

# Day5 - Project

02476 Machine Learning Operations

Nicki Skafte Detlefsen, Associate Professor, DTU Compute

January 2026

# The job

 You (and your group) are just hired as an MLOps engineers at a start-up.

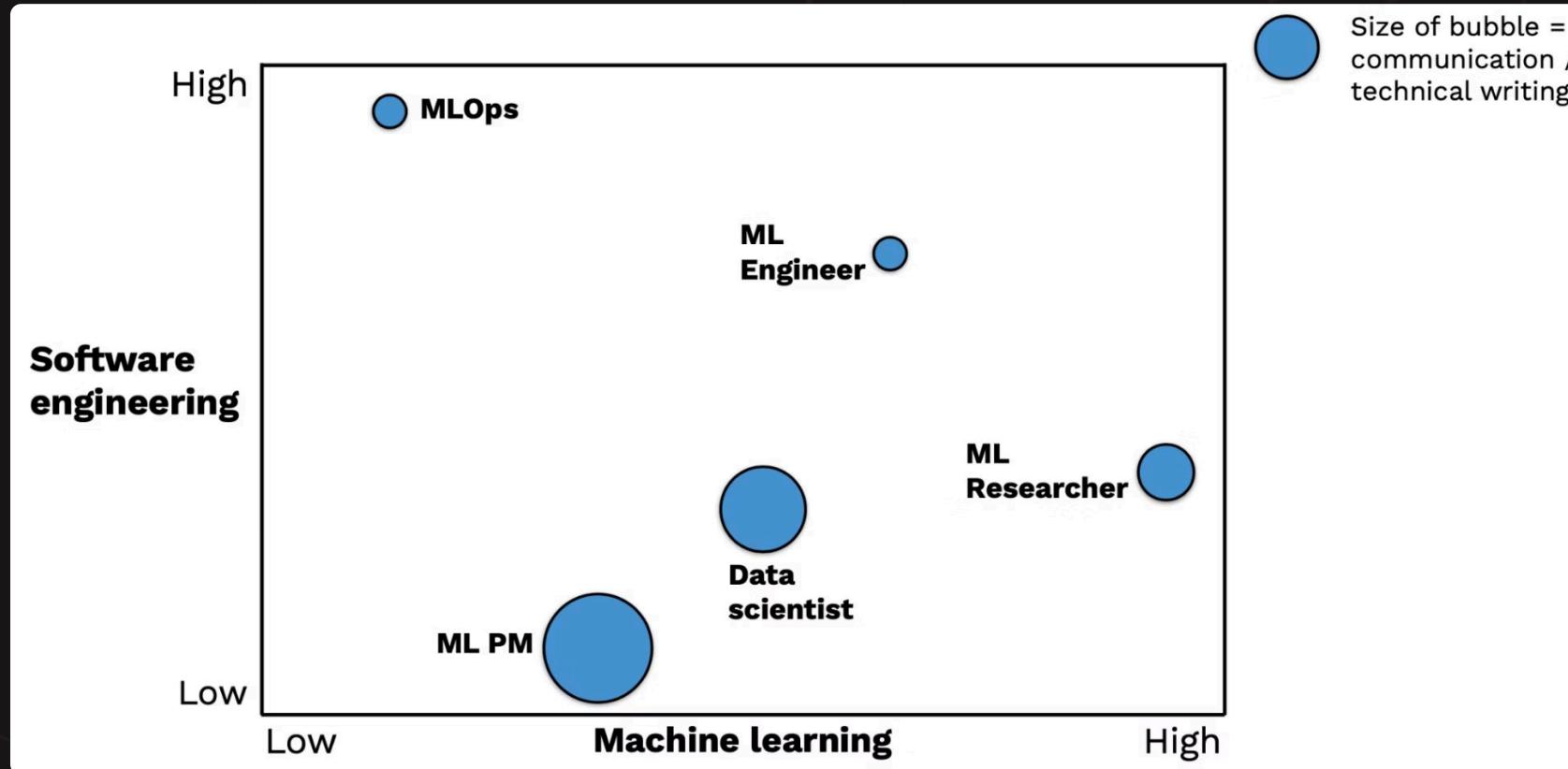
Your first job:

- *Develop an MLOps pipeline to solve a specific task for the company*

 Importantly: You are judged not by how great the model is but how fast you can setup a pipeline to solve the task.

