

---

# **SMART Documentation**

***Release 0.0.1***

**Rob Chew**

**Sep 28, 2018**



1	Feature Highlights	3
2	Quick Start	5



SMART is an open source application designed to help data scientists and research teams efficiently build labeled training datasets for supervised machine learning tasks.



# CHAPTER 1

---

## Feature Highlights

---

- **Active Learning** algorithms for selecting the next batch of data to label.
- **Inter-rater reliability** metrics to help determine a human-level baseline and the understand the test validity of your labeling task.
- **Admin dashboard** and other project management tools to help oversee the labeling process and coder progress.
- **Multi-user coding**, for parallel annotation efforts within a project.
- **Self-hosted installation**, to keep sensitive data secure within your organization's firewall.





## CHAPTER 2

---

### Quick Start

---

```
$ git clone [github repo]
$ cd smart/envs/dev/
$ docker-compose build
$ docker volume create --name=vol_smart_pgdata
$ docker volume create --name=vol_smart_data
$ docker-compose up -d
```

Open your browser to <http://localhost:8000>

## 2.1 Part 1: Installation

---

**Note:** Additional installation instructions and developer notes are available at the SMART Github code repository [\[LINK\]](#).

---

To begin installing SMART, first clone the code repository to the local directory of your choice:

```
$ git clone [github repo]
```

SMART uses Docker in development to aid in dependency management. First, install [Docker](#) and [Docker Compose](#). Then navigate to `envs/[dev|prod]` and run `docker-compose build` to build all the images:

```
$ cd smart/envs/dev
$ docker-compose build
```

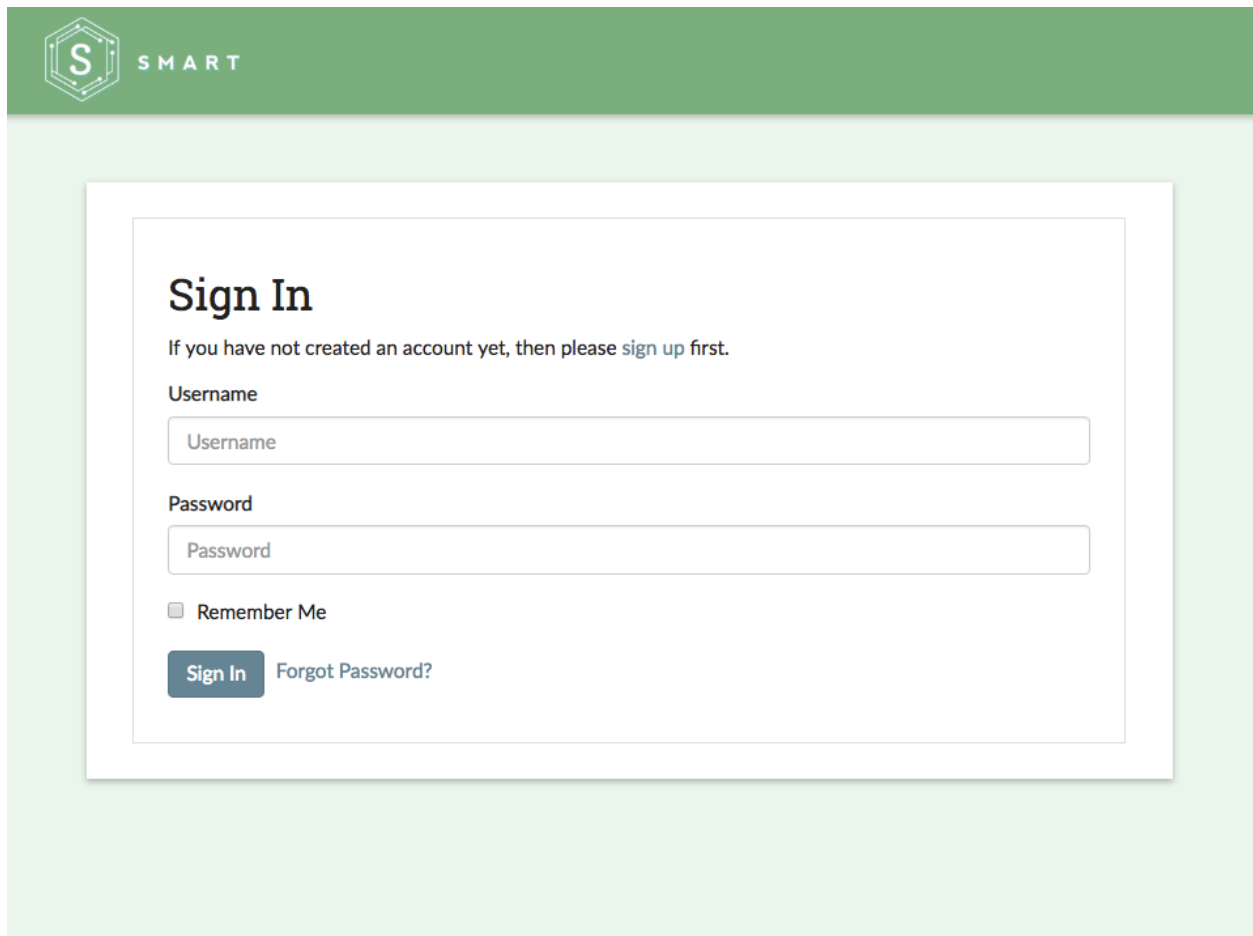
Next, create the docker volumes where persistent data will be stored: `docker volume create --name=vol_smart_pgdata` and `docker volume create --name=vol_smart_data`.

```
$ docker volume create --name=vol_smart_pgdata
$ docker volume create --name=vol_smart_data
```

Lastly, run `docker-compose up` to start all docker containers. This will start up the containers in the foreground so you can see the logs. If you prefer to run the containers in the background use `docker-compose up -d`. When switching between branches there is no need to run any additional commands (except build if there is dependency change).

```
$ docker compose up -d
```

To see SMART in action, navigate to <http://localhost:8000> in your web browser of choice. You should be welcomed by the SMART login screen:

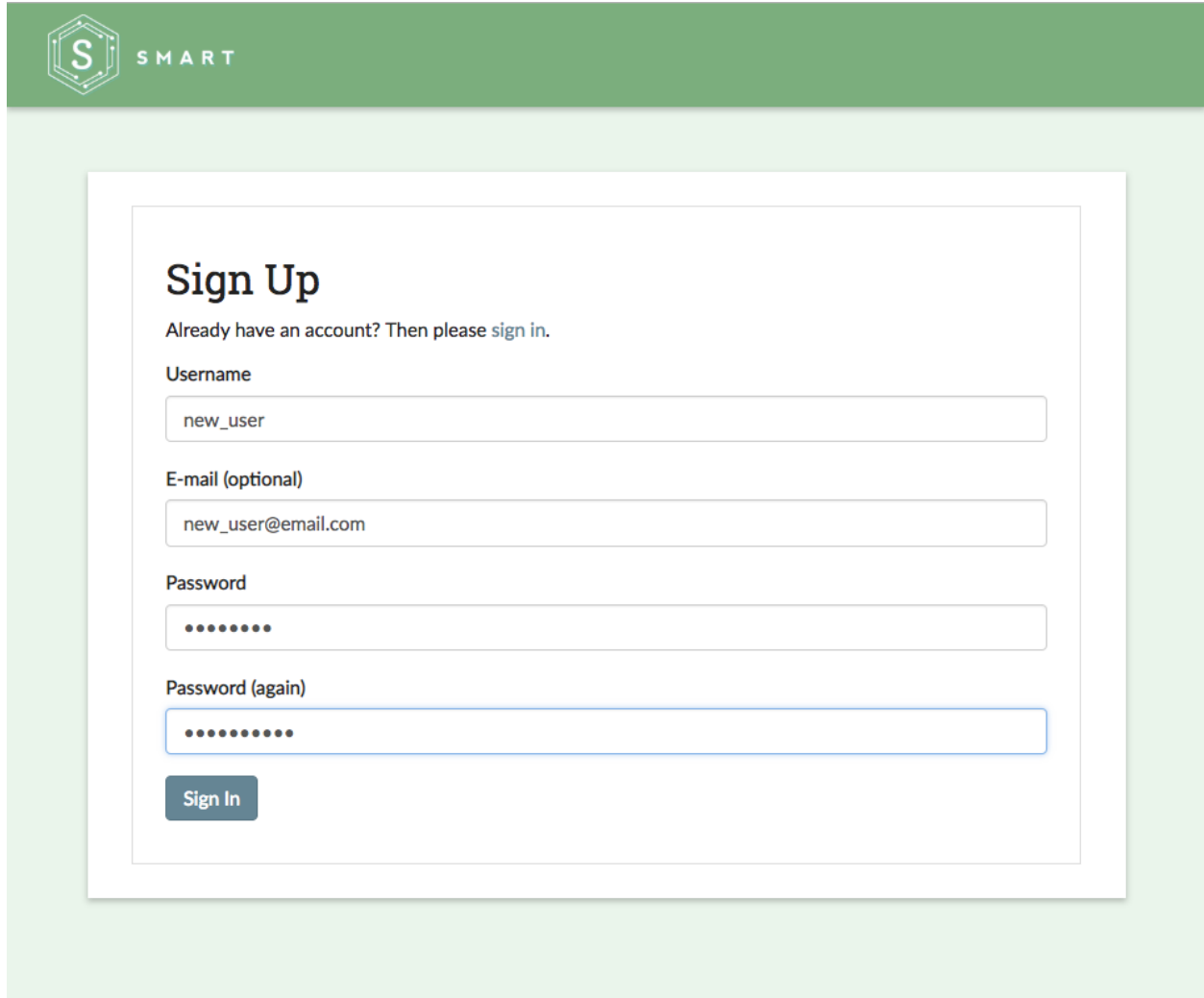
The image shows the SMART login interface. At the top, there is a green header bar with the SMART logo (a stylized 'S' inside a hexagon) and the word 'SMART' in white. Below the header, the main content area has a light green background. In the center, there is a white rectangular box with a thin gray border. Inside this box, the title 'Sign In' is displayed in a large, bold, black font. Below the title, a message reads: 'If you have not created an account yet, then please [sign up](#) first.' Below this message are two input fields: 'Username' and 'Password', each with a placeholder text of the same name. Under the password field, there is a checkbox labeled 'Remember Me'. At the bottom of the form, there is a dark blue 'Sign In' button and a blue link labeled 'Forgot Password?'.

