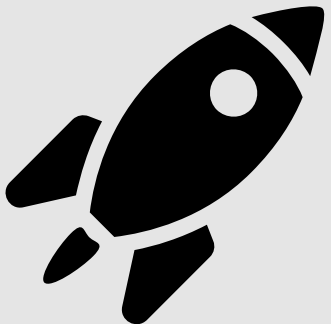


Frontiers of Statistics

BERDC Special Topics Talk 11



DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson
Biostatistics, Epidemiology,
and Research Design Core

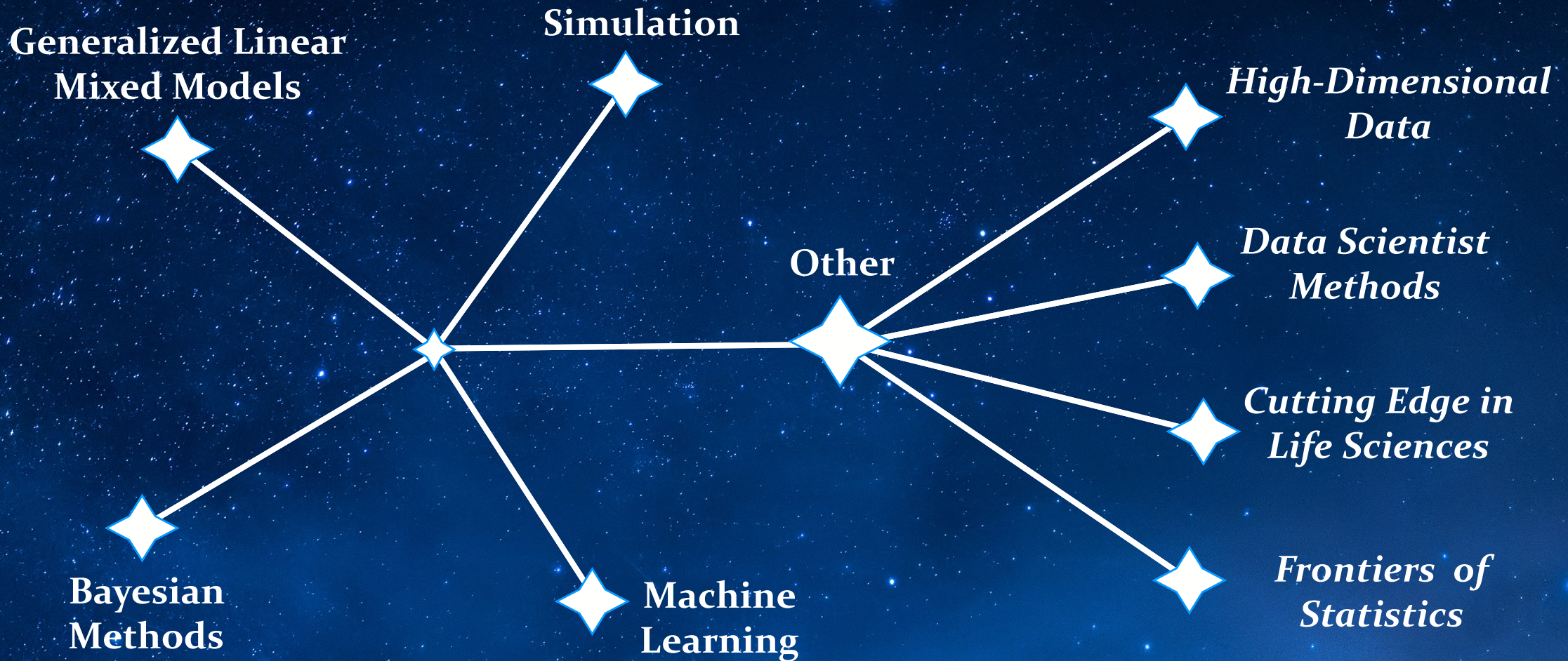
Opening

Goal: Explore the frontiers of statistical methods and techniques

- ◆ Go over some of the current state-of-the-art procedures
- ◆ Stare down looming problems
- ◆ Describe a framework of potential solutions
- ◆ Equip ourselves with resources
- ◆ Strike out towards new statistical worlds through scientific paper examples



Current Status



Current Status

◆ Generalized Linear Mixed Models

- ◆ Ordinary least squares → Maximum Likelihood [1]
- ◆ **Generalized:** beyond Gaussian distribution (Poisson, Beta, Logistic, Negative Binomial, etc.)
- ◆ **Mixed:** Including fixed and random effects
- ◆ Everything in one place (GLMMs for Everything)

◆ Bayesian Methods

- ◆ Does not depend on sample size
- ◆ Always open to new data
- ◆ Modern computing allows for posterior simulation (MCMC)
- ◆ Lots of activity (Bayesian Analysis Modules)
- ◆ For every frequentist method, there is usually a Bayesian equivalent

