

# **Introduction to Applied Science**

Rex W. Douglass

11/15/22

# Table of contents

<b>Preface</b>	<b>10</b>
<b>Overviews</b>	<b>11</b>
References . . . . .	12
 <b>I   Introduction</b>	 <b>13</b>
Introduction	14
Overview Literature	15
Sociology of Inference	99
Practice Coding	100
Salaries	101
 <b>II   Presentation</b>	 <b>102</b>
Markdown	103
 <b>III  ——Computation——</b>	 <b>104</b>
 <b>IV  Computation</b>	 <b>105</b>
<b>Computation</b>	<b>106</b>
git . . . . .	106
<b>Bash</b>	<b>107</b>
<b>R</b>	<b>108</b>
Tidyverse . . . . .	108

<b>Python</b>	<b>109</b>
Numpy . . . . .	109
Pandas . . . . .	109
<b>jax</b>	<b>110</b>
<b>Numpyro</b>	<b>111</b>
<b>Stan</b>	<b>112</b>
brms . . . . .	112
<b>pyro</b>	<b>113</b>
<b>tensorflow</b>	<b>114</b>
<b>SQL</b>	<b>115</b>
<b>V Data management</b>	<b>116</b>
<b>Filter</b>	<b>117</b>
<b>Joins</b>	<b>120</b>
<b>Regex</b>	<b>121</b>
<b>Fuzzy Recording Matching</b>	<b>122</b>
<b>VI ———Certainty———</b>	<b>123</b>
<b>VII Mathematical Objects</b>	<b>124</b>
<b>Set</b>	<b>125</b>
<b>List (Sequence)</b>	<b>126</b>
<b>Vector/Matrix/Tensor</b>	<b>129</b>
<b>Table</b>	<b>134</b>
Introduction . . . . .	134

<b>Function</b>	<b>138</b>
Introduction . . . . .	138
Frequentist . . . . .	138
Bayesian . . . . .	140
 <b>VIII Operations of Logic</b>	 <b>141</b>
<b>And</b>	<b>142</b>
Tensorflow . . . . .	145
 <b>IX Operations of Arithmetic</b>	 <b>147</b>
<b>Addition</b>	<b>148</b>
Introduction . . . . .	148
Frequentist . . . . .	148
Bayesian . . . . .	149
 <b>Subtraction</b>	 <b>151</b>
Introduction . . . . .	151
Frequentist . . . . .	151
Bayesian . . . . .	152
 <b>Multiplication</b>	 <b>154</b>
Introduction . . . . .	154
Frequentist . . . . .	154
Bayesian . . . . .	156
 <b>Division</b>	 <b>157</b>
Introduction . . . . .	157
Frequentist . . . . .	157
Bayesian . . . . .	158
 <b>Modulo</b>	 <b>160</b>
Tensorflow . . . . .	163
 <b>X Operations of Algebra</b>	 <b>164</b>
<b>Dot product</b>	<b>165</b>
Introduction . . . . .	165

Bayesian . . . . .	166
<b>XI ———Uncertainty———</b>	<b>168</b>
<b>XII Probability</b>	<b>169</b>
Probability distribution	170
Random Variable	172
Probability Mass/Density Function (PMF/PDF)	174
<b>XIIIMoments of a Distribution</b>	<b>176</b>
Mean	177
Introduction . . . . .	177
Frequentist . . . . .	177
Bayesian . . . . .	181
<b>XIVDistributions</b>	<b>182</b>
<b>XV Information</b>	<b>183</b>
Entropy	184
<b>XVIIInference</b>	<b>185</b>
Bayesianism	186
Frequentism	187
<b>XVIEstimation</b>	<b>188</b>
Performance	189
Out of Sample Performance	190

<b>Regularization</b>	<b>192</b>
<b>P Values</b>	<b>193</b>
<b>Bias Variance Tradeoff</b>	<b>194</b>
<b>Bias</b>	<b>195</b>
<b>Variance</b>	<b>196</b>
<b>Asymptotics</b>	<b>197</b>
Introduction . . . . .	197
Frequentist . . . . .	197
Bayesian . . . . .	198
<b>XVIII Domain / Generalizability / Transportability</b>	<b>200</b>
<b>Outliers</b>	<b>202</b>
<b>Regime Change</b>	<b>203</b>
<b>Internal Validity</b>	<b>204</b>
<b>Transportability</b>	<b>205</b>
<b>External Validity</b>	<b>206</b>
<b>Matching</b>	<b>207</b>
<b>Poststratification</b>	<b>208</b>
<b>Outcome regressions / Response Surface Modeling</b>	<b>209</b>
<b>XIX Likelihood</b>	<b>210</b>
<b>Likelihood Function</b>	<b>211</b>
<b>Maximum Likelihood Estimation (MLE)</b>	<b>214</b>
<b>XX ———Measurement———</b>	<b>215</b>

<b>XXI Data and Measurement</b>	<b>216</b>
Data Validity	217
Data Reliability	218
<b>XXI Research Design</b>	<b>219</b>
Research Design Directed Acyclic Graphs (DAGs)	220
Potential Outcomes	221
Project Design	222
Unit of Analysis	223
Estimand	224
Identification	225
Garden of Forking Paths	226
Confounder	227
Unobserved Confounding	228
<b>XXI Prediction/Forecasting</b>	<b>229</b>
<b>XXI Counterfactual causal inference</b>	<b>230</b>
<b>XXV Causality the 12 Assumptions</b>	<b>231</b>
No unmeasured confounders	232
Correct model specification	233
No conditioning on a collider	234
No conditioning on Mediators	235

Positivity	236
Consistency	237
No Interference	238
No Relevant Effect Modification	239
Collapsibility	240
Compliance	241
Missing Data Mechanism	242
Transparency	243
Recoverability	244
Testability	245
No Relevant Measurement Error	246
<b>XXIV Exogeneity</b>	<b>247</b>
Random Control Trials	248
Instrumental Variables	249
Difference in Difference	250
Placebo Tests	251
Regression Discontinuity (RDD)	252
Fixed Effects	253
Synthetic Controls	254
<b>XXV Supervised Learning</b>	<b>255</b>
OLS	256
Interactions	257



Decision Trees	258
Random Forest	259
Gradientboosting	260
Gaussian Processes	261
<b>XXVII</b> Reinforcement Learning	262
Reinforcement Learning	264
<b>XXIX</b> Unsupervised Learning	265
Distance Metrics	266
<b>XXX</b> Neural Networks	267
Neural Network Debugging	268

# Preface

[UNDER CONSTRUCTION NOT FOR CIRCULATION]

Mathematics is the language of the Universe, and like any foreign language, it's not something you learn, it's something you get used to. This book is an artifact resulting from getting used to it over ten years as working scientists and before that 9 years of Phd/MA/BA. Its design philosophy is not to be a coherent text but rather a reference guide. The reader should be able to quickly turn to a topic (or find it through the search bar), and see multiple representations of that idea in english, formalism, code, and relevant papers. It is not designed to teach these topics for the first time, rather it is to help a working applied scientist quickly relearn and apply a topic.

# Overviews

[Regression and Other Stories](#) (Gelman, Hill, and Vehtari 2020)

[Applied Bayesian Modelling](#) (Congdon 2014)

[Causal Inference The Mixtape](#) (Cunningham 2021)

[Introduction to the concept of likelihood and its applications](#) (Etz 2017)

[Introduction to Probability for Data Science](#) (**IntroductionProbabilityData?**)

[A Review of Generalizability and Transportability](#) (Degtiar and Rose 2023)

[Regression and Causality](#) (Schomaker 2021)

[What are the most important statistical ideas of the past 50 years?](#) (Gelman and Vehtari 2021)

<https://statquest.org/>

[The Effect: An Introduction to Research Design and Causality](#) (**huntington-kleinEffectIntroductionResearch?**)

[Natural Language Processing Advancements By Deep Learning: A Survey](#) (Torfi et al. 2021)

[Minimum Viable Study Plan for Machine Learning Interviews](#) (Pham [2020] 2022)

<https://www.bradyneal.com/causal-inference-course> [Introduction to Causal Inference](#) (Neal 2020a)

[What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory](#) (Lundberg, Johnson, and Stewart 2021)

[applied-ml](#)

[ML use cases by company](#)

<https://github.com/khangich/machine-learning-interview/blob/master/extra.md>

[The Ultimate Guide to Machine Learning Job Interviews](#)

[InfoQ](#)

[Introducing the Facebook Field Guide to Machine Learning video series](#)

[Advice On Interviewing With Amazon](#)

Do you see companies asking to implement algorithm from scratch? Yes, some companies like LinkedIn, Intuit did. The common questions include: implement kmeans, linear/logistic regression. You can find the code here. [backprop https://github.com/khangich/machine-learning-interview/blob/master/sample/backprop.py](#) [kmeans https://github.com/khangich/machine-learning-interview/blob/master/sample/kmeans.ipynb](#) [logit https://github.com/khangich/machine-learning-interview/blob/master/sample/logistic\\_regression.ipynb](#)

[BAYES AND FREQUENTIST](#)

[Machine Learning engineer onsite interview: one week checklist](#)

(Pearl and Mackenzie 2018) [The Book of Why: The New Science of Cause and Effect](#)<http://bayes.cs.ucla.edu/WHY/jmdewhy-review2018.pdf>

(“Computational Linear Algebra - YouTube” n.d.) [Computational Linear Algebra for Coders](#)(*Fastai/Numerical-Linear-Algebra* [2017] 2022)

[Which causal inference book you should read A flowchart and a list of short book reviews](#)(Neal 2019)

[Detexify](#)

[https://bookdown.org/kevin\\_davisross/probsim-book/](https://bookdown.org/kevin_davisross/probsim-book/)(Ross 2022)

## References

# **Part I**

## **Introduction**

# Introduction

This is a book created from markdown and executable code.

See (**knuth84?**) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

# Overview Literature

Minimum Viable Study Plan for Machine Learning Interviews

<https://github.com/khangich/machine-learning-interview>

Causal Inference: The Mixtape <https://mixtape.scunning.com/index.html>

Bayesian Workflow Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, Martin Modrák <https://arxiv.org/abs/2011.01808>

How to avoid machine learning pitfalls: a guide for academic researchers Michael A. Lones <https://arxiv.org/abs/2108.02497>

Information geometry and divergences <https://franknielsen.github.io/IG/#bookIG>

Statistical Rethinking: A Bayesian Course with Examples in R and Stan (& PyMC3 & brms) <https://xcelab.net/rm/statistical-rethinking/> <https://www.youtube.com/playlist?list=PLDcUM9US4XdMROZ57-OIRtIK0aOynbgZN>

ML Frameworks Interoperability Cheat Sheet <http://blo.ocks.org/miguelusque/raw/f44a8e729896a96d0a3e4b07b517>

Regression and Other Stories, Andrew Gelman, Jennifer Hill, Aki Vehtari copy of the book <https://users.aalto.fi/~ave/ROS.pdf>

[tidybayes: Bayesian analysis + tidy data + geoms](#)

[Graphical Data Analysis with R](#) Antony Unwin

[Data Visualization](#) A practical introduction, Kieran Healy

[Bayes Rules! An Introduction to Applied Bayesian Modeling](#), Alicia A. Johnson, Miles Q. Ott, Mine Dogucu, 2021-12-01

[Bayesian Statistics Independent readings course on Bayesian statistics with R and Stan](#), Andrew Heiss and Meng Ye, Fall 2022 <https://bayesf22-notebook.classes.andrewheiss.com/rethinking/>

<https://bayesf22-notebook.classes.andrewheiss.com/bayes-rules/>

[Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis](#)

[An Introduction to Proximal Causal Learning](#)

[A Selective Review of Negative Control Methods in Epidemiology](#)

[Backpropagation is not just the chain rule%2C%20to%20predict%20y.\)](#)

[Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs Andrew Gelman & Guido Imbens](#)

R Markdown Cookbook Yihui Xie, Christophe Dervieux, Emily Riederer 2022-11-07 <https://bookdown.org/yihui/rmarkdown-cookbook/>

Understanding Machine Learning: From Theory to Algorithms  
<https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>

<https://simplystatistics.org/>

[Estimation Prediction, Estimation, and Attribution](#)

[The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant](#)

[A Parsimonious Tour of Bayesian Model Uncertainty](#)

[Causal Inference for the Brave and True](#)

<https://bayesiancomputationbook.com/welcome.html>

Measurement error and the replication crisis The assumption that measurement error always reduces effect sizes is false  
<https://www.science.org/doi/10.1126/science.aal3618>

<https://journals.sagepub.com/doi/abs/10.1177/00031224211004187#:~:text=The%20estimand%20is%20the%20t>

Exploring the Dynamics of Latent Variable Models  
<https://www.cambridge.org/core/journals/political-analysis/article/abs/exploring-the-dynamics-of-latent-variable-models/CBE116F37900DAE957B2D7EB53DB0907#.X7h7GMnwHwM.twitter>

<https://github.com/HenrikBengtsson/matrixStats>

[Let's Git started](#)



<https://github.com/facebookresearch/StarSpace>

<https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/>

What's Wrong With My Time Series Blog post by Alex Smolyanskaya ALEX SMOLYANSKAYA February 28, 2017 - San Francisco, CA Tweet this post! Post on LinkedIn What's wrong with my time series? Model validation without a hold-out set <https://multithreaded.stitchfix.com/blog/2017/02/28/whats-wrong-with-my-time-series/>

ggRandomForests: Exploring Random Forest Survival <https://arxiv.org/pdf/1612.08974.pdf>

<https://districtdatalabs.silvrback.com/time-maps-visualizing-discrete-events-across-many-timescales>

Explained Visually <https://setosa.io/ev/>

<https://github.com/google/BIG-bench/blob/main/docs/paper/BIG-bench.pdf>

[Two Experiments in Peer Review: Posting Preprints and Citation Bias](#)

Random Walk: A Modern Introduction Gregory F. Lawler and Vlada Limic

Can Transformers be Strong Treatment Effect Estimators? <https://arxiv.org/pdf/2202.01336v1.pdf>

Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition <https://bookdown.org/content/4857/>

Patches Are All You Need? <https://openreview.net/forum?id=TVHS5Y4dNvM>

The validate R-package makes it super-easy to check whether data lives up to expectations you have based on domain knowledge. It works by allowing <https://github.com/data-cleaning/validate>

Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong <https://journals.sagepub.com/doi/10.1080/07388940500339167>

autoxgboost <https://github.com/ja-thomas/autoxgboost>

1,500 scientists lift the lid on reproducibility <https://www.nature.com/articles/533452a>

Methodology over metrics: current scientific standards are a disservice to patients and society [https://www.jclinepi.com/article/S0895-4356\(21\)00170-0/fulltext](https://www.jclinepi.com/article/S0895-4356(21)00170-0/fulltext)

bper: Bayesian Prediction for Ethnicity and Race <https://github.com/bwilden/bper>

Automatic Differentiation Variational Inference <https://www.jmlr.org/papers/volume18/16-107/16-107.pdf>

What are the most important statistical ideas of the past 50 years? Andrew Gelman, Aki Vehtari <https://arxiv.org/pdf/2012.00174.pdf>

Why Propensity Scores Should Not Be Used for Matching <https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-for-matching>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/>

PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R <https://cran.r-project.org/web/packages/PRROC/vignettes/PRROC.pdf>

On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives <https://arxiv.org/abs/1902.10286>

['Trust Us': Open Data and Preregistration in Political Science and International Relations] <https://osf.io/preprints/metaarxiv/8h2bp/>

pals [https://cran.r-project.org/web/packages/pals/vignettes/pals\\_examples.html](https://cran.r-project.org/web/packages/pals/vignettes/pals_examples.html)

Greedy Function Approximation: A Gradient Boosting Machine <https://jerryfriedman.su.domains/ftp/trebst.pdf>

Natural Scales in Geographical Patterns <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5379183/>  
<https://daattali.com/shiny/timevis-demo/>

<https://www.extremetech.com/computing/151980-inside-ibms-67-billion-sage-the-largest-computer-ever-built>

Faux peer-reviewed journals: a threat to research integrity <http://deevybee.blogspot.com/2020/12/?m=1>

<https://github.com/mmxgn/spacy-clausie>

<http://deevybee.blogspot.com/2020/12/?m=1>

<http://www.deeplearningbook.org>

Statistical Nonsignificance in Empirical Economics <https://www.aeaweb.org/articles?id=10.1257/aeri.20190252&fr>

Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions\* Seth J. Hill† Margaret E. Roberts‡ October 25, 2021 [http://www.margaretroberts.net/wp-content/uploads/2021/10/hillroberts\\_acqbiaspoliticalbeliefs.pdf](http://www.margaretroberts.net/wp-content/uploads/2021/10/hillroberts_acqbiaspoliticalbeliefs.pdf)

The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable <https://www.nature.com/articles/s41591-021-01535-y>

<https://www.math.uzh.ch/pages/varrank/index.html>

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing <https://arxiv.org/pdf/2107.13586.pdf>

How should variable selection be performed with multiply imputed data? <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3177>

Feature Interactions in XGBoost <https://arxiv.org/abs/2007.05758>

Landscape of R packages for eXplainable Artificial Intelligence by Szymon Maksymiuk, Alicja Gosiewska, Przemysław Biecek <https://arxiv.org/pdf/2009.13248.pdf>

Feature Engineering and Selection: A Practical Approach for Predictive Models <https://bookdown.org/max/FES/>

Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC300808/>

xgboost.surv <https://github.com/bcjaeger/xgboost.surv>

DoubleML The Python and R package DoubleML provide an implementation of the double / debiased machine learning framework of Chernozhukov et al. (2018). The Python package is built on top of scikit-learn (Pedregosa et al., 2011) and the R package on top of mlr3 and the mlr3 ecosystem (Lang et al., 2019). <https://docs.doubleml.org/stable/index.html>

Preplication, Replication: A Proposal to Efficiently Upgrade Journal Replication Standards Get access Arrow Michael Colaresi <https://academic.oup.com/isp/article-abstract/17/4/367/2528282?redirectedFrom=fulltext>

<https://deepmind.com/blog/article/using-jax-to-accelerate-our-research>

<https://github.com/tidyverts/fable>

The Effect: An Introduction to Research Design and Causality  
<https://theeffectbook.net/>

<https://github.com/dedupeio/dedupe>

<https://arxiv.org/abs/2205.07407> What GPT Knows About Who is Who Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, Chris Tanner

An Introduction to Ontology Engineering <https://people.cs.uct.ac.za/~mkeet/files/OEbook.pdf>

R Packages for Item Response Theory Analysis: Descriptions and Features <https://www.tandfonline.com/doi/full/10.1080/15366367.2019.1586404>

Accuracy vs Explainability of Machine Learning Models [NIPS workshop poster review] <https://www.inference.vc/accuracy-vs-explainability-in-machine-learning-models-nips-workshop-poster-review/>

<https://arxiv-sanity-lite.com/>

Attitudes toward amalgamating evidence in statistics\* Andrew Gelman† Keith O'Rourke‡ <http://www.stat.columbia.edu/~gelman/research/unpublished/Amalgamating6.pdf>

An overview of gradient descent optimization algorithms  
<https://ruder.io/optimizing-gradient-descent/>

<https://codeocean.com/>

ClustGeo: an R package for hierarchical clustering with spatial constraints <https://arxiv.org/pdf/1707.03897.pdf>

An Algorithmic Framework for Bias Bounties Ira Globus-Harris, Michael Kearns, Aaron Roth <https://arxiv.org/abs/2201.10408>

On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis <https://arxiv.org/pdf/1707.01780.pdf>

Fast TreeSHAP: Accelerating SHAP Value Computation for Trees Jilei Yang <https://arxiv.org/abs/2109.09847>

Comparing interpretability and explainability for feature selection Jack Dunn, Luca Mingardi, Ying Daisy Zhuo <https://arxiv.org/abs/2105.05328>

Training Deep Nets with Sublinear Memory Cost Tianqi Chen, Bing Xu, Chiyuan Zhang, Carlos Guestrin <https://arxiv.org/abs/1604.06174>

ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R <https://arxiv.org/pdf/1508.04409.pdf>

A Survey of Recent Abstract Summarization Techniques Diyah Puspitaningrum <https://arxiv.org/abs/2105.00824>

UNDERSTANDING RANDOM FORESTS from theory to practice <https://arxiv.org/pdf/1407.7502.pdf>

Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology <https://arxiv.org/pdf/1809.03006.pdf>

Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO Ray Bai, Veronika Rockova, Edward I. George <https://arxiv.org/abs/2010.06451>

Representation Tradeoffs for Hyperbolic Embeddings Christopher De Sa† Albert Gu† Christopher Re† Frederic Sala† <https://arxiv.org/pdf/1804.03329.pdf>

Ratios: A short guide to confidence limits and proper use V.H. Franz\* October, 2007 <https://arxiv.org/pdf/0710.2024.pdf>

The Endogeneity of Historical Data Posted on August 28, 2020 by Adam Slez <https://broadstreet.blog/2020/08/28/the-endogeneity-of-historical-data/>

A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251194>

Post model-fitting exploration via a “Next-Door” analysis Leying GUAN<sup>1\*</sup> and Robert TIBSHIRANI<sup>2</sup> <https://tibshirani.su.domains/ftp/nextDoor.pdf>

Understanding BERT Transformer: Attention isn’t all you need A parsing/composition framework for understanding Transformers <https://medium.com/synapse-dev/understanding-bert-transformer-attention-isnt-all-you-need-5839ebd396db>

Einstein VI: General and Integrated Stein Variational Inference in NumPyro Ahmad Salim Al-Sibahi, Ola Rønning, Christophe Ley, Thomas Wim Hamelryck <https://openreview.net/forum?id=nXSDybDWV3>

Dream Investigation Results Official Report by the Minecraft Speedrunning Team <https://mcspeedrun.com/dream.pdf>

Improving Parameter Estimation of Epidemic Models: Likelihood Functions and Kalman Filtering 39 Pages Posted: 8 Aug 2022 [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4165188](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4165188)

Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment Published online by Cambridge University Press: 09 March 2021 <https://www.cambridge.org/core/journals/political-analysis/article/abs/do-namebased-treatments-violate-information-equivalence-evidence-from-a-correspondence-audit-experiment/56C6846518DDADE6EAF92DAE11552BDF>

How Much Should We Trust Staggered Difference-In-Differences Estimates? European Corporate Governance Institute – Finance Working Paper No. 736/2021 Rock Center for Corporate Governance at Stanford University Working Paper No. 246 Journal of Financial Economics (JFE), Forthcoming [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3794018](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3794018)

Building useful models for industry—some tips Jim Savage January 2017 <https://khakieconomics.github.io/2017/01/01/Building-useful-models-for-industry.html>

An Introduction to Proximal Causal Learning <https://arxiv.org/pdf/2009.10982.pdf>

First Things First: Assessing Data Quality before Model Quality Anita Gohdes and Megan Price [meganp@benetech.org](mailto:meganp@benetech.org) View all authors and affiliations [https://journals.sagepub.com/doi/full/10.1177/0022002712459708?casa\\_token=xXfXTkmn9p94f4BFh60b0eH\\_PE](https://journals.sagepub.com/doi/full/10.1177/0022002712459708?casa_token=xXfXTkmn9p94f4BFh60b0eH_PE)

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans <https://www.nature.com/articles/s42256-021-00307-0>

Why and How We Should Join the Shift From Significance Testing to Estimation <https://www.preprints.org/manuscript/202112.0235/v1>

How to make replication the norm <https://www.nature.com/articles/d41586-018-02108-9>

Applied Bayesian Statistics Using Stan and R <https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/applied-bayesian-statistics/>

<https://seeing-theory.brown.edu/index.html>

<https://www.brodrigues.co/>

FINDING ECONOMIC ARTICLES WITH DATA AND SPECIFIC EMPIRICAL METHODS <http://skranz.github.io/r/2021/01/05/FindingEconomicArticles4.html>

Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2849145](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145)

Machine vision on historical maps <https://weinman.cs.grinnell.edu/research/maps.shtml>

Enhancing Validity in Observational Settings When Replication Is Not Possible [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2543525](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2543525)

1.1 Billion Taxi Rides with SQLite, Parquet & HDFS <https://tech.marksblogg.com/billion-nyc-taxi-rides-sqlite-parquet-hdfs.html>

Understanding the Bias-Variance Tradeoff <http://scott.fortmann-roe.com/docs/BiasVariance.html>

Is the LKJ(1) prior uniform? “Yes” <http://srmart.in/is-the-lkj1-prior-uniform-yes/>

Informative priors for correlation matrices: An easy approach <http://srmart.in/informative-priors-for-correlation-matrices-an-easy-approach/>

A Tutorial on Spectral Clustering <https://arxiv.org/pdf/0711.0189v1.pdf>

Automated Geocoding of Textual Documents: A Survey of Current Approaches <https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12212>

Sparklyr <https://spark.rstudio.com/>

The AAA Tranche of Subprime Science Andrew Gelman and Eric Loken <http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics10.pdf>

Never trust rownames of a dataframe June 16th, 2015 by Ankur Gupta | <https://www.perfectlyrandom.org/2015/06/16/never-trust-the-row-names-of-a-dataframe-in-R/>

GRAPH ALGORITHMS <http://www.martinbroadhurst.com/tag/igraph>

Groundhog: Addressing The Threat That R Poses To Reproducible Research <http://datacolada.org/95>

CS231n Convolutional Neural Networks for Visual Recognition <https://cs231n.github.io/neural-networks-3/>

Implementing Variational Autoencoders in Keras: Beyond the Quickstart Tutorial <http://louistiao.me/posts/implementing-variational-autoencoders-in-keras-beyond-the-quickstart-tutorial/>

Hypothesis Testing in Econometrics <http://home.uchicago.edu/amshaikh/webfiles/testingreview.pdf>

“Why Should I Trust You?” Explaining the Predictions of Any Classifier <https://arxiv.org/pdf/1602.04938v3.pdf>

Yes, but Did It Work?: Evaluating Variational Inference [http://www.stat.columbia.edu/~gelman/research/published/Evaluating\\_Variational\\_Inference.pdf](http://www.stat.columbia.edu/~gelman/research/published/Evaluating_Variational_Inference.pdf)  
<https://statmodeling.stat.columbia.edu/2018/06/27/yes-work-evaluating-variational-inference/>

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets <https://arxiv.org/abs/2103.12028>

One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV Joshua Angrist, Michal Kolesár <https://arxiv.org/abs/2110.10556>

Underspecification Presents Challenges for Credibility in Modern Machine Learning <https://arxiv.org/abs/2011.03395>

A Survey of Predictive Modelling under Imbalanced Distributions <https://arxiv.org/pdf/1505.01658.pdf>

Varying Slopes Models and the CholeskyLKJ distribution in TensorFlow Probability <https://adamhaber.github.io/post/varying-slopes/>

Shapley Decomposition of R-Squared in Machine Learning Models <https://arxiv.org/pdf/1908.09718.pdf>



Understanding Global Feature Contributions With Additive Importance Measures Ian Covert, Scott Lundberg, Su-In Lee  
<https://arxiv.org/abs/2004.00668>

True to the Model or True to the Data? <https://arxiv.org/abs/2006.16234>

When to Impute? Imputation before and during cross-validation Byron C. Jaeger\*1 | Nicholas J. Tierney2 | Noah R. Simon3 <https://arxiv.org/pdf/2010.00718.pdf>

