Data Science as a Team Sport

Gabriela de Queiroz

@gdequeiroz | linktr.ee/gdq

slides: bit.ly/kroz-talks

Image: www.freepik.com/vectors/people

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

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

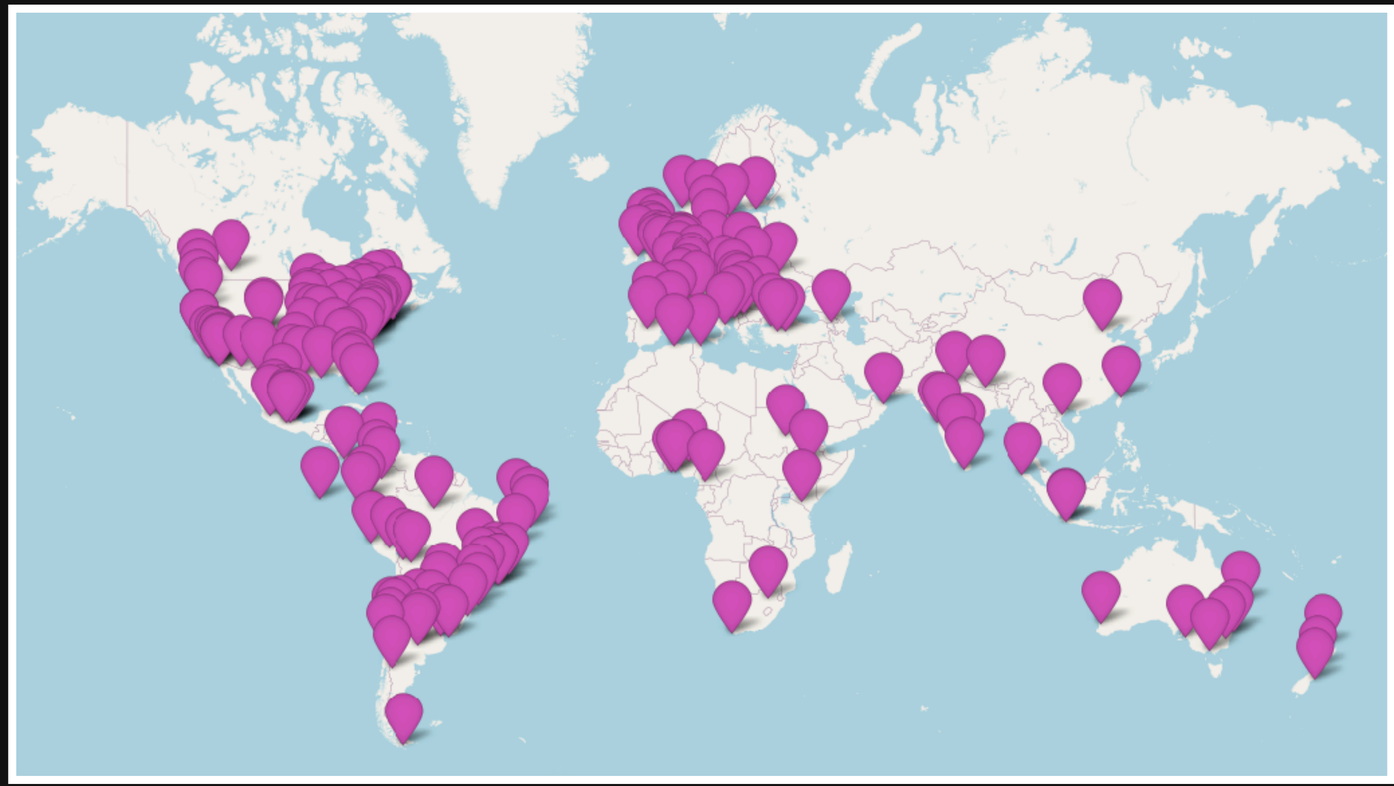
rladies.org

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

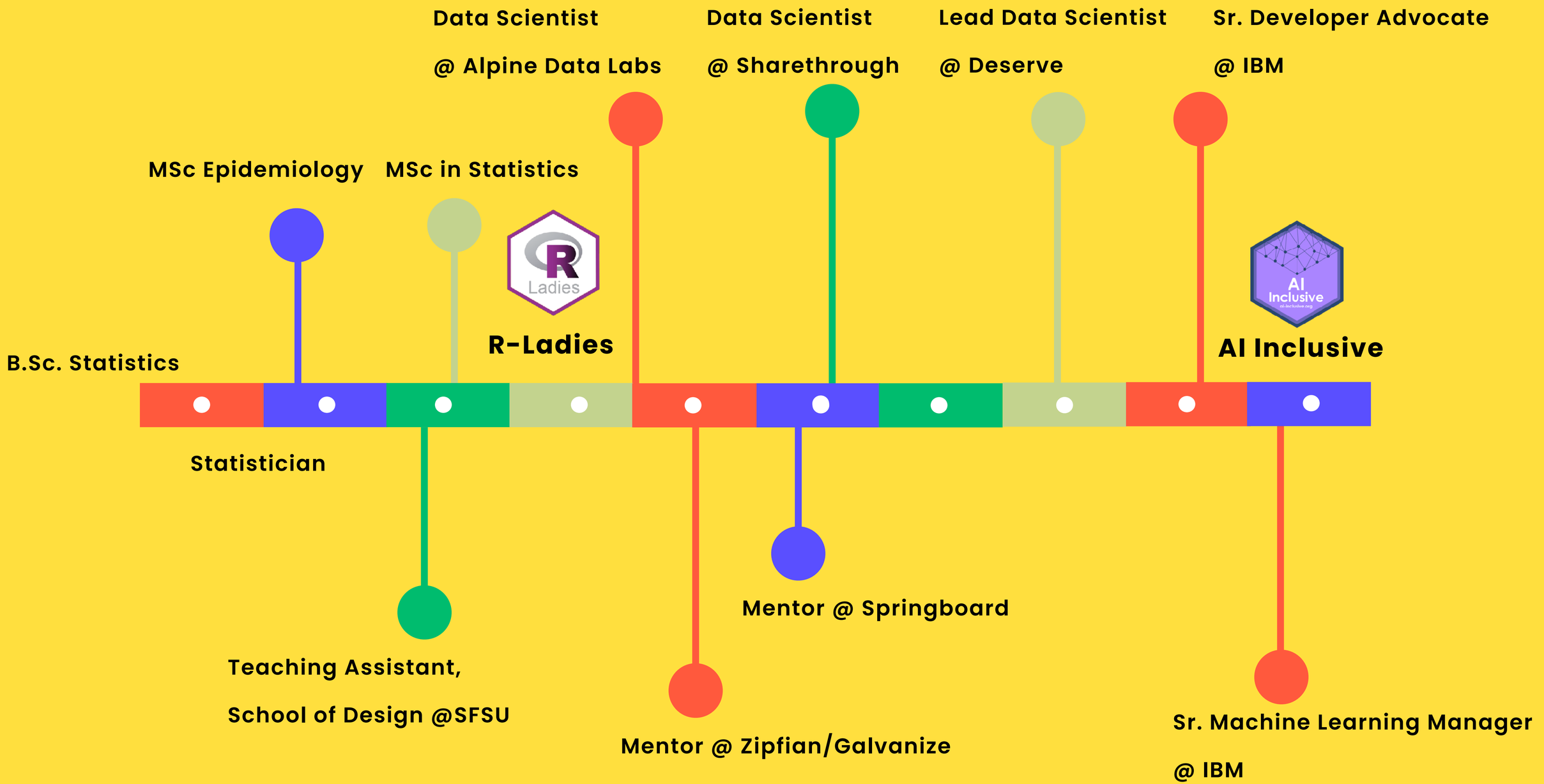
Dataquest & AI Inclusive Scholarship Program

