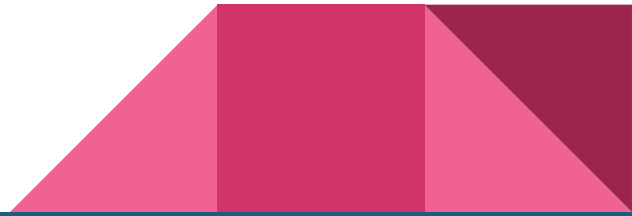
Technical Aspects in ML

Marcin Walkowski
Gradient Science Club 2022

Plan for Today

- Technical Aspects in ML
  - Which ML framework to choose?
  - Model deployment - NVIDIA Triton Inference Server
  - Useful ML tools
- Housekeeping
  - FOKA & projects
  - Upcoming meetings
  - Gradient board elections
  - Budget plan
  - AI Bay seminar



Resources

- Coursera – MLOps Specialization
- NVIDIA Triton Inference Server developer page
- NVIDIA TensorRT developer page

Which ML framework to choose?

Technical Aspects in ML

Which ML framework to choose?

First things first - what is our task?

- ANNs and deep learning
- Datascience
- Data analysis
- Reinforcement learning
- Graph NNs
- NLP
- ...

Which ML framework to choose?

First things first - what is our task?

- ANNs and deep learning
- **Datascience**
- Data analysis
- Reinforcement learning
- Graph NNs
- NLP
- ...

Which ML framework to choose?

First things first - what is our task?

- ANNs and deep learning
- Datascience
- **Data analysis**
- Reinforcement learning
- Graph NNs
- NLP
- ...

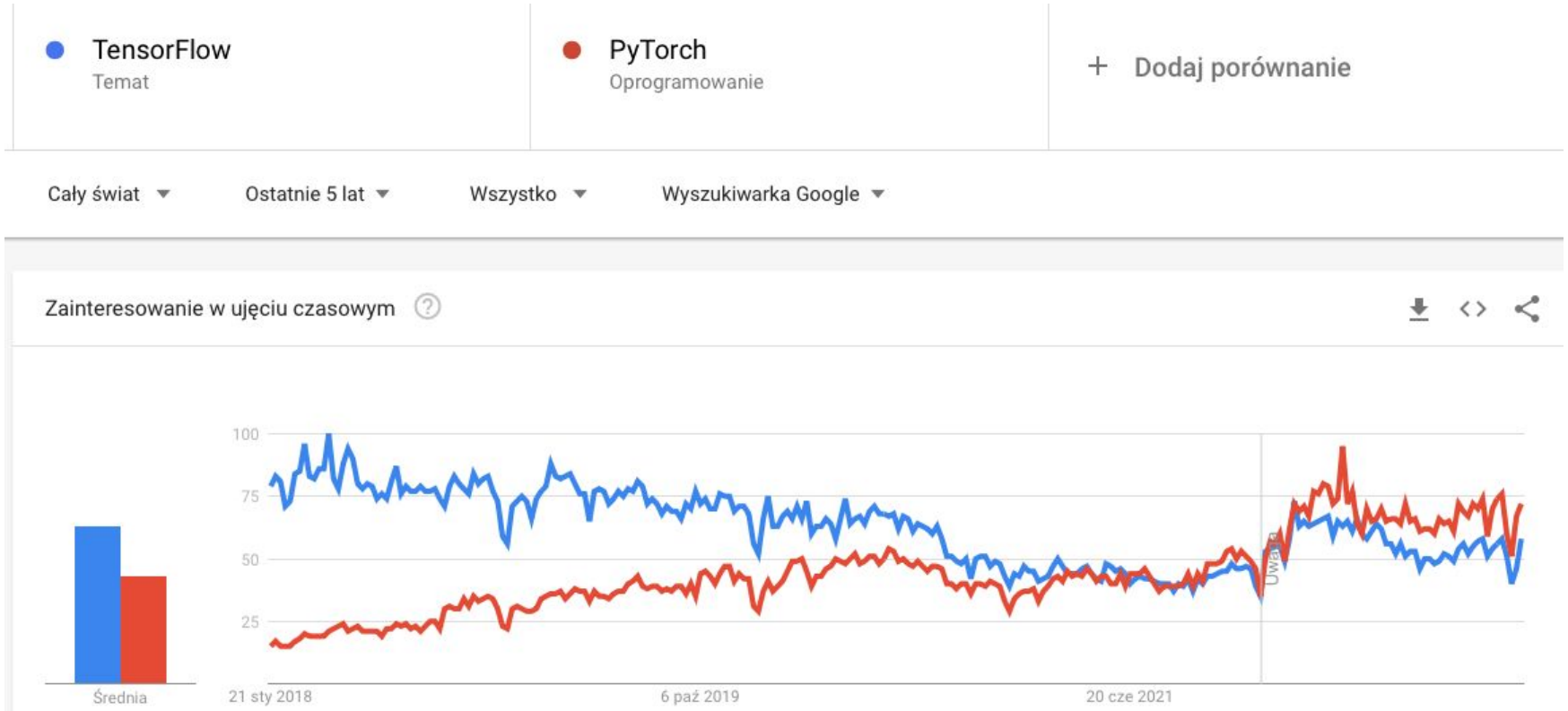
Which ML framework to choose?

