# STAT 430/830: Final Project

DUE: Monday August 16 by 11:59pm EST

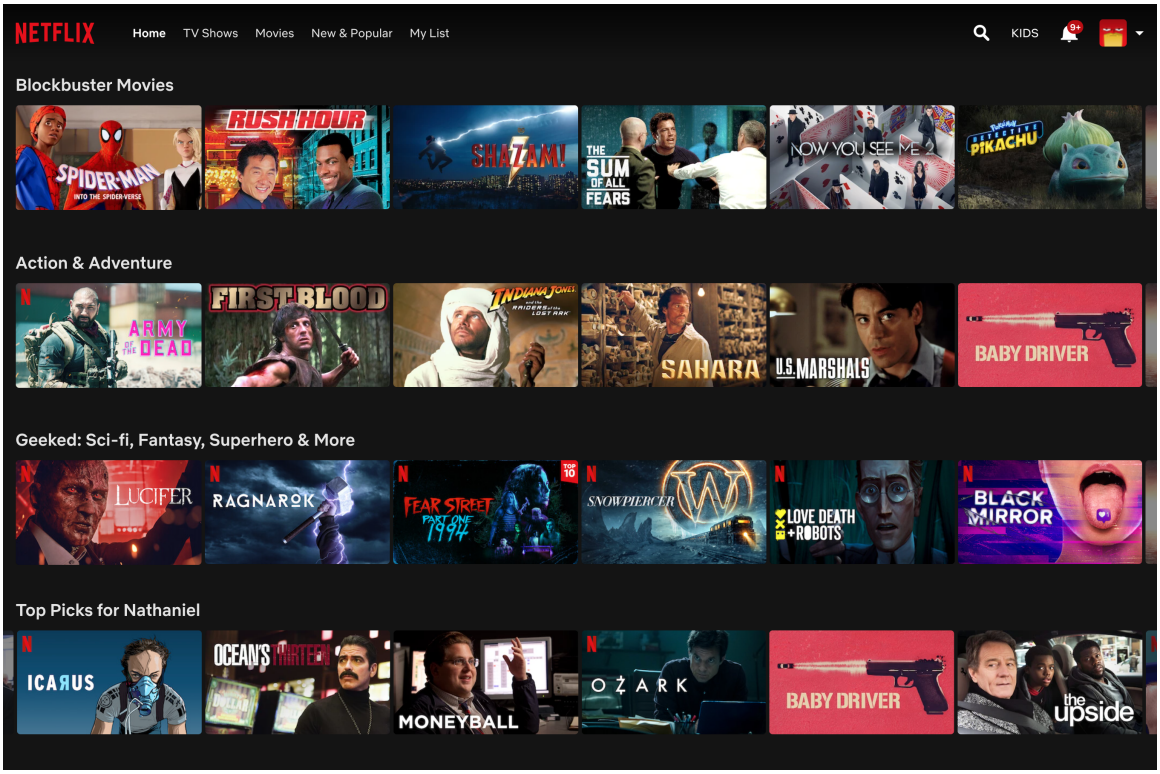
## PREAMBLE

Netflix, at one time just an online DVD rental service, has become a titan in the entertainment industry. While predominantly a streaming service, Netflix has also become well-known for its original programming such as the *Stranger Things* television series, the Oscar-nominated film *Marriage Story*, and the ludicrous documentary series *Tiger King*.

The success of Netflix is due, in part, to their well-known data-driven culture. Enmeshed within this culture is a strong appreciation for, and exploitation of, designed experiments. Netflix's home-grown ABlaze experimentation platform is well-known in the industry for its sophistication and the "wins" it has helped them achieve. It is perhaps unsurprising, then, that Netflix is a leader in online-experimentation. Though not recent, this job ad from 2016 for a Senior Data Scientist illustrates the organization's experimental maturity. In this role, you would "design, run, and analyze A/B and multivariate tests", "analyze experimental data with statistical rigor", and "adapt existing methods such as Response Surface Methodology (RSM) to online A/B testing".

In this project you will embark on a Netflix-inspired experimental journey with a hypothetical problem and a web-based response surface simulator.

