



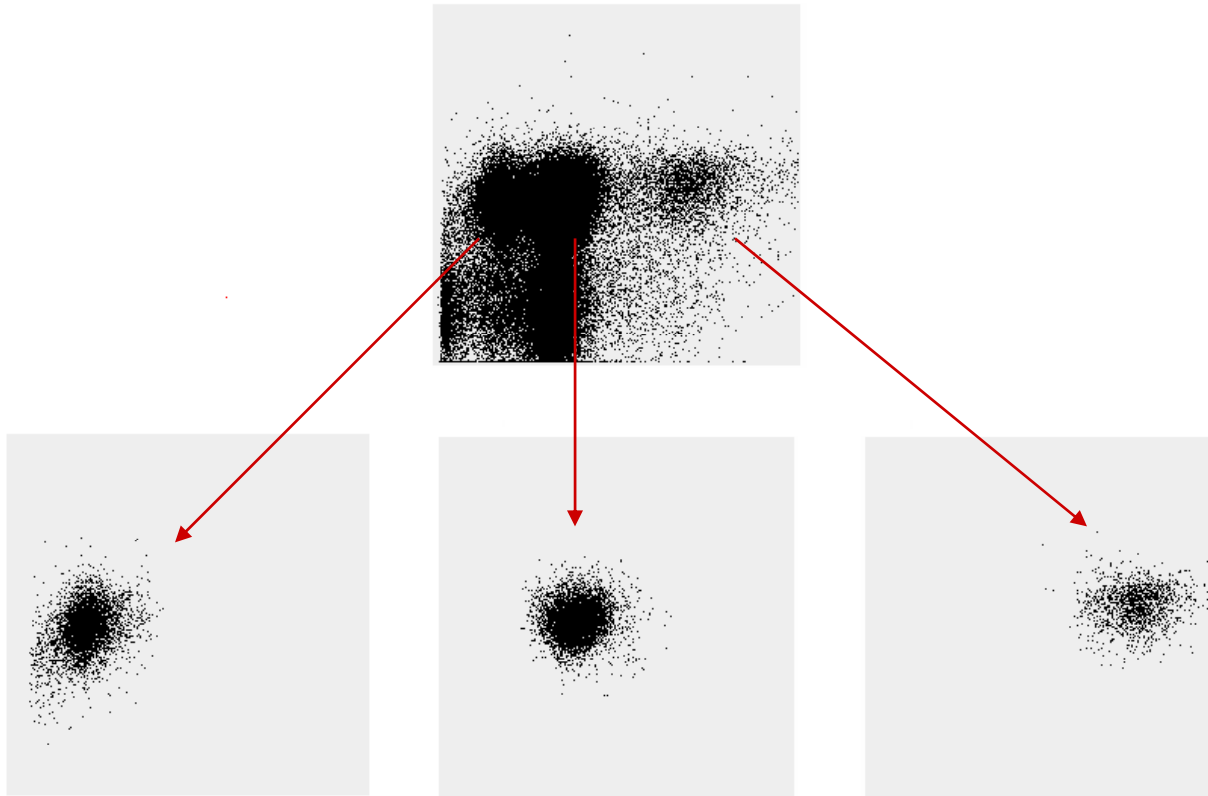
FLOCK: A Density-Based Clustering Method for Automated Identification and Comparison of Cell Populations in High-Dimensional Flow Cytometry Data

Max Yu Qian, Ph.D.

Division of Biomedical Informatics and Department of Pathology
University of Texas Southwestern Medical Center, Dallas, TX
September 21, 2010

Why Computation Is Necessary

- Segregating overlapping cell populations



Solution: Clustering

- Assumption: Cells of the same *population* express ALL biological markers similarly

Related Work in Clustering

- Density-based (such as DBSCAN)
- Partitioning approaches (such as K-means)
- Hierarchical approaches (such as HAC)
- Grid-based approaches (such as STING)

J. Han, M. Kamber, A. K. H. Tung, “Spatial Clustering Methods in Data Mining: A Survey”

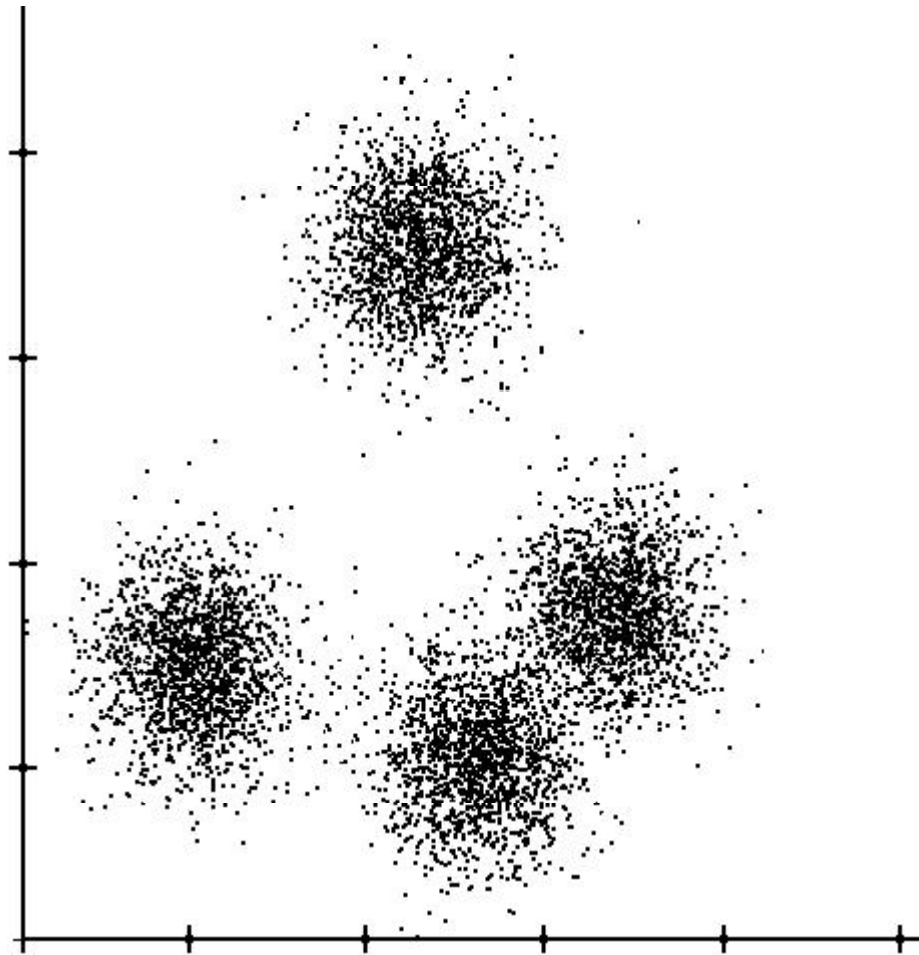
There is another category called Model-based Clustering, such as the EM method.

Previous Methods not Directly Applicable

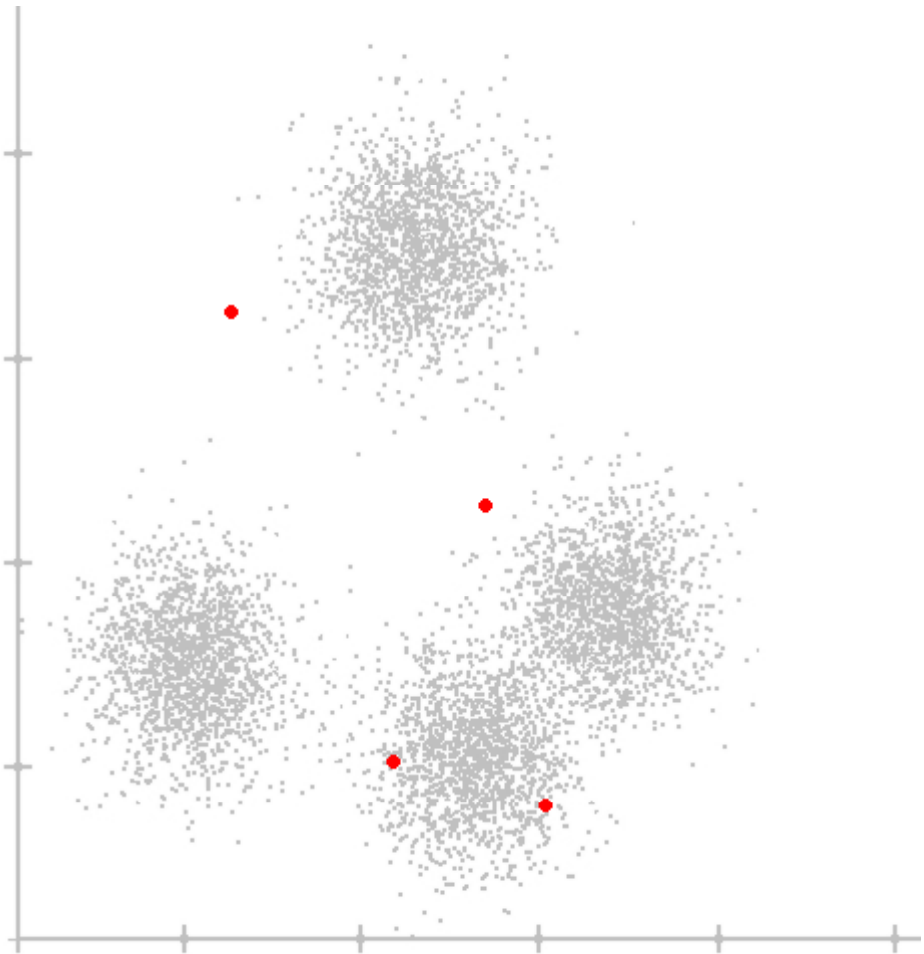
FCM data requires the clustering method to be:

- 1) Efficient
- 2) Able to handle high-dimensionality
- 3) Easy setting parameters

