Appendix: User Study Instructions

Here we provide the instructions presented to participants for the user study, separately showing the list interface, and the scatterplot interface.

Instructions

Here we provide instructions for the study you are about to take in using visualization to help bootstrapped information extraction. We will go over a description of the task, the dataset that you will be interacting with, a description of the visual interface and how to perform interactions, followed by a short demo that you will have to complete before moving on to the main experiment.

Task Objective

Your objective in this task is to accurately assign labels to entities as fast as possible, and as many as you can.

Task Description

The goal of this study is to assess the effectiveness of different user interfaces that enable a human to provide labels for the purposes of training a machine learning system. The task we are considering is that of **automatic labeling of entities**. An entity is a noun phrase in an expression that is distinguished by its *category*, for instance an entity could be an affiliation, a location, an organization, or a person.

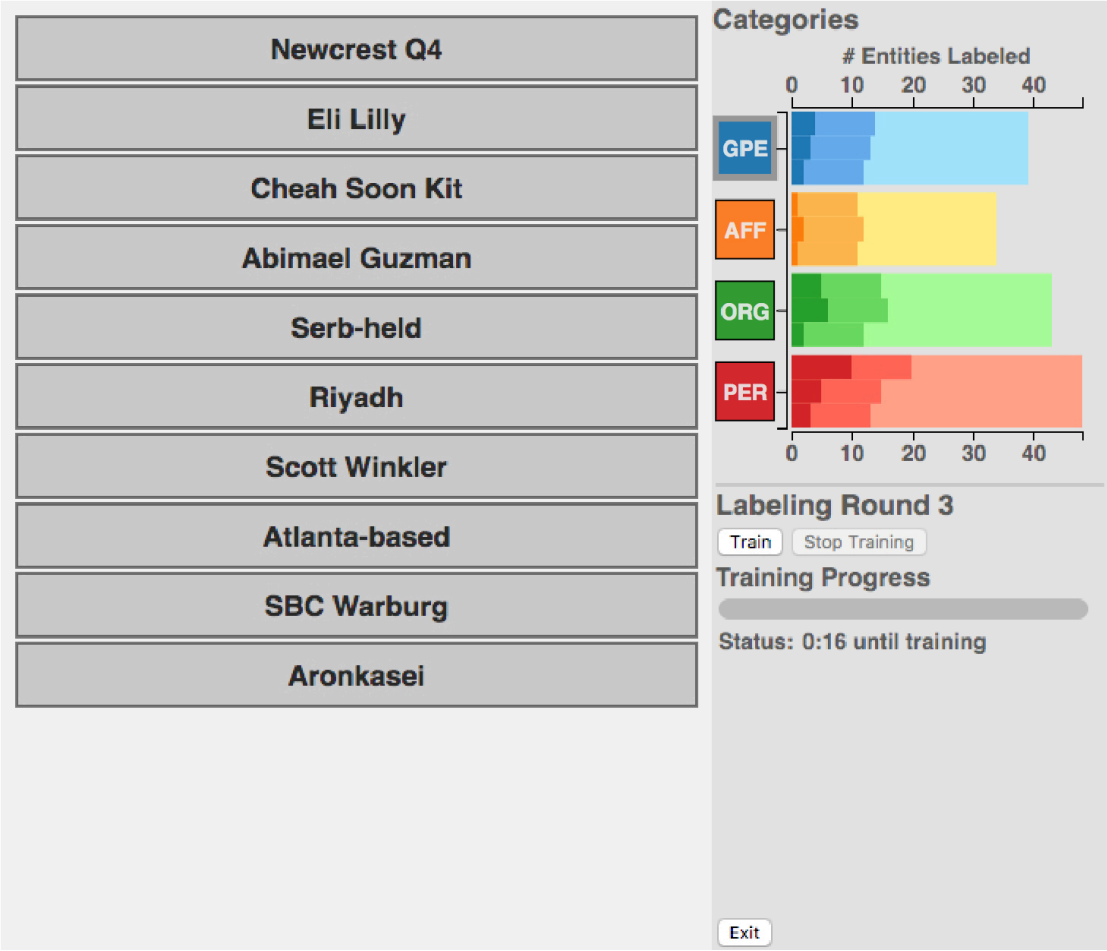
Training a machine learning system amounts to computing a model given a set of labeled entity examples, which will be your responsibility to provide. For instance, if you see the entity "George Washington", then the label you would provide would be **Person**. However, you will not provide a complete set of labels in one sitting. Instead, you will provide labels **in rounds**. Namely, you will label entities for at most 1 minute, and then the model will update, based on the labels you just provided as well as automatically discovering other labels. After the update, the next round begins, and you will be presented with a new view of the entities for labeling. You will label entities for a total of 10 rounds.

