UNIVERSITY of WASHINGTON

# Software Design for Data Science

## Project Proposals

*DATA 515 Students*
*University of Washington*
*January 25, 2024*

# Proposals & Team Formation

During project proposals:

- Take notes on what projects sound interesting to you

After project proposals:

- You'll have time in class to talk to each other and form teams around a project proposal
- 3-4 people per team (4 is optimal)
- If a team has 1-2 people and can't find more people or a project, I will help coordinate
- If you're not comfortable with this process, let me know and I'll help you find a team

# Projects

Alumni Referral Portal
Analysing/Forecasting Stock Market Trends using Financial News Sentiment
BookMark
BOOK RECOMMENDATION TOOL
Box Office Collection Predictor
Brain Tumor - Understanding Computer Vision techniques
Celestial Tracker: Charting Meteor Landings and Comet Orbits
Classical Music Exploration
ClimbFinder: Your Next Rock Climbing Adventure
Energy Forecasting
EquiTrack: Dynamic Analysis & Sentiment Tracker
Examining Perceived Health To “Actual” Health Over Time

# Projects, Part 2

Fit Plate! Restaurant recommender based on user preferences and health goals
Flight Delay Prediction Tool
Flight-Forecast Predict Flight Delays based on Weather Conditions
Game of Thrones: Dialogue Analysis
Global Climate-Resilient Agriculture
Global Renewable Energy Consumption Analytics
The Happiness Project
Healthcare Cost Prediction
HIKING TRAILS IN NATIONAL PARKS
Housing Prices Near Transit Stops
Housing Price Predictor
Huberman Lab Lookup

# Projects, Part 3

Job application with LLM
Lie Detection Tool
Live Light & Well
Long Gone Summer
Neighborhood Scorer
NOMINATOR: OSCAR NOMINATION PREDICTOR
Nose knows/ Spirited away
NutriCart Explorer
Player Trade Value Analysis
The Polytope Permutation Puzzle
Qualitative Survey/Text Analysis with Sentiment and Contextual Similarity
RECIPE RECOMMENDER SYSTEM

# Projects, Part 4

RentWise
Seattle Area Restaurants Analysis Tool
Sentiment Analysis using LLM
The single-cell synchronizer
Ski Genius
StartUp Researcher
Summarization of Transcript tool with LLM
Travel tracker
US Presidential Election Data Explorer
Wildfire Outlooks in the U.S.
Workout Goals tracker

# ALUMNI REFERRAL PORTAL

**Shivam Agarwal**

## Project Type: Platform

## The what and the why ?

**So for the last few months I have been on the search for an internship and I believe that a lot of our seniors are in search for full time opportunities. What I have felt is that even though you apply through the normal application process, it is not very certain that you will get a shot for the interview process. But I have also found that referrals are a very good way to get your applications noticed. But getting a referral is a task within itself. And this is where my tool will come into play. This tool will have the details of the alumnus of the program (the university can be the future scope) and would allow the users i.e the present students of the program to send of an email to them seeking referrals.**

## Features:

- **A predefined mail template which can be updated by leveraging LLM's (OpenAI's API calls)**
- **Ability to track the referrals - given, pending or declined to provide**
- **A way to expand the network**

## Datasource:

The most tricky part of the project would be to secure the dataset. The current student data is readily available but the alumni database access is somewhat tricky. If we could get permissions to get access to that then that would serve as the primary database. Also we can float a general form for recruiters and other alumni to join the platform

# Analysing/Forecasting Stock Market Trends using Financial News Sentiment

**News plays a crucial role in influencing investor sentiment and stock demand. It can impact expectations and confidence. Sensationalized headlines may lead to panic selling, while optimistic news can fuel investment hype and drive buying activity, sometimes exceeding fundamental values.**

**Goals -**

- To use **NLP** and **time series modeling** to analyze the correlation between positive and negative sentiment in news and the corresponding movements in company stocks.
- Explore forecasting methods to predict potential stock changes based on sentiment data (maybe)
- Develop a user-friendly tool with filtering options for company names, date ranges, and sentiment parameters adjustment.
- Design a dashboard that visualizes trends/forecasts alongside relevant news snippets, and key market indicators for a comprehensive overview.

**Datasets -**

- **Stock prices:** Historical data from Yahoo Finance for chosen companies and date ranges.
- **News sentiment:** Financial news articles via scraping or existing APIs like News API.

