Research Statement

The main question I have been addressing in my research career is “How does one extract information from complex data?” Answers range from designing efficient algorithms and proving their efficiency to establishing performance bounds.

My research focus is at the interface of statistics, information theory, applied probability and computer engineering. Over the years, both statistics and information theory have deeply fascinated me, as has the interplay between these disciplines, and how this interplay applies to answering questions in data analysis and machine learning. I am interested in developing methods, mostly algorithms, but I am equally interested in proving mathematical results associated with these algorithms. As main author, most of my publications contain simultaneously a theoretical contribution as well as a practical one. Ethical considerations are of prime importance for me. I choose my research questions very carefully, and always bear in mind the scientific, technological and societal importance raised by these questions, as well as my capacity to improve upon the state of the art. Opportunities to collaborate with other researchers are also a defining factor. To date, my most significant achievements may have come from fruitful collaborations with other researchers who are exceptional thinkers and who work in complementary specialties. Over the years, I have grown increasingly interested in building larger teams of researchers, which include undergraduate and graduate students as well as junior and senior scientists, to explore broader scientific questions.

Initiated approximately thirty years ago [1], Pattern Theory is aimed at analyzing patterns from a statistical point of view in all signals generated by the world, whether they be visual, acoustic, textual, molecular (e.g., DNA strings), neural, and so forth. Patterns are described using hidden variables, together with their probability distributions, whereas signals, or relevant functions of the signals, are modeled conditionally on the hidden variables. In principle, the detection of patterns in noisy and ambiguous samples can then be achieved by applying Bayes's rule. The philosophy of pattern theory is a driving force which has guided my research efforts. Within each application, the conceptual workflow is as follows: building a prior model for a pattern, building the likelihood function for the observed data or functions of the observed data given the pattern, designing an efficient algorithm for computing an estimator of the pattern and establishing the validity and the computational performance of the resulting procedure. I have carried all or part or a variant of this workflow for the following patterns: one dimensional structures within 2D images [2], the texture of skin within 2D images [3], faces within 2D images [7], surgical tools within video-sequences [10], words within spoken text [4], spike trains from calcium imaging [5], anatomical landmarks within Brain magnetic resonance imaging (MRI) [8], granulomas in Tuberculosis infected lungs from joint structural and molecular imaging [9,6] and the time progression of Alzheimer's disease [12].

Here are two collaborative projects I plan to pursue. The first one is a continuation of previous work and concerns variants of the game of 20 questions (a.k.a. the Ulam-Renyi game) and their applications. The second one is a new project about computational medicine and specifically the computational study of neurodegenerative disease.

The game of 20 questions or Ulam-Renyi game: Let us consider the following set-guessing problem. Let Ω be a fixed set and $S \subset \Omega$ be a subset containing k elements. One can sequentially choose subsets of Ω, notated $V_1, V_2, \ldots$ and query the number of elements in the intersections $V_1 \cap S, V_2 \cap S, \ldots$. The goal is to find S as accurately as possible while minimizing the number of subsets, or questions. The famous Ulam-Renyi game was raised by Ulam [15] and first investigated by Renyi [16]. Carole (an anagram for oracle) chooses a number between one and N, and Paul (a.k.a. Paul Erdős) chooses a sequence of subsets to query in order to find this number. Moreover, Carole can answer either YES or NO and is allowed to lie for at most a given number of times. What is then the minimum number of questions needed in the worst case or in average for Paul to find out Carole's number? The hand-waiving answer is that in both cases a number of questions proportional to $\ln N$ is necessary and sufficient. This is a remarkable result which is closely related to the original results obtained in sparse signal recovery [13, 14]. Classical applications include group testing and block coding. Together with my collaborators, we have contributed to solving variants of the Ulam-Renyi game using information theory and optimal control [11]. We have also opened a new field of applications in machine perception, which we briefly describe next.

I consider machine perception as an efficient mechanism aimed at reducing uncertainty; as do others. A concrete example is as follows: consider the task of locating a front facing standard size face within an image, this location being by definition fully characterized by the pixel location of the nose. As in Bayesian statistics, this location is described by a random variable, which distribution over the set pixel locations has large entropy. We assume that a collection of unit cost questions are available. Each question is parameterized with the coordinates of a sub-image. The answer which is a numerical value is obtained by computing a function of the image values within this sub-image and is modeled as a noisy answer to the question "does the face belong to this sub-image?" Which sub-images should then be chosen and in which order such that one would detect the face while minimizing the average or worse case number of queries? I have contributed in applying this framework to the tracking of roads from remote sensing data [1], the detection of faces from single images [7], the tracking of surgical tools from video sequences [10], and the development of efficient protocols in electron microscopy [17]. I Am currently applying the same framework for scene annotation [19].

I plan to further study problems in optimal search, detection, and interrogation in a Bayesian decision-making framework, developing search algorithms with both theoretical guarantees and strong empirical performance. Using the 20 questions mathematical framework I plan to study specific problems in computer vision focusing on target tracking and detection, in screening, in simulation optimization, in machine learning and in experimental psychology for understanding visual search.

