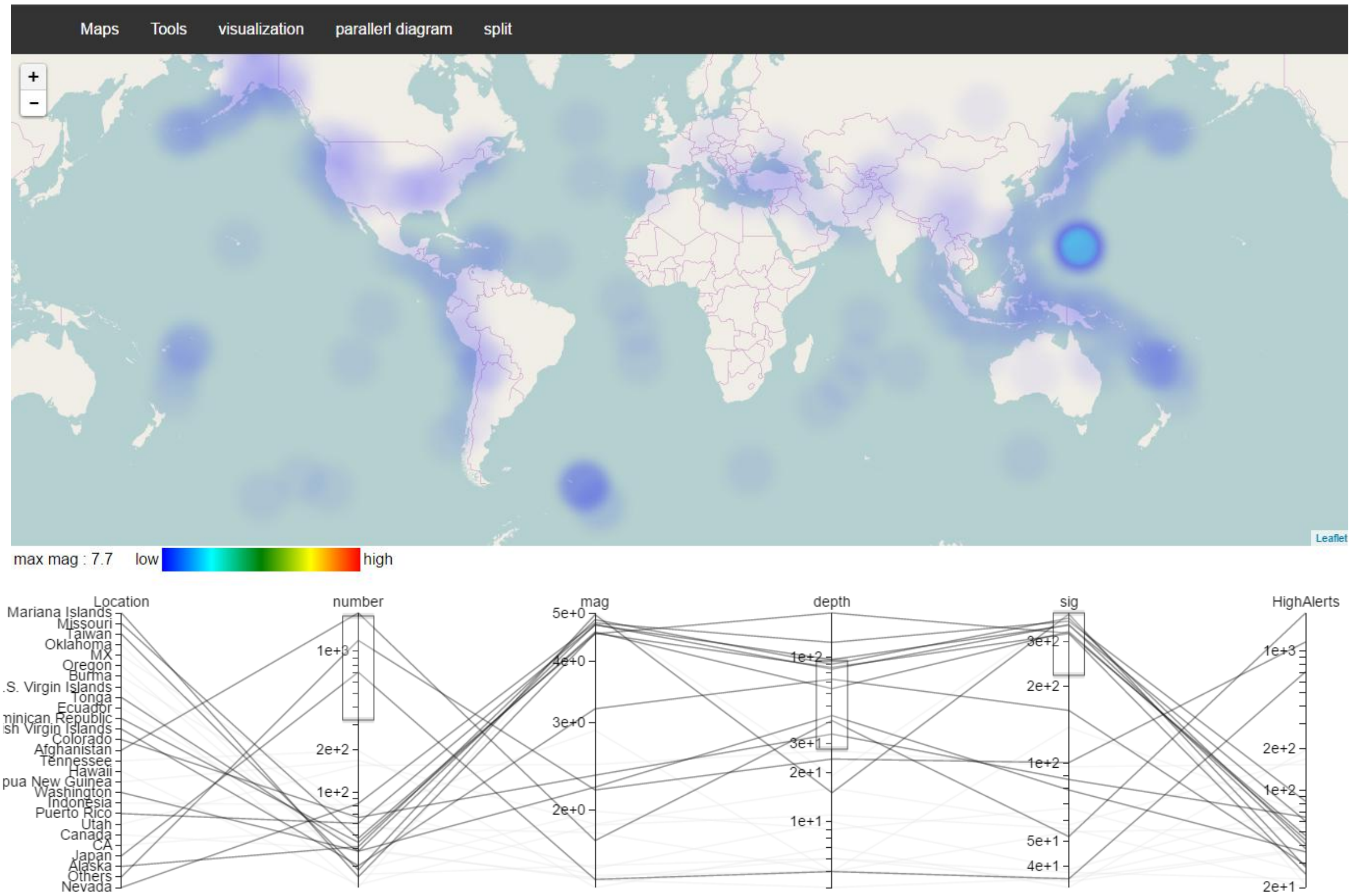
Data Visualization course

Submitters :

samer shahin
ebraheem kshkoosh



The database :

source :
http://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php

Description :

the database contains all earthquakes that happened around the world in the past month .
the database may be acquired in a GeoJson format , spreadsheet , KML ,ATOM and QuakeML . .

GeoJSON is a format for encoding a variety of geographic data structures. A GeoJSON object may represent a geometry, a feature, or a collection of features. GeoJSON uses the JSON standards. The GeoJSONP feed uses the same JSON response, but the GeoJSONP response is wrapped inside the function call, eqfeed_callback .

Different data formats have different fields , what can be found in GeoJson format and in spreadsheet format are not the same . for example the Sig field which is unique to the GeoJson format .

we Have chosen to work with GeOJson format , because its suites our application (JavaScript / d3 Web application) and it has the most valuable fields .

What can be found in the data :

```
{
  type: "FeatureCollection",
  metadata: {
    generated: Long Integer,
    url: String,
    title: String,
    api: String,
    count: Integer,
    status: Integer
  },
  bbox: [
    minimum longitude,
    minimum latitude,
    minimum depth,
    maximum longitude,
    maximum latitude,
    maximum depth
  ],
  features: [
    {
      type: "Feature",
      properties: {
        mag: Decimal,
        place: String,
        time: Long Integer,
        updated: Long Integer,
        tz: Integer,
        url: String,
        detail: String,
        felt:Integer,
        cdi: Decimal,
        mmi: Decimal,
        alert: String,
        status: String,
        tsunami: Integer,
        sig:Integer,
        net: String,
        code: String,
        ids: String,
        ids: String,
        sources: String,
        types: String,
        nst: Integer,
        dmin: Decimal,
        rms: Decimal,
        gap: Decimal,
        magType: String,
        type: String
      },
      geometry: {
        type: "Point",
        coordinates: [
          longitude,
          latitude,
          depth
        ]
      },
      id: String
    },
    …
```

There are a huge number of fields in this database , we have chosen to work with these fields:

- **Mag :**
  The magnitude for the event . a decimal number [-1.0, 10.0]
- **depth :**
  The depth of the event in kilometers .a decimal number between [0, 1000]
- **Sig**
  A number between [0, 1000] describing how significant the event is. Larger numbers indicate a more significant event. This value is determined on a number of factors, including: magnitude, maximum MMI, felt reports, and estimated impact.