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

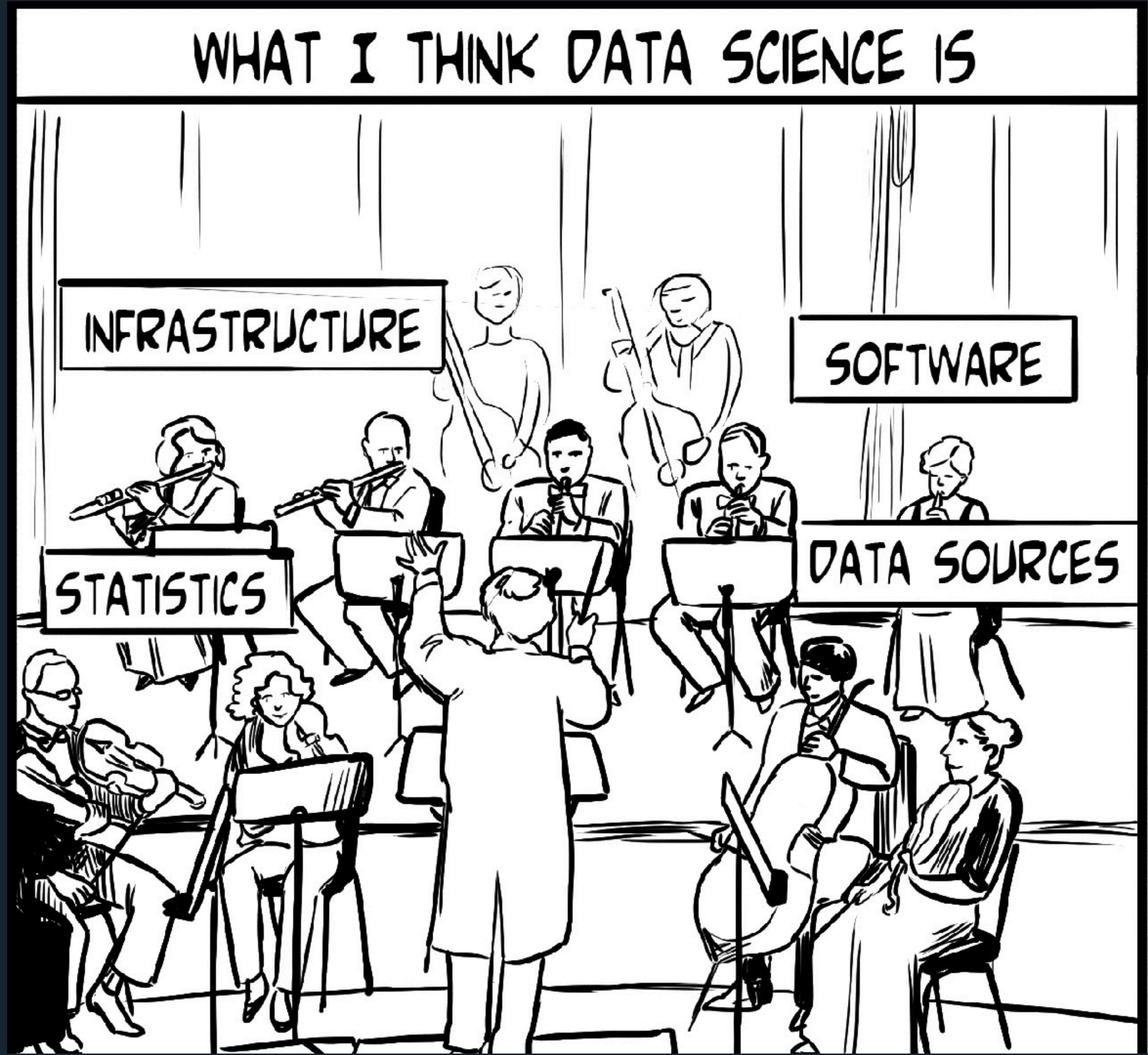
🗒️ Applications close on December 18th


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

The Data Science Career

What is Data Science?

What is Data Science?




Data Scientists are **adaptable** and
flexible professionals

Companies

What is the role of a data scientist?

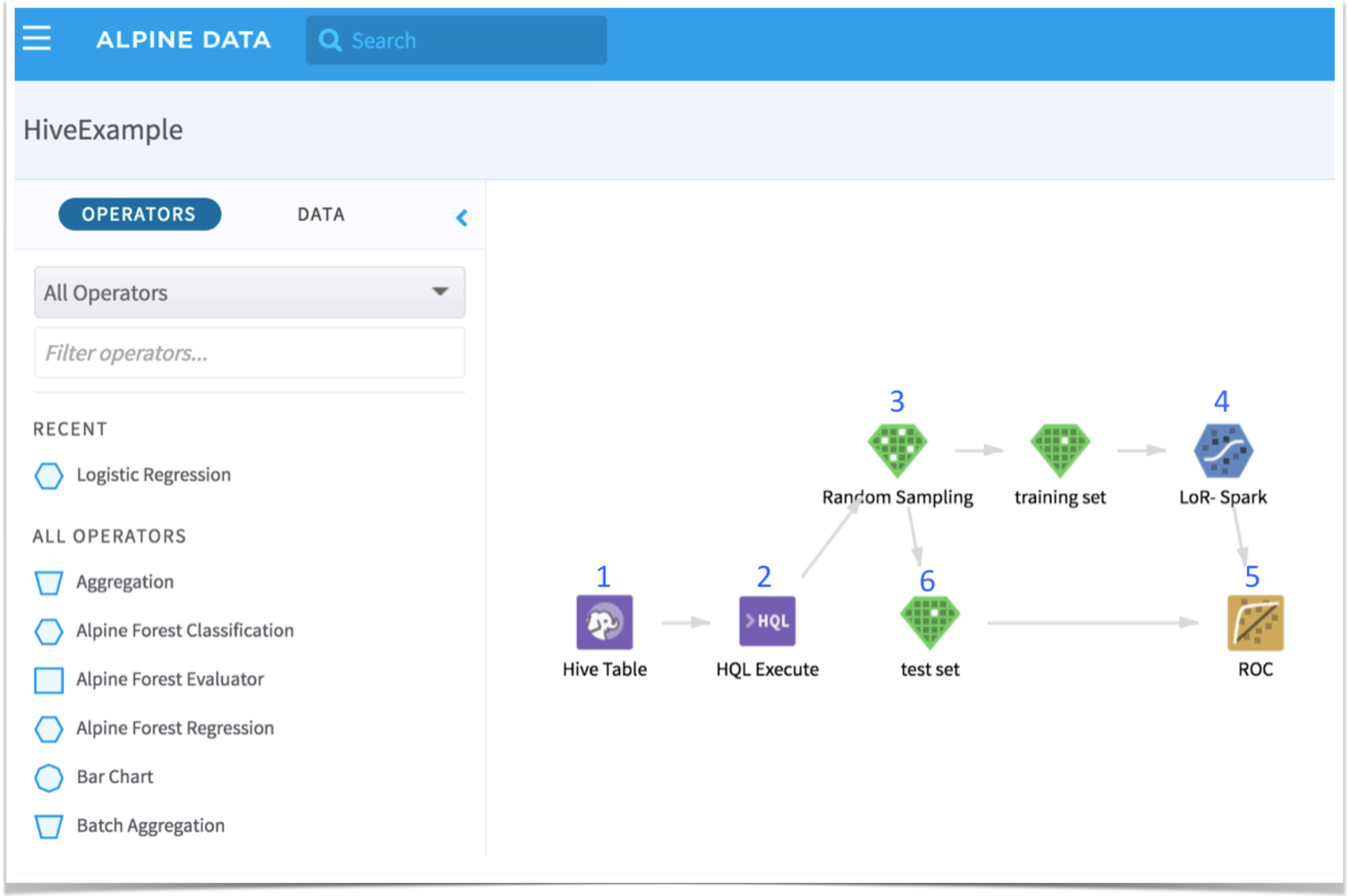
Data Scientists can have different roles in different companies

slides: **bit.ly/kroz-talks**

Advanced Analytics Platform for Big Data

ALPINE DATA™

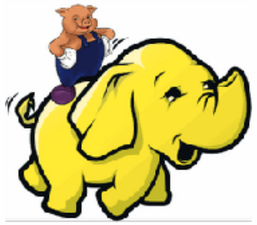
Empowers business users to define and participate in data science projects and gives data scientists the tools they need to create value from data



