Memorial Sloan Kettering
Cancer Center

IMPACT annotator

September 26, 2018

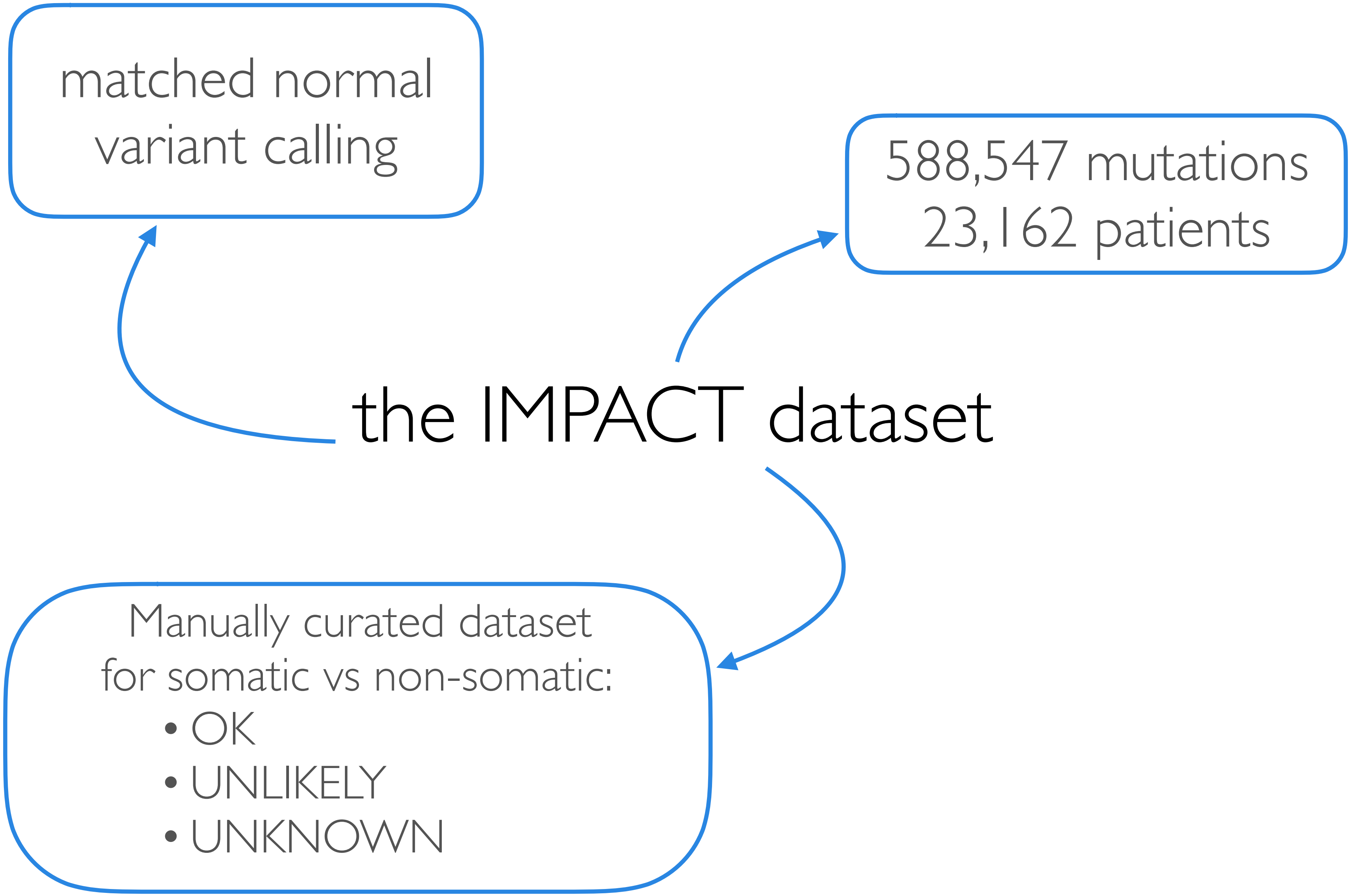
Pierre Guilmin
Elsa Bernard

Goal

Create a tool that classifies variant automatically

- somatic vs non-somatic OR driver vs passenger
- using Supervised Machine Learning Classification
- on the IMPACT dataset

IMPACT



matched normal variant calling

588,547 mutations
23,162 patients

the IMPACT dataset

Manually curated dataset for somatic vs non-somatic:

- OK
- UNLIKELY
- UNKNOWN

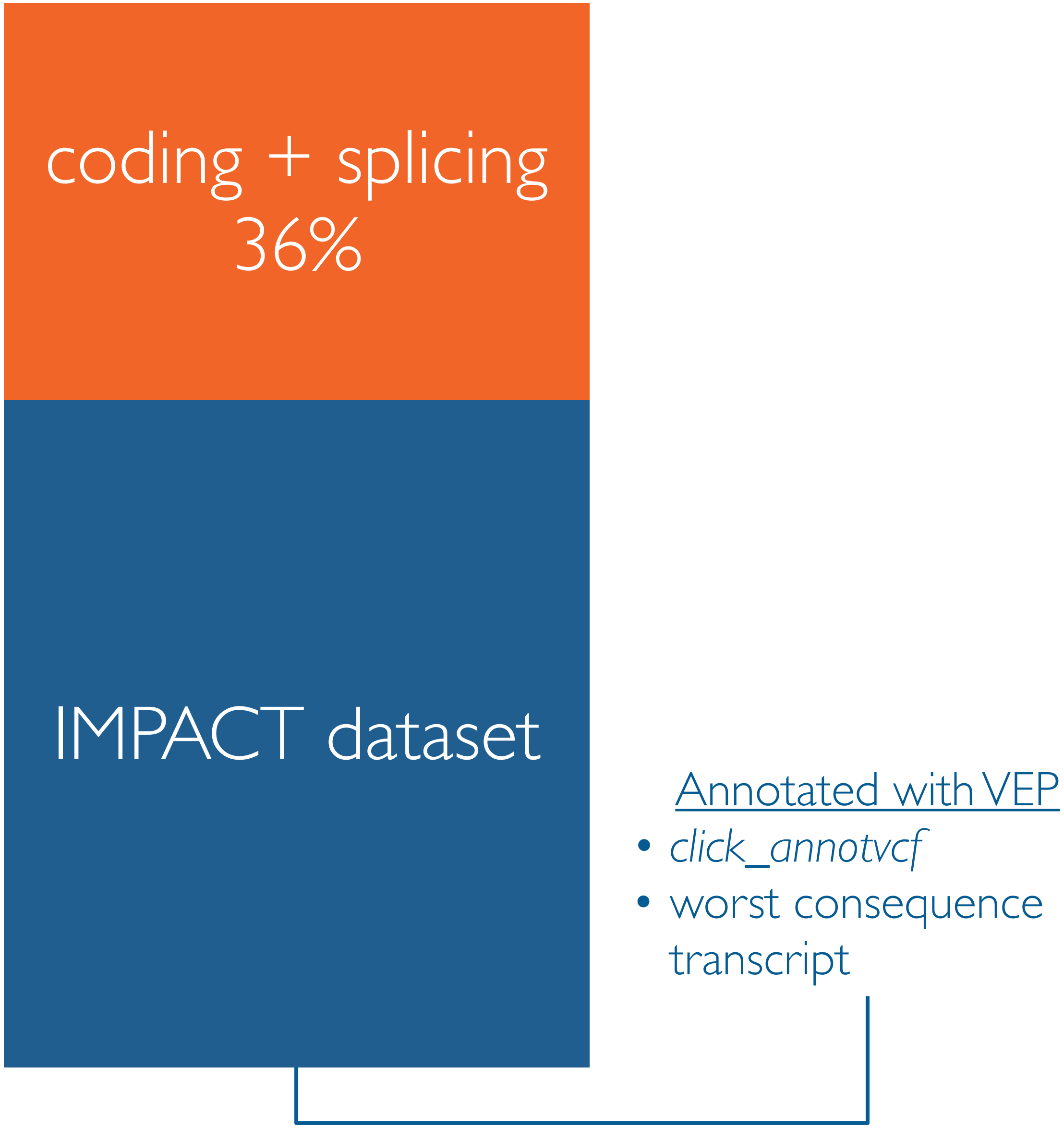
The IMPACT sub-dataset used

IMPACT dataset

Annotated with VEP

- *click_annotvcf*
- worst consequence transcript

The IMPACT sub-dataset used



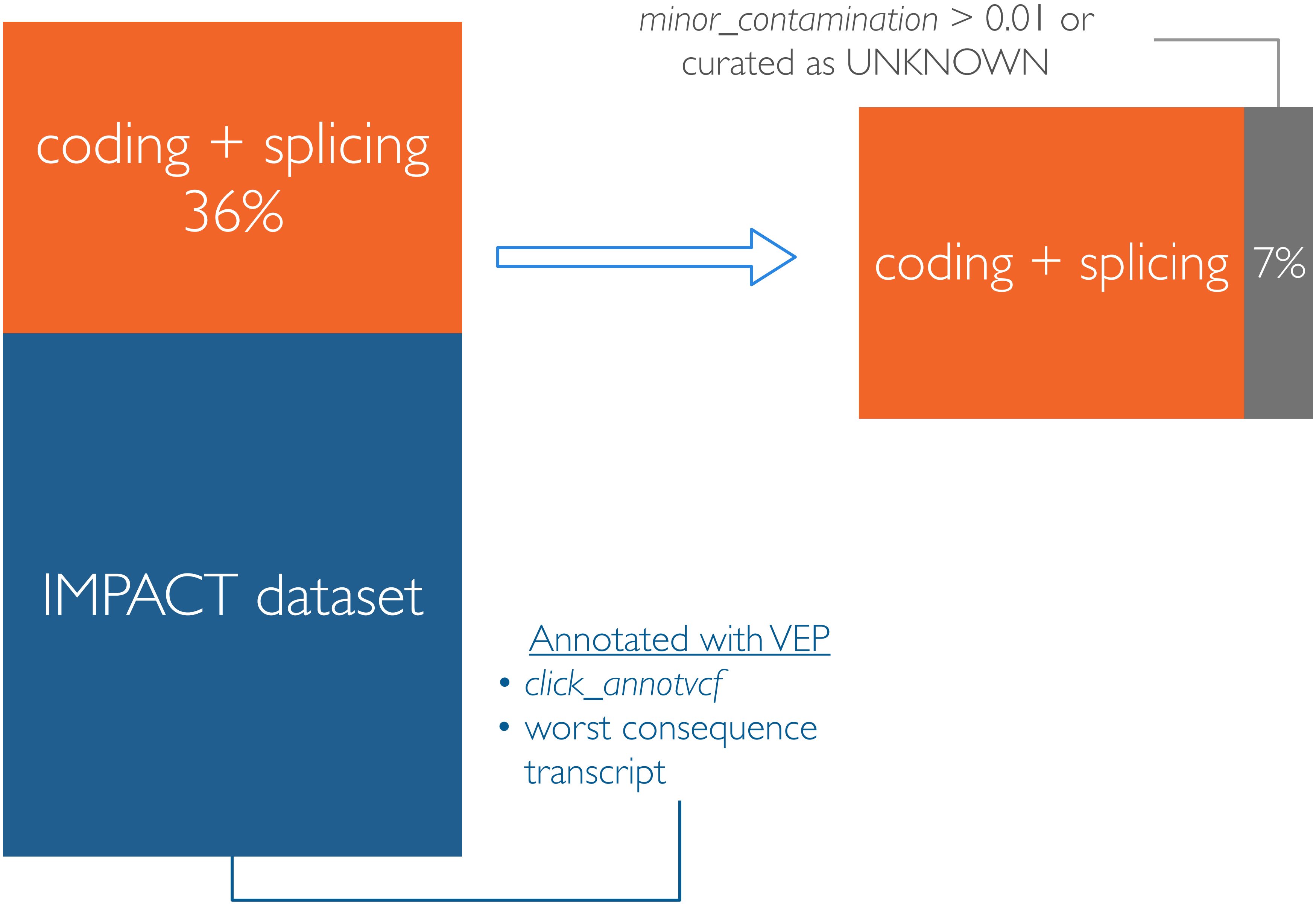
coding + splicing
36%

IMPACT dataset

Annotated with VEP

- *click_annotvcf*
- worst consequence transcript

The IMPACT sub-dataset used



coding + splicing 36%

IMPACT dataset

Annotated with VEP

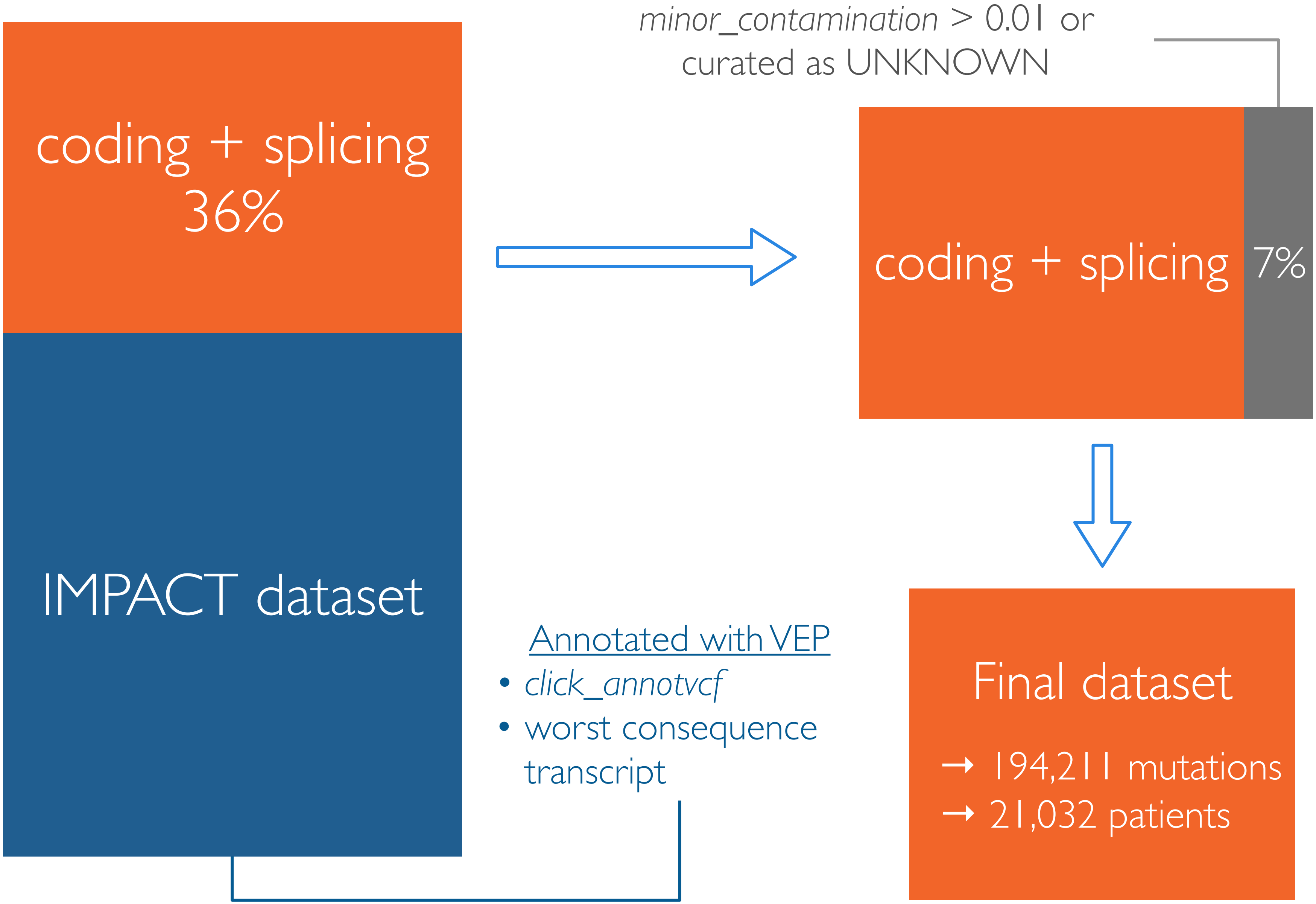
- *click_annotvcf*
- worst consequence transcript