A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications Hongyun Cai, Vincent W. Zheng, Kevin Chen-Chuan Chang <https://arxiv.org/abs/1709.07604>

Comparing methods addressing multi-collinearity when developing prediction models <https://arxiv.org/abs/2101.01603>

Nonparametric causal effects based on incremental propensity score interventions <https://arxiv.org/abs/1704.00211>

Deep learning generalizes because the parameter-function map is biased towards simple functions Guillermo Valle-Pérez, Chico Q. Camargo, Ard A. Louis <https://arxiv.org/abs/1805.08522>

Bayesian Item Response Modeling in R with brms and Stan <https://arxiv.org/pdf/1905.09501.pdf>

Bayesian Inference for a Covariance Matrix <https://arxiv.org/pdf/1408.4050.pdf>

Cross-validation Confidence Intervals for Test Error Pierre Bayle, Alexandre Bayle, Lucas Janson, Lester Mackey <https://arxiv.org/abs/2007.12671>

Comparing Published Scientific Journal Articles to Their Preprint Versions <https://arxiv.org/pdf/1604.05363.pdf>

End-to-End Weak Supervision Salva Rühling Cachay, Benedikt Boecking, Artur Dubrawski <https://arxiv.org/abs/2107.02233>

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests\* <https://arxiv.org/pdf/1510.04342.pdf>

Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift <https://arxiv.org/pdf/1801.05134.pdf>

A review on outlier/anomaly detection in time series data <https://arxiv.org/abs/2002.04236>

Entropic Out-of-Distribution Detection: Seamless Detection of Unknown Examples David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano L. I. Oliveira, Teresa Ludermir <https://arxiv.org/abs/2006.04005>

An Exploratory Characterization of Bugs in COVID-19 Software Projects Akond Rahman, Effat Farhana <https://arxiv.org/abs/2006.00586>

Be Careful What You Backpropagate: A Case For Linear Output Activations & Gradient Boosting Anders Oland, Aayush Bansal, Roger B. Dannenberg, Bhiksha Raj <https://arxiv.org/abs/1707.04199>

Introducing Stan2tfp - a lightweight interface for the Stan-to-TensorFlow Probability compiler May 21, 2020 4 min read <https://adamhaber.github.io/post/stan2tfp-post1/>

L2 Regularization versus Batch and Weight Normalization Twan van Laarhoven <https://arxiv.org/abs/1706.05350>

Unsupervised Discovery of Temporal Structure in Noisy Data with Dynamical Components Analysis David G. Clark, Jesse A. Livezey, Kristofer E. Bouchard <https://arxiv.org/abs/1905.09944>

Monte Carlo Gradient Estimation in Machine Learning Shakir Mohamed, Mihaela Rosca, Michael Figurnov, Andriy Mnih <https://arxiv.org/abs/1906.10652>

Large-scale linear regression: Development of high-performance routines Alvaro Frank, Diego Fabregat-Traver, Paolo Bientinesi <https://arxiv.org/abs/1504.07890>

The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions Raj Agrawal, Jonathan H. Huggins, Brian Trippe, Tamara Broderick <https://arxiv.org/abs/1905.06501>

TensorFlow Distributions Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, Rif A. Saurous <https://arxiv.org/abs/1711.10604>

Asymptotically Exact, Embarrassingly Parallel MCMC Willie Neiswanger, Chong Wang, Eric Xing <https://arxiv.org/abs/1311.4780>

Python for Data Science <https://aeturrell.github.io/python4DS/welcome.html>

Using the flextable R package <https://ardata-fr.github.io/flextable-book/>

Coding for Economists <https://aeturrell.github.io/coding-for-economists/intro.html>

When Should You Adjust Standard Errors for Clustering? Get access Arrow Alberto Abadie, Susan Athey, Guido W Imbens, Jeffrey M Wooldridge <https://academic.oup.com/qje/advance-article-abstract/doi/10.1093/qje/qjac038/6750017>

Awesome Deep Learning for Natural Language Processing (NLP) <https://github.com/brianspiering/awesome-dl4nlp>

R for applied epidemiology and public health <https://epirhandbook.com/en/index.html>

COVID 19: Reduced forms have gone viral, but what do they tell us?\* [https://drive.google.com/file/d/1ERjcGXD2jvfDFXdl0\\_NtF4X95UeQ5f4W/view](https://drive.google.com/file/d/1ERjcGXD2jvfDFXdl0_NtF4X95UeQ5f4W/view)

Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology <https://elifesciences.org/articles/67995>

Taking Uncertainty Seriously: Bayesian Marginal Structural Models for Causal Inference in Political Science <https://github.com/ajnafa/Latent-Bayesian-MSM>

Generalized Linear Models <https://data.princeton.edu/wws509/notes/c7s4>

genieclust: Fast and Robust Hierarchical Clustering with Noise Point Detection <https://genieclust.gagolewski.com/>

Awesome Graph Classification <https://github.com/benedekrozemberczki/awesome-graph-classification>

parallelDist <https://github.com/alexreckert/parallelDist>

Interpretable Machine Learning A Guide for Making Black Box Models Explainable Christoph Molnar <https://christophm.github.io/interpretable-ml-book/>

The Inverse CDF Method [https://dk81.github.io/dkmathstats\\_site/prob-inverse-cdf.html](https://dk81.github.io/dkmathstats_site/prob-inverse-cdf.html)

HamiltonianMC <https://chi-feng.github.io/mcmc-demo/app.html#HamiltonianMC,banana>

End-to-End Balancing for Causal Continuous Treatment-Effect Estimation <https://assets.amazon.science/5b/71/fa078e6f4f97a76a2a622c767dd5/end-to-end-balancing-for-causal-continuous-treatment-effect-estimation.pdf>

A tour of probabilistic programming language APIs What does it look like to do MCMC in different frameworks? <https://colcarroll.github.io/ppl-api/>

Probabilistic Programming & Bayesian Methods for Hackers <https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/>

Beyond Multiple Linear Regression Applied Generalized Linear Models and Multilevel Models in R <https://bookdown.org/roback/bookdown-bysh/>

Maybe a section on hyperparameters?

Does batch size matter? <https://blog.janestreet.com/does-batch-size-matter/>

The Much Quieter Revolution of Synthetic Control: Episode I [https://causalinf.substack.com/p/the-much-quieter-revolution-of-synthetic?utm\\_campaign=post&utm\\_medium=web&utm\\_source=](https://causalinf.substack.com/p/the-much-quieter-revolution-of-synthetic?utm_campaign=post&utm_medium=web&utm_source=)

User-friendly introduction to PAC-Bayes bounds <https://arxiv.org/pdf/2110.11216.pdf>

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results <https://journals.sagepub.com/doi/full/10.1177/2515245917747646>

The RecordLinkage Package: Detecting Errors in Data [https://journal.r-project.org/archive/2010-2/RJournal\\_2010-2\\_Sariyar+Borg.pdf](https://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf)

<https://grow.google/certificates/interview-warmup/>

The inverse-transform method for generating random variables in R <https://heds.nz/posts/inverse-transform/>

The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology <https://journals.sagepub.com/doi/10.1177/1948550616673876>

Evolution of Reporting P Values in the Biomedical Literature, 1990-2015 <https://jamanetwork.com/journals/jama/fullarticle/2503172>

SHAP (SHapley Additive exPlanations) <https://github.com/slundberg/shap>

The h-index is no longer an effective correlate of scientific reputation <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0253397>

Prior Choice Recommendations Andrew Gelman edited this page on Apr 17, 2020 · 51 revisions <https://github.com/standev/stan/wiki/Prior-Choice-Recommendations#prior-for-a-covariance-matrix>

Institute for Replication (I4R) <https://i4replication.org/index.html>

How Life Sciences Actually Work: Findings of a Year-Long Investigation <https://guzey.com/how-life-sciences-actually-work/>

Efficient Neural Causal Discovery without Acyclicity Constraints <https://github.com/phlippe/ENCO>

awesome-text-summarization <https://github.com/mathsyouth/awesome-text-summarization>

(Ir)Reproducible Machine Learning: A Case Study <https://reproducible.cs.princeton.edu/irreproducibility-paper.pdf>

THE MYTH OF THE EXPERT REVIEWER <https://parameterfree.com/2021/07/06/the-myth-of-the-expert-reviewer/>

Understanding and Choosing the Right Probability Distributions <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119197096.app03>

Spatial Interdependence and Instrumental Variable Models <https://osf.io/preprints/socarxiv/pgrcu/>

The case for formal methodology in scientific reform <https://royalsocietypublishing.org/doi/10.1098/rsos.200805>

Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3603970](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3603970)

Pandas Comparison with R / R libraries [https://pandas.pydata.org/docs/getting\\_started/comparison/comparison.html](https://pandas.pydata.org/docs/getting_started/comparison/comparison.html)

Non-Standard Errors <https://orbilu.uni.lu/bitstream/10993/48686/1/SSRN-id3961574.pdf>

Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors <https://pubmed.ncbi.nlm.nih.gov/26186114/>

The (lack of) impact of retraction on citation networks Charisse R Madlock-Brown 1, David Eichmann <https://pubmed.ncbi.nlm.nih.gov/24668038/>

The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature <https://psyarxiv.com/pbrdk/>

Bayesian Estimation of Correlation Matrices of Longitudinal Data Riddhi Pratim Ghosh, Bani Mallick, Mohsen Pourahmadi <https://projecteuclid.org/journals/bayesian-analysis/volume-16/issue-3/Bayesian-Estimation-of-Correlation-Matrices-of-Longitudinal-Data/10.1214/20-BA1237.full>

Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals <https://osf.io/preprints/socarxiv/cfdb/>

How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It <https://osf.io/preprints/socarxiv/453jk/>

Lost in Aggregation: Improving Event Analysis with Report-Level Data Scott J. Cook, Nils B. Weidmann <https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12398>

Frequentist versus Bayesian approaches to multiple testing Arvid Sjölander & Stijn Vansteelandt <https://link.springer.com/article/10.1007/s10654-019-00517-2>

Research note: Examining potential bias in large-scale censored data <https://misinfreview.hks.harvard.edu/article/research-note-examining-potential-bias-in-large-scale-censored-data/>

When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? Kosuke Imai, In Song Kim <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12417>

Runtime warnings and convergence problems Stan Development Team <https://mc-stan.org/misc/warnings.html>

Dirichlet Process Gaussian mixture model via the stick-breaking construction in various PPLs This page was last updated on 29 Mar, 2021. <https://luiarthur.github.io/TuringBnpBenchmarks/dpsbgmm>

xgboost: “Hi I’m Gamma. What can I do for you?” — and the tuning of regularization <https://medium.com/data-design/xgboost-hi-im-gamma-what-can-i-do-for-you-and-the-tuning-of-regularization-a42ea17e6ab6>

PostGIS In Action <https://livebook.manning.com/book/postgis-in-action-second-edition/about-this-book/>

Stan User's Guide <https://mc-stan.org/docs/stan-users-guide/index.html>

Smoothing Terms in GAM Models <https://maths-people.anu.edu.au/~johnm/r-book/xtras/autosmooth.pdf>

Designing a Deep Learning Project [https://medium.com/\(erogol/designing-a-deep-learning-project-9b3698aef127?\)](https://medium.com/(erogol/designing-a-deep-learning-project-9b3698aef127?))

PyTorch With Baby Steps: From  $y=x$  To Training A Convnet <https://lelon.io/blog/pytorch-baby-steps>

Bayesian inference with Stan: A tutorial on adding custom distributions Jeffrey Annis, Brent J. Miller & Thomas J. Palmeri <https://link.springer.com/article/10.3758/s13428-016-0746-9>

Bayes Rules! An Introduction to Applied Bayesian Modeling <https://www.bayesrulesbook.com/>

Graduate Qualitative Methods Training in Political Science: A Disciplinary Crisis Published online by Cambridge University Press: 21 November 2019 <https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/graduate-qualitative-methods-training-in-political-science-a-disciplinary-crisis/7B0EEB76E1CC234AFED7EED8DA71BA35>

Time Series Analysis by State Space Methods (Oxford Statistical Science Series) [https://www.amazon.com/dp/0198523548/ref=cm\\_sw\\_r\\_tw\\_apa\\_fabc\\_0MWV12PSS3K9NV](https://www.amazon.com/dp/0198523548/ref=cm_sw_r_tw_apa_fabc_0MWV12PSS3K9NV)

Hyperparameters and tuning strategies for random forest [https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1301?casa\\_token=\\_zNb\\_GkfYAUAAAAA%3AszhI](https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1301?casa_token=_zNb_GkfYAUAAAAA%3AszhI)

Your Cross Validation Error Confidence Intervals are Wrong — here's how to Fix Them <https://towardsdatascience.com/your-cross-validation-error-confidence-intervals-are-wrong-heres-how-to-fix-them-abbfe28d390>

Probabilistic Programming with Variational Inference: Under the Hood <https://willcrichton.net/notes/probabilistic-programming-under-the-hood/>

How to Measure Statistical Causality: A Transfer Entropy Approach with Financial Applications <https://towardsdatascience.com/causality-931372313a1c>

Kullback-Leibler Divergence Explained <https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained>

Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance <https://www.cs.purdue.edu/homes/lintan/publications/variance-ase20.pdf>

How (Not) to Reproduce: Practical Considerations to Improve Research Transparency in Political Science <https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/abs/how-not-to-reproduce-practical-considerations-to-improve-research-transparency-in-political-science/32E7CF5D975C081BA666D3BD475D7913>

Quantifying Bias from Measurable and Unmeasurable Confounders Across Three Domains of Individual Determinants of Political Preferences Published online by Cambridge University Press: 22 February 2022 <https://www.cambridge.org/core/journals/political-analysis/article/quantifying-bias-from-measurable-and-unmeasurable-confounders-across-three-domains-of-individual-determinants-of-political-preferences/D1D2DEE9E7180BDCFC592885BE66E9AF>

5 Levels of Difficulty — Bayesian Gaussian Random Walk with PyMC3 and Theano <https://towardsdatascience.com/5-levels-of-difficulty-bayesian-gaussian-random-walk-with-pymc3-and-theano-34343911c7d2>

Single-Parameter Models | Pyro vs. STAN <https://towardsdatascience.com/single-parameter-models-pyro-vs-stan-e7e69b45d95c>

Partial Identification in Econometrics Elie Tamer <https://scholar.harvard.edu/files/tamer/files/pie.pdf>

LightGBM for Quantile Regression Understand Quantile Regression <https://towardsdatascience.com/lightgbm-for-quantile-regression-4288d0bb23fd>

Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach [https://strathprints.strath.ac.uk/59463/1/Gallop\\_Weschle\\_PSRM\\_2016\\_Assessing\\_the\\_impact\\_of\\_non\\_rand](https://strathprints.strath.ac.uk/59463/1/Gallop_Weschle_PSRM_2016_Assessing_the_impact_of_non_rand)

yardstick is a package to estimate how well models are working using tidy data principles. See the package webpage for more information. <https://yardstick.tidymodels.org/index.html>

The Three Faces of Bayes <https://slackprop.wordpress.com/2016/08/28/the-three-faces-of-bayes/>



Evaluating Random Forests for Survival Analysis using Prediction Error Curves <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194196/>

The role of metadata in reproducible computational research  
<https://www.sciencedirect.com/science/article/pii/S2666389921001707>

Ecological Inference in the Social Sciences <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885825/>

Two Wrongs Make a Right: Addressing Underreporting in Binary Data from Multiple Sources <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667662/>

On the low reproducibility of cancer studies <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6599599/>

Quarto with Python <https://www.meyerperin.com/using-quarto/>

Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015  
<https://www.nature.com/articles/s41562-018-0399-z>

Bayesian analysis of tests with unknown specificity and sensitivity\* Andrew Gelman† and Bob Carpenter‡  
<https://www.medrxiv.org/content/10.1101/2020.05.22.20108944v3.full.pdf>

Notes on the Negative Binomial Distribution [https://www.johndcook.com/negative\\_binomial.pdf](https://www.johndcook.com/negative_binomial.pdf)

The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!) <https://www.joelonsoftware.com/2003/10/08/the-absolute-minimum-every-software-developer-absolutely-positively-must-know-about-unicode-and-character-sets-no-excuses/>

Automatic Differentiation Variational Inference <https://www.jmlr.org/papers/volume18/16-107/16-107.pdf>

IZA DP No. 13233: The Influence of Hidden Researcher Decisions in Applied Microeconomics <https://www.iza.org/publications/dp/13233/the-influence-of-hidden-researcher-decisions-in-applied-microeconomics>

cdfquantreg: An R Package for CDF-Quantile Regression  
<https://www.jstatsoft.org/article/view/v088i01>

<https://techdevguide.withgoogle.com/>

What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes David M. W. Powers <https://arxiv.org/abs/1503.06410>

When LOO and other cross-validation approaches are valid  
<https://statmodeling.stat.columbia.edu/2018/08/03/loo-cross-validation-approaches-valid/>

Hamiltonian Monte Carlo explained [http://arogozhnikov.github.io/2016/12/19/markov\\_chain\\_monte\\_carlo.html](http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html)

Controlling for Unobserved Confounds in Classification Using Correlational Constraints Virgile Landeiro, Aron Culotta  
<https://arxiv.org/abs/1703.01671>

The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies Scott E. Maxwell <https://statmodeling.stat.columbia.edu/wp-content/uploads/2017/07/maxwell2004.pdf>

You need 16 times the sample size to estimate an interaction than to estimate a main effect <https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/>

Machine Learning of Sets [http://akosiorek.github.io/ml/2020/08/12/machine\\_learning\\_of\\_sets.html](http://akosiorek.github.io/ml/2020/08/12/machine_learning_of_sets.html)

Weak Supervision: A New Programming Paradigm for Machine Learning <http://ai.stanford.edu/blog/weak-supervision/>

The earth is flat ( $p > 0.05$ ): Significance thresholds and the crisis of unreplicable research <https://peerj.com/preprints/2921/>

Advanced Natural Language Processing with TensorFlow 2: Build effective real-world NLP applications using NER, RNNs, seq2seq models, Transformers, and more  
[https://www.amazon.com/Advanced-Natural-Language-Processing-TensorFlow/dp/1800200935?encoding=UTF8&qid=1558444444&sr=8&linkCode=sl1&tag=kirkdborne-20&linkId=4448e1a0cd126f52a2aba844c4bdb78e&language=en\\_US&ref=as\\_li\\_ss\\_tl](https://www.amazon.com/Advanced-Natural-Language-Processing-TensorFlow/dp/1800200935?encoding=UTF8&qid=1558444444&sr=8&linkCode=sl1&tag=kirkdborne-20&linkId=4448e1a0cd126f52a2aba844c4bdb78e&language=en_US&ref=as_li_ss_tl)

3 reasons why you can't always use predictive performance to choose among models <https://statmodeling.stat.columbia.edu/2015/10/23/26857/>

Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models Arthur Lewbel  
<https://www.tandfonline.com/doi/full/10.1080/07350015.2012.643126>

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift Sergey Ioffe, Christian Szegedy <https://arxiv.org/abs/1502.03167>

Gradient Boosting explained [demonstration] [http://arogozhnikov.github.io/2016/06/24/gradient\\_boosting\\_explained.html](http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html)


Clustered standard errors vs. multilevel modeling [https://statmodeling.stat.columbia.edu/2007/11/28/clustered\\_s](https://statmodeling.stat.columbia.edu/2007/11/28/clustered_s)

Advanced R <https://adv-r.hadley.nz/index.html>

Regression to the mean continues to confuse people and lead to errors in published research <https://statmodeling.stat.columbia.edu/2018/06/24/regression-mean-continues-confuse-people-lead-errors-published-research/>

The statistical significance filter leads to overoptimistic expectations of replicability <https://statmodeling.stat.columbia.edu/2018/05/22/statistical-significance-filter-leads-overoptimistic-expectations-replicability/>

How to cross-validate PCA, clustering, and matrix decomposition models <http://alexhwilliams.info/itsneuronalblog/2018/02/26/crossval/?mlreview>

Inference in Experiments Conditional on Observed Imbalances in Covariates Per Johansson  & Mattias Nordin <https://www.tandfonline.com/doi/full/10.1080/00031305.2022.2054859>

Scientific progress despite irreproducibility: A seeming paradox Richard M. Shiffrin, Katy Borner, Stephen M. Stigler <https://arxiv.org/abs/1710.01946>

On Statistical Non-Significance Alberto Abadie <https://arxiv.org/abs/1803.00609>

On the number of signals in multivariate time series Markus Matilainen, Klaus Nordhausen, Joni Virta <https://arxiv.org/abs/1801.04925>

Data Science vs. Statistics: Two Cultures? Iain Carmichael, J.S. Marron <https://arxiv.org/abs/1801.00371>

The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions George Philipp, Dawn Song, Jaime G. Carbonell <https://arxiv.org/abs/1712.05577>

Theory of Deep Learning III: explaining the non-overfitting puzzle Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, Hrushikesh Mhaskar <https://arxiv.org/abs/1801.00173>

On overfitting and post-selection uncertainty assessments Liang Hong, Todd A. Kuffner, Ryan Martin <https://arxiv.org/abs/1712.02379>

A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results Beau Coker, Cynthia Rudin, Gary King <https://arxiv.org/abs/1804.08646>

Labelling as an unsupervised learning problem Terry Lyons, Imanol Perez Arribas <https://arxiv.org/abs/1805.03911>

Structural Breaks in Time Series Alessandro Casini, Pierre Perron <https://arxiv.org/abs/1805.03807>

On consistency and inconsistency of nonparametric tests Mikhail Ermakov <https://arxiv.org/abs/1807.09076>

A New Angle on L2 Regularization Thomas Tanay, Lewis D Griffin <https://arxiv.org/abs/1806.11186>

On the Robustness of Interpretability Methods David Alvarez-Melis, Tommi S. Jaakkola <https://arxiv.org/abs/1806.08049>

Identifying Causal Effects with the R Package causaleffect Santtu Tikka, Juha Karvanen <https://arxiv.org/abs/1806.07161>

How Does Batch Normalization Help Optimization? Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Madry <https://arxiv.org/abs/1805.11604>

The effect of the choice of neural network depth and breadth on the size of its hypothesis space Lech Szymanski, Brendan McCane, Michael Albert <https://arxiv.org/abs/1806.02460>

Is preprocessing of text really worth your time for online comment classification? Fahim Mohammad <https://arxiv.org/abs/1806.02908>

Geometric Understanding of Deep Learning Na Lei, Zhongxuan Luo, Shing-Tung Yau, David Xianfeng Gu <https://arxiv.org/abs/1805.10451>

Cross validation residuals for generalised least squares and other correlated data models Ingrid Annette Baade <https://arxiv.org/abs/1809.01319>

Out-of-Distribution Detection Using an Ensemble of Self Supervised Leave-out Classifiers Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, Theodore L. Willke <https://arxiv.org/abs/1809.03576>

Handling Imbalanced Dataset in Multi-label Text Categorization using Bagging and Adaptive Boosting Genta Indra Winata, Masayu Leylia Khodra <https://arxiv.org/abs/1810.11612>

On the Art and Science of Machine Learning Explanations Patrick Hall <https://arxiv.org/abs/1810.02909>

Causal inference under over-simplified longitudinal causal models Lola Etievant, Vivian Viallon <https://arxiv.org/abs/1810.01294>

Revisiting the Gelman-Rubin Diagnostic Dootika Vats, Christina Knudson <https://arxiv.org/abs/1812.09384>

A Survey on Data Collection for Machine Learning: a Big Data – AI Integration Perspective Yuji Roh, Geon Heo, Steven Eui-jong Whang

A Fundamental Measure of Treatment Effect Heterogeneity Jonathan Levy, Mark van der Laan, Alan Hubbard, Romain Pirracchio <https://arxiv.org/abs/1811.03745>

Causal Discovery Toolbox: Uncover causal relationships in Python Diviyan Kalainathan, Olivier Goudet <https://arxiv.org/abs/1903.02278>

Dying ReLU and Initialization: Theory and Numerical Examples Lu Lu, Yeonjong Shin, Yanhui Su, George Em Karniadakis <https://arxiv.org/abs/1903.06733>

ROC and AUC with a Binary Predictor: a Potentially Misleading Metric John Muschelli <https://arxiv.org/abs/1903.04881>

Gamification in Science: A Study of Requirements in the Context of Reproducible Research Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, Albrecht Schmidt <https://arxiv.org/abs/1903.02446>

Matrix factorization for multivariate time series analysis Pierre Alquier, Nicolas Marie <https://arxiv.org/abs/1903.05589>

On the complexity of logistic regression models Nicola Bulso, Matteo Marsili, Yasser Roudi <https://arxiv.org/abs/1903.00386>

On Heavy-user Bias in A/B Testing Yu Wang, Somit Gupta, Jiannan Lu, Ali Mahmoudzadeh, Sophia Liu <https://arxiv.org/abs/1902.02021>

DeepMoD: Deep learning for Model Discovery in noisy data  
Gert-Jan Both, Subham Choudhury, Pierre Sens, Remy Kusters

Learning Causality: Synthesis of Large-Scale Causal Networks from High-Dimensional Time Series Data Mark-Oliver Stehr, Peter Avar, Andrew R. Korte, Lida Parvin, Ziad J. Sahab, Deborah I. Bunin, Merrill Knapp, Denise Nishita, Andrew Poggio, Carolyn L. Talcott, Brian M. Davis, Christine A. Morton, Christopher J. Sevinsky, Maria I. Zavodszky, Akos Vertes <https://arxiv.org/abs/1905.02291>

Text Classification Algorithms: A Survey Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, Donald E. Brown <https://arxiv.org/abs/1904.08067>

The Information Complexity of Learning Tasks, their Structure and their Distance Alessandro Achille, Giovanni Paolini, Glen Mbeng, Stefano Soatto <https://arxiv.org/abs/1904.03292>

Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches Shane Storks, Qiaozi Gao, Joyce Y. Chai <https://arxiv.org/abs/1904.01172>

Evaluating A Key Instrumental Variable Assumption Using Randomization Tests Zach Branson, Luke Keele <https://arxiv.org/abs/1907.01943>

Model selection for high-dimensional linear regression with dependent observations Ching-Kang Ing <https://arxiv.org/abs/1906.07395>

Doubts on the efficacy of outliers correction methods Marjorie Fonnesu, Nicola Kuczewski

The Design of Global Correlation Quantifiers and Continuous Notions of Statistical Sufficiency Nicholas Carrara, Kevin Vanslette

An Econometric Perspective on Algorithmic Subsampling Sokbae Lee, Serena Ng <https://arxiv.org/abs/1907.01954>

Factor Analysis for High-Dimensional Time Series with Change Point Xialu Liu, Ting Zhang <https://arxiv.org/abs/1907.09522>

Causal Regularization Dominik Janzing <https://arxiv.org/abs/1906.12179>

The exact form of the ‘Ockham factor’ in model selection  
Jonathan Rougier, Carey Priebe <https://arxiv.org/abs/1906.11592>

Measuring Average Treatment Effect from Heavy-tailed Data  
Jason (Xiao)Wang, Pauline Burke <https://arxiv.org/abs/1905.09252>

The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial Benyamin Ghogh, Mark Crowley <https://arxiv.org/abs/1905.12787>

Statistical methods research done as science rather than mathematics James S. Hodges <https://arxiv.org/abs/1905.08381>

Regression Analysis of Unmeasured Confounding Brian Knaeble, Braxton Osting, Mark Abramson

