SLAC NATIONAL ACCELERATOR LABORATORY

Crystallography Labeling Using Machine Learning

Abdallah AbuHashem[1], Chuck Yoon[2+]

[1]Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

[+]Contact: yoon82@stanford.edu

Introduction

In crystallography experiments, the size of data received is too big. In addition, the number of events/images is in the magnitude of hundreds of thousands. However, scientists usually are interested in a small part of the overall events : single hits and sometimes multiple hits.

Already existing software, Psocake, helped in cutting down the number of events, yet it did not have an interface to classify them, so a GUI had to be developed.

Moreover, the process of labeling was still slow, so to make it faster, a Machine Learning algorithm using the concept of diffusion maps was developed to construct manifolds out of the different events data.

Keywords: Crystallography, Psocake, Labeling, Machine Learning, Diffusion Maps, GUI.

Research

Diffusion Map Manifold

Overview:

Diffusion maps is a feature extraction algorithm. In this research, we use crystallography events data to construct manifolds (E.g. Fig. 1). The distances between points in the manifold represents the diffusion distances or similarity between different points/events.

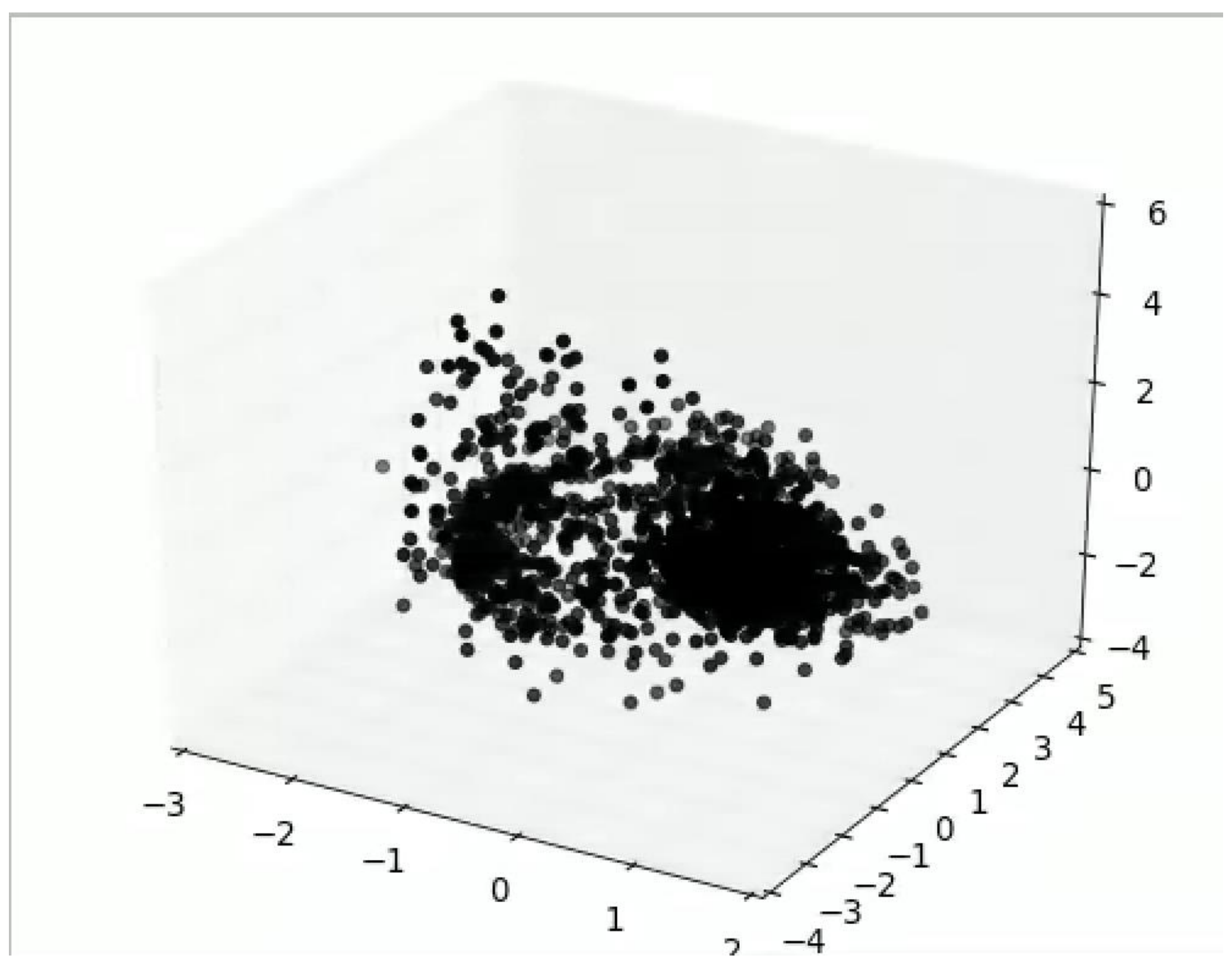

Fig. 1 – Diffusion Map Manifold
Exp. Amo06516 Runs 93-97

Result:

From the manifold in Fig. 1, it is noticeable that there are two main clusters and a bunch of sparse points in the Euclidean space around the two clusters. Every cluster is dominated by a specific type of events. The one to the right has a majority of Single hit events, while the smaller cluster to the left has a majority of water or unknown hits, which are not important nor interesting for scientists. For the sparse points, they are not dominated by any type of hits, but most of multiple hits exist in these areas.

Labeling GUI and Use

A GUI was developed to get use of the diffusion map algorithm and the constructed manifolds. The GUI consisted of two parts:

- First Part (Fig. 2) included the manifold, and few controlling buttons.
- Second Part (Fig. 3) showed the picked events and their labeled. It also allowed labeling them.

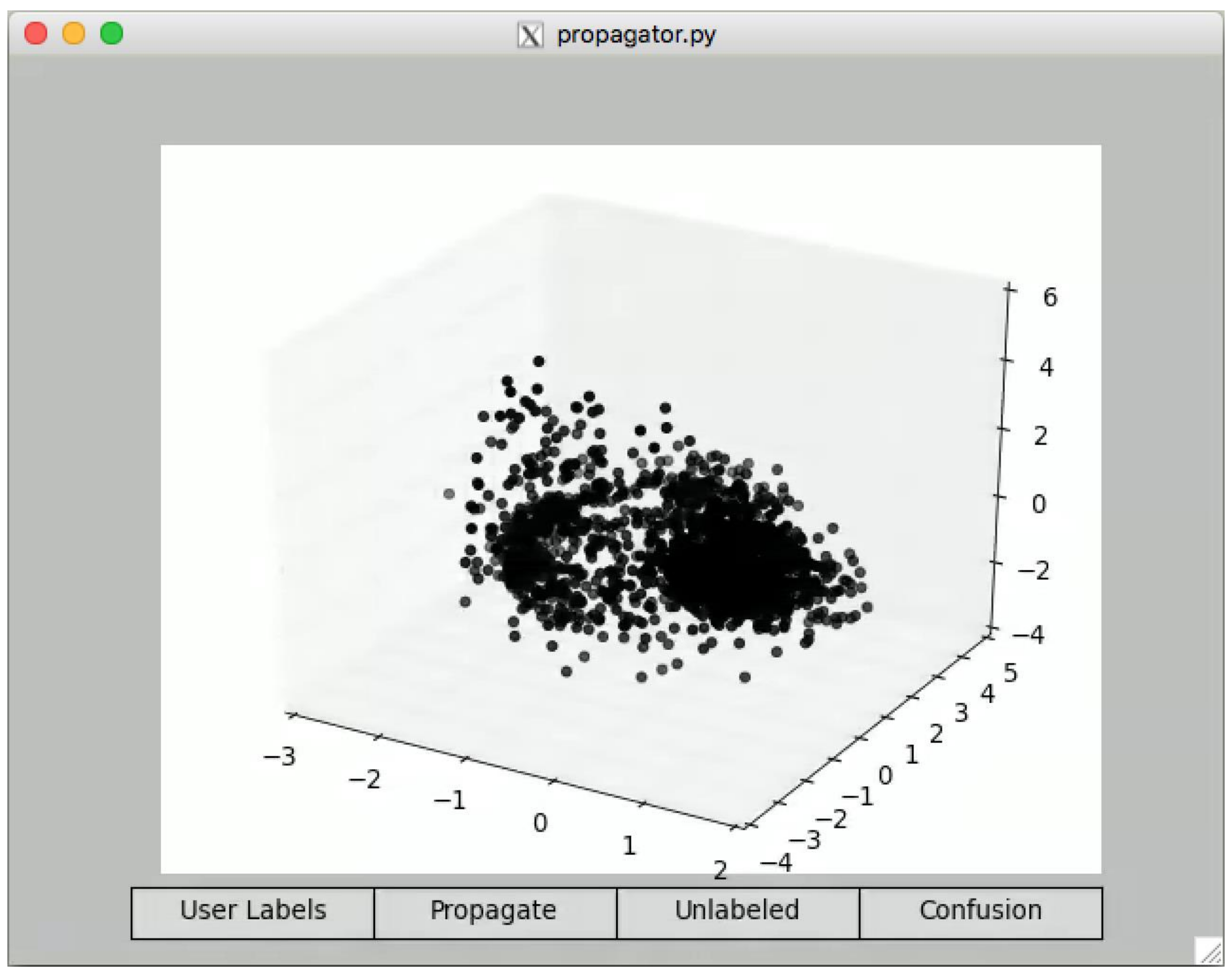

Fig. 2 – Part One of The GUI

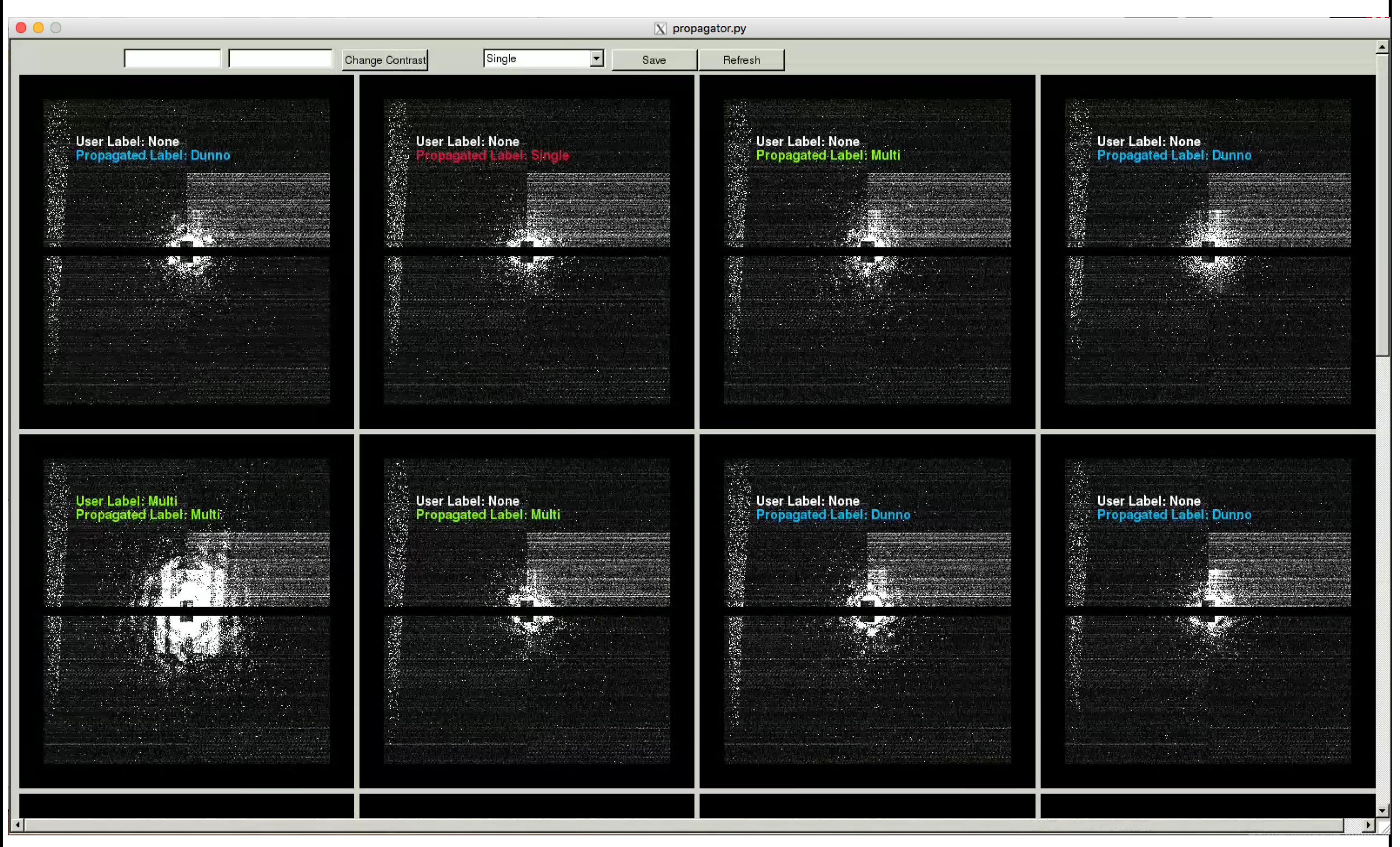

Fig. 3 – Part Two of The GUI

How it works:

A label propagating algorithm was developed based on the manifold distances. As soon as the user starts labeling, the algorithm starts spreading the labels. The spread labels are then shown on the manifold as different colors*. If the user found out that some of the spread labels are not spread correctly, he can correct them, and the spreading algorithm will fix the rest of the labels to achieve a better accuracy.

