Introduction to Data Science - Python

ENSISA CPB2

Ali El Hadi Ismail Fawaz
Germain Forestier

ENSISA, Université Haute-Alsace

September 6, 2025

Une école d'ingénieurs de l'Université de Haute-Alsace

Introduction

What is Data Science ?

Introduction

What is Data Science ?

- Extracting information from a given dataset.

Introduction

What is Data Science ?

- Extracting information from a given dataset.
- These information are often used in domains susch as:

Introduction

What is Data Science ?

- Extracting information from a given dataset.
- These information are often used in domains susch as:
  - Artificial Intelligence
  - Machine Learning
  - Deep Learning
  - Data Mining
  - Big Data

Introduction

What is Data Science ?

- Extracting information from a given dataset.
- These information are often used in domains susch as:
  - Artificial Intelligence
  - Machine Learning
  - Deep Learning
  - Data Mining
  - Big Data
  - . . .

Introduction

What is Data Science ?

- Extracting information from a given dataset.
- These information are often used in domains susch as:
  - Artificial Intelligence
  - Machine Learning
  - Deep Learning
  - Data Mining
  - Big Data
  - ...

Data Science can also be used in: Business Intelligence, Data Analytics, Visualization etc.

Introduction

Data Science

Introduction

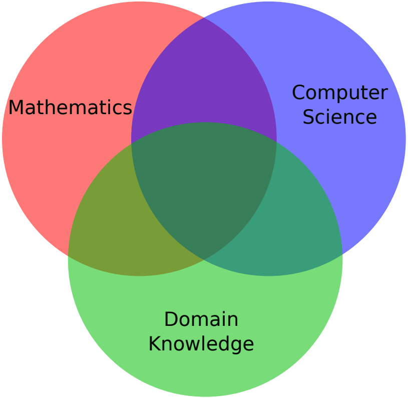
Data Science

- Originated in the late 90s

Introduction

Data Science

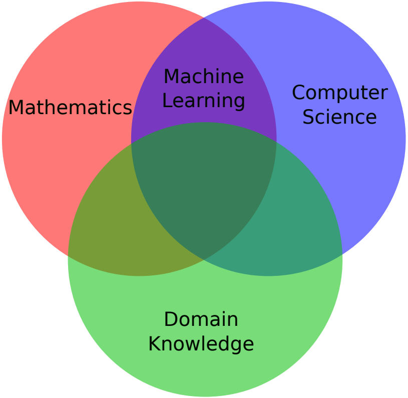
- Originated in the late 90s



Introduction

Data Science

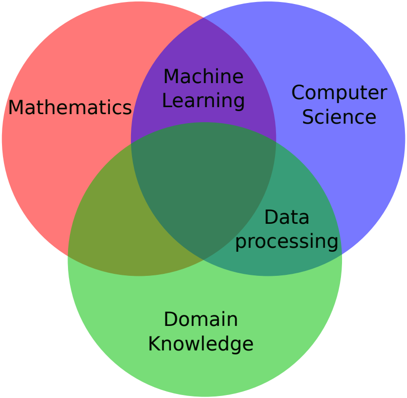
- Originated in the late 90s



Introduction

Data Science

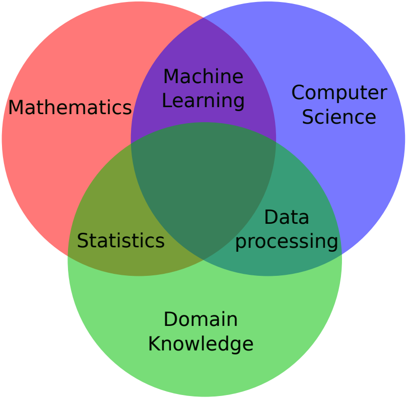
- Originated in the late 90s



Introduction

Data Science

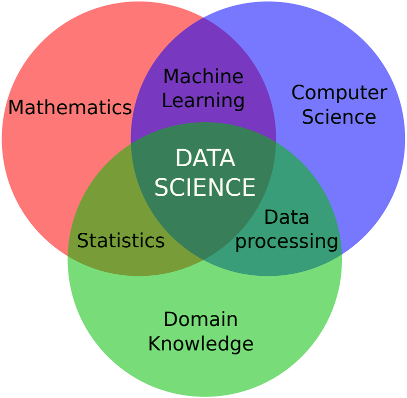
- Originated in the late 90s



Introduction

Data Science

- Originated in the late 90s



Introduction

Data Science

- Originated in the late 90s

- Its a collection of algorithms used in order to analyse, understand, process some data