- **Longitude**
  typical values [-180,180
  Decimal degrees longitude. Negative values for western longitudes.
- **Latitude**
  typical values [-180 ,180]
  Decimal degrees latitude. Negative values for southern latitudes
- **Place (string)**
  Textual description of named geographic region near to the event. This may be a city name, or a Flinn-Engdahl Region name.

There are a lot more fields that have usefull information but we didn't include in our visualization , for various reasons ., such as:

- alert : “green”, “yellow”, “orange”, “red” .
  this field is unreliable , most of the inputs are missing (null)
- Felt : The total number of felt reports submitted to the DYFI system.
  we didn’t include this parameter because in our visualization we work on a group of earthquakes , and the average of their data . having the average of felt reports for a group of events would give a false information . and this field needs another type of data visualization .
- Accuracy parameters : parameters that evaluate the accuracy of the given data . like magnitude Error , depth Error , horizontal, vertical error .
  (see more in parallel diagram section)

A sample of the data :

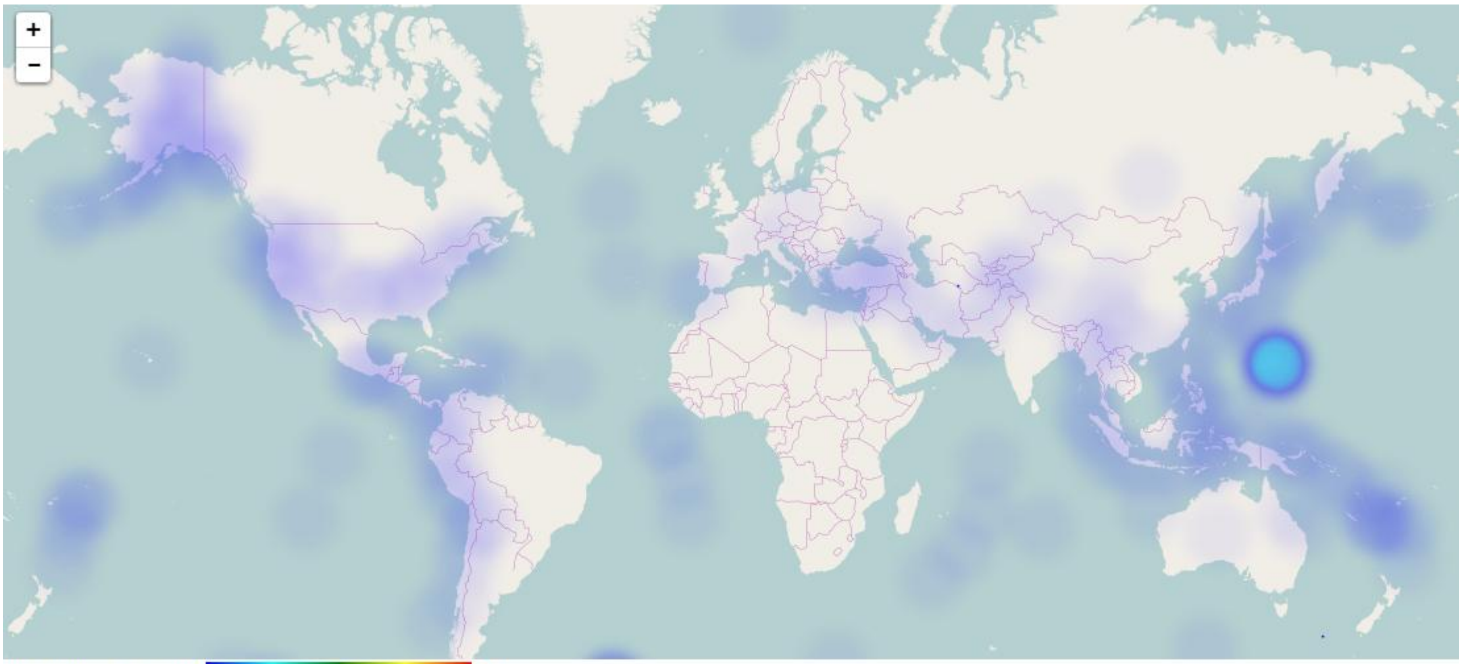

```
{
  "type":"Feature",
  "properties":{
    "mag":6.7,
    "place":"39km E of Namatanai, Papua New Guinea",
    "time":1472613096060,
    "updated":1472654483000,
    "tz":600,
    "url":"http://earthquake.usgs.gov/earthquakes/eventpage/us10006iy0",
    "detail":"http://earthquake.usgs.gov/earthquakes/feed/v1.0/detail/us10006iy0.geojson",
    "felt":2,
    "cdi":2.9,
    "mmi":2.95,
    "alert":"green",
    "status":"reviewed",
    "tsunami":1,
    "sig":691,
    "net":"us",
    "code":"10006iy0",
    "ids":",us10006iy0,pt16244050,gcmt20160831031136,",
    "sources":",us,pt,gcmt,",
    "types":",cap,dyfi,geoserve,impact-link,losspager,moment-tensor,origin,phase-data,shakemap,",
    "nst":null,
    "dmin":0.796,
    "rms":0.74,
    "gap":20,
    "magType":"mwp",
    "type":"earthquake",
    "title":"M 6.7 - 39km E of Namatanai, Papua New Guinea"
  },
  "geometry":{
    "type":"Point",
    "coordinates":[
      152.7879,
      -3.6914,
      499.06
    ]
  },
  "id":"us10006iy0"
},
{
```

User Task :

