Latent Dirichlet Allocation to Integrate Single-Cell Targeted Proteomics Methods

Pratheepa Jeganathan

Postdoctoral Scholar
Holmes Group

Department of Statistics, Stanford University

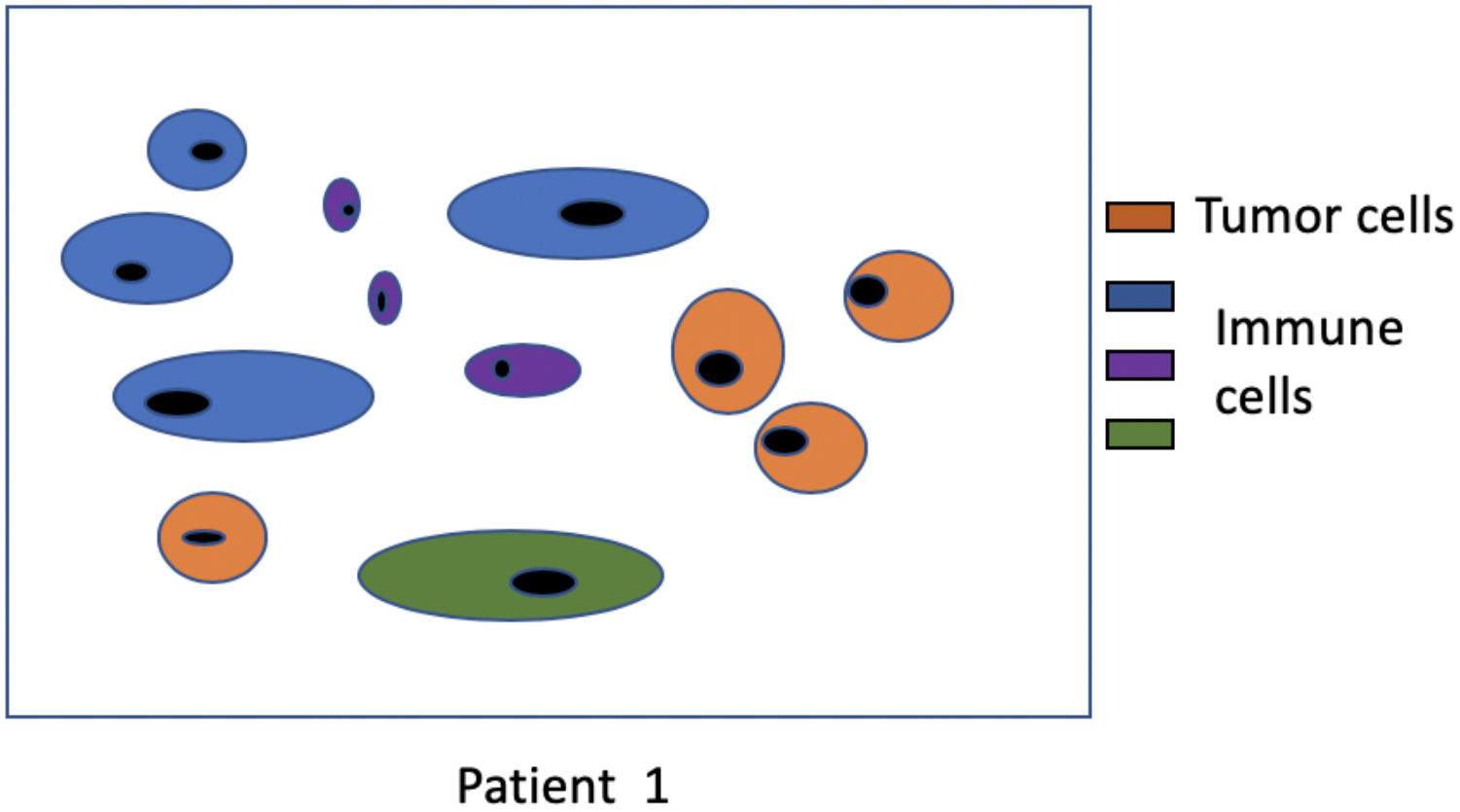
BIRS 2020

06/16/2020

Outline

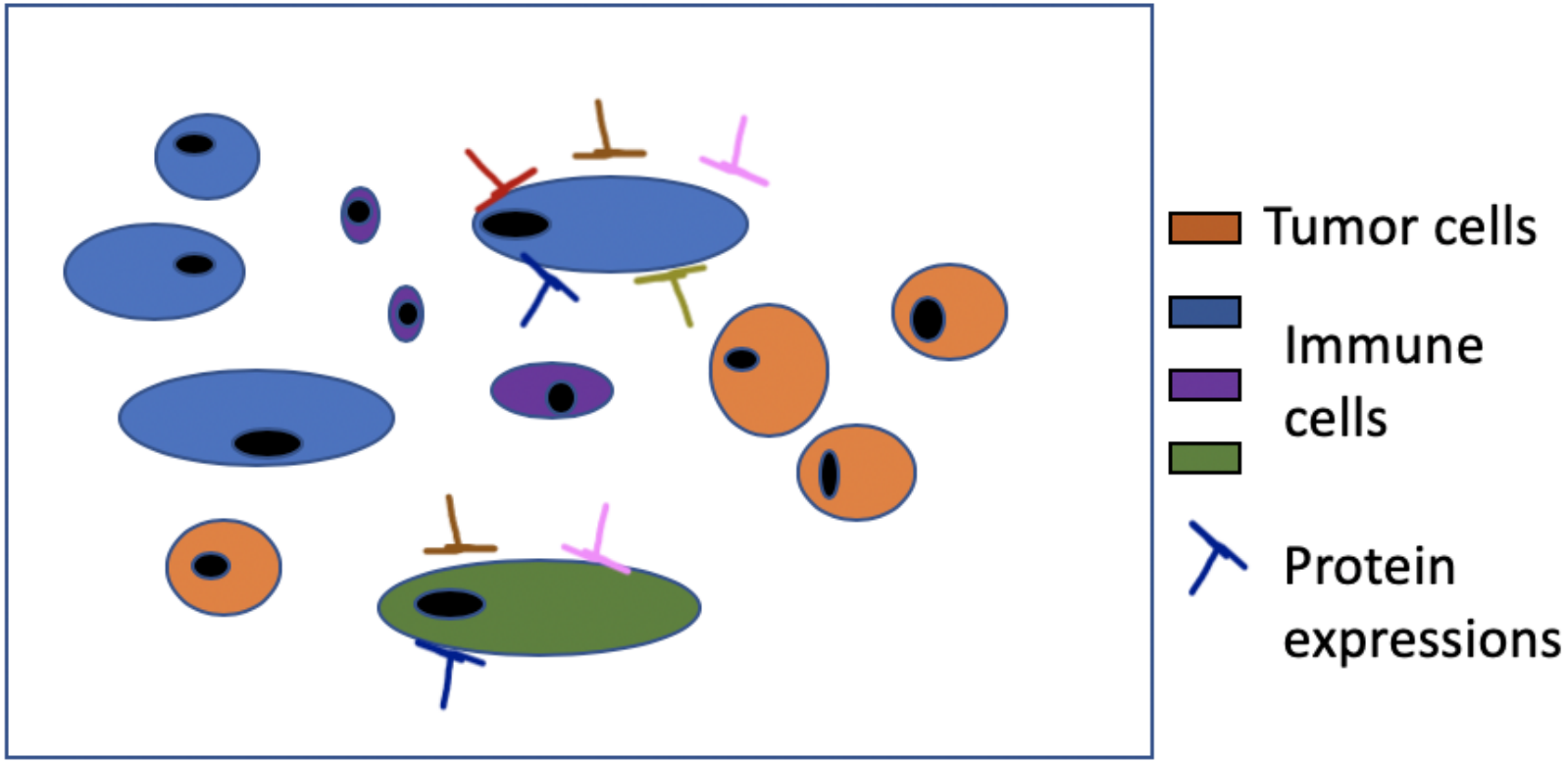
1. Motivation
2. Single-cell proteomics data
3. Topic modeling
4. Results
5. Conclusion