**Navya Eedula**

UNIVERSITY of WASHINGTON

# BookMark

**Project Type**: Tool

**What is it?**
A personalized book recommendation tool based on the user preferences.

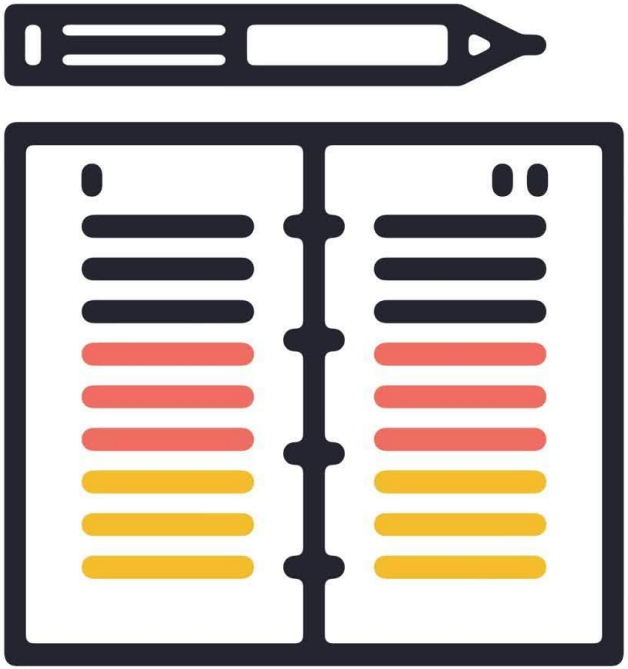
**What it does?**

- User inputs preferences such as learning goals, mood, author, subject etc.
- Analyze the books for reading history and ratings to recommend a book based on user preference.
- Provide amazon or goodreads links if book is hosted on these & provide information about free access.
- For academic preferences, provide links to available research papers along with books.

**Data Sources:**

- **Books data**: https://data.seattle.gov/Community/Library-Collection-Inventory/6vkj-f5xf/about_data
- **Amazon kindle books data**: https://www.kaggle.com/datasets/asaniczka/amazon-kindle-books-dataset-2023-130k-books
- **Goodreads data:** https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks
- **Data on academic papers**: https://www.kaggle.com/datasets/nechbamohammed/research-papers-dataset

- Manasa Shivappa Ronur



# BOOK RECOMMENDATION TOOL

Lawrie Brunswick

**Data:**

- Book Rating by ISBN:
- Filename – BX-Book-Ratings.csv
- https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset/data
- Book Details:
- Filename – BX_Books.csv
- https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset/data

*Source: Book-Crossing: User review ratings*

# Box Office Collection Predictor

**Name**: Apratim Tripathi

**Project Type**: Tool

**Objectives**:

- Develop a tool for predicting box office collections of hypothetical movies.
- Utilize two datasets: one containing data of over 5000 movies with complete cast and crew details, and another with their respective box office information.
- Train a machine learning model on these datasets to understand the correlation between cast/crew composition and box office success.
- Enable users to input a desired cast and crew for a tentative movie.
- Use the trained model to predict the potential box office collection of this hypothetical movie.

**Data Source(s):**

- https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- A database containing the box office information of the above movies

# Brain Tumor - Understanding Computer Vision techniques

Diane Chiang

**Aim**: I would like to take a deep dive into computer vision techniques such as instance and semantic segmentation to detect brain tumors from a diverse range of brain tumor images, which contributes to the early detection and diagnosis of brain tumors

- Perhaps we can build a tool where the users upload/select a similar brain tumor image and our tool will predict whether the patient is likely to have a brian tumor or not. Although that cannot serve as a ultimate truth, we can encourage patients to seek professional help.

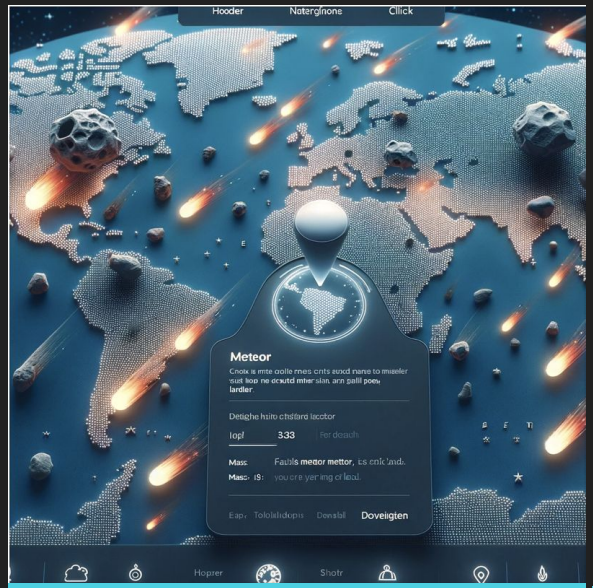
**Data Sources (Kaggle)** :

- Medical Image DataSet: Brain Tumor Detection
- Brain Tumor Image DataSet: Instance Segmentation
- Brain Tumor Image DataSet : Semantic Segmentation

