

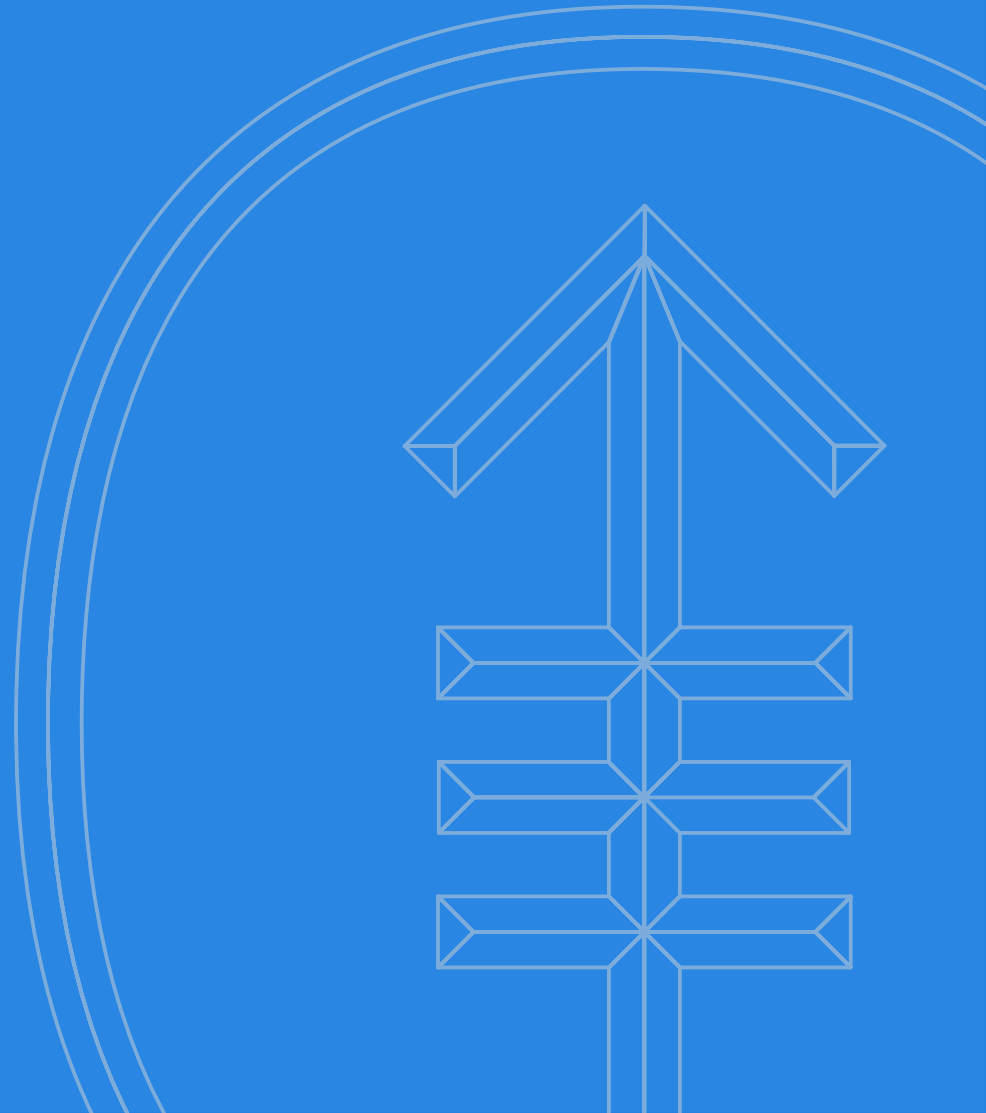


Memorial Sloan Kettering
Cancer Center

IMPACT annotator

September 11, 2018

Pierre Guilmin
Graduate Research Assistant



Recurrent synonymous mutations



# of patients	Description	Gene	Gene Type	HGVSp	COSMIC count
64	17:7579312 C/T ⁽³⁰⁾ & C/A ⁽²⁹⁾ & C/G ⁽⁵⁾	TP53	Tumor suppressor	p.T125T	71 ^(30 & 37 & 4)
21	18:60985833 & 18:60985834 G/A ⁽¹¹⁾ & C/T ⁽¹⁰⁾	BCL2	Oncogene	p.K22K & p.L23L	19 ^(13 & 6)
18	6:26056117 C/T	HIST1H1C	Unknown	p.A180A	8
13	20:41101171 G/A	PTPRT	Tumor suppressor	p.I395I	6
11	20:40790142 C/T	PTPRT	Tumor suppressor	p.T863T	0
11	17:7578177 C/T	TP53	Tumor suppressor	p.E224E	15
11	20:9525085 C/T	PAK5	Unknown	p.R600R	0
10	21:39795342 G/A	ERG	Oncogene	p.I126I	0

