

# Data Science as a Team Sport

**Gabriela de Queiroz**

Sr. Machine Learning Manager, IBM

@gdequeiroz | [linktr.ee/gdq](https://linktr.ee/gdq)



# Gabriela de Queiroz

Sr. Machine Learning Manager, IBM California

- Founder of **R-Ladies** ([rladies.org](http://rladies.org))
- Founder of **AI Inclusive** ([ai-inclusive.org](http://ai-inclusive.org))
- Member of **R Foundation** ([r-project.org](http://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](http://rladies.org)

198

Chapters



56

Countries

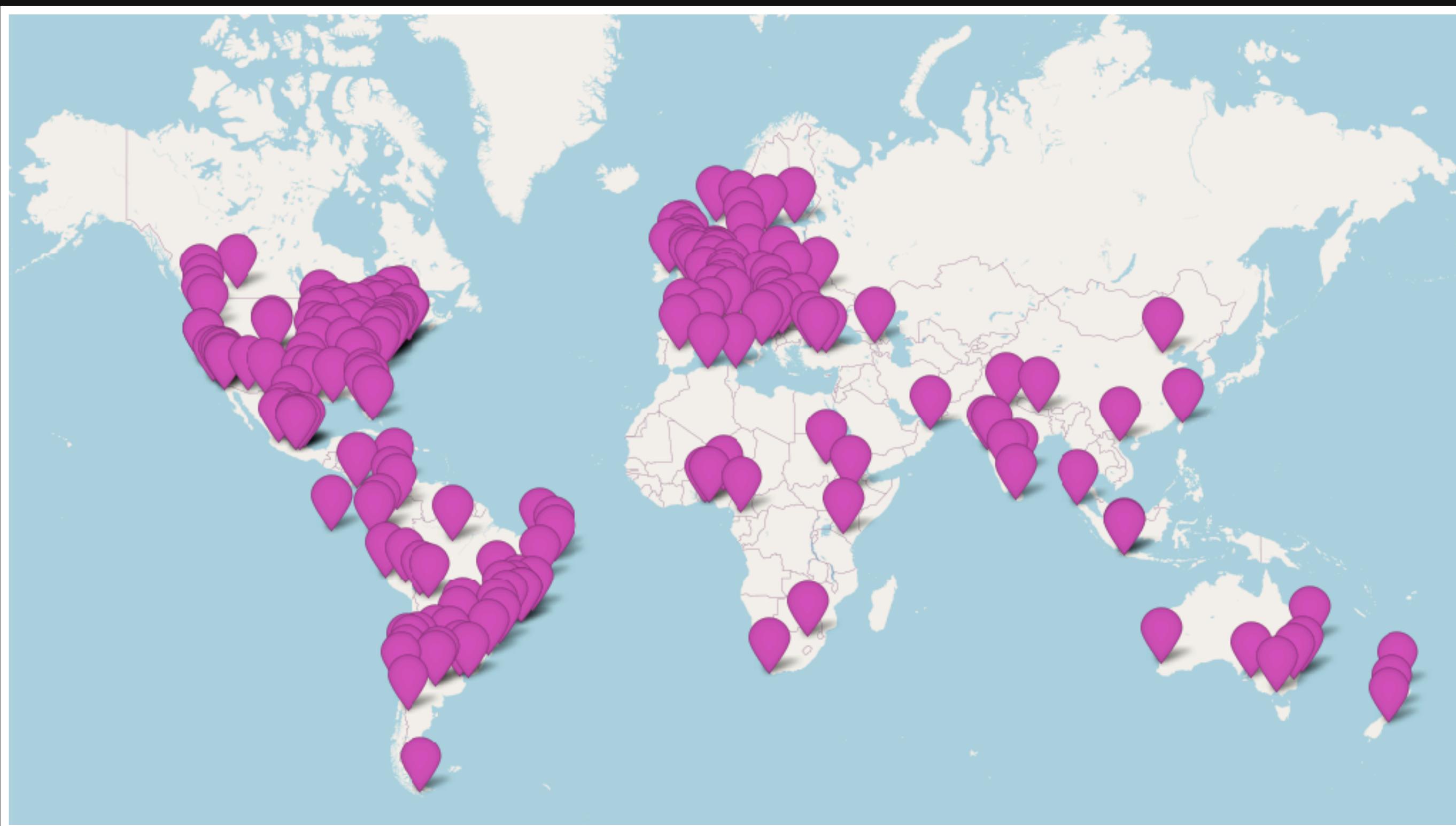


83614

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.