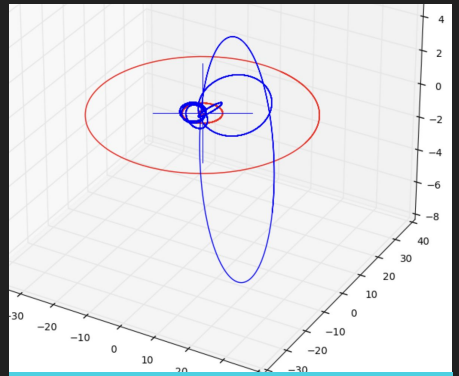
# Celestial Tracker: Charting Meteor Landings and Comet Orbits

-Ananya Bajaj (Project Type: Tool)

**Proposal**: Create an educational-interactive dashboard/website for exploring meteor landings, allowing users to view images, details, and Wikipedia links (click/web scraping) through interactive hover features. Additionally, the site will classify meteors into 9 distinct classes (classification task). A secondary page will utilize the nearest Earth comet datasets to display comet orbits, highlighting those posing a collision risk with the earth, and/or alternatively, visualizing/warning users of close-proximity objects with NASA's Near-Earth Object dataset.




**Dashboard prototype visualized**



**Comet orbits visualized**

Dataset Links:

Meteor Landings (Nasa's Open Data Portal)

Neo Earth Close Approaches (JPL Lab)

Near Earth Comets (Kaggle Dataset)

Nearest Earth Objects (Kaggle Dataset)

## Additional functionalities (optional/exploratory):

**Search Functionality**: allow users to find specific meteorites, add filters

**Timeline Slider**: Animate the history of meteor landings and comet approaches over time.

**User-Submitted Stories**: Allow users to submit personal stories or local legends about meteor sightings.

**Space Weather Alerts (more complex)**: Integrate real-time space weather alerts, informing users of meteor showers or comet sightings.

**Gamification**: Quiz about comets/meteors

# Classical Music Exploration

Elaine Zhang

- Project Output: research and reusable data (visualization)
  - Choose a route on the visualization: explore by performance history or explore by composers
    - Explore performance history:
      - Analyze which composers/works/time periods/soloist types have been the most popular/least popular in the New York Philharmonic's past performances
      - Explore how these have changed over time and during different seasons
    - Explore composers:
      - Display composer stats (popularity by NY Philharmonic plays, total works composed, plays to works ratio, etc.)
      - Are there composers/works that are played more during specific months of the year?
- Data:
  - New York Philharmonic Performance History
  - Open Opus API - Composer Metadata

# ClimbFinder: Your Next Rock Climbing Adventure

**Project Type:** Tool

## Key Features

- **Innovative App for Climbers:** Tailored for the rock climbing community.
- **Data-Driven Discovery:** Utilizes open-source data for finding ideal climbing spots routes tailored to you.
- **Customizable Searches:** Filter by grade, type, difficulty, and more.
- **Ease of Access:** Simplified process to identify your next climbing challenge.
- **Connects to Climbing Resources:** Links to comprehensive climbing guides and information.

## Data Source

- **OpenBeta - Areas & Routes:** Two datasets providing information on North American climbing areas and specific routes, including type, difficulty, and protection details, with integration capabilities via latitude and longitude coordinates.

# Energy Forecasting

As energy landscape pivot towards sustainable energy resources, managing the unpredictability of these resources becomes crucial. This project aims to combine historical weather data, and energy production records to implement time series forecasting models that supports grid stability.

Datasets:
U.S. Energy Information administration
Open EI (Open Energy Information)

Salah Makky Elbakri

# **EquiTrack**: Dynamic Analysis & Sentiment Tracker

**Project Type**: Research Tool

**Capabilities of the Tool**:

- **Custom Data Selection & Organization**: Allows users to selectively organize a wide range of stock data along with the technical indicators to suit their needs.
- **Financial Analysis Features**: Offers capabilities for both quantitative and qualitative analysis across various timeframes, facilitating deeper insights into individual stocks and aggregated groupings.
- **Sentiment Analysis Integration**: Integrates sentiment analysis derived from the latest tweets, offering a nuanced understanding of public opinion on particular stocks.

**Key Benefits**:

- **Empower User Decision-Making**: The tool's customizability in data organization and analysis allows users to make more informed and tailored investment decisions.
- **Comprehensive Market View**: Combines traditional stock metrics with sentiment analysis, providing a holistic market perspective.
- **Data Export Capability**: Enhances utility by allowing users to export their personalized data sets in multiple formats such as CSV, JSON, and PDF, catering to diverse analytical needs.

**Data Sources**:

- Stock Data: yfinance API for real-time and historical stock data, includes both quantitative and qualitative data.
- Sentiment Data: Twitter API for gathering recent tweets about specific stocks for sentiment analysis.

