Memorial Sloan Kettering Cancer Center

IMPACT annotator

November 5, 2018

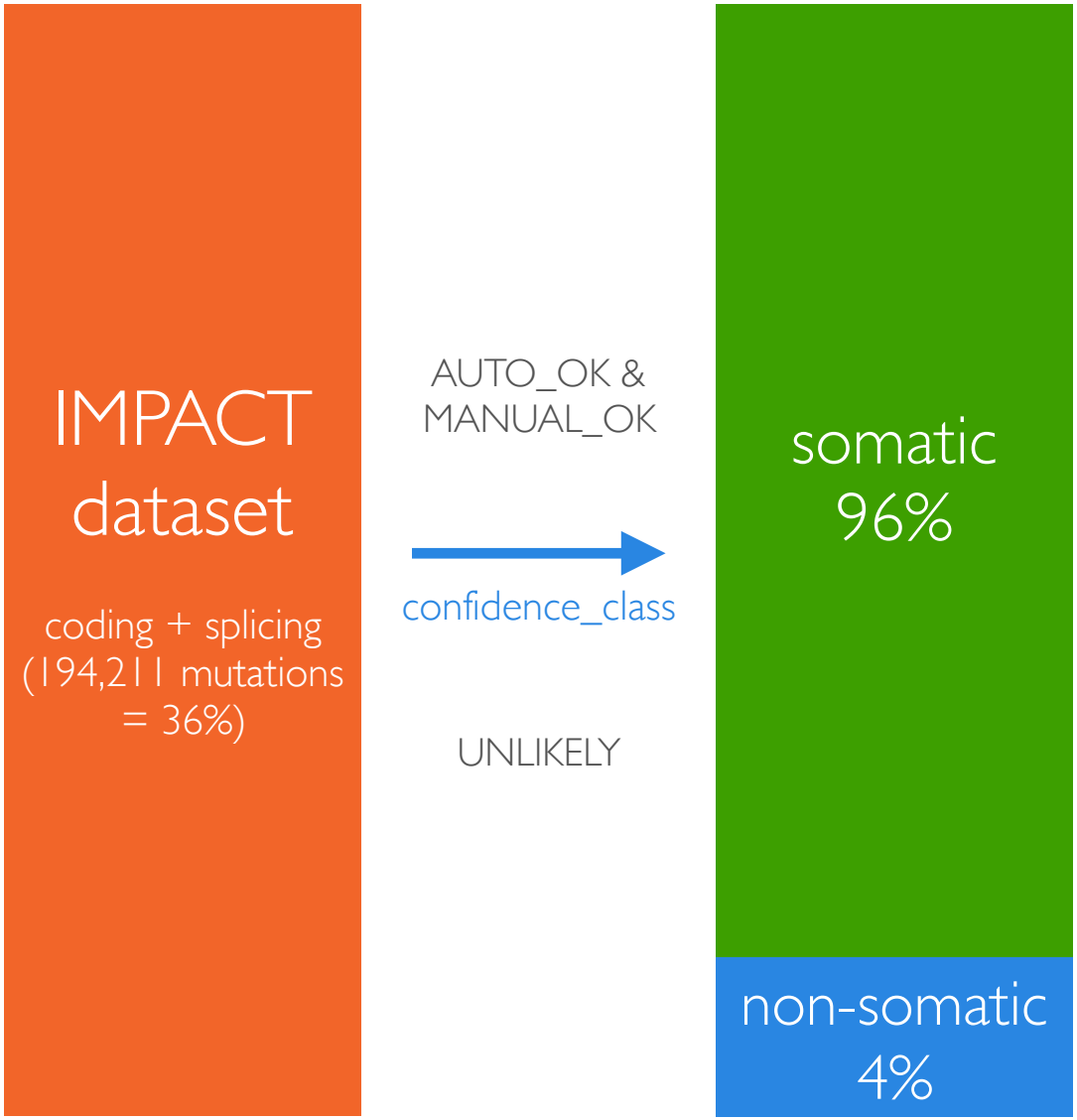
Pierre Guilmin
Elsa Bernard

Different classes

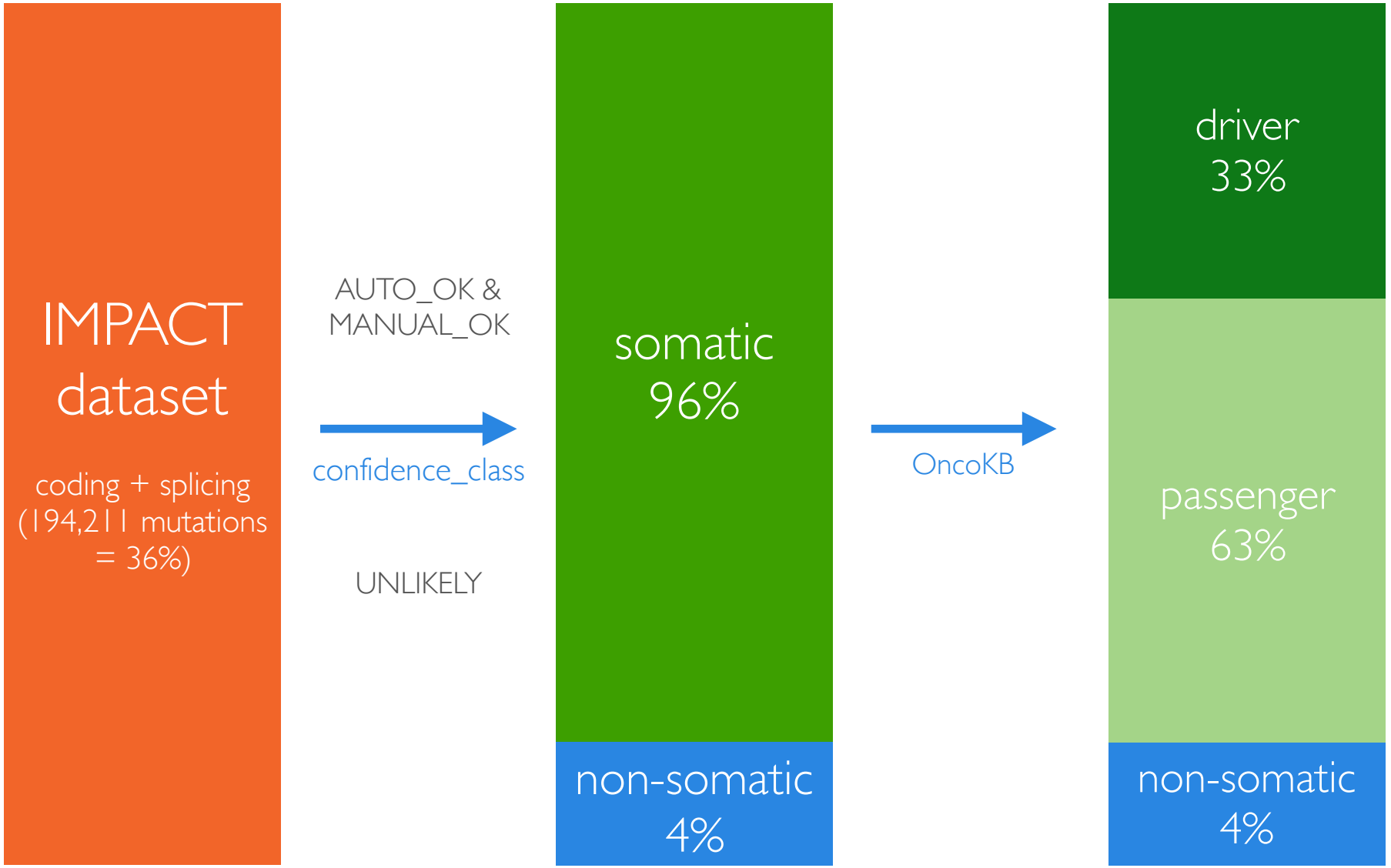
IMPACT dataset

coding + splicing
(194,211 mutations
= 36%)