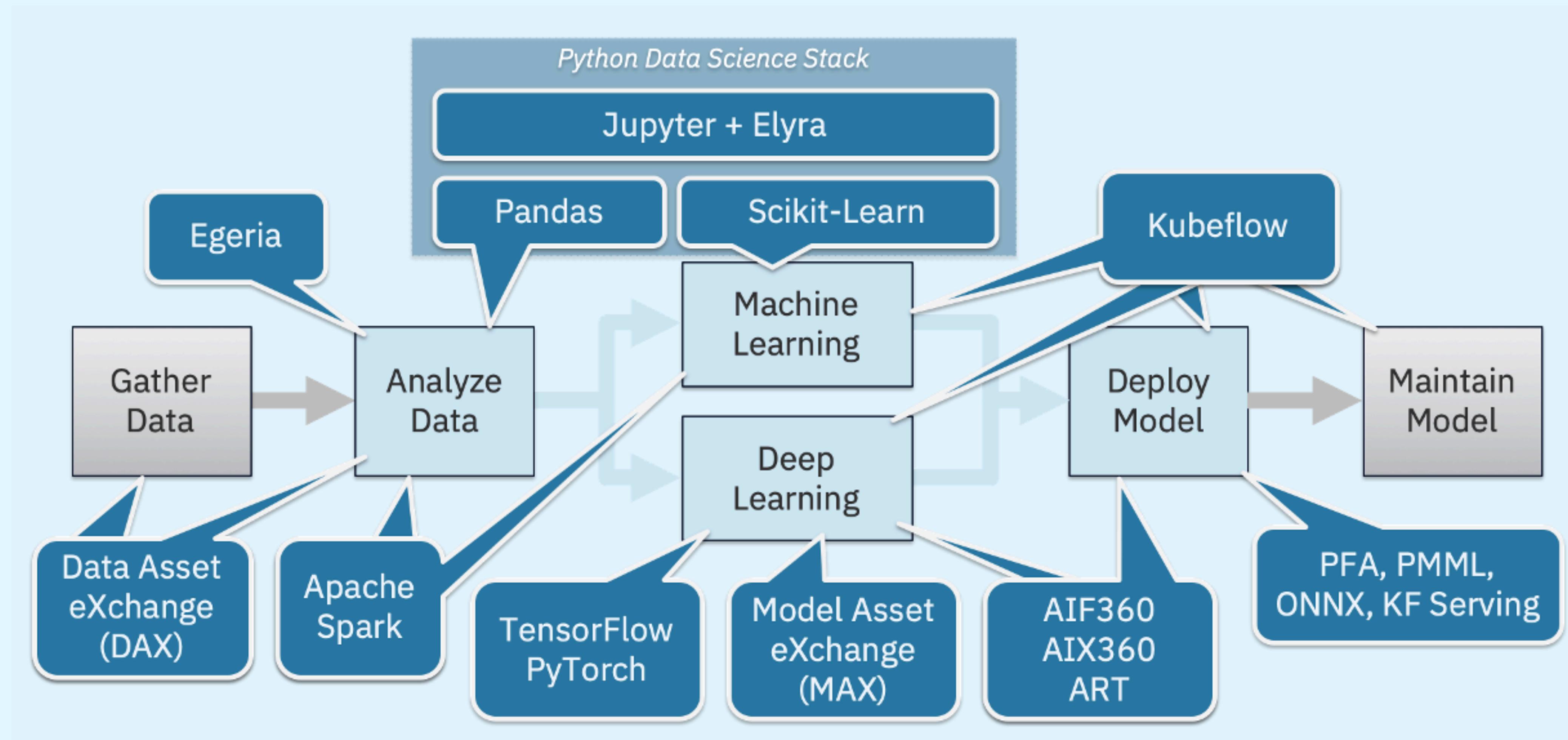
**Mission:** Increase the representation and participation of minority groups in Artificial Intelligence

- Website: [ai-inclusive.org](http://ai-inclusive.org)
- Twitter: [bit.ly/ai-inclusive-twitter](https://bit.ly/ai-inclusive-twitter)
- Instagram: [bit.ly/ai-inclusive-instagram](https://bit.ly/ai-inclusive-instagram)



# We build tools to make AI accessible and available to everybody

([codait.org](http://codait.org))



Open Source @ IBM

# Some Projects

# Model Asset eXchange (MAX)

**Website:**  
[ibm.biz/model-exchange](http://ibm.biz/model-exchange)

## Model Asset eXchange

Try the tutorial



Join the community



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

[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

[View all models](#)

# 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.  
Mar 29, 2019 →

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 →

# ibm.biz/model-exchange

Model Deployable, Trainable

## Object Detector

Localize and identify multiple objects in a single image.