Dyadic Regression Bryan S. Graham <https://arxiv.org/abs/1908.09029>

Illusion of Causality in Visualized Data Cindy Xiong, Joel Shapiro, Jessica Hullman, Steven Franconeri <https://arxiv.org/abs/1908.00215>

“multiColl”: An R package to detect multicollinearity Román Salmerón, Catalina García, José García <https://arxiv.org/abs/1910.14590>

All of Linear Regression Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Junhui Cai <https://arxiv.org/abs/1910.06386>

What is the Value of Data? On Mathematical Methods for Data Quality Estimation Netanel Raviv, Siddharth Jain, Jehoshua Bruck <https://arxiv.org/abs/2001.03464>

Imputation for High-Dimensional Linear Regression Kabir Aladin Chandrasekher, Ahmed El Alaoui, Andrea Montanari <https://arxiv.org/abs/2001.09180>

On Model Evaluation under Non-constant Class Imbalance Jan Brabec, Tomáš Komárek, Vojtěch Franc, Lukáš Machlica <https://arxiv.org/abs/2001.05571>

Identifying Mislabeled Data using the Area Under the Margin Ranking Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, Kilian Q. Weinberger <https://arxiv.org/abs/2001.10528>

Expanding the scope of statistical computing: Training statisticians to be software engineers Alex Reinhart, Christopher R. Genovese <https://arxiv.org/abs/1912.13076>

Learning under Model Misspecification: Applications to Variational and Ensemble methods Andres R. Masegosa <https://arxiv.org/abs/1912.08335>

Explaining the Explainer: A First Theoretical Analysis of LIME Damien Garreau, Ulrike von Luxburg <https://arxiv.org/abs/2001.03447>

Algorithms for Heavy-Tailed Statistics: Regression, Covariance Estimation, and Beyond Yeshwanth Cherapanamjeri, Samuel B. Hopkins, Tarun Kathuria, Prasad Raghavendra, Nilesh Tripuraneni <https://arxiv.org/abs/1912.11071>

Markov Chain Monte Carlo Methods, a survey with some frequent misunderstandings Christian P. Robert (U Paris Dauphine and U Warwick), Wu Changye (U Paris Dauphine) <https://arxiv.org/abs/2001.06249>

Valid p-Values and Expectations of p-Values Revisited Albert Vexler <https://arxiv.org/abs/2001.05126>

Counterexamples to “The Blessings of Multiple Causes” by Wang and Blei Elizabeth L. Ogburn, Ilya Shpitser, Eric J. Tchetgen Tchetgen <https://arxiv.org/abs/2001.06555>

Identifying Mislabeled Instances in Classification Datasets Nicolas Michael Müller, Karla Markert <https://arxiv.org/abs/1912.05283>

Randomized p-values for multiple testing and their application in replicability analysis Anh-Tuan Hoang, Thorsten Dickhaus <https://arxiv.org/abs/1912.06982>

Over-parametrized deep neural networks do not generalize well Michael Kohler, Adam Krzyzak <https://arxiv.org/abs/1912.03925>

Re-Evaluating Strengthened-IV Designs: Asymptotic Efficiency, Bias Formula, and the Validity and Power of Sensitivity Analyses Siyu Heng, Bo Zhang, Xu Han, Scott A. Lorch, Dylan S. Small <https://arxiv.org/abs/1911.09171>

Unbiased variable importance for random forests Markus Loecher <https://arxiv.org/abs/2003.02106>

Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited Wesley J. Maddox, Gregory Benton, Andrew Gordon Wilson <https://arxiv.org/abs/2003.02139>



A Multi-Way Correlation Coefficient Benjamin M. Taylor  
<https://arxiv.org/abs/2003.02561>

Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding Victor Veitch, Anisha Zaveri <https://arxiv.org/abs/2003.01747>

The Implicit and Explicit Regularization Effects of Dropout Colin Wei, Sham Kakade, Tengyu Ma <https://arxiv.org/abs/2002.12915>

Natural Language Processing Advancements By Deep Learning: A Survey Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, Edward A. Fox <https://arxiv.org/abs/2003.01200>

An Evaluation of Change Point Detection Algorithms Gerrit J.J. van den Burg, Christopher K.I. Williams <https://arxiv.org/abs/2003.06222>

Complexity Measures and Features for Times Series classification Francisco J. Baldán, José M. Benítez <https://arxiv.org/abs/2002.12036>

Computing Shapley Effects for Sensitivity Analysis Elmar Plischke, Giovanni Rabitti, Emanuele Borgonovo <https://arxiv.org/abs/2002.12024>

Bayesian Posterior Interval Calibration to Improve the Interpretability of Observational Studies Jami J. Mulgrave, David Madigan, George Hripcsak <https://arxiv.org/abs/2003.06002>

Demystify Lindley's Paradox by Interpreting P-value as Posterior Probability Guosheng Yin, Haolun Shi <https://arxiv.org/abs/2002.10883>

Estimation of causal effects with small data in the presence of trapdoor variables Jouni Helske, Santtu Tikka, Juha Karvanen <https://arxiv.org/abs/2003.03187>

Dimensional Analysis in Statistical Modelling Tae Yoon Lee, James V. Zidek, Nancy Heckman <https://arxiv.org/abs/2002.11259>

Causal bounds for outcome-dependent sampling in observational studies Erin E. Gabriel, Michael C. Sachs, Arvid Sjölander <https://arxiv.org/abs/2002.10519>

cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R <https://arxiv.org/abs/2002.09209>

A New Framework for Online Testing of Heterogeneous Treatment Effect Miao Yu, Wenbin Lu, Rui Song <https://arxiv.org/abs/2002.03277>

Combining Observational and Experimental Datasets Using Shrinkage Estimators Evan Rosenman, Guillaume Basse, Art Owen, Michael Baiocchi <https://arxiv.org/abs/2002.06708>

A confidence interval robust to publication bias for random-effects meta-analysis of few studies M. Henmi, S. Hattori, T. Friede <https://arxiv.org/abs/2002.07598>

Boosting Simple Learners Noga Alon, Alon Gonen, Elad Hazan, Shay Moran <https://arxiv.org/abs/2001.11704>

Analytic Study of Double Descent in Binary Classification: The Impact of Loss Ganesh Kini, Christos Thrampoulidis <https://arxiv.org/abs/2001.11572>

Fast Bayesian Estimation of Spatial Count Data Models Prateek Bansal, Rico Krueger, Daniel J. Graham <https://arxiv.org/abs/2007.03681>

High-recall causal discovery for autocorrelated time series with latent confounders Andreas Gerhardus, Jakob Runge <https://arxiv.org/abs/2007.01884>

Estimating the Prediction Performance of Spatial Models via Spatial k-Fold Cross Validation Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen, Jukka Heikkonen <https://arxiv.org/abs/2005.14263>

Validating Label Consistency in NER Data Annotation Qingkai Zeng, Mengxia Yu, Wenhao Yu, Tianwen Jiang, Meng Jiang <https://arxiv.org/abs/2101.08698>

Learning Prediction Intervals for Model Performance Benjamin Elder, Matthew Arnold, Anupama Murthi, Jiri Navratil <https://arxiv.org/abs/2012.08625>

Dive into Decision Trees and Forests: A Theoretical Demonstration Jinxiong Zhang <https://arxiv.org/abs/2101.08656>

Self-semi-supervised Learning to Learn from Noisy Labeled Data Jiacheng Wang, Yue Ma, Shuang Gao <https://arxiv.org/abs/2011.01429>

Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases Yu Gu, Sue Kase, Michelle

Vanni, Brian Sadler, Percy Liang, Xifeng Yan, Yu Su  
<https://arxiv.org/abs/2011.07743>

Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, Alice Xiang <https://arxiv.org/abs/2011.07586>

A Survey on Data Augmentation for Text Classification Markus Bayer, Marc-André Kaufhold, Christian Reuter  
<https://arxiv.org/abs/2107.03158>

A Survey on Automated Fact-Checking Zhijiang Guo, Michael Schlichtkrull, Andreas Vlachos <https://arxiv.org/abs/2108.11896>

The Benchmark Lottery Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, Oriol Vinyals <https://arxiv.org/abs/2107.07002>

Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, Philip Resnik <https://arxiv.org/abs/2107.02173>

The Modern Mathematics of Deep Learning Julius Berner, Philipp Grohs, Gitta Kutyniok, Philipp Petersen <https://arxiv.org/abs/2105.04026>

Biases in human mobility data impact epidemic modeling Frank Schlosser, Vedran Sekara, Dirk Brockmann, Manuel Garcia-Herranz <https://arxiv.org/abs/2112.12521>

Sparse-softmax: A Simpler and Faster Alternative Softmax Transformation Shaoshi Sun, Zhenyuan Zhang, BoCheng Huang, Pengbin Lei, Jianlin Su, Shengfeng Pan, Jiarun Cao  
<https://arxiv.org/abs/2112.12433>

Clean or Annotate: How to Spend a Limited Data Collection Budget Derek Chen, Zhou Yu, Samuel R. Bowman  
<https://arxiv.org/abs/2110.08355>

How many labelers do you have? A closer look at gold-standard labels Chen Cheng, Hilal Asi, John Duchi  
<https://arxiv.org/abs/2206.12041>

Eliciting and Learning with Soft Labels from Every Annotator  
<https://arxiv.org/abs/2207.00810>

Quantified Reproducibility Assessment of NLP Results Anya Belz, Maja Popović, Simon Mille <https://arxiv.org/abs/2204.05961>

SHAP and LIME Python Libraries: Part 2 - Using SHAP and LIME <https://www.dominodatalab.com/blog/shap-lime-python-libraries-part-2-using-shap-lime>

Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations Sander Greenland, corresponding author Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/>

The Mythos of Model Interpretability <https://arxiv.org/pdf/1606.03490v1.pdf>

Lessons Learned Reproducing a Deep Reinforcement Learning Paper Apr 6, 2018 <http://amid.fish/reproducing-deep-rl>

Spatial autocorrelation: bane or bonus? View ORCID ProfileMatt. D. M. Pawley, Brian H. McArdle doi: <https://doi.org/10.1101/385526> <https://www.biorxiv.org/content/10.1101/385526v1>

On Reality and the Limits of Language Data Nigel H. Collier, Fangyu Liu, Ehsan Shareghi <https://arxiv.org/abs/2208.11981>

Open Information Extraction from 2007 to 2022 – A Survey Pai Liu, Wenyang Gao, Wenjie Dong, Songfang Huang, Yue Zhang <https://arxiv.org/abs/2208.08690>

Colah's blog <http://colah.github.io/>

Causal Reasoning: Fundamentals and Machine Learning Applications <http://causalinference.gitlab.io/book/>

<http://courses.d2l.ai/berkeley-stat-157/units/index.html#>

A Compendium of Clean Graphs in R [http://shinyapps.org/apps/RGraphCompendium/index.php?utm\\_content=](http://shinyapps.org/apps/RGraphCompendium/index.php?utm_content=)

Bayesian Inference an interactive visualization [https://rpsychologist.com/d3/bayes/?utm\\_content=buffera5352&u](https://rpsychologist.com/d3/bayes/?utm_content=buffera5352&u)

Fitting distributions with R <https://www.magesblog.com/post/2011-12-01-fitting-distributions-with-r/>

Concerns About Bots on Mechanical Turk: Problems and Solutions <https://www.cloudresearch.com/resources/blog/concerns-about-bots-on-mechanical-turk-problems-and-solutions/>

Reproducible Research with R & RStudio 2nd Edition Christopher Gandrud <http://christophergandrud.github.io/RepResR-RStudio/>

Regression and Causality <https://arxiv.org/pdf/2006.11754.pdf>

Introduction to Causal Inference Fall 2020 <https://www.bradyneal.com/causal-inference-course>

Tensorflow 2.0 Pitfalls A list of commonly seen issues along with solutions. [http://blog.ai.ovgu.de/posts/jens/2019/001\\_tf20\\_pitfalls/index.html](http://blog.ai.ovgu.de/posts/jens/2019/001_tf20_pitfalls/index.html)

Cold Case: The Lost MNIST Digits Chhavi Yadav, Léon Bottou <https://arxiv.org/abs/1905.10498>

Automated Text Classification of News Articles: A Practical Guide Published online by Cambridge University Press: 09 June 2020 <https://www.cambridge.org/core/journals/political-analysis/article/abs/automated-text-classification-of-news-articles-a-practical-guide/10462DB284B1CD80C0FAE796AD786BC6>

How to Use t-SNE Effectively <https://distill.pub/2016/misread-tsne/>

Locality Sensitive Hashing in R <http://dsnotes.com/post/locality-sensitive-hashing-in-r-part-1/>

Identification of and Correction for Publication Bias Isaiah Andrews <https://www.aeaweb.org/articles?id=10.1257/aer.20180310>

Mediation Analysis is Counterintuitively Invalid <http://datacolada.org/103>

Dive into Deep Learning <http://d2l.ai/>

CS224d: Deep Learning for Natural Language Processing <http://cs224d.stanford.edu/syllabus.html>

Regression Models for Count Data: beyond the Poisson model <http://cursos.leg.ufpr.br/rmcd/>

p-hacking fast and slow: Evaluating a forthcoming AER paper deeming some econ literatures less trustworthy <http://datacolada.org/91>

Attention Is All You Need Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin <https://arxiv.org/abs/1706.03762>

When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? Kosuke Imai Harvard University In Song Kim <https://imai.fas.harvard.edu/research/files/FEmatch.pdf>

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? <https://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf>

P values in display items are ubiquitous and almost invariably significant: A survey of top science journals <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197440>

R Workflow for Reproducible Data Analysis and Reporting <http://hbiostat.org/rflow/>

The reusable holdout: Preserving validity in adaptive data analysis <https://ai.googleblog.com/2015/08/the-reusable-holdout-preserving.html>

The science that's never been cited Nature investigates how many papers really end up without a single citation. [https://www.nature.com/articles/d41586-017-08404-0?WT.mc\\_id=TWT\\_NA\\_1711\\_FHNEWSFNEVERCITED\\_PORTFOLIO](https://www.nature.com/articles/d41586-017-08404-0?WT.mc_id=TWT_NA_1711_FHNEWSFNEVERCITED_PORTFOLIO)

didimputation The goal of didimputation is to estimate TWFE models without running into the problem of staggered treatment adoption. <https://github.com/kylebutts/didimputation>

Methods Matter: P-Hacking and Causal Inference in Economics <https://docs.iza.org/dp11796.pdf>

CloudForest <https://github.com/ryanbressler/CloudForest>

The idea for Artificial Contrasts is based on: Eugene Tuvand and Kari Torkkola's "Feature Filtering with Ensembles Using Artificial Contrasts" <http://enpub.fulton.asu.edu/workshop/FSDM05-Proceedings.pdf#page=74> and Eugene Tuv, Alexander Borisov, George Runger and Kari Torkkola's "Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination" [http://www.researchgate.net/publication/220320233\\_Feature\\_Selection\\_with\\_Ensembles\\_Artificial\\_Variables](http://www.researchgate.net/publication/220320233_Feature_Selection_with_Ensembles_Artificial_Variables)

The idea for growing trees to minimize categorical entropy comes from Ross Quinlan's ID3: [http://en.wikipedia.org/wiki/ID3\\_algorithm](http://en.wikipedia.org/wiki/ID3_algorithm)

“The Elements of Statistical Learning” 2nd edition by Trevor Hastie, Robert Tibshirani and Jerome Friedman was also consulted during development.

Methods for classification from unbalanced data are covered in several papers: <http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf> <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163175/> <http://www.biomedcentral.com/1471-2105/11/523> <http://bib.oxfordjournals.org/content/early/2012/03/08/bib.b> <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067863>

Denisty Estimating Trees/Forests are Discussed: <http://users.cis.fiu.edu/~lzhen001/activities/KDD2011Program/> [http://research.microsoft.com/pubs/158806/CriminisiForests\\_FoundTrends\\_2011.pdf](http://research.microsoft.com/pubs/158806/CriminisiForests_FoundTrends_2011.pdf)  
The later also introduces the idea of manifold forests which can be learned using down stream analysis of the outputs of leafcount to find the Fiedler vectors of the graph laplacian.

An introduction to Git and how to use it with RStudio <http://r-bio.github.io/intro-git-rstudio/>

Probability and Statistics Cookbook <https://pages.cs.wisc.edu/~tdw/files/cookbook-en.pdf>

The Plain Person’s Guide to Plain Text Social Science <https://plain-text.co/>

Causal Graphical Views of Fixed Effects and Random Effects Models <https://psyarxiv.com/cxd2n/>

Beware Default Random Forest Importances <https://explained.ai/rf-importance/index.html>

Tools and guides to put R models into production <https://putrinprod.com/>

’Metrics Monday: You Can’t Compare OLS with 2SLS PUBLISHED NOVEMBER 20, 2017 <http://marcfbellemare.com/wordpress/12723>

Causal Inference Animated Plots <https://nickchk.com/causalgraphs.html#iv>

Scaling Data from Multiple Sources [https://www.cambridge.org/core/journals/political-analysis/article/abs/scaling-data-from-multiple-sources/1F9D30D8DDCE44379E8B962C29DADBAB?utm\\_source](https://www.cambridge.org/core/journals/political-analysis/article/abs/scaling-data-from-multiple-sources/1F9D30D8DDCE44379E8B962C29DADBAB?utm_source)

GAM: The Predictive Modeling Silver Bullet <https://multithreaded.stitchfix.com/blog/2015/07/30/gam/>

Generalized Full Matching Published online by Cambridge University Press: 23 November 2020 [https://www.cambridge.org/core/journals/political-analysis/article/abs/generalized-full-matching/3DA71D8BEDA6F02B5D36457E114C79B6?utm\\_source=hootsuite](https://www.cambridge.org/core/journals/political-analysis/article/abs/generalized-full-matching/3DA71D8BEDA6F02B5D36457E114C79B6?utm_source=hootsuite)

A Deep Dive Into How R Fits a Linear Model <http://madrury.github.io/jekyll/update/statistics/2016/07/20/lm-in-R.html>

A ModernDive into R and the Tidyverse <https://moderndive.com/>

INSTRUMENTAL VARIABLES REGRESSIONS INVOLVING SEASONAL DATA David E.A. GILES [http://web.uvic.ca/~dgiles/blog/Giles\\_FWL.pdf](http://web.uvic.ca/~dgiles/blog/Giles_FWL.pdf)

The Book of Statistical Proofs <https://statproofbook.github.io/>

An econometric method for estimating population parameters from non-random samples: An application to clinical case finding [http://www-personal.umich.edu/~zmclaren/mclaren\\_tbprevalence.pdf](http://www-personal.umich.edu/~zmclaren/mclaren_tbprevalence.pdf)

Parallelizing neural networks on one GPU with JAX <http://willwhitney.com/parallel-training-jax.html>

<https://wrdrd.github.io/docs/>

Learning interactions via hierarchical group-lasso regularization Michael Lim\* and Trevor Hastie\* June 21, 2014 [https://hastie.su.domains/Papers/glinternet\\_jcgs.pdf](https://hastie.su.domains/Papers/glinternet_jcgs.pdf)

On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data <http://web.mit.edu/insong/www/pdf/FEmatch-twoway.pdf>

Backprop is not just the chain rule AUG 18, 2017 <http://timvieira.github.io/blog/post/2017/08/18/backprop-is-not-just-the-chain-rule/>

HOW TO PLOT XGBOOST TREES IN R <http://theautomatic.net/2021/04/28/how-to-plot-xgboost-trees-in-r/>

Rectangling <https://tidyr.tidyverse.org/articles/rectangle.html>

R Packages (2e) <https://r-pkgs.org/>

Prior distributions for variance parameters in hierarchical models <http://www.stat.columbia.edu/~gelman/research/published/taumain.pdf>

A visual introduction to machine learning <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

Prior distribution Andrew Gelman Volume 3, pp 1634–1637 [http://www.stat.columbia.edu/~gelman/research/published/p039\\_o.pdf](http://www.stat.columbia.edu/~gelman/research/published/p039_o.pdf)

P-curve.com <http://www.p-curve.com/>



Model Tuning and the Bias-Variance Tradeoff <http://www.r2d3.us/visual-intro-to-machine-learning-part-2/>

How (and why) to create a good validation set <https://www.fast.ai/posts/2017-11-13-validation-sets.html>

The Promise and Pitfalls of Differences-in-Differences: Reflections on ‘16 and Pregnant’ and Other Applications <https://www.nber.org/papers/w24857>

Applied Bayesian Modeling <http://www.leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:cogdon.pdf>

Gaussian Distributions are Soap Bubbles <https://www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/>

Interpreting Instrumented Difference-in-Differences <http://www.mit.edu/~liebers/DDIV.pdf>

Probability, log-odds, and odds [https://www.montana.edu/rotella/documents/502/Prob\\_odds\\_log-odds.pdf](https://www.montana.edu/rotella/documents/502/Prob_odds_log-odds.pdf)

TRANSFORMERS FROM SCRATCH <https://peterbloem.nl/blog/transformers>

How to Examine External Validity Within an Experiment <https://www.nber.org/papers/w24834>

Program Evaluation <https://www.lecy.info/program-evaluation/>

Facing Imbalanced Data Recommendations for the Use of Performance Metrics [https://sites.pitt.edu/~jeffcohn/biblio/Jeni\\_Metrics.pdf](https://sites.pitt.edu/~jeffcohn/biblio/Jeni_Metrics.pdf)

The Art and Practice of Economics Research: Lessons from Leading Minds <https://static1.squarespace.com/static/56ec62678a65e20b89da5f33/t/6164758b00bbcb015c12dd53>

Statistics: P values are just the tip of the iceberg <https://www.nature.com/articles/520612a>

Random Forests, Decision Trees, and Categorical Predictors: The “Absent Levels” Problem <https://www.jmlr.org/papers/volume19/16-474/16-474.pdf>

Can transparency undermine peer review? A simulation model of scientist behavior under open peer review Federico Bianchi, Flaminio Squazzoni <https://academic.oup.com/spp/article/49/5/791/6602348?login=false>

The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation <https://aclanthology.org/2021.emnlp-main.97/>

Channeling Fisher: Randomization Tests and the Statistical In-significance of Seemingly Significant Experimental Results Get access Arrow Alwyn Young <https://academic.oup.com/qje/article-abstract/134/2/557/5195544?redirectedFrom=fulltext>

Advanced R <https://adv-r.hadley.nz/index.html>

Causal Machine Learning: A Survey and Open Problems <https://ai.papers.bar/paper/460ac86ef8e611ecb9b9d35608ee6155>

On the Meaning of Within-Factor Correlated Measurement Er-rors <https://academic.oup.com/jcr/article-abstract/11/1/572/1822756>

Trustworthy Machine Learning <http://www.trustworthymachinelearning.com/>

Statistical Modeling: The Two Cultures Author(s): Leo Breiman <http://www2.math.uu.se/~thulin/mm/breiman.pdf>

Estimating misclassification error with small samples via boot-strap cross-validation <https://academic.oup.com/bioinformatics/article/21/9/1979/409121?login=true>

Critical appraisal of artificial intelligence-based prediction mod-els for cardiovascular disease <https://academic.oup.com/eurheartj/article/43/31/2921/6593474?login=false>

The Only Probability Cheatsheet You'll Ever Need <http://www.wzchen.com/probability-cheatsheet/>

Come back and scrape these <http://www.wzchen.com/data-science-books>

Download the Datasaurus: Never trust summary statistics alone; always visualize your data <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>

Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis [https://abhsarma.github.io/pubs/Prior\\_Setting\\_CHI2020.pdf](https://abhsarma.github.io/pubs/Prior_Setting_CHI2020.pdf)

50 Years of Data Science David Donoho <https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734>

The Assessment of Intrinsic Credibility and a New Argument for  $p < 0.005$  Leonhard Held <https://arxiv.org/abs/1803.10052>

Arbitrariness of peer review: A Bayesian analysis of the NIPS experiment Olivier Francois <https://arxiv.org/abs/1507.06411>

Diaries of Social Data Research <https://anchor.fm/diaries-soc-data-research/episodes/The-Evolution-of-Computational-Social-Science-from-a-Sociology-Perspective-with-Chris-Baile17vikf>

pipecleaner <https://alistaire47.github.io/pipecleaner/>

Cross-Validation for Correlated Data <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2020.1801451>

The causal hype ratchet <https://statmodeling.stat.columbia.edu/2018/12/21/causal-hype-ratchet/>

A Permutation Test for the Regression Kink Design Peter Ganong & Simon Jäger <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2017.1328356#.XEx7z89KjXF>

Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions <https://arxiv.org/abs/1604.07125>

High-Dimensional Convex Geometry [https://amitrajaraman.github.io/notes/convex-geometry/Convex\\_Geometry.pdf](https://amitrajaraman.github.io/notes/convex-geometry/Convex_Geometry.pdf)

Data Science Interviews During the 2020 Pandemic <https://alexgude.com/blog/interviewing-for-data-science-positions-in-2020/>

Tech Interviews: Respect Everyone's Time <https://alexgude.com/blog/interviews-respect-time/>

Distribution-Free Prediction Intervals with Conformal Inference using R <https://arelbundock.com/posts/conformal/>

Robustness checks <https://statmodeling.stat.columbia.edu/2018/11/14/robustness-checks-joke/> <https://statmodeling.stat.columbia.edu/2017/11/29/what-point-robustness-check/>

Synthetically generated text for supervised text analysis Andrew Halterman [https://andrewhalterman.com/files/Halterman\\_synthetic\\_text.pdf](https://andrewhalterman.com/files/Halterman_synthetic_text.pdf)

EPP: interpretable score of model predictive power Alicja Gosiewska, Mateusz Bakala, Katarzyna Woznica, Maciej Zwolinski, Przemyslaw Biecek <https://arxiv.org/abs/1908.09213>

Measuring Calibration in Deep Learning Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, Dustin Tran <https://arxiv.org/abs/1904.01685>

What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems Oliver J. Maclaren, Ru-anui Nicholson <https://arxiv.org/abs/1904.02826>

Comparing Spike and Slab Priors for Bayesian Variable Selection Gertraud Malsiner-Walli, Helga Wagner <https://arxiv.org/abs/1812.07259>

Time-uniform, nonparametric, nonasymptotic confidence sequences Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, Jasjeet Sekhon <https://arxiv.org/abs/1810.08240>

Open Science in Software Engineering Daniel Méndez Fernández, Daniel Graziotin, Stefan Wagner, Heidi Seibold

Safe Testing We develop the theory of hypothesis testing based on the E-value, a notion of evidence that, unlike the p-v <https://arxiv.org/abs/1906.07801>

A Mini-Introduction To Information Theory Edward Witten <https://arxiv.org/abs/1805.11965>

An Introduction to Deep Reinforcement Learning Vincent Francois-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, Joelle Pineau <https://arxiv.org/abs/1811.12560>

On the cross-validation bias due to unsupervised pre-processing Amit Moscovich, Saharon Rosset <https://arxiv.org/abs/1901.08974>

Troubling Trends in Machine Learning Scholarship Zachary C. Lipton, Jacob Steinhardt <https://arxiv.org/abs/1807.03341>

The Role of the Propensity Score in Fixed Effect Models Dmitry Arkhangelsky, Guido Imbens <https://arxiv.org/abs/1807.02099>

Proxy Controls and Panel Data Ben Deaner <https://arxiv.org/abs/1810.00283>

Structural Breaks in Time Series Alessandro Casini, Pierre Perron <https://arxiv.org/abs/1805.03807>

