1 Collaboration and Evaluation

Our team consists of University of California San Diego (UCSD) faculty members from the departments of Computer Science and Engineering (Yoav Freund), Electrical and Computer Engineering (Piya Pal, Peter Gerstoft), and Mathematics (Alex Cloninger, Rayan Saab), directly representing three of the four communities of HDR TRIPODS. Additionally, all team members are founding members of UCSD's Halicioglu Data Science Institute (HDSI), a new academic unit at UCSD with a focused mission. Its mission, one that it shares to a large extent with HDR TRIPODS, is to lay the groundwork for the scientific foundations of this emerging discipline, develop new methods and infrastructure, and train students, faculty and industrial partners to use data science in ways that will allow them to solve some of the world's most pressing problems `https://datascience.ucsd.edu/` the HDSI website.

Having all our team members located on the same campus and regularly interacting through HDSI will greatly enhance our ability to collaborate closely on the proposed projects. In addition to the direct and frequent interaction of the faculty members involved in this proposal, with the NSF's support, we intend to jointly supervise PhD students and to co-mentor a postdoctoral fellow.

Additionally, we note that our team includes theoreticians who also have a deep interest in and knowledge of applications, and all have worked closely and published papers with practitioners. In the context of this proposal, Co-PI Gerstoft's intimate knowledge of sensors and sensor networks will allow us to account for practical issues that arise in data science, including (among others) its multi-modality and incompleteness. Our awareness of practical issues will hopefully allow us to maximize the impact of the algorithms and methodologies that result from our work.

In what follows, we will describe our expertise, plans for advising joint students, for collaborating, as well as for evaluating the results of our work.

2 Expertise

- **PI Freund's** expertise is in Computational Learning Theory, and statistics. Among his theoretical work are Boosting [22], statistical analysis of the generalization error for ensemble classifiers [23, 5], online learning [2, 7], learning and game theory [4, 8, 6] Learning low dimensional structures [3]. In addition, PI Freund has worked on applications of machine learning to image analysis [15] and, in particular, image analysis for biological microscopy [12, 16, 1, 13, 25].
- **Co-PI Cloninger's** expertise is in applied harmonic analysis and the analysis of high dimensional data. He focuses on approaches that model the data as being locally lower dimensional, including data concentrated near manifolds or subspaces. These types of problems arise in a number of scientific disciplines, including imaging, medicine, and artificial intelligence, and the techniques developed relate to a number of machine learning and statistical algorithms, including deep learning, network analysis, and measuring distances between probability distributions.
- **Co-PI Gerstoft's** focus on data-driven computational geophysics and within these fields I further cover the subtopics: applied statistical signal processing, inverse methods [9], mathematical models, extracting information from noise [21, 10], and machine learning[18, 14]. I currently focus on developing new sensing techniques using large array sensor data. Data science methods (big data/machine learning) and compressive sensing[24, 11] is a main focus for sensing the physical environment. These techniques are applied to observing tsunamis, earthquakes, traffic, Antarctic signals, as well as extracting environmental information from just noise.
- **Co-PI Pal's** primary research interests are high dimensional statistical signal processing and big data analysis, energy-efficient sensing, high resolution imaging, and optimization techniques for solving inverse problems in signal processing with applications in radar and sonar signal processing, biomedical and molecular imaging, and machine learning. More specifically, I focus on developing new energy-efficient sampling

and sensing techniques for acquiring and processing high dimensional signals, and understanding their fundamental performance limits.

- **Co-PI Saab's** expertise is in applied and computational harmonic analysis, and in mathematical signal processing. He is interested in, and has published extensively on questions regarding efficient acquisition, quantization, representation, and processing of data. He has studied these problems in classical contexts, such as those of band limited functions, but also in modern contexts, such as compressed sensing of structured signals, be they sparse vectors, low-rank matrices, or signals from arbitrary sets. He has also worked on signal and data processing problems related to multiple sensors, including the blind source separation problem. In his work, Saab uses and develops a variety of tools from high dimensional probability theory, applied harmonic analysis, convex analysis and optimization, and quantization theory, among others.

