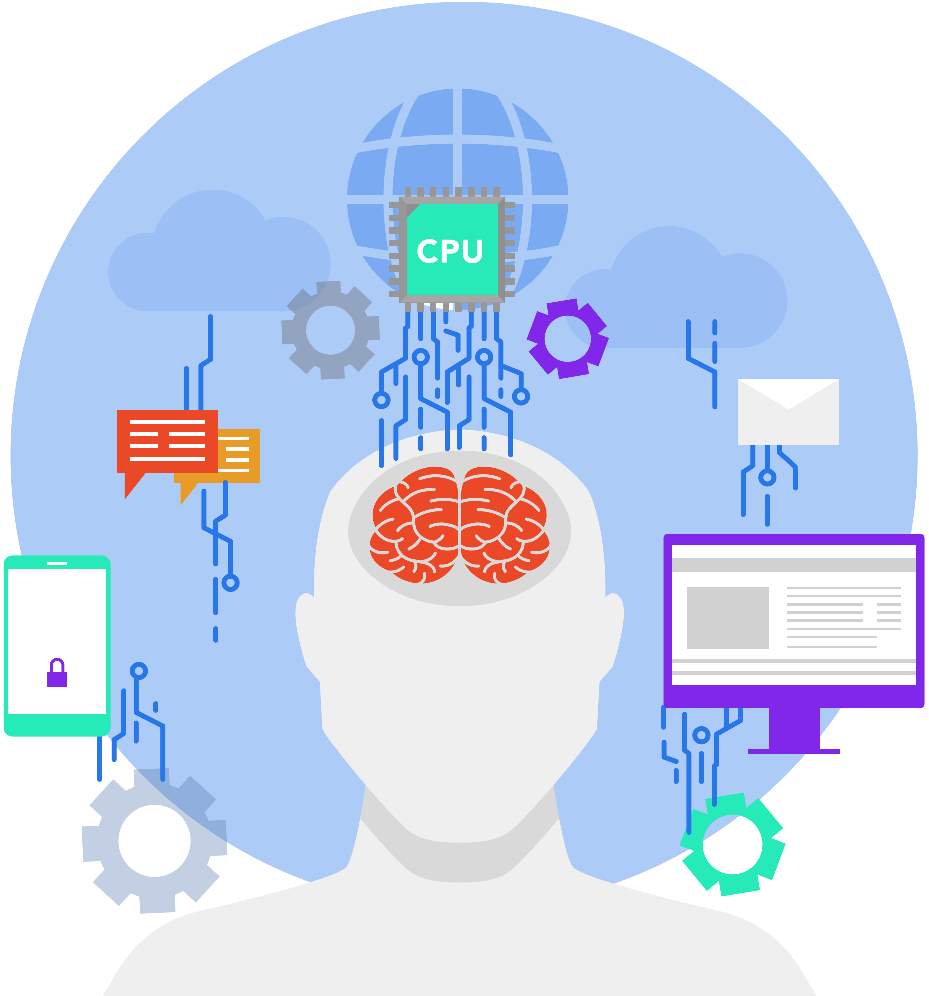


Data Acquisition & Data Understanding

OUTLINE

- ❑ **Data**
  - ❑ Data Source
  - ❑ Data Structure
  - ❑ Data Type
  - ❑ Data Item Type
  - ❑ Data Model
- ❑ **Data Acquisition**
  - ❑ Manual Data Acquisition from Repository
  - ❑ Data Acquisition via Public API
  - ❑ Data Acquisition with Web Scraping
  - ❑ Data Acquisition from a Relational Database
- ❑ **Data Understanding**

Data

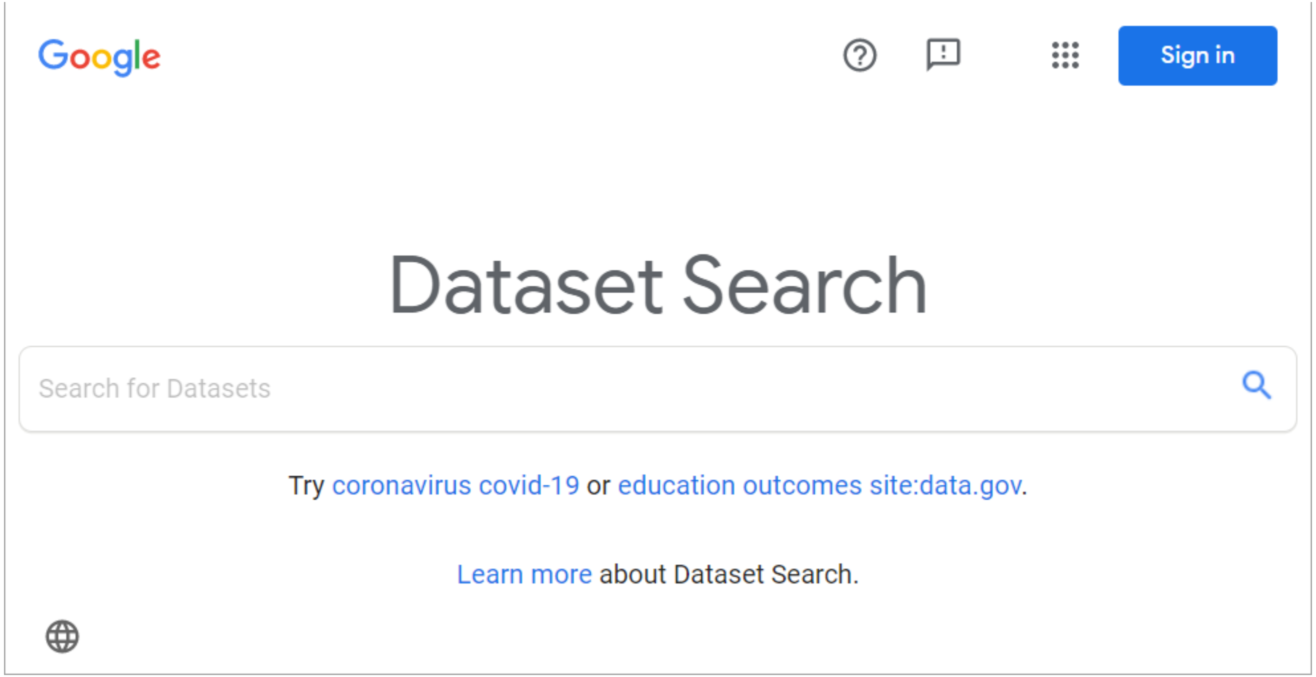
- Data is **a collection of information or facts** in the form of numbers, symbols, words, images, and others obtained through observing variables or searching certain sources.
- Data is the **raw material** for AI solutions.
- Data from different sources can not be used directly because:
  - the aims and objectives of the data are different.
  - the original state is separate or even tightly integrated and complex.
  - different levels of richness.
  - different levels of reliability.

Data Sources

Internal sources	Spreadsheets (Excel, CSV, JSON, etc.)
	Databases: can be queried via SQL, etc.
	Text documents
	Multimedia documents (audio, image, video)
External sources	Public domain web pages
	Open data repositories

Google Dataset Search

Data available on the Web can be searched using the Google Dataset Search service: https://datasetsearch.research.google.com



Dataset Search

Try coronavirus covid-19 or education outcomes site:data.gov.

