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

Data S1

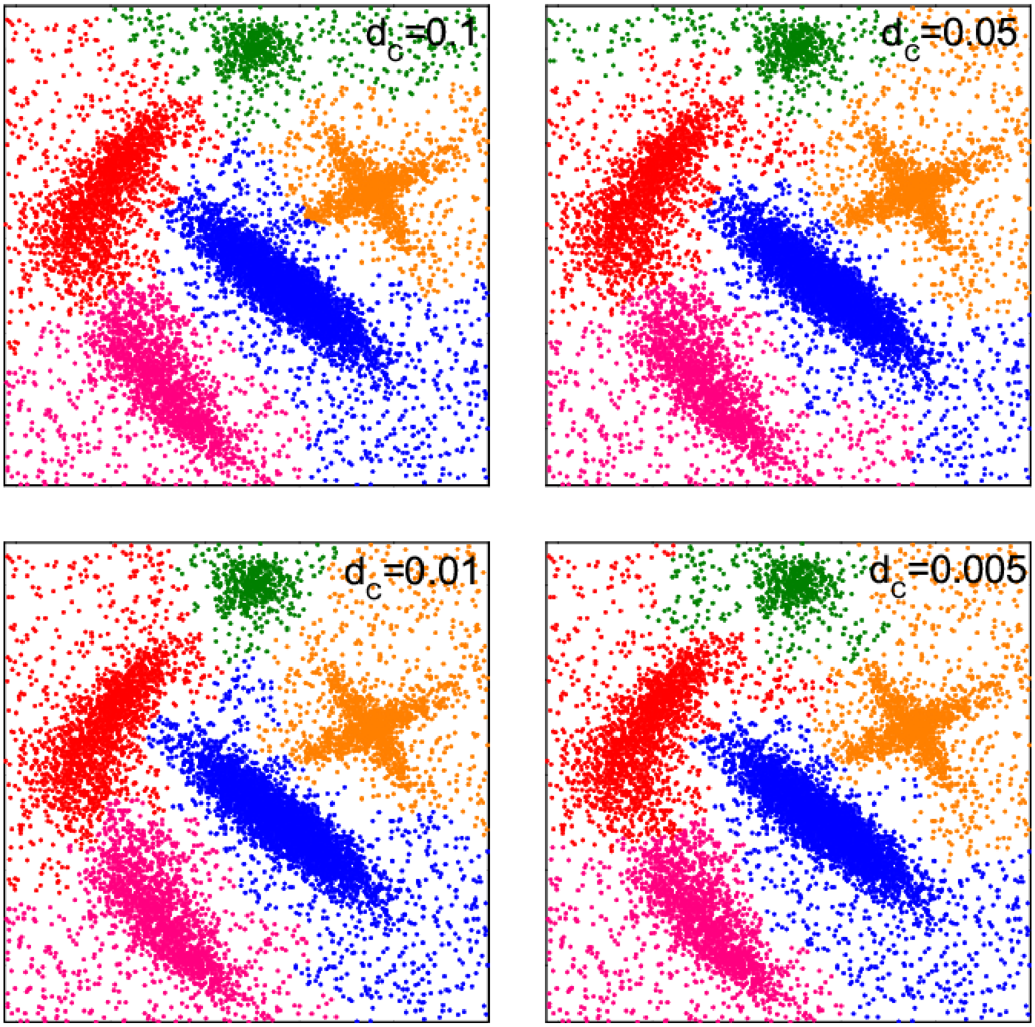

Fig. S1.

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.

