

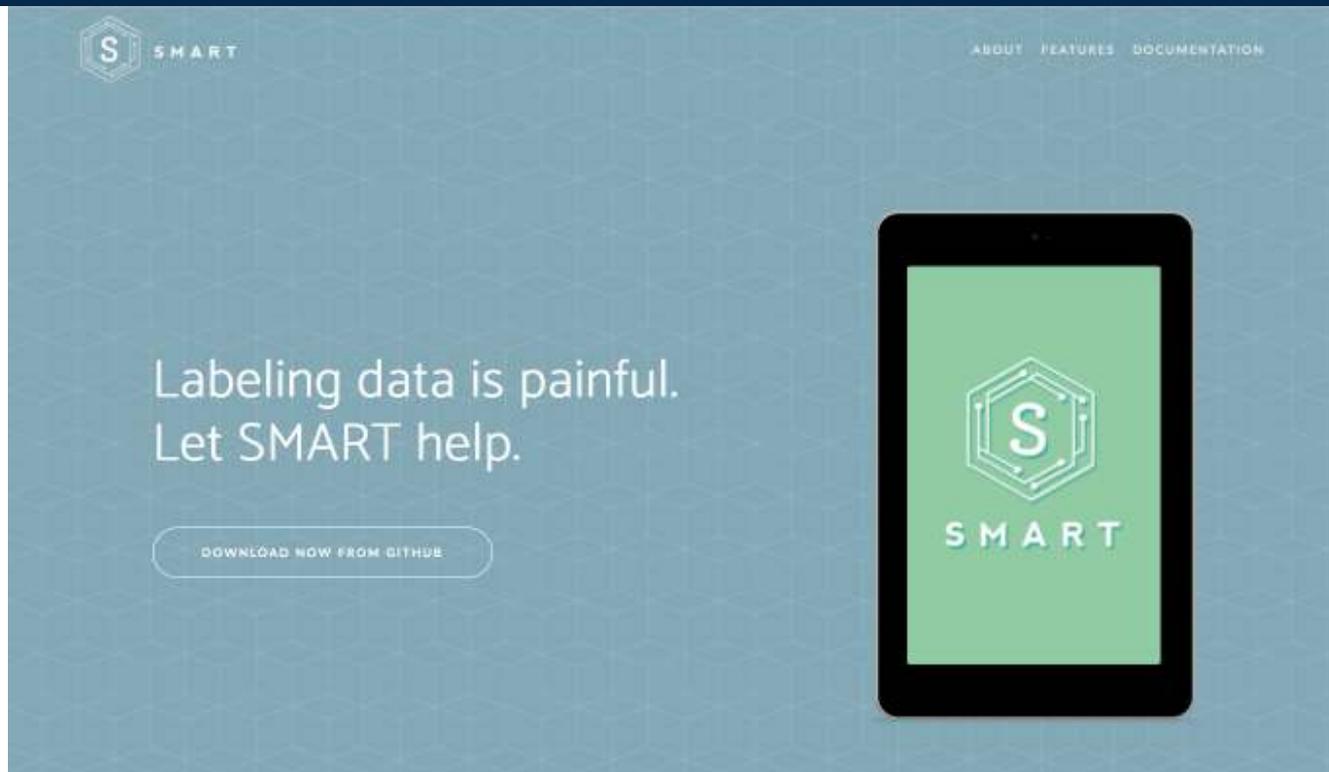
SMART: An Open Source Tool to Facilitate Auto-Coding



Caroline Kery

Government Advances in Statistical Programming (GASP!) Workshop

Sep 23rd, 2019



Motivation – Unstructured Text data

Institutional Support for Cloud Services Survey

Institutional information and/or digital literacy policy

6. Does your institution have a formal information/digital literacy policy?

- Yes
- No
- Not sure

6a. If yes, does it cover use of Cloud services?

