Extended synopsis of the GUDHI project

Geometry understanding in higher-dimensions. The central goal of this proposal is to settle the algorithmic foundations of geometry understanding in dimensions higher than 3. We coin the term *geometry understanding* to encompass a collection of tasks including the computer representation and approximation of geometric structures, and the inference of geometric or topological properties from sampled shapes.

The need to understand geometric structures is ubiquitous in science and has become an essential part of *scientific computing* and *data analysis*. Geometry understanding is by no means limited to three dimensions and many applications in physics, biology, and engineering require a keen understanding of the geometry of a variety of *higher dimensional spaces*. Let us mention phase space in particle physics, invariant manifolds in dynamical systems, configuration spaces of mechanical systems, conformational spaces of molecules, image manifolds, and shape spaces, to name a few. In *data analysis and manifold learning*, data are often thought as points in some high-dimensional metric space [Donoho, 2000]. Those points are usually not uniformly distributed in the embedding space but lie close to some *low-dimensional manifold*, which reflects the fact that the physical system that produced the data has a moderate number of degrees of freedom. Understanding the geometry of this manifold is key to understanding the underlying system.

In many applications, these manifolds are highly nonlinear and have a non trivial topology : a precise understanding of their geometry is out of reach of current techniques. Let us illustrate our motivation and objectives through the paradigmatic example of *energy landscapes of molecules*. Understanding energy landscapes is a major challenge in chemistry and biology but, despite a lot of efforts and a wide variety of approaches, little is understood about the actual structures of these landscapes. The case of cyclo-octane C_8H_{16}, a cyclic alkane used in manufacture of plastics, is instructive. This relatively simple molecule has been studied in chemistry for over 40 years but it is only very recently that its conformational space has been fully understood. By analyzing a dataset of 1M points in $\mathbb{R}^{72}$, each describing a cyclo-octane conformation, it has been shown that the conformation space of cyclo-octane has the unexpected geometry of a multi-sheeted 2-dimensional surface composed of a sphere and a Klein bottle, intersecting in two rings [Martin et al., 2010]. Besides its fundamental interest, such a discovery opens new avenues for understanding the energy landscape of cyclo-octane. Extending this type of analysis to large molecules, in particular to proteins, would have tremendous implications. Many other examples with high potential remain to be solved in domains as varied as neurosciences, medical imaging, speech recognition and astrophysics. *The crux is to have access to robust and efficient data structures and algorithms to represent and analyze geometry in higher dimensions.* This is the grand challenge the Gudhi project wants to take up.

The curses of higher-dimensional geometry. Many difficulties have to be faced when processing and analyzing high-dimensional geometries. First, the dimensionality severely restricts our intuition and ability to visualize the data. Understanding higher dimensional shapes must hence rely on *automated* methods and tools that produce *provably correct* results under *realistic circumstances*.

Second, major difficulties come from the fact that the complexity of data structures and algorithms used to approximate shapes rapidly grows as the dimensionality increases, which makes them intractable in high dimensions. This phenomenon, referred to as the *curse of dimensionality*, prevents, in particular, subdividing the ambient space, as is usually done in 3-space, since the size of any such subdivision depends exponentially on the ambient dimension. Instead, any practical method must be *sensitive to the intrinsic dimension* (usually unknown) of the shape under analysis. As already mentioned, and observed in the cyclo-octane example, the intrinsic dimension can often be assumed to be much lower than the ambient dimension. This is a powerful assumption we refer to as the *manifold model*.[1]

In addition, high-dimensional data often suffer from significant *defects*, including sparsity, noise, and outliers that may hide the intrinsic dimension of the underlying structure. This is particularly so in the case of

[1] We use the term manifold in a rather liberal way, including stratified manifolds.

biological data, such as high throughput data from microarrays or other sources.

The emergence of geometry understanding in higher dimensions. The last decade has seen tremendous progress in the understanding of geometry in high-dimensional spaces. In signal and image processing, and in machine learning, a variety of techniques, known as *nonlinear dimensionality reduction* have been proposed to reduce the dimension of data, to learn nonlinear manifolds and to cluster data [Lee and Verleysen, 2007]. Such techniques are widely used but have limited guarantees and impose strong constraints on the dimension or topology of the shapes they can successfully handle.