Comparing interpretability and explainability for feature selection Jack Dunn, Luca Mingardi, Ying Daisy Zhuo <https://arxiv.org/abs/2105.05328>

Cross-validation: what does it estimate and how well does it do it? Stephen Bates, Trevor Hastie, Robert Tibshirani <https://arxiv.org/abs/2104.00673>

On the implied weights of linear regression for causal inference Ambarish Chattopadhyay, Jose R. Zubizarreta <https://arxiv.org/abs/2104.06581>

A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification Anastasios N. Angelopoulos, Stephen Bates <https://arxiv.org/abs/2107.07511>

Out-of-distribution Generalization in the Presence of Nuisance-Induced Spurious Correlations Aahlad Puli, Lily H. Zhang, Eric K. Oermann, Rajesh Ranganath <https://arxiv.org/abs/2107.00520>

A Tutorial on VAEs: From Bayes' Rule to Lossless Compression Ronald Yu <https://arxiv.org/abs/2006.10273>

Common Limitations of Image Processing Metrics: A Picture Story <https://arxiv.org/abs/2104.05642>

On the Inductive Bias of Masked Language Modeling: From Statistical to Syntactic Dependencies Tianyi Zhang, Tatsunori Hashimoto <https://arxiv.org/abs/2104.05694>

A large-scale study on research code quality and execution Ana Trisovic, Matthew K. Lau, Thomas Pasquier, Mercè Crosas <https://arxiv.org/abs/2103.12793>

Explaining by Removing: A Unified Framework for Model Explanation Ian Covert, Scott Lundberg, Su-In Lee <https://arxiv.org/abs/2011.14878>

What is Entropy? A new perspective from games of chance Sarah Brandsen, Isabelle Jianing Geng, Gilad Gour <https://arxiv.org/abs/2103.08681>

Instrumental variables, spatial confounding and interference Andrew Giffin, Brian J. Reich, Shu Yang, Ana G. Rappold <https://arxiv.org/abs/2103.00304>

On Linear Identifiability of Learned Representations Geoffrey Roeder, Luke Metz, Diederik P. Kingma <https://arxiv.org/abs/2007.00810>

Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, Yejin Choi <https://arxiv.org/abs/2009.10795>

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks Curtis G. Northcutt, Anish Athalye, Jonas Mueller <https://arxiv.org/abs/2103.14749>

Contamination Bias in Linear Regressions Paul Goldsmith-Pinkham, Peter Hull, Michal Kolesár

Towards optimal doubly robust estimation of heterogeneous causal effects Edward H. Kennedy <https://arxiv.org/abs/2004.14497>

When are Non-Parametric Methods Robust? Robi Bhattacharjee, Kamalika Chaudhuri <https://arxiv.org/abs/2003.06121>

When Is Parallel Trends Sensitive to Functional Form? Jonathan Roth, Pedro H. C. Sant'Anna <https://arxiv.org/abs/2010.04814>

Optimal Regularization Can Mitigate Double Descent Preetum Nakkiran, Prayaag Venkat, Sham Kakade, Tengyu Ma

Valid Causal Inference with (Some) Invalid Instruments Jason Hartford, Victor Veitch, Dhanya Sridhar, Kevin Leyton-Brown <https://arxiv.org/abs/2006.11386>

A Survey on Knowledge Graphs: Representation, Acquisition and Applications Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, Philip S. Yu <https://arxiv.org/abs/2002.00388>

The MCC-F1 curve: a performance evaluation technique for binary classification Chang Cao, Davide Chicco, Michael M. Hoffman <https://arxiv.org/abs/2006.11278>

Causal Inference and Data Fusion in Econometrics Paul Hünermund (Copenhagen Business School), Elias Bareinboim (Columbia University) <https://arxiv.org/abs/1912.09104>

Learning to Induce Causal Structure Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Anirudh Goyal, Jorg Bornschein, Melanie Rey, Theophane Weber, Matthew Botvinic, Michael Mozer, Danilo Jimenez Rezende <https://arxiv.org/abs/2204.04875>

Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data Spencer Frei, Niladri S. Chatterji, Peter L. Bartlett <https://arxiv.org/abs/2202.05928>

Causal Inference Through the Structural Causal Marginal Problem Luigi Gresele, Julius von Kügelgen, Jonas M. Kübler,

Elke Kirschbaum, Bernhard Schölkopf, Dominik Janzing  
<https://arxiv.org/abs/2202.01300>

Benefits and costs of matching prior to a Difference in Differences analysis when parallel trends does not hold Dae Woong Ham, Luke Miratrix <https://arxiv.org/abs/2205.08644>

Causal influence, causal effects, and path analysis in the presence of intermediate confounding Iván Díaz  
<https://arxiv.org/abs/2205.08000>

Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra <https://arxiv.org/abs/2201.02177>

Deep Learning Interviews: Hundreds of fully solved job interview questions from a wide range of key topics in AI Shlomo Kashani, Amir Ivry <https://arxiv.org/abs/2201.00650>

Better Uncertainty Calibration via Proper Scores for Classification and Beyond Sebastian Gruber, Florian Buettner  
<https://arxiv.org/abs/2203.07835>

Sensitivity Analysis of Individual Treatment Effects: A Robust Conformal Inference Approach Ying Jin, Zhimei Ren, Emmanuel J. Candès <https://arxiv.org/abs/2111.12161>

Towards a Unified Information-Theoretic Framework for Generalization Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, Daniel M. Roy <https://arxiv.org/abs/2111.05275>

Learning in High Dimension Always Amounts to Extrapolation Randall Balestriero, Jerome Pesenti, Yann LeCun  
<https://arxiv.org/abs/2110.09485>

Understanding Dataset Difficulty with V-Usable Information Kawin Ethayarajh, Yejin Choi, Swabha Swayamdipta  
<https://arxiv.org/abs/2110.08420>

Batch Normalization Explained Randall Balestriero, Richard G. Baraniuk <https://arxiv.org/abs/2209.14778>

Bayesian Online Changepoint Detection <https://arxiv.org/pdf/0710.3742.pdf>

Impact of subsampling and pruning on random forests.  
<https://arxiv.org/pdf/1603.04261.pdf>

Selection Collider Bias in Large Language Models Emily McMilin <https://arxiv.org/abs/2208.10063>

On the Factory Floor: ML Engineering for Industrial-Scale Ads Recommendation Models Rohan Anil, Sandra Gadanho, Da Huang, Nijith Jacob, Zhuoshu Li, Dong Lin, Todd Phillips, Cristina Pop, Kevin Regan, Gil I. Shamir, Rakesh Shivanna, Qiqi Yan <https://arxiv.org/abs/2209.05310>

Selective review of offline change point detection methods <https://arxiv.org/pdf/1801.00718.pdf>

How Much More Data Do I Need? Estimating Requirements for Downstream Tasks Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, Marc T. Law <https://arxiv.org/abs/2207.01725>

Snorkel: Rapid Training Data Creation with Weak Supervision <https://arxiv.org/pdf/1711.10160.pdf>

Defining and Characterizing Reward Hacking Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krashennnikov, David Krueger <https://arxiv.org/abs/2209.13085>

On Leave-One-Out Conditional Mutual Information For Generalization Mohamad Rida Rammal, Alessandro Achille, Aditya Golatkar, Suhas Diggavi, Stefano Soatto <https://arxiv.org/abs/2207.00581>

Formal Algorithms for Transformers Mary Phuong, Marcus Hutter <https://arxiv.org/abs/2207.09238>

Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, Armen Aghajanyan <https://arxiv.org/abs/2205.10770>

Why do tree-based models still outperform deep learning on tabular data? Léo Grinsztajn (SODA), Edouard Oyallon (ISIR, CNRS), Gaël Varoquaux (SODA) <https://arxiv.org/abs/2207.08815>

Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, Preetum Nakkiran <https://arxiv.org/abs/2207.06569>



Towards Understanding Grokking: An Effective Theory of Representation Learning Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, Mike Williams <https://arxiv.org/abs/2205.10343>

Pen and Paper Exercises in Machine Learning Michael U. Gutmann <https://arxiv.org/abs/2206.13446>

Introduction to DiD with Multiple Time Periods Brantly Callaway and Pedro H.C. Sant’Anna 2022-07-19 <https://bcallaway11.github.io/did/articles/multi-period-did.html>

Applications of Deep Neural Networks with Keras Jeff Heaton Fall 2022.0 <https://arxiv.org/pdf/2009.05673.pdf>

Joint Distributions for TensorFlow Probability DAN PIPONI†, DAVE MOORE† & JOSHUA V. DILLON, Google Research <https://arxiv.org/pdf/2001.11819.pdf>

Descending through a Crowded Valley — Benchmarking Deep Learning Optimizers <https://arxiv.org/pdf/2007.01547.pdf>

Geographic Difference-in-Discontinuities Kyle Butts <https://arxiv.org/pdf/2109.07406.pdf>

Pre-trained Models for Natural Language Processing: A Survey Xipeng Qiu\*, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai & Xuanjing Huang <https://arxiv.org/pdf/2003.08271.pdf>

Knowledge Graphs on the Web – an Overview <https://arxiv.org/pdf/2003.00719.pdf>

Knowledge Graphs <https://arxiv.org/pdf/2003.02320.pdf>

TOPOLOGY OF DEEP NEURAL NETWORKS GREGORY NAITZAT, ANDREY ZHITNIKOV, AND LEK-HENG LIM <https://arxiv.org/pdf/2004.06093.pdf>

Noise-Induced Randomization in Regression Discontinuity Designs <https://arxiv.org/pdf/2004.09458.pdf>

Markov Chain Monte Carlo Methods, a survey with some frequent misunderstandings <https://arxiv.org/pdf/2001.06249.pdf>

Learning Dependency Structures for Weak Supervision Models <https://arxiv.org/pdf/1903.05844.pdf>

Approximate leave-future-out cross-validation for Bayesian time series models <https://arxiv.org/pdf/1902.06281.pdf>

Relational Representation Learning for Dynamic (Knowledge) Graphs: A Survey <https://arxiv.org/pdf/1905.11485v1.pdf>

Statistical methods research done as science rather than mathematics James S. Hodges <https://arxiv.org/pdf/1905.08381.pdf>

R Tip: use `isTRUE()` <https://win-vector.com/2018/06/11/r-tip-use-istrue/>

The tidymodels Package <https://www.tidyverse.org/blog/2018/08/tidymodels-0-0-1/>

Regular expressions are tricky. RegExplain makes it easier to see what you're doing. <https://www.garrickadenbuie.com/project/regexplain/>

The ability of different peer review procedures to flag problematic publications <https://link.springer.com/article/10.1007/s11192-018-2969-2>

gglabeller [https://github.com/AliciaSchep/gglabeller?utm\\_content=buffer552f9&utm\\_medium=social&utm\\_source=twitter](https://github.com/AliciaSchep/gglabeller?utm_content=buffer552f9&utm_medium=social&utm_source=twitter)

The `{targets}` R package user manual <https://books.ropensci.org/targets/>

How Regularization Works <https://e2eml.school/regularization.html>

Don't be tricked by the Hashing Trick <https://booking.ai/dont-be-tricked-by-the-hashing-trick-192a6aae3087>

How to Use Catboost with Tidymodels <https://blog.rmhogervorst.nl/blog/2020/08/28/how-to-use-catboost-with-tidymodels/>

R Markdown: The Definitive Guide <https://bookdown.org/yihui/rmarkdown/>

37 Reasons why your Neural Network is not working <https://blog.slavv.com/37-reasons-why-your-neural-network-is-not-working-4020854bd607>

Time to assume that health research is fraudulent until proven otherwise? <https://blogs.bmj.com/bmj/2021/07/05/time-to-assume-that-health-research-is-fraudulent-until-proved-otherwise/>

Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review Rida Shahzad<sup>1</sup>, Bushra Ayub<sup>2</sup>, <http://orcid.org/0000-0001-5100-3189> M A Rehman Siddiqui<sup>3</sup> <https://bmjopen.bmj.com/content/12/9/e061519.abstract>

Running R Scripts on a Schedule with GitHub Actions By Simon P. Couch DECEMBER 27, 2020 <https://www.simonpcouch.com/blog/r-github-actions-commit/>

Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction Shangzhi Hong & Henry S. Lynn <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01080-1>

spatialRF: Easy Spatial Regression with Random Forest <https://blasbenito.github.io/spatialRF/>

Supervised Clustering: How to Use SHAP Values for Better Cluster Analysis <https://www.aidancooper.co.uk/supervised-clustering-shap-values/>

Exploring Neural Networks Visually in the Browser <https://cprimozic.net/blog/neural-network-experiments-and-visualizations/>

How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on Over 60 Replicated Studies\* [https://yiqingxu.org/papers/english/2021\\_iv/LLXZ.pdf](https://yiqingxu.org/papers/english/2021_iv/LLXZ.pdf)

Feathr: LinkedIn's feature store is now available on Azure Posted on April 12, 2022 Xiaoyong Zhu <https://azure.microsoft.com/en-us/blog/feathr-linkedin-s-feature-store-is-now-available-on-azure/>

A Survey of Learning on Small Data Xiaofeng Cao, Weixin Bu, Shengjun Huang, Yingpeng Tang, Yaming Guo, Yi Chang, Ivor W. Tsang <https://arxiv.org/abs/2207.14443>

Ontology-based industrial data management platform Sergey Gorshkov, Alexander Grebeshkov, Roman Shebalov <https://arxiv.org/abs/2103.05538>

How to Speed Up XGBoost Model Training <https://www.anyscale.com/blog/how-to-speed-up-xgboost-model-training>

Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy <https://www.datasciencecentral.com/markov-chain-monte-carlo-methods-for-bayesian-data-analysis-in/#w6JI5>

How to Easily Draw Neural Network Architecture Diagrams  
<https://towardsdatascience.com/how-to-easily-draw-neural-network-architecture-diagrams-a6b6138ed875>

L2 Regularization and Batch Norm <https://blog.janestreet.com/l2-regularization-and-batch-norm/>

Trust in LIME: Yes, No, Maybe So? <https://www.dominodatalab.com/blog/trust-in-lime-local-interpretable-model-agnostic-explanations>

Inside Manifold: Uber's Stack for Debugging Machine Learning Models [https://towardsai.net/p/l/inside-manifold-ubers-stack-for-debugging-machine-learning-models?utm\\_source=twitter&utm\\_medium=social&utm\\_campaign=rop-content-recycle](https://towardsai.net/p/l/inside-manifold-ubers-stack-for-debugging-machine-learning-models?utm_source=twitter&utm_medium=social&utm_campaign=rop-content-recycle)

Data validation for machine learning JUNE 5, 2019 ~ ADRIAN COLYER <https://blog.acolyer.org/2019/06/05/data-validation-for-machine-learning/>

Multiprocessing vs. Threading in Python: What Every Data Scientist Needs to Know <https://blog.floydhub.com/multiprocessing-vs-threading-in-python-what-every-data-scientist-needs-to-know/>

A Comprehensive Guide to Machine Learning <https://www.eecs189.org/static/resources/comprehensive-guide.pdf>

A Concrete Introduction to Probability (using Python) <https://github.com/norvig/pytudes/blob/main/ipynb/Probability.ipynb>

An Interactive Guide To The Fourier Transform <https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/>

Identity Crisis [https://betanalpha.github.io/assets/case\\_studies/identifiability.html](https://betanalpha.github.io/assets/case_studies/identifiability.html)  
<https://betanalpha.github.io/writing/>

Bayes Sparse Regression Michael Betancourt March 2018  
[https://betanalpha.github.io/assets/case\\_studies/bayes\\_sparse\\_regression.html#1\\_fading\\_into\\_irrelevance](https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html#1_fading_into_irrelevance)

An Introduction to Stan Michael Betancourt March 2020  
[https://betanalpha.github.io/assets/case\\_studies/stan\\_intro.html](https://betanalpha.github.io/assets/case_studies/stan_intro.html)  
<https://developers.google.com/machine-learning>

Prior Modeling Michael Betancourt September 2021  
[https://betanalpha.github.io/assets/case\\_studies/prior\\_modeling.html](https://betanalpha.github.io/assets/case_studies/prior_modeling.html)

Colorized Math Equations <https://betterexplained.com/articles/colorized-math-equations/>

Towards A Principled Bayesian Workflow Michael Betancourt  
April 2020 [https://betanalpha.github.io/assets/case\\_studies/principled\\_bayesian\\_workflow.html](https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html)

Ordinal Regression Michael Betancourt May 2019 [https://betanalpha.github.io/assets/case\\_studies/ordinal\\_regression.html](https://betanalpha.github.io/assets/case_studies/ordinal_regression.html)

Analysing continuous proportions in ecology and evolution: A  
practical introduction to beta and Dirichlet regression Jacob C.  
Douma, James T. Weedon <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13234>

slider <https://davisvaughan.github.io/slider/index.html>

torch.manual\_seed(3407) is all you need: On the influence of  
random seeds in deep learning architecture for computer vision  
[https://davidpicard.github.io/pdf/lucky\\_seed.pdf](https://davidpicard.github.io/pdf/lucky_seed.pdf)

A Day in the Life of a Silicon Valley Data Engineer  
<https://towardsdatascience.com/a-day-in-the-life-of-a-google-data-engineer-722f1b2206cc>

ICML 2018 Notes [https://david-abel.github.io/blog/posts/misc/icml\\_2018.pdf](https://david-abel.github.io/blog/posts/misc/icml_2018.pdf)

ICML 2019 Notes [https://david-abel.github.io/notes/icml\\_2019.pdf](https://david-abel.github.io/notes/icml_2019.pdf)

Keep using plate notation <https://davidrushingdewhurst.com/blog/2020-07-28keep-using-plate-notation.html>

Data Visualization <https://dataviz21.classes.andrewheiss.com/content/>

DO YOU KNOW THE 4 TYPES OF ADDITIVE VARIABLE  
IMPORTANCES? [https://datajms.com/post/variable\\_importance\\_feature\\_attribution/](https://datajms.com/post/variable_importance_feature_attribution/)

geostan: Bayesian spatial analysis <https://connordonegan.github.io/geostan/>

Using Observational Study Data as an External Control  
Group for a Clinical Trial: an Empirical Comparison  
of Methods to Account for Longitudinal Missing Data  
[https://www.researchgate.net/publication/357609855\\_Using\\_Observational\\_Study\\_Data\\_as\\_an\\_External\\_Control\\_Group\\_for\\_a\\_Clinical\\_Trial](https://www.researchgate.net/publication/357609855_Using_Observational_Study_Data_as_an_External_Control_Group_for_a_Clinical_Trial)

Selective Ignorability Assumptions in Causal Inference Mar-  
shall M. Joffe , Wei Peter Yang and Harold I. Feldman  
<https://www.degruyter.com/document/doi/10.2202/1557-4679.1199/html>

Expressing Regret: A Unified View of Credible Intervals Kenneth Rice  & Lingbo Ye <https://www.tandfonline.com/doi/abs/10.1080/00031305.2022.2039764>

Efficient Identification in Linear Structural Causal Models with Instrumental Cutsets <https://causalai.net/r49.pdf>

Polymatching algorithm in observational studies with multiple treatment groups <https://www.sciencedirect.com/science/article/abs/pii/S0167947321001985>

An Introduction to Statistical Learning [https://hastie.su.domains/ISLR2/ISLRv2\\_website.pdf](https://hastie.su.domains/ISLR2/ISLRv2_website.pdf)

Core concepts in pharmacoepidemiology: Confounding by indication and the role of active comparators <https://onlinelibrary.wiley.com/doi/10.1002/pds.5407>

Aim for Clinical Utility, Not Just Predictive Accuracy Sachs, Michael C.a; Sjölander, Arvidb; Gabriel, Erin E.b [https://journals.lww.com/epidem/Fulltext/2020/05000/Aim\\_for\\_Clinical\\_Utility,\\_Not\\_Just\\_Predictive.8.aspx?article-sam-container](https://journals.lww.com/epidem/Fulltext/2020/05000/Aim_for_Clinical_Utility,_Not_Just_Predictive.8.aspx?article-sam-container)

Monitoring Machine Learning Models in Production A Comprehensive Guide <https://christophergs.com/machine%20learning/2020/03/14/how-to-monitor-machine-learning-models/>

\*args and \*\*kwargs in Python <https://towardsdatascience.com/args-kwargs-python-d9c71b220970>

Waiting for Event Studies: A Play in Three Acts Sun and Abraham (2020) Explainer <https://causalinf.substack.com/p/waiting-for-event-studies-a-play>

Computational Socioeconomics <https://arxiv.org/abs/1905.06166>

Is Peer Review a Good Idea? Remco Heesen and Liam Kofi Bright <https://www.journals.uchicago.edu/doi/10.1093/bjps/axz029>

Opening the Black Box: a motivation for the assessment of mediation Danella M Hafeman 1, Sharon Schwartz <https://pubmed.ncbi.nlm.nih.gov/19261660/>

Invited Commentary: Propensity Scores Marshall M. Joffe, Paul R. Rosenbaum <https://academic.oup.com/aje/article/150/4/327/98791>

Advances in propensity score analysis Peter C Austin [peter.austin@ices.on.ca](mailto:peter.austin@ices.on.ca) View all authors and affiliations <https://journals.sagepub.com/doi/full/10.1177/0962280219899248>

Analysis in an imperfect world Michael Wallace First published:  
29 January 2020 <https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2020.01353.x>

Central Limit Theorem [http://mfviz.com/central-limit/?utm\\_content=buffer918f&utm\\_medium=social&utm\\_s](http://mfviz.com/central-limit/?utm_content=buffer918f&utm_medium=social&utm_s)

Adjusting for Covariates in Randomized Clinical Trials for  
Drugs and Biological Products Draft Guidance for Industry  
<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products>

Machine learning for improving high-dimensional proxy  
confounder adjustment in healthcare database studies: An  
overview of the current literature <https://onlinelibrary.wiley.com/doi/10.1002/pds.5500>

A Gentle Introduction to tidymodels <https://rviews.rstudio.com/2019/06/19/a-gentle-intro-to-tidymodels/>

Simultaneous Variable and Covariance Selection with the Multi-  
variate Spike-and-Slab LASSO [https://arxiv.org/pdf/1708.08911.pdf?utm\\_content=bufferb1cd5&utm\\_medium=](https://arxiv.org/pdf/1708.08911.pdf?utm_content=bufferb1cd5&utm_medium=)

Transformers Explained Visually — Not Just How, but Why  
They Work So Well <https://towardsdatascience.com/transformers-explained-visually-not-just-how-but-why-they-work-so-well-d840bd61a9d3>

Easy Bayesian Bootstrap in R [https://www.sumsar.net/blog/2015/07/easy-bayesian-bootstrap-in-r/?utm\\_content=buffer53c16&utm\\_medium=social&utm\\_source=twitter.com&utm\\_cam](https://www.sumsar.net/blog/2015/07/easy-bayesian-bootstrap-in-r/?utm_content=buffer53c16&utm_medium=social&utm_source=twitter.com&utm_cam)

Leave-future-out cross-validation for time-series models  
<https://discourse.mc-stan.org/t/leave-future-out-cross-validation-for-time-series-models/12954/2>

PCA in a tidy(verse) framework [https://tbradley1013.github.io/2018/02/01/pca-in-a-tidy-verse-framework/?utm\\_content=bufferfaf31&utm\\_medium=social&utm\\_source=twitter.com&utm\\_ca](https://tbradley1013.github.io/2018/02/01/pca-in-a-tidy-verse-framework/?utm_content=bufferfaf31&utm_medium=social&utm_source=twitter.com&utm_ca)

Visualising Residuals [https://drsimonj.svbtle.com/visualising-residuals?utm\\_content=bufferdb80e&utm\\_medium=social&utm\\_source=twitter.com&utm\\_campaign=buffer](https://drsimonj.svbtle.com/visualising-residuals?utm_content=bufferdb80e&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer)

Covariate adjustment for randomized controlled trials revisited  
Jixian Wang <https://onlinelibrary.wiley.com/doi/full/10.1002/pst.1988?campaign=wolearlyview>

STAT 545 Data wrangling, exploration, and analysis with R  
<https://stat545.com/index.html>

Topics in Econometrics: Advances in Causality and Foundations of Machine Learning <https://maxkasy.github.io/home/TopicsInEconometrics2019/>

A nontechnical explanation of the counterfactual definition of confounding Martijn J.L. Bours [https://www.jclinepi.com/article/S0895-4356\(19\)30173-8/pdf](https://www.jclinepi.com/article/S0895-4356(19)30173-8/pdf)

Discovering Reliable Correlations in Categorical Data <https://deepai.org/publication/discovering-reliable-correlations-in-categorical-data>

Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations <https://deepai.org/publication/dealing-with-disagreements-looking-beyond-the-majority-vote-in-subjective-annotations>

<https://papers.labml.ai/papers/weekly>

Generalized Principal Component Analysis <https://deepai.org/publication/generalized-principal-component-analysis>

Deeptime: a Python library for machine learning dynamical models from time series data <https://deepai.org/publication/deeptime-a-python-library-for-machine-learning-dynamical-models-from-time-series-data>

Delving into Deep Imbalanced Regression <https://deepai.org/publication/delving-into-deep-imbalanced-regression>

Causality-based Feature Selection: Methods and Evaluations <https://deepai.org/publication/causality-based-feature-selection-methods-and-evaluations>

Causal Inference Through the Structural Causal Marginal Problem 02/02/2022 by Luigi Gresele, et al. <https://deepai.org/publication/causal-inference-through-the-structural-causal-marginal-problem>

Causal Discovery from Incomplete Data: A Deep Learning Approach <https://deepai.org/publication/causal-discovery-from-incomplete-data-a-deep-learning-approach>

AutoML: A Survey of the State-of-the-Art <https://deepai.org/publication/automl-a-survey-of-the-state-of-the-art>



CatBoostLSS – An extension of CatBoost to probabilistic forecasting <https://deepai.org/publication/catboostlss-an-extension-of-catboost-to-probabilistic-forecasting>

Breiman’s “Two Cultures” Revisited and Reconciled 05/27/2020 Subhadeep, et al. <https://deepai.org/publication/breiman-s-two-cultures-revisited-and-reconciled>

Graphical Representation of Missing Data Problems Felix Thoemmes<sup>1</sup> and Karthika Mohan<sup>2</sup> [https://ftp.cs.ucla.edu/pub/stat\\_ser/r448-reprint.pdf](https://ftp.cs.ucla.edu/pub/stat_ser/r448-reprint.pdf)

Judea Pearl\* Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes [https://ftp.cs.ucla.edu/pub/stat\\_ser/r483-reprint.pdf](https://ftp.cs.ucla.edu/pub/stat_ser/r483-reprint.pdf)

Gene name errors are widespread in the scientific literature Mark Ziemann, Yotam Eren & Assam El-Osta <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7>

A hypothesis is a liability Itai Yanai & Martin Lercher <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w>

gganimate extends the grammar of graphics as implemented by ggplot2 to include the description of animation. <https://gganimate.com/>

Geocomputation with R <https://geocompr.robinlovelace.net/index.html>

ftfy: fixes text for you <https://ftfy.readthedocs.io/en/latest/>

ggforce <https://ggforce.data-imaginist.com/index.html>

Geographically based Economic data (G-Econ) <https://gecon.yale.edu/>

The packages dtw for R and dtw-python for Python provide the most complete, freely-available (GPL) implementation of Dynamic Time Warping-type (DTW) algorithms up to date. <https://dynamictimewarping.github.io/>

