Memorial Sloan Kettering
Cancer Center

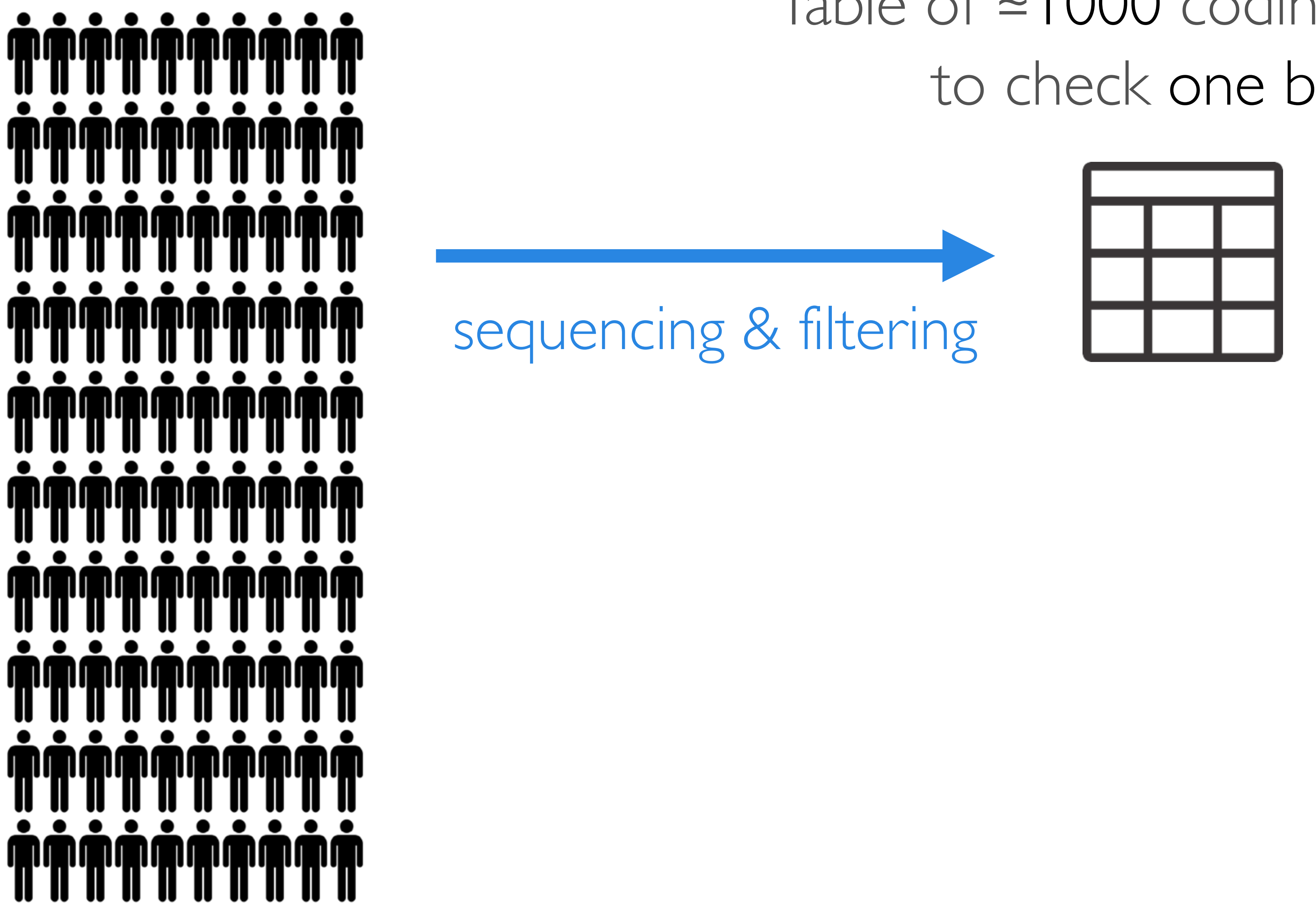
Variant classification

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?