Computational study of neurodegenerative disease: Neurodegenerative diseases include Alzheimer's (AD), Parkinson's (PD), Amyotrophic lateral sclerosis (ALS), (often referred to as Lou Gehrig's disease), Huntington's (HD) as well as other rarer diseases. In the USA, 1 in 3 seniors dies with AD or another dementia [20]. These diseases are characterized by the degeneration and death of neurons and despite huge efforts and investments, there is still no disease modifying treatment for neurodegeneration. However, the medical community has recently released large and comprehensive datasets which open the way to the computational study of neurodegenerative diseases. There is at least one de-identified publicly available large free dataset for each of the diseases above. The data are very rich and complex. They contain 100 – 10,000 subjects, often several visits (1-20) per subject, clinical and historical medical records, protein concentrations, brain MRI – both raw and pre-processed with volumes, thickness and shape measurements, functional imaging – functional MRI and molecular imaging, multiple cognitive tests as well as genetic data. In particular, whole genome sequencing is currently available for more than 800 subjects of the AD neuroimaging initiative (ADNI) cohort.

The scientific goal is to understand the disease process by combining these measurements. This would allow for (1) reducing the costs of clinical trials by identifying disease sub-types and detect subtle changes in disease progression, and (2) assessing exams and treatment options on an individual basis and predict the time to clinical events such as the time of onset of dementia.

I have led such effort for AD. I have been able to verify from statistical analysis of the ADNI data the "Cliff Jack curves" [12] which reveal a cascade of biomarkers leading to the AD dementia. I plan to further develop this research and identify (Ulf Grenander's) patterns in neurodegenerative diseases from large and heterogeneous data.

References:

1. Grenander, Ulf (1994). General Pattern Theory. Oxford Science Publications.
2. D. Geman and B. Jedynak, ``An active testing model for tracking roads from satellite images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 18, pp. 1-14, 1996.
3. B. Jedynak, H. Zheng and M. Daoudi, ``Skin Detection using Pairwise Models", Image and Vision Computing, 23(13), pp. 1122-1130, Nov. 2005.
4. Jedynak, Bruno M., and Sanjeev Khudanpur. "Maximum likelihood set for estimating a probability mass function." Neural computation 17(7) (2005): 1508-1530.
5. J. Vogelstein, B. Watson, A. Packer, B. Jedynak, R. Yuste, and L. Paninski, "Spike inference from Calcium Imaging using Sequential Monte Carlo Methods", Biophysical journal, 97(2), 2009. PMCID: PMC2711341.
6. Stephanie L. Davis, Eric L. Nuermberger, Peter Um, Camille Vidal, Bruno Jedynak, Martin G. Pomper, William R. Bishail, and Sanjay K. Jain, "Non-invasive pulmonary [18F]-2-uoro-deoxy-D-glucose positron emission tomography correlates with bactericidal activity of tuberculosis drug treatment", Antimicrobial Agents and Chemotherapy, 53(11), 2009. PMCID: PMC2772305
7. Raphael Sznitman and Bruno Jedynak , "Active Testing for Face Detection and Localization", IEEE Pattern analysis and Machine Intelligence, 32(10), 2010. PMID: 20479494.
8. C. Vidal and B. Jedynak, "Learning to Match: Deriving optimal template-Matching algorithms from Probabilist Image Models", International Journal of Computer Vision, 88(2), 2010.
9. C. Vidal, D. Beggs, L. Younes, S. K. Jain and B. Jedynak, "Incorporating User Input in Template-Based Segmentation" , proceedings of ISBI'11.
10. R. Sznitman, R. Richa, R. Taylor, B. Jedynak and G. Hager. "Unified detection and tracking of instruments during retinal microsurgery". IEEE TPAMI: 35 (5), 1263-1273, 2013.
11. B. Jedynak, P. Frazier and R. Sznitman, "Twenty questions with noise: Bayes optimal policies for entropy loss", Journal of Applied Probability, 49(1), March 2012.

12. B. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B. T. Wyman, D. Raunig, David, C. P. Jedynak, B. Caffo and J. Prince for the Alzheimer's Disease Neuroimaging Initiative, "A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort", NeuroImage, 63(2), August 3rd 2012, PMID: 22885136.
13. EJ Candès, J. Romberg, T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", IEEE Trans. Information Theory, 52 (2), 489-509.
14. Donoho, David L. "Compressed sensing." IEEE Trans. on Information Theory, 52(4) (2006): 1289-1306.
15. S. Ulam, Adventures of a Mathematician. New York: Charles Scibner's Sons, 1976.
16. A. Renyi, A Diary on Information Theory. Akademiai Kiado, 1984.
17. R. Sznitman, A. Lucchi, P. I. Frazier, B. Jedynak and P. Fua, "An Optimal Policy for Target Localization with Application to Electron Microscopy", International Conference of Machine Learning (ICML), 2013.
18. Jack Jr, Clifford R., et al. "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade." The Lancet Neurology 9(1) (2010): 119-128.
19. Erdem Yoruk, Ehsan Jahangiri, Rene Vidal, Bruno Jedynak, Laurent Younes, and Donald Geman, "Entropy Pursuit for Scene Annotation", work in progress.
20. Alzheimer's Association 2013 Alzheimer's Disease Facts and Figure.