Company Numbers

- **40** Employees
- **10** Engineers
- **5** Data Scientist

Tools

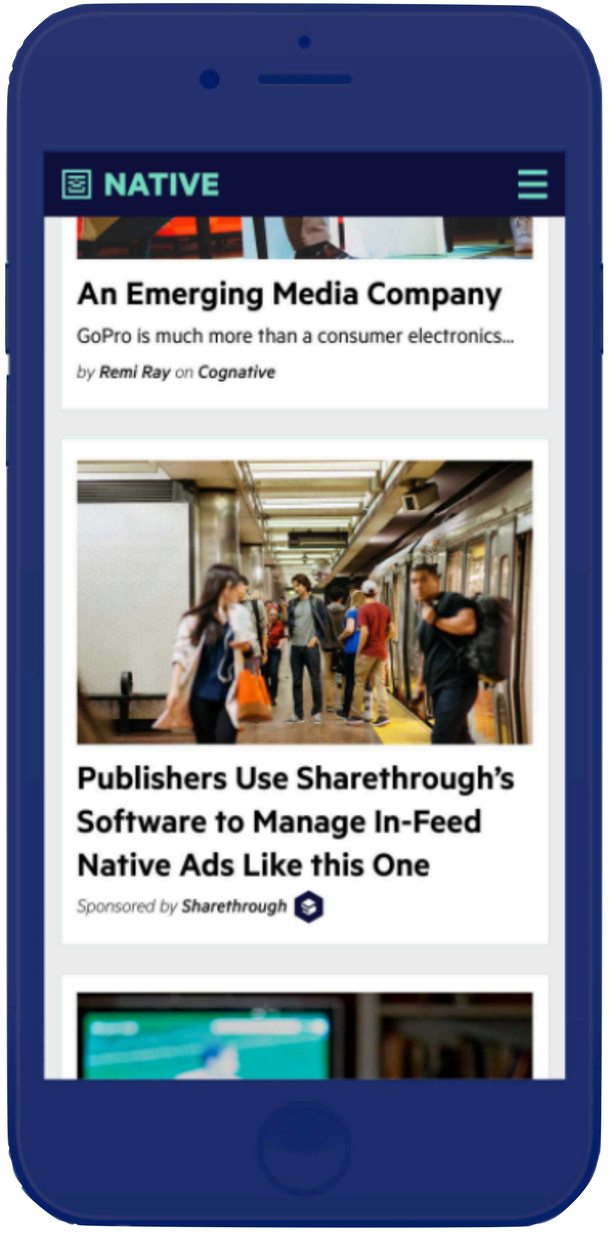


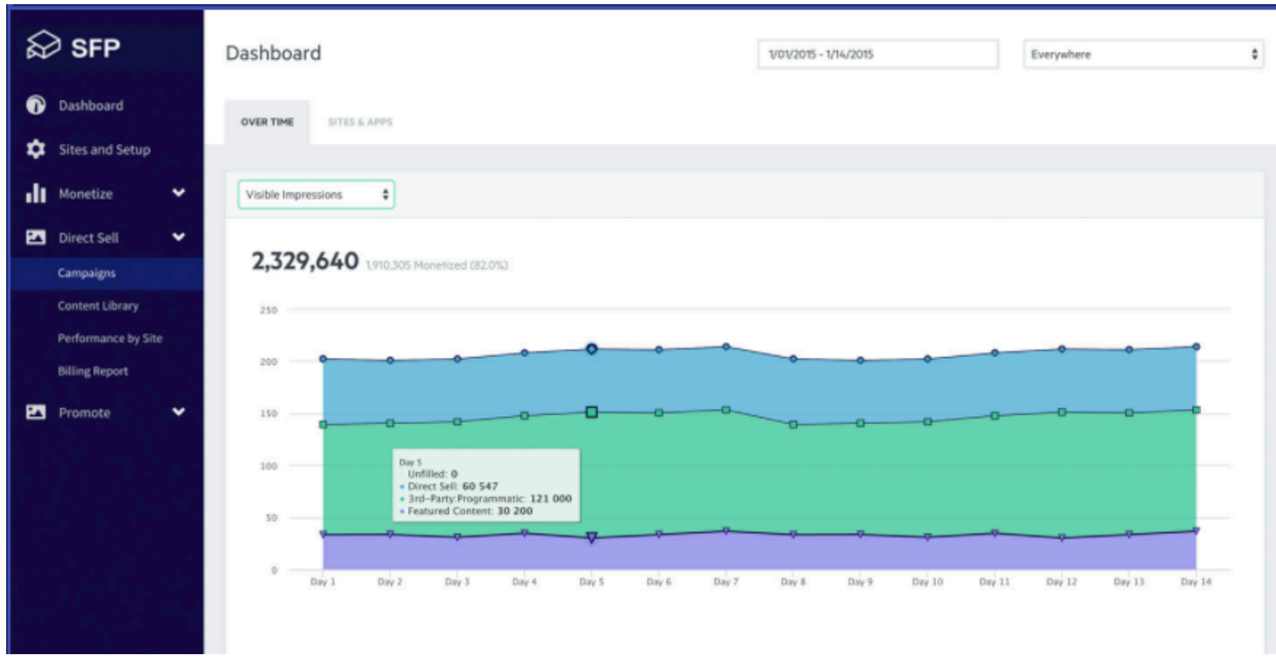
Key Responsibilities:

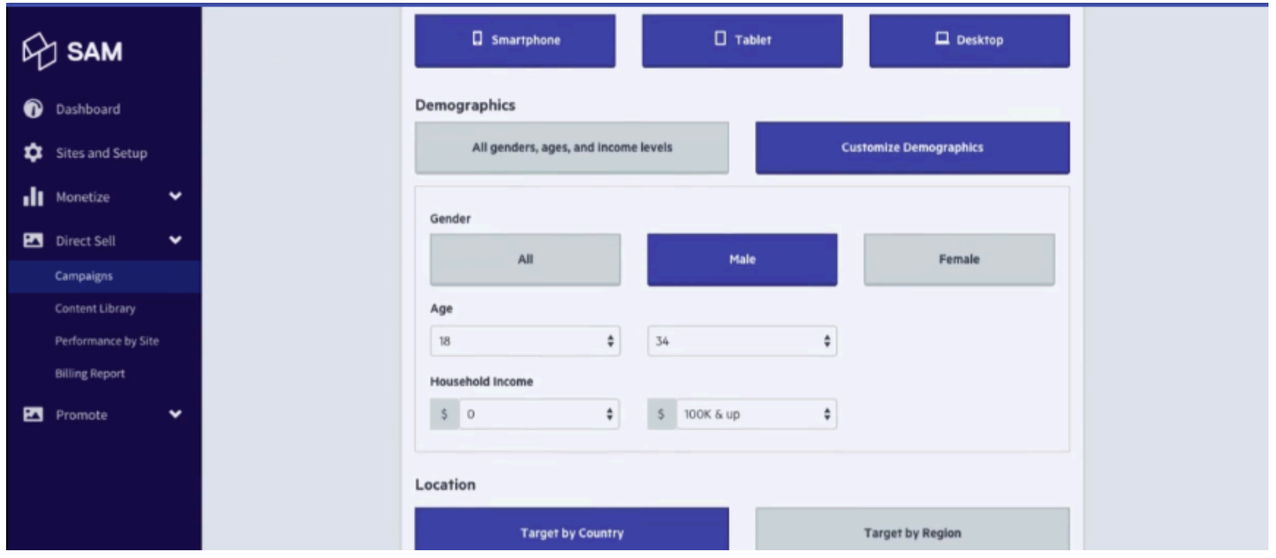
- Consultant
- Write Documentation
- Train Customers
- Prioritization

