Data Gathering & *Acquisition*

Eng- Mohamed Khaled Idris
Eng- Mayar Swilam


Step -1
Collection of Data from Various source
Step -2
Data cleaning and Feature Engineering
Step -3
Model building for selecting correct ML Algorithm
Step -4
Evaluate Model
Step -5
Model Deployment

DATA GATHERING

The First Phase and the most Important One

- This is specialized for the Data Engineers

HOW ?

- First of all, depends on your problem, ask about the task
  - ❑ Is it Natural Language Processing?
  - ❑ Is it Traditional Machine Learning?
  - ❑ Is it Computer Vision?
  - ❑ Is it a Time Series problem?
- Define your objective
  - ❑ Is it classification?
  - ❑ Is it regression?
  - ❑ Is it text generation/translation/summarization?
  - ❑ Is it an unsupervised learning task?depending

TYPES OF DATASETS

Labeled Dataset	Unlabeled Dataset
Labeled dataset means that there is a label that classify each record to a specific class or type, usually the last column but not always	Unlabeled dataset means that there is no label that define to which class/type that the record belongs to

Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Class_att
95.48023	46.55005	59	48.93018	96.6839	77.28307	0.778848	29.838	14.5939	12.2245	-21.6958	38.7849	Abnormal
74.09473	18.82373	76.03216	55.271	128.4057	73.38822	0.910886	13.1813	10.1368	8.49572	-0.33713	11.6844	Abnormal
87.67909	20.36561	93.82242	67.31347	120.9448	76.73063	0.574775	23.8665	13.0473	9.41012	5.212541	28.6308	Abnormal
48.25992	16.41746	36.32914	31.84246	94.88234	28.3438	0.388445	16.1775	15.0636	13.79474	-8.04464	21.6135	Abnormal
38.50527	16.9643	35.11281	21.54098	127.6329	7.986683	0.396364	34.8106	12.7802	15.24996	-28.8339	18.0442	Normal
54.92086	18.96843	51.60146	35.95243	125.8466	2.001642	0.106175	7.3907	11.3014	8.37076	-17.7235	9.8711	Normal
44.36249	8.945435	46.9021	35.41706	129.2207	4.994195	0.537574	33.0601	7.808	11.3766	-5.20236	33.2503	Normal
48.31893	17.45212	48	30.86681	128.9803	-0.91094	0.744322	36.6194	14.635	11.6271	-28.5988	10.3379	Normal
45.70179	10.65986	42.57785	35.04193	130.1783	-3.38891	0.990034	26.6333	18.3694	16.44832	-13.4388	34.2846	Normal

Spine Dataset

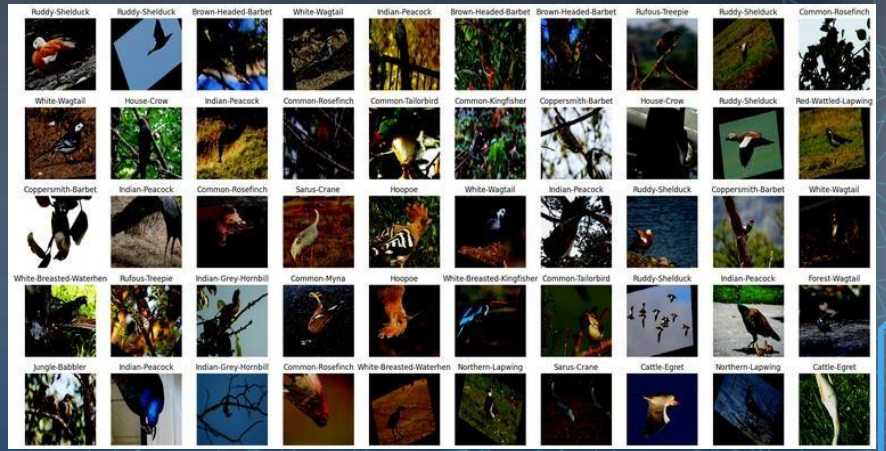
WHAT ABOUT LABELED DATA?

- **Note!** It is used for Supervised Learning tasks
- It could be a toy, random or real-world dataset, on our projects we are trying to use real-world data to be consistent

3560	الخدته الهانم تاخده مساحه السلالم
3561	الى تسيبه الهانم تاخده مساحة السلالم
3562	الى سابته الهانم تاخده مساحة السلالم
3563	الى سابته الهانم تاخده مساحة السلالم
3564	الى خدته الهانم تخده مساحه السلالم
3565	عدوك العاقل ولا حبيبك العبيط
3566	صباح الخير يا جاري، أنت في دارك وأنا في داري: أي لا يختلط الإنسان بالجار الشرير، إنما يكفي السلام «صباح الخير»، لكن كل إنسان في حاله (بلا اختلاط)
3567	خاف من عدوك مرة ومن صديقك ألف مرة
3568	خاف من عدوك مره ومن صديقك الف مره
3569	خاف من عدوك مرة ومن صديقك ألف مرة
3570	اجود من الدهب من يجود بالدهب
3571	ضربت الودع ملقتش صاحب جدع
3572	لو فيه خير ما عافه الطير
3573	ارميه في السطوح وان كان لك فيه خير ما يروح
3574	ازرع كل يوم تاكل كل يوم
3575	ما شافهومش وهما بيسرقوا شافوهم وهما بيتحاسبوا
3576	ما شافوهمش وهم بيسرقوا شافوهم وهم بيتحاسبوا

6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
39	0	2	138	220	0	1	152	0	0	1	0	2	1
58	0	0	130	197	0	1	131	0	0.6	1	0	2	1
47	1	2	130	253	0	1	179	0	0	2	0	2	1
35	1	1	122	192	0	1	174	0	0	2	0	2	1
58	1	1	125	220	0	1	144	0	0.4	1	4	3	1
56	1	1	130	221	0	0	163	0	0	2	0	3	1
56	1	1	120	240	0	1	169	0	0	0	0	2	1
55	0	1	132	342	0	1	166	0	1.2	2	0	2	1
41	1	1	120	157	0	1	182	0	0	2	0	2	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
67	1	0	160	286	0	0	108	1	1.5	1	3	2	0



The most common data sources to collect data for a ML model:

1. Open Source Datasets
2. Web Scraping
3. Synthetic Datasets
4. Manual Data Generation

STARTING TO COLLECT DATA

OPEN-SOURCE DATASETS

1. Google Cloud Public Datasets

https://cloud.google.com/public-datasets

Google is not just a search engine, it's much more! There are many public data sets that you can access on the Google cloud and analyze to obtain new insights from this data. There are more than 100 datasets and all of them are hosted by BigQuery and Cloud Storage.