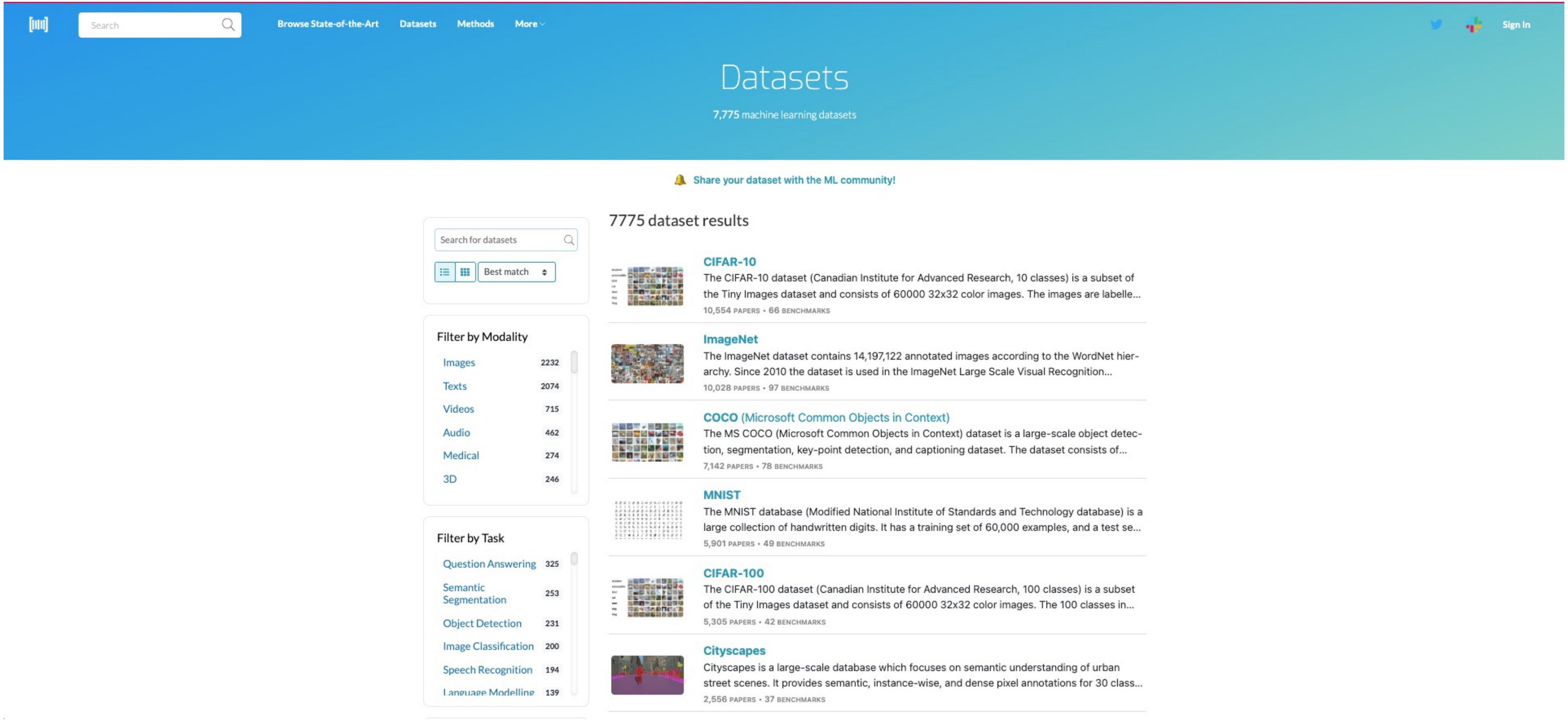
Learn more about Dataset Search.

Open Data Repositories

- Portal Satu Data Indonesia (https://data.go.id)
- Portal Data Jakarta (https://data.jakarta.go.id)
- Portal Data Bandung (http://data.bandung.go.id)
- Badan Pusat Statistik (https://www.bps.go.id)
- Badan Informasi Geospasial (https://tanahair.indonesia.go.id/)
- UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php)
- Kaggle (https://www.kaggle.com/datasets)
- World Bank Open Data (https://data.worldbank.org)
- UNICEF Data (https://data.unicef.org)
- WHO Open Data (https://www.who.int/data)
- IBM Data Asset eXchange (https://developer.ibm.com/exchanges/data/)
- DBPedia (https://www.dbpedia.org/resources/)
- Wikidata (https://www.wikidata.org/)

Papers with Code

Data available on the AI state-of-the-art research website:
https://paperswithcode.com/datasets



Datasets

7,775 machine learning datasets

Share your dataset with the ML community!

Filter by Modality

Images	2232
Texts	2074
Videos	715
Audio	462
Medical	274
3D	246

Filter by Task

Question Answering	325
Semantic Segmentation	253
Object Detection	231
Image Classification	200
Speech Recognition	194
Language Modelling	139

7775 dataset results

CIFAR-10
The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The images are labelle...
10,554 PAPERS • 66 BENCHMARKS

ImageNet
The ImageNet dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition...
10,028 PAPERS • 97 BENCHMARKS

COCO (Microsoft Common Objects in Context)
The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of...
7,142 PAPERS • 78 BENCHMARKS

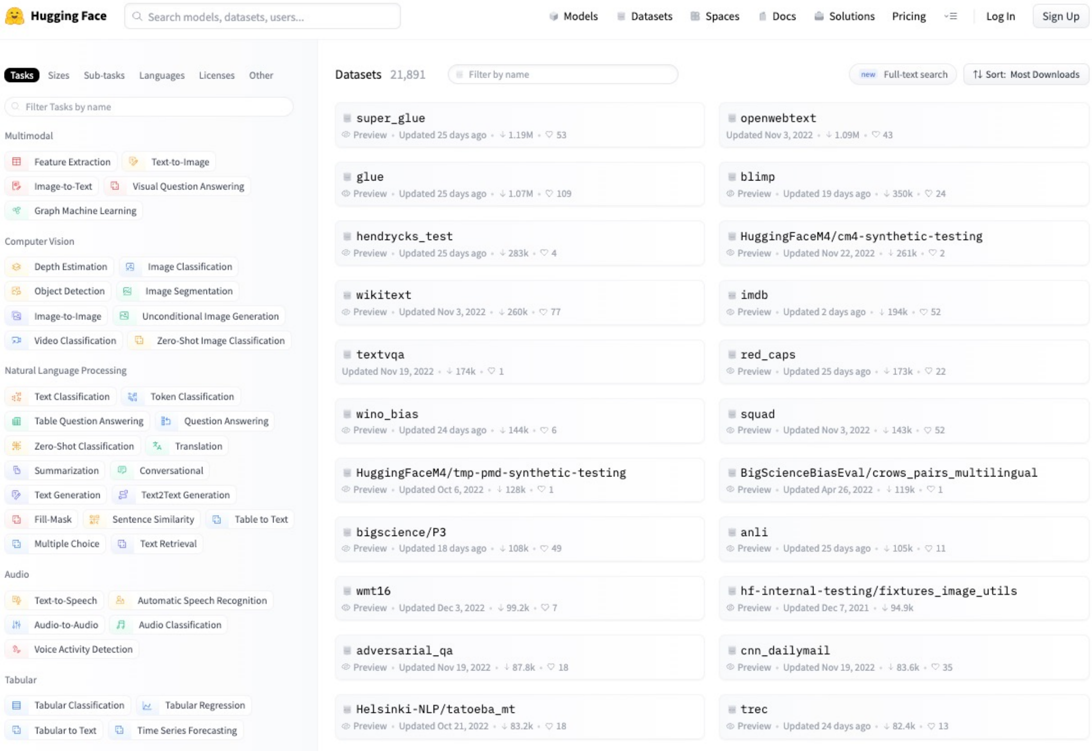
MNIST
The MNIST database (Modified National Institute of Standards and Technology database) is a large collection of handwritten digits. It has a training set of 60,000 examples, and a test se...
5,901 PAPERS • 49 BENCHMARKS

