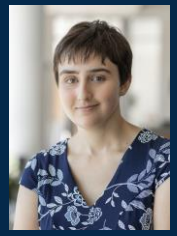
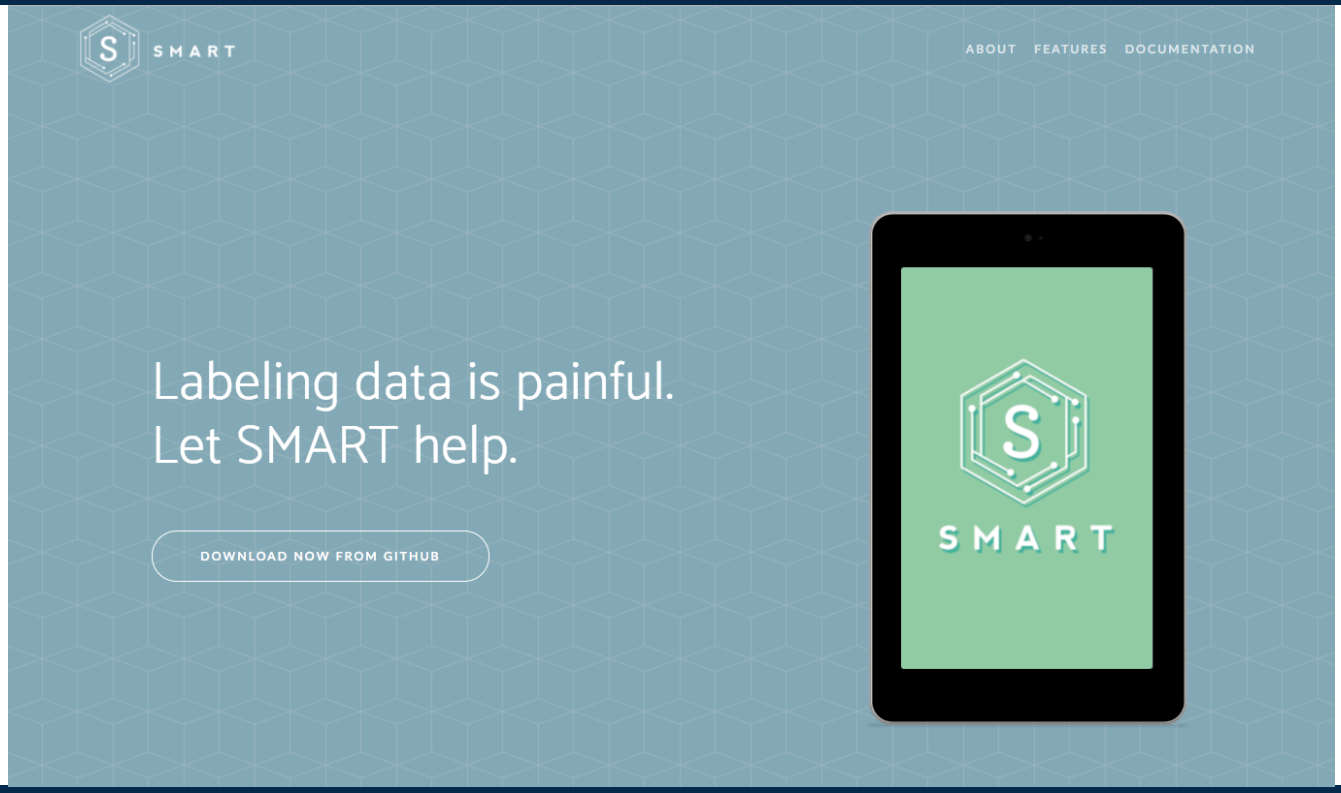
SMART: An Open Source Tool to Facilitate Auto-Coding



