Interspecies machine learning: Using animal training data to classify humans neurons

LUKA BLAGOJEVIC, DEPARTMENT OF NETWORK AND DATA SCIENCE – 1.4.2021.

DATA MINING AND BIG DATA ANALYTICS PROJECT

1. Introduction and dataset

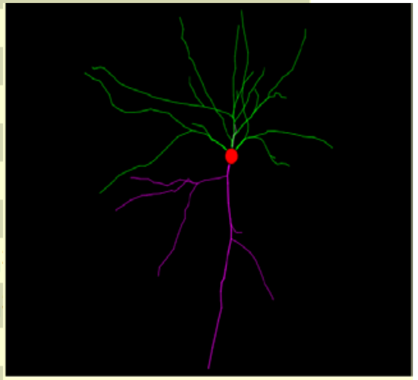
- **Project goal: By using animal training, classify human neuron cell types**
- **Why?** A lot of easily obtainable animal data – useful since human data is harder to obtain, control for and it's more expensive
- Neuromorpho.org - animal and human neuron data

(22 features, I removed highly correlated ones – 19 used)

- Two classes:
  1) **primary cell type** (13 of them)
  2) **secondary cell type** (31 of them)

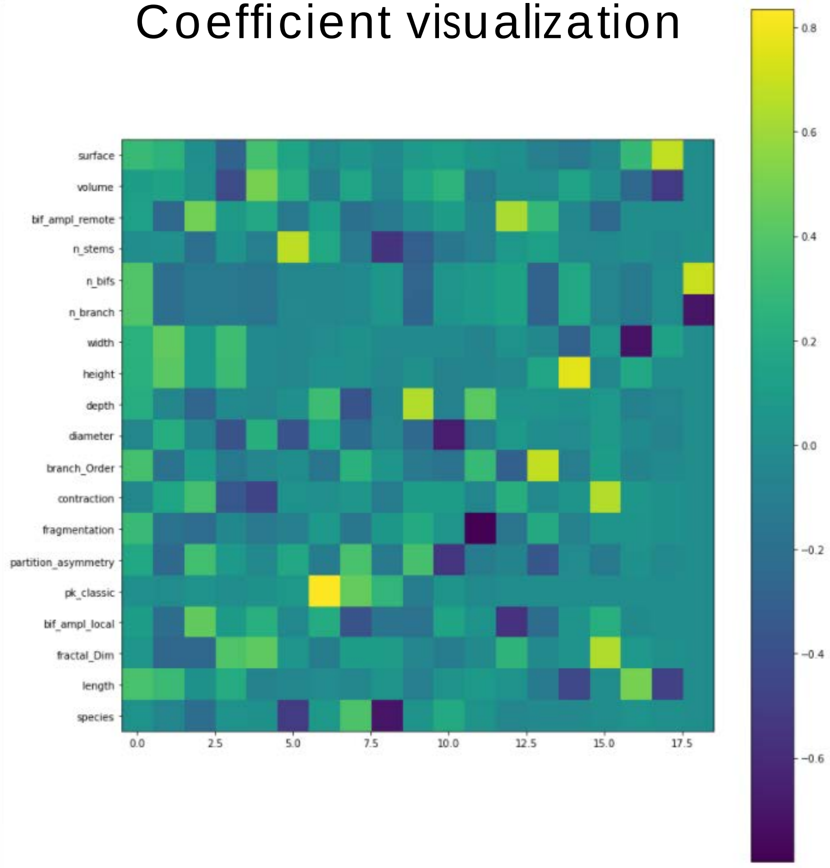
Species	Number of data points
Human	7150
Monkey	2950
Rat	7450
Mouse	7450

Measurements	
Soma Surface :	1350.39 μm^2
Number of Stems :	5
Number of Bifurcations :	19
Number of Branches :	43
Overall Width :	306.94 μm
Overall Height :	409.98 μm
Overall Depth :	107.38 μm
Average Diameter :	1.06 μm
Total Length :	3041.4 μm
Total Surface :	10207.5 μm
Total Volume :	4050.87 μm
Max Euclidean Distance :	283.54 μm
Max Path Distance :	292.86 μm
Max Branch Order :	6
Average Contraction :	0.95
Total Fragmentation :	248
Partition Asymmetry :	0.51
Average Rall's Ratio :	1.22
Average Bifurcation Angle Local :	60.19°
Average Bifurcation Angle Remote :	60.07°
Fractal Dimension :	1.01