Yasovar Tammareddy

# Examining Perceived Health To “Actual” Health Over Time

Ted Liu

**Project Type:** Research

We all have perceptions on our health – whether we view ourselves as healthy or maybe not so healthy. How accurate are these perceptions compared to objective metrics (weight, cholesterol, etc.)? Have the accuracy of these perceptions changed over time?

I hope to investigate how accurate our perception of individual health is to objective measures over time using NHANES data sets.

**Datasets**:

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2007

# Fit Plate!

## Restaurant recommender based on user preferences and health goals

### Project output: Tool (web app)

- Input: User will provide health goals and diet preferences
- Cluster user into categories according to the goal they want to achieve
- ML model to recommend restaurants and top dishes
- The dishes recommended would be according to users health goal, eg: someone wanting to build lean muscle would be suggested a restaurant with more protein-rich foods
- Incorporate location of the user to suggest the closest and healthiest restaurant option

### Data

- Restaurant data: 93 popular chain of restaurants across USA, with ~26k records for 2022
  - https://www.menustat.org/data.html
- Nutrition data: This data contains essential nutrient information such as protein, fats, carbs, etc for base ingredients
  - https://www.kaggle.com/code/leogenzano/nutrientes-an-lisis-exploratorio-eda/input
  - https://www.kaggle.com/datasets/sonalishanbhag/dietaryhabitssurvey
  - https://www.kaggle.com/datasets/thedevastator/healthy-diet-recipes-a-comprehensive-dataset
- Web scraping for additional food ingredients if required: https://www.nutritionvalue.org/

Swarali Desai

# Flight Delay Prediction Tool

Bruno Barreto

In the U.S., flight delays are regular yet unpredictable occurrences that affect many people. However, an abundance of airline on-time data exists that could help travelers anticipate them. This project aims to create a tool that informs travelers about how likely delays for their route/carrier and how severe of a delay they should expect.

- Sources:
  - Marketing Carrier Delay Dataset
  - Reporting Carrier Delay Dataset
  - Carrier Summary Data (4 Datasets)


- Note:  The majority of the BTS' data is stored in Microsoft Access format, which is difficult to open on Mac.

# Flight-Forecast

## Predict Flight Delays based on Weather Conditions

**Project Type:** Tool
**Project Area:** Predictive Analytics

**Idea:**
One of the most common causes for flight delays is bad weather. This tool aims to leverage historical data on flight delays in conjunction with past weather data. By analyzing this existing data and integrating with weather forecast APIs, this tool calculates the probability of delays for a given date and time.

**Possible Data Sources:**

1. Flight Data: https://transtats.bts.gov/databases.asp?Z1qr_VQ=E&Z1qr_Qr5p=N8vn6v10&f7owrp6_VQF=D
2. Weather Data: https://www.wunderground.com/history

**Project proposal by:** Sushma Vankayala

# Game of Thrones: Dialogue Analysis

Raagul Nagendran

- **Project type:** Reusable data

- **Short pitch:** A breakdown of how every character contributed to the show dialogue wise, what their favourite words were (obviously we'll have to censor out a few words), and dialogue complexity of each character. Visualize the outcome using cool charts and text visualizations based on sentiment and complexity.

- **Data :**
  - https://www.kaggle.com/datasets/mylesoneill/game-of-thrones
  - https://www.kaggle.com/datasets/gunnvant/game-of-thrones-srt

# Global Climate-Resilient Agriculture

Name: Trisha Banerjee

Project Type: Research

Objectives:

1. Analyze the correlation between temperature changes and food production over time.
2. Identify regions where climate change has the most significant impact on food supply.
3. Assess the sustainability of current food production methods in the context of climate change.
4. Assess the vulnerability of different food crops and livestock to changing climate conditions.

Data Sources:

1. https://www.kaggle.com/datasets/dorbicycle/world-foodfeed-production
2. https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
3. https://www.kaggle.com/datasets/brsdincer/all-natural-disasters-19002021-eosdis

# Global Renewable Energy Consumption Analytics

Trisha Prasant

**Project Type: Analysis Tool**

As temperatures rise and storms grow more fierce, improving the efficiency and increasing the use of renewable energy sources is critical. In turn, understanding which nations are leading the way and which require more immediate improvement will help target efforts. Creating an analysis tool to help individuals realize these trends could be helpful in reaching global goals.

**Potential Questions:**

- Which types of renewables are improving the fastest?
- Which countries using which types of renewables?
- How long will it take to meet global demands and eliminate non-renewables?

