# Yao Yao

U.S. Citizen, Available for Relocation, Seeking Data Science Position; Data Challenge
Project Summaries: https://www.linkedin.com/in/yaoya0/
Projects: https://github.com/yaowser/

10218 Spectrum
Irvine, CA 92618
(925) 395-3640
yao.y89@gmail.com


## EDUCATION & AWARDS

Southern Methodist University, **Masters of Science in Data Science**, ML/NLP focus, GPA: 4.0 — Aug 2018

Oakland University: Completed MBA courses in business analytics, finance, statistics — May 2016

University of Michigan, **Bachelors of Science in Material Science Engineering**, GPA: 3.36 — May 2012

**Awards:** Second Place for Capstone Poster (SMU), Awarded $25,000: Michigan Clean Energy Challenge (Warmilu)

## TECHNICAL AND LANGUAGE SKILLS

**Software:** Python, R, Databricks, Multiprocessing, Azure, Computer Vision, C++, Keras, Tensorflow
**Languages:** English (first language), Mandarin Chinese (fluent), German (working proficiency)
**Certifications:** R, Python, SQL, machine learning, predictive analysis, linear algebra, Processing 3, SQL, C++, HTML, CSS, Probability, Statistics, VBA, and SAS via Udemy, Udacity, EDX, Khan Academy, and Codecademy

## EXPERIENCE

**Data Scientist** — Irvine, CA
*Kanopy* — Oct 2019 – Jun 2020

- Created a pricing model for more film selection and less licensing fees while maintaining revenue growth
- Created a hybrid recommendation engine per user based on 1) metadata NLP word similarity, 2) collaborative filtering and individual watch history, pending playlists, and film ratings, 3) location and unsupervised clustering
- Algorithmically generated curated lists of films that reflect that of university courses and commonly watched
- Generated key insights for projected film viewership trends, revenue, and what film licenses to purchase

**Data Scientist** — Indianapolis, IN
*Viral Launch* — Nov 2018 – Oct 2019

- Soloed 8 projects from data orchestration, workflow design, to production code for online software release
- Used local and Azure cloud multiprocessing to forecast time series predictions for 50+ million search terms
- Optimized key features for ad campaigns to generate best ROI for ad bid, ad budget, and sales margins
- Used feature importance to find top search terms that generated most revenue for top 20+ million products
- Applied computer vision and split testing to optimize product pictures to generate best sales conversion

**Graduate Student Researcher: 11 Major Projects + Capstone (Published)** — Dallas, TX
*Southern Methodist University* — Jan 2017 – Aug 2018

- Team leader, writer, and presenter for dataset, problem, workflow methods, interpretations, and applications
- Yelp's Review Filtering Algorithm (Thesis), ML Audio Source Separation, Taught Databricks & Blockchain
- Mastered: time series, hyperparameter optimization, classification, feature importance, validation, regression, visualization, imputation, proportional sampling, reduction, clustering, Naive Bayes, Stanford NLP, decision trees, random forest, XGBoost, SVM, neural networks, Hidden Markov, OOP, polymorphism, data scraping

**Graduate NLP and Semantics Researcher** — Chicago, IL
*Univ. of Illinois Chicago Big Data Symposium/CRIM* — Mar 2016 – Dec 2017

- Researched new product and market creation by analyzing large scale, online interactions on Reddit
- Parsed unstructured data into relational databases by applying NLP and semantic tagging via Python
- Created timeseries models to predict product prices through mentorship with Sears data analyst

**Product Technology – Management Associate** — Troy, MI
*U. S. Steel* — Sept 2012 – Dec 2014

- Automated spreadsheets and summary reports for lab data, revenue, and shipping logistics using VBA
- Created Java forms to enhance data entry at manufacturing plants to track scraps, rejects, and costs
- Developed new geometries for existing parts, accompanied by progressive stamping methods using CAD
- Investigated the formability, weight distribution, and costs of steel geometries with that of aluminum

**Co-founder, Design and Test Engineer (Simulations)** Ann Arbor, MI
*Warmilu, Warming Blanket for Premature Infants* Sept 2011 – May 2012

- Evaluated material heat transfer, heat insulation, and price optimization for blanket prototype fabrication
- Simulated thermodynamic design using 3D FEA to keep hypothermic body temperature at 37ºC for 4 hours
- Ran thermocouple testing with LabView to verify heat transfer and retention rates of working prototype
- Developed a blanket that complies with medical regulations in accordance to Mott's Children's Hospital

**Undergraduate Solar Cell Research** Ann Arbor, MI
*Polymer Solar Cell Synthesis with Professor Jinsang Kim, MSE Department* Sept 2010 – May 2012

- Maximized product yield following delicate reaction and extraction procedures
- Optimized power output and chemical synthesis difficulty for production feasibility
- Synthesized, characterized, and spin-coated organic solar cells that increased energy output from 12 to 17%