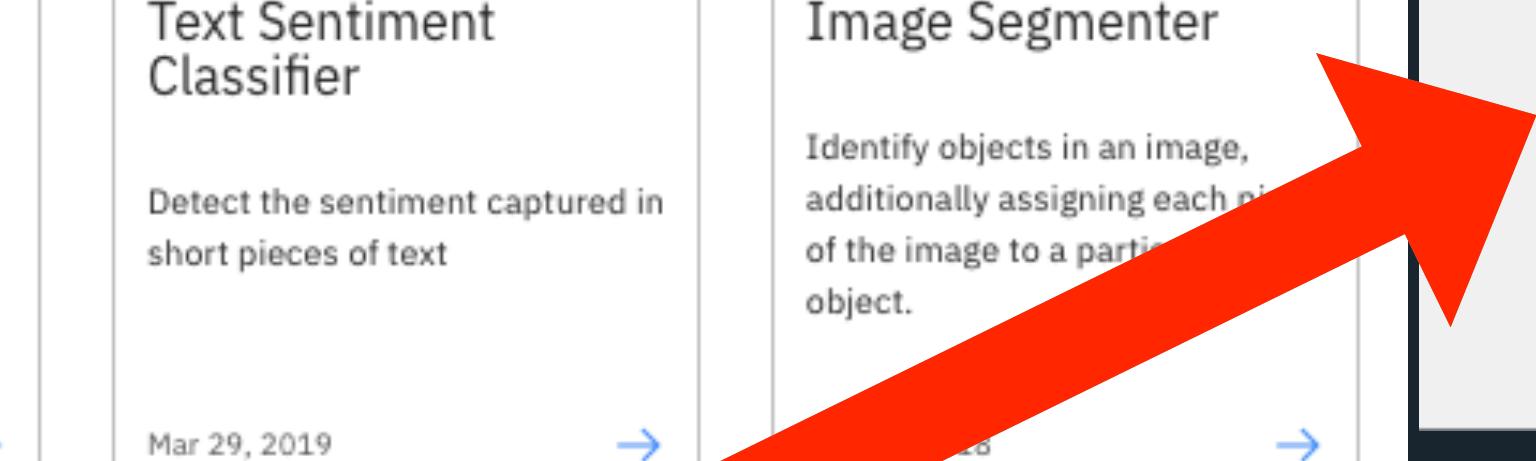
Get this model



Try the API →

Try the web app →

Try in a Node-RED flow →



# Data Asset eXchange (DAX)

**Website:**  
[ibm.biz/data-exchange](http://ibm.biz/data-exchange)

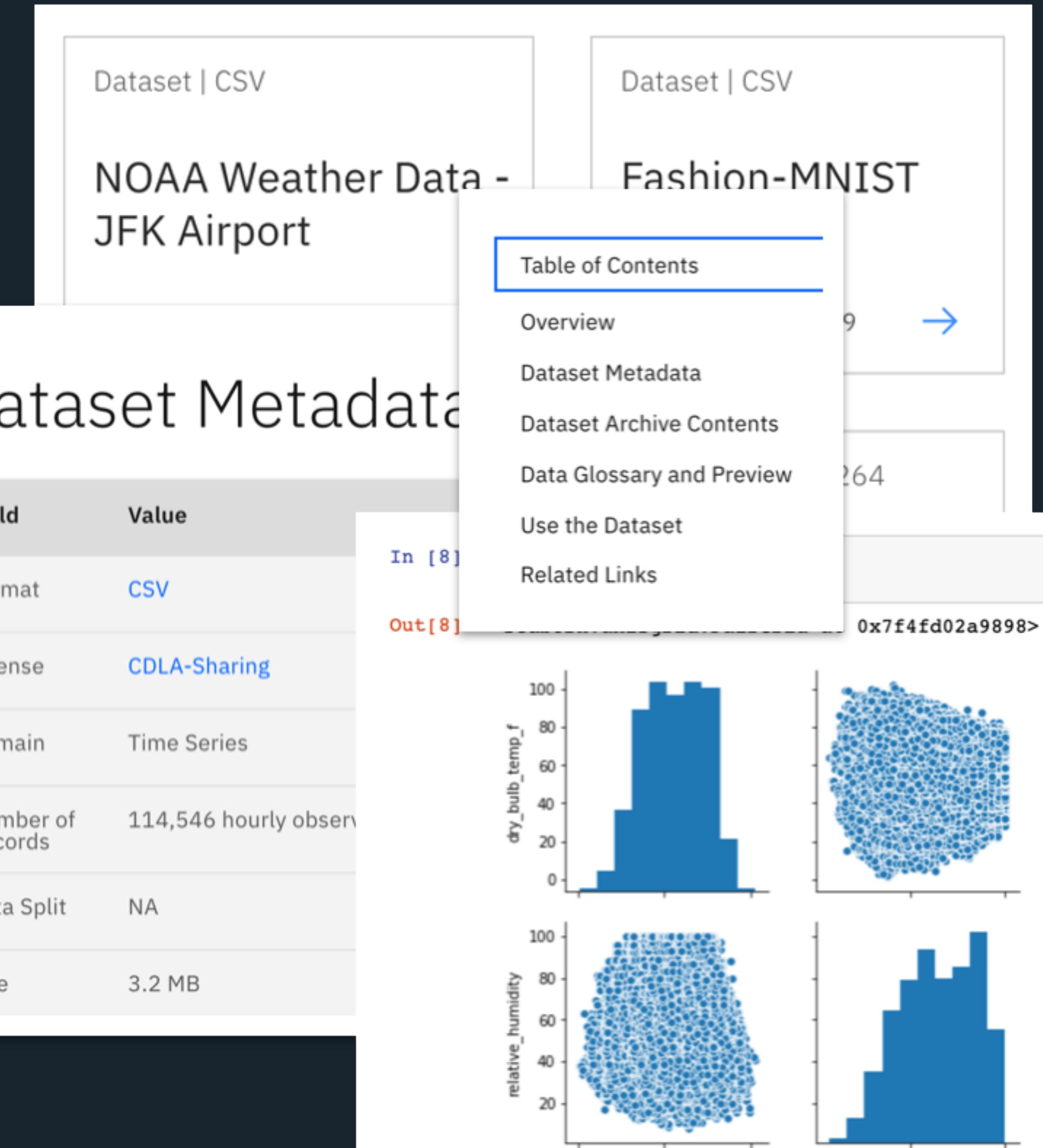
## Data Asset eXchange

Explore useful and relevant data sets for enterprise data science

Dataset   CSV NOAA Weather Data - JFK Airport August 11, 2020 →	Dataset   TSV (normal) Groningen Meaning Bank - Modified May 14, 2020 →	Dataset   CSV Fashion-MNIST August 17, 2020 →
Dataset   JPG, JSON PubLayNet August 15, 2020 →	Dataset   WAV TensorFlow Speech Commands September 28, 2020 →	Dataset   PNG, JSON PubTabNet August 11, 2020 →
Dataset   JSON, HDF5 Oil Reservoir Simulations August 11, 2020 →	Dataset   CoNLL-U Finance Proposition Bank August 11, 2020 →	Dataset   CoNLL-U Contracts Proposition Bank August 11, 2020 →

# 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](http://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

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	<a href="#">National Oceanic and Atmospheric Administration (NOAA)</a>
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. <i>Agriculture</i> 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. <i>Retail</i> Estimate outdoor retail foot traffic based on weather condition and temperature predictions.
Business Use Case	