Dataset

The OntoNotes v5.0 (2012) dataset is used for this study. This dataset is a collection of text from news, broadcast, talk shows, weblogs, usenet newsgroups, and conversational telephone speech. We use 4 of the 11 categories from the dataset to distinguish entities:

- **Affiliation**: This category is largely comprised of affiliations, including nationalities, religious groups, or political groups. For instance "American" is an affiliation.
- **Geopolitical Entity (GPE)**: This is the name of a location. It could be a city, state, or country. For instance, "San Francisco" and "California" would be locations.
- **Organization**: An organization could be a company, agency, institution, or a department, among others. For instance "Microsoft", "Doubletree", "Democrats", and "Treasury" are all organizations, as well as sports teams.
- **Person**: This is the name of a person. It could be ones full name, first name only, or last name only. For example, "Michael Smith" would be an entity with category person, but so are "Michael" and "Smith". Fictional names are also included.

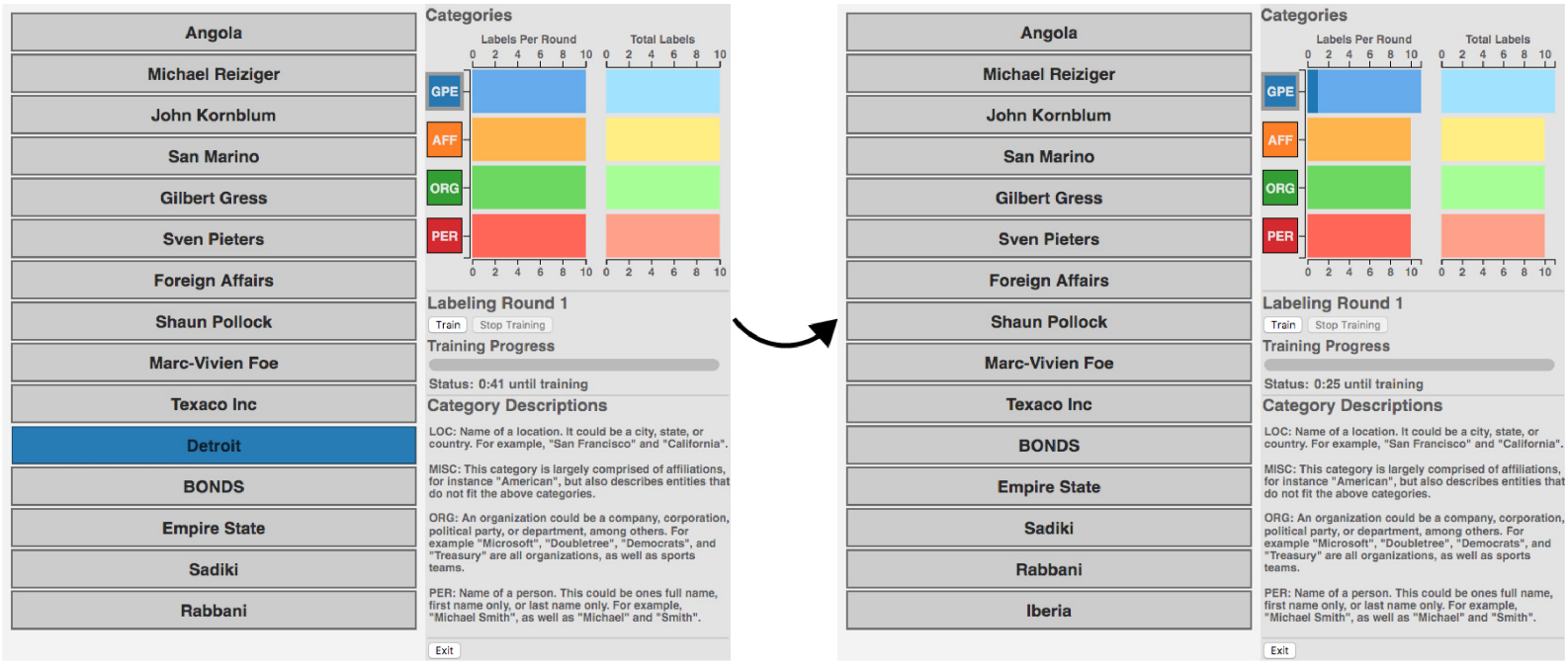
Visual Interactions



We have provided a visual interface for you to label entities. Above we show an example of the interface, where a list of entities determined as most informative by the model are shown on the left.

