EdLab Research Task 2
Analyzing Real-Time Location Tracking Data

Henry Williams

February 17, 2019

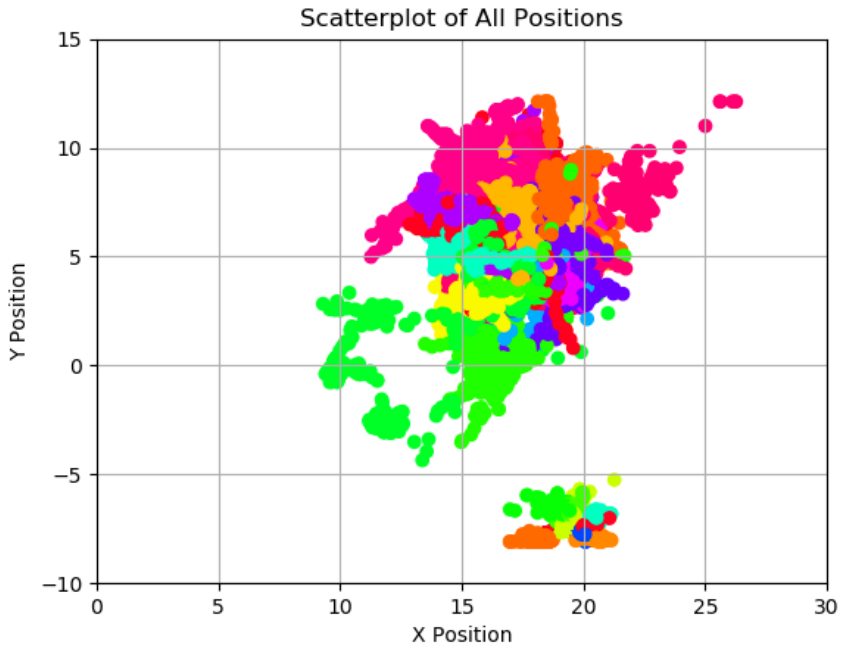

Figure 1: Scatterplot of all recorded positions, colored by tag ID

Research Task Background

The Smith Learning Theater (LT) at the Gottesman Libraries, Teachers College is equipped with a real-time location tracking system - Quuppa. The Quuppa system tracks the position of each participant (wearing a tracking tag) in the 3D space according to a defined frequency. So each tracking record has the ID number of the tracking tag, 3D coordinates - (x,y,z), and a timestamp. Attached is a sample of tracking log file in json format. For the purpose of this assignment, please use variable "id" for tag ID, "smoothedPosition" for location, and "positionTS" for timestamp. Your task is to process and analyze the log file. There is no fixed research questions for this exercise, and you can analyze the data using any methods you see fit. Share your results in a write up along with your code.

Introduction

Finding compelling ways to analyze and process time-series position data is a complex task. The multi-dimensional nature of the data, as well as the complexity inherent in dynamic processes which evolve over time, presents distinct challenges. There appear to be two main approaches one could take here: descriptive, and prescriptive. Prescriptive intending to simulate future activity in the Smith Learning Theater (LT hereafter), and descriptive intending to describe the conditions in the LT in the time covered by this dataset. I chose to focus more heavily on a descriptive approach, using the available data to identify patterns in these individual's movements and to develop methods which could be used on broader datasets, particularly in visualizing the the distribution of people (Figure 1 for example) across the space during the observed time and finding groupings of positions for specific individuals and between multiple people over time that seem to be meaningfully related.

I was exploratory in my analysis, trying to look at the provided data in a variety of ways to see what conclusions might naturally proceed from observation. My central finding was that the movement of an individual through the LT over time will often be localized to a set of primary "clusters" (meaning their movement for some stretch of time will center around one of these clusters) which can be identified with a variety of computational tools. In addition, I constructed several heatmaps which pinpointed areas of greatest activity within the LT. I also created dynamic animations of my visualizations which can be cycled through to observe patterns in where individuals spend the majority of their time in the LT. Finally, I identified possible applications of my analysis in creating a path charting the major locations visited within the room for a given individual, or a program which would infer probable interactions between individuals by identifying cluster centers that are close to eachother between the two at the same time.

