INL/CON-23-72474-Revision-0

Serving Machine Learning Models in a Production Environment

April 2023

Brandon Samuel Biggs Jr

Changing the World's Energy Future

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/CON-23-72474-Revision-0

Serving Machine Learning Models in a Production Environment

Brandon Samuel Biggs Jr

April 2023

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

April 28, 2023

Brandon Biggs
INL/CON-23-72474

Serving Machine Learning Models in a Production Environment

Battelle Energy Alliance manages INL for the
U.S. Department of Energy's Office of Nuclear Energy

Idaho National Laboratory

Roadmap

- Motivation
- Challenges
- Goals
- Machine Learning Operations (MLOps)
- Model Repositories
- Hosting
  - Online vs Offline (Batch)
- Considerations
- Future Considerations

Motivation

- Inference can be powerful (and fun!)
- Save time
- Collaborate
- Generate new ideas

Hosted inference API

Text-to-Image

A dog trying to fix a computer

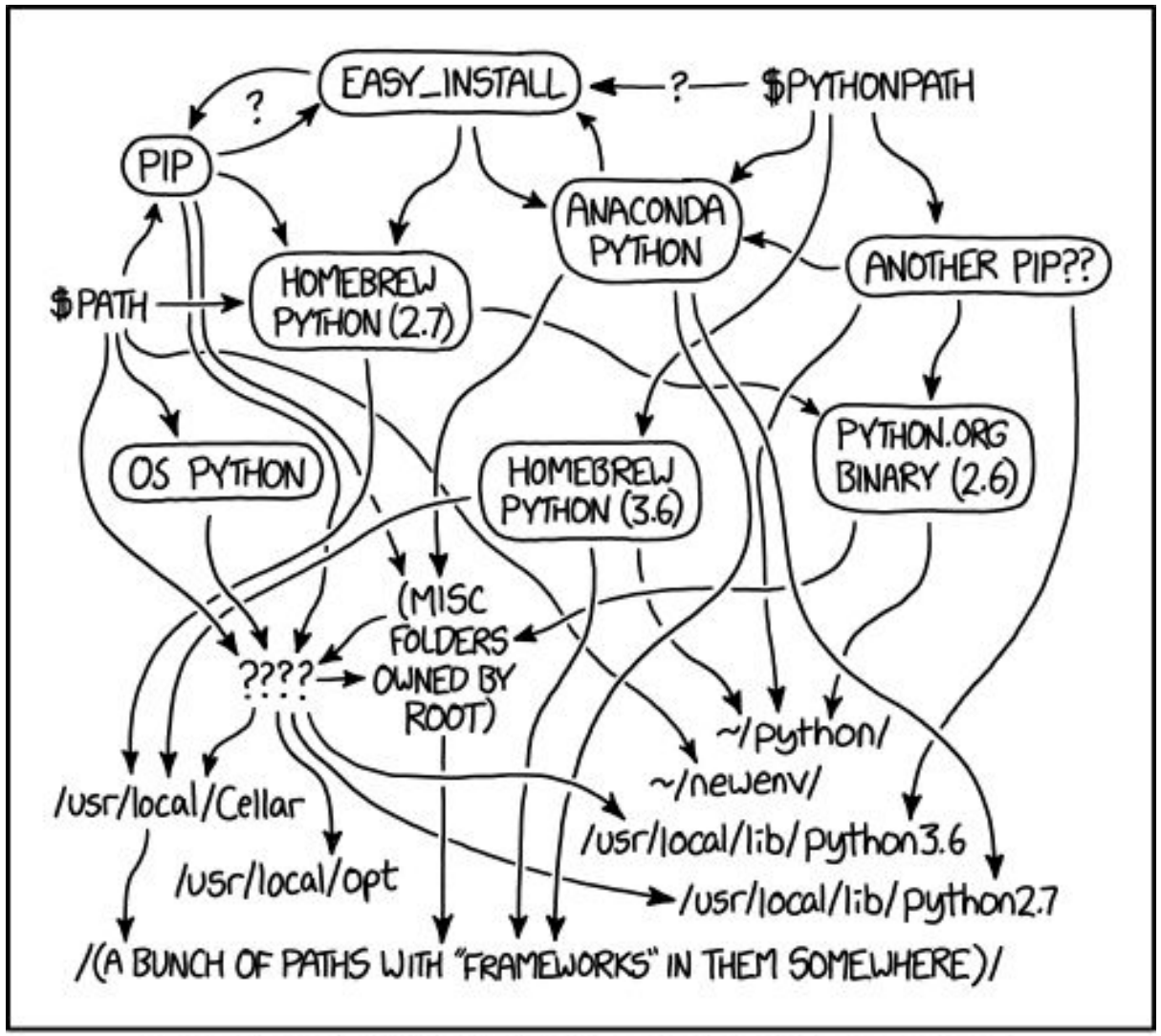
Compute

Computation time on gpu: 11.832 s

https://huggingface.co/stabilityai/stable-diffusion-2

Challenges

- Machine learning is rapidly advancing.
- Managing dependencies is hard.
  - Pip, pipx, conda, poetry, pipenv for Python
  - Containers?
  - Accelerator packages?
- Between 8 and 90 days for companies to deploy a single model [1]
- 2,473 organizations surveyed and found that a significant portion of their attempted AI deployments fail [1]
- Can we make machine learning useful?

