

# SOCR 2024 MDP Project Summaries

The one-page summaries below describe the main SOCR MDP R&D Projects for 2024  
[https://www.socr.umich.edu/html/SOCR\\_Research.html](https://www.socr.umich.edu/html/SOCR_Research.html)

**GDrive:** [https://drive.google.com/drive/folders/1k7PtsvjJwUO\\_1lgIG7MX2lpBUV9CYHQQ](https://drive.google.com/drive/folders/1k7PtsvjJwUO_1lgIG7MX2lpBUV9CYHQQ)

**GSlides:** <https://docs.google.com/presentation/d/14Ex5-xLx88lzS-5gjz5mGCuiHs0LR91jsP8D0J4ZsQM/>

## SOCR Project Leaders:

- Programming: Simeone Marino, Yongxiang Zhao, Alex Kalinin, Ivo Dinov
- Methods (CBDA, GrayRain/VH, DataSifter): Simeone Marino
- Analytics: Simeone Marino, Mingyi Tang, Yang Chen, Ivo Dinov
- Spacekime Analytics: Yueyang Shen, TBD, Ivo Dinov
- AI/ML: Yueyang Shen, TBD, Ivo Dinov

## SOCR Trainees/Students

### Project Summaries

Project Area	Skills	Likely Majors
<b>Programming Subteam: <a href="#">SOCRAT</a></b> (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)
<b>TensorFlow.JS</b>  <a href="#">UKBB t-SNE</a> , <a href="#">BrainViewer</a>	<a href="https://js.tensorflow.org">https://js.tensorflow.org</a> <a href="https://js.tensorflow.org/api/latest/">https://js.tensorflow.org/api/latest/</a> <a href="https://codepen.io/pen?&amp;editors=1011">https://codepen.io/pen?&amp;editors=1011</a>	Computer Science (CSE/CS-LSA)
<b>Methods (CBDA &amp; DataSifter)</b> (4 students) <a href="#">DataSifter</a> & <a href="#">CBDA</a> & <a href="#">GrayRain/VH</a>	Technical math background, R-computing	Math, CS, Eng, Physics, Stats, STEM
<b>AI/ML Methods</b> (2-4 students)	Develop Reinforcement Learning, and deep networks methods	Math, CS, Eng, Physics, Stats, STEM
<b>Analytics</b> (4 students) <a href="#">TDA</a> <a href="#">Biomed/Health Applications</a> (see <a href="#">Case-Studies</a> )	R/Python, statistical modeling, high-throughput data analytics, machine learning	Statistics, Biostatistics, Bioinformatics Math Computer Science (CSE/CS-LSA)
<b>Spacekime Analytics</b> (sub-team – working directly with the PI) (4 students) <a href="http://www.spacekime.org">www.spacekime.org</a>	Information measures, entropy KL divergence, PDEs, Dirac’s bra-ket, Wheeler-DeWitt Equation, operators. See <a href="#">The Enigmatic Kime: Time Complexity in Data Science</a> at the <a href="#">University of Michigan Institute for Data Science (MIDAS) Seminar Series</a> , <a href="#">Slidedeck</a> , <a href="#">YouTube video of this seminar</a>	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](http://rcompute.nursing.umich.edu)
- SOCR-Lighthouse: <https://lighthouse.arc-ts.umich.edu> ([Lighthouse User Guide](#))

# SOCR AI Bot Project

**SOCR Project Leaders:** Simeone Marino, Shihang Li, Yongxiang Zhao, Yueyang Shen, Ivo Dinov

**Website:** [https://rcompute.nursing.umich.edu/SOCR\\_AI\\_Bot/](https://rcompute.nursing.umich.edu/SOCR_AI_Bot/)

**GitHub:** <https://github.com/SOCR/>

**GDrive:** <https://drive.google.com/drive/folders/1pfekobc2oz4v7rO4PD7HasWekC1UGd04>

## 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

- Review the provided references, discuss with team-members, and unleash your creative side!

