Are You Using Airflow Or Similar SW For ML Pipelining?
You're Doing It All Wrong.

ClearML @ MLLifecycleConf, Jan 26 , 2021

Ariel Biller (@LSTMeow)
Evangelist

Disclaimers:
Tried to balance tech/mgmt
In depth Airflow vs. Others - offline🙏

FEEDBACK PLS

https://twitter.com/LSTMeow

https://www.linkedin.com/in/LSTMeow/

https://www.reddit.com/user/LSTMeow

https://www.facebook.com/ariel.biller.LSTMeow

CLEAR|ML

“AIRFLOW” or similar software

Apache Airflow - “A platform to programmatically author, schedule, and monitor workflows”

“Airflow works best with workflows that are mostly static and slowly changing. When DAG structure is similar from one run to the next, it allows for clarity around unit of work and continuity”.

Kubeflow - “A *machine learning toolkit for Kubernetes*”

(not the same “flow” - this one comes from Kubernetes+TensorFlow)

Amazing tools!
For static workflows…

CLEAR|ML

Outline

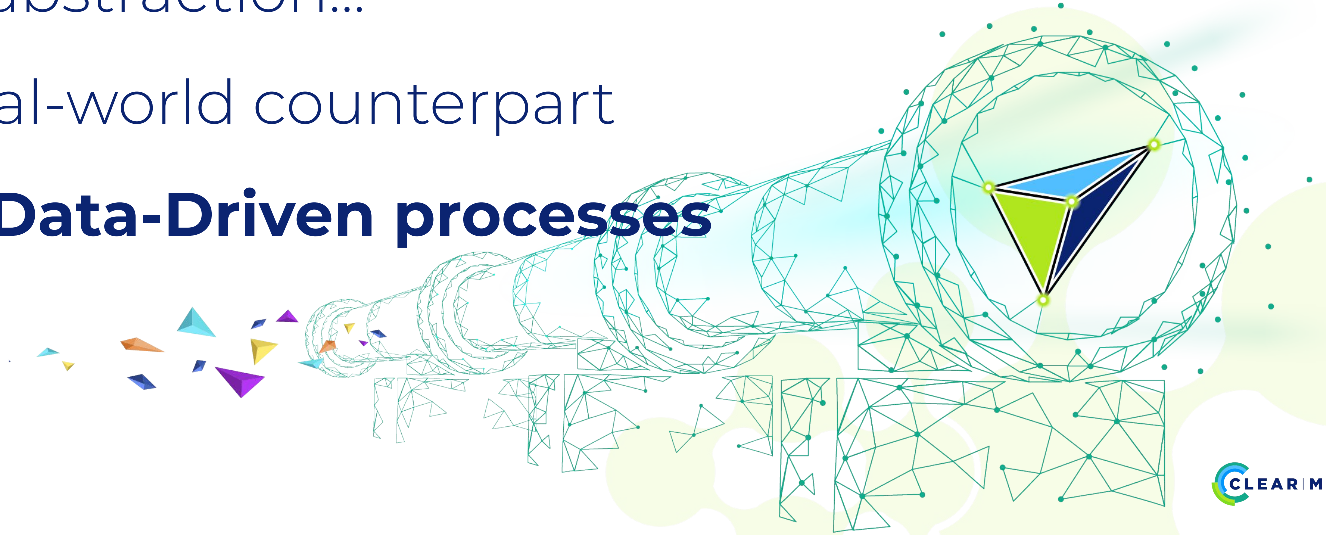
- Machine Learning Pipelines
- From Research to Production (More Than ML Pipelines)
- The case for Pipelines in R&D
- Wrong Pipe - why not “Airflow”

Part I - ML Pipelines

What is so special about pipelines?

"...ordered stages to process sequence of input values..."

- **Universal programming paradigm**
  - Helpful abstraction...
  - ...with real-world counterpart
- **Really fits Data-Driven processes**



A process is “considered” as a ML pipeline if:

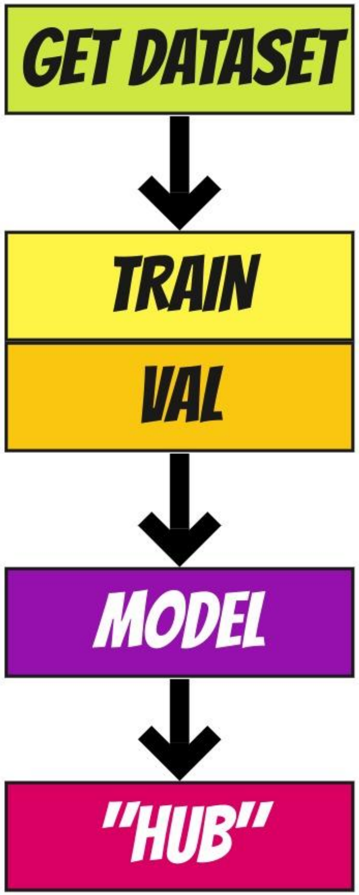
- “Consumes” data
- Multiple steps
- Inter-step dependency is data/model
- Takes a while...
- Result is a model

The Default ML Pipeline:

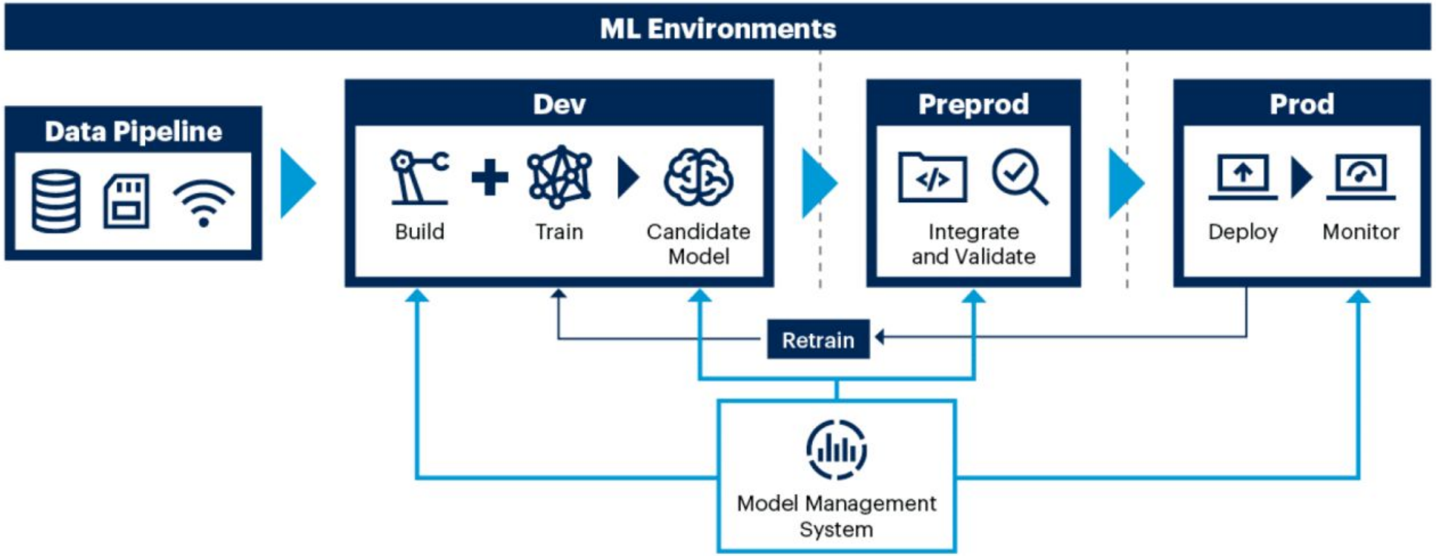
- Dataset in, Model out (DIMO?)
- “Get Dataset” : “80% of the work”
- Often another (data) pipeline
- Most use-cases are manual
- Can be SOTA, but error prone!

Training, validating,

and... storing for further use.



“Typical” “Production” “end-to-end” ML pipeline:



100% Automated

Source: NVIDIA MLOps blogpost

CLEAR|ML

Automating: Coffee Pipelines

Manual pipeline:

The French Press

- **Coffee grounds** (how did you get these?)
- **Hot water** (what temperature?)
- **Fill** (how much?)
- **Wait** (how long?)
- **Press** (how fast?)
- **Coffee (irreproducible)**

Photo by Sonny Ravesteijn on Unsplash

Automated, scaled:

Capsule Coffee

- **Capsules** (predetermined content)
- **Water** (just fill the tank)
- **Press Button**
- **Coffee (same every time)**

Is this a compromise? Yes.

Massively reproducible ? Yes.

Photo by Felipe Giacometti on Unsplash

CLEAR|ML

ML-Pipeline in “production”:

“The <u>element</u> that turns the data into models”

- **Data** (how much?)
- **Press <u>Button</u>** (trigger)
- **Model (how good?)**

“no serviceable parts”

(no more research)

Photo by Crystal Kwok on Unsplash

From Research to Production:
“Who is in charge of this monstrosity?”

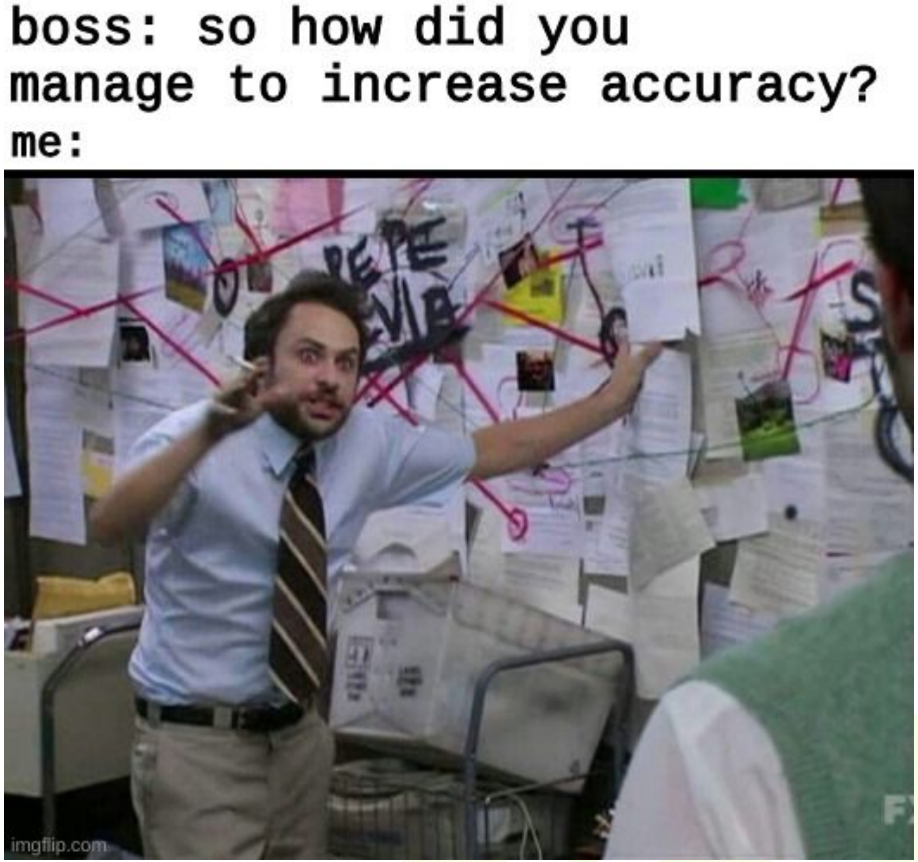
Photo by Najib Kalil on Unsplash

Part 2 -
From Research to Production
“More Than Just ML Pipelines”

