**dataviscourse-pr-visualize\_material\_informatics**

CS 6660 project

**## Basic Info**

Title: Visualize material informatics

Team Members: Yucheng Yang, Steven Kauwe

Email: dominate0704@gmail.com, jkkauwe@gmail.com

UID: u0878896, u0742731

Repository: [https://github.com/kaaiian/dataviscourse-pr-visualize\\_material\\_informatics](https://github.com/kaaiian/dataviscourse-pr-visualize%5C_material%5C_informatics)

**## Background and Motivation**

Discuss your motivations and reasons for choosing this project, especially any background or research interest that may have influenced your decisions.

This project comes from the research of Steven Kauwe who investigates the use of materials informatics for predicting material properties as a Ph.D student in the Department of Materials Science and Engineering. The current state of research in that area has focused on showing proof of concept machine learning using various data types. Because materials science is inherently chemistry based, researchers have found it is useful to represent chemical systems based on their constituent elements. This is the foundation for what we will refer to as the composition-based feature vector (CBFV). The CBFV is a common method of vectorizing chemistry space, allowing a representation of physical composition to be parsed into traditional machine learning frameworks. Because of the relatively new interest in this area, some fundamental ideas have yet to be implemented in an open and accessible way. In particular, to our knowledge there exists no central tool for researchers to evaluate the effect constituent elements have in prediction error. This project aims to provide a proof of concept tool for just such composition based error exploration.

**## Project Objective**

Questions we would like to answer with this visualization.

How does the relationship between elemental properties affect the emergent properties of the molecule?

What elements have the most error associated with them?

Are non-real compositions less accurate that real compositions for this data?

Are there patterns in error relative to feature values, elements?

Do compounds with similar elements cluster close to each other in feature space, and predictions?

How does proximity to similar elements affect the error of a compound?

What would you like to learn and accomplish? List the benefits

We would like to learn how whether certain chemistries are more prone to error in our machine learning models. We would also like to know if this is an error is caused by the chemistry itself by plotting and visualizing it across different properties. This project will improve our ability to work as a team creating a complex and interactive web page. Presenting the project will develop our ability to show off the features and practical uses of the tool we create.

**## Data**

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

Data is collected from the computational materials property repositories: Automatic - Flow for Materials Discovery (AFLOW), and The Materials Project (MP) who do high throughput density functional calculations (DFT) to compute material properties as a part of the Materials Genome Initiative. The AFLOW data was retrieved directly from their search tool after selecting the required electronic property, Band Gap (http://aflowlib.org/advanced.php?search=). The MP data was obtained via a third party, Citrine Informatics, at the web address https://citrination.com/datasets/150675/. The data is then vectorized via composition and used in a machine learning scheme that saves the cross-validation predictions and trains a model for future predictions. The error associated with the cross-validation results comprise the interactive aspect of the project.

**## Data Processing**

Do you expect to do substantial data cleanup?

The data may have duplicate entries and compositions that do not represent real compounds. These will need to be addressed. The duplicates values can be solved by averaging all reported values. The non-real entries will not be rejected from the data set, but will instead be marked to let researchers know how this effects model error. To do this, we will need to cross-checked all AFLOW and MP data against a crystal structure database. This will be done with the Pearson Crystal Database (PCD).

What quantities do you plan to derive from your data?

We expect to derive a predicted value of band gap from our data. Error metrics for all cross-validation predictions will also be obtained during model creation. Before use online, we will need to classify the data entries by the constituent elements and whether we could confirm if a real structure is reported for the composition. The compositions will also be vectorized in a composition-based feature vector (CBFV) which is used in the machine learning. This will start as a vector with approximately 250 elements, however principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) will reduce this down to 2-5 dimensions for human interaction.

How will data processing be implemented?

Data processing will be implemented in python 3 using the pandas and sk-learn libraries. All data will be processed offline then compactly represented as CSV’s or JSON’s.

**## Visualization Design**

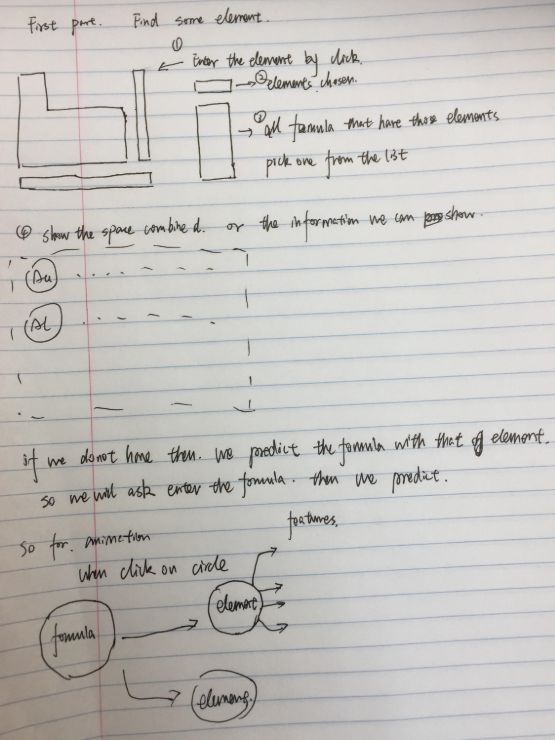
How will you display your data?

