SWIFT: High-resolution High-dimensional Analysis of Flow-Cytometry Data

GMM Based lustering

Gaurav Sharma*,
*Dept. of Electrical and Computer Engineering
University of Rochester

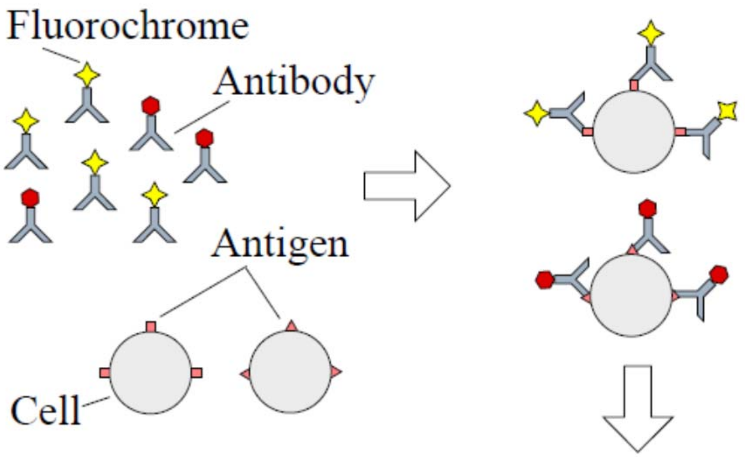
Acknowledgements

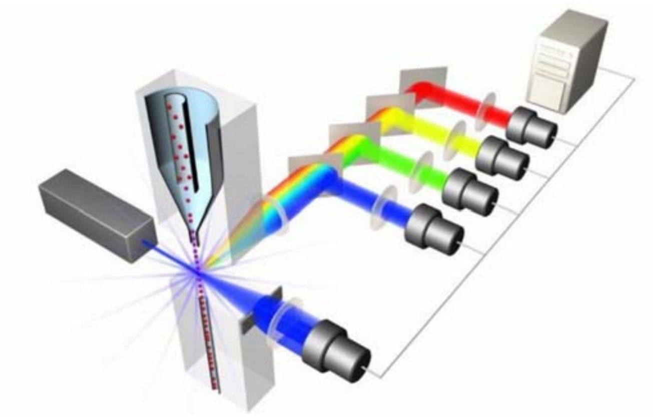
Collaborators:

Tim Mosmann
Iftekhar Naim
Jonathan Rebhahn
Suprakash Datta
James Cavenaugh
Juilee Thakar
Sally Quataert
Alexandra Livingstone
Alex Rosenberg
Jason Weaver

Flow Cytometry (FC) Overview

- Measure multiple properties of individual cells
- Immunology Applications
  - Quantify different antigens in individual cells
  - Immunopheotyping: classification of infectious diseases
  - Quantifying effect of vaccines, treatment, …

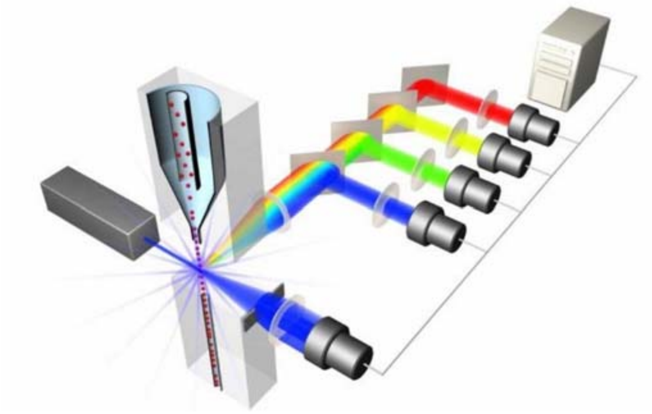




Ref: http://flow.csc.mrc.ac.uk

Complexity of Flow Cytometry data

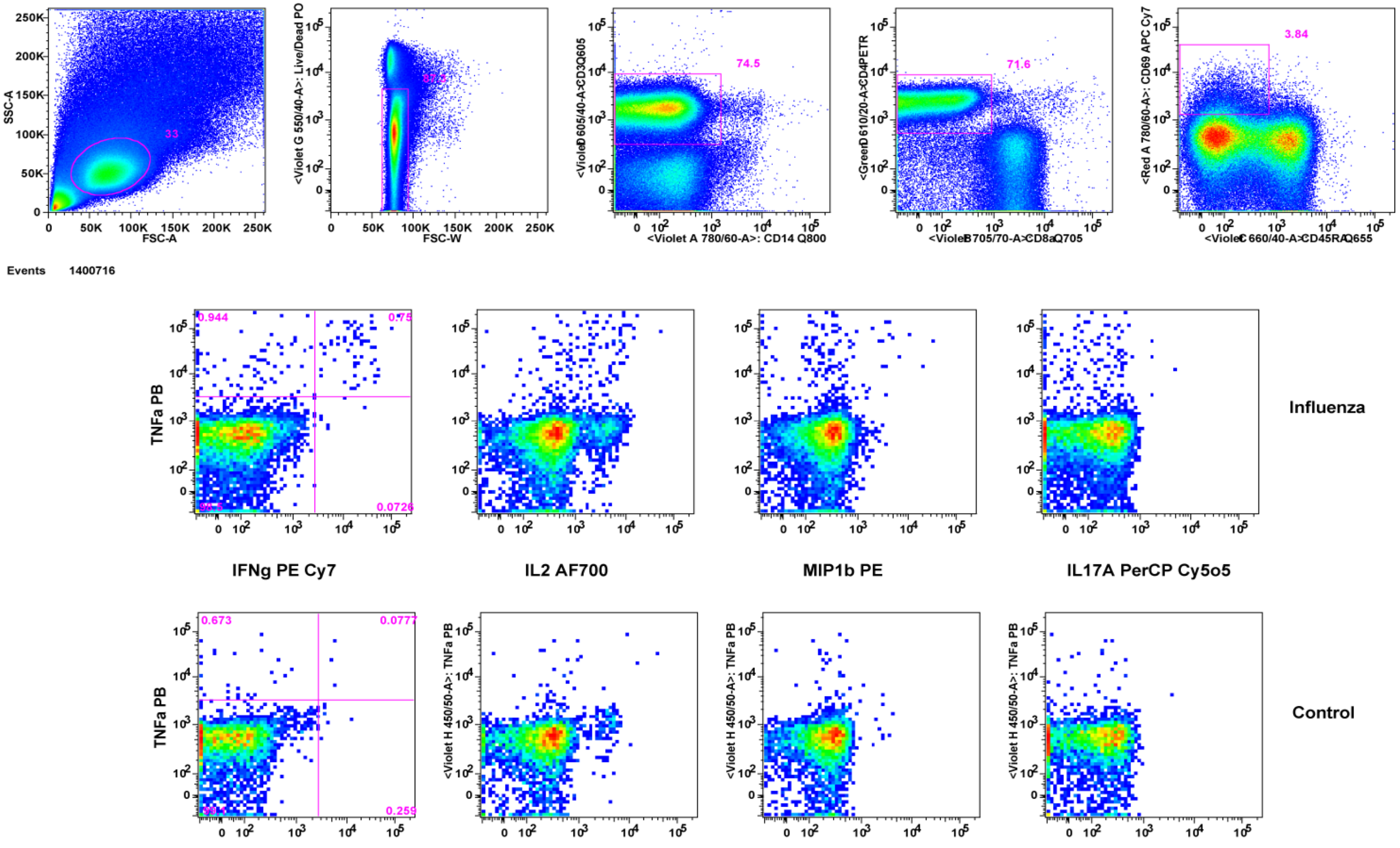
- **High dimensional data**
  - **Conventional Fluorescence flow cytometry: 16-18 colors**
    - **Pratip Chattopadhyay, 27 colors**
  - **CyTOF Mass Spectrometry flow cytometry: >33 parameters**
  - **Spectral fluorescence cytometry, >20 parameters**
  - **ChipCytometer, 40 parameters**
- **Large datasets: ~ 1 million cells/sample, pooled samples common (~ 20 million cells)**