Preprints: An underutilized mechanism to accelerate outbreak science Michael A. Johansson ,Nicholas G. Reich,Lauren Ancel Meyers,Marc Lipsitch Published: April 3, 2018 <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002549>

Randomization does not imply unconfoundedness <https://drive.google.com/file/d/1nV8QMLxwXi-iWSqiwRN4KnMSWfoWJned/view>

Bayesian Gaussian Graphical Models <https://donaldrwilliams.github.io/BGGM/index.html>

DoWhy: Addressing Challenges in Expressing and Validating Causal Assumptions <https://drive.google.com/file/d/1i81CnMd683A788RYtEb8KSowhhPJn3Z6/view>

Plotting background data for groups with ggplot2 <https://drsimonj.svbtle.com/plotting-background-data-for-groups-with-ggplot2>

Benign Overfitting in Linear Regression 06/26/2019 <https://deepai.org/publication/benign-overfitting-in-linear-regression>

Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data <https://deepai.org/publication/amortized-causal-discovery-learning-to-infer-causal-graphs-from-time-series-data>

Accelerating Deep Learning by Focusing on the Biggest Losers 10/02/2019 <https://deepai.org/publication/accelerating-deep-learning-by-focusing-on-the-biggest-losers>

A Class of Algorithms for General Instrumental Variable Models 06/11/2020 <https://deepai.org/publication/a-class-of-algorithms-for-general-instrumental-variable-models>

A Survey of Parameters Associated with the Quality of Benchmarks in NLP 10/14/2022 <https://deepai.org/publication/a-survey-of-parameters-associated-with-the-quality-of-benchmarks-in-nlp>

A study of uncertainty quantification in overparametrized high-dimensional models 10/23/2022 <https://deepai.org/publication/a-study-of-uncertainty-quantification-in-overparametrized-high-dimensional-models>

DeclareDesign Blog The trouble with ‘controlling for blocks’ <https://declaredesign.org/blog/biased-fixed-effects.html>

A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning 06/09/2022 <https://deepai.org/publication/a-critical-review-on-the-use-and-misuse-of-differential-privacy-in-machine-learning>

A Comprehensive Survey of Image Augmentation Techniques for Deep Learning 05/03/2022 <https://deepai.org/publication/a-comprehensive-survey-of-image-augmentation-techniques-for-deep-learning>

Stance Detection: A Survey ACM Computing Surveys Volume 53 Issue 1 January 2021 <https://dl.acm.org/doi/abs/10.1145/3369026>

PyTorch: An Imperative Style, High-Performance Deep Learning Library 12/03/2019 <https://deepai.org/publication/pytorch-an-imperative-style-high-performance-deep-learning-library>

On Causally Disentangled Representations 12/10/2021 <https://deepai.org/publication/on-causally-disentangled-representations>

A Visual Exploration of Gaussian Processes <https://distill.pub/2019/visual-exploration-gaussian-processes/>

Principled Machine Learning: Practices and Tools for Efficient Collaboration <https://dev.to/robogeek/principled-machine-learning-4eho>

Rules of Machine Learning:

bookmark\_border Best Practices for ML Engineering Martin Zinkevich <https://developers.google.com/machine-learning/guides/rules-of-ml>

The Variability of Model Specification 10/06/2021 <https://deepai.org/publication/the-variability-of-model-specification>

Taxonomy of Benchmarks in Graph Representation Learning 06/15/2022 <https://deepai.org/publication/taxonomy-of-benchmarks-in-graph-representation-learning>

Recognizing Variables from their Data via Deep Embeddings of Distributions 09/11/2019 <https://deepai.org/publication/recognizing-variables-from-their-data-via-deep-embeddings-of-distributions>

Relaxed Softmax for learning from Positive and Unlabeled data 09/17/2019 <https://deepai.org/publication/relaxed-softmax-for-learning-from-positive-and-unlabeled-data>

On Quantitative Evaluations of Counterfactuals 10/30/2021 <https://deepai.org/publication/on-quantitative-evaluations-of-counterfactuals>

Learning Neural Causal Models from Unknown Interventions  
10/02/2019 <https://deepai.org/publication/learning-neural-causal-models-from-unknown-interventions>

Memorizing without overfitting: Bias, variance, and interpolation in over-parameterized models 10/26/2020  
<https://deepai.org/publication/memorizing-without-overfitting-bias-variance-and-interpolation-in-over-parameterized-models>

InceptionTime: Finding AlexNet for Time Series Classification 09/11/2019 <https://deepai.org/publication/inceptiontime-finding-alexnet-for-time-series-classification>

Learning from Positive and Unlabeled Data by Identifying the Annotation Process 03/02/2020 <https://deepai.org/publication/learning-from-positive-and-unlabeled-data-by-identifying-the-annotation-process>

Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning 12/22/2019 <https://deepai.org/publication/lessons-from-archives-strategies-for-collecting-sociocultural-data-in-machine-learning>

Identification In Missing Data Models Represented By Directed Acyclic Graphs 06/29/2019 <https://deepai.org/publication/identification-in-missing-data-models-represented-by-directed-acyclic-graphs>

rrtools: Tools for Writing Reproducible Research in R  
<https://github.com/benmarwick/rrtools>

fastshap The goal of fastshap is to provide an efficient and speedy (relative to other implementations) approach to computing approximate Shapley values which help explain the predictions from machine learning models.

Monitoring Data Quality at Scale with Statistical Modeling  
May 7, 2020 <https://www.uber.com/blog/monitoring-data-quality-at-scale/>

This standard operating procedure (SOP) document describes the default practices of the experimental research group led by Donald P. Green at Columbia University.  
<https://github.com/acoppock/Green-Lab-SOP>

ggplot2 extensions <https://exts.ggplot2.tidyverse.org/gallery/>

RemixAutoML website and reference manual <https://github.com/AdrianAntico/RemixAutoML>

Inference in Linear Regression Models with Many Covariates and Heteroscedasticity Matias D. Cattaneo, Michael Jansson, and Whitney K. Newey [https://eml.berkeley.edu/~mjansson/Publications/Cattaneo-Jansson-Newey\\_2018\\_JASA.pdf](https://eml.berkeley.edu/~mjansson/Publications/Cattaneo-Jansson-Newey_2018_JASA.pdf)

Derivation of front door adjustment without intervention on the mediator [https://figshare.com/articles/journal\\_contribution/Derivation\\_of\\_front\\_door\\_adjustment\\_without\\_intervention/Derivation\\_of\\_front\\_door\\_adjustment\\_without\\_intervention/12345678](https://figshare.com/articles/journal_contribution/Derivation_of_front_door_adjustment_without_intervention/Derivation_of_front_door_adjustment_without_intervention/12345678)

Satellite image datasets <https://eod-grss-ieee.com/dataset-search>

Point of View: How should novelty be valued in science? Barak A Cohen <https://elifesciences.org/articles/28699>

The Softmax function and its derivative <https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/>

Why software projects take longer than you think: a statistical model 2019-04-15 <https://erikbern.com/2019/04/15/why-software-projects-take-longer-than-you-think-a-statistical-model.html>

MOVE IT OR LOSE IT: INTRODUCING PSEUDO-EARTH MOVER DIVERGENCE AS A CONTEXT-SENSITIVE METRIC FOR EVALUATING AND IMPROVING FORECASTING AND PREDICTION SYSTEMS <https://events.barcelonagse.eu/live/files/2912-pemdivbarcelonapdf>

The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use [https://eprints.soton.ac.uk/434156/1/The\\_spatial\\_allocation\\_of\\_population.pdf](https://eprints.soton.ac.uk/434156/1/The_spatial_allocation_of_population.pdf)

<https://www.youtube.com/playlist?list=PL8PYTP1V4I8D0UkqW2fEhgLrnlDW9QK7z>

Robust misinterpretation of confidence intervals Rink Hoekstra & Richard D. Morey & Jeffrey N. Rouder & Eric-Jan Wagenmakers <https://ejwagenmakers.com/inpress/HoekstraEtAlPBR.pdf>

Transformers from Scratch Brandon Rohrer <https://e2eml.school/transformers.html>

Evidence on Research Transparency in Economics Edward Miguel <https://www.aeaweb.org/articles?id=10.1257/jep.35.3.193>

The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions  
<https://journals.sagepub.com/doi/10.1177/0049124118799376>

The Generalizability of Survey Experiments\* Published online by Cambridge University Press: 12 January 2016  
<https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/abs/generalizability-of-survey-experiments/72D4E3DB90569AD7F2D469E9DF3A94CB>

Preregistering qualitative research Tamarinde L. HavenORCID Icon & Dr. Leonie Van Grootel  
<https://www.tandfonline.com/doi/full/10.1080/08989621.2019.1580147>

Categorical Perception of p-Values V. N. Vimal Rao, Jeffrey K. Bye, Sashank Varma  
<https://onlinelibrary.wiley.com/doi/10.1111/tops.12589>

On the Practice of Lagging Variables to Avoid Simultaneity† William Robert Reed  
<https://onlinelibrary.wiley.com/doi/10.1111/obes.12088>

Phantom Counterfactuals Tara Slough  
<https://onlinelibrary.wiley.com/doi/10.1111/ajps.12715>

“Don’t Know” Responses, Personality, and the Measurement of Political Knowledge\* Published online by Cambridge University Press: 19 June 2015  
<https://www.cambridge.org/core/journals/political-science-research-and-methods/article/abs/dont-know-responses-personality-and-the-measurement-of-political-knowledge/C28B2FF6AD8181F9F60651C0933E5620>

The influence of hidden researcher decisions in applied microeconomics  
<https://onlinelibrary.wiley.com/doi/10.1111/ecin.12992>

Inference in Experiments Conditional on Observed Imbalances in Covariates Per JohanssonORCID Icon & Mattias Nordin  
<https://www.tandfonline.com/doi/full/10.1080/00031305.2022.2054859>

Research Replication: Practical Considerations Published online by Cambridge University Press: 04 April 2018  
<https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/research-replication-practical-considerations/B744967268CDAA3F44103AA5C8539EA2>

The self-fulfilling prophecy of post-hoc power calculations Christos Christogiannis Stavros Nikolakopoulos Nikolaos Pandis Dimitris Mavridis  
[https://www.ajodo.org/article/S0889-5406\(21\)00697-1/fulltext](https://www.ajodo.org/article/S0889-5406(21)00697-1/fulltext)

Equinox is a JAX library based around a simple idea: represent parameterised functions (such as neural networks) as PyTrees. <https://docs.kidger.site/equinox/>

P-Hacking, Data Type and Data-Sharing Policy <https://docs.iza.org/dp15586.pdf>

UpSetR generates static UpSet plots. <https://github.com/hms-dbmi/UpSetR>

The purpose of the future package is to provide a very simple and uniform way of evaluating R expressions asynchronously using various resources available to the user. <https://github.com/HenrikBengtsson/future>

miceRanger: Fast Imputation with Random Forests <https://github.com/farrellday/miceRanger>

scoringRules An R package to compute scoring rules for fixed (parametric) and simulated forecast distributions. <https://github.com/FK83/scoringRules>

bayesdfa implements Bayesian Dynamic Factor Analysis (DFA) with Stan. <https://github.com/fate-ewi/bayesdfa>

dtplyr provides a data.table backend for dplyr. <https://github.com/tidyverse/dtplyr>  
performance <https://github.com/easystats/performance>

BorutaShap is a wrapper feature selection method which combines both the Boruta feature selection algorithm with shapley values. <https://github.com/Ekeany/Boruta-Shap>

D-Lab's Introduction to Machine Learning with tidymodels <https://github.com/dlab-berkeley/R-Machine-Learning>

ggVennDiagram <https://github.com/gaospecial/ggVennDiagram>

The {InteractionPoweR} package conducts power analyses for regression models in cross-sectional data sets where the term of interest is an interaction between two variables, also known as 'moderation' analyses. <https://github.com/dbaranger/InteractionPoweR>

varimpact uses causal inference statistics to generate variable importance estimates for a given dataset and outcome. <https://github.com/ck37/varImpact>

tidystringdist <https://github.com/ColinFay/tidystringdist>

<https://github.com/CenterForPeaceAndSecurityStudies/IntroductiontoMachineLearning>

Probabilistic Programming and Bayesian Methods for Hackers

Chapter 1 [https://github.com/CamDavidsonPilon/Probabilistic-](https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_TFP)

[Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1\\_Introduction/Ch1\\_Introduction\\_TFP](https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_TFP)

Papers about Causal Inference and Language <https://github.com/causaltext/causal-text-papers>

diffobj - Diffs for R Objects <https://github.com/brodieG/diffobj>

CatBoost <https://github.com/catboost/catboost>

Awesome Public Datasets <https://github.com/awesomedata/awesome-public-datasets/blob/master/README.rst>

BlackJAX is a library of samplers for JAX that works on CPU as well as GPU. <https://github.com/blackjax-devs/blackjax>

sensemakr: Sensitivity Analysis Tools for OLS <https://github.com/carloscinelli/sensemakr>

TorchArrow: a data processing library for PyTorch <https://github.com/pytorch/torcharrow>

causaleffect is a Python library for computing conditional and non-conditional causal effects. <https://github.com/pedemonte96/causaleffect>

Dynamic State Space Models in JAX. <https://github.com/probml/dynamax>

numpy-hilbert-curve <https://github.com/PrincetonLIPS/numpy-hilbert-curve>

Bayesian optimization in JAX <https://github.com/PredictiveIntelligenceLab/JAX-BO>

splines\_in\_stan.pdf [https://github.com/milkha/Splines\\_in\\_Stan/blob/master/splines\\_in\\_stan.pdf](https://github.com/milkha/Splines_in_Stan/blob/master/splines_in_stan.pdf)

This is a repository that makes an attempt to empirically take stock of the most important concepts necessary to understand cutting-edge research in neural network models for NLP. <https://github.com/neulab/nn4nlp-concepts>

The EloML package provides Elo rating system for machine learning models. Elo Predictive Power (EPP) score helps to assess model performance based Elo ranking system. <https://github.com/ModelOriented/EloML>



SPTAG (Space Partition Tree And Graph) is a library for large scale vector approximate nearest neighbor search <https://github.com/microsoft/SPTAG>

fixest: Fast and user-friendly fixed-effects estimation <https://github.com/lrberge/fixest/>

tidybayes: Bayesian analysis + tidy data + geoms <https://github.com/mjskay/tidybayes>

Milvus is an open-source vector database built to power embedding similarity search and AI applications. <https://github.com/milvus-io/milvus>

bpCausal: Bayesian Causal Inference with Time-Series Cross-Sectional Data R package for A Bayesian Alternative to the Synthetic Control Method. <https://github.com/liulch/bpCausal>

priors.pdf <https://github.com/lmbastos/Delay/blob/master/Code/priors.pdf>

cheat\_sheet-slabinterval.pdf [https://github.com/mjskay/ggdist/blob/master/figures-source/cheat\\_sheet-slabinterval.pdf](https://github.com/mjskay/ggdist/blob/master/figures-source/cheat_sheet-slabinterval.pdf)

tidypolars is a data frame library built on top of the blazingly fast polars library that gives access to methods and functions familiar to R tidyverse users. <https://github.com/markfairbanks/tidypolars>

ftfy: fixes text for you <https://github.com/rspeer/python-ftfy>

ggannotate <https://github.com/MattCowgill/ggannotate>

Conducting and Visualizing Specification Curve Analyses The goal of `specr` is to facilitate specification curve analyses (Simonsohn, Simmons & Nelson, 2019; also known as multiverse analyses, see Steegen, Tuerlinckx, Gelman & Vanpaemel, 2016). <https://github.com/masurp/specr>

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction. <https://github.com/lmcinnes/umap>

survminer: Survival Analysis and Visualization <https://github.com/kassambara/survminer>

A curated list of resources dedicated to Natural Language Processing <https://github.com/keon/awesome-nlp>

vip: Variable Importance Plots <https://github.com/koalaverse/vip/>

How To Make Your Data Analysis Notebooks More Reproducible <https://github.com/karthik/rstudio2019>

Awesome Self-Supervised Learning <https://github.com/jason718/awesome-self-supervised-learning>

dagitty Graphical Analysis of Structural Causal Models  
<https://github.com/jtextor/dagitty>

Awesome Machine Learning <https://github.com/josephmisiti/awesome-machine-learning#computer-vision-5>

Replication, Replication <https://gking.harvard.edu/files/abs/replication-abs.shtml>

Ecological Regression with Partial Identification <https://gking.harvard.edu/publications/ecological-regression-partial-identification>

biglasso: Extend Lasso Model Fitting to Big Data in R  
<https://github.com/YaohuiZeng/biglasso>

marginalEffects package for R <https://github.com/vincentarelbundock/marginalEffects>

Feature Engineering and Selection by Max Kuhn and Kjell Johnson (2019). <https://github.com/topepo/FES>

tidyposterior <https://github.com/tidymodels/tidyposterior>

R Data Science Tutorials <https://github.com/ujjwalkarn/DataScienceR>

janitor <https://github.com/sfirke/janitor>

Conformal Inference R Project Maintained by Ryan Tibshirani  
<https://github.com/ryantibs/conformal>

semTools: Useful tools for structural equation modeling  
<https://github.com/simsem/semTools/wiki>

collapse is a C/C++ based package for data transformation and statistical computing in R. Its aims are:  
<https://github.com/SebastianKrantz/collapse>

latex2exp <https://github.com/stefano-meschiari/latex2exp>

Tracking Progress in Natural Language Processing <https://github.com/sebastianruder/NLP-progress>

Introduction to R Package Idealstan Robert Kubinec December 27, 2021 <https://github.com/saudiwin/idealstan>

Google's Compact Language Detector 3 is a neural network model for language identification and the successor of CLD2 (available from) CRAN. T <https://github.com/ropensci/cld3>

terra is an R package for spatial data analysis. <https://github.com/rspatial/terra>

rmcelreath stat\_\_rethinking\_2022 [https://github.com/rmcelreath/stat\\_\\_rethinking\\_2022#calendar-topical-outline](https://github.com/rmcelreath/stat__rethinking_2022#calendar-topical-outline)

charlatan makes fake data, inspired from and borrowing some code from Python's faker (<https://github.com/joke2k/faker>) <https://github.com/ropensci/charlatan>

skimr provides a frictionless approach to summary statistics which conforms to the principle of least surprise, displaying summary statistics the user can skim quickly to understand their data. It handles different data types and returns a <https://github.com/ropensci/skimr>

assertr <https://github.com/ropensci/assertr>

Explaining Models by Propagating Shapley Values of Local Components Hugh Chen, Scott Lundberg, Su-In Lee <https://arxiv.org/abs/1911.11888>

Explaining Models by Propagating Shapley Values of Local Components Hugh Chen, Scott Lundberg, Su-In Lee

Visualizing a Million Time Series with the Density Line Chart <https://idl.cs.washington.edu/files/2018-DenseLines-arXiv.pdf>

What is GANs? GANs(Generative Adversarial Networks) are the models that used in unsupervised machine learning <https://hollobit.github.io/All-About-the-GAN/>

Explaining machine learning models with SHAP and SAGE <https://iancovert.com/blog/understanding-shap-sage/>

The Dozen Things Experimental Economists Should Do (More of) <https://ideas.repec.org/p/feb/artefa/00648.html>

Synthetic Control Using Lasso (scul) <https://hollina.github.io/scul/>

Everything is fucked: The syllabus <https://thehardestscience.com/2016/08/11/everything-is-fucked-the-syllabus/>

Regression Modeling With Proportion Data (Part 1) Predicting Attendance in the German Handball-Bundesliga <https://hansjoerg.me/2019/05/10/regression-modeling-with-proportion-data-part-1/>

Conditional independences and causal relations implied by sets of equations Tineke Blom, Mirthe M. van Diepen, Joris M. Mooij; 2 <https://jmlr.org/papers/v22/20-863.html>

Researcher Degrees of Freedom Analysis <https://joachim-gassen.github.io/rdfanalysis/>

Evidence-Based Medicine—An Oral History Richard Smith, MBChB, CBE, FMedSci, FRCPE, FRCGP1; Drummond Rennie, MD, FRCP2 <https://jamanetwork.com/journals/jama/article-abstract/1817042>

Geocomputation with R's guide to reproducible spatial data analysis <https://jakubnowosad.com/ogh2022/#/title-slide>

Tutorial: JAX 101 <https://jax.readthedocs.io/en/latest/jax-101/index.html>

Autodidax: JAX core from scratch <https://jax.readthedocs.io/en/latest/autodidax.html>

Conda: Myths and Misconceptions <https://jakevdp.github.io/blog/2016/08/25/conda-myths-and-misconceptions/>

A Statistical Method for Empirical Testing of Competing Theories Kosuke Imai Princeton University Dustin Tingley <https://imai.fas.harvard.edu/research/files/mixture.pdf>

The Influence of Data Pre-processing and Post-processing on Long Document Summarization Xinwei Du, Kailun Dong, Yuchen Zhang, Yongsheng Li, Ruei-Yu Tsay <https://arxiv.org/abs/2112.01660>

COLLIDER: A Robust Training Framework for Backdoor Data Hadi M. Dolatabadi, Sarah Erfani, Christopher Leckie <https://arxiv.org/abs/2210.06704>

Time Series Data Augmentation for Deep Learning: A Survey Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, Huan Xu <https://arxiv.org/abs/2002.12478>

Bayesian Changepoint Detection in (Num)Pyro Posted on Tue 08 June 2021 in probabilistic programming, changepoint detection, Bayesian <https://irustandi.github.io/bayesian-changepoint-detection-in-numpyro.html>

Modeling Regime Shifts in Multiple Time Series Etienne Gael Tajeuna, Mohamed Bouguessa, Shengrui Wang <https://arxiv.org/abs/2109.09692>

Why negative results? Publication of negative results is difficult in most fields, but in NLP the problem is exacerbated by the near-universal focus on improvements in benchmarks. <https://insights-workshop.github.io/>

Small Data, Big Decisions: Model Selection in the Small-Data Regime Jorg Bornschein, Francesco Visin, Simon Osindero <https://arxiv.org/abs/2009.12583>

Quantifying With Only Positive Training Data Denis dos Reis, Marcílio de Souto, Elaine de Sousa, Gustavo Batista <https://arxiv.org/abs/2004.10356>

Superbloom: Bloom filter meets Transformer John Anderson, Qingqing Huang, Walid Krichene, Steffen Rendle, Li Zhang <https://arxiv.org/abs/2002.04723>

Selection via Proxy: Efficient Data Selection for Deep Learning Cody Coleman, Christopher Yeh, Stephen Musmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, Matei Zaharia <https://arxiv.org/abs/1906.11829>

Many Proxy Controls Ben Deaner <https://arxiv.org/abs/2110.03973>  
<https://datatalks.club/slack.html>

The Reparameterization “Trick” As Simple as Possible in TensorFlow [https://medium.com/\(llionj/the-reparameterization-trick-4ff30fe92954?\)](https://medium.com/(llionj/the-reparameterization-trick-4ff30fe92954?))

Mixed Models for Big Data GAM MIXED MODELS BIG DATA BAYESIAN Explorations of a fast penalized regression approach with mgcv <https://m-clark.github.io/posts/2019-10-20-big-mixed-models/>

I saw your RCT and I have some worries! FAQs Macartan Humphreys 6 September 2021 [https://macartan.github.io/i/notes/rct\\_faqs.html](https://macartan.github.io/i/notes/rct_faqs.html)

Avoiding technical debt in social science research <https://medium.com/pew-research-center-decoded/avoiding-technical-debt-in-social-science-research-54618194790a>

Confounder Selection: Objectives and Approaches F. Richard Guo, Anton Rask Lundborg, Qingyuan Zhao <https://math.papers.bar/paper/cc98597a2c2e11edaa66a71c10a887e7>

Diagnosing Biased Inference with Divergences Michael Betancourt January 2017 [https://mc-stan.org/users/documentation/case-studies/divergences\\_and\\_bias.html](https://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html)

Regression and Causality Michael Schomaker <https://math.papers.bar/paper/7e46323aaf3d11eb9864394904658322>

Mathematical Proof Between Generations <https://math.papers.bar/paper/347f685c018d11edb9b9d35608ee6155>

Document Deduplication with Locality Sensitive Hashing May 23, 2017 <https://mattilyra.github.io/2017/05/23/document-deduplication-with-lsh.html>

Mastering Shiny <https://mastering-shiny.org/>

How to be a modern scientist <https://leanpub.com/modernscientist>

Blind 75 LeetCode Questions <https://leetcode.com/discuss/general-discussion/460599/blind-75-leetcode-questions>

How Exactly UMAP Works And why exactly it is better than tSNE <https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668>

MATH 342 (Time Series), <https://lbelzile.github.io/timeseRies/>

Generative vs. Discriminative; Bayesian vs. Frequentist <https://lingpipe-blog.com/2013/04/12/generative-vs-discriminative-bayesian-vs-frequentist/>

All Bayesian Models are Generative (in Theory) <https://lingpipe-blog.com/2013/05/23/all-bayesian-models-are-generative-in-theory/>

Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly <https://journals.sagepub.com/doi/full/10.1177/2515245919858072>

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant <https://journals.sagepub.com/doi/full/10.1177/0956797611417632>

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone Uri SimonsohnView all authors and affiliations [https://journals.sagepub.com/doi/pdf/10.1177/0956797613480366?casa\\_token=r3DLe47WVEcAAp8JHISIDfwfyYHGtnWyqSw](https://journals.sagepub.com/doi/pdf/10.1177/0956797613480366?casa_token=r3DLe47WVEcAAp8JHISIDfwfyYHGtnWyqSw)

The national accounting paradox: how statistical norms corrode international economic data Daniel Mügge <https://orcid.org/0000-0001-9408-7597> d.k.muegge@uva.nl and Lukas Linsiview all authors and affiliations <https://journals.sagepub.com/doi/full/10.1177/1354066120936339>

Intellectual contributions meriting authorship: Survey results from the top cited authors across all science categories Gregory S. Patience ,Federico Galli,Paul A. Patience,Daria C. Boffito <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198117>

The goal of gluedown is to ease the transition from R's powerful vectors to formatted markdown text. <https://kiernann.com/gluedown/>

The Nine Circles of Scientific Hell NeuroskepticView all authors and affiliations <https://journals.sagepub.com/doi/10.1177/1745691612459519>

The Temporal Structure of Scientific Consensus Formation Uri Shwed shwed@bgu.ac.il and Peter S. BearmanView all authors and affiliations <https://journals.sagepub.com/doi/10.1177/0003122410388488>

Efficient estimation of generalized linear latent variable models <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216129#:~:text=Generalized%20linear%20lat>

The Phantom Menace: Omitted Variable Bias in Econometric Research Kevin A. ClarkeView all authors and affiliations <https://journals.sagepub.com/doi/10.1080/07388940500339183>

Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects Michael Gordon ,Domenico Viganola ,Anna Dreber,Magnus Johannesson,Thomas Pfeiffer Published: April 14, 2021 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0248780>

Why we publish where we do: Faculty publishing values and their relationship to review, promotion and tenure expectations Meredith T. Niles ,Lesley A. Schimanski,Erin C. McKiernan,Juan Pablo Alperin Published: March 11, 2020 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228914>

Reappraising the utility of Google Flu Trends Sasikiran Kandula ,Jeffrey Shaman <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007258>