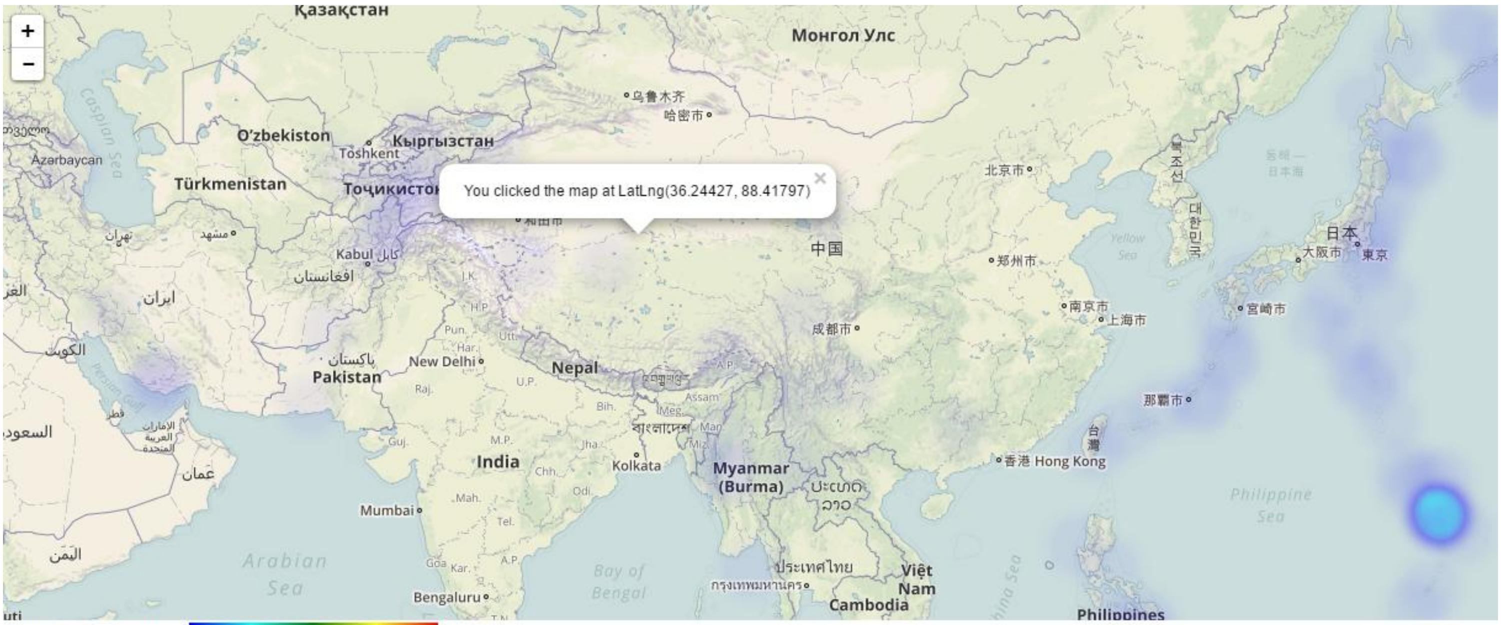
User Task is mainly exploration . Having such a huge database with well over 6000 events , by Viewing the database alone a user can't tell what there is in the database , can't find relation and patterns in the database . Moreover it would take him a lot of time just to view it .
Therefore our visualization came to help Users to explore the database in a very easy way and in a short amount of time .

by viewing the heatmap the user can tell where the earquakes events happened , how strong the events are , at which depth , how significant the event was and more .
and by viewing the parallel diagram , the user can get a much accurate information , actuall numbers and enables him to explore patterns in the data .

Heat Map



max mag : 7.7 low high



max mag : 7.7 low high


You clicked the map at LatLng(18.65145, 145.14587)


You clicked the map at LatLng(18.37017, 144.05823)


+
-

Mapping :

1. Position : The position of the earthquakes is taken from the fields longitude and latitude of the earthquake . the position of the dots on the map is calculated relatively to the size of the map and position on the screen .
Note : we are using a Two Dimensional maps , these maps are inaccurate because the don't show the actual size ratio of the lands .
for example Greenland is stretched and its shown bigger than it's actually is .
to correct this , a glob map or 3d map should be used instead .

2. color :

this field depends on a number of values :
first , the more number of dots in the same area the more redish the area is .
And , the closer the distance between the different donts in the area , the more red it is .
by zooming in and out on the map we can see that the color of the area changes .

The main parameter **Magnitude ,** the magnitude is not linearly calculated to affect the color , meaning : a higher magnitues gives a much bigger effect than lower magnitude (e.g 1 dot with magnitude of 5 equals more than 5 dots with magnitude of 1) .
the magnitude ratio is calculated by this formula :(http://earthquake.usgs.gov/learn/topics/how_mch_bigger.php)

for example : A magnitude 8.7 earthquake is 794 times BIGGER on a seismogram than a magnitude 5.8 earthquake. The magnitude scale is logarithmic, so

```
(10**8.7)/(10**5.8) = (5.01*10**8)/(6.31*10**5)
              = .794*10**3
              = 794
             OR
              = 10**(8.7-5.8)
              = 10**2.9
              = 794.328
```

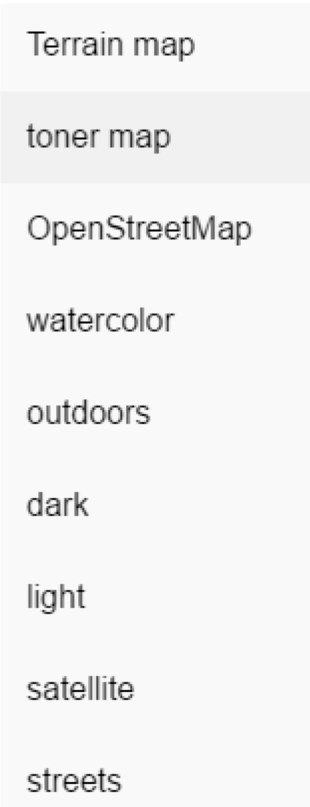
as shown in the formula above , we took the magnitude of the earthquakes and calculated their actual ratio , then to avoid getting a ridiculously high and low numbers , we normalized the data by the Highest magnitude in the map (or the selected area) .
by doing this we maintained the ratio between the magnitudes , and have given the higher magnitudes more effect on the color .
finally we defined a sensitivity field , for the porous of letting the user to change the sensitivity of the colors . by multiplying the resulted magnitude with this variable .

this variable is initialized with 1900 , and could be changed in a steps of 250 , by the user .