Methods

In investigating the dataset, I quickly discovered that the z coordinate is unused for all of the position data it contains, meaning the primary analytical task is finding interesting and compelling patterns in 2D positions, associated timestamps, and the tag IDs these positions are associated with.

This program was written entirely in Python, with the goal being rapid iteration and the ability to utilize the swiss army knife of tools available within Python's ecosystem. It primarily made use of the following libraries:

1. Scikit-Learn [1]
   (a) K-Means Clustering
   (b) Agglomerative Hierarchical Clustering
   (c) Mean-Shift Clustering [2]
2. Matplotlib [3]
3. NumPy [4]

All figures were generated on a 2014 Macbook Pro, and the code was written with a function-oriented paradigm in mind. The code can be found alongside this document in a file called "edlab.python."

Data Visualization

I utilized Matplotlib to visualize patterns in the movement of individuals in several different primary subdivisions: all the locations associated with a given tag, all the locations of all tags, all the locations of all tags over a given time interval, all the locations of a given tag over a given time interval, and all the locations of 2 or more tags over a given time interval. I started out by exploring what different visualizations of locations looked like, some using time for the third dimension of the plot, before settling on heatmaps and scatterplots and the easiest way to identify clustering and see density of individuals within the room.

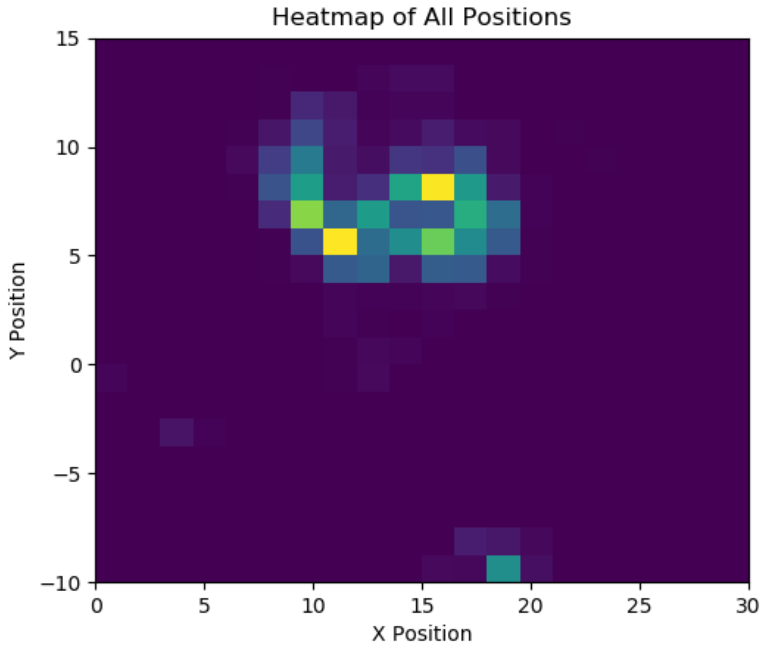

Figure 2: Heatmap of all recorded positions

Creating Data Animations

The Matplotlib library has an associated animations library which can be used to create self-updating plots that animate to show change over time and between different tags. Two such animations are attached in the files "HeatmapAnimation.mp4" and "ScatterplotAnimation.mp4." These animations helped me narrow down what approaches I would take and could be expanded to other applications in analyzing this data, primarily tracking the movement of a particular individual over time with a line that would follow the different clusters they move between.

Clustering Techniques

Soon after creating solid visualizations of position distributions within the LT, I consulted resources on statistical analysis which led me to see clustering as the best approach to finding patterns in this particular dataset. The scikit-learn library's documentation describes clustering as "Automatic grouping of similar objects into sets," in this case the sets are sets of points which are closely located in a specific time interval, and which segment the movement of a person over time into discrete

clusters of positions which simplify analyzing them. I explored many different techniques for doing this clustering, but settled on three in particular in writing up this report.

K-Means Clustering

The algorithm for K-means clustering clusters data by separating samples into n groups of equal variance, minimizing a criterion known as the within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified before running. It scales well to large number of samples which initial drew me to using it. It is also simple and low-cost computationally. I did my analysis assuming a boilerplate 4 clusters for each individual, which I decided on through visual observation, but this gave overall questionable results. Later on, when I implemented the mean-shift clustering algorithm (which fits a number of clusters without pre-specifying) I used that number of clusters to repeat my K-means analysis to markedly better results.

Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering differs from k-means in that rather than choosing a number of clusters and starting out with random centroids, we instead begin with every point in our dataset as a potential "cluster." The algorithm then finds the two closest points and moves to combine them into a cluster. We continue to find the next closest points, and clustering them, and repeating the process until we only have one cluster, later selecting out some of these subclusters based on observation. An illustration of the structure of these results can be seen in Figure 3, in a dendrogram showing how each cluster inherits from smaller subclusters. This algorithm is computationally intensive, however, and did not yield particularly excellent results, leading me to continue searching for a better approach.

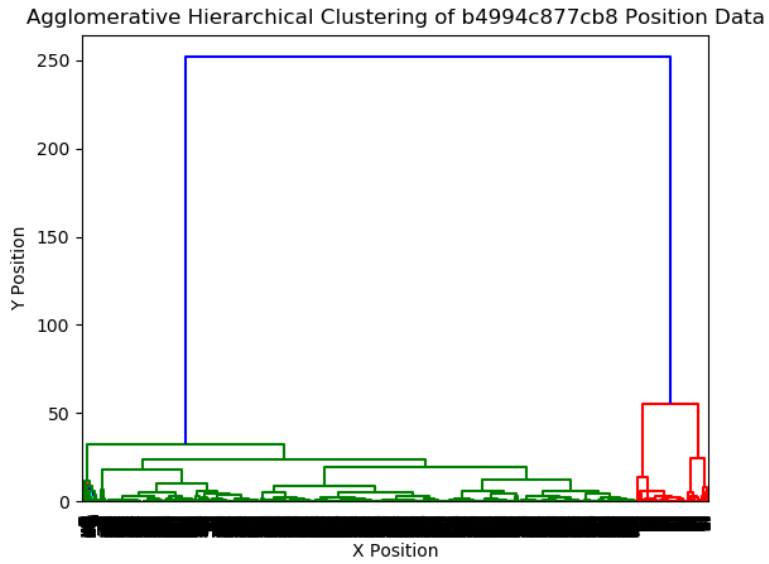

Figure 3: Example of a agglomerative hierarchical dendrogram

Mean-Shift Clustering

By far the best approach to clustering this dataset I employed was the mean-shift clustering algorithm. The approach of this clustering algorithm is to identify blobs in a medium-to-large number of samples without specifying an expected number of clusters beforehand. It is based on centroids (the intersection of lines connecting groups of points), updating possible candidates to be the mean of the points in a region. These are then continually weeded out in a later processing step to minimize duplicates and to form the final set of clusters and cluster-centers they are built around. Another benefit of this algorithm is that it provides discrete center points for clusters that can be used for other analysis.

Results

Observing the heatmap of all recorded positions (Figure 2) and especially the scatterplots and heatmaps for all the individual tags (Figures 4 and 5 respectively), we can see that the area just above the center of the theater exhibits the highest density of position datapoints. This activity is particularly clustered around two centers seen at around the coordinates $(16, 9)$ and $(12, 5)$. This leads me to believe that these are primary loci of activity within the theater, perhaps being near the **stage** or **main passageway** to the available seating.

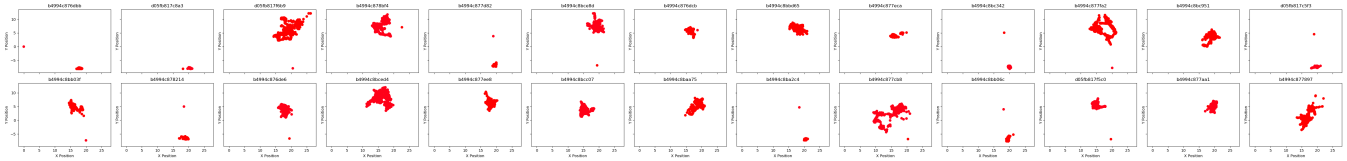

Figure 4: Scatterplots of positions for all recorded tags

Observation of the scatterplots of positions for each tag yields the interesting fact that the position data is nearly always gathered into one to five (but most often one) distinct areas where all the points are closely packed together. In addition these areas are generally located in similar portions of the LT as previously discussed. The heatmaps of positions tell a similar story, with the added information that most individuals spent by far the majority of their time in the LT localized around one-to-two small areas, seen in yellow. This could be that individuals **seat**, or simply the portion of the theater they stood in for a long period of time.

