# BIGDATA: Collaborative Research:F: Manifold Hypotheses and Foundations of Data Science with Applications

It is often said that progress in science is characterized by successive steps of measurement, *arithmetization*, *algorithmization*, and *algebraization* – each step representing in a succinct manner the intuitions collected in the earlier step. In sciences, various breakthrough in technology, e.g., sequencing, high-throughput measurement of DNA/RNA abundance, electron and scanning tunneling microscopy, astronomical studies, space telescopes, collection of social-media data, on-line observations of human interactions, etc., have made it possible to obtain a quantitative arithmetic picture of the "states" of a complex structure (cell, organism, population, social groups, universe) at a certain instant and under certain conditions. As the complexity of the systems studied have scaled, the "bigness" of the data has grown spectacularly; many scalable exact and approximate algorithms have been proposed; a unifying foundational study of the emerging "data science" has become prominent, and yet, it shies away from the final step of the algebraization of data sciences. Such an approach could center around the so-called "Manifold Hypothesis," that seeks a differential algebraic structure in the state-space – to be inferred from the sampled data point clouds.

We wish to build on "Manifold Hypothesis" to create an algebraic (geometric/topological) investigation of existing and emerging Big Data approaches in computer science, statistics, computational science, and mathematics, along with innovative applications in domain sciences, namely, cancer biology, linguistics and the physical sciences that lead towards the further development of the field of data science. Thus, the main emphasis of our study is "Foundational" (F): focusing on fundamental theories, techniques, methodologies, technologies of broad applicability to Big Data problems.

**Intellectual Merit:** The vision for this project rests on the growing importance of data science and its multifaceted impacts, such as on genomics, Internet, society, astronomy and cosmology, where the engineered system's ability to generate quantitative data far supersedes the algorithmic and computational resources, and on the belief that the most efficient and effective means for engineering more powerful domain-agnostic analysis must build on a geometric (or topological) foundation of data science. The current project will introduce new paradigms, theories and tools for "Manifold Hypothesis," connect it to related efforts in topological data analysis (TDA), machine learning, deep learning neural nets, etc., and study its suitability by applying the framework to cancer biology (somatic evolution), linguistics (creolization) and cryo-EM (structural biology). It will build on new algorithmic techniques for geometric and topologic reasoning about an *ensemble of data points* as they arise in different contexts. The proposed solution will aim to seamlessly combine formalisms and techniques from differential and algebraic geometry, computational geometry, computational topology, information-theory, machine learning and statistical estimation theories.

**Broader Impact:** The research in this project is devoted to providing computer scientists with powerful new tools for designing and understanding the Big Data they create and deploy. Solving this problem alone will have immediate economic and scientific benefits. At the same time, this challenge is intended to illustrate the deeper level of scientific inquiry that the combination of techniques from mathematics, information theory, machine learning and statistical estimation theory is expected to enable. The proposal calls for the training and mentoring of undergraduate, graduate students and post doctoral researchers. Outreach includes the development of new course material in Data Science to be deployed at Cornell, MIT, NYU, and University of Washington.

**Keywords:** Manifold Hypothesis, Data Science, Cancer, Linguistics and Cryo-EM

**BIGDATA: Collaborative Research:F: Manifold Hypotheses and Foundations of Data Science with Applications**

# 1 Objectives

**Theme of the proposal: A Unified Algorithmic Framework for for Data Science** The proposal seeks to create a mathematical framework to unify diverse algorithmic techniques in data science to approach various issues raised by Big Data Problems coming from many different fields of applications. Our approach is founded on "Manifold Hypothesis," which claims that

> "High dimensional data tend to lie in the vicinity of a low dimensional manifold, thus providing the basis of *manifold learning*. The goal of data science is then to develop algorithms (with accompanying complexity guarantees) for fitting a manifold to an unknown probability distribution supported in a separable Hilbert space, only using *i.i.d* samples from that distribution. More precisely, our setting is the following. Suppose that data are drawn independently at random from a probability distribution $P$ supported on the unit ball of a separable Hilbert space $H$. Let $G(d, V, \tau)$ be the set of submanifolds of the unit ball of $H$ whose volume is at most $V$ and reach (which is the supremum of all $r$ such that any point at a distance less than $r$ has a unique nearest point on the manifold) is at least $\tau$. Let $L(M, P)$ denote mean-squared distance of a random point from the probability distribution $P$ to $M$.
>
> We wish to obtain algorithms that test the manifold hypothesis in the following sense. Any such algorithm takes i.i.d random samples from $P$ as input, and determines which of the following two is true (at least one must be):
>
> 1. There exists $M \in G(d, CV, C\tau)$ such that $L(M, P) \leq C\epsilon$.
> 2. There exists no $M \in G(d, V/C, C\tau)$ such that $L(M, P) \leq \epsilon/C$.
>
> The answer is correct with probability at least $1 - \delta$."

Examples of low-dimensional manifolds embedded in high-dimensional spaces include: image vectors representing 3D objects under different illumination conditions, camera views, phonemes in speech signals, mutational data from tumors from multiple patients, measurements from Cryo-EM or vectorized representations of words in language/dialect belonging to a speech community. The low-dimensional structure typically arises due to constraints arising from certain dynamics: determined by physical laws or the evolutionary processes. A recent empirical study [1] of a large number of $3 \times 3$ images represented as points in $\mathbb{R}^9$ revealed that they approximately lie on a two-dimensional manifold known as the Klein bottle.

One of the characteristics of high-dimensional data of the type studied by data scientists is that the number of dimensions is comparable, or larger than, the number of samples. This has the consequence that the sample complexity of function approximation can grow exponentially. On the positive side, the data exhibits the phenomenon of "concentration of measure" [2, 3] and asymptotic analysis of statistical techniques are possible.

Standard dimensional reduction techniques such as Principal Component Analysis and Factor Analysis, work well when the data lies near a linear subspace of high-dimensional space. They do not work well when the data obey more complex dynamics, as is the case in many applications that are beginning to be tackled. Among many other approaches proposed, two relatively successful competing approaches that are commonly used to handle complex data science problems are: (i) *Topological Data Analysis* (based on Persistent Homology) and (ii) *Deep Neural Networks* (based on Multiple Levels of Abstraction). How are these techniques related to each other? How can successful

theories, theorems and techniques developed in one framework be translated to the others? How can they be unified into one generalized framework? How can one select the most successful framework to a particular instance of data science for a specific application? What are the scopes and limits of each of these frameworks?

We also build a bridge to the future by investigating each of these framework's algorithmic complexity and feasibility when faced with data sets obeying certain assumptions. These considerations bring into focus various mathematical and theoretical computer science techniques that underpin these frameworks.

The work proposed here builds on our collective experience in the areas of differential/algebraic geometric, topological, algorithmic and information theoretic expertise. The team consists of applied and theoretical computer scientists as well as mathematicians, many of whom are also involved in domain specific data sciences and possess considerable experience in applied physical sciences, systems, computational biology and linguistics.

**Practical Applications: Cancer, Linguistics and Cryo-EM data** The research proposed here overlaps with the several research themes previously and currently being developed in various applied data science areas such as computational biology, astronomy, social media (e.g., natural language processing), image processing and physical sciences. There are voluminous data sets in problems addressing cancer genomics (e.g., TCGA, the Cancer Genome Atlas), linguistics (e.g., text corpora) and microscopy (e.g., cryo-EM). We believe that by studying these data through the lenses of manifold hypothesis, we will have new insights on the underlying dynamics: somatic evolution in cancer (e.g., evolution-by-duplication (EBD) and various selective pressures related to cancer hallmarks), creolization (e.g., parametrization of a universal language) or processes affecting viral populations. The research here, though not immediately, connects these issues to other areas in computer science: algorithmic complexity, analysis of temporal data, and design of supervisory controllers (e.g., therapy design for cancer).

