Scientific computing for structure in big or complex datasets

University of Miami, Department of Computer Science, & the Center for Computational Science

List of Projects

—

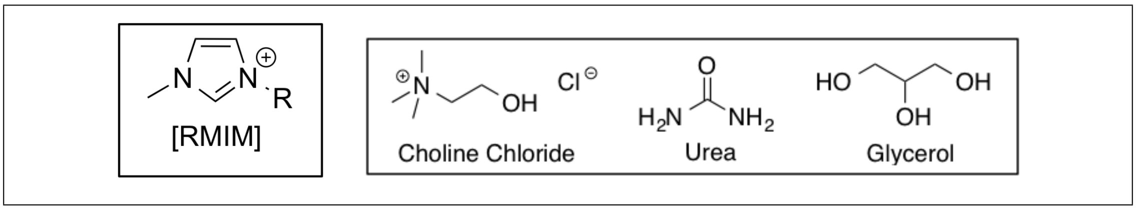
13 Feburary 2017

Introduction. These are the mentored projects for grant CNS 1659144, *REU Site: Scientific Computing for Structure in Big or Complex Datasets*, March 1 , 2017 and ends February 29, 2020. The grant responds to Program Solicitation NSF 13-542, Research Experiences for Undergraduates (REU) Sites and Supplements.

<u>*Project:*</u> *Develop an open source GPU-enabled Monte Carlo program to aid in the design of environmentally friendly chemical solvents, Orlando Acevedo, Associate Professor of Chemistry.*

With the need for more efficient and cleaner technologies becoming increasingly important, the search for alternatives to the most costly and environmentally damaging solvents has become an immediate priority. Conventional solvents account for the vast majority of waste in synthesis and processing, and are generally toxic, flammable, and/or corrosive.[1]. As such, this project allows an REU undergraduate student to participate in our currently funded computational study [4] that seeks to gain a deeper understanding into the use of ionic liquids as alternatives to organic solvents. In addition, a novel solvent class, deep eutectic solvents, will be studied as a potential alternative to ionic liquids.

Ionic liquids are molten salts characteristically at room temperature.[3] Different ion



combinations can dramatically alter solvent properties, including the melting point, viscosity, density, electrical conductance, solvent polarity, and gas solubility.[5] In recent years, a new family of ionic fluids called deep eutectic solvents (DES) has emerged as an environmentally safer and relatively inexpensive alternative to ionic liquids.[6] DESs are made by mixing together a solid quaternary ammonium salt and a hydrogen bond donor, which like ionic liquids, often have melting points below 25 °C.

The REU student's role in the project will be to help develop an open source GPU-enabled mixed quantum and molecular mechanical (QM/MM) Monte Carlo program for simulating these environmentally friendly solvents. We propose the creation of a program, MCGPU (Monte Carlo on Graphics Processing Units), to take advantage of the dramatic power of the GPU [2]. The development of open source MCGPU code should have a large broader impact to the scientific simulation community. Most computational software packages that incorporate QM/MM methodology are not open source and often expensive, particularly to 4-year and masters granting colleges.

Project: Computational Projects in Neuroscience, Vance Lemmon, Professor, Department of Neurological Surgery; Prof. Victor Milenkovic, Computer Science; Prof. Stefan Wufty, Computer Science.

(1) Research in next generation sequencing projects to examine changes in gene expression in neurons after injury and during various regeneration paradigms. The very large data sets are analyzed to examine regulatory regions such as transcription factor binding sites and miRNA binding sites, and to predict the regulatory cascades and key drivers. This project is with Prof. Stefan Wufty of Computer Science.

(2) The development of ontologies and tools to mine and uncover hidden knowledge in the original literature on neuroregeneration and neuroprotection, allowing laboratory scientists to explore the literature and identify novel hypothesis for testing. Students are involved in ontology building and developing ways to identify and transfer findings from the literature to the RegenBase knowledgebase.

(3) Research to visualize, manipulate and analyze very large 3-D images from the Miami Project Light Sheet Fluorescent, and their manipulation. These large images are displayed in realm-time using the specialized hardware and new visualization software at the UM VizLab. The project extends successful display in to 2D to 3D using LEAP IR technology and develop algorithms for identifying and measuring axons. This project is with Prof. Victor Milenkovic of Computer Science.