Characterizing the tumor ecosystem



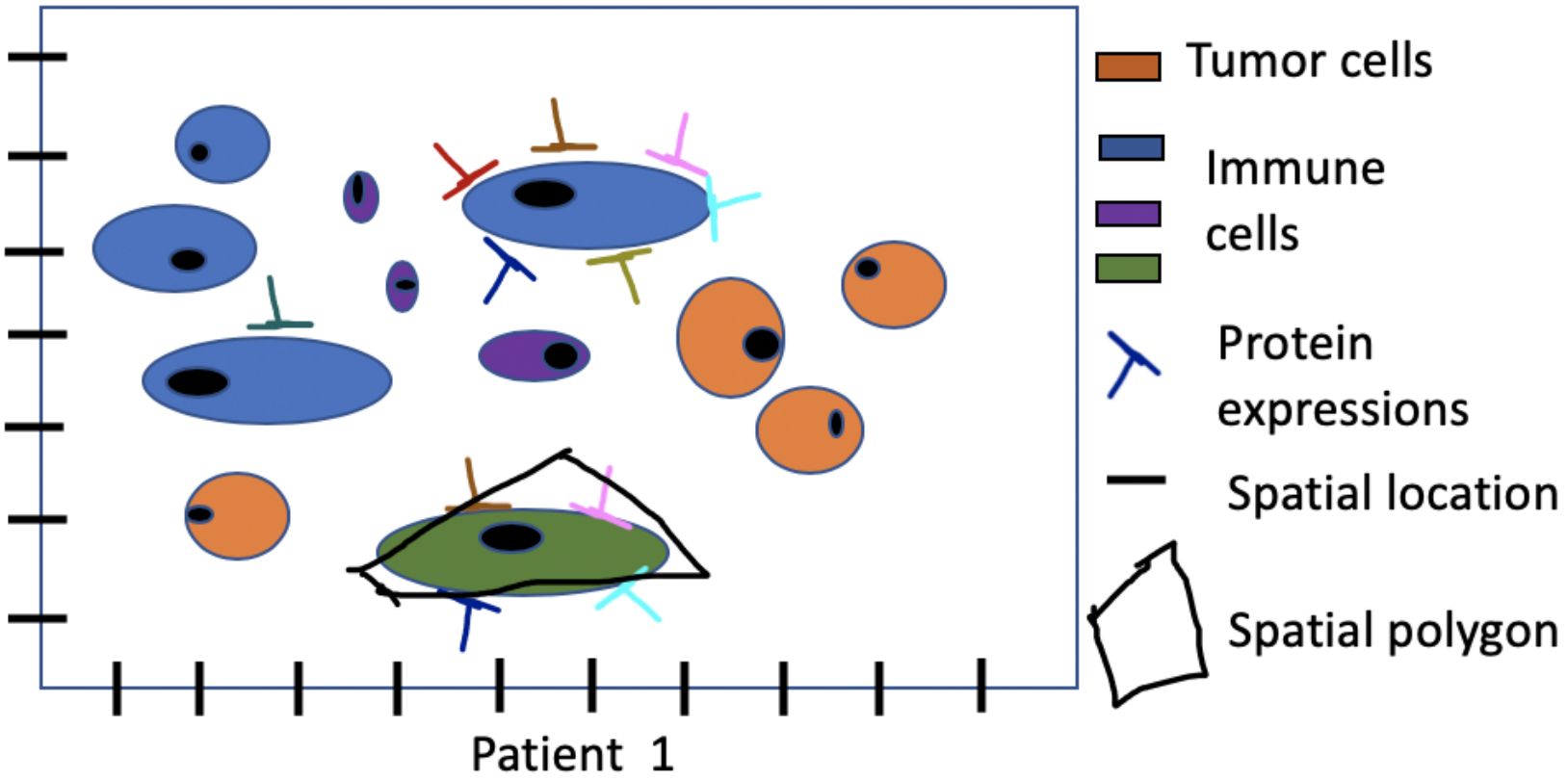
- To identify prognosis and treatment in breast cancer patients.

Single-cell targeted proteomics methods[1]



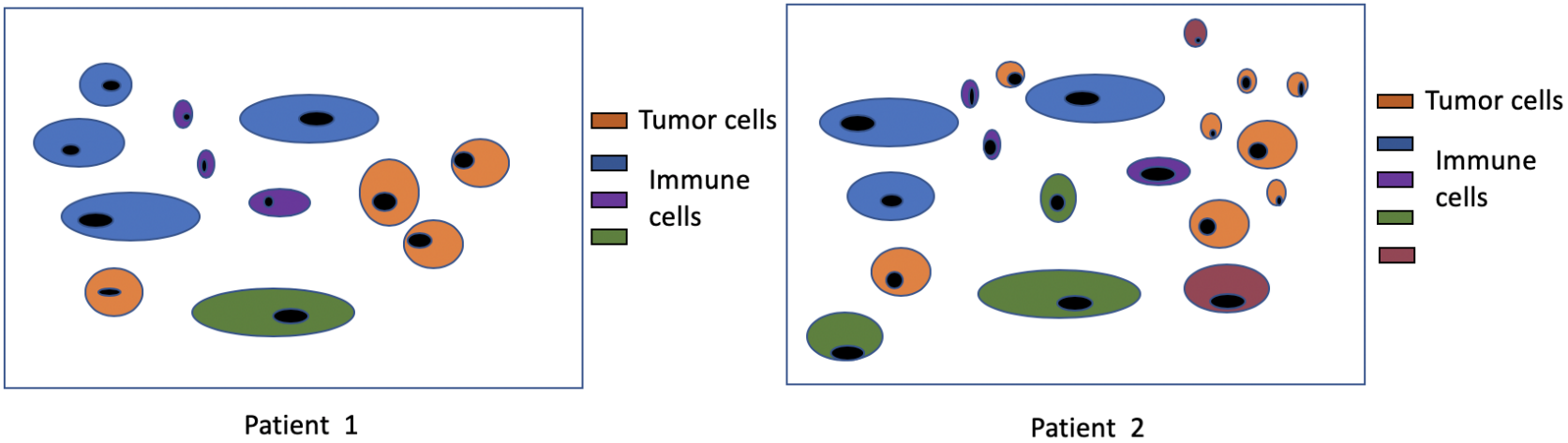
[1] CyTOF (Wagner et al., 2019)

Single-cell targeted proteomics methods[2]



