Data Science as a Team Sport

Gabriela de Queiroz
Sr. Machine Learning Manager, IBM

@gdequeiroz | linktr.ee/gdq

Image: www.freepik.com/vectors/people

Gabriela de Queiroz

Sr. Machine Learning Manager, IBM California

- 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

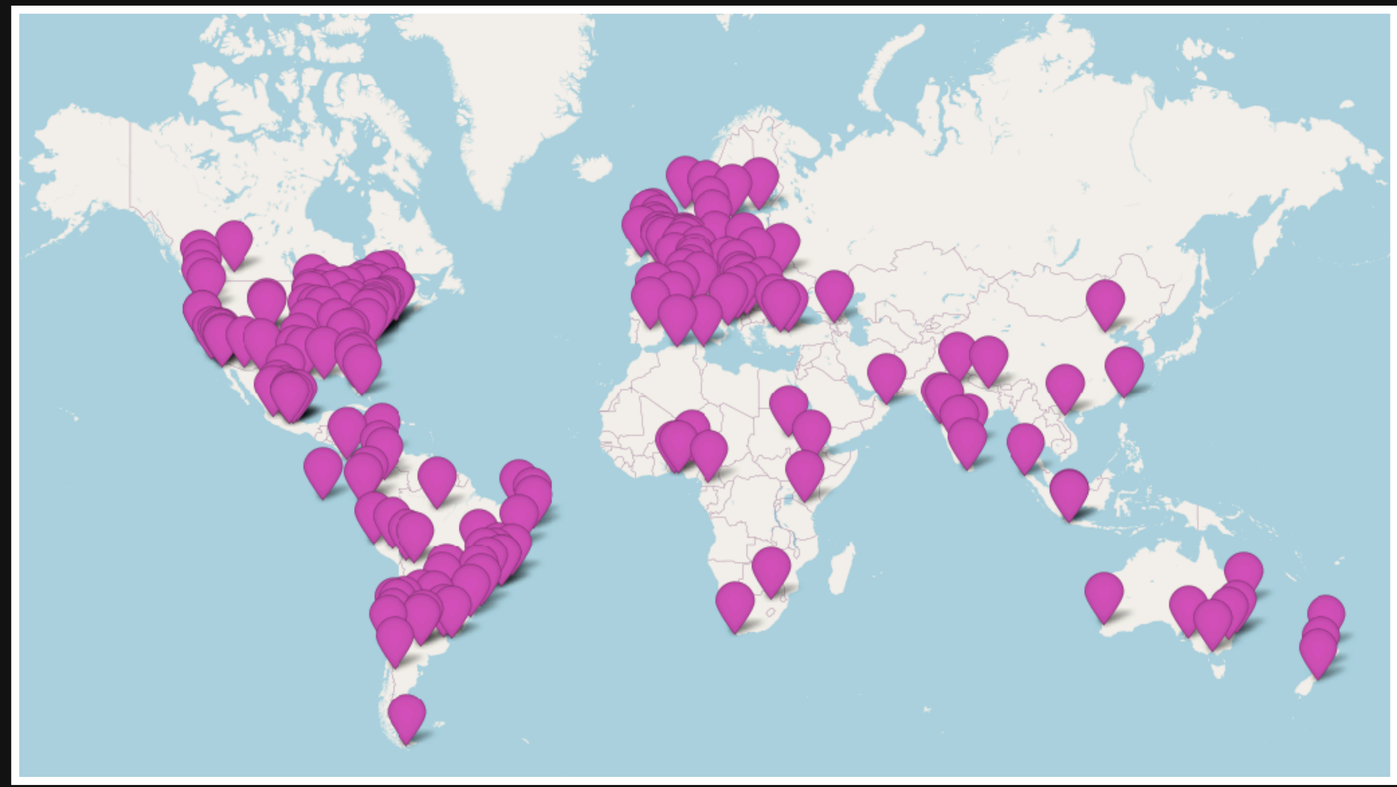
rladies.org

198
Chapters

56
Countries

83614
Members

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.

