Reverse EXPO 2022

UNIVERSITY
OF ALBERTA

Hello!

I’m Andrea, a Master’s student in Computing Science.

My primary research area is NLP, specifically in applying **data science** and **machine learning** to **social communication**.

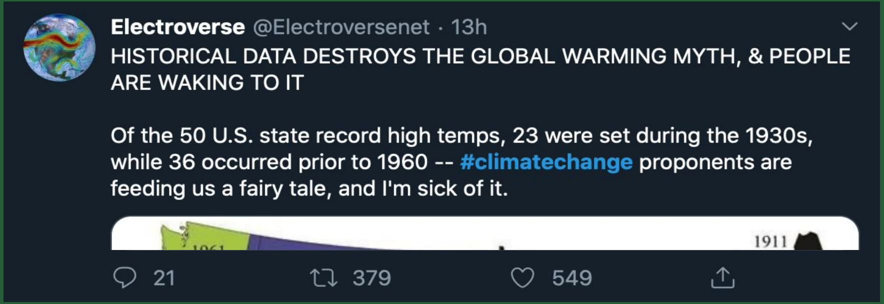
Framing Climate Change

"To frame is to select some **aspects** of a perceived reality and make them more **salient** in a communicating text, in such a way as to promote a particular **problem** definition, **causal** interpretation, **moral** evaluation, and/or **treatment** recommendation for the item described."

- frames are manifested through **frame elements**: *keywords*, *stock phrases*, stereotyped images, *sources of information*, and *sentences*
- receivers' responses are clearly affected if they perceive and process information about *one interpretation* and possess little or incommensurable data about *alternatives*

Framing Climate Change

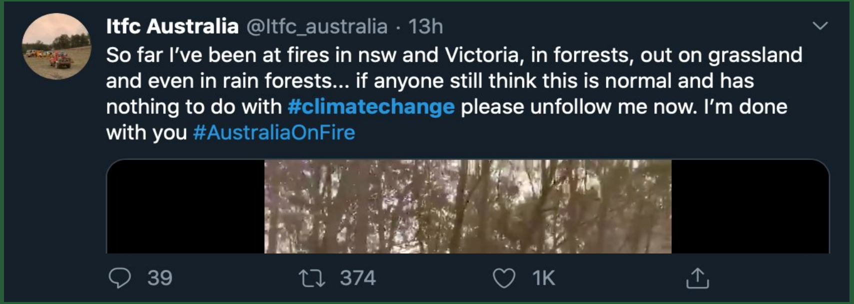
- How are climate change issues framed on social media?
- **Problem**: no scalable model of general and topical issue frames, nor a flexible methodology for creating one
- **Approach**: using ADaPT-ML to implement frame elements as labelling functions



Electroverse @Electroversenet · 13h

HISTORICAL DATA DESTROYS THE GLOBAL WARMING MYTH, & PEOPLE ARE WAKING TO IT

Of the 50 U.S. state record high temps, 23 were set during the 1930s, while 36 occurred prior to 1960 -- #climatechange proponents are feeding us a fairy tale, and I'm sick of it.



Itfc Australia @Itfc_australia · 13h

So far I've been at fires in nsw and Victoria, in forrests, out on grassland and even in rain forests... if anyone still think this is normal and has nothing to do with #climatechange please unfollow me now. I'm done with you #AustraliaOnFire

ADaPT-ML

A Data Programming Template for Machine Learning

Overview

1. Background
2. Why make ADaPT-ML? What is it?
3. System walkthrough
4. ADaPTing to new use cases

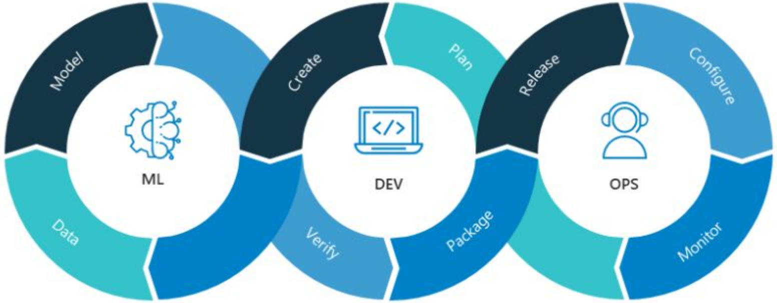
Background

Machine Learning (ML)

- Machine learning algorithms build a **model** based on sample data, known as **training data**, in order to make **predictions** or decisions without being explicitly programmed to do so.