---

**Note:** By default, SMART will use port 8000 for the front-end and port 5432 for the back-end processes. See the SMART code repository README for instructions on how to change the default ports.

---

Finally, create a profile to start your own new labelling projects or to be added to an existing one:



The image shows a web interface for signing up. At the top, there is a green header bar with the SMART logo on the left, which consists of a hexagon containing an 'S' and the word 'SMART' to its right. Below the header, the main content area has a light green background. Centered in this area is a white rectangular box with a subtle drop shadow. Inside this box, the title 'Sign Up' is displayed in a large, bold, black font. Below the title, a line of text reads 'Already have an account? Then please [sign in](#).' in a smaller black font. The form contains four input fields: 'Username' with the text 'new\_user', 'E-mail (optional)' with the text 'new\_user@email.com', 'Password' with seven dots, and 'Password (again)' with seven dots. The 'Password (again)' field is highlighted with a blue border. At the bottom left of the form is a dark blue button with the text 'Sign In' in white.

**Sign Up**

Already have an account? Then please [sign in](#).

Username

new\_user

E-mail (optional)

new\_user@email.com

Password

.....

Password (again)

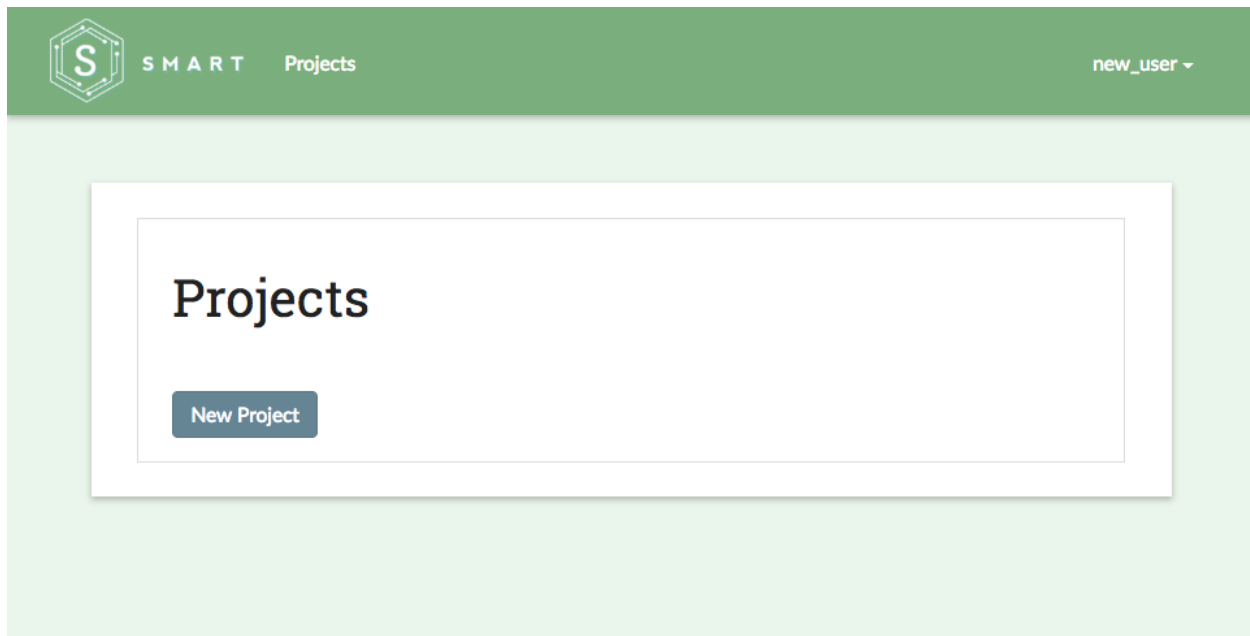
.....

Sign In

## 2.2 Part 2: Creating a New Project

Starting a new labelling project in SMART is as easy as pressing the “New Project” button on the SMART landing page. All users have the ability to start their own coding projects, though they may be restricted to modifying or deleting existing projects depending on their user roles.

For new users without any existing projects, the SMART landing page should look like this:



For the purpose of this tutorial, we'll create a project to classify tweets as being either *Hotdog* or *Not Hotdog* related (hat tip to HBO's *Silicon Valley*). This project is called the "Not Hotdog" Classifier.

### 2.2.1 Project Description

The first step in creating your project is to provide a project name and description. The name will be the internal reference for the project which users will see on their landing pages and the description will be available for users on the project Details page. Below, we fill out the name and description for our "Not Hotdog" classifier:

A screenshot of the 'Project Info' form. The title 'Project Info' is at the top. Below it is a 'Name' label and a text input field containing 'Not Hotdog'. Then there is a 'Description' label and a text area containing '"What would you say if I told you there is an app on the market that tell you if you have a hotdog or not a hotdog..."'. At the bottom left is a 'Next Step' button, and at the bottom right is the text 'Step 1 of 6'.

## 2.2.2 Creating Label Definitions

In the Labels section, we will create categories for labeling. These labeled observations will be used to train a classification model that predicts what category (*Hotdog* or *Not Hotdog*) a new observation belongs. To add new categories, just fill-in the names of the categories you're interested in predicting into the input boxes. If you have more than two labels, use the “add label” link to add more rows to the form. If you decide that you want to remove a label after adding it, use the “remove label” link to remove the label name.

Name	Description	Delete
<input type="text" value="hot dog"/>	<input type="text" value="This is a hotdog"/>	
<input type="text" value="not hotdog"/>	<input type="text" value="This is not a hotdog"/>	

[add label](#)

1. Info [Next Step](#) Step 2 of 6

### Note:

- SMART requires at least two category labels and the labels must be unique.
- If you plan on uploading a data file that contains labels, the label categories in the file must match those provided on this page.

### Warning:

- You cannot add, remove, or update any labels for a project after the project is created.

## 2.2.3 Project Permissions

To help organize your labeling projects, you can assign special permissions to other project members. Project members can be assigned one of two user-roles:

- **Admins** are able to update the project description, upload additional data, control project permissions, and annotate data.
- **Coders** are able to view project details and annotate data.

In this panel, you can select project members and assign their role types. Clicking the “add permissions” link adds more rows to the form. If you decide that you want to remove a permission after adding it, click the “remove permission” link next to the inputs to remove the permission. If an intended project member is not listed below, please check to see if they have created an account.

In the development environment, SMART includes three user profiles for testing purposes (`root`, `user1`, and `test_user`). Inviting additional users to a project is optional. For the purposes of this tutorial, we will add `user1` as a coder:

Profile	Permission	Delete
<input type="text" value="user1"/>	<input type="text" value="Coder"/>	<a href="#">remove permission</a>

[add permissions](#)

1. Info

2. Labels

Next Step

Step 3 of 6

---

**Note:**

- The project creator is always assigned Admin privileges.
  - Each user profile can only be assigned one permission type.
  - Each row must be completely filled in with both a profile and permission.
  - You can update permissions after creating a project.
- 

### 2.2.4 Advanced Settings

The Advanced Settings page allows you to customize your labelling experience and utilize advanced features such as *Active Learning* or *Inter-rater Reliability (IRR)*. For the tutorial, we’ll keep the default settings, but please reference the *Advanced Feature Details* section of the documentation to learn more about these and other options.

# Advanced Settings

## Active Learning

☒ Use Active Learning to select data to label

☒ By Uncertainty using Least Confident

☐ By Uncertainty using the Margin

☐ By Uncertainty using Entropy

## Model Selection

Choose the classification algorithm used by active learning algorithms and for general prediction.

☒ Logistic Regression (default)

☐ Support Vector Machine (warning: slower for large datasets)

☐ Random Forest

☐ Gaussian Naive Bayes

## Inter-rater Reliability (IRR) Settings

Under IRR, a certain percentage of the data is labeled by multiple coders. The project admin can then examine the consistency of the labels across different coders. The options below allow you to set what percentage of the data is coded multiple times, and how many coders must code data designated for IRR before it is analyzed.

☐ Use Inter-rater Reliability

## Batch Settings

Use default size (10 times the number of labels): ☒

1. Info 2. Labels 3. Permissions Next Step

Step 4 of 6

## 2.2.5 Adding Codebook

This page gives you the opportunity to upload extra information for coders that maybe be helpful for clarifying the labelling task (ex: tips for differentiating categories, examples of labeled data, etc.). This is particularly useful if the

categories you're interested in labelling are numerous or nuanced.

A demo codebook for the tutorial can be found in the `smart/demo/` directory. To upload the codebook, click the “Choose File” button and select `hotdog-codebook.pdf`:

## Codebook Upload Instructions

### Description

This page gives you the opportunity to upload extra information for coders (ex: tips or hints differentiating different labels, examples of labeled data, etc.). To upload, the file must pass the following checks:

The file must be a pdf

#### Data

Choose File

hotdog-codebook

1. Info

2. Labels

3. Permissions

4. Advanced

Next Step

Step 5 of 6

---

**Note:** The codebook file must be a PDF.

---

## 2.2.6 Upload Data

Time to upload your data!

To upload, the data file must pass the following checks:

- The file needs to have either a `.csv`, `.tsv`, or `.xlsx` file extension.
- The file requires the data to be formatted into two columns, with header names `Text` and `Label` OR three columns with header names `ID`, `Text`, and `Label`.
- The largest file size supported is 4GBs.
- The (optional) ID column should contain a unique identifier for your data. The identifiers should be no more than 128 characters.

The `Text` column should contain the text you wish users to label. For our “Not Hotdog” classifier, the `Text` column would contain the tweet text.

The `Label` column should contain any pre-existing labels for the corresponding text. If none of your data contains existing labels, then this column can be left blank. Extending our example, if a lead coder has already annotated some tweets as *Hotdog* or *Not Hotdog*, this column would contain those labeled records.

The data used in this tutorial is shipped with SMART and can be found in the `smart/demo/` directory. To upload this file, click the “Choose File” button and select `hotdog-example.csv`:



*Data Upload Notes:*

SMART restricts your project to having two million unique records.

If there are multiple rows with the same text, only one of the records will be saved.

SMART will keep up to two million unique records per data set.