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

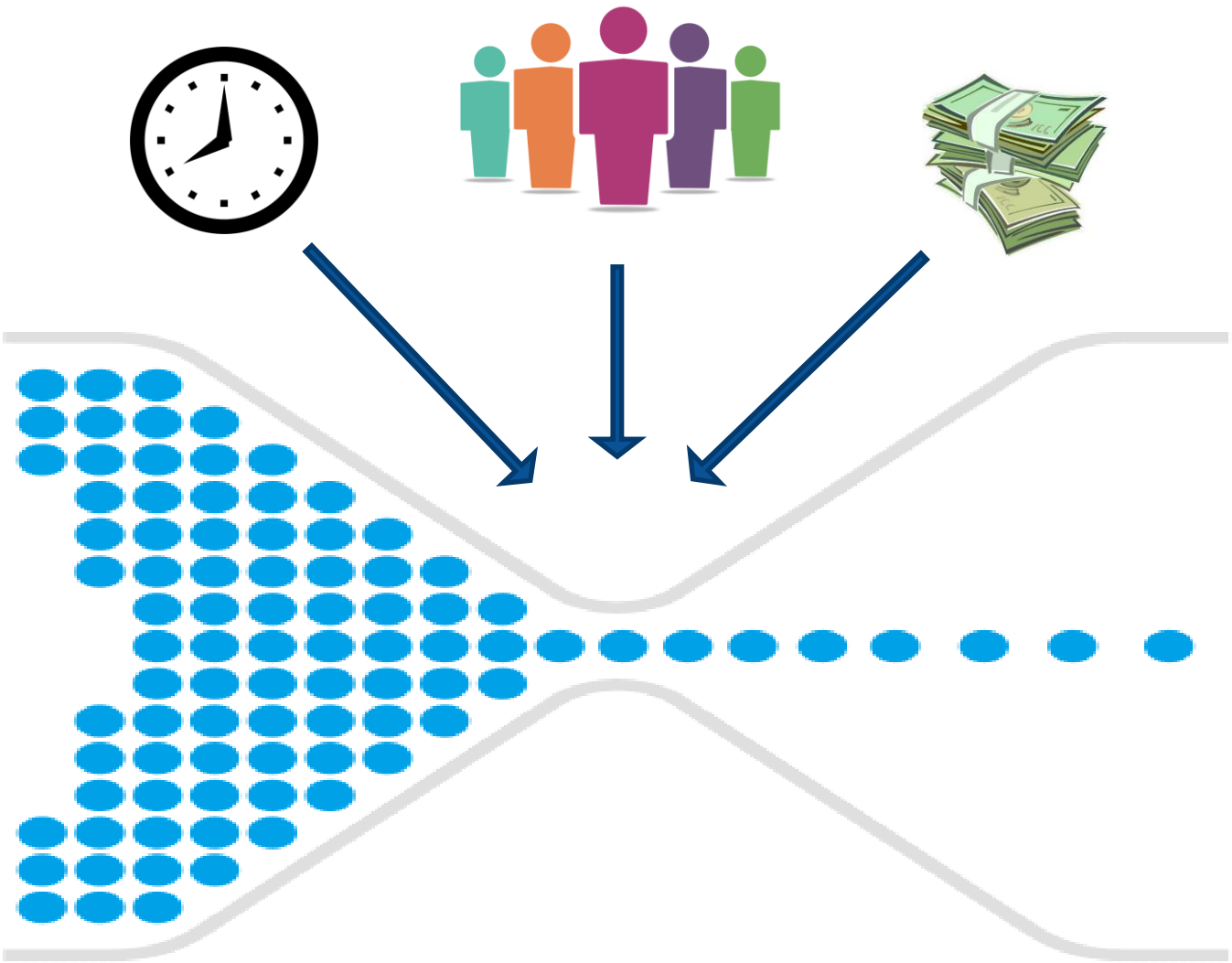
Further information

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

Coding

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

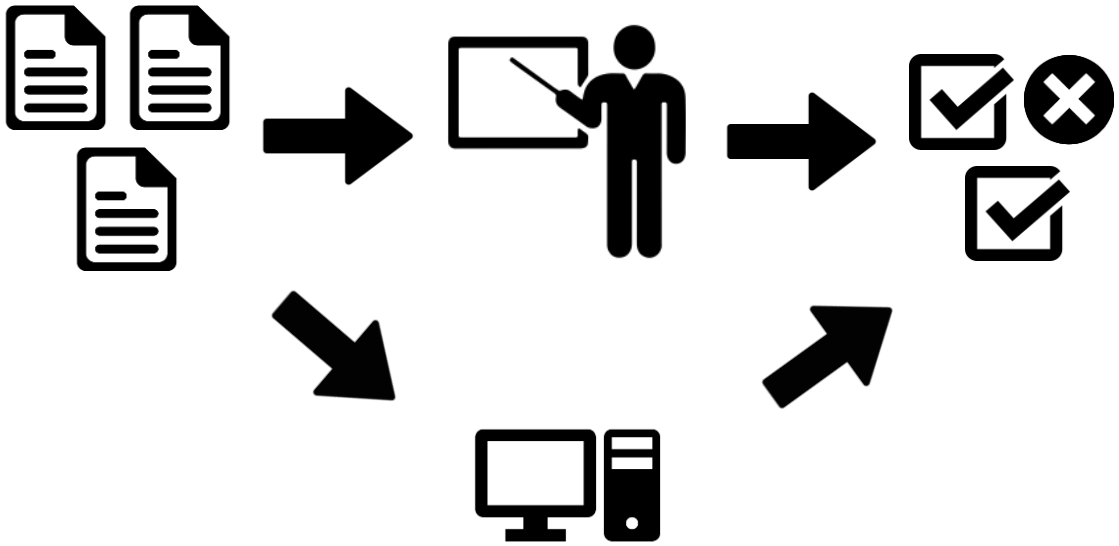
Bottlenecks of Data Labeling



Bottleneck by Electric Cloud

Coding and Machine learning

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



Example: Auto-coders