CLEAR|ML

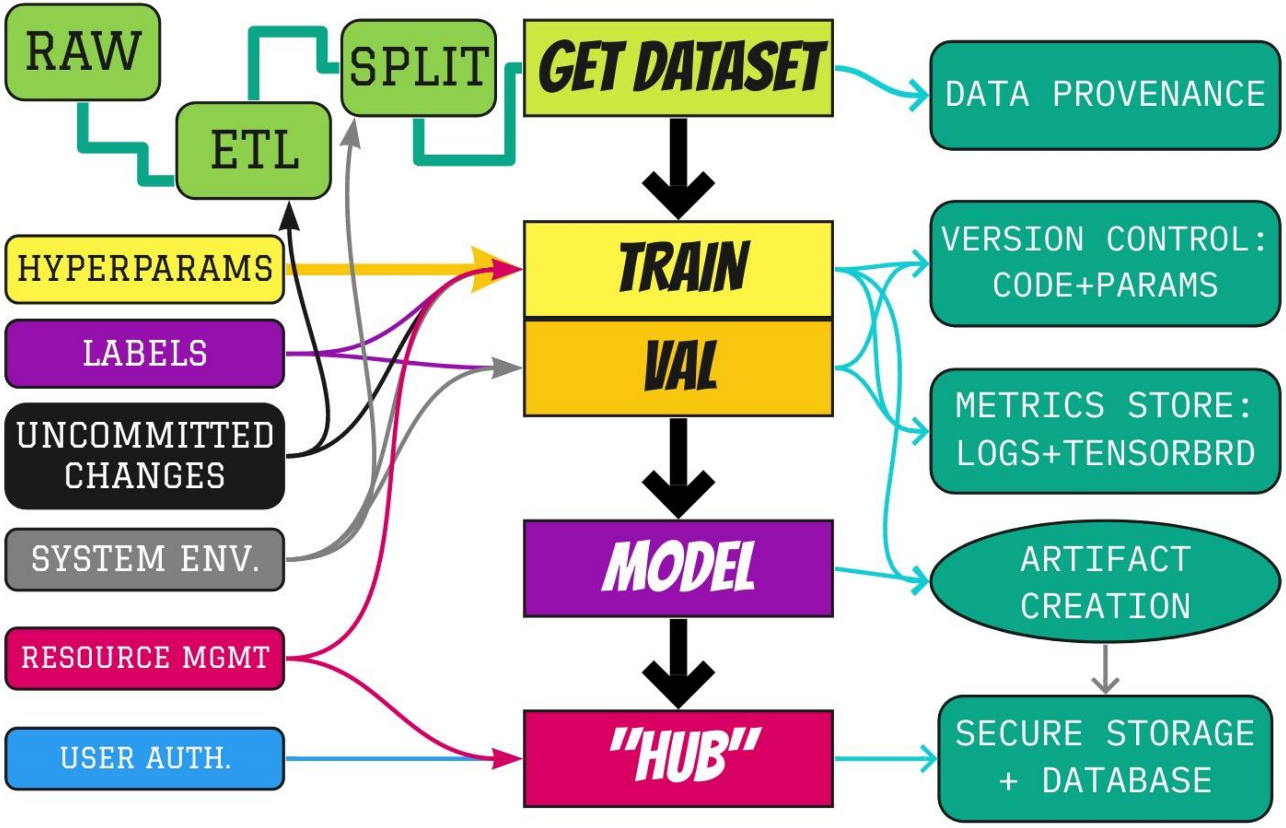
This is ML/DL R&D

- “80% is data wrangling”
- fast paced, multi env.
- “best model” approach
- beyond traditional CI/CD
- Too Many Experiments
- “who deploys this”



ML/DL Research is inherently messy

CLEAR|ML



THIS IS WHY YOU NEED A PLATFORM

CLEAR|ML

Production: “Everything is MLOPs”

- Model serving
- Data Preprocessing
- *etc*... (hardware!?)

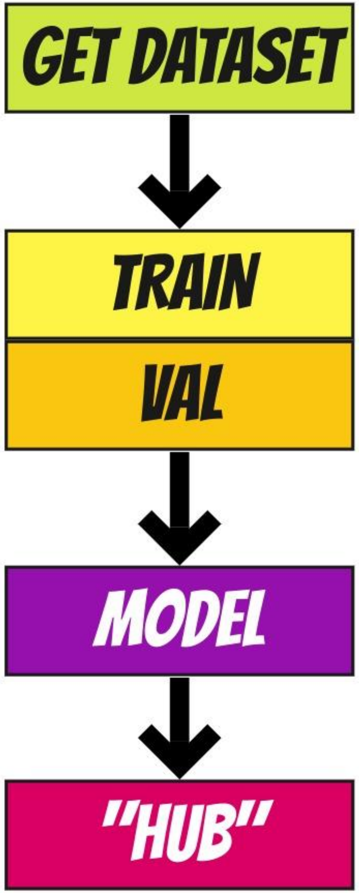
MLOps for R&D:

- Automation
- Orchestration
- Reproducibility
- **Integrates with workflow**

Part 3 - R&D
The case for pipelines

Can we use ”static” pipelines for research?

