

Data Science as a Team Sport



Gabriela de Queiroz

@gdequeiroz | linktr.ee/gdq

slides: bit.ly/kroz-talks

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**



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.



Wednesday, October 31, 2012
31 OCT

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!

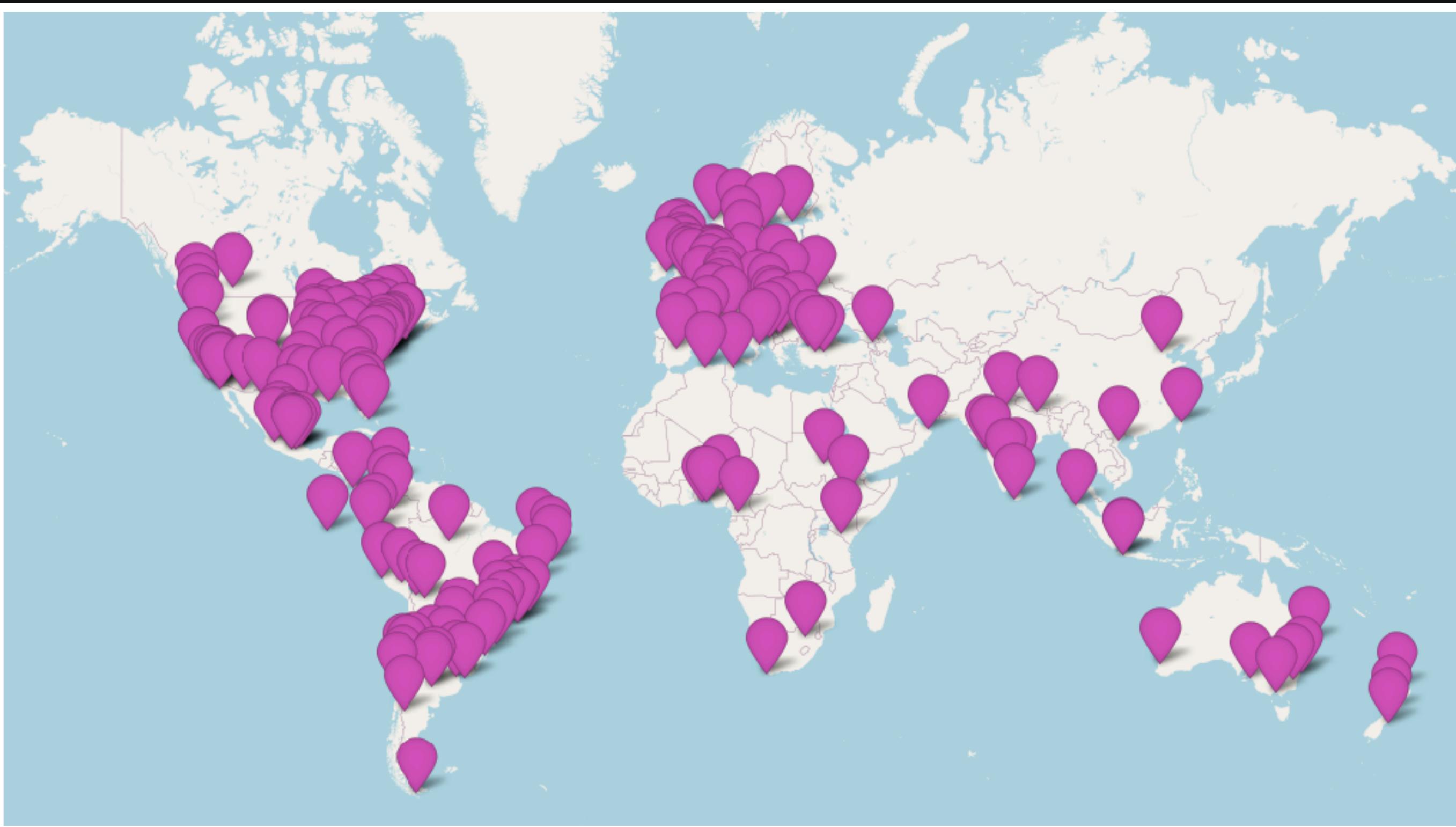


R-Ladies

rladies.org



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





AI Inclusive

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

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

- 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



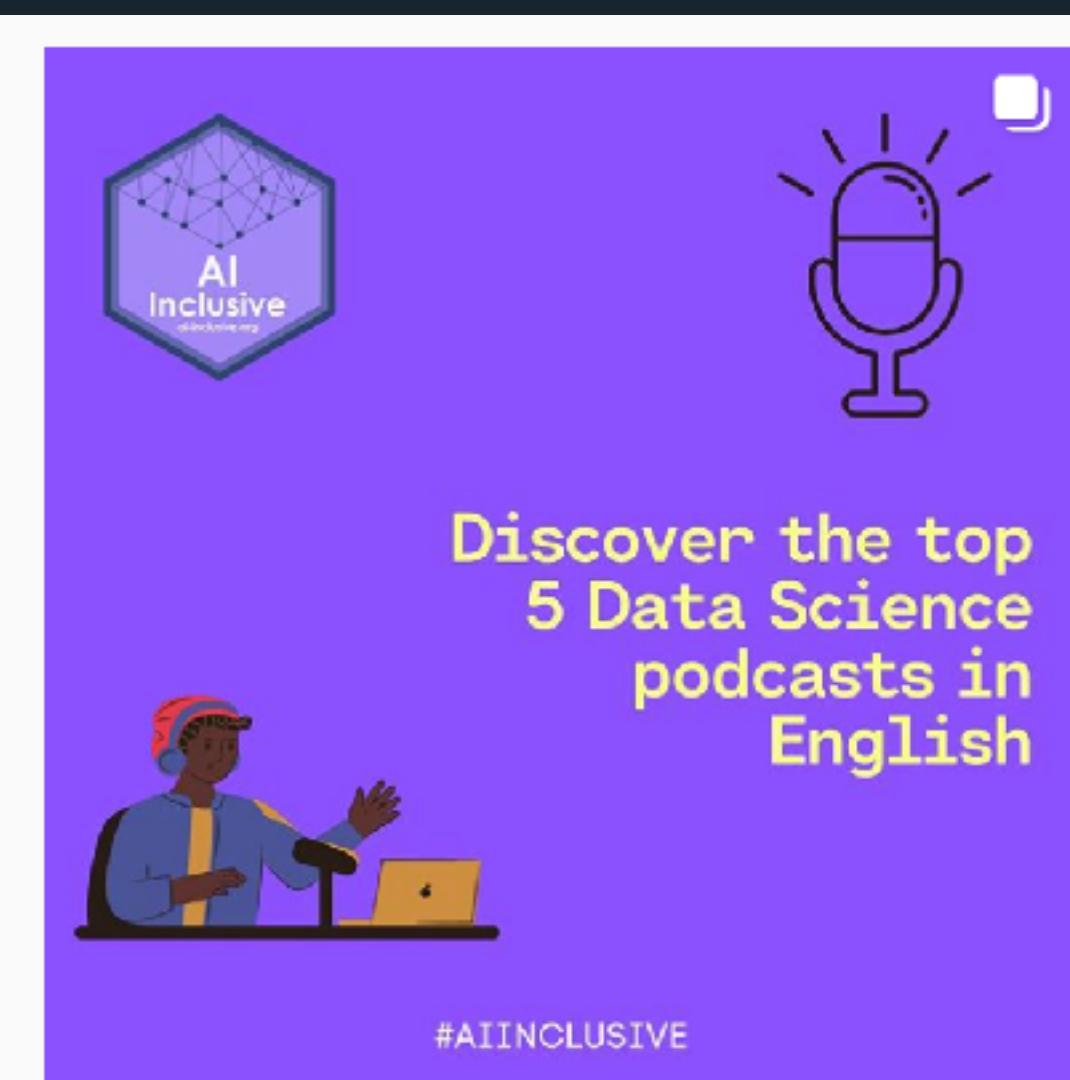
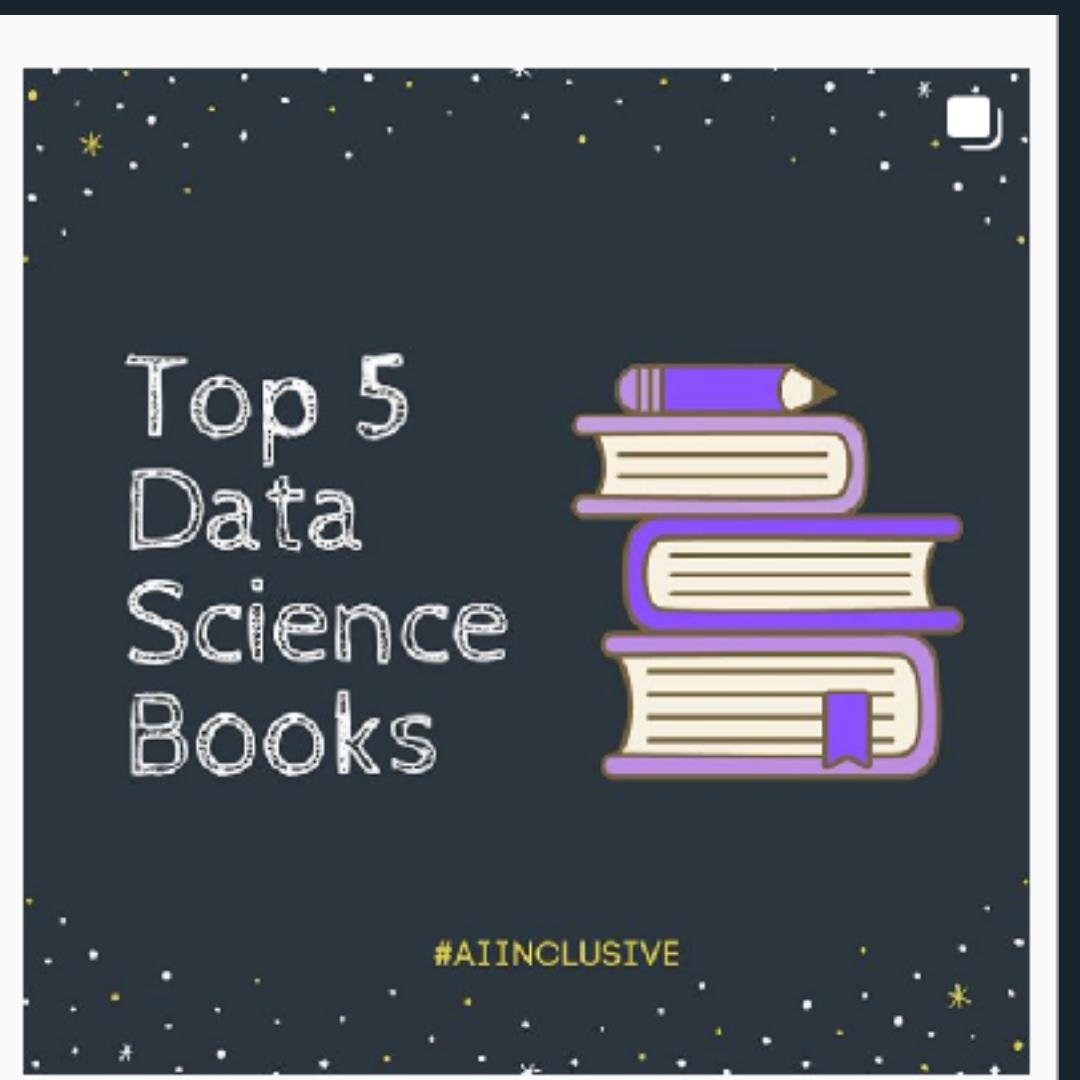
Follow US:

bit.ly/ai-inclusive-instagram

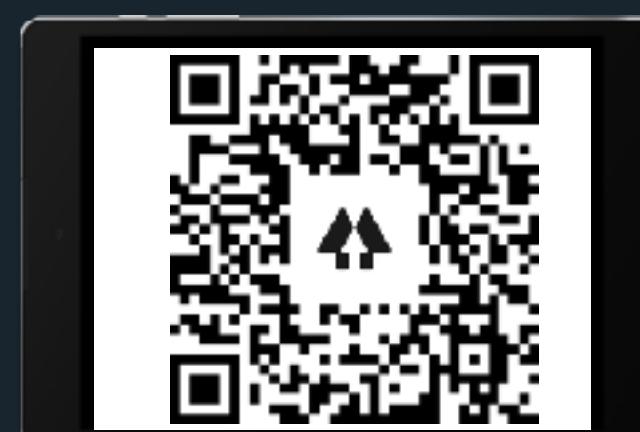
bit.ly/ai-inclusive-twitter



ai-inclusive.org



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





Dataquest & AI Inclusive Scholarship Program



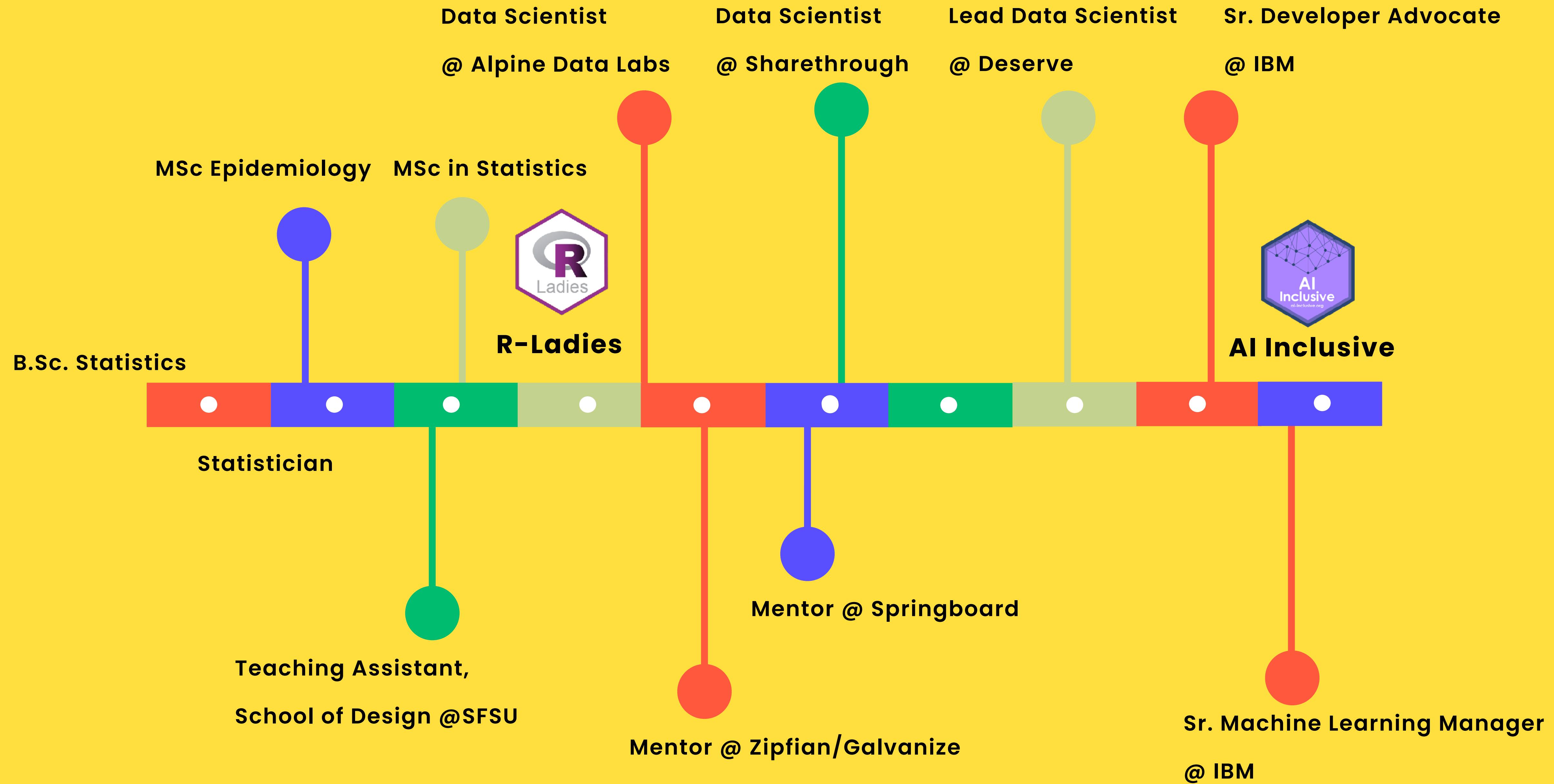
Round 1: Women and Underrepresented Genders

- You'll learn all the skills you need to land a job in Data Science
- You will have access to all courses for FREE
- Learn R, Python, and SQL 100% online!

Apply now:
dataquest.io/scholarship



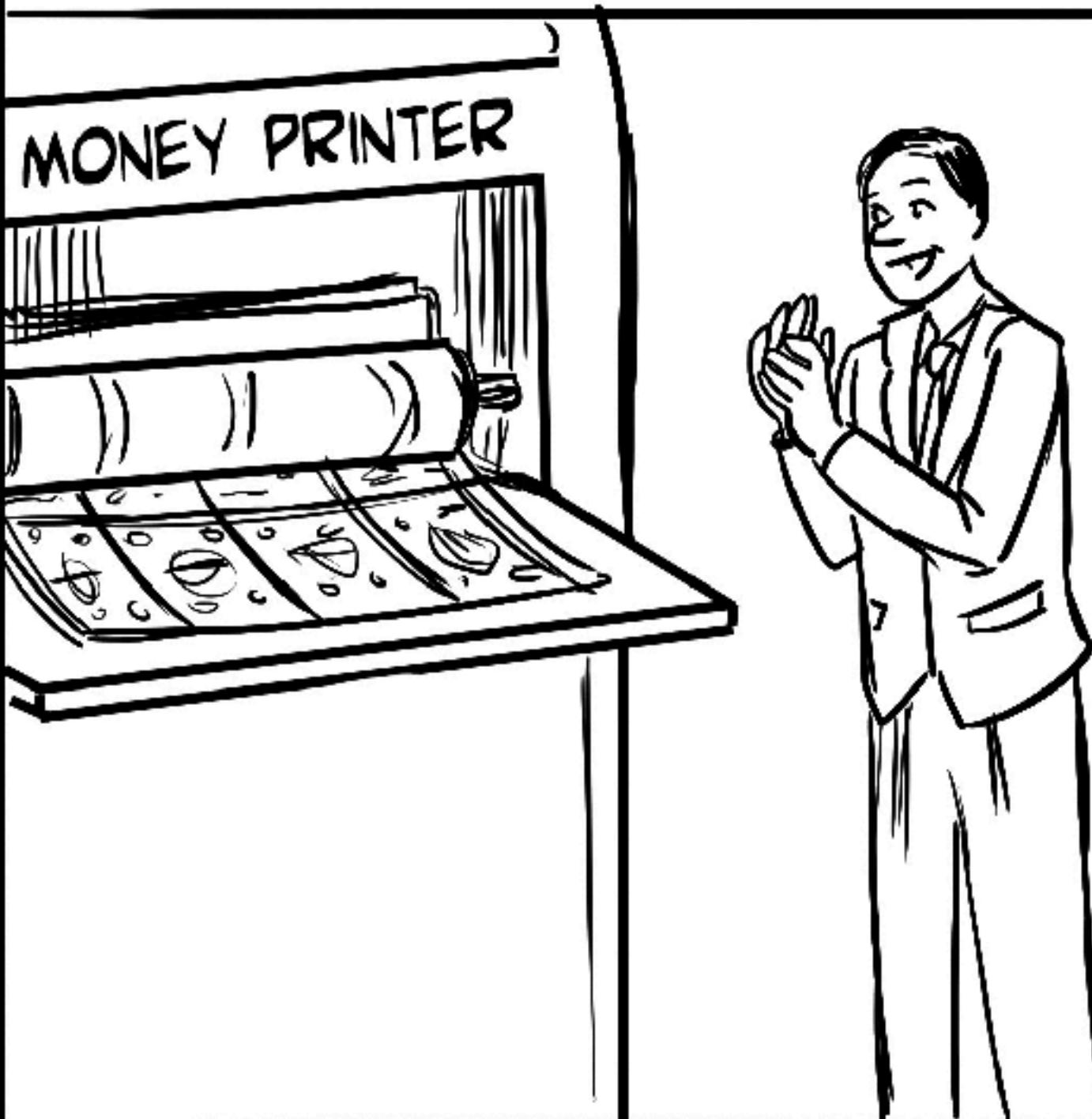
Applications close on December 18th



The Data Science Career

What is Data Science?

