Novel Differential Flow Cytometry Data Analyses Method Using Data Nuggets Compression and Projection Pursuit Algorithms

Davit Sargsyan

Nonclinical Statistics Conference 2024

Wiesbaden, Germany

September 27 (Day 3, 9:10-9:30AM)

Johnson&Johnson
Innovative Medicine

Our Team

Prof. Javier Cabrera's Lab:

- Javier Cabrera
- Yajie Duan
- Mahan Dastgiri
- Ge Cheng
- Chun-Pang Lin

Prof Ah-Ng Kong's Lab:

- Ah-Ng Kong
- Rebecca Mary Peters
- PoChung Chou

J&J Colleagues:

- Marguerite Emrich
- Jocelyn Sendecki
- Helena Geys
- Kanaka Tatikola

14 Stones of Ryoan-ji

- A 15^{th} century Japanese temple of Ryoan-ji featuring a meditation stone garden

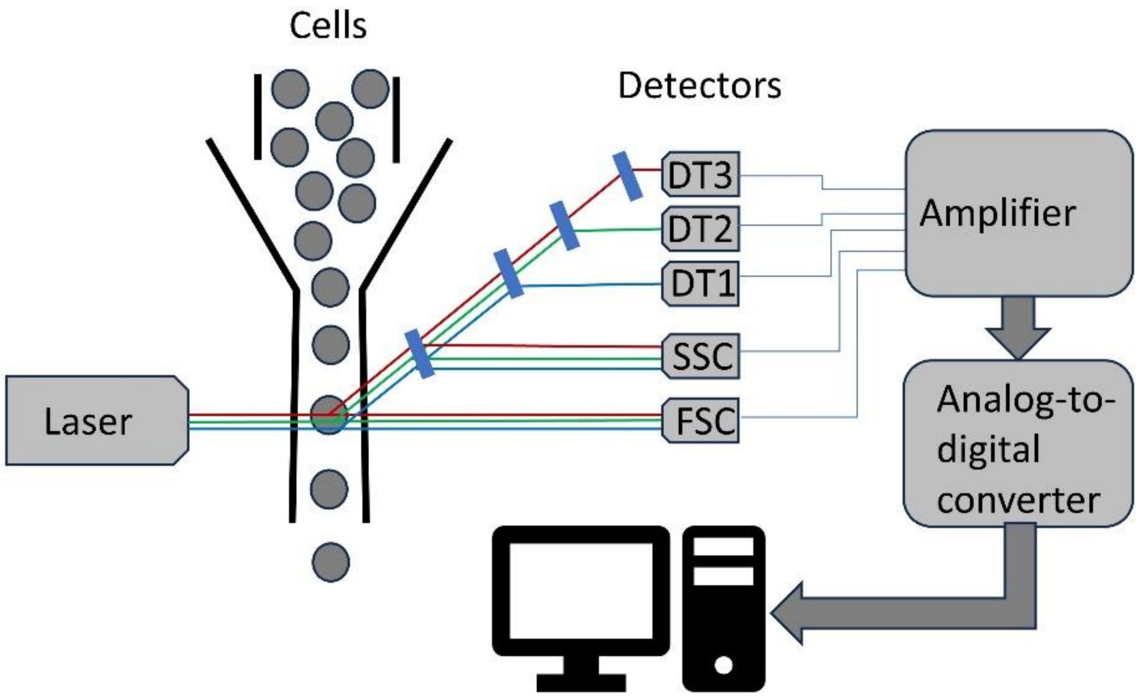
- Veranda that wraps around the garden opens to a view of 14 large stones

15 (!) Stones of Ryoan-ji

- A 15th century Japanese temple of Ryoan-ji featuring a meditation stone garden
- Veranda that wraps around the garden opens to a view of 14 large stones
- Moving from one sitting spot to another, a careful observer will soon realize that the number of the stones is, in fact 15
- But at no point will all 15 stones be revealed to the observer at ones!

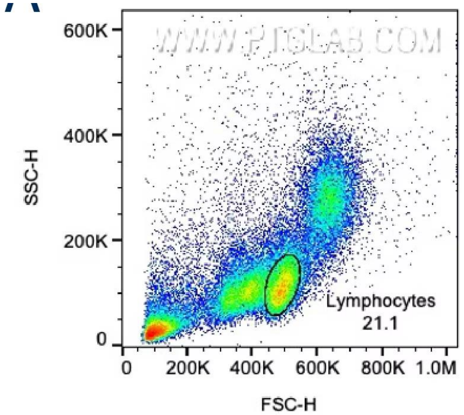
Study Design

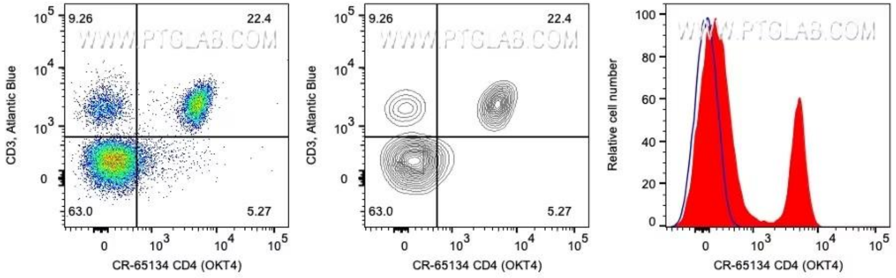
- Blood samples from 44 infants: 24 Unexposed (UE) and 20 Exposed to HIV (HEU), i.e., mothers diagnosed with HIV
- Samples either untreated (Unstimulated) or treated with one of six compounds. We examined Untreated and LPS (lipopolysaccharide) treated samples only.
- Each sample analyzed on a flow cytometer. Data published as .FSC files on a public repository.
- Combined data had >42M rows (cells or particles). Out of these, ~14M identified as lymphocytes.
- Delta[*i, j,k*] = LPS[*i, j,k*] – mean(Unstimulated)[*, *j,k*] for *j*-th subject and *k*-th marker (protein)
- Remaining ~6.9M lymphocytes (=LPS group)