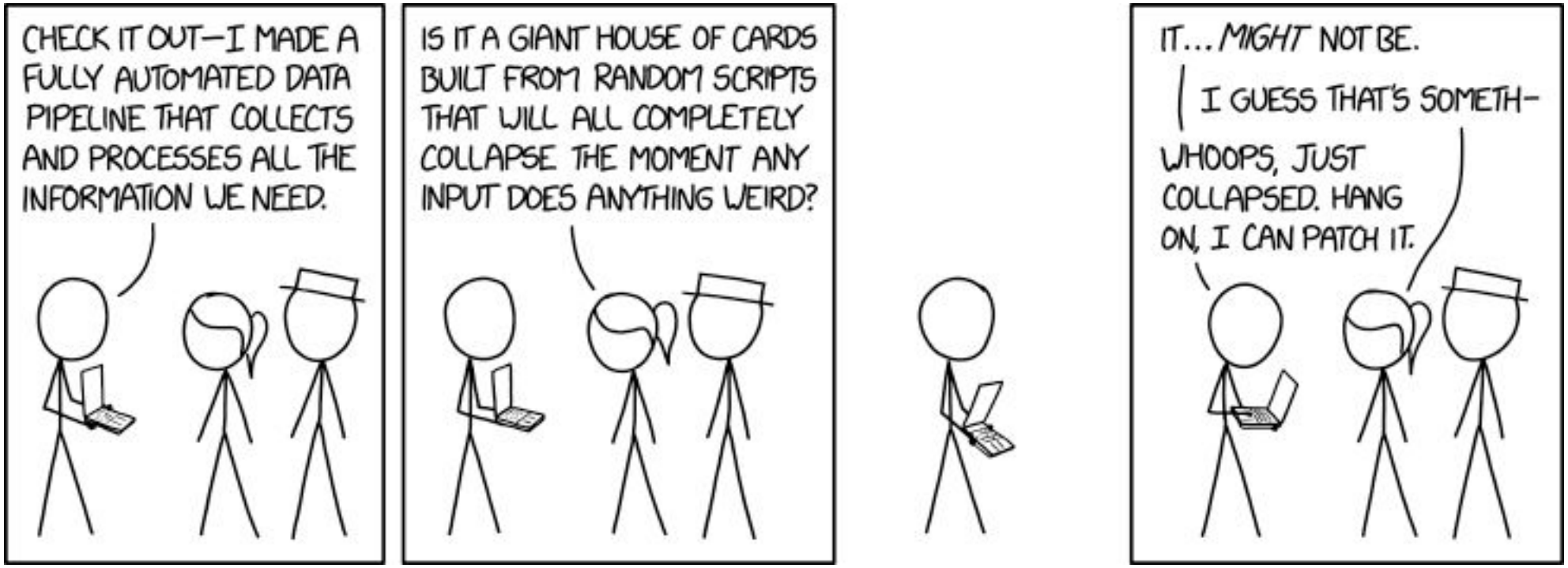


MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

[1] https://arxiv.org/pdf/2011.09926.pdf
[2] https://xkcd.com/1987/

Goals

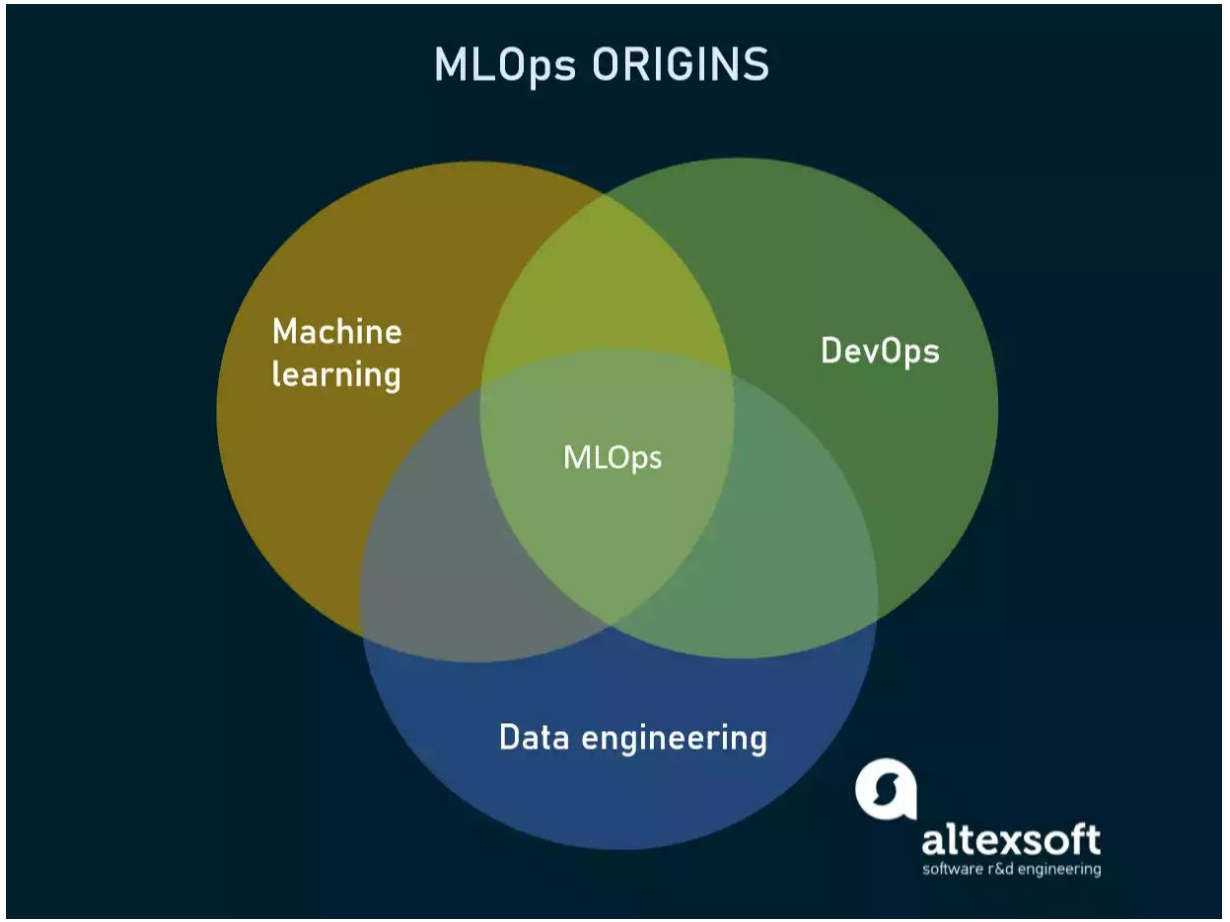
- Make machine learning more accessible.
- Provide a way for people to utilize the efforts of data scientists, researchers, etc.



https://xkcd.com/2054/

Machine Learning Operations - MLOps

- Like DevOps but for machine learning.
- Has a lot of the same practices but adapted for ML specific challenges.
  - Data
  - Model Training
  - Model Serving



https://www.altexsoft.com/blog/mlops-methods-tools

Hosting

- Online and/or Offline
- Cloud vs On-Prem

Hosting Tools - Cloud

- AWS SageMaker
- Google Vertex AI
- Azure Machine Learning Endpoints
- Many more

Hosting Tools - Local

- NVIDIA Triton
- BentoML
- TensorFlow Serving
- Torch Serve
- Mlflow
- Kubeflow
- Building your own with an API framework!

```
from transformers import ViTImageProcessor, ViTForImageClassification
from PIL import Image
import requests
from fastapi import FastAPI

processor = ViTImageProcessor.from_pretrained('google/vit-base-patch16-224')
model = ViTForImageClassification.from_pretrained('google/vit-base-patch16-224')

app = FastAPI()

@app.post("/classify_via_url/")
async def classify_via_url(image_url):
    image = Image.open(requests.get(image_url).raw)
    inputs = processor(images=image, return_tensors="pt")
    outputs = model(**inputs).logits
    prediction_id = logits.argmax(-1).item()
    label = model.config.id2label[prediction_id]
    return label
```

Image Classification Projects description..

GET	`/classify/{image-url:path}`	Image By Url

Text-to-Speech this is a description

GET	`/text-to-speech/{text}`	Get Audio From Text

Image Generation Projects description..

GET	`/stable-diffusion/text-to-image/{text}`	Image From Text

Model Repositories

- Hugging Face (cloud)
- BentoML (on prem)
- Weights and Biases (Paid)
- DagsHub
- MLRun
- More!

