Human Computer Interaction
Part III – Interactive prototype
Erik Beitel

MNST Data analyzed by kMeans for Handwriting Recognition (Training)

1. Data:
   I used the MNIST dataset (with 60,000 data points) to visualize and train the model. I choose this dataset as it is both a good introduction to AI, and it is very clean. This makes it easy to work with, and allow me to focus on giving the user insight into the algorithm. For the visualizations, I did consider reducing the number of points displayed, but felt that it would be detrimental to the performance of the algorithm, and there were better ways to make it more clear. I chose to manipulate the opacity and size of the data points to effectively visualize such a large dataset.

2. Transformation:
   I used two transformations on the data to show meaningful patterns to the user. One was principle component analysis (PCA). PCA is a classic dimensionality reduction technique that allows us to project data in a higher dimensional space (in the case of MNIST, 784 dimensions) into a space with low enough dimensions we can visualize it (I choose to project into 3). It does so in a manner that captures the most information. Along with assigning different colors based on label, the human eye can see clusters in the data that correspond to different digits.
   I also applied kMeans, visualizing the centroids moving through the image and how they change. I used kMeans as it is a fairly effective classifier, as well as fairly intuitive – I felt that I could use kMeans to give a lay person a good introduction to machine learning and how it can work.

3. Color scheme
   My color scheme was primarily grayscale, for everything except the data. This follows first from my data itself – MNIST just shows intensity, and color is irrelevant. With that in mind, grayscale as a primary color scheme also bring focus to any area that has color. In this case, I used it to focus attention on the visualized data, as well as the slider moving across. I ensured the data points themselves were a vibrant color so that they could still be easily seen with a low opacity. I choose RGB(255,255,255) for the centroids, as it shows most clearly through the data

4. InfoViz System
   I choose to use a 3D scatter plot to represent 3 principal components of my data. This is both a very natural representation, and also one that allows humans to see the patterns, and watch the centroids converge. Additionally, because my project's focus is on the centroids and giving the lay person insight into how kMeans converges, I have a separate panel displaying the current state of the centroids, so that the user can visually see the algorithm training.
   I aimed for a fairly minimalistic visual hierarchy that focuses attention on the data itself, and conveys the relevant information at a glance. This led to two layouts (see screenshots) that I would use A/B testing to choose between. Both have the main visualizations centered and easily accessible to the user. That design principle led to the choice of not rendering the image for a given data point when it is clicked or hovered on,

and rather showing the label. In my user testing, I noticed that the user would become frustrated because there was so much data they were never able to click on the same point twice, and it led to frustration and less intuition by the user. When I changed it from rendering the explicit handwritten digit to the label it was supposed to be, user experience was significantly improved.
The animation moves at a slower clip than I would prefer, due to resource constraints on the digital ocean droplet I am using. This slower refresh led to making a greater step size so the user could actually see changes in the rendered centroids.

5. Link: http://167.99.42.95:8050/ for A, http://167.99.42.95:1337/ for B
   Note: The digital ocean droplet is fragile, and I have noticed it hanging if there is not enough activity for a long enough time. If the given URL does not render, please contact me at erik.beitel@epfl.ch or +41 076 412 0373 and I will restart the instance.
   If images are not loading, or the page freezes, please let me know and I will temporarily increase the resources allocated to the instance. If possible, please let me know a timerange to test in advance, as I do not have the funding to keep the instance up with the resources for optimal performance 24/7. Thank you for your consideration.

