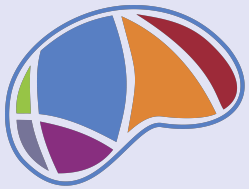
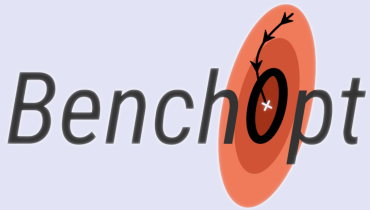
AI benchmarking infrastructures: lessons from Benchopt

Thomas Moreau



MIND



The Era of Benchmarks: AI as an empirical science

The ImageNet competition

- Annual competition since 2010
- Evaluate image classification methods with 14M labeled images among 1k categories
- Top-1/Top-5 accuracy reported on a fixed test set

The ImageNet competition

- Annual competition since 2010
- Evaluate image classification methods with 14M labeled images among 1k categories
- Top-1/Top-5 accuracy reported on a fixed test set
- **Boosted AI and Deep Learning research when Alex Krizhevsky won in 2012.**

The ImageNet competition

- Annual competition since 2010
- Evaluate image classification methods with 14M labeled images among 1k categories
- Top-1/Top-5 accuracy reported on a fixed test set
- **Boosted AI and Deep Learning research when Alex Krizhevsky won in 2012.**

$\Rightarrow$ Demonstrates the importance of benchmarks to drive research in AI.

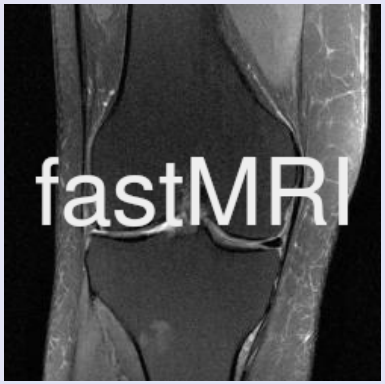
Many benchmarks in AI

Many benchmarks followed ImageNet:

- Natural Language Processing: GLUE, SuperGLUE
- Reinforcement Learning: Atari, MuJoCo, OpenAI Gym
- Others: fastMRI, DAWNBench, MLPerf, etc.

GLUE

Gymnasium



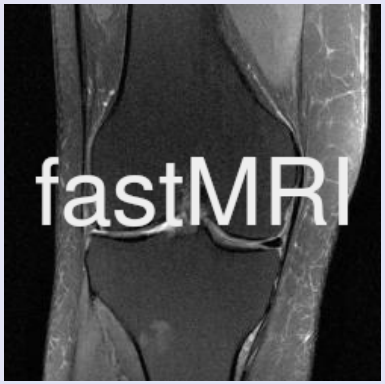
Also many dataset repos:

- OpenML, UCI, PhysioNet, HuggingFace, etc.

Many benchmarks in AI

Many benchmarks followed ImageNet:

- Natural Language Processing: GLUE, SuperGLUE
- Reinforcement Learning: Atari, MuJoCo, OpenAI Gym
- Others: fastMRI, DAWNBench, MLPerf, etc.



Also many dataset repos:

- OpenML, UCI, PhysioNet, HuggingFace, etc.

$\Rightarrow$ "Benchmarks" are now ubiquitous in AI research.

What makes a benchmark?

A benchmark is defined by three components:

- **Objective**: what is being measured?
- **Dataset**: on what evidence?
- **Solvers/Methods**: What are we comparing?

$\Rightarrow$ A benchmark is not only a dataset with any type of evaluation protocol.

What makes a benchmark?

A benchmark is defined by three components:

- **Objective**: what is being measured?
- **Dataset**: on what evidence?
- **Solvers/Methods**: What are we comparing?

Different benchmark goals emphasize different components.

- **Challenge benchmarks:** Fixed dataset + single metric $\rightarrow$ stresses solver comparison. Focus on performance

What makes a benchmark?

A benchmark is defined by three components:

- **Objective**: what is being measured?
- **Dataset**: on what evidence?
- **Solvers/Methods**: What are we comparing?

Different benchmark goals emphasize different components.

- **Challenge benchmarks:** Fixed dataset + single metric $\rightarrow$ stresses solver comparison.
  Focus on performance
- **SOTA tracking benchmarks:** Multiple metrics and datasets, compare solvers to measure progress over time.
  Risk: test-set overfitting / cherry-picked metrics.

What makes a benchmark?

A benchmark is defined by three components:

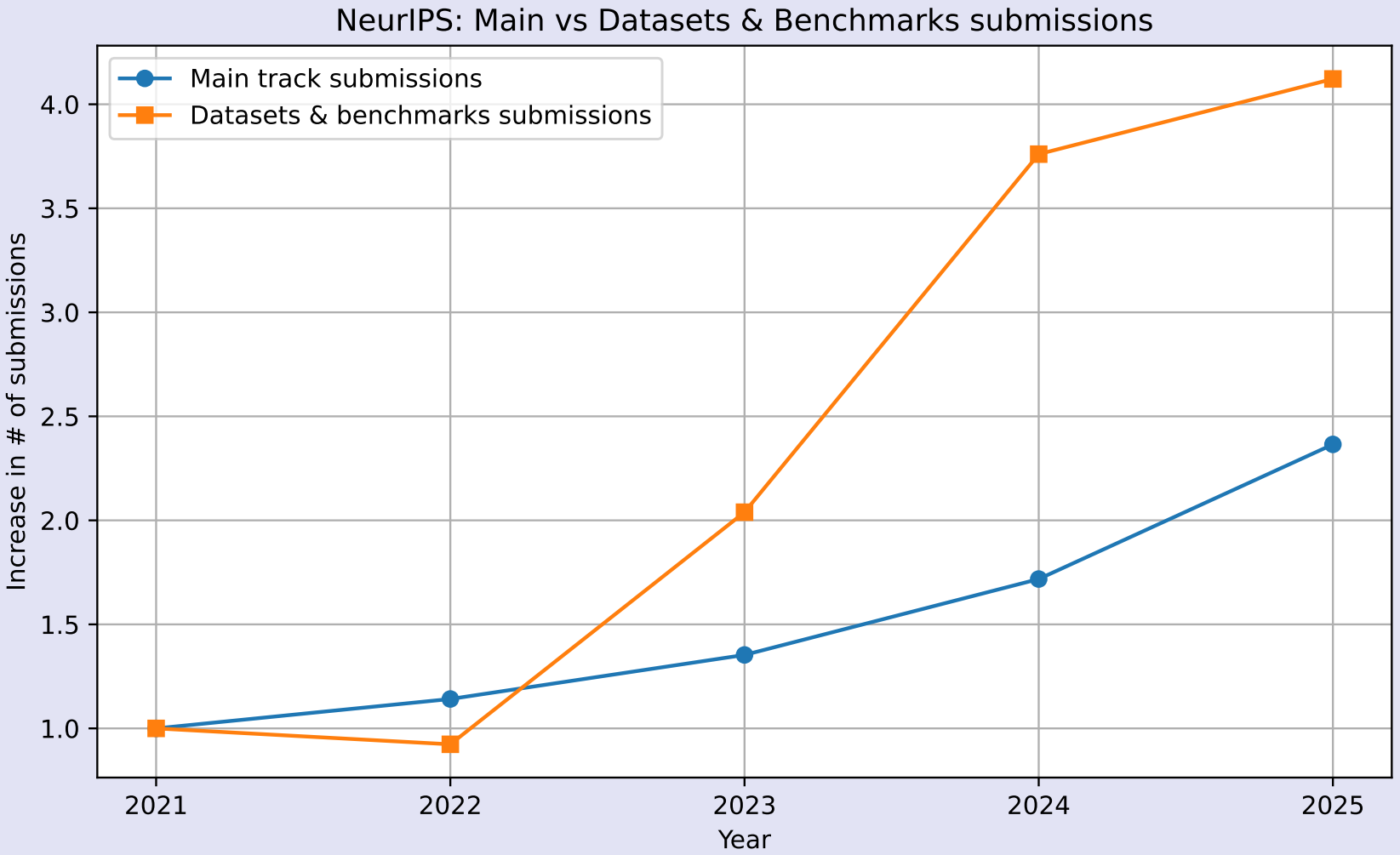
- **Objective**: what is being measured?
- **Dataset**: on what evidence?
- **Solvers/Methods**: What are we comparing?