# Why you do not need to care about the model?

That is a job for the ML research not MLOps engineer

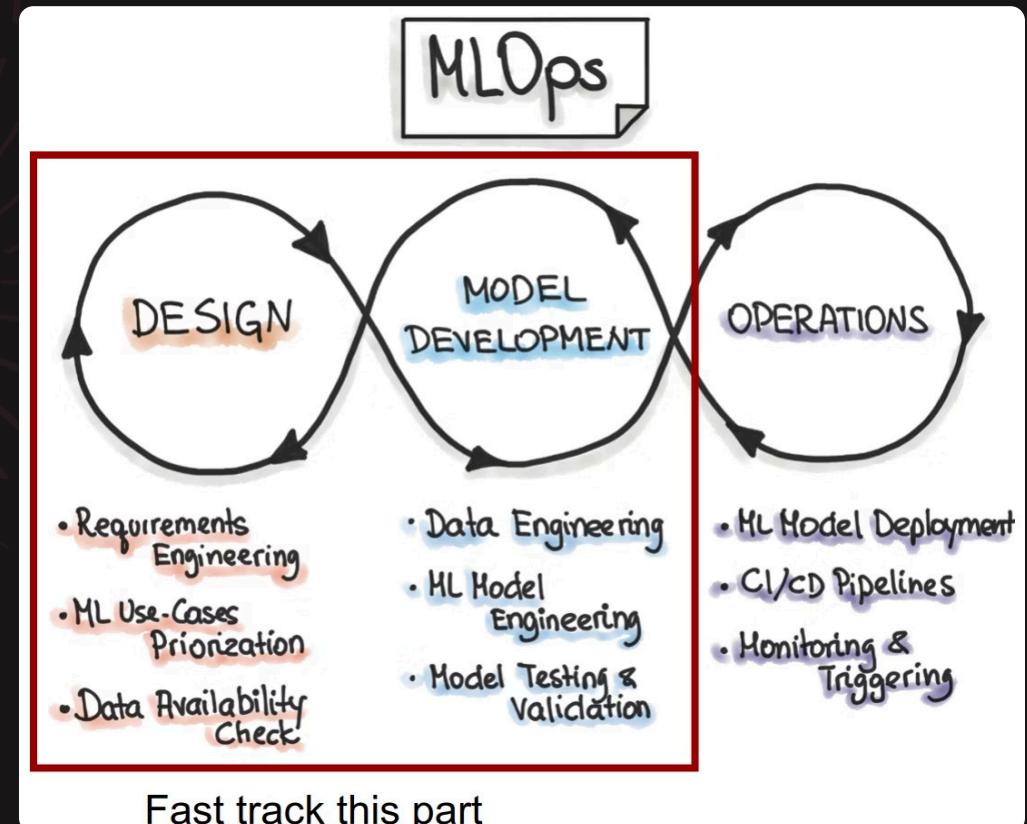


# How to solve the problem?

💡 You already have all the tools for the pipeline, you just need a good starting model.