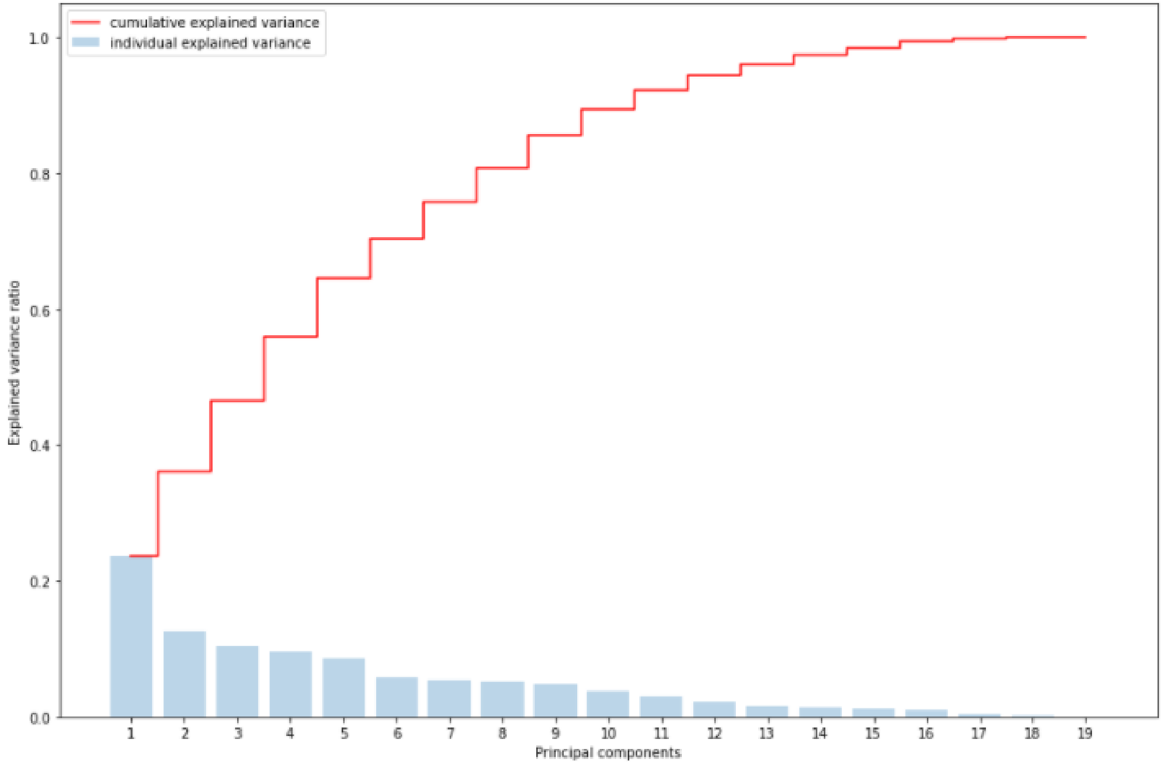


2. Principal component analysis

Coefficient visualization

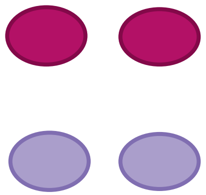


Explained variance



3. Do classes, across species, refer to morphologically similar things?

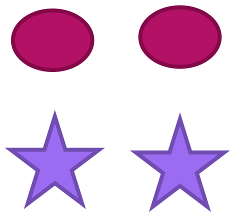
Hypothesis: Class labels refer to the similar morphological objects across species **(same name – same shape)**



Human data: Neuron class A

Mouse data: Neuron class A

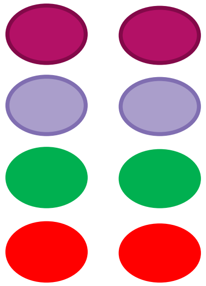
Rejected hypothesis **(same name – different shape)**



Human data: Neuron class A

Mouse data: Neuron class A

Intraclass similarity between species **TEST – Do class labels referal to objects in of similar neuromorphological shape across species**