Part 1 - Data Cleaning

Part 2 - Data Analysis

Part 3 - Time Series Forecasting

## NOAA Weather Data – JFK Airport

```
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:

- [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](#)

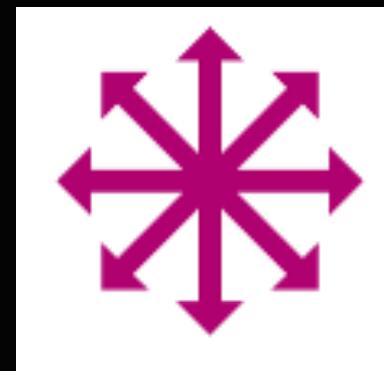
# Trusted AI Lifecycle through Open Source

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](http://art-demo.mybluemix.net)

Is it fair?



FAIRNESS

AI Fairness 360  
↳ (AIF360)

DEMO: [aif360.mybluemix.net](http://aif360.mybluemix.net)

Is it easy to understand?

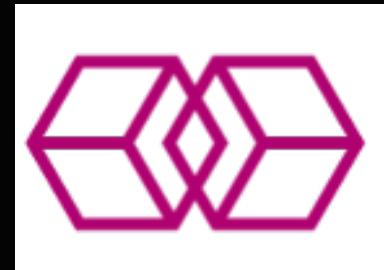


EXPLAINABILITY

AI Explainability 360  
↳ (AIX360)

DEMO: [aix360.mybluemix.net](http://aix360.mybluemix.net)

Is it accountable?



LINEAGE

AI FactSheets 360

DEMO: [aifs360.mybluemix.net](http://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

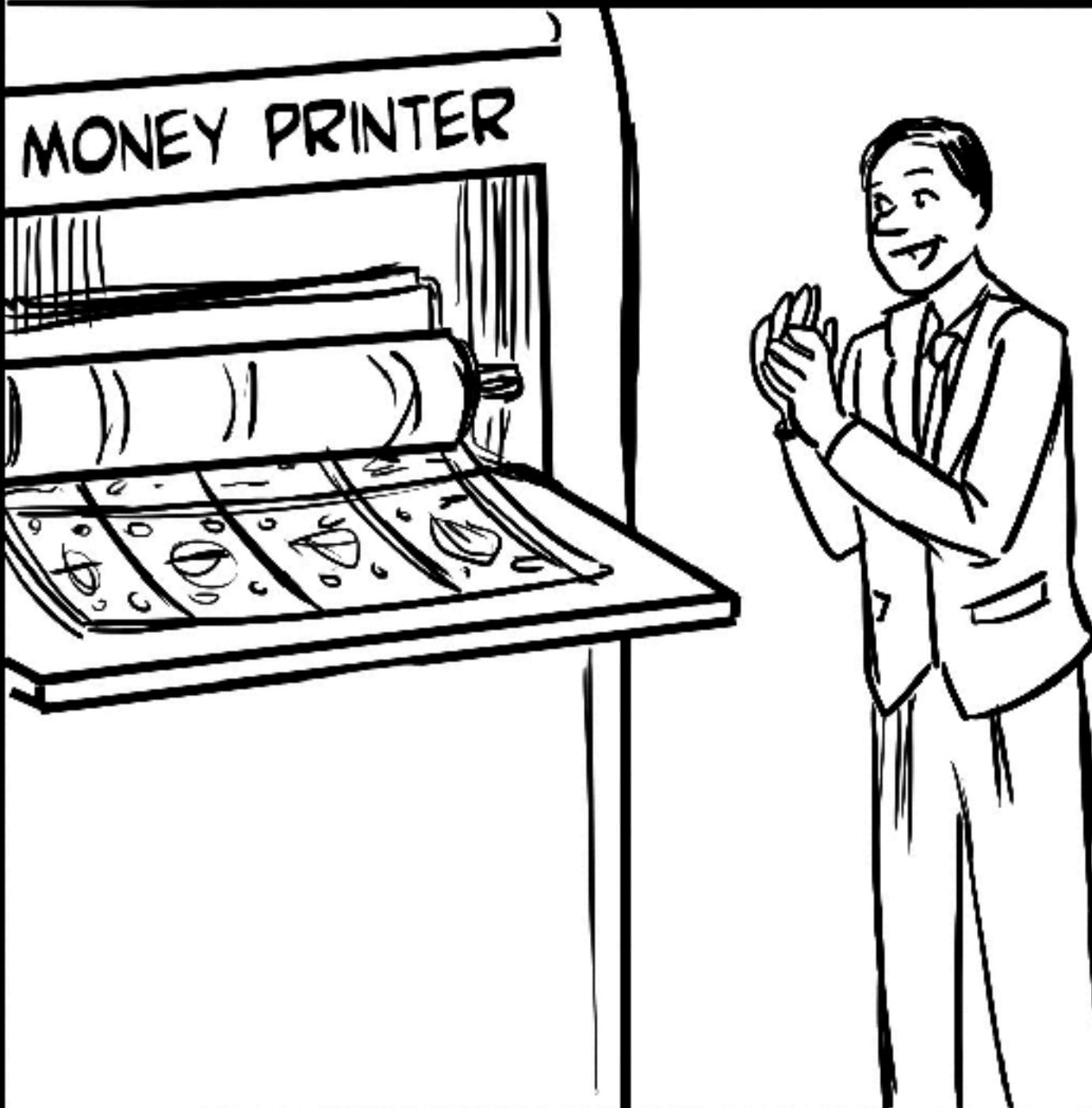
**AIX360**  
Interpretability and explainability of data and machine learning models  
Python 621 136