## THE PROBLEM



In this project you will be concerned with optimizing the www.netflix.com homepage by way of minimizing *browsing time*. For those unfamiliar with Netflix, a screenshot of the homepage is included above. As is depicted in the screenshot, the homepage is laid out in a grid system in which movies and TV shows appear as tiles with rows differing with respect to some categorization. Though not depicted in the screenshot, when one hovers their mouse over a tile, its size is enlarged and a preview of the show/movie is automatically played in the enlarged window.

When faced with so many viewing options, Netflix users often experience choice-overload and can be overcome by a psychological phenomenon known as decision paralysis. The problem is that it becomes harder to make a decision, and it takes longer to make a decision, when faced with a large number of options to choose from. Decision paralysis negatively impacts Netflix because a user may become overwhelmed by all of the options and fatigued by the prospect of making a choice, and may ultimately lose interest and not watch anything.

To overcome this, Netflix tries to help you choose what to watch, and by a variety of mechanisms tries to help you choose quickly. Of relevance is browsing time – the length of time a user spends browsing (as opposed to watching) Netflix. Ideally, browsing time and, in particular, average browsing time would be small. In this project you will conduct a series of experiments to learn *what* influences browsing time and *how* that may be exploited in order to minimize average browsing time. There are infinitely many things that likely influence the amount of time someone spends browsing Netflix, but just four factors will be explored in this project. Each is related to the "Top Picks For..." row of the Netflix homepage. This row contains recommendations algorithmically curated for the specific user.

- **Tile Size:** The ratio of a tile's height to the overall screen height. Note the tile's aspect ratio is fixed so changing this factor changes the size of the tile, but not its shape. Smaller values correspond to a larger number of tiles visible on the screen, and larger values correspond to fewer visible tiles.

- **Match Score:** A prediction of how much you will enjoy watching the show or movie, based on your viewing history. This is recorded as a percentage, with larger values indicating a higher likelihood of enjoyment.

- **Preview Length:** The duration (in seconds) of a show or movie's preview.

- **Preview Type:** The type of preview that is autoplayed.

The table below summarizes the region of operability for each of these factors, and the default values they take on when not being experimented with.

| Factor | Code Name | Region of Operability | Default Value |
|---|---|---|---|
| Tile Size | `Tile.Size` | [0.1,0.5] | 0.2 |
| Match Score | `Match.Score` | [0,100][a] | 95 |
| Preview Length | `Prev.Length` | [30, 120][b] | 75 |
| Preview Type | `Prev.Type` | {`TT`, `AC`}[c] | `TT` |

[a] For purposes of experimentation `Match.Score` must be an integer
[b] For purposes of experimentation `Prev.Length` can only be changed in increments of 5 seconds
[c] `TT` stands for *teaser/trailer* and `AC` stands for *actual content*

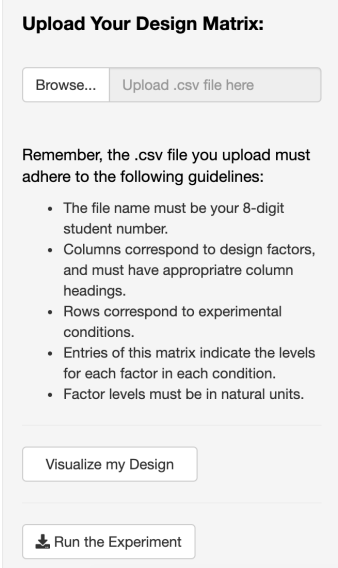
Through a series of experiments you will seek to determine which of these factors significantly influences browsing time, and you will attempt to find an optimal configuration of them that minimizes expected browsing time. You will do this by interacting with a web-based simulator, into which you will submit experimental designs and out of which you will receive response observations.

The remainder of this document provides guidelines for using the simulator, an overview of the sequential experimentation process you will undertake, and a description of the deliverable that you must submit. An outline of the marking scheme is included as an Appendix to make clear my expectations and to make transparent the manner in which you will be graded.

