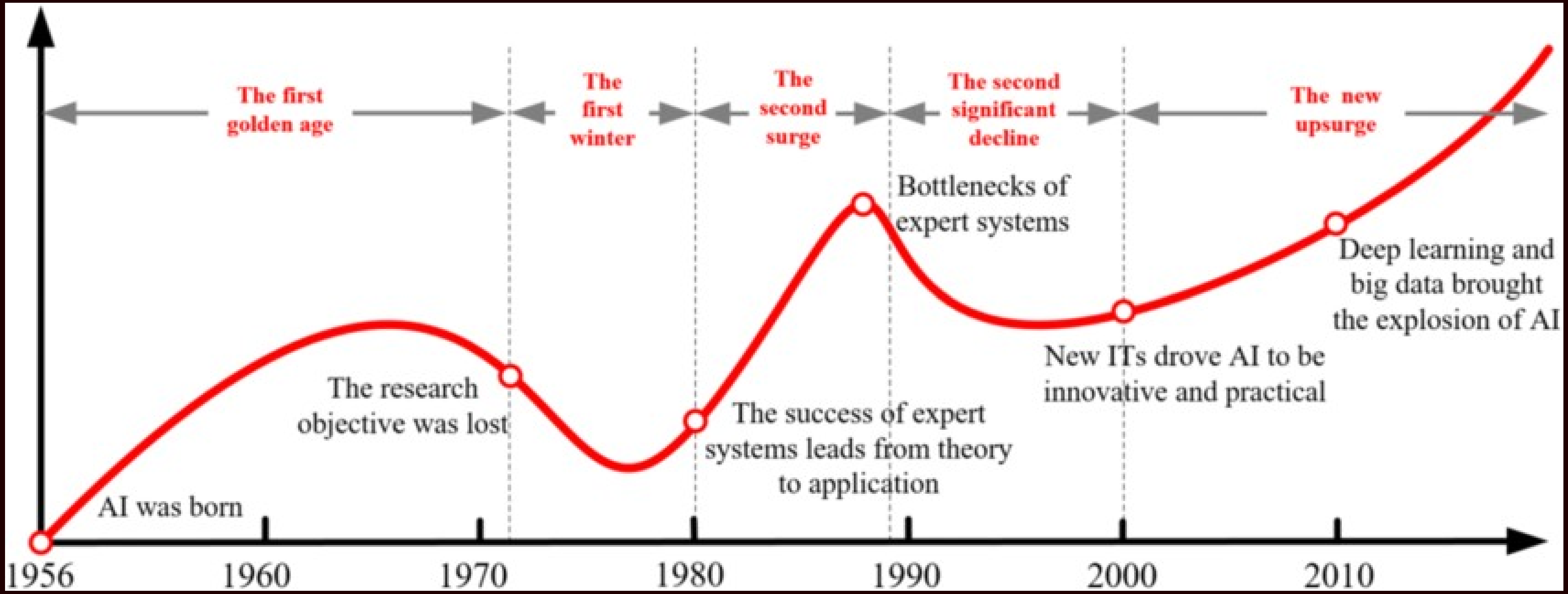
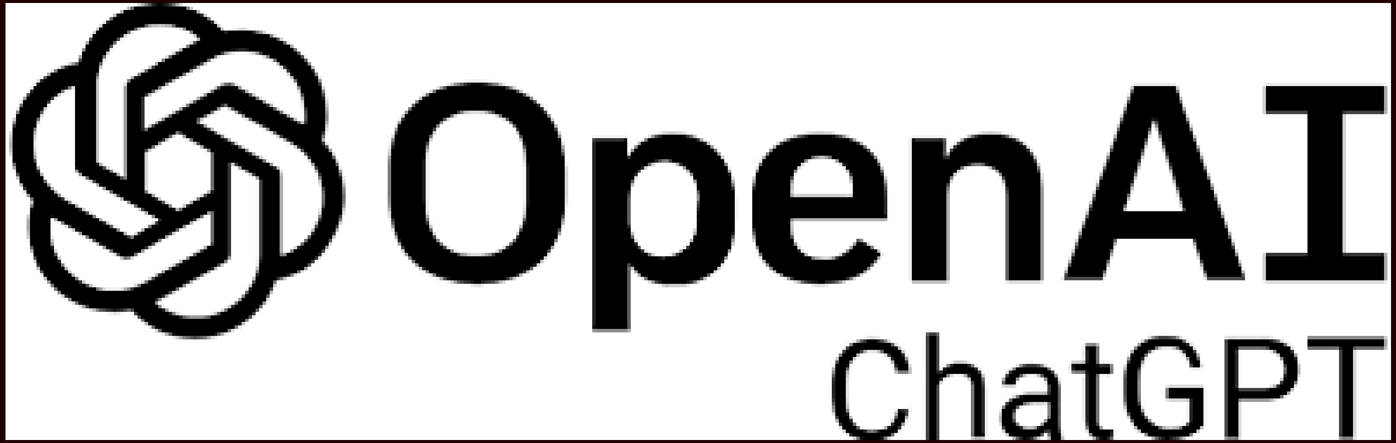
Effortless AI Scaling: Red Hat OpenShift and LLM Integration Production-grade AI

