SOCR 2026 MDP Project Summaries

The project profiles below describe the main SOCR MDP R&D Projects for 2026
https://www.socr.umich.edu/html/SOCR_Research.html

GDrive: https://drive.google.com/drive/folders/1w8SIGnL6G44kK3_WA9CUxRHSdiV0-UfC
GSlides: https://docs.google.com/presentation/d/1d8bSbOLn_fg0JBdeHXX4Mwdn9u4QYcHAu6fXseg7wNw/edit?usp=drive_link
SOCR Project Leaders:

- Programming: Simeone Marino, Alex Kalinin, Ivo Dinov
- Methods (CBDA, GrayRain/VH, DataSifter): Simeone Marino
- Analytics: Simeone Marino, Wei-Cheng Chiang, Ivo Dinov
- Spacekime Analytics: Yueyang Shen, Ivo Dinov
- AI/ML: Yueyang Shen, Ryan Kwon, Achu Shankar, Ivo Dinov

SOCR Trainees/Students

Project Summaries

Project Area	Skills	Likely Majors
Programming Subteam: SOCRAT (Charts, Wrangler, Modeler, Analyses, Tools) (2-3 students)	UI/UX design, HTML5, JavaScript, Adobe Illustrator, Canvas	Computer Science (CSE/CS-LSA) School of Information (SI)
TensorFlow.JS UKBB t-SNE, BrainViewer	https://js.tensorflow.org https://js.tensorflow.org/api/latest/ https://codepen.io/pen?&editors=1011	Computer Science (CSE/CS-LSA)
Methods (CBDA & DataSifter) (4 students) DataSifter & CBDA & GrayRain/VH	Technical math background, R-computing	Math, CS, Eng, Physics, Stats, STEM
AI/ML Methods (2-4 students)	Develop Reinforcement Learning, and deep networks methods	Math, CS, Eng, Physics, Stats, STEM
Analytics (4 students) TDA Biomed/Health Applications (see Case-Studies)	R/Python, statistical modeling, high-throughput data analytics, machine learning	Statistics, Biostatistics, Bioinformatics Math Computer Science (CSE/CS-LSA)
Spacekime Analytics (sub-team – working directly with the PI) (4 students) www.spacekime.org	Information measures, entropy KL divergence, PDEs, Dirac's bra-ket, Wheeler-DeWitt Equation, operators. See *The Enigmatic Kime: Time Complexity in Data Science* at the University of Michigan Institute for Data Science (MIDAS) Seminar Series, Slidedeck, YouTube video of this seminar	Physics, math, or engineering background is preferred

SOCR Computing servers:

- ARC-TS: https://arc-ts.umich.edu/open-ondemand/
- SOCR-RShiny: rcompute.nursing.umich.edu
- SOCR-Lighthouse: https://lighthouse.arc-ts.umich.edu (Lighthouse User Guide)

SOCR AI Bot Project

SOCR Project Leaders: Simeone Marino, Ryan Kwon, Achu Shankar, Yueyang Shen, Ivo Dinov
Website: https://rcompute.nursing.umich.edu/SOCR_AI_Bot/
GitHub: https://github.com/SOCR/
GDrive: https://drive.google.com/drive/folders/1pfekobc2oz4v7rO4PD7HasWekC1UGd04
https://socr.umich.edu/GAIM/

Description

Review this new DSPA2 Appendix (**Appendix 9**: OpenAI Synthetic Text, Image and Code Generation), which illustrates the basics of using GPT3 (and Chat-OpenAI-GPT3) for fast, some-what-impressive, and application-ready (pragmatic) demonstrations. Here are the key points:

- We want to quickly develop an RShiny app that allows us to deploy a more interactive app with users able to enter their own text/instructions to drive the synthetic text, image and code simulation on-the-fly. In the DSPA2 Appendix 9, we demonstrate that, but the code/examples are frozen in the HTML page, only the SVG plotly graphics are dynamic. Ideally, we make this all a UX/UI.
- Review this simple syntax for calling the OpenAI API (which can be done in Python or R, albeit for brevity, we are demoing only the R version in Appendix 9). I am hoping you and the team will be able to take this to the next phase – make it more reliable, get better results, generalize to 3D volumes, compare against the GAN/DNN 2D image & 3D volume synth results we already have, and think of other creative ideas!
- We need your review and suggestions on how to expand, improve, augment or better utilize the OpenAI infrastructure in all SOCR projects.
- Mind that each OpenAI user gets only 3-months and $18 credits to run jobs for free (and then billing starts) …. This is unfortunate, but this is their business-model! I'm counting on you to figure out a way for us to run a SOCR-only OpenAI service, where we get the RCompute server to listen to our OpenAI calls (so that we don't need to be charged for each query/request). How can we do that? We need to dive deeper into this and discuss follow-up calls.
- The **Source Rmd file** that generated the DSPA2 Appendix 9 is included as a template. Please don't share this and don't run too many instances with the SOCR OpenAI API account credentials (see instructions on the top of Rmd source, follow these guidelines, establish your account and use your secure-API-keys for testing/validation. For some reason the embedding of the keys in my **.Renviron** file did not initially work (Why?).

Student Skills

- Stats/DS/ML/AI, EECS, Computer Science (CSE/CS-LSA) and School of Information (SI)
- UI/UX design, HTML5, JavaScript

Deliverables

- Expanded functionality, integrate with SOCR GR VH Webapp and SOCR Clustering app
- Improve the synth gen-AI generation (text, images and code).
- Explore alternative free AI API Cloud services to gain access to more powerful AI pretrained models
- Customize these modern, e.g., transfer learning, to medical images (2D) and brain volumes (3D)

Team Activities

- Weekly team Zoom meetings
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

- Video1, Video2
- GSlidedeck
- Review the provided references, discuss with team-members, and unleash your creative side!
- *Tools*: https://monai.io/, https://build.nvidia.com/nvidia/maisi, https://github.com/Project-MONAI/tutorials
- *Datasets*: https://github.com/openmedlab/Awesome-Medical-Dataset?tab=readme-ov-file

SOCR MDP Project: SOCRAT

SOCR Project Leaders: Alex Kalinin / Ivo Dinov
Website: https://socr.umich.edu/HTML5/SOCRAT/
GitHub: https://github.com/SOCR/SOCRAT
Training Modules: https://github.com/SOCR/socr-tutorials
GDrive: https://drive.google.com/drive/folders/1UrNpNDI5sWoXW61YwP02NSv3PBbxfvpC

Description

The Statistics Online Computational Resource Analytics Toolbox (SOCRAT) is a Dynamic Web Toolbox for Interactive Data Processing, Analysis, and Visualization. It's purely built using HTML5 standards and JavaScript (core library) as well as node.js,

Student Skills

- EECS, Computer Science (CSE/CS-LSA) and School of Information (SI)
- UI/UX design, HTML5, JavaScript

Project Goals

- Add most classical parametric and nonparametric statistical tests! To transfer the old SOCR Java Analyses to HTML5/JS.
    - https://socr.umich.edu/html/ana/
    - Java code: https://github.com/SOCR/SOCR-Java/tree/master/src/edu/ucla/stat/SOCR
- Go through the Training Modules, practice HTML/JS/Angular/Node programming
- Get your GitHub domain going and pull current SOCRAT branch
- Choose 1-2 deliverables, go over current design, start expansion, include unit tests, pilot development
- Coordinate with team

Deliverables

- Expanded collection of Charts
- Expanded collection of Data-Modelers
- Expanded collection of (parametric and non-parametric) Statistical Analyses
- Expanded collection of machine learning classification, prediction, clustering and analytics modules.

Team Activities

- Weekly team Zoom meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

- Review the websites
- Alexandr A. Kalinin, Selvam Palanimalai, and Ivo D. Dinov. 2017. SOCRAT Platform Design: A Web Architecture for Interactive Visual Analytics Applications. In Proceedings of HILDA'17, Chicago, IL, USA, May 14, 2017, 6 pages. DOI:10.1145/3077257.3077262

IDE for development (Eclipse, WebStorm, IntelliJ, Netbeans, …, RStuio, Spyder/Py)

SOCR MDP Project: Methods: CBDA

SOCR Project Leaders: Simeone Marino

Website:	http://socr.umich.edu/HTML5/CBDA/
GitHub:	https://github.com/SOCR/CBDA
C-RAN Package:	https://cran.r-project.org/web/packages/CBDA
Training Modules:	http://socr.umich.edu/HTML5/CBDA/
GDrive:	https://drive.google.com/drive/folders/1hjwtgz64A_IUsnRK1gv7mGSJ3HdBHaRW

Description

The SOCR Compressive Big Data Analytics (CBDA) Project conducts research and implements efficient computational algorithms to tackle the Big Data problems of representation and analysis of complex heterogeneous information. Big Data cannot be loaded and processed as a whole. CBDA implements a real-time efficient divide-and-conquer strategy to deconstruct the Big Data into meaningful pieces of information that can be eventually reconstructed for actionable knowledge and predictive analytics.