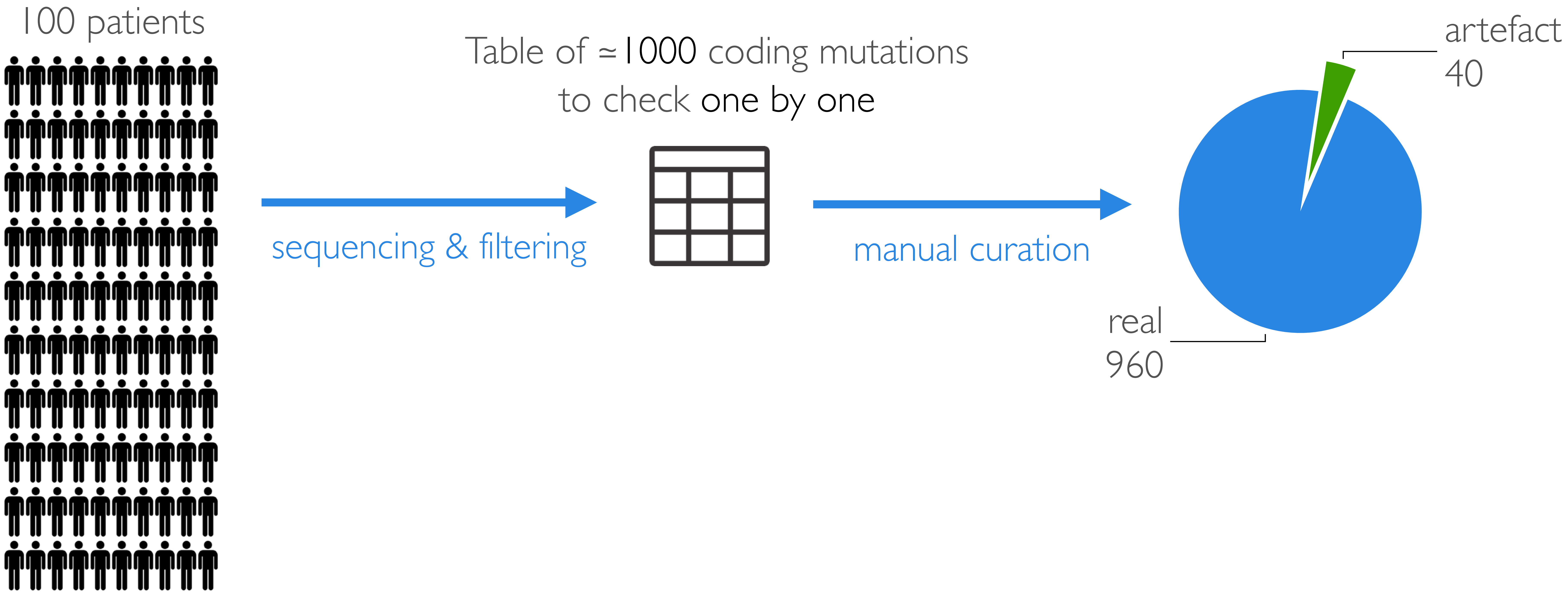


100 patients

sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

A variant classifier, why is it important?



100 patients

Table of ≃1000 coding mutations
to check one by one

sequencing & filtering

manual curation

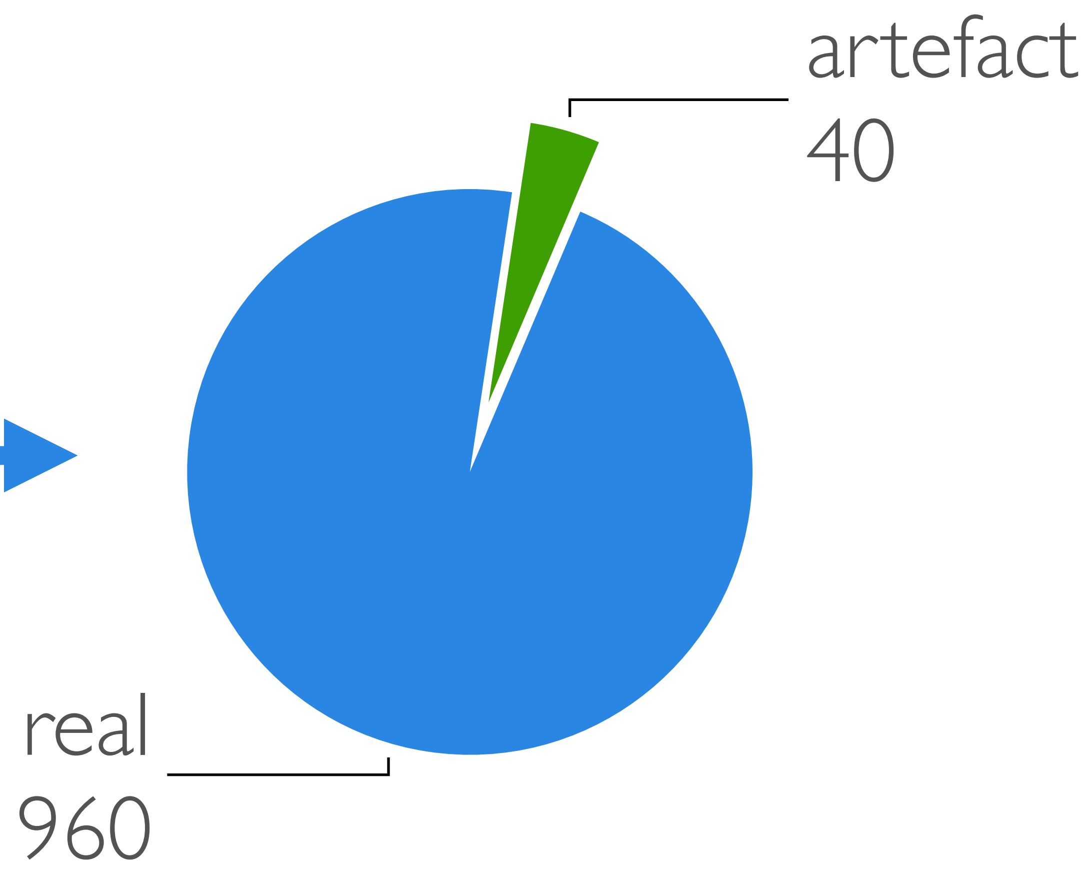
A variant classifier, why is it important?