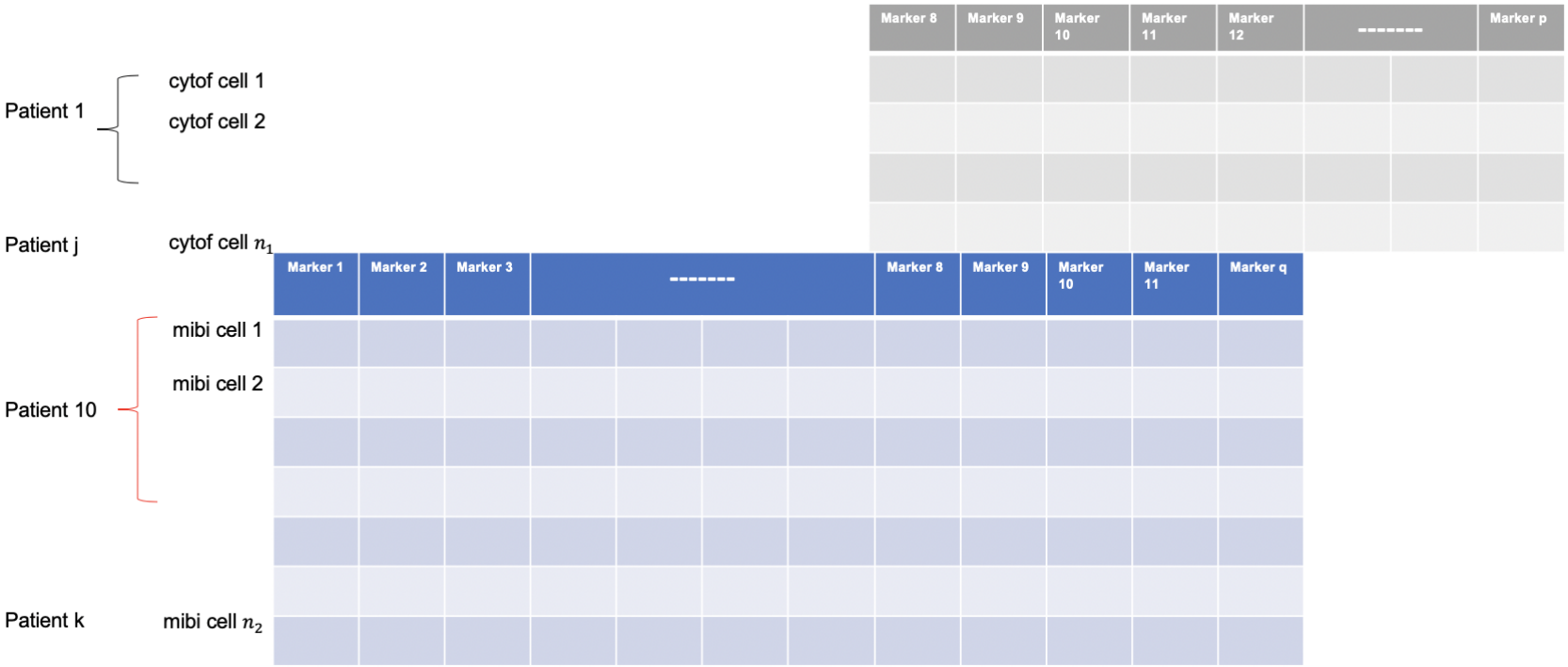
[2]MIBI-TOF (Keren et al., 2018)

Heterogeneous tumor ecosystem



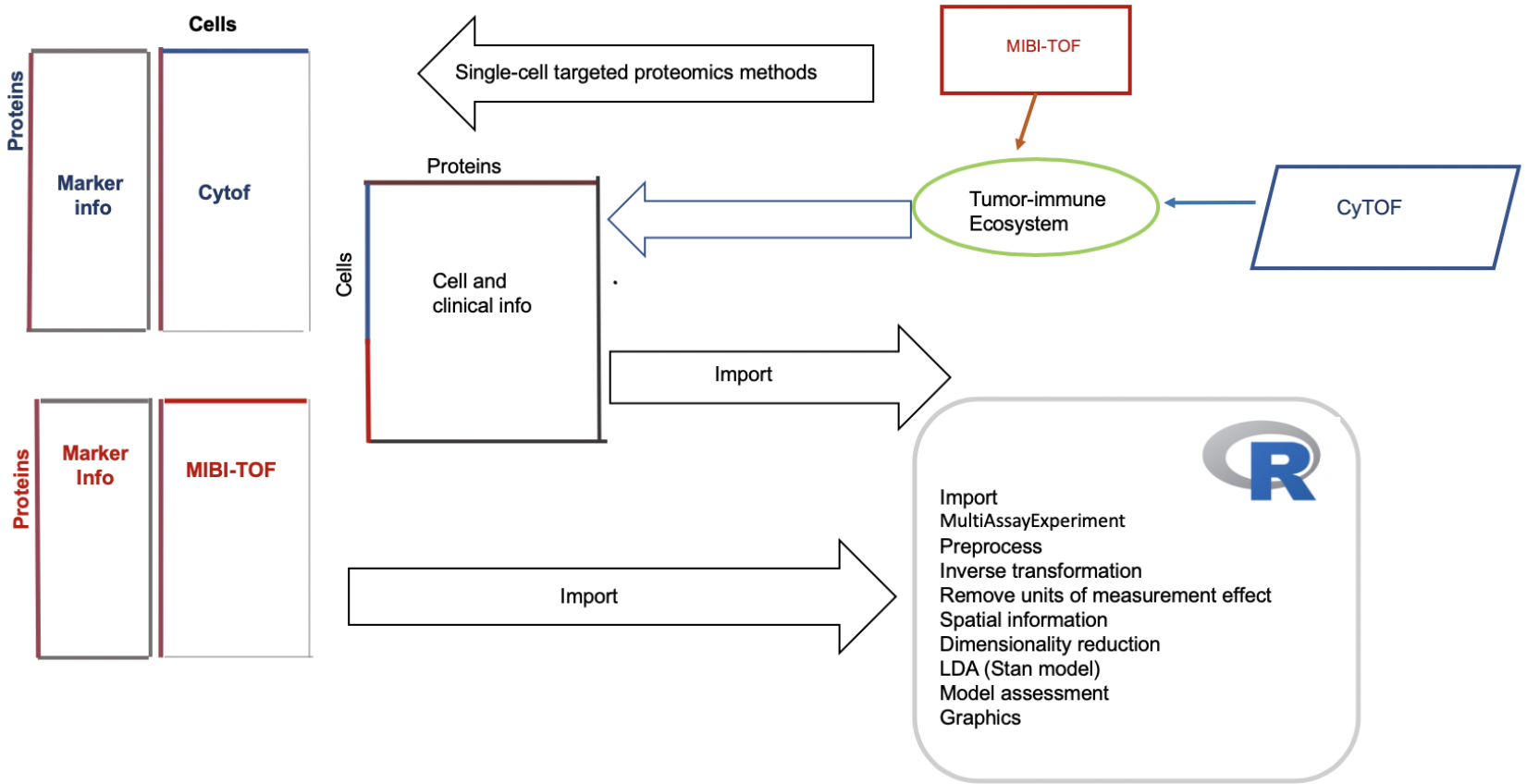
- Tumor ecosystem is phenotypically and functionally heterogeneous.
- Integrate multiple targeted proteomics methods.

Goal 1: How should we approach integrating partially-overlapping proteomic data collected on different patients with similar phenotypes?