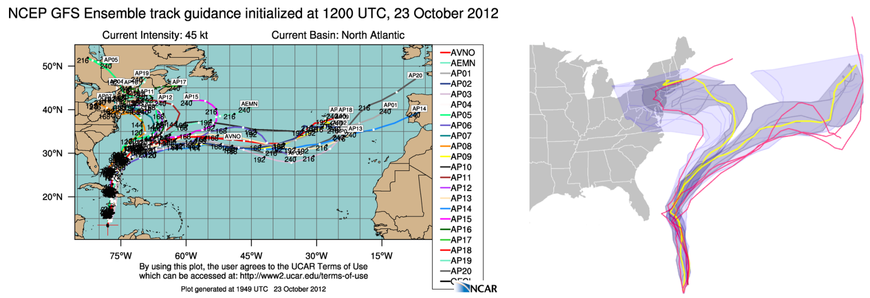
Project: Modeling big behavioral data, Daniel Messinger, Professor of Pediatrics, Electrical and Computer Engineering, and Music Engineering.

Infants develop in the context of social interaction with parents and others. Study of these social interactions typically involves behavioral observation using expert coding systems that rely on subjective judgements. To produce objective measurements of infant and parent behavior, my lab employs computer vision and pattern recognition approaches. These methods produce "big behavioral data" — fine-grained records of expression and other movements in time. To make sense of the underlying interactive process in these observations, we employ computational approaches such as inverse optimal control and statistical approaches such as dynamic time-series models. The ultimate goal is understanding and predicting both typical development and developmental disturbances such as autism spectrum disorder (ASD). To this end, we incorporate common genetic (e.g., dopaminergic) variants in our models of interaction and development.

Students in the lab will be exposed to interdisciplinary research at the borders of behavioral science, neuroimaging, electrical and computer engineering, and human genetics. They will continue to contribute to research topics such as the objective modeling of early emotion expressions, the impact of common genetic variants on empathy and joint attention development, face-to-face infant-parent interactions, the development of attachment, and preschoolers' self-segregation by gender.

Project: Exploring Visualization in Making Sense of Ensemble Data, Mahsa Mirzargar, Assistant Professor Computer Science.

Recent advances in scientific computing have revolutionized data production and analysis. Data-driven analysis is now an integral component of scientific and engineering disciplines especially when computational methods or predictive simulations are used. However, due to various sources of error and uncertainty, it is not uncommon to use multiple outcomes of a model or simulation that are called ensembles. Various scientific disciplines such as uncertainty quantification [4], perception [5], and visualization [2] have demonstrated that proper analysis, summarization, and presentation of ensemble data can significantly facilitate the knowledge-acquisition and decision making process. The goal of this project is to provide an introduction in the state-of-the-art computational and visualization tools and techniques that are widely used invarious disciplines to explore, find, and summarize features and structures in an ensemble. The ultimate goal is to prepare the REU student(s) for real-world data-intensive problems and hopefully motivate them to pursue a graduate level career in this new and demanding area.



(a) (b)

FIGURE 1. Sandy Hurricane ensemble publicly available as part of Tropical Cyclone Guidance Project at NCAR [6]: (a) Direct ensemble visualization, (b) Curve boxplot visualization [3] as one of the state-of-the-art techniques that is capable of capturing and highlighting the multimodal nature of the data.

This research experience aims to demonstrate the value of effective visual exploration to assist experiment design, quality control of data, hypothesis testing, and overall understanding of data. In this project, the mentor will utilize and adopt selective stat-of-the-art ensemble exploration and visualization techniques to study structures in ensemble data. The datasets for this project will be selected from publicly available datasets from UCI machine learning database (`http://archive.ics.uci.edu/ml/`), weather, and hurricane forecast datasets available from NOAA (`http://www.spc.noaa.gov/exper/`) and NCAR (`http://www.ral.ucar.edu/guidance/`) as examples of real-life datasets. Through hands-on-experience, the students will learn concepts such as data cleaning, statistical summarization, basic machine learning algorithms, along with their use-cases.

The mentor has previous experience in working with REU students as part of her post-doctoral training that has resulted in a publication at International Journal for Uncertainty Quantification with Ms. Hudson (REU student) [1]. Depending on the level of familiarity and interest in analytical and software development aspects, students can participate in the following activities:

- Using available data visualization tools to help identify and explore different patterns in ensemble data.
- Utilizing the state-of-the-art machine learning algorithms (e.g., clustering) to find structures in heterogeneous datasets.
- Utilize GPU programming toolkits (e.g., ArrayFire [7]) to accelerate selective set of data exploration and visualization techniques to analyze large-scale ensemble data.
- Working collaboratively with the mentor in building visual exploration environment to enable analysis of ensemble data.

Project: Community attacks in GPU enabled cryptography, Burton Rosenberg, Associate Professor Computer Science.

When deploying a cryptographic primitive, such as public-key cryptography, issues not apparent when focused on only the mathematics or protocol, can arise through unintended structure appearing in the scale up. Password strength is often measured by entropy, but this measure does not fully account for the attacker's game — not the guess of a particular password but the guess of some password. Other measures of entropy, such as guesswork, account for the ensemble structure of the password data [5].