- No.
- Configuration Overload!
- Breaks Workflow (slow)
- Yet Another Tool...
- Generally Non Parameterizable



CLEAR|ML

MLOps prerequisites:

You are already (right?) using an experiment tracking platform:

- Full tracking (incl. pipelines)
- Offload to remote execution
- Parametrization Interface
- Easy to use!!!

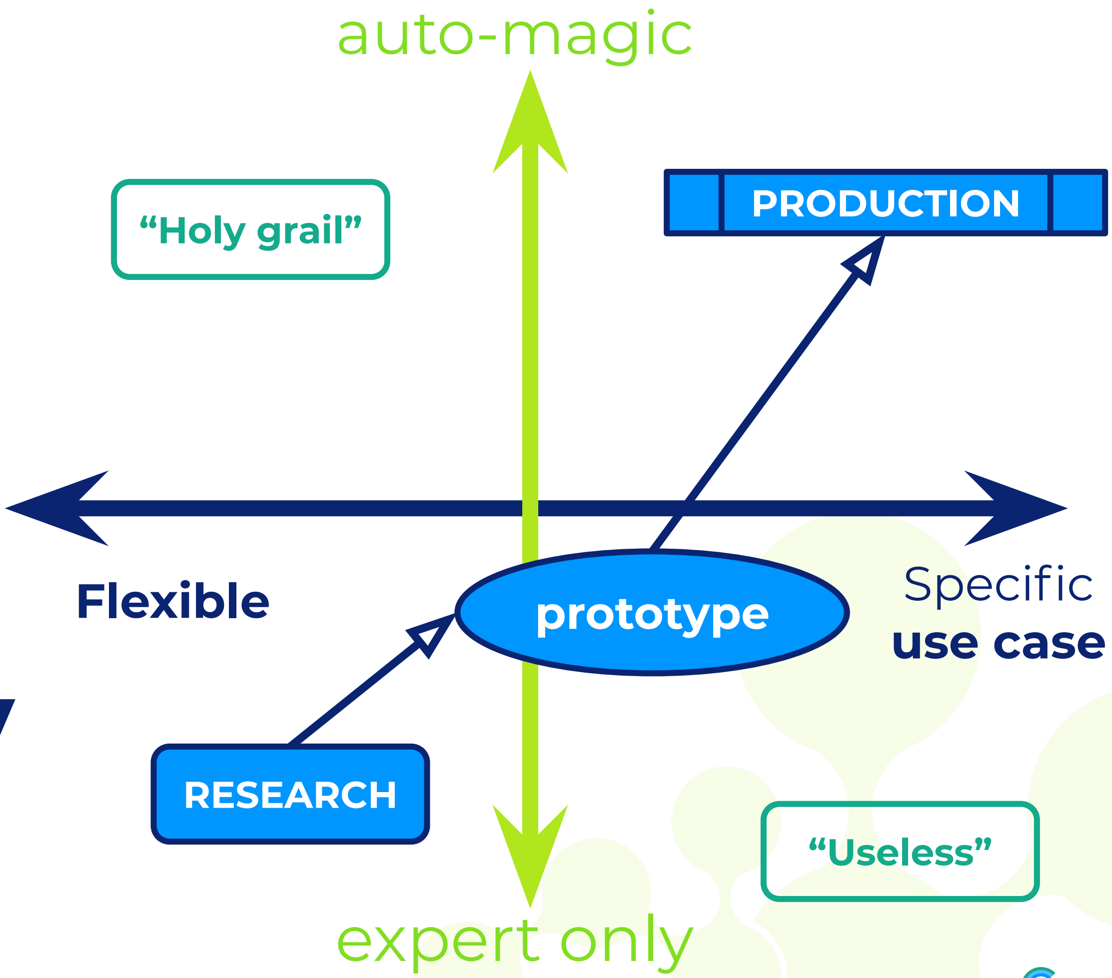
Does your platform have pipelines?

CLEAR|ML

Interface

vs.