coding + splicing

7%

minor_contamination > 0.01 or curated as UNKNOWN

The IMPACT sub-dataset used



coding + splicing
36%

IMPACT dataset

minor_contamination > 0.01 or
curated as UNKNOWN

coding + splicing

7%

Annotated with VEP

- *click_annotvcf*
- worst consequence transcript

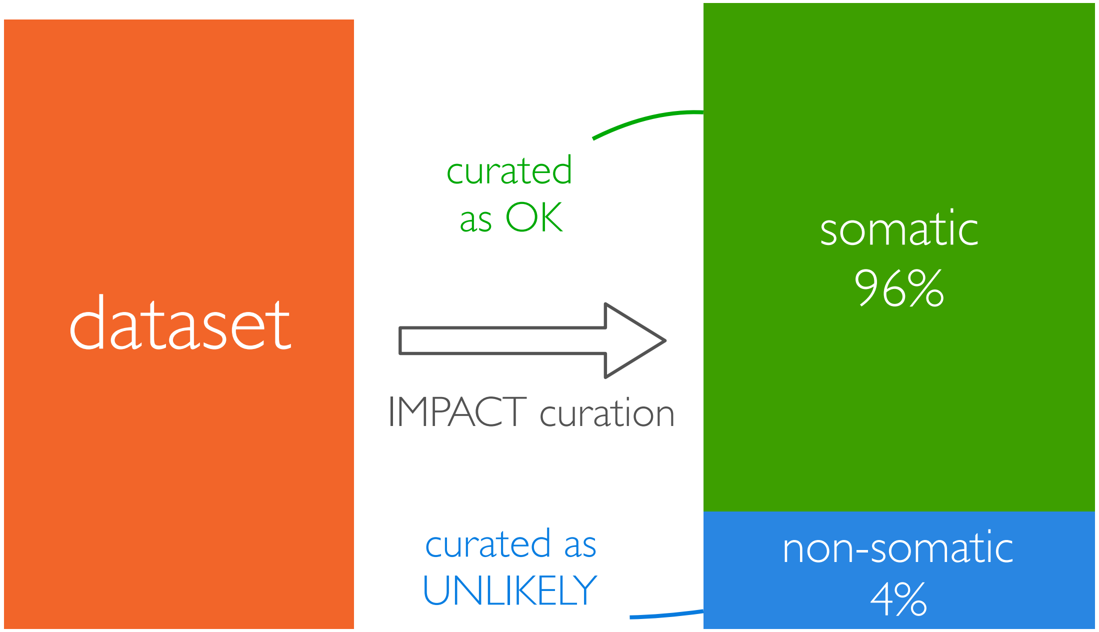
Final dataset

→ 194,211 mutations
→ 21,032 patients

Different classes

dataset

Different classes



dataset

curated as OK

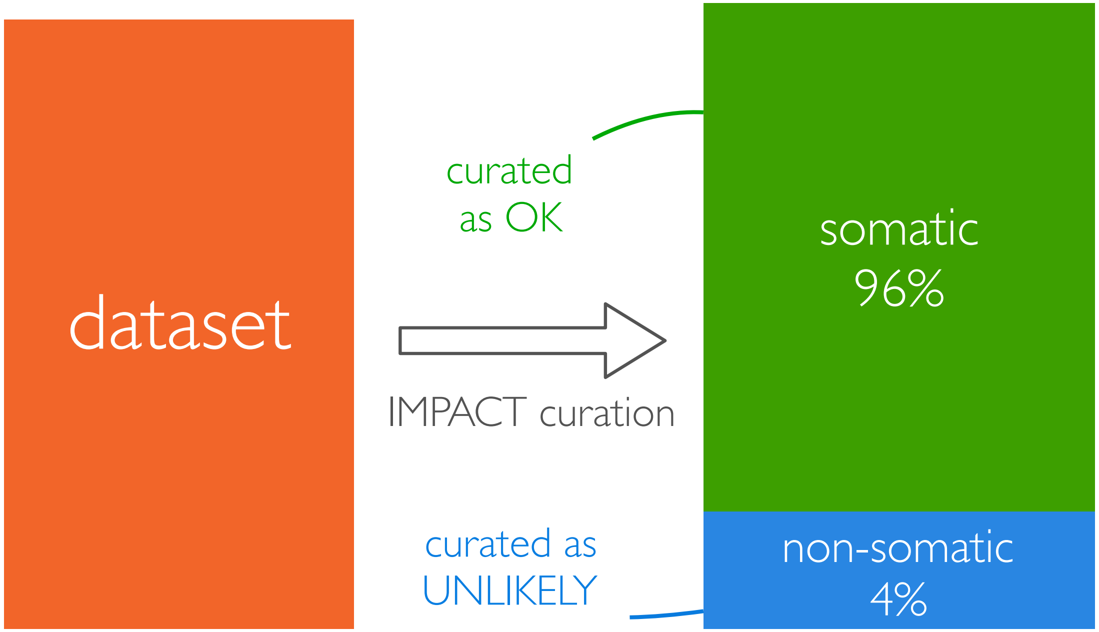
IMPACT curation

curated as UNLIKELY

somatic 96%

non-somatic 4%

Different classes



dataset

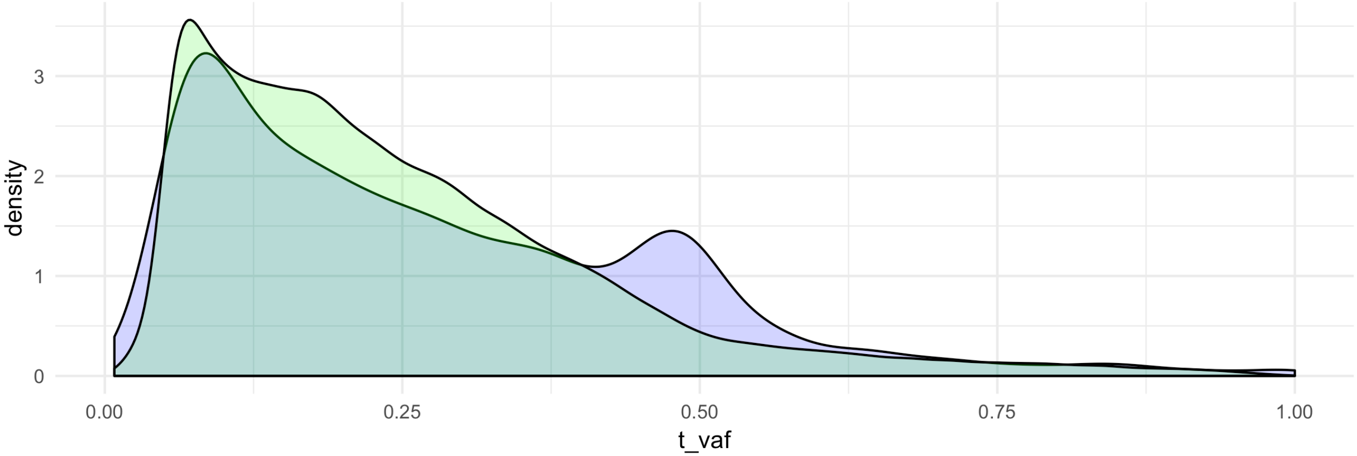
IMPACT curation

curated as OK

somatic 96%

curated as UNLIKELY