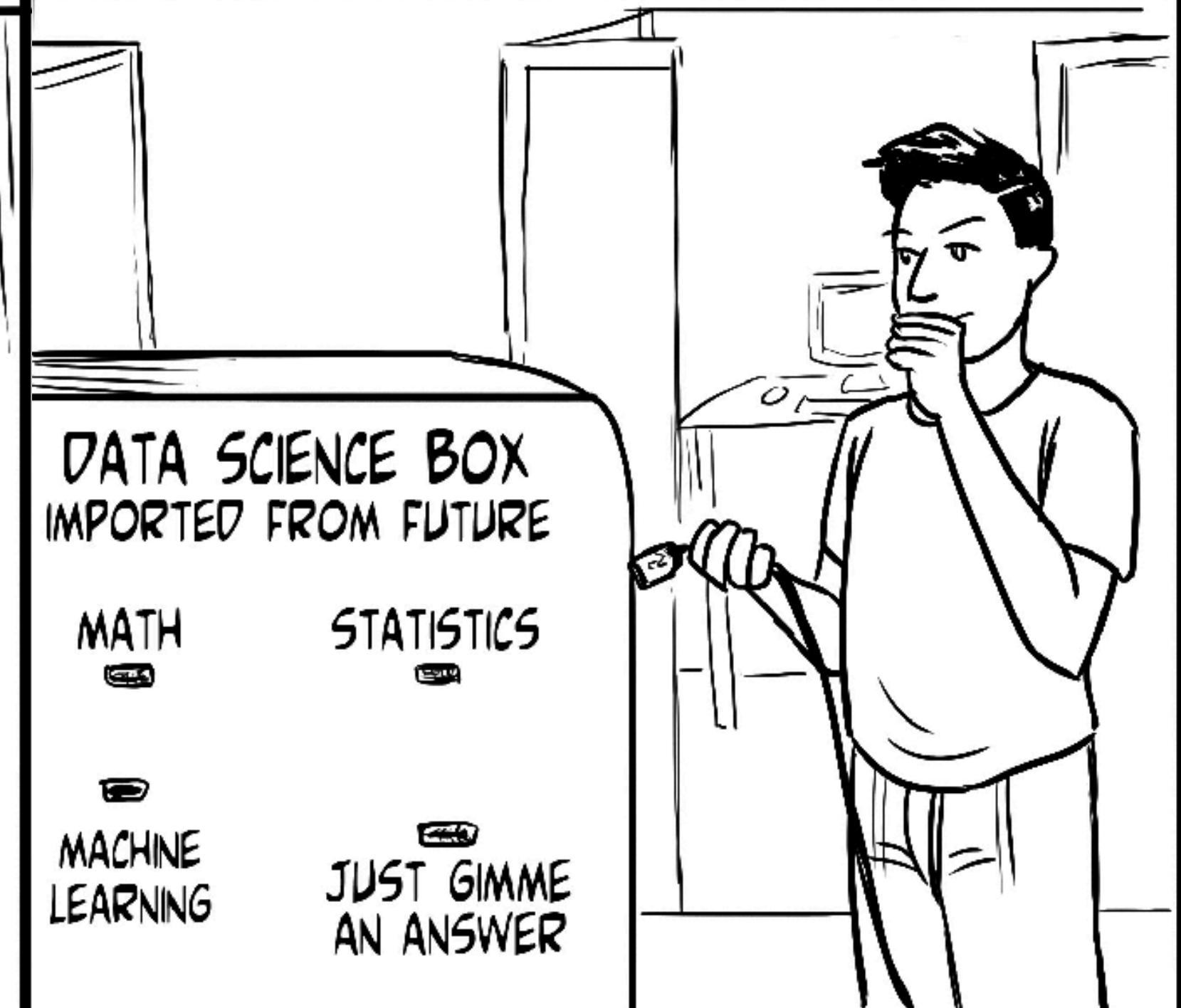
WHAT MY BOSS THINKS DATA SCIENCE IS



WHAT MY CUSTOMERS THINK DATA SCIENCE IS

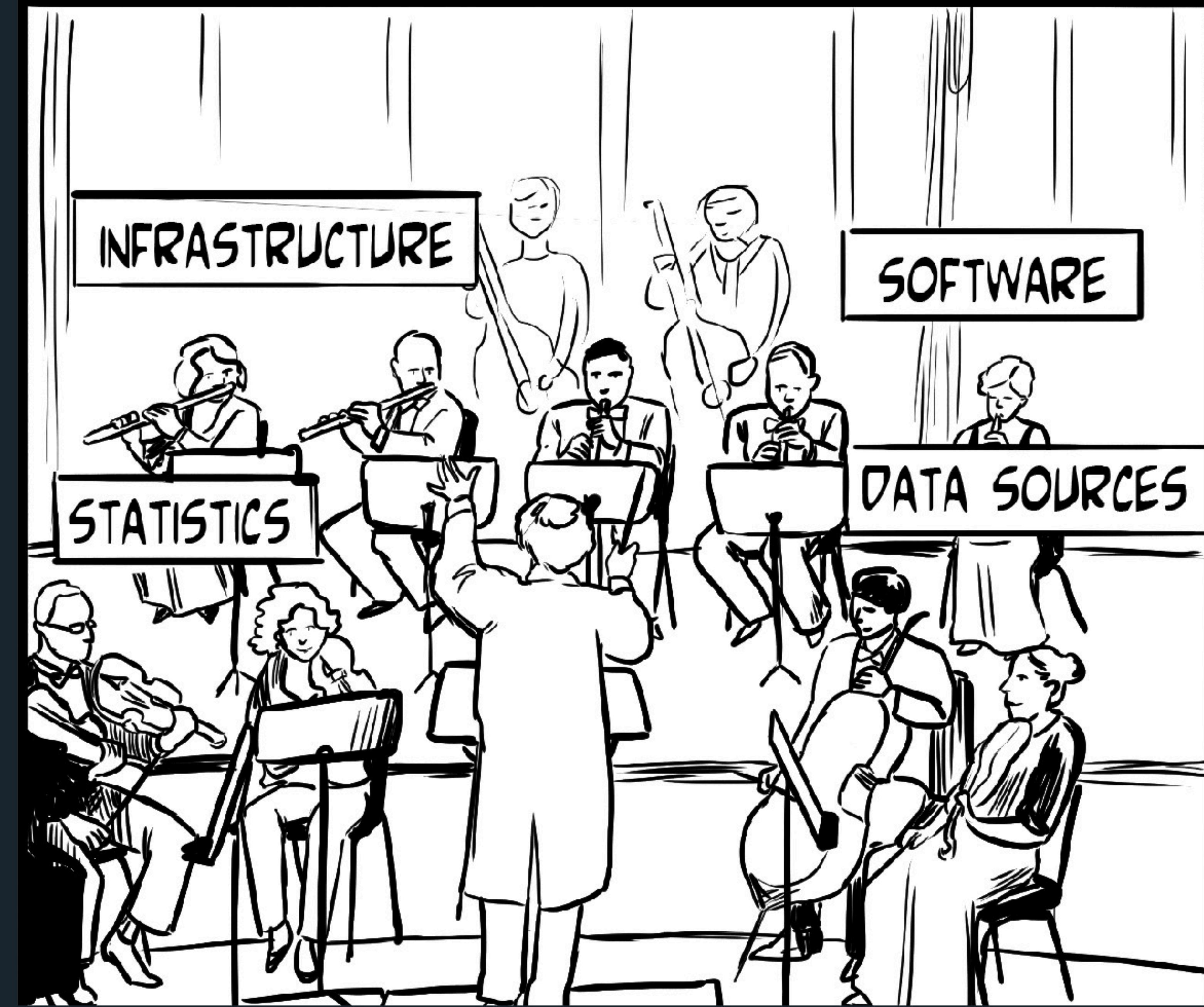


WHAT SOFTWARE ENGINEERS THINK DATA SCIENCE IS



What is Data Science?

WHAT I THINK DATA SCIENCE IS



The background of the image is a light-colored wall covered with numerous cowboy hats, arranged in a grid pattern. The hats are light brown or tan in color and have a classic western style with a prominent band and a dark, rounded crown.

Data Scientists are **adaptable** and
flexible professionals

Companies

What is the role of a data scientist?

Data Scientists can have different roles in different companies

Advanced Analytics Platform for Big Data



Empowers business users to define and participate in data science projects and gives data scientists the tools they need to create value from data

The image shows the Alpine Data platform interface and a data pipeline diagram.

