Toby Dylan Hocking — Curriculum Vitae

Contact and General Info

NAU Building 90, Office 210
1295 S. Knoles Dr.
Flagstaff, AZ 86011
E-mail: toby.hocking@nau.edu

Birth: 17 March 1984 in Newport Beach, California
Citizenship: USA.
Languages: English (native), French (fluent since 2009).
Web: http://tdhock.github.io, Erdős number = 3.

Research Interests

Machine learning algorithms, statistical software, and data visualization techniques. Emphasis on efficient algorithms for large datasets, based on constrained optimization (regression, classification, ranking, clustering, segmentation, changepoint detection, survival analysis). Application domains include medicine, genomics, neuroscience, recommendation systems, image and text analysis.

Professional Experience

Northern Arizona University, Flagstaff, Arizona, USA (2018-present).

Tenure-Track Assistant Professor, School of Informatics, Computing, and Cyber Systems.
"Optimization algorithms for machine learning and interactive data analysis."

McGill University, Montreal, Canada (2014-2018).

Postdoc with Guillaume Bourque, Department of Human Genetics.
"Changepoint detection and regression models for peak detection in genomic data."

Tokyo Institute of Technology, Tokyo, Japan (2013).

Postdoc with Masashi Sugiyama, Department of Computer Science.
"Support vector machines for ranking and comparing."

Sangamo BioSciences, Richmond, CA, USA (2006-2008).

Research Assistant with Jeff Miller in the Zinc Finger Technology group.
"A web app for visualization and statistical analysis of experimental data."

Education

École Normale Supérieure, Cachan, France (2009-2012).

Math Ph.D. with Francis Bach, Département d'Informatique; Jean-Philippe Vert, Institut Curie.
"Learning algorithms and statistical software, with applications to bioinformatics."

Université Paris 6, Paris, France (2008-2009).

Master of Statistics, internship at INRA with Mathieu Gautier and Jean-Louis Foulley.
"A Bayesian Outlier Criterion to Detect SNPs under Selection in Large Data Sets."

University of California, Berkeley, CA, USA (2002-2006).

Double B.A. in Statistics, Molecular and Cell Biology; thesis in Statistics with Terry Speed.
"Chromosomal copy number analysis using SNP microarrays and a binomial test statistic."

Honors and Awards (Selected)

Co-PI on National Science Foundation grant, $3,000,000, Sept 2021 to Aug 2026. "MIM: Discovering in reverse – using isotopic translation of omics to reveal ecological interactions in microbiomes." Role: supervise PHD student working on machine learning analysis of new metabolic flux data, in order to infer new types of interactions between microbes.

Senior personnel on NAU sub-contract of Department of Energy grant, $9,000,000, 2022-2024, Lawrence Livermore National Laboratory Science Focus Area program entitled "Microbes Persist: Systems Biology of the Soil Microbiome" led by PI Jennifer Pett-Ridge. Role: machine learning analysis of microbiome interaction networks, identifying taxa and traits that are associated with soil carbon cycling processes.

Senior personnel on Missouri Department of Elementary and Secondary Education grant, $1,509,570,

July 2021–June 2022, contract entitled "MMD-DCI Research, Development, & Leadership" led by PI Ronda Jenson. Role: summer salary and mentoring two graduate students on interpretable machine learning algorithms for Predictive Modeling Framework for District Continuous Improvement.

Air Force Research Laboratory, Summer Faculty Fellowship, May–July 2021, "Machine learning algorithms for understanding physically unclonable functions based on resistive memory devices."

PI on R Consortium Grant, $34,000, Jan–Dec 2020, "RcppDeepState: an easy way to fuzz test compiled code in R packages."

"Mobilité entrant" travel award to do research with Guillem Rigaill in Université Evry, France, 2016.

International useR conference, Best Student Poster Award, "Adding direct labels to plots," 2011.

