Rend. Sem. Mat. Univ. Politec. Torino
Vol. 70, 4 (2012), 323 – 345

D. Mumford, S. Chiossi and S. Salamon

HOW DOES A MATHEMATICIAN THINK ABOUT SPACE AND SHAPE?*

Abstract. The lecture on which this article was based was addressed to non-mathematicians and therefore the ideas behind recent work in medical imaging are explained in elementary terms in much of this paper. We consider especially the infinite dimensional spaces of curves and shapes in two or three dimensional Euclidean space, acted on by the group of diffeomorphisms. Various types of distinguished paths can be defined in these spaces, and they correspond to warping one shape to another in special ways. Along the way, we explain the relevance of concepts from differential geometry such as tangent vectors, Riemannian metrics, lengths, geodesics and curvature.

1. Introduction

There has been a huge explosion in biomedical imaging, in the analysis of spatial and temporal data, and in the effective visualization of this data. Facial recognition software is one example. Recently, new more sophisticated mathematical tools have entered into the game. It is important to have some basic understanding of these tools, to know what to expect of them, and to enquire what prospects lie ahead.

In order to explain the tools, we start by recalling Euclid and the Pythagorean theorem which are the basis of measurements in space. This leads us in Section 4 to the concept of a metric space. Many examples of this can be given if we define suitably the 'distance' between two objects which could be either conventional points, or shapes in the plane or in space. The scene is then set for a more rigorous analysis of the deformation or warping of one shape into another.

Geometrical techniques, based on an infinitesimal Pythagorean theorem that were originally introduced by Gauß in finite dimensions, can be applied in this very general setting. As Riemann proposed 150 years ago, we can apply these methods in particular to spaces of shapes and groups of diffeomorphisms of the ambient space in which the shapes lie. The concept of minimizing "energy" leads one to define appropriate geodesics in these spaces.

We survey a few aspects of the general theory, and attempt to explain what the new tools may be good for. To this aim, we conclude with some illustrations from medicine and cartography.

This article is based on the first author's lecture upon receiving an honorary doctorate at the University of Turin, a lecture intended primarily for a non-mathematical audience. In this sense, it is unlike the more conventional contributions to this journal, and only provides a glimpse of his extensive work in this area. Those who wish to discover more should refer to the material that is freely available on the website [11].

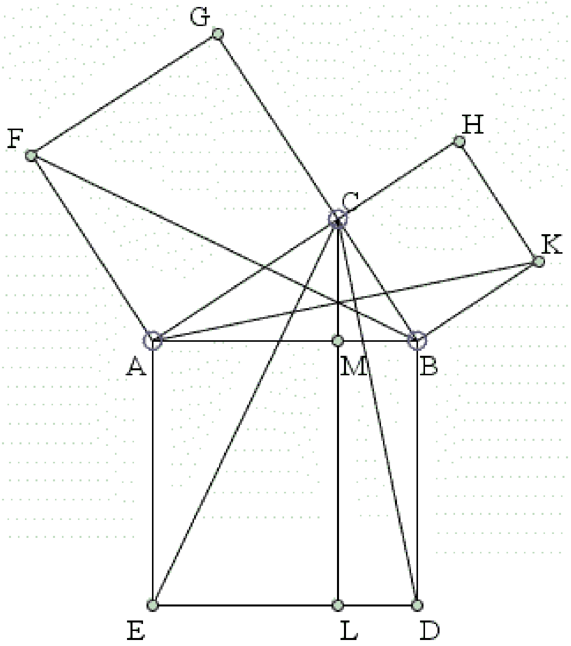
*This is an expanded version of the first author's Lectio Magistralis given in Turin on 3 May 2012

Acknowledgements. The text of David Mumford's original lecture was edited by Simon Chiossi and Simon Salamon, who took the liberty of expanding selected sections, and incorporating material that the first author had presented a few days later at the Brain and Spine Institute (ICM) in Paris.

2. Euclid's towering influence

The birth of the mathematical analysis of space is traditionally placed in ancient Greece during the first millennium BC, although it is now known that very similar ideas had been developed earlier by the Babylonians, the Indians and the Chinese.

All of us will have heard of the theorem of "Pythagoras" but may not realize that it is not just an obscure fact about right triangles, but the key property that governs distances in space. The figure below illustrates Euclid's first proof, though not the most transparent proof – see [2] for 100 proofs:



Proving Pythagoras' theorem

It was Euclid (circa 300 BC) who first axiomatized geometry by prescribing the rules that space has to obey. His opus *The Elements*, which covered plane geometry, the theory of numbers, irrationals, and solid geometry, was to become "the" standard work, and provided the foundations on which the pillars of geometry have been built over the subsequent twenty-three centuries.

To understand how widely Euclid's work had spread, we can read this excerpt from Matteo Ricci. Ricci was a brilliant Italian Jesuit priest, who spent the second half of his life trying convert the Chinese, reaching Beijing in 1598. This quote is from the Preface of his translation of the first five books of *The Elements* into Chinese, dating from 1605[1]:

> *At my university, I above all got to know one name: Euclid. He brought mathematical theory to great perfection, and he towers high above his predecessors. He has opened new avenues and has enlightened the path for later generations. ...In the books he wrote ...there is not a single thing that can be doubted. Especially his Elements is very exact and can rightly be called a standard work. ...Everything is contained in his theory and there is nothing that does not follow from it.*

3. Absolute and relative space

The space of Euclid's geometry seemed so natural and intuitive, that its absoluteness (from the Latin: free, independent of anything) was considered obvious, its existence beyond the shadow of a doubt. This had profound repercussions well beyond mathematics.

The German philosopher Immanuel Kant (1724–1804) argued in the *Critique of Pure Reason* (1781) that Euclidean space was an *a priori* category. Roughly put, the notion of space is not discovered by the human mind, but is instead an unavoidable – in-built/*a priori* – systematic framework for organizing our experiences, and any denial regarding its essence makes no sense. We quote from [6, ch. I, part I, §3]:

> *Geometry is a science which determines the properties of space synthetically, and yet* a priori *...It must be originally intuition ...found in the mind* a priori*, that is, before any perception of objects.*

For instance, it should be "obvious" that space is 3-dimensional, for physical objects can be measured in three different directions (length, width, depth). Another fact we are taught and convinced of in school is that the sum of the angles of a triangle equals two right angles. However, in many rigorous treatments of plane geometry, this so-called fact is derived from a list of (apparently) more basic axioms including Euclid's parallel postulate. Playfair's 1795 version of the latter states that given a line and a point not on it, at most one line parallel to the given line can be drawn through the point in the plane.

For centuries, the status of the parallel postulate remained constantly in doubt: was it really needed as an extra postulate? An affirmative answer came in the early 1800's, and enabled mathematicians to throw off the philosopher's chains with the advent of the aptly-called "non-Euclidean" geometries. The cautious German Gauß

[1] His translation was part of his campaign to persuade the Chinese Mandarins that the West had something to offer them and thus open the door to their possible conversion.

(1777–1855), the swashbuckling Austro-Hungarian Bolyai (1802–1860), and the pioneering Russian Lobachevsky (1792–1856) all hit on the idea that Pythagoras' theorem *does not necessarily have to hold* in the real world, indeed quite the opposite.