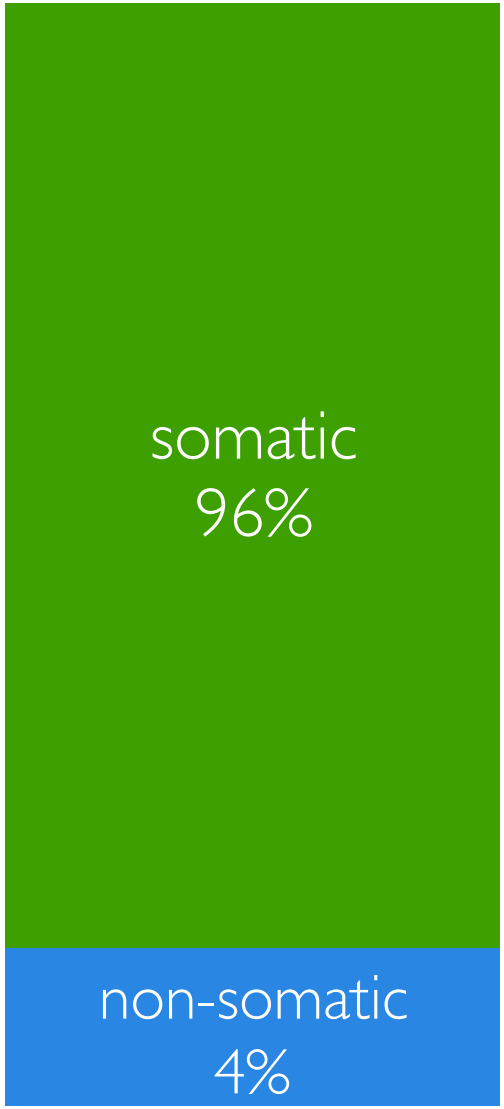
Different classes

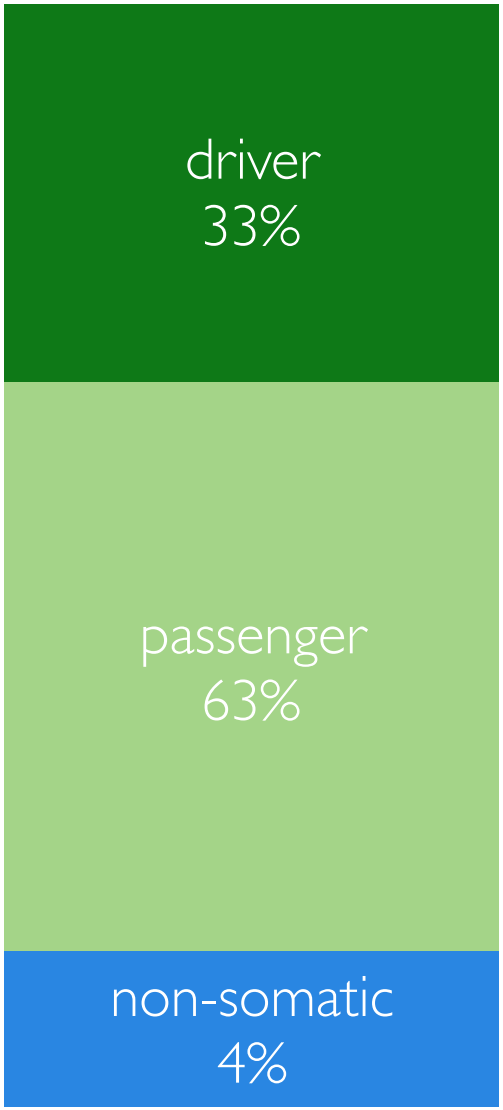


Different classes

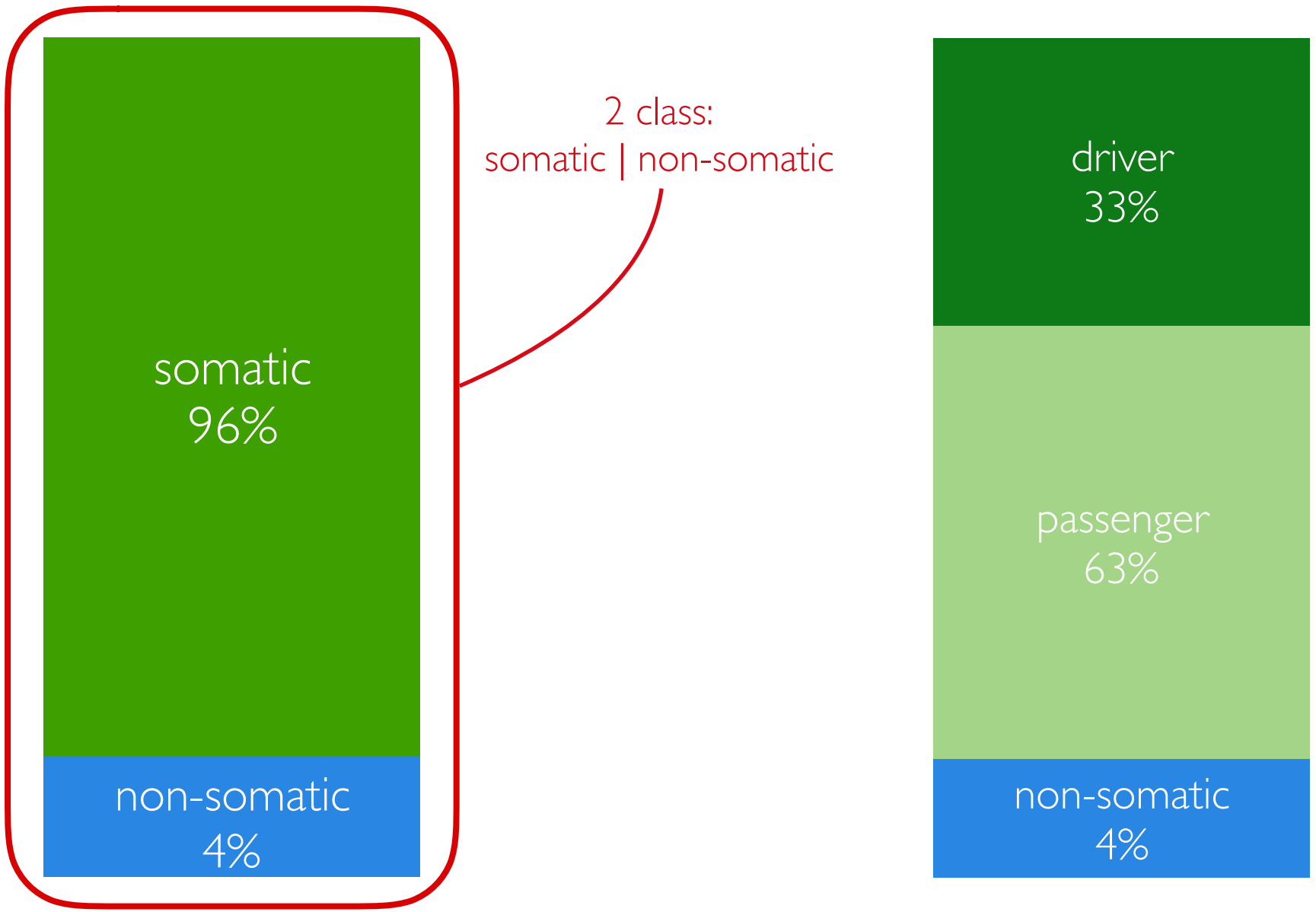


Supervised learning

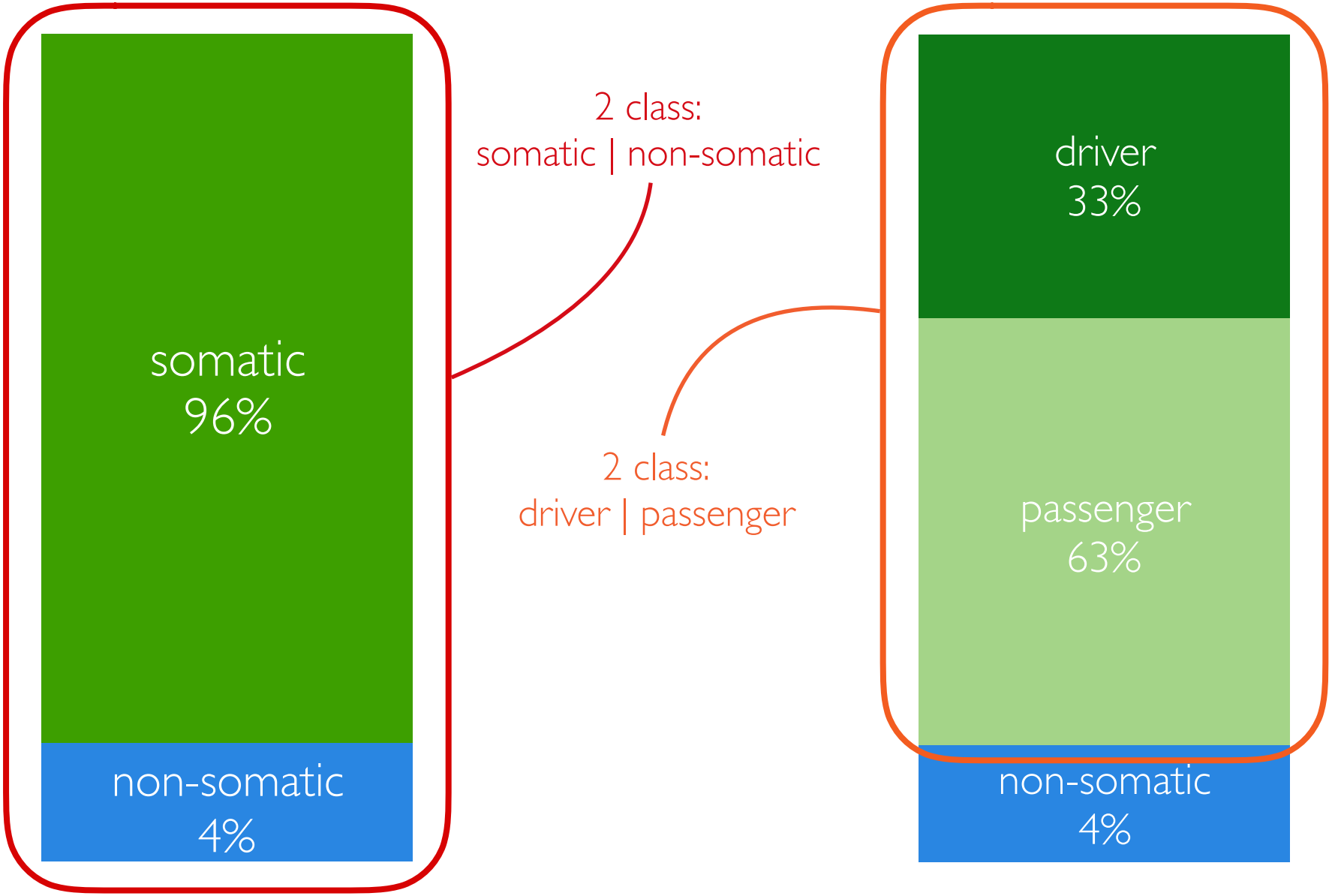




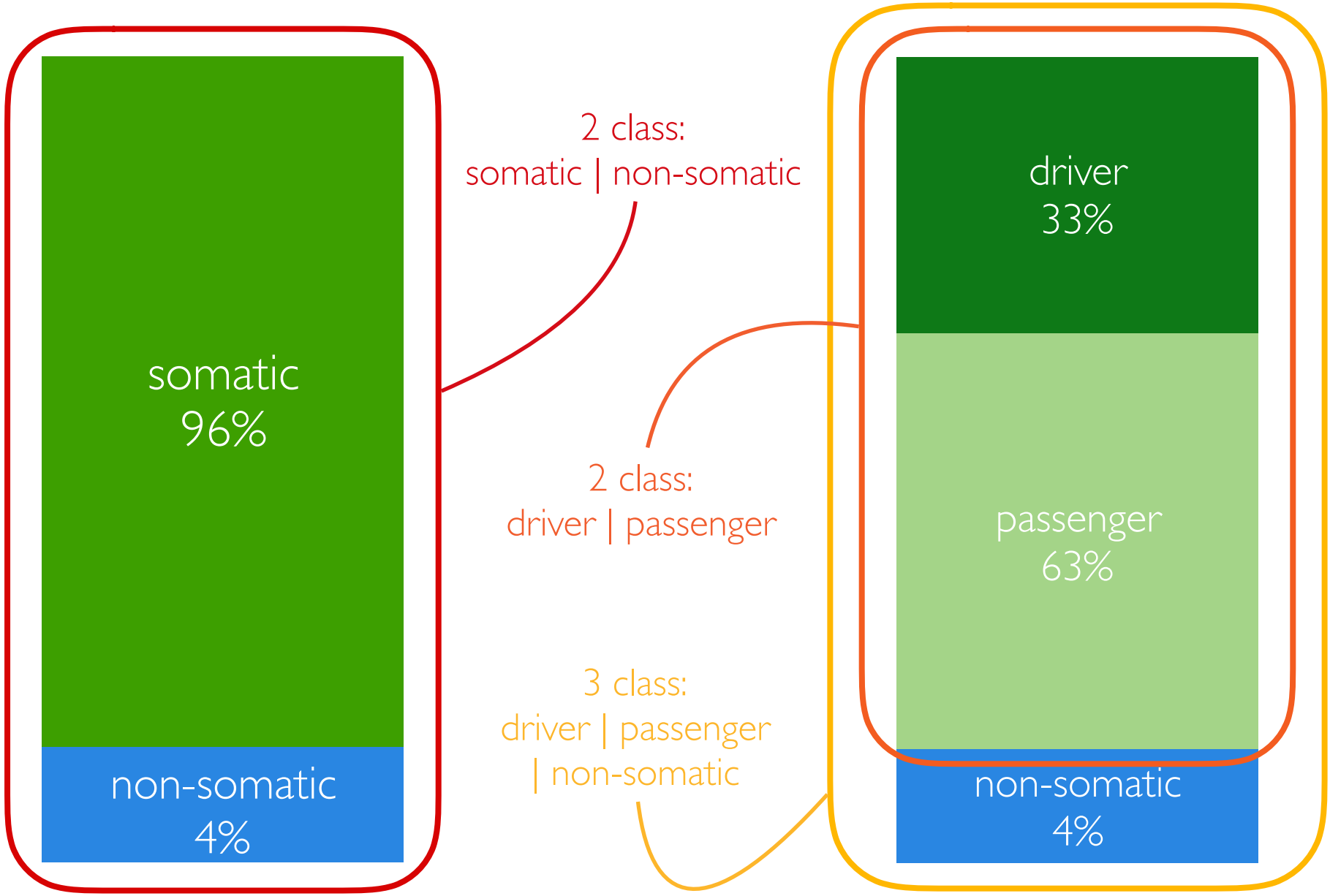
Supervised learning



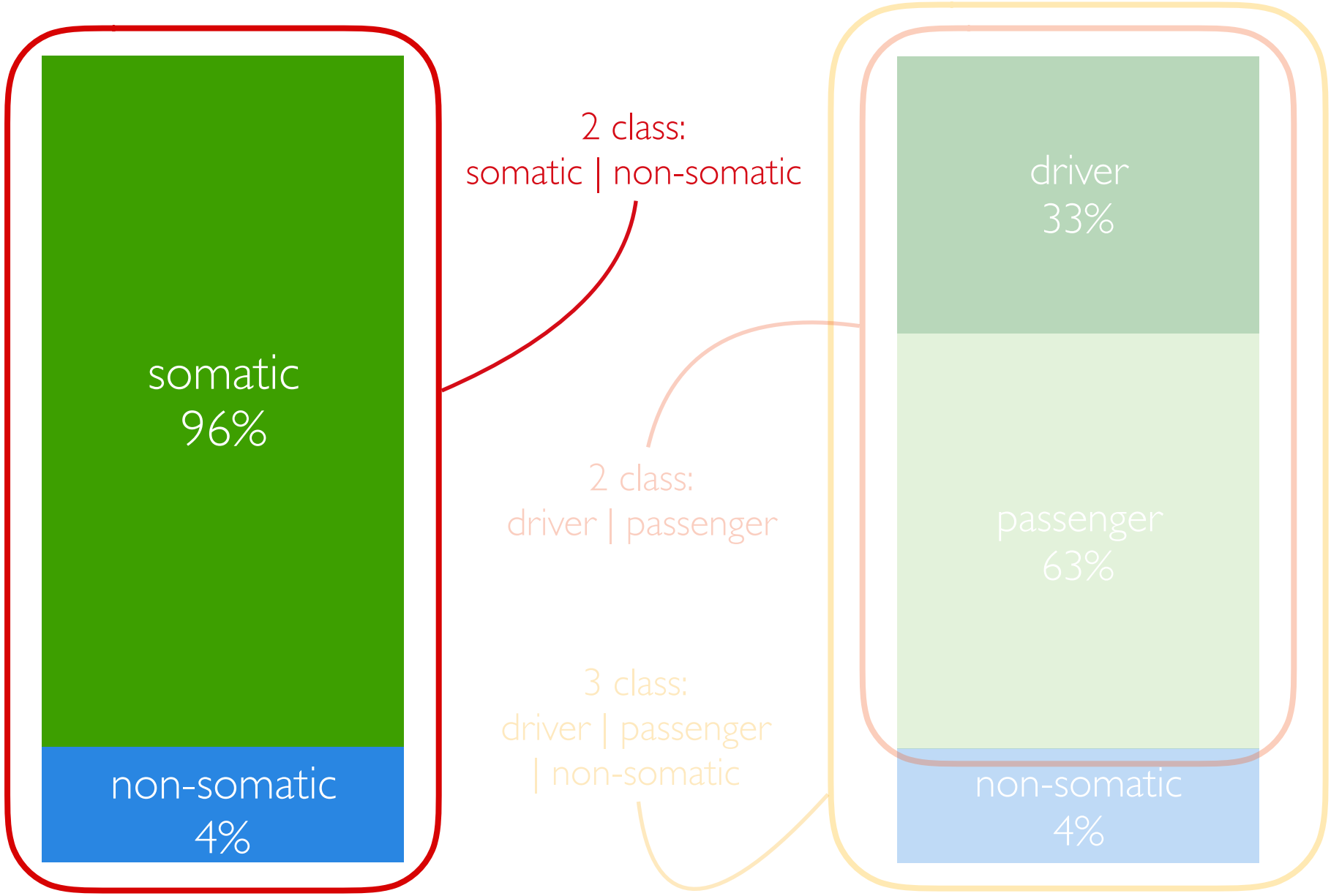
Supervised learning



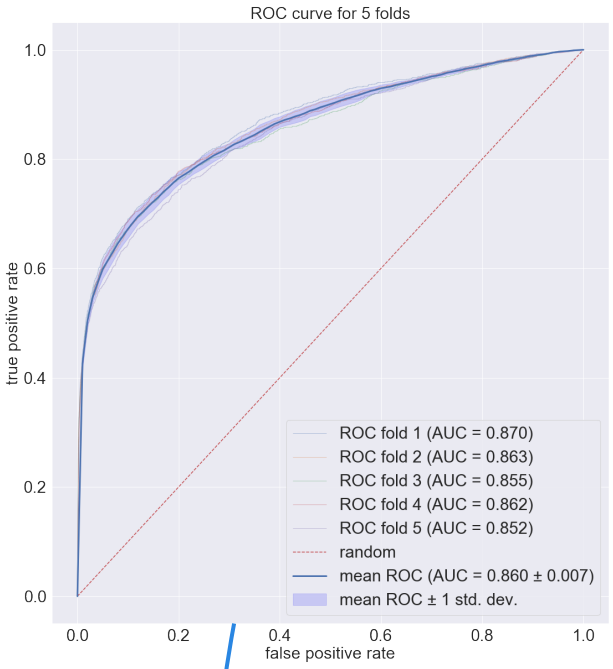
