Analyzing Basic and High Throughput (Cytometry) Datasets

Pier Federico Gherardini, PhD
Director of Informatics
Parker Institute for Cancer Immunotherapy

PARKER INSTITUTE FOR CANCER IMMUNOTHERAPY

Overview

- Why is this data important?
- What are some technology considerations to be aware of?
- How is this data usually (under)-utilized?
- How can we fully leverage this data?

IO Drugs are delivered systemically and they have systemic effects

- Systemic immunity is required for effective immunotherapy
  - Spitzer et al. (2017). Cell 168 (3), 487-502. e15
- CPI treatment results in infiltration of new clones from the periphery into the tumor
  - Yost et al. (2019) Nat Med. Aug; 25(8): 1251–1259.
  - Wu et al. (2020) Nature 579(7798):274-278
- A single cycle of CPI induces peripheral T cell turnover and these dynamics are associated with response
  - Valpione et al. (2020) Nat Cancer. 1, 210-221
- Blood provides the opportunity for longitudinal sampling

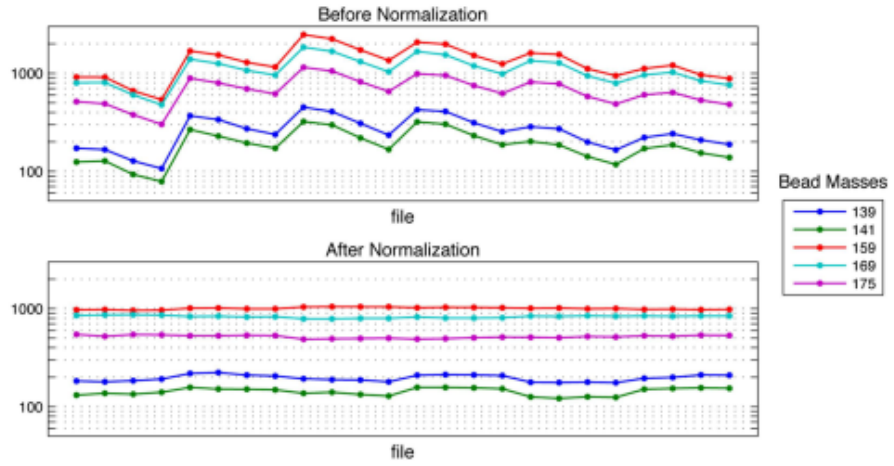
Blood analysis by Cytometry is complementary to single-cell sequencing

	Cytometry	Single-cell sequencing
Number of cells	1M+ (Only constrained by sample size)	20000, less than that if samples are multiplexed
Number of analytes	40+	~2000
Reliability of measurement	High but requires antibodies	Subjected to random-sampling (drop-outs and false negatives)
Cost per sample	Lower	Higher

Many of the concepts I will talk about apply to both kinds of data

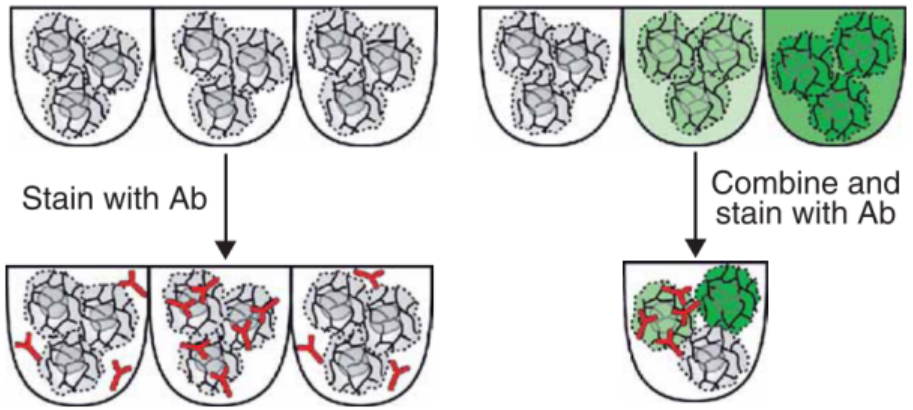
There are two main sources of variability in cytometry data

Machine sensitivity



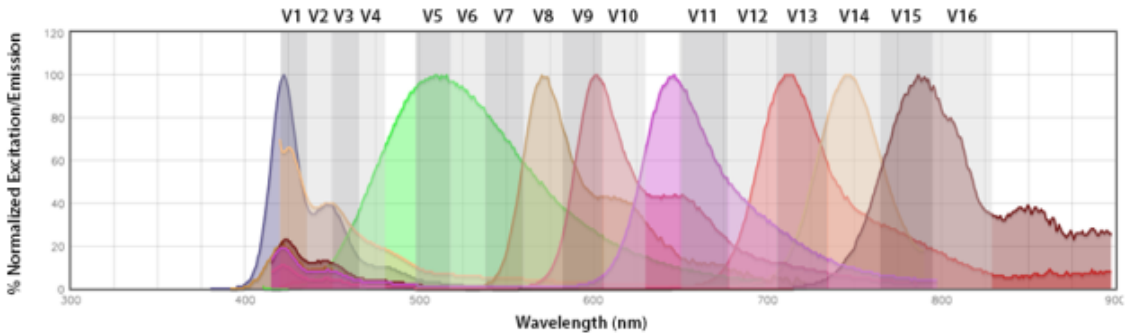
Beads provide reference synthetic standards

