The 29th Annual International Astronomical Data Analysis Software & Systems (ADASS) Conference

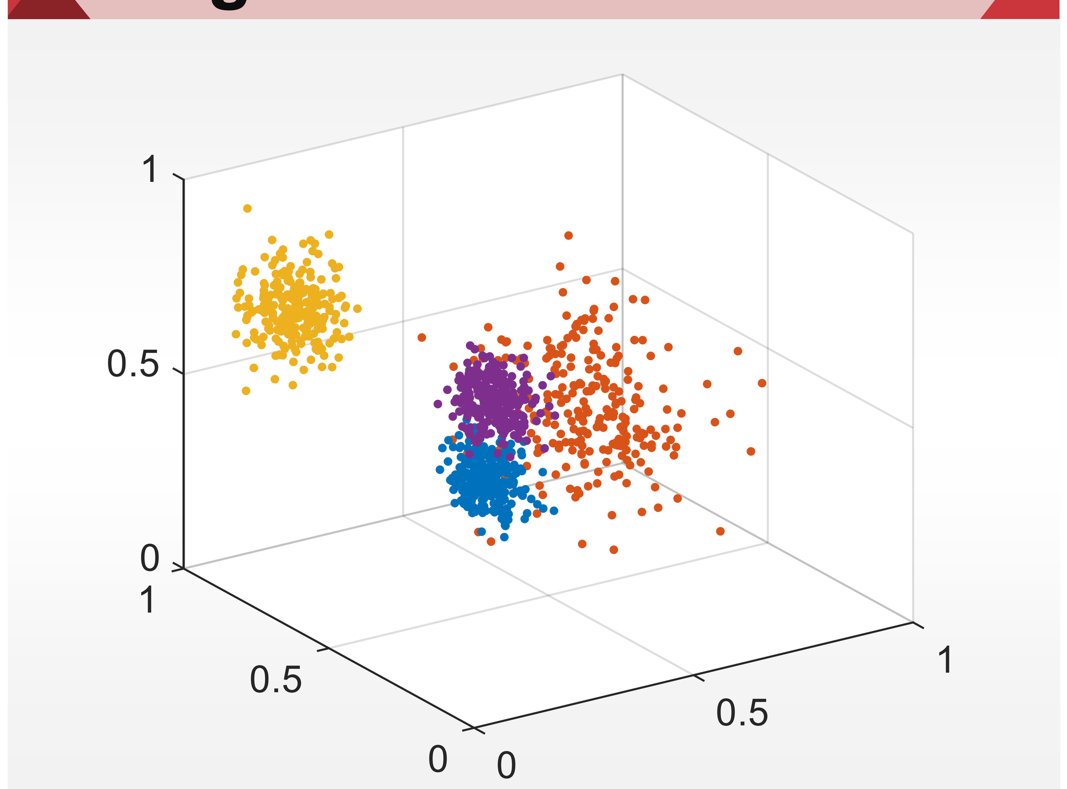
Visualizing High-Dimensional Chemical Abundance Space in GALAH DR2

Youngjoo Kim[1, 2] Scott C. Trager[2] Alexandru C. Telea[1] Jos B. T. M. Roerdink[1]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen
[2]Kapteyn Astronomical Institute, University of Groningen

Aim: Visualize high-dimensional data to find interesting patterns and underlying structures

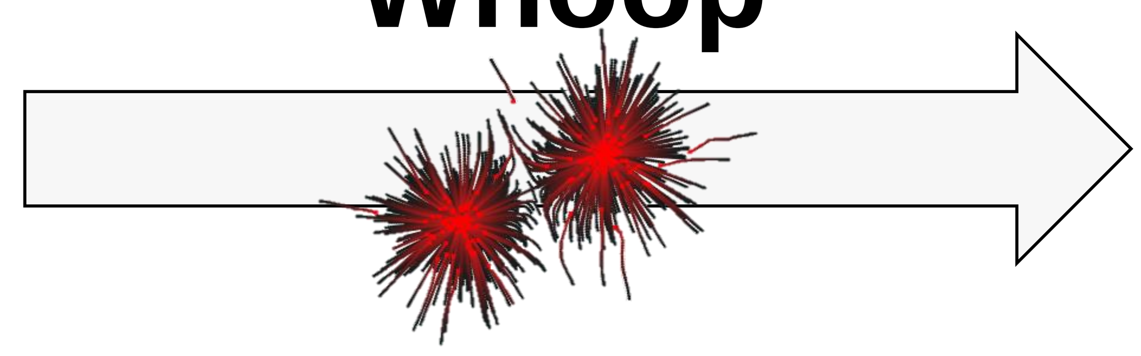
High-Dimensional Data



Gaussian random data with four clusters in 3D (also applicable to *n*D)