Supervised learning

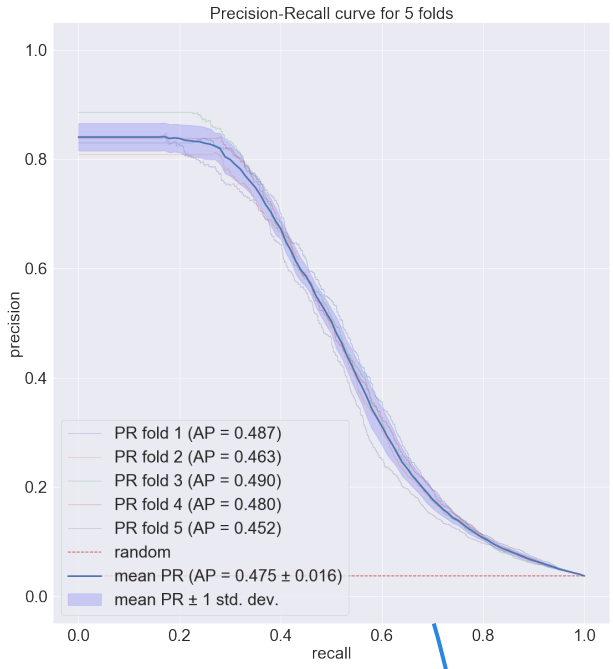


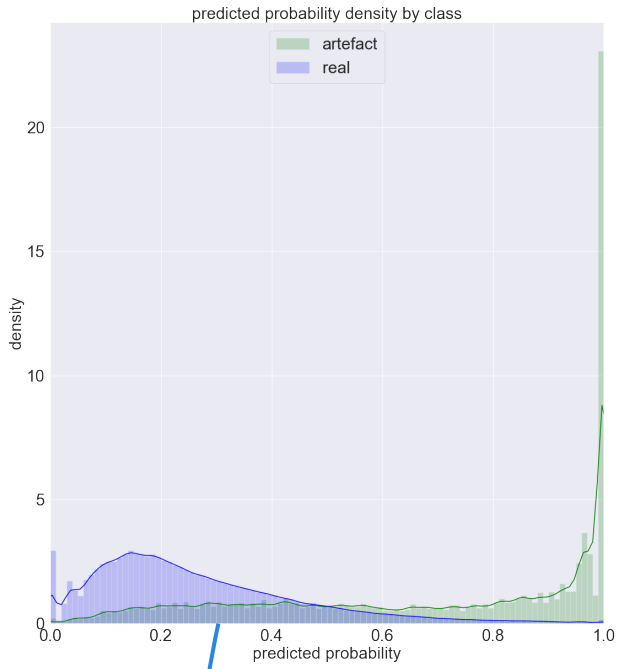
Supervised learning



Main metrics used







ROC AUC

area under the ROC curve

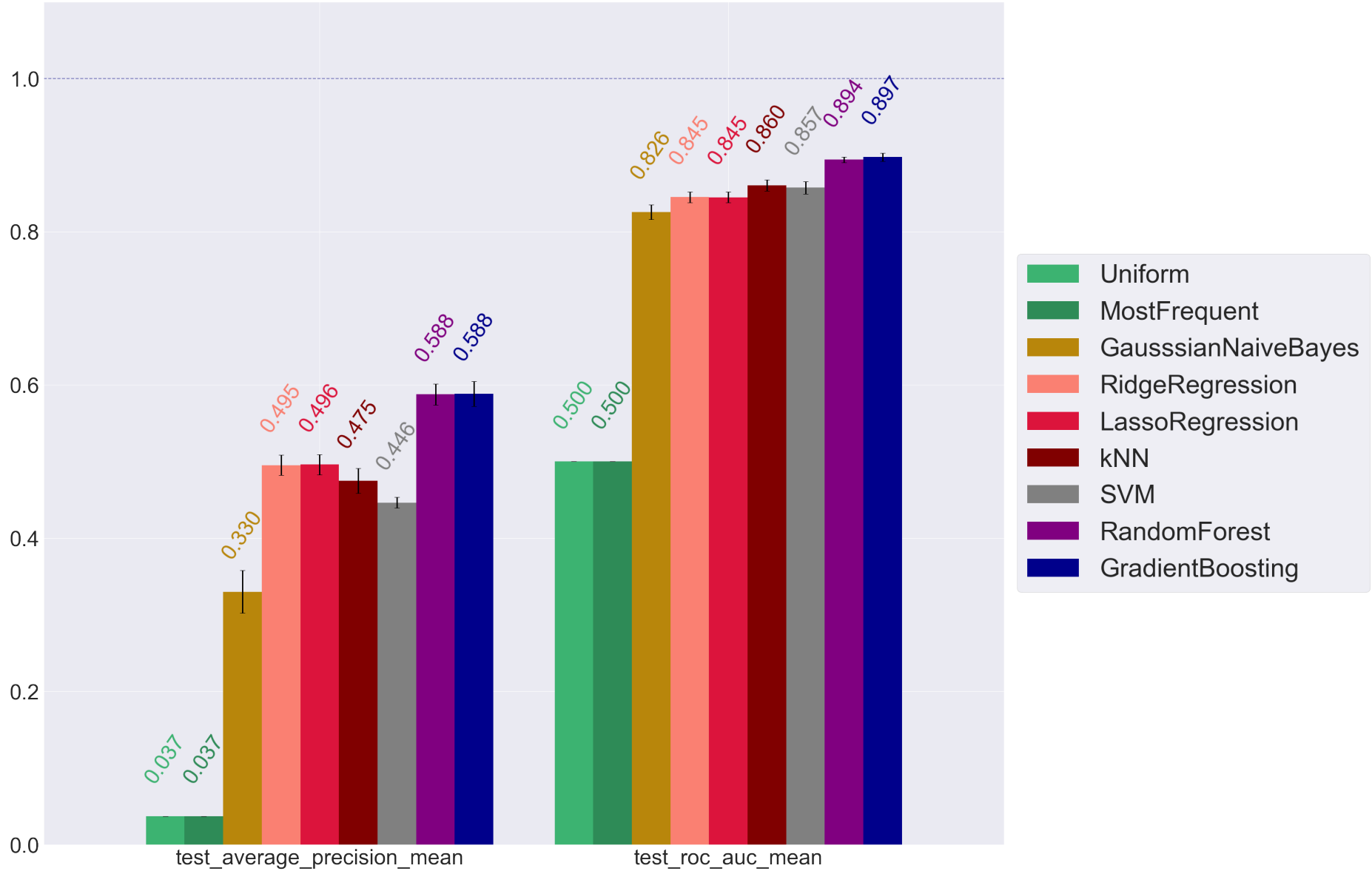
average precision

area under the precision-recall curve

Probability distribution

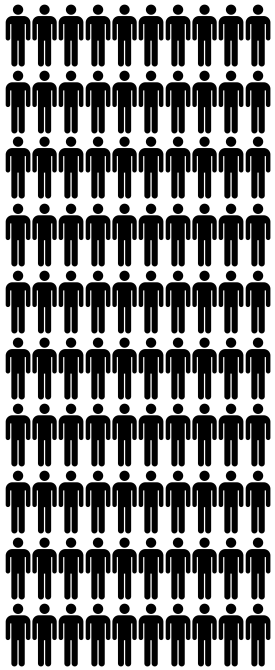
probability predicted by the classifier for each class