Platform Interface:

- Header:** ALPINE DATA, Search bar.
- Project:** HiveExample.
- Operators Tab:** OPERATORS (selected), DATA.
- Operator Selection:** All Operators dropdown, Filter operators... button.
- Recent Operators:** Logistic Regression.
- All Operators List:**
 - Aggregation
 - Alpine Forest Classification
 - Alpine Forest Evaluator
 - Alpine Forest Regression
 - Bar Chart
 - Batch Aggregation

Data Pipeline Diagram:

```
graph LR; 1[Hive Table] --> 2[HQL Execute]; 2 --> 3[Random Sampling]; 3 --> 4[training set]; 3 --> 6[test set]; 4 --> 5[LoR- Spark]; 5 --> 6; 6 --> 5; 5 --> 7[ROC]
```

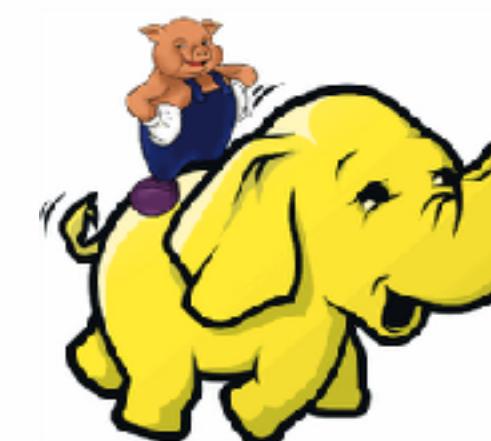
The pipeline consists of the following steps:

1. Hive Table
2. HQL Execute
3. Random Sampling (produces both training set and test set)
4. training set
5. LoR- Spark
6. test set
7. ROC

Company Numbers

- 40 Employees
- 10 Engineers
- 5 Data Scientist

Tools



Key Responsibilities:



Consultant



Write Documentation



Train Customers

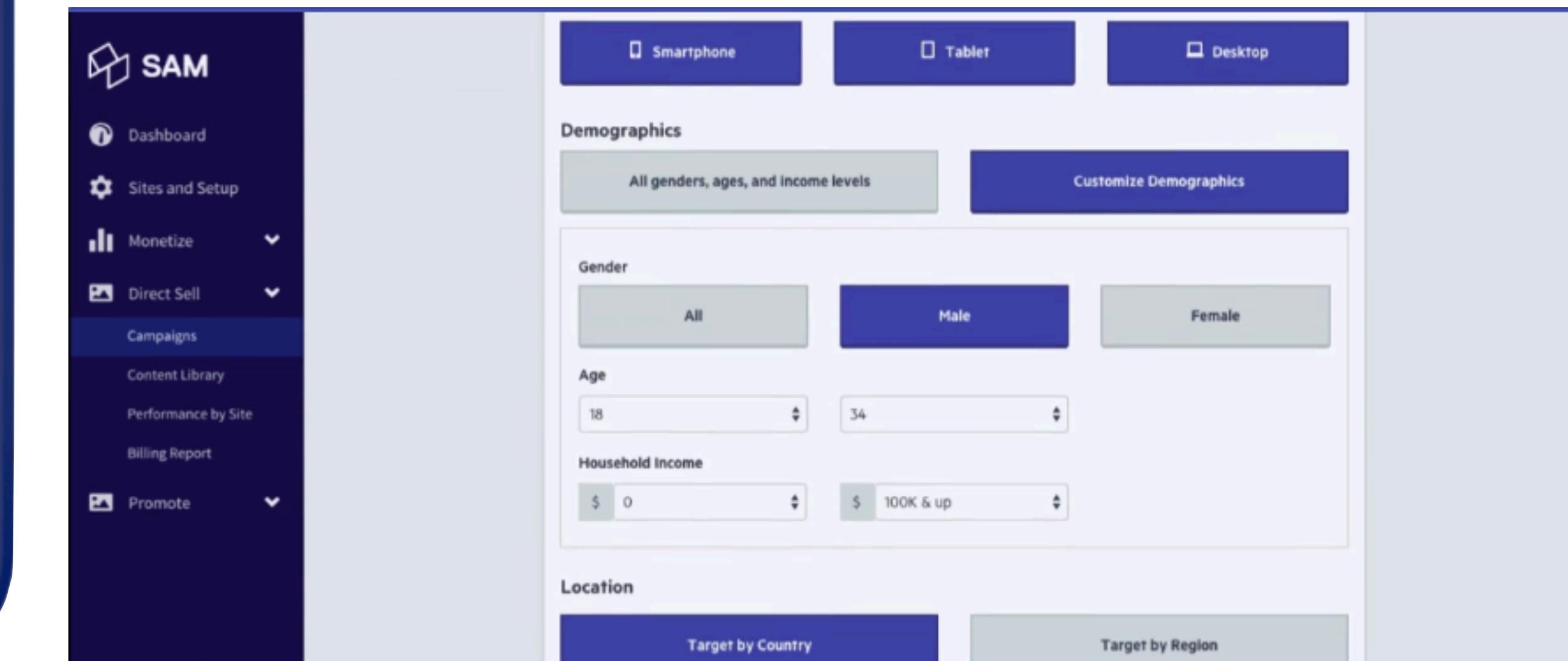
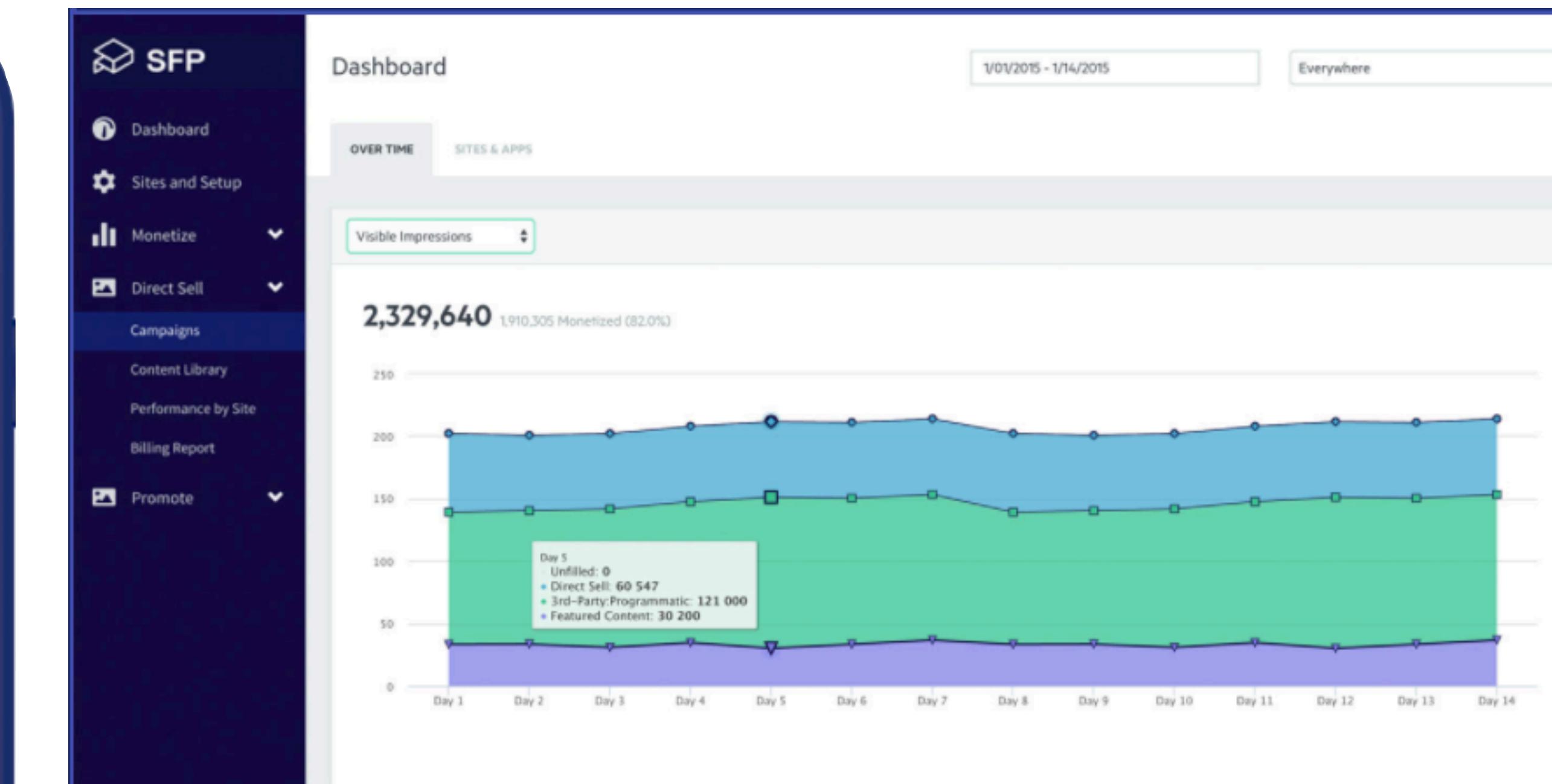
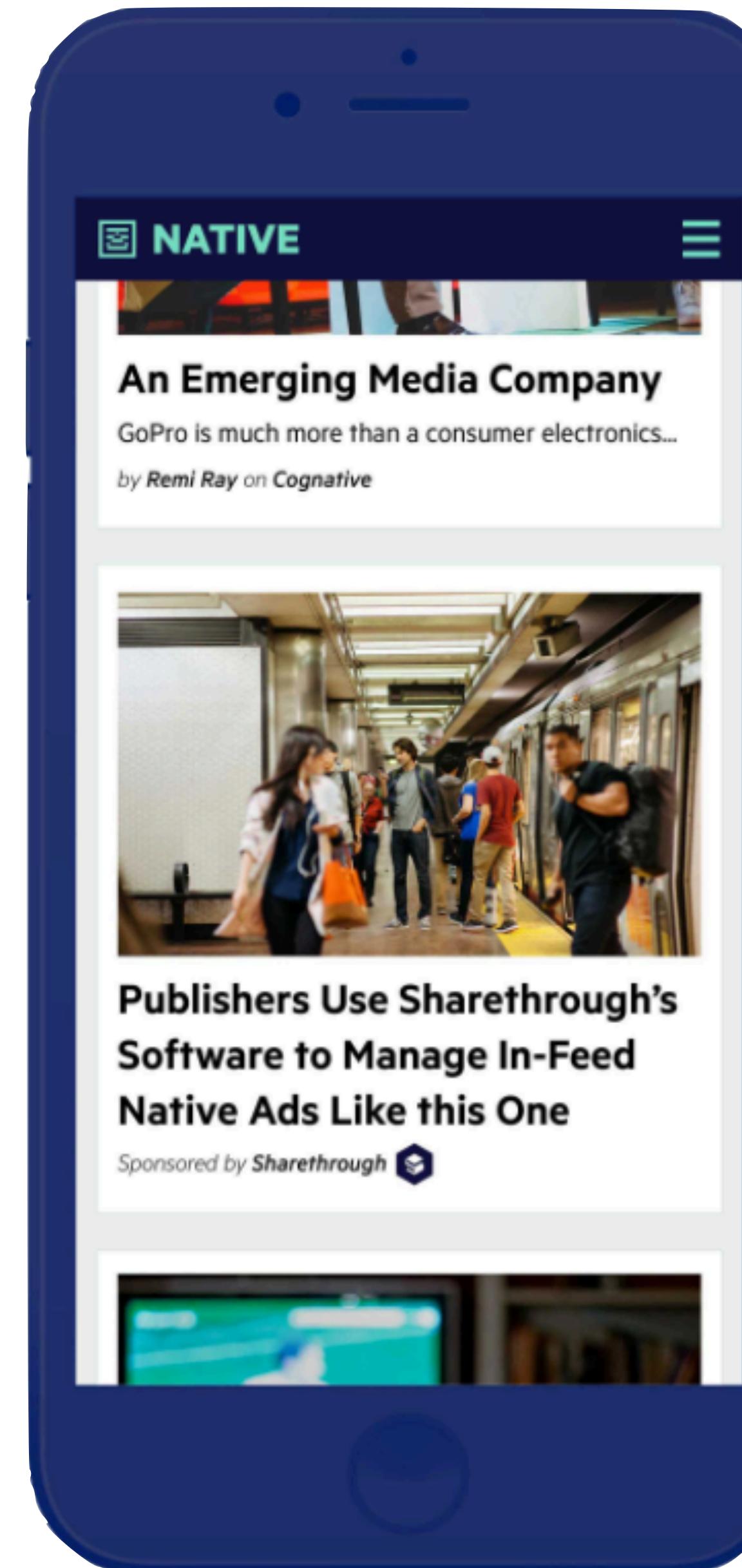