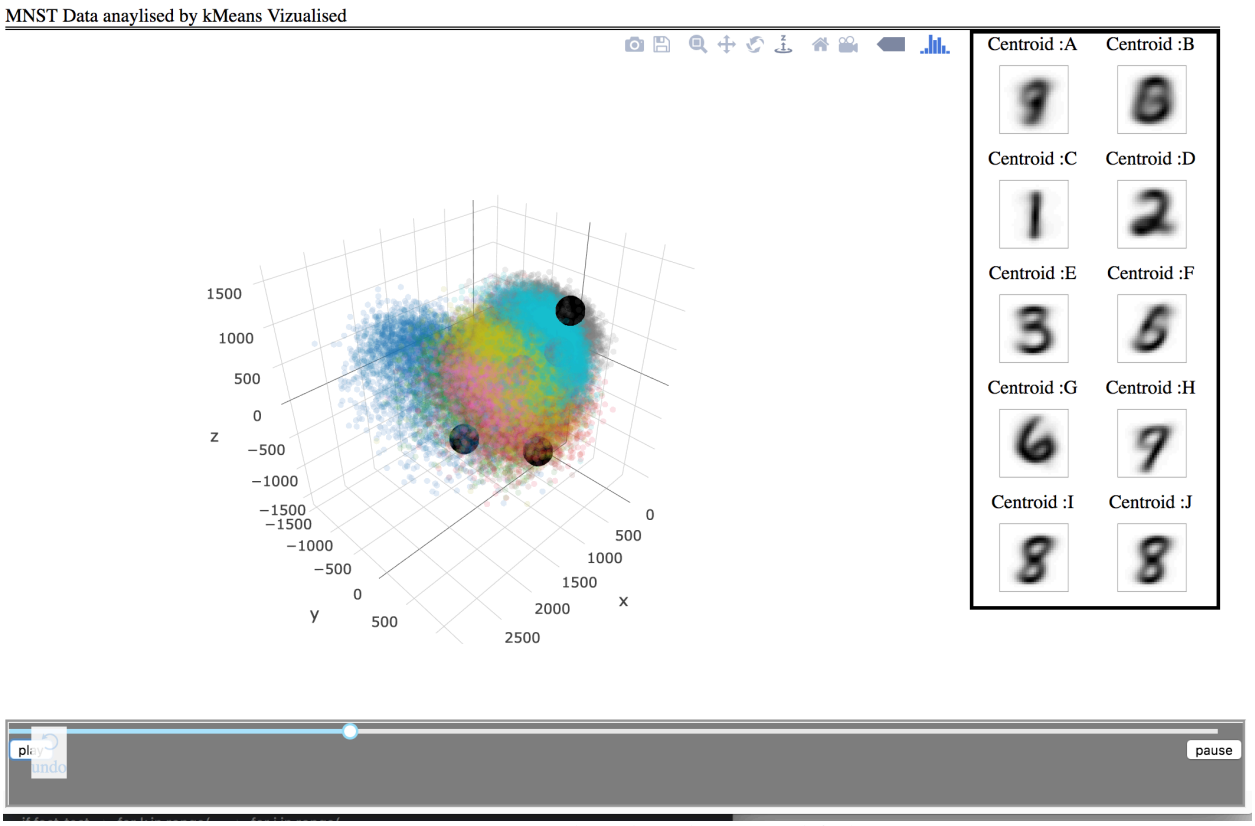

Figure 1: Design A midway through training, default view

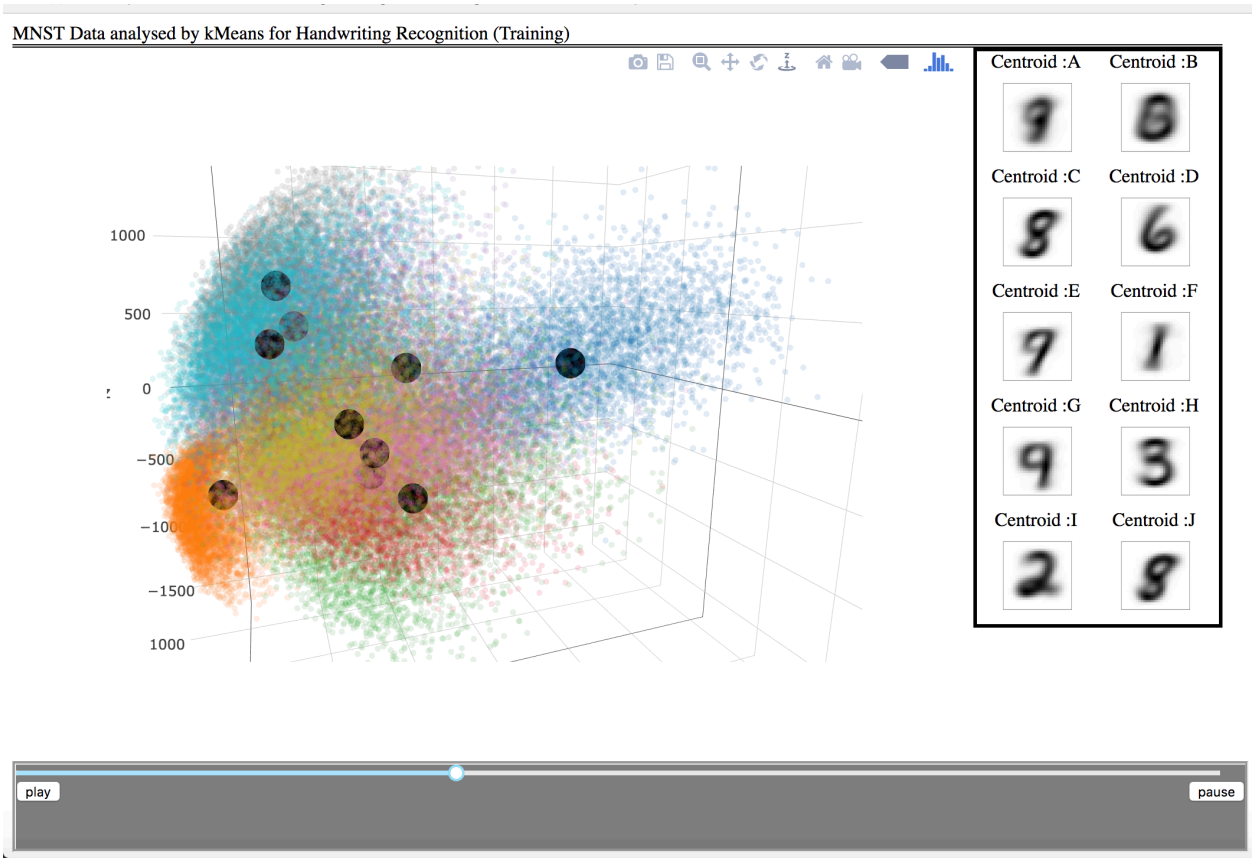