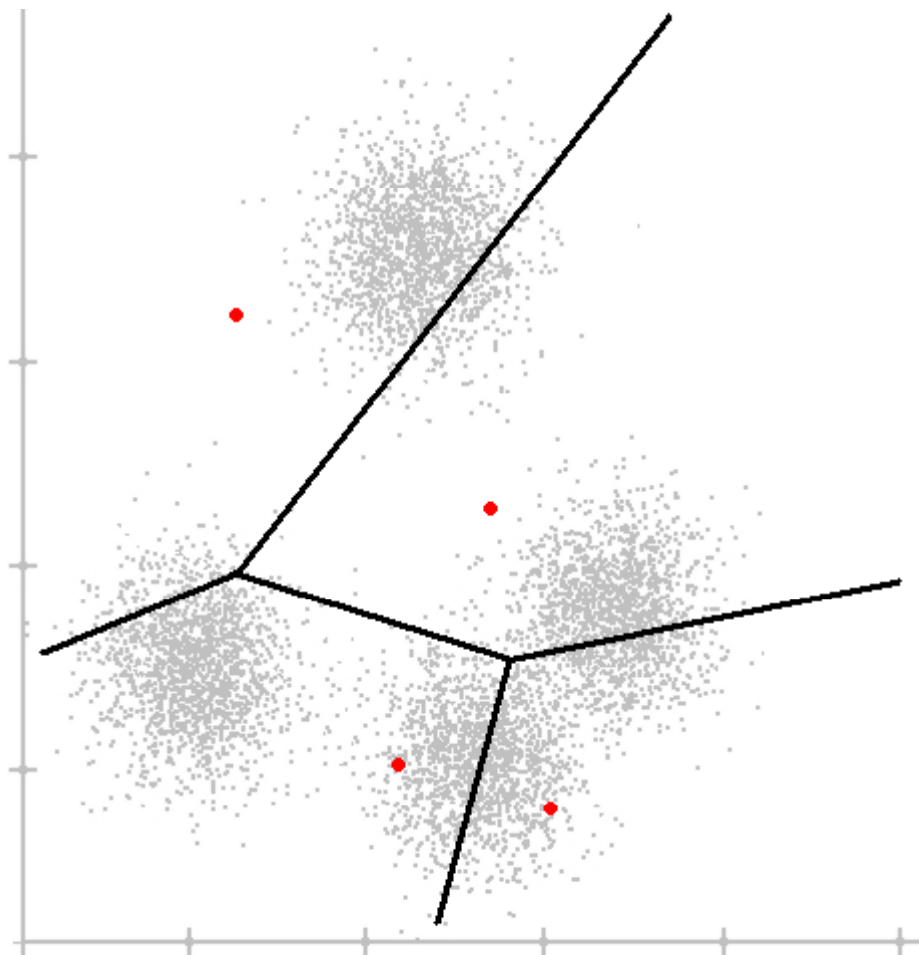
Four
populations
on 2D
display



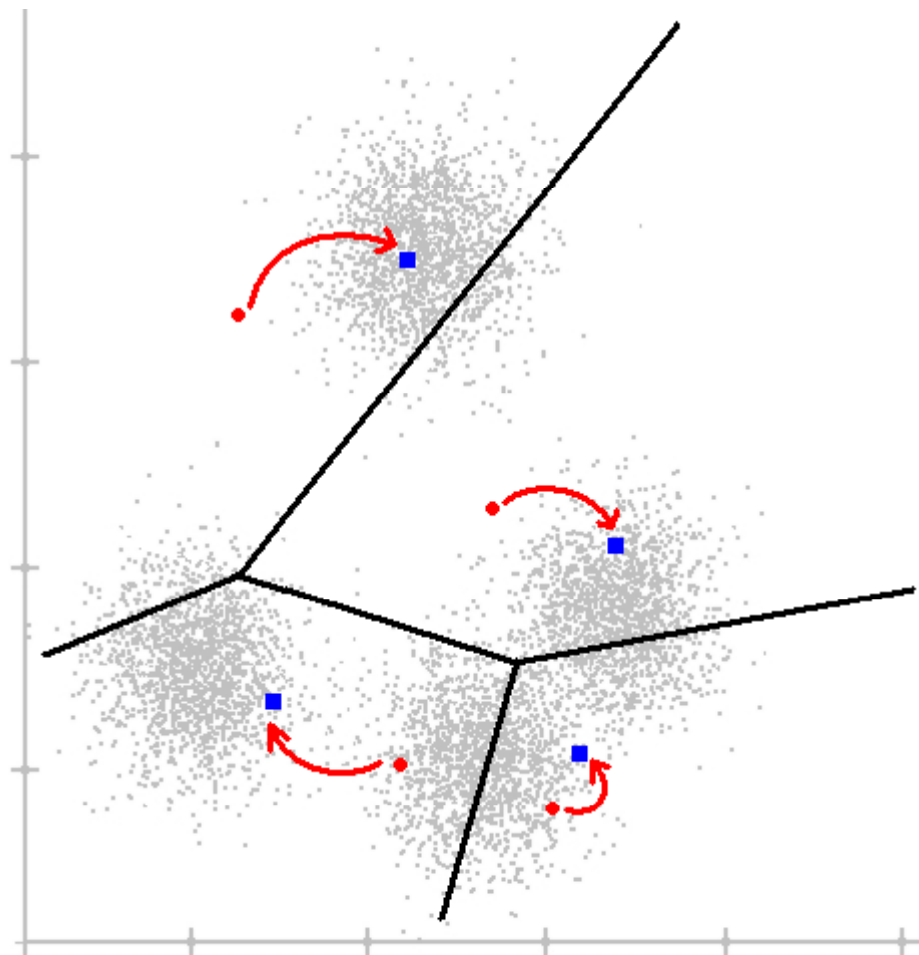
Let $K=4$;
Select random
seeds



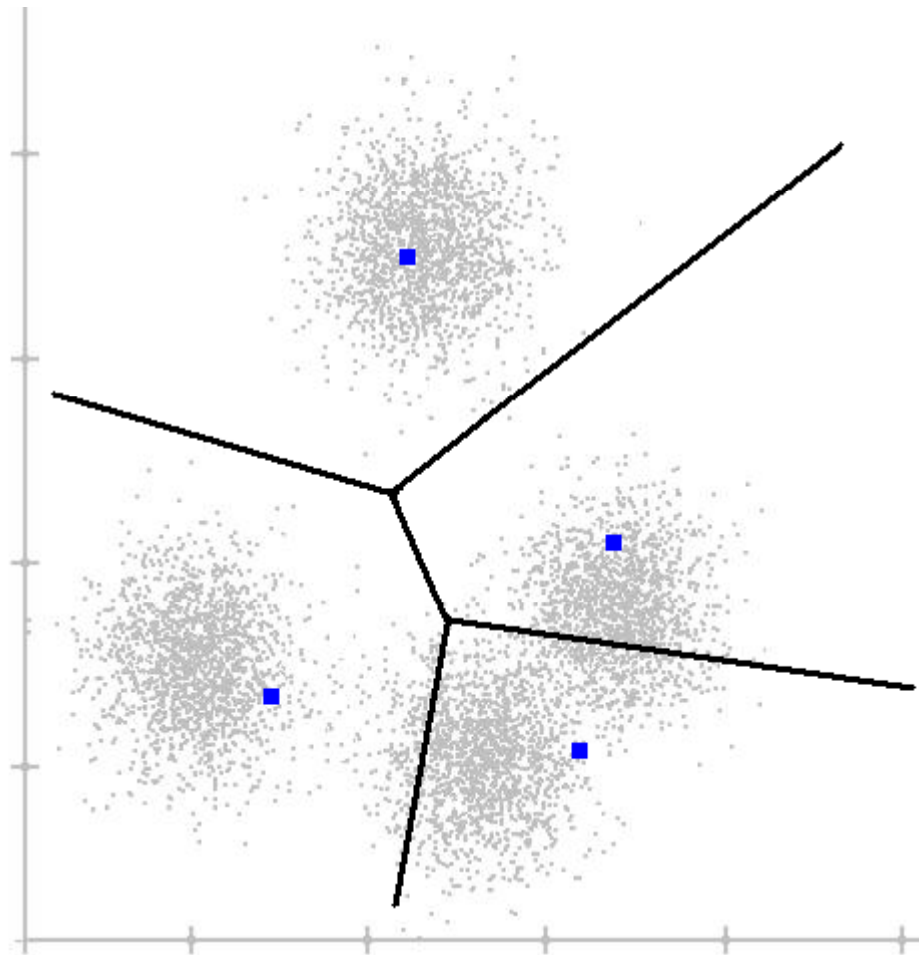
Space
partitioning
based on
centroids



Recalculate
centroids

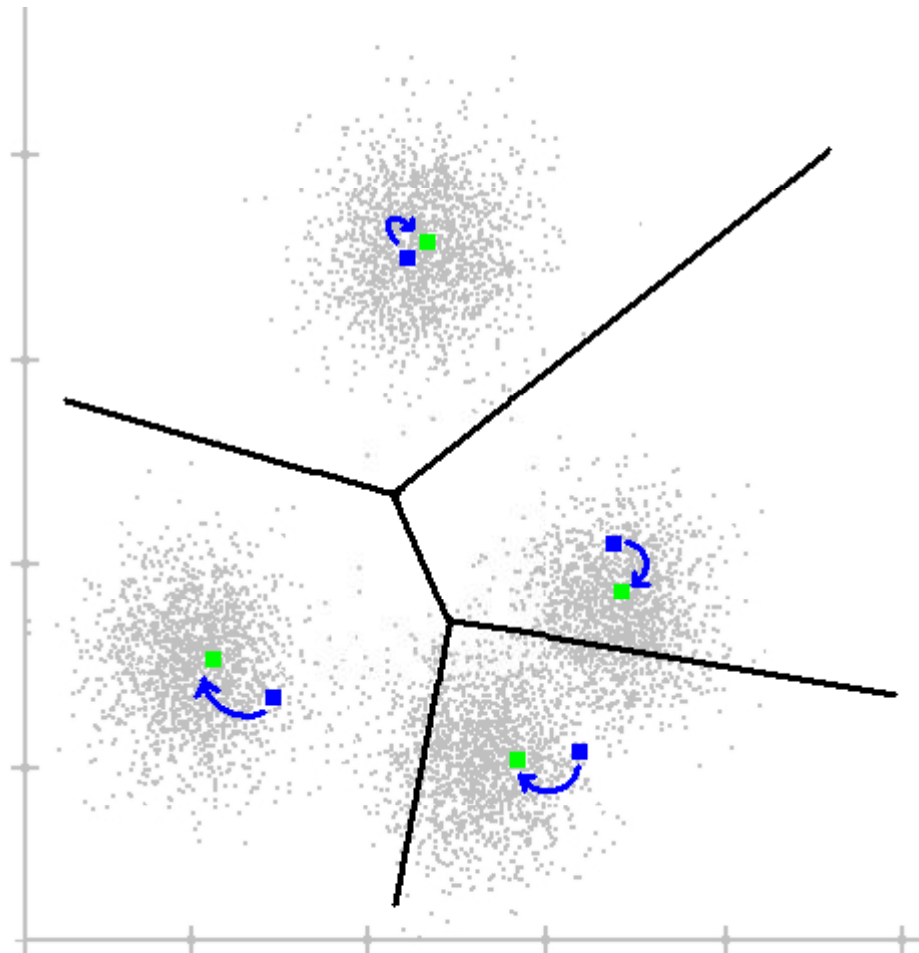


Repartition
based on
new
centroids

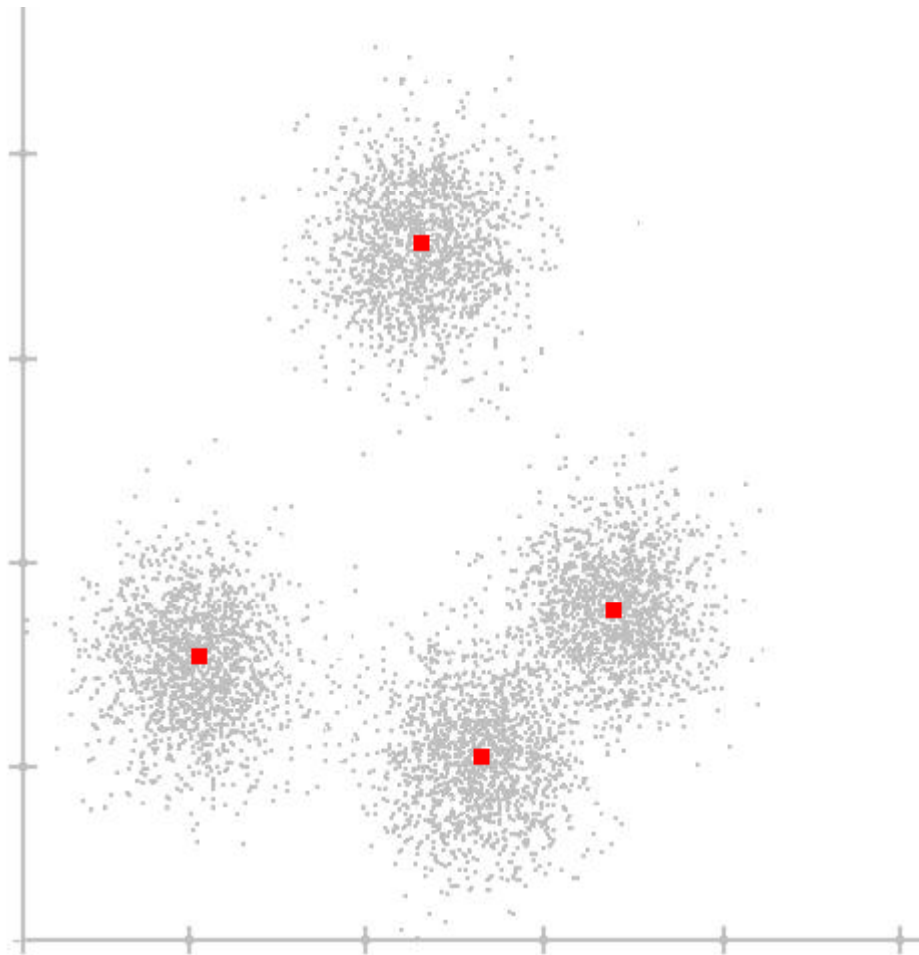


Repeat the
procedure
many times

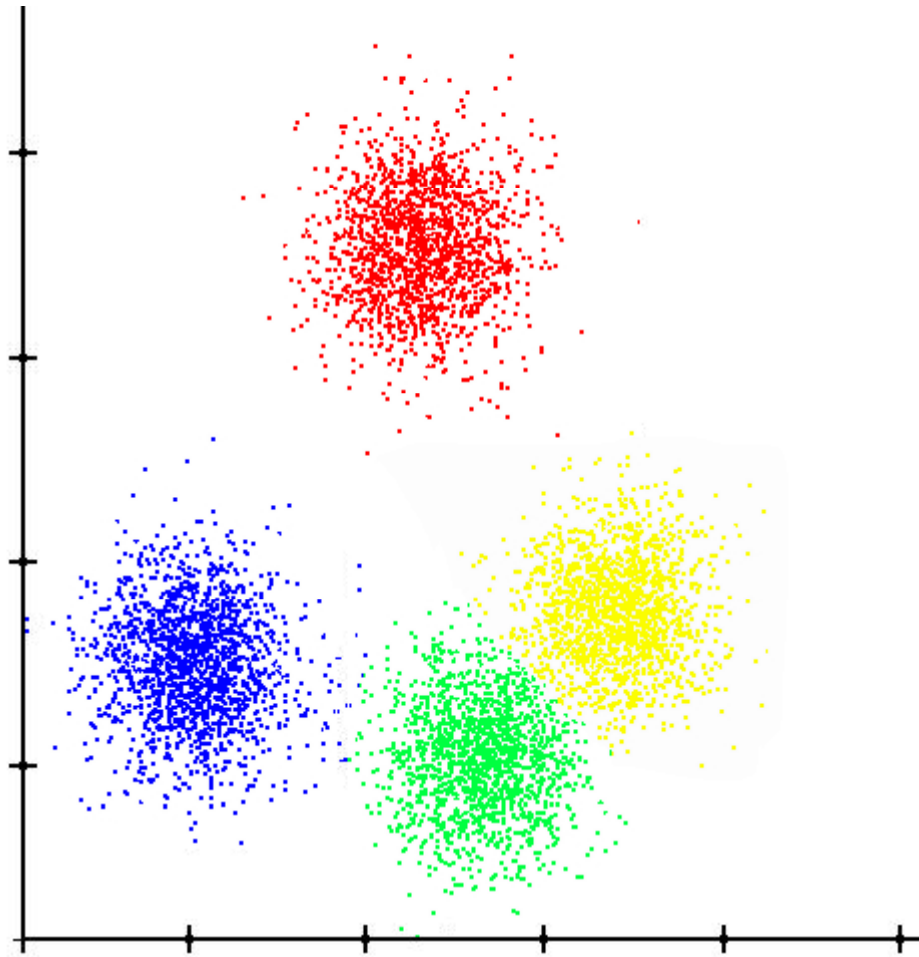
... ..



