@gdequeiroz | linktr.ee/gdq

slides, links and materials: **bit.ly/dc-thurs-ibm**

Gabriela de Queiroz

Sr. Machine Learning Manager, IBM

- Founder of **R-Ladies** (rladies.org)
- Founder of **AI Inclusive** (ai-inclusive.org)
- Member of **R Foundation** (r-project.org)

- B.S. in Statistics
- MSc. in Epidemiology
- MSc. in Statistics

Data Scientist + Developer Advocate + Open Source Developer + Manager +

Statistician + Epidemiologist + Community Builder + Mentor + Speaker + Educator

Population: ~200 million

Official Language: Portuguese

Population: ~6 million

Researcher/Statistician

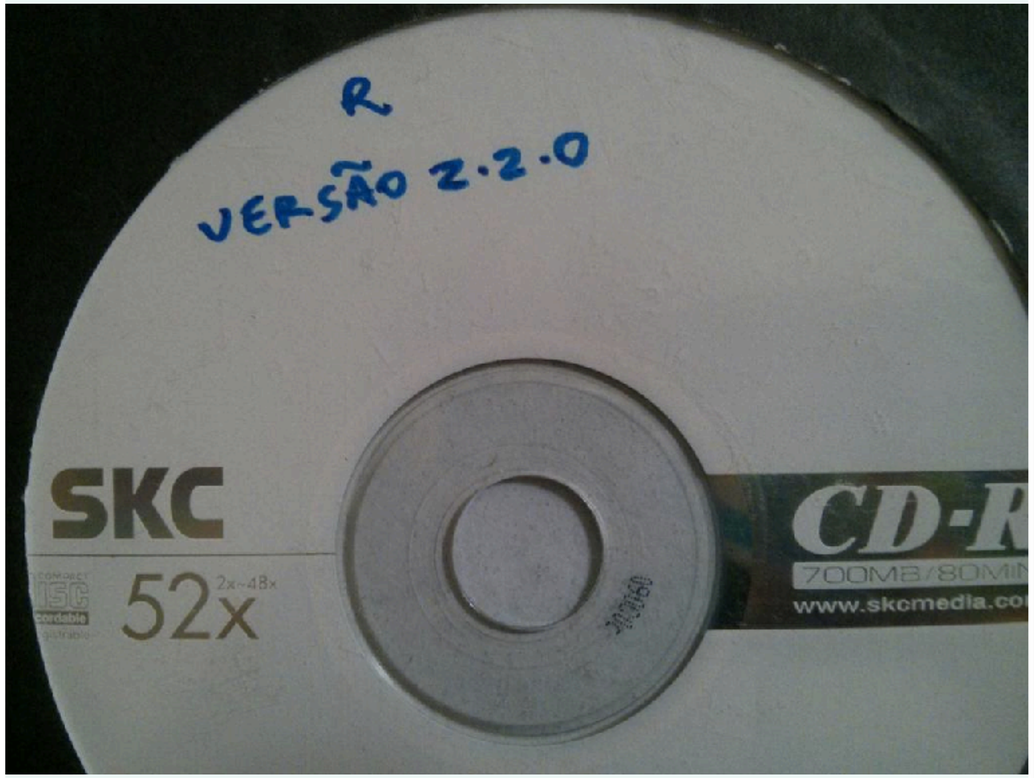
Undergraduate

- STATE UNIVERSITY
- PUBLIC AND FREE UNIVERSITY
- **BACHELOR IN STATISTICS**

Grad School

- SCIENTIFIC INSTITUTION FOR RESEARCH
- PUBLIC AND FREE
- **MSC. IN EPIDEMIOLOGY**

❤️ First exposure to R ❤️



R 2.2.0 (October, 2005)

R 3.1.0 (April, 2014)
R 3.0.3 (March, 2014)
R 3.0.2 (September, 2013)
R 3.0.1 (May, 2013)
R 3.0.0 (April, 2013)
R 2.15.3 (March, 2013)
R 2.15.2 (October, 2012)
R 2.15.1 (June, 2012)
R 2.15.0 (March, 2012)
R 2.14.2 (February, 2012)
R 2.14.1 (December, 2011)
R 2.14.0 (November, 2011)
R 2.13.2 (September, 2011)
R 2.13.1 (July, 2011)
R 2.13.0 (April, 2011)
R 2.12.2 (February, 2011)
R 2.12.1 (December, 2010)
R 2.12.0 (October, 2010)
R 2.11.1 (May, 2010)
R 2.11.0 (April, 2010)
R 2.10.1 (December, 2009)
R 2.10.0 (October, 2009)
R 2.9.2 (August, 2009)
R 2.9.1 (June, 2009)
R 2.9.0 (April, 2009)
R 2.8.1 (December, 2008)
R 2.8.0 (October, 2008)
R 2.7.2 (August, 2008)
R 2.7.1 (June, 2008)
R 2.7.0 (April, 2008)
R 2.6.2 (February, 2008)
R 2.6.1 (November, 2007)
R 2.6.0 (October, 2007)
R 2.5.1 (July, 2007)
R 2.5.0 (April, 2007)
R 2.4.1 (December, 2006)
R 2.4.0 (October, 2006)
R 2.3.1 (June, 2006)
R 2.3.0 (April, 2006)
R 2.2.1 (December, 2005)
R 2.2.0 (October, 2005)

R 4.0.3 (October, 2020)
R 4.0.2 (June, 2020)
R 4.0.1 (June, 2020)
R 4.0.0 (April, 2020)
R 3.6.3 (February, 2020)
R 3.6.2 (December, 2019)
R 3.6.1 (July, 2019)
R 3.6.0 (April, 2019)
R 3.5.3 (March, 2019)
R 3.5.2 (December, 2018)
R 3.5.1 (July, 2018)
R 3.5.0 (April, 2018)
R 3.4.4 (March, 2018)
R 3.4.3 (November, 2017)
R 3.4.2 (September, 2017)
R 3.4.1 (June, 2017)
R 3.4.0 (April, 2017)
R 3.3.3 (March, 2017)
R 3.3.2 (October, 2016)
R 3.3.1 (June, 2016)
R 3.3.0 (April, 2016)
R 3.2.5 (April, 2016)
R 3.2.4 (March, 2016)
R 3.2.3 (December, 2015)
R 3.2.2 (August, 2015)
R 3.2.1 (June, 2015)
R 3.2.0 (April, 2015)
R 3.1.3 (March, 2015)
R 3.1.2 (October, 2014)
R 3.1.1 (July, 2014)

The R Inferno

Patrick Burns[1]

30th April 2011

Contents

List of Figures

List of Tables

1 Falling into the Floating Point Trap

2 Growing Objects

3 Failing to Vectorize
3.1 Subscripting
3.2 Vectorized if
3.3 Vectorization impossible

4 Over-Vectorizing

An Introduction to R

Notes on R: A Programming Environment for Data Analysis and Graphics
Version 2.5.1 (2007-06-27)

W. N. Venables, D. M. Smith
and the R Development Core Team

1 Introduction and preliminaries
1.1 The R environment
1.2 Related software and documentation
1.3 R and statistics
1.4 R and the window system
1.5 Using R interactively
1.6 An introductory session
1.7 Getting help with functions and features
1.8 R commands, case sensitivity, etc.
1.9 Recall and correction of previous commands
1.10 Executing commands from or diverting output to a file
1.11 Data permanency and removing objects

2 Simple manipulations; numbers and vectors
2.1 Vectors and assignment
2.2 Vector arithmetic
2.3 Generating regular sequences
2.4 Logical vectors
2.5 Missing values
2.6 Character vectors
2.7 Index vectors; selecting and modifying subsets of a data set
2.8 Other types of objects

3 Objects, their modes and attributes
3.1 Intrinsic attributes: mode and length
3.2 Changing the length of an object
3.3 Getting and setting attributes
3.4 The class of an object

4 Ordered and unordered factors

HEI

HEALTH EFFECTS INSTITUTE

STATEMENT

Synopsis of Research Report 171

Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study)

BACKGROUND

For nearly two decades, scientists seeking to understand the role that air pollution might play in population health effects have relied heavily on epidemiologic studies known as time-series studies. Time-series studies use information on daily changes in air pollutant concentrations and daily counts of mortality and morbidity. Although initially conducted at the individual city level, coordinated analyses across many cities have recently emerged as the tool of choice for developing more reliable and comparable estimates of the short-term effects of air pollution on health in regions around the world. HEI has a long-standing interest in these coordinated analyses; it has funded studies such as the National Morbidity, Mortality, and Air Pollution Study; Air Pollution and Health: A European and North American Approach; and Public Health and Air Pollution in Asia.

The present study, referred to hereafter by its Spanish acronym ESCALA (Estudio de Salud y Contaminación del Aire en Latinoamérica), was initiated to address underlying data and methodologic limitations in the epidemiologic literature on the health effects of air pollution in Latin America that had been identified in a 2005 review by the Pan American Health Organization. The William and Flora Hewlett Foundation, which has a strong interest in understanding air pollution and health in Latin America, provided HEI with supplemental support to address gaps in the evidence necessary to inform regulatory decisions, and in the process to build a network of health experts capable of carrying out research on air pollution in the future. The multicenter study was led by Dr. Isabelle Romieu, then at the Instituto Nacional de Salud Pública in México, in collaboration with Dr. Nelson Gouveia in Brazil and Dr. Luis Cifuentes in Chile.

APPROACH

The primary objective of the ESCALA study was to estimate the effect of daily exposures to PM_{10} (particulate matter ≤ 10 µm in aerodynamic diameter) and to ozone on daily mortality from several causes (all natural causes, cardiopulmonary disease, respiratory disease, cardiovascular disease, cerebrovascular–stroke, and chronic obstructive pulmonary disease) and for several age groups (all-age, ≥ 65 years, < 1 year, 1–4 years, 1–14 years) in nine Latin American cities, and for the region as a whole, using a common analytic framework. The nine cities were Monterrey, Toluca, and Mexico City in México; Rio de Janeiro, São Paulo, and Porto Alegre in Brazil; and Santiago, Concepción, and Temuco in Chile. Of these, three cities (Porto Alegre, Concepción, and Temuco) were excluded from the ozone analyses because of the lack of adequate ozone monitoring data.

