Análisis de datos con Python

Sivana Hamer - sivana.hamer@ucr.ac.cr
Escuela de Ciencias de la Computación, Universidad de Costa Rica

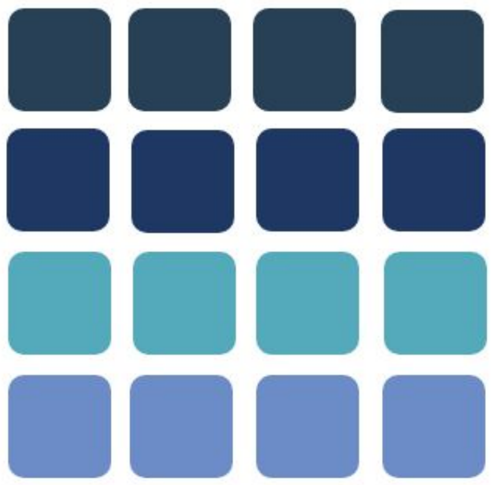
“Un dato es una **representación simbólica** (numérica, alfabética, algorítmica, espacial, etc.) **de un atributo o variable cuantitativa o cualitativa**. Los datos describen hechos empíricos, sucesos y entidades. Es un valor o referente que recibe el computador por diferentes medios, los datos representan la información que el programador manipula en la construcción de una solución o en el desarrollo de un algoritmo.” @Wikipedia

Utilizamos y generamos datos todo el tiempo

https://hipaatrek.com/6-steps-hipaa-compliant-social-media/

Existen distintos tipos de datos…

Estructurado



- Lista de números de teléfono
- Temperatura en los distritos

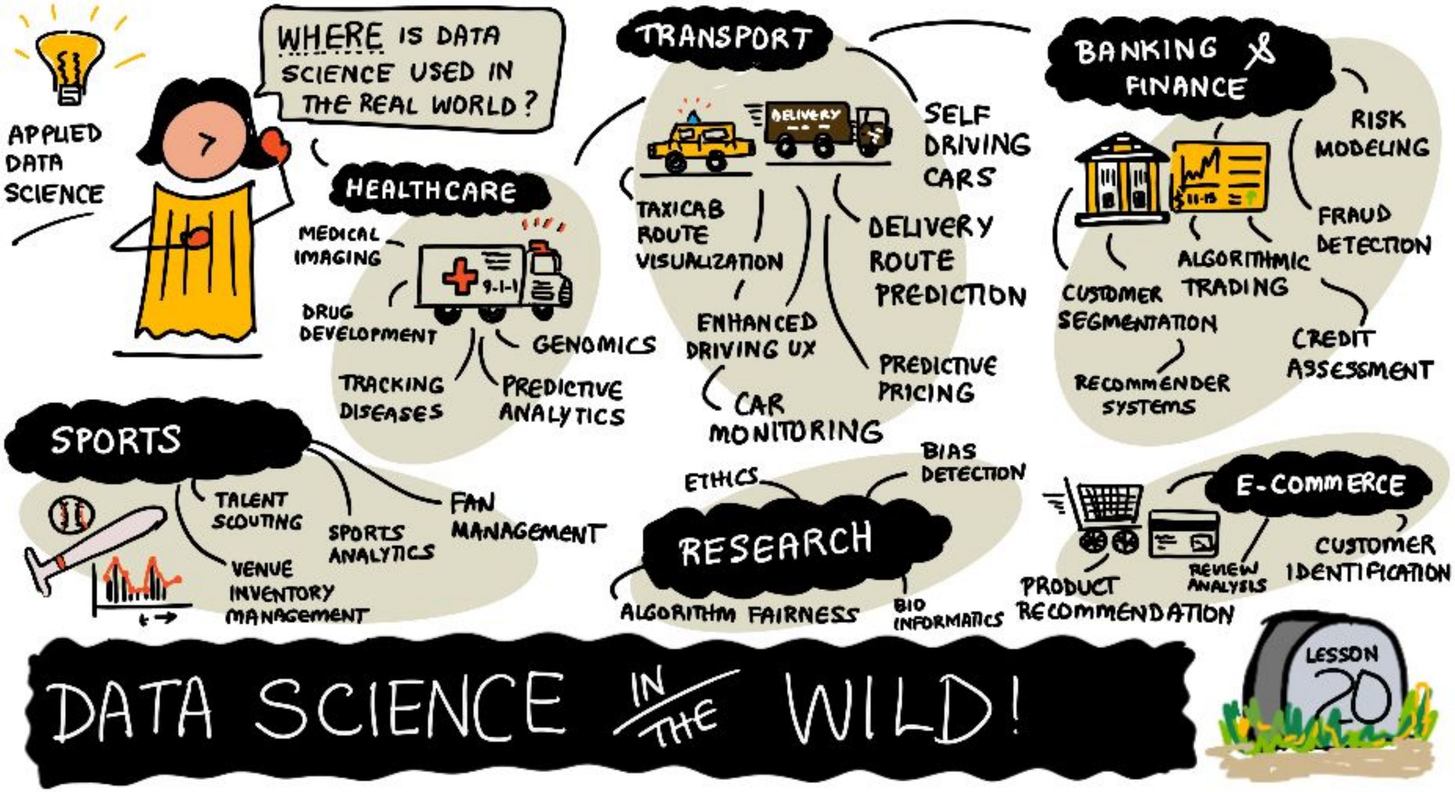
Desestructurada

- Diccionarios
- Imagenes de tráfico

“El análisis de datos es un **proceso** que consiste en inspeccionar, limpiar y transformar datos con el **objetivo de resaltar información útil**, para sugerir conclusiones y apoyo en la toma de decisiones. El análisis de datos tiene múltiples facetas y enfoques, que abarca diversas técnicas en una variedad de nombres, en diferentes negocios, la ciencia, y los dominios de las ciencias sociales.” @Wikipedia