Indy Van Mol
Piros

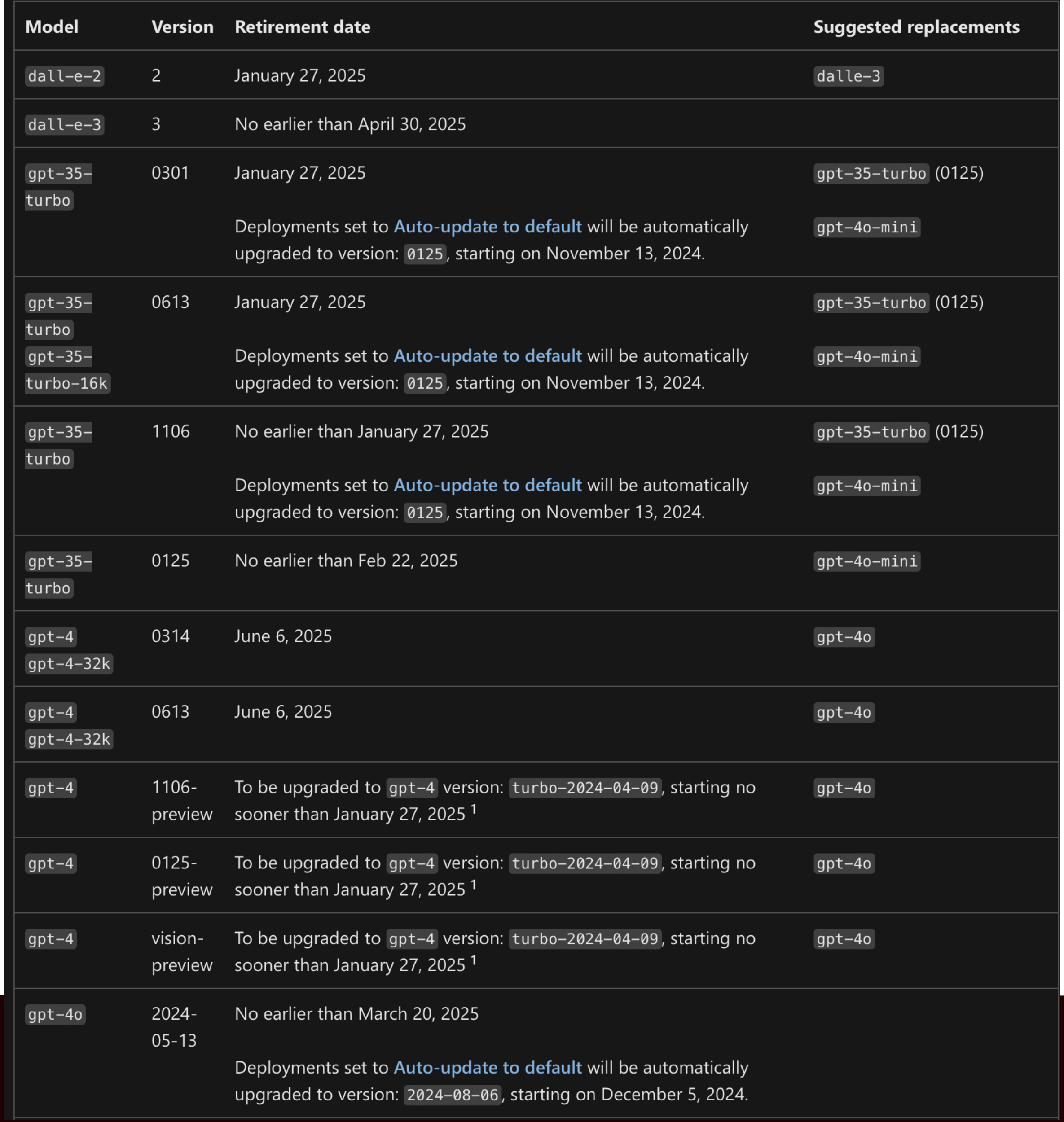
The explosion of AI and its integration into all business applications