Student Skills

- Probability, stats, math, numerical methods, optimization
- R programming with RStudio (IDE) experience

Project Goals

- Go through the provided materials and references
- Download the CBDA Package
- Practice with test-cases (https://umich.instructure.com/courses/38100/files/folder/Case_Studies)
- Identify specific R&D direction to go deeper into, and meaningfully contribute to CBDA
- Coordinate with team

Deliverables

- New CBDA methods
- Expanded collection of machine learning forecasting, prediction, classification, clustering methods to expand the available CBDA algorithms
- Release new versions of CBDA R package and publish CBDA #2 manuscript
- Python/Perl scripts to speed up the subsampling strategy with Big Data > 100Gb-1Tb

Team Activities

- Weekly team face-to-face/Zoom meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

- Review the websites
- Marino S, Xu J, Zhao Y, Zhou N, Zhou Y, Dinov, ID. (2018) Controlled feature selection and compressive big data analytics: Applications to biomedical and health studies, PLoS ONE 13(8): e0202674, DOI: 10.1371/journal.pone.0202674.
- Marino, S, Zhao, Y, Zhou, N, Zhou, Y, Toga, AW, Zhao, L, Jian, Y, Yang, Y, Chen, Y, Wu, Q, Wild, J, Cummings, B, Dinov, ID. (2020). Compressive Big Data Analytics: An ensemble meta-algorithm for high-dimensional multisource datasets, PLoS ONE, 15(8):e0228520, DOI: 10.1371/journal.pone.0228520.

SOCR MDP Project: Methods: DataSifter

SOCR Project Leaders: Simeone Marino

Website:	https://DataSifter.org
GitHub:	https://github.com/SOCR/DataSifter
C-RAN Package:	(lite version pending)
Training Modules:	https://DataSifter.org
GDrive:	https://drive.google.com/drive/folders/1jVT5pTa_n8xHjUszn1u5gwTzyvPLtszj

Description

The SOCR DataSifter is a novel method, and an efficient R package, for on-the-fly de-identification of structured Clinical/Epic/PHI data. This approach provides complete administrative control over the risk of data identification when sharing large clinical cohort-based medical data. At the extremes, the data-governor may specify that either null data or completely identifiable data is generated and shared with the data-requester. This decision may be based on data-governor determined criteria about access level, research needs, etc. For instance, to stimulate innovative pilot studies, the data office may dial up the level of protection (which may naturally devalue the information content in the data), whereas for more established and trusted investigators, the data governors may provide a more egalitarian dataset that balances preservation of information content and sensitive-information protection.

Student Skills

- Probability, stats, math, numerical methods, optimization
- R programming with RStudio (IDE) experience

Project Goals

- Go through the provided materials and references
- Download the DataSifter-lite Package
- Practice with test-cases (https://umich.instructure.com/courses/38100/files/folder/Case_Studies)
- Identify specific R&D direction to go deeper into and meaningfully contribute to DataSifter methods, implementation and/or validation
- Coordinate with team
- Examine the open-source Generative Pre-trained Transformer 3 (GPT-3), autoregressive language DL model, which has shown some impressive results in synthetic generation of human-like text/speech.
    - Try to integrate with DataSifterText (perhaps for the next paper)?
    - https://github.com/openai/gpt-3
    - https://en.wikipedia.org/wiki/GPT-3
    - https://arxiv.org/abs/2005.14165

Deliverables

- Test this Python **SDV** package https://ealizadeh.com/blog/sdv-library-for-modeling-datasets https://sdv.dev/SDV/user_guides/evaluation/index.html
- Remember we use the R-package charlatan: https://cran.r-project.org/web/packages/charlatan/index.html
- See: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-00977-1
- Implement these new privacy metrics: https://sdv.dev/SDV/user_guides/evaluation/single_table_metrics.html#privacy-metrics
- New DataSifter methods/algorithms (e.g., addressing text, time-varying, graph data organizations)
- Release new versions of DataSifter R package
- Coordinate/support collaborators

Team Activities

- Weekly team face-to-face/Zoom meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

New DataSifter Paper - *DataSifterLevel: A Non-parametric Decision-making in Statistical Obfuscattion of Sensitive Information*

Goal: Develop a new method for solving a critical outstanding problem in the practical utilization of the DataSifter Algorithm - determination of an appropriate level of obfuscation. Our prior work has demonstrated and validated the power of the DataSifter technique to desensitize sensitive information and facilitate data sharing that balances preservation of data *utility* (data value or energy) and diminishing the *risks* of leakage of sensitive information (data privacy protection). We have extended the core DataSifter method to handle both time-varying (longitudinal) and unstructured (text) data elements. We employ the same strategy used in *parametric* Fisherian statistical inference and *non-parametric* randomization inference for bijective mapping between observed data-driven statistics (*critical values*) and their corresponding likelihoods (*probability values*). Specifically, in this new study, we propose and validate a new algorithm to quantify the relation between (investigator-controlled) *level-of-obfuscation (LOO) parameter* and a likelihood probability measure corresponding to the relative benefits-to-risk (BTR) ratio (utility/risk). The bivariate relation between LOO↔BTR resembles the classical relation between statistics and p-values and the ROC curve relation between true-positive-rate (sensitivity) and the false-positive-rate (1-specificity). We use resampling and randomization to provide data-driven estimation and untangle the LOO↔BTR relationship.

Also productize and report/publish the **DataSifter Longitudinal Obfuscator (DSLO)**

References

- Review the websites
- Marino, S, Zhou, N, Zhao, Yi, Wang, L, Wu Q, and Dinov, ID. (2019) DataSifter: Statistical Obfuscation of Electronic Health Records and Other Sensitive Datasets, Journal of Statistical Computation and Simulation, 89(2): 249–271, DOI: 10.1080/00949655.2018.1545228.
- Zhou, N, Wang, L, Marino, S, Zhao, Y, Dinov, ID. (2022) , DataSifter II: Partially Synthetic Data Sharing of Sensitive Information Containing Time-varying Correlated Observations, Journal of Algorithms & Computational Technology, 15:1–17, DOI: 10.1177/17483026211065379.
- Zhou, N., Wu, Q., Wu, Z., Marino, S., Dinov, ID. (2022) DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes, Journal of Medical Systems, 46(96):1-14, DOI: 10.1007/s10916-022-01880-6.

SOCR MDP Project:

SOCR RShiny Pilot Project Longitudinal Data Modeling & Visualization

SOCR Project Leaders: Simeone Marino, Ivo Dinov

Website:	TBD
GitHub:	TBD
Training Modules:	TBD
GDrive:	TBD

Description

This is a testing/pilot project with specific guidelines, action items, and resources to build a fully functioning RShiny application as a pilot project.

Goal: To develop a PoC (Proof-of-Concept) web app (RShiny) that ingests, visualizes & models longitudinal data.

Resources: We provide a fully functional RShiny template (SOCR_LM_Hist_Summary_RShinyApp_test.r) with 4 simple functionalities (*tabs*): Data, ScatterPlot, Distribution, and Summary. Also attached is a relevant demo dataset as RData (Data_pilot_test_Shushun.RData) for testing the longitudinal data visualization and analysis.

New app requirements (initial draft)

1. The Data tab can be very similar (may consider expanding to allow for importing/loading data from different file types, e.g., json, xml).
2. The Scatterplot, Distribution and Summary tabs are just (working) placeholders.
3. We want the Scatterplot to be called DataVisualization and to be designed mimicking all the functionalities of this existing SOCR app: https://socr.umich.edu/HTML5/MotionChart/, which is build using pure JavaScript/HTML5, where as the new pilot app will be built as an RShiny app using pure R.
4. The (new) DataVisualization tab should have a semi-automated pull-down menu where all the longitudinal variables present in the Data are listed and available for selection. Also there should be a mechanism to select other (time-static/cross-sectional) variables, just like in the MotionChart app. In this tab, there should be a time-dynamic (user-controllable) way to display up to 7D data using X/Y coordinates and glyph/blob appearance (shape/form/color/size/boundary).
5. A new tab should be created and named "Modeling", where we select the longitudinal variables to model (Y) and user-specified longitudinal-model to use (ARMA, ARIMAX, LMM, GEE), OR, as in the ScatterPlot tab of the template RShiny app, automatically apply 2-4 models to the data and display all the predictions for comparison and selection of optimal model. Perhaps report different quantitative metrics of model performance.
6. A quantitative Summary of each model should be displayed in the Summary tab.
7. The app should provide some default dataset(s), as well as user specification/loading of local or URL accessible data (e.g., from SOCR datasets archives on Canvas), and the app be responsive to users updates of the datasets.

Additional Goals

- Add a tab named "Obfuscation" where we use our patented obfuscation algorithm *DataSifter* to de-identify the uploaded data (see this RShiny app as an example).
- Add "reticulate" package to the RShiny app to provide access to Python functionalities.

Pilot Data Specs

Total of 23 subjects data ~ 70,000 records

```
## Short description of data fields
 # "ParticipantResearchID": random de-identified ID
 # "Date_Of_Observation" : date as YYYY|MM|DD
 # "Hour_Of_Observation" : hr range in 24 hrs (so 1= midnight to 1am,
 #                       9 = 8-9AM, 21 = 8-9PM,...)
 # A=Active, R=Resting, EB=Energy Burnt
 # Frequency=# of events , Secs=length in seconds
```

Model to fit

Y=Active Energy Burnt [AEB] over time (autoregressive only for now).
We will eventually provide static auxiliary variables to include in the model for better performance.