Specificity

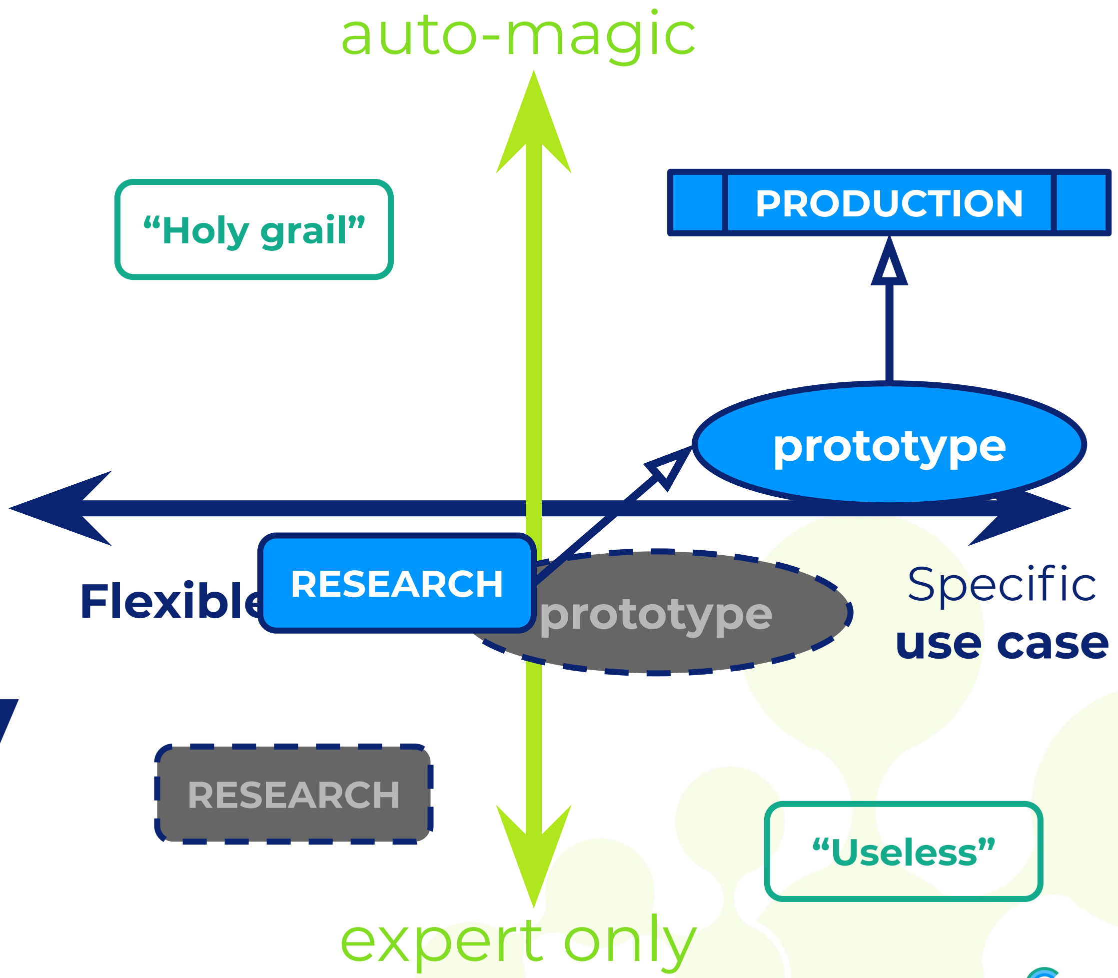


Pipelined!

Interface

VS.

Specificity



Pipelines for R&D:

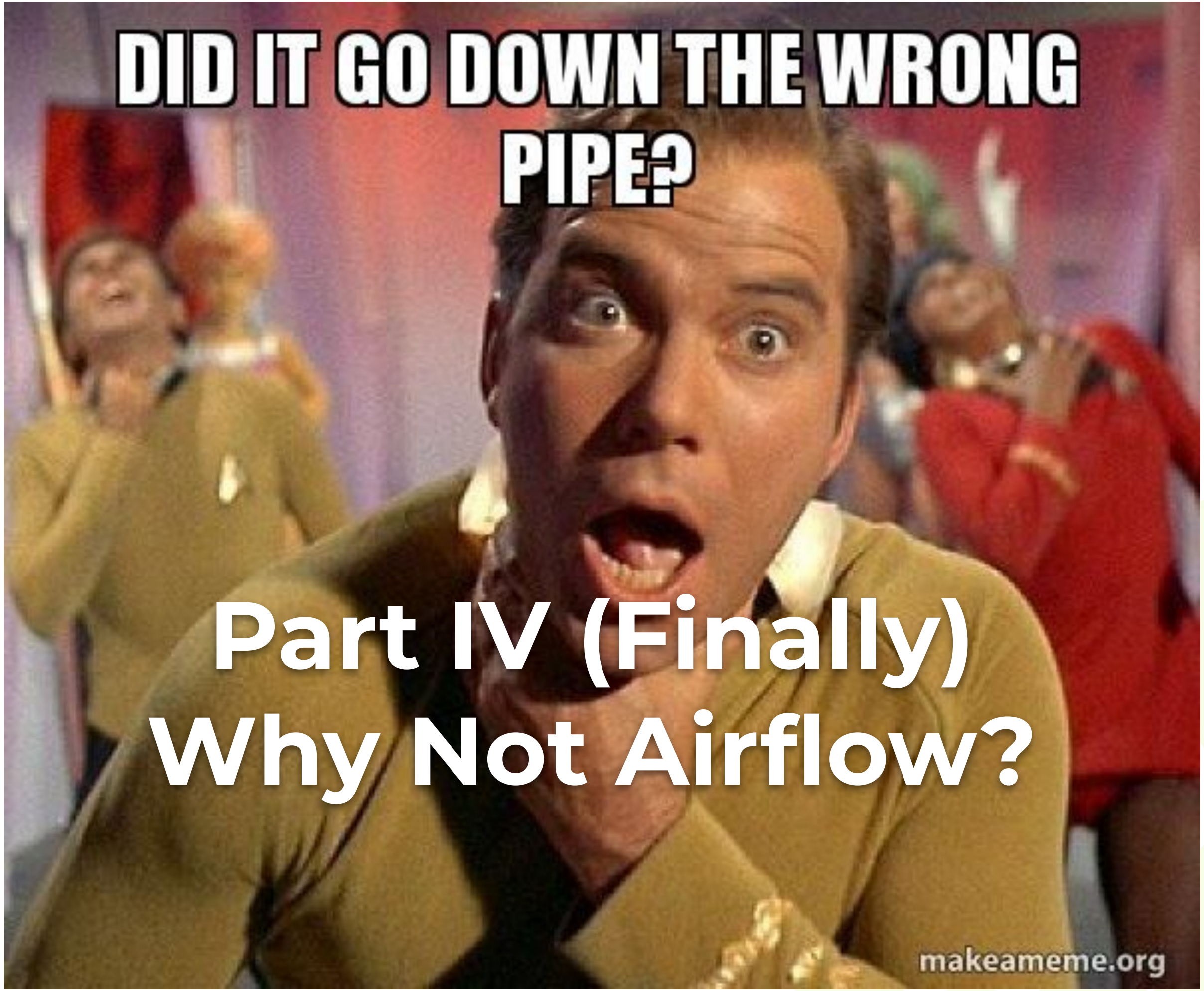
Rapid, reproducible iterations on complex experiments