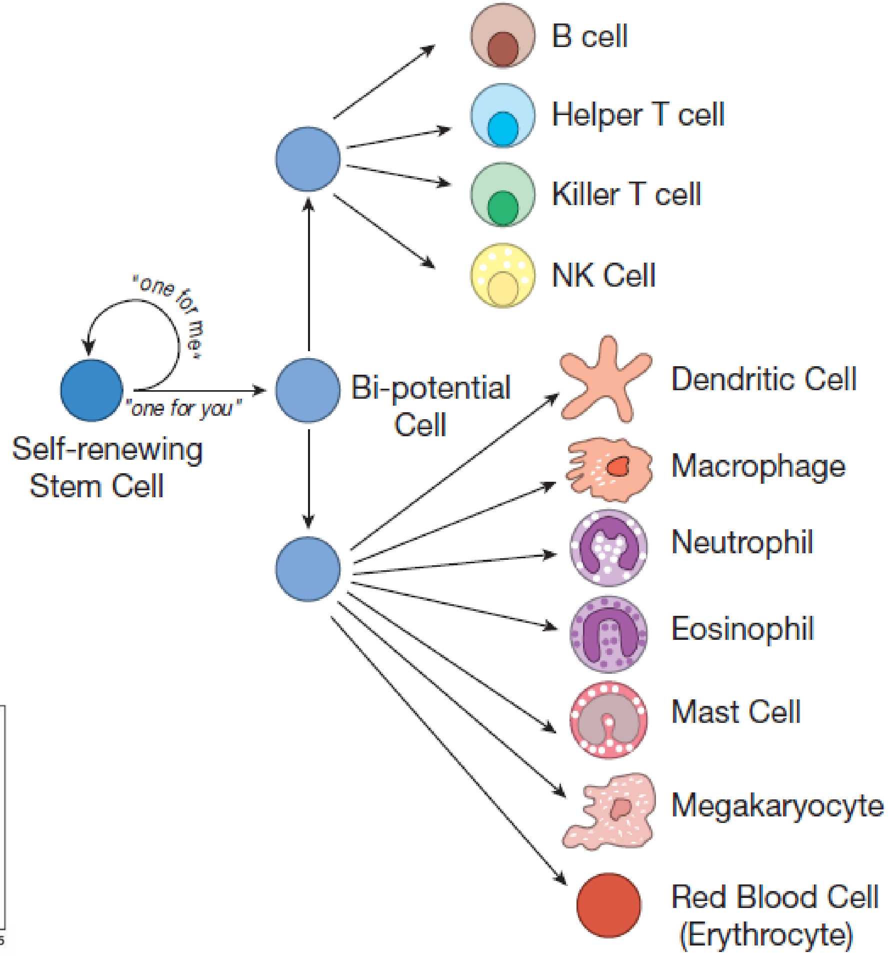
Flow Panel Design

Fluorophore	Marker	Description
FSC-A	Size	
SSC-A	Granularity	
FITC	IFNa	Pro-inflammatory cytokine (Th cell response)
PerCP-Cy5-5	MHCII	Expressed by APCs (B cells, Mono/Macs, DCs), upregulated upon infection and presentation of antigen
APC-Cy7-A	IL6	Pro-inflammatory cytokine (Th cell response)
Pac Blue	IL12	Pro-inflammatory cytokine (Th cell response)
Alexa Fluor 700	TNFa	Pro-inflammatory cytokine (Th cell response)
PE	CD123	Dendritic cells
PE-Cy7	CD14	Monocytes/Macrophages
APC-A	CD11c	Dendritic cells



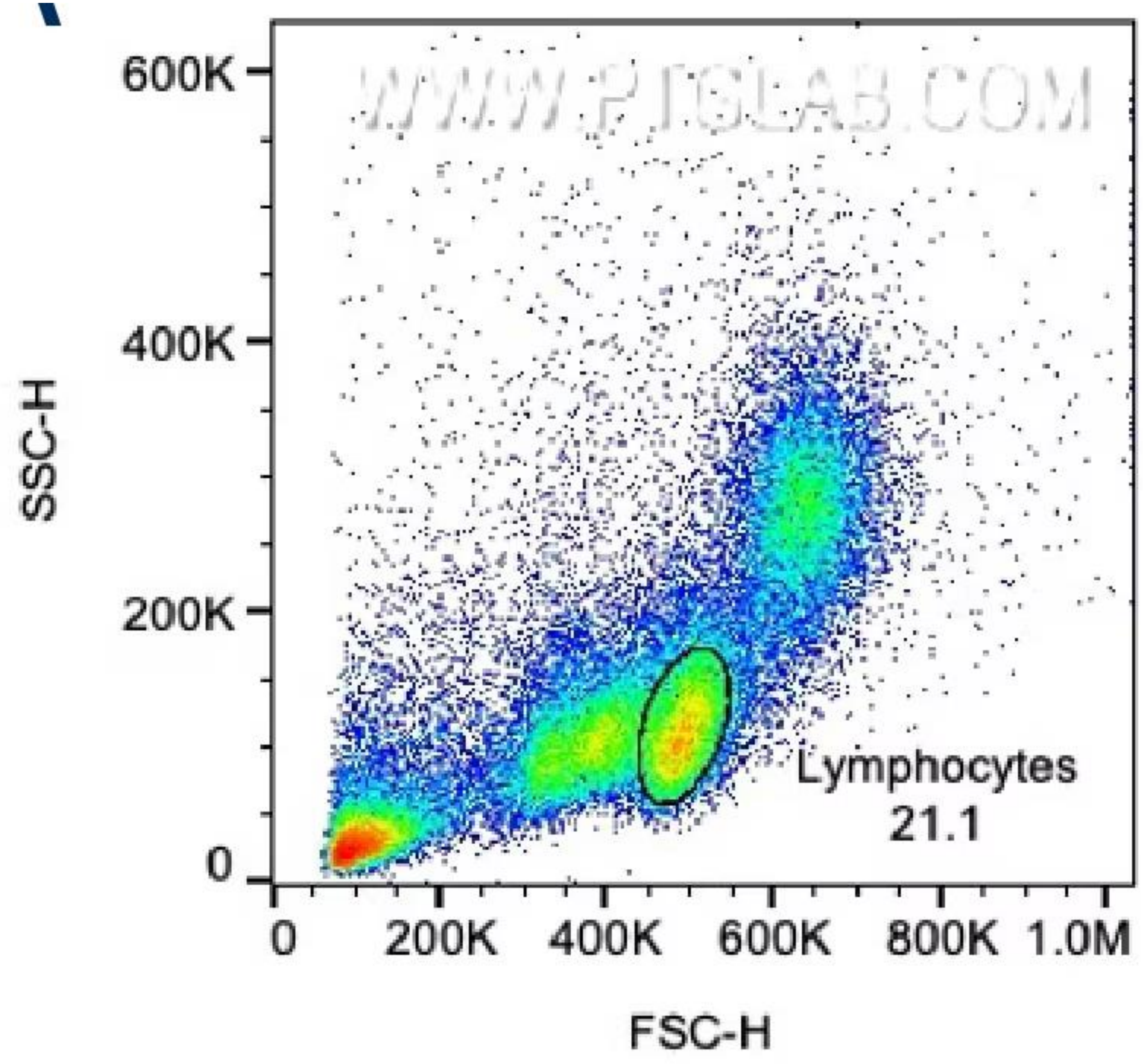


https://www.ptglab.com/news/blog/flow-cytometry-gating-for-beginners/

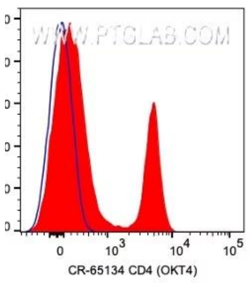


How the Immune System Works. L Sompayrac, Wiley 2019

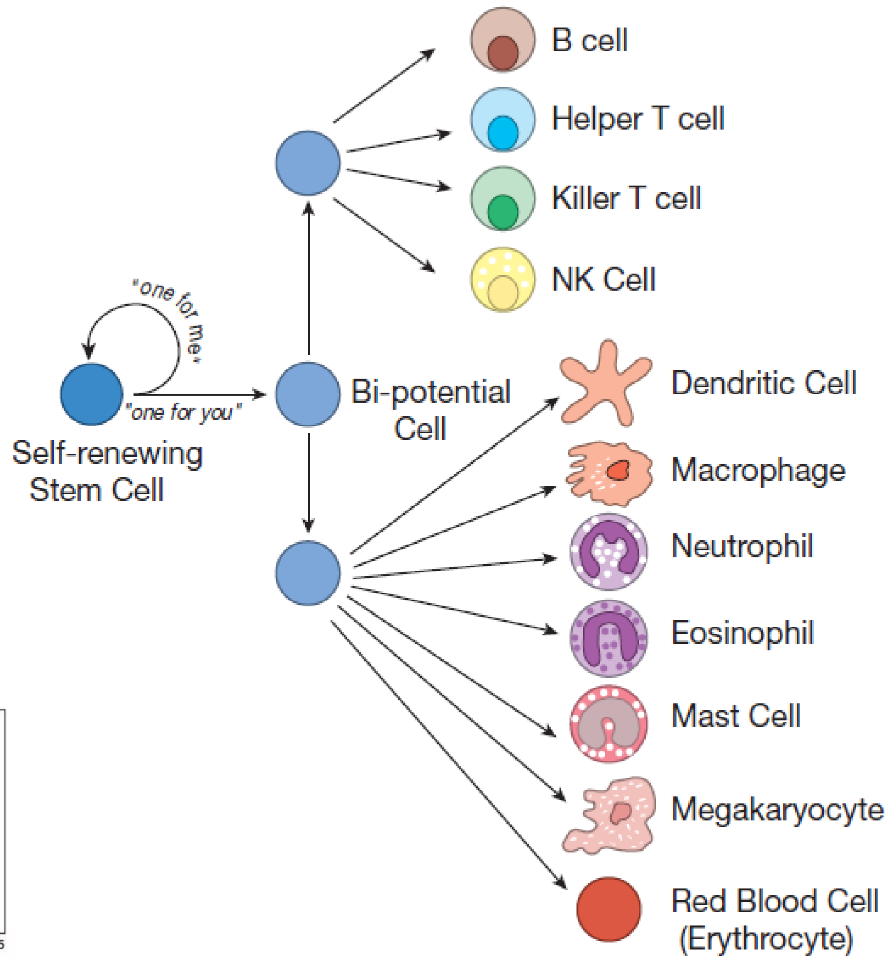
Flow Panel Design



onse)
, DCs), ation of
onse)
onse)
onse)