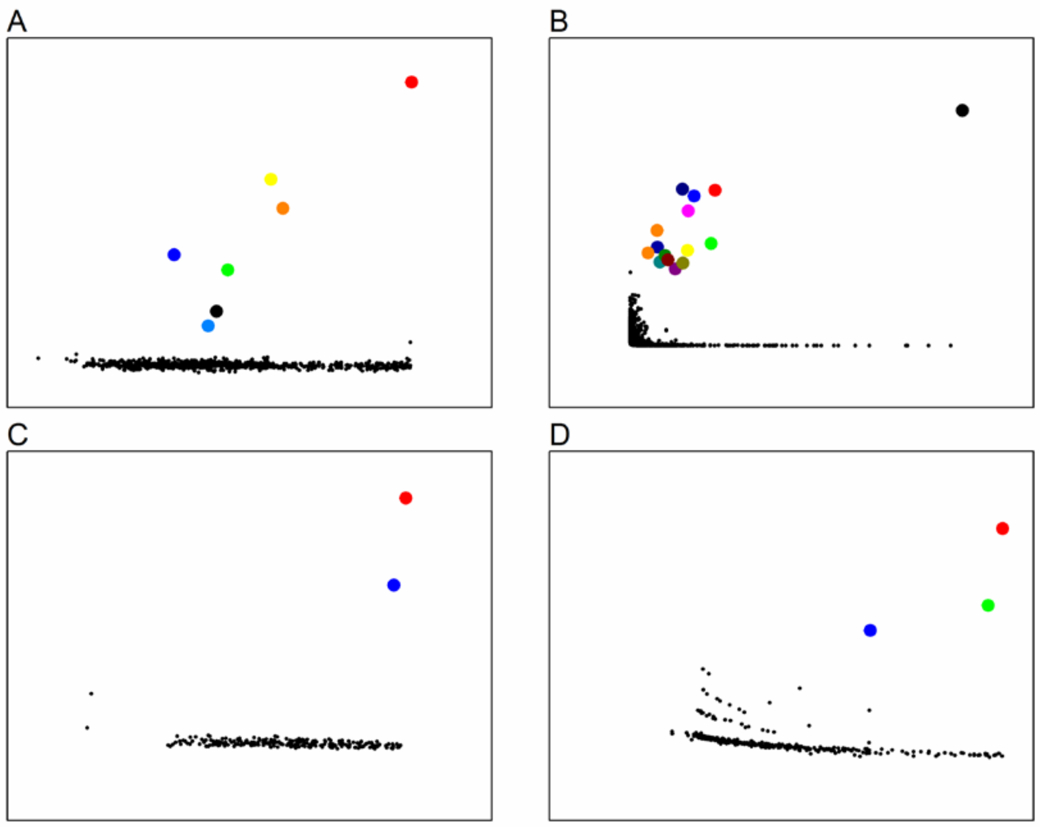

Fig. S2
Decision graphs for the data point distributions in Figure 3

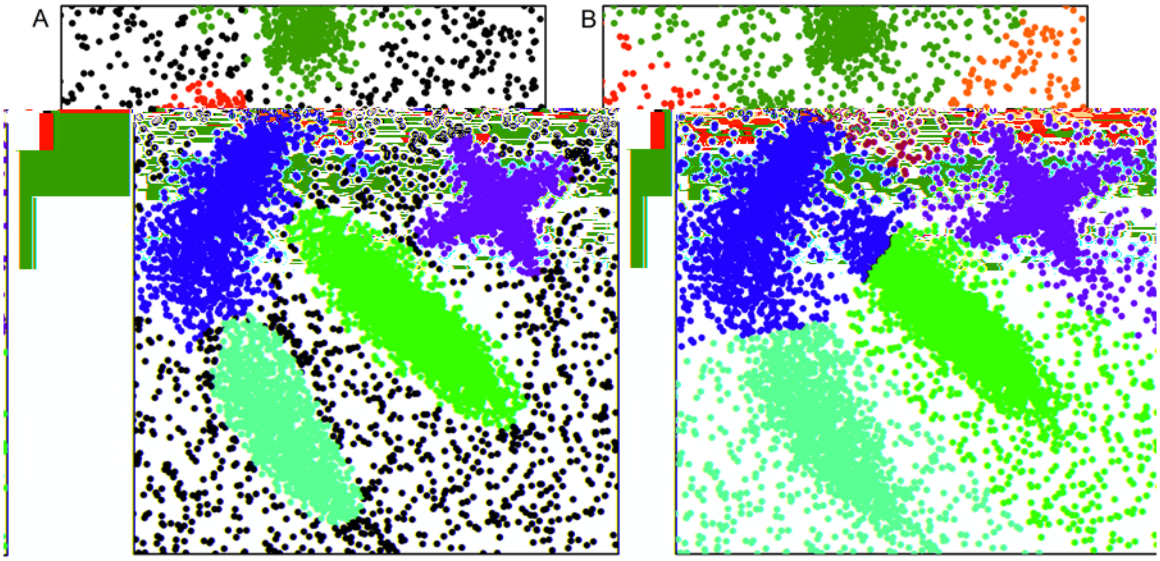

Fig. S3

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.