Filter high-dimensional data

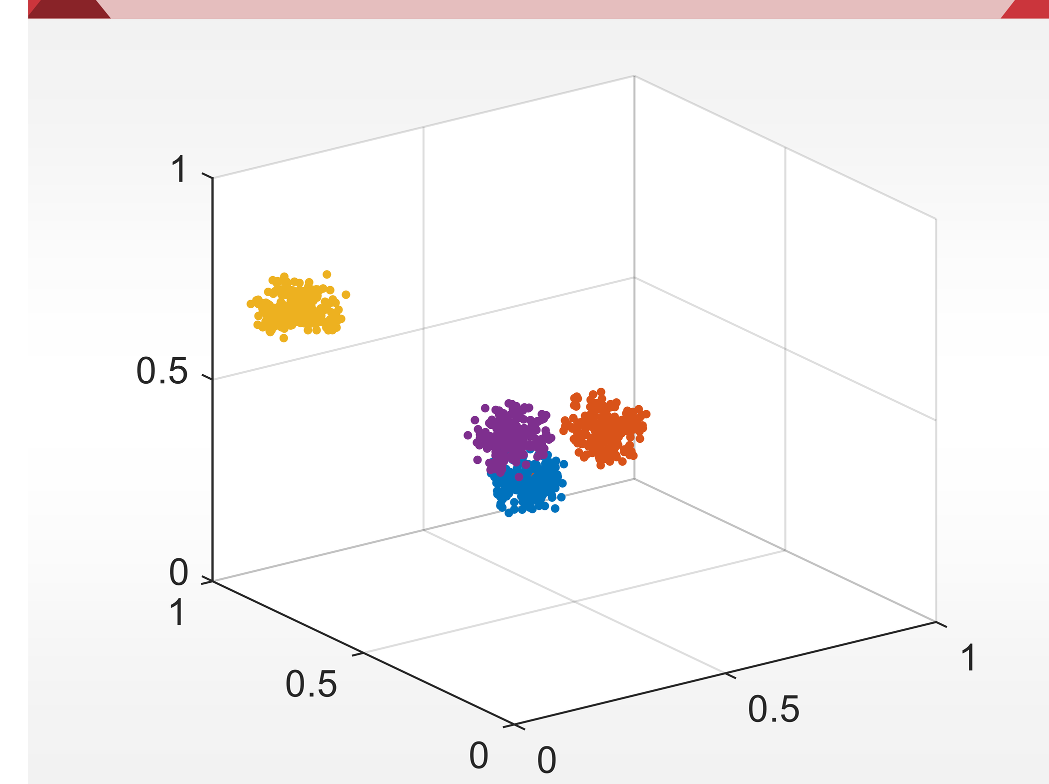
"Whoop"



Local Gradient Clustering (LGC)

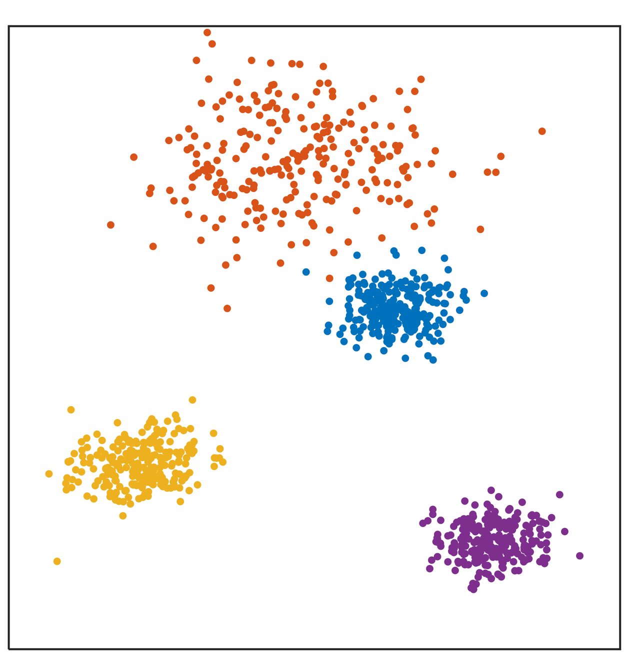
Shift points along the gradient of the kernel density estimator

