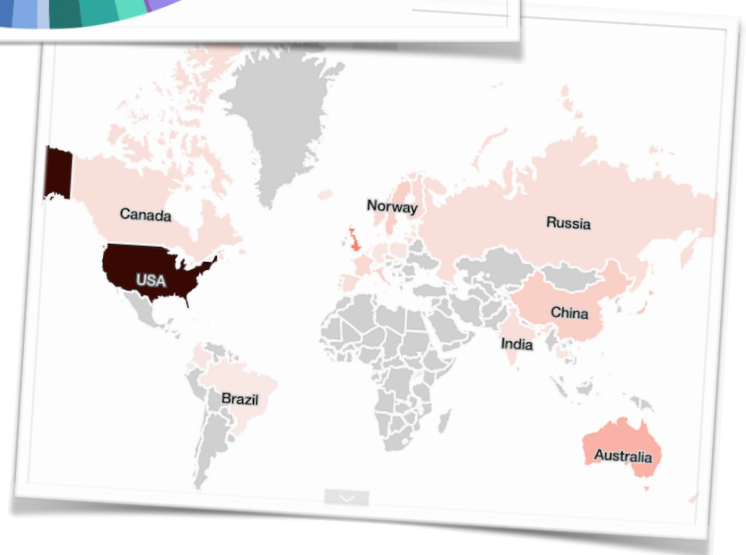
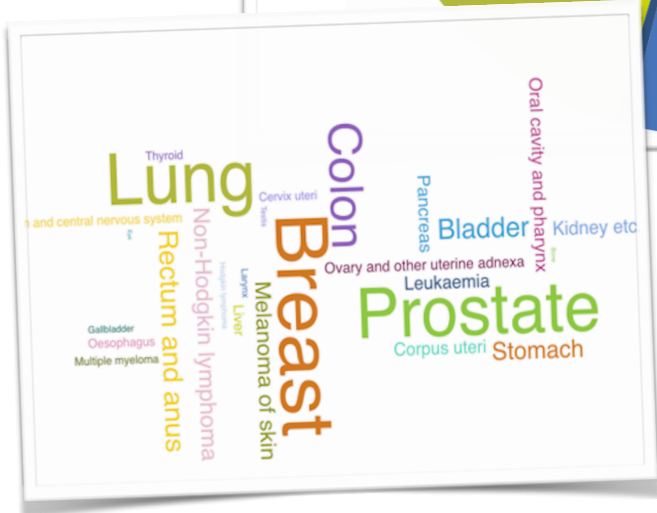
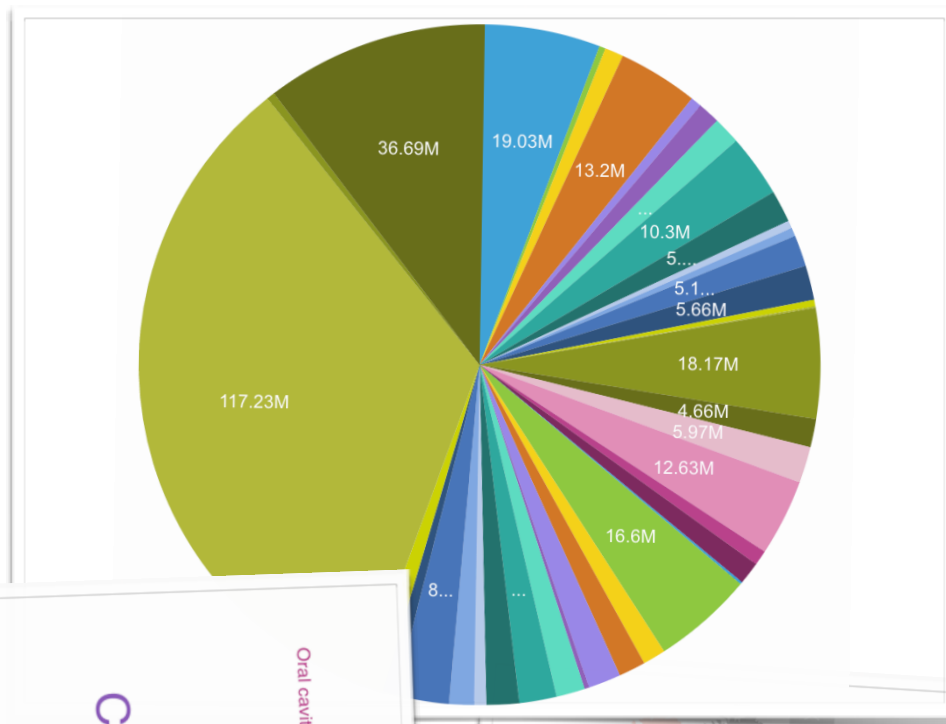