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

Credit Card designed for Generation Z

(INCLUDING INTERNATIONAL STUDENTS)

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.

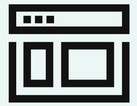
Company Numbers

- **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

python™

docker

R

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

Responsibilities: (MANY!)

MACHINE LEARNING TEAM

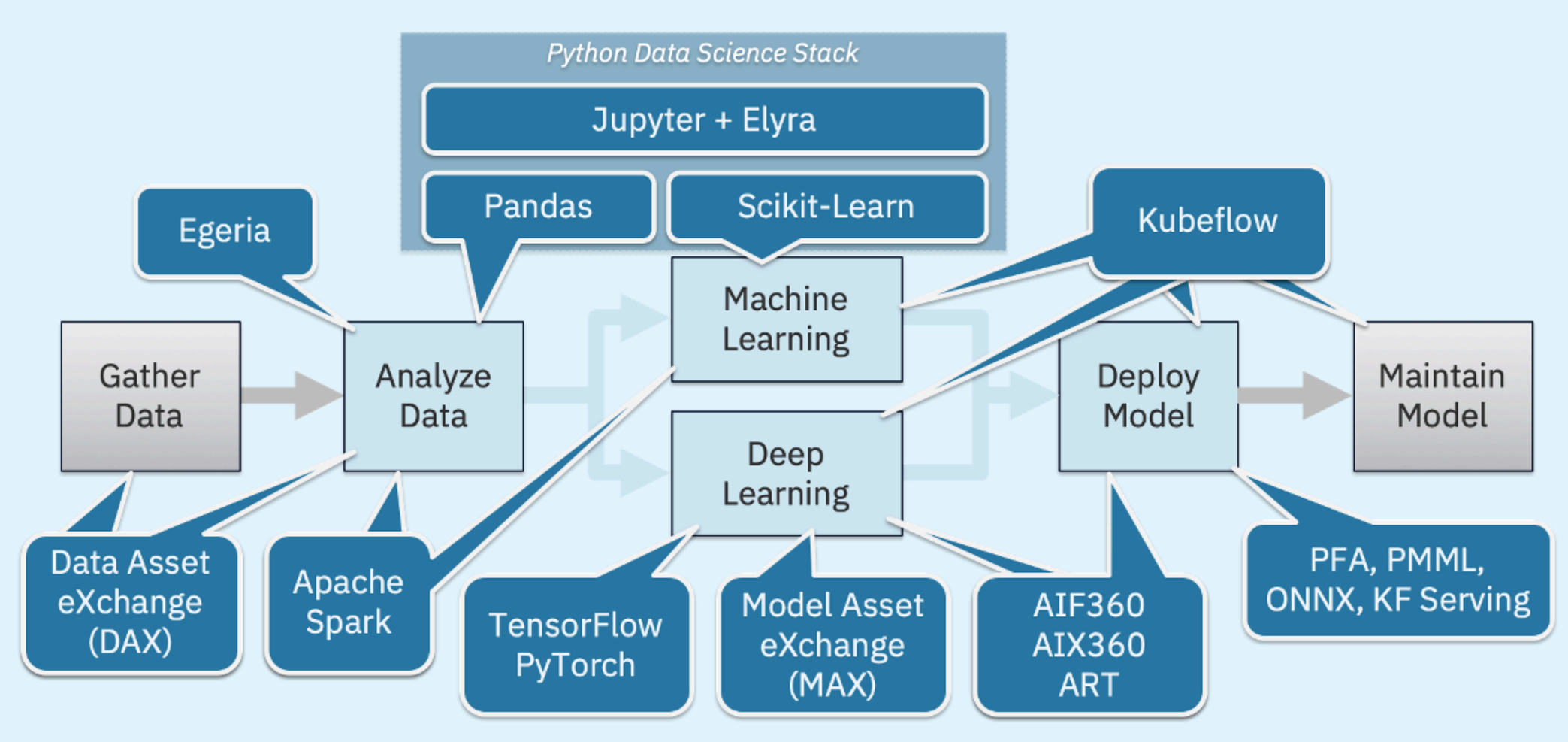
10 open source developers

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)