100 patients

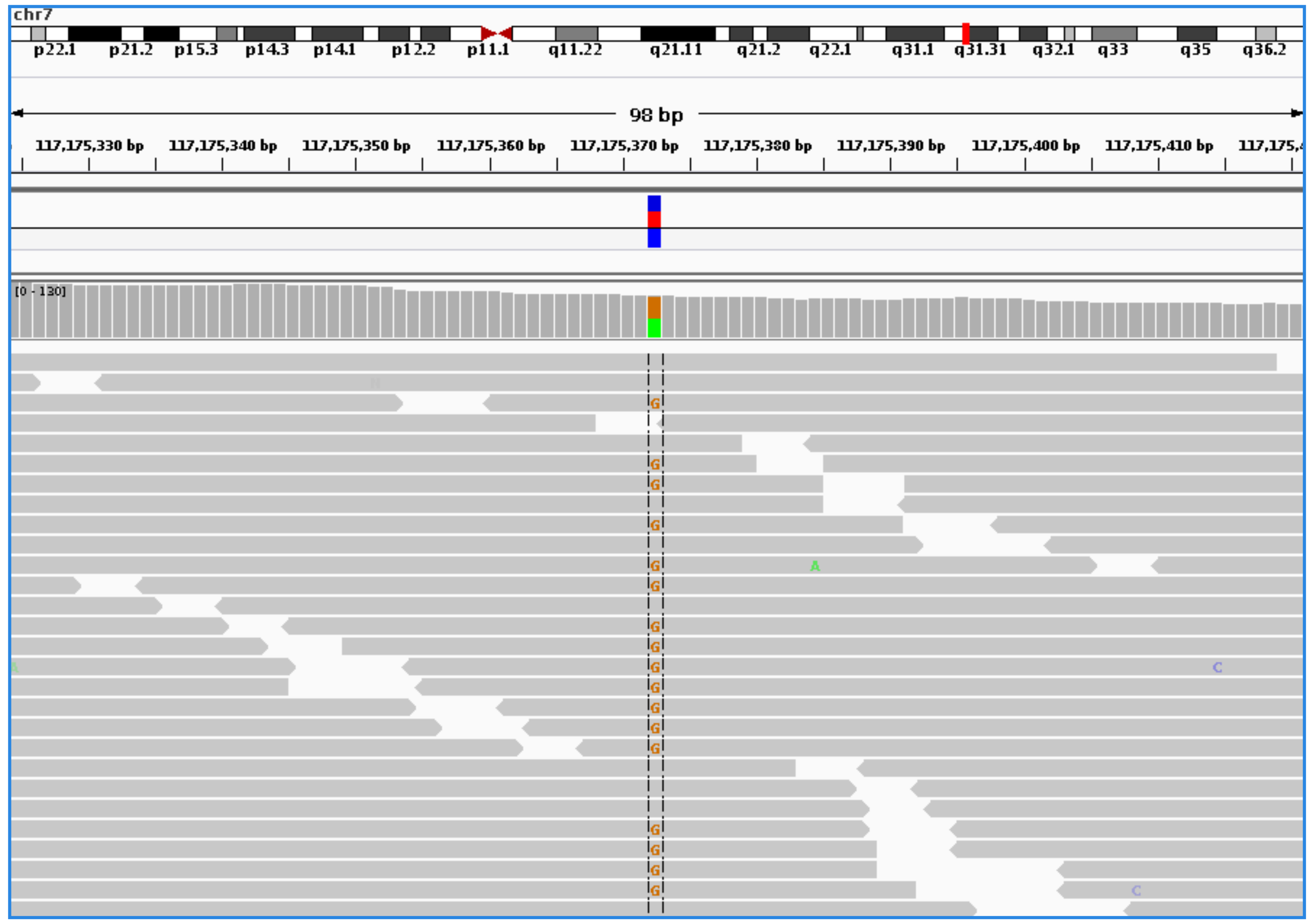
sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation

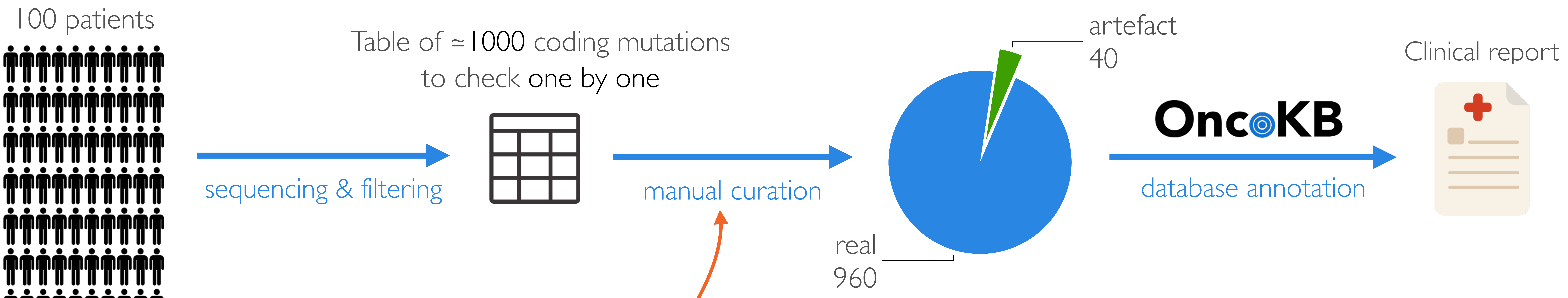


Expert work



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

artefact
40

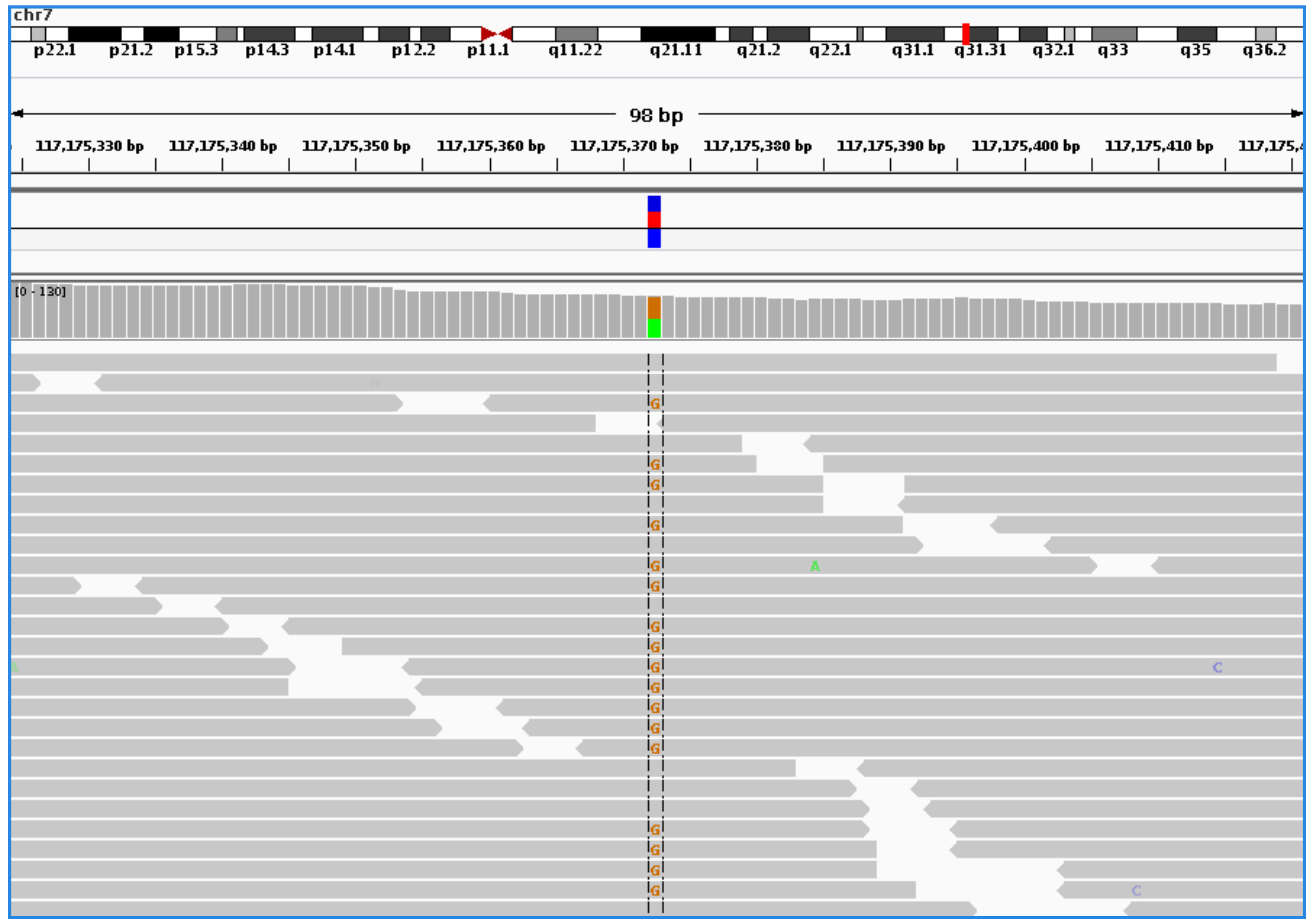
real
960

OncoKB

database annotation

Clinical report

Expert work



IGV snapshot

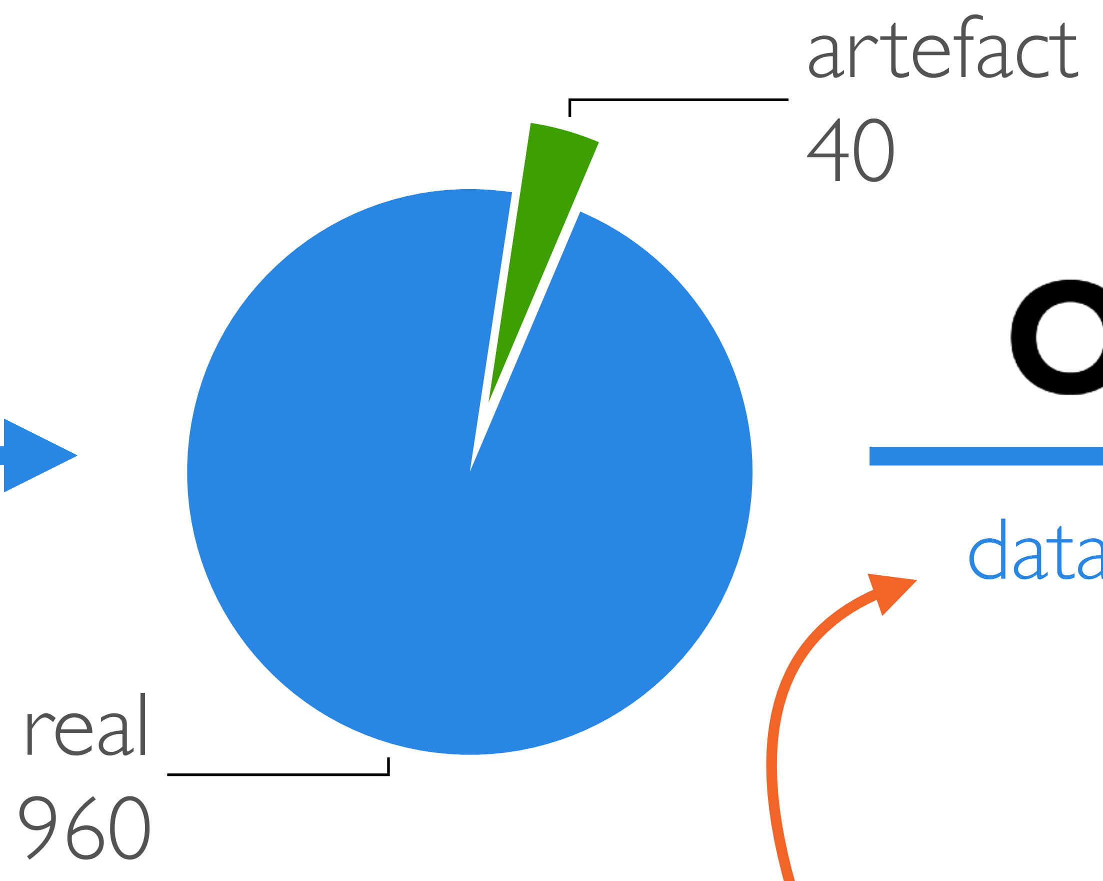
A variant classifier, why is it important?

