ERC Advanced Grant 2012
research proposal (PART B2)

Algorithmic Foundations of Geometry Understanding in Higher Dimensions (GUDHI)

1 The research proposal

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 approximation and computer representation of geometric structures, and the inference of geometric or topological properties of sampled shapes.

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. This is also the central assumption of *nonlinear dimensionality reduction* techniques which have been flourishing recently in signal and image processing, and in machine learning. The geometric and topological approach developed in this project is complementary. By exploiting the *intrinsic properties* of the objects, we will produce data structures and algorithms that are sensitive to the intrinsic dimension, and therefore will *break the current computational bottleneck of geometric and topological methods.* At the same time, we will be able to process and analyze shapes whose complexities in terms of geometry and topology are beyond what dimensionality reduction techniques permit.

We now detail the scientific strategy and the organization of the Gudhi project. In view of the current state of the art, we identify four main scientific objectives : to develop *scalable representations* of shapes in the form of *simplicial complexes*; to design *practical algorithms* to approximate *highly nonlinear shapes*; to design robust algorithms to infer geometric and topological properties from data subject to significant *defects* and under *realistic conditions*; to develop *an open source software platform* to foster research and impact. Accordingly, the research work will be organized in four focus areas.

1.1 State of the art and scientific objectives

Dimensionality reduction. A powerful and widely accepted assumption in scientific computing and data analysis to bypass the curse of dimensionality is that the objects of interest can be modeled as *low-dimensional manifolds*, even if they are embedded in high-dimensional ambient spaces. Low here is to be understood with respect to the ambient dimension and may be significantly higher than 3. This powerful assumption is supported by the fact that data are most of the time associated with physical systems that have relatively few degrees of freedom. This assumption is valid across science and engineering and is at the heart of dimensionality reduction and manifold learning. *Dimensionality reduction* techniques intend to infer the intrinsic dimensionality of the data, as well as to provide structure-preserving mappings of the data into lower-dimensional spaces. *Nonlinear* dimensionality reduction techniques [Lee and Verleysen, 2007] are capable of discovering nonlinear structures and have been successfully applied to analyze data in a wide range of applications. Nevertheless, these methods come with no or very limited guarantees. For example, Isomap [Tenenbaum et al., 2000] provides a correct embedding only if the manifold is isometric to a convex open set of $\mathbb{R}^k$, where k is the dimension of the manifold, and LLE [Roweis and Saul, 2000] can only

reconstruct topological balls. *Geometric and topological methods* are complementary to dimension reduction methods. They intend to better capture the geometry and the topology of manifolds by representing them as *simplicial complexes.*

Scalable representations in the form of simplicial complexes. Simplicial complexes are combinatorial structures, a special case of hypergraphs, that encode proximity relationships between subsets of points. Simplicial complexes have been extensively used in $\mathbb{R}^3$ to produce fine meshes providing a precise approximation of the geometry (for the Hausdorff distance) and of the topology (homeomorphism) of nonlinear manifolds. Such precise approximations are well suited for scientific computing and visualization purposes. In computational topology, simplicial complexes are also used to infer topological properties of sampled shapes, most importantly *homological invariants* [Hatcher, 2002] that count the number of connected components, tunnels, voids etc. Well known examples of simplicial complexes used in this context are the Čech and the Rips complexes. Those complexes depends on a parameter ε that plays the role of a scale. By varying ε, we obtain nested families of simplicial complexes, known as *filtrations.* Multiscale topological analysis based on *filtrations* is at the heart of the recent theory of *persistent homology* [Edelsbrunner and Harer, 2010, Zomorodian, 2009].

Simplicial complexes have been studied for a long time by the mathematical community but the *algorithmic* side of the theory is still in its infancy. Most of the known simplicial complexes are either extremely difficult to compute, even in moderate dimensions, because their construction involves high-degree algebra (e.g., the Čech complex) or are easy to compute (e.g., the Rips complex), but their size is then so big that they cannot be constructed from real high-dimensional data. Current research aims at finding new types of simplicial complexes that are easy to compute while still capturing the essential geometric and topological features of shapes. Recent progress has been made with the discovery of Delaunay-like complexes, such as the witness complex [Carlsson and de Silva, 2004] and the Delaunay tangential complex [Boissonnat and Ghosh, 2010a], that offer new tradeoffs between complexity and approximation quality.

Another issue is the design of *data structures* to encode simplicial complexes. Little work has been done on this subject. Brisson [Brisson, 1989] and Lienhardt [Lienhardt, 1994] have introduced data structures to represent d-dimensional cell complexes. These data structures are general and powerful, but very redundant, and they do not scale to large data sets in high dimensions. In contrast, one could store only the 1-skeleton of the complex (edges and vertices). This saves a lot of memory space at the penalty of having to recover the full complex when needed, e.g. when one wants to store a filtration. This expansion can be done in a purely combinatorial way, thus very efficiently, in the special case of flag complexes. Elaborating upon this idea, Attali et al. [Attali et al., 2011] have proposed a concise data structure that is efficient when applied to simplicial complexes that are *close* to flag complexes. Still, designing efficient data structures to represent and process general simplicial complexes is a widely open area in dire need of advances in combinatorics, new algorithmic paradigms and new analyses. See [Boissonnat and Maria, 2012] for a recent contribution.

