Memorial Sloan Kettering Cancer Center

Variant classifier

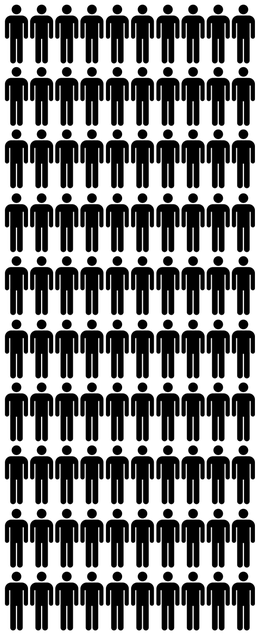
Developing a knowledge-based approach using IMPACT data

November 8, 2018

Papaemmanuil Lab
Pierre Guilmin | Elsa Bernard
In collaboration with A. Zehir, R. Ptashkin and C. Debyani

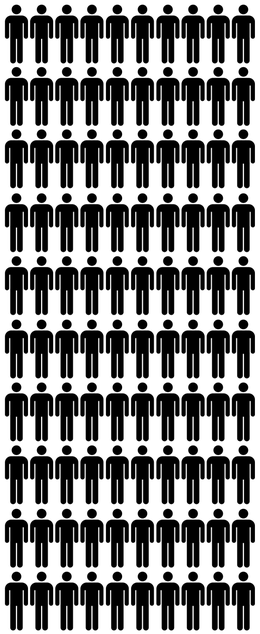
A variant classifier, why is it important?

100 patients



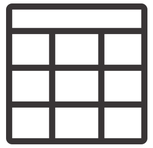
A variant classifier, why is it important?

100 patients

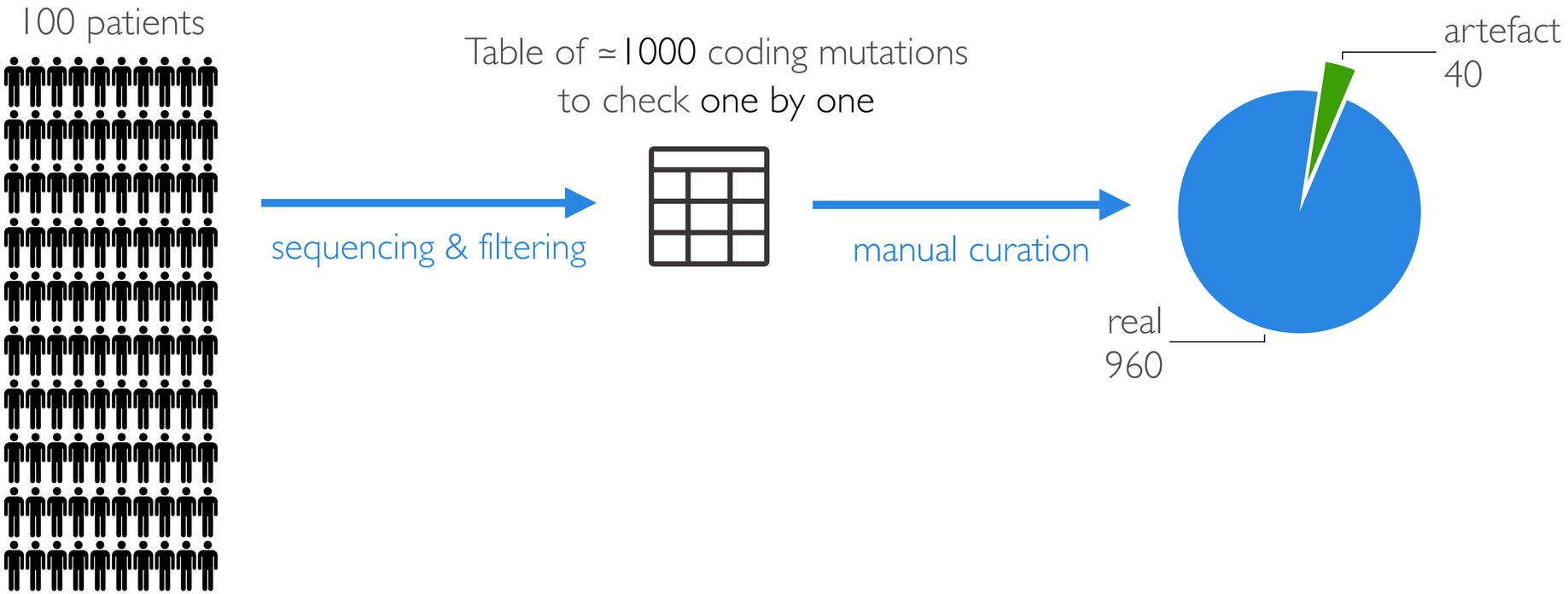


sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

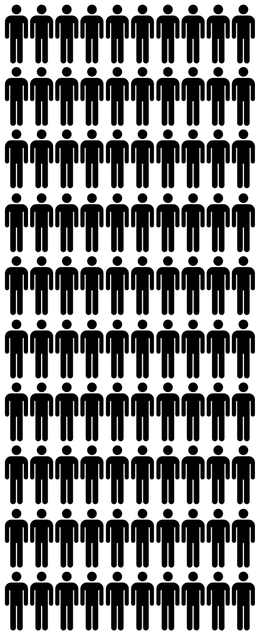


A variant classifier, why is it important?



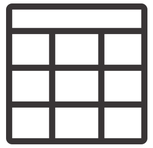
A variant classifier, why is it important?

100 patients



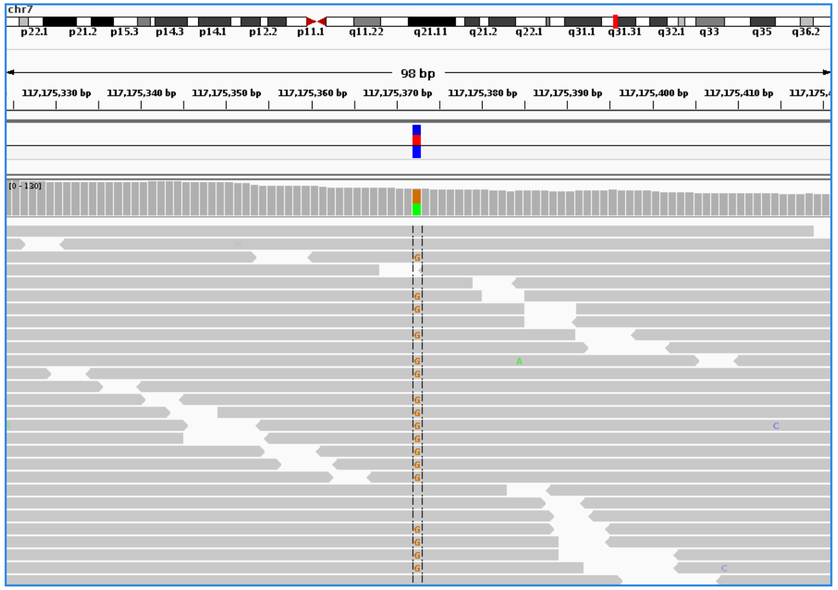
sequencing & filtering

Table of ≃1000 coding mutations
to check one by one



manual curation

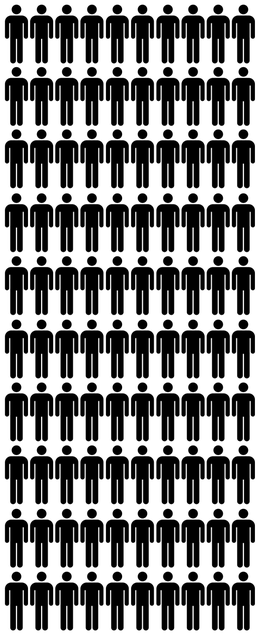
Expert work



IGV snapshot

A variant classifier, why is it important?