# Watson Analytic

Data Visualization on Global Trends on Cancer Incidence An Application of IBM Watson Analytics



# Watson Analytic

## Data Visualization on Global Trends on Cancer Incidence An Application of IBM Watson Analytic[1]

### **Purpose:**

Using the IBM Watson Analytics to implement the CI5 Cancer Database from WHO cancer registry. Try to build the visualization of data and explore the data distribution and trends.

### **What is Watson?**

Watson is an IBM supercomputer that combines artificial intelligence (AI) and sophisticated analytical software for optimal performance as a “question answering” machine.[2]

### **What is Watson Analytics?**

A smart data discovery service available on the cloud, it guides data exploration, automates predictive analytics and enables effortless dashboard and info-graphic creation.

## Setup your first account:

Free	Plus	Professional
Upload spreadsheets, get visualizations, discover insights and build dashboards-all on your own.	Get all the features of Free plus more storage and data sources, including databases and Twitter.	Get all the features of Plus plus a multi-user tenant to collaborate, more storage and more data.
\$ 0 <sup>00</sup> USD	Starting at \$ 30 <sup>00</sup> USD* per month per user	Starting at \$ 80 <sup>00</sup> USD* per month per user
<a href="#">Try free edition</a>	<a href="#">Purchase now</a>	<a href="#">Purchase now</a>
1 user	1 user	1 or more users
1 MB of storage included	2 GB of storage included	100 GB of storage included

It's quite easy for you to register an account for IBM cloud server. It will only take about 6 minutes to fill all the information it needs. There are three types of the account for Watson Analytics. The types and price are above. What's more, it provides a new user a 30 days free trial of professional version. In my opinion, a normal free version has so limited storage that you can do anything.[3]

## Where do we start from?

Here is a link to the tutorial to the book. It teaches us how to get ready to run our database on the IBM cloud.

<https://community.watsonanalytics.com/wp-content/uploads/2017/03/Tutorial-about-Watson-Analytics-2017-04-10.pdf>[4]

### Cancer Data:

Our cancer data comes from WHO official website[5]. The cancer data has 181 different types which can be grouped by 28 different groups according to human physiological structure. The data comes from 191 different cities of different countries in the world.

CANCERID	name
1	All sites but non-melanoma skin
2	Lip
3	Tongue
4	Mouth
5	Salivary glands
6	Tonsil
7	Other oropharynx
8	Nasopharynx
9	Hypopharynx
10	Pharynx unspecified
11	Oesophagus
12	Squamous cell carcinoma
13	Adenocarcinoma
14	Other specified carcinomas
15	Unspecified carcinoma
16	Sarcoma
17	Other specified morphology
18	Unspecified morphology
19	Stomach
20	Small intestine
21	Colon
22	Rectum
23	Anus
24	Anus: Squamous cell carcinoma
25	Anus: Basaloid and cloacogenic carcinom
26	Anus: Adenocarcinoma
27	Anus: Other carcinoma
28	Anus: Unspecified carcinoma
29	Anus: Melanoma
30	Anus: Other morphology