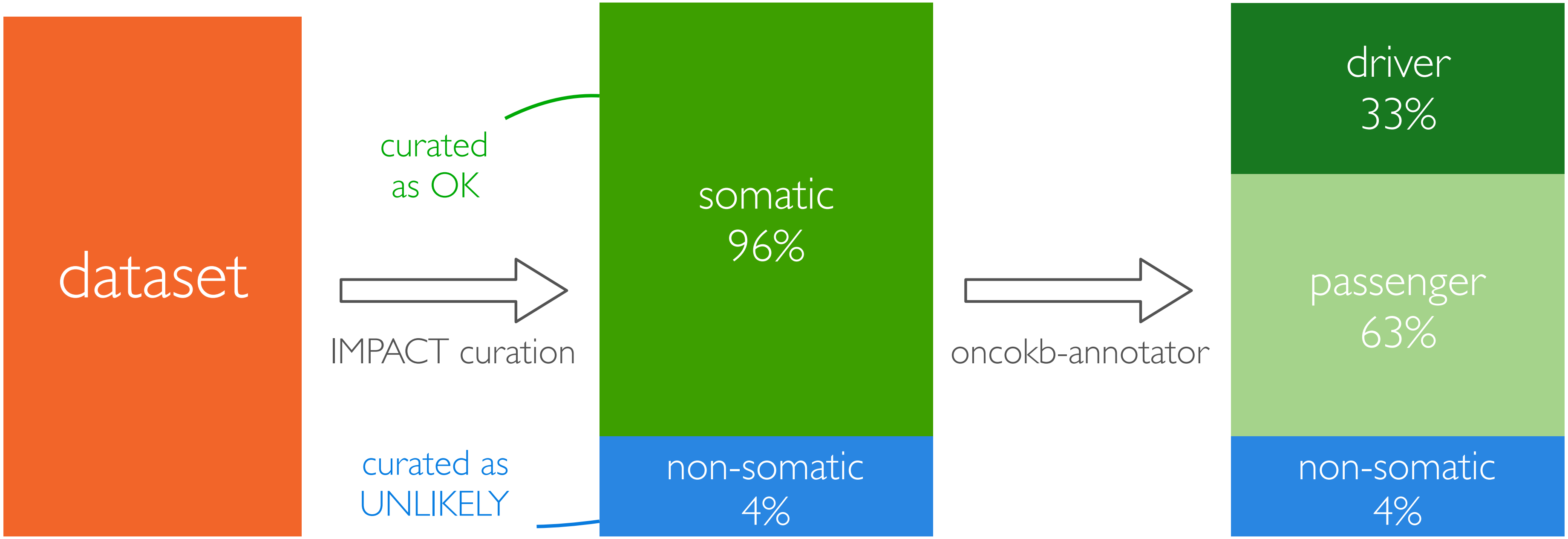
non-somatic 4%



Tumor VAF for somatic vs non-somatic

non-somatic = artefact + germline

Different classes



dataset

curated as OK

IMPACT curation

curated as UNLIKELY

somatic 96%

non-somatic 4%

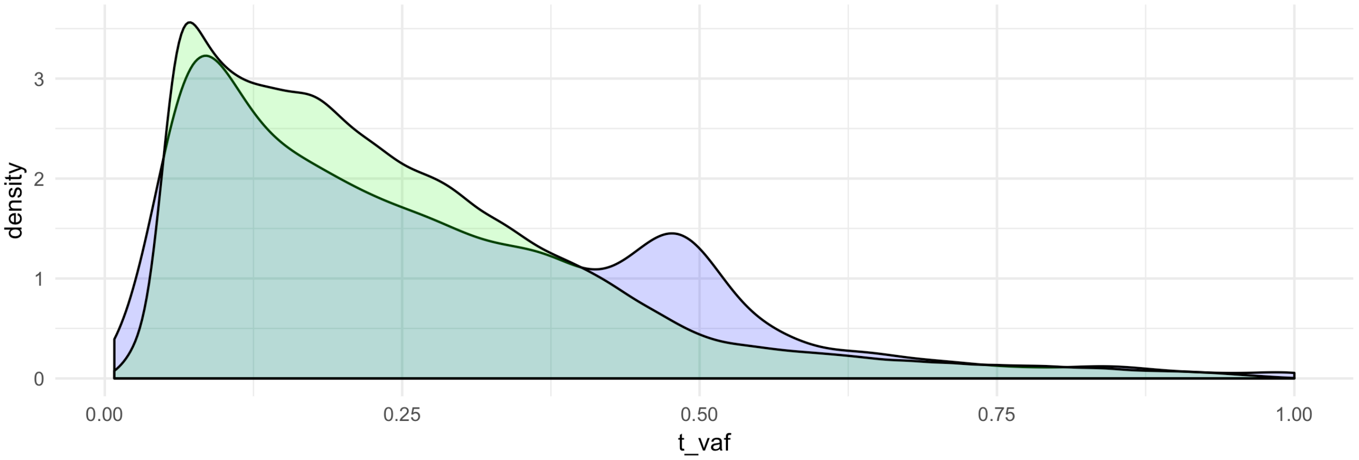
oncokb-annotator

driver 33%

passenger 63%

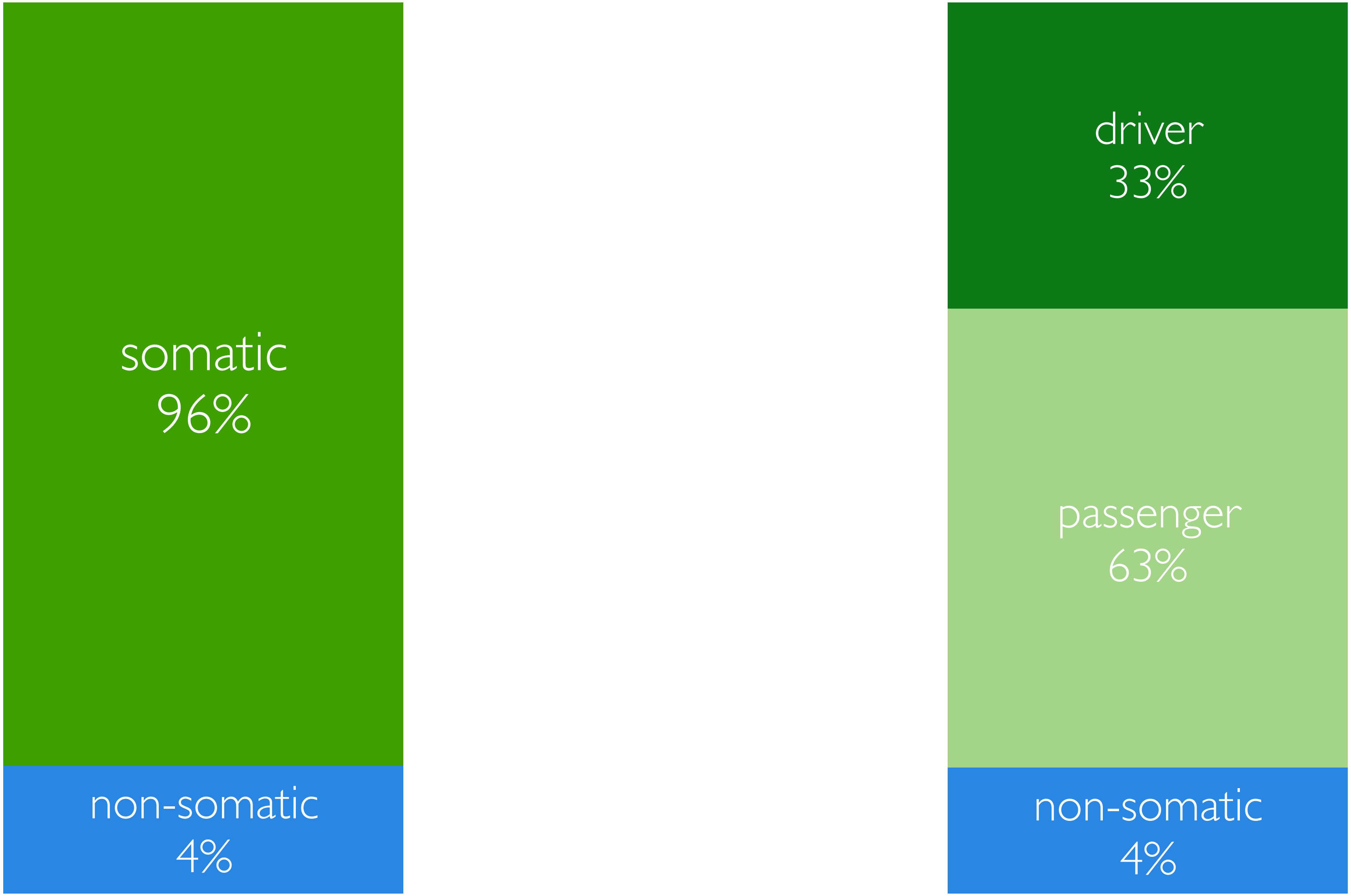
non-somatic 4%

Tumor VAF for somatic vs non-somatic

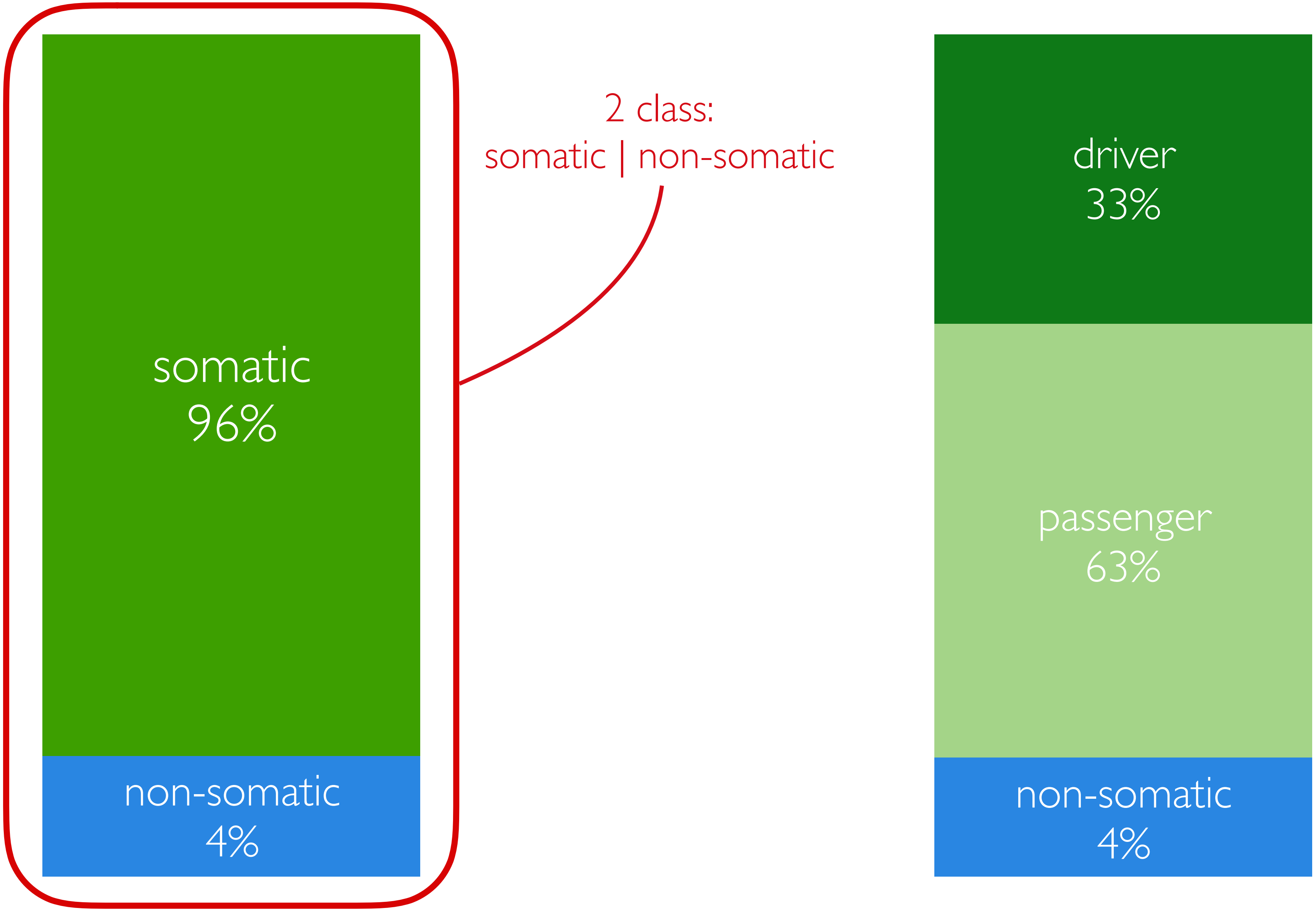


→ non-somatic = artefact + germline

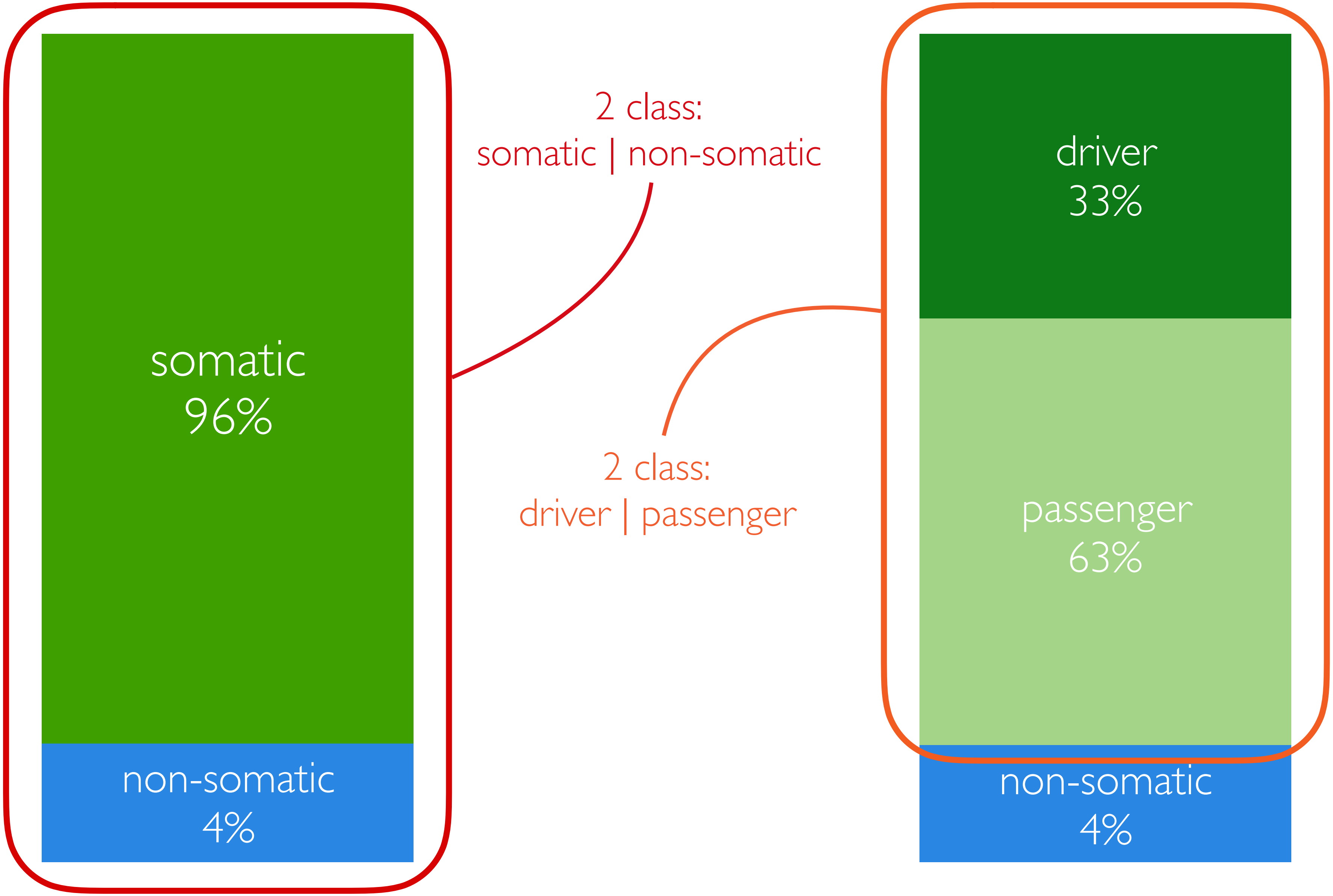
Supervised classification



Supervised classification



Supervised classification



somatic
96%

non-somatic
4%

2 class:
somatic | non-somatic

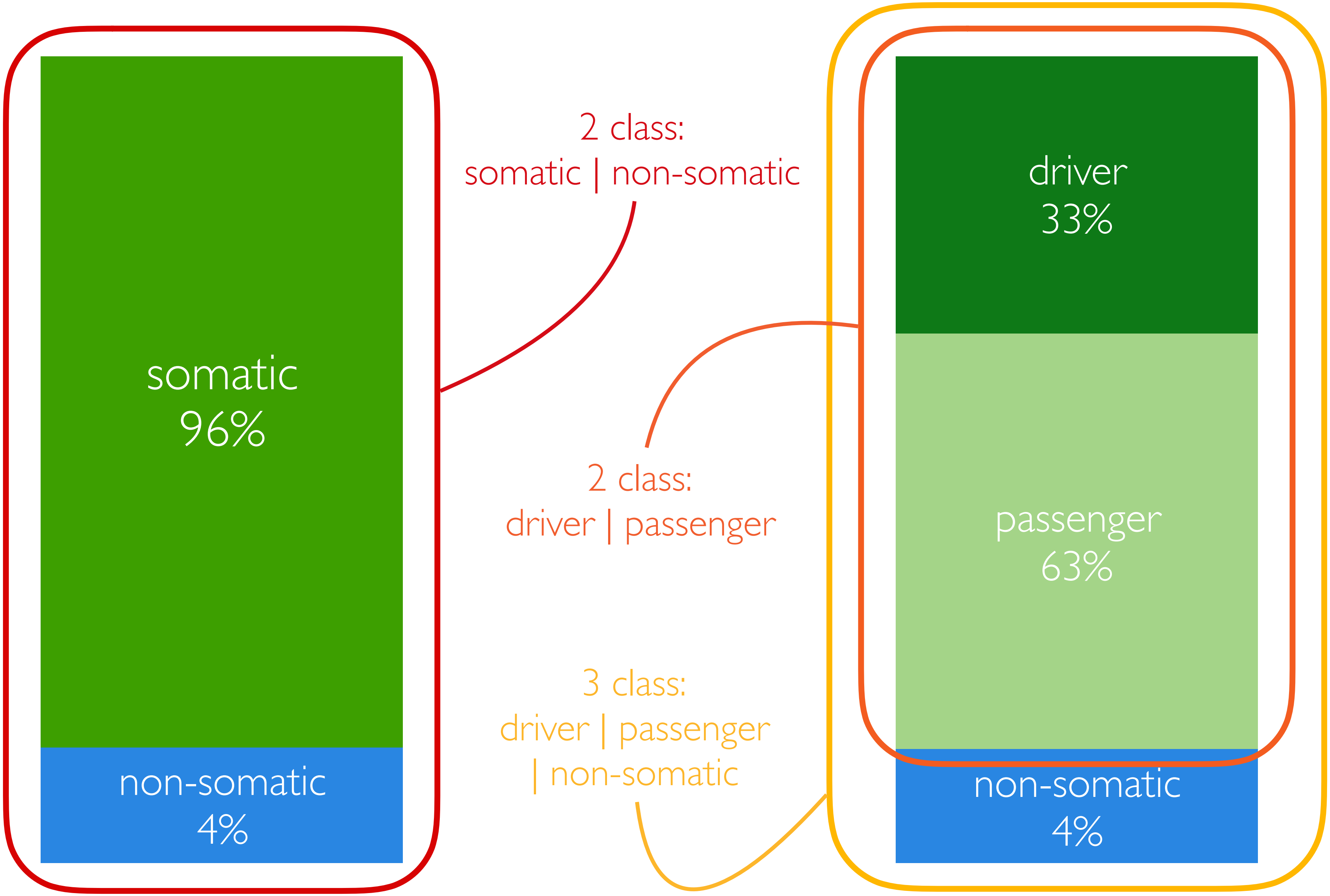