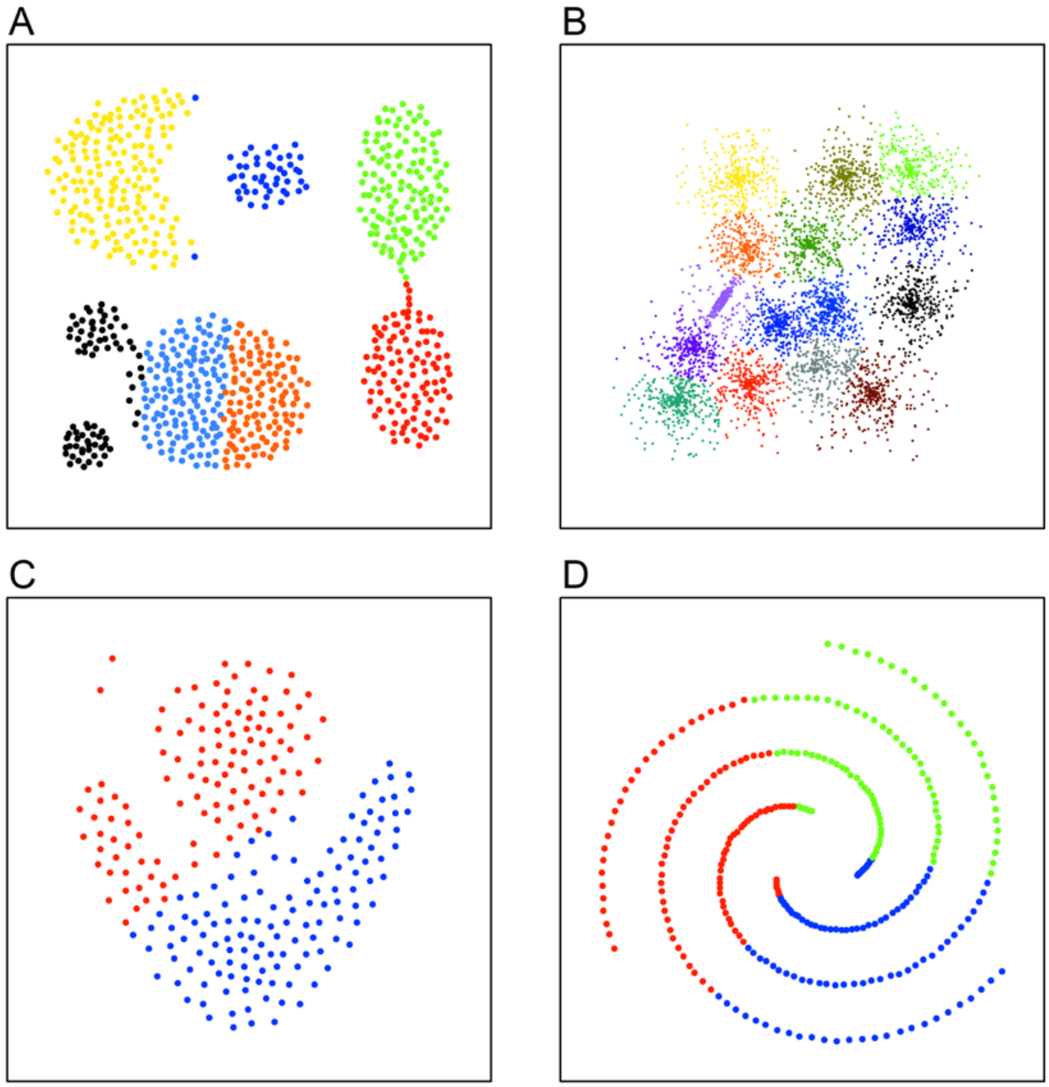

Fig. S4

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.