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

codait.org

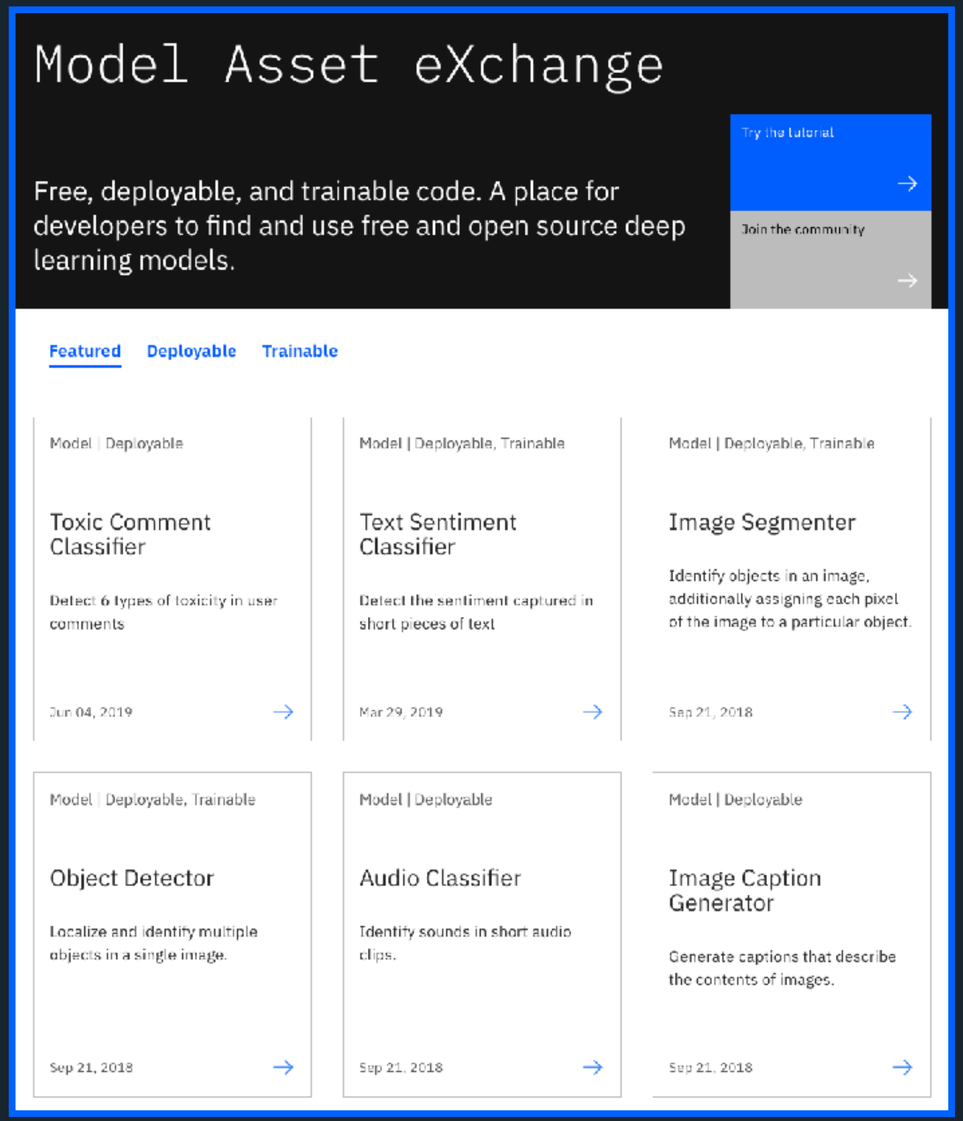
Open Source @ IBM

Some Projects

slides: bit.ly/kroz-talks

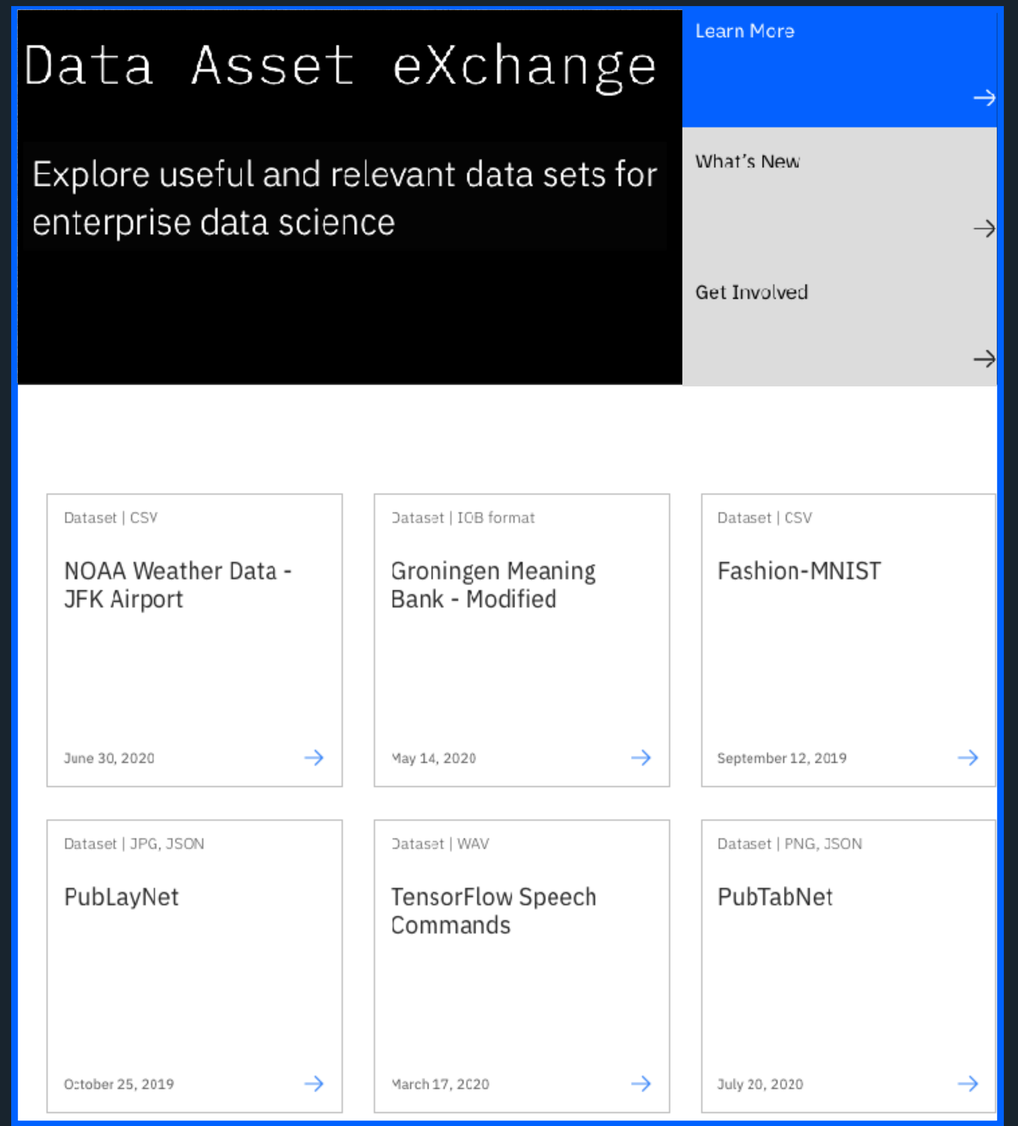
Model Asset eXchange

ibm.biz/model-exchange



Data Asset eXchange

ibm.biz/data-exchange

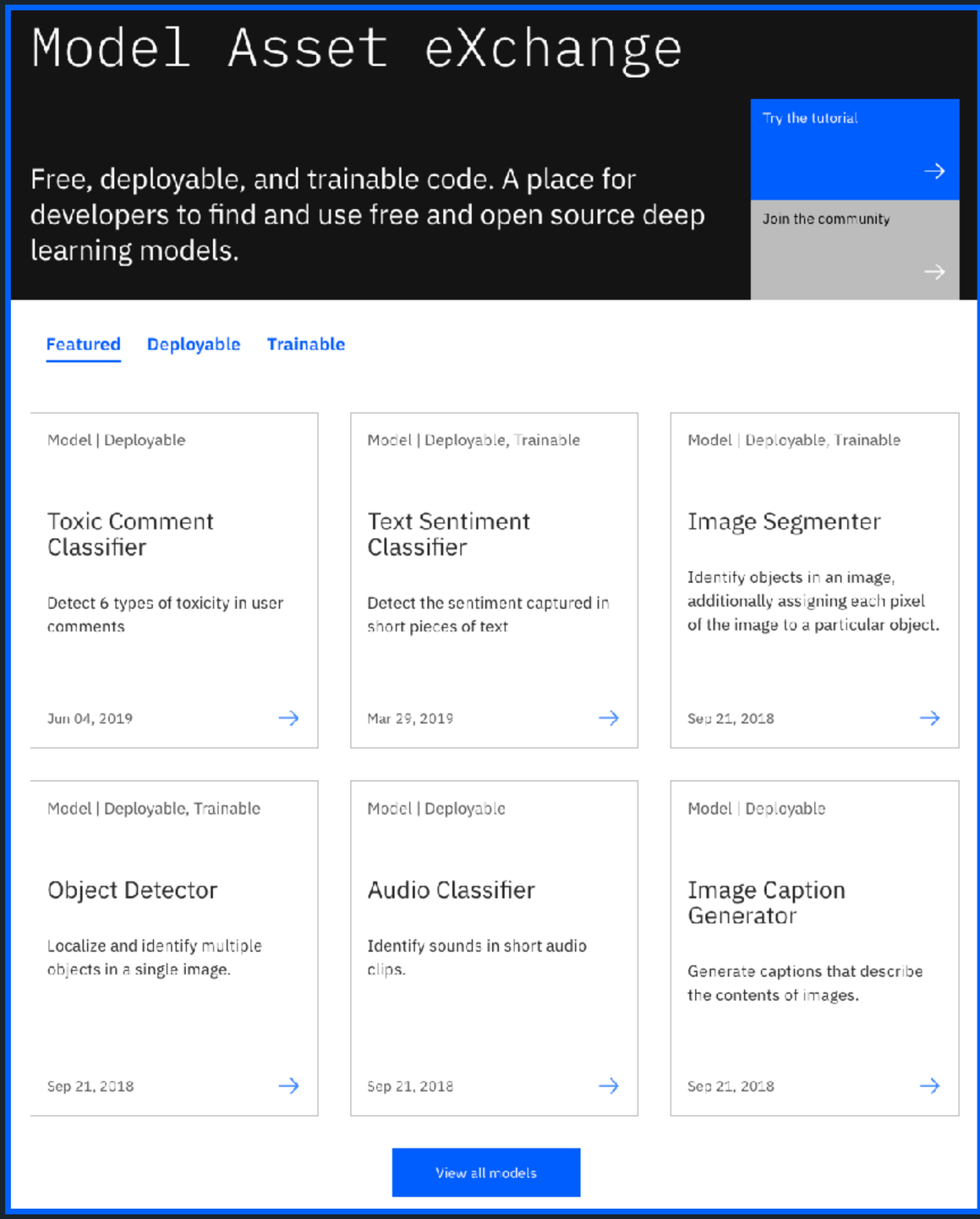


Model Asset eXchange (MAX)

Place for developers/data scientists to find and use

free and **open source** deep learning models

ibm.biz/model-exchange

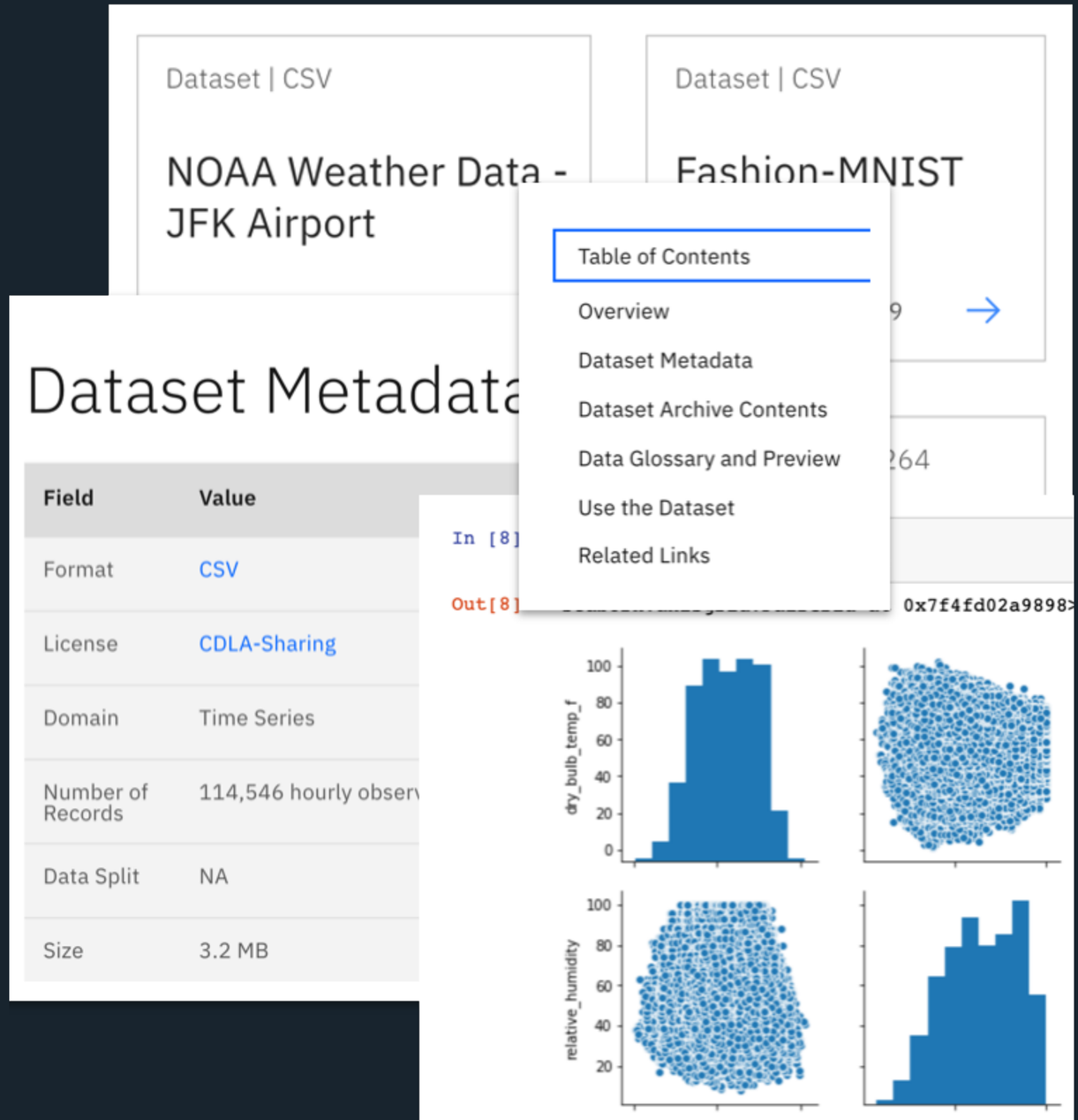


slides: bit.ly/kroz-talks


MAX Object Detector
Upload an image