Machine Learning Operations (MLOps)

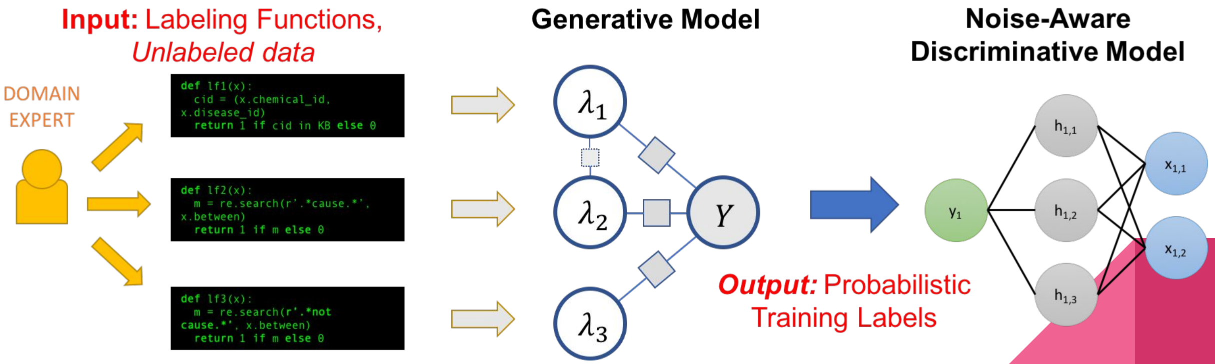
- a set of practices that aims to **deploy** and maintain machine learning **models** in production **reliably and efficiently**.



Background

Data Programming

- “a paradigm for the programmatic **creation of training sets** called data programming in which users express **weak supervision strategies** or domain **heuristics** as **labeling functions**, which are programs that label subsets of the data, but that are noisy and may conflict. We show that by explicitly representing this training set labeling process as a **generative model**, we can "**denoise**" the generated training set”



```
def lf1(x):
  cid = (x.chemical_id,
x.disease_id)
  return 1 if cid in KB else 0
```

```
def lf2(x):
  m = re.search(r'.*cause.*',
x.between)
  return 1 if m else 0
```

```
def lf3(x):
  m = re.search(r'.*not
cause.*', x.between)
  return 1 if m else 0
```

Why make ADaPT-ML?

Have you needed to model a phenomenon for which there is **insufficient training data**, and **no resources** to acquire it?

> *I have!*

Is there an **open-source** MLOps platform out there with **data programming** at its core?

> *There was not!*

Could you manage by simply using **Snorkel** as a **standalone tool**?

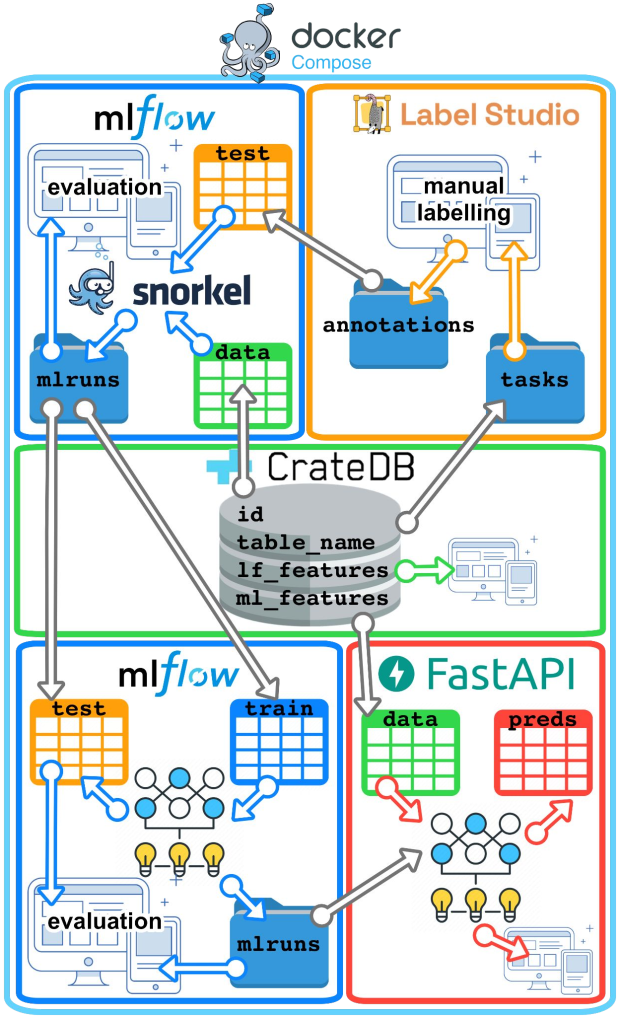
> *You could, but we all know that not having full integration causes headaches.*

So, what exactly is ADaPT-ML?

It is a **multimodal**-ready **MLOps** system that covers the data processing, **data labelling**, model design, model training and optimization, and endpoint deployment.