2 class:
driver | passenger

driver
33%

passenger
63%

non-somatic
4%

Supervised classification



somatic
96%

non-somatic
4%

2 class:
somatic | non-somatic

2 class:
driver | passenger

3 class:
driver | passenger
| non-somatic

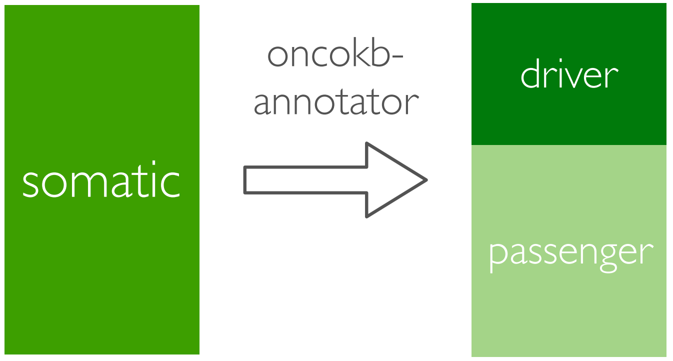
driver
33%

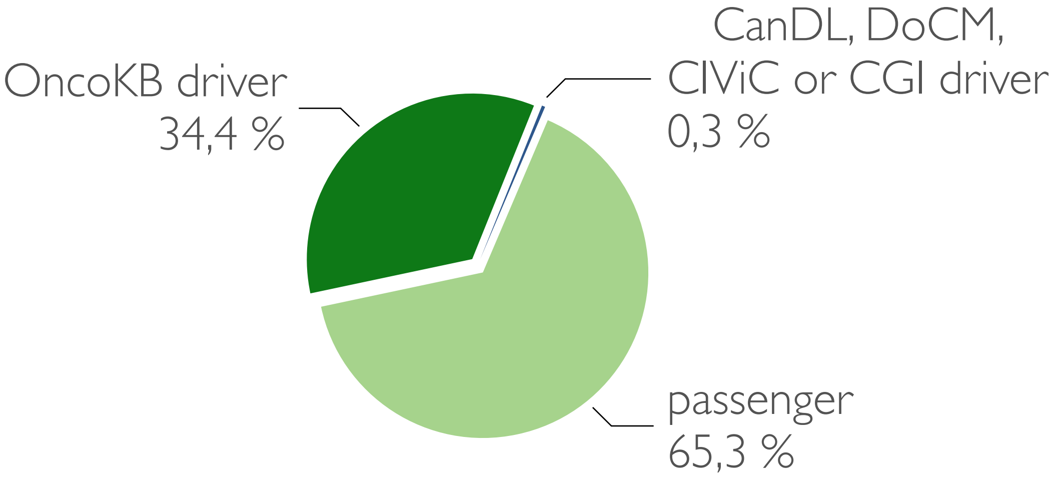
passenger
63%

non-somatic
4%

Literature

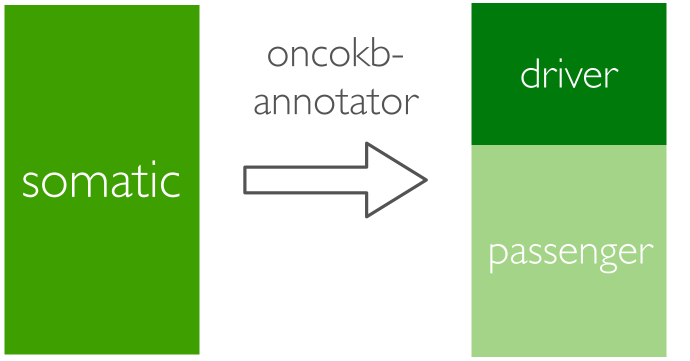
Variant annotation databases

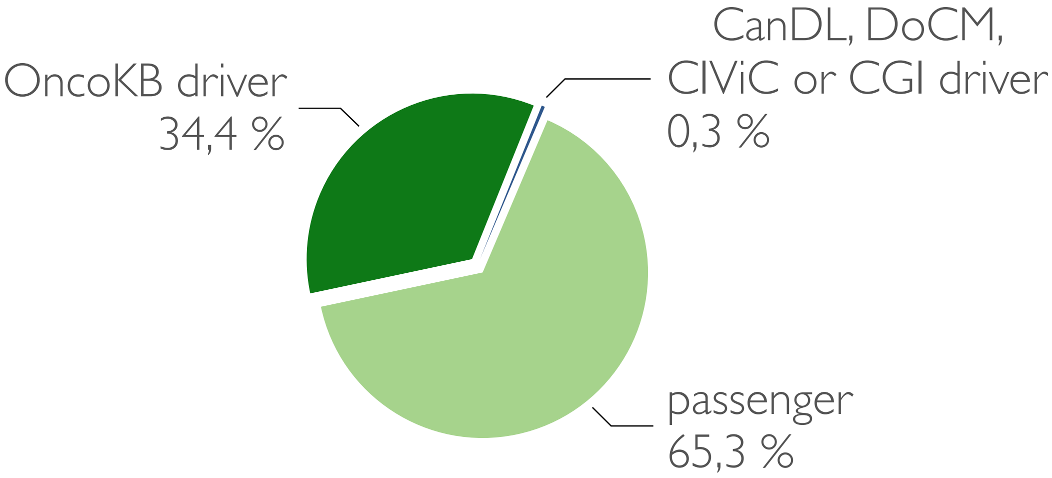




Literature

Variant annotation databases





Existing algorithms

→ SIFT, PolyPhen-2: predicts whether an amino-acid substitution affects protein function

- not cancer specific