◆ Simulation

- ◆ Modern computing allows for powerful uses
- ◆ Can generate confidence intervals by bootstrapping, use MC to investigate the performance of statistical procedures, and generate estimate posterior distributions [2]
- ◆ More simply, can create mock data with known characteristics
- ◆ Also, a core component of resampling methods [3]

◆ Machine Learning

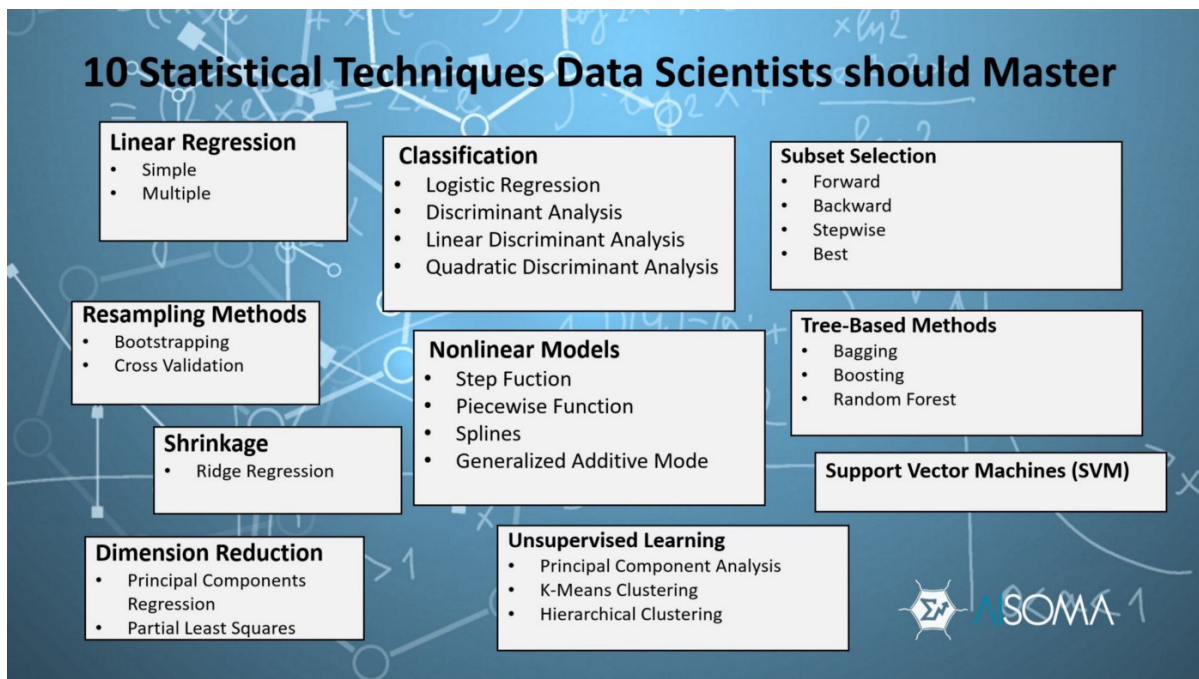
- ◆ Branch of AI focusing on use of data and algorithms to improve with more data
- ◆ Notably used in classification, prediction, and data mining [4-5]
- ◆ Again, lots of activity (future Special Topics talk planned for later this year)

Current Status

◆ High-dimensional data

- ◆ Many more variables than observations (no-no in standard regression)
- ◆ **Examples:** healthcare data, financial data, genomics [6]
- ◆ **Methods:** Ridge, Lasso, & Principal Components regression, etc. [7]

◆ Statistical Techniques Data Scientists should Master [8]



◆ Cutting-Edge Statistical Methods for a Life-Course Approach [9]

- ◆ Regression with covariates
- ◆ Hazard modeling
- ◆ Individual growth modeling
- ◆ Structural equation modeling
- ◆ Propensity score analysis
- ◆ Degression discontinuity analysis

◆ Frontiers of Statistics [10]

- ◆ Semiparametric Modeling
- ◆ Nonparametric Models
- ◆ Statistical Learning and Bootstrap
- ◆ Longitudinal Data Analysis
- ◆ Statistics in Science and Technology
- ◆ Financial Econometrics
- ◆ Parametric Techniques and Inferences

Looming Problems



**Overly Simple / Outdated
Models**



Large Datasets



**Replication and
Significance**



Computational Limits

Looming Problems

◆ Overly Simple / Outdated Models

- ◆ Basic statistics where better modeling would be helpful
- ◆ ‘Nothing is really normally distributed’
- ◆ Lots of noise or small signal needs careful parsing
- ◆ Research Question -> Statistical Model -> Computer coding problem

◆ Large Datasets [11-12]

- ◆ High dimensionality and large sample size (computer)
- ◆ Multiple comparisons / spurious correlations
- ◆ Multiple aggregations -> heterogeneity and experimental design
- ◆ Coherent stories and results (metagenomics problems) -> use imagination