First things first - what is our task?

- **ANNs and deep learning** - we need auto differentiation
- Datascience
- Data analysis
- Reinforcement learning
- Graph NNs
- NLP
- ...

ANNs and deep learning

VS



Source: https://trends.google.pl/trends/

PyTorch vs TensorFlow - similarities

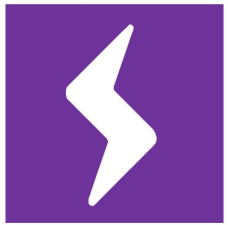
- Open-source frameworks used for deep learning
- Built-in support for automatic differentiation
- Wide range of APIs for creating, training and deploying models
- Strong GPU support

PyTorch vs TensorFlow - differences

- PyTorch is said to be more pythonic
- TensorFlow is said to be more notebook friendly
- PyTorch has easier data parallelism
- Tensorflow has Tensorboard - data visualization library ...
- ... which is also supported by PyTorch
- TensorFlow has Keras ...

… PyTorch Lightning

- Quick prototyping
- Easy access to advanced training strategies
- Dataset code encapsulated
  in Lightning Data Modules
- Built-in support for many loggers



PyTorch
Lightning

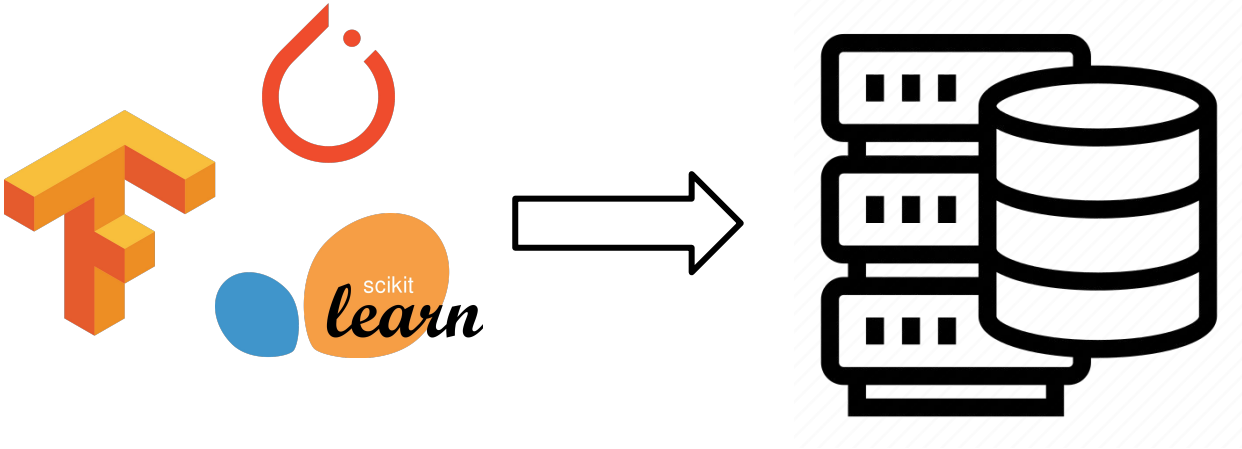
JAX and Flax

- Google open-source project
- Automatic differentiation
  of native Python and NumPy functions
- Accelerated linear algebra -
  running NumPy programs on GPUs and TPUs
- Flax - ANN library for JAX

Model deployment – NVIDIA Triton Inference Server

Technical Aspects in ML

Persistence



Persistence

```
# TF2 code
# Save the weights
model.save_weights('./model_weights/my_model_weights')

# Create a new model instance
model = create_model()

# Restore the weights and evaluate
model.load_weights('./model_weights/my_model_weights')
loss, acc = model.evaluate(test_images, test_labels, verbose=2)
```

```
# TF2 code
# Save the entire model as a SavedModel.
model.save('models/my_model')

# Load the entire model and evaluate
new_model = tf.keras.models.load_model('saved_model/my_model')
loss, acc = model.evaluate(test_images, test_labels, verbose=2)
```