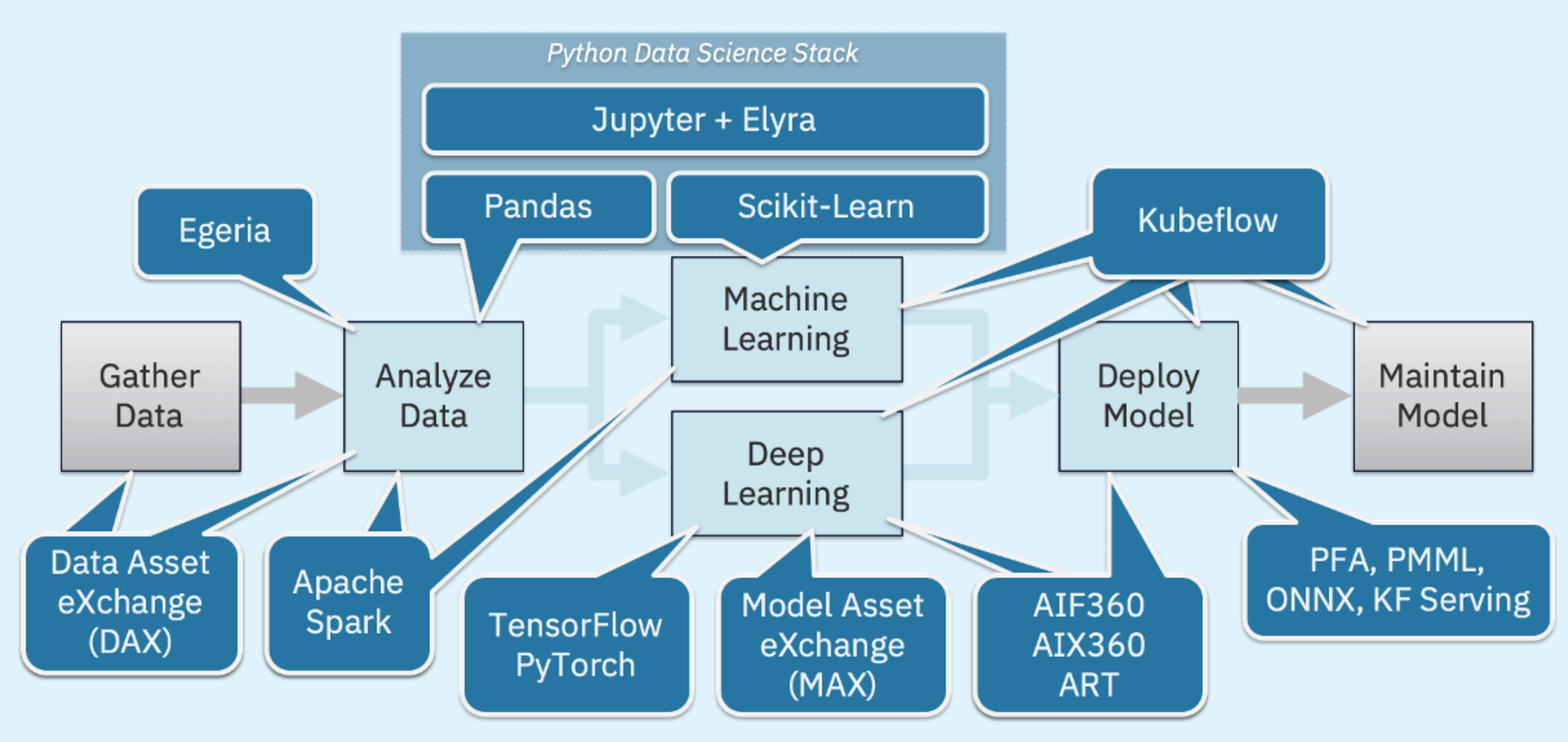
Mission: Increase the representation and participation of minority groups in Artificial Intelligence

- Website: ai-inclusive.org
- Twitter: bit.ly/ai-inclusive-twitter
- Instagram: bit.ly/ai-inclusive-instagram

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

We build tools to make AI accessible and available to everybody

(codait.org)