**Data**

hotdog-example.csv

1. Info

2. Labels

3. Permissions

4. Advanced

5. CodeBook

Submit

Step 6 of 6

**Tip:**

- SMART will keep up to two million unique records per data set.
- If there are multiple rows with the same text, only one of the records will be saved.
- You may add a dataset that already contains labelled observations. However, all labels present in the upload file must be in list of categories assigned in the *Creating Label Definitions* step.

## 2.3 Part 3: Reviewing Projects & Editing Project Settings

### 2.3.1 Projects Page

The projects page serves as the central page for a SMART user. The page provides a list of all projects the user is on, and provides links to major parts of each project. Users with admin privileges will be able to see links to a project's respective *Admin Dashboard* as well as the *Download Labeled Data and/or Model* button. Coders will only see the *Details Page* and *Annotate Data Page* links. This is also the page where you go to *Create a New Project*.

In the image below, the new user is an admin for all projects except for the *Pepsi* or *Coke?* project.

## Projects

Not Hotdog	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	<a href="#">Download Model and Labeled Data</a>
Not Hotdog (plus hamburger)	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	<a href="#">Download Labeled Data</a>
Pepsi or Coke?	<a href="#">Details</a>	<a href="#">Annotate</a>		
Whizzo Butter or a Dead Crab?	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	No Labeled Data to Download

New Project

### 2.3.2 Details Page

The Details page provides an overview of the information and settings for your project. Each project has its own Details page, which is created when you start a new project. You can navigate to any project Details page from the [Projects Page](#) or to a specific Details page by pressing the “Details” link on the top navigation bar when on a project [Annotate Data Page](#) or [Admin Dashboard](#) page.

**The Details page lets you review:**

- The project Description

Description

"What would you say if I told you there is an app on the market that tell you if you have a hotdog or not a hotdog..."

- What permissions have been assigned to what users

Permissions


Admin(s)

new\_user


Coder(s)

user1


- The advanced settings (i.e. [Active Learning](#), [Inter-rater Reliability \(IRR\)](#), classifier, batch size)

Advanced Project Settings 	
Selection Algorithm	least confident
Batch Size	20
Percent IRR	5.0%
Number of users for IRR	2

- The labels being used and their descriptions (if applicable)

Labels 	
hot dog	This is a hotdog
not hotdog	This is not a hotdog

- A sample of your data

<u>Data</u> 	
@BenSimmons25	would you rather eat a hotdog for every meal for the rest of your life or be a hotdog?
RT @matthewcpinsent:	Trump heard about the Kushner/RUS meeting, then didnt, then it was unimpo...
Now that was fun!	<a href="https://t.co/UvC8hMb54x">https://t.co/UvC8hMb54x</a>
Add Beauty & Charm to Your Home with New Wood Flooring	<a href="https://t.co/v81GCTPLDT">https://t.co/v81GCTPLDT</a> <a href="https://...">https://...</a>
"Don't choke on a hotdog."	- @PlayerEssence 2018

At the bottom of the Details page, there are buttons to delete the project, edit

the project settings, or download the labeled data and (if applicable) trained model. These buttons are only visible to users with admin privileges for the project.

[Return to Projects](#)[Update Project](#)[Delete Project](#)[Admin Page](#)[No Labeled Data to Download](#)

### 2.3.3 Updating a Project

The Update Project page is accessible from the *Details Page* of a project. This page can be used for the following operations:

- Edit the project name and description
- Add additional data to label
- Add or change the codebook file
- Add, remove, or change project permissions
- Edit label *descriptions*

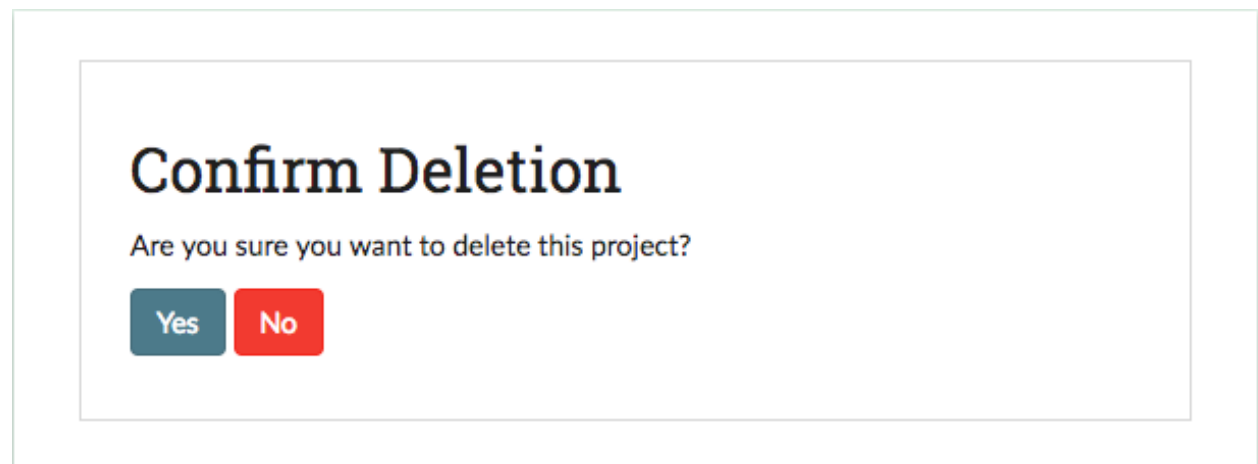
---

**Tip:**

- SMART allows up to two million records total. This includes additional data added later.
  - New data is checked against existing data for duplication.
- 

### 2.3.4 Deleting a Project

The button to delete a project can be found on the *Details Page* of a project. To delete a project, click this button and then select “yes” at the prompt.



## 2.4 Part 4: Annotating Data

Once your project has been created, you are ready to start labeling your data! To begin, you can navigate to any project Annotation page from the *Projects Page* or to a specific Annotation page by pressing the “Annotate” link on the top navigation bar when on a project *Details Page* or *Admin Dashboard* page.

The Annotate page consists of either two or five tabs, depending on your permission User. The sections below are marked by either ADMIN (available only to those with admin privileges) or ALL (available to everyone with at least coder privileges).

**Note:** If a user with admin privileges is on the annotation page, then other admin will be unable to access admin-only tabs until the first admin has left the page. This is to prevent multiple admin from labeling the same data simultaneously. Coders and Admin can always access the Annotate Data and History pages. See [user-roles](#) for a chart of user permissions.

In addition, each tab has access to the project's [Label Guide \(feature\)](#) and [Codebook \(feature\)](#) using the buttons shown below:

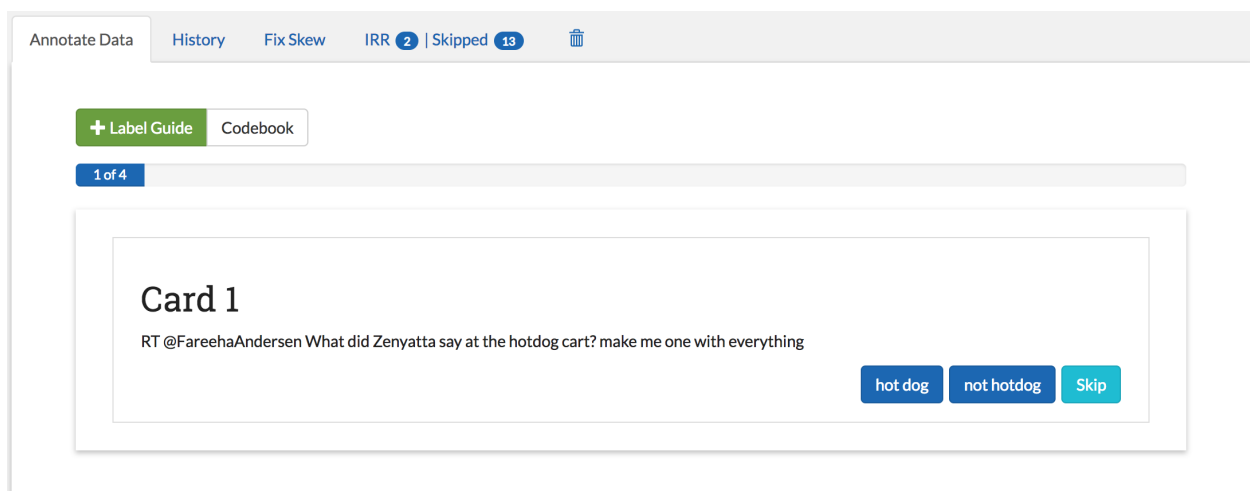


## 2.4.1 Annotate Data Page

**User: ALL**

The Annotate Data tab is where most users will spend a majority of their time. When you enter this page, SMART will pass you a portion of the current batch as a deck of “cards”, to be presented to you one at a time. You can then choose one of two actions:

- *label*: Assign a label to the piece of data by clicking on the button corresponding to the desired label. If the data is not being used for [Inter-rater Reliability \(IRR\)](#), then this data will be marked as labeled and removed from the pool of unlabeled data. If data is IRR, then it may still be presented to additional coders on the project, but will not be presented to you again.
- *skip*: Skip the data. This option is used when you are unsure of what label to choose. Skipped data that is not IRR is sent to the Admin Annotation page to be reviewed by any user with admin privileges.