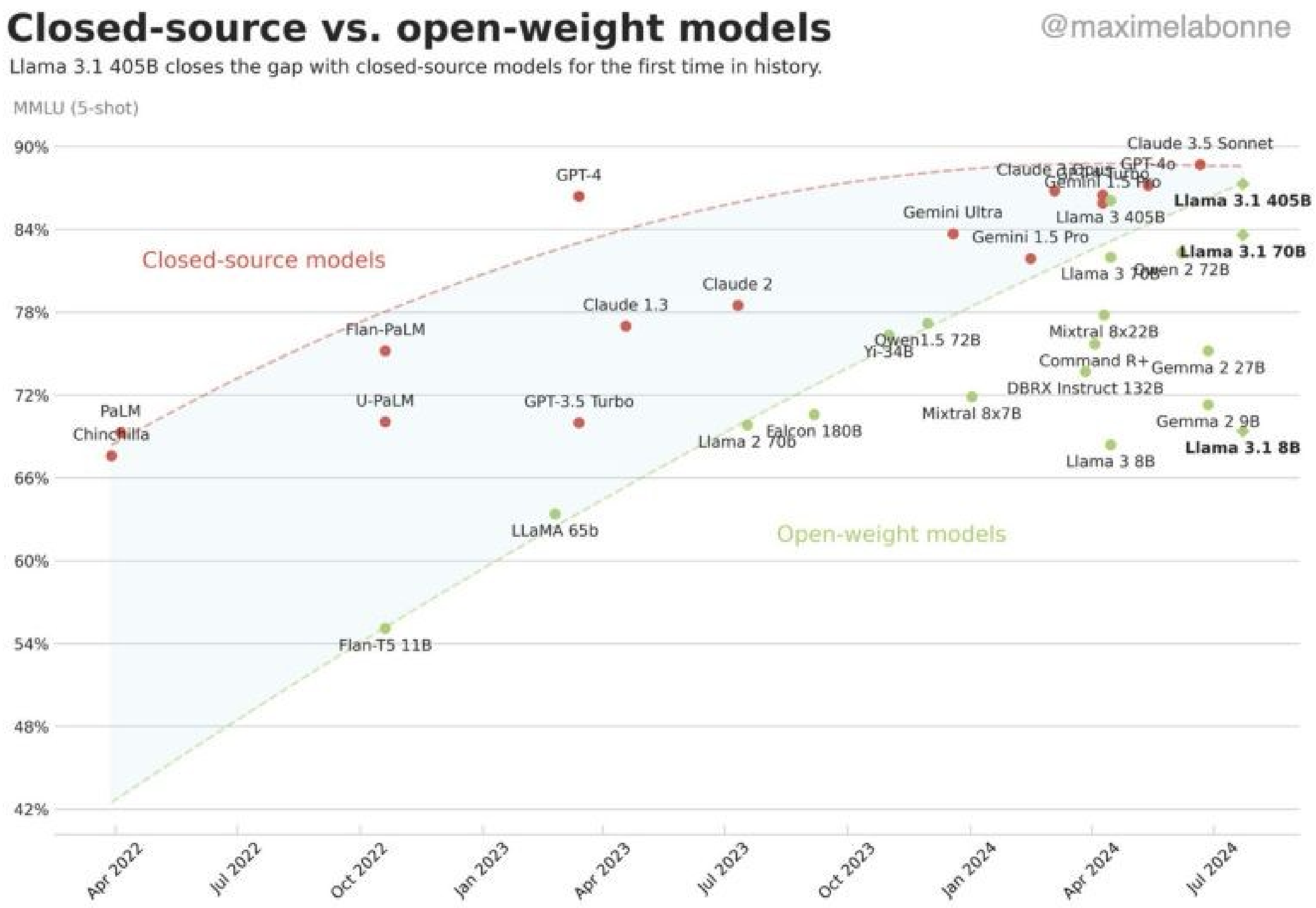
The LLM take over



LLMs



Model	Version	Retirement date	Suggested replacements
`dall-e-2`	2	January 27, 2025	`dalle-3`
`dall-e-3`	3	No earlier than April 30, 2025	
`gpt-35-turbo`	0301	January 27, 2025 Deployments set to Auto-update to default will be automatically upgraded to version: `0125`, starting on November 13, 2024.	`gpt-35-turbo` (0125) `gpt-4o-mini`
`gpt-35-turbo` `gpt-35-turbo-16k`	0613	January 27, 2025 Deployments set to Auto-update to default will be automatically upgraded to version: `0125`, starting on November 13, 2024.	`gpt-35-turbo` (0125) `gpt-4o-mini`
`gpt-35-turbo`	1106	No earlier than January 27, 2025 Deployments set to Auto-update to default will be automatically upgraded to version: `0125`, starting on November 13, 2024.	`gpt-35-turbo` (0125) `gpt-4o-mini`
`gpt-35-turbo`	0125	No earlier than Feb 22, 2025	`gpt-4o-mini`
`gpt-4` `gpt-4-32k`	0314	June 6, 2025	`gpt-4o`
`gpt-4` `gpt-4-32k`	0613	June 6, 2025	`gpt-4o`
`gpt-4`	1106-preview	To be upgraded to `gpt-4` version: `turbo-2024-04-09`, starting no sooner than January 27, 2025 [1]	`gpt-4o`
`gpt-4`	0125-preview	To be upgraded to `gpt-4` version: `turbo-2024-04-09`, starting no sooner than January 27, 2025 [1]	`gpt-4o`
`gpt-4`	vision-preview	To be upgraded to `gpt-4` version: `turbo-2024-04-09`, starting no sooner than January 27, 2025 [1]	`gpt-4o`
`gpt-4o`	2024-05-13	No earlier than March 20, 2025 Deployments set to Auto-update to default will be automatically upgraded to version: `2024-08-06`, starting on December 5, 2024.	



Closed-source vs. open-weight models

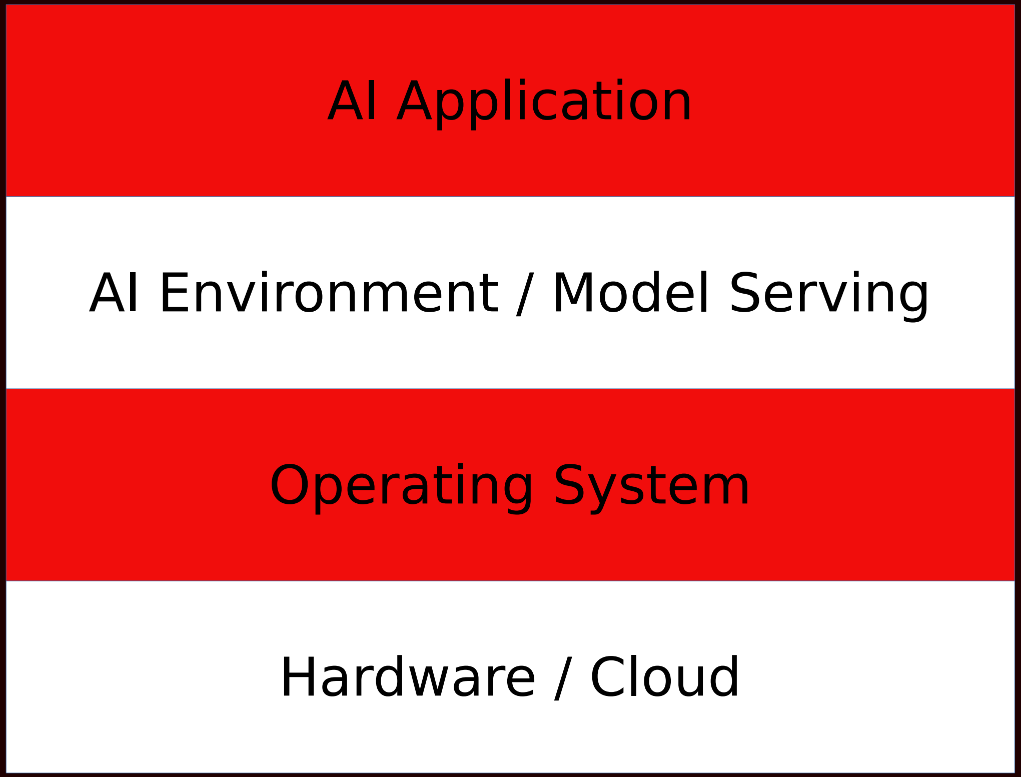
Llama 3.1 405B closes the gap with closed-source models for the first time in history.