100 patients

sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation

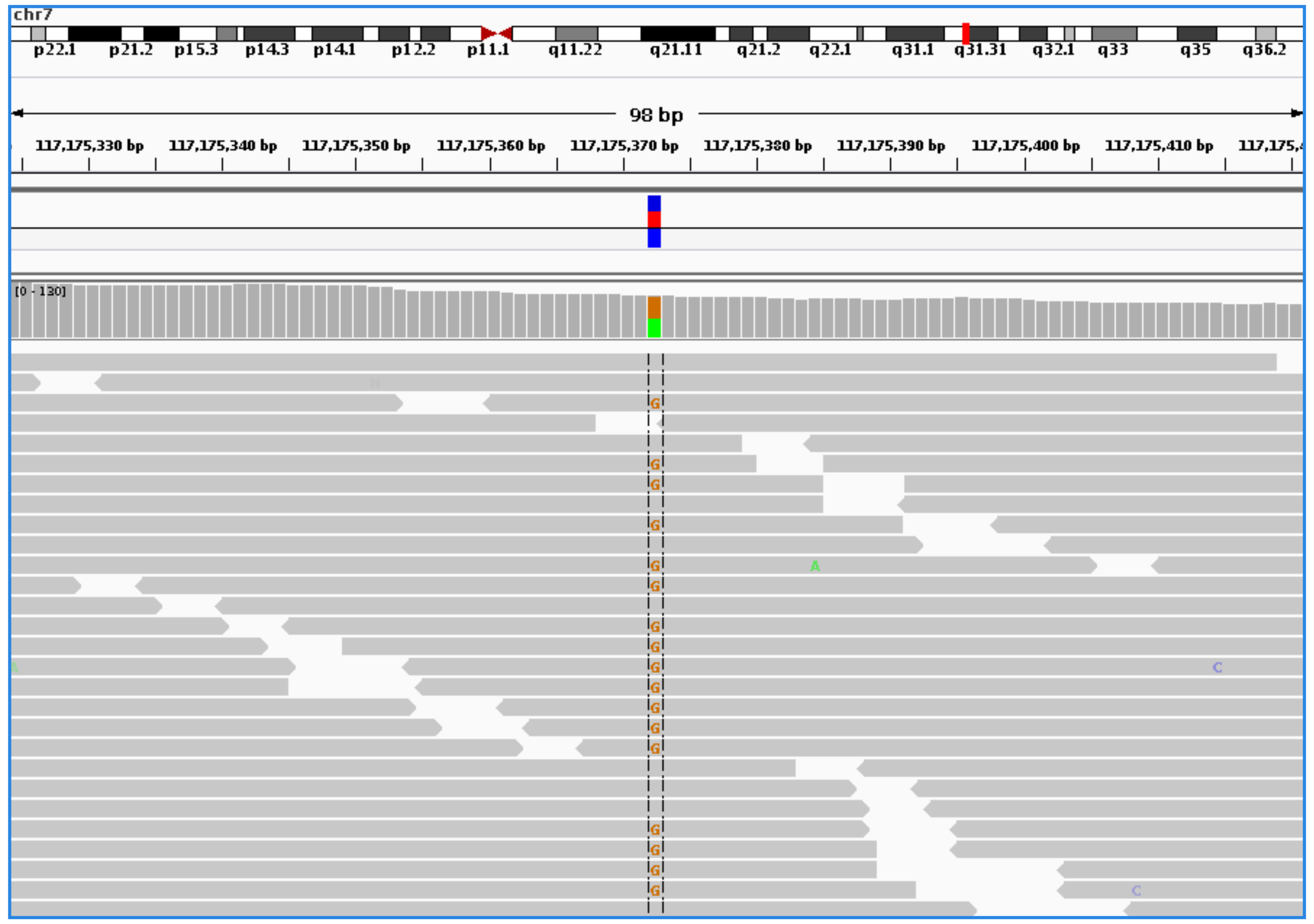


OncoKB

database annotation

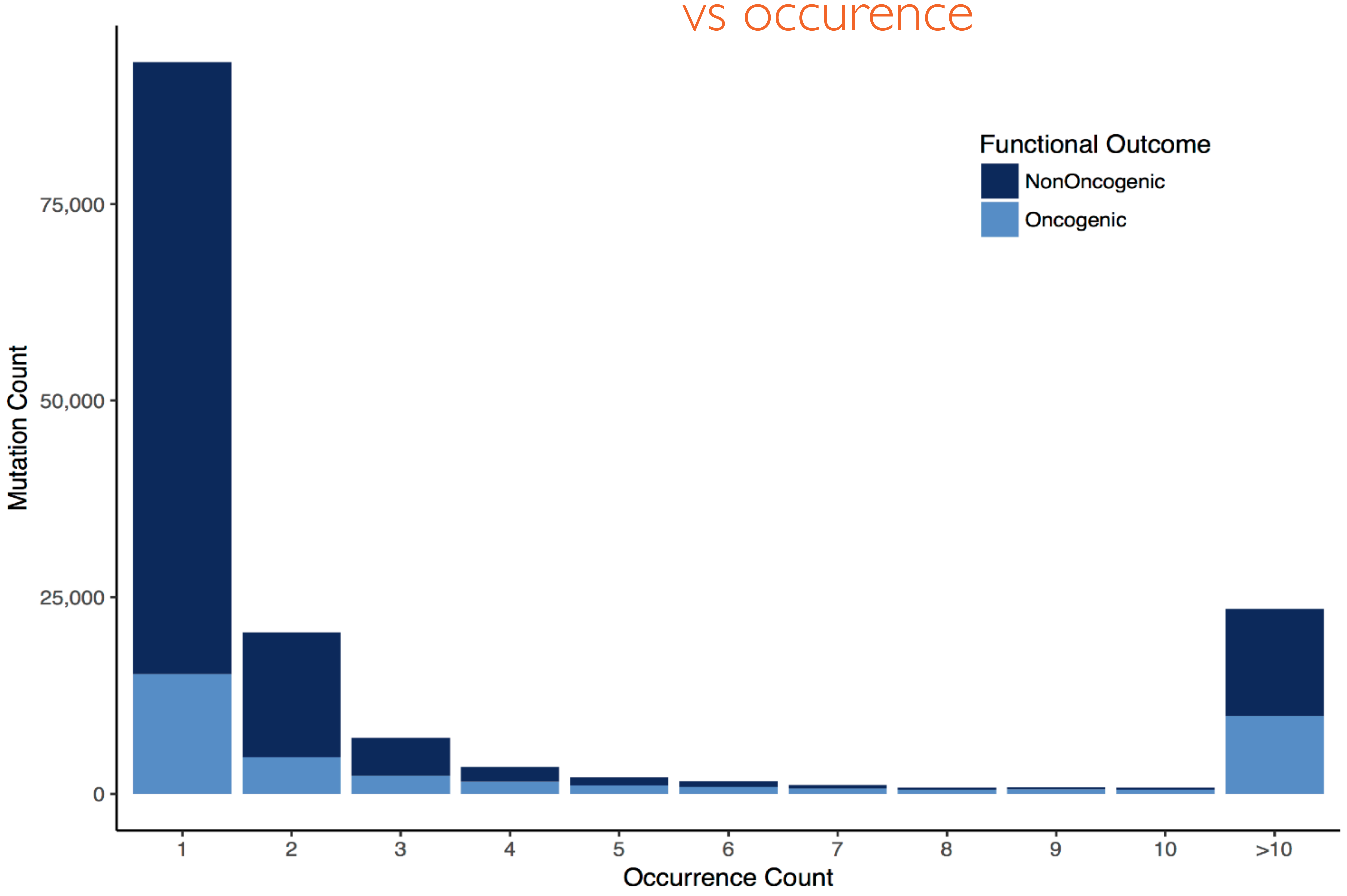
Clinical report

Expert work



IGV snapshot