Antibody staining

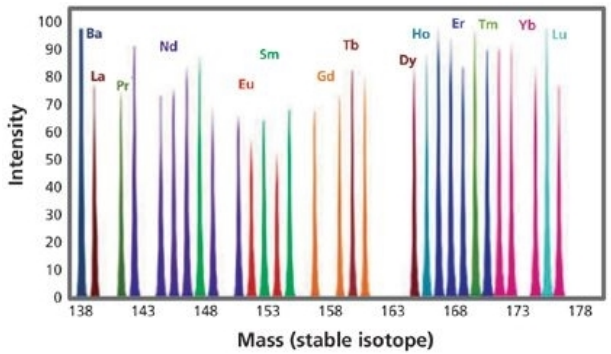


Barcoding of samples pre-staining

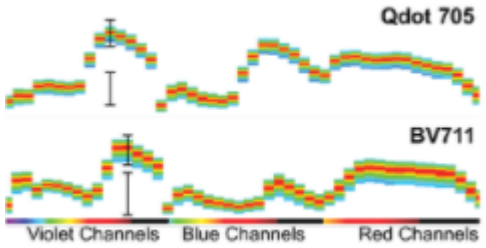
Technological advances in this field revolve around the signal detection methodology



BD – FACS Symphony

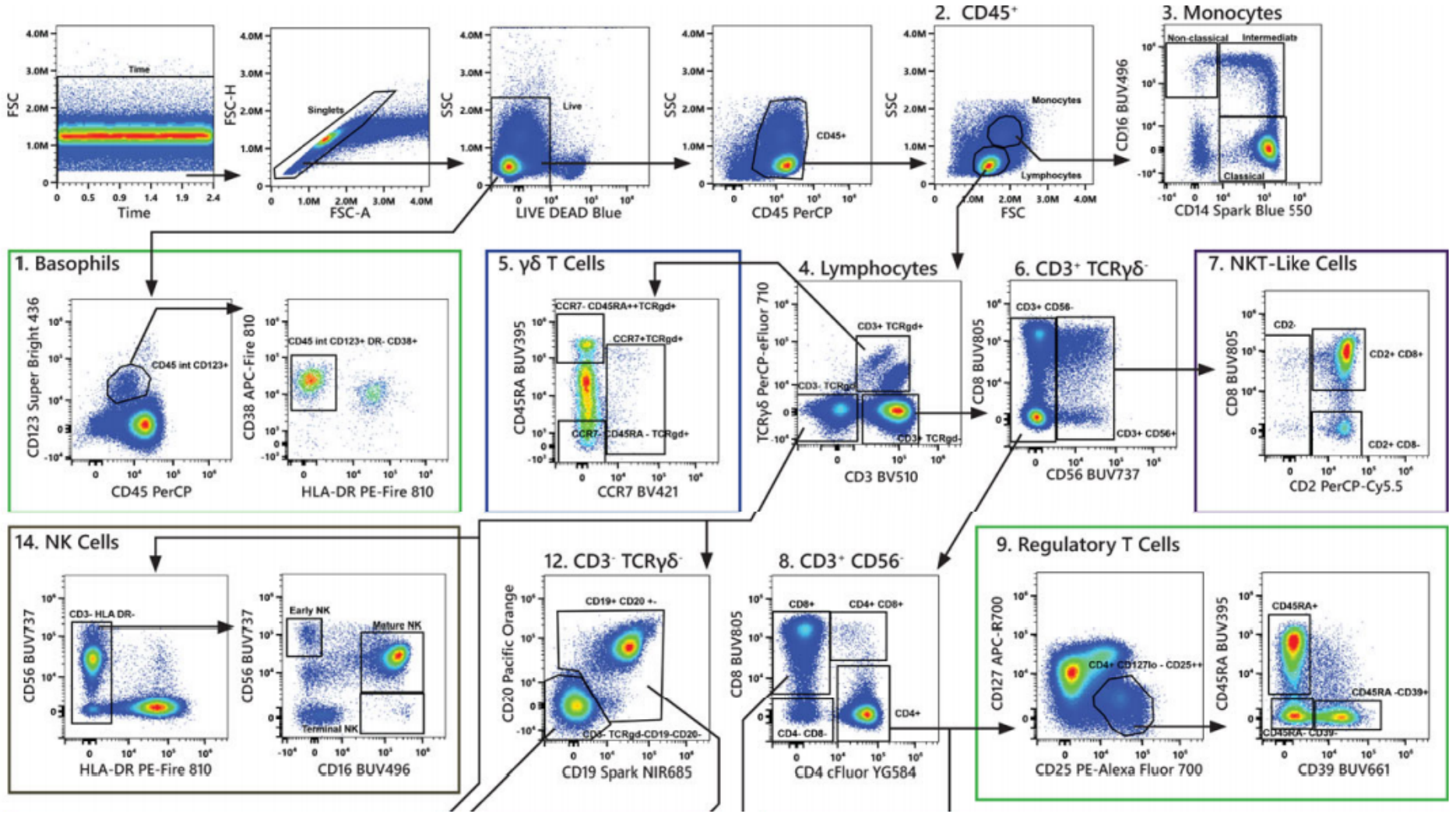