Tools :
1. change color sensitivity : changes the scale of the color , by a factor of 250 in each step .

2. Zoom in and out on the map . to see either individual events or a joined ones .

3 . change map . a user can select the map that best suits him .
for example . the terrain map and satellite maps   is more suitable for exploring the GeoLocations of the earthquakes , and to see mountains , rivers , forest , to zoom in on buildings and cities , so he can know if the earthquake location was close to populated area or not etc ..
tuner maps . dark and light allows the user to see a clean map with no extra information . so he can see the colors of the heat map earthquakes .
moreover , the user can zoom in on an area in satellite map , to see the terrains and the inhabitant , and then look at the same view in tuner map , to better see the colors of the earthquakes , without changing the zoom

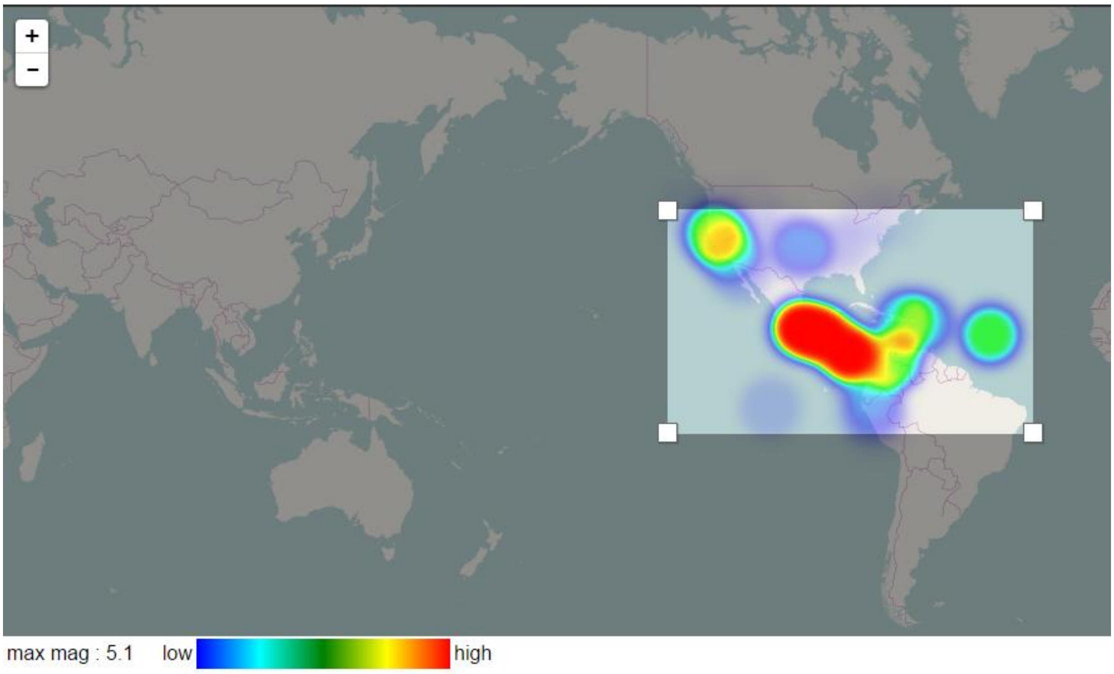


4. Select area :

This is an important tool , that took a lot of effort in coding .

this tool allows the user to select an area , and when selecting the area , it removes all other Earthquakes that are not selected, leaving only the user selection .

the user selection is then recalculated and normalized for the selections only And the maximum magnitude in the selected area is shown on screen .



How it's Done :

We divided the map into a small squares , and initialized an array to contain for each square , all the dots that are inside it's borders .

when selecting an area ,an array of all the selected dots is returned .

by checking if each dot Long and latitude falls inside the square border .

(that's a simplification of a complex process)

Parallel Visualization :

Axis :

Location axis :
this axis indicates the location of the earthquakes (wow) 0.
in our data there are well over 250 different locations,
and that is expected because the earthquakes happen across the entire globe .
showing all different locations in the parallel diagram would give a very dense and unreadable data .
(insert picture here)
How to Divide the world map ?

- by continents: which will give us a good clean parallel diagram, but with only 7 continents, we would not have much information, and it wouldn't be relevant because for example, we can see that the vast majority of the earthquakes happened in east and Middle East of Asia, west America (+ Alaska) . And in Europe Africa there are only a few . the magnitude and depth are calculated as average in the specific location. which in large areas such as the continents, these variables have no meaning .

- by grid:
we divide the map of the world into grids as many as we want and in what ever size for example 9 parts 15 or 36.
this method might be a little confusing for the users (viewers) ,

- by country: the parallel diagram would have 200-250 lines but it will give a much accurate information and would emphasize any patterns in the data,
moreover the majority of the quakes are mainly concentrated in 20-40 countries,
which in the diagram these countries would be separated from the other 200+ (because each have a different data) , but with a limited size axes, the lines would be very close to each other and visually unpleasant. Especially the location axes, that would be unreadable .

Solution :
Divide the World map into countries, but join the countries that have too few earthquakes into one Element We called it **others. This** way we were able to reduce the number of displayed countries to 20-30 , and it's important to note that these are the interesting countries. Which the user would look for when querying the information .
hurray …

Number of events :
the total number of events that happened in or around the Location .
We can see that the highest number of earthquakes happened in japan , Alaska , Afghanistan , Utah, Tennessee and pue New Guinea .
"Others" Element also have a high number of earthquakes , which is to be expected .

