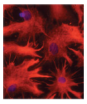
EXTENDED PDF FORMAT
SPONSORED BY

Travel Grants Available
Neuroscience 2014
Application Deadline: July 31st
>Apply Now
R&D systems
a bio-techne brand
www.rndsystems.com

Clustering by fast search and find of density peaks

Alex Rodriguez and Alessandro Laio

Science **344**, 1492 (2014);
DOI: 10.1126/science.1242072

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of June 26, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/344/6191/1492.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2014/06/25/344.6191.1492.DC1.html

This article **cites 14 articles**, 1 of which can be accessed free:
http://www.sciencemag.org/content/344/6191/1492.full.html#ref-list-1

This article appears in the following **subject collections:**
Computers, Mathematics
http://www.sciencemag.org/cgi/collection/comp_math

Downloaded from www.sciencemag.org on June 26, 2014

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2014 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

intrinsic and extrinsic contributions depends on the sample quality (such as the doping density and the amount of disorder). Studies of the dependence on temperature and on disorder are therefore required to better understand the doping density dependence of the VHE. Furthermore, a more accurate determination of σ_H that takes into account the fringe fields in our Hall bar device may be needed for a better quantitative comparison.

REFERENCES AND NOTES

1. A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, A. K. Geim, *Rev. Mod. Phys.* **81**, 109–162 (2009).
2. A. Rycerz, J. Tworzydlo, C. W. J. Beenakker, *Nat. Phys.* **3**, 172–175 (2007).
3. A. R. Akhmerov, C. W. J. Beenakker, *Phys. Rev. Lett.* **98**, 157003 (2007).
4. D. Xiao, W. Yao, Q. Niu, *Phys. Rev. Lett.* **99**, 236809 (2007).
5. W. Yao, D. Xiao, Q. Niu, *Phys. Rev. B* **77**, 235406 (2008).
6. D. Xiao, G.-B. Liu, W. Feng, X. Xu, W. Yao, *Phys. Rev. Lett.* **108**, 196802 (2012).
7. Y. J. Zhang, T. Oka, R. Suzuki, J. T. Ye, Y. Iwasa, *Science* **344**, 725–728 (2014).
8. K. F. Mak, C. Lee, J. Hone, J. Shan, T. F. Heinz, *Phys. Rev. Lett.* **105**, 136805 (2010).
9. A. Splendiani *et al.*, *Nano Lett.* **10**, 1271–1275 (2010).
10. T. Cao *et al.*, *Nat. Commun.* **3**, 887 (2012).
11. K. F. Mak, K. He, J. Shan, T. F. Heinz, *Nat. Nanotechnol.* **7**, 494–498 (2012).
12. G. Sallen *et al.*, *Phys. Rev. B* **86**, 081301 (2012).
13. S. Wu *et al.*, *Nat. Phys.* **9**, 149–153 (2013).
14. H. Zeng, J. Dai, W. Yao, D. Xiao, X. Cui, *Nat. Nanotechnol.* **7**, 490–493 (2012).
15. D. Xiao, M.-C. Chang, Q. Niu, *Rev. Mod. Phys.* **82**, 1959–2007 (2010).
16. X. Li, F. Zhang, Q. Niu, *Phys. Rev. Lett.* **110**, 066803 (2013).
17. S. Murakami, N. Nagaosa, S.-C. Zhang, *Science* **301**, 1348–1351 (2003).
18. J. Sinova *et al.*, *Phys. Rev. Lett.* **92**, 126603 (2004).
19. Y. K. Kato, R. C. Myers, A. C. Gossard, D. D. Awschalom, *Science* **306**, 1910–1913 (2004).
20. J. Wunderlich, B. Kaestner, J. Sinova, T. Jungwirth, *Phys. Rev. Lett.* **94**, 047204 (2005).
21. J. Wunderlich *et al.*, *Science* **330**, 1801–1804 (2010).
22. N. Nagaosa, J. Sinova, S. Onoda, A. H. MacDonald, N. P. Ong, *Rev. Mod. Phys.* **82**, 1539–1592 (2010).
23. Materials and methods are available as supplementary materials on *Science* Online.
24. The skew scattering contribution, which is important only for high-mobility devices (*22*), is neglected in MoS_2 devices with relatively low mobility.
25. T. Cheiwchanchamnangij, W. R. L. Lambrecht, *Phys. Rev. B* **85**, 205302 (2012).
26. B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis, *Nat. Nanotechnol.* **6**, 147–150 (2011).
27. L. J. van der Pauw, *Philips Techn. Rev.* **20**, 220–224 (1958).
28. B. W. H. Baugher, H. O. H. Churchill, Y. Yang, P. Jarillo-Herrero, *Nano Lett.* **13**, 4212–4216 (2013).
29. G. Kioseoglou *et al.*, *Appl. Phys. Lett.* **101**, 221907 (2012).
30. O. Lopez-Sanchez, D. Lembke, M. Kayci, A. Radenovic, A. Kis, *Nat. Nanotechnol.* **8**, 497–501 (2013).
31. Strictly speaking, a bilayer device with slightly broken inversion symmetry by the substrate and/or by unintentional doping could also produce a finite VHE, but these effects are expected to be much smaller as compared with the VHE in monolayer devices (*13*).
32. Here, our assumption that only the photoexcited electrons contribute to the Hall response is reasonable because the holes are much more vulnerable to traps than are the electrons, given our highly n-doped device. Unlike the electron side, the VHE and the SHE become equivalent on the hole side owing to the spin-valley coupled valence band.
33. H.-A. Engel, B. I. Halperin, E. I. Rashba, *Phys. Rev. Lett.* **95**, 166605 (2005).

ACKNOWLEDGMENTS

We thank D. C. Ralph for his insightful suggestions and J. W. Kevek for technical support. We also thank J. Shan for many fruitful discussions and Y. You for private communications regarding the optical data on monolayer MoS_2. This research was supported by the Kavli Institute at Cornell for Nanoscale Science and the Cornell Center for Materials Research [National Science Foundation (NSF) DMR-1120296]. Additional funding was provided by the Air Force Office of Scientific Research (FA9550-10-1-0410) and the Nano-Material Technology Development Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (2012M3A7B4049887). Device fabrication was performed at the Cornell NanoScale Science and Technology Facility, a member of the National Nanotechnology Infrastructure Network, which is supported by NSF (grant ECCS-0335765). K.L.M. acknowledges support from the NSF Integrative Graduate Education and Research Traineeship program (DGE-0654193) and the NSF Graduate Research Fellowship Program (DGE-1144153).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/344/6191/1489/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
References (*34*–*37*)

24 December 2013; accepted 23 May 2014
10.1126/science.1250140

MACHINE LEARNING

Clustering by fast search and find of density peaks

Alex Rodriguez and Alessandro Laio

Cluster analysis is aimed at classifying elements into categories on the basis of their similarity. Its applications range from astronomy to bioinformatics, bibliometrics, and pattern recognition. We propose an approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. This idea forms the basis of a clustering procedure in which the number of clusters arises intuitively, outliers are automatically spotted and excluded from the analysis, and clusters are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. We demonstrate the power of the algorithm on several test cases.

Clustering algorithms attempt to classify elements into categories, or clusters, on the basis of their similarity. Several different clustering strategies have been proposed (*1*), but no consensus has been reached even on the definition of a cluster. In K-means (*2*) and K-medoids (*3*) methods, clusters are groups of data characterized by a small distance to the cluster center. An objective function, typically the sum of the distance to a set of putative cluster centers, is optimized (*3*–*6*) until the best cluster centers candidates are found. However, because a data point is always assigned to the nearest center, these approaches are not able to detect nonspherical clusters (*7*). In distribution-based algorithms, one attempts to reproduce the observed realization of data points as a mix of predefined probability distribution functions (*8*); the accuracy of such methods depends on the capability of the trial probability to represent the data.

Clusters with an arbitrary shape are easily detected by approaches based on the local density of data points. In density-based spatial clustering of applications with noise (DBSCAN) (*9*), one chooses a density threshold, discards as noise the points in regions with densities lower than this threshold, and assigns to different clusters disconnected regions of high density. However, choosing an appropriate threshold can be nontrivial, a drawback not present in the mean-shift clustering method (*10*, *11*). There a cluster is defined as a set of points that converge to the same local maximum of the density distribution function. This method allows the finding of nonspherical clusters but works only for data defined by a set of coordinates and is computationally costly.

Here, we propose an alternative approach. Similar to the K-medoids method, it has its basis only in the distance between data points. Like DBSCAN and the mean-shift method, it is able to detect nonspherical clusters and to automatically find the correct number of clusters. The cluster centers are defined, as in the mean-shift method, as local maxima in the density of data points. However, unlike the mean-shift method, our procedure does not require embedding the data in a vector space and maximizing explicitly the density field for each data point.

The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. For each data point i, we compute two quantities: its local density ρ_i and its distance δ_i from points of higher density. Both these quantities depend only on the distances d_{ij} between data points, which are assumed to satisfy the triangular inequality. The local density ρ_i of data point i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad (1)$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is a cutoff distance. Basically, ρ_i is equal to the number of points that are closer than d_c to point i. The algorithm is sensitive only to the relative magnitude of ρ_i in different points, implying that, for large data sets, the results of the analysis are robust with respect to the choice of d_c.

SISSA (Scuola Internazionale Superiore di Studi Avanzati), via Bonomea 265, I-34136 Trieste, Italy.
E-mail: laio@sissa.it (A.L.); alexrod@sissa.it (A.R.)

δ_i is measured by computing the minimum distance between the point i and any other point with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \quad (2)$$

For the point with highest density, we conventionally take $\delta_i = \max_j(d_{ij})$. Note that δ_i is much larger than the typical nearest neighbor distance only for points that are local or global maxima in the density. Thus, cluster centers are recognized as points for which the value of δ_i is anomalously large.