Different benchmark goals emphasize different components.

- **Challenge benchmarks:** Fixed dataset + single metric $\rightarrow$ stresses solver comparison. Focus on performance
- **SOTA tracking benchmarks:** Multiple metrics and datasets, compare solvers to measure progress over time. Risk: test-set overfitting / cherry-picked metrics.
- **Research benchmarks:** Fixed set of methods evaluated, with broad range of metrics. Risk: incomplete / quickly outdated

Too many benchmarks in AI?

In the recent years, many benchmarks have been proposed:



$\Rightarrow$ Most of them don't have long-term maintainance plan.

Benchmark goals in AI

kaggle

	Short-term	**Long-term**
Task-specific	Challenge/Competition $\rightarrow$ push limits quickly	SOTA tracking $\rightarrow$ measure progress
Generalizable	Research question $\rightarrow$ empirical study	Field Benchmark $\rightarrow$ extensible

NEURAL INFORMATION PROCESSING SYSTEMS

Benchmark goals in AI

kaggle

	Short-term	**Long-term**
Task-specific	Challenge/Competition $\rightarrow$ push limits quickly	SOTA tracking $\rightarrow$ measure progress
Generalizable	Research question $\rightarrow$ empirical study	Field Benchmark $\rightarrow$ extensible

NEURAL INFORMATION PROCESSING SYSTEMS

Takeaway

Most attention goes to the top-left quadrant for fast progress, but solid science requires the bottom-right.

Benchmark goals in AI

	Short-term	**Long-term**
Task-specific	Challenge/Competition $\rightarrow$ push limits quickly	SOTA tracking $\rightarrow$ measure progress
Generalizable	Research question $\rightarrow$ empirical study	Field Benchmark $\rightarrow$ extensible

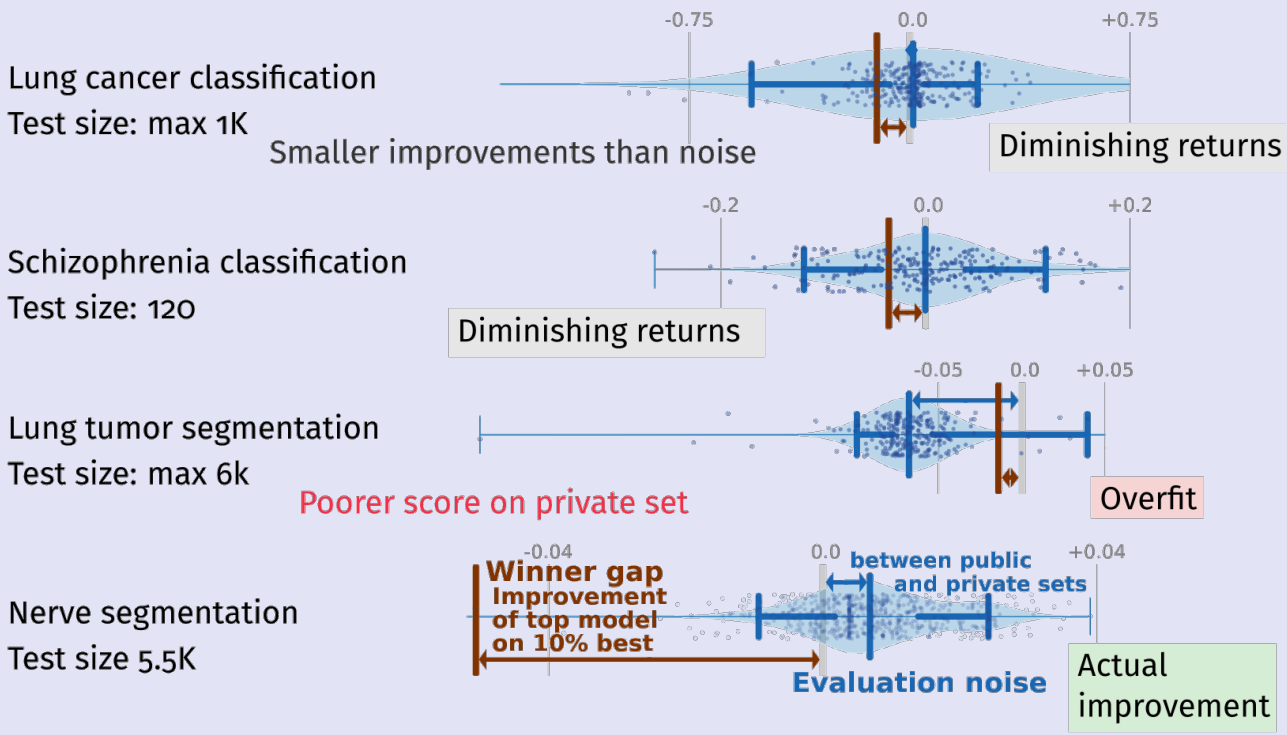
Takeaway

Most attention goes to the top-left quadrant for fast progress, but solid science requires the bottom-right.

Challenges in AI benchmarking

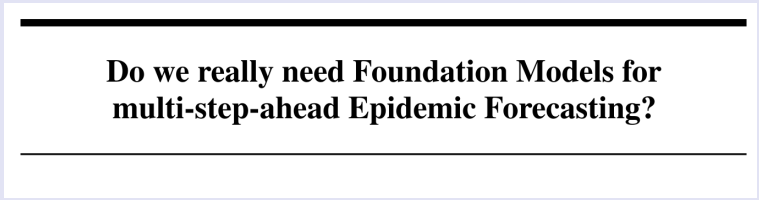
- Futile benchmarks

[Varoquaux and Cheplygina 2022]



Challenges in AI benchmarking

- Futile benchmarks
- Lack of proper baselines hinders scientific progress.


Do we really need Foundation Models for multi-step-ahead Epidemic Forecasting?

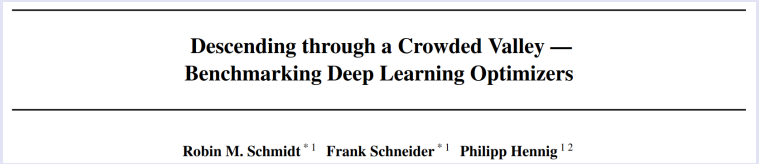

Position: Quo Vadis, Unsupervised Time Series Anomaly Detection?

M. Saquib Sarfraz [1 2] Mei-Yen Chen [1] Lukas Layer [1] Kunyu Peng [2] Marios Koulakis [2]


PNAS RESEARCH ARTICLE | COMPUTER SCIENCES OPEN ACCESS

Implicit data crimes: Machine learning bias arising from misuse of public data

Efrat Shimron [a,1], Jonathan I. Tamir [b,c,d], Ke Wang [a], and Michael Lustig [a]


Descending through a Crowded Valley — Benchmarking Deep Learning Optimizers

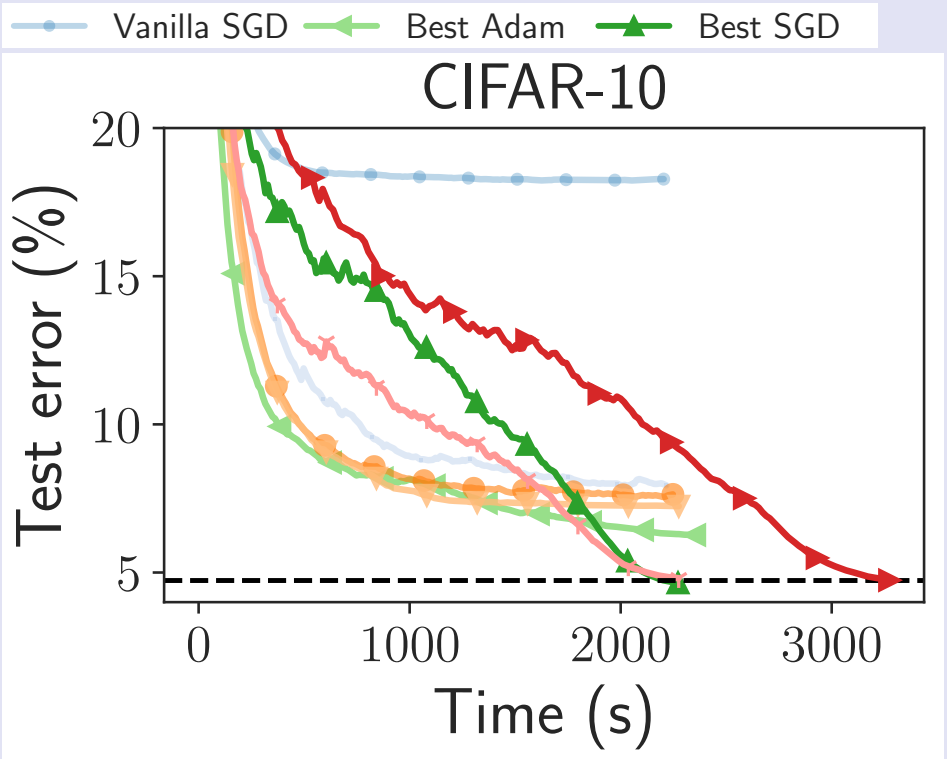
Robin M. Schmidt [* 1] Frank Schneider [* 1] Philipp Hennig [1 2]

Unclear improvement!

