Nan Xiao

Genomic Data Scientist
Seven Bridges Genomics
Cambridge, Massachusetts

me@nanx.me
https://nanx.me
GitHub | LinkedIn

Qualifications

- Machine learning researcher with 5 years' experience and published learning methods for high-dimensional data analysis, data fusion, and translational bioinformatics.
- R developer with 8 years' of R engineering experience. Author and contributor of 20+ open source R packages and Shiny applications. Journal referee for *The R Journal*.
- Data science practitioner with experience building and leading data science teams; managing and guiding complex data analysis projects; coordinating data sharing and collaborations across engineering and product teams to serve the executive team.

Work Experience

2016 – Now Genomic Data Scientist. Seven Bridges Genomics, Inc. Cambridge, MA

Program Analyst Team Lead. Led a data science team to:

- Apply intensive quantitative and analytical skills to internal/external data to design a new pricing model that can reduce customers' AWS instance costs significantly.
- Provide direct decision support to the Chief Strategy Officer, Business Development, Product, and Marketing teams to help shape optimized data-driven company strategy.
- Deliver Shiny web applications for internal data visualization, reporting, and consulting.
- Develop the R API client package and related software for accessing, analyzing, and democratizing petabyte-scale genomic data on cloud-based Seven Bridges Platform.

R Packages Authored

My R packages for machine learning, data visualization, and dynamic reporting.

sevenbridges-r

2016 – Seven Bridges API client, CWL schema, metadata schema, and SDK helper in R.

msaenet

2016 – Multi-step adaptive elastic-net algorithm for high-dimensional feature selection.
Integrated by Max Kuhn's *caret* package for streamlined machine learning modeling.

ggsci

2016 – Scientific journal and sci-fi themed color palettes for ggplot2.
Downloaded 10k+/month. Top 2% of 11,000+ R packages on CRAN.

liftr

2015 – Containerize R Markdown documents with Docker.
DockerCon 2017 talk invited by Docker, Inc.

enpls

2014 – Ensemble partial least squares algorithm for feature screening and outlier detection.
Integrated by Max Kuhn's *caret* package for streamlined machine learning modeling.

OHPL

2017 – Ordered homogeneity pursuit lasso algorithm for group feature selection.

hdnom

2015 – Benchmarking and visualization toolkit for high-dimensional survival modeling.

protr

2012 – Efficient protein sequence feature extraction for machine learning modeling.

Rcpi

2013 – Integrative molecular feature extraction for computational drug discovery.

RECA

2012 – Relevant component analysis algorithm for supervised distance metric learning.

grex

2016 – Gene ID mapping for Genotype-Tissue Expression (GTEx) data.

R Packages Contributed

mxnet-r

2015 – Contributor of the R binding for Amazon-backed deep learning framework MXNet.

ashr

2015 – Empirical Bayes approach for large-scale hypothesis testing and FDR estimation.

dml

2015 – Distance metric learning toolkit for dimensionality reduction in computer vision.

Web Applications Authored

2013 – 6 Shiny web applications for biological, pharmaceutical, and image data analysis.

- hdnom.io (selected as Shiny User Showcase by RStudio, Inc.)
- dockflow.org | imgsvd.com | targetnet.org | protr.org | p-values.org

Publications

2013 –

- 9 journal papers on statistical machine learning methodology, high-dimensional sparse regression, feature selection, and recommender systems. [Google Scholar]
- 4 translated books on predictive modeling, data visualization, and R programming.

Education

2015 – 2016 Ph.D. Student (Human Genetics). The University of Chicago. Chicago, IL

2012 – 2018 Ph.D. Candidate (Statistics). Central South University. Changsha, China

2008 – 2012 Bachelor of Science (Statistics). Central South University. Changsha, China

Skills

Extensive research and engineering experience in: R, Shiny, Docker, statistical machine learning, large-scale linear models, recommender systems, and data visualization.

Last revision: December 2017.