This observation, which is the core of the algorithm, is illustrated by the simple example in Fig. 1. Figure 1A shows 28 points embedded

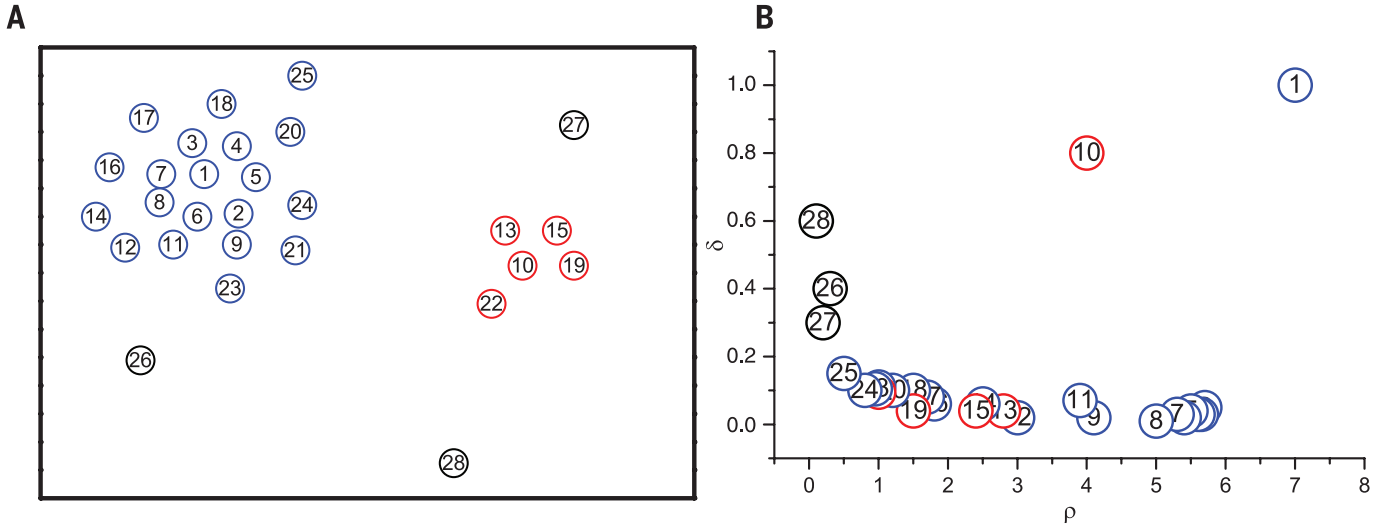


Fig. 1. The algorithm in two dimensions. (**A**) Point distribution. Data points are ranked in order of decreasing density. (**B**) Decision graph for the data in (A). Different colors correspond to different clusters.

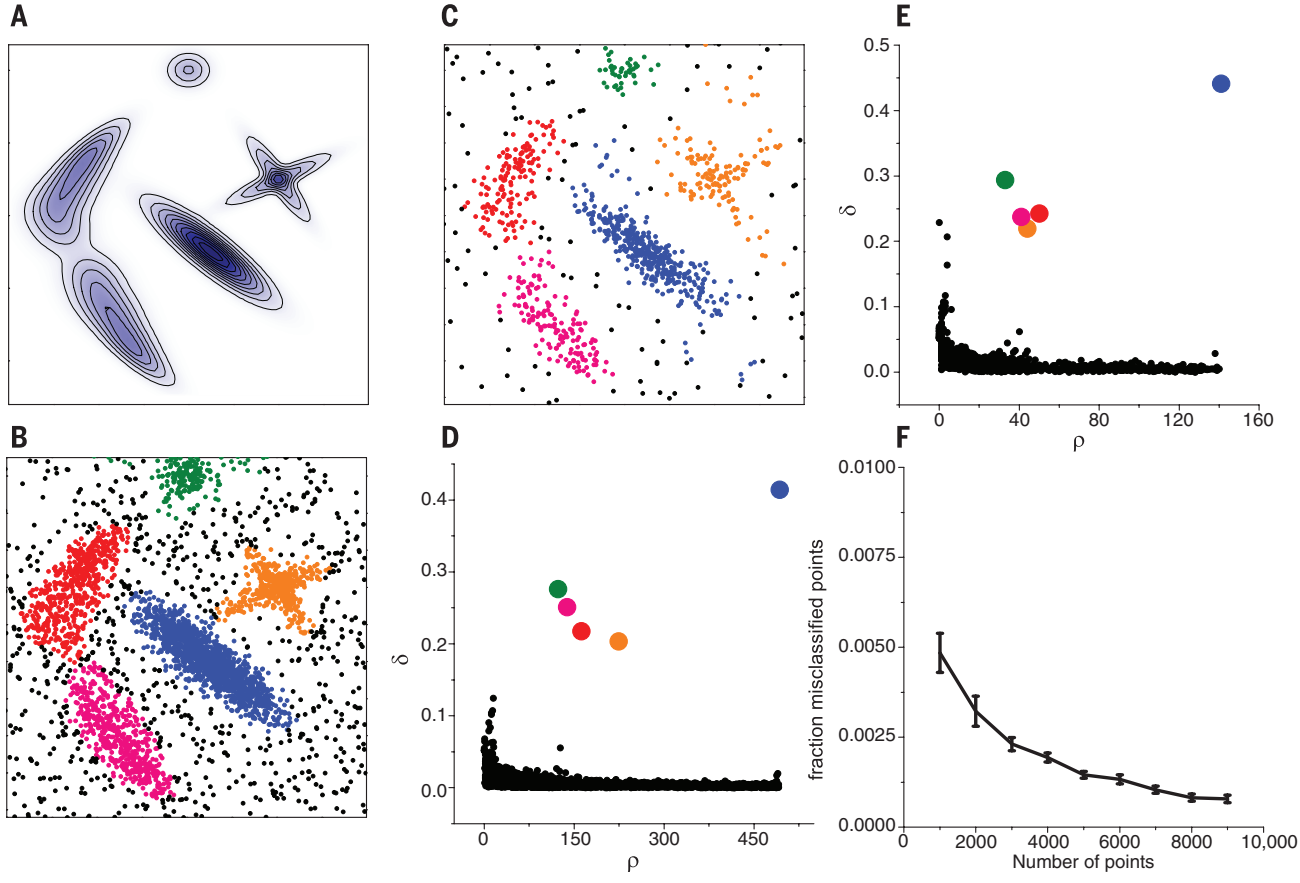


Fig. 2. Results for synthetic point distributions. (**A**) The probability distribution from which point distributions are drawn. The regions with lowest intensity correspond to a background uniform probability of 20%. (**B** and **C**) Point distributions for samples of 4000 and 1000 points, respectively. Points are colored according to the cluster to which they are assigned. Black points belong to the cluster halos. (**D** and **E**) The corresponding decision graphs, with the centers colored by cluster. (**F**) The fraction of points assigned to the incorrect cluster as a function of the sample dimension. Error bars indicate the standard error of the mean.

in a two-dimensional space. We find that the density maxima are at points 1 and 10, which we identify as cluster centers. Figure 1B shows the plot of δ_i as a function of ρ_i for each point; we will call this representation the decision graph. The value of δ for points 9 and 10, with similar values of ρ, is very different: Point 9 belongs to the cluster of point 1, and several other points with a higher ρ are very close to it, whereas the nearest neighbor of higher density of point 10 belongs to another cluster. Hence, as anticipated, the only points of high δ and relatively high ρ are the cluster centers. Points 26, 27, and 28 have a relatively high δ and a low ρ because they are isolated; they can be considered as clusters composed of a single point, namely, outliers.

After the cluster centers have been found, each remaining point is assigned to the same cluster as its nearest neighbor of higher density. The cluster assignment is performed in a single step, in contrast with other clustering algorithms where an objective function is optimized iteratively (*2*, *8*).

In cluster analysis, it is often useful to measure quantitatively the reliability of an assignment. In approaches based on the optimization of a function (*2*, *8*), its value at convergence is also a natural quality measure. In methods like DBSCAN (*9*), one considers reliable points with density values above a threshold, which can lead to low-density clusters, such as those in Fig. 2E, being classified as noise. In our algorithm, we do not introduce a noise-signal cutoff. Instead, we first find for each cluster a border region, defined as the set of points assigned to that cluster but being within a distance d_c from data points belonging to other clusters. We then find, for each cluster, the point of highest density within its border region. We denote its density by ρ_b. The points of the cluster whose density is higher than ρ_b are considered part of the cluster core (robust assignation). The others are considered part of the cluster halo (suitable to be considered as noise).

In order to benchmark our procedure, let us first consider the test case in Fig. 2. The data points are drawn from a probability distribution with nonspherical and strongly overlapping peaks (Fig. 2A); the probability values corresponding to the maxima differ by almost an order of magnitude. In Fig. 2, B and C, 4000 and 1000 points, respectively, are drawn from the distribution in Fig. 2A. In the corresponding decision graphs (Fig. 2, D and E), we observe only five points with a large value of δ and a sizeable density. These points are represented in the graphs as large solid circles and correspond to cluster centers. After the centers have been selected, each point is assigned either to a cluster or to the halo. The algorithm captures the position and shape of the probability peaks, even those corresponding to very different densities (blue and light green points in Fig. 2C) and nonspherical peaks. Moreover, points assigned to the halo correspond to regions that by visual inspection of the probability distribution in Fig. 2A would not be assigned to any peak.

To demonstrate the robustness of the procedure more quantitatively, we performed the analysis by drawing 10,000 points from the distribution in Fig. 2A, considering as a reference the cluster assignment obtained on that sample. We then obtained reduced samples by retaining only a fraction of points and performed cluster assignment for each reduced sample independently. Figure 2F shows, as a function of the size of the reduced sample, the fraction of points assigned to a cluster different than the one they were assigned to in the reference case. The fraction of misclassified points remains well below 1% even for small samples containing 1000 points.