2. Amazon Web Services Open Data Registry

https://registry.opendata.aws/

Amazon Web Services have a large number of data sets on their open data registry. You can download these data sets and use them on your own system or you can analyze the data on the Amazon Elastic Compute Cloud (Amazon EC2).

199 results

About COVID-19 Public Datasets
BigQuery Public Datasets Program

Getting started with COVID-19 Public Datasets

About Cymbal: Google Cloud's demo brand
Cymbal Group

Synthetic datasets across industries showcasing Google Cloud.

AFSC Open Data Portal
NOAA

Fisheries research data for the Alaska region

American Community Survey (ACS)
United States Census Bureau

Detailed US demographic data at various geographic resolutions

Area Deprivation Index (ADI)
BroadStreet

ADI: An index of socioeconomic status for communities

Austin Crime Data
City of Austin

City of Austin crime data for 2014 and 2015

Band Protocol Data
Cloud Public Datasets - Finance

Band Protocol data loaded into BigQuery

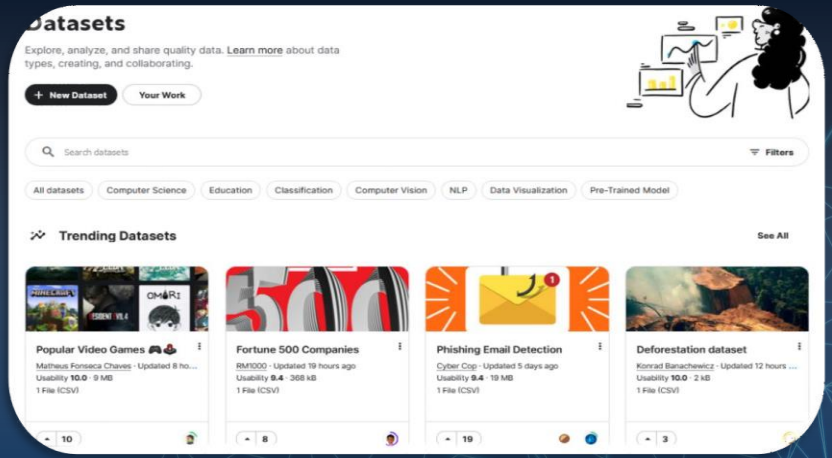
Births Data Summary
Centers for Disease Control

Natality Data from CDC Births

OPEN-SOURCE DATASETS (CONT.)

3. Kaggle

There are around 23,000 public datasets on Kaggle that you can download for free. In fact, many of these datasets have been downloaded millions of times already. You can use the search box to search for public datasets on whatever topic you want ranging from health to science to popular cartoons! You can also create new public datasets on Kaggle and those may earn you medals and also lead you towards advanced Kaggle titles like Expert, Master, and Grandmaster.



OPEN-SOURCE DATASETS (CONT.)

- https://datasetsearch.research.google.com/
- https://www.re3data.org/search
- https://search.datacite.org/
- https://ieee-dataport.org/datasets
- https://data.mendeley.com/
- https://sn-scigraph.figshare.com/browse

WHAT ABOUT TERM OPEN-SOURCE?THE

Term open-source means any one can reach to, deal with and republish it depend on the desire, so the question is, if there is open-source datasets, so it should be there are private datasets ?

The answer is YES ! Some hospitals, enterprises and companies offers their data for VERY HIGH prices, this data maybe for patients, clients or even customers, so there is a contract to SELL/TRADE this data.

Ethics ?

TO KNOW

- There are many file extensions for the datasets:
  - Comma-separated Values (CSV)
  - Tap-Separated Values (TSV)
  - Microsoft Excel XML spreadsheet (XLSX)
  - Database (DB)
  - Extensible Markup Language (XML)
  - JavaScript Object Notation (JSON)
  - Text files (TXT)

They are all supported by Python

WEB SCRAPING

Collecting data from websites using an automated process is known as **web scraping**. Some websites explicitly forbid users from scraping their data with automated tools like the ones that you'll create in Python. Websites do this for two possible reasons:

1. The site has a good reason to protect its data. For instance, Google Maps doesn't let you request too many results too quickly.

2. Making many repeated requests to a website's server may use up bandwidth, slowing down the website for other users and potentially overloading the server such that the website stops responding entirely.

