

Weighted Sum-of-Trees Model for Clustered Data

Kevin McCoy^{1,2}, Zachary Wooten³, Katarzyna Tomczak⁴,
Christine B. Peterson²

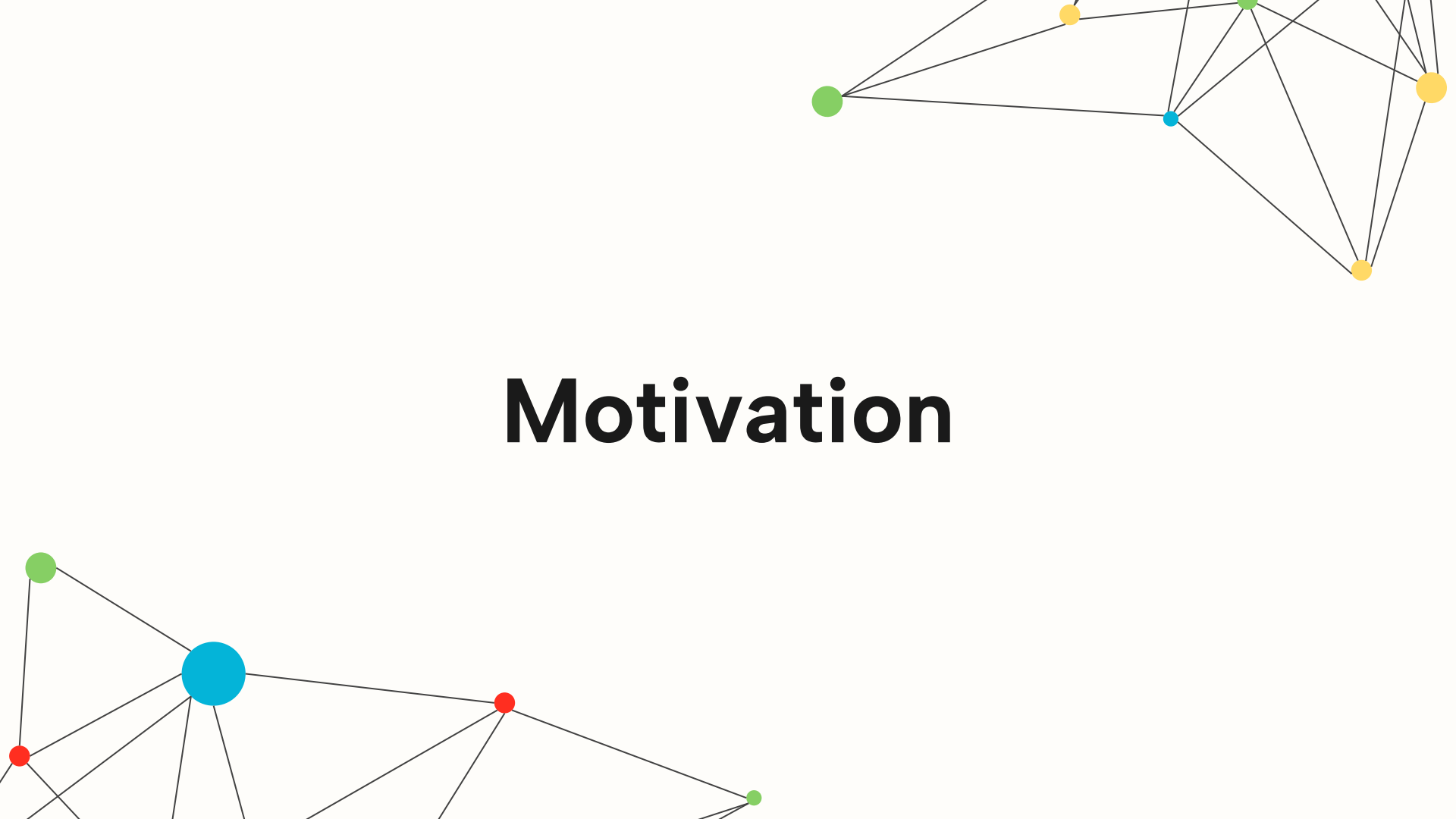
¹Department of Statistics, Rice University

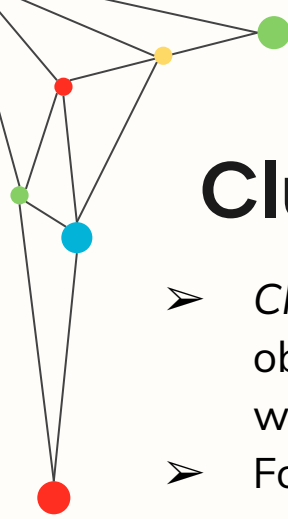
²Department of Biostatistics, UT MD Anderson Cancer Center

³Department of Biostatistics, St. Jude Children's Research Hospital

⁴Department of Translational Molecular Pathology, UT MD Anderson Cancer Center

Motivation

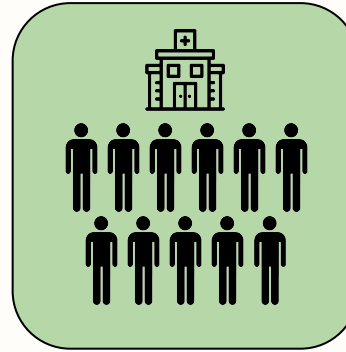




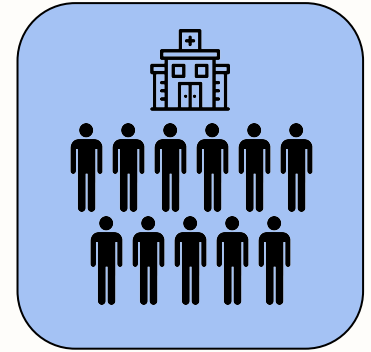
Clustered Data

- *Clustered data* arise when observations from a sample are nested within groups.
- For example:
 - A primary school wants to analyse test scores of children from different classrooms / teachers.
 - A new drug is being tested in cancer patients across the country at a few different hospitals, with different standards of care.