Objective 1: *To extend current knowledge of simplicial complexes, most notably their combinatorial and algorithmic properties. To fully understand Delaunay-like simplicial complexes. To design new compact data structures and algorithms to encode and process simplicial complexes.*

Geometric approximation of highly nonlinear shapes. In low dimensions, computing a geometric approximation of a given geometric object is a well-understood problem [Boissonnat and Teillaud, 2006,

Dey, 2007], and good approximations can be efficiently constructed. The situation is quite different in higher dimensions. Although the mathematical literature on triangulation of manifolds is abundant, few effective algorithms have been proposed and evaluated. The main issue is to avoid computing a subdivision of the ambient space since this would lead to algorithms that scale exponentially with the ambient dimension. Very few attempts have been made to exploit the low-dimensional manifold assumption. To analyze dynamical systems in science and engineering, higher-dimensional *continuation methods* have been proposed to mesh solution manifolds [Henderson, 2002]. These methods are however lacking guarantees and are restricted in practice to very low dimensional manifolds. Recently, we made progress towards a better understanding of the complexity issues when triangulating smooth submanifolds and provided a provably correct algorithm that scales *linearly* with the ambient dimension [Boissonnat and Ghosh, 2010b]. It is however still a challenge to turn this theoretical result into a fully practical algorithm.

A natural way to bypass the curse of dimensionality is to directly work on the manifold and to resort to *intrinsic geometry*. What is lacking however is a full development of Riemannian computational geometry. A central question in our context, which has been elusive for a long time, is to provide sampling conditions asserting the existence of intrinsic Delaunay triangulations on Riemannian manifolds. Previously announced sampling criteria [Leibon and Letscher, 2000] for the existence of intrinsic Delaunay triangulations have recently been demonstrated to be insufficient, and sufficient criteria were introduced [Boissonnat et al., 2012]. This opens stimulating avenues for new developments on Delaunay triangulations in Riemannian manifolds and other non Euclidean spaces. Here too, there is a dire need for an algorithmic side of the theory and for the development of efficient, possibly approximate, provably correct algorithms.

Another fundamental problem is *shape reconstruction* which consists in computing an approximation of a geometric structure given a point sample. In low dimensions, effective reconstruction techniques exist that can provide precise geometric approximations very efficiently and under reasonable sampling conditions [Dey, 2007]. However, almost all these methods rely on the computation of a triangulation of the ambient space and previous attempts to extend them to higher dimensions led to algorithms whose complexities depends exponentially in the ambient dimension. Despite some very recent results [Boissonnat and Ghosh, 2010a], designing practical algorithms that can reconstruct smooth submanifolds of high-dimensional spaces under mild sampling conditions remains widely open. Extension beyond smooth manifolds is even more challenging.

Objective 2 : *To develop new algorithms to* triangulate non Euclidean metric spaces*. To study intrinsic Delaunay triangulations of Riemannian manifolds. To reconstruct submanifolds using Delaunay-like simplicial complexes. To construct crude approximations with quality guarantees. To extend current knowledge on processing non-smooth manifolds.*

Robust geometric and topological inference. Since computing precise geometric approximations is currently only possible under strong assumptions that may not be met in some applications, we can look for *cruder information* that can still unveil some of the properties of the structures underlying the data. A prominent example is *homology* that can be computed without a precise reconstruction and under less restrictive conditions [Chazal et al., 2009b, Niyogi et al., 2008]. The rapidly growing theory of *computational homology* [Kaczynski et al., 2003] and, in particular, of *persistent homology* [Edelsbrunner and Harer, 2010, Ghrist, 2008, Zomorodian, 2009] was recently introduced as a powerful tool for the study of the topological invariants of sampled shapes. The approach consists of building a simplicial complex whose elements

are filtered by some user-defined function. The filtration is then used to remove topological noise and to report the stable topological features. These advances in computational topology attracted interest in the mathematical community and in several fields like neurosciences, computer vision or sensor networks [Fekete et al., 2009, Carlsson et al., 2008, de Silva and Ghrist, 2007]. What still prevents these new methods from highly impacting applications is the lack of efficient data structures and algorithms to construct simplicial complexes and filtrations in high dimensions (Objective 1).

In addition, practitioners need methods that are stable with respect to *noise, sparsity, outliers and other defects* that corrupt data. The geometric approaches have so far only considered very restrictive noise models [Niyogi et al., 2008]. Larger families of noise models have recently been considered and statistical approaches have been proposed to provide geometric approximations that are stable with respect to those models [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. A major challenge is to design unifying frameworks that embrace *statistical approaches* and deterministic methods, and offer topological guarantees.