Introduction

Data Science

- Originated in the late 90s

- Its a collection of algorithms used in order to analyse, understand, process some data
- Its a rapidly evolving science

Introduction

Most common domains that use Data Science:

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:
  - The concept of training a machine to achieve a given task

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
    - Simulate the human intelligence
    - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:
    - The concept of training a machine to achieve a given task
    - It is constrained on having examples to learn from, i.e. data

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:
  - The concept of training a machine to achieve a given task
  - It is constrained on having examples to learn from, i.e. data
- Deep Learning:

Introduction

Most common domains that use Data Science:

- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:
  - The concept of training a machine to achieve a given task
  - It is constrained on having examples to learn from, i.e. data
- Deep Learning:
  - A learning approach that is adapted to neural networks

Introduction

Most common domains that use Data Science:

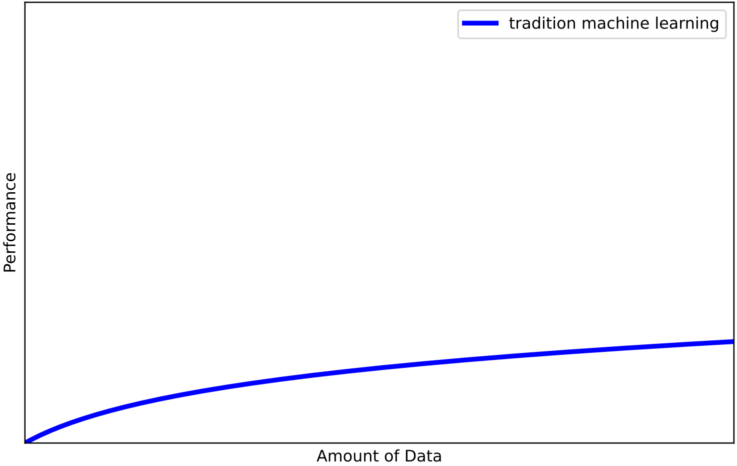
- Artificial Intelligence:
  - Simulate the human intelligence
  - It is found in robotics, game bots, chat bots, etc.
- Machine Learning:
  - The concept of training a machine to achieve a given task
  - It is constrained on having examples to learn from, i.e. data
- Deep Learning:
  - A learning approach that is adapted to neural networks
  - It is constrained on having a **lot** of examples, i.e. large amount of data