# SOCR Multi-Table Join Harmonization, Aggregation & Computing Project

**SOCR Project Leaders:** Simeone Marino, Ivo Dinov

**Website:** ...

**GitHub:** ...

**GDrive:** [https://drive.google.com/drive/folders/1bF6\\_repW9HIIAJWpIGLfiWedyC6wNkCA](https://drive.google.com/drive/folders/1bF6_repW9HIIAJWpIGLfiWedyC6wNkCA)

## Description

Using the [Barcelona case-study](#), we want to develop a generic Rmd electronic-notebook illustrating end-to-end the process of generating a single computable data object using a large number of flat-files, tables, SQL query dumps, and other spreadsheets including different types of data elements describing one specific process, phenomenon, or a study. For instance, consider an EHR system that may have upwards of 10K tables. Upon running an SQL query, we get a set of dozens or hundreds of tables describing the clinical states of a number of study participants. The aim is to use the Barcelona case-study, which does not have any individual or sensitive information as a template to solve the more general problem of deriving computable data objects out of a large number of separate spreadsheets/tables.

## Student Skills

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

## References

- Review the provided references, discuss with team-members, and unleash your creative side!

# SOCR 2024 MDP Project: SOCRAT

**SOCR Project Leaders:** Shihang Li, Yongxiang, Zhao, ..., 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

- 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](https://doi.org/10.1145/3077257.3077262)

**IDE** for development (Eclipse, WebStorm, IntelliJ, Netbeans, ..., RStudio, Spyder/Py)

# SOCR 2024 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/1hjwqz64A\\_IUsnRK1qv7mGSJ3HdBHaRW](https://drive.google.com/drive/folders/1hjwqz64A_IUsnRK1qv7mGSJ3HdBHaRW)

## 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](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 2024 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](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](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](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://bmcmmedresmethodol.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](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)

## 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 2024 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-3B1g14YGIbbOCg>

### **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 spacetime analytics (TCIU R package)

### **References**

- See [GDrive partition](#)

...



# SOCR 2024 MDP Project: Webapps & Data Analytics

**SOCR Project Leaders:** Simeone Marino, Shihang Li, 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](https://github.com/SOCR/ALS_PA)>  
**Training Modules:** <http://DSPA.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](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\\_/](https://drive.google.com/file/d/1-u9adsMIYmMkcPD9W_6BbfC1IMETsHF_/))
- Practice with test-cases ([https://umich.instructure.com/courses/38100/files/folder/Case\\_Studies](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-between-modules.html>
- <https://engineering-shiny.org/structuring-project.html>

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

# SOCR 2024 MDP Project: Data Analytics - MIMIC-III/IV

**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](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<sup>1</sup> 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.

<sup>1</sup>SOCR-Flux Compute server:

[https://docs.google.com/document/d/1UmBq\\_BMiMeUcijvKUCzPeG3tKZaWkinVtKrVWenPK1Y](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-ofzNzqsTT1cxLHlegPB0nGmwKaM0>

# SOCR 2024 MDP Project: SOCR TensorFlow/TensorBoard Apps

**SOCR Project Leader:** Peijing Li, Haodong Luo, Alex Kalinin, Ivo Dinov

**Website:** [https://socr.umich.edu/HTML5/SOCR\\_TensorBoard\\_UKBB](https://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB)

**GitHub:** <https://github.com/SOCR/97-tensorflowjs-quick-start>

**Training Modules:** <https://js.tensorflow.org/tutorials/>

**GDrive:** [https://drive.google.com/drive/folders/1wJY8539tpLmYiJc\\_vKZvl6oDVDAHTQu9](https://drive.google.com/drive/folders/1wJY8539tpLmYiJc_vKZvl6oDVDAHTQu9)

## Description

The SOCR TensorFlowJS/TensorBoardJS project aims to design, build, validate and release new webapps based on the ML TensorFlow framework. For example, students will dive deep into TensorFlow.JS (<https://js.tensorflow.org>, <https://js.tensorflow.org/api/latest/>, <https://codepen.io/pen?&editors=1011>) and TensorBoard.JS (<https://github.com/tensorflow/tensorboard>, [https://www.tensorflow.org/guide/summaries\\_and\\_tensorboard](https://www.tensorflow.org/guide/summaries_and_tensorboard)).