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

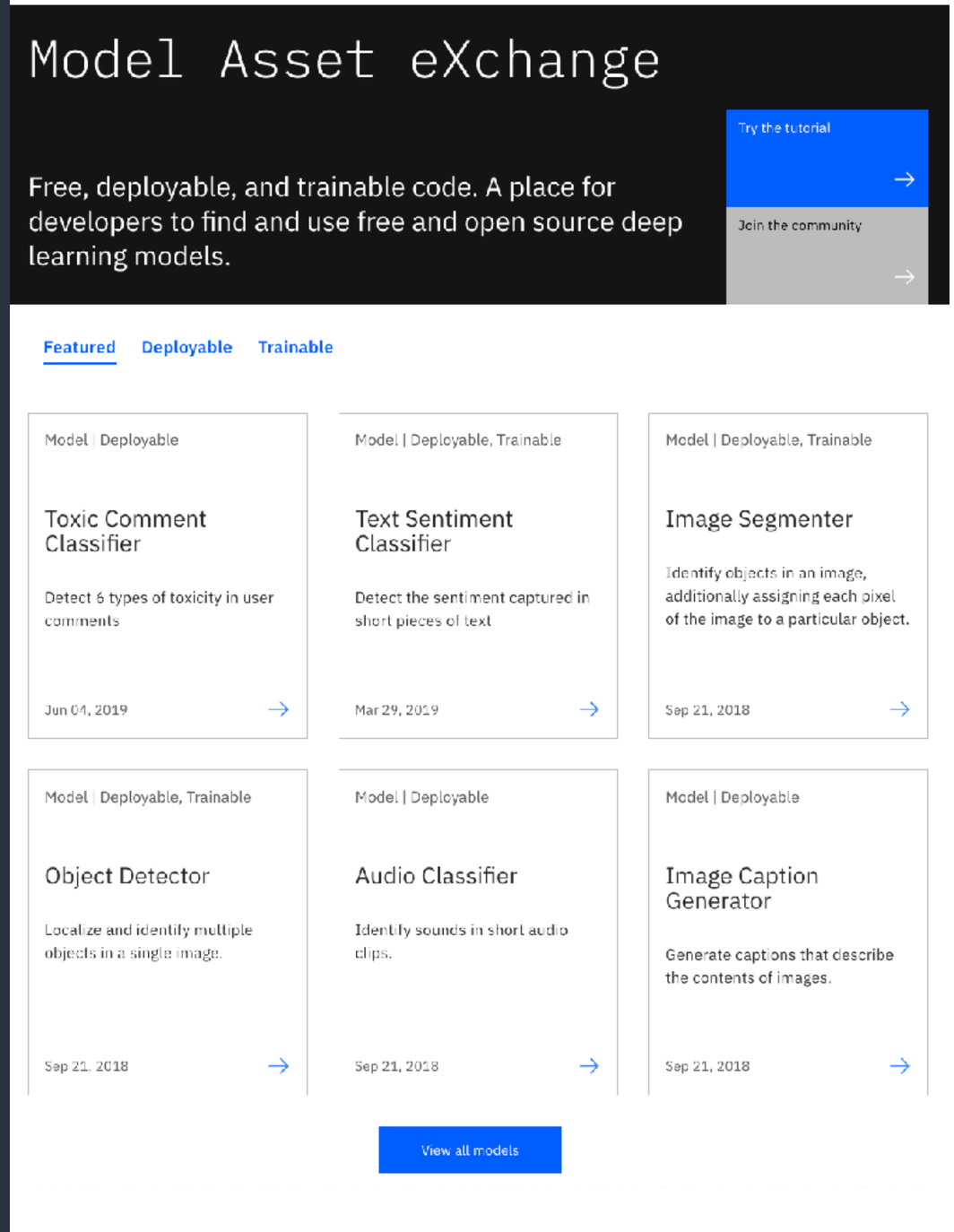
codait.org

Open Source @ IBM

Some Projects

Model Asset eXchange (MAX)

Website:
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 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