The techniques developed in *computational geometry and topology* are complementary. They aim at processing and analyzing shapes with non trivial geometry and topology [Edelsbrunner and Harer, 2010]. Emblematic problems such as mesh generation and surface reconstruction in 3-dimensions are now well-understood and several provably correct and highly efficient solutions are available [Boissonnat and Teillaud, 2006, Dey, 2007]).

Attempts to analyze higher dimensional shapes led to the development of beautiful pieces of theory with deep roots in various areas of mathematics like Riemannian geometry, geometric measure theory, differential and algebraic topology. Let us mention the new and rapidly growing fields of geometric inference [Chazal et al., 2011], persistent homology [Edelsbrunner and Harer, 2010], and topological data analysis [Carlsson, 2009]. These advances attracted interest from several fields like chemistry [Martin et al., 2010], astrophysics [van de Weygaert et al., 2011], biological data analysis [Fekete et al., 2009], or sensor networks [Ghrist, 2008]. However, until now, the applications have been limited to small data sets or moderate dimensions. Interestingly, G. Carlsson, a mathematician from Stanford university recently launched a start up company that applies some of these techniques to visualization of data sets (`(http://www.ayasdi.com/`).

The grand challenge : settling the algorithmic foundations. Since many of the most promising applications are in higher dimensions, there is a pressing need to develop more effective tools that scale to real problems. We identify the lack of algorithmic foundations for geometry understanding in higher dimensions as the main cause of the current limited impact of geometric and topological methods. Settling such foundations is a challenge of great theoretical and practical significance at the heart of the Gudhi project.

A *tenet* of this proposal is that, to take up the challenge, we need a *global approach* involving tight and long-standing interactions between *mathematical research, algorithmic design* and *advanced software development.* We believe that this is key to obtaining methods with built-in robustness, scalability and guarantees, which are the best promises for *impact in the long run.*

We strongly believe that, by following this paradigm, our ambitious objective is realistic and can be reached. To pave the way towards this goal, we have identified four main scientific challenges.

Scientific challenge 1 : Choosing the right representation. As discussed above, dimensionality reduction techniques cannot provide precise approximation of complicated shapes (as required in scientific computing) nor compute essential features of a shape like its topological invariants. More expressive representations of shapes are provided by *simplicial complexes*, the higher dimensional analogue of triangulations. Simplicial complexes are combinatorial structures (a special type of hypergraphs) that encode relationships between subsets of points. Simplicial complexes can be used to produce manifold meshes well suited to scientific computing purposes [Henderson, 2002], or much cruder approximations still useful to infer some important features of shapes such as their homology or some local geometric properties [Niyogi et al., 2008]. A *central tenet* in this project is to regard *simplicial complexes as a unifying representation of shapes for*

geometry understanding in higher dimensions.

Many types of simplicial complexes, e.g. the Čech or Rips complexes, can be used and a first issue is to choose the appropriate type. This choice depends on the *combinatorial and algorithmic complexities* of the complex, as well as on its *power to approximate* a shape. The choice of the underlying *metric* is another fundamental issue that determines the type and quality of an approximation. The simple Euclidean distance in the ambient space, while easy to deal with, is often not the right choice. As already mentioned, when working in high dimensional spaces, the objects of interest have often an intrinsic dimension much smaller than the ambient dimension. It is thus important to exploit the *intrinsic geometry* of the objects. Computational intrinsic geometry has not been seriously tackled yet and even a basic question like the existence of Delaunay triangulations on Riemannian manifolds has been elusive so far. This question is of utmost importance for anisotropic mesh generation and optimal approximations. Another important situation is when data are not provided as point clouds in some Euclidean space, but rather as a matrix of pairwise distances (i.e., a *discrete metric space*). Although such data may not be sampled from geometric subsets of Riemannian manifolds, they may still carry some interesting topological structures that need to be understood. Lastly, let us mention other *pseudo-distances* such as Kullback-Leibler, Itakura-Saito or Bregman divergences that may be preferred in information theory, and signal and image processing. These divergences are usually not genuine distances (they may not be symmetric nor satisfy the triangular inequality) and geometric data structures and algorithms need to be revisited in this context (see [Boissonnat et al., 2010]).