**Tentative Data Sources:**

https://www.kaggle.com/datasets/jamesvandenberg/renewable-power-generation
https://www.kaggle.com/datasets/ramjasmaurya/global-powerplants


# The Happiness Project

**Project Type**: Research / Tool

**Goals:**

- Explore factors associated with "happiness"
- Interactive tool for policymakers to estimate "happiness" impact of policy tradeoffs (*e.g.* environment vs. economy)

**Data:**

- World Happiness Report
  - Survey data of life satisfaction by country and year
  - Published by Our World in Data
- World Development Indicators
  - Economic, demographic indicators by country and year
  - Published by WorldBank
- State of Global Air
  - Air pollution & its health impacts by country and year
  - Published by Our World in Data

**Proposed by:** Sue Boyd

# Healthcare Cost Prediction

Cindy Lyu

- Project Type: cerate a tool
- Goal: The project is to predict patient's healthcare costs and to identify factors contributing to this prediction.
  - Give your information (age, sex, region, smoker)
  - Predict your medical cost
  - List hospitals near you and their average cost
- Data:
  - Healthcare Insurance Expenses
    - Medical insurance costs incurred by the insured person.
    - https://www.kaggle.com/datasets/mirichoi0218/insurance
  - Hospital Charges for Inpatients
    - Hospitals charges in the US for the top 100 diagnoses.
    - https://www.kaggle.com/datasets/speedoheck/inpatient-hospital-charges

# HIKING TRAILS IN NATIONAL PARKS

Build a tool to help people visiting national parks in the US chose a hiking trail based off elevation gain, time of year / weather, trail features (no dogs, paved, …) from All trails app data. This will help people narrow down the trails and apply more filters for just national park trails and help plan.

Datasets:

Weather:
https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-year

All Trails:
https://github.com/j-ane/trail-data/blob/master/alltrails-data.csv

Ellie Holden

# Housing Prices Near Transit Stops

Project type: research & tool (web app)

- Research how proximity to transit stops affects housing prices in King County
- Build map visualization with filters to compare housing prices in relation to nearby transit stops
- Users input their preferred walking distance and the maximum affordable price
- The tool ranks and recommends top 5 houses based on user preferences

Data:

- King County House Sales
- United States House Listings
- Transit Stops for King County Metro
- Seattle Public Garages or Parking Lots

Kelly Wang

# Housing Price Predictor (Ashwin N)

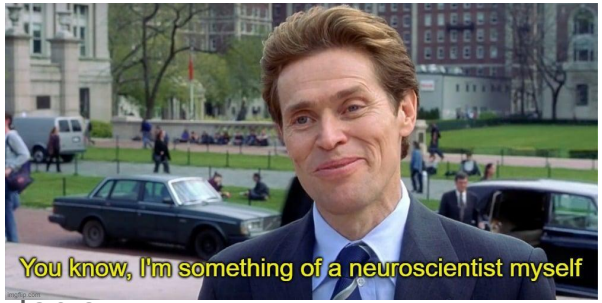
Project Type: Reusable Data/Tool

Goal: Build a regression model for house prices taking into account variables like land area, floor space, house specs (bedrooms/bathrooms), location, and more.

Data Sources:

- HOUSE PRICE PREDICTION - SEATTLE (kaggle.com) Data from Seattle in 2022
- king_country_houses_aa (kaggle.com) Data from King County in 2014-2015

# Huberman Lab Lookup




Ed Seryozhenkov

**Project type:** Tool

**Idea:** A streamlit app for searching the Huberman Lab podcast by topic.

**Example:**

- Input: [a search string containing the word “sleep”]
- Output: [a list of podcasts, timestamps, + youtube links to when sleep was discussed on the Huberman Lab]

**Data Sources:**

1. The RSS feed (contains links to .mp3 audio, metadata, tags, and timestamps)
2. Transcripts of the Podcasts (generate from audio using Whisper or Azure)

**Stretch Goal:** LLM summarization:

https://docs.streamlit.io/knowledge-base/tutorials/llm-quickstart

# Job application with LLM

Nguyen Ha

- Project output: a tool (an LLM data loader & a web app for visualization and application tracker)
  - Upload your resume & input job posts you're interested in
  - Scrape job posts to gather information
  - Load data into an LLM
  - Use LLM to analyze and match your experience with opening jobs
  - Show a chart of which job titles fit most with your experience
  - Compare how much a job requirement fits with your profile
  - Track your application status
  - Other attributes to display:
    - Does this company provide visa sponsorship?
    - Does this job require relocation?
    - Does the salary range fit your expectation?
- Data:
  - LinkedIn
    - Company information
    - Job title, description, location
  - Glassdoor
    - Job title, description, location
    - Market salary range
    - Interview questions

# Lie Detection Tool

Zongze Li

**Project Type** - Tool development

**Short Pitch** -

- Lies are an important part in our life. Sometimes we have to tell white lies to others to not hurt their feelings, sometimes we just lied for our own benefits, or we over exaggerated things when needs to do so. There are already many previous studies regarding lies and we may explore something new from there. We lies in different occurrences as well. What if we can develop a specific lie detecting tool that is used for a specific scenario? Like for a specific lying games where people are encouraged to lie? I am interested in design a unique lying game in which people need to lie to get better score and a corresponding tool for the detection using mostly facial expressions and body languages captures.