In [1]:
```
from bs4 import BeautifulSoup
from urllib.request import urlopen
```

In [5]:
```
url = "https://en.wikipedia.org/wiki/Data_science"
page = urlopen(url)
html = page.read().decode("utf-8")
soup = BeautifulSoup(html, "html.parser")
```

In [6]:
```
soup.get_text()
```

```
analyzing astronomical survey data acquired by a space telescope, the Wide-field Infrared Survey Explorer.\nData science is
an interdisciplinary academic field [1] that uses statistics, scientific computing, scientific methods, processes, algorithm
s and systems to extract or extrapolate knowledge and insights from noisy, structured, and unstructured data.[2]\nData scien
ce also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology,
and medicine).[3] Data science is multifaceted and can be described as a science, a research paradigm, a research method, a
discipline, a workflow, and a profession.[4]\nData science is a "concept to unify statistics, data analysis, informatics, an
d their related methods" to "understand and analyze actual phenomena" with data.[5] It uses techniques and theories drawn fr
om many fields within the context of mathematics, statistics, computer science, information science, and domain knowledg
e.[6] However, data science is different from computer science and information science. Turing Award winner Jim Gray imagine
d data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted t
hat "everything about science is changing because of the impact of information technology" and the data deluge.[7][8]\nA  da
ta scientist is a professional who creates programming code and combines it with statistical knowledge to create insights fr
om data.[9]\n\n\nFoundations[edit]\nData science is an interdisciplinary field[10] focused on extracting knowledge from typi
cally large data sets and applying the knowledge and insights from that data to solve problems in a wide range of applicatio
n domains.[11] The field encompasses preparing data for analysis, formulating data science problems, analyzing data, develop
ing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. A
s such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, info
rmation visualization, data sonification, data integration, graphic design, complex systems, communication and business.[1
2][13] Statistician Nathan Yau, drawing on Ben Fry, also links data science to human–computer interaction: users should be a
```

WEB SCRAPING FREE TOOLS

- Scrapy
- ProWebScraper
- ScraperAPI

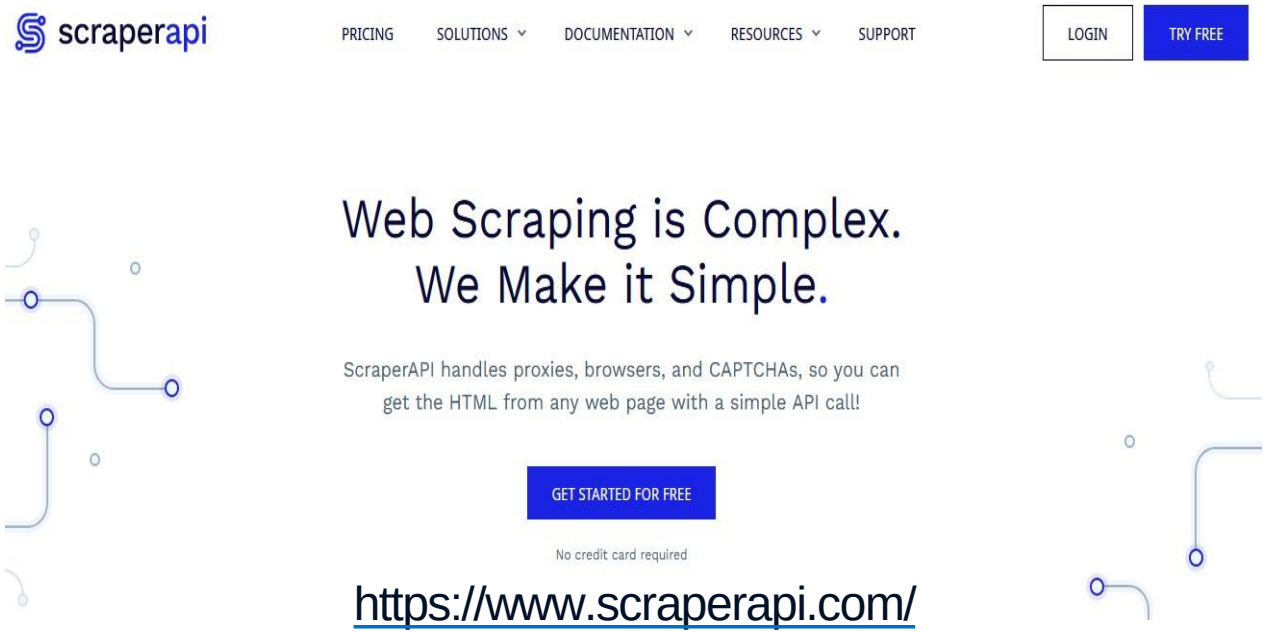
```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://www.zyte.com/blog/']

    def parse(self, response):
        for title in response.css('.oxy-post-title'):
            yield {'title': title.css('::text').get()}

        for next_page in response.css('a.next'):
            yield response.follow(next_page, self.parse)
EOF
$ scrapy runspider myspider.py
```

https://scrapy.org/

SCRAPY

https://prowebscraper.com/

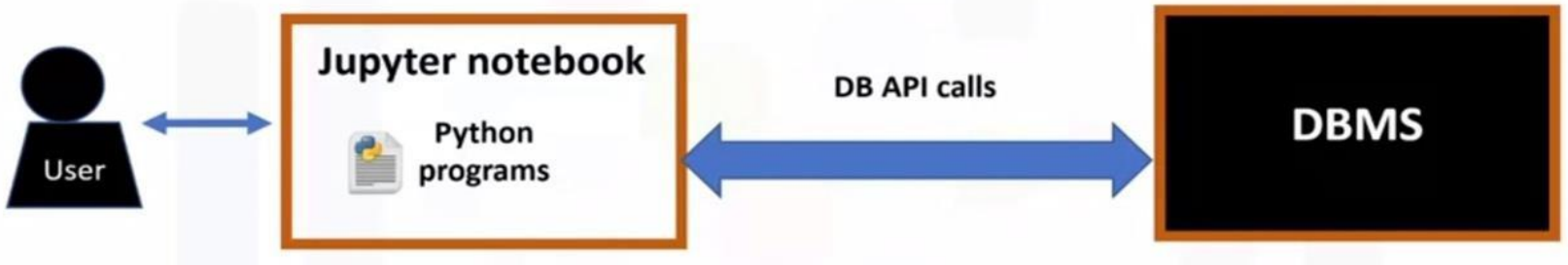
PROWEBSCRAPER


scraperapi
PRICING
SOLUTIONS
DOCUMENTATION
RESOURCES
SUPPORT
LOGIN
TRY FREE
Web Scraping is Complex.
We Make it Simple.
ScraperAPI handles proxies, browsers, and CAPTCHAs, so you can get the HTML from any web page with a simple API call!
GET STARTED FOR FREE
No credit card required
https://www.scraperapi.com/

SCRAPERAPI

WHAT ABOUT DATABASES?

