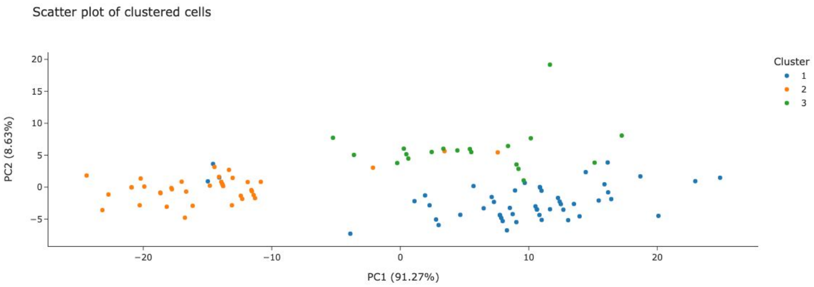
cluster_pval

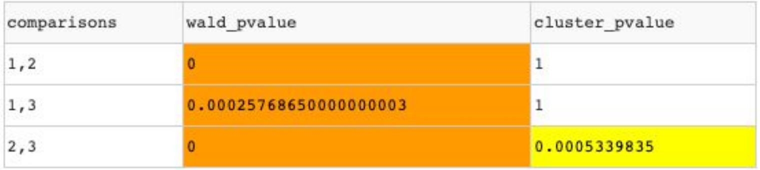
S. Jannetty, C. Miller, A. Mounsey, S. Pollack & L. Droog

In a world without inflated Type 1 error...

- Unsupervised clustering is commonly used on scRNAseq data to sort cells into cell types.
- Standard tests for differences in means between cell type clusters ignore that clusters were inferred from the data, inflating Type 1 error
- Gao et al (2021) proposed a new method for calculating p values that controls for this type 1 error
- We created a tool that clusters scRNAseq data and calculates adjusted and unadjusted p values for differences in means between clusters