## Student Skills

- EECS, Computer Science (CSE/CS-LSA) and School of Information (SI)
- AngularJS, TensorFlowJS, TensorBoard, JavaScript, HTML5

## Project Goals

- Go through the Training Modules, practice HTML/JS/Angular/Node programming
- Get your GitHub domain going and start pilot testing various applications
- Use SOCR Data to experiment
- Review Vegi's SOCR t-SNE TensorFlow Webapp ([http://socr.umich.edu/HTML5/SOCR\\_TensorBoard\\_UKBB](http://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB))
- Coordinate with team
- Rapid RDD (research, development and deployment) is needed in this project

## Deliverables

- 2-5 new SOCR TF/TB Apps
- ...

## 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

# SOCR 2024 MDP Project: Interactive Graphical Probability Distribution Calculator

**SOCR Project Leader:** Shihang Li (Mark Bobrovnikov, Jared Chai), Ivo Dinov

**Website:** <http://Distributome.org> & [https://shiny.med.umich.edu/apps/dinov/SOCR\\_DistribCalc\\_RShiny\\_App/](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\\_SOxlo4MyLVtang9xLY](https://drive.google.com/drive/folders/184p8VNSOumYEG_SOxlo4MyLVtang9xLY)

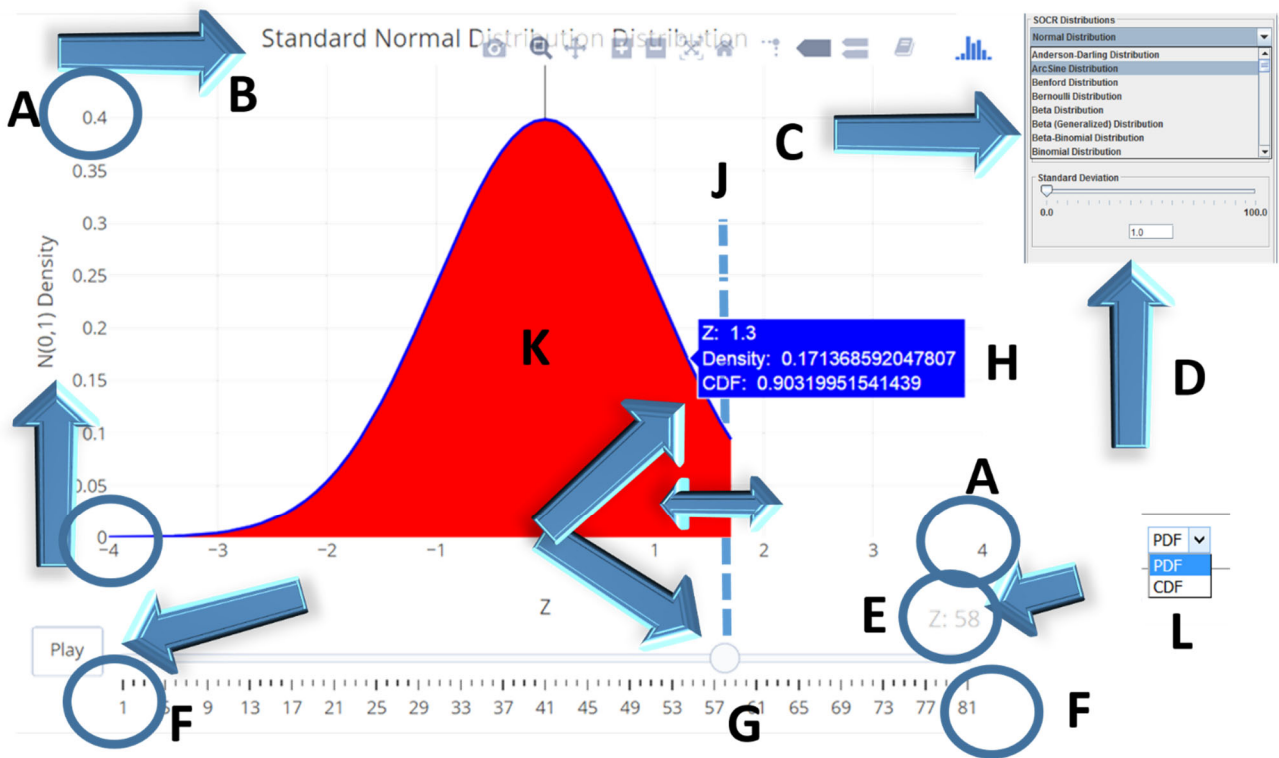
**Goal:** Stand-alone RMD source and a demo HTML (RMD-output) that address the goal of the challenge, provide the desired functionality, and implement it efficiently, without any back-end support (e.g., no shiny server apps).