This software was created especially for any researcher with:

- Some programming experience or interest in learning how to write code based off of examples.
- Access to large amounts of unlabelled data that is constantly changing, such as social media data.
- Domain expertise or an intuition about how they would follow rules, heuristics, or use knowledge bases to annotate the unlabelled data.

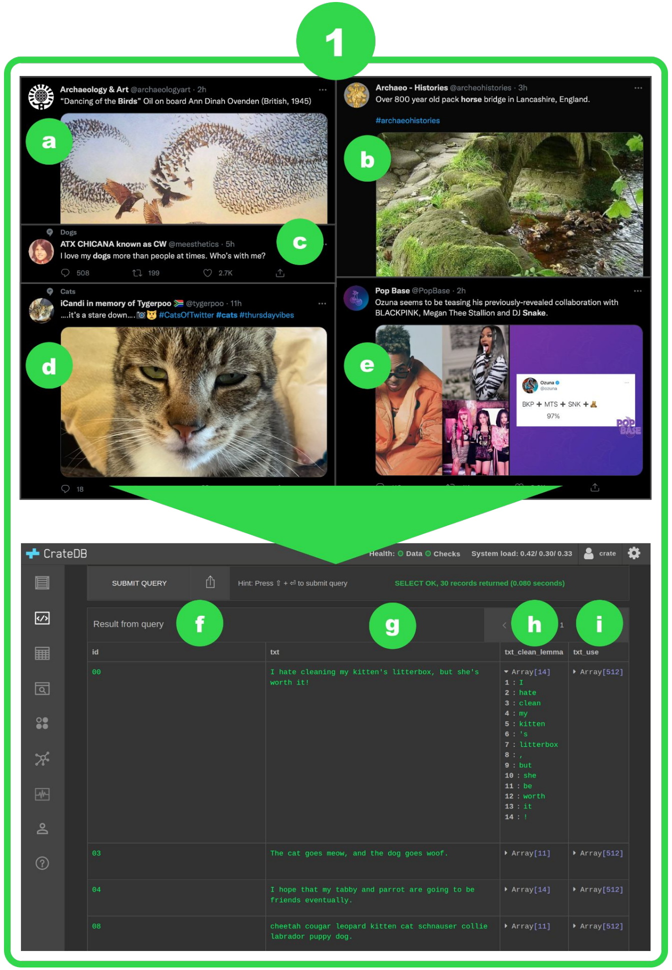


Questions?

System Walkthrough

Data Prerequisites

- Large amounts of unlabelled data in CrateDB
- Features extracted for labelling functions
- Feature vectors for machine learning



Archaeology & Art @archaeologyart · 2h
"Dancing of the **Birds**" Oil on board Ann Dinah Ovenden (British, 1945)

Archaeo - Histories @archeohistories · 3h
Over 800 year old pack **horse** bridge in Lancashire, England.

#archaeohistories

Dogs
ATX CHICANA known as CW @meesthetics · 5h
I love my **dogs** more than people at times. Who's with me?
508 199 2.7K

Cats
iCandi in memory of Tygerpoo 🇿🇦 @tygerpoo · 11h
....it's a stare down....📷😾 #CatsOfTwitter #**cats** #thursdayvibes

Pop Base @PopBase · 2h
Ozuna seems to be teasing his previously-revealed collaboration with BLACKPINK, Megan Thee Stallion and DJ **Snake**.

CrateDB

Health: ○ Data ○ Checks System load: 0.42/ 0.30/ 0.33 crate

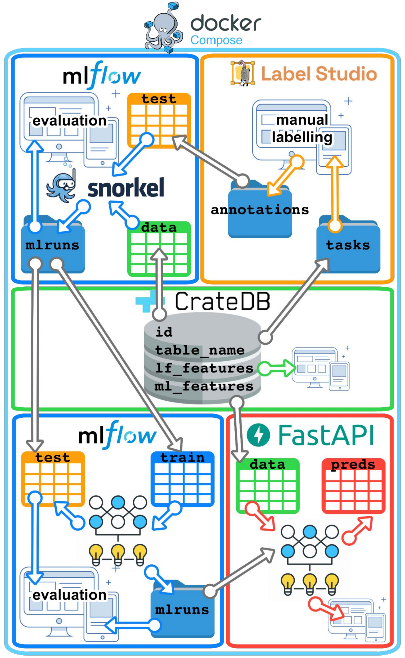
SUBMIT QUERY

Hint: Press ⇧ + ⏎ to submit query

SELECT OK, 30 records returned (0.080 seconds)

