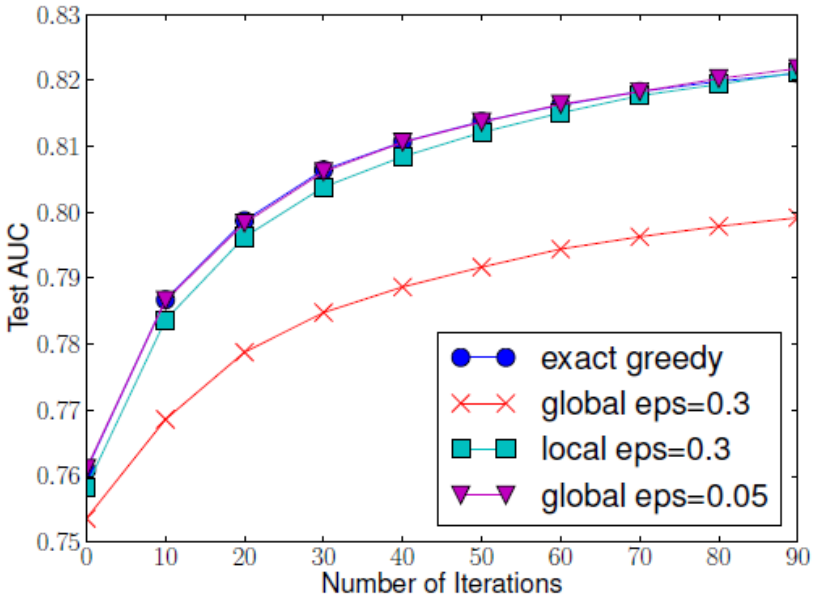

Figure 3: XGBoost: A Scalable Tree Boosting System
http://delivery.acm.org/10.1145/2940000/2939785/p785-chen.pdf?ip=47.215.149.166&id=2939785&acc=CHORUS&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1572684697_7ef0b6fff5d6284c0ea7ec3c06a68b8e

The inputs to this system is the Higgs 10M dataset available from http://opendata.cern.ch/record/328, and the goal of the system is to determine the area under the curve of exact and approximate split finding algorithms.

The Higgs boson dataset is very large and the article does not specify what column in the dataset they used. So I would design the system to be interchangeable in the column to determine the closest to the provided graph.

This project would require me to study split-finding algorithms and concepts such as AOC for the final graph. The data would need to be analyzed by dimensions to determine the proper data splits to reach the final result.

Psudocode:

1: Greedy algorithm based on algorithm 1 of page

2: Approximate algorithm based on algorithm 2 of page

3: AOC calculations

4: Plot

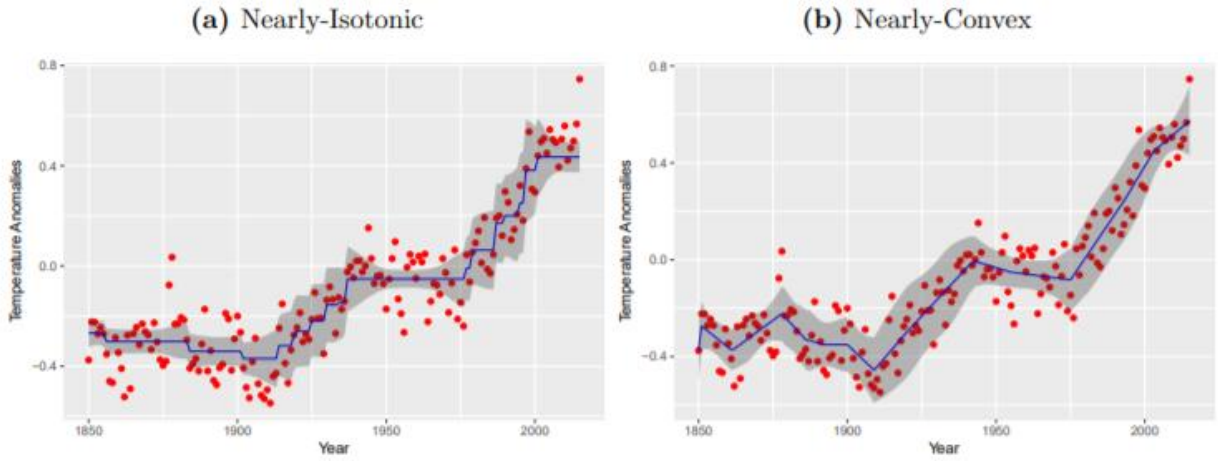

Figure 5: CVXR: An R Package for Disciplined Convex Optimization
https://web.stanford.edu/~boyd/papers/pdf/cvxr_paper.pdf

This image used temperature anomalies from multiple years to help determine an approximate change in temperature over time. The dataset is from the Carbon Dioxide Information Analysis Center (CDIAC), and can be found here: https://cdiac.ess-dive.lbl.gov/ftp/trends/temp/jonescru/global.txt. The dataset provides information based on a per month and annual temperature anomalies in columns.

This report would require me to study isotonic and convex fits to data regression. Additionally it would help in understanding the data points interaction with the lines and confidence intervals derived from them. It would also help in understanding how non-linear regressions can be derived and implemented in real world datasets. Finally, the provided information was shown coded in R, this project would help with reproducible research by having to translate code from one language to another.

I will need to study the methodology and concepts further before I can establish a firm psudocode, but here is a basic method:

1 Import data

2 average datapoints in a basic regression

3 create nearly-isotonic graph

4 create nearly-convex graph

5 create confidence intervals

6 plot everything