Challenges in AI benchmarking

- Futile benchmarks
- Lack of proper baselines hinders scientific progress.
- Reproducing benchmarks is hard.

[Moreau et al. 2022]



Challenges in AI benchmarking

- Futile benchmarks
- Lack of proper baselines hinders scientific progress.
- Reproducing benchmarks is hard.
- Statistical validity is often missing.

[Christodoulou et al. 2024]

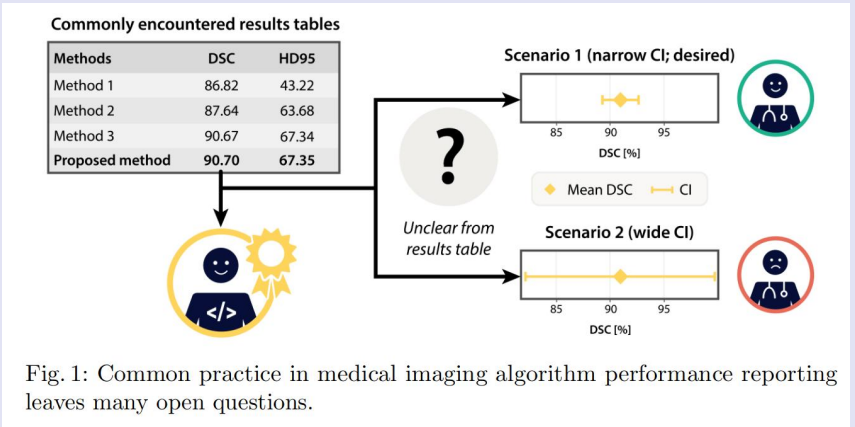
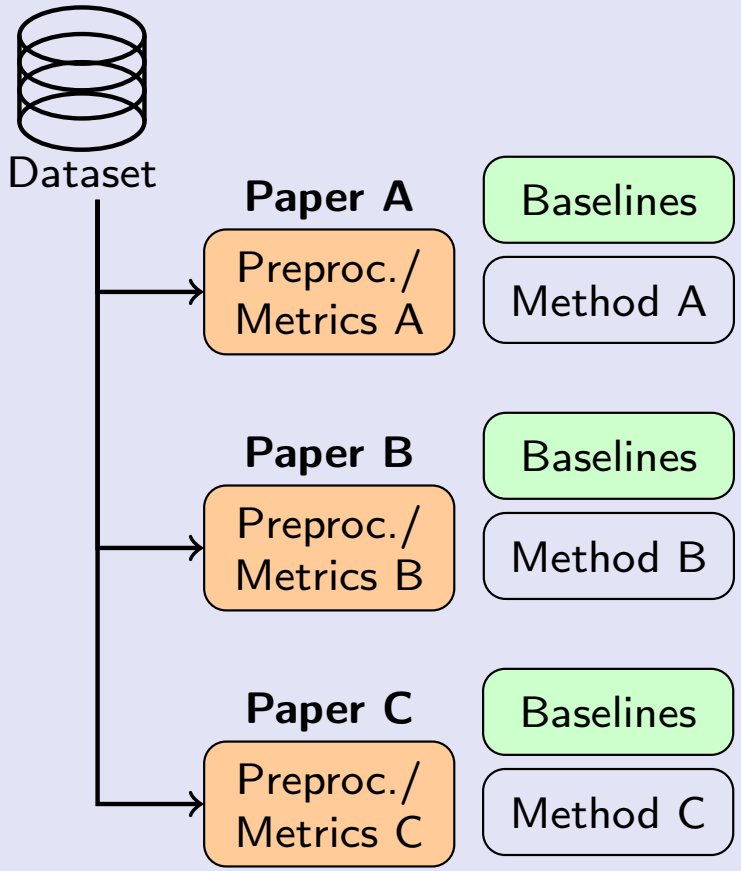


Fig. 1: Common practice in medical imaging algorithm performance reporting leaves many open questions.

Challenges in AI benchmarking

- Futile benchmarks
- Lack of proper baselines hinders scientific progress.
- Reproducing benchmarks is hard.
- Statistical validity is often missing.
- Benchmarking cost is duplicated across groups.

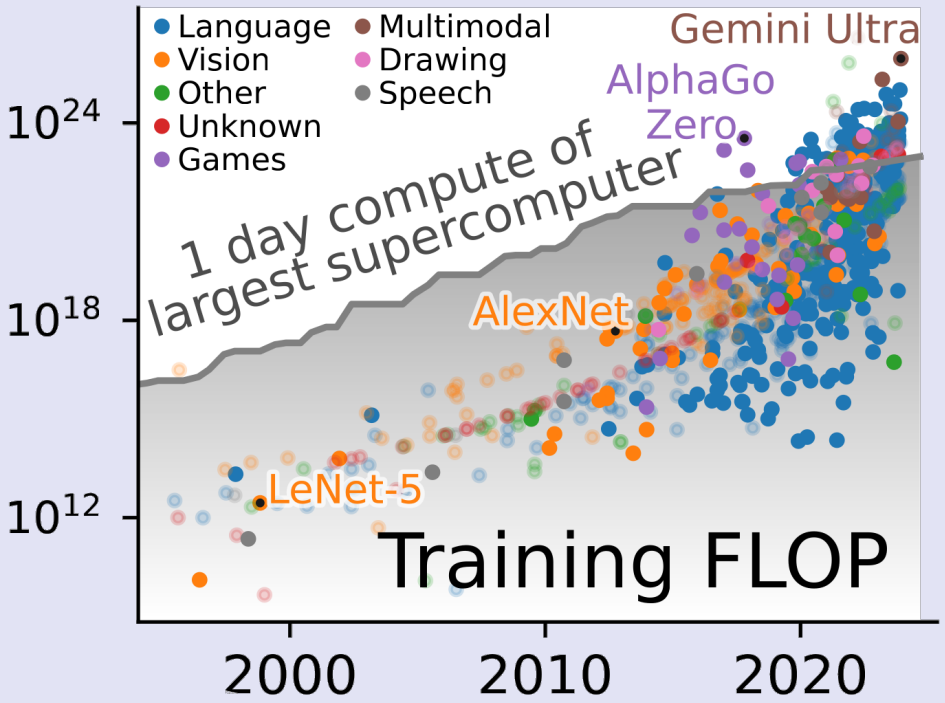


Each paper independently rebuilds preprocessing, baselines, and evaluation → duplicated cost.

Challenges in AI benchmarking

- Futile benchmarks
- Lack of proper baselines hinders scientific progress.
- Reproducing benchmarks is hard.
- Statistical validity is often missing.
- Benchmarking cost is duplicated across groups.
- Using AI infrastructure is complicated for many researchers.

[Varoquaux et al. 2025]



What AI research needs: shared evaluation infrastructure

Scientific fields mature when they build **shared infrastructure**:

- **Compute** – HPC clusters, cloud, SLURM, ...
- **Data** – HuggingFace, OpenML, UCI
- **Evaluation** – ???