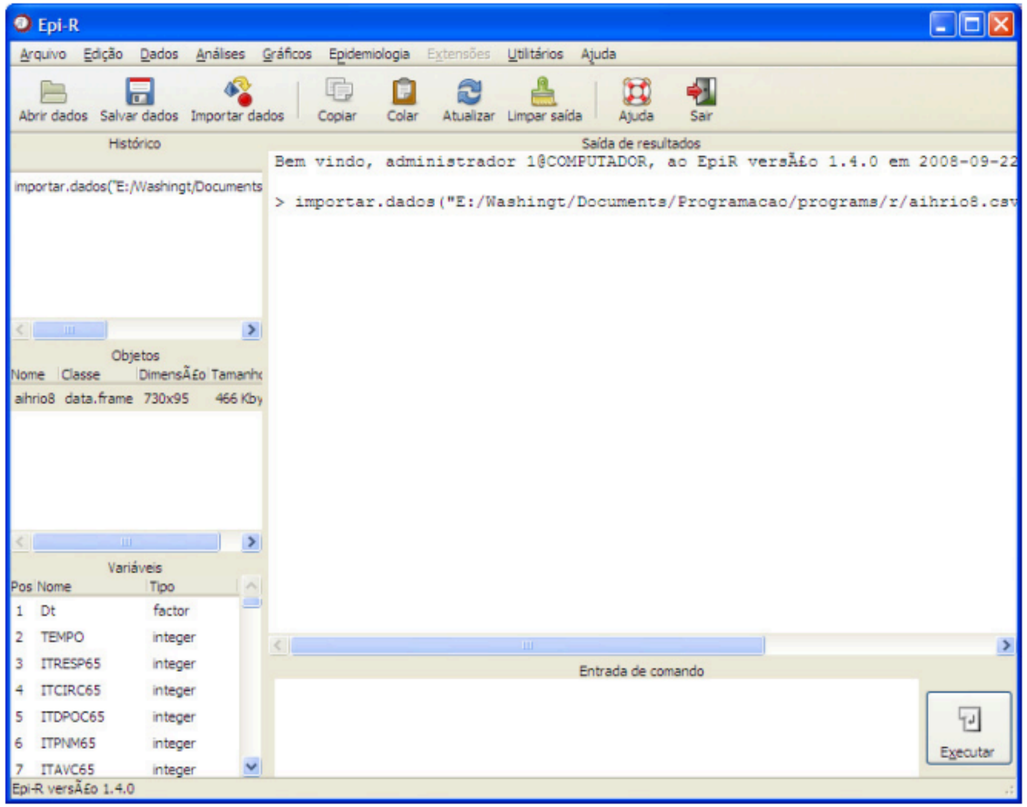
In the first stage of the analyses, the investigators estimated the percentage change in the risk of mortality per 10-µg/m^3 increase in PM_{10} or ozone for each combination of age group and cause of death for the individual cities in each country. They followed a common protocol for fitting the widely used Poisson regression models to the air pollution and mortality time-series data in each city while controlling for other factors that might also explain the temporal patterns of mortality (e.g., temperature, humidity, season, day-of-the-week, holidays). The investigators also carried out analyses to test the sensitivity of the results to various details of the models. Ultimately, the final models used in the individual cities were chosen to fit specific patterns of mortality in those cities.

With the individual city data, the investigators also explored two-pollutant models, in which PM_{10} results were controlled for the presence of ozone and vice versa; whether the association of ozone with mortality differed by warm and cold season;

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Isabelle Romieu, at the Instituto Nacional de Salud Pública, Morelos, México, and colleagues. The complete report, *Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study)* (© 2012 Health Effects Institute), can be obtained from HEI or our Web site (see last page).

Epi-R

A graphic user interface oriented to epidemiological data analysis




Meetup
Data Visualization Group in the Bay Area
San Francisco Hadoop Users
SF Bayarea Machine Learning
SF DATA MINING
#SFMySQL Meetup

R- ladies

It was founded in October 2012.

The idea was to give back to the community and create a place where people would feel comfortable, safe and welcome.

A place where people could ask questions, learn together and share.

31 OCT

Wednesday, October 31, 2012

Introduction to R (beginners and pre-beginners)

Hosted by
Gabriela de Queiroz

Details

Hello R-ladies!

The first meetup will take place on October 31st at the Google office in San Francisco.

For this first meetup, we'll do an introduction to R. We'll go over the following topics:

installing R setting up an R environment (RStudio) basic commands (open files, simple dataset manipulation, simple plots, etc) loading packages the help function and how to read its output

All you need is your laptop and charger.

We look forward to seeing you!

R-Ladies

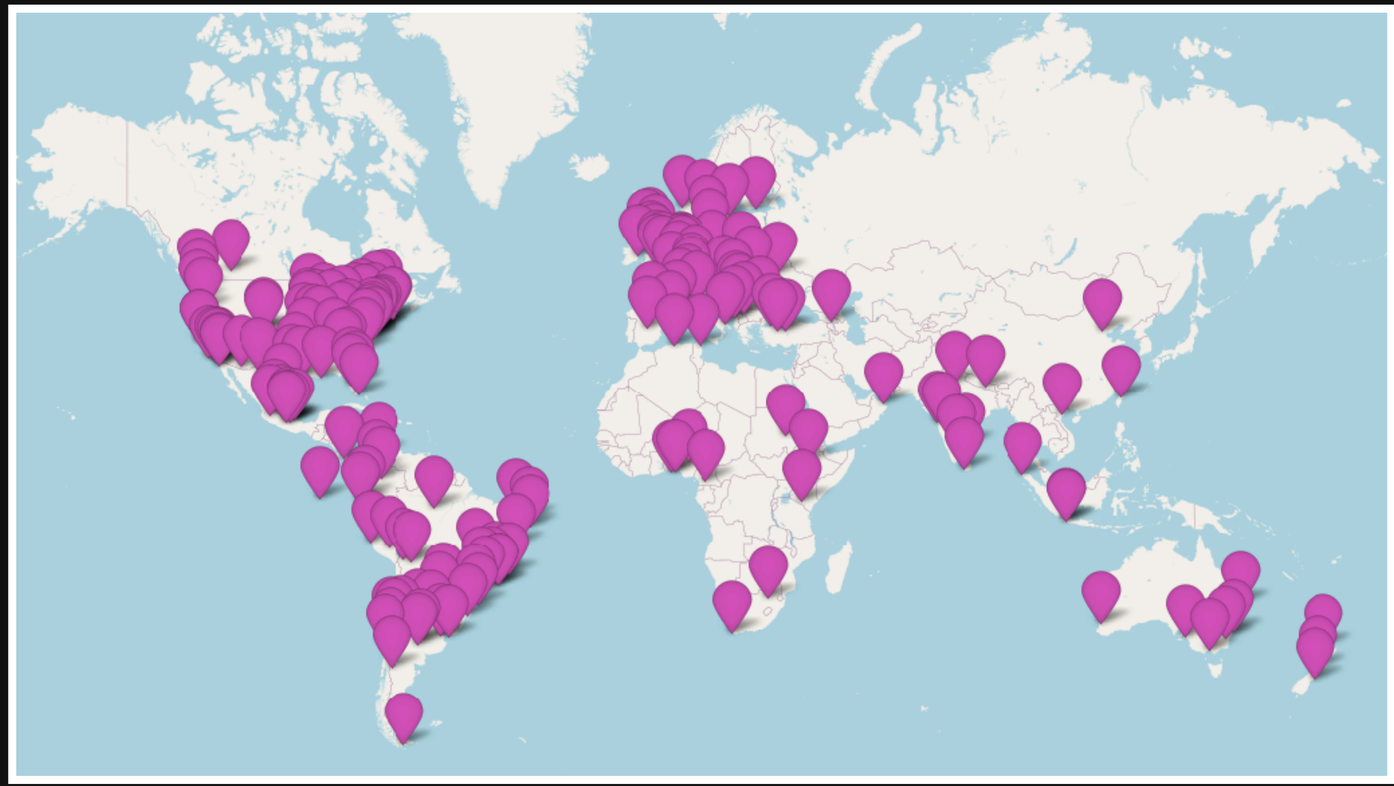
rladies.org

54
R-Ladies Countries

193
R-Ladies Cities

71618
R-Ladies members on meetup.com

Worldwide organization that promotes diversity in the R community via meetups and mentorship in a friendly and safe environment



AI Inclusive

> Together, we are building a community to make **AI** more **inclusive** to everyone.

Mission: Increase the representation and participation of minority groups in Artificial Intelligence

If you want to start a chapter, send us an email: info@ai-inclusive.org

- Website: ai-inclusive.org
- Twitter: bit.ly/ai-inclusive-twitter
- Facebook: bit.ly/ai-inclusive-facebook
- Instagram: bit.ly/ai-inclusive-instagram
- Youtube: bit.ly/ai-inclusive-youtube

Follow us:

bit.ly/ai-inclusive-instagram

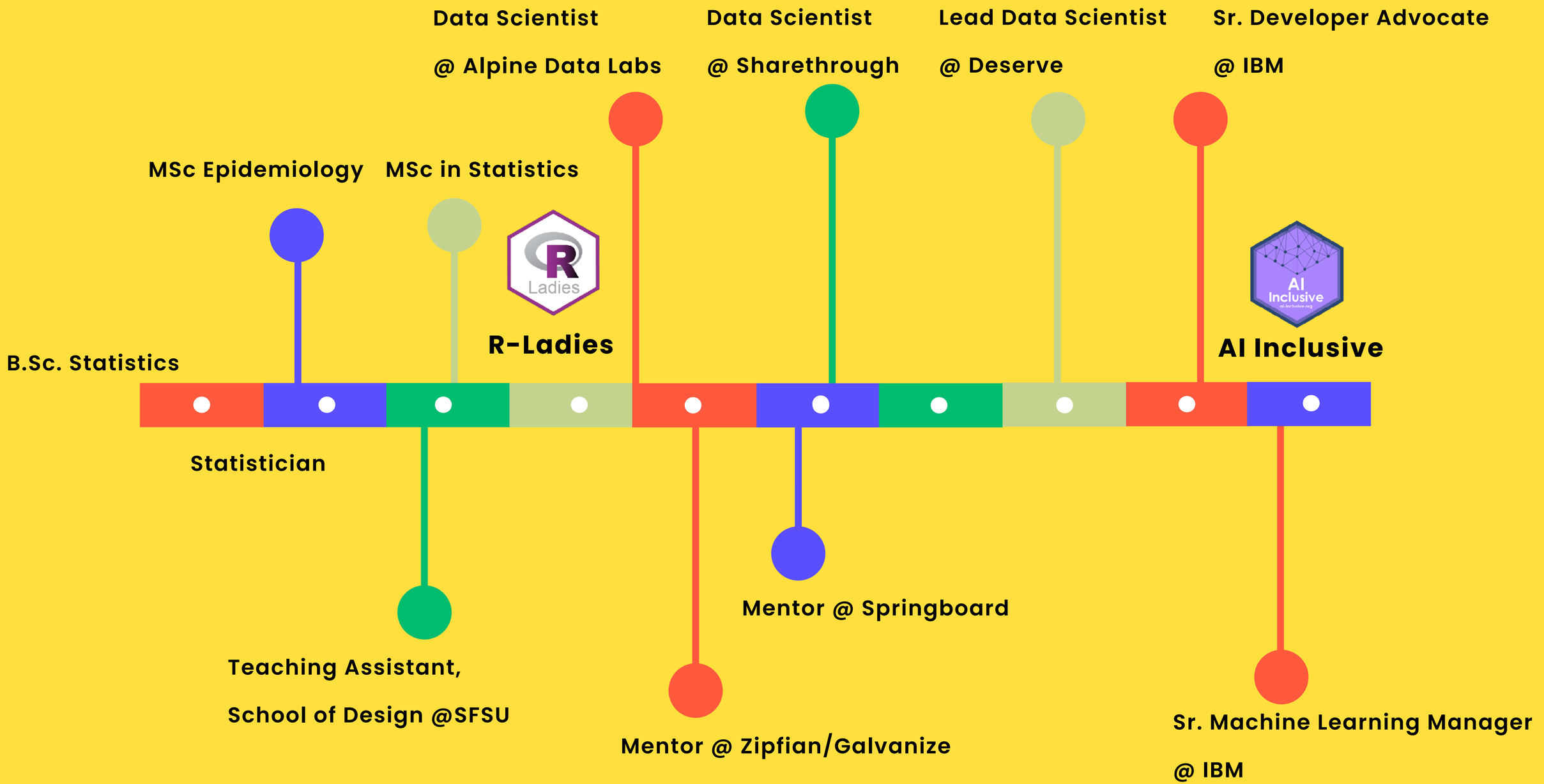
bit.ly/ai-inclusive-twitter

ai-inclusive.org

Resources on AI, DS, ML
Events, Free Tickets and much more

Soon:

Scholarships for Data Science Courses


Data Scientist
@ Alpine Data Labs
Data Scientist
@ Sharethrough
Lead Data Scientist
@ Deserve
Sr. Developer Advocate
@ IBM
MSc Epidemiology
MSc in Statistics
R
Ladies
R-Ladies
AI
Inclusive
AI Inclusive
B.Sc. Statistics
Statistician
Teaching Assistant,
School of Design @SFSU
Mentor @ Springboard
Mentor @ Zipfian/Galvanize
Sr. Machine Learning Manager
@ IBM

Center for Open Source Data & AI Technologies (CODAIT)

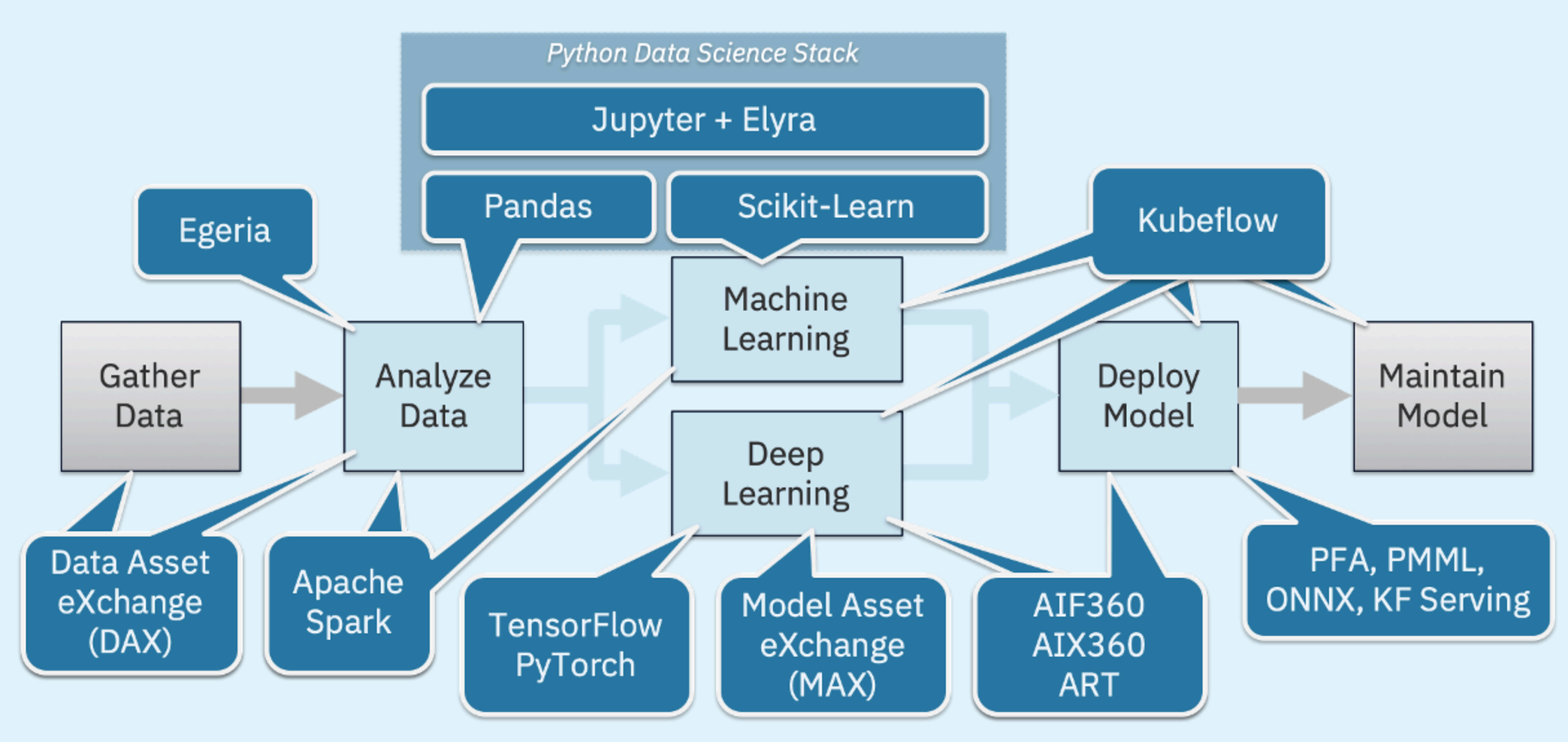
We build tools to make AI accessible and available to everybody

30+ Open Source Developers

codait.org

We build tools to make AI accessible and available to everybody

(codait.org)