- **Workflow Orchestration**
- **Workflow Version Control**
- **Workflow Parametrization**
- **Modular (standalone elements)**

“Pipelined” Research - superior vantage point towards production (🤞)

Photo by bradley pisney on Unsplash

CLEAR|ML



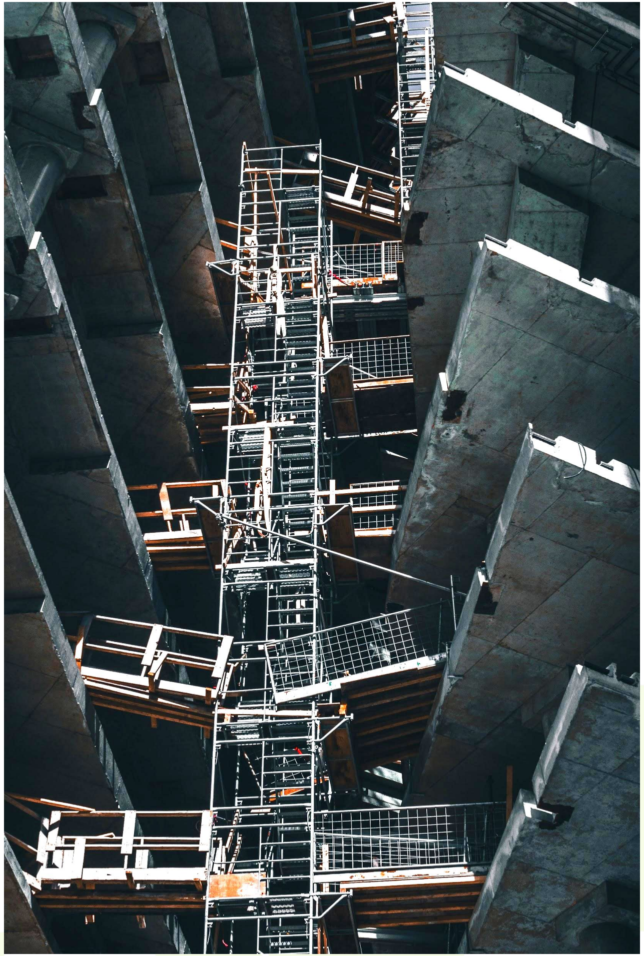
Part IV (Finally)
Why Not Airflow?

But how to "add pipelines"?

- ~~D.I.Y.~~
- **Adapt existing Pipelines**
- **Wrap using dedicated Tool**
- **Grow using existing MLOps**

Top-Down Pipeline Design

1. **Conclude your research work**
2. **Add the necessary code snippets everywhere and define the DAG**
3. **Lay down all the interfaces and connections**
4. **Fix all the bugs…**
5. **If you want to change anything, start from scratch…**



Bottom-up Pipeline Design

1. **Integrate with platform for remote execution (0-2 LOC)**
2. **Remote execution works (0 LOC)**
3. **Successful Experiment id is now template - valid pipeline stage.**
4. **Populate pipeline by repeating (3)**
5. **Change and iterate at will!**

CLEAR|ML

Wrap vs. Grow

Existing pipeline	Top-down (Wrap)	Bottom-up (Grow)
Not the best fit for the job	**Tailor-made to workflow**	
Robust scheduling and execution	**Scales well from single seat -> team**	**Only as good as your platform :)**
Hard to iterate on pipeline design Questionable flexibility		**Everything always clicks <u>by design</u>**

Summary - Why not airflow?

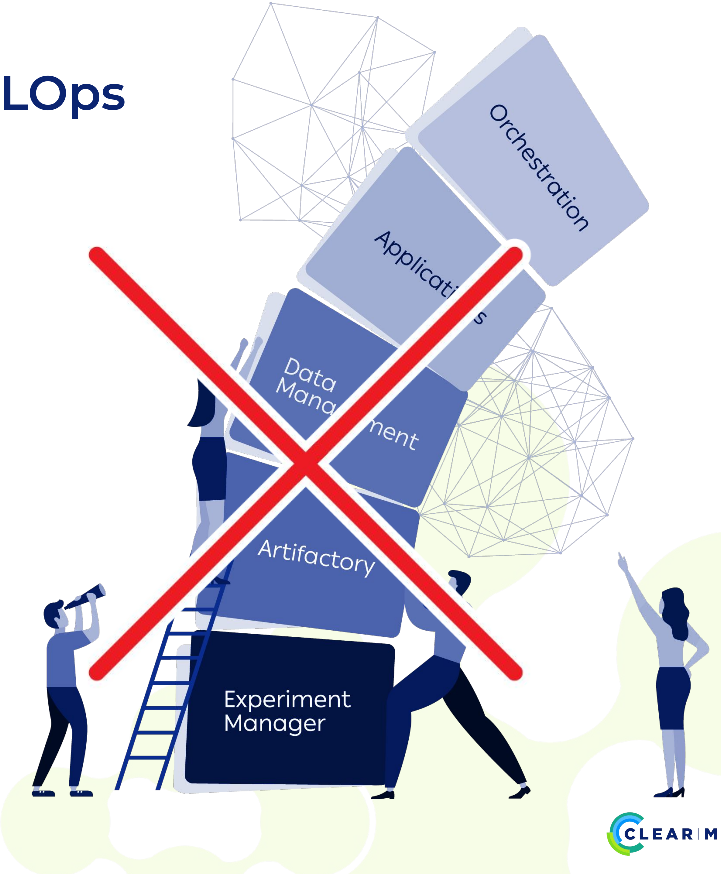
- You *will* use ML pipelines in production
- You *should* use ML pipelines in R&D
- These are not necessarily the same kind of ML pipelines
- Build ML pipelines during research
- By who? Researchers!
- And Scale, Data pipes? Engineers.