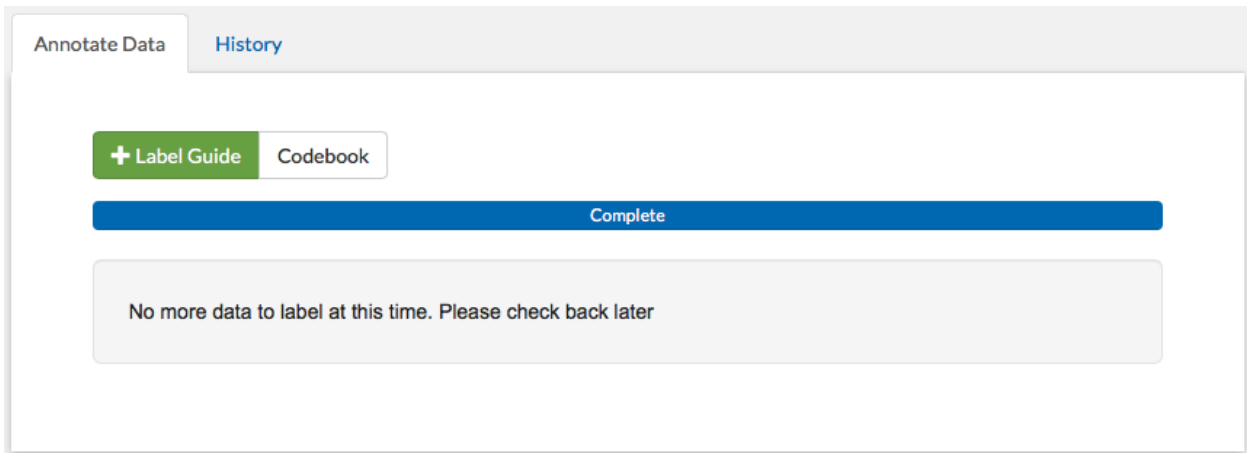
### Refilling the Batch

A user's card deck will continue to refill itself from the batch until it is empty. Once a batch has been coded or skipped, a new batch of unlabeled data will be requested from SMART. This batch will be selected using the chosen

active learning algorithm, or randomly, depending on if Active Learning was enabled in [Advanced Settings](#). The batch may also be selected randomly for three other reasons:

- It is the first batch
- Each possible label has not been used at least once
- There has not been a full batch worth of data marked as labeled (possibly some was skipped or is IRR and waiting for additional labels)

If a model is currently running, then this batch will be delayed until the model has finished running, and you will be presented with the message in the image below. Note that this does not apply to projects that have disabled having a model. Projects that have disabled Active Learning but have a model will still have to wait for the model to run, but it will be done faster as predictions will not have to be generated for the unlabeled data (see [Part 5: Administrator Dashboard](#) for more details).



---

**Tip:** If you are seeing the message above, try refreshing the page. The batch might have become available after the application was last queried. If the message is still there, then wait a few minutes for the model to finish and refresh again.

---

## 2.4.2 History Page

**User:** ALL

Perhaps you have been happily coding your data and you accidentally click the wrong label. Now you have data labeled “hotdog” which is decidedly *not about hotdogs!* Or perhaps you have labeled a number of items when your project leader announces that from this day forth, *chedderwursts will also be counted as hotdogs!* The history tab exists for scenarios like these ones. In this tab, you are able to view and edit your past labels. This page includes all data that has been labeled by you personally, and provides the text, past label, and date/time of the most recent label.

The history table is automatically sorted by the date to provide the most recent labels first [see [Searching and Sorting \(feature\)](#)].

Annotate Data
History
Fix Skew
IRR 0 | Skipped 6

## Instructions

This page allows a coder to change past labels.

To annotate, click on a data entry below and select the label from the expanded list of labels. The chart will then update with the new label and current timestamp

**NOTE:** Data labels that are changed on this page will not effect past model accuracy or data selected by active learning in the past. The training data will only be updated for the next run of the model

+ Label Guide
Codebook

	Data	Old Label	Date/Time
▶	RT @overlyexclusive: believe in yourself, believe in what youre doing. believe i...	not hotdog	2018-08-16, 12:27.49
▶	RT @Jack_Stewart242 Barney finally cought the elusive hotdog unicorn, the...	not hotdog	2018-08-16, 12:27.48
▶	RT @FareehaAndersen What did Zenyatta say at the hotdog cart? make me o...	not hotdog	2018-08-16, 12:27.48
▶	RT @Arielmonaye LMAO! No that is not a hotdog and lucky charms https://t....	hot dog	2018-08-16, 12:27.47
▶	@MelAHaughty Let's make it happen ____! been saying this since you got it ...	not hotdog	2018-08-15, 21:42.51
▶	SeaWorld was fun but damn one hotdog was \$11.99 chicken strips was 13 d...	hot dog	2018-08-15, 21:42.48

To save space, the history table only includes enough text for each data sample to fit the page width. To expand a row for reading and editing, click on the arrow to the left of the text. This will open up a subrow with the entire text and the label/skip options. Note that changing a label to skip will remove it from the history table as you have effectively given up responsibility for it.

▼
RT @FareehaAndersen What did Zenyatta say at the hotdog cart? make me o...
not hotdog
2018-08-16, 12:27.48

RT @FareehaAndersen What did Zenyatta say at the hotdog cart? make me one with everything

hot dog
not hotdog
Skip

**Note:** *Inter-rater Reliability (IRR)* data labels can be changed in the history table up until the point where enough people have labeled/skipped it and it is processed. At this point, the data is effectively “labeled by everyone” (either from consensus or from an admin resolving a dispute) and will no longer be editable on anyone’s history table. Expanding a resolved IRR datum will simply show a message (see below):

RT @FunnyAsianDude The Not Hotdog app is nominated for an Emmy. Seriously. big ups to the entire staff at @HBO #Emmys2018 #nothotdog @SiliconHBO https://t.co/t0bGzzlXdo

Note: This is Inter-rater Reliability data and is not editable.

**Warning:** *For Active Learning Users:* Active learning algorithms use past labeled data to select future batches. Data labels changed retroactively will appear in the training data for the next batch, but will not effect past batches or the current batch. Excessive amounts of label changing may confuse active learning algorithms and make them less effective (see [Active Learning](#) for more details)

### 2.4.3 Fix Skew Page

User: ADMIN

Suppose your project not only includes the labels “hot dog” and “not hotdog”, but also “hamburger” (since hamburgers are the natural partner of hotdogs and therefore a separate category). The only problem is that hamburgers are fairly rare in your data, and nobody has seen one yet! You know your classifier won’t even run until a hamburger has been found (see [Refilling the Batch](#)), but you are worried that waiting for random selection to find a hamburger might take a while. The “Fix Skew” page exists for this scenario. In this tab, users with admin privileges may search unlabeled data directly for examples of rare labels. The graph on the right side of the page shows the current counts for each label (see image below).



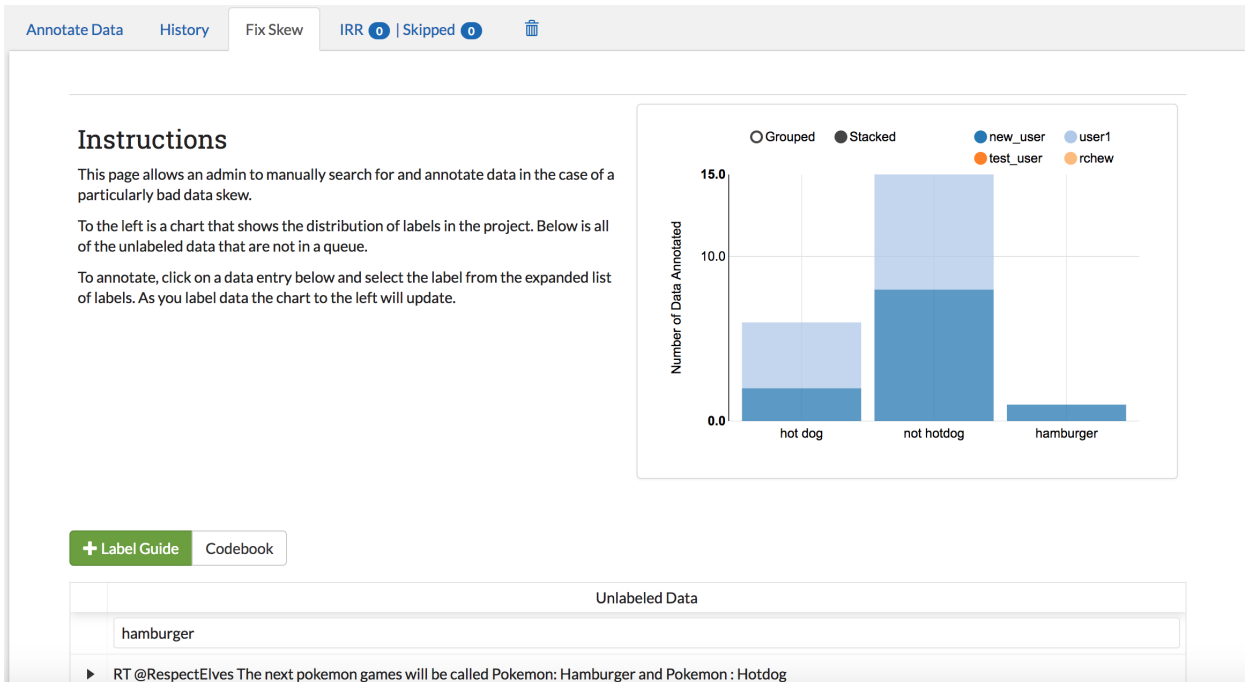
To fix a skew, follow these steps:

1. Use the search bar below “Unlabeled Data” to search the data for keywords [see [Searching and Sorting \(feature\)](#) for more information].
2. Click on the arrow to the left of the row to expand
3. Assign a label to the data



Unlabeled Data	
hamburger	
▶ RT @RespectElves The next pokemon games will be called Pokemon: Hamburger and Pokemon : Hotdog	
▼ I want a hamburger & a hotdog.	
I want a hamburger & a hotdog.	
<div>hot dog not hotdog hamburger</div>	

Once data has been labeled, the graph at the top will show the change in label counts:



**Warning:** The Fix Skew tab should *NOT* be used in place of the Annotate Data tab. The Fix Skew tab does not use *Inter-rater Reliability (IRR)*, or allow the option of skipping data. Excessive use of this page can also undermine the effects of Active Learning, or introduce unintended bias (since the data is chosen consciously by the user).

## 2.4.4 Admin Annotation Page

**User: ADMIN**

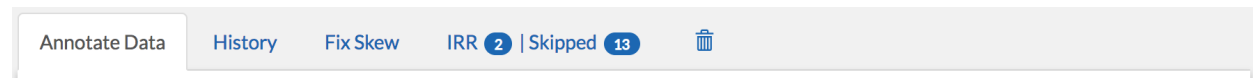
The Admin Annotation page lets users with admin User privileges resolve ambiguous data. There are two types of ambiguous data that could end up in this table.

1. Normal (not *Inter-rater Reliability (IRR)*) data that was skipped
2. *Inter-rater Reliability (IRR)* data that has been annotated/skipped by enough people, where there was either a disagreement between the assigned labels, or at least one coder skipped it (this counts as a disagreement).