“La ciencia de datos es un **campo** interdisciplinario que involucra métodos **científicos**, procesos y sistemas **para extraer conocimiento o un mejor entendimiento de datos** en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático, y la analítica predictiva.” @Wikipedia

Se puede aplicar para muchas áreas…



https://github.com/microsoft/Data-Science-For-Beginners

Tiene relación con muchas otras disciplinas…

Bases de datos

Big data

Aprendizaje automático (*machine learning*)

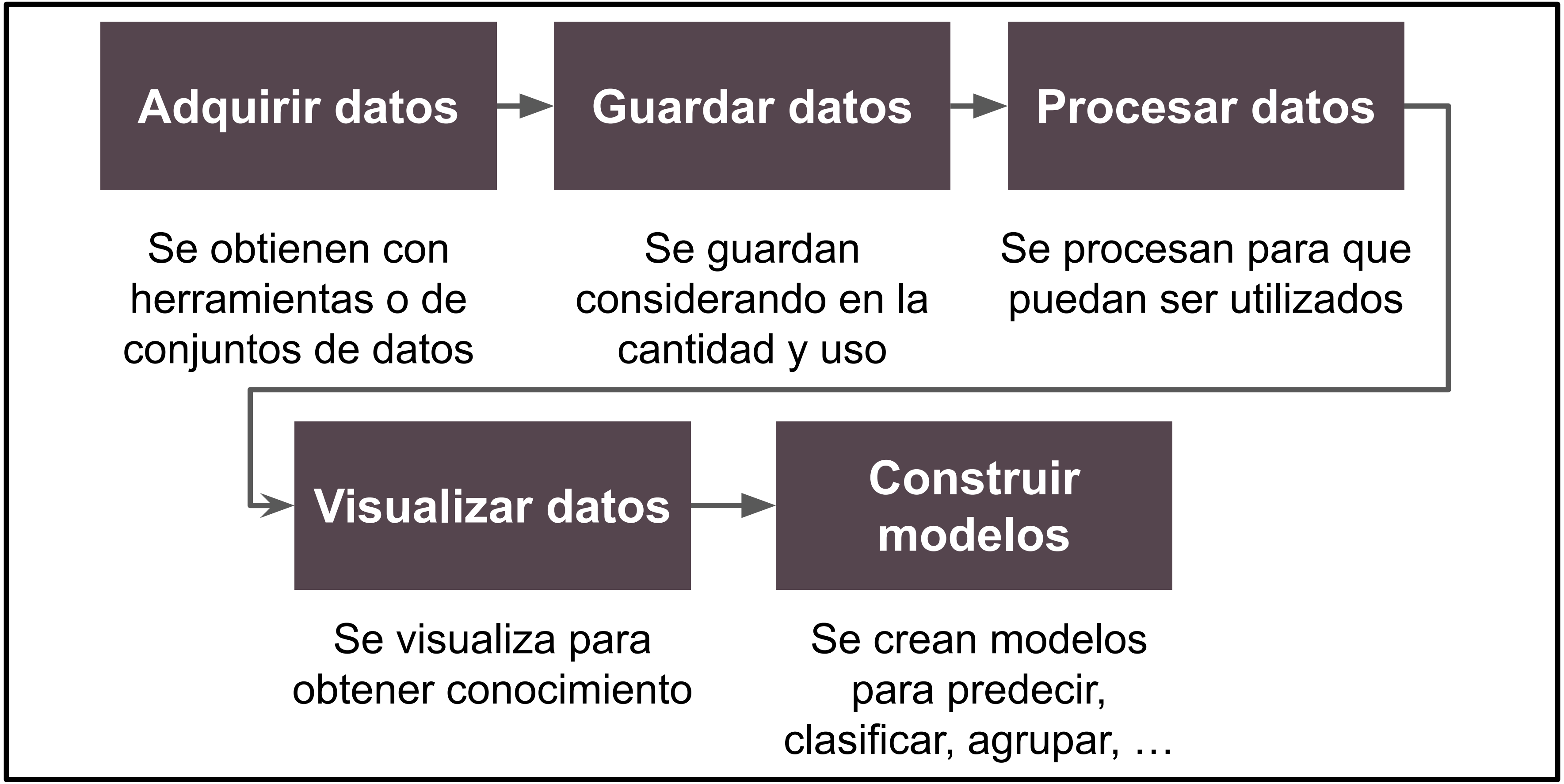
