Introduction to Hugging Face

WORKING WITH HUGGING FACE

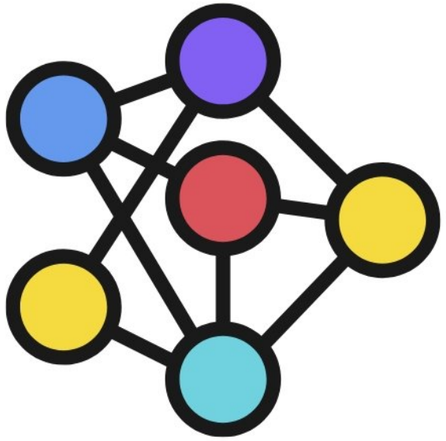
Jacob H. Marquez
Lead Data Engineer

datacamp

The home of the AI community...

Hugging Face

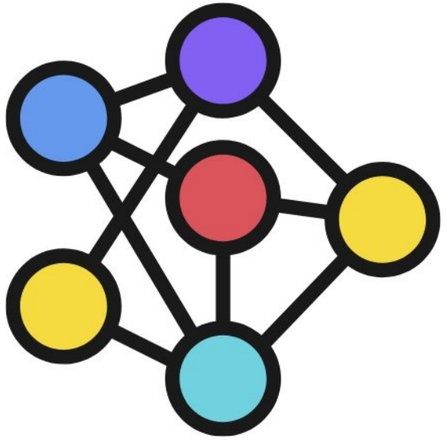
The home of the AI community...



Models

Hugging Face

The home of the AI community...

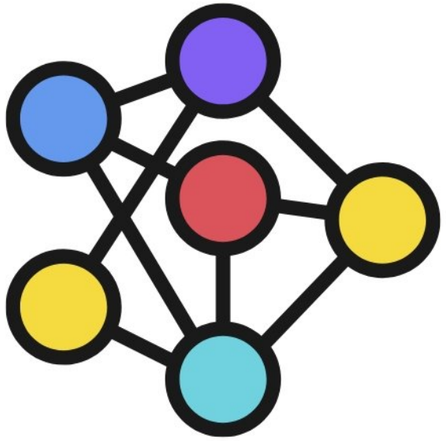


Models

Hugging Face

Datasets

The home of the AI community...



Models

Hugging Face

Datasets

Applications

[1] https://huggingface.co/docs/hub/index

Core ML Libraries

- **Transformers**
  State-of-the-art ML for PyTorch, TensorFlow, JAX
- **Diffusers**
  State-of-the-art Diffusion models in PyTorch
- **Datasets**
  Access & share datasets for any ML tasks
- **Transformers.js**
  State-of-the-art ML running directly in your browser
- **Tokenizers**
  Fast tokenizers optimized for research & production
- **Evaluate**
  Evaluate and compare models performance
- **timm**
  State-of-the-art vision models: layers, optimizers, and utilities
- **Sentence Transformers**
  Embeddings, Retrieval, and Reranking

Training & Optimization

- **PEFT**
  Parameter-efficient finetuning for large language models
- **Accelerate**
  Train PyTorch models with multi-GPU, TPU, mixed precision
- **Optimum**
  Optimize HF Transformers for faster training/inference
- **AWS Trainium & Inferentia**
  Train/deploy Transformers/Diffusers on AWS
- **TRL**
  Train transformers LMs with reinforcement learning
- **Safetensors**
  Safe way to store/distribute neural network weights
- **Bitsandbytes**
  Optimize and quantize models with bitsandbytes
- **Lighteval**
  All-in-one toolkit to evaluate LLMs across multiple backends

