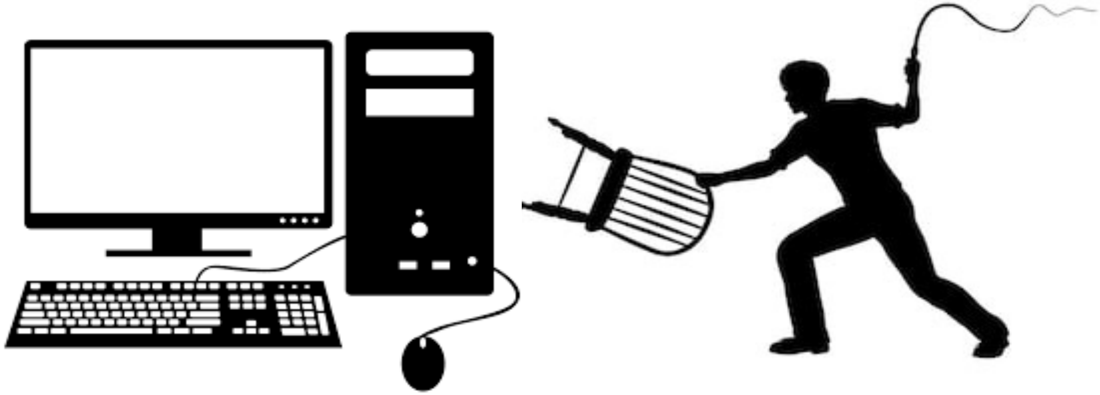
Bioinformatics:

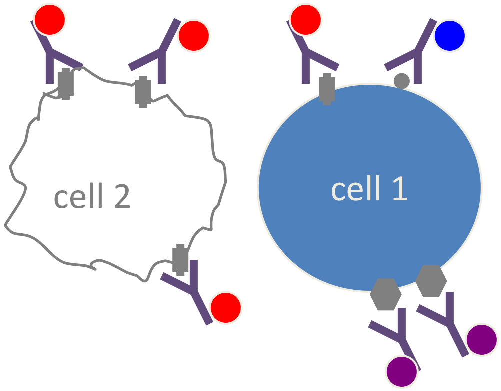
taming computers to make sense out of big data in biology and medicine



Maxime Jan, Julien Dorier, Christian Iseli & Nicolas Guex

BIOENG-519 December 20th, 2024

Flow Cytometry



	FSC	SSC	color1	color2	color3
cell 1	100	100	100	100	200
cell 2	100	500	300	0	0
cell 3	110	100	100	90	220
cell 4	100	510	290	5	0

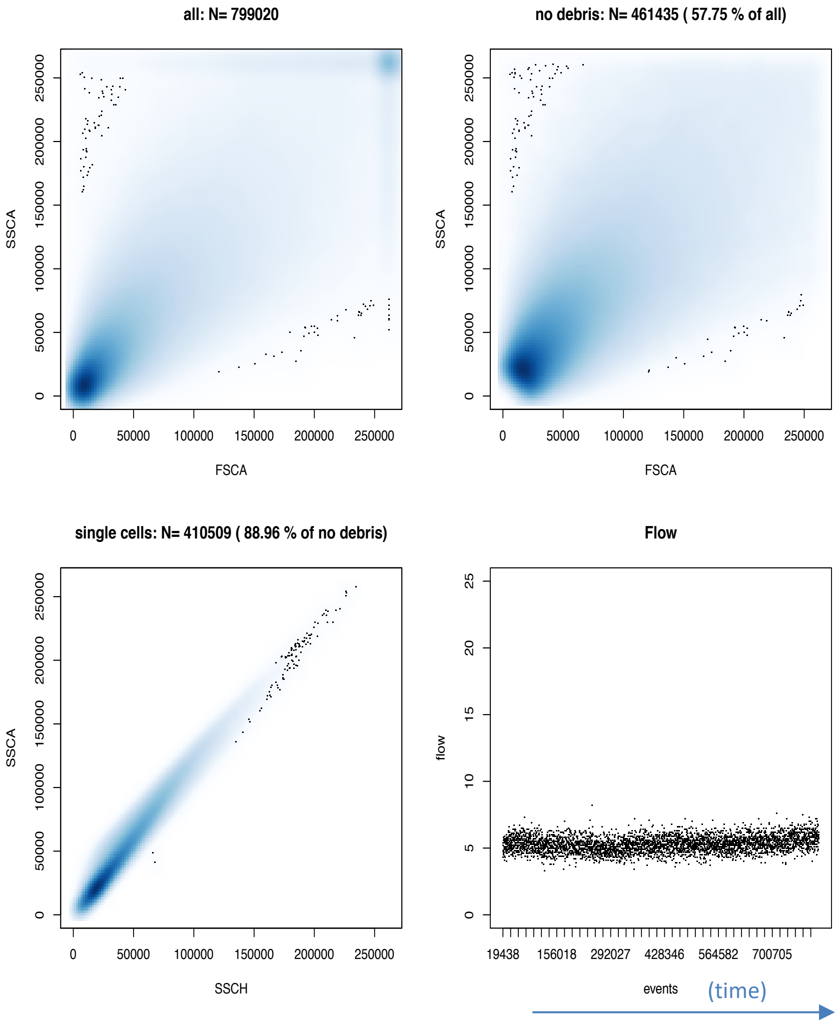
$$\sqrt{10^2+0^2+0^2+10^2+20^2} = 24.5$$

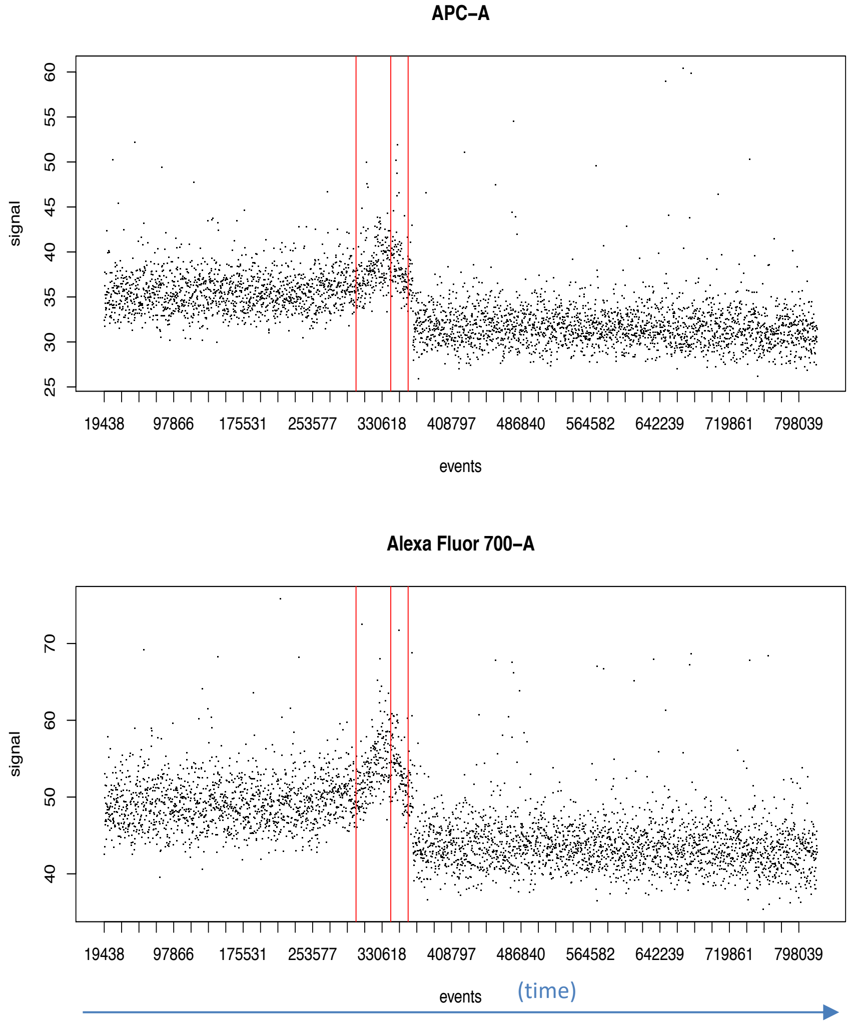
$$\sqrt{10^2+410^2+190^2+95^2+200^2} = 494.3$$

Acquisition QC and filtering

Debris, doublets, flow

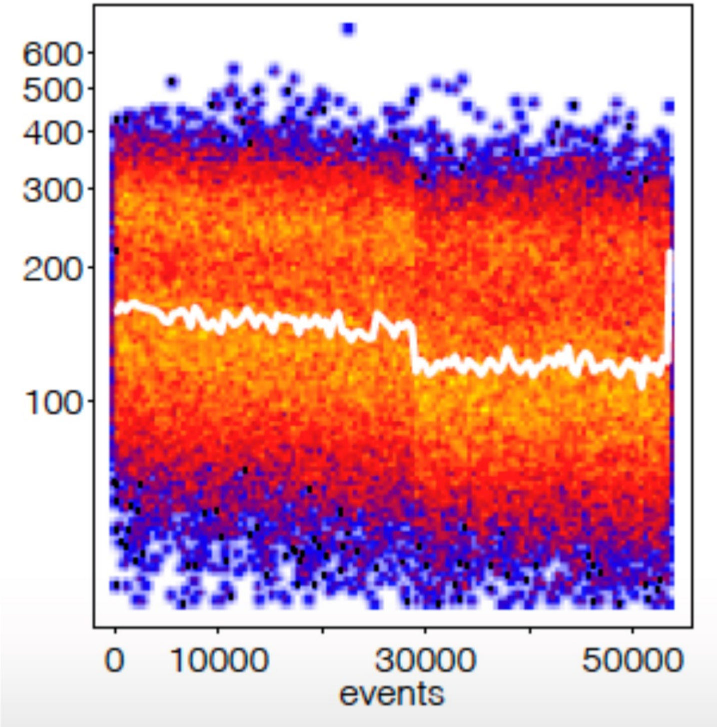
… However, drop in intensity for few channels



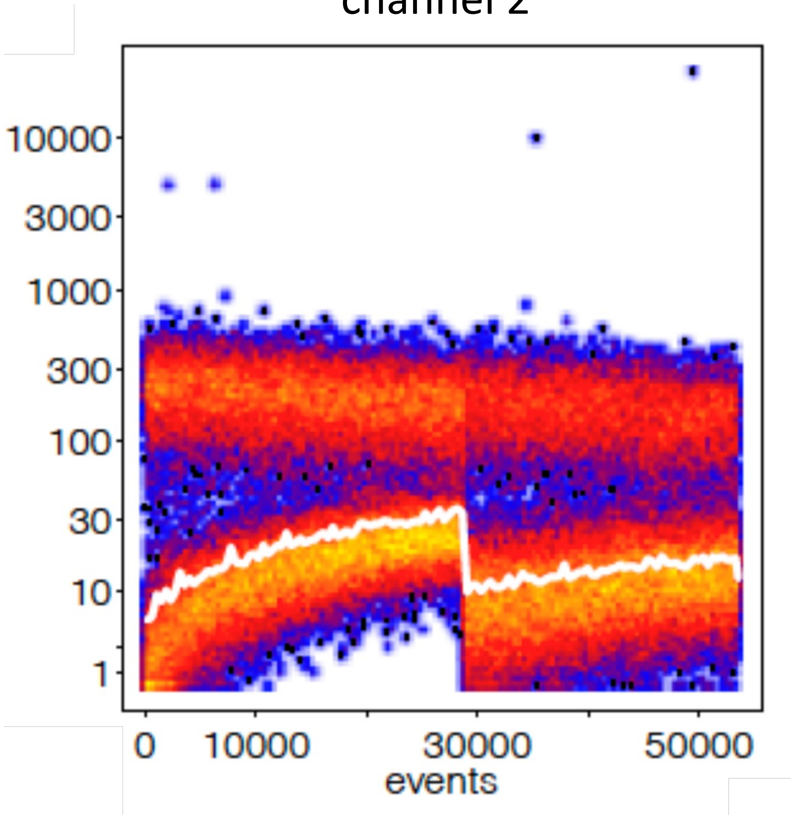


Cytof Drift

channel 1



channel 2

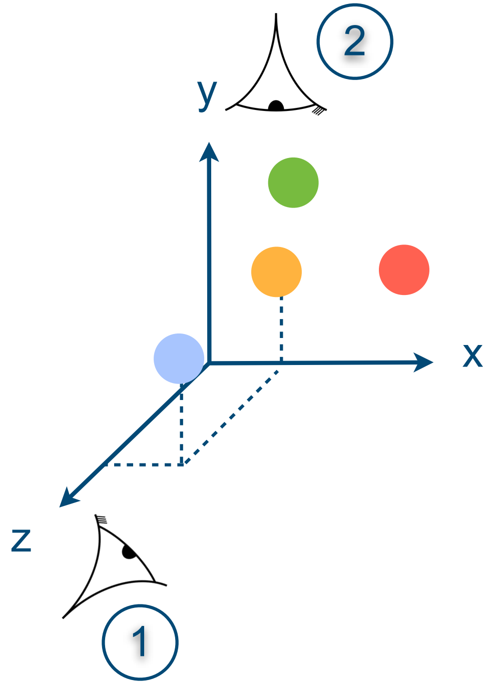


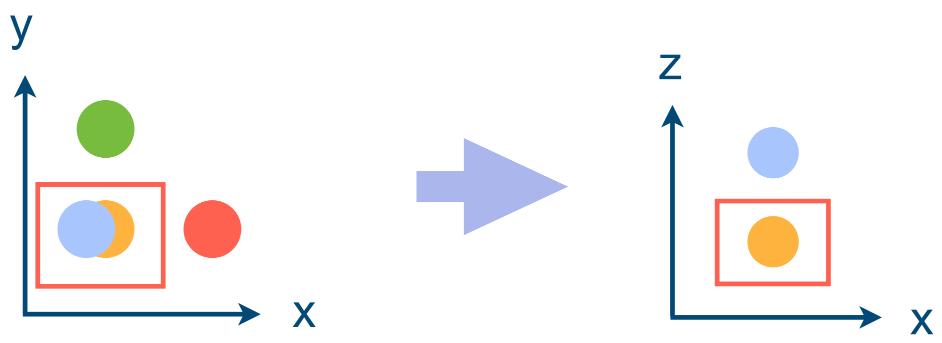
Large scale Data Analysis

Clustering

- applied to flow cytometry
- characteristics of various algorithms

Manual gating uses a sequential approach to address the multidimensionality of the data





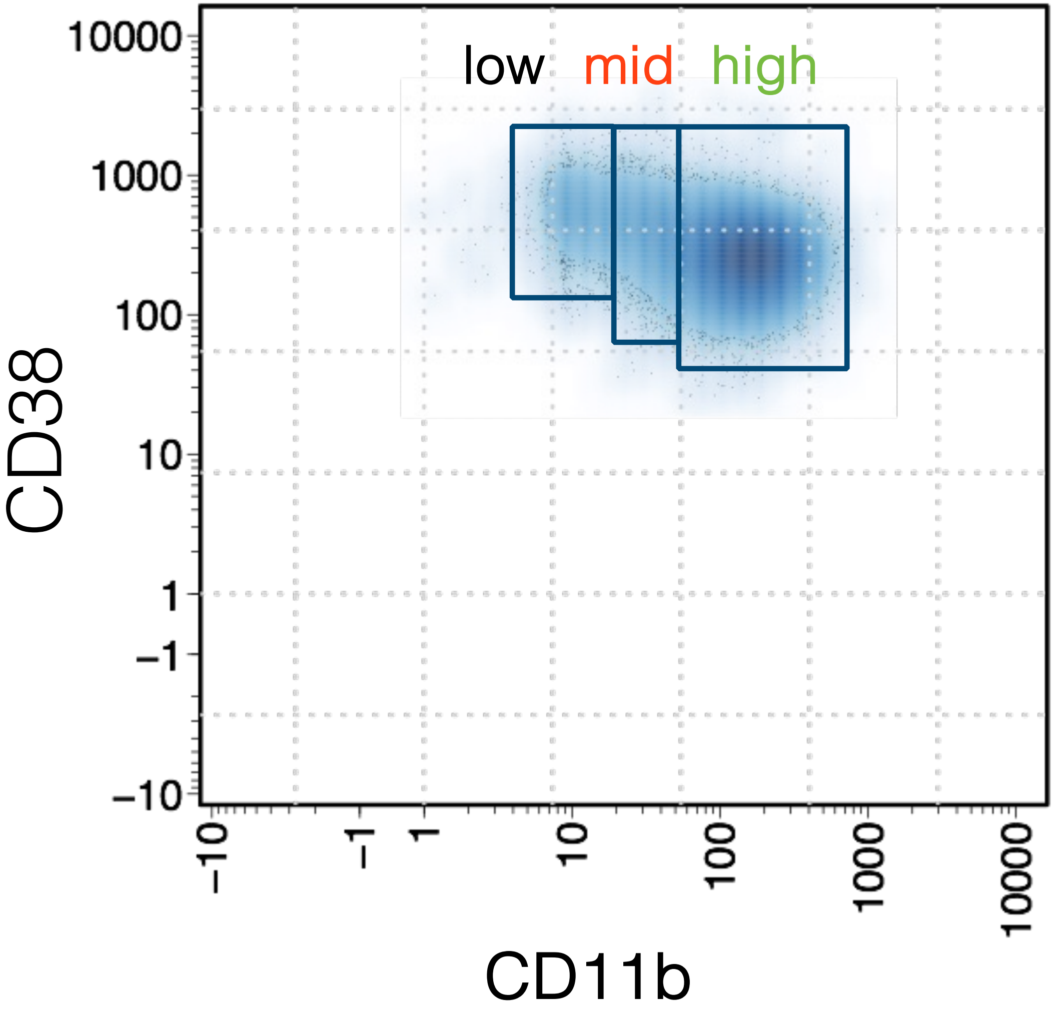
Does not scale well on large data sets (i.e. samples and markers)