A	B	C	D	E
CANCERID	LABEL			
1	All sites but non-melanoma skin			
2	Oral cavity and pharynx			
3	Oesophagus			
4	Stomach			
5	Colon			
6	Rectum and anus			
7	Liver			
8	Gallbladder			
9	Pancreas			
10	Larynx			
11	Lung			
12	Bone			
13	Melanoma of skin			
14	Breast			
15	Cervix uteri			
16	Corpus uteri			
17	Ovary and other uterine adnexa			
18	Prostate			
19	Testis			
20	Kidney etc.			
21	Bladder			
22	Eye			
23	Brain and central nervous system			
24	Thyroid			
25	Non-Hodgkin lymphoma			
26	Hodgkin lymphoma			
27	Multiple myeloma			
28	Leukaemia			

Registry	number	Location
3602	99	Australia
3603	99	Australia
3604	99	Australia
3605	99	Australia
3606	99	Australia
3607	99	Australia
4007	99	Australia
4008	99	Australia
7602	99	Brazil
12403	99	Canada
12406	99	Canada
12413	99	Canada
15602	99	China
15607	99	China
15630	99	China
17001	99	Colombia
18800	99	Costa Rica
19100	99	Croatia
20300	99	Czech Republic
20800	99	Denmark
21801	99	Ecuador
23300	99	Estonia
24600	99	Finland
25001	99	France
25002	99	France
25003	99	France
25004	99	France
25005	99	France

REGISTRY	ETHNIC	GFYEAR	SEX	CANCERID	NO_4	N5_9	N10_14	N15_19	N20_24	N25_29	N30_34	M
3602	99	1983	1	1	47	19	29	47	97	82	130	
3602	99	1983	1	2	0	0	0	0	3	3	5	
3602	99	1983	1	3	0	0	0	1	0	0	0	
3602	99	1983	1	4	0	1	0	0	1	0	0	
3602	99	1983	1	5	0	0	0	0	0	0	1	
3602	99	1983	1	6	0	0	0	0	0	0	0	
3602	99	1983	1	7	0	0	0	0	0	0	0	
3602	99	1983	1	8	0	0	0	0	0	1	1	
3602	99	1983	1	9	0	0	0	0	0	0	0	
3602	99	1983	1	10	0	0	0	0	1	0	0	
3602	99	1983	1	11	0	0	0	0	0	0	1	
3602	99	1983	1	12	0	0	0	0	0	0	0	
3602	99	1983	1	13	0	0	0	0	0	0	1	
3602	99	1983	1	14	0	0	0	0	0	0	0	
3602	99	1983	1	15	0	0	0	0	0	0	0	
3602	99	1983	1	16	0	0	0	0	0	0	0	
3602	99	1983	1	17	0	0	0	0	0	0	0	
3602	99	1983	1	18	0	0	0	0	0	0	0	
3602	99	1983	1	19	0	0	0	0	1	0	6	
3602	99	1983	1	20	0	0	0	0	0	1	0	
3602	99	1983	1	21	0	0	0	0	2	2	4	
3602	99	1983	1	22	0	0	0	0	3	1	5	
3602	99	1983	1	23	0	0	0	1	0	0	0	
3602	99	1983	1	24	0	0	0	0	0	0	0	
3602	99	1983	1	25	0	0	0	0	0	0	0	

The first graph shows the CancerID and cancer name. Second one shows CancerGroupID and cancer category. The third one shows that we use registryID and ethnicID to indicate the specific location. The fourth graph is the detail information of the total amount of cancer in age N?? based on registryID, ethnic-ID, year, sex, cancerID. Our continued charts are based on the second, third and fourth graph.