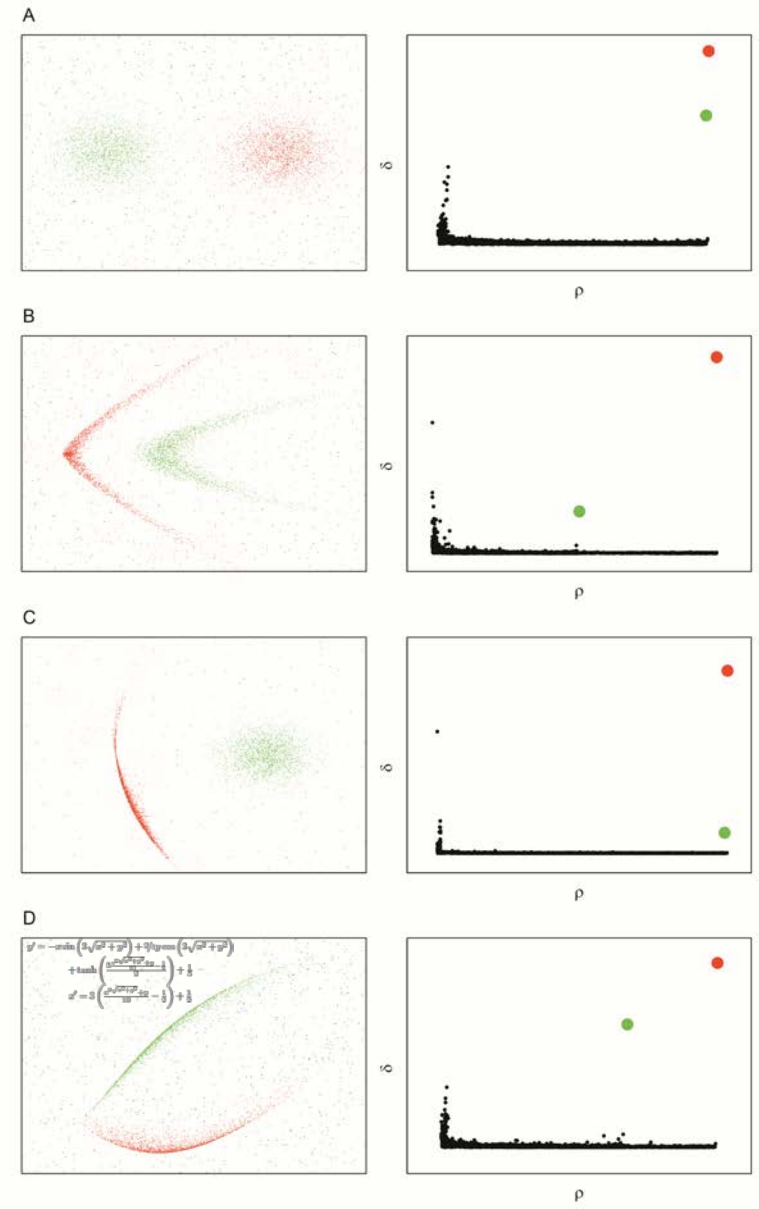

Fig. S5

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.