100 patients



sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

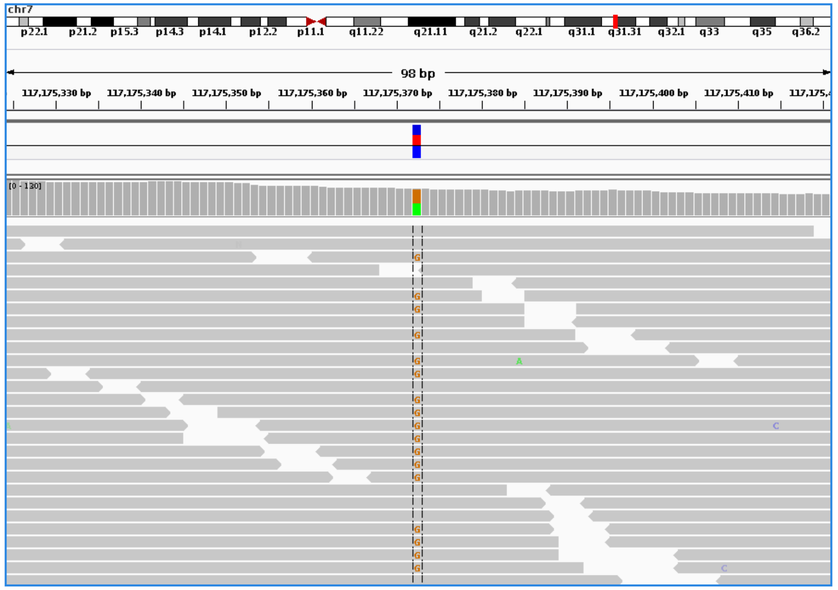
manual curation

OncoKB

database annotation

Clinical report

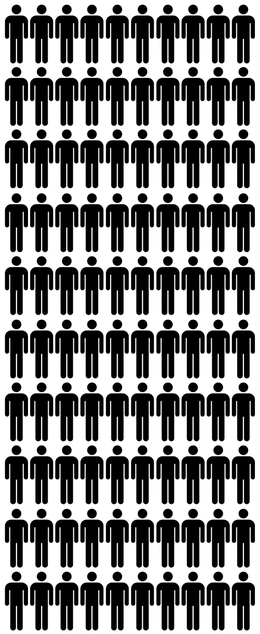
Expert work



IGV snapshot

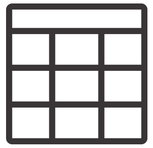
A variant classifier, why is it important?

100 patients

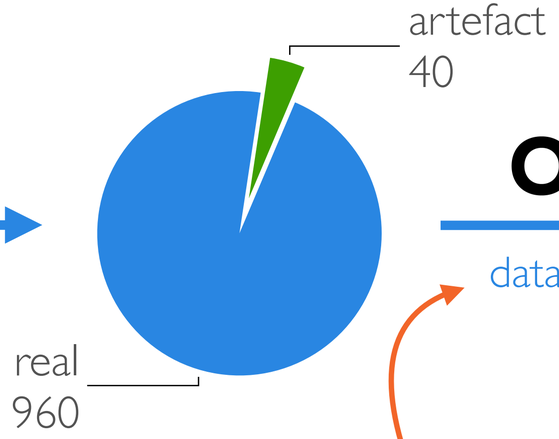


sequencing & filtering

Table of ≃1000 coding mutations
to check one by one



manual curation

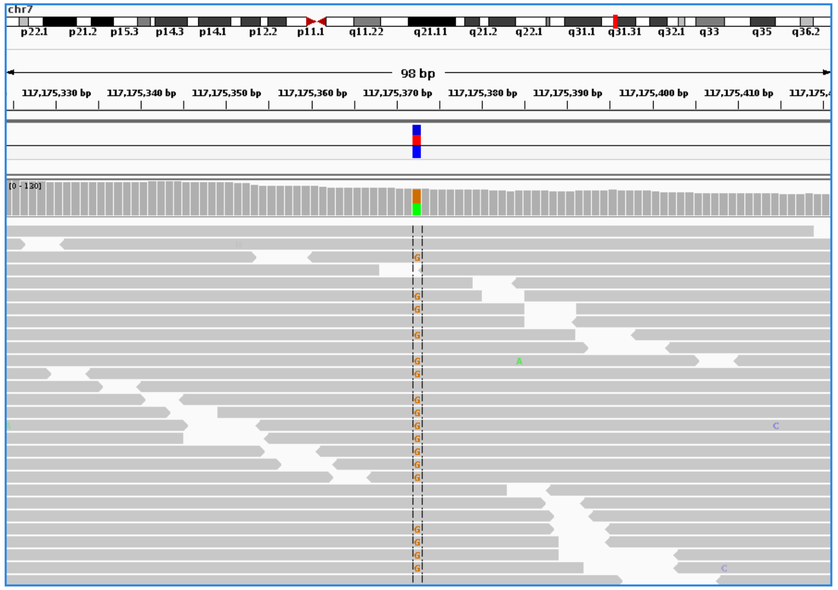


OncoKB

database annotation

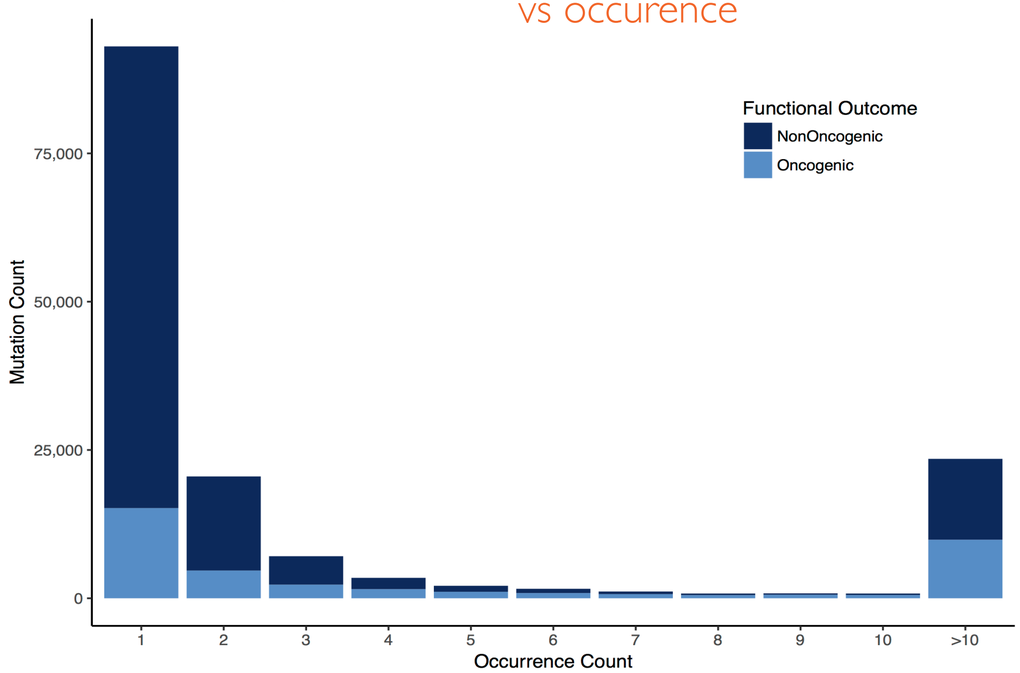
Clinical report

Expert work



IGV snapshot

MSK-IMPACT: oncogenicity
vs occurence



The goal

Create a tool that classifies variant automatically

- real vs artefact OR driver vs passenger
- using Supervised Machine Learning Classification
- all cancer, all mutation type
- on the IMPACT dataset