- Sequential → inaccuracies propagate and amplify in the downstream steps

Practical advantages of looking at all dimensions (1)

CD11b low/mid/high monocyte (sub)populations at resting state

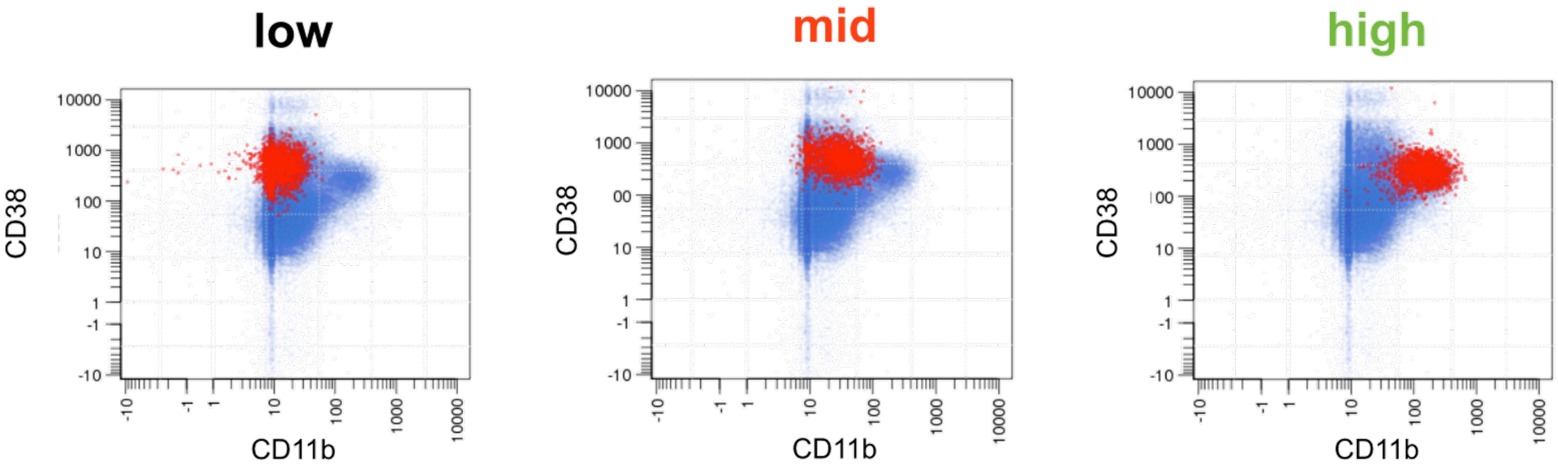
2D gating

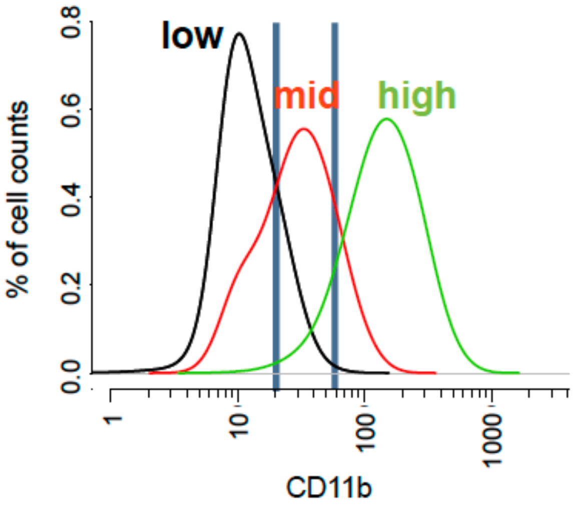


Practical advantages of looking at all dimensions (2)

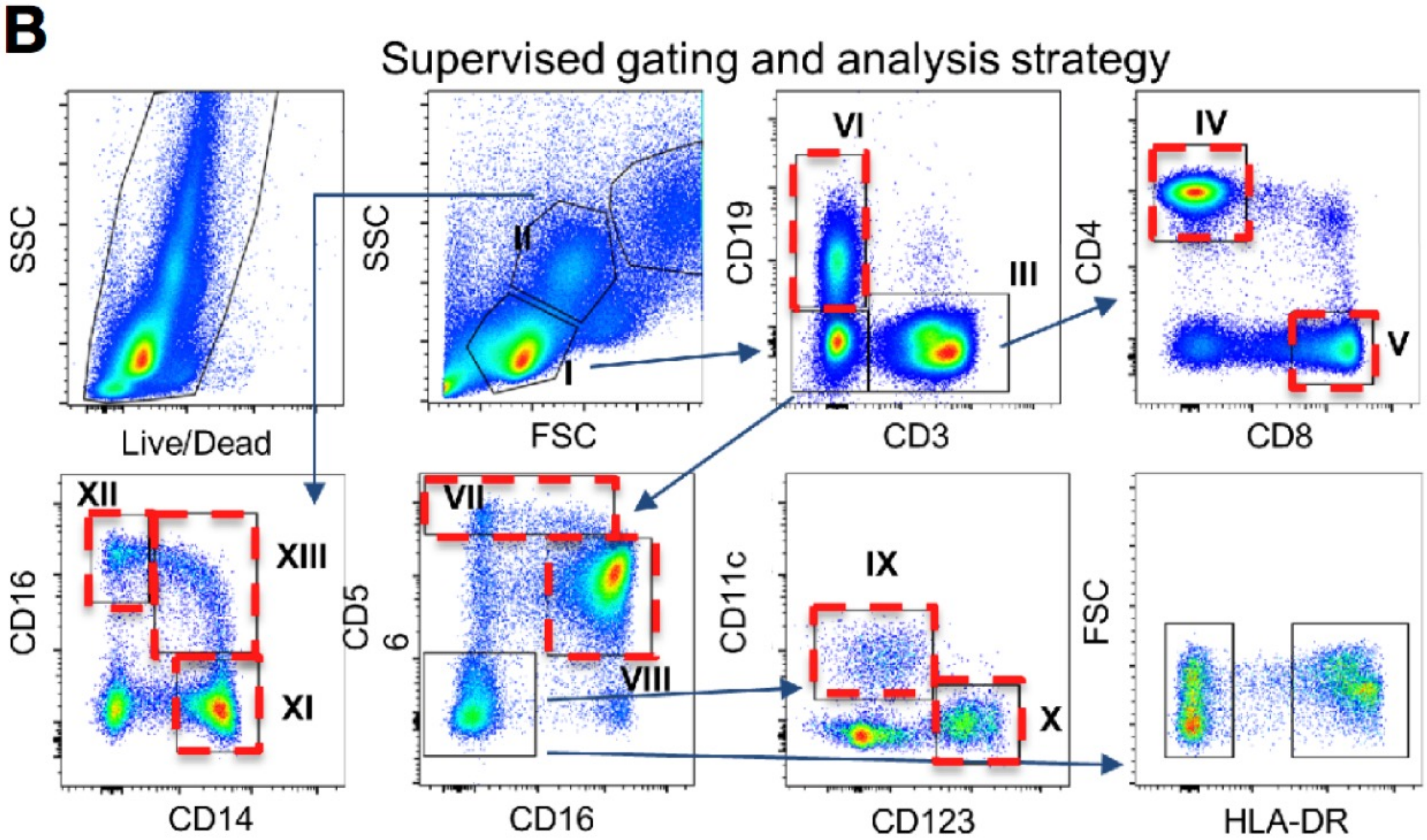
CD11b low/mid/high monocyte (sub)populations at resting state

Multi-dimentional



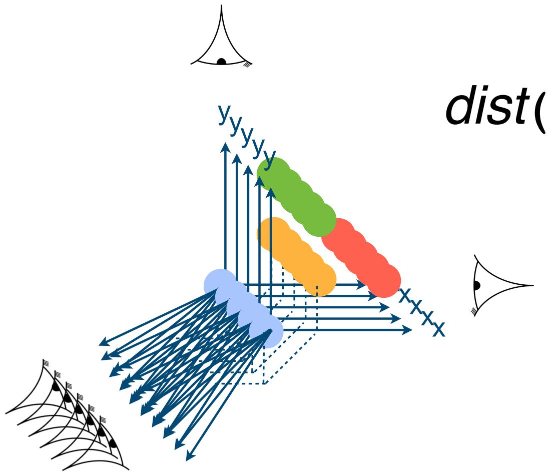


Manual Gating Strategy

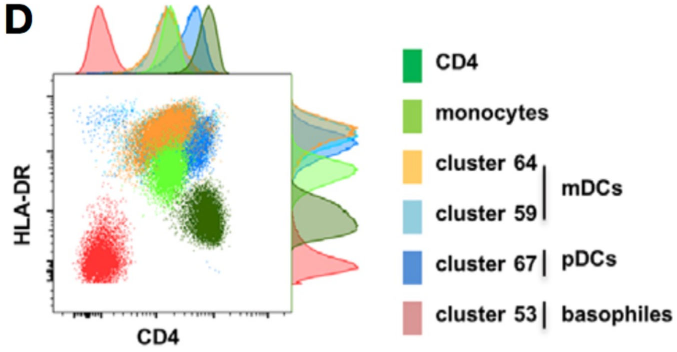


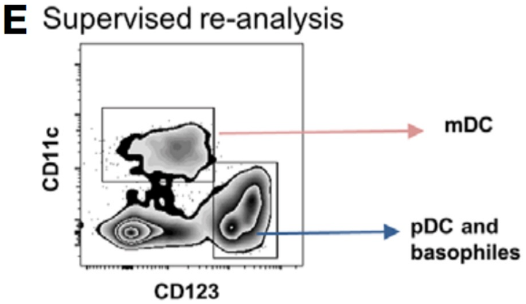
doi: 10.3389/fimmu.2021.633910

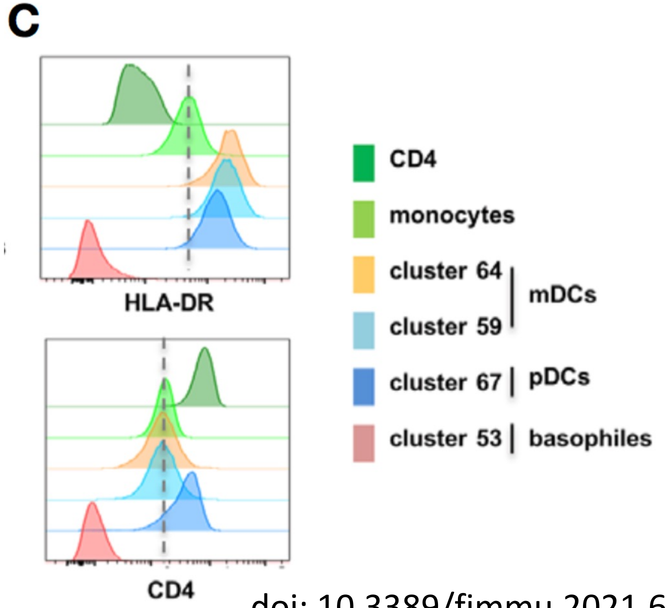
Clustering can consider all data at the same time

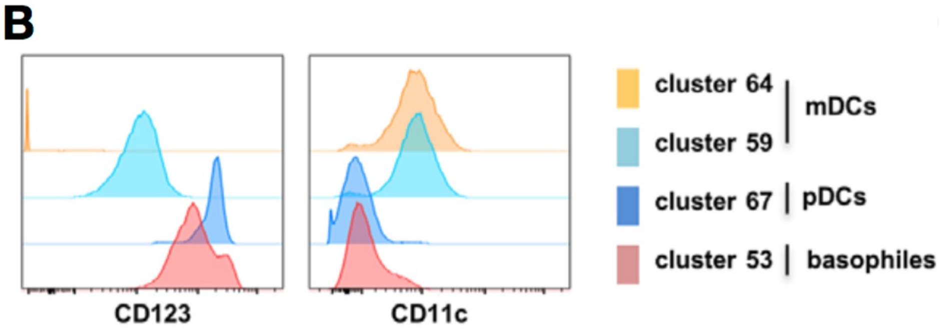


$$dist(\vec{A}, \vec{B}) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$



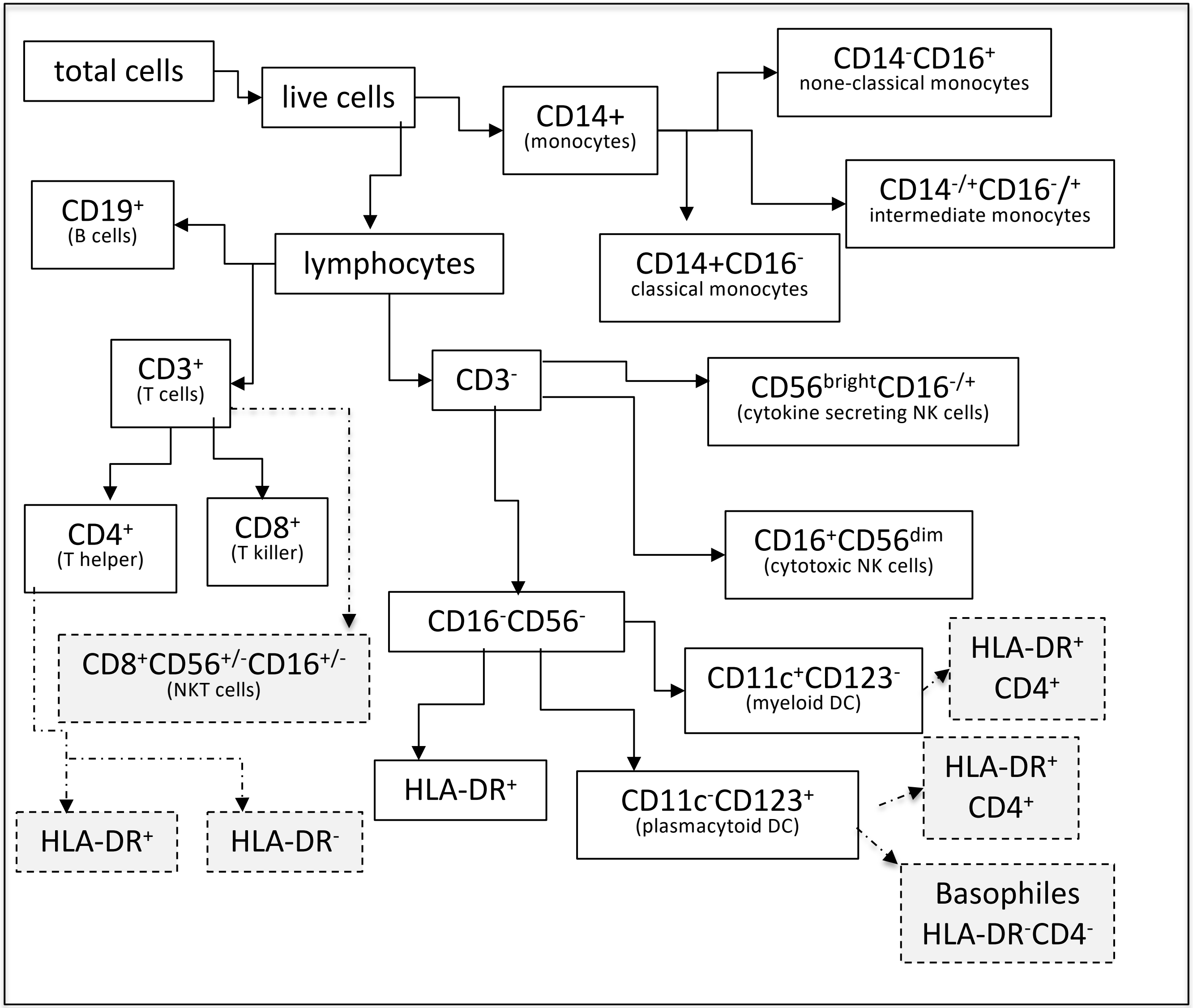






doi: 10.3389/fimmu.2021.633910

Revised gating strategy, considering new populations



total cells

live cells

CD14+
(monocytes)

CD14$^-$CD16$^+$
none-classical monocytes

CD14$^{-/+}$CD16$^-$/$^+$
intermediate monocytes

CD14+CD16$^-$
classical monocytes

CD19$^+$
(B cells)

lymphocytes

CD3$^+$
(T cells)

CD3$^-$

CD56^{bright}CD16$^{-/+}$
(cytokine secreting NK cells)

CD4$^+$
(T helper)

CD8$^+$
(T killer)

CD16$^+$CD56^{dim}
(cytotoxic NK cells)

CD16$^-$CD56$^-$

CD8$^+$CD56$^{+/-}$CD16$^{+/-}$
(NKT cells)

CD11c$^+$CD123$^-$
(myeloid DC)

HLA-DR$^+$
CD4$^+$

HLA-DR$^+$

CD11c$^-$CD123$^+$
(plasmacytoid DC)

