Stephen Y. Hwang

✆ (224) 730-9044
✉ stephenyounghwang@gmail.com
gnawhnehpets.github.io
linkedin.com/in/hwangstephen

Skills

Programming	R, Python, SQL, Perl, bash, Shiny
ETL tools	dplyr, pandas, Dataflow, Dataprep, BigQuery, Redshift, BigTable, Snowflake, MongoDB, job orchestration and monitoring (Cloud Composer, MWAA), messaging (Pub/Sub, SNS, SQS), managed Hadoop/Spark services (Dataproc, EMR),
Other technologies	Git, Tableau, Jira, RMarkdown, knitr, Jupyter, ggplot, Data Studio, Confluence, REDCap Working knowledge of Docker, Kubernetes

Professional Experience

Senior Bioinformatics Scientist, RTI International, Research Triangle Park, NC — Jun 2015 - present

- Supervise multiple teams of analysts and developers as program manager and technical task lead
- Design and build extract, transform & load (ETL) pipelines on AWS & GCP to ingest and warehouse data
- Use R to implement data transformations (dplyr, tidyr) and to create novel visualizations (ggplot)
- Utilize Snowpipe to ingest Twitter JSON data into Snowflake
- Create intermediate tables from raw data to support ad-hoc analyses and reporting
- Design and create tools and pipelines to support analytics needs and facilitate analysis and insights
- Interface with technical and non-technical stakeholders to resolve complex business and technical issues

Data Engineer (Consultant), Animal Cancer DX, LLC, Raleigh, NC — Jun 2021 - present

- Designed and created an ETL pipeline to aggregate and stage experimental data on S3, ingest data into Redshift; application hosted on EC2 automatically generates and distributes reports with statistical analyses to stakeholders
- Designed and created a Shiny application to be hosted in a Docker container
- Designed and built data browser dashboard using Shiny and ggplot

Bioinformatics Data Scientist, Johns Hopkins Medical Institutions, Baltimore, MD — Aug 2013 - Jun 2015

- Completed 2-year data mining project in 10 weeks using R, SQL, and bash
- Designed and created an ETL batch process to ingest and analyze public health datasets
- Used Python to implement data transformations (pandas) and create novel visualizations (matplotlib, ggplot)
- Designed and implemented analysis pipelines for large genomic datasets for exploratory analysis and publication
- Established unsupervised machine learning methods into quality control process

Bioinformatics Analyst, Johns Hopkins Medical Institutions, Baltimore, MD — Oct 2011 - Aug 2013

- Designed and created an ETL pipeline to ingest and transform public health data sources for analytic use, increasing the efficiency by 700% over the legacy workflow

Bioinformatics Researcher, Putonti Lab, Loyola University, Chicago, IL — Jan 2010 - Jan 2011

- Created pipeline to aggregate and analyze genomic data extracted from NCBI repository

Education

Johns Hopkins University — Baltimore, MD
M.Sc. Bioinformatics — Aug 2013

Loyola University — Chicago, IL
B.Sc. Bioinformatics — May 2011

Publications

Gern, J. E., Jackson, D. J., Lemanske, R. F., Seroogy, C. M., Tachinardi, U., Craven, M., **Hwang, S. Y.**, ... Bacharier, L. B. (2019). The Children's Respiratory and Environmental Workgroup (CREW) Birth Cohort Consortium: Design, methods, and study population. *Respiratory Research, 20*(1), 115. doi:10.1186/s12931-019-1088-9

Vaz, M., **Hwang, S. Y.**, Kagiampaki, I., Phallen, J., Patil, A., O'Hagan, H. M., ... Baylin, S. B. (2017). Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. *Cancer Cell, 32*, 360–376. doi:10.1016/j.ccell.2017.08.006