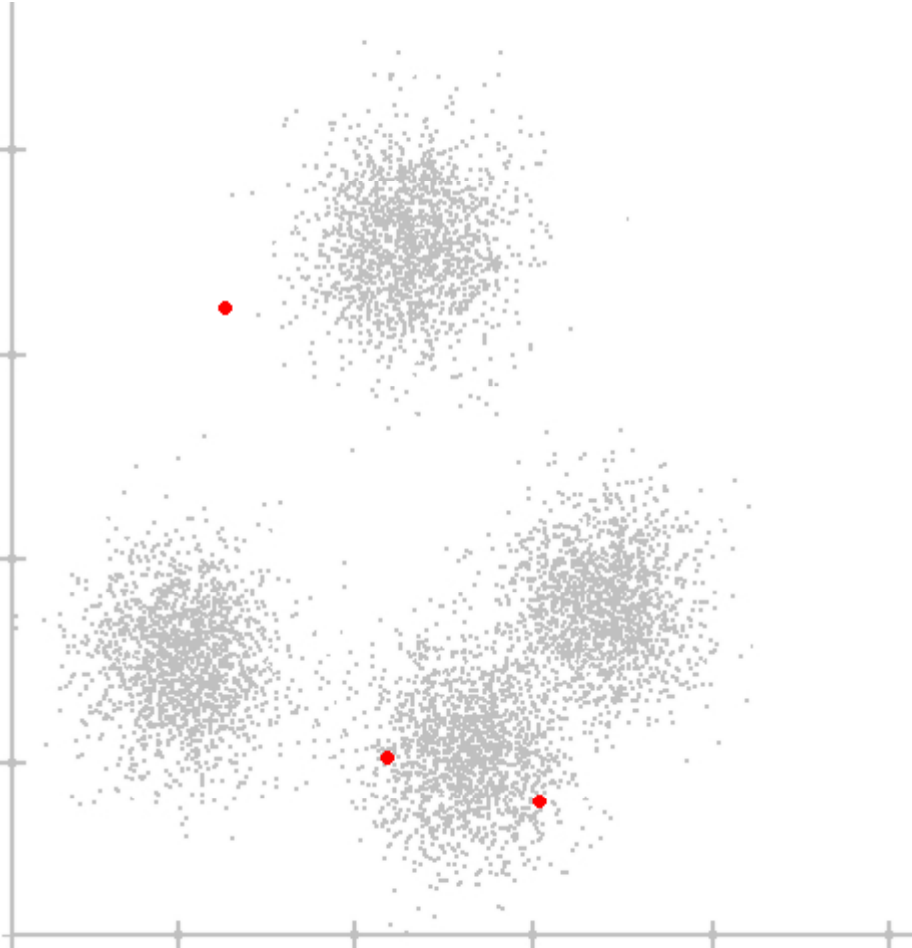
Final
centroids



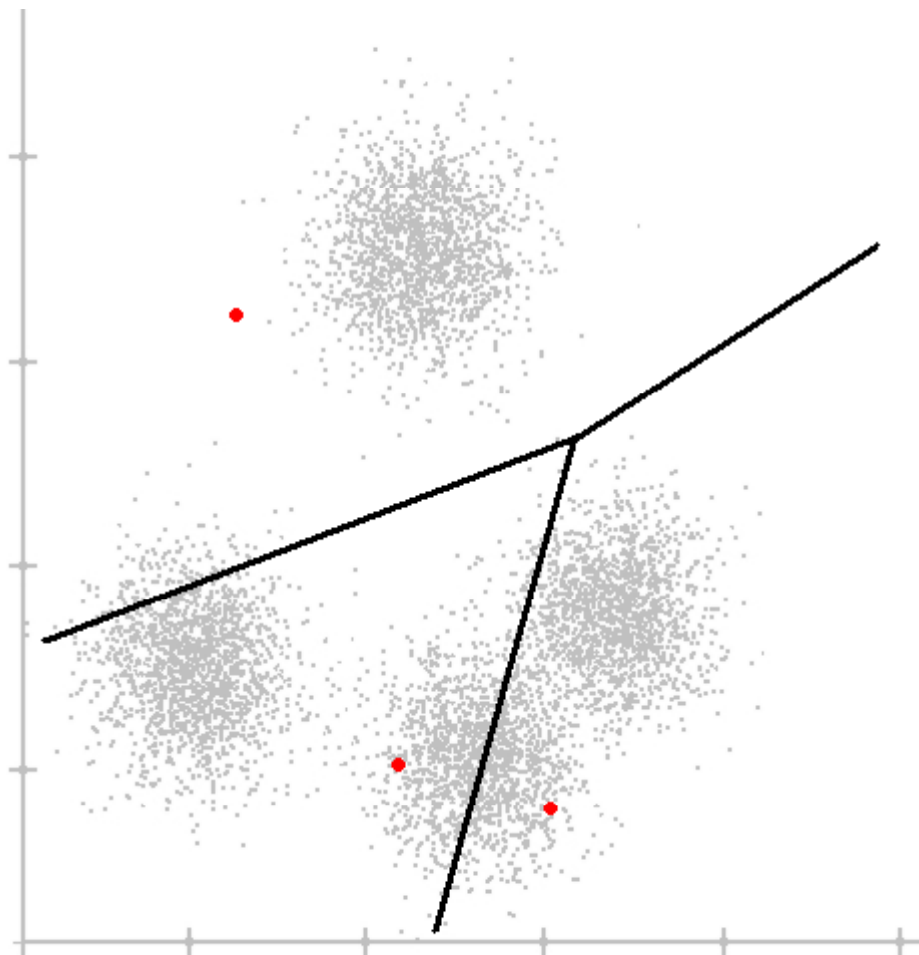
Final
clustering
results



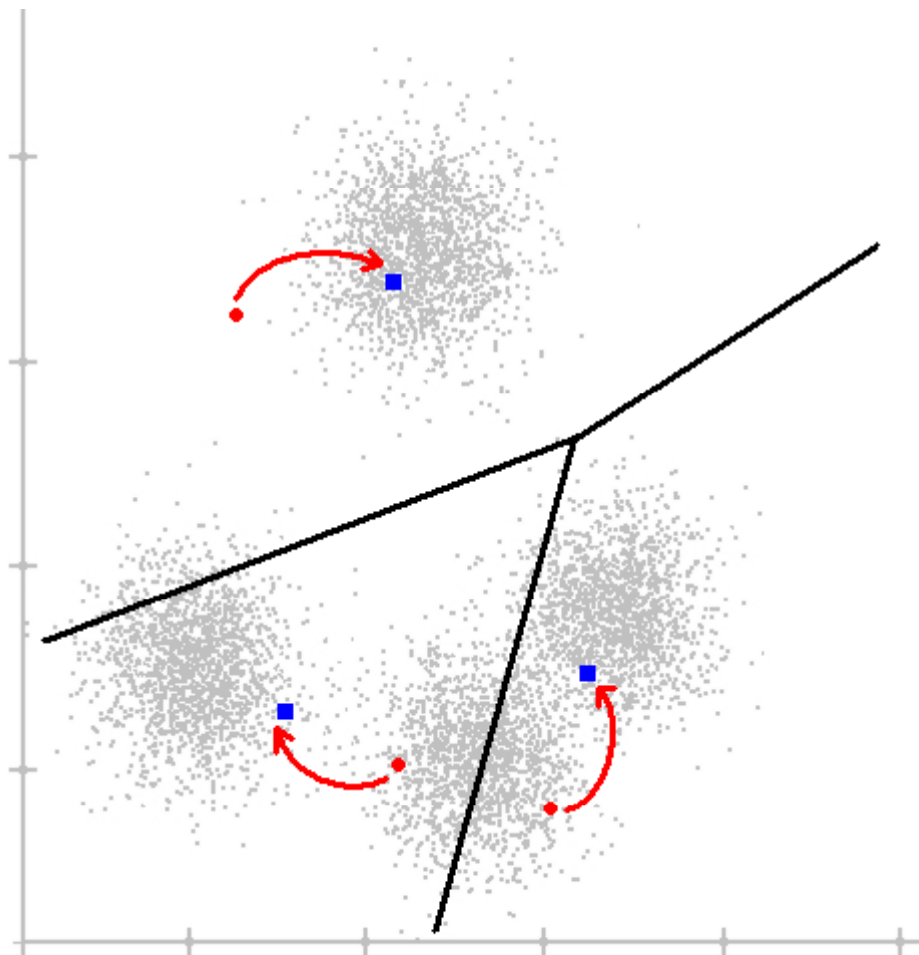
Let $K=3$



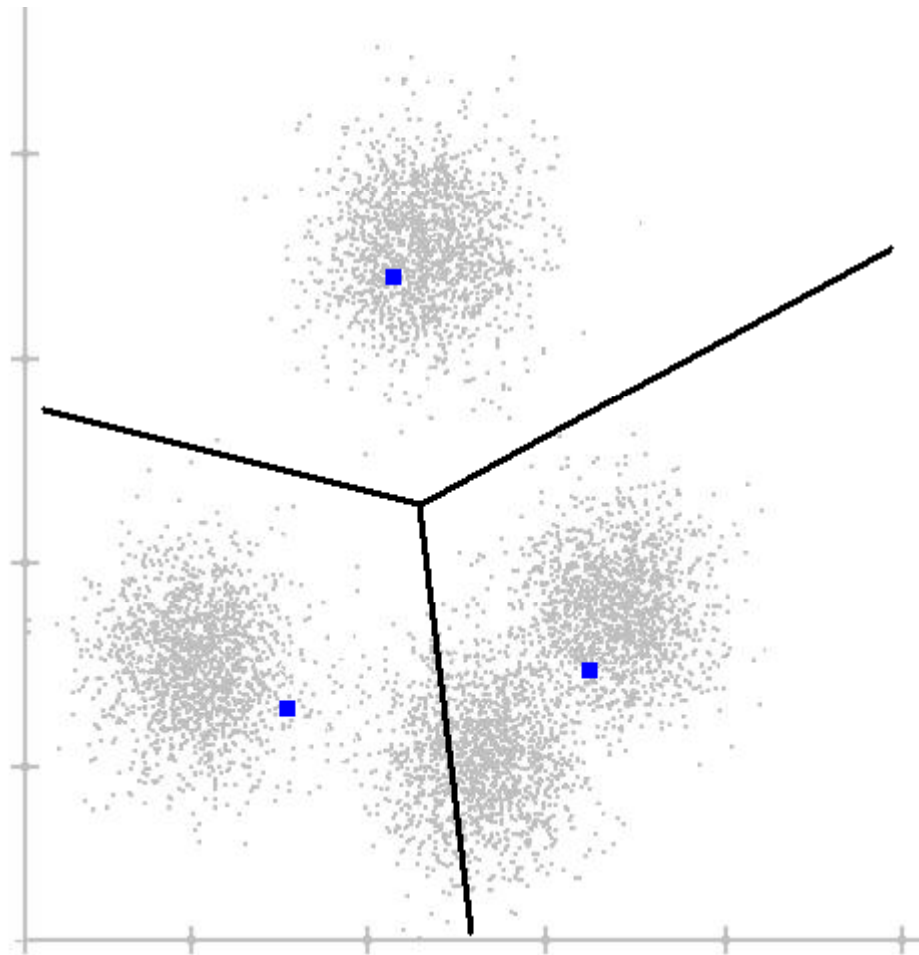
Space
partitioning
based on
centroids



Recalculate
centroids

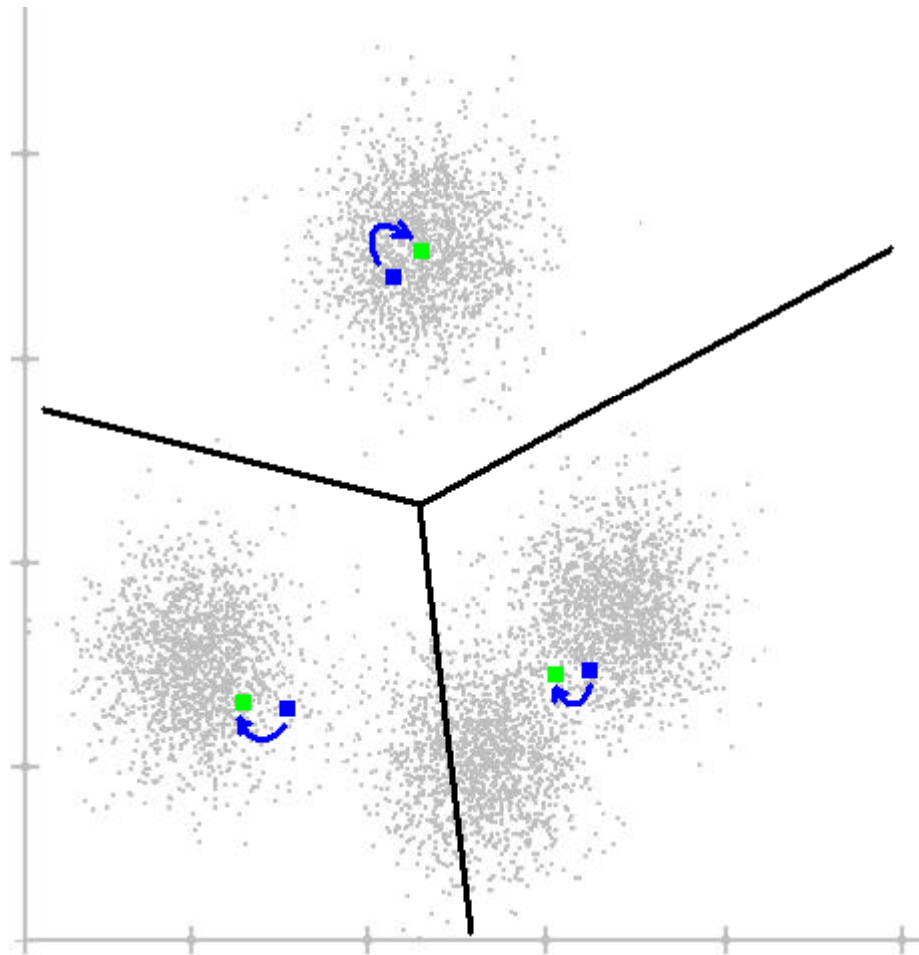


Repartition
based on
new
centroids

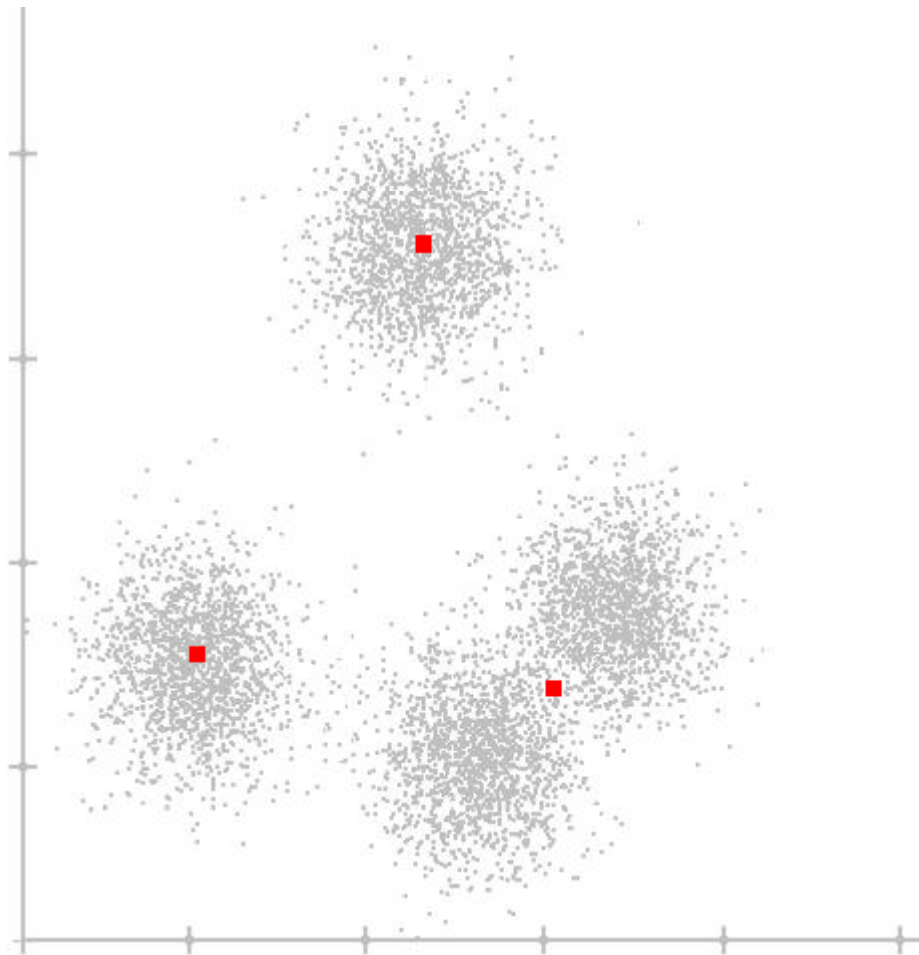


Repeat the
procedure

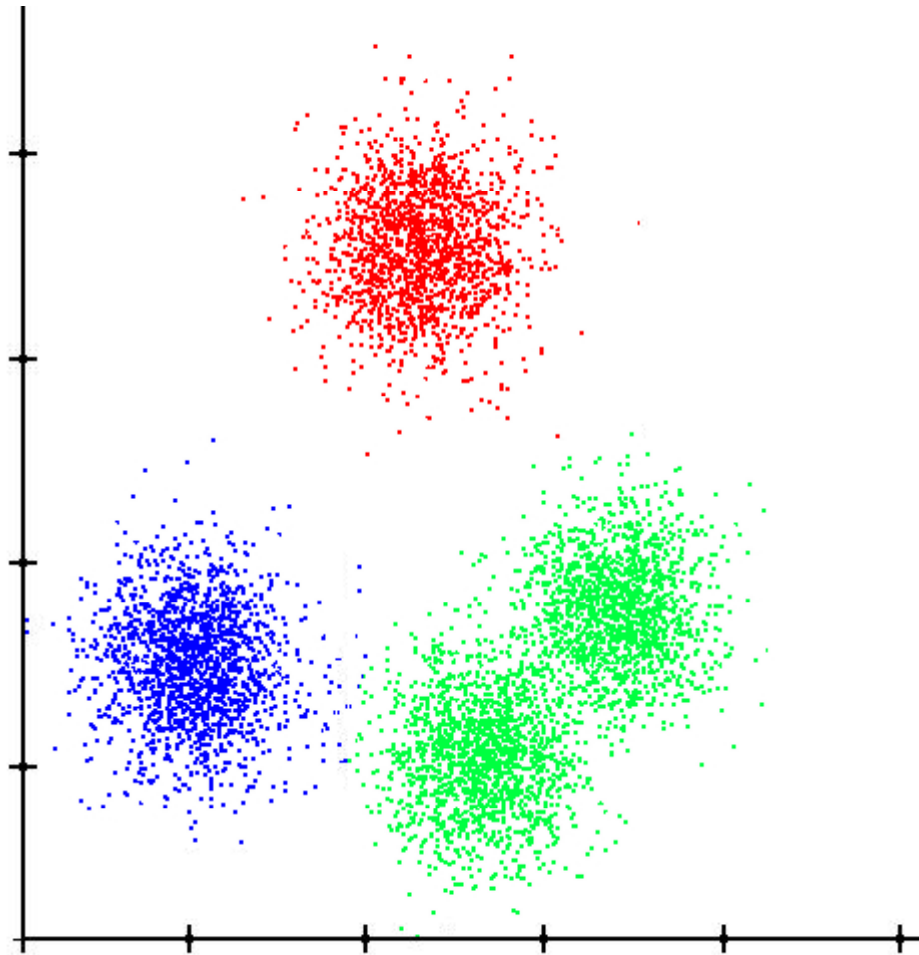
... ..