Fluidigm – Helios

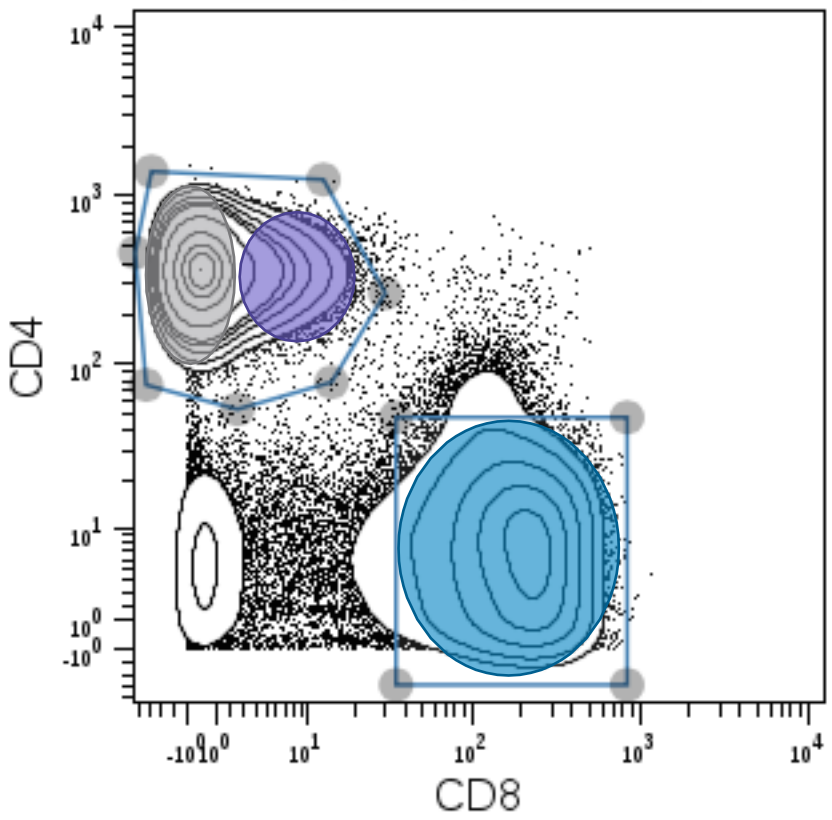


Cytek – Aurora

Gating is the standard for basic analysis of this data



Clustering is the process of identifying cell populations automatically

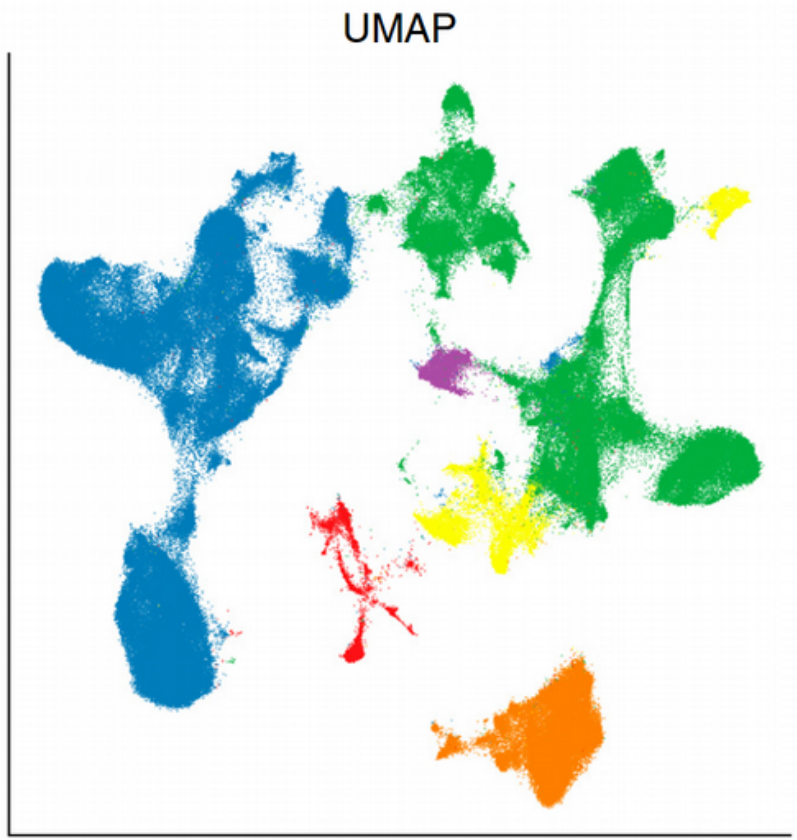