**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

[https://www.socr.umich.edu/html/SOCR\\_Research.html](https://www.socr.umich.edu/html/SOCR_Research.html)

- **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:** Yongxiang Zhao; Shihang Li; 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](https://github.com/SOCR/ProbDistCalc_RShiny)

# SOCR 2024 MDP Project:

## Data Science Fundamentals: Spacekime Analytics

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

**Website:** <https://tciu.predictive.space> , [www.spacekime.org](http://www.spacekime.org)

**GitHub:** <https://github.com/SOCR/TCIU>

**Training Modules:** ODE/PDE, Kaluza-Klein Theory ([https://en.wikipedia.org/wiki/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](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](http://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 2024 MDP Project: AI for Qualitative-and-Mixed 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 2024 MDP Project: Synthetic Image/Volume Generation

**SOCR Project Leaders:** Kaiming Cheng, Yueyang Shen, Yongkai Qiu, Ivo Dinov

**GitHub:** [https://github.com/SOCR/DL\\_ZNet\\_3D\\_BrainSeg](https://github.com/SOCR/DL_ZNet_3D_BrainSeg)  
[https://github.com/SOCR/DL\\_ZNet\\_2D\\_BrainSeg](https://github.com/SOCR/DL_ZNet_2D_BrainSeg)  
[https://github.com/SOCR/DL\\_AFUnet\\_3D\\_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\\_ZyilFNCPIA6OmFrFuW-A-i0H](https://drive.google.com/drive/folders/1adD9Kp3s_ZyilFNCPIA6OmFrFuW-A-i0H)





## 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>

## 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-identified/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
    -  ImageNet
    -  MNIST
    -  CIFAR-100

-  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
  - [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://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](https://socr.umich.edu/DSPA2/DSPA2_notes/14_DeepLearning.html)

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

# SOCR 2024 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-6uJIPbObl4rVc\\_QDp9i5tD](https://drive.google.com/drive/folders/1hvzoihZMO-6uJIPbObl4rVc_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/](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>

NEW VH feature – **VirtualHospital – Data Marketplace Exchange (DMX) BlockchainContracts** (Vinayak)

- 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/1dtQcVKVkJFSKQVPdAboSesQOy2qLBsTg>
- [PharmaLedger Blockchain Platform](#)
- ...

