Correlated Histograms Clustering

A novel unsupervised learning technique that leverages the underlying statistics of a dataset across its different dimensions to identify cluster centroids

Brice Brosig, Lone Star Analysis

Co-Author: Randy Allen, Lone Star Analysis

Agenda

- Background
- Motivation
- Correlated Histograms Clustering
- Application – Training Syllabus Outcomes
- Application – Robustness Against Noisy Data
- Summary
- Future Work
- Acknowledgements

Background

- Supervised learning
  - Dataset with known, “ground truth” labels.
  - The Task is to fit a model to the data such that you fit new, unseen data well.
- Unsupervised learning (*clustering*)
  - The practitioner has a dataset *without* labels.
  - The task is typically to learn one or more of the following:
    - The number of classes each instance falls into.
    - Where those categories are in the domain of the dataset (borders and/or centroids).
    - What instances fall into which categories.
- Semi-Supervised learning
  - Combination of the two – some of the data is labeled and unsupervised learning tasks can aid in the supervised learning.

Motivation

- Unlabeled and / or noisy data
  - Practitioners mostly deal with data that is “messy”. From sensors to surveys, we must make use of data that is not easily visualized or categorized – if at all.
- No a priori knowledge of the number of clusters
  - We often can’t assume things about the data before hand. Often, the number of clusters is one of the things we want to find out when using a clustering technique.
- An interest in the centroids
  - Centroids express the actual characteristics of the different clusters. These characteristics can be more useful than just knowing which instances fall into what category.
- New clustering / neighborhood metric
  - Almost all other clustering techniques use *distance* as the metric to build clusters. This requires consideration and normalization of the data.

Motivation

Method name	Parameters	Geometry (metric used)
K-Means	number of clusters	Distances between points
Affinity propagation	damping, sample preference	Graph distance (e.g., nearest-neighbor graph)
Mean-shift	bandwidth	Distances between points
Spectral clustering	number of clusters	Graph distance (e.g., nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Any pairwise distance
DBSCAN	neighborhood size	Distances between nearest points
OPTICS	minimum cluster membership	Distances between points
Gaussian mixtures	many	Mahalanobis distances to centers
BIRCH	branching factor, threshold, optional global clusterer.	Euclidean distance between points
Bisecting K-Means	number of clusters	Distances between points

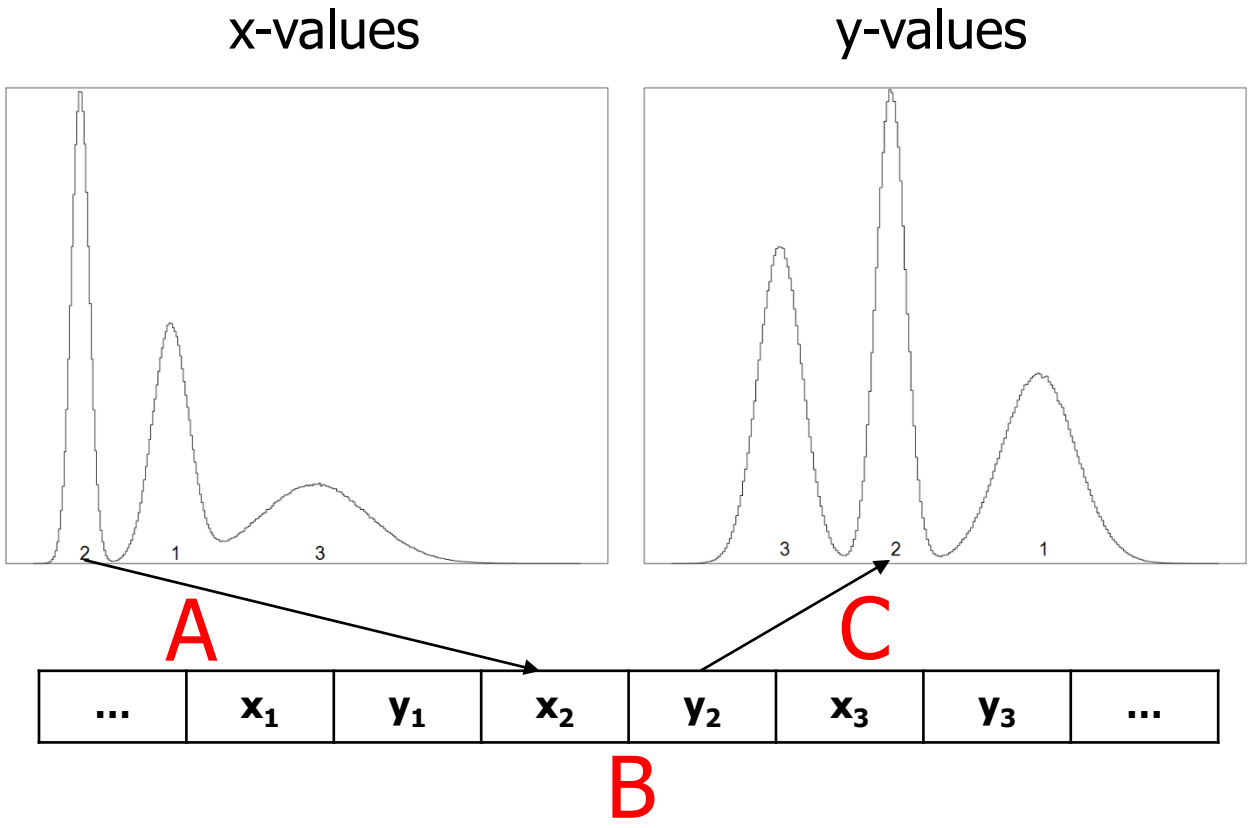
Correlated Histograms Clustering

Modes

A. Create a histogram or density of each dimension
   A. Prefer Harrel-Davis Quantile respective Density Estimate
B. Use those histograms to compute the modality and locations of each mode
   A. Prefer the Lowland Modality Technique to identify modes

