## SOCR 2019 MDP Project Summaries

The one-page summaries below describe the main SOCR MDP R&D Projects for 2019 (January-December)

**GDrive:** [https://drive.google.com/drive/folders/16ya0jRQkKU37e9OAhmAfQEFM4qArojrF](https://drive.google.com/drive/folders/16ya0jRQkKU37e9OAhmAfQEFM4qArojrF) (include GSlides)

**SOCR Project Leaders:**
- SOCRAT: Alex Kalinin / others
- BlueML: Syed Husain
- CBDA: Simeone Marino
- DataSifter: Nina Zhou / Simeone Marino
- Data Analytics: Brandon Cummings, Jerome Choi, Yuming Sun, Nina Zhou, Ivo Dinov
- Data Science Fundamentals: Ivo Dinov

**SOCR Trainees/Students**
[https://docs.google.com/spreadsheets/d/1iAUaO0a3Z-P55Cls9higJtabsYb0YxkpjShYZ0jgLqU](https://docs.google.com/spreadsheets/d/1iAUaO0a3Z-P55Cls9higJtabsYb0YxkpjShYZ0jgLqU)

<table>
<thead>
<tr>
<th>Project Summaries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Area</strong></td>
</tr>
<tr>
<td>Programming Subteam: <strong>SOCRAT</strong> (Charts, Wrangler, Modeler, Analyses, Tools)</td>
</tr>
<tr>
<td><strong>TensorFlow.JS</strong></td>
</tr>
<tr>
<td>Analytics</td>
</tr>
<tr>
<td>CBDA TDA DataSifter Biomed/Health Applications (see Case-Studies)</td>
</tr>
<tr>
<td>Data Science Fundamentals</td>
</tr>
<tr>
<td>See <a href="https://medium.com/data-science-club/the-enigmatic-kime-time-complexity-in-data-science-64f8f3e5e6e4">The Enigmatic Kime: Time Complexity in Data Science</a> at the University of Michigan Institute for Data Science (MIDAS) Seminar Series, Slidedeck, YouTube video of this seminar</td>
</tr>
<tr>
<td>498/599 Programming</td>
</tr>
</tbody>
</table>
SOCR 2019 MDP Project: SOCRAT

SOCR Project Leaders: Alex Kalinin / Syed Husain / Ivo Dinov
Website: http://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/1UrNpNDl5sWoXW61YWp02NSv3PBbxfypC

Description
The Statistics Online Computational Resource Analytics Toolbox (SOCRAT) is a Dynamic Web Toolbox for Interactive Data Processing, Analysis, and Visualization. It's purely build 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 BlueJeans meetings
● Code review (pull/push Github requests)
● Join/present the SOCR All-Hands Calls (twice a month, SOCR BlueJeans channel)

References
● Review the websites
SOCR 2019 MDP Project: BlueML

SOCR Project Leader: Syed Husain

Website: ....
GitHub: https://github.com/SOCR/BlueML
Training Modules: https://github.com/SOCR/socr-tutorials
GDrive: https://drive.google.com/drive/folders/1VpHCi25cbNHmMYCKst36XGSRppR5wZsK

Description
The SOCR BlueML project is purely built using HTML5 standards and JavaScript and includes a core library for applying machine learning to high sampling-rate longitudinal data like waveform EEG and EKG. 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). Another example is the Dynamic Visualization Toolkit (https://github.com/SOCR/DVT)

Syed - please check the GitHub security vulnerability message!!! ...
SOCR 2019 MDP Project: 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/1hjwgtgz64A_lUsnRK1gv7mG SJ3HdBHaRW

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 an 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/BlueJeans meetings
● Code review (pull/push Github requests)
● Join/present the SOCR All-Hands Calls (twice a month, SOCR BlueJeans channel)

References

● Review the websites
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 a meaningfully contribute to DataSifter methods, implementation and/or validation
● Coordinate with team

Deliverables
● 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/BlueJeans meetings
● Code review (pull/push Github requests)
● Join/present the SOCR All-Hands Calls (twice a month, SOCR BlueJeans channel)

References
● Review the websites
The SOCR 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 and Python/Jupyter Notebook and the SOCR-Flux Compute Server (https://docs.google.com/document/d/1UmBq_BMiMeUcijvKUCzPeG3tKZaWkinVtKrVWenPK1Y).

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

**References**
- Review the websites and listed resources
SOCR 2019 MDP Project: **Data Analytics - MIMIC-III**

**SOCR Project Leaders:** Brandon Cummings, 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
- Onboarding references: https://drive.google.com/drive/u/1/folders/1Y6Yqq1CuTkHQ5rZg-C9r8_je18nM886l

**GDrive:** https://drive.google.com/drive/folders/1sN1fLYA0oLf1I4e1REJRhlaMD0jXBS7w

**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 Server¹ 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.

¹SOCR-Flux Compute server: https://docs.google.com/document/d/1UmBq_BMiMeUciyjvKUCzPeG3tKZaWkinVtKrVWenPK1Y

**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/BlueJeans meetings
- Code review (pull/push Github requests)
- Join/present the SOCR All-Hands Calls (twice a month, SOCR BlueJeans channel)
SOCR Project Leader: Syed Husain, Chiranjeevi Vegi <vegi@umich.edu>, Ivo Dinov

Website: http://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

Description

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

References
● Review the websites
SOCR 2019 MDP Project: Data Science Fundamentals

SOCR Project Leader: Ivo Dinov

Website: pending
GitHub: NA
GDrive: https://drive.google.com/drive/folders/1PMMBR2bzBPubYMpywLkcTkJPxOKQ4Ag

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.


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

Project Goals
● 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 an 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/BlueJeans meetings
● Code review Join/present the SOCR All-Hands Calls (twice a month, SOCR BlueJeans 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