INRIA/INRA (French computer science and agricultural research institutes), Ph.D. scholarship, 2009 (declined).

UC Berkeley, Department of Statistics VIGRE research scholarship, 2001.

Papers in progress and under review

Tao F, Huang Y, Hungate BA, Manzoni S, Frey SD, Schmidt MWI, Reichstein M, Carvalhais N, Ciais P, Jiang L, Lehmann J, Mishra U, Hugelius G, **Hocking TD**, Lu X, Shi Z, Viatkin K, Vargas R, Yigini Y, Omuto C, Malik AA, Peralta G, Cuevas-Corona R, Di Paolo LE, Luotto I, Liao C, Liang YS, Saynes VS, Huang X, Luo Y. Microbial carbon use efficiency promoting global soil carbon storage. Under review in *Nature*.

Hillman J, **Hocking TD**. Optimizing ROC Curves with a Sort-Based Surrogate Loss Function for Binary Classification and Changepoint Detection. Under review in *Journal of Machine Learning Research*.

Runge V, **Hocking TD**, Romano G, Afghah F, Fearnhead P, Rigaill G. gfpop: an R Package for Univariate Graph-Constrained Change-point Detection. Under review in *Journal of Statistical Software*, arXiv:2002.03646.

Venuto D, **Hocking TD**, Spanurattana S, Sugiyama M. Support vector comparison machines. Under review in *Machine Learning*, arXiv:1401.8008.

Barnwal A, Cho H, **Hocking TD**. Survival regression with accelerated failure time model in XGBoost. Accepted in *Journal of Computational and Graphical Statistics*.

Hocking TD, Srivastava A. Labeled Optimal Partitioning. Accepted in *Computational Statistics*.

Peer-reviewed journal papers

Hocking TD, Rigaill G, Fearnhead P, Bourque G. Generalized Functional Pruning Optimal Partitioning (GFPOP) for Constrained Changepoint Detection in Genomic Data. *Journal of Statistical Software*, Vol. 101, Issue 10 (2022).

Chaves AP, Egbert J, **Hocking TD**, Doerry E, Gerosa MA. Chatbots language design: the influence of language use on user experience. *ACM Transactions on Computer-Human Interaction* 29, 2, Article 13 (2022).

Hocking TD, Vargovich J. Linear Time Dynamic Programming for Computing Breakpoints in the Regularization Path of Models Selected From a Finite Set. *Journal of Computational and Graphical Statistics* (2021), doi:10.1080/10618600.2021.2000422.

Hocking TD. Wide-to-tall data reshaping using regular expressions and the nc package. *R Journal* (2021), doi:10.32614/RJ-2021-029.

Liehrmann A, Rigaill G, **Hocking TD**. Increased peak detection accuracy in over-dispersed ChIP-seq data with supervised segmentation models. *BMC Bioinformatics* 22, Article number: 323 (2021).

Fotoohinasab A, **Hocking TD**, Afghah F. A Greedy Graph Search Algorithm Based on Changepoint Analysis for Automatic QRS-Complex Detection. *Computers in Biology and Medicine* 130 (2021).

Abraham A, Prys-Jones T, De Cuyper A, Ridenour C, Hempson G, **Hocking TD**, Clauss M, Doughty C. Improved estimation of gut passage time considerably affects trait-based dispersal models. *Functional Ecology* (2021); 35: 860-869.

Hocking TD, Rigaill G, Fearnhead P, Bourque G. Constrained dynamic programming and supervised penalty learning algorithms for peak detection in genomic data. *Journal of Machine Learning Research* 21(87):1–40, (2020).

Hocking TD. Comparing namedCapture with other R packages for regular expressions. *R Journal* (2019). doi:10.32614/RJ-2019-050

Jewell S, **Hocking TD**, Fearnhead P, Witten D. Fast Nonconvex Deconvolution of Calcium Imaging Data. *Biostatistics* (2019), doi: 10.1093/biostatistics/kxy083.