CLEAR|ML

ClearML: Made With 💙 From Allegro AI

- **Single solution towards "lean-stack" MLOps**
  - Experiment management
  - Workload orchestration
  - Data management
  - More coming...

- **"Bottom-Up" design**
  - Easy integration with ML/DL code
  - Log and keep, compare everything
  - One-click orchestration
  - Workflow versioning (pipelines!)
  - Dataset and model management
  - Remote session work/debug



Open source & trusted by leading brands

AI teams in over 1,000 organizations rely upon ClearML for their development and deployment

CLEAR|ML

Thank You!

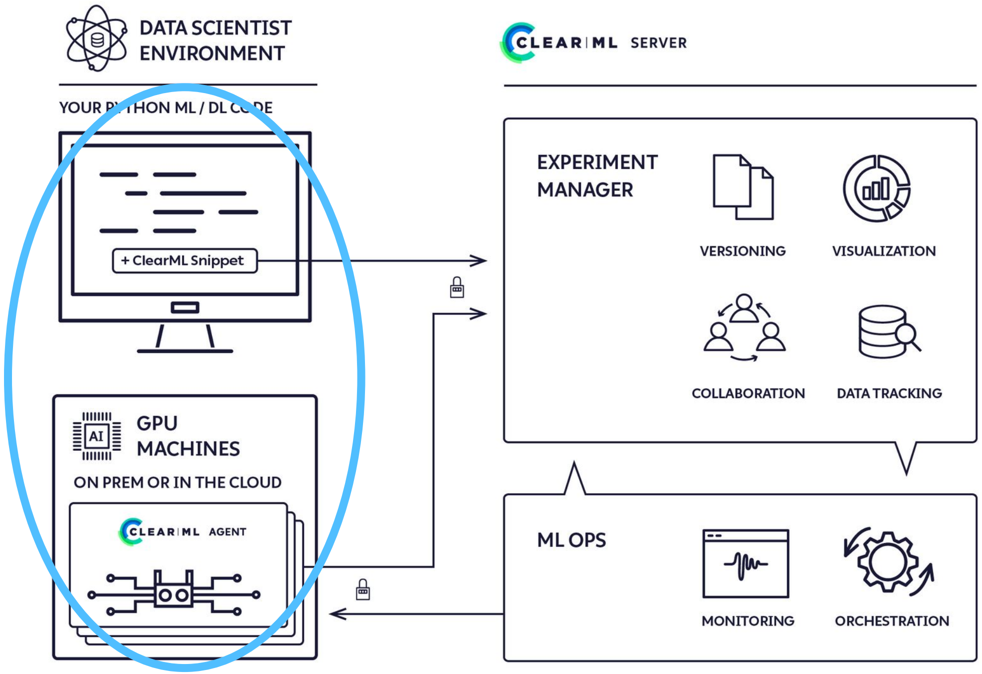
ariel@clear.ml

@LSTMeow

Appendix

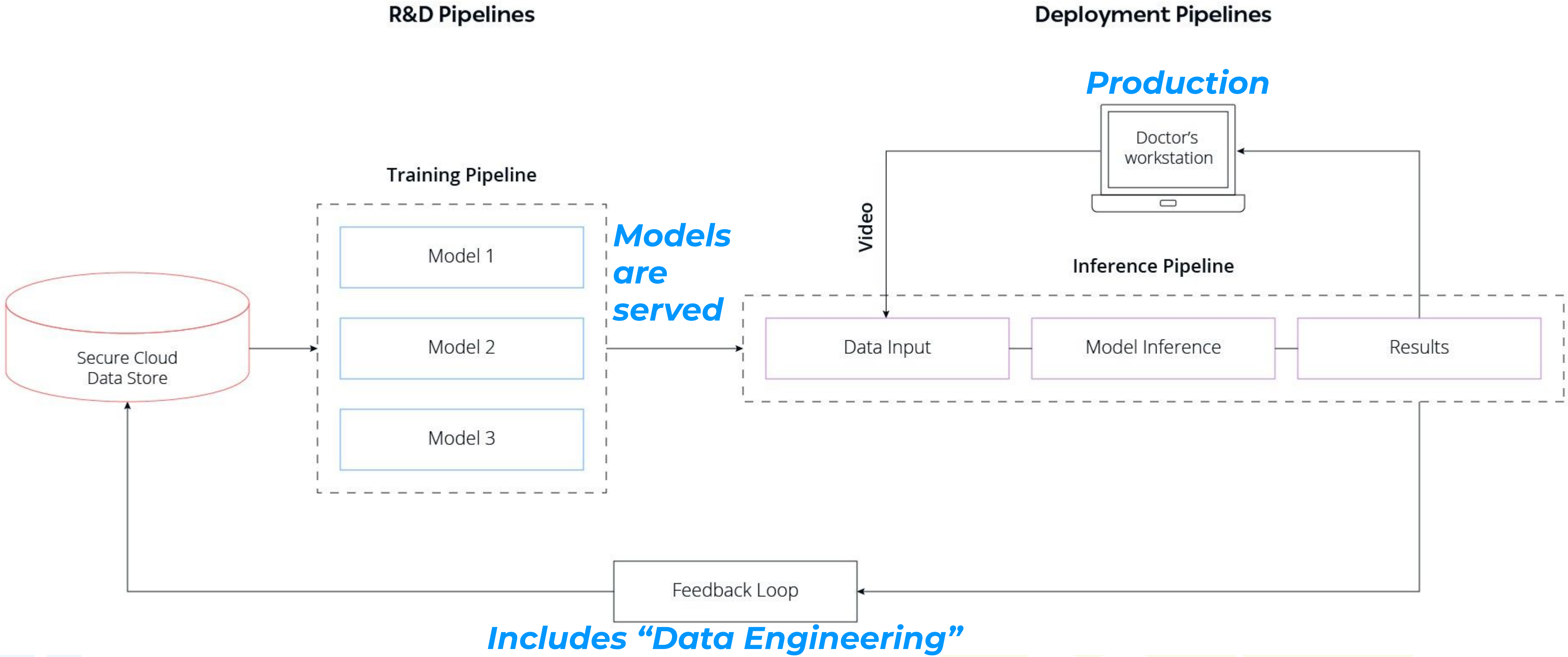
(real world examples)