## THE SIMULATOR

The response surface simulator can be accessed at the following URL:

https://nathaniel-t-stevens.shinyapps.io/Netflix_Simulator_v2/




The interface (pictured above) and the manner in which you interact with it is straightforward: you upload a design matrix and then collect your results. Interaction with the simulator should include three distinct steps:

1. Upload a `.csv` file containing your design matrix. The `.csv` file **must** adhere to the following formatting guidelines:
    - The file name must be your 8-digit student number, i.e., `20208083.csv`. Any file name other than this will result in an error.
    - The columns correspond to design factors with headings `Tile.Size`, `Match.Score`, `Prev.Length`, `Prev.Type`. Any heading other than these will result in an error. The order of the headings does not matter. You do not need to experiment with every factor in every experiment, in which case not all columns (and headings) are required.
    - Each row corresponds to a distinct experimental condition, and each element indicates the level of the corresponding factor.
    - Factor levels must be in natural units.
2. Click the "Visualize my Design" button. This will render a plot of the design space and indicate the experimental conditions you plan to run.
    - If the design is not the one you intended, you may reset the simulator (by clicking the "Reset" button) and upload a different design matrix.
    - If there is anything amiss with the file you uploaded, an error (instead of a plot) will be returned.
    - **IMPORTANT:** Make sure to click the "Reset" button prior to every subsequent `.csv` upload.
3. Supposing you are happy with the design, click the "Run the Experiment" button. This will generate $n = 100$ browsing times (recorded in minutes) for each condition. The results will be automatically downloaded in a `.csv` file.
    - Remark 1: This mimics the random assignment of $n = 100$ users to each condition and the observation of their response variable.
    - Remark 2: You may assume without justification that $n = 100$ is a sufficient sample size in each condition for the task at hand.
    - Remark 3: You may assume that browsing time observations do not include the amount of time spent watching previews; browsing time records only the time spent scrolling and searching.

## THE EXPERIMENTS

Your experimental journey will consist of the phases described below. Note that STAT 430 students may ignore the `Prev.Type` factor for the entirety of this project, and they need only complete Phases I-III. The STAT 830 students, however, must consider all four factors and conduct all four phases.

### PHASE I: Factor Screening

Use a two-level experiment (i.e., $2^K$ factorial or $2^{K-p}$ fractional factorial) to determine *which* factors significantly influence the response. A factor deemed insignificant can be ignored in all subsequent phases of experimentation.

#### STAT 430 Instructions

You will experiment with three factors: `Tile.Size`, `Match.Score`, `Prev.Length`. The *low* and *high* levels of these factors (for **this** experiment) are shown below.

| Factor | Low | High |
|---|---|---|
| `Tile.Size` | 0.1 | 0.3 |
| `Match.Score` | 80 | 100 |
| `Prev.Length` | 100 | 120 |

Using the data collected from your two-level experiment, determine which factors significantly influence browsing time. Be sure to include formal hypothesis tests and main effect plots in your analysis.

#### STAT 830 Instructions

The STAT 430 Instructions apply for you, but you will *also* consider the `Prev.Type` factor whose *low* and *high* levels (for **this** experiment) are shown in the table below.

| Factor | Low | High |
|---|---|---|
| `Prev.Type` | `TT` | `AC` |

### PHASE II: Method of Steepest Descent

Considering only those factors deemed to significantly influence browsing time in PHASE I, perform a *method of steepest descent* analysis to move from the initial region of experimentation toward the vicinity of the optimum. Note that this may require intermediate two-level designs to reorient toward the optimum. You will find tests for curvature and a plot of average browsing time vs. step number useful.

**NOTE:** the initial region of experimentation is *not* in the vicinity of the optimum, and embarking down the path of steepest descent is necessary. You may use this fact without justification.

### PHASE III: Response Optimization

Once you are confident that you are in the vicinity of the optimum, conduct a central composite design and use a second order response surface model to identify the location of the optimum (i.e., the factor levels that minimize expected browsing time). Report the estimate and a 95% confidence interval for the expected browsing time at this location.

### PHASE IV: Confirmation (STAT 830 ONLY)

Read "Design and Analysis of Confirmation Experiments" by Stevens and Anderson-Cook. Choose one of the confirmation techniques discussed in the article, and use it to determine whether the optimal operating conditions identified in PHASE III achieve the good performance predicted by your response surface model.