The PI is a computer scientist with a theoretical training (formal methods, algorithms and complexity), but with an extensive practical experience in many applied computational areas: business, finance, control theory, robotics, genomics, and systems biology. Thus, the PI has a broad view of the ways various multidisciplinary issues are interrelated, possess capabilities to develop the theory to abstract the details and enjoys direct access to scientists, technologists and business leaders from constituent fields, who would be needed to build the most suitable realizetions.

**Impact of the proposal: A Unified Framework for Data Science** *The mission of the proposed Center for Algorithms in Data Sciences (CADS) is to lead the development of a cross-cutting and transformative algebraic framework for data science capable of dealing with far more complex data in a domain-agnostic manner that will equip mathematicians, information scientists and applied scientists with the tools and conceptual frameworks to better understand the mechanisms driving the data-generating processes.* We are motivated by the successful usage of data sciences in social media applications (e.g., ad targeting and optimizing return on investment (ROI)), computational systems biology (e.g., discovery of cancer bio-markers), astronomy (e.g., detection of exoplanets from Kepler data), and many more. While computer science, both in academia and industry, has attempted various engineering approaches to address each of these problems separately, there is lacking a more unified theory to understand and alleviate the challenges, which if left unsolved, could have a crippling effect on scientific progress.

The results of this research are expected to have applications far beyond Computer Science. We believe that the research produced by this proposal will lead to a significantly improved analysis of

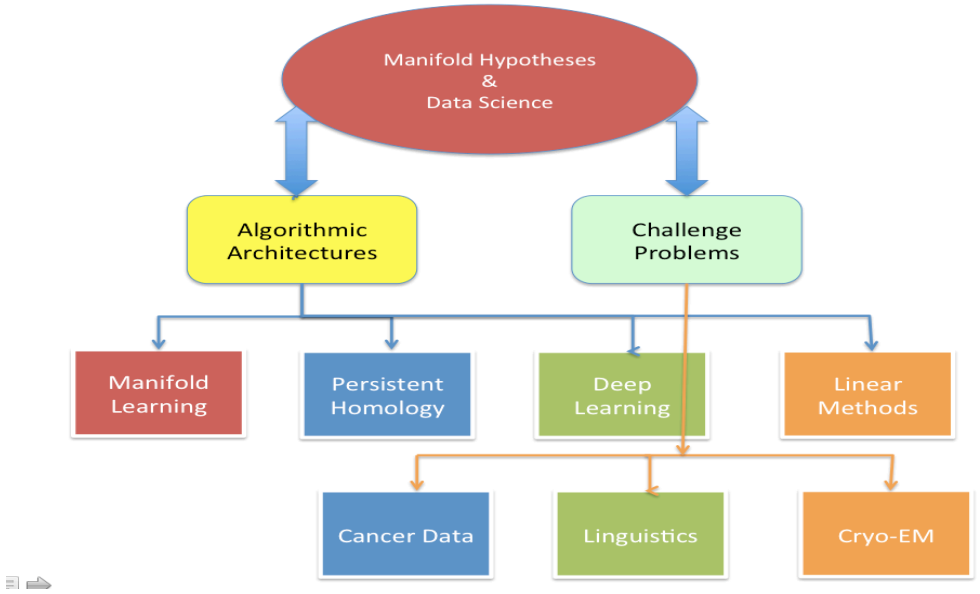


Figure 1: The Center for Algorithms in Data Sciences (CADS): An Overview. While the proposal ultimately aims to build to a bridge to the future (a unified algorithmic framework for data science), the current proposal only focuses on unifying currently successful theories and models to shape the framework. The architectural research and its implementation is outside the scope of the current proposal.

Big Data produced by government, and scientific institutions, thus paving the way for new systems focusing on healthcare systems, internet applications, law enforcement, financial systems career and skill markets, education and justice systems. Motivated by our own backgrounds and immediate needs, we plan to focus our impacts on three specific areas: Cancer Therapy, Natural Language Processing, and Structural Biology. However, the overarching goal of the Center is to develop the mathematical and algorithmic machinery necessary to make this vision a reality.

As stated before, and to bring focus to the research, educational and outreach efforts of CADS, and to serve as instruments that both drive and evaluate the theoretical (algorithmic and mathematical) research, we have identified the following challenge problems in the three specific applied areas. These challenge problems, and the underlying integrative research theme, are intended to foster a research climate that nurtures scientific excitement and creativity, informed risk taking, and a true team-building, collaborative effort: in other words, a far-reaching effort whose goal would be to train the next generation of computer science researchers. The focus on challenge problems also introduces flexibility into the research efforts of the Center: as progress is made on a particular challenge problem, more resources will be devoted to it as other researchers in team are attracted.

The PI's and their collaborators are uniquely positioned to address the research and educational challenges inherent in a Center-level activity. The PI Mishra has devoted a significant portion of his career in trying to understand and improve complex systems: Systems and Evolutionary Biology, Modeling Tumorigenesis Processes, Financial Markets, Social Networks and Cyber Security, where his research has built upon a synthesis of mathematical, algorithmic and game theoretic techniques. The collaborator Mitter is a highly regarded world-class leader in information theory; Berwick is an expert linguist working closely with Chomsky; Doerschuk is a well respected computational biomedical scientist; and Narayanan is a rising star focusing on mathematical aspects of data science.

**Management** The management structure of the proposed Center will involve the PI Mishra and all the collaborators Berwick, Doerschuk, Mitter, and Narayanan. In particular, Narayanan and Mitter will collaborate with the PI on the questions related to Manifold Learning, whereas, similarly, Berwick and Doerschuk will collaborate on Linguistics and Cryo-EM, respectively. In addition, since Berwick has significant background on Linguistics and Doerschuk on Cryo-EM, they will also pay attention to the proposal's long-term practical feasibility. Professors Mitter and Narayanan will oversee the research on mathematical and information theoretic aspects of this research.

The center will primarily train graduate students in the basic data science and analysis areas, but also through exposure to local industry and applications, make them skilled in effectively developing and transferring the technology to the real-world applications.

# 2 Research Directions

**Motivation, described informally:** We are intrigued by a plethora of competing data science frameworks, whose interrelations, we poorly understand, at present. We wish to explore how these methods can be organized and interconnected on the foundations of the "Manifold Hypothesis," as described earlier. We also wish to examine the underlying generative processes, which will contribute to creation of better phenomenological and/or mechanistic models, thus providing new scientific insights.

We consider a number of methods for each of these frameworks. These methods have a topological or geometric flavor. By topological methods, we generally mean methods whose qualitative behavior is invariant under homeomorphisms of the data. By geometric methods we mean methods that involve distances in an essential way.

## 2.1 Methods

The first method, *persistent homology* can be considered to be a general topological method, while the others should be considered geometric.

1. *Persistent Homology*: The idea here is to look for features that are invariant under a range of scales. A reasonable assumption would be that such features would be robust with respect to noise. Taking as input a finite metric space with $n$ points, for each element $d_i$ in a set of $\binom{n}{2}$ distances, we connect all pairs within $d_i$ of each other, and consider the homology groups of the resulting clique complexes. The inclusion maps of the simplicial complexes translate into homomorphisms of the homology groups. The image of such a homomorphism is a persistent homology group. We would like to learn the persistent homology groups of data. Please see [4–8].

2. *Manifold Learning*: This is a subfield of machine learning based on the hypothesis that data lies in the vicinity of a low dimensional manifold. We would like to learn the underlying manifold from data, as described in more detail earlier. Please see [9–12].