HUGGING FACE

The Future of AI is Open Source

Red Hat

How to build you AI envoirment



AI Application

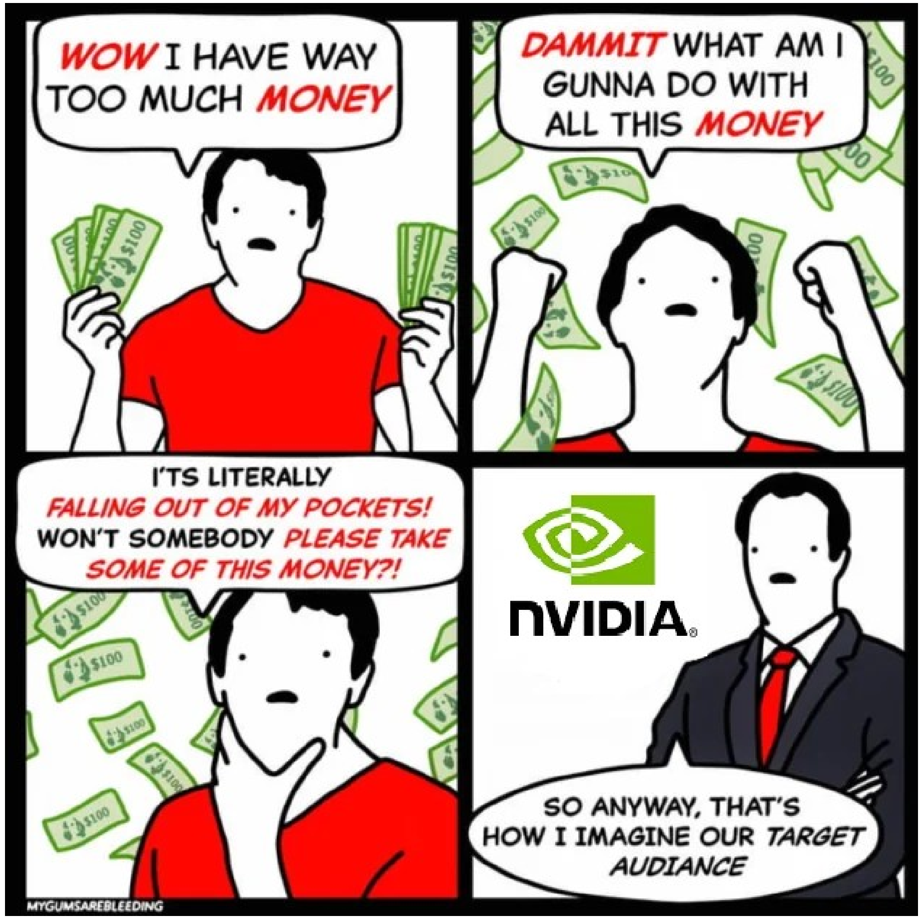
AI Environment / Model Serving

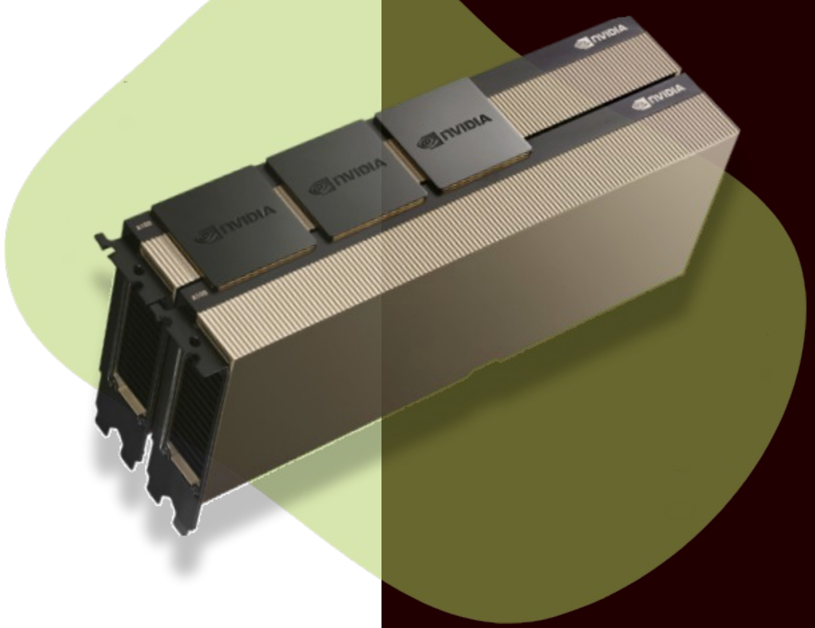
Operating System

Hardware / Cloud

AI Accelerators

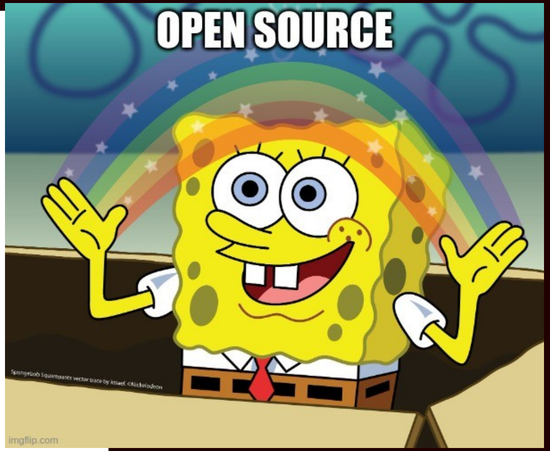
- A100
- H100
- H200
- Blackwell
- L40s


WOW I HAVE WAY TOO MUCH MONEY
DAMMIT WHAT AM I GUNNA DO WITH ALL THIS MONEY
I'TS LITERALLY FALLING OUT OF MY POCKETS! WON'T SOMEBODY PLEASE TAKE SOME OF THIS MONEY?!
NVIDIA
SO ANYWAY, THAT'S HOW I IMAGINE OUR TARGET AUDIANCE
MYGUMSAREBLEEDING



Nvidia Nvidia Nvidia

WHEN EVERYONE DIGS FOR GOLD
SHOVELS $99
NVIDIA
Microsoft
Meta
SELL
SHOVELS

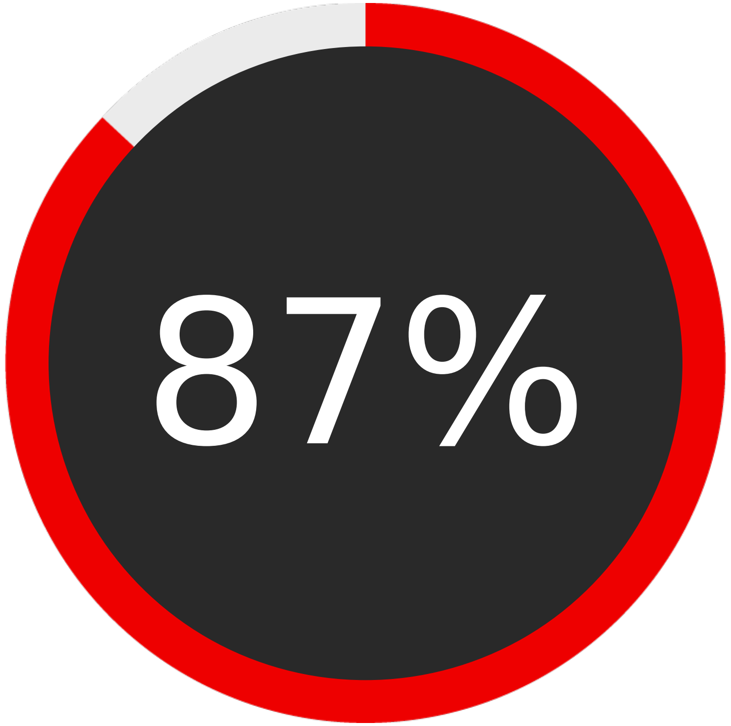


What are the challenges that are impeding your progress?

Bridging environments
How do you bridge the data scientist and app developer environments?

