Introduction to Computational High-Dimensional Flow Analysis and Practical Considerations

Paul D. Simonson, M.D., Ph.D.

Australasian Cytometry Society 2024 Conference

October 20, 2024

Outline

- Why use computational methods for high-dimensional flow data?
- Flow clustering algorithms
- FlowSOM
- Dimensionality reduction
  - t-SNE, UMAP
- Visualization in flow software

What is high-dimensional flow data?

- It’s relative!
- Yesterdays 4-color is now 10-color, which will soon be 12-color…
- CyTOF data: usually ~40 markers
- Spectral flow cytometry

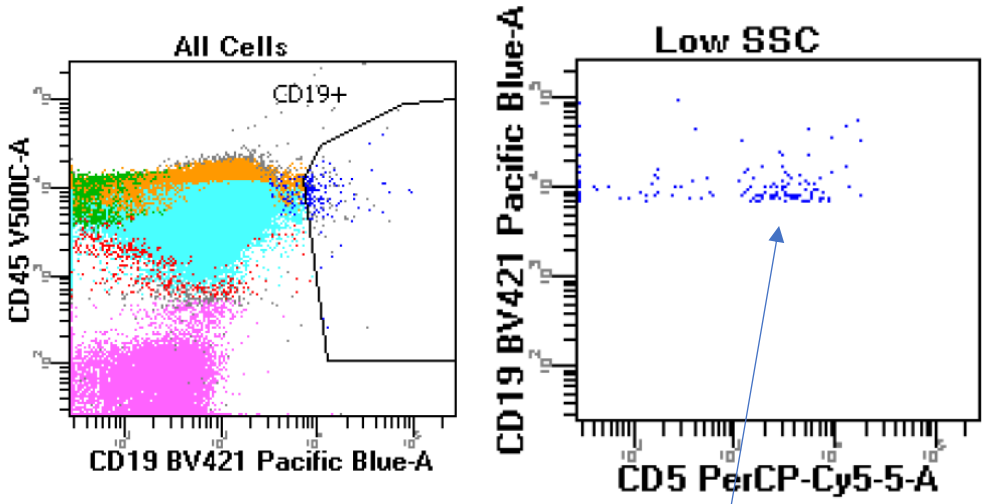
Example application: 18 color flow cytometry to evaluate T cell subsets

- Your lab has purchased an 18-color flow cytometer
- You now want to offer a new T cell panel (22 antibodies) for immunomonitoring in clinical trials, and, eventually, clinical use.
- You would like to be able to parse the cells into T cell subsets for identification and quantification
  - Minimize subjectivity
  - Include the ability to identify unexpected subsets
- **You have decided to employ computational methods in addition to traditional gating to help in the analysis.**

Why use new computational approaches?

- Increasing numbers of flow channels means increased complexity.
  - Adoption of spectral flow cytometry further increases complexity!
  - More colors allows identification of more cell subsets within the data.
  - That's a lot of 2x2 plots to look at!
- Increasing numbers of gates leads to increasing chances of spillage of cell subsets into the wrong gates.
- Gating creates bias that can result in missing unexpected populations.
- Gating requires some subjective decisions, limiting reproducibility.

- **Computational approaches can result in a less biased, more reproducible approach to flow cytometry analysis.**



T cells that spilled into the CD19+ cells gate, *not* CD5+ B cells

Basic components of computational analysis

Data cleanup

Specific considerations of data to be analyzed

Clustering of data into cell types

Data visualization

Inspection of clusters and giving them names

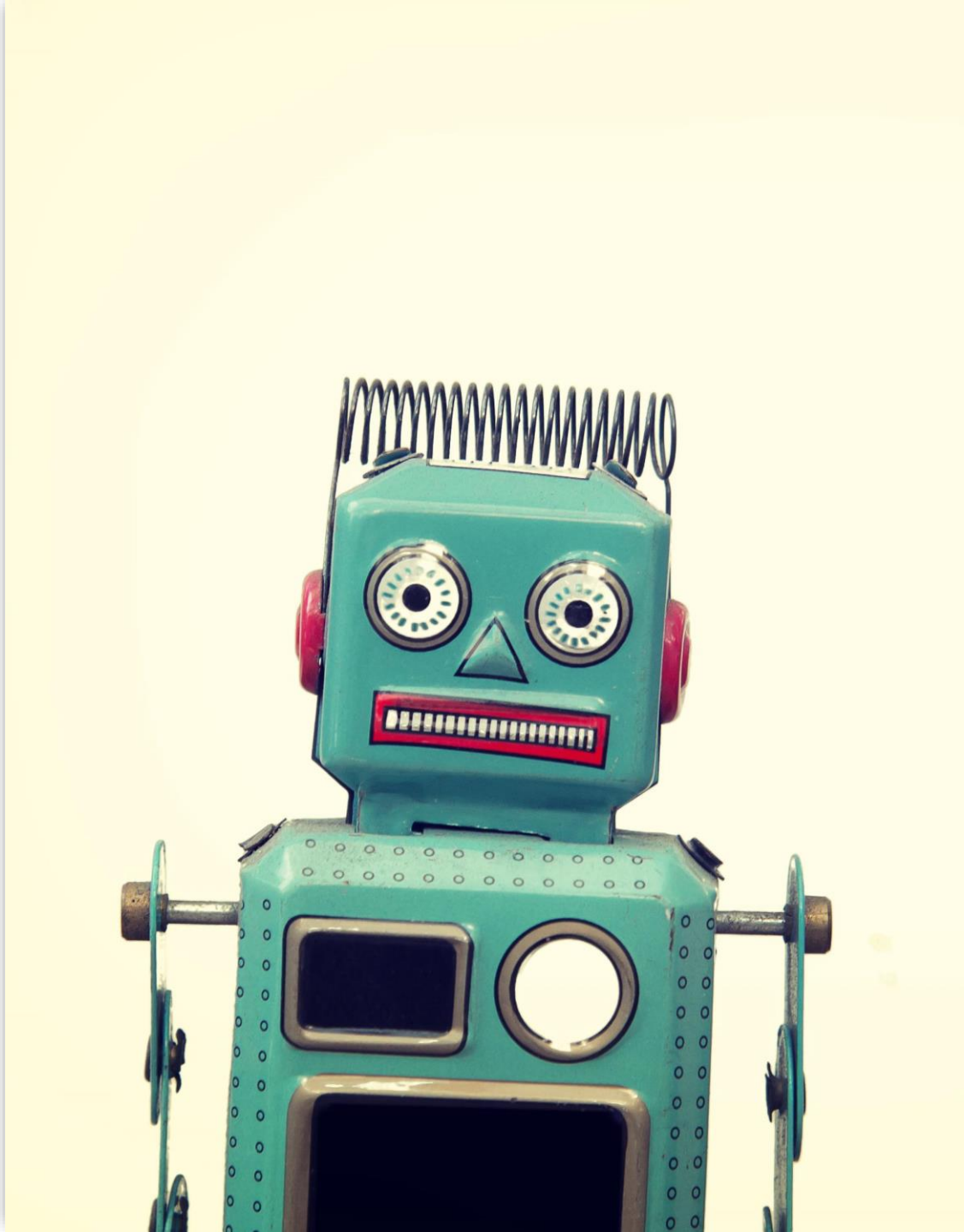
Further downstream analysis…

Clustering algorithms help identify cell populations in a less subjective way than gating

- Clustering helps identify groups of cells that are similar to each other.
  - Clustering algorithms can "see" all the cells' features at once; it’s not limited by 2D hierarchical gating
  - Can identify unexpected clusters that might be missed by usual gating strategy

Many clustering algorithms exist

- flowMeans, FlowSOM, PhenoGraph, SPADE3, SWIFT, DBSCN, HDBSCN, MegaClust, X-Shift, ADICyt, SamSPECTRAL, FLOCK, FLAME, FlowDensity, Accense, DEPECHE, kmeans, LDA, ACDC, Flock2, etc., etc., etc.
- Supervised vs. unsupervised vs. semi-supervised
- How to choose a clustering algorithm?
  - Accurate and reproducible
    - Similar cell populations are found in different specimens
  - Meets the needs of the problem at hand
  - Are others using it?



Comparisons of clustering algorithms

- FlowCAP I challenge compared unsupervised clustering algorithms
  - Great challenge but lacked high-dimensional data

- Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A. 2016 Dec;89(12):1084-1096. PMID: 27992111.
  - Compared 18 clustering algorithms
  - Used 6 well documented/gated data sets
  - Evaluated ability to identify major cell populations and single rare cell population, based on expert gating for comparison
  - Excluded doublets, debris, and dead cells and performed asinh transformation on data
  - Used default algorithm parameters where available, and aimed for 40 clusters when user input was needed

Results from Weber et al.