MSK-IMPACT: oncogenicity
vs occurence

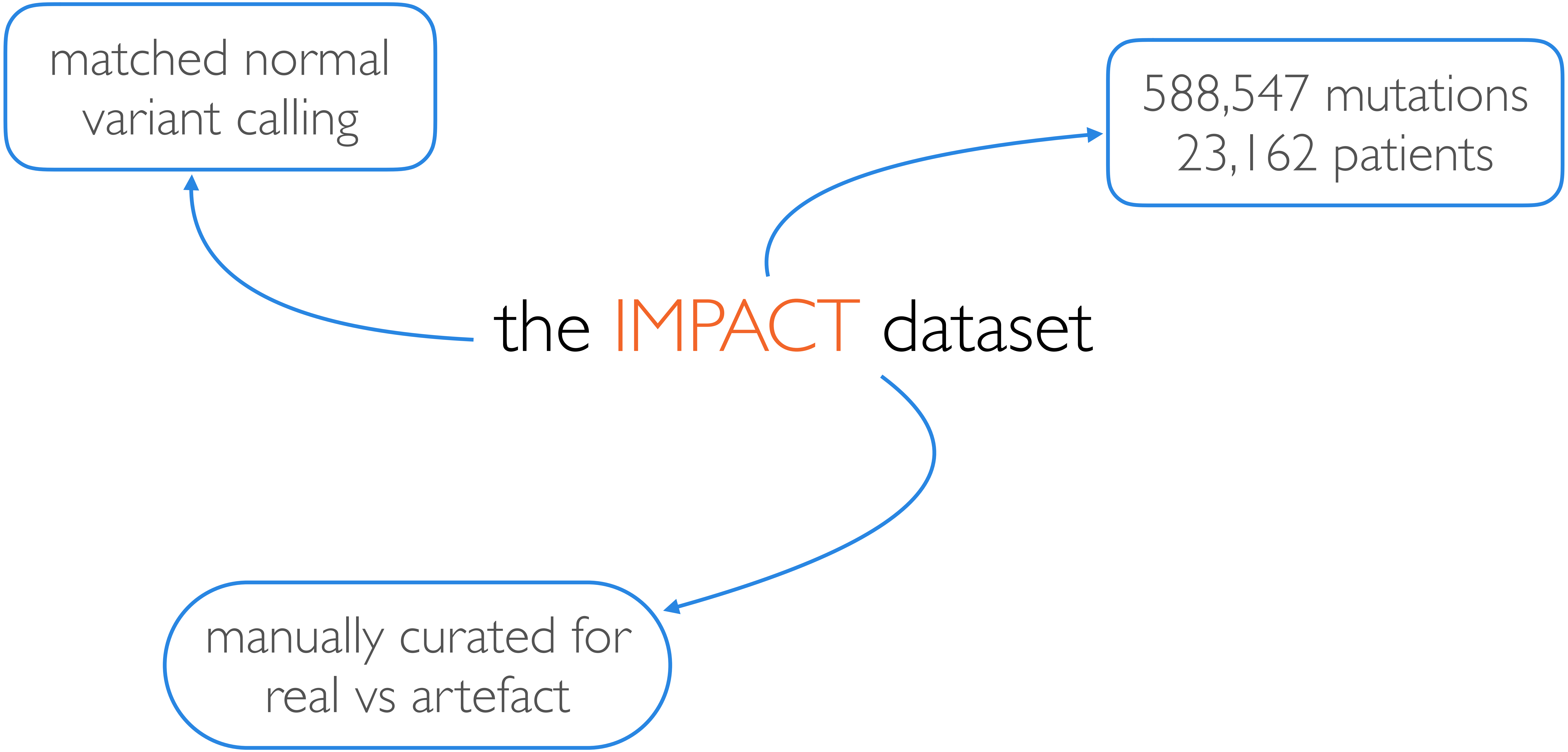


The goal

Create a tool that classifies variant automatically

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

IMPACT, the dataset