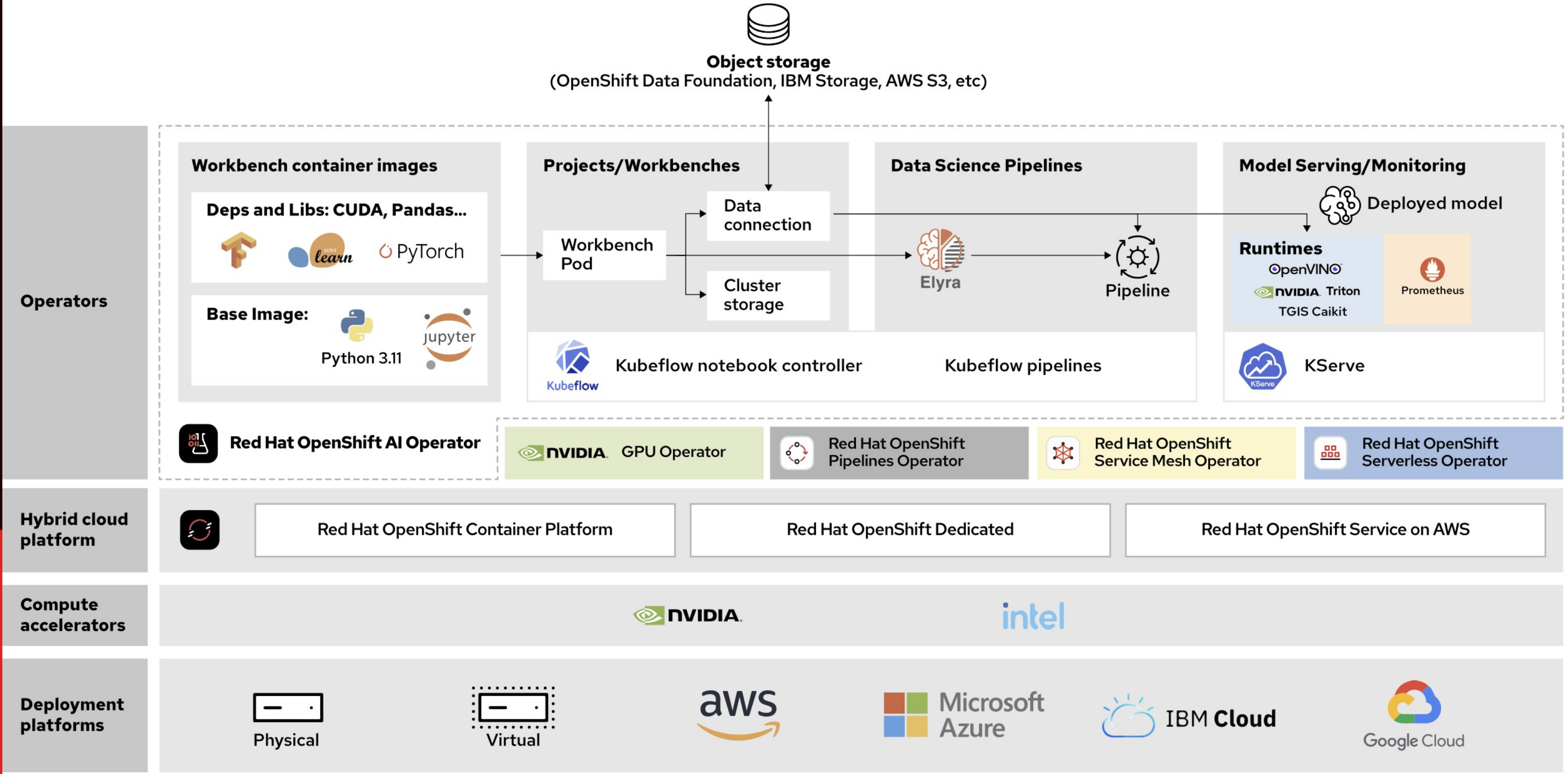
Infrastructure complexity
How do you make the complexity of infrastructure invisible?

Realizing value
How do you safely realize the business benefits of AI innovation?



87% of AI projects never make it into production

Red Hat OpenShift Container Platform



Object storage
(OpenShift Data Foundation, IBM Storage, AWS S3, etc)

Operators

Workbench container images

Deps and Libs: CUDA, Pandas...

learn

PyTorch

Base Image:

Python 3.11

jupyter

Projects/Workbenches

Workbench Pod

Data connection

Cluster storage

Data Science Pipelines

Elyra

Pipeline

Model Serving/Monitoring

Deployed model

Runtimes

OpenVINO

NVIDIA Triton

TGIS Caikit

Prometheus

Kubeflow

Kubeflow notebook controller

Kubeflow pipelines

KServe

Red Hat OpenShift AI Operator

NVIDIA GPU Operator

Red Hat OpenShift Pipelines Operator

Red Hat OpenShift Service Mesh Operator

Red Hat OpenShift Serverless Operator

Hybrid cloud platform

Red Hat OpenShift Container Platform

Red Hat OpenShift Dedicated

Red Hat OpenShift Service on AWS

Compute accelerators

NVIDIA

intel

Deployment platforms

Physical

Virtual

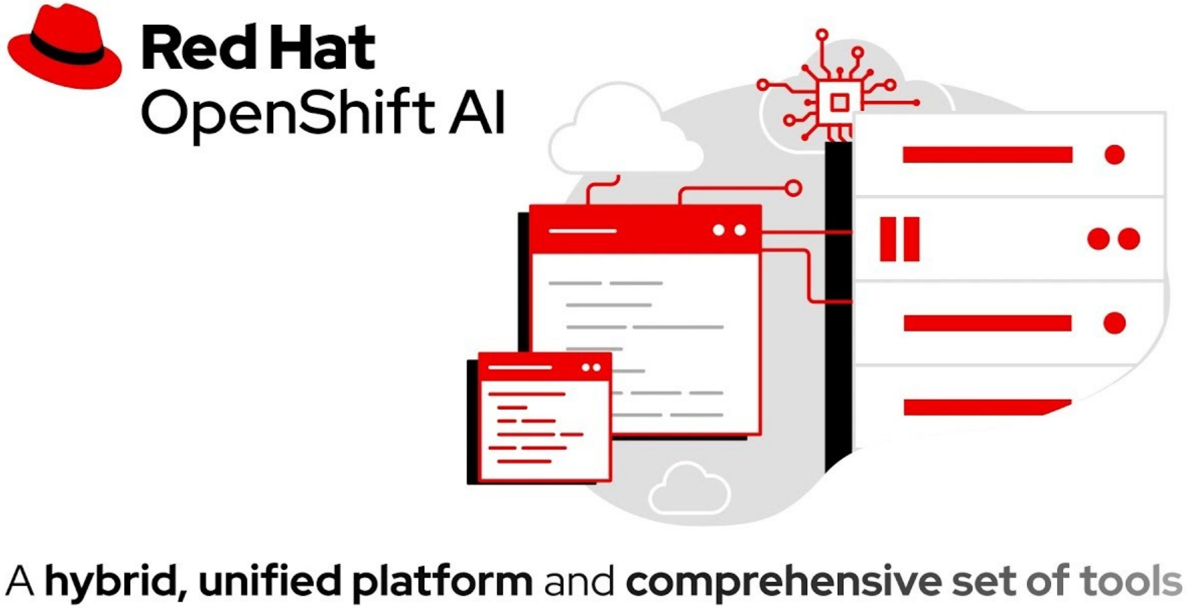
aws

Microsoft Azure

IBM Cloud

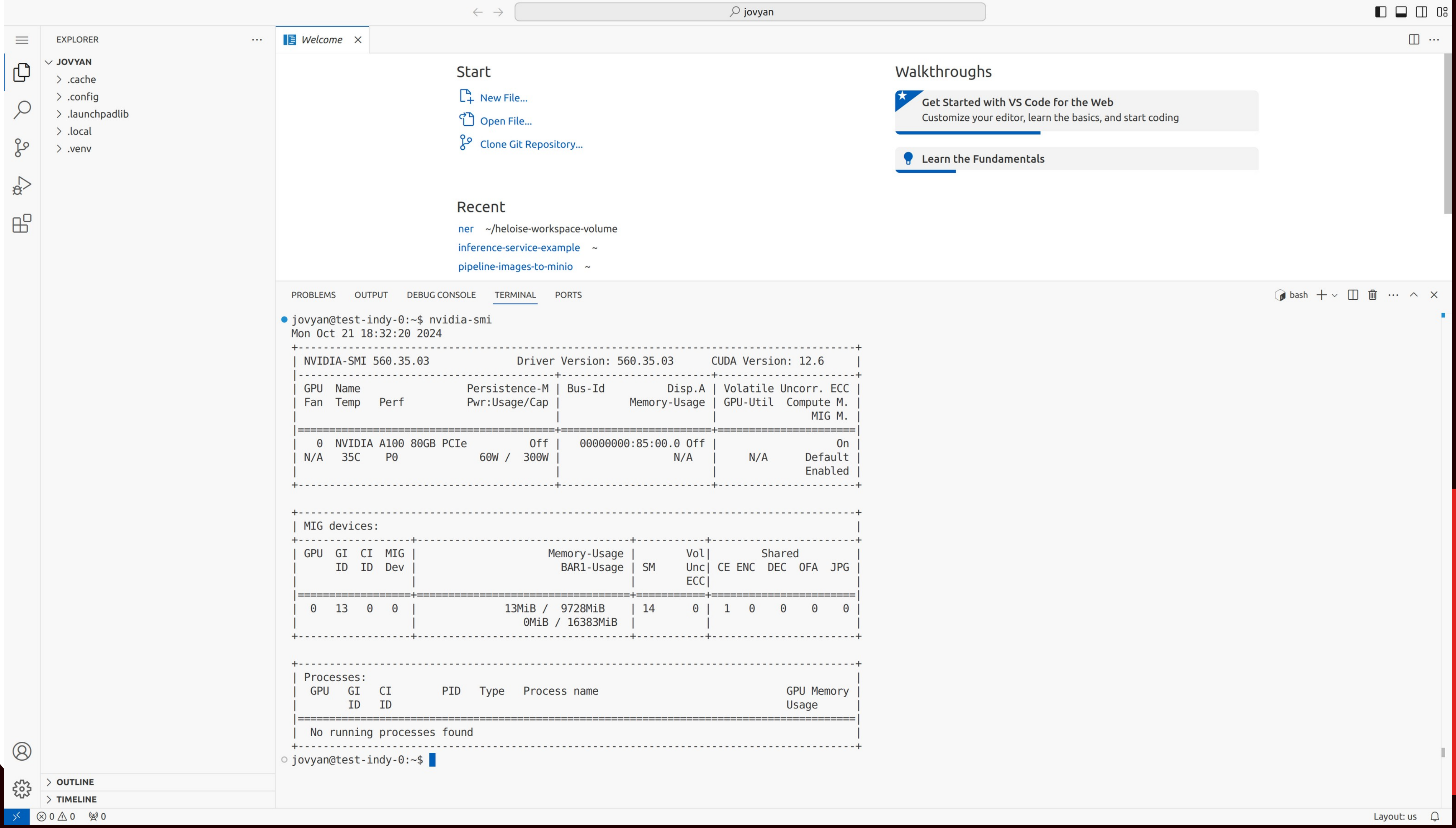
Google Cloud

The Development Platform

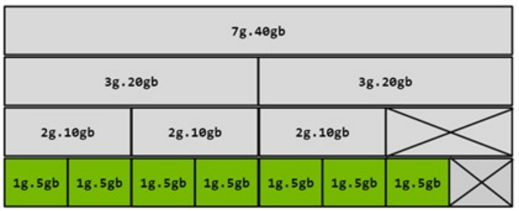


Kubeflow

VSCode

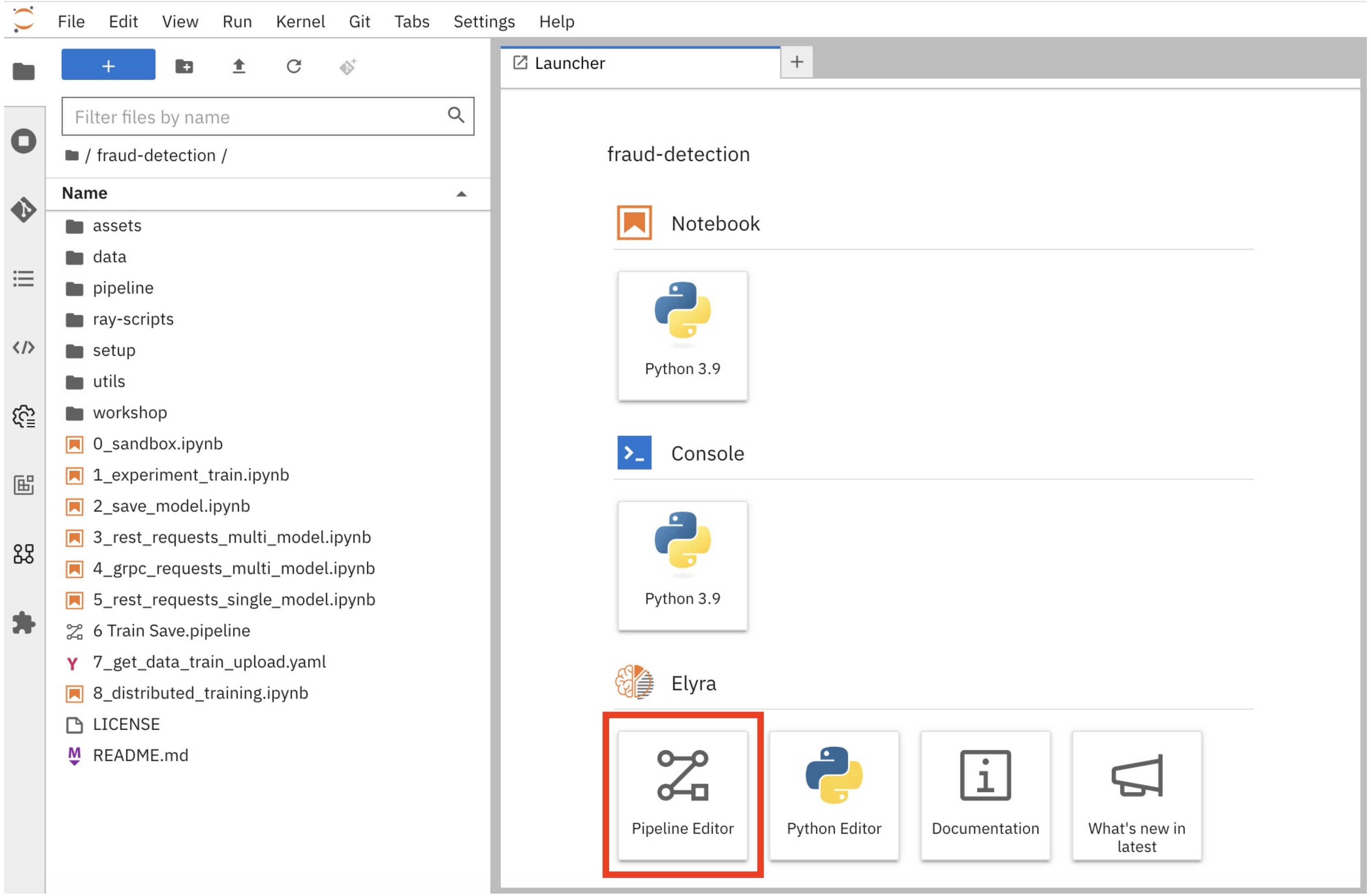


```
jovyan@test-indy-0:~$ nvidia-smi
Mon Oct 21 18:32:20 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 560.35.03              Driver Version: 560.35.03      CUDA Version: 12.6     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA A100 80GB PCIe          Off |   00000000:85:00.0 Off |                   On |
| N/A   35C    P0             60W /  300W |                  N/A   |     N/A      Default |
|                                         |                        |              Enabled |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| MIG devices:                                                                            |
+------------------+----------------------------------+-----------+-----------------------+
| GPU  GI  CI  MIG |                     Memory-Usage |        Vol|        Shared         |
|      ID  ID  Dev |                       BAR1-Usage | SM     Unc| CE ENC  DEC  OFA  JPG |
|                  |                                  |        ECC|                       |
|==================+==================================+===========+=======================|
|  0   13   0   0  |              13MiB /  9728MiB    | 14      0 |  1   0    0    0    0 |
|                  |                 0MiB / 16383MiB  |           |                       |
+------------------+----------------------------------+-----------+-----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  No running processes found                                                             |
+-----------------------------------------------------------------------------------------+
jovyan@test-indy-0:~$
```