- Yes
- No
- Not sure

Further information

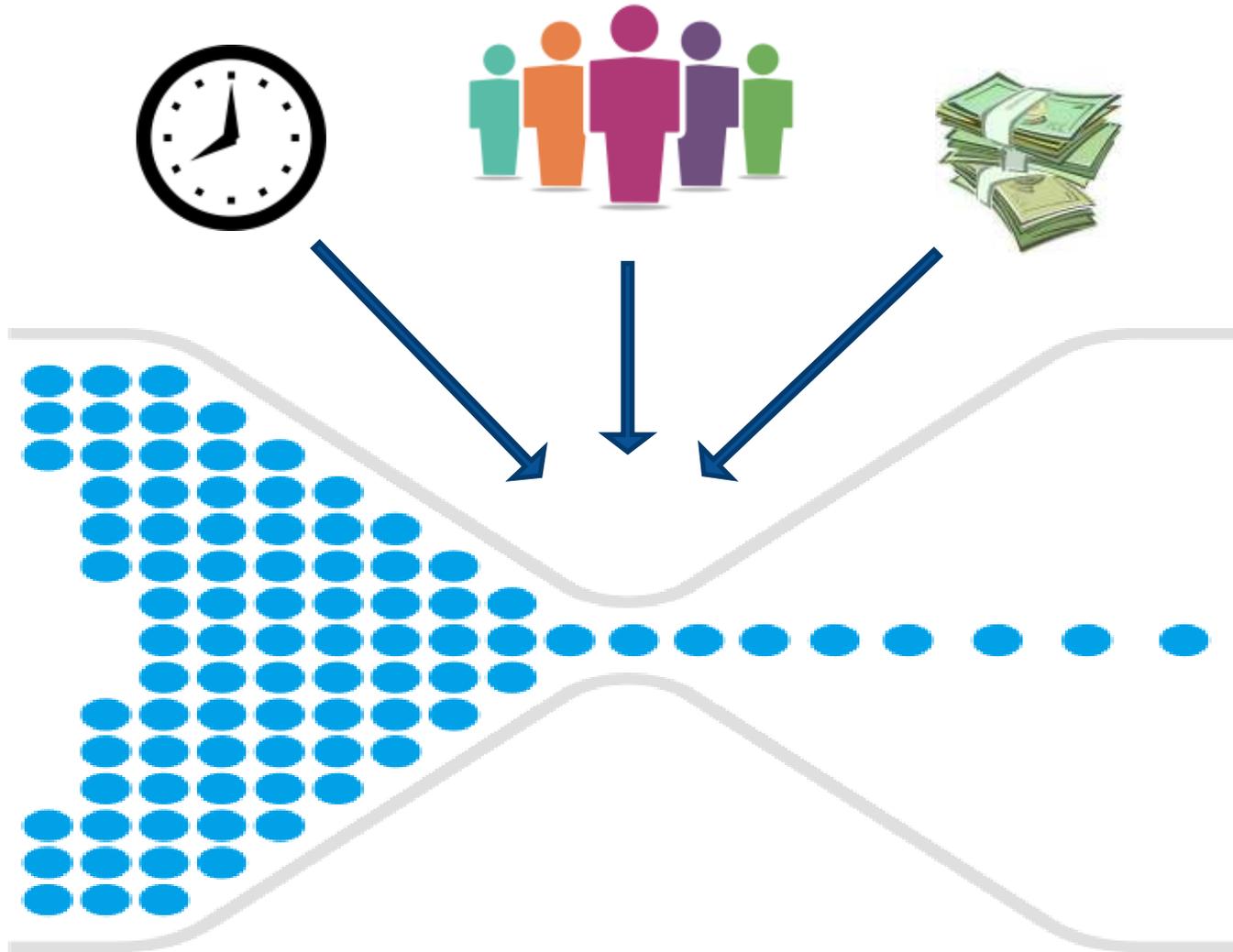
6b. If yes, does it address the needs of staff and researchers who wish to continue using IT services when they leave the institution?