Result from query

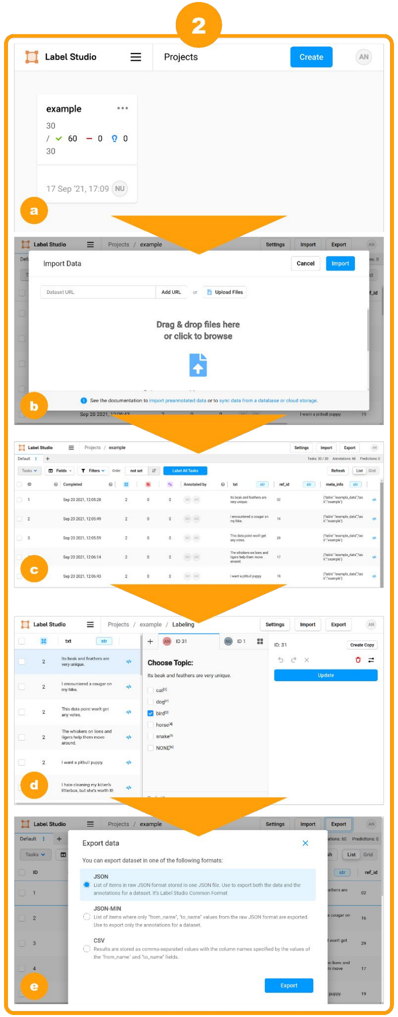
id	txt	txt_clean_lemma	txt_use
00	I hate cleaning my kitten's litterbox, but she's worth it!	▾ Array[14] 1 : I 2 : hate 3 : clean 4 : my 5 : kitten 6 : 's 7 : litterbox 8 : , 9 : but 10 : she 11 : be 12 : worth 13 : it 14 : !	▸ Array[512]
03	The cat goes meow, and the dog goes woof.	▸ Array[11]	▸ Array[512]
04	I hope that my tabby and parrot are going to be friends eventually.	▸ Array[14]	▸ Array[512]
08	cheetah cougar leopard kitten cat schnauser collie labrador puppy dog.	▸ Array[11]	▸ Array[512]



System Walkthrough

Create a Human-annotated Testing Dataset

- Sample some data points to manually label in Label Studio
- Customize the labelling configuration to handle images, text, audio, conversations, timelines, etc.
- Export the annotations and determine the inter-annotator agreement
- Choose how you want to handle annotator disagreements to finalize one set of gold labels



example
30 / ✓ 60 – 0 0
30
17 Sep '21, 17:09 NU

Projects / example

Import Data

Cancel | Import

Dataset URL | Add URL | or | Upload Files

Drag & drop files here
or click to browse

See the documentation to import preannotated data or to sync data from a database or cloud storage.

Default | + | Tasks: 30 / 30 | Annotations: 60 | Predictions: 0

Tasks | Fields | Filters | Order | not set | Label All Tasks | Refresh | List | Grid

ID	Completed				Annotated by	txt	ref_id	meta_info
1	Sep 20 2021, 12:05:28	2	0	0	NU AN	Its beak and feathers are very unique.	02	{"table":"example_data","task":"example"}
2	Sep 20 2021, 12:05:49	2	0	0	NU AN	I encountered a cougar on my hike.	16	{"table":"example_data","task":"example"}
3	Sep 20 2021, 12:05:59	2	0	0	NU AN	This data point won't get any votes.	29	{"table":"example_data","task":"example"}
	Sep 20 2021, 12:06:14	2	0	0	NU AN	The whiskers on lions and tigers help them move around.	17	{"table":"example_data","task":"example"}
	Sep 20 2021, 12:06:43	2	0	0	NU AN	I want a pitbull puppy.	19	{"table":"example_data","task":"example"}

Projects / example / Labeling

	txt
2	Its beak and feathers are very unique.
2	I encountered a cougar on my hike.
2	This data point won't get any votes.
2	The whiskers on lions and tigers help them move around.
2	I want a pitbull puppy.
	I hate cleaning my kitten's litterbox, but she's worth it!

AN ID 31 | NU ID 1

Choose Topic:

Its beak and feathers are very unique.

- □ cat[1]
- □ dog[2]
- ☑ bird[3]
- □ horse[4]
- □ snake[5]
- □ NONE[6]

ID: 31 | Create Copy

Update

Export data

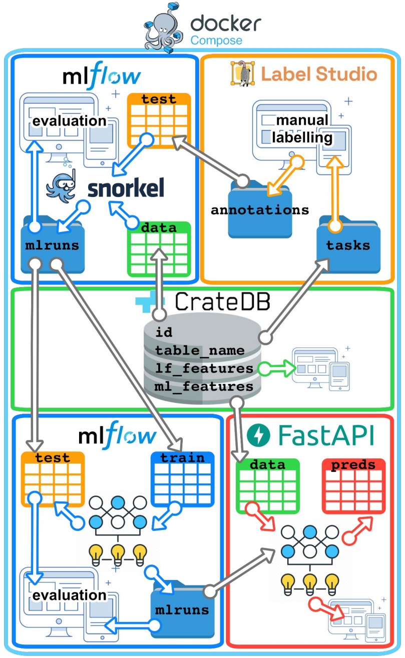
You can export dataset in one of the following formats:

● JSON
List of items in raw JSON format stored in one JSON file. Use to export both the data and the annotations for a dataset. It's Label Studio Common Format

○ JSON-MIN
List of items where only "from_name", "to_name" values from the raw JSON format are exported. Use to export only the annotations for a dataset.