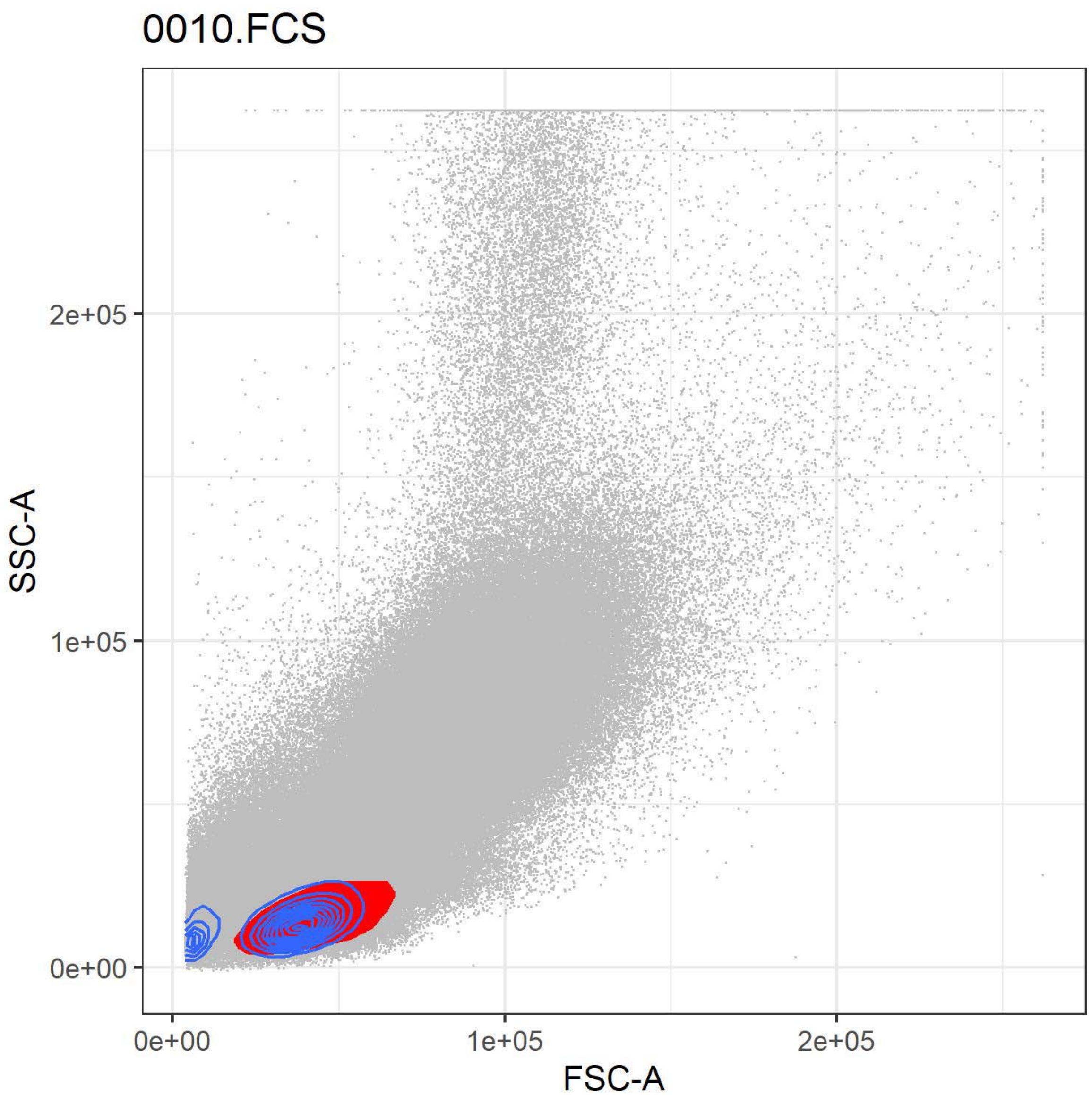
flow-



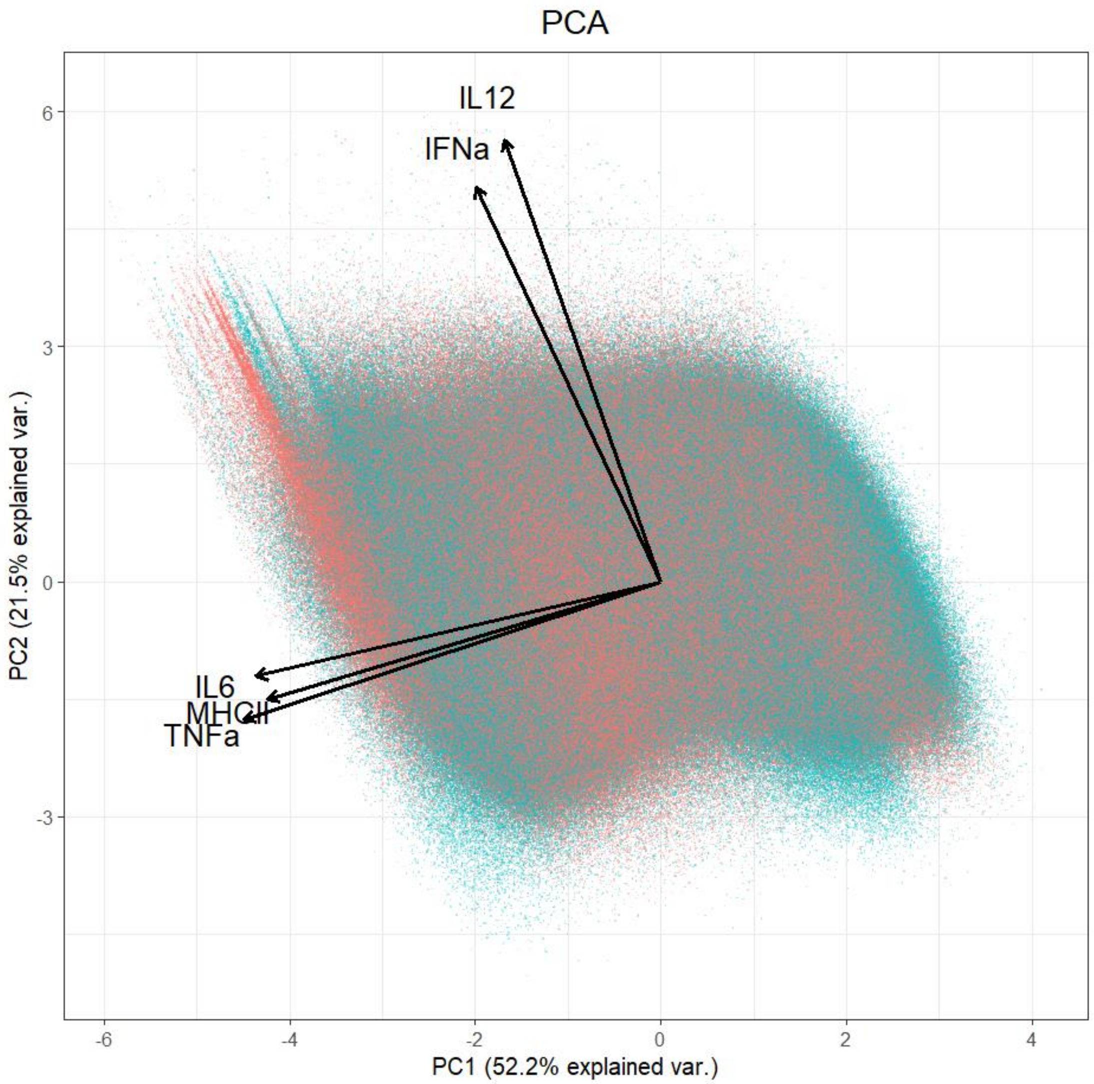
How the Immune System Works. L Sompayrac, Wiley 2019