The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets Takaya Saito ,Marc Rehmsmeier Published: March 4, 2015 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432>

Statistically Controlling for Confounding Constructs Is Harder than You Think Jacob Westfall ,Tal Yarkoni Published: March 31, 2016 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152719>

Ten simple rules for collaboratively writing a multi-authored paper Marieke A. Frassl ,David P. Hamilton,Blaize A. Denfeld,Elvira de Eyto,Stephanie E. Hampton,Philipp S. Keller,Sapna Sharma,Abigail S. L. Lewis,Gesa A. Weyhenmeyer,Catherine M. O'Reilly,Mary E. Lofton,Núria Catalán Published: November 15, 2018 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006508>

Analyzing Selection Bias for Credible Causal Inference When in Doubt, DAG It Out [https://journals.lww.com/epidem/fulltext/2019/07000/analyzing\\_selection\\_bias\\_for\\_credib](https://journals.lww.com/epidem/fulltext/2019/07000/analyzing_selection_bias_for_credib)

Selective publication of antidepressant trials and its influence on apparent efficacy: Updated comparisons and meta-analyses of newer versus older trials Erick H. Turner ,Andrea Cipriani,Toshi A. Furukawa,Georgia Salanti,Ymkje Anna de Vries Published: January 19, 2022 <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003886#sec018>

Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time Robert M. Kaplan ,Veronica L. Irvin Published: August 5, 2015 <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132382>

Test-Negative Designs Differences and Commonalities with Other Case–Control Studies with “Other Patient” Controls [https://journals.lww.com/epidem/Abstract/2019/11000/Test\\_Negative\\_Designs\\_\\_\\_Differences\\_and.10.aspx](https://journals.lww.com/epidem/Abstract/2019/11000/Test_Negative_Designs___Differences_and.10.aspx)

Examining linguistic shifts between preprints and publications David N. Nicholson,Vincent Rubinetti,Dongbo Hu,Marvin Thielk,Lawrence E. Hunter,Casey S. Greene Published: February 1, 2022 <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001470>

Break Down: Model Agnostic Explainers for Individual Predictions <https://pbiecek.github.io/breakDown/>

‘Trust Us’: Open Data and Preregistration in Political Science and International Relations <https://osf.io/preprints/metaarxiv/8h2bp/>



The Methodological Divide of Sociology - Evidence From Two  
Decades of Journal Publications <https://osf.io/preprints/socarxiv/s59bp/>

Shapley Residuals: Quantifying the limits of the Shapley value  
for explanations. <https://par.nsf.gov/biblio/10187138-shapley-residuals-quantifying-limits-shapley-value-explanations>

Activation Functions <https://paperswithcode.com/methods/category/activation-functions>

Software citation principles <https://peerj.com/articles/cs-86/>

Causality Redux: The Evolution of Empirical Methods in  
Accounting Research and the Growth of Quasi-Experiments  
[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3935088](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3935088)

Large-Scale Study of Curiosity-Driven Learning <https://pathak22.github.io/large-scale-curiosity/>

Bayes and big data: the consensus Monte Carlo algorithm  
[https://orsociety.tandfonline.com/doi/full/10.1080/17509653.2016.1142191?casa\\_token=AaNmx-7IVb4AAAAA%3Af\\_Zh3iwRXbyNvI4Tz5Erf0UrxkvftTGLN2EXwtvBu5Je0ejMp3fOYbYpUT9R6vBlgbwU2hoid9wZtaIZHA](https://orsociety.tandfonline.com/doi/full/10.1080/17509653.2016.1142191?casa_token=AaNmx-7IVb4AAAAA%3Af_Zh3iwRXbyNvI4Tz5Erf0UrxkvftTGLN2EXwtvBu5Je0ejMp3fOYbYpUT9R6vBlgbwU2hoid9wZtaIZHA)

Finance is Not Excused: Why Finance Should Not Flout Basic  
Principles of Statistics Forthcoming, Significance (Royal Statistical Society), 2021 [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3895330](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3895330)

Bayesian Time Series Forecasting with Change Point and  
Anomaly Detection <https://openreview.net/forum?id=rJLTTe-0W>

How to translate a verbal theory into a formal model  
<https://osf.io/preprints/metaarxiv/n7qsh/>

Does Regression Produce Representative Estimates of Causal  
Effects? <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12185>

Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variables <https://papers.nips.cc/paper/2020/file/c6a01432c8138d46ba39957a8250e027-Paper.pdf>

Specification Curve: Descriptive and Inferential Statistics on  
All Reasonable Specifications [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2694998](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694998)

The Standard Errors of Persistence Morgan Kelly [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3398303](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3398303)

<https://papers.labml.ai/lists>

The International Political Economy Data Resource  
[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2534067](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2534067)

LightGBM: A Highly Efficient Gradient Boosting Decision Tree  
<https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>

Notes on Changing from Rmarkdown/Bookdown to Quarto  
<https://www.njtierney.com/post/2022/04/11/rmd-to-qmd/>

Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017)  
<https://osf.io/preprints/metaarxiv/9sz2y/>

Teaching Safe-Stats, Not Statistical Abstinence [https://nhorton.people.amherst.edu/mererenovation/17\\_Wickham](https://nhorton.people.amherst.edu/mererenovation/17_Wickham)

Quantile Regression With LightGBM [https://notebook.community/ethen8181/machine-learning/ab\\_tests/quantile\\_regression/quantile\\_regression](https://notebook.community/ethen8181/machine-learning/ab_tests/quantile_regression/quantile_regression)

Forecasting: Principles and Practice <https://otexts.com/fpp3/>

Comparison of Preregistration Platforms <https://osf.io/preprints/metaarxiv/zry2u>

<https://nips.cc/Conferences>

<https://opensyllabus.org/>

Deep Learning Yoshua Bengio

An Introduction to Statistical Learning Gareth James, Daniela Witten, Trevor Hastie

Critical Questions for Big Data Danah Boyd, Kate Crawford

The Elements of Statistical Learning Trevor Hastie

Mostly Harmless Econometrics Joshua D. Angrist

Counterfactuals and Causal Inference Stephen L. Morgan

Machine Learning: A Probabilistic Perspective Kevin P. Murphy

Causality: Models, Reasoning, and Inference Judea Pearl

Methods to Estimate Causal Effects - An Overview on IV, DiD and RDD and a Guide on How to Apply them in Practice  
<https://osf.io/preprints/socarxiv/usvta/>

WINNER'S CURSE? ON PACE, PROGRESS, AND EMPIRICAL RIGOR <https://openreview.net/pdf?id=rJWF0Fywf>

NLP Highlights Podcast <https://open.spotify.com/show/4tGHZmicSHIVU3ksf5iYv8>

A method to streamline p-hacking <https://open.lnu.se/index.php/metapsychology/article/view/2529>

Machine Learning Theory - Part 3: Regularization and the Bias-variance Trade-off <https://mostafa-samir.github.io/ml-theory-pt3/>

NeetCode <https://neetcode.io/>

Detecting p-Hacking <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18583>

Is Temperature Exogenous? The Impact of Civil Conflict on the Instrumental Climate Record in Sub-Saharan Africa Kenneth A. Schultz, Justin S. Mankin First published: 28 March 2019 <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12425>

Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets Rhian Daniel, Jingjing Zhang, Daniel Farewell First published: 14 December 2020 <https://onlinelibrary.wiley.com/doi/full/10.1002/bimj.201900297>

vtree a flexible R package for displaying nested subsets of a data frame <https://nbarrowman.github.io/vtree.html>

Election polling errors across time and space Will Jennings & Christopher Wlezien <https://www.nature.com/articles/s41562-018-0315-6>

Bayesian Estimation of Signal Detection Models <https://mvuorre.github.io/posts/2017-10-09-bayesian-estimation-of-signal-detection-theory-models/>

Add to feature engineering The xspliner package is a collection of tools for training interpretable surrogate ML models. <https://modeloriented.github.io/xspliner/index.html>

Observational Studies <https://muse.jhu.edu/issue/48885>

Bringing more causality to analytics <https://motifanalytics.medium.com/bringing-more-causality-to-analytics-d378108bb15>

Nice stats note <https://moultano.wordpress.com/2013/08/09/logs-tails-long-tails/>

Figuring out why my object detection model is underperforming with FiftyOne, a great tool you probably haven't heard of  
<https://mlops.systems/redactionmodel/computervision/tools/debugging/jupyter/2022/03/12/fiftyone-computervision.html>

A ModernDive into R and the Tidyverse <https://moderndive.com/index.html>

Hopfield Networks is All You Need <https://ml-jku.github.io/hopfield-layers/>

Rediscovering Bayesian Structural Time Series June 7, 2020  
<https://minimizeregret.com/post/2020/06/07/rediscovering-bayesian-structural-time-series/>

Prophet [https://www.youtube.com/watch?v=pOYAXv15r3A&feature=emb\\_logo](https://www.youtube.com/watch?v=pOYAXv15r3A&feature=emb_logo)

A paper is the tip of an iceberg <https://minhlab.wordpress.com/2017/03/18/a-paper-is-the-tip-of-the-iceberg/>

Geometric Intuition for Training Neural Networks <https://seadl.org/2019/11/25/geometric-intuition-for-training-neural-networks/>

Latent Variable Modelling in brms January 20, 2020  
<https://scottclaessens.github.io/blog/2020/brmsLV/>

Robust Empirical Bayes Confidence Intervals <https://scholar.princeton.edu/mikkelpm/ebci>

Bias of OLS Estimators due to Exclusion of Relevant Variables and Inclusion of Irrelevant Variables Deepankar Basu  
[https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1257&context=econ\\_workingpaper](https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1257&context=econ_workingpaper)

scikit-survival <https://scikit-survival.readthedocs.io/en/latest/index.html>

Scikit-learn's Defaults are Wrong <https://ryxcommar.com/2019/08/30/scikit-learns-defaults-are-wrong/>

Selecting on the DV Design, Inference, and the Strategic Logic of Suicide Terrorism: A Rejoinder <https://scholar.princeton.edu/sites/default/files/rejoinder3.pdf>

Sampling from weird probability distributions Alan R. Pearse 6 July 2019 [https://rpubs.com/a\\_pear\\_9/weird\\_distributions](https://rpubs.com/a_pear_9/weird_distributions)

Outliers: Love'em or leave'em João Neto April 2020  
<https://rpubs.com/jpn3to/outliers>

Synthetic controls with staggered adoption Eli Ben-Michael, Avi Feller, Jesse Rothstein <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12448>

Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends <https://pubs.aeaweb.org/doi/pdfplus/10.1257/aeri.20210236?cookieSet=1>

(PROTOTYPE) INTRODUCTION TO NAMED TENSORS

IN PYTORCH Author: Richard Zou [https://pytorch.org/tutorials/intermediate/named\\_tensor\\_tutorial.html](https://pytorch.org/tutorials/intermediate/named_tensor_tutorial.html)

Comparing meta-analyses and preregistered multiple-laboratory replication projects <https://pubmed.ncbi.nlm.nih.gov/31873200/>

Does retraction after misconduct have an impact on citations? A pre-post study <https://pubmed.ncbi.nlm.nih.gov/33187964/>

Sparsity information and regularization in the horseshoe and other shrinkage priors Juho Piironen, Aki Vehtari <https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-2/Sparsity-information-and-regularization-in-the-horseshoe-and-other-shrinkage/10.1214/17-EJS1337SI.full>

A Word of Caution about Many Labs 4: If You Fail to Follow Your Preregistered Plan, You May Fail to Find a Real Effect <https://psyarxiv.com/ejubn>

Discrepancies between meta-analyses and subsequent large randomized, controlled trials <https://pubmed.ncbi.nlm.nih.gov/9262498/>

Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective <https://research.facebook.com/publications/applied-machine-learning-at-facebook-a-datacenter-infrastructure-perspective/>

An overview of systematic reviews found suboptimal reporting and methodological limitations of mediation studies investigating causal mechanisms <https://pubmed.ncbi.nlm.nih.gov/30904567/>

That's a lot to Process! Pitfalls of Popular Path Models Julia M. Rohrer Paul Hünemann Ruben C. Arslan Malte Elson <https://psyarxiv.com/paeb7/>

The Matrix-F Prior for Estimating and Testing Covariance Matrices Joris Mulder, Luis Raúl Pericchi

<https://projecteuclid.org/journals/bayesian-analysis/volume-13/issue-4/The-Matrix-F-Prior-for-Estimating-and-Testing-Covariance-Matrices/10.1214/17-BA1092.full>

Sample Size Justification Daniel Lakens <https://psyarxiv.com/9d3yf>

A unified view on Bayesian varying coefficient models  
Maria Franco-Villoria, Massimo Ventrucchi, Håvard Rue  
<https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-13/issue-2/A-unified-view-on-Bayesian-varying-coefficient-models/10.1214/19-EJS1653.full>

Introduction to the concept of likelihood and its applications  
Alexander Etz <https://psyarxiv.com/85ywt>

Tapped Out or Barely Tapped? Recommendations for How to Harness the Vast and Largely Unused Potential of the Mechanical Turk Participant Pool  
Jonathan Robinson, Cheskie Rosenzweig, Aaron J. Moss, Leib Litman <https://psyarxiv.com/jq589>

Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy  
Anna Laurina, Vichyute Himanshu Yadav, Shravan Vasishth  
<https://psyarxiv.com/hf297/>

Play with Generative Adversarial Networks (GANs) in your browser! <https://poloclub.github.io/ganlab/>

Probabilistic Machine Learning: An Introduction <https://probml.github.io/pml-book/book1.html>

How Good are FiftyThree Forecasts <https://projects.fivethirtyeight.com/checking-our-work/>

plotnine is an implementation of a grammar of graphics in Python, it is based on ggplot2 <https://plotnine.readthedocs.io/en/stable/>

Doubly Robust Difference-in-Differences <https://psantanna.com/DRDID/>

Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors  
James R. Faulkner, Vladimir N. Minin <https://projecteuclid.org/journals/bayesian-analysis/volume-13/issue-1/Locally-Adaptive-Smoothing-with-Markov-Random-Fields-and-Shrinkage-Priors/10.1214/17-BA1050.full>

Introduction to Probability for Data Science <https://probability4datascience.com/>

The Design Space of Computational Notebooks: An Analysis of 60 Systems in Academia and Industry [https://pg.ucsd.edu/publications/computational-notebooks-design-space\\_VLHCC-2020.pdf](https://pg.ucsd.edu/publications/computational-notebooks-design-space_VLHCC-2020.pdf)

Analyzing Minard's Visualization Of Napoleon's 1812 March <https://thoughtbot.com/blog/analyzing-minards-visualization-of-napoleons-1812-march>

A course in Time Series Analysis [https://web.stat.tamu.edu/~suhasini/teaching673/time\\_series.pdf](https://web.stat.tamu.edu/~suhasini/teaching673/time_series.pdf)

When and How Should One Use Deep Learning for Causal Effect Inference [https://technionmail-my.sharepoint.com/personal/urishalit\\_technion\\_ac\\_il/\\_layouts/15/onedrive](https://technionmail-my.sharepoint.com/personal/urishalit_technion_ac_il/_layouts/15/onedrive)

A Primer on Pólya-gamma Random Variables - Part II: Bayesian Logistic Regression <https://tiao.io/post/polya-gamma-bayesian-logistic-regression/>

treeheatr <https://trang1618.github.io/treeheatr/>

DiD Reading Group <https://taylorjwright.github.io/did-reading-group/>

Why is it that natural log changes are percentage changes? What is about logs that makes this so? <https://stats.stackexchange.com/questions/244199/why-is-it-that-natural-log-changes-are-percentage-changes-what-is-about-logs-th>

STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION [https://statistics.fas.harvard.edu/files/statistics-2/files/statistical\\_paradises\\_and\\_paradoxes.pdf](https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf)

Prior predictive checks for Bayesian regression. [https://engineeringdecisionanalysis.shinyapps.io/Priors/?\\_ga=2.25513452691.1612547156](https://engineeringdecisionanalysis.shinyapps.io/Priors/?_ga=2.25513452691.1612547156)

No, you can't explain what a p-value is with one sentence (Parts I, II) <https://statsepi.substack.com/p/no-you-cant-explain-what-a-p-value>

Does it make sense to log-transform the dependent when using Gradient Boosted Trees? <https://stats.stackexchange.com/questions/262114/does-it-make-sense-to-log-transform-the-dependent-when-using-gradient-boosted-tr/263753#263753>

Why is Euclidean distance not a good metric in high dimensions? <https://stats.stackexchange.com/questions/99171/why-is-euclidean-distance-not-a-good-metric-in-high-dimensions>

Plotting partial pooling in mixed-effects models <https://www.tjmahr.com/plotting-partial-pooling-in-mixed-effects-models/>

A Monte Carlo study on methods for handling class imbalance

<https://static1.squarespace.com/static/58a7d1e52994ca398697a621/t/5a2833cec83025cca6b99ff8/1512584144990/>

Billion-scale semantic similarity search with FAISS+SBERT

<https://towardsdatascience.com/billion-scale-semantic-similarity-search-with-faiss-sbert-c845614962e2>

The caret Package Max Kuhn 2019-03-27 <https://topepo.github.io/caret/index.html>

Custom Loss Functions for Gradient Boosting Optimize what matters <https://towardsdatascience.com/custom-loss-functions-for-gradient-boosting-f79c1b40466d>

What is the trade-off between batch size and number of iterations to train a neural network? <https://stats.stackexchange.com/questions/164876/what-is-the-trade-off-between-batch-size-and-number-of-iterations-to-train-a-neu/236393#236393>

Generalized Instrumental Variables <https://arxiv.org/pdf/1301.0560.pdf>

experimenter demand effects (EDEs)—bias that occurs when participants infer the purpose of an experiment and respond so as to help confirm a researcher’s hypothesis <https://www.cambridge.org/core/journals/american-political-science-review/article/abs/demand-effects-in-survey-experiments-an-empirical-assessment/043386DC63A69098E859414EF9932EBC>

An Overview of Google’s Work on AutoML and Future Directions Jun 14, 2019 <https://slideslive.com/38917526/an-overview-of-googles-work-on-automl-and-future-directions?locale=en>

The other kind of machine learning regression — unmeasured method performance <https://stuart-reynolds.medium.com/the-other-kind-of-machine-learning-regression-unmeasured-method-performance-81b7eb00efda>

The tidyverse style guide <https://style.tidyverse.org/index.html>



STOP CONFOUNDING YOURSELF! STOP CONFOUND-  
ING YOURSELF! <https://slatestarcodex.com/2014/04/26/stop-confounding-yourself-stop-confounding-yourself/>

Causal model and theory Suparna Chaudhry and Andrew Heiss  
2021-05-26 [https://stats.andrewheiss.com/donors-crackdowns-aid/00\\_causal-model-theory.html](https://stats.andrewheiss.com/donors-crackdowns-aid/00_causal-model-theory.html)

Stanza – A Python NLP Package for Many Human Languages  
<https://stanfordnlp.github.io/stanza/>

Inference for deterministic simulation models: The Bayesian  
melding approach <https://sites.stat.washington.edu/raftery/Research/PDF/poole2000.pdf>

Unofficial guidance on various topics by Social Science Data Ed-  
itors <https://social-science-data-editors.github.io/guidance/>

Welcome to The Advanced Matrix Factorization Jungle  
<https://sites.google.com/site/igorcarrron2/matrixfactorizations>

Exploring Enterprise Databases with R: A Tidyverse Approach  
<https://smithjd.github.io/sql-pet/>

Even with randomization, mediation analysis can still be  
confounded <https://www.r-bloggers.com/2019/04/even-with-randomization-mediation-analysis-can-still-be-confounded/>

Inference and Prediction Diverge in Biomedicine [https://www.cell.com/patterns/fulltext/S2666-3899\(20\)30160-4](https://www.cell.com/patterns/fulltext/S2666-3899(20)30160-4)

Algebra, Topology, Differential Calculus, and Optimiza-  
tion Theory For Computer Science and Machine Learning  
<https://www.cis.upenn.edu/~jean/math-deep.pdf>

The conditional nature of publication bias: a meta-regression  
analysis Published online by Cambridge University Press: 11  
May 2020 <https://www.cambridge.org/core/journals/political-science-research-and-methods/article/abs/conditional-nature-of-publication-bias-a-metaregression-analysis/40C0A166F3ED1516A051C5ED270D1650>

A Practical Guide to Weak Instruments Michael Keane† and  
Timothy Neal† [https://uc822f03065d525621a0034d9737.dl.dropboxusercontent.com/cd/0/inline2/Bwkck59j4aGcZhAQXypGx3CI1GMrv7IzuLn3qml-\\_qg3e7n9WFySMETeOh8YarvEvY0co5iYeI7ah1ppzWGgI3CLOk-5aStOsOeAY9TicEABvPqXkGLXgZd6eXOFKRBv-OldPL3mcixiiBC2OoLXuBymI3IyQIzTE2BPwCLdFAMijckVG6tTEng7yAEiJwOGXnwoFk6gB7td51Loi\\_1f26t\\_3zcBSHgpjBD\\_yVRhmb\\_R\\_Nt0kxdh3Nvhm0rueJcbzfl-](https://uc822f03065d525621a0034d9737.dl.dropboxusercontent.com/cd/0/inline2/Bwkck59j4aGcZhAQXypGx3CI1GMrv7IzuLn3qml-_qg3e7n9WFySMETeOh8YarvEvY0co5iYeI7ah1ppzWGgI3CLOk-5aStOsOeAY9TicEABvPqXkGLXgZd6eXOFKRBv-OldPL3mcixiiBC2OoLXuBymI3IyQIzTE2BPwCLdFAMijckVG6tTEng7yAEiJwOGXnwoFk6gB7td51Loi_1f26t_3zcBSHgpjBD_yVRhmb_R_Nt0kxdh3Nvhm0rueJcbzfl-)

gkqXGgZApK5Rc3JdAi7woThAAkD1hGko0HSYQT7SIIdRBZZ28FpMer2sZVBkFXpY\_9o-  
nefwJiFcbyIaiuqQVvQckMw6QWx\_L4nJRL1Btd7ztnss1dJ\_YA/file

Why You Should (or Shouldn't) be Using Google's JAX in  
2022 <https://www.assemblyai.com/blog/why-you-should-or-shouldnt-be-using-jax-in-2022/>

A guide to working with country-year panel data and Bayesian  
multilevel models <https://www.andrewheiss.com/blog/2021/12/01/multilevel-models-panel-data-guide/>

Statistical Significance Annual Review of Statistics and Its Ap-  
plication <https://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-031219-041051>

Bayesian Additive Regression Trees: A Review and Look  
Forward Annual Review of Statistics and Its Application  
<https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-031219-041110>

Bad Data Handbook: Cleaning Up The Data So You Can Get  
Back To Work <https://www.amazon.ca/Bad-Data-Handbook-Cleaning-Back/dp/1449321887>

Data Replication with Code Ocean – A How-To Guide for PA  
Authors Simon Heuberger October 2, 2019 [https://www.american.edu/spa/data-science/upload/authors\\_how\\_to.pdf](https://www.american.edu/spa/data-science/upload/authors_how_to.pdf)

Statistical Significance, p-Values, and the Reporting of Uncer-  
tainty Guido W. Imbens <https://www.aeaweb.org/articles?id=10.1257/jep.35.3.157>

Methods Matter: P-Hacking and Publication Bias in Causal  
Analysis in Economics By Abel Brodeur, Nikolai Cook, and An-  
thony Heyes <https://www.aeaweb.org/content/file?id=12747>

Using Synthetic Controls: Feasibility, Data Require-  
ments, and Methodological Aspects Alberto Abadie  
<https://www.aeaweb.org/articles?id=10.1257/jel.20191450&from=f>

Star Wars: The Empirics Strike Back Abel Brodeur Mathias Lé  
Marc Sangnier Yanos Zylberberg AMERICAN ECONOMIC  
JOURNAL: APPLIED ECONOMICS VOL. 8, NO. 1, JAN-  
UARY 2016 <https://www.aeaweb.org/articles?id=10.1257/app.20150044>

Beware performative reproducibility Well-meant changes to im-  
prove science could become empty gestures unless underlying

values change. <https://www.nature.com/articles/d41586-021-01824-z>

Plausibly Exogenous Timothy G. Conley, Christian B. Hansen, Peter E. Rossi <https://direct.mit.edu/rest/article-abstract/94/1/260/57981/Plausibly-Exogenous>

Deep Learning on Electronic Medical Records is doomed to fail  
Originally posted 2022-03-22 [https://www.moderndescartes.com/essays/deep\\_learning\\_emr/](https://www.moderndescartes.com/essays/deep_learning_emr/)

Collider bias undermines our understanding of COVID-19 disease risk and severity <https://www.nature.com/articles/s41467-020-19478-2>

Bayesian analysis of tests with unknown specificity and sensitivity Andrew Gelman, Bob Carpenter <https://www.medrxiv.org/content/10.1101/2020.05.22.20108944v3>

The One Standard Error Rule for Model Selection: Does It Work? by Yuchen Chen <sup>1</sup> and Yuhong Yang <sup>2,\*</sup>  
<https://www.mdpi.com/2571-905X/4/4/51>

MODELING COVARIANCE MATRICES IN TERMS OF STANDARD DEVIATIONS AND CORRELATIONS, WITH APPLICATION TO SHRINKAGE John Barnard, Robert McCulloch and Xiao-Li Meng [https://www.jstor.org/stable/24306780#metadata\\_info\\_tab\\_contents](https://www.jstor.org/stable/24306780#metadata_info_tab_contents)

Regression and Other Stories, with Andrew Gelman, Jennifer Hill & Aki Vehtari podcast <https://learnbayesstats.com/episode/20-regression-and-other-stories-with-andrew-gelman-jennifer-hill-aki-vehtari/>

Causal Inference: What If (the book) [https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2022/10/hernanrobins\\_WhatIf\\_15sep22.pdf](https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2022/10/hernanrobins_WhatIf_15sep22.pdf)

Statistical Comparisons of Classifiers over Multiple Data Sets  
<https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf>

CRITICAL VALUES ROBUST TO P-HACKING <https://www.pascalmichaillat.org/12.html>

P-value, compatibility, and S-value Author links open overlay panel Mohammad Ali Mansournia Maryam Nazemipoura Mahyar Etminan <https://www.sciencedirect.com/science/article/pii/S2590113322000153?via%3Dihub>

The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data Author links open overlay panel Gilles E. Gignac Marcin Zajenkowski <https://www.sciencedirect.com/science/article/abs/pii/S0160289620300271>

Natural Scales in Geographical Patterns Telmo Menezes<sup>1,\*</sup>  
and Camille Roth<sup>1,2,3</sup>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5379183/>

Making Sense of Sensitivity: Extending Omitted Variable Bias  
January 2018 [https://www.researchgate.net/publication/322509816\\_Making\\_Sense\\_of\\_Sensitivity\\_Extending\\_](https://www.researchgate.net/publication/322509816_Making_Sense_of_Sensitivity_Extending_)

A Survey of Methods for Time Series Change Point De-  
tection Samaneh Aminikhanghahi and Diane J. Cook  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5464762/>

Negative Controls: A Tool for Detecting Confounding and Bias  
in Observational Studies <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3053408/>

Introduction to Facebook AI Similarity Search (Faiss)  
<https://www.pinecone.io/learn/faiss-tutorial/>

Is probabilistic bias analysis approximately Bayesian?  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3257063/>