Depuydt P, Koster J, Boeva V, **Hocking TD**, Speleman F, Schleiermacher G, De Preter K. Meta-mining of copy number profiles of high-risk neuroblastoma tumors. *Scientific Data* (2018).

Alirezaie N, Kernohan KD, Hartley T, Majewski J, **Hocking TD**. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. American Journal of Human Genetics (2018). doi:10.1016/j.ajhg.2018.08.005

Sievert C, Cai J, VanderPlas S, Khan F, Ferris K, **Hocking TD**. Extending ggplot2 for linked and dynamic web graphics. *Journal of Computational and Graphical Statistics* (2018).

Depuydt P, Boeva V, **Hocking TD**, *et al.* Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High-risk Neuroblastoma Patients. *Journal of the National Cancer Institute* (2018). DOI:10.1093/jnci/djy022.

Hocking TD, Goerner-Potvin P, Morin A, Shao X, Pastinen T, Bourque G. Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning. *Bioinformatics* (2017) 33 (4): 491-499.

Shimada K, Shimada S, Sugimoto K, Nakatochi M, Suguro M, Hirakawa A, **Hocking TD**, Takeuchi I, Tokunaga T, Takagi Y, Sakamoto A, Aoki T, Naoe T, Nakamura S, Hayakawa F, Seto M, Tomita A, Kiyoi H. Development and analysis of patient-derived xenograft mouse models in intravascular large B-cell lymphoma. *Leukemia* (2016).

Chicard M, Boyault S, Colmet-Daage L, Richer W, Gentien D, Pierron G, Lapouble E, Bellini A, Clement N, Iacono I, Bréjon S, Carrere M, Reyes C, **Hocking TD**, Bernard V, Peuchmaur M, Corradini N, Faure-Conter C, Coze C, Plantaz D, Defachelles A-S, Thebaud E, Gambart M, Millot F, Valteau-Couanet D, Michon J, Puisieux A, Delattre O, Combaret V, Schleiermacher G. Genomic copy number profiling using circulating free tumor DNA highlights heterogeneity in neuroblastoma. *Clinical Cancer Research* (2016).

Maidstone R, **Hocking TD**, Rigaill G, Fearnhead P. On optimal multiple changepoint algorithms for large data. *Statistics and Computing* (2016). doi:10.1007/s11222-016-9636-3

Suguro M, Yoshida N, Umino A, Kato H, Tagawa H, Nakagawa M, Fukuhara N, Karnan S, Takeuchi I, **Hocking TD**, Arita K, Karube K, Tsuzuki S, Nakamura S, Kinoshita T, Seto M. Clonal heterogeneity of lymphoid malignancies correlates with poor prognosis. *Cancer Sci.* (2014) Jul;105(7):897-

904.

Hocking TD, Boeva V, Rigaill G, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Richer W, Bourdeaut F, Suguro M, Seto M, Bach F, Vert J-P. SegAnnDB: interactive Web-based genomic segmentation. *Bioinformatics* (2014) 30 (11): 1539-1546. DOI:10.1093/bioinformatics/btu072

Hocking TD, Wutzler T, Ponting K and Grosjean P. Sustainable, extensible documentation generation using inlinedocs. *Journal of Statistical Software* (2013), 54(6), 1-20. DOI:10.18637/jss.v054.i06

Hocking TD, Schleiermacher G, Janoueix-Lerosey I, Boeva V, Cappo J, Delattre O, Bach F, Vert J-P. Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinfo.* (2013), 14:164. DOI:10.1186/1471-2105-14-164

Gautier M, **Hocking TD**, Foulley JL. A Bayesian outlier criterion to detect SNPs under selection in large data sets. *PloS ONE* 5 (8), e11913 (2010).

Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, **Hocking TD**, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature biotechnology* 26 (6), 702-70 (2008).