◆ Replication and Significance

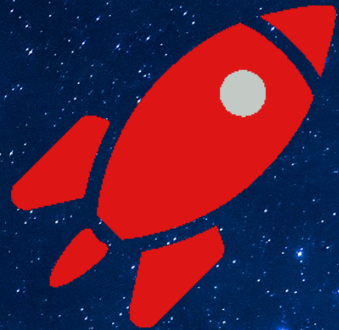
- ◆ Replication crisis
- ◆ Large random errors
- ◆ Underpowering
- ◆ Problems with p-values [13]

◆ Computational Limits [14]

- ◆ Large, complex datasets are becoming common
- ◆ Even with powerful computing, some problems take too long for a brute solution (ex. phylogenetics)
- ◆ Other take a lot of time to train (AI models)
- ◆ Hardware supplies are also limited
- ◆ Likely a shift from computing power to algorithmic innovation

Potential Solutions

Change Framework



New Models



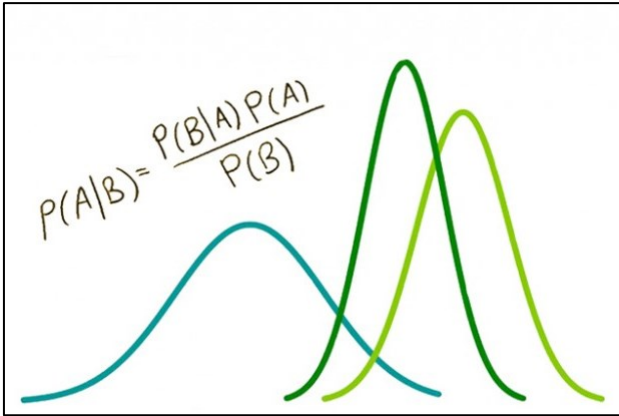
Better Models



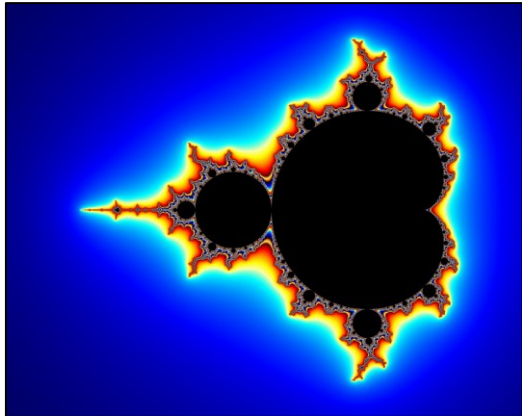
Potential Solutions

◆ Change Framework

- ◆ Bayes
- ◆ Mandelbrot



[15]



[16]



[17]

◆ Better Models

- ◆ Simplify -> GLMMs, better experimental design, model selection
- ◆ Complexify -> nonlinear, multivariate, machine learning, etc.
- ◆ Adapt -> take methods and approaches from other disciplines and use them in your own

◆ New Models

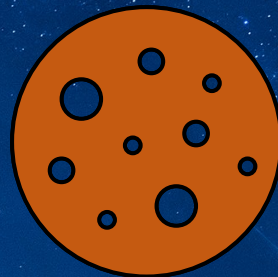
- ◆ Will depend on the specific problems and discipline
- ◆ Keep on lookout in journals & other areas
- ◆ The real frontier...

Resources



Journals

**Websites/
Videos**



Books

Papers



Resources

♦ Journals

- ♦ The R Journal
- ♦ Journal of Statistical Software
- ♦ Frontiers in Applied Mathematics and Statistics

♦ Websites/Videos

- ♦ Big Problems in Statistics [18]
- ♦ Solve Every Statistics Problem with One Weird Trick [3]

♦ Books

- ♦ An Introduction to Statistical Learning [19]
- ♦ The Elements of Statistical Learning [20]
- ♦ Frontiers of Statistics [10]

♦ Papers

- ♦ What are the Open Problems in Bayesian Statistics? [21]
- ♦ Cutting-Edge Statistical Methods for a Life-Course Approach [9]
- ♦ Multilevel Methods and Statistics: The Next Frontier [22]
- ♦ SSP: an R package to estimate sampling effort in studies of ecological communities [23]
- ♦ New Models and Methods for Applied Statistics: Topics in Computer Experiments and Time Series Analysis [24]
- ♦ Gamma-ray blazer variability: new statistical methods of time-flux distributions [25]
- ♦ New Author Guidelines for Displaying Data and Reporting Data Analysis and Statistical Methods in Experimental Biology [26]