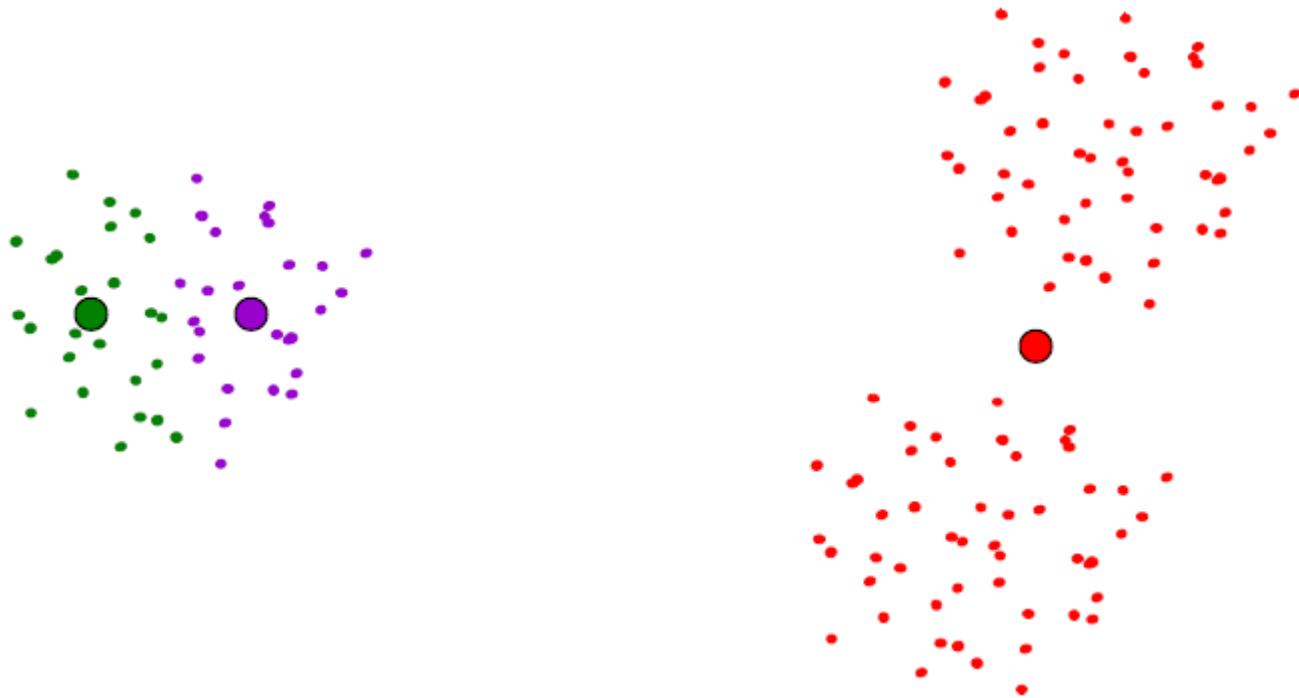


Final
Centroids



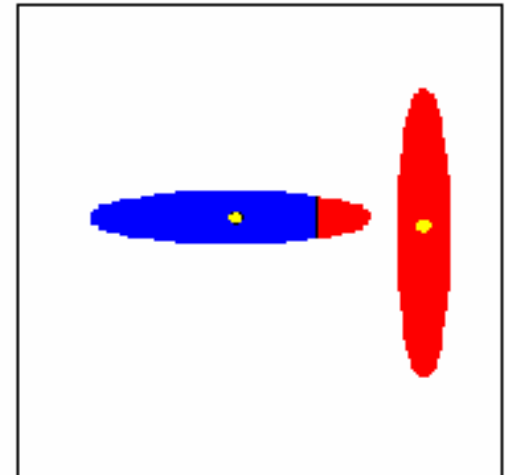
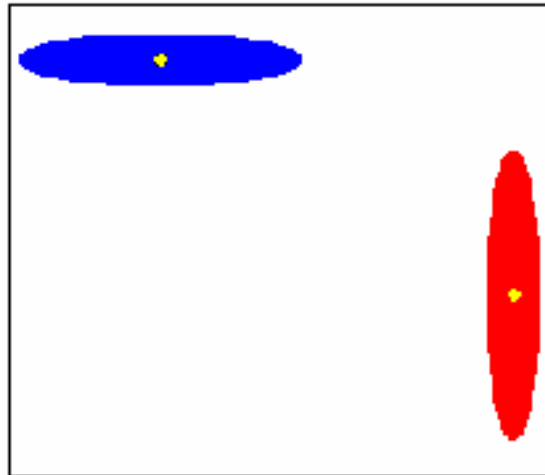
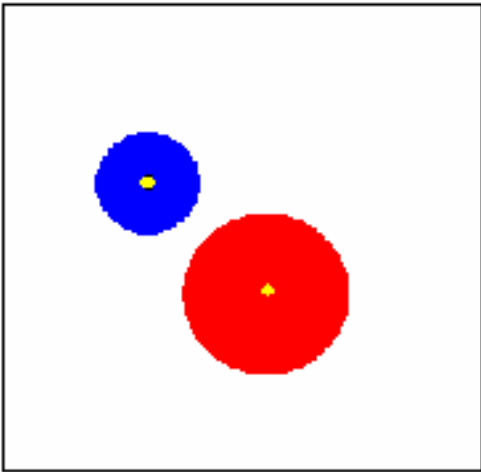
Final
clustering
results



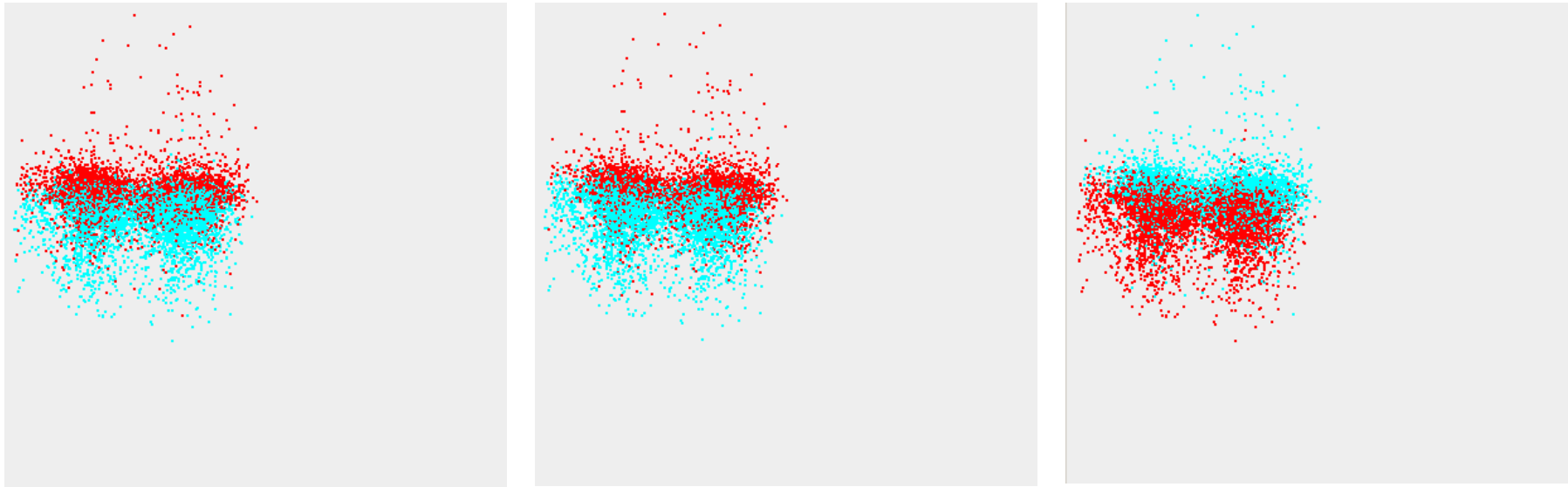


Seeds trapped in local optimum even if K is correct

Non-spherical populations



K-means Applied to High-Dimensional Data



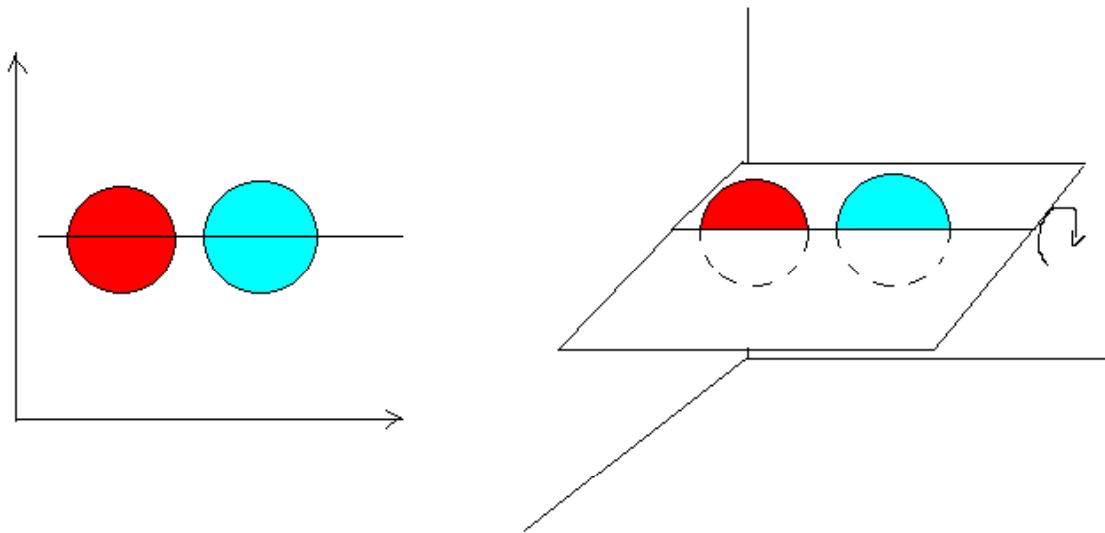
Three different ways used to generating random seeds

Number of Iterations = 1000, $K=2$

“For high dimensional data clustering, standard algorithms such as EM and K-means are often trapped in local minimum”

Ding C, He X, Zha H, Simon HD. Adaptive dimension reduction for clustering high dimensional data. In: *Proceedings of IEEE International Conference on Data Mining*.

Bradley PS, Fayyad UM. Refining initial points for K-means clustering. In: *Proceedings of the Fifteenth International Conference on Machine Learning*.



When number of dimension increases, there are more and more local optimum traps. This is also called ***Curse of Dimensionality***.

Therefore

Dimensions need to be reduced

However, the relationship between dimension selection and clustering is *chicken-egg*:

- to cluster high-dimensional data, dimensionality must be reduced (due to curse of dimensionality)
- it is more effective to select dimensions within individual data clusters than for whole dataset



Flow cytometry clustering
without K

The Procedure of flock

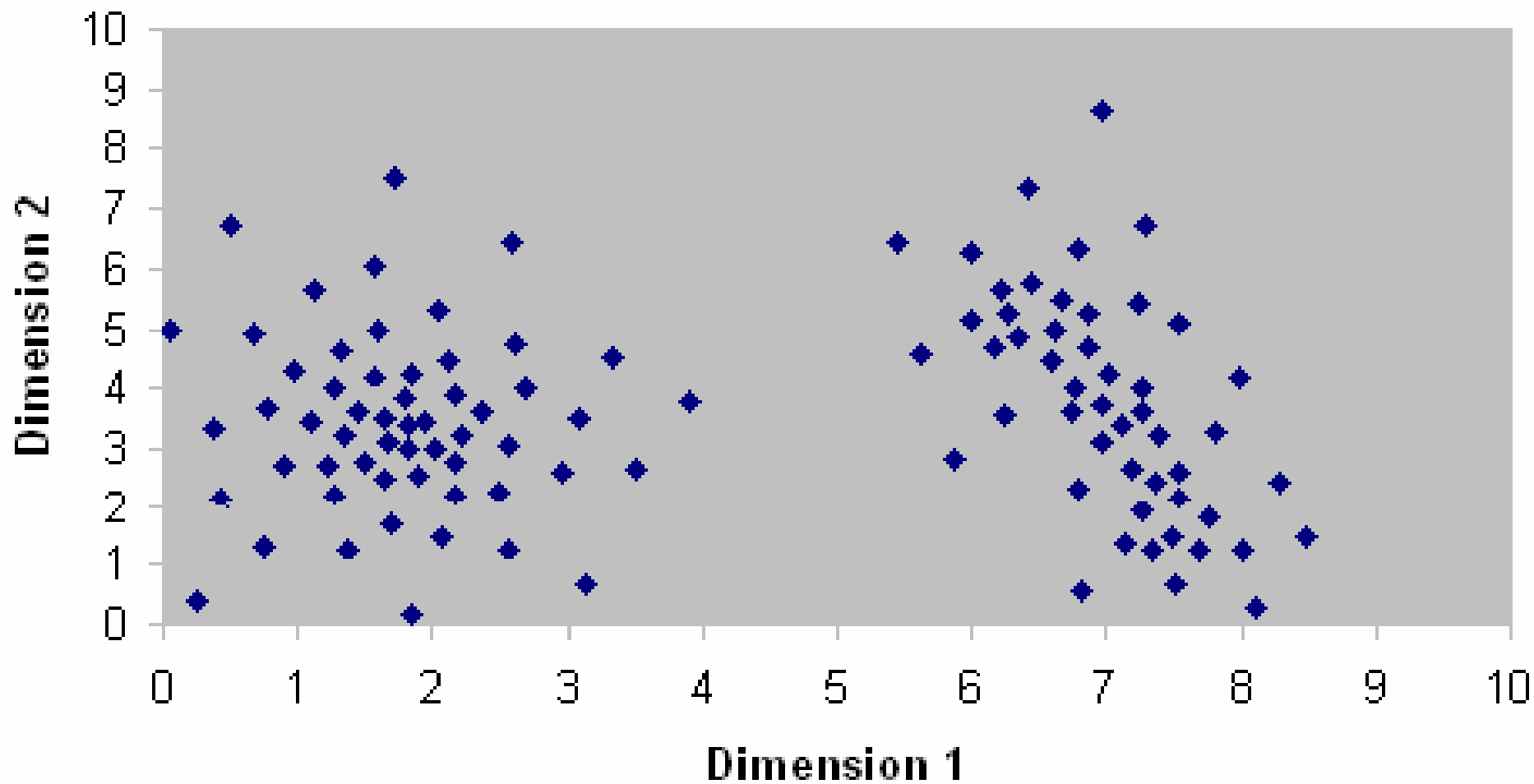
Flow cytometry clustering
without K

- 1) Generate initial clusters (yes, chicken first!)
 - Parameter selection
- 2) Normalize dimensions within clusters
- 3) Select dimensions for initial clusters
- 4) Partition and merge the initial clusters in their selected subspaces
- 5) Output the final clusters

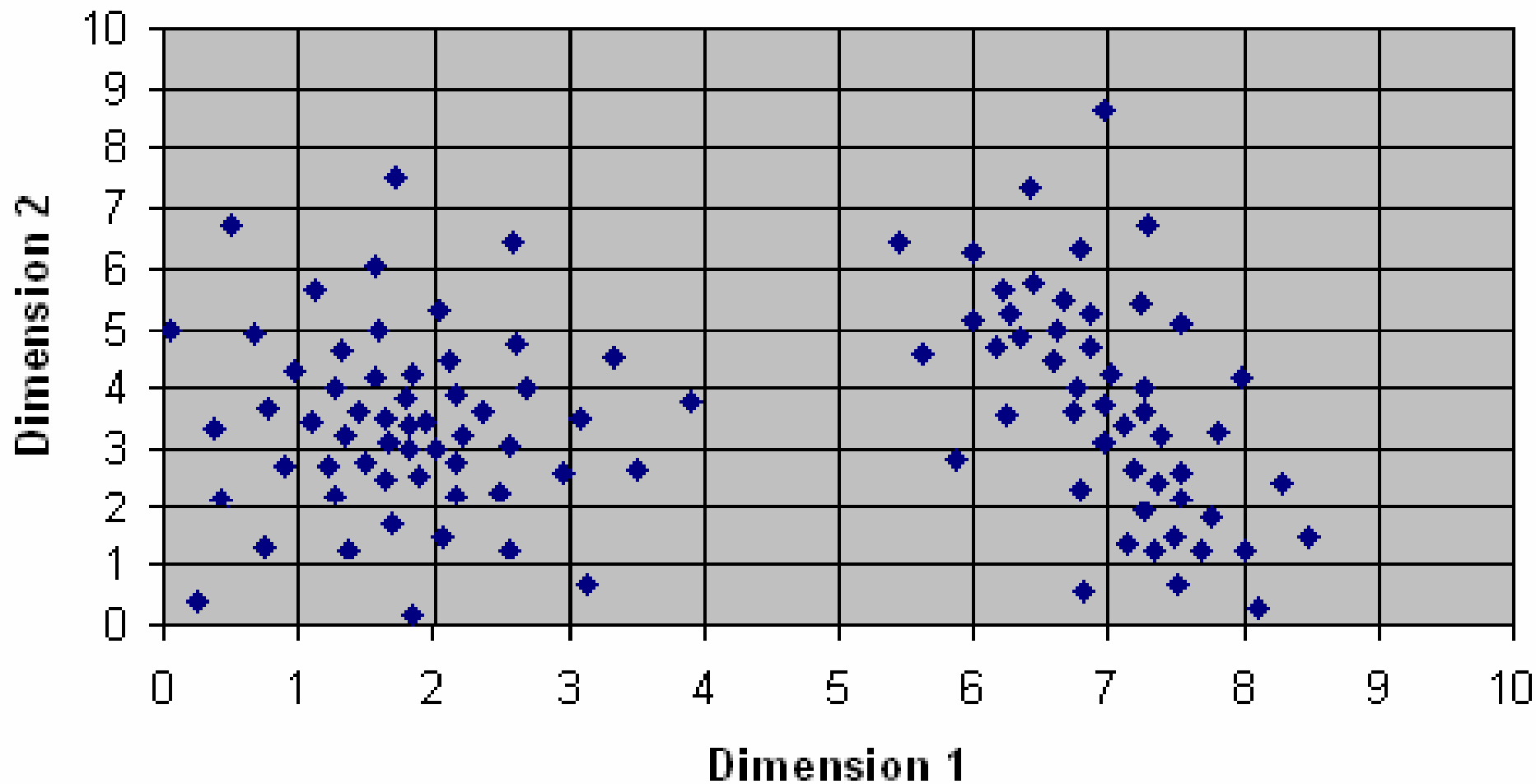
*Details of each step in following slides

Generation of Initial Clusters

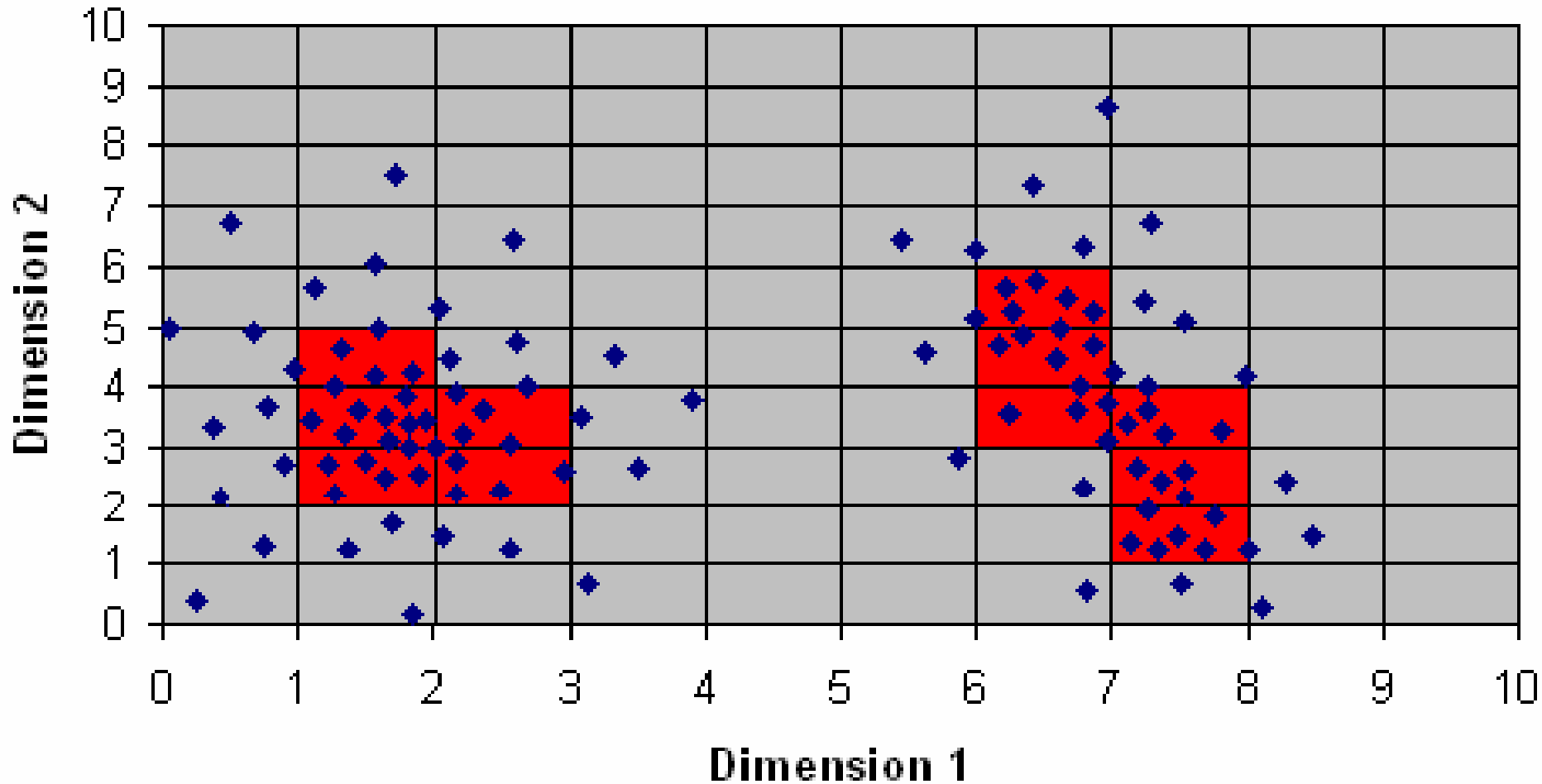
2D example



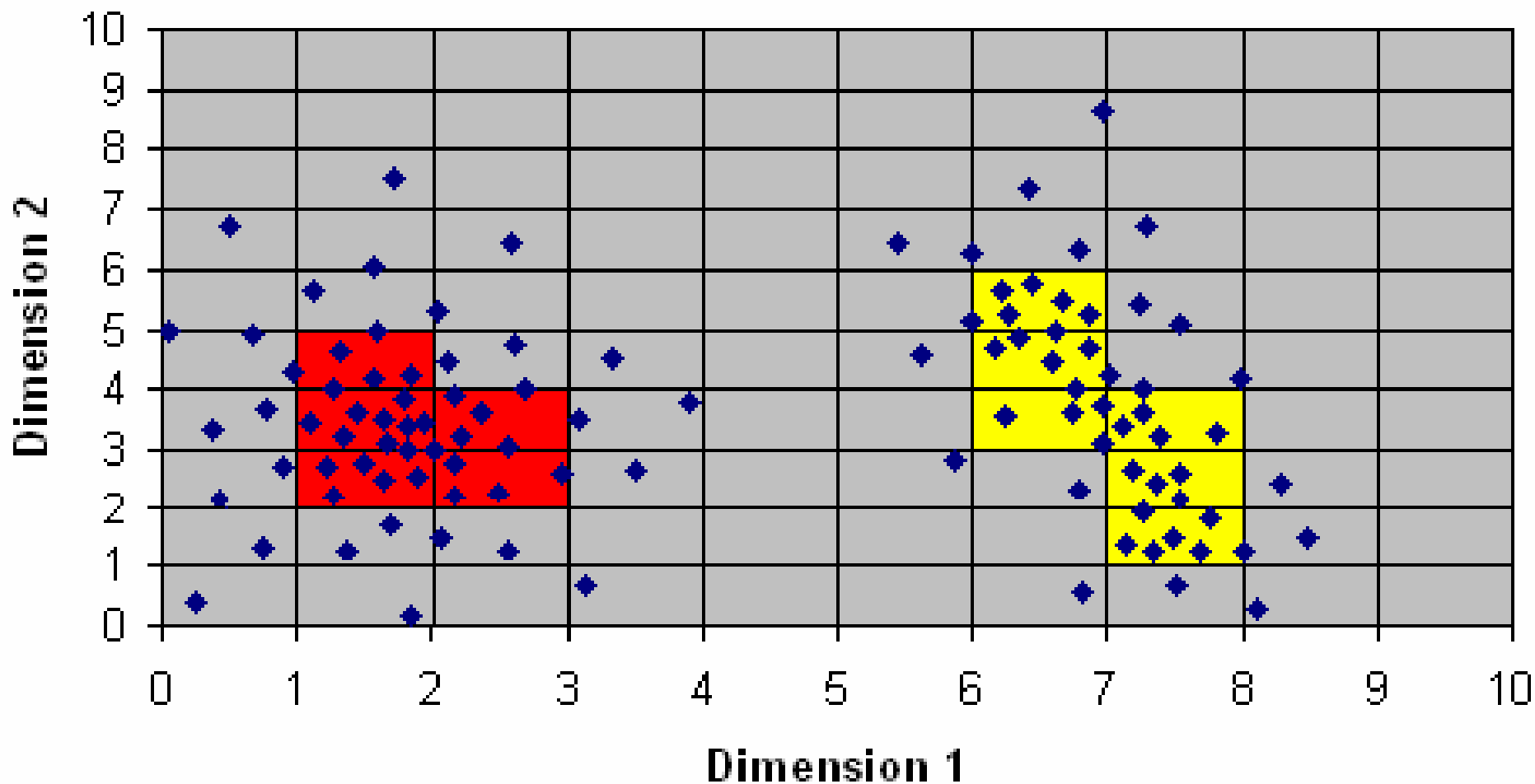
Divide with hyper-grids



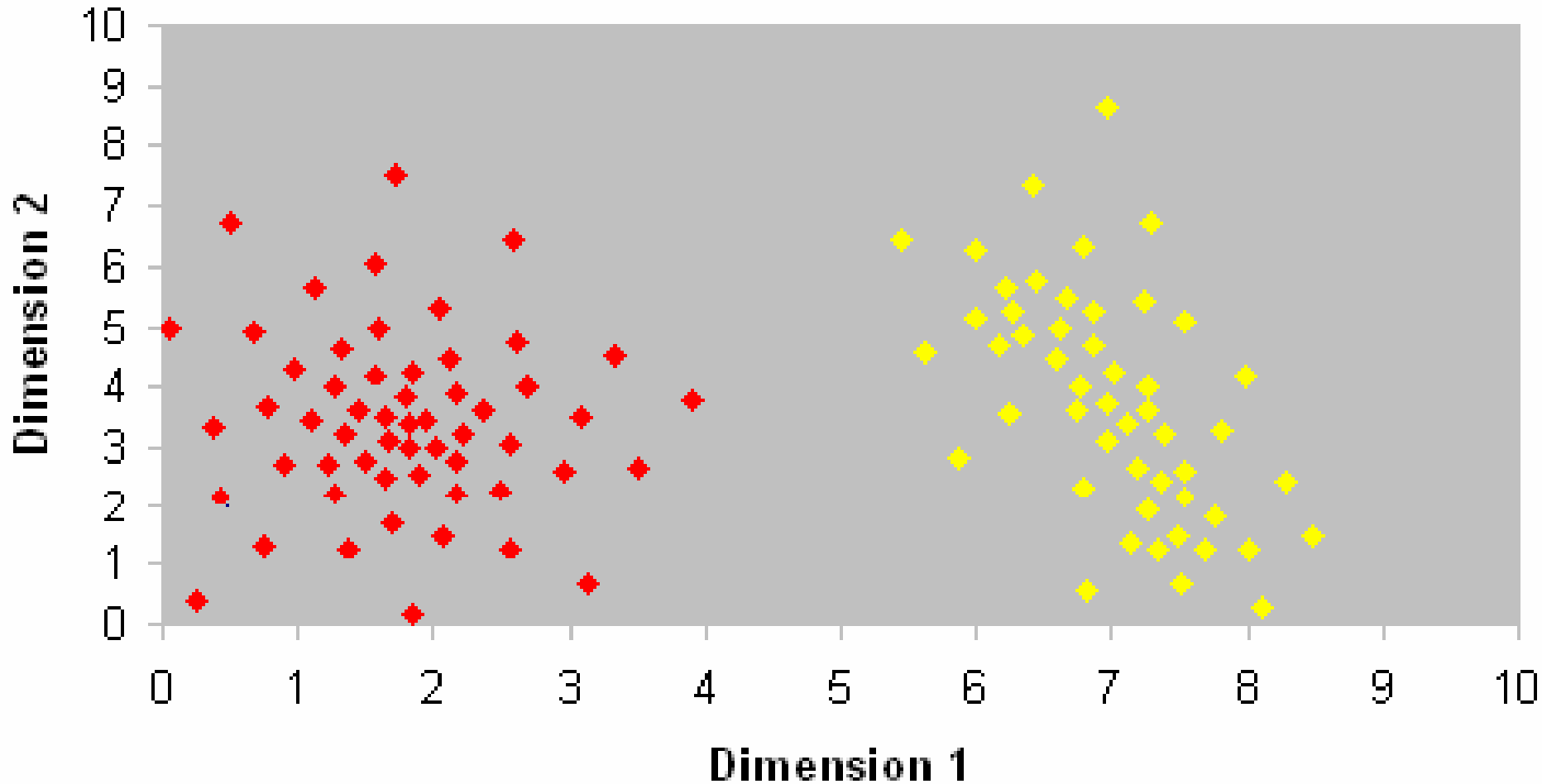
Find dense hyper-regions



Merge neighboring dense hyper- regions



Clustering based on region centers



Bin selection methods

Goal is to minimize the Mean Squared Error

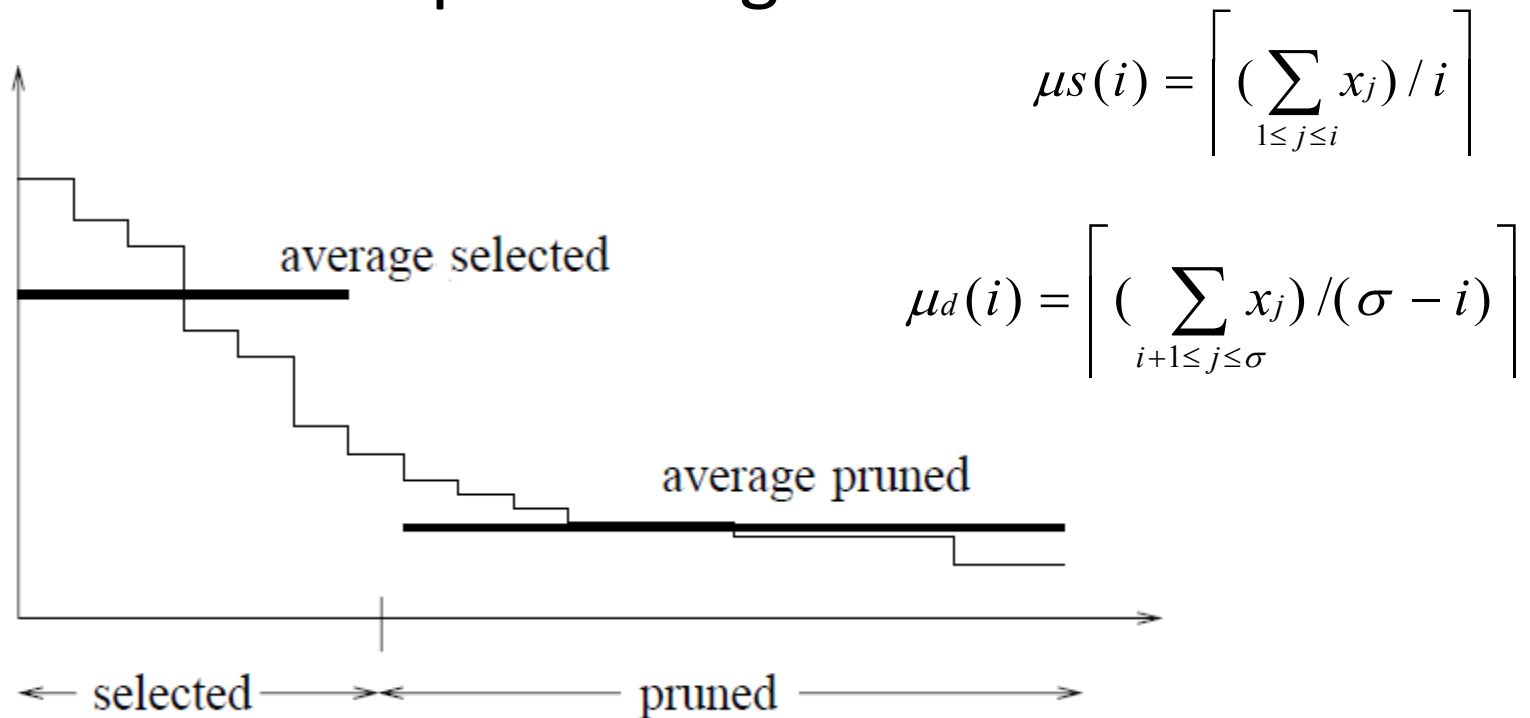
$$L(h(x), f(x)) = \int (h(x) - f(x))^2.$$

- Scott's method $v_{\text{scott}} = 3.49sN^{-1/3}.$
- **Stone's method** $K(v, M) = \frac{1}{v} \left(\frac{2}{N-1} - \frac{N+1}{N-1} \sum_{m=1}^M \pi_i^2 \right)$
- Knuth's method, to maximize

$$N \log M + \log \Gamma\left(\frac{M}{2}\right) - M \log \Gamma\left(\frac{1}{2}\right) - \log \Gamma\left(N + \frac{M}{2}\right) + \sum_{k=1}^M \log \Gamma\left(n_k + \frac{1}{2}\right) + K$$

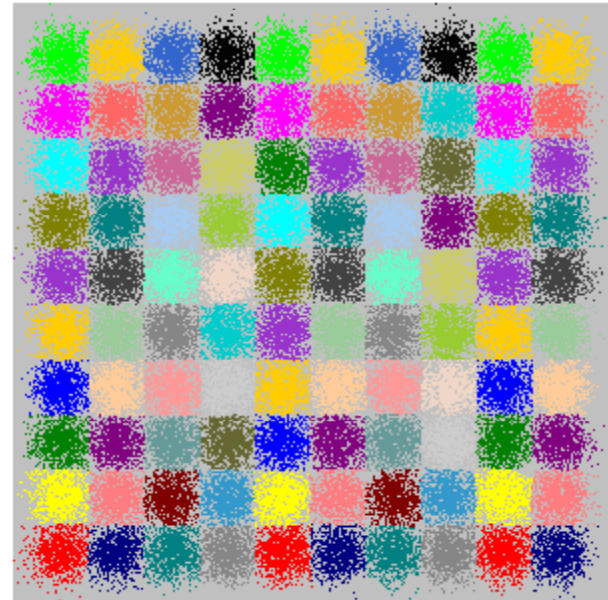
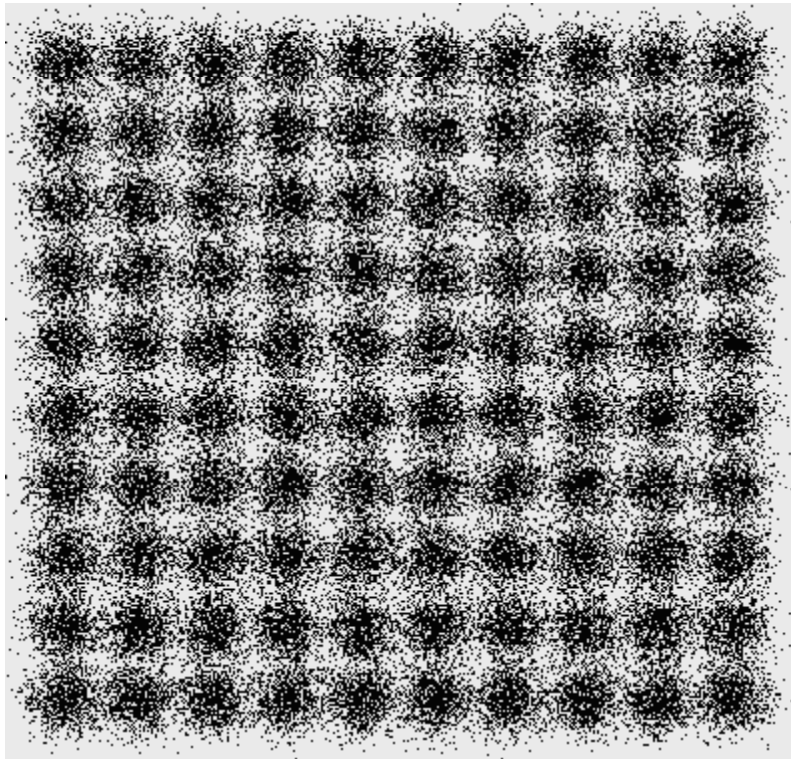
Density threshold selection

- Minimum description length



$$L(i) = \log_2(\mu_s(i)) + \sum_{1 \leq j \leq i} \log_2(|x_j - \mu_s(i)|) + \log_2(\mu_d(i)) + \sum_{i+1 \leq j \leq \sigma} \log_2(|x_j - \mu_d(i)|)$$

Simulation Study



Birch dataset (Zhang et al, SIGMOD 1996)

Two assumptions with the above model

- 1) The center area is denser than the surrounding area in a population
- 2) There is only one group of adjacent hyper-regions in one population

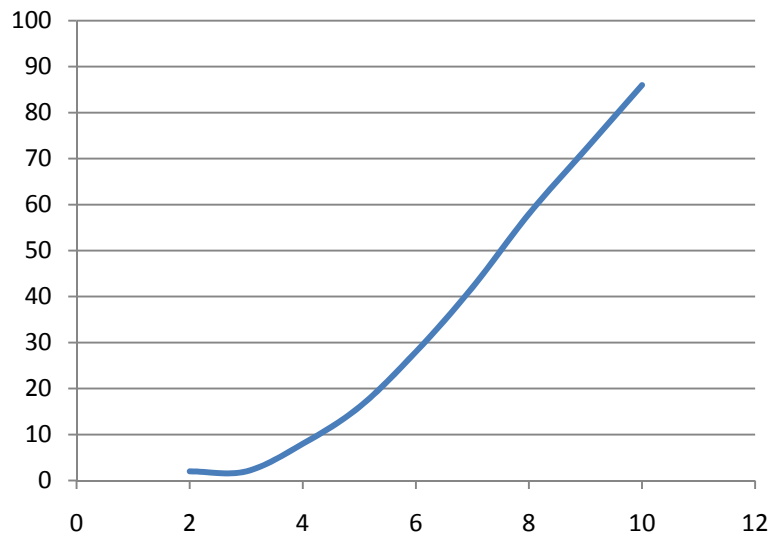
When number of dimensions increases:

- 1) Assumption 1 may not hold for a sparse population; further partitioning to identify the sparse population may be necessary
- 2) There could be multiple adjacent hyper-regions within one population; they need to be merged.

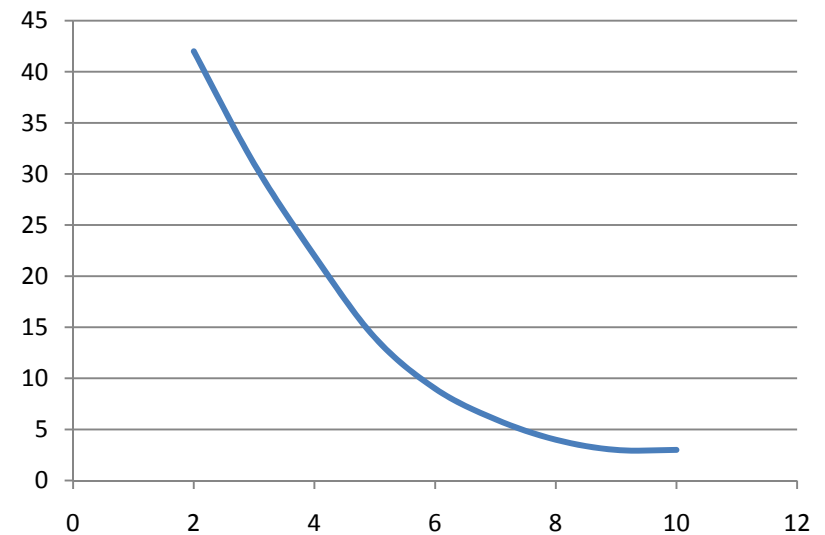
Merging and partitioning will be done in the reduced-dimensional space

Density Variability in High-Dimensional Data Space

Fix the number of bins and density threshold, and use a Gaussian simulator to simulate 2-d, ..., 10-d data with 2 Gaussian clusters



X-axis: Number of dimensions
Y-axis: Number of groups of adjacent hyper-regions



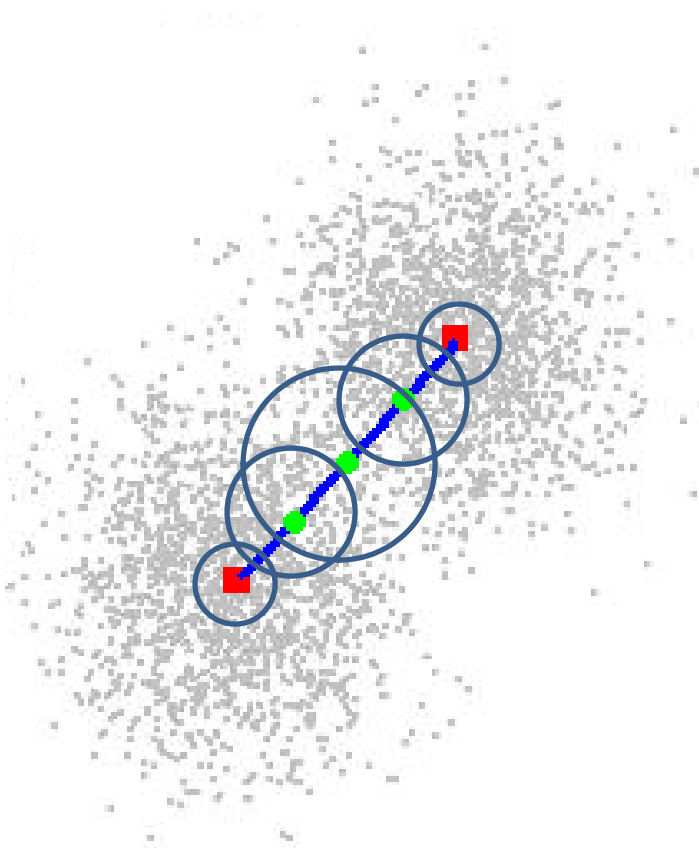
X-axis: Number of dimensions
Y-axis: Number of bins selected by Stone's Method

Dimension Selection and Cluster Merging

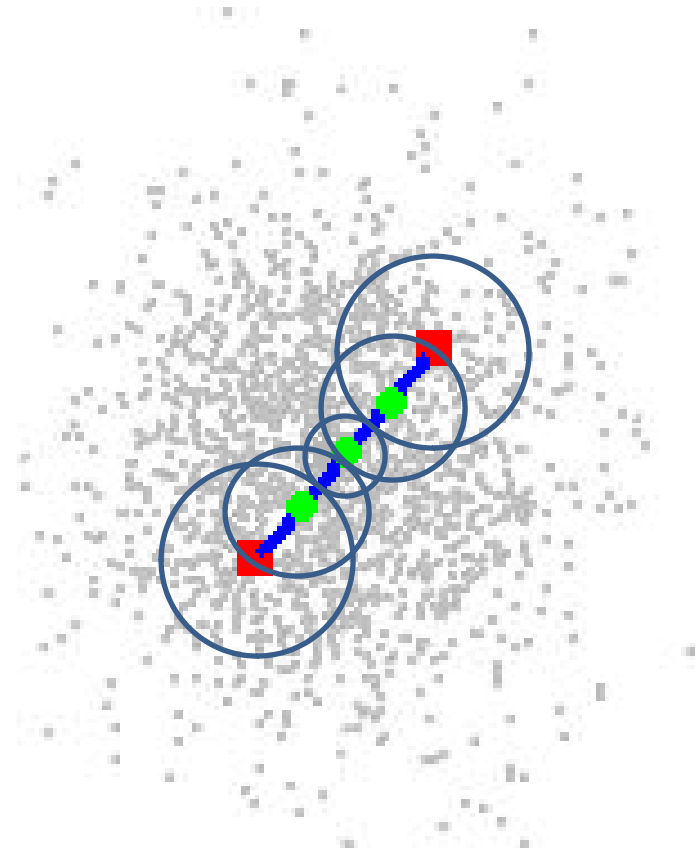
- 1) 0-1 column-wise normalize each cluster
- 2) Select 3 dimensions for each cluster based on standard deviations (if number of dimensions < 3 , all dimensions are used)
- 3) Partition a cluster into two, *if necessary* (this step can be optionally repeated)
- 4) 0-1 column-wise normalize each pair of partitions
- 5) Select 3 dimensions for each pair of partitions
- 6) Starting from the pair that are closest in the 3-dimensional space, merge a pair of partitions, *if necessary*
- 7) Repeat Steps 4) to 6) until there is no pair to merge

Merging/Partitioning Criteria

The most common approach is nearest/mutual neighbor graph, but it is very slow ($O(N^2)$).



Two partitions should not be merged



Two partitions should be merged

Results

FlowCAP Challenges

- Challenge 1 (fully automated)
- Challenge 2 (tuned parameters allowed)
- Challenge 3 (number of clusters known)
- Challenge 4 (manual gating results of a couple of files known)

Evaluation criteria: manual gating

Data: diffuse large B-cell lymphoma, graft versus host disease, normal donors, symptomatic west nile virus, and hematopoietic stem cell transplant

FlowCAP Data

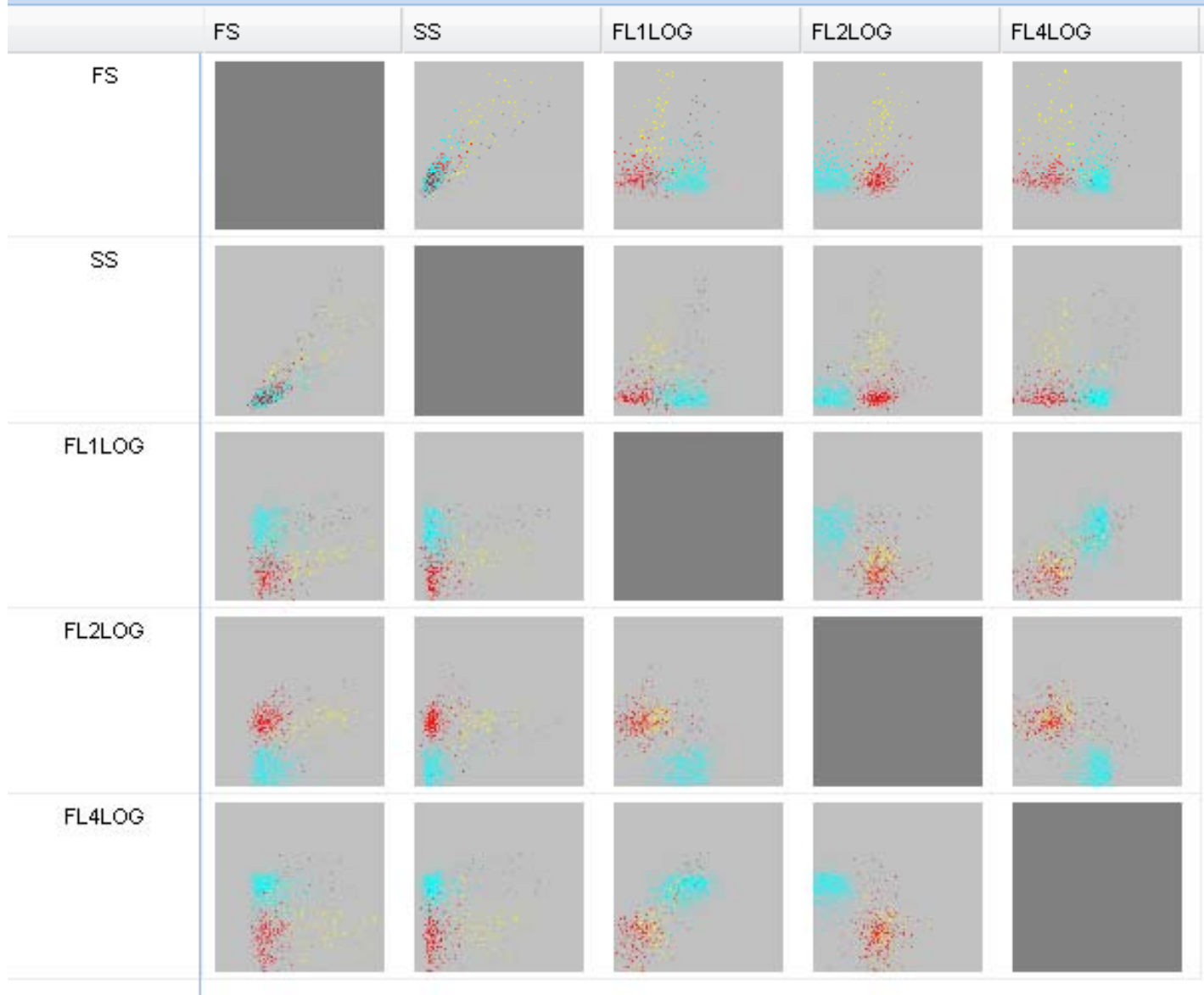
Dataset	#Samples	#Events	Analyte	Detector	Reporter	Provided By
GvHD	12	14,000	CD4 CD8b CD3 CD8	Anti-CD4 Anti-CD8b Anti-CD3 Anti-CD8	FITC PE PerCP APC	BCCRC & TreeStar
DLBCL	30	5,000	CD3 CD5 CD19	Anti-CD3 Anti-CD5 Anti-CD19	CY5 FITC PE	BCCRC
ND	30	17,000	CD56 CD8 CD45 CD3/CD14	Proprietary Proprietary Proprietary Proprietary Anti-CD56 Proprietary Anti-CD8 Proprietary Anti-CD45 Anti-CD3/CD14	FITC PerCPCy5 PacificBlue PacificOrange Qdot605 APC Alexa700 PE PECy5 PECy7	Amgen
WNV	13	100,000	IFN γ CD3 CD4 IL17 CD8 Free Amines	Anti-IFN γ Anti-CD3 Anti-CD4 Anti-IL17 Anti-CD8 NA	PEA PECy5 PECy7 APC AlexaFluor700 CFSE	McMaster
HSCT	30	10,000	CD45.1 Ly65/Mac1 Dead Cells CD45.2	Anti-CD45.1 Anti-Ly65/Mac1 NA Anti-CD45.2	FITC PE PI APC	BCCRC

Challenge 1 (auto)

	Rank Score	Total Runtime
FlowVB	30.5	03:02:23:09
CDP	17	00:01:48:06
FEK	21	00:15:25:00
FLOCK	42.3	00:00:37:38
flowMeans	45.3	00:04:23:27
flowClust/Merge	23	10:13:00:00
FLAME	41.8	00:05:31:12
MM&PCA	26.5	00:00:04:35
MM	25.5	-
SamSPECTRAL	41.5	00:07:21:44
SWIFT	15.5	05:23:24:30

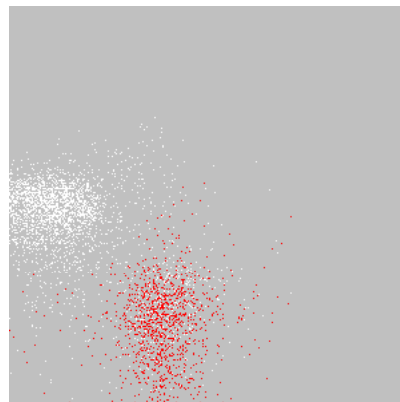
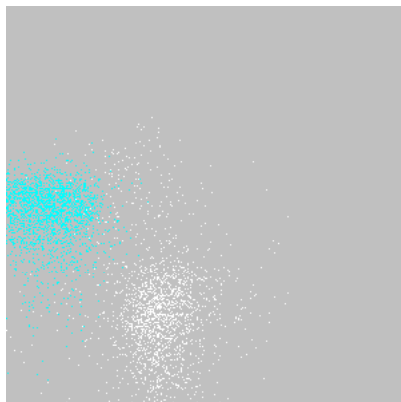
DLBCL_001

Click on any thumbnail image to view and adjust populations of taskID=1 File=001

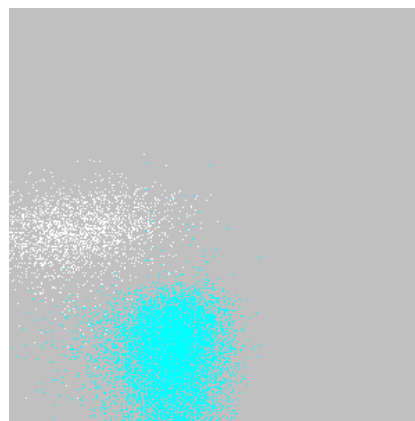
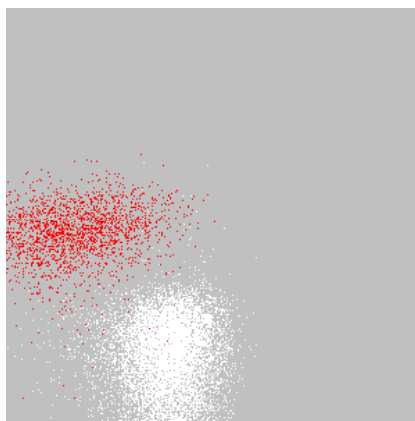


X: FL2; Y: FL4

DLBCL_001

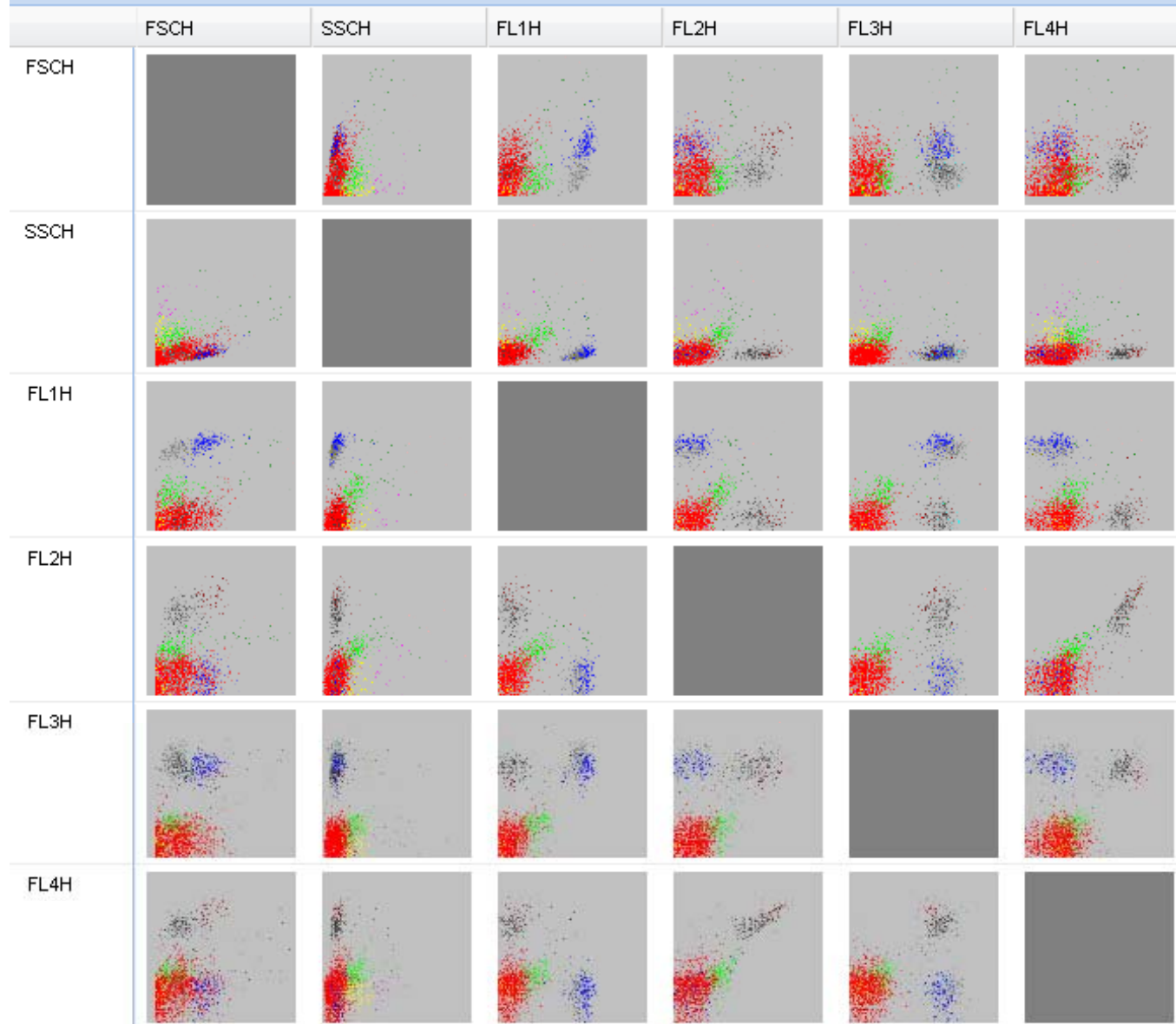


DLBCL_006



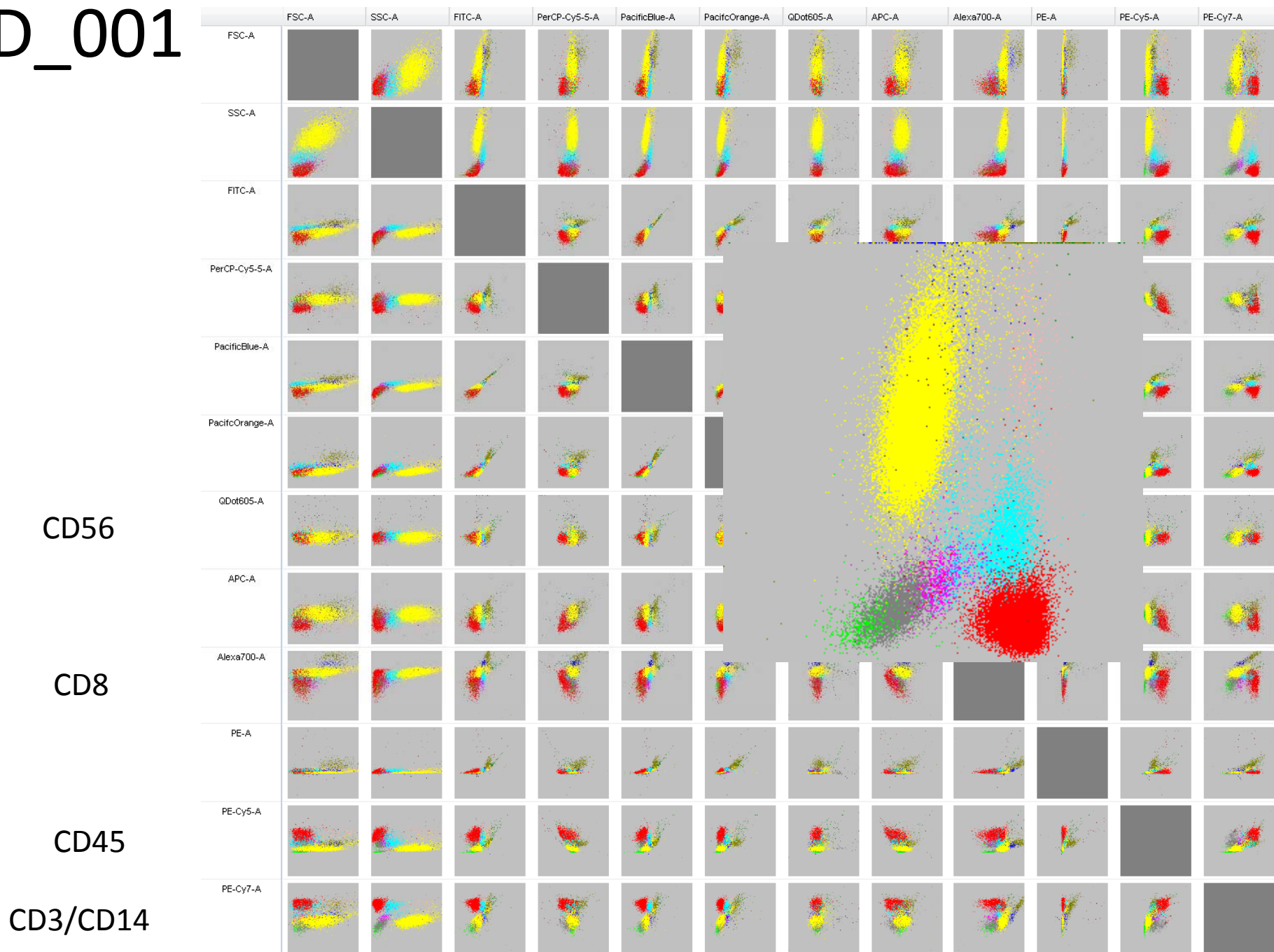
GvHD_001

on any thumbnail image to view and adjust populations of taskID=1 File=001



High-dimensional Data

ND_001



Challenge 2 (tuned)