Mag x 10:
10 x the average magnitude of all the events that happened in the event .
this ranges between 0 and 100 .
we can see that the Locations with the highest magnitude average are the ones with the fewer number of earthquakes .
that doesn't indicate that these locations have more events with higher magnitude than the locations with the higher number of earthquakes .
for that there is the axis , High Mags Or high alerts .

depth :
the average depth of all the earthquakes in kilometers .the values ranges between -1 (undefined) and 1000 .
Missouri have the highest depth average

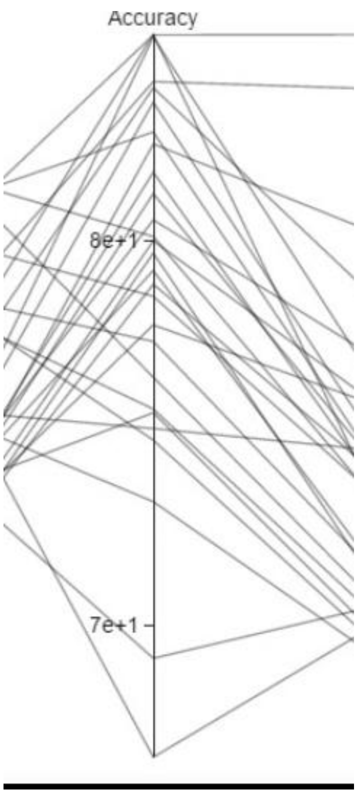
Sig:
the average significances of the event .
High alerts / high magnitude :
The number of events in that accrued in the location , with a magnitude over 4.5 .
this field is essential to determine the actual number of the significant earthquakes . and it should be noted that in the axis average mag we get accurately get this information .

Accuracy : as previously mentioned , we removed this axis from the diagram.

We calculated this field using :

- magnitude error ,depth error horizontal and vertical error , Dmin
- nst ([1,200], the number of seismic stations used to determine earthquake location (the higher the better)
- gap : [0,180] the largest azimuthally gap between azimuthally adjacent stations (in degrees) , the smaller this number, the more reliable is the calculated earthquake position.

  There is a connection between all these parameters , but we lack the equation to accurately calculate it . and thus we deemed our calculation unreliable .



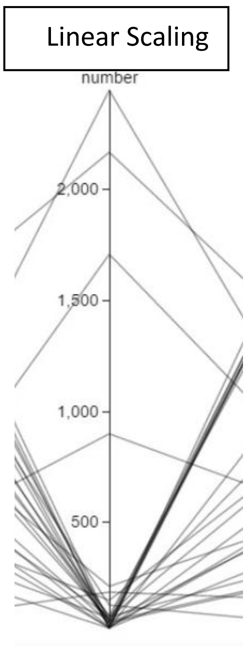
Patterns that could be found in the data :
we can see that the event with High significance and high alert levels (high magnitude) have a high **depth** .
for example in Alaska : there are 100-200 earthquakes with average mag ~ 5 and average sig ~600 and with average depth ~ 200 kilometers

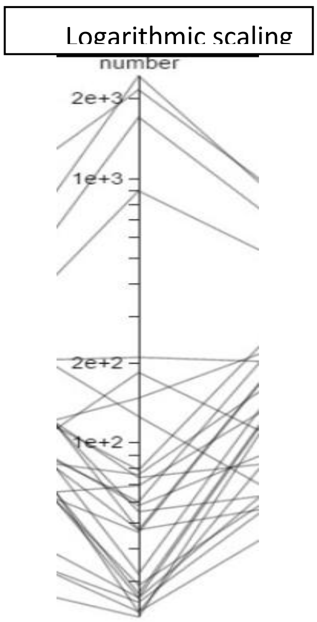
- ⇨ earthquakes with higher depth have a higher magnitude .
- ⇨ the number of earthquakes and the number of high alerts in the locations with high depth and high magnitude are close . which might indicate that high alerts events happen usually together with low alert events , but this is only an assumption .
  adding the time factor would prove this theory (actually we know that it's true)

scale :

the data ranges between [0,2600] and it's not evenly distributed among the sale . for example in the axis "number of earthquakes " and "depth " we can see that the lines mainly concentrate on both ends of the axis .
showing that in linaer scaling would give a bad indication and bad visualization of the data . since the axis is limited in size , many of the line wouldn't be visually seperable and there would be a large unused space in the axis (in the center).
therefore , we used logarithmic scaling. It solves the problem of space by almost evenly distributing the data across the axis . this is shown in the "number " and "depth " axis. we can visually separate the data lines.

the magnitude axis : the average magnitude range is [0,10] . and the values of the data lines are very close to each other . we can see a concentration in the range 4-5.
To better show the data we multiplied this axis by 10 .



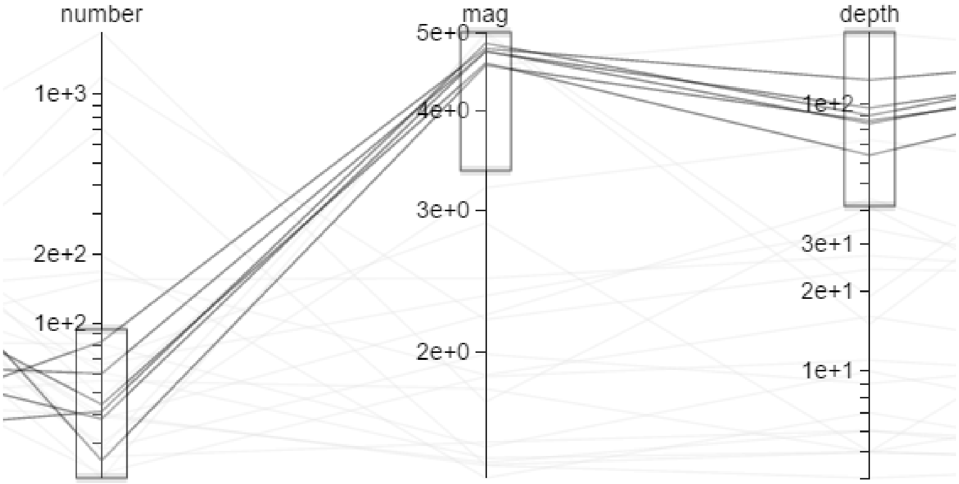


Parallel diagram Tools and features :

there are a number of tools available in this open source .