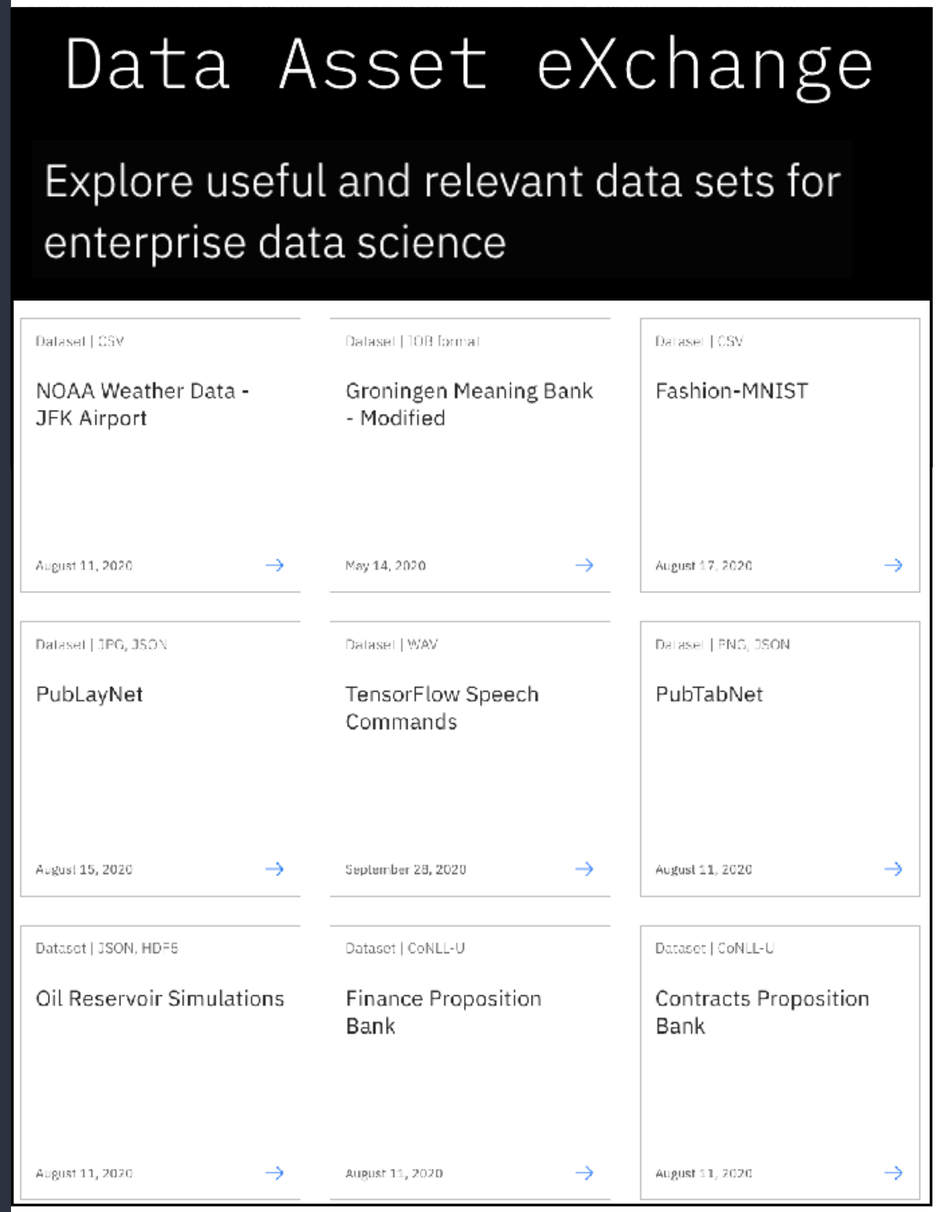
Try the API

Try the web app

