Novel machine learning approach to differential flow cytometry analysis base on projection pursuit

Mahan Dastgiri 1, Yajie Duan1, Davit Sargsyan2,3,4, Abraham Adkwei4,5, Rebecca Mary Peter2,3, PoChung Chou2,3, Ge Cheng1, Chun-Pang Lin1, Jocelyn Sendecki4, Helena Geys4, Kanaka Tatikola4, Ah-Ng Kong2,3 and Javier Cabrera1

1Department of Statistics, School of Arts and Sciences, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

2Department of Pharmaceutics, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

3Graduate Program in Pharmaceutical Science, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

4Janssen Pharmaceuticals, Johnson and Johnson, Spring House, PA, USA, and Beerse, BE

5Epidemiology and Statistics Department, University of Georgia, GA, USA

**SUBMIT TO: Special Issue of the Journal of Biopharmaceutical Statistics (JBS) special issue on AI/ML**

**Correspondence**

Professor Javier Cabrera

Rutgers, the State University of New Jersey

School of Arts and Sciences, Room XXX

XXXXXXXX, Piscataway, NJ 08854

Phone: +1-XXX-XXX-XXXX

Email: [xavier.cabrera@gmail.com](mailto:xavier.cabrera@gmail.com)

RESOURCES (delete before submission)

1. Wiki page:

[https://en.wikipedia.org/wiki/Flow\_cytometry#](https://en.wikipedia.org/wiki/Flow_cytometry)

1. FCS standards (also, see publications):

<https://isac-net.org/>

Table of Contents

[Abstract 2](#_Toc137151707)

[1. Background 2](#_Toc137151708)

[2. Materials and Methods 2](#_Toc137151709)

[2.1 Experimental Design 2](#_Toc137151710)

[3 Results 2](#_Toc137151711)

[4 Discussion 2](#_Toc137151712)

[5 Figures and Tables 2](#_Toc137151713)

[6. References 2](#_Toc137151714)

# Abstract

Summary here. Write last.

# Background

## 1.1 Key principals of flow cytometry

* Multicolor flow cytometry (FC) is a laboratory technique used in biological disciplines including cancer biology, molecular biology, immunology and infectious disease monitoring to measure individual cell properties such as size, granularity and molecular composition. It is proven to be a highly sensitive quality control tool of fluorescent conjugates used in immunoassays (PMID:27936034) , with wide applications in immunophenotyping to simultaneously characterize mixed population of cells from blood, bone marrow and solid tissues based of cluster of differentiation (CD) markers, chemokine receptor markers, cytokines and antigen specific markers (PMID: [29512141](https://pubmed.ncbi.nlm.nih.gov/29512141)). Furthermore, fluorescent protein analysis revolutionized in vivo detection of bacterial or viral infections, transplanted cells and gene knockout in cells (PMID: 24813442). Cell cycle analysis is another critical assay to measure dynamics of pathophysiological processes by utilizing DNA binding dyes (PMID: 32441778

).

Immunophenotyping and antigen specific response are frequently employed in immunological application of flow cytometry. Immunophenotyping applies fluorochrome-conjugated antibodies to target antigens on the cellular surface, which are defined by cluster of differentiation (CD) numbers. The investigation of CD markers on immune cells indicates the specific population of tested cells. Antigen specific response works via fluorescent major histocompatibility or MHC multimers which carry the desired antigen and bound to tested immune cells recognizing the antigen. Following the binding, the response of the antigen is measured and evaluated (PMID: 29512141).