IMPACT, the dataset

matched normal
variant calling

588,547 mutations
23,162 patients

the IMPACT dataset

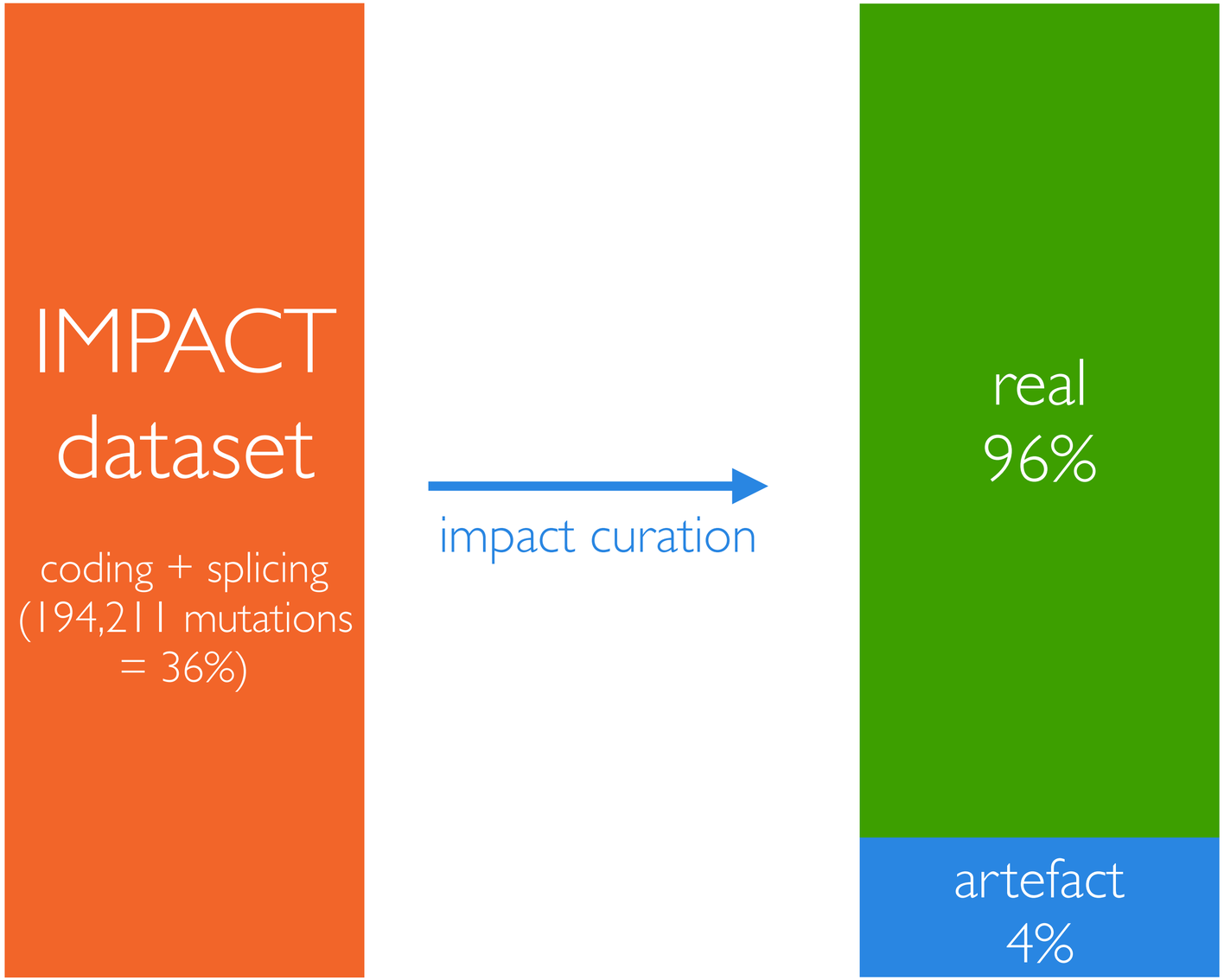
manually curated for
real vs artefact

Two steps classification

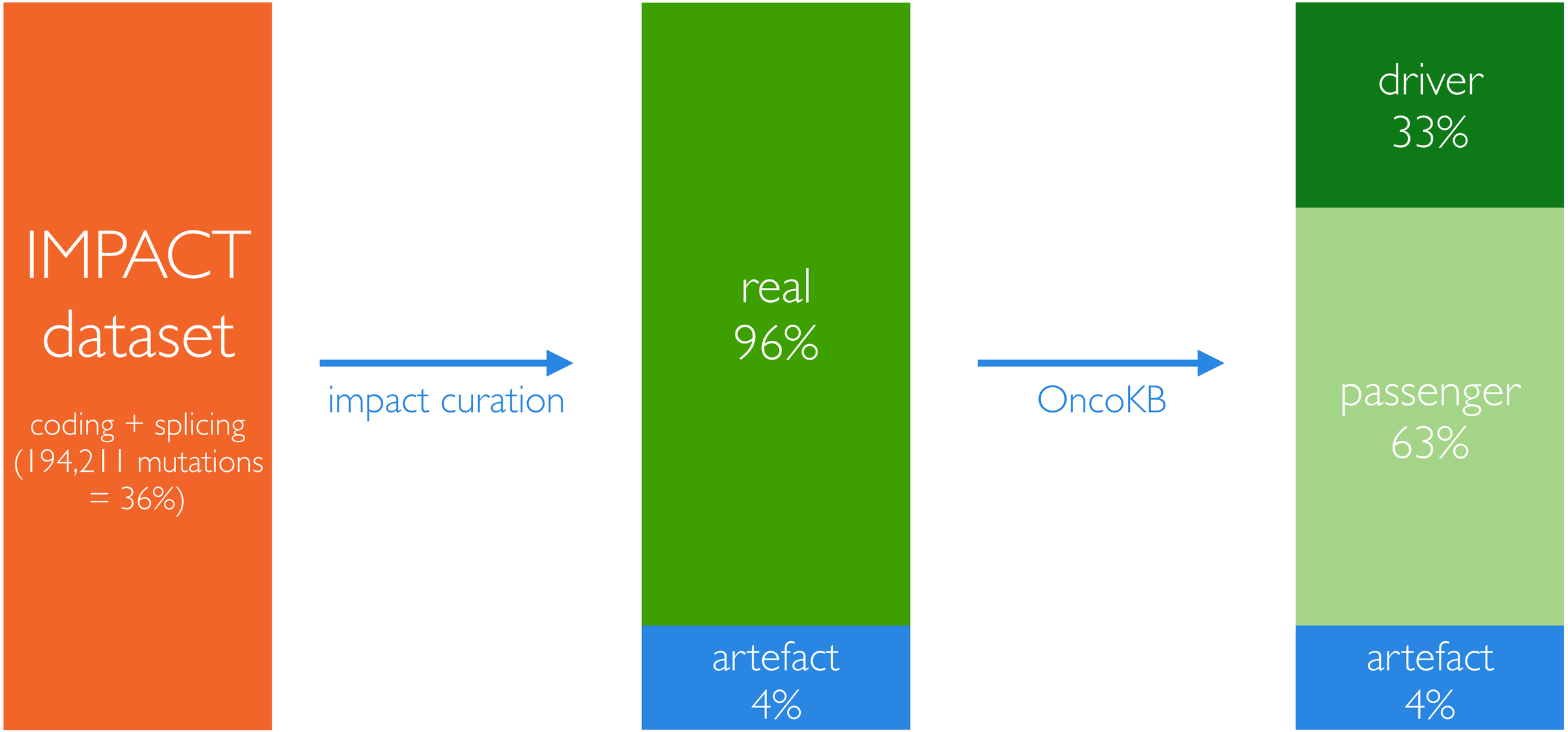
IMPACT dataset

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