- Text data often needs to be organized (labeled or coded) for further analysis to be possible

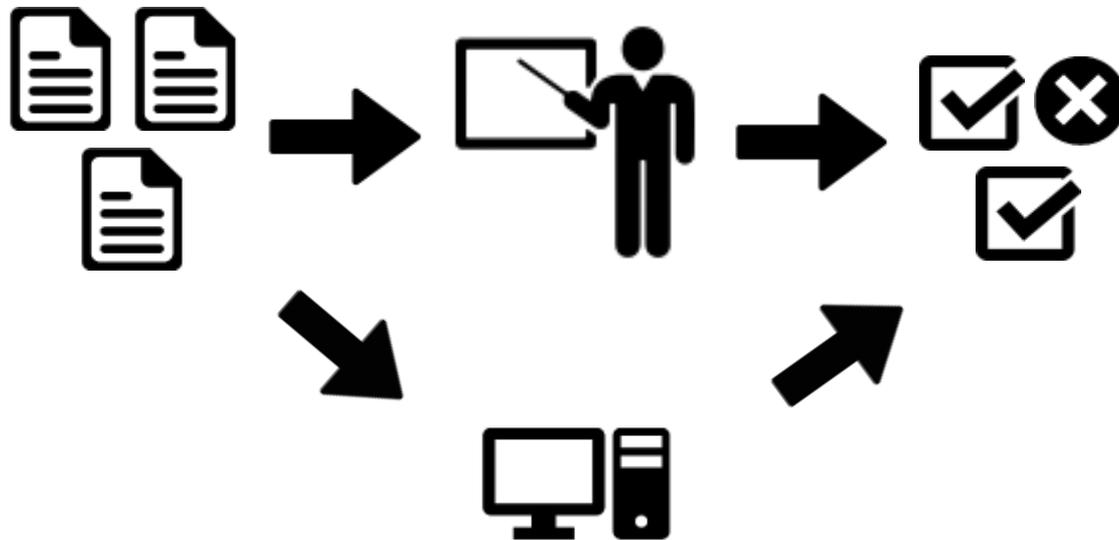


Bottlenecks of Data Labeling



Coding and Machine learning

- Possible solution, train a machine to label things for you!