Correlation

A. Pick up a mode from some dimension and find the nearest point
B. Look at the other components of that point
C. Find the nearest modes to the other components
D. Repeat step A for all modes and dimensions



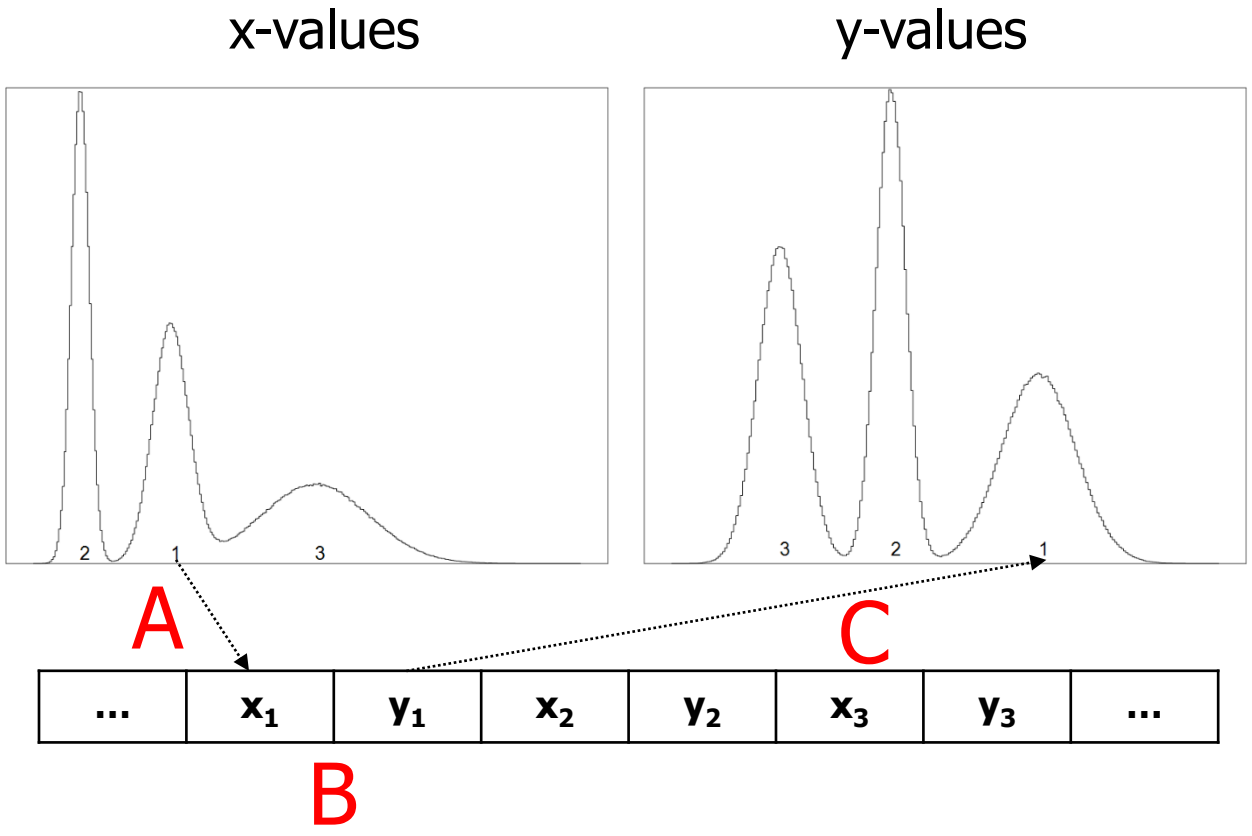
Correlated Histograms Clustering

Modes

A. Create a histogram or density of each dimension
   A. Prefer Harrel-Davis Quantile respective Density Estimate
B. Use those histograms to compute the modality and locations of each mode
   A. Prefer the Lowland Modality Technique to identify modes

Correlation

A. Pick up a mode from some dimension and find the nearest point
B. Look at the other components of that point
C. Find the nearest modes to the other components
D. Repeat step A for all modes and dimensions