CIFAR-100
The CIFAR-100 dataset (Canadian Institute for Advanced Research, 100 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The 100 classes in...
5,305 PAPERS • 42 BENCHMARKS

Cityscapes
Cityscapes is a large-scale database which focuses on semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 class...
2,556 PAPERS • 37 BENCHMARKS

Hugging Face

Data available on the AI state-of-the-art research website:
https://huggingface.co/datasets



🤗 Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Docs | Solutions | Pricing | Log In | Sign Up

Tasks | Sizes | Sub-tasks | Languages | Licenses | Other

Filter Tasks by name

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Visual Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity
- Table to Text
- Multiple Choice
- Text Retrieval

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification
- Tabular Regression
- Tabular to Text
- Time Series Forecasting

Datasets 21,891 | Filter by name | new Full-text search | Sort: Most Downloads

super_glue
Preview • Updated 25 days ago • 1.19M • 53

openwebtext
Updated Nov 3, 2022 • 1.09M • 43

glue
Preview • Updated 25 days ago • 1.07M • 109

blimp
Preview • Updated 19 days ago • 350k • 24

hendrycks_test
Preview • Updated 25 days ago • 283k • 4

HuggingFaceM4/cm4-synthetic-testing
Preview • Updated Nov 22, 2022 • 261k • 2

wikitext
Preview • Updated Nov 3, 2022 • 260k • 77

imdb
Preview • Updated 2 days ago • 194k • 52

textvqa
Updated Nov 19, 2022 • 174k • 1

red_caps
Preview • Updated 25 days ago • 173k • 22

wino_bias
Preview • Updated 24 days ago • 144k • 6

squad
Preview • Updated Nov 3, 2022 • 143k • 52

HuggingFaceM4/tmp-pmd-synthetic-testing
Preview • Updated Oct 6, 2022 • 128k • 1

BigScienceBiasEval/crows_pairs_multilingual
Preview • Updated Apr 26, 2022 • 119k • 1

bigscience/P3
Preview • Updated 18 days ago • 108k • 49

anli
Preview • Updated 25 days ago • 105k • 11

wmt16
Preview • Updated Dec 3, 2022 • 99.2k • 7

hf-internal-testing/fixtures_image_utils
Preview • Updated Dec 7, 2021 • 94.9k

adversarial_qa
Preview • Updated Nov 19, 2022 • 87.8k • 18

cnn_dailymail
Preview • Updated Nov 19, 2022 • 83.6k • 35

Helsinki-NLP/tatoeba_mt
Preview • Updated Oct 21, 2022 • 83.2k • 18

trec
Preview • Updated 24 days ago • 82.4k • 13

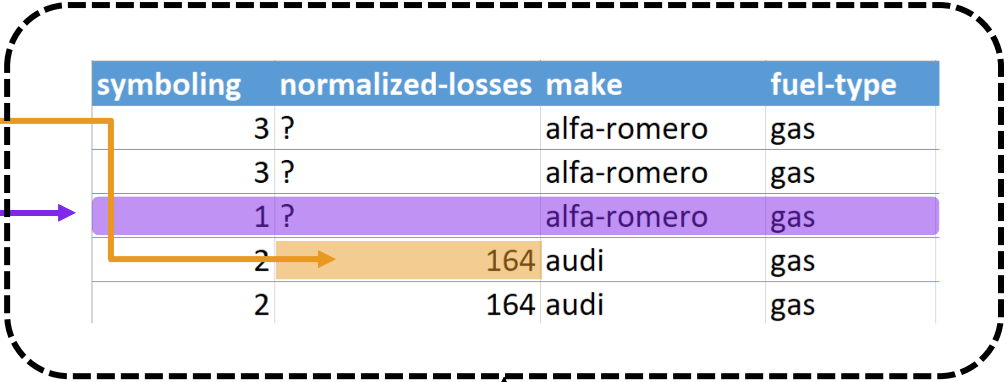
Data Structure

Data item (*datum*): the smallest unit of data; one value for one specific variable

Data: a collection of data items that have a certain unity of meaning (describing one object).

Dataset: a collection of data

Metadata: data that describes other data



symboling	normalized-losses	make	fuel-type
3	?	alfa-romero	gas
3	?	alfa-romero	gas
1	?	alfa-romero	gas
2	164	audi	gas
2	164	audi	gas

"make":
- type: string,
- description: name of the vehicle manufacturer

Data Types based on Structure

	Structured Data	Unstructured Data
Characteristic	• The data model used is known or determined before the data is created or constructed. • The data item format is (usually) text. • Each data item is clearly distinguished. • Direct extraction/querying/processing is pretty straightforward.	• The data model used is not predefined before. • The data item format is (usually) text, image, sound, video, and other formats. • Each data item is not clearly distinguished because of irregularity and ambiguity. • Direct extraction/querying/processing is quite difficult.
Example	Tabular data, object-oriented data, time-series data	Text data in free text documents, audio data, video data.

Semi-structured data: Structured data that does not follow the tabular structure model as in relational databases, but still contains tags or other markers that can separate semantic elements in the data and set hierarchies between the data items.

Data Item Types

	Nominal/ Categorical	Ordinal	Interval	Ratio
Properties of the original set	Discrete, not sorted	Discrete, in order	Continuous/numeric, ordered, distinction indicate differences	Continuous/numeric, ordered, values indicate the ratio to the quantity of units of the same type
Example	Color (red, green, blue)	Student letter grades (A, B, C, D, E)	Temperature in Celsius, date in specific calendar, location in the Cartesian coordinate	Length of the road, Temperature in Kelvin
Data measure states	Membership	Membership, comparison/level	Membership, comparison/level, difference	Membership, comparison/level, difference, magnitude
Mathematical operations	=, ≠	=, ≠, <, >	=, ≠, <, >, +, -	=, ≠, <, >, +, -, ×, ÷

Data Item Types

	Nominal/ Categorical	Ordinal	Interval	Ratio
Typical value representation (central tendency)	Mode	Mode, median	Mode, median, arithmetic mean	Mode, median, arithmetic mean, geometric mean, harmonic mean
Distribution representation	Grouping	Grouping, range, interquartile range	Grouping, range, interquartile range, variance, standard deviation	Grouping, range, interquartile range, variance, standard deviation, coefficient of variation
Has absolute zero which represents the lowest absolute value	No	No	No	Yes

Data Model

The data model represents an **abstraction of form or structure that underlies how data items are organized into a single meaning**.