SOCR MDP Project:

SOCR TypeScript + React HTML5 Dynamic Webapp Developments

SOCR Project Leaders: Simeone Marino, Ivo Dinov

Website:	https://socr.umich.edu/HTML5/
GitHub:	https://github.com/SOCR (e.g., socr-qi-app, socr-clt-webapp-transition, grayrain-website-transition, grayrain-website-new, socr-clt-webapp, gray-rain-website, etc.)
Training Modules:	https://www.simplilearn.com/tutorials/reactjs-tutorial, https://react.dev/learn/tutorial-tic-tac-toe, https://github.com/benawad/lireddit
GDrive:	TBD

Description

This project will require building strong fullstack programming experience to update some of the early SOCR Webapps that are based on JavaScript/TypeScript and React frameworks. Check the appropriate GitHub repositories, try the live applications, and consider enhancements, functionality expansions, optimizations, and general improvements of the code and user-experiences.

SOCR MDP Project:
R&D: Dynamic Mode Decomposition (DMD) Spacetime Analytics

SOCR Project Leaders: Yueyang Shen, Ivo Dinov

Website: http://dmdbook.com/
GitHub: TBD
Training Modules:
GDrive: https://drive.google.com/drive/folders/1hwEq72iKOHpBzxk0i-3B1g14YGlbbOCg

Description

Dynamic mode decomposition (DMD) is a technique for modeling spatiotemporal dynamical systems. DMD decomposes complex flows into a simple representation based on spatiotemporal coherent structures. DMD is an equation-free data-driven method capable of providing an accurate decomposition of a complex system into spatiotemporal coherent structures that may be used for short-time futurestate prediction and control.

Student Skills

- Math, Physics, computational stats
- Engineering/STEM

Project Goals

- Implement a new DMD R package
- Compare DMD against spacekime analytics (TCIU R package)

References

- See GDrive partition

…

SOCR MDP Project: Webapps & Data Analytics

SOCR Project Leaders: Simeone Marino, Ivo Dinov

Website: <many, e.g., http://socr.umich.edu/HTML5>
GitHub: https://github.com/SOCR <many, e.g., https://github.com/SOCR/ALS_PA>
Training Modules: http://DSPA2.predictive.space
GDrive: https://drive.google.com/drive/folders/1sN1fLYA0oLf1I4e1REJRthaMD0jXBs7w

Description

The SOCR Webapp & Data Analytics projects are focused on interrogating massive amounts of complex biomedical and health data. Each project tackles multiple case-studies using R/RMD/RStudio, RShiny Services, and Python/Jupyter Notebook and the SOCR-Flux Compute Server (https://docs.google.com/document/d/1UmBq_BMiMeUcijvKUCzPeG3tKZaWkinVtKrVWenPK1Y). The webapp development will use R markdown notebook, RShiny web-servers, and Google BigQuery Datasets.

Student Skills

- Biostats, quantitative analytics, probability, stats, math, numerical methods, optimization
- R programming with RStudio (IDE) experience, and/or Python/Jupyter Notebook

Project Goals

- Go through the provided materials and references
- Review the SOCR Data Analytics Publications (http://socr.umich.edu/people/dinov/publications.html)
- Review the SOCR R-environment (https://drive.google.com/file/d/1-u9adsMIYmMkcPD9W_6BbfC1IMETsHF_/)
- Practice with test-cases (https://umich.instructure.com/courses/38100/files/folder/Case_Studies)
- Identify specific case-study and an R&D direction to go deeper into, and meaningfully contribute
- Coordinate with team

Deliverables

- New SOCR end-to-end data analytics protocols
- Analytical results, abstracts, publications, presentations, research findings, etc.
- MIMIC-III analytics
- Baby-growth and mother-obesity relations
- Data Value Metric (DVM)
- European Economics Indicators (longitudinal analytics)
- 2D, 3D, 4D Visualization of complex data
- Coordinate/support collaborators

Team Activities

- Weekly team face-to-face/Zoom meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

- Review the websites and listed resources
- https://rcompute.nursing.umich.edu/
- http://myumi.ch/O49zG
- …

We are looking for 1-2 students to develop a new **SOCR model for an RShiny module architecture** that we can have a very light-weight RShiny apps start, and as the user requests more/higher-level functionality, we dynamically load in additional modules (functions) reflecting user needs, requests, and access type.

In essence, we are looking to mimic how RStudio starts just with R base and then all other functionalities needed for analytics are to be dynamically loaded via package/library load.

Here are some sources that demonstrate modular design of RShiny apps:

- https://www.r-bloggers.com/2019/01/the-shiny-module-design-pattern/
- https://shiny.rstudio.com/articles/modules.html
- https://shiny.rstudio.com/articles/communicate-bet-modules.html
- https://engineering-shiny.org/structuring-project.html

If you are interested in this project, please contact Simeone directly.

SOCR MDP Project: Data Analytics - MIMIC III/IV & UKBB

SOCR Project Leaders: Simeone Marino, Ivo Dinov

Website: TBD
GitHub: https://github.com/SOCR
Training Modules:

- Data Science & Predictive Analytics: http://DSPA.predictive.space
- Previous SOCR Data Analytics Publications: http://socr.umich.edu/people/dinov/publications.html
- Gaining access to the dataset requires an online training module; see onboarding materials below https://drive.google.com/drive/u/1/folders/1Y6Yqq1CuTkHQ5rZg-C9r8_je18nM886l

GDrive: https://drive.google.com/drive/folders/1sN1fLYA0oLf1I4e1REJRthaMD0jXBs7w

Description

This SOCR Data Analytics project is focused on interrogating the MIMIC-III database, a large collection of ~43,000 critical care patients from an ICU in Boston, MA. We will use R/RStudio, Python/Jupyter, and the SOCR-Flux Compute Server[1] to digest the vital signs, laboratory results, free-text data, and waveforms available in this unique dataset and predict clinical outcomes via statistical modeling tools.

[1] SOCR-Flux Compute server:
https://docs.google.com/document/d/1UmBq_BMiMeUcijvKUCzPeG3tKZaWkinVtKrVWenPK1Y

Student Skills

- Biostats, quantitative analytics, probability, stats, math, numerical methods
- Programming experience in R (with RStudio) or Python (with Jupyter Notebook)
- Relational databases & structured query language (SQL)

Project Goals

- Review the provided materials and references (see above)
- Request access to the MIMIC-III dataset (https://mimic.physionet.org/gettingstarted/access/)
    - This involves an online but comprehensive human subjects research ethics course
- Practice with demo dataset (https://physionet.org/works/MIMICIIIClinicalDatabaseDemo/) and the MIMIC Query Builder (https://querybuilder-lcp.mit.edu/dashboard.cgi)
- Identify specific research aims and questions of interest to the team
- Coordinate with team to create a reproducible, accessible answer to these specific aims

Deliverables

- New SOCR end-to-end data analytics protocols
- Data extraction & time-alignment tools for the MIMIC-III dataset
- Build statistical models to predict meaningful clinical outcomes
- Analytical results, abstracts, publications, presentations, research findings, etc.
- Visualization of complex, multidimensional data

Team Activities

- Weekly team face-to-face/Zoom meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)
- Start using the latest MIMIC-IV multimodal data
    - https://physionet.org/content/mimiciv/2.0/
    - https://colab.research.google.com/drive/1REu-ofzNzqsTT1cxLHIegPB0nGmwKaM0