Algorithm comparison



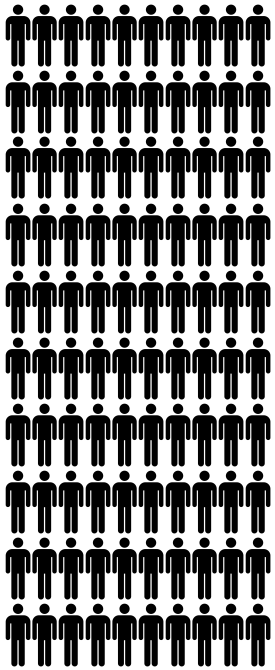
IMPACT annotator

100 patients



IMPACT annotator

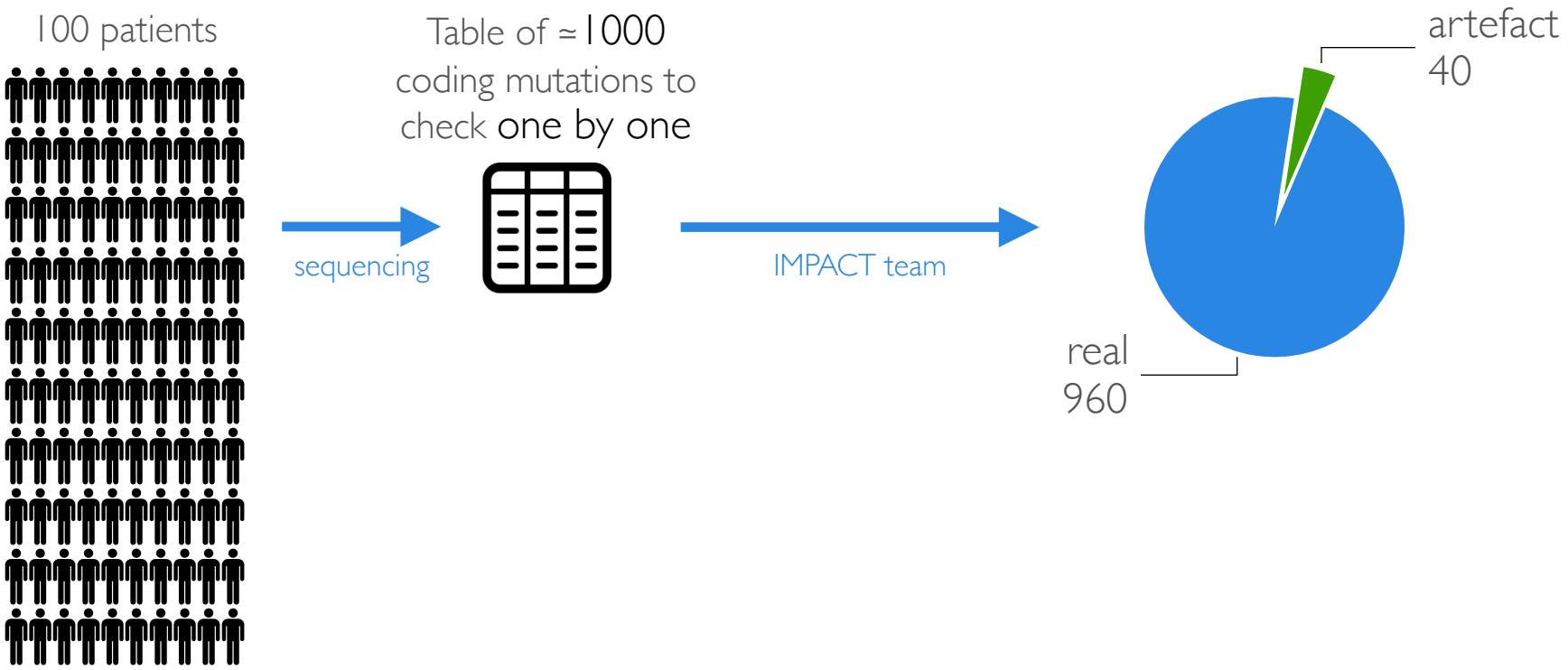
100 patients



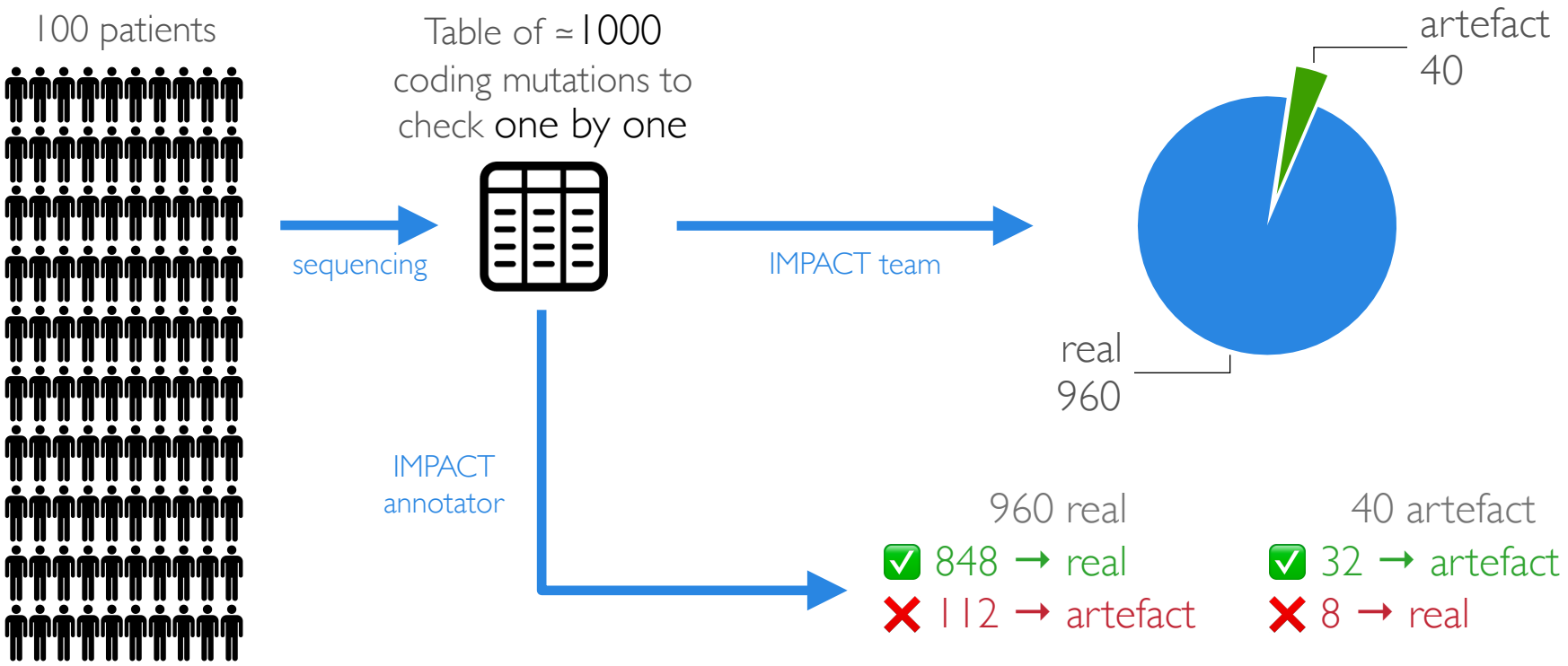
sequencing

Table of ≃1000 coding mutations to check one by one

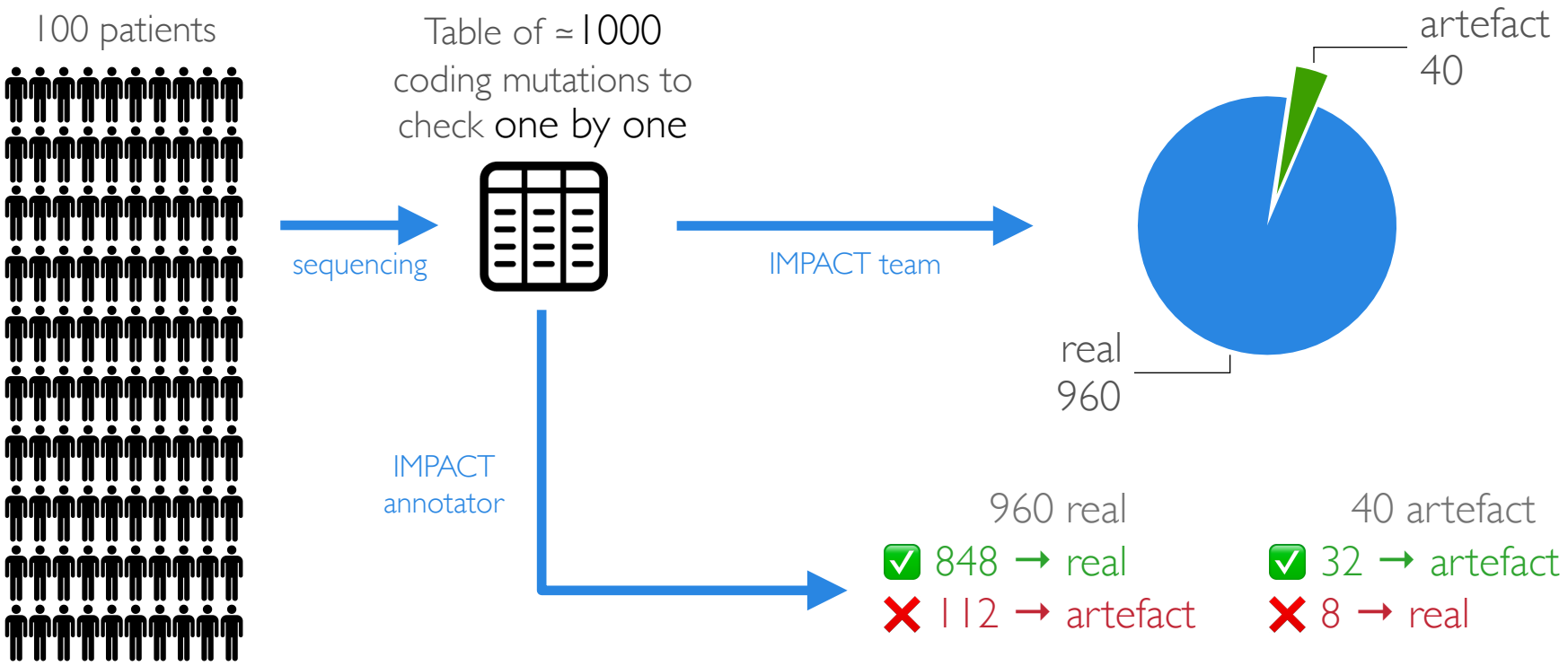
IMPACT annotator



IMPACT annotator



IMPACT annotator



IMPACT team

1000/1000 mutations to check one by one

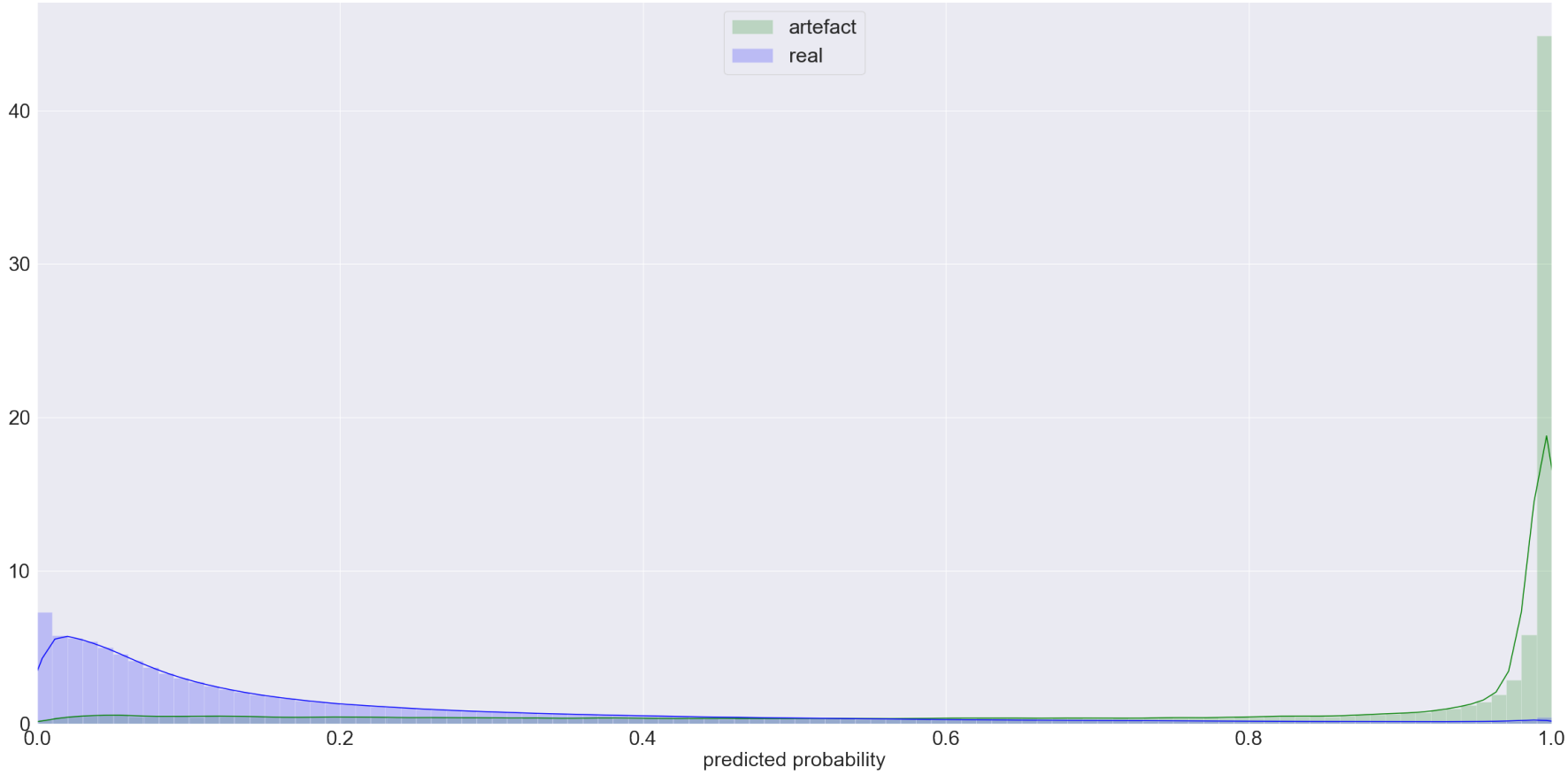
❌ 0/40 artefacts considered as real

IMPACT annotator

144/1000 mutations to check one by one

❌ 8/40 artefacts considered as real

Threshold choice

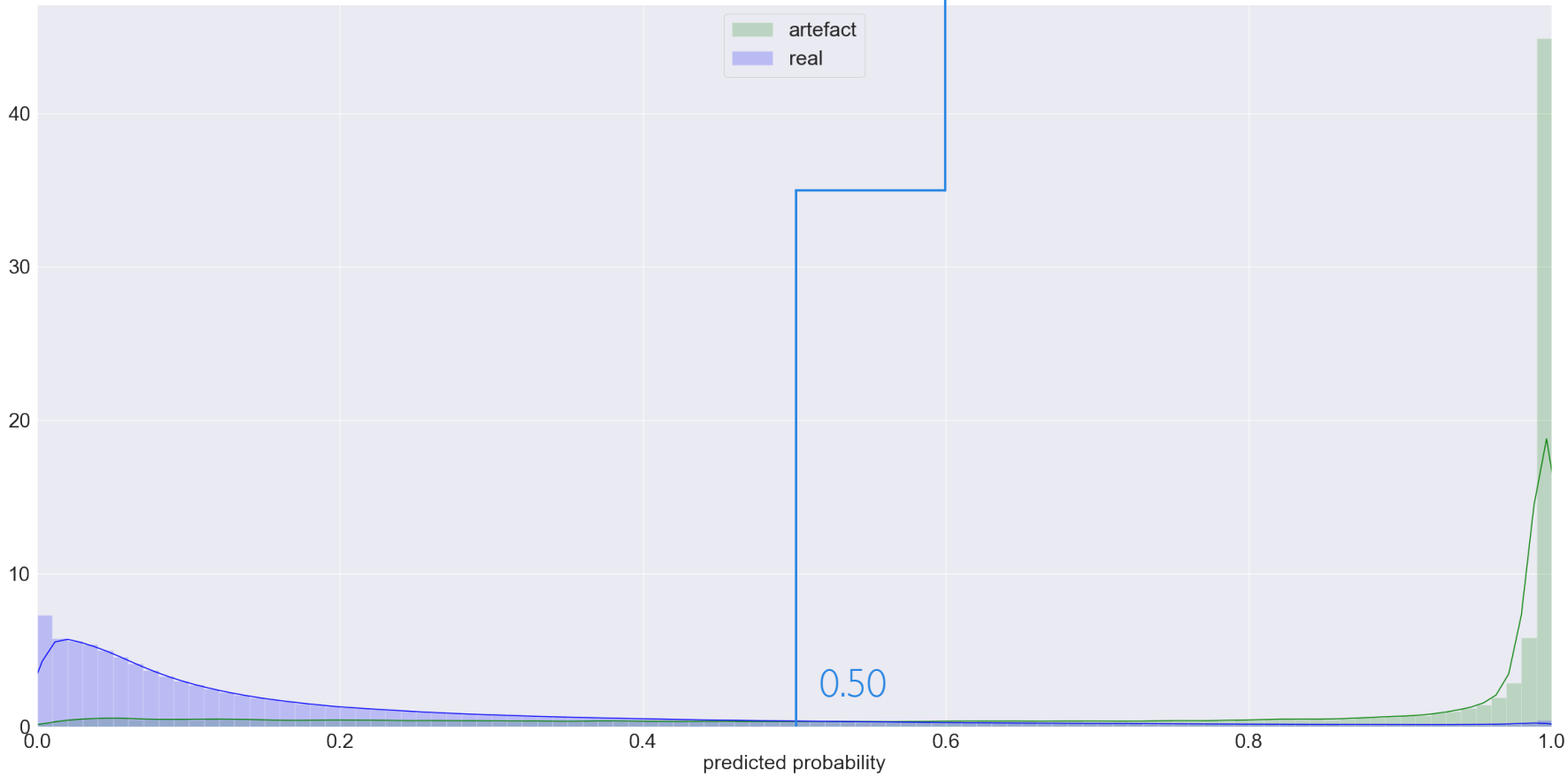


Threshold choice

threshold = 0.5

📝 144 mutations to check

960 real	40 artefact
✅ 848 → real	✅ 32 → artefact
❌ 112 → artefact	❌ 8 → real



Threshold choice

threshold = 0.25

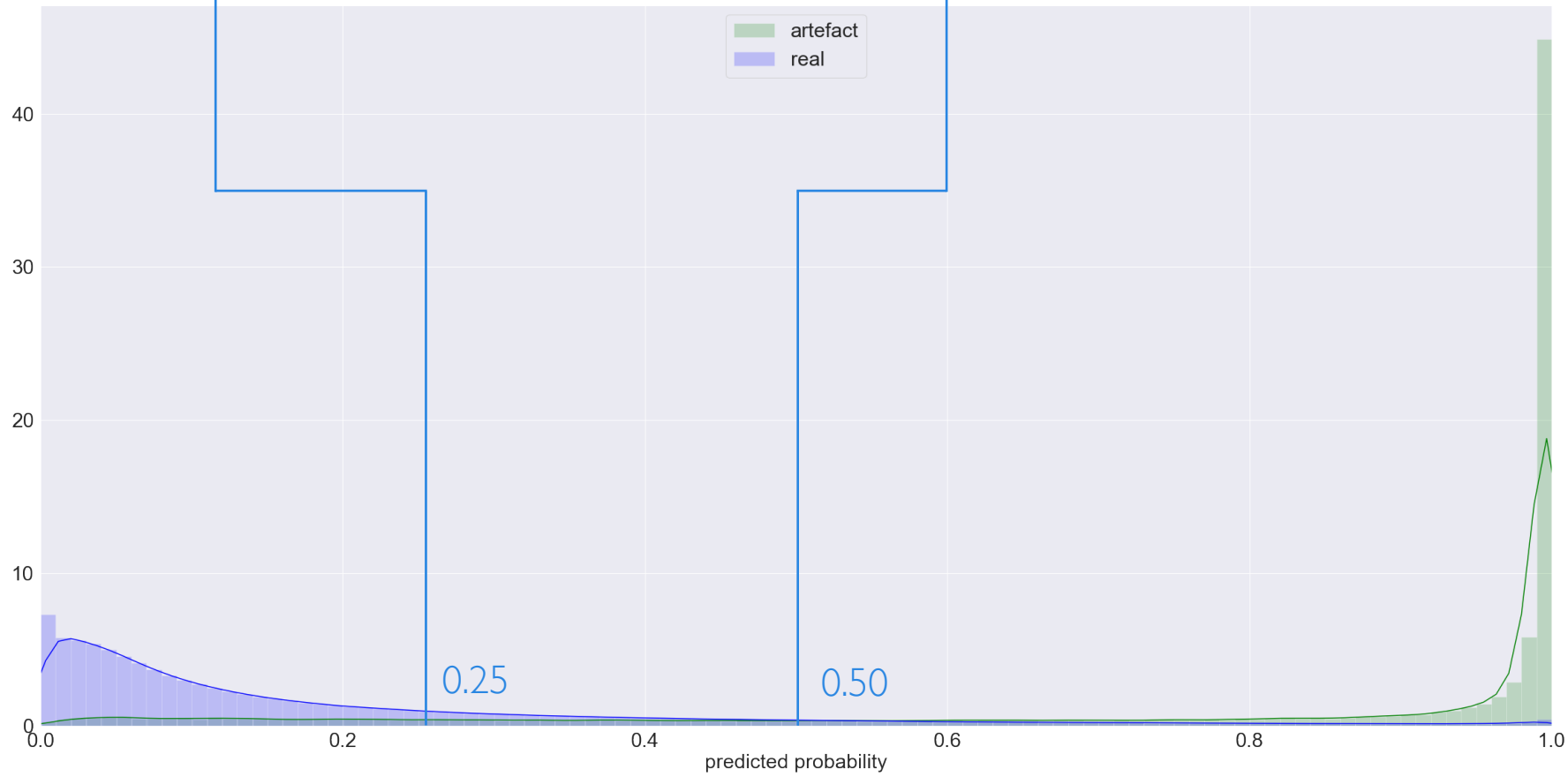
296

960 real	40 artefact
✅ 699	✅ 35
❌ 261	❌ 5

threshold = 0.5

144 mutations to check

960 real	40 artefact
✅ 848 → real	✅ 32 → artefact
❌ 112 → artefact	❌ 8 → real



Threshold choice

threshold = 0.25
296

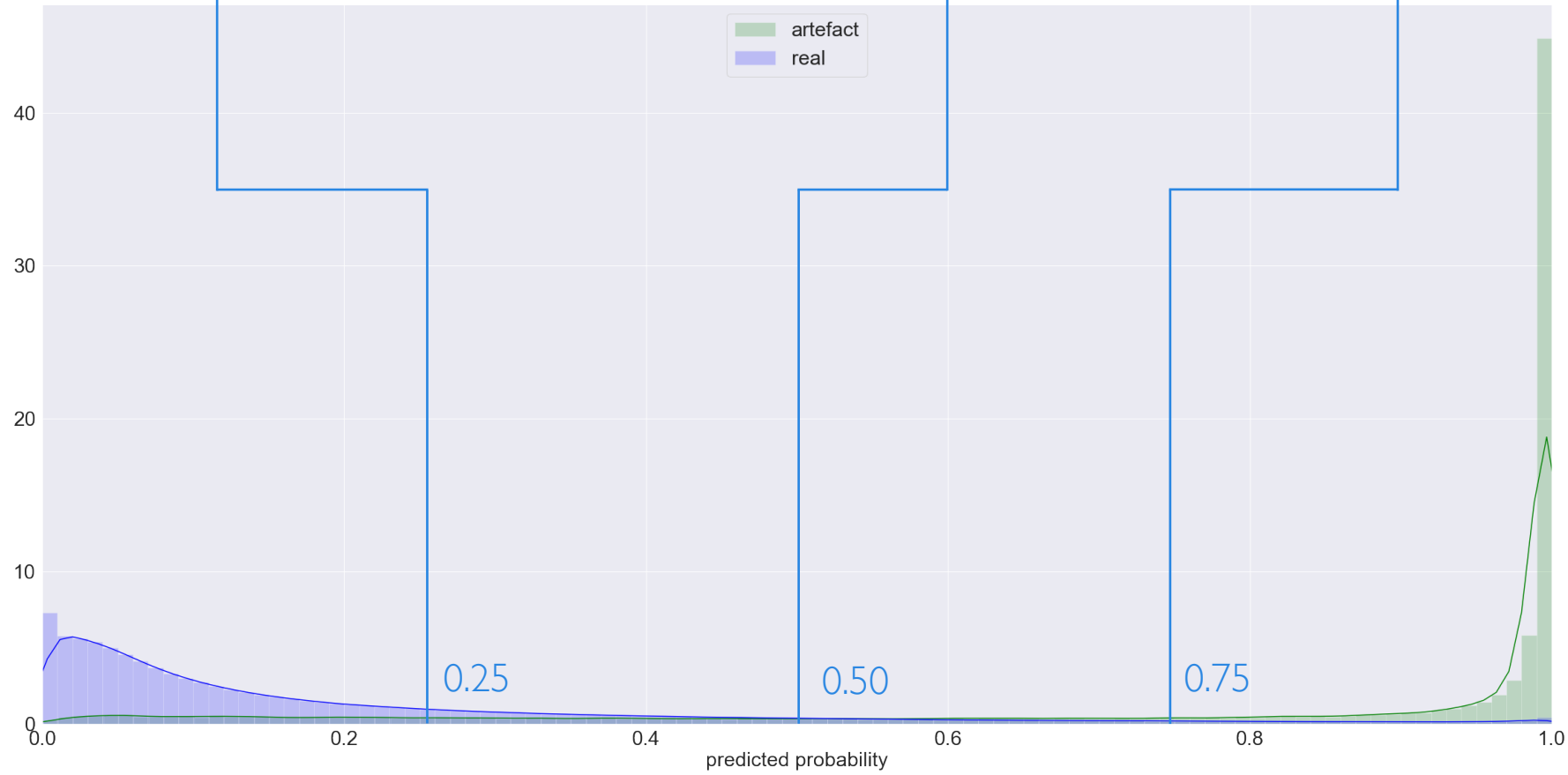
960 real	40 artefact
✅ 699	✅ 35
❌ 261	❌ 5

threshold = 0.5
144 mutations to check

960 real	40 artefact
✅ 848 → real	✅ 32 → artefact
❌ 112 → artefact	❌ 8 → real

threshold = 0.75
72

960 real	40 artefact
✅ 916	✅ 28
❌ 44	❌ 12



The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

somatic driver (x 4)

- COSMIC count
- is a hotspot, is a 3d hotspot