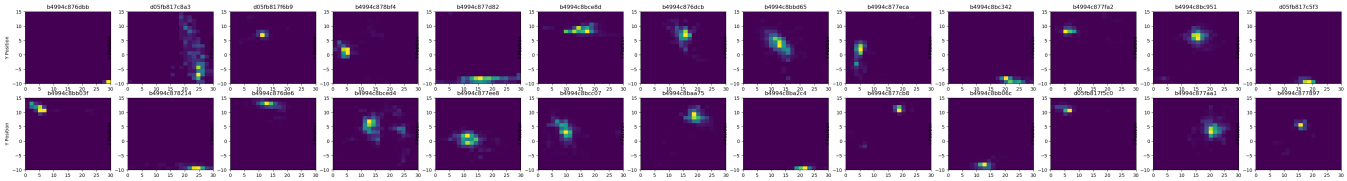

Figure 5: Heatmaps of positions for all recorded tags

These heatmaps and scatterplots alone, however, do not provide much clue into the time evolution of positions or how exactly they are grouped within the room. For these purposes, rather that looking at all the data at once, it is easier to focus on two illustrative examples of my analysis pipeline that show what information can be gleaned thruogh the aforementioned clustering techniques:

Example: Tag ID b4994c877cb8

In doing my analysis, I started with the above information in order to identify the best ways forward, coming to the conclusion that clustering techniques would be best suited to finding where an individual spent substantial portions of time in the LT.

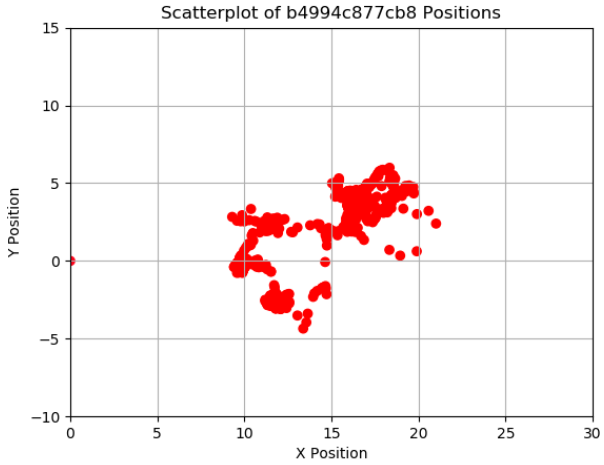

Figure 6: Scatterplot of positions for Tag ID b4994c877cb8

The scatterplot of positions for Tag ID b4994c877cb8 (Figure 6) tells an interesting and somewhat divergent story. Unlike the far more heavily clustered plots for the other tags, this one is more spread out, and in a less-traveled portion of the LT. My hypothesis was that clustering algorithms would show that rather that just a stroll from their seat, this individual probably spent long portions of time at multiple places within the room, maybe changing seats or just not sitting still.

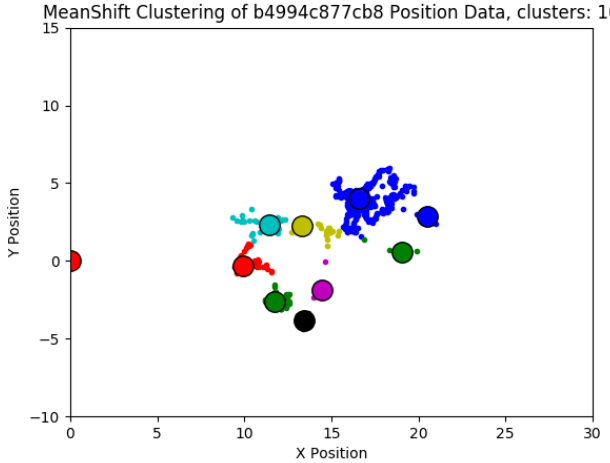

Figure 7: Mean-shift clusters of positions for Tag ID b4994c877cb8

The mean-shift clusters of positions for Tag ID b4994c877cb8 (Figure 7) appear to confirm this hypothesis, showing ten distinct points around which these locations are clustered, each one representing a meaningful portion of time spent surrounding that location. Of course it is not certain that these particular points in the room are all signifigant, but exploring the locations identified

in this data from this data within the LT itself might correspond to some points of interest in the real world. The coordinates of the centers of the algorithmically-identified clusters are shown in the table below:

Cluster:	1	2	3	4	5	6	7	8	9	10
X-coordinate	16.59	11.75	9.91	11.42	14.44	13.30	13.40	20.5	19.06	12.45
Y-coordinate	4.017	-2.64	-0.28	2.35	-1.88	2.25	-3.8	2.87	0.55	-0.4

Example: Tag ID b4994c8bced4

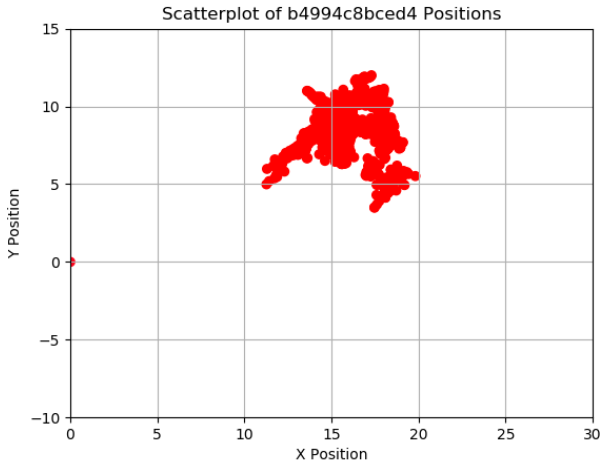

Figure 8: Scatterplot of positions for Tag ID b4994c8bced4

The prior example focused on the jump between initial observations and final results, but as described in the methods section, there were many steps between these. In addition, these steps led me to refine my approach and research alternate techniques (including mean-shift clustering) to improve these results.

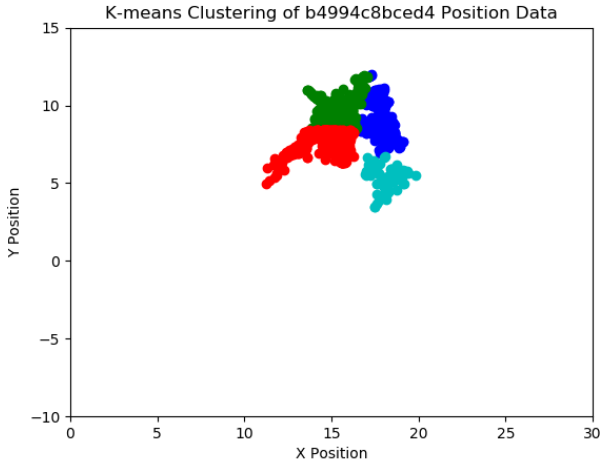

Figure 9: K-Means clusters ($nclusters = 4$) of positions for Tag ID b4994c8bced4

The scatterplot of positions for Tag ID b4994c8bced4 (Figure 8) shows that this is a much more typical case when compared to the other individuals, but focusing on these cases was how I developed my analysis. My interest in these cases was how exactly do individuals localize their movement around particular areas within this biggest cluster, and could their overall movements be seperated into "phases" or "areas" similar to the more outlying examples.