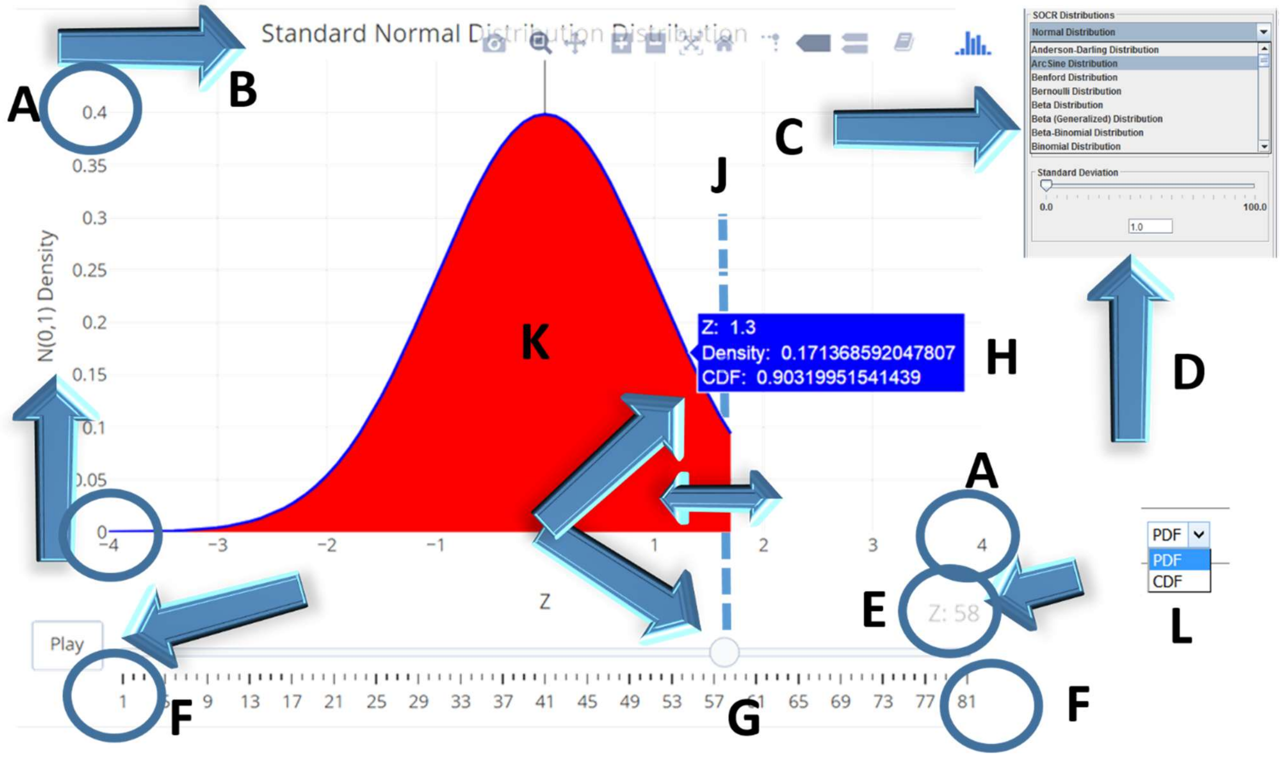
SOCR MDP Project:
Interactive Graphical Probability Distribution Calculator

SOCR Project Leader: Pranjal Srivastava, …, (Mark Bobrovnikov, Jared Chai), Ivo Dinov

Website: http://Distributome.org & https://shiny.med.umich.edu/apps/dinov/SOCR_DistribCalc_RShiny_App/
GitHub: https://github.com/distributome
Training Modules: https://github.com/SOCR/socr-tutorials & http://dspa.predictive.space/
GDrive: https://drive.google.com/drive/folders/184p8VNSOumYEG_SOxlo4MyLVtanq9xLY
Goal: Stand-alone RMD source and a demo HTML (RMD-output) that address the goal of the challenge, provide the desired functionality, and implement it efficiently, without any back-end support (e.g., no shiny server apps).
Deliverables: Stand-alone RMD source and a demo HTML (RMD-output) that address the goal of the challenge, provide the desired functionality, and implement it efficiently.
Background:

- Review DSPA Chapter 2 (for probability distributions) and Chapter 5 (for plot_ly demos)
- See this DSPA Interactive Normal Probability CDF calculation tool.
- Review, experiment and play with the Probability Distributome Calculators. Try at least 2-dozen distributions - what works well and what can be improved there? Mind the selection of parameters and the choice for function to plot (PDF, CDF).

Desired Functionality: The schematic below illustrates the core functionality of the interactive probability distribution calculator. Be creative in your solution.



- Include a drop-down list for the user to select the distribution
- Include an effective strategy to specify the parameters of the selected distribution, mind that the parameter number, interpretation and values will be different for different distributions.
- Make sure you keep the interactive aspect of the interface (*plot_ly* style interactivity)
- Make sure all axes are appropriately scaled, labeled and drawn.

Functionality Annotations

- **A**: x and y axes ranges and labels
- **B**: Appropriate title (placed to avoid overlaps
- **C**: Selection of the specific distribution - should include at least 20 distributions, see class notes the Distributome Calculators.
- **D**: appropriate section of the specific distribution parameters

- **E**: cut off of the critical value (Z)
- **F**: Ranges of the animation slider should match the x-axis range (Z-values range)
- **G**: Play button and the animation point provide the manual user control over the critical value cut off
- **H**: Report the appropriate Z-values, density curve height, and cumulative distribution up to Z (i.e., P(X<Z))
- **J**: There should be a light-colored vertical line at the animation index == Z-value and extending up to the corresponding density height
- **K**: shaded area represents the integral CDF value .....
- **L**: Drop-down selected for plotting PDF, CDF or inverse-CDF (quantile) function to plot.

Starting R Code: The basic skeleton of one solution (using "**plot_ly**") is included below. Many solutions are possible and you can start with anything you like, including this initial script.

```
library(magrittr); library(plotly)
select the right user-specified distribution (drop down list)
# Assuming Std Normal N(0,1) going down
# define the range
z<-seq(-4, 4, 0.1)
# points from -4 to 4 in 0.1 steps
# Define the quantile levels for the inverse-CDF (quantile) function)
q<-seq(0.001, 0.999, 0.01)
# probability quantile values from 0.1% to 99.9% in 0.1% steps
# define a DF containing Z, PDF and CDF
dStandardNormal <- data.frame(Z=z, Density=dnorm(z, mean=0, sd=1),
Distribution=pnorm(z, mean=0, sd=1))
# define an index feature
dStandardNormal$ID <- seq.int(nrow(dStandardNormal))
# Aggregate frames for interactive plot
aggregate_by <- function(dataset, feature) {
feature <- lazyeval::f_eval(feature, dataset)

levels <- plotly:::getLevels(feature)
aggData <- lapply(seq_along(levels), function(x) {
cbind(dataset[feature %in% levels[seq(1, x)], ], frame = levels[[x]])
})
dplyr::bind_rows(aggData)
}

# Apply the aggregate to ID index
dStandardNormal <- dStandardNormal %>% aggregate_by(~ID)

# generate the Plot_ly object
plotMe <- dStandardNormal %>% plot_ly( x = ~Z, y = ~Density, frame = ~frame,
type = 'scatter', mode = 'lines', fill = 'tozeroy', fillcolor="red",
line = list(color = "blue"), text = ~paste("Z: ", Z, "
Density: ", Density, "CDF: ", Distribution), hoverinfo = 'text' ) %>%
layout( title = "Standard Normal Distribution Distribution",
# Specify the right distribution and its parameters!!!
yaxis = list( title = "N(0,1) Density", range = c(0,0.45),
zeroline = F, tickprefix = "" # density value
),
xaxis = list( title = "Z", range = c(-4,4), zeroline = T, showgrid = T ) ) %>%
animation_opts( frame = 100, transition = 1, redraw = FALSE ) %>%
animation_slider( currentvalue = list( prefix = "Z: " ) )
# display interactive plot
plotMe
```

The optimal solution will include RMD (source) and HTML output (webapp).

Distributome RShiny R-package

Team: …; Ivo Dinov

Develop a new SOCR team "*Distributome*" R package, read this free ebook on how to document, build vignettes, construct, package, and validate a new R package that can be submitted to CRAN for peer review and inclusion in the CRAN archive. I hope we can include at least 40+ distributions, and provide as much information for each, at least as much as it's available in the Distributome meta-data XML and as much as is available on Wikipedia.

- See the current release of the **Distributome RShiny App**: https://rcompute.nursing.umich.edu/DistributionCalculator/
- GitHub partition: https://github.com/SOCR/ProbDistCalc_RShiny
- SOCR Distributions Java Code: https://github.com/SOCR/SOCR-Java/tree/master/src/edu/ucla/stat/SOCR/distributions
- Look at "`paramEstimate(double[] distData)`" functions in each SOCR and UAH distribution, e.g., Poisson Distribution param Estimation

Expand SOCRAT Modeler and Probability Distribution Calculator to a new SOCR RShiny Modeler, improving on the SOCR Java Prob Modeler

The current app (RShiny) includes examples of fitting of 2 Prob-Model Distributions to numerical -- *Normal* & *Gamma* distributions – these models show in the 3rd "Distributions" tab of the Newly skeletonized SOCR Modeler RShinyApp, which extends the DistributionCalculators and Histogram apps.

https://drive.google.com/drive/folders/1WzsUe0HD0iTWMiXvpc1I-Q2DwXJi6LYX

This is a great pilot project for any new student to complete the model fitting of all 78 distributions that are listed in the Global.R file.

Tasks:

- Expand the list of distribution model fits from Normal/Gamma to all 78 distributions
- Make the app robust (avoiding runtime errors)
- Handle - categorical, string, incomplete datasets
- Manage tensors (2D, 3D+ arrays, etc. using our SOCR Copula model for multidimensional distributions
  - https://socr.umich.edu/HTML5/BivariateNormal/BVN2/
- Integrate with Clustering app, VH App, and other SOCR RShiny Apps
- Integrate these 2 SOCR RShiny Apps:
  - https://rcompute.nursing.umich.edu/SOCR_Modeler_app/
  - https://rcompute.nursing.umich.edu/SOCR_Clustering/
  - (and may be even this one) https://rcompute.nursing.umich.edu/DistributionCalculator/

Here are the JAVA files that contain for each distribution (file name) a method called "paramEstimate()", which we can quickly export from the Java source into R.

https://github.com/SOCR/SOCR-Java/tree/master/src/edu/ucla/stat/SOCR/distributions

Also the attached R-package source (ExtDist) also has the parameter estimation functions (e.g., eUniform()) for about 20-25 distributions.