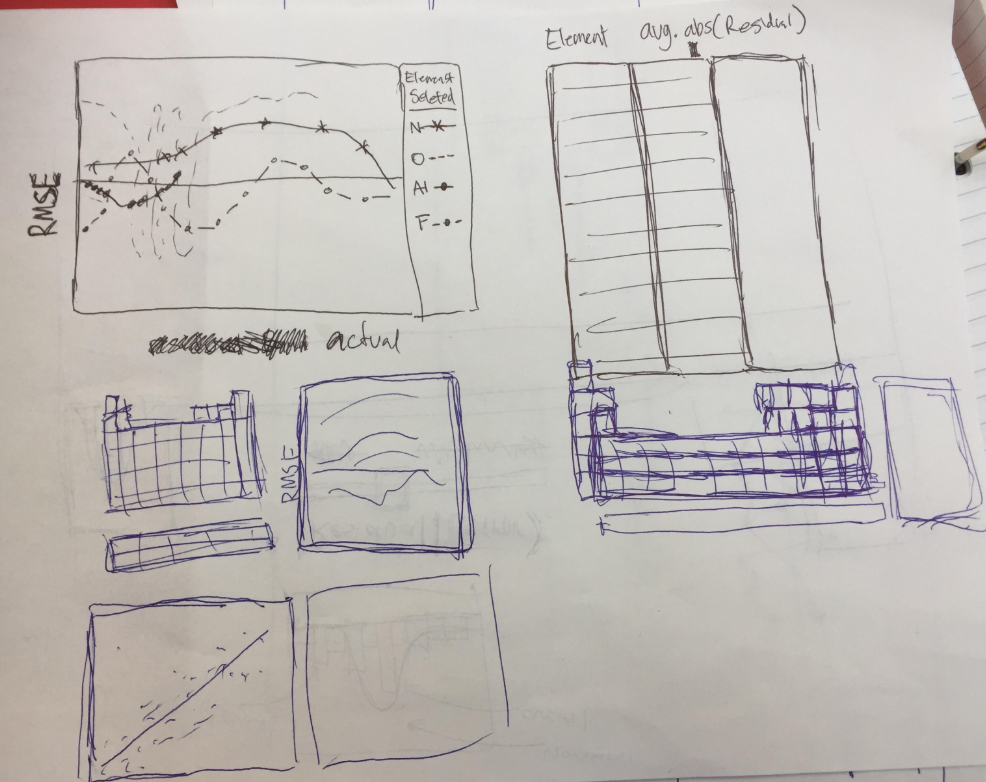
This data will mostly be displayed using scatter plots,distribution plots, trees, and a interactive periodic table. The markers will typically represent individual chemical compounds.

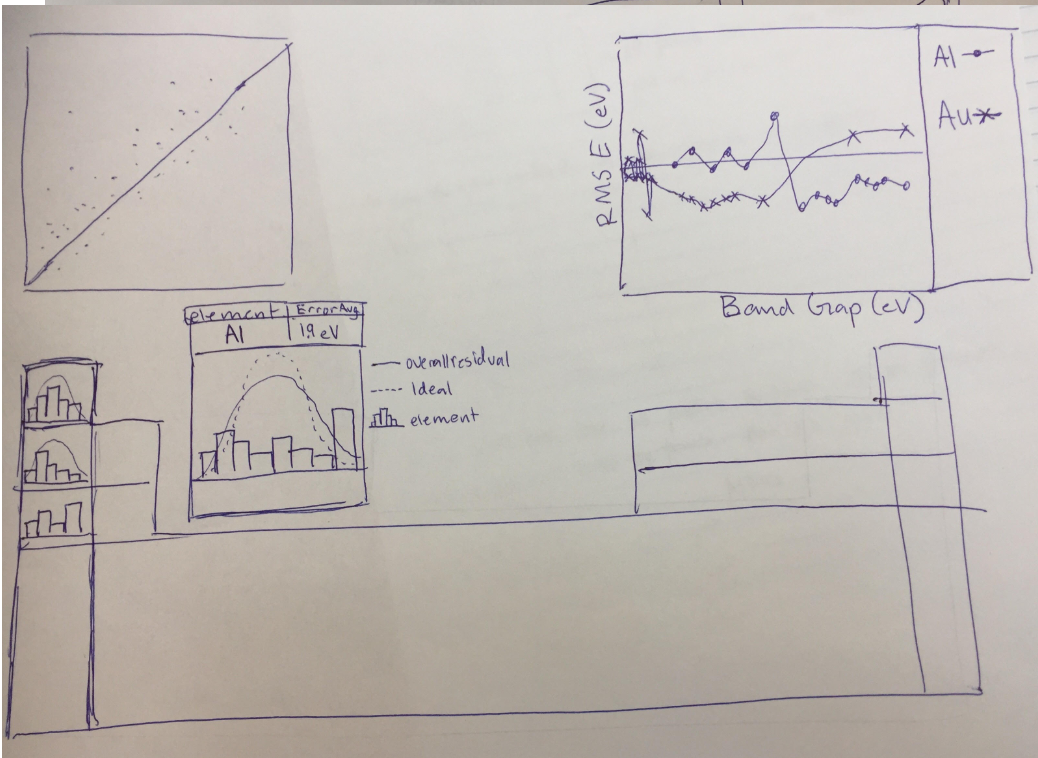
Provide some general ideas that you have for the visualization design.