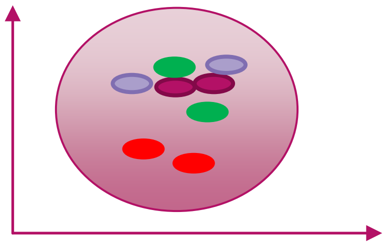
Human data: Neuron class A

Human data: Neuron class A

Human data: Neuron class A

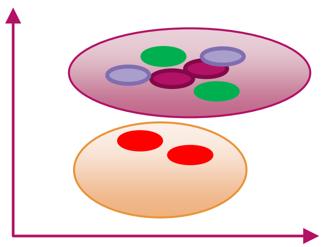
Human data: Neuron class A

1 cluster found

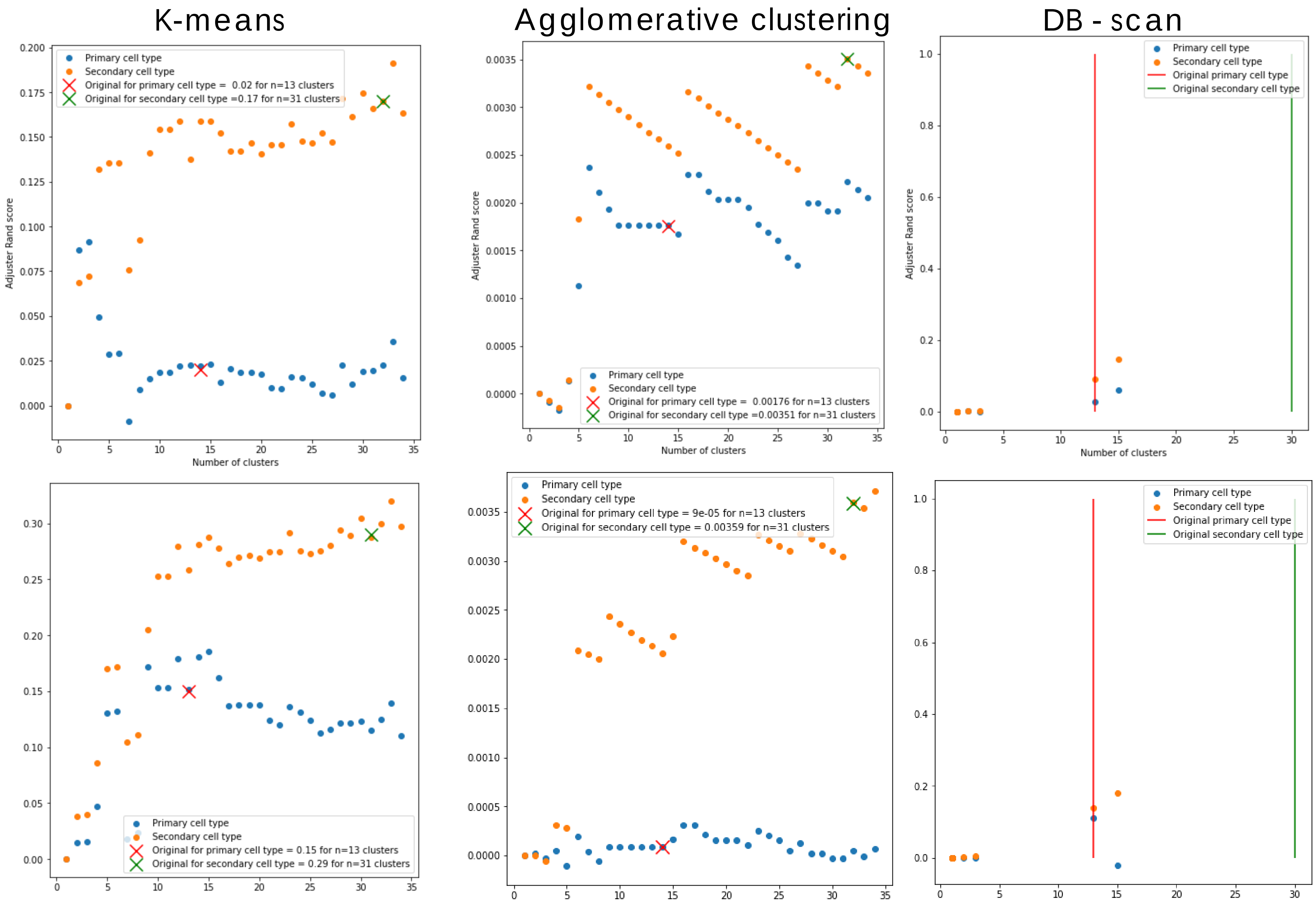


Hypothesis **true**

2 or more clusters found



Hypothesis **rejected**



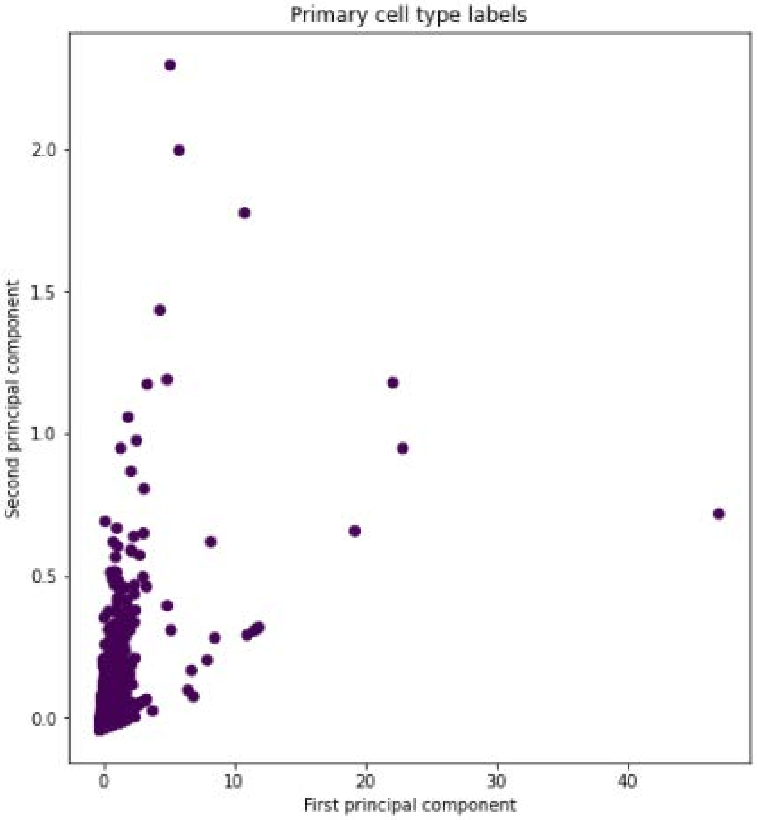