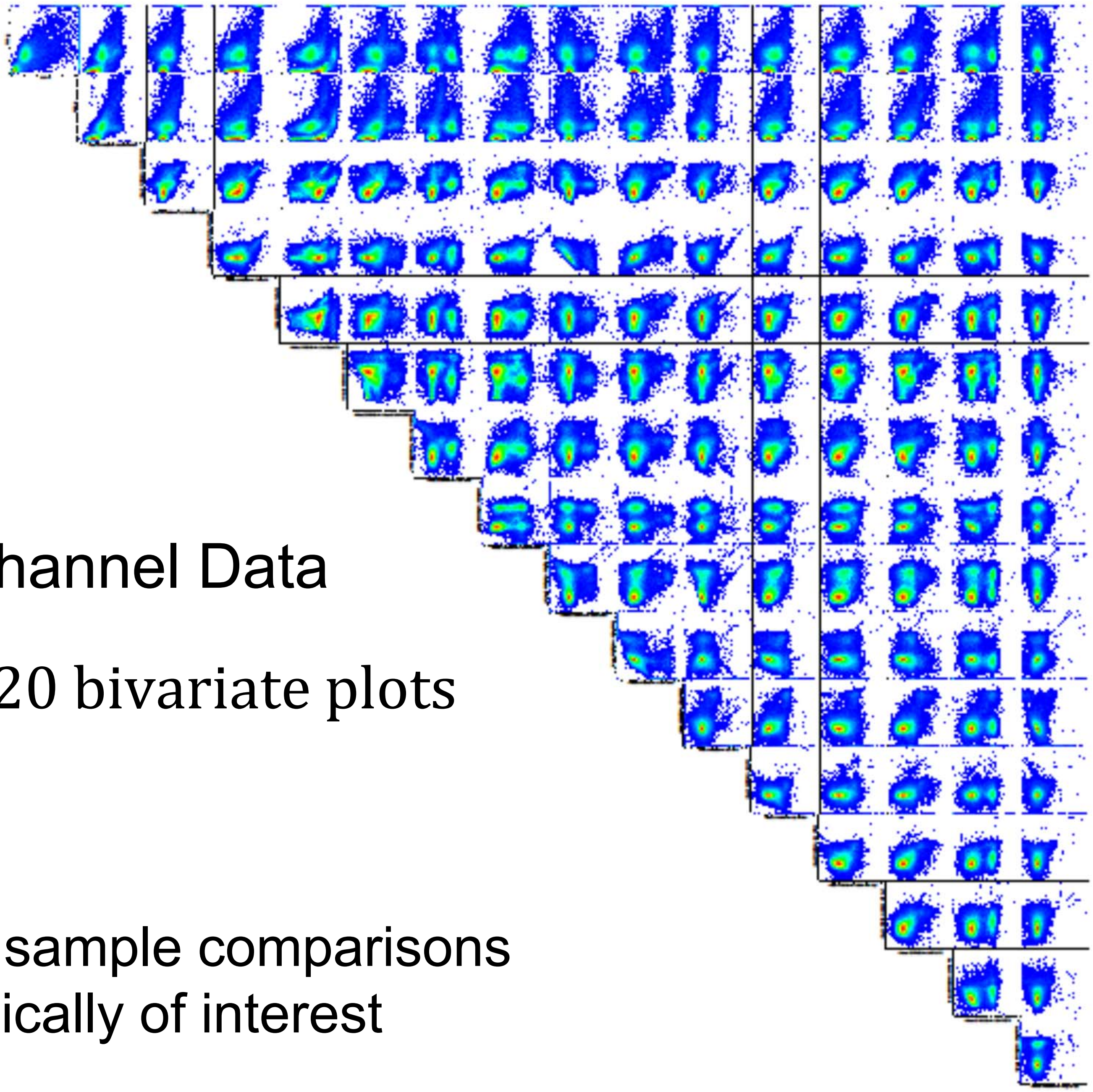
Ref: http://flow.csc.mrc.ac.uk

Flow Cytometry Data Analysis via Manual Gating

- Bi-variate gating (selection): 2 axes at a time
- Example:



Manual Gating Analyses is limited



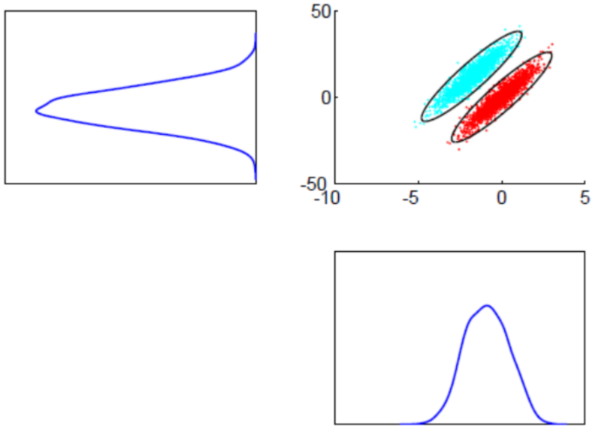
16 Channel Data

$$\binom{16}{2} = 120 \text{ bivariate plots}$$

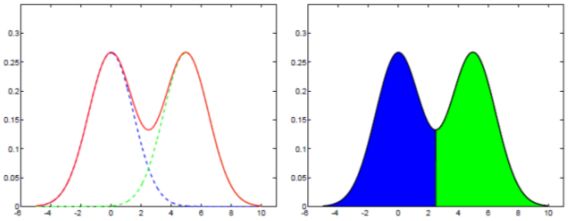
Multiple sample comparisons typically of interest

Limitations of Manual Gating

- **Inferences are based on partial view of data**
  - **Only subset of bi-variate option is explored**
- **Focused on identifying presence of absence of specific cell sub-populations**
  - **Primarily hypothesis testing rather than discovery**
- **Partial views can mask subpopulations**



- **Gating does not comprehend overlaps**
- **Subjective and limited repeatability**



Overlapping sub-populations

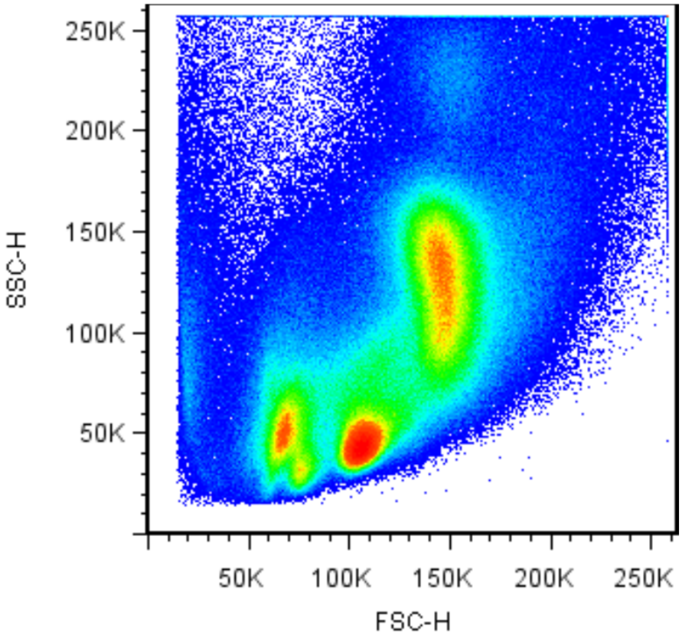
Automated FC Data Analysis

- SWIFT Processing pipeline
  - Clustering
    - Identifying homogenous subpopulations of cells
  - Templating
    - Facilitating efficient cross sample comparison for inference
  - Competition
    - Resolving subpopulation shifts

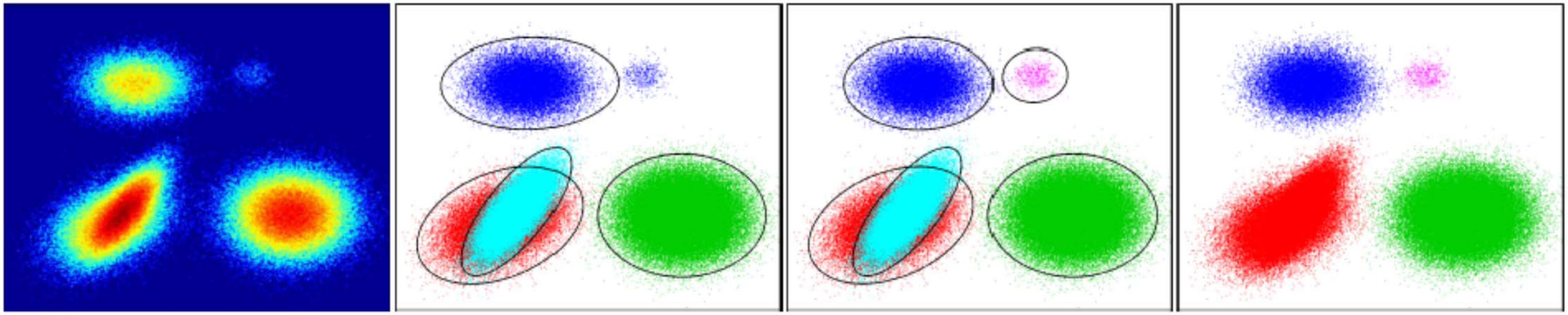
I. Naim, S. Datta, J. Rebhahn, J. S. Cavenaugh, T. R. Mosmann, and G. Sharma, "SWIFT - scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets: Part 1 - Algorithm design," Cytometry, Part A, vol. 85, no. 5, pp. 408-421, May 2014.
T. R. Mosmann, I. Naim, J. Rebhahn, S. Datta, J. S. Cavenaugh, J. M. Weaver, and G. Sharma, "SWIFT - scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets: Part 2 - Biological evaluation," Cytometry, Part A, vol. 85, no. 5, pp. 422-433, May 2014.