Everything you always wanted to know about eval-  
uating prediction models (but were too afraid to ask)  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2997853/>

Should We Trust Clustered Standard Errors? A Comparison  
with Randomization-Based Methods Lourenço S. Paz & James  
E. West <https://www.nber.org/papers/w25926>

The Value of Statistical Life: A Meta-analysis of Meta-analyses  
H. Spencer Banzhaf <https://www.nber.org/papers/w29185>

A large-scale study on research code quality and execution Ana  
Trisovic, Matthew K. Lau, Thomas Pasquier & Mercè Crosas  
<https://www.nature.com/articles/s41597-022-01143-6>

SELECTION INTO IDENTIFICATION IN FIXED EF-  
FECTS MODELS, WITH APPLICATION TO HEAD START  
[https://www.nber.org/system/files/working\\_papers/w26174/w26174.pdf](https://www.nber.org/system/files/working_papers/w26174/w26174.pdf)

Fast and effective pseudo transfer entropy for bivariate data-  
driven causal inference Riccardo Silini & Cristina Masoller  
Scientific Reports volume 11, Article number: 8423 (2021)  
<https://www.nature.com/articles/s41598-021-87818-3>

Variable selection in the presence of missing data: resampling  
and imputation Qi Long\* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5156376/>

Consensus features nested cross-validation Saeid Parvande,1,2  
Hung-Wen Yeh,3 Martin P Paulus,4 and Brett A McKinney1,5  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7776094/>

Bayesian Approaches for Missing Not at Random Outcome  
Data: The Role of Identifying Restrictions Antonio R. Linero\*  
and Michael J. Daniels† <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936760/>

Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking  
and Publication Bias? Abel Brodeur Nikolai Cook Jonathan  
Hartley Anthony Heyes <https://osf.io/preprints/metaarxiv/uxf39/>

Elements of Information Theory, 2nd Edition <https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959>

Why Data Is Never Raw <https://www.thenewatlantis.com/publications/why-data-is-never-raw>

When U.S. air force discovered the flaw of averages  
<https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html>

A History of Scientific Journals Publishing at the Royal Society,  
1665-2015 Aileen Fyfe, Noah Moxham, Julie McDougall-Waters,  
and Camilla Mørk Røstvik <https://www.uclpress.co.uk/products/187262>

Assessing the Statistical Analyses Used in Basic and Applied  
Social Psychology After Their p-Value Ban Ronald D. Fricker  
Jr., Katherine Burke, Xiaoyan Han & William H. Woodall  
<https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1537892>

Oh No! I Got the Wrong Sign! What Should I Do? Peter E.  
Kennedy <https://www.tandfonline.com/doi/abs/10.3200/JECE.36.1.77-92>

Prediction, Estimation, and Attribution Bradley Efron  
<https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1762613>

The Gaussian Graphical Model in Cross-Sectional and Time-  
Series Data Sacha Epskamp, Lourens J. Waldorp, René Mõttus  
& Denny Borsboom Pages 453-480 | Published online: 16 Apr  
2018 <https://www.tandfonline.com/doi/full/10.1080/00273171.2018.1454823>

Bruce E. Hansen Jackknife Standard Errors for Clustered Re-  
gression August 2022 <https://www.ssc.wisc.edu/~bhansen/papers/tcauchy.pdf>

A Five-Star Guide for Achieving Replicability and Reproducibility When Working with GIS Software and Algorithms  
<https://www.tandfonline.com/doi/abs/10.1080/24694452.2020.1806026?journalCode=raag21>

Some useful equations for nonlinear regression in R Andrea Onofri 2019-01-08 <https://www.statforbiology.com/nonlinearregression/usefulequations>

Random Forests for Spatially Dependent Data Arkajyoti Saha, Sumanta Basu & Abhirup Datta Received 02 Dec 2020  
<https://www.tandfonline.com/doi/abs/10.1080/01621459.2021.1950003>

Jackknife Standard Errors for Clustered Regression Bruce E. Hansen\* University of Wisconsin† August, 2022  
<https://www.ssc.wisc.edu/~bhansen/papers/tcauchy.pdf>

Social Science Reproduction Platform (SSRP) is an openly licensed platform that facilitates the sourcing, cataloging, and review of attempts to verify and improve the computational reproducibility of social science research.  
<https://www.socialsciencereproduction.org/about>

Non-Standard Errors <https://orbilu.uni.lu/bitstream/10993/48686/1/SSRN-id3961574.pdf>

Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices. <https://psycnet.apa.org/record/2023-14088-001>

Bayesian inference with INLA <https://becarioprecario.bitbucket.io/inla-gitbook/index.html>

Measurement Models [http://cfariss.com/documents/FarissKenwickReuning2019\\_MesurmentModels.pdf](http://cfariss.com/documents/FarissKenwickReuning2019_MesurmentModels.pdf)

Distinguishing cause from effect using observational data: methods and benchmarks <https://arxiv.org/abs/1412.3773v3>

The Effect: An Introduction to Research Design and Causality <https://theeffectbook.net/>

Feature Engineering and Selection: A Practical Approach for Predictive Models <https://bookdown.org/max/FES/>

Feature Interactions in XGBoost <https://arxiv.org/abs/2007.05758>

How should variable selection be performed with multiply imputed data? <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3177>

Statistical Nonsignificance in Empirical Economics <https://www.aeaweb.org/articles?id=10.1257/aeri.20190252&fr>

Deep Learning <https://www.deeplearningbook.org/>

On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives <https://arxiv.org/abs/1902.10286>

Why Propensity Scores Should Not Be Used for Matching <https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-for-matching>

What are the most important statistical ideas of the past 50 years? <https://arxiv.org/pdf/2012.00174.pdf>

Automatic Differentiation Variational Inference <https://www.jmlr.org/papers/volume18/16-107/16-107.pdf>

Methodology over metrics: current scientific standards are a disservice to patients and society [https://www.jclinepi.com/article/S0895-4356\(21\)00170-0/fulltext](https://www.jclinepi.com/article/S0895-4356(21)00170-0/fulltext)

Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong <https://journals.sagepub.com/doi/10.1080/07388940500339167>

Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition <https://bookdown.org/content/4857/>

Random Walk: A Modern Introduction <https://math.uchicago.edu/~lawler/srwbook.pdf>

On the reliability of published findings using the regression discontinuity design in political science <https://arxiv.org/abs/2109.14526>

Exploring the Dynamics of Latent Variable Models <https://www.cambridge.org/core/journals/political-analysis/article/abs/exploring-the-dynamics-of-latent-variable-models/CBE116F37900DAE957B2D7EB53DB0907#.X7h7GMnwHwM.twitter>

Cross-validation: what does it estimate and how well does it do it? <https://arxiv.org/abs/2104.00673>

What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory <https://journals.sagepub.com/doi/abs/10.1177/00031224211004187#:~:text=>

Measurement error and the replication crisis <https://www.science.org/doi/10.1126/science.aal3618>

Bayesian Modeling and Computation in Python <https://bayesiancomputationbook.com/welcome.html>

The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2011.00525.x>

Causal Inference for The Brave and True <https://matheusfacure.github.io/python-causality-handbook/landing-page.html>

A Parsimonious Tour of Bayesian Model Uncertainty <https://arxiv.org/abs/1902.05539>

The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant <http://www.stat.columbia.edu/~gelman/research/published/signif4.pdf>

Prediction, Estimation, and Attribution <https://efron.ckirby.su.domains//papers/2019PredictEstimatAttribut.pdf>

Difference-in-Differences with a Continuous Treatment [https://psantanna.com/files/Callaway\\_Goodman-Bacon\\_SantAnna\\_2021.pdf](https://psantanna.com/files/Callaway_Goodman-Bacon_SantAnna_2021.pdf)

A Practical Introduction to Regression Discontinuity Designs: Foundations <https://arxiv.org/pdf/1911.09511.pdf>

The influence of decision-making in tree ring-based climate reconstructions <https://www.nature.com/articles/s41467-021-23627-6>

The Influence of Hidden Researcher Decisions in Applied Microeconomics <https://docs.iza.org/dp13233.pdf>

Introducing geofacet <https://ryanhafen.com/blog/geofacet/>

The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432>

Cross-validation FAQ Aki Vehtari First version 2020-03-11. Last modified 2022-07-30. <https://avehtari.github.io/modelselection/CV-FAQ.html>

Shapley values for feature selection: The good, the bad, and the axioms Daniel Fryer, Inga Strümke, Hien Nguyen <https://arxiv.org/abs/2102.10936>

A Crash Course in Good and Bad Controls Carlos Cinelli\* Andrew Forney† Judea Pearl [https://ftp.cs.ucla.edu/pub/stat\\_ser/r493.pdf](https://ftp.cs.ucla.edu/pub/stat_ser/r493.pdf)

Reinforcement Learning in R Nicolas Pröllochs, Stefan Feuerriegel <https://arxiv.org/abs/1810.00240>



Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.02881>

NumPyro <https://github.com/pyro-ppl/numpyro>

Replacing the do-calculus with Bayes rule <https://arxiv.org/pdf/1906.07125.pdf>

Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results <https://academic.oup.com/qje/article-abstract/134/2/557/5195544?redirectedFrom=fulltext&login=false>

A Survey on Societal Event Forecasting with Deep Learning <https://arxiv.org/pdf/2112.06345.pdf>

Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the “Many Analysts, One Data Set” Project <https://journals.sagepub.com/doi/full/10.1177/23780231211024421>

Political Event Coding as Text to Text Sequence Generation [https://yaoyaodai.github.io/files/CASE\\_2022.pdf](https://yaoyaodai.github.io/files/CASE_2022.pdf)

Bayesian Thinking for Toddlers <https://psyarxiv.com/w5vbp/>

The Dunning-Kruger Effect is Autocorrelation <https://economicsfromthetopdown.com/2022/04/08/the-dunning-kruger-effect-is-autocorrelation/>

Causal Inference and Its Applications in Online Industry <https://alex deng.github.io/causal/>

Bayesian Workflow <https://arxiv.org/abs/2011.01808>

Papers about Causal Inference and Language <https://github.com/causaltext/causal-text-papers>

Achieving Statistical Significance with Control Variables and Without Transparency <https://www.cambridge.org/core/journals/political-analysis/article/abs/achieving-statistical-significance-with-control-variables-and-without-transparency/1E867C357835019E0C9322B918414045>

Questionable research practices among researchers in the most research-productive management programs <https://onlinelibrary.wiley.com/doi/10.1002/job.2623>

The problem of the missing dead Sophia Dawkins <https://orcid.org/0000-0002-2609-0820> [sophia.dawkins@yale.edu](mailto:sophia.dawkins@yale.edu) View all authors and affiliations

<https://declaredesign.org/>

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty  
<https://www.pnas.org/doi/full/10.1073/pnas.2203150119>

How to avoid machine learning pitfalls: a guide for academic researchers <https://arxiv.org/pdf/2108.02497.pdf>

Multiple Imputation Through XGBoost <https://arxiv.org/abs/2106.01574>

TACRED is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text from the corpus used in the yearly TAC Knowledge Base Population (TAC KBP) challenges. <https://nlp.stanford.edu/projects/tacred/>

Understanding lime [https://cran.r-project.org/web/packages/lime/vignettes/Understanding\\_lime.html](https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html)

Race to the bottom: Spatial aggregation and event data  
[https://www.tandfonline.com/doi/abs/10.1080/03050629.2022.2025365?casa\\_token=wrWE-FIlIAAAAA%3AU5Dsm6FMC\\_1wN1GKsdbEyveqc7XKFEe2beBsBxSVjVopzFgrJdYgfQ9gvW0nL17UUSyAIR5](https://www.tandfonline.com/doi/abs/10.1080/03050629.2022.2025365?casa_token=wrWE-FIlIAAAAA%3AU5Dsm6FMC_1wN1GKsdbEyveqc7XKFEe2beBsBxSVjVopzFgrJdYgfQ9gvW0nL17UUSyAIR5)

Inference and Prediction Diverge in Biomedicine [https://www.cell.com/patterns/fulltext/S2666-3899\(20\)30160-4](https://www.cell.com/patterns/fulltext/S2666-3899(20)30160-4)

I saw your RCT and I have some worries! FAQs Macartan Humphreys [https://macartan.github.io/i/notes/rct\\_faqs.html](https://macartan.github.io/i/notes/rct_faqs.html)

Measuring the landscape of civil war: Evaluating geographic coding decisions with historic data from the Mau Mau rebellion  
<https://journals.sagepub.com/eprint/dRCkdD4ZWSp99x8cinAV/full>

LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale <https://arxiv.org/abs/2208.07339>

Understanding Machine Learning: From Theory to Algorithms  
<https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf>

Known broken ones that didn't get sorted

Confounder Selection: Objectives and Approaches F. Richard Guo, Anton Rask Lundborg, Qingyuan Zhao

[2011.01808](<https://arxiv.org/abs/2011.01808>) Article identifier not recognized [2108.02497](<https://arxiv.org/abs/2108.02497>)  
Article identifier not recognized

# **Sociology of Inference**

# Practice Coding

NeetCode <https://neetcode.io/>

<https://leetcode.com/>

<https://www.hackerrank.com/>

# Salaries

[Harnham 2022 Guide](#)

[levels.fyi](#)

## **Part II**

# **Presentation**

# Markdown

[R Markdown Cookbook](#)

## **Part III**

# **——-Computation——-**



**Part IV**

**Computation**

# Computation

## git

<https://git-scm.com/doc>

# Bash

Bash scripting cheatsheet <https://devhints.io/bash>

# R

<https://www.r-project.org/other-docs.html>

[Hands-On Programming with R, Garrett Golemund](#)

## **Tidyverse**

<https://www.tidyverse.org/>

[R for Data Science](#)

[The Tidyverse Cookbook](#)

# Python

<https://docs.python.org/3/>

## Numpy

<https://numpy.org/>

## Pandas

<https://pandas.pydata.org/docs/>

Effective Pandas <https://store.metasnake.com/effective-pandas-book>

**jax**

<https://github.com/google/jax>

# Numpyro

<https://github.com/pyro-ppl/numpyro>

## Stan

<https://mc-stan.org/users/documentation/>

## brms

<https://github.com/paul-buerkner/brms>



## pyro

<https://pyro.ai/>

[The StatQuest Introduction to PyTorch](#)

**tensorflow**

<https://www.tensorflow.org/>

# SQL

postgresql <https://www.postgresql.org/docs/>

For practice <https://www.hackerrank.com/domains/sql>

## **Part V**

# **Data management**

# Filter

**Instance of:** Higher-order function

**AKA:** Subset

**Distinct from:**

**English:**

**Formalization:**

**Cites:** [Wikipedia](#) ; [Wikidata](#)

**Code**

**R**

**Base**

[subset](#): Subsetting Vectors, Matrices and Data Frames

**Dplyr**

[Subset rows using column values](#)

**DataTable**

[Subsetting Rows](#)

## Python

Documentation: [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Joins



# Regex

R [Regular expressions](#)

# Fuzzy Recording Matching

Name Match

## **Part VI**

# **——-Certainty——-**

## **Part VII**

# **Mathematical Objects**

# Set

Cites: [Wikipedia](#); [Wikidata](#); [PlanetMath](#)

# List (Sequence)

AKA: Sequence,  $a_n$  where  $n$  is the  $n$ th element, (1,2,3, ....)

Distinct from: Set

Measure of:

Description: A list is a collection of objects with a specific ordering and where the same object can appear more than once. Call each object an element, and its location its index or rank. An index is a natural number counting upward from the first element in the list. Whether counting begins at 0 or 1 depends on local conventions.

Formalization:

Algorithm:

Cites: [Wikipedia](#) [Wikidata](#) [Encyclopedia Of Math](#) [Wolfram PlanetMath](#)

## R

Documentation:

[list: Lists – Generic and Dotted Pairs](#)

Examples:

```
example_list = list(1,2,3)
example_list
```

```
[[1]]
```

```
[1] 1
```

```
[[2]]
```

```
[1] 2
```

```
[[3]]
```

```
[1] 3
```

## Python

Documentation:

[More on Lists](#)

Examples:

```
example_list = [1,2,3]
example_list
```

```
[1, 2, 3]
```

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
dbListTables(con)
```

```
character(0)
```

```
dbWriteTable(con, "mtcars", mtcars)
dbListTables(con)
```

```
[1] "mtcars"
```

```
create table StatisticalNumbers(
  value int
)
```

```
SELECT * FROM mtcars LIMIT 5;
```

Table 1: 5 records

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
18.7	8	360	175	3.15	3.440	17.02	0	0	3	2



# Vector/Matrix/Tensor

**Instance of:** algebraic object / data structure

**AKA:** array, matrices

**Distinct from:** list

**English:** Vectors, matrices, and tensors are like lists in that they are a collection of objects which are indexed. They differ in that the index can be multi-dimensional, where vectors are 1-d indexed, matrices are 2-d indexed, and tensors are m-d indexed. They also are typically constrained to have objects that share the same type, e.g. numbers or strings.

**Formalization:**

**Cites:**

Array:

[Wikipedia](#)

[3Blue1Brown: Vectors | Chapter 1, Essence of linear algebra](#)

[3Blue1Brown: Linear combinations, span, and basis vectors | Chapter 2, Essence of linear algebra](#)

Matrix:

[Wikipedia](#)

[3Blue1Brown: Linear transformations and matrices | Chapter 3, Essence of linear algebra](#)

Tensor:

[Wikipedia](#)

**Code**

## R

### Vector

Note unlike matrix and array, the basic vector function initializes an empty vector and you have to actually use `as.vector` to coerce something else to vector as the constructor.

vector: [Vectors](#)

```
example_vector <- as.vector(c(1,2,3,4))
class(example_vector)
```

```
[1] "numeric"
```

```
example_vector
```

```
[1] 1 2 3 4
```

### Matrix

Note we can choose which direction to fill the matrix with, either by `row1-col1`, `row1-col2`, `row1-col3`, `row1-col4`

matrix: [Matrices](#)

```
example_matrix <- matrix(c(1,2,3,4,"A","B","C","D"), nrow = 2, ncol = 4, byrow = TRUE,
                        dimnames = list(c("row1", "row2"),
                                         c("C.1", "C.2", "C.3", "C.4")))
class(example_matrix)
```

```
[1] "matrix" "array"
```

```
example_matrix
```

```
      C.1 C.2 C.3 C.4
row1 "1"  "2"  "3"  "4"
row2 "A"  "B"  "C"  "D"
```

## Arrays

Note array dimensions are ordered, row, column, depth, ..., M, and elements are filled row1-col1-depth1, row2-col1-depth1, row1-col2-depth1,... and so on. Note this was coerced to a string because any of the elements were a string.

array: Multi-way Arrays

```
example_tensor= array(c(1,2,3,4,"A","B","C","D","+","-","*","/"),dim=c(2,3,2,2))
class(example_tensor)
```

```
[1] "array"
```

```
example_tensor
```

```
, , 1, 1
```

```
      [,1] [,2] [,3]
[1,] "1"  "3"  "A"
[2,] "2"  "4"  "B"
```

```
, , 2, 1
```

```
      [,1] [,2] [,3]
[1,] "C"  "+"  "*"
[2,] "D"  "-"  "/"
```

```
, , 1, 2
```

```
      [,1] [,2] [,3]
[1,] "1"  "3"  "A"
[2,] "2"  "4"  "B"
```

```
, , 2, 2
```

```
      [,1] [,2] [,3]
[1,] "C"  "+"  "*"
[2,] "D"  "-"  "/"
```

## Python

Documentation:

Examples:

## SQL

Documentation:

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

**Jax**

**Torch**

```
import torch
```

# Table

## Introduction

**Instance of:** arrangement of information or data

**AKA:** Dataframe

**Distinct from:**

**English:** A collection of rows and columns, where rows represent specific instances (AKA records, k-tuple, n-tuple, or a vector), and columns represent features (AKA variables, parameters, properties, attributes, or stanchions). The intersection of a row and column is called a sell.

**Formalization:**

**Cites:** [Wikipedia Table \(information\)](#) ; [Wikipedia Table Table \(database\)](#) ; Wikidata ; Wolfram

[ML Frameworks Interoperability Cheat Sheet](#)

**Code**

**R**

**Documentation:** [data.frame: Data Frames](#)

Examples:

```
df=data.frame(a=c(1,2,3,4), b=c('a','b','c','d'))
df
```

```
   a b
1 1 a
2 2 b
3 3 c
4 4 d
```

## Python

Documentation: [pandas.DataFrame](#)

Examples:

```
import pandas as pd
df = pd.DataFrame({'a': [1, 2,3,4], 'b': ['a','b','c','d']})
df
```

```
   a b
0  1 a
1  2 b
2  3 c
3  4 d
```

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
```

```
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

```
DROP TABLE IF EXISTS df;
```

```
CREATE TABLE IF NOT EXISTS df (
  a INTEGER,
  b CHAR
);
```

```
INSERT INTO df (a, b)
VALUES
  (1, 'a'),
  (2, 'b'),
  (3, 'c'),
  (4, 'd');
```

```
SELECT * FROM df;
```

Table 2: 4 records

a	b
1	a
2	b
3	c
4	d



## Torch

```
import torch
```

# Function

## Introduction

Instance of:

## Frequentist

**AKA:**  $f : X \mapsto Y$ ,  $f(x)$ , map, mapping, linear map, linear function, transformation, morphism

**Distinct from:**

**English:** A function from a set X to a set Y is an assignment of an element of Y to each element of X. The set X is the domain, and the set Y is the codomain.

**Formalization:**

The formalization is annoying in that there are multiple conventions for writing a function which are equivalent.

Pseudo code

```
f = function(X){
  Y=X+1 #some operation
  return(Y)
}
```

Where the domain is all floating point numbers and so is the codomain (up to the precision of the computer).

That can also be written.

$$f : X \mapsto Y$$

Where  $f$  is the name of the function,  $\mapsto$  is the “maps to” or “Maplet” symbol.

Cites: Wikipedia ; Wikidata ; Wolfram

## Code

### R

Examples:

### Python

Examples:

### SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

Code

**Part VIII**

**Operations of Logic**

# And

**Instance of:** logic

**AKA:**

**Distinct from:**

**English:**

**Formalization:**

**Cites:** Wikipedia ; Wikidata ; Wolfram

**Code**

## R

Examples:

```
a=TRUE  
b=FALSE  
a & b
```

```
[1] FALSE
```

## Python

Examples:

Note do not use & or you will get a different result

```
a=True
b=False
a and b
```

False

Note do not use & or you will get a different result &' is a bitwise operator in Python that acts on bits and performs bit by bit operation

<https://www.geeksforgeeks.org/difference-between-and-and-in-python/#:~:text=and%20is%20a%20Logical%20AND,otherwise%20True%20when%20using%20logically.>

```
a=14
b=4
a & b
```

4

## Numpy

[https://numpy.org/doc/stable/reference/generated/numpy.logical\\_and.html](https://numpy.org/doc/stable/reference/generated/numpy.logical_and.html)

Examples:

```
import numpy as np
a=np.array(True)
b=np.array(False)
np.logical_and(a, b)
```

False

## Jax

Examples:

```
import jax.numpy as jnp
a=jnp.array(True)
```

WARNING:jax.\_src.lib.xla\_bridge:No GPU/TPU found, falling back to CPU. (Set TF\_CPP\_MIN\_LOG\_LEVEL=0 for verbose output)

```
b=jnp.array(False)
jnp.logical_and(a, b)
```

DeviceArray(False, dtype=bool)

## SQL

[https://www.w3schools.com/sql/sql\\_operators.asp](https://www.w3schools.com/sql/sql_operators.asp) <https://www.databasesstar.com/sql-boolean-data-type/#:~:text=SQL%20Server%20Boolean,TRUE%20and%200%20for%20FALSE.>

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
```



```
con <- dbConnect(RPostgres::Postgres())
```

```
SELECT TRUE AND FALSE;
```

Table 3: 1 records

?column?
FALSE

```
SELECT TRUE AND TRUE;
```

Table 4: 1 records

?column?
TRUE

## Torch

[https://pytorch.org/docs/stable/generated/torch.logical\\_and.html](https://pytorch.org/docs/stable/generated/torch.logical_and.html)

```
import torch
a = torch.tensor(True)
b = torch.tensor(False)
torch.logical_and(a, b)
```

```
tensor(False)
```

## Tensorflow

[https://www.tensorflow.org/api\\_docs/python/tf/math/logical\\_and](https://www.tensorflow.org/api_docs/python/tf/math/logical_and)

```
import tensorflow as tf
a = tf.constant(True)
b = tf.constant(False)
tf.math.logical_and(a, b)
```

```
<tf.Tensor: shape=(), dtype=bool, numpy=False>
```

## **Part IX**

# **Operations of Arithmetic**

# Addition

## Introduction

Instance of: operation of arithmetic

## Frequentist

AKA: + ; add

Distinct from:

English:

Formalization:

Cites: [Wikipedia](#) ; [Wikidata](#) ; Wolfram

Code

R

Documentation: [mean](#): [Arithmetic Mean](#)

Examples:

Python

Documentation: [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

**Code**

# Subtraction

## Introduction

Instance of: operation of arithmetic

## Frequentist

AKA: - ; minus

Distinct from:

English:

Formalization:

Cites: [Wikipedia](#) ; [Wikidata](#) ; Wolfram

Code

R

Documentation: [mean](#): [Arithmetic Mean](#)

Examples:

Python

Documentation: [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:



Cites:

**Code**

# Multiplication

## Introduction

**Instance of:** operation of arithmetic

## Frequentist

**AKA:**  $*$  ;  $\times$  ; ; multiply

**Distinct from:**

**English:**

**Formalization:**

**Cites:** Wikipedia ; Wikidata ; Wolfram

[3Blue1Brown: Matrix multiplication as composition | Chapter 4, Essence of linear algebra](#) [3Blue1Brown: Cross products in the light of linear transformations | Chapter 11, Essence of linear algebra](#)

**Code**

**R**

**Documentation:** [mean: Arithmetic Mean](#)

**Examples:**

## Python

Documentation: [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

Code

# Division

## Introduction

Instance of: operation of arithmetic

## Frequentist

AKA:  $/$  ;  $\frac{numerator}{denominator}$  ;  $\div$

Distinct from:

English:

Formalization:

Cites: [Wikipedia](#) ; [Wikidata](#) ; Wolfram

Code

R

Documentation: [mean](#): [Arithmetic Mean](#)

Examples:

Python

Documentation: [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

**Code**

# Modulo

**Instance of:** modular arithmetic

**AKA:**

**Distinct from:**

**English:** A modulo operation retrieves the remainder following euclidean division. Euclidean division is division with a remainder. The quotient is an integer count of the number of times the whole divisor can be placed in the dividend. The remainder is the dividend minus the quotient times the divisor.