Likewise, flaws in the creation of RSA keys, which are innocuous in small datasets, become fatal when the resulting community of keys is evaluated ensemble [6]. Ensemble thinking about cryptography also reminds the algorithm designer that pre-computation trade-offs can be rethought to provide avenues of attack when attacking the entire community rather than a single instance [1].

As a result, it is good to review cryptographic algorithms for those that might benefit from thinking about the community of data. Among cryptographic algorithms, this project will focus on the subfield elliptic curve cryptography, as there are several new opportunities for speed up, for instance the use of Edwards Curves [3],[4] and Twisted Edwards curves [2].

Project participants will write code and contribute to libraries that support GPU programming for elliptic curve cryptography, and relate these advances to security by applying the high performance applications to the ensemble of deployed data.

Project: Neural processing of visual scenes with large-scale deep neural networks, Odelia Schwartz, Associate Professor Computer Science.

Visual processing in the brain has traditionally been studied with simple stimuli such as bars of light. While a lot has been learned by this approach, neurons are highly nonlinear, and it is difficult to extrapolate from studies with simple stimuli to understanding how the brain processes complex natural scenes in the environment. Understanding how the brain processes natural scenes has in the past been considered a barrier in the field. However, there is exciting potential to overcome this barrier based on two recent advances.

First, we have recently applied scene statistics models to understanding how primary visual cortex neurons process natural scenes, with particular emphasis on modeling neural nonlinearities via unsupervised learning of a canonical computation known as divisive normalization [1][3].Second, there has been tremendous recent progress in scene recognition in machine learning, particularly with deep convolutional neural networks [5]. Such deep learning approaches have shown remarkable ability to explain cortical processing in high level visual object recognition areas [4] [6].

Deep learning intriguingly benefits from divisive normalization. However, the nonlinearities in such models have been chosen in a fixed and somewhat ad hoc manner and not learned from data. We propose that the brain has adopted a more optimal strategy for choosing nonlinearities based on the statistical regularities of its inputs. Students will work with models that integrate deep convolutional networks (e.g., AlexNet of Krizhevsky et al. 2012) with our approach for learning divisive normalization. Students will explore different versions of model, including comparisons with purely unsupervised learning and different forms of divisive normalization [2].

More broadly, our group is engaged in testing the models experimentally, via collaborations. In addition to an ongoing neurophysiology collaboration with Albert Einstein College of Medicine, we plan to apply fMRI analysis techniques with Psychology faculty at University of Miami, who have experience running fMRI tasks with complex visual stimuli such as faces and other objects. Students working on this project will be able to also interact with our experimental collaborators and thus be exposed to the broader neuroscience field.

The machine learning is done on large-scale image databases such as Imagenet, using Python and Matlab. For deep learning, we are working with the Berkeley Caffe and with TensorFlow. The unsupervised learning of divisive normalization is currently implemented by our group in Matlab. All the large-scale simulations require GPU.

<u>Project:</u> Brain connectivity in typical and atypical development, Lucina Q. Uddin, Assistant Professor of Psychology.

The emerging consensus that dynamic interactions between brain regions constitute the neural basis of cognition has made the examination of large-scale brain networks a priority for cognitive neuroscience research. While the majority of extant studies have focused on localizing cognitive functions to specific brain regions, it is increasingly evident that a network approach can provide unique insights into complex brain and cognitive processes. The overarching goal of our research program is to understand the neural bases of high-level cognitive processes. Within a cognitive neuroscience framework, our research combines functional connectivity analyses of resting-state functional magnetic resonance imaging data and structural connectivity analyses of diffusion tensor imaging data to non-invasively examine the organization of large-scale human brain networks. Current projects focus on understanding dynamic network interactions underlying cognitive flexibility in neurodevelopmental disorders such as autism.

Students will be mentored on projects utilizing publicly available neuroimaging datasets (`https://www.humanconnectome.org/`). In collaboration with graduate students and postdoctoral fellows in the lab, students will learn to formulate hypotheses, analyze neuroimaging data using open source software, create publication-quality figures, and present results in written and oral format.

References Cited

References

References (Prof. Acevedo)

[1] Paul Anastas and Nicolas Eghbali. Green chemistry: Principles and practice. *Chem. Soc. Rev.*, 39:301–312, 2010.

[2] Orlando Acevedo et al. Monte carlo on graphics processing units. `http://www.github.com/orlandoacevedo/MCGPU`, 2015.

[3] Michael Freemantle. *An Introduction to Ionic Liquids*. The Royal Society of Chemistry, 2009.

[4] Orlando Acevedo PI. Chemical theory, models, and computation. *National Science Foundation CHE–1562205*, July 15, 2015 — June 30, 2018. $325,000.