What AI research needs: shared evaluation infrastructure

Scientific fields mature when they build **shared infrastructure**:

- **Compute** – HPC clusters, cloud, SLURM, ...
- **Data** – HuggingFace, OpenML, UCI
- **Evaluation** – ???

Evaluation is currently **rebuilt from scratch** for every paper: challenges are tackled individually

What AI research needs: shared evaluation infrastructure

Scientific fields mature when they build **shared infrastructure**:

- **Compute** – HPC clusters, cloud, SLURM, ...
- **Data** – HuggingFace, OpenML, UCI
- **Evaluation** – ???

Evaluation is currently **rebuilt from scratch** for every paper: challenges are tackled individually

The missing layer

A community-maintained, extensible, validated framework for

- defining evaluation protocols
- contributing datasets & baselines
- reproducing & extending results

⇒ This is what Benchopt provides.

Reproducible method comparison with Benchopt






. . .

References

Making runnable benchmarks with `benchopt`

`benchopt` provides a framework to organize and run benchmarks

Examples of existing benchmarks:

- **Image Classification (resnet)**
- **NanoGPT Optimization**
- **TSFMs evaluation**
- **Inverse problem resolution**
- **Unsup. Domain Adaptation**
- **Bilevel Optimization**
- **Brain Computer Interface**
- **...**

Benchopt scope and contributors

	Short-term	Long-term
Task-specific	Challenge	SOTA tracking
General.	Research Q.	**Field Benchmark**

⇒ Community-maintained, multi-institution, extensible benchmarks

Contributors from...

Structure of a benchmark

3 components: Objective, Datasets, Solvers

```
benchmark/
├── objective.py
├── datasets/
│   ├── dataset1.py
│   └── dataset2.py
└── solvers/
    ├── solver1.py
    └── solver2.py
```

Modular & extendable

- New metric? modify objective

Structure of a benchmark

3 components: Objective, Datasets, Solvers

```
benchmark/
├── objective.py
├── datasets/
│   ├── dataset1.py
│   └── dataset2.py
└── solvers/
    ├── solver1.py
    └── solver2.py
```

Modular & extendable

- New metric? modify objective
- New dataset? add a file

Structure of a benchmark

3 components: Objective, Datasets, Solvers

```
benchmark/
├── objective.py
├── datasets/
│   ├── dataset1.py
│   └── dataset2.py
└── solvers/
    ├── solver1.py
    └── solver2.py
```

Modular & extendable

- New metric? modify objective
- New dataset? add a file
- New solver? add a file

Structure of a benchmark

3 components: Objective, Datasets, Solvers

```
benchmark/
├── objective.py
├── datasets/
│   ├── dataset1.py
│   └── dataset2.py
└── solvers/
    ├── solver1.py
    └── solver2.py
```

Modular & extendable

- New metric? modify objective
- New dataset? add a file
- New solver? add a file

Structure of a benchmark

3 components: Objective, Datasets, Solvers

```
benchmark/
├── objective.py
├── datasets/
│   ├── dataset1.py
│   └── dataset2.py
└── solvers/
    ├── solver1.py
    └── solver2.py
```

Modular & extendable

- New metric? modify objective
- New dataset? add a file
- New solver? add a file

Template to create a new benchmark:

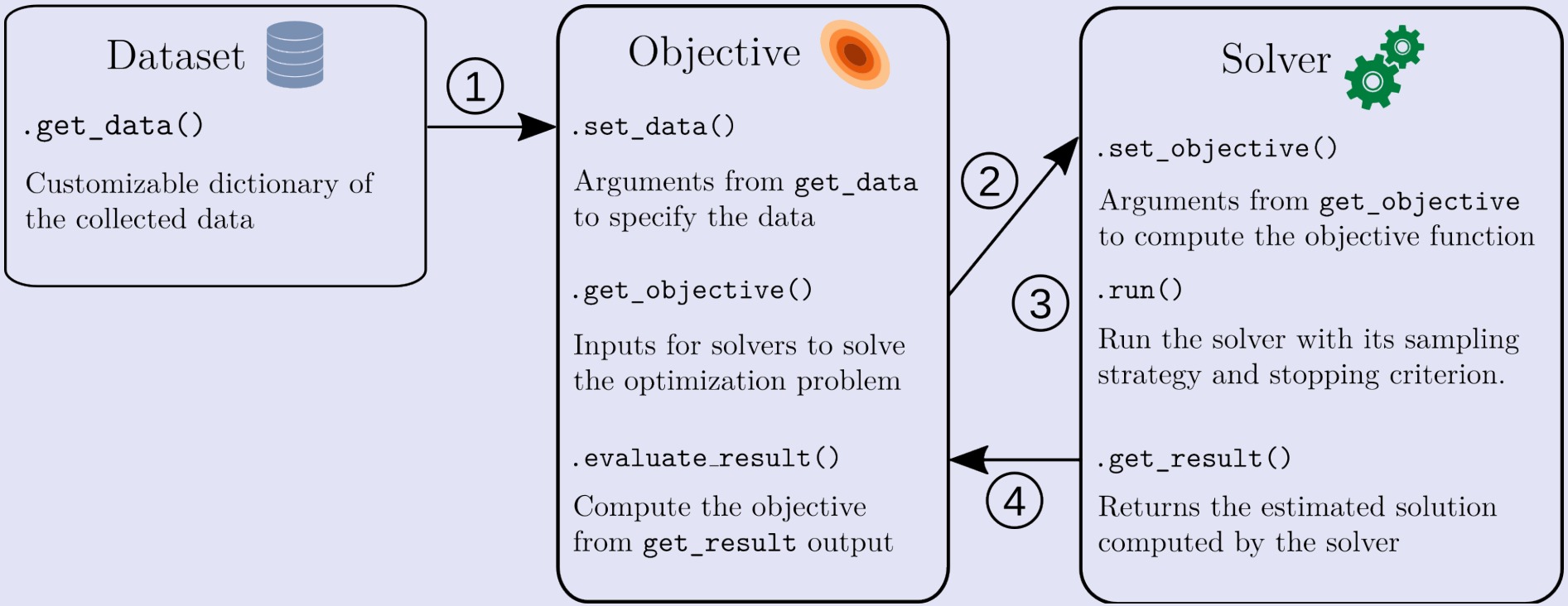
`https://github.com/benchopt/template_benchmark`

`https://github.com/benchopt/template_benchmark_ml`

A modular framework to create benchmarks

3 components: Objective, Dataset, Solver

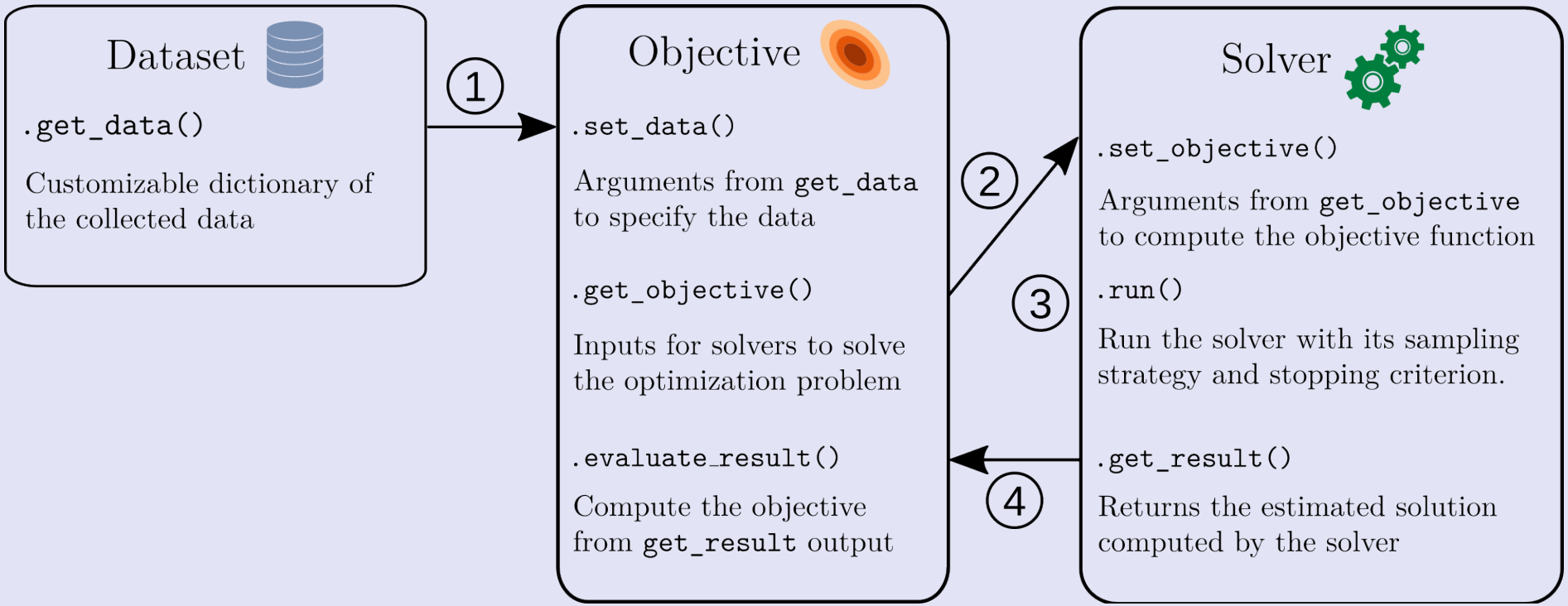
Dependency relation between Dataset - Objective - Solver