→ CHASM, FATHMM, CanDrA, CScape, rDriver, …: supervised learning for driver classification

- no sequencing features (features: SIFT, PolyPhen-2, conservative features, genetic features, other prediction algorithm features…)
- smaller and not "real" dataset (driver: COSMIC / passenger: synthetic, dbSNP, …)
- never done on somatic vs non-somatic: applied on real somatic mutations only

non-synonymous SNVs only (no indels)

The features

NGS features

- VAF
- strand ratio
- …

Frequency in normals

Population AF

- Kaviar
- gnomAD

COSMIC score

Impact scores

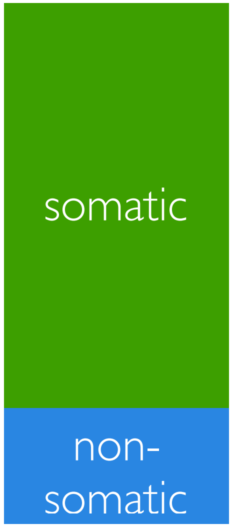
- SIFT
- PolyPhen-2
- …

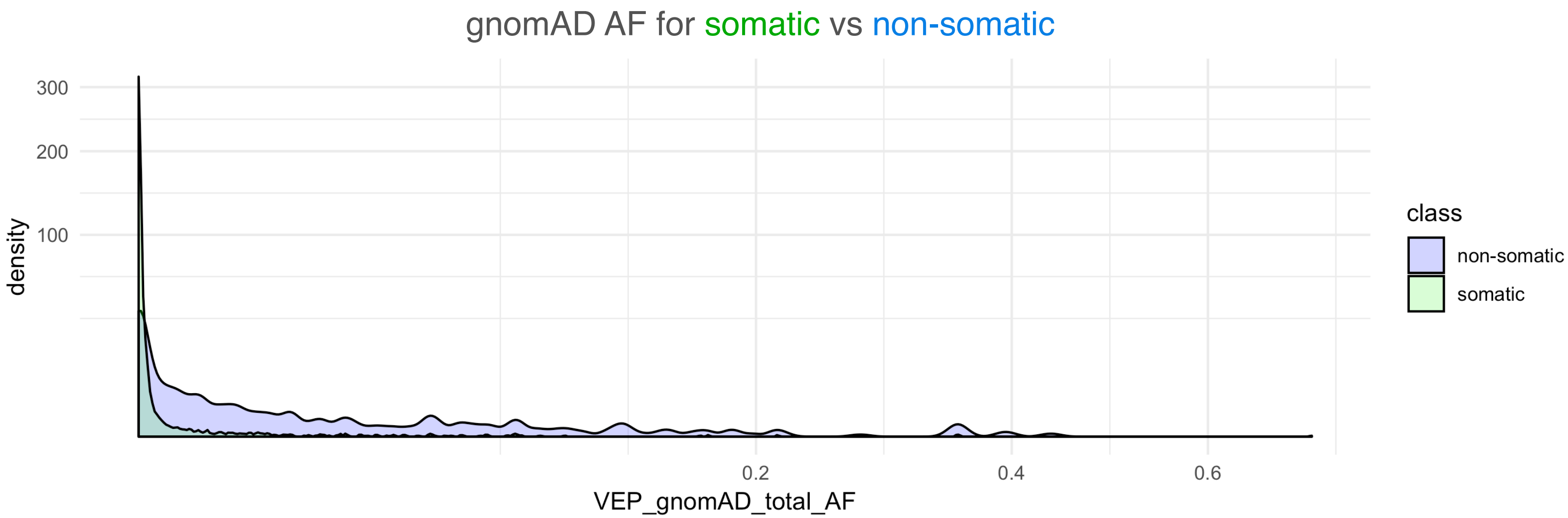
Genetic features

- gene and gene type
- mutation effect
- …

Any ideas?