[5] Hermann Weingärtner. Understanding Ionic Liquids at the Molecular Level: Facts, Problems, and Controversies. *Angewandte Chemie International Edition*, 47(4):654–670, January 2008.

[6] Qinghua Zhang, Karine De Oliveira Vigier, Sebastien Royer, and Francois Jerome. Deep eutectic solvents: syntheses, properties and applications. *Chem. Soc. Rev.*, 41:7108–7146, 2012.

References (Prof. Mirzargar)

[1] Yanyan He, Mahsa Mirzargar, Sophia Hudson, Robert M Kirby, and Ross T Whitaker. An uncertainty visualization technique using possibility theory: Possibilistic marching cubes. *International Journal for Uncertainty Quantification*, 5(5), 2015.

[2] Suresh K Lodha, Catherine M Wilson, and Robert E Sheehan. Listen: sounding uncertainty visualization. In *Proceedings of the 7th conference on Visualization'96*, pages 189–195. IEEE Computer Society Press, 1996.

[3] Mahsa Mirzargar, Ross T Whitaker, and Robert M Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE transactions on visualization and computer graphics*, 20(12):2654–2663, 2014.

[4] William L Oberkampf and Christopher J Roy. *Verification and validation in scientific computing*. Cambridge University Press, 2010.

[5] Chris MR Smerecnik, Ilse Mesters, Loes TE Kessels, Robert AC Ruiter, Nanne K De Vries, and Hein De Vries. Understanding the positive effects of graphical risk information on comprehension: measuring attention directed to written, tabular, and graphical risk information. *Risk Analysis*, 30(9):1387–1398, 2010.

[6] J. L. Vigh. Real-time operational tropical cyclone guidance data. Tropical Cyclone Guidance Project (TCGP), National Center for Atmospheric Research, Research Applications Laboratory, Boulder, Colorado. `http://www.ral.ucar.edu/hurricanes/repository`, 2016.

[7] Pavan Yalamanchili, Umar Arshad, Zakiuddin Mohammed, Pradeep Garigipati, Peter Entschev, Brian Kloppenborg, James Malcolm, and John Melonakos. ArrayFire - A high performance software library for parallel computing with an easy-to-use API, 2015.

References (Prof. Rosenberg)

[1] David Adrian, Karthikeyan Bhargavan, Zakir Durumeric, Pierrick Gaudry, Matthew Green, J. Alex Halderman, Nadia Heninger, Drew Springall, Emmanuel Thomé, Luke Valenta, Benjamin VanderSloot, Eric Wustrow, Santiago Zanella-Béguelin, , and Paul Zimmermann. Imperfect forward secrecy: How diffie-hellman fails in practice. In *ACM Conference on Computer and Communications Security (CCS'15)*, 2015. to appear.

[2] Daniel J. Bernstein, Peter Birkner, Marc Joye, Tanja Lange, and Christiane Peters. Twisted edwards curves. In *Proceedings of the Cryptology in Africa 1st International Conference on Progress in Cryptology*, AFRICACRYPT'08, pages 389–405, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Daniel J. Bernstein, Peter Birkner, Tanja Lange, and Christiane Peters. ECM using edwards curves. *Math. Comput.*, 82(282), 2013.

[4] Daniel J. Bernstein, Tien-Ren Chen, Chen-Mou Cheng, Tanja Lange, and Bo-Yin Yang. Ecm on graphics cards. In *Proceedings of the 28th Annual International Conference on Advances in Cryptology: The Theory and Applications of Cryptographic Techniques*, EUROCRYPT '09, pages 483–501, Berlin, Heidelberg, 2009. Springer-Verlag.

[5] Joseph Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 538–552, Washington, DC, USA, 2012. IEEE Computer Society.

[6] Arjen K. Lenstra, James P. Hughes, Maxime Augier, Joppe W. Bos, Thorsten Kleinjung, and Christophe Wachter. Ron was wrong, whit is right. Cryptology ePrint Archive, Report 2012/064, 2012. `http://eprint.iacr.org/2012/064`.

References (Prof. Schwartz)

[1] RC Cagli, A Kohn, and O Schwartz. Flexible gating of contextual modulation during natural vision. *Nature Neuroscience*, 8(11):1648–55, 2015.

[2] RC Cagli and O Schwartz. The impact on mid-level vision of statistically optimal divisive normalization in v1. *Journal of Vision*, 13(8):1–20, 2013.

[3] M Carandini and D Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[4] N. Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446, 2015.

[5] A Krizhevsky, I Sutskever, and GE Hinton. Imagenet classification with deep convolutional neural networks. *I. Neural Information Processing Systems*, 25:1097–1105, 2012.

[6] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, and J. J. Seibert, D.and DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.