# CALIFORNIA STATE SCIENCE FAIR
# 2013 PROJECT SUMMARY

**Name(s)**
Sam Kumar

**Project Number**
33812

**Project Title**
## True or Fake: A Stylometric Approach to Checking Originality

**Abstract**

**Objectives/Goals**
The objective of this work is to identify the true authorship of student writings using an authorial invariant, a metric based on word and punctuation usage. This project examines if three different metrics of a text - the proportion of function words, average number of function words per sentence, and punctuation frequency - could be used as viable authorial invariants.

**Methods/Materials**
A total of 56 essays, two each from 28 students, were analyzed. A computer program was developed and used to calculate the above metrics for each essay.
To test if each metric is consistent for different essays of the same student, Essay 1 of each student was compared to Essay 2 of the same student (Control Group). In all of these comparisons, the ideal metric would be the same for both essays; Success Rate 1 is the proportion of these comparisons for which each metric was not significantly different. To test if each metric is different for essays of different students, Essay 1 of each student was compared to Essay 1 of every other student (Experimental Group). In all of these comparisons, the ideal metric would be different between the two essays; Success Rate 2 is the proportion of these comparisons for which each metric was significantly different. By using 99% confidence in the statistical tests (alpha = 0.01), special care was exercised to minimize false positives.

**Results**
For the three metrics, namely the proportion of function words, average number of function words per sentence, and punctuation frequency, Success Rate 1 was 1.00, 0.96, and 0.93, and Success Rate 2 was 0.13, 0.22, and 0.29, respectively. Success Rate 1 is high for all three metrics, while Success Rate 2 is low. In an effort to achieve high values of both success rates, different combinations of the metrics were formulated and tested. This did not improve the initial outcome.

**Conclusions/Discussion**
A high Success Rate 1 indicates that the three metrics would almost never give a false positive result. However, the three metrics will only distinguish between essays written by different students 20% of the time. Nevertheless, this authorial invariant approach could be applied to situations where false positives must be expressly avoided, such as detecting ghostwriting in academics. Further research is needed to find metrics that have a high success rate for both criteria.

**Summary Statement**
This project investigated the ability of metrics to identify authorship based on word and punctuation usage in essays, and found that such methods could potentially be used to detect ghostwriting in academia.

**Help Received**
My parents looked at my poster board and presentation and gave me feedback to help me improve. My science teacher and parents looked at my experimental design and gave me suggestions for improvement.