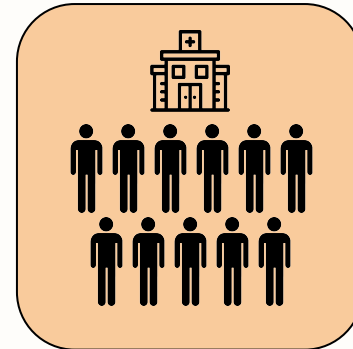
Hospital A



Hospital B



Hospital C



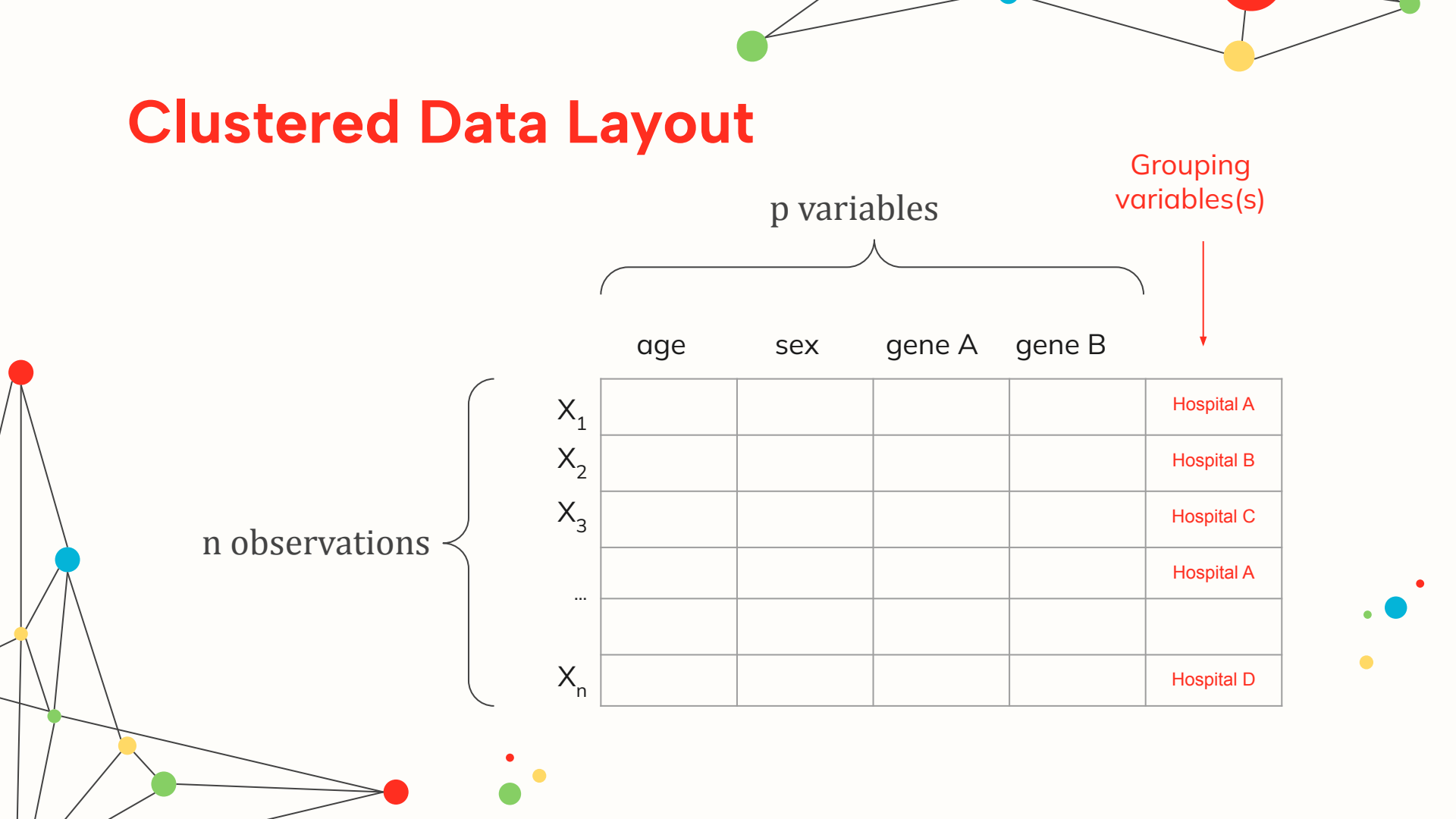
Standard Data Layout

n observations

p variables

	age	sex	gene A	gene B
X_1				
X_2				
X_3				
...				
X_n				

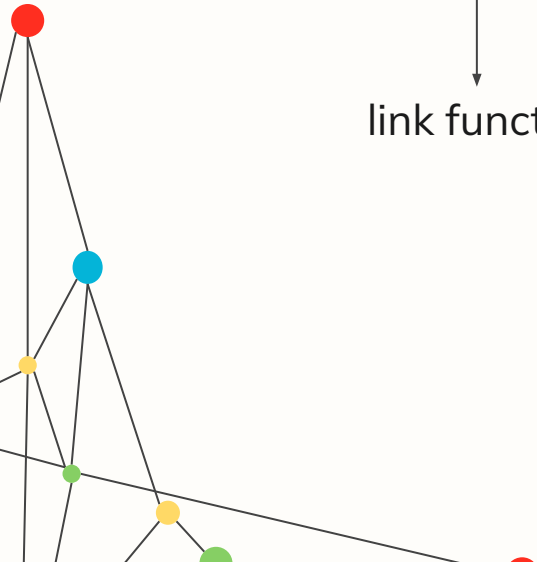
Clustered Data Layout





Generalized Linear Mixed Model (GLMM)