Varying d_c for the data in Fig. 2B produced mutually consistent results (fig. S1). As a rule of thumb, one can choose d_c so that the average number of neighbors is around 1 to 2% of the total number of points in the data set. For data

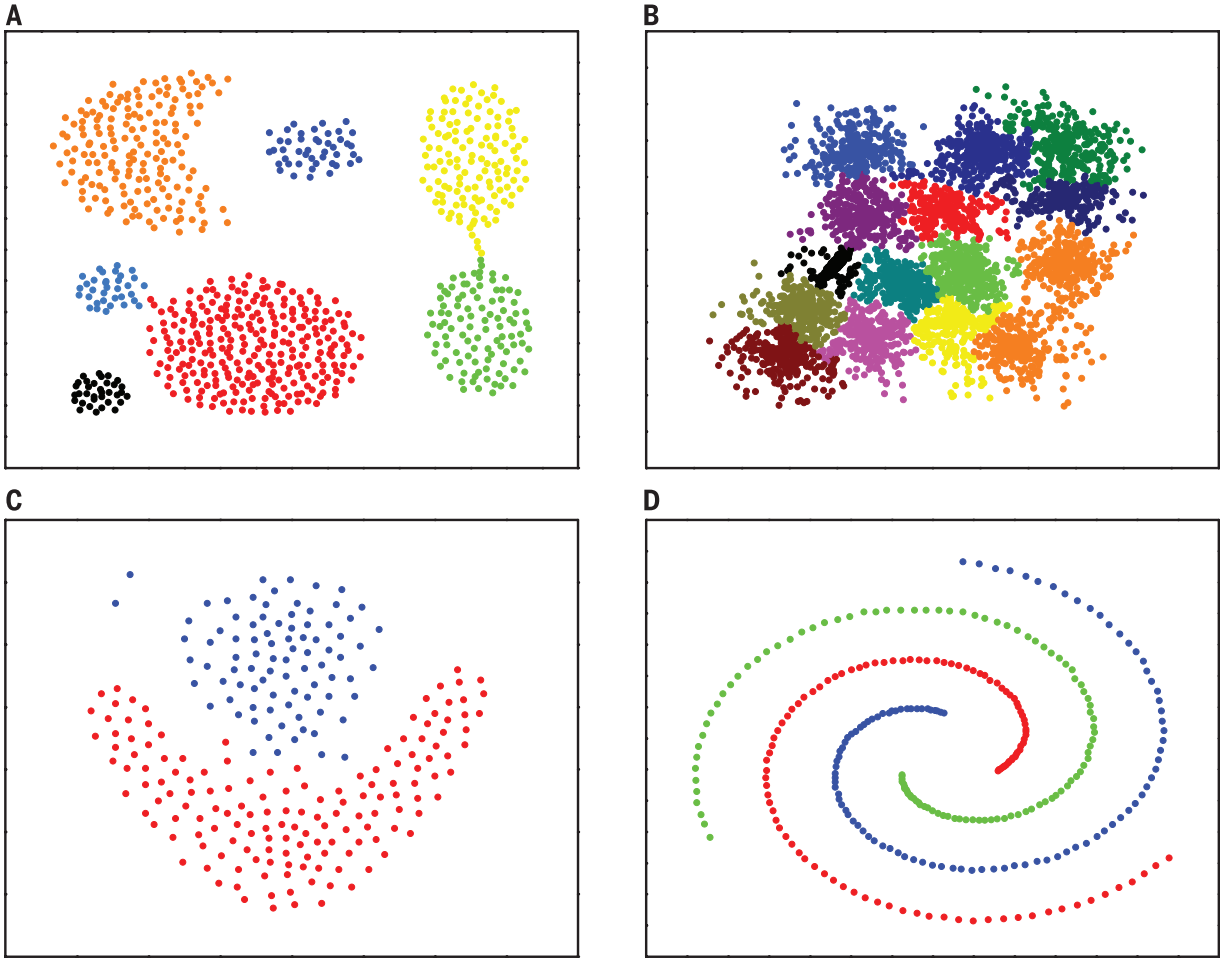


Fig. 3. Results for test cases in the literature. Synthetic point distributions from (*12*) (**A**), (*13*) (**B**), (*14*) (**C**), and (*15*) (**D**).

sets composed by a small number of points, ρ_i might be affected by large statistical errors. In these cases, it might be useful to estimate the density by more accurate measures (*10*, *11*).

Next, we benchmarked the algorithm on the test cases presented in Fig. 3. For computing the density for cases with few points, we adopted the exponential kernel described in (*11*). In Fig. 3A, we consider a data set from (*12*), obtaining results comparable to those of the original article, where it was shown that other commonly used methods fail. In Fig. 3B, we consider an example with 15 clusters with high overlap in data distribution taken from (*13*); our algorithm successfully determines the cluster structure of the data set. In Fig. 3C, we consider the test case for the FLAME (fuzzy clustering by local approximation of membership) approach (*14*), with results comparable to the original method. In the data set originally introduced to illustrate the performance of path-based spectral clustering (*15*) shown in Fig. 4D, our algorithm correctly finds the three clusters without the need of generating a connectivity graph. As comparison, in figs. S3 and S4 we show the cluster assignations obtained by K-means (*2*) for these four test cases and for the example in Fig. 2. Even if the K-means optimization is performed with use of the correct value of K, the assignations are, in most of the cases, not compliant with visual intuition.

The method is robust with respect to changes in the metric that do not significantly affect the distances below d_c, that is, that keep the density estimator in Eq. 1 unchanged. Clearly, the distance in Eq. 2 will be affected by such a change of metric, but it is easy to realize that the structure of the decision graph (in particular, the number of data points with a large value of δ) is a consequence of the ranking of the density values, not of the actual distance between far away points. Examples demonstrating this statement are shown in fig. S5.

Our approach only requires measuring (or computing) the distance between all the pairs of data points and does not require parameterizing a probability distribution (*8*) or a multidimensional density function (*10*). Therefore, its performance is not affected by the intrinsic dimensionality of the space in which the data points are embedded. We verified that, in a test case with 16 clusters in 256 dimensions (*16*), the algorithm finds the number of clusters and assigns the points correctly (fig. S6). For a data set with 210 measurements of seven x-ray features for three types of wheat seeds from (*17*), the algorithm correctly predicts the existence of three clusters and classifies correctly 97% of the points assigned to the cluster cores (figs. S7 and S8).

We also applied the approach to the Olivetti Face Database (*18*), a widespread benchmark for machine learning algorithms, with the aim of identifying, without any previous training, the number of subjects in the database. This data set poses a serious challenge to our approach because the "ideal" number of clusters (namely of distinct subjects) is comparable with the number of elements in the data set (namely of different images, 10 for each subject). This makes a reliable estimate of the densities difficult. The similarity between two images was computed by following (*19*). The density is estimated by a Gaussian kernel (*11*) with variance $d_c = 0.07$. For such a small set, the density estimator is unavoidably affected by large statistical errors; thus, we assign images to a cluster following a slightly more restrictive criterion than in the preceding examples. An image is assigned to the same cluster of its nearest image with higher density only if their distance is smaller than d_c. As a consequence, the images further than d_c from any other image of higher density remain unassigned. In Fig. 4, we show the results of an analysis performed for the first 100 images in the data set. The decision graph (Fig. 4A) shows the presence of several distinct density maxima. Unlike in other examples, their exact number is not clear, a consequence of the sparsity of the data points. A hint for choosing the number of centers is provided by the plot of $\gamma_i = \rho_i \delta_i$ sorted in decreasing order (Fig. 4B). This graph shows that this quantity, that is by definition large for cluster centers, starts growing

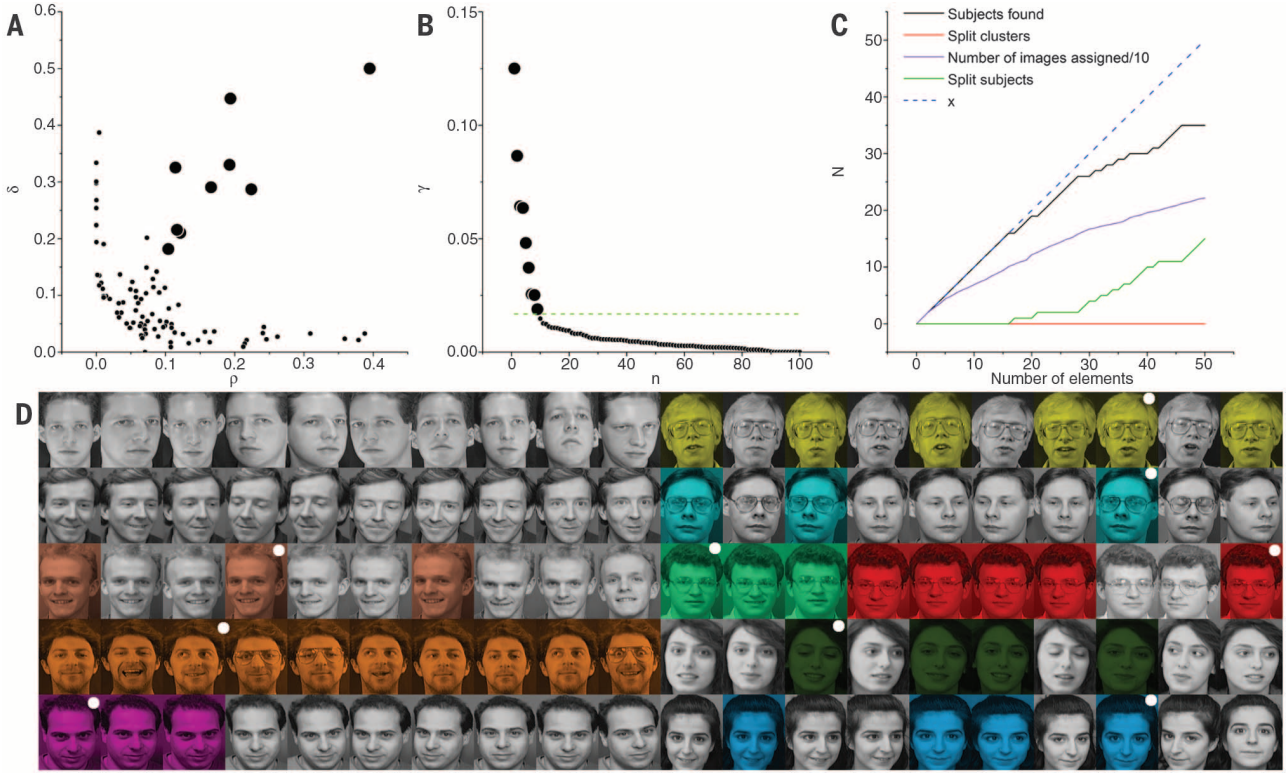


Fig. 4. Cluster analysis of the Olivetti Face Database. (**A**) The decision graph for the first hundred images in the database (*18*). (**B**) The value of $\gamma_i = \rho_i \delta_i$ in decreasing order for the data in (A). (**C**) The performance of the algorithm in recognizing subjects in the full database as a function of the number of clusters: number of subjects recognized as individuals (black line), number of clusters that include more than one subject (red line), number of subjects split in more than one cluster (green), and number of images assigned to a cluster divided by 10 (purple). (**D**) Pictorial representation of the cluster assignations for the first 100 images. Faces with the same color belong to the same cluster, whereas gray images are not assigned to any cluster. Cluster centers are labeled with white circles.