💡 Your base framework is Pytorch

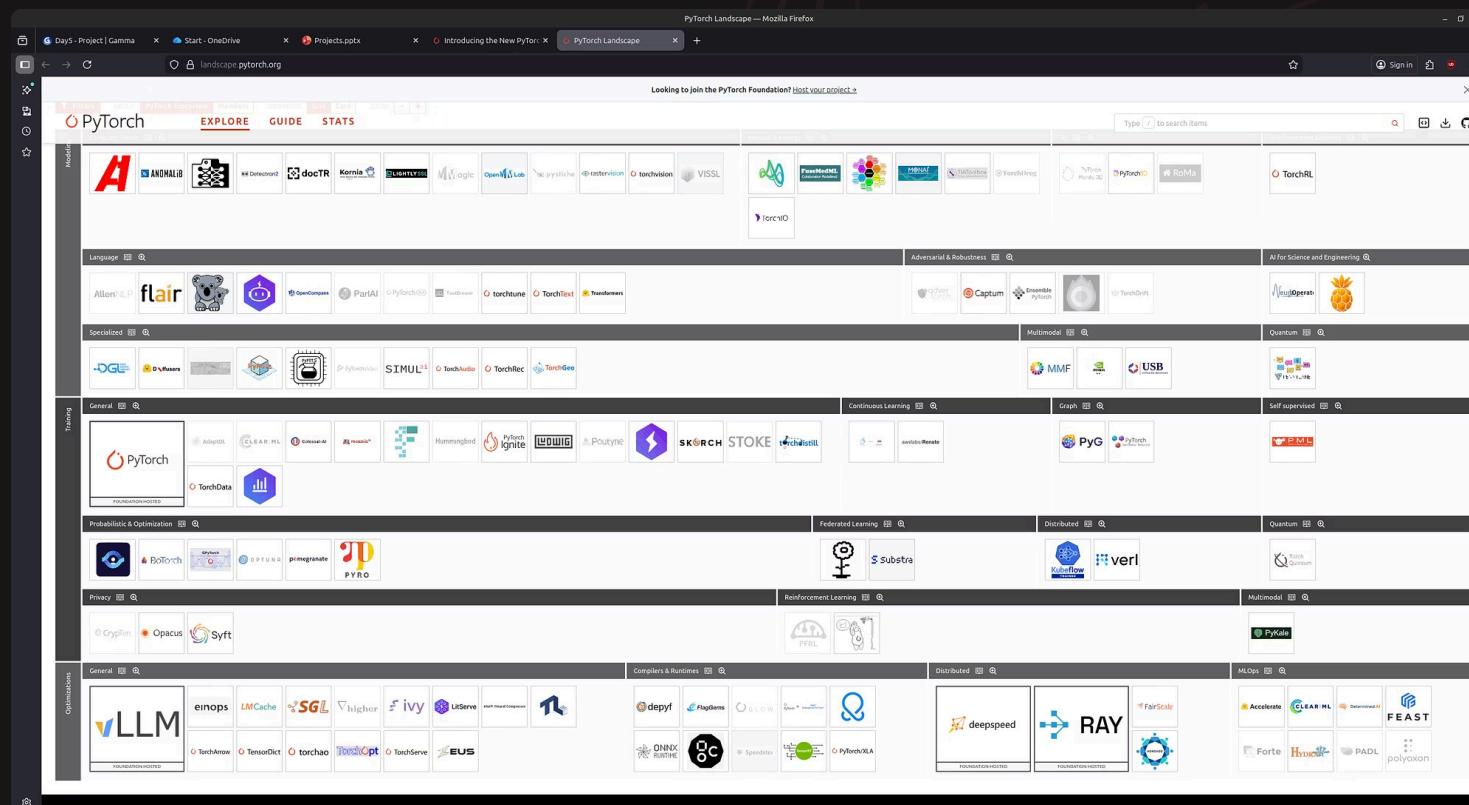
💡 You turn your attention towards open-source projects build on top of Pytorch



# The Pytorch Landscape

 Collection of frameworks build to be used in collaboration with Pytorch <https://landscape.pytorch.org/>

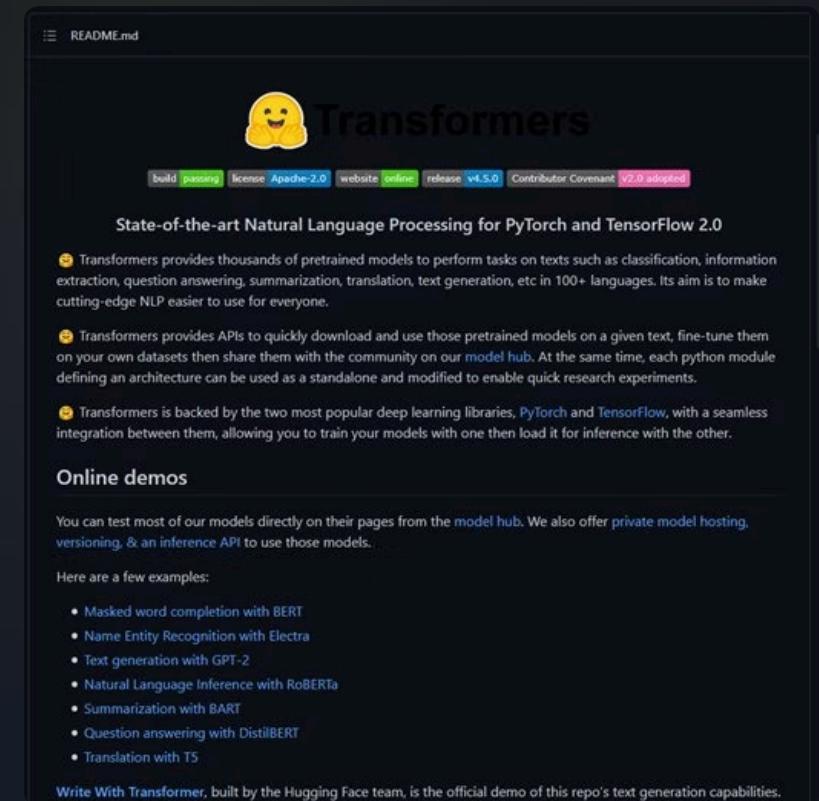
 It is not a complete list of all great frameworks



# Example 1: Transformers

<https://github.com/huggingface/transformers>

Provides state-of-the-art NLP models for both Pytorch, Jax and Tensorflow.