### How to setup our database?

We have quite a lot of different excel and txt file to create our database together. However, Watson Analytic can only deal with one sum-up table once. So, we tried to build up the our own database by mysql and then reshape our data and upload to our Watson cloud.

```
1 drop database aiDBSum;
2 create database aiDBSum;
3 use aiDBSum;
4
5 create table cancer_summary
6 (cancer_groupId int,
7 cancer_label varchar(64),
8 primary key(cancer_groupId));
9
10 create table registry
11 (registry_id int,
12 ethnic_group int,
13 location varchar(64),
14 primary key(registry_id,ethnic_group));
15
16
17 create table summary
18 (registry_id int,
19 ethnic_group int,
20 year int,
21 sex int,
22 cancer_groupId int,
23 total int,
24 n0_4 int,
25 n5_9 int,
26 n10_14 int,
27 n15_19 int
```

Here is the DDL file to build up mysql.

To connect our database to the Watson Analytic cloud server and upload our data on it. We need to establish a security gateway. First, we add gateway which will create an ID and security token for connection.

Gateway ID	Security Token
JSsEcdHKFpV_prod_ng 	eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJjb25rr 

### Add Gateway ✕

  
 Require security token to connect clients ⓘ  Token Expiration:  days ⓘ

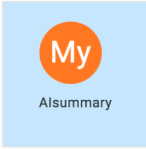
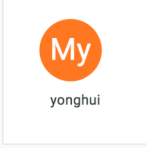
Then we need to set the ACL(access control allow) accessible to Watson Cloud server.

After connecting to the IBM Watson server, we need to pick our table which we want to upload and reshape it.

Add data ✕

Import Connection Local file

Shape before Upload now

Connections	Schemas	Tables and Views	Selected (3)
 Alsummary	aiDBSum <span>3 of 3</span>	<input checked="" type="checkbox"/> cancer_summary <input checked="" type="checkbox"/> registry <input checked="" type="checkbox"/> summary	registry summary cancer_summary
 yonghui			

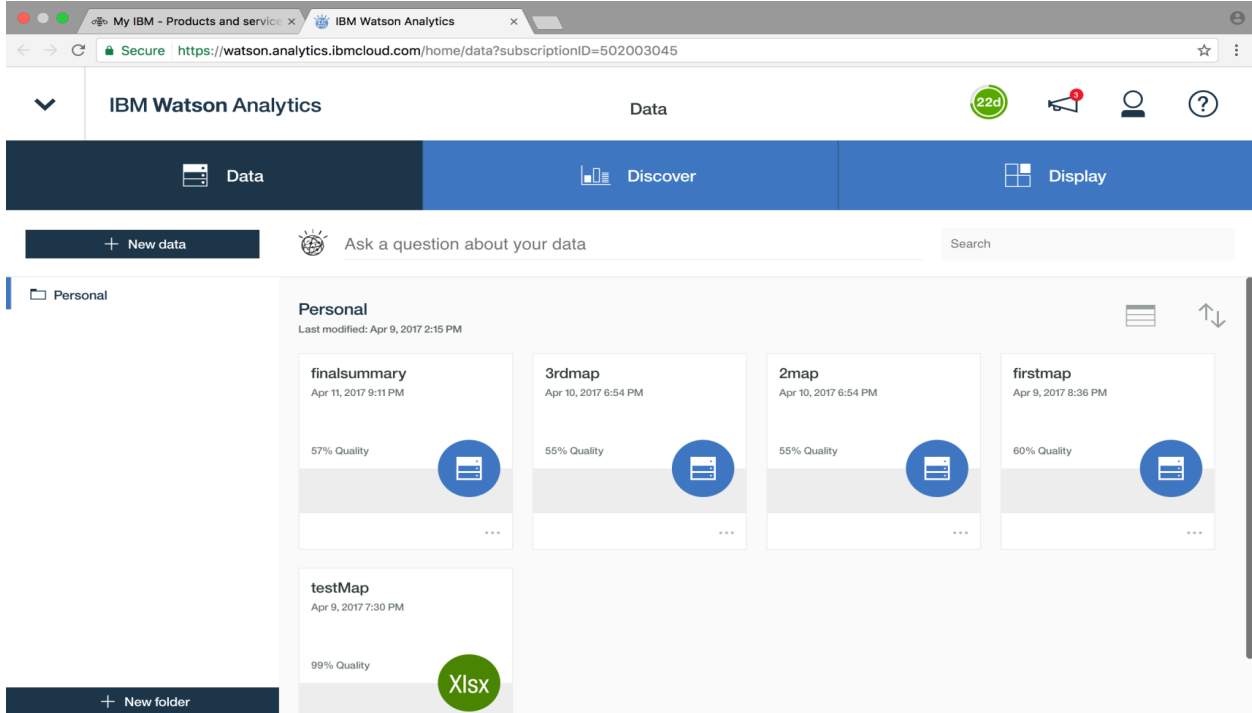
## Overview of our data on the cloud server:

After finishing reshaping and uploading, we now get our data on cloud. If your database is quite big, it will take a while for uploading and analyzing. Please be patient.

The screenshot shows a dialog box titled "Add data" with a close button (X) in the top right corner. Below the title bar is a dark header with the text "Join Data" and two buttons: "CANCEL" and "JOIN". The main content area has a light gray background and contains the following text: "Choose how to join the data by selecting join keys. You can also select which columns to include from table registry." Below this, there are two data sources: "A. summary" (26 of 26) and "B. registry" (3 of 3). Two join key pairs are shown: "registry\_id" from table A joined to "registry\_id" from table B, and "ethnic\_group" from table A joined to "ethnic\_group" from table B. A plus sign icon and the text "Add another join key pair" are positioned between the two pairs. At the bottom, there is a checked checkbox labeled "Matches are case sensitive" and a dropdown menu showing "A + matching rows".

The screenshot shows a dialog box titled "Add data" with a close button (X) in the top right corner. Below the title bar is a dark header with the text "Join Data" and two buttons: "CANCEL" and "JOIN". The main content area has a light gray background and contains the following text: "Choose how to join the data by selecting join keys. You can also select which columns to include from table cancer\_summary." Below this, there are two data sources: "A. summary" (25 of 25) and "B. cancer\_summary" (2 of 2). One join key pair is shown: "cancer\_groupId" from table A joined to "cancer\_groupId" from table B. A plus sign icon and the text "Add another join key pair" are positioned to the right of the pair. At the bottom, there is a checked checkbox labeled "Matches are case sensitive" and a dropdown menu showing "A + matching rows".

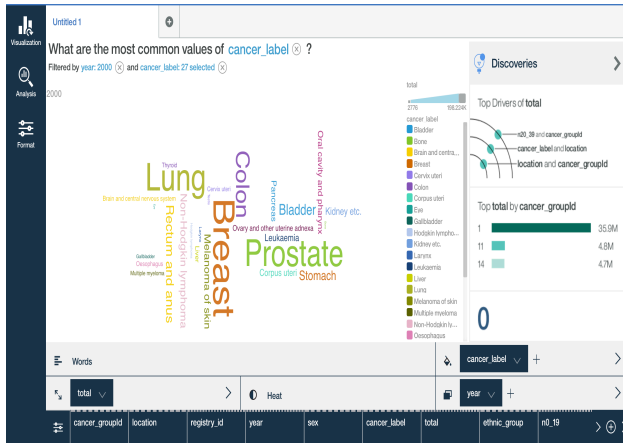
Here is the final cloud server of our account.