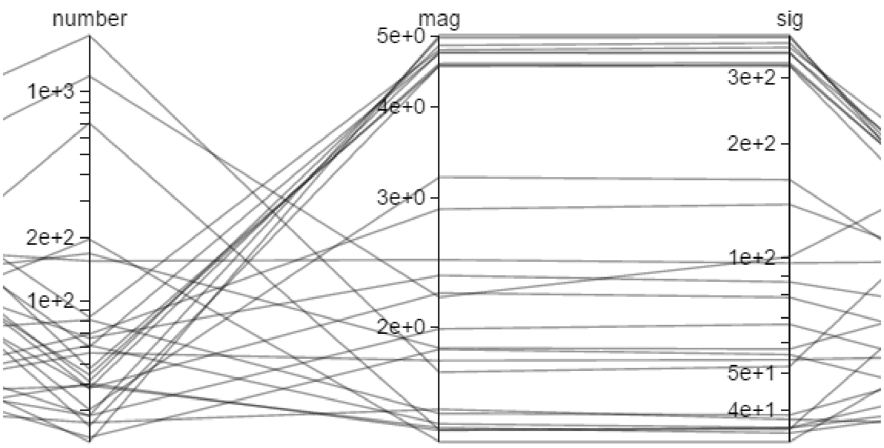
1. The selectors :

By selecting the lines it highlights them and it lowers the opacity of the unselected lines



There are 4 options for selecting the data .
none, 1D-axes ,angular , 2D-stums.

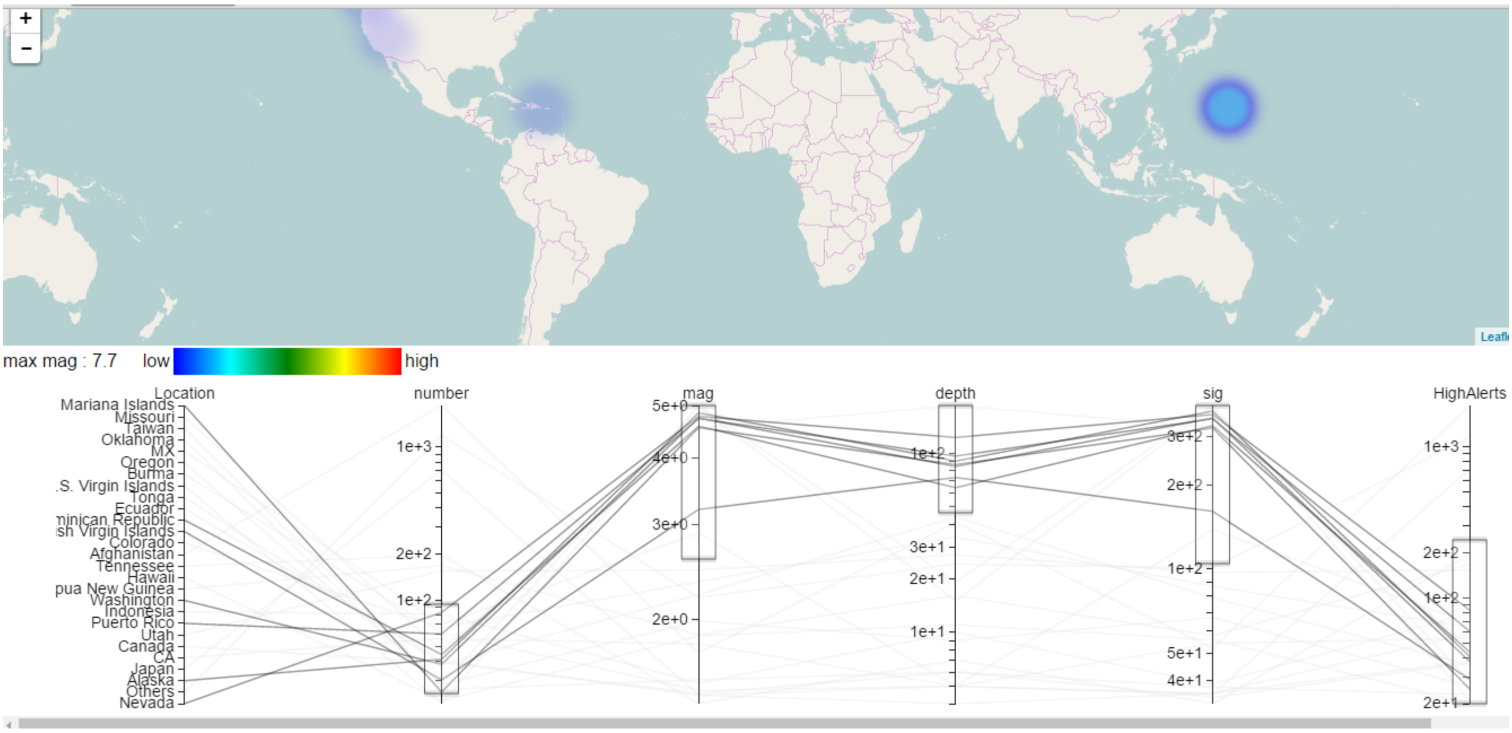
with the option to select "And" ,"OR" on the selected linens .
+ a rest button is available .
2. change the axes positions by dragging them .



3. by double clicking on the axis , the axes values changes from high to low and vise versa .