Persistence

```
# TF2 code
# Create a callback that saves the model's weights
cp_callback = tf.keras.callbacks.ModelCheckpoint(filepath='path/to/checkpoints',
                                                 save_weights_only=True,
                                                 verbose=1)

# Train the model with the new callback
model.fit(train_images,
          train_labels,
          epochs=10,
          validation_data=(test_images, test_labels),
          callbacks=[cp_callback])
```

Deployment – easy solution

```
# Python-like code
from flask import Flask
import tensorflow as tf

app = Flask(__name__)

@app.route('/predict', methods=['POST'])
def predict():
    data = request.json['data']

    model = tf.keras.models.load_model('saved_models/my_model')

    prediction = model.predict(data)

    return jsonify({'prediction': list(prediction)})


if __name__ == '__main__':
    app.run(port=8080)
```

Deployment - dedicated inference server

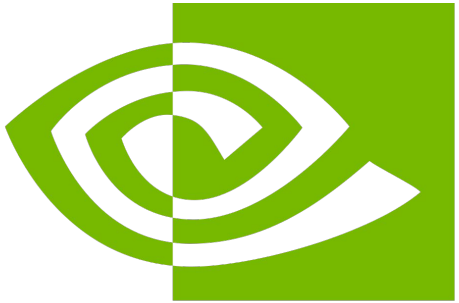
- TorchServe
- TensorFlow Server
- NVIDIA Triton Inference Server
- BentoML
- Coretex
- KFServing
- ...

Deployment - dedicated inference server

- TorchServe
- TensorFlow Server
- **NVIDIA Triton Inference Server**
- BentoML
- Coretex
- KFServing
- ...

NVIDIA Triton Inference Server

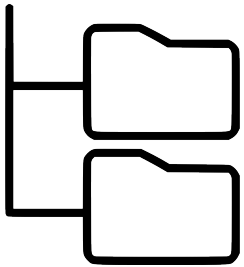
- Open-source project
- Cloud or local deployment
- GPU and CPU inference
- Built-in optimization
- Client libs, web and local APIs



NVIDIA®

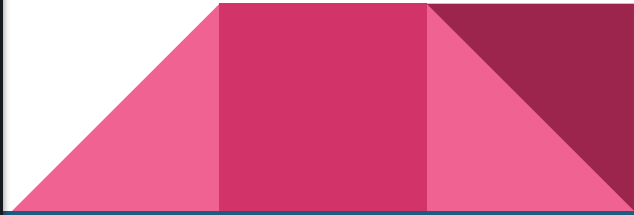
NVIDIA Triton Inference Server - structure

- Source code / container
- **Model repository**



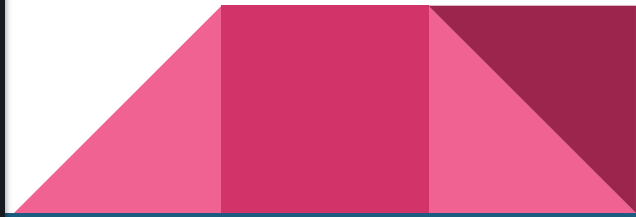
NVIDIA Triton Inference Server – model repository

```
<model-repository-path>/
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  ...
```



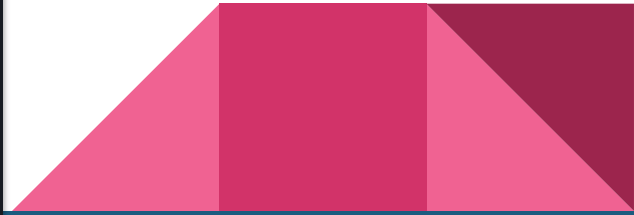
NVIDIA Triton Inference Server – model repository

```
<model-repository-path>/
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  ...
```



NVIDIA Triton Inference Server – model repository

```
<model-repository-path>/
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  <model-name>/
    [config.pbtxt]
    [<output-labels-file> ...]
    <version>/
      <model-definition-file>
    <version>/
      <model-definition-file>
    ...
  ...
```