Shape descriptors are used in a variety of applications, including shape classification, shape retrieval, shape matching, shape registration, and symmetry detection. A fundamental question is *how much information* about a shape can be recovered from descriptors. Most existing work in shape analysis only provides lower bounds on a shape distance (e.g., the Gromov-Hausdorff distance) based on descriptor distance. There are very few exceptions to this rule, and the upper bounds on shape distances obtained so far are quite loose [Bronstein et al., 2006, Mémoli and Sapiro, 2005]. Deriving tight upper bounds is thus still widely open. Building on the persistent homology theory new topological descriptors have been recently introduced. These descriptors can be robustly estimated from filtrations built on top of measurement data points and open new directions to explore [Chazal et al., 2009a, Skraba et al., 2010].

Objective 3 : *To study new* robust models for geometric and topological inference. *To combine statistical and geometric and topological approaches for data analysis. To study topological signatures of shapes.*

Theory versus practice. Geometric and topological methods are well behind dimensionality reduction techniques in terms of software development and applications. To go beyond low-dimensional examples, one needs efficient and robust software to construct and manipulate simplicial complexes in dimensions higher than 3. Only a very few such software exist. *Qhull* (`http://www.qhull.org/`) can compute convex hulls and Delaunay triangulations in moderate dimensions, but is of little use in the context of geometry understanding since it only constructs full-dimensional triangulations. *Multifario* (`http://multifario.sourceforge.net/`) is a set of subroutines and data structures dedicated to *meshing* manifolds that occur in dynamical systems. The software can in principle construct geometric approximations of manifolds of arbitrary codimension, but examples are only reported in very low dimensions. Interestingly, the algorithm implemented in Multifario uses a multiple parameter continuation approach and has some similarity with our algorithm [Boissonnat and Ghosh, 2010b], which has theoretical guarantees and whose complexity is only linear in the ambient dimension. *Polymake* (`http://polymake.org/`) can handle several types of complexes, build Voronoi diagrams, and compute advanced topological characteristics of objects like a finite representation of the fundamental group. However, it is more oriented towards an interactive use for mathematical experimentation rather than routine automation at large scales.

Several software libraries exist for homology computation. RedHom (`http://redhom.ii.uj.edu.pl/`)

computes Betti numbers and torsion coefficients of cubical sets, simplicial complexes and general, regular CW complexes (http://redhom.ii.uj.edu.pl/). Two implementations of *persistent homology* algorithms are currently available, PLEX a package for Matlab (`http://comptop.stanford.edu/u/programs/jplex/`), and Dionysus (`http://www.mrzv.org/software/dionysus/`). They both offer the construction of several types of simplicial complexes. PLEX has been successfully applied in low dimensions [Fekete et al., 2009, Ghrist, 2008]. Dionysus offers advanced functionalities but has no documentation and a limited diffusion. As evidenced by recent experiments [Boissonnat and Maria, 2012], huge improvements are to be expected.

Objective 4 : *To devise an open source* software platform for geometric understanding in high dimensions. *To apply our techniques to a few key problems where they may have a huge impact. To disseminate our methods and to benchmark our results on real data from various applied fields.*

1.2 Research roadmap

The proposal is structured into four focus areas that address the objectives mentioned above.

Focus Area 1: Data structures sensitive to the intrinsic dimension

A central tenet in this project is that simplicial complexes are the appropriate representation of shapes for geometry understanding in higher dimensions. They will be at the heart of all the developments undertaken in all four focus areas. The challenge is to devise small simplicial complexes with good approximation properties, to analyze their complexity and to propose succinct data structures to encode them. We will work along the following directions.

Delaunay-like simplicial complexes. Simplicial complexes have recently been derived from the Delaunay triangulation. Given a set of point P lying close to a manifold $\mathbb{M}$, the intent is to construct simplicial complexes on P that provide better tradeoffs between complexity, approximation quality and topological faithfulness than the Rips and the Čech complexes. Two such simplicial complexes are of special interest for meshing and reconstructing manifolds in high dimensional spaces (Focus Area 2), namely the witness complex [Carlsson and de Silva, 2004], and the tangential Delaunay complex [Boissonnat and Ghosh, 2010a]. Both can be considered as approximations of the restricted Delaunay complex, a subcomplex of the Delaunay triangulation that is known to be topology equivalent to $\mathbb{M}$ under appropriate conditions on P [Dey, 2007]. While the restricted Delaunay complex is extremely difficult or even impossible to compute, the tangential Delaunay and the witness complexes are good candidates for practical implementations.

Both structures use ideas that can be combined. The tangential complex constructs local triangulations in a way that is reminiscent of dimensionality reduction techniques. Since dimensionality reduction has been flourishing during the last decade, many variations on the basic idea of the tangential complex are conceivable, with the the potential of enlarging its usability. In addition to being easy to compute, the witness complex brings the idea of landmarking (subsampling), which can also be used in conjunction with the tangential Delaunay complex. We will further look for new variants combining dimensionality reduction, landmarking and Delaunay triangulations. We want to understand the various properties of Delaunay-based simplicial complexes as well as their relationships. See [Boissonnat et al., 2012] for a first step in that direction.