Peer-reviewed conference papers

In addition to peer-reviewed journals, I publish papers at highly competitive computer science conferences like *ICML* and *NeurIPS*, with double-blind peer reviews, and ≈20% acceptance rates.

Kolla AC, Groce A, **Hocking TD**. Fuzz Testing the Compiled Code in R Packages. IEEE 32nd International Symposium on Software Reliability Engineering (ISSRE 2021), pp. 300-308, doi: 10.1109/ISSRE52982.2021.00040.

Fotoohinasab A, **Hocking TD**, Afghah F. A Graph-Constrained Changepoint Learning Approach for Automatic QRS-Complex Detection. *Asilomar Conference on Signals, Systems, and Computers* (2020).

Fotoohinasab A, **Hocking TD**, Afghah F. A Graph-constrained Changepoint Detection Approach for ECG Segmentation. *42th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2020).*

Hocking TD, Bourque G. Machine Learning Algorithms for Simultaneous Supervised Detection of Peaks in Multiple Samples and Cell Types. *Pacific Symposium on Biocomputing* 25:367-378 (2020).

Drouin A, **Hocking TD**, Laviolette F. Maximum margin interval trees. *Neural Information Processing Systems (NeurIPS)*, 2017.

Hocking TD, Rigaill G, Bourque G. PeakSeg: constrained optimal segmentation and supervised penalty learning for peak detection in count data. *International Conference on Machine Learning (ICML)*, 2015.

Hocking TD, Rigaill G, Bach F, Vert J-P. Learning sparse penalties for change-point detection using max-margin interval regression. *International Conference on Machine Learning (ICML)*, 2013.

Hocking TD, Joulin A, Bach F, Vert J-P. Clusterpath: an Algorithm for Clustering using Convex Fusion Penalties. *International Conference on Machine Learning (ICML)*, 2011.

Books, Chapters, Manuals

Hocking TD and Killick R. *Changepoint detection algorithms and applications in R.* Textbook in preparation.

Hocking TD. Introduction to Machine Learning and Neural Networks. Chapter in textbook *Land Carbon Cycle Modeling: Matrix Approach, Data Assimilation, and Ecological Forecasting* edited by

Yiqi Luo. (expected publication Jan 2022)

Hocking TD. Animated interactive data visualization using the grammar of graphics (The animint2 Manual), 17 web pages/chapters with interactive graphics and exercises. (2018)

Conference Tutorials

Hocking TD, Killick R. Introduction to optimal changepoint detection algorithms, *useR* 2017.

Hocking TD, Ekstrøm CT. Understanding and creating interactive graphics, *useR* 2016.

Invited talks (selected)

Keynote for IEEE conference in Prescott Arizona, University of Arizona Math/Stats seminar (2022); ASU West ML Day, TRIPODS University of Arizona, IEEE NJACS (2021); NAU Math Department Colloquium (2018); University of Waterloo, Université de Montréal, Sainte-Justine Children's Hospital, University of Québec à Montréal, Polytechnique Montréal (2017); Université Laval Centre for Big Data Research (2016); McGill Barbados epigenomics workshop (2015); Sapporo Japan Workshop on Machine Learning and Applications to Biology (2013); Google Research New York, Université Rennes, Université Angers, INRIA Lille (2012); Institut de Biologie de Lille (2011).

Consulting (selected)

Acronis SCS, cybersecurity company in Phoenix (2022). Interpretable and non-linear machine learning algorithms which use source code analysis to predict software vulnerabilities.

Plotly, data visualization startup in Montreal (2015). Original author of ggplot functionality in plotly R package.

Teaching

All of my course materials are freely available online, https://tdhock.github.io/teaching/

Spring 2022, Northern Arizona University, CS570, Deep Learning.

Fall 2021, Northern Arizona University, CS499/599, Unsupervised Learning.

Summer 2021, 60 minute lecture "Introduction to Machine Learning and Neural Networks" for summer school on "New Advances in Land Carbon Cycle Modeling."