The labeling interface is shown to the right. In the upper portion of the interface the set of categories are shown, along with statistics on the number of entities that have thus far been labeled, per category. This is shown per-round by individual bars with round increasing from top to bottom. Bars colored with a dark hue indicate entities that were labeled by the user, while bars with a medium hue indicate entities that were automatically labeled by the bootstrapping technique. A single large bar, with light hue, is shown that encodes the total number of entities labeled over all rounds. This visualization is dynamically updated as rounds proceed, and as the user provides labels.

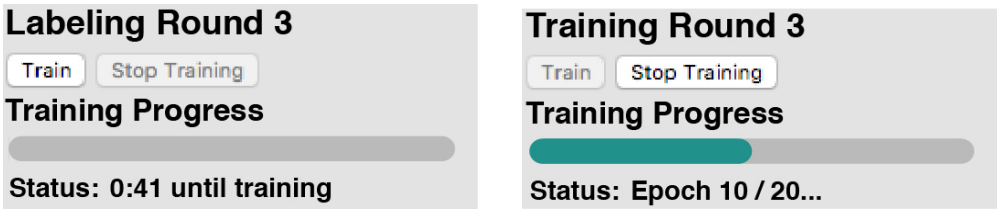
Labeling Entities



You can provide a categorical label to an entity by first clicking on the category of choice in the Category menu, and then clicking on the entity in the list. Upon clicking on the entity, the list will update to show the next most informative entity. This is shown in the above figure.

You may also select multiple categories by holding the Shift key, and clicking on entities. Clicking on a selected entity will deselect the entity from that category. Upon releasing Shift, the entities will be assigned the selected category, and the list will update, as shown above.

Updating the Model



In the above figure we show the labeling interface for the model update. The user has at most 1 minute to label entities, during which time the model is not updated. On the left we show the corresponding status: the title at the top shows the mode ("Labeling") and round number, and a timer countdown is included to show how long until the next round of updates occurs. At the end of 1 minute the model will update, incorporating the labeled entities that you provided. The figure on the right shows the update status, displaying how many iterations the update has performed and how many more iterations it has to go. If you wish to update the model prior to the prespecified maximum number of iterations, you may click on the "Train" button to initiate the model update. When the model updates, the list of most informative entities will change. Updates will occur over a certain number of iterations, but if you wish to stop training then click the "Stop Training" button. A potential cue to determine when to stop training is how much the list of informative entities changes per model update.

Tutorial

We have provided below a simple tutorial for you to get accustomed to the user interface. In this tutorial, you will be responsible for assigning each entity to its corresponding category. Once all entities have been assigned, the list view will be empty and the "Train" button will be enabled. Clicking this button will show the amount of entities that you correctly labeled. Furthermore, shortly after clicking the button, the main interface will load and the study will begin. The entities in the tutorial will serve as the initial set of labeled entities for the model.

Instructions

Here we provide instructions for the study you are about to take in using visualization to help bootstrapped information extraction. We will go over a description of the task, the dataset that you will be interacting with, a description of the visual interface and how to perform interactions, followed by a short demo that you will have to complete before moving on to the main experiment.

Task Objective

Your objective in this task is to accurately assign labels to entities as fast as possible, and as many as you can.

Task Description

The goal of this study is to assess the effectiveness of different user interfaces that enable a human to provide labels for the purposes of training a machine learning system. The task we are considering is that of **automatic labeling of entities**. An entity is a noun phrase in an expression that is distinguished by its *category*, for instance an entity could be an affiliation, a location, an organization, or a person.