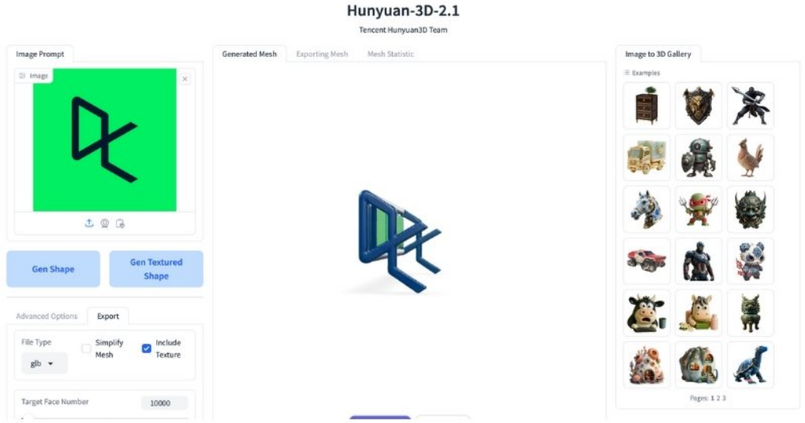
Community and open-source heroes

Models

- deepseek-ai/DeepSeek-R1
  Text Generation • 685B • Updated Mar 27 • 885k • 12.4k
- black-forest-labs/FLUX.1-dev
  Text-to-Image • Updated 10 days ago • 1.53M • 10.8k
- CompVis/stable-diffusion-v1-4
  Text-to-Image • Updated Aug 23, 2023 • 3.46M • 6.86k
- stabilityai/stable-diffusion-xl-base-1.0
  Text-to-Image • Updated Oct 30, 2023 • 2.56M • 6.71k
- meta-llama/Meta-Llama-3-8B
  Text Generation • 8B • Updated Sep 27, 2024 • 357k • 6.24k
- bigscience/bloom
  Text Generation • 176B • Updated Jul 28, 2023 • 6.66k • 4.92k
- stabilityai/stable-diffusion-3-medium
  Text-to-Image • Updated Aug 12, 2024 • 11.6k • 4.79k
- hexgrad/Kokoro-82M
  Text-to-Speech • Updated Apr 10 • 1.58M • 4.61k
- openai/whisper-large-v3
  Automatic Speech Recognition • 2B • Updated Aug 12, 2024 • 2.9M • 4.6k
- meta-llama/Llama-2-7b-chat-hf
  Text Generation • 7B • Updated Apr 17, 2024 • 1.05M • 4.48k

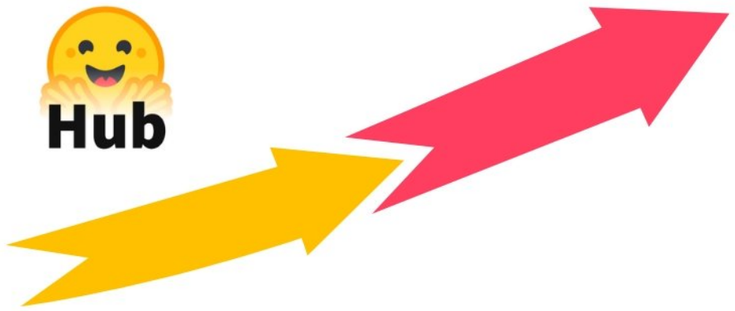
Datasets

- m-a-p/FineFineWeb
  Viewer • Updated Dec 19, 2024 • 4.89B • 3.28M • 65
- huggingface/documentation-images
  Viewer • Updated 3 days ago • 52 • 3M • 73
- permutans/fineweb-bbc-news
  Viewer • Updated Jan 27 • 15.9M • 1.5M • 20
- hallucinations-leaderboard/results
  Updated Oct 31, 2024 • 1.42M • 2
- huggingface/badges
  Viewer • Updated Apr 8 • 1 • 1.39M • 45
- jat-project/jat-dataset-tokenized
  Viewer • Updated Dec 22, 2023 • 32M • 1.25M • 1
- KakologArchives/KakologArchives
  Updated 1 minute ago • 999k • 16
- lavita/medical-qa-shared-task-v1-toy
  Viewer • Updated Jul 20, 2023 • 64 • 911k • 19
- huggingchat/models-logo
  Viewer • Updated May 27 • 13 • 905k • 4
- adams-story/datacomp200m
  Viewer • Updated Jul 19, 2023 • 213M • 793k • 1
- TAUR-Lab/Taur_CoT_Analysis_Project___gpt-…
  Viewer • Updated Oct 19, 2024 • 82.2k • 741k
- Salesforce/wikitext
  Viewer • Updated Jan 4, 2024 • 3.71M • 691k • 469
- openai/gsm8k
  Viewer • Updated Jan 4, 2024 • 17.6k • 513k • 789
- nvidia/PhysicalAI-Robotics-GR00T-X-Embodi…
  Updated May 15 • 475k • 133
- agents-course/course-images
  Viewer • Updated 17 days ago • 2 • 468k • 13
- princeton-nlp/SWE-bench_Verified
  Viewer • Updated Feb 18 • 500 • 438k • 183