Automated Gating for Data Preprocessing

- In each sample, estimate density in FSC vs SSC plot
- Identify landmarks (“hill tops”)
- Lymphocytes are the bottom right cluster
- Identify boundary of the cluster by pixels above a threshold
- Delete stand-alone points on the periphery of the cluster (pixel with most neighboring pixels being below the threshold)
- Convex hull (“rubber band model”) to smooth the boundaries
- Map pixels in density plot back to cells; delete all but lymphocytes

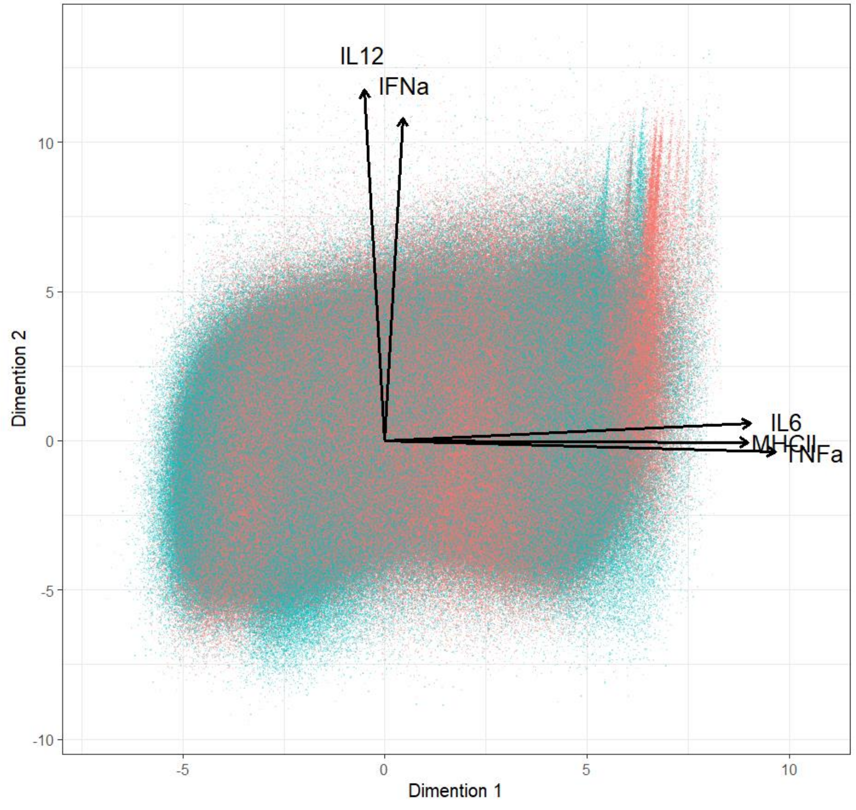


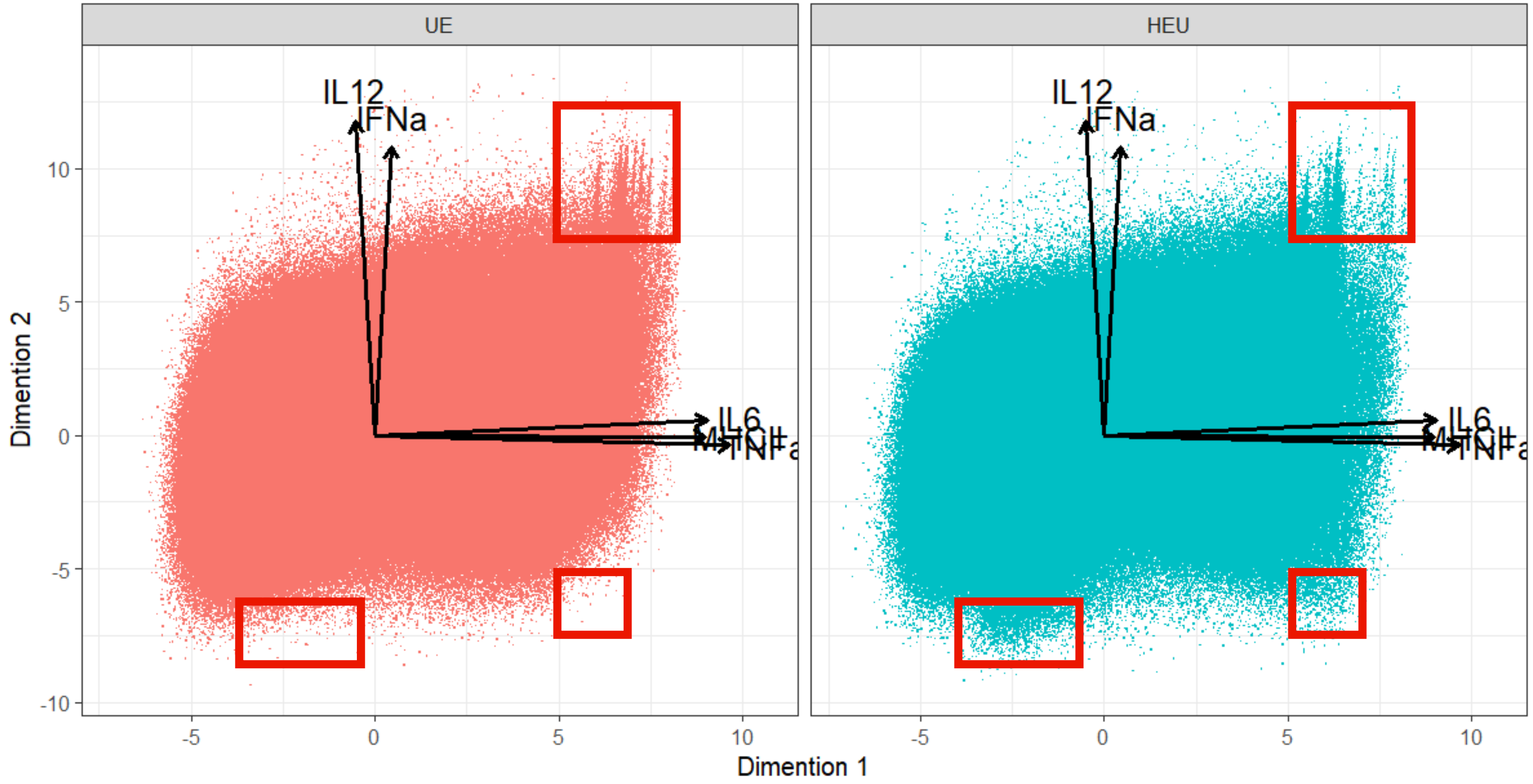
Principal Components Analysis



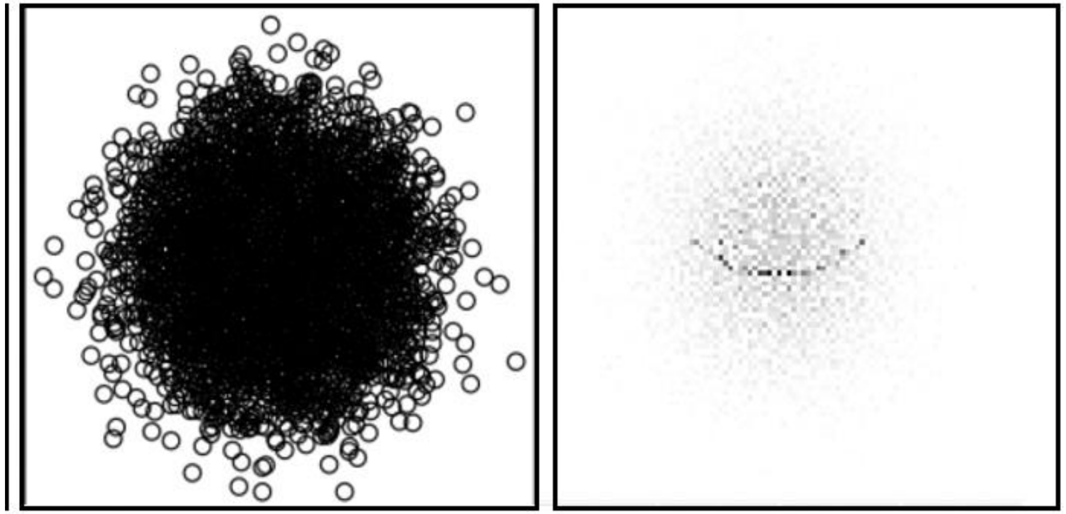
- PC1 explained 52.2% of variability; PC2 21.5%
- MHCII, IL6 and TNFa drove the differences in PC1 direction; IL12 and INFa in PC2 direction
- Majority of points from UE and HEU overlap. However, we are interested in profiles of cells that are different between UE and HEU
- Therefore, a different approach is needed – instead of looking for max variability (PCA), we want to look for max difference (PP)