- A standard that allows to write a single program
  - that works with multiple kinds of relational databases
  - instead of writing a separate program for each one



PYTHON DATABASE API

MySQLdb

- is an interface for connecting to a MySQL database from Python
- implements the **Python DB API** v2.0

```
import MySQLdb

# Open database connection
db = MySQLdb.connect("localhost","testuser","test123","TESTDB" )
```

MYSQLDB LIBRARY

1) Connection Objects

- used to connect to a database and manage transactions
- The methods used with connection objects:
  - **cursor() method** returns a new cursor object using the connection
  - **commit() method** commit any pending transaction to the database
  - **rollback() method** roll back the database to the start of any pending transaction
  - **close() method** close a database connection

2) Cursor Objects

- used to run queries
- used to scan through the results of a database
- works similar to a cursor in a text processing system
  - where you scroll down in your result set and get your data into the application

Sr.No	Method & Description
1	**callproc()** This method is used to call existing procedures MySQL database.
2	**close()** This method is used to close the current cursor object.
3	**Info()** This method gives information about the last query.
4	**executemany()** This method accepts a list series of parameters list. Prepares an MySQL query and executes it with all the parameters.
5	**execute()** This method accepts a MySQL query as a parameter and executes the given query.
6	**fetchall()** This method retrieves all the rows in the result set of a query and returns them as list of tuples. (If we execute this after retrieving few rows it returns the remaining ones)
7	**fetchone()** This method fetches the next row in the result of a query and returns it as a tuple.
8	**fetchmany()** This method is similar to the fetchone() but, it retrieves the next set of rows in the result set of a query, instead of a single row.

SQLITE

- **SQLITE** is used to connect on database file but not a server, it is such a quick way to manage an offline database file to extract data, especially with python
- You can use the ***SQLITE3*** API to connect to any local SQL file and manage it

```python
# ----------------------
# database functions Module
# ----------------------

import sqlite3

DB_FILENAME = 'Exam system db.db'

################## To get personal info field ######################


def personInfo_examiner(examiner_id):
    conn = sqlite3.connect(DB_FILENAME)
    res = conn.execute(
        f"""select examiner.examiner_id, examiner.examiner_name, examiner.examiner_
        , examiner.examiner_gen, examiner.examiner_email, examiner.examiner_phone,
        from examiner, department  where examiner_id = {examiner_id} and examiner.c
        """)
    res = list(res)
    conn.close()
    return res[0]


```

method, then calls the cursor's execute method with the parameters given.

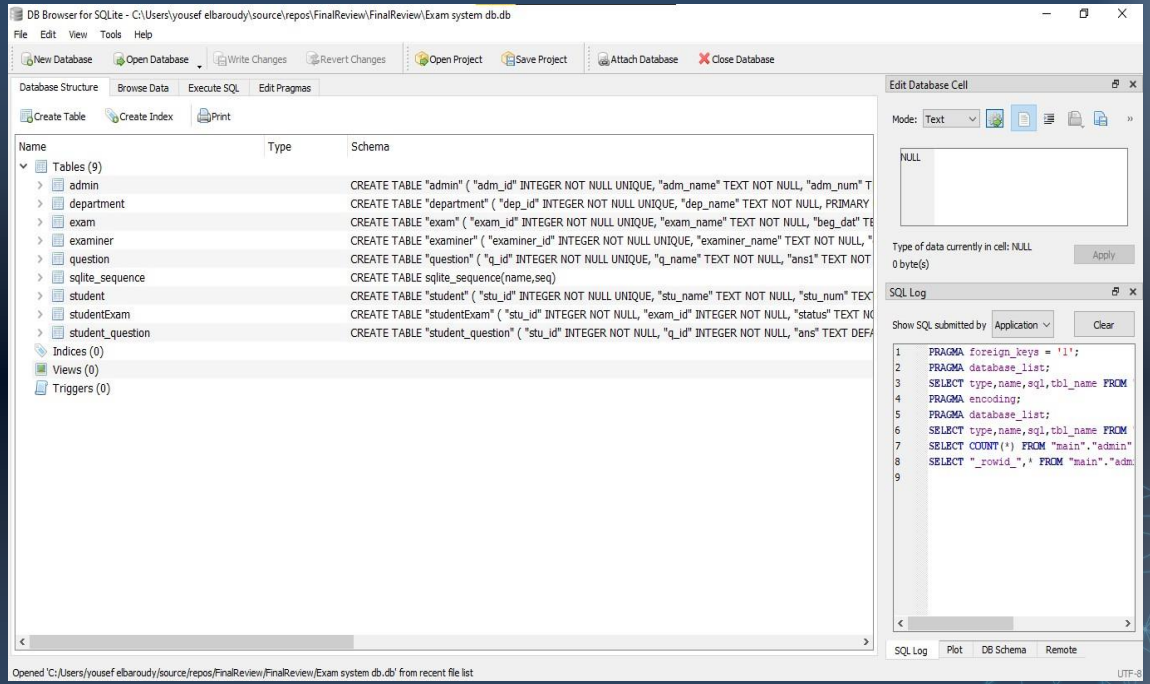
5	**cursor.executemany(sql, seq_of_parameters)** This routine executes an SQL command against all parameter sequences or mappings found in the sequence sql.
6	This routine is a shortcut that creates an intermediate cursor object by calling the cursor method, then calls the cursor.s executemany method with the parameters given.
7	This routine executes multiple SQL statements at once provided in the form of script. It issues a COMMIT statement first, then executes the SQL script it gets as a parameter. All the SQL statements should be separated by a semi colon (;).
8	**connection.executescript(sql_script)** This routine is a shortcut that creates an intermediate cursor object by calling the cursor method, then calls the cursor's executescript method with the

BASIC TUTORIAL OF SQLITE3 WITH PYTHON

https://www.tutorialspoint.com/sqlite/sqlite_python.htm

YOU MAY BROWSE YOUR DATA FIRST

- Using **DB Browse for SQLite** you can browse your data freely and to see what tables, columns and records it consists of.