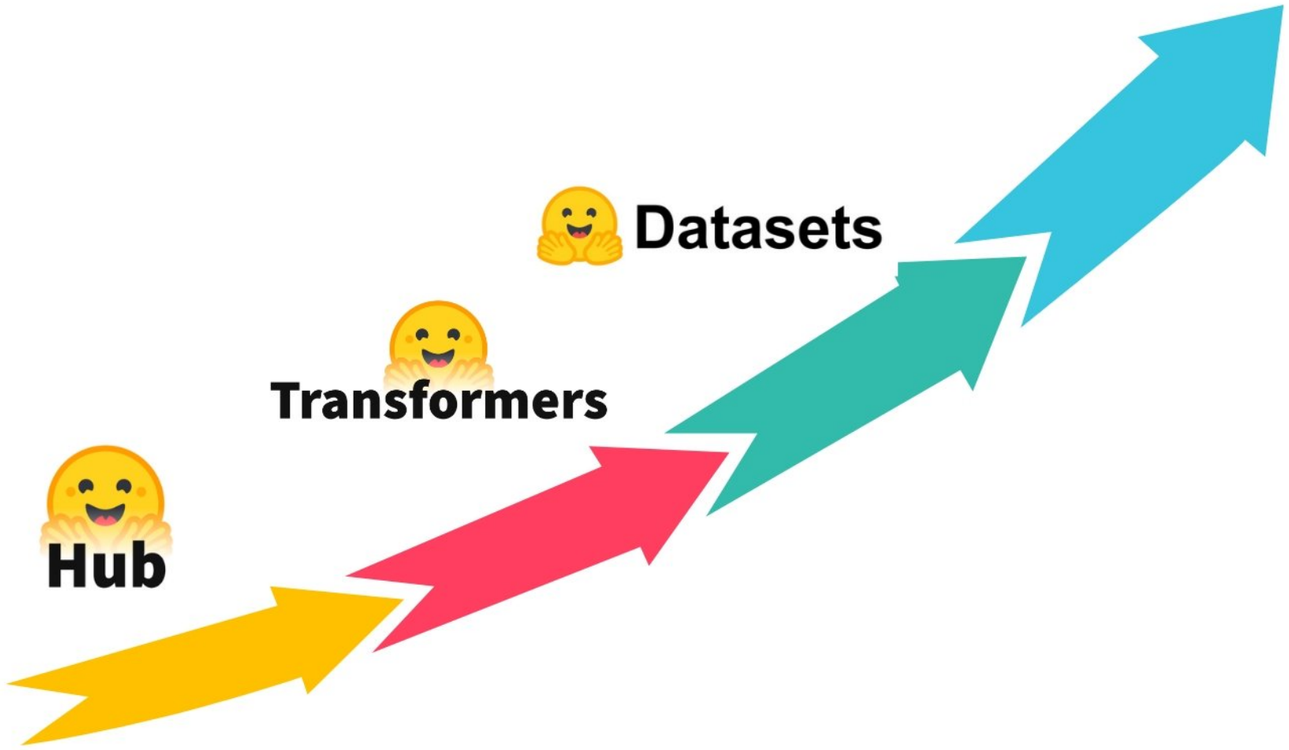


Applications

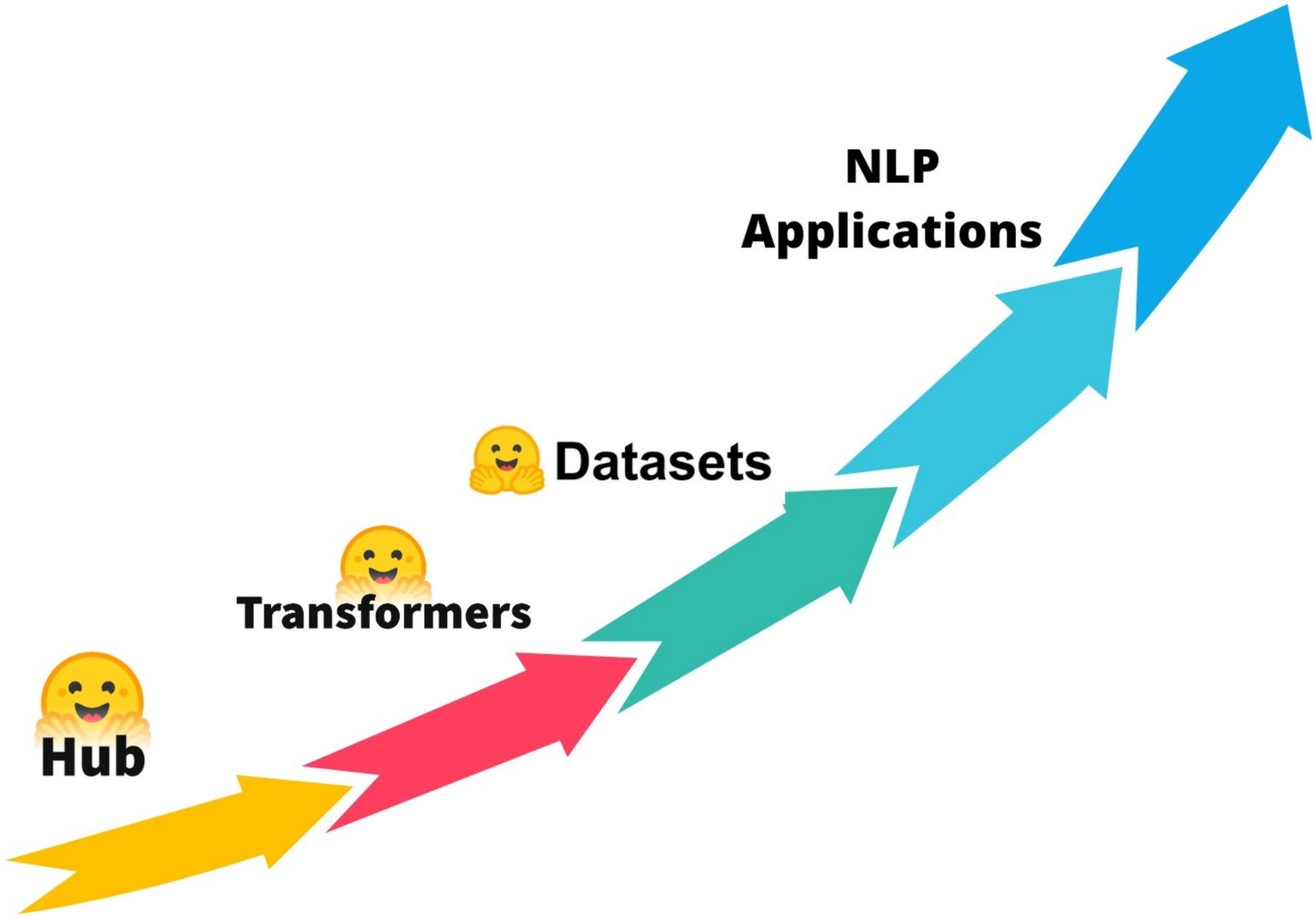
Coming up...