Try in a Node-RED flow

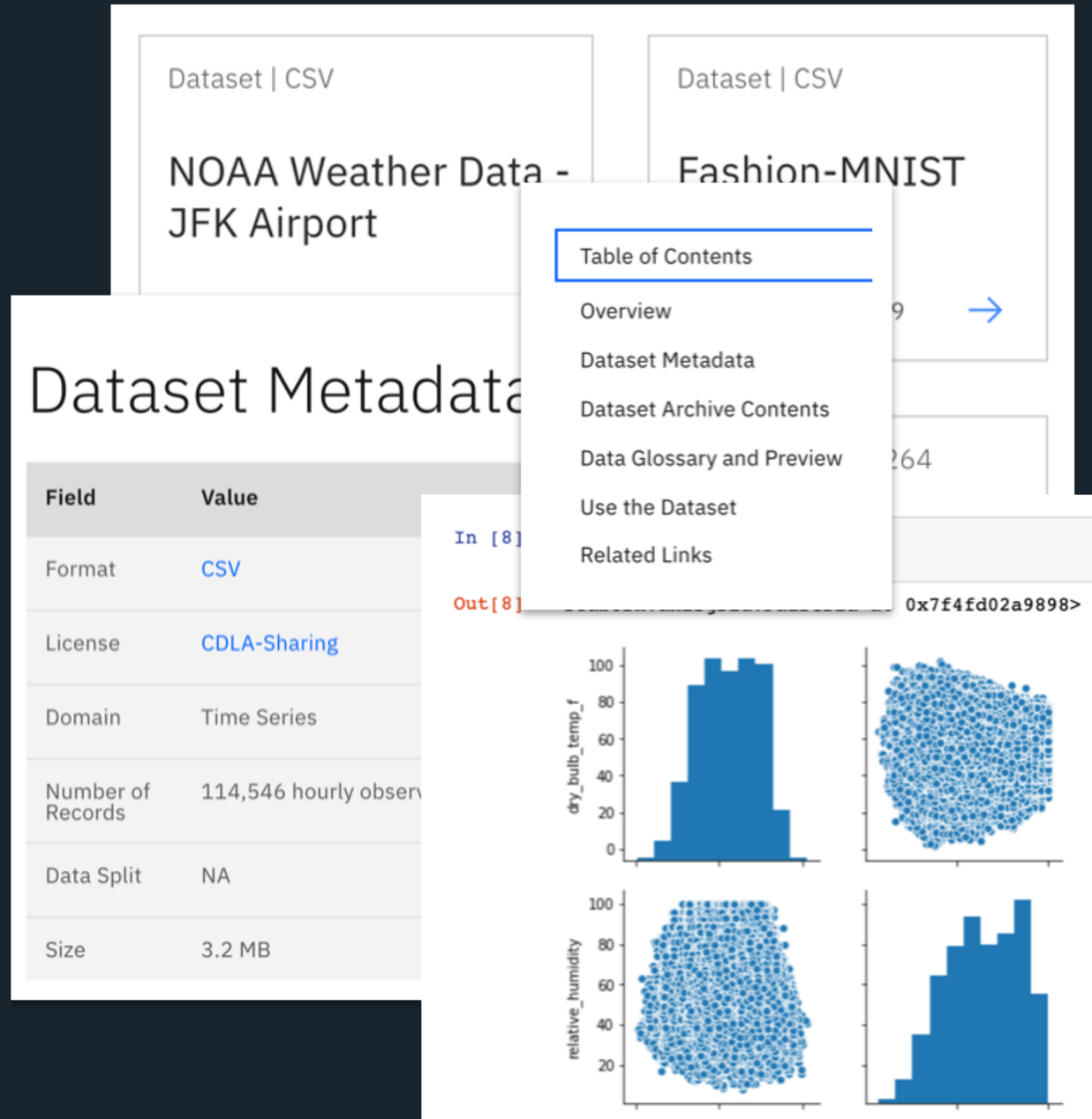
Data Asset eXchange (DAX)