BUILD SYNTHETIC DATASET

As the term already suggests, synthetic datasets are "***synthetic***" in the sense that they are generated through computer programs, instead of being composed through the documentation of real-world events.

So why should we even consider using synthetic data for ML models, when they don't contain real-world data?

BUILD SYNTHETIC DATASET (CONT.)

- Synthetic datasets are particularly useful when adequate real world data
1. cannot be obtained (or is very hard to obtain).

- Another key benefit of using synthetic data is that you are able to clearly define a number of features, such as the scope, format, and amount of noise within the dataset.

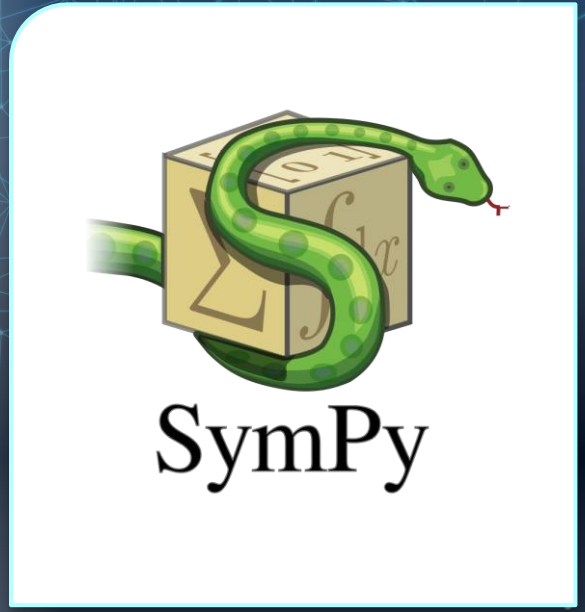
- Another benefit of using synthetic data that shouldn't be underestimated is that it eliminates the risk of any copyright infringement or privacy issues.

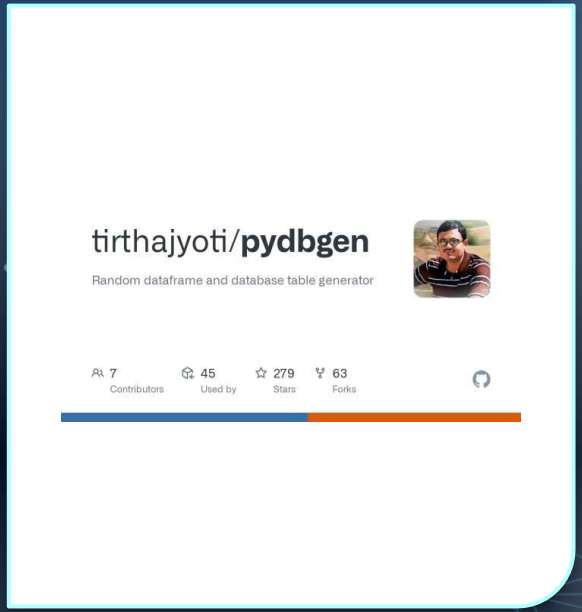
- This is especially interesting if you need some sort of personally identifiable

BECAREFUL!

the usage of synthetic datasets has some significant downsides.

- First, the creation of synthetic data is a big engineering burden, especially when you're alone or in a small team.
- Secondly, you run the risk of introducing **BIAS** in your data. As of its current status, artificial
  data alone is not enough to train advanced machine learning algorithms.





TOOLS FOR SYNTHETIC DATASET

MANUAL DATA GENERATION

This technique is very similar to the synthetic datasets with the exception, that it contains real data, and you need to generate the data manually instead of automatically.

you'll need to either set up sensors or conduct surveys to collect the data you need.

DIAMOND QUESTION

Why anyone should generate their own data, if there are so many datasets available on the internet for free as well as an abundance of web scraping tools?