Prioritization

Native Advertising software for publishers, app developers & advertisers.



sharethrough



Company Numbers

- 150 Employees
- 20 Engineers
- 1 Data Scientist

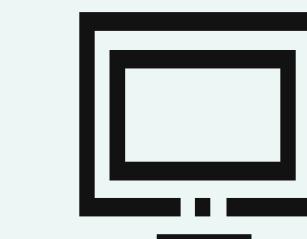
Tools



Key Responsibilities:



Infrastructure



Models / Optimization



Experimental Design



Education



Through machine learning and alternative data, Deserve is helping millennials and Gen Z's, the next wave of credit owners, gain financial independence through access to fair credit products.

Credit Card designed for Generation Z

(INCLUDING INTERNATIONAL STUDENTS)



Company Numbers

- 35 Employees
- 7 Engineers
- 4 Data Scientist

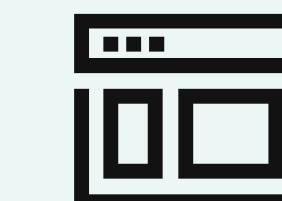
Tools



Key Responsibilities:



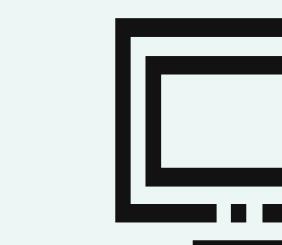
Infrastructure



Dashboard



R packages



Models

Open Source @ IBM

CODAIT

Center for Open Source Data
and AI Technologies

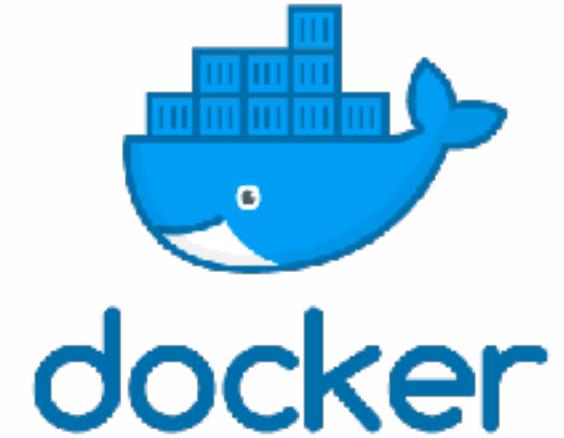


Open Source, Data & AI Technologies

Company Numbers

- +300k Employees
- ? Engineers
- ? Data Scientist

Tools



Key Responsibilities:



Senior Open Source Developer - DL/ML/AI Developer and Advocate

Key Responsibilities:

- 1) Write great code in key open source communities
- 2) Democratize AI by building tools, launching new open source projects, and improving existing ones
- 3) Gain eminence in the community by socializing your work, and speaking at events
- 4) Work with offering managers and product teams on applications
- 5) Guide and mentor clients to become self-sufficient open source developers
- 6) Be authentic; mentor others, and be open to mentoring by others
- 7) Read and comment on more code than you write; fix bugs, test cases, and documentation
- 8) Etc ...

Sr. Machine Learning Manager

	Sun 3/22	Mon 3/23	Tue 3/24	Wed 3/25	Thu 3/26	Fri 3/27	Sat 3/28
PST							
9am		9 – 10 Meeting with X Team	9 – 10 Meeting with X Team	9 – 10 Meeting with Z Team	9 – 10 Meeting with X Team	9 – 10 Sprint Retrospective	
10am		10 – 11 Project Planning Session	10 – 11 Press Interview	10 – 11 Project Planning Session	10 – 11 Corporate Event update	10 – 11 Project Planning session	
11am		11 – 12p Call with East Coast office	11 – 12:30p Presentation to executive team	11 – 12p Coaching session	11 – 12p Presentation to Mobile team	11 – 12:30p Training session	
12pm		12p – 1p Management Luncheon	12:30p – 1:30p Call with Tony	12p – 1p Lunch with CEO	12p – 1p Team Lunch	12:30p – 1:30p Conference Call	
1pm		1p – 2p 1 on 1 with Tony		1p – 2p Project Z Meeting	1p – 2p 1 on 1 with Chris	1:30p – 2:30p Project Z Meeting	
2pm		2p – 3p Review session		2p – 3p 1 on 1 with Amanda	2p – 3p 1 on 1 with Tara	2:30p – 3:30p Check in with Rich	
3pm		3p – 5p Leadership team planning session		3p – 5p Company All Hands	3p – 4p Meeting with engineering	3:30p – 5p Senior leadership checkin	
4pm					4p – 5p New project Kickoff		
5pm		5p – 6p Call with Australia office	5p – 6p Drinks with Thomas	5p – 6p Leadership Team Dinner	5p – 6p Speak at Event	5p – 6p Team happy hour	
6pm							

Responsibilities: (MANY!)

MACHINE LEARNING TEAM

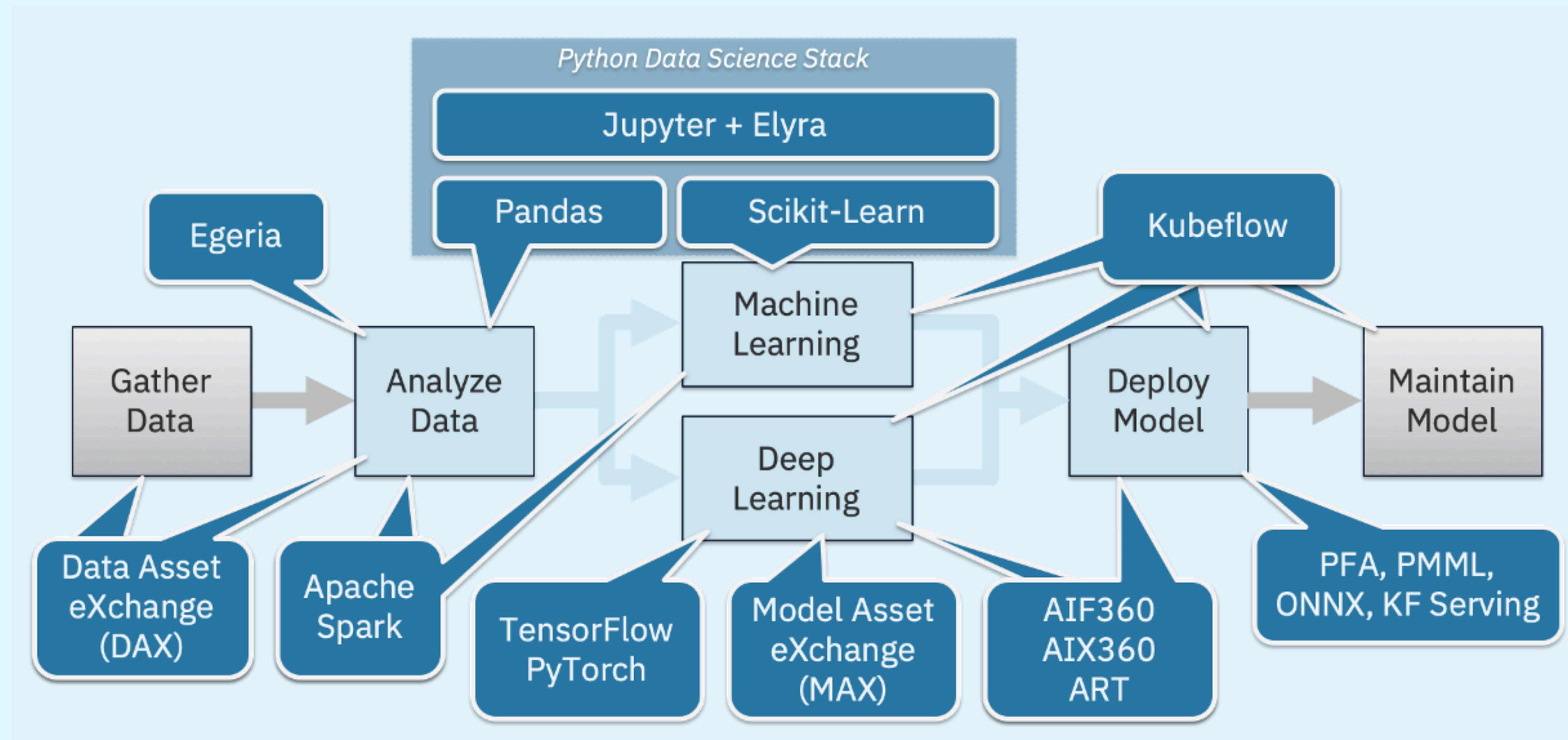
10 open source developers

Data Scientists & Software Engineers

Backgrounds in Stats, Math, Engineering, Computer Science,
Finance.

We build tools to make AI accessible and available to everybody

(codait.org)





Open Source @ IBM

Some Projects

Model Asset eXchange



ibm.biz/model-exchange

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 Segmente

Identify objects in an image, additionally assigning each pixel of the image to a particular object.

Sep 21, 2018



Model | Deployable, Trainable

Object Detector

Localize and identify multiple objects in a single image.