ELSEVIER

Accident Analysis & Prevention
Volume 88, March 2016, Pages 117-123

Comparison of methods for auto-coding causation of injury narratives

S.J. Bertke [a] ✉, A.R. Meyers [b], S.J. Wurzelbacher [b], A. Measure [c], M.P. Lampl [d], D. Robins [d]

NIOSH Industry & Occupation Computerized Coding System (NIOCCS)

ELSEVIER

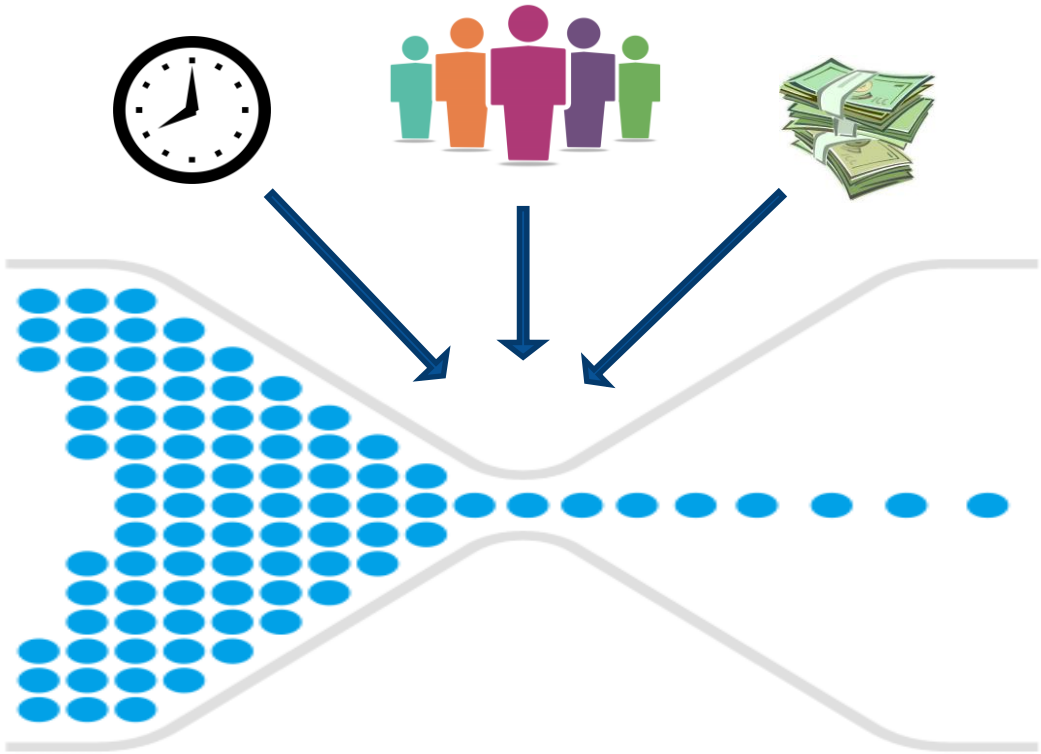
Journal of Safety Research
Volume 57, June 2016, Pages 71-82