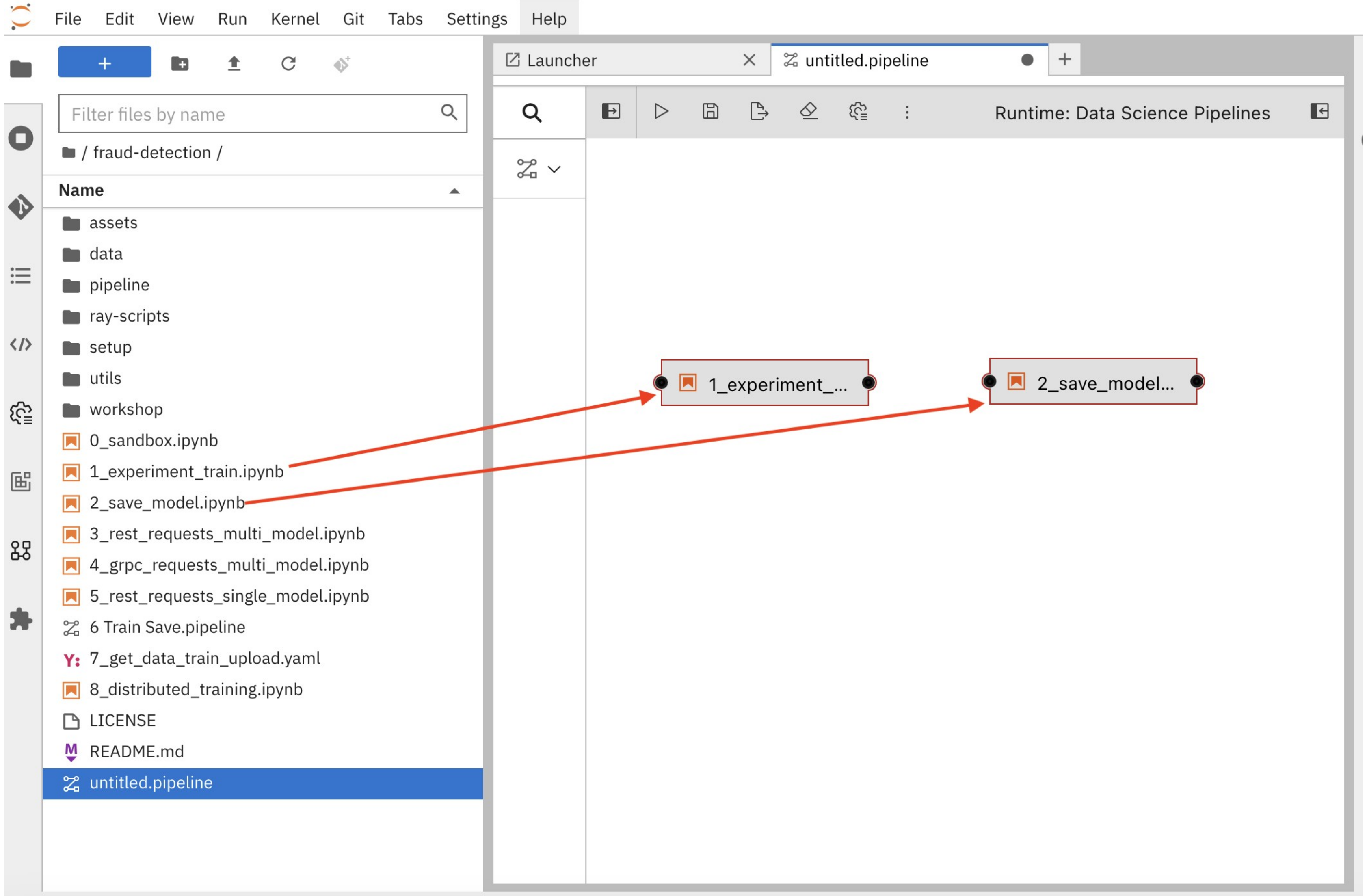
- 1 x 7g.40gb
  or
- 2 x 3g.20gb
  or
- 3 x 2g.10gb
  or
- 7 x 1g.5gb