- OncoKB oncogenic

The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

AF (x 12)

- in dbSNP
- gnomAD total AF and AF by population (AFR, AMR, ASJ, …)

genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

somatic driver (x 4)

- COSMIC count
- is a hotspot, is a 3d hotspot
- OncoKB oncogenic

The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

AF (x 12)

- in dbSNP
- gnomAD total AF and AF by population (AFR, AMR, ASJ, …)

genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

freq (x 1)

- frequency in normals

somatic driver (x 4)

- COSMIC count
- is a hotspot, is a 3d hotspot
- OncoKB oncogenic

The features group

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor +/- strand count
- sample_coverage

AF (x 12)

- in dbSNP
- gnomAD total AF and AF by population (AFR, AMR, ASJ, …)

genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

freq (x 1)

- frequency in normals

somatic driver (x 4)

- COSMIC count
- is a hotspot, is a 3d hotspot
- OncoKB oncogenic

Consequence (x 6)

- gene type
- mutation consequence (stopgain, frameshift, …)
- VEP_IMPACT, VEP_CLIN_SIG
- SIFT and PolyPhen class

The features group

AF (x 12)

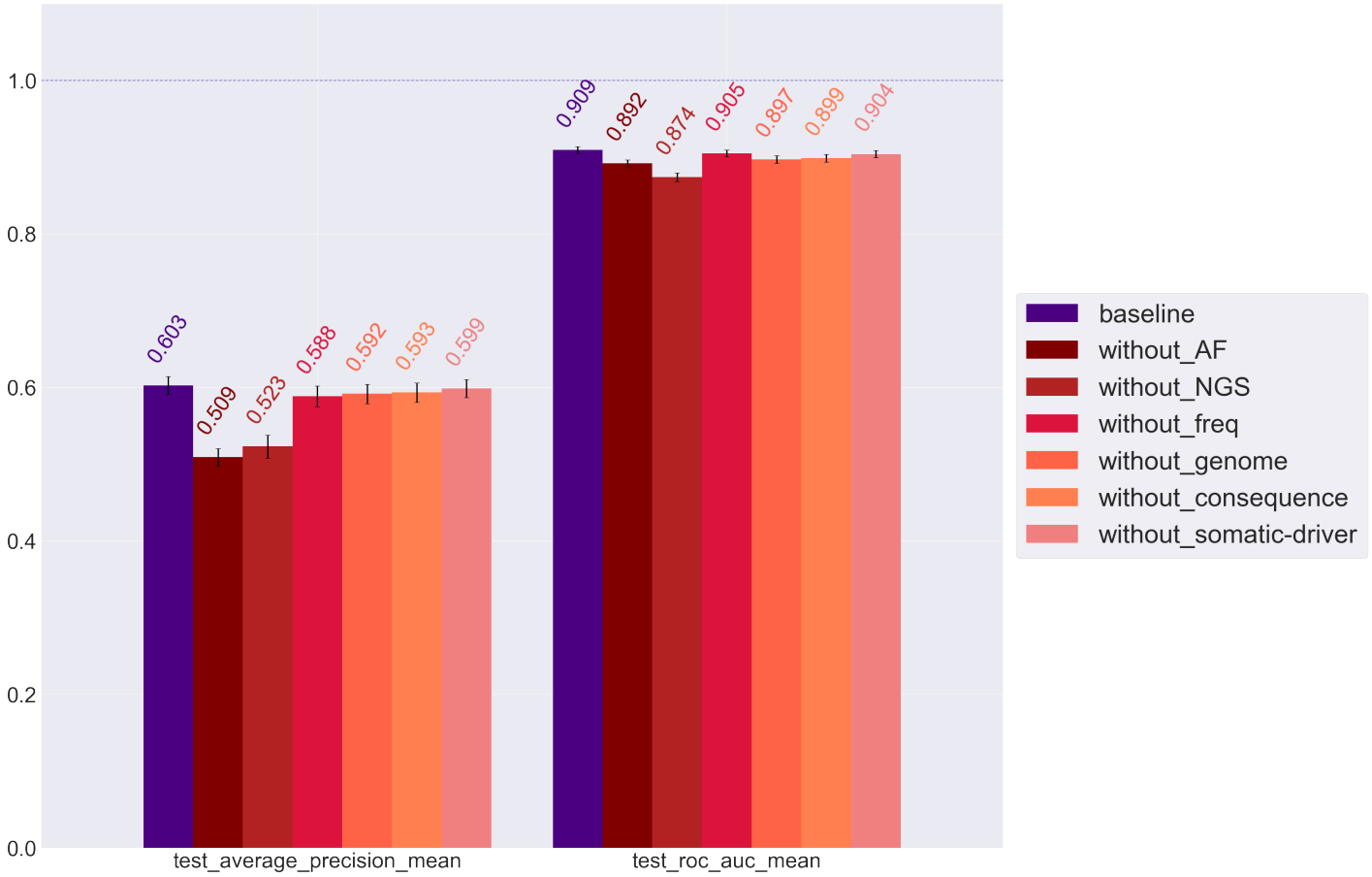
- in dbSNP
- gnomAD total AF and AF by population (AFR, AMR, ASJ, …)

NGS (x 11)

- tumor & normal depth, vaf and count
- tumor positive and negative count
- sample_coverage

freq (x 1)

- frequency in normals



genome (x 3)

- Chromosome
- Hugo Symbol
- Variant Class

Consequence (x 6)