NVIDIA Triton Inference Server – config file

```
platform: "tensorrt_plan"
max_batch_size: 8
input [
  {
    name: "input0"
    data_type: TYPE_FP32
    dims: [ 16 ]
  },
  {
    name: "input1"
    data_type: TYPE_FP32
    dims: [ 16 ]
  }
]
output [
  {
    name: "output0"
    data_type: TYPE_FP32
    dims: [ 16 ]
  }
]
```

NVIDIA Triton Inference Server – config file

```
instance_group [
  {
    count: 2
    kind: KIND_CPU
  }
]
```

```
instance_group [
  {
    count: 1
    kind: KIND_GPU
    gpus: [ 0 ]
  },
  {
    count: 2
    kind: KIND_GPU
    gpus: [ 1, 2 ]
  }
]
```

NVIDIA Triton Inference Server – config file

```
optimization { execution_accelerators {
  gpu_execution_accelerator : [ {
    name : "tensorrt"
    parameters { key: "precision_mode" value: "FP16" }
    parameters { key: "max_workspace_size_bytes" value: "1073741824" }
    }]
}}
```

```
dynamic_batching {
  preferred_batch_size: [ 4, 8 ]
  max_queue_delay_microseconds: 100
}
```

```
optimization { execution_accelerators {
  cpu_execution_accelerator : [ {
    name : "openvino"
  }]
}}
```

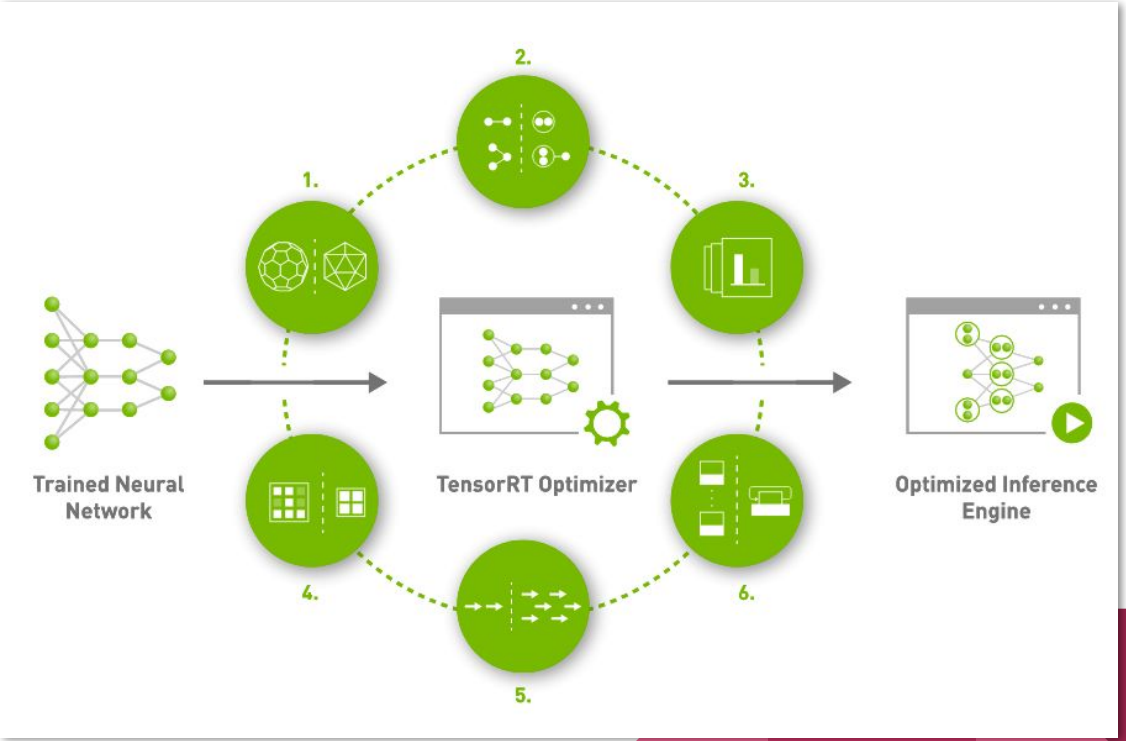
NVIDIA Triton Inference Server – supported model formats