Pipelines



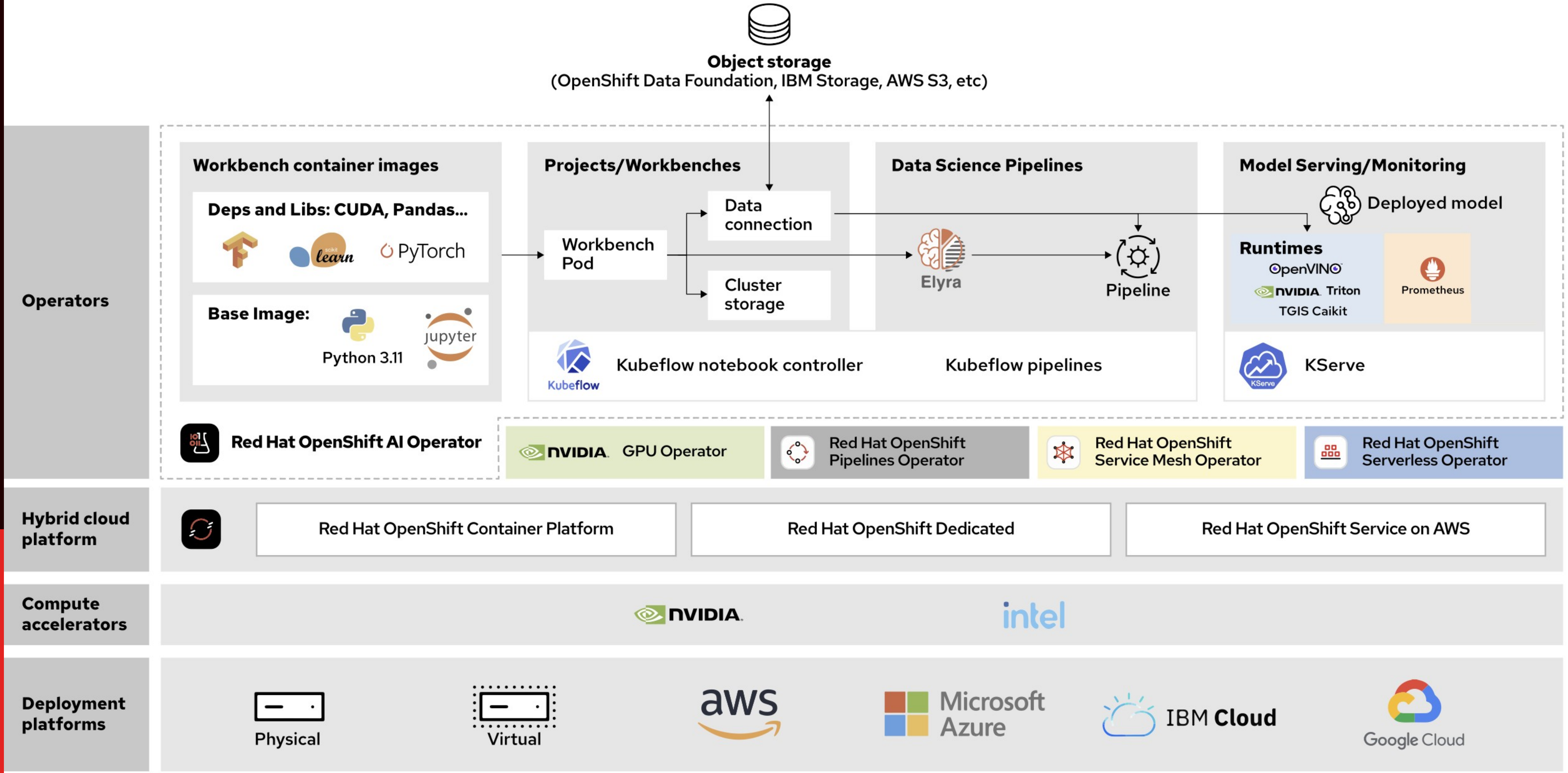
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Filter files by name
/ fraud-detection /
Name
assets
data
pipeline
ray-scripts
setup
utils
workshop
0_sandbox.ipynb
1_experiment_train.ipynb
2_save_model.ipynb
3_rest_requests_multi_model.ipynb
4_grpc_requests_multi_model.ipynb
5_rest_requests_single_model.ipynb
6 Train Save.pipeline
7_get_data_train_upload.yaml
8_distributed_training.ipynb
LICENSE
README.md
Launcher
fraud-detection
Notebook
Python 3.9
Console
Python 3.9
Elyra
Pipeline Editor
Python Editor
Documentation
What's new in latest

Pipelines



- assets
- data
- pipeline
- ray-scripts
- setup
- utils
- workshop
- 0_sandbox.ipynb
- 1_experiment_train.ipynb
- 2_save_model.ipynb
- 3_rest_requests_multi_model.ipynb
- 4_grpc_requests_multi_model.ipynb
- 5_rest_requests_single_model.ipynb
- 6 Train Save.pipeline
- 7_get_data_train_upload.yaml
- 8_distributed_training.ipynb
- LICENSE
- README.md
- untitled.pipeline

Production Interference



Object storage
(OpenShift Data Foundation, IBM Storage, AWS S3, etc)

Operators

Workbench container images

Deps and Libs: CUDA, Pandas...

PyTorch

Base Image:

Python 3.11

jupyter

Projects/Workbenches

Workbench Pod

Data connection

Cluster storage

Data Science Pipelines

Elyra

Pipeline

Model Serving/Monitoring

Deployed model

Runtimes

OpenVINO

NVIDIA Triton

TGIS Caikit

Prometheus

Kubeflow

Kubeflow notebook controller

Kubeflow pipelines

KServe

Red Hat OpenShift AI Operator

NVIDIA GPU Operator

Red Hat OpenShift Pipelines Operator

Red Hat OpenShift Service Mesh Operator

Red Hat OpenShift Serverless Operator

Hybrid cloud platform

Red Hat OpenShift Container Platform

Red Hat OpenShift Dedicated

Red Hat OpenShift Service on AWS

Compute accelerators

NVIDIA

intel

Deployment platforms

Physical

Virtual

aws

Microsoft Azure

IBM Cloud

Google Cloud

Wrap Up

- Buy Some Hardware with Nvidia GPUs
- OpenShift for Scaleability
- Develop your LLM solution on top of the OpenShift AI platform
- Move quickly from development to Production with OpenShift AI endpoint
- Congratulations you have integrated Open Source AI that adds value to your Business

We Are Piros

Let’s talk about your challenges

Get in touch

CALL US

+32 (0)475 757 998

FIND US

Gaston Geenslaan 11, 3001 Leuven

Piros.be

GROENEVELD
KU Leuven
HEVERLEE
Jezuietenpark
Piros.
Sint-Jansbergsesteenweg
Gaston Geenslaan
Veldstraat
Jos de Haeslaan
Remiseboswg
Steengroevenlaan
Celestijnenlaan
Kapeldreef
Dyle
Kardinaal Mercierlaan
Jules Vandenbemptlaan
Hertogstraat
Goswin III Laan
Priorijlaan
Ter Elstlaan
Tervuursevest
Groenstraat
Schapenstraat