**Data** -

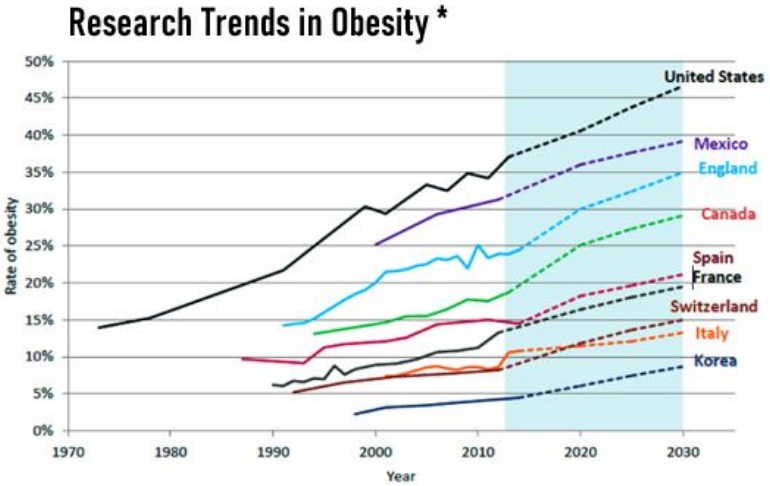
P.S. since we need to design a game to collect data, there aren't much existing ones for reference

https://www.kaggle.com/datasets/dbthapa/online-reviews-deception-detection

https://www.kaggle.com/code/ist597/ok-cupid-deception-detection

UNIVERSITY of WASHINGTON

# Live Light & Well




**Project output:**

Research & Design a tool (to make Recommendations):

- Analyze an individual's risk of obesity by considering their demographic information, dietary habits, and physical activity levels.
- Create a tool that provides personalized recommendations. Suggest practical tips and recipes tailored to people who are obese, to promote healthier lifestyle choices.

**Data:**

- Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System | Data | Centers for Disease Control and Prevention (cdc.gov)
- Nutritional value in Food ingredients: FoodData Central (usda.gov) and food-nutrition-dataset (kaggle.com)
- Ingredients to Recipes mapping: Food.com Recipes and Interactions (kaggle.com)


**Parvati Jayakumar**

* Reference: OECD projections assuming that BMI will continue to rise as a linear... | Download Scientific Diagram (researchgate.net)

# Long Gone Summer

- Project Output: Reusable Data / A Tool (Web App)
  - Fans input desired teams or stadiums and time frame
  - Tool will create potential schedules and routes
    - Tool may also provide time estimates for travel time via driving or flights
- Data
  - MLB Schedule Data
    - Available via baseball-reference.com
  - Stadium/City Location Data
    - Data from simplemaps
  - Flight Data OR Road/Driving Time Data
    - Google Maps API

# Neighborhood Scorer – Alexander Schad

Have you ever searched for an apartment in Seattle and been confused on which areas are the best?

This tool project will combine datasets such as apartment prices, walk score, available transportation, and neighborhood descriptions to suggest to the user which neighborhoods might be the best options to explore based on input from the users.

Sources:

- https://www.apartmentlist.com/rent-report/wa/seattle
- https://www.walkscore.com/score/seattle
- https://data-seattlecitygis.opendata.arcgis.com/

# NOMINATOR: OSCAR NOMINATION PREDICTOR

**Idea from:** Anurag Agarwal
**Project Type:** Tool
**What it does:** Picture this – you're at a gathering, passionately debating which movies will make it to the next Oscar Awards. The excitement is palpable, but the discussions often lead to nowhere. Fear not! With Nominator, you now have a sleek interface to elevate your movie predictions to the next level.
**How it works:** Users simply input their favorite movie, and like magic, the Nominator churns out a probability value or a normalized score predicting its chances of gracing the esteemed Oscars stage.
**Why it's cool:** Say goodbye to endless debates with your friends that lead to no conclusion. Nominator transforms your movie discussions into a thrilling experience. Whether you're rooting for the underdog or betting against the crowd, this tool empowers you to support your favorite movie or challenge the status quo.
**Datasets to be used:**

1. Previous list of Oscar movie nominations
2. Metadata related to movies

# Nose knows/ Spirited away

- Project output: A recommendation tool
  - Recommend perfumes based on user's preference for fragrance notes, intensity and silage
  - Build a basic collaborative filtering recommender / use cosine similarity to find perfumes most similar to a user's taste
  - Include an option to filter based on price range and customer reviews (if data is available)
- Data
  - Fragrantica.com
  - Web scrapping some popular perfume website (scraper code ready to use)

# *NutriCart Explorer*

UNIVERSITY of WASHINGTON

## Project Output: A Reusable Tool (Web app)

- The tool will allow users to select any recipe to know about its ingredients and nutrient constituents like percentage of carbohydrate, fat etc
- It will also create a shopping cart with all needed ingredients along with the quantity of each.
- The tool will also allow users to select multiple recipes at a time to get a consolidated cart and nutrient chart.