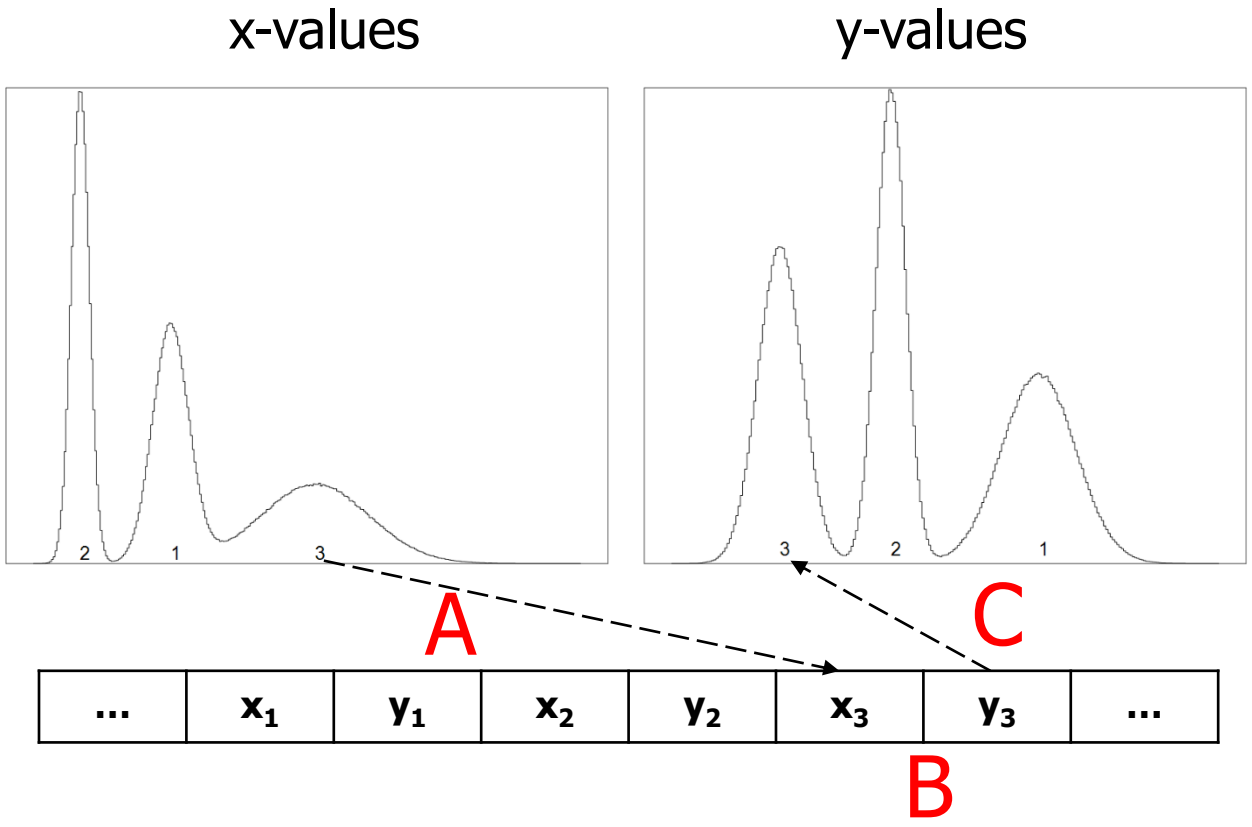
Correlated Histograms Clustering

Modes

A. Create a histogram or density of each dimension
   A. Prefer Harrel-Davis Quantile respective Density Estimate
B. Use those histograms to compute the modality and locations of each mode
   A. Prefer the Lowland Modality Technique to identify modes

Correlation

A. Pick up a mode from some dimension and find the nearest point
B. Look at the other components of that point
C. Find the nearest modes to the other components
D. Repeat step A for all modes and dimensions