Example: Auto-coders

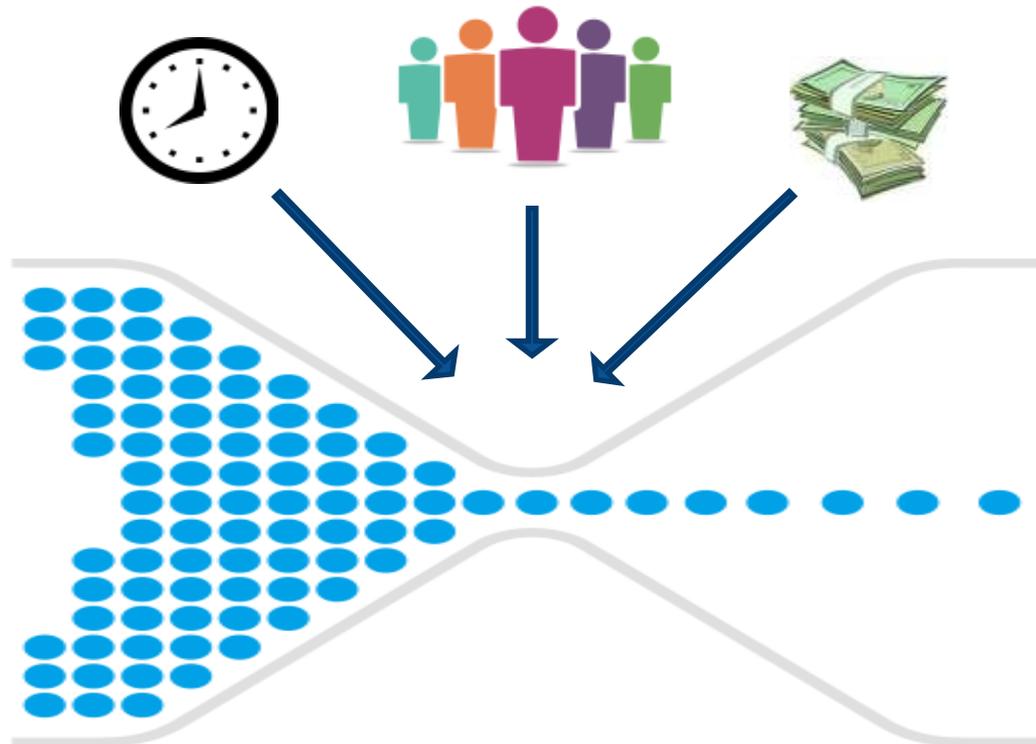


NIOSH Industry &
Occupation Computerized
Coding System (NIOCCS)



But still...

- Many machine learning labelers use **supervised learning** which leverages existing labeled data to learn how to label new data.
- In practice, this means that to create successful auto-coders, we still need large amounts of manually labeled data.