Bayesian decision support for coding occupational injury data

Gaurav Nanda [a] ✉, Kathleen M. Grattan [b] ✉, MyDzung T. Chu [b], Letitia K. Davis [b], Mark R. Lehto [a, 1] ✉

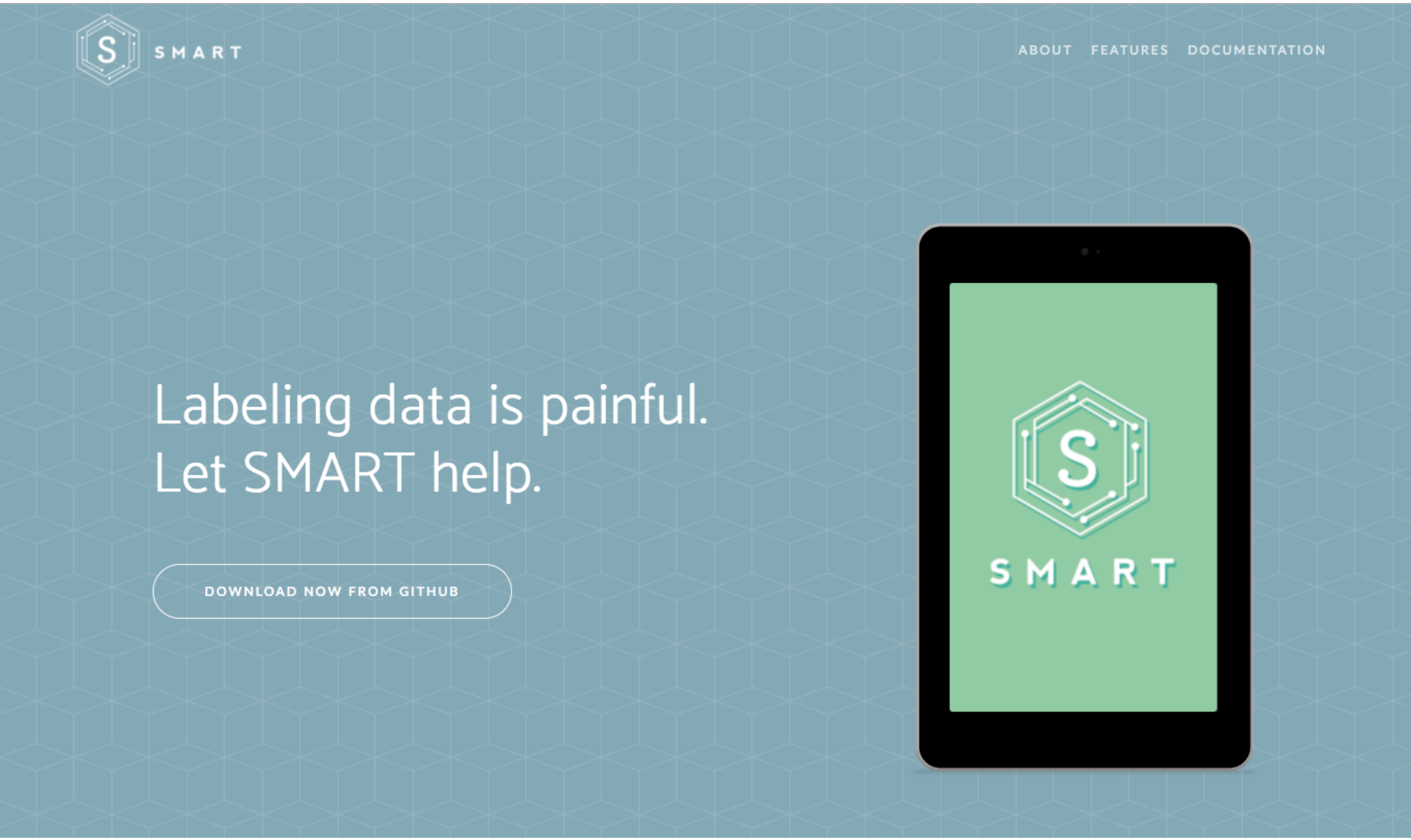
But still…