anomalously below a rank order 9. Therefore, we performed the analysis by using nine centers. In Fig. 4D, we show with different colors the clusters corresponding to these centers. Seven clusters correspond to different subjects, showing that the algorithm is able to "recognize" 7 subjects out of 10. An eighth subject appears split in two different clusters. When the analysis is performed on all 400 images of the database, the decision graph again does not allow recognizing clearly the number of clusters (fig. S9). However, in Fig. 4C we show that by adding more and more putative centers, about 30 subjects can be recognized unambiguously (fig. S9). When more centers are included, the images of some of the subjects are split in two clusters, but still all the clusters remain pure, namely include only images of the same subject. Following (*20*) we also computed the fraction of pairs of images of the same subject correctly associated with the same cluster (r_{true}) and the fraction of pairs of images of different subjects erroneously assigned to the same cluster (r_{false}). If one does not apply the cutoff at d_c in the assignation (namely if one applies our algorithm in its general formulation), one obtains $r_{true} \sim 68\%$ and $r_{false} \sim 1.2\%$ with ~42 to ~50 centers, a performance comparable to a state-of-the-art approach for unsupervised image categorization (*20*).

Last, we benchmarked the clustering algorithm on the analysis of a molecular dynamics trajectory of trialanine in water at 300 K (*21*). In this case, clusters will approximately correspond to kinetic basins, namely independent conformations of the system that are stable for a substantial time and separated by free energy barriers, that are crossed only rarely on a microscopic time scale. We first analyzed the trajectory by a standard approach (*22*) based on a spectral analysis of the kinetic matrix, whose eigenvalues are associated with the relaxation times of the system. A gap is present after the seventh eigenvalue (fig. S10), indicating that the system has eight basins; in agreement with that, our cluster analysis (fig. S10) gives rise to eight clusters, including conformations in a one-to-one correspondence with those defining the kinetic basins (*22*).

Identifying clusters with density maxima, as is done here and in other density-based clustering algorithms (*9*, *10*), is a simple and intuitive choice but has an important drawback. If one generates data points at random, the density estimated for a finite sample size is far from uniform and is instead characterized by several maxima. However, the decision graph allows us to distinguish genuine clusters from the density ripples generated by noise. Qualitatively, only in the former case are the points corresponding to cluster centers separated by a sizeable gap in ρ and δ from the other points. For a random distribution, one instead observes a continuous distribution in the values of ρ and δ. Indeed, we performed the analysis for sets of points generated at random from a uniform distribution in a hypercube. The distances between data points entering in Eqs. 1 and 2 are computed with periodic boundary conditions on the hypercube. This analysis shows that, for randomly distributed data points, the quantity $\gamma_i = \rho_i \delta_i$ is distributed according to a power law, with an exponent that depends on the dimensionality of the space in which the points are embedded. The distributions of γ for data sets with genuine clusters, like those in Figs. 2 to 4, are strikingly different from power laws, especially in the region of high γ (fig. S11). This observation may provide the basis for a criterion for the automatic choice of the cluster centers as well as for statistically validating the reliability of an analysis performed with our approach.

REFERENCES AND NOTES

1. R. Xu, D. Wunsch 2nd, *IEEE Trans. Neural Netw.* **16**, 645–678 (2005).
2. J. MacQueen, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam, J. Neyman, Eds. (Univ. California Press, Berkeley, CA, 1967), vol. 1, pp. 281–297.
3. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344 (Wiley-Interscience, New York, 2009).
4. B. J. Frey, D. Dueck, *Science* **315**, 972–976 (2007).
5. J. H. Ward Jr., *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
6. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition* (Wiley, New York, 1999).
7. A. K. Jain, *Pattern Recognit. Lett.* **31**, 651–666 (2010).
8. G. J. McLachlan, T. Krishnan, *The EM Algorithm and Extensions* (Wiley Series in Probability and Statistics vol. 382, Wiley-Interscience, New York, 2007).
9. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, E. Simoudis, J. Han, U. Fayyad, Eds. (AAAI Press, Menlo Park, CA, 1996), pp. 226–231.
10. K. Fukunaga, L. Hostetler, *IEEE Trans. Inf. Theory* **21**, 32–40 (1975).
11. Y. Cheng, *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 790 (1995).
12. A. Gionis, H. Mannila, P. Tsaparas, *ACM Trans. Knowl. Discovery Data* **1**, 4, es (2007).
13. P. Fränti, O. Virmajoki, *Pattern Recognit.* **39**, 761–775 (2006).
14. L. Fu, E. Medico, *BMC Bioinformatics* **8**, 3 (2007).
15. H. Chang, D.-Y. Yeung, *Pattern Recognit.* **41**, 191–203 (2008).
16. P. Fränti, O. Virmajoki, V. Hautamäki, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1875–1881 (2006).
17. M. Charytanowicz *et al.*, *Information Technologies in Biomedicine* (Springer, Berlin, 2010), pp. 15–24.
18. F. S. Samaria, A. C. Harter, in *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision* (IEEE, New York, 1994), pp. 138–142.
19. M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, M. K. Markey, *IEEE Trans. Image Process.* **18**, 2385–2401 (2009).
20. D. Dueck, B. Frey, *ICCV 2007. IEEE 11th International Conference on Computer Vision* (IEEE, New York, 2007), pp. 1–8.
21. F. Marinelli, F. Pietrucci, A. Laio, S. Piana, *PLOS Comput. Biol.* **5**, e1000452 (2009).
22. I. Horenko, E. Dittmer, A. Fischer, C. Schütte, *Multiscale Model. Simulation* **5**, 802–827 (2006).

ACKNOWLEDGMENTS

We thank E. Tosatti, D. Amati, F. Laio, F. Marinelli, A. Maritan, R. Allen, J. Nasica, and M. d'Errico for stimulating discussion. We acknowledge financial support from the grant Associazione Italiana per la Ricerca sul Cancro 5 per mille, Rif. 12214, and Fondo per gli Investimenti della Ricerca di Base–Accordo di programma, Rif. RBAP11ETKA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/344/6191/1492/suppl/DC1
Figs. S1 to S11
Data S1

18 June 2013; accepted 23 May 2014
10.1126/science.1242072

NANOFLUIDICS

Observing liquid flow in nanotubes by 4D electron microscopy

Ulrich J. Lorenz and Ahmed H. Zewail*

Nanofluidics involves the study of fluid transport in nanometer-scale structures. We report the direct observation of fluid dynamics in a single zinc oxide nanotube with the high spatial and temporal resolution of four-dimensional (4D) electron microscopy. The nanotube is filled with metallic lead, which we melt in situ with a temperature jump induced by a heating laser pulse. We then use a short electron pulse to create an image of the ensuing dynamics of the hot liquid. Single-shot images elucidate the mechanism of irreversible processes, whereas stroboscopic diffraction patterns provide the heating and cooling rates of single nanotubes. The temporal changes of the images enable studies of the viscous friction involved in the flow of liquid within the nanotube, as well as studies of mechanical processes such as those that result in the formation of extrusions.

Advances in nanofabrication have made it possible to reduce the size of microfluidic devices and to study fluid flow at the nanometer scale (*1*, *2*). Nanoscale fluid dynamics and transport properties are dominated by surface effects and may substantially differ from those occurring at larger scales. For water in carbon nanotubes, for example, flow rates have been reported to exceed the predictions of classical continuum theory by several orders of magnitude (*3*–*5*). However, the degree of the enhancement remains a point of discussion (*6*). The study of a single nanochannel, rather than a large ensemble, should reduce the experimental uncertainty and provide an opportunity to visualize mechanical and fluid dynamics at the nanoscale. Such experiments not only incur the challenge of preparing

Physical Biology Center for Ultrafast Science and Technology, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, CA 91125, USA.
*Corresponding author. E-mail: zewail@caltech.edu

www.sciencemag.org/content/344/6191/1492/suppl/DC1

Supplementary Materials for

Clustering by fast search and find of density peaks

Alex Rodriguez and Alessandro Laio

E-mail: laio@sissa.it (A.L.); alexrod@sissa.it (A.R.)