Introduction

We need more data

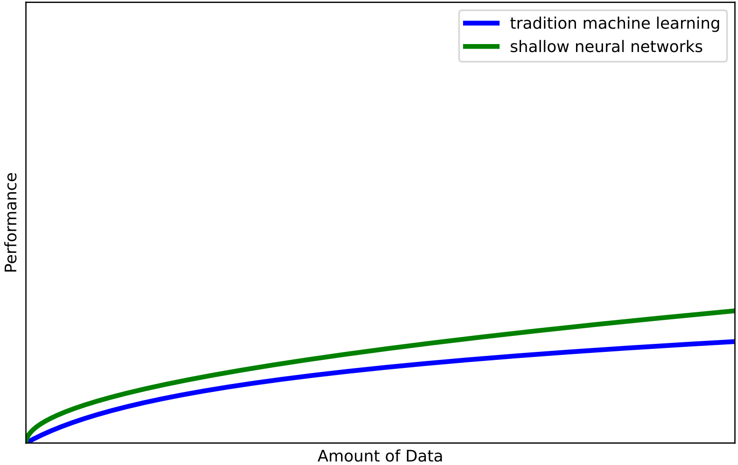
Introduction

We need more data



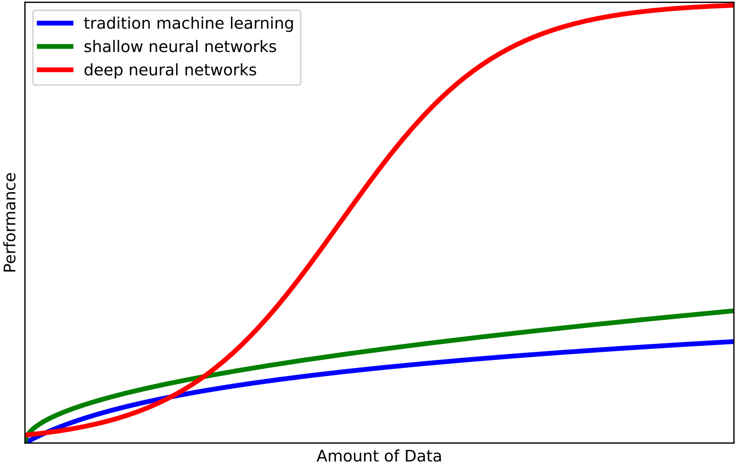
Introduction

We need more data



Introduction

We need more data



Introduction

Data Mining - Big Data

Introduction

Data Mining - Big Data

- Data Mining:

Introduction

Data Mining - Big Data

- Data Mining:
  - Extracting information from datasets

Introduction

Data Mining - Big Data

- Data Mining:
    - Extracting information from datasets
    - Use Machine Learning tools to analyse the data

Introduction

Data Mining - Big Data

- Data Mining:
  - Extracting information from datasets
  - Use Machine Learning tools to analyse the data
- Big Data:

Introduction

Data Mining - Big Data

- Data Mining:
  - Extracting information from datasets
  - Use Machine Learning tools to analyse the data
- Big Data:
  - A current definition of massive amount of data

Introduction

Data Mining - Big Data

- Data Mining:
  - Extracting information from datasets
  - Use Machine Learning tools to analyse the data
- Big Data:
  - A current definition of massive amount of data
  - Raises a question for the usage of existing learning methods

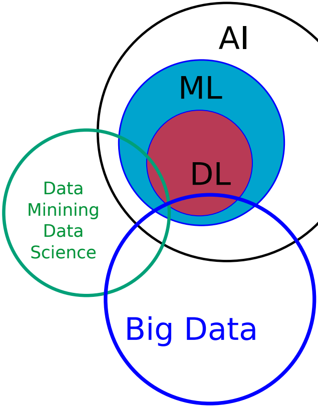
Introduction

Data Mining - Big Data

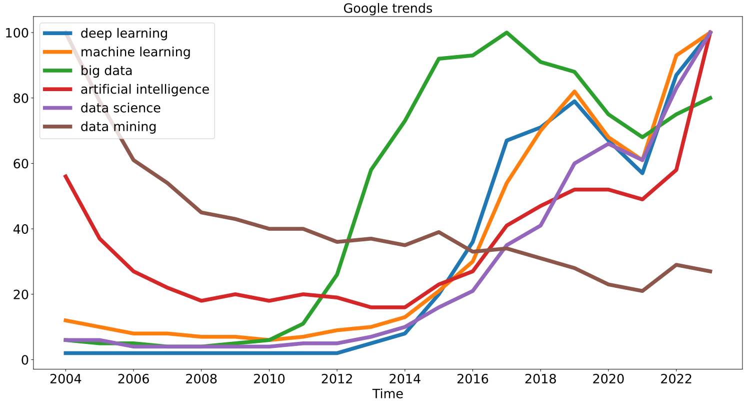
- Data Mining:
  - Extracting information from datasets
  - Use Machine Learning tools to analyse the data
- Big Data:
  - A current definition of massive amount of data
  - Raises a question for the usage of existing learning methods
  - The more we have data the better models we can learn

Introduction

Categories of Data Science



Categories of Data Science



Introduction

- **Most recently domains**

Introduction

- **Most recently domains**
    - Machine Learning, Big Data, AI, Data Science, Deep Learning: used a lot in industry and the news

Introduction

- **Most recently domains**
  - Machine Learning, Big Data, AI, Data Science, Deep Learning: used a lot in industry and the news
  - Can be easily over hyped

Introduction

- **Most recently domains**
  - Machine Learning, Big Data, AI, Data Science, Deep Learning: used a lot in industry and the news
  - Can be easily over hyped
  - Can be easily bashed for no reason

Introduction

- **Most recently domains**
  - Machine Learning, Big Data, AI, Data Science, Deep Learning: used a lot in industry and the news
  - Can be easily over hyped
  - Can be easily bashed for no reason
- **Be careful of fake news**

Introduction

- **Most recently domains**
  - Machine Learning, Big Data, AI, Data Science, Deep Learning: used a lot in industry and the news
  - Can be easily over hyped
  - Can be easily bashed for no reason
- **Be careful of fake news**
  - How Big Data will help feeding 9 billion person
  - AI can now foresee cancer years before it develops
  - Checkout how AI models can generate a Breaking Bad episode
  - Become a billionaire with Big Data ?