Same for the EstimationTools R package …. (their source is online).

Multidimensional distributions:

Bobrovnikov, M, Chai, JT, and **Dinov**, ID. (2022) Interactive Visualization and Computation of 2D and 3D Probability Distributions, SN Computer Science, 3, 327, DOI: 10.1007/s42979-022-01206-w.

- https://socr.umich.edu/HTML5/BivariateNormal/BVN2/

Add Hazard and Survival function plots to PDF and CDF

Having analytical closed forms for the PDF (f) and CDF (F), the *Survival* and *Hazard* functions are defined as:

- The **survival function** is *S(t) = 1 − F(t)*, or the probability that a person (machine, process, or organization) lasts/survives longer than t time units, where F(t) is the usual CDF, i.e., the (tail) probability that the process lasts less than or equal to t time units.
- The **hazard function** is *λ(t) = f(t)/S(t)*, is the probability that a person (machine, process, or organization) dies/fails in the next instant, given that it lasted/survived up to the current time t.

We'll implement these modifications for the Probability Distributome Calculators and the SOCR Distribution Calculator & Modeler.

SOCR MDP Project:
Data Science Fundamentals: Spacekime Analytics

SOCR Project Leader: Yueyang Shen, Milen Velev, Ivo Dinov

Website: https://tciu.predictive.space , www.spacekime.org
GitHub: https://github.com/SOCR/TCIU
Training Modules: ODE/PDE, Kaluza-Klein Theory (https://en.wikipedia.org/wiki/Kaluza–Klein_theory)
GDrive: https://drive.google.com/drive/folders/1PMMBR2bzBPubYMpywLkcTkJPyxOKQ4Aq

Description

The SOCR Data Science Fundamentals project will explore new theoretical representation and analytical strategies to understand large and complex data. It will utilize information measures, entropy KL divergence, PDEs, Dirac's bra-ket operators. This fundamentals of data science research project will explore time-complexity and inferential uncertainty in modeling, analysis and interpretation of large, heterogeneous, multi-source, multi-scale, incomplete, incongruent, and longitudinal data.

See The Enigmatic Kime: Time Complexity in Data Science (https://midas.umich.edu/event/midas-seminar-series-presents-ivo-d-dinov-phd-university-of-michigan/) at the University of Michigan Institute for Data Science (MIDAS) Seminar Series, Slidedeck (http://socr.umich.edu/docs/uploads/2018/Dinov_TCIU_Kime_MIDAS_2018.pdf).

Student Skills
- Physics, math or engineering background
- R programming with RStudio (IDE) experience, and/or Python/Jupyter Notebook, Python

Project Goals
- Develop, validate and share a new TCIU Python Package, also see TCIU GitHub repo
- Go through the provided materials and references
- Review the current platform (will be provided)
- Perform 3D and 4D Plot_Ly visualization of complex manifolds, including 5D space-kime and 2D-curved Kime.
- Identify specific case-study and an R&D direction to go deeper into, and meaningfully contribute
- Coordinate with team

Deliverables
- Visualization protocols
- Math proofs of various physics properties in 5D Minkowski spacekime

Team Activities
- Weekly team face-to-face/Zoom meetings
- Code review Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

Key points
- *What is the problem?* Use complex-time physics to formulate data science theory & practice
- *Why is it important?* There is currently no canonical theory for Big Data discovery science
- *What is the SOCR Solution?* Blend transdisciplinary knowledge to build a new Data Analytic method
- *It's real; here it is (in a pilot form) … demo …* See TCIU Video
- *Why should you consider joining this SOCR-MDP Project?* High-risk/high-potential yield project.

References
- Review the websites and listed resources
- TCIU Website: http://tciu.predictive.space/
- TCIU GitHub: https://github.com/SOCR/TCIU/
- Spacekime: www.SpaceKime.org
- The new Data Science/TCIU/Spacekime book (fully accessible) on the publisher, De Gruyter, website (via UMich IP address) and on UM Library (login required) site. More info is available on the TCIU website.

SOCR MDP Project: AI for Qualitative & Mixed-Type Data

SOCR Project Leaders: Ivo Dinov
Website: …<TBD>...
GitHub: …<TBD>...
Training Modules: <see references below>
GDrive: https://drive.google.com/drive/folders/1SGRi8axPaiWIAzYJXwdmSNJ7wXbwt0ay

Description

Develop new SOCR/DSPA AI/ML tools for representation, modeling, analysis, interrogation, visualization, and interpretation of (1) qualitative data, (2) mixed qualitative-and-quantitative data, and (3) meta-analyses

Student Skills

- EECS, Computer Science (CSE/CS-LSA) and School of Information (SI)
- AI/ML
- UI/UX design, HTML5, JavaScript

Project Goals

- Build and validate AI/ML tools for qualitative data,
- Build and validate AI/ML tools for mixed qualitative-and-quantitative data,
- Build and validate AI/ML tools for meta-analyses

Deliverables

- Stand-alone Rmd notebooks with methods formulation, computational implementation and example utilization using simulated and real (observed) datasets.

Team Activities

- Weekly team meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

SOCR MDP Project: Synthetic 2D Image & 3D Volume Generation

SOCR Project Leaders: Achu Shankar, Ryan Kwon, Kaiming Cheng, Yueyang Shen, Ivo Dinov
GitHub: https://github.com/SOCR/DL_ZNet_3D_BrainSeg
https://github.com/SOCR/DL_ZNet_2D_BrainSeg
https://github.com/SOCR/DL_AFUnet_3D_BrainSeg

Training Modules: see background papers, R-packages, and tutorials listed below
GDrive: https://drive.google.com/drive/folders/1adD9Kp3s_ZyiIFNCPIA6OmFrfuwA-i0H
Pilot RSHiny App: https://rcompute.nursing.umich.edu/SOCR_ImgGenApp/
https://socr.umich.edu/GAIM/,

Description:

This project aims to build a set of ML/AI tools and complete protocols that provide pragmatic mechanisms to generate realistic 2D images (e.g., faces or animals) or 3D volumes (e.g., brain MRI data). A number of different techniques can be used to generate synthetic data (including images/volumes, e.g., VAE, GAN, diffusion methods, Stochastic corruption, generative, and discriminative models, etc. Start by reading several of the papers listed below and playing with the code provided in them. Then, try some of the 2D, 3D and 4D hyper-volumes (data) referenced below and synthetically generate similar simulated cases.

References:

- Papers
  - https://paperswithcode.com/paper/diffusion-models-beat-gans-on-image-synthesis/review/?hl=32098
  - https://paperswithcode.com/paper/score-matching-model-for-unbounded-data-score-1/review/?hl=38037
  - https://paperswithcode.com/paper/score-matching-model-for-unbounded-data-score-1
  - See DSPA Chapter 22 (Deep Learning)
  - Papers with Code: https://paperswithcode.com/task/image-generation

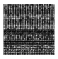
Additional Resources.

- Papers with Code: https://paperswithcode.com/task/image-generation
- Datasets:
  - Brain Images:
    - Neuroimaging of a large group of healthy individuals from the community (138 subjects), as well as samples of individuals diagnosed with schizophrenia (58), bipolar disorder (49), and ADHD (45): https://openneuro.org/datasets/ds000030/versions/1.0.0. It's easy to get an account and download the data (80GB). All data is de-IDed/anonymized.

  - **3D data**: We can use the 3D Brain Tumor Segmentation (BraTS) image dataset for training and testing …. Brain MR dataset contains 257 training images with corresponding labels and the dimensions of these MR images are 240*240 with 155 slices and 4 different imaging modalities including T1 (T1-weighted), T1C (contrast enhanced T1-weighted), T2 (T2-weighted), and FLAIR (Fluid Attenuation Inversion Recovery).