We contribute to and advocate for the **open-source** technologies

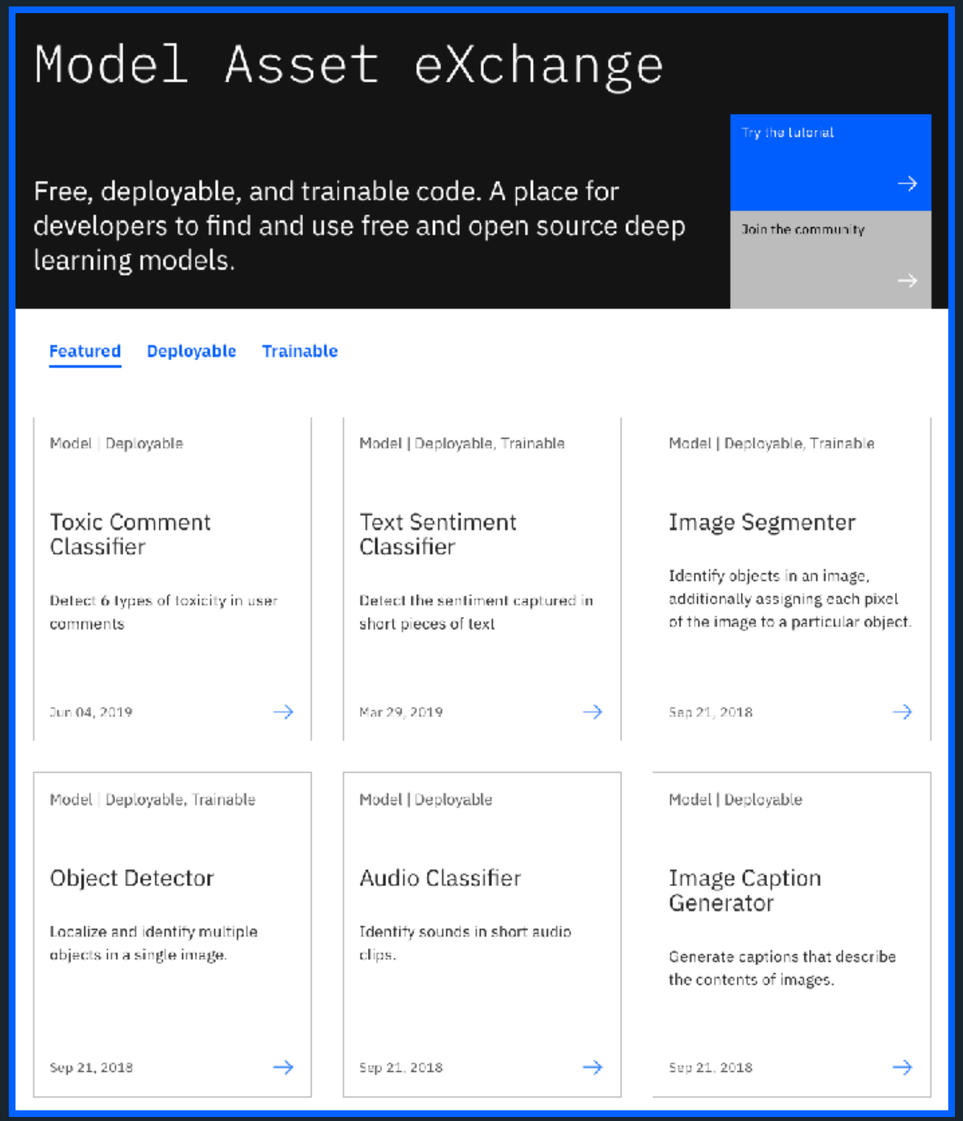
codait.org

Open Source @ IBM

Some Projects

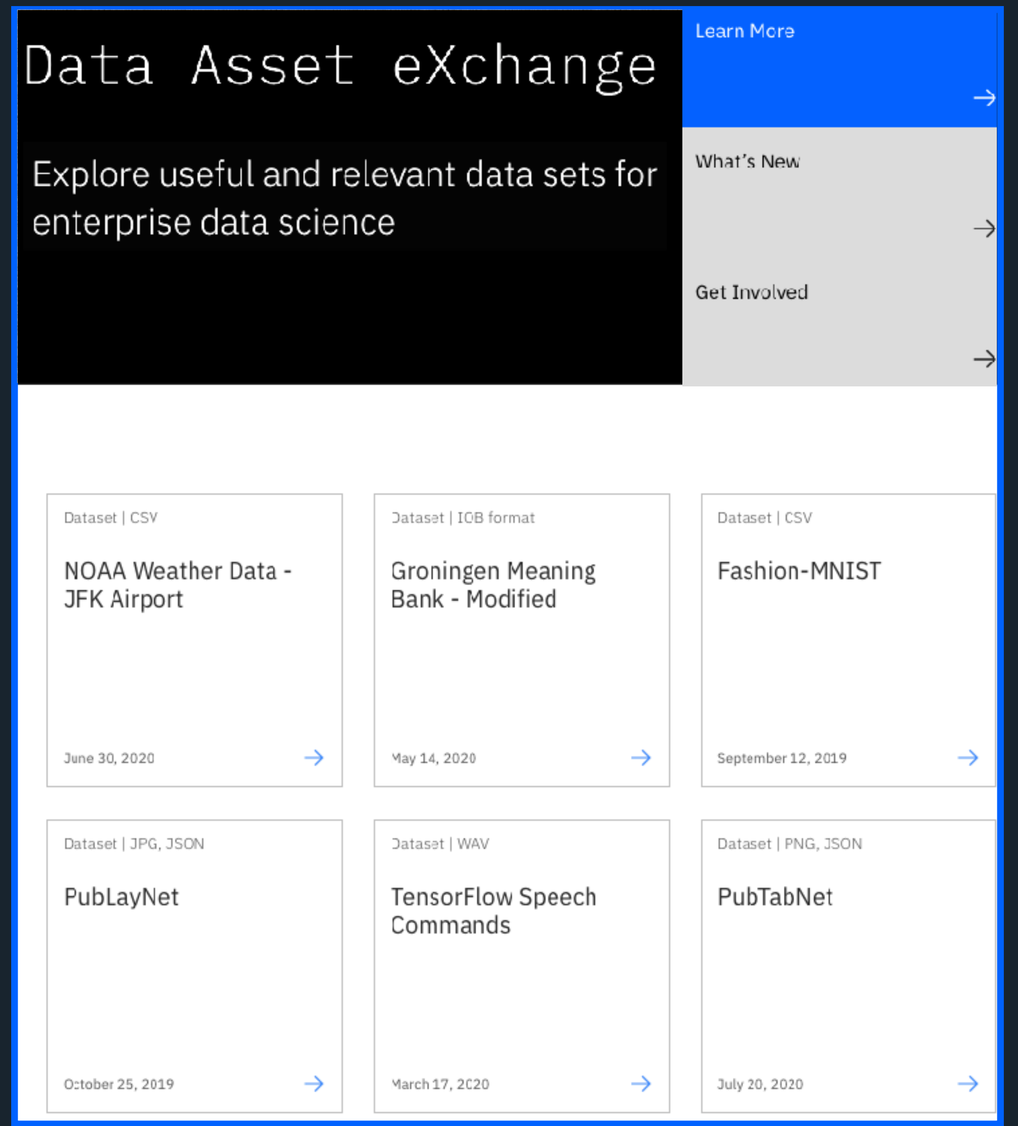
Model Asset eXchange

ibm.biz/model-exchange



Data Asset eXchange

ibm.biz/data-exchange

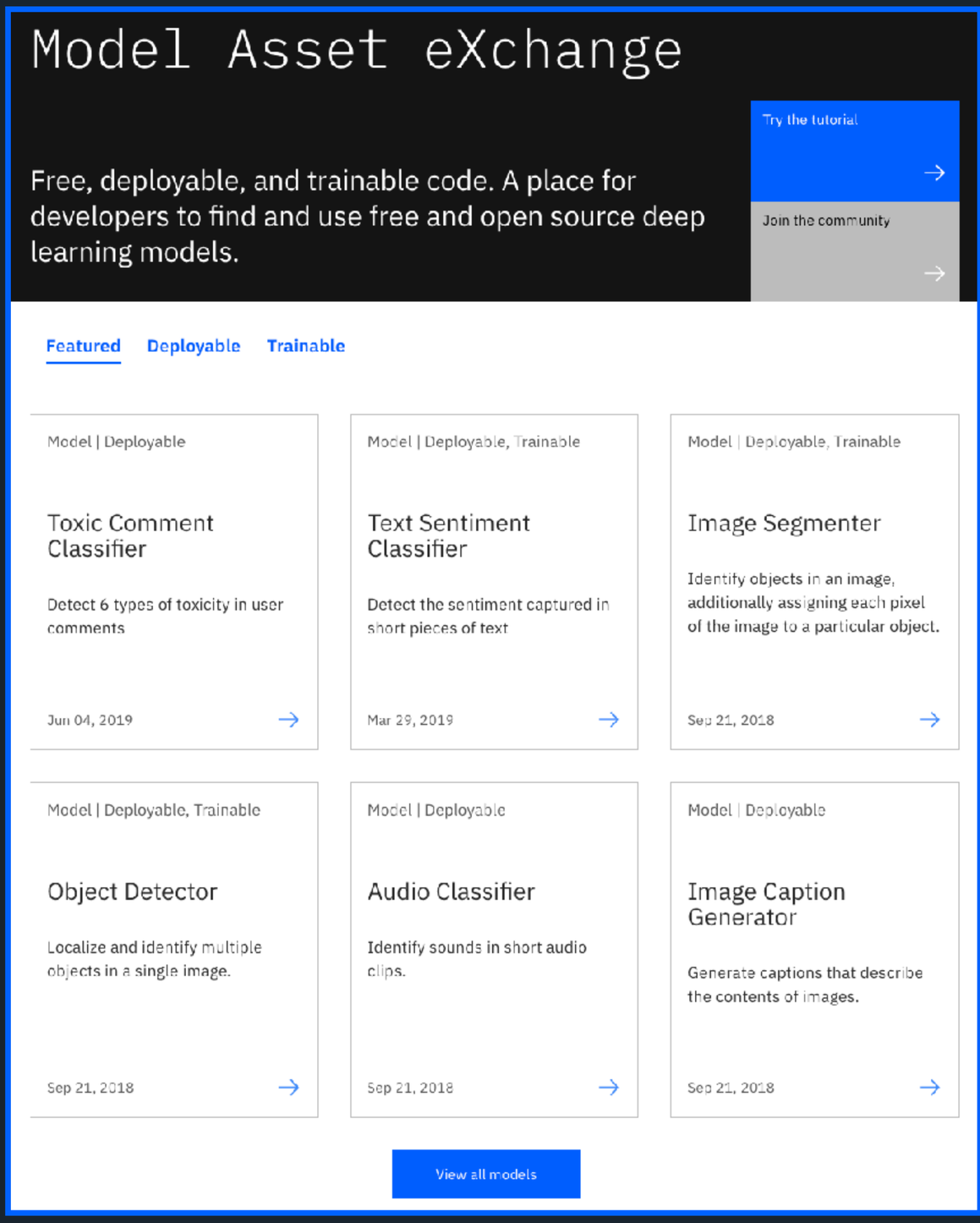


Model Asset eXchange (MAX)

Place for developers/data scientists to find and use

free and **open source** deep learning models

ibm.biz/model-exchange



MAX Object Detector

Upload an image

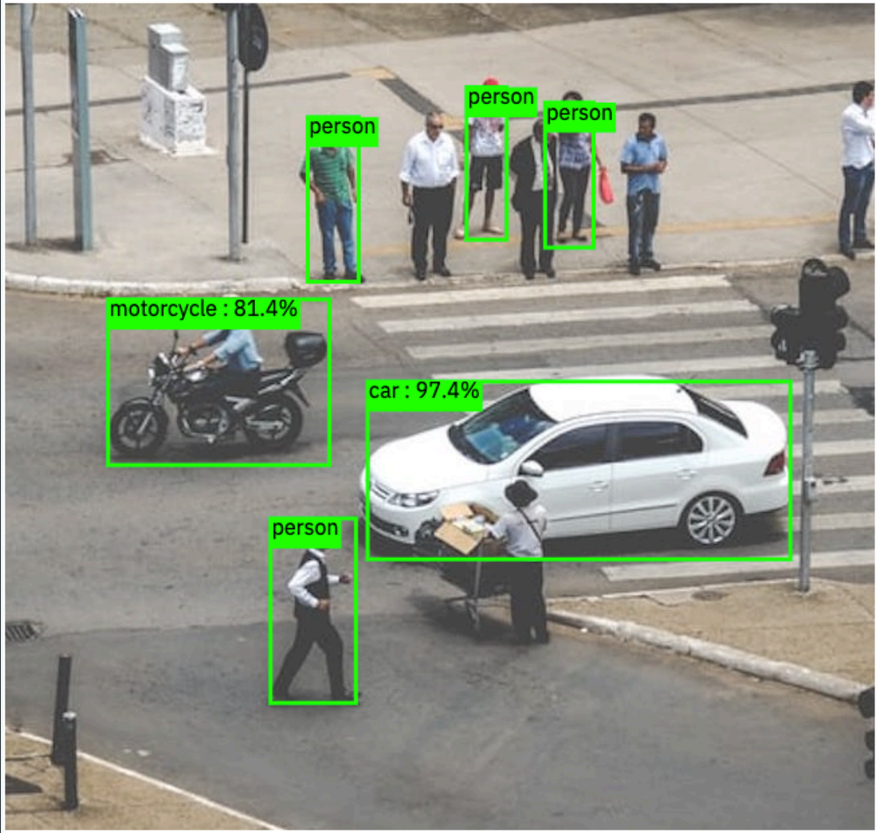
Choose File No file chosen

Submit Use your webcam

Filter detected objects ⓘ

Probability Threshold: 50%

Labels Found ⓘ



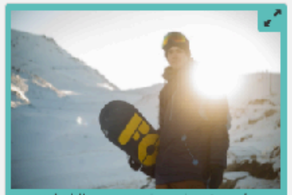
MAX Image Caption Generator

Upload A New Image ⓘ

Choose Files No file chosen

Submit

Deselect All Select All Delete Uploaded Images


a man holding a snowboard on top of a snow covered slope .