3. *Spectral Analysis*: PCA and ICA are two exemplar methods in this category. PCA can be viewed as a special case of manifold learning wherein a linear subspace is learnt that is the best fit in terms of mean squared loss with respect to the data. Please see [13–17].

4. *SVMs and Large Margin Classifiers*: These are ways of classifying data by a separating hyperplane or regressing data by a linear function respectively. In the case of classification, the margin between the positive and negative samples is often maximized – thus giving rise to a

large margin classifier. In the case of regression, Drucker et al. [18] describe a scheme (SVR) wherein one learns a linear hyperplane with $\ell_\infty$ loss $\epsilon$ and minimizes the magnitude of the gradient of the linear function among all candidates. Please see [19–22].

5. *Kernel Methods*: Given two points $x$ and $y$, we denote a kernel function by $k(x, y)$. The hyperplane in a kernel SVM is defined using this kernel function $k(x, y)$, and thus is, generally speaking, not a linear hyperplane in $\mathbb{R}^n$ with respect to the canonical inner product. This approach gives additional flexibility and power, but involves only a subclass of general manifolds. Please see [23].

6. *Deep learning*: In deep learning one starts with defining a function that extracts features called the encoder $f_\theta$. For each $x$ from a data set $X$, let $c = f_\theta(x)$ be the code constructed from $x$. A decoder $g_\theta$ maps a representation $c$ onto an element $r$ of the input space. The encoder and decoder are trained to minimize the reconstruction error $dist(x, r)$ over training samples. In order to constrain $g_\theta f_\theta$ away from the identity, the range of $f_\theta$ is forced to be low dimensional. In the framework of denoising auto-encoders, the objective of the decoder, when composed with the encoder, is to map a corrupted input back to the original.

   Suppose that data is drawn from a distribution supported on a low dimensional manifold $\mathcal{M}$ embedded in high dimensional Euclidean space $\mathbb{R}^n$. Suppose that i.i.d $N(0, \sigma^2 I)$ gaussian noise is added to the data. We would like to design a denoising auto-encoder that takes $\tilde{x} = x + \zeta$, $x \in \mathcal{M}$ being a data point and $\zeta \in \mathbb{R}^n$ being a noise vector, and maps it onto a point $r$, where $\mathbb{E}[(x - r)^2]$ is small. One way to accomplish this goal is by considering $r(\tilde{x}) - \tilde{x}$ as half the gradient of the log-density of the noisy data. More sophisticated variants may be explored, such as a gradient path following approach, rather than taking the gradient in one shot.

   A natural question for further investigation is *how manifold learning can aid deep learning*. Please see [24].

### 2.2 Research Questions

The questions below should be answered using the analysis of real-world data.

**For what problems are the above methods suitable?** The suitability of manifold learning for a particular set of data is related to the generating process underlying the data. One reason for data to lie near a manifold might be the presence of a low dimensional Lie group acting on the data. Another might be that it lies on the invariant manifold of a dynamical system, for example submanifolds in oceanographic data. Stochastic stable manifolds also exist for noisy dynamical systems. When data consists of digits or images of faces, there are a few basic transformations which can be used to go from one data point to another. This structure results in data lying close to a manifold.

Manifold learning generally assumes some form of smoothness everywhere. If data does not possess smoothness everywhere, it might be modeled not as lying on a single manifold, but as lying on the union of several intersecting manifolds.

Persistent Homology works with a data set as a whole rather than the points themselves. It is an unsupervised method and hence is seemingly unsuited to classification tasks. However, in applications it is possible for the points themselves to have additional structure by virtue of their position. For example, points might correspond to medical images. In fact [8] have used it to classify medical images by performing persistent homology calculations on the images themselves.

**How effective are the above methods in practice, e.g. in the context of applied problems discussed later?** While we are excited by the possibility of new theoretical results connecting various algebraic and algorithmic ideas currently flourishing within the machine learning community, it still remains unclear how our research would influence practitioners, who also need to derive mechanistic insights. For example, a deep neural net model of cancer progression only provides a machinery that is phenomenologically accurate (and likely to be successful in creating advanced diagnostic and prognostic tools), and yet it does not connect to mechanistic processes, necessary for finding drug-targets, rational drug-design and therapy design. Thus, we wish to pay equal attention to domain-specific applications, where Big Data and their analysis must be tied to domain-expertise in understanding how best tangible utilities can be derived from Data Science.

More specifically, we focus on questions of the following kind.

1. How can we find the dynamics governing somatic evolution and driving cancer progression by using a variant of least squares fitting, built on the foundation of "Manifold Hypothesis?" How much improvement does it produce, when compared to models derived from SVM or graphical model-based analysis?

2. How best can we map words spatially onto a manifold, and then use this embedding to predict the variable scope's influence on words in a sequence? Could we have arrived at similar insights by just using persistent homology here?

3. When we fit a manifold to the different views of a 3D molecule obtained through Cryo-EM, what insight can we get from such a manifold? Can one identify the manifold up to homeomorphism using Persistent homology?

We provide illustrative examples of the far-reaching research we plan to pursue (Section 2.3). Research in these areas will play a cross-cutting and fundamental role in the research pertaining to the Challenge Problems (Section 2.4).

### 2.3 Core Open Problems

**Algorithmic Questions for Manifold Learning.** One of the primary obstacles to learning a manifold from data is the absence of an obvious way of describing a manifold having high (even infinite) co-dimension. This was overcome in (Fefferman et al., "Testing the Manifold Hypothesis", 2013) [11] by approximating a manifold with bounded reach, dimension and volume by another manifold that is contained in a finite dimensional affine subspace. (The reach of $\mathcal{M}$ is the supremum over all $r$ such that any point at a distance $r$ from $\mathcal{M}$ has a unique nearest point on $\mathcal{M}$. ) This latter manifold was then expressed as a section of a vector bundle over another putative manifold. This putative manifold was described as the set of zeroes of a section of a vector bundle whose base space was a neighbourhood of the data.

The techniques introduced in the above paper make the questions below amenable to analysis.

- Manifold fitting: Assume that data is drawn i.i.d from a probability measure supported near a manifold. Fit a manifold of bounded reach to this data. Obtain bounds on the computational complexity of this task.

- Mapping: Given prescribed data points, and an abstract manifold in terms of charts, find a map from the abstract manifold to a Hilbert space, such that the reach of the image of the map is bounded below and sum of the squares of the distances of the data points to the image is minimized under the reach constraint.

Example: In the case of Cryo-EM there are 2D projection images of a 3D object taken in various projection directions. These projection directions can be associated with the abstract Lie group $SO_3$. Therefore the views may be labeled by points in $SO_3$. The images taken from these views may be represented as points in the Hilbert space $L^2([0,1] \times [0,1])$. Thus we would like to to fit a copy of $SO_3$ to data in a Hilbert space.

- Going from one representation of a manifold to another:

  One representation of an embedded manifold could be as the set of zeroes of a section of a vector bundle whose base space is a neighborhood of the data. We might like to construct a "distance oracle" for a neighborhood of the manifold, which when presented with a point, outputs the distance of the point to the manifold. This would be a second representation. The gradient of the squared distance can be used to find the nearest point on the manifold to the presented point.

- Sampling: How to sample from an interpolated manifold. We would like to perform this step by constructing a Markov chain using charts that mixes within a prescribed time. This chain could be used to generate additional random samples synthetically. Rejection sampling may be used for manifolds with boundary.

- Using samples to infer characteristics of manifolds: Random samples can be used to construct a fine net, which can be in turn used for a variety of purposes such as homology computations and volume computation.