# Example 2: Monai

<https://github.com/Project-MONAI/MONAI>

Models for healthcare imaging

The screenshot shows the GitHub repository page for 'MONAI' (Public). Key features visible include:

- Repository Overview:** Shows 23 people watching, 11 branches, 114 tags, and a 'Code' button.
- Recent Activity:** A list of recent commits, such as 'Replace pyupgrade with builtin Ruff's UP rule (#8606)' and 'Consolidating Version Bumps (#8681)'.
- Files:** A sidebar showing files like .github, docs, monai, tests, .clang-format, .coderabbit.yaml, .deepsource.toml, dockerignore, .gitattributes, .gitignore, .pre-commit-config.yaml, .readthedocs.yml, CHANGELOG.md, CITATION.cff, CODE\_OF\_CONDUCT.md, CONTRIBUTING.md, Dockerfile, LICENSE, MANIFEST.in, README.md, SECURITY.md, environment-dev.yaml, and pyproject.toml.
- Details:** Includes sections for About (AI Toolkit for Healthcare Imaging), Releases (1.5.1 latest on Sep 22, 2025), and Used by (4.4k).
- Contributors:** Shows 257 contributors and a link to +243 contributors.

# Example 3: PyTorch geometric

[https://github.com/pyg-team/pytorch\\_geometric](https://github.com/pyg-team/pytorch_geometric)

Graph Neural Network Library for PyTorch to work on irregular data such as graphs and points.



PyTorch Geometric makes implementing Graph Neural Networks a breeze (see [here](#) for the accompanying tutorial). For example, this is all it takes to implement the `edge convolutional layer`:

```
import torch
from torch.nn import Sequential as Seq, Linear as Lin, ReLU
from torch_geometric.nn import MessagePassing
```

# A open-source framework can usually get you 80% of the way

Open-source frameworks provide a robust foundation for MLOps, covering most common functionalities and significantly reducing development effort.



## Pre-built Models & Algorithms

Access state-of-art, often pre-trained, models and algorithms ready for fine-tuning.

The remaining 20% focuses on unique differentiation:



## Battle-tested Code

Community-maintained, extensively tested, and optimized for higher reliability.



## Strong Community Support

Extensive documentation, tutorials, and forums make learning and troubleshooting accessible.

### Customization

Tailor models to unique business logic, data types, and performance needs.

### Integration

Connect ML pipelines with existing systems, data sources, and applications.

### Deployment

Adapt to infrastructure, set up monitoring, scaling, and security protocols.

# Your first task



## Find a Dataset

Identify a compelling dataset that aligns with your interests and project goals. Consider its structure, size, and relevance.



## Choose a Model

Select a suitable machine learning or deep learning model. Explore various architectures and their applications for your chosen dataset.



## Set the Right Scope

Aim for a challenge that is harder than basic benchmarks (e.g., MNIST, CIFAR) but easier than training a large language model (LLM) from scratch. Find your sweet spot!

# How to get an good idea?

Look at the **used by** section on github

The screenshot shows the GitHub repository page for `kornia.org`. The top navigation bar includes options for switching branches (master), viewing 11 branches, 16 tags, Go to file, Add file, and Code. The main content area displays a list of recent commits from user `edgarriba`, followed by a detailed view of the repository's structure and activity.

**Commits:**

- update new kornia logo (e36ca3d, 2 days ago)
- upgrade ci workflow with pytorch 1.8 (#892) (.circleci, 29 days ago)
- Create CODEOWNERS (#947) (.github, 2 days ago)
- [Feat] Add tpu support for the losses module (#834) (docker, 3 months ago)
- update new kornia logo (docs, 2 days ago)
- Updated doc & example for augmentation (#583) (examples, 8 months ago)
- Fixed the issue of NaN gradients by adding epsilon in focal loss (#924) (kornia, 2 days ago)
- remove pytorch version variable (packaging, 8 months ago)
- Deprecate some augmentation functionals (#943) (test, 2 days ago)
- Fixed tests and docs (#654) (tutorials, 7 months ago)
- Create .codecov.yml (#735) (.codecov.yml, 6 months ago)
- reorganize color module (.gitconfig, 2 years ago)
- Update gitignore to avoid version.py (.gitignore, 2 years ago)
- create CHANGELOG and update for 0.4.1 (#726) (CHANGELOG.md, 6 months ago)
- Create CITATION.md (#949) (CITATION.md, 2 days ago)
- add code of conduct file (CODE\_OF\_CONDUCT.md, 2 years ago)
- Update CONTRIBUTING.rst (#316) (CONTRIBUTING.rst, 17 months ago)