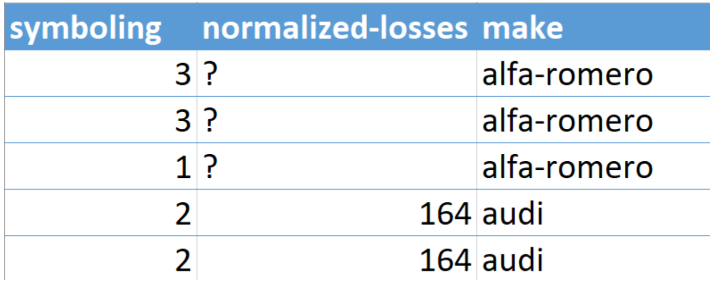
- The data model also determines how these data items are related to one another, and how these data items are related to entities in the real world.
- For example, a data model for representing an individual student consists of several data items, such as the student's primary number, name, and study program.

The term data model itself is sometimes used to **express abstractions of objects and relations that are relevant to a particular application domain**.

- For example, a data model for an e-commerce company usually contains abstractions that represent the concept of customers, products, and goods purchase transactions.

Model Data: Tabular

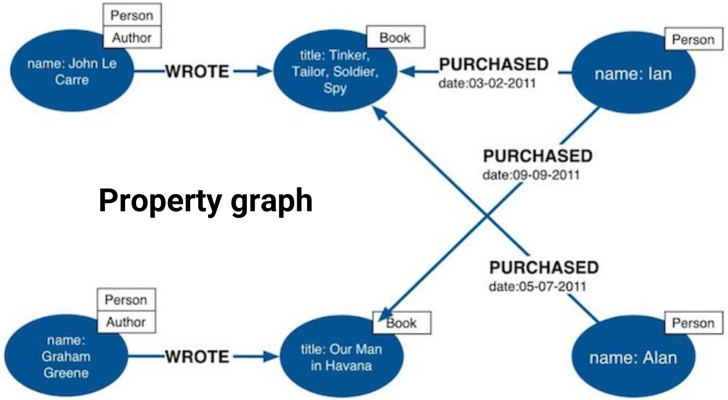
- Consists of N records.
- Each record contains D attributes.
- Records = rows, data points, instances, examples, transactions, tuples, entities, objects, feature vectors.
- Attributes = columns, fields, dimensions, features.
- The same attribute for each record is usually assumed to have the same data item type.
- Structures can be strict (i.e. relational database) or loose (i.e. Excel spreadsheets).
- Depending on the tightness of the structure, there can be a formal query language to access the data items in it (i.e. SQL).

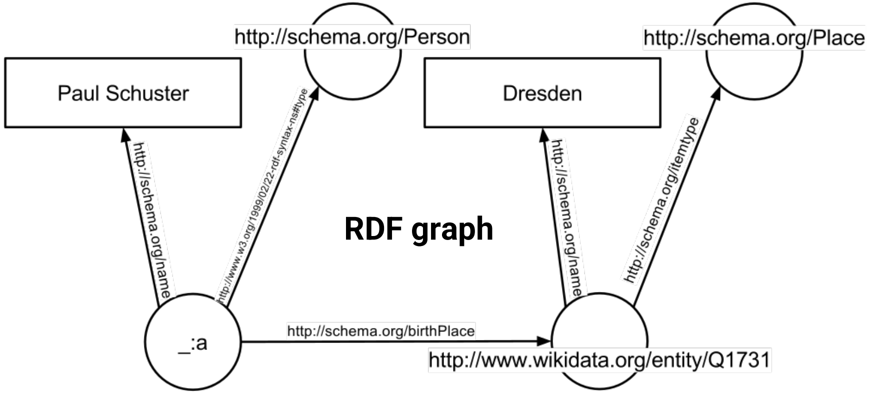


symboling	normalized-losses	make
3	?	alfa-romero
3	?	alfa-romero
1	?	alfa-romero
2	164	audi
2	164	audi

Model Data: Graph

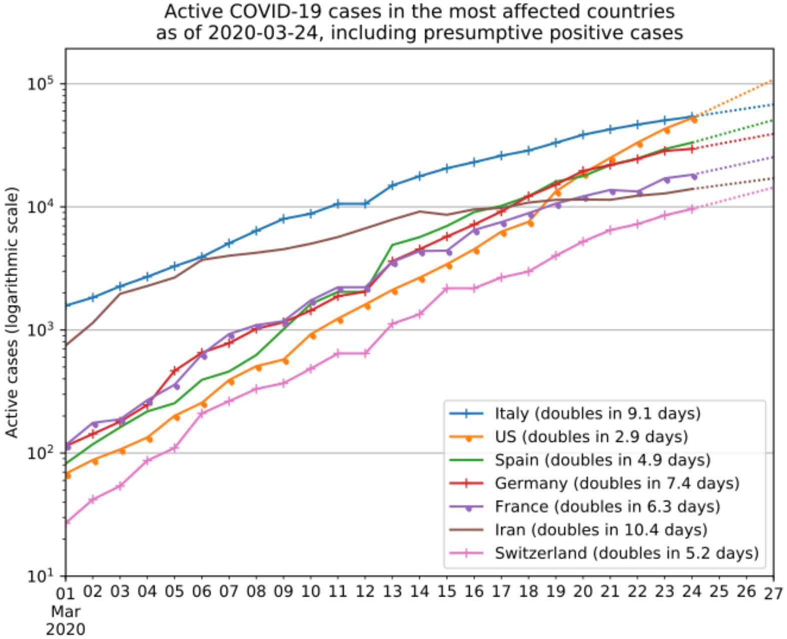
- Consists of vertices (nodes) and sides/ connections between vertices (edges).
- One node (usually) represents one record.
- Can express relations between records explicitly.
- Included in the graph data model are hierarchical/tree data models and object-oriented data models.
- Modern graph data models:
  - Property graph
  - Resource description framework (RDF)
- Graph database implementations: Neo4j, Apache Tinkerpop, GraphDB, Virtuoso, AllegroGraph, Oracle Spatial and Graph, etc.
- Graph databases usually have their own query language, i.e. Cypher, Gremlin, GraphQL, SPARQL.
- Some of these query languages have been designated as standard by various standards bodies.


Property graph


RDF graph
http://schema.org/Person
http://schema.org/Place
Paul Schuster
Dresden
http://schema.org/name
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://schema.org/name
http://schema.org/itemtype
_:a
http://schema.org/birthPlace
http://www.wikidata.org/entity/Q1731

Model Data: Sequence

- Consists of sequentially connected records.
- Example: data from a temperature sensor over a period of time.
- The implied structure of the order in which the records appear.
- Audio and video recordings can be viewed as sequence data, but each record itself is unstructured.
- Attributes in sequence data can be classified into contextual attributes and behavioral attributes.
- The contextual attributes define the implied dependency base (i.e. time stamp on a temperature sensor)
- Behavioural attributes define data items whose values are obtained in a certain context (i.e. temperature).
- If the contextual attribute is time/time stamp, then the sequence data is called a time series.

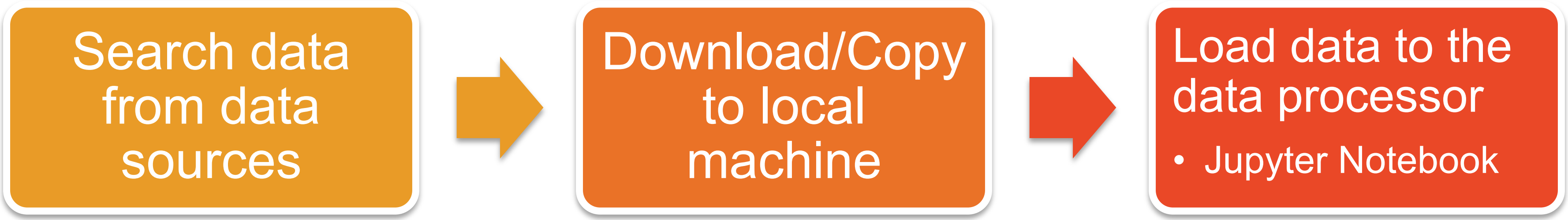


Data Acquisition

Acquisition data from data sources from both internal and external organizations, there are at least four modes of access:

1. Access manually by downloading data files directly or obtaining them via certain communication channels such as e-mail or sending via chat applications
2. Access programmatically via the Application Programming Interface (API).
3. Access programmatically by extracting directly from Web pages (Web scraping)
4. Access programmatically to relational databases within the organization.