Challenges for Automated Clustering

- Large size and high dimensionality of datasets
  - Efficiency is important
- Small subpopulations are often important
  - Antigen specific T-cells (<100 out of 10^6)
- Subpopulation distributions are skewed
  - Non-ellipsoidal
- Overlapping subpopulations and background noise



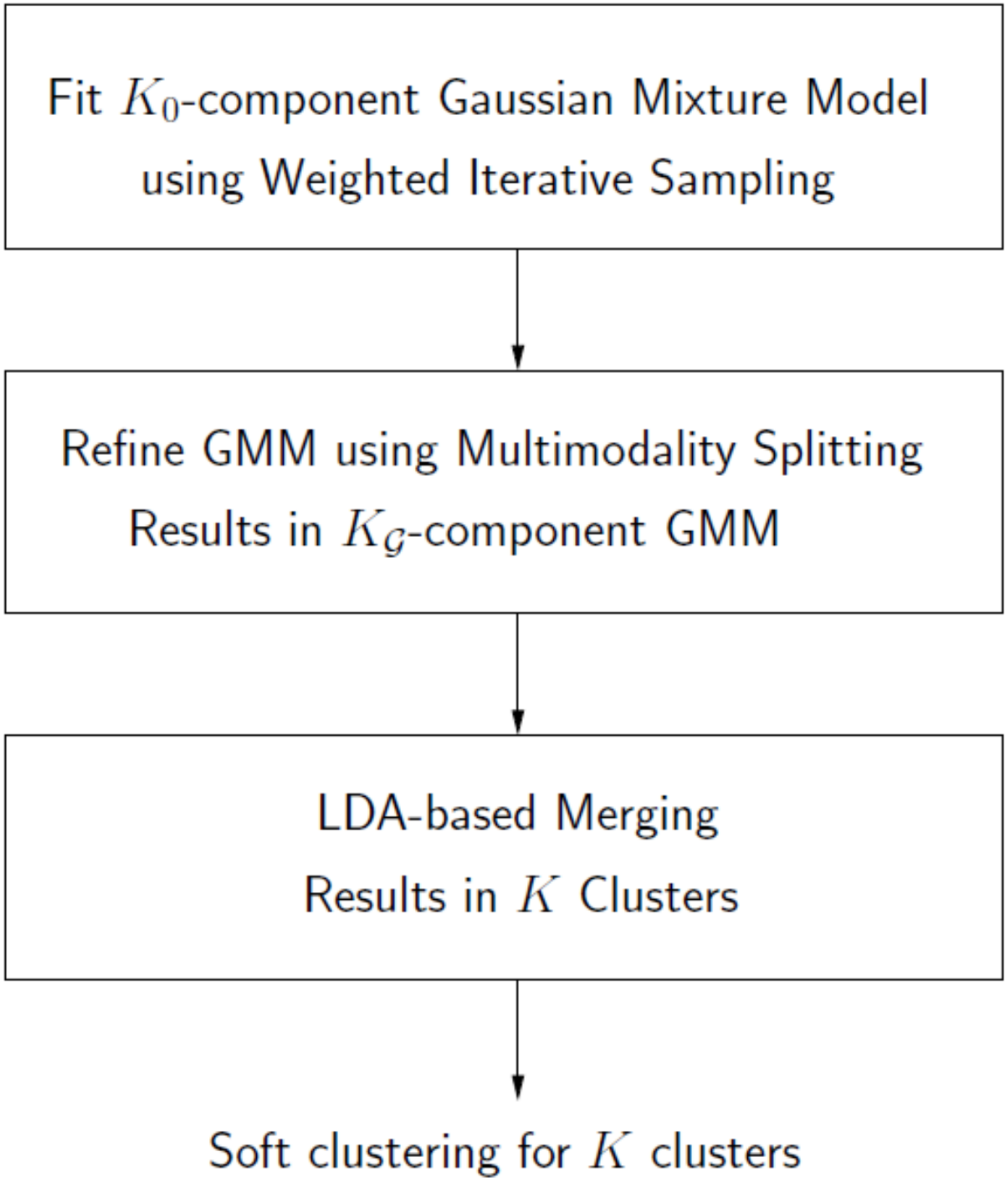
SWIFT Clustering: 3 Stage Framework



(a) Initial Population (b) Initial GMM (c) Splitting (d) LDA Merging

- Weighted Iterative Sampling based EM : Gaussian mixture model clustering + novel weighted iterative sampling
  - Scalability to large datasets (~ 20 million cells, 20 dimensions)
- Multimodality Splitting: Refines initial multimodal clusters
  - Identification of rare populations
- LDA-based Merging: Merge overlapping Gaussians using Linear Discriminant Analysis (LDA)

SWIFT Clustering: 3 Stage Framework



Fit K_0-component Gaussian Mixture Model using Weighted Iterative Sampling

Refine GMM using Multimodality Splitting
Results in $K_{\mathcal{G}}$-component GMM

LDA-based Merging
Results in K Clusters

Soft clustering for K clusters

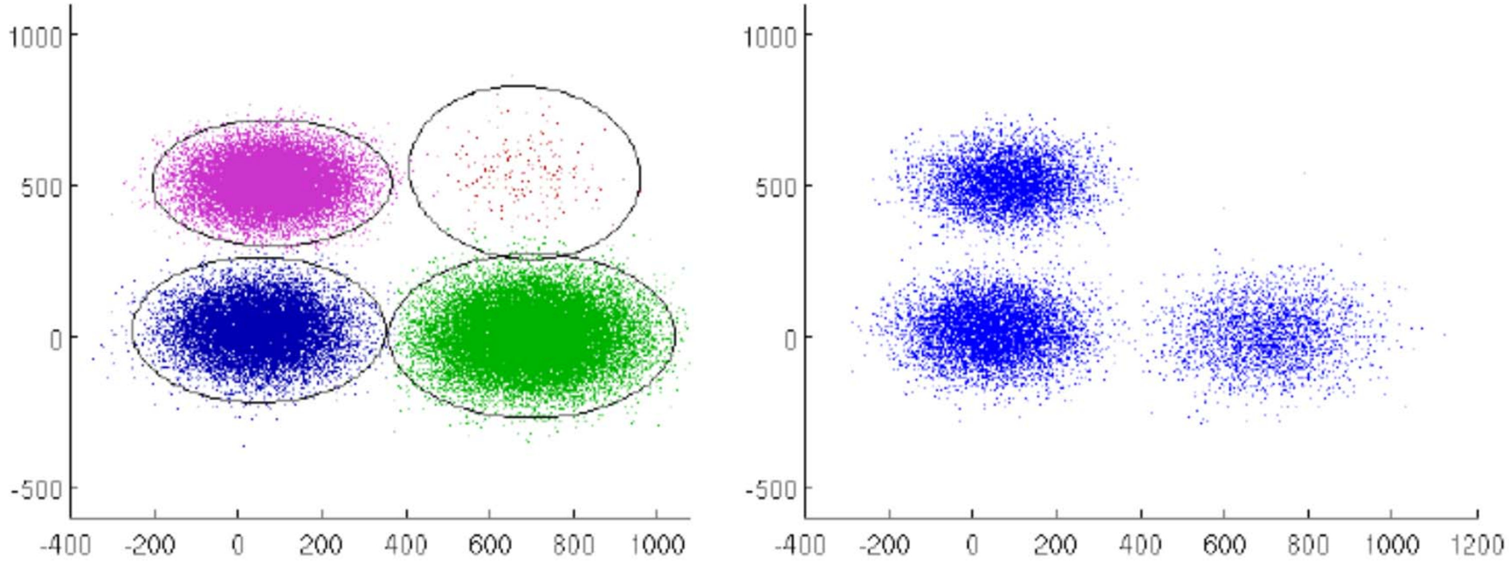
SWIFT Clustering: Stage 1 GMM

- Gaussian mixture model (GMM) clustering is chosen among the model based methods
  - ➢ Faster than other model based clustering methods
- Expectation Maximization (EM) algorithm for parameter Estimation
- Computational complexity of each iteration: $O(NK_0d^2)$
  - ➢ N = the number of data-vectors in the dataset ($\sim 10^6$)
  - ➢ K_0= is the number of Gaussian components ($\sim 10^2$)
  - ➢ d = is the dimension of each data-vectors (~ 20)