last
exon
base

The IMPACT sub-dataset used

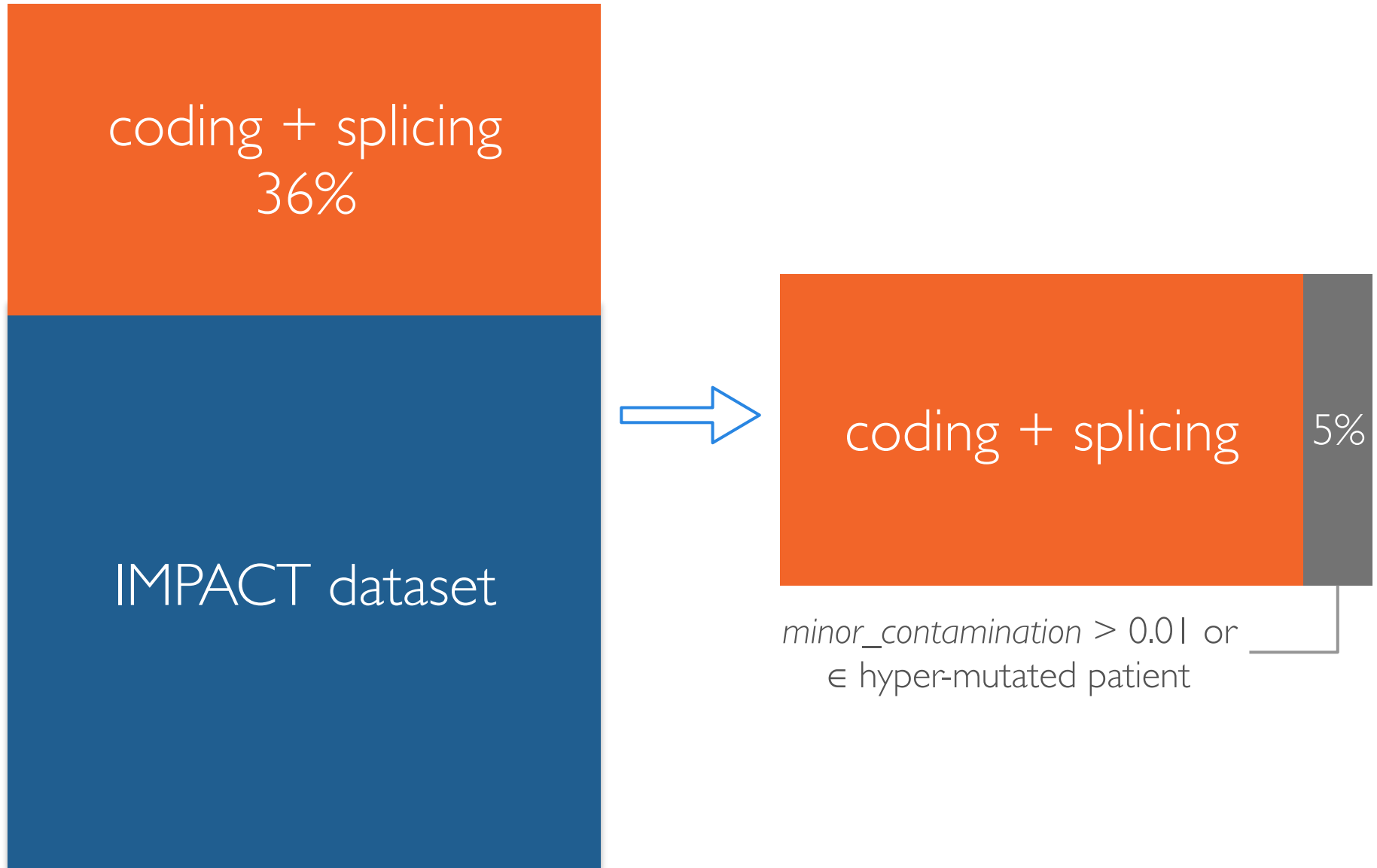


1884

coding + splicing
36%

IMPACT dataset

The IMPACT sub-dataset used

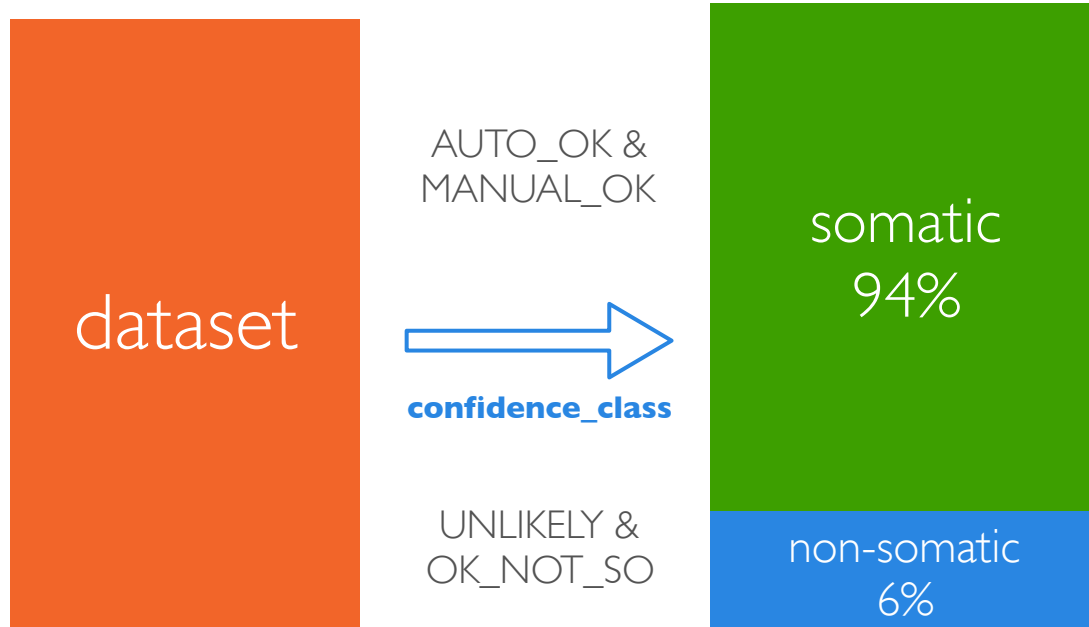


Different classes

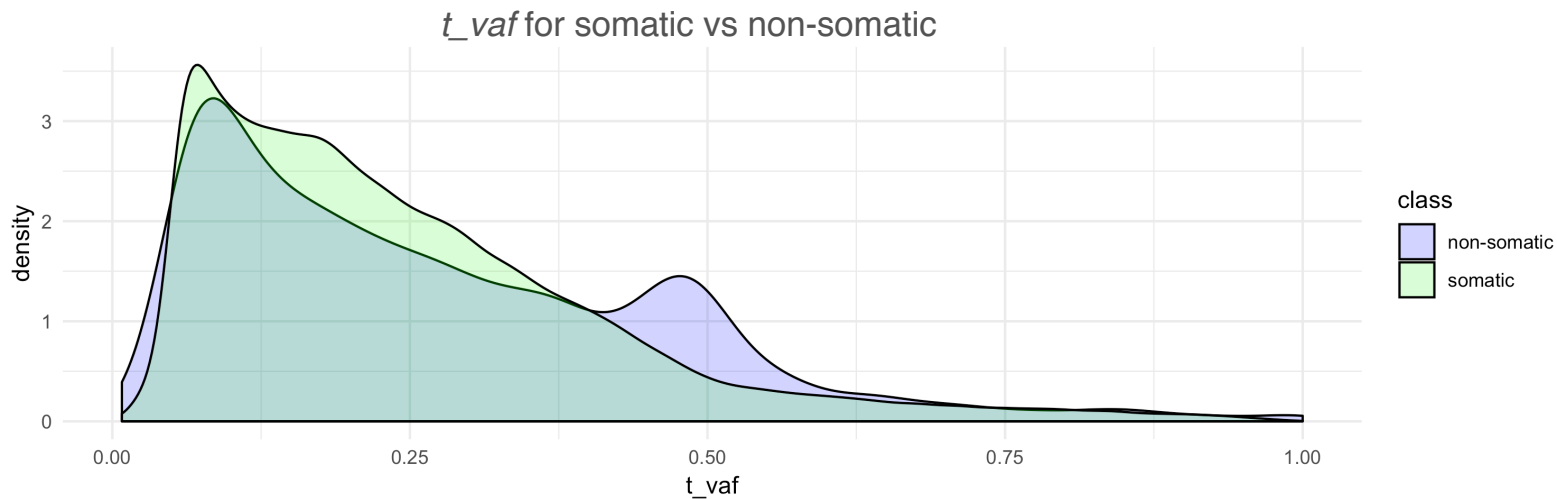
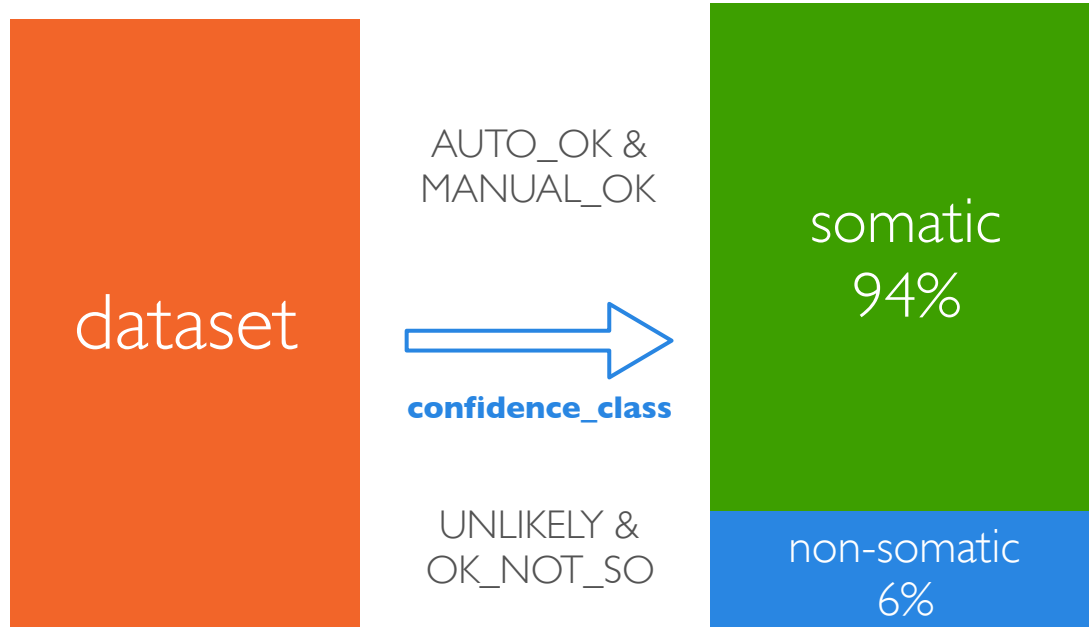


