

Getting CLARITY

Albert Lee

Sahithi Garikapati

Wei Wang

The Art of Data Science
3/10/2016

OPPORTUNITY

❖ CLARITY

- Chemical transformation of intact biological tissues into a hydrogel-tissue hybrid
- Studied under light
- Retains fine structure and native biological molecules
- Extracts information without disassembly

- ❖ Analyzing **CLARITY** data will provide the much-needed across-scales standardization on a global perspective, leading to fundamental insights into psychiatric disease.

SIGNIFICANCE

- ❖ Psychiatric disease represents the leading cause of disability both in the U.S. and worldwide!
- ❖ Lack of circuit-level understanding of psychiatric diseases is causing pharmaceutical companies to shut down psychiatry programs.
- ❖ The effort made to study **CLARITY** models has the potential to influence the lives of millions of people across the globe!!!

GAP

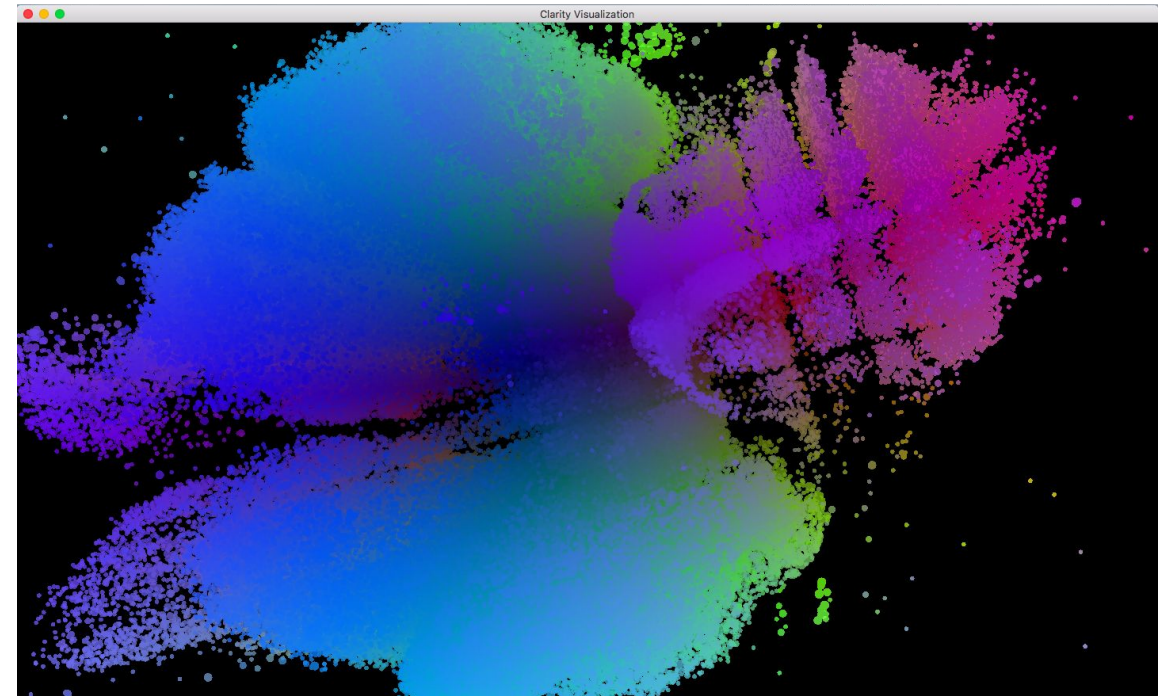
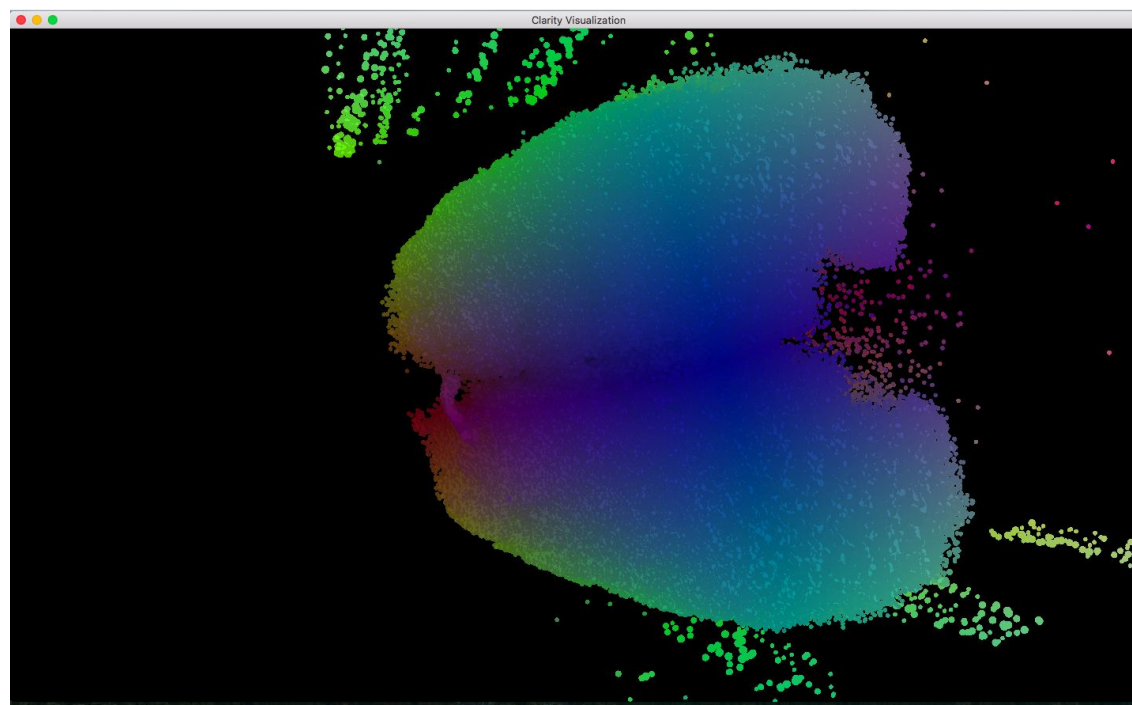
- ❖ Presently no other imaging techniques have the ability to link causal information with behavioral significance of connections in the brain.
- ❖ No standardized parameters exist to distinguish behavior based on neural structural dynamics.