- CyTOF : intensity is divided by five and arcsinh transformed.
- MIBI -TOF: Counts are divided by cell size, arcsinh transformed, and standardized across markers.
- Note: MIBI-TOF markers have transformed values more than 4.2

MultiAssayExperiment[3]

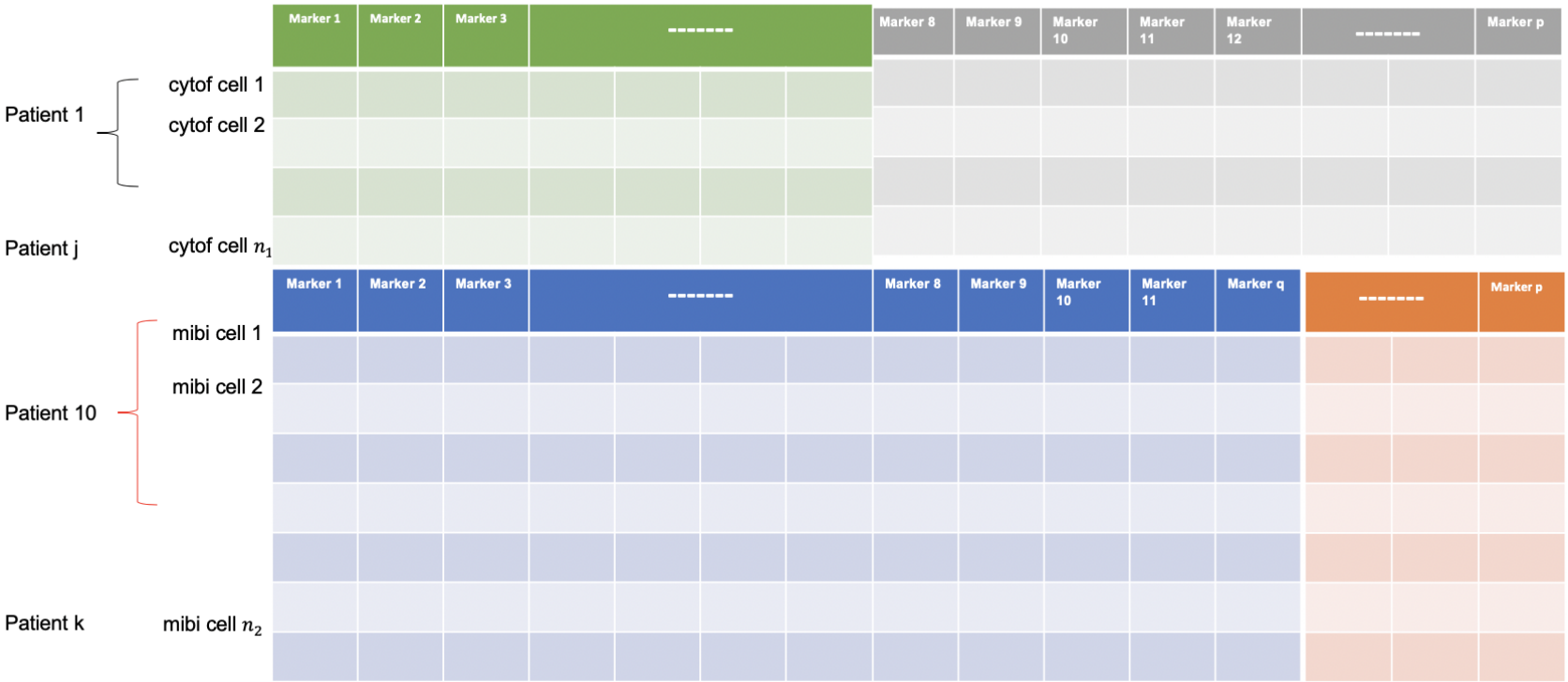


[3] (Ramos et al., 2017)

Remove units of measurement effect

- **Z**: Inverse transformation of CyTOF.
- **X**: Given transformed data of MIBI-TOF.
- For each marker i in MIBI-TOF, $y_{ij} = \left(\frac{x_{ij} - x_{i,\text{min}}}{x_{i,\text{range}}} \right) \times u$, where u is the upper limit of the rescaled variable.
  - u is the maximum of **Z**.
- (**Z**, **Y**): Round to integer.
- Note: We can match the quantiles if there is no concern on the given transformation.

Impute on rescaled data



- K-nearest neighbour averaging[4].

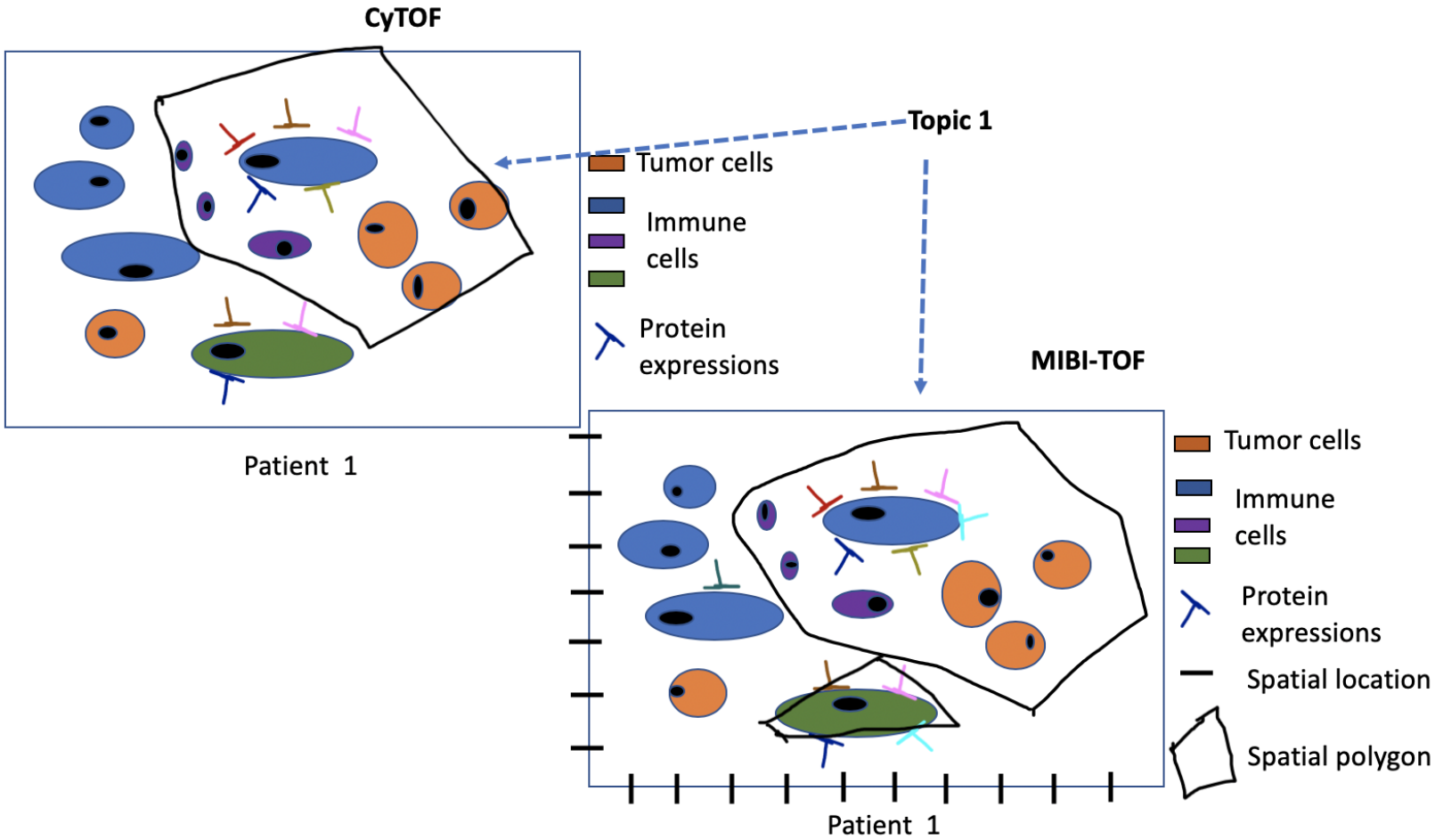
[4] (Hastie, Tibshirani, Narasimhan, & Chu, 2019)

Why topic modeling

- Given (**Z**, **Y**).
- **Z** does not have spatial information of cells.
- **Y** has spatial information of cells.