At the dawn of the 20th century, Einstein's special theory of relativity had established, building on Maxwell's theory of electromagnetism, that the laws of physics needed to be understood in a 4-dimensional space that incorporates time as the fourth dimension. The mathematics had been supplied by Hermann Minkowski, who famously wrote:

> *The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.* ([10, p. 75])

Minkowski's space was however still "flat", and close in spirit to Euclid's. However, after the acceptance of Einstein's general theory of relativity, physics had adopted the abstract theory of non-Euclidean geometries as models of the universe. The total eclipse observations of 1919 had confirmed Einstein's discovery that space is warped by mass and energy, so that Pythagoras' theorem, although approximately true, *does not hold exactly*. The universe we live in has a more complex geometry than that envisioned by Euclid. Nowadays, this non-Euclidean correction is even incorporated into GPS navigation systems.

Einstein's field equations are expressed in the language of Riemann, whose curvature tensor adapts the concept of the Gaußian curvature of a 2-dimensional surface to n dimensions. Although Einstein's theory was expressed with $n = 4$ so as to provide a curved version of Minkowski space, more speculative theories of supergravity use very similar equations with $n = 11$.

4. A more abstract perspective

Let us move to the modern view of space. *The key point, in the present understanding and also in this article, is that any collection of things at all can be called the "points" of a "space"*. If geometry is ultimately, and etymologically, about measuring objects and shapes, the best way to endow such a space with geometry is to define the *distance* $d(P,Q)$ between any pair of points P and Q.

A few restrictions on the numbers $d(P,Q)$ are necessary for things to work. The necessary rules are incorporated in the definition of a *metric space*, a concept that dates back to Fréchet [4]. One would want the distance between any two points to be non-negative:

$$d(P,Q) \geqslant 0.$$

Moreover, it should be zero precisely when the two points coincide:

$$d(P,Q) = 0 \text{ if only if } P = Q.$$

Another important property is that measuring the distance going from P to Q or backwards should not make any difference:

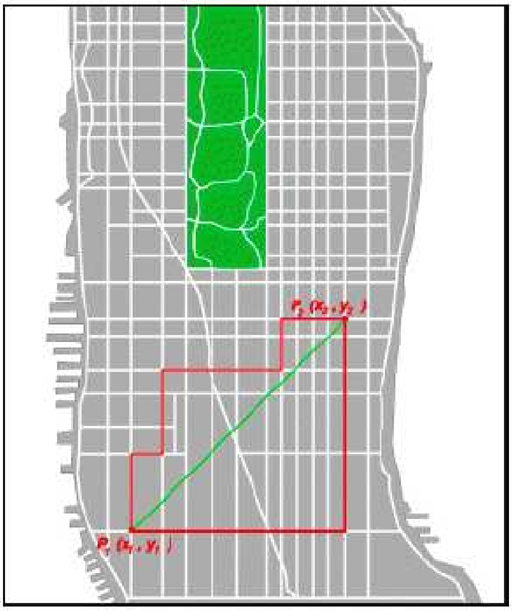
$$d(P,Q) = d(Q,P).$$

This means that d is a “symmetric function” of the two points P,Q. Finally, we want a triangle with vertices P,Q,R to satisfy the property

$$d(P,R) + d(R,Q) \geqslant d(P,Q). \tag{1}$$

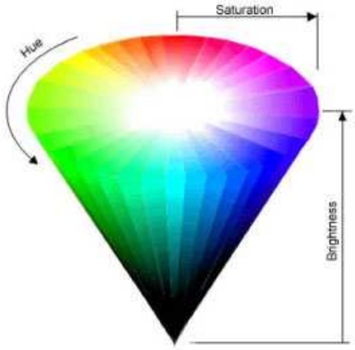
Thus no side of the triangle cannot be longer than the sum of the other two sides.

Introducing such rules enables one to consider unusual definitions for the distance $d(P,Q)$ between two points P,Q, ones that can be completely at odds with the idea of a straight line from P to Q. Defining d to be the *walking distance* along the streets of Manhattan, rather than distance as the crow flies, is such a case. In fact, this distance which is illustrated in the next figure, is used by *Google maps* to calculate shortest paths, either by car or by foot. The diagonal line is how a crow gets from P_1 to P_2 but humans must take one of many zig-zag paths, two of which are shown in the figure. We can then *define* $d(P_1,P_2)$ to be the length of any such zig-zag path instead of the length of the diagonal path.:



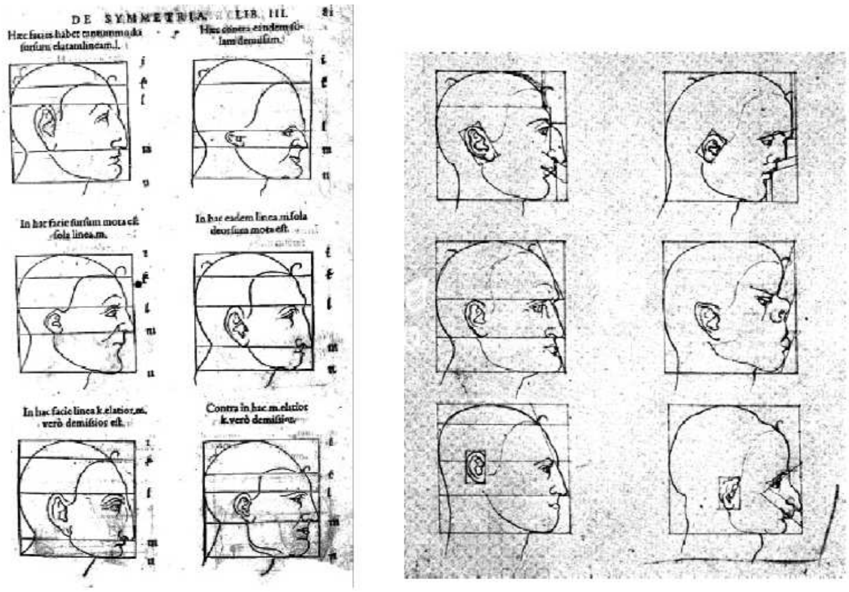
Walking in Manhattan

Another advantage of this approach is that one can take unconventional objects to be “points”, for example take (psychophysical) colors. In this case, our space is the set of colors that humans perceive, and we can take the distance d between two colors to be that given by the so-called psychophysical metric, an estimate of how different they are perceived to be by humans with normal color vision. This is based on 3 coordinates: hue, brightness and saturation, which place each color as a point on a circular cone (see, e.g., [5] from where the figure is taken). In this simple model though, in which our space of colors has been represented as a subset of ordinary space, d has reverted to the familiar Euclidean distance.



The cone of colors

Making a leap of faith, we can be more radical and decide to consider the totality of human faces as points varying in a space of faces, as Dürer did:


DE SYMMETRIA LIB. III.
Hæc facies habet tantummodo sursum elatam lineam.l.
Hæc contra eandem solam demissam.
In hac facie sursum mota est sola linea.m.
In hac eadem linea.m.sola deorsum mota est.
In hac facie linea k.elatior,m. verò demissior est.
Contra in hac m.elatior k.verò demissior.

Albrecht Dürer: *Vier Bücher von menschlicher Proportion*, 1528

5. Warping and comparison of shapes

There is a long tradition in the natural sciences of classifying animals by their shapes. The famous book *On Growth and Form* [14] by the mathematical biologist D'Arcy Thompson (1860–1948), introduced a major new idea: that of comparing shapes of animals by computing a *warping* of the plane or space that brings one animal into registration with the other animal. Thompson illustrated this idea by warping pictures into one another, for example various mammals' skulls, or fish, using simple mathematical functions:

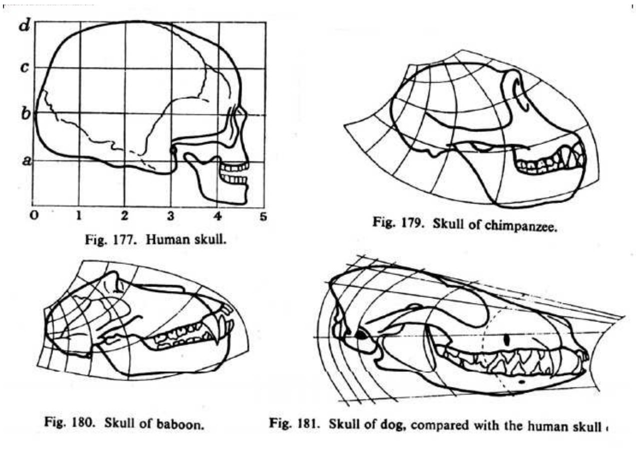


Fig. 177. Human skull.

Fig. 179. Skull of chimpanzee.

Fig. 180. Skull of baboon.

Fig. 181. Skull of dog, compared with the human skull

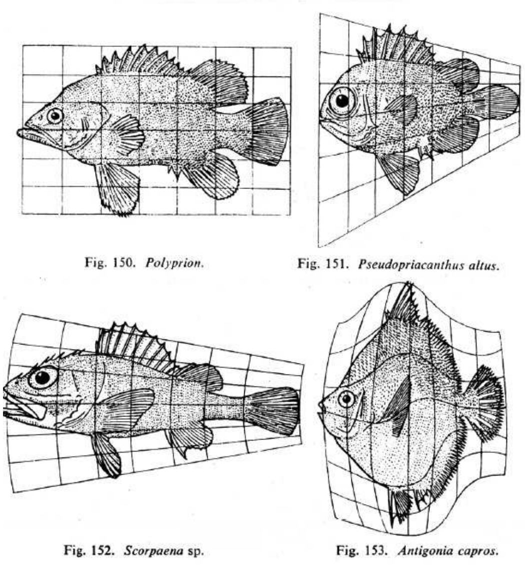


Fig. 150. *Polyprion.*

Fig. 151. *Pseudopriacanthus altus.*

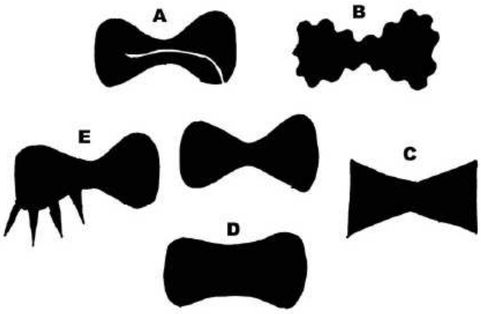
Fig. 152. *Scorpaena* sp.

Fig. 153. *Antigonia capros.*

D'Arcy Thompson: *On Growth and Form*, 1917

However, Thompson did not introduce a numerical measure for the difference between the shapes of two animals. The first to attempt this was the statistician David Kendall, followed by Kanti Mardia and many others seeking to adapt statistical tests to data sets of shapes. Their approach was to describe the shape of objects of interest by choosing sets of significant ‘landmark points’ on each shape and defining distances between two such sets (see, e.g. [8, 7, 1]).

But what makes two shapes seem similar is not so simple. There are many different ways to thinking about similarity. Consider the following six two-dimensional shapes (from a paper of B. Kimia):



The central shape is similar in various respects to all five around it, but which is it closest to and which is farthest from it? Emphasizing different features of the shapes leads to different orderings of the distance from the five shapes A-E and the central reference shape:

- looking at the area of overlap, the distances satisfy $A < B,C < D,E$;
- looking at points in each shape farthest away from any point in the central shape and vice versa (‘Hausdorff distance’) plus the same for points outside the shapes leads to: $B < C,D < A$;
- looking at similarity of the slope of the boundary: $D < B,C < A,E$;
- looking at corners or similarity of curvature: $D < A,B < C,E$;
- if you are willing to discard some bad bits, E is identical to the middle.

In the first case, the distance arises from a so-called L^1 norm, which measures a difference of areas. In the second case, it is an L^∞ norm that considers only extreme distances. The third and fourth cases combine these concepts with first and second derivatives respectively. We begin to see a collection of metrics analogous to the choices for function spaces.

On the other hand, it is also clear that D is distinguished by being more of a “single” object; the others all consist of a left part and a right part joined together. But this is a more subjective observation that is harder to quantify mathematically.

6. Ingredients from differential geometry

Following the ideas of Bernhard Riemann (1826–66), to analyse the space of shapes mathematically, one needs coordinates.

The idea of cartography is central to this theory. A *manifold* M is a geometric object which has local charts, that is subsets where a flat set of coordinates is given, just like on the page of an atlas. (There might not be global charts because the whole manifold may be curved and bend back on itself, like the surface of a sphere or the semi-circular canals in the ear.) The manifold is said to be *smooth* if comparisons of different charts can be carried out using differentiable functions of their respective local coordinates.

The expression *atlas* is used by mathematicians to denote any set of charts that completely cover the manifold:

The pages of an atlas give systems of local coordinates

Our goal is to represent shapes as points of some manifold M. Various technical features are necessary to measure the “distance” between two different shapes, represented as points P,Q of M. These are:

Local coordinates. Having chosen a chart surrounding a given point P, we will have coordinates $x_1,\ldots,x_n$ with which to describe P, where n is the dimension of the manifold. Thus, one can associate to P a set of n numbers:

$$P \longmapsto (x_1(P),x_2(P),\ldots,x_n(P)).$$

Tangent spaces. One can associate to each point P of M its tangent space T_PM, which in coordinates is the vector space generated by the instantaneous displacements $\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n}$ parallel to directions defined by the local coordinates. More explicitly, consider a curve $\gamma : [a,b] \to M$ in which we regard $\gamma(t)$ as the point occupied by a particle moving on M at time t. Then the tangent vector to γ at $\gamma(t)$ is the vector

$$\gamma'(t) = \left(\frac{d(x_1(\gamma(t)))}{dt}, \ldots, \frac{d(x_n(\gamma(t)))}{dt}\right).$$

It is an element of T_PM.

In addition to these objects, which are implicit in the definition of smooth manifold, we also need the concept of

Riemannian metric. This is a way of measuring size in T_PM by resurrecting Pythagoras and requiring that small distances, in the limit, should satisfy his theorem. This means that in suitably chosen local coordinates, an infinitesimal distance ds is equal to the sum of squares of the displacements in each coordinate direction:

$$ds^2 = \sum_i^n dx_i^2, \tag{2}$$

In any set of coordinates, this distance becomes a quadratic expression:

$$ds^2 = \sum_{i,j=1}^n g_{ij}(P) dx_i dx_j, \tag{3}$$

where $g_{ij}(P)$ is a positive-definite symmetric matrix. Only in a flat locally Euclidean setting, may we find coordinates such that $(g_{ij}) = (\delta_{ij})$ equals the identity matrix at every point. As P varies on M, the g_{ij} become functions, and for any value of t, we may define the *norm* of the tangent vector $\gamma'(t)$ by

$$\|\gamma'(t)\| = \sqrt{\sum_{i,j=1}^n g_{ij}(\gamma(t)) \frac{d(x_i(\gamma(t)))}{dt} \frac{d(x_j(\gamma(t)))}{dt}}.$$

Provided t continues to represent time, this quantity is simply the instantaneous *speed* of the particle moving along the curve.

Length of paths. When speed is constant, distance equals speed times time. When the speed varies, one computes the *arc length* of the curve γ by integrating the speed:

$$\ell(\gamma) = \int_a^b \|\gamma'(t)\| \, dt. \tag{4}$$

This equation is a basic starting point for any geometric analysis of shape. Much of this theory is due to Gauß, at least when M is 2-dimensional.

EXAMPLE 1. Take $M = \mathbb{R}^2$ to be the Euclidean plane. The curve $\gamma(t) = (t^2, t^3)$ is a so-called semi-cubical parabola, and has a cusp shape "$\prec$". Its velocity is

$$\gamma'(t) = (2t,\ 3t^2),$$

and this vanishes when $t = 0$, indicating that the vertex $(0,0)$ of the cusp is a singular point in which there is no well-defined tangent direction (it can point neither left nor right!). The curve's arc length, measured from $P = (0,0)$ to $\gamma(1) = (1,1)$ is

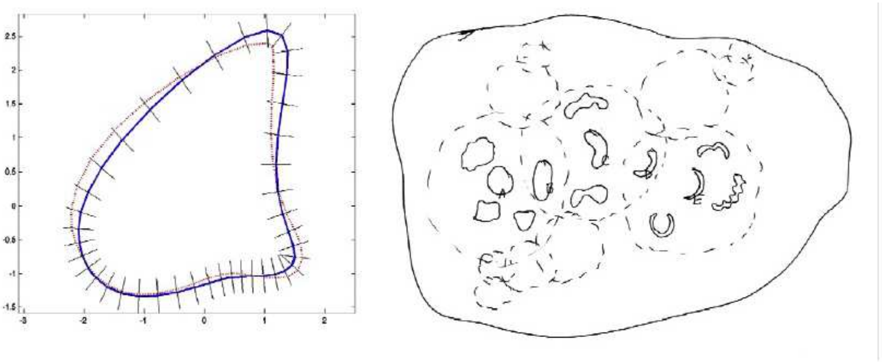
$$\ell(\gamma) = \int_0^1 t\sqrt{4+9t^2}\, dt = \frac{13\sqrt{13}-8}{27} \approx 1.44.$$

For many other curves (parametrized even more simply), it is impossible to find such an explicit expression for the arc length function.

7. An infinite-dimensional manifold of closed curves

In this section, we shall consider the set of all simple closed curves in the plane. Let us fix one such curve C_0, parametrized by a periodic differentiable function $\gamma : \mathbb{R} \to \mathbb{R}^2$ such that $\gamma(s+1) = \gamma(s)$ for all s. We assume $\gamma'(s) \neq 0$ for all s.

The set $\mathcal{C}$ of all such smooth closed curves C forms a manifold. We shall also refer to C as a "shape". One can compare nearby shapes (as in the left figure) by placing small perpendicular "hairs" on the basic shape, and describing all nearby shapes by how far one has to move along each hair. This is shown on the left in the figure below. The whole blob on the right is a cartoon of the entire space of plane curves, and the dotted lines represent coordinate charts that arise as small deformations of a central shape – forming a single coordinate chart. Passing across charts, we see a sequence (A,B,C,D,E) of points along a curve in the space of shapes connecting a circle to a banana to a new moon:



Let's describe the coordinate charts explicitly. If $f : \mathbb{R} \to \mathbb{R}$ is a periodic real-valued function with period 1, then we can define another curve by

$$\gamma_f(s) = \gamma(s) + f(s)\mathbf{n}(s), \tag{5}$$

where $\mathbf{n}(s)$ is the unit normal to C_0 (one of the hairs in the figure for each fixed value of s). Let C_f denote the image of γ_f, and we assume $\sup \|f\|$ is small enough so that this is a new smooth curve. Note that although we could make s the arc length parametrization of C_0, s would not generally be the arc length parametrization of C_f.

Then

$$U_\gamma = \{f \mid \gamma_f \text{ is smooth}\}$$

is a subset of the vector space of all maps f, and the assignment

$$U_\gamma \longrightarrow \mathcal{C}, \qquad f \longmapsto C_f,$$

is a chart. Therefore,

$$\mathcal{C} = \bigcup_\gamma U_\gamma$$

describes an atlas (albeit with an uncountable infinity of pages) for the space of plane curves or shapes.

A curve β on $\mathcal{C}$ itself (better if we call β a *path* to avoid confusion) is a mapping from an interval $[a,b]$ to $\mathcal{C}$, so the image of each number u with $a \leqslant u \leqslant b$ is itself a closed curve. This means that β describes a means to pass continuously from the plane curve labeled by the parameter a to that labeled by b. It is therefore a "warping" of one shape into the other!

The tangent space $T_C\mathcal{C}$ at the point corresponding to the curve C is the vector space of all normal vector fields $f(s)\mathbf{n}(s)$ along C. A Riemannian metric is given by assigning an inner product to this vector space, or equivalently a norm $\|f\,\mathbf{n}\|$ to each such vector field. One possibility is to define

$$\|f\,\mathbf{n}\|_\gamma^2 = \int_C f(s)^2\,\|\gamma'(s)\|ds.$$

(The derivative of γ is needed to make the norm independent of the parametrization.) Another possibility is to use an expression involving the derivatives of $f(s)$.

Riemann put together the theory of n-dimensional manifolds in his Habilitation lecture in 1854 [13]. It is not so well known that here he also imagined the infinite-dimensional version:

> *There are however manifolds in which the fixing of position requires not a finite number but either an infinite series or a continuous manifold of determinations of quantity. Such manifolds are constituted for example by . . . the possible shapes of a figure in space, etc.*

This infinite-dimensional perspective will be central for us.

8. Shapes under diffeomorphism

A warping of the sort considered in Section 5 is an example of a *diffeomorphism*, meaning a bijective smooth map from the Euclidean space $\mathbb{R}^n$ to itself, with a non-zero Jacobian (the determinant of the matrix of partial derivatives) at each point. In the present treatment, we can restrict attention to $n = 2$ or $n = 3$, so that we are dealing with diffeomorphisms of the plane (as in Section 5), and of space, respectively.

Since all human bodies of a fixed gender are warpings of a "textbook body" of that gender (modulo amputations, tumors and so on), we obtain the mathematicians' new age slogan:

> *We are all diffeomorphisms of one another!*

The punchline is that diffeomorphisms (of $\mathbb{R}^n$) form an infinite-dimensional manifold. If we fix a diffeomorphism $\mathbf{x} \mapsto \phi(\mathbf{x})$, nearby diffeomorphisms have the form

$$\mathbf{x} \longmapsto \phi(\mathbf{x}) + \mathbf{v}(\mathbf{x}), \tag{6}$$

where $\mathbf{x} \mapsto \mathbf{v}(\mathbf{x})$ is any vector field sufficiently small that (6) is bijective and so on. (Note that, unlike ϕ, the map $\mathbf{v}$ does not need to be bijective.)

The set of all diffeomorphisms forms a group (more precisely, a *Lie group* G), because diffeomorphisms can be "multiplied together" by composing them as maps. The group of diffeomorphisms acts transitively on the space of shapes $\mathcal{C}$, meaning that a diffeomorphism $\mathbf{x} \mapsto \phi(\mathbf{x})$ can always be found mapping any closed curve in $\mathcal{C}$ to any other. Having fixed a shape or "template curve" C_0, there will be a subgroup H of G consisting of those diffeomorphisms that leave C_0 fixed. An arbitrary curve C in $\mathcal{C}$ can now be identified with the coset

$$\phi H = \{\phi\psi \mid \psi \in H\},$$

where ϕ is *any* diffeomorphism that maps C_0 to C. Passing to the coset (itself a subset of G) eliminates the ambiguity in the choice of ϕ.

By way of conclusion, the set of shapes $\mathcal{C}$ can now be identified with the set, denoted G/H, of cosets of H in G. This object is a geometric *quotient* of the group of diffeomorphisms.

9. Minimizing energy

We first consider the elastic strain energy. If ϕ is a deformation of a body B, the deformation gradient $D\phi$ is the matrix of partial derivatives of ϕ. The entries of

$$G(\phi) = (D\phi)^t (D\phi),$$

a symmetric matrix, may be regarded as the metric g_{ij} on B defined by pulling back the Euclidean metric on $\phi(B)$. Elasticity theory is based on *strain energy*

$$E(\phi) = \int_B e(G(\phi), \mathbf{x})\, d\mathbf{x},$$

in which the *energy density* $e(G(\phi), \mathbf{x})$ measures the energy needed to deform the bit of the body near $\mathbf{x}$ by the given amount.

The total energy E is minimized when ϕ is a Euclidean motion. If boundary conditions are imposed as well, then ϕ would have to be the identity diffeomorphism I that leaves every point $\mathbf{x}$ fixed. This approach is physically natural and allows to compute deformations from the template to all other shapes of least energy.

An alternative ("liquid") approach is to use minimum path length instead of the strain to define E. The latter then represents a kind of kinetic energy and, as opposed to the elasticity approach, particles can be said to 'forget where they came from'.

Consider a vector field

$$\mathbf{v}\colon\ \mathbf{x} \longmapsto \mathbf{v}(\mathbf{x}, t)$$

depending on time t. Here, $\mathbf{x}$ is a point of $\mathbb{R}^3$, or more generally a manifold M. We can "integrate" this field in order to find a path

$$\gamma\colon\ [a,b] \longrightarrow G, \qquad t \longmapsto \phi_t, \tag{7}$$

where G is the group of diffeomorphisms of $\mathbb{R}^3$ or M. This means that $\phi_0 = I$ is the identity diffeomorpism and for each t, ϕ_t is a diffeomorphism $\mathbf{x} \mapsto \phi_t(\mathbf{x})$ satisfying

$$\frac{\partial \phi_t(\mathbf{x})}{\partial t} = \mathbf{v}(\phi_t(\mathbf{x}), t). \tag{8}$$

Mimicking (4), we can define the length of the path (7) by

$$\ell(\gamma) = \int_a^b \|\mathbf{v}(\cdot, t)\| \, dt,$$

where $\| \cdot \|$ is an appropriate norm defined on vector fields. This gives a metric on the group of diffeomorphisms

$$d(\psi, \psi') = \min\{\ell(\gamma)\} \text{ taken over all paths (7) such that } \psi = \phi_a,\ \psi' = \phi_b.$$

It is easy to see that $d(\psi, \psi') = d(\psi \circ \phi, \psi' \circ \phi)$, i.e. d is a 'right-invariant' metric. We then set $E(\phi) = d(I, \phi)$.

One can equally well work within the space $\mathcal{C}$ of curves, by defining the length of a path in $\mathcal{C}$ by

$$\ell(\gamma) = \int_a^b \|f(\cdot, t)\|_{\gamma(t)} \, dt,$$

where (for each t) $f(\cdot, t)$ is the function defined on the curve $\gamma(t)$, defining the tangent to the path γ at t as in definition (5). One can then define the distance

$$d(C, C') = \min\{\ell(\gamma)\} \text{ taken over all } C_t \text{ such that } C = C_a,\ C' = C_b.$$

10. Geodesics and curvature

Navigating the Earth, a shortest path is seldom a straight line: one must weave to avoid hills and valleys, as illustrated overleaf.

More generally, whenever a manifold has an inner product defined on its tangent spaces, one can seek the shortest path from one point P to another Q, and this is called a *geodesic*. Formally, the latter is defined as a solution of the equation

$$\delta \left(\int_0^1 \left\| \frac{d\mathbf{x}(t)}{dt} \right\| dt \right) = 0, \tag{9}$$

where δ represents the derivative relative to an arbitrary variation of the path $t \mapsto \mathbf{x}(t)$.

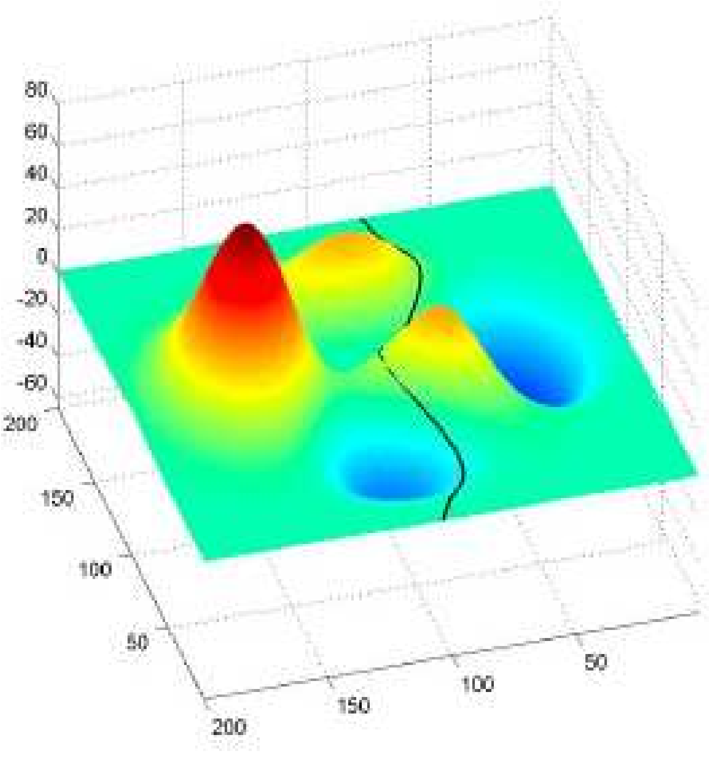
A standard technique from the calculus of variations allows one to convert this into a differential equation that characterizes a geodesic $t \mapsto \mathbf{x}(t)$. This is the *Euler-Lagrange equation* associated to (9), and on a manifold with local coordinates $x^1, x^2, \ldots, x^n$, it turns out that it can be written like this:

$$\frac{d^2 x_i}{dt^2} + \sum_{j,k=1}^{n} \Gamma^i_{jk}(\mathbf{x}) \frac{dx_j}{dt} \frac{dx_k}{dt} = 0. \tag{10}$$

Here $\Gamma^i_{ij}(\mathbf{x})$ are the so-called *Christoffel symbols* that are computed from the functions $g_{ij}(\mathbf{x})$ and their derivatives by the formula

$$\Gamma^i_{jk} = \frac{1}{2}\sum_{m=1}^{n} g^{im}\left(\frac{\partial g_{mj}}{\partial x_k} + \frac{\partial g_{mk}}{\partial x_j} - \frac{\partial g_{jk}}{\partial x_m}\right), \tag{11}$$

where (g^{ij}) is the inverse matrix of (g_{ij}). In (10), each $x_i = x_i(t)$ is a function of t as are $\Gamma^i_{jk}(\mathbf{x}(t))$.



A path on a surface minimizing distance

EXAMPLE 2. In the Euclidean setting, the functions $g_{ij} = \delta_{ij}$ are constant, and so $\Gamma^i_{jk} = 0$ for all i, j, k (see (3) and (11)). The geodesic equations (10) simplify to $d^2x_i/dt^2 = 0$ for all i. Their solutions are given by

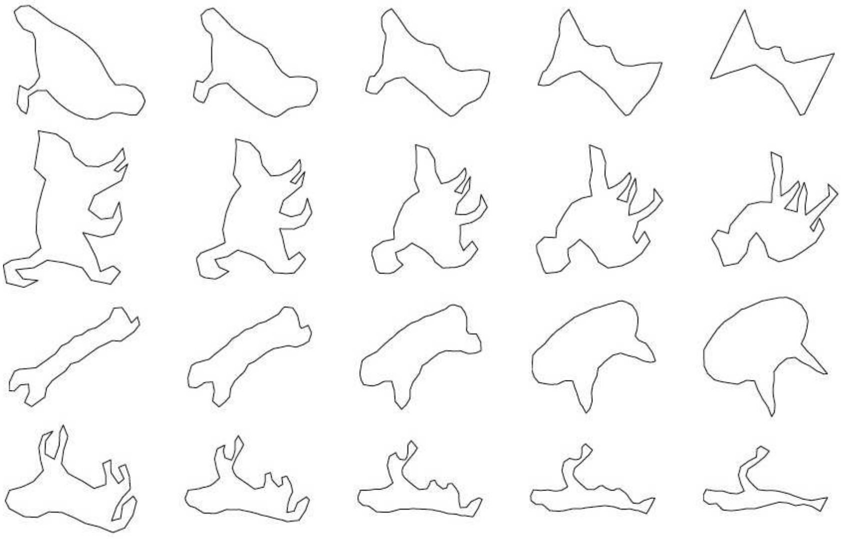
$$x_i(t) = a_it + b_i,$$

where a_i, b_i are constants, and represent straight lines in $\mathbb{R}^n$. For arbitrary metrics g_{ij}, geodesics are not straight lines, a concept that anyway makes no sense in an arbitrary manifold.

One can define an analogue of the geodesic equation on an infinite-dimensional space of diffeomorphisms, by replacing the velocity $d\mathbf{x}/dt = (dx_1/dt, \ldots, dx_n/dt)$ by

$$\mathbf{v}(\mathbf{x},t) = \frac{\partial\phi}{\partial t}\big(\phi^{-1}(\mathbf{x},t), t\big), \tag{12}$$

which is merely a restatement of (8).



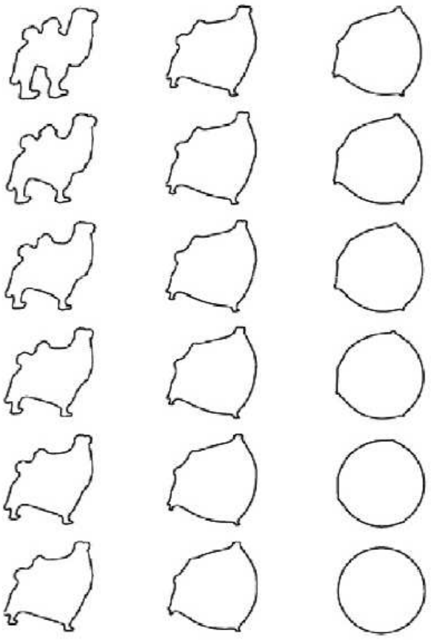
Four examples of geodesics in the space of 2-dimensional shapes (Laurent Younes)

Equation (10) is replaced by an equation of the form:

$$\frac{\partial \mathbf{v}}{\partial t} = \text{a quadratic expression in } \mathbf{v} \text{ and its derivatives, integrals}$$

The main ingredient needed to define geodesics is the Riemannian metric, or assigning a norm on infinitesimal motions. However, there are many choices for a metric. As we already mentioned, for plane curves, one could use $\int f^2\,ds$, but then the metric would collapse, as was shown in [9].

An option to avoid the collapsing phenomenon is to strengthen the metric by penalizing the curvature of the shape, using derivatives of f (at least one is needed). The next figure shows a geodesic derived using this approach with, in a sense, 3/2 derivatives of f (follow the changes from top to bottom then left to right):



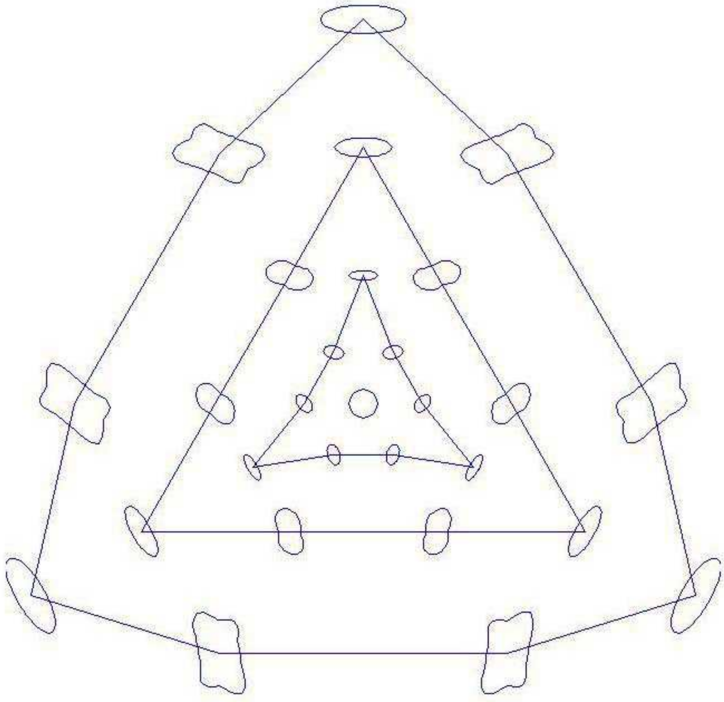
A geodesic with the 'Weil-Petersen' distance measure (Matt Feiszli)

Another approach is to measure the size of normal vector fields to a curve by taking their extension to a vector field to all of $\mathbb{R}^n$ with minimum Sobolev norm of some order. The geodesics for this metric will be projections of special geodesics on the space Diff($\mathbb{R}^n$) of diffeomorphisms of Euclidean space. The relevant norm turns out to be given by

$$\|\mathbf{v}\|^2 = \int_{\mathbb{R}^n} (L\mathbf{v}, \mathbf{v})\, d\mathrm{vol},$$

where L is a suitable self-adjoint, positive-definite differential operator. There are lots more possibilities.

The classical sign of negative curvature is that triangles have angle sums less than π and their sides tend to curve in towards the center. Positive curvature implies the opposite: angle sums greater than π and sides bulging out away from the center. In a suitable metric on plane curves, curvature is negative for small or jagged curves and positive for large smooth ones, leading to the next figure (for details, see [9]). This shows three triangles whose vertices are ellipses oriented at 60 degrees to each other and connected by geodesics in this metric. For the small ellipses, the sum of the angles is about 102° and the geodesic sides nearly go back to the circle in the middle. For the large ellipses, on the other hand, the angle sum is about 207° and along the geodesics, two protrusions grow while two shrink which can be seen as bulging away from a circle. The triangle in the middle represents the critical Euclidean case in which angles add up to exactly 180°.



Geodesic triangles in the space of plane curves for the Michor-Mumford metric in [9]

The concept of curvature is crucial to understanding the previous figure. In a 2-dimensional manifold M, it is determined by a single function on M, the Gaußian curvature K. On a convex surface such as a sphere or ellipsoid, K is everywhere positive, but on a pseudo-sphere or hyperboloid, K is everywhere negative. In n dimensions, curvature is much more sophisticated and is characterized by a complicated tensor introduced by Riemann. It is best understood as a function on 2-planes in the tangent spaces T_PM to M called *sectional curvature*. Describing 2-planes by 2-forms in $\Lambda^2 T_PM$, it is given at each $P \in M$ by a quadratic function on these 2-forms.

In the presence of positive curvature, geodesics tend not to be unique if one extends them beyond what's called the cut locus (the antipodal point to the starting point in the case of a sphere). On the other hand, if the sectional curvature is everywhere negative, geodesics between any two points are unique but the space is in some sense vaster, volume of spheres growing exponentially with their radius, and it's easy to "get lost" because going around a city block, you don't come back to where you started.

Curvature is a big obstacle to doing analysis and statistics on nonlinear spaces. At the same time it reflects the nonlinear nature of these big spaces: shapes and diffeomorphisms do not live in vector spaces but have their own inherent geometry, which up to now is only partially understood.

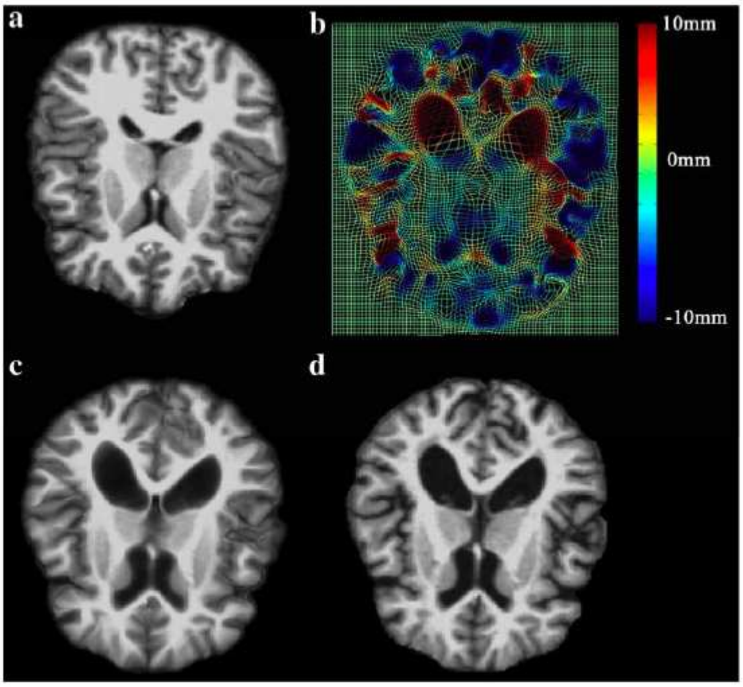
11. A shape breakthrough in medicine

The 3-dimensional version of the theory has striking applications to medicine, in particular anatomy and diagnostics.

First of all note that all vertebrates are (more or less) diffeomorphic, and all healthy male and all healthy female humans are really diffeomorphic with only moderate distortion. Would it then be possible to form an ideal 3D computer model of a human? And is it possible, for each MRI or other scan of each patient, to find an optimal diffeomorphism of the scanned region with the ideal model?

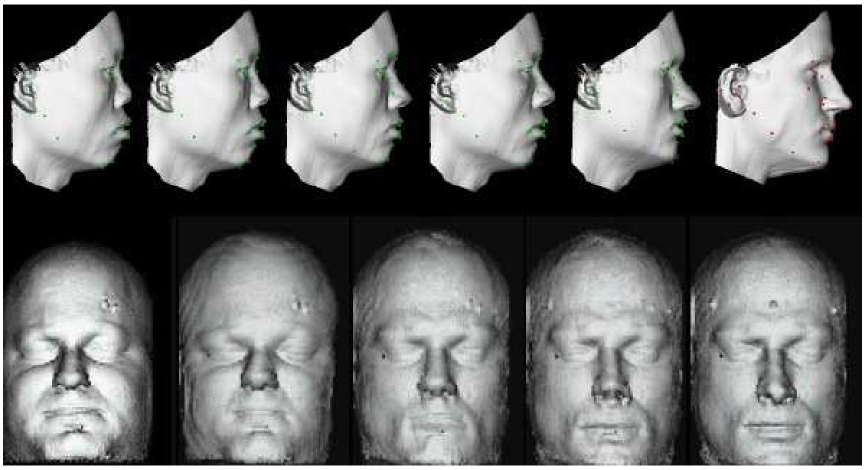
To see what we want to match up in general, bear in mind that the body has organs, bones, vessels with boundaries, well-defined points (e.g. traditional points on the skull, with curious names such as nasion, menton or gonion), muscle fibre and nerve fibre tracts (giving line fields). The diffeomorphism should be constrained to respect these boundaries, curves, points, orientations, maybe even some densities.

Ulf Grenander and Michael Miller introduced a method for computing the optimal warping of the ambient Euclidean space based on modeling the warping as resulting from a flow in all of space and the differential geometric idea of geodesics. This has been used extensively by Miller's group to register pairs of MRI scans and other medical images as well as to estimate average normal body configurations.



Whole brain diffeomorphic metric mapping (Du, Younes, Qiu [3])

The application of these ideas is apparent in the brain scans on the preceding page. Figure **a** shows a healthy brain, figure **d** a senile brain with shrunken white matter and enlarged ventricles, while **c** is a warp of **a** matching **d**.



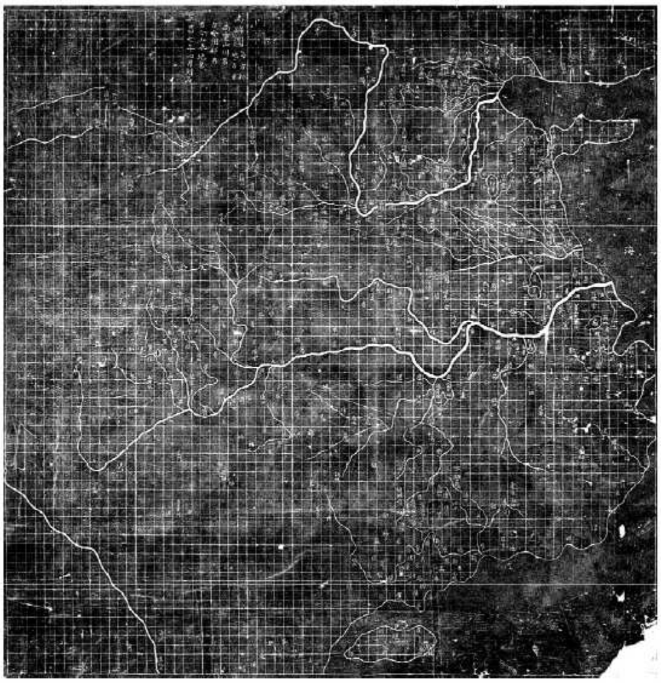
Geodesics between faces: shortest paths in the space of diffeomorphisms carrying one face to the other (Vaillant, Trouve, Younes)

12. Chinese cartography

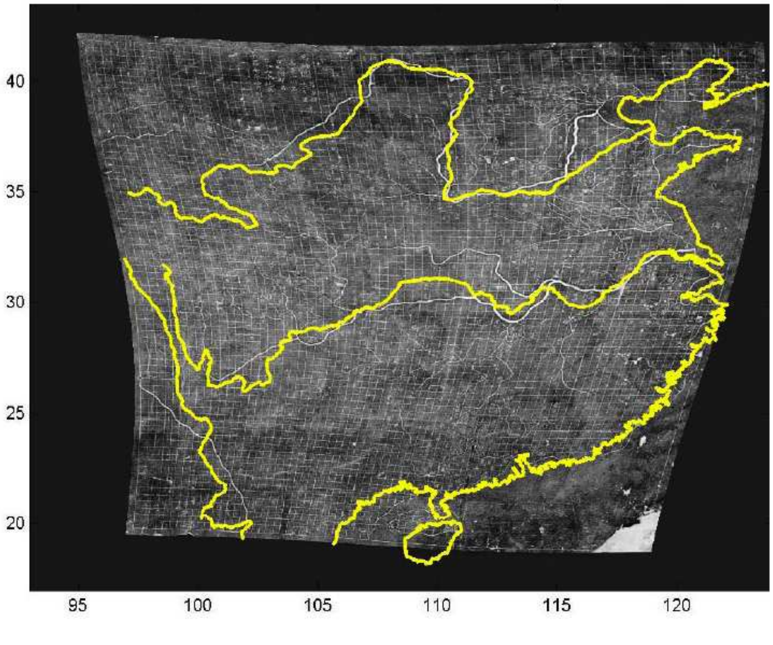
In the year 1137 AD, a remarkable map of China was carved in stone in Xian covering about one square meter. It was called the *Yu Ji Tu*, the map (*Tu*) of the trails (*Ji*) of the legendary Xia Dynasty tribal leader *Yu*. It's likely that it is based on maps drawn some fifty years earlier in the Northern Song dynasty, such as the now lost monumental map of Shen Kua that covered all of China with 43 panels drawn at a scale of $1 : 900{,}000$ (the *Yu Ji Tu* has a scale of about $1 : 3{,}000{,}000$).

Apart from the characters marking locations of prominent cities, mountains and lakes, what is astonishing is that the map is overlaid on a precise NS/EW Cartesian grid, and all the major features of the coastline, the Yellow and Yangtze rivers are immediately recognizable and their shapes correspond closely to those in modern maps.

It was evidently based on the work of meticulous surveyors who achieved a remarkable level of accuracy surpassing even the high point of Ptolemy's world map. The grid has constant spacing, consisting of 71 vertical lines and 74 horizontal lines which are described as all being 100 li apart (a li being the standard Chinese unit for measuring long distances). On the other hand, the southern limit (modern Hanoi) of the map lies at about latitude 18° and the northern limit (in Inner Mongolia) at about 43°. This creates a simple contradiction: meridian lines must converge as you move north, their separation being proportional to the cosine of the latitude. And $\cos(43^\circ)/\cos(18^\circ)$ is about 0.77. Thus true meridian lines in a map covering this area are far from being a constant distance apart.



The *Yu Ji Tu* from a rubbing in the US Library of Congress



Modern map in yellow, underlying *Yu Ji Tu* map in white

To analyze what compromises were made by the makers of this map, in [12] we applied the same warping ideas used for medical images to "geo-reference" the whole area of the *Yu Ji Tu* to latitude and longitude coordinates. The results can be seen on the preceding page (work with Alexander Akin). Note the angle formed by the north-south gridlines in the bottom image.

Chinese culture has an ancient tradition of thinking of the Earth as a large flat square, and the *Yu Ji Tu* was drawn in accordance with this long established cartographic method (going back to Pei Xiu's "Cartesian" coordinates, 267 AD). There is no indication that the draftsmen were aware of the curvature of the Earth. Its east-west lines are quite accurate: these could be found using the ancient Chinese method of measuring the angle of the sun at its solsticial meridian crossing. But its north-south lines are heavily distorted to keep the desired 100 li spacing. It's hard not to believe that the empirical data would have revealed that the north-south lines were not true. Did some Chinese know the Earth was round and small enough to cause such errors? Cosmogonies with a round Earth had been discussed but were always in the realm of speculation and, when quantified, were unrealistically huge. Along with the absence of geometric models of the Earth, there were also no geometric models of the sun, moon, planets and stars. In particular this made it difficult to understand how to incorporate lunar parallax into their predictions of solar eclipses, a matter of great consequence to the legitimacy of their emperors.

13. Outlook

What have we gained from this approach?

We can consider the set of all colors, all faces, all fish, all human cortices or all maps of some country as points of a space. In all but the first of these cases, the resulting space is infinite dimensional and not an easy mathematical entity to describe.

Nonetheless, we can introduce measures of distance on these spaces meant to quantify differences and find shortest paths warping one shape into another. This is accomplished by adapting geometrical techniques introduced over 150 years ago, but ones (such as the deep theory of geodesics in an infinite-dimensional setting) that are being studied in other branches of contemporary mathematical research.

All this brings the very intuitive but imprecise idea of shape (and yet shapes that are instantly recognizable) into a clear mathematical setting for further analysis and application.

References

[1] Barden, D., Carne, T. K., Kendall, D. G., and Le, H. *Shape and Shape Theory*. Wiley Series in Probability and Statistics. Wiley, 1999.

[2] Bogomolny, A., et al. Pythagorean theorem. http://www.cut-the-knot.org/pythagoras.

[3] Du, J., Younes, L., and Qiu, A. Whole brain diffeomorphic metric mapping via integration of sulcal and gyral curves, cortical surfaces, and images. *Neuroimage 56*, 1 (2011), 162–173.

[4] FRÉCHET, M. Sur quelques points du calcul fonctionnel. *Palermo Rend. 22* (1906), 1–74.

[5] JEWETT, T. Color tutorial. http://www.tomjewett.com/colors/hsb.html.

[6] KANT, I. *Critique of Pure Reason*. Available at http://www.gutenberg.org/ebooks/4280.

[7] KENDALL, D. G. A survey of the statistical theory of shape. *Statistical Sci. 4*, 2 (1989), 73–185.

[8] MARDIA, K. V., MCDONNELL, P., AND LINNEY, A. D. Penalized image averaging and discrimination with facial and fishery applications. *J. Appl. Stat. 33*, 3 (2006), 339–371.

[9] MICHOR, P., AND MUMFORD, D. Riemannian geometries on spaces of plane curves. *J. Eur. Math. Soc. 8*, 1 (2006), 1–48.

[10] MINKOWSKI, H. Space and time. In *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*, H. A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl, Eds. Dover, New York, 1952, pp. 75–91.

[11] MUMFORD, D. Archive for reprints, notes and talks. http://www.dam.brown.edu/people/mumford.

[12] MUMFORD, D. The Song dynasty "Yu Ji Tu" and the curvature of the earth. See [11].

[13] RIEMANN, B. On the hypotheses which lie at the foundation of geometry. *Nature 8* (1873), 14–17, 36–37. translated from German by W. K. Clifford, freely available at http://www.emis.de/classics.

[14] THOMPSON, D. W. *On Growth and Form*. Dover, 1992. A reprint of the 1942 2nd edition (1st ed. 1917).

AMS Subject Classification: 54C56; 53C22, 58D05

David MUMFORD
Division of Applied Mathematics, Brown University
182 George Street, Providence, RI 02912, USA
e-mail: David_Mumford@brown.edu

Simon CHIOSSI
Dipartimento di Scienze Matematiche, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino (TO), ITALIA
e-mail: simon.chiossi@polito.it

Simon SALAMON
Department of Mathematics, King's College London
Strand, London WC2R 2LS, UK
e-mail: simon.salamon@kcl.ac.uk

Lavoro pervenuto in redazione il 04.05.2012 e, in forma definitiva, il 19.11.2012