Sep 21, 2018



Model | Deployable

Audio 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



Data Asset eXchange



ibm.biz/data-exchange

Data Asset eXchange

Explore useful and relevant data sets for enterprise data science

[Learn More](#)



[What's New](#)



[Get Involved](#)



Dataset | CSV

NOAA Weather Data - JFK Airport

June 30, 2020

Dataset | ICB format

Groningen Meaning Bank - Modified

May 14, 2020

Dataset | CSV

Fashion-MNIST

September 12, 2019



Dataset | JPG, JSON

PubLayNet

October 25, 2019

Dataset | WAV

TensorFlow Speech Commands

March 17, 2020

Dataset | PNG, JSON

PubTabNet

July 20, 2020



Model Asset eXchange (MAX)

Place for developers/data scientists to find and use
free and **open source** deep learning models



ibm.biz/model-exchange

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 pixel of the image to a particular object.

Sep 21, 2018

Model | Deployable, Trainable

Object Detector

Localize and identify multiple objects in a single image.

Sep 21, 2018

Model | Deployable

Audio 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

[View all models](#)



MAX Object Detector

Upload an image

Choose File No file chosen

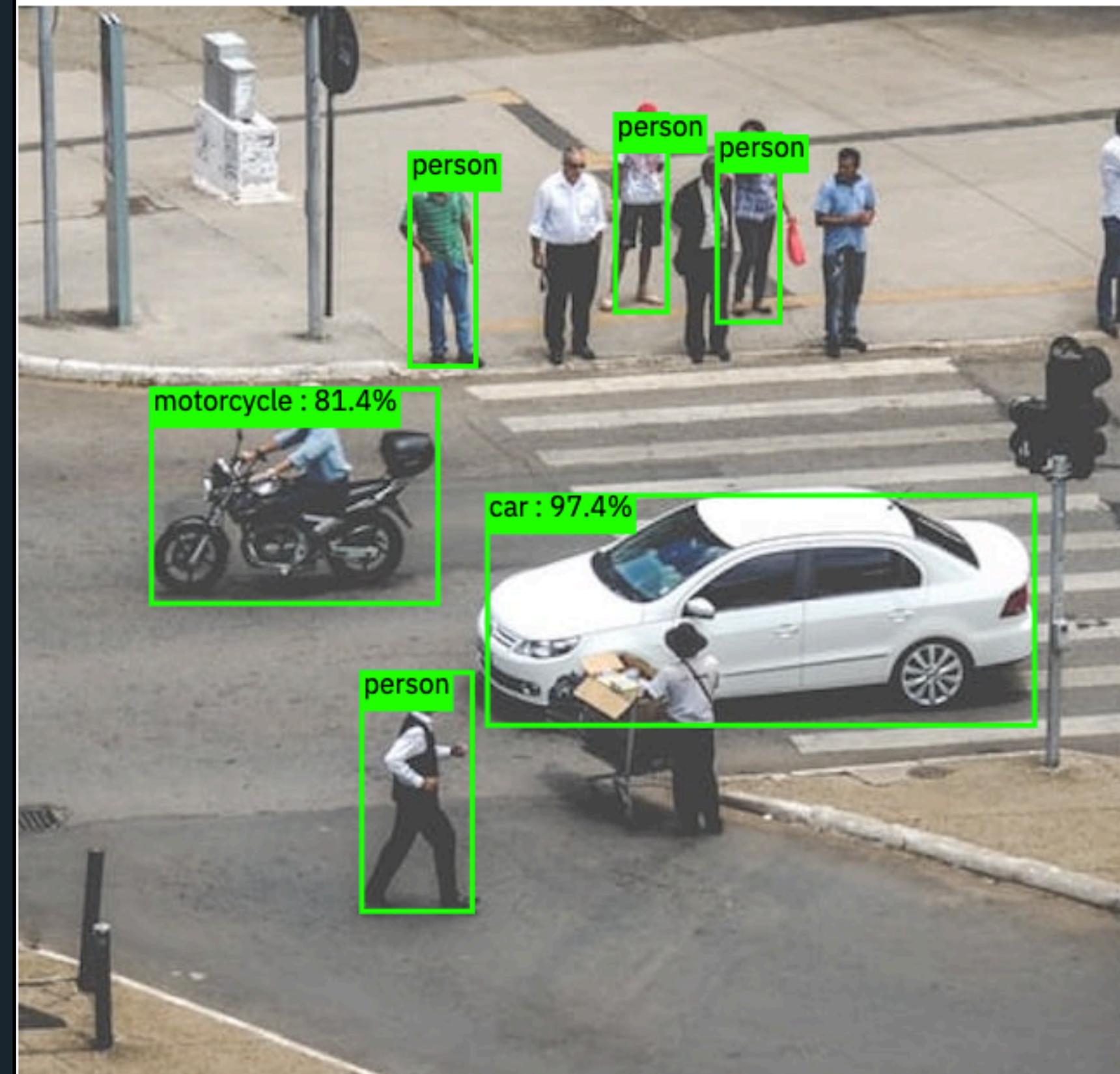
Submit

Use your webcam

Filter detected objects ⓘ

Probability Threshold: 50%

Labels Found ⓘ



MAX Image Caption Generator

Upload A New Image ⓘ

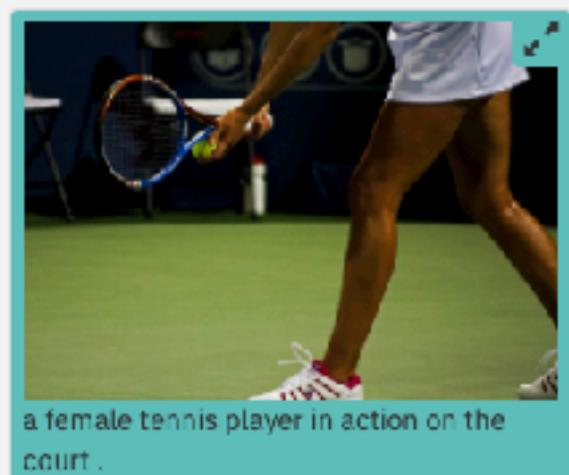
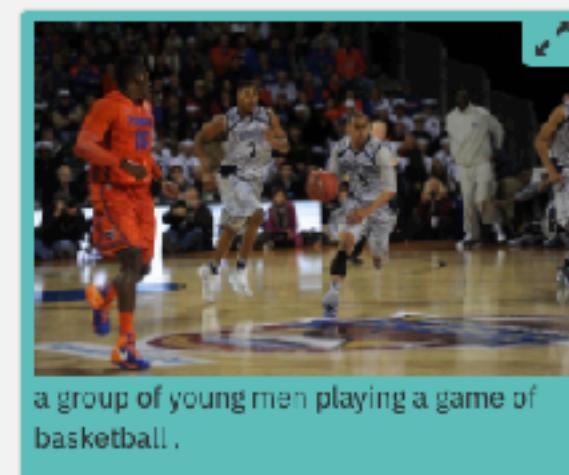
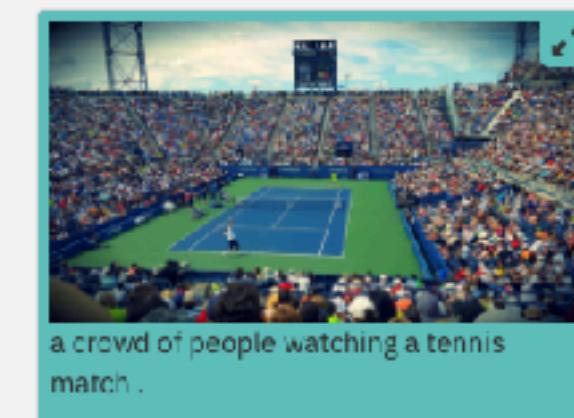
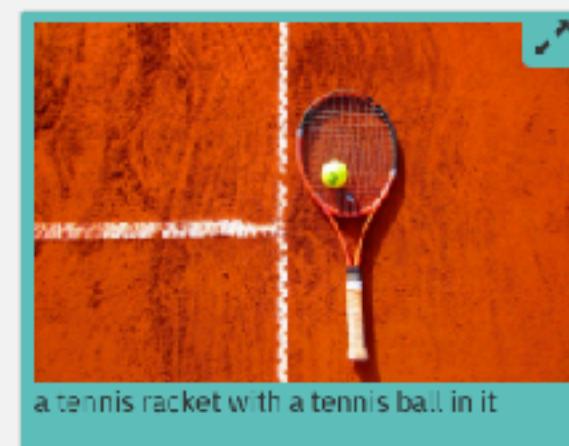
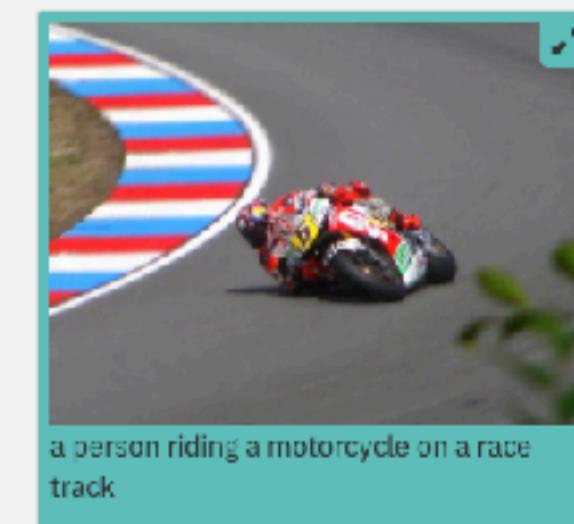
Choose Files No file chosen