$$g(\mathbb{E}[y|X]) = X\beta + Z\alpha$$



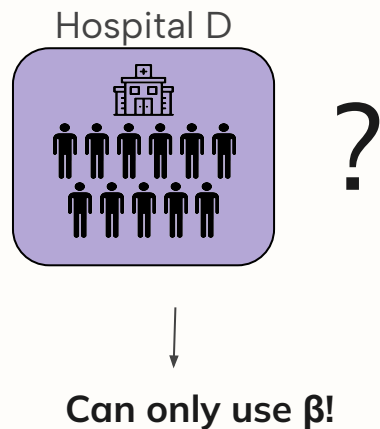
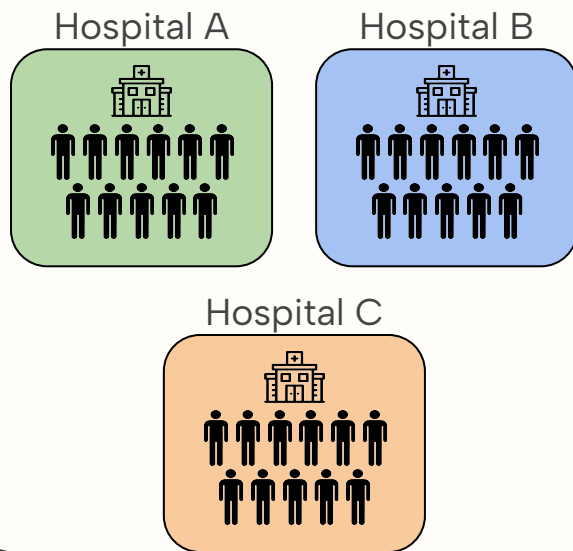
link function

fixed effects random effects

GLMM Limitations

$$g(\mathbb{E}[y|X]) = X\beta + Z\alpha$$

- The random part $Z\alpha$ is strictly linear and additive.
- Assumes observations in different clusters are independent
- **Can only use the fixed effects to make out-of-sample predictions.**



Tree Based Methods

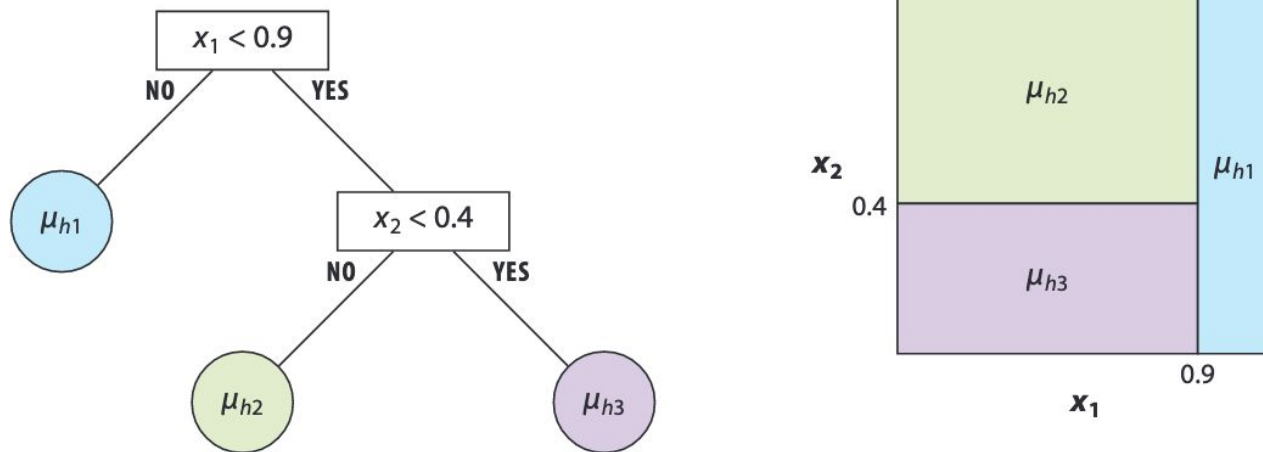


Figure 1

(a) An example binary tree, with internal nodes labeled by their splitting rules and leaf nodes labeled with the corresponding parameters μ_{ht} . (b) The corresponding partition of the sample space and the step function $g(\mathbf{x}, T_b, M_b)$.

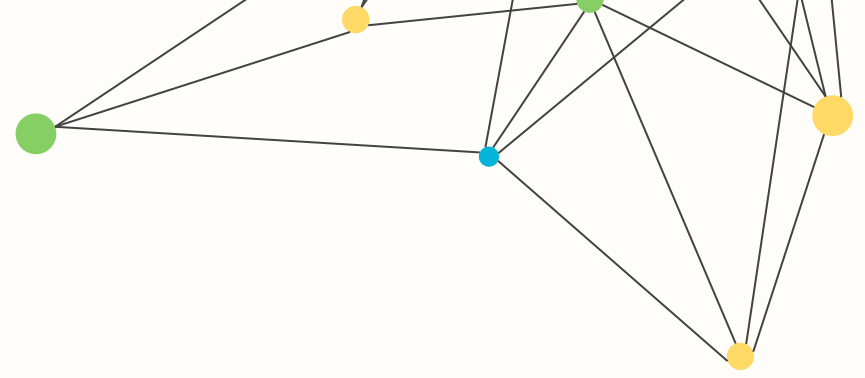


Overarching Goal

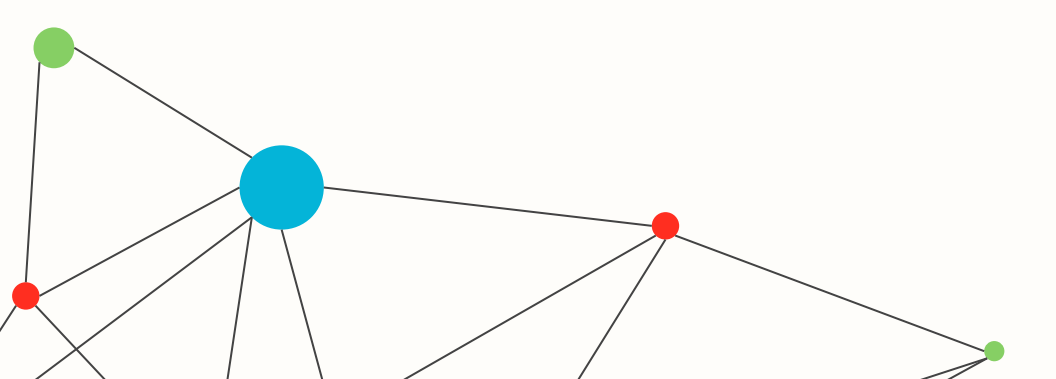
- Some work has been done to combine the mixed modeling approach of GLMMs and predictive performance of tree based methods:
 - Mixed Effects Regression Trees (MERT), (Hajjem 2011)
 - Mixed Effects Regression Forest (MERF), (Hajjem 2014)
 - mixed BART, (Spanbauer 2021)
- However, *they all fail to address the third limitation of GLMMs in that they cannot predict out-of-sample.*