Example:

The developed algorithms were applied on different runs in experiment amo06516 (A crystallography experiment on virus PR772). The included Manifold shows runs 93-97. In Fig. 4, the colored* points represent user labeled events. Fig. 5 shows how it was propagated. The percentage of user provided labels is 20%. The number of points in the manifold is 1770, and the accuracy that was achieved is 87%.

*Red = Single, Green = Multiple, Blue = Water/Unknown

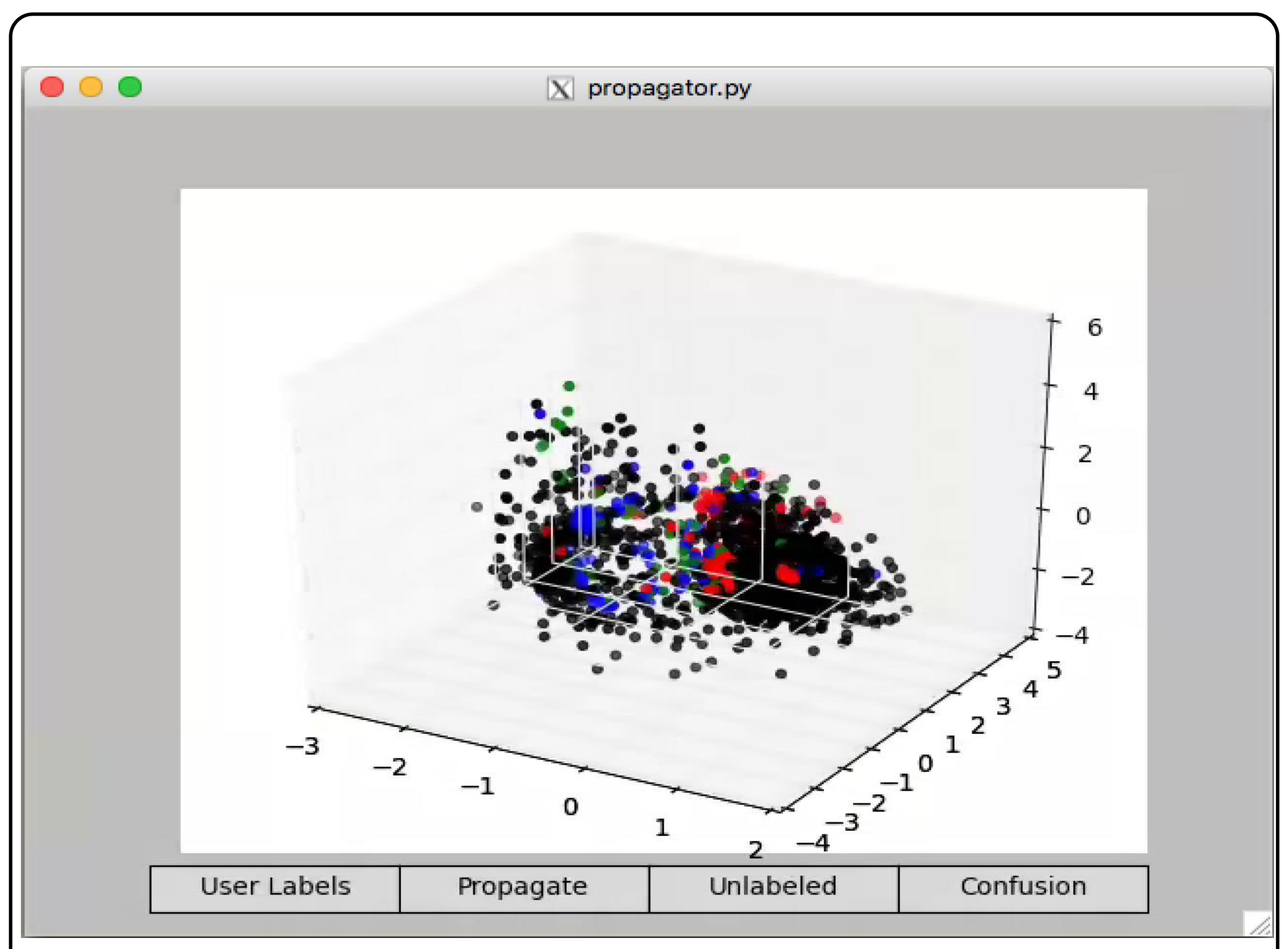

Fig. 4 – User Provided Labels

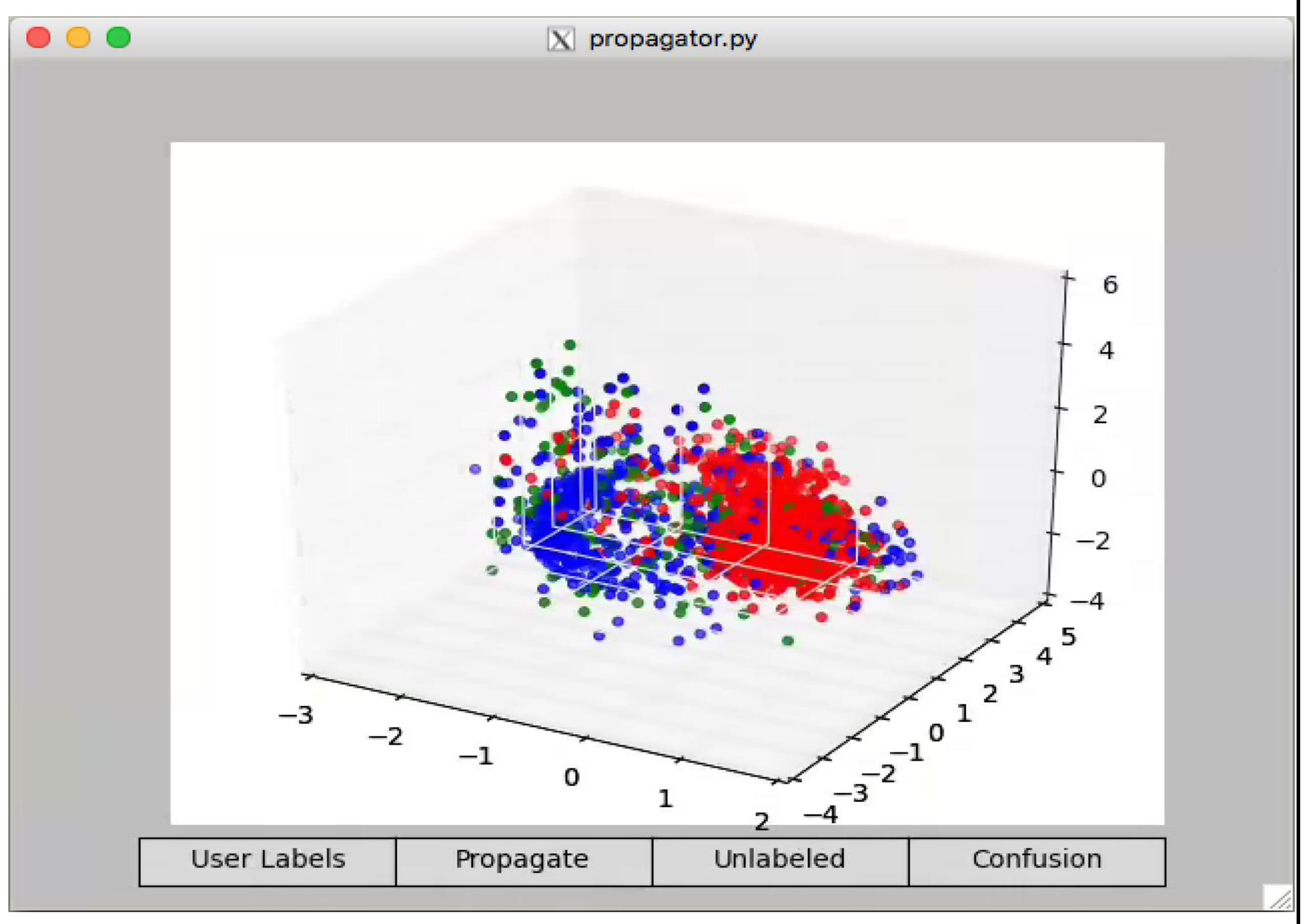

Fig. 5 – Propagated Labels

Recommender System:

In order to make the accuracy of the label propagation algorithm higher, a recommender system was implemented. The system would show 10 images at a time for the user to label. The shown images are the most confusing images for the propagating algorithm.

Conclusions

The results we got are really interesting, and can help scientists a lot. However, it was possible to do things in a better way.

•Getting use of the manifold's labels and data in other runs or other experiments' manifolds. Although this is currently possible between small number of runs, it does not work on large numbers.

•The label propagating works at an acceptable accuracy of about 85%. However, a higher accuracy is always better. In addition, it would be better if it required less than 20% user labeled events.

Acknowledgments

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

A special thanks to my mentor, Chuck Yoon, for providing countless advice and support on the coding and algorithm development process.

Date: 09/07/2016