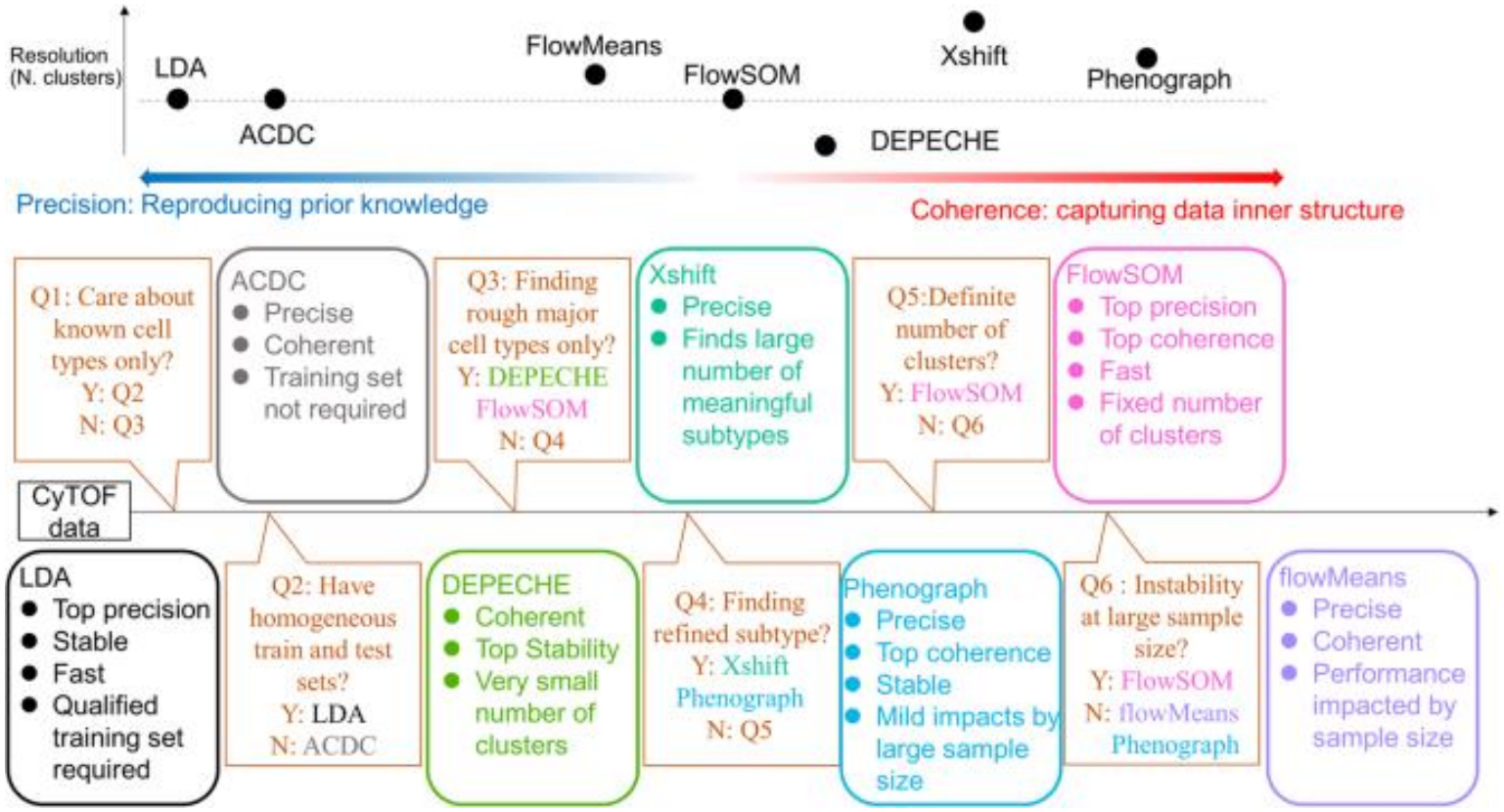
	MULTIPLE POPULATIONS OF INTEREST								SINGLE RARE POPULATION OF INTEREST			
	LEVINE_32DIM		LEVINE_13DIM		SAMUSIK_01		SAMUSIK_ALL		NILSSON_RARE		MOSMANN_RARE	
	MEAN F1	RUNTIME HH:MM:SS	MEAN F1	RUNTIME HH:MM:SS	MEAN F1	RUNTIME HH:MM:SS	MEAN F1	RUNTIME HH:MM:SS	F1	RUNTIME HH:MM:SS	F1	RUNTIME HH:MM:SS
ACCENSE	0.494	00:05:32	0.358	00:04:48	0.517	00:06:21	0.502	00:05:32	0.445	00:06:11	0.021	00:04:37
ClusterX	**0.682**	**01:57:02**	**0.474**	**03:50:51**	0.571	01:52:09	0.603	02:02:08	0.132	00:29:00	0.004	01:56:13
DensVM	0.660	08:30:13	0.448	08:11:09	0.239	07:34:49	0.496	07:55:14	0.153	03:19:36	0.004	07:55:34
FLOCK	**0.727**	**00:03:43**	0.379	00:00:29	0.608	00:00:35	**0.631**	**00:14:28**	0.089	00:00:08	0.102	00:01:06
flowClust	NA	NA	0.416	02:59:27	0.612	06:04:13	0.610	11:56:58	**0.461**	**04:20:24**	0.080	03:32:41
flowMeans	**0.769**	**02:34:01**	**0.518***	**00:04:09**	0.625	04:13:12	**0.653**	**02:03:17**	**0.488**	**00:01:06**	0.104	00:03:57
flowMerge	NA	NA	0.247	07:45:41	0.452	09:56:25	0.341	03:21:40	0.111	09:41:02	0.159	11:06:45
flowPeaks	0.237	00:05:19	0.215	00:00:21	0.058	00:07:05	0.323	00:16:39	0.016	00:00:08	0.001	00:02:18
FlowSOM	**0.780***	**00:00:41**	**0.495**	**00:00:15**	**0.707***	**00:00:19**	**0.702***	**00:02:13**	**0.447**	**00:00:08**	**0.665**	**00:02:14**
FlowSOM_pre	0.502	00:00:35	0.422	00:00:10	0.583	00:00:14	0.528	00:02:08	**0.447**	**00:00:03**	**0.665**	**00:01:32**
immunoClust	0.413	03:20:51	0.308	02:57:27	0.552	01:35:10	0.523	02:06:40	0.371	00:06:57	0.563	01:51:23
k-means	0.420	00:00:13	0.435	00:00:02	**0.650**	**00:00:05**	0.590	00:00:26	0.243	00:00:01	0.103	00:00:11
PhenoGraph	0.563	00:37:00	**0.468**	**00:12:09**	**0.671**	**00:05:55**	**0.653**	**05:30:35**	0.229	00:01:58	0.498	00:43:43
Rclusterpp	0.605	01:13:04	0.465	00:17:54	**0.637**	**00:08:32**	0.613	00:14:05	0.360	00:00:17	**0.737**	**02:12:32**
SamSPECTRAL	0.512	04:24:05	0.253	00:24:01	0.263	00:34:42	0.138	00:39:26	0.088	00:01:52	**0.618**	**03:42:28**
SPADE	NA	NA	0.127	00:04:46	0.169	00:03:02	0.130	00:53:39	0.180	00:00:52	0.027	00:12:12
SWIFT	0.177	02:27:39	0.179	01:07:03	0.202	02:19:30	0.208	02:50:08	0.390	00:11:26	0.484	00:34:34
X-shift	**0.691**	**04:45:26**	**0.470**	**00:48:17**	**0.679**	**00:24:54**	**0.657**	**03:48:27**	**0.531***	**00:04:37**	**0.802***	**03:18:20**

Results show the mean F1 score for data sets with multiple cell populations of interest, and F1 score for data sets with a single rare cell population of interest; as well as runtimes. For each data set, the best-performing method is indicated with a star (*), and the top five methods are displayed in bold. Runtimes are not precisely comparable between methods due to differences in subsampling, number of processor cores, and hardware specifications (Supporting Information Tables S1 and S4); however they are included in order to provide users with information about order-of-magnitude differences. NA = not available, due to errors or non-completion (Supporting Information Table S1).

Weber LM, Robinson MD. Cytometry A. 2016 Dec;89(12):1084-1096. PMID: 27992111

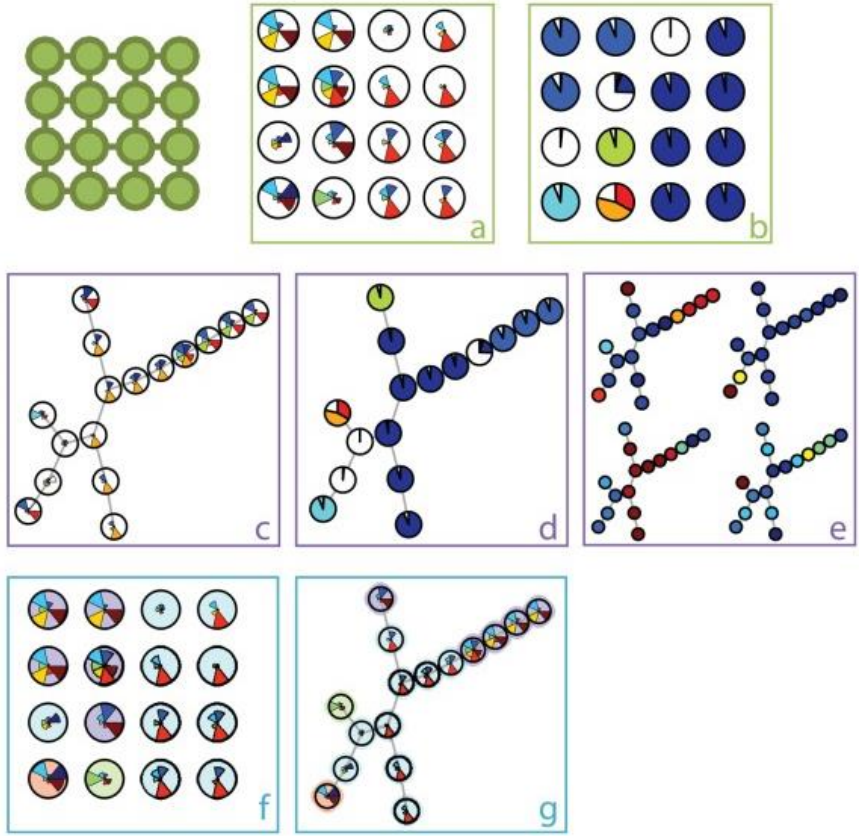
Another comparison study

- Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, Ding X. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biol. 2019 Dec 23;20(1):297. PMID: 31870419.
  - Seven unsupervised methods (Accense, Xshift, PhenoGraph, FlowSOM, flowMeans, DEPECHE, and kmeans) and two semi-supervised methods (Automated Cell-type Discovery and Classification and linear discriminant analysis (LDA)) tested on six mass cytometry datasets.
  - FlowSOM and Phenograph were deemed the top performing unsupervised clustering methods.



FlowSOM:

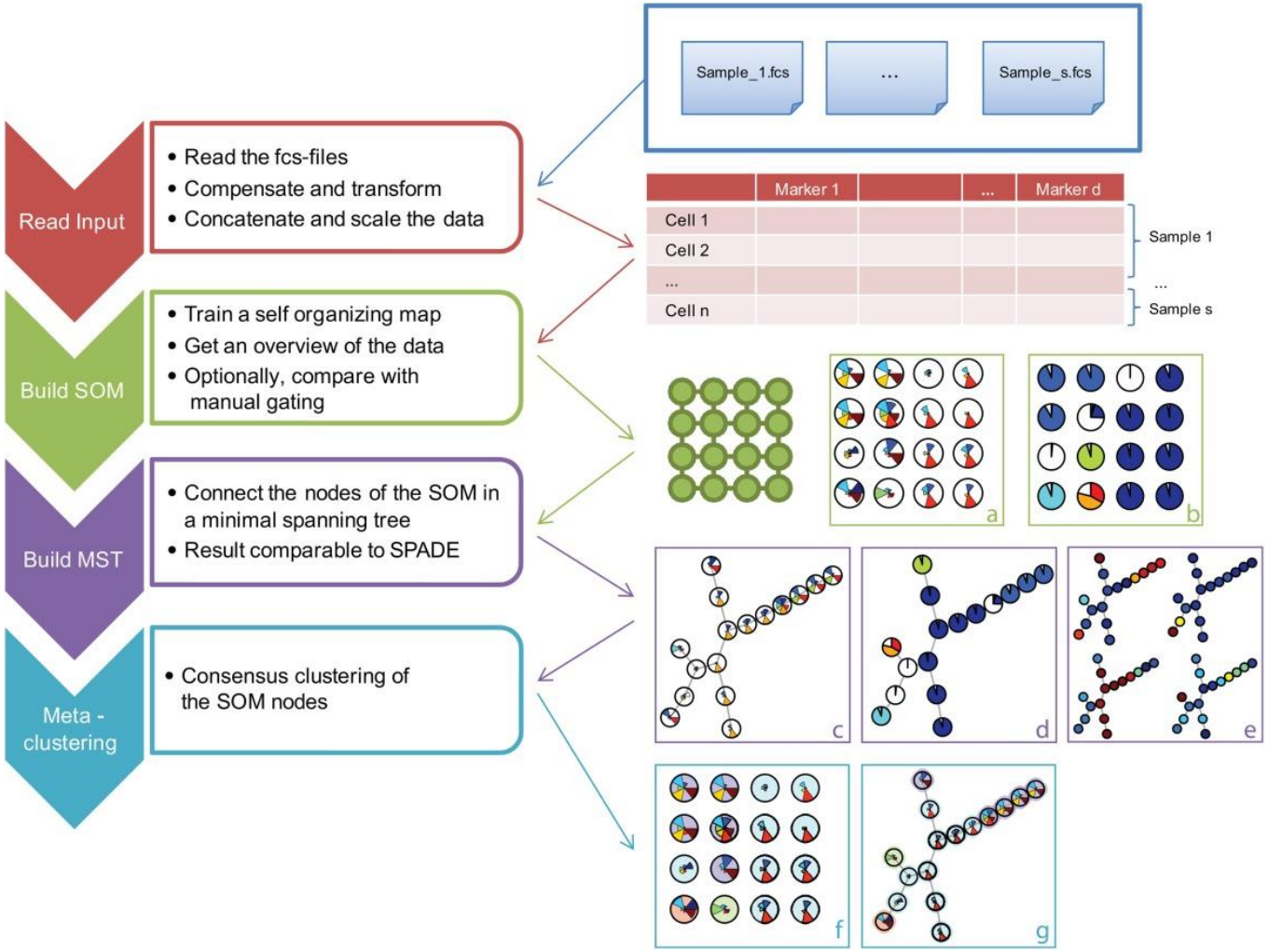
- Introduced in 2015 (Van Gassen et al.)
- Finds clusters in an unsupervised way
- Software package does clustering *and* visualization
- Cluster types can be applied to new cases
- Computationally fast
- Can be run on most computers
- Widely adopted (cited by >1260 papers)
- Disadvantages
  - Might miss very small populations
  - Uses lots of computer memory