a person riding a snowboard down a snow covered slope .


a man riding skis down a snow covered slope .


a person riding a motorcycle on a race track

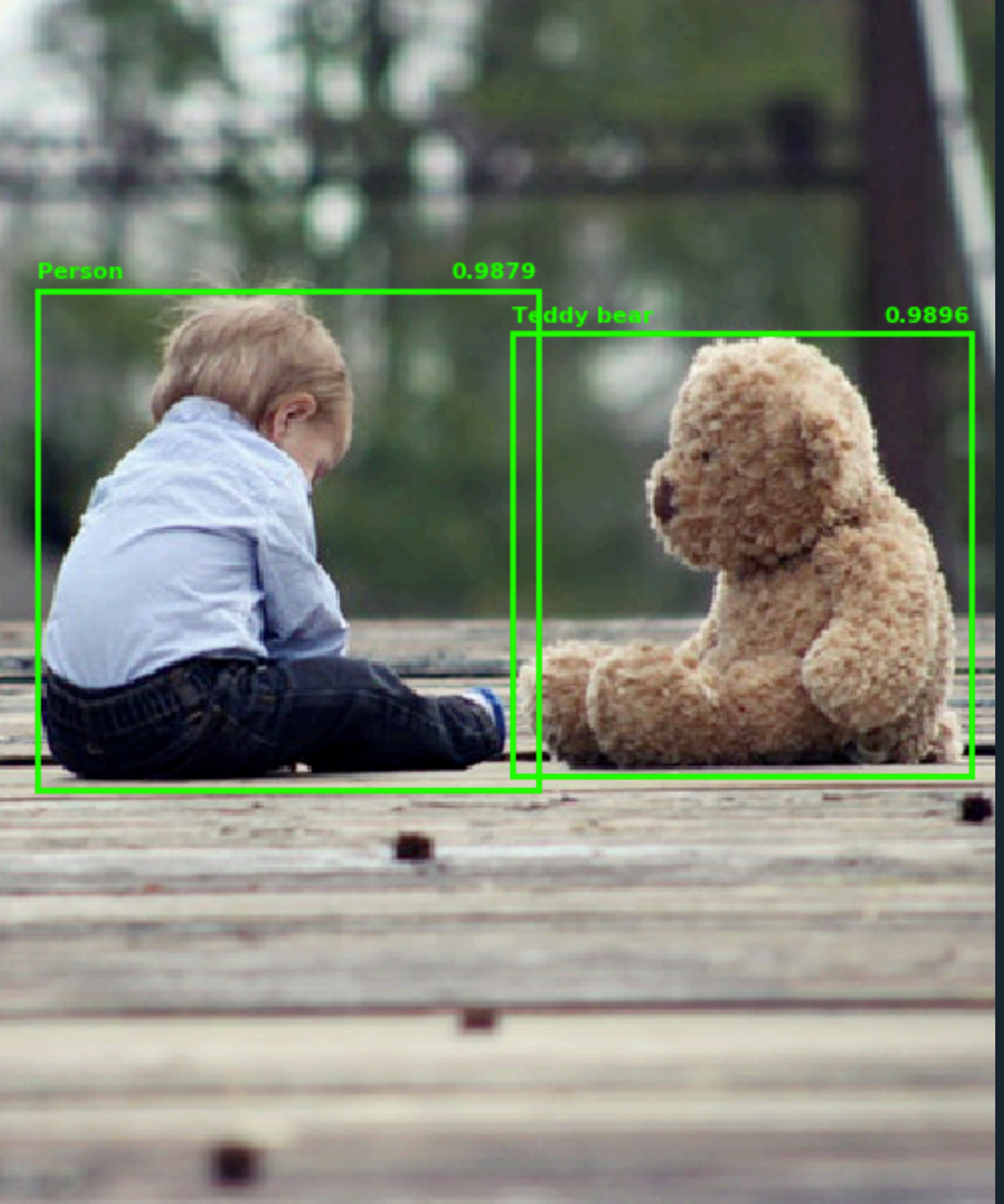

a tennis racket with a tennis ball in it


a group of baseball players standing on top of a field .


a crowd of people watching a tennis match .


a group of young men playing a game of basketball .


a female tennis player in action on the court .



OBJECT DETECTOR

Localize and **identify** multiple objects in a single image

Model Asset eXchange

Free, deployable, and trainable code. A place for developers to find and use free and open source deep learning models.

Try the tutorial

Join the community

Featured Deployable Trainable

Model | Deployable

Toxic Comment Classifier

Detect 6 types of toxicity in user comments

Jun 04, 2019

Model | Deployable, Trainable

Text Sentiment Classifier

Detect the sentiment captured in short pieces of text

Mar 29, 2019

Model | Deployable, Trainable

Image Segmenter

Identify objects in an image, additionally assigning each pi of the image to a parti object.

Model | Deployable, Trainable

Object Detector

Localize and identify multiple objects in a single image.

Sep 21, 2018

Model | Deplo

dio Classifier

Identify sounds in short audio clips.

Sep 21, 2018

Model | Deployable

Image Caption Generator

Generate captions that describe the contents of images.

Sep 21, 2018

ibm.biz/model-exchange

Model | Deployable, Trainable

Object Detector

Localize and identify multiple objects in a single image.

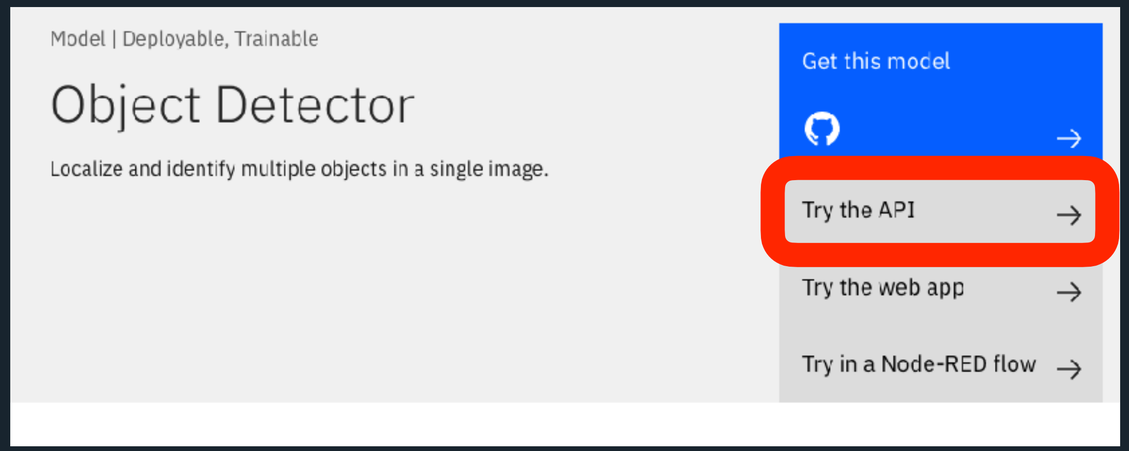
Get this model

Try the API

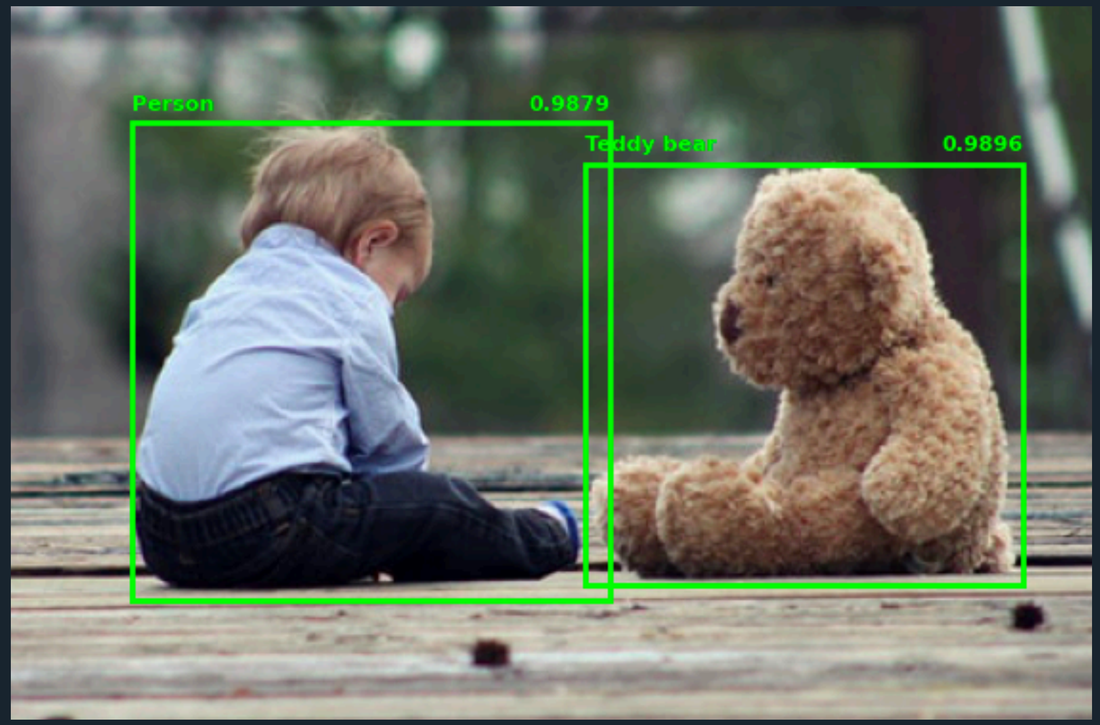
Try the web app

Try in a Node-RED flow

Access the API via Python



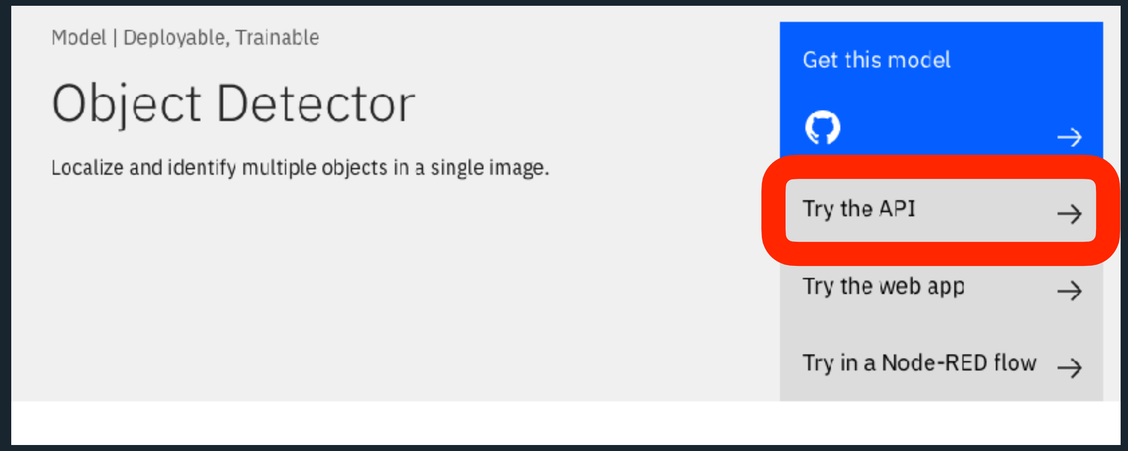

```
# Model
url = 'http://max-object-detector.codait-prod-41208c73af8fca213512856c7a09
db52-0000.us-east.containers.appdomain.cloud/'
model_endpoint =  'model/predict'
complete_url = url + model_endpoint

# Upload an image to the MAX model's rest API
path_to_input_image = 'baby-bear.jpg'

with open(path_to_input_image, 'rb') as file:
    file_form = {'image': (path_to_input_image, file, 'image/jpeg')}
    # Post the image to the rest API using the requests library
    r = requests.post(url=complete_url, files=file_form)
    # Return the JSON
    response = r.json()

IPython.display.Image(path_to_input_image, width = 450)
```