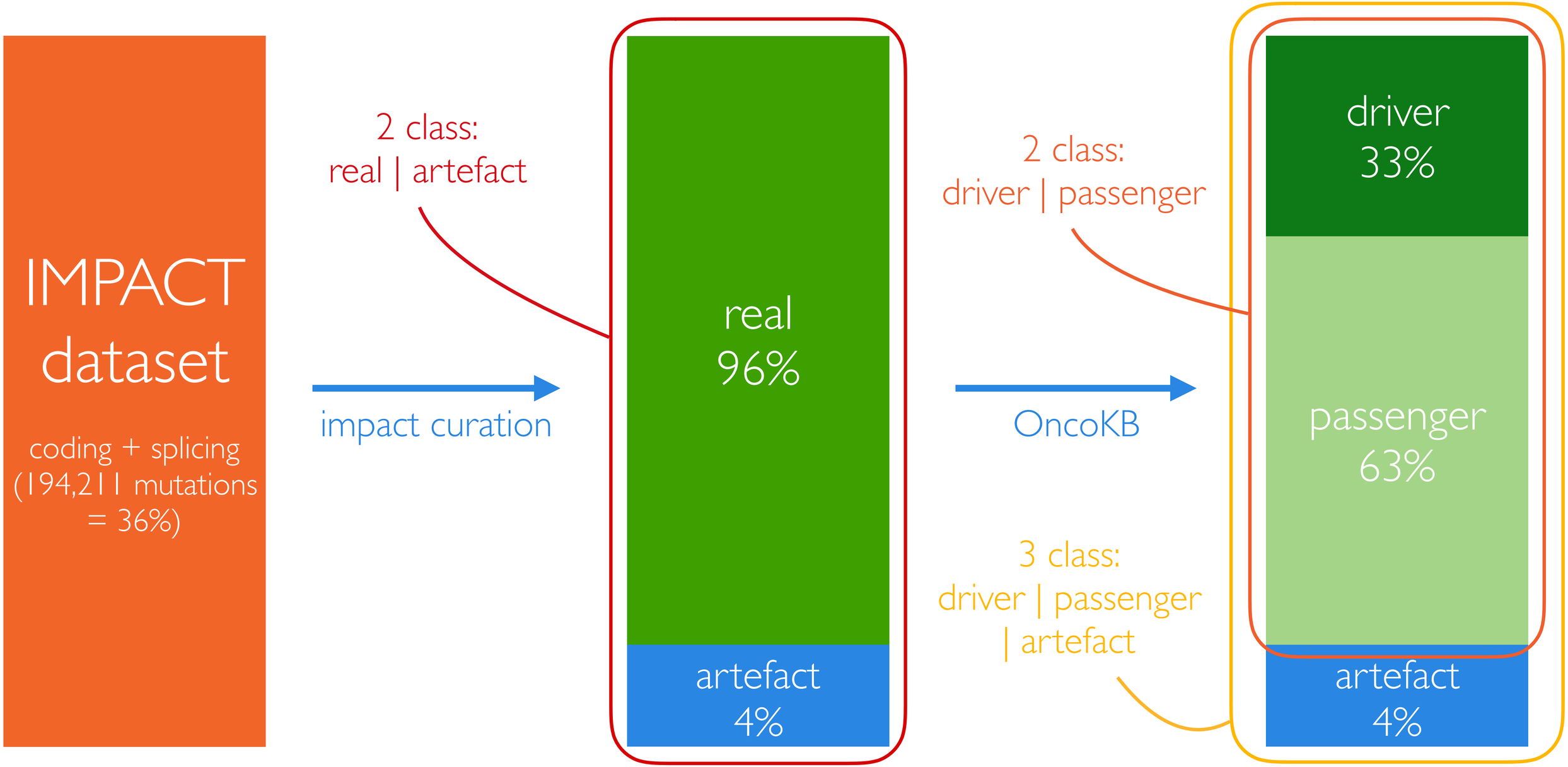
Two steps classification



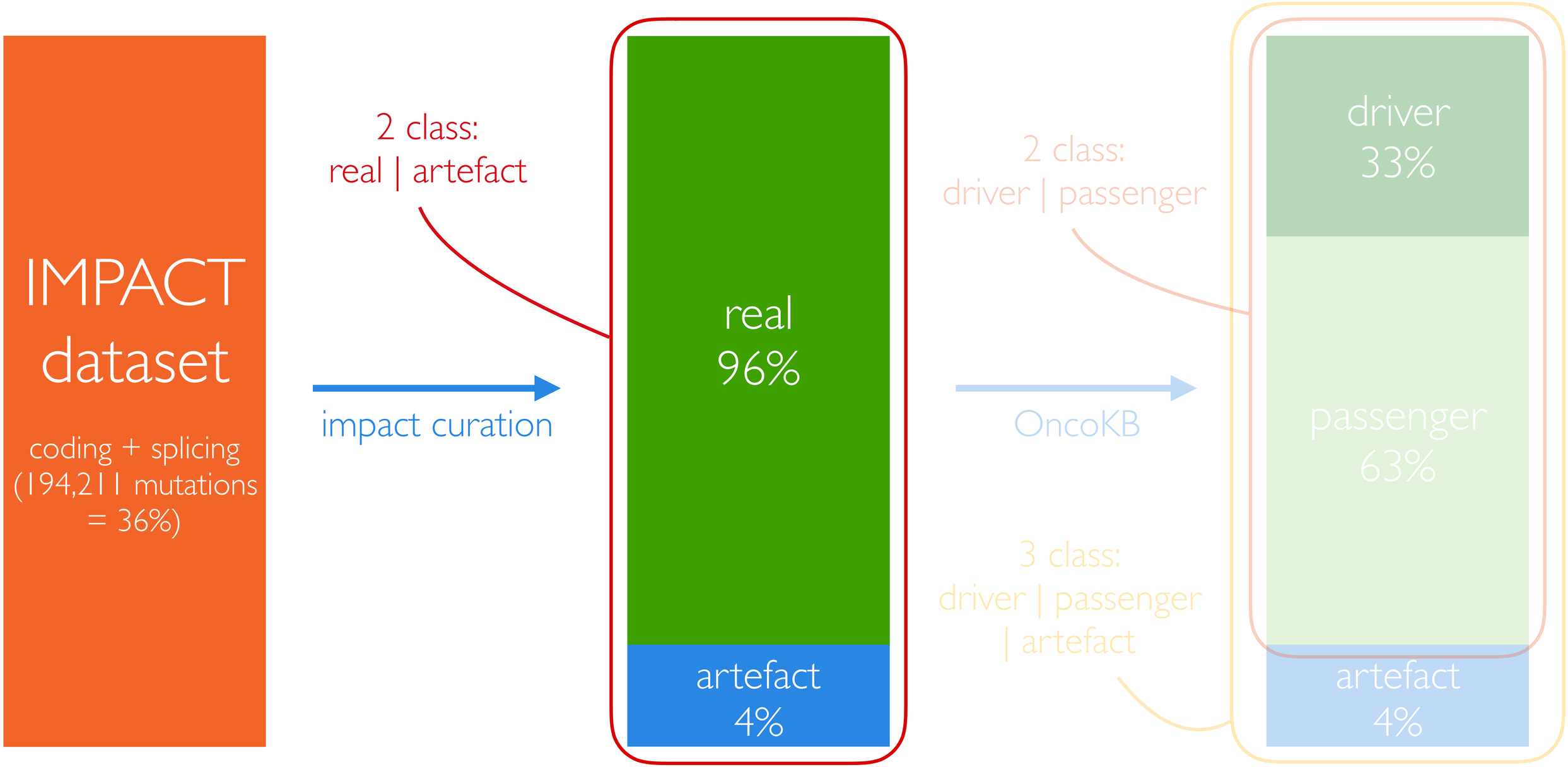
Two steps classification



Two steps classification



Two steps classification



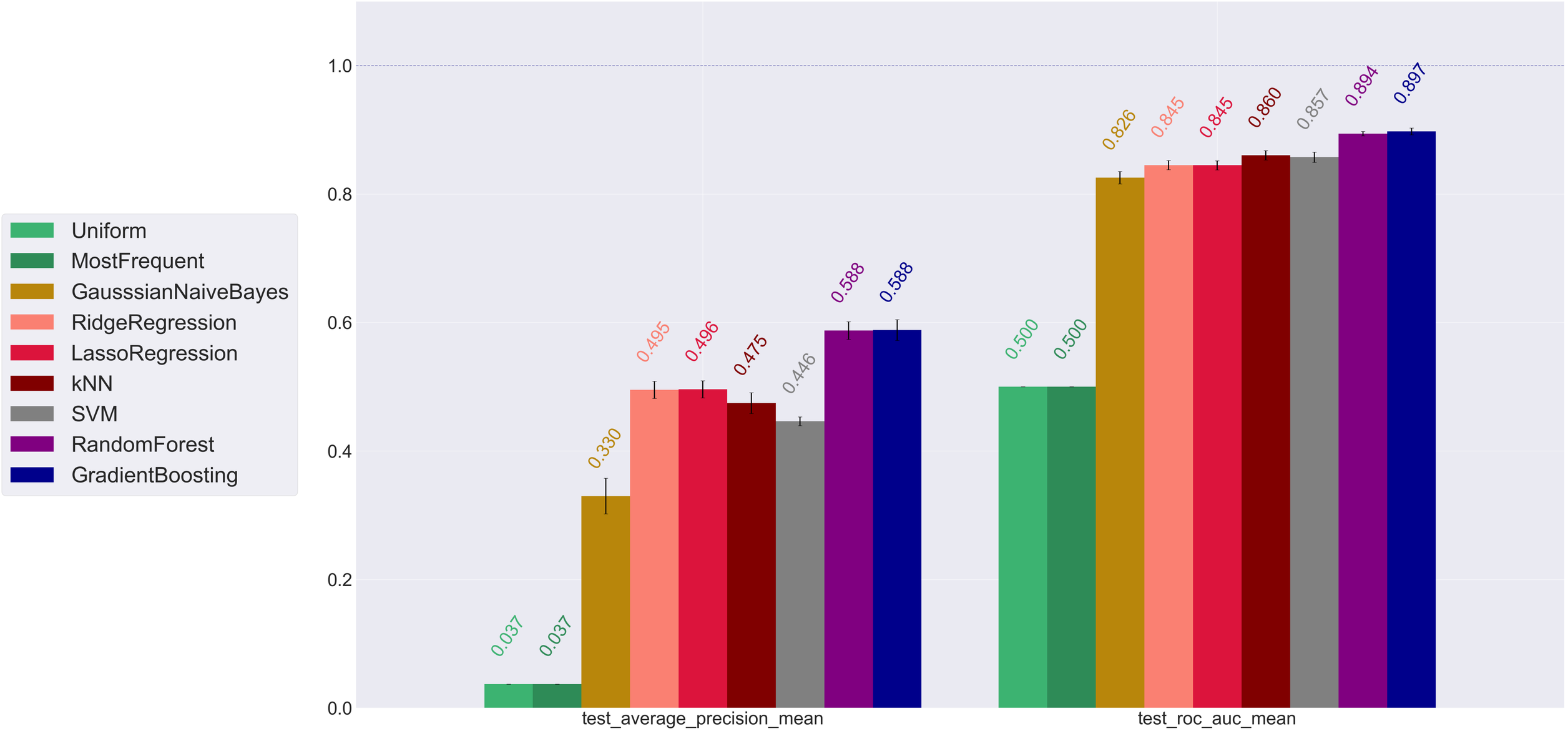
The features used in our model

- Sequencing features (n = 11)
  Tumor VAF, tumor depth
- Genomic coordinates (n = 3)
  Chromosome, Hugo Symbol
- Control population (n = 12)
  GnomAD allele frequency
- Cancer population (n = 4)
  COSMIC count, OncoKB
- Normal control (n = 1)
  Frequency in normals
- Mutation consequence (n = 6)
  Protein effect, SIFT & PolyPhen class