Published 27 June 2014, *Science* **344**, 1492 (2014)
DOI: 10.1126/science.1242072

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
available at www.sciencemag.org/content/344/6191/1492/suppl/DC1

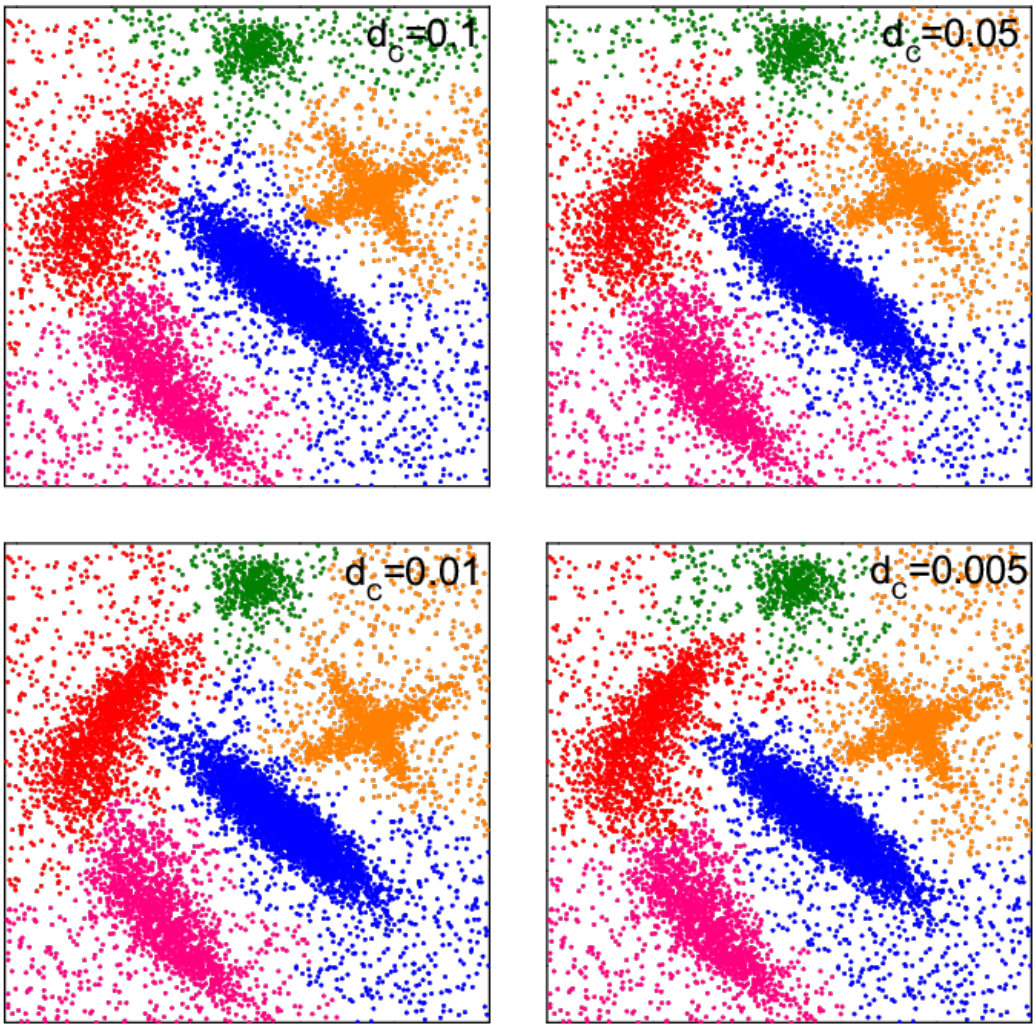


Fig. S1.

Comparison of the assignation for several values of d_c for the example in Fig. 2. Although the value of d_c varies by a factor of 20 and, consequently, the average number of neighbours varies between 11 % and 0.2 %, the assignations are very similar.

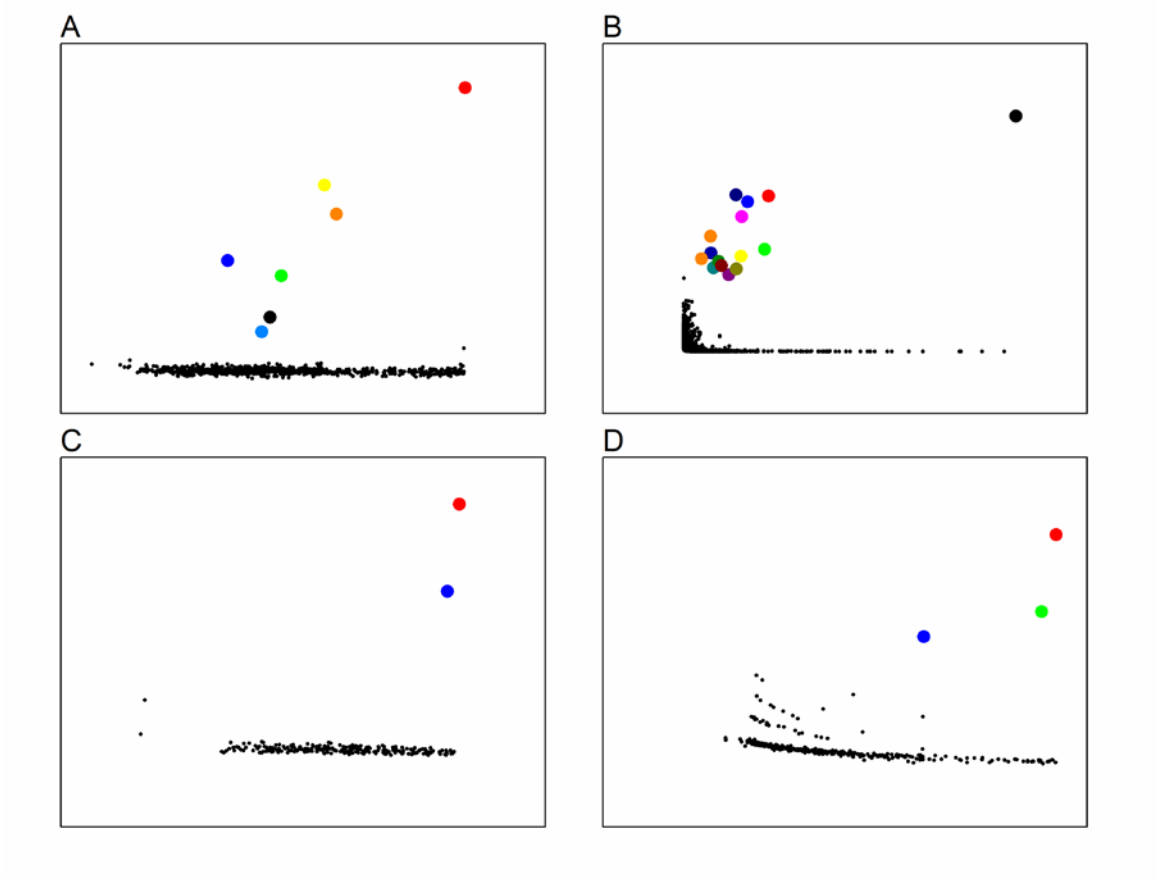


Fig. S2
Decision graphs for the data point distributions in Figure 3

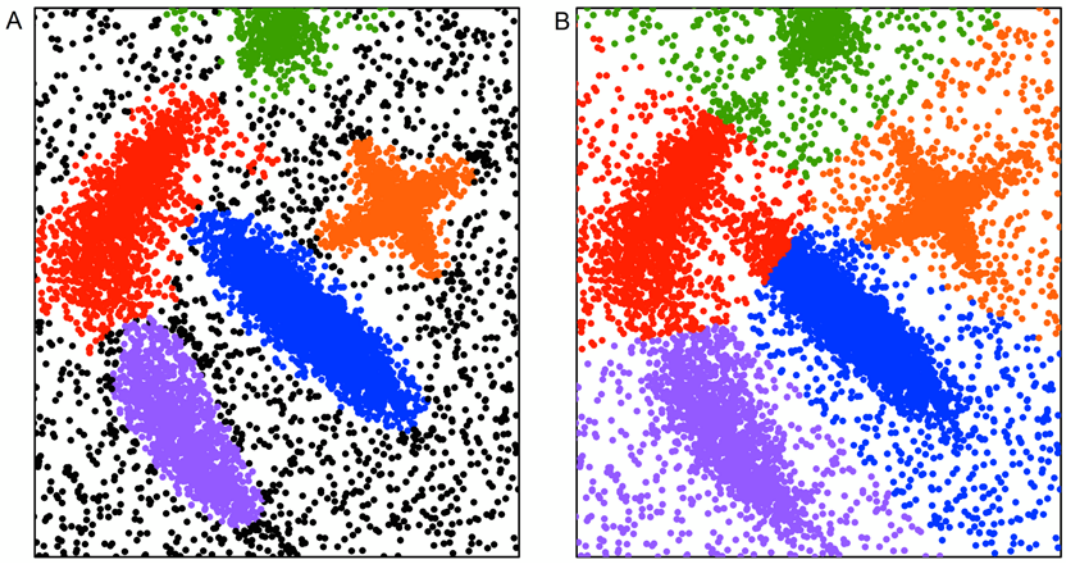


Fig. S3

Comparison between the present method (panel A) and K-means (panel B) for 10000 points harvested from the probability distribution shown in Fig. 2A. Following Ref. 4 K-means results have been obtained by running 10000 times the algorithm and taking the best solution according to the objective function. The value of K has been set to the 5.

