

STEPHEN J.C. MUCHOVEJ

Ph.D. Astrophysics – Columbia University, 2008

B.A. Physics and Astronomy – UC Berkeley, 2002

stephen.muchovej@gmail.com

www.linkedin.com/in/muchovej

www.github.com/muchovej

917.545.2182

POSITIONS HELD

Chief Data Scientist, Quottly, Inc., October 2015 – Present

Insight Data Science Fellow, June-August 2015

Senior Scientist/National Science Foundation AAP Fellow, Caltech, 2008-2015

Graduate Researcher/National Science Foundation GSR Fellow, Columbia University, 2002-2008

EXPERIENCE

Data Science/Analytics

Methods: Machine Learning (neural networks, NLP, LDA, word2vec, regression, outlier analysis, predictive analytics, classification, decision trees, MCMC, recommender systems), Business Analytics, Signal processing, Image Processing, Mosaicking, Fourier Analysis, physics

Stack: SQL, Python (numpy, scipy, IPython, pandas, scikit-learn, nltk, gensim), R, MATLAB, d3.js

- Created course and transfer database for Quottly to help students find the right course to transfer to their university (current DB covers 4M+ classes and 5M+ course transfers across 650 universities).
- Developed ML algorithms to increase efficiency in scraping university schedules and requirements.
- Constructed a college course-similarity model (using NLP, tf-idf, and LDA) to predict course transferability (adding 2M potential course matches to the Quottly database).
- Developed outlier rejection routines to excise corrupted time-ordered data.
- Created data visualization tools to monitor database quality and health.
- Built Image Processing routines to characterize galaxy clusters by combining images (obtained in image and Fourier Space) from thousands of diverse data sets.
- Reduced and mosaicked hundreds of telescope “images” and analyzed the resulting sky maps to place constraints on cosmological parameters governing the formation of structure in the Universe.

Data Engineering/Computer Science

Methods: Data Mining, Meta-Programming, Distributed systems, Scalable Systems

Stack: Ruby, RoR, Redis, C/C++, HTML, AWS, Azure, web-scraping, github, csh, bash, Unix, Linux

- Using message queues and VMs, created and deployed a cloud-based data ingestion pipeline for collecting and updating course information and formatting it into PostGRES tables – currently used to collate 1M+ course entries at 650 universities on a 3-day cadence.
- Using meta-programming, built generalized stand-alone scraping library for web data; enabling a junior programmer to write a webpage scraper in less than 30 minutes.
- Profiled the Quottly codebase, improving query efficiency by 73%.
- Collected legislator and legislative data for the state of CA using multiple APIs and web-scraping.
- Developed pipeline to calibrate multi-terabyte time-ordered astronomical data sets.
- Architected a complete pipeline for the reduction, characterization, and visualization of interferometric data, used by several academic research groups.
- Adapted and extended real-time distributed software systems used for control, monitoring, and data acquisition for 3 telescopes, supporting various communication protocols.

Communication/Management

- Directed the distributed Quottly data science team: responsible for task allocation and code reviews.
- Regularly presented data-driven market analysis to the product and sales team at Quottly.
- Coordinated a data analysis team consisting of collaborators in the US, UK, and South Africa.
- Directed the public outreach efforts for Caltech’s Owens Valley Radio Observatory.
- Ran the Family Astro project at Columbia U, teaching astronomy to children (6-12) and their parents.