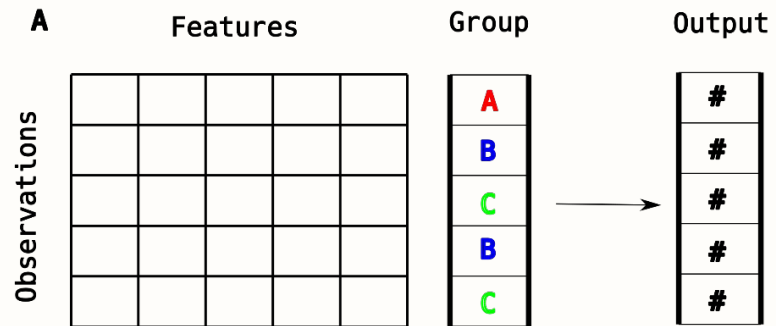
Our Goal: Develop a tree-based method with good predictive performance on out-of-sample groups.



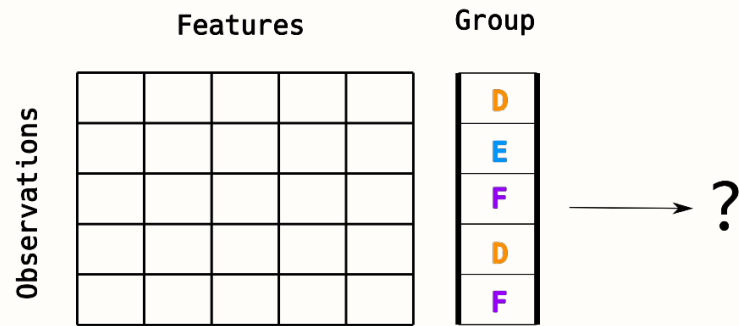
Our Method



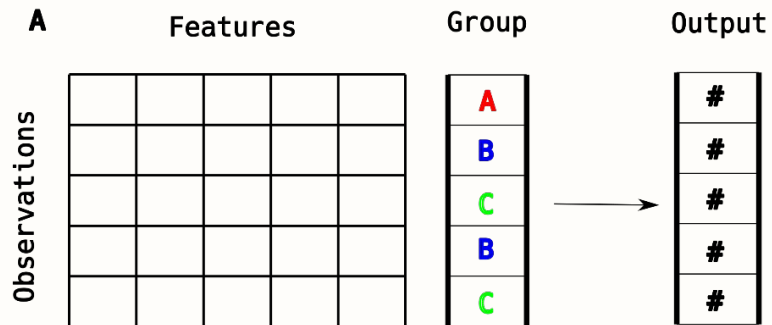
Training Data



Test Data

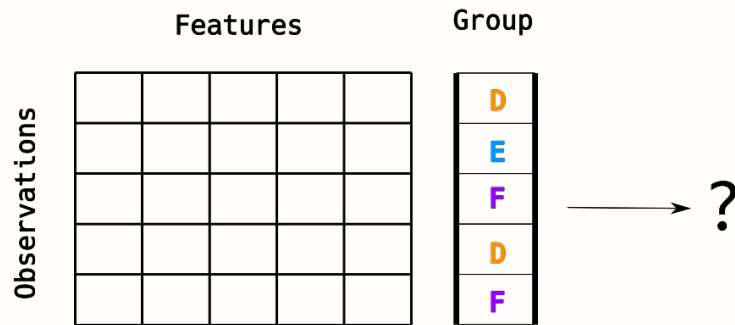


Training Data



B Construct $f : X \rightarrow \text{group}$
and extract group probabilities

Test Data

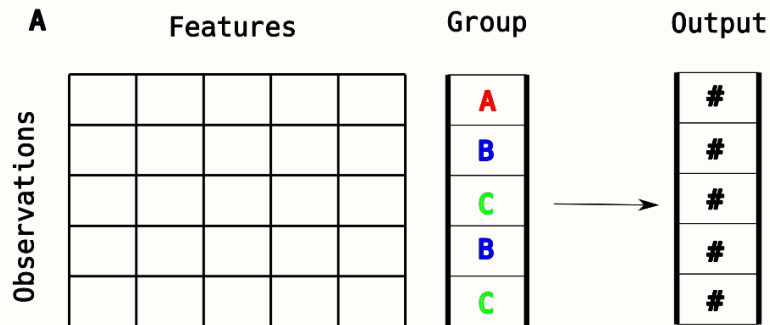


$$\mathbb{P}(A) = 0.90$$

$$\mathbb{P}(B) = 0.05$$

$$\mathbb{P}(C) = 0.05$$

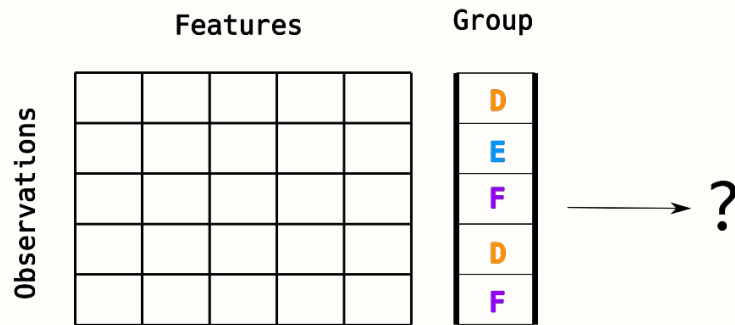
Training Data



B Construct $f : X \rightarrow \text{group}$
and extract group probabilities

C Construct trees $f_A, f_B, \dots$ on all
groups. Calculate final output via
mixture with group probabilities
as weights.