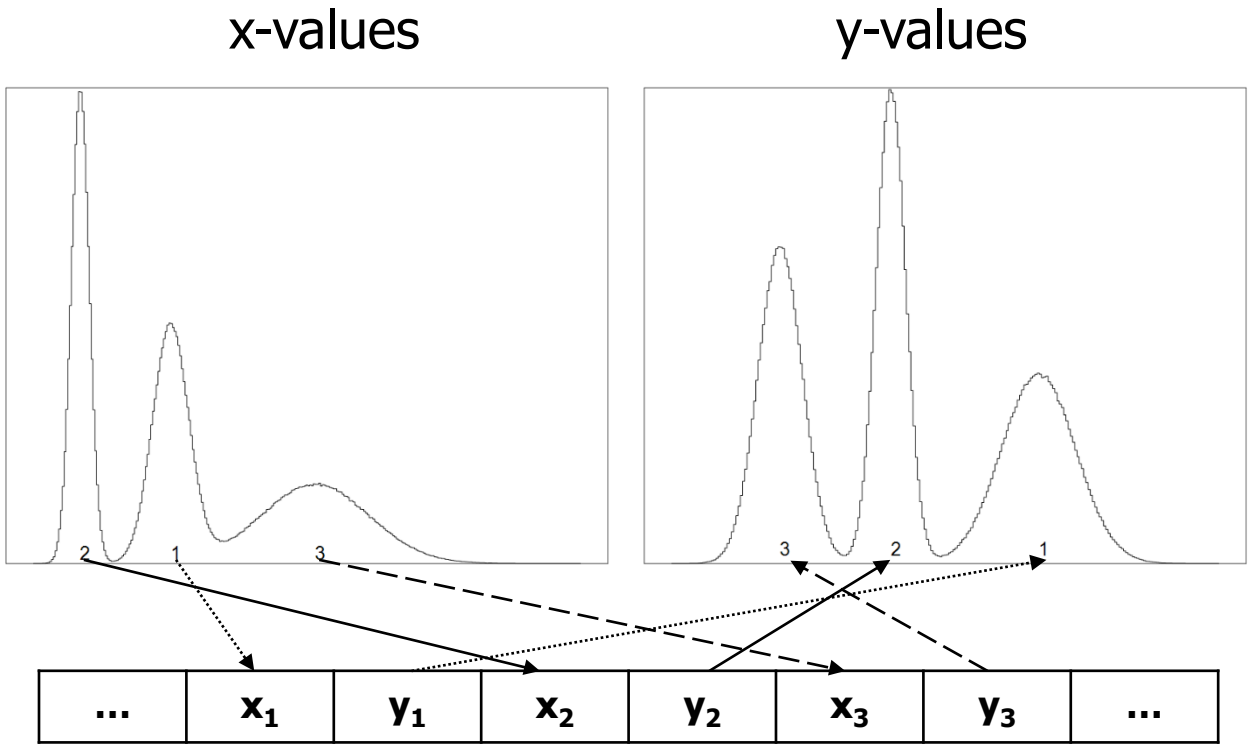
Correlated Histograms Clustering

Modes

A. Create a histogram or density of each dimension
   A. Prefer Harrel-Davis Quantile respective Density Estimate
B. Use those histograms to compute the modality and locations of each mode
   A. Prefer the Lowland Modality Technique to identify modes

Correlation

A. Pick up a mode from some dimension and find the nearest point
B. Look at the other components of that point
C. Find the nearest modes to the other components
D. Repeat step A for all modes and dimensions



* To find the modes we make use of Andrey Akinshin's Lowland Modality technique alongside the Harrel-Davis Quantile Respective Density Estimate.

Application – Training Syllabus Outcomes

- Suppose we have implemented a pilot training syllabus and evaluated trainees on several metrics.
- Knowing the number of outcomes and the characteristics of those outcomes can be useful.
- It is likely that one has more than 3 metrics to evaluate and therefore a visualization is difficult.

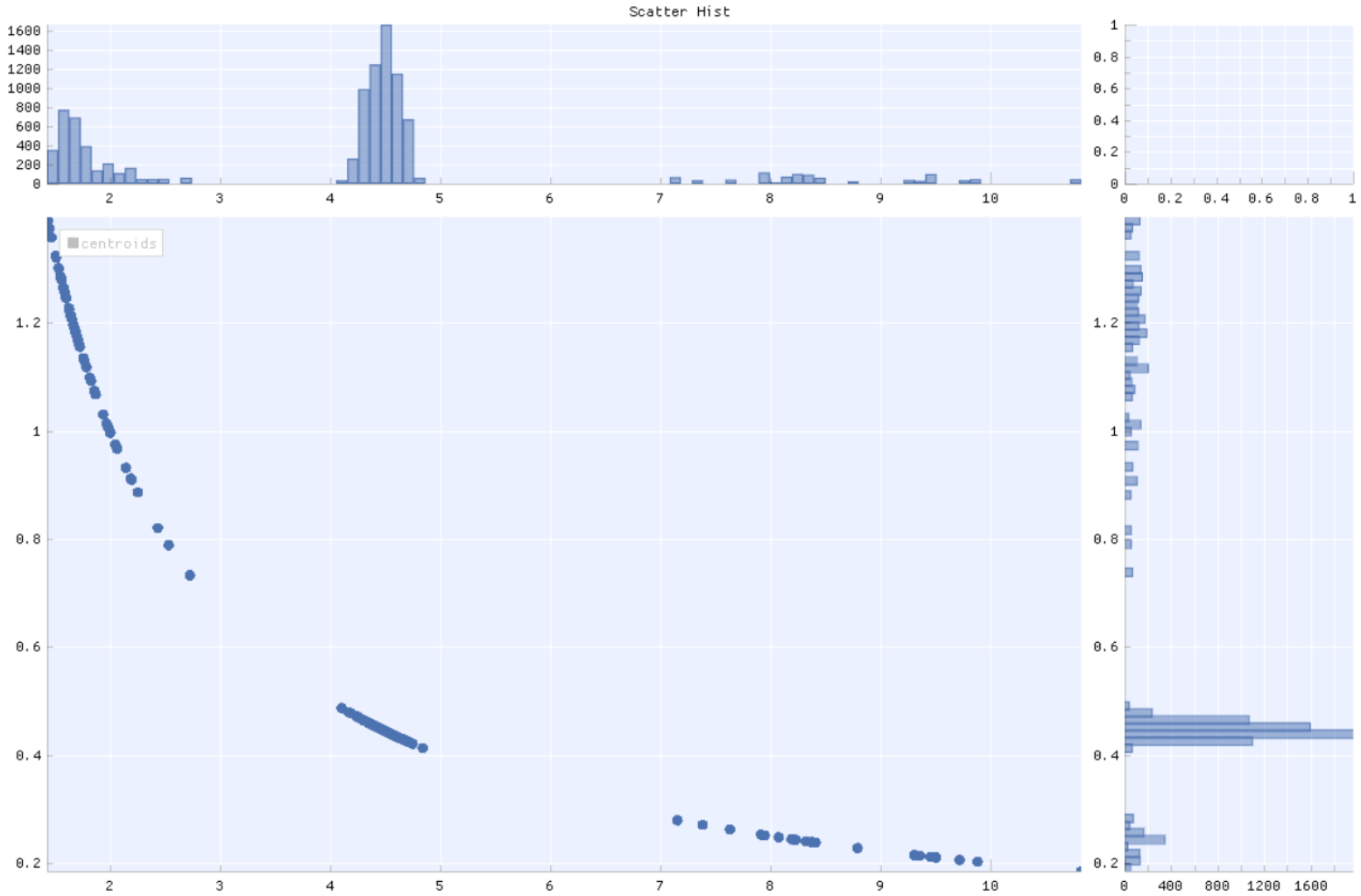
What is needed:

- Discovery of the number of clusters (how many outcomes).
- The centroids of each cluster (characteristics of outcomes).
- Ability to do so with n-dimensional data (more than 3 metrics).