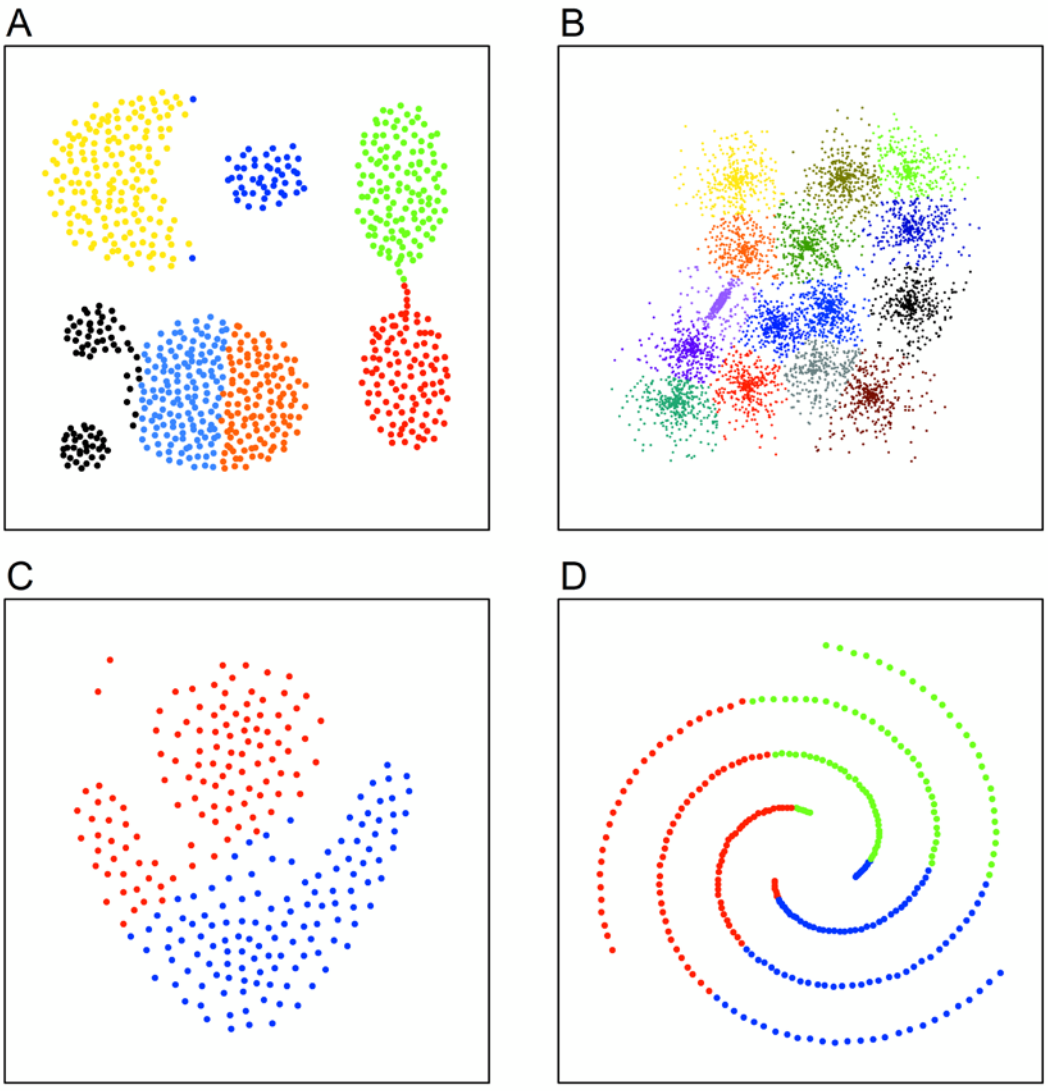


Fig. S4

K-means assignations for the data point distributions in Figure 3. Following Ref. 4, K-means results have been obtained by running 10000 times the algorithm and taking the best solution according to the objective function. In all the cases, the value of K has been chosen by visual inspection . Thus, K=7, 15, 2 and 3 for panel A, B, C and D respectively.

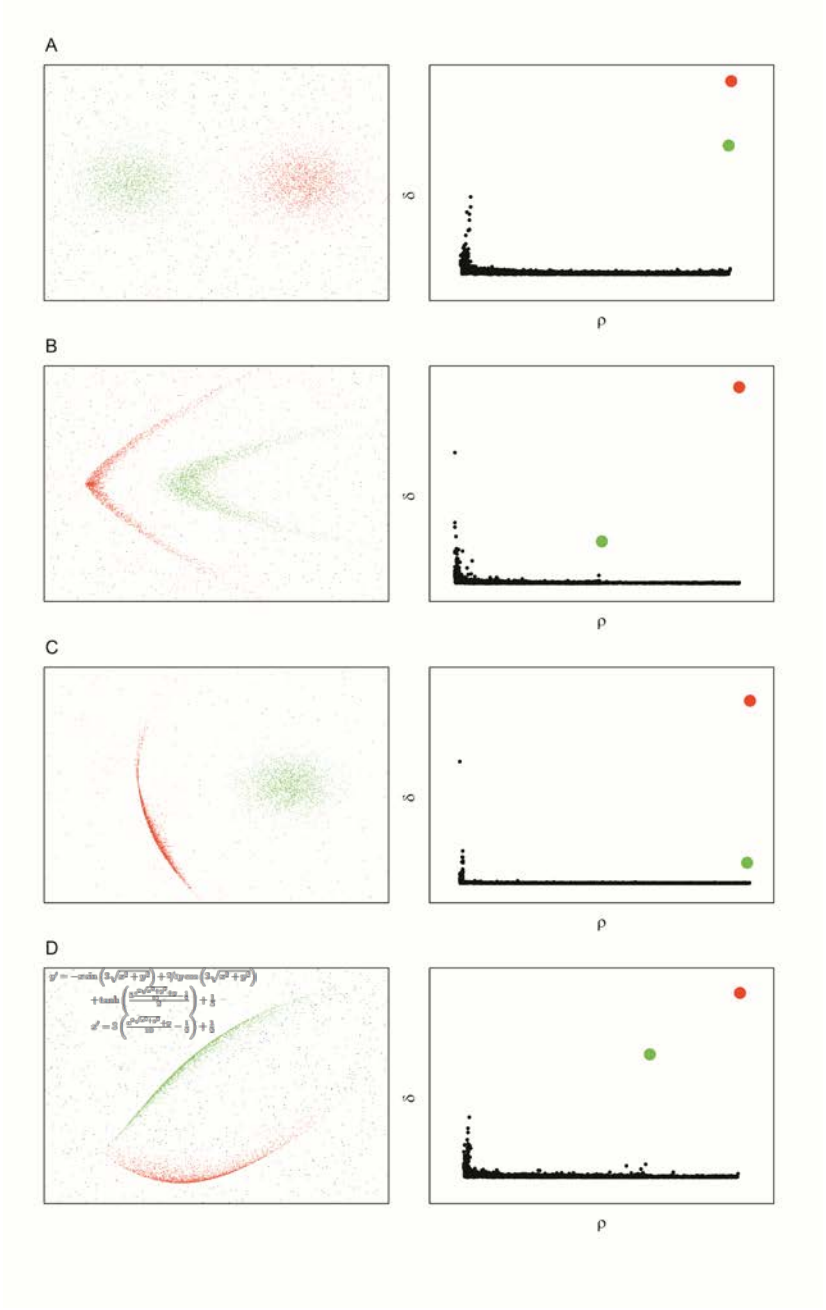


Fig. S5

With the aim of illustrate the robustness of the method with respect to changes in the metric, the algorithm has been applied on the trivial spherical distribution shown in panel A, as well to the sets generated by non-linear transformations and shown in panels B,C, and D. In the last case, the form of the transformation is explicitly shown in the inset. Applying the algorithm with the same density estimator we obtain a correct identification of the clusters, as well as qualitatively similar decision graphs, all consistent with the presence of two clusters.

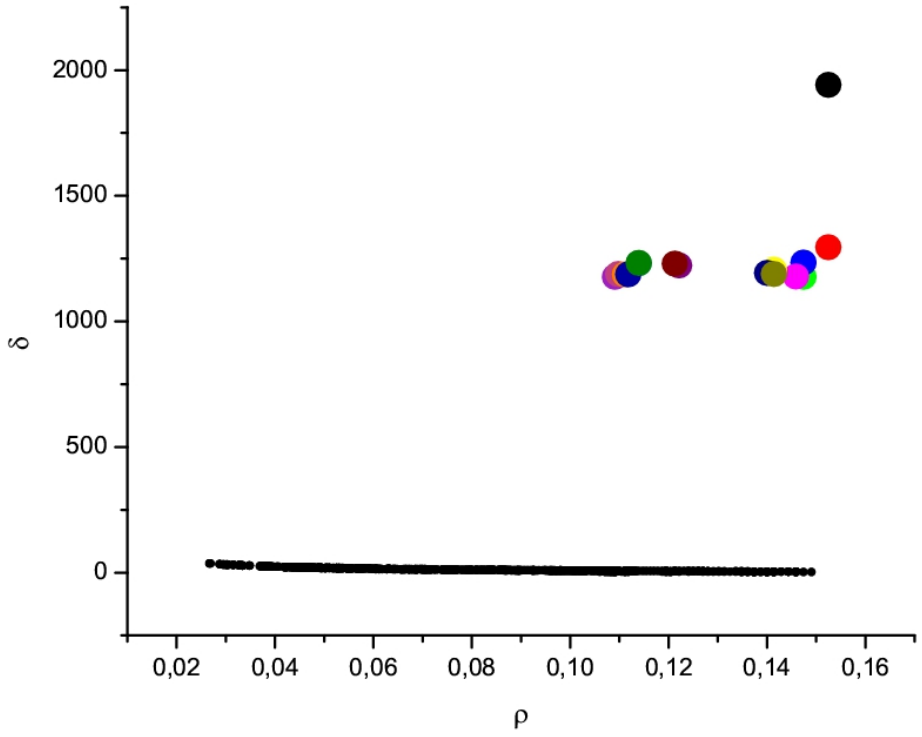


Fig. S6

Decision graph for the syntetic data example with 16 cluster in 256 dimensions from ref. [16]

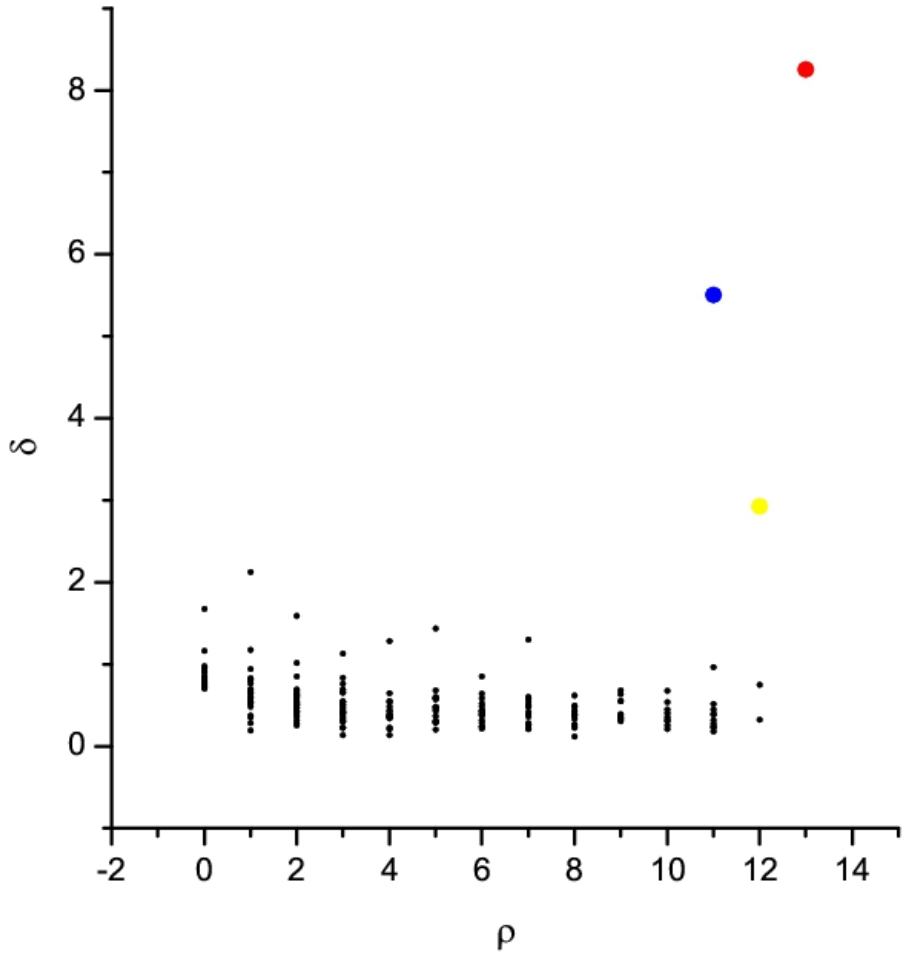


Fig. S7
Decision graph for the seeds data set from ref. [17]