    - See this recent pub: https://www.sciencedirect.com/science/article/pii/S1746809421000550

  - 2D images:
    - CIFAR-10 ; CIFAR-100
    - ImageNet
    - MNIST

        - Cityscapes
        - CelebA
        -  Fashion-MNIST
        - CUB-200-2011
        - STL-10
        - Oxford 102 Flower
        - 100*100 grid/table of 2D PNG brain (MRI) images

- **MONAI - Medical Open Network for Artificial Intelligence** (100K 2D images and 3D volumes for AI/ML)
    - GitHub: https://github.com/Project-MONAI/tutorials
    - Project Site: https://monai.io/

- Review these references
    - See SOCR GitHub:
        - https://github.com/SOCR/DL_AFUnet_3D_BrainSeg
        - https://github.com/SOCR/DL_ZNet_2D_BrainSeg
        - https://github.com/SOCR/DL_ZNet_3D_BrainSeg
    - See SOCR Papers/references
        - 10.3390/bioengineering10050581,
        - 10.1109/JTEHM.2022.3176737,
        - …
    - AI/ML in meta-analyses – please see the attached PDF (AI_MetaAnalyses_ToolsSummary.pdf), which includes a dozen tools and an extensive article on the topic
    - AI/ML for qualitative data –
        - See DSPA Chapter 11 (Association rules) and Chapter 19 (Text Mining/Sentiment analysis)
        - See the examples and tools in this article (AI_QualitativeData_SentimentAnalysis.pdf)
        - See the R package RQDA (https://cran.r-project.org/src/contrib/Archive/RQDA/, http://rqda.r-forge.r-project.org/)
    - AI/ML for mixed methods (qualitative-quantitative) data –
        - https://m-clark.github.io/mixed-models-with-R/random_intercepts.html
        - https://cran.r-project.org/web/packages/boral/index.html
        - http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html
    - Go over DSPA Chapter 14 (Deep Learning): https://socr.umich.edu/DSPA2/DSPA2_notes/14_DeepLearning.html

IDE for development (Eclipse, WebStorm, IntelliJ, Netbeans, …, **RStuio/RMarkdown**, **Spyder/Py**)

SOCR MDP Project: VirtualHospital/Synthetic Data

SOCR Project Leaders: Simeone Marino, Johnny Liu, Ronak Shetty, Ivo Dinov

Website:	https://gray-rain.com
GitHub:	...
Training Modules:	https://www.edx.org/course/health-informatics-data-and-interoperability-stand
GDrive:	https://drive.google.com/drive/folders/1hvzoihZMO-6uJIPbObI4rVc_QDp9i5tD

Description

This project involves developing a Virtual Hospital (VH) and Synthetic Patient (SP) capability to simulate realistic electronic health records including categorical, discrete, continuous, imaging, text and biomedical specimen data. This relates to the unstructured-DataSifter, synthetic text generation, text-obfuscation, and text-mining/inference.

Student Skills

- EECS, Stats/Biostats/Math, Computer Science (CSE/CS-LSA) and School of Information (SI)
- UI/UX design, HTML5, JavaScript

Project Goals

- Go through the Training Modules, practice HTML/JS/Angular/Node programming
- Get your GitHub domain going
- Choose 1-2 deliverables, go over current design, start expansion, include unit tests, pilot development
- Coordinate with team

Deliverables

- HL7/FHIR Interface (see below)
- Learn the current state of the project (Simeone)
- Examine the prior datasets (NHANES, MIMIC, etc.)
- Choose categorical, discrete, continuous, imaging, text and biomedical specimen data and illustrate examples
- Experiment and demo 1,000 VH/SP cases
- Validate VH/SP using machine learning classification, prediction, clustering and analytics modules.

Team Activities

- Weekly team meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR Zoom channel)

References

- Adopt the HL7 XML Format, examine the "*Persona*" virtual clinical casebooks, and ensure 2 easy pathways (data portals):
- From-Anywhere Into-VirtualHospital, and
- From VirtualHospital Out-to-Anywhere.
- Data Standards and Formats for Information exchange - HL7 Format - *Fast Healthcare Interoperability Resources* (**FHIR**)
- Instrucitons https://fire.ly/2017/10/31/make-your-first-fhir-client-in-r-within-one-hour/
    - R Packages:
        - RonFHIR
        - FHIR/Github, install.packages("remotes"); remotes::install_github("TPeschel/fhiR")
        - Fhircrackr see the fhircrackr vignettes for many great examples, may need to install the more advanced dev-version: devtools::install_github("POLAR-fhiR/fhircrackr").

  - FHIR is a very useful standard to describe and exchange medical data in an interoperable way. FHIR is not useful for statistical analyses of data, since FHIR data is stored in many nested and interlinked resources instead of matrix-like DF structures.
  - Use the available public servers, https://hapi.fhir.org/baseR4 or http://fhir.hl7.de:8080/baseDstu3 as FHIR server endpoint to connect VH clients to.
  - Example of executing a **FHIR search** of the form [base]/[type]?parameter(s), where [type] refers to the type of resource you are looking for, and [parameter(s)] characterize specific data-search properties:
    - https://hapi.fhir.org/baseR4/Patient?gender=female
- **Documentation/Training**
  - Read this simple overview or R-based LH7 data XML representation
  - Read this PDF (RonFHIR-Overview-2018-11-15.pdf) …. A must read!
  - Paper (R/**RShiny**/FHIR): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5939961/

- Synthetic data generation using Variational Auto-encoders/auto-decoders (AE-AD).

- We can explore data augmentation and synthetic generation variational AE-AD neural networks to generate synthetic cases (using the decoder process, i.e., predict function of the NN model) and explore the similarities between the joint distributions of the original data and the synth-data.

- https://towardsdatascience.com/how-to-generate-new-data-in-machine-learning-with-vae-variational-autoencoder-applied-to-mnist-ca68591acdcf

- https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726

- https://arxiv.org/ftp/arxiv/papers/1808/1808.06444.pdf

- https://towardsdatascience.com/autoencoders-for-the-compression-of-stock-market-data-28e8c1a2da3e

- SOCR example https://socr.umich.edu/HTML5/ABIDE_Autoencoder/

- **Charlatan R package:** https://cran.r-project.org/web/packages/charlatan/index.html
- **synthpop** package provides synthetic –data generation: https://www.r-bloggers.com/generating-synthetic-data-sets-with-synthpop-in-r/ and saves the synthText in diff formats. See this paper.
- **stringdist** package allows us to compare strings/text: https://cran.r-project.org/web/packages/stringdist
- GAN (generative Adversarial Network) models for synthetic image generation: https://www.r-bloggers.com/conditional-generative-adversarial-network-with-mxnet-r-package/
  - Also see: https://becominghuman.ai/generative-adversarial-networks-for-text-generation-part-1-2b886c8cab10
- See this medical text-GAN mtGAN (Python) paper: https://arxiv.org/pdf/1812.02793.pdf
- Autoencoder approach: https://github.com/stas-semeniuta/textvae and https://www.aclweb.org/anthology/D17-1066.pdf
- Update the **DataSifterText synthesizer** to allow obfuscation and de-novo synth medical notes generation using this newly released Open Asclepius-Synthetic-Clinical (HuggingFace) LLM.
  - https://huggingface.co/datasets/starmpcc/Asclepius-Synthetic-Clinical-Notes
  - **Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes**
  - Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, Edward Choi
  - Note that there is lots of real clinical/medical text data available here: https://mtsamples.com, but these appear to be only available one at a time (no bulk download unfortunately), which we can use for various SOCR unstructured-text processing/analytics ….

NEW VH feature – **VirtualHospital – Data Marketplace Exchange (DMX) Blockchain Contracts**

- Use the Ether R-package: https://cran.r-project.org/web/packages/ether/index.html

- Goal: VH becomes the mediator between data-governors (owning data) and domain-scientists/data-analysis (experts). Neither of these 2 stakeholders trust each other (or GR/VH for that matter). However, using blockchain technology, GR/VH allows the 2 parties to sign contracts using GR VH/DS technology for data-sharing via statistical obfuscation.

- Once a contract is in place and added to the Ether-contract block-chain, the data-governor can use the VH to obfuscate the data (according to the contract agreement terms, initially likely $\eta = 1$, for sharing synth data) and share the de-sensitized data with the partner (domain expert).

- Next, the Data-Analyst (expert) can see the data in their GR/VH profile/storage and begin the analytics.

- Finally, the analysts send back reports of their findings to the corresponding data-governors for review

- Based on the results, re-negotiation of contract terms may take place, potentially lowering the obfuscation level ($0 \leq \eta \leq 1$). This can strengthen partnerships between the stakeholders.

See **these blockchain materials**:

- https://drive.google.com/drive/folders/1dtQcVKVkJFSKQVPdAboSesQOy2gLBsTg
- PharmaLedger Blockchain Platform
- …

SOCR MDP Project: HTML5/JavaScript

SOCR Project Leaders: Ivo Dinov
Website: various
GitHub: https://github.com/SOCR
Training Modules: https://github.com/SOCR/socr-tutorials
GDrive: https://drive.google.com/drive/folders/1uZaLGej8NICGfL9NgiFURPvkNIJshQ5F

ReadMe File: https://docs.google.com/document/d/1nyJiJqrDq8wRjEjJumScp50g6ns1SL2CoH8p0fAhsFM/

I. Kime_CircleCloud_WebApp_index.html

This is a pure JavaScript/HTML5 app that needs some bug-fixing, improvements, and feature enhancements to show the dynamics of the natural attraction-repelling forces.

II. SOCR_Vase_TCIU_Model.html

This is an R-source (Rmd) that is knitted/exported as pure HTML5/JS.

III. SOCR_BivariateNormalDistribution Webapp

How to expand the Bivariate-Normal to any Bivariate Distribution? Define the 2 marginal distributions and use this protocol to specify their association (in terms of the correlation) to derive the joint Bivariate distribution PDF:
https://academic.oup.com/mnras/article/406/3/1830/977873

This is an R-source (Rmd) that is knitted/exported as pure HTML5/JS. Below is a starting JavaScript implementing the basic Bivariate Normal Distribution functions.