dataset

Different classes

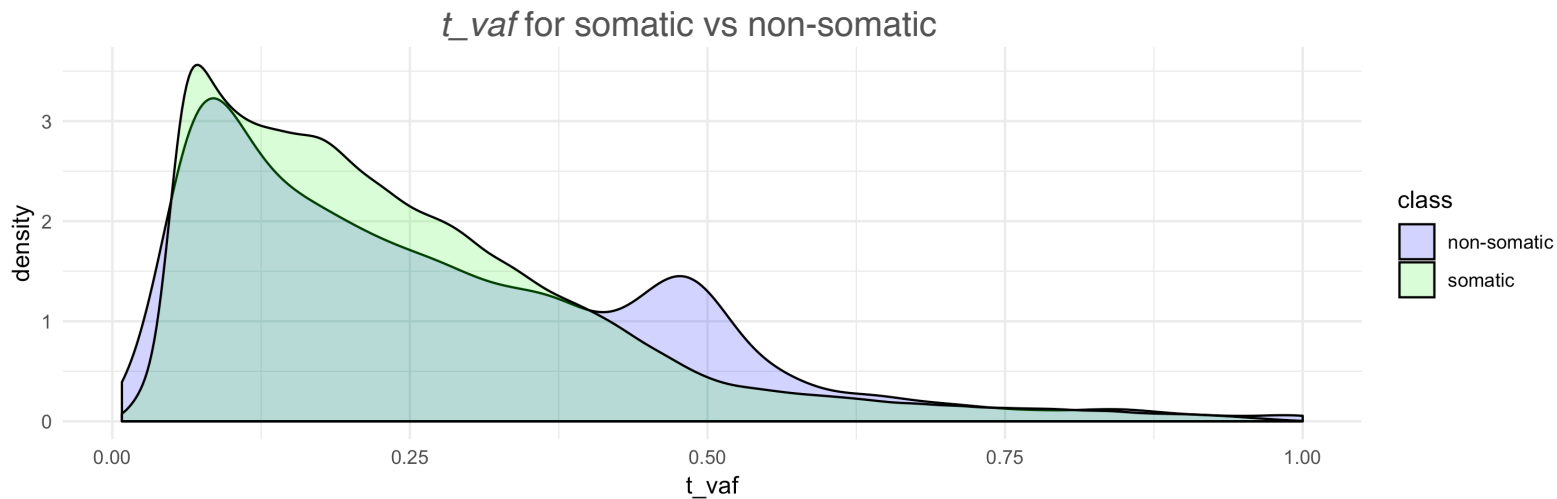
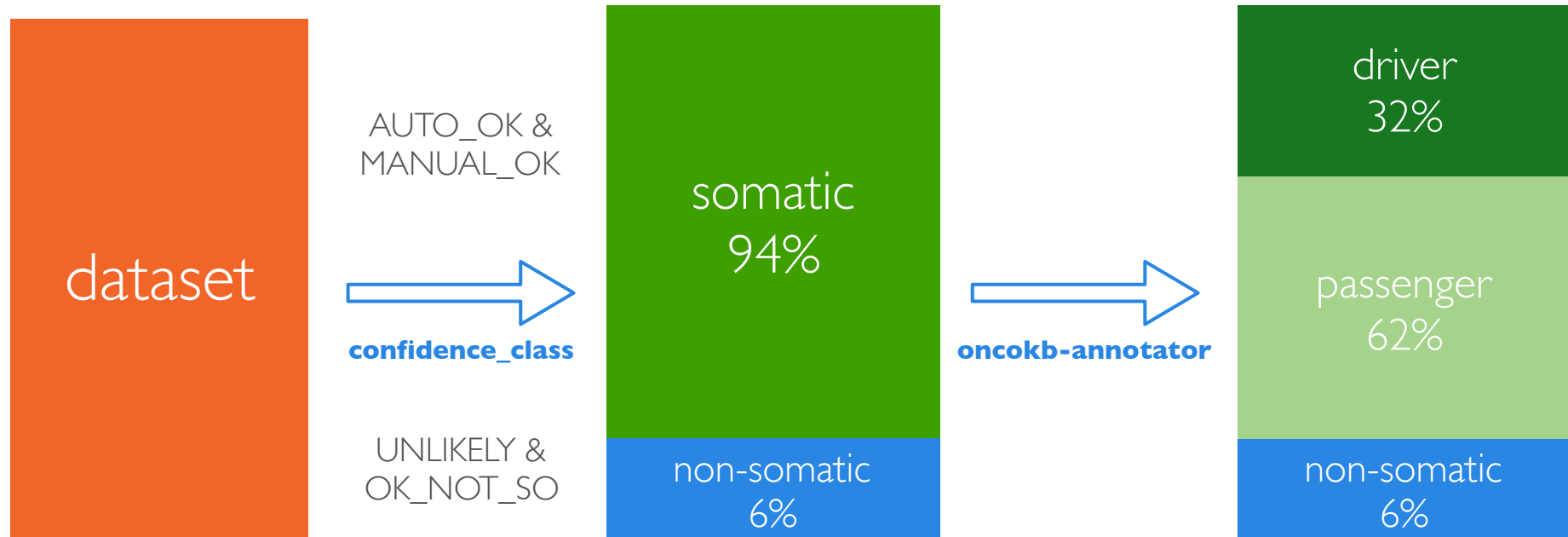


