

Data Analysis Competition

There are an enormous quantity of interrelated variables that drive enrollment, retention, and graduation rates in higher education. Please help us explore the big, complex, and interrelated data affecting these important metrics!

In this competition, you are asked to analyze data on thousands of higher education institutions in the United States across 2013, 2014, 2015, and 2016 in order to make predictions about enrollment, retention, and graduation rates for hundreds of other universities. In short, you will create predictive models that will *learn* from complete training datasets on many US universities in order to predict responses for new inputs.

Training Data

You will use this training data set to develop/train predictive models in order to predict enrollment, retention, and graduation for hundreds of universities.

The training data set consists of four comma-delimited CSV files, each corresponding to a different year (2013-2016).

Download Training Data here:

- [training2013.csv](#)
- [training2014.csv](#)
- [training2015.csv](#)
- [training2016.csv](#)

Each of the four files contains 1491 of rows, each of which corresponds to a U.S. college or university in addition to a header row. The names of each institution have been removed. Depending on the year, each row contains between 29 to 32; different variables that explain something about the university or the state in which that university resides.

All variables are defined in this Data Dictionary:

- [DataDictionary.xlsx](#)

NOTE: For those that are using Machine Learning methods to develop a predictive model, you will likely need to partition these data set into both 1) training and 2) test data sets in order to test and validate your models prior to applying your models against the *Submission Data* (below).

Prediction Data

Once you have developed, trained, tested, validated, and improved your predictive model(s), you will run your trained models against this "new" Prediction dataset in order to estimate response variables related to: Enrollment, Retention, and Graduation Rates.

Download Prediction Data here:

- [prediction2013.csv](#)
- [prediction2014.csv](#)
- [prediction2015.csv](#)
- [prediction2016.csv](#)

The Prediction Data is formatted similarly to the Training data in four CSV files corresponding to the years 2013-2016. These four data files contain 298 rows each including a header row. Each row describes an institution that was purposely excluded from the Training Data, and the prediction dataset is missing data for the three columns related to Enrollment, Retention, and Graduation Rate response variables.

Your job is to fill in these missing variables for these new institutions using predictions derived from your predictive model.

Submission Data

You will submit a single CSV data file that characterizes the predicted relative change in Enrollment as well as Retention, and Graduation for the years 2013-2016 for hundreds of higher ed institutions.

Specifically, you will submit a single CSV file with 298 rows (*including one header row*) that contains the following columns:

Data Viz Competition

Use any data you like, and represent them in the medium of your choice. Participants are encouraged to visualize the information in creative ways that reveal interesting features and stories found in your dataset. You may use any tool(s) of your choice to explore the data and produce visualizations, for example: Excel, [OpenRefine](#), [Tableau Public](#), [Observable](#), or [Jupyter Notebook](#).

Your submissions can be static or interactive visualizations of any type that can be accessed by the judges. Submissions will be accepted in any reasonable format that does not require judges to install extensive dependencies or new software. If possible, make your visualization publicly available on the web, using services such as [Tableau Public](#), [GitHub](#), [nbviewer](#), [BLOcks](#), [RPubs](#), [Google Charts / Fusion Tables](#), or others. In that case, when making a submission, please do not upload a file, but provide a link and description in the submission form caption box. Alternatively, formats such as PDF, PNG, JPG, or ipynb can uploaded on the submission form. Please describe the submission file in the caption box.

[Detailed Rules and Guidelines](#)[Upload Your Submission](#)

DataViz Resources

If you need some inspiration or direction, try exploring some of the data viz resources below—the best way to get ideas of good methods to visualize data is to look at lots of examples.