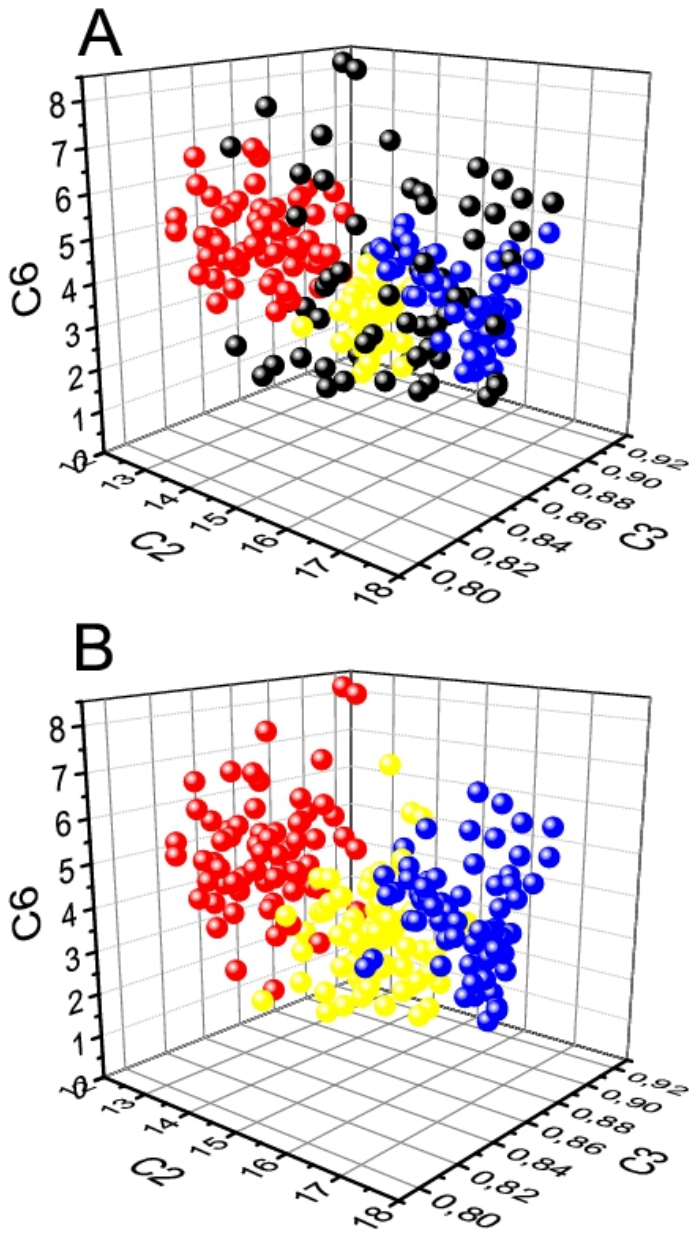


Fig. S8

Assignation for the seeds data set (panel A) compared with the different species (panel B) projected in a three dimensional subspace. In panel A each color corresponds to a different cluster, with halo points coloured in black. In panel B each color corresponds to a different species.

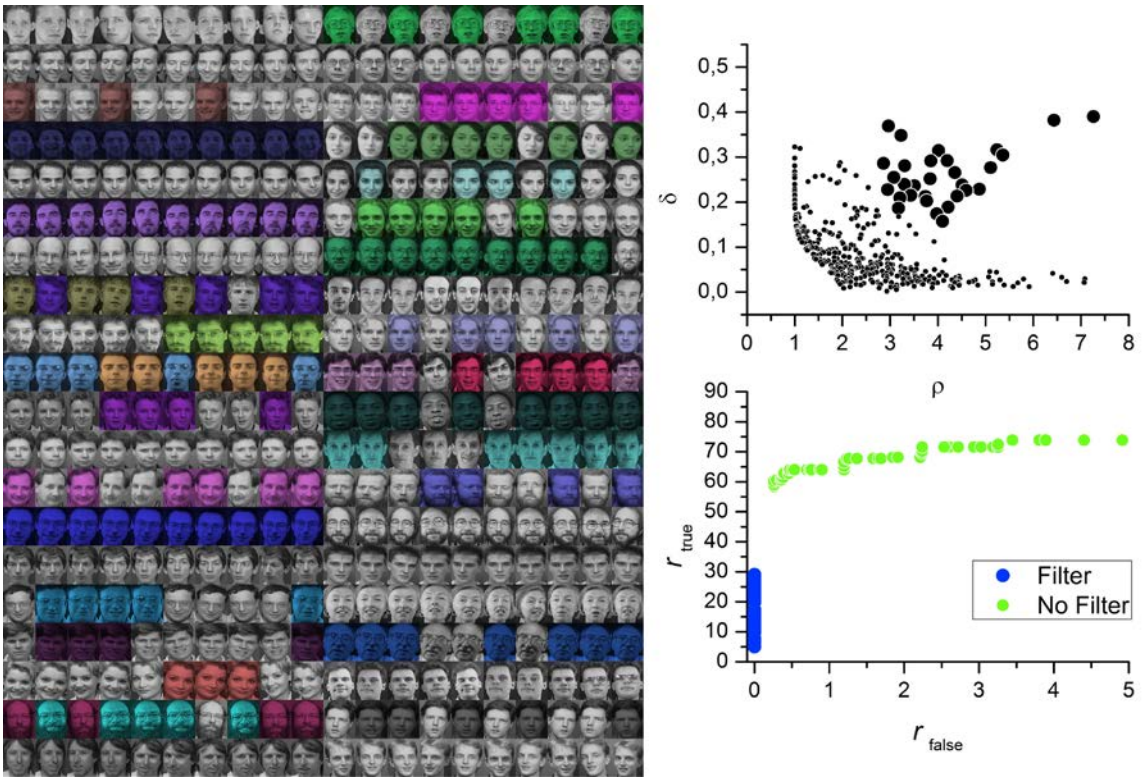


Fig. S9

Cluster analysis performed on the whole Olivetti face database. Top-Right panel: The decision graph. For this dataset the "ideal" number of clusters is 40, with only 10 elements for each cluster. Since each putative density peak includes only 10 elements, the estimator of the density is unavoidably affected by a large statistical error. In these conditions, it can be difficult to deduce from the decision graph the exact number of density peaks. In figure we highlight the 30 data points with the highest value of $\gamma_i = \rho_i \delta_i$. Left panel: Images in the database colored by cluster for the case of 30 centers. Light grey images are not assigned to any cluster. Notice that a few subjects are split in two clusters, but not a single cluster includes images of two different subjects. In Fig 4 of the manuscript we report the performance of the algorithm in recognizing the subjects for different numbers of centers. Bottom-Right panel: the fraction of pair of images of the same subject correctly associated to the same cluster (r_{true}) as a function of the fraction of pair of images of different subjects erroneously assigned to the same cluster (r_{false}). Each point corresponds to a different number of putative centers. Blue points: an image is assigned to the same cluster of its nearest image with higher density only if their distance is smaller than d_c =0.07. The assignation in the left panel has been obtained applying this criterion. Notice that with this filter one finds r_{false}=0 for any number of clusters. Since several images are not assigned to any cluster, one finds values of r_{true} of 30 % or less. Green points: an image is always assigned to the same cluster of its nearest image with higher density. In this manner all the images are assigned to a cluster, like in the k-medoids approach and in the affinity propagation approach. In this case, some of the clusters can contain images from different subjects and $r_{false} \neq 0$, but the typical values of r_{true} are much larger than in the former case.

