jacobschreiber

https://jmschrei.github.io/

whoami

jmschreiber91@gmail.com
@jmschreiber91
jmschreiber91
@jmschrei

awards

Stanford Dean's Award Fellowship
Stanford Genome Training Program Fellowship
NSF IGERT Big Data Fellowship
ACM-BCB Best Paper 2020

education

2020-Now	**Post-doctoral Researcher** Anshul Kundaje	Stanford University
2020-2020	**Post-doctoral Researcher** William Stafford Noble	University of Washington
2016-2020	**Ph.D.** Computer Science and Engineering William Stafford Noble	University of Washington
2014-2016	M.S. Computer Science and Engineering	University of Washington
2009–2013	B.S. Cum Laude Biomolecular Engineering	University of California, Santa Cruz

experience

2018- **Independent Consultant**
A machine learning consultancy that inspects data, investigates hypotheses, develops predictive models, and summarizes the results for presentations. Customers include FOXO Technologies (Brian Chen <bchen@foxotechnologies.com>), Talus Bioscience (Lindsay Pino <lpino@talus.bio>), and Basepaws (signed agreement, no billed hours yet) (Damian Kao <damian@basepaws.com>).

Editorial Roles
I am on the editorial board of the Journal of Machine Learning Research (JMLR, `https://jmlr.csail.mit.edu/`), the Journal of Open Source Software (JOSS, `https://joss.theoj.org/`), the Stanford AI Lab Blog (SAIL, `https://ai.stanford.edu/blog/`).

2017 **Core Developer, scikit-learn**
Reference: Gael Varoquaux <gael.varoquaux@inria.fr>
For a year I served as a core developer on the scikit-learn team, focusing on the tree code but also reviewing issues and PRs related to probabilistic models.

2017 **Research Intern, Autopilot Maps, Tesla**
Reference: Nathan Jones <najones@tesla.com>
This internship focused on exploring new ways that machine learning can improve Tesla AutoPilot. The projects involved processing terabytes of fleet data, doing exploratory data analysis, and building working machine learning prototypes.

2016 **Research Intern, Aspen Technology**
Reference: Mike Noskov <Mike.Noskov@aspentech.com>
This internship focused developing a machine learning implementation that could be deployed in-house to analyze internal data and make structured predictions.

2015 **Software Engineering Intern, Neurospin, INRIA**
Reference: Olivier Grisel <olivier.grisel@inria.fr>
This internship focused on speeding up the gradient boosting implementation in scikit-learn and resulting in speedups for most tree-based models.

publications (selected)

S. Whalen*, **J. Schreiber***, W.S. Noble, K. Pollard. Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics, 2021. `https://www.nature.com/articles/s41576-021-00434-9` *co-first authors

J. Schreiber*, R. Singh. Machine learning for profile prediction in genomics. Current Opinion in Chemical Biology, 2021. https://www.sciencedirect.com/science/article/pii/S1367593121000600

J. Schreiber, D. Hedge, and W.S. Noble. Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics. Proceedings of the ACM Conference on Bioinformatics, Computational Biology, and Health Informatics 2020 **(Best Paper)** https://www.biorxiv.org/content/10.1101/801183v2.full

J. Schreiber, J. Bilmes, and W.S. Noble. Priotizing transcriptomic and epigenomic experiments by using an optimization strategy that leverages imputed data. Bioinformatics, 2021 https://academic.oup.com/bioinformatics/article/37/4/439/5910545

J. Schreiber, R. Singh, J. Bilmes, and W.S. Noble. A pitfall for machine learning methods aiming to predict across cell types, bioRxiv (under review at Genome Biology), 2020 https://www.biorxiv.org/content/10.1101/512434v2

J. Schreiber, T. Durham, J. Bilmes, and W.S. Noble. Avocado: Multi-scale Deep Tensor Factorization Learns a Latent Representation of the Human Epigenome. Genome Biology, 2020 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-01977-6

M. Gasperini, A.J. Hill, J.L. McFaline-Figueroa, B. Martin, S. Kim, M.D. Zhang, D. Jackson, A. Leith, **J. Schreiber**, W.S. Noble, C. Trapnell, N. Ahituv, and J. Shendure. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell, 2019 (https://www.sciencedirect.com/science/article/pii/S009286741831554X)

J. Schreiber, Z. L. Wescoe, R. Abu-shumays, J. T. Vivian, B. Baatar, K. Karplus, and M. Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands, Proceedings of the National Academy of Science, 2013 https://www.pnas.org/content/110/47/18910.short

software

pomegranate (2,923 stars, 538 forks, used by 707 repos, >1.8M downloads as of 7/31/2022)
pomegranate is a Python package for probabilistic modeling that is a NumFOCUS Affiliated Project (https://numfocus.org/sponsored-projects/affiliated-projects). It extends scikit-learn by offering a more flexible API for building and training complex probabilistic models, such as Bayesian networks, hidden Markov models, and mixture models. Users can build models with the many pre-defined distributions or easily implement their own custom ones. https://github.com/jmschrei/pomegranate

apricot (440 stars, 41 forks, >54k downloads as of 7/31/2022)
apricot is a Python package that implements submodular optimization for the purpose of summarizing massive data sets into non-redundant subsets that still represent the space of the full data. The package follow the format of scikit-learn so that selection can be done easily and without background knowledge and dropped into existing pipelines. https://github.com/jmschrei/apricot

Avocado (90 stars, 17 forks, 33,319 downloads as of 8/3/2021)
Avocado is a Python package that implements deep tensor factorization for the purpose of modeling large, but incomplete, compendia of epigenomic data. The model both learns a low-dimensional representation that is broadly useful and can be used to impute the missing values in the tensor. https://github.com/jmschrei/avocado

scikit-learn (>46.7k stars, >21.7k forks, used by >243k repos, >434M downloads as of 5/5/20)
scikit-learn is a Python package that implements classic supervised and unsupervised machine learning algorithms as well as many components of the machine learning ecosystem, such as model evaluation, hyperparamer selection, and data preprocessing steps. I contributed for several years and was a core contributor for around a year, focusing on the tree-based methods (specifically gradient boosting) and probabilistic models. I am now an emeritus core developer because I do not regularly contribute right now. https://github.com/scikit-learn/scikit-learn

talks

Avocado Learns a Latent Representation of the Human Epigenome
UW Research Affiliates Day (2017-2019), Stanford Center for Genomics and Personalized Medicine (2018, Invited), ISMB (2018), Biological Data Science (2018), ASHG (2019), HudsonAlpha (2019)

pomegranate: probabilistic modeling in python
UW eScience (2015/6/7), PyData Chicago (2016), Moore-Sloan Data Science Summit (2016/7), Seattle DAML Meetup (2017), Data Intelligence (2017), scipy (2017), Strata Data Conference (2017), PyData NYC (2017), NYU CDS (2017), ODSC East (2017, Invited), Tesla Autopilot Maps (2017, Invited), University of California, ODSC West (2018, Invited) ODSC East (2019, Invited), ODSC West (2019, Invited)

apricot: submodular optimization for machine learning
scipy (2019), Moore-Sloan Data Science Summit (2019)