Sample preparation includes three approaches: (a) transfection and expression of fluorescent proteins, (b) staining with fluorescent dyes, or (c) with fluorescently conjugated antibodies (PMID: 29512141). To measure specific proteins on the surface or inside a cell, fluorescent chemical compounds called fluorochromes or fluorophores are added to the suspension. The fluorochromes are attached to molecules with affinity to specific proteins, hence labeling these proteins. Cells are first separated and suspended in a liquid, and the suspension is pass through narrow tubes, one cell at a time. A universal flow cytometer is composed of three systems, including fluidics (transporting samples to the laser intercept point), optics (inducing the visible and fluorescent light signals), and electronics (converting light signals to visualized readout in a computer) systems (PMID: 29512141). The instruments contain a large number of such tubes for parallel processing. As a cell moves through a tube, it is hit by a beam of light from a lamp or a laser. The light-excites fluorochromes then emit light in a relatively narrow band of wavelengths. The emitted light passes through a series of optical filters and dichroic mirrors deflecting it onto detectors (Figure X <DRAW CYTOMETER PICTURE>) @ Davit. Besides measuring light emitted by fluorochromes, flow cytometers also detect light scattered by the cells forward or to the side (FSC and SSC, respectively). The FSC and SSC measurements provide information about the cells’ physical properties and are used to separate single, live cells from cell clusters and debris during data preprocessing. Additionally, the instruments are able to measure electrical current impedance, i.e., the opposition to alternative current as the cells travel through the tubes. This allows for calculation of the cell size and additional physical properties. As of 2023, flow cytometers may contain as many as 10 lasers and up to 30 fluorochrome detectors.

The detectors convert the analog signal into digital and send the data to the instrument’s computer. The data collection process in flow cytometer is called *acquisition*. The data is typically saved in Flow Cytometry Standards (FCS) format as a matrix, with rows representing individual cells and columns the markers [1]. FCS specifications were developed and are maintained by the international Society for Advancement of Cytometry (ISAC).

Flow cytometry is used in biology to achieve a variety of goals including cell genotyping, sorting and studying apoptosis but, in this work, only one specific, widely used experimental design will be considered, namely, studying treatment effects on immune cell differentiation. Administering potent test compounds to naïve immune cells leads to their specialization that is manifested through changes in cell surface markers. Cytometers identify and quantify these markers allowing for differential analysis of the samples across treatment groups.

Following acquisition, the data is processed, traditionally using a technique called *gating*. Specialized tools such as FlowJo and … import FSC files and plot the data, 2 dimensions at the time. The investigator draws areas of interests, or gates, to manually identify clusters of cells that they are interested in. This process of gating goes on sequentially as the investigator focuses on specific subpopulations of cells. The gating strategy follows current understanding of differentiation process, with major differentiating proteins gated first (Figure X <ADD PLOT EXAMPLE OF GATING STRATEGY>). Once gating is completed, the software will count the number of cells in each gate and output a processed data file. Often, the interest is not or not only the counts, but the ratios of child-parent populations as defined by the gating strategy, i.e., frequencies.

# 2 Materials and Methods

## 2.1 Data Source and Experimental Design

@Mahan

The HIV-exposed-uninfected versus unexposed (HEUvsUE) dataset was obtained from the Flow Repository website (<http://flowrepository.org/id/FR-FCM-ZZZU>). The data was collected to find cell populations that can be used to differentiate between HIV-exposed-uninfected (HEU) and unexposed (UE) infants. For this purpose, blood samples were taken from infants 6 months after birth, where some were stimulated with six Toll-like–receptor ligands and some were left unstimulated for control [5]. There are 308 FCS files from 40 patients. In this study, we aimed to find the region within the data where stimulated and unstimulated cells differ the most. We selected data files from two patients, one HIV-exposed-uninfected patient and one unexposed patient. For each patient, we considered [Lipopoly](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2196288/)saccharide (LPS)-stimulated and unstimulated, making it 4 data files in total and 1,754,745 cells in combination. Each data file provides information on eight cell markers, FSC, and SSC. The details can be seen in Table 1.