Manual Data Acquisition



Search data from data sources

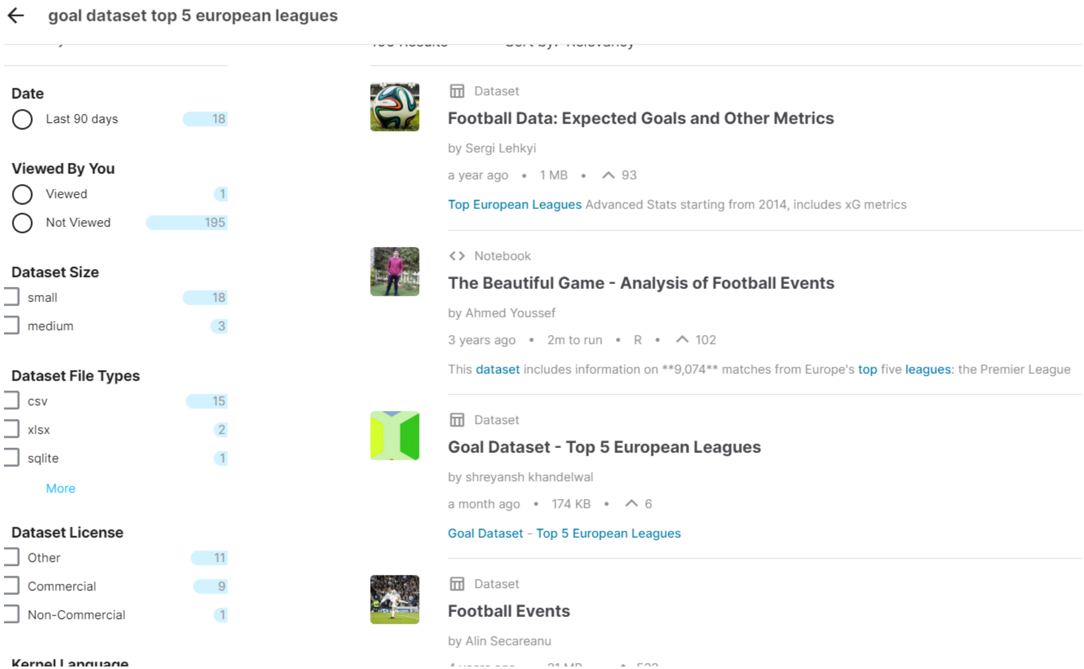
Download/Copy to local machine

Load data to the data processor

- Jupyter Notebook

Data Acquisition (Manually) from Kaggle

- For example, we will access data from "**Goal Dataset – Top 5 European Leagues**" from Kaggle.
- Visit the Kagle web page https://www.kaggle.com then log in (create an account if needed)
- Searching for "goal dataset top 5 European leagues" or any desired keyword.
- Click "Goal Dataset – Top 5 European Leagues"



← goal dataset top 5 european leagues

Date

○ Last 90 days 18

Viewed By You

○ Viewed 1

○ Not Viewed 195

Dataset Size

□ small 18

□ medium 3

Dataset File Types

□ csv 15

□ xlsx 2

□ sqlite 1

More

Dataset License

□ Other 11

□ Commercial 9

□ Non-Commercial 1

Kernel Language

Dataset

Football Data: Expected Goals and Other Metrics

by Sergi Lehkyi

a year ago • 1 MB • ^ 93

Top European Leagues Advanced Stats starting from 2014, includes xG metrics

<> Notebook

The Beautiful Game - Analysis of Football Events

by Ahmed Youssef

3 years ago • 2m to run • R • ^ 102

This dataset includes information on **9,074** matches from Europe's top five leagues: the Premier League

Dataset

Goal Dataset - Top 5 European Leagues

by shreyansh khandelwal

a month ago • 174 KB • ^ 6

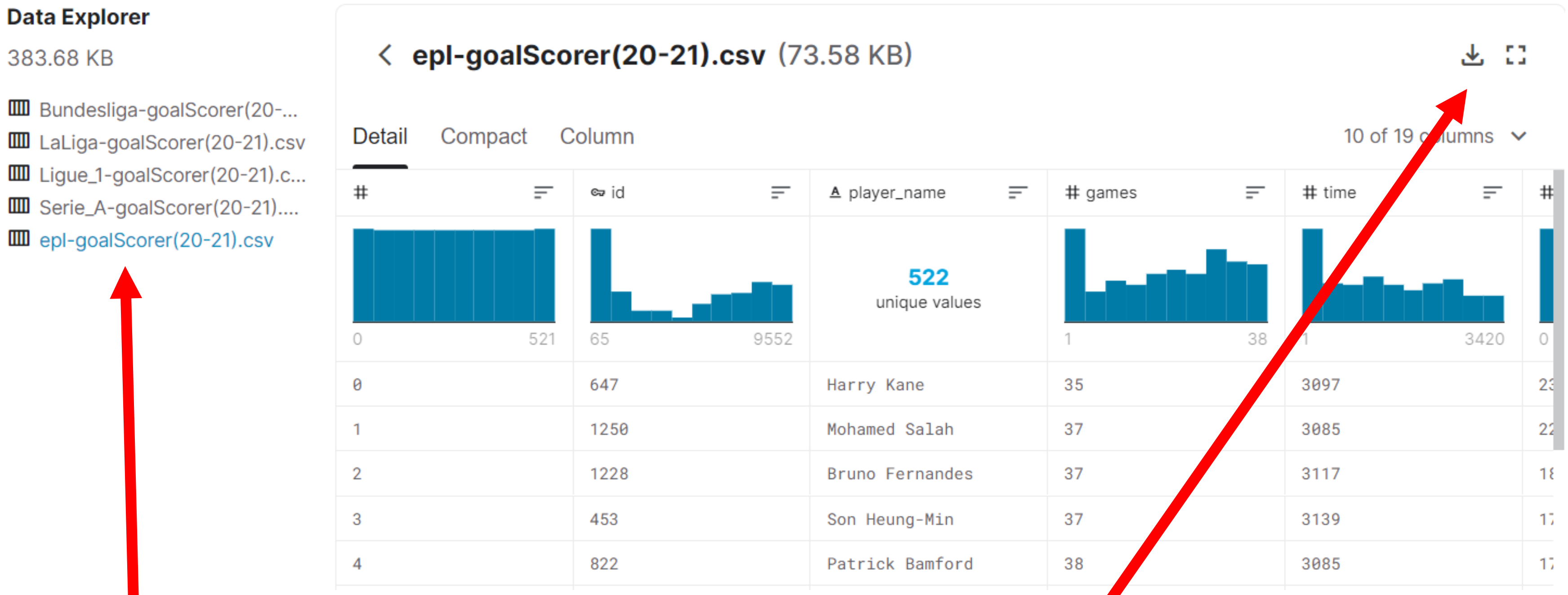
Goal Dataset - Top 5 European Leagues

Dataset

Football Events

by Alin Secareanu

Data Acquisition (Manually) from Kaggle



Data Explorer

383.68 KB

- Bundesliga-goalScorer(20-...
- LaLiga-goalScorer(20-21).csv
- Ligue_1-goalScorer(20-21).c...
- Serie_A-goalScorer(20-21)....
- epl-goalScorer(20-21).csv

