Nan Xiao

Genomic Data Scientist
Seven Bridges Genomics
me@nanx.me
https://nanx.me

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 collaboration across engineering and product teams to serve the Chief Strategy Officer.

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 cost significantly.
- Provide direct decision support to the Executive Team, Business Development, Product, and Marketing team to help shape optimized data-driven company strategies.
- Deliver interactive web apps 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

sevenbridges-r

2016 – Seven Bridges API client, CWL schema, metadata schema, and SDK helper in R.
https://sbg.github.io/sevenbridges-r/ | https://bioconductor.org/packages/sevenbridges/

mxnet-r

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

msaenet

2016 – Multi-step adaptive elastic-net algorithm for high-dimensional feature selection.
https://msaenet.com | https://cran.r-project.org/package=msaenet
Integrated by the *caret* package for streamlined machine learning modeling.

ggsci

2016 – Scientific journal and sci-fi themed color palettes for ggplot2.
https://ggsci.net | https://cran.r-project.org/package=ggsci
Downloaded 12k/month. Top 2% of 11,000+ R packages on CRAN.

liftr

2015 – Containerize R Markdown documents with Docker.
https://liftr.me | https://cran.r-project.org/package=liftr
DockerCon 2017 talk invited by Docker, Inc.

OHPL

2017 – Ordered homogeneity pursuit lasso algorithm for group feature selection.
https://OHPL.io | https://cran.r-project.org/package=OHPL

hdnom

2015 – Benchmarking and visualization toolkit for high-dimensional survival modeling.
https://hdnom.org | https://cran.r-project.org/package=hdnom

enpls

2014 – Ensemble partial least squares algorithm for feature screening and outlier detection.
https://enpls.org | https://cran.r-project.org/package=enpls

protr

2012 – Efficient protein sequence feature extraction for machine learning modeling.
https://nanx.me/protr/ | https://cran.r-project.org/package=protr

Rcpi

2013 – Integrative molecular feature extraction for computational drug discovery.
https://nanx.me/Rcpi/ | https://bioconductor.org/packages/Rcpi/

RECA

2012 – Relevant component analysis algorithm for supervised distance metric learning.
https://nanx.me/RECA/ | https://cran.r-project.org/package=RECA

grex

2016 – Gene ID mapping for Genotype-Tissue Expression (GTEx) data.
https://nanx.me/grex/ | https://cran.r-project.org/package=grex

Web Applications

DockFlow

2017 – Bioconductor workflow containerization and orchestration using Docker and liftr.
https://dockflow.org

hdnom.io

2015 – Shiny app for benchmarking and visualizing high-dimensional survival models.
http://hdnom.io
Selected as RStudio Shiny User Showcase.

ImgSVD

2014 – Shiny app for image compression via singular value decomposition.
http://imgsvd.com
Joint work with *Yihui Xie*, *Yixuan Qiu*, and *Tong He*.

TargetNet

2014 – Shiny app for drug target identification by learning from binding affinities data.
http://targetnet.org

ProtrWeb

2013 – Shiny app for efficient protein sequence feature extraction.
http://protr.org

Signify

2015 Shiny app for making your (>0.05) *p*-values sound significant.
http://p-values.org

Publications

Preprints

2016 **Nan Xiao**, Q.-S. Xu, and M.-Z. Li (2016). hdnom: Building nomograms for penalized Cox models with high-dimensional survival data. *bioRxiv*. doi: 10.1101/065524.

Journal articles

2015 **Nan Xiao** and Q.-S. Xu (2015). Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. *Journal of Statistical Computation and Simulation*. doi: 10.1080/00949655.2015.1016944.

2015 **Nan Xiao**, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequence. *Bioinformatics*. doi: 10.1093/bioinformatics/btv042.

2015 D.-S. Cao*, **Nan Xiao***, Q.-S. Xu and A. F. Chen (2015). Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions, *Bioinformatics*. *Joint first authors. doi: 10.1093/bioinformatics/btu624.