3 Collaboration Plan

Joint Students and Postdoctoral Scholar

In addition to its focus on the proposed scientific content, a main goal of this proposal is to recruit and train highly qualified personnel for the workforce of tomorrow. Exploiting the fact that our team members are all UCSD faculty with affiliations to HDSI, we will recruit and advise graduate and post-graduate students as a group. We will actively recruit students who have a strong mathematical background and are able and willing to implement algorithms as reusable code. Equally importantly, we will actively seek out students from underrepresented groups in STEM.

The Post Doc will assist PI Freund in keeping track of all activities and work with each of the four PIs on a per need base as well as connect with some of the graduate students. We will make sure that post Doc is actively involved in the project.

Collaboration

The PIs' close proximity by virtue of their affiliation to UCSD and HDSI will facilitate collaboration through regular meetings. Furthermore, the graduate students will glue our team together and each graduate student will have at least two PI-advisers. Each graduate student will have a one hour meeting with their two advisors each week. The student and both advisors will also have regular bi-weekly meetings to assess progress and ensure continuing synergy on each project. In addition, we will have a weekly group meeting with all PIs, postdoc, and students, and where one of the team members will present their work.

Finally, we will have a kick-off retreat and followed by annual retreats where all the team members meet to assess our progress on the proposed work.

Datasets:

PI Gerstoft are familiar with all these data sets and will help the other PIs to access the data. We will use several benchmarks and datasets to evaluate the algorithm we develop:

(1) We intend to compare our performance on the DCASE challenge dataset[17]: Two tasks in acoustics and audio signal processing in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge. see http://dcase.community/challenge2019/index This is an annual event, so new datasets will be available, See http://dcase.community/challenge2019/index for updates on the 2019 Challenge

(2) Closely related to item (1), but giving us better freedom to demonstrate our methods, will we collect audio data in Prof. Christiansen's instrumented living quarter (at Computer Science at UCSD). We will evaluate out algorithms on the collected data.

(3) For seismic data it is now common to use massive sensor data. We have access to the 5000 geophones collected in a 7x10 km area Long Beach over months at a sampling rates of 500 Hz. We have worked on the data in e.g. [20]. One week of the data used takes 2TB of disk space and apart from geophysical studies

can be used to monitoring traffic[19]. Many other data sets are available at the Iris website www.iris.edu, as NSF requires all seismic data obtained with NSF funding to be uploaded on Iris. This data is frequently used by PI Gerstoft and will share it with the PIs.
(4) For electromagnetic data we have access to MIMO array data capable of taking phase coherent measurements for the determination and characterization of electromagnetic (EM) channels (1–6 GHz). The array aperture is intended to be flexible, with a maximum inter element spacing around 5 wavelengths. The MIMO array consists of an 8 element source array and 50 element receiver array. The array data is available and will be Collected by PI Gerstoft.
(5) An extreme data sensor network is available by using an ad-hoc network of cell-phones

bf Kaggle data challenge
Forecasting earthquakes is one of the most important problems in Earth science because of their devastating consequences. Kaggle has launched a Change based on earthquake prediction on laboratory experiments https://www.kaggle.com/c/LANL-Earthquake-Predictionhttps://www.kaggle.com/c/LANL-Earthquake-Prediction. We envision that we can can combine our method with our own data set for a Kaggle Challenge.

Industrial collaboration PIs Gerstoft and Feund collaborate with startup company Occuspace on Predictive model of campus foot traffic. Since August 2017, the UC San Diego Library had been beta testing a program called Waitz that installed occupancy sensors on each floor. Waitz sensors are unique in that they provide occupancy estimates by measuring mobile device signals (similar to bluetooth/wifi beacons). This information was put on the library's homepage to try to drive student traffic to less occupied areas for study space.

Evaluation
We plan to submit paper to the following conferences and journals: Science, Nature, COLT, ICML, AIstat, JMLR, IEEE Signal Processing, IEEE Information theory, Applied and Computational Harmonic Analysis, Foundations of Computational Mathematics, Information and Inference, SIAM Journal on Mathematics of Data Science, please add.
We Will implement our algorithms in python and jupyter notebooks and make them available through GitHUB.
We will evaluate our performance in terms of:

1. Number and quality of publications.
2. Number and quality of (open source) implementations.
3. Performance of our methods on benchmark data.