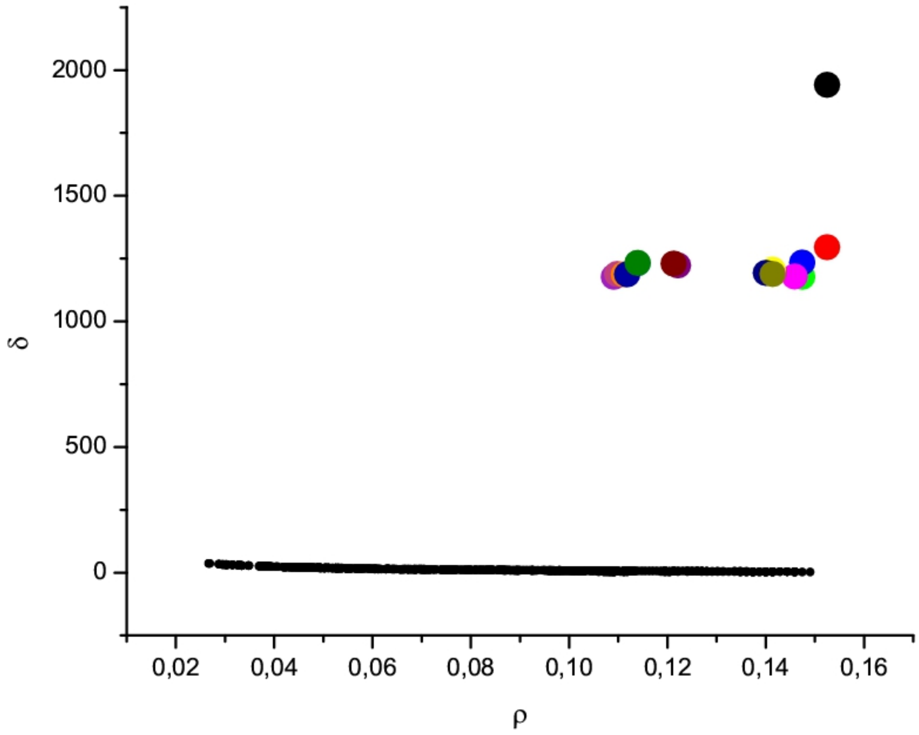

Fig. S6

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

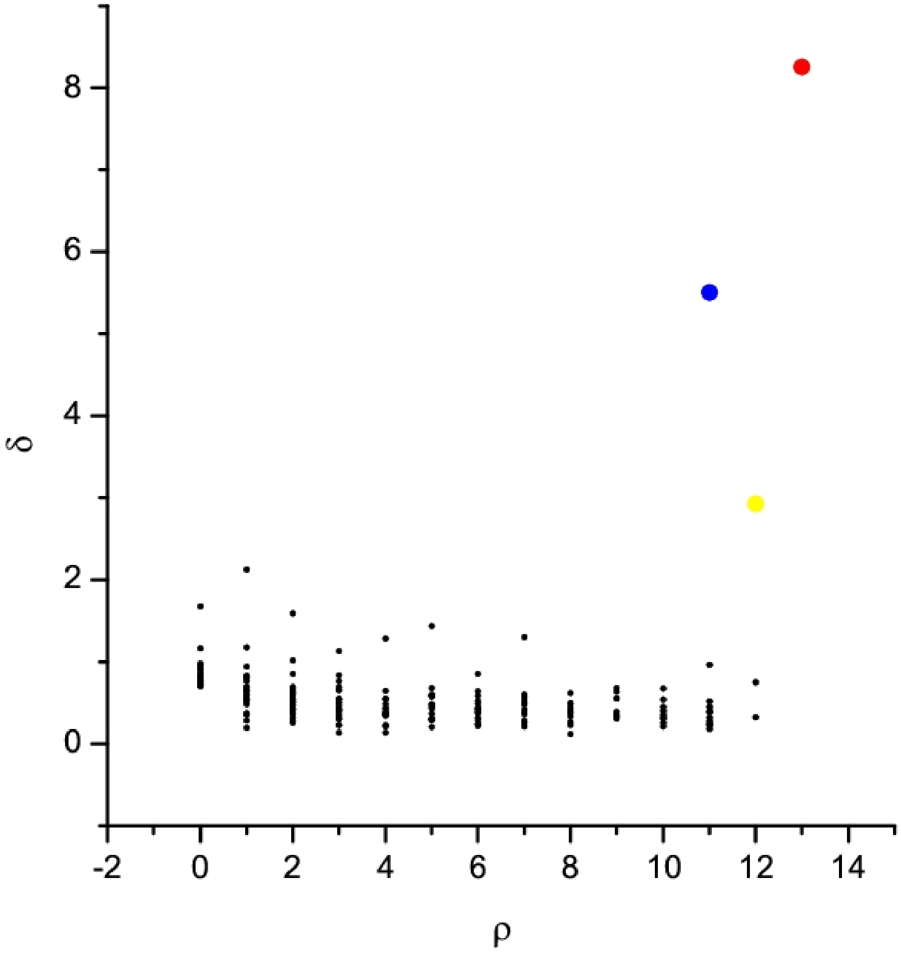

Fig. S7