Filtered Data



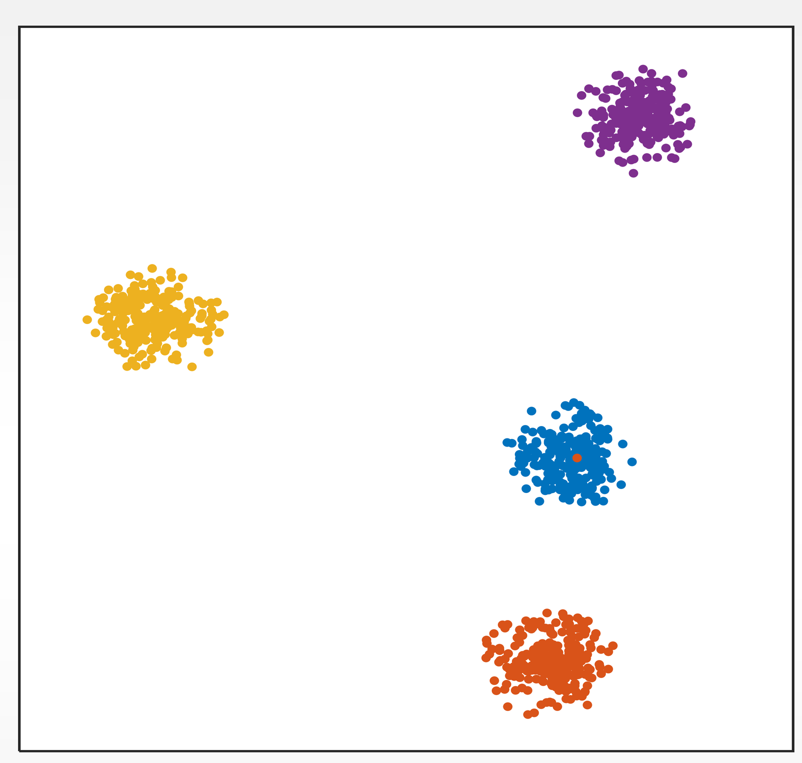
Clusters are separated in 3D

Dimensionality Reduction



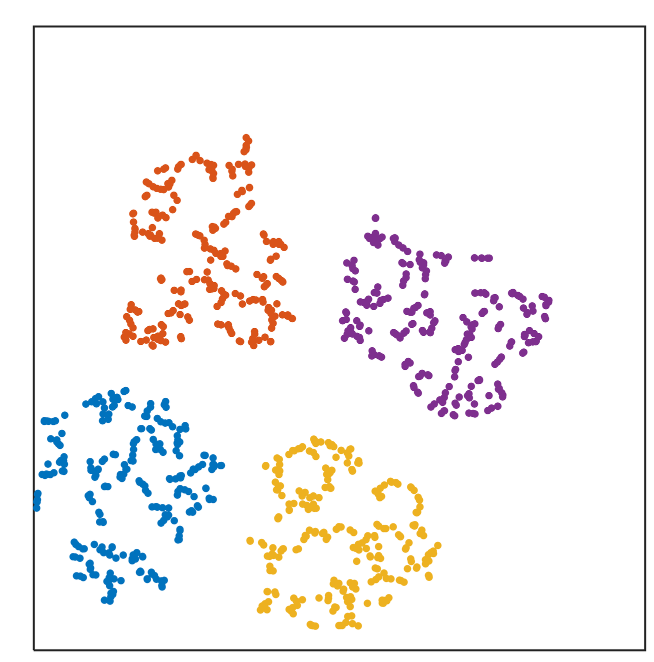
Landmark Multidimensional Scaling (LMDS [1]):
Clusters are **not well separated**.
Method is **fast**.

VS.