Capability and capacity for a potential Phase II
We suggest to call our phase II institute for **Institute for Learning Sensor Network** All PIs are members of the Haligliou Data Science Institute (HDSI). Thus we have the space and logistic support to form an effective phase II effort.
We plan to join forces with researchers in the areas of embedded computers and robotics for sensor development. Domain science researchers at both UCSD Scripps Institute of Oceanography and medical sensing will benefit from our developments.
We collaborate with many researchers nationally and internationally, both in mathematics/computer science as well as Domain science in acoustics, seismics, and electromagnetics. We expect that establishing this **Institute for Learning Sensor Network** will increase the collaboration with many researchers visiting the institute.

References

[1] P. A., L. D., K. D., B. W., F. Y., and M. W. Visualization of individual scr mrnas during drosophila embryogenesis yields evidence for transcriptional bursting. *Current Biology*, Nov. 2009.

[2] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, May 1997.

[3] S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7), July 2009.

[4] Y. Freund, M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, and R. E. Schapire. Efficient algorithms for learning to play repeated games against computationally bounded adversaries. In *36th Annual Symposium on Foundations of Computer Science*, pages 332–341, 1995.

[5] Y. Freund, Y. Mansour, and R. E. Schapire. Generalization bounds for averaged classifiers. *The Annals of Statistics*, 2004.

[6] Y. Freund and M. Opper. Drifting games and Brownian motion. *Journal of Computer and System Sciences*, 64:113–132, 2002.

[7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, Aug. 1997.

[8] Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

[9] P. Gerstoft. Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions. *The Journal of the Acoustical Society of America*, 95(2):770–782, 1994.

[10] P. Gerstoft, K. G. Sabra, P. Roux, W. Kuperman, and M. C. Fehler. Green's functions extraction and surface-wave tomography from microseisms in southern california. *Geophysics*, 71(4):SI23–SI31, 2006.

[11] P. Gerstoft, A. Xenaki, and C. Mecklenbräuker. Multiple and single snapshot compressive beamforming. *J. Acoust. Soc. Am.*, 138(4):2003–2014, 2015.

[12] G. Giannone, B. J. Dubin-Thaler, O. Rossier, Y. Cai, O. Chaga, G. Jiang, W. Beaver, H.-G. Dobereiner, Y. Freund, G. Borisy, and M. P. Sheetz. Lamellipodial actin mechanically links myosin activity with adhesion-site formation. *Cell*, 128:561–575, February 2007.

[13] V. I., S. A.Y., D. J.D., M. D.W., F. Y., and K. D. Automatic identification of fluorescently labeled brain cells for rapid functional imaging. *Journal of Neurophysiology*, July 2010.

[14] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft. Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, 90(1):3–14, 2018.

[15] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *Proceedings of the International Conference on Computer Vision*, 2003.

[16] R. Liu, Y. Freund, and G. Spraggon. Image-based crystal detection: a machine-learning approach. *ActaCryst D.*, 64(12):1187–1195, 2008.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *Worksh. Detect. Class. Acoust. Scenes and Events*, 2017.

[18] H. Niu, E. Reeves, and P. Gerstoft. Source localization in an ocean waveguide using supervised machine learning. *The Journal of the Acoustical Society of America*, 142(3):1176–1188, 2017.

[19] N. Riahi and P. Gerstoft. The seismic traffic footprint: Tracking trains, aircraft, and cars seismically. *Geophysical Research Letters*, 42(8):2674–2681, 2015.

[20] N. Riahi and P. Gerstoft. Using graph clustering to locate sources within a dense sensor array. *Signal Processing*, 132:110–120, 2017.

[21] K. G. Sabra, P. Gerstoft, P. Roux, W. Kuperman, and M. C. Fehler. Surface wave tomography from microseisms in southern california. *Geophysical Research Letters*, 32(14), 2005.

[22] R. E. Schapire and Y. Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

[23] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

[24] A. Xenaki, P. Gerstoft, and K. Mosegaard. Compressive beamforming. *J. Acoust. Soc. Am.*, **136**(1):260–271, 2014.

[25] C. Y, M. LE, T. AS, F. D, F. B, M. PP, K. HJ, F. Y, and K. D. An active texture-based digital atlas enables automated mapping of structures and markers across brains. *Nature Methods*, 16(9), April 2019.