# IST 652

# Team Member

* Thomas Bahng
* Meng Synn

# Topic

Exploration of popularity data of historic characters through Pantheon Project. The dataset is coming from <https://www.kaggle.com/mit/pantheon-project> and contains 17 columns and about 11,300 rows of data. The data appears to be structured and complete in a single comma separated values file.

|  |  |
| --- | --- |
| **Field Name** | **Sample Data** |
| article\_id | 1844 |
| full\_name | Archimedes |
| sex | Male |
| birth\_year | -287 |
| city |  |
| state | Syracuse |
| country | Italy |
| continent | Europe |
| latitude | 37.08333 |
| longitude | 15.28333 |
| occupation | Mathematician |
| industry | Math |
| domain | Science & Technology |
| article\_languages | 129 |
| page\_views | 22022077 |
| average\_views | 170714 |
| historical\_popularity\_index | 30.9907 |

# Research Questions:

* Which historic characters are the most popular?
* When did they live and where are they from?
* What factors could have generated their popularity?
* Are there any observable trends in the categorical data provided?

# Data Preparation Plan:

1. Load the data into Python
2. Explore and clean the data for anomalies, missing data, format inconsistency, etc.
3. Resolve any issues related to the previous step, including deciding how to handle missing data or lack of format uniformity
4. Filter, gather, and export data into a smaller data set that can help the analysis through smaller tasks

# Analysis and Visualization

1. Start the analysis
2. Aggregate data by country, continent, and sex, etc.
3. Visualization of results