Try yourself here:

ibm.biz/max-notebook

Access the API via R




```r
library(dplyr)
library(httr)

# Endpoint
endpoint <- 'http://max-object-detector.codait-prod-41208c73af8fca213512856c7a09db52-0000.us-east.containers.appdomain.cloud/'
# endpoint <- 'http://localhost:5000'  # if running docker locally or docker hub

object_detector <- function(path_to_img, endpoint) {
  model_endpoint <- paste0(endpoint, 'model/predict')  # Model endpoint
  # POST
  response <- httr::POST(url = model_endpoint,
                         body = list(image = upload_file(path_to_img,
                                                         type = "image/jpeg")),
                         encode = c("multipart")
  ) %>% content()
  response$predictions
}

# Get the image file from GH
download.file(url = "http://github.com/IBM/MAX-Object-Detector/blob/master/samples/baby-bear.jpg?raw=true",
              'baby-bear.jpg', mode = 'wb')


object_detector("baby-bear.jpg", endpoint)
```

Access the API via Swagger

MAX Object Detector 1.4.0

[Base URL: /]
http://max-object-detector.codait-prod-41208c73af8fca213512856c7a09db52-0000.us-east.containers.appdomain.cloud/swagger.json

Localize and identify multiple objects in a single image.

model Model information and inference operations

- GET `/model/labels` Return the list of labels that can be predicted by the model
- GET `/model/metadata` Return the metadata associated with the model
- POST `/model/predict` Make a prediction given input data

POST `/model/predict` Make a prediction given input data

Parameters — Cancel

Name	Description
image * required file *(formData)*	An image file (encoded as PNG or JPG/JPEG) Choose File traffic.jpeg
threshold number *(query)*	Probability threshold for including a detected object in the response in the range [0, 1] (default: 0.7). Lowering the threshold includes objects the model is less certain about. 0.7

Execute | Clear

Responses — Response content type: application/json

Curl

```
curl -X POST "http://max-object-detector.codait-prod-41208c73af8fca213512856c7a09db52-0000.us-east.containers.appdomain.cloud/model/predict?threshold=0.7" -H "accept: application/json" -H "Content-Type: multipart/form-data" -F "image=@traffic.jpeg;type=image/jpeg"
```

Request URL

```
http://max-object-detector.codait-prod-41208c73af8fca213512856c7a09db52-0000.us-east.containers.appdomain.cloud/model/predict?threshold=0.7
```

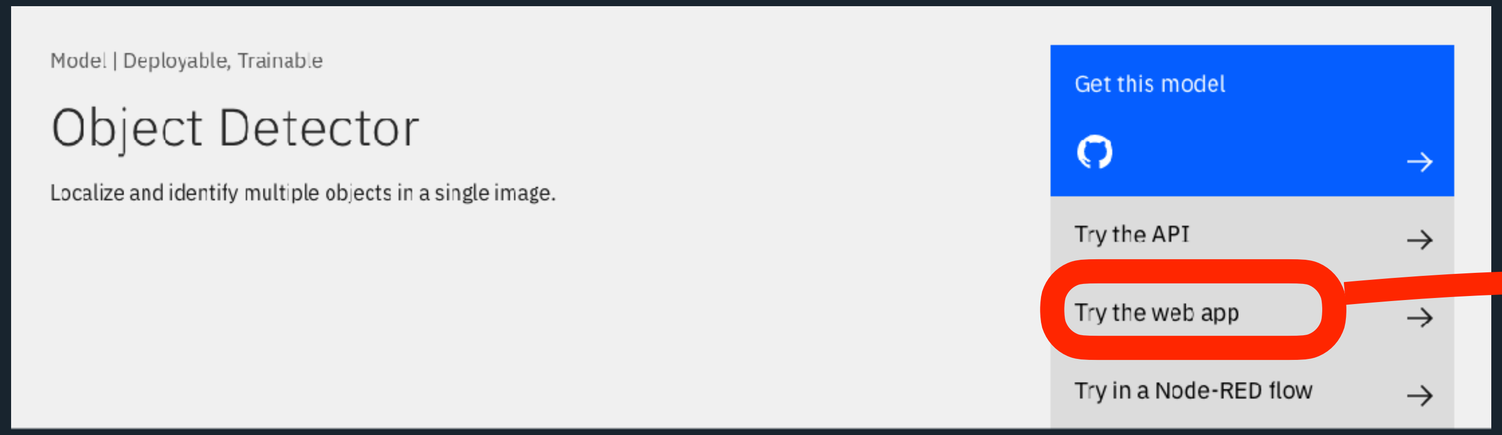
Server response

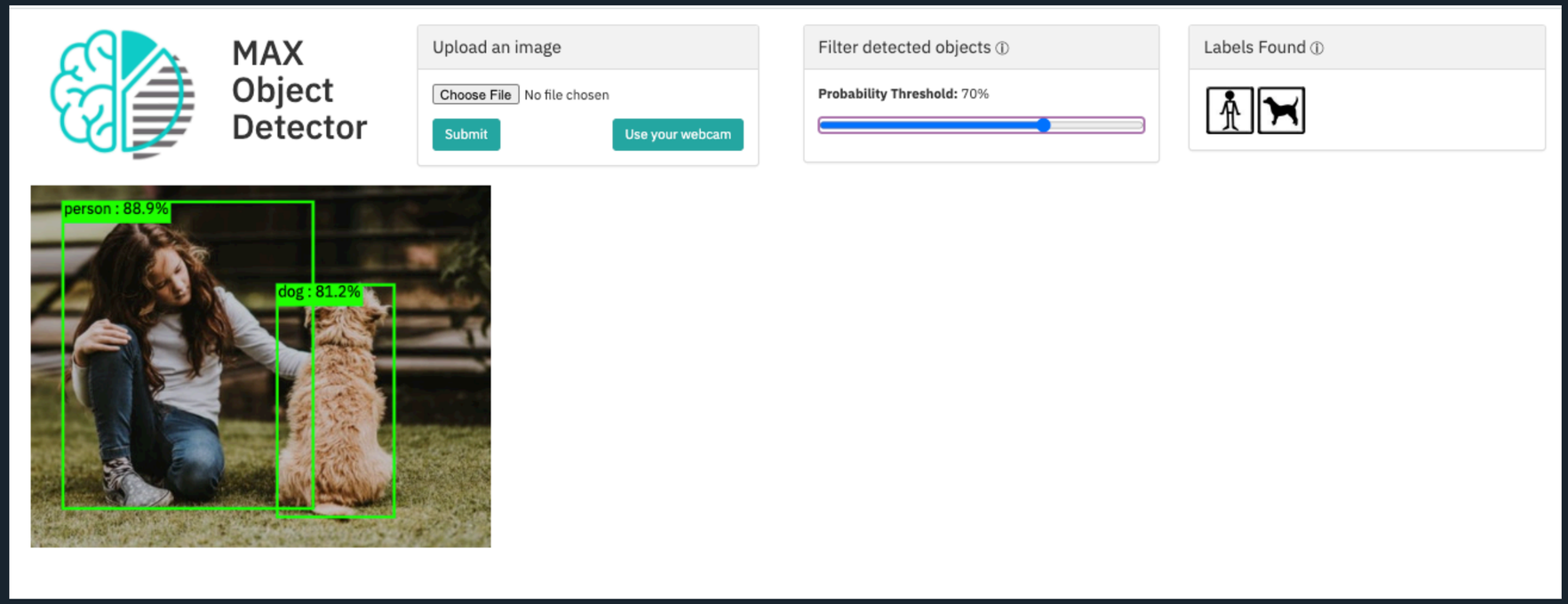
Code: 200 — Details

Response body


```
{
  "status": "ok",
  "predictions": [
    {
      "label_id": "3",
      "label": "car",
      "probability": 0.9741619627685547,
      "detection_box": [
        0.45756405559196472,
        0.41685003042222107,
        0.6725007891654968,
        0.9029390215873718
      ]
    },
    {
      "label_id": "1",
      "label": "person",
      "probability": 0.8824045658111572,
      "detection_box": [
        0.6231690645217896,
        0.30522748827934265,
        0.8462619781494141,
        0.4034259617328644
      ]
    },
    {
      "label_id": "4",
      "label": "motorcycle",
      "probability": 0.8161902685165405,
      "detection_box": [
        0.3583410978317261,
        0.11843161284923553,
        0.5584568117141724,
        0.37282349338531494
      ]
    },
```