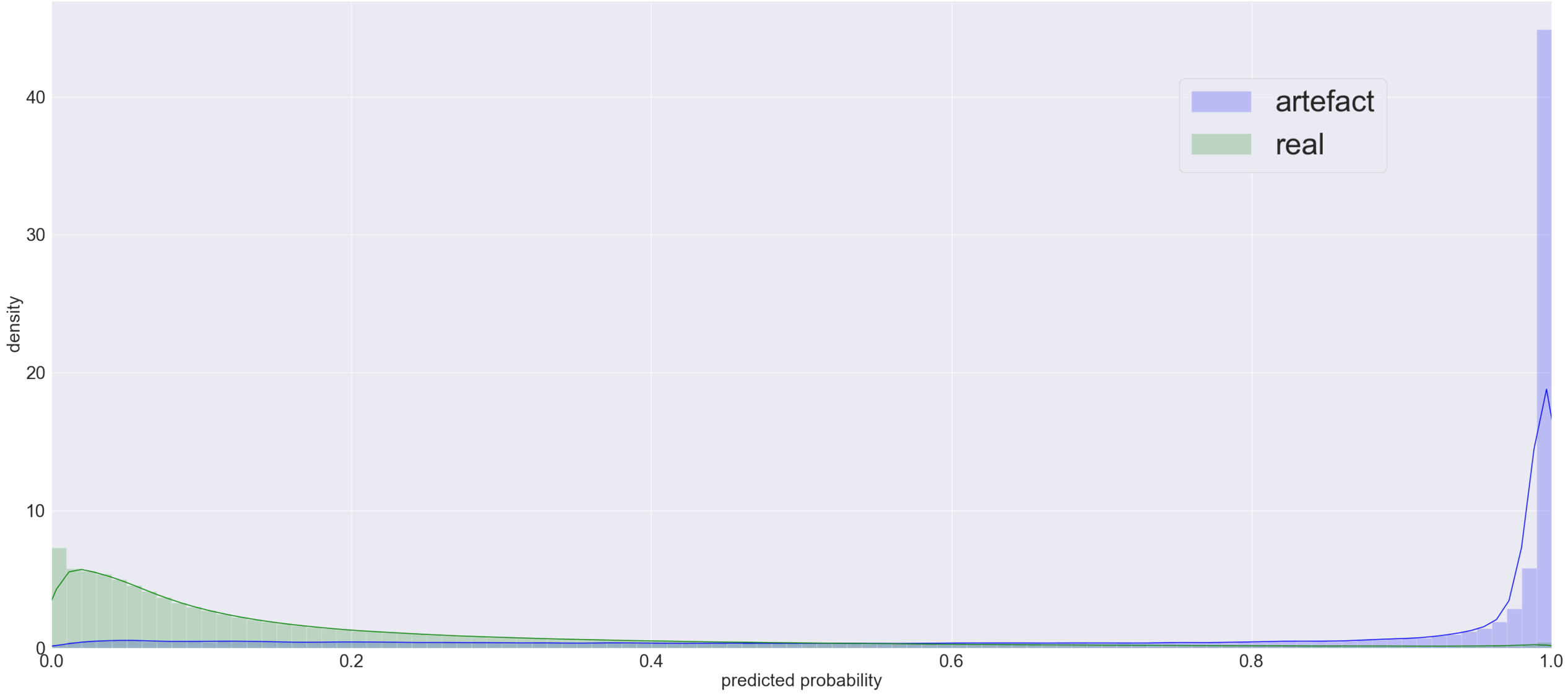
Algorithm comparison

- Uniform
- MostFrequent
- GausssianNaiveBayes
- RidgeRegression
- LassoRegression
- kNN
- SVM
- RandomForest
- GradientBoosting