Choose File No file chosen
Submit
Use your webcam
Filter detected objects
Probability Threshold: 50%
Labels Found
person
person
person
person
motorcycle : 81.4%
car : 97.4%
person
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 .
surfboard
man
riding
top
covered
player
tennis
slope
snow
baseball
field
ball
street
skateboard
standing
person
snowboard
holding
group
court
people
ramp
skis
wave
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

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.

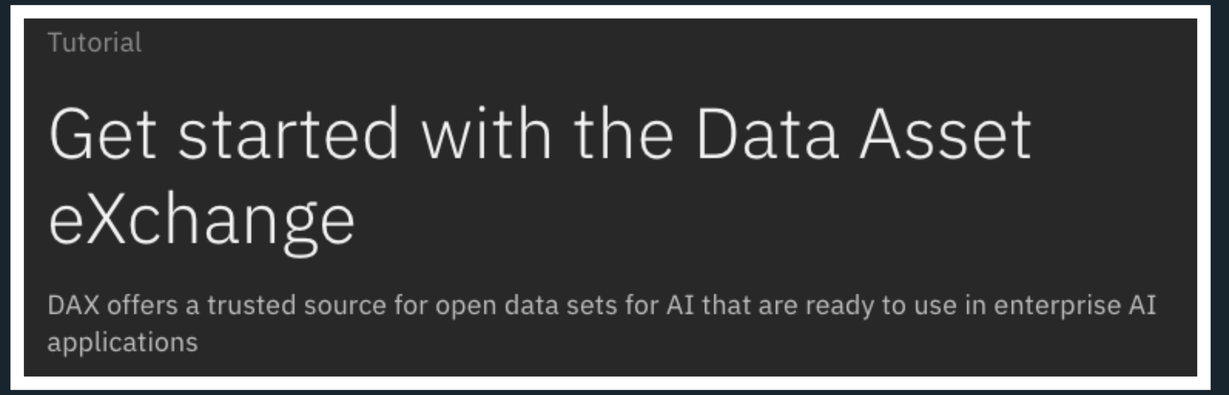
In this notebook, we clean the raw dataset by:

- 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:

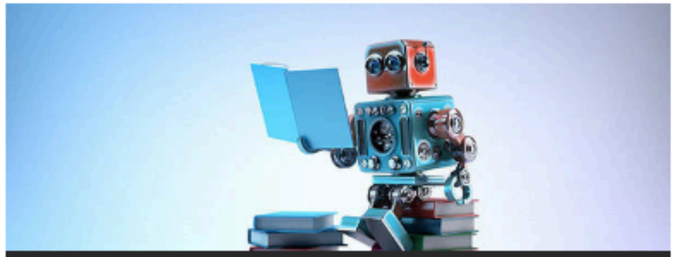
ibm.biz/max-tutorial

ibm.biz/dax-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

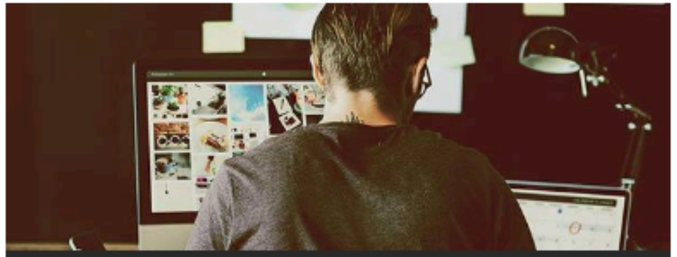
Build a web app that recognizes yoga poses using a model from the Model Asset Exchange

Oct 03, 2019

Code Pattern

Use your arms to make music

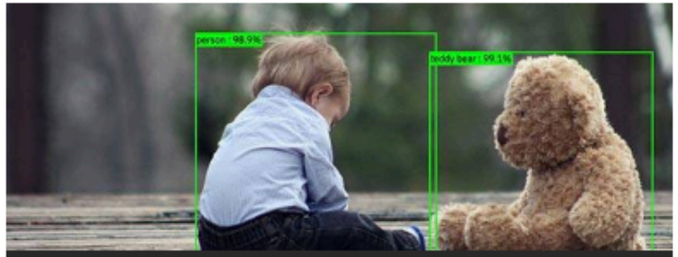
Apr 22, 2019



Code Pattern

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

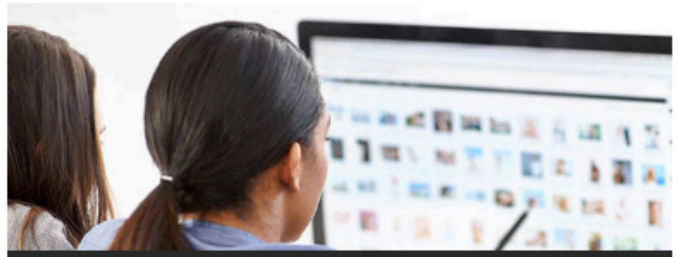
Mar 28, 2019



Code Pattern

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

Mar 28, 2019



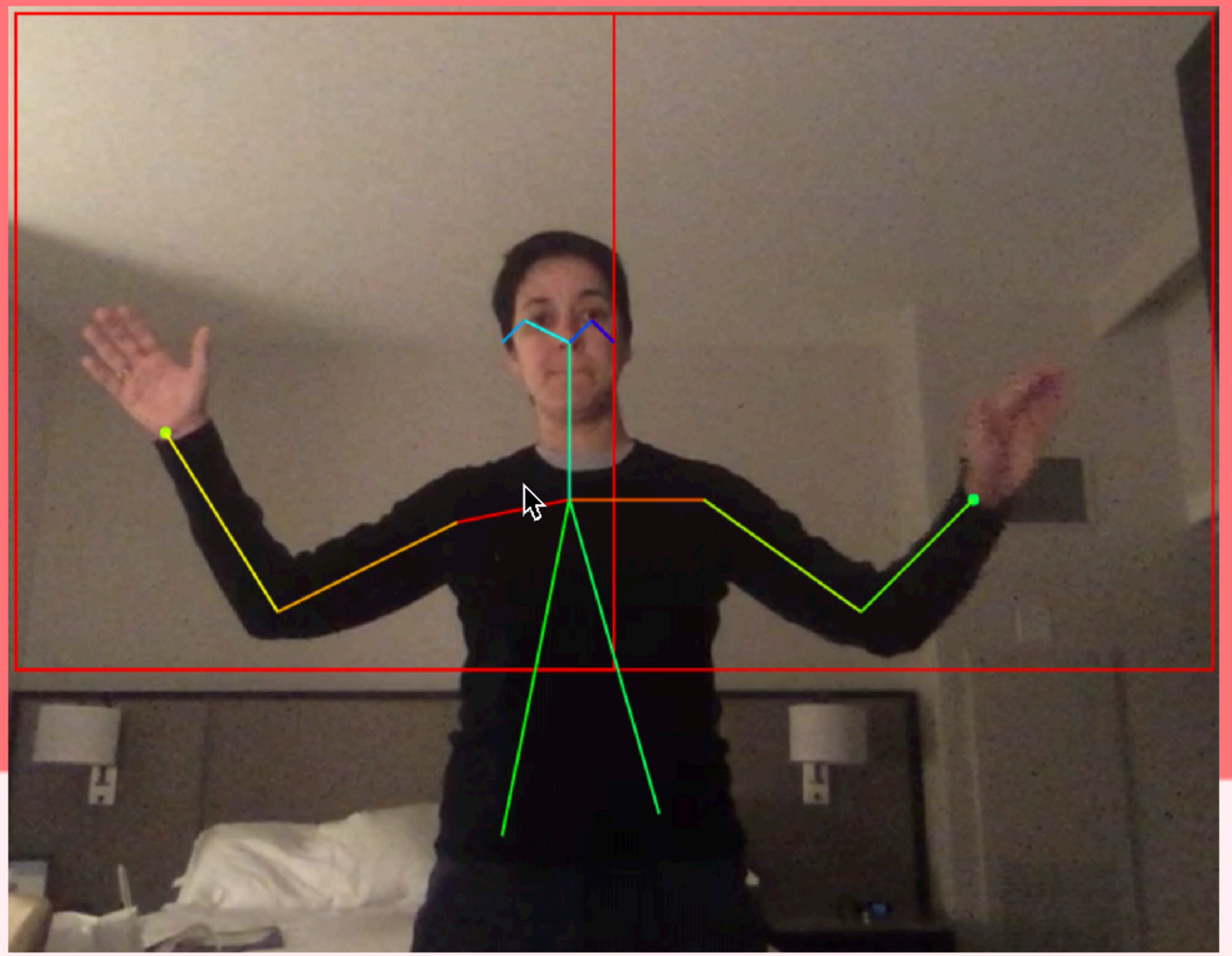
Code Pattern

Deploy a deep learning-powered 'Magic cropping tool'

Mar 28, 2019

veremax

a video theremin using OpenPose



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

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

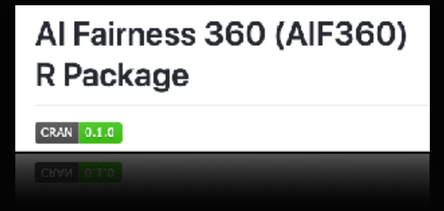
Is it accountable?