**About:** Open Source Differentiable Computer Vision Library for PyTorch

**Tags:** machine-learning, computer-vision, image-processing, pytorch

**Readme:**

**View license:**

**Releases:** 16

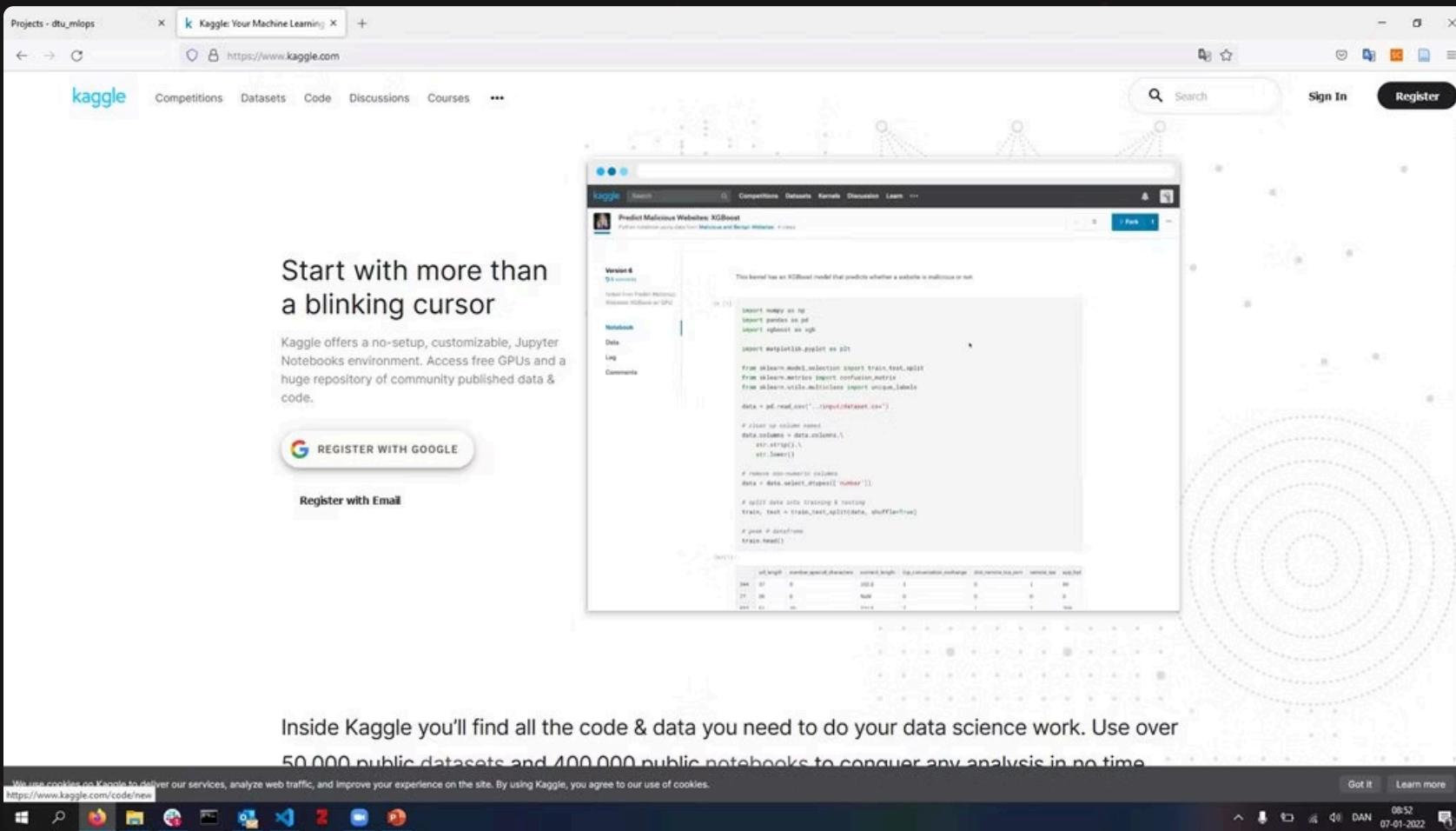
Morphological operators, Deep learning, + 15 releases