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

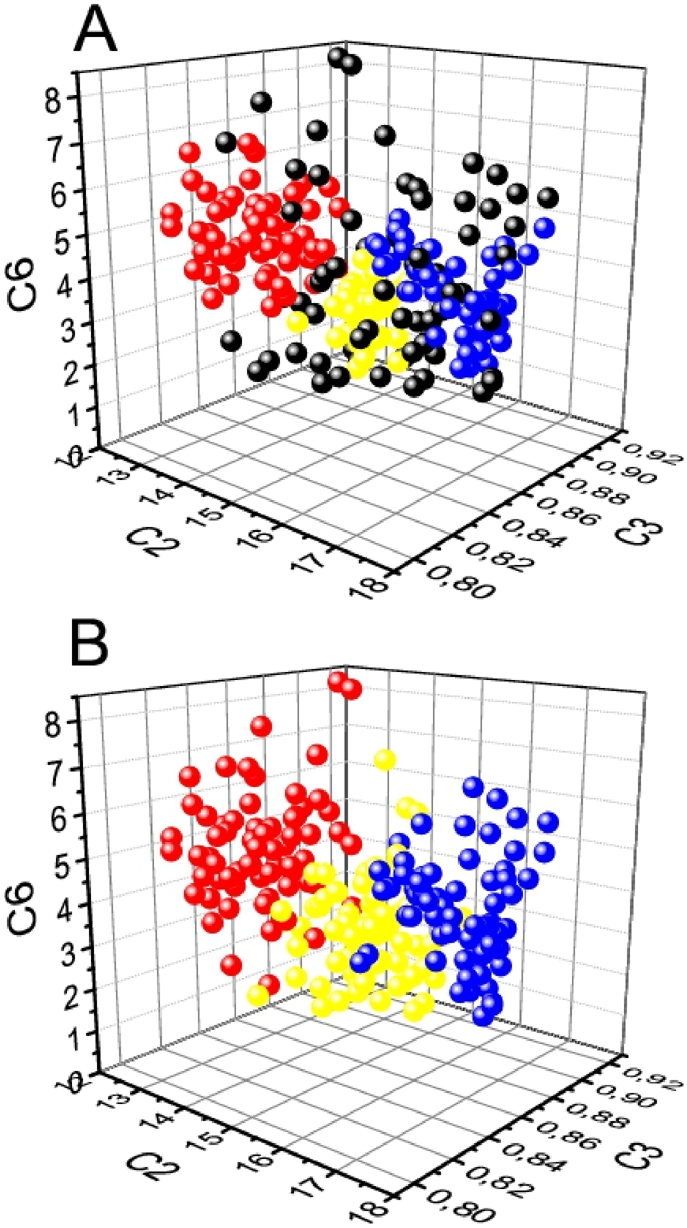

Fig. S8

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.

