



# Tengyu Song

 st3nv

 st3nv.github.io

 ts3464@columbia.edu

 New York, NY

## Key Skills

---

### Data Science

Python, Pandas, Scikit-learn,  
OpenCV, PyTorch, TensorFlow,  
Tableau, Machine Learning

### Statistics

R, SAS, SPSS, Linear regression,  
A/B testing, Bayesian analysis,  
Causal inference

### Other

SQL, Git, HTML/CSS, JavaScript,  
MATLAB, VB.NET, Microsoft  
Office, Googling

## Work Experience

---

### Data Analyst, EGSC Columbia

09/2023 – 05/2023

- Revamped the Python data processing codebase for the annual QoL survey, boosting code's readability and maintainability, while also resolving over 20 existing bugs.
- Created a dynamic visualization pipeline by seamlessly integrating Python (pandas, matplotlib) with Tableau, producing compelling visual narratives for presentation.

### Data Engineer, JD.com

09/2021 – 11/2021

- Managed daily data extraction tasks using SQL and designed a comprehensive visualization dashboard that illuminated essential performance metrics, enhancing operational efficiency for trend analysis.
- Designed and implemented a machine learning-driven coupon grading system utilizing Machine learning models (GBDT/Random Forest) for decision making, resulting in a 10% surge in the company's 7-day Gross Merchandise Volume (GMV).
- Led the anomaly detection project on R&D process flow data. Leveraged Isolation Forest and other algorithms to achieve an 80% detection accuracy.

### Data Analyst, Yum China Holdings

04/2021 – 08/2021

- Authored and executed complex data queries using SQL with Hive and Impala, accumulating over 1000 lines of SQL code during the internship.
- Investigated user behavior hypotheses through rigorous data mining and visualization, delivering crucial insights for strategic decisions.
- Led the mini project "Cross Analysis of User Acquisition and Retention", identifying critical factors in customer engagement, improving the company's SMS marketing efficiency by 20%.

## Selected Projects / Research

---

### Dynamic Pricing for MTA Subway System, Python, Scikit-learn, Statistical Analysis

- Introduced a dynamic hybrid pricing model consisting of Random Forest and ARIMA model to solve the issue of over-crowdedness and increase fare revenue. Also addressed the issue of price discrepancy in nearby stations by implementing spectral clustering on geographical and ridership data.

### Automatic Defect Detection of PV Cell Panels, Python, OpenCV, PyTorch, Computer Vision

- Performed multi-category defect detection on photographed images of PV cell appearance using finetuned Mask-RCNN model.
- Devised an effective image cropping algorithm using OpenCV to reduce detection difficulty.

### Relationship between Human Imagination and Perception, Python, Statistics, Psychopy

- Built sophisticated data pipelines in Python to process and analyze complex eye-tracking and EEG datasets.
- Conducted in-depth data analysis using R to uncover correlations between perception and imagination patterns in different tasks, utilizing advanced statistical models and machine learning techniques.

### ChatGPT vs Human on Coding Problems, Python, PyTorch, Huggingface, LLMs

- Used API to collect ChatGPT's answers on multiple coding problem, assessing the model's accuracy on different tasks in the data science domain.

- Developed RoBERTa-based NLP model to distinguish between human-authored and GPT-generated code snippets.

**SUFE Rating Desktop Version, Database, SQL, VB.NET**

- Led a team of 15 that created the desktop version of most influential professor rating platform on campus, with up to 20,000+ users and 15,000+ highest number of visits in a single day.

**More Projects / Research**

**Bagging Enhanced Sparse Recovery Algorithms, Python, Numpy**

- Optimized signal recovery performance by applying bagging techniques to refine Orthogonal Matching Pursuit and Matching Pursuit algorithms.
- Executed rigorous experiments on simulated datasets. Evaluated performance of different bagging strategies and tuned hyperparameters to identify the optimal configuration.

**Differential Privacy under Robust M-estimators**

- Formulated a concentration bound for sensitivity curve of M-estimators based on robust statistics theory.
- Further simplified the noise tuning process for robust estimators in  $(\epsilon, \delta)$ -differential privacy framework by building connection with the smoothed local sensitivity of datasets.

**RE-TESTR Text Detection, Python, PyTorch**

- Reimplemented Text Spotting Transformers(TESTR) using PyTorch framework. Carried out ablation studies on multi-language datasets.

**Microblog Rumor Diffusion and Debunking Dynamics, Python, Beautiful Soup, Pandas**

- Carried out data collection process. Crawled over 10,000 sets of rumor-related content from Weibo using Python.
- Using SIIR model to simulate counter-propagation rumor spreading dynamics, employing Random Forest to quantify the influence of various determinants in rumor mitigation.

**Extracurricular / Teaching Experience**

**Teaching Assistant, Introduction to Statistic** 01/2023 – 05/2023

- Hosted bi-weekly office hours, providing tailored assistance for students' homework and final project, boosted students' performance in the final exam by 5%.
- Collaborated with the instructor to devise comprehensive solutions for students and detailed rubrics for grader, ensuring fast and transparent grading processes.
- Developed lab materials on R data analysis and conducted monthly lab sessions to 80 students, highly praised by students and faculty.

**Mentor Leader, Summer Training for Aspiring Researchers (STAR) Program** 05/2023 – 07/2023

- Led the mentor group to provide academic support for undergraduate mentees. Designed and delivered innovative math/statistics refresher and problem sets.
- Organized social engagement events, such as board game afternoon and chess breaks, enhancing connections between mentors and mentees.

**Selected Honors / Awards**

Chair's List of Academic Achievement	2023
JSM Award	2023
First Place of Columbia Data Science Hackathon	2023
People's Scholarship	2019

**Education**

<b>Columbia University</b> , M.A. Statistics, GPA: 4.0/4.0	09/2022 – 12/2023
<b>Shanghai University of Finance and Economics</b> , B.Sc. Statistics, GPA: 3.8/4.0	09/2022 – 12/2023