Different classes



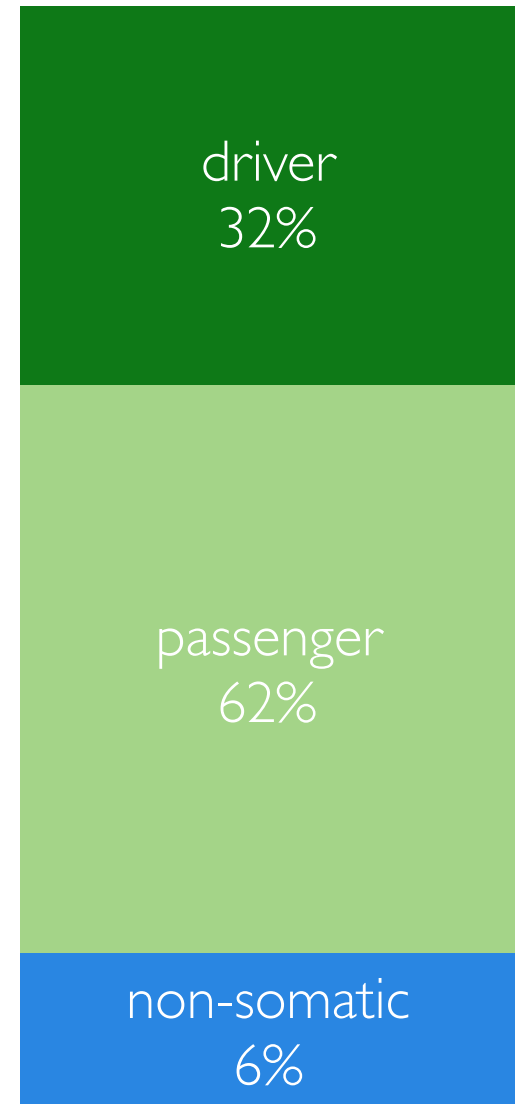
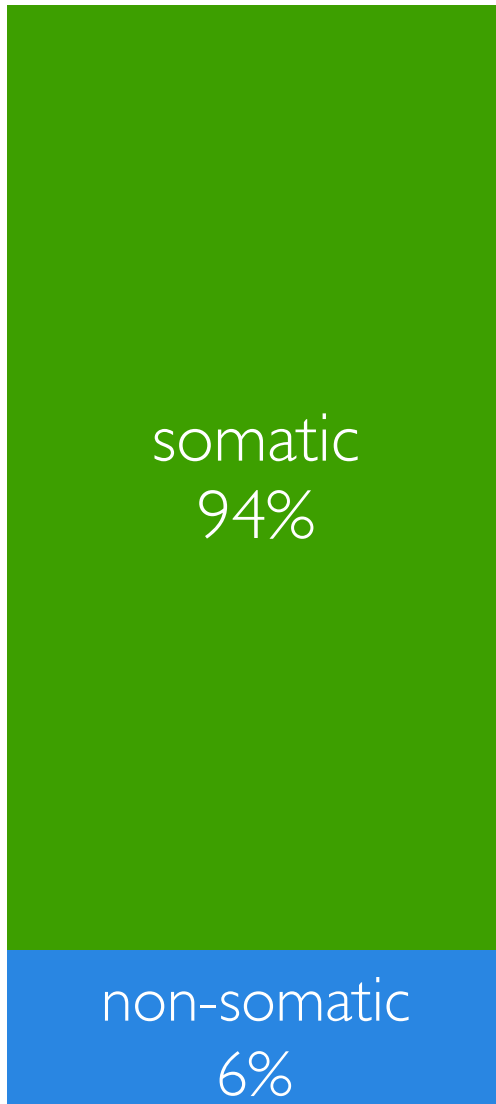
non-somatic =
artefact +
germline