Coming up...

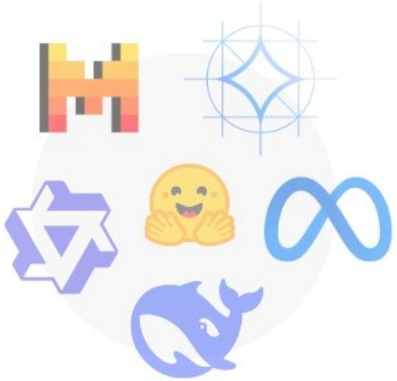


Coming up...

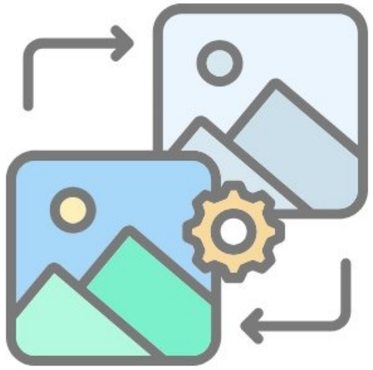


The journey beyond!

Working with Hugging Face



Introduction to LLMs in Python



Multi-Modal Models with Hugging Face

Efficient Model Training with PyTorch

Let's practice!

WORKING WITH HUGGING FACE

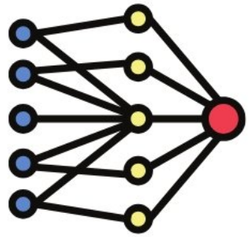
Running Hugging Face models

WORKING WITH HUGGING FACE

Jacob H. Marquez
Lead Data Engineer

datacamp

Inference with Hugging Face

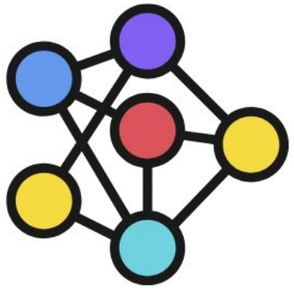


Dog Image Classifier

Pomeranian

Complete the story:

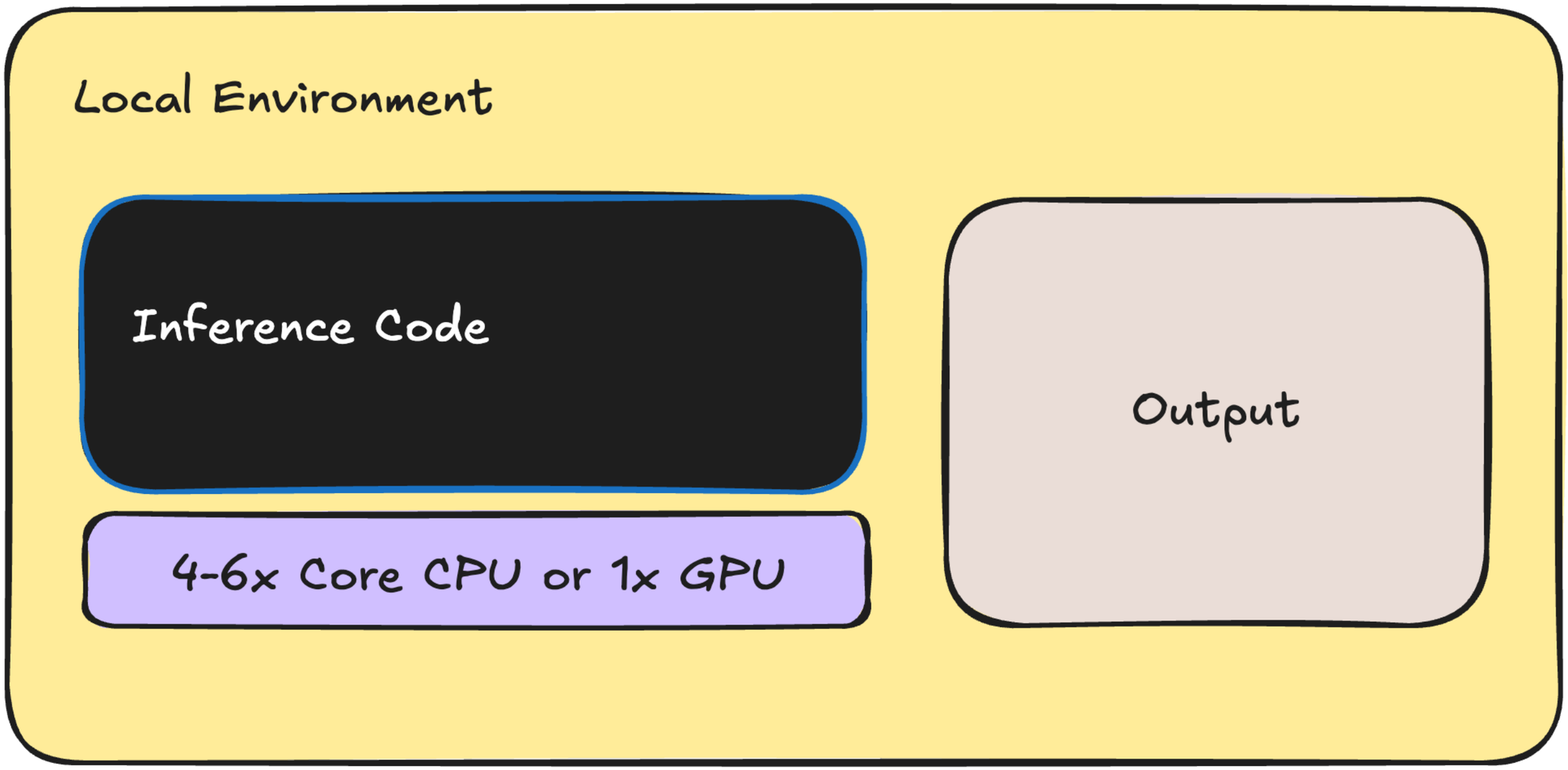
Jack went to school one day...