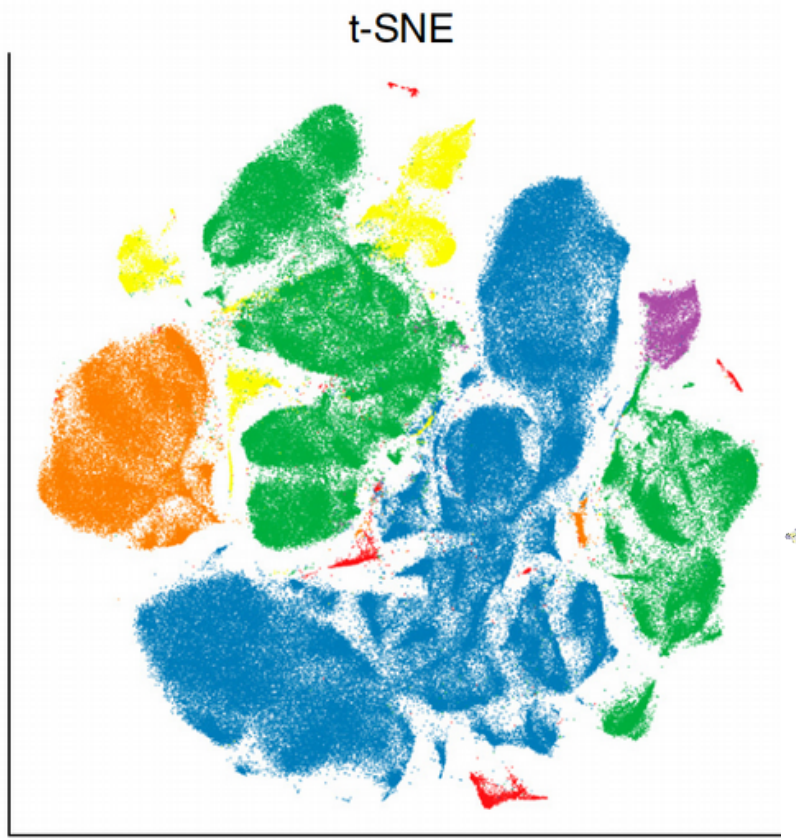
Clustering is not a well-defined problem and there are multiple methods to solve it

Examples specifically for cytometry data: FlowSOM, Vortex, Phenograph

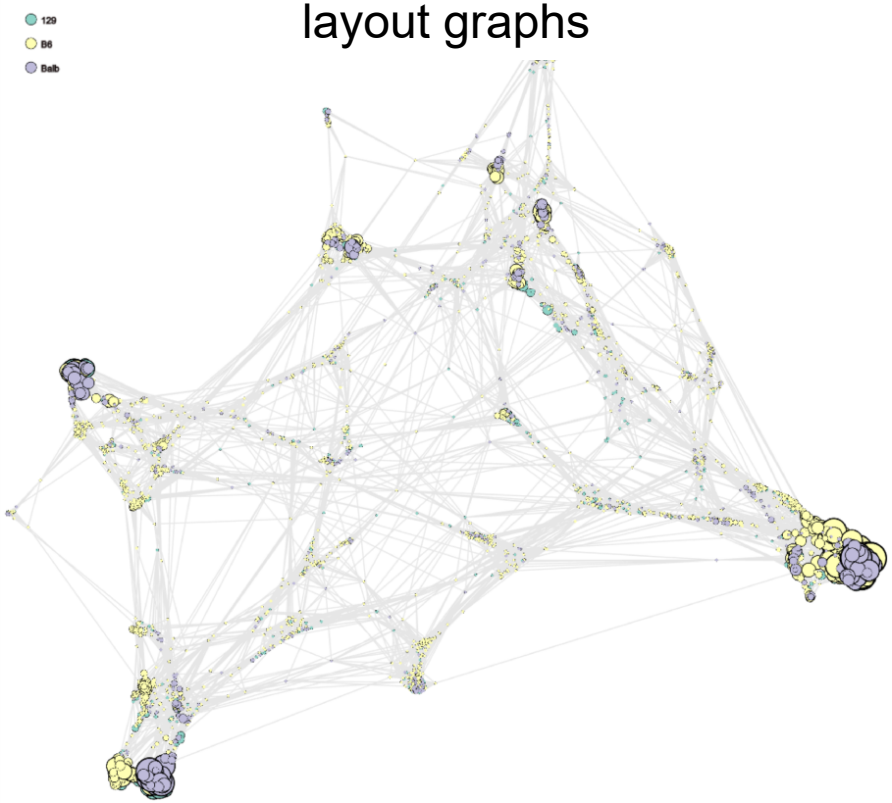
Clustering enables us to discover novel cell populations