Examples

The  Journal
Volume 13/2, December 2021

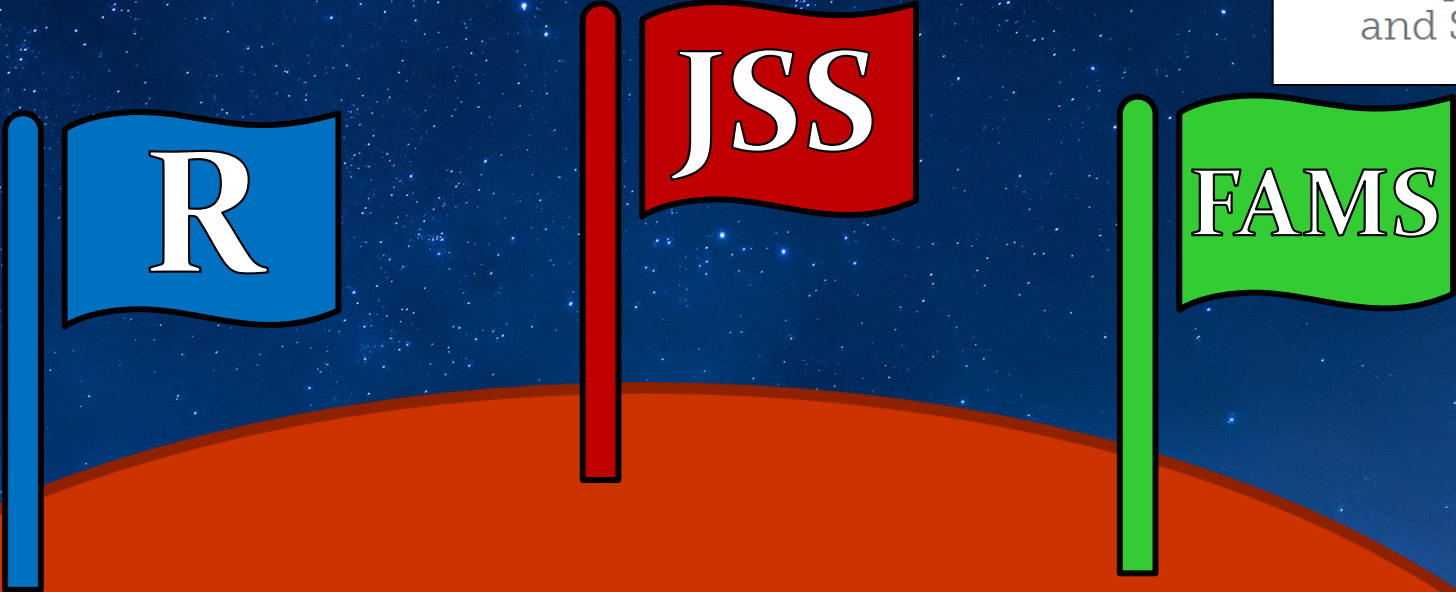
A peer-reviewed, open-access publication of the
R Foundation for Statistical Computing



Journal of Statistical Software
January 2022, Volume 101, Issue 1. doi: 10.18637/jss.v101.i01

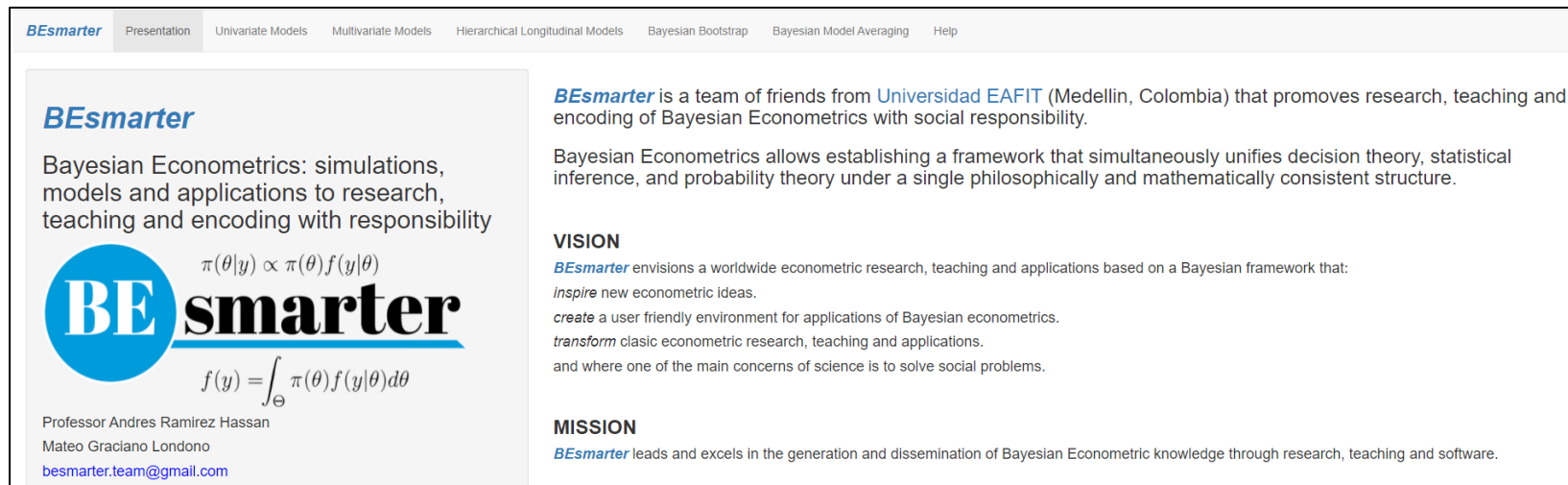


frontiers
in Applied Mathematics
and Statistics



Examples: R Journal

◆ GUIDed tour of Bayesian regression [27]