## Data:

- Kaggle Nutrition Data
- Epicurious

Baisakhi Sarkar

# Player Trade Value Analysis (Interactive Tool)

Amit Peled

- **Evaluate players' statistics to determine their value in trades.**
- **This could involve creating a model to rank players based on various performance metrics or give them a grade out of 100 to assess multiple player transfers.**
- **Allow users to select players being traded to analyze the strength of the trade.**

## Datasets:

2021-2022 NBA Player Stats: This dataset provides comprehensive statistics for NBA players for the 2021-2022 season. [Kaggle Link]

NBA Players Data (1950 to 2022): This extensive dataset covers NBA players' data from 1950 to 2022, offering a historical perspective for long-term trend analysis. [Kaggle Link]

# Priyam Gupta: The Polytope Permutation Puzzle

**Aim**: To solve each of the given permutation puzzles in the fewest number of moves. A permutation puzzle comprises a solution state, an initial state, and a set of allowed moves. The solution state and initial state are arrangements of symbols we call colors, while the moves for a puzzle correspond to certain permutations of these arrangements.

- What is the computational complexity of solving a permutation puzzle? Are there specific types of puzzles that are inherently more complex than others?
- Are there general principles or heuristics that apply to a wide range of such problems?
- Are there algorithms that are particularly well-suited for specific types of puzzles?

**Data Sources**:

- https://www.kaggle.com/competitions/santa-2023/data

# Qualitative Survey/Text Analysis with Sentiment and Contextual Similarity - Janice Kim


Project Type: Reusable Data & Tool

- Basic visualization and Summary of the Data
- Data Recreation Option
    - Eliminating Stop words only
    - With similar group of dictionary, filter out by a theme
    - Pulling out Sentiments, Topic Words, Transformation to Numeric Data

- Classification of the text as a topic
- Textual Feature by group (quantitative analysis)
- Sentimental Analysis
- Similar Text (ex) Find similar example essay and check the score and level)
    - Provides contextual information instead of simple keyword matching

https://www.kaggle.com/datasets/arsenycheplukov/raw-ielts-essays

https://www.kaggle.com/datasets/yassershrief/goggle-play-data?select=user_reviews.csv

# RECIPE RECOMMENDER SYSTEM

-SAIKRIPA MOHAN

Project Type: Tool

I want to build a tool that recommends recipes based on the user's nutritional requirements.

We all get bored of eating the same food week after week. This application will help users explore various recipes while helping them meet their health goals.

Input details such as High protein, High in Vitamin C, Low Cholesterol etc.

Datasets:

https://www.kaggle.com/code/ngohoantamhuy/food-recommendation-systems/input

https://www.kaggle.com/datasets/shrutisaxena/food-nutrition-dataset

# RentWise

Project Output: Research and a Tool (with visualizations)

- Use Machine Learning algorithms to predict a fair bidding price
- Develop a tool to filter and provide listings based on weather and crime rates
- Create interactive visualizations to show listings with crime statistics, walkability scores, and climate patterns for different neighborhoods

Data:

1. Zillow Housing Data (Kaggle or Zillow Website)
2. NOAA
3. Walkability Index
4. FBI Crime Data Explorer

# Seattle Area Restaurants Analysis Tool

Reusable Data Project
Jake Flynn

**Aim:** To develop an interactive map-based web app that allows users to a gain a comprehensive understanding of Seattle area restaurants by combining customer review data from multiple sources (i.e. Yelp and Google Reviews) with Food Establishment Inspection Reports from King County

Questions that the tool can enable users to answer:

- Have restaurants with high ratings also consistently scored well in health inspections?
- What factors contribute most to a successful rating both in health inspections and customer reviews?
- Are there any restaurants with recent health violations in my area?
- How do ratings from different sources compare with each other for a given restaurant?