- 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.*

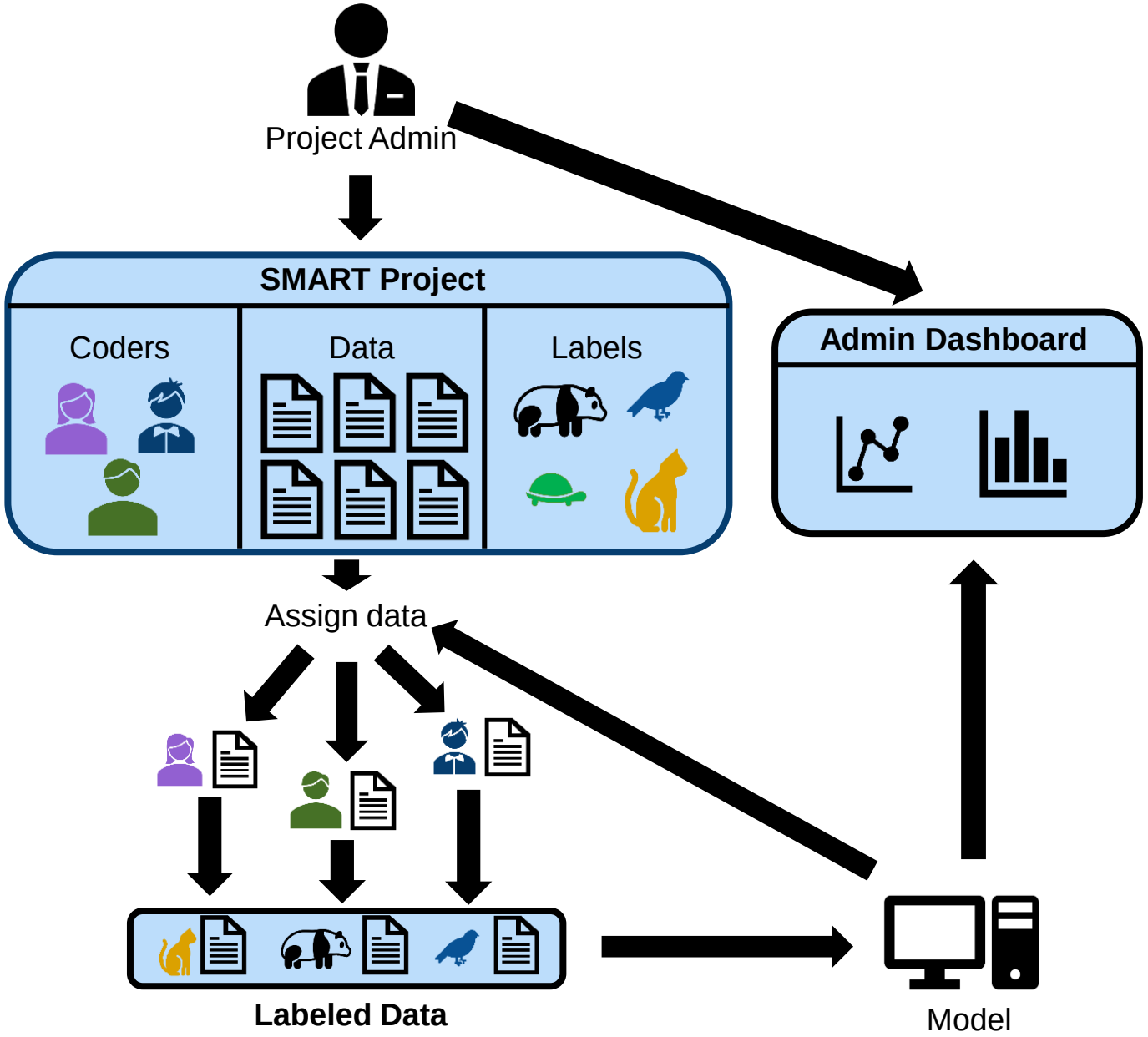


Bottleneck by Electric Cloud

SMART


SMART
ABOUT FEATURES DOCUMENTATION
Labeling data is painful.
Let SMART help.
DOWNLOAD NOW FROM GITHUB
SMART