Table 1: List of reporter and analytes for the datasets.

|  |
| --- |
| Channel Reagent |
| FSC-A  SSC-A  FITC-A IFNa  PE-A CD123  PerCP-Cy5-5-A MHCII  PE-Cy7-A CD14  APC-A CD11c  APC-Cy7-A IL6  Pacific Blue-A IL12  Alex 700-A TNFa |

## 2.2 Data compression with Data Nuggets. (Javier and Kanaka)

Our calculations in the rest of this paper are not possible to do based on the raw data because when you have 1.5 million observations any calculation of order n2 is not computable. For this reason we apply a compression algorithm call data nuggets (see Beavers e.a. 2022) that represents a dataset of millions of observations with a weighted set of a few thousand observation that are called data-nugets. Data-nuggets preserve the structure of the data much better than random samples and for this reason they are more suitable to find the true data structures using low dimensional projections. In summary data nuggets compression reduces a large dataset into a smaller collection of data nuggets while preserving the underlying structure.

## 2.3 Projection pursuit and differential projection pursuit(Javier)

Projection Pursuit is a technique that searchers for projections of multivariate p-dimensional data into lower d-dimensional projections containing the main structure of the data. By main structure we mean clusters, outliers and any other low dimensional nonlinear structure. These methods were introduced by Friedman, Tukey (1972) for finding structure, while exploring a 9-dimensional data from particle physics. Later Friedman (1982) introduce Friedman index as an example of projection pursuit (PP) indices. Cook e.a (1993) made substantial progress in this area by introducing several new PP indices. In particular, they introduced the natural Hermite index that became very popular and will be the center tool of this paper. The natural Hermite index measures the distance between any *d*-dimensional distribution and a *d*-dimensional normal distribution .

Text

Description automatically generated

The computational burden of these indices is satisfactory for small to moderate data sets but not attainable for large datasets. Therefore Duan, Cabrera (2023) introduced weighted versions of the PP indices computed over data nuggets. They showed that the “most interesting” projections found by the Natural Hermite index on large datasets are identical to those found by the weighted version of the index over data-nuggets.

In our case we are interested in comparing d-dimensional projections of two p-dimensional samples where d < p. For this we are introducing a differential version of the Hermite index for comparing any two d-dimensional distributions and . Let be the estimate of combined data distribution. It could be estimated by combined data or by the average of two distributions . The differential Hermite index for two distributions is given by the following formula:

This index is maximized over all d-dimensional projections of our data, where and are estimated by the d-dimensional Kernel density estimators of our projected data and . The steps are

1. Apply a projection matrix *P* to our two data sets: and .
2. Calculate density estimators and from our projected data , and the combined density from their average or combined projected data.
3. Calculate the index for projection P using and .
4. Repeat step (i)-(iii) to maximize over all projections P.

This procedure usually finds a local maximum, so it needs to be repeated a few times with different initial P’s to obtain some good local maxima. Usually, we will obtain 3 or 4 local maxima projections.

When we want to compare *k* >2 samples we require to evaluate *k*(*k*-1)/2 integrals. But in Weigle e.a.(2023) it was shown that

=c ,

which requires to evaluate only *k* integrals. Therefore the algorithm to find the optimal projection for the difference between two groups can be extended to *k* groups.

Based on the new Natural Hermite index for big data proposed by Duan et al., the differential Hermite index could be applied on large data sets by constructing data nuggets. For big data sets, data nuggets are firstly constructed and refined with centers, weights and scales. After sphering the data nugget centers with weights, let be the m projected nugget centers for two data sets k = 1, 2, and let be the scale vector of data nuggets. The density estimators and in the differential Hermite index can be calculated based on the nuggets:

where with a pre-determined minimal scale level . The combined density could be estimated by the data nuggets constructed by the combined data sets, or alternatively approximated by the average of and .

## 2.4 Factor analysis and clustering

@Javier, @Yajie, @Mahan

Once the optimal projections are obtained, Varimax rotation is conducted, and the orthogonal basis is rotated to express the two main axes of the projections as a function of a few of the proteins (surface markers). SVM is then performed to predict the blue and red clusters of cells, which corresponds to positive and negative difference between the two densities, respectively. In order to find the cluster of cells that differ the most between groups, weighted K-means clustering is performed.