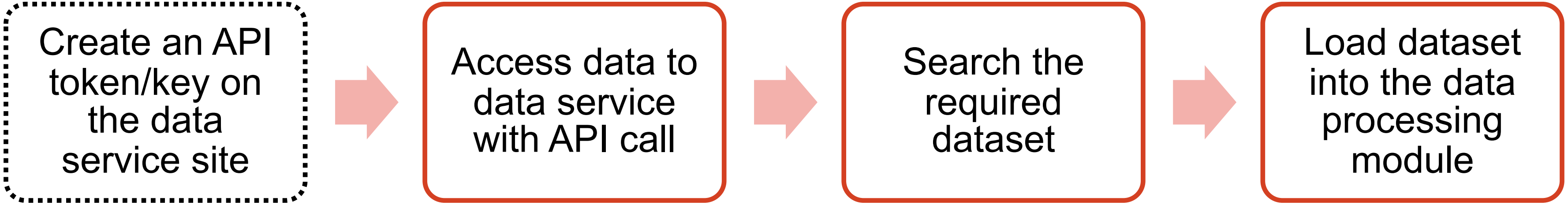
epl-goalScorer(20-21).csv (73.58 KB)

#	id	player_name	# games	# time	#
0 – 521	65 – 9552	522 unique values	1 – 38	1 – 3420	0
0	647	Harry Kane	35	3097	23
1	1250	Mohamed Salah	37	3085	22
2	1228	Bruno Fernandes	37	3117	18
3	453	Son Heung-Min	37	3139	17
4	822	Patrick Bamford	38	3085	17

- In the data explorer page, select "**epl-goalScorer (20-21).csv**"
- Download the data by clicking the **download button** on the right and save it in your working folder.

Data Acquisition via Public API

- Data can be obtained by utilizing the **public Application Programming Interface (API)** provided by several data services, such as Kaggle, One Data Indonesia Portal, or Bandung Data Portal.
- API token/key (maybe) is needed to access data via API.
- The process for generating API tokens/keys (if needed) is detailed in the documentation for each service.



Create an API token/key on the data service site

Access data to data service with API call

Search the required dataset

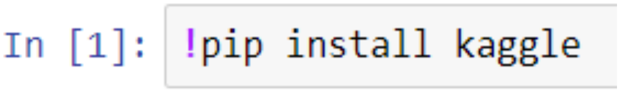
Load dataset into the data processing module

Data Acquisition with API from Kaggle (1)

Kaggle (https://www.kaggle.com) provides a Python-based API to access the data in it. This API can be run on Jupyter Notebook.

- Start Jupyter Notebook in your folder and open or create a new script (Python 3).

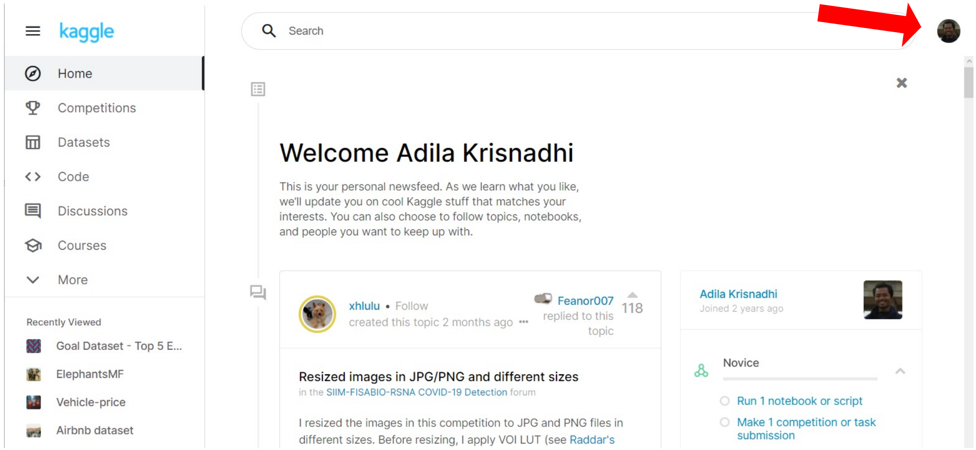
- Install `kaggle` library (ex: with pip)



```
In [1]: !pip install kaggle
```

Data Acquisition with API from Kaggle (2)

- Log in to Kaggle, click on your **profile photo** (top right), and then click '**Your Profile**' to open your profile page.
- On your profile page, click the '**Account**' tab. Swipe down a bit, and you will find the '**Create New API Token**' button.



Welcome Adila Krisnadhi

This is your personal newsfeed. As we learn what you like, we'll update you on cool Kaggle stuff that matches your interests. You can also choose to follow topics, notebooks, and people you want to keep up with.

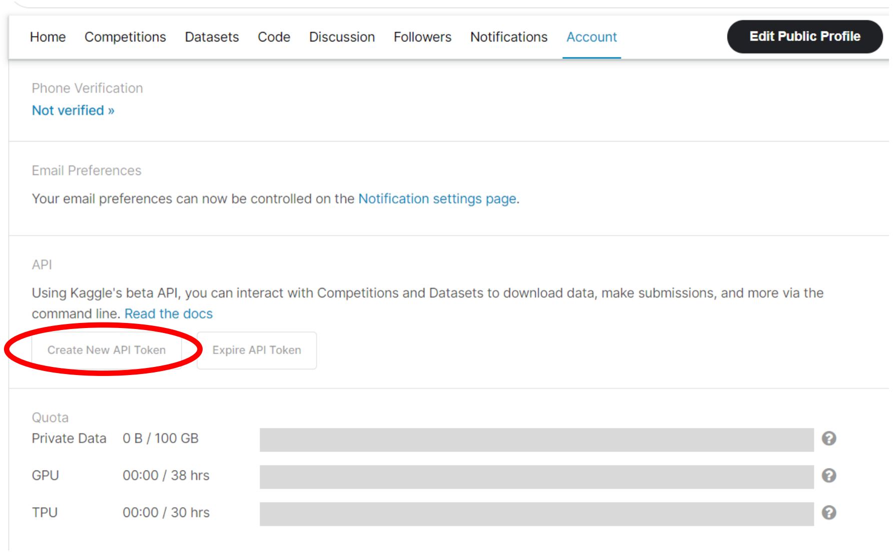
created this topic 2 months ago

replied to this topic

Resized images in JPG/PNG and different sizes

in the SIIM-FISABIO-RSNA COVID-19 Detection forum

I resized the images in this competition to JPG and PNG files in different sizes. Before resizing, I apply VOI LUT (see Raddar's



Phone Verification

Not verified »

Email Preferences

Your email preferences can now be controlled on the Notification settings page.