Different classes

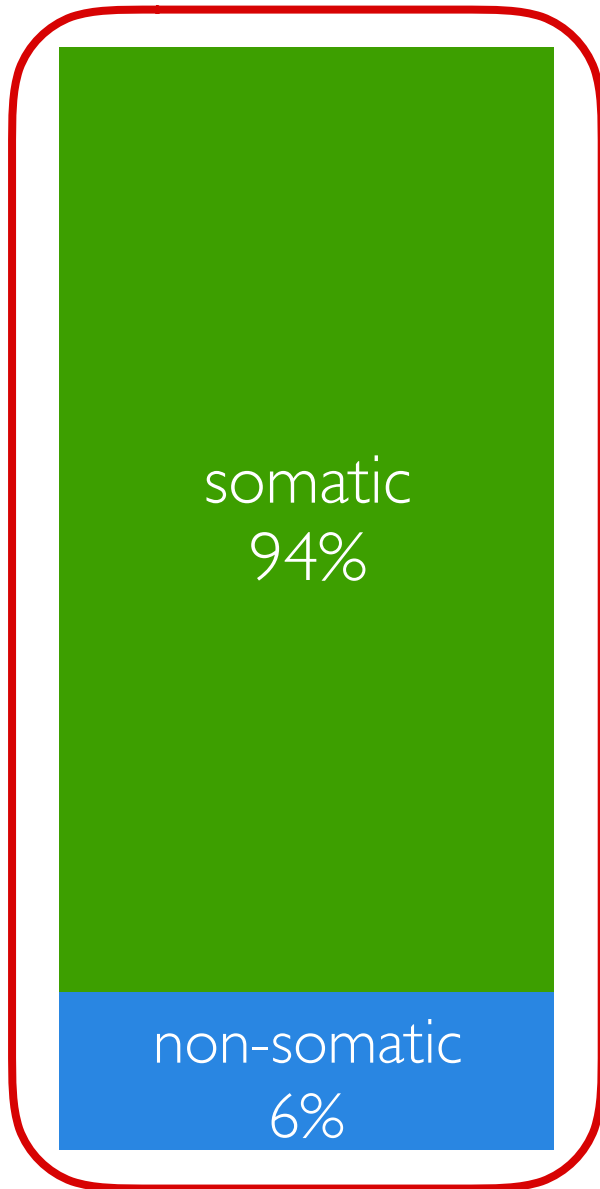


non-somatic =
artefact +
germline

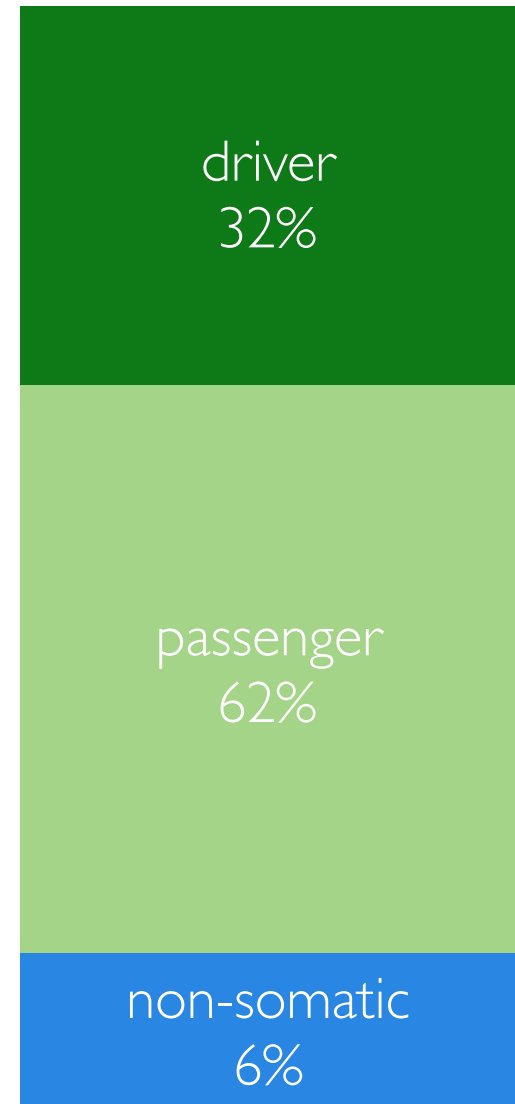
Supervised learning



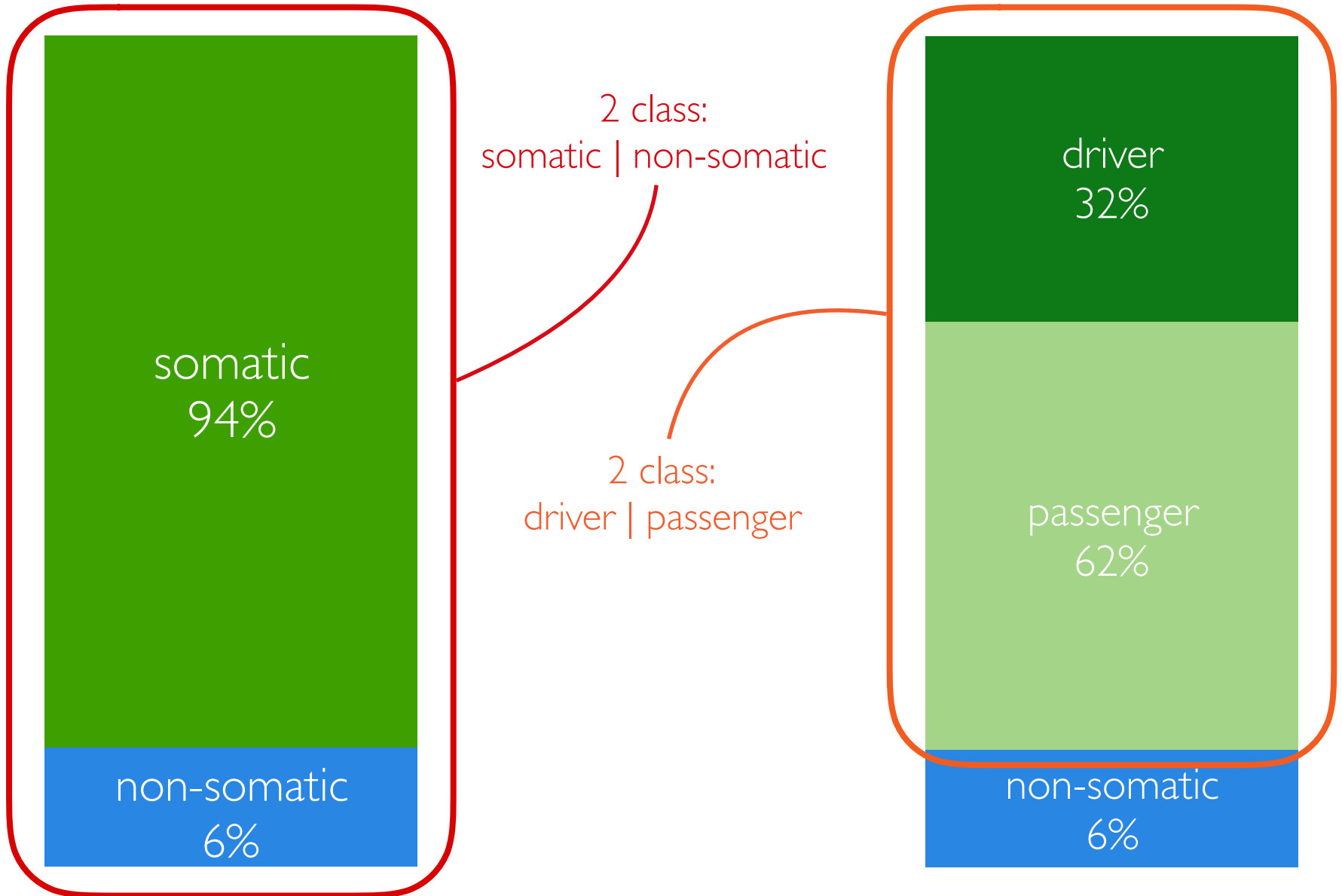
Supervised learning



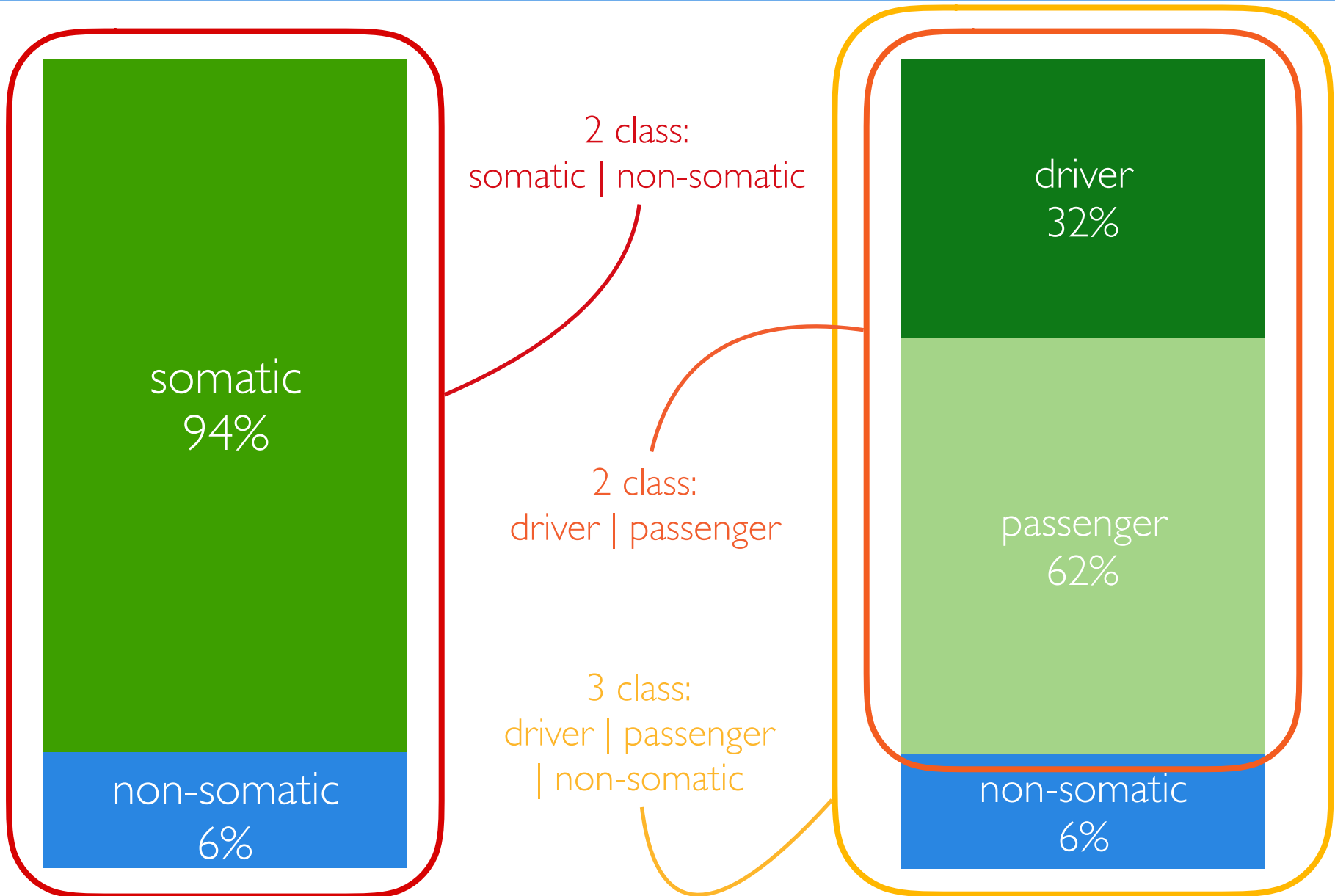
2 class:
somatic | non-somatic



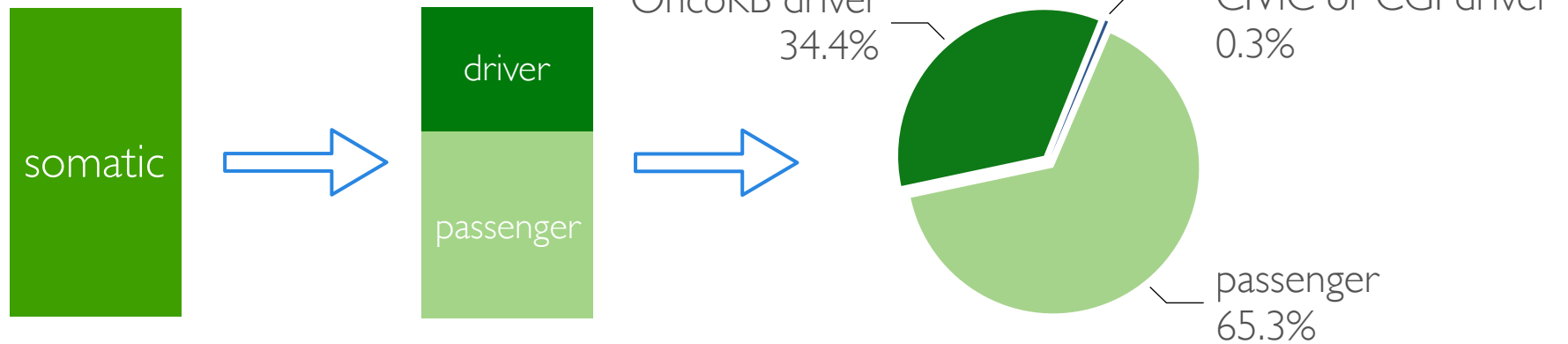
Supervised learning



Supervised learning



