Ankai Jie

W: ankaijie.com **P**: 415-623-8076
E: ankaijie@gmail.com

EDUCATION

University of Waterloo
Sept 2015 — Present
Year 4

Bachelor of Computer Science, Major in Data Science with Business Option
Deans Honours List, 3.9 GPA
Marsh Memorial Scholarship, President's Scholarship, Research Scholarship

SKILLS

Languages
Python, Scala, Java, C++, SQL, Bash, R, Javascript

Data Analytics
Hadoop, Hive, HBase, Spark, MapReduce, Kafka, Apache Giraph, Redshift, MySql, PostgreSQL

Infrastructure
AWS — EC2, EMR, RDS, ECS, Azure, Docker, Jenkins, Airflow

Other
SBT, Node.js, JQuery, D3.js AngularJS, Git, Maven

EXPERIENCE

Sept 2018 — Present

ML Engineering
Yelp
San Francisco, CA, USA

- Create XGBoost feature with Word2Vec on review text, reducing ad click prediction (CTR) error by 5%
- Develop MapReduce job to split metrics by feature and ad category, allowing for more detailed evaluation of ad CTR model

Jan 2018 — April 2018

Software Engineering
Coursera
Mountain View, CA, USA

- Created Scala job for key Coursera partners to dynamically regrade learners for updated assessments
- Developed REST resources and database logic in Scala for group project assessment system used in master and undergraduate level degree courses

May 2017 — Aug 2017

Data Engineering
The Meet Group
San Francisco, CA, USA

- Developed MapReduce jobs to approximate billion record distinct counts using KMV sketchsets
- Engineered scalable data pipeline on Hadoop to analyze time series metrics on HBase and Redshift
- Integrated Python/Ansible to automate jobs, ETLs, and application deployment on AWS

May 2016 — Aug 2016

Data Engineering
Scotiabank
Toronto, ON, Canada

- Developed data pipeline for behavioural analysis on large scale temporal graphs with visuals in D3
- Implemented all-pairs shortest paths in Java for time series data aggregation using Apache Giraph

PROJECTS

311 Service Analysis
Python, Scikit-learn, R

- Applied k-means to geographically cluster 311 service requests and used time series forecasting
- Provided meaningful suggestions to improve 311 resolution efficiency through low-cost decisions
- 1st place winner at Citadel's Data Open Waterloo for $20 000

Citation Analyzer
Undergrad Researcher
Python, Celery, Flask, SQL

- Theory-crafted graph processing algorithms with graduate studies professor to identify influential and fraudulent researchers in Academia
- Developed pipeline in Python to ingest Scopus data and visualize million record datasets on the web

Hey Kanye!
Python, SQL Server, Azure Cognitive Services

- Rap song generator that uses Markov chains and parse trees to create/sing custom hip-hop lyrics
- Winner of Hack the North, Microsoft Azure award

AWARDS & COURSES

Awards

- 1st Place Citadel Data Open Waterloo, 2018
- UWaterloo Datafest Winner, 2018
- Hack the North Winner, 2016
- Microsoft Azure Cloud Award 2016
- Euclid Math contest Honour Roll, 2015

Coursework

- Data Structures (Enriched) and Algorithms
- Computational Statistics and Data Analysis
- Introduction to Machine Learning
  - Regression, SVMs, Neural nets, RNNs, CNNs, GBTs
- Oracle Java SE 8 Certification