Spring 2021, Northern Arizona University, CS470, Artificial Intelligence.

Fall 2020, Northern Arizona University, CS499/599, Unsupervised Learning.

Summer 2020, 90 minute lecture "Introduction to Machine Learning and Neural Networks" for summer school on "New Advances in Land Carbon Cycle Modeling."

Spring 2020, Northern Arizona University, CS499, Deep Learning.

Fall 2019, Northern Arizona University, CS/EE599, Reproducible Machine Learning Research.

Spring 2019, Northern Arizona University, CS499, Optimization algorithms for machine learning.

Professional Service

2022, Member of steering committee for R project in Google Season of Docs.

2021–present, machine learning editor for rOpenSci Statistical Software.

2018–present, editor for Journal of Statistical Software.

2021–present, co-author of R Development Guide and member of R Contribution Working Group (resources for making it easy/accessible to contribute improvements to base R).

2012–present, co-administrator and mentor for R project in Google Summer of Code (teaching how to create and improve R packages).

2017–2018, president of organizing committee for "R in Montreal 2018" conference.

2010–present: peer reviewer for Technometrics, International Conference on Machine Learning (ICML), Advances in Neural Information Processing Systems (NeurIPS), Journal of Machine Learning Research (JMLR), Artificial Intelligence Review, Journal of Computational and Graphical Statistics (JCGS), R Journal, Bioinformatics, PLOS Computational Biology, BMC Bioinformatics, IEEE Transactions on Pattern Analysis and Machine Intelligence, Information and Inference, Journal of Statistical Computation and Simulation, Computo, Genome Biology.

Software Online (Selected)

Numerous free/open-source software contributions using R, C, C++, Python, and JavaScript.

R: contributions to base R regex functionalty and data reshaping in data.table package. Maintainer of numerous R packages (17 on CRAN as of Feb 2022) for machine learning (changepoint detection, classification, regression, ranking, etc), directlabels for labeled figures, animint2 for animated interactive figures, inlinedocs for documentation generation.

Python: contributions to pandas module for data manipulation (str.extractall regex functionality), maintainer of GUI/web server software for labeling and changepoint detection in genomic data (annotate_regions, SegAnnDB, PeakLearner).

References

Yiqi Luo, collaborator on research and teaching (summer school, textbook).
Regents' Professor, Department of Biological Sciences, Northern Arizona University
Phone: +1 (928) 523-1925, E-mail: yiqi.luo@nau.edu
Web: https://www2.nau.edu/luo-lab/?member_info&id=52

Eck Doerry, collaborator on research and familiar with my teaching.
Professor, School of Informatics, Computing and Cyber Systems, Northern Arizona University
Phone: +1 (928) 523-9377, E-mail: Eck.Doerry@nau.edu, Web: https://www.cefns.nau.edu/~edo

Jarrett "Jay" Barber, colleague familiar with my teaching.
Associate Professor, School of Informatics, Computing and Cyber Systems,
Northern Arizona University, Phone: +1 (928) 523-6869, E-mail: Jarrett.Barber@nau.edu

Alex Groce, collaborator on research papers and grants.
Associate Professor, School of Informatics, Computing and Cyber Systems,
Northern Arizona University, Phone: +1 (928) 523-8263, E-mail: Alex.Groce@nau.edu

Fatemeh Afghah, collaborator on research papers and grants.
Associate Professor, Clemson University, Phone: +1 (864) 656-0100, E-mail: fafghah@clemson.edu,
Web: https://fafghah.people.clemson.edu/

Guillem Rigaill, collaborator on research papers and grants.
Researcher at INRAE (French Agronomy Research Institute)
Phone: +33 (0) 1 64 85 35 44, E-mail: guillem.rigaill@inrae.fr
Web: http://www.math-evry.cnrs.fr/members/Grigaill/welcome