Cytometry Pt A, Volume: 87, Issue: 7, Pages: 636-645, First published: 08 January 2015, DOI: (10.1002/cyto.a.22625)

How does FlowSOM work?

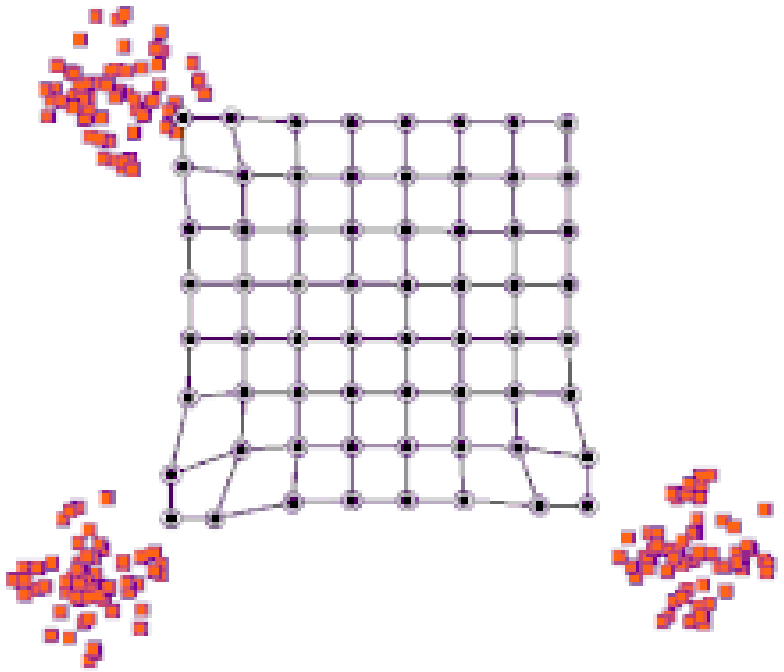
- Creates a self-organizing map (SOM)
- Creates a minimal spanning tree graph (mostly for visualization)
- Applies a “consensus clustering” algorithm to organize the nodes into larger clusters



Cytometry Pt A, Volume: 87, Issue: 7, Pages: 636-645, First published: 08 January 2015, DOI: (10.1002/cyto.a.22625)

Generating the self-organizing map (SOM)

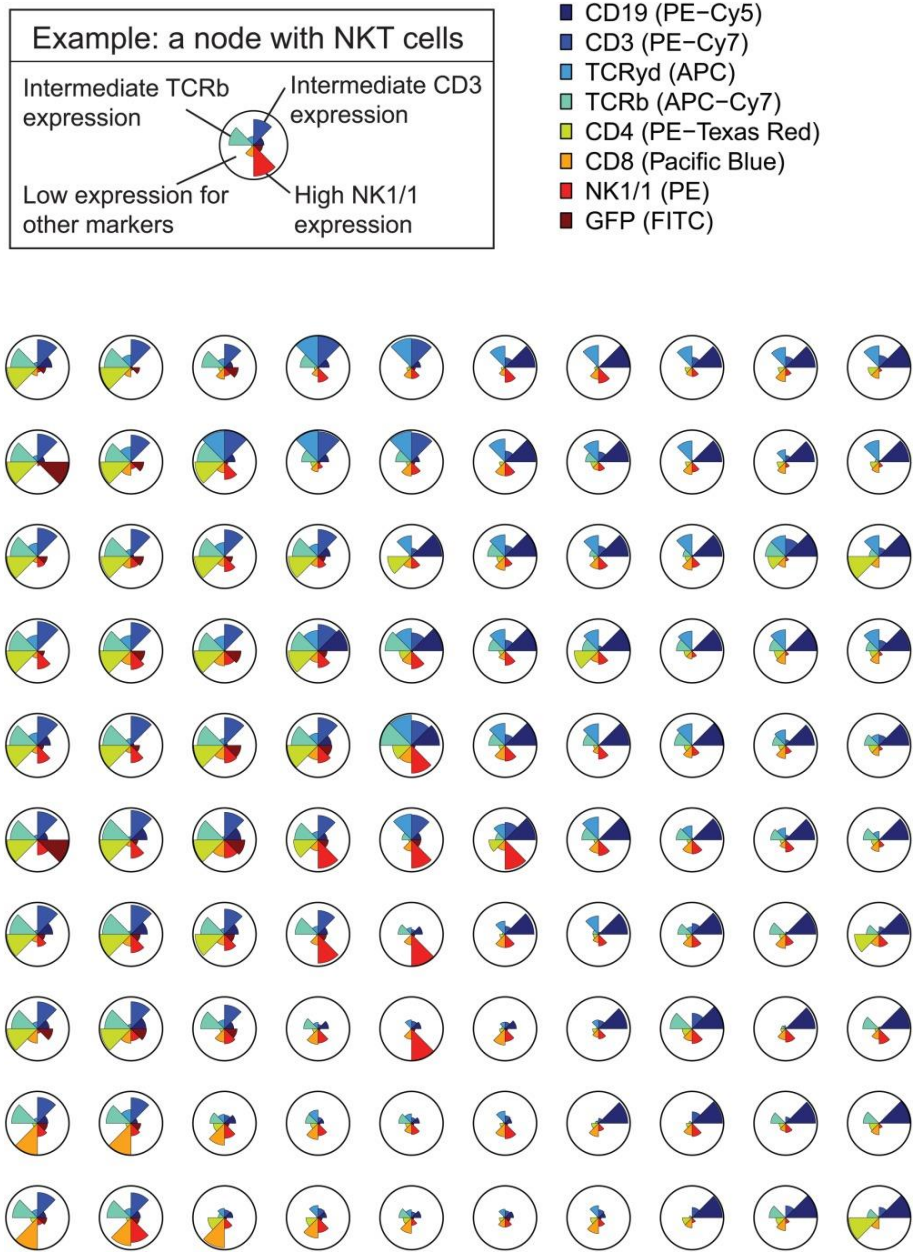
- The map consists of “nodes” that are iteratively moved around until the clusters of similar cells are mapped out.
- The number of nodes is chosen to be greater than the number of real clusters we expect to find (nodes are grouped into clusters in the final step).
- The greater the number of nodes, the greater the “purity” of cells in a node.
- More nodes are needed to be able to identify small populations.



https://en.wikipedia.org/wiki/Self-organizing_map

Plots generated using the SOMs can give insight into the heterogeneity of the data

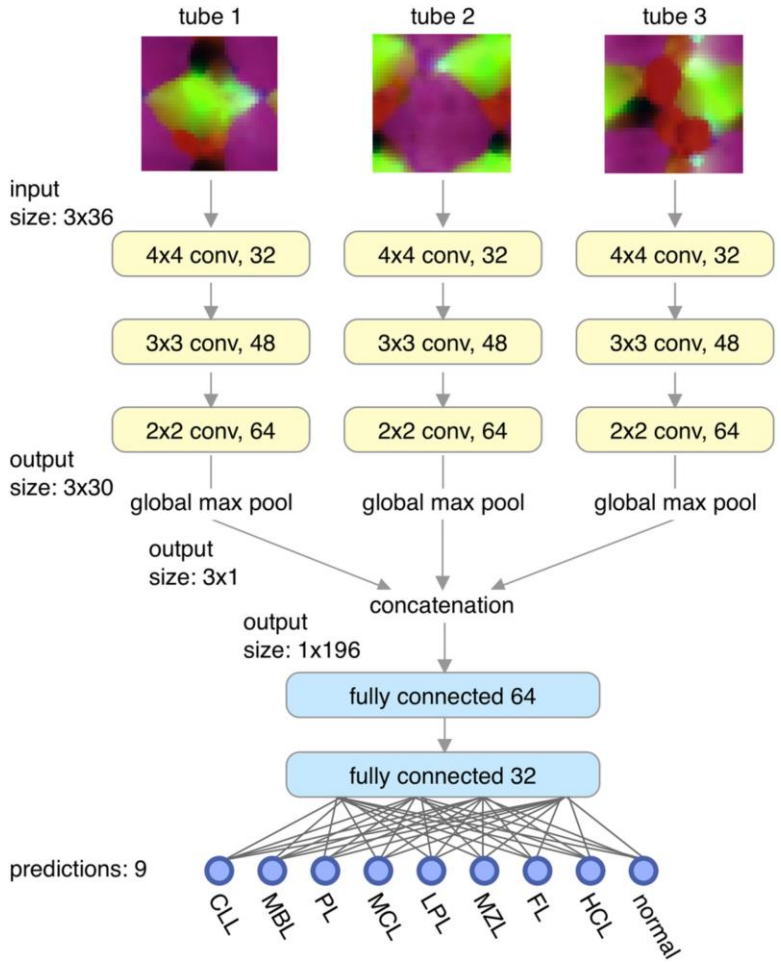
- "Star charts" demonstrate the relative marker expression intensity of each node.
- Heterogeneity in nodes can prompt closer inspection (e.g., with standard 2D plots)
- Other plots are also possible (labels found by traditional gating, relative number of cells in each node, etc.)