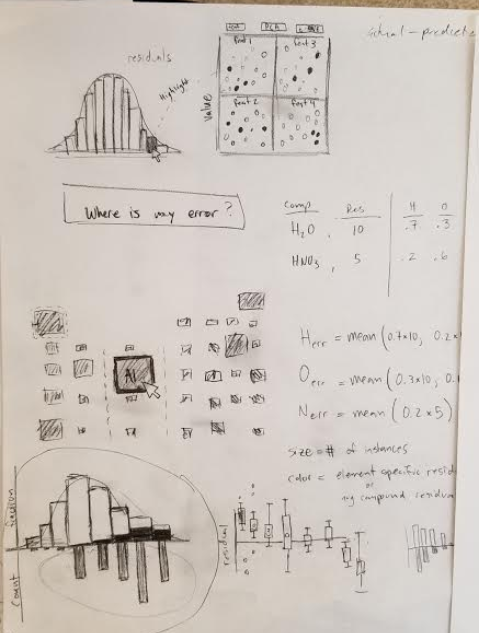
The user should be able to interact with a periodic table to select compounds of interest. These, in turn, will be highlighted in the other views to allow the user to quickly and intuitively explore the interaction of composition with error. The user could interact with a distribution plot to filter for values they would like to consider. It could also be possible for the user to put in their own composition and generate a prediction.

Develop three alternative prototype designs for your visualization.

