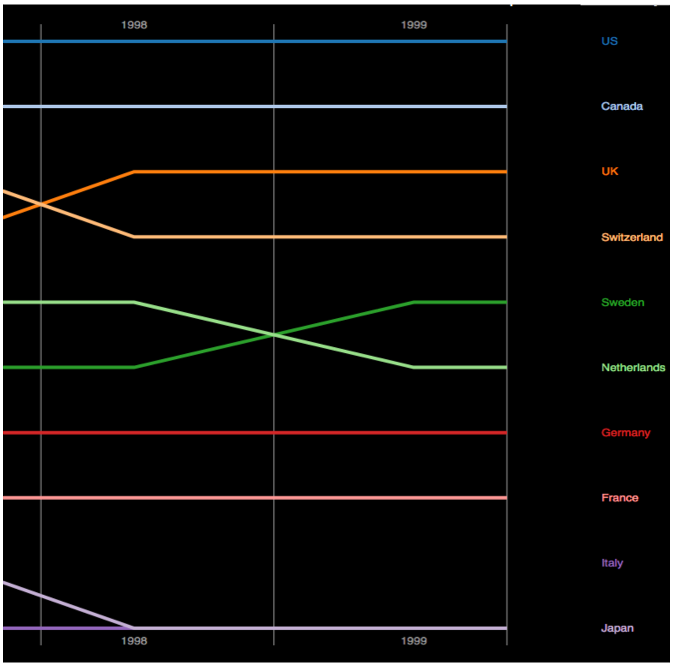


[Image from Git Repo,]

The figure above shows how the country yearly orinality ranking varies from 1975 to 1999. Note that the top axis bears the year. Also note the the the corresponding state is highlighted.

Clicking on the interface expands the expand the horizontal lines revealing more details:

Figure 5.4: Originality ranking Illustration



[Image from Git Repo,]

Lastly, the user can have fun by interactively clicking and unclicking the interface to reveal the ranking of countries and also feel their artistic self!

5.2 Peer Assessment

1. Preparation were they prepared during team meetings? Yes. Each member of our group answered yes to this question

2. Contribution did they contribute productively to the team discussion and work? Yes. Each member of our group answered yes

3. Respect for others ideas did they encourage others to contribute their ideas? Yes, each member of the group answered yes to this question

4. Flexibility were they flexible when disagreements occurred? Yes . Each member of our group answered yes to this question

5.3 Conclusion

In this chapter the Originality ranking of the data set was discussed. Further details can be found on the git hub.

Bibliography

[1] B.H. Hall, A.B JAffe, M Trajtenberg. The nber patent citations data file:lessons, insights and methodological tools. *NBER Working Paper*, 8498, 2001.

[2] Web page. Wealth of nations. `http://romsson.github.io/dragit/example/nations.html`, 2015. Accessed Dec 24, 2017.

[3] Web page. Chris pudney. `http://www.vislives.com/`, 2012. Accessed Dec 24, 2017.

[4] Tamara Munzner. Visualization analysis and design, 2014.