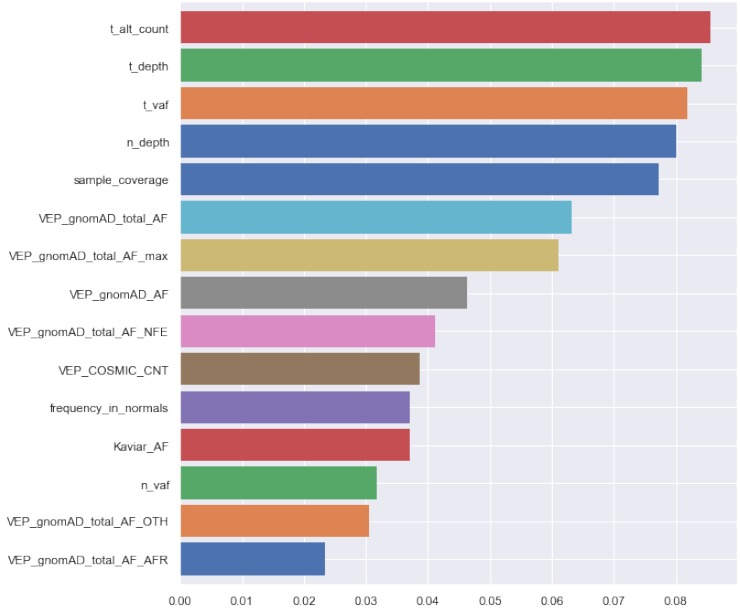
- gene type
- Mutation consequence (stopgain, frameshift, …)
- VEP_IMPACT,VEP_CLIN_SIG
- SIFT and PolyPhen class

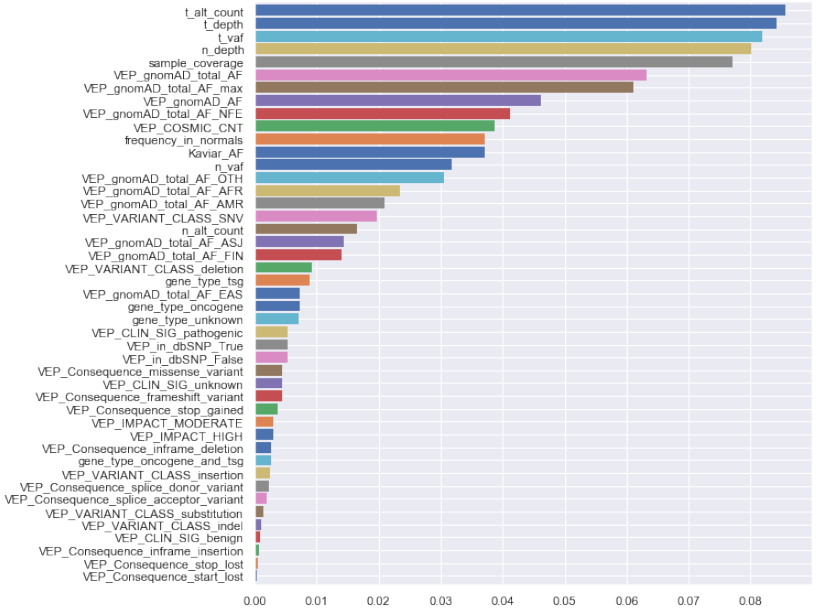
somatic driver (x 4)

- COSMIC count
- is a hotspot, is a 3d hotspot
- OncoKB oncogenic

Individual features importance

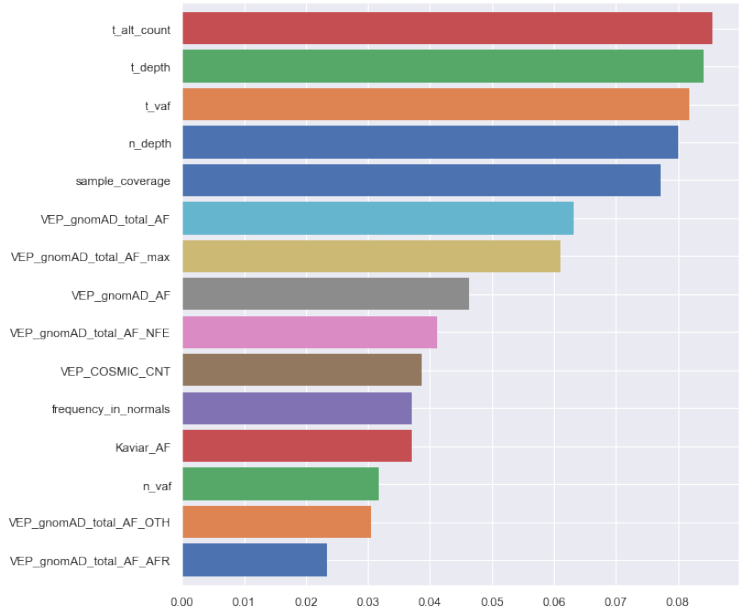
Random Forest (1000 trees)

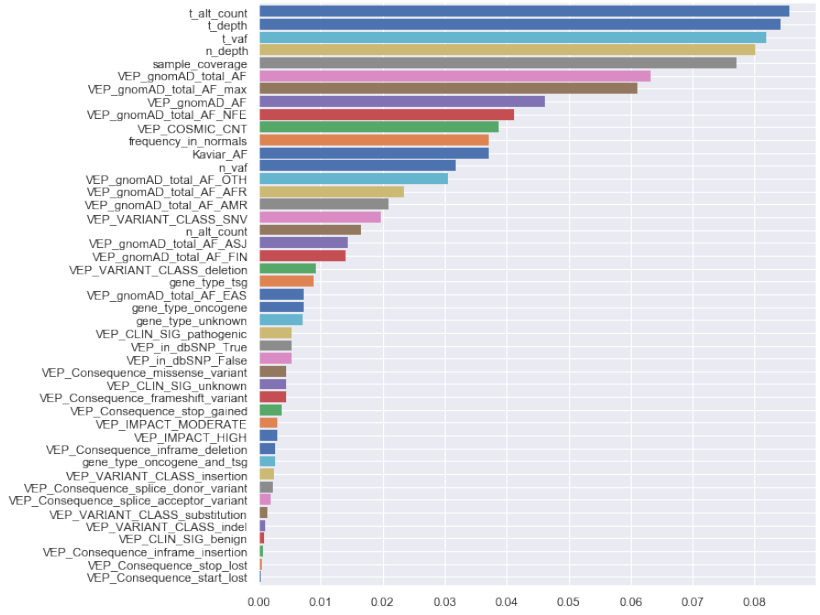




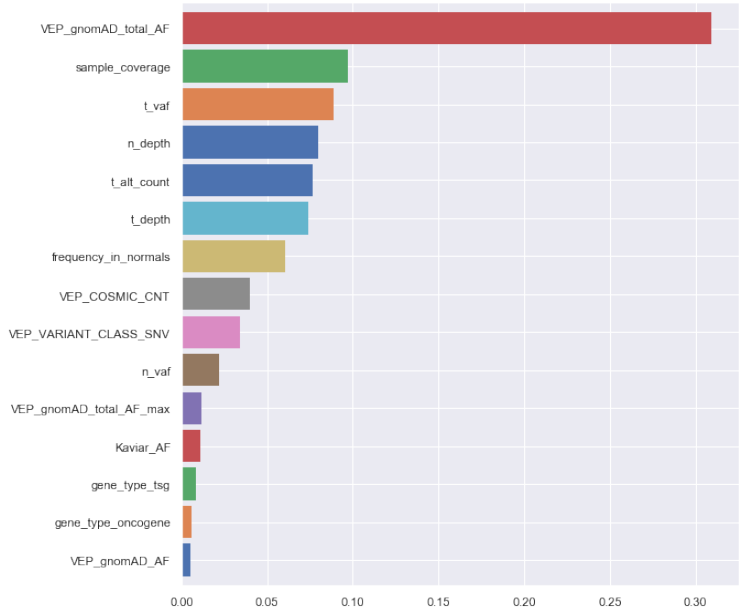
Individual features importance

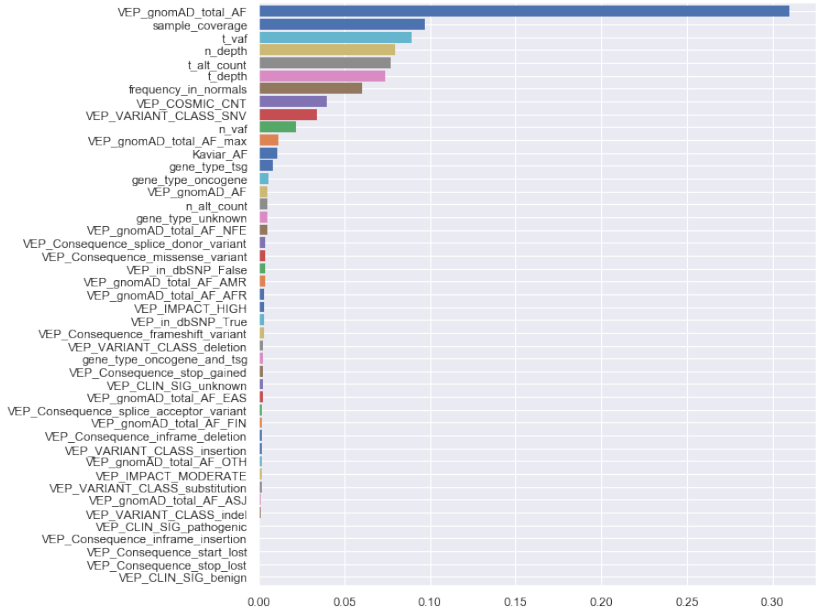
Random Forest (1000 trees)



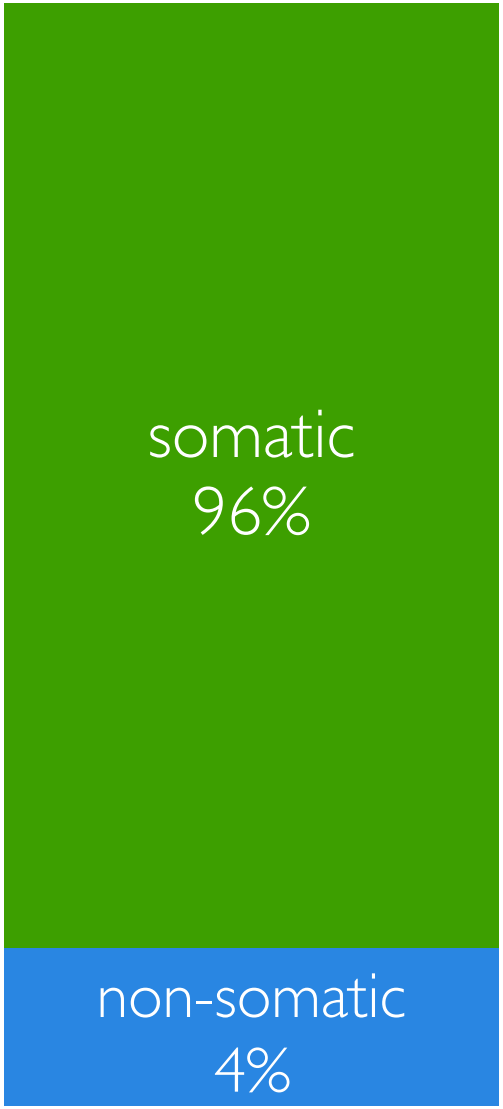


Gradient Boosting (1000 trees)