**Data Sources:**

- King County Food Establishment Inspection Data
- Kaggle Seattle Area Restaurants (collected from Yelp.com)
- King County Foodborne Illness Outbreaks Dataset
- Food establishment closures in King County

# Sentiment Analysis using LLM

- ★ **Project type**
  - Research
- ★ **What will you do?**
  - Use LLM models like BERT /LLama-2 for identifying sentiments in product reviews
  - Compare performance of LLMs for this task vs using traditional methods
- ★ **Dataset**
  - Walmart product review dataset (https://data.world/opensnippets/walmart-products-reviews-dataset)
- ★ **Why is this cool?**
  - Get hands-on experience on GenAI techniques


Sayani Boral

# The single-cell synchronizer

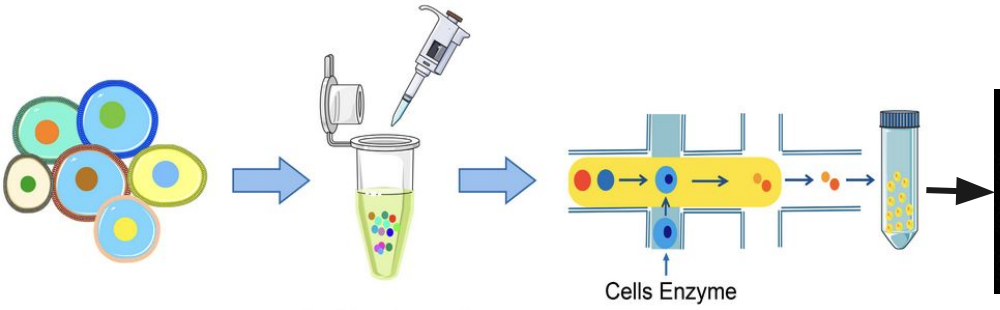
By: Elliott Sanger

Want to integrate cells from different samples?

Want to  integrate cells from different studies?

Integrate using many publicly available tools (Seurat, Harmony, LIGER, etc)

**This tool would run several integration methods and compare effectiveness!**




| | Gene1 | Gene2 | Gene3 | Gene4 | Gene5 | Gene6 | Gene7 |
|---|---|---|---|---|---|---|---|
| Cell1 | 0 | 10 | 0 | 0 | 0 | 0 | 3 |
| Cell2 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cell3 | 0 | 8 | 0 | 0 | 0 | 0 | 0 |

Datasets: https://ega-archive.org/datasets/EGAD00001006049

# Ski Genius

Kyle Sorstokke

**Type:** Tool

**Pitch:**

- Compare snow quality and amount between ski resorts (recent and historical trends)
- Factor in financial and geographic variables
- Help determine where to ski locally or where to plan bigger trips

**Data:**

Prices and Season Info

Ski-Resort Stats

# StartUp Researcher by Mark Ralston Daniel

**Project Type:** Research/Tool

**Short Pitch:** StartUpInsights is a comprehensive research project focused on gaining valuable insights into the startup ecosystem. The project aims to answer fundamental questions about the factors contributing to the success or failure of startups. A tool might be a search engine to explore the database.

**Datasets:**

- Crunchbase data on kaggle, https://www.kaggle.com/datasets/justinas/startup-investments
- More data https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

# TL;DR - Summarization of Transcript tool with LLM

Joobee Jung

- **Project Type**
  : Building a Summarization Tool for Transcript Datasets

- **Short Pitch**
  - Process input transcript data and generate concise summaries of episodes.
  - Addresses the challenge of grasping the content of lengthy transcripts, especially when overviews may not capture the actual substance.
  - Aims to save time and enhance comprehension and can be applied to various content sources, like YouTube videos through URLs.

- **Data Sources** :
  - https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset (TED Talks)
  - https://www.kaggle.com/datasets/sentinel3734/skeptoid-podcast-transcripts (Skeptoid Podcast)

- **Technology**
  : Python, LangChain, GPT-3

# Travel tracker

(Fiona) Fang Yu Lim

- Project output:  reusable data (simple web app)
  - Allows travelers to find essential information about airline on-time performance and TSA wait times at specific airports.
    - Flight density heatmaps
  - Based on past performance, predict whether the travelers' flight might be delayed.

- Data
  - Bureau of Transportation Statistics Airline On-Time Performance Data
  - TSA Wait Times

# US Presidential Election Data Explorer

Kyle Bretherton

## Project output: reusable data

- Ask a question about presidential election results
  - Ask about a specific election or about the change between two elections
- Output a list and map showing the answers to the query

## Data

- MIT Election Data and Science Lab
  - Presidential election results by county, 2000-2020
- Spatial data about the locations of the counties from Census Bureau

# Wildfire Outlooks in the U.S.

Jacob Peterson

- Type of project: Tool
- Summary: Create an interface that looks at past wildfire occurrence data and drought areas overlap in the U.S. to assess future wildfire trends. The tool would be helpful for wildfire mitigators and people assessing places to live.
- Datasets:
  - Wildfire Dataset- Kaggle
  - U.S. Drought Monitor

# Workout Goals tracker

## A new interface for your Health data

**Project Type**: Reusable Data

**Project Area**: Data visualization

**Idea**: A simple to use, dashboard with user configurable widgets that uses data pulled from apple/Samsung/google health apps to help set and visualize goals and trends based on your needs!

**Dataset**: Extract of the data from the health app. (Available as xml files)

- Abhinav Duvvuri