ClearML: Open Source MLOPs solution

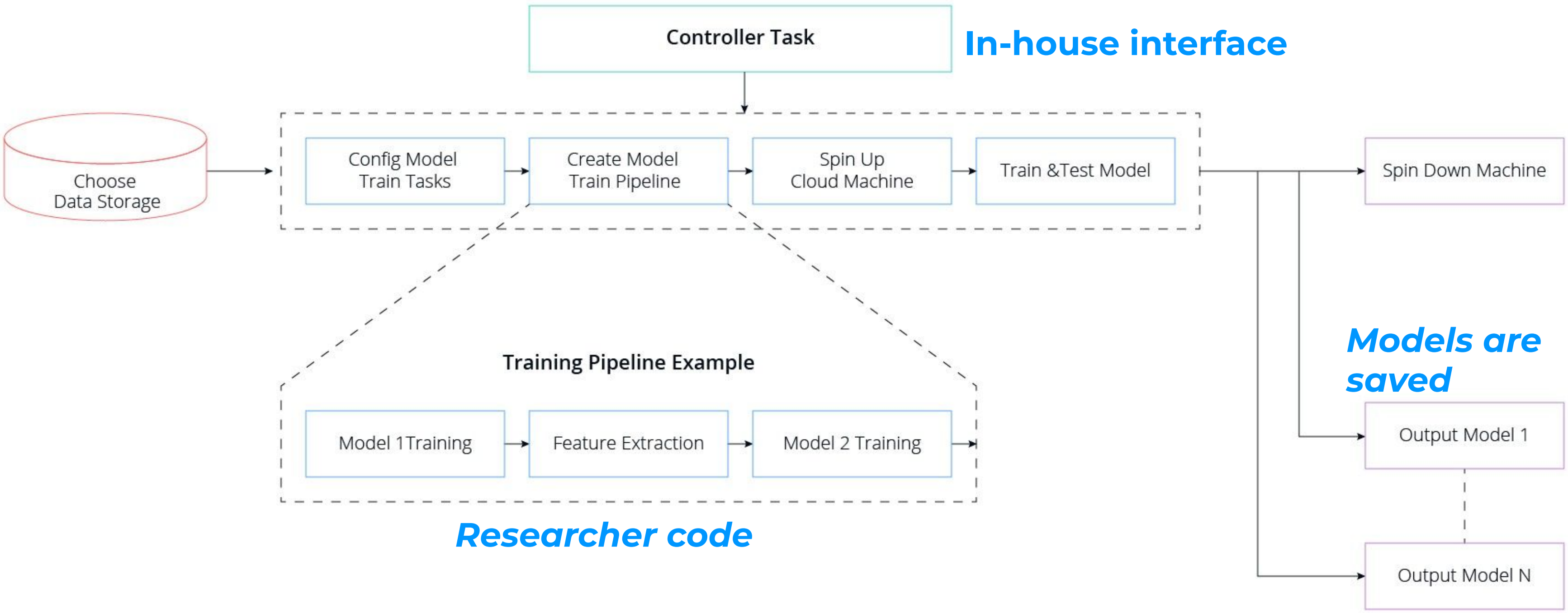


- Experiment Management
- Dataset Management & Lineage
- Model Management & Lineage
- Orchestration + Scheduling
- Remote Development Support
- Hyperparameter Optimization
- Pipelines

Deep Learning in production @ Theator



Orchestration in research @ Theator



Choose your MLOps story:

- Build from scratch (x2)
- Use experiment tracking
    - Only part of the story
- Assemble stack from OSS
    - Rewrite code
    - Add YAMLs
    - Orchestration still difficult
- **Use ClearML for free**

Finally: **If you are reading this - let's chat :)**

Join the ClearML Community:

- For feature requests or bug reports, see [ClearML GitHub Issues].
- If you have *any* questions, post on the clearml [Slack Channel].
- Or, tag your questions on [stackoverflow] with the clearml tag.
- You can always find me at ariel@clear.ml