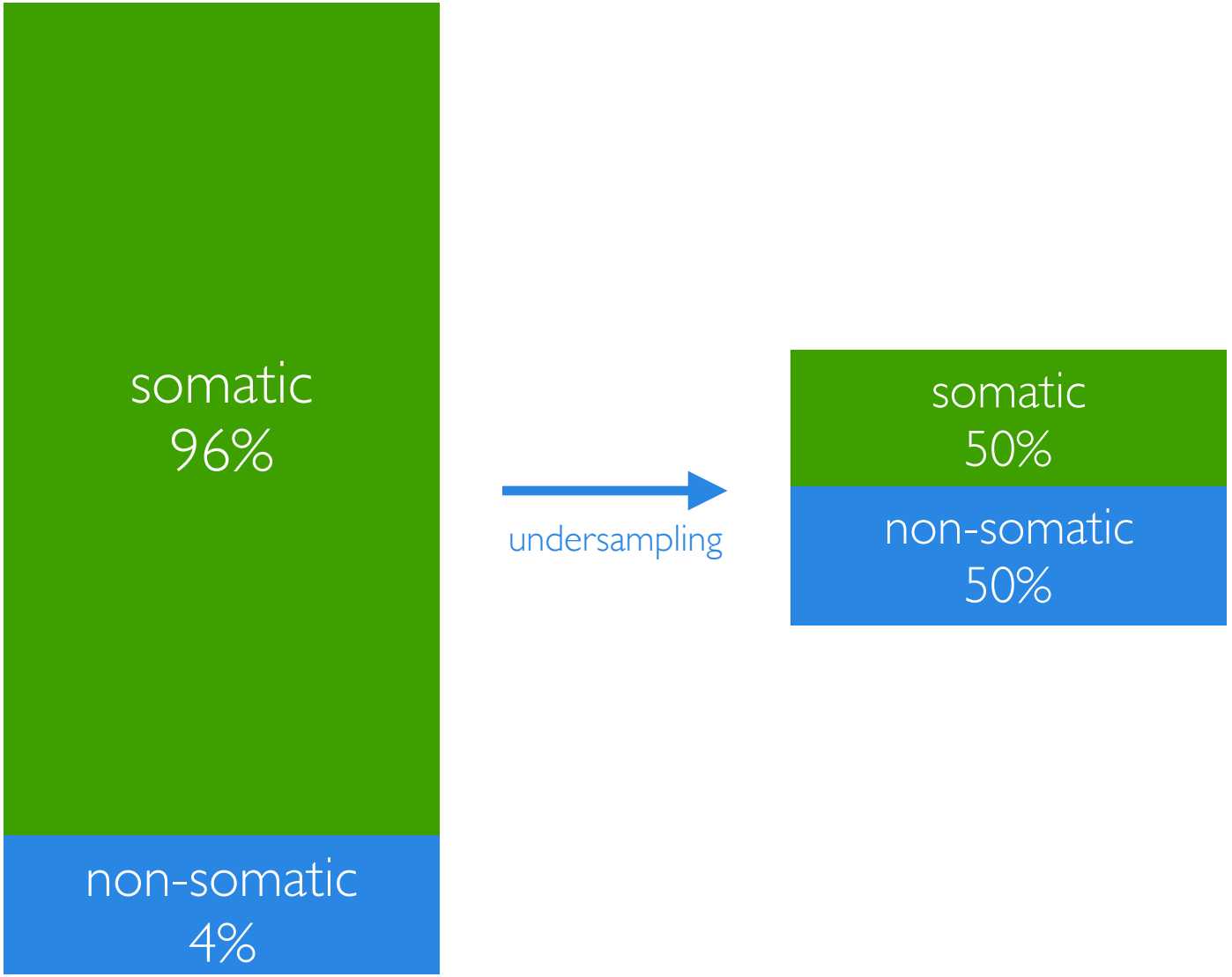




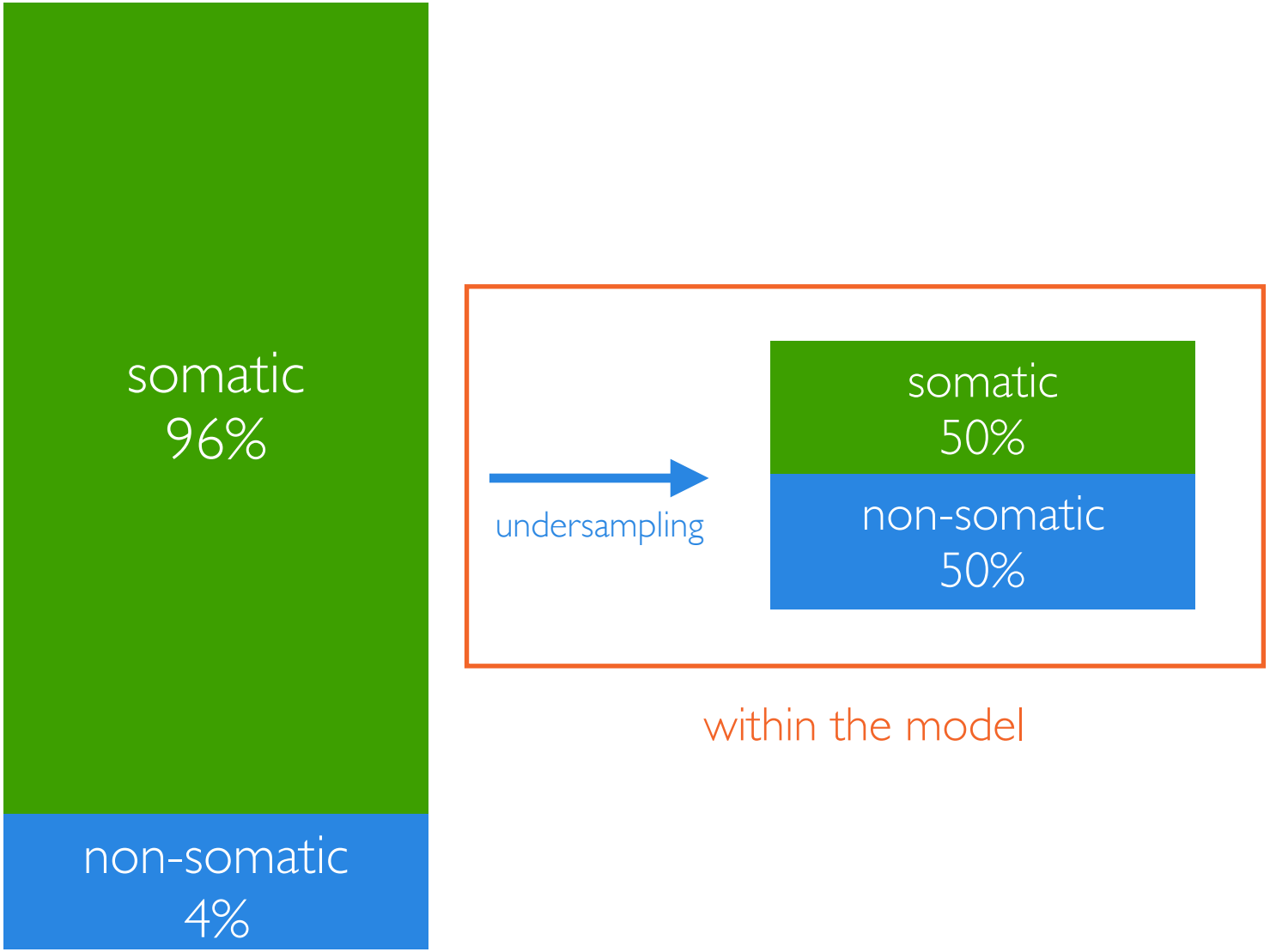
Imbalanced learning



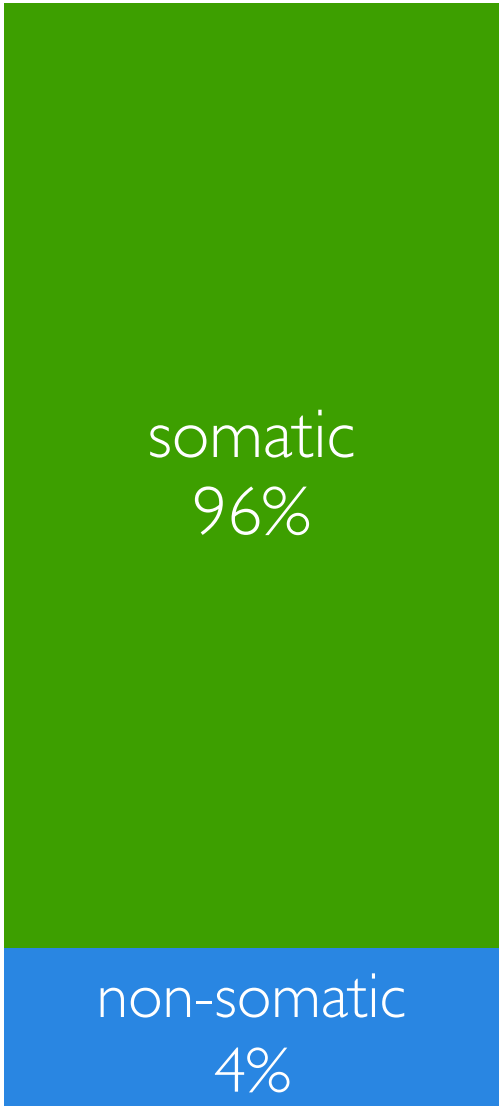
Imbalanced learning

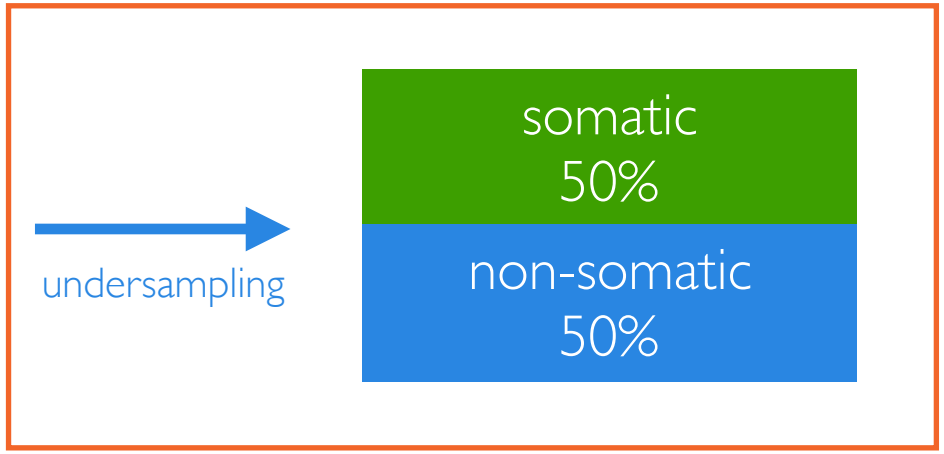


Imbalanced learning



Imbalanced learning

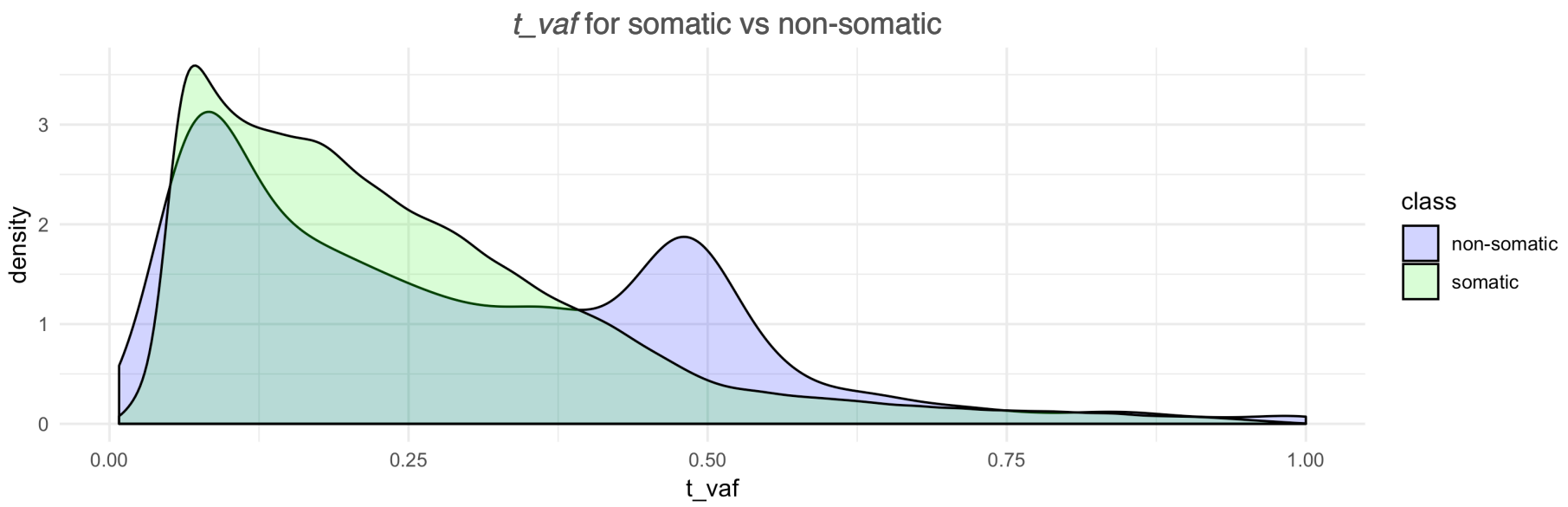




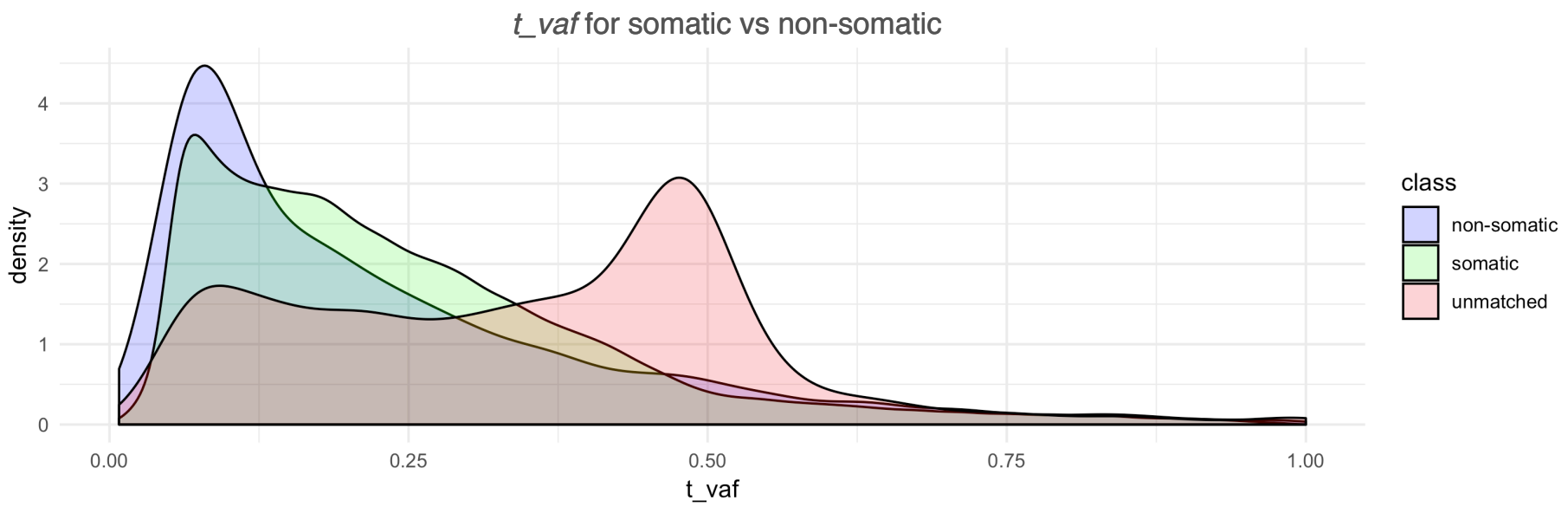
within the model

So the dataset used to train the model is only 8% of the whole dataset (7,199 real and 7,199 artefact mutations)