Proposed method (LGC+LMDS):
Clusters are **well separated** in the 2D projection.
Method is **fast**.

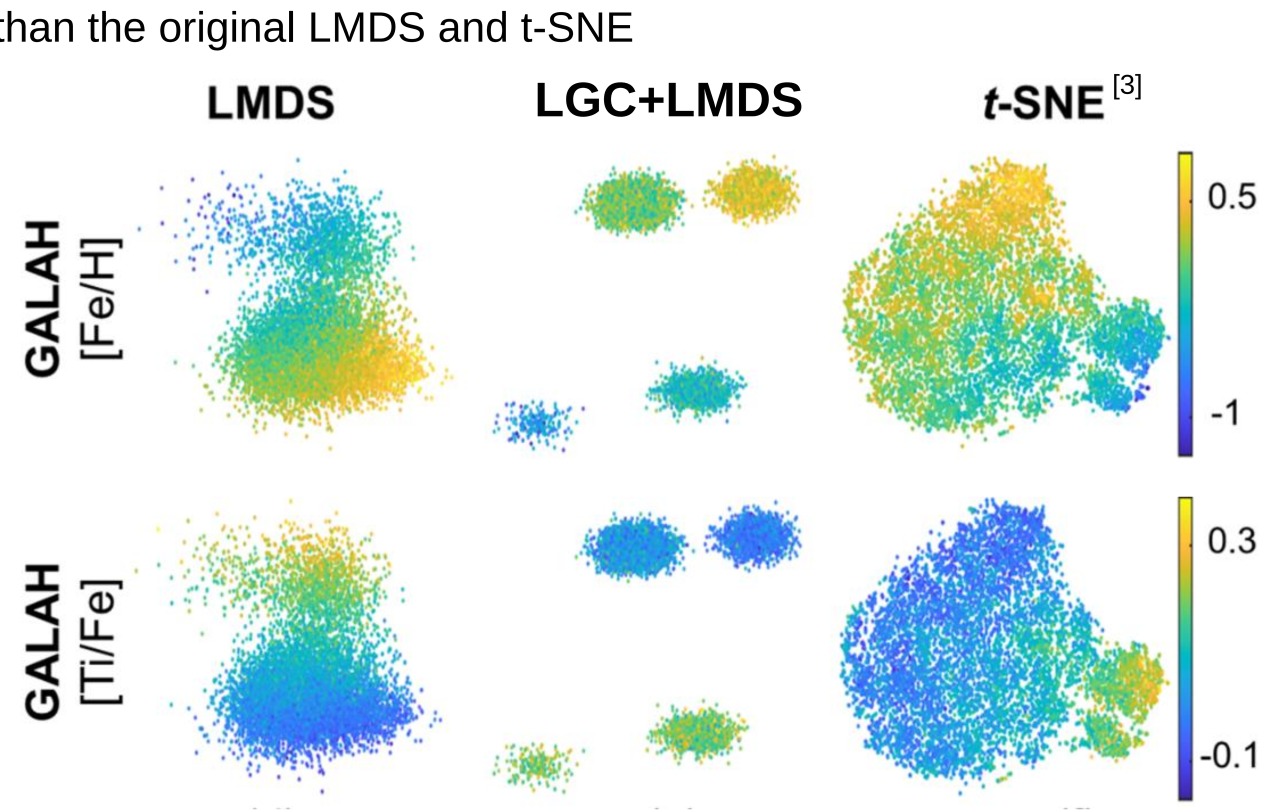
VS.



t-Stochastic Neighbor Embedding (t-SNE [2]):
Clusters are **well separated**.
Method is **slow**.

GALAH DR2

- **Dataset:** 10K observations are randomly chosen from the second data release of GALactic Archaeology with HERMES survey (GALAH DR2) [4] cross-matched with *Gaia* DR2 [5-6]. 10-D data set that consists of the following 10 stellar abundances are used: [Fe/H], [Mg/Fe], [Al/Fe], [Si/Fe], [Ca/Fe], [Ti/Fe], [Cu/Fe], [Zn/Fe], [Y/Fe], and [Ba/Fe]
- **Results:** LGC+LMDS shows cleaner separation of substructures in the 2D abundance-space than the original LMDS and t-SNE