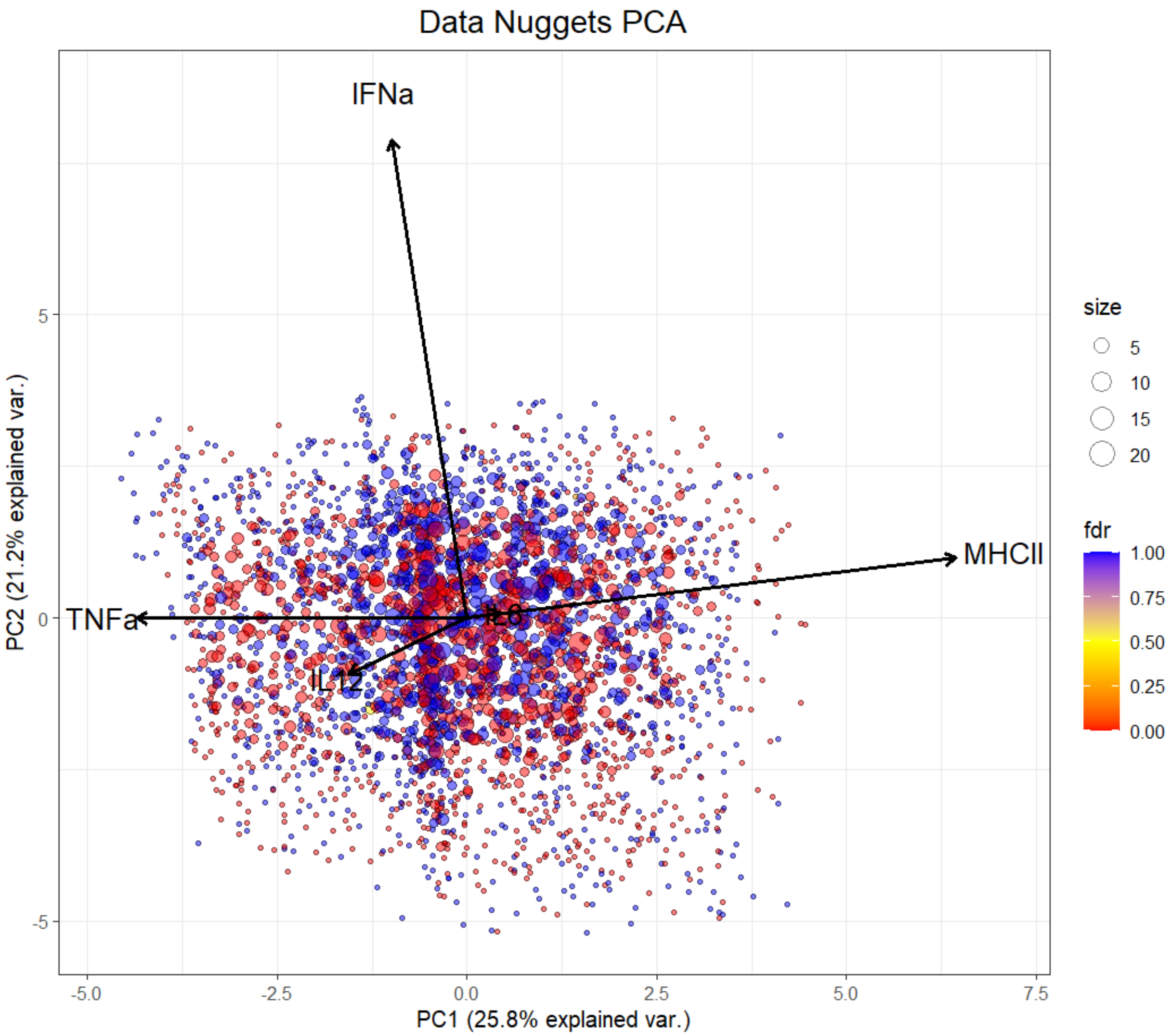
Rotated Principal Components and Differentially Populated Regions





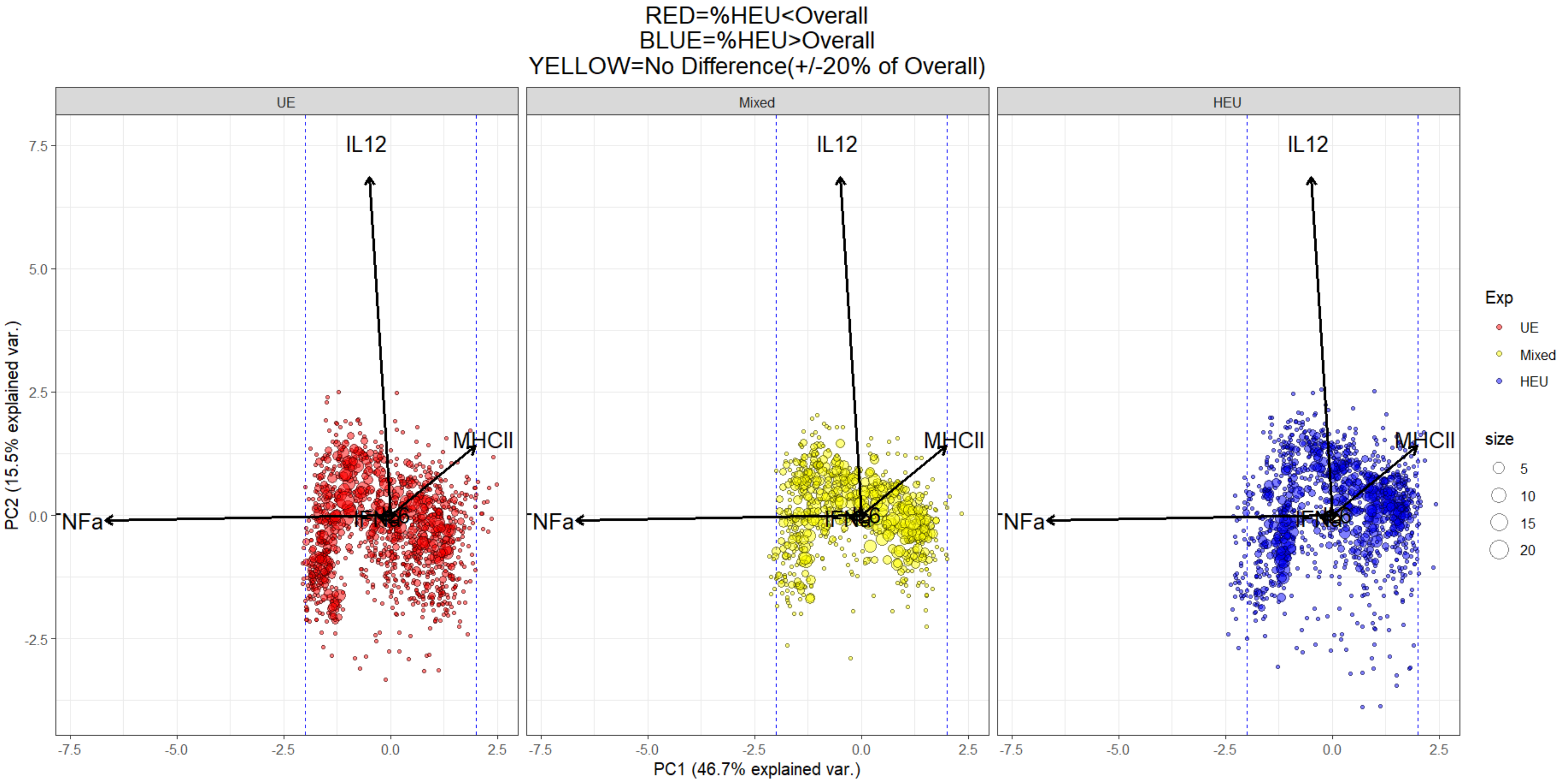
Data Compression with Data Nuggets





Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. Beavers et al, arXiv 2024

Data Nuggets Biplot by %HEU in Each Nugget vs. in Total (40.5%)