Combinatorial and algorithmic complexity. Since the main limitation to using simplicial complexes is their combinatorial and algorithmic complexity, we will devote effort at better understanding those questions, both theoretically and experimentally. A central challenge is to obtain *complexity bounds* for simplicial complexes of well sampled substructures such as submanifolds. We will consider various types of random or deterministic simplicial complexes. We intend to measure the effect of *perturbations* (either noise or computed perturbations) on the mathematical properties and combinatorial complexity of those structures, and to develop *probabilistic analyses*. In addition to their combinatorial complexity, the *complexity of algorithms* that construct the simplicial complexes is to be precisely analyzed under realistic models. In particular, expected complexity and output-sensitive complexity analyses will be performed in addition to worst-case analysis. We will also develop parallel as well as out-of-core algorithms.

Compact data structures to encode simplicial complexes. We will develop efficient data structures to encode *general* simplicial complexes. An open problem is to establish bounds on the minimal size of data structures encoding simplicial complexes. Another major challenge is to design succinct data structures in the spirit of what has been done for trees and graphs [Ferragina et al., 2005, Munro and Raman, 2002]. Theoretical guarantees and large scale experimentation on various types of simplicial complexes are mandatory. As a first encouraging step in this direction, we have devised a tree structure that can store all the simplices of any simplicial complex in a compact way while it can support fast queries. Simplicial complexes of 500 million simplices can be routinely constructed and stored on a laptop computer [Boissonnat and Maria, 2012]. This data structure already represents a significant leap forward from the state of the art. It deserves further analysis and optimisation, and further research is expected to yield encodings capable of sizes orders of magnitude higher still. In addition to designing compact data structures, we will also consider *simplifications* of simplicial complexes that preserve geometric and topological properties.

Focus Area 2: Triangulation of non Euclidean metric spaces

The challenge is to develop algorithms that can construct simplicial complexes approximating manifolds and spaces, and to do this efficiently and with guarantees on the approximation qualities of the output. An example of a traditionally desired guarantee is the requirement that the constructed complex be homeomorphic to the original space.

Intrinsic Delaunay triangulations of Riemannian manifolds. The Delaunay paradigm has proven to be central to the development and understanding of meshing algorithms, whether the domain of interest is a full dimensional subset of $\mathbb{R}^d$, or a more general manifold. Locally a manifold can be well approximated by a Euclidean space, and this is the main idea exploited by dimension reduction and dimension detection algorithms. In fact, the same idea can be found at the heart of anisotropic meshing algorithms and global meshing and reconstruction algorithms. The problem then boils down to ensuring that the locally constructed complexes knit together into a coherent whole. This issue is tackled by manipulating the complex either through strategic refinement or perturbation of the weights defining the neighbour relations.

The underlying problem is that the Delaunay paradigm is being applied where the meaningful metric is not Euclidean. In order to strengthen the theoretical foundations for tackling these issues, we intend to develop a deeper understanding of the intrinsic Delaunay complex defined by a Riemannian metric.

The conditions for the existence of such intrinsic Delaunay triangulations have so far only been established through extrinsic criteria related to manifolds embedded in Euclidean space [Boissonnat et al., 2012]. We plan to establish sampling criteria based solely on intrinsic properties of the manifold. In this abstract setting, the techniques required to demonstrate a homeomorphism with the original manifold will be different, but likely to be more general than those so far developed for specific substructures of an ambient Delaunay triangulation. We will then devise an algorithm for constructing the intrinsic Delaunay complex. Since exact computation of geodesic distances is out of reach in many cases, in particular when the manifold is only known through a point sample, we will develop algorithms that are *robust* with respect to approximate intrinsic distance computations. We will take inspiration from our recent work on anisotropic mesh generation [Boissonnat et al., 2008]. We expect that these results will lead to insights into the meshing of manifolds equipped with *alternatives to a metric*, such as Bregman divergences with locally defined potential functions.

Manifold reconstruction using Delaunay-like structures. The tangential complex paved the way for a manifold reconstruction algorithm that does not depend exponentially on the ambient dimension [Boissonnat and Ghosh, 2010a]. We would like to extend this success to witness complex based techniques which are free of expensive geometric predicates: so far, the complexity of witness complex based manifold reconstruction is exponential in the ambient dimension d. As a first step in that direction, we established sufficient (albeit quite restrictive) conditions under which the witness complex, and the restricted Delaunay triangulation are identical [Boissonnat et al., 2011]. As yet, no algorithm for landmark selection has been devised which meets these required sampling conditions. We plan to tackle this problem through perturbation techniques, using the tangential complex paradigm to avoid dependence on the ambient dimension.

Crude models. The homeomorphism guarantees obtained for triangulating a manifold still require in general sampling criteria which may clash with the practical reality. Even if the manifold is known to sufficient accuracy to allow for such sampling, light-weight representations may be preferred in some applications. In order to progress towards practical algorithms with meaningful guarantees, satisfactory tradeoffs must be discovered.

We plan to explore two avenues to address the problem. On the one hand, we will strive to attain relaxed, parameterisable, approximation quality metrics that yield a meaningful comparison between the algorithmic output and the underlying manifold represented by the given input data. This approach is applicable in the case where the manifold is known to high precision, and only the output representation is crude, due to a fixed vertex budget, for example. Evaluations based on Gromov-Hausdorff distance, and Wasserstein-type distances will be investigated.