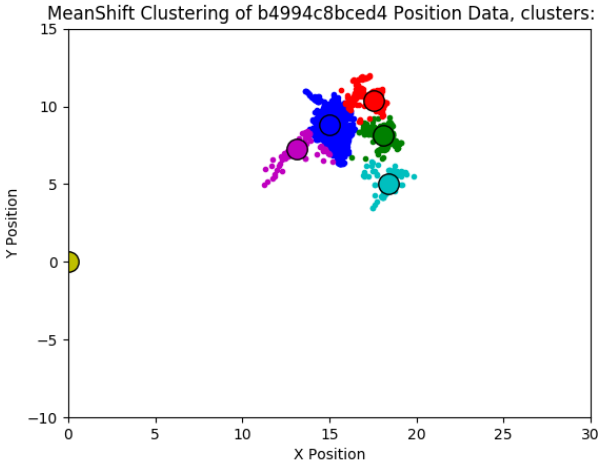

Figure 10: Mean-shift clusters of positions for Tag ID b4994c8bced4

My first approach, as described in the methods section, was K-means clustering (Figure 9), an approach which was primarily flawed because it required pre-selecting an anticipated number of clusters, which I eyeballed around four between the various tags, but which was of course a matter of guesswork. The general wonky results of this approach led me to mean-shift clustering (Figure 10) which provided remarkbly superior results. Interestingly, discounting vastly outlying clusters (which I hypothesize are centered around the entrance or exit) the average number of clusters fit by the method was around six, though there are of course outlying examples like Tag ID b4994c8bced4 from before.

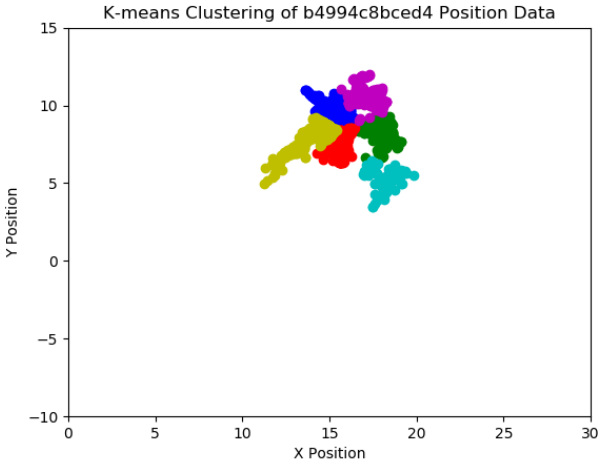

Figure 11: Adjusted K-Means clusters ($nclusters = 6$) of positions for Tag ID b4994c8bced4

Armed with the mean-shift algorithm to fit an ideal number of clusters to each individuals movement, I re-ran my k-means approach with the number of clusters fit by the mean-shift approach. This yielded results which were remarkably similar to the mean-shift method given the far less sophisticated approach of k-means, leading me to believe that the application of a hybrid method of fitting clusters and then running k-means clustering might be a dynamic solution if one were interested in fitting very large datasets or doing many such clusterings over small time intervals.

Conclusion

My work focused on observing patterns within the data through visualization and clustering and I came to several meaningful conclusions about this dataset. The challenges inherent in learning meaningfully from a complex dataset of this kind with minimal context and on a short time interval are many, so I thought it best to create tools which could be expanded on and improved in future work. Primarily, I found that movement of an individual through the LT does appear to be meaningfully localized around central points of clusters which may correspond to real-world hotspots. I also found that there is substantial overlap in which individuals were most often found within the LT. Finally I created tools for visualizing and animating these results.

In addition to these conclusions, I have identified two possible future applications which could expand on the tools I have built to pursue a particular functionality:

Potential Application: Tracking an Individual's Path

Given the success of applying different clustering techniques to the positional data associated with a given tag (an individual in the theater), there is an interesting possible application which could be extended from these results. By ordering the centers of each cluster in time, and drawing lines between them, one could find an individuals primary path through the room. Comparing these paths between a large sample of individuals could identify the highest-traffic corridors and when inefficient routes are taken to a certain location, as well as when individuals do unnecessary backtracking or are caught up in lines on egress or ingress. Additionally we could combine these paths with other tracking information about the location of particular attractions within the theater or other points of interest to see how many individuals engaged with them and in what order.

Potential Application: Inferring Interactions Between Individuals

Since my analysis showed groupings of locations in clusters which appear to show where someone spent a substantial portion of their time within the LT, studying the overlap of these clusters between individuals in a certain time interval could allow us to infer that they might have interacted or crossed paths. This could be assigned a probability based on the amount of time they spent in closely located clusters, and these probabilities could be used to make an overall report on the probability that two people interacted, which could be used to improve the layout and design of the theater to better facillitate these types of interactions if so desired.

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] Dorin Comaniciu and Peter Meer, "Mean Shift: A robust approach toward feature space analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. pp. 603-619.

[3] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," in *Computing in Science Engineering*, vol. 9, no. 3, pp. 90-95, May-June 2007. doi: 10.1109/MCSE.2007.55

[4] Travis E, Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).