Key Technology

- **PP** searches multivariate *p*-dimensional data for lower *d*-dimensional projections, revealing the main structure of the data, i.e., clusters, outliers, and any other low-dimensional nonlinear structure (see Friedman and Tukey 1974).
- PP indices (e.g., Natural Hermite index) are functions to numerically measure features of low-dimensional projections
- Higher values of PP index = more interesting structures
- For PP index optimization, used Grand Tour Simulated Annealing (GTSA) algorithm

- The **Natural Hermite index** measures the distance between the *d*-dimensional distribution $f(y)$ and the *d*-dimensional normal distribution $\phi(y)$:

$$I^N = \int_{\mathbb{R}^d} [f(y)\text{-}\phi(y)]^2 \phi(y)dy$$

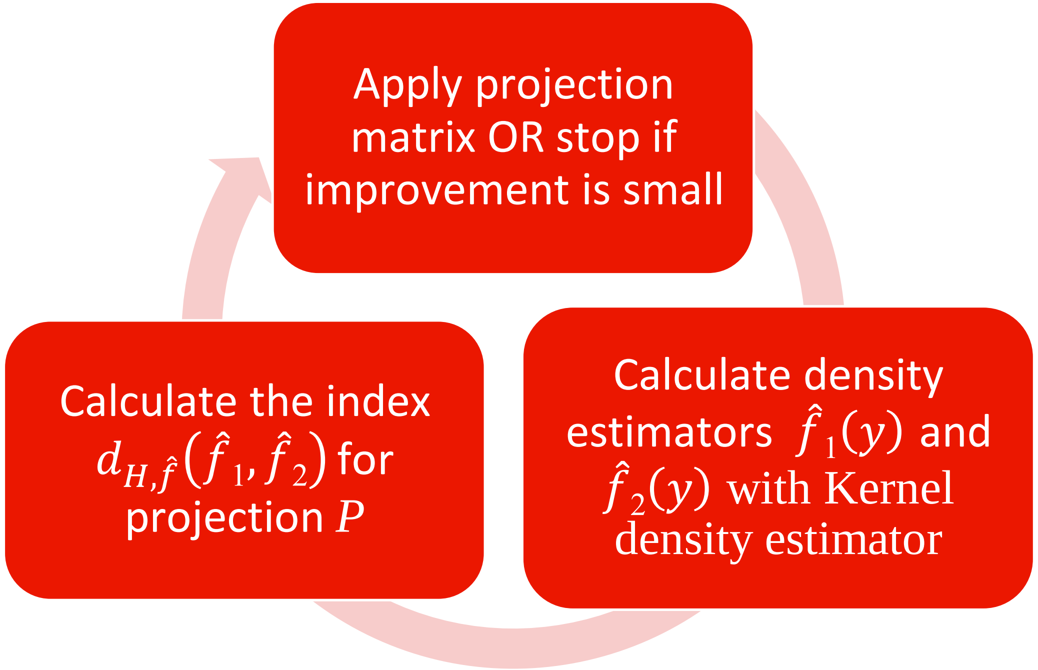
- Grand Tour algorithm assigns a sequence of projections onto (usually) 2-dimensional planes to any given dimension of Euclidean space.
- Flipping through the sequence of projection creates “data movie”

Differential Projection Pursuit

Let's define **Differential Natural Hermite** dissimilarity for *k* *d*-dimensional distributions:

- Let $f_1(x), \ldots, f_k(x)$ be a set of *k* density functions
- Let $f(x) = \frac{w_1 f_1(x) + \cdots + w_k f_k(x)}{w_1 + \cdots + w_k}$ be the weighted average
- For every pair of densities $f_i(x)$, $f_j(x)$ the differential Natural Hermite dissimilarity with respect to $f(y)$ is defined by:

$$d_f(f_i, f_j) = \left| \int_{\mathbb{R}^d} [f_i(x) - f_j(x)]^2 f(x)\, dx \right|^{\frac{1}{2}}$$

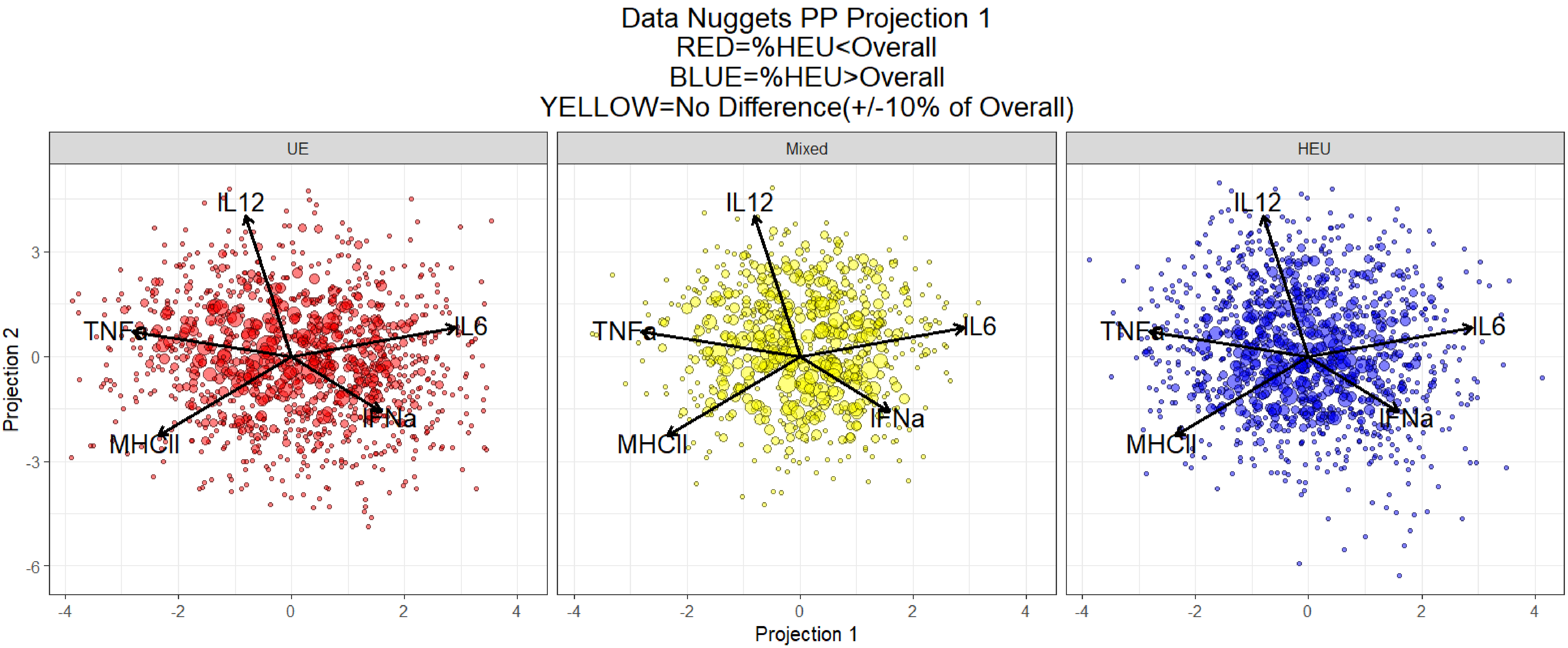


$$\hat{f}_B(\mathbf{y}) = \sum_{i=1}^{m} \frac{w_i}{\sum_{i=1}^{m} w_i} |\mathbf{S_i}|^{-1/2} \phi\left(\mathbf{S_i}^{-1/2}(\mathbf{y} - \mathbf{y}_i)\right)$$