○ CSV
Results are stored as comma-separated values with the column names specified by the values of the "from_name" and "to_name" fields.

Export

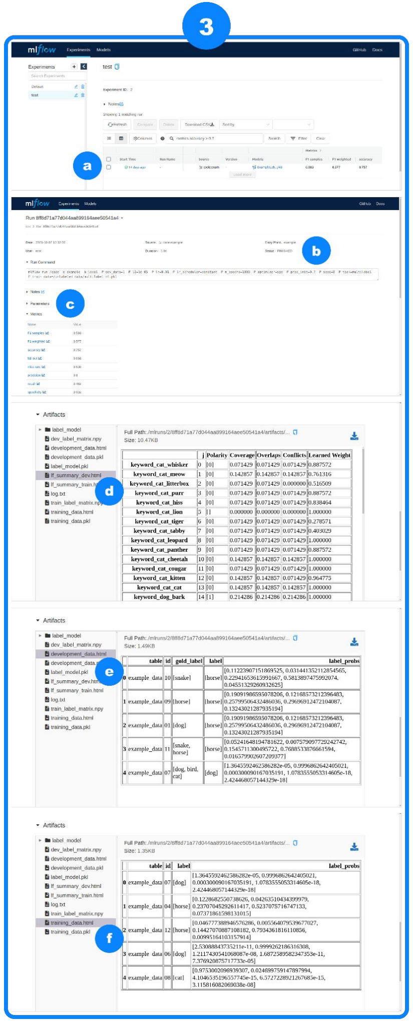


amwhitta@ualberta.ca

System Walkthrough

Programmatically Create Training Data

- Sample some data points from CrateDB that represent your domain
- Run an experiment with a set of base and custom parameters, including your labelling function features
  - The testing dataset is used to validate the label model
- Monitor the results with the MLflow UI



Run 8ff8d71a77d044aa899164aee50541a4

Start Time	Run Name	Source	Version	Models	F1 samples	F1 weighted	accuracy
14 days ago	-	code:exam	-	ExampleLab../48	0.583	0.577	0.757

Status: FINISHED

- label_model
- dev_label_matrix.npy
- development_data.html
- development_data.pkl
- label_model.pkl
- lf_summary_dev.html
- lf_summary_train.html
- log.txt
- train_label_matrix.npy
- training_data.html
- training_data.pkl

Full Path:./mlruns/2/8ff8d71a77d044aa899164aee50541a4/artifacts/...
Size: 10.47KB

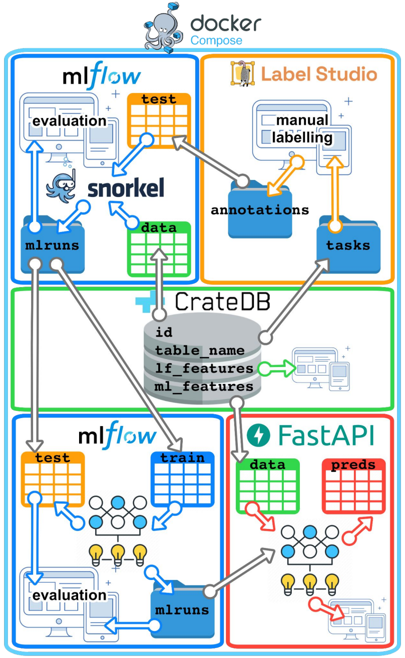
	j	Polarity	Coverage	Overlaps	Conflicts	Learned Weight
keyword_cat_whisker	0	[0]	0.071429	0.071429	0.071429	0.887572
keyword_cat_meow	1	[0]	0.142857	0.142857	0.142857	0.761316
keyword_cat_litterbox	2	[0]	0.071429	0.071429	0.000000	0.516509
keyword_cat_purr	3	[0]	0.071429	0.071429	0.071429	0.887572
keyword_cat_hiss	4	[0]	0.071429	0.071429	0.071429	0.838464
keyword_cat_lion	5	[]	0.000000	0.000000	0.000000	1.000000
keyword_cat_tiger	6	[0]	0.071429	0.071429	0.071429	0.278571
keyword_cat_tabby	7	[0]	0.071429	0.071429	0.071429	0.403029
keyword_cat_leopard	8	[0]	0.071429	0.071429	0.071429	1.000000
keyword_cat_panther	9	[0]	0.071429	0.071429	0.071429	0.887572
keyword_cat_cheetah	10	[0]	0.142857	0.142857	0.142857	1.000000
keyword_cat_cougar	11	[0]	0.071429	0.071429	0.071429	1.000000
keyword_cat_kitten	12	[0]	0.142857	0.142857	0.071429	0.964775
keyword_cat_cat	13	[0]	0.142857	0.142857	0.142857	1.000000
keyword_dog_bark	14	[1]	0.214286	0.214286	0.214286	1.000000