Access the API via Web App





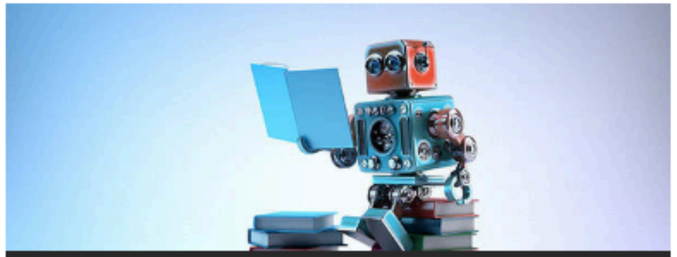
Try yourself here:

ibm.biz/object-detector-webapp

ibm.biz/max-tutorial

Examples on how to easily consume MAX models

ibm.biz/max-code-patterns



Code Pattern

Create a machine learning powered web app to answer questions

Nov 05, 2019

Code Pattern

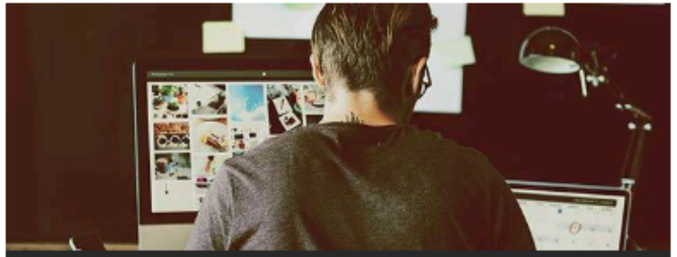
Build a web app that recognizes yoga poses using a model from the Model Asset Exchange

Oct 03, 2019

Code Pattern

Use your arms to make music

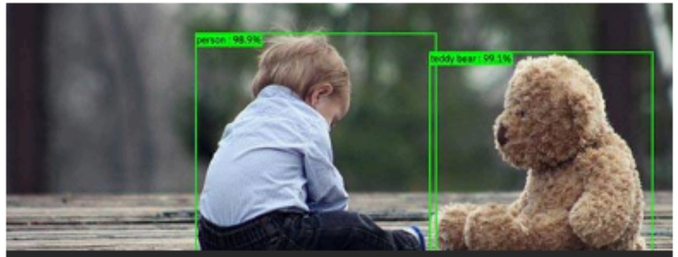
Apr 22, 2019



Code Pattern

Create a web app to interact with machine learning generated image captions

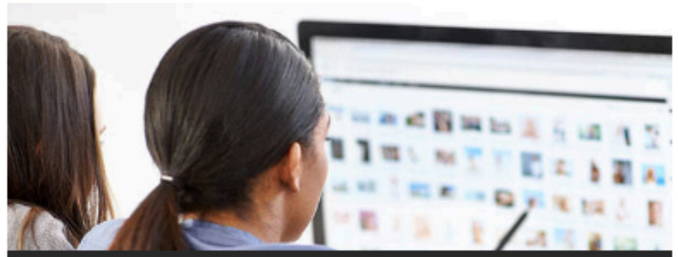
Mar 28, 2019



Code Pattern

Create a web app to visually interact with objects detected using machine learning

Mar 28, 2019



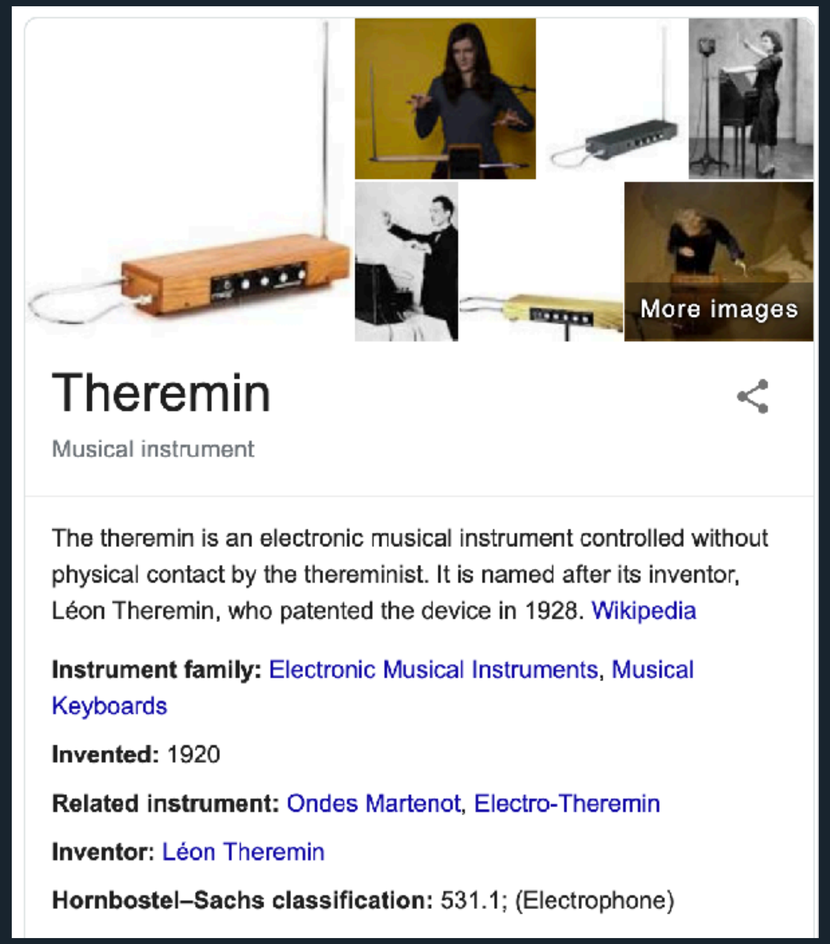
Code Pattern

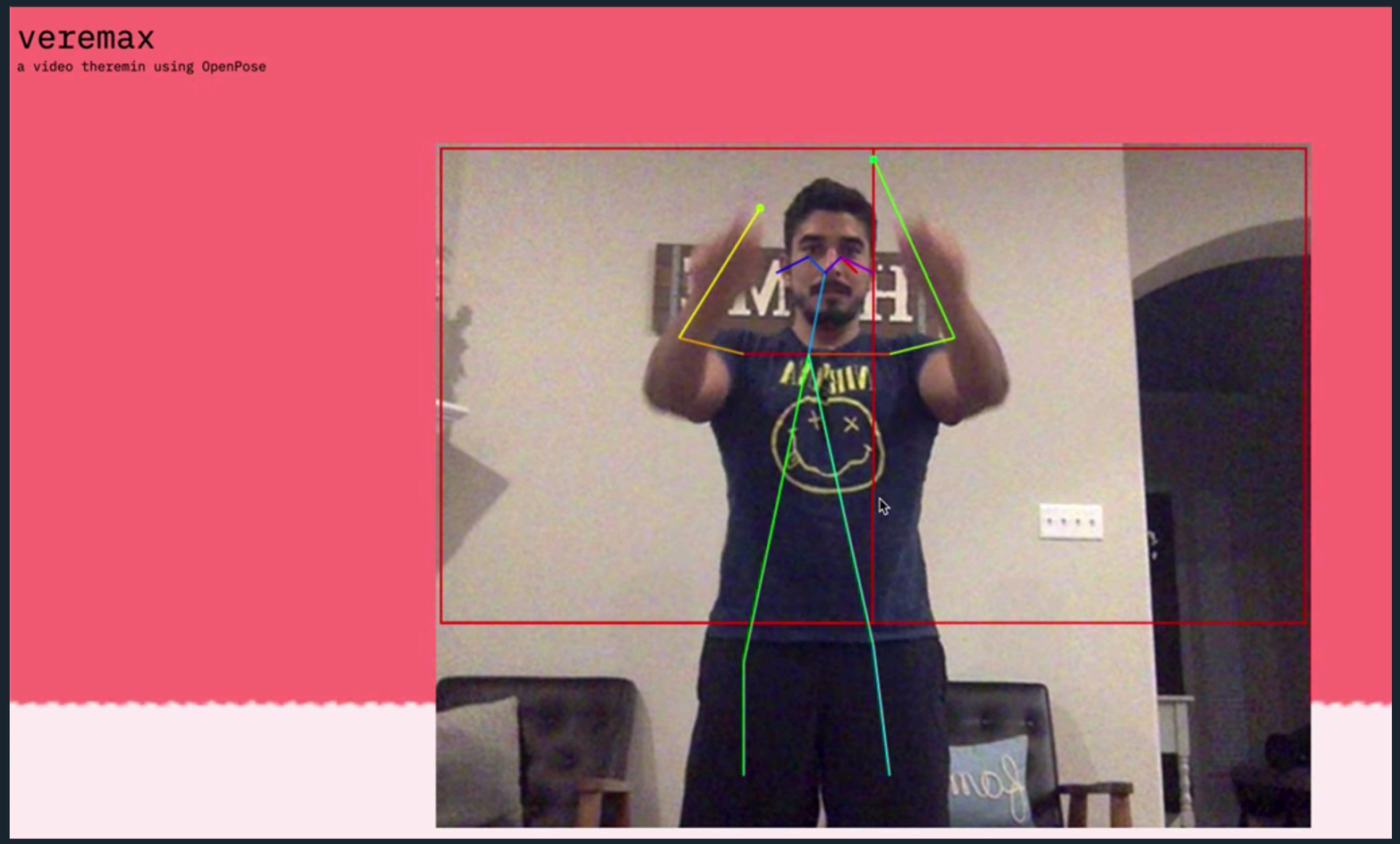
Deploy a deep learning-powered 'Magic cropping tool'

