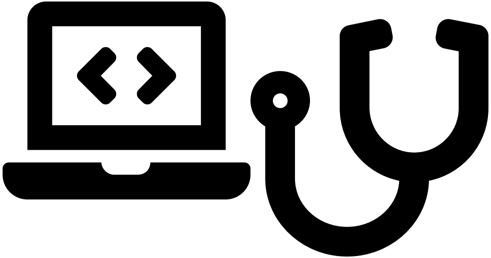
Model evaluation, a machine-learning bottleneck

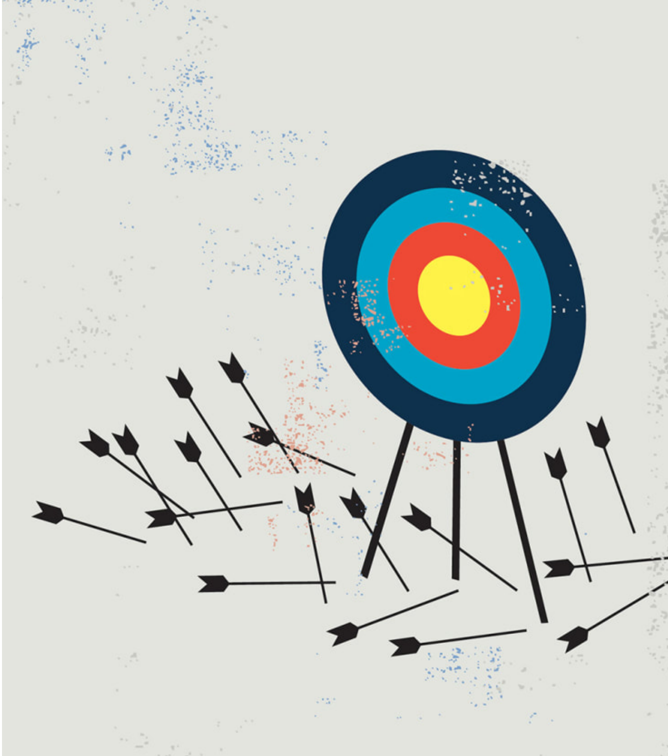
Gaël Varoquaux
Inria

See also [Varoquaux and Colliot 2022]



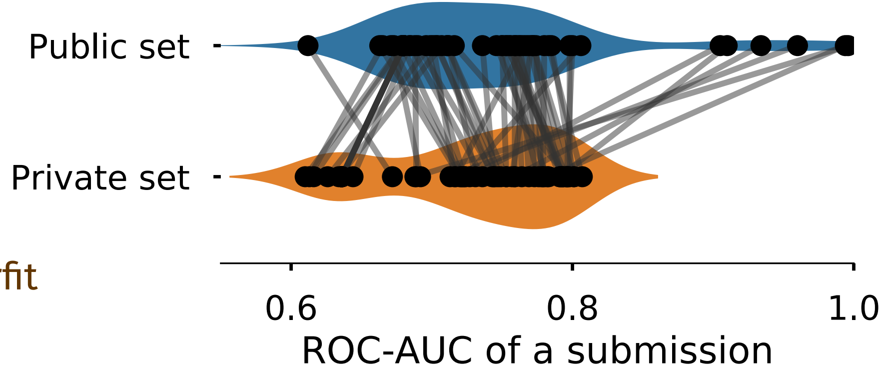
Model evaluation is the Achilles heel of machine learning

Machine learning
has become
an empirical science



Prediction challenge: Autism status

- 10 000 € incentives



Analysts overfit the public set

- Best performer: linear models on graph features
- Graph neural networks performed poorly

Machine learning in medical imaging [Varoquaux and Cheplygina 2022]

Kaggle competitions

Lung cancer classification

Test size: max 1K

Schizophrenia classification

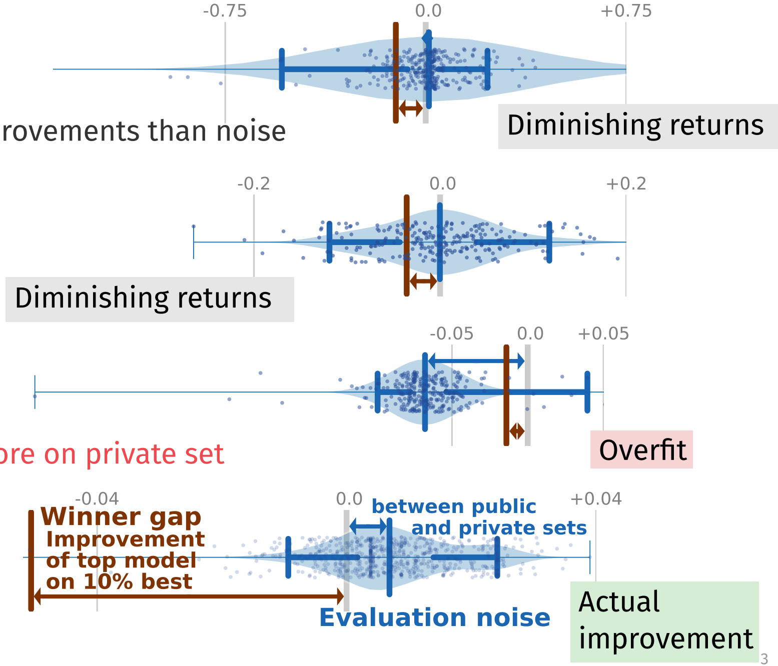
Test size: 120

Lung tumor segmentation

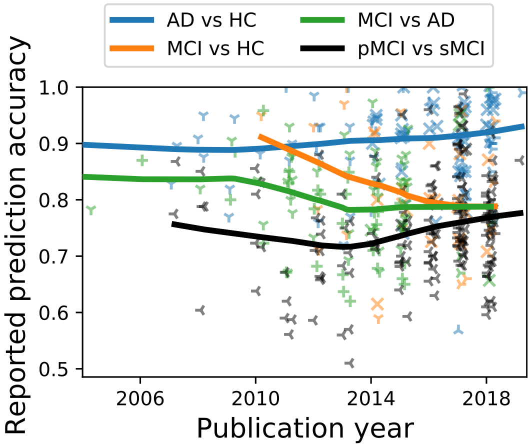
Test size: max 6k

Nerve segmentation

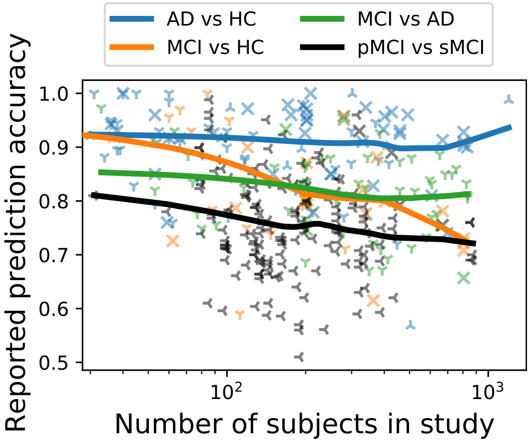
Test size 5.5K



Little progress: publications on Alzeihmer's disease diagnostic



Over time



Poorer performance on larger (more real-life) cohorts

Beyond the performance number

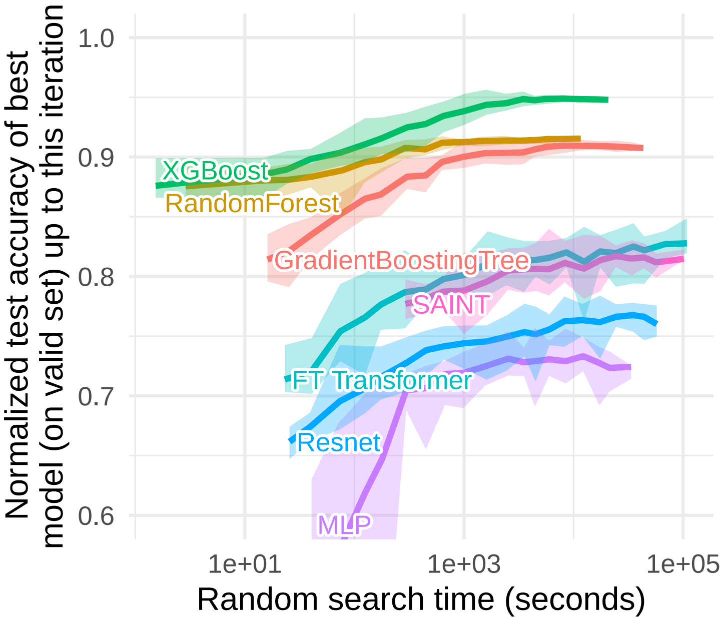
- Useless predictions using doctor’s marks
- Training on automated labels extracted with bias

Models bring no value to the clinic

Deep learning on tabular data

Promising publications in serious venues & labs

But classic tree-based methods perform best



More *valid* benchmarks

Reflect and capture
- the application setting
- the generalization error

This talk:

1 Meaningful classification metrics

2 Quantifying generalization error

1 Meaningful classification metrics

Metrics must capture and reflect application

Going further for images analysis [Maier-Hein... 2022]

Meaningful metrics in imbalanced settings

Accuracy, balanced accuracy?

Accuracy
uninformative under class imbalance

90% of class 0
$\Rightarrow$ predicting only class 0 gives Acc=90%

Truth:

Predicted:	$T+$	$T-$
$P+$	**TP**	**FP**
$P-$	**FN**	**TN**

Balanced accuracy: errors on class 0 and class 1

- Sensitivity (also called recall): fraction of class 1 retrieved. $\frac{\text{TP}}{\text{TP + FN}}$
- Specificity: fraction of class 0 actually classified as 0. $\frac{\text{TN}}{\text{TN + FP}}$
- Balanced accuracy: $\frac{1}{2}$ (sensitivity + specificity)

Sensitivity: $\mathbb{P}(P{+}|T{+})$

Specificity: $\mathbb{P}(P-|T-)$

Asking the right question: $\mathbb{P}(P+|T+)$ vs $\mathbb{P}(T+|P+)$

Positive predictive value (via Bayes' theorem):

Truth

$$\mathbb{P}(T+ \mid P+) = \frac{\text{sensitivity} \times \text{prevalence}}{(1 - \text{specificity}) \times (1 - \text{prevalence}) + \text{sensitivity} \times \text{prevalence}}.$$

Predictive positive

Summary metric: Markedness: PPV + NPV - 1

Drawback: depends on prevalence

$\Rightarrow$ Characterizes not only the classifier, but also the dataset

Odds ratios for invariance to sampling [Varoquaux and Colliot 2022]

Definition: Odds of a $$\mathbb{O}(a) = \frac{\mathbb{P}(a)}{1 - \mathbb{P}(a)}$$

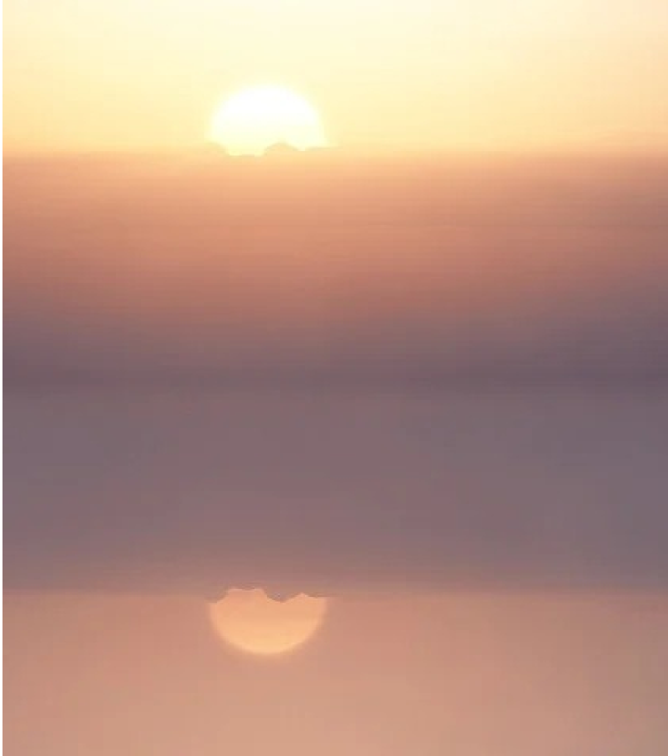
Likelihood ratio of positive class:

$$\text{LR+} = \frac{\mathbb{O}(T{+}|P{+})}{\mathbb{O}(T{+})} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

- Independent of class prevalence
- Use prevalence on target population to compute $\mathbb{O}(T{+})$

Useful to extrapolate across test-sets of different prevalence

Confidence score and calibration



Interpreting classifier score as a probability? – Calibration

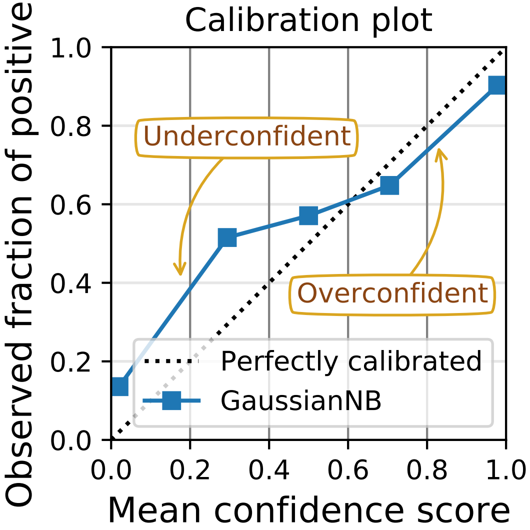
Calibration

Average error rate for all samples with score s is s

Computed in bins on score s

ECE:
expected calibration error

Average error on bins of score s

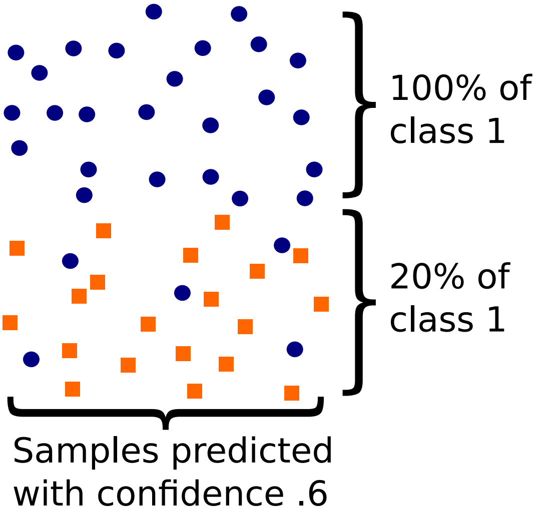


⚠ Average error

Calibration is not enough

[Perez-Lebel... 2022]

Average error rate for all samples with score s is s



A calibrated classifier can assign **a score of .6** to individuals, but be **100% accurate on a subgroup**, and **20%** on another.

⚠ Calibration does not control individual probabilities

Metrics controlling individual probabilities

Does the classifier approach $\mathbb{P}(y|X)$?

Proper scoring rules

$$\text{Brier score} = \sum_i (\underbrace{\hat{s}_i}_{\text{Confidence score}} - \overbrace{y_i}^{\text{Observed (binary) label}})^2$$

(also log-loss)

Minimal for $\hat{s} = P(y|X)$

Drawbacks

- cannot be interpreted as an error rate
- no scale

Decomposing scoring rules into error rates [Perez-Lebel... 2022]

- Classifier output: $S = f(X)$
- Label probabilities: $Q = \mathbb{P}[Y|X]$
- Calibrated score[1]: $C = \mathbb{E}\Big[\mathbb{P}[Y|X] \Big| S\Big]$

1 Knowing the classifier output, what's the label probabilities

Scoring rule decomposition

Expected label

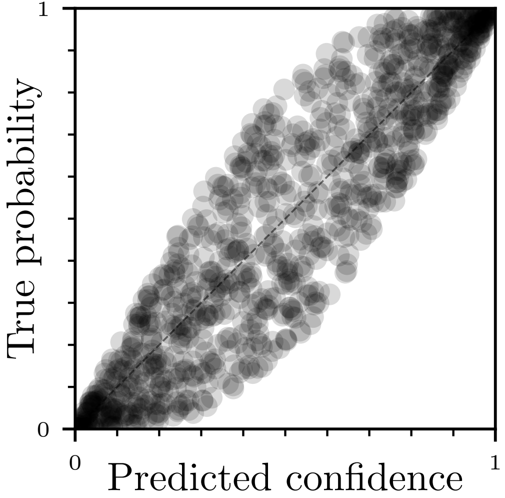
Calibrated score

$$\mathbb{E}\left[d(S,Y)\right] = \underbrace{\mathbb{E}\left[d(S,C)\right]}_{\text{Calibration error}} + \underbrace{\mathbb{E}\left[d(C,Q)\right]}_{\text{Grouping error}} + \underbrace{\mathbb{E}\left[d(Q,Y)\right]}_{\text{Irreducible error}}$$

Classifier output

Label distribution

An oracle calibration plot



No calibration error

$\frac{\text{On average}}{\text{predicted confidence}}$ = true probability

Grouping error

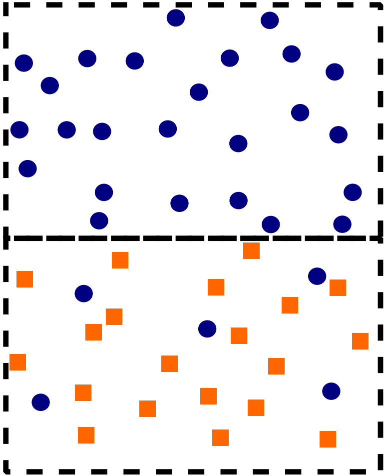
Classifier over-confident on some samples, under-confident on others

Measures the *dispersion* of scores

Requires access to true probabilities 😬

Estimating the grouping loss

[Perez-Lebel... 2022]



$P(Y{=}1|X) = 1$

$P(Y{=}1|X) = .2$

Estimating true probabilities on well-chosen bins

(and controlling errors due to binning)

Meaningful classification metrics

- Machine learning research chases metrics

  These should reflect application as well as possible

- Think in terms of $\mathbb{P}(T+\,|P+)$
  - Accuracy reasonable proxy only for balanced classes
  - LR+ interesting to keep in mind

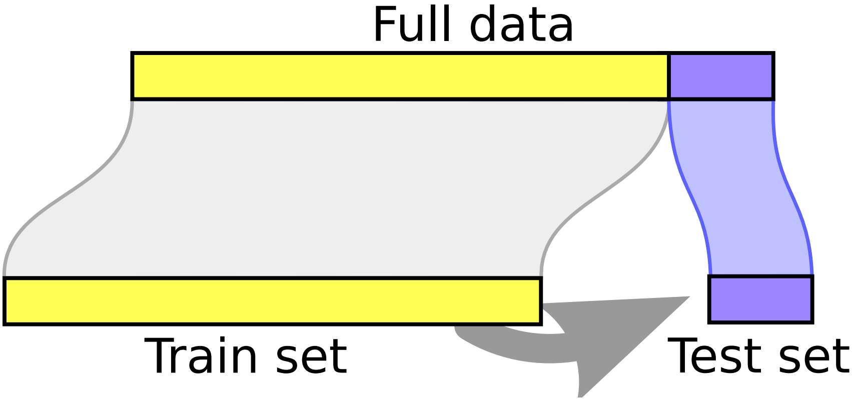
- Think in terms of uncertainty
  - Calibration quantifies average errors
  - Grouping loss: error on individual uncertainty

- A single number does not tell the whole story

2 Quantifying generalization error

Corresponding research paper: [Bouthillier... 2021]

How we do model evaluation



Definitions: what are we benchmarking?

Senario 1: *a prediction rule*:

We are given $f : \mathcal{X} \rightarrow \mathcal{Y}$

Senario 2: *a training procedure*:

We are given: a procedure that outputs a prediction rule $\hat{f}$ from training data $(\mathbf{X}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$

Definitions: what are we benchmarking?

Senario 1: *a prediction rule*:

We are given $f : \mathcal{X} \rightarrow \mathcal{Y}$

For application claims: *eg* medicine

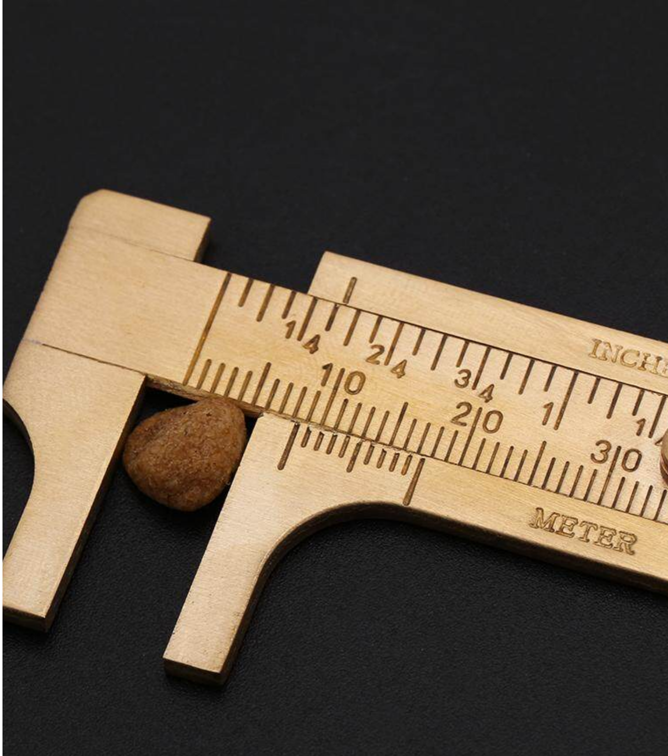
Senario 2: *a training procedure*:

We are given: a procedure that outputs a prediction rule $\hat{f}$ from training data $(\mathbf{X}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$

For machine-learning research (claims on algorithms)

1

Benchmarking a prediction rule



External validation

We are given $f : \mathcal{X} \rightarrow \mathcal{Y}$

X_{test} different enough from X_{train}

- No repeated acquisition of same individual in train & test

[Little... 2017]

- Ideally: show generalization to new site, later in time...

X_{test} representative of target population

Sample X_{test}:

- To match statistical moments
- To minimize a confounding association (shortcuts)

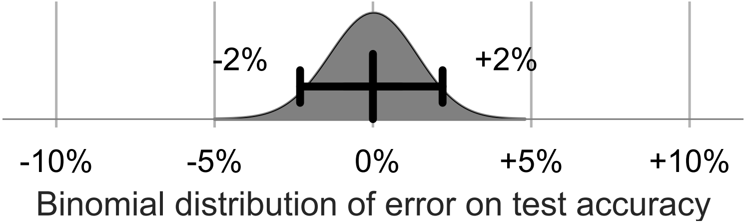
[Chyzhyk... 2018]

Evaluation error: Sampling noise on test set

Evaluation quality is limited by number of test examples

[Varoquaux 2018]

Sampling noise[1] for $n_{\text{test}} = 1000$:



The data at hand (*eg* the test set) is just a small sample of the full population "in the wild", and sampling other data will lead to other results.

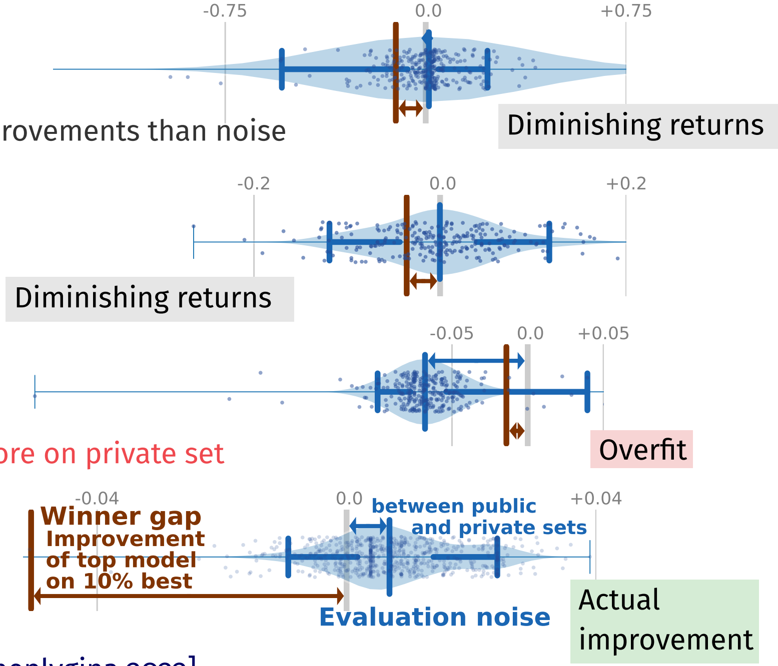
Evaluation noise is not negligible – in Kaggle competitions

Lung cancer classification
Test size: max 1K

Schizophrenia classification
Test size: 120

Lung tumor segmentation
Test size: max 6k

Nerve segmentation
Test size 5.5K



[Varoquaux and Cheplygina 2022]

Uncertainty due to finite test set

Know when to stop, what to trust

(diminishing returns, creeping complexity)

Confidence interval[1]: Range of values compatible with the observations

[1] Technically not making the difference with a credible interval

N	65%	80%	90%	95%
100	[-9.0% 9.0%]	[-8.0% 8.0%]	[-6.0% 5.0%]	[-5.0% 4.0%]
1000	[-3.0% 2.9%]	[-2.5% 2.4%]	[-1.9% 1.8%]	[-1.4% 1.3%]
10000	[-0.9% 0.9%]	[-0.8% 0.8%]	[-0.6% 0.6%]	[-0.4% 0.4%]
100000	[-0.3% 0.3%]	[-0.2% 0.2%]	[-0.2% 0.2%]	[-0.1% 0.1%]

Table from [Varoquaux and Colliot 2022]

2
Benchmarking learning procedures

Benchmarking to conclude on good training procedures

- We are given: a procedure that outputs a prediction rule $\hat{f}$ from training data $(\mathbf{X}, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$

We want machine-learning research claims

(novel frobnicate improves prediction)

- Many arbitrary components

`torch.manual_seed(3407)` ?? [Picard 2021]

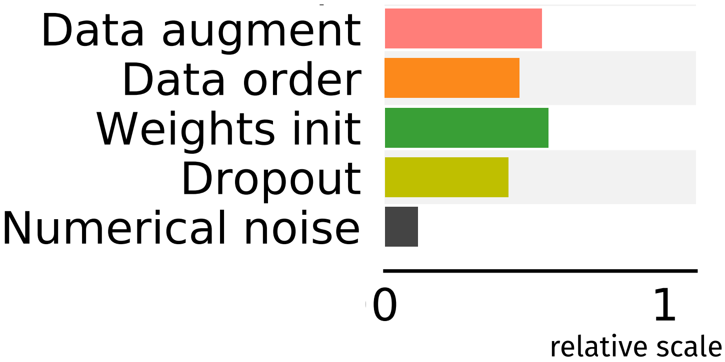
Useless to tune random seeds

(for weights init, dropout, data augmentation)

will not carry over to new training data

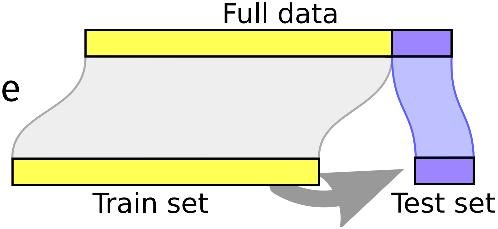
Benchmarking learning procedures: additional sources of variance

Variance when rerunning an evaluation,
modifying arbitrary elements:



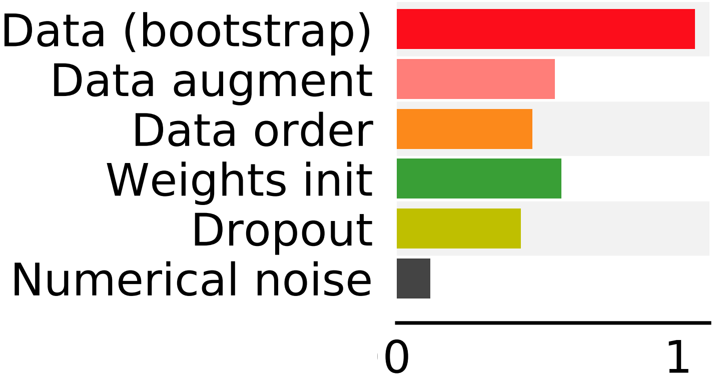
Across various computer vision and NLP tasks [Bouthillier... 2021]

Uncertainty due to test set sampling



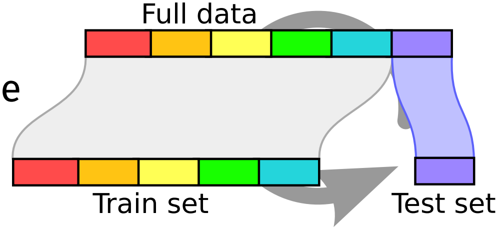
The test set remains a limited sample of the population

The train-test split is an arbitrary choice



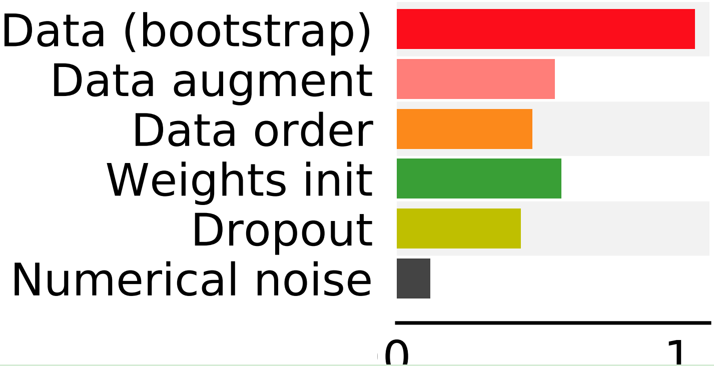
[Bouthillier... 2021]

Uncertainty due to test set sampling



The test set remains a limited sample of the population

The train-test split is an arbitrary choice



[Bouthillier... 2021]

Better evaluation

Sample multiple times these arbitrary choices: **cross-validation**

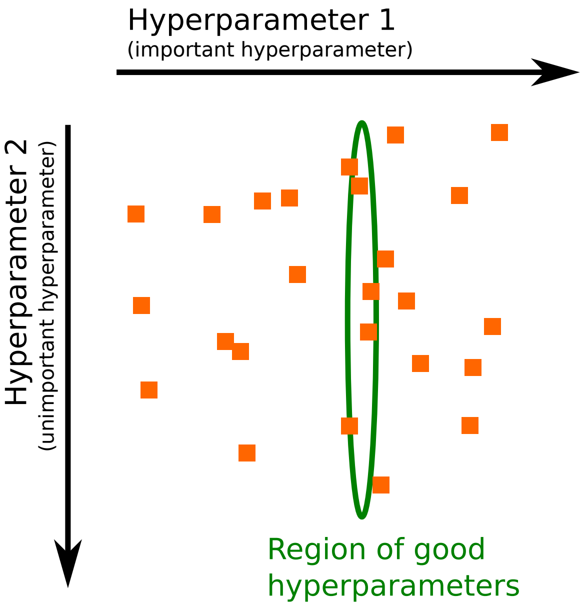
Benchmark also hyper-parameter selection

Sub-optimal hyper-parameters on models routinely lead to invalid conclusions

See refs in [Bouthillier... 2021]

Random search

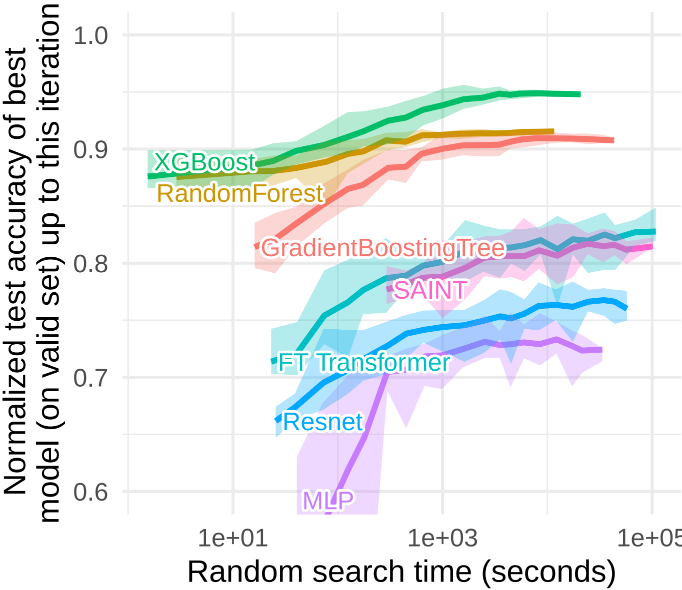
[Bergstra and Bengio 2012]



Benchmark also hyper-parameter selection

Sub-optimal hyper-parameters on models routinely lead to invalid conclusions

See refs in [Bouthillier... 2021]



Random search

[Bergstra and Bengio 2012]

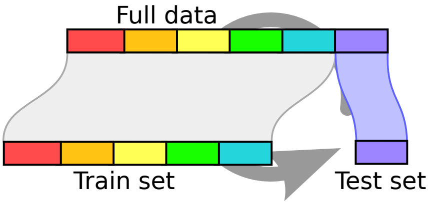
Draw subsets to estimate variance

[Grinsztajn... 2022]

Benchmarking training procedures (*eg* to compare them)

Control arbitrary fluctuations (that will not generalize)

Sample all:



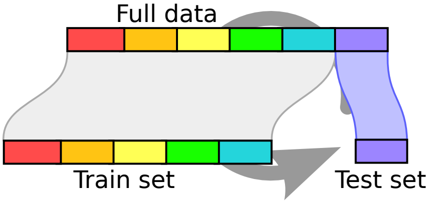
- data sampling
  Multiple train-test splits (cross-validation)

- arbitrary choices (seeds)
  Randomize them all

- hyper-parameters
  Hyper-parameter optimization   Too expensive to fully randomize

Accounting for variance in conclusions

Confidence intervals & statistical testing



Statistical testing with multiple folds

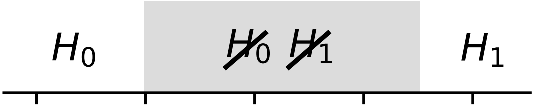
Challenge: folds are not independent

⚠ t-test/Wilcoxon across folds are not valid

Don't divide std by number of folds

Solution: Neyman-Pearson-like approach [Bouthillier... 2021]

- Test on $\mathbb{P}(p_1 > p_2) > \delta$



- Evaluate $\mathbb{P}(p_1 > p_2)$ by resampling

Randomize everything: data splits, seeds,...

Gaussian approximation: compare differences to standard deviations

More valid benchmarks [Varoquaux and Colliot 2022]

Meaningful performance metrics

- Should be suited to the application setting
- Machine learning does metric chasing 😬
- $\mathbb{P}$(true label | predicted label) $\mathbb{P}$(label | input)

Evaluation procedures

- Account for variance
- Difference between applying prediction rules & learning them

Careful benchmarking is crucial

- Optimistic flukes will not generalize
- What is our purpose? External validity ✊

@GaelVaroquaux

References I

A. I. Bandos, H. E. Rockette, and D. Gur. A permutation test sensitive to differences in areas for comparing roc curves from a paired design. *Statistics in medicine*, 24:2873, 2005.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281, 2012.

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, ... Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

D. Chyzhyk, G. Varoquaux, B. Thirion, and M. Milham. Controlling a confound in predictive models with a test set minimizing its effect. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE, 2018.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

J. Demšar. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML*, page 65. Citeseer, 2008.

References II

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

R. Dror, B. G., Bogomolov, M., and R. Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 2017.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

E. Lesaffre. Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU hospital for joint diseases*, 66(2), 2008.

M. A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording. Using and understanding cross-validation strategies. perspectives on saeb et al. *GigaScience*, 6(5):1–6, 2017.

References III

L. Maier-Hein, A. Reinke, E. Christodoulou, B. Glocker, P. Godau, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, ... Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv preprint arXiv:2206.01653*, 2022.

A. Makarova, H. Shen, V. Perrone, A. Klein, J. B. Faddoul, A. Krause, M. Seeger, and C. Archambeau. Overfitting in bayesian optimization: an empirical study and early-stopping solution. *arXiv preprint arXiv:2104.08166*, 2021.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine learning*, 52(3): 239–281, 2003.

A. Perez-Lebel, M. L. Morvan, and G. Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. *arXiv:2210.16315*, 2022.

D. Picard. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv:2109.08203*, 2021.

N. Traut, K. Heuer, G. Lemaître, A. Beggiato, D. Germanaud, M. Elmaleh, A. Bethegnies, L. Bonnasse-Gahot, W. Cai, S. Chambon, ... Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery. *NeuroImage*, 255:119171, 2022.

References IV

G. Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage*, 180:68–77, 2018.

G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8, 2022.

G. Varoquaux and O. Colliot. Evaluating machine learning models and their diagnostic value. `https://hal.archives-ouvertes.fr/hal-03682454/`, 2022.