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
 - non-synonymous SNVs only
 - smaller and not “real” dataset
 - never done on somatic vs non-somatic

The features

NGS features

- vaf
- strand ratio
- ...

Frequency in normals

Population AF

- Kaviar
- GnomAD

COSMIC score

Impact scores

- SIFT
- PolyPhen-2
- ...

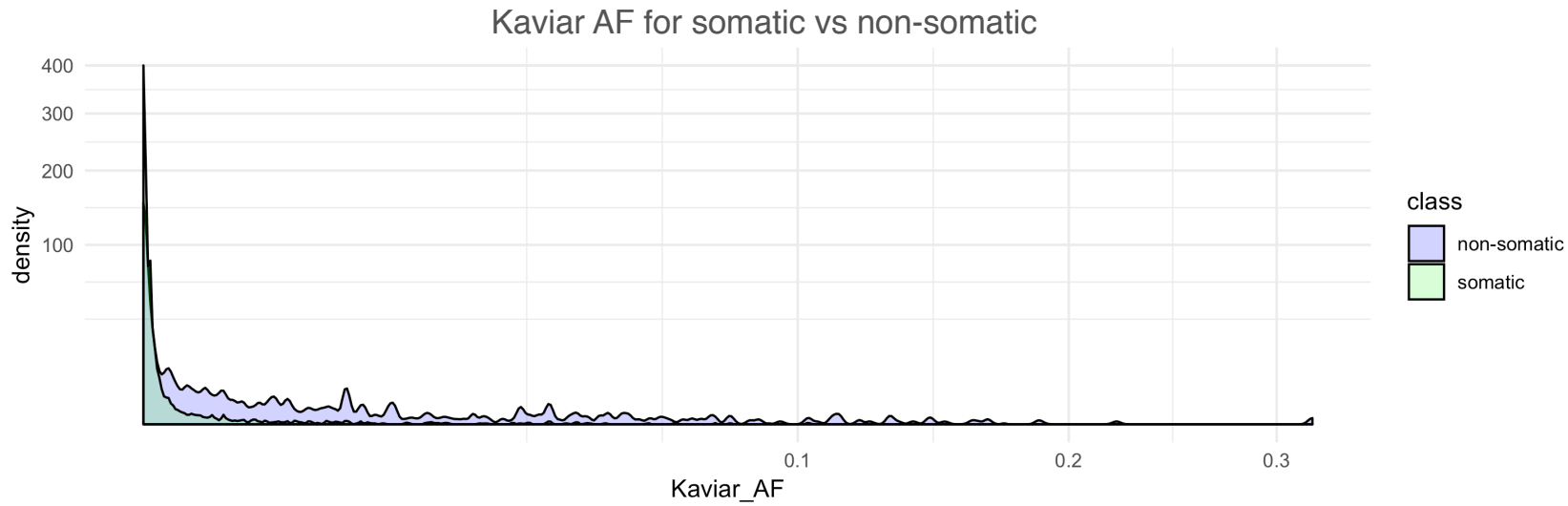
Genetic features

- gene and gene type
- mutation effect
- ...