- Integration of forms over a manifold: The volume and the Euler characteristic of a Riemannian manifold without boundary can be respectively expressed as integrals of the volume form and the Gaussian curvature respectively. The tangential Delaunay complex of Boissonnat et al. [25] could be useful for this purpose.

- Regression on data from a manifold: Consider a $C^m$ submanifold of a Hilbert space with boundary whose reach is greater than 0. Assume that the boundary is a $C^m$ submanifold whose boundary is greater than 0 as well. Suppose $f : \mathcal{M} \to \mathbb{R}$ is an unknown $C^k$ function corrupted with additive gaussian noise, what are the optimal rates of estimating $\mathcal{M}$? We would like to draw a parallel with the literature on compressed sensing, in particular, the Dantzig selector of Candes and Tao [26] . Results of Fefferman [27, 28] allow us to write the minimum $C^k-$ norm of any function satisfying equality constraints corresponding to some data points as the minimum of the objective in a finite dimensional convex program.

- All of the above in the presence of noise.

  An alternative way of stating the problem of estimating a manifold from noisy samples involves parametrizing the manifolds using reach $\tau$ and covering number at a scale of $\tau$. We can infer the dimension from the projection map.

## 2.4 Research in the Challenge Problem Areas

- **Cancer Data:**

  In the near future, cancer research is likely to become much more data-centric, primarily because of the rapid growth and ready availability of vast amount of cancer patient genomic data, as well as because of advances in single-molecule single-cell technologies. Nonetheless, it remains impossible to track the tumor progression in any single patient over time, though

emerging technology for noninvasive analysis of circulating tumor cells and cell free DNA (in blood and urine) is beginning to paint an incomplete, but useful, picture. Motivated by these possibilities, we seek to use the currently available aggregated data from multiple patients to infer an approximate phenomenological "shape" of cancer progression, which will ultimately build on the similarities among data-points at different scales and encoding them as "barcodes" (e.g., in terms of persistent homologies). Less intensely studied but of equal importance would be an analysis of these data in light of the "Manifold Hypothesis," which could shed important light on the underlying dynamics governing the somatic evolution. In particular, we wish to infer causal relations among various mutational events occurring in the course of cancer progression, organizing them in terms of "variational" and "selectivity" relations, and linking them to our understanding of various intra- and inter-cellular pathways.

Thus, we seek to understand initiation and progression of cancer in terms of "chronological" and "causal/selectivity" relations among somatic alterations as they occur in the genomes and manifest as point mutations, structural alterations, DNA methylation and histone modification changes. For example, if through some initial mutations (e.g. in EGFR) a cell acquires the ability to ignore anti-growth signals, this cell-type may enjoy a clonal expansion (modeled as a discrete state of the cancers progression and marked by the acquisition of a set of genetic events). However, such a state of affairs may result in a Malthusian pressure on the population of all the cell-types in terms of deregulation of glutamine metabolism and thus, set the stage for clonal expansion of a new cell-type that can disable G1-S checkpoint (e.g., a "selected" mutation in CDK). Such causal structures is likely to be implicit in the genomic data from multiple patients, some involving tumor populations with just EGFR-cell-types and some others with a heterogeneous population with EGFR+CDK-cell-types, etc.

Such a structure can be summarized in terms of a directed acyclic graph, $G = (V, E)$, where the vertices $V$ encode the mutational events and the edges $E$ describe the "selectivity" relations among the effected vertex and its selective parent vertices (mutations). When a vertex is connected to multiple parents, the selectivity structure may need to be described by a logical relation: e.g., singular (only one parent), conjunctive (all parent events are necessary), disjunctive (any parent event is sufficient), or even more complex relations (but limited to propositional or modal logic expressions). Such a graph, of course, ignores the exact geometry of the time and only expresses the "temporal priorities" in a topological sense. A selectivity graph (SBCN: Suppes Bayes Causal Network), as described here, can construct a temporal possible-world model, which is amenable to temporal logic analysis (via model checking), thus allowing the data-scientists to propose more complex hypotheses describing various evolutionary forces in cancer progression. See [29, 30].

Nonetheless, rigorous algorithmic tools to infer such selectivity and temporal relations from the topology of the data, which is induced by these genetic events and drives cancer progression, have remained largely elusive. The main reason for this state of affairs is that information directly revealed in the data lacks direct temporal measurements but also contains large amount of irrelevant structures, complicated by heterogeneity in cell-types and non-selective "bystander/passenger" mutations. See [31], and reference therein.

As a proof-of-concept, we wish to reanalyze the patient data for glioblastoma (GBM). In particular, we will create tree and DAG cancer progression models for glioblastoma and relate our inferred prima facie causes to the shapes inferred by persistent homologies and also to the moduli space behavior of the GBM evolution. In addition, we will develop new model checking algorithms to incorporate the topological properties available from the algebraic analysis (e.g.,

moduli space behavior and persistent homologies) to improve our classification algorithms (e.g., the module to separate genuine causes from spurious causes).

The selectivity relations inferred from the phenomenological models are obtained from the patient data and require mechanistic explanation in terms of various biochemical pathways, if they have to be used in therapy design or drug discovery. For this purpose, we will create in silico models of a population of tumor cells, where the behavior of individual cells can be simulated in terms of the known regulatory and signaling pathways, and evolved over time to validate (or refute) our inferred selectivity relations. In order for efficient simulation, cell-autonomous processes, involved in cancer progression, will be abstracted (simplified/approximated) by using "approximate-bisimulation" relations and can be utilized efficiently in population-level simulation. At the population level, each cell could be viewed as an agent interacting strategically with the other cells in a game-theoretic setting. The therapy design algorithms we have developed earlier can then be implemented on the resulting cancer hybrid automata (CHA) model using techniques from supervisory control theory and theories of games against nature. See [32].

- **Linguistic Data:**

  With the growing ubiquity of social media, there now exist a very large number of massive text corpora in multiple languages and being shaped by diverse groups of speech communities. The basic building blocks of these linguistic Big Data are words. SVD and tensor word encodings have been found to be very useful in mapping $n$-gram type word associations in the form of latent variables. However, this linear analysis does not deal with the hierarchical, compositional structure of language. Much information in language exists in the form of operator-variable structures that resemble the application of functions to arguments in the lambda calculus, or the hierarchical environment frames of a programming language like Scheme. See [33–35].

  Thus an important subject of study could be based on the notion of variable scopes, akin to the scope-rules of a programming language. Just as in programming languages, proper scope cannot always be ignored. For example, even if we have the precisely corrected bigram probabilities for the word sequence *what who bought*, this does not suffice to fix its meaning, because when the "operator" *John knows* is hierarchically composed with this sequence, it yields only an ill-formed structure, *John knows what who bought*. Replacing *John* with the typed operator *who* rescues the meaning structure, correctly so in this case even though the bigram probabilities have not been altered from the ill-formed example because the linear sequence *what who bought* is fixed. Many other examples in human languages follow this pattern of drawing on hierarchical, rather than linear structure, to fix meaning. Also, see [34, 36].

  While deep learning methods have been applied to certain of these problems in language processing, they have not had close to the same success as in visual object recognition. In part this anomaly is due to the fact that such approaches have not exploited the notion of scope within the known manifold structure of natural languages. Based on the foundations suggested by "Manifold Hypothesis," we propose to study this problem more rigorously in this project: exploit the known, empirically verified high-dimensional structure of composed phrases. In addition to suggesting better NLP (Natural Language Processing) algorithm, this approach will also clarify the connections between manifold learning and deep learning.

