

< Company > | < Data Scientist >

< Location >

To whom it may concern:

Please find attached my CV in application for the < Data Scientist > role with < Company >. I am an applied statistician with a strong balance of theoretical understanding and implementation ability in

1. *TensorFlow+Keras*: model building/fitting (*tf.keras.optimizers*) and Variational Inference (VI) for
 - Generative Modeling – Variational Auto-Encoders (VAEs), Semi-Supervised Learning, and Normalizing Flows (*TensorFlow-Probability.bijectors*)
 - Epistemic and Aleatoric Uncertainty Modeling + Bayesian Deep Learning / Neural Networks
2. *PyMC4+TF-Probability*: probabilistic programming for rapid Bayesian model development with either Hamiltonian Monte Carlo or approximate (VI) posterior inference
3. *Scikit-Learn*: rapid off the shelf regression and classification (Supervised) and Unsupervised ML
 - ensemble methods (bagging, *stacking*, random forests, boosting), SVMs, Ridge/Lasso, KNNs
 - clustering, mixture, latent factor (PCA/SVD/NMF) & generative models (LDA²+Naive Bayes)
4. *statsmodels+R*: classical statistical analyses – GLMs, testing, power analysis, causal inference, etc.
5. *Bokeh*: interactive data dashboards, e.g., here's a graph of the *evolution of the current covid situation*

I additionally often work in *bash*, *github*, and *SQL* contexts and I am familiar with *C++*, *AWS*, and *Docker*.

After completing a PhD in Statistics I worked for 5 years providing data and computational workflow management as well as analysis support for research scientists in various genomics contexts. For the past 4 years I have worked in data science training and application roles. My work as a data scientist has involved evaluating medical practitioner performance and analyzing transaction data for PFM applications.

I greatly enjoy developing problem understanding and solution conceptualization, and I am highly motivated by productively contributing to my teams. I pride myself on being adaptable and responsive, exceptionally service oriented, and communicating effectively and openly in written, verbal, and visual mediums.

My wife and I are currently in Stockholm, Sweden and would love to settle down here; but, we are also open to other locations with a similar balance of striking natural beauty and metropolitan cultural richness.

Many thanks for your consideration.

Scott Schwartz, PhD



2006-2010 | Ph.D., M.S. Statistics

Duke University, NC

2001-2006 | B.S., B.A. Computer Science, Mathematics

Trinity University[†], TX

2019-2020 | Data Scientist

Tink, Stockholm, Sweden

- Developed model performance monitoring system for prioritizing data collection expenditures
- Developed exhaustive-featurization methodology to guide transaction categorization cold-start
- Developed methodology underlying subscription management and income verification products

2018-2019 | Senior Data Scientist

Covera/Spremo Health, NY

- Developed a PyMC3 implementation of a multivariate-multinomial Bayesian analysis modeling provider performance in diagnosis overcall and undercall rates for >300 pathology conditions
- Used Bayesian model comparison (marginal likelihood calculations via numerical integration) to identify and parsimoniously model associated outcomes from the >300 pathology conditions
- Determined sample size requirements for the above from currently available effect size estimates
- Developed a scikit-learn workflow to predictively model the indicated clinical impact expected for given error configurations present within the battery of diagnosis assessments modeled above

2016-2017 | Instructor, Senior Data Scientist

Galvanize, TX & NY

- Developed and taught multi-lecture Python content for >35 machine learning topics, including
 - cross-validation, confusion and cost/benefit matrices, loss functions, and regularization
 - ensemble methods, gradient boosting, SVMs, CNNs, clustering, NLP, and SVD/NMF
 - GLMs, model diagnostics, (nonparametric) hypothesis testing, and Bayesian analysis

2014-2016 | Research Associate

Integrative Biology, University of Texas, TX

- Managed lab bioinformatics pipeline development, data infrastructure, and statistical analysis
 - Research used mapping, bulk segregation and other experimental populations
 - RAD-Seq and WGS genotyping; RNA/TAG-Seq, and allele specific counting

2011-2014 | Bioinformatic Analyst

TxGen, Texas A&M AgriLife, TX

- Lead bioinformatic analysis and consulting services; managed data QC and delivery pipeline
- fastQC/X/cutadapt, BWA/bowtie/SAMtools, GATK/HTseq/TopHat, SAM/GFF/VCF/IGV

2010-2011 | Research Associate

Postdoctoral Fellowship, Texas A&M University, TX

- Provided statistical and genomic data analysis support as a member of a basic science wet lab