Introduction

Types of Data:

Introduction

Types of Data: Structured vs Unstructured Data

Introduction

Types of Data: Structured vs Unstructured Data

- Structured Data: easy to look for, highly organized with a specific format.

Introduction

Types of Data: Structured vs Unstructured Data

- Structured Data: easy to look for, highly organized with a specific format.
- Unstructured Data: unorganized with no specific format, very hard to search for: images, text, video etc.

Introduction

Types of Data: Structured vs Unstructured Data

- Structured Data: easy to look for, highly organized with a specific format.
- Unstructured Data: unorganized with no specific format, very hard to search for: images, text, video etc.

Example of structured data:

	Attribute 1	**Attribute 2**	**Attribute 3**
Instance 1	1.1	dog	True
Instance 2	2.1	cat	False
Instance 3	3.0	lion	True

Introduction

Types of Data: Structured vs Unstructured Data

- Structured Data: easy to look for, highly organized with a specific format.
- Unstructured Data: unorganized with no specific format, very hard to search for: images, text, video etc.

Example of structured data:

	Attribute 1	**Attribute 2**	**Attribute 3**
Instance 1	1.1	dog	True
Instance 2	2.1	cat	False
Instance 3	3.0	lion	True

- Types of data can be real, integer, character, string, boolean etc.

Introduction

Types of Data: Structured vs Unstructured Data

- Structured Data: easy to look for, highly organized with a specific format.
- Unstructured Data: unorganized with no specific format, very hard to search for: images, text, video etc.

Example of structured data:

	Attribute 1	**Attribute 2**	**Attribute 3**
Instance 1	1.1	dog	True
Instance 2	2.1	cat	False
Instance 3	3.0	lion	True

- Types of data can be real, integer, character, string, boolean etc.
- Can be found in databses, clouds etc.

Introduction

Dataset Example: IRIS

Introduction

Dataset Example: IRIS

- IRIS samples: given the length and width of sepal and petal of an IRIS

Introduction

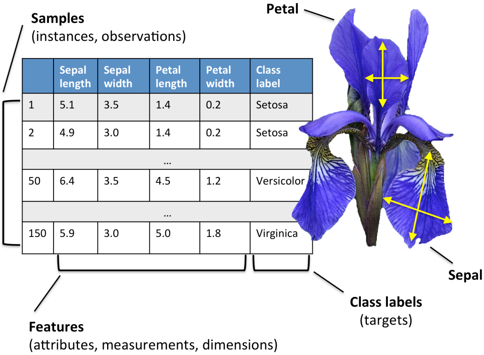
Dataset Example: IRIS

- IRIS samples: given the length and width of sepal and petal of an IRIS
- $\xrightarrow{goal}$ predict the iris type

Introduction

Dataset Example: IRIS

- IRIS samples: given the length and width of sepal and petal of an IRIS
- $\xrightarrow{goal}$ predict the iris type



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
…					
50	6.4	3.5	4.5	1.2	Versicolor
…					
150	5.9	3.0	5.0	1.8	Virginica

source: https://rpubs.com/wjholst/322258

Introduction

Some IRIS samples:

Introduction

Some IRIS samples:

sepal length	sepal width	petal length	petal width	iris type
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
7.3	2.9	6.3	1.8	virginica

Introduction

Some IRIS samples:

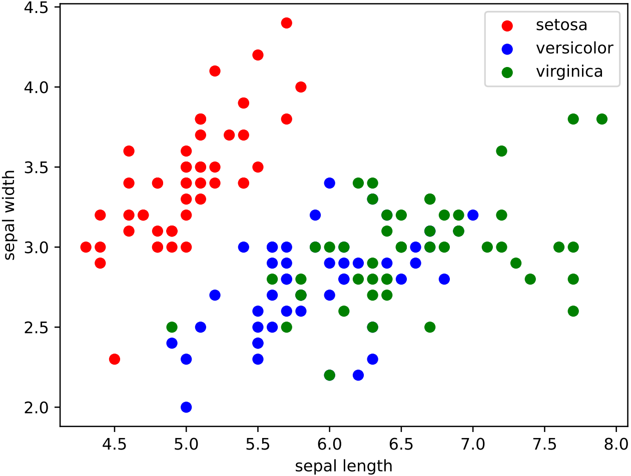
sepal length	sepal width	petal length	petal width	iris type
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
7.3	2.9	6.3	1.8	virginica
7.7	2.6	6.9	2.3	?