**Packages:** No packages published

**Used by:** 290

+ 282

# How to get an good idea?



# How to get an good idea?

The screenshot shows a Mozilla Firefox browser window with the Hugging Face Datasets page open. The URL in the address bar is `huggingface.co/datasets`. The page has a dark theme with orange highlights. On the left, there's a sidebar with filters for 'Main' tasks, 'Libraries', 'Languages', 'Licenses', 'Other', 'Modalities' (3D, Audio, Document, Geospatial, Image, Tabular, Text, Time-series, Video), 'Size (rows)' (1K to 1T), 'Format' (json, csv, parquet, optimized-parquet, imagefolder, soundfolder, webdataset, text, arrow), and 'Evaluation' (Benchmark). The main content area displays a grid of dataset cards. Each card includes the dataset name, a small icon, a brief description, and metrics like 'Updated 3 days ago', '43.7k', and '107'. Some cards also show 'Viewer' and 'Preview' options. The first few cards include 'genrobot2025/10Kn-RealMind-OpenData', 'facebook/research-plan-gen', 'OpenDataArena/00A-Mixture-500k', 'Anthropic/hh-rllf', 'nvidia/Nemotion-Math-v2', 'llm-jp/jhle', 'OpenDataArena/00A-Mixture-100k', 'WNT3D/Ultimate-Offensive-Red-Team', 'OpenDataArena/00A-Math-460k', 'nvidia/PhysicalAI-Autonomous-Vehicles', 'Goutieff/ReActor', 'wikimedia/Wikipedia', 'missvector/linux-commands', '123lop/binance-futures-ohlcv-2018-2026', 'DanielleSry/TransPhy3D', 'opendatabl/ScienceMetaBench', 'Nanbeige/ToolMind', 'MLCommons/people\_speech', 'roneneldan/TinyStories', 'Idavidrein/gpqa', 'xihuuywh/DRIM-VisualReasonHard', 'TeichAI/claudie-4.5-opus-high-reasoning-250x', 'Lewandoński/OpenVE-3M', 'Salesforce/wikitext', 'bshada/open-schematics', and 'TeichAI/glm-4.7-2000x'. A message at the top says 'Hugging Face is way more fun with friends and colleagues! Join an organization'.

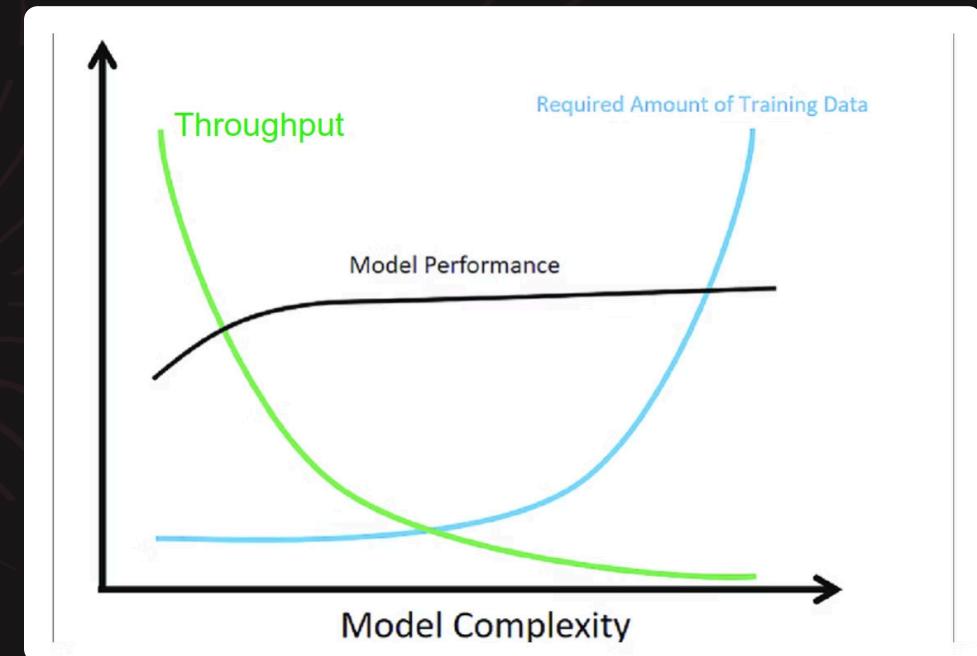
# General recommendations

## Data

- Choose where data loading is not too complex
- <10 GB (else work on a subset)