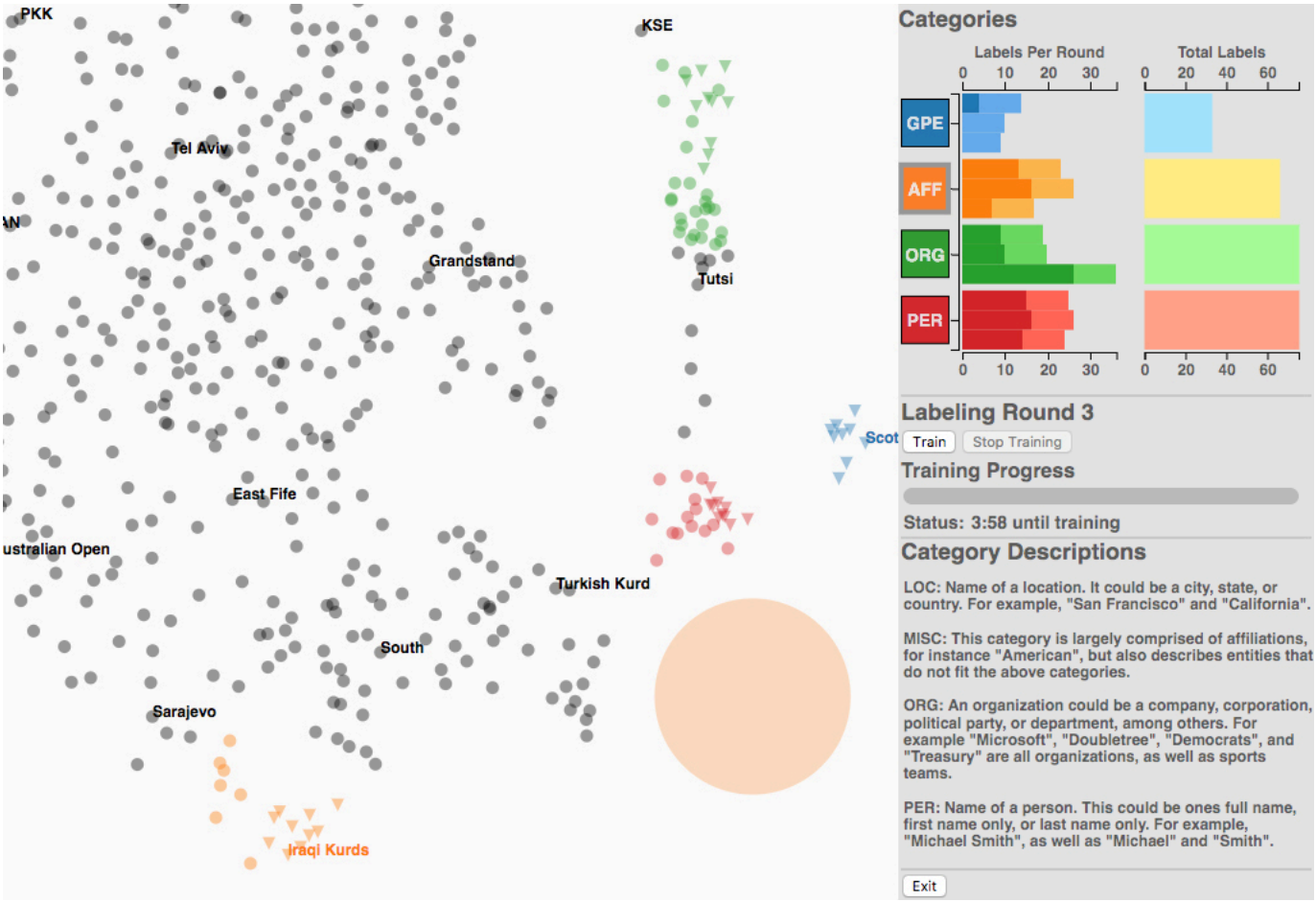
Training a machine learning system amounts to computing a model given a set of labeled entity examples, which will be your responsibility to provide. For instance, if you see the entity "George Washington", then the label you would provide would be **Person**. However, you will not provide a complete set of labels in one sitting. Instead, you will provide labels **in rounds**. Namely, you will label entities for at most 1 minute, and then the model will update, based on the labels you just provided as well as automatically discovering other labels. After the update, the next round begins, and you will be presented with a new view of the entities for labeling. You will label entities for a total of 10 rounds.