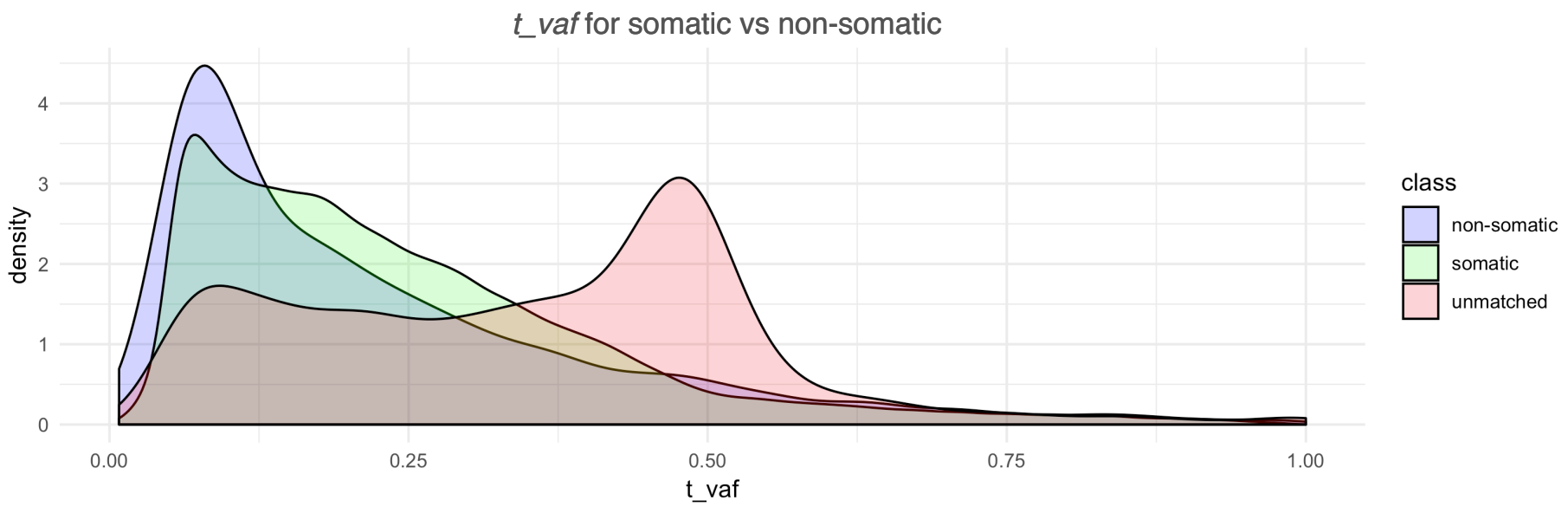
Removing the unmatched samples



Removing the unmatched samples



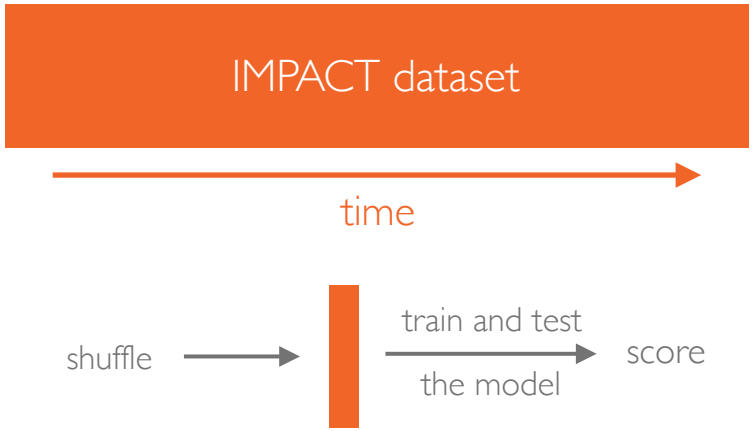
Removing the unmatched samples



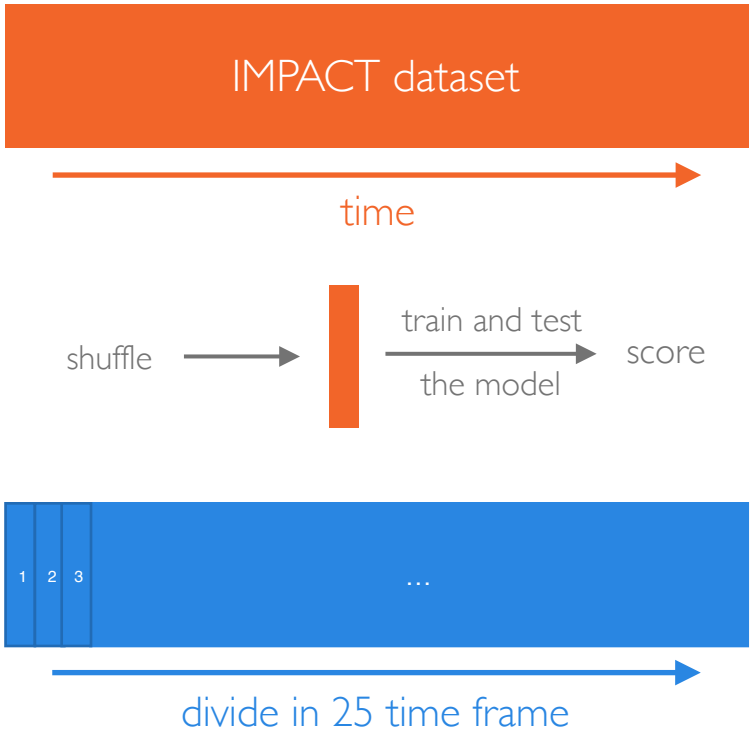
Without the unmatched artefacts (and missing rows):
4,477 artefacts (instead of 7,199)

The sorted dataset phenomenon

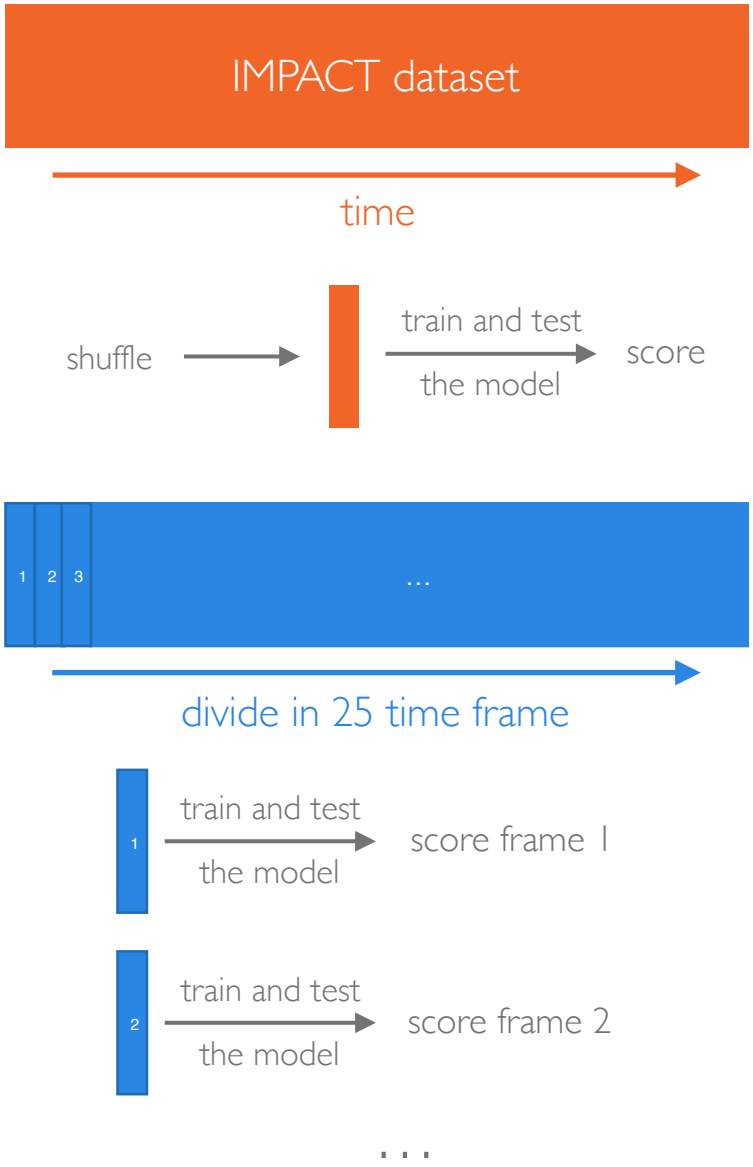
The sorted dataset phenomenon



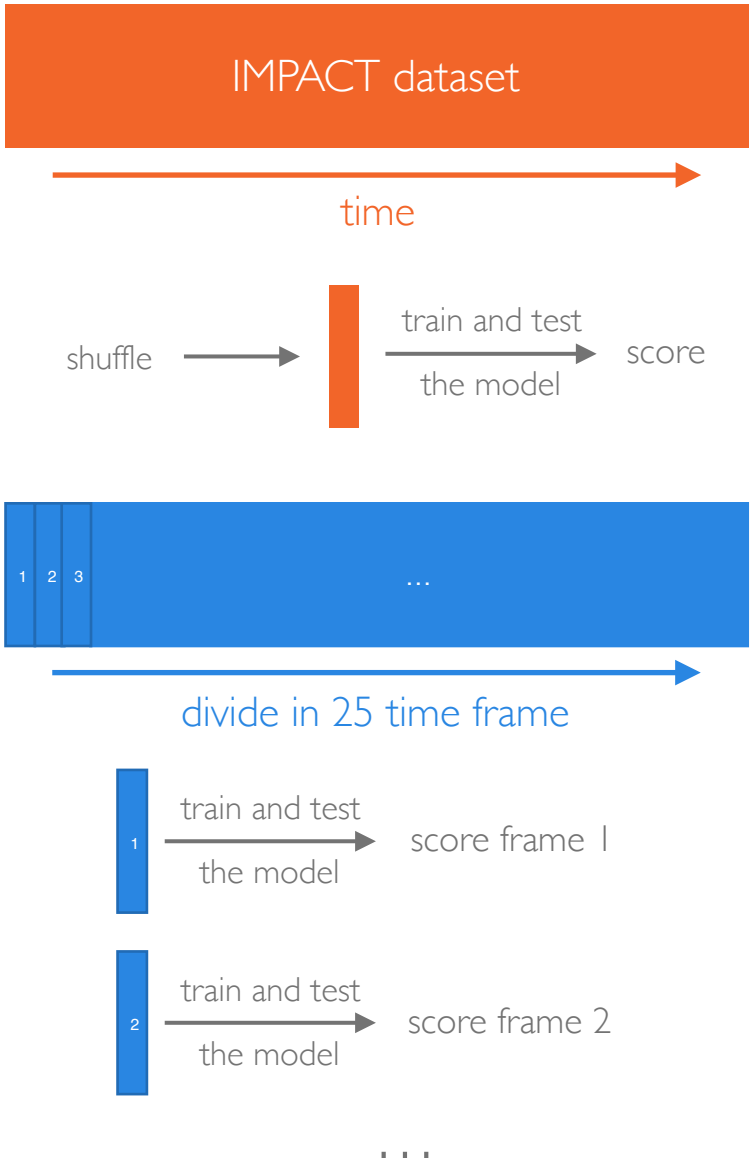
The sorted dataset phenomenon

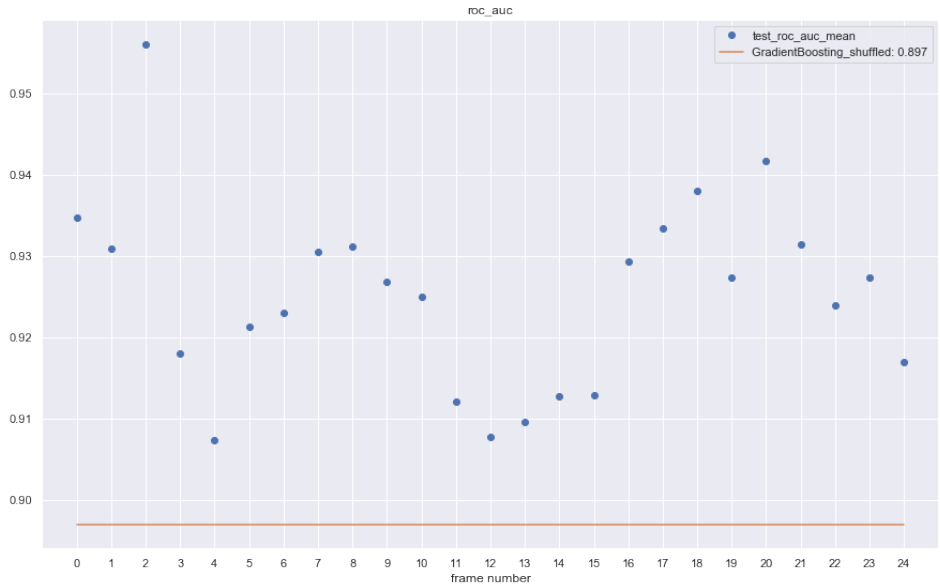


The sorted dataset phenomenon



The sorted dataset phenomenon





The sorted dataset phenomenon