Scientific challenge 2 : Bypassing the curse of dimensionality. Simplicial complexes have been known and studied for a long time in mathematics, but not so much from a computational point of view. This is however a main issue since the complexity of many geometric algorithms and data structures grows exponentially with increasing dimension. It is thus not possible to partition a high dimensional space, which rules out most, if not all, geometric algorithms developed in low dimensions. Hence, extending computational geometry in high dimensions cannot be done in a straightforward manner and one has to take advantage of additional structural properties of the problem. In this project, we will address the curse of dimensionality by focusing on the inherent structure in the data which we assume to be of relative *low intrinsic dimension.* We will put the emphasis on *output-sensitive* algorithms and on *average-case* analysis. First investigations led to very promising results, such as the design of new simplicial complexes with low complexity and approximation algorithms that scale well with the dimension [Boissonnat and Ghosh, 2014].

Scientific challenge 3 : Searching for stable models. When dealing with approximations and samples, one needs stability results to ensure that the quantities that are computed are good approximations of the real ones. This is especially true in higher-dimensions where data are usually corrupted by various types of noise. When the noise magnitude is small, methods have been proposed to robustly estimate topological and geometric properties of shapes [Niyogi et al., 2011]. The recent and fast developing theory of *persistent homology* provides a powerful tool to study the homology of sampled spaces and to remove topological noise [Edelsbrunner and Harer, 2010, Oudot et al., 2013]. However, in many applications, the noise is non local and the previous methods fail. Recently, larger families of noise models have been considered and statistical approaches have been proposed to provide shape approximations that are stable with respect to those types of noise [Genovese et al., 2011]. These methods however do not provide topological guarantees on the approximation and the question of designing computationally tractable estimators converging at an optimal rate remains open. New ideas such as those in [Chazal et al., 2011] are required to design unifying frameworks that *embrace statistical approaches and deterministic methods*, and offer topological guarantees.

Scientific challenge 4 : Turning theory into practice. A major challenge, if not the most important, is to develop *theory* that is *of practical significance* for applications. To take up the challenge, we will undertake the development of a *software platform* devoted to geometry understanding in higher dimensions. We consider such a platform as central to our research for three main reasons. *First*, the software platform will allow *large scale experimentation*, which is mandatory to design the right models and data structures [Boissonnat and Maria, 2012]. We believe that this will revitalize the current theory and open *new vistas for research*, both of a practical and a theoretical nature, leading towards a virtuous cycle between theory and experimental research. This has proven to be highly fruitful when developing the CGAL library (`http://www.cgal.org`). The same will be even more true in high dimensional geometry.

Second, maintaining such a platform will help further efforts and enable *consolidation in the long run*. Having a library with interoperable modules will allow us to incrementally add more and more sophisticated tools based on solid foundations. This is consistent with our long-term vision and our conviction that it is only through such a long standing effort that true impact, both theoretical and applied, can be gained.

Third, the platform will serve as a unique tool to *communicate* with the computational geometry community and with researchers from other fields. In return, we will get feedback from practitioners which will help shape the theoretical models and the software platform.

Objectives and research roadmap. The ambition of this proposal is to settle the *algorithmic foundations of geometry understanding in dimensions higher than 3*. We intend to develop *scalable representations in the form of simplicial complexes* and *practical algorithms* to approximate *highly nonlinear shapes*, and to infer geometric and topological properties from data subject to significant *defects* and under *realistic conditions*. As is common in many applications across science and engineering, we will assume that the objects of interest can be modeled as *low-dimensional manifolds* embedded in possibly high-dimensional spaces. By exploiting the *intrinsic properties* of the objects, and inventing data structures and algorithms that are sensitive to the intrinsic dimension, we intend to *break the current computational bottleneck*.