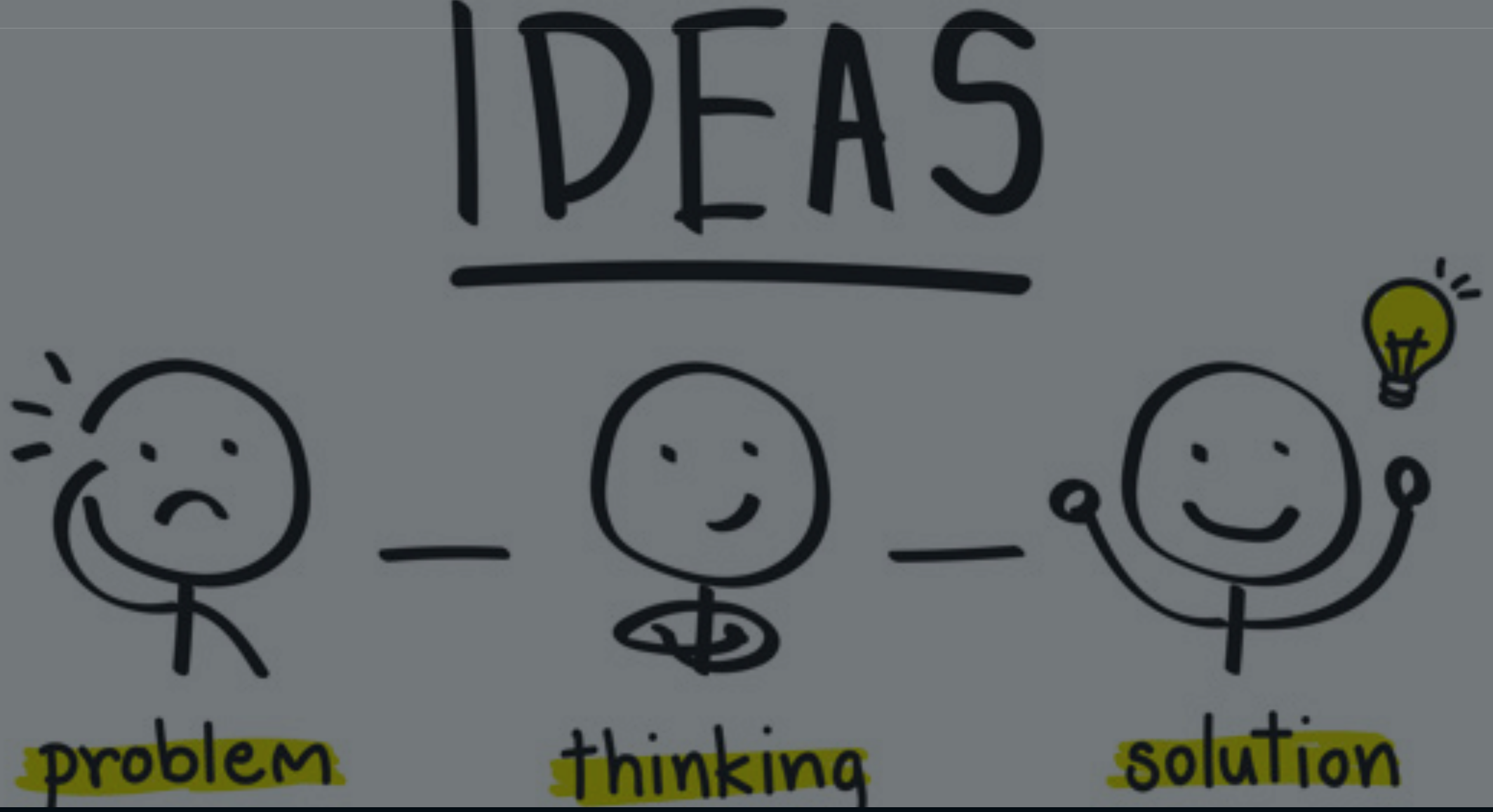
LINEAGE

AI FactSheets 360

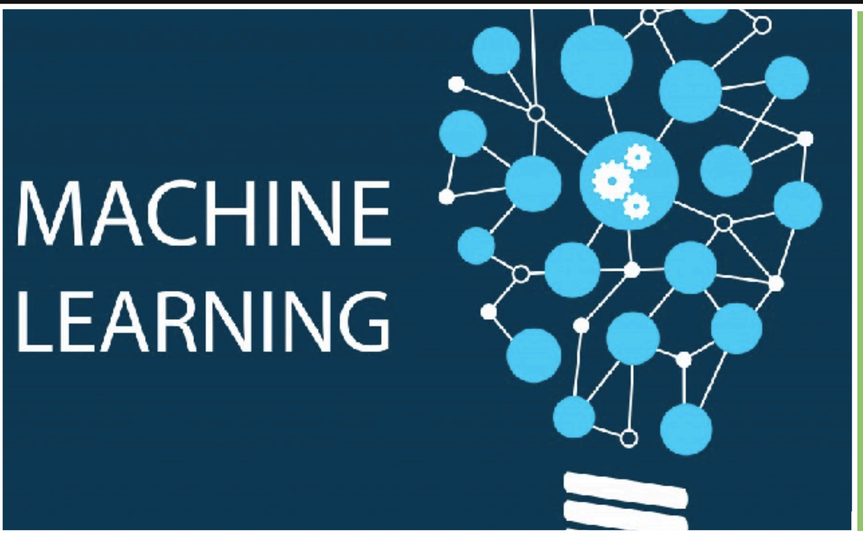
DEMO: aifs360.mybluemix.net

Available in R too!



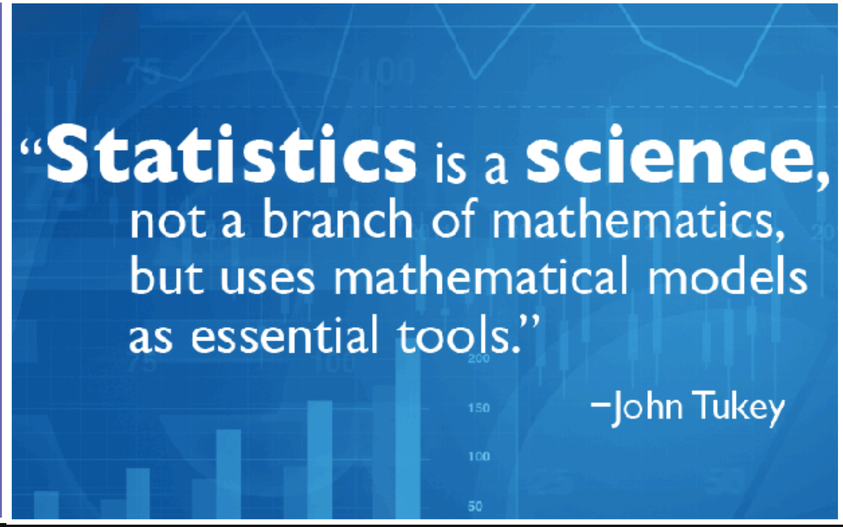


IMPORTANT SKILLS FOR A DATA SCIENTIST



Machine Learning

Statistics



Programming

Communication

Curiosity
(keep asking why)

Flexibility

Critical Thinking

Ethics

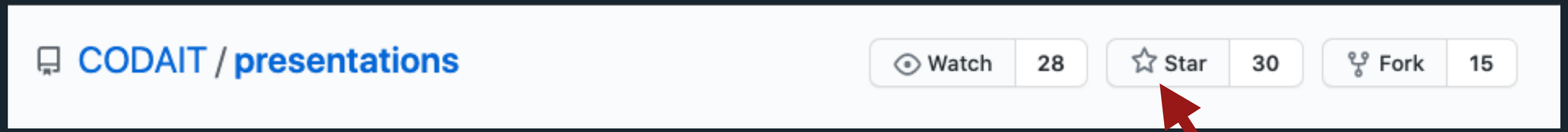
Be yourself

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



Don’t forget to star the repo

dataquest.io/scholarship