On the other hand, an appropriate approach, when crude input data is the only explicit information about the underlying manifold, is to assume that the manifold belongs to some restricted family of manifolds which can be distinguished from each other on the basis of little information. We thus wish to be able to prove that the output represents the "projection" of the true manifold into a restricted space of manifolds.

Stratified manifolds. While manifold triangulation and reconstruction in higher dimensions already represent a challenge for effective practical algorithms, there is a need for advancing the knowledge on more complicated objects than manifolds. Stratified manifolds represent a potentially tractable yet flexible generalisation that can model many known naturally occurring structures. Examples include conformation

spaces of molecules, such as that discovered for cyclo-octane [Martin et al., 2010], and also the invariant sets that appear in dynamical systems [Henderson, 2002]. Methods have been developed for meshing and reconstructing surfaces with boundaries. Also, algorithms have been proposed for separating the strata of stratified manifolds [Bendich et al., 2007]; the resulting strata being manifolds with boundaries.

We intend to devise algorithms for meshing and reconstructing manifolds with boundary with an aim towards applications to stratified manifolds. New ideas will be required to develop sampling conditions which can yield guarantees for these structures. The algorithms will draw upon the theory developed for pure manifolds, as well as upon the geometric inference algorithms discussed in Focus Area 3. In particular, linear subspace clustering approaches will be required to locally resolve different strata.

Focus Area 3: Robust models for geometric and topological inference

The goal is to infer geometric and topological properties from defect-laden data. A major challenge is to combine statistical approaches relying on powerful models of noise and deterministic methods coming with topological guarantees.

Homology inference. Building upon the distance function approach, algorithms have been developed to infer the homology of general shapes from Čech or Rips complexes built on top of the data. These algorithms are efficient and stable under small Hausdorff noise [Chazal and Oudot, 2008]. The basic idea is that the topology of the sampled shape is carried by the topology of the sublevel sets of the distance function to the data points, which in turn is related to the topology of the Čech and the Rips complexes.

To comply with the presence of noise and outliers in the data, we intend to explore different approaches inspired from these algorithms. We will in particular focus on a new paradigm for point cloud data analysis that was recently proposed by researchers from my group. Point clouds are no longer considered as mere compact sets but rather as *empirical measures*. A notion of distance to such probability measures has been defined and shown to be stable with respect to perturbations of the probability measure [Chazal et al., 2011]. It has also been shown that the sublevel sets of this distance function carry the geometric information about the probability measure. If we consider a model where the data is generated from a probability distribution on a shape corrupted by some noise, the sublevel-sets then carry the information about the shape itself. The distance to an empirical measure can easily be computed pointwise in the case of a point cloud (by averaging the squared distances to the k nearest neighbors). Unfortunately, the sublevel-sets remain hard to compute or approximate, and a challenge is to design efficient algorithms to compute or approximate them in high dimensions. Such algorithms would naturally find applications in topological inference in the presence of significant noise and outliers, but also in other less obvious contexts such as stable clustering. The current bottleneck stems from the fact that there exists no equivalent of a union of balls nor of Čech complexes for the distance to a measure. Our first goal will be to define such equivalents that will allow to infer the homology of the underlying shape or more generally the topological persistence of the distance to measure functions.

Remarkably, the measure-theoretic approach adopted to introduce and study the distance to a measure is well suited to take the statistical nature of data into account and to develop inference models that combine statistical and geometric methods. This new framework opens new promising research directions. It has already given rise to fruitful applications in density estimation, providing topological guarantees on the level sets of the density estimates [Biau et al., 2011], and in geometric deconvolution allowing the recovery of

topological features of shapes sampled with a huge amount of (known) noise [Caillerie et al., 2011].

Clustering with a geometric prior. Clustering may be viewed as the most basic homology inference problem, since it consists in inferring the connected components in the data set. Typical methods for clustering data sets in high dimensions, *e.g.,* spectral clustering [Shi and Malik, 1997], work well under three specific assumptions. First, the clusters should be sufficiently connected, for example, the second eigenvalue of their graph Laplacian should be large enough. Second, they should be well separated, that is, the interpoint distances between different clusters should be large enough on average. Third, the clusters should be balanced enough. We intend to develop methods that would take advantage of situations where the clusters have additional properties, such as being nearly convex or smooth, which seems reasonable in the large number of cases where manifold learning techniques apply. A promising strategy would be to use geometric regularity measures stemming from the geometric sampling theory recently introduced in my group [Chazal et al., 2009b]. One challenge is to design these regularity measures in such a way that they are both easy to compute and amenable to classical clustering strategies. We expect the resulting clustering schemes to outperform classical spectral clustering when the data exhibit some form of geometric regularity. Byproducts of this effort could also lead to efficient algorithms for assessing the degree of geometric regularity present in real data.