Test Data



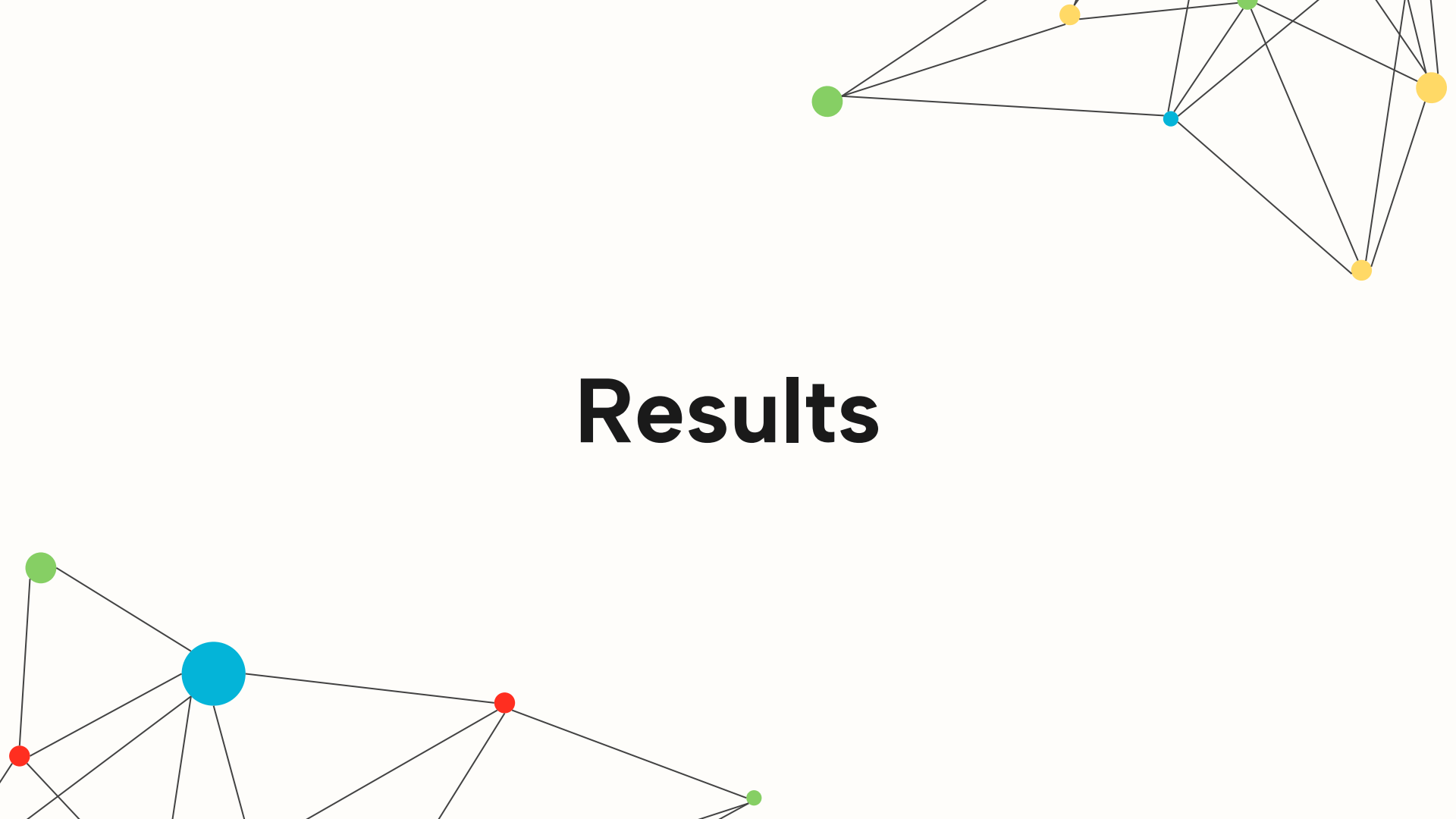
$$\mathbb{P}(A) = 0.90$$

$$\mathbb{P}(B) = 0.05$$

$$\mathbb{P}(C) = 0.05$$

$$y = 0.9f_A(X) + 0.05f_B(X) + 0.05f_C(X)$$