Website:
ibm.biz/data-exchange



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
- Many 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 →

DAX Dataset Preview | Notebook Preview | Run Notebook in Watson Studio | Dataset Homepage

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.
Business Use Case	*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 temperature. Remotely adjust air conditioning configs to boost energy efficiency based on temperature shifts. *Retail* Estimate outdoor retail foot traffic based on weather condition and temperature predictions.

DAX Dataset Preview | Notebook Preview | Run Notebook in Watson Studio | Dataset Homepage

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 proje
ct 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

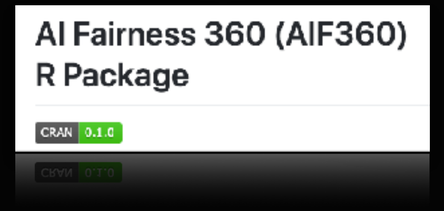
Is it accountable?

LINEAGE

AI FactSheets 360

DEMO: aifs360.mybluemix.net

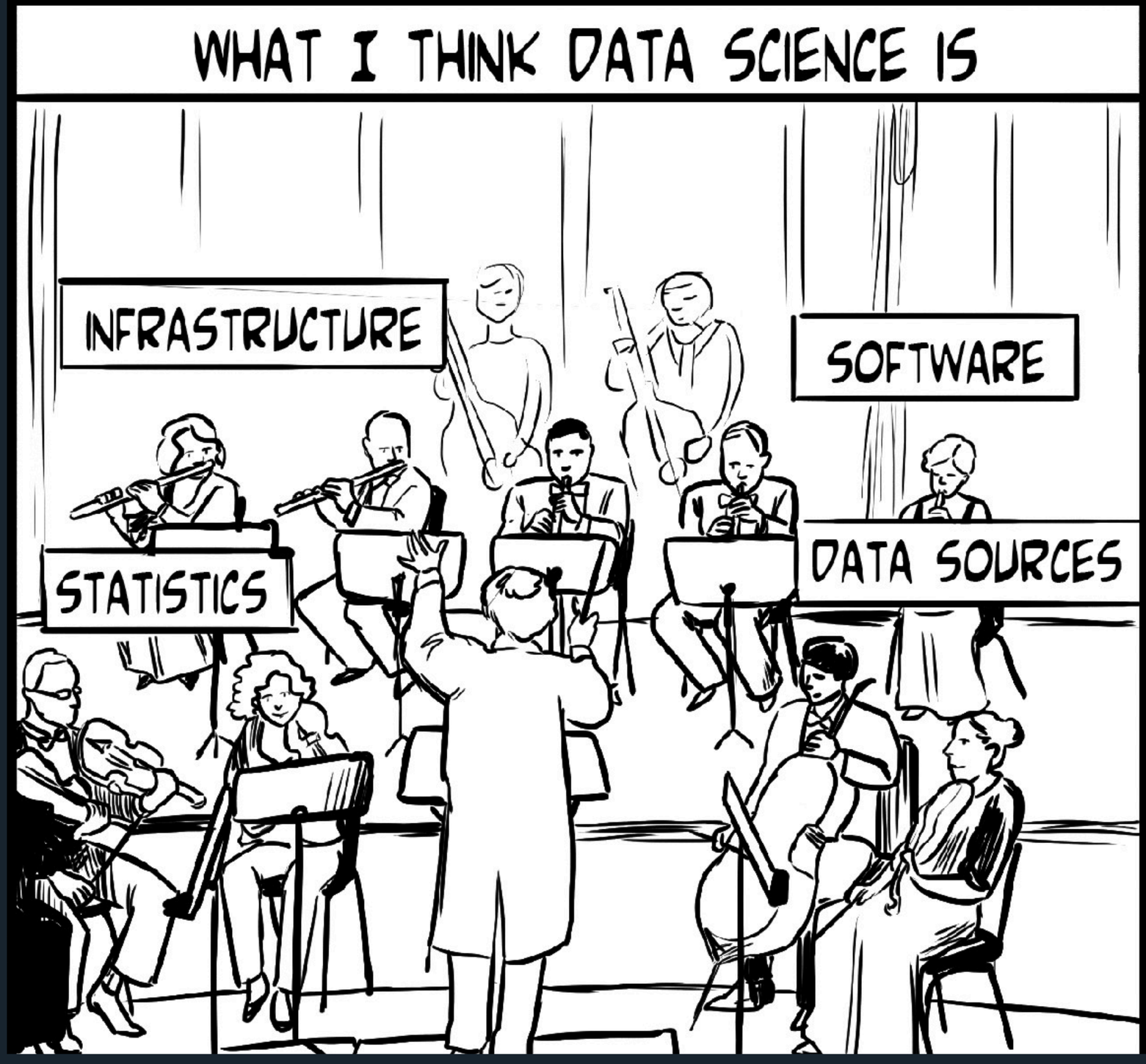
Available in R too!



What is Data Science?

https://towardsdatascience.com/explaining-data-science-ai-ml-and-deep-learning-to-management-a-presentation-and-a-script-4968491eb1e5

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

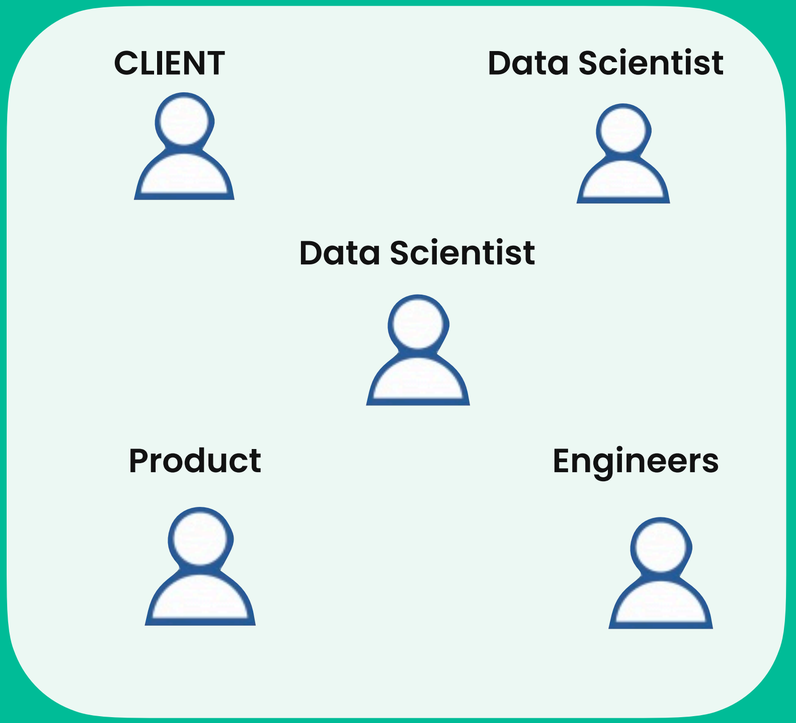
Company Numbers

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

Tools

Key Responsibilities:

- Consultant
- Write Documentation
- Train Customers
- Prioritization



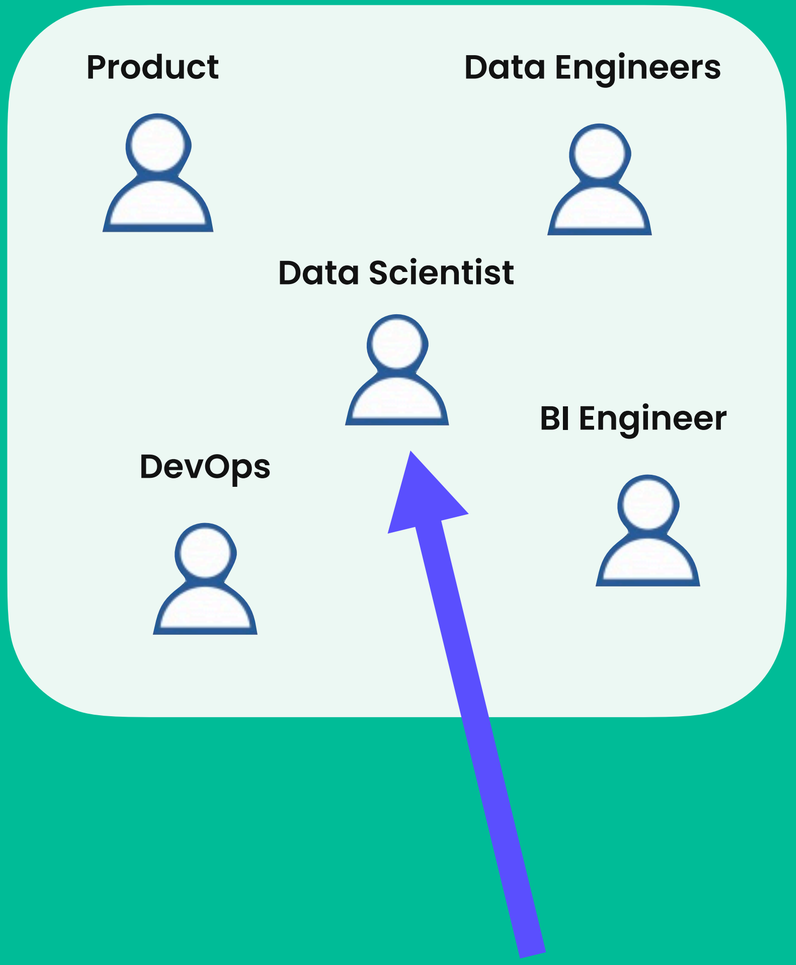
Company Numbers

- **150** Employees
- **20** Engineers
- **1** Data Scientist

Tools

Key Responsibilities:

- Infrastructure
- Models / Optimization
- Experimental Design
- Education



Only DS in the company

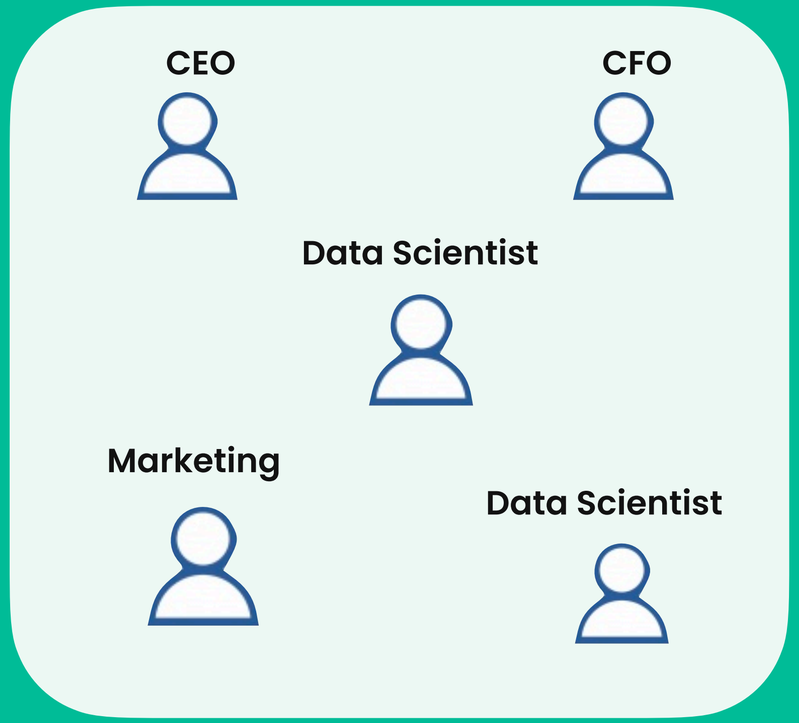
Company Numbers

- **35** Employees
- **7** Engineers
- **4** Data Scientist

Tools

Key Responsibilities:

- Infrastructure
- Dashboard
- R packages
- Models



Company Numbers

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

Tool

python™

R

docker

OPENSHIFT

Key Responsibilities:

OS Developer

Developer Advocate

Data Scientist

Researchers

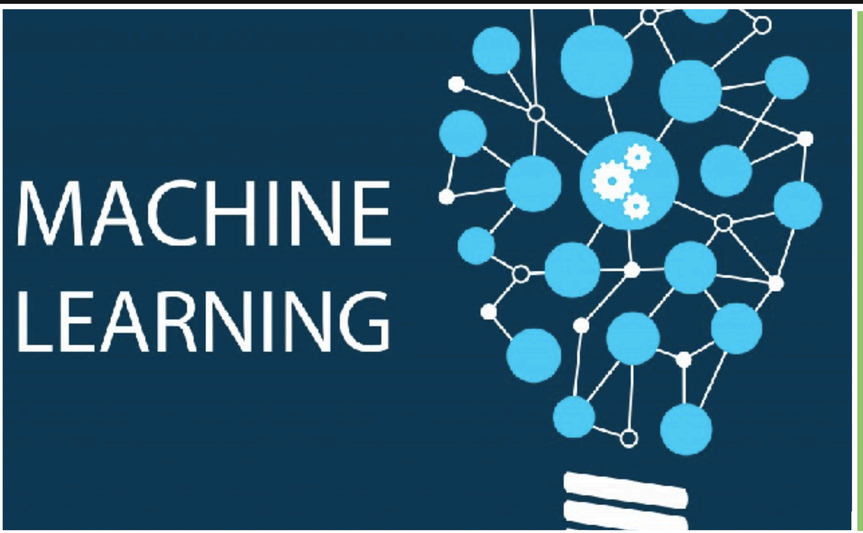
Data Scientist

AI Ethics

Data Science requires a **team of players** in different positions

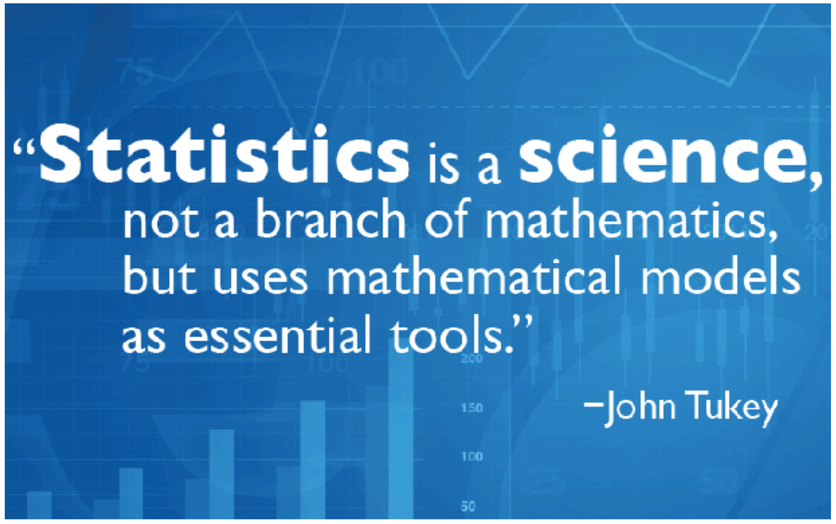
Image: www.freepik.com/vectors/background

IMPORTANT SKILLS FOR A DATA SCIENTIST



Machine Learning

Statistics



Programming

Communication

Curiosity
(keep asking why)

Flexibility

Critical Thinking

Ethics

Be yourself

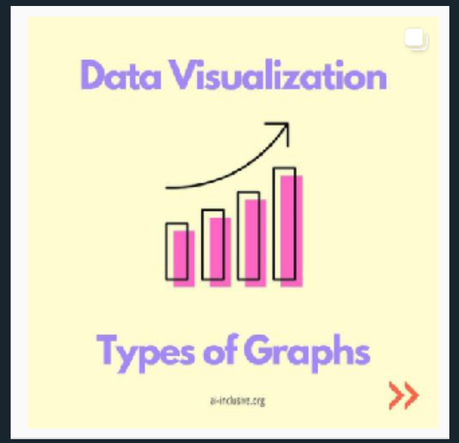
Data Science is a Team Work

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

Thank you!

@gdequeiroz | linktr.ee/gdq

Follow us: bit.ly/ai-inclusive-instagram



AI Inclusive
ai-inclusive.org

ai-inclusive.org

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