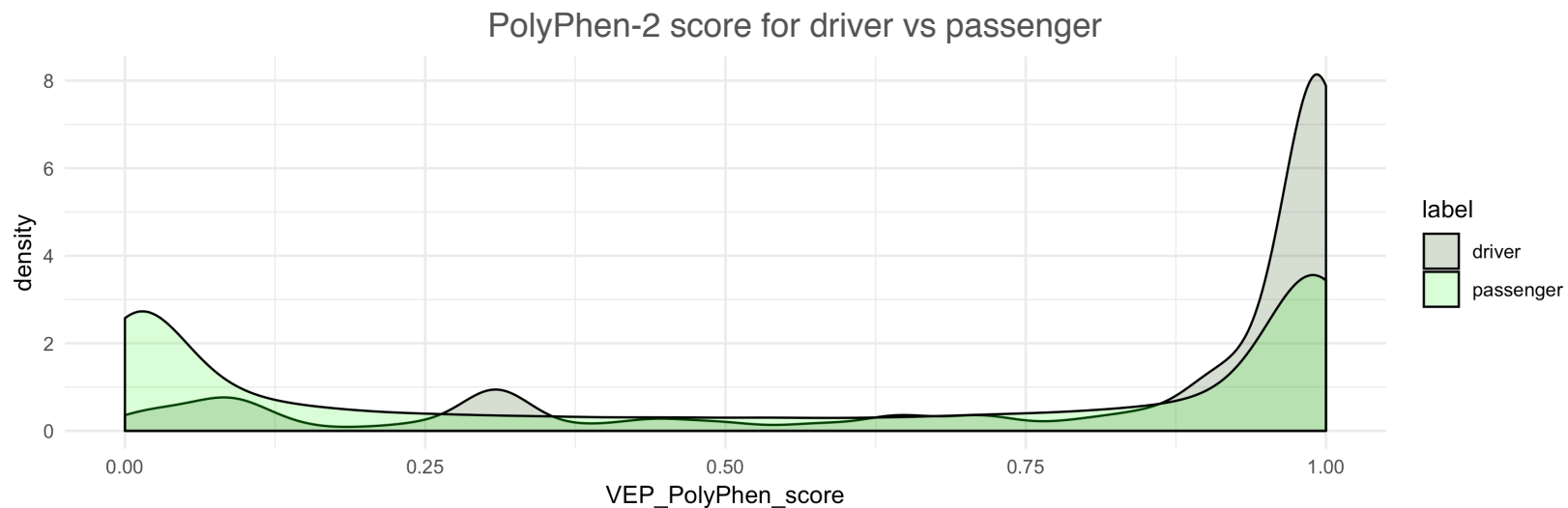
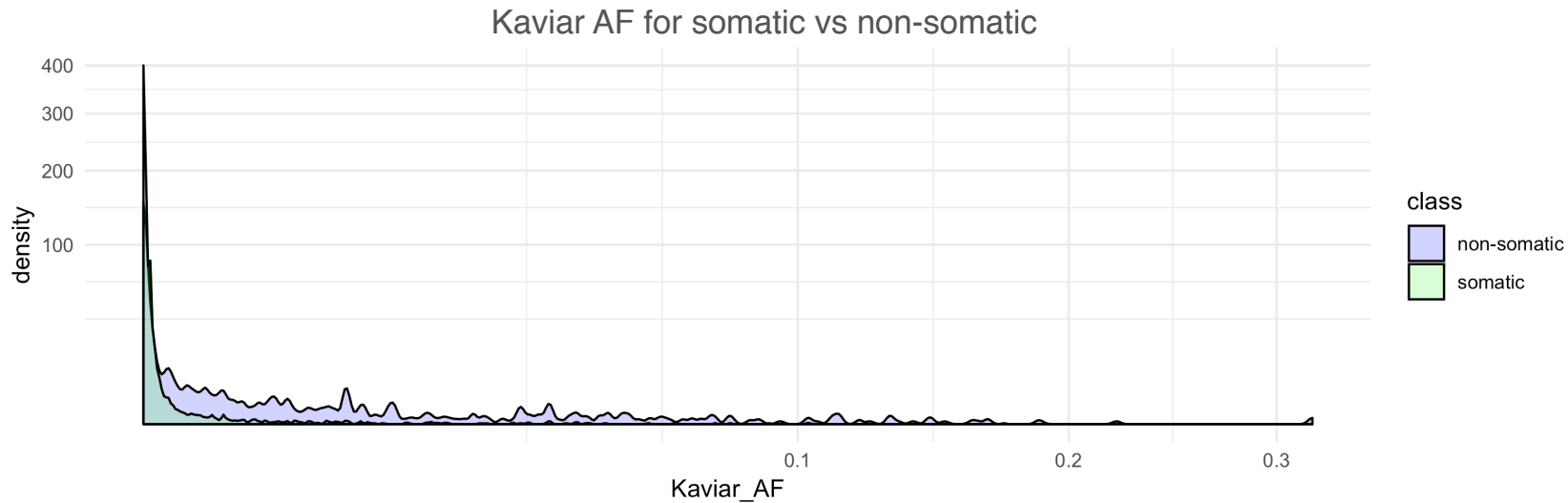
Cancer type?

Any ideas?