Introduction

Understanding the data starts by visualizing it:

Introduction

Understanding the data starts by visualizing it:

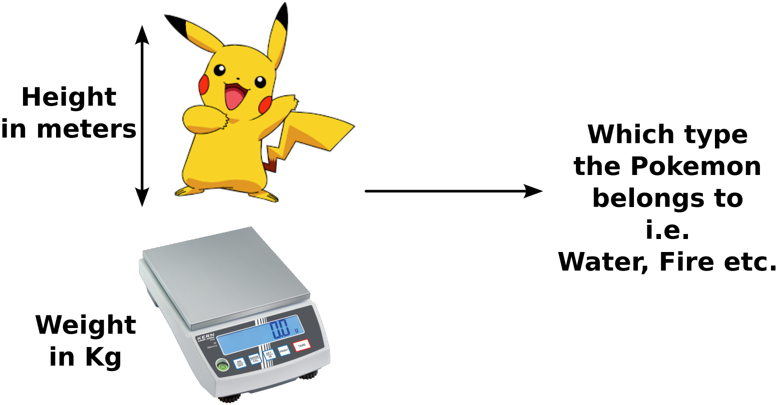


Introduction

Data Example 2: Pokemon Types

Introduction

Data Example 2: Pokemon Types



Height in meters

Weight in Kg

Which type the Pokemon belongs to i.e. Water, Fire etc.

Introduction

Pokemon Example

Introduction

Pokemon Example

Pokemon Name	Height	Weight	Type
Bulbasaur	0.7	6.9	Grass
Charmander	0.6	8.5	Fire
Squirtle	0.5	9.0	Water
Caterpie	0.3	2.9	Bug

Introduction

Pokemon Example

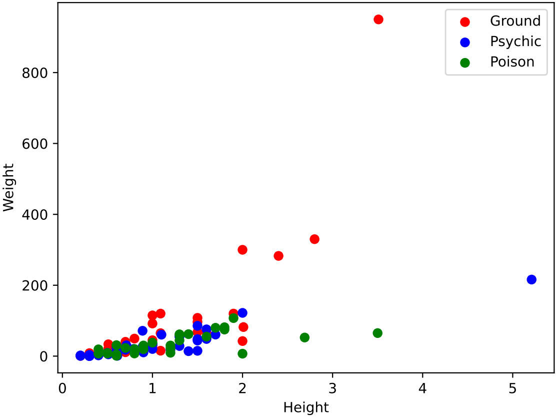
Pokemon Name	Height	Weight	Type
Bulbasaur	0.7	6.9	Grass
Charmander	0.6	8.5	Fire
Squirtle	0.5	9.0	Water
Caterpie	0.3	2.9	Bug
Charizard	1.7	90.5	?

Introduction

Visualization of types: Ground, Psychic and Poison.

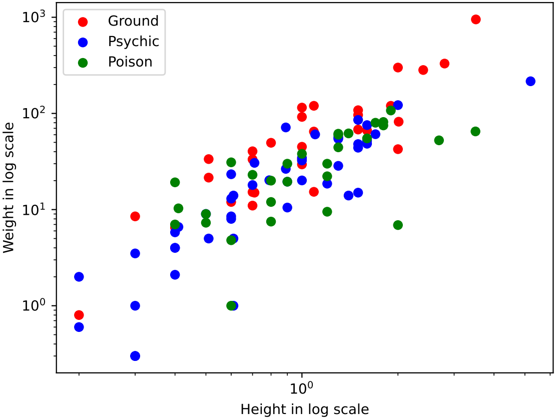
Introduction

Visualization of types: Ground, Psychic and Poison.



Introduction

Visualization of types: Ground, Psychic and Poison.