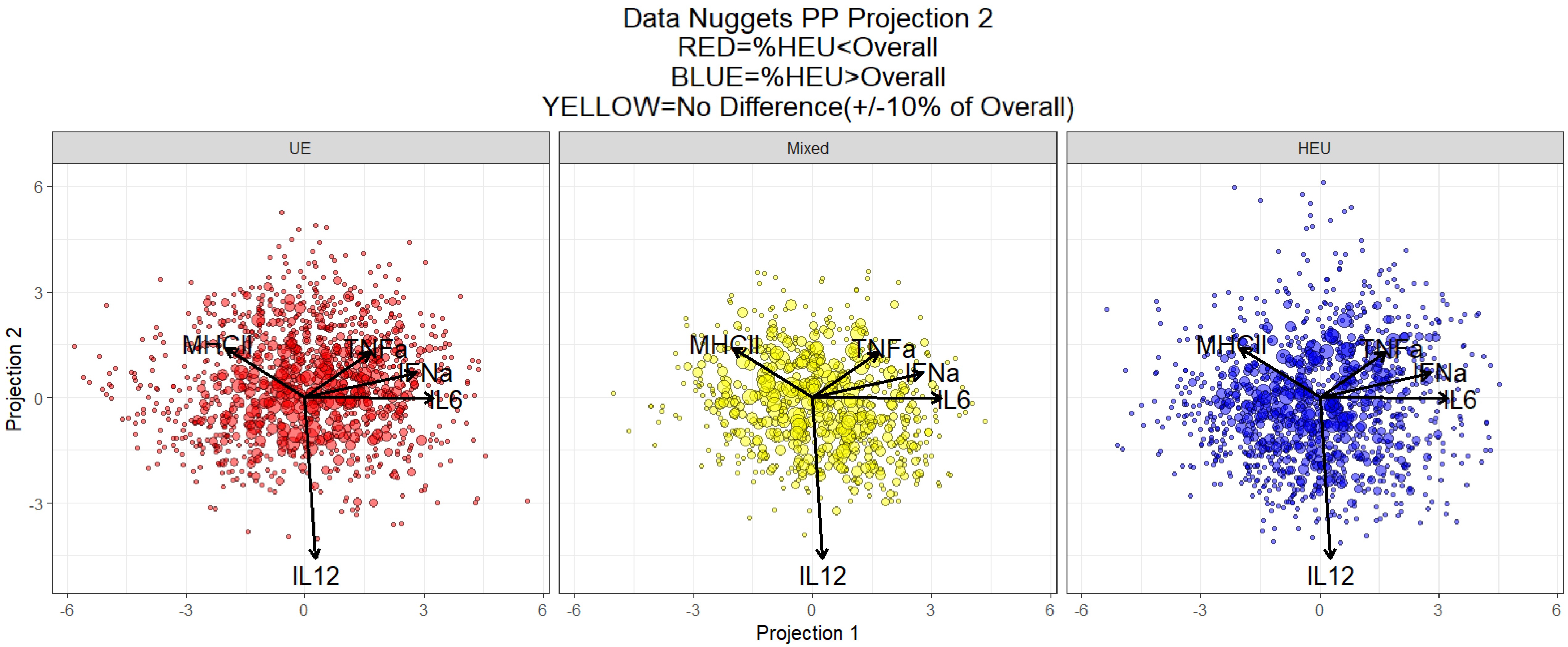
where $\mathbf{S_i} = \max\{s_i^2, \delta^2\}\mathbf{I}_d$ with a pre-determined minimal scale level δ.

Density estimator for big data sets based on data nuggets. Duan et al 2024 :