API

Using Kaggle's beta API, you can interact with Competitions and Datasets to download data, make submissions, and more via the command line. Read the docs

Quota

Private Data	0 B / 100 GB	
GPU	00:00 / 38 hrs	
TPU	00:00 / 30 hrs	

Data Acquisition with API from Kaggle (3)

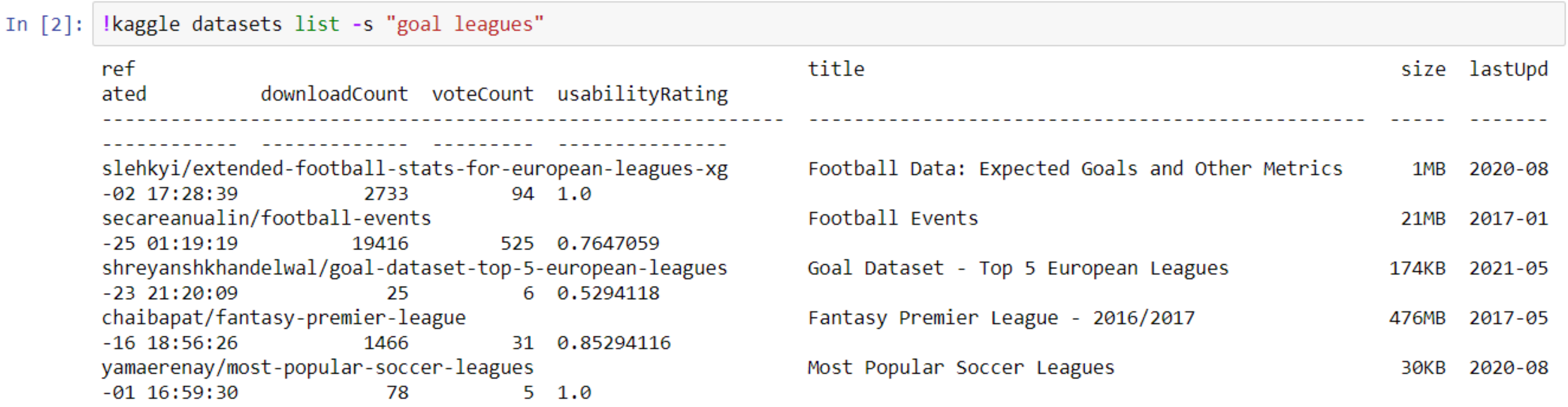
- Click '**Create New API Token**'. If the button doesn't work, click 'Expire API Token' first.
  - The browser will download the `kaggle.json` file to your Downloads folder.
- The Kaggle API by default assumes that the `kaggle.json` file is located in the folder:
  - `~/.kaggle/` (*Linux/Mac*)
  - `C:\Users\<Windows-username>\.kaggle\` (*Windows*)
    - If the folder doesn't exist, create it first with the `mkdir` command in the shell/command line.
    - Move the `kaggle.json` file to that folder (using File/Windows Explorer or via the mv or move command in the shell).
- The `kaggle.json` file contains the Kaggle username and the key string associated with that username. Therefore, in practice, this file must be secured so that it is not accessed by unauthorized parties.

Data Acquisition with API from Kaggle (4)

- The Kaggle API has four commands:
  - `kaggle competitions {list, files, download, submit, submissions, leaderboard}`
  - `kaggle datasets {list, files, download, create, version, init}`
  - `kaggle kernels {list, init, push, pull, output, status}`
  - `kaggle config {view, set, unset}`
- Kaggle API documentation can be viewed at https://github.com/Kaggle/kaggle-api

Data Acquisition with API from Kaggle (5)

- To perform a dataset search: `kaggle datasets list -s <keyword>`
  - If there is a problem with access failure, etc., you can try by regenerating the API Token.
- The dataset name/identifier is in the ref column of the search output table. For example, we want to download "Goal Dataset – Top 5 European Leagues", then the name of the dataset is `shreyanshkhandelwal/goal-dataset-top-5-european-leagues`

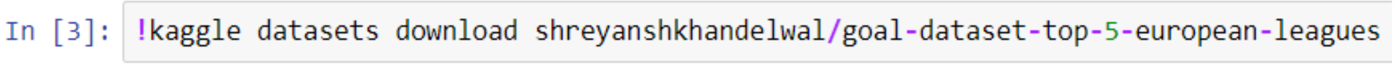


```
In [2]: !kaggle datasets list -s "goal leagues"
```

```
ref                                                            title                                               size  lastUpd
ated          downloadCount  voteCount  usabilityRating
-------------------------------------------------------------  --------------------------------------------------  -----  -------
------------  -------------  ---------  ---------------
slehkyi/extended-football-stats-for-european-leagues-xg        Football Data: Expected Goals and Other Metrics       1MB  2020-08
-02 17:28:39           2733         94  1.0
secareanualin/football-events                                  Football Events                                      21MB  2017-01
-25 01:19:19          19416        525  0.7647059
shreyanshkhandelwal/goal-dataset-top-5-european-leagues        Goal Dataset - Top 5 European Leagues               174KB  2021-05
-23 21:20:09             25          6  0.5294118
chaibapat/fantasy-premier-league                               Fantasy Premier League - 2016/2017                  476MB  2017-05
-16 18:56:26           1466         31  0.85294116
yamaerenay/most-popular-soccer-leagues                         Most Popular Soccer Leagues                          30KB  2020-08
-01 16:59:30             78          5  1.0
```

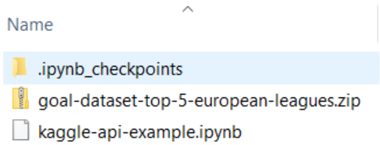
Data Acquisition with API from Kaggle (6)

- Download the desired dataset with the `kaggle datasets download` command



```
In [3]: !kaggle datasets download shreyanshkhandelwal/goal-dataset-top-5-european-leagues
```

- The dataset will be downloaded in the active folder as a compressed zip file.



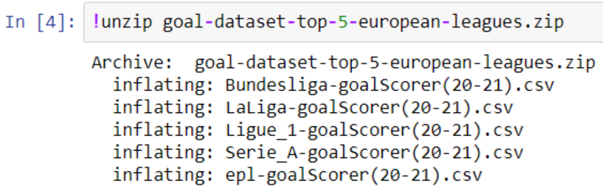
Name

.ipynb_checkpoints

goal-dataset-top-5-european-leagues.zip

kaggle-api-example.ipynb

- Next, we extract the dataset with the unzip command, and the dataset is in the form of csv files ready for use.



```
In [4]: !unzip goal-dataset-top-5-european-leagues.zip

Archive:  goal-dataset-top-5-european-leagues.zip
  inflating: Bundesliga-goalScorer(20-21).csv
  inflating: LaLiga-goalScorer(20-21).csv
  inflating: Ligue_1-goalScorer(20-21).csv
  inflating: Serie_A-goalScorer(20-21).csv
  inflating: epl-goalScorer(20-21).csv
```

- The csv file can be directly loaded into Pandas DataFrame.

Data Acquisition with Web Scraping

- **Web scraping** is the process of extracting data directly and automatically from a web page.
- This is one method for data scientists to obtain data that is only available on a web page and not available from other, more accessible sources.
- Challenges that will be encountered when doing web scraping:
  - The web scraping method is very dependent on the structure of the web page to be scraped.
  - The content and structure of websites often change dynamically.
  - Website content is generally opened within the scope of a certain access license.