All these wonderful tools improve the user experience and data mining .

The Interaction between HeatMap and the Parallel visualization :



by choosing the split View from the Tools tabs , we enter a split view between the heatmap and the parallel visualization .
the haetmap shows the specific location and better visualize it . because what's easier than seeing the location of the earthquakes on the maps to determent their positions . and it shows the magnitude .
on the other hand the parallel visualization shows more data with actual numbers .
We Enabled the user to select an area in the parallel diagram , and only the selected area is shown on the map , and vise versa .
when the user select an area on the map , only the relevant data lines are highlighted in the parallel diagram , while the others (unselected) are shown with lower opacity .
the changes are shown in runtime and happens instantly .
this interaction between the heatmap and the parallel visualization allows the user to better query for information (better datamining method) .

discussion :

the locations are displayed by names , and not all user know the location of each country on the map . moreover , if a user want to select all the countries for example in America , he would have to handle select each and every country in the Location Axis , which will take a lot of time , and the user might miss a country or two .
on the other hand , in the heatmap , the user can see the map of the earth , and can see the locations (name and place) on the map .
the selection feature makes it a lot easier for the user to concentrate on a specific location , as oppose to the parallel diagram .

the earthquakes also occur in oceans and international water (not included in any country) .
in the heatmap there are only the magnitude and position , which are enough to compare number location and magnitude of groups of earthquakes , but the user needs more data . in the parallel diagram the user can get a lot more data , actual numbers , like average depth , magnitude ,high alerts , significance which are not available in the haetmap .
as previously mentioned , the color in the heatmap is very reflective , because it combines the location of the earthquakes , the distance between them and the logarithmically calculated magnitude value , to best reflect the actual difference between a clusters of earthquakes and on the individual level as well (by zooming in on the map).

we would like to have a field that describe the earthquakes damage, either financial damage , death tolls , damaged building number and status .either one of these parameters would have given our visualization more value , in determining the damage that are caused by earthquakes and not just data about them . but we couldn't find an open database containing such data

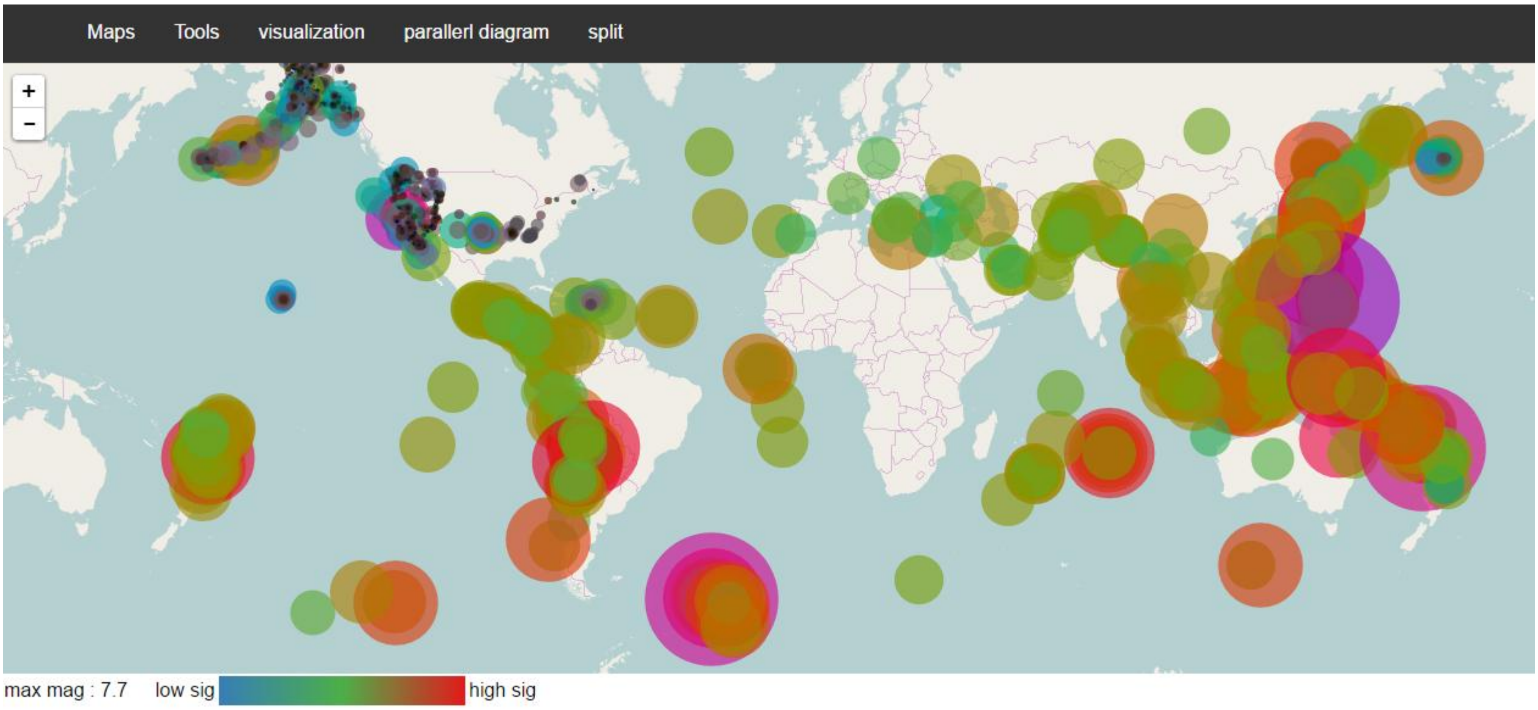
with a little more updates and changes , our visualization could be used by professional seismologist .
another thing is , the data source can be changed to be dynamic .
to take the data directly from the website .

by viewing the heat map with satellite map , views can explore the Geolocation of the earthquakes , and might find direct relation and pattern between the Geolocaction and earthquakes .
for example : it's less likely for earthquakes to happen in middle of the cotenants' , and a lot more likely to happen around costal areas .
and earthquakes with greater depth tend to happen near high mountains.(I guess)

Note : the earthquake location , magnitude and depth can be measured to some accuracy range . and there is always an error factor in these measurements due to hardware constrains , and the placement of the seismic stations that measure the earthquakes data .
for each earthquake there are a number of parameters that evaluate the error , like depth error magnitude error and such (see the accuracy field that was previously discussed)

Allover , what lack in the heatmap can be found in the parallel and vise versa . these two visualization nicely interact together to give the user the most Visualization Value .