A modular framework to create benchmarks

3 components: Objective, Dataset, Solver

Dependency relation between Dataset - Objective - Solver



$\Rightarrow$ Benchopt defines the interface between components.

Creating a benchmark: the scientist's task

Define the 3 components for your problem:

Benchmark	**Dataset**	**Solver**	**Objective**
Lasso regression	(X, y)	sparse solver	suboptimality gap
DL optimization	dataloader + network	optimizer	val. accuracy
Brain–Computer Interface	EEG recordings + trial labels	decoder	classif. accuracy on 5-folds
Storage infra. evaluation	Large file dataset	dataloading strategy	throughput

Creating a benchmark: the scientist's task

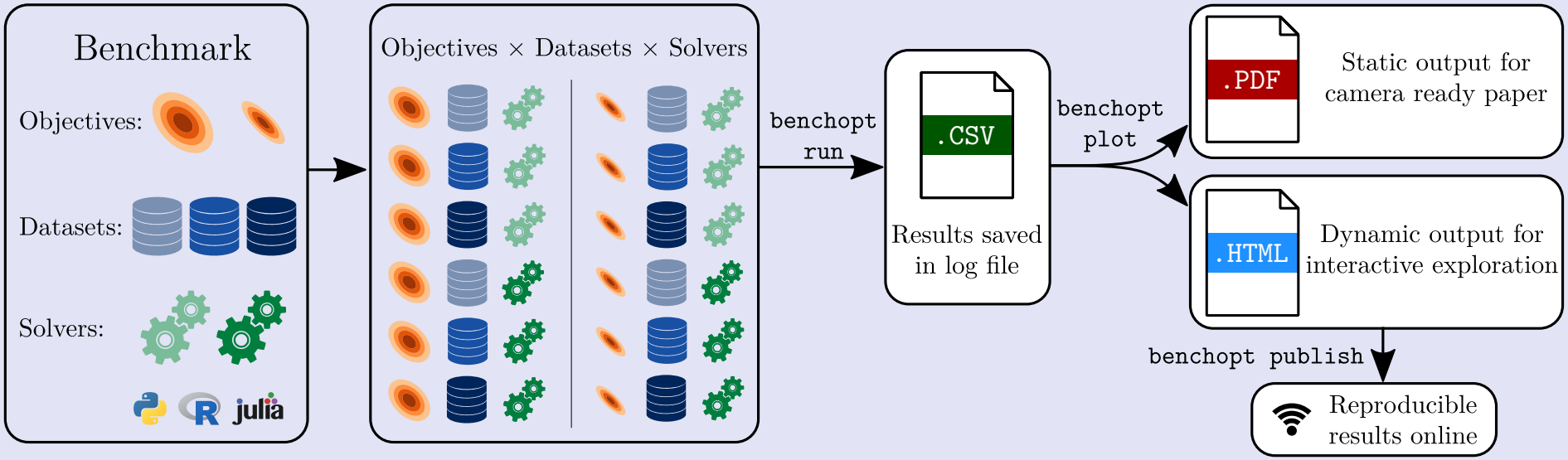
Define the 3 components for your problem:

Benchmark	**Dataset**	**Solver**	**Objective**
Lasso regression	(X, y)	sparse solver	suboptimality gap
DL optimization	dataloader + network	optimizer	val. accuracy
Brain–Computer Interface	EEG recordings + trial labels	decoder	classif. accuracy on 5-folds
Storage infra. evaluation	Large file dataset	dataloading strategy	throughput

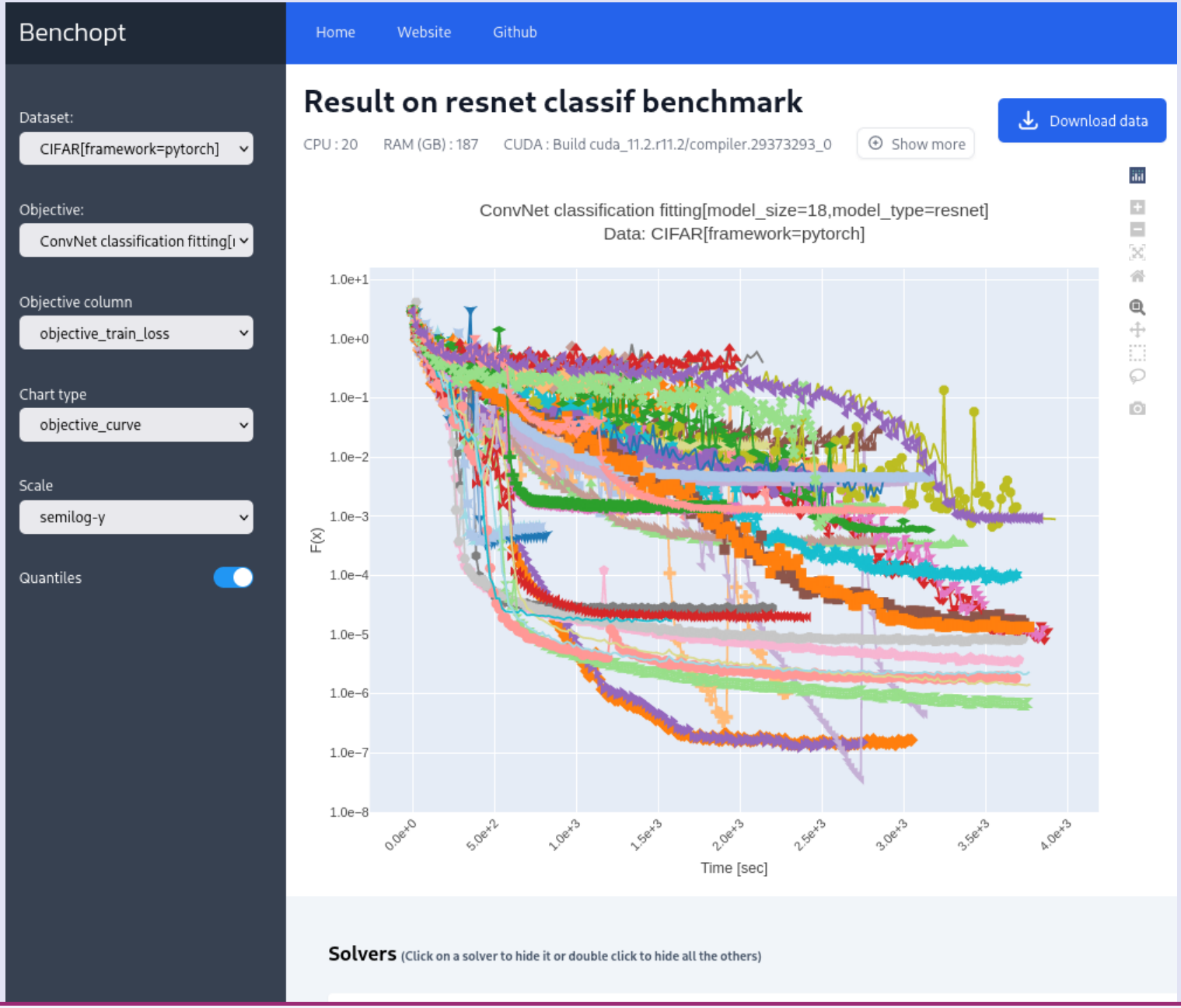
$\Rightarrow$ The benchmark structure is the same: only the components change.