## Model

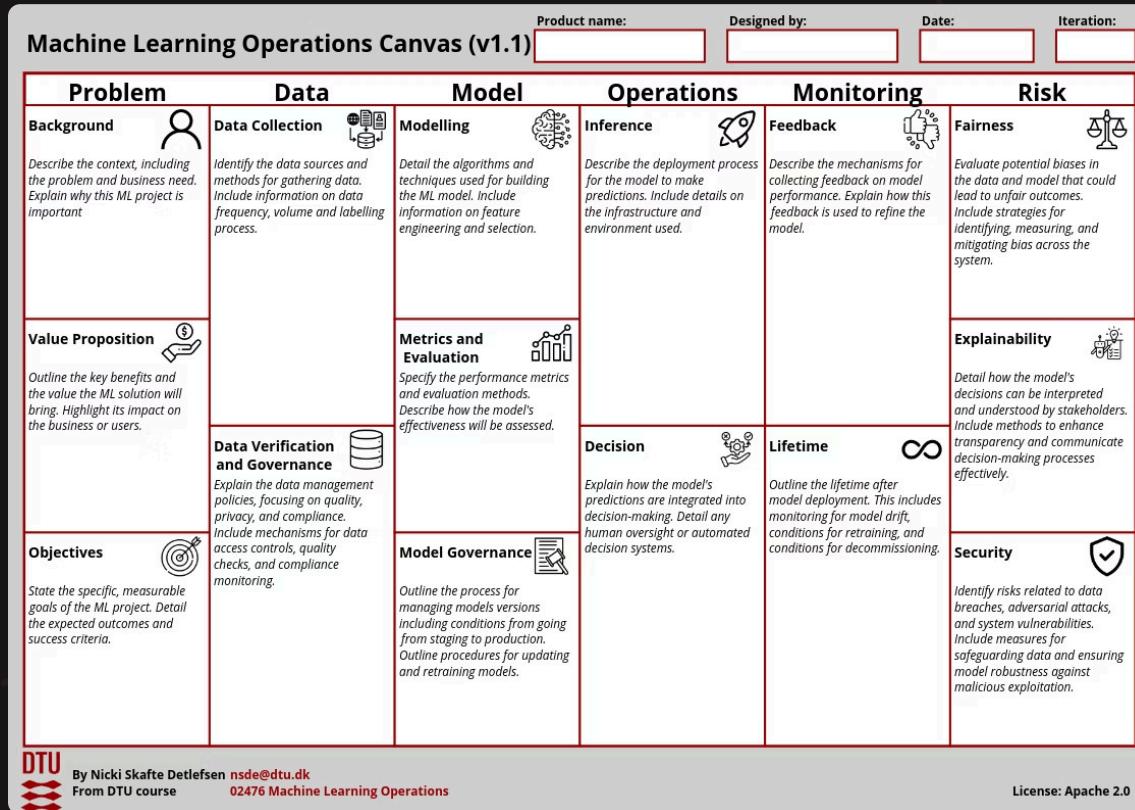
- Start out with a public baseline model if possible
- Choose smaller models over large models



# Summary

1. Pick a dataset you would like to work with
2. Pick a model you would like to work with
3. Write a small project description containing
  - a. Overall goal of the project
  - b. What data are you going to run on (initially, may change). Describe overall number of samples, size, modality...
  - c. What models do you expect to use
4. Create project repository
5. Upload project description as part of **[README.md](#)** file
6. Work on the rest of project...

# ML Canvas for staying organized and thinking ahead



A structural framework for staying organized for large machine learning projects and making sure all the different phases are aligned

[https://github.com/SkafteNicki/dtu\\_mlops/tree/main/canvas](https://github.com/SkafteNicki/dtu_mlops/tree/main/canvas)

Made with **GAMMA**

# Project checklist

⚠ You do not need to do everything to pass, the list is meant to be exhaustive

## Week 1

- Create a git repository
- Make sure that all team members have write access to the github repository
- Create a dedicated environment for your project to keep track of your packages (using conda)
- Create the initial file structure using cookiecutter
- Fill out the `make_dataset.py` file such that it downloads whatever data you need and
- Add a model file and a training script and get that running
- Remember to fill out the `requirements.txt` file with whatever dependencies that you are using
- Remember to comply with good coding practices (`pep8`) while doing the project
- Do a bit of code typing and remember to document essential parts of your code
- Setup version control for your data or part of your data
- Construct one or multiple docker files for your code
- Build the docker files locally and make sure they work as intended
- Write one or multiple configurations files for your experiments
- Used Hydra to load the configurations and manage your hyperparameters
- When you have something that works somewhat, remember at some point to do some profiling and see if you can optimize your code
- Use wandb to log training progress and other important metrics/artifacts in your code
- Use pytorch-lightning (if applicable) to reduce the amount of boilerplate in your code

# How is the project evaluated?

✓ We look at how well you can use the tools and techniques from the material in your project