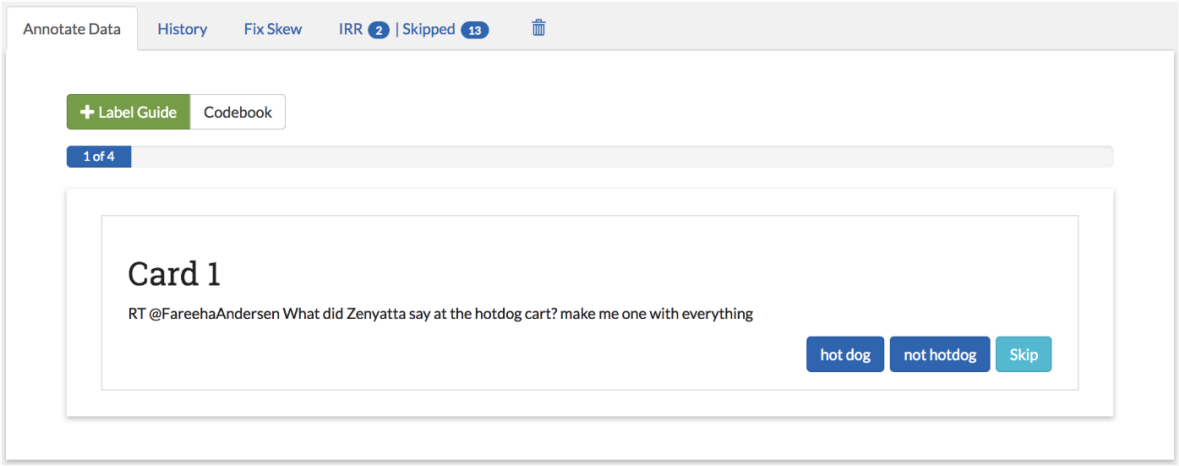
SMART Overview



SMART – Organizing labeling tasks

Multi-user Coding

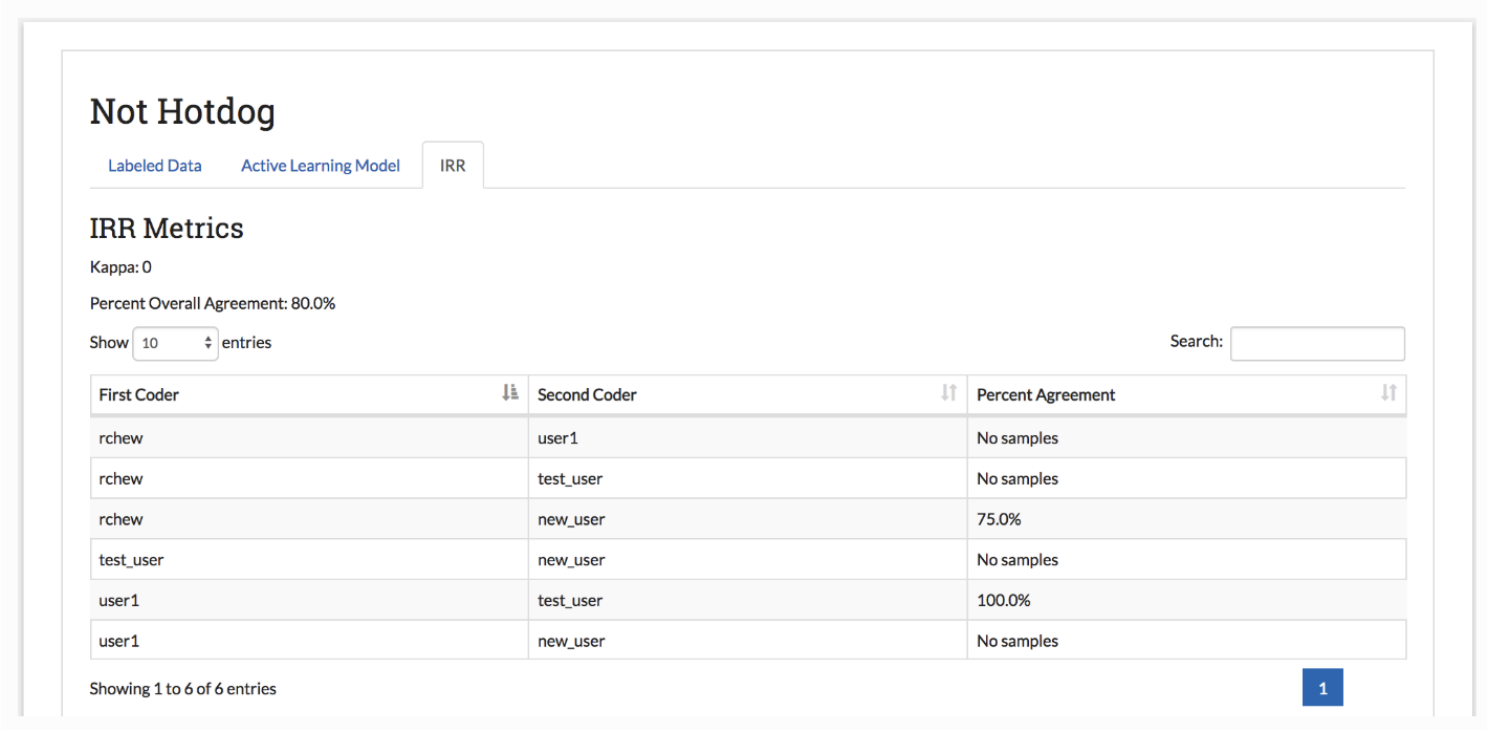
Allow parallel annotation efforts within a project.