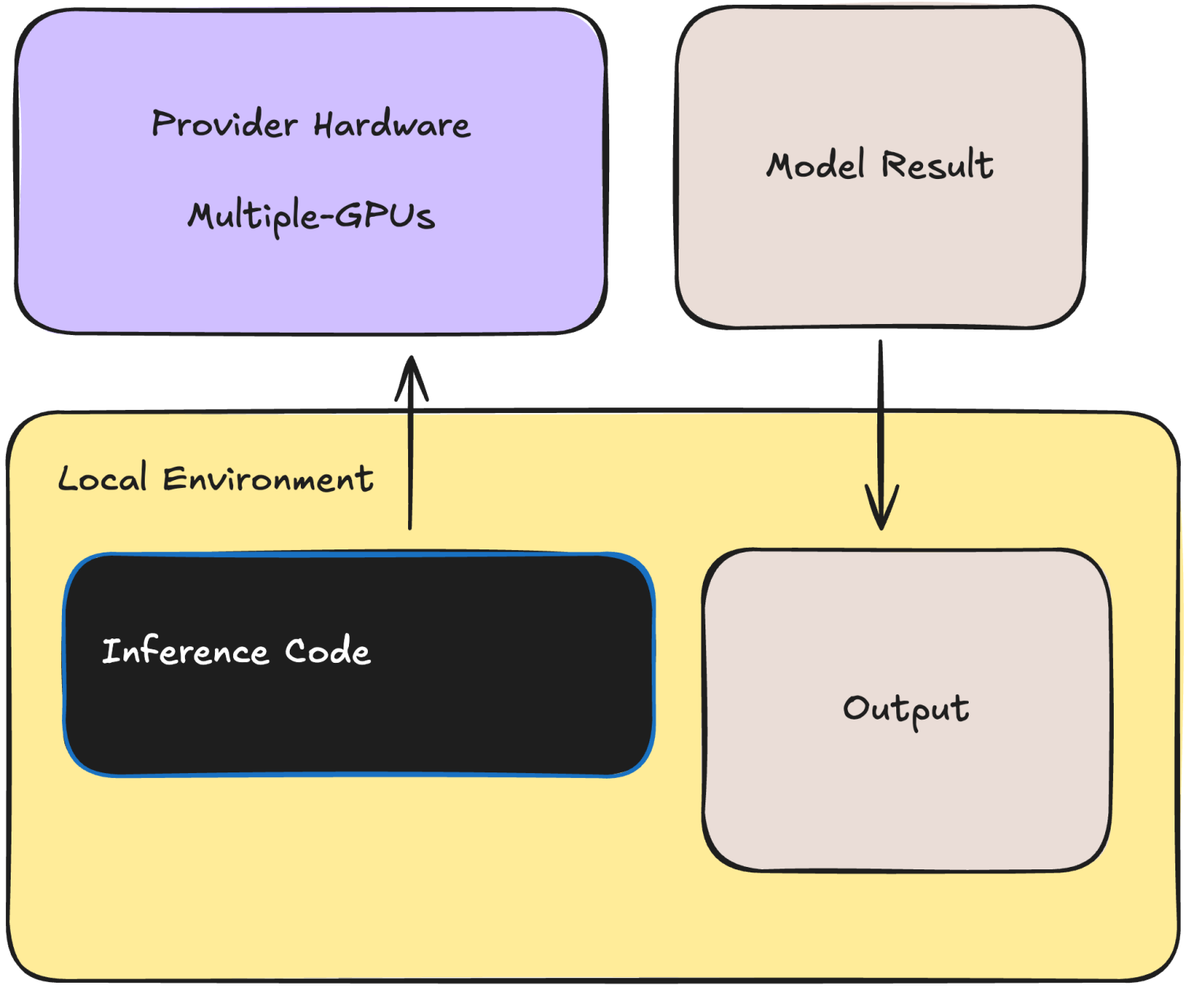
Text Generator

...with a flutter in his stomach and a notebook full of dreams.