Adjusted Rand score
0 – random clustering
1 – perfect match (clustered classes and original classes)

Adjusted Mutual Information score
0 – random clustering
1 – perfect match (clustered classes and original classes)

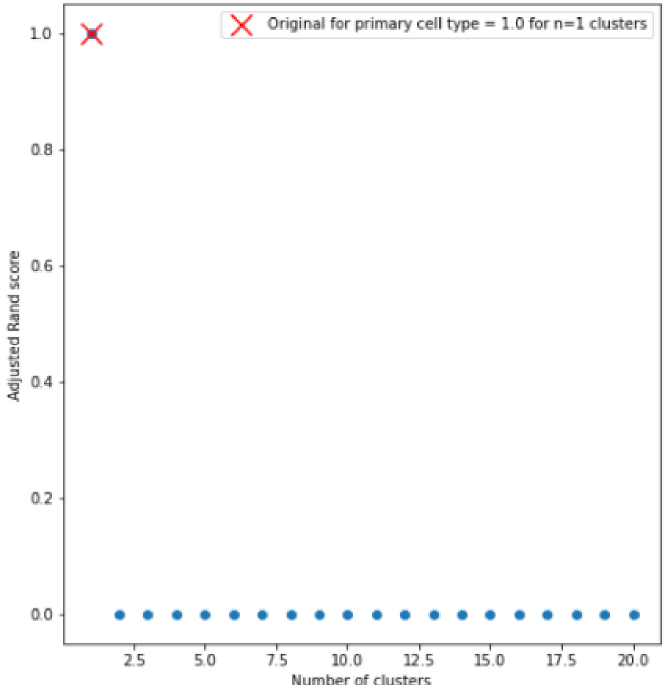
Start with 13 primary cell types that are common to humans and animals