## THE DELIVERABLE

You will prepare and submit a report via Crowdmark by the due date listed at the top of this document. The report will consist of the following components (each of which must begin on a separate page):

- Component #1: Executive Summary (1 page max)
  - Summary of the problem, your experimental journey, and the ensuing findings.
  - Be sure to state the location and value of the optimum.

non-technical

- Component #2: Introduction (2 pages max)
  - Describe *in your own words* the problem you are trying to solve
  - Describe *in your own words* the goals of response surface methodology

- Component #3: Factor Screening (2 pages max)
  - Explain your factoring screening experiment through the lens of QPDAC. State the objective, explain your design, collect the data, analyze the data, and draw a conclusion.
  - Be sure to *justify any decisions you made* in either the design or the analysis. For instance, why did you use a $2^K$ factorial experiment as opposed to a $2^{K-p}$ fractional factorial experiment (or vice versa)?
  - Be sure to include visual and/or tabular summaries of the experiment.

- Component #4: Method of Steepest Descent (2 pages max)
  - Explain your MSD experiments through the lens of QPDAC. State the objective, explain your design, collect the data, analyze the data, and draw a conclusion.
  - Be sure to *justify any decisions you made* in either the design or the analysis. For instance, how did you choose your step sizes? How did you know when to stop?
  - Be sure to include visual and/or tabular summaries of the experiment.

- Component #5: Response Optimization (2 pages max)
  - Explain your response surface experiment through the lens of QPDAC. State the objective, explain your design, collect the data, analyze the data, and draw a conclusion.
  - Be sure to *justify any decisions you made* in either the design or the analysis. For instance, how did you choose *low* and *high* levels of the factors? How did you choose where to place your axial conditions?
  - Be sure to include visual and/or tabular summaries of the experiment.

- Component #6 (STAT 830 ONLY): Confirmation (1 page max)
  - Explain your confirmation experiment through the lens of QPDAC. State the objective, explain your design, collect the data, analyze the data, and draw a conclusion.
  - Be sure to *justify any decisions you made* in either the design or the analysis. For instance, why did you choose an interval approach instead of the comparative probability metric approach (or vice versa)?

**IMPORTANT:** Your report *will not contain* R code or R output. Discussion of your analyses should be succinct, and analysis results should be included as figures and/or nicely formatted tables. Note that figures and tables count toward the page limit. Include only that which is necessary to tell your story and to justify your decisions.

## APPENDIX: Marking Scheme

STAT 430 projects will be marked out of 50 points; STAT 830 projects will be marked out of 60 points. The points are allocated as follows.

### REPORT [45 points (STAT 430); 55 points (STAT 830)]

- Executive Summary [5 points]
  - [2] Grammar, professionalism
  - [3] Clarity, relevance
- Introduction [10 points]
  - [2] Grammar, professionalism
  - [3] Clarity of problem recapitulation
  - [5] Clarity, coverage/depth, relevance of RSM discussion

Each of the following sections is worth 10 points to be allocated as indicated below.

- Factor Screening [10 points]
- Method of Steepest Descent [10 points]
- Response Optimization [10 points]
- Confirmation (STAT 830 ONLY) [10 points]
  - [2] Grammar, professionalism
  - [1] Clarity of question
  - [3] Suitability of design and clarity of design choices
  - [3] Suitability of analysis and clarity of analysis choices
  - [1] Suitability and clarity of conclusions

### ACCURACY & EFFICIENCY [5 points]

- Accuracy of Optimum [3 points]
  - [3] The optimum you've identified is **very close** to the true optimum.
  - [2] The optimum you've identified is **close** to the true optimum.
  - [1] The optimum you've identified is **somewhat close** to the true optimum.
  - [0] The optimum you've identified is **not at all close** to the true optimum.
- Efficiency of Experimentation[1] [2 points]
  - [2] If the total number of experimental conditions performed is $\leq 20$ (STAT 430), $\leq 45$ (STAT 830)
  - [1] If the total number of experimental conditions performed is $> 20$ and $\leq 30$ (STAT 430), $> 45$ and $\leq 60$ (STAT 830)
  - [0] If the total number of experimental conditions performed is $> 30$ (STAT 430), $> 60$ (STAT 830)

[1] If you want to play around with the simulation without sacrificing your condition count, feel free to play with `20203083`, but note that it has a different underlying response surface than yours. Exploring it will not provide any insight for your surface.