Estadística

Inteligencia artificial

Visualización

…

Con el objetivo de obtener información útil



Adquirir datos

Se obtienen con herramientas o de conjuntos de datos

Guardar datos

Se guardan considerando en la cantidad y uso

Procesar datos

Se procesan para que puedan ser utilizados

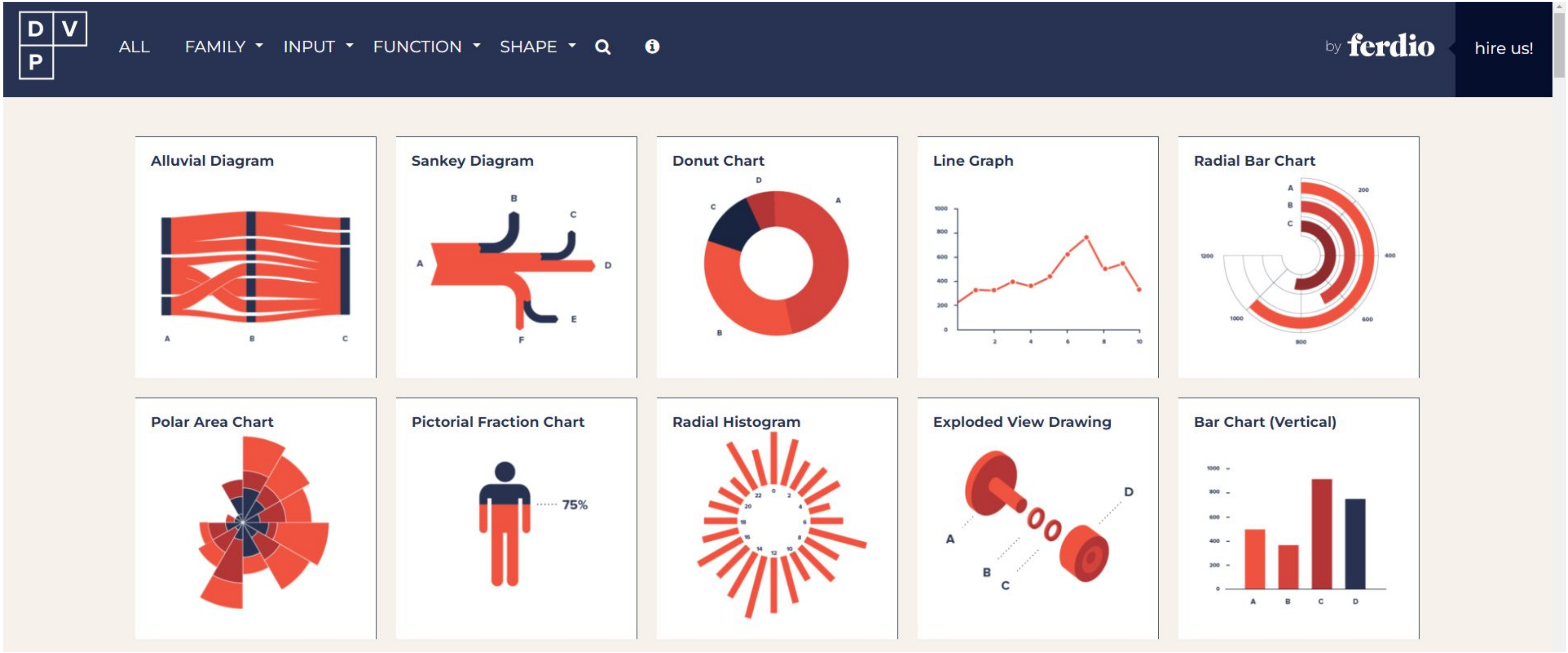
Visualizar datos

Se visualiza para obtener conocimiento

Construir modelos

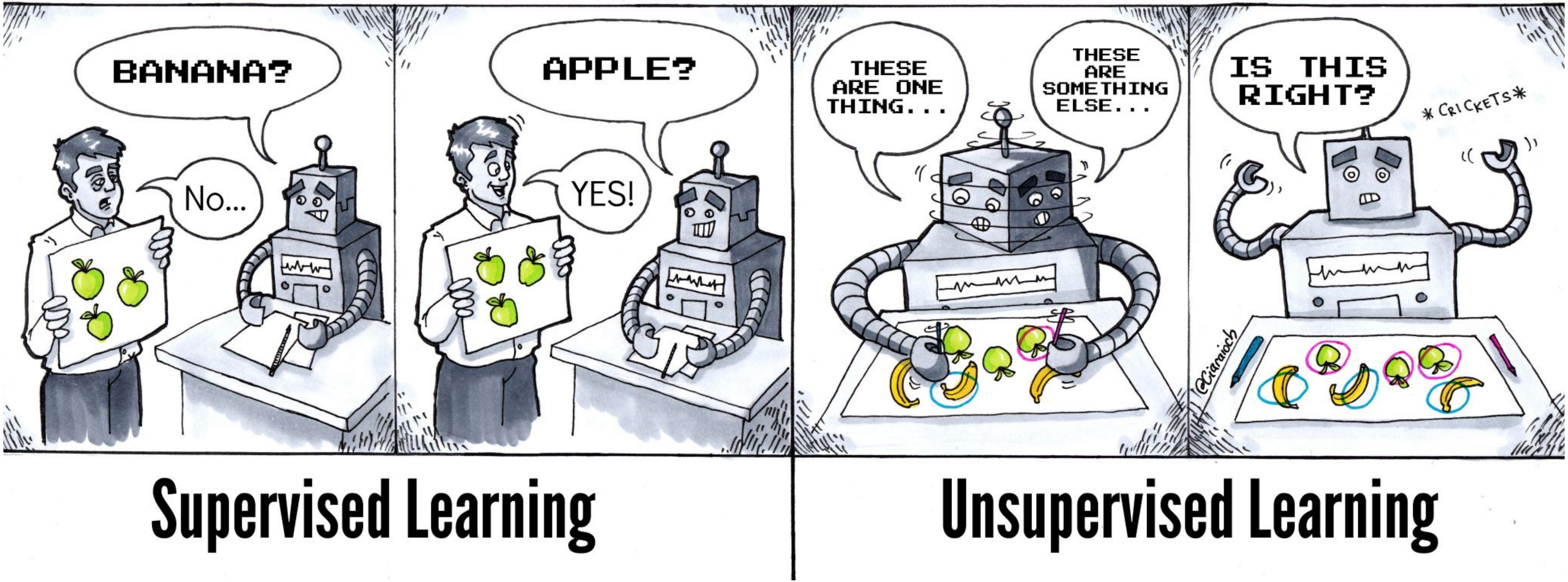
Se crean modelos para predecir, clasificar, agrupar, …