Example of informative feature

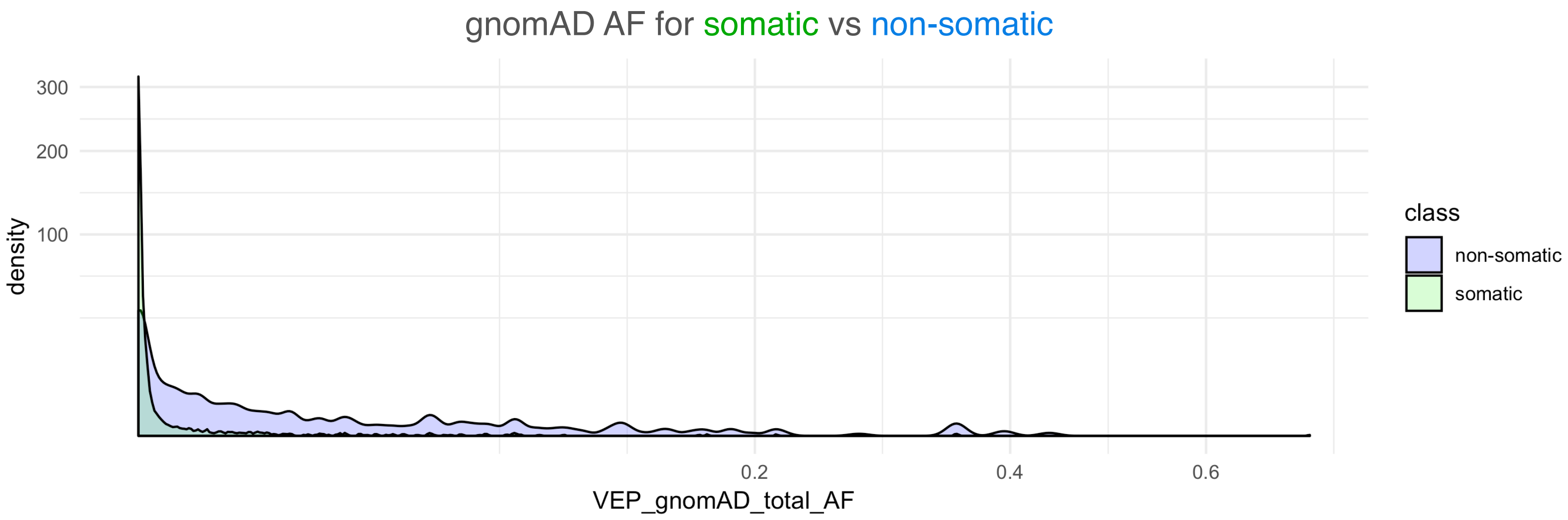



gnomAD AF for somatic vs non-somatic

Example of informative feature

somatic

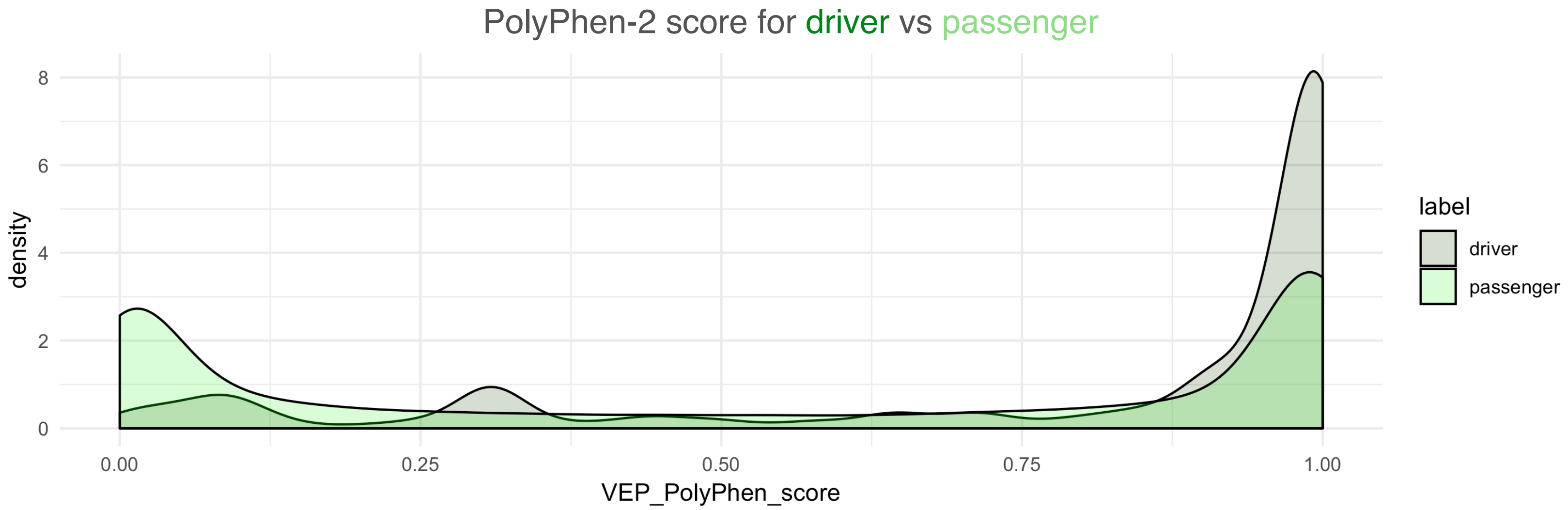
non-somatic



gnomAD AF for somatic vs non-somatic

driver

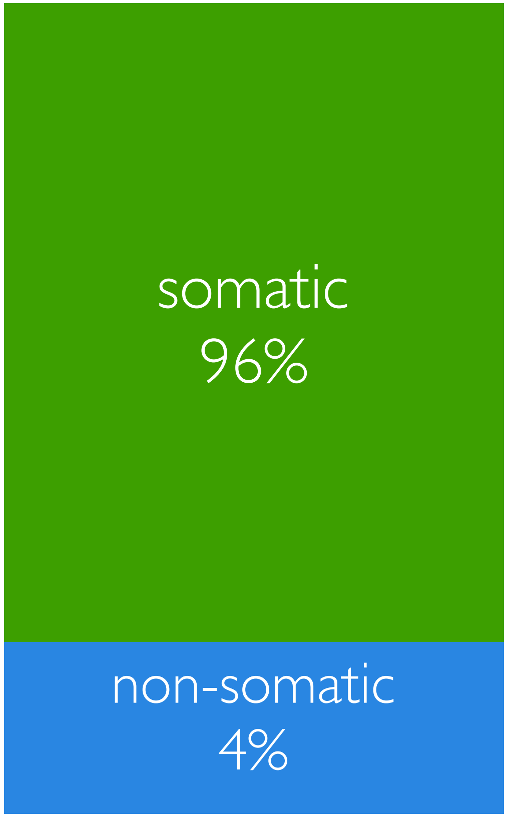
passenger



PolyPhen-2 score for driver vs passenger

Dataset complexity for machine learning

Inconsistencies

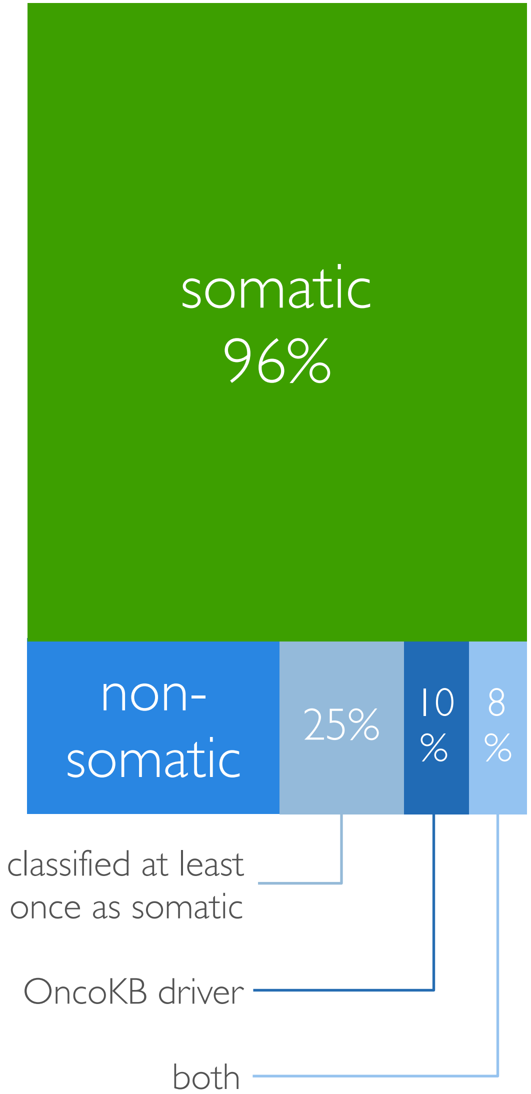


somatic
96%

non-somatic
4%

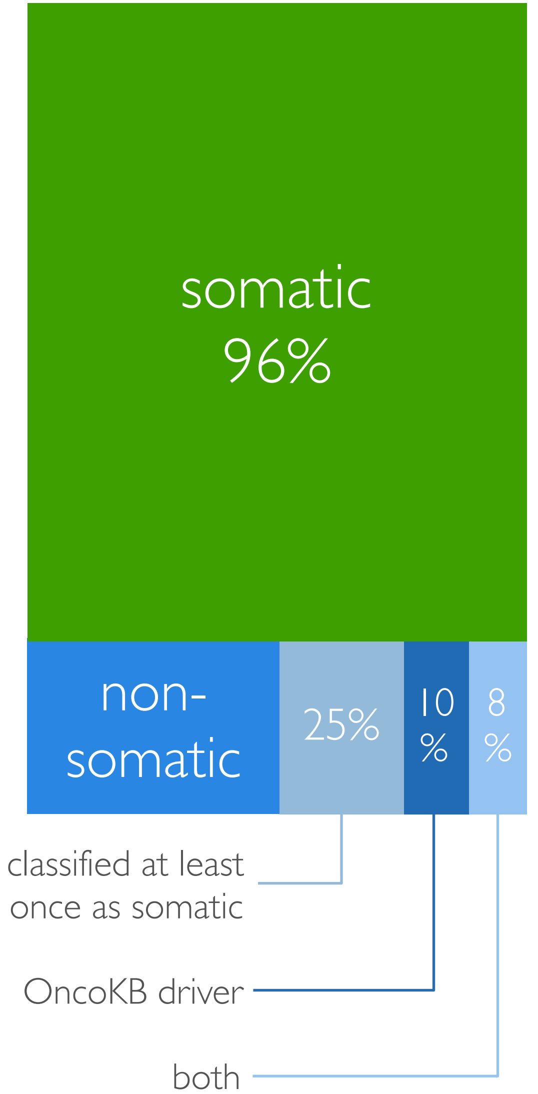
Dataset complexity for machine learning

Inconsistencies

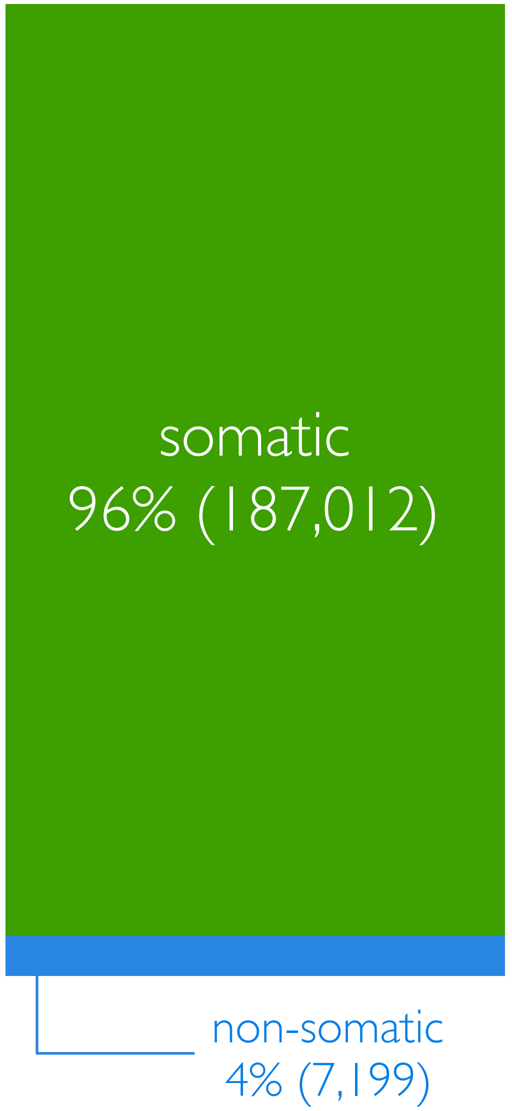


Dataset complexity for machine learning

Inconsistencies

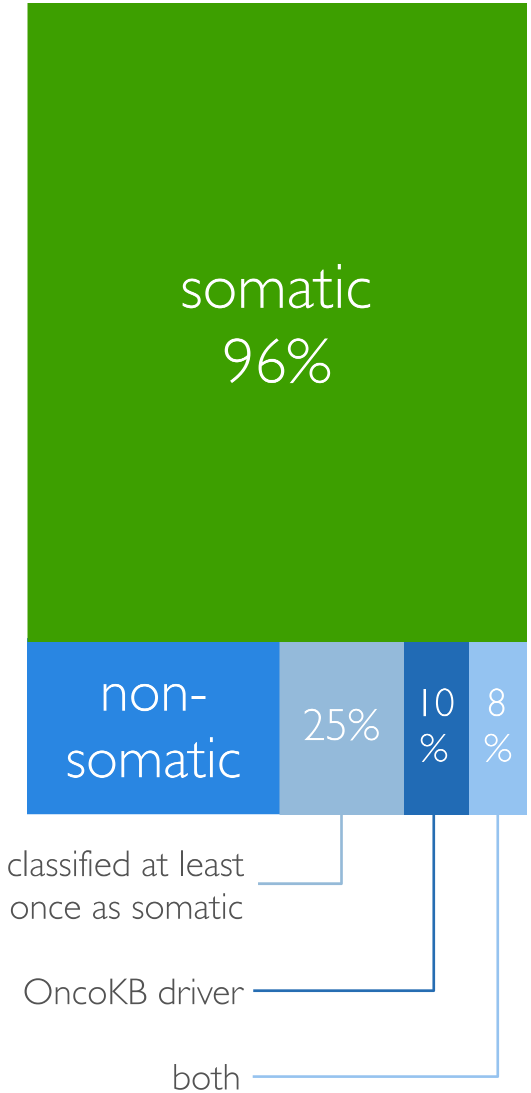


Imbalanced dataset

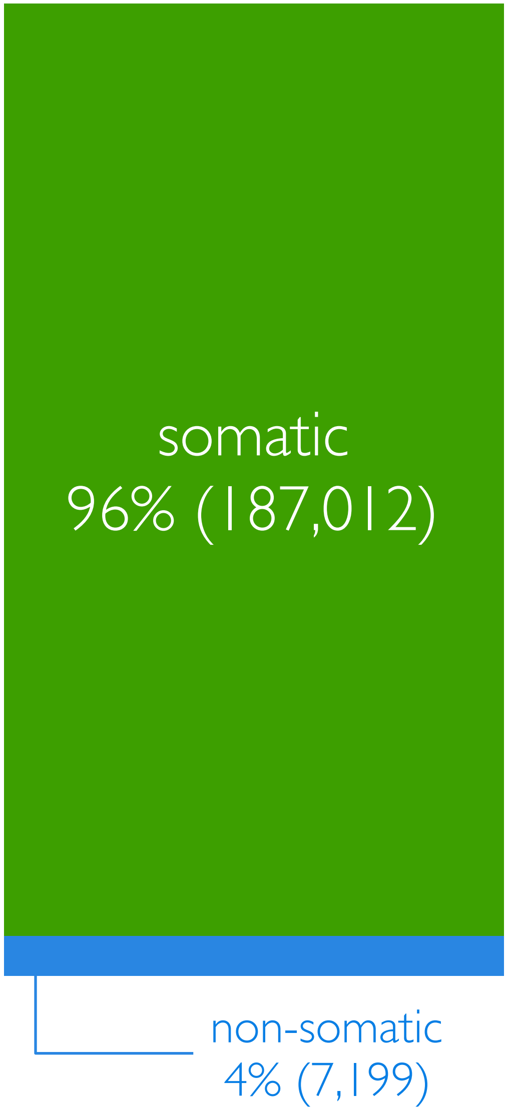