Data Acquisition with Web Scraping

General steps (detailed examples can be found at https://realpython.com/beautiful-soup-web-scraper-python/)

- Specify the URL of the web page (HTML) to scrape.
- Use the `requests.get` function to access the URL. The HTML text will be stored in the **text attribute** of the object returned by `requests.get`
- Perform **HTML parsing** with the `beautifulsoup` library to obtain the desired data table (by extracting the relevant HTML elements).

Data Acquisition from a Relational Database (RDB)

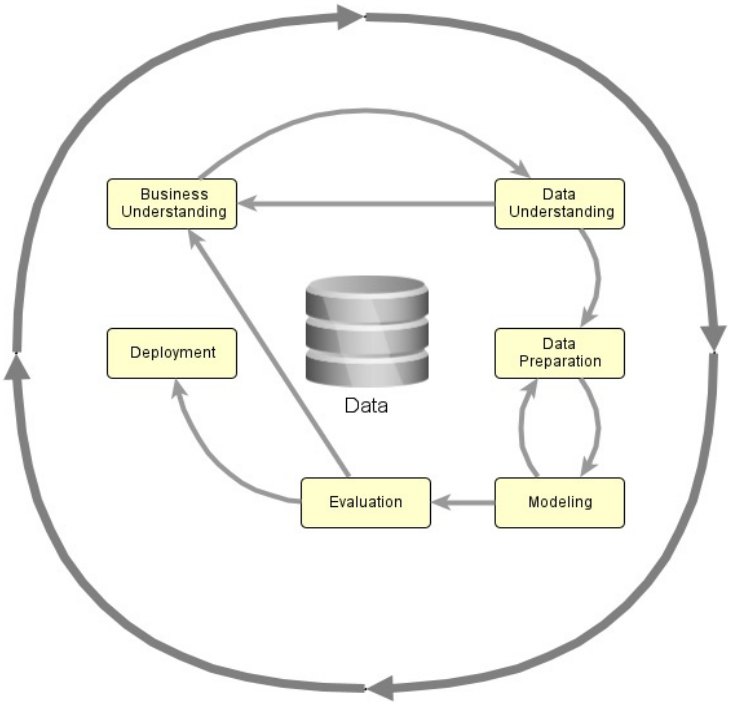
- Data can also be sourced from an organization's relational database (RDB).
- In practice, this is often done with the help of data engineers in organizations who are more competent in managing data and its infrastructure within the organization.
  - A data analyst can submit a data request to a data engineer who then fetches the data from an organization's internal database.
  - The data engineer will do this by executing SQL queries to the existing database system. The results can then be submitted to the data analyst manually (i.e. in the form of one or several CSV spreadsheet files).
- However, a data analyst can also do it himself if he has direct access to the relevant database.
- A data analyst can use a Python library called **SQLAlchemy**, or alternatively, use a bridge library specific to a particular relational database engine.

Data Acquisition from a Relational Database (RDB)

- General steps:
  - Import `pandas`
  - Import the RDB connector library, eg: `mysql.connector` for MySQL
  - Use the `connect` method of the RDB connector to open a connection to RDB.
  - Prepare SQL queries in strings.
  - Use `pandas.read_sql` with SQL query string arguments and an RDB connection to execute the SQL and load the results into a `DataFrame`.
  - Close connection.
- The process between opening and closing a connection is usually placed in a `try-except` block.
- Opening the connection requires credentials (username, password) to the RDB which are hardcoded directly. This can be hidden with security techniques not discussed here.
- A short example can be seen at: https://medium.com/analytics-vidhya/importing-data-from-a-mysql-database-into-pandas-data-frame-a06e392d27d7

Data Science Methodology

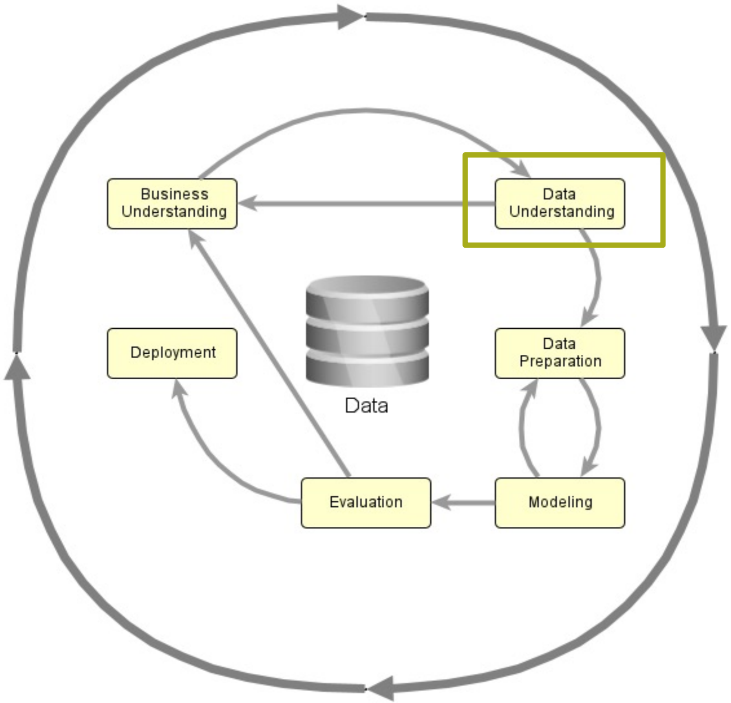
- The application of Artificial Intelligence (AI) in the real world is often carried out within the framework of a data science methodology which is also commonly adopted as steps for developing AI solutions.
- The data science methodology aims to systematically extract useful knowledge for solving business problems encountered.
- The formulation of the data science methodology adopts the **Cross-Industry Standard Process for Data Mining (CRISP-DM)** which can be stated in the several steps.
- In general, the entire methodology consists of a series of iterative processes.



CRISP-DM
Data Science Methodology

Data Understanding

- The data understanding stage is carried out after the business problem is defined as the result of the business understanding stage.
- In the data understanding stage, data collection and data analysis activities are carried out with the aim of **obtaining a complete picture of the data** that can be obtained as material for solving the business problem.
- Proceed to data preparation, if the initial understanding of the data is sufficient or return to business understanding if the definition of a business problem must be revised.



CRISP-DM
Data Science Methodology

Data Understanding

- **Data understanding** is a stage in data science methodology and AI development that aims to obtain an initial understanding of the data needed to solve a given business problem.
- Well-defined business problems serve as the basis for determining what data is needed.
- If an AI solution is developed to solve these business problems, then **data can be analogous to the raw material needed to build the AI solution**.
- Data understanding provides an initial description of:
    - the strengths of data,
    - deficiencies and limitations on data use,
    - the level of suitability of the data with the business problem to be solved,
    - data availability (open/closed, access fees, etc.)

Data Understanding Process Steps

Identify the parts in the business process where the data (existing or not) can affect the running of the business process.

Determine the organization's internal and external data sources, access mechanisms, and other things that can help or obstruct the acquisition of this data.

Assess each set of data specified above to determine the added business value that can be achieved if AI solutions can be realized with this data.

Identify other data from both internal and external sources of the organization that can bring improvements ot business processes through the built of AI solutions.

Data Understanding

- The realization of the four steps requires mastery of data collection and data analysis techniques.
- The first, second, and fourth steps involve a lot of **data collection** techniques, while the third step can be realized with the help of **data analysis** techniques.
- **Data analysis** techniques use **statistical** and **visualization** methods.

THANK YOU