The screenshot shows the BEsmarter website with a navigation bar at the top containing links for Presentation, Univariate Models, Multivariate Models, Hierarchical Longitudinal Models, Bayesian Bootstrap, Bayesian Model Averaging, and Help. The main content area is divided into two columns. The left column features the BEsmarter logo, which includes the text $\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$ and $f(y) = \int_{\Theta} \pi(\theta)f(y|\theta)d\theta$, along with the names of Professor Andres Ramirez Hassan and Mateo Graciano Londono, and the email address besmarter.team@gmail.com. The right column contains a description of BEsmarter as a team from Universidad EAFIT, a paragraph about Bayesian Econometrics, a VISION section with three bullet points, and a MISSION section.

BEsmarter

Bayesian Econometrics: simulations, models and applications to research, teaching and encoding with responsibility

$\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$

BE smarter

$f(y) = \int_{\Theta} \pi(\theta)f(y|\theta)d\theta$

Professor Andres Ramirez Hassan
Mateo Graciano Londono
besmarter.team@gmail.com

BEsmarter is a team of friends from **Universidad EAFIT** (Medellin, Colombia) that promotes research, teaching and encoding of Bayesian Econometrics with social responsibility.

Bayesian Econometrics allows establishing a framework that simultaneously unifies decision theory, statistical inference, and probability theory under a single philosophically and mathematically consistent structure.

VISION

BEsmarter envisions a worldwide econometric research, teaching and applications based on a Bayesian framework that:

- inspire* new econometric ideas.
- create* a user friendly environment for applications of Bayesian econometrics.
- transform* classic econometric research, teaching and applications.

and where one of the main concerns of science is to solve social problems.

MISSION

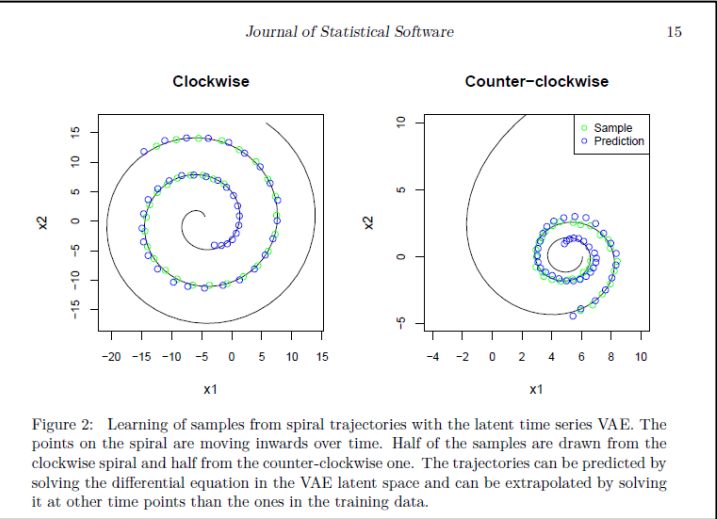
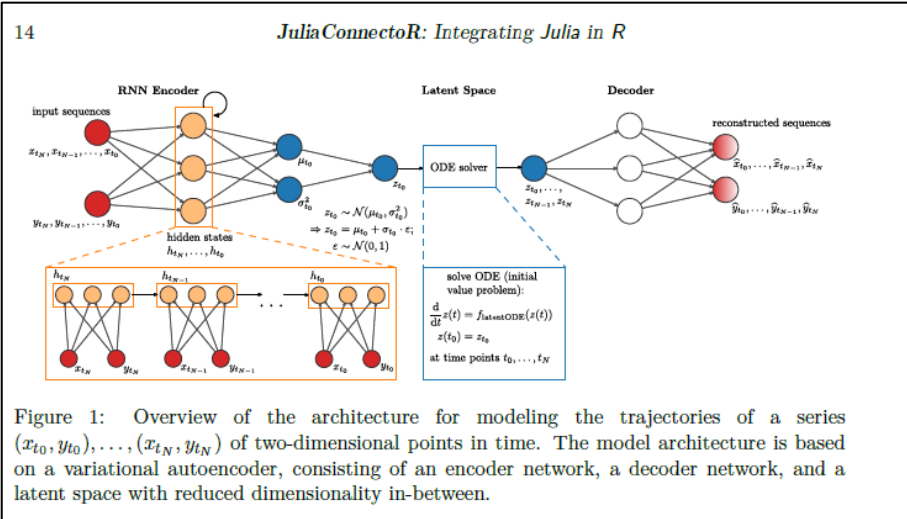
BEsmarter leads and excels in the generation and dissemination of Bayesian Econometric knowledge through research, teaching and software.