Existen muchas maneras en que se pueden visualizar datos



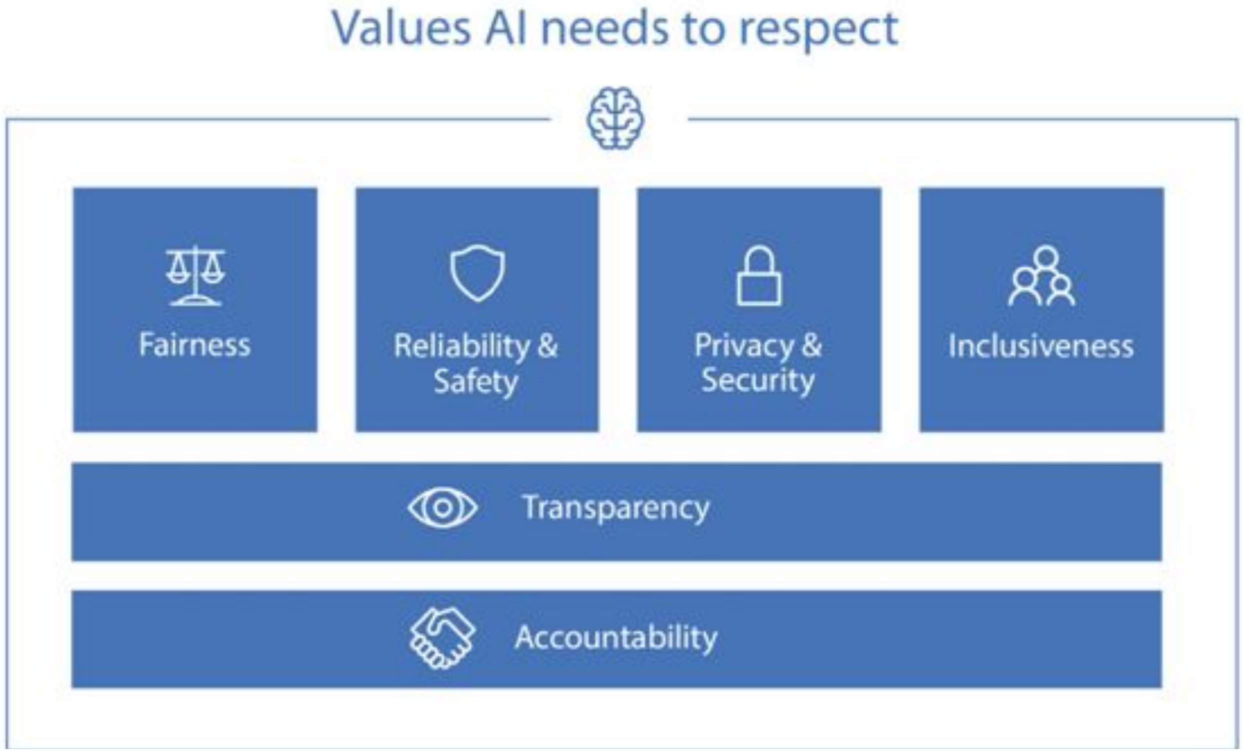
https://datavizproject.com/

Hay tipos de técnicas de aprendizaje automatizado…



https://twitter.com/athena_schools/status/1063013435779223553

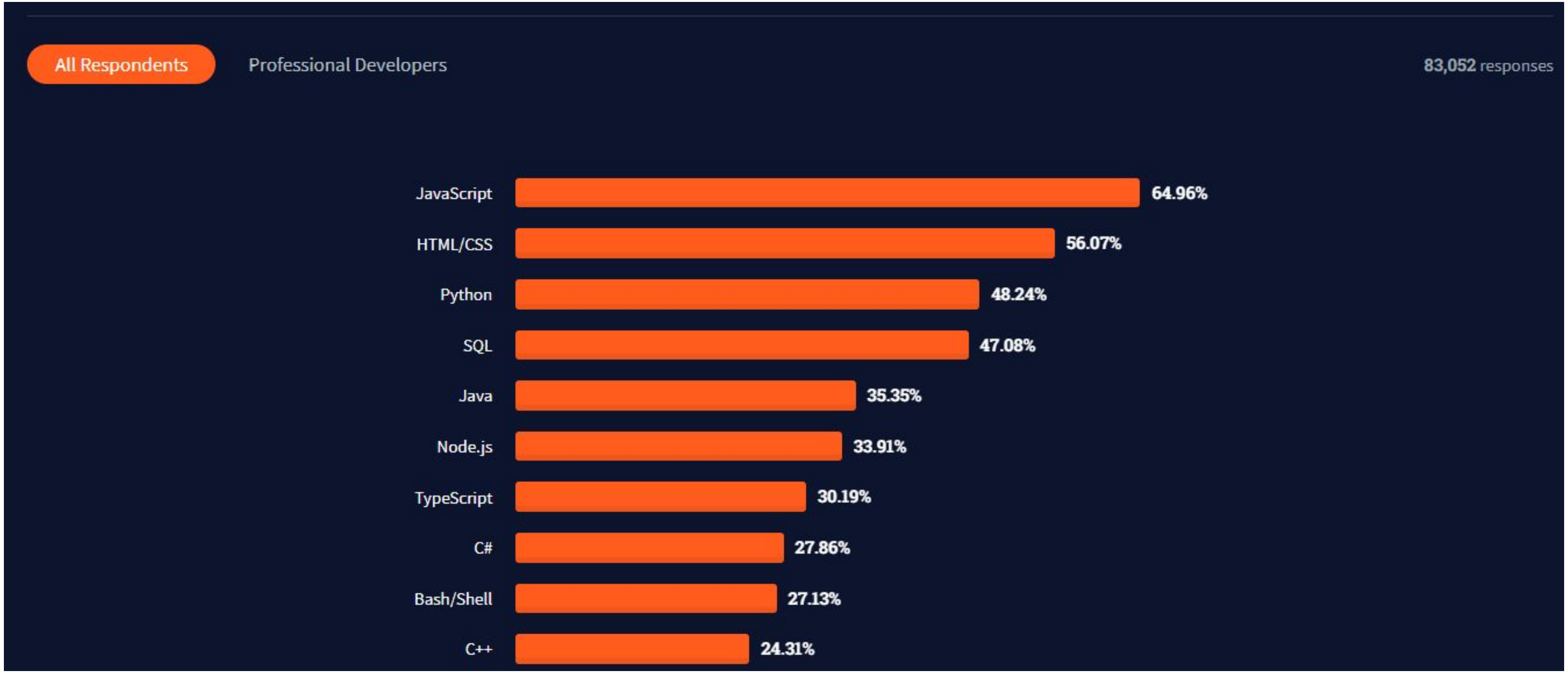
Hay que tener ciertas consideraciones al usar estos modelos