Full Path:./mlruns/2/8ff8d71a77d044aa899164aee50541a4/artifacts/...
Size: 1.49KB

	table	id	gold_label	label	label_probs
0	example_data	10	[snake]	[horse]	[0.11223907151869525, 0.031441352112854565, 0.22941653615991667, 0.5813897475992074, 0.04551329260932625]
1	example_data	09	[horse]	[horse]	[0.19091986595078206, 0.12168573212396483, 0.25799506432486036, 0.29696912472104087, 0.13243021287935194]
2	example_data	01	[dog]	[horse]	[0.19091986595078206, 0.12168573212396483, 0.25799506432486036, 0.29696912472104087, 0.13243021287935194]
3	example_data	11	[snake, horse]	[horse]	[0.05241648194781622, 0.00757909772924742, 0.1545711300495722, 0.7688533876661594, 0.01657990260720937 7]
4	example_data	07	[dog, bird, cat]	[dog]	[1.3645592462586282e-05, 0.9996862642405021, 0.0003000901670335191, 1.0783555053314605e-18, 2.4244680571443 29e-18]

Full Path:./mlruns/2/8ff8d71a77d044aa899164aee50541a4/artifacts/...
Size: 1.35KB

	table	id	label	label_probs
0	example_data	07	[dog]	[1.3645592462586282e-05, 0.9996862642405021, 0.0003000901670335191, 1.0783555053314605e-18, 2.4244680571443 29e-18]
1	example_data	04	[horse]	[0.1228682550738626, 0.04263510434399979, 0.2370704529261141 7, 0.5237075716747133, 0.07371861598131015]
2	example_data	12	[horse]	[0.04677738894657628 6, 0.00556407953967702 7, 0.14427070887108182, 0.7934361816110856, 0.00995164103157914]
3	example_data	06	[dog]	[2.530888437352 11e-11, 0.9999262186316308, 1.2117430541068087e-08, 1.6872589582347353e-11, 7.376920875717733e-05]
4	example_data	08	[cat]	[0.9753002096939307, 0.024699759147897994, 4.1046535196557745e-15, 6.5727228921267685e-15, 3.115816082069038e-08]