◆ We Need Trustworthy R Packages [28]

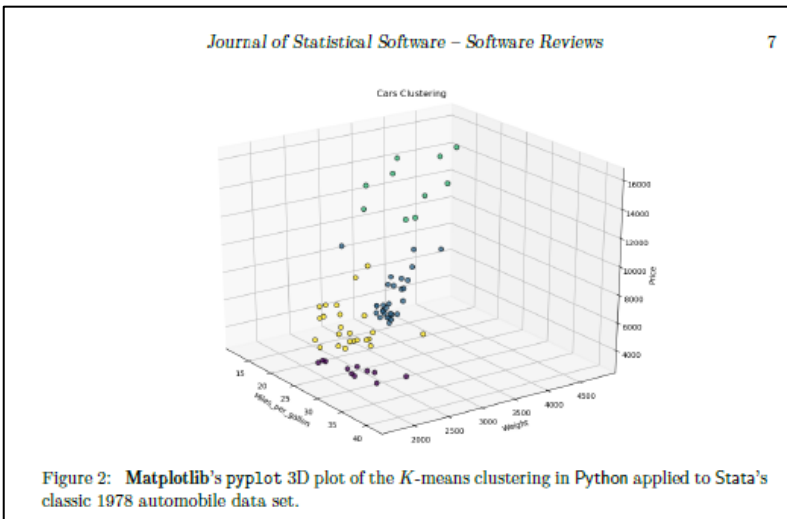
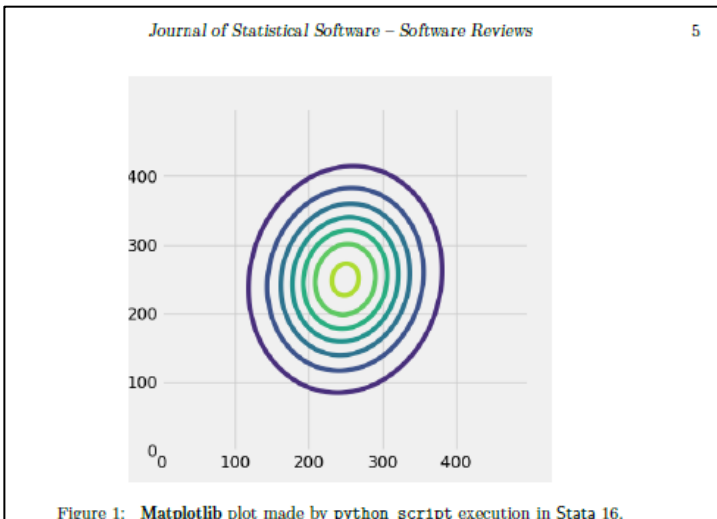
“There is a need for rigorous software engineering in R packages, and there is a need for new research to bridge scientific computing with more traditional computing. Automated tools, interdisciplinary graduate courses, code reviews, and a welcoming developer community will continue to democratize best practices. Democratized software engineering will improve the quality, correctness, and integrity of scientific software, and by extension, the disciplines that rely on it.”

Examples: JSS

◆ The JuliaConnectoR: A Functionally-Oriented Interface for Integrating Julia in R [29]



◆ Data Science in Stata 16: Frames, Lasso, and Python Integration [30]



Examples: FAMS

◆ Dawoud-Kibria Estimator for Beta Regression Model: Simulation and Application [31]