# 3 Results

@Mahan @Javier and @Yajie to provide the results, @Davit and others to interpret.

Since performing the Differential Projection Pursuit is computationally expensive in the presence of 1,754,745 cells, we first conducted the mentioned data nugget method. After obtaining 3385 refined data nuggets from the HEUvsUE dataset, we obtained six projections and conducted varimax rotations on them. We have estimated the density of the stimulated cells as , the unstimulated cells as , and the difference between them as − , which are plotted in Figure 1. The column of graphs on the right shows several blue clusters where the stimulated cells were more abundant than the unstimulated. Alternatively, the red clusters are the regions where the unstimulated cells are more abundant than the stimulated.

In Table 2, we see the proportion of stimulated and unstimulated cells from the total in the predicted blue region for each projection, as well as the percentage and counts of the stimulated and unstimulated cells for the region. For the rest of the analysis, we focused on projections 1 and 3. Figure 2 pictures the projected data nuggets for the first and third projections displayed in Figure 1. The left plots are the projected data nuggets, while the right plots are the projected raw data using the same projection matrices. In the middle, we have the estimated densities using the data nuggets with contours.

Figure 3 shows the protein expression levels for the predicted blue and red regions we obtained. The loading for each surface protein for projections one and three is represented in the loading plots shown in Figure 4. For the first projection, proteins FITC-A and PerCP-Cy5-5-A contribute significantly to the first direction, while proteins APC-A, APC-Cy7-A, and PE-Cy7-A strongly correlate with the second direction. For the third projection, proteins APC-Cy7-A and PE-A made significant contributions to the first direction, and proteins Alex 700-A, APC-A, and PerCP-Cy5-5-A have strong relationships with the second direction.

In order to detect clusters of interest, weighted K-means was applied to the predicted blue regions for the first and third projections, where w one three respectively. The clustering results are picturedin Figure 5 and Figure 6, accordingly The box plots in Figures 5 and 6 show the protein expression levels for each cluster, while the plots on the left picture the projected data nuggets where the clusters we obtained within the predicted blue regions are in colors. Tables 3 and 4 represent the proportion of stimulated and unstimulated cells from the total in each cluster of the predicted blue region for projections 1 and 3, as well as the percentage and counts of the stimulated and unstimulated cells.

The same analysis can be performed for the red region. In figure 7, we display the clusters we obtained from the red region of the differential projection pursuit. The analysis which consists of the boxplot of the clusters and tables with the proportion and percentage of cells in each cluster is shown in supplement to the paper in Appendix 1.

A screenshot of a computer generated image

Description automatically generated

Figure 1: Density plot of Stimulated (first column), Unstimulated (second column) data nuggets, and the difference between them (third column) for 6 projections that were obtained by optimizing the data nuggets projection pursuit Hermite index.

Table 2: Proportion of the stimulated and unstimulated cells in the predicted blue region. The three numbers in the third column are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each of the six projections.

(the two tables are the same with different layout; will keep one)

|  |  |  |  |
| --- | --- | --- | --- |
| Projection |  | Predicted blue cluster based on 2-d projection | Predicted blue cluster based on nugget centers |
| 1 | Stimulated | 0.020 (53.5%) 10870 | 0.087 (53%) 46624 |
| Unstimulated | 0.010 (46.5%) 9464 | 0.044 (47%) 41328 |
| 2 | Stimulated | 0.014 (51.5%) 7678 | 0.110 (51.9%) 59130 |
| Unstimulated | 0.008 (48.5%) 7238 | 0.059 (48.1%) 54700 |
| 3 | Stimulated | 0.017 (52.9%) 8879 | 0.253(49.2%) 135400 |
| Unstimulated | 0.009 (47.1%) 7921 | 0.15 (50.8%) 139596 |
| 4 | Stimulated | 0.008 (54.4%) 4539 | 0.075 (54.4%) 40324 |
| Unstimulated | 0.004 (45.6%) 3804 | 0.036 (45.6%) 33823 |
| 5 | Stimulated | 0.017 (50.5%) 8850 | 0.196 (49.6%) 104937 |
| Unstimulated | 0.009 (49.5%) 8682 | 0.115 (50.4%) 106823 |
| 6 | Stimulated | 0.032 (51.4%) 17221 | 0.100 (53%) 53714 |
| Unstimulated | 0.017 (48.6%) 16256 | 0.051 (47%) 47675 |



