**Predicting Economic Indicators**

**Grace Gee, Eugene Wang**

Our previous notebook is attached below.

**Background and Motivation.**

We’ve pivoted after not being able to find a complete historical data set on sovereign bond indicators. The only sovereign bond indicators public dataset was from Moody’s which has sporadic data for different countries and does not include the most recent values. We thought of innovative ways to extract bond ratings data such as scrapping historical archive of wikipedia (<http://en.wikipedia.org/wiki/List_of_countries_by_credit_rating>). However the free internet archive only included data as far back as 2010. This doesn’t really work since most bond ratings are only updated semi-annually, which gives us a dearth of data to visualize.

In light of this, we switched to related economic time series, specifically 6 indicators that are related to the sovereign bonds ratings:

* Budget balance as a % of GDP
* Foreign debt as a % of GDP
* Real GDP Growth
* Inflation
* Current account as a % of GDP
* Unemployment rate

These indicators are primarily used to show the fiscal and monetary health of a country, which are close to our original intent for the project on sovereign bond ratings.

We are very interested in the field of financial indicators and have done previous work in sentimental analysis of FOMC statements on gold. Our previous work involved building a classifier that predicts the directionality of price movements. However, this project will be our first attempt to use time series to predict continuous values.

**Project Objectives.** Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

Economic indicators are often interrelated. For example, the unemployment rate of a country is famously modelled by the phillips curve (inverse relationship between the two).

The purpose of the project is to allow our users to manually construct their own multilinear model to predict their chosen time series data (out of the 6) using the other 5 unchosen time series. This provides an exploration tool for those interested in discovering the relationship between indicators (e.g. they can increase the coefficient of one indicator and see if there is a better fit between the predicted and actual value for chosen indicator). On a deeper level, we also want to show our users that it is not easy getting a fit for a multilinear regression model manually. Intuitively, we want to show that the coefficients for a linear regression model with more than 1 independent variable is hard to obtain by inspection.

**Data.** From where and how are you collecting your data? If appropriate, provide a link to your data sources.

We will collect time series data on

* Budget balance as a % of GDP
* Foreign debt as a % of GDP
* Real GDP Growth
* Inflation
* Current account as a % of GDP
* Unemployment rate

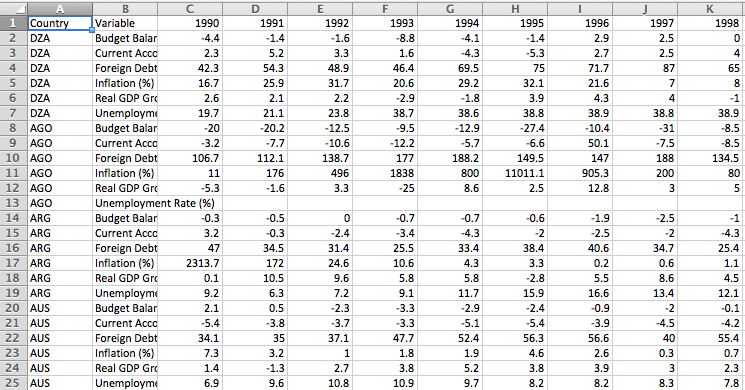
The data spans 1990 - 2013 and will be collected from <https://www-countrydata-com.ezp-prod1.hul.harvard.edu/index.php/customer/countrydata/>

The data contains all the 6 time series for all the major countries in the world. By ensuring that we obtain data on different countries from the same source, we can minimize discrepancies between numbers when we are comparing between countries, since they use the same methodology.

**Data Processing.** Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

When we first obtain the data, it is in csv format with the first column as country names and second column as the indicator time series. Though the data is already in a pretty format, we still did the following wrangling in javascript (check data\_wrangling.html)

1. Convert country names to alpha3 code. This part of the wrangling was carried out manually. We obtain a list of country names and alpha3 code from (<http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3>) and convert the names accordingly.
2. We used javascript to convert the flat csv data into json. The top node of the json are the countries’ alpha3 code followed by the indicators they have. Lastly, each date is a key attached to a specific value for the indicator.

Here’s a screenshot of the data we obtained after we switch out the alpha3 code.

**Visualization.** How will you display your data? Provide some general ideas that you have for the visualization design. Include sketches of your design.

1. Visual map -

Since this includes data for countries, we used a world map. We included zoom, tooltip, clickable, and also showed a heat map depending on the relative value of the indicator.

Most importantly, since the indicators’ values change across time, we added a simulation component when users can click play and see how the relative values (shown as colors) change across time in a continuous simulation!