Cytometry Pt A, Volume: 87, Issue: 7, Pages: 636-645, First published: 08 January 2015, DOI: (10.1002/cyto.a.22625)

Another use for SOM nodes data: supervised machine learning

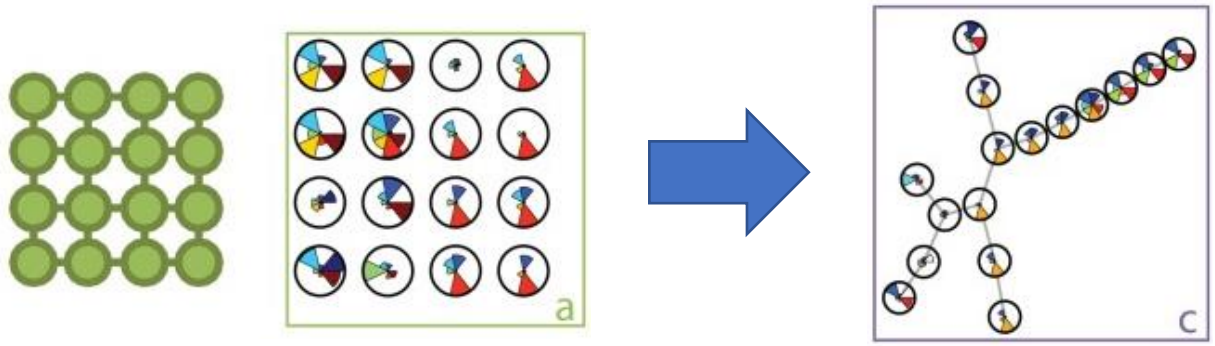
- Supervised machine learning generally requires reducing flow data to population level representations, like FlowSOM data, that can be passed to a classifier (CNN, random forest, etc.)
- Examples:
  - Identifying B cell neoplasms by machine learning (Zhao M et al. Cytometry A. 2020 Oct;97(10):1073-1080)
  - Identifying MDS (Duetz C et al. Cytometry A. 2021 Aug;99(8):814-824.)



Zhao M et al. Cytometry A. 2020 Oct;97(10):1073-1080

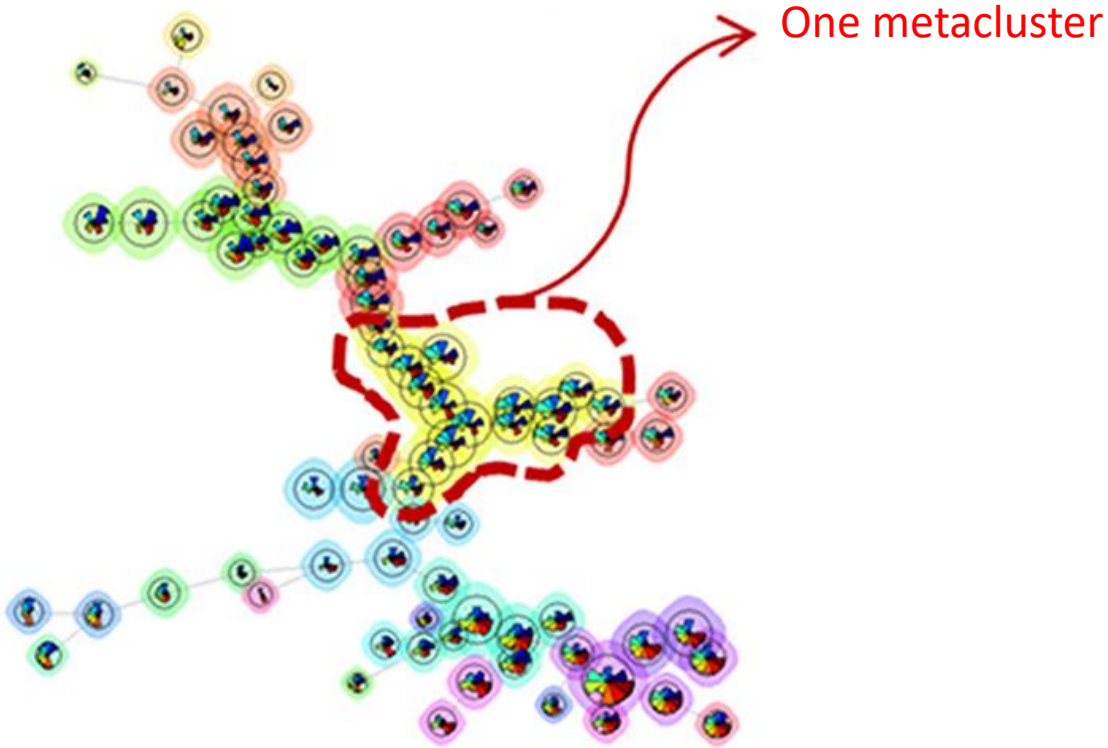
Minimal spanning trees provide another way to visualize how nodes relate to each other

- Nodes that are most like one another are linked to each other.
- Loops are not allowed.

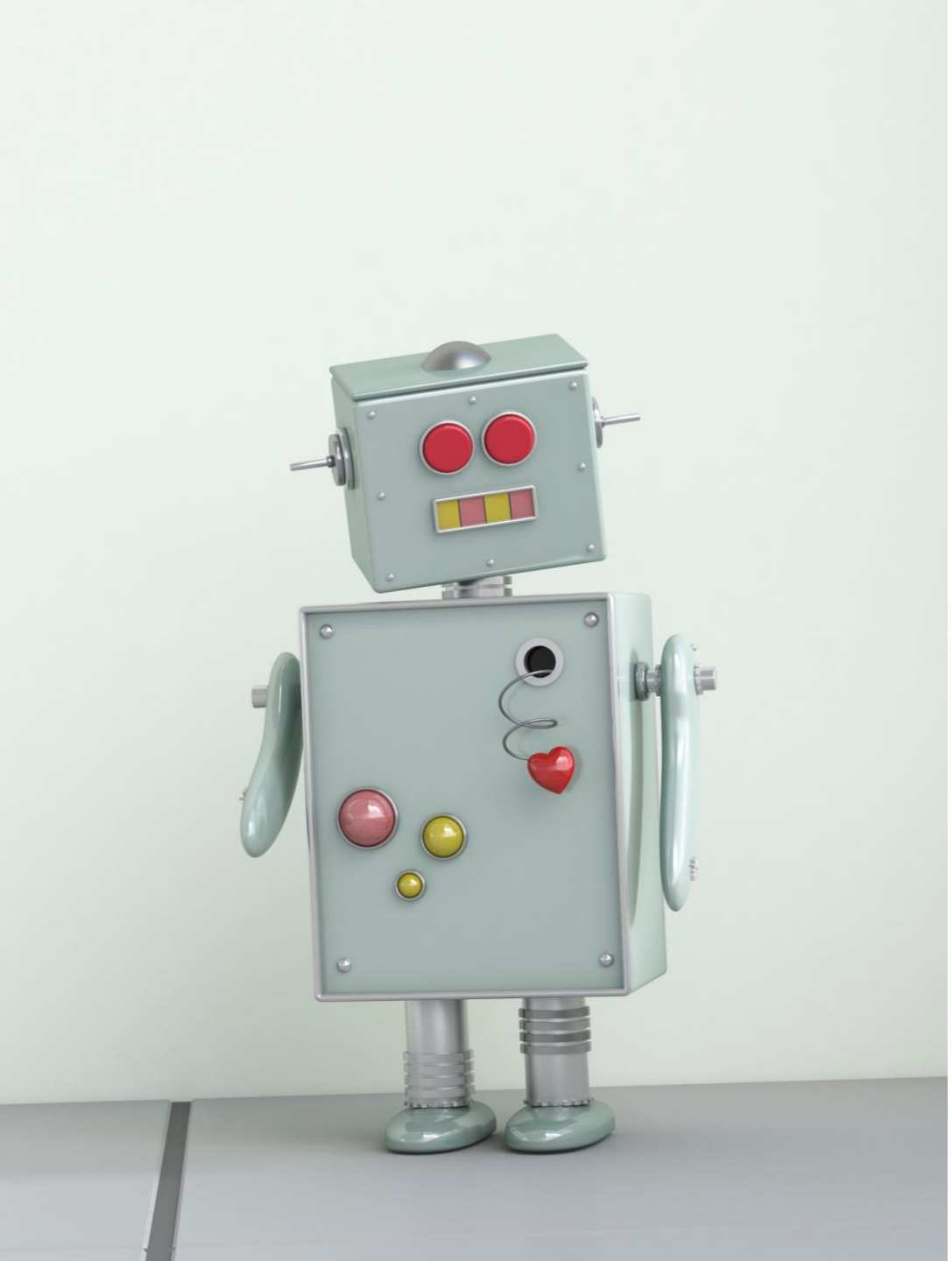


Nodes are grouped into larger clusters (or “metaclusters”)

- The nodes themselves are grouped into metaclusters using a **consensus hierarchical clustering algorithm**.
- Marker expression patterns can be inspected to give names to clusters (e.g., neutrophil, eosinophil).