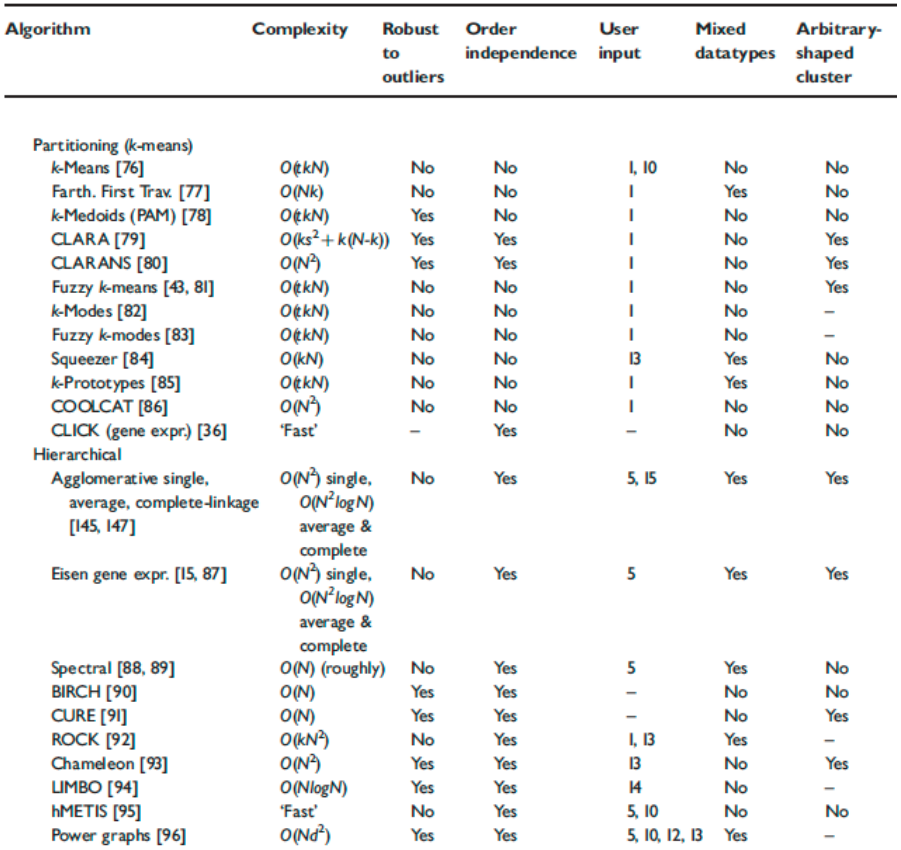
HLA-DR$^+$
CD4$^+$

HLA-DR$^+$

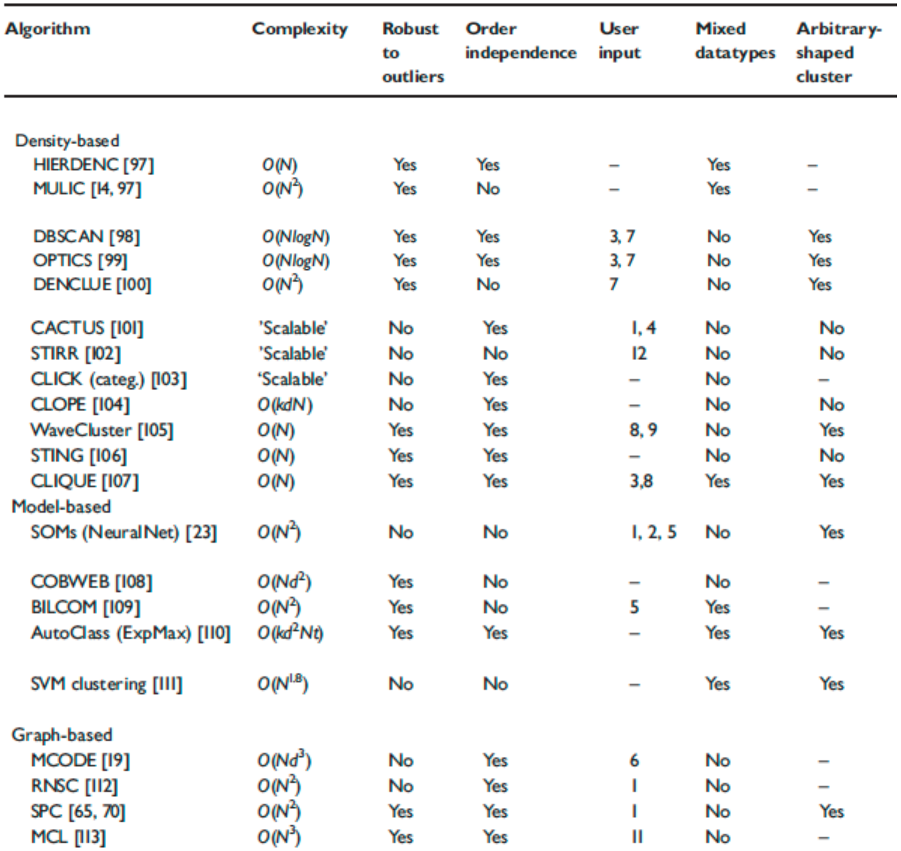
HLA-DR$^-$

Basophiles
HLA-DR$^-$CD4$^-$

Clustering Algorithms



Algorithm	Complexity	Robust to outliers	Order independence	User input	Mixed datatypes	Arbitrary-shaped cluster
Partitioning (*k*-means)						
k-Means [76]	$O(tkN)$	No	No	1, 10	No	No
Farth. First Trav. [77]	$O(Nk)$	No	No	1	Yes	No
k-Medoids (PAM) [78]	$O(tkN)$	Yes	No	1	No	No
CLARA [79]	$O(ks^2+k(N-k))$	Yes	Yes	1	No	Yes
CLARANS [80]	$O(N^2)$	Yes	Yes	1	No	Yes
Fuzzy *k*-means [43, 81]	$O(tkN)$	No	No	1	No	Yes
k-Modes [82]	$O(tkN)$	No	No	1	No	–
Fuzzy *k*-modes [83]	$O(tkN)$	No	No	1	No	–
Squeezer [84]	$O(kN)$	No	No	13	Yes	No
k-Prototypes [85]	$O(tkN)$	No	No	1	Yes	No
COOLCAT [86]	$O(N^2)$	No	No	1	No	No
CLICK (gene expr.) [36]	'Fast'	–	Yes	–	No	No
Hierarchical						
Agglomerative single, average, complete-linkage [145, 147]	$O(N^2)$ single, $O(N^2 logN)$ average & complete	No	Yes	5, 15	Yes	Yes
Eisen gene expr. [15, 87]	$O(N^2)$ single, $O(N^2 logN)$ average & complete	No	Yes	5	Yes	Yes
Spectral [88, 89]	$O(N)$ (roughly)	No	Yes	5	Yes	No
BIRCH [90]	$O(N)$	Yes	Yes	–	No	No
CURE [91]	$O(N)$	Yes	Yes	–	No	Yes
ROCK [92]	$O(kN^2)$	No	Yes	1, 13	Yes	–
Chameleon [93]	$O(N^2)$	Yes	Yes	13	No	Yes
LIMBO [94]	$O(NlogN)$	Yes	Yes	14	No	–
hMETIS [95]	'Fast'	No	Yes	5, 10	No	No
Power graphs [96]	$O(Nd^2)$	Yes	Yes	5, 10, 12, 13	Yes	–



Algorithm	Complexity	Robust to outliers	Order independence	User input	Mixed datatypes	Arbitrary-shaped cluster
Density-based						
HIERDENC [97]	$O(N)$	Yes	Yes	–	Yes	–
MULIC [14, 97]	$O(N^2)$	Yes	No	–	Yes	–
DBSCAN [98]	$O(NlogN)$	Yes	Yes	3, 7	No	Yes
OPTICS [99]	$O(NlogN)$	Yes	Yes	3, 7	No	Yes
DENCLUE [100]	$O(N^2)$	Yes	No	7	No	Yes
CACTUS [101]	'Scalable'	No	Yes	1, 4	No	No
STIRR [102]	'Scalable'	No	No	12	No	No
CLICK (categ.) [103]	'Scalable'	No	Yes	–	No	–
CLOPE [104]	$O(kdN)$	No	Yes	–	No	No
WaveCluster [105]	$O(N)$	Yes	Yes	8, 9	No	Yes
STING [106]	$O(N)$	Yes	Yes	–	No	No
CLIQUE [107]	$O(N)$	Yes	Yes	3,8	Yes	Yes
Model-based						
SOMs (NeuralNet) [23]	$O(N^2)$	No	No	1, 2, 5	No	Yes
COBWEB [108]	$O(Nd^2)$	Yes	No	–	No	–
BILCOM [109]	$O(N^2)$	Yes	No	5	Yes	–
AutoClass (ExpMax) [110]	$O(kd^2Nt)$	Yes	Yes	–	Yes	Yes
SVM clustering [111]	$O(N^{1.8})$	No	No	–	Yes	Yes
Graph-based						
MCODE [19]	$O(Nd^3)$	No	Yes	6	No	–
RNSC [112]	$O(N^2)$	No	Yes	1	No	–
SPC [65, 70]	$O(N^2)$	Yes	Yes	1	No	Yes
MCL [113]	$O(N^3)$	Yes	Yes	11	No	–

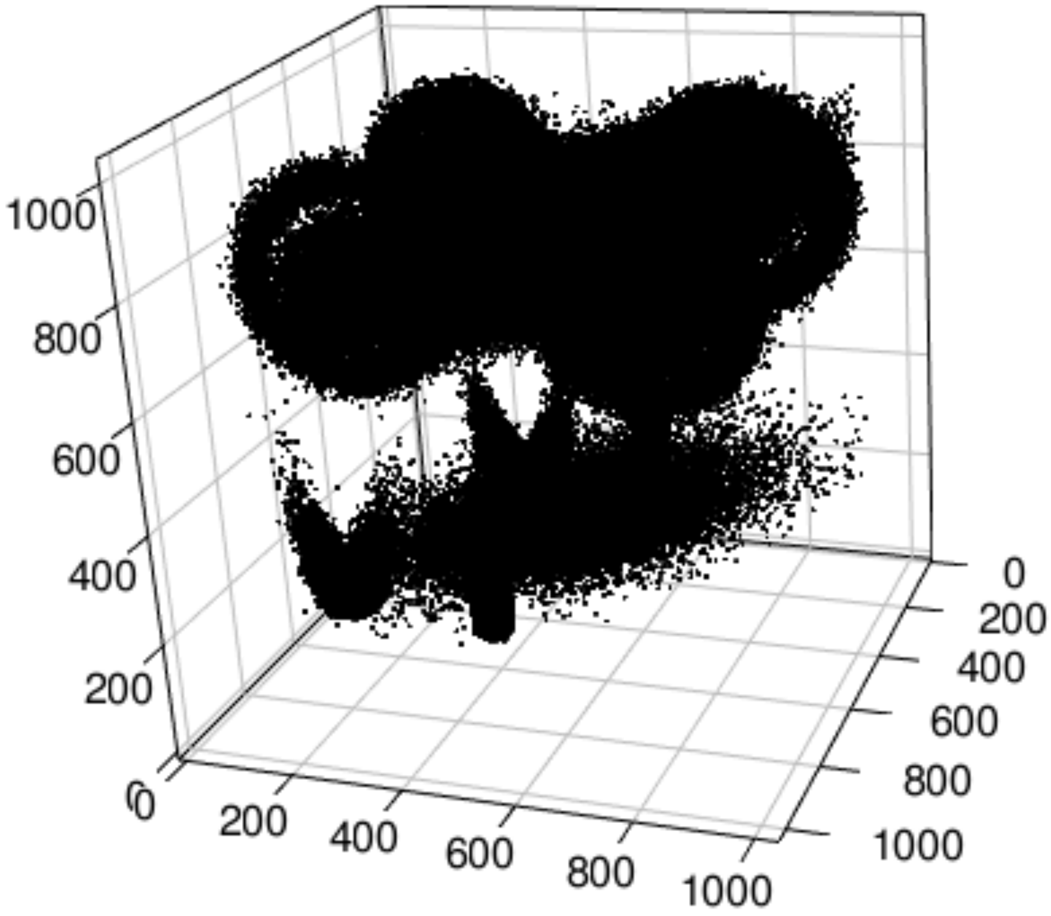
Source: A roadmap of clustering algorithms: finding a match for a biomedical application
B.Andreopoulos et al; Briefings in Bioinformatics (2009) VOL 10. NO 3. 297-314

Clustering Wishlist

- Unaffected by the order in which the data are presented
- Not assume any specific cluster shape
- Proper separation of overlapping distributions
- Automatically discover the “ideal” number of clusters
- Resistant to noise (e.g. not assign outliers in clusters)
- Capable of clustering millions of observations

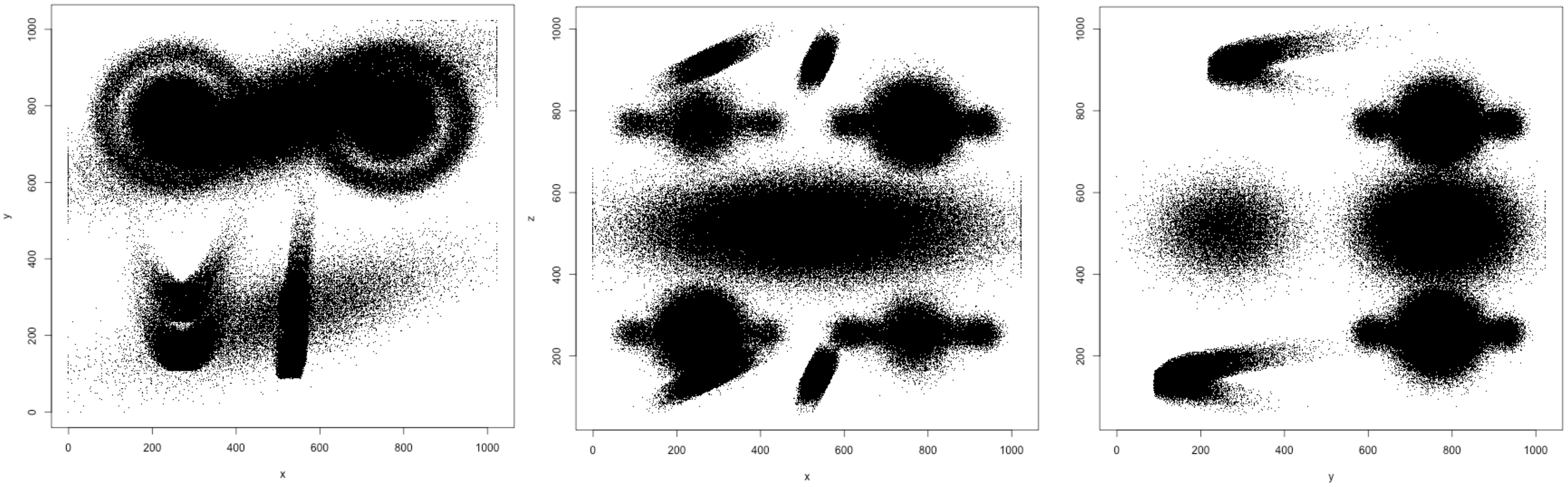
Tests with synthetic data

comprising a total of 799,956 points



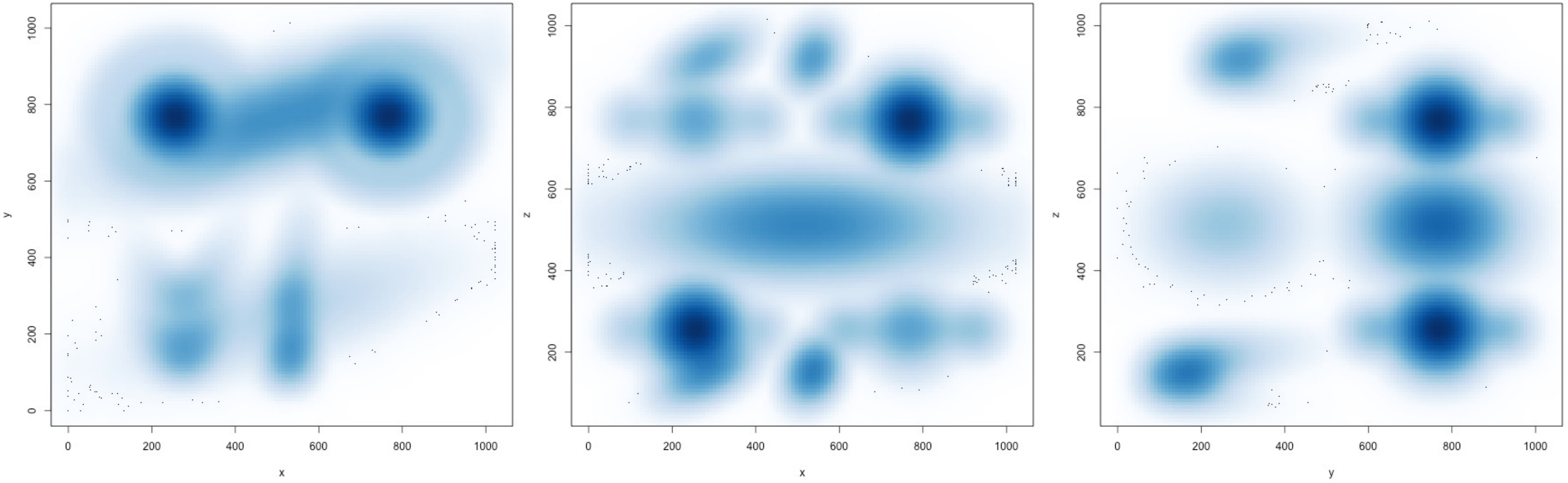
Tests with synthetic data

comprising a total of 799,956 points



Tests with synthetic data

comprising a total of 799,956 points



Tests with synthetic data

INPUT: 14 shapes, comprising a total of 799,956 points

Spheres 1 (blue) and 4 (dark blue) and ellipse 1 (yellow) have 200,000 points

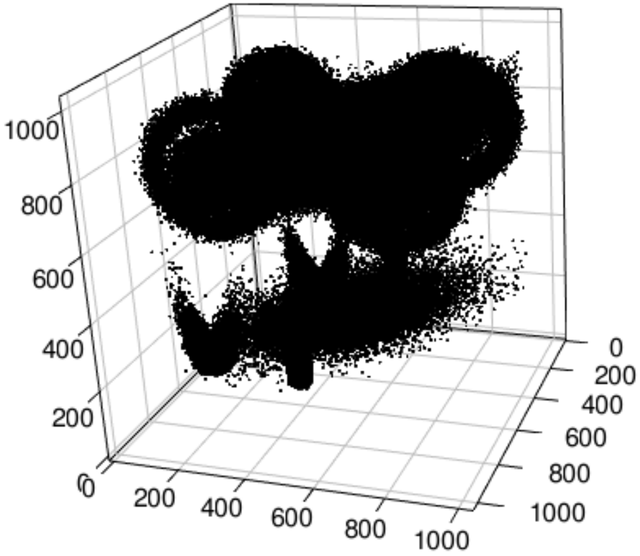
Spheres 2 (cyan) and 3 (light blue) and ellipse 2 (brown) have 20,000 points