  Revisiting our example, in the case of simple "noun phrases" such as *the multidimensional aspect of learning*, it is possible to specify a 56-dimensional space that spans all possible variation in the several hundred contemporary Indo-European languages, including English, Spanish, Russian,French, Hindi, Farsi, etc. We aim to investigate whether we can use this manifold structure to quickly move from analysis in one language to another, approximating the very

low sample complexity that native language learners exhibit. Existing methods that use MCMC sampling or EM within a Bayesian framework build on formulations that are provably NP-hard with respect to sample's embedding dimension, and remain intractable, even when approximate solutions are sought. Recent proposals suggesting that small scale sampling will suffice to resolve this difficulty all rely on onerous i.i.d. assumptions of perfect knowledge of posterior distributions that cannot be met in practice. Additional references can be found in [36].

Here we take a different approach. We will apply the manifold learning method to resolve this complexity problem. We will also exploit a second recently discovered constraint regarding natural languages. It turns out that the "operator" that assembles hierarchical structure and so the scope/environment frames in natural language, can be represented simply as set union. This has implications for recovering hierarchical structure from linear word strings, because this new reformulation not only covers typical compositions where two hierarchical structures are assembled into one, e.g., *John* and *saw the radio* into a larger hierarchical object sentence, *John saw the radio*, but also the ubiquitous and more difficult appearance of displaced syntactic units, e.g., *The radio, John saw*. Specifically, we will examine how the usual matrix-like composition operation that comprises the basis of virtually all current statistical parsing methods trained on corpora (e.g., the CKY algorithm) could be modified to use this more empirically accurate operator. See also [35, 37].

- **Cryo-EM Data:**

  An important approach to studying biological nanomachines is structural biology, which focuses on the geometric shape of the object at resolutions as small as atomic resolution and on the relationship between biological function and geometry. A technique of increasing importance is single-particle cryo electron microscopy (cryo EM). In cryo EM, a aqueous film containing thousands of unoriented objects is flash frozen to cryogenic temperatures and imaged. The image is basically a 2-D projection of the 3-D electron scattering intensity distribution of the specimen. Primarily because of damage by the electron beam, two choices are made in high spatial-resolution studies. First, the electron microscope beam current is minimized leading to highly-noisy (SNR $< 0.1$) images. Second, only one projection image is recorded and, due to the low SNR and the unoriented nature of the objects in the film, the projection direction for any particular instance of the object is not known and cannot be determined from the image. So, instead of reconstructing based on a full set of oriented projection images of a single object, as is done in x-ray computed tomography in medical imaging, many images each of different instances of the object and with different and unknown projection directions must be computationally combined to compute the reconstruction [38].

  There are three problems of increasing difficulty that we propose to address with the manifold learning ideas at the core of this proposal. Suppose that the electron scattering intensity distribution ($\rho(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^3$) of an individual instance of the object is represented by a weighted ($c_\alpha$) sum of basis functions ($\psi_\alpha(\mathbf{x})$):

$$\rho(\mathbf{x}) = \sum_\alpha c_\alpha \psi_\alpha(\mathbf{x}). \tag{1}$$

  The goal of 3-D reconstruction is then characterization of the weights $c_\alpha$. For $\boldsymbol{\chi} \in \mathbb{R}^2$, let the $i$th image be denoted by $\sigma_i(\boldsymbol{\chi})$. Because of the projection-slice theorem, it is natural to work with Fourier transforms: $\sigma_i(\boldsymbol{\chi}) \leftrightarrow \Sigma_i(\boldsymbol{\kappa})$, $\rho(\mathbf{x}) \leftrightarrow P(\mathbf{k})$, and $\psi_\alpha(\mathbf{x}) \leftrightarrow \Psi_\alpha(\mathbf{k})$. Using standard

first-order image formation theory [39–41], $\Sigma_i(\boldsymbol{\kappa})$ can be expressed in the form

$$\Sigma_i(\boldsymbol{\kappa}) = \exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(|\boldsymbol{\kappa}|)P\left(R^{-1}_{\alpha_i,\beta_i,\gamma_i}\begin{bmatrix}\boldsymbol{\kappa}\\0\end{bmatrix}\right) \tag{2}$$

where the electron microscope's optics are described by the contrast transfer function $G$, the fact that the image is not centered is described by the complex exponential of the offset $\boldsymbol{\chi}_{0,i} \in \mathbb{R}^2$, and the projection-slice theory is used in the form of a $z$-directed projection of the object after a rotation by $R \in \mathbb{R}^{3\times3}$ ($R^T = R^{-1}$, $\det R = +1$) parameterized by Euler angles $(\alpha_i, \beta_i, \gamma_i)$ though other parameterizations (e.g., quaternions) are equally useful.

For noise-free pixelized images arrayed as a vector $\mathbf{s}_i$ with the weights $c_\alpha$ also arrayed as a vector $\mathbf{c}$, Eqs. 1 and 2 imply

$$\mathbf{s}_i = L(\alpha_i, \beta_i, \gamma_i, \boldsymbol{\chi}_{0,i})\mathbf{c} \tag{3}$$

where the elements of $L$ are $\exp(-i2\pi\boldsymbol{\kappa}^T\boldsymbol{\chi}_{0,i})G(|\boldsymbol{\kappa}|)\Psi_\alpha(R^{-1}_{\alpha_i,\beta_i,\gamma_i}(\boldsymbol{\kappa}^T,0)^T)$ where (discretized) $\boldsymbol{\kappa}$ indexes rows and $\alpha$ indexes columns. A typical size of problem is $10^3$–$10^5$ images, $200\times200$ pixels per image (so $\mathbf{s}_i \in \mathbb{R}^{40{,}000}$), and $\mathbf{c} \in \mathbb{R}^{N_c}$ with $N_c$ as small as $10^3$ [42] or as large as $100\times100\times100 = 10^6$ (a voxel representation of a object of size 200Å at 4Å spatial resolution using voxels of size $2\times2\times2$Å$^3$).

Eq. 3 displays the manifold explicitly: each image $\mathbf{s}_i$ is related to a constant unknown vector $\mathbf{c}$ by a matrix $L$ whose structure is known but whose parameters are unknown. In the simplest case of no offsets ($\boldsymbol{\chi}_{0,i} = \mathbf{0}$), the explicit parameterization is just $SO_3$. If the manifold can be learned from the noisy data and a map constructed to $SO_3$, then the reconstruction problem can be solved by many methods since the projection orientation of each image is known.

A more challenging problem adds discrete classes representing the fact that biological-chemical-physical methods are sometimes unable to distinguish between classes of objects and so the images are an unlabeled mixture of images showing instances of all classes. Reasons for such heterogeneity include relatively discrete steps in the maturation pathway of virus particles such as the bacteriophage Hong Kong 97. Let $\eta_i$ be the class of the $i$th instance. Then Eq. 3 is replaced by

$$\mathbf{s}_i = L(\alpha_i, \beta_i, \gamma_i, \boldsymbol{\chi}_{0,i})\mathbf{c}^{(\eta_i)}. \tag{4}$$

In this case, learning the manifold described by the $\mathbf{s}_i$ vectors mixes classes and it is necessary to describe that manifold as the union of a set of manifolds, one manifold for each class. The basic approach is that the underlying manifolds should be smoother, e.g., in the approach of Fefferman-Mitter-Narayanan [11], the confidence is greater ($\delta$ is smaller), the reach $\tau$ is greater, or the error $e$ is smaller.