A group of images of data nuggets

Description automatically generated

Figure 2: Projected data nuggets (first column), estimated density plot for the data nuggets (second column), and projected raw data (third column) for the first and third projections

A diagram of different colored lines

Description automatically generated with medium confidence A diagram of different colored lines

Description automatically generated

Figure 3: Expression level for each protein in the predicted blue and red regions of the first projection (left picture), and the third projection (right picture)

A graph with numbers and symbols

Description automatically generated with medium confidenceA graph with black text

Description automatically generated with medium confidence

Figure 4: Loading plots for the first (left) and third (right) projection.

A graph with colored dots

Description automatically generatedA group of colorful boxes with text

Description automatically generated with medium confidence

Figure 5: Five clusters of the predicted blue region for projection one (left picture) and protein expression level for each cluster (right picture)

A graph showing a number of blue clusters

Description automatically generatedA chart of different colored boxes

Description automatically generated with medium confidence

Figure 6: Five clusters of the predicted blue region for projection three (left picture) and protein expression level for each cluster (right picture)

Table 3: Proportion of the stimulated and unstimulated cells in each 5 clusters of the predicted blue region of the first projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| Stimulated | 0.0038 (55.2%) 2013 | 0.0098 (53.9%) 5259 | 0.0011 (51.1%) 582 | 0.0039 (51.3%) 2098 | 0.0017 (54.0%) 920 |
| Unstimulated | 0.0018 (44.8%) 1634 | 0.0048 (46.1%) 4498 | 0.0006 (48.9%) 558 | 0.0021 (48.7%) 1990 | 0.0008 (46.0%) 784 |

Table 4: Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted blue region of the third projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
| Stimulated | 0.0041 (50.5%) 2216 | 0.0036 (59.3%) 1904 | 0.0012 (55.2%) 665 | 0.0027 (46.3%) 1428 | 0.0018 (56.5%) 977 | 0.0013 (58.2%) 703 | 0.0018 (49.9%) 986 |
| Unstimulated | 0.0023 (49.5%) 2170 | 0.0014 (40.7%) 1307 | 0.0006 (44.8%) 539 | 0.0018 (53.7%) 1659 | 0.0008 (43.5%) 752 | 0.0005 (41.8%) 504 | 0.0011 (50.1%) 990 |

A graph showing a red cluster

Description automatically generated A graph with colored dots

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

# 4. Discussion

Gating approach to flow data cytometry was determined in part by biology but also by limitation of computing power and tools that would allow multidimensional data visualization and analysis. Plotting and clustering such data two dimensions at the time went around these limitations. However, such projections can present severely distorted images of a multidimensional object, masking important patterns. Additionally, results from manual gating are highly dependent on the investigator’s perception and experience and are almost certainly non-reproducible. Automated gating can improve reproducibility, but it still does not address the dimensionality issue. In this worked, we stepped back from gating and instead examined the data in its true dimensional space. Applying data nuggets reduced the amount of data by grouping individual cells into typical groups. Projection pursuit found optimal projections that revealed the most information about the data. Finally, by comparing projections of samples with different treatments, we identified cell subpopulations that had significantly different densities between the treatment groups. Some of these subgroups were identifiable using current classification of immune cells based on surface markers and physical characteristics of the cells while several of the subgroups could represent new subtypes or reveal previously unknown mechanism. The latter would require more careful examination and interpretation by biologists, as well as conformation from other data sets.