Results



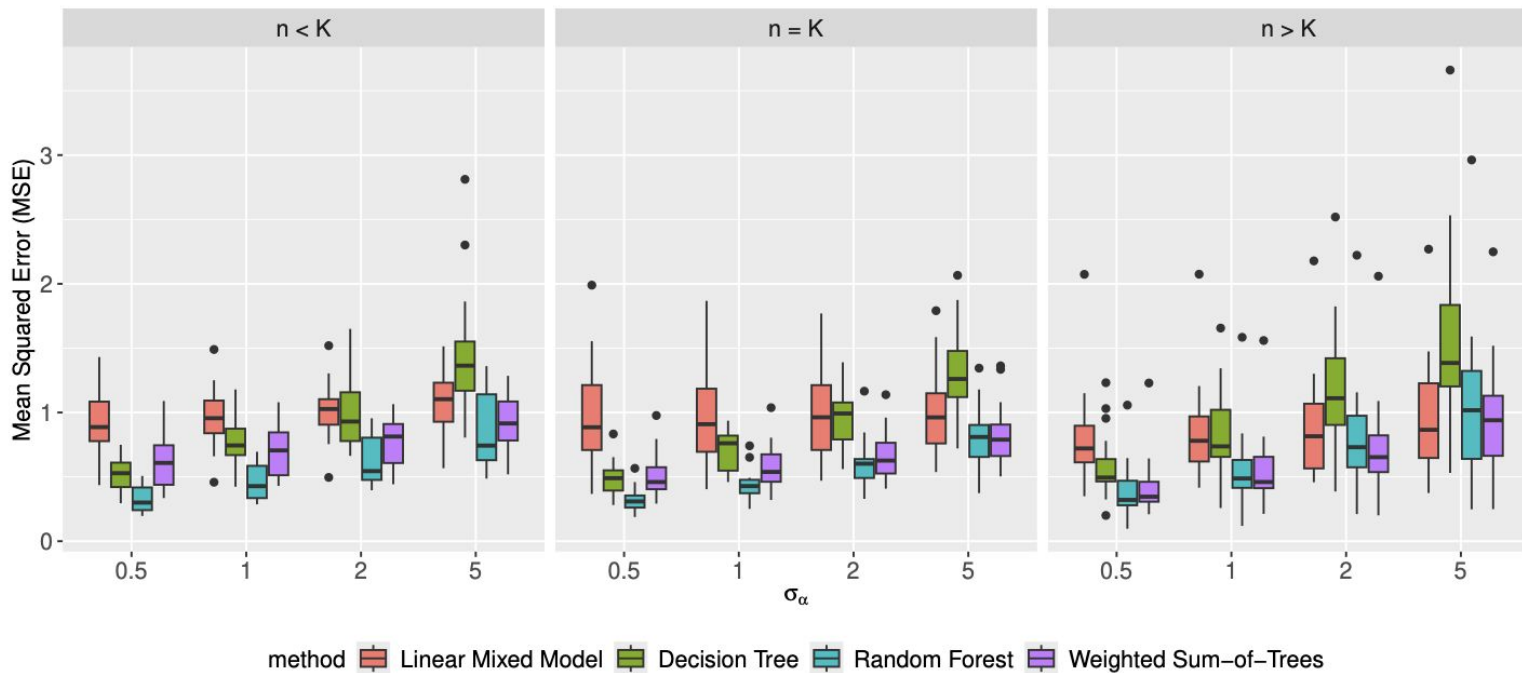
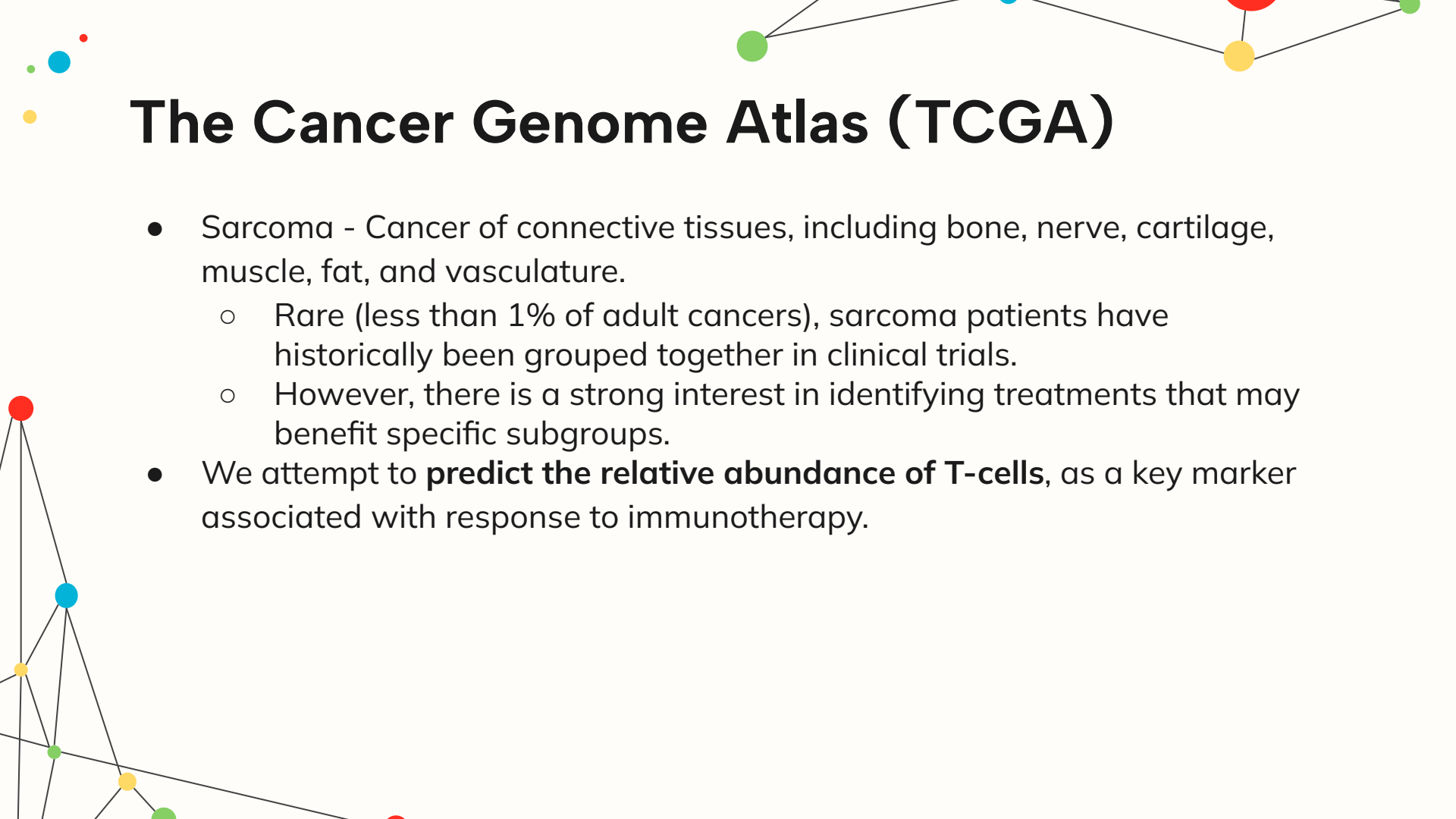
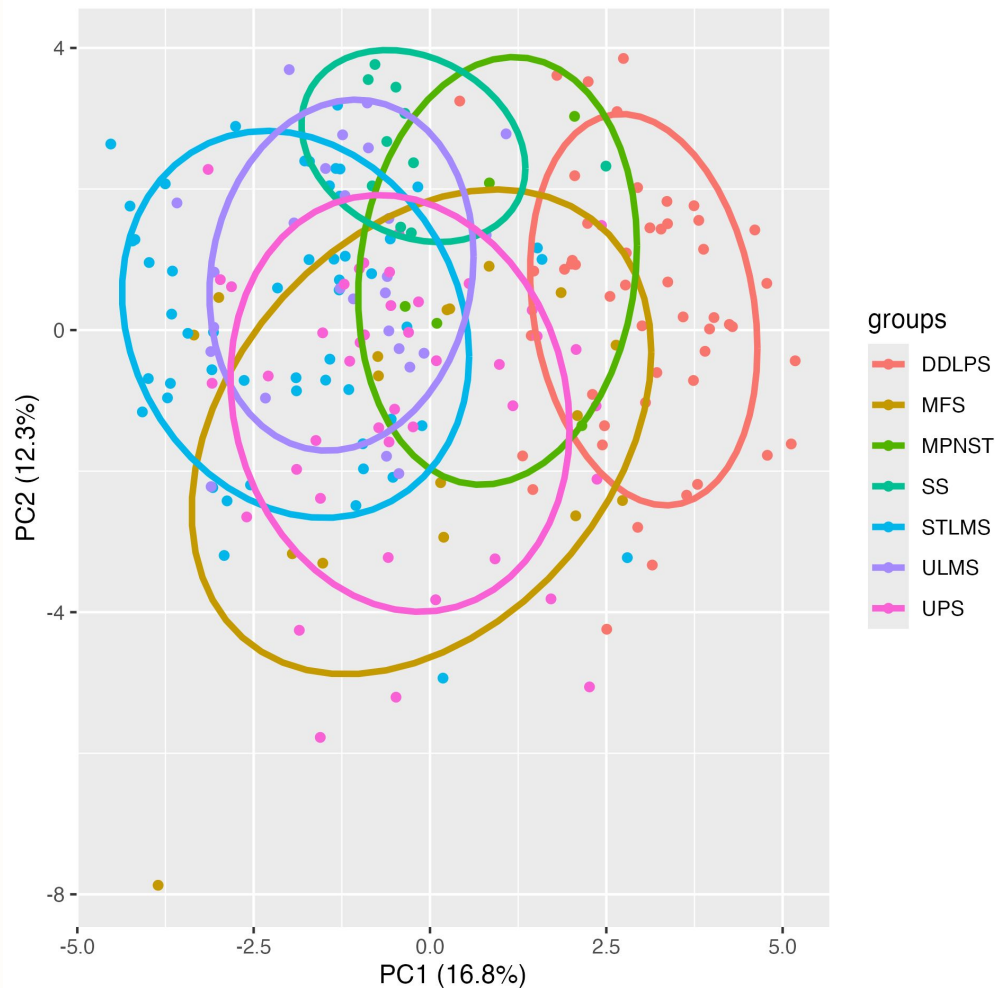