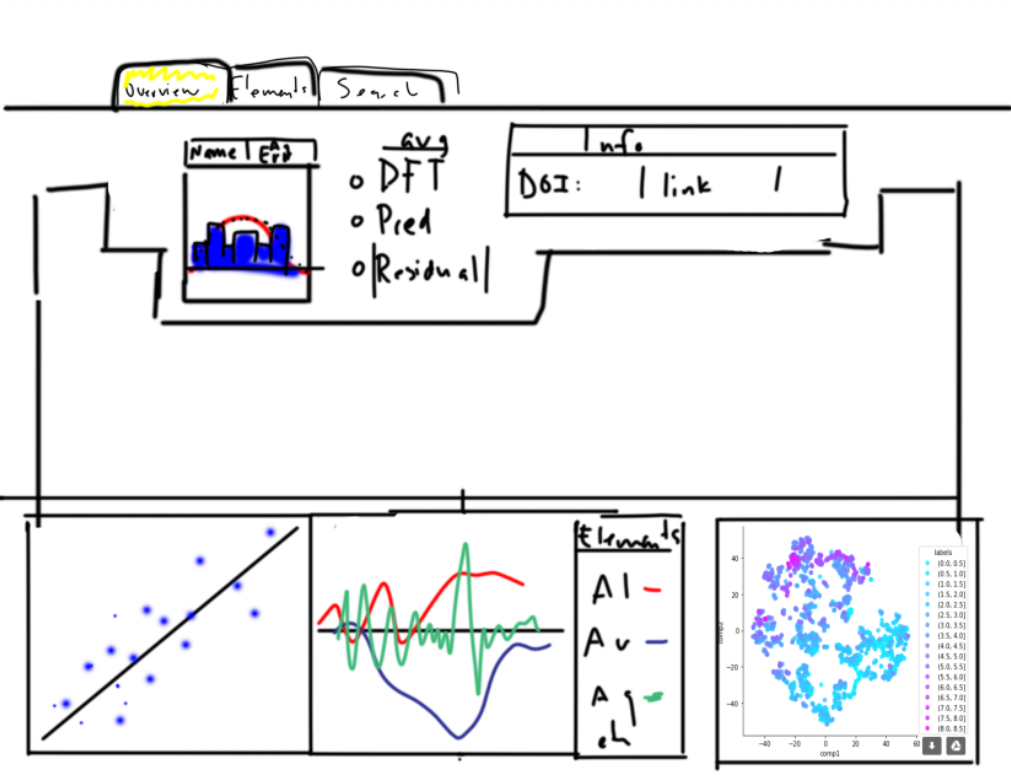


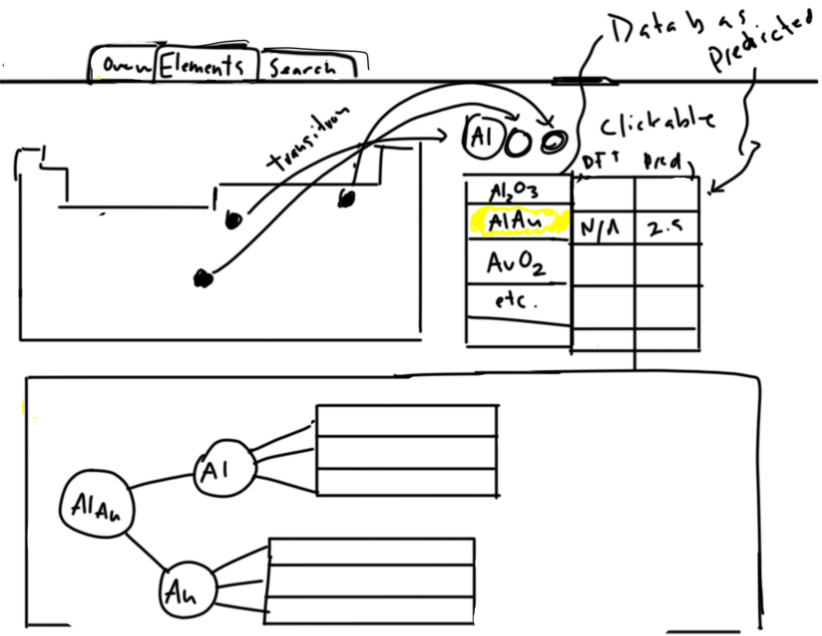


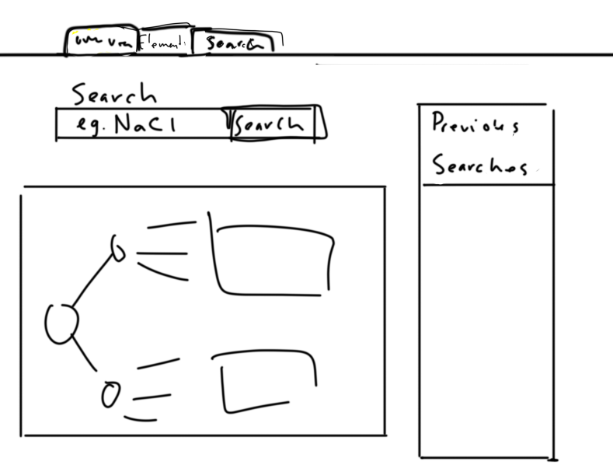




Create one final design that incorporates the best of your three designs.







Describe your designs and justify your choices of visual encodings.

There will be three pages: overview, elements, and search. Each page will independently operate, providing a specific use for the user. The overview will show overall trends in the data and allow for high level exploration. The elements is designed to let users to see specific information associated with the individual elements and the formulae they are a part of. The search page allows users to see previous searches and generate new predictions.

This three page design is made to be captivating while simple to interact with. It gives the user enough choices to spend significant time on the webpage, but also limits the available information to prevent a feeling of being overwhelmed. The representation of data is standard for most publications in the field and is appropriate.

**## Must-Have Features**

List the features without which you would consider your project to be a failure

Completion of the overview page would constitute success. This is the main aspect of the project and provides the user with the most functional information.

## Optional Features

The element and search pages aim to present our more detail in the data set. These are less complicated and serve a secondary purpose of which is helpful for evaluating individual formula, but not critical for gauging the error in general.

## Project Schedule

Oct 27- Nov 2

Make a relation map for each visualization features

Organize features

Make a workflow

Complete a description for the outline of the overview page

Nov 3-9 (MILESTONE DUE)

Transform data to an appropriate format

Structure the project architecture

Outline of our web page (coded) \*sit down and program together

Assign features to each member

Creator & Tester for each feature (code review)

Write up milestone report

Nov 10-16

Code up the features for the overview page

Nove 17-23

Code features for elements and search pages

Finish coding and integrate all pages together

Nove 24-30 (Project Due)

Make changes to features to meet user need based on user testing