2015 D.-S. Cao, **Nan Xiao**, Y.-J. Li, W.-B. Zeng, Y.-Z. Liang, A.-P. Lu, Q.-S. Xu, A. F. Chen (2015). Integrating multiple evidence sources to predict adverse drug reactions based on systems pharmacology model. *CPT: Pharmacometrics & Systems Pharmacology*. doi: 10.1002/psp4.12002.

2017 Y.-W. Lin, **Nan Xiao**, L.-L. Wang, C.-Q. Li, Q.-S. Xu (2017). Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2017.07.004.

2015 J.-B. Wang, D.-S. Cao, M.-F. Zhu, Y.-H. Yun, **Nan Xiao**, Y.-Z. Liang (2015). *In silico* evaluation of $\log D_{7.4}$ and comparison with other prediction methods. *Journal of Chemometrics*. doi: 10.1002/cem.2718.

2015 L. Shen, D.-S. Cao, Q.-S. Xu, X. Huang, **Nan Xiao**, Y.-Z. Liang (2015). A novel local manifold-ranking based k-NN for modeling the regression between bioactivity and molecular descriptors. *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2015.12.005.

Book translations

2016 Max Kuhn and Kjell Johnson (2016). *Applied Predictive Modeling*. (Hui Lin, Yi-Xuan Qiu, En-Chi Ma, **Nan Xiao**, & Vivian Zhang, Trans.). China Machine Press (Original work published in 2013). ISBN: 978-7-1115-3342-9.

2014 Winston Chang (2014). *R Graphics Cookbook*. (**Nan Xiao**, Yi-Shuo Deng, Tai-Yun Wei, & Yi-Xuan Qiu, Trans.). Posts and Telecom Press (Original work published in 2013). ISBN: 978-7-115-34227-0.

2013 Hadley Wickham (2013). *ggplot2: Elegant Graphics for Data Analysis*. (Tai-Yun Wei, Yi-Xuan Qiu, **Nan Xiao**, Tao Gao, & Wei-Cheng Zhu, Trans.). Xi'an Jiaotong University Press (Original work published in 2010). ISBN: 978-7-5605-4969-9.

2013 Robert Kabacoff (2013). *R in Action: Data Analysis and Graphics with R*. (Tao Gao, **Nan Xiao**, & Gang Chen, Trans.). Posts and Telecom Press (Original work published in 2011). ISBN: 978-7-115-29990-1.

Selected Talks

2017 *Reproducible Dynamic Report Generation with Docker and R*
Invited talk. DockerCon 2017, Austin, TX. April 2017.

2017 *Persistent Reproducible Reporting with Docker and R*
Invited talk. The 10th China R Conference, Tsinghua University, China. May 2017.

2016 *hdnom.io: High-Dimensional Survival Modeling with Shiny*
Invited talk. RStudio Shiny Developer Conference, Stanford University. January 2016.

2015 *Introduction to Reproducible Research in Bioinformatics*
Invited talk. CRI Annual Bioinformatics Workshop, Center for Research Informatics, The University of Chicago. December 2015.

Selected Posters

2017 *DockFlow: Bioconductor Workflow Containerization and Orchestration with liftr*
Nan Xiao, Tengfei Yin, and Miaozhu Li.
BioC 2017, Dana-Farber Cancer Institute, Boston, MA. July 2017.

2017 *The Deep Connection between Drugs and Side Effects*
Nan Xiao.
ISCB Art in Science Competition. ISMB/ECCB 2017, Prague, Czech Republic. July 2017.

2015 *liftr: Reproducible Bioinformatics and Statistical Data Analysis with Docker, Rabix, and knitr*
Nan Xiao, Tengfei Yin, and Miaozhu Li.
BioC 2015, Fred Hutchinson Cancer Research Center, Seattle, WA. July 2015.

Education

2015 – 2016	Ph.D. Student (Human Genetics). The University of Chicago.	Chicago, IL
2012 – Now	Ph.D. Candidate (Statistics). Central South University.	Changsha, China
2008 – 2012	Bachelor of Science (Statistics). Central South University.	Changsha, China

Journal Referee

Journal of Statistical Computation and Simulation
Chemometrics and Intelligent Laboratory Systems
Genetic Epidemiology
The R Journal

Last revision: September 2017