When users click on a country, we show a button asking them if they want to predict indicator values for the specific country.

2. Line graph -

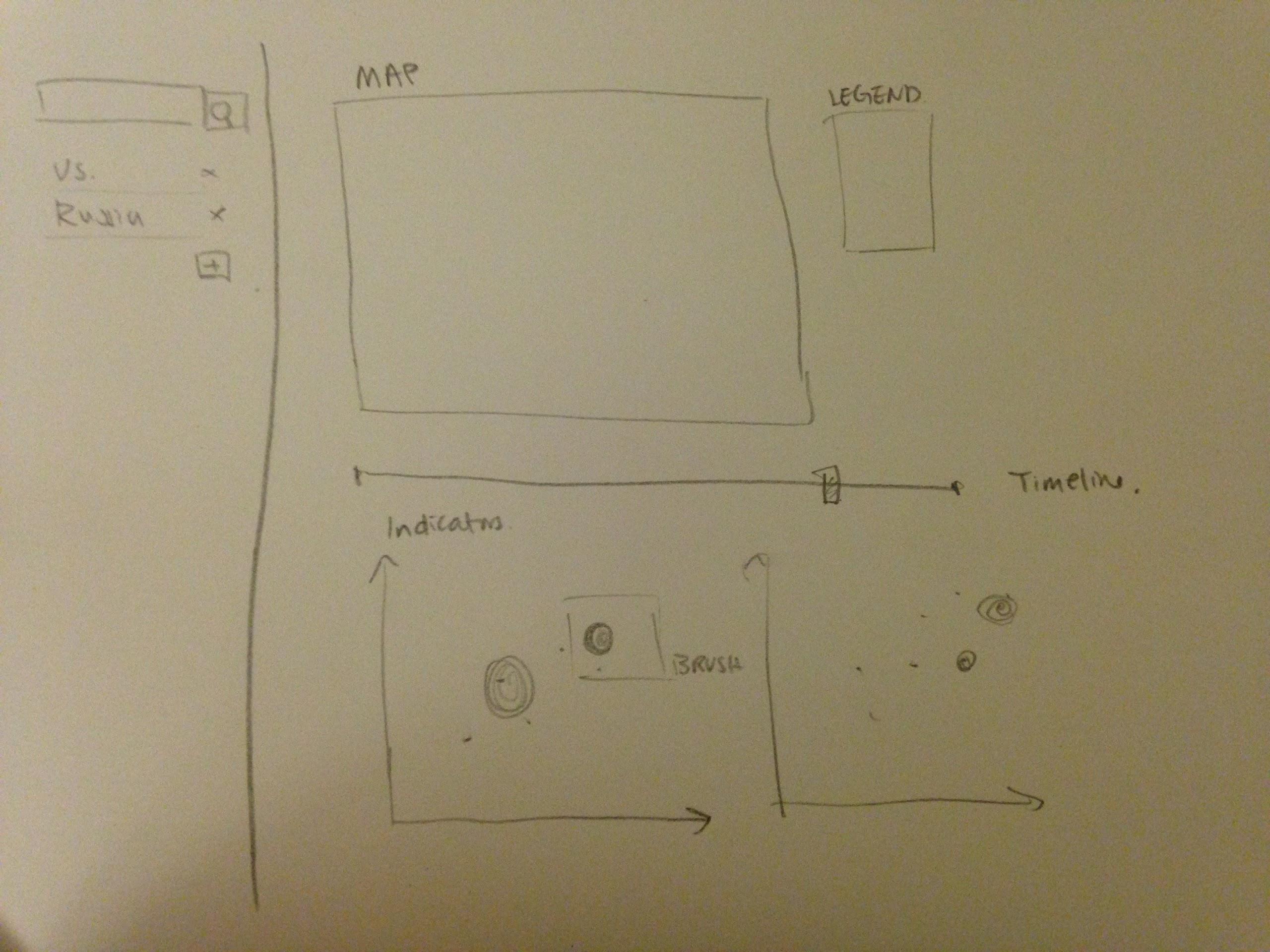
We include a line graph that shows both the actual value for the specific country’s economic indicator across time as well as a predicted value. They will be distinguished by color and has awesome features like scroll zoom, click zoom, tooltip, highlight the line and even clickable legends!

Most importantly, the line graphs have awesome transition animations!

We also have a multilinear regression model equation above the line graph. Users can manually change the coefficients, hit enter and see how the predicted line graph changes. The goal is for the users to try all sorts of coefficients and try to bring the predicted value as close as possible to the actual value graph.