Duetz C et al. Cytometry A. 2021 Aug;99(8):814-824

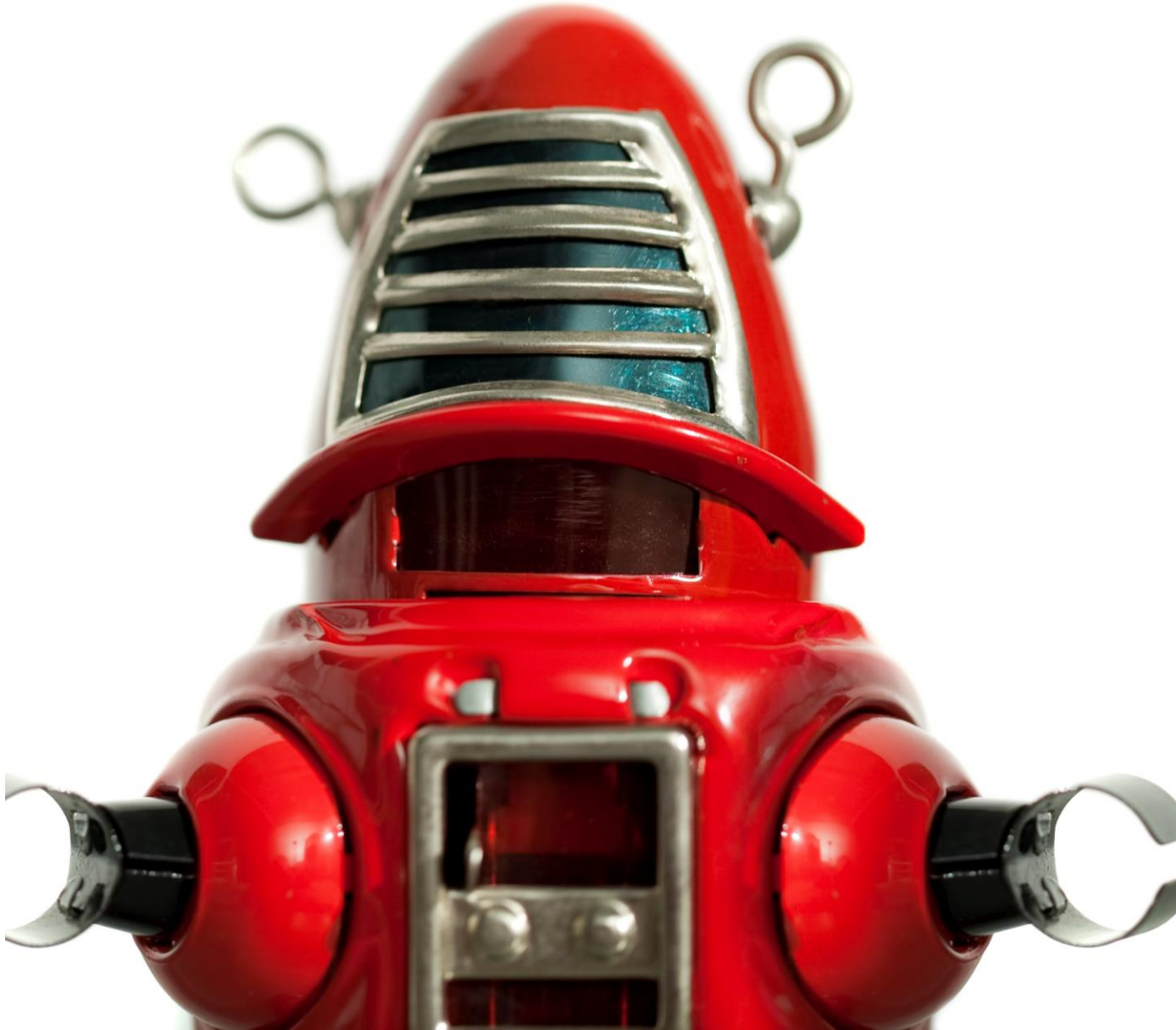


Practical considerations in applying FlowSOM

- Optimize the preanalytical variables
  - Minimize batch-to-batch variability
  - Use calibration controls
- Preprocessing data
  - Remove non-viable cells, doublets, etc.
  - Apply compensations.
  - Transform data using logicle, asinh, etc.
- Have enough (and the right kind of) data to represent the full range of immunophenotypes
  - Consider combining data files from different batches
- Computation is not instantaneous
  - Development of a software pipeline or use of a commercial package can help.
- Consider running the algorithm multiple times.
- Good place to get started: Quintelier K et. Analyzing high-dimensional cytometry data using FlowSOM. Nat Protoc. 2021 Aug;16(8):3775-3801. PMID: 34172973.

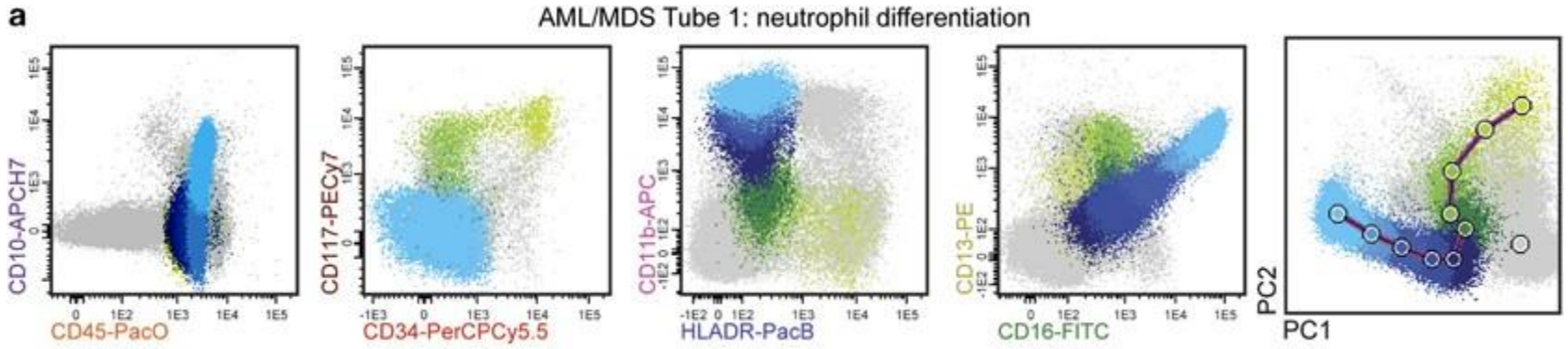
How does one know the clusters are real?

- Options
  - ~~Blindly trust the clustering algorithm~~
  - Try multiple clustering algorithms to see whether the same clusters are recurrently found
  - Try re-running the clustering algorithm (with a different random number seed)
  - Visually inspect the clustering using standard 2x2 plots
  - Apply dimensionality reduction algorithms to visualize (more to follow)



Dimensionality reduction can help in visualizing the overall data distribution

- For high-dimensional flow data, this can help us get the big picture without all the 2x2 scatter plots.
- Dimensionality reduction maps the data to a lower dimensionality (usual two-dimensions for plotting) embedding, manifold, or topology.
- Popular dimensionality reduction algorithms:
  - PCA (principal component analysis)
  - t-SNE (t-distributed stochastic neighbor embedding)
  - UMAP (uniform manifold approximation and projection)
- Dimensionality reduction does not necessary result in clusters.



van Dongen JJ et al. Leukemia. 2012 Sep;26(9):1908-75. PMID: 22552007.

What is t-SNE?

- Developed in 2008.
- Maps high-dimensional data to clusters in two-dimensions
  - Calculates probability distributions of cells being close to each other in high-dimensional space.
  - It then tries to distribute cells in 2D space by moving cells until similar probability distributions are achieved.
- Dissimilar clusters are (generally) farther apart
- Available in some commercial flow cytometry software

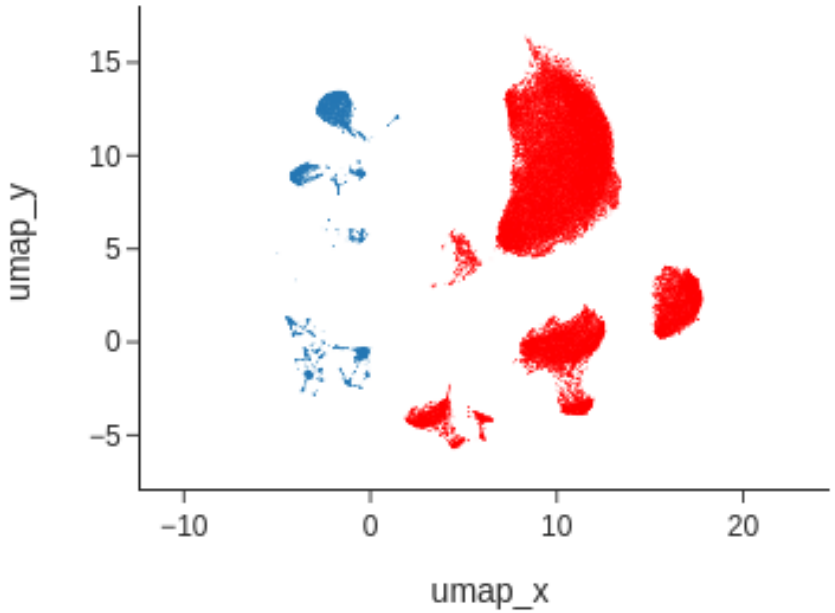
https://www.oreilly.com/content/an-illustrated-introduction-to-the-t-sne-algorithm/

UMAP

- Reduces high-dimensional data to two-dimensional representations
- Better preserves relationships between cells and clusters
- "Embeddings" can be saved and used again with data from new samples
  --> cell populations will show up in the expected locations

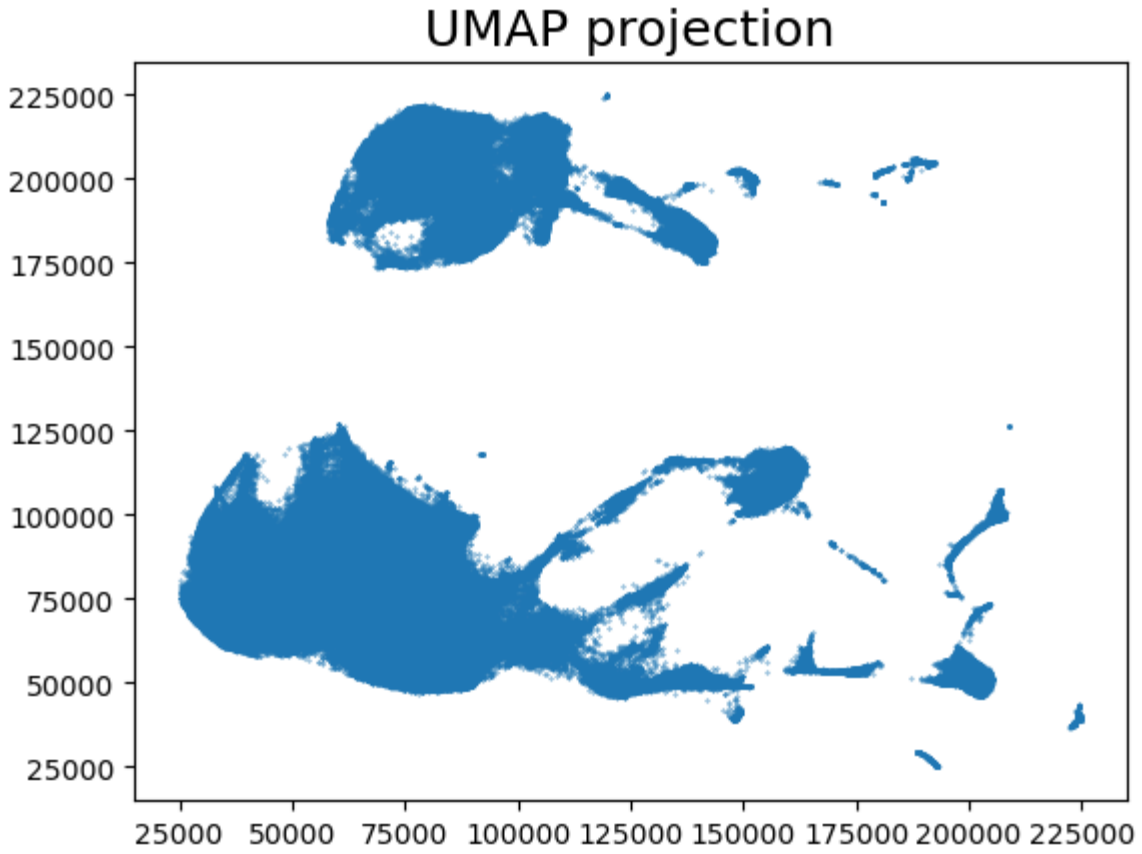
- Recommended tutorial: https://umap-learn.readthedocs.io/en/latest/basic_usage.html