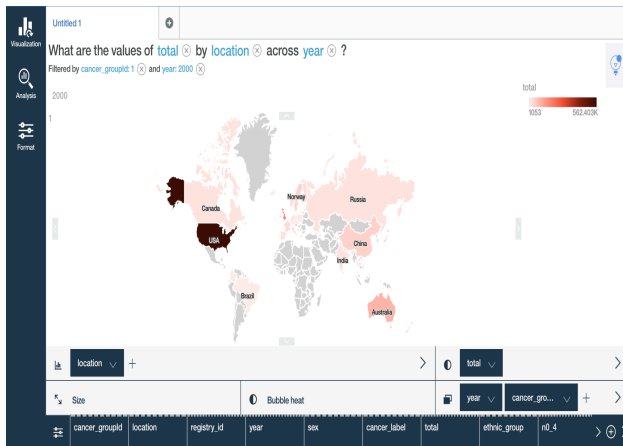


### Get our visualized data:

The usage of Watson analytic is quite similar to the excels. Most operation can be done by just clicking mouse.



Here is a graph shows the total amount of the cancer case of 27 groups in the year of 2000. We use different color to show different cancer category. The size of the letter shows the total amount of the cancer case.

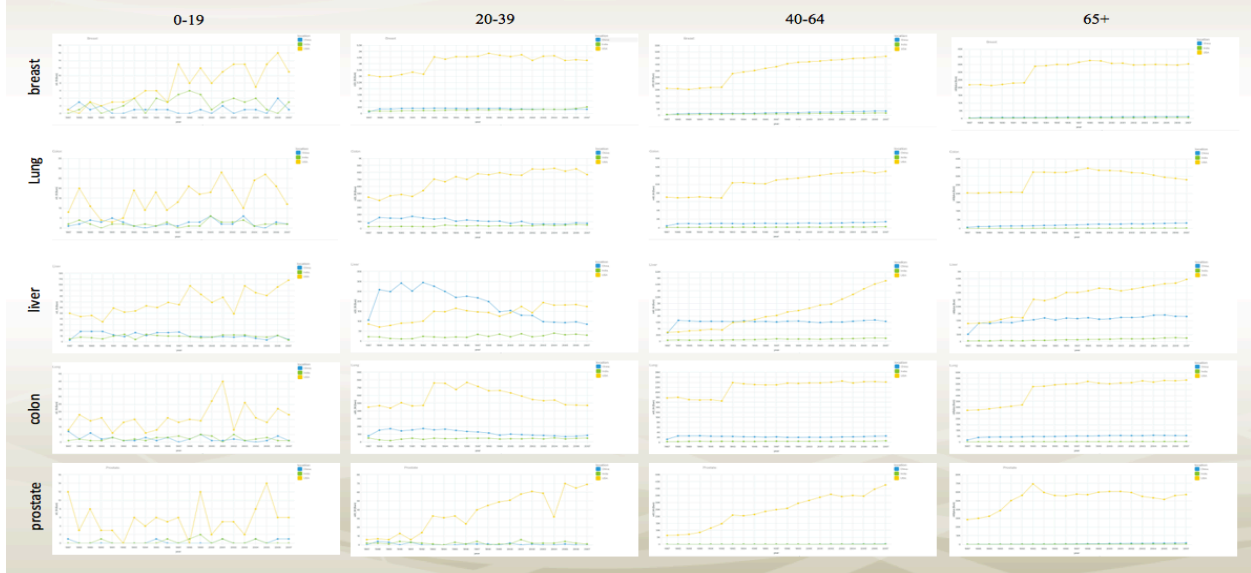
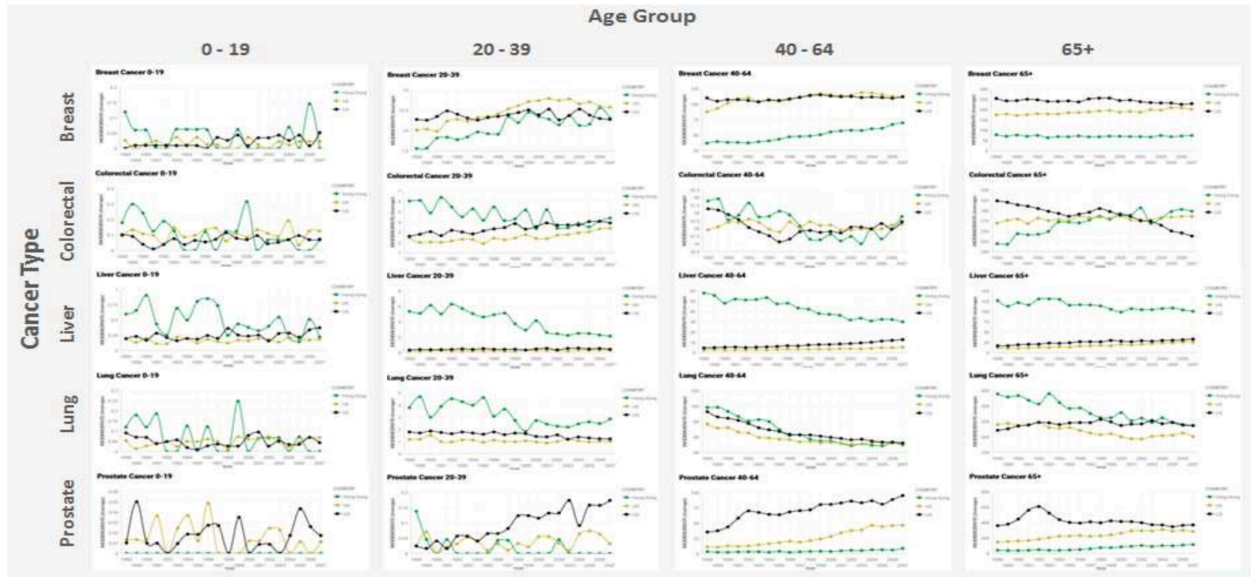