Correlated Histograms does all of these!

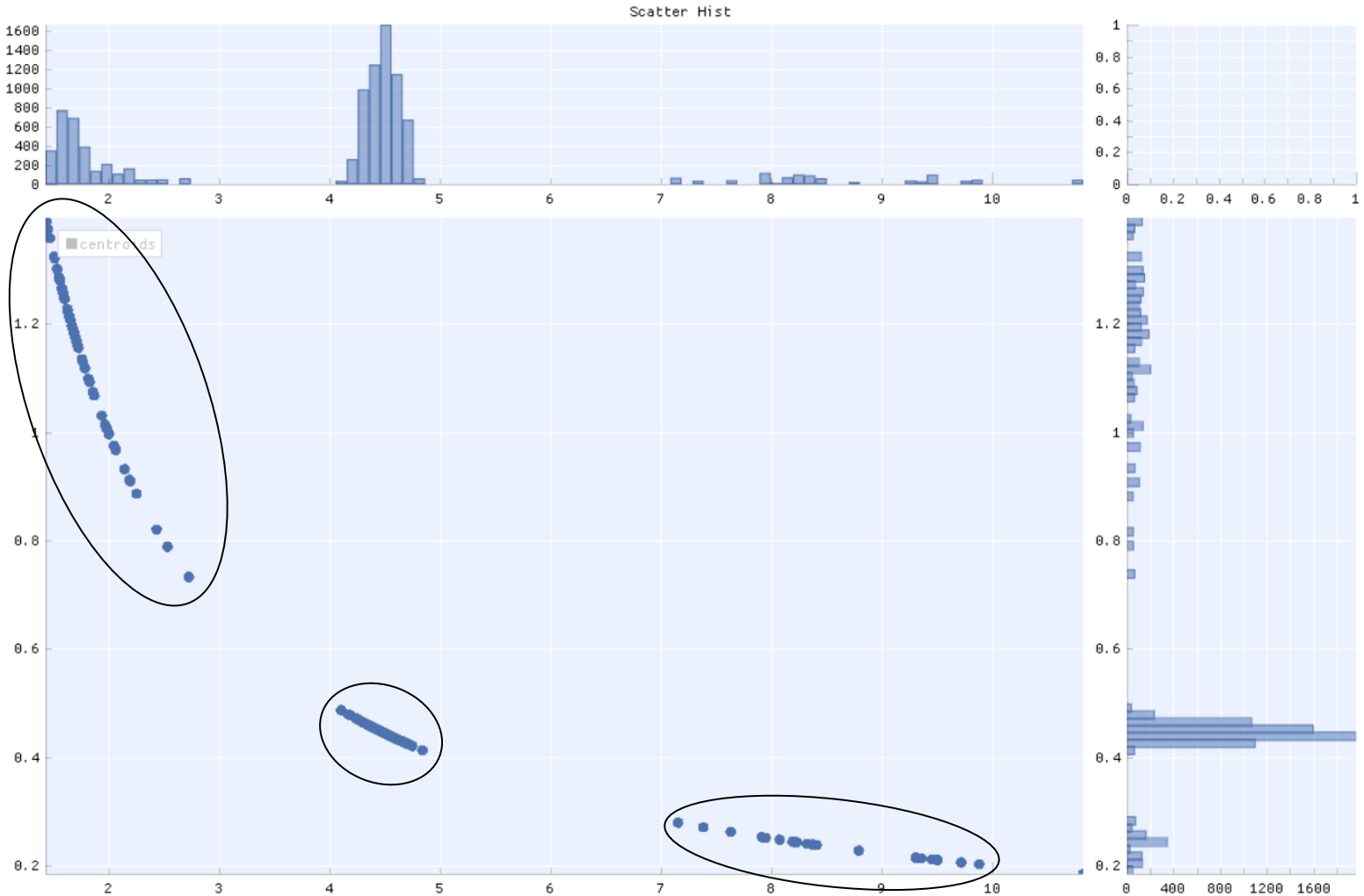
Application – Training Syllabus Outcomes