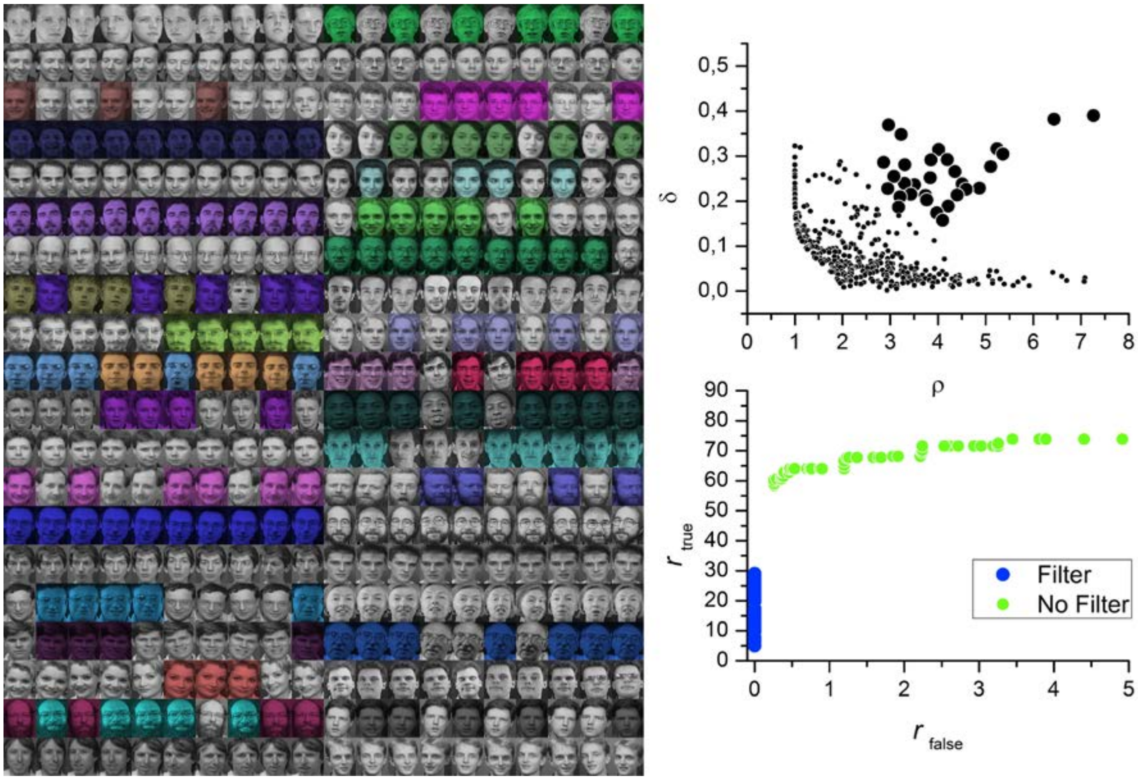

Fig. S9

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.

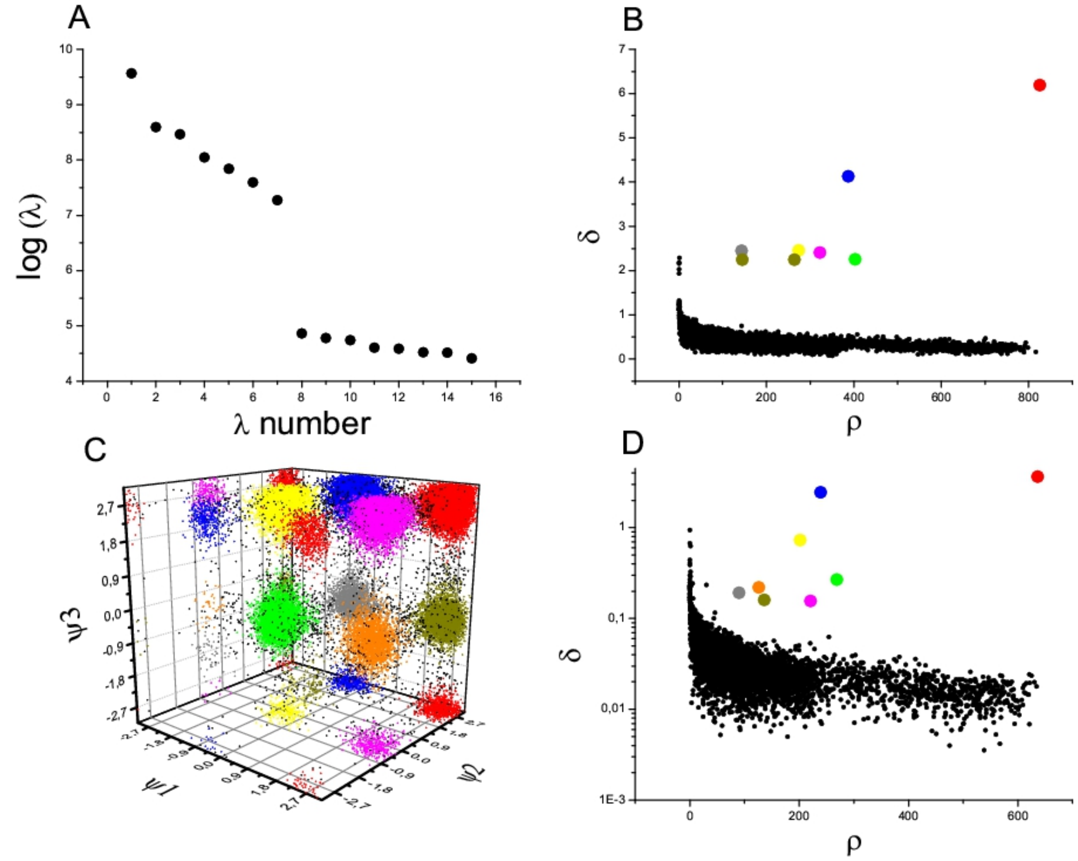

Fig. S10

Cluster analysis of a 2 *μ*s 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.