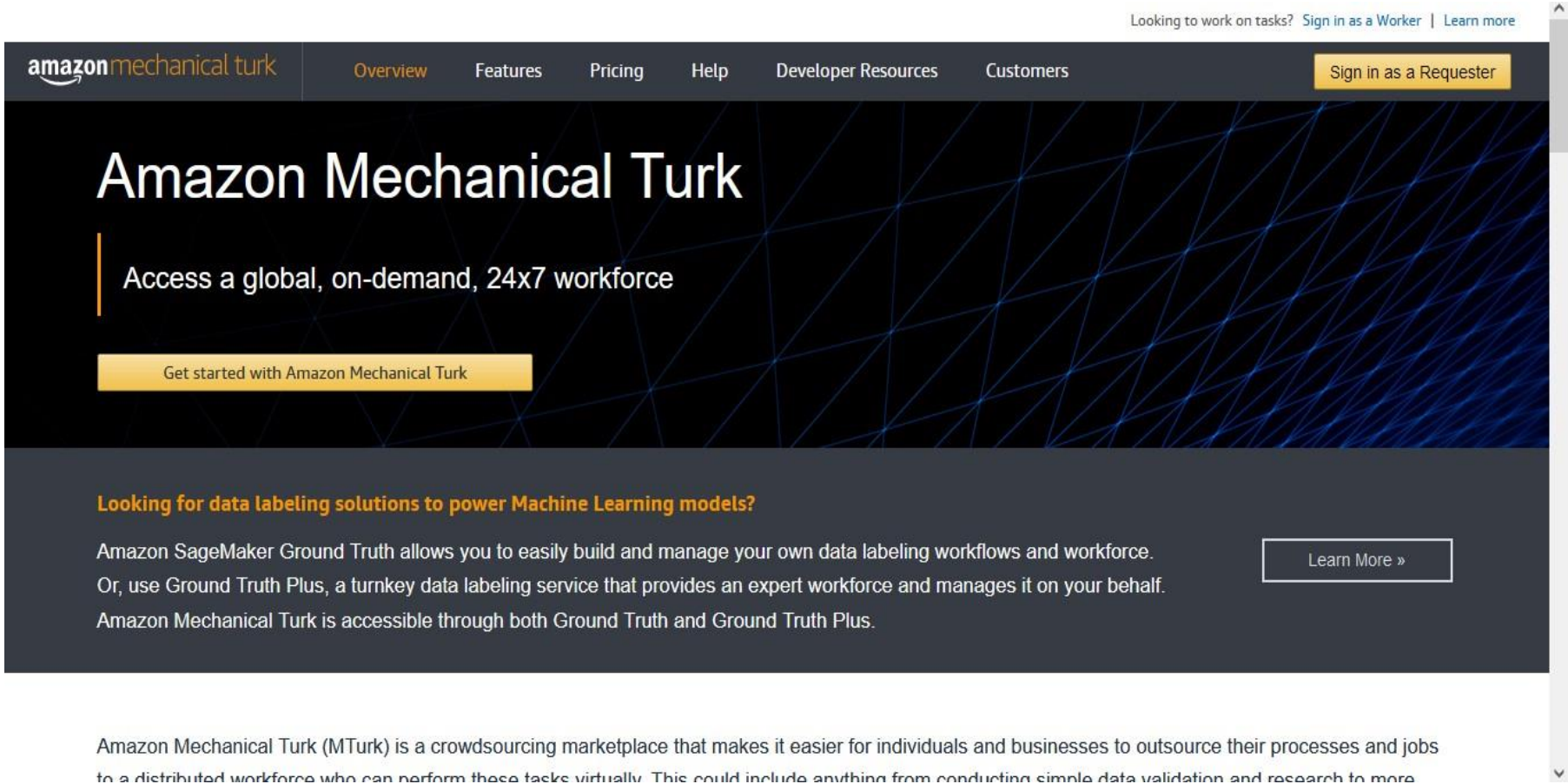
USE CASE

What about a smart factory that utilize Machine Learning to perform a product quality control whenever there is a new product or a new defect to detect ?

CROWDSOURCING

- **Crowdsourcing** means that human workers are given tasks to gather the necessary bits of data that collectively become the generated dataset.
- here is a wide range of crowdsourcing tasks from simple ones like labeling images up to complex ones like collaborative writing that involve multiple steps.

The most popular platform for crowdsourcing is **Amazon Mechanical Turk** where tasks are assigned to human workers, who are compensated for finishing the tasks.



https://www.mturk.com/

MANUAL DATA GENERATION (CONT.)

Now, as you might assume, this manual data generation has a lot of downsides.

First of all, extracting and formatting data on your own is very complex, which means it requires a lot of time, cost and knowledge. Especially for businesses, the usage of internally sourced data raises many concerns around privacy, especially when it involves **customer personally identifiable information (PII).**

MORE APPROACHES TO GAIN DATA ?

Data Augmentation

Transfer Learning

For NLP: Synonym Replacement, Replacing Entities, Bigram Flipping … etc

DATA AUGMENTATION

Let’s imagine for a second that we were not able to find a dataset that would meet all our requirements, BUT at the same time, we have a certain amount of basic data. Can we work with it?

DATA AUGMENTATION

Data augmentation is the increase of an existing training dataset's size and diversity without the requirement of manually collecting any new data source.

The process of data augmentation means that the input data will undergo a set of transformations and this way, thanks to the variations of data samples, our dataset will become richer.

EXAMPLE FOR IMAGES

- For example, if we deal with images, the number of augmentations that we can utilize is sufficient, because an image can be cut, mirrored, turned upside down, etc. Moreover, we can change the color settings with the help of brightness, saturation, contrast, clarity, and blur. These are the so-called ‘photometric transformations’.

TensorFlow

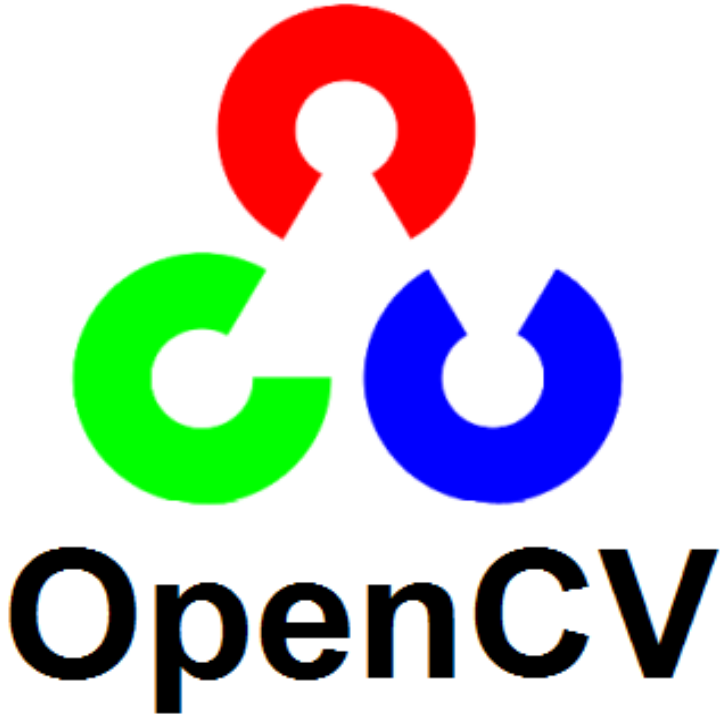
TENSORFLOW FOR DATA AUGMENTATION

- TensorFlow – allows to set ranges for rotation angles, brightness, zoom, rescale, etc. There's an option to turn on a built-in transformer feature in the generation flow of new samples.

scikit-image
image processing in python

SCIKIT IMAGE FOR DATA AUGMENTATION

- Scikit Image – a great library which helps not only to conduct basic operations with images, but also works with color spaces and allows you to apply filters.