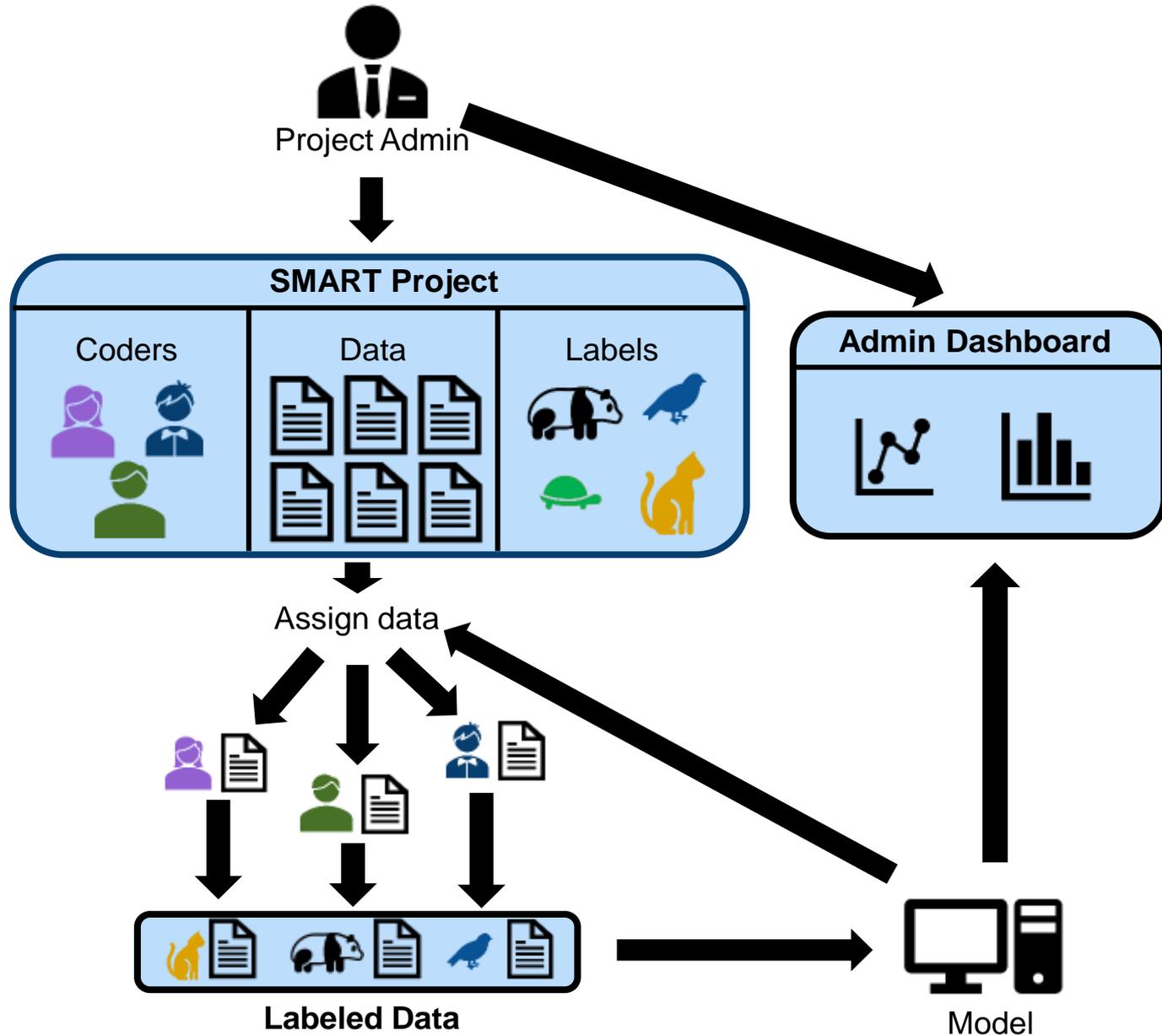


Labeling data is painful.
Let SMART help.

[DOWNLOAD NOW FROM GITHUB](#)



SMART Overview

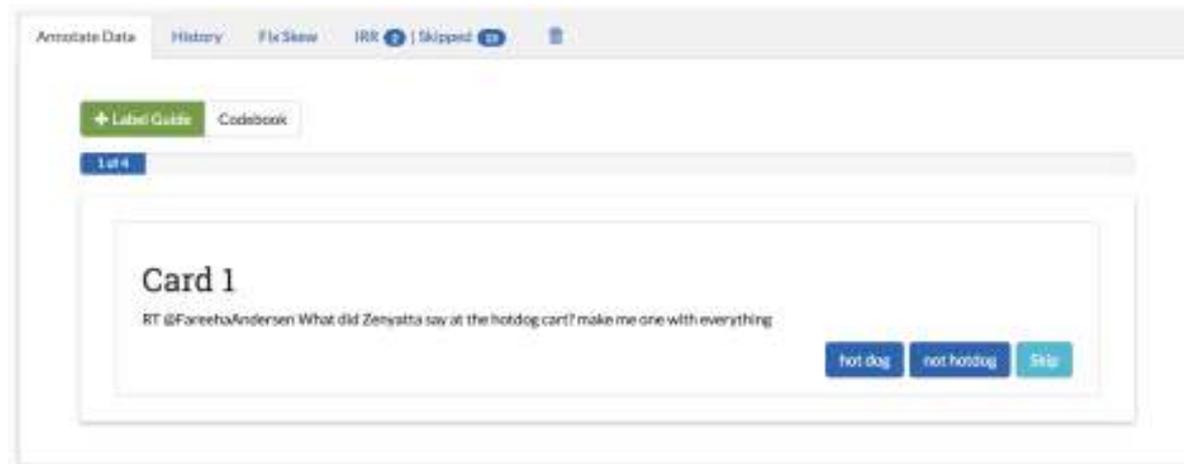


SMART – Organizing labeling tasks



Multi-user Coding

Allow parallel annotation efforts within a project.



SMART – Inter-rater reliability



Inter-rater Reliability

Get your team on the same page and ensure quality labels.

Not Hotdog

Labeled Data Active Learning Model **IRR**

IRR Metrics

Kappa: 0
Percent Overall Agreement: 80.0%

Show entries Search:

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

Showing 1 to 6 of 6 entries 1

SMART – Inter-rater reliability cont.



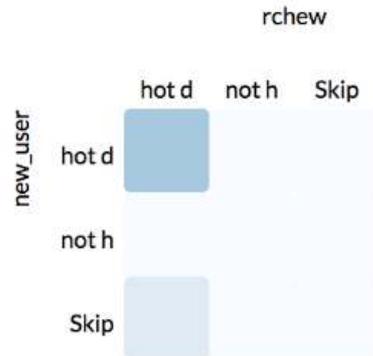
Inter-rater Reliability

Get your team on the same page and ensure quality labels.