A yet more challenging problem is continuous heterogeneity within each class. Reasons for such heterogeneity include the fact that such huge multicomponent objects such as viruses, ribosomes, or nuclear pore complexes are flexible. Then Eq. 4 is replaced by

$$\mathbf{s}_i = L(\alpha_i, \beta_i, \gamma_i, \boldsymbol{\chi}_{0,i})\mathbf{c}^{(i)} \tag{5}$$

where the set of $\mathbf{c}$ vectors corresponding to a single class $\eta_0$, i.e., $C_{\eta_0} = \{\mathbf{c}^{(i)} | \eta(i) = \eta_0\}$ is somehow "clustered" around a nominal vector $\bar{\mathbf{c}}^{(\eta_0)}$. In this case the tradeoff between the accuracy with which the manifold fits the experimental data and the rapidity of fluctuation in the manifold in comparison with the fluctuation expected based on the resolution of the experimental data may provide insight into the size of the set containing $C_{\eta_0}$, e.g., the sample covariance of the vectors in $C_{\eta_0}$.

In Refs. [43–45] the PIs describe a model-based statistical estimation approach to these three problems that is based on Gaussian measurement noise and Gaussian mixtures to describe the continuous heterogeneity present in multiple classes. The approach has been used, e.g., Ref. [42, 46, 47]. The estimators are computed via an expectation maximization algorithm where the nuisance parameters are the parameters of $L$ and the expectations are computed numerically which is a large computational burden. The Gaussian assumptions, especially in the third problem, are crucial. However, they are poorly justified. For example, some recent "direct electron detectors" [48] act as digital counters of incident electrons in each pixel so that at least the part of the measurement noise due to low beam current is probably better described by a counting process. What is probably the most popular software system, Relion [49, 50], uses the estimation formulation described in Ref. [45] and therefore has these challenges. The potential of the methods described in this proposal is two fold: First, at a practical level, they provide an entirely different tradeoff between computational complexity and performance. Second, at the level of principle, they do not require the Gaussian assumptions and therefore offer the hope of more robust performance in the presence of realistic large noise signals.

# 3 Summary of Research Component

The research component of this project pursues an ambitious but manageable agenda that integrates algorithmic and experimental methods to create a comprehensive theory to advance data sciences; it is based on a framework building upon Manifold Hypothesis. Our approach, focusing on *geometric (differential/algebraic) and topological techniques*, will lay the foundations for complex, and yet, readily applicable engineering systems of the future.

### Deliverables and Dissemination Plans

The following lists our deliverables and estimated man-years (MY) of effort, to be divided amongst the investigators, collaborators and their research teams.

- A Theory for Manifold Learning : 1 MY
- Connections to Topological Data Analysis: 1 MY
- Connections to Deep Learning: 1 MY
- Devising Generalized Learning Algorithms: 1 MY
- Feasibility Analysis (in the oncogenomic setting): 1 MY
- Feasibility Analysis (in the linguistic setting): 1 MY
- Feasibility Analysis (in the cryo-EM setting): 1 MY

# 4 Broader Impacts of the Proposed Work:

### Intellectual Contributions:

The proposed research will make contributions to Computer Science and Mathematics, leading further to a symbiotic application to data science, in the context of cancer, linguistics and physical sciences. The foundational basis provided by Manifold Hypothesis will introduce many new algorithmic questions for solving Big Data Analysis problems by combining, for the first time, our understanding of complex topological and geometric constraints imposed by manifold hypothesis. The proposed approach to tame algorithmic complexity and data overfitting will be the first to combine *differential algebraic features such as curvature, reach, and volume*, *topological features such as Betti numbers* and *deep learning features such depth and hierarchy.* Finally, we will demonstrate

these methods to design new domain-specific informatics solutions, for instance, to develop improved cancer therapy, linguistic creolization and cryo-EM analysis, aimed at public health, social media and engineering.

**Broader Impacts:**
The proposed research will create a new algorithmic sub-discipline within the fields of data science and pave the way for new tools for practitioners. Potential applications of our research include the design of systems applied to healthcare, social networks, and physical sciences. Our governing philosophy — that the design of more sophisticated and data science algorithms – can be formulated and solved in the context of Manifold Hypothesis, with applications that exemplify the use of Computational Thinking. Finally, our research will lay the groundwork for testing novel hypotheses about learning and evolution and for exploring the Shared Principles Between the Computing, Social and Biological Sciences [51].

# 5 Education and Outreach

Our goals are to integrate education and entrepreneurship activities for this project with our research activities and to blur the line among them. Toward this end, our proposed plan includes not only traditional activities like new courses and outreach to K-12 programs but also coordinated efforts to *merge graduate and undergraduate education and research opportunities* through workshops; dedicated *education research* to develop new approaches to teaching complex skills like domain-specific modeling; and outreach to underrepresented groups, the general public, and industry through high-impact mechanisms. Our plan also will build on and extend the highly successful education and outreach program from the CMACS Expeditions in Computing (EiC) award with which the PI of the current proposal was very closely involved.

Berwick has developed software for both secondary school and university use, combined with distance learning which touches on the themes of the proposal. In the main evolutionary biology course for MIT, Berwick's software captures visually the complex dynamical system effects in populations undergoing stochastic selection, migration, and drift. This software is in the process of being made available as part of the widely-used text by Prof. Matthew Hamilton at Georgetown, as a website application. A suitably modified version of this software is being tested as part of the AP Biology program at Boston and Cambridge area public schools. This follows on from Berwick's STEM expertise in developing a physics distance learning program for secondary school women, based on the simulated construction of Ferris Wheels.

A highlighted feature of our educational initiatives is multiple *meetings, meet ups and seminars* at different levels. As part of these activities, a large number of members of the Silicon Alley Community met regularly in a Tuesday Spamhaus meetings where they had opportunity to discuss data science applications to Ad-exchanges, attribution analysis, illiquid markets, market defaults (and prepayments), securitization, natural language processing, data in clinical genomics, onco-genomics, data markets, clinical trials, job and skill markets, causality theory, graph theory (random graphs, diffusion on graphs, agony-based distances), game theory, but much more specifically, data science, especially the ones with need for advanced machine learning. At NYU, the PI taught a graduate-level course on Social Networks describing the applications of data science to social media data and collaborators. The PI Mishra also taught a 10-days course at SEI/CMU on data science applications to cyber security.

Our educational activities also will have strong ties to *new degree programs*. The group can provide support and infrastructure for new Masters programs in Data Science (Courant/NYU) and

Boot Camps for young technologists in New York Area.

## 6 Description of Team Members

**Bob Berwick:** with expertise in Computational linguistics, computational complexity theory, and parameterized learning will focus on the integration of current linguistic theory into manifold learning and the topological structure of high dimensional linguistic descriptions. **Peter Doerschuk:** with expertise in computational biophysics, biomedical research and biophysical data science will focus on Cryo-EM data analysis using manifold learning. **Bud Mishra:** with expertise in statistics, data science, systems biology and cancer biology will focus on Manifold Hypotheses, algorithms for topological data analysis and manifold learning and applications to cancer biology. **Sanjoy Mitter:** with expertise in control and information theory will focus on Manifold Learning, bio-physical data science and linguistic applications. **Hari Narayanan:** with expertise in Manifold Learning, Convex Optimization, MCMC and Complexity will focus on Manifold Learning, Deep learning and Cryo-EM applications.

## 7 Results from prior NSF support

PI Mishra has a long and successful history of NSF funding. He has been a PI on following recent awards (NSF CCF-0836649, 09/15/08-08/31/12, 'Collaborative Research: CDI-Type II: Discovery of Succinct Dynamical Relationships in Large-Scale Biological Data Sets'; NSF CCF-0926166, 09/01/09-08/31/14, 'Collaborative Research: Next Generation Model Checking and Abstract Interpretation with a Focus on Embedded Control and Systems Biology'; NSF IGERT-0333389,10/01/03-09/30/12 'IGERT: Program in Computational Biology (COB)'). These awards have resulted in more than 15 publications [52–68], 12 Ph.D. dissertations, and several new courses (Computational Systems Biology (GWAS), BioInformatics (Signals and Cancer), Computational Systems Biology (Model Checking and Systems Biology), Heuristic Problem Solving (HPS) and Social Networks (Signaling Games)).