To reach these objectives, the guiding principle will be to foster a symbiotic relationship between theory and practice, and to address *fundamental research* issues along three parallel advancing fronts. We will simultaneously develop *mathematical approaches* providing theoretical guarantees, *effective algorithms* that are amenable to theoretical analysis and rigorous experimental validation, and *perennial software* development.

The proposal is structured into the following four *focus areas* that address the four scientific challenges listed above. **A1** – *Data structures sensitive to the intrinsic dimension* – will address Challenges 1 and 2 by extending current knowledge on the combinatorial and algorithmic properties of simplicial complexes. **A2** – *Triangulation of non Euclidean metric spaces* – will address Challenges 1 and 2 by developing effective algorithms to mesh or reconstruct manifolds equipped with various metrics. **A3** – *Robust models for geometric and topological inference* – will address Challenge 3 by providing the crucial algorithms for topological data analysis. **A4** – *Software platform for geometric understanding in high dimensions* – will address Challenge 4 by providing the software environment for experimenting with our new data structures and algorithms, for integrating them in a library of interoperable modules, and for diffusing our results to applied fields.

References

[Boissonnat and Ghosh, 2014] Boissonnat, J.-D. and Ghosh, A. (2014). Manifold reconstruction using tangential Delaunay complexes. *Discrete and Computational Geometry*, 51(1):221–267.

[Boissonnat and Maria, 2012] Boissonnat, J.-D. and Maria, C. (2012). A data structure to represent simplicial complexes. In *Proc. of the 20th European Symposium on Algorithms*, ESA 2012.

[Boissonnat et al., 2010] Boissonnat, J.-D., Nielsen, F., and Nock, R. (2010). Bregman Voronoi diagrams. *Discrete and Computational Geometry*, 44(2).

[Boissonnat and Teillaud, 2006] Boissonnat, J.-D. and Teillaud, M., editors (2006). *Effective Computational Geometry for Curves and Surfaces*. Springer-Verlag.

[Carlsson, 2009] Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc., 46*, pages 255–308.

[Chazal et al., 2011] Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011). Geometric inference for probability measures. *Journal on Foundations of Computational Mathematics*, 11(6):733–751.

[Dey, 2007] Dey, T. (2007). *Curve and Surface Reconstruction : Algorithms with Mathematical Analysis*. Cambridge University Press.

[Donoho, 2000] Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *Mathematical Challenges of the 21st Century*. AMS.

[Edelsbrunner and Harer, 2010] Edelsbrunner, H. and Harer, J. (2010). *Computational topology*. American Mathematical Society.

[Fekete et al., 2009] Fekete, T., Pitowsky, I., Grinvald, A., and Omer, D. (2009). Arousal increases the representational capacity of cortical tissue. *Journal of Neurosciences*, 27:211–227.

[Genovese et al., 2011] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2011). Minimax manifold estimation. *arXiv:1007.0549v3*.

[Ghrist, 2008] Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc., 45(1)*, pages 61–75.

[Henderson, 2002] Henderson, M. E. (2002). Multiple parameter continuation: computing implicitly defined k-manifolds. *Int. Journal of Bifurcation and Chaos*, 12:451–476.

[Lee and Verleysen, 2007] Lee, J. A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer.

[Martin et al., 2010] Martin, S., Thompson, A., Coutsias, E. A., and Watson, J.-P. (2010). Topology of cyclo-octane energy landscape. *The journal of chemical physics*, 132(234115).

[Niyogi et al., 2008] Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete and Computational Geometry*, 39(1):419–441.

[Niyogi et al., 2011] Niyogi, P., Smale, S., and Weinberger, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM J. Comput.*, 40:646–663.

[Oudot et al., 2013] Oudot, S. Y., De Silva, V., and Chazal, F. (2013). Persistence Stability for Geometric complexes. *Geometriae Dedicata*, on-line first(on-line first):on–line first.

[van de Weygaert et al., 2011] van de Weygaert et al., R. (2011). Alpha, Betti and the megaparsec universe: on the homology and topology of the cosmic web. In *Trans. on Comp. Science XIV*, Lecture Notes in Computer Science Vol. 6970. Springer-Verlag.