Local inference



Inference providers



[1] https://huggingface.co/docs/inference-providers/en/index

Inference with Hugging Face

Local Inference

- Free
- 😄 Convenient
- 😴 Slow and resource-intensive

Inference Providers

- Fast
- Free to get started

[1] https://huggingface.co/docs/inference-providers/en/index

Introduction to the Transformers Library

- Simplifies working with **pre-trained models**

Transformers

[1] https://github.com/huggingface/transformers

The pipeline

```python
from transformers import pipeline

gpt2_pipeline = pipeline(task="text-generation", model="openai-community/gpt2")

print(gpt2_pipeline("What if AI"))
```

```
[{'generated_text': 'What if AI wouldn\'t be used?"\n\nI had to agree with your
  theory. If a machine\'s learning algorithm is a perfect match for all of a
  human\'s needs, then you may not have a problem with it. My problem was whether'}]
```

[1] Model Card: https://huggingface.co/openai-community/gpt2

Adjusting Pipeline Parameters

```python
from transformers import pipeline

gpt2_pipeline = pipeline(task="text-generation", model="openai-community/gpt2")

results = gpt2_pipeline("What if AI", max_new_tokens=10, num_return_sequences=2)

for result in results:
    print(result['generated_text'])
```

```
What if AI had never existed?
What if AI could be really smarter than us?
```

Using inference providers

```
import os
from huggingface_hub import InferenceClient

client = InferenceClient(
    provider="together",
    api_key=os.environ["HF_TOKEN"],
)
```

[1] https://huggingface.co/docs/inference-providers/en/index

```
completion = client.chat.completions.create(
    model="deepseek-ai/DeepSeek-V3",
    messages=[
        {
            "role": "user",
            "content": "What is the capital of France?"
        }
    ],
)
```

[1] https://huggingface.co/docs/inference-providers/en/index

```python
print(completion.choices[0].message)
```

```
The capital of France is **Paris**. It is known for its iconic landmarks such as
the Eiffel Tower, the Louvre Museum, and Notre-Dame Cathedral.

Would you like any additional information about Paris or France?
```

Let's practice!

WORKING WITH HUGGING FACE

Hugging Face Datasets

WORKING WITH HUGGING FACE

Jacob H. Marquez
Lead Data Engineer

datacamp

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Community | Docs | Pricing

Main | Tasks | Libraries | Languages | Licenses | Other

Modalities

3D | Audio | Document | Geospatial | Image | Tabular | Text | Time-series | Video

Size (rows)

< 1K — > 1T

Format

json | csv | parquet | imagefolder | soundfolder | webdataset | text | arrow

Datasets 443,200 | Filter by name | Full-text search | Sort: Trending

fka/awesome-chatgpt-prompts
Viewer • Updated Jan 6 • 203 • 22.9k • 8.19k

HuggingFaceFW/fineweb-2
Viewer • Updated 11 days ago • 5.02B • 38.3k • 574

facebook/seamless-interaction
Updated 10 days ago • 1 • 89

marcelbinz/Psych-101
Viewer • Updated Nov 2, 2024 • 60.1k • 218 • 75

black-forest-labs/kontext-bench
Viewer • Updated 11 days ago • 1.03k • 38

FreedomIntelligence/ShareGPT-4o-Image
Viewer • Updated 6 days ago • 92.3k • 75 • 70

sequelbox/Celestia3-DeepSeek-R1-0528
Viewer • Updated 5 days ago • 91k • 19

institutional/institutional-books-1.0
Viewer • Updated 21 days ago • 983k • 38.2k • 220