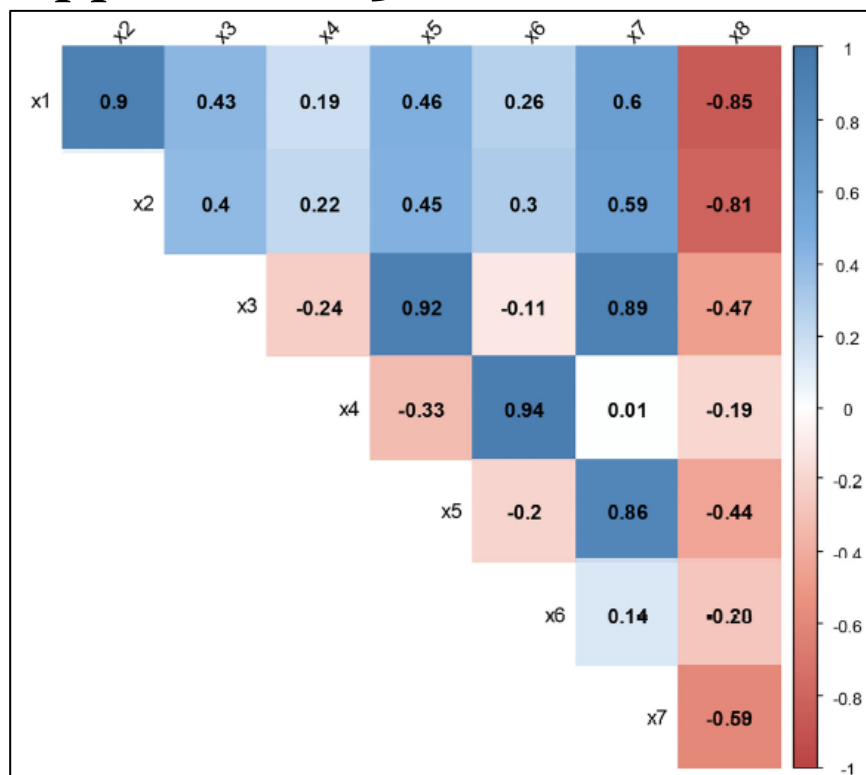


FIGURE 1 | Visualization of the correlation matrix.

◆ Editorial on Machine Learning in Natural Complex Systems [32]

TABLE 1 | List of commonly used non-metric proximity measures in various domains.

Measure	Application field
Dynamic Time Warping (DTW) (6)	Time series or spectral alignment
Inner distance (7)	Shape retrieval e.g., in robotics
Compression distance (8)	Generic used also for text analysis
Smith Waterman Alignment (5)	Bioinformatics
Divergence measures (9)	Spectroscopy and audio processing
Generalized Lp norm (10)	Time series analysis
Non-metric modified Hausdorff (11)	Template matching
(Domain-specific) alignment score (12)	Mass spectrometry

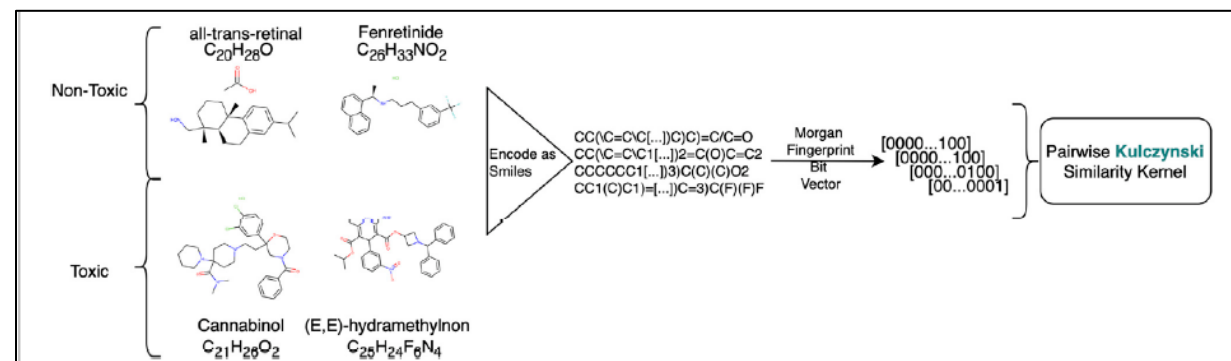
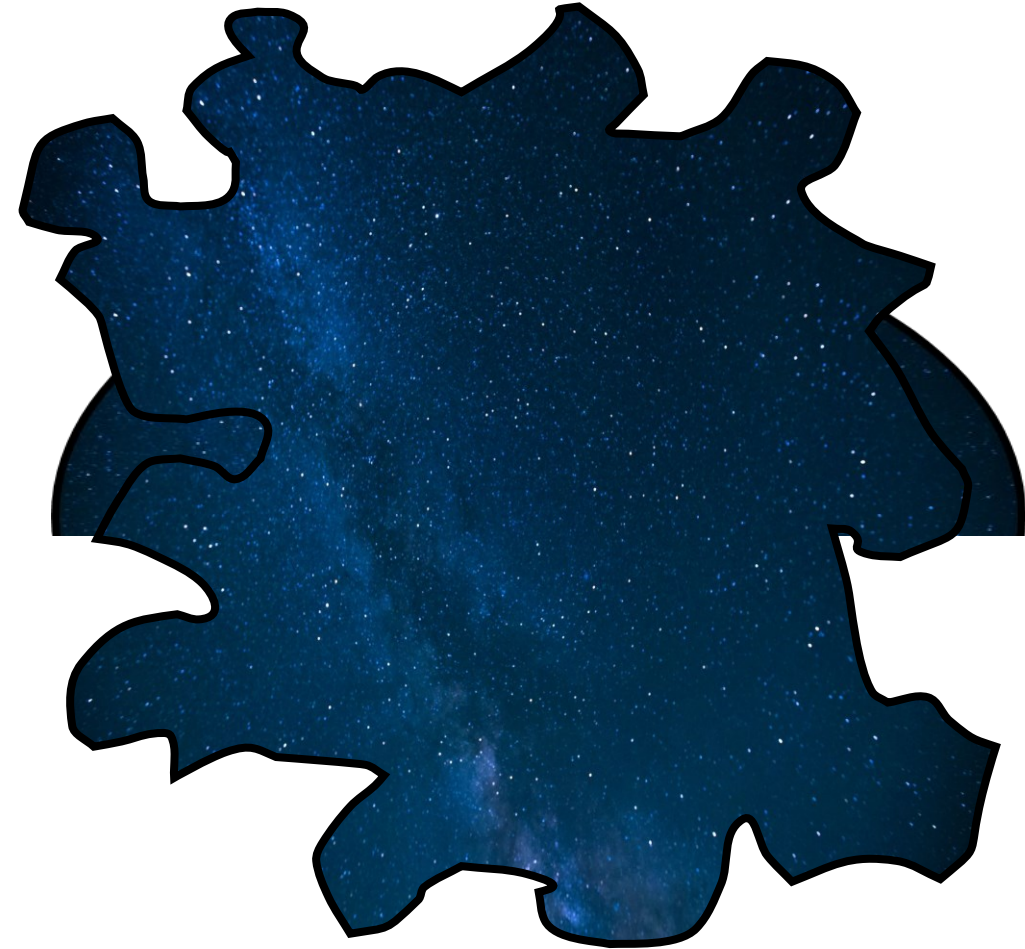
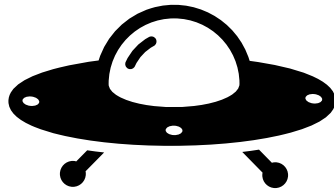