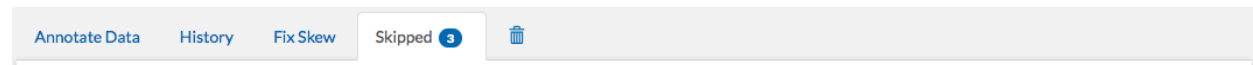
**Tip:** Coders are not given any indication of which data is being used for IRR. If you are using IRR in your project, and cannot find a specific skipped datum in the admin table, it may be IRR data that has not been seen by enough

people yet.

The Admin Annotation tab is marked with badges showing the total number of unaddressed items. For a project that uses IRR, it will look like the tab in the image below with two sections:

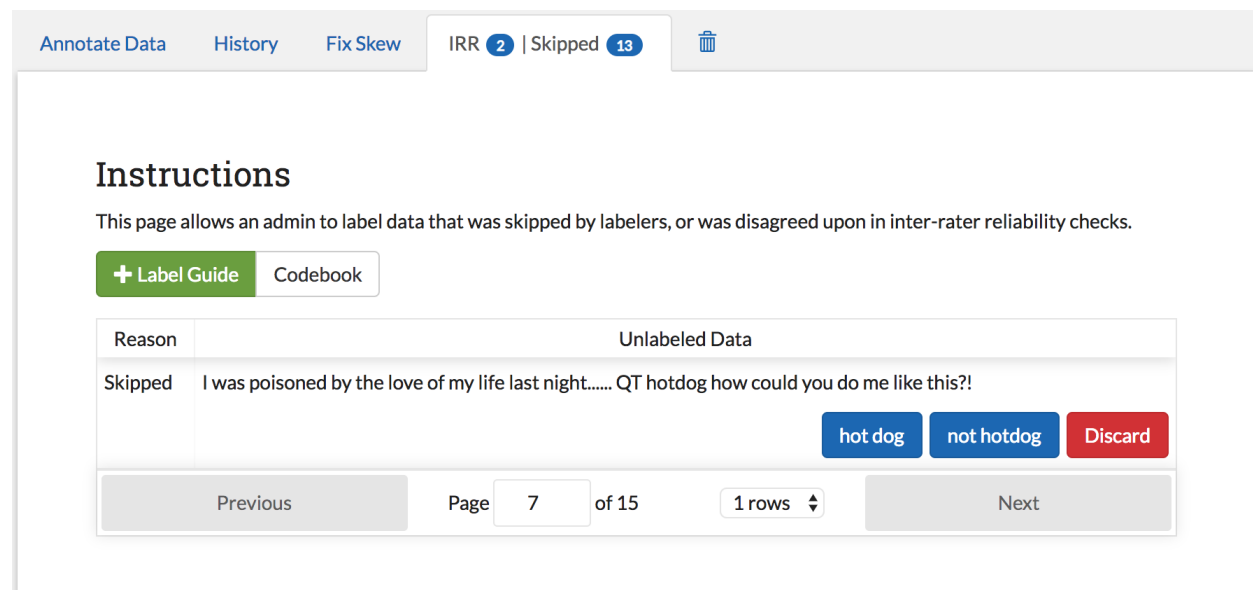


Projects that do not utilize IRR will only show the Skipped count:



The Admin Annotation page consists of a table with two columns. The first shows the reason data ended up in the table (IRR or skipped). The second gives the text for the data and provides options for how the data should be processed. The admin has two options for any data in this table:

- *label*: By clicking on one of the label buttons, the data is assigned the selected label and becomes part of the training set. If this data was skipped, then it will also become available in the admin's [History Page](#) if they want to change it later. If the data is IRR, it will also appear in their history table, but will **NOT** be editable by any user.
- *discard*: This option exists for data that is simply un-codable and should not be included in the project. Clicking this option will remove the data from any IRR records, the [Fix Skew Page](#), and any consideration for future batches. (Note that the data can be restored on the [Recycle Bin Page](#)).



## 2.4.5 Recycle Bin Page

**User: ADMIN**

The Recycle Bin page acts much like a recycle bin or trash folder for most computers. Any data that was discarded in the [Admin Annotation Page](#) will appear on this page:

**Tip:** You can search the Recycle Bin table for specific data [see [Searching and Sorting \(feature\)](#)]

Annotate Data
History
Fix Skew
IRR 1 | Skipped 6

## Instructions

This page displays all data that has been discarded by an admin.

All data in this table has been removed from the set of unlabeled data to be predicted, and will not be assigned to anyone for labeling.

To add a datum back into the project, click the Restore button next to the datum.

+ Label Guide
Codebook

Discarded Data	
▶	RT @JunkFoodMunchie Stopped Up At @ZombieDogz & Ordered A Juan Of The Dead, Victim 13 & The Walking Dead ...
▶	If you have to use two hands to hold a hotdog, its too big.
▶	I was poisoned by the love of my life last night..... QT hotdog how could you do me like this?!
▶	@Patrick_D_Ng Hello Patrick. Thanks for stopping by. Please note that Costco Canada is not removing or replacing any...

Previous
Page 1 of 1
Next

Data in the table will only be shown up to the width of the page to maximize the number of rows shown on the screen. To expand data, click the arrow on the left of the row. This will open a subrow with the entire text and a “Restore” button. Clicking on this button will remove the data from the Recycle Bin and place it back in the pool of unlabeled data for consideration.

▼
@Patrick\_D\_Ng Hello Patrick. Thanks for stopping by. Please note that Costco Canada is not removing or replacing any...

@Patrick\_D\_Ng Hello Patrick. Thanks for stopping by. Please note that Costco Canada is not removing or replacing any of its hot dogs. We will continue to offer both all beef hotdog and all beef polish. Have a great rest of the day! - Jonathan

Restore

**Note:** Restoring data will *not* restore any past records for this data. If data was marked for *Inter-rater Reliability (IRR)*, was discarded from the admin table, and then restored, any past labels or skips will not be restored with it and the data will not be marked for IRR unless it is chosen again later.

## 2.4.6 Label Guide (feature)

User: ALL

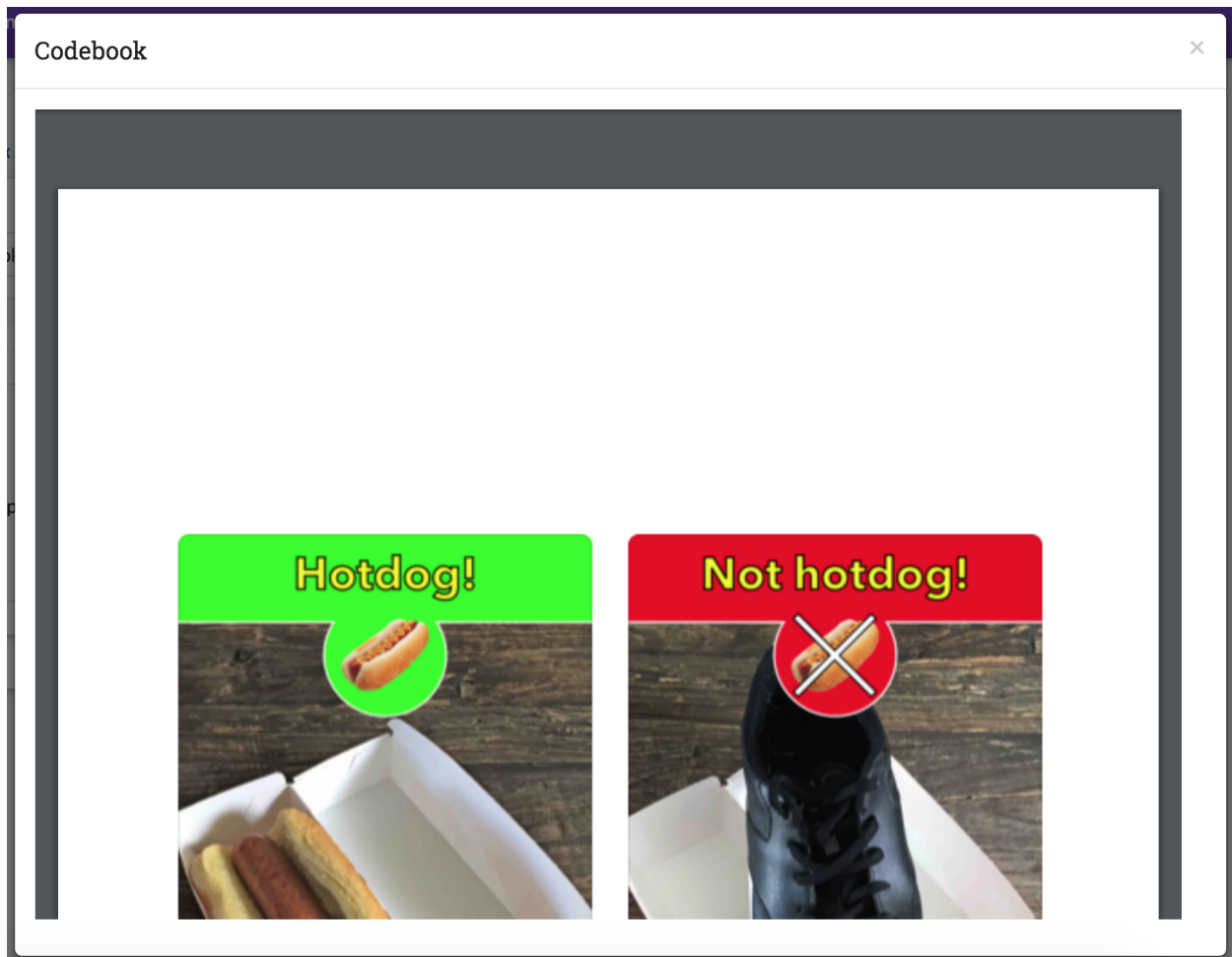
The label guide contains the list of possible labels and their descriptions as set by the project creator or updater. This guide is placed on every tab of the *Annotate Data Page* for the user’s convenience. To open the tab, click on the green + Label Guide button (see *Annotate Data Page*). The button will turn red with a minus sign as long as the guide is open (as shown below). To close, click the button again.

Label Guide	Codebook
hot dog This is a hotdog	
not hotdog This is not a hotdog	

## 2.4.7 Codebook (feature)

User: ALL

When creating or updating a project, a creator or admin has the option to add a codebook (see [Adding Codebook](#)). If a codebook has been uploaded, then in addition to the [Label Guide \(feature\)](#), a codebook button will be available on each tab of the [Annotate Data Page](#) page. To open, click the codebook button. This will open a pdf viewer on the application with the file. To close, either click the x in the top right corner of the popup, or click anywhere on the screen outside of the codebook.