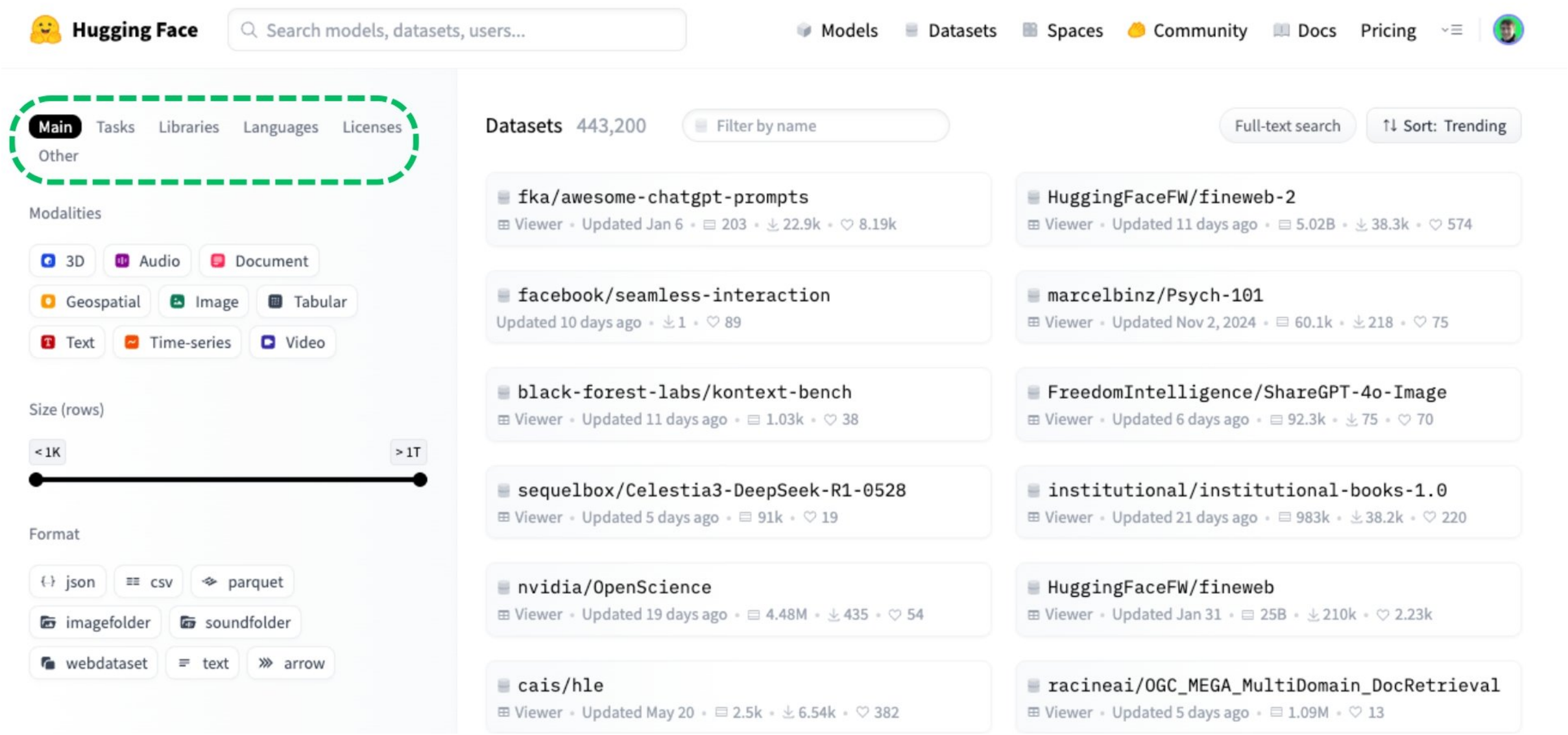
nvidia/OpenScience
Viewer • Updated 19 days ago • 4.48M • 435 • 54

HuggingFaceFW/fineweb
Viewer • Updated Jan 31 • 25B • 210k • 2.23k

cais/hle
Viewer • Updated May 20 • 2.5k • 6.54k • 382

racineai/OGC_MEGA_MultiDomain_DocRetrieval
Viewer • Updated 5 days ago • 1.09M • 13