Figure 2:Design A, rotated to show centroids

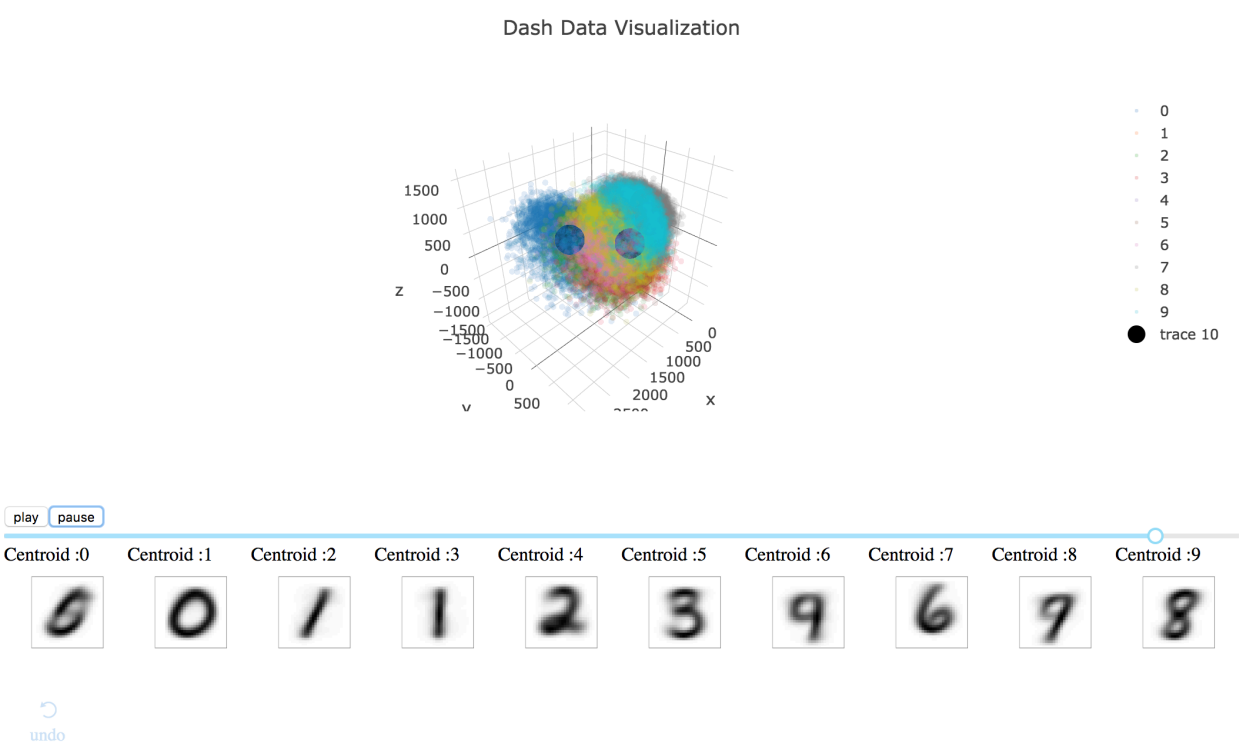

Figure 3: Design B, Default view