Bone marrow involved by CLL
(86% of cellularity)



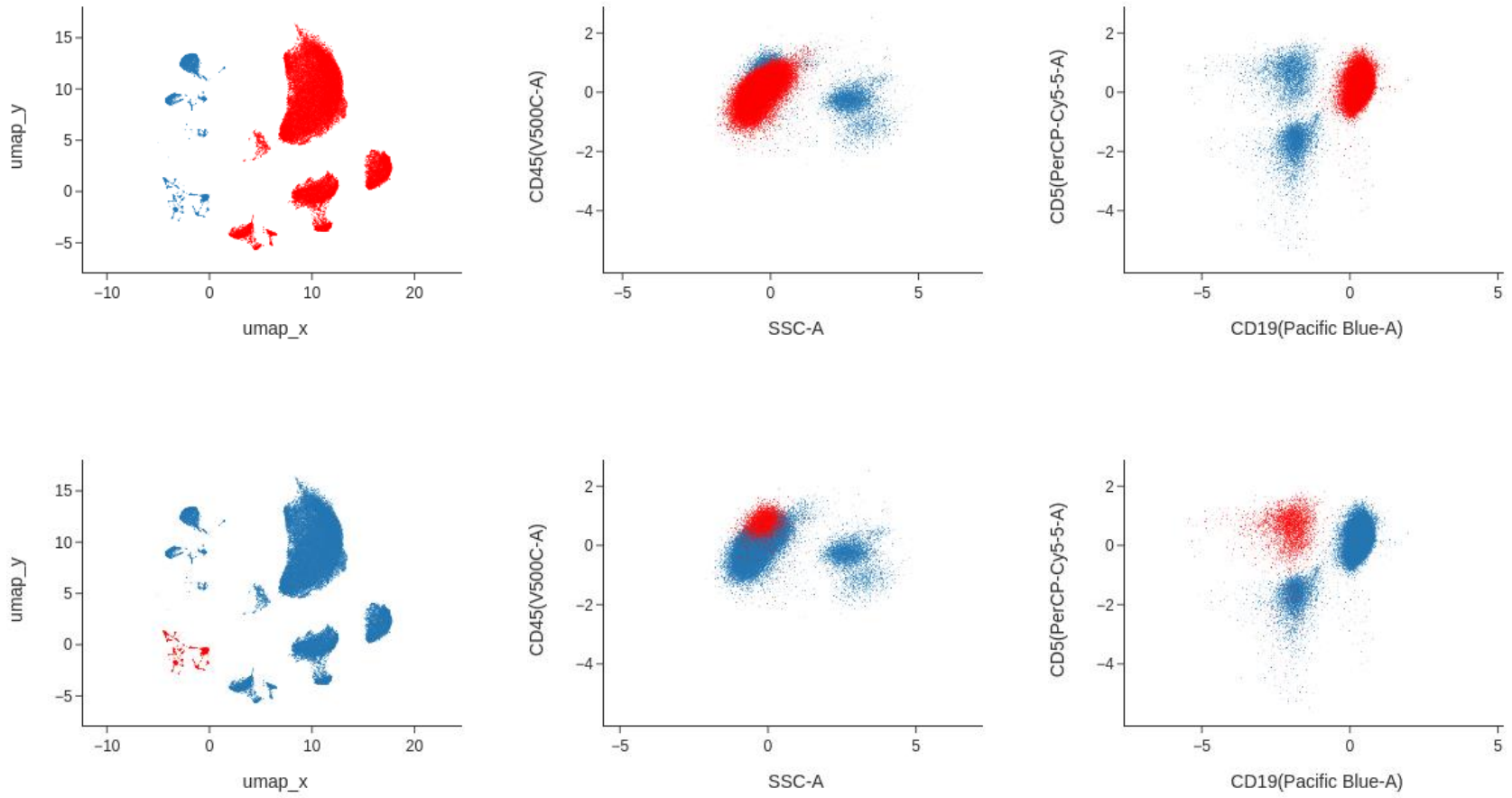
UMAP: Pros and cons

- Pros:
  - Can apply the same manifold to new cases, out-of-the-box
  - Can represent large-scale relationships between data somewhat better than t-SNE
- Cons
  - Plots data along a manifold, not necessarily in clusters

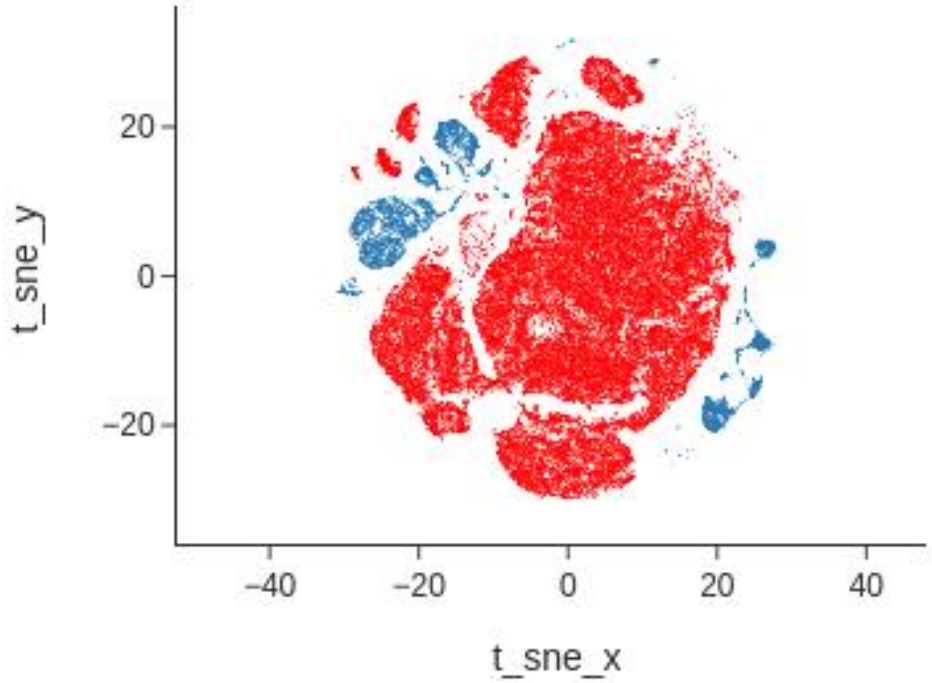


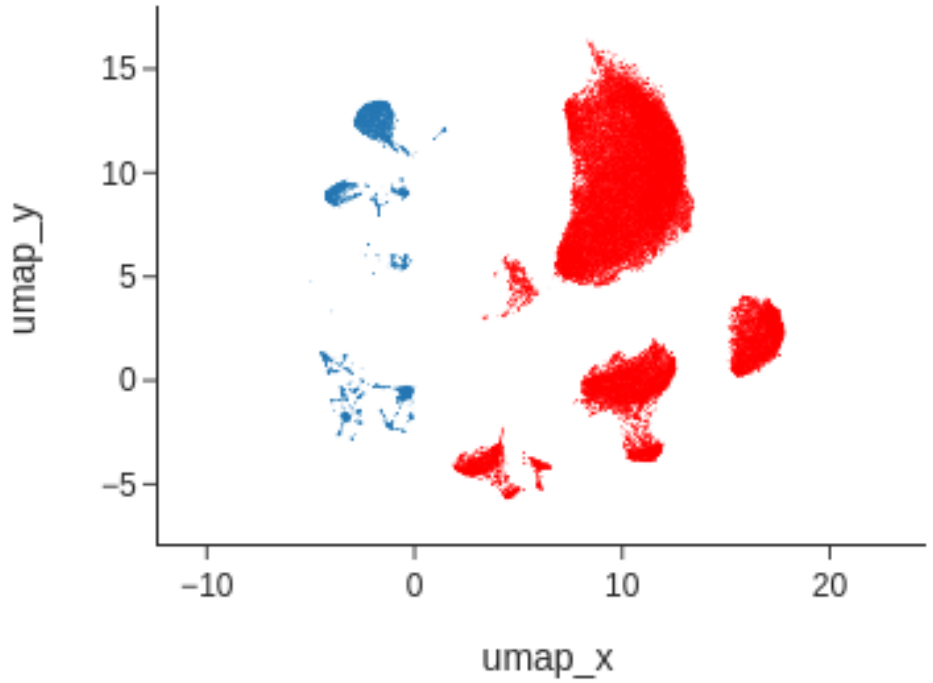
Example gating using UMAP

- UMAP calculated using*:
  - FSC-H
  - SSC-H
  - sKappa
  - sLambda
  - CD5
  - CD23
  - CD10
  - CD20
  - CD19
  - CD45