Submit

Deselect All

Select All

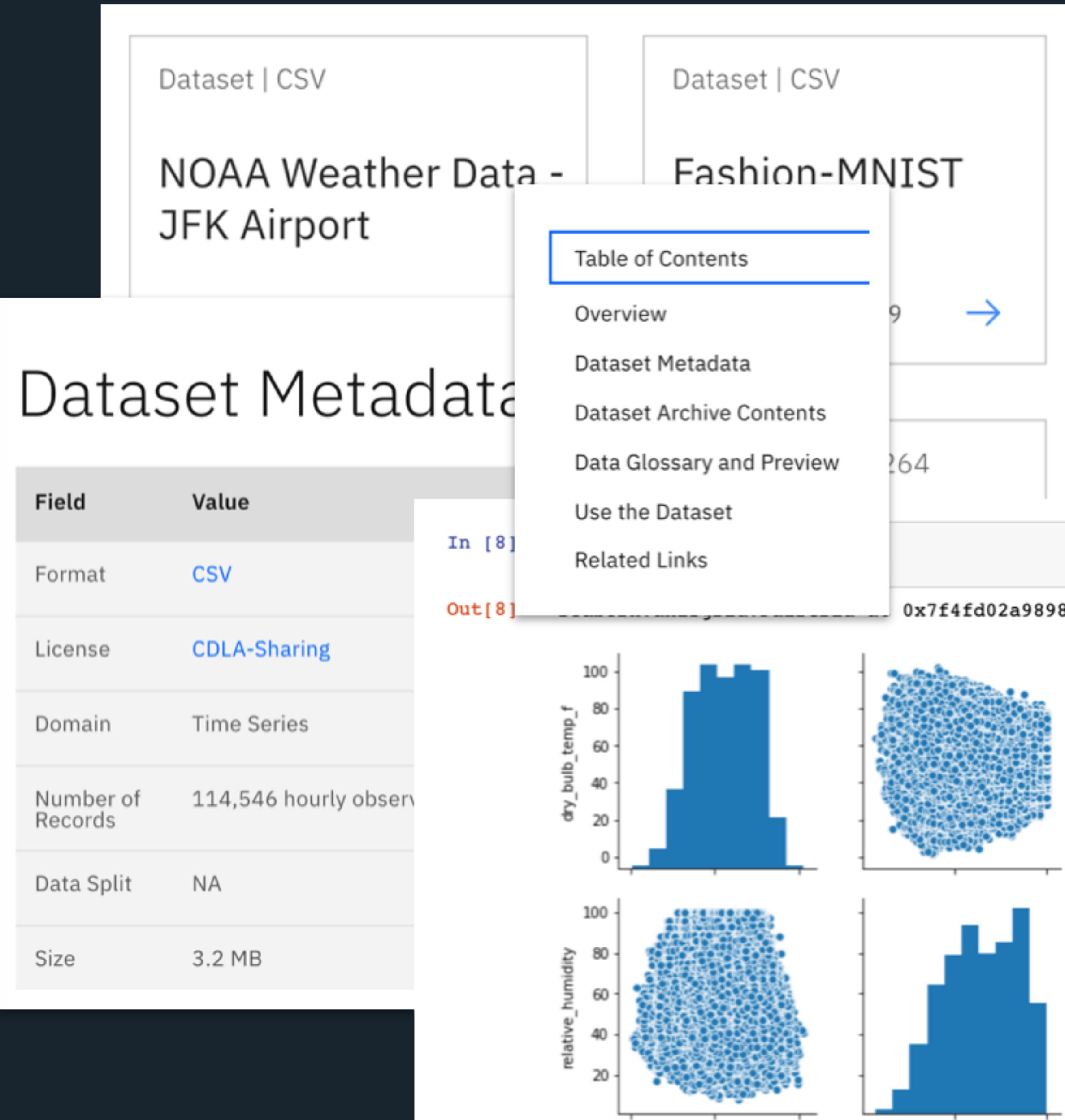
Delete Uploaded Images



surfboard
pitching
players
bike
motorcycle
carrying
basketball
playing
track
close
ramp
man
player
tennis
people
watching
top
covered
holding
group
court
side
skis
wave
action
young
race
swinging
city
snowboard
fire
Street
person
snowbaseball
crowd
bikes
slope
standing
racket
skateboard
match
skiing
hydrant
beach