Fig. 2: Mean squared error (MSE) over a range of noise values σ_α for settings where $\mathbf{U} = \mathbf{I}$. The scale of the response variable y is standardized. Each boxplot represents MSE values across 20 simulated data sets. Simulations were performed for the settings $n = 10$, $K = 40$ (left), $n = K = 20$ (center), and $n = 40$, $K = 10$, (right).



The Cancer Genome Atlas (TCGA)

- Sarcoma - Cancer of connective tissues, including bone, nerve, cartilage, muscle, fat, and vasculature.
 - Rare (less than 1% of adult cancers), sarcoma patients have historically been grouped together in clinical trials.
 - However, there is a strong interest in identifying treatments that may benefit specific subgroups.
- We attempt to **predict the relative abundance of T-cells**, as a key marker associated with response to immunotherapy.

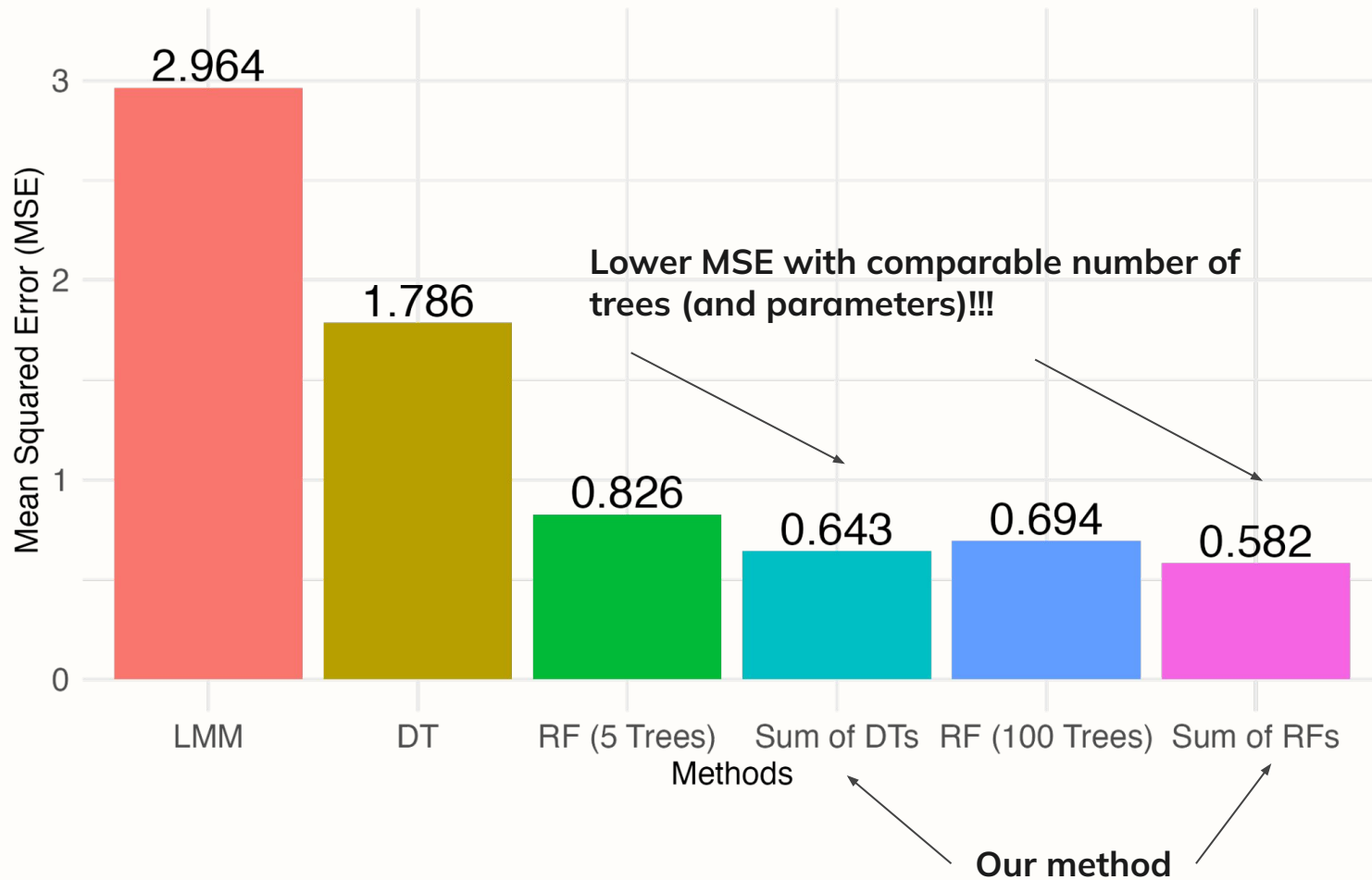


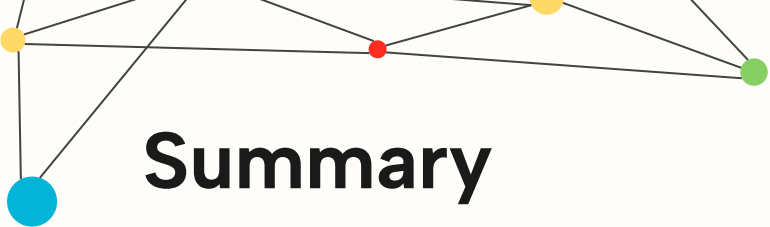
Data: (206 patients x 37 Features)