Benchmark workflow

Steps: Install/Prepare, Test, Run, Explore, Publish

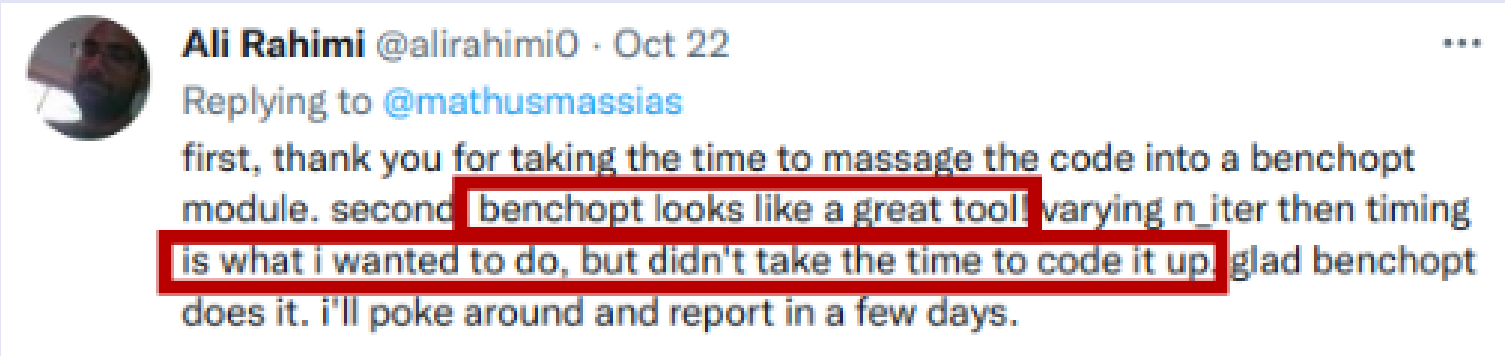


Interactive results exploration



Benchopt makes your life easy

- **Integrate broadly:** use implementations from Python, R, Julia, or binaries.
- **Scale experiments:** run in parallel locally or on HPC clusters, natively integrate with SLURM, ...
- **Save time:** cache results to avoid recomputing unchanged runs.
- **Trust comparisons:** control randomness with seeds and stable protocols.
- **Share outcomes:** merge and publish results from multiple runs, easy interactive visualization.

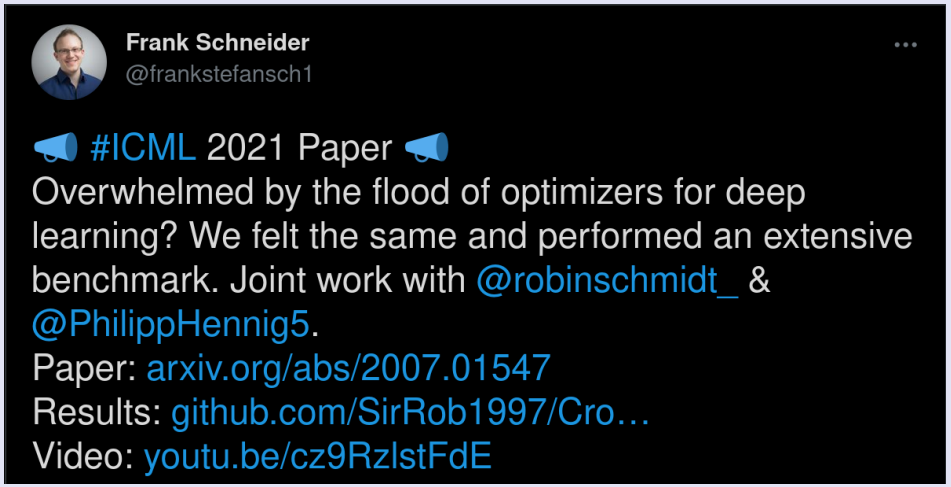


Typical case: deep learning optimization

A. List of optimizers and schedules considered

Table 2: List of optimizers considered for our benchmark. This is only a subset of all existing methods for deep learning.

Name	Ref.	Name	Ref.
AcceleGrad	(Levy et al., 2018)	HyperAdam	(Wang et al., 2019b)
ACClip	(Zhang et al., 2020)	K-BFGS/K-BFGS(L)	(Goldfarb et al., 2020)
AdaAlter	(Xie et al., 2019)	KF-QN-CNN	(Ren & Goldfarb, 2021)
AdaBatch	(Devarakonda et al., 2017)	KFAC	(Martens & Grosse, 2015)
AdaBayes/AdaBayes-SS	(Aitchison, 2020)	KFLR/KFRA	(Botev et al., 2017)
AdaBelief	(Zhuang et al., 2020)	L4Adam/L4Momentum	(Rolínek & Martius, 2018)
AdaBlock	(Yun et al., 2019)	LAMB	(You et al., 2020)
AdaBound	(Luo et al., 2019)	LaProp	(Ziyin et al., 2020)
AdaComp	(Chen et al., 2018)	LARS	(You et al., 2017)
Adadelta	(Zeiler, 2012)	LHOPT	(Almeida et al., 2021)
Adafactor	(Shazeer & Stern, 2018)	LookAhead	(Zhang et al., 2019)
AdaFix	(Bae et al., 2019)	M-SVAG	(Balles & Hennig, 2018)
AdaFom	(Chen et al., 2019a)	MADGRAD	(Defazio & Jelassi, 2021)
AdaFTRL	(Orabona & Pál, 2015)	MAS	(Landro et al., 2020)
Adagrad	(Duchi et al., 2011)	MEKA	(Chen et al., 2020b)
ADAHESSIAN	(Yao et al., 2020)	MTAdam	(Malkiel & Wolf, 2020)
Adai	(Xie et al., 2020)	MVRC-1/MVRC-2	(Chen & Zhou, 2020)
AdaLoss	(Teixeira et al., 2019)	Nadam	(Dozat, 2016)
Adam	(Kingma & Ba, 2015)	NAMSB/NAMSG	(Chen et al., 2019b)
Adam$^+$	(Liu et al., 2020b)	ND-Adam	(Zhang et al., 2017a)
AdamAL	(Tao et al., 2019)	Nero	(Liu et al., 2021b)
AdaMax	(Kingma & Ba, 2015)	Nesterov	(Nesterov, 1983)
AdamBS	(Liu et al., 2020c)	Noisy Adam/Noisy K-FAC	(Zhang et al., 2018)
AdamNC	(Reddi et al., 2018)	NosAdam	(Huang et al., 2019)
AdaMod	(Ding et al., 2019)	Novograd	(Ginsburg et al., 2019)
AdamPSGDP	(Heo et al., 2021)	NT-SGD	(Zhou et al., 2021b)
AdamT	(Zhou et al., 2020)	Padam	(Chen et al., 2020a)
AdamW	(Loshchilov & Hutter, 2019)	PAGE	(Li et al., 2020b)
AdamX	(Tran & Phong, 2019)	PAL	(Mutschler & Zell, 2020)
ADAS	(Eliyahu, 2020)	PolyAdam	(Orvieto et al., 2019)
AdaS	(Hosseini & Plataniotis, 2020)	Polyak	(Polyak, 1964)
AdaScale	(Johnson et al., 2020)	PowerSGD/PowerSGDM	(Vogels et al., 2019)
AdaSGD	(Wang & Wiens, 2020)	Probabilistic Polyak	(de Roos et al., 2021)
AdaShift	(Zhou et al., 2019)	ProbLS	(Mahsereci & Hennig, 2017)
AdaSqrt	(Hu et al., 2019)	PStorm	(Xu, 2020)
Adathm	(Sun et al., 2019)	QHAdam/QHM	(Ma & Yarats, 2019)
AdaX/AdaX-W	(Li et al., 2020a)	RAdam	(Liu et al., 2020a)
AEGD	(Liu & Tian, 2020)	Ranger	(Wright, 2020b)
ALI-G	(Berrada et al., 2020)	RangerLars	(Grankin, 2020)
AMSBound	(Luo et al., 2019)	RMSProp	(Tieleman & Hinton, 2012)
AMSGrad	(Reddi et al., 2018)	RMSterov	(Choi et al., 2019)
AngularGrad	(Roy et al., 2021)	S-SGD	(Sung et al., 2020)
ArmijoLS	(Vaswani et al., 2019)	SAdam	(Wang et al., 2020b)
ARSG	(Chen et al., 2019b)	Sadam/SAMSGrad	(Tong et al., 2019)
ASAM	(Kwon et al., 2021)	SALR	(Yue et al., 2020)
AutoLRS	(Jin et al., 2021)	SAM	(Foret et al., 2021)
AvaGrad	(Savarese et al., 2019)	SC-Adagrad/SC-RMSProp	(Mukkamala & Hein, 2017)
BAdam	(Salas et al., 2018)	SDProp	(Ida et al., 2017)
BGAdam	(Bai & Zhang, 2019)	SGD	(Robbins & Monro, 1951)
BPGrad	(Zhang et al., 2017b)	SGD-BB	(Tan et al., 2016)
BRMSProp	(Aitchison, 2020)	SGD-G2	(Ayadi & Turinici, 2020)
BSGD	(Hu et al., 2020)	SGDEM	(Ramezani-Kebrya et al., 2021)
C-ADAM	(Tutunov et al., 2020)	SGDHess	(Tran & Cutkosky, 2021)
CADA	(Chen et al., 2021)	SGDM	(Liu & Luo, 2020)
Cool Momentum	(Borysenko & Byshkin, 2020)	SGDR	(Loshchilov & Hutter, 2017)
CProp	(Preechakul & Kijsirikul, 2019)	SHAdagrad	(Huang et al., 2020)
Curveball	(Henriques et al., 2019)	Shampoo	(Anil et al., 2020; Gupta et al., 2018)
Dadam	(Nazari et al., 2019)	SignAdam++	(Wang et al., 2019a)
DeepMemory	(Wright, 2020a)	SignSGD	(Bernstein et al., 2018)
DGNOpt	(Liu et al., 2021a)	SKQN/S4QN	(Yang et al., 2020)
DiffGrad	(Dubey et al., 2020)	SM3	(Anil et al., 2019)
EAdam	(Yuan & Gao, 2020)	SMG	(Tran et al., 2020)