## Project Coordination Plan

Bud Mishra (NYU) is the PI and coordinator of the research. The PI Mishra and his collaborators have been assigned primary and secondary roles according to the research and educational tasks (see table below). Additionally, we have requested funds to support graduate students who will assist in the execution of the research plan. Undergraduate researchers will also be recruited using REU and similar sources of funding. The following coordination mechanisms will be put into place to ensure the successful coordination of all project activities.

- Project personnel (PIs, co-PIs, students, etc.) will participate in bi-weekly video conferences/online meetings (with shared workspace/desktop using e.g. www.gotomeeting.com).
- We will host annual project meetings for project personnel and other interested parties, with the meeting location rotating between New York and Boston. Travel funds have been allocated in the budget to support these activities.
- We will have graduate student (possibly, postdocs, not funded by this project) exchanges between the participating institutions. Travel funds have been allocated in the budget to support these activities.
- We will use the Subversion (SVN) version control system as an inter-institutional shared repository for source code, web pages, and documentation.
- We will establish a Manifold Hypothesis Web Site where project personnel can find (and post) project-relevant documents, collaborative Wikis, timetables, blogs, links to relevant web sites, etc.

The following is our anticipated time line:

| | | |
|---|---|---|
| A Theory for Manifold Learning : | P: HN, S: BM & SM | 2016–2017 |
| Connections to Topological Data Analysis: | P: BM, S: HN | 2016–2017 |
| Connections to Deep Learning: | P: HN, S: BM | 2016–2017 |
| Devising Generalized Learning Algorithms: | P: BM, S: ALL | 2017–2018 |
| Feasibility Analysis (in the oncogenomic setting): | P: BM, S: SM | 2016–2020 |
| Feasibility Analysis (in the linguistic setting): | P: RB, S: BM | 2016–2020 |
| Feasibility Analysis (in the cryo-EM setting): | P: PD, S: SM | 2016–2020 |

## Data Management Plan

This project is foundational in nature, but may lead to the following kinds of data (as defined in OMB Circular A-110) :

- Original data (observations, measurements etc.) aggregated from multiple sources.
- Metadata, including experimental protocols and software code for implementing the algorithms.
- Curriculum materials, including lecture notes, slides, assignments, and exam questions.

The original data, experimental protocols, and samples will be documented according to the guidelines specified in "Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences (2003)" `http://www.nap.edu` to ensure reproducibility of the results. Software will be documented for use by end-users. Software tools will be freely distributed via a common web-site that will be created and maintained by the personnel on this project. Curriculum materials will be freely distributed via web-sites for the respective courses.

Our policy for access and sharing of data with other researchers will be to provide it at no more than incremental cost and within a reasonable time after publication and, when appropriate, after patent applications have been filed. All data will be retained for at least five years from the end of the project. Experimental data will be archived at NYU on local file systems. Software (and documentation) will be replicated and archived at NYU, MIT, Cornell and UW on local file systems. Obligatory semi-annual exchanges will ensure that all PIs have access to updated software. A publicly accessible website for the project will be hosted at one of the partner institutions, or on free alternatives (e.g., GoogleSites). Requests for data will be made by contacting the PI (Mishra) and/or corresponding authors on publications resulting from this research. Such requests will be reviewed by the PI as part of regular collaboration meetings and/or informal discussions via phone or email.

All participants of this project will have access to all data, and all will be responsible for ensuring that the data they generate is available, both internally and for external researchers for a period at least five years from the end of the project. Should any PI leave the project, the remaining PIs will take responsibility for archiving, managing, and disseminating the departing PI's data.

## References


[1] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, pp. 255–308, 2009.

[2] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality. aide-memoire of a lecture at," in *AMS Conference on Math Challenges of the 21st Century*, 2000.

[3] M. Ledoux, *The Concentration of Measure Phenomenon*. Math. Surveys and Monographs, AMS, 2001.

[4] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.

[5] H. Edelsbrunner and J. Harer, "Persistent homology — a survey.," *Surveys on Discrete and Computational Geometry. Contemporary Mathematics*, vol. 453, no. 2, pp. 257–281, 2008.

[6] R. J. Adler, O. Bobrowski, M. S. Borman, E. Subag, and S. Weinberger, "Persistent homology for random fields and complexes," *Institute of Mathematical Statistics Collections*, vol. 6, pp. 124–143, 2010.

[7] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete and Computational Geometry*, vol. 50, pp. 330–353, 2013.

[8] O. Dunaeva, H. Edelsbrunner, A. Lukyanov, M. Machin, and D. Malkova, "The classification of endoscopy images with persistent homology," in *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2014 16th International Symposium*, pp. 565–570, 2014.

[9] J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st ed., 2007.

[10] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[11] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *arxiv*, no. 1310.0425, 2013.

[12] H. Narayanan and S. Mitter, "Sample complexity of testing the manifold hypothesis," *NIPS*, 2010.

[13] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[14] H. H., "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[15] I. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, 2nd ed., Springer NY, 2002.

[16] P. Comon, "Independent component analysis: a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[17] J. V. Stone, *Independent component analysis : a tutorial introduction.* MIT Press,Cambridge, Mass., 2004.

[18] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999.

[19] A. Shmilovici, "Support vector machines," *Data Mining and Knowledge Discovery Handbook*, pp. 231–247, 2010.

[20] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press New York, NY, USA, 2000.

[21] C.Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] Edited by Alexander J. Smola, Peter Bartlett, Bernhard Scholkopf and Dale Schuurmans, *Advances in Large-Margin Classifiers.* MIT Press, Cambridge, Mass., September 2000.

[23] B. S. Thomas Hofmann and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[24] A. C. Y. Bengio and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. PAMI, special issue Learning Deep Architectures*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] J. Boissonnat and A. Ghosh, "Manifold reconstruction using tangential delaunay complexes," *Discrete & Computational Geometry*, vol. 51, no. 1, pp. 221–267, 2014.

[26] E. Candes and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, pp. 2313–2351, 2007.

[27] C. Fefferman, "Interpolation and extrapolation of smooth functions by linear operators," *Rev. Mat. Iberoamericana*, vol. 21, no. 1, pp. 313–348, 2005.

[28] C. Fefferman, "The $c^m$ norm of a function with prescribed jets ii," *Rev. Mat. Iberoamericana*, vol. 25, no. 1, pp. 275–421, 2009.

[29] I. Korsunsky, D. Ramazzotti, G. Caravagna, and B. Mishra, "Inference of cancer progression models with biological noise," *CoRR*, vol. abs/1408.6032, 2014.

[30] L. O. Loohuis, G. Carvagna, A. Graudenzi, D. Ramazzotti, G.-C. Mauri, M. Antoniotti, and B. Mishra, "Inferring tree causal models of cancer progression with probability raising," *PLoS One*, vol. 9, no. 10, 2014.

[31] L. O. Loohuis, A. Witzel, and B. Mishra, "Improving detection of driver genes: Power-law null model of copy number variation in cancer," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 11, no. 6, pp. 1260–1263, 2014.

[32] L. O. Loohuis, A. Witzel, and B. Mishra, "Cancer hybrid automata: Model, beliefs and therapy," *Inf. Comput.*, vol. 236, pp. 68–86, 2014.

[33] R. C. Berwick, "Songs to syntax: Cognition, combinatorial computation, and the origin of language," *IJCINI*, vol. 5, no. 4, pp. 22–32, 2011.