Useful ML Tools

Technical Aspects in ML

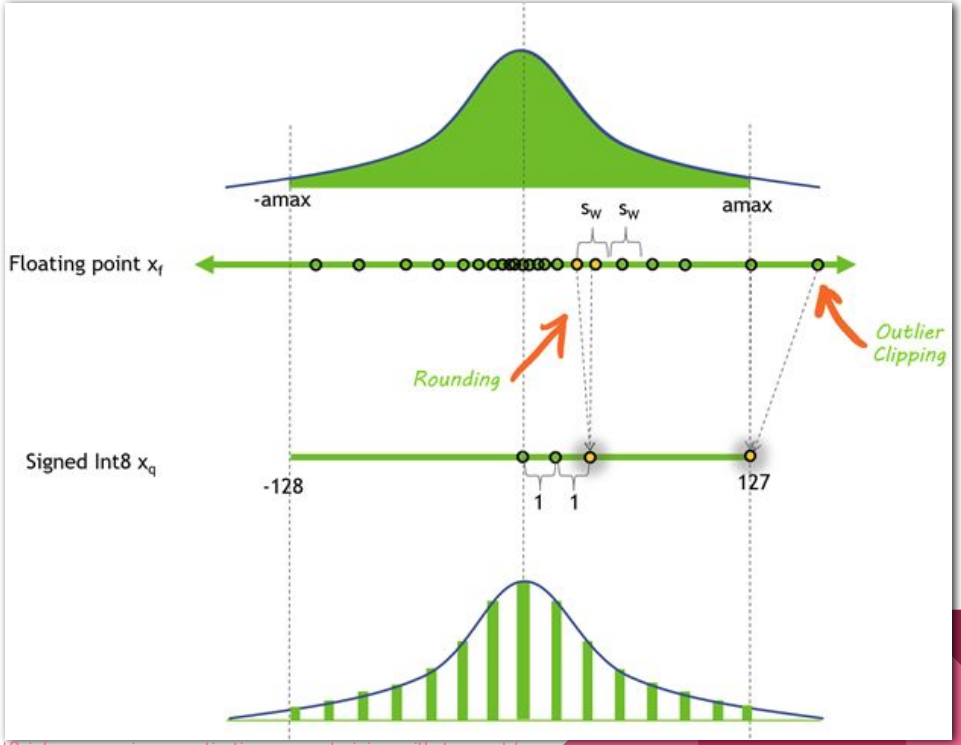
NVIDIA TensorRT

- Reduced Precision
- Layer and Tensor Fusion
- Kernel Auto-Tuning
- Dynamic Tensor Memory
- Multi-Stream Execution
- Time Fusion



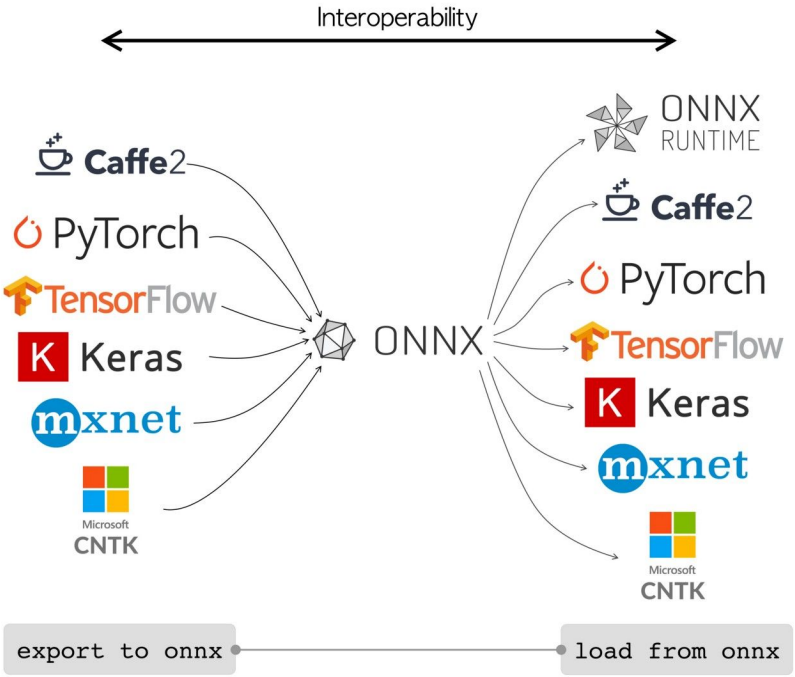
NVIDIA TensorRT

- **Reduced Precision**
- Layer and Tensor Fusion
- Kernel Auto-Tuning
- Dynamic Tensor Memory
- Multi-Stream Execution
- Time Fusion



Source: https://developer.nvidia.com/blog/achieving-fp32-accuracy-for-int8-inference-using-quantization-aware-training-with-tensorrt/

ONNX - Open Neural Network Exchange



Source: https://towardsdatascience.com/onnx-preventing-framework-lock-in-9a798fb34c92