4. Donut chart - After users are happy with their model, they can click submit and we will show an animated donut chart telling them how much mean percentage error their model has.

This sketch is from the previous process book but the general gist of the idea is still in place in our new design.



**Must-Have Features.** These are features without which you would consider your project to be a failure.

1. Map of countries with bond ratings highlighted in key. Hovering gives more country-specific detail (eg country name, significant indicator numbers). Users should be able to toggle between different indicator maps.
2. A ‘play’ button to simulate and show animations of the changes in map values (through colors) over time.
3. For each indicator, a line graph of the actual value vs predicted value for the indicators.
4. Prediction accuracy of the multilinear regression model, as shown in a animated donut chart.
5. Scatterplot to show the target indicator to predict and the other indicators the user is trying to build the model off. This is more of a helper tool to let the user get an intuition of the relationships between variable.

**Optional Features.** Those features which you consider would be nice to have, but not critical.

1. We can explore even more time series indicators or models such as neural networks or regression based decision tree. However, admittedly, this may be less interactive since some of these classifiers may be quite foreign to the user.

**Project Schedule.** Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

April 6 - April 13

\*propotype due before april 10

Nov 26 - Dec 3

Build Classifiers

Eugene - Run words through Linear Regression classifier

Grace - Run words through SVM or PCA

John - Run words through Neural Networks

April 13 - April 20

April 20 - April 27

April 27 - May 4

May 4 - May 8

Analysis and Writeup

***Older Process Book***

**Sovereign Bond Rating Prediction**

**Grace Gee, Eugene Wang**

**Background and Motivation.** Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

Sovereign bond ratings signal the amount of risk in investing in a particular country. They tend to incorporate quantitative measures of political situations, economic status, historical resilience amongst many other factors.

Our project will be composed of 2 components:

1. Analyses - Prediction of sovereign bond ratings given historical time series data. We will explore different classifiers, such as SVMs and neural networks that are well suited for classifying fixed categories.
2. Visual Exploration - Since we will have just 2 strong indicators of interest

We are very interested in the field of financial indicators and have done previous work in sentimental analysis of FOMC statements on gold. Our previous work involved building a classifier that predicts the directionality of price movements. However, this project will be our first attempt to use time series to predict fixed classes.

If time allows,

1. model accuracy

2. maps

3. more detailed country info such as matching dates.

**Project Objectives.** Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

What economic indicators and time series to predict sovereign bond ratings.

Countries with bond ratings over time

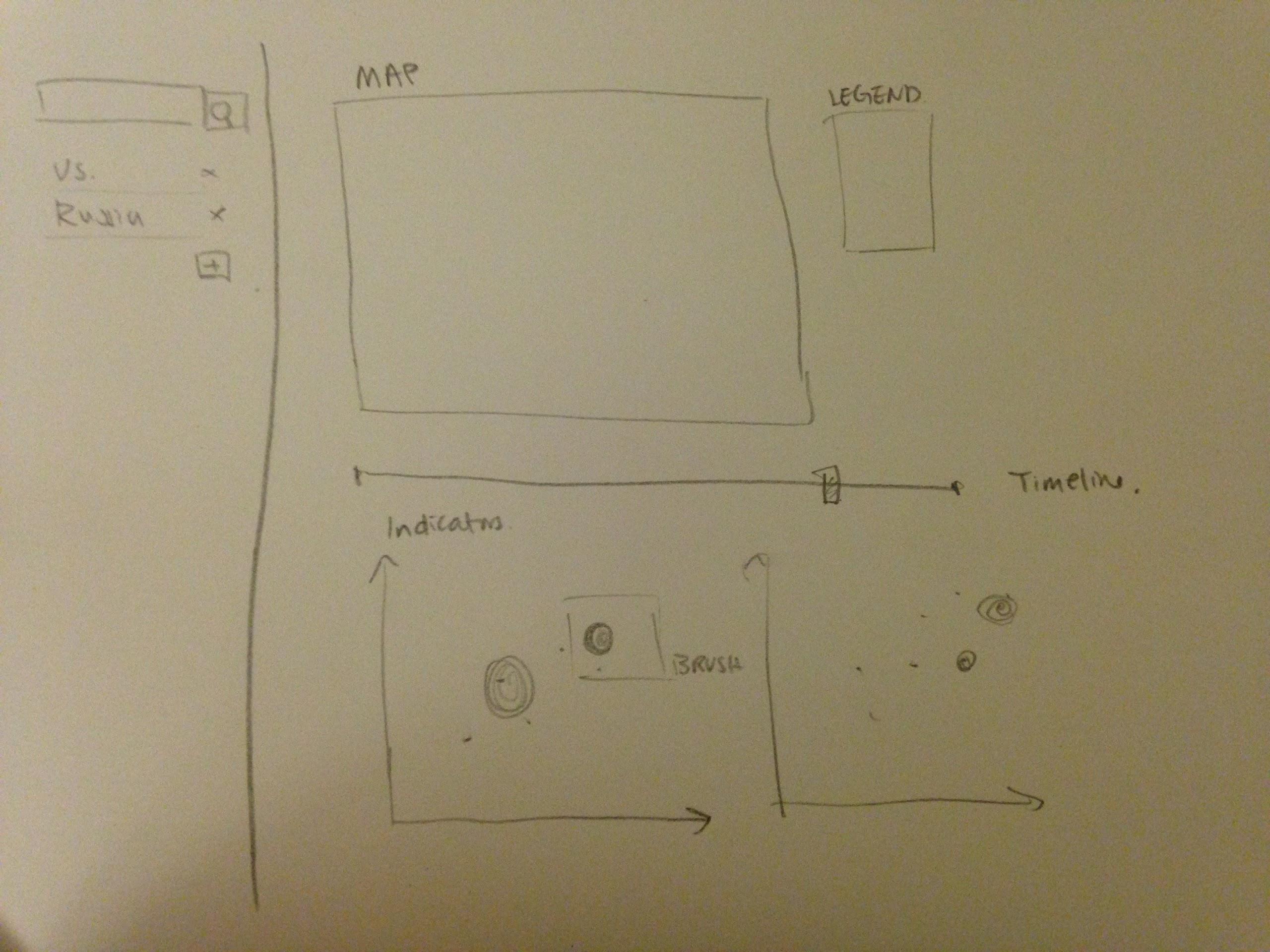
**Data.** From where and how are you collecting your data? If appropriate, provide a link to your data sources.