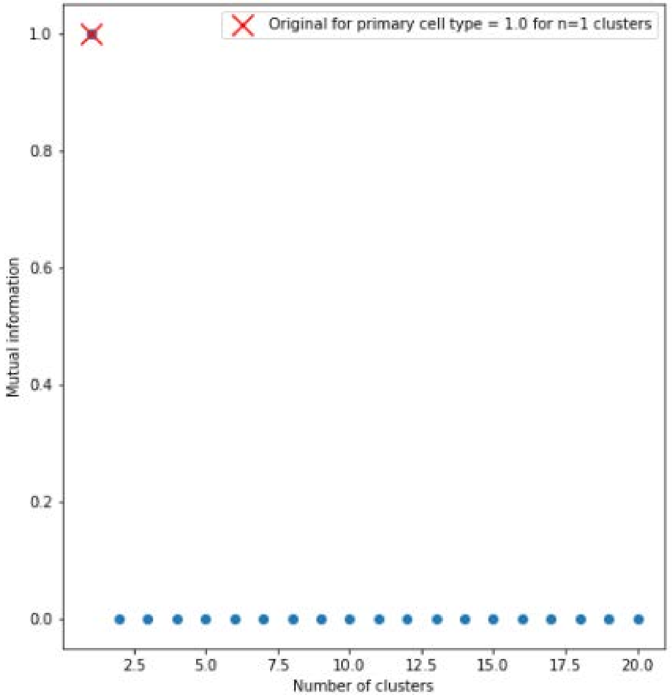
Filter the data, one cell type per run (and >20 samples)

Select next class to compute



Compute k-means for $k \in [1,20]$





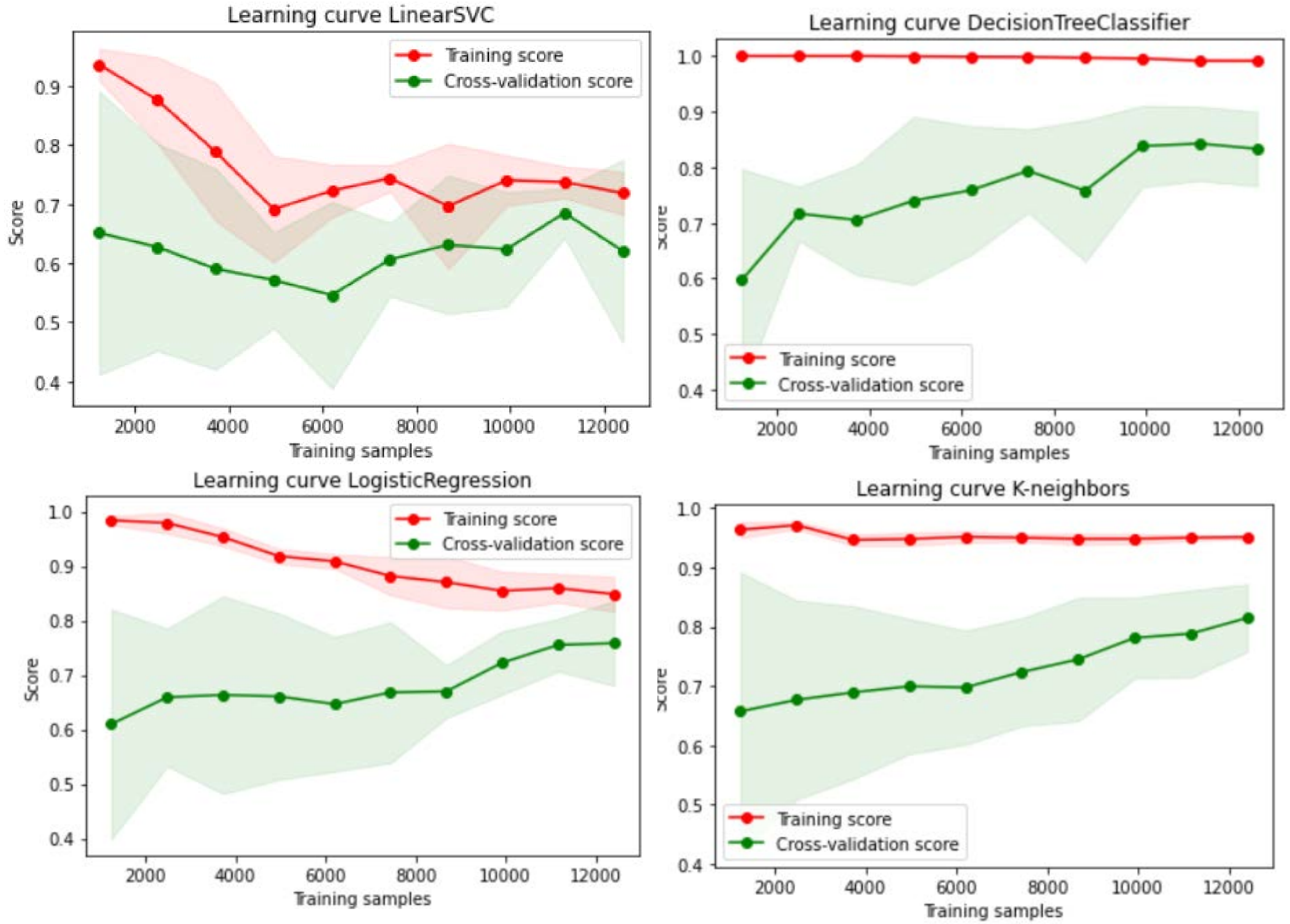
I obtained 1 cluster for each of cell classes

