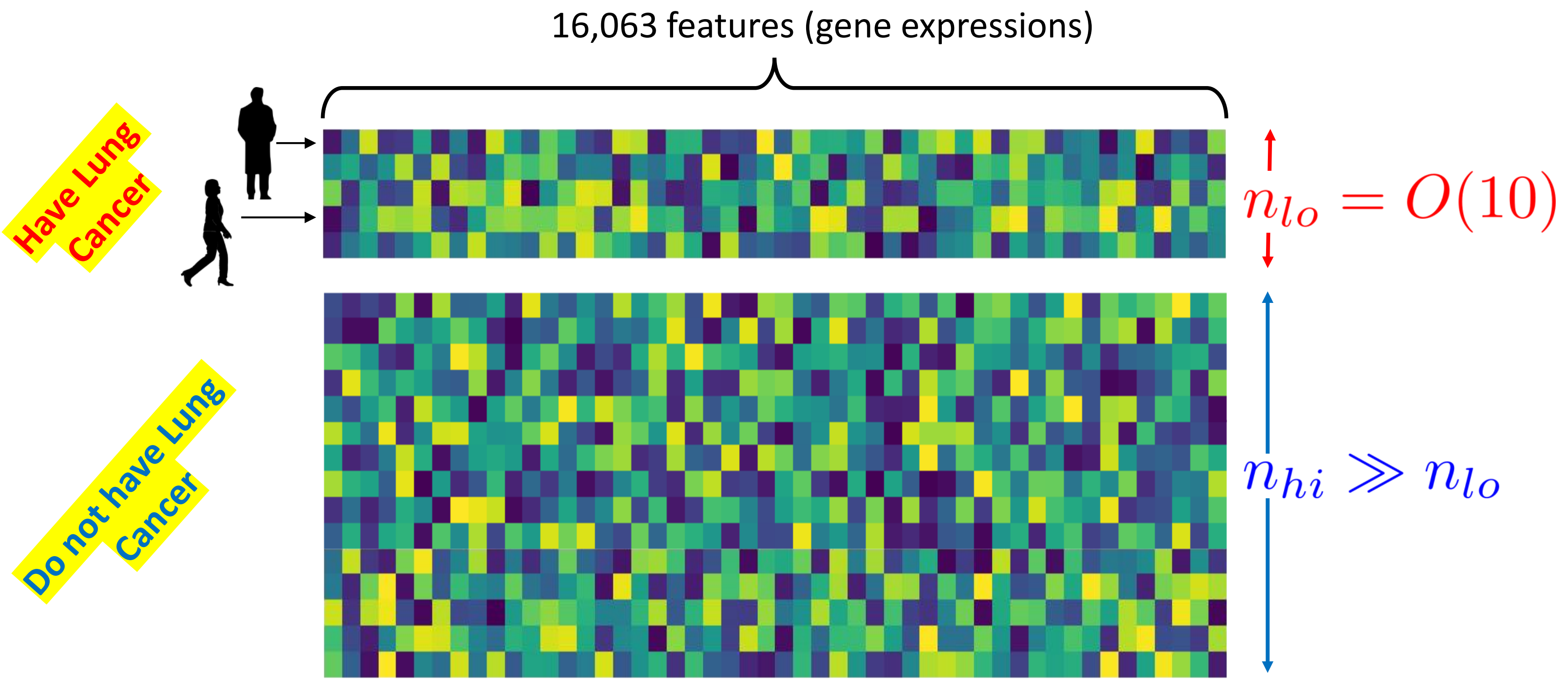
ROBUST HIGH-DIMENSIONAL CLASSIFICATION FROM FEW POSITIVE EXAMPLES

Deepayan Chakrabarti (deepay@utexas.edu)
Benjamin Fauber (ben.fauber@dell.com)

PROBLEM



How can we classify tumors from genes?

High-dimensional Limited-data Imbalanced ^ Binary Classification

$O(10^5)$ features $n_{lo} = O(10)$ $n_{lo} \ll n_{hi}$

Existing approaches

Modify the data: sample, then train

But samples built from limited data can be biased

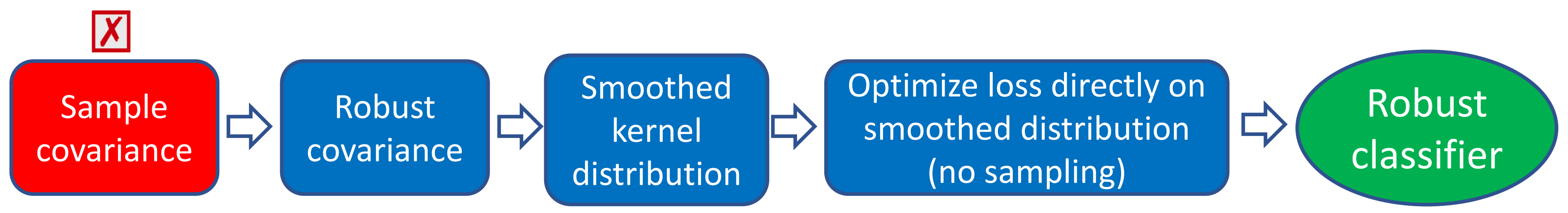
Ensemble methods: many repetitions of the above

Can overfit due to limited data and high dimensionality

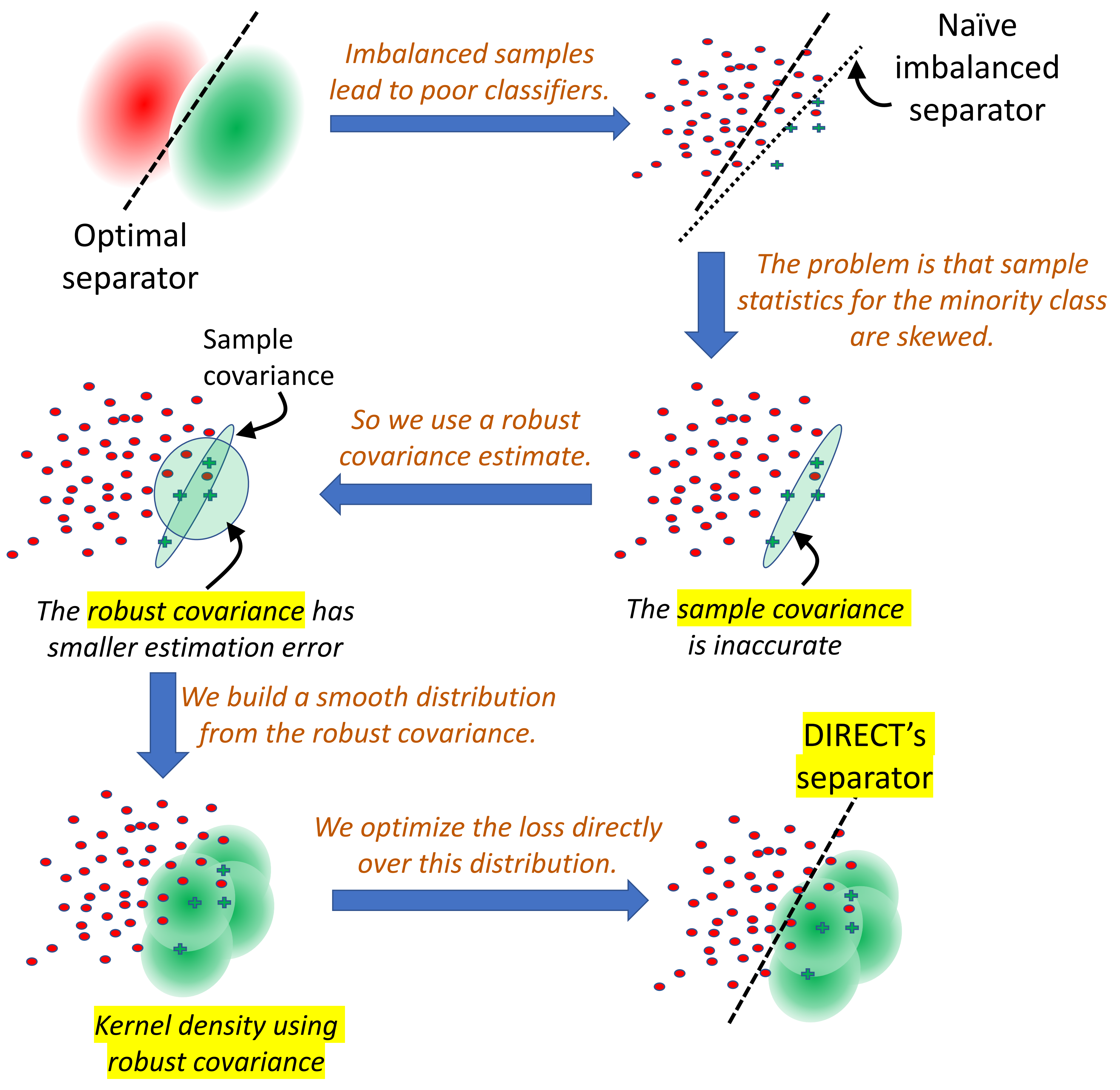
Cost-sensitive methods: modify the loss

Underperform for limited-size datasets [Cunha+/21]

APPROACH



Sample covariance ✗ ⇨ Robust covariance ⇨ Smoothed kernel distribution ⇨ Optimize loss directly on smoothed distribution (no sampling) ⇨ Robust classifier



Imbalanced samples lead to poor classifiers.

The problem is that sample statistics for the minority class are skewed.

So we use a robust covariance estimate.

The robust covariance has smaller estimation error

The sample covariance is inaccurate

We build a smooth distribution from the robust covariance.

We optimize the loss directly over this distribution.

Kernel density using robust covariance

DIRECT is fast, parameter-free, and accurate

https://github.com/deepayan12/direct

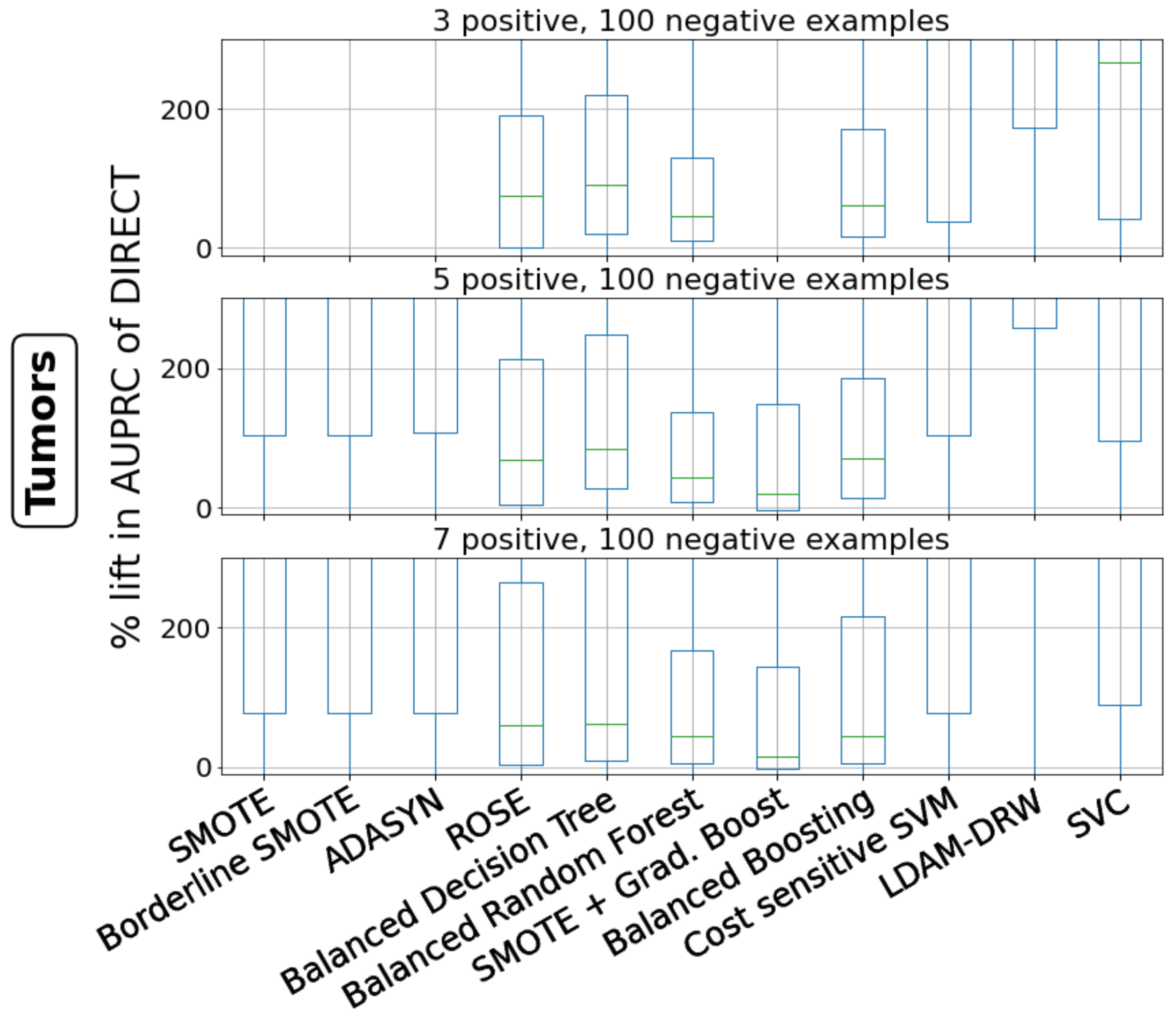
RESULTS

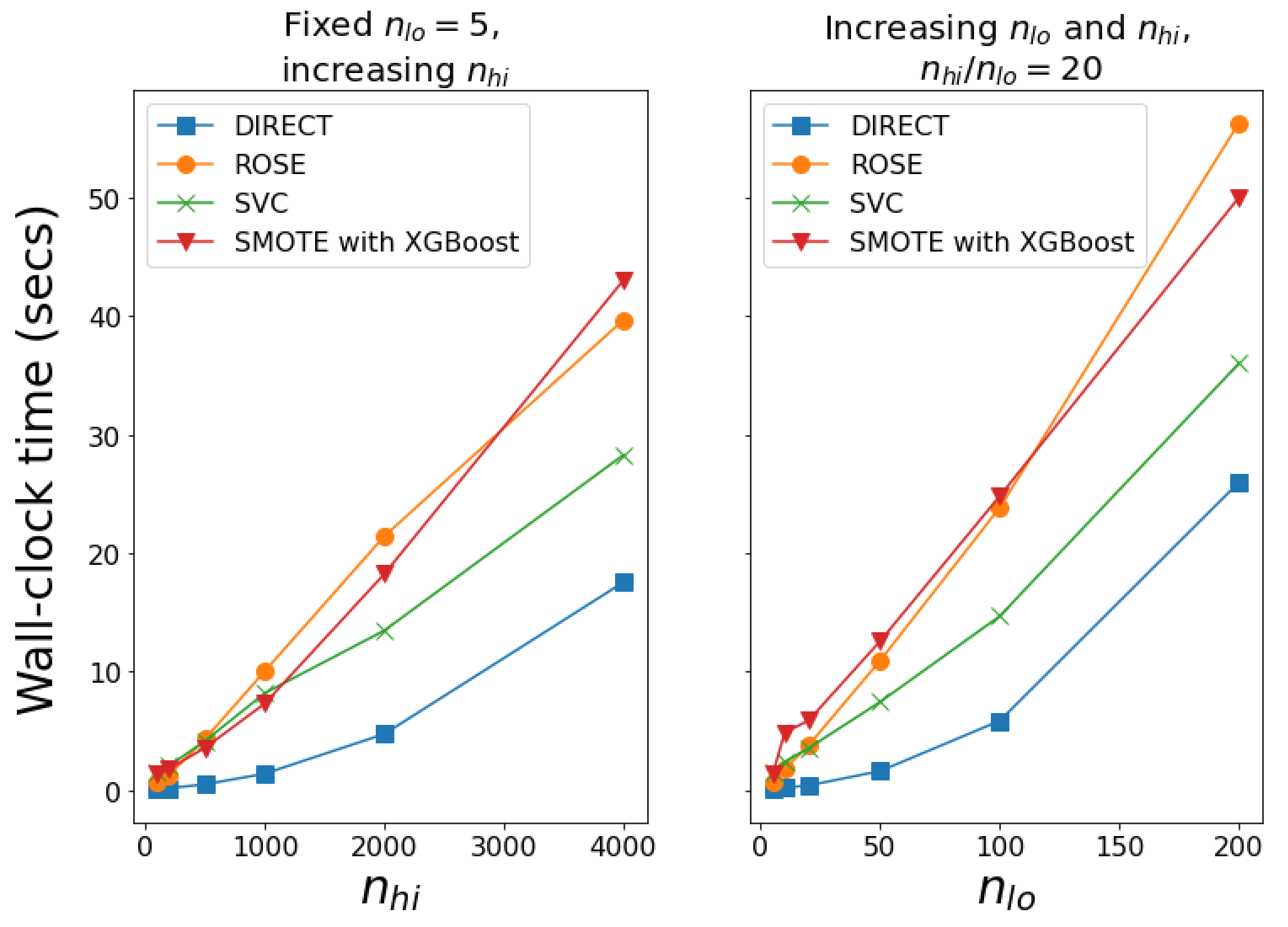
Datasets

1 medical (16K features)
2 image
5 text (10K-100K features)
20 UCI datasets

Metric

Area under the Precision-Recall curve (AUPRC)


Tumors
% lift in AUPRC of DIRECT
3 positive, 100 negative examples
5 positive, 100 negative examples
7 positive, 100 negative examples
SMOTE
Borderline SMOTE
ADASYN
ROSE
Balanced Decision Tree
Balanced Random Forest
SMOTE + Grad. Boost
Balanced Boosting
Cost sensitive SVM
LDAM-DRW
SVC


Fixed $n_{lo} = 5$, increasing n_{hi}
Increasing n_{lo} and n_{hi}, $n_{hi}/n_{lo} = 20$
DIRECT
ROSE
SVC
SMOTE with XGBoost
Wall-clock time (secs)
n_{hi}
n_{lo}