Here is a world map showing the total amount of the cancer case. It's a special functionality of Watson. To do this, we need to check whether that location is in the Watson's map library and change our location to their format, e.g FL.USA.



Here is a graph shows the total amount of the cancer case by year and location. We use different location to show different countries's value.

**Comparison to research paper:**



We tried to build a similar output as the research paper. Here is five cancer categories' line charts between children, young people, middle aged people and elder people. Since the original graph is quite hard to figure out the accurate location they use. We change our data to China, USA and India.

### **Advantage of Watson analytics:**

- ▶ Watson has a nice User Interface
- ▶ easy to use
- ▶ support multiple languages
- ▶ cover most countries in the world while doing mapping
- ▶ query system allow to draw graph by natural language

### **Some deficiencies of Watson:**

- ▶ Watson Analytics can not combine multiple format of data together.
- ▶ Two excel files can not be merged even though they both have a column with the same name.
- ▶ limited mathematic operations.
- ▶ Analysis system is not quite accurate.
- ▶ Comparing to mysql, you need to store quite a lot of redundant data on cloud.

### **Summary:**

In our study, we described data visualization with the IBM Watson Analytics platform to explore the open-sourced data on global cancer trends. We included 28 cancers from different geographic regions. An interactive interface was applied to plot a choropleth map to show global cancer distribution, and line charts to demonstrate historical cancer trends over 20 years. And we also found some advantages and disadvantages of the Watson analytics.

## Reference:

[1]. Tsoi, Kelvin Kf, et al. "Data Visualization on Global Trends on Cancer Incidence An Application of IBM Watson Analytics." Proceedings of the 50th Hawaii International Conference on System Sciences. 2017.

[2]. Watson (computer) - Wikipedia. (n.d.). Retrieved April 28, 2017, from [https://en.wikipedia.org/wiki/Watson\\_\(computer\)](https://en.wikipedia.org/wiki/Watson_(computer))

[3]. IBM. (n.d.). IBM Watson Analytics. Retrieved from <https://www.ibm.com/us-en/marketplace/watson-analytics/purchase#product-header-top>

[4]. IBM Corporation 2015, 2017, Getting started with Watson Analytics

[5].World Health Organization. (n.d.). CI5: CANCER INCIDENCE IN FIVE CONTINENTS. Retrieved from <http://ci5.iarc.fr/Default.aspx>