References

[1] V. De Silva and J. B. Tenenbaum, "Sparse multidimensional scaling using landmark points," Technical report, Stanford University, Vol. 120, 2004.
[2] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of machine learning research, No. 9, pp. 2579-2605, 2008.
[3] F. Anders *et al.,* "Dissecting stellar chemical abundance space with t-sne," Astronomy & Astrophysics, Vol. 619, No. A125, 2018.
[4] S. Buder et al., "The GALAH Survey: Second data release," Monthly Notices of the Royal Astronomical Society, Vol. 478, 2018.
[5] Gaia Collaboration. "The Gaia mission," Astronomy & Astrophysics, Vol. 595, No. A1, 2016.
[6] Gaia Collaboration. "Gaia Data Release 2-Summary of the contents and survey properties," Astronomy & Astrophysics, Vol. 616, No. A1, 2018.
[7] M. Muja and D. G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", International Conference on Computer Vision Theory and Applications (VISAPP'09), 2009.
[8] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," Theory of Probability and its Applications, Vol. 14, No.1, pp. 153-158, 1969.
[9] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Transactions on information theory, Vol. 21, No. 1, pp. 32-40, 1975.

Summary

Key idea

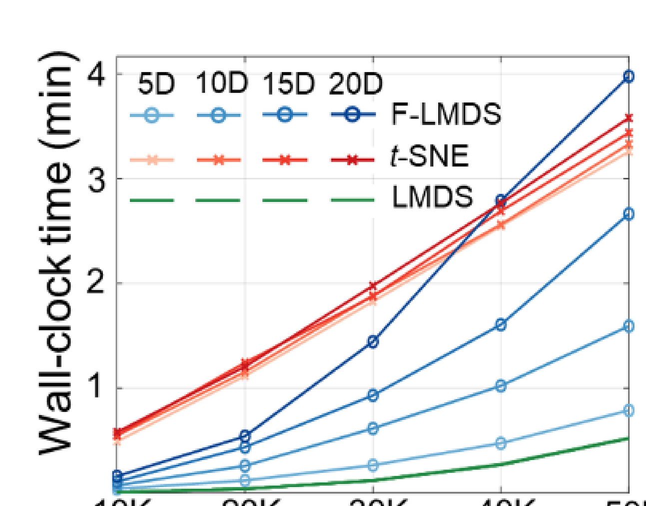
Filter the high-dimensional data so that potential clusters are well separated even after dimensionality reduction

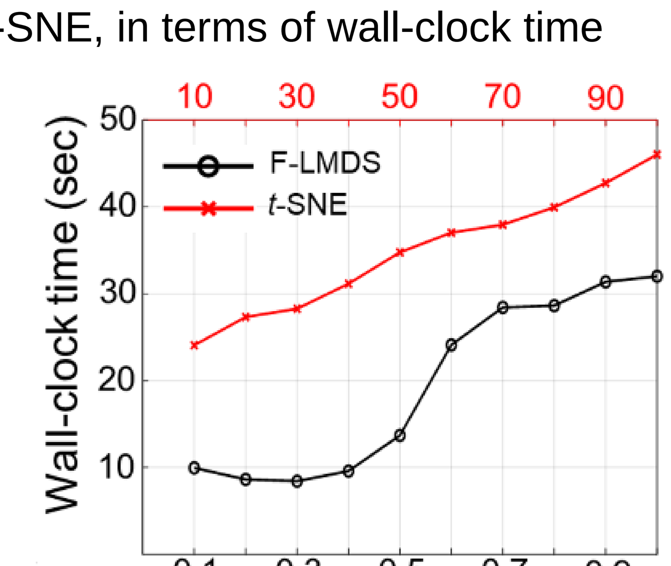
Method

I. Estimate density using Epanechnikov kernel [7-8]
II. Shift points upstream in kernel density gradient, resulting in cluster contraction [9]
III. Perform LMDS [1]

Advantages

- Clusters are **well separated** after the projection by preprocessing the data with local-based gradient clustering
- **Predictable** outcome with one parameter
- More **computationally scalable** than *t*-SNE, in terms of wall-clock time





Future Work

- A more sophisticated analysis of the different substructures gained from the LGC+LMDS results using GALAH DR2

Contact: lyoungjookiml@gmail.com