**Formalization:**

**Cites:** Wikipedia ; Wikidata ; Wolfram

**Code**

**R**

Examples:

```
dividend=10
divisor=3
remainder= 10 %% 3
remainder
```

[1] 1



## Python

Examples:

```
dividend=10
divisor=3
remainder= 10 % 3
remainder
```

1

## Numpy

Examples:

```
import numpy as np
dividend=np.array(10)
divisor=np.array(3)
remainder= np.mod(10,3)
remainder
```

1

## Jax

Examples:

```
import jax.numpy as jnp
dividend=jnp.array(10)
```

WARNING:jax.\_src.lib.xla\_bridge:No GPU/TPU found, falling back to CPU. (Set TF\_CPP\_MIN\_LOG\_LEVEL=0 for verbosity)

```
divisor=jnp.array(3)
remainder= jnp.mod(10,3)
remainder
```

DeviceArray(1, dtype=int32, weak\_type=True)

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

```
SELECT MOD(10, 3);
```

Table 5: 1 records

mod
1

## Torch

```
import torch
dividend = torch.tensor(10)
divisor = torch.tensor(3)
remainder=torch.remainder(dividend, divisor)
remainder
```

tensor(1)

## Tensorflow

```
import tensorflow as tf
dividend = tf.constant(10)
divisor = tf.constant(3)
remainder=tf.math.floormod(dividend, divisor)
remainder
```

<tf.Tensor: shape=(), dtype=int32, numpy=1>

**Part X**

**Operations of Algebra**

# Dot product

## Introduction

**Instance of:** algebraic operation

**AKA:** scalar product; inner product ; projection product ;  $\cdot$  ;  $\cdot$

**Distinct from:**

**English:**

**Formalization:**

$$a \cdot b$$

**Cites:** [Wikipedia](#) ; Wikidata ; Wolfram

[3Blue1Brown: Dot products and duality | Chapter 9, Essence of linear algebra](#)

**Code**

**R**

**Documentation:**

Examples:

**Python**

**Documentation:** [numpy.mean](#)

Examples:

## SQL

Documentation: [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
#* deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

**Code**

## Part XI

# ——-Uncertainty——-



**Part XII**

**Probability**

# Probability distribution

**Instance of:** function

**AKA:**

**Distinct from:**

**English:** A probability distribution is a function that assigns probabilities to events. A normal distribution, beta distribution, and gamma distribution, for example, are all probability distributions.

There are three Kolmogorov axioms that a probability function must meet.

1. Probability of every event must be nonnegative.
2. No probability must exceed 1.
3. The probability of either of two disjoint events occurring must equal the sum of each of their individual probabilities of occurring.

Do not confuse a random variable for its distribution. A probability distribution is mathematical description of the long run behavior of a process, and a random variable may be usefully described by an existing well known probability distribution (Ross 2022).

**Formalization:**

It's a function

$$P : \mathcal{A} \mapsto \mathbb{R}$$

The three Kolmogorov axioms are formalized as

Axiom 1

$$P(X \in E) \geq 0 \forall E \in \mathcal{A}$$

Where  $\mathcal{A}$  is the input space, similar to a sample space.

Axiom 2

$$P(X \in E) \leq 1 \forall E \in \mathcal{A}$$

Axiom 3

$$P(X \in \bigsqcup_{i \in I} E_i) = \sum_i P(X \in E_i)$$

**Cites:** [Wikipedia](#) ; Wikidata ; Wolfram

[Chapter 4 Distributions](#) (rossIntroductionProbabilitySimulation?) Where  $\bigsqcup$  is the symbol for [disjoint union](#).

# Random Variable

**Instance of:** measurable function

**AKA:** random quantity, aleatory variable, or stochastic variable

**Distinct from:**

**English:** A variable which takes values from a sample space, where a probability distribution describes which values/sets of values are more likely to be taken.

**Formalization:**

A random variable is just a function mapping outcomes to some measurement space.

$$X : \Omega \mapsto E$$

$$P(X \in S) = P(\omega \in \Omega | X(\omega) \in S)$$

**Cites:** [Wikipedia](#) ; Wikidata ; Wolfram

**Code**

**R**

Examples:

**Python**

Examples:

The measurement space is usually the reals,  $\mathbb{R}$ . The outcomes are formally supposed to be [probability spaces](#), which are defined as triples  $(\omega, \mathcal{F}, P)$ . Where  $\mathcal{F}$  could be sets of more than 1 of the possible outcomes, and  $P$  maps each possible set to a 0-1 probability.

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

# Probability Mass/Density Function (PMF/PDF)

**Instance of:** function

**AKA:** Discrete density function; density; Probability Function

**Distinct from:**

**English:** A function that takes in a value, and returns the relative likelihood that a random variable takes on that value. Probability mass functions refer to discrete distributions, e.g. what is the probability that a 6 sided die lands on 5? Probability density functions refer to continuous probability distributions and are usually discussed in terms of ranges or cut points, e.g., what is the probability that a sample from a random normal value is above 2?

**Formalization:**

In the continuous case, the probability that a random variable  $X$  takes on a value between  $a$  and  $b$  is the integration of its density over that range.

$$Pr[a \leq X \leq b] = \int_a^b f_x(x)dx$$

**Cites:** [Wikipedia](#) ; Wikidata ; Wolfram

**Code**

**R**

Examples:

Where  $a$  and  $b$  are the range of values, and  $f_x$  is the density of the random variable.

## Python

Examples:

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## **Part XIII**

# **Moments of a Distribution**



# Mean

## Introduction

**Measure of:** Central tendency

## Frequentist

**AKA:** Arithmetic mean; average;  $\bar{x}$  (sample mean);  $\mu$  (population mean);  $\mu_x$  (population mean)

**Distinct from:** Geometric mean (GM); Harmonic mean (HM); generalized mean/ Power mean; weighted arithmetic mean

**English:** Take a list of numbers, sum those numbers, and then divide by the number of numbers.

**Formalization:**

$$\bar{x} = \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \dots + x_n}{n}$$

**Cites:** [Wikipedia](#) ; [Wikidata](#) ; [Wolfram](#)

**Code**

**R**

**Documentation:** [mean](#); [Arithmetic Mean](#)

Examples:

```
x = c(1,2,3,4)
x
```

```
[1] 1 2 3 4
```

```
#Algorithm
x_bar = sum(x, na.rm=T)/length(x)
x_bar
```

```
[1] 2.5
```

```
#Base Function
x_bar = mean(x, na.rm=T)
x_bar
```

```
[1] 2.5
```

## Python

Documentation: [numpy.mean](#)

Examples:

```
x = [1,2,3,4]
print(x)
```

```
[1, 2, 3, 4]
```

```
#Algorithm
x_bar= sum(x)/len(x)
x_bar
```

```
2.5
```

```
#statistics Function
import statistics
x_bar = statistics.mean(x)
x_bar
```

```
2.5
```

```
#scipy Function
#<string>:1: DeprecationWarning: scipy.mean is deprecated and will be removed in SciPy 2.0.0
import scipy
x_bar = scipy.mean(x)
```

<string>:1: DeprecationWarning: scipy.mean is deprecated and will be removed in SciPy 2.0.0, us

```
x_bar
```

2.5

```
#numpy Function
import numpy as np
x = np.array(x)
x_bar = x.mean()
x_bar
```

2.5

## SQL

**Documentation:** [PostgreSQL AVG Function](#)

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
```

```
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

```
DROP TABLE IF EXISTS t1;
```

```
CREATE TABLE IF NOT EXISTS t1 (
  id serial PRIMARY KEY,
  amount INTEGER
);
```

```
INSERT INTO t1 (amount)
VALUES
  (10),
  (NULL),
  (30);
```

```
SELECT * FROM t1;
```

Table 6: 3 records

id	amount
1	10
2	NA
3	30

```
SELECT AVG(amount)::numeric(10,2)
FROM t1;
```

Table 7: 1 records

avg
20

## Torch

```
import torch
```

## Bayesian

[Bayesian average; Solving an age-old problem using Bayesian Average](#); [Of bayesian average and star ratings](#); [Bayesian Average Ratings](#) ;

**English:** The Bayesian average is the weighted average of a prior and the observed sample average. When would you want this? When you have strong beliefs about the true mean, or when sample size is too small to reliably calculate a mean. For example a movie rating website where a movie may have only a single 5 star rating and so would rank higher than the Godfather with over a 100 almost all 5 star ratings.

**Formalization:**

$$\bar{x} = \frac{C * m + (\sum_{i=1}^n x_i)}{c + n}$$

Where  $m$  is a prior for true mean, and  $C$  is a constant representing how many elements would be necessary to reliably estimate a sample mean.

**Code**

**Part XIV**

**Distributions**

# **Part XV**

## **Information**

# Entropy



**Part XVI**

**Inference**

# Bayesianism

# Frequentism

AKA:

[Introduction to the Concept of Likelihood and Its Applications](#)(Etz 2017)

# **Part XVII**

## **Estimation**

## Performance

# Out of Sample Performance

Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, Carsten F. Dormann <https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.02881>

[Cross-validation FAQ, Aki Vehtari](#)

[When to Impute? Imputation before and during cross-validation](#)

[Leakage and the Reproducibility Crisis in ML-based Science](#)

[Rescaling and other forms of unsupervised preprocessing introduce bias into cross-validation, Amit Moscovich, Saharon Rosset](#)

[Approximate leave-future-out cross-validation for Bayesian time series models](#)

[groupdata2](#)

[How Cross-Validation Can Go Wrong and What to Do About it.](#)

[Moving cross-validation from a research idea to a routine step in Bayesian data analysis](#)

[Model selection tutorials and talks, Aki Vehtari Underspecification Presents Challenges for Credibility in Modern Machine Learning \(Paper Explained\) -Has a neat example of holding out on camera model shows massive degradation in medical mission vision application.](#)

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Cross-validation: what does it estimate and how well does it do it? Stephen Bates, Trevor Hastie, Robert Tibshirani

Consensus features nested cross-validation

Your Cross Validation Error Confidence Intervals are Wrong — here's how to Fix Them

# Regularization

Ridge Regression Can Produce Misleading Inferences in the Presence of Strong Confounders: The Case of Mass Polarization

Fast Penalized Regression and Cross Validation for Tall Data with the oem Package



# P Values

Bayesian estimation supersedes the t test

Sequential sampling and testing Safe, anytime-valid inference:  
confidence sequences, p-values/e-values, and e-processes

Some papers about p values

The Difference Between “Significant” and “Not Significant” is  
not Itself Statistically Significant

# Bias Variance Tradeoff

[On the Bias-Variance Tradeoff: Textbooks Need an Update  
A Modern Perspective Posted on January 5, 2020](#)(Neal 2020b)

TL;DR: It is not always necessary to trade bias for variance when increasing model complexity.

## Bias

## Variance

# Asymptotics

## Introduction

**Instance of:** mathematical analysis

## Frequentist

**AKA:** asymptotic analysis

**Distinct from:**

**English:** Describing the behavior of the mathematical system in the limit, e.g. as an index goes to infinity.

**Formalization:**

**Cites:** [Wikipedia](#) ; [Wikidata](#) ; Wolfram

**Code**

**R**

Examples:

**Python**

Examples:

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Bayesian

English: Formalization:

Cites:

**Code**

## **Part XVIII**

# **Domain / Generalizability / Transportability**



Surveys (Degtiar and Rose 2023)

CHANNELLING FISHER: RANDOMIZATION TESTS AND  
THE STATISTICAL INSIGNIFICANCE OF SEEMINGLY  
SIGNIFICANT EXPERIMENTAL RESULTS

An Automatic Finite-Sample Robustness Metric: When Can  
Dropping a Little Data Make a Big Difference?

Outlier

# Outliers

An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?

## Regime Change

## **Internal Validity**

## **Transportability**

## **External Validity**

# Matching

AKA: Subclassification

[5 Matching and Subclassification](#)(Cunningham 2021)

[Chapter 14 - Matching](#)(([huntington-kleinEffectIntroductionResearch?](#)))

## Poststratification



# Outcome regressions / Response Surface Modeling

(Degtiar and Rose 2023) “fit an outcome regression in study sample data to estimate conditional means, then obtain PATEs by marginalizing over (i.e., standardizing to) the target sample covariate distribution by predicting counterfactuals for the target sample If the target sample is not a simple random sample from the target population, this would be a weighted average using sampling weights (Kim et al., 2018).”

## **Part XIX**

# **Likelihood**

# Likelihood Function

**Instance of:** joint probability

**AKA:** Likelihood

**Distinct from:**

**English:** The joint probability of the data, conditional on the parameters. How likely are we to observe these data, given this parameter is the True one.

(Etz 2017)

- Likelihood is not probability, but is proportional to a probability.
- Likelihoods are relative, scaled by an arbitrary constant, and need not sum to one.
- Under likelihood the data are fixed, and the hypothesis vary.

**Formalization:**

$$\mathcal{L}(\theta|X)$$

Also written in terms of a probability of observing  $X$  given the a parameter value

Where  $\theta$  are the parameters, and  $X$  is the evidence.

$$\mathcal{L}(\theta) = K \times P(X|\theta)$$

**Cites:** [Wikipedia](#) ; Wikidata ; Wolfram

**Code**

Where  $\theta$  are the parameters, and  $X$  is the evidence, and  $K$  is an arbitrary scaling constant.

## R

Examples:

## Python

Examples:

## SQL

```
library(DBI)
# Create an ephemeral in-memory RSQLite database
#con <- dbConnect(RSQLite::SQLite(), dbname = ":memory:")
#dbListTables(con)
#dbWriteTable(con, "mtcars", mtcars)
#dbListTables(con)

#Configuration failed because libpq was not found. Try installing:
## deb: libpq-dev libssl-dev (Debian, Ubuntu, etc)
#install.packages('RPostgres')
#remotes::install_github("r-dbi/RPostgres")
#Took forever because my file permissions were broken
#pg_lsclusters
require(RPostgres)
```

Loading required package: RPostgres

```
# Connect to the default postgres database
#I had to follow these instructions and create both a username and database that matched my
#https://www.digitalocean.com/community/tutorials/how-to-install-postgresql-on-ubuntu-20-04-
con <- dbConnect(RPostgres::Postgres())
```

## Torch

```
import torch
```

## Maximum Likelihood Estimation (MLE)

## **Part XX**

# **——-Measurement——-**

## **Part XXI**

# **Data and Measurement**



# Data Validity

English: A measure is valid to the degree that it represents what you are trying to measure. The validity of a measuring process can be measured as representing what you're trying to measure in some aggregate, e.g. on average (Gelman, Hill, and Vehtari 2020).

# Data Reliability

English: A reliable measure is one that is precise and stable, when you measure the same thing again you get the same answer (even it's wrong, you get the same wrong answer again) (Gelman, Hill, and Vehtari 2020).

**Part XXII**

**Research Design**

# Research Design Directed Acyclic Graphs (DAGs)

[Directed Acyclic Graphs](#)(Cunningham 2021)

(Pearl 2018)

[3.1 - Graphical Models \(Intro and Outline\)](#)(“3.1 - Graphical Models (Intro and Outline) - YouTube” n.d.) [The Flow of Association and Causation in Graphs](#), Brady Neal(Neal, n.d.a)

[Single World Intervention Graphs \(SWIGs\): A Unification of the Counterfactual and Graphical Approaches to Causality](#)(Richardson and Robins 2013)

[4.7 - Structural Causal Models SCMs](#)(Brady Neal - Causal Inference 2020b)

(Janzing et al. 2013)

(Pearl 2019)

(Pearl 2015)

# Potential Outcomes

[4 Potential Outcomes Causal Model](#)(Cunningham 2021)

(Hernán and Taubman 2008)


[2.1 - What are Potential Outcomes?](#)(Brady Neal - Causal Inference 2020a)

[Potential Outcomes, Brady Neal](#)(Neal, n.d.b)

## **Project Design**

# Unit of Analysis

Ecological Inference in the Social Sciences <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885825/>

Race to the bottom: Spatial aggregation and event data Scott J. Cook  Nils B. Weidmann  
<https://www.tandfonline.com/doi/abs/10.1080/03050629.2022.2025365>

Extremal Behavior of Aggregated Data with an Application to Downscaling Sebastian Engelke, Raphael de Fondeville, Marco Oesting <https://arxiv.org/abs/1712.09816>

# Estimand

What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory



# Identification

[Partial Identification in Econometrics](#)

Identity Crisis [https://betanalpha.github.io/assets/case\\_studies/identifiability.html](https://betanalpha.github.io/assets/case_studies/identifiability.html)

[Identifiability of Path-Specific Effects](#)(Avin, Shpitser, and Pearl, n.d.)

# Garden of Forking Paths

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time\*, Andrew Gelman† and Eric Loken

Achieving Statistical Significance with Covariates and without Transparency

# Confounder

Confounder Selection: Objectives and Approaches (Guo, Lundborg, and Zhao 2022)

# Unobserved Confounding

AKA: Sensitivity Analysis

(Veitch and Zaveri 2020)

[A causal framework for distribution generalization](#)(**A?** causal framework for distribution generalization)

[Invariant Risk Minimization](#)(Arjovsky et al. 2020)

## **Part XXIII**

# **Prediction/Forecasting**

## **Part XXIV**

# **Counterfactual causal inference**

**Part XXV**

**Causality the 12  
Assumptions**

**No unmeasured confounders**



## **Correct model specification**

**No conditioning on a collider**

## No conditioning on Mediators

# Positivity

## Consistency

## No Interference

## **No Relevant Effect Modification**

## **Collapsibility**



## Compliance

## Missing Data Mechanism

# Transparency

## Recoverability

## Testability

## **No Relevant Measurement Error**

**Part XXVI**

**Exogeneity**

# Random Control Trials

AKA: RCT

Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015

CHANNELLING FISHER: RANDOMIZATION TESTS AND THE STATISTICAL INSIGNIFICANCE OF SEEMINGLY SIGNIFICANT EXPERIMENTAL RESULTS

## Warnings

- Dicing RCT results up by coverates or in with a regression model instead of doing a simple T test can generate spurious results from a few high leverage outlier observations (Young 2019).



# Instrumental Variables

How Much Should We Trust Instrumental Variable Estimates  
in Political Science? Practical Advice Based on Over 60 Repli-  
cated Studies

Chapter 19 - Instrumental Variables(huntington-kleinEffectIntroductionResearch?)

Deep IV: A Flexible Approach for Counterfactual Predic-  
tion(Hartford et al. 2017)

# Difference in Difference

How Much Should We Trust Differences-In-Differences Estimates?

How Much Should We Trust Staggered Difference-In-Differences Estimates?

When Is Parallel Trends Sensitive to Functional Form?\*

Chapter 18 - Difference-in-Differences((**huntington-kleinEffectIntroductionResearch?**))

9 Difference-in-Differences(Cunningham 2021)

(Kahn-Lang and Lang 2018)

## Placebo Tests

Do We Really Know the WTO Cures Cancer? <https://www.cambridge.org/core/journals/british-journal-of-political-science/article/abs/do-we-really-know-the-wto-cures-cancer/B84A6FCF516FAE3ED7E0C20FE3DA42CF>

# Regression Discontinuity (RDD)

6 Regression Discontinuity (Cunningham 2021)

Chapter 20 - Regression Discontinuity (([huntington-kleinEffectIntroductionResearch?](#)))

Regression Discontinuity Designs in Economics (Lee and Lemieux 2010)

# Fixed Effects

[Chapter 16 - Fixed Effects](#)([huntington-kleinEffectIntroductionResearch?](#))

# Synthetic Controls

10 Synthetic Control(Cunningham 2021)

## **Part XXVII**

# **Supervised Learning**

**OLS**



# Interactions

(Lim and Hastie 2015)

(“You Need 16 Times the Sample Size to Estimate an Interaction Than to Estimate a Main Effect | Statistical Modeling, Causal Inference, and Social Science” n.d.)

## Decision Trees

## Random Forest

# Gradientboosting

## Videos

StatQuest with Josh Starmer [Gradient Boost Part 1 \(of 4\): Regression Main Ideas](#) [XGBoost Part 1 \(of 4\): Regression](#)  
[XGBoost LightGBM](#)

# Gaussian Processes

[GPJax](#) “GPJax aims to provide a low-level interface to Gaussian process (GP) models in Jax, structured to give researchers maximum flexibility in extending the code to suit their own needs.”

## **Part XXVIII**

# **Reinforcement Learning**

Overviews \* Reinforcement Learning: A Six Part Series (Mutual Information 2022)

# Reinforcement Learning

Overviews \* Reinforcement Learning: A Six Part Series (Mutual Information 2022)



## **Part XXIX**

# **Unsupervised Learning**

## Distance Metrics

**Part XXX**

# **Neural Networks**

# Neural Network Debugging

“Since this seriously helped a friend the other day: If your network doesn’t learn, reduce the dataset to a small batch (eg 32 examples) and try to overfit on it. The loss should be near zero (because the network should be able to memorize it); if not, there’s a bug in your code.”

“3.1 - Graphical Models (Intro and Outline) - YouTube.” n.d. Accessed November 15, 2022. [https://www.youtube.com/watch?v=Go4EkHN\\_PcA&list=PLoazKTcS0Rzb6bb9L508cyJ1z-U9iWkA0&index=20](https://www.youtube.com/watch?v=Go4EkHN_PcA&list=PLoazKTcS0Rzb6bb9L508cyJ1z-U9iWkA0&index=20).

Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. “Invariant Risk Minimization.” arXiv. <https://doi.org/10.48550/arXiv.1907.02893>.

Avin, Chen, Ilya Shpitser, and Judea Pearl. n.d. “Identifiability of Path-Specific Effects,” 7.

Brady Neal - Causal Inference, dir. 2020a. *2.1 - What Are Potential Outcomes?* <https://www.youtube.com/watch?v=q8x9aetyok0>.

———, dir. 2020b. *4.7 - Structural Causal Models SCMs*. <https://www.youtube.com/watch?v=dQeRqb0N6gs>.

“Computational Linear Algebra - YouTube.” n.d. Accessed November 15, 2022. <https://www.youtube.com/playlist?list=PLtmWHNX-gukIc92m1K0P6bIONZb-mg0hY>.

Congdon, Peter. 2014. *Applied Bayesian Modelling*. John Wiley & Sons. <https://books.google.com?id=ImejAwAAQBAJ>.

Cunningham, Scott. 2021. *Causal Inference*. Yale University Press. <https://books.google.com?id=DZ4REAAAQBAJ>.

Degtiar, Irina, and Sherri Rose. 2023. “A Review of Generalizability and Transportability.” *Annual Review of Statistics and Its Application* 10 (1): annurev-statistics-042522-103837. <https://doi.org/10.1146/annurev-statistics-042522-103837>.

- Etz, Alexander. 2017. “Introduction to the Concept of Likelihood and Its Applications.” PsyArXiv. <https://doi.org/10.31234/osf.io/85ywt>.
- Fastai/Numerical-Linear-Algebra. (2017) 2022. fast.ai. <https://github.com/fastai/numerical-linear-algebra>.
- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. 1st ed. Cambridge University Press. <https://doi.org/10.1017/9781139161879>.
- Gelman, Andrew, and Aki Vehtari. 2021. “What Are the Most Important Statistical Ideas of the Past 50 Years?” arXiv. <http://arxiv.org/abs/2012.00174>.
- Guo, F. Richard, Anton Rask Lundborg, and Qingyuan Zhao. 2022. “Confounder Selection: Objectives and Approaches.” arXiv. <https://doi.org/10.48550/arXiv.2208.13871>.
- Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. “Deep IV: A Flexible Approach for Counterfactual Prediction.” In *Proceedings of the 34th International Conference on Machine Learning*, 1414–23. PMLR. <https://proceedings.mlr.press/v70/hartford17a.html>.
- Hernán, M. A., and S. L. Taubman. 2008. “Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions.” *International Journal of Obesity* 32 (3, 3): S8–14. <https://doi.org/10.1038/ijo.2008.82>.
- Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2013. “Quantifying Causal Influences.” *The Annals of Statistics* 41 (5). <https://doi.org/10.1214/13-AOS1145>.
- Kahn-Lang, Ariella, and Kevin Lang. 2018. “The Promise and Pitfalls of Differences-in-Differences: Reflections on ‘16 and Pregnant’ and Other Applications.” Working Paper. Working Paper Series. National Bureau of Economic Research. <https://doi.org/10.3386/w24857>.
- Lee, David S, and Thomas Lemieux. 2010. “Regression Discontinuity Designs in Economics.” *Journal of Economic Literature* 48 (2): 281–355. <https://doi.org/10.1257/jel.48.2.281>.
- Lim, Michael, and Trevor Hastie. 2015. “Learning Interactions via Hierarchical Group-Lasso Regularization.” *Journal of Computational and Graphical Statistics* 24 (3): 627–54. <https://doi.org/10.1080/10618600.2014.938812>.

- Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86 (3): 532–65. <https://doi.org/10.1177/00031224211004187>.
- Mutual Information, dir. 2022. *Reinforcement Learning: A Six Part Series*. [https://www.youtube.com/watch?v=NFo9v\\_yKQXA](https://www.youtube.com/watch?v=NFo9v_yKQXA).
- Neal, Brady. 2019. “Which Causal Inference Book You Should Read.” November 23, 2019. <https://www.bradyneal.com/which-causal-inference-book>.
- . 2020a. “Introduction to Causal Inference from a Machine Learning Perspective.” *Course Lecture Notes (Draft)*. [https://www.bradyneal.com/Introduction\\_to\\_Causal\\_Inference-Dec17\\_2020-Neal.pdf](https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf).
- . 2020b. “On the Bias-Variance Tradeoff: Textbooks Need an Update (Blog Post).” January 5, 2020. <https://www.bradyneal.com/bias-variance-tradeoff-textbooks-update>.
- . n.d.a. “Graph Terminology Bayesian Networks and Causal Graphs The Basic Building Blocks of Graphs The Flow of Association and Causation,” 141.
- . n.d.b. “What Are Potential Outcomes? The Fundamental Problem of Causal Inference Getting Around the Fundamental Problem of Causal Inference A Complete Example with Estimation,” 198.
- Pearl, Judea. 2015. “TRYGVE HAAVELMO AND THE EMERGENCE OF CAUSAL CALCULUS.” *Econometric Theory* 31 (1): 152–79. <https://doi.org/10.1017/S0266466614000231>.
- . 2018. “Does Obesity Shorten Life? Or Is It the Soda? On Non-manipulable Causes.” *Journal of Causal Inference* 6 (2): 20182001. <https://doi.org/10.1515/jci-2018-2001>.
- . 2019. “On the Interpretation of Do(x).” *Journal of Causal Inference* 7 (1). <https://doi.org/10.1515/jci-2019-2002>.
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited. <https://books.google.com?id=EmY8DwAAQBAJ>.
- Pham, Khang. (2020) 2022. *Minimum Viable Study Plan for Machine Learning Interviews*. <https://github.com/>

- [khangich/machine-learning-interview](#).
- Richardson, Thomas S., and James M. Robins. 2013. “Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality.” *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128 (30): 2013.
- Ross, Kevin. 2022. *An Introduction to Probability and Simulation*. [https://bookdown.org/kevin\\_davisross/probsim-book/](https://bookdown.org/kevin_davisross/probsim-book/).
- Schomaker, Michael. 2021. “Regression and Causality.” arXiv. <http://arxiv.org/abs/2006.11754>.
- Torfi, Amirshina, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. 2021. “Natural Language Processing Advancements By Deep Learning: A Survey.” arXiv. <https://doi.org/10.48550/arXiv.2003.01200>.
- Veitch, Victor, and Anisha Zaveri. 2020. “Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding.” arXiv. <https://doi.org/10.48550/arXiv.2003.01747>.
- “You Need 16 Times the Sample Size to Estimate an Interaction Than to Estimate a Main Effect | Statistical Modeling, Causal Inference, and Social Science.” n.d. Accessed November 11, 2022. <https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/>.
- Young, Alwyn. 2019. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results\*.” *The Quarterly Journal of Economics* 134 (2): 557–98. <https://doi.org/10.1093/qje/qjy029>.