Topological signatures for shapes. Using topological quantities deduced from measurements to design signatures for shapes is a relatively new idea. The bottom line of the approach is the following: given a finite sampling of the shape, build some filtered simplicial complex on top of the point cloud, and use the topological structure of this filtration (encoded as a planar diagram called a *persistence diagram*) as a signature for the point cloud [Chazal et al., 2009a, Skraba et al., 2010]. This construction is well-suited to finite metric spaces, and the obtained signatures are known to be stable under small perturbations of the spaces in the Gromov-Hausdorff distance. As in many applications data only come with a measure of similarity between the pairs of data points (that does not satisfy the triangle equality), a major remaining challenge is to extend the construction to *general spaces endowed with a similarity measure*, and to prove the stability of its topological structure with respect to perturbations of the space in some Gromov-Hausdorff-like distance.

A more fundamental question is how much information about a shape can be recovered from descriptors. As discussed in the state of the art, deriving tight upper bounds on shape distances based on shape descriptors is still widely open. We intend to tackle the problem as follows. Thanks to the virtually infinite variety of filtrations that can be built on top of a shape, it is easy to enrich the pool of signatures used for that shape. We can thus restrict the possibility of false positives in the shape comparison process. It is then a question of how large a family of filtrations should be to guarantee that different shapes lead to different signatures. From an algorithmic perspective, the problem boils down to identifying small samplings of this family of filtrations that can be used as proxies for a better (if not perfect) assessment of the similarity between two shapes. As we intend to consider our signatures on large sets of shapes, it will also be important to design algorithms to efficiently compute these signatures and to compare them.

Finally, the difficulty of matching two shapes is intimately tied to matching a shape to itself — shapes with many natural self maps (symmetries) can be difficult to match because of the ambiguities symmetries create (reflected in duplicate descriptors, etc.). It may be interesting to define the analog of a *condition number* for a shape, which would capture the intrinsic difficulty of characterizing or matching against that shape.

Focus Area 4 : A software platform for geometry understanding in higher dimensions

We intend to develop an open source software platform of highest quality for geometric understanding in high dimensions. The goal is not to provide code tailored to the numerous potential applications but rather to provide the central data structures and algorithms that underly any application in geometry understanding in higher dimensions.

The development of such a platform will serve to benchmark and optimize new algorithmic solutions resulting from our theoretical work. Such development will necessitate a whole line of research on software architecture and interface design, heuristics and fine-tuning optimization, robustness and arithmetics issues, and visualization. We aim at providing a full programming environment following the very recipes that made up the success story of the CGAL library.

An important aspect of the work is also to educate students and young researchers and to offer them a multidisciplinary education covering the mathematical, algorithmic and implementational aspects of the subject.

Data structures. The software platform will provide tools to extract simplicial complexes from point clouds. We will focus our attention on simplicial complexes that are relevant for geometric understanding, in the first place Rips, tangential Delaunay and witness complexes. The platform will provide efficient implementations of data structures to encode those complexes. In particular, the compact data structures developed in Focus Area 1 will be fully tested and benchmarked against other encodings.

Basic algorithms. The platform will offer a well chosen set of basic algorithms at the heart of geometry understanding in higher dimensions. It will include in particular algorithms to *sample*, *mesh* and *reconstruct smooth and stratified manifolds*, compute the *persistent homology* of simplicial complex filtrations, *cluster data*, compute *signatures of shapes* (see Focus Areas 2 and 3). The platform will also provide some visualization tools.

Parallel implementations as well as out-of-core versions of some critical algorithms will make it possible to handle huge data sets in high dimensions.

Applications. The results of the Gudhi project will be used and benchmarked against real data. We have selected two main applications, one in structural biology and the other in astrophysics, and intend to collaborate with renowned experts in those fields.

We will establish strong collaborations with Prof. E. Coutsias (University of New Mexico) and F. Cazals, two leaders in computational chemistry and computational structural biology. A protein or any macromolecule with n atoms is a flexible system with $3n$ degrees of freedom. Dynamic molecular simulations or Monte Carlo simulations are able to sample the *energy landscape of a molecular system*. A better understanding of the sampled conformation space of the molecule is likely to foster our understanding of transitions between stable states, whence of molecular functions. We also intend to study the persistent homology of the level sets of those molecular energy landscapes. A key question is to derive collective coordinates describing large amplitude - low frequency deformations, a process reminiscent of dimensionality reduction. We aim at identifying collective coordinates incorporating the homological constraints – a topic which has not yet been addressed [Wales, 2003].

We will also collaborate with R. van de Weijgaert, an astrophysicist from Groningen University, aiming at understanding the phase space dynamics of *cosmic structure formation*. In principle n-body simulations are $6D$ datasets: each particle has 3 spatial coordinates and 3 velocity components. In fact, lately the interest in the phase-space structure of cosmological datasets has increased significantly. The idea is that matter of evolving cosmic structure defines an intricate strongly non-uniform distribution in phase-space, marked by matter streams which define ever more complicated manifolds. At any location in space, there may be far more than 1 stream, and this can be far better understood by looking at 6-dimensional phase space. This is of high interest in the search for dark matter. Regretfully, the current observational cosmological datasets – the galaxy surveys – do only provide $3D$ information. However, a very interesting higher dimensional dataset will be produced by the European Gaia satellite in the coming years: after its schedule launch of 2013, this satellite will provide an unprecedented accurate map of the structure and distribution of stars in our Galaxy and will shed new light on its formation history.