CHALLENGE

- ❖ Massive datasets leading to computational burden
- ❖ Small number of datasets
- ❖ Lack of prior studies on **CLARITY** data
- ❖ Difficulty in processing data

Obtaining statistically significant results may be difficult.

CLARITY image
visualization
of a rat brain
under the
influence of
cocaine.



FORMAL STATEMENT OF PROBLEM

Model

$F(i,j) \sim$ Clarity image of a particular section

$$F(l,k) = \{F(l|k) * F(k)\}$$

$F(k) = \{0,1,2\} \rightarrow 0 = \text{Control Image}, 1 = \text{Cocaine Image}, 2 = \text{Fear Image}$

Histogram Equalization

$$F(i,j) = \text{floor}((L-1) \sum F(i,j) P_n)$$

P_n = Number of pixels of intensity n / Total number of pixels

$F(l) = \text{mean}(F(i,j))$ % mean gray value in the image

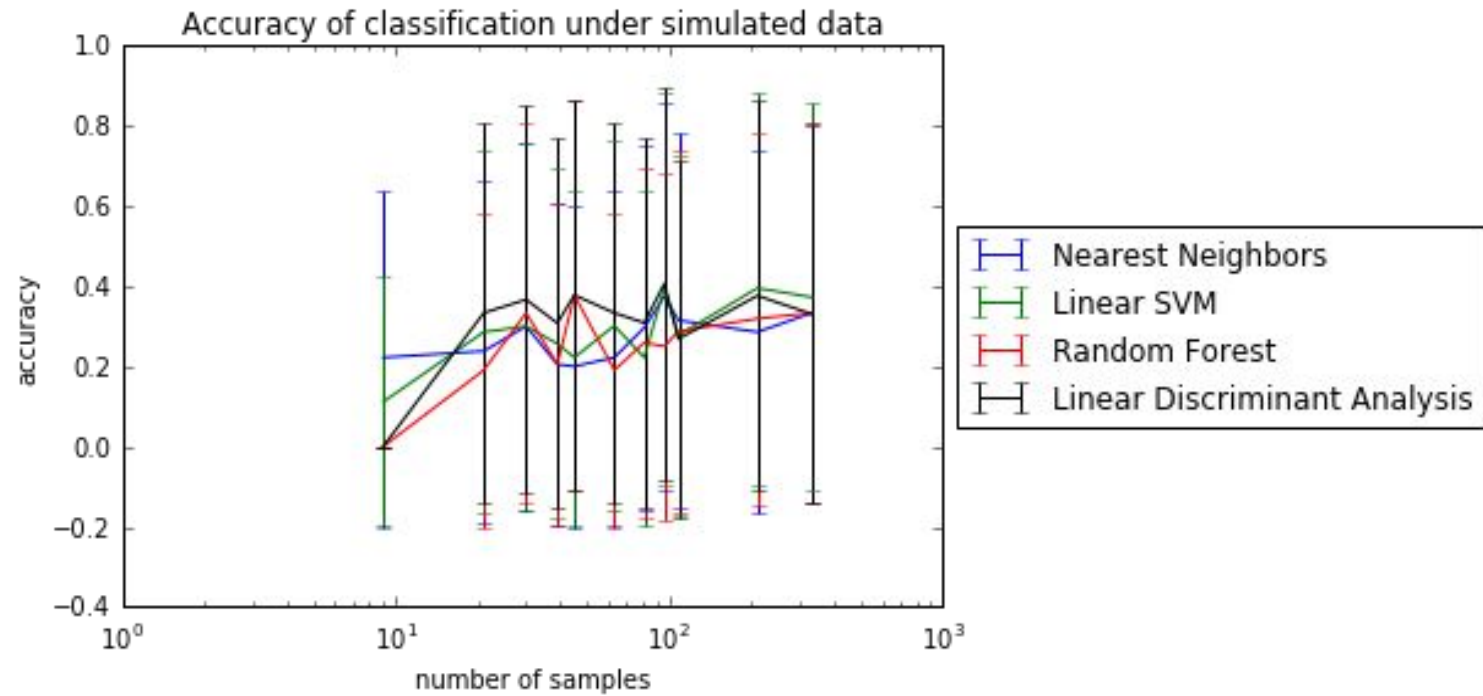
MODEL ASSUMPTIONS

- ❖ Histograms are sampled according to: $x_i \sim \text{iid} F$
- ❖ Independent data points: $F_{X|0} = \text{Norm}(\mu_0, \sigma_0) V \times V$.
- ❖ Class conditional difference across conditions = $\{\text{Control, Cocaine, Fear}\}$

FORMAL STATEMENT OF ALGORITHM

- ❖ Linear Discriminant Analysis = LDA
- ❖ Random Forest = RF
- ❖ K-Nearest Neighbors = KNN
- ❖ Support Vector Machine = SVM

RESULTS



Accuracy of classification with real data:

Accuracy of Nearest Neighbors: 0.00 (+/- 0.00)

