Stephen Y. Hwang

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

Skills

Programming	R (dplyr, tidyr, ggplot), Python, Perl, bash, Shiny, RMarkdown, knitr, Jupyter
	Working knowledge of Pub/Sub, Dataproc, Dataflow, Cloud Composer (Airflow), Spark, Hadoop, Dialogflow, AutoML
Warehousing	SQL, BigQuery, Snowflake
	Working knowledge of Redshift, MongoDB, Datastore, Hive, BaseX
Other technologies	Google Data Studio, Tableau, Jira, Confluence, Git, Unix/Linux environments, AWS, 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 AWS S3, ingest data into RDS; 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 stage public health datasets and ingest into a SQL database
- 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 M.Sc. Bioinformatics	Baltimore, MD Aug 2013
Loyola University B.Sc. Bioinformatics	Chicago, IL May 2011