https://docs.microsoft.com/en-gb/azure/cognitive-services/personalizer/media/ethics-and-responsible-use/ai-values-future-computed.png

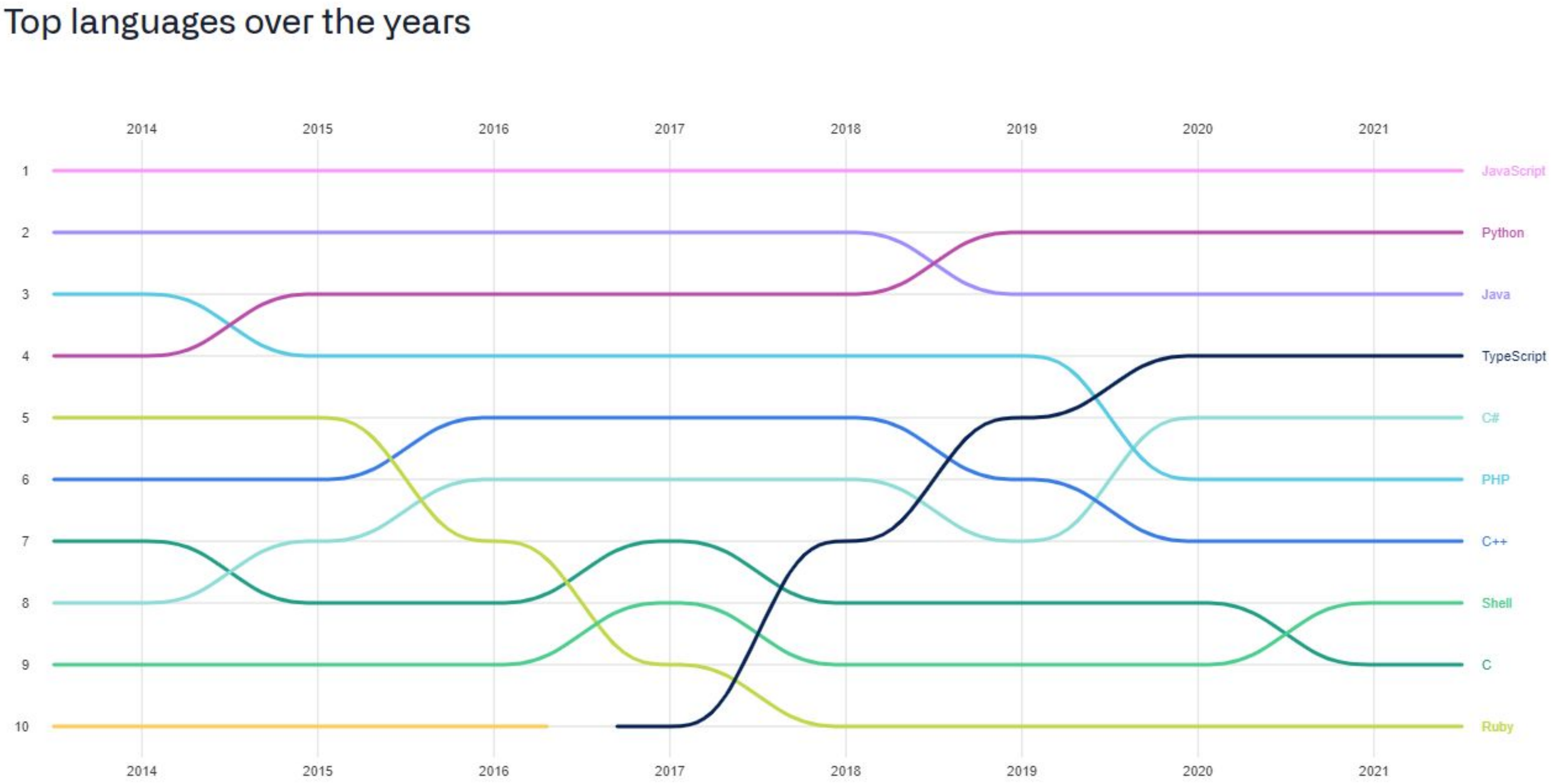
Python es un lenguaje de programación orientado a objetos