SWIFT Clustering: Sampling for Scalability

- Idea: Operate on smaller subsample of dataset for better computational performance
- Challenge: Poor representation of smaller subpopulations

➢ Overcome via weighted iterative sampling

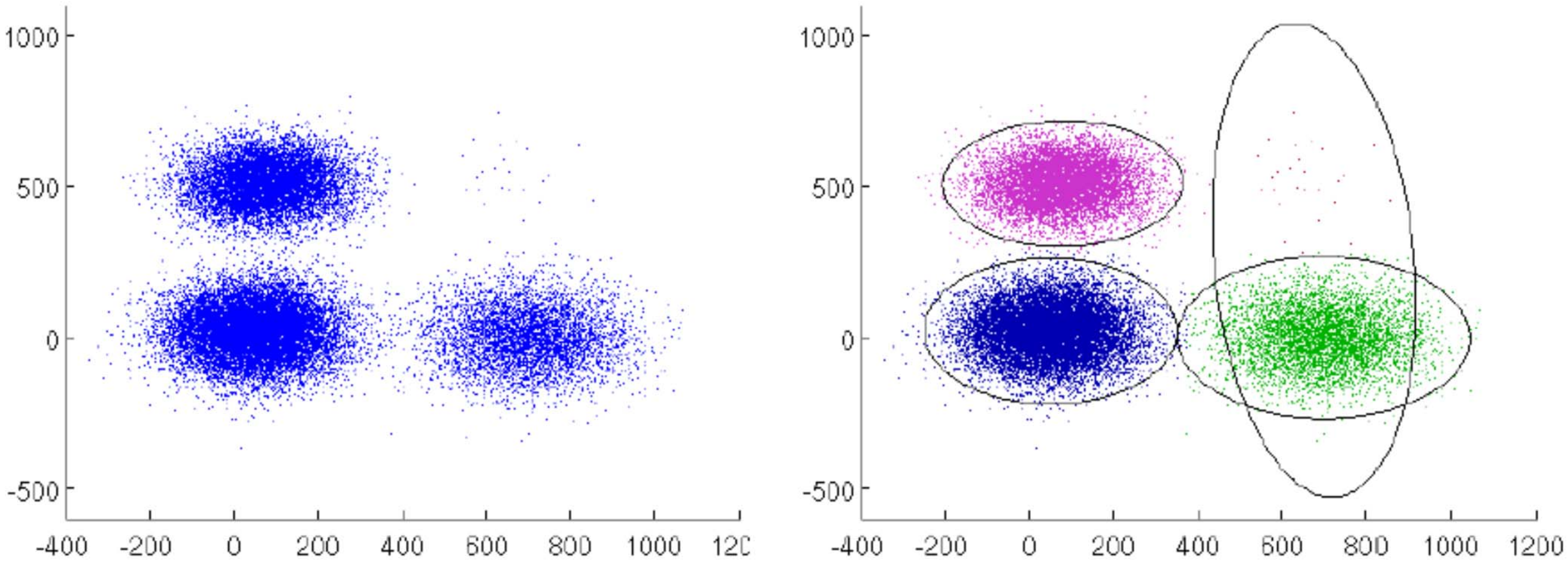


(g) 4 Gaussians with 150K, 100K, 50K and 150 datapoints

(h) After 10% sampling

SWIFT Clustering: Weighted Iterative Sampling

- First sample: Random

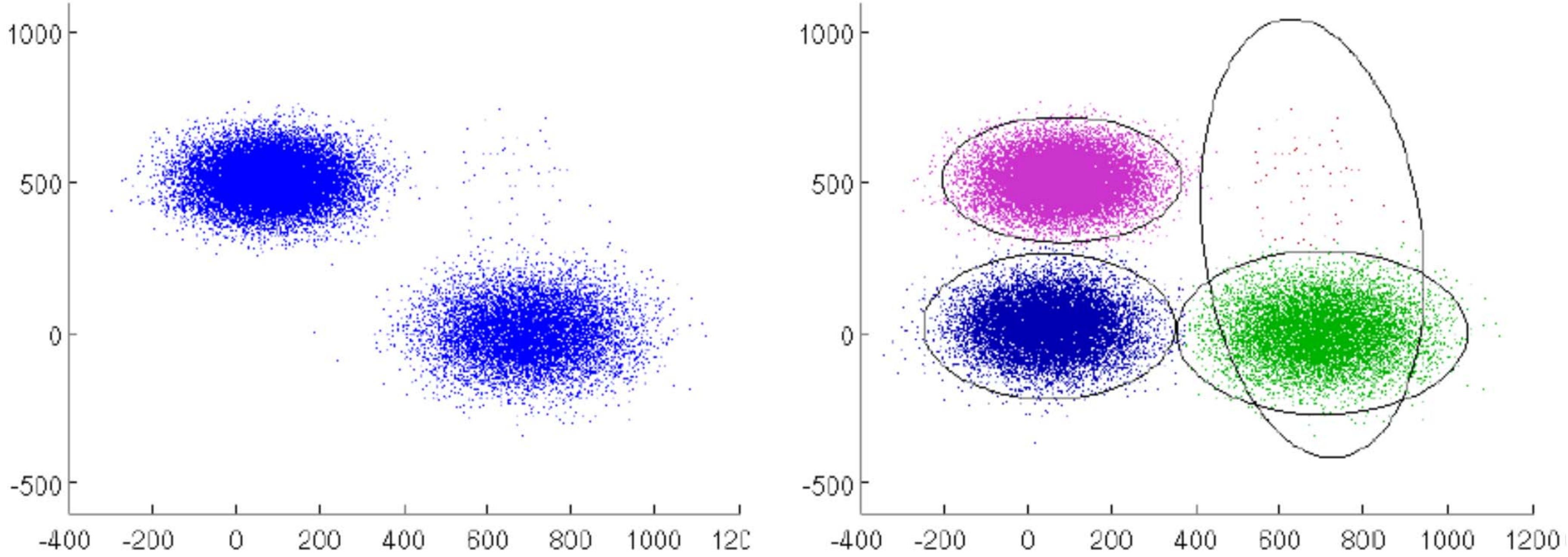


(a) First sample

(b) Clustering first sample

SWIFT Clustering: Weighted Iterative Sampling

- Second sample: Sampling Probability $(1 - \sum_{l \in \{1\}} \gamma_{il})$

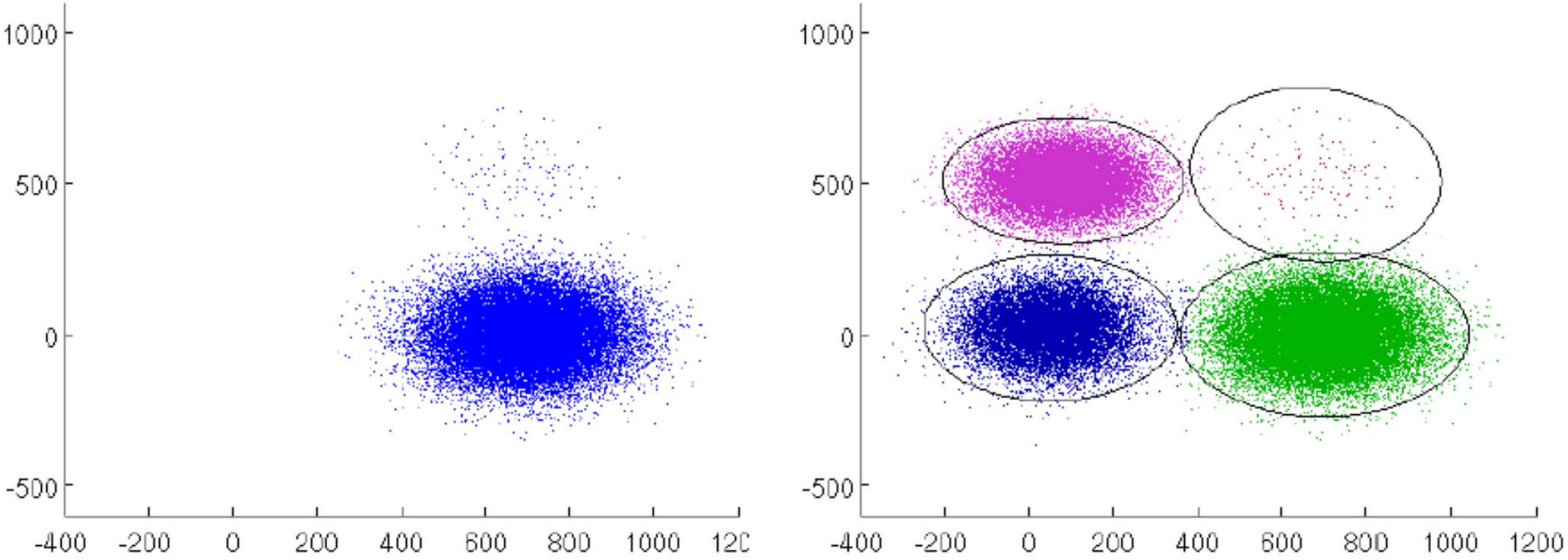


(c) Second sample

(d) Clustering second sample

SWIFT Clustering: Weighted Iterative Sampling

- Third sample: Sampling Probability $(1-\sum_{l\in\{1,2\}}\gamma_{il})$

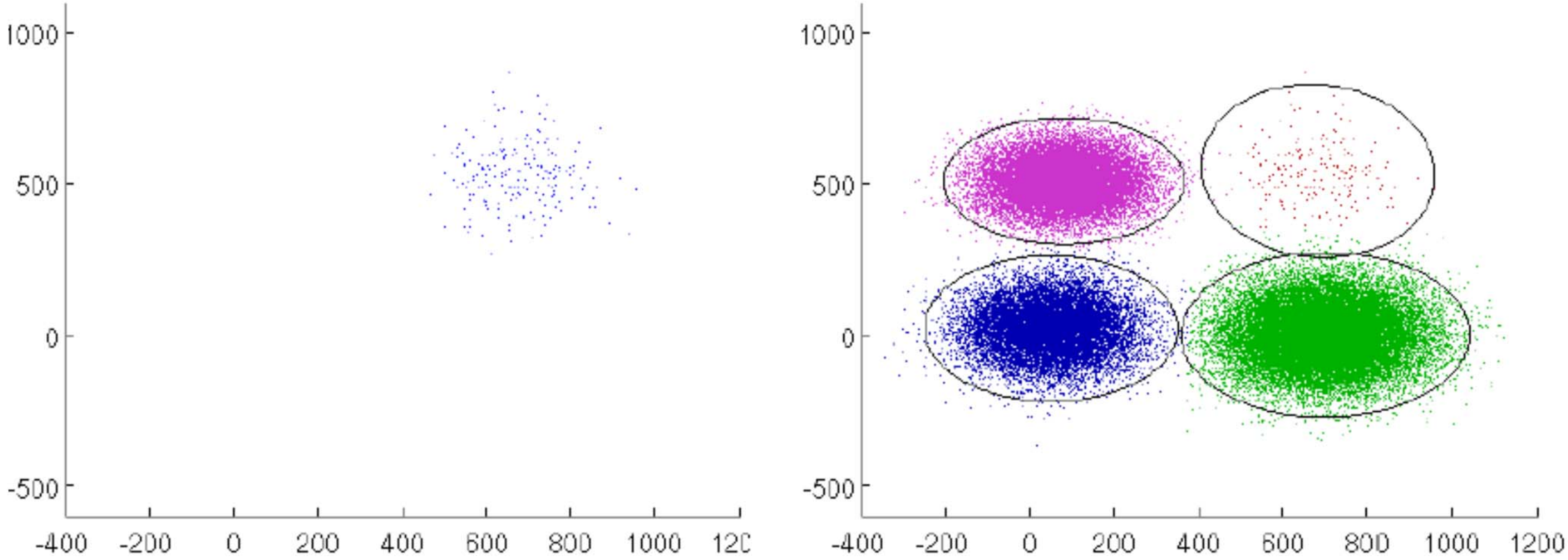


(e) Third sample

(f) Clustering third sample

SWIFT Clustering: Weighted Iterative Sampling

- Final sample: Sampling Probability $(1 - \sum_{l \in \{1,2,3\}} \gamma_{il})$



(g) Last sample