OPENCV FOR DATA AUGMENTATION

- OpenCV – a pioneer of Computer Vision. In this Python-based library, there are tools for rotation, scaling, filters, cropping, etc.

TRANSFER LEARNING “LAZY LEARNING”

Another ‘magic wand’ for cases when it’s hard to “flesh out” the training dataset is *Transfer Learning*.

Transfer learning is an area in ML that utilizes the knowledge gained while solving one problem to solve a different, but related problem.

It’s just the way the human brain works: it’s easier for us to learn new things if we’ve had similar experiences in the past.

EXAMPLE

Let’s say, it’s easy to learn to ride a bike if you mastered a bike with training wheels before that. Learning a new programming language when you’ve been programming using other languages also shouldn’t be as hard.

TRANSFER LEARNING "LAZY LEARNING" (CONT.)

- Along with the rise of Computer Vision in recent years, the use of pre-trained models for object classification and identification has become a thing.
- Even now, in order to train a model for image classification, it will take days of processing.
- Taking into account the iterative and repetitive nature of Data Science, the search for the best model parameters can drag on for months.
- That is why the use of pre-trained models can save a lot of time and effort for data scientists in cases when you need a lot of input data

Here are some of the great examples of pre-trained models for Image Classification:

- Oxford' VGG-16, year 2014
- Microsoft's ResNet50, year 2015
- Google's InceptionV3, year 2015
- Google's EfficientNet, year 2019

SOME PRE-TRAINED MODELS (TRANSFER LEARNING)

Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Time (ms) per inference step (CPU)	Time (ms) per inference step (GPU)
Xception	88	79.0%	94.5%	22.9M	81	109.4	8.1
VGG16	528	71.3%	90.1%	138.4M	16	69.5	4.2
VGG19	549	71.3%	90.0%	143.7M	19	84.8	4.4
ResNet50	98	74.9%	92.1%	25.6M	107	58.2	4.6
ResNet50V2	98	76.0%	93.0%	25.6M	103	45.6	4.4
ResNet101	171	76.4%	92.8%	44.7M	209	89.6	5.2
ResNet101V2	171	77.2%	93.8%	44.7M	205	72.7	5.4
ResNet152	232	76.6%	93.1%	60.4M	311	127.4	6.5
ResNet152V2	232	78.0%	94.2%	60.4M	307	107.5	6.6
InceptionV3	92	77.9%	93.7%	23.9M	189	42.2	6.9
InceptionResNetV2	215	80.3%	95.3%	55.9M	449	130.2	10.0
MobileNet	16	70.4%	89.5%	4.3M	55	22.6	3.4
MobileNetV2	14	71.3%	90.1%	3.5M	105	25.9	3.8
DenseNet121	33	75.0%	92.3%	8.1M	242	77.1	5.4
DenseNet169	57	76.2%	93.2%	14.3M	338	96.4	6.3
DenseNet201	80	77.3%	93.6%	20.2M	402	127.2	6.7
NASNetLarge	343	82.5%	96.0%	88.9M	533	344.5	20.0
EfficientNetB0	29	77.1%	93.3%	5.3M	132	46.0	4.9
EfficientNetB1	31	79.1%	94.4%	7.9M	186	60.2	5.6
EfficientNetB2	36	80.1%	94.9%	9.2M	186	80.8	6.5
EfficientNetB3	48	81.6%	95.7%	12.3M	210	140.0	8.8
EfficientNetB4	75	82.9%	96.4%	19.5M	258	308.3	15.1
EfficientNetB5	118	83.6%	96.7%	30.6M	312	579.2	25.3
EfficientNetB6	166	84.0%	96.8%	43.3M	360	958.1	40.4
EfficientNetB7	256	84.3%	97.0%	66.7M	438	1578.9	61.6
EfficientNetV2B0	29	78.7%	94.3%	7.2M	-	-	-
EfficientNetV2B1	34	79.8%	95.0%	8.2M	-	-	-
EfficientNetV2B2	42	80.5%	95.1%	10.2M	-	-	-
EfficientNetV2B3	59	82.0%	95.8%	14.5M	-	-	-
EfficientNetV2S	88	83.9%	96.7%	21.6M	-	-	-
EfficientNetV2M	220	85.3%	97.4%	54.4M	-	-	-
EfficientNetV2L	479	85.7%	97.5%	119.0M	-	-	-
ConvNeXtTiny	109.42	81.3%	-	28.6M	-	-	-

THERE IS DATA INTEGRATION TOO

- You can bring many **Related** datasets and integrate them together
- It may require some tasks:
  - Data Preparation
  - Data Cleaning
  - Normalization
  - Unifying Columns and targets/classes

- The objective is to increase data as much as possible

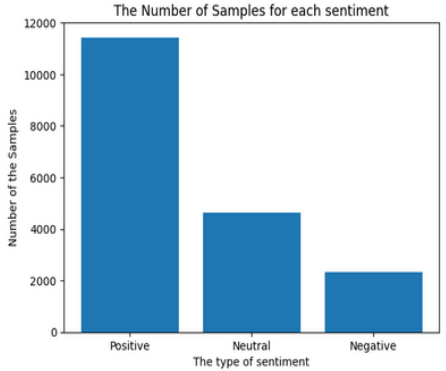
2.2 Getting youtube statistics dataset

In [103]:
```
youtube_comments.head()
```

Out[103]:

	Unnamed: 0	Video ID	Comment	Likes	Sentim
0	0	wAZZ-UWGVHI	Let's not forget that Apple Pay in 2014 requir...	95.0	1.0
1	1	wAZZ-UWGVHI	Here in NZ 50% of retailers don't even have co...	19.0	0.0
2	2	wAZZ-UWGVHI	I will forever acknowledge this channel with t...	161.0	2.0
3	3	wAZZ-UWGVHI	Whenever I go to a place that doesn't take App...	8.0	0.0
4	4	wAZZ-UWGVHI	Apple Pay is so convenient, secure, and easy t...	34.0	2.0

In [104]:
```
counting = youtube_comments.Sentiment.value_counts()
plt.bar(['Positive', 'Neutral', 'Negative'],counting)
plt.xlabel('The type of sentiment')
plt.ylabel('Number of the Samples')
plt.title("The Number of Samples for each sentiment")
plt.show()
```



DATA INTEGRATION

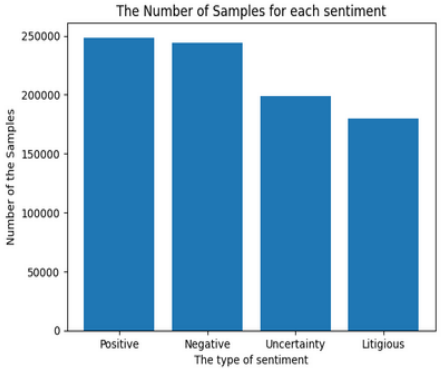
2.4 Getting sentiment dataset with 1 million tweets

In [111]:
```
sentiment_tweets.head()
```

Out[111]:

	Text	Language	Label
0	@Charlie_Corley @Kristine1G @amyklobuchar @Sty...	en	litigious
1	#BadBunny: Como dos gotas de agua: Joven se di...	es	negative
2	https://t.co/YJNiO0p1JV Flagstar Bank disclose...	en	litigious
3	Rwanda is set to host the headquarters of Unit...	en	positive
4	OOPS. I typed her name incorrectly (today's br...	en	litigious

In [113]:
```
counting = sentiment_tweets.Label.value_counts()
plt.bar(['Positive','Negative','Uncertainty','Litigious'],counting)
plt.xlabel('The type of sentiment')
plt.ylabel('Number of the Samples')
plt.title("The Number of Samples for each sentiment")
plt.show()
```



DATA INTEGRATION

2.6 Getting them all Together

In [119]:

```
# Concatenating the First & Second corpuses into one df
df = pd.concat([First_corpus,Second_corpus],ignore_index=True)
print(df.Label.value_counts())
print(f'The number of total samples after concatenating: {len(df)}')
df.head(10)
```

```
1.0    258395
0.0    245460
Name: Label, dtype: int64
The number of total samples after concatenating: 503855
```

Out[119]:

	Text	Label
0	Here in NZ 50% of retailers don't even have co...	0.0
1	I will forever acknowledge this channel with t...	1.0
2	Whenever I go to a place that doesn't take App...	0.0
3	Apple Pay is so convenient, secure, and easy t...	1.0
4	We only got Apple Pay in South Africa in 2020/...	1.0
5	In the United States, we have an abundance of ...	1.0
6	Wow, you really went to town on the PSU test r...	1.0
7	The lab is the most exciting thing in IT I've ...	1.0
8	Linus, I'm an engineer and love the LMG conten...	1.0
9	There used to be a time where Linus was the sm...	1.0

DATA INTEGRATION

WHAT ABOUT UNLABELED DATASETS ?

UNLABELED DATASETS

- It means that there is no target column. In other words, the data has no class to belong to.

- Usually used with Unsupervised Learning. **Why/How ?**

Data Labeling

DATA LABELING

Data Labeling – it's the process of data tagging or annotation for use in machine learning.

Labels are different and unique for each specific dataset, depending on the task at hand. The same dataset can have different meanings of labels and use them for various tasks.

For example, the classification of cats and dogs can turn into the classification of animals that have spots on the fur and the ones that don't.

- **Crowdsourcing**: a third-party gives a platform for individuals and businesses to outsource their processes and jobs;
- **Outsourcing**: hiring freelancers or contractors;
- **Specialized teams**: hiring teams that work in the field of Data Labeling and are trained and managed by third-party organization;
- **In-house teams**: giving tasks of Data Labeling to the internal team of workers or data scientists.

DATA LABELING TECHNIQUES

Each of these has its own pros and cons(such as the quality of the results, the cost of the job, or the speed in which labeling is

completed), and one method that suits one endeavor may not work for another. Moreover, you can combine them as you go.

NOTE !

DATA LABELING (CONT.)

If you cannot afford to hire a dedicated team for Data Labeling and you've decided to do everything in-house, you can't do without software tools to help with your task:

- **LabelBox, Annotorious, VGG Image Annotator, VoTT, ML Kit for Firebase** – images annotation tools
- **Anvil, VoTT, VGG Image Annotator, CVAT** – video annotation tools
- **Stanford CoreNLP, Brat, Dataturks, Tagtog** – text annotation tools
- **Prodigy, EchoML, Praat** – audio annotation tools

https://github.com/heartexlabs/awesome-data-labeling

THE CONCEPT OF SEMI-SUPERVISED LEARNING ?

To Be Explained Later

SOURCES

https://waverleysoftware.com/blog/data-collection-for-machine-learning-guide/

https://medium.com/codex/how-to-collect- data-for-a-machine-learning-model- 2b152752a15b

https://www.geeksforgeeks.org/top-8- free-dataset-sources-to-use-for-data- science-projects/

Tasks

(Web Scraping)

1- Name any of web scraping tool and scrap any website of your choose
2- save your result in txt file and upload it

(Data Labeling)

1- Use the following dataset on link **https://www.kaggle.com/datasets/paultimothymooney/sample-images-for-kaggle-demos?select=1928768_1037975187515_2166_n.jpg**
2- label each image for which animal it is (Put each image in a folder with the name of the animal and put all files in one folder of your name in a ZIP File)
3- Upload your ZIP file
** You can use an automatic labeling tool, or you can do it manually*

QUESTIONS


Thank You