Example of informative feature



Example of informative feature



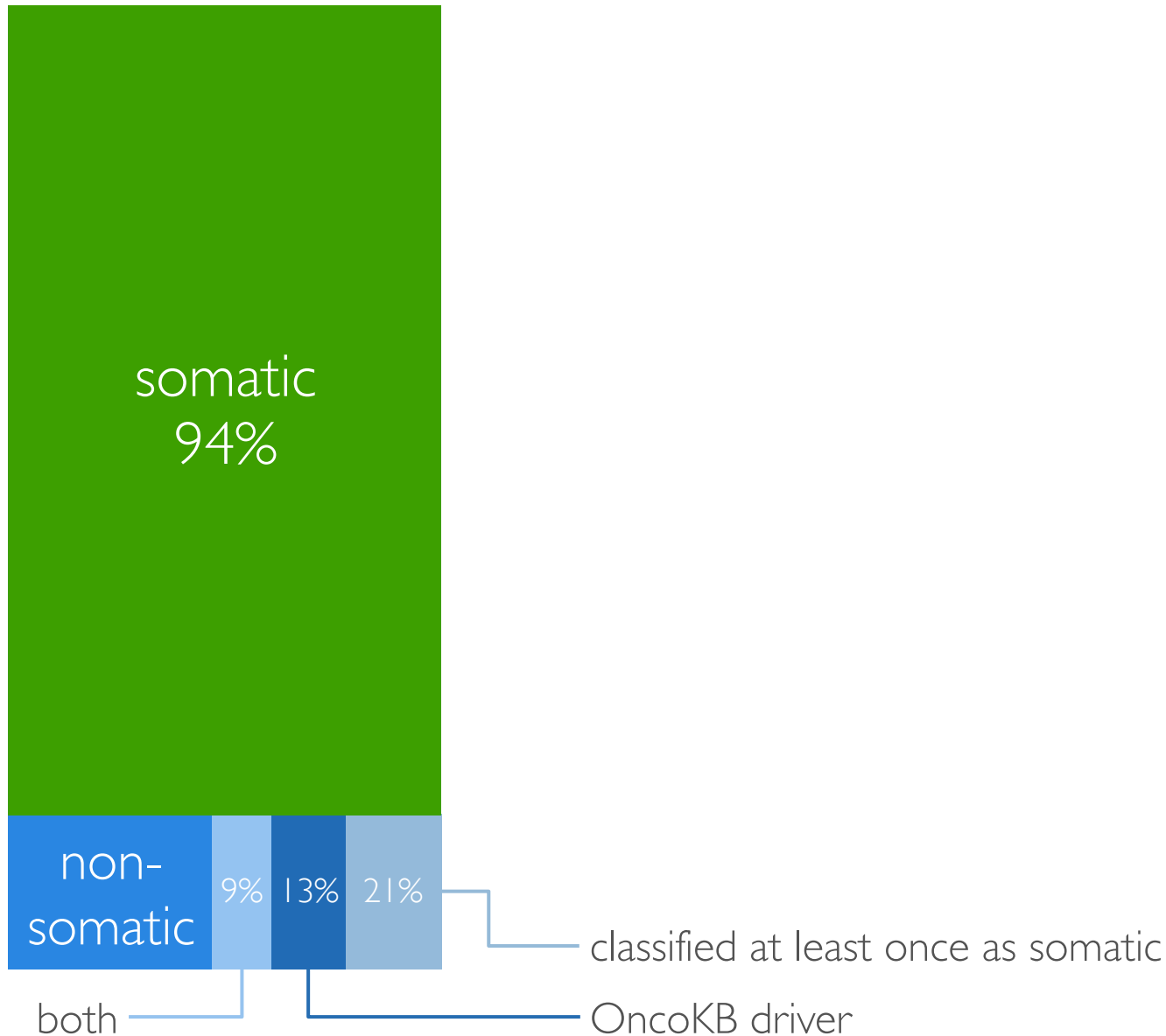
somatic vs non-somatic



somatic
94%

non-somatic
6%

somatic vs non-somatic



somatic vs non-somatic

somatic
94%

non-
somatic

9%

13%

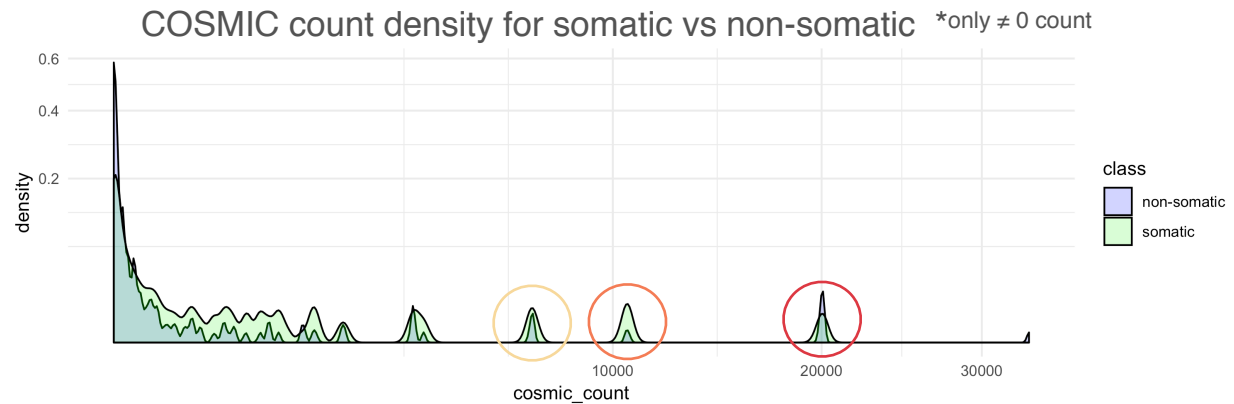
21%

both

classified at least once as somatic

OncoKB driver

Ex: hotspots classified as non-somatic



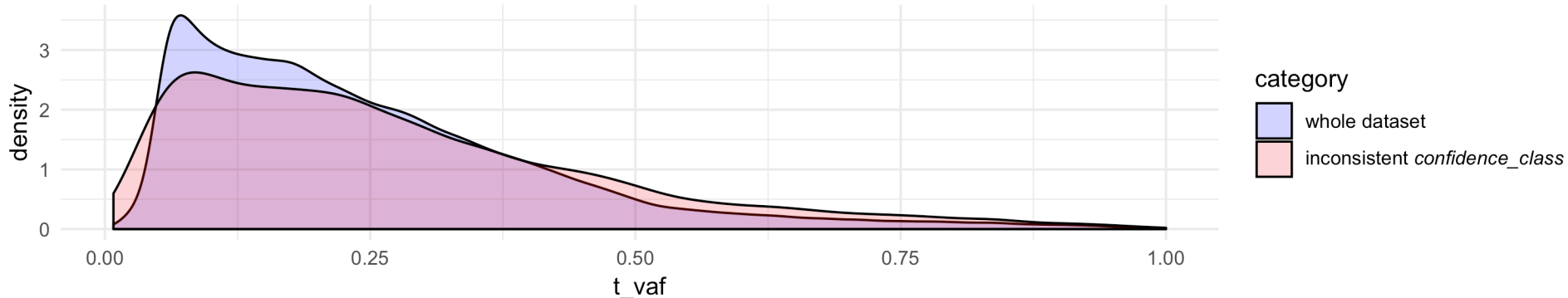
Description	Gene	HGVSp	COSMIC count	somatic count	non-somatic count	OncoKB driver
7:140453136 A/T	BRAF	p.V600E	20034	622	32	✓
12:25398284 C/T	KRAS	p.G12D	10579	1110	2	✓
12:25398284 C/A	KRAS	p.G12V	7055	896	10	✓



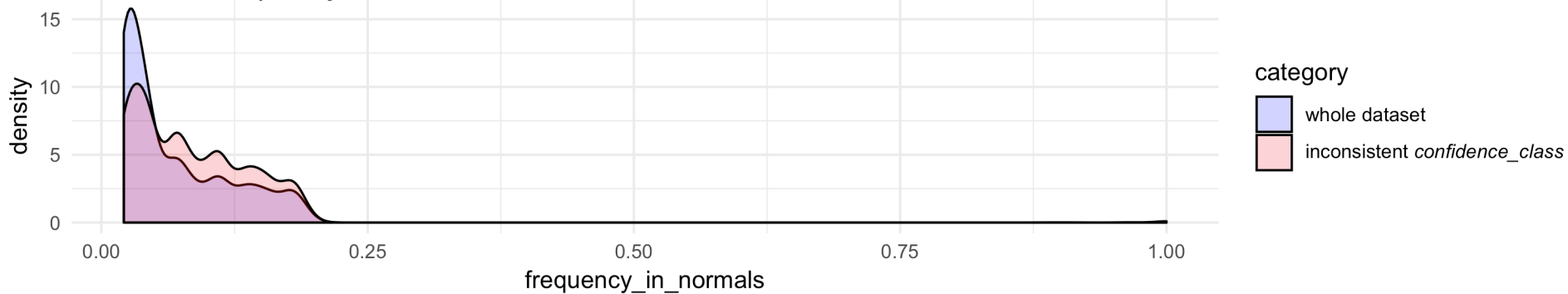
Inconsistent somatic vs non-somatic metrics



t_vaf for mutations with inconsistent *confidence_class* vs the whole dataset



frequency in normals for mutations with inconsistent *confidence_class* vs the whole dataset



strand ratio for mutations with inconsistent *confidence_class* vs the whole dataset