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

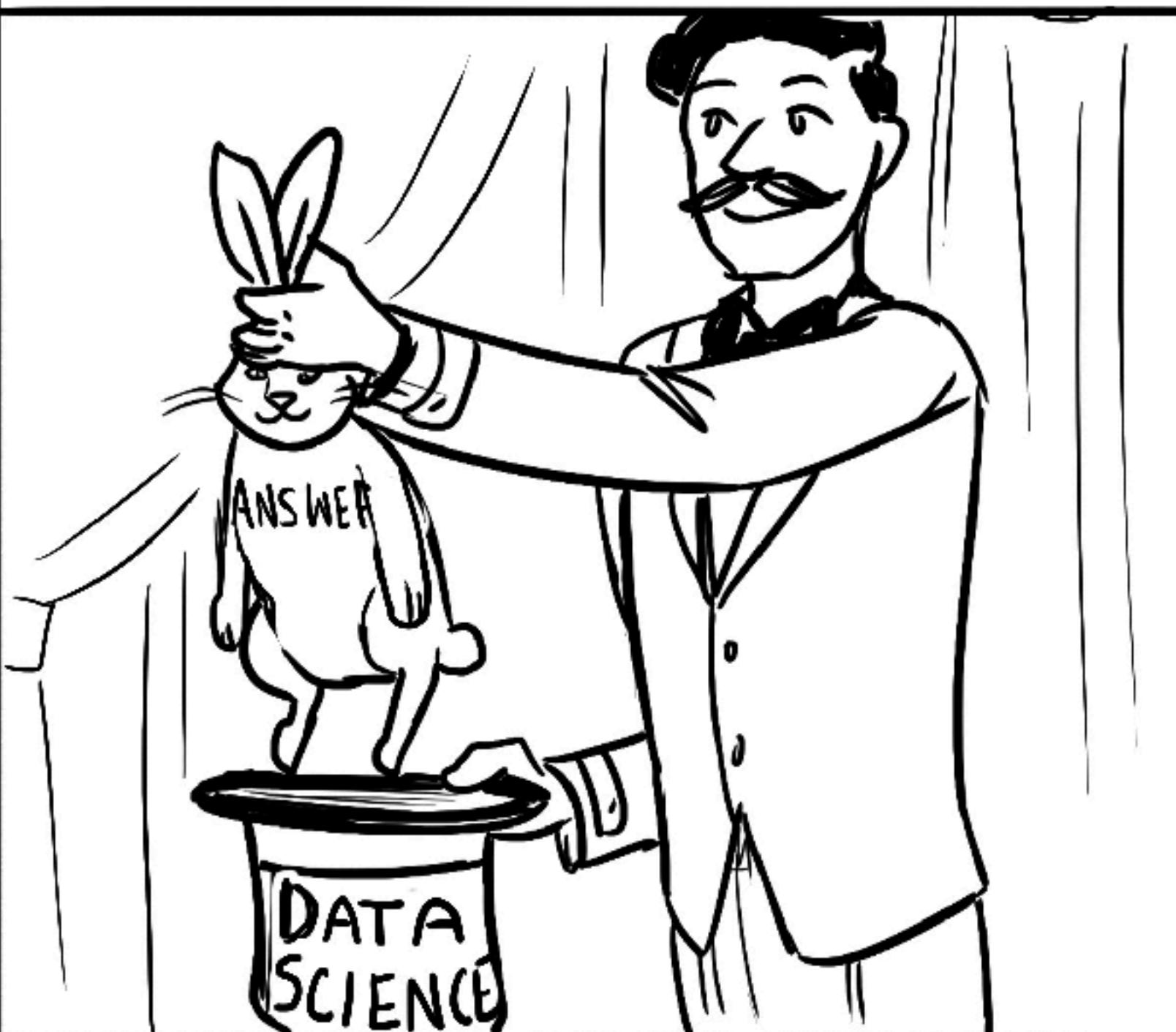
Available in R too!

# What is Data Science?

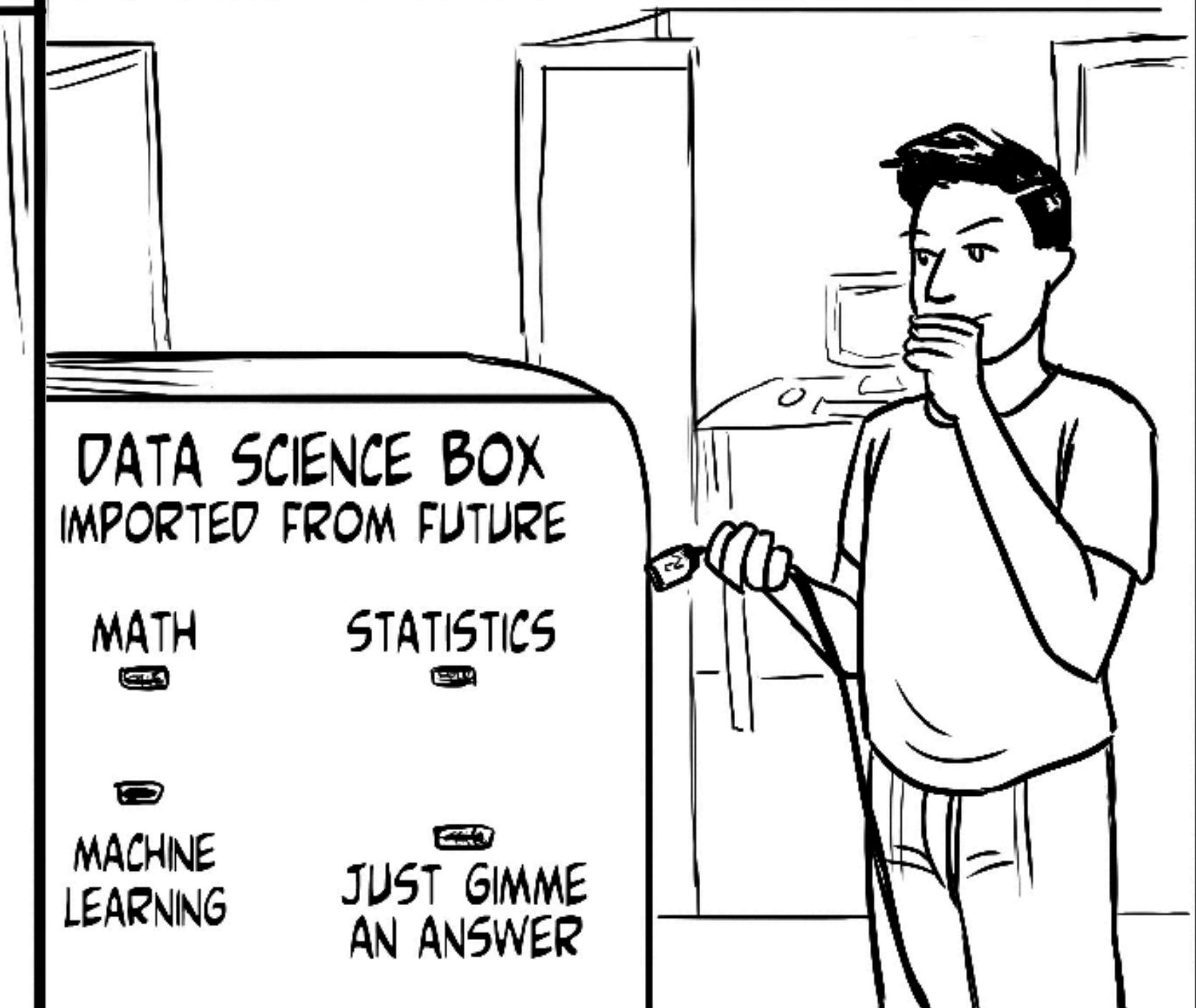
WHAT MY BOSS THINKS DATA SCIENCE IS



WHAT MY CUSTOMERS THINK DATA SCIENCE IS

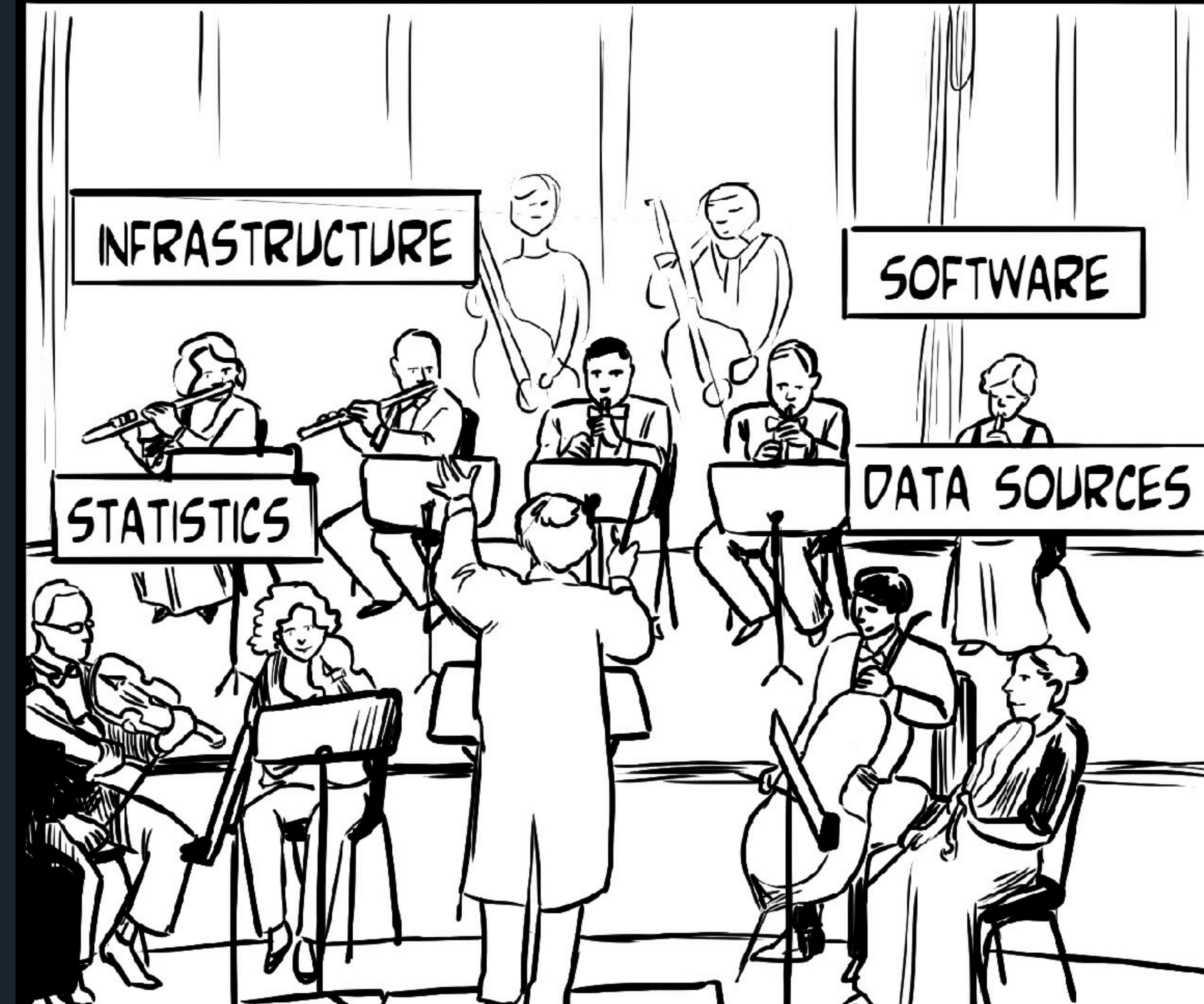


WHAT SOFTWARE ENGINEERS THINK DATA SCIENCE IS



# What is Data Science?

WHAT I THINK DATA SCIENCE IS





Data Scientists are  
**adaptable and flexible**  
professionals

# Companies

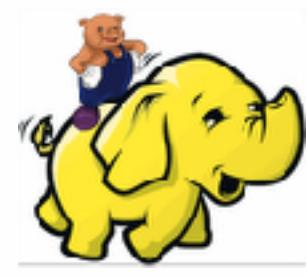
What is the role of a data scientist?

Data Scientists can have different roles in different companies

## Company Numbers

- 40 Employees
- 10 Engineers
- 5 Data Scientist

## Tools



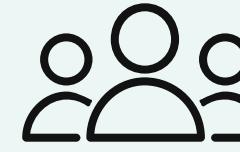
## Key Responsibilities:



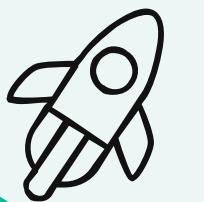
Consultant



Write Documentation



Train Customers



Prioritization

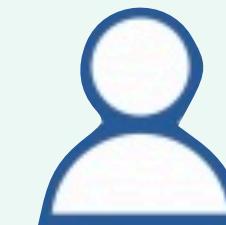
CLIENT



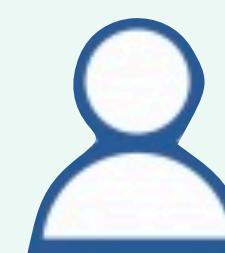
Data Scientist



Data Scientist