- Our dataset of training metrics scattered and histogrammed



Application – Training Syllabus Outcomes

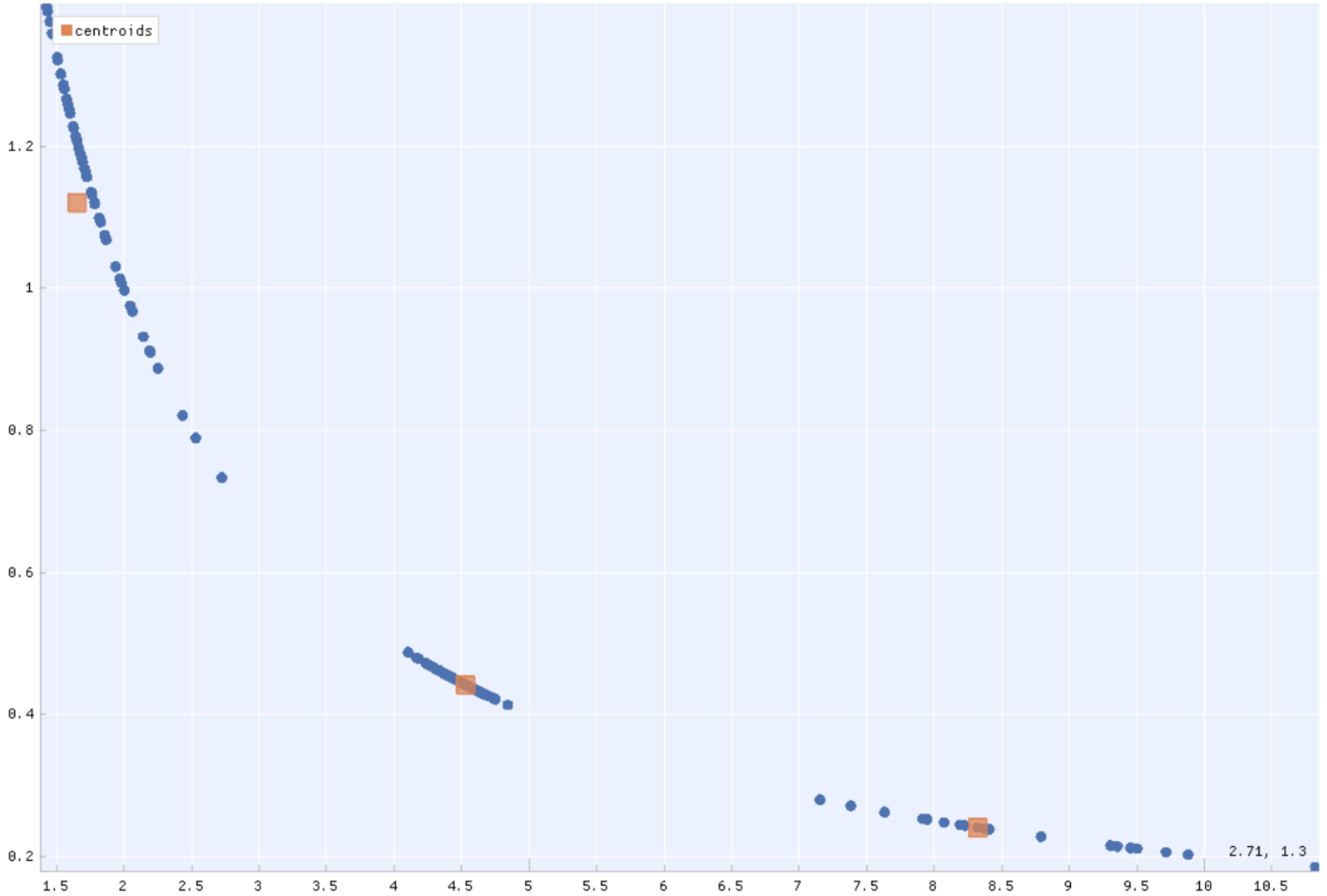
- Our dataset of training metrics scattered and histogrammed



Application – Training Syllabus Outcomes

- Identification of the cluster centroids gives us:
  - The number of outcomes.
  - A datapoint associated with each outcome:
    - (1.654, 1.120)
    - (4.533, 0.441)
    - (8.324, 0.240)