Es uno de los lenguajes más utilizados en computación



https://insights.stackoverflow.com/survey/2021

Es uno de los lenguajes más utilizados en software abierto



https://octoverse.github.com/

Python es un ecosistema de librerías

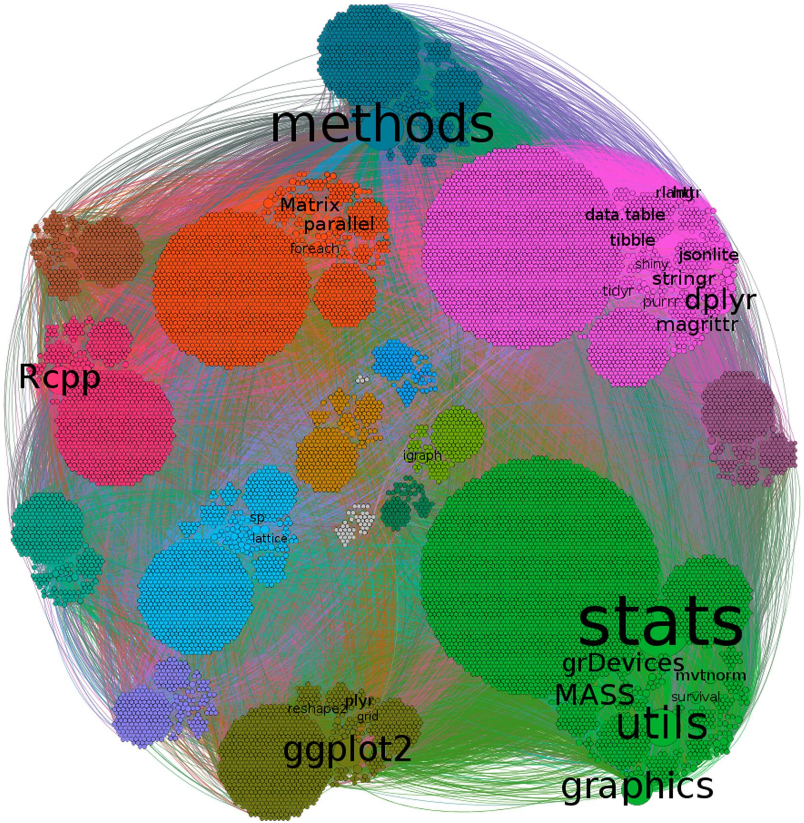


Imagen del ecosistema de R de https://www.sciencedirect.com/science/article/pii/S0164121220301709

Para el análisis de datos, hay muchas librerías…

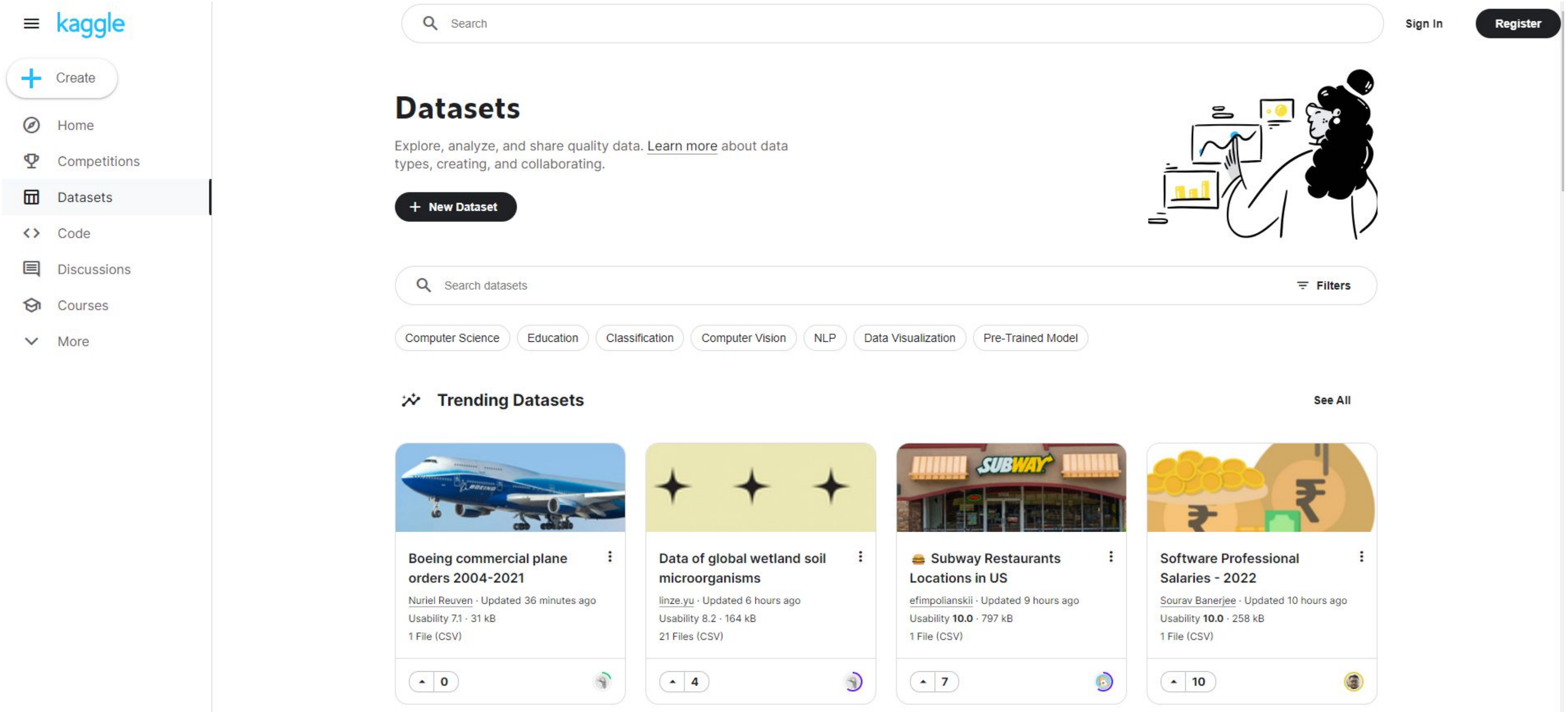
pandas

NumPy

scikit learn

matplotlib

Vamos a utilizar datos de Kaggle



Datasets

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.