Product



Engineers



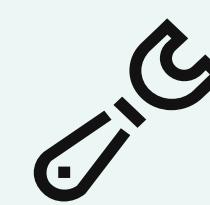
## Company Numbers

- 150 Employees
- 20 Engineers
- 1 Data Scientist

## Tools



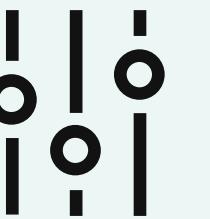
## Key Responsibilities:



Infrastructure



Models / Optimization

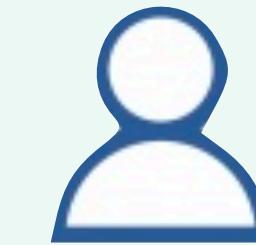


Experimental Design



Education

Product



Data Engineers



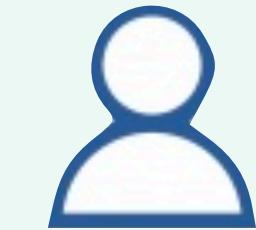
Data Scientist



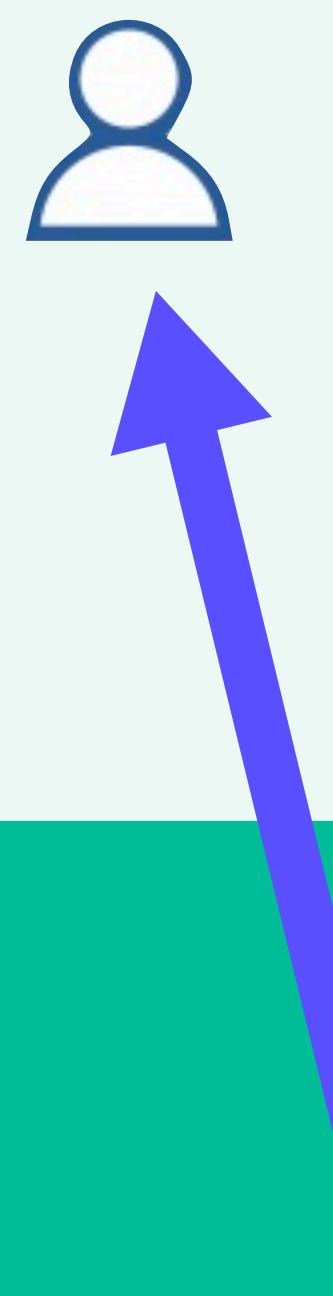
BI Engineer



DevOps



Only DS in the company



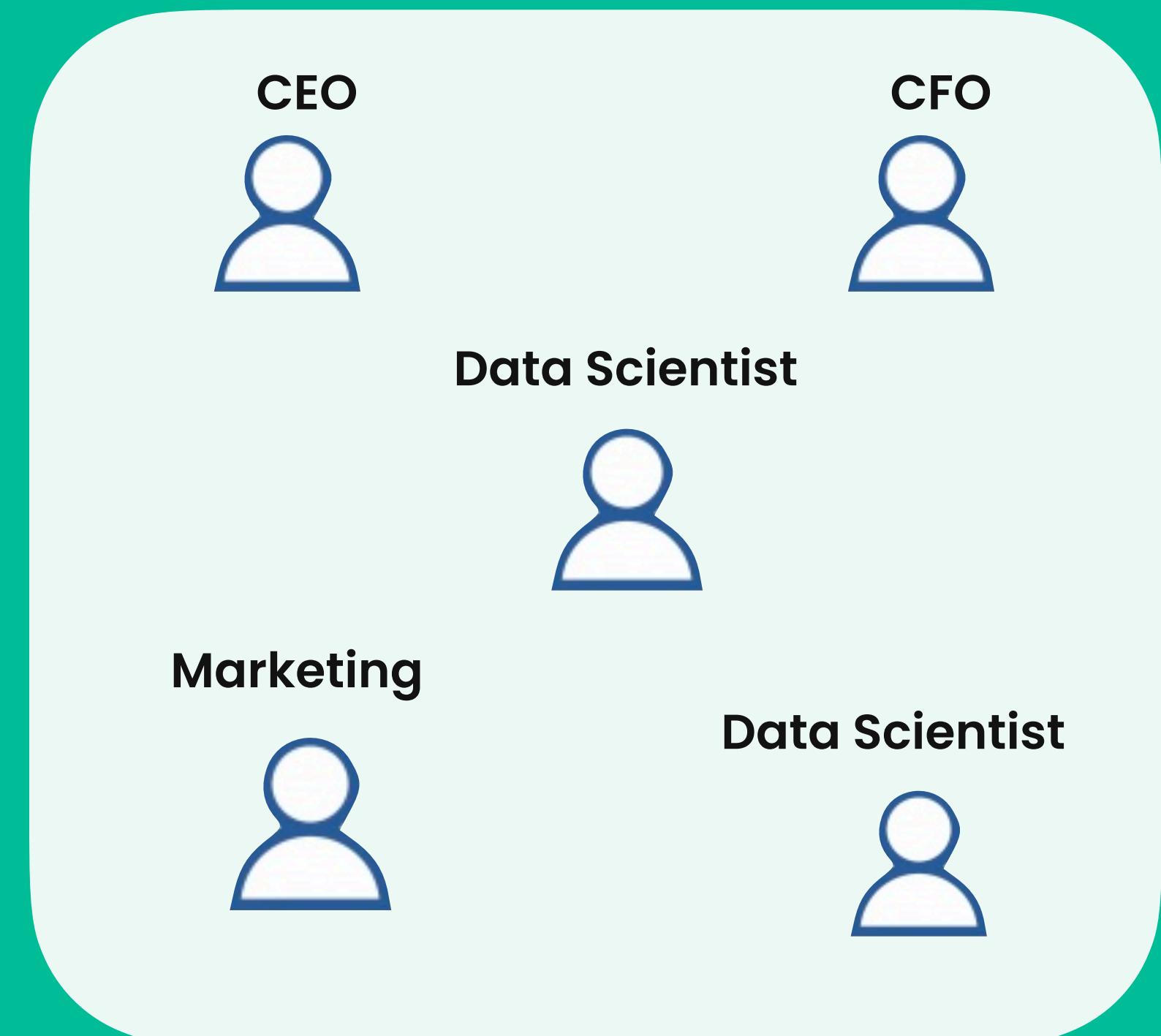
## Company Numbers

- 35 Employees
- 7 Engineers
- 4 Data Scientist



## Key Responsibilities:

- Infrastructure
- Dashboard
- R packages
- Models



## Company Numbers

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

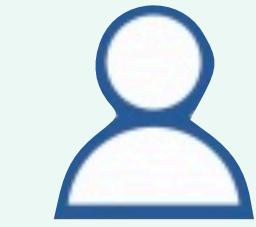
## Tool



## Key Responsibilities:



OS Developer



Developer Advocate



Data Scientist



Data Scientist



Researchers



AI Ethics



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