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):



SMART – Admin Page: Monitor labeling



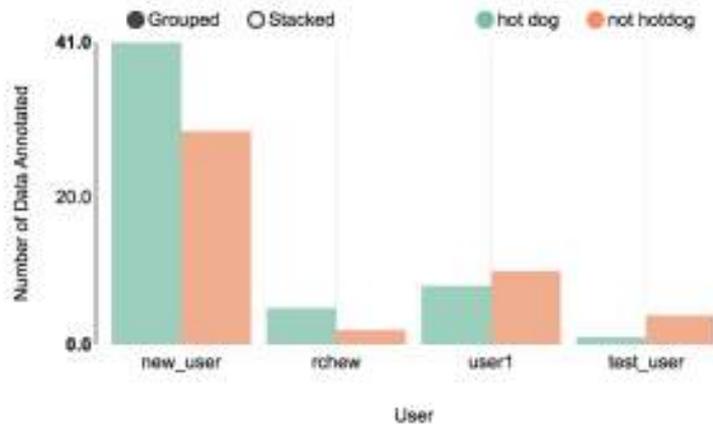
Admin Dashboard

Manage the labeling process and monitor coder progress.

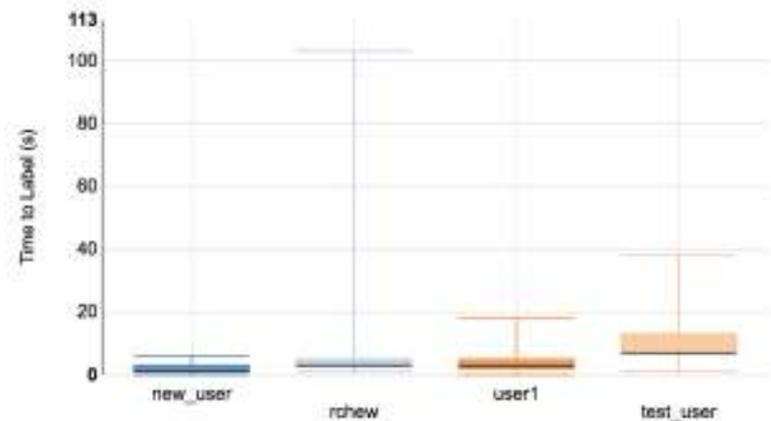
Not Hotdog

Labeled Data Active Learning Model IRR

Label Distribution



Time To Label



SMART – Monitor model progress



Admin Dashboard

Manage the labeling process and monitor coder progress.