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document 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

Audio

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

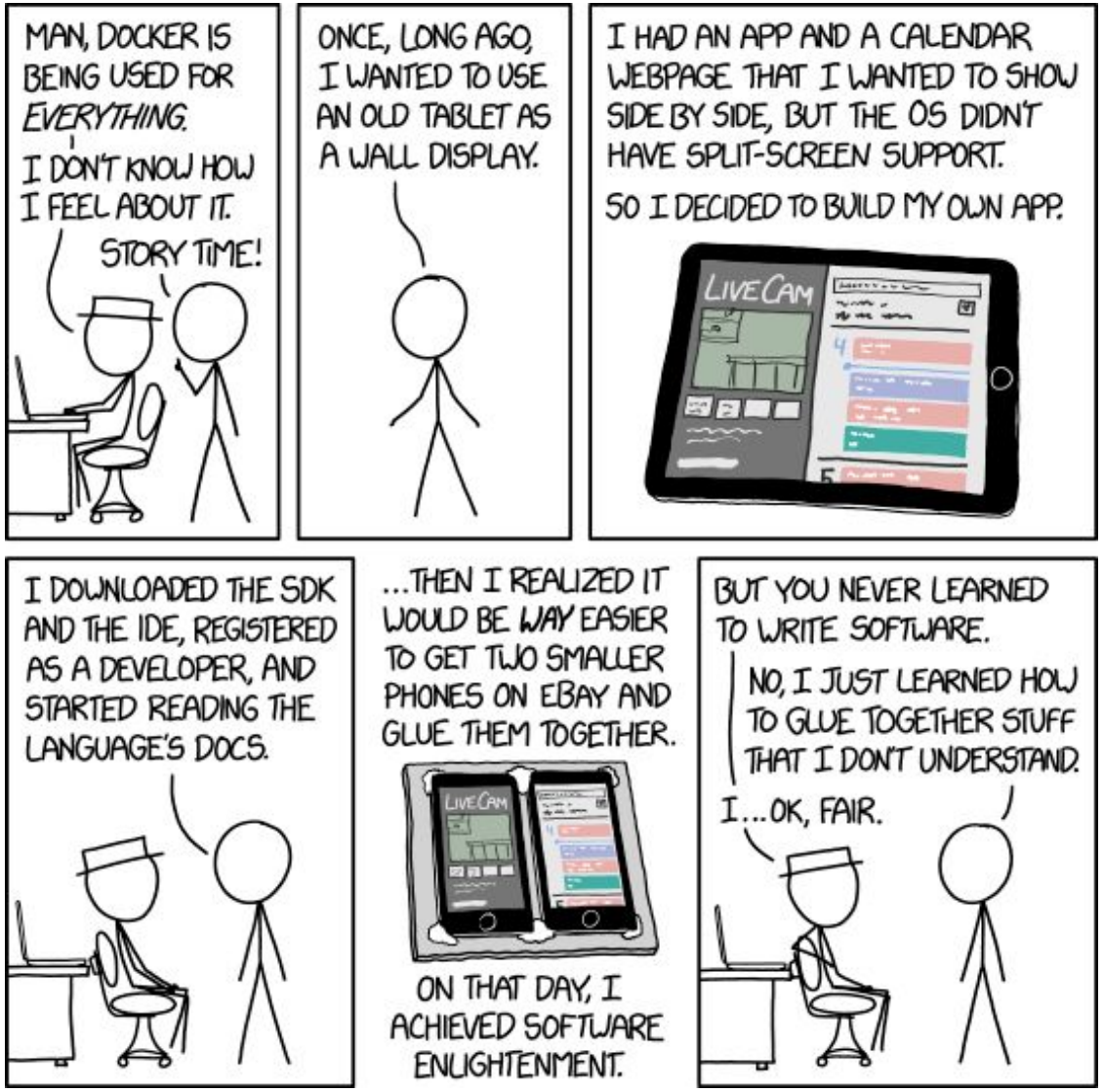
https://huggingface.co/models

Considerations

- Containers
- Hardware
- Security
- Monitoring

Considerations - Containers

- Very useful for dependencies
- Lots of already created containers with dependencies
  - NVIDIA (many)
  - AMD Infinity Hub (few)
- Don’t have to use Docker, Singularity/Apptainer have been well tested.



https://xkcd.com/1988/

Considerations - Hardware

- A lot of work in the NLP space going into optimizing large language models for CPUs/non-datacenter GPUs
- Some models just require a lot of GPU memory
- Tools exist for testing on GPUs before deploying models somewhere
  - Google Colab

Considerations – Security and Monitoring

- You're building an API for people to access. Watch out for abuse and unauthorized access.
- Keep an eye on models being used to generate toxic content

Future Considerations

- Data Versioning
- Experiment Tracking
- Model Registry
- API Scaling

Questions?

- Brandon.Biggs@inl.gov

Idaho National Laboratory

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.