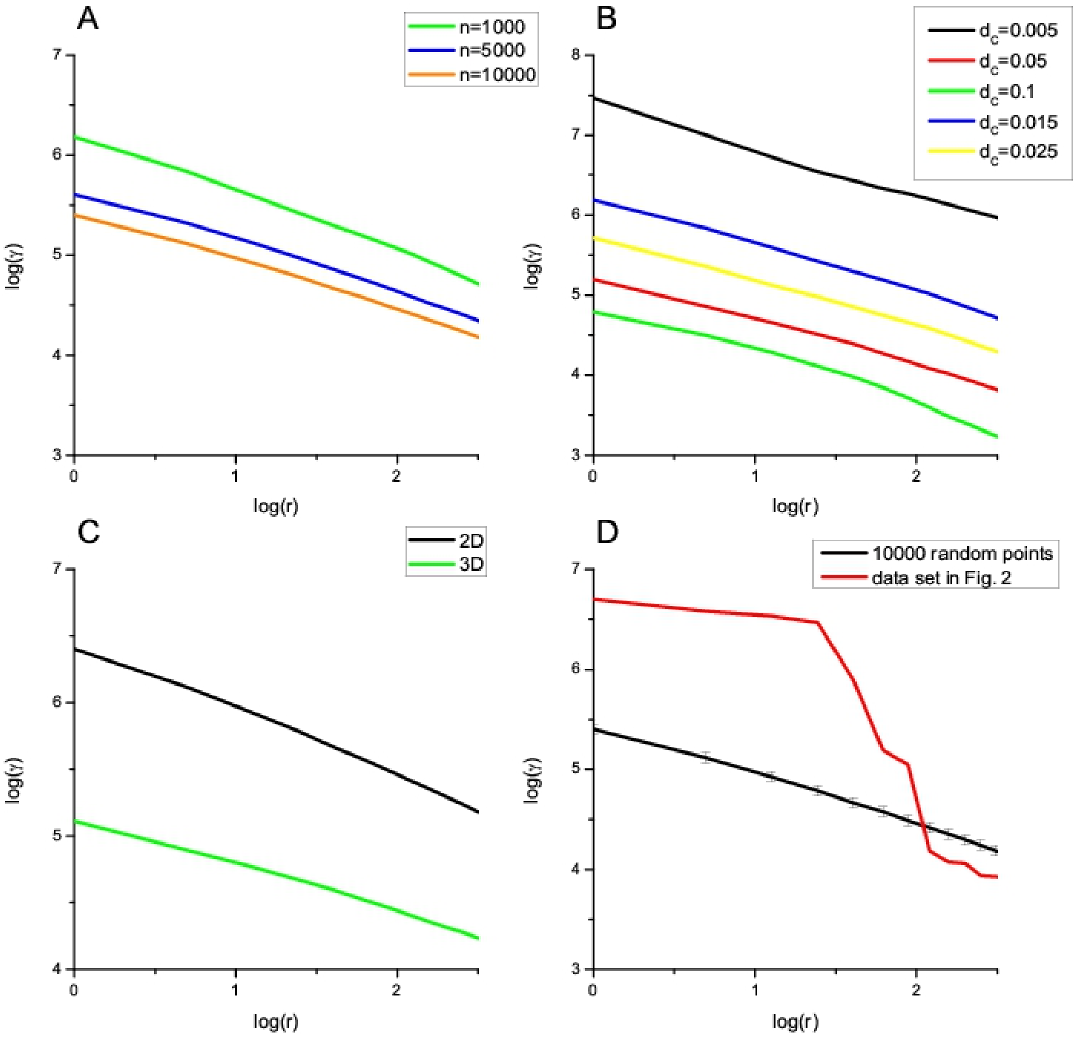

Fig. S11

The value of γ_i= ρ_i δ_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