In addition to these two major applications, we foresee other applications in neurosciences, medical imaging and dynamical systems. We will leap at opportunities arising from our new results and tools.

Datasets. To benchmark our algorithms, it is important to have access to datasets that are representative of realistic tasks. Some datasets exist in the machine learning community (e.g. http://archive.ics.uci.edu/ml/). We will also use datasets from structural biology and astrophysics. We will create a public repository for new datasets, including synthetic datasets with controlled parameters that will be found useful for benchmarking against our data structures and algorithms. Providing easy access to such data sets is an important service to the research community of geometry understanding in higher dimensions and will help build a standard for benchmarking algorithms in the field.

Infrastructure and diffusion. The development infrastructure will include a web site, comprehensive documentation, an svn repository, and regular execution of test suites on several hardware architectures. We will set up an editorial board in charge of the review of new packages submitted for inclusion in the platform. One strategic goal is to give to the platform sufficient momentum that it will attract interest of a large community of researchers in computational geometry and topology, and various applicative fields, and serve as a standard for experimental research. This is important for the development of new tools as well as for the perennial maintenance of the software. In addition to a vigorous effort towards code diffusion, we will set up appropriate communication channels towards developers inside and outside the Gudhi project, including mailing lists, developer workshops, publications and talks at conferences in applied domains.

2 Resources

Research environment. The PI and his research team, Geometrica, are part of INRIA, the French National Institute for Research in Mathematics and Informatics. Part of the group, including the PI, is located in the INRIA research center in Sophia Antipolis (38 research groups) and part of the group is hosted by the INRIA research center in Saclay in Paris's area (26 research teams). Two members of the research center in Sophia Antipolis are members of the French Academy of Science, two have received an ERC advanced grant and two have received an ERC junior grant. Because of its partial location in Saclay, the group benefits from tight collaborations with the univerity of Orsay, the Ecole Normale Supérieure and

the Ecole Polytechnique. In particular, 4 members of the group teach at these prestigious institutions. Geometrica currently includes 10 permanent researchers, 6 postdoctoral researchers, 11 Ph.D. students, and 1 research engineer.

The team members. The team members who are directly involved in this proposal are the PI (J-D. Boissonnat) and 2 permanent researchers of the Geometrica team : Frédéric Chazal and Mariette Yvinec. J-D. Boissonnat will conduct and supervise the research activities of Gudhi and will be involved in the project for at least 70% of his time. Frédéric Chazal and Mariette Yvinec will each devote 20% of their time to this project to co-supervise with the PI the research and implementation work of the students, postdocs and engineers to be engaged in this project. Frédéric Chazal, located in Saclay, is a world expert in geometric inference and computational topology. Mariette Yvinec, located in Sophia Antipolis, is a member of the CGAL Editorial Board. She will bring her unique expertise in geometric computing. Other members of Geometrica, not financially supported by this project, will also contribute to the ideas and expertise of Gudhi: D. Cohen-Steiner, O. Devillers, M. Glisse and S. Oudot.

External collaborators. We will establish strong collaborations with Prof. Coutsias (University of New Mexico) and F. Cazals (INRIA), two leaders in computational chemistry and structural biology, to work on the understanding of energy landscapes of macromolecules. We will also collaborate with R. van Weijgaert, an astrophysicist from Groningen University, to work on the understanding of phase space dynamics of cosmic structure formation (see Focus Area 4).

Available resources. Our European ICT Fet-Open project Computational Geometric Learning (CG-Learning) will still be active until november 2013 and will ensure a smooth start of Gudhi (total amount of 606,000 Euros for 3 years). The ANR Présage (182,487 euros) will provide additional resources for the activities of Geometrica on probabilistic techniques in geometry.

The Geometrica team is equipped with numerous PCs and has access to a large PC cluster owned by INRIA Sophia Antipolis.

Requested resources: personnal costs.
– 70% of PI's salary over 5 years with a 70% commitment of his time
– 20% of 2 PI's close collaborators over 5 years with a 20% commitment of their time
– 2 full-time post-doctoral researcher during 2.5 years covering the 5 years of the project
– 2 full-time research engineers during 2.5 years covering the 5 years of the project
– 1 full-time Ph.D. student during years 1, 2, 3
– 2 full-time Ph.D. students during years 2, 3, 4
– 4 men-months of invited professors in years 1-5.

The funded 3 PhD students will have their research devoted to the fundamental aspects of the 3 first Focus Areas A1-A3 of this proposal. An additional Ph.D. student, not funded by the Gudhi project, will work on the applications, co-advised with F. Cazals (Focus Area 4). There will be a lot of synergy between the works of the Ph.D. students, in particular in relation with the development of the platform. The funded research engineers will help stabilize the software modules, as well as for the construction of new datasets to be made available to the scientific community.