Honorable mentions

- kaggle
- Weights & Biases
- Huffing Face
- seaborn
- Gradio
- Shout-out to #ciekawe-linki

Questions
&
Discussion

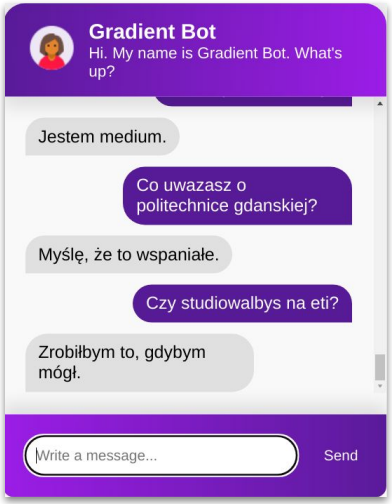
FOKA & projects

Housekeeping

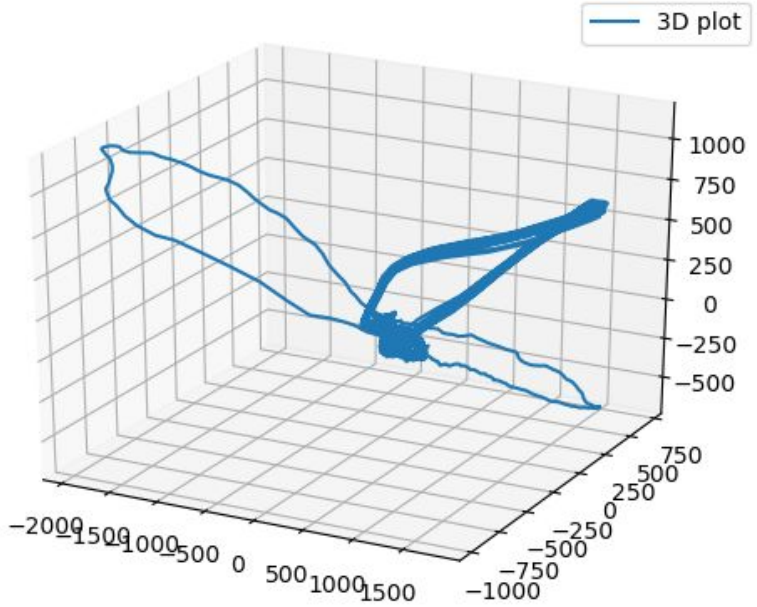
FOKA & projects

- FOKA - Forum Organizacji i Kół Akademickich
- Takes place on 21/03/2023 at PG
- Chance to show off our projects

FOKA & projects



GUMed VCG project



Plot author: Patryk Utkała

#vcg-gumed

#ideas-and-team-building

Upcoming meetings

Housekeeping

Upcoming meetings

- Guest lectures
- Gradient paper reading club
- Your idea 🤔

Gradient board elections

Housekeeping

Gradient board members

- Dawid Krefta
- Jakub Dembski
- Franciszek Górski
- Marcin Walkowski
- Bazyli Polednia

Gradient board members

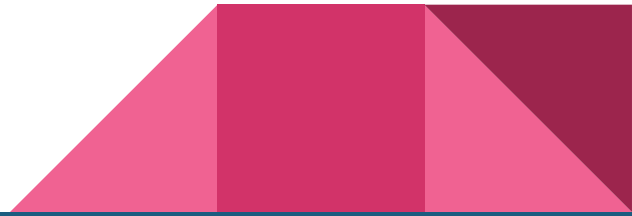
- Dawid Krefta ❌
- Jakub Dembski ❌
- Franciszek Górski ❓
- Marcin Walkowski 🔜❌
- Bazyli Polednia 🔜❌

Become Gradient board member

- Approach me (Marcin Walkowski) in person / on Discord
- Submit your application to gradientpg@gmail.com

  Application should answer the following two questions

  - Who you are?
  - What is your motivation?



Budget plan

Housekeeping

2023 budget ideas

- Gradient merch
- Conference trips
- GPUs

AI Bay seminar

Housekeeping

AI Bay seminar

- 26/01/2023 15:00 to 17:30 ⏰
- PG NE AUD 1 (Prawe) 📍
- Agenda 📣
  - IDEAS NCBR – Nowy ośrodek badawczo-rozwojowy w obszarze AI
    by Piotr Sankowski, IDEAS NCBR Warszawa
  - Uczenie w trybie ciągłym
    by Sebastian Cygert, WETI PG and IDEAS NCBR Warszawa

Thank you!
See you after winter break