**Undergraduate eBay Research (Abstract Published)** Ann Arbor, MI
*Behavioral Game Theory on eBay with Professor Romesh Saigal, IOE department* Sept 2009 – Aug 2010

- Wrote programs in VBA and MATLAB that imported user data from eBay's API into support vectors
- Derived perceived values of all eBay items using Nash Equilibrium and Game Theory decision trees
- Formulated the seller reliability index of every eBay user based on seller attributes with machine learning

# Academic Records

Yao Yao

Search | Enroll | My Academics

## My Course History

Select Display Option
Hide courses from My Planner
Show courses from My Planner

Sort results by
Then by
Sort

Test | Taken | Transferred | In Progress

| Course | Description | Term | Grade | Units | Status |
|---|---|---|---|---|---|
| MSDS 6110 | IMMERSION | Spring 2017 | A | 1.50 | |
| MSDS 6120 | CAPSTONE 1A | Spring 2018 | A | 1.00 | |
| MSDS 6130 | Capstone 1b | Summer 2018 | A | 1.00 | |
| MSDS 6306 | INTRO TO DATA SCIENCE | Spring 2017 | A | 3.00 | |
| MSDS 6370 | STATISTICAL SAMPLING | Spring 2018 | A | 3.00 | |
| MSDS 6371 | EXPERIMENTAL STATISTICS I | Spring 2017 | A | 3.00 | |
| MSDS 6372 | STATISTICS II | Summer 2017 | A | 3.00 | |
| MSDS 6390 | VISUALIZATION OF INFORMATION | Spring 2018 | A | 3.00 | |
| MSDS 7330 | FILE ORGAN DATA BASE MAN | Summer 2017 | A | 3.00 | |
| MSDS 7331 | DATA MINING | Fall 2017 | A | 3.00 | |
| MSDS 7333 | Quantifying the World | Summer 2018 | A | 3.00 | |
| MSDS 7335 | Machine Learning | Summer 2018 | A | 3.00 | |
| MSDS 7349 | DATA AND NETWORK SECURITY | Fall 2017 | A | 3.00 | |

**Yelp's Review Filtering Algorithm (Capstone): 2nd place in best poster**
Jan 2018 – Aug 2018

- Data scraping, cluster, stratify
- Feature creation from metadata, NLP sentiment, spelling, readability, deceptive and extreme text classifiers
- Balance and scale, logistic regression, feature selection, final model
- Find features that correspond to Yelp's Algorithm and evaluate

Presentation: https://www.slideshare.net/YaoYao44/yelps-review-filtering-algorithm-powerpoint
Poster: https://www.slideshare.net/YaoYao44/yelps-review-filtering-algorithm-poster
Paper: https://scholar.smu.edu/datasciencereview/vol1/iss3/3/

**Quantifying The World: 7 Case Studies**
May 2018 – Aug 2018

Case Study 1: Data Imputation by Markov Chains in SAS
Case Study 2: ARIMA Stock Prices Time Series in Python
Case Study 3: Real Location Mapping by MAC Addresses in R

Case Study 4: Modeling Age and Runner Time distribution in R
Case Study 5: Spam Classification Optimization with Naive Bayes, Decision Trees, Random Forest, XGBoost, Linear SVM, Polynomial SVM, and Radial SVM in R
Case Study 6: Higgs Boson parameter optimization with Neural Networks in Python's Keras and Tensorflow
Case Study 7: Unlabeled Dataset Modeling and Prediction
Bonus Case Study 8: Multiclass Prediction Challenge with Random Forest

**Machine Learning Audio Separation Comparison: Clustering Repeating Period and Hidden Markov Model**
May 2018 – Jul 2018

1) Gaussian Mixture Smoothing 2) Spectra Signature Mapping 3) Clustering Repeating Period + Optimization: winner by time, results, and generality 4) Supervised Hidden Markov Model (Overlay + Sequencing)

Presentation: https://www.slideshare.net/YaoYao44/audio-separation-comparison-clustering-repeating-period-and-hidden-markov-model
Paper: https://www.slideshare.net/YaoYao44/audio-separation-comparison-clustering-repeating-period-and-hidden-markov-model-101442471

**Visualization of Information: 10 Assignments**
Jan 2018 – Apr 2018

Code: https://github.com/yaowser/viz-hw

1) Bauhaus movement recreation (1/15/18)
2) Self Portrait (1/22/18)
3) Kinetic Art (1/29/18)
4) Airline Misery Weather Dashboard (2/5/18)
5) Image Eye Spy Filter (2/12/18)
6) Twitter Net Visualization Word Cloud Winter Olympics (2/19/18)
7) OOP Design Interactive Scatterplot (3/12/18)
8) OOP Glassdoor Job Search Interactive Map with Abstract Visualizations and Pie and Bar Graph Income Kept Adjusted for Location: Taxes, Mortgage, Commute time, Standard of Living Index (3/19/18)
9) OOP Polymorphism Abstract Superclass Subclass Turret Game (4/2/18)
10) Custom 3D Music Visualizer by Decibel Amplitude and Sample Rate in Bins (Java) (4/16/18)