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

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

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='...')
```

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

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:

- [0. Prerequisites](#)
- [1. Read the Raw Data](#)
- [2. Clean the Data](#)
 - [2.1 Select data columns](#)
 - [2.2 Clean up precipitation column](#)
 - [2.3 Convert columns to numerical types](#)
 - [2.4 Reformat and process data](#)
 - [2.5 Create a fixed interval dataset](#)
 - [2.6 Feature encoding](#)
 - [2.7 Rename columns](#)

[**ibm.biz/max-tutorial**](#)

Series

Learning Path: An introduction to the Model Asset Exchange

Learn how to use state-of-the-art deep learning models in your applications or services

[**ibm.biz/dax-tutorial**](#)

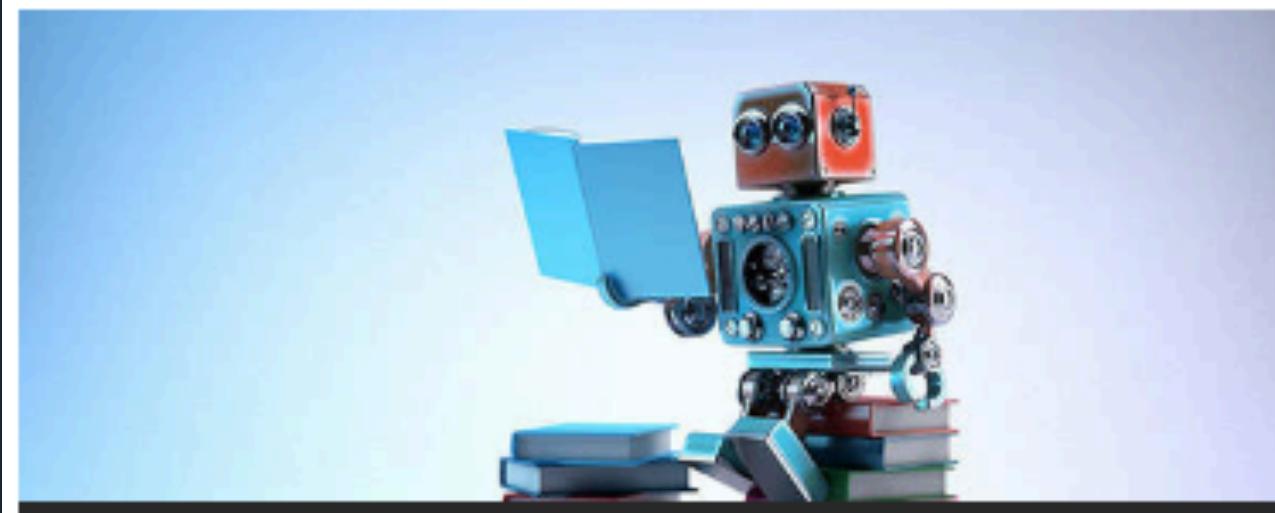
Tutorial

Get started with the Data Asset eXchange

DAX offers a trusted source for open data sets for AI that are ready to use in enterprise AI applications

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 →



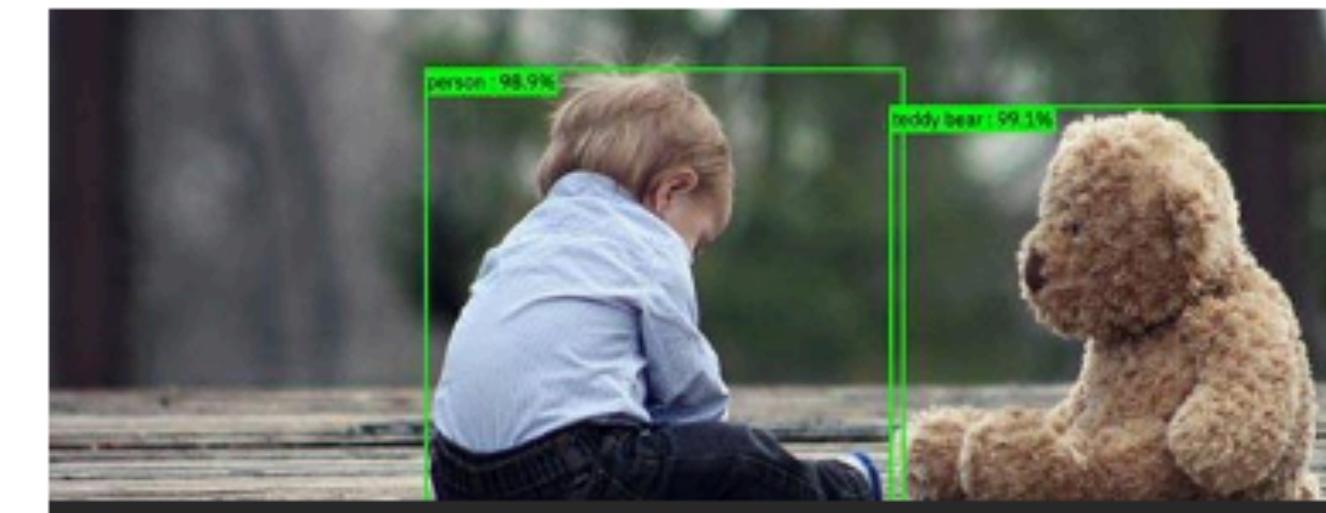
Code Pattern
Build a web app that recognizes yoga poses using a model from the Model Asset Exchange
Oct 03, 2019 →



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 →



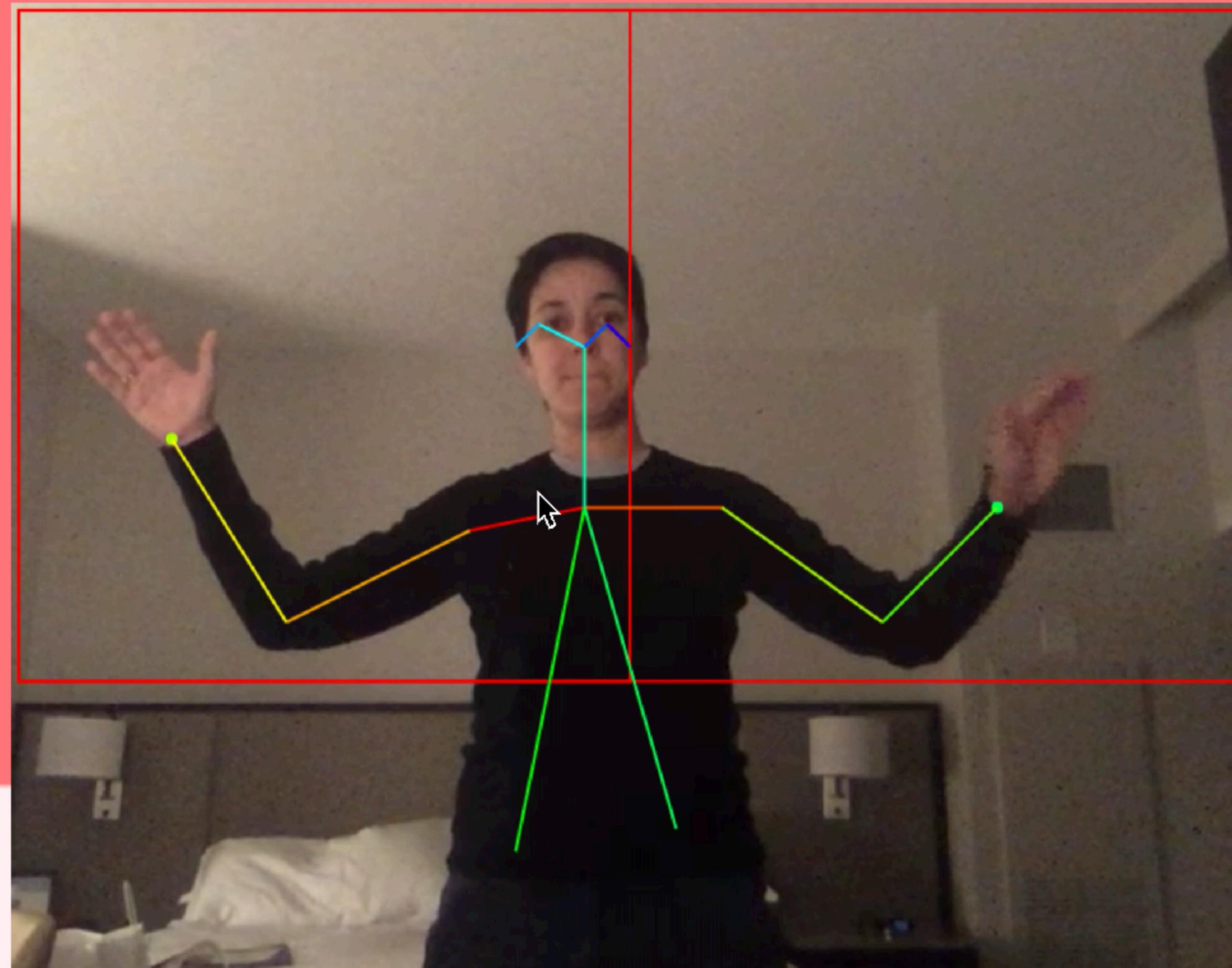
Code Pattern
Create a web app to visually interact with objects detected using machine learning
Mar 28, 2019 →



Code Pattern
Deploy a deep learning-powered 'Magic cropping tool'
Mar 28, 2019 →

veremax

a video theremin using OpenPose



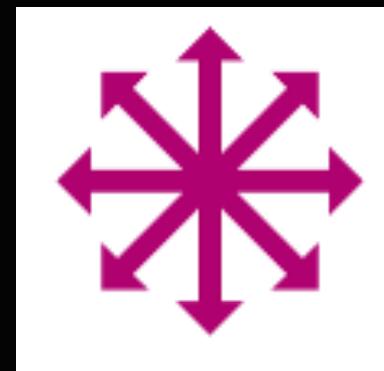
Trusted AI Lifecycle through Open Source

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?

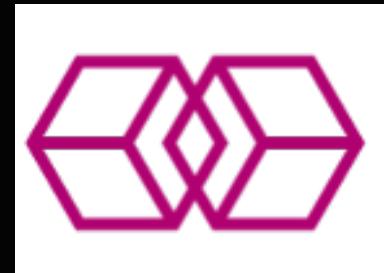


EXPLAINABILITY

AI Explainability 360
↳ (AIX360)

DEMO: aix360.mybluemix.net

Is it accountable?



LINEAGE

AI FactSheets 360

DEMO: aifs360.mybluemix.net

Trusted-AI

This GitHub org hosts LF AI Foundation projects in the category of Trusted and Responsible AI.

IBM @LFAI_Foundation info@lfaifoundation.org

Repositories 4 Packages People Projects

Pinned repositories

adversarial-robustness-toolbox
Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference
Python 1.7k 480

AIF360
A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models.
Python 1k 340

AIX360
Interpretability and explainability of data and machine learning models
Python 621 136

AI Fairness 360 (AIF360) R Package
CRAN 0.1.0 CRAN 0.1.0

Available in R too!

IDEAS



IMPORTANT SKILLS FOR A
DATA SCIENTIST

MACHINE LEARNING

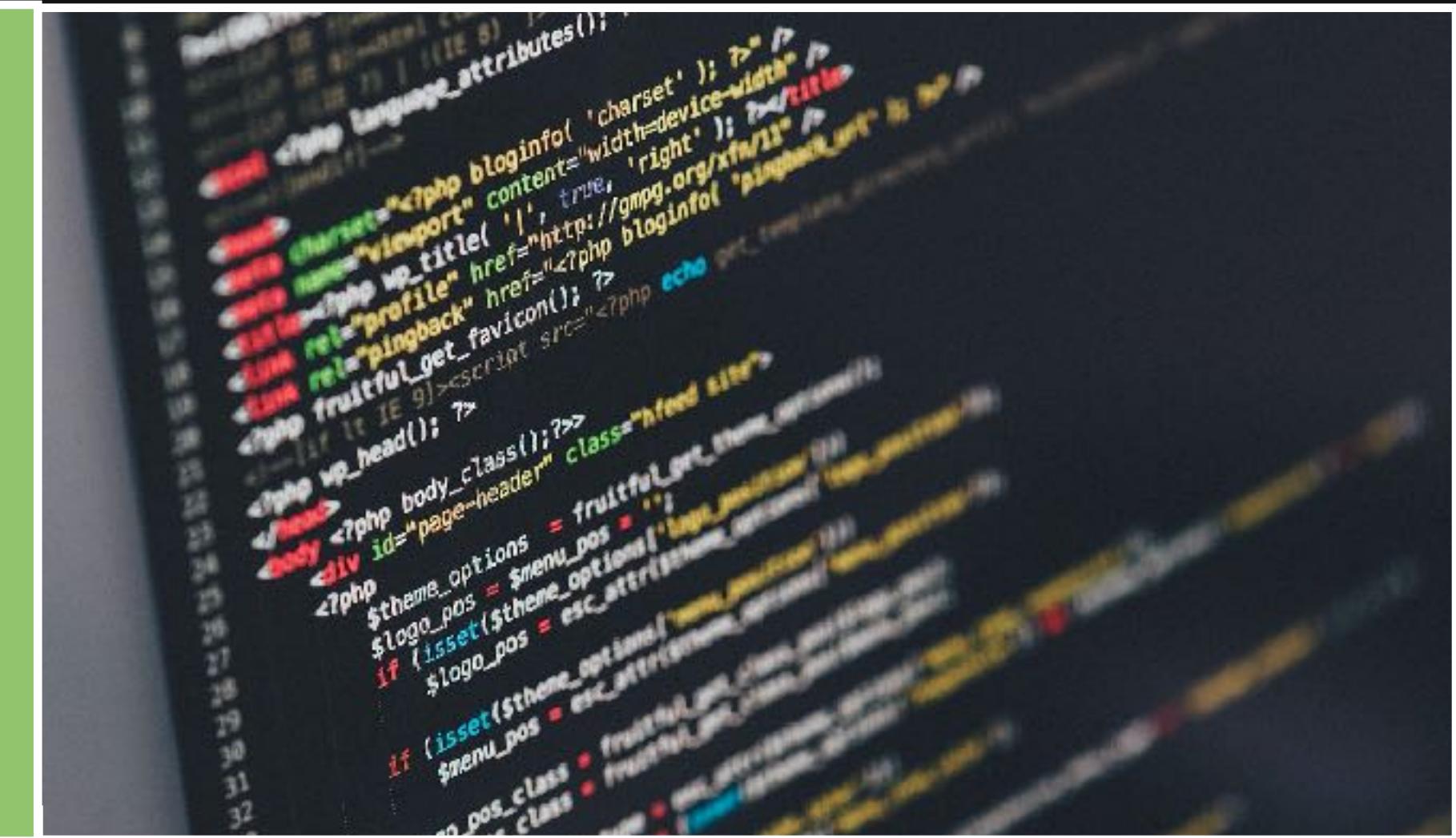


Machine Learning

Statistics

“Statistics is a **science**,
not a branch of mathematics,
but uses mathematical models
as essential tools.”

—John Tukey



```
1 <?php language_attributes(); ?>
2 <?php bloginfo( 'charset' ); ?>
3 <?php wp_head(); ?>
4 <?php wp_title( '|', true, 'right' ); ?>
5 <?php bloginfo( 'xhtml1' );
6 <?php wp_viewport() ?>
7 <?php wp_meta_tags(); ?>
8 <?php wp_head(); ?>
9 <?php wp_head(); ?>
10 <?php wp_head(); ?>
11 <?php wp_head(); ?>
12 <?php wp_head(); ?>
13 <?php wp_head(); ?>
14 <?php wp_head(); ?>
15 <?php wp_head(); ?>
16 <?php wp_head(); ?>
17 <?php wp_head(); ?>
18 <?php wp_head(); ?>
19 <?php wp_head(); ?>
20 <?php wp_head(); ?>
21 <?php wp_head(); ?>
22 <?php wp_head(); ?>
23 <?php wp_head(); ?>
24 <?php wp_head(); ?>
25 <?php wp_head(); ?>
26 <?php wp_head(); ?>
27 <?php wp_head(); ?>
28 <?php wp_head(); ?>
29 <?php wp_head(); ?>
30 <?php wp_head(); ?>
31 <?php wp_head(); ?>
32 <?php wp_head(); ?>
```

Programming



Communication



Critical Thinking



Curiosity
(keep asking why)



Ethics



Flexibility



Be yourself

Data Science is a Team Work

Individuals with different skill sets, backgrounds, views, ideas, where they will support each step of the data science process.



Thank you!

slides & materials: bit.ly/kroz-talks



dataquest.io/scholarship

Don't forget to star the repo