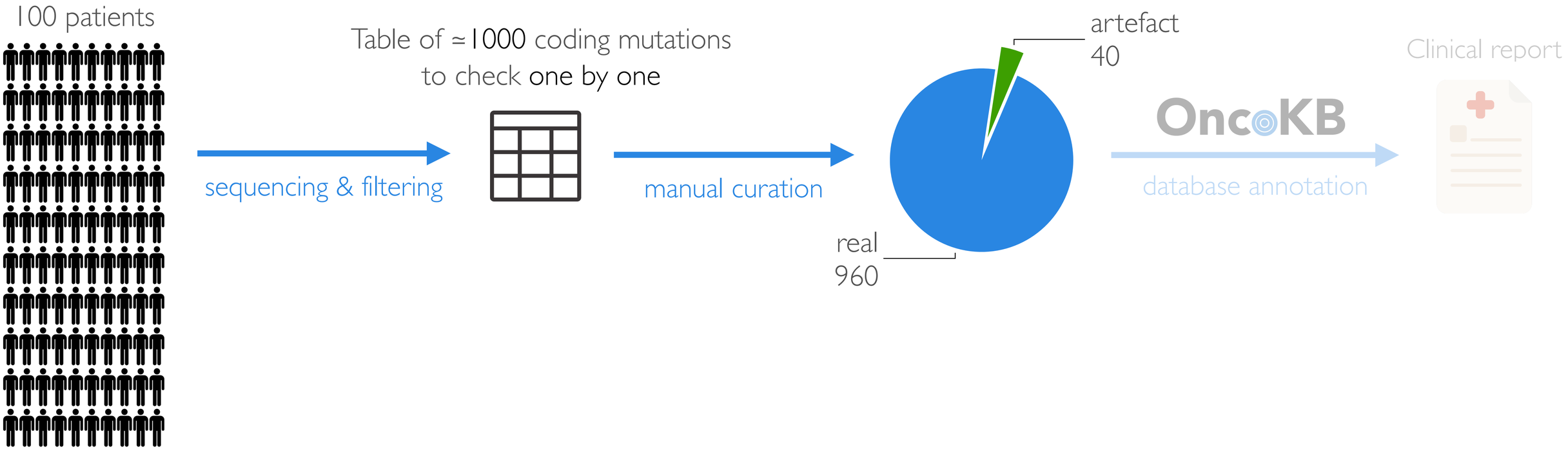
Algorithm comparison



Best algorithm probability output

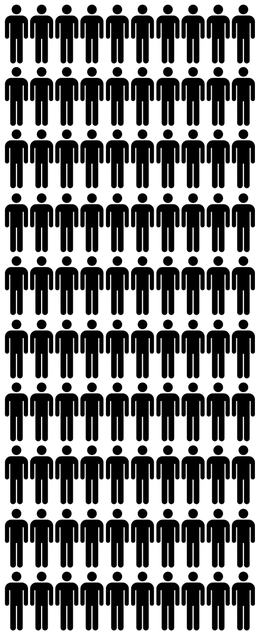


The variant classifier performances



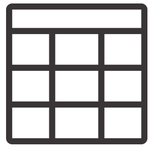
The variant classifier performances

100 patients



sequencing & filtering

Table of ≃1000 coding mutations
to check one by one



manual curation

OncoKB

database annotation

Clinical report

variant classifier

960 real

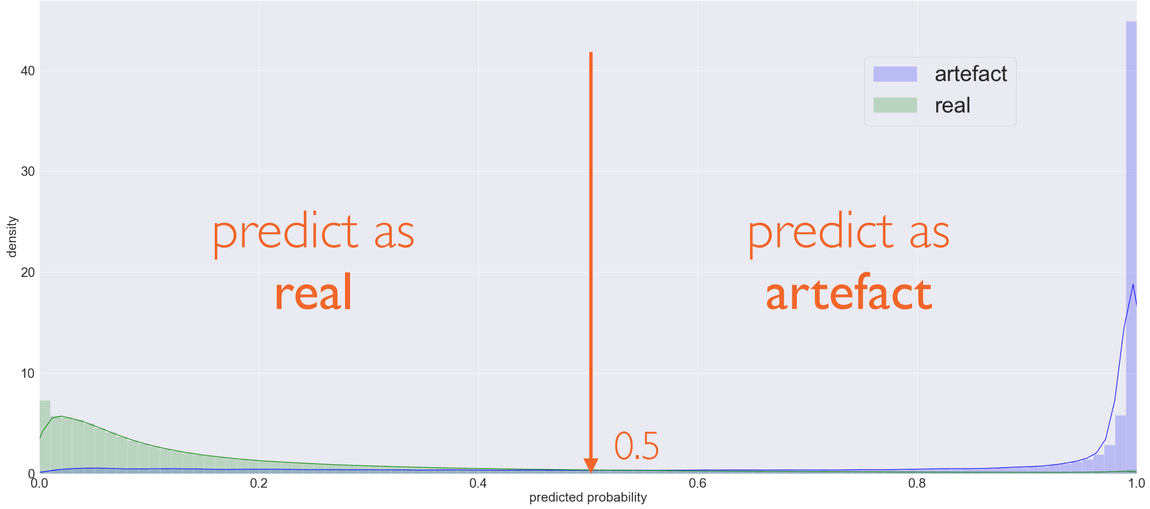
✅ 848 → real

❌ 112 → artefact

40 artefact

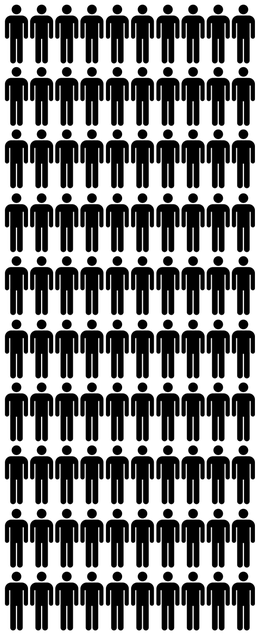
✅ 32 → artefact

❌ 8 → real



The variant classifier performances

100 patients



sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation

OncoKB

database annotation

Clinical report

variant classifier

960 real

✅ 848 → real

❌ 112 → artefact

40 artefact

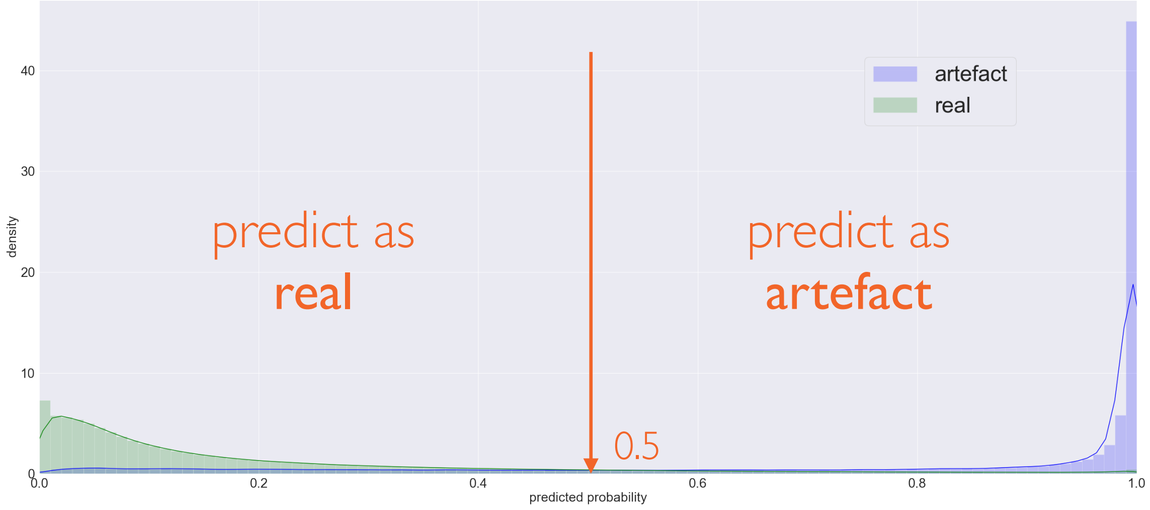
✅ 32 → artefact

❌ 8 → real

📝 8x less work

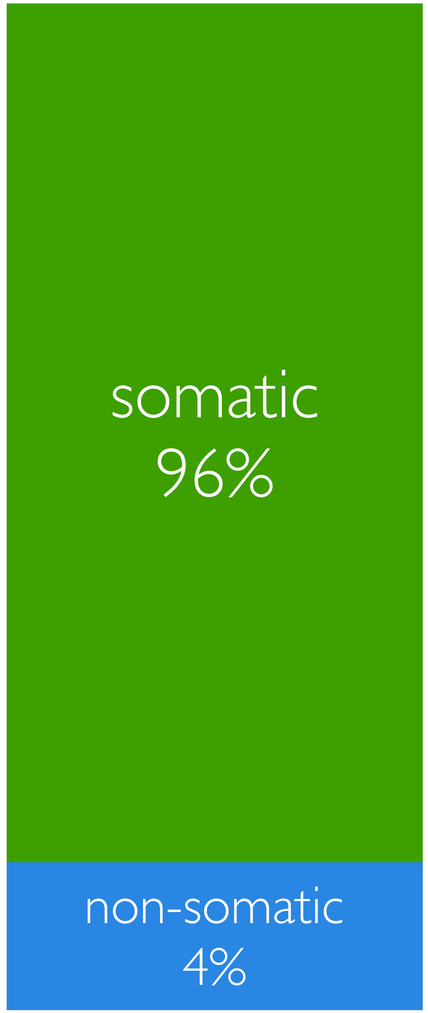
144/1000 mutations to check one
by one instead of 1000/1000