Hypothesis is true

4. Hyperparameter optimization and learning curve for classifiers

- With Grid search, I found optimal hyperparameters for 4 classifiers on mixed data:

1. Linear SVC
2. Logistic Regression
3. Decision Trees
4. K-neighbors Decision Trees and K-neighbors tended to overfit, so I didn't use them for the final evaluation



7. Results – Test data: Human neurons

It has potential, but not as good as human training data

Linear SVC

Training data: Human

	precision	recall	f1-score	support
principal cell	0.94	0.95	0.94	11000
pyramidal	0.90	0.94	0.92	4910
sensory	0.98	0.87	0.92	960
interneuron	0.78	0.40	0.53	420
glia	0.46	0.12	0.19	100
magnopyramidal	0.95	0.93	0.94	350
von economo neuron	0.87	0.47	0.61	70
accuracy			0.93	17810
macro avg	0.84	0.67	0.72	17810
weighted avg	0.92	0.93	0.92	17810

Training data: Animal

	precision	recall	f1-score	support
principal cell	0.42	0.38	0.40	4397
pyramidal	0.00	0.00	0.00	1963
sensory	0.26	0.71	0.39	383
interneuron	0.47	0.16	0.24	167
glia	0.02	1.00	0.04	40
magnopyramidal	0.00	0.00	0.00	142
von economo neuron	0.00	0.00	0.00	29
accuracy			0.28	7121
macro avg	0.17	0.32	0.15	7121
weighted avg	0.28	0.28	0.27	7121

Training data: Monkey

	precision	recall	f1-score	support
principal cell	0.60	0.71	0.65	4397
pyramidal	0.49	0.22	0.30	1963
sensory	0.00	0.00	0.00	383
interneuron	0.67	0.01	0.02	167
glia	0.04	0.97	0.07	40
magnopyramidal	0.00	0.00	0.00	142
von economo neuron	0.00	0.00	0.00	29
accuracy			0.50	7121
macro avg	0.26	0.27	0.15	7121
weighted avg	0.52	0.50	0.49	7121

Log Reg.

Training data: Human

	precision	recall	f1-score	support
principal cell	0.94	0.95	0.95	11000
pyramidal	0.90	0.94	0.92	4910
sensory	0.98	0.88	0.92	960
interneuron	0.79	0.40	0.53	420
glia	0.48	0.10	0.17	100
magnopyramidal	0.96	0.93	0.94	350
von economo neuron	1.00	0.50	0.67	70
accuracy			0.93	17810
macro avg	0.86	0.67	0.73	17810
weighted avg	0.92	0.93	0.92	17810

Training data: Animal

	precision	recall	f1-score	support
principal cell	0.33	0.25	0.28	4397
pyramidal	0.00	0.00	0.00	1963
sensory	0.15	0.93	0.26	383
interneuron	0.70	0.17	0.27	167
glia	0.03	1.00	0.05	40
magnopyramidal	0.00	0.00	0.00	142
von economo neuron	0.00	0.00	0.00	29
accuracy			0.21	7121
macro avg	0.17	0.34	0.12	7121
weighted avg	0.23	0.21	0.20	7121

Training data: Monkey

	precision	recall	f1-score	support
principal cell	0.62	0.75	0.68	4397
pyramidal	0.50	0.21	0.30	1963
sensory	0.00	0.00	0.00	383
interneuron	0.00	0.00	0.00	167
glia	0.04	1.00	0.08	40
magnopyramidal	0.00	0.00	0.00	142
von economo neuron	0.00	0.00	0.00	29
accuracy			0.53	7121
macro avg	0.17	0.28	0.15	7121
weighted avg	0.52	0.53	0.50	7121

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

High recall – negatives are most likely (true) **negatives**

Low recall – negatives are most likely **not** (true) **negatives**

High precision – positives are most likely are (true) **positives**

Low precision – positives are most likely **not** (true) **positives**

Thank you for the attention!