(h) Final clustering

Weighted Iterative Sampling Advantages

- Improves resolution of small subpopulations
  - ➢ Increased weights for small clusters while resampling
  - ➢ Traditional EM shows poor convergence in the presence of high dynamic range in mixing coefficients
- Scalability in both memory and computation time
  - ➢ Complexity of each EM iteration reduced from $O(NK_0d^2)$ to $O(nK_0d^2)$
  - ➢ n = Sample size
  - ➢ Simulation results show 18-fold speed up $N = 2\times10^6$, $n = 2\times10^4$
- Extensible to other soft clustering methods
  - ➢ Mixture of t, skewed t distributions, or fuzzy clustering

Weighted Iterative Sampling

- Mathematical Analysis:

A. The Weighted Iterative Sampling preserves the stationary points of likelihood function, under two assumptions:

1 Parameters of the large fixed clusters converged to true values

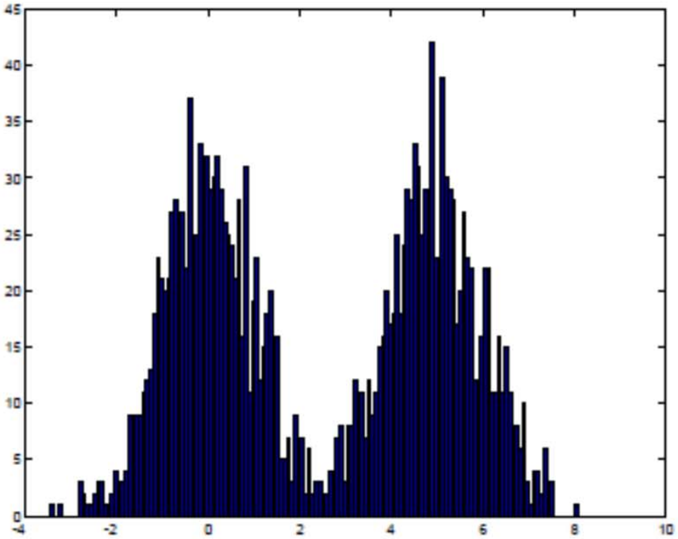
2 The estimated membership probabilities for the large fixed clusters are accurate

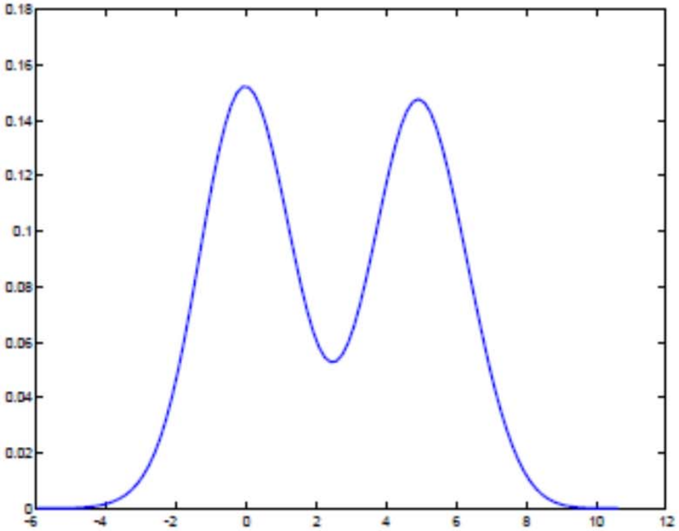
B. Condition number of Hessian at true parameters worse under high dynamic range