Why topic modeling

Goal 2: Without including the spatial x-y coordinate data, how well can we predict cell co-location.

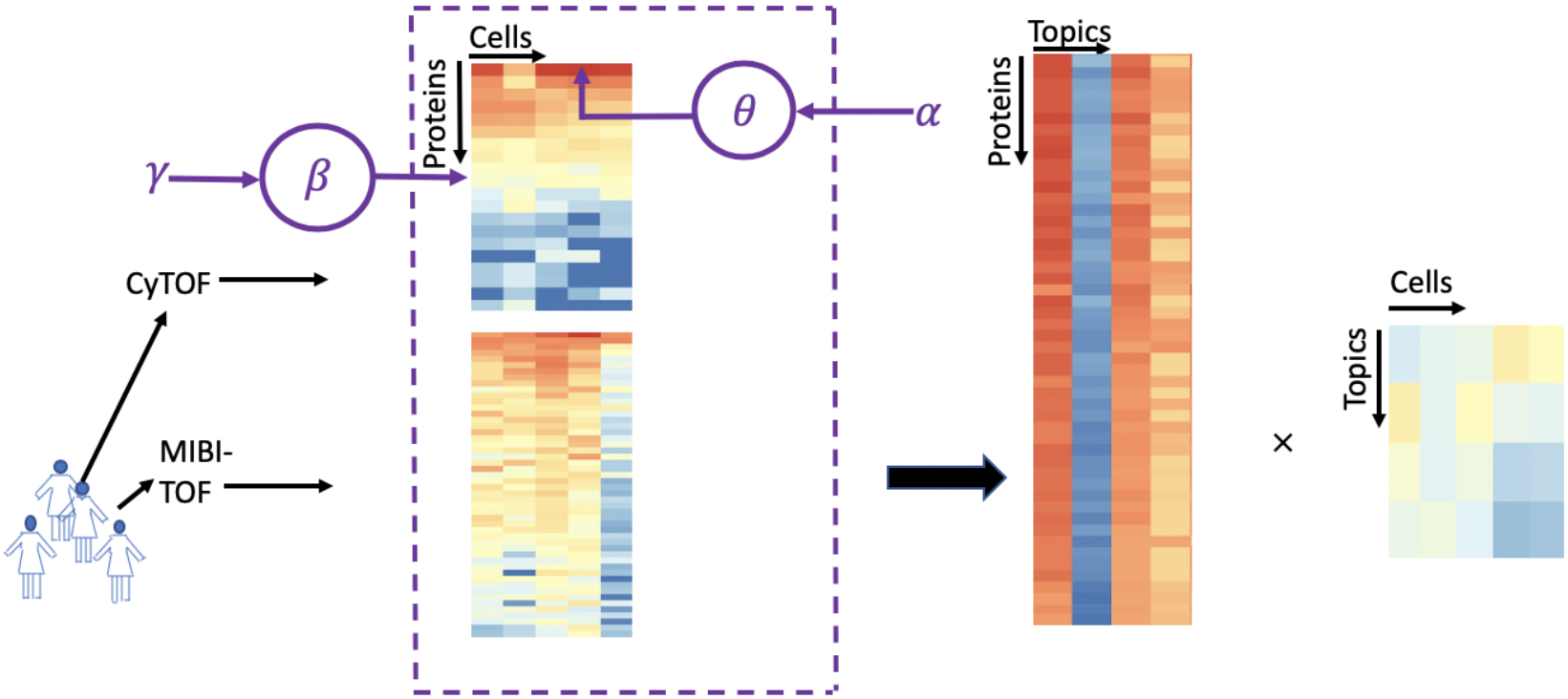


Why topic modeling

Goal 3: Can we predict the spatial expression patterns of proteins measured on CyTOF but not measured in the MIBI-TOF data?

Simulate from the fitted topic model.

Latent Dirichlet Allocation (LDA)[5]



[5](Blei, Ng, & Jordan, 2003)

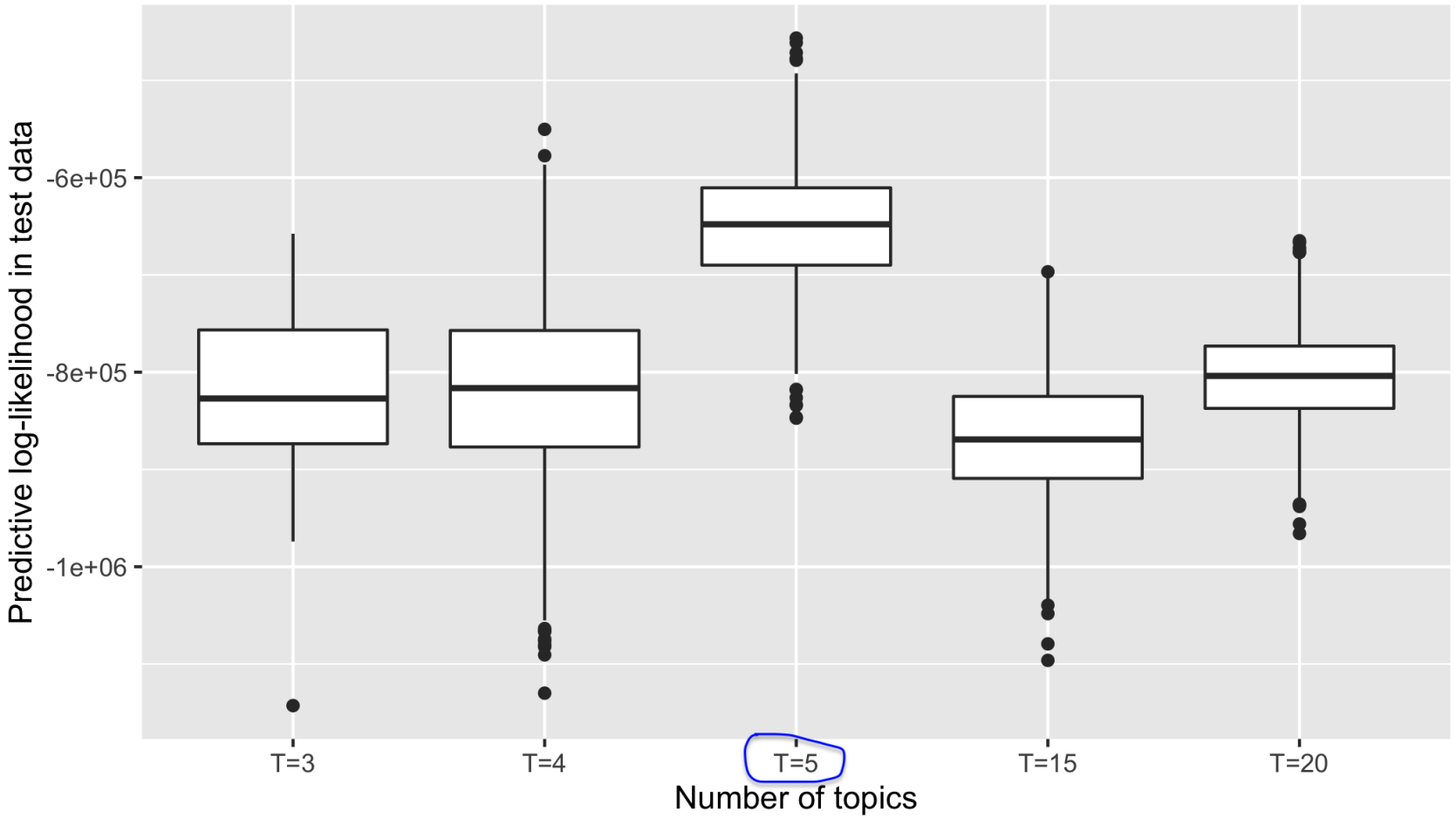
Hamiltonian Monte Carlo - No U-Turn Sampler (NUTS)