**Warning:** This feature makes use of the browser's built in pdf viewer. For most modern browsers like Firefox, Chrome, or Safari, this viewer will include a print or download button. However, if you are using an outdated browser, this might not be available.

## 2.4.8 Searching and Sorting (feature)

User: ALL

You can sort any table on an annotation page by a desired column by clicking on the column header.

One click will sort it in ascending order (indicated by a grey bar at the top of the column name).

	Data	Old Label	Date/Time
▶	@alyssaavd I think imma just eat a hotdog bun and go to bed	hot dog	2018-08-15, 21:41.22
▶	@badrepbecca So are you gorgeous _□	not hotdog	2018-08-15, 21:40.20
▶	@bayley_johnson part 2 Dude my eyeball looks like a hotdog SOMEBODY HELP THIS POOR GIRL BEFORE SHE HURTS ...	hot dog	2018-08-15, 21:41.25
▶	@BeyondRGaming @Donnie_Mnemonic @TopsRTs @FatalRTs @FlyRTs @Retweet_Twitch @Twitch_RT Happy Birthday H...	hot dog	2018-08-15, 21:41.40

A second click will sort it in descending order (indicated by the grey bar below the text).

[+ Label Guide](#) [Codebook](#)

	Data	Old Label	Date/Time
▶	When #Lunchables was a highlight in your lunchbox at school. Remember the #HotDog & #Pizza packs? #GrowingUpBriti...	hot dog	2018-08-15, 21:42.41
▶	what do you put on your hotdog as a condiment, like for ketchup retweet for mustard tryna prove a point.	hot dog	2018-08-15, 21:41.08
▶	What did Zenyatta say at the hotdog cart? make me one with everything	hot dog	2018-08-15, 21:41.43

The tables on the [History Page](#), [Fix Skew Page](#), and [Recycle Bin Page](#) can be filtered using the text boxes under each column header. When text is entered in one of these boxes, only the rows containing the entered text will be displayed.

	Data	Old Label	Date/Time
	sandwich		
▶	RT @talkfastwebb is a hotdog a sandwich	hot dog	2018-08-15, 21:41.37
▶	RT @MattRAck3 Is a hotdog a sandwich? __	hot dog	2018-08-15, 21:40.56
▶	RT @coco_konski Is a hotdog a sandwich?	hot dog	2018-08-15, 21:41.33

## 2.5 Part 5: Administrator Dashboard

If you are a project admin, you may want some way to keep track of how your project is doing. The administrator dashboard allows users with admin privileges track their project's progress. Depending on the project settings, this page will have one, two, or three tabs to let them track different aspects of the project. Each of the sections below specifies what projects they apply to.

Each of the Administrator Dashboard tabs includes a table. These tables can be filtered using the text box at located at the top right. They can also be sorted by column by clicking on the column header (exactly as with the [annotation page tables](#) but instead of a grey bar there is an arrow and stair icon).

Show  entries Filter Text:

Text	Label	Coder

## 2.5.1 Labeled Data Page

### Visible for all projects

The labeled data page is designed to provide a summary of how the coders are doing comparatively in terms of speed and label distribution. There are three main features of the page:

- The bar chart on the top left permits project admin to compare at a glance how many items coders have labeled and the distribution of the labels for each coder. This lets admin see which users are labeling more and detect possible overuse or underuse of labels by particular users.
- The box and whisker chart on the top right lets project admin see how long each coder is taking to annotate the data. This helps admin detect coders who may be having trouble or coders who are simply clicking through their data.



- The Labeled Data table at the bottom of the page contains all data that has been officially assigned a label. The table includes a snippet of the text, the assigned label, and the user responsible for the label. This lets admin see how much data has been collected in total so far, and get a sense for their labeled data without [Part 6: Downloading Labeled Data and/or Model](#).

Labeled Data Table

Show 10 entries

Filter Text:

Text	Label	Coder
@alyssaavd I think imma just eat a hotdog bun and go to bed	hot dog	new_user
@askBrii I'd like to find out	not hotdog	test_user
@badrepbecca So are you gorgeous_□	not hotdog	new_user
@bayley_johnson part 2 Dude my eyeball looks like a hotdog SOMEBODY HELP THIS POOR GIRL BEFORE SHE HURTS HERSE...	hot dog	new_user
@BeyondRGaming @Donnie_Mnemonic @TopsRTs @FatalRTs @FlyRts @Retweet_Twitch @Twitch_RT Happy Birthday Hotdog...	hot dog	new_user
@buzsybee @chaandlerrr lmao i was part of the hot dog gang... sometimes a single hotdog with a regular piece of bread as a bun...	hot dog	user1
@CBCOlympics Awwwww... Who needs love when ya got Rick Tulsie coming all the way from the Saddledome, to give me a heart...	hot dog	user1
@Chedardon ____ always showin love	not hotdog	new_user
@chloebennet sounds like a great time. sometimes I do the same. yesterday went to Costco for gas, a hotdog and browsing for a...	hot dog	new_user
@EricDSnider If anyone tries to put Idris anywhere but on my big screen, they're going to catch my hands. Ali is am_ https://t.co/...	not hotdog	user1

Showing 1 to 10 of 100 entries

1

2

3

4

5

10

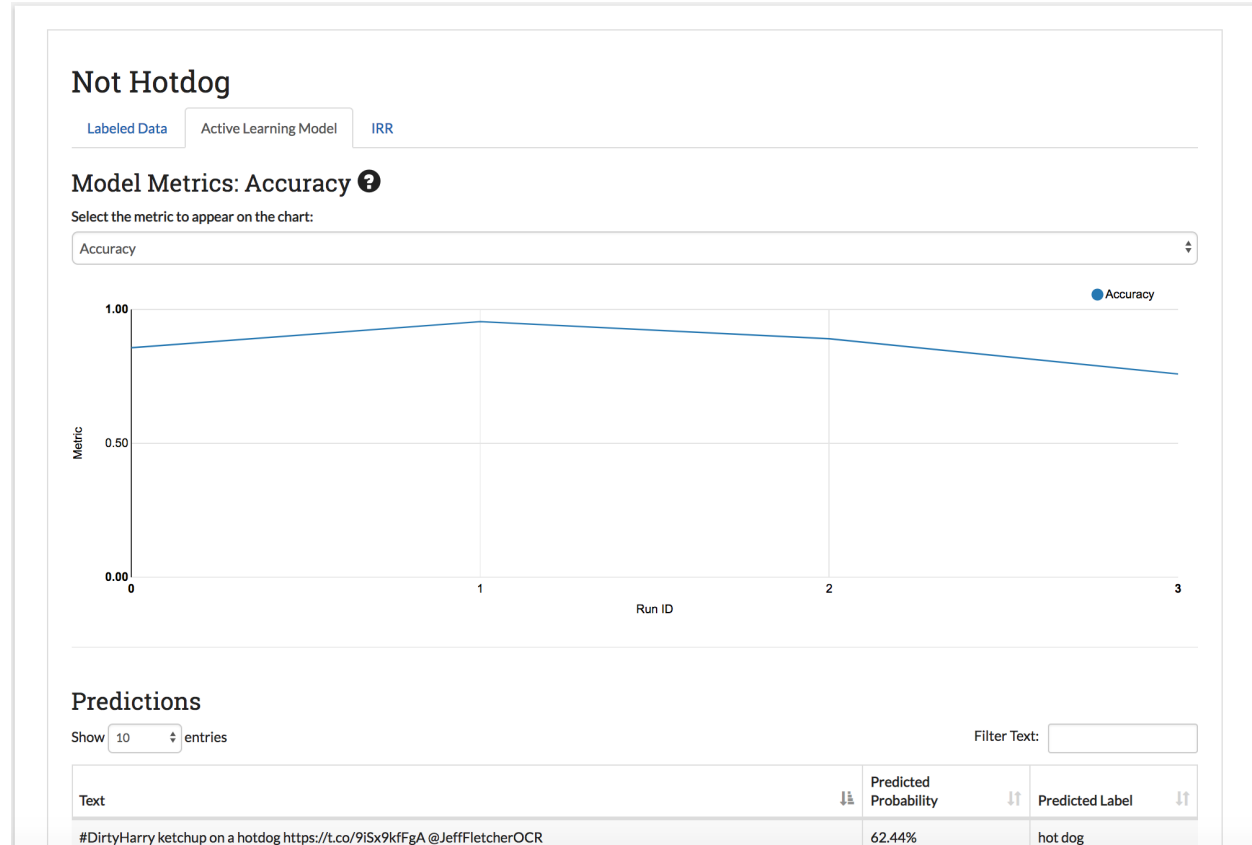
Next

## 2.5.2 Active Learning Model Page

### Visible for projects with a model

The model page lets project admin track how well the classifier trained on their labeled data is performing. After each batch of data is labeled, the model retrains on the entire labeled data set. The model page has two main parts:

- **The model metrics chart:** This chart shows the change in model accuracy, F1 score, Precision, or Recall (see [Active Learning Metrics](#) for more information) after each successive batch is labeled. These scores are calculated by running five-fold [cross validation](#) on the labeled data. You can change which metric is being displayed using the dropdown above the chart. You can also get a formal definition of the displayed metric by hovering over the (?) symbol next to the title.



- **The prediction table:** (only for projects using active learning) Each time a model is run, SMART then predicts the likelihood of the unlabeled data belonging to each class. The Predictions table shows the label with the highest probability for each unlabeled piece of data. If your project uses Uncertainty-based Active Learning (entropy, margin, or least confident), then the data in the table with lower probabilities (the data where the model is the most “uncertain”) is more likely to be chosen for the next batch.