- Additional considerations required in practice

SWIFT Clustering: Stage 2 Multimodality Splitting

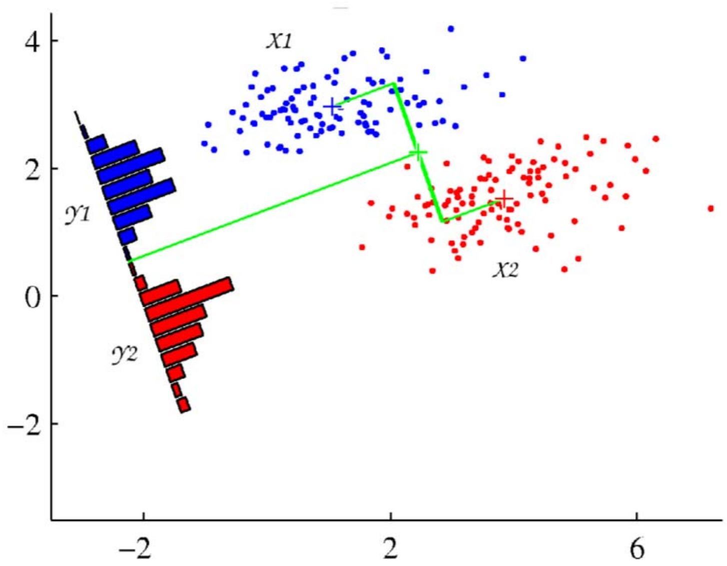
- Split clusters that are multimodal
  - Multimodality Detection: 1-D Kernel Density Estimation
  - Any data dimension or PCA dimension
- Unimodality is typically biologically significant
- Outcome: significant improvement in resolution of small clusters



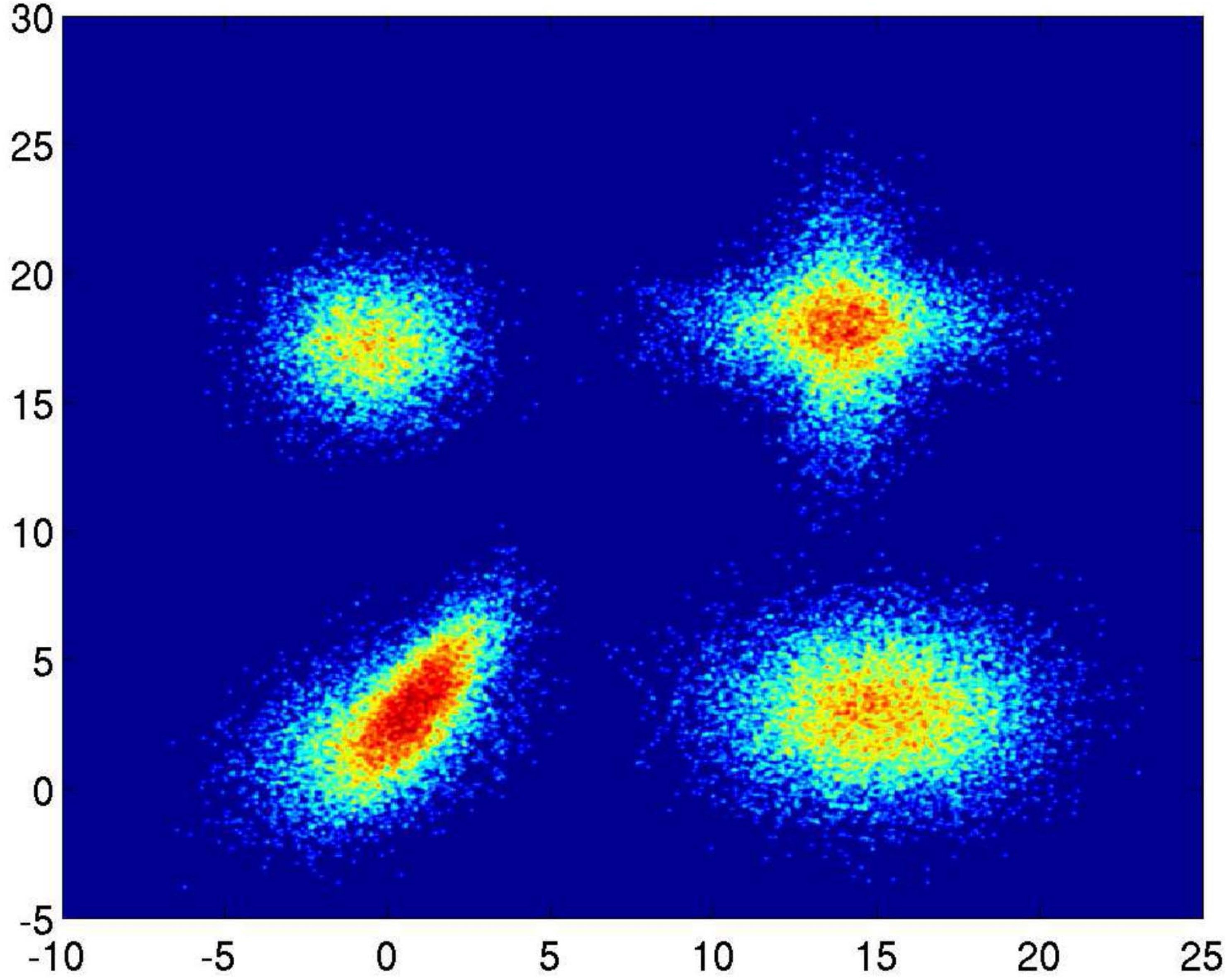


SWIFT Clustering: Stage 3 Agglomerative Merging

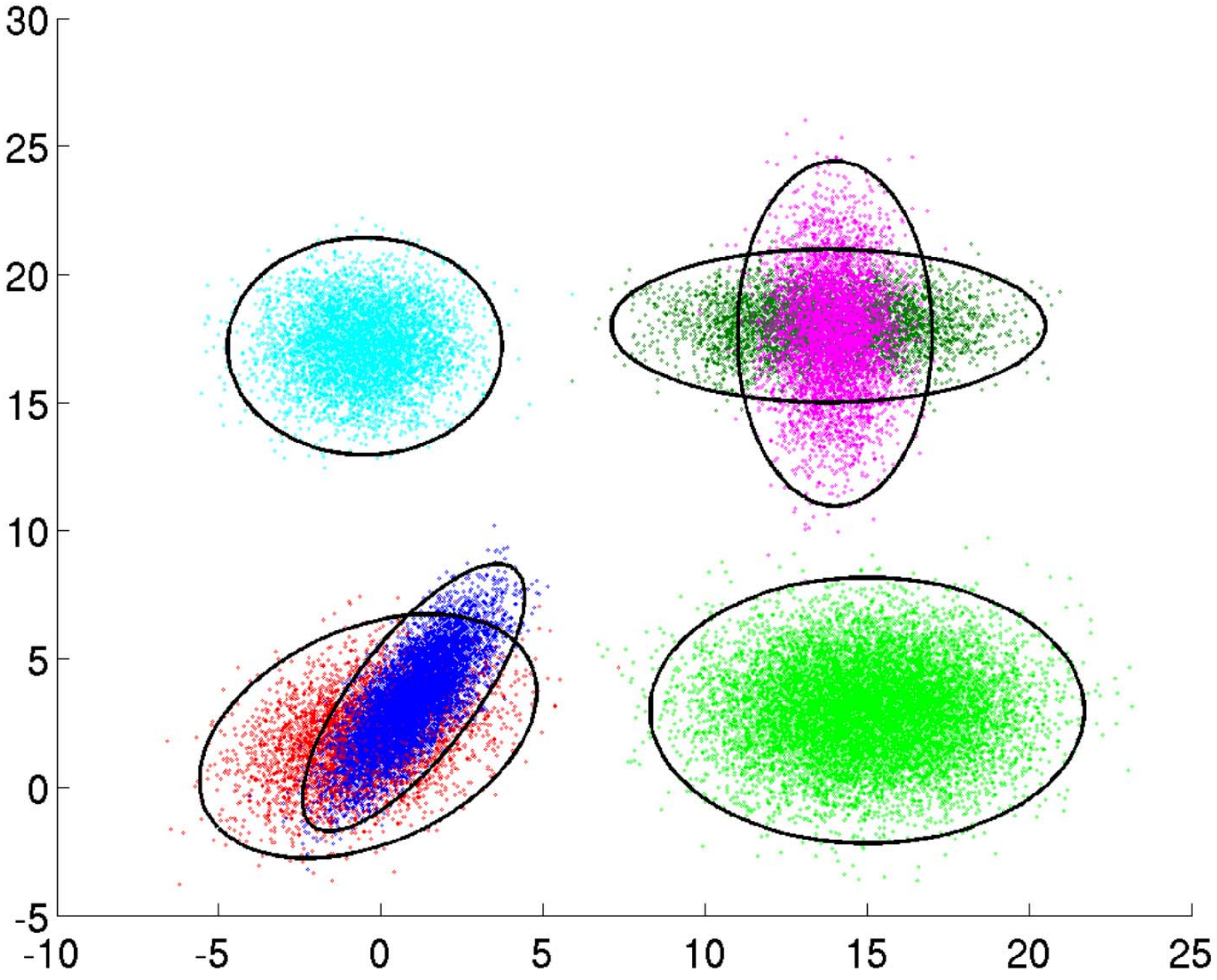
- Gaussian Mixture Model: Each cluster follows a multivariate Gaussian distribution
- ➢ Symmetric, ellipsoidal clusters
- FC datasets have skewed clusters
- ➢ Not well-explained by a single Gaussian
- Merge pairs of clusters while honoring unimodality
- ➢ Examine modality along LDA dimension