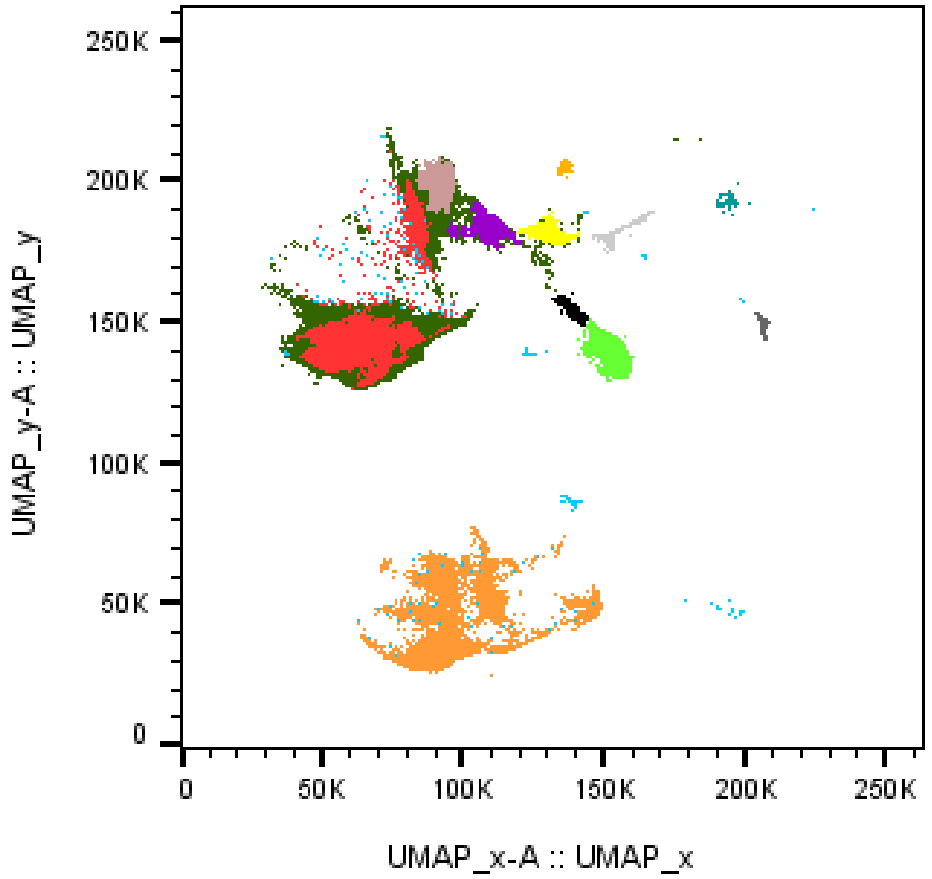
Comparing t-SNE and UMAP



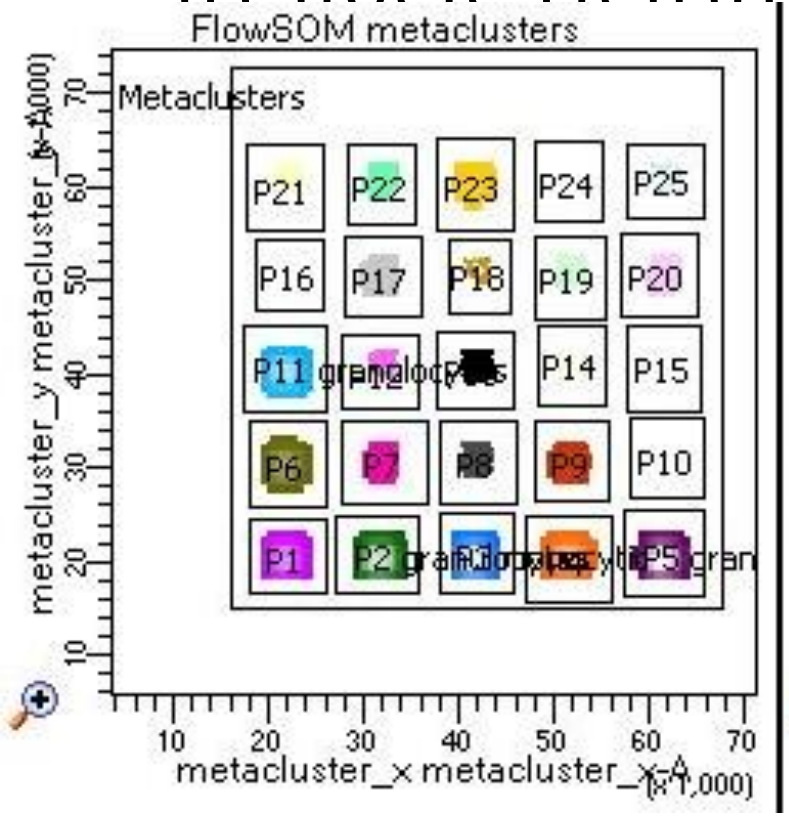


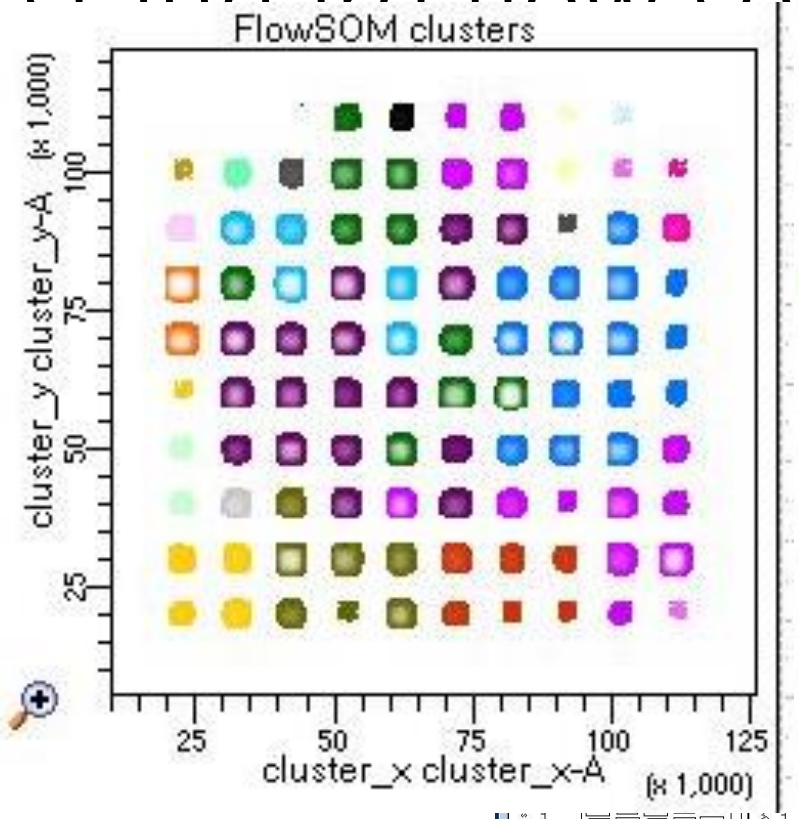
Basic implementation

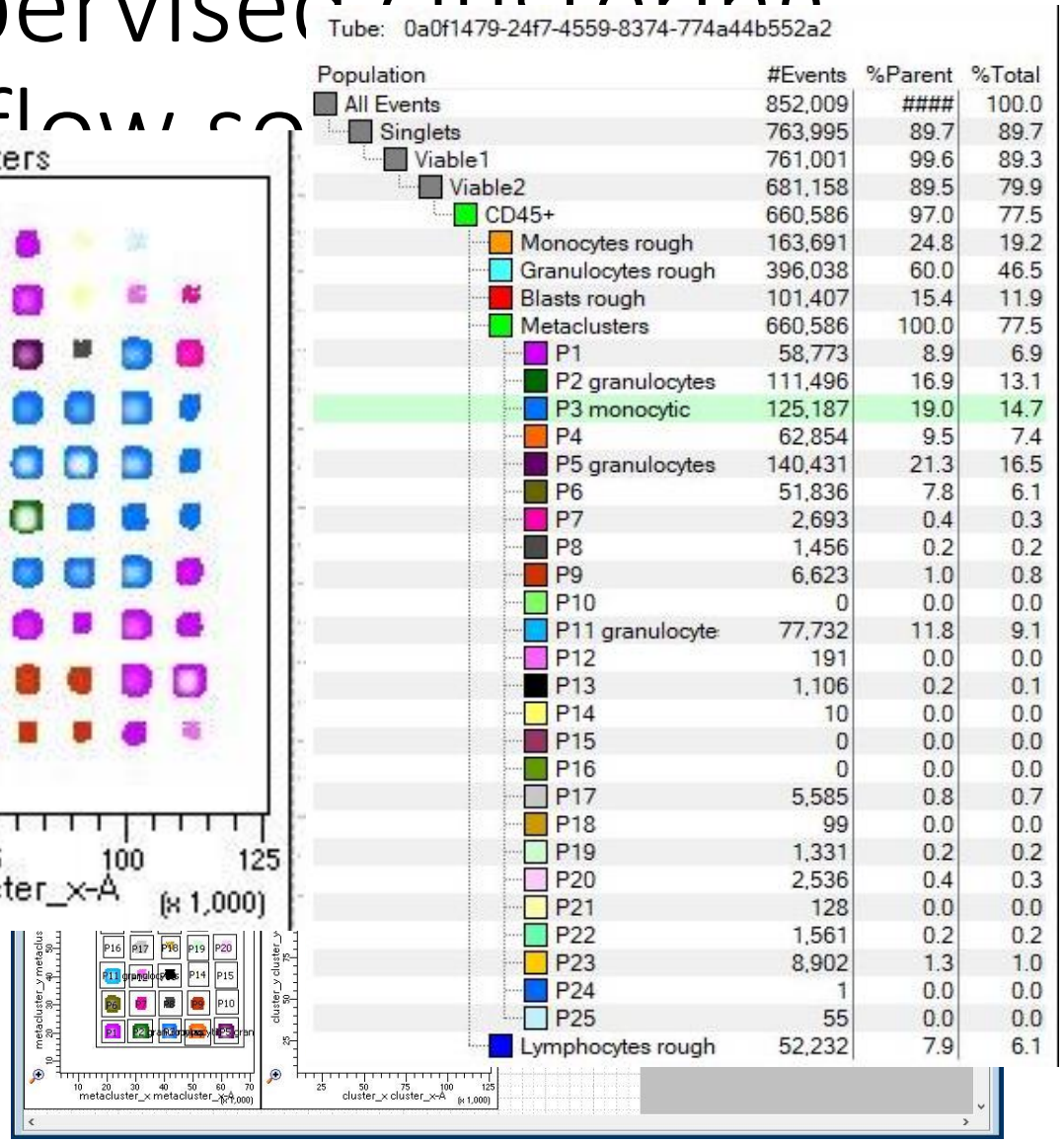
- Computer on local network looks for new FCS files on file server.
- With new FCS file, run UMAP and clustering algorithms.
- Create new FCS with UMAP and/or t-SNE coordinates and cluster labels added as additional channels.
- Compare clustering and embedding/manifold with standard software and gating



Plotting results of unsupervised clustering analysis using standard flow so







https://github.com/SimonsonLab/add-labels-to-fcs

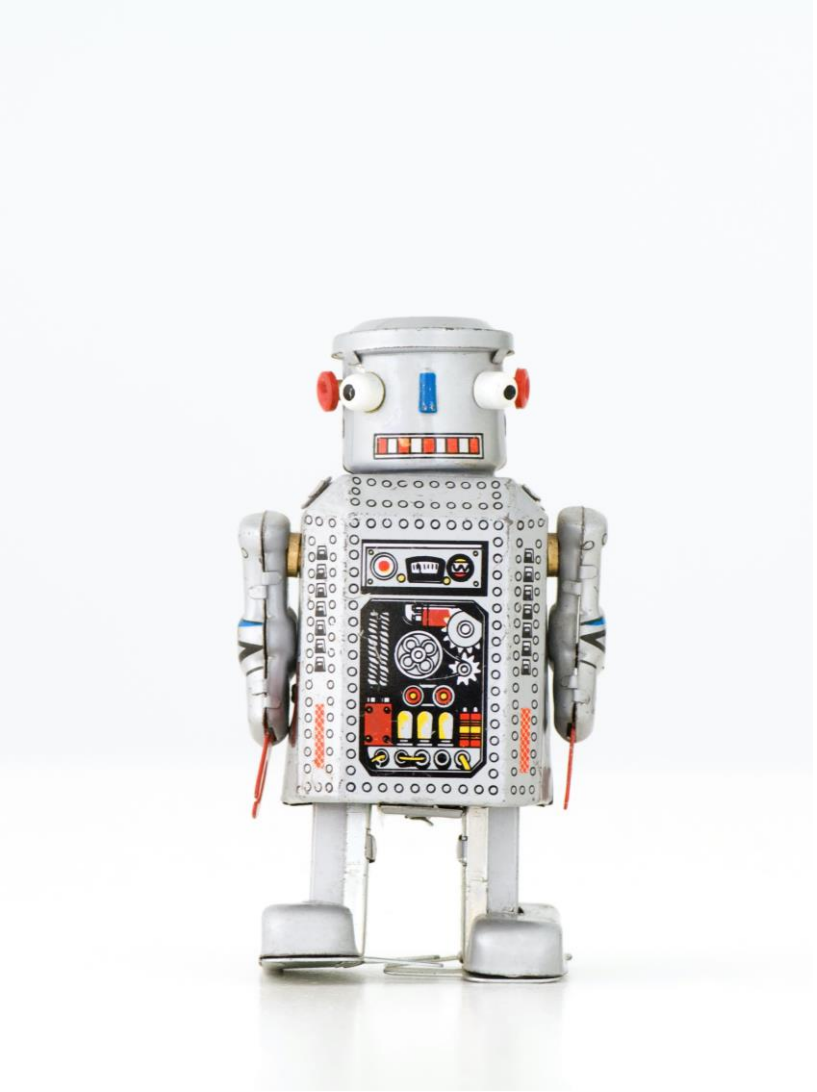
Example application: 18 color flow cytometry to evaluate T cell subsets