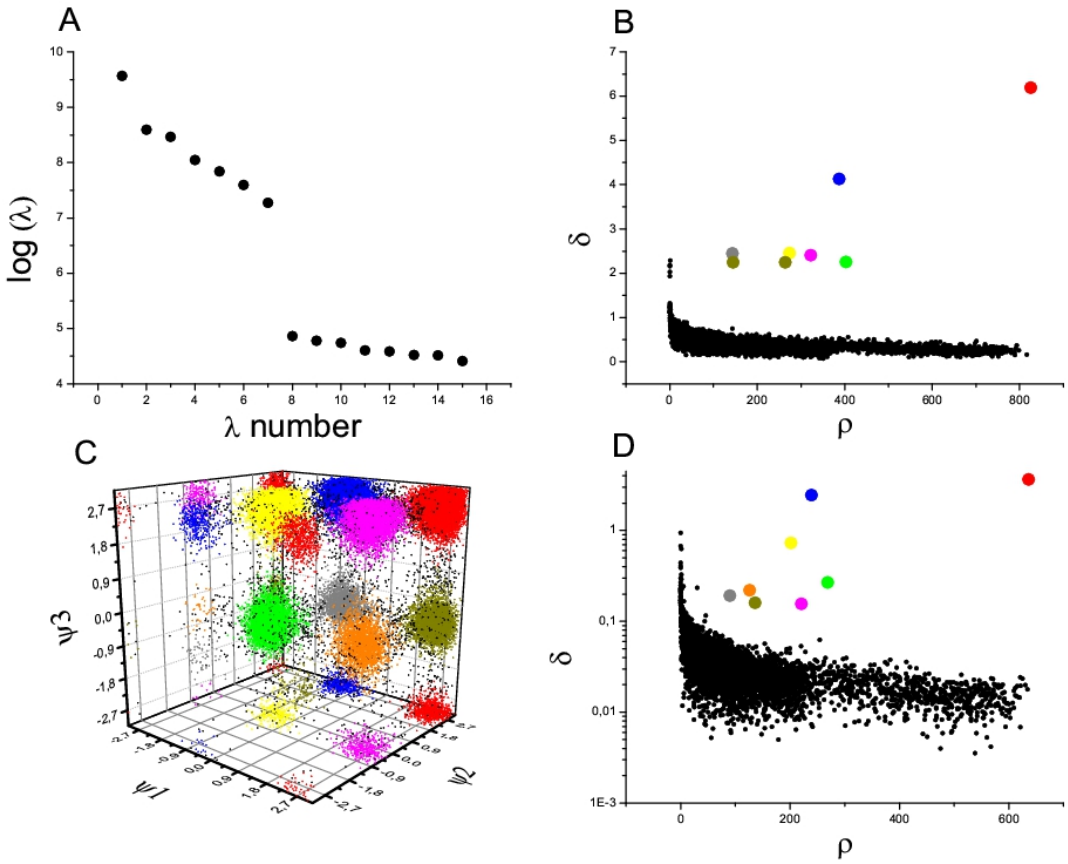


Fig. S10

Cluster analysis of a 2 μs molecular dynamics trajectory of trialanine in water solution from [21]. The eigenvalues of the kinetic matrix (panel A) show that there are seven relevant eigenvectors, indicating that the system has 8 kinetic basins [21, 22]. We then performed the cluster analysis on a data set generated by computing the values of the 6 backbone dihedrals from the trajectory. The distance between two configurations is estimated from the root mean square difference between these dihedral angles, with the differences computed taking into account the periodicity. The decision graph is shown in panel B. The number of clusters is eight, in perfect agreement with the kinetic analysis. In panel C the clusters are shown as a function of the three ψ angles, indicating that different clusters are distinguished by the value of the three ψ angles, once again consistently with the standard kinetic analysis [21]. Moreover, if we adopt a radically different metric to define the distance between two configurations, namely the root mean square deviation (RMSD) of the Cartesian coordinates of the backbone atoms, the method still detects the same eight clusters found using the dihedral distance. The decision graph obtained using this metric is shown in panel D.

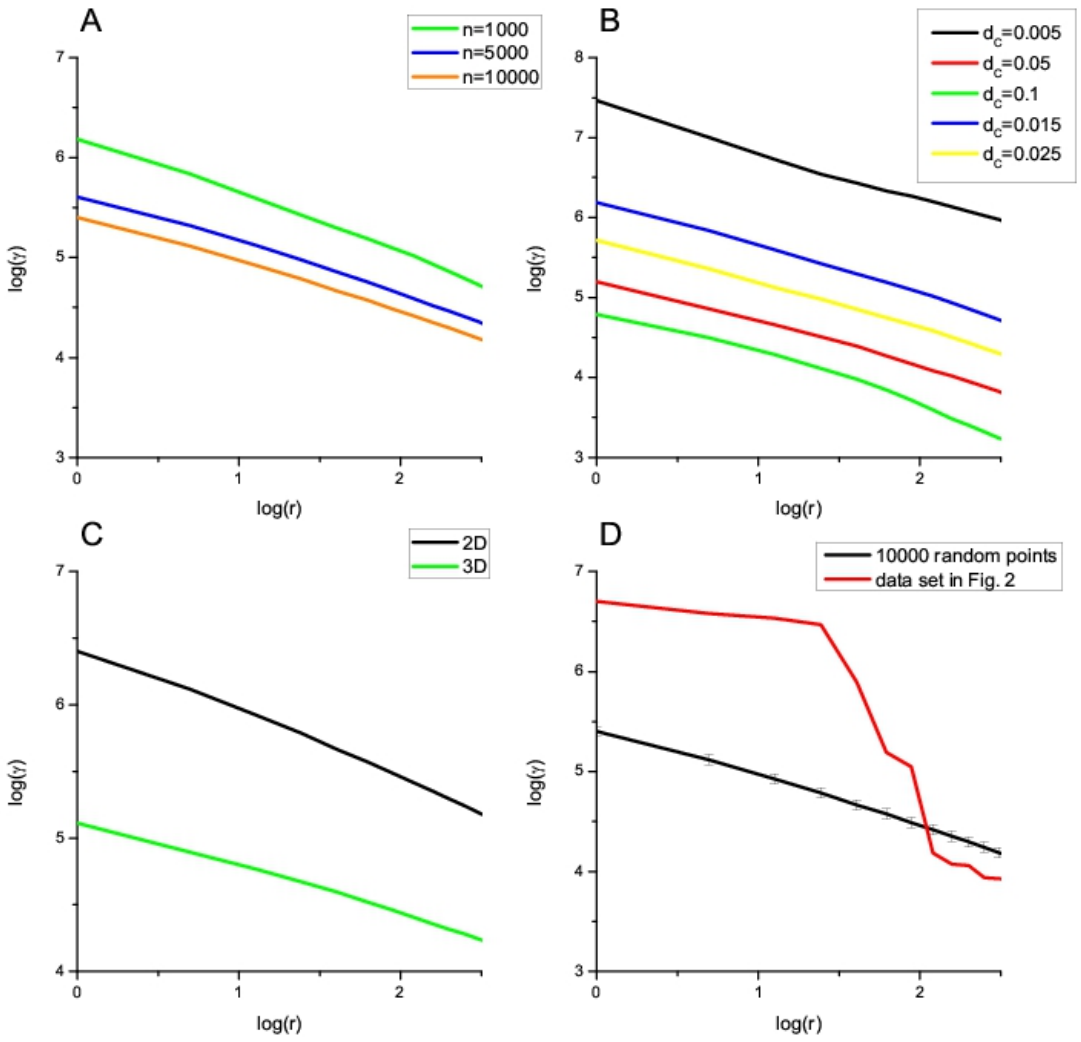


Fig. S11

The value of $\gamma_i = \rho_i \delta_i$ sorted in decreasing order as a function of the rank r_i of data point i in the sorted list. The data points in panel A,B and C are all distributed at random in a hypercube of size one. The distances between data points are computed with periodic boundary conditions. The curves are the average over 500 independently realizations. Panel A: The distribution of γ_i for different sample sizes in a two-dimensional box. Panel B: The distribution for different choices of the cutoff parameter d_c in eq. 1. Panel C: The distribution for points in a two dimensional box and in a three dimensional box. Panel D: Comparison between the distribution of γ_i in the set in Fig. 2 and for a sample of randomly distributed points. The error bars are estimated from the standard deviation of the distribution over the 500 independent realizations.

References and Notes

1. R. Xu, D. Wunsch 2nd, Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**, 645–678 (2005). Medline doi:10.1109/TNN.2005.845141

2. J. MacQueen, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam, J. Neyman, Eds. (Univ. California Press, Berkeley, CA, 1967), vol. 1, pp. 281–297.

3. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344 (Wiley-Interscience, New York, 2009).

4. B. J. Frey, D. Dueck, Clustering by passing messages between data points. *Science* **315**, 972–976 (2007). Medline doi:10.1126/science.1136800

5. J. H. Ward Jr., Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963). doi:10.1080/01621459.1963.10500845

6. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition* (Wiley, New York, 1999).

7. A. K. Jain, Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666 (2010). doi:10.1016/j.patrec.2009.09.011

8. G. J. McLachlan, T. Krishnan, *The EM Algorithm and Extensions* (Wiley Series in Probability and Statistics vol. 382, Wiley-Interscience, New York, 2007).

9. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, E. Simoudis, J. Han, U. Fayyad, Eds. (AAAI Press, Menlo Park, CA, 1996), pp. 226–231.

10. K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **21**, 32–40 (1975). doi:10.1109/TIT.1975.1055330

11. Y. Cheng, Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 790 (1995). doi:10.1109/34.400568

12. A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation. *ACM Trans. Knowl. Discovery Data* **1**, 4, es (2007). doi:10.1145/1217299.1217303

13. P. Fränti, O. Virmajoki, Iterative shrinking method for clustering problems. *Pattern Recognit.* **39**, 761–775 (2006). doi:10.1016/j.patcog.2005.09.012

14. L. Fu, E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics* **8**, 3 (2007). Medline doi:10.1186/1471-2105-8-3

15. H. Chang, D.-Y. Yeung, Robust path-based spectral clustering. *Pattern Recognit.* **41**, 191–203 (2008). doi:10.1016/j.patcog.2007.04.010

16. P. Fränti, O. Virmajoki, V. Hautamäki, Fast agglomerative clustering using a k-nearest neighbor graph. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1875–1881 (2006). Medline doi:10.1109/TPAMI.2006.227

17. M. Charytanowicz *et al*., *Information Technologies in Biomedicine* (Springer, Berlin, 2010), pp. 15–24.

18. F. S. Samaria, A. C. Harter, in ‚*Proceedings of 1994 IEEE Workshop on Applications of Computer Vision* (IEEE, New York, 1994), pp. 138–142.

19. M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, M. K. Markey, Complex wavelet structural similarity: A new image similarity index. *IEEE Trans. Image Process.* **18**, 2385–2401 (2009). Medline doi:10.1109/TIP.2009.2025923

20. D. Dueck, B. Frey, *ICCV 2007. IEEE 11th International Conference on Computer Vision* (IEEE, New York, 2007), pp. 1–8.

21. F. Marinelli, F. Pietrucci, A. Laio, S. Piana, A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLOS Comput. Biol.* **5**, e1000452 (2009). Medline doi:10.1371/journal.pcbi.1000452

22. I. Horenko, E. Dittmer, A. Fischer, C. Schütte, Automated model reduction for complex systems exhibiting metastability. *Multiscale Model. Simulation* **5**, 802–827 (2006). doi:10.1137/050623310