



UNIVERSITY OF  
CAMBRIDGE

# Multi-Modal Training Data Creation with Snorkel

Open-Source Project

Presented by Dmitry Kazhdan

# Background

- Huge interest in Deep Learning
- Deep Learning needs large labelled training datasets

# Snorkel System

## Benefits:

- User-friendly (supported by study)
- Quicker than hand-labelling
- Flexible
- Conceptually intuitive

## But:

- Needs a context hierarchy
- Only evaluated on textual datasets



# Project Goals

- Apply Snorkel to a non-textual dataset
- Apply Snorkel to a Multi-Modal dataset
- Train a Multi-Modal classifier
- Show Snorkel's ability to exploit dataset correlations

# Tasks

- Pick a suitable benchmark multi-modal task
- Pick available labelled dataset
- Pick benchmark model
- Regenerate the labels of the dataset, using Snorkel
- Compare the two training datasets (directly and using the model)

# Project Extensions

- Labelling function can in principle be of any form
- Simply a partial mapping of inputs to outputs
- Extension is to investigate which other approaches may be used as LFs, e.g.:
  - Clustering
  - Semi-supervised learning
  - Weaker classifiers
  - ...



# Plan

- Study relevant online resources
- Develop proof-of-concept approach with a non-textual dataset
- Select a well-known multi-modal task
- Use Snorkel to label the corresponding dataset (key step)
- Evaluate approach
- Work on extensions (if time permits)



**Questions?**