[34] A. Villavicencio, M. Idiart, R. C. Berwick, and I. Malioutov, "Language acquisition and probabilistic models: keeping it simple," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pp. 1321–1330, 2013.

[35] R. C. Berwick, P. Pietroski, B. Yankama, and N. Chomsky, "Poverty of the stimulus revisited," *Cognitive Science*, vol. 35, no. 7, pp. 1207–1242, 2011.

[36] I. Malioutov and R. C. Berwick, "Improving statistical parsing by linguistic regularization," in *10th International Conference on Intelligent Systems Design and Applications, ISDA 2010, November 29 - December 1, 2010, Cairo, Egypt*, pp. 1071–1076, 2010.

[37] P. Niyogi and R. C. Berwick, "A markov language learning model for finite parameter spaces," in *32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings.*, pp. 171–180, 1994.

[38] G. J. Jensen, ed., *Cryo-EM, Parts A–C*, vol. 481–483 of *Methods in Enzymology*. Elsevier Inc., 2010.

[39] H. P. Erickson, "The Fourier transform of an electron micrograph—First order and second order theory of image formation," in *Advances in Optical and Electron Microscopy (Volume 5)* (R. Barer and V. E. Cosslett, eds.), pp. 163–199, London and New York: Academic Press, 1973.

[40] J. Lepault and T. Pitt, "Projected structure of unstained, frozen-hydrated T-layer of *bacillus brevis*," *The EMBO Journal*, vol. 3, no. 1, pp. 101–105, 1984.

[41] C. Toyoshima and N. Unwin, "Contrast transfer for frozen-hydrated specimens: Determination from pairs of defocused images," *Ultramicroscopy*, vol. 25, no. 4, pp. 279–291, 1988.

[42] Q. Wang, T. Matsui, T. Domitrovic, Y. Zheng, P. C. Doerschuk, and J. E. Johnson, "Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps," *Journal of Structural Biology*, vol. 181, pp. 195–206, Mar. 2013.

[43] Y. Zheng, Q. Wang, and P. C. Doerschuk, "3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object," *Journal of the Optical Society of America A*, vol. 29, pp. 959–970, June 2012.

[44] C. J. Prust, P. C. Doerschuk, G. C. Lander, and J. E. Johnson, "*Ab initio* maximum likelihood reconstruction from cryo electron microscopy images of an infectious virion of the tailed bacteriophage P22 and maximum likelihood versions of Fourier Shell Correlation appropriate for measuring resolution of spherical or cylindrical objects," *Journal of Structural Biology*, vol. 167, pp. 185–199, 2009.

[45] P. C. Doerschuk and J. E. Johnson, "*Ab initio* reconstruction and experimental design for cryo electron microscopy," *IEEE Transactions on Information Theory*, vol. 46, pp. 1714–1729, Aug. 2000.

[46] J. Tang, B. M. Kearney, Q. Wang, P. C. Doerschuk, T. S. Baker, and J. E. Johnson, "Dynamic and geometric analyses of *Nudaurelia capensis* $\omega$ virus maturation reveal the energy landscape of particle transitions," *J. Molecular Recognition*, vol. 27, pp. 230–237, 10 February 2014.

[47] T. Domitrovic, N. Movahed, B. Bothner, T. Matsui, Q. Wang, P. C. Doerschuk, and J. E. Johnson, "Virus assembly and maturation: Auto-regulation through allosteric molecular switches," *J. Molecular Biology*, vol. 425, pp. 1488–1496, 13 May 2013.

[48] W. Kühlbrandt, "The resolution revolution," *Science*, vol. 343, pp. 1443–1444, 28 March 2014.

[49] S. H. W. Scheres, "A Bayesian view on cryo-EM structure determination," *Journal of Molecular Biology*, vol. 415, pp. 406–418, 13 January 2012.

[50] S. H. W. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination," *Journal of Structural Biology*, vol. 180, pp. 519–530, 2012.

[51] R. Greenspan, M. Mitchell, and J. A. Wise, "Shared principles between the computing and biological sciences," *National Science Foundation*, p. 15, 2011.

[52] I. Korsunsky, D. Ramazzotti, G. Caravagna, and B. Mishra, "Inference of cancer progression models with biological noise," *CoRR*, vol. abs/1408.6032, 2014.

[53] G. Narzisi, B. Mishra, and M. C. Schatz, "On algorithmic complexity of biomolecular sequence assembly problem," in *Algorithms for Computational Biology - First International Conference, AlCoB 2014, Tarragona, Spain, July 1-3, 2014, Proceedigns*, pp. 183–195, 2014.

[54] L. O. Loohuis, A. Witzel, and B. Mishra, "Improving detection of driver genes: Power-law null model of copy number variation in cancer," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 11, no. 6, pp. 1260–1263, 2014.

[55] L. O. Loohuis, A. Witzel, and B. Mishra, "Cancer hybrid automata: Model, beliefs and therapy," *Inf. Comput.*, vol. 236, pp. 68–86, 2014.

[56] J. Jee, L. C. Klippel, M. S. Hossain, N. Ramakrishnan, and B. Mishra, "Discovering the ebb and flow of ideas from text corpora," *IEEE Computer*, vol. 45, no. 2, pp. 73–77, 2012.

[57] A. Sundstrom, S. Cirrone, S. Paxia, C. Hsueh, R. Kjolby, J. K. Gimzewski, J. Reed, and B. Mishra, "Image analysis and length estimation of biomolecules using AFM," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1200–1207, 2012.

[58] F. Vezzi, G. Narzisi, and B. Mishra, "Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons," *CoRR*, vol. abs/1210.1095, 2012.

[59] S. Kleinberg and B. Mishra, "The temporal logic of causal structures," *CoRR*, vol. abs/1205.2634, 2012.

[60] L. O. Loohuis, A. Witzel, and B. Mishra, "Towards cancer hybrid automata," in *Proceedings First International Workshop on Hybrid Systems and Biology, HSB 2012, Newcastle Upon Tyne, UK, 3rd September 2012.*, pp. 137–151, 2012.

[61] F. Menges, G. Narzisi, and B. Mishra, "Totalrecaller: improved accuracy and performance via integrated alignment and base-calling," *Bioinformatics*, vol. 27, no. 17, pp. 2330–2337, 2011.

[62] G. Narzisi and B. Mishra, "Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons," *Bioinformatics*, vol. 27, no. 2, pp. 153–160, 2011.

[63] A. Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 8, no. 3, pp. 775–784, 2011.

[64] S. Kleinberg and B. Mishra, "The temporal logic of token causes," in *KR*, 2010.

[65] A. Mitrofanova, S. Kleinberg, J. Carlton, S. Kasif, and B. Mishra, "Predicting malaria interactome classifications from time-course transcriptomic data along the intraerythrocytic developmental cycle," *Artificial Intelligence in Medicine*, vol. 49, no. 3, pp. 167–176, 2010.

[66] A. Mitrofanova, M. Farach-Colton, and B. Mishra, "Efficient and robust prediction algorithms for protein complexes using gomory-hu trees," in *Pacific Symposium on Biocomputing*, pp. 215–226, 2009.

[67] B. Mishra, "Technical perspective - where biology meets computing," *Commun. ACM*, vol. 52, no. 3, p. 96, 2009.

[68] S. Tadepalli, N. Ramakrishnan, L. T. Watson, B. Mishra, and R. F. Helm, "Simultaneously segmenting multiple gene expression time courses by analyzing cluster dynamics," *J. Bioinformatics and Computational Biology*, vol. 7, no. 2, pp. 339–356, 2009.