We want to expand the current app (see it live here and the code (BVN.zip) is here), to include in addition to the MCMC simulation-based (approximate) calculations, which is already implemented, ADD an exact PDF/CDF based calculations. Include a new check-box to allow the user to specify exact vs. approximate calculation (just like we have for with/without using WebGL for the visualization.

```
function normalCDF(X){
      // using Hastings algorithm with maximal error=10^{-6}
      var T=1/(1+.2316419*Math.abs(X));
      var D=0.3989423*Math.exp(-X*X/2);
      var Prob=D*T*(.3193815+T*(-0.3565638+T*(1.781478+
                   T*(-1.821256+T*1.330274))));
      if (X>0) {Prob=1-Prob}
      return Prob
}

function binormalCDF(x,y,R){     // P(X>x,Y>y;R)
    with (Math){
           var s=(1-normalCDF(x))*(1-normalCDF(y));
           var sqr2pi=sqrt(2*PI);
           var h0=exp(-x*x/2)/sqr2pi;
           var k0=exp(-y*y/2)/sqr2pi;
           var h1=-x*h0;
           var k1=-y*k0;
           var factor=R*R/2;
           s=s+R*h0*k0+factor*h1*k1;
           var n=2;
           while ((n*(1-abs(R))<5)&&(n<101)) {
                factor=factor*R/(n+1);
                h2=-x*h1-(n-1)*h0;
                k2=-y*k1-(n-1)*k0;
                s=s+factor*h2*k2;
                h0=h1; k0=k1; h1=h2;
                k1=k2; n=n+1;
           }
           var v=0;
           if (R>.95) {
                v=1-normalCDF(max(h,k))
                s=v+20*(s-v)*(1-R);
           } else if ((R<-.95)&&(h+k<0)) {
                v=abs(normalCDF(h)-normalCDF(k))
                s=v+20*(s-v)*(1+R);
           }
    }
    return s;
}

function BVN( ) {
// The following user inputs are necessary
    X,Y, M1, M2 (means), and S1, S2 (sigma1 and sigma 2),
     and Rho (correlation)
    Prob="NaN";
    if ((S1<=0)||(S2<=0)){
      alert("The standard deviations must be positive.")
    } else if ((R<-1)||(R>1)){
      alert("The correlation coefficient must be between -1 and +1.");
    } else {
          h=-(X-M1)/S1;
          k=-(Y-M2)/S2;
          Prob=binormalCDF(h,k,R);
          Prob=Math.round(100000*Prob)/100000;
     }
    return(Prob);
}
```

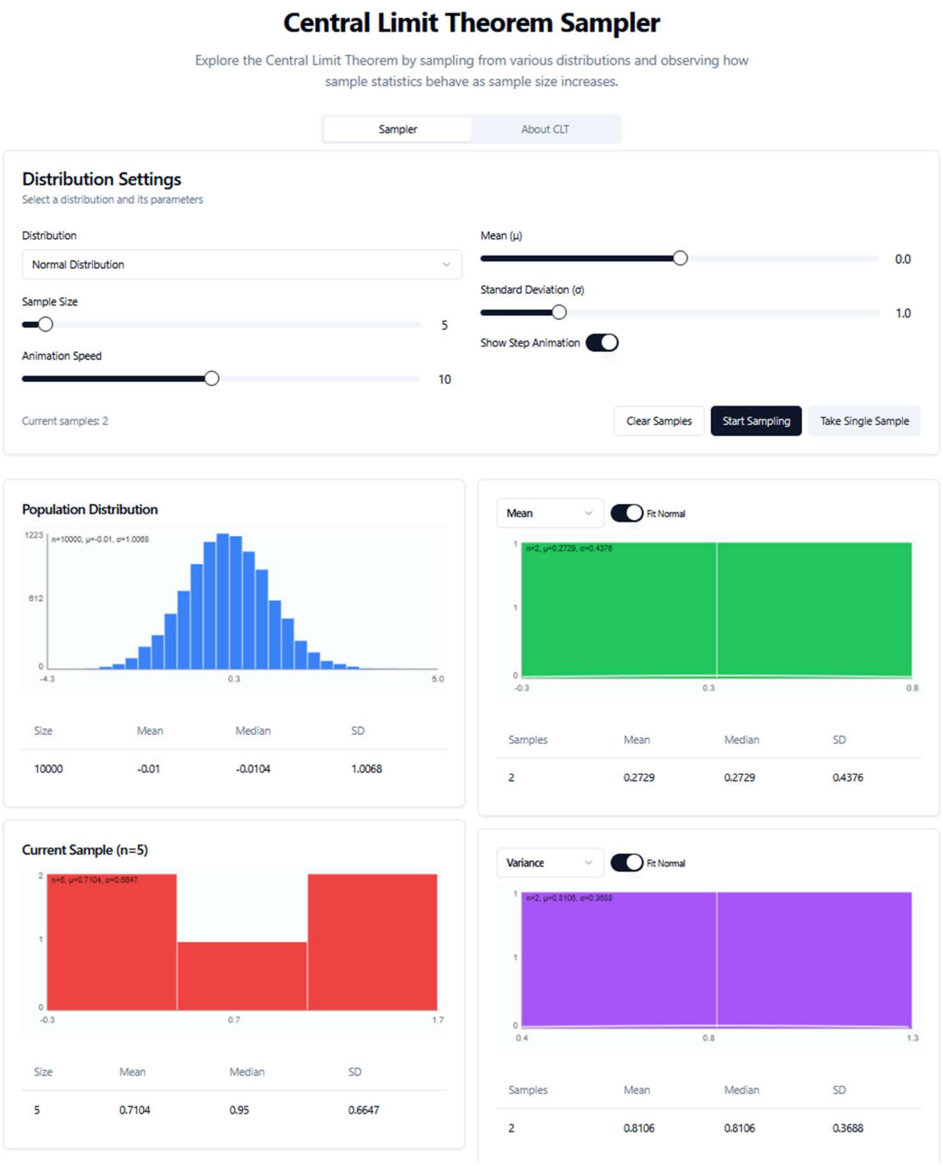
SOCR MDP Project: Java Applets Code → HTML/JS Apps

SOCR Project Leaders: Ivo Dinov
Website: various
GitHub: https://github.com/SOCR & https://github.com/SOCR/SOCR-Java

Convert and modernize some of the SOCR Java Applets to modern HTML5/JavaScript apps

Pilot Project: Start with the **SOCR Sampling Distribution (CLT) Java Applet**
Only viewable in MS IE (not edge!): (http://socr.ucla.edu/htmls/SOCR_Experiments.html)

- https://jse.amstat.org/v16n2/dinov.html
- http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem
- **HTML5/JavaScript Code**: https://github.com/SOCR/socr-clt-webapp
- **Deployed** at: https://socr-clt-webapp.lovable.app/



Central Limit Theorem Sampler

Explore the Central Limit Theorem by sampling from various distributions and observing how sample statistics behave as sample size increases.

Distribution Settings

Select a distribution and its parameters

Distribution: Normal Distribution

Sample Size: 5

Animation Speed: 10

Mean (μ): 0.0

Standard Deviation (σ): 1.0

Show Step Animation

Current samples: 2

Population Distribution

Size	Mean	Median	SD
10000	-0.01	-0.0104	1.0068

Mean

Samples	Mean	Median	SD
2	0.2729	0.2729	0.4376

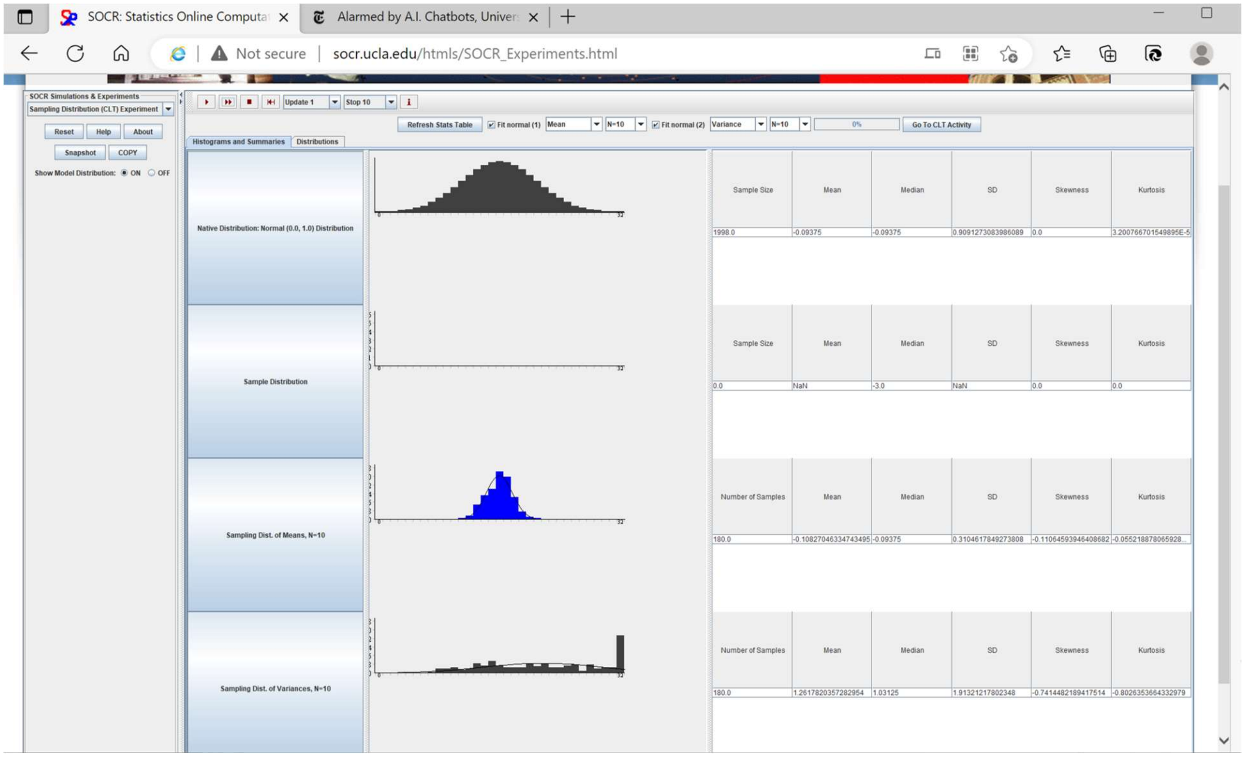
Current Sample (n=5)