Trending Datasets

Boeing commercial plane orders 2004-2021
Nuriel Reuven · Updated 36 minutes ago
Usability 7.1 · 31 kB
1 File (CSV)
0

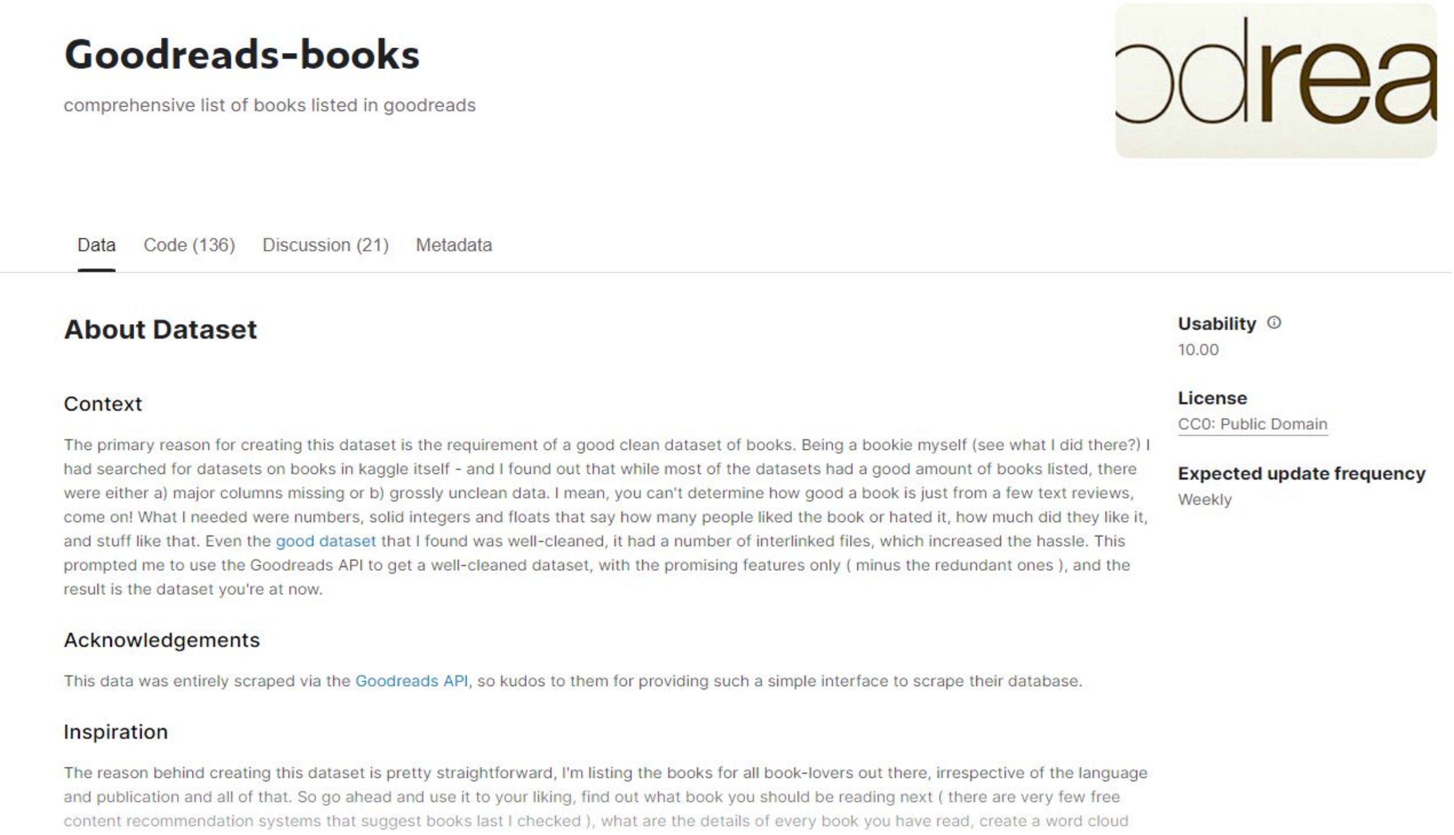
Data of global wetland soil microorganisms
linze.yu · Updated 6 hours ago
Usability 8.2 · 164 kB
21 Files (CSV)
4

Subway Restaurants Locations in US
efimpolianskii · Updated 9 hours ago
Usability 10.0 · 797 kB
1 File (CSV)
7

Software Professional Salaries - 2022
Sourav Banerjee · Updated 10 hours ago
Usability 10.0 · 258 kB
1 File (CSV)
10

https://www.kaggle.com/datasets

El primer *dataset* es sobre libros en Goodreads



Goodreads-books

comprehensive list of books listed in goodreads

About Dataset

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Weekly

Context

The primary reason for creating this dataset is the requirement of a good clean dataset of books. Being a bookie myself (see what I did there?) I had searched for datasets on books in kaggle itself - and I found out that while most of the datasets had a good amount of books listed, there were either a) major columns missing or b) grossly unclean data. I mean, you can't determine how good a book is just from a few text reviews, come on! What I needed were numbers, solid integers and floats that say how many people liked the book or hated it, how much did they like it, and stuff like that. Even the good dataset that I found was well-cleaned, it had a number of interlinked files, which increased the hassle. This prompted me to use the Goodreads API to get a well-cleaned dataset, with the promising features only (minus the redundant ones), and the result is the dataset you're at now.