# 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 wp_get_favicon(); ?>
6 <?php wp_get_theme(); ?>
7 <?php wp_get_stylesheet(); ?>
8 <?php wp_get_script(); ?>
9 <?php wp_get_header(); ?>
10 <?php wp_get_sidebar(); ?>
11 <?php wp_get_content(); ?>
12 <?php wp_get_footer(); ?>
13 <?php wp_get_scripts(); ?>
14 <?php wp_get_styles(); ?>
15 <?php wp_get_menus(); ?>
16 <?php wp_get_themes(); ?>
17 <?php wp_get_plugins(); ?>
18 <?php wp_get_widgets(); ?>
19 <?php wp_get_posts(); ?>
20 <?php wp_get_comments(); ?>
21 <?php wp_get_terms(); ?>
22 <?php wp_get_taxonomies(); ?>
23 <?php wp_get_categories(); ?>
24 <?php wp_get_tags(); ?>
25 <?php wp_get_pages(); ?>
26 <?php wp_get_menus(); ?>
27 <?php wp_get_themes(); ?>
28 <?php wp_get_plugins(); ?>
29 <?php wp_get_widgets(); ?>
30 <?php wp_get_posts(); ?>
31 <?php wp_get_comments(); ?>
32 <?php wp_get_terms(); ?>
33 <?php wp_get_taxonomies(); ?>
34 <?php wp_get_categories(); ?>
35 <?php wp_get_tags(); ?>
36 <?php wp_get_pages(); ?>
```