2007-2010 | Instructor and Consultant

Statistical Science, Duke University, NC

- Taught and supported statistics courses; consulted on experimental design and data analysis

2006-2007 | Research Assistant

Children's Environmental Health Duke University, NC

- Collaborated with diverse interdisciplinary team in an applied, translational research setting
-

-
- **2020 | Genome Biology.** *Promoter scanning during transcription initiation in *Saccharomyces cerevisiae*: Pol II in the “shooting gallery”.* Qiu C, Jin H, Vvedenskaya I, Llenas J, Zhao T, Malik I, Visbisky A, Schwartz S, Cui P, Cabart P, Han K, Lai W, Metz R, Johnson C, Sze S, Pugh B, Nickels B, Kaplan C.
 - **2019 | Patent Application**, Docket No. 60518-0011. *Computer-implemented detection and statistical analysis of errors by healthcare providers.* Elgort D, Schwartz S, Sweeney E, Dubbin G, Langseth G, Ciollaro M, Andre A.
 - **2019 | Plant Cell and Environment**, 42(7), 2165-2182. *Complex interactions between day length and diurnal patterns of gene expression drive photoperiodic responses in a perennial C4 grass.* Weng X, Lovell J, Schwartz S, Changde C, Haque T, Zhang L, Razzaque S, Juenger T.
 - **2017 | Cancer Prevention Research**, 10(10), 553-562. *Early Exposure to a High Fat/High Sugar Diet Increases the Mammary Stem Cell Compartment and Mammary Tumor Risk in Female Mice.* Lambertz I, Luo L, Berton T, Schwartz S, Hursting S, Conti C, Fuchs-Young R.
 - **2016 | Plant Physiology**, 172(2), 734-48. *Promises and challenges of eco-physiological genomics in the field: tests of drought responses in Switchgrass.* Lovell J, Shakirov E Schwartz S, Lowry D, Aspinwall A, Taylor S, Bonnette J, Hawkes C, Fay P, Juenger T.
 - **2016 | Genome Research**, 26(4), 510-18. *Drought responsive gene expression and regulatory divergence between upland and lowland ecotypes of a perennial C4 grass.* Lovell J, Schwartz S, Lowry D, Shakirov E, Wang M, Johnson J, Sreedasyam A, Plott C, Jenkins J, Schmutz J, Juenger T.
 - **2016 | BMC Genomics**, 17(202). *Colletotrichum graminicola mutant deficient in the establishment of biotrophy reveals early transcriptional events in the maize anthracnose disease interaction.* Torres M, Ghaffari N, Buiate E, Moore N, Schwartz S, Johnson C, Vaillancourt L.
 - **2014 | Nature Biotechnology**, 32, 903-14. *A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the SEQC Consortium.* Su Z, et al.
 - **2013 | Zoonoses Public Health**, 60(5), 327-35. *Identification and phylogenetic analysis of the first pandemic (H1N1) 2009 in influenza virus from feral swine.* Clavijo A, Nikooienejad A, Shahrokh M, Metz R, Schwartz S, Atashpaz-Gargariz E, Deliberto T, Lutman M, Pedersen K, Bazan L, Swenson S, Koster L, Zang M, Beckham T, Johnson C, Bonpheng M.
 - **2012 | Genome Biology**, 13(4). *A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response.* Schwartz S, Ivanov I, Davidson L, Goldsby J, Dahl D, Dougherty E, Herman D, Donavan S, Chapkin R.
 - **2012 | Statistics in Medicine**, 31(10), 949-62. *Sensitivity analysis for unmeasured confounding in principal stratification.* Schwartz S, Li F, Reiter J.
 - **2011 | Physiological Genomics**, 43(10), 640-54. *Integrated microRNA and mRNA expression profiling in a rat colon carcinogenesis model: Effect of a chemoprotective diet.* Shah M, Schwartz S, Zhao C, Davidson L, Zhou B, Lupton J, Ivanov I, Chapkin R.
 - **2011 | Journal of the American Statistical Association**, 106(496), 1331-44. *Dirichlet processes for flexible modeling of continuous intermediate variables using principal stratification.* Schwartz S, Li F, Mealli F.
 - **2010 | Statistics in Medicine**, 29(16), 1710-23. *Joint Bayesian analysis of birthweight and censored gestational age using finite mixture models.* Schwartz S, Gelfand A, Miranda M.
 - **2010 | Dissertation, Duke University**, advisors: Drs. Fan Li and Jerome P. Reiter. *Bayesian Mixture Modeling Approaches for Intermediate Variables and Causal Inference.* Schwartz S.