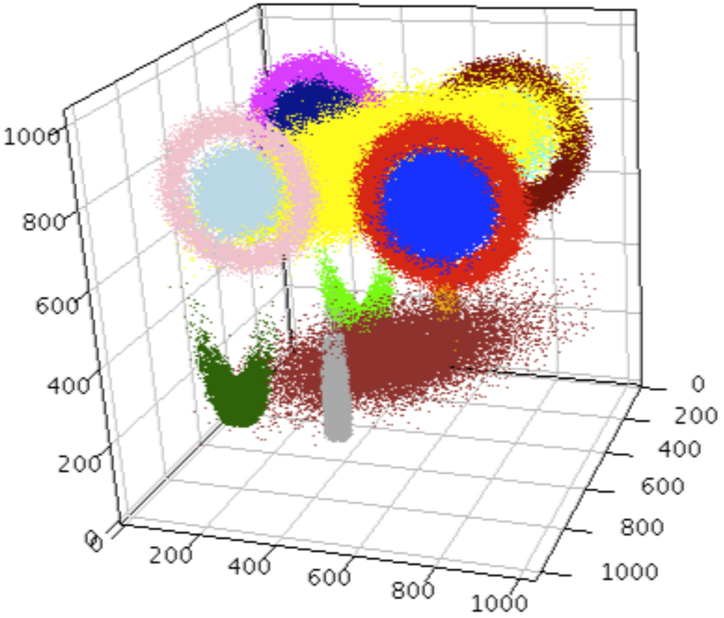
Tores 1 (red) and 2 (dark red) have 16,652 points

Tores 3 (pink) and 4 (magenta) have 8,326 points

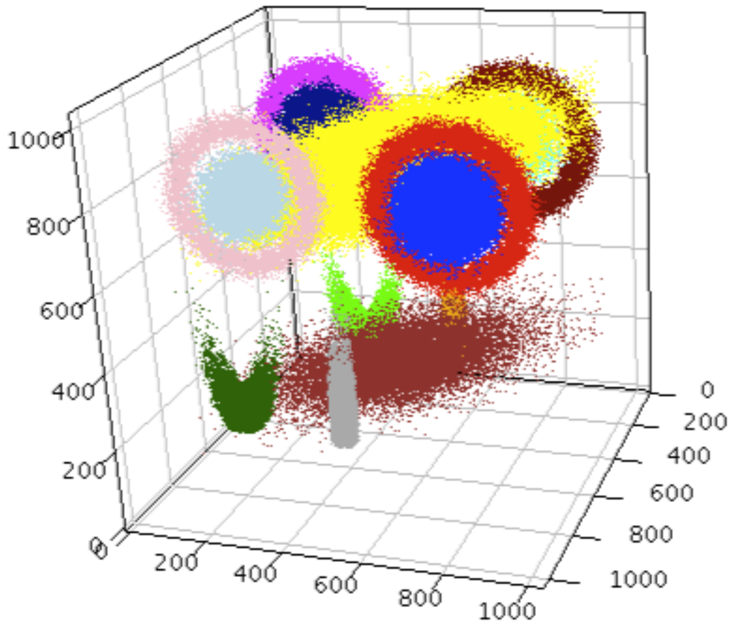
Bananas 1 (green) and 2 (orange) have 30,000 points and different curvatures

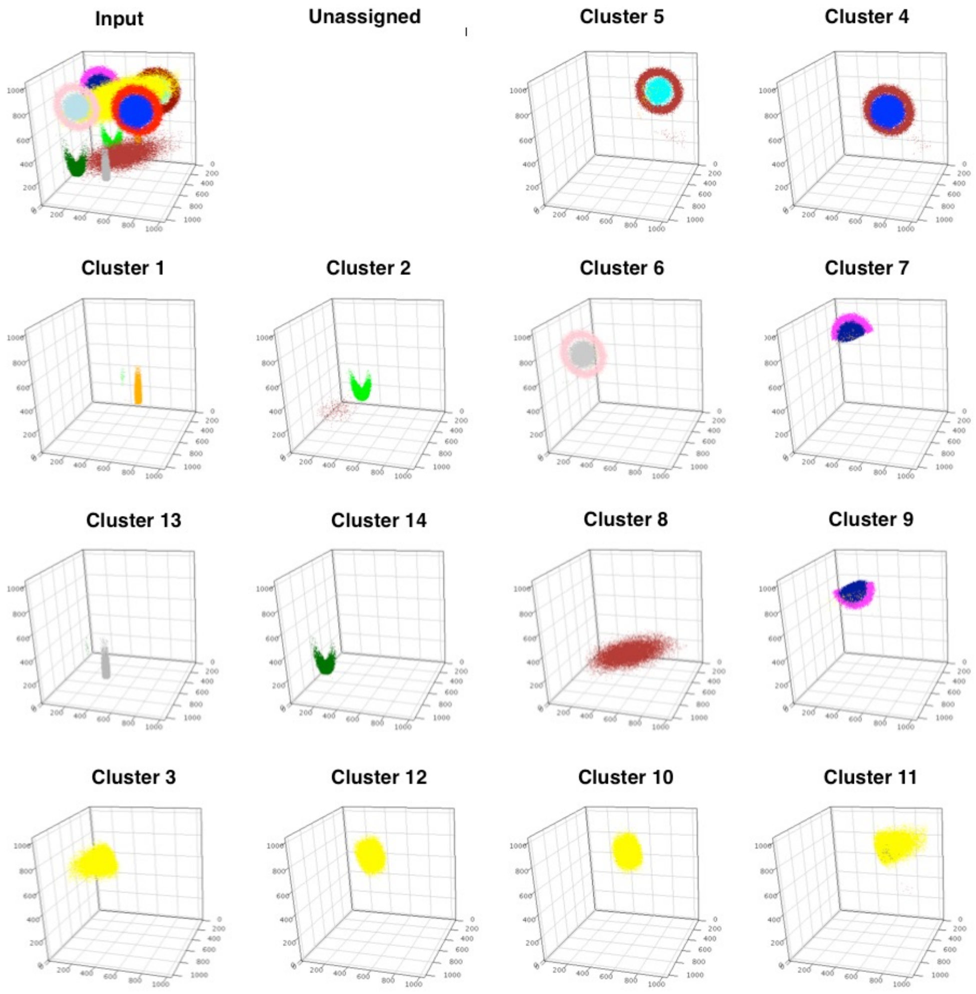
Bananas 3 (dark green) and 4 (dark grey) have 15,000 points and different curvatures





K - mean





Input

Unassigned

Cluster 5

Cluster 4

Cluster 1

Cluster 2

Cluster 6

Cluster 7

Cluster 13

Cluster 14

Cluster 8

Cluster 9

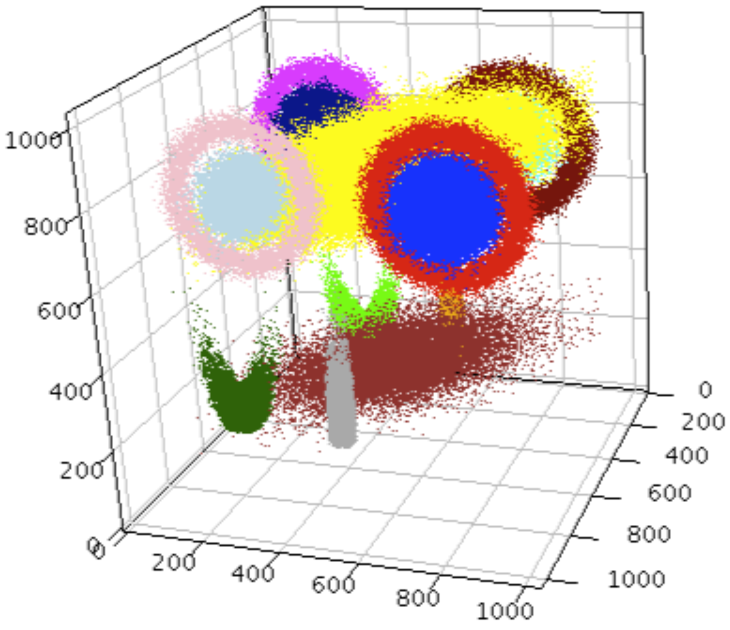
Cluster 3

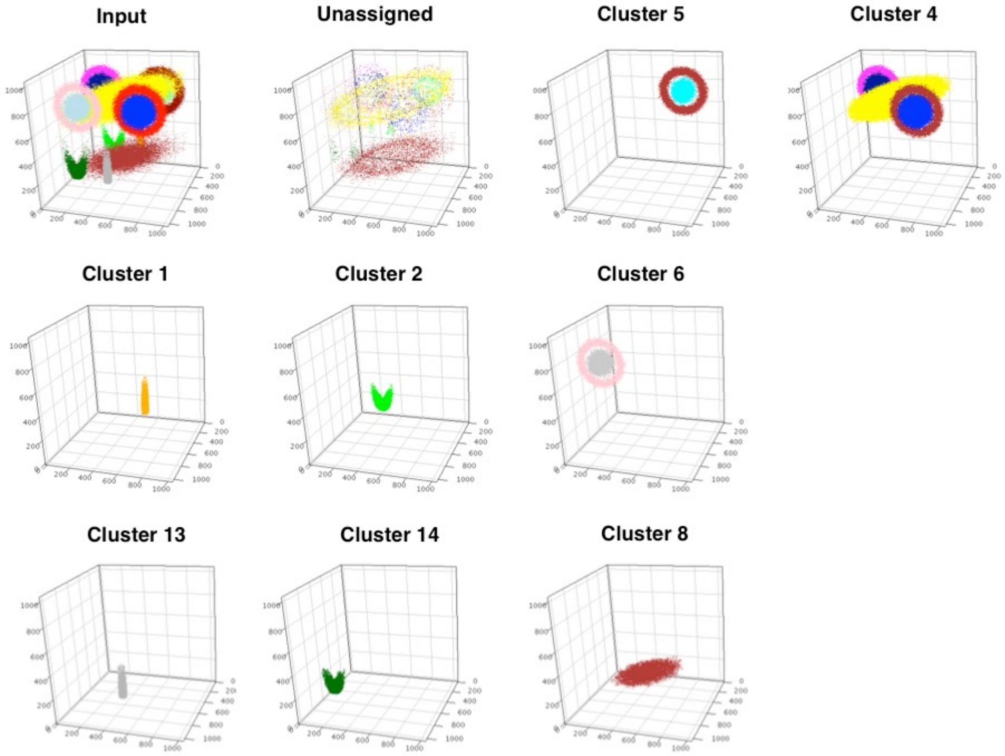
Cluster 12

Cluster 10

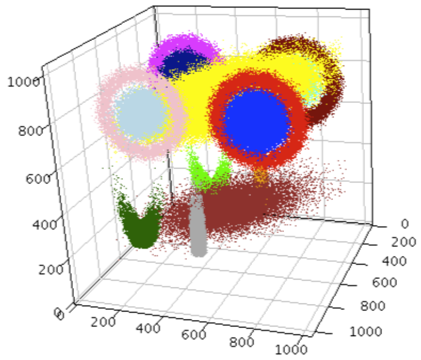
Cluster 11

DBscan

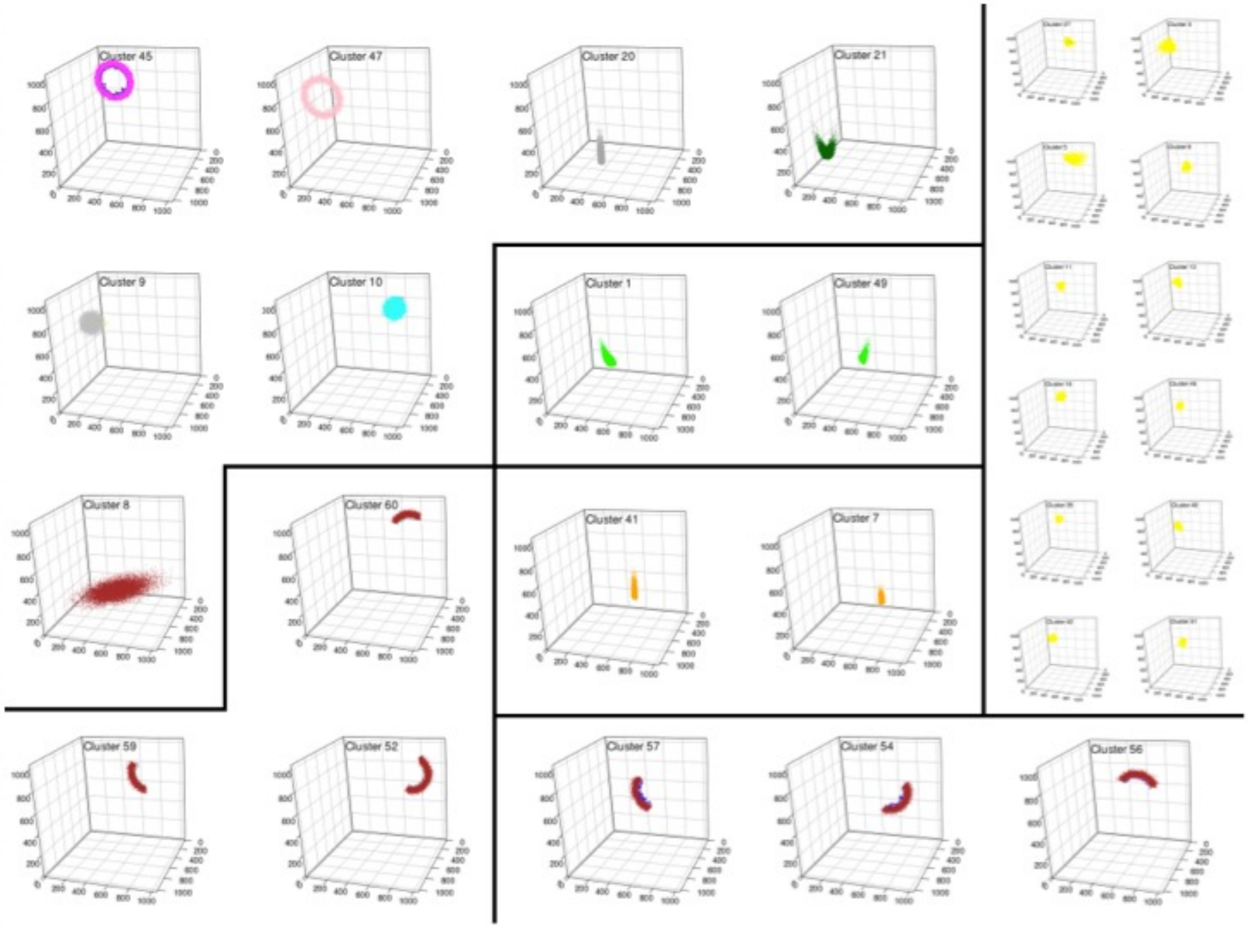

1000
800
600
400
200
0
0
200
400
600
800
1000
0
200
400
600
800
1000


Input
Unassigned
Cluster 5
Cluster 4
Cluster 1
Cluster 2
Cluster 6
Cluster 13
Cluster 14
Cluster 8