- Your lab has purchased a new 18-color flow cytometer
- You now want to offer a new T cell panel (22 antibodies) for immunomonitoring in clinical trials, and, eventually, clinical use.
- You would like to be able to parse the cells into T cell subsets for identification and quantification
  - Minimize subjectivity
  - Include the ability to identify unexpected subsets
- **You have decided to employ computational methods in addition to traditional gating to help in the analysis.**

Additional considerations

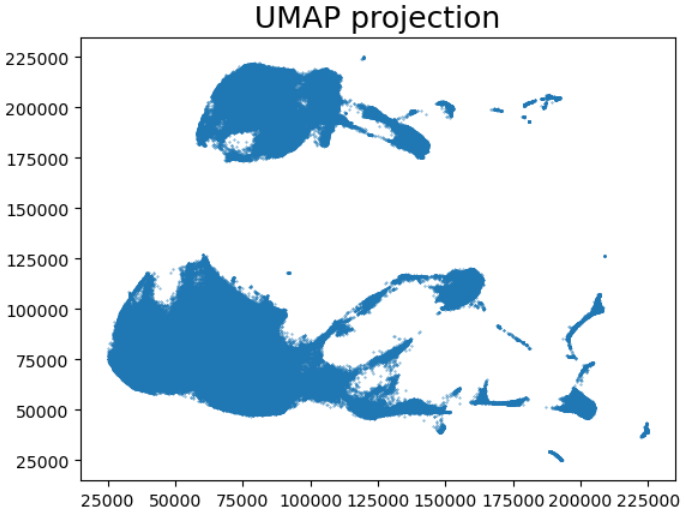
- Comparing cases
  - Combine cases into one data set and create embeddings
  - Create embeddings and apply to additional cases
- How many cells do I really need?
- What kind of computer power do I need?
- Should I hire a data scientist?
- Establish a pipeline
- Additional software packages
  - Bioconductor
  - Scanpy
  - Seurat
  - pathML
- Who will sign the report?

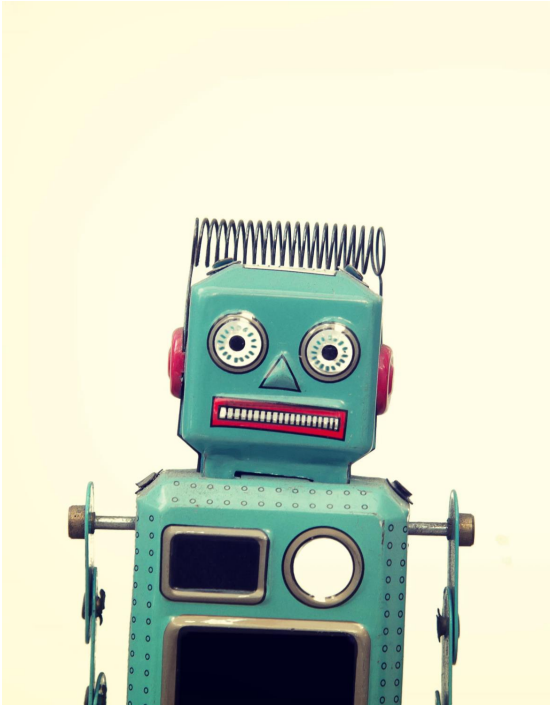
PathML

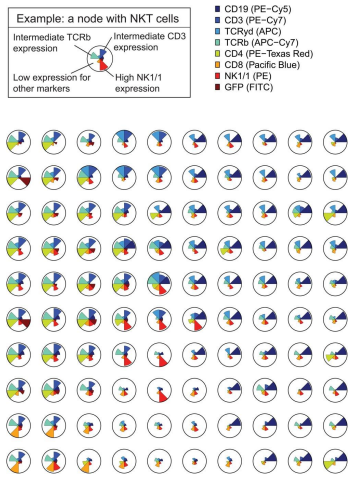


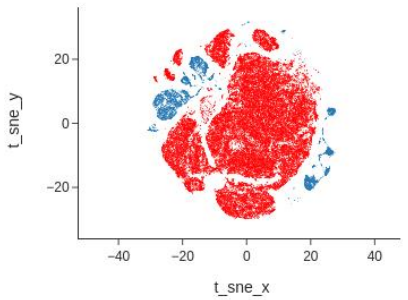
Summary

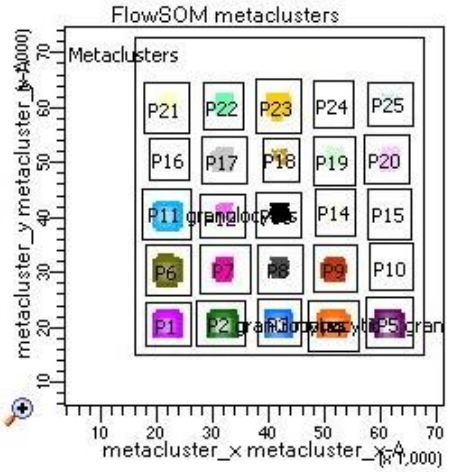
- High-dimensional flow cytometry is becoming more commonplace and presents challenges for analysis by standard gating.
- Clustering algorithms, like FlowSOM, can help detect cell clusters in an unsupervised, less biased manner.
- Dimensionality reduction algorithms, including t-SNE and UMAP, help in visualizing the overall distribution and heterogeneity of cells within two-dimensional plots.
- Evaluation of immunophenotypes is important for verifying and labeling clustering, which can be done using visualization software (including standard flow cytometry software).











References

- Aghaeepour N, Finak G; FlowCAP Consortium; DREAM Consortium; Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods. 2013 Mar;10(3):228-38. doi: 10.1038/nmeth.2365. Epub 2013 Feb 10. Erratum in: Nat Methods. 2013 May;10(5):445. PMID: 23396282; PMCID: PMC3906045. *FlowCAP I Challenge results.*
- *Baumgaertner P, Sankar M, Herrera F, Benedetti F, Barras D, Thierry AC, Dangaj D, Kandalaft LE, Coukos G, Xenarios I, Guex N, Harari A. Unsupervised Analysis of Flow Cytometry Data in a Clinical Setting Captures Cell Diversity and Allows Population Discovery. Front Immunol. 2021 Apr 30;12:633910. doi: 10.3389/fimmu.2021.633910. PMID: 33995353; PMCID: PMC8119773. Evaluates MegaClust algorithm on a small clinical trial cohort and demonstrates comparison with conventional gating.*
- Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science(), vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14. *Introduction of HDBSCAN algorithm.*
- *Cheung M, Campbell JJ, Thomas RJ, Braybrook J, Petzing J. Assessment of Automated Flow Cytometry Data Analysis Tools within Cell and Gene Therapy Manufacturing. Int J Mol Sci. 2022 Mar 17;23(6):3224. doi: 10.3390/ijms23063224. PMID: 35328645; PMCID: PMC8955358. Demonstrates the use of synthetic data sets to compare clustering algorithms.*
- Duetz C, Van Gassen S, Westers TM, van Spronsen MF, Bachas C, Saeys Y, van de Loosdrecht AA. Computational flow cytometry as a diagnostic tool in suspected-myelodysplastic syndromes. Cytometry A. 2021 Aug;99(8):814-824. doi: 10.1002/cyto.a.24360. Epub 2021 May 12. PMID: 33942494; PMCID: PMC8453916. *Supervised machine learning approach with FlowSOM and interpretability analysis.*
- Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, Ding X. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biol. 2019 Dec 23;20(1):297. doi: 10.1186/s13059-019-1917-7. PMID: 31870419; PMCID: PMC6929440. *A nice paper that compares several clustering methods and explores the relative strengths of each.*
- Quintelier K, Couckuyt A, Emmaneel A, Aerts J, Saeys Y, Van Gassen S. Analyzing high-dimensional cytometry data using FlowSOM. Nat Protoc. 2021 Aug;16(8):3775-3801. doi: 10.1038/s41596-021-00550-0. Epub 2021 Jun 25. PMID: 34172973. Excellent protocol article for using FlowSOM (using R).
- Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016 Jul;16(7):449-62. doi: 10.1038/nri.2016.56. Epub 2016 Jun 20. PMID: 27320317.
- Sörensen T, Baumgart S, Durek P, Grützkau A, Häupl T. immunoClust--An automated analysis pipeline for the identification of immunophenotypic signatures in high-dimensional cytometric datasets. Cytometry A. 2015 Jul;87(7):603-15. doi: 10.1002/cyto.a.22626. Epub 2015 Apr 7. PMID: 25850678.
- Van der Maaten L and Hinton G. Visualizing data using t-SNE. J Machine Learning Research. 2008; 9:2579-2605.
- van Dongen JJ, Lhermitte L, Böttcher S, Almeida J, van der Velden VH, Flores-Montero J, Rawstron A, Asnafi V, Lécrevisse Q, Lucio P, Mejstrikova E, Szczepański T, Kalina T, de Tute R, Brüggemann M, Sedek L, Cullen M, Langerak AW, Mendonça A, Macintyre E, Martin-Ayuso M, Hrusak O, Vidriales MB, Orfao A; EuroFlow Consortium (EU-FP6, LSHB-CT-2006-018708). EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia. 2012 Sep;26(9):1908-75. doi: 10.1038/leu.2012.120. Epub 2012 May 3. PMID: 22552007; PMCID: PMC3437410.
- Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, Saeys Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. 2015 Jul;87(7):636-45. doi: 10.1002/cyto.a.22625. Epub 2015 Jan 8. PMID: 25573116. *Original introduction of FlowSOM algorithm.*
- Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A. 2016 Dec;89(12):1084-1096. doi: 10.1002/cyto.a.23030. Epub 2016 Dec 19. PMID: 27992111.
- Zhao M, Mallesh N, Höllein A, Schabath R, Haferlach C, Haferlach T, Elsner F, Lüling H, Krawitz P, Kern W. Hematologist-Level Classification of Mature B-Cell Neoplasm Using Deep Learning on Multiparameter Flow Cytometry Data. Cytometry A. 2020 Oct;97(10):1073-1080. doi: 10.1002/cyto.a.24159. Epub 2020 Jun 9. PMID: 32519455.

Questions?