WHAT IS . . .

Data Mining

Mauro Maggioni

Data collected from a variety of sources has been accumulating rapidly. Many fields of science have gone from being data-starved to being data-rich and needing to learn how to cope with large data sets. The rising tide of data also directly affects our daily lives, in which computers surrounding us use data-crunching algorithms to help us in tasks ranging from finding the quickest route to our destination considering current traffic conditions to automatically tagging our faces in pictures; from updating in near real time the prices of sale items to suggesting the next movie we might want to watch.

The general aim of *data mining* is to find useful and interpretable patterns in data. The term can encompass many diverse methods and therefore means different things to different people. Here we discuss some aspects of data mining potentially of interest to a broad audience of mathematicians.

Assume a sample data point x_i (e.g., a picture) may be cast in the form of a long vector of numbers (e.g., the pixel intensities in an image): we represent it as a point in $\mathbb{R}^D$. Two types of related goals exist. One is to detect patterns in this set of points, and the other is to predict a function on the data: given a training set $(x_i, f(x_i))_i$, we want to predict f at points outside the training set. In the case of text documents or webpages, we might want to automatically label each document as belonging to an area of research; in the case of pictures, we might want to recognize faces; when suggesting the next movie to watch given past ratings of movies by a viewer, f consists of ratings of unseen movies. Typically, x_i is noisy (e.g., noisy pixel values), and so is $f(x_i)$ (e.g., mislabeled samples in the training set).

Of course mathematicians have long concerned themselves with high-dimensional problems. One example is studying solutions of PDEs as functions in infinite-dimensional function spaces and performing efficient computations by projecting the problem onto low-dimensional subspaces (via discretizations, finite elements, or operator compression) so that the reduced problem may be numerically solved on a computer. In the case of solutions of a PDE, the model for the data is specified: a lot of information about the PDE is known, and that information is exploited to predict the properties of the data and to construct low-dimensional projections. For the digital data discussed above, however, typically we have little information and poor models. We may start with crude models, measure their fitness to the data and predictive ability, and, those being not satisfactory, improve the models. This is one of the key processes in statistical modeling and data mining. It is not unlike what an applied mathematician does when modeling a complex physical system: he may start with simplifying assumptions to construct a "tractable" model, derive consequences of such a model (e.g., properties of the solutions) analytically and/or with simulations, and compare the results to the properties exhibited by the real-world physical system. New measurements and real-world simulations may be performed, and the fitness of the model reassessed and improved as needed for the next round of validation. While physics drives the modeling in applied mathematics, a new type of intuition, built on experiences in the world of high-dimensional data sets rather than in the world of physics, drives the intuition of the

Mauro Maggioni is assistant professor of mathematics and computer science at Duke University. His email address is mauro.maggioni@duke.edu.

DOI: http://dx.doi.org/10.1090/noti831

mathematician set to analyze high-dimensional data sets, where "tractable" models are geometric or statistical models with a small number of parameters.

One of the reasons for focusing on reducing the dimension is to enable computations, but a fundamental motivation is the so-called *curse of dimensionality*. One of its manifestations arises in the approximation of a 1-Lipschitz function on the unit cube, $f : [0,1]^D \to \mathbb{R}$ satisfying $|f(x) - f(y)| \le ||x - y||$ for $x, y \in [0,1]^D$. To achieve uniform error ϵ, given samples $(x_i, f(x_i))$, in general one needs at least one sample in each cube of side ϵ, for a total of ϵ^{-D} samples, which is too large even for, say, $\epsilon = 10^{-1}$ and $D = 100$ (a rather small dimension in applications). A common assumption is that either the samples x_i lie on a low-dimensional subset of $[0,1]^D$ and/or f is not simply Lipschitz but has a smoothness that is suitably large, depending on D (see references in [3]). Taking the former route, one assumes that the data lies on a low-dimensional subset in the high-dimensional ambient space, such as a low-dimensional hyperplane or unions thereof, or low-dimensional manifolds or rougher sets. Research problems require ideas from different areas of mathematics, including geometry, geometric measure theory, topology, and graph theory, with their tools for studying manifolds or rougher sets; probability and geometric functional analysis for studying random samples and measures in high dimensions; harmonic analysis and approximation theory, with their ideas of multiscale analysis and function approximation; and numerical analysis, because we need efficient algorithms to analyze real-world data.

As a concrete example, consider the following construction. Given n points $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ and $\epsilon > 0$, construct $W_{ij} = \exp(-\frac{||x_i - x_j||^2}{2\epsilon})$, $D_{ii} = \sum_j W_{ij}$, and the Laplacian matrix $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ on the weighted graph G with vertices $\{x_i\}$ and edges weighted by W. When x_i is sampled from a manifold $\mathcal{M}$ and n tends to infinity, L approximates (in a suitable sense) the Laplace-Beltrami operator on $\mathcal{M}$ [2], which is a completely intrinsic object. The random walk on G, with transition matrix $P = D^{-1}W$, approximates Brownian motion on $\mathcal{M}$. Consider, for a time $t > 0$, the so-called diffusion distance $d_t(x,y) := ||P^t(x,\cdot) - P^t(y,\cdot)||_{L^2(G)}$ (see [2]). This distance is particularly useful for capturing clusters/groupings in the data, which are regions of fast diffusion connected by bottlenecks that slow diffusion. Let $1 = \lambda_0 \ge \lambda_1 \ge \cdots$ be the eigenvalues of P and φ_i be the corresponding eigenvectors (φ_0, when G is a web graph, is related to Google's pagerank). Consider a diffusion map Φ_d^t that embeds the graph in Euclidean space, where $\Phi_d^t(x) := (\sqrt{\lambda_1^t}\varphi_1(x), \ldots, \sqrt{\lambda_d^t}\varphi_d(x))$, for some $t > 0$ [2]. One can show that the Euclidean distance between $\Phi_d^t(x)$ and $\Phi_d^t(y)$ approximates $d_t(x,y)$, the diffusion distance at time scale t between x and y on the graph G.

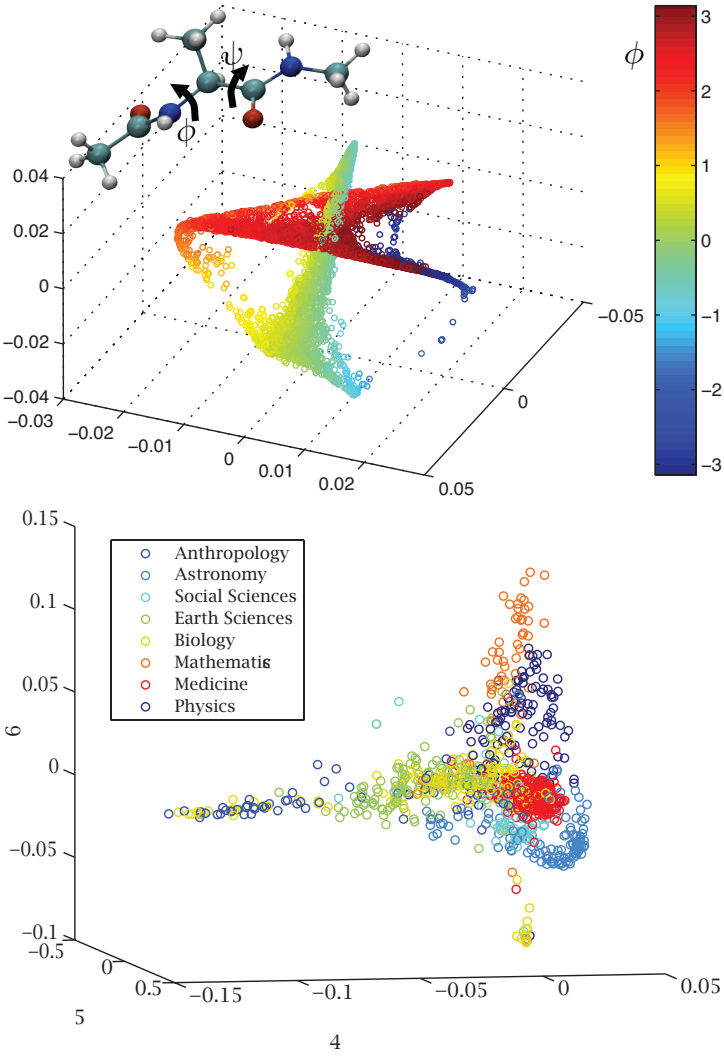
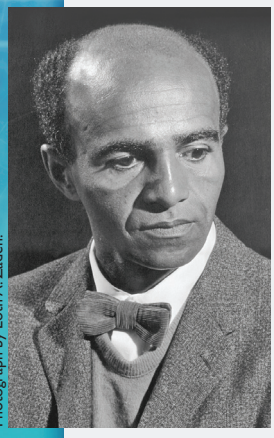