[1] https://huggingface.co/datasets

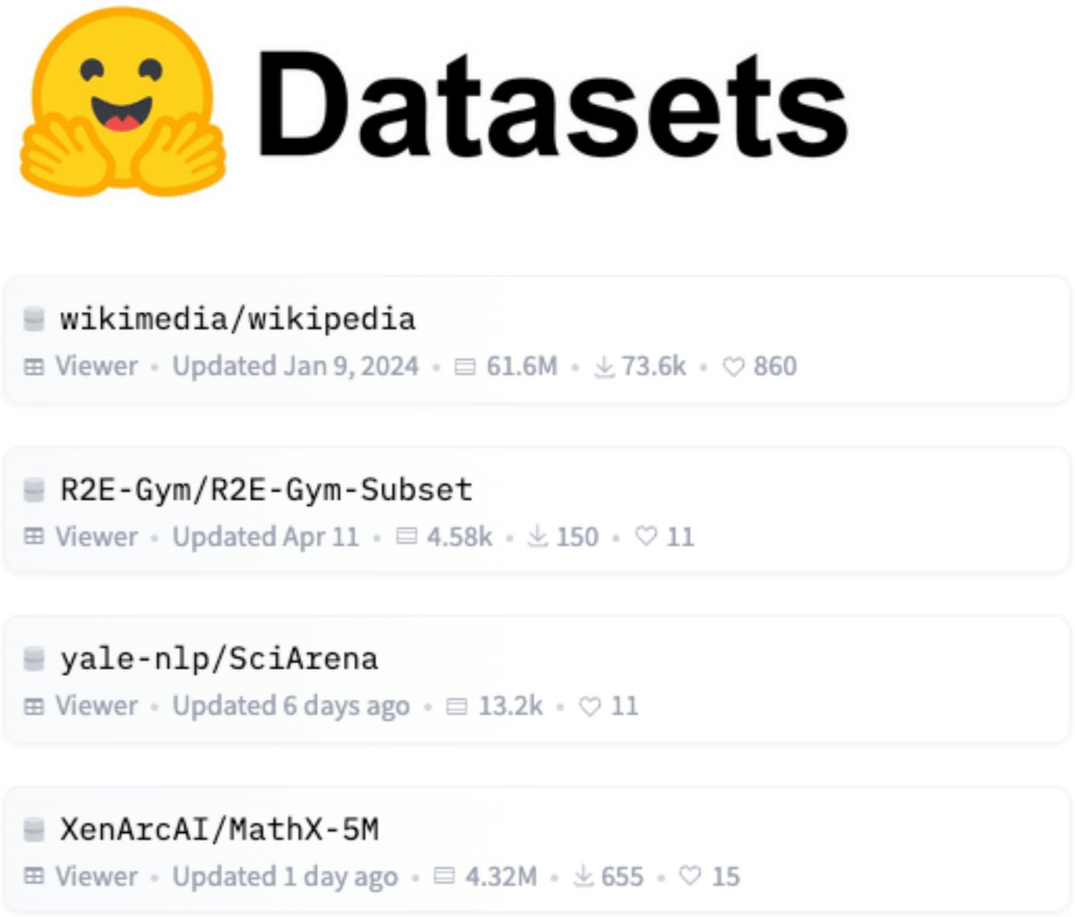


[1] https://huggingface.co/datasets

Installing Datasets Package

```
pip install datasets
```

- Access
- Download
- Use
- Share



[1] https://huggingface.co/docs/datasets/loading

Downloading a dataset

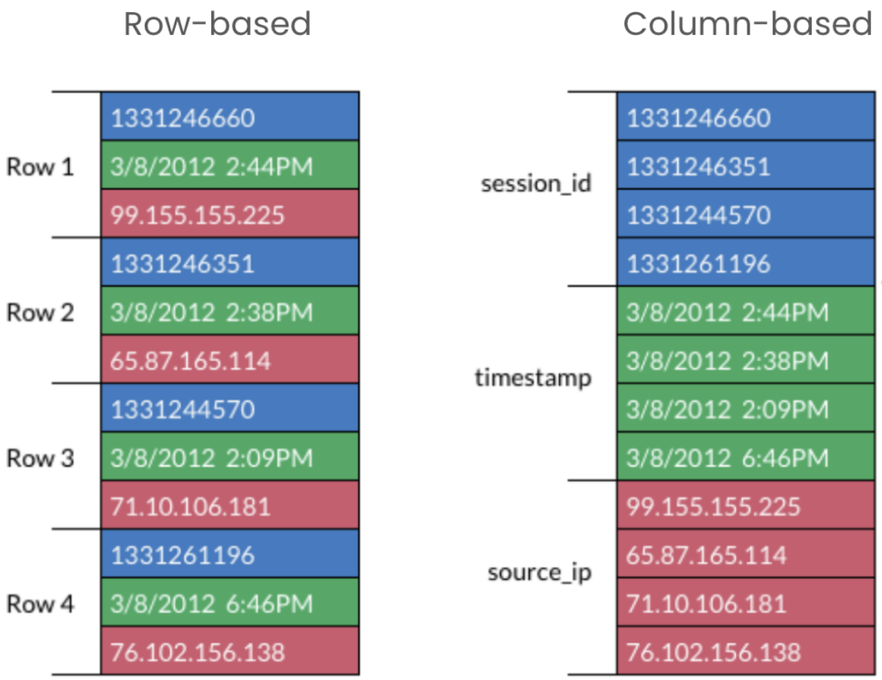
```
from datasets import load_dataset

data = load_dataset("IVN-RIN/BioBERT_Italian")
```

Split parameter

```
data = load_dataset("IVN-RIN/BioBERT_Italian", split="train")
```

[1] https://huggingface.co/docs/datasets/v2.15.0/loading

Apache Arrow dataset formats



[1] https://arrow.apache.org/overview/

Data manipulation

```
data = load_dataset("IVN-RIN/BioBERT_Italian", split="train")

# Filter for pattern " bella "
filtered = data.filter(lambda row: " bella " in row['text'])
print(filtered)
```

```
Dataset({
    features: ['text'],
    num_rows: 1122
})
```

[1] https://huggingface.co/docs/datasets/process#select-and-filter

Data manipulation

```
# Select the first two rows
sliced = filtered.select(range(2))

print(sliced)
```

```
Dataset({features: ['text'], num_rows: 2})
```

```
# Extract the 'text' for the first row
print(sliced[0]['text'])
```

```
Concentrazioni atmosferiche di PCDD/PCDF...
```

[1] https://huggingface.co/docs/datasets/process#select-and-filter

Let's practice!

WORKING WITH HUGGING FACE