# SOCR 2024 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/1uZaLGej8NICGfL9NqiFURPvkNIJshQ5F>

**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 2024 MDP Project: Java Applets Code → HTML/JS Apps

SOCR Project Leaders: Ivo Dinov

Website: various

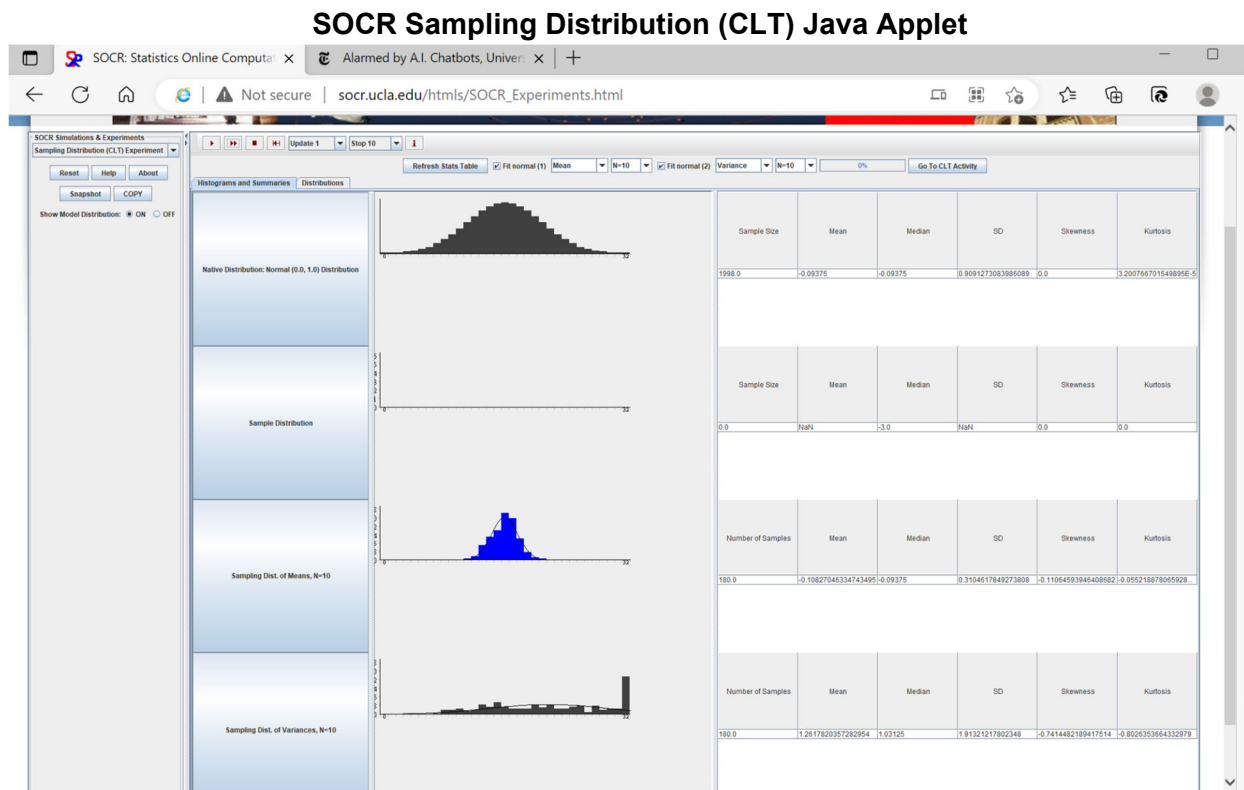
GitHub: <https://github.com/SOCR> & <https://github.com/SOCR/SOCR-Java>

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](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](http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem)



## 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://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ExperimentsActivities)
  - [http://socr.ucla.edu/htmls/exp/LLN\\_Simple\\_Experiment.html](http://socr.ucla.edu/htmls/exp/LLN_Simple_Experiment.html)
- CLT
  - [http://wiki.stat.ucla.edu/socr/index.php/SOCR\\_EduMaterials\\_ExperimentsActivities](http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ExperimentsActivities)
  - [http://socr.ucla.edu/htmls/exp/Sampling\\_Distribution\\_CLT\\_Experiment.html](http://socr.ucla.edu/htmls/exp/Sampling_Distribution_CLT_Experiment.html)
- $\pi$  and  $e$  stochastic estimation experiments:
  - [http://socr.ucla.edu/htmls/exp/Uniform\\_E-Estimate\\_Experiment.html](http://socr.ucla.edu/htmls/exp/Uniform_E-Estimate_Experiment.html)
- Polynomial model fitting: <http://socr.ucla.edu/Applets.dir/SOCRCurveFitter.html>