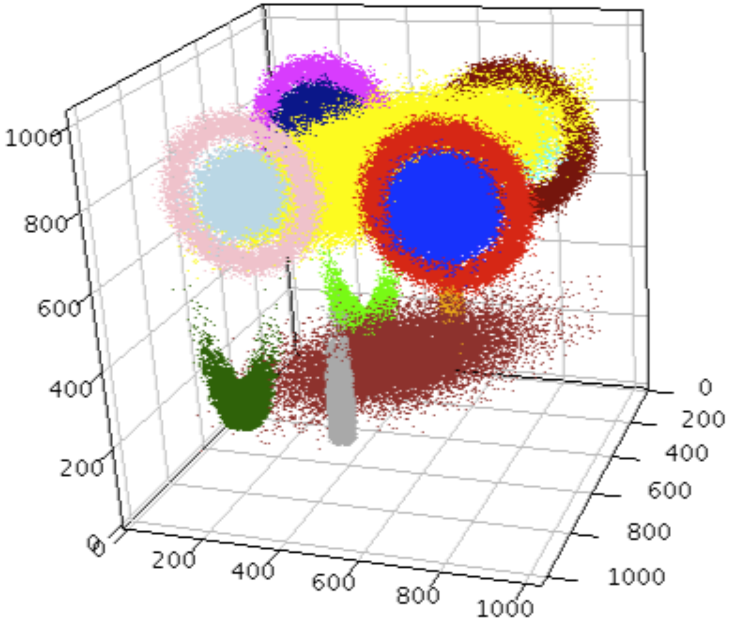
Phenograph



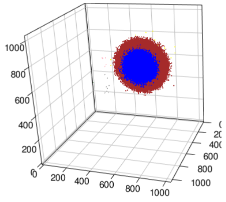
- Identified 61 clusters
- Correctly identified 7 out of 14 shapes
- Oversplit the others
- Dense spheres each split in 16 clusters



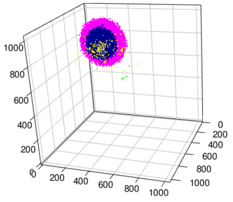
FlowSOM (run #1)



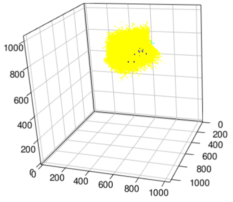
cluster 1



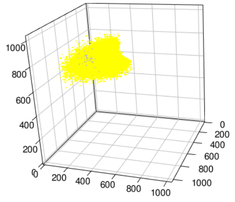
cluster 2



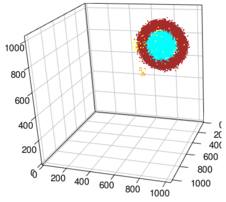
cluster 3



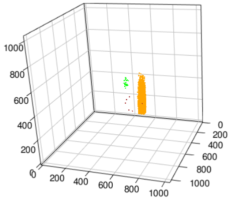
cluster 4



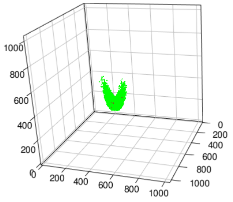
cluster 5



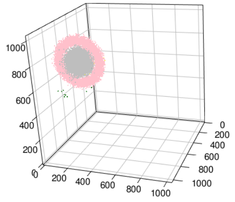
cluster 6



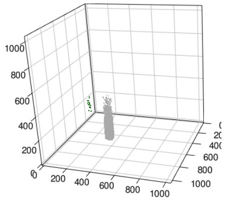
cluster 7



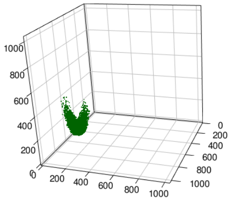
cluster 8



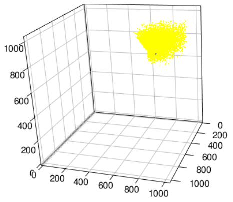
cluster 9



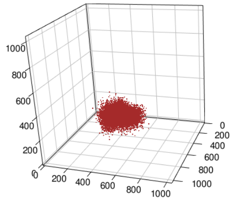
cluster 10



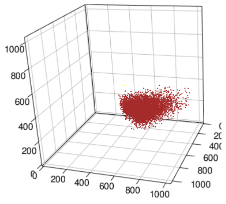
cluster 11



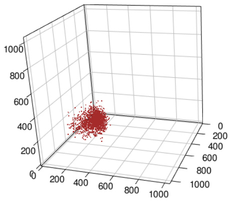
cluster 12



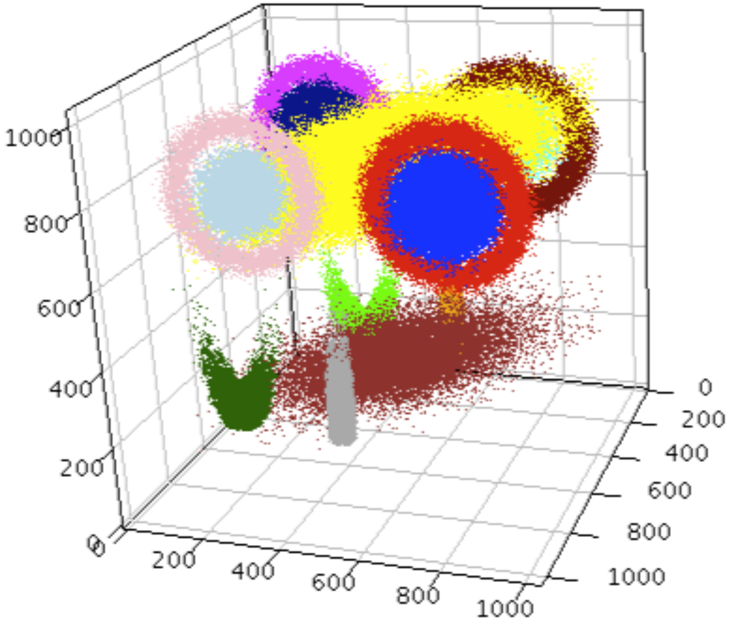
cluster 13



cluster 14

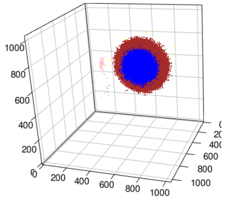


FlowSOM (run #2)

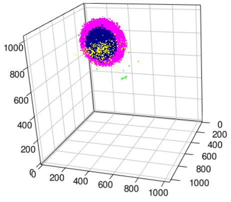


RUN #1	RUN #2
216882	216964
208557	208644
111741	130095
74487	54045
36688	36772
30008	29998
29967	29967
28328	28223
15014	16856
14968	15270
13325	15024
10927	14957
7428	1623
1635	1518

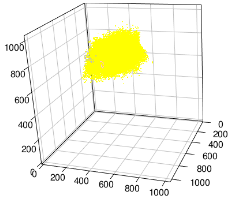
cluster 1



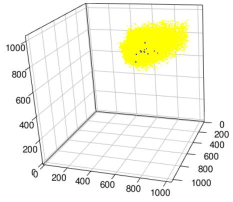
cluster 2



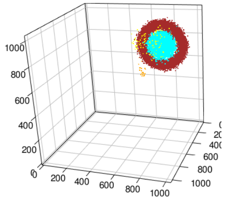
cluster 3



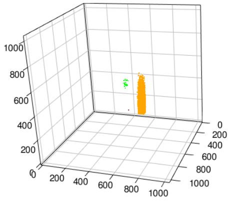
cluster 4



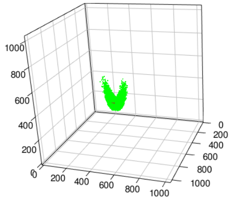
cluster 5



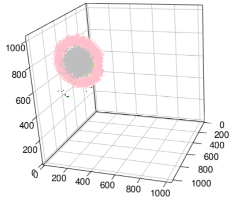
cluster 6



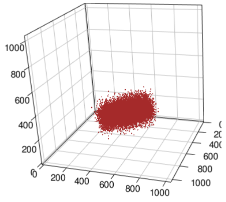
cluster 7



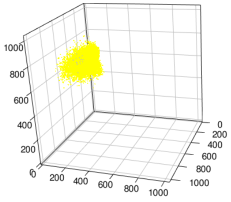
cluster 8



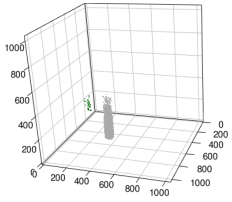
cluster 9



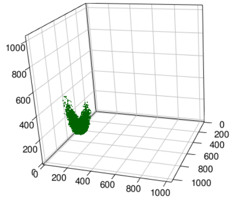
cluster 10



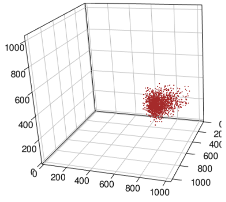
cluster 11



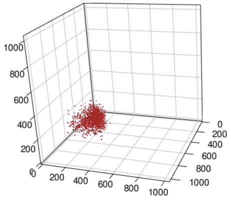
cluster 12



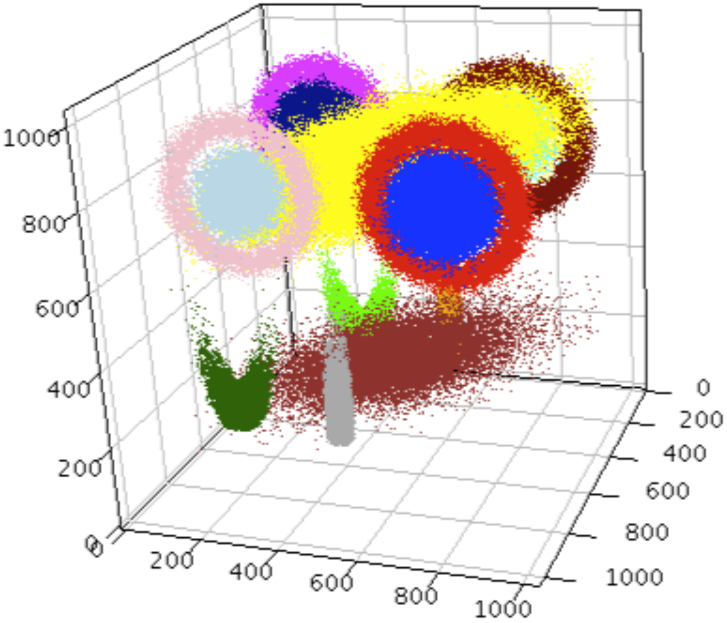
cluster 13

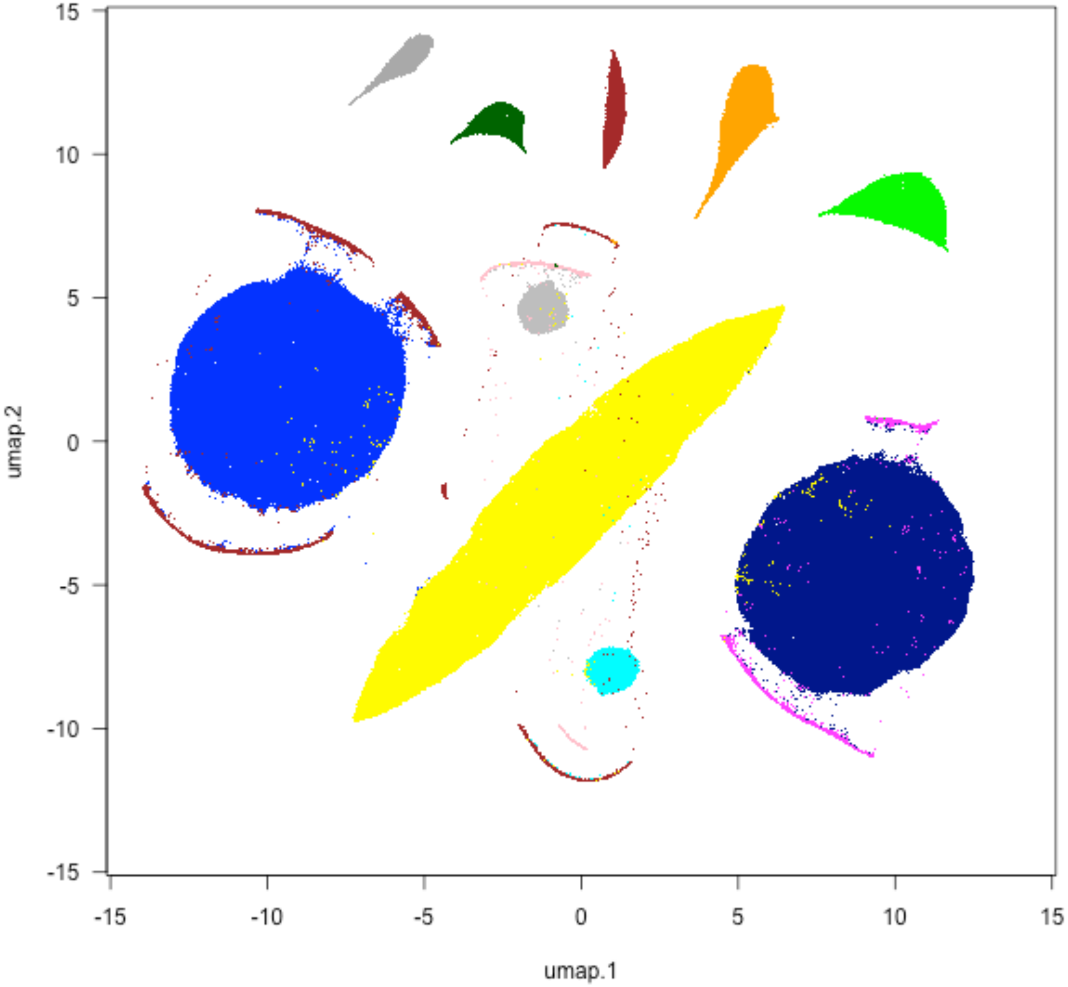


cluster 14



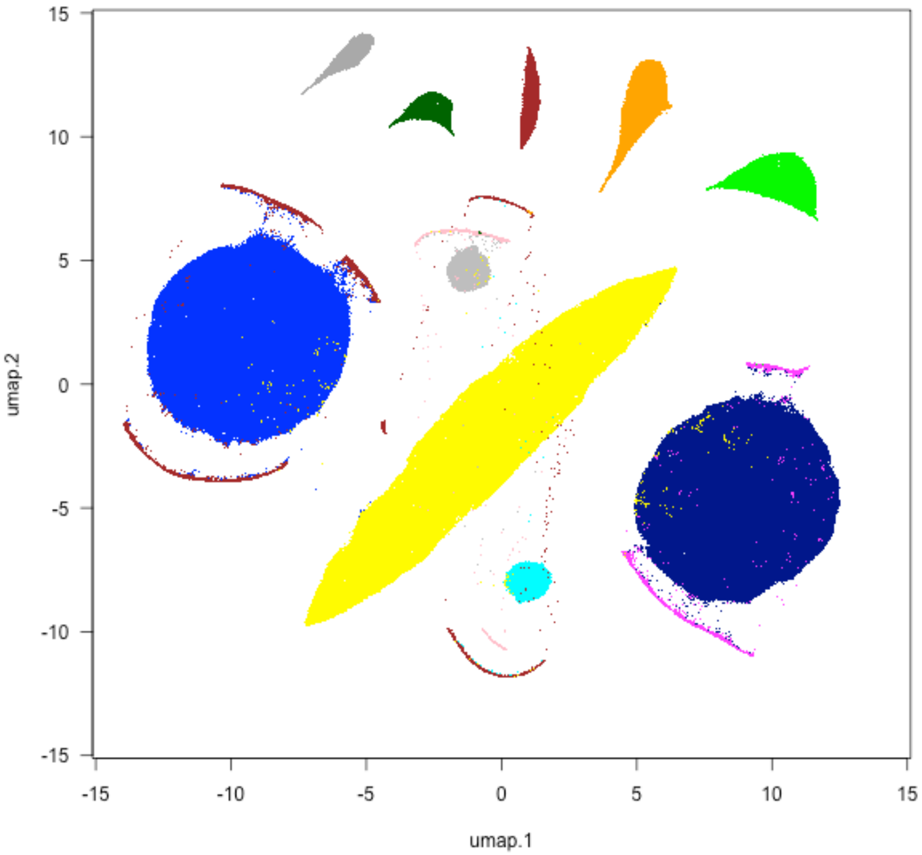
small parenthesis : UMAP



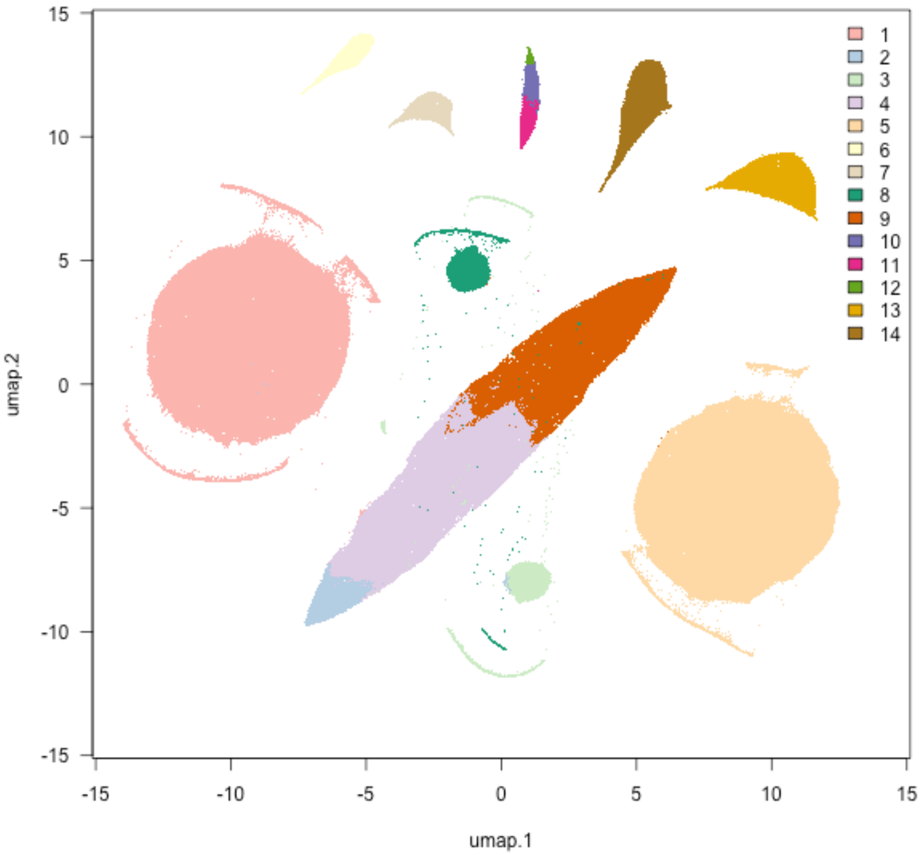


small parenthesis : UMAP

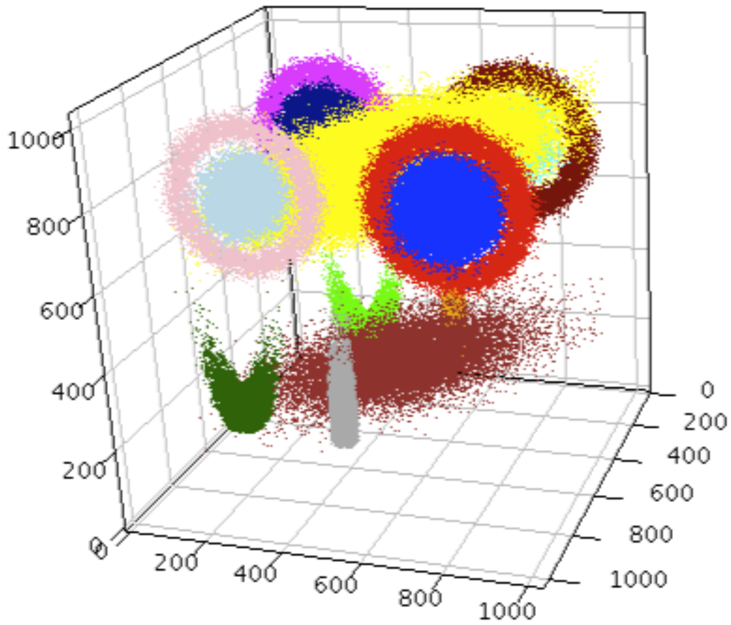
3D Gold Standard Projected in 2D

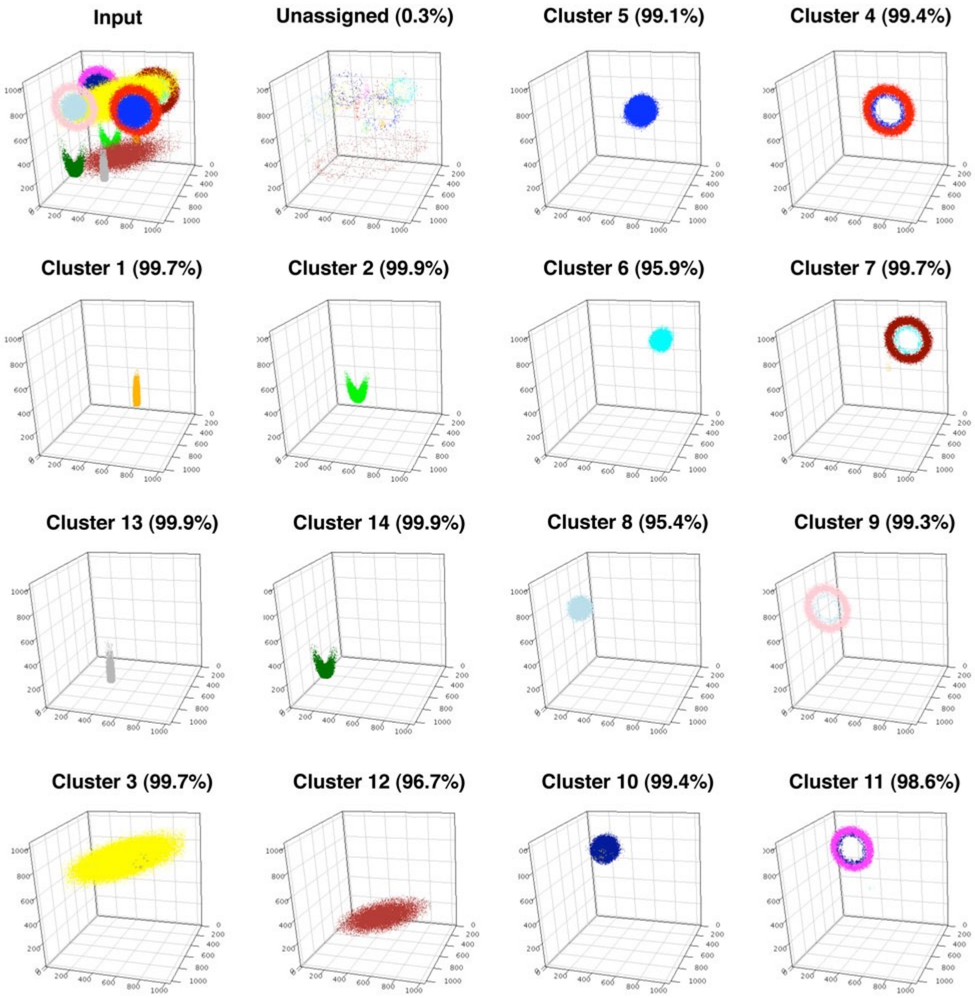


FlowSOM run #1 clusters



Megaclust / hdbscan





Input

Unassigned (0.3%)

Cluster 5 (99.1%)

Cluster 4 (99.4%)

Cluster 1 (99.7%)

Cluster 2 (99.9%)

Cluster 6 (95.9%)

Cluster 7 (99.7%)

Cluster 13 (99.9%)

Cluster 14 (99.9%)

Cluster 8 (95.4%)

Cluster 9 (99.3%)

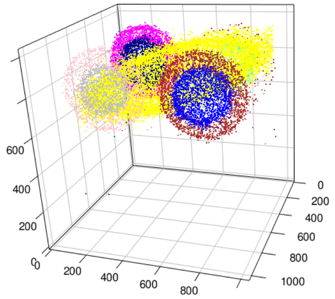
Cluster 3 (99.7%)

Cluster 12 (96.7%)

Cluster 10 (99.4%)

Cluster 11 (98.6%)

hdbscan unassigned (3%)



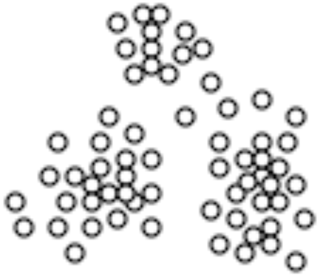
Hierarchical Clustering

- Number of Computations
- Memory

  1e6 * 1e6 * 4 = 4Tb

observations	computation	time
10	50	60 us
100	5000	6 ms
1000	500000	600 ms
10000	50000000	1 mn
100000	5000000000	1 h 40 mn
1000000	500000000000	7 days

- Parallelization



Density-based hierarchical clustering

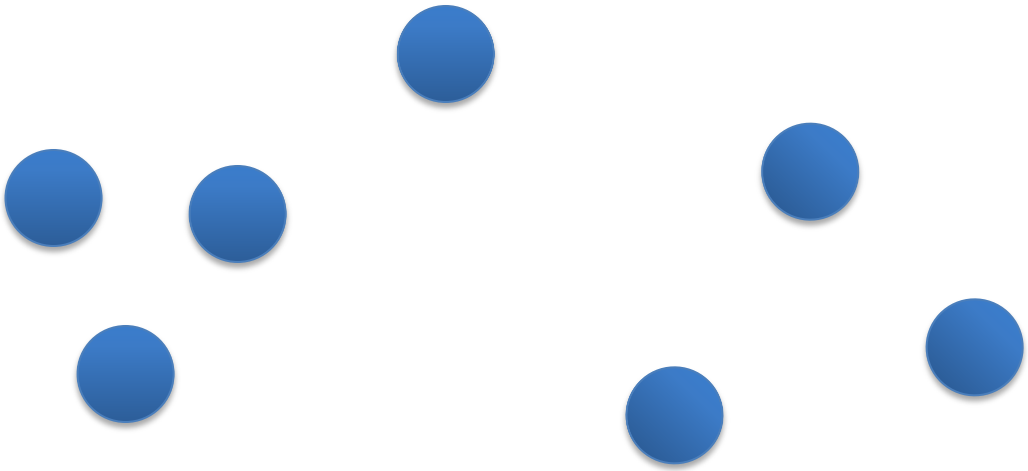
- Compute all pairwise distances and retain only those that are equal or smaller than a given distance threshold *T*

- A cluster is formed by single linkage and retained if it contains at least *N* points. Clusters too small are ignored

- Hierarchical clustering is obtained by repeating the clustering for increasing distance thresholds

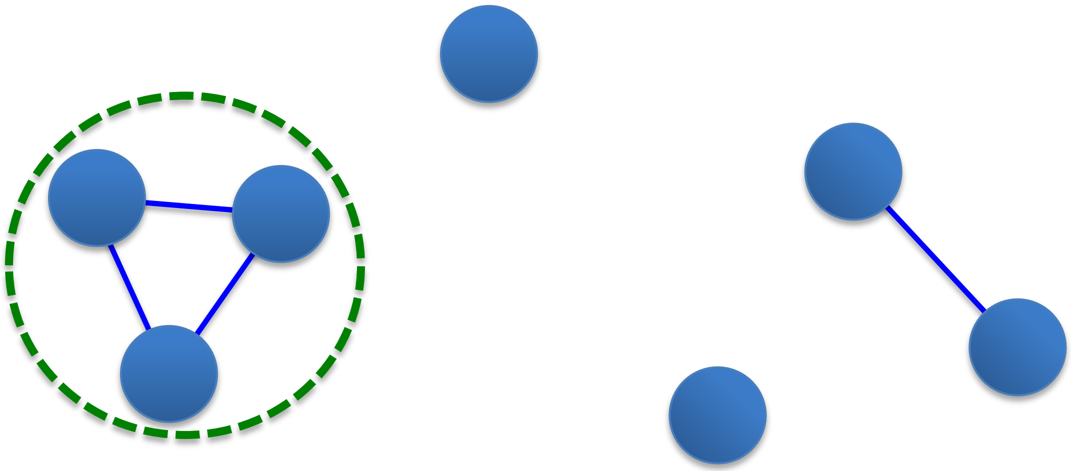
T from T_{min} to T_{max} in s steps

Density-based hierarchical clustering



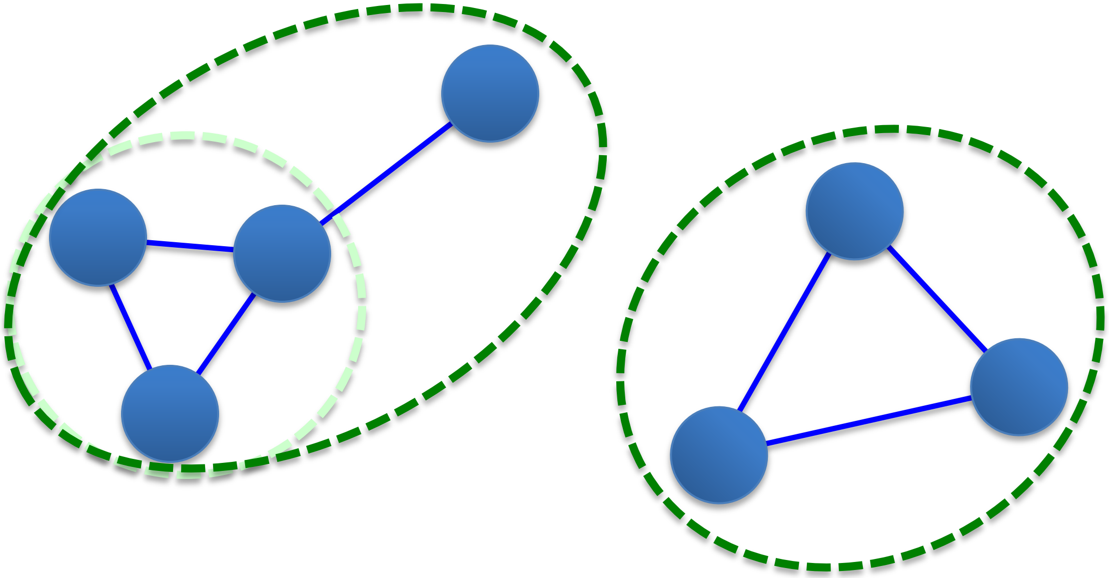
Density-based hierarchical clustering

- $D \leq 1$ and $N \geq 3$



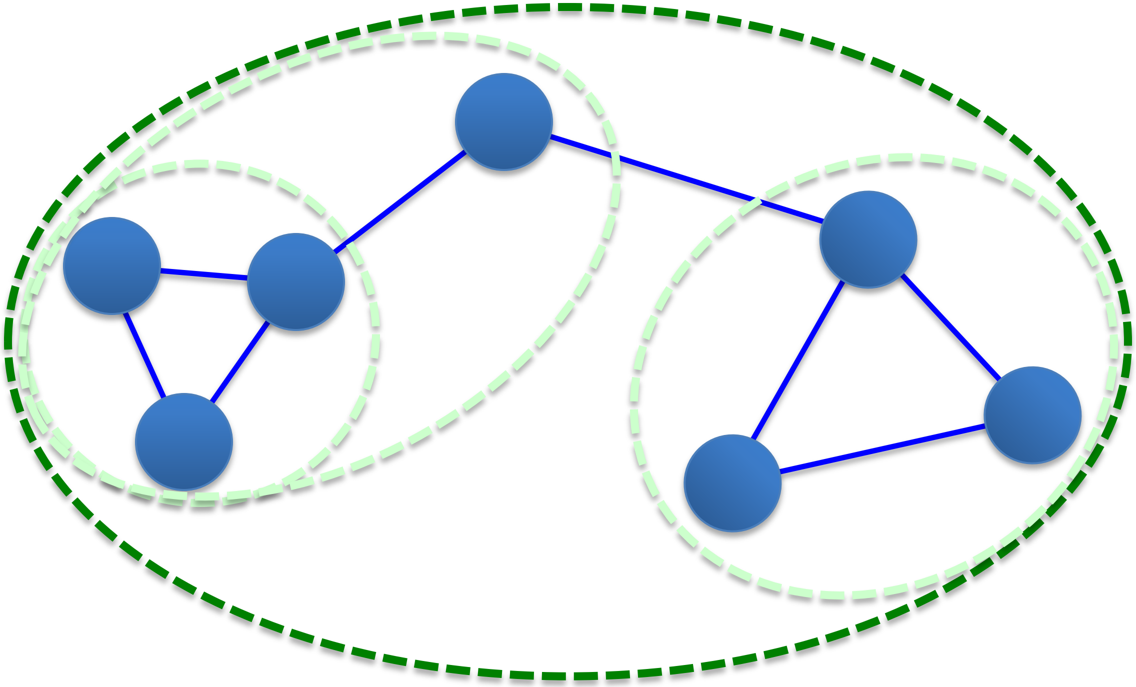
Density-based hierarchical clustering

- $D \leq 2$ and $N \geq 3$

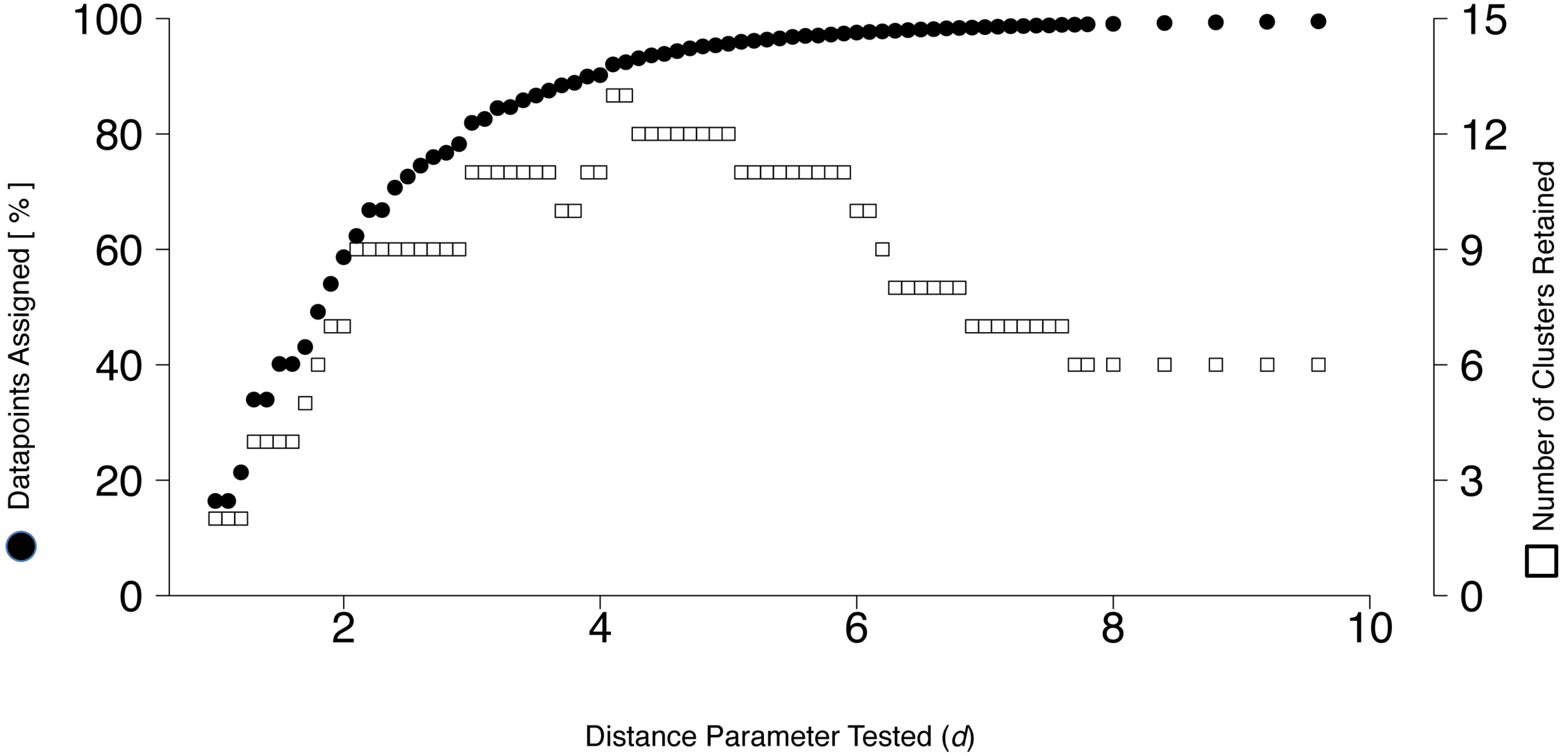


Density-based hierarchical clustering

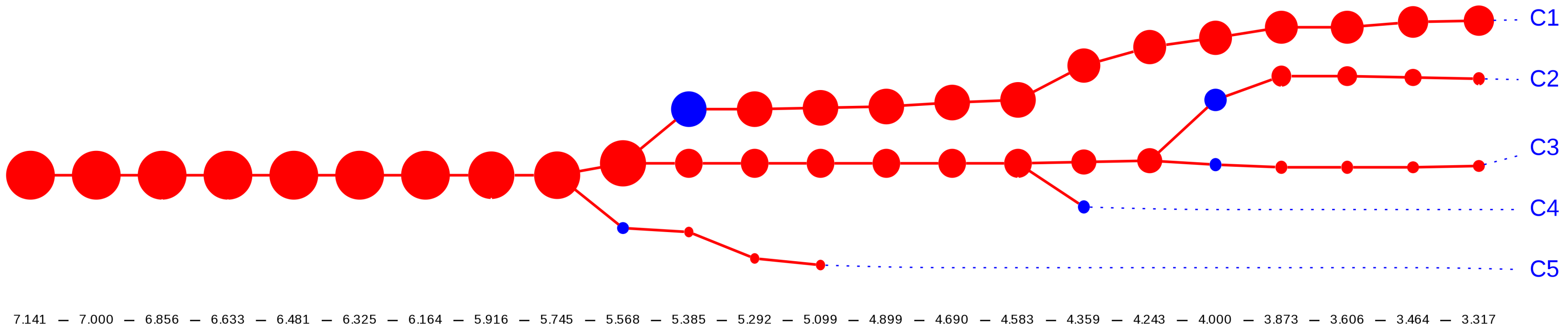
- $D \leq 3$ and $N \geq 3$



Density-based hierarchical clustering

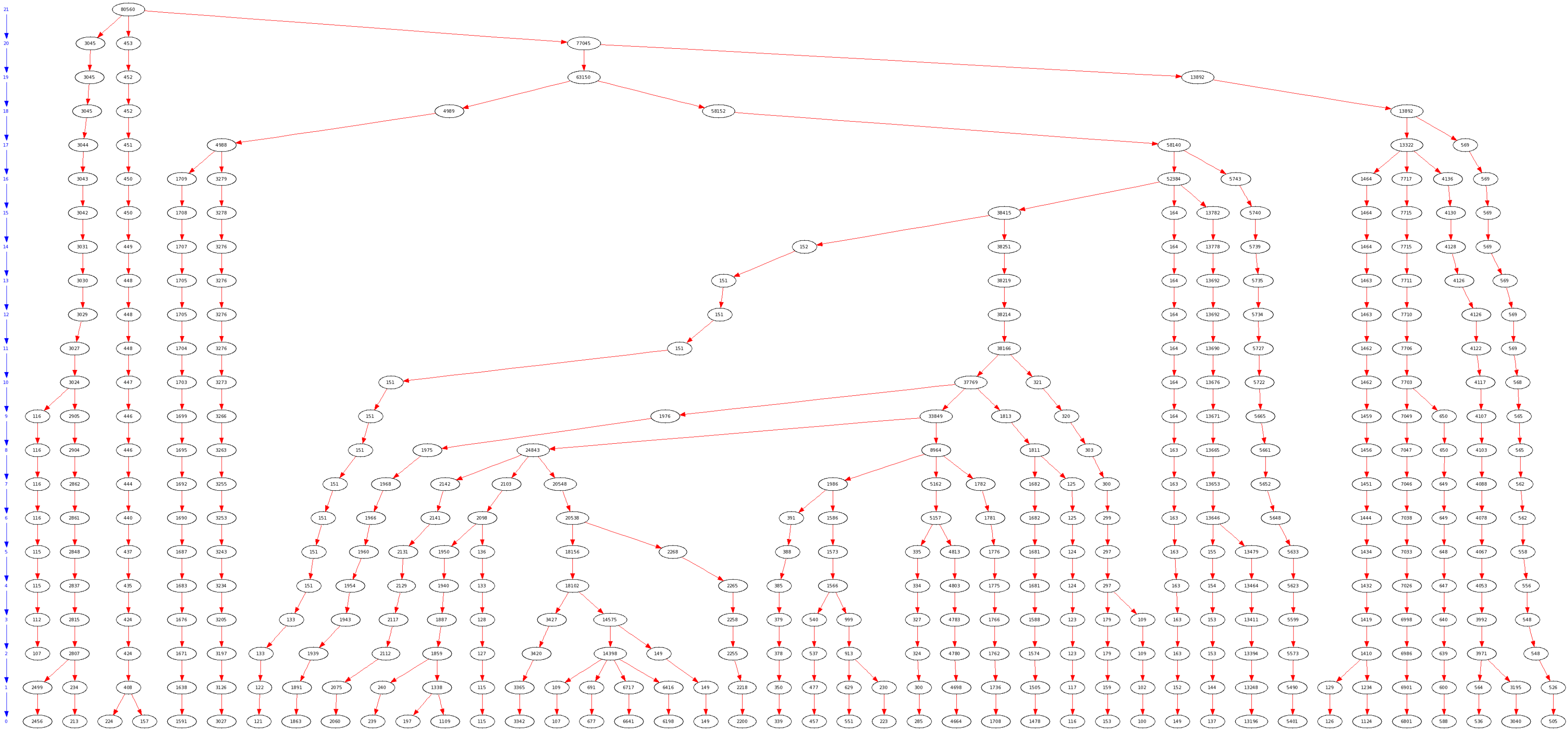


Assignment to the final partition

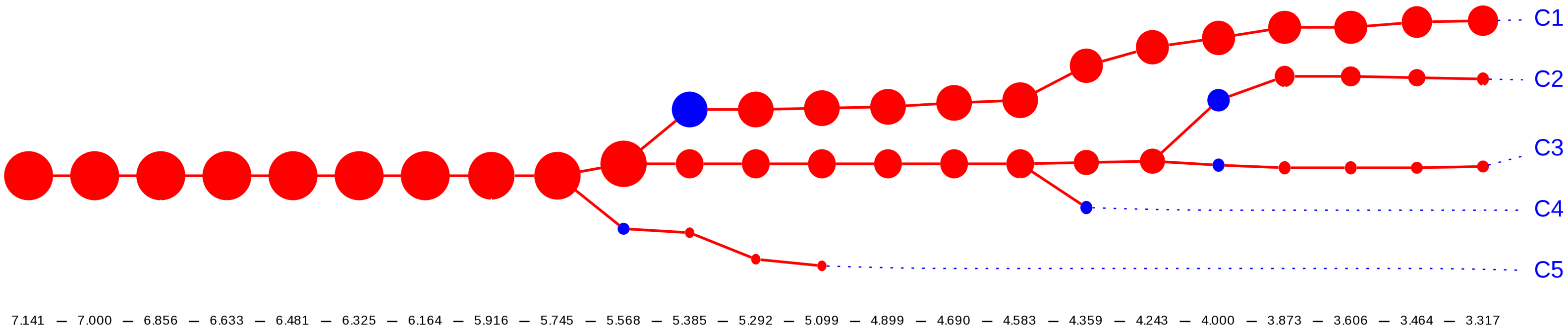


- At the end of the hierarchical clustering, seed clusters are determined and the points are re-attributed to the closest seed

A more realistic output

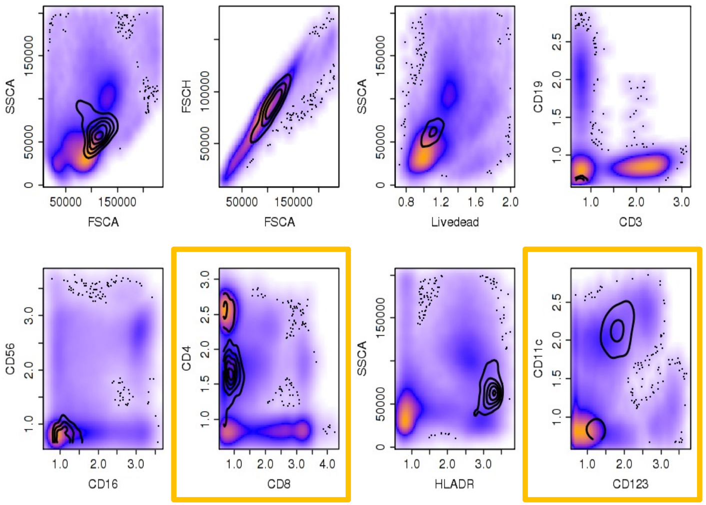


Parameters affecting results

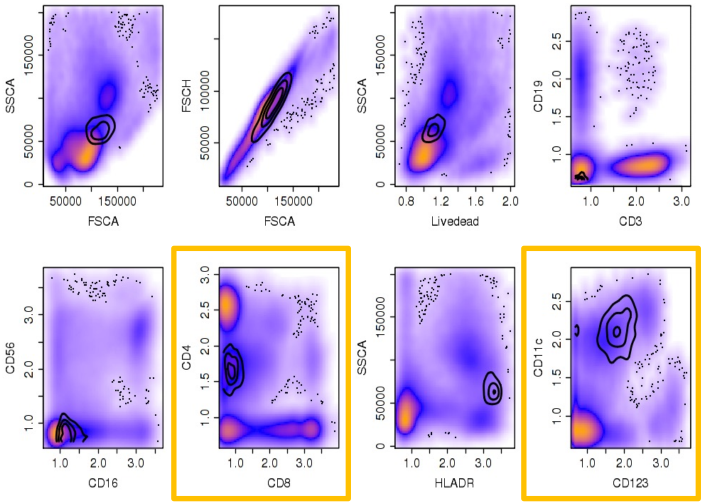


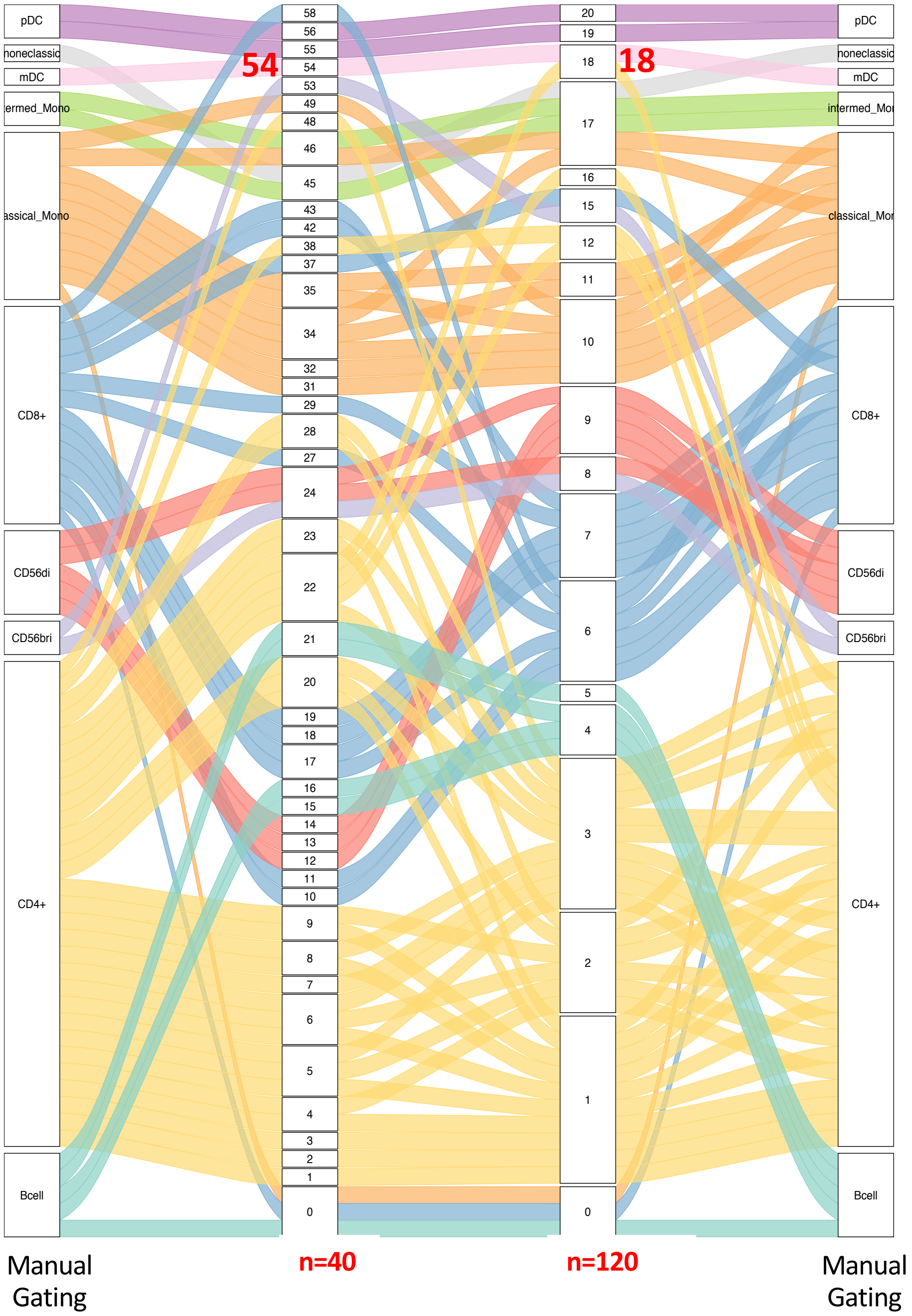
- -f first distance to test
- -l last distance to be tested
- -s step increment for the distance test
- -k minimum percent of events needed to retain a cluster
- -n minimum number of events needed to retain a cluster
- -p pctAssigned (Stop sampling as soon as pctAssigned events have been assigned)

@n120, Cluster 18 (1197 events, 0.005% live cells)



@n40, Cluster 54 (890 events, 0.003% live cells)

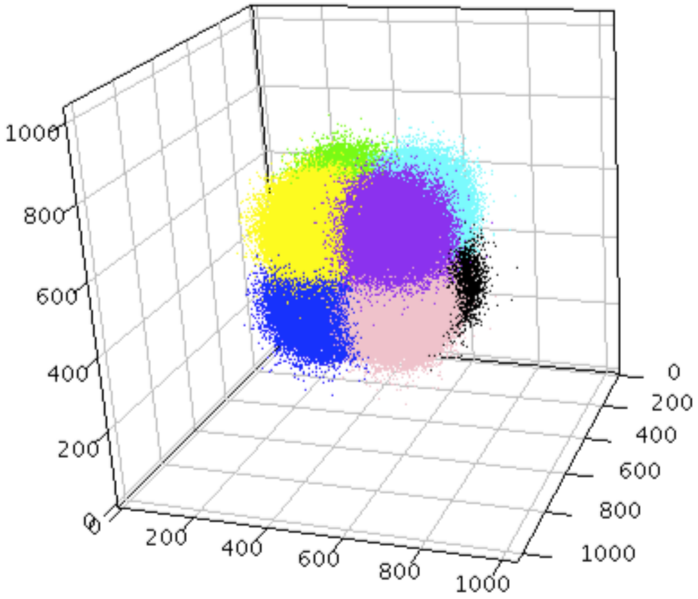


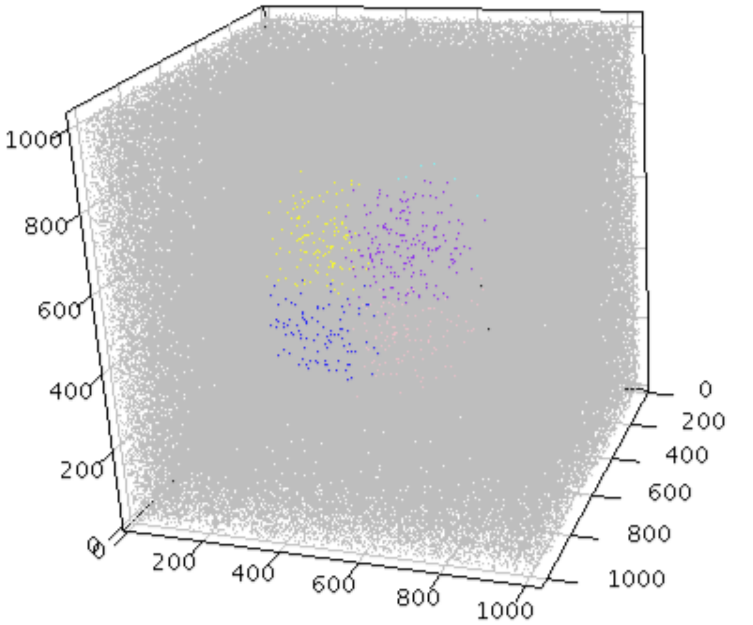


Tests with noise

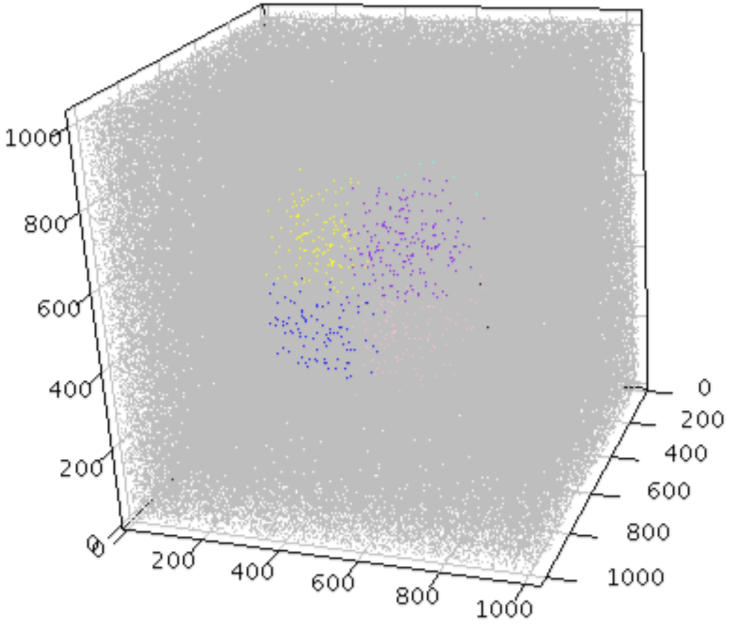
INPUT:

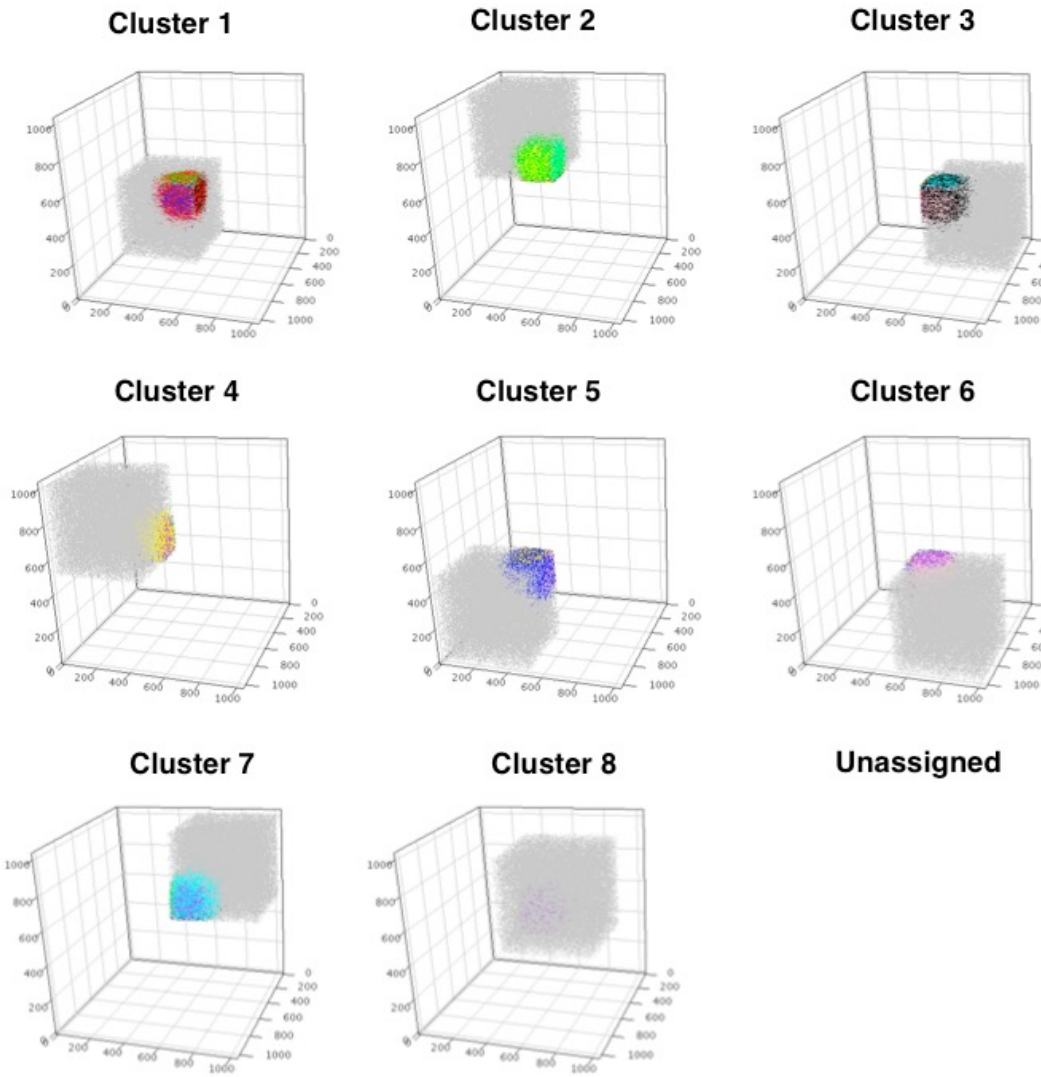
- 8 overlapping spheres with 125,000 points, comprising a total of 1e6 points
- random noise 500,000 points





K - means




Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

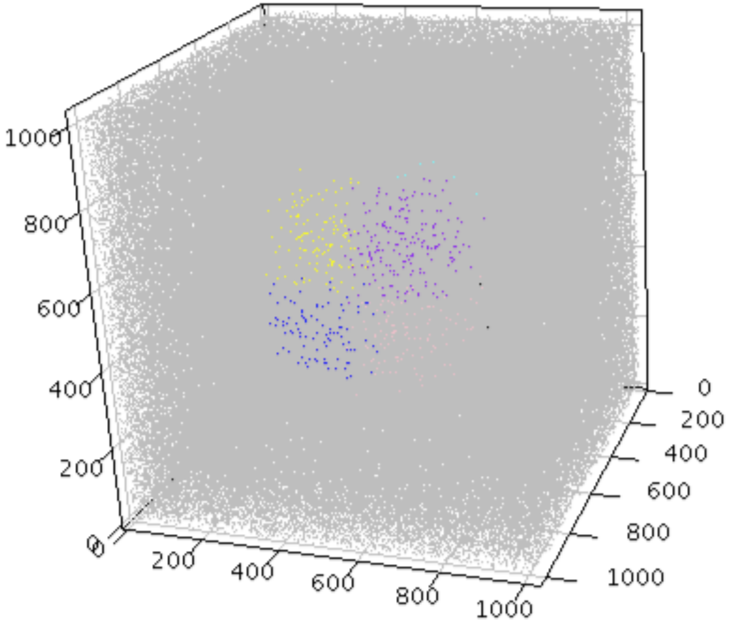
Cluster 6

Cluster 7

Cluster 8

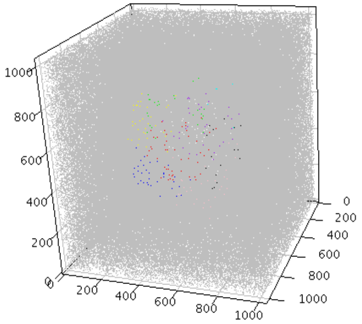
Unassigned

DBscan



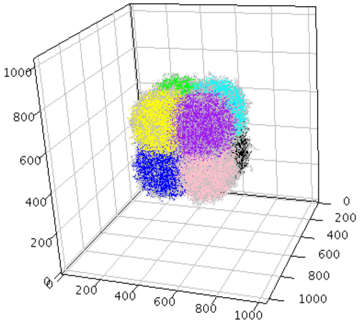
Unassigned

367’162 points



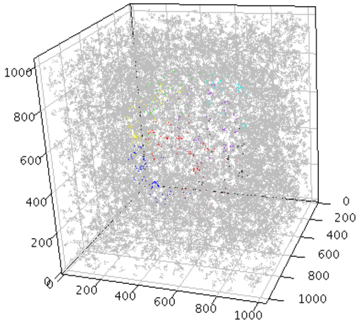
Cluster 1

1’051’091 points

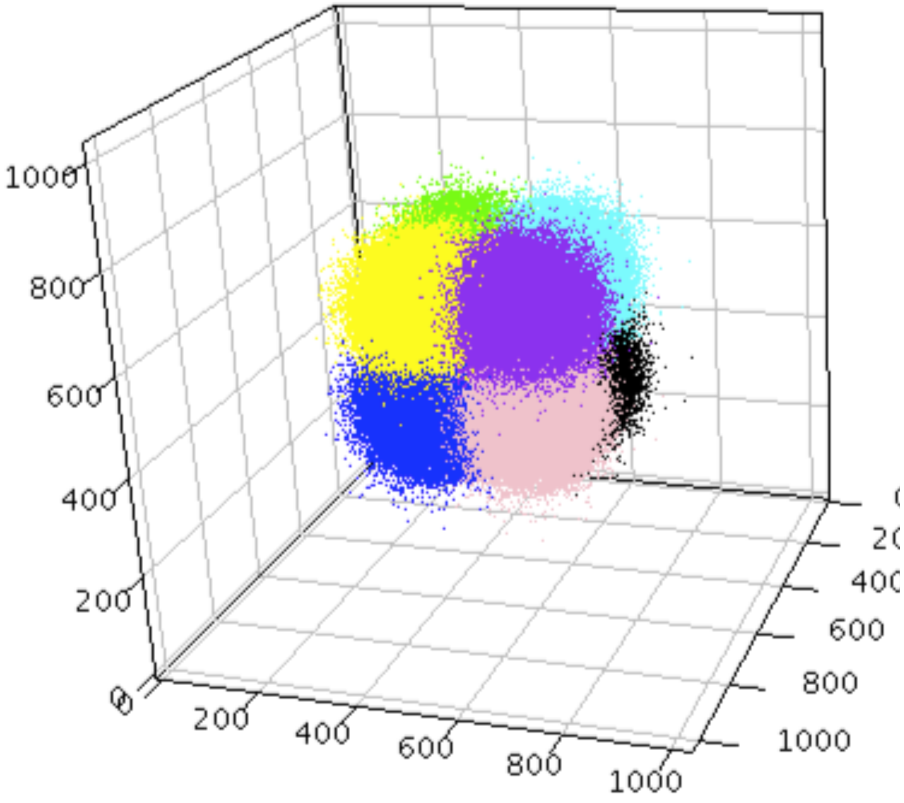


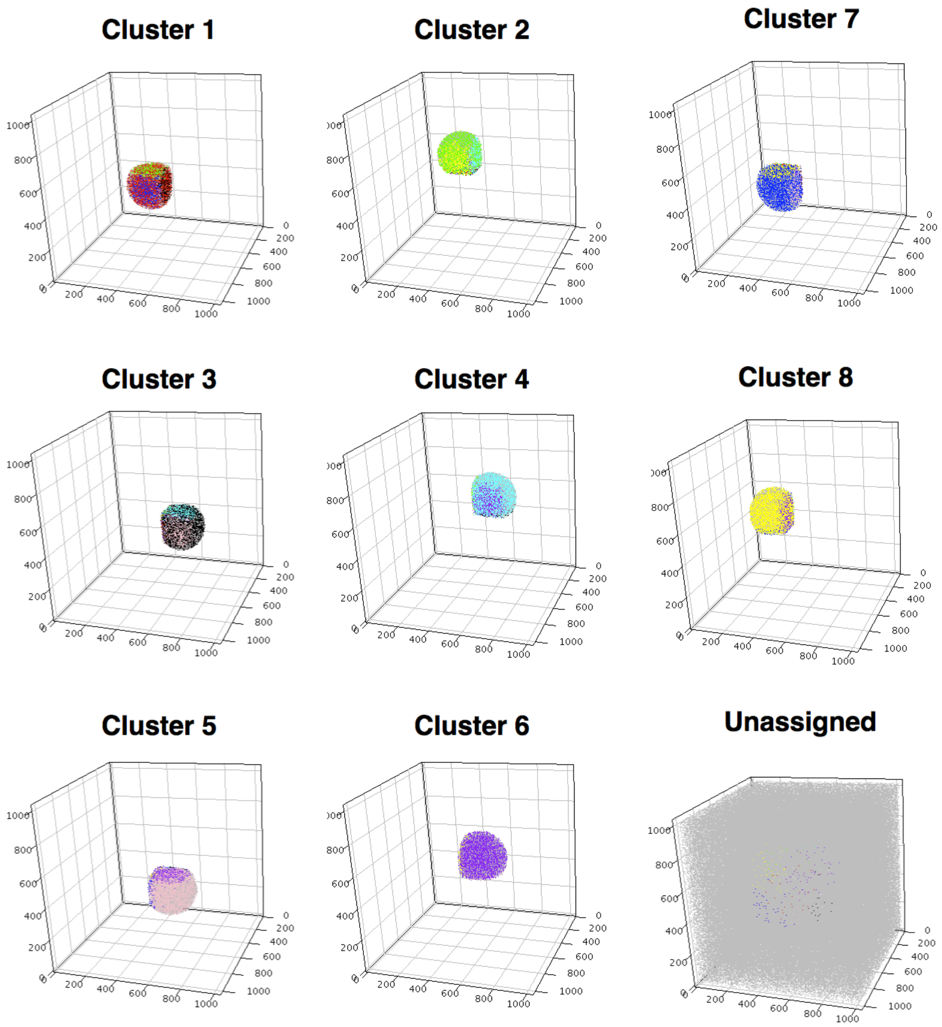
pool of 5’346 other clusters:

81’747 points



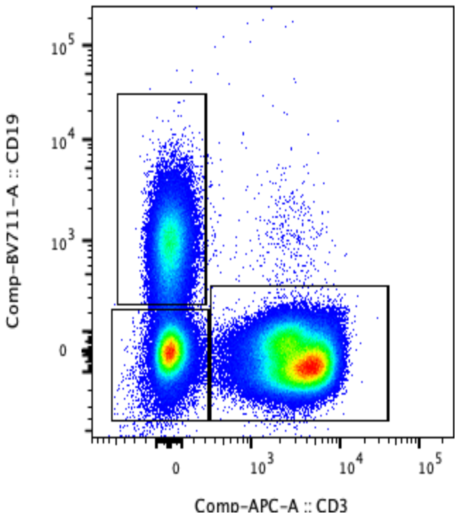
Megaclust


1000
800
600
400
200
0
200
400
600
800
1000
0
200
400
600
800
1000

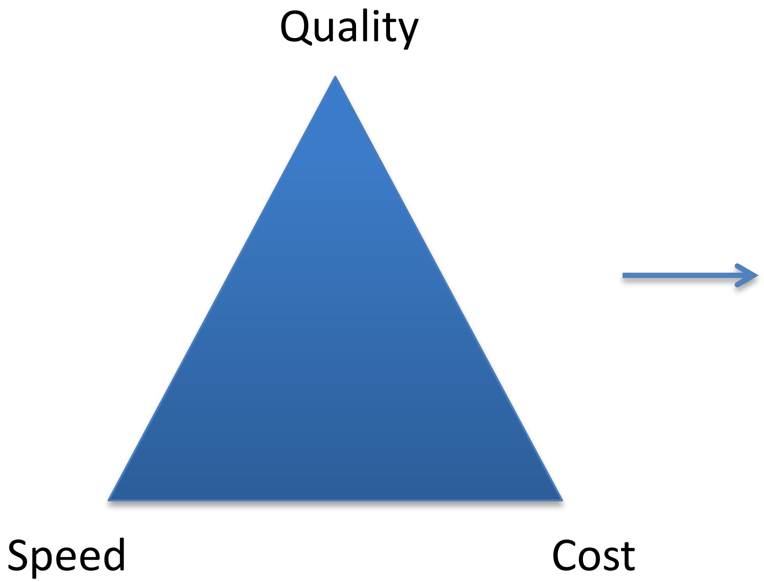

Cluster 1
Cluster 2
Cluster 7
Cluster 3
Cluster 4
Cluster 8
Cluster 5
Cluster 6
Unassigned

Input and parameters affecting results

- numerical range of markers
  - will each marker contribute equally to the distance metric ?
- data sampling
  - will each sample contribute equally to the cluster discovery ?
- data acquisition calibration
  - can samples be mixed in a single run ?
- input quality
  - are acquisitions stable ?



Data Analysis Tradeoffs



Practical considerations

consider computer time
consider human time (to code or operate)
consider delivery time
consider desired outcome
consider quality *really* necessary