Not Hotdog

Labeled Data

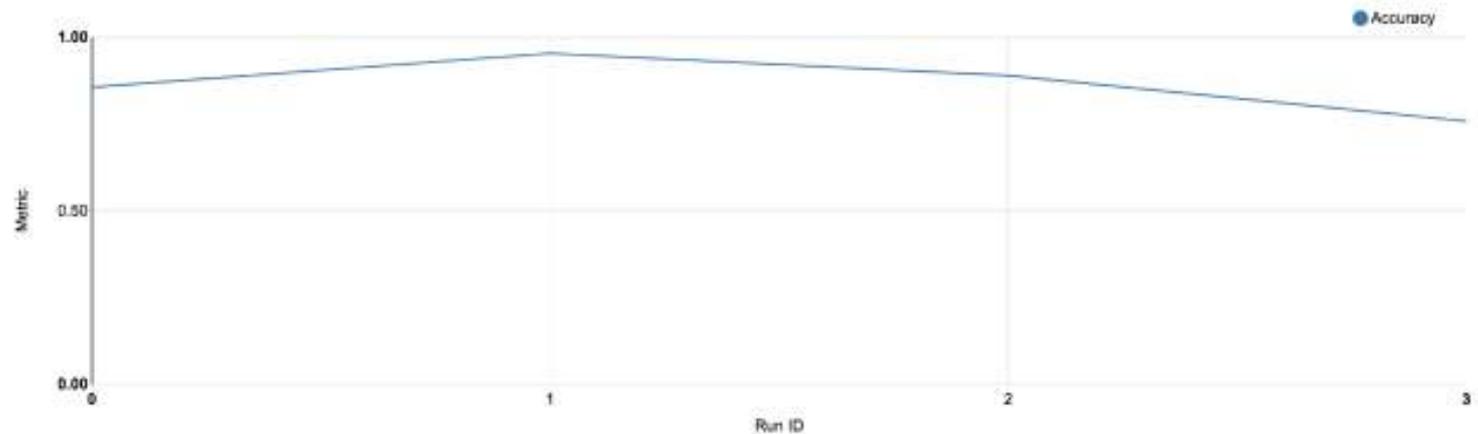
Active Learning Model

xxx

Model Metrics: Accuracy ?

Select the metric to appear on the chart:

Accuracy



SMART - Properties



Open Source

Made available under the permissive MIT License.



On-Premise Install

Keep sensitive data secure within your organization's firewall.

SMART Features – Bringing it together



Active Learning

Label observations more likely to improve model performance.



Inter-rater Reliability

Get your team on the same page and ensure quality labels.



Admin Dashboard

Manage the labeling process and monitor coder progress.



Multi-user Coding

Allow parallel annotation efforts within a project.



On-Premise Install

Keep sensitive data secure within your organization's firewall.



Open Source

Made available under the permissive MIT License.

<https://rtiinternational.github.io/SMART/>



SMART

latest

TUTORIAL

Part 1: Installation

Part 2: Creating a New Project

Part 3: Reviewing Projects & Editing Project Settings

Part 4: Annotating Data

Part 5: Administrator Dashboard

Part 6: Downloading Labeled Data and/or Model

ADVANCED FEATURES

Advanced Feature Details

ABOUT SMART

Frequently Asked Questions (FAQs)

Release Notes and Change Log

License

Read the Docs

v: latest ▾

Docs » SMART User Docs

Edit on GitHub

SMART User Docs

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

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.

Quick Start

```
$ git clone https://github.com/RTIIInternational/SMART.git
$ cd smart/envs/dev/
$ docker-compose build
$ docker volume create --name=vol_smart_pgdata
$ docker volume create --name=vol_smart_data
$ docker-compose run --rm smart_backend ./migrate.sh
$ docker-compose up -d
```

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

Tutorial

- [Part 1: Installation](#)
- [Part 2: Creating a New Project](#)
- [Part 3: Reviewing Projects & Editing Project Settings](#)



[Home Page](#)

[Papers](#)

[Submissions](#)

[News](#)

[Editorial Board](#)

[Announcements](#)

[Proceedings](#)

[Open Source
Software](#)

[Search](#)

SMART: An Open Source Data Labeling Platform for Supervised Learning

Rob Chew, Michael Wenger, Caroline Kery, Jason Nance, Keith Richards, Emily Hadley, Peter Baumgartner, 20(82):1–5, 2019.

Abstract

SMART is an open source web application designed to help data scientists and research teams efficiently build labeled training data sets for supervised machine learning tasks. SMART provides users with an intuitive interface for creating labeled data sets, supports active learning to help reduce the required amount of labeled data, and incorporates inter-rater reliability statistics to provide insight into label quality. SMART is designed to be platform agnostic and easily deployable to meet the needs of as many different research teams as possible. The project website <https://rtiinternational.github.io/SMART/> contains links to the code repository and extensive user documentation.

[\[abs\]](#)[\[pdf\]](#)[\[bib\]](#) [\[code\]](#)

© JMLR 2019. (edit, beta)

Questions?

delivering **the promise of science**
for global good



Caroline Kery
Data Scientist
RTI International