Try some of these catalogs:

- [The Data Visualisation Catalogue](#)
- [Periodic Table of Visualization Methods](#)
- [TimeViz Browser](#)
- [A Visual Bibliography of Tree Visualization 2.0](#)

Browse these resources:

- [Flowingdata](#)
- [Infovis Wiki](#)
- [visualising data](#)
- [Gapminder](#)

Play with some web based tools:

- [Raw Graphs](#) (quick charts based on D3.js)
- [Data Wrapper](#) (simple charts designed for journalists)
- [Tableau Public](#) (free public version of a popular and powerful enterprise viz tool, see [Iron Viz competition](#) examples)
- [Vega Voyager](#) (open alternative to Tableau in development)
- [Highcharts Cloud](#) (free web editor version of popular commercial js library)

Or get to know some libraries:

- Python: [matplotlib](#) (you will probably want to use it along with [Pandas](#) to manipulate the data. Alternatively, checkout [Bokeh](#), [seaborn](#), [plotly](#), or [ggplot](#))
- R: [ggplot](#) (usually used in conjunction with other [Tidyverse](#) packages. Alternatively, checkout [ggvis](#), [Shiny](#))
- HTML+JS: [D3](#) (Alternatively, checkout [C3.js](#), [dygraphs](#), [chartjs](#), or [Vega](#))

[illegible]

Competition Scoring and Rules

The 2019 Data Science Competition is now open! (March 15th - April 15th)

Please join us for the Competition Celebration on April 30, 3:30 pm-5:00 pm in the IRIC's first floor atrium. We'll have refreshments and awards—the final scoring will be revealed and the winners will be announced!

Please note: The data visualization submissions will be graded by our judges.

Data Analysis & Machine Learning

After you submit your solution as a single CSV with *predicted* response values, those will be compared to *actual* known values.

Scores are computed as the *sum* of Root Mean Squared Errors (RMSE) for each of the 12 response variable columns in the submitted dataset.

RMSE is computed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$


We then sum across the RMSE for each column to get a single additive measure of overall prediction error.

NOTE: Multiple submissions are allowed, but only your best score will count.

The top three student scorers will be awarded prizes:

- **Golden Vandal:** University of Idaho scholarship of \$500.
- **Silver Vandal:** University of Idaho scholarship of \$300.
- **Bronze Vandal:** University of Idaho scholarship of \$200.

Contestants with scores in the top 20 will receive additional fun prizes to be awarded at the Competition Celebration on April 30th.

 Upload Your Submission

Data Visualization

For the data visualization competition, your submission will be reviewed by the judging committee. We will be looking for you to show us surprising insights in a beautiful way. Above all, we want you to tell a story about your data (use a dataset of *your* choice). The most illustrative and informative visualizations will earn the highest scores. The richer and more surprising, the better the overall score. Make the competition as fun of an experience as possible!

Winners will be determined by a panel of judges. The judges decisions are final.

Detailed Rules:

1. Only 1 submission per person.
2. Collaborative submissions are encouraged (such as a scientist collaborating with an artist).
3. If the submission is interactive, the judges will only view the first 5 minutes, or use the interactive for only 5 minutes. *No feature length documentaries, please.*
4. Submissions using media that cannot be digitally submitted are still eligible, but the piece must be portable enough that it can be brought to the judges. A small sculpture is OK. A 2 ton statue is probably not eligible.


Submission Guidelines:

1. Each visualization must be supported with the following written information:
 - A caption of no more than 200 words that describes the submission.
 - A statement of no more than 1 sentence that describes the intended audience for the visualization.
 - A statement of no more than 100 words that describes the source of the data or information that is depicted in the visualization.
2. Each visualization must be your own work.
3. Eligible media include (but are not limited to):
 - Figures (such as graphs) that represent data.
 - Infographics that represent data or a process.
 - Videos and animations (less than 5 minutes).
 - 3D physical and digital sculptures.
 - Illustrations.

The top three student scorers will be awarded prizes:

- **Golden Viz:** University of Idaho scholarship of \$500.
- **Silver Viz:** University of Idaho scholarship of \$300.
- **Bronze Viz:** University of Idaho scholarship of \$200.

Contestants with scores in the top 20 will receive additional fun prizes to be awarded at the Competition Celebration on April 30th.

 Upload Your Submission