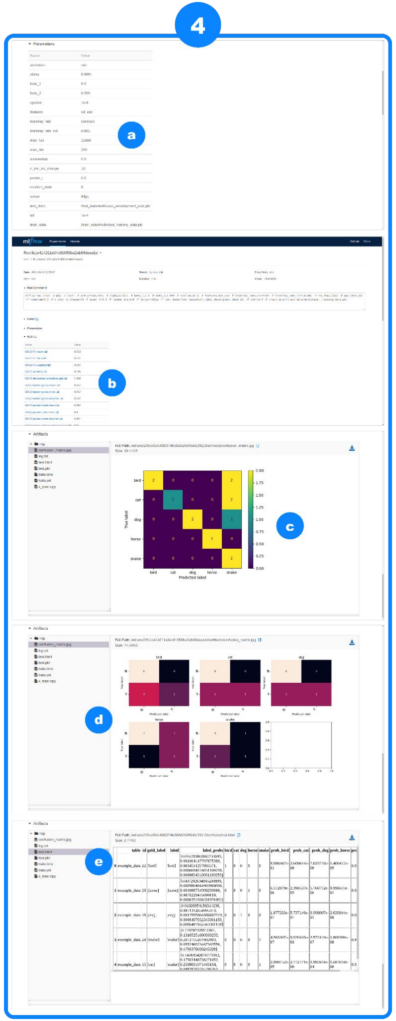


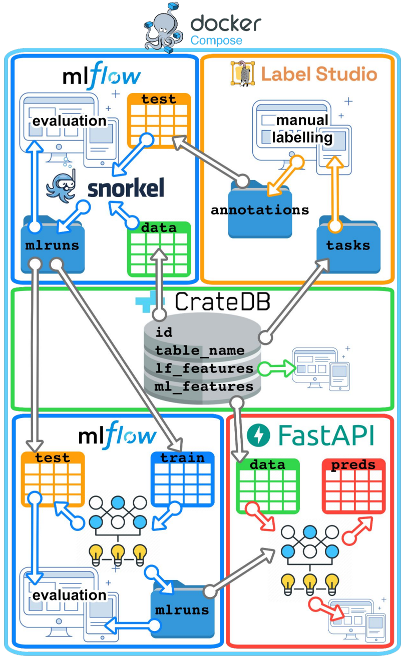
amwhitta@ualberta.ca

System Walkthrough

Create the Discriminative End Model

- Run an experiment using a suitable machine learning algorithm and your feature vectors
  - The model will be evaluated using the gold labels and label model's labels
- Observe the results with the MLflow UI

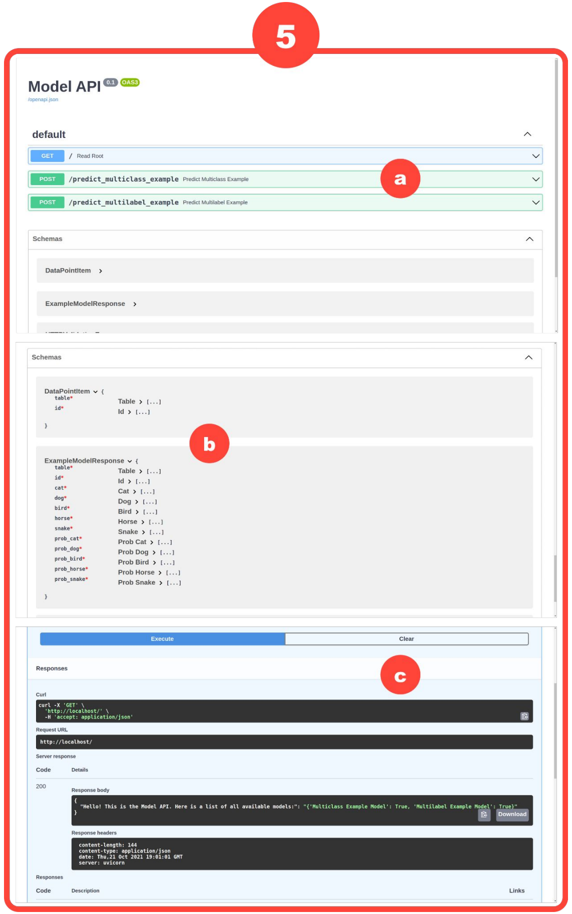




System Walkthrough

Deploy the Best Model

- Get predictions for any data point, needing only to supply the table name and id
- Can keep track of which model has been deployed within the MLflow UI



Model API 0.1 OAS3
/openapi.json

default

GET / Read Root

POST /predict_multiclass_example Predict Multiclass Example

POST /predict_multilabel_example Predict Multilabel Example

Schemas

DataPointItem ›

ExampleModelResponse ›

Schemas

```
DataPointItem ⌄ {
   table*          Table › [...]
   id*             Id › [...]
}

ExampleModelResponse ⌄ {
   table*          Table › [...]
   id*             Id › [...]
   cat*            Cat › [...]
   dog*            Dog › [...]
   bird*           Bird › [...]
   horse*          Horse › [...]
   snake*          Snake › [...]
   prob_cat*       Prob Cat › [...]
   prob_dog*       Prob Dog › [...]
   prob_bird*      Prob Bird › [...]
   prob_horse*     Prob Horse › [...]
   prob_snake*     Prob Snake › [...]
}
```

Execute | Clear

Responses

Curl

```
curl -X 'GET' \
  'http://localhost/' \
  -H 'accept: application/json'
```

Request URL

```
http://localhost/
```

Server response

Code | Details

200

Response body

```
{
  "Hello! This is the Model API. Here is a list of all available models:": "{'Multiclass Example Model': True, 'Multilabel Example Model': True}"
}
```

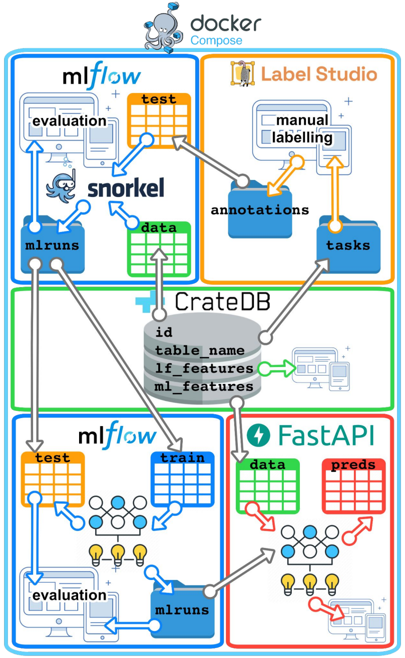
Download

Response headers

```
 content-length: 144
 content-type: application/json
 date: Thu,21 Oct 2021 19:01:01 GMT
 server: uvicorn
```

Responses

Code | Description | Links



amwhitta@ualberta.ca

Questions?