References

[Attali et al., 2011] Attali, D., Lieutier, A., and Salinas, D. (2011). Efficient data structure for representing and simplifying simplicial complexes in high dimensions. In *Proc. of the 27th Symp. on Computational geometry*, SoCG '11.

[Bendich et al., 2007] Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Morozov, D. (2007). Inferring local homology from sampled stratified spaces. In *Proc. of the IEEE Symp. on Foundations of Computer Science*, pages 536–546.

[Biau et al., 2011] Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., and Rodrguez, C. (2011). A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237.

[Boissonnat et al., 2012] Boissonnat, J.-D., Dyer, R., and Ghosh, A. (2012). Stability of Delaunay-type structures for manifolds. In *Proc. of the 27th Symp. on Computational Geometry*, SoCG '12, pages 229–238.

[Boissonnat et al., 2011] Boissonnat, J.-D., Dyer, R., Ghosh, A., and Oudot, S. (2011). Equating the witness and restricted Delaunay complexes. Technical Report CGL-TR-24. http://cgl.uni-jena.de/Publications/WebHome.

[Boissonnat and Ghosh, 2010a] Boissonnat, J.-D. and Ghosh, A. (2010a). Manifold reconstruction using tangential Delaunay complexes. In *Proc. 26th Annual Symposium on Computational Geometry*. INRIA Report 7712.

[Boissonnat and Ghosh, 2010b] Boissonnat, J.-D. and Ghosh, A. (2010b). Triangulating smooth submanifolds with light scaffolding. *Mathematics in Computer Science*, 4(4):431–461.

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

[Boissonnat et al., 2008] Boissonnat, J.-D., Wormser, C., and Yvinec, M. (2008). Locally uniform anisotropic meshing. In *Proc. of the 24th Symp. on Computational Geometry*, pages pages 270–277. INRIA Report 7712.

[Brisson, 1989] Brisson, E. (1989). Representing geometric structures in d dimensions: topology and order. In *Proc. of the 5th Symp. on Computational geometry*, SCG '89, pages 218–227, New York, NY, USA. ACM.

[Bronstein et al., 2006] Bronstein, A. M., Bronstein, M. M., and Kimmel, R. (2006). Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proc. National Academy of Sciences (PNAS)*, 103:1168–1172.

[Caillerie et al., 2011] Caillerie, C., Chazal, F., Dedecker, J., and Michel, B. (2011). Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423.

[Carlsson and de Silva, 2004] Carlsson, G. and de Silva, V. (2004). Topological estimation using witness complexes. In *Symposium on Point-Based Graphics*.

[Carlsson et al., 2008] Carlsson, G., Ishkhanov, T., and de Silva, V. (2008). On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1).

[Chazal et al., 2009a] Chazal, F., Cohen-Steiner, D., Guibas, L. J., Mémoli, F., and Oudot, S. Y. (2009a). Gromov-Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, pages 1393–1403.

[Chazal et al., 2009b] Chazal, F., Cohen-Steiner, D., and Lieutier, A. (2009b). A sampling theory for compact sets in euclidean space. *Discrete Comput. Geom.*, 41(3):461–479.

[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.

[Chazal and Oudot, 2008] Chazal, F. and Oudot, S. Y. (2008). Towards Persistence-Based Reconstruction in Euclidean Spaces. In *Proc. ACM Symp. on Computational Geometry*, pages 232–241.

[de Silva and Ghrist, 2007] de Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, (7):339–358.

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

[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.

[Ferragina et al., 2005] Ferragina, P., Luccio, F., Manzini, G., and Muthukrishnan, S. (2005). Structuring labeled trees for optimal succinctness, and beyond. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 184–196, Washington, DC, USA. IEEE Computer Society.

[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.

[Hatcher, 2002] Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press.

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

[Kaczynski et al., 2003] Kaczynski, T., Mischaikow, K., and Mrozek, M. (2003). *Computational Homology*. Springer.

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

[Leibon and Letscher, 2000] Leibon, G. and Letscher, D. (2000). Delaunay triangulations and Voronoi diagrams for Riemannian manifolds. In *Symp. Comp. Geom.*, pages 341–349.

[Lienhardt, 1994] Lienhardt, P. (1994). N-dimensional generalized combinatorial maps and cellular quasi-manifolds. *Int. J. Comput. Geometry Appl.*, 4(3):275–324.

[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).

[Mémoli and Sapiro, 2005] Mémoli, F. and Sapiro, G. (2005). A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5:313–346.

[Munro and Raman, 2002] Munro, J. I. and Raman, V. (2002). Succinct representation of balanced parentheses and static trees. *SIAM J. Comput.*, 31:762–776.

[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.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326.

[Shi and Malik, 1997] Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905.

[Skraba et al., 2010] Skraba, P., Ovsjanikov, M., Chazal, F., and Guibas, L. (2010). Persistence-based segmentation of deformable shapes. In *CVPR Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*.

[Tenenbaum et al., 2000] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

[Wales, 2003] Wales, D. (2003). *Energy Landscapes*. Cambridge University Press.

[Zomorodian, 2009] Zomorodian, A. (2009). *Topology for Computing*. Cambridge University Press.