Visualization methods enable us to get a high-level overview of a dataset

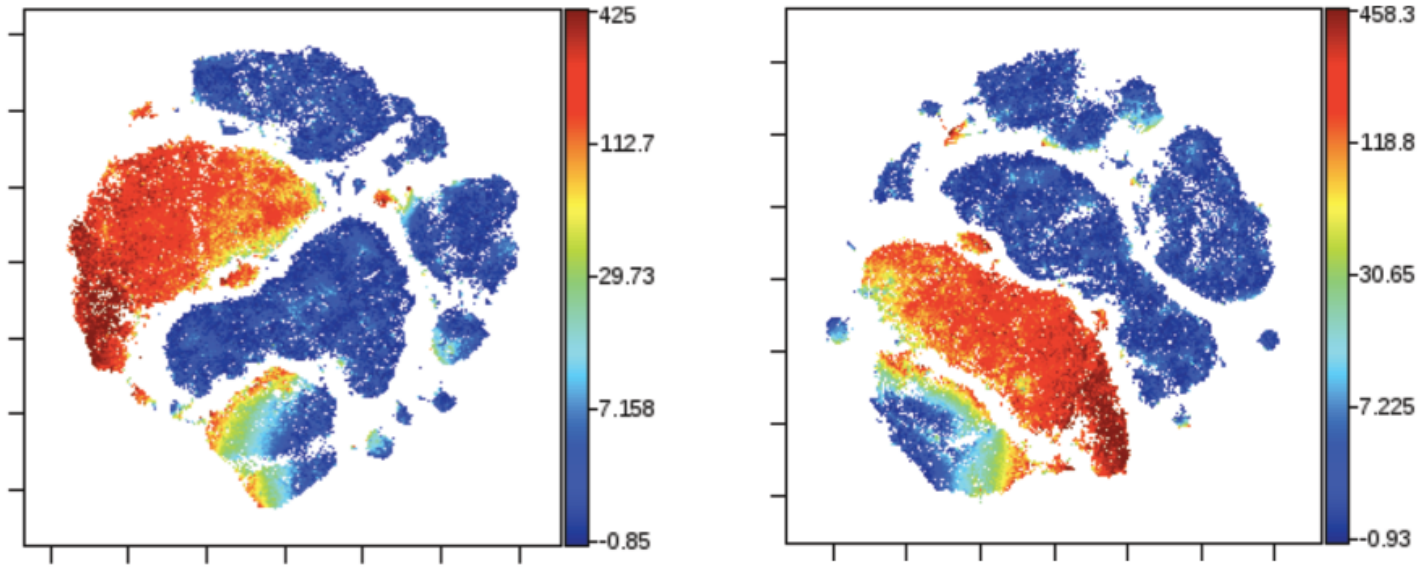




Force-directed layout graphs



Unsupervised visualization are not oriented



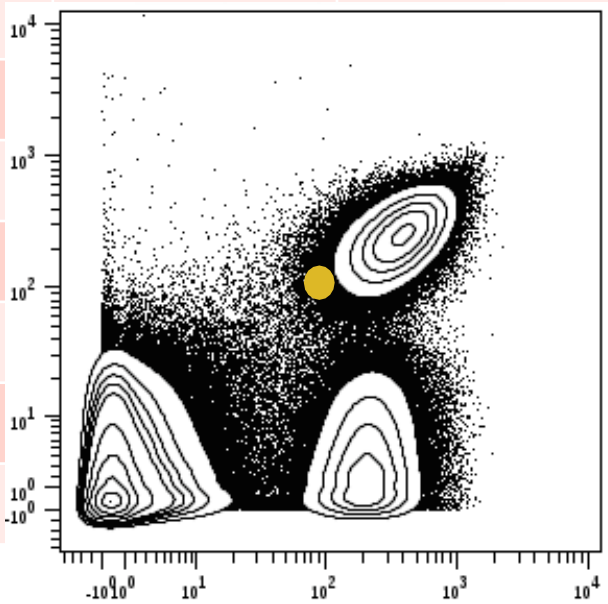
Building statistical models with this data

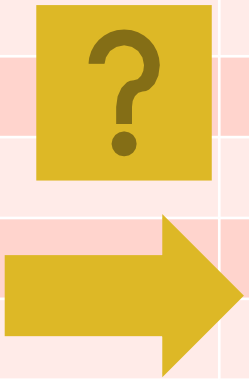
Establishing an analogy with gene expression data

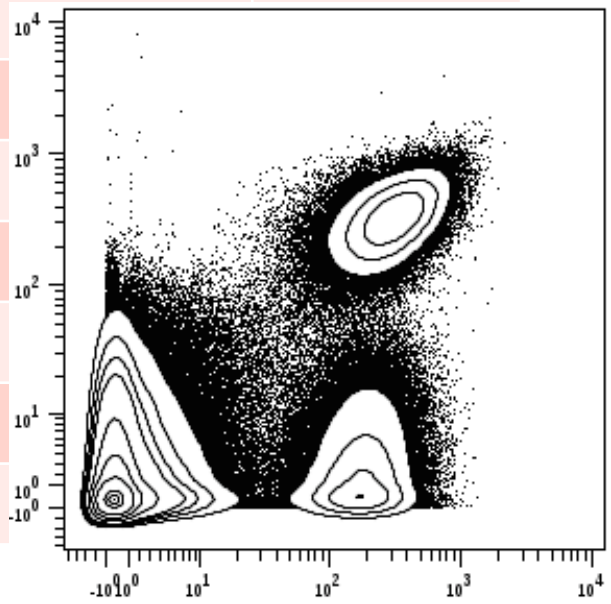
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene 1					
Gene 2					
Gene 3					
Gene 4					
…					
…					
…					
…					