Dataset

The OntoNotes v5.0 (2012) dataset is used for this study. This dataset is a collection of text from news, broadcast, talk shows, weblogs, usenet newsgroups, and conversational telephone speech. We use 4 of the 11 categories from the dataset to distinguish entities:

- **Affiliation**: This category is largely comprised of affiliations, including nationalities, religious groups, or political groups. For instance "American" is an affiliation.
- **Geopolitical Entity (GPE)**: This is the name of a location. It could be a city, state, or country. For instance, "San Francisco" and "California" would be locations.
- **Organization**: An organization could be a company, agency, institution, or a department, among others. For instance "Microsoft", "Doubletree", "Democrats", and "Treasury" are all organizations, as well as sports teams.
- **Person**: This is the name of a person. It could be ones full name, first name only, or last name only. For example, "Michael Smith" would be an entity with category person, but so are "Michael" and "Smith". Fictional names are also included.

Visual Interactions



We have provided a visual interface for you to label entities. Above we show an example of the interface, where the entities are shown in the left in a 2D plot. Each entity is represented by either a circle or a triangle, and its specific text shown slightly offset to its upper right.

The labeling interface is shown to the right. In the upper portion of the interface the set of categories are shown, along with statistics on the number of entities that have thus far been labeled, per category. This is shown per-round by individual bars with round increasing from top to bottom. Bars colored with a dark hue indicate entities that were labeled by the user, while bars with a medium hue indicate entities that were automatically labeled by the bootstrapping technique. A single large bar, with light hue, is shown that encodes the total number of entities labeled over all rounds. This visualization is dynamically updated as rounds proceed, and as the user provides labels.

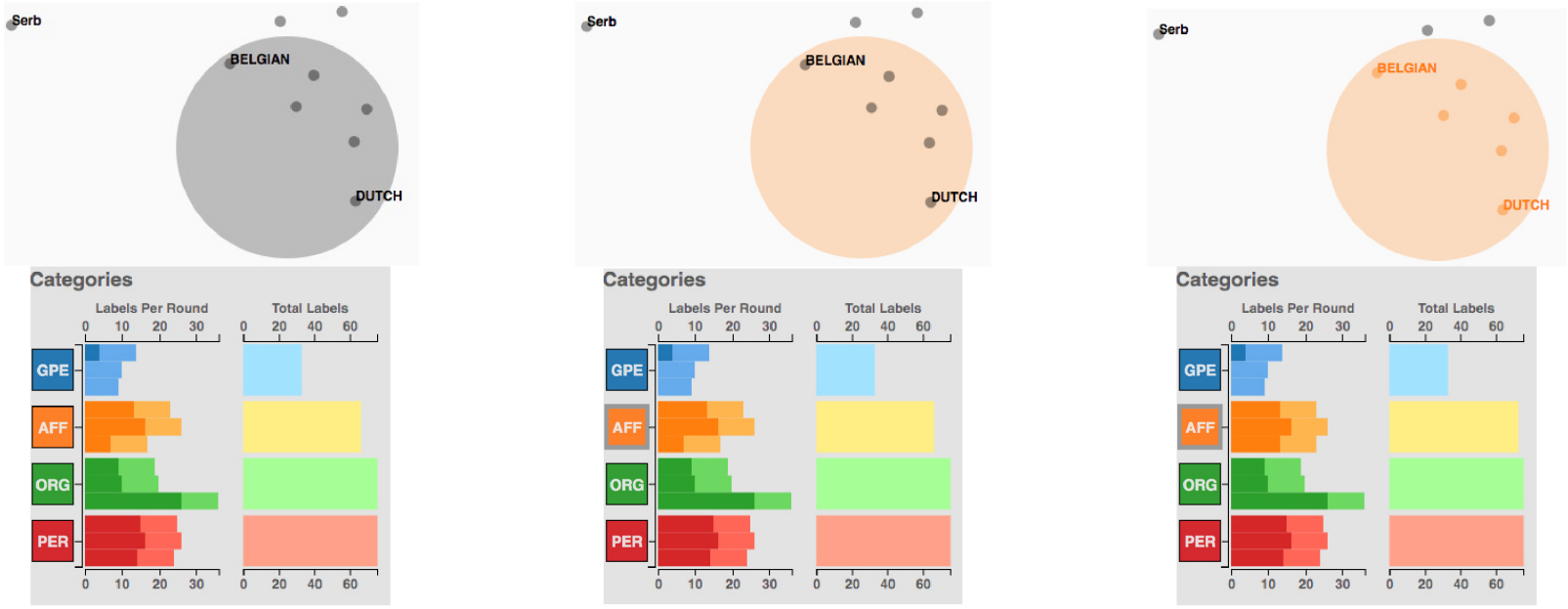
Changing the View

You can browse the 2D plot and change its view with the following set of controls:

- **Left click drag**: This will translate the view of the plot.
- **Mouse wheel**: This will zoom in to / out of the plot.