Introduction

Different tasks to solve in Data Mining

Introduction

Different tasks to solve in Data Mining

- Classification: predict the discrete value of the class label

Introduction

Different tasks to solve in Data Mining

- Classification: predict the discrete value of the class label
- Regression: predict the continuous value of the label

Introduction

Different tasks to solve in Data Mining

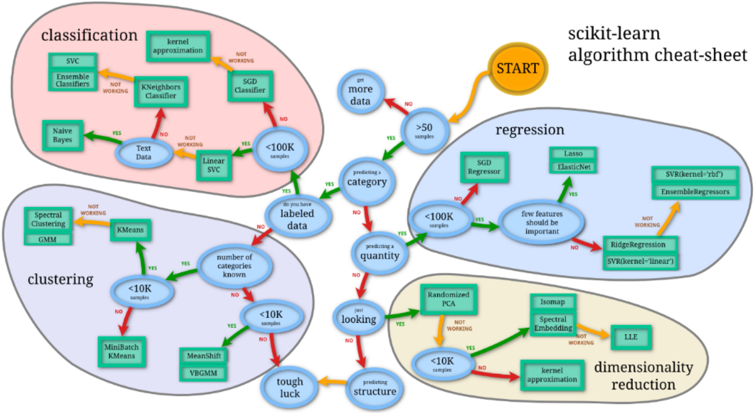
- Classification: predict the discrete value of the class label
- Regression: predict the continuous value of the label
- Clustering: discover partitions without having the labels

Introduction

Different tasks to solve in Data Mining

- Classification: predict the discrete value of the class label
- Regression: predict the continuous value of the label
- Clustering: discover partitions without having the labels

source: https://scikit-learn.org/



Introduction

Where to find data ?

Introduction

Where to find data ?

- Specialized websites: `https://www.kaggle.com/`

Introduction

Where to find data ?

- Specialized websites: `https://www.kaggle.com/`
- Open data websites: `https://www.data.gouv.fr/fr/`

Introduction

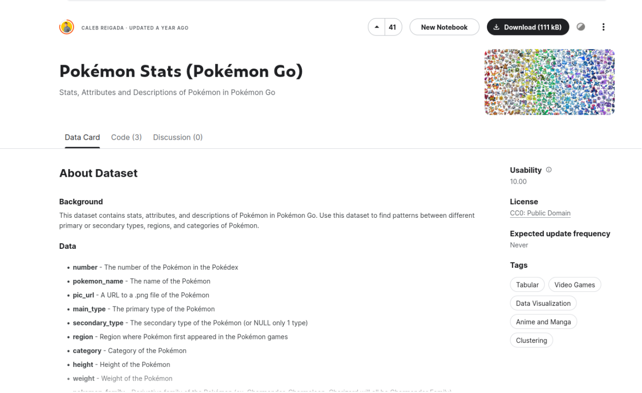
Where to find data ?

- Specialized websites: `https://www.kaggle.com/`
- Open data websites: `https://www.data.gouv.fr/fr/`
- Research data: `https://datasetsearch.research.google.com/`

Introduction

Where to find data ?

- Specialized websites: `https://www.kaggle.com/`
- Open data websites: `https://www.data.gouv.fr/fr/`
- Research data: `https://datasetsearch.research.google.com/`



Pokémon Stats (Pokémon Go)

Stats, Attributes and Descriptions of Pokémon in Pokémon Go

About Dataset

Background

This dataset contains stats, attributes, and descriptions of Pokémon in Pokémon Go. Use this dataset to find patterns between different primary or secondary types, regions, and categories of Pokémon.

Data

- **number** - The number of the Pokémon in the Pokédex
- **pokemon_name** - The name of the Pokémon
- **pic_url** - A URL to a .png file of the Pokémon
- **main_type** - The primary type of the Pokémon
- **secondary_type** - The secondary type of the Pokémon (or NULL only 1 type)
- **region** - Region where Pokémon first appeared in the Pokémon games
- **category** - Category of the Pokémon
- **height** - Height of the Pokémon
- **weight** - Weight of the Pokémon
- **pokemon_family** - Derivative family of the Pokémon (ex. Charmander, Charmeleon, Charizard will all be Charmander Family)

Usability

10.00

License

CC0: Public Domain

Expected update frequency

Never

Tags

source: https://www.kaggle.com/datasets/calebreigada/pokemon