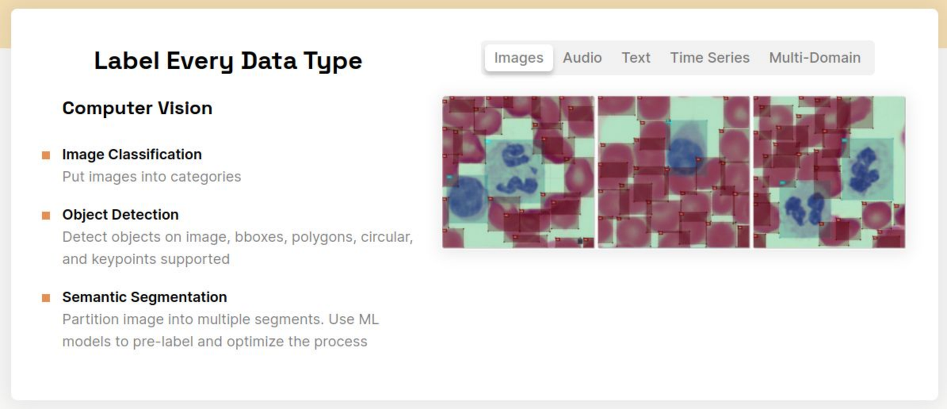
ADaPTing to New Use Cases * = future features

Label Studio

- Define classification task name and categories
- Format the Labelling Configuration

★ Implement methods for taking representative samples

★ Expand data formatting for more modalities



Label Every Data Type

Computer Vision

- **Image Classification**
  Put images into categories
- **Object Detection**
  Detect objects on image, bboxes, polygons, circular, and keypoints supported
- **Semantic Segmentation**
  Partition image into multiple segments. Use ML models to pre-label and optimize the process

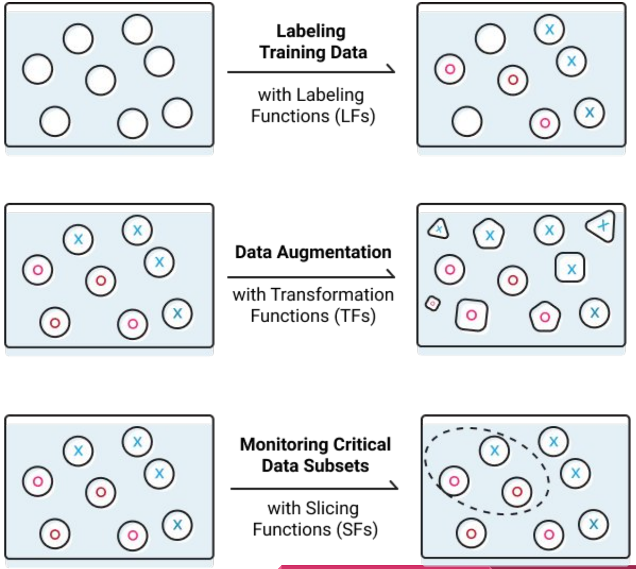
ADaPTing to New Use Cases

* = future features

Data Programming

- Write your labelling functions
- Create an MLflow endpoint
  - parameters for labelling functions
  - loader functions for labelling function features

★ Integrate slicing and transformation functions



Labeling Training Data
with Labeling Functions (LFs)
Data Augmentation
with Transformation Functions (TFs)
Monitoring Critical Data Subsets
with Slicing Functions (SFs)

ADaPTing to New Use Cases

* = future features

Model Creation and Deployment

- Define the response format and endpoint

★ Implement more machine learning algorithms

Thank you for your interest in this project!

https://github.com/U-Alberta/ADaPT-ML

amwhitta@ualberta.ca

Resources

https://academic.oup.com/joc/article/43/4/51-58/4160153	(Entman's publication on framing)
https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/	(MLOps diagram)
https://en.wikipedia.org/wiki/Machine_learning	(definition of ML)
https://en.wikipedia.org/wiki/MLOps	(definition of MLOps)
https://arxiv.org/abs/1605.07723	(publication describing data programming)
https://www.snorkel.org/blog/snorkel-programming	(diagram and description of Snorkel)
https://labelstud.io/	(Label Studio's website)
https://www.snorkel.org/features/	(more description of Snorkel)