Size	Mean	Median	SD
5	0.7104	0.95	0.6647

Variance

Samples	Mean	Median	SD
2	0.8106	0.8106	0.3688

SOCR Sampling Distribution (CLT) Java Applet


SOCR: Statistics Online Computa
Alarmed by A.I. Chatbots, Univer
Not secure
socr.ucla.edu/htmls/SOCR_Experiments.html
SOCR Simulations & Experiments
Reset
Help
About
Snapshot
COPY
Refresh Stats Table
Go To CLT Activity
0%
Histograms and Summaries
Distributions
Native Distribution: Normal (0.0, 1.0) Distribution
Sample Distribution
Sampling Dist. of Means, N=10
Sampling Dist. of Variances, N=10
0
32

Sampling Distribution (CLT) Experiment

Show Model Distribution: ● ON ○ OFF

Update 1 · Stop 10

☑ Fit normal (1) Mean N=10 · ☑ Fit normal (2) Variance N=10

Sample Size	Mean	Median	SD	Skewness	Kurtosis
1998.0	-0.09375	-0.09375	0.9091273083986089	0.0	3.200766701549895E-5

Sample Size	Mean	Median	SD	Skewness	Kurtosis
0.0	NaN	-3.0	NaN	0.0	0.0

Number of Samples	Mean	Median	SD	Skewness	Kurtosis
180.0	-0.10827046334743495	-0.09375	0.3104617849273808	-0.11064593946408682	-0.055218878065928...

Number of Samples	Mean	Median	SD	Skewness	Kurtosis
180.0	1.2617820357282954	1.03125	1.91321217802348	-0.7414482189417514	-0.8026353664332979

Java app source-code:

- https://github.com/SOCR/SOCR-Java/tree/master/src/edu/ucla/stat/SOCR/experiments

Additional critical Java Apps to convert to HTML5/JS:

- LLN:
  - http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ExperimentsActivities
  - http://socr.ucla.edu/htmls/exp/LLN_Simple_Experiment.html
- CLT
  - http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ExperimentsActivities
  - http://socr.ucla.edu/htmls/exp/Sampling_Distribution_CLT_Experiment.html
- π and e stochastic estimation experiments:
  - http://socr.ucla.edu/htmls/exp/Uniform_E-Estimate_Experiment.html
- Polynomial model fitting: http://socr.ucla.edu/Applets.dir/SOCRCurveFitter.html

SOCR MDP Project: Convolutional Graph Neural Networks

SOCR Project Leaders: TBD, Ivo Dinov
Website: TBD
GitHub: TBD

Description:

This research project aims to explore graph-based neural networks (e.g., graph convolutional networks, GCN) to analyze biomedical data.

References:

- Papers
  - https://arxiv.org/abs/1901.00596
  - https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-019-0069-y
  - https://github.com/thunlp/GNNPapers
- Tutorials
  - https://tkipf.github.io/graph-convolutional-networks/
  - https://petar-v.com/talks/GNN-Wednesday.pdf
  - https://neptune.ai/blog/graph-neural-network-and-some-of-gnn-applications
- Code
  - https://paperswithcode.com/method/gcn
  - https://paperswithcode.com/search?q_meta=&q=graph+neural+networks
- Test data: https://umich.instructure.com/courses/38100/files/folder/Case_Studies

SOCR MDP Project: Topological Data Analysis (TDA), Persistent Homology, and Betti Numbers for Point-cloud Data

SOCR Project Leaders: Yueyang Shen, Ivo Dinov
Website: TBD
GitHub: TBD

Training Modules: see background papers, R-packages, and tutorials listed below
GDrive: https://drive.google.com/drive/folders/1QZLjH_xN_SSsrjF2jsc9mDQBUCdDKPlY

Description:

This project aims to expand the SOCR lab data analytical capabilities using advanced topological representations. Data are viewed as high-dimensional point clouds. These are interpreted as samples through a high-dimensional manifold which will be modeled using simplicial complexes. Using the simplicial decomposition, we can compute persistent homology, Betti numbers, and derive other topological metrics that can be used for classification, regression and clustering.

References:

- Papers
  - https://www.ams.org/journals/notices/201905/rnoti-p686.pdf
  - https://arxiv.org/pdf/1705.02037.pdf
  - https://arxiv.org/pdf/1812.02987.pdf
- Tutorials
  - See DSPA Chapter 5: http://www.socr.umich.edu/people/dinov/courses/DSPA_notes/05_DimensionalityReduction.html
  - http://www.stat.cmu.edu/~jisuk/files/20180613_SoCG_Jisu_KIM_TDA_slide.pdf
- R-Code
  - https://cran.r-project.org/web/packages/TDA/TDA.pdf
  - https://cran.r-project.org/web/packages/TDAstats/TDAstats.pdf
- Test data: https://umich.instructure.com/courses/38100/files/folder/Case_Studies

Topological, Fiber-Bundles, and Differential-Geometric approaches to Data Science

- *Flag Manifolds*, see nested Flag vector spaces and Canonical Correlation Analysis (CCA) (DS applications)
- *Grassmann Manifolds*, see; Foundations of Grassmann manifold
- Comparing datasets using flag-manifolds.

Algebraic Data Analysis (ADA)

- Similarly to TDA, try to design a new Dataset → Algebraic Group, Ring or Field mapping that transforms each dataset into a mathematical object (e.g., a Lie group like SO(n,F) with a corresponding Lie algebra so(n,F) …
- See:
  - https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9259191
  - http://www.cs.technion.ac.il/~ron/PAPERS/Conference/RosBroBroKim3DOR2012.pdf
  - https://arxiv.org/pdf/1912.00396.pdf

SOCR MDP Project: Novel AI/ML Techniques

SOCR Project Leaders: Yueyang Shen, Simeone Marino, Ivo Dinov
Website: TBD
GitHub: TBD

Training Modules: see background papers, R-packages, and tutorials listed below
GDrive: https://drive.google.com/drive/folders/1CSg6xxcwYsjRSV1AjgFNjm1dfTQ5jxeS

Description:

This project aims to build new reinforcement-learning-based and deep neural network-based techniques for representation of complex high-dimensional data and for clustering, classification, interpolation, extrapolation, forecasting, and prediction.

References:

- Papers
  - See Project partition: https://drive.google.com/drive/folders/1CSg6xxcwYsjRSV1AjgFNjm1dfTQ5jxeS
  - See this talk: https://www.youtube.com/watch?v=8QLDXlTiRlI
  - https://h2o.gitbooks.io/h2o-tutorials/content/tutorials/ensembles-stacking/ (DNN + SuperLearner)
- Tutorials
  - See DSPA Appendix: https://dspa.predictive.space
- R-Code
  - ...
- **Test data**:
  - https://HealthGym.ai/datasets/
  - https://umich.instructure.com/courses/38100/files/folder/Case_Studies
  - https://github.com/openmedlab/Awesome-Medical-Dataset?tab=readme-ov-file
- Review these concepts:
  - https://en.wikipedia.org/wiki/Backpropagation
  - https://en.wikipedia.org/wiki/Invariant_subspace
  - Learn to compress and compress to learn
  - DSPA: dspa2.predictive.space

Demos: Here are some specific examples we can consider implementing using QL in the Appendix:

- Puzzle: https://en.wikipedia.org/wiki/15_puzzle
- Nonlinear function optimization like this DSPA example: https://socr.umich.edu/people/dinov/courses/DSPA_notes/21_FunctionOptimization.html#93_Convexity
- Games:
  - *2048*: https://towardsdatascience.com/a-puzzle-for-ai-eb7a3cb8e599 and https://github.com/voice32/2048_RL
  - *Snake*: https://towardsdatascience.com/how-to-teach-an-ai-to-play-games-deep-reinforcement-learning-28f9b920440a
  - *Atari*: https://becominghuman.ai/lets-build-an-atari-ai-part-1-dqn-df57e8ff3b26 and datasets (https://paperswithcode.com/task/atari-games)
- Some harder problems are included here: https://analyticsindiamag.com/reinforcement-learning-top-state-of-the-art-games-alphago/