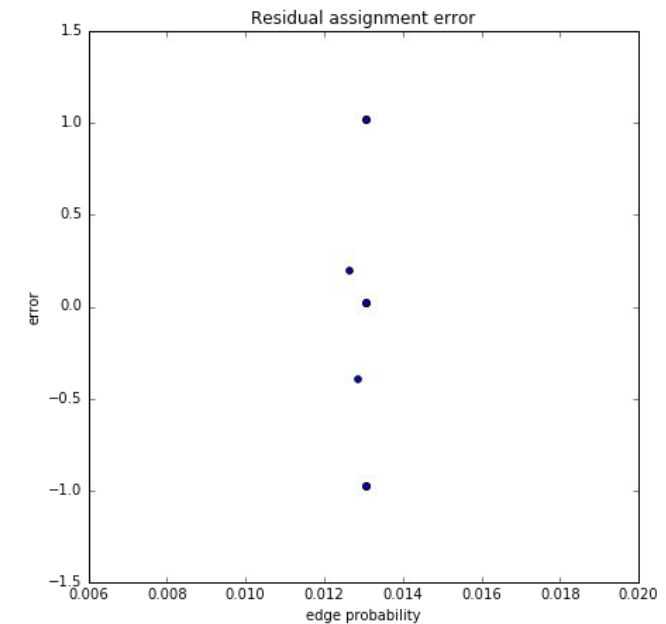
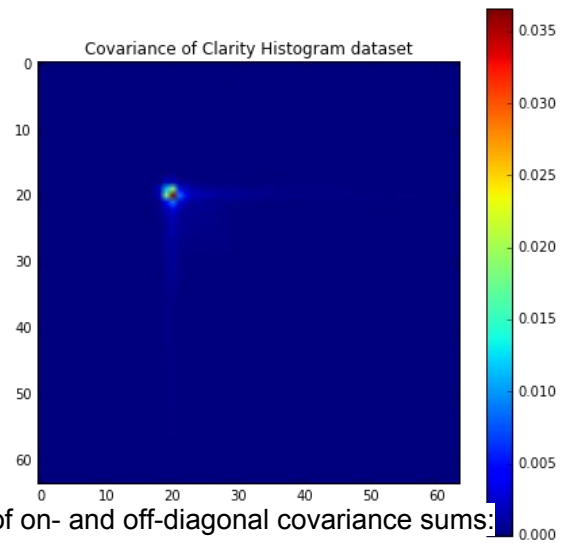
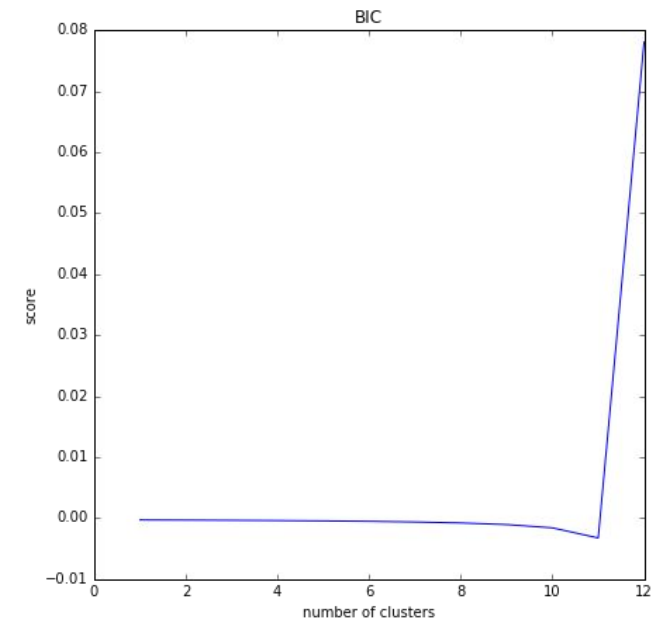
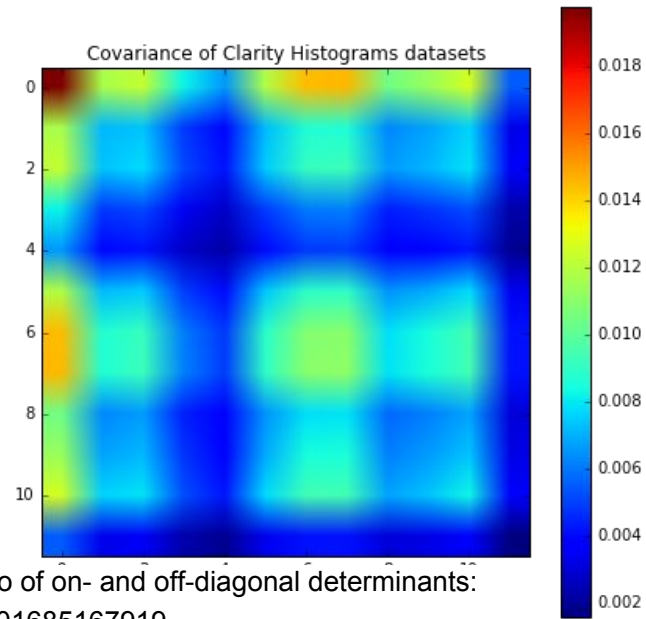
Accuracy of Linear SVM: 0.25 (+/- 0.87)

Accuracy of Random Forest: 0.08 (+/- 0.55)

Accuracy of Linear Discriminant Analysis: 0.33 (+/- 0.94)

The low accuracy with high error was expected due to the small sample size.

MODEL CHECKING



RESOLUTION

- ❖ Because of a small number of datasets most statistical tests do not produce meaningful results.
- ❖ We intend to return to exploratory data analysis and re-evaluate our model.
- ❖ Additionally, instead of histograms, 3D adjacency matrixes and/or axis-aligned graphs should be considered.