SMART – Inter-rater reliability

Inter-rater Reliability

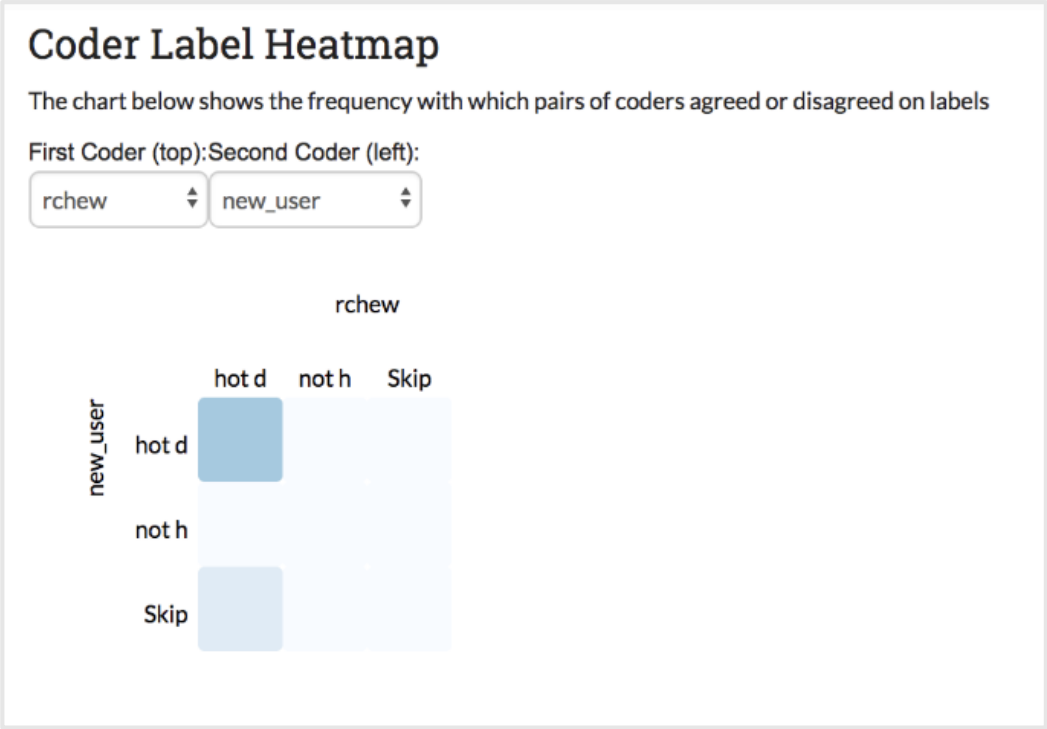
Get your team on the same page and ensure quality labels.



SMART – Inter-rater reliability cont.