### Predictions

Show 10 entries Filter Text:

Text	Predicted Probability	Predicted Label
#DirtyHarry ketchup on a hotdog <a href="https://t.co/9iSx9kfFgA">@JeffFletcherOCR</a>	62.44%	hot dog
#eatgodaLA When you get to LA. Must visit and tryout the best hotdog in town. The Pinks Hotdogs since 1939. #eatgoda #pink...	57.57%	hot dog
#KatherineWaterston #DanFogler ____ Now Jacob makes his way past his first encounter with Tina while she is trying to have h...	54.27%	hot dog
#MLB #Baseball: LA #LA #Angels Ugly Christmas Sweater Baseball Hat New MLB <a href="https://t.co/nVMRxQKNtd">https://t.co/nVMRxQKNtd</a> <a href="https://t.co/yQvZ...">https://t.co/yQvZ...</a>	62.06%	not hotdog
"Don't choke on a hotdog." - @PlayerEssence 2018	65.09%	hot dog
"Jelena is a hotdog, I guess"	86.72%	hot dog
(strained) Pauling here. I'm at a hotdog eating contest and Engie made me a stomach teleporter, but...oogh...it's not working.	58.14%	hot dog
.@izmir3, tell Five Guys they need to get on the home delivery trend.	57.77%	not hotdog
.@GolicAndWingo I think what goes in/on the bread determines what it is. Burger, sandwich, hotdog, Shawarma, etc	59.07%	hot dog
10.whatcha want us to be?	54.54%	not hotdog

Showing 1 to 10 of 846 entries

1 2 3 4 5 85 Next



## 2.5.3 IRR Page

### Visible for projects that are using IRR

The Inter-Rater Reliability (IRR) page lets admin explore the results of having multiple users label the same data (see *Inter-rater Reliability (IRR)* for a full explanation of IRR). The IRR tab includes four parts:

- **Kappa:** The first value below the IRR Metrics Title is a kappa score. This is a common metric for evaluating IRR. This score is calculated using *Cohen's kappa* if the number of required coders is two, and *Fleiss's kappa* if the number of required coders is higher than two.
- **Percent Overall Agreement:** The next value below the kappa gives the percent of IRR data where all coders agreed (note that everyone skipping it does not count as agreement).
- **Pairwise Percent Agreement Table:** Below the numeric metrics, a table is provided with the percent agreement between each pair of coders. In the case where a particular pair has never coded the same IRR data (since there may be more coders on a project then required for IRR), the message “No samples” is displayed.

The screenshot shows the IRR Metrics page for a project named "Not Hotdog". At the top, there are three tabs: "Labeled Data", "Active Learning Model", and "IRR", with "IRR" being the active tab. Below the tabs, the "IRR Metrics" section displays "Kappa: 0" and "Percent Overall Agreement: 80.0%". There is a "Show 10 entries" dropdown and a "Search:" input field. Below this is a table with three columns: "First Coder", "Second Coder", and "Percent Agreement". The table contains six rows of data. At the bottom of the table, it says "Showing 1 to 6 of 6 entries".

First Coder	Second Coder	Percent Agreement
rchew	user1	No samples
rchew	test_user	No samples
rchew	new_user	75.0%
test_user	new_user	No samples
user1	test_user	100.0%
user1	new_user	No samples

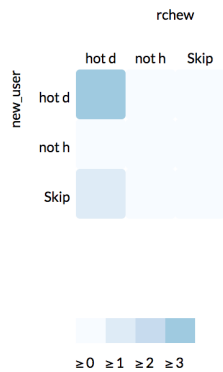
- **Coder Label Heatmap:** The pairwise relationship in the pairwise percent agreement table can be explored in more detail in the Coder Label Heatmap. An admin can examine how often two coders agreed or disagreed on a label and pinpoint areas of disagreement between coders. You can select two coders to compare using the two dropdowns labeled *First Coder* (left) and *Second Coder* (top) above the chart. The legend on the bottom of the chart corresponds to the number of pieces of data involved.

### Coder Label Heatmap

The chart below shows the frequency with which pairs of coders agreed or disagreed on labels

First Coder (top): Second Coder (left):

rchew new\_user



If you select two coders with no samples between them, the heat map will not display:

### Coder Label Heatmap

The chart below shows the frequency with which pairs of coders agreed or disagreed on labels







First Coder (top): Second Coder (left):

rchew user1

## 2.6 Part 6: Downloading Labeled Data and/or Model

So you have been working hard labeling your data and have accumulated a respectable amount. How do you get the data out of the application and onto your computer? SMART provides a download function that works one of three ways depending on the state and settings of your project:

1. If your project has no data labeled, then the download button does nothing and will display “No Labeled Data to Download”.
2. If your project is not using a model or the requirements for a model to run have not yet been met (see [Refilling the Batch](#)), then the download button will display “Download Labeled Data” and output a comma separated value (.csv) file of the labeled data with the columns `ID` (for the unique ID of the data), `Text`, and `Label`. The data is sorted by `Label`.
3. If your project has a model, then the download button will display “Download Model and Labeled Data”. This will output a zip file with:
  - (a) The labeled data file (see number 2)
  - (b) A csv with the labels and their internal ID’s assigned by the application
  - (c) A pickle (.pkl) file with the preprocessed version of your input data as a TFIDF matrix
  - (d) A pickle file with the trained classifier model
  - (e) A pickle file with the trained Vectorizer used to preprocess data into the TFIDF format
  - (f) A README with detailed descriptions of the files and sample code on how to preprocess new data and predict it with your trained model.

 project_4_labeled_data.csv	Today, 6:23 PM	14 KB	Comm...t (.csv)
 project_4_labels.csv	Today, 6:23 PM	38 bytes	Comm...t (.csv)
 project_4_tfidf_matrix.pkl	Yesterday, 8:36 PM	2.8 MB	Micros...cument
 project_4_training_3.pkl	Today, 12:27 PM	2 KB	Micros...cument
 project_4_vectorizer.pkl	Yesterday, 8:36 PM	152 KB	Micros...cument
 README.pdf	Aug 14, 2018, 1:06 PM	66 KB	PDF Document

This button is available in one of two places.

- The Projects Page:

## Projects

Not Hotdog	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	<a href="#">Download Model and Labeled Data</a>
Not Hotdog (plus hamburger)	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	<a href="#">Download Labeled Data</a>
Whizzo Butter or a Dead Crab?	<a href="#">Details</a>	<a href="#">Annotate</a>	<a href="#">Admin Dashboard</a>	No Labeled Data to Download

New Project

- The bottom of the *Details Page*:

Return to Projects
Update Project
Delete Project

Admin Page
No Labeled Data to Download

OR

Return to Projects
Update Project
Delete Project

View CodeBook
Admin Page
Download Labeled Data

OR

Return to Projects
Update Project
Delete Project

View CodeBook
Admin Page
Download Model and Labeled Data

## 2.7 Advanced Feature Details

### 2.7.1 Active Learning

#### What is Active Learning?

The process of creating annotated training data for supervised machine learning models is often expensive and time-consuming. **Active Learning** is a branch of machine learning that seeks to minimize the total amount of data required for labeling by strategically sampling observations that provide new insight into the problem. In particular, *Pool-based Active Learning* algorithms seek to select diverse and informative data for annotation (rather than random observations) from a pool of unlabeled data. Active learning algorithms are a cornerstone of the SMART platform, allowing users to utilize these methodologies with minimal costs to projects.

Due to the lack of a universal “one size fits all” active learning algorithm, SMART provides a number of options, allowing users to select the configuration that works best for their situation. Additionally, the first batch is always chosen randomly, both because there must be sufficient training data for Active Learning to work and to mitigate initial bias.

To learn more about active learning, read Settles (2010)<sup>1</sup> for an excellent survey of the active learning literature.

<sup>1</sup> Settles, B. (2012). Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), 1-114.

## Enabling Active Learning in SMART

Users can enable active learning and select the method for measuring uncertainty in the Advanced Settings page when creating a new project (see [Advanced Settings](#) for more details).

As of this release, SMART supports **Uncertainty Sampling** with three different measures of uncertainty: *Least Confident*, *Margin*, and *Entropy*. Uncertainty Sampling works by training the model on the existing labeled data and then calculating the probability that each piece of unlabeled data belongs to each possible label. The algorithm returns the most “uncertain” data to be correctly labeled by the coders. The algorithm uses one of three methods to select the unlabeled data that the classifier is the most “uncertain” about:

- **Least Confident** (default): The algorithm chooses the data with the lowest probability for the most likely label using the equation:

$$*arg max x_{LC}^* =_x 1 - P_{\theta}(\hat{y}|x)$$

where:

$$\hat{y} =_y P_{\theta}(y|x)$$

- **Margin Sampling**: The algorithm chooses the data with the smallest difference between the probability of the most likely and least likely labels using the equation:

$$x_M^* =_x [P_{\theta}(\hat{y}_2|x) - P_{\theta}(\hat{y}_1|x)]$$

where:

yhat\_1 and yhat\_2 are the first and second most likely predictions under the model.

- **Entropy**: The algorithm chooses the most uncertain or “disordered” data by taking the data with the highest score for the entropy equation:

$$x_H^* =_x - \sum_y P_{\theta}(\hat{y}|x) * \log P_{\theta}(\hat{y}|x)$$

where:

y ranges over all possible labelings of x.

## Active Learning Metrics

An important consideration in active learning is model performance. To assess your models as your team labels data, SMART provides the following classification model evaluation metrics in the Active Learning Model page of the Admin dashboard:

- **Accuracy**: proportion of observations that the model correctly labeled.

$$\frac{TP + TN}{TP + FN + TN + FP}$$

- **Precision**: Indicates how precise the model is at correctly predicting a particular category.

$$\frac{TP}{TP + FN}$$

- **Recall:** Indicates how comprehensive the model is at identify observations of a particular category.

$$\frac{TP}{TP + FP}$$

- **F1-score:** the harmonic mean of Precision and Recall.

$$2 \cdot \frac{(Precision \cdot Recall)}{Precision + Recall}$$

where:

TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

## 2.7.2 Inter-rater Reliability (IRR)

### What is IRR?

SMART is designed to support labeling projects that may utilize many labelers. When many coders are working on a project, it becomes crucial that coders agree on what labels should apply to what data. Inter-rater Reliability (IRR) is a set of metrics that measures how consistently coders agree with each other, and it is common for a labeling project to require a minimum score for a particular IRR metric for the data to be deemed usable. IRR metrics are calculated from having coders label the same data and examining the results.

### Enabling IRR in SMART

Project creators can enable IRR in their SMART projects through the [Advanced Settings](#) page of project creation. Once IRR is enabled, two additional settings are available:

1. *The percentage of a batch that will be IRR* – This number signifies how much of the data per batch will be used to calculate IRR metrics. This data must be either labeled or skipped by a minimum number of coders before it can be processed.
2. *The minimum number of coders participating in IRR activities* – This number signifies the minimum number of coders that would need to either skip or annotate a piece of IRR data before it can be processed.