This methodology has the potential to create a platform for analyzing large and complex data revealing multiple aspects of immune system and disease biology such as cytokine release by stimulated immune cells, different states of signal transduction pathways and complex cell functions. It can help identify specific subsets of cells in normal and pathological conditions with unique biological functions, considering enormous heterogeneity of cells.

@Davit:

Based on discussion with Maggie, discuss experimental design - Stain Index, color assignment, compensation, Data from different instruments is not comparable as each instrument have its own setting, and lasers need to be calibrated whole the time…

@Javier:

1. The use of data nuggets is to make this work with very large data tables
2. Cons of using 2D for N-D data
3. WH use PP? PP finds best projections and is not attached to individual markers although we try to find projections with axis using minimal number of markers
4. Discuss data: different treatment groups so we are interested in differential analysis
5. Did we find any markers (in the results) that are interesting? Are any of the clusters interpretable? Can we find an example (or synthetic data) where a cluster is masked if you look at simple projections but revealed in PP? E.g., 4D data with 4 clusters along the 4D diagonal, e.g., a cylinder along the 4D diagonal. The clusters will be masked in simple 2D projections but if w project on the diagonal, the difference is visible even in 1D. More complicated – 4D diagonal and some linear combination. Individual 2D scatter plots will look bad; hence, gating would not work.

@ALL: please contribute

# 6. References

1 Spidlen, J., Moore, W., Parks, D., Goldberg, M., Bray, C., Bierre, P., Gorombey, P., Hyun, B., Hubbard, M., Lange, S., Lefebvre, R., Leif, R., Novo, D., Ostruszka, L., Treister, A., Wood, J., Murphy, R.F., Roederer, M., Sudar, D., Zigon, R., and Brinkman, R.R.: ‘Data File Standard for Flow Cytometry, version FCS 3.1’, Cytometry A, 2010, 77, (1), pp. 97-100

2. Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory

data analysis*. IEEE Transactions on computers,* 100(9), 881-890.

3. Friedman, J. H. (1987). Exploratory projection pursuit. Journal of the American statistical association, 82(397), 249-266.

4. Cook, D., Buja, A., & Cabrera, J. (1993). Projection pursuit indices based on orthonormal function expansions. Journal of Computational and Graphical Statistics, 2(3), 225-250.

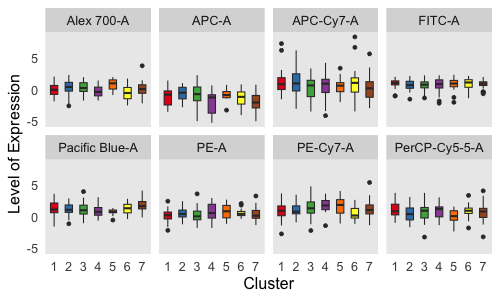
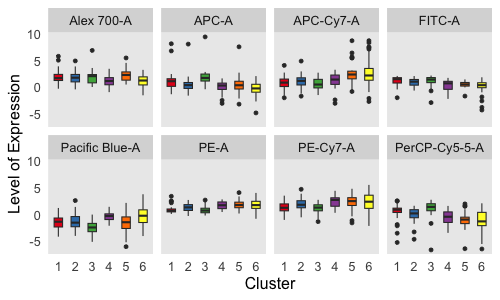
5. Duan Y, Cabrera J, A New Projection Pursuit Index for Big Data. (in revision)

6. Beavers, T., Cabrera, J., Chen G, Duan Y, Lubomirski, Tiggler, S, M. (2019). Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. (in revision)

# **Supplement**

## Appendix I Results of the analysis of the red region of the projections one and three

The box plots in Figure 8 show the protein expression levels for each cluster of the red region in the first and third projections. Tables 5 and 6 represent the proportion of stimulated and unstimulated cells from the total in each cluster of the predicted red region for projections one and three, as well as the percentage and counts of the stimulated and unstimulated cells.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |