A MapReduce-based approach for data clustering

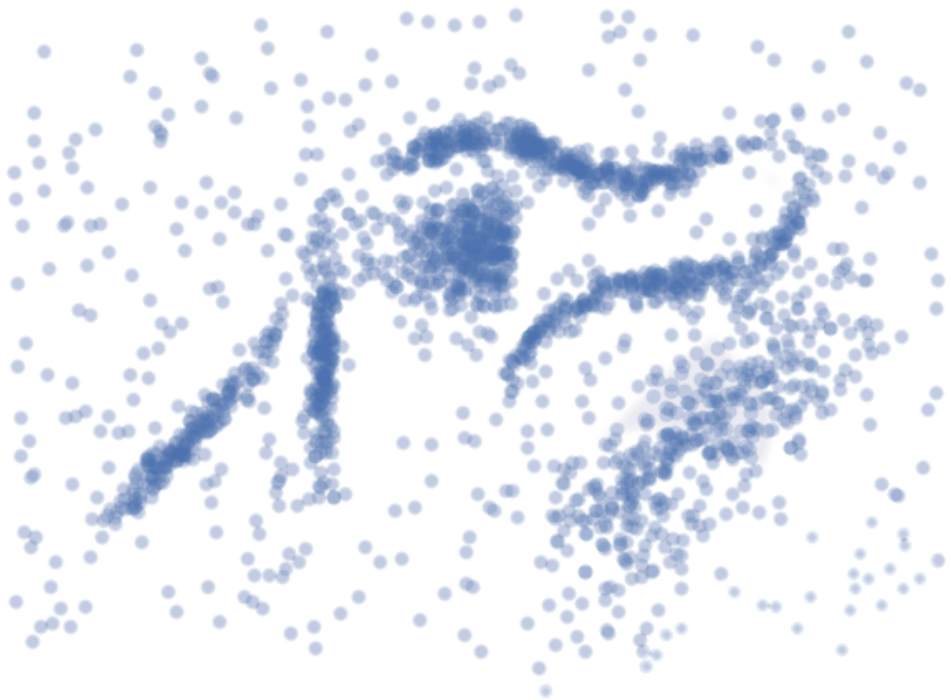
Magdalena Wiercioch

Faculty of Mathematics and Computer Science & Faculty of Physics, Astronomy and Applied Computer Science
Jagiellonian University
e-mail: *magdalena.wiercioch@uj.edu.pl*

SIGML, April 2018

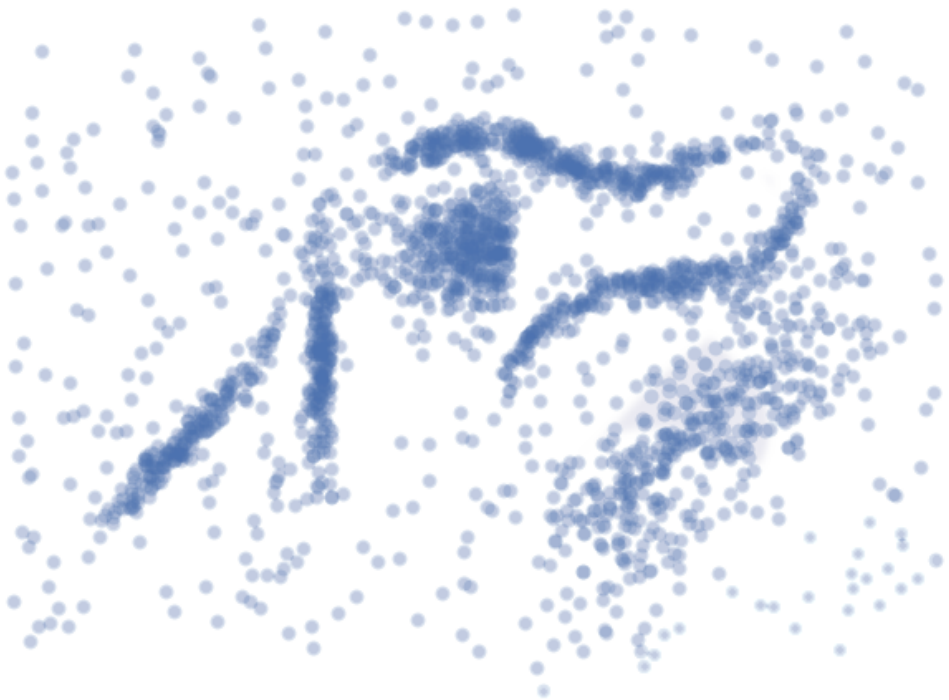
Outline

Problem - big data challenges



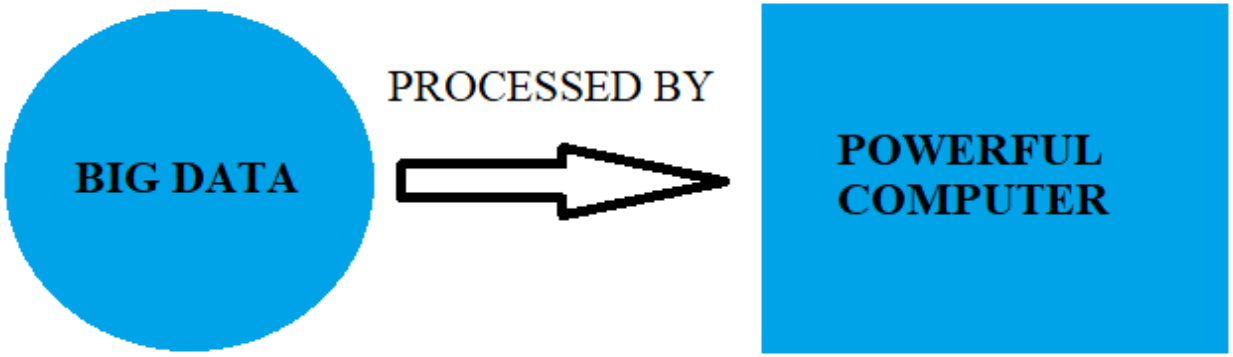
- Find similar regions.

Problem - big data challenges

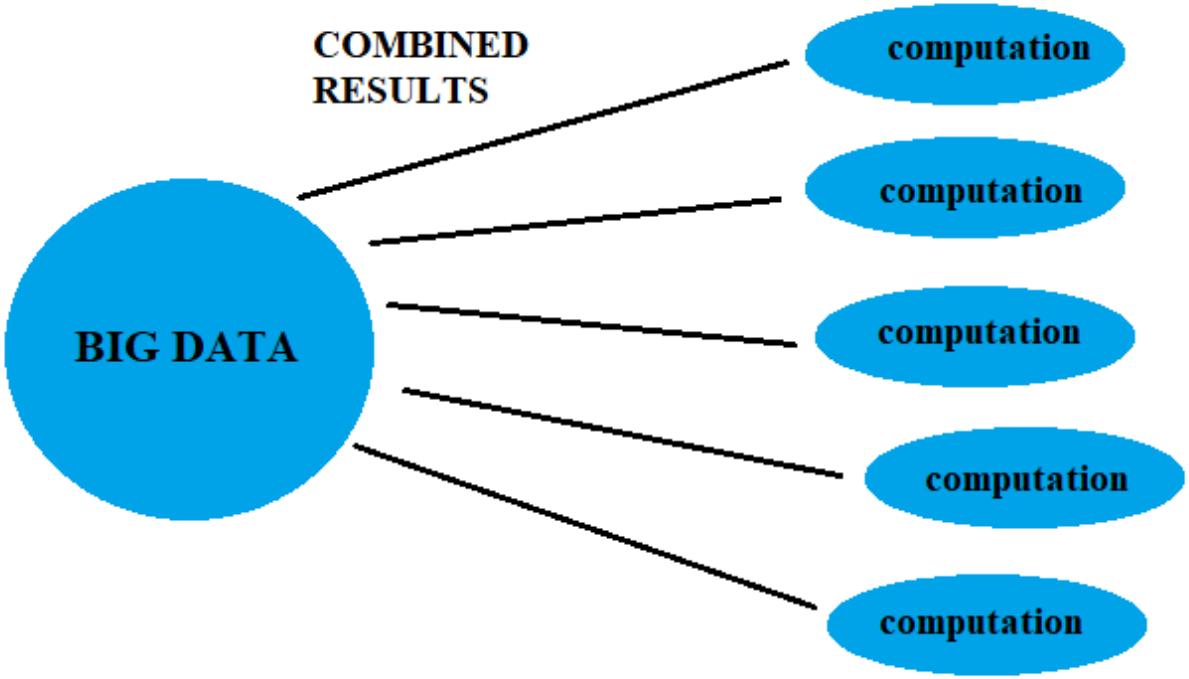


- Find similar regions.
- For big data (velocity, volume, variety).

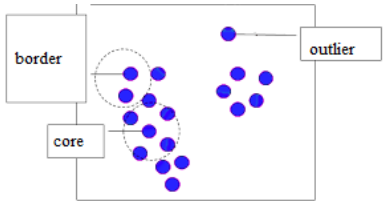
Map Reduce: introduction



Map Reduce: introduction



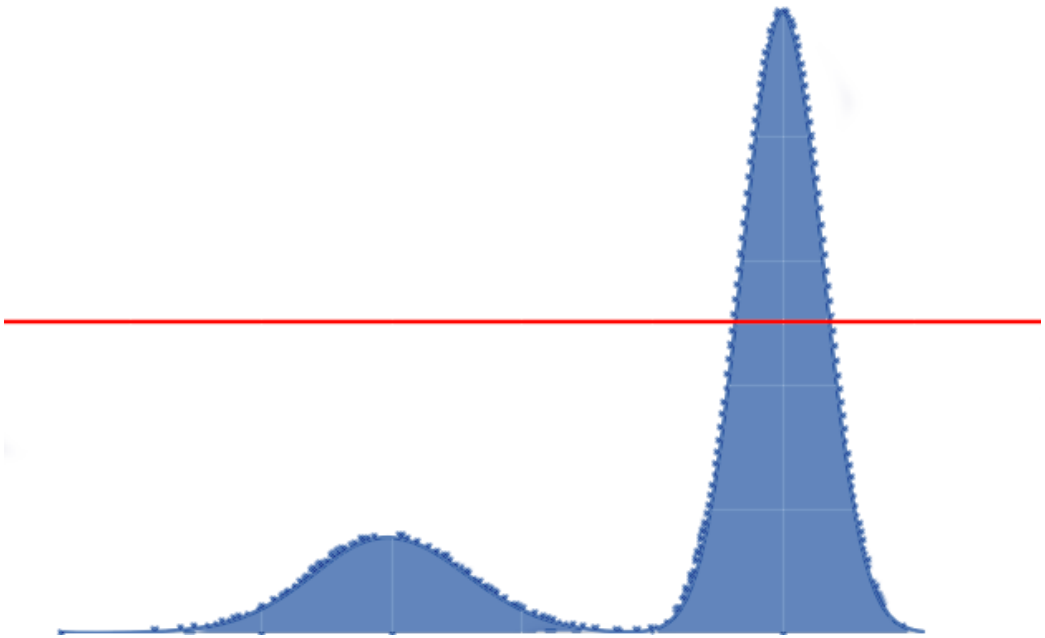
DBSCAN



- We have *dense* clusters of points separated from each other by areas of low density.

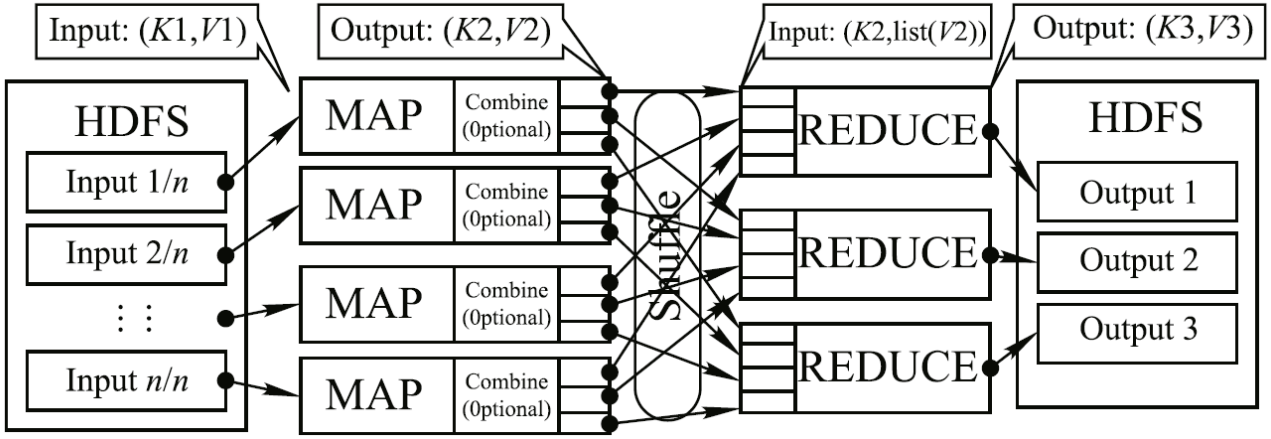
The idea: set a threshold for density g with level p and split the final set into connected components.
$clusters = connectedcomponentsof\{x : g(x) \geq p\}$ The elements with density less than p are taken as outliers.

DBSCAN



- It favors small clusters with high density.
- How to choose the *right* value of p?

Architecture



Conclusions

- We have presented the overview of Map Reduce technique.
- We have shown some disadvantages of DBSCAN.
- Our algorithm is being tested.

Thank you.

References

Yaobin He, Haoyu Tan, Wuman Luo, Shengzhong Feng, and Jianping Fan. *MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data*. Frontiers of Computer Science, 2014.

Xiaojuan Hu and Lei Liu and Ningjia Qiu and Di Yang, and Meng Li. *A MapReduce-based improvement algorithm for DBSCAN*. Journal of Algorithms & Computational Technology, 2018.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. *A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. AAAI Press, 1996.