- Hamiltonian Monte Carlo (HMC) is a Markov chain Monte Carlo method (MCMC).
  - Avoids random walk behaviour, takes series of steps informed by first-order gradient information (of log posterior density).
  - Use when direct sampling is difficult (approximation to the posterior).
  - Two tuning parameters: step size and number of steps.
  - HMC - NUTS method: user does not need to specify the above two tuning parameters.
- Use rstan R/Stan (probabilistic programming language for statistical inference) package.

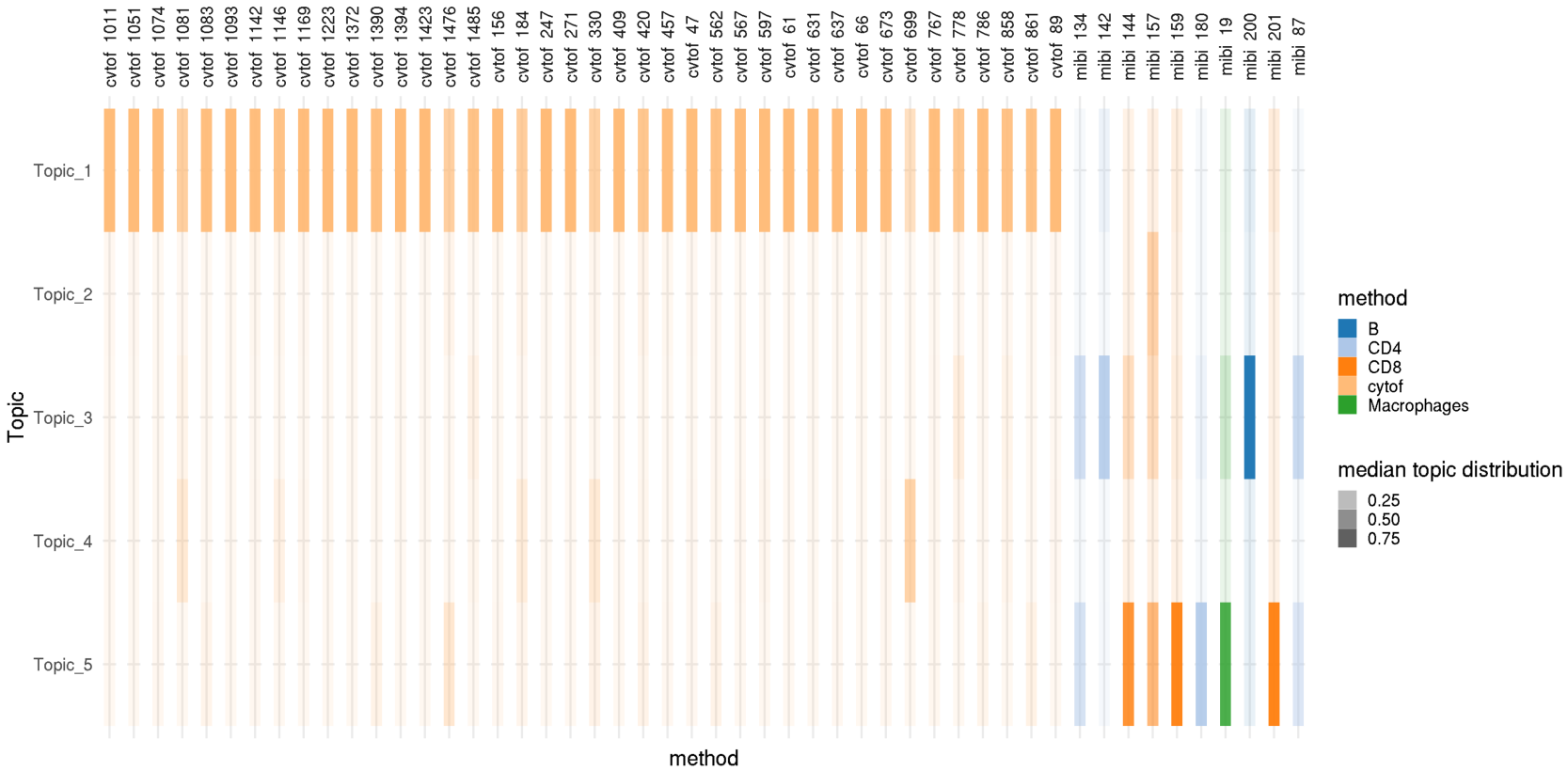
Results on subset of data

- Chose one patient from CyTOF (Live cells: immune panel)
- Chose one patient from MIBI-TOF. Ideal is to use MIBI-TOF data from the same patient.
- Kept 10% cells from MIBI-TOF for the test data to choose number of topics based on the posterior log-likelihood.
- Model assessment: simulate data from the fitted model and plot the distribution of the median of protein expression.