A dark-themed code editor window displaying a block of PHP code. The code includes various WordPress-related functions like wp\_head(), wp\_title(), and wp\_get\_favicon(). The code is color-coded for syntax highlighting, with red for tags, green for strings, and blue for variables.

# 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.

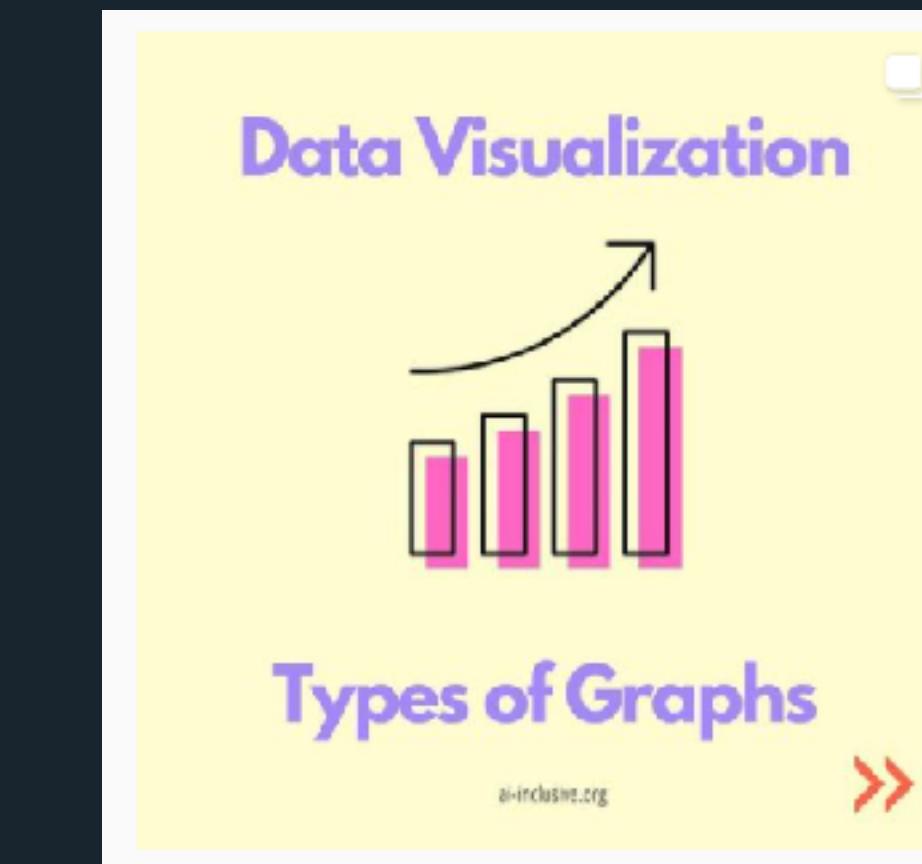
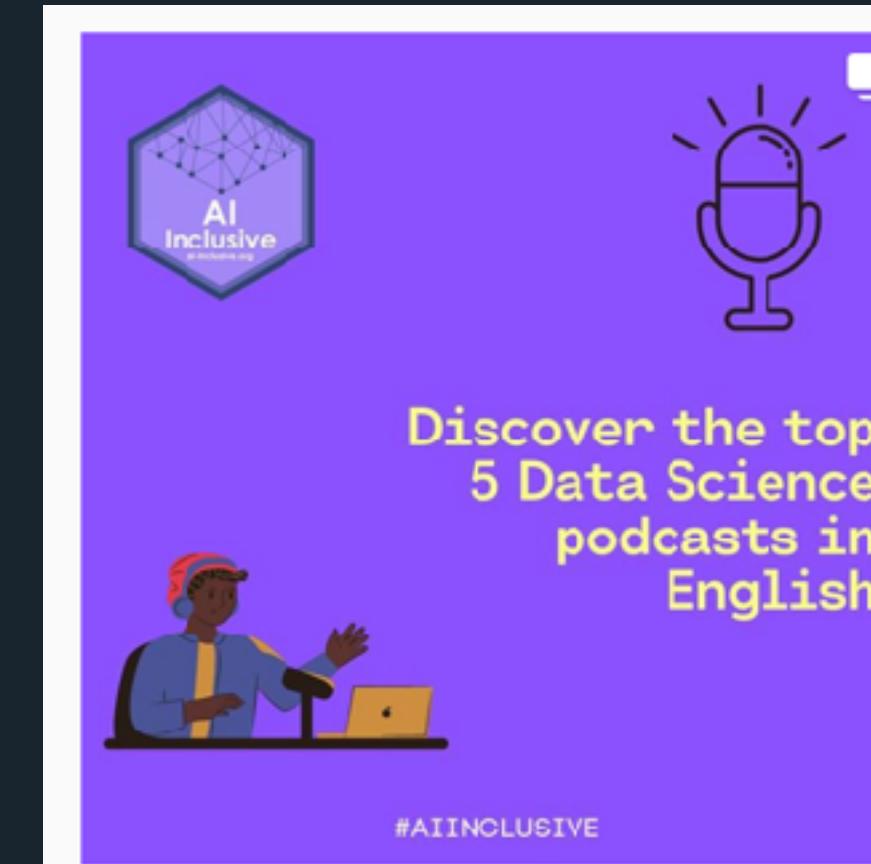


# Thank you!



@gdequeiroz | linktr.ee/gdq

Follow US: [bit.ly/ai-inclusive-instagram](https://bit.ly/ai-inclusive-instagram)



ai-inclusive.org

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