Sovereign debt ratings are provided by Moody’s, Standard & Poor’s and Fitch. We will be collecting Moody’s ratings (the leading rating agency) from Bloomberg at HBS.

**Data Processing.** Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

1. Features: Government debt-GDP ratio, Household debt-income, GDP, Inflation, Bond yields, Unemployment rate, PE ratio, Debt-Equity ratio
2. Target: Sovereign bond ratings are categorized into 9 sets of (AAA, AA, A, BAA, BA, B, CAA, CA, C).
3. Methodologies:

**Visualization.** How will you display your data? Provide some general ideas that you have for the visualization design. Include sketches of your design.



**Must-Have Features.** These are features without which you would consider your project to be a failure.

Analyses

1. Test methods of interpolation of monthly, quarterly, weekly data & methods of filling missing data
2. Preprocessing features such as SVD and PCA to reduce the dimensionality of the feature set to circumvent the problem of curse of dimensionality.
3. Categorical classifiers such as decision trees, clustering, neural network and nearest neighbors.

Visualizations

1. Map of countries with bond ratings highlighted in key. Hovering gives more country-specific detail (eg country name, significant indicator numbers). Users should be able to toggle between “Actual” Moody’s ratings and “Predicted” ratings maps.
2. For each feature, scatter plots of countries (eg feature vs rating).
3. Prediction accuracy of algorithm in predicting sovereign bond ratings. This number changes with slider event.
4. Timeline slider to show prediction and actual rating differences over time. Event triggers scatter plots and model accuracy.

**Optional Features.** Those features which you consider would be nice to have, but not critical.

1. If time permits, we will explore the relationship between the sovereign bond rating and significant economic events over time.
2. We can explore even more time series features or classifiers such as kernel-based functions including RBF neural networks.
3. Features sidebar to allow users to add/remove indices. Map and accuracy will reflect the changes in prediction from included features.

**Project Schedule.** Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

March 23 - March 30

Data preprocessing & Set up Feature Analysis

Eugene - Fill in missing data/interpolate data; Literature of Classifiers used in rating classifications

Grace - Bloomberg download; Build classifier modules that take in time series features

March 30 - April 6

Visualization of results

Eugene - Build map

Grace - Build scatter plot queries

April 6 - April 13

\*propotype due before april 10

Nov 26 - Dec 3

Build Classifiers

Eugene - Run words through Linear Regression classifier

Grace - Run words through SVM or PCA

John - Run words through Neural Networks

April 13 - April 20

April 20 - April 27

April 27 - May 4

May 4 - May 8

Analysis and Writeup