FIGURE 2 | Preprocessing workflow for creating the Tox-21 datasets. Chemicals represented as SMILE codes are translated to Morgan Fingerprints. The kernel is created by using an application related pairwise similarity measure on the Morgan Fingerprints, in this case so-called *Kulczynski*.

Conclusions

- ◆ Statistics is a wide topic used in many applications
- ◆ There is no, single, monolithic frontier
- ◆ Rather, there are frontiers on many edges
- ◆ Whether reframing, modifying, or developing new models, there are a lot of exciting possibilities
- ◆ Start your journey today



References 1

- [1] [http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes MSM.pdf](http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes_MSM.pdf)
- [2] <https://statpower.net/Content/MLRM/Lecture%20Slides/SimulationIntro.pdf>
- [3] <https://www.youtube.com/watch?v=BhY-un6JURA>
- [4] <https://www.ibm.com/cloud/learn/machine-learning>
- [5] <https://online.stanford.edu/courses/sohs-ystatslearning-statistical-learning>
- [6] <https://www.statology.org/high-dimensional-data/>
- [7] [http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes MSM.pdf](http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes_MSM.pdf)
- [8] <https://medium.com/nerd-for-tech/10-statistical-techniques-data-scientists-should-master-c52772f8e1cc>
- [9] <https://pubmed.ncbi.nlm.nih.gov/24425722/>
- [10] <https://stat.princeton.edu/frontiers/Content.pdf>
- [11] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236847/>
- [12] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2905821/>
- [13] <http://webhome.auburn.edu/~tds0009/Articles/Johnson1999.pdf>
- [14] <https://cset.georgetown.edu/wp-content/uploads/AI-and-Compute-How-Much-Longer-Can-Computing-Power-Drive-Artificial-Intelligence-Progress.pdf>
- [15] <https://lacol.net/wp-content/uploads/2018/09/bayesia-no-words-wordpreess-4-750x410.jpg>
- [16] <https://www.rumahcoding.co.id/wp-content/uploads/2013/04/39368795-mandelbrot.jpg>

References 2

- [17] [https://www.goodreads.com/book/show/242472.The Black Swan](https://www.goodreads.com/book/show/242472.The_Black_Swan)
- [18] <https://stats.stackexchange.com/questions/2379/what-are-the-big-problems-in-statistics>
- [19] <https://www.statlearning.com/>
- [20] <https://hastie.su.domains/ElemStatLearn/>
- [21] https://www.stat.berkeley.edu/~aldous/157/Papers/Bayesian_open_problems.pdf
- [22] <https://journals.sagepub.com/doi/full/10.1177/1094428120959827>
- [23] <https://onlinelibrary.wiley.com/doi/epdf/10.1111/ecog.05284>
- [24] <https://rucore.libraries.rutgers.edu/rutgers-lib/55790/PDF/1/play/>
- [25] <https://academic.oup.com/mnras/article/508/1/1446/6368871?login=true>
- [26] <https://jpet.aspetjournals.org/content/372/1/136.abstract>
- [27] <https://journal.r-project.org/archive/2021/RJ-2021-081/RJ-2021-081.pdf>
- [28] <https://journal.r-project.org/archive/2021/RJ-2021-109/RJ-2021-109.pdf>
- [29] <https://www.jstatsoft.org/article/view/v10i06>
- [30] <https://www.jstatsoft.org/article/view/v09s01>
- [31] <https://www.frontiersin.org/articles/10.3389/fams.2022.775068/full>
- [32] <https://www.frontiersin.org/articles/10.3389/fams.2020.553000/full>

Acknowledgements



- ◆ The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- ◆ For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. *"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)"*.

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