Dataset complexity for machine learning

Inconsistencies



Imbalanced dataset



Highly recurrent mutations

78% of the driver mutations are unique across the dataset

BUT

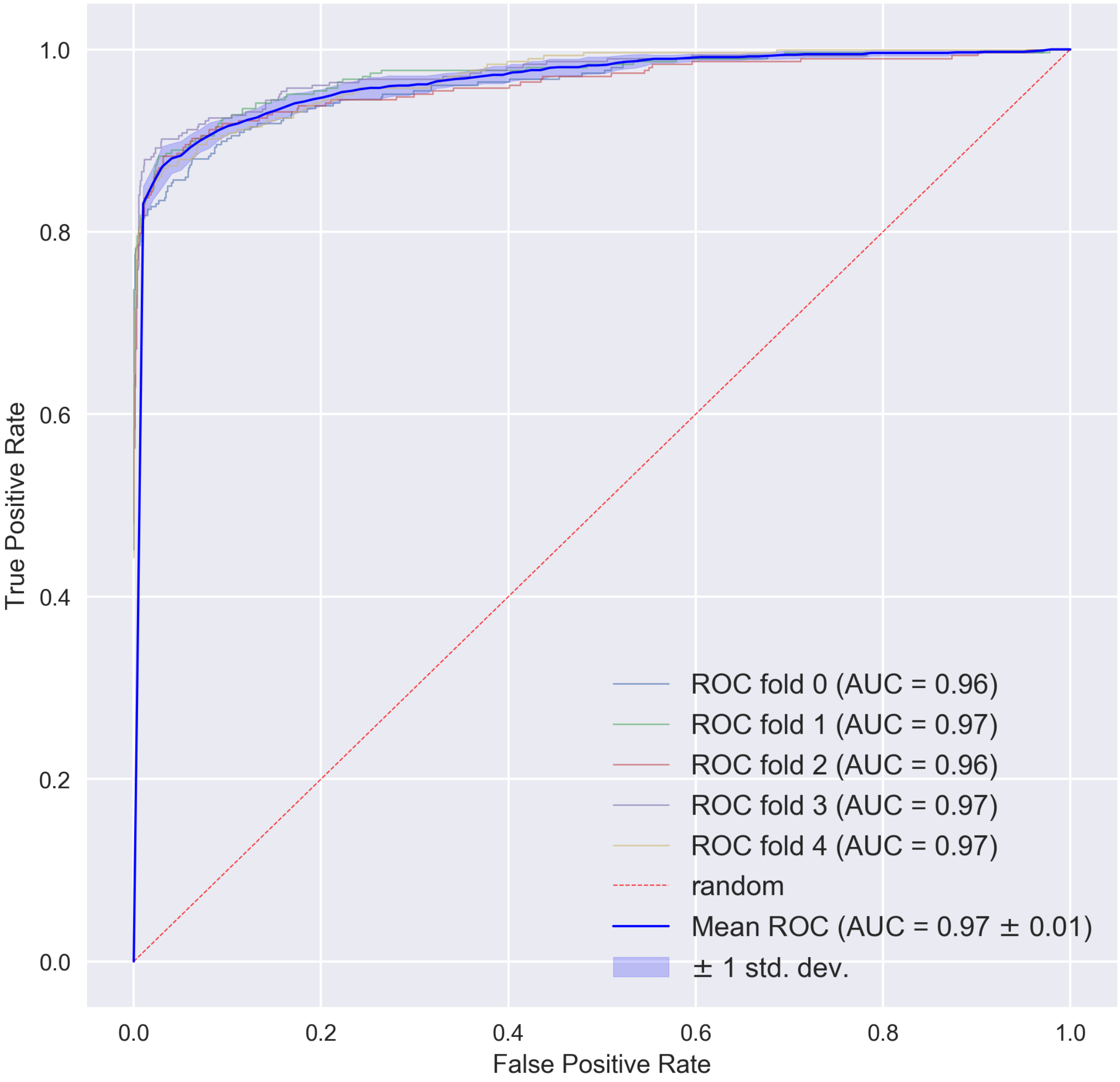
100 mutations are shared by 45 patients or more
(ex: KRAS p.G12D, >1000 patients)

↓

- Overfitting on hotspots?
- Performance on « rare » drivers?

First results

- not representative yet
- logistic regression (l2-regularization | ridge regression)
- 10,000 samples
- 5-fold stratified cross-validation
- driver vs passenger


True Positive Rate
False Positive Rate
ROC fold 0 (AUC = 0.96)
ROC fold 1 (AUC = 0.97)
ROC fold 2 (AUC = 0.96)
ROC fold 3 (AUC = 0.97)
ROC fold 4 (AUC = 0.97)
random
Mean ROC (AUC = 0.97 ± 0.01)
± 1 std. dev.