	WNV	ND	Rank Score	Total Runtime
NMF-curvHDR	0.81 (0.77, 0.84)	0.83 (0.82, 0.84)	15	07:23:04:00
CDP	0.75 (0.71, 0.78)	0.86 (0.84, 0.88)	11	00:00:33:30
FLOCK	0.84 (0.82, 0.86)	0.89 (0.87, 0.91)	25.5	00:00:28:31
flowClust/Merge	0.77 (0.75, 0.79)	0.88 (0.81, 0.92)	21.5	10:13:00:00
FLAME	0.84 (0.82, 0.85)	0.87 (0.86, 0.87)	22.5	-
SamSPECTRAL	0.85 (0.83, 0.88)	0.91 (0.91, 0.92)	25.5	00:13:00:00
SamSPECTRAL-FixK	0.76 (0.71, 0.81)	0.92 (0.91, 0.93)	19	00:08:26:10

WNV	ND
0.81 (0.78, 0.83)	0.85 (0.84, 0.86)
0.71 (0.67, 0.74)	0.86 (0.81, 0.89)
0.78 (0.75, 0.81)	0.81 (0.80, 0.82)
0.83 (0.80, 0.86)	0.91 (0.89, 0.92)
0.88 (0.86, 0.90)	0.85 (0.76, 0.92)
0.77 (0.74, 0.79)	0.73 (0.60, 0.85)
0.80 (0.76, 0.84)	0.90 (0.89, 0.91)
0.64 (0.52, 0.72)	0.76 (0.76, 0.77)
0.69 (0.60, 0.75)	0.75 (0.74, 0.76)
0.75 (0.60, 0.85)	0.92 (0.92, 0.93)
0.69 (0.64, 0.74)	0.87 (0.86, 0.88)

Compared with
Challenge 1

FLOCK in ImmPort (www.immport.org)



About ImmPort

Access Data

Tools

Resources

Ne

About ImmPort

ImmPort, the Immunology Database and Analysis Portal, is a one stop shop to access reference and experiment data for immunologists. ImmPort provides advanced information technology support in the production, analysis, archiving, and exchange of scientific data for the diverse community of life science researchers supported by NIAID/DAIT.

[What is ImmPort](#)

DATA SOURCES

NIAID/DAIT Investigators

Experimental Data

Clinical Study Data

Public Reference Databases



IMMPORT DATABASE

Data Standardization

Quality Control

Data Curation

Maps to Ontologies

IMMPORT TOOLS

Search Data

Visualization

Data Analysis

What You Can Do:

Search Data

Visualize Data

Analyze Data



[ImmPort Flow Cytometry Analysis Component \(FLOCK\)](#)

Identify distinct cell populations in multicolor flow cytometry data with ImmPort's novel clustering algorithms.

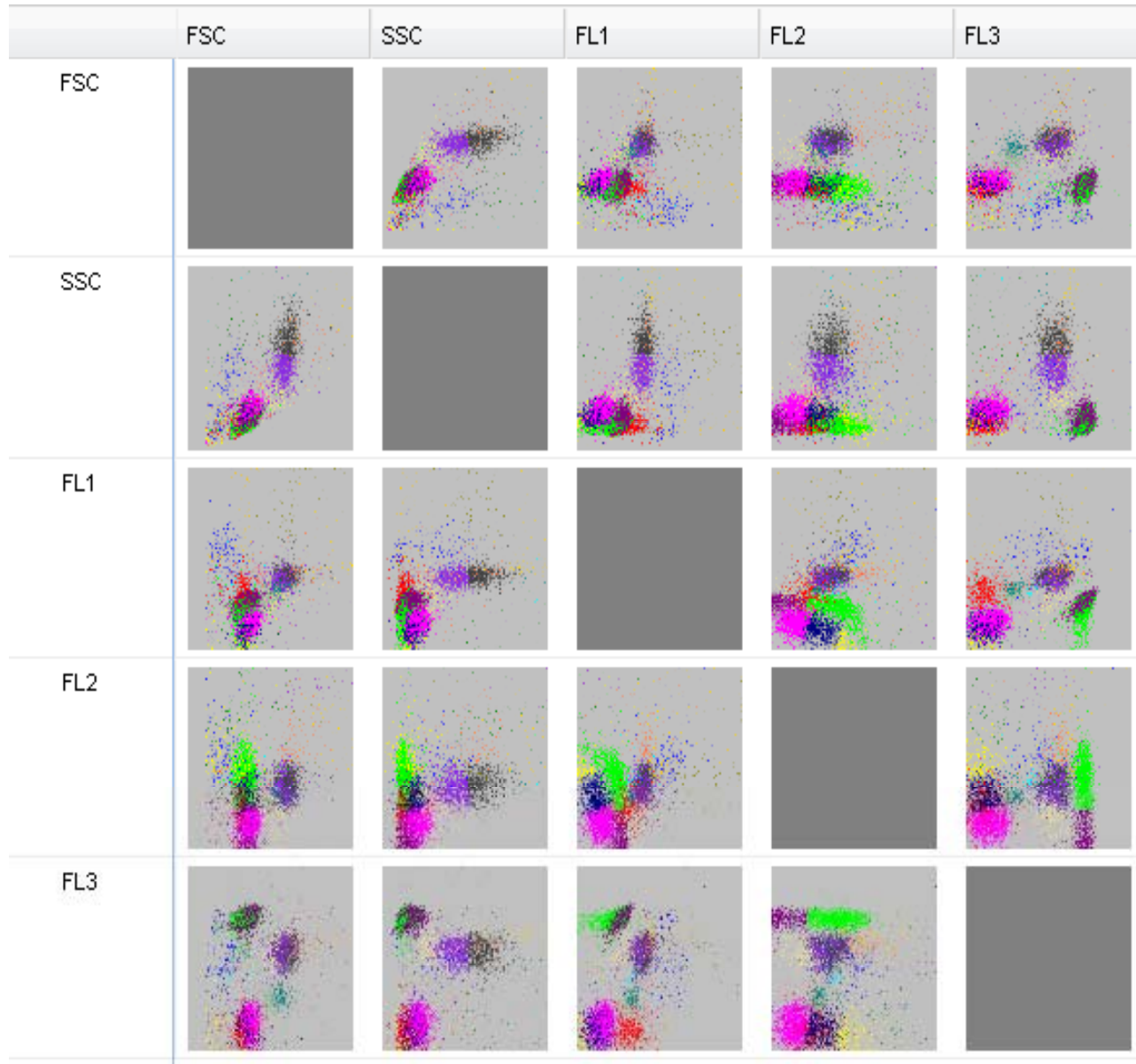


[ImmPort Genetic Analysis Tools \(IGAT\)](#)

[tag SNP](#)

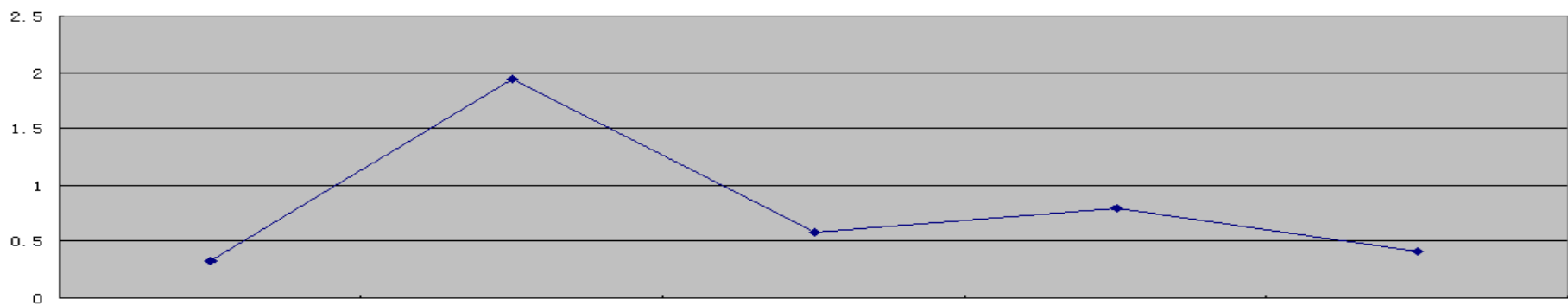
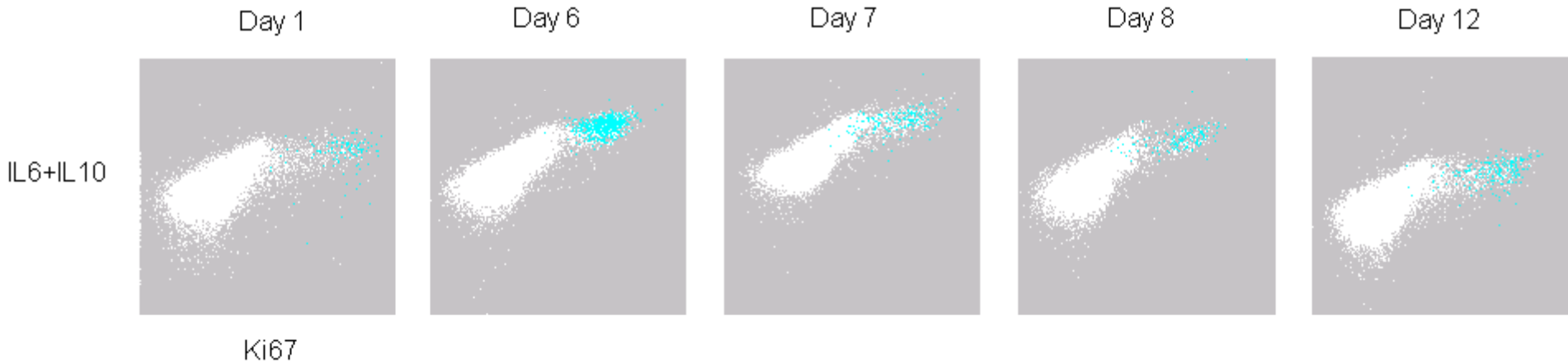
Identify representative SNPs of a genomic region based on linkage disequilibrium characteristics using

Automated Identification of Cell Populations



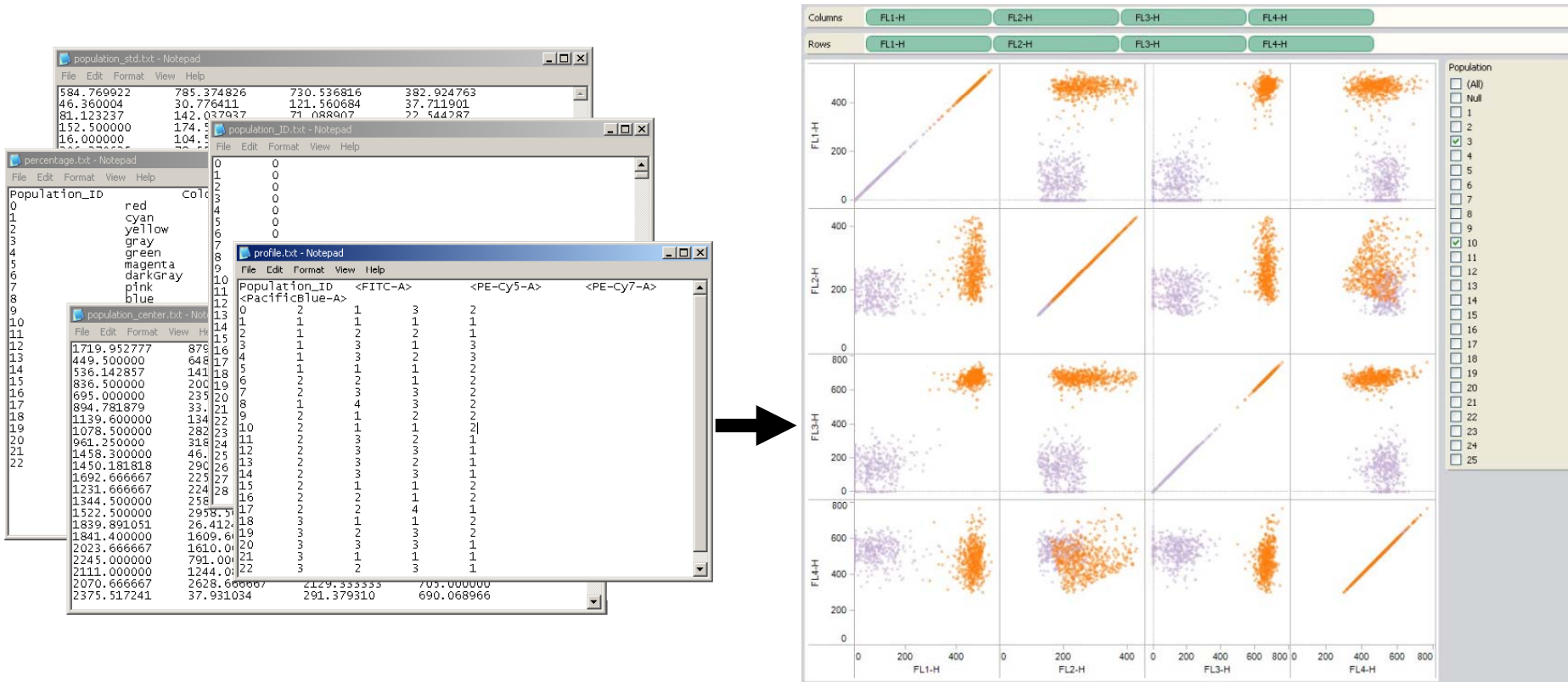
FCM data from Montgomery Lab, Yale Univ.

Cross-Sample Comparison with FLOCK



Proportion change of PlasmaBlasts at different days with Tetanus study

Download FLOCK Results to Your Own Software



Casale FCM data from Immune Tolerance Network
Visualization Software: Tableau

Discussion

- Computational analysis most needed for high-dimensional dataset
- Preprocessing is also important
- FlowCAP2 can include cross-sample comparison, since the alignment and mapping is also challenging
- From cluster to population

Conclusions

FLOw Clustering without K - FLOCK

- Identifies cell populations within multi-dimensional space
- Automatically determines the number of unique populations present using a rapid binning approach
- Can handle non-spherical hyper-shapes
- Maps populations across independent samples
- Calculates useful summary statistics
- Reduces subjective factors in gating
- Implemented in ImmPort and freely available

Acknowledgment

UT Southwestern

Richard Scheuermann

Megan Kong

Paula Guidry

David Dougall

Eva Sadat

Jamie Lee

Jennifer Cai

Jie Huang

Nishanth Marthandan

Diane Xiang

Young Kim

Adam Seegmiller

Nitin Karandikar

Northrop Grumman

John Campbell

Yue Liu

Liz Thompson

Patrick Dunn

Jeff Wiser

Mike Atassi

Immune Tolerance Network

Dave Parrish

Keith Boyce

Tom Casale

Jason Liu

FlowCAP Organization Committee

Rochester

Iñaki Sanz

Chungwen Wei

Eun Hyung Lee

Tim Mosmann

Jessica Halliley

Chris Tipton



Supported by NIH N01 AI40076 (BISC)