Circles (bunus)



Mapping :

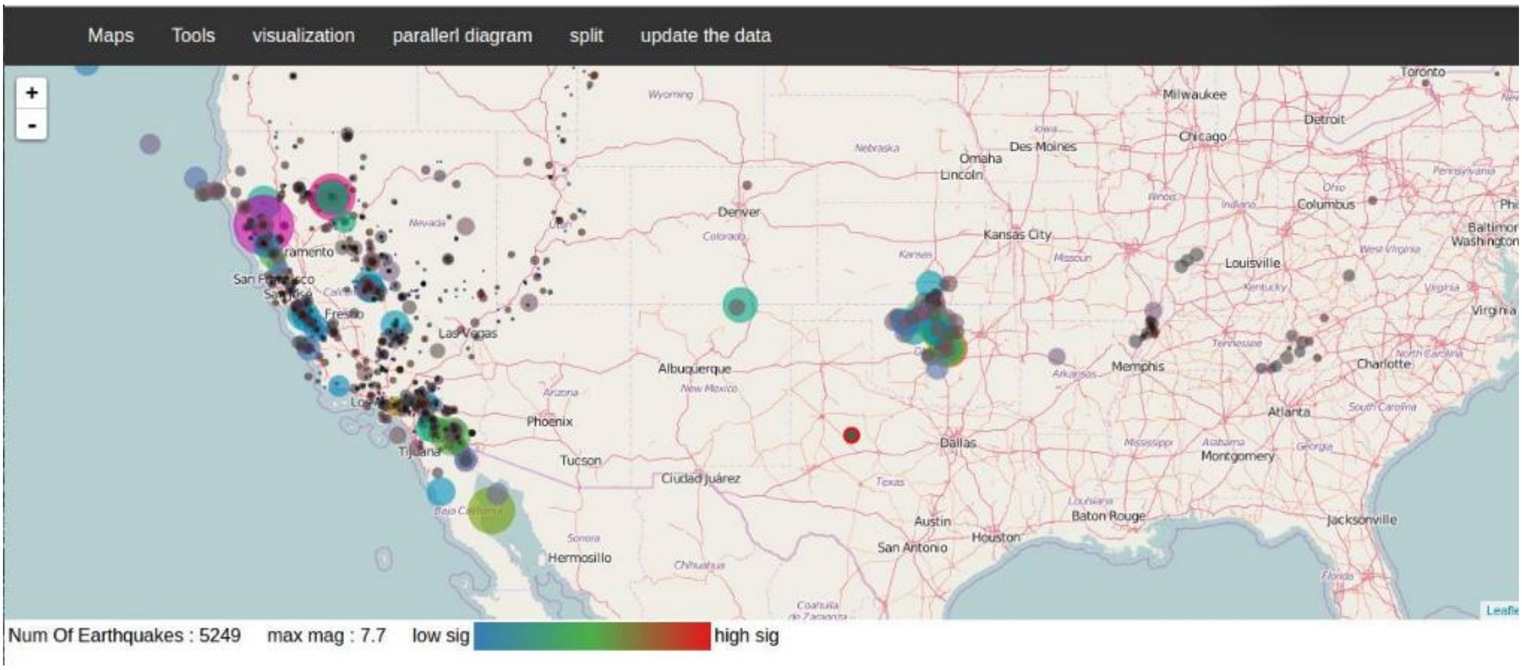
- radius of the circles -> (magnitude^2) , not like the heat map.
  this is in order to amplify the difference in magnitude betteen two different earthquakes . although this might give the user a wrong impression
- Color -> sig field
  starting from blue (lowest significancy) to red (high alert earthquakes) .
  the black colors mean that there was no information about the earthquake.
- Position of the center of the circle - > the average location of the joined earthquakes positions (longitude and latitude

In this visualization we can compare the significance of the earthquakes and the magnitude . by zooming in on the map the circles get separated , and we can then clearly see individual earthquakes .
the circles have opacity , so that One circles doesn't completely block the others , moreover the resulted color of many circles on top of each others is also meaningful .

cons : there are no actual numbers shown in there .
and its difficult for the viewer to determine the center of the circles (unless he zooms in enough)
When there are many circles on top of each others , their color becomes constant .

And the user no longer can tell whats going on there

Note : this is a separate visualizations was created as a bonus , it doesn't run with the parallel visualization , and doesn't have the select area feature like the heat .



Final thoughts :

Our visualization is nice clean, easy to user and understand .
it saves the user a lot of time in querying the data . it shows patterns , data behavior and more .
the visualization gives a fun experience and an interactive way for data mining . it holds value in highlighting the important data (e.g earthquakes with high magnitude)
We had fun creating it and a lot of experience in try and learn mistakes .
we built the code in a way to enable expansion in the future .
adding fields and variables is easily applied and with a few changes .

Cons :

1. The map we are using is a 2d – map , it is stretched in north and south , and thus not showing the actual size ration of the lands.
2. The parallel diagram missing the feature of , when selecting a line (or data) it shows its actual stats .
3. The color of the parallel is black . by mapping the data lines (paths) into different colors , it would enable a better separation of the lines .
4. the selection on the map and parallel visualization still have some bugs . .

Sources :

http://earthquake.usgs.gov/earthquakes/feed/v1.0/geojson.php
https://github.com/heyman/leaflet-areaselect
https://syntagmatic.github.io/parallel-coordinates/ .