Individual cells cannot be identified consistently between samples

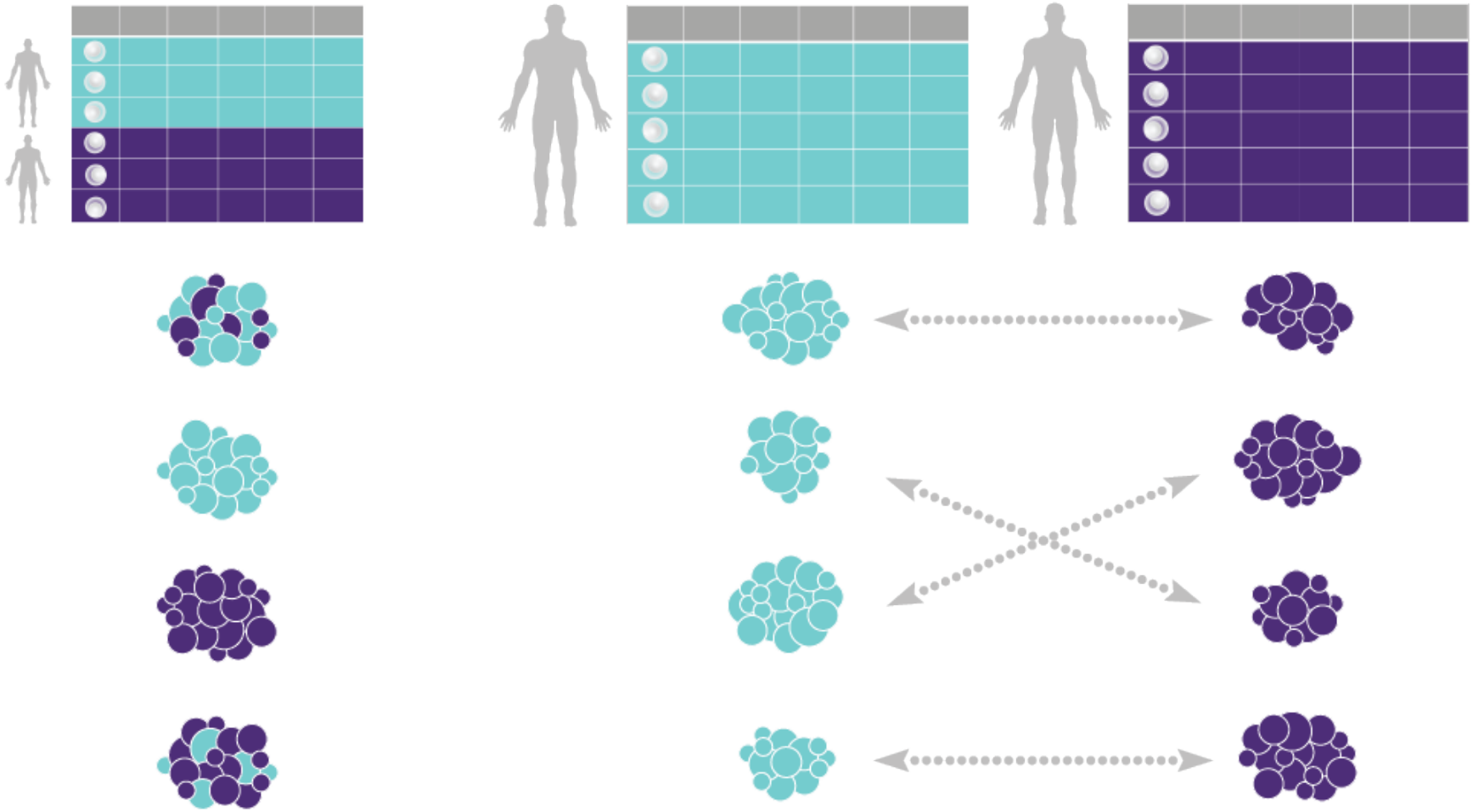
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Cell 1					
Cell 2					
Cell 3					
Cell 4					
…					
…					
…					
…					







Data needs to be pooled before clustering to enable statistical modeling



Clusters defined on pooled data can be identified consistently across samples

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Cluster 1 (Abundance)					
Cluster 1 (Marker A)					
Cluster 1 (Marker B)					
Cluster 2 (Abundance)					
Cluster 2 (Marker A)					
Cluster 2 (Marker B)					
…					
…					

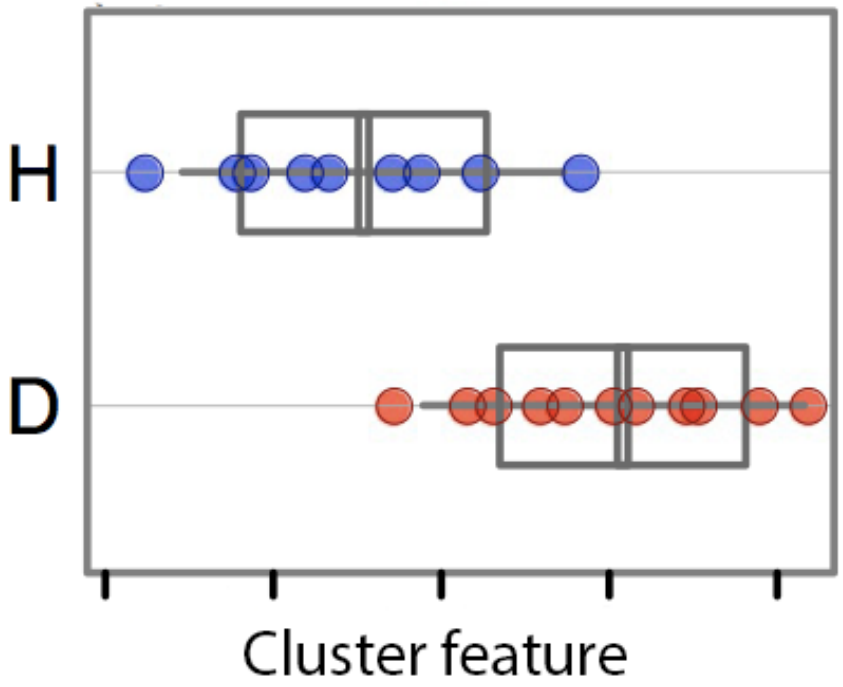
Once you get data in this shape you can borrow from a variety of existing modeling methods

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Group	Healthy	Healthy	Disease	Disease	Disease
Survival time	5	7	2	4	10
Cluster 1 (Abundance)					
Cluster 1 (Marker A)					
Cluster 1 (Marker B)					
Cluster 2 (Marker B)					
…					
…					