dPP Projection 1

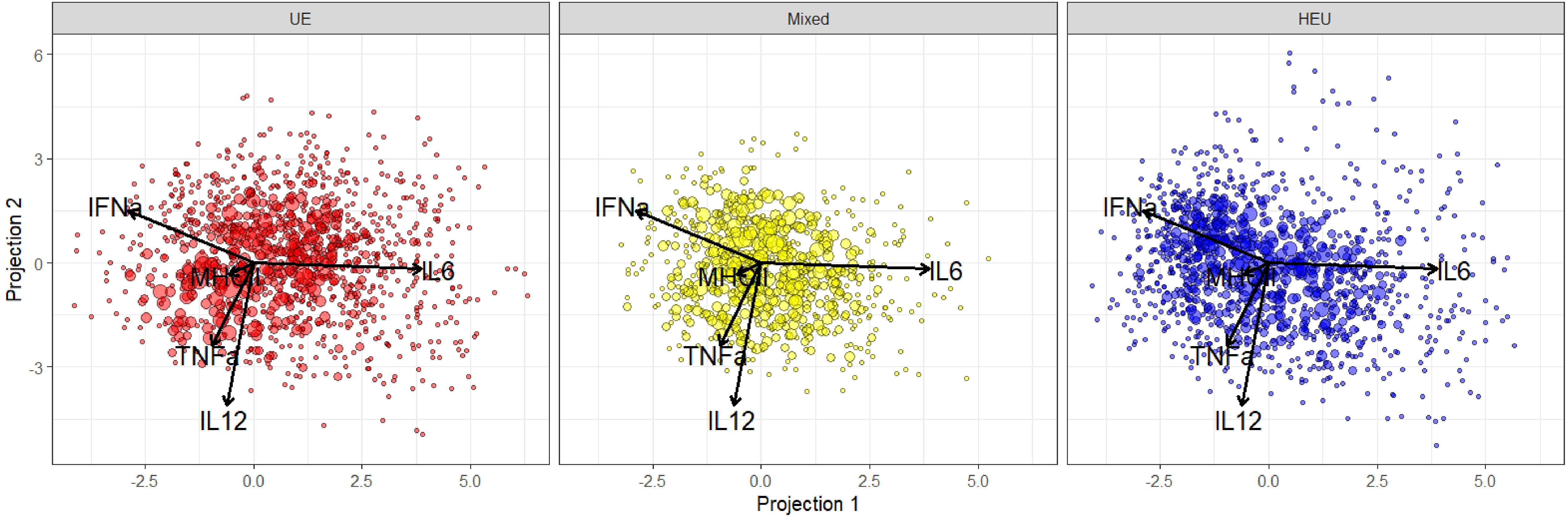


dPP Projection 2

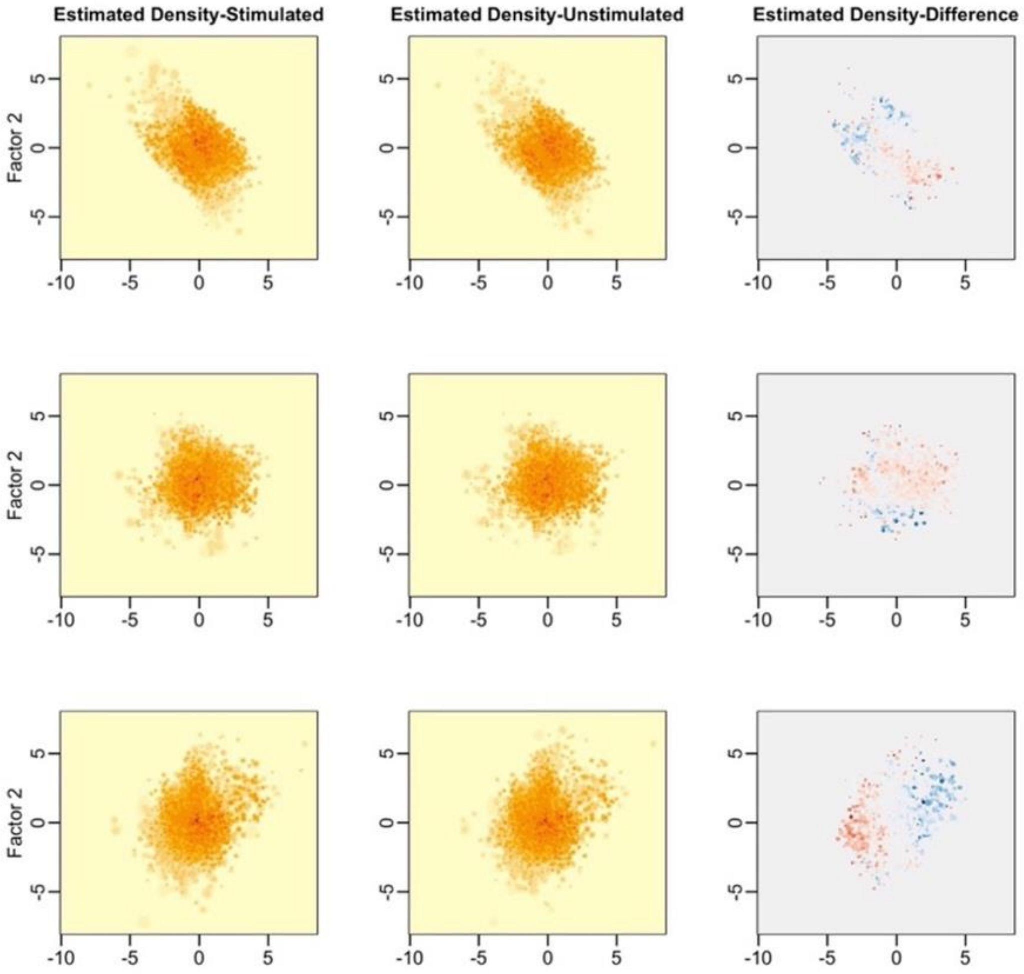


dPP Projection 3

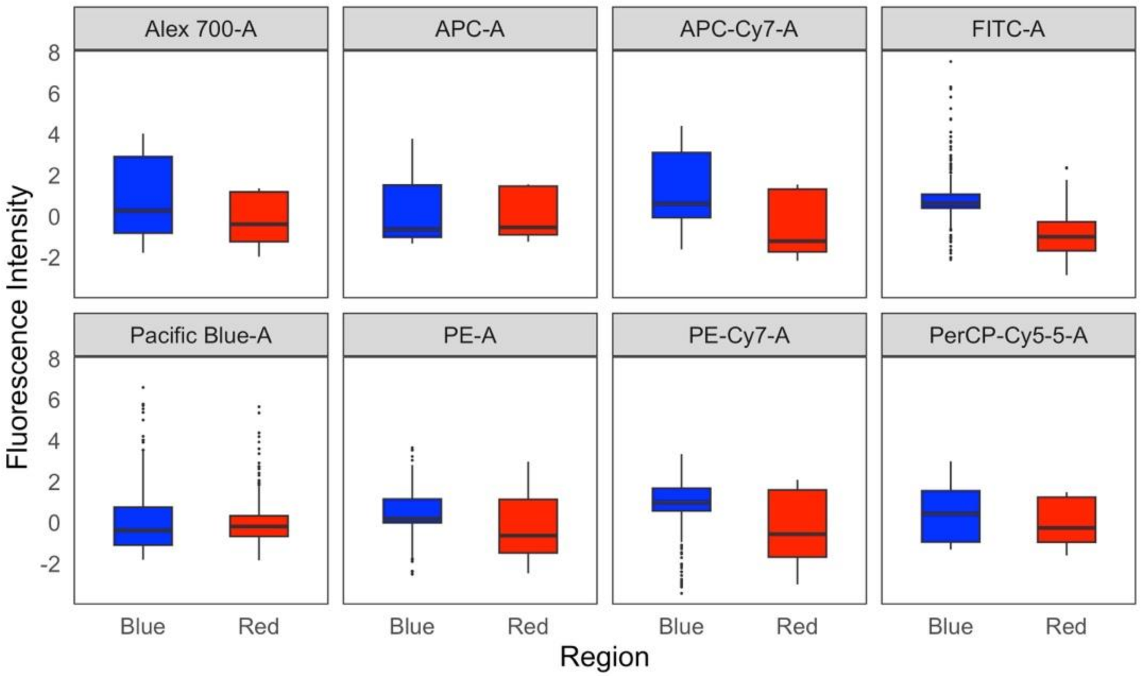
Data Nuggets PP Projection 3
RED=%HEU<Overall
BLUE=%HEU>Overall
YELLOW=No Difference(+/-10% of Overall)



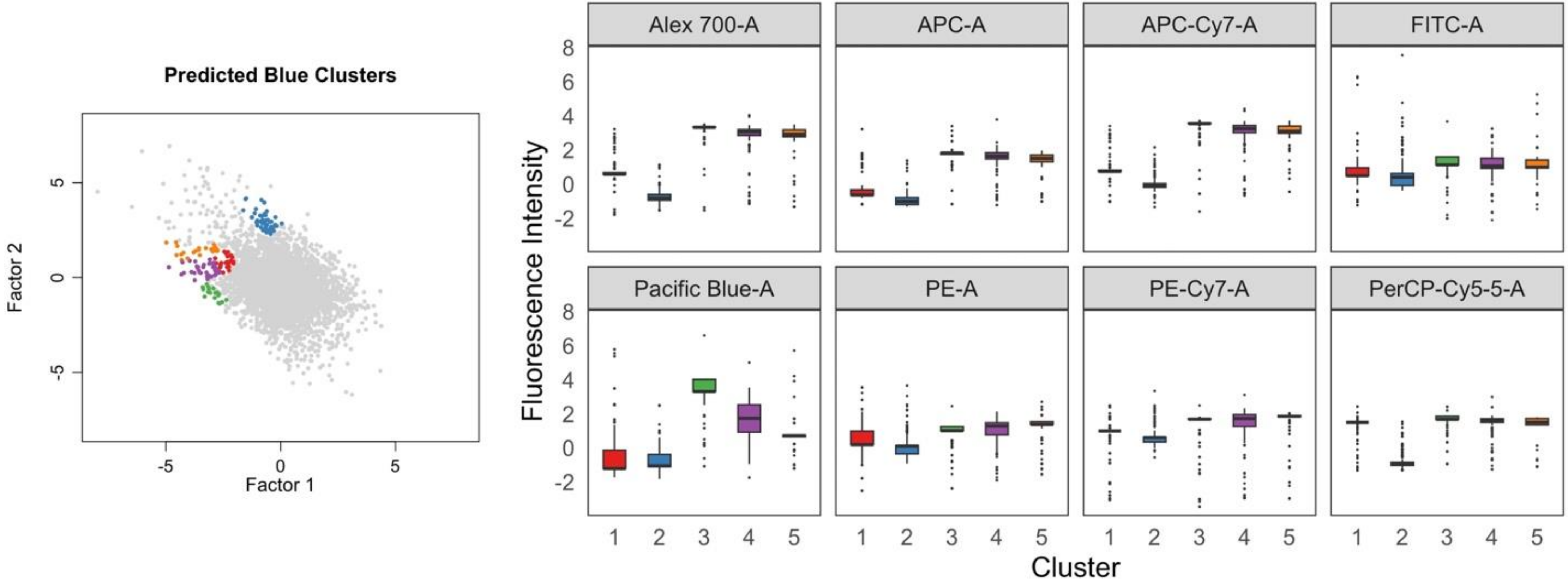
LPS vs Untreated Projections



- 3 projections of LPS vs. Untreated, optimized for dPP
- In the 3rd columnBLUE = LPS>Untreated; RED = LPS<Untreated difference between the two densities



Profiling Cells in Differentially Populated Regions



(a) (b)

Novel machine learning approach to differential flow cytometry analysis base on projection pursuit. Dastgiri et al, 2024 (submitted)

Conclusion

- Manual or automated gating of flow cytometry data might not be able to capture the structure of multidimensional data

- Differential Projection Pursuit creates 2D views complex multidimensional structures, optimized for maximal separation between the experimental groups

- The scientists and the statisticians must work as a team to correctly design, analyze and interpret the results of the experiments

Johnson&Johnson
Innovative Medicine

References and Publications

- *A New Projection Pursuit Index for Big Data.* Yajie Duan, Javier Cabrera, arXiv 2021 (under revision for JCGS)

- *Novel Machine Learning Approach to Differential Flow Cytometry Analysis base on differential Projection Pursuit*. Mahan Dastgiri, Yajie Duan, Davit Sargsyan, Abraham Adkwei, Rebecca Mary Peters, PoChung Chou, Ge Cheng, Chun-Pang Lin, Jocelyn Sendecki, Helena Geys, Kanaka Tatikola, Ah-Ng Kong and Javier Cabrera (under revision for JBS)

- *Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure*. Traymon E. Beavers, Ge Cheng, Yajie Duan, Javier Cabrera, Mariusz Lubomirski, Dhammika Amaratunga and Jeffrey E. Teigler. arXiv 2024

Thank you!

Johnson&Johnson
Innovative Medicine