Example: Four Cluster Dataset (two skewed)

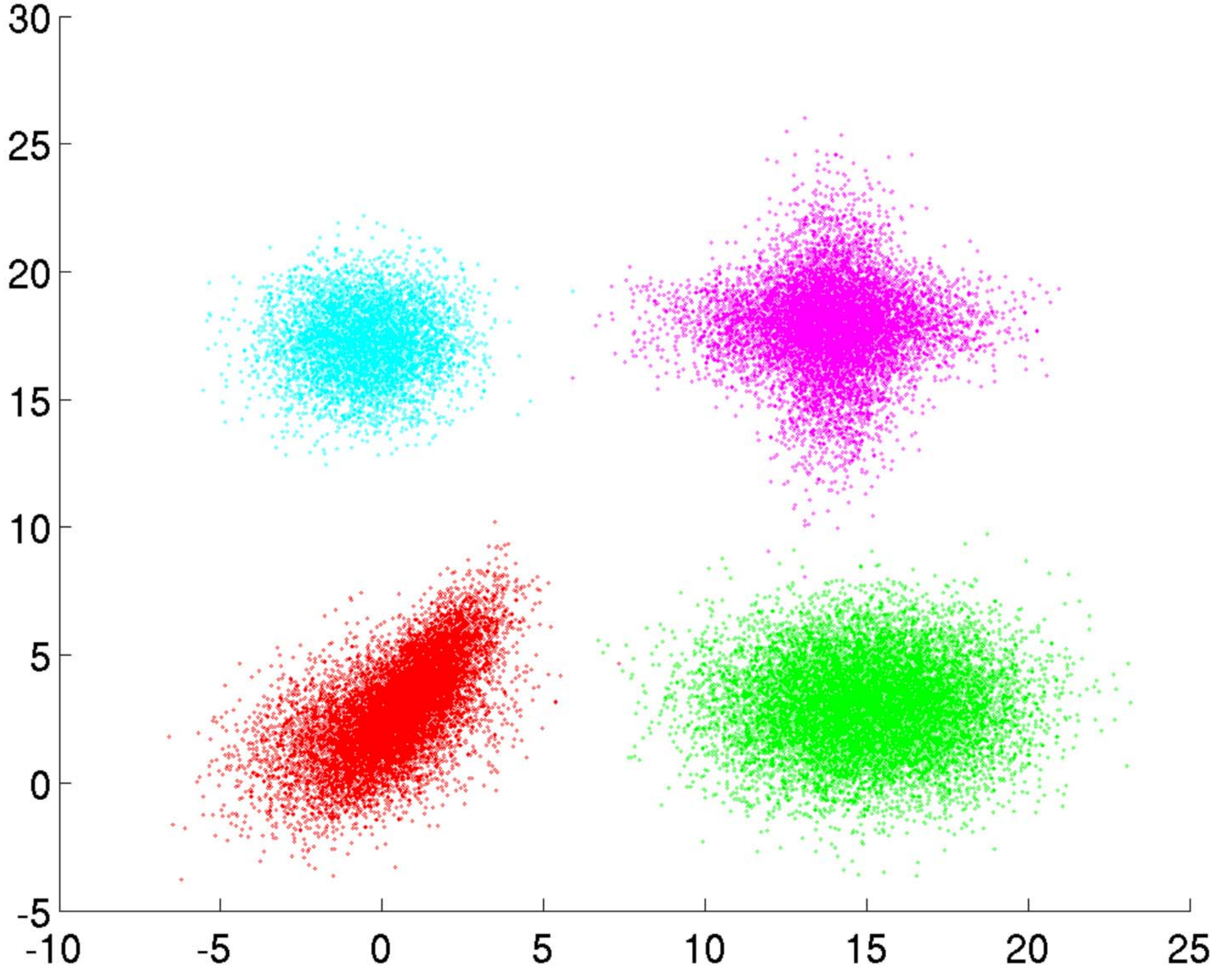


Example: Four Cluster Dataset (two skewed)



Initial GMM Fit

Example: Four Cluster Dataset (two skewed)



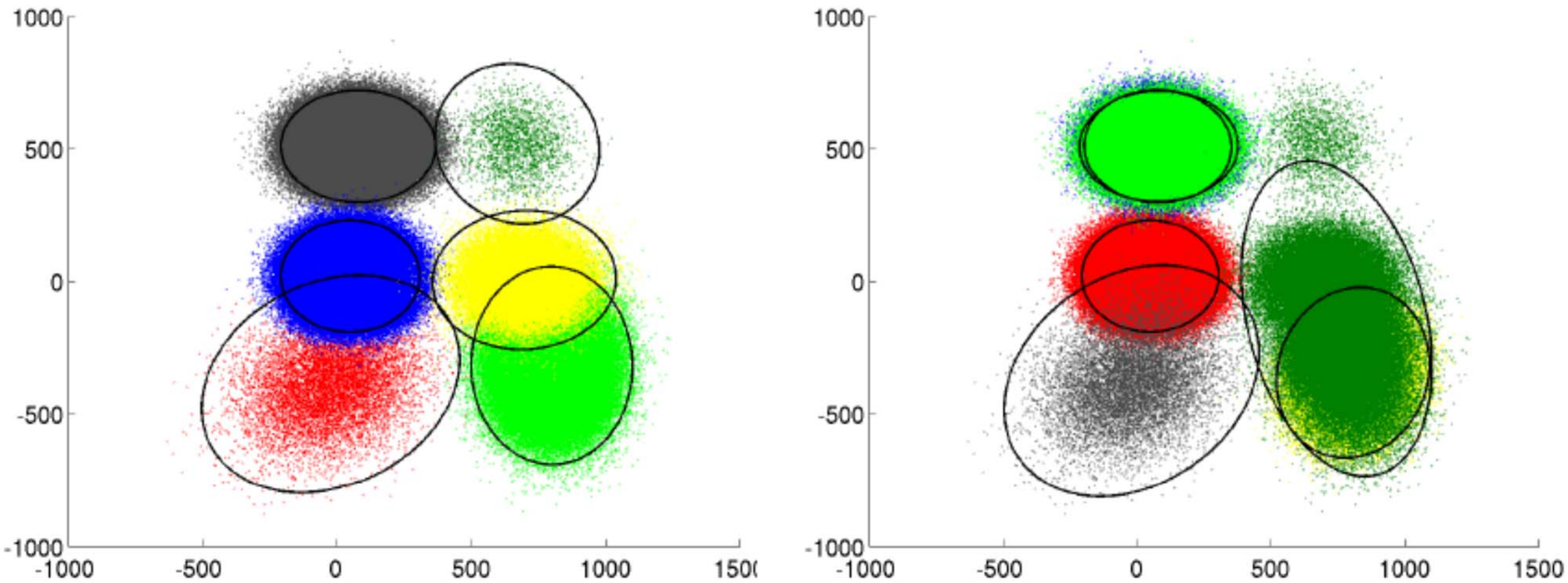
Merged Clusters

Results on Synthetic Data

- Validation of the clustering methods for FC data is challenging
  - ➢ Ground truth datasets are rarely available
  - ➢ Visual validation is difficult for high dimensional data clustering

- Initial validation on synthetic data
  - ➢ Synthetic mixture of 6 (overlapping) bivariate Gaussians
  - ➢ Total Size: 2.002 million events
  - ➢ High dynamic range in population sizes
  - ➢ Largest cluster: 1×10^6 events
  - ➢ Smallest cluster: 2×10^3 events