This approach can also be used with gated data

Example R packages: glmnet, SAM, DEseq2, siggenes

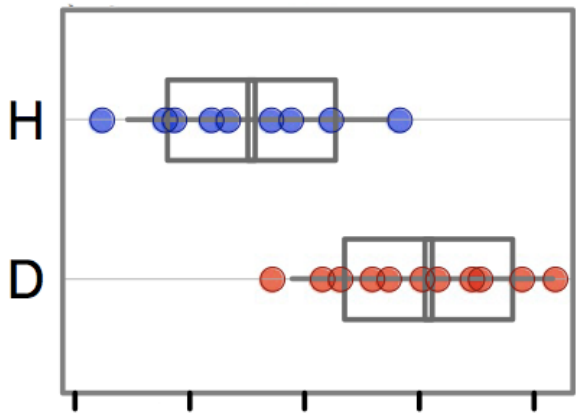
The shape of the result depends on the endpoint of interest



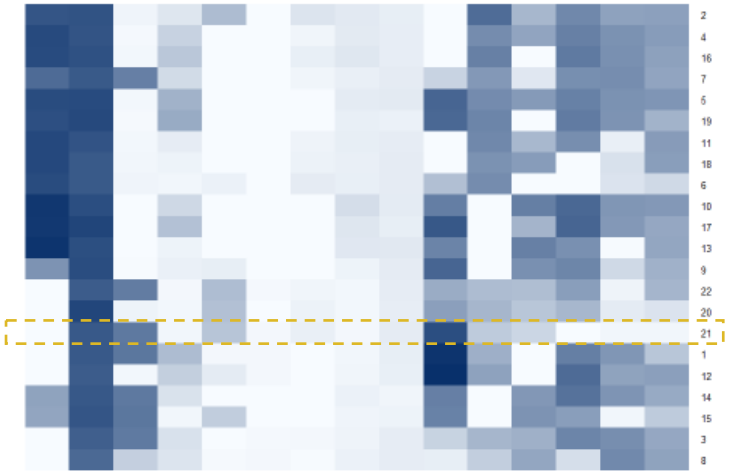
Categorical

Continuous

If a result is real, one should usually be able to bring it back to gated data

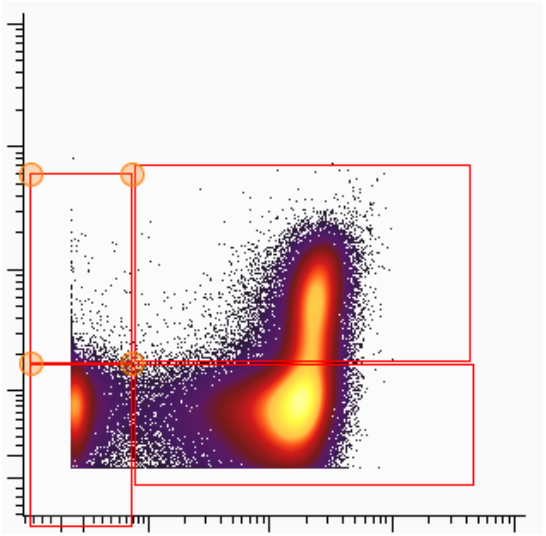


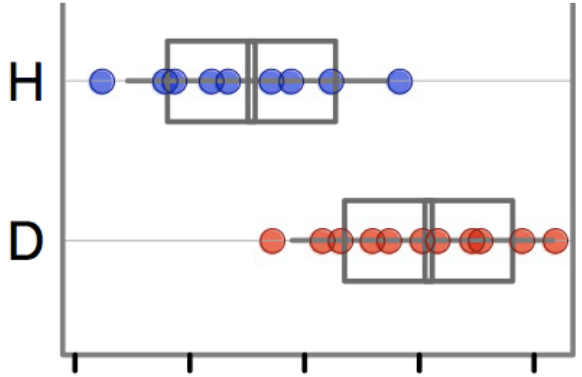
Cluster 123



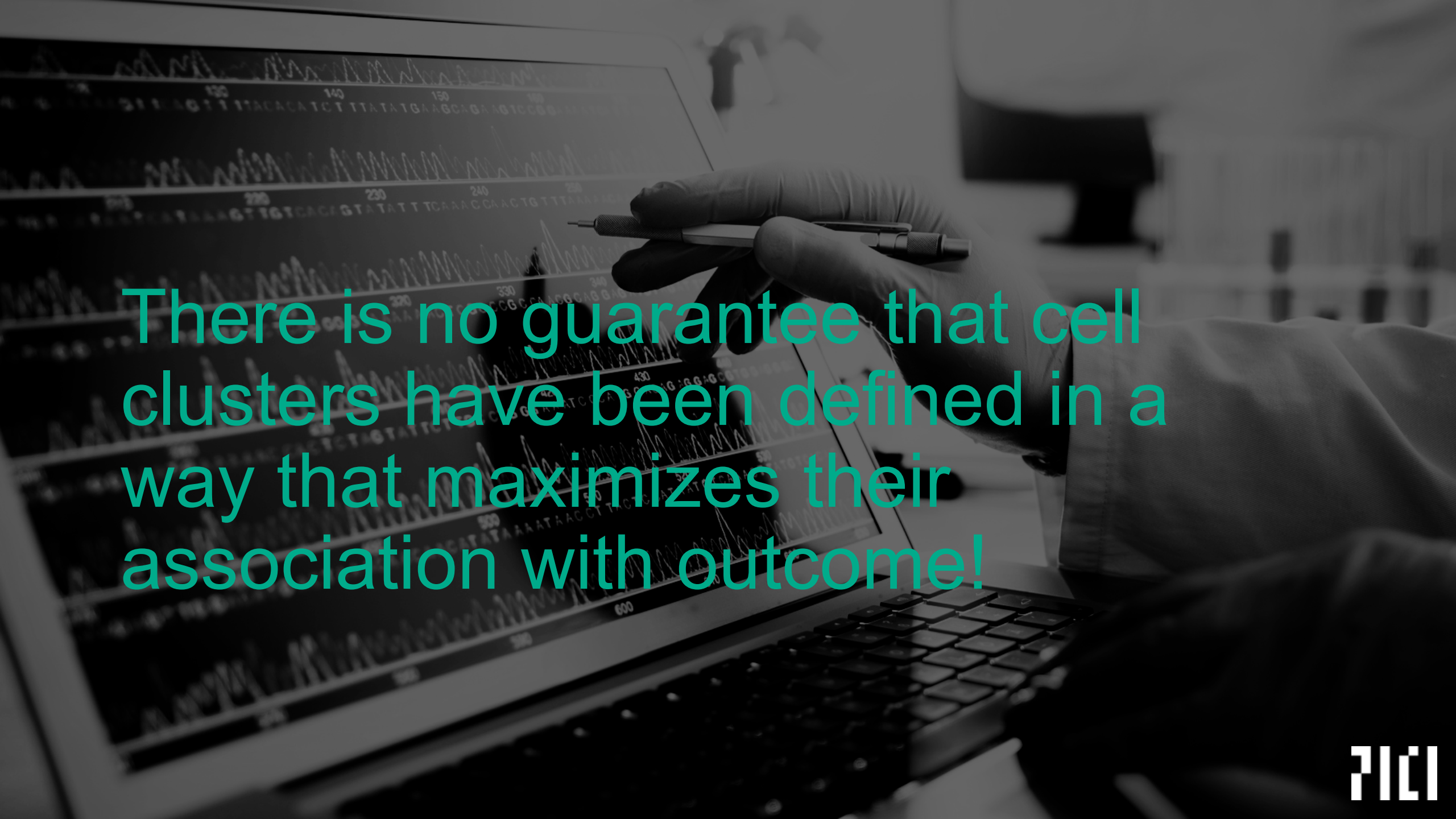
Markers

Clusters

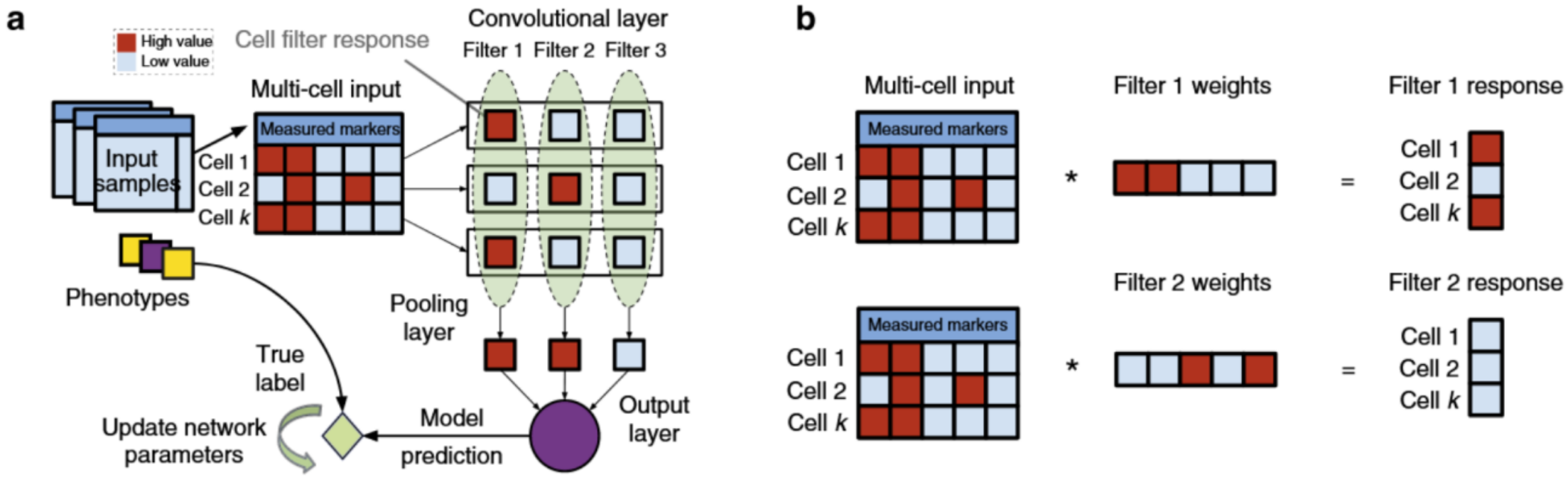




Tbet+ CD8 T cell


There is no guarantee that cell clusters have been defined in a way that maximizes their association with outcome!

Simultaneous learning of cell types and their association with outcome



Arvaniti et al. Nat Commun. 2017 Apr 6;8:14825. doi: 10.1038/ncomms14825.

Summary

- Peripheral blood data can provide key (longitudinal) insights into mechanism of action and patterns of response / resistance
- Different technologies exist to measure 28-40+ markers in millions of cells
- This data is often underutilized with analysis restricted to the identification of cell types known a-priori
- Unsupervised clustering analysis can be used to discover new cell types in this data
- Statistical models can be built to associate the presence/absence of specific cell populations with outcomes of interest

https://github.com/ParkerICI/flow-analysis-tutorial