❌ 8/40 artefacts considered as real



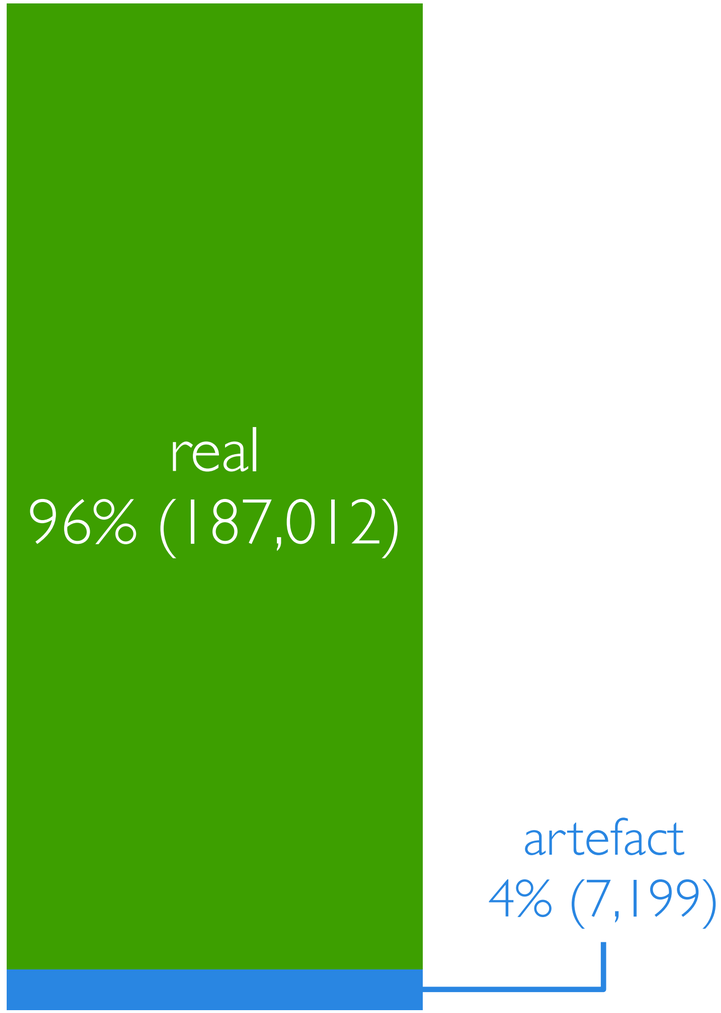
Main challenges

Imbalanced dataset



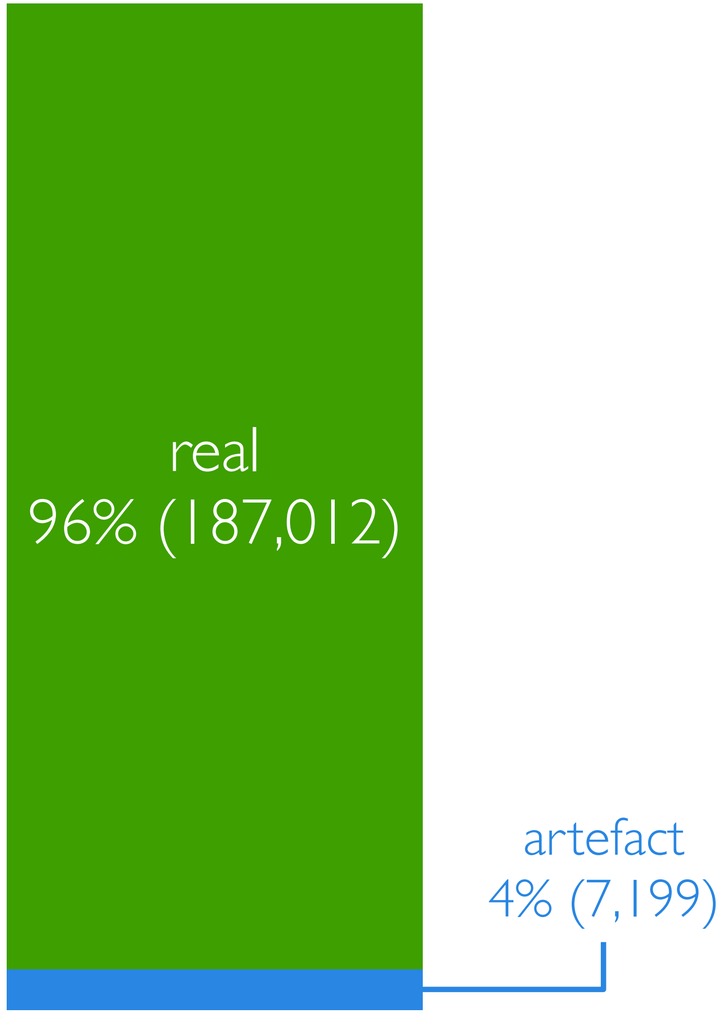
Main challenges

Imbalanced dataset

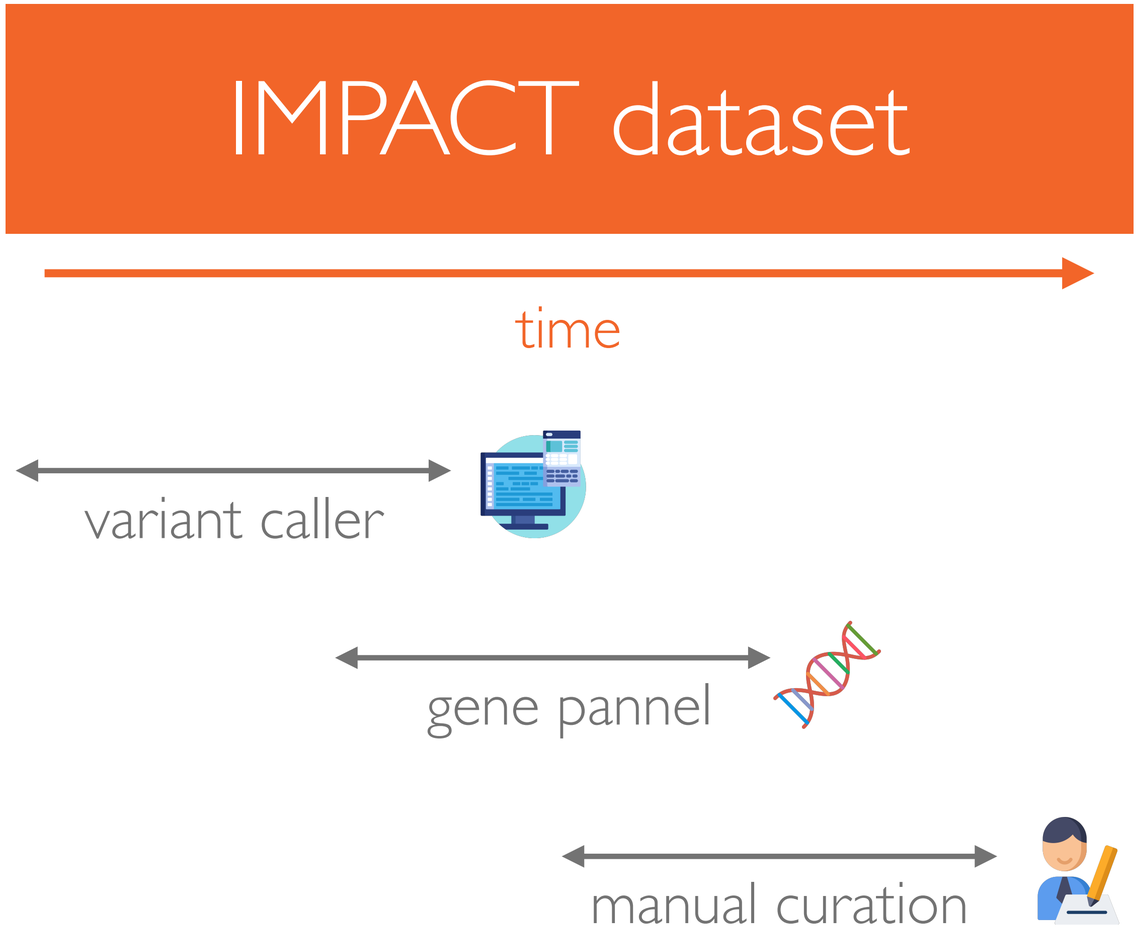


Main challenges

Imbalanced dataset

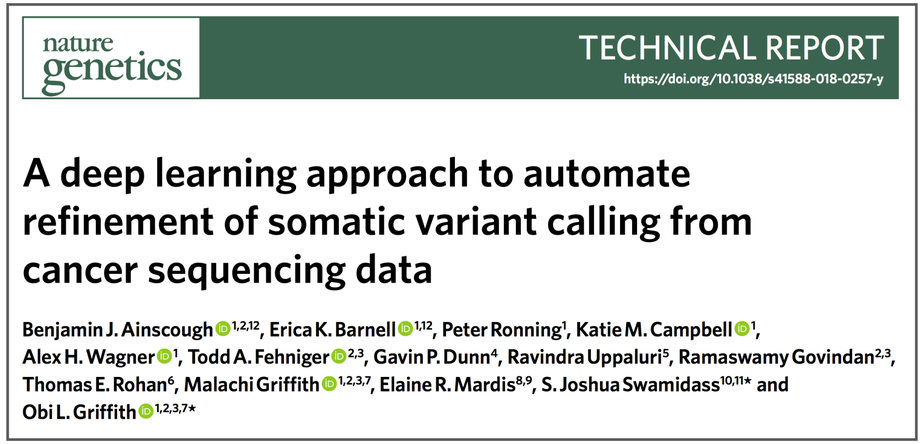


Evolution over time



Next steps

Compare with new paper


nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-018-0257-y

A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data

Benjamin J. Ainscough[1,2,12], Erica K. Barnell[1,12], Peter Ronning[1], Katie M. Campbell[1], Alex H. Wagner[1], Todd A. Fehniger[2,3], Gavin P. Dunn[4], Ravindra Uppaluri[5], Ramaswamy Govindan[2,3], Thomas E. Rohan[6], Malachi Griffith[1,2,3,7], Elaine R. Mardis[8,9], S. Joshua Swamidass[10,11]* and Obi L. Griffith[1,2,3,7]*

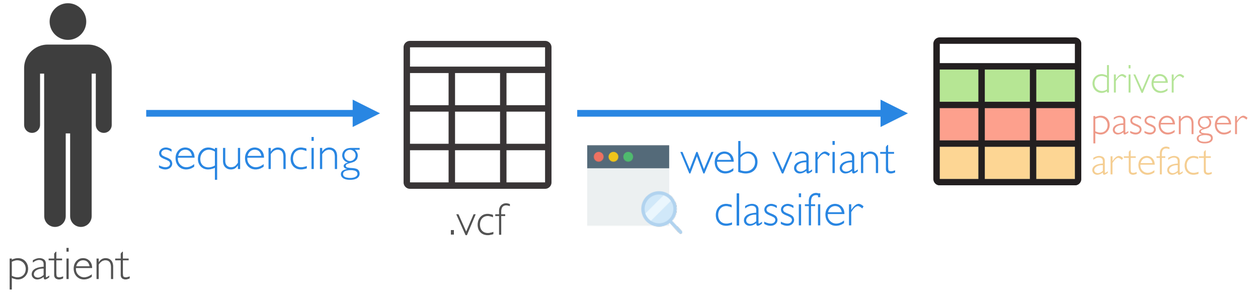
Get detailed calling features

Variant caller flags, read mapping quality, …

Explore new methods

- Deep learning
- Improved under-sampling strategy
- …

Create a 2-steps web-based classifier



Optimisation of clinical heme panel