As an example, if a project creator chooses 100% for the percentage of the batch that will be IRR and 3 for the minimum number of coders participating in IRR activities, all data in each batch would be required to be labeled by three coders before it could be processed.

---

**Tip:** Setting the percentage to 0% is the same as disabling IRR.

---

### IRR Data Flow

If the project creator has enabled IRR, additional steps are added to the data pipeline. First, when the project provides a batch of unlabeled data to label, the previously specified IRR percentage is taken out and marked as IRR. When a user opens the annotation page to begin labeling, SMART first checks if there is any IRR data that SMART has not yet seen. This data is pulled first, and the rest of the deck is filled with non-IRR data. This deck is then shuffled before being presented to the user to make it harder to know what data is IRR. SMART tracks what IRR data has been labeled/skipped by which users. Skips are automatically recorded in the IRR Log table, while labels are placed in the same label table as non-IRR data (though the training set will not incorporate them as they are marked IRR). Once IRR data has sufficient skips or labels, two outcomes can happen:

1. If everyone labeled the datum and these labels were the same, then the datum is added with the agreed upon label to the training set.

2. If any coder skipped the datum, or coders disagreed upon the label, the datum is sent to the admin for the final label.

After a datum is processed, the labels from all coders are recorded in the IRR Log table.

---

**Note:**

- If an admin chooses to discard an IRR datum as unusable, all records of this datum will be flushed from the IRR Log table.
- 

## IRR Metrics

To evaluate the reliability of coders, several metrics are calculated for the project admins. This includes percent overall agreement (how many often did everyone give the same label), pairwise percent agreement (how much did two users in particular agree), and a heat map showing the frequency where one coder chose label A and another chose label B (see [IRR Page](#) for more information). In addition, SMART provides a kappa score, which is a common IRR metric. The kappa score comes from one of the two types below:

### Cohen's kappa

This metric is used when there are two coders. Cohen's kappa is most commonly used for categorical items<sup>2</sup>. The general formula is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where

$$p_o = \text{accuracy}$$

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

and where N is the number of data points, k represents the number of possible labels, and n is a matrix of label counts of category by labeler (or how many times did each coder choose each label)<sup>2</sup>.

$p_e$  is the hypothetical probability of agreeing by chance.

### Fleiss's kappa

This metric is the counterpart to Cohen's kappa for more than three coders. The formula is the ratio between the degree of agreement that is attainable above chance, and the degree of agreement actually achieved<sup>3</sup>. The general formula is:

$$\kappa = \frac{\hat{P} - \hat{P}_e}{1 - \hat{P}_e}$$

$$\hat{P} = \frac{1}{N} \sum_{i=1}^N \left( \frac{1}{l(l-1)} * \sum_{j=1}^k n_{ij}(n_{ij} - 1) \right)$$

$$\hat{P}_e = \sum_{j=1}^k \left( \frac{1}{l(N)} * \sum_{i=1}^N n_{ij} \right)$$

Where N is the number of data points, k represents the number of possible labels, l is the number of labels for each piece of data, and n is a matrix of data points by the number of votes per label<sup>3</sup>.

<sup>2</sup> [https://en.wikipedia.org/wiki/Cohen%27s\\_kappa](https://en.wikipedia.org/wiki/Cohen%27s_kappa)

<sup>3</sup> [https://en.wikipedia.org/wiki/Fleiss%27\\_kappa](https://en.wikipedia.org/wiki/Fleiss%27_kappa)

## 2.7.3 Fix Skewed Label Distributions

In many applied settings, the distribution of categories the user may be interested in labelling is not well balanced. In particular, if one or more categories of interest occur rarely, labeling observations at random will be particularly inefficient and can quickly exhaust a project's labelling budget. To help combat this issue, SMART implements a version of the *guided learning* strategy outlined in Attenberg and Provost (2010)<sup>4</sup>. This approach treats active learning as a search problem, allowing the user to utilize prior context to identify relevant observations of the rare category, effectively initializing the training batch with a set of relevant rare examples. Findings in Attenberg and Provost (2010)<sup>4</sup> indicate an 8x reduction of real annotation cost per instance using this method on imbalanced data sets when compared to other active learning strategies studied.

See [Fix Skew Page](#) for more information on using this feature.

## 2.8 Frequently Asked Questions (FAQs)

### 2.8.1 General Questions

#### What does SMART stand for?

SMART stands for “**S**marter **M**anual Annotation for **R**esource-constrained collection of **T**raining data”, an acronym so contrived that it qualifies as both a [backronym](#) and a [recurvise acronym](#).

#### What was the genesis for SMART?

The creation of SMART stems directly from a data-focused business problem that the core development team encounter repeatedly at [RTI International](#); projects that could greatly benefit from having a well trained supervised ML classifier often don't yet have the labelled data to build one. From our experience as practicing data scientists and researchers, the data annotation effort can be particularly arduous in the social science and public health domains where the underlying categories are difficult for humans to categorize and/or are ambiguous. Several of our projects utilizing social media, online news articles, legislative texts, etc. require careful coordination with project staff or crowdsourced workers to get labelled data that is timely and reliable. SMART was a means to help make this process more efficient, especially in cases where using a crowdsourcing platform isn't possible (e.g., when the data is proprietary or sensitive in nature).

More generally, building labeled data sets is also a commonly reported bottleneck for groups doing applied data science and machine learning in both research and industry. With generous support from the [National Consortium for Data Science](#) and [RTI International](#), we were able to build SMART as an open source project to share with the larger data science and applied ML community.

#### Can I add or remove new class labels after a project is created?

SMART does not allow users to add, remove, or modify class labels after a project is created. This is mainly to prevent awkward interactions with the active learning models. For example, the model evaluation metrics and their associated visualizations become difficult to compare if class definitions are frequently changing.

That being said, we recognize that determining meaningful categories for a new labeling project can be non-trivial. To help users iterate during early-stage labeling projects when model categories are still being decided, we support exporting labeled data from any existing projects and recommend creating a new project (with modified codebook, label descriptions, etc.) to provide a clean annotation experience for your coding team.

<sup>4</sup> Attenberg, J., & Provost, F. (2010). Why label when you can search?: Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 423-432). ACM.

## I accidentally mislabeled an observation. How do I correct my mistake?

If you accidentally mislabel a document during the coding process, you can re-label the observation in the “History” tab of the annotation page. Re-labelling is only unavailable if:

- An Admin has provided a final label on a skipped document,
- An Admin has provided a final label on an IRR document with coder disagreement.

**Warning:** When using active learning, data labels modified on the History tab will not change the model accuracy metrics of past batches displayed on the Active Learning tab of the Admin page, but instead, will update the data for the next model re-training.

## What functionality do I get as a Coder? Admin?

SMART has two levels of user: `coder` and `admin`. The following chart summarizes what operations are accessible to what level of user:

User Type		Admin	Coder
Annotation Page	Label Data	✓	✓
	Skip Data	✓	✓
	Discard Data	✓	
	Approve Data	✓	
	History Table	✓	✓
	Fix Skew	✓	
	View Codebook	✓	✓
Admin Page	Coder Metrics	✓	
	Active Learning Metrics	✓	
	IRR Metrics	✓	
Details Page	View Project Details	✓	✓
	Update Projects	✓	
Download	Download Labeled Data	✓	
	Download Trained Model	✓	

## 2.8.2 Technical Questions

### What kinds of supervised ML tasks does SMART support?

As of this version, SMART only supports text classification tasks. However, we hope to extend the platform to support other types of media (images, video, etc.), and perhaps other types of modeling tasks beyond classification in future



releases.

### What features underly the active learning models?

Currently, SMART uses a term frequency inverse document frequency (TF-IDF) matrix to structure text data. Other popular representations (embeddings, pre-trained language models, etc.) are not currently available in SMART but are welcome additions in future releases.

### Can I code for multiple text classification tasks at the same time?

The only way currently to annotate for multiple modeling tasks is to create multiple projects (one for each task). Though multi-task active learning (learning how to select observations that best jointly learn multiple modeling tasks simultaneously) is an exciting area of research, there are not plans for supporting it in the near future.

### Do I have to use Active Learning, IRR, etc.?

Depending on your modeling goals, many of the options provided in SMART (active learning, IRR, etc.) may be unnecessary or overkill for your use case. To allow users to customize their data labelling experience, users are encouraged to add or remove project features in the Advanced Settings page during project creation.

### What are the metrics on the Active Learning page?

The model evaluation metrics presented on the Active Learning section can help you and your team diagnose how a model is performing as more data is labelled. Definitions for the classification evaluation metrics can be found in the *Active Learning* section of the *Advanced Feature Details* page.

### What active learning strategies does SMART support?

The active learning strategies implemented in SMART can be found in the *Active Learning* section of the *Advanced Feature Details* page.

### Why support labeling data in batches?

We implemented an option to label data in batches due to its practicality. While many active learning strategies assume a sequential back-and-forth between the model and the labeller, waiting for the model to train and predict new examples after every new labeled observation can be prohibitively slow when models are complex or when the underlying data set is large. Additionally, labeling observations in batches more easily allows the labeling process to be spread out among multiple people working on a batch in parallel.

To provide assistance for just this scenario, researchers have developed *batch-mode active learning* algorithms that help assemble batches containing both informative and diverse examples, reducing the chance that observations within a batch will provide redundant information. While effective on large batch sizes, initial tests comparing batch-mode active learning models against simpler non-batch active learning strategies showed similar performance on more modest batch sizes [[link to notebook](#)]. Due to the complexity of many batch-mode active learning models and similar performance on smaller batch sizes, we forego including batch-mode active learning models in the initial release.

### What's the tech stack used to build SMART?

It consists of a Django web application backed by a Postgres database and a Redis store. The application uses Docker in development to aid in dependency management.

## 2.9 Release Notes and Change Log

### 2.9.1 Release v.0.0.1

### 2.9.2 Contributors

- Peter Baumgartner
- Rob Chew
- Emily Hadley
- Caroline Kery
- Lucy Liu
- Joey Morris
- Jason Nance
- Keith Richards
- Michael Wenger

## 2.10 License

The MIT License (MIT)

Copyright <YEAR> <COPYRIGHT HOLDER>

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.