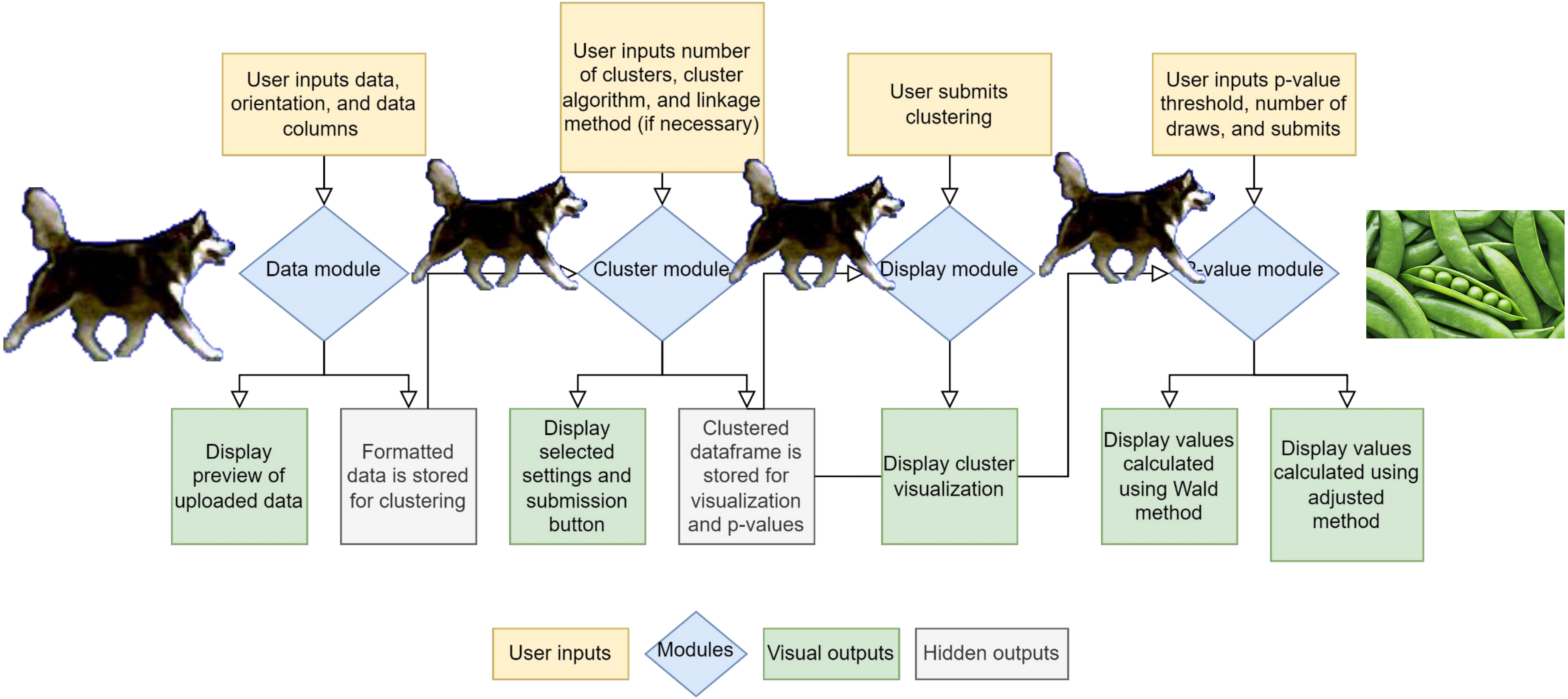
Scatter plot of clustered cells



comparisons	wald_pvalue	cluster_pvalue
1,2	0	1
1,3	0.00025768650000000003	1
2,3	0	0.0005339835

Demo

Design



User inputs data, orientation, and data columns

Data module

Display preview of uploaded data

Formatted data is stored for clustering

User inputs number of clusters, cluster algorithm, and linkage method (if necessary)

Cluster module

Display selected settings and submission button

Clustered dataframe is stored for visualization and p-values

User submits clustering

Display module

Display cluster visualization

User inputs p-value threshold, number of draws, and submits

P-value module

Display values calculated using Wald method

Display values calculated using adjusted method

User inputs

Modules

Visual outputs

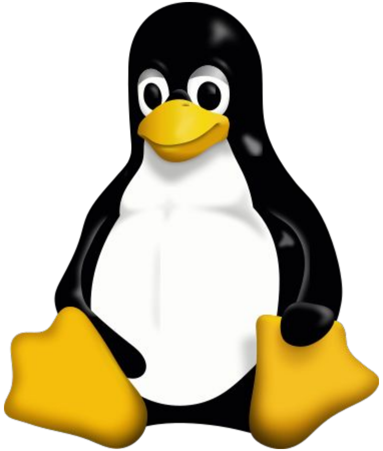
Hidden outputs

Technologies employed

- Pandas to import and process data
- SciKit Learn for hierarchical and k-means clustering
- Plotly to create figures
- Dash to create GUI
- Unittests for testing
- Heroku to create Dash app
- Git action for continuous integration

Data used for testing

- Penguin data [1]
- scRNAseq data for tests using representative data [2 & 3]



[1] Horst AM, Hill AP, Gorman KB (2020). Palmer penguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. [https://allisonhorst.github.io/palmerpenguins/]. Doi:10.5281/zenodo.3960218.

[2] Gala HP, Lanctot A, Jean-Baptiste K, Guiziou S et al. A single-cell view of the transcriptome during lateral root initiation in Arabidopsis thaliana. Plant Cell 2021 Aug 13;33(7):2197-2220. PMID: 33822225

[3] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., ... & Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. Nature communications, 8(1), 1-12.

Collaboration strategies

- Slack channel
  - Easy communication in different areas of the project
  - Able to troubleshoot and set up meeting times

- Meetings before class

- No branching for the ClusterClub since we were working in individual modules

320 commits

Challenges

- Were not able to use Travis CI
- Some issues with merge conflicts
- Lack of ability to test GUI through unit testing
- Troubles with getting R and Python to cooperate
- Slow when using large datasets
  - Paywall issues with Heroku
- Stochastic nature of adjusted p value calculation

Lessons Learned

- Don't use Travis CI
- Division of labor and good communication make for successful teamwork
- How to structure a python package with modules and tests
- How to structure a GitHub repository so that other people can use it
- About the variety of technologies available for clustering and building GUIs

PYTORCH

Seurat

fastcluster: Fast Hierarchical Clustering Routines for R and 'Python'

Thanks for watching!
Go Dawgs!