Results on Synthetic Data: Typical Result



(a) Weighted Iterative Sampling-based EM (b) Traditional EM (on full dataset)

- EM shows poor convergence in the presence of high dynamic range in mixing coefficients

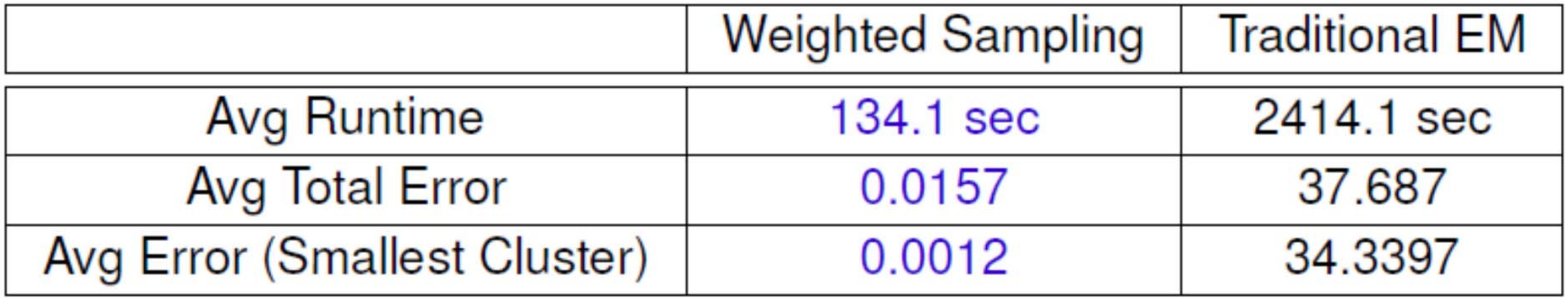
➢ Can be explained this using the Hessian-based convergence analysis of EM (Xu and Jordan [1996])

Results on Synthetic Data: Metrics

- Error measured using symmetric Kullback-Leibler divergence from true parameter values
- EM converges to local optima

➢ Convergence time and accuracy are sensitive to initialization

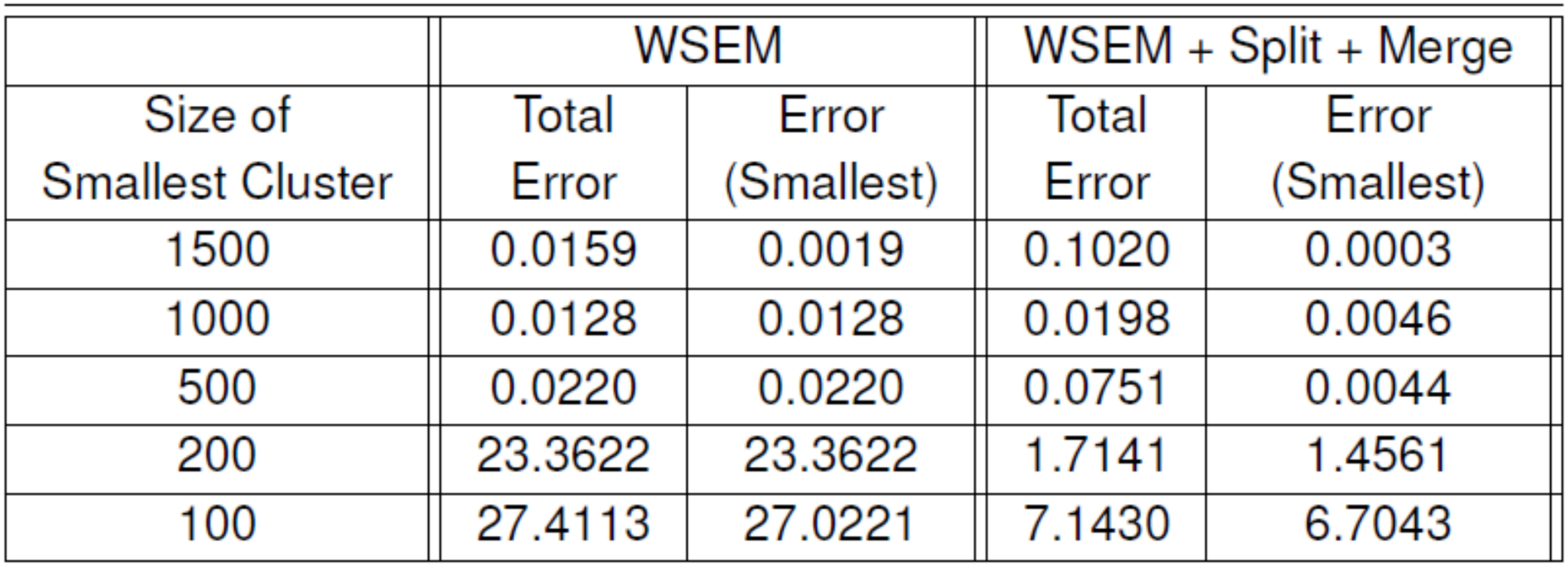
➢ Compare accuracy/runtime averaged over 10 independent runs (each with 10 repetitions to avoid local optima)



	Weighted Sampling	Traditional EM
Avg Runtime	134.1 sec	2414.1 sec
Avg Total Error	0.0157	37.687
Avg Error (Smallest Cluster)	0.0012	34.3397

Results on Synthetic Data: Small Cluster Resolution

- Smallest cluster size decreasing with other populations fixed with a size of 2 million cells



	WSEM		WSEM + Split + Merge	
Size of Smallest Cluster	Total Error	Error (Smallest)	Total Error	Error (Smallest)
1500	0.0159	0.0019	0.1020	0.0003
1000	0.0128	0.0128	0.0198	0.0046
500	0.0220	0.0220	0.0751	0.0044
200	23.3622	23.3622	1.7141	1.4561
100	27.4113	27.0221	7.1430	6.7043

SWIFT Clustering Summary

- Scalable algorithm for FC data clustering

➢ Weighted Sampling based EM + Multimodality Splitting + LDA-based Merging

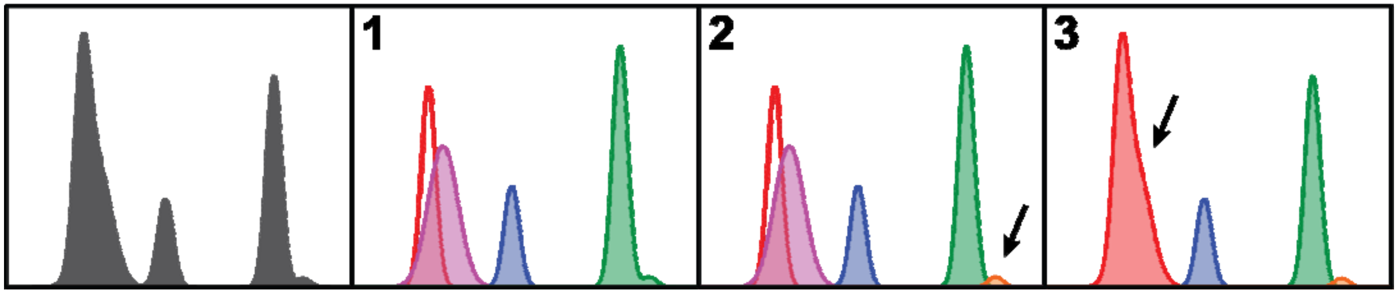
➢ Scales to large datasets (> 15 million cells, 20 dimensions)

- Integrated in a problem-aware manner

✓ Modality aware representation of overlapping populations

✓ Ability to resolve small subpopulations (< 100 cells out of 10 millions)

✓ Semantics of data representation preserved unlike dimensionality reduction methods



Acknowledgements

Flow Cytometry analysis:
Iftekhar Naim
Jonathan Rebhahn
Gaurav Sharma
Suprakash Datta
Alex Rosenberg
Jason Weaver
James Cavenaugh
Sally Quataert
Alexandra Livingstone
Juilee Thakar

LAVA Cell/population Simulations
Jonathan Rebhahn
Sui Huang
Gaurav Sharma

Influenza responses:
Jason Weaver
David Roumanes
Xi Li
Yilin Qi
EunHyung Lee
Martin Zand
Hulin Wu
John Treanor

FlowCAP Consortium:
Ryan Brinkman
Tim Mosmann
Nima Aghaeepour
Raphael Gottardo
Richard H. Scheuermann
Greg Finak

Rochester Human Immunology Center
CEIRS/New York Influenza Center of Excellence
Center for Biodefense Immune Modeling