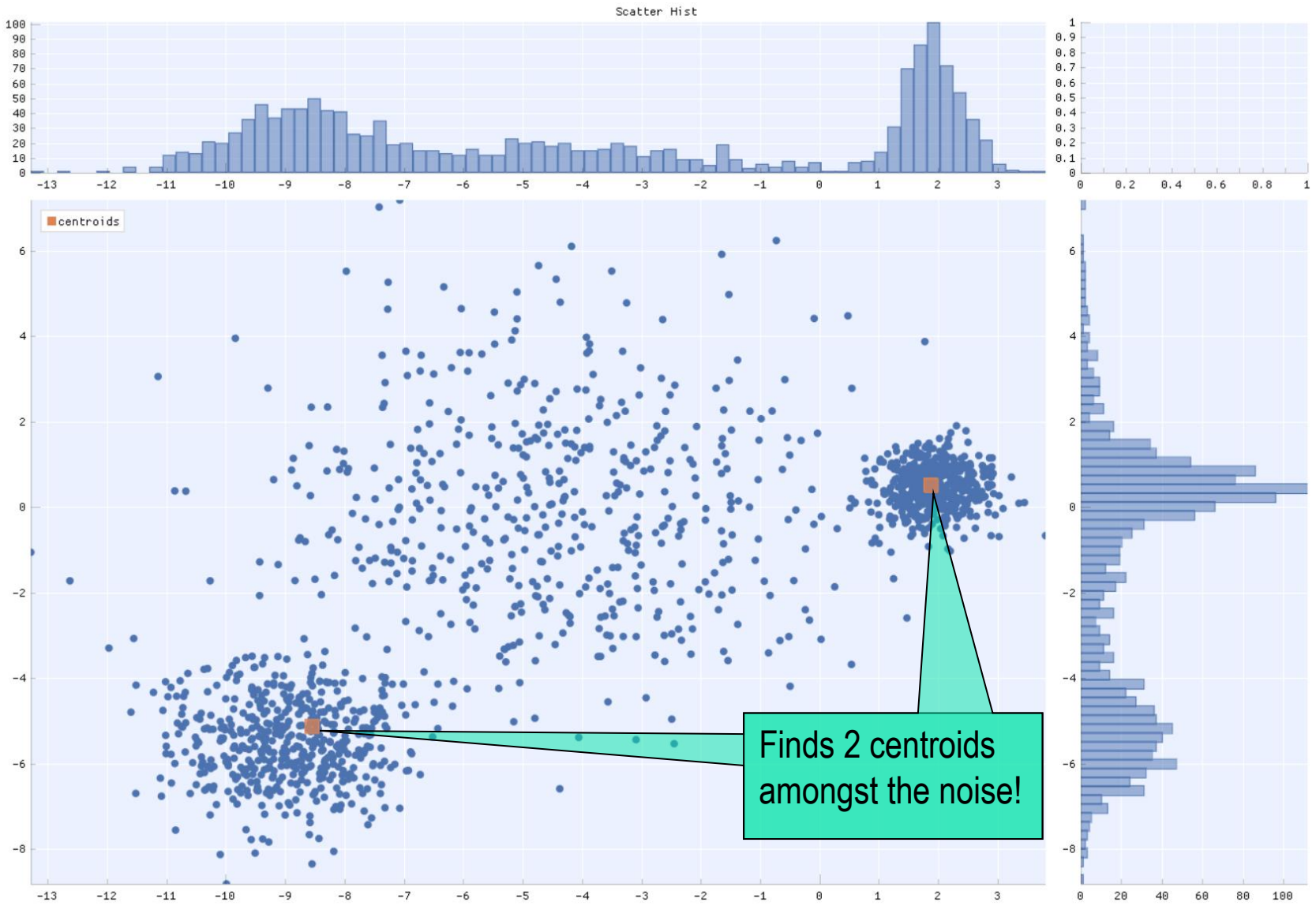
We walk away knowing the number of trainees our syllabus produces *and* ways to describe each type!



Application – Robustness Against Noisy Data

- Another Scenario
- Centroids:
  - (-8.638, -5.119)
  - (1.845, 0.537)

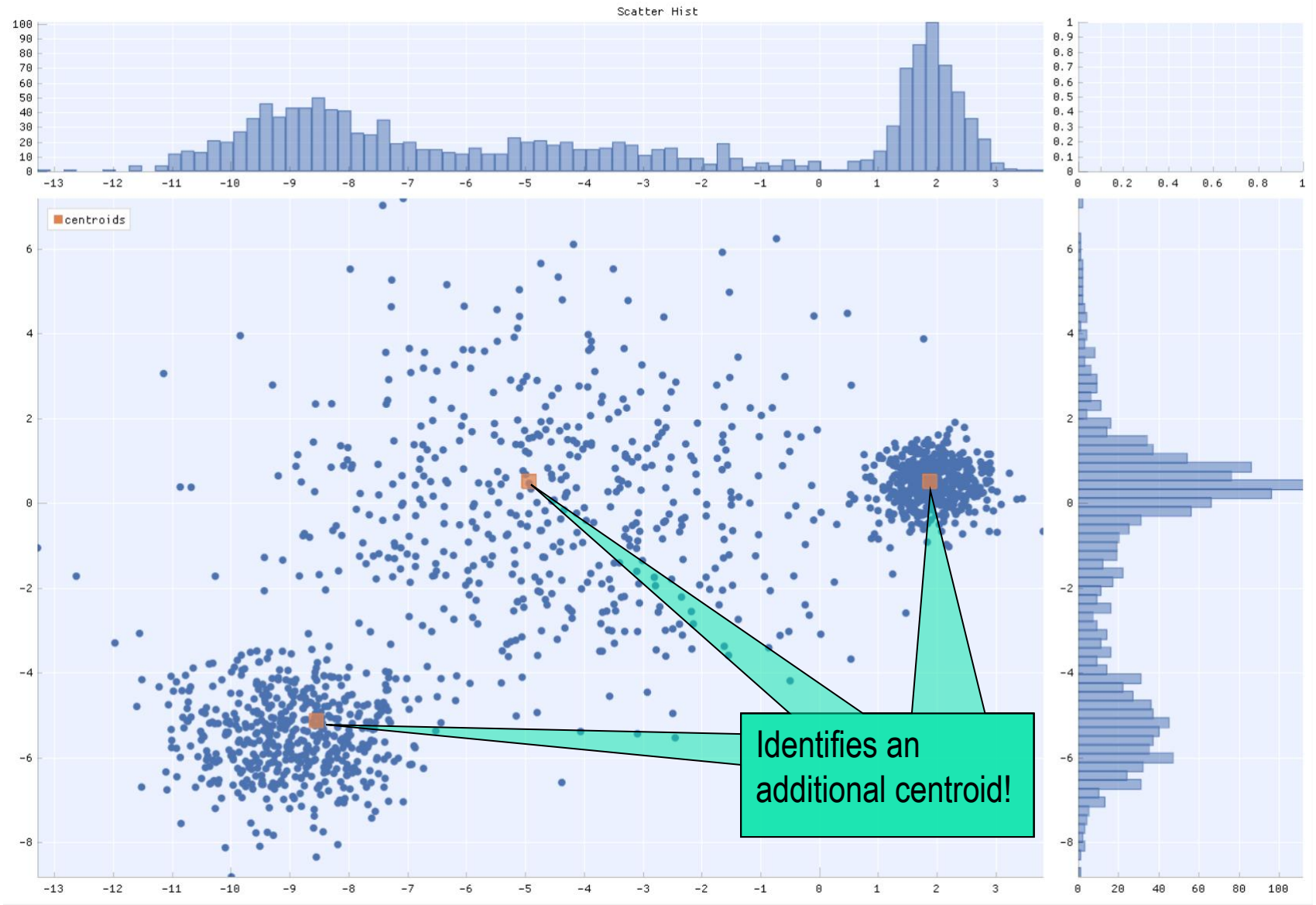
Low Sensitivity!