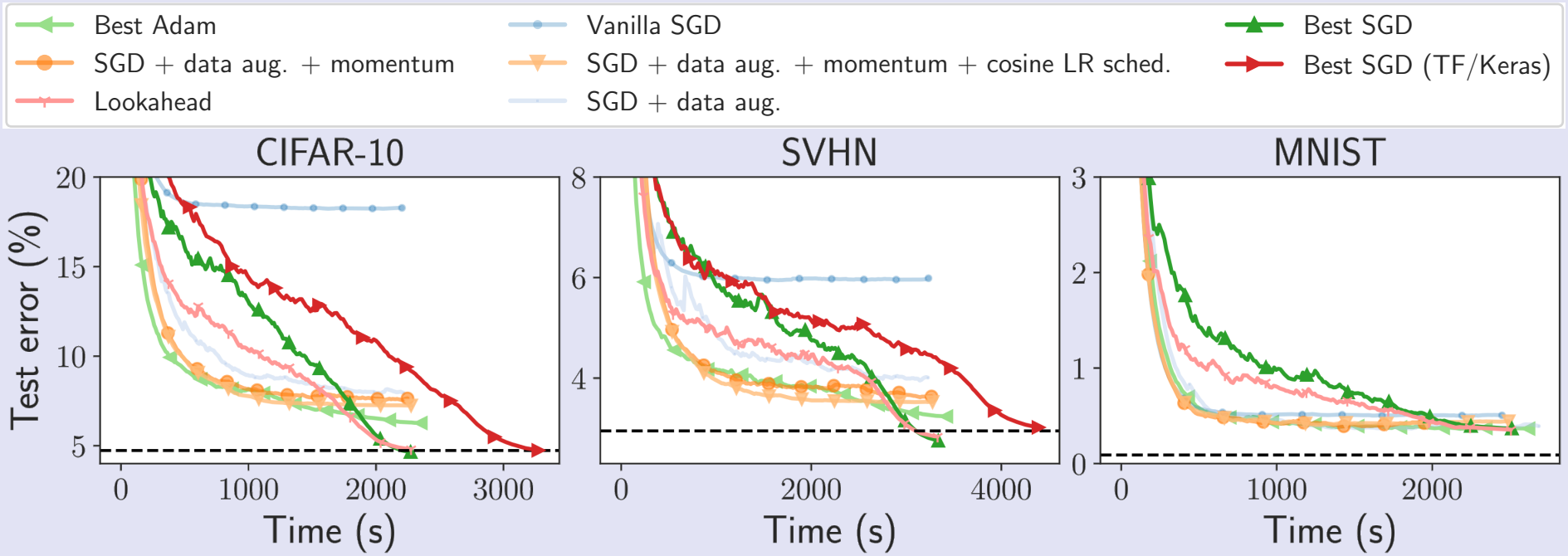
$\Rightarrow$ Many novel methods but unclear improvements.

$\Rightarrow$ But this benchmark cannot be easily reproduced!

Example: Optimization for ResNet on image classification

- Image classification with resnet18
- Various optimization strategies
- Compare `pytorch` and `tensorflow`
- Publish reproducible SOTA for baselines

Z. Ramzi



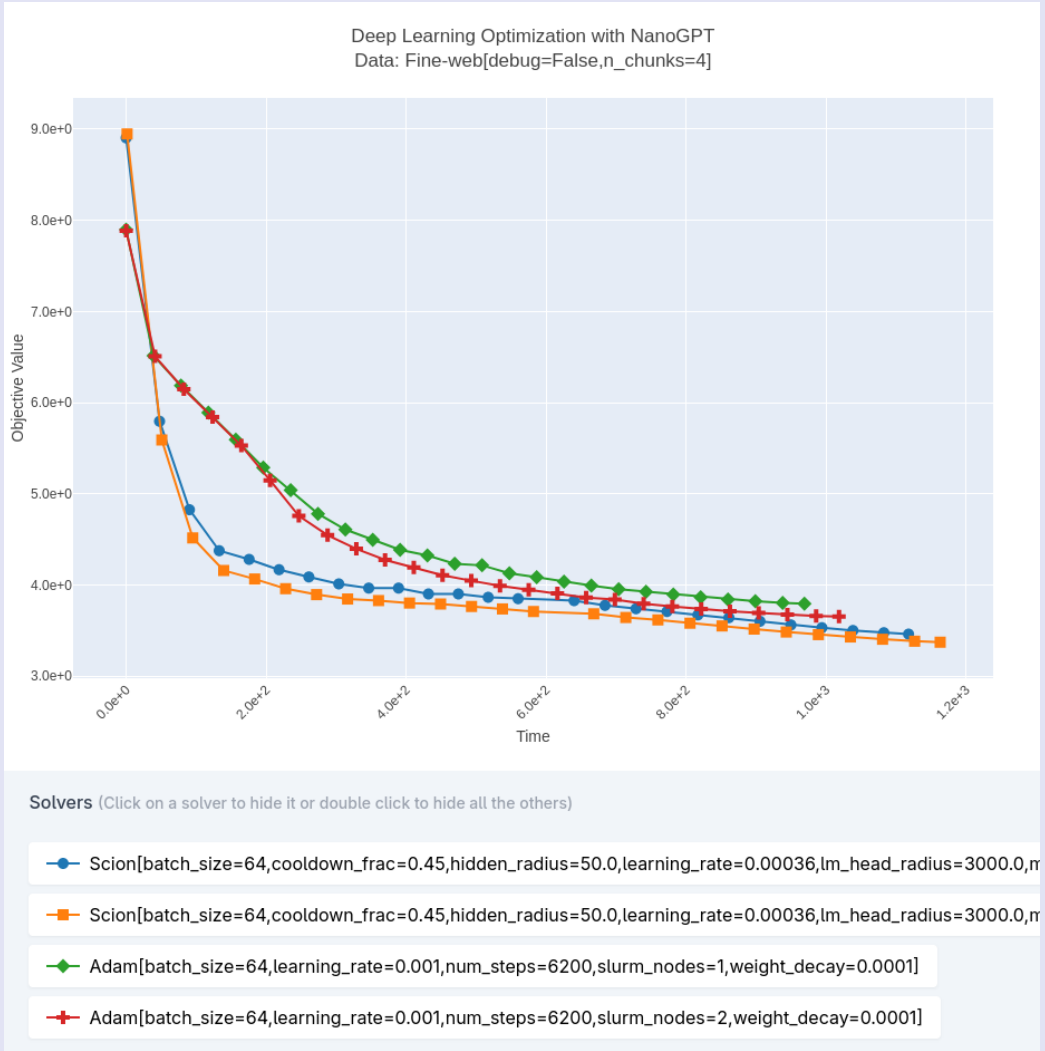
`https://github.com/benchopt/benchmark_resnet_classif/`