Acknowledgements

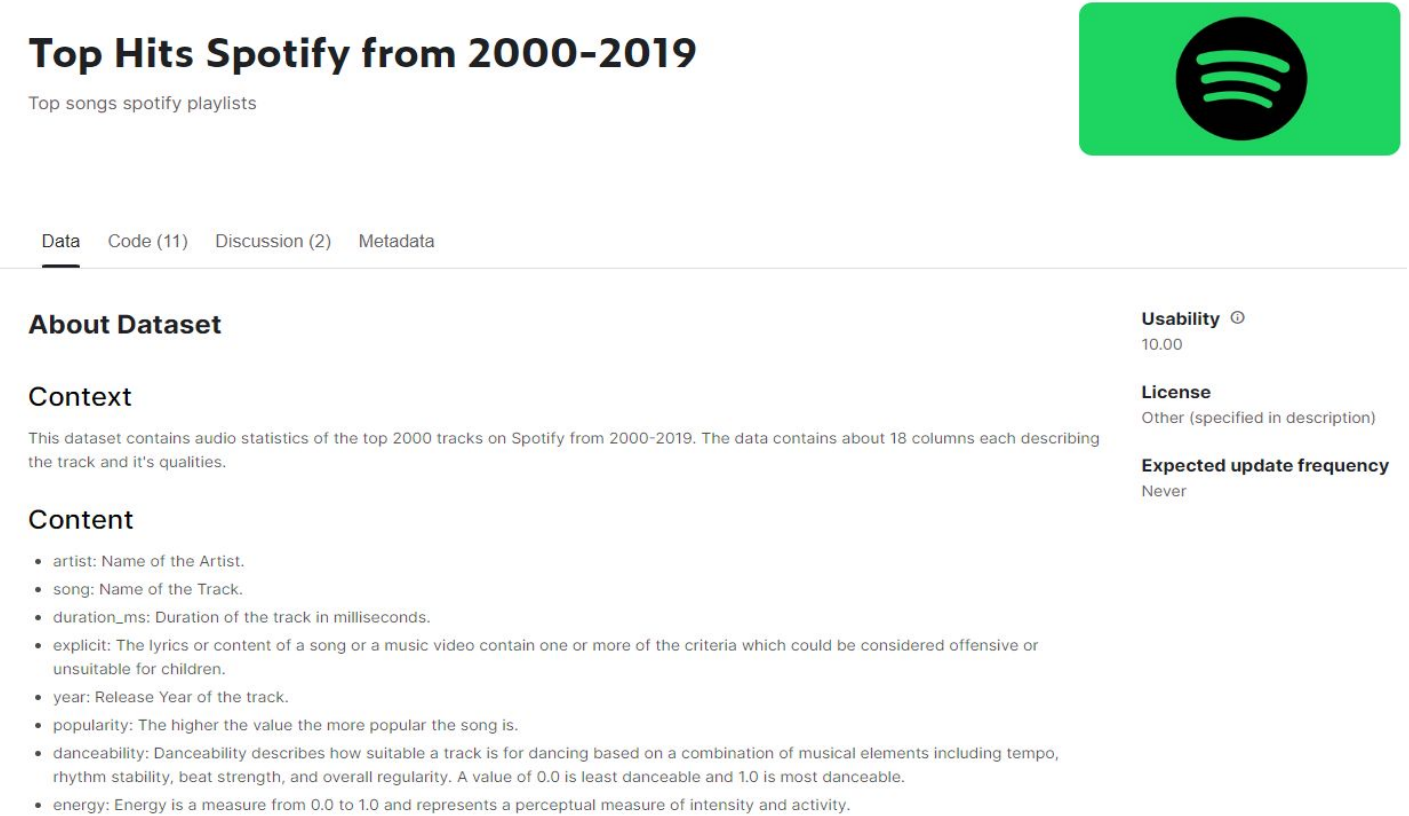
This data was entirely scraped via the Goodreads API, so kudos to them for providing such a simple interface to scrape their database.

Inspiration

The reason behind creating this dataset is pretty straightforward, I'm listing the books for all book-lovers out there, irrespective of the language and publication and all of that. So go ahead and use it to your liking, find out what book you should be reading next (there are very few free content recommendation systems that suggest books last I checked), what are the details of every book you have read, create a word cloud

https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks

El segundo *dataset* es sobre canciones de Spotify



Top Hits Spotify from 2000-2019

Top songs spotify playlists

About Dataset

Usability ⓘ
10.00

License
Other (specified in description)

Expected update frequency
Never

Context

This dataset contains audio statistics of the top 2000 tracks on Spotify from 2000-2019. The data contains about 18 columns each describing the track and it's qualities.

Content

- artist: Name of the Artist.
- song: Name of the Track.
- duration_ms: Duration of the track in milliseconds.
- explicit: The lyrics or content of a song or a music video contain one or more of the criteria which could be considered offensive or unsuitable for children.
- year: Release Year of the track.
- popularity: The higher the value the more popular the song is.
- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019

Manos
a la
obra

Referencias

- Microsoft. (2022). *Data Science for Beginners - A Curriculum.* Recuperado de: https://github.com/microsoft/Data-Science-For-Beginners
- Microsoft. (2022). *Machine Learning for Beginners - A Curriculum.* Recuperado de: https://github.com/microsoft/ML-For-Beginners
- Peng y Matsui. (2017). The Art of Data Science. Recuperado de: https://github.com/waldronlab/The-Art-of-Data-Science
- AWESOME DATA SCIENCE. (2022). Recuperado de: https://github.com/academic/awesome-datascience
- Krishnamurthy (2019). Understanding Data Bias. Recuperado de: https://towardsdatascience.com/survey-d4f168791e57