Shea Garrison-Kimmel

Data Scientist & Software Engineer

Los Angeles, CA 90025 | sheagk.github.io

610-731-6378 | sheagk@gmail.com

github | linkedin | google scholar

Experience

Foursquare (via Factual)

Los Angeles, CA

Data Scientist & Software Engineer

Aug. 2019 - present

- Practiced at developing big data jobs in the Hadoop ecosystem and the pipelines that orchestrate them
- Responsibilities include building new ETL pipelines, maintaining and upgrading legacy jobs, answering open-ended questions, applying supervised learning to big data problems, and owning several products
- Tech stack is mostly Python, Scala, & Java; Spark, MapReduce, & Hive; and Airflow & Luigi; with PostgreSQL, Docker, k8s, EMR, Grafana, & other tools mixed in as necessary
- Selected highlights include:
  - Designed, planned, and led the development of a Spark pipeline that ingests and combines more than 10 billion graph edges and over a billion nodes into a clustered graph
  - Built a Python package that interfaces with Solr, PostgreSQL, and Airflow to track quality metrics & present regressions in an easily digestible report
  - Maintained & updated an ETL pipeline that processes ~150 TB on a bi-weekly cadence, including integration efforts following the FSQ/Factual merger & migrating jobs from an on-prem cluster to AWS
  - Unified & extended a variety of internal tools to create an autonomous pipeline that measures the fill rate and accuracy of several Places attributes against external competitors
  - Developed a technique to convert a spam model's scores into "cumulative" ones (such that X% of entities with a converted score $\geq X$ are real), giving customers an easy way to select a sample of a given purity
  - Used a mix of public & paid APIs, along with human validation, to identify permanently closed Places in countries around the world, and prototyped an ML model to identify such Places automatically

Caltech & the University of California, Irvine

Pasadena, CA & Irvine, CA

Einstein Postdoctoral Fellow & Ph.D. Researcher

Aug. 2009 - July 2019

- Prolific & influential computational astrophysicist with roughly 70 publications (9 of which I led), more than 5,000 citations, and dozens of talks & presentations at conferences & universities around the world
- Built & tested galaxy formation models with parallel simulations run in super-computing environments
- Wrote scripts & libraries (primarily in Python) to create plots and visualizations from simulation outputs, then used those results to bolster and refine my scientific arguments
- Won multiple grants, prizes & CPU allocations totaling over $350,000 and 60MM hours of supercomputer time
- Accomplished teacher & advisor, with experience in leading lectures and in individual research mentoring
- Leveraged techniques such as bootstrapping, Monte Carlo sampling, and A/B testing to understand, *e.g.*, galaxy morphologies, how the disk of the Milky Way impacts dwarf galaxies, and constrain theoretical models
- Drove collaborations across Southern California by organizing several conferences on galaxy formation theory

Education

University of California, Irvine

Irvine, CA

Ph.D. in Physics & Astronomy (Thesis: *Galaxy Formation in the Local Group*)

Dec. 2010 - June 2015

Haverford College

Haverford, PA

B.S. in Astronomy and Physics with a concentration in Computer Science

Aug. 2005 - May 2009