Example: Large scale-optimization for Deep learning

- Use modern large-scale datasets and models
- Classical optimization (schedulers, momemtum, . . .)
- Distributed training and mixed precision

S. Vaiter

T. Silveti-Falls



`https://github.com/benchopt/benchmark_nanogpt/`

Example: Inverse problem resolution

- Provide a series of tasks: one dataset + one physics
- An easy interface to evaluate reconstruction models
- Target to auto-run and upload on huggingface hub

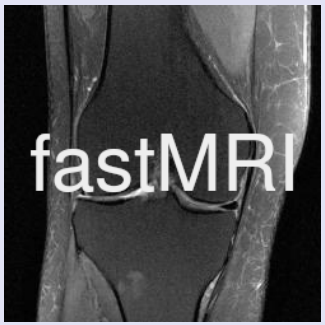
```python
from dbench import run_benchmark
import deepinv as dinv

# Create reconstruction method
my_solver = dinv.models.RAM()
run_benchmark(
    my_solver, "benchmark_name"
)
```

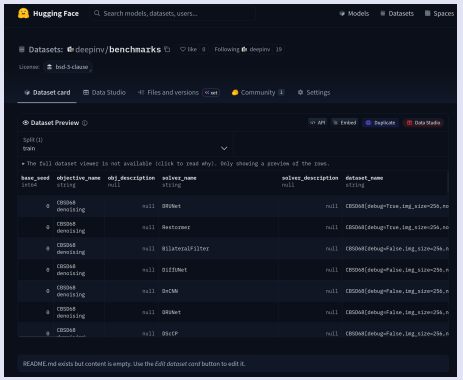
J. Tachella

Deep Inverse

Dataset example



HF hub leaderboard



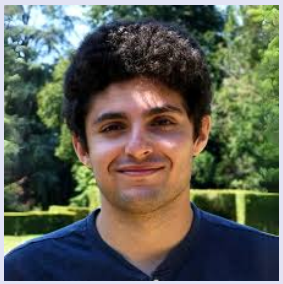
https://github.com/deepinv/benchmarks

Example: Large-scale Inverse problem resolution

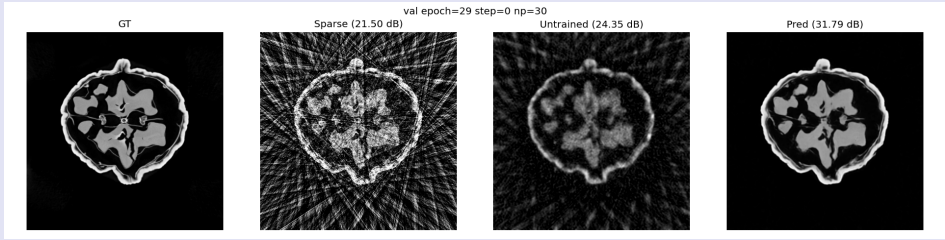
- A few very large-scale tasks: radio-astronomy, CT, ...
- Inference, Training and workflows
- Target measuring distributed performances

Inference **Training**

Workflow



B. Malezieux



`https://github.com/bmalezieux/benchmark_invprob_largescale`

An overlooked aspect: longer term maintenance

Many research results are not maintained after publication:

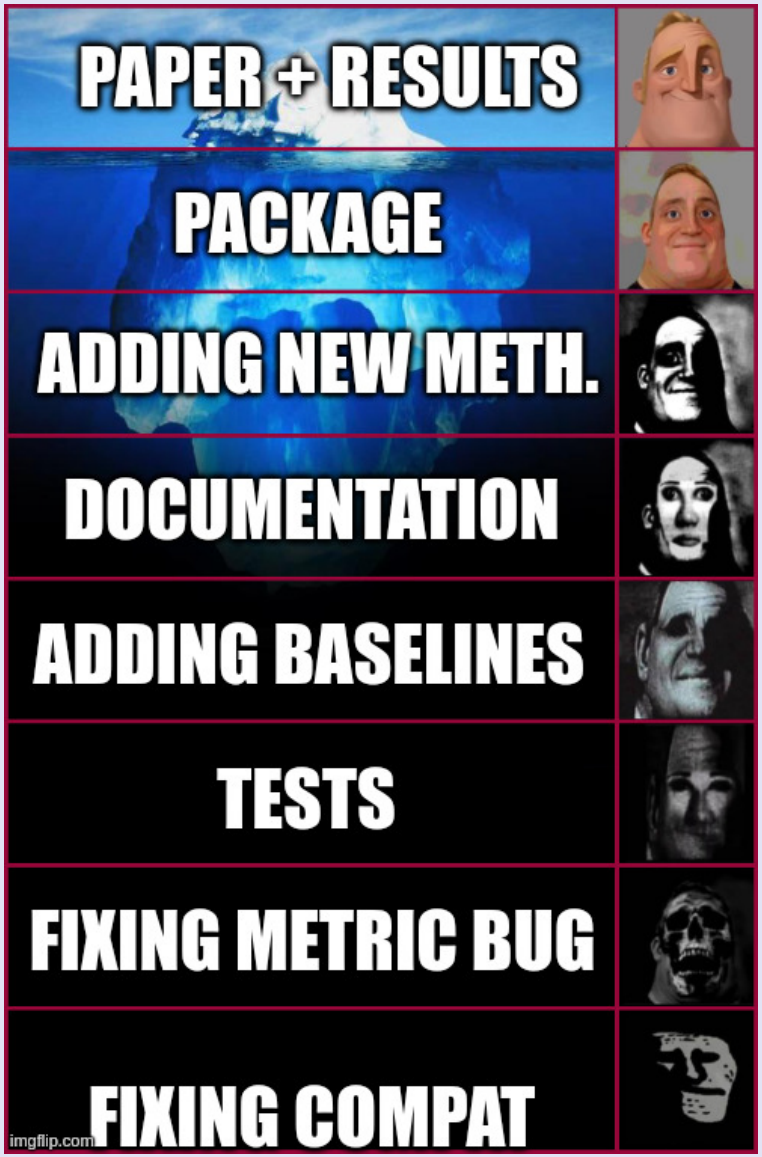
- Every PhD creates a package;
- Every post-doc abandons one;
- The ecosystem grows horizontally, not vertically

An overlooked aspect: longer term maintenance

Many research results are not maintained after publication:

- Every PhD creates a package;
- Every post-doc abandons one;
- The ecosystem grows horizontally, not vertically

$\Rightarrow$ Limited incentives in AI to maintain a codebase



The benchopt roadmap

Grow the benchmark collection

- Zero-th order optimization
- Benchmarking dataloading strategies on HPC infrastructures
- . . . and more community contributions

⇒ Open question: how to benchmark foundation models?

Improve benchmarking methodology

- **Statistical validity:** how many samples / CV splits to trust a ranking?
  → work w/ C. Eve and G. Varoquaux
- **Foundation model evaluation:** few-shot, prompted, fine-tuned – incomparable by standard metrics.
- **Frugality:** benchmark under compute constraints, not just final accuracy.

Thank you for your attention!

Three asks:

1. **Contribute** a solver or dataset to an existing benchmark.
2. **Consider Benchopt** as infrastructure for you next benchmark, in a paper or with broader scope.
3. **Star the repo** – it matters for visibility!

Want to run a benchmark sprint?
Come talk to me!