Application – Robustness Against Noisy Data

- Another Scenario
- Centroids:
  - (-8.476, -5.357)
  - (-4.506, 0.483)
  - (1.847, 0.483)

High Sensitivity!



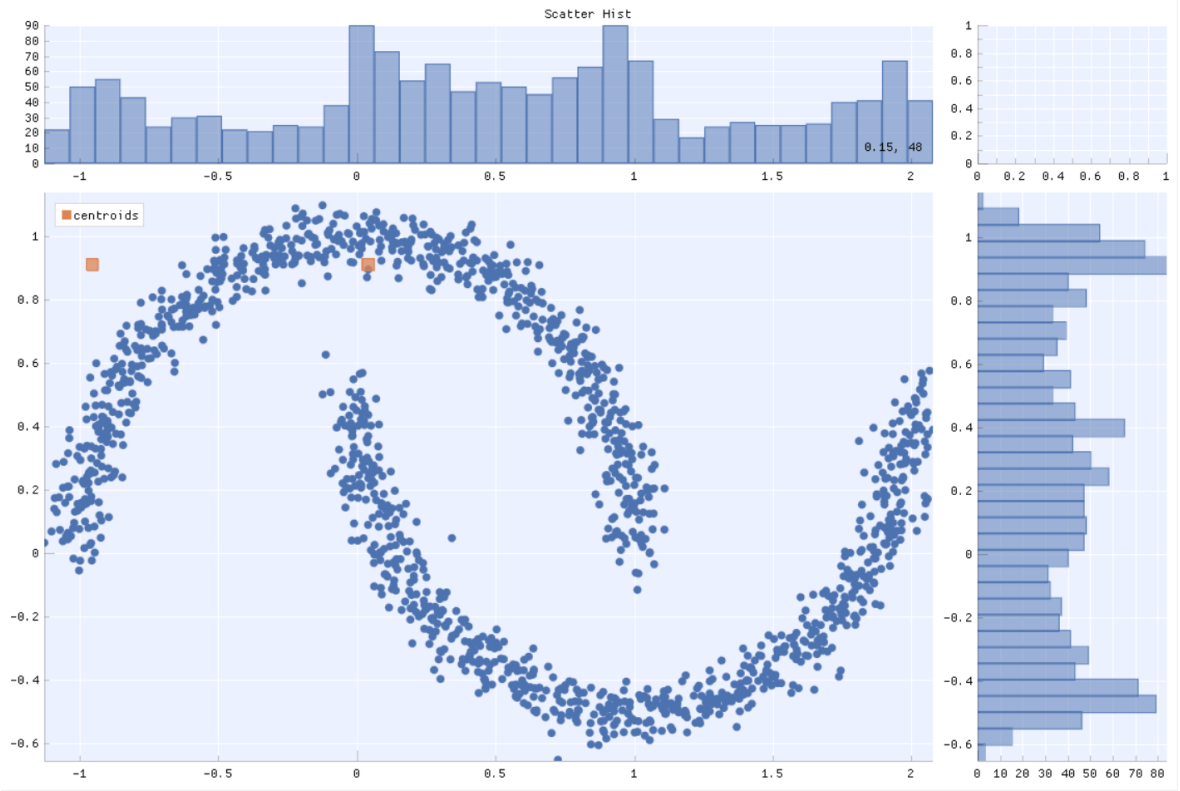
Summary

- Correlated Histograms is an unsupervised learning technique that has applications anywhere that clustering is the task at hand.
- Differences being:
  - You get centroids of clusters rather than classification of instances.
  - These centroids are derived from the underlying statistics of the data rather than distances between points.

Key take-away: Statistics is largely underutilized as a metric in classical clustering techniques!
Correlated Histograms leverages statistics and
can lead to great insights into messy, unfamiliar data.

Future Work

- Handling data that is “oddly shaped” with respect to the orthogonal vectors.
- Swapping out the Harrel-Davis QRDE for other density estimates, classic histograms, or “adaptive histograms” (also from Andrey Akinshin).
- Other Modality detection techniques.
- Checking agreement between some number of nearest points.



Acknowledgements

- Dr. Randy Allen for mentoring me and for co-authoring this paper.