⚠ We do not look at how good model performance you get

⚠ We do not look at how complex a model and dataset you are using

I am specifically looking at

💻 How well are your code, data, experiments version controlled and reproducible

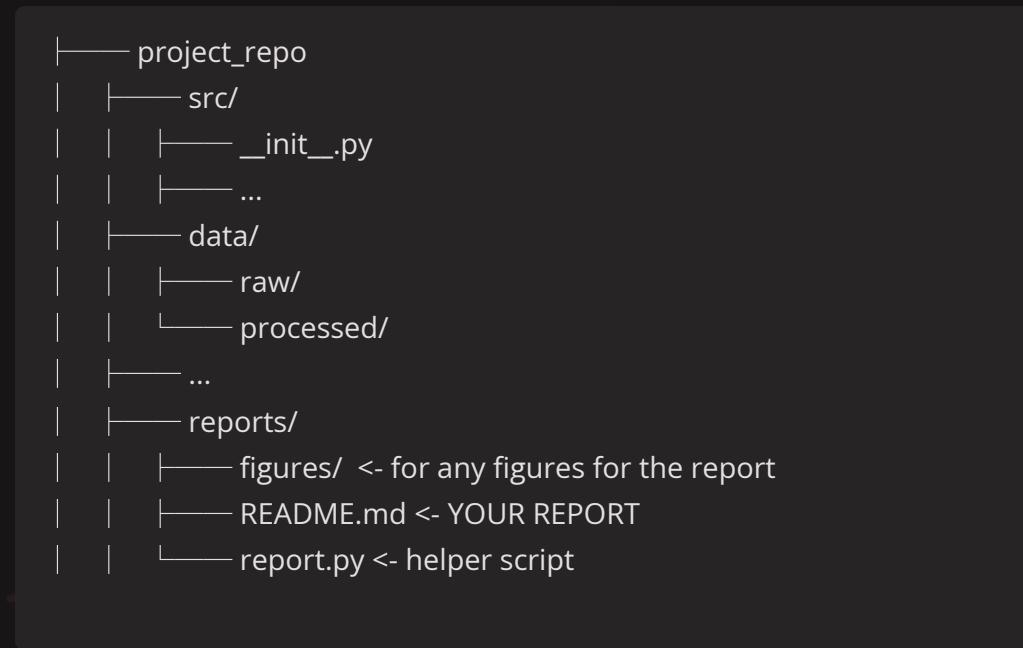
♻️ Is appropriate continues integration implemented for automatization of tasks

📦 Is a final model deployed online and able to be interacted with a end user

🤝 How well does it look like you have collaborated on the project

# Exam report template

Add this to your public project repository



The README.md content is as follows:

```
Exam template for 02476 Machine Learning Operations

This is the report template for the exam. Please only remove the text formatted with as three dashes in front and behind like:
--- question 1 fill here ---
where you instead should add your answers. Any other changes may have unwanted consequences when your report is auto generated in the end of the course. For questions where you are asked to include images, start by adding the image to the figures subfolder (please only use .png, .jpg or .jpeg) and then add the following code in your answer:
!{my_image}(figures/<image>.extension)
```

[https://github.com/SkafteNicki/dtu\\_mllops/tree/main/reports](https://github.com/SkafteNicki/dtu_mllops/tree/main/reports)

I will scrape your reports and repositories on the 23/1 at 23:59.

# Hand-in for today

Should be handed in before midnight today

- If all have access to learn, signup to a group and hand-in
- If only one or more group members are missing from learn, still hand-in as a group and send a email with remaining student ids to me

The screenshot shows a table of project groups. A red arrow points from the 'Assignment' column of the table to a 'Text Submission 1' card.

	Groups	Members	Assignment	Discussions	Locker
<input type="checkbox"/>	MLOps 1	4	Project reposi... <small>?</small>		
<input type="checkbox"/>	MLOps 2	4	Project reposi... <small>?</small>		
<input type="checkbox"/>	MLOps 3	4	Project reposi... <small>?</small>		
<input type="checkbox"/>	MLOps 4	4	Project reposi... <small>?</small>		
<input type="checkbox"/>	MLOps 5	1	Project reposi... <small>?</small>		

**Text Submission 1**

Unevaluated Friday, 5 January 2024 3:22 PM

<https://github.com/> Username /E project\_repo

# Meme of the day

[https://skaftenicki.github.io/dtu\\_mlops/pages/projects/](https://skaftenicki.github.io/dtu_mlops/pages/projects/)

**When someone asks why you never stops  
talking about machine learning**