As you adjust the zoom level of the plot, text labels will be dynamically shown in order to minimize clutter. However, hovering the mouse over individual circles will display the text for those circles, and hide the text for any other occluding labels.

Labeling Entities

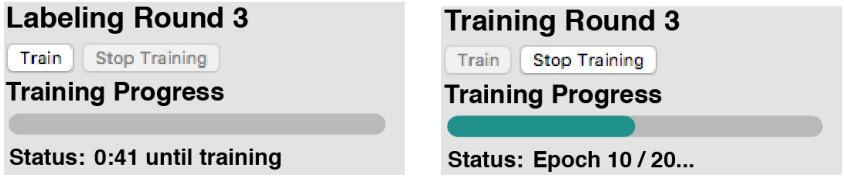


You can provide a categorical label to an entity by first clicking on the category of choice in the Category menu, and then "brush" the label onto the entities in the 2D plot. Brushing is controlled by a circle tool, which has the following options:

- **Shift + mouse movement**: This will show the area of the brush, but will not yet assign labels. It will be colored by the selected category. Above, on the left we show this when no category is selected, which will declare an entity as unlabeled. In the middle the user has clicked on the "Miscellaneous" **FIX** category box, and the brush area's color has been updated to reflect this category's color.
- **Shift + left click drag**: This will assign the selected category to the set of entities that lie within the brush selection. On the right the user has assigned to the set of entities the category "Miscellaneous". **FIX** The entity colors are update accordingly, and the plot showing the number of entities currently labeled is also updated.
- **Shift + mouse wheel**: This will adjust the size of the brush.

Furthermore, if you make a mistake in labeling, simply click the category currently selected. Brushing in this mode will assign entities to be unlabeled, or not yet belonging to any category.

Updating the Model



In the above figure we show the labeling interface for the model update. The user has at most 1 minute to label entities, during which time the model is not updated. On the left we show the corresponding status: the title at the top shows the mode ("Labeling") and round number, and a timer countdown is included to show how long until the next round of updates occurs. At the end of 1 minute the model will update, incorporating the labeled entities that you provided. The figure on the right shows the update status, displaying how many iterations the update has performed and how many more iterations it has to go. If you wish to update the model prior to the prespecified maximum number of iterations, you may click on the "Train" button to initiate the model update. When the model updates, both the entities and their positions will dynamically update. Updates will occur over a certain number of iterations, but if you wish to stop training then click the "Stop Training" button. A potential cue to determine when to stop training is how much the entities are changing in their position in the 2D plot - this suggests that the model is not significantly changing.

When should I advance to the next round?

As mentioned above in the task objective, we want you to assign as many labels as possible, as quickly as possible. In particular, you may move on to the next round at any point prior to the 1 minute duration. To help you determine when you should move on, we want to emphasize **group labeling** of entities. We suggest prioritizing labels based on **groups of entities** that you think belong to the same label, and assigning all of the entities to a label with a single brush. This will maximize the amount of labels that you can provide. If at any time you are unable to identify groups of entities, as nearby entities appear to belong to different categories, and you begin to assign labels to single entities, then we suggest clicking the "Train" button.

Note that as rounds proceed, the model will ideally improve and provide better groupings of entities, thus making it easier to perform group-wise labeling. Thus, do not hesitate to click the "Train" button if you find it difficult to accurately provide lots of labels in the 2D plot. Further rounds should help provide better groupings.

Tutorial

We have provided below a simple tutorial for you to get accustomed to the user interface. In this tutorial, you will be responsible for assigning each entity to its corresponding category, modifying the view and performing labeling as appropriate. Once all entities have been assigned a label the "Train" button will be enabled, and clicking this button will show the amount of entities that you correctly labeled. Furthermore, shortly after clicking the button, the main interface will load and the study will begin. The entities in the tutorial will serve as the initial set of labeled entities for the model.