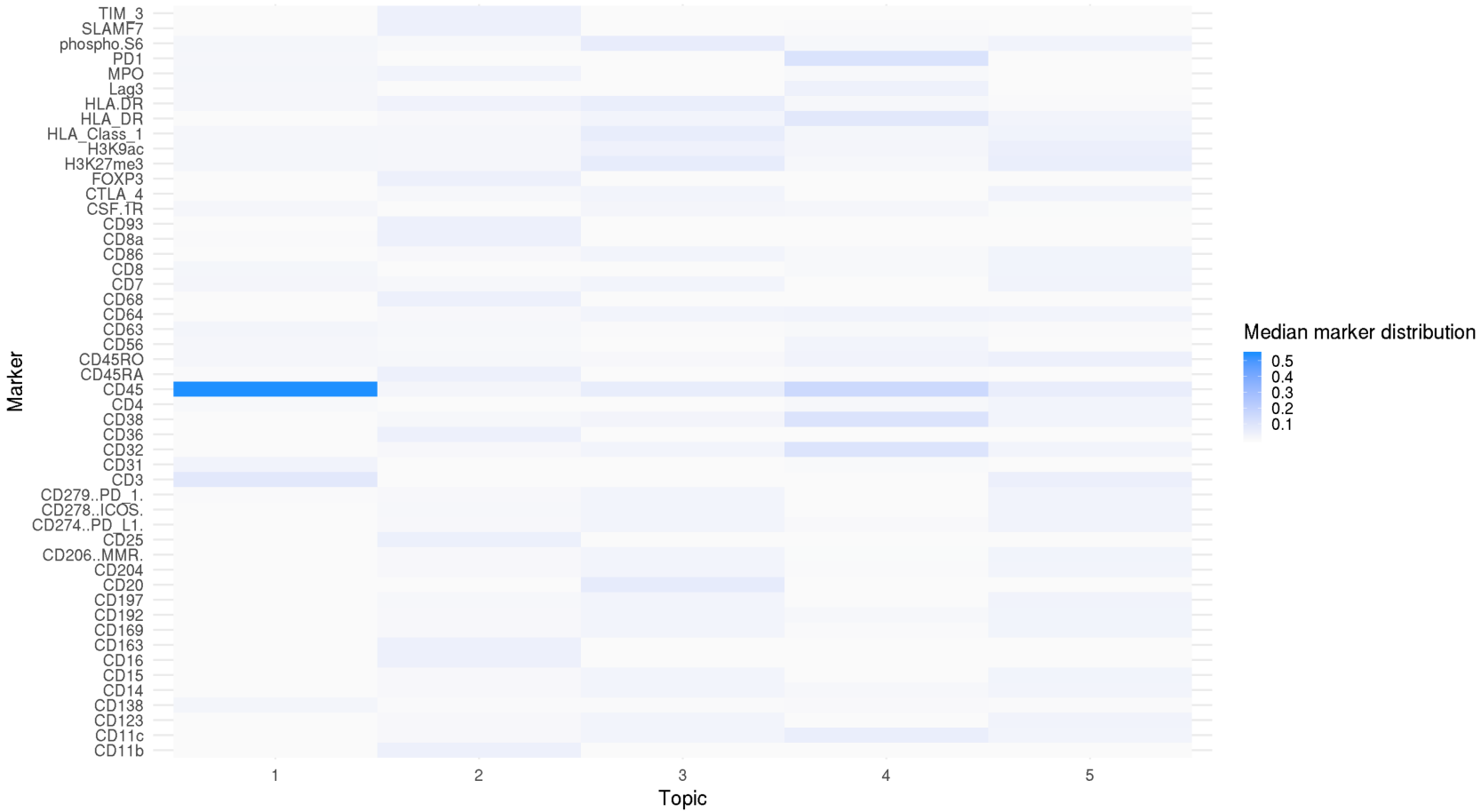
Choosing number of topics



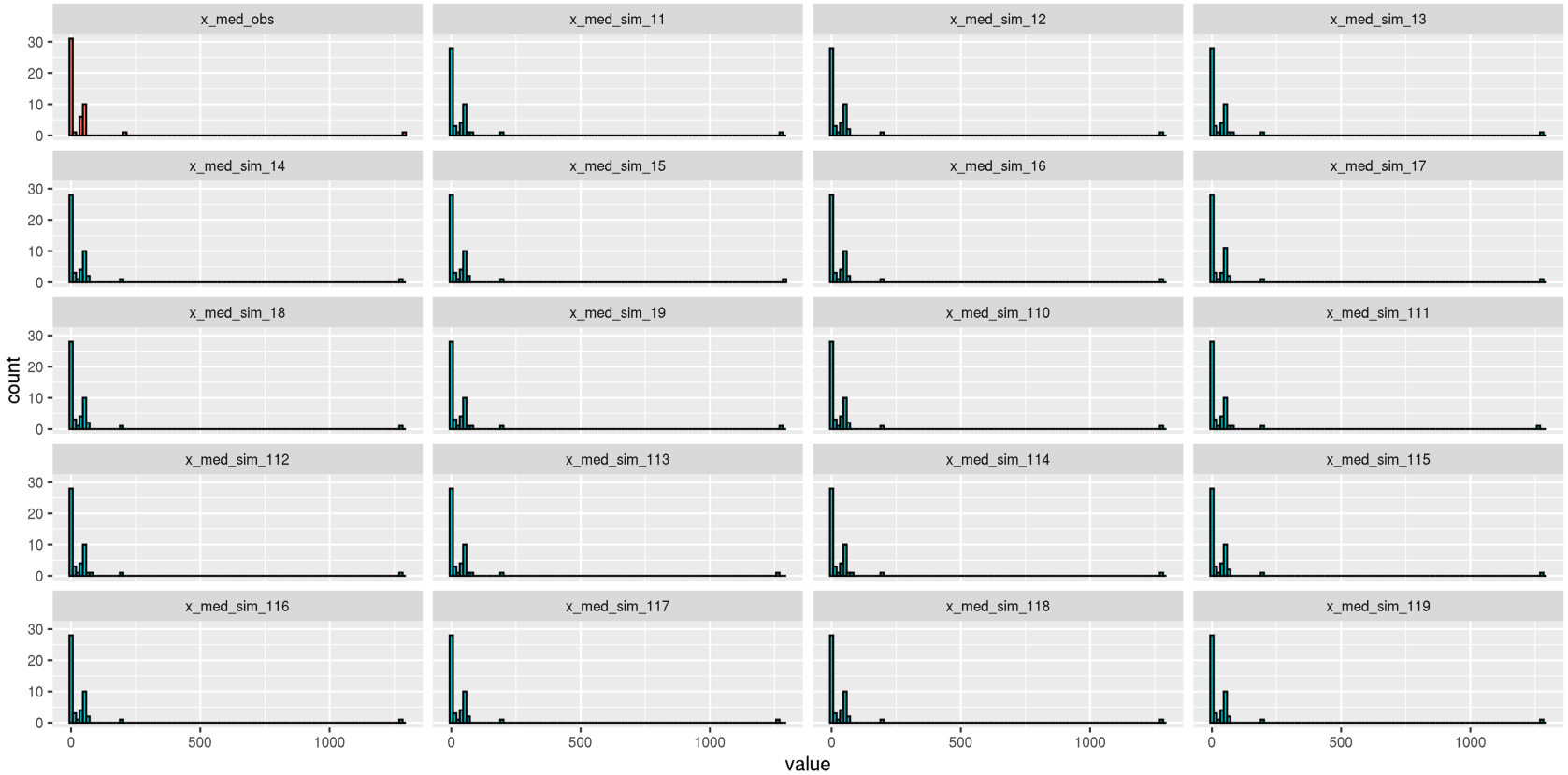
Estimated topic distribution



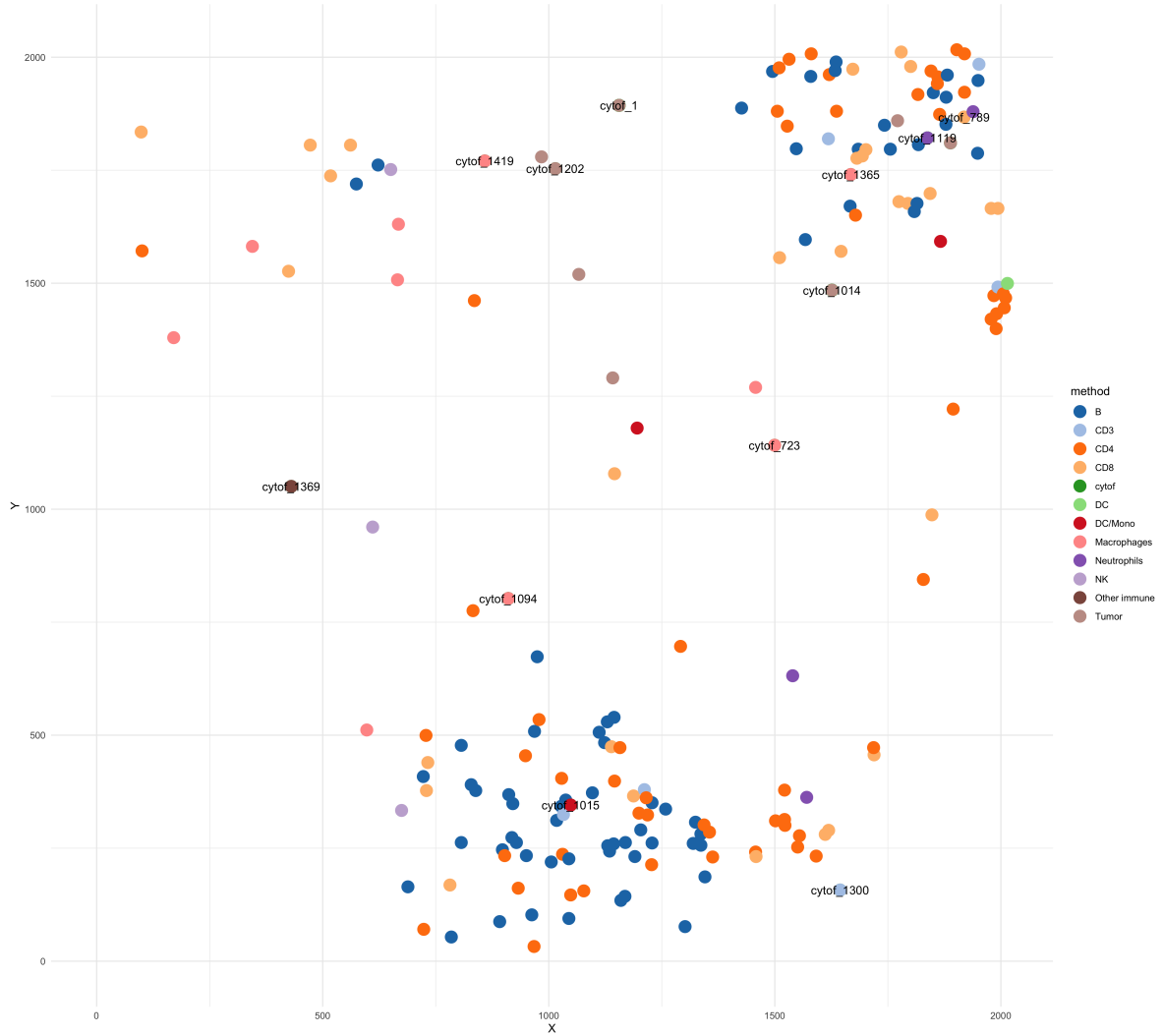
Estimated marker distribution in each topic



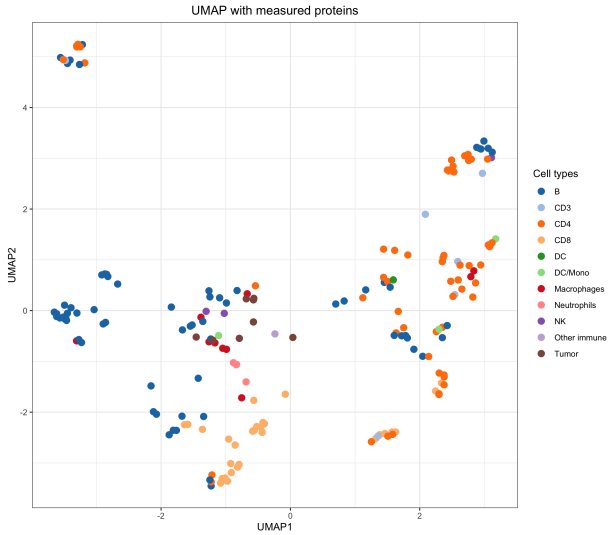
Model assessment

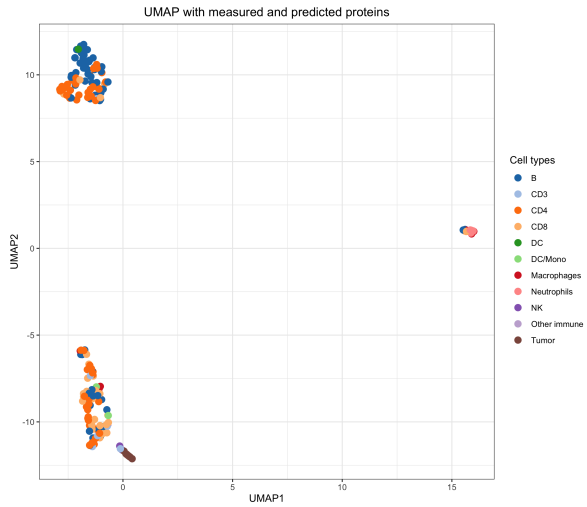


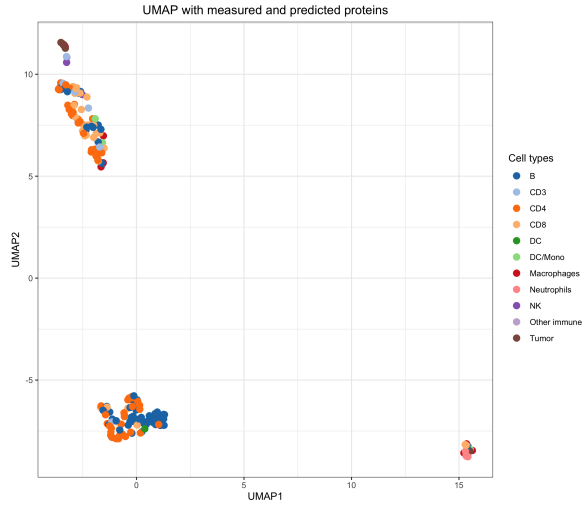
Infer spatial co-location of CyTOF cells



Predict spatial pattern of proteins not measured in the MIBI-TOF







Conclusion

- Goal 1: Integrate partial-overlapping proteomic data using MultiAssayExperiment.
- Goal 2: Leverage topic modeling to infer spatial co-location of cells of CyTOF data.
- Goal 3: Leverage topic modeling to predict spatial expression pattern of proteins not measured in the MIBI-TOF data.

Challenges and future work

- Transformation.
- Units of measurements are different across platforms.
- Different subjects in different platforms.
- Topic modeling
  - Expensive computational cost for four chain and 2000 iterations with 1000 warmup iterations.
  - Computational cost for fixing label switching
  - Recompute $\hat{R}$, effective sample size (ESS), posterior log likelihood on train data.
  - Identify best range for the number of topics.
- Infer spatial polygons.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Hastie, T., Tibshirani, R., Narasimhan, B., & Chu, G. (2019). impute: impute: Imputation for microarray data [Computer software manual]. (R package version 1.60.0)

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., ... others (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, *174*(6), 1373–1387.

Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., ... others (2017). Software for the integration of multiomics experiments in bioconductor. *Cancer research*, *77*(21), e39–e42.

Wagner, J., Rapsomaniki, M. A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., ... others (2019). A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, *177*(5), 1330–1345.

Thank You!

Professor Susan Holmes
(Stanford Statistics)
Postdoc Mentor

Dr. Kris Sankaran
(Former Holmes Lab Member)