**Sampling: Estimating the Initial Mean Number of Views for Videos to be on Youtube's Trending List**
Mar 2018 – Mar 2018

Proportional allocation after design effect is the best method to estimate the mean views -- lowest average absolute difference from true mean and 80% that the true mean is within CI
- 'Music' and 'Comedy' may be harder categories to get on trending
- 'Nonprofits & Activism' and 'News & Politics' may be easier categories to get on trending
- Use other social platforms and increase social interactions to get on trending

Presentation: https://www.slideshare.net/YaoYao44/estimating-the-initial-mean-number-of-views-for-videos-to-be-on-youtubes-trending-list-95090196
Paper: https://www.slideshare.net/YaoYao44/estimating-the-initial-mean-number-of-views-for-videos-to-be-on-youtubes-trending-list-95090200

**Blockchain Security and Demonstration**
Sep 2017 – Dec 2017

Paper and Presentation of How Blockchain works, implementation, security risks, application survey, Quorum case study, and demonstration

Paper: https://www.slideshare.net/YaoYao44/blockchain-security-and-demonstration-86062973
Presentation: https://www.slideshare.net/YaoYao44/blockchain-security-and-demonstration

**Machine Learning algorithms on Zillow real estate data set**
Sep 2017 – Dec 2017

From the Zillow real estate data set of properties in the southern California area, conduct the following data cleaning, data analysis, predictive analysis, and machine learning algorithms:

https://github.com/yaowser/data_mining_group_project

Lab 1: Data cleaning, exploration, removal of outliers, Correlation of Continuous Variables and Log Error (Target Variable), scatterplot analysis, adding new data features, Categorical and Continuous Feature Importance

Mini-lab 1: Stochastic Gradient Descent classifier, Optimizing Logistic Regression Model Performance, Optimizing Support Vector Machine Classifier, Accuracy of results and efficiency, Logistic Regression Feature Importance, interpretation of support vectors, Density Graph

Lab 2: Classification and Regression Prediction Models, training and testing splits, optimization of K Nearest Neighbors (KD tree), optimization of Random Forest, optimization of Naive Bayes (Gaussian), advantages and model comparisons, feature importance, Feature ranking with recursive feature elimination, Two dimensional Linear Discriminant Analysis

Lab 3: Attribute Visualization, Continuous Variable Correlation Heatmap, Train and Adjust Parameters for KMeans, Spectral Clustering, and Agglomerative Clustering while evaluating metrics, deployment, Interactive Heatmap

**Prediction of Future Employee Turnover via Logistic Regression**
Jul 2017 – Aug 2017

Using logistic regression, can we use continuous and categorical data sets with interaction, selection, and validation to predict employee turnover?

Paper: https://www.slideshare.net/YaoYao44/prediction-of-future-employee-turnover-via-logistic-regression
https://github.com/yaowser/logistic-regression-employee-turnover

**Databases: Teaching Apache Spark: Demonstrations on the Databricks Cloud Platform**
May 2017 – Aug 2017

Paper and Video lecture Tutorial to Apache Spark via Databricks: Cloud Computing, Structured Streaming, Unified Analytics Integration, End-to-End Applications
https://github.com/yaowser/learn-spark/
Presentation: https://www.slideshare.net/YaoYao44/teaching-apache-spark-demonstrations-on-the-databricks-cloud-platform/YaoYao44/teaching-apache-spark-demonstrations-on-the-databricks-cloud-platform

**Data Reduction and Classification for Lumosity Data**
Jun 2017 – Jul 2017

Can the randomization grouping of participants in the original study be predicted? Utilizing cognitive ability measurements, participant activity measurements, and participants' ages, we attempt to predict randomization grouping utilizing linear discriminant analysis and principal component analysis techniques.

Paper: https://www.slideshare.net/YaoYao44/data-reduction-and-classification-for-lumosity-data
https://github.com/yaowser/LDA-PCA-Lumosity-Categorical-Prediction

**Predicting Sales Price of Homes Using Multiple Linear Regression**
Apr 2017 – Jun 2017

https://github.com/yaowser/MLR-iowa-housing

Develop a multiple regression model to predict final selling price of homes based on explanatory variables in the Ames Housing dataset. The analysis was performed on Ames housing data that were collected from the Ames Assessor's Office for individual residential properties sold in Ames, IA from 2006 to 2010.

**Distribution of Chess Wins: Expected Values from Random Moves**
Mar 2017 – Apr 2017

From 500k chess matches, is it statistically significant that White wins from random moves simply because it goes first?

Presentation: https://www.slideshare.net/YaoYao44/api-python-chess-distribution-of-chess-wins-based-on-random-moves
https://github.com/yaowser/python-chess

**Job Scraper Key Words for Data Scientist Positions**
Feb 2017 – Mar 2017

https://github.com/yaowser/MSDS-6306-job-scraper/blob/master/cybercoders.pdf