Inter-rater Reliability

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



SMART – Admin Page: Monitor labeling

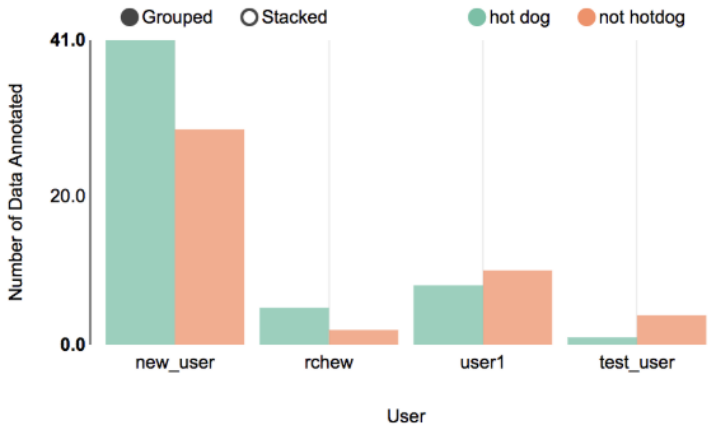
Admin Dashboard

Manage the labeling process and monitor coder progress.

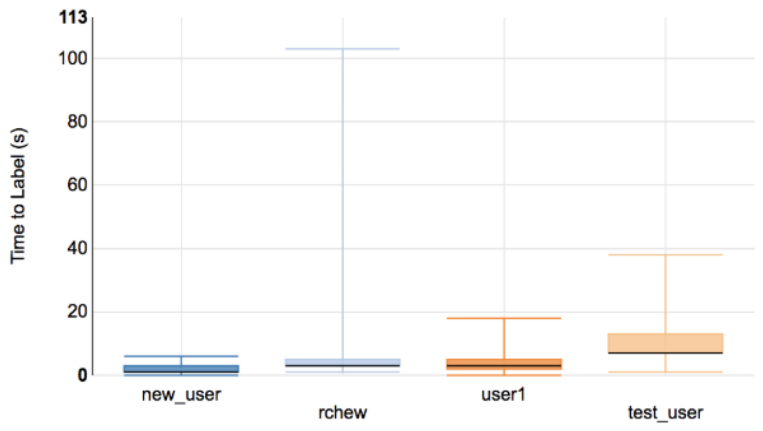
Not Hotdog

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.

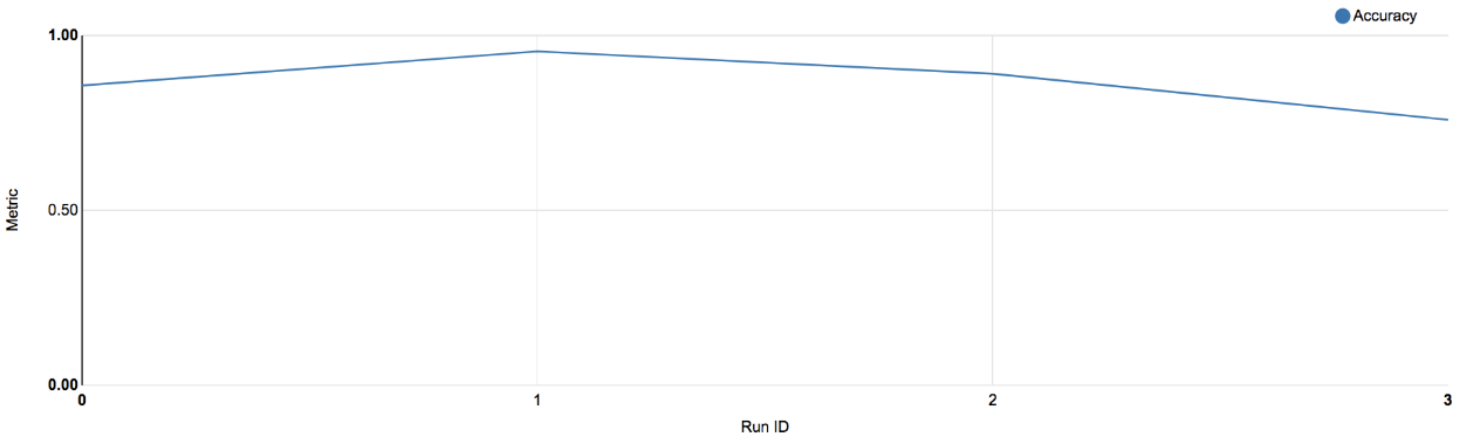
Not Hotdog

Labeled Data | Active Learning Model | IRR

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/

User Documentation - https://smart-app.readthedocs.io/

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/RTIInternational/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

There's also a paper!

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]

Questions?

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

RTI INTERNATIONAL

Caroline Kery
Data Scientist
RTI International