7 Subtypes of Sarcoma:

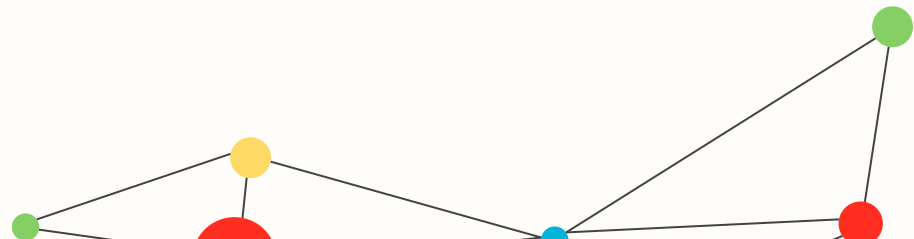
- 80 leiomyosarcoma (LMS)
 - 53 soft tissue LMS (STLMS)
 - 27 gynecologic LMS (ULMS)
- 50 dedifferentiated liposarcoma (DDLPS)
- 44 undifferentiated pleomorphic sarcoma (UPS)
- 17 myxofibrosarcoma (MFS)
- 10 synovial sarcoma (SS)
- 5 malignant peripheral nerve sheath tumor (MPNST)

Comparison of Methods on TCGA Sarcoma Data



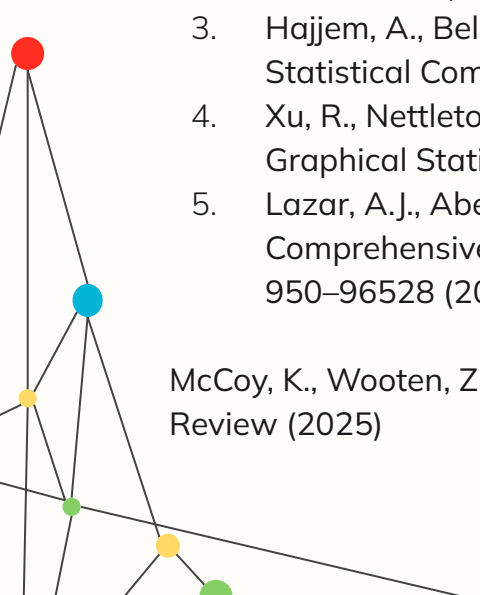


Summary

- **Clustered data are incredibly common in clinical, education, and social science research.**
 - **Our work shows that constructing decision trees and forests can be improved when clustered or grouped observations are present.**
 - **Furthermore, our method performs well out-of-sample in both simulated and real-world data problems.**
- 



References



1. Hajjem, A., Bellavance, F., Larocque, D.: Mixed effects regression trees for clustered data. *Statistics & Probability Letters* 81(4), 451–459 (2011)
2. Spanbauer, C., Sparapani, R.: Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. *Statistics in Medicine* 40(11), 2665–2691 (2021)
3. Hajjem, A., Bellavance, F., Larocque, D.: Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84(6), 1313–1328 (2014)
4. Xu, R., Nettleton, D., Nordman, D.J.: Case-specific random forests. *Journal of Computational and Graphical Statistics* 25(1), 49–65 (2016)
5. Lazar, A.J., Abeshouse, A., Adebamowo, C., Adebamowo, S.N., Akbani, R., Akeredolu, T., et al.: Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* 171(4), 950–965 (2017)

McCoy, K., Wooten, Z., Tomczak, K., Peterson, C.: Weighted Sum-of-Trees Model for Clustered Data. *Under Review* (2025)

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under **Grant No. 1842494**, as well as The Ken Kennedy Institute Computational Science and Engineering Recruiting Fellowship, funded by the Energy HPC Conference. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.



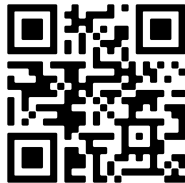
National
Science
Foundation

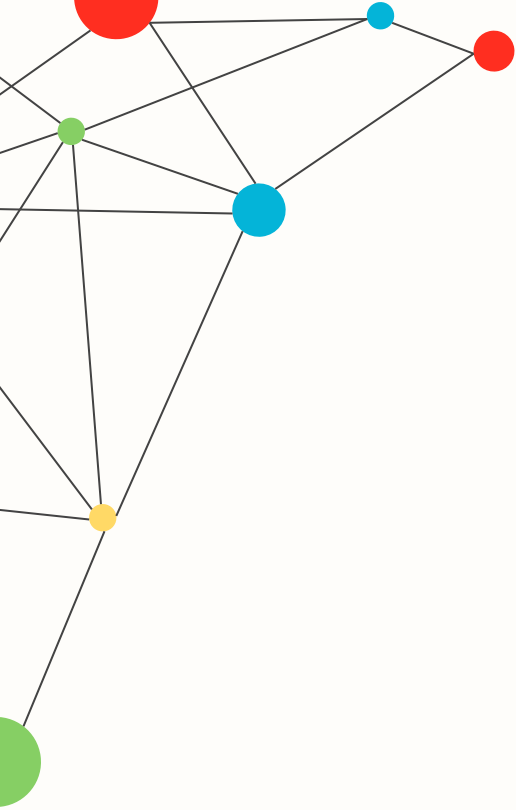


RICE KEN KENNEDY
INSTITUTE
AI, Data, and Computing for Global Impact

Q&A

Email: kmm12@rice.edu
Website: kmccoy.net ->





CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