Mar 28, 2019

Use your arms to make music

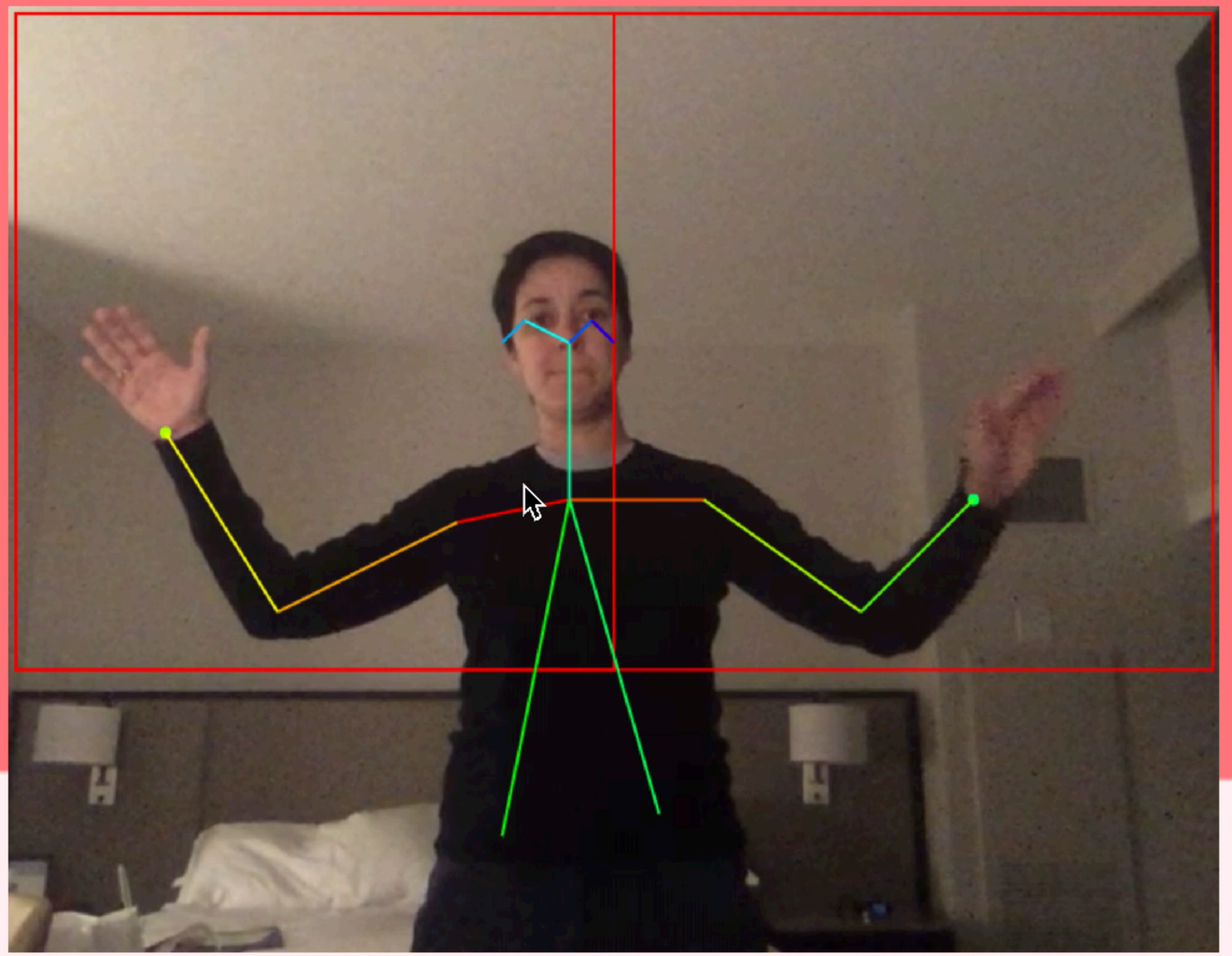
Create music with your arms using the Model Asset eXchange (MAX) **human pose estimator** model and **TensorFlow**





veremax

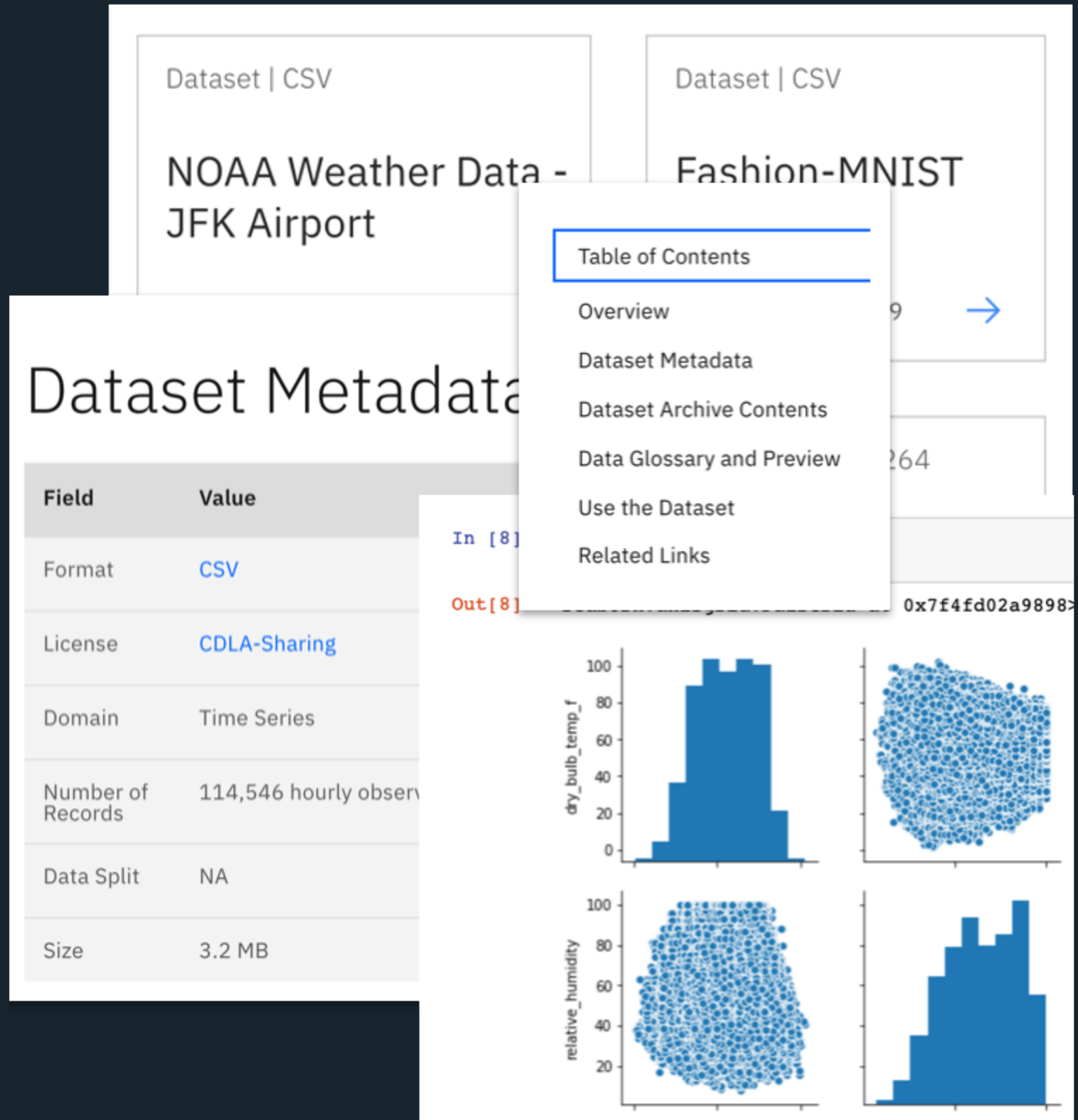
a video theremin using OpenPose



powered by the MAX Human Pose Estimator | made possible with TensorFlow.js | learn more about the code

Data Asset eXchange (DAX)

- Curated repository for **open** datasets from IBM Research and third-parties
- Published under data friendly licenses
- Standardized dataset formats and metadata
- Data sets include starter **notebooks** (cleansing, data exploration, analysis)



ibm.biz/data-exchange

CDLA-Sharing | CSV

NOAA Weather Data – JFK Airport

Local climatological data originally collected at JFK airport.

Save Like

Get this dataset →

Run dataset notebooks →

Preview the data & notebooks →

NOAA Weather Data – JFK Airport

Dataset Metadata | Dataset Preview | Dataset Glossary

Format	CSV
License	CDLA-Sharing
Domain	Time Series
Number of Records	114,546 hourly observations
Data Split	NA
Size	3.2 MB
Data Origin	National Oceanic and Atmospheric Administration (NOAA)
Dataset Version	Version 2 – September 12, 2019 Version 1 – July 16, 2019
Dataset Coverage	Location: New York City Dates: 2010-01-01 through 2018-07-27 Note: To download raw data from NOAA for a different region or date span, follow the steps outlined in the data archive's README.txt.
	Agriculture Detect unseasonal temperature change and alert farmers about potential damage to plants. Energy Regulate solar cell charging hours based on weather type condition and temperature. Regulate wind turbine operation based on wind speed and wind direction. Generate energy demand alerts based on

NOAA Weather Data – JFK Airport

Part 1 - Data Cleaning | Part 2 - Data Analysis | Part 3 - Time Series Forecasting

In [1]:
```
# @hidden_cell
# The project token is an authorization token that is used to access project resources like data sources, connections, and used by platform APIs.
from project_lib import Project
project = Project(project_id='...', project_access_token='...')
```

Cleaning NOAA Weather Data of JFK Airport (New York)

This notebook relates to the NOAA Weather Dataset - JFK Airport (New York). The dataset contains 114,546 hourly observations of 12 local climatological variables (such as temperature and wind speed) collected at JFK airport. This dataset can be obtained for free from the IBM Developer Data Asset Exchange.

In this notebook, we clean the raw dataset by:

- removing redundant columns and preserving only key numeric columns
- converting and cleaning data where required
- creating a fixed time interval between observations (this aids with later time-series analysis)
- filling missing values
- encoding certain weather features

Table of Contents:

Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application

IBM and LFAI move forward on trustworthy and responsible AI

IBM donates Trusted AI toolkits to the Linux Foundation AI

Did anyone tamper with it?

ROBUSTNESS

Adversarial Robustness 360
↳ (ART)

DEMO: art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360
↳ (AIF360)

DEMO: aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

AI Explainability 360
↳ (AIX360)

DEMO: aix360.mybluemix.net

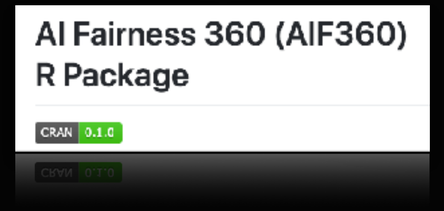
Is it accountable?

LINEAGE

AI FactSheets 360

DEMO: aifs360.mybluemix.net

Available in R too!



Follow us:

bit.ly/ai-inclusive-instagram

bit.ly/ai-inclusive-twitter

ai-inclusive.org

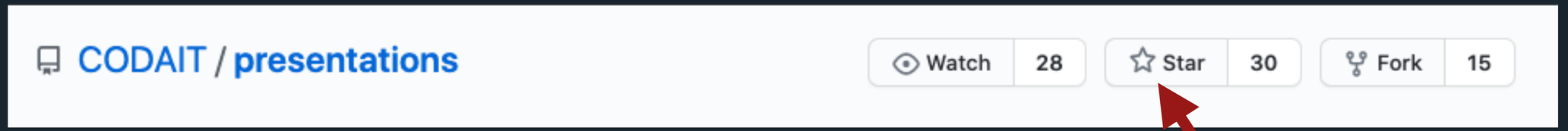
Resources on AI, DS, ML
Events, Free Tickets and much more

Soon:

Scholarships for Data Science Courses

Thank you!

slides & materials: bit.ly/dc-thurs-ibm



Don’t forget to star the repo

ai-inclusive.org