# SOCR 2024 MDP Project: Convolutional Graph Neural Networks

**SOCR Project Leaders:** TBD, Ivo Dinov

**Website:** TBD

**GitHub:** TBD

## **Description:**

This 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://computational-social-networks.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](https://paperswithcode.com/search?q_meta=&q=graph+neural+networks)
- Test data: [https://umich.instructure.com/courses/38100/files/folder/Case\\_Studies](https://umich.instructure.com/courses/38100/files/folder/Case_Studies)

# **SOCR 2024 MDP Project: Topological Data Analysis (TDA), Persistent Homology, and Betti Numbers for Point-cloud Data**

**SOCR Project Leaders:** TBD, 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\\_SSrjF2jisc9mDQBUCdDKPIY](https://drive.google.com/drive/folders/1QZLjH_xN_SSrjF2jisc9mDQBUCdDKPIY)

## **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.socr.umich.edu/people/dinov/courses/DSPA_notes/05_DimensionalityReduction.html)
  - [http://www.stat.cmu.edu/~jisu/files/20180613\\_SoCG\\_Jisu\\_KIM\\_TDA\\_slide.pdf](http://www.stat.cmu.edu/~jisu/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](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 2024 MDP Project: Novel AI/ML Techniques

**SOCR Project Leaders:** Yueyang, Simeone, Kaiming, Ivo Dinov

**Website:** TBD

**GitHub:** TBD

**Training Modules:** see background papers, R-packages, and tutorials listed below

**GDrive:** <https://drive.google.com/drive/folders/1CSg6xxcwYsjRSV1AjqFNjm1dfTQ5jxeS>

## 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/1CSg6xxcwYsjRSV1AjqFNjm1dfTQ5jxeS>
  - See this talk: <https://www.youtube.com/watch?v=8QLDXiTiRII>
  - <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://umich.instructure.com/courses/38100/files/folder/Case\\_Studies](https://umich.instructure.com/courses/38100/files/folder/Case_Studies)
- Review these concepts:
  - <https://en.wikipedia.org/wiki/Backpropagation>
  - [https://en.wikipedia.org/wiki/Invariant\\_subspace](https://en.wikipedia.org/wiki/Invariant_subspace)
  - Learn to compress and compress to learn
  - DSPA: [dspa.predictive.space](https://dspa.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](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](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](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/>

# SOCR 2024 MDP Project: Synthetic Image/Volume Generation

**SOCR Project Leaders:** Kaiming Cheng, Ivo Dinov

**GitHub:** TBD

**Training Modules:** see background papers, R-packages, and tutorials listed below

**GDrive:** [https://drive.google.com/drive/folders/1adD9Kp3s\\_ZyilFNCPIA6OmFrfuwA-i0H](https://drive.google.com/drive/folders/1adD9Kp3s_ZyilFNCPIA6OmFrfuwA-i0H)










## 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>

## Additional Resources.

- 
- See [DSPA Chapter 22 \(Deep Learning\)](#)
- Papers with Code: <https://paperswithcode.com/task/image-generation>
- Datasets:
  - Brain Images:
    - 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).
    - 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.
  - 2D images:
    -  CIFAR-10
    -  ImageNet
    -  MNIST
    -  CIFAR-100
    -  Cityscapes
    -  CelebA
    -  Fashion-MNIST
    -  CUB-200-2011
    -  STL-10

-  Oxford 102 Flower
- [100\\*100 grid/table of 2D PNG brain \(MRI\) images](#)