Figure 1. Top: Diffusion map embedding of the set of configurations of a small biomolecule (alanine dipeptide) from its 36-dimensional state space. The color is one of the dihedral angles φ, ψ of the molecule, known to be essential to the dynamics [4]. This is a physical system where (approximate) equations of motion are known, but their structure is too complicated and the state space too high-dimensional to be amenable to analysis. Bottom: Diffusion map of a data set consisting of 1161 *Science News* articles, each modeled by a 1153-dimensional vector of word frequencies, embedded in a low-dimensional space with diffusion maps, as described in the text and in [2].

David Blackwell

Memorial Conference

April 19–20, 2012
Howard University
Washington, DC



Photograph by Lotfi A. Zadeh.

Join us on April 19–20, 2012, at Howard University in Washington, DC for a special conference honoring David Blackwell (1919–2010), former President of the Institute of Mathematical Statistics and the first African-American to be inducted into the National Academy of Sciences.

This conference will bring together a diverse group of leading theoretical and applied statisticians and mathematicians to discuss advances in mathematics and statistics that are related to, and in many cases have grown out of, the work of David Blackwell. These include developments in dynamic programming, information theory, game theory, design of experiments, renewal theory, and other fields. Other speakers will discuss Blackwell's legacy for the community of African-American researchers in the mathematical sciences.

The conference is being organized by the Department of Mathematics at Howard University in collaboration with the University of California, Berkeley, Carnegie Mellon University, and the American Statistical Association. Funding is being provided by the National Science Foundation and the Army Research Office. To learn more about the program, invited speakers, and registration, visit: **https://sites.google.com/site/conferenceblackwell**.

In Figure 1 we apply this technique to two completely different data sets. The first one is a set of configurations of a small peptide, obtained by a molecular dynamics simulation: a point $x_i \in \mathbb{R}^{12\times 3}$ contains the coordinates in $\mathbb{R}^3$ of the 12 atoms in the alanine dipeptide molecule (represented as an inset in Figure 1). The forces between the atoms in the molecule constrain the trajectories to lie close to low-dimensional sets in the 36-dimensional state space. In Figure 1 we apply the construction above[1] and represent the diffusion map embedding of the configurations collected [4]. The second one is a set of text documents (articles from *Science News*), each represented as a $\mathbb{R}^{1153}$ vector whose kth coordinate is the frequency of the kth word in a 1153-word dictionary. The diffusion embedding in low dimensions reveals even lower-dimensional geometric structures, which turn out to be useful for understanding the dynamics of the peptide considered in the first data set and for automatically clustering documents by topic in the case of the second data set. Ideas from probability (random samples), harmonic analysis (Laplacian), and geometry (manifolds) come together in these types of constructions.

This is only the beginning of one of many research avenues explored in the last few years. Many other exciting opportunities exist, for example the study of stochastic dynamic networks, where a sample is a network and multiple samples are collected in time: quantifying and modeling change requires introducing sensible and robust metrics between graphs.

Further reading: [5, 3, 1] and the references therein.

References

1. *Science: Special issue: Dealing with data*, February 2011, pp. 639–806.
2. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, *Proc. Natl. Acad. Sci. USA* **102** (2005), no. 21, 7426–7431.
3. D. Donoho, *High-dimensional data analysis: The curses and blessings of dimensionality*, "Math Challenges of the 21st Century", AMS, 2000.
4. M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, Determination of reaction coordinates via locally scaled diffusion map, *J. Chem. Phys.* **134** (2011), 124116.
5. J. W. Tukey, The Future of Data Analysis, *Ann. Math. Statist.* **33**, Number 1 (1962), 1–67.

[1] *We use here a slightly different definition of the weight matrix W, which uses distances between molecular configurations up to rigid affine transformations, instead of Euclidean distances.*