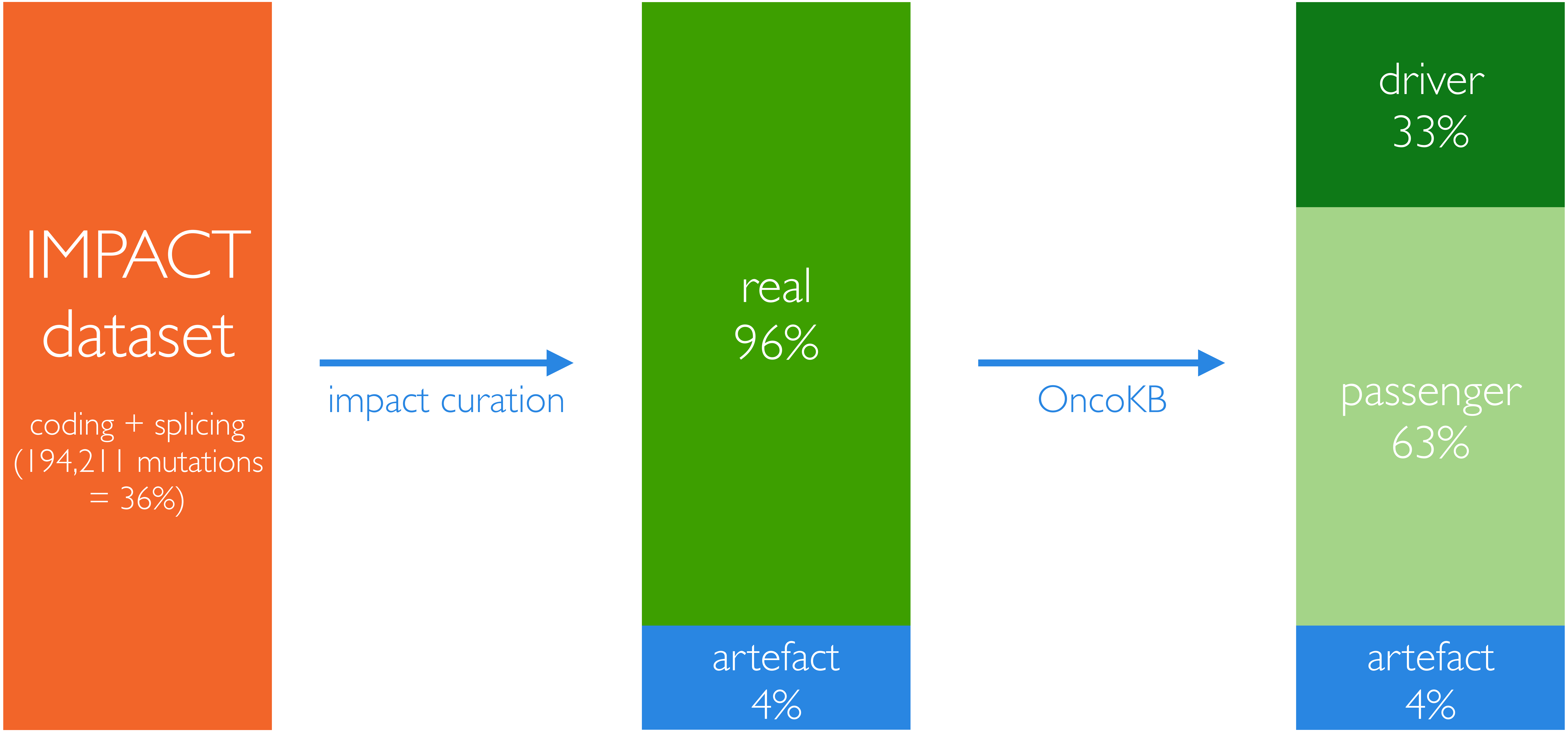
matched normal variant calling

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%)

impact curation

real 96%

artefact 4%

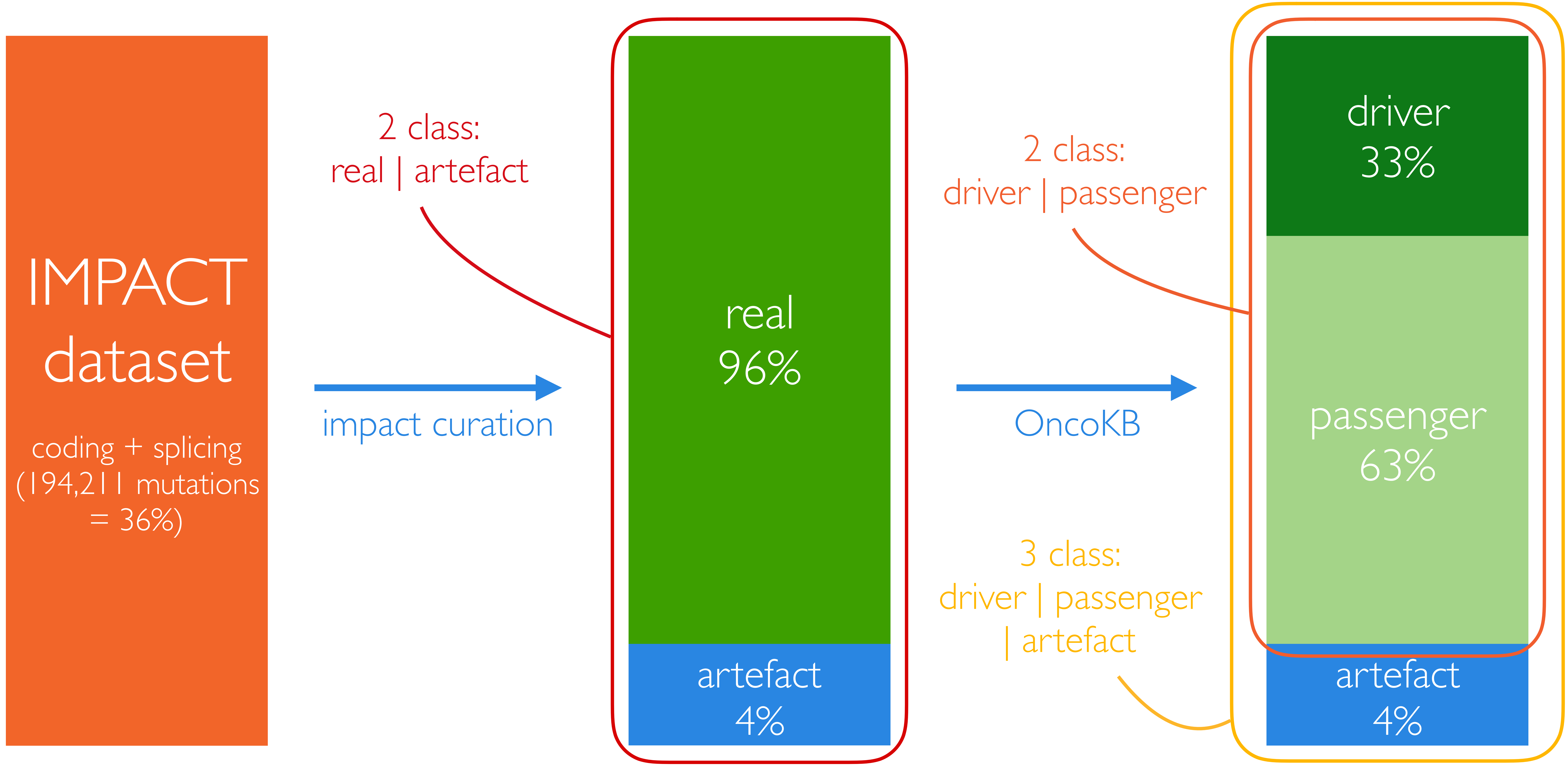
OncoKB

driver 33%

passenger 63%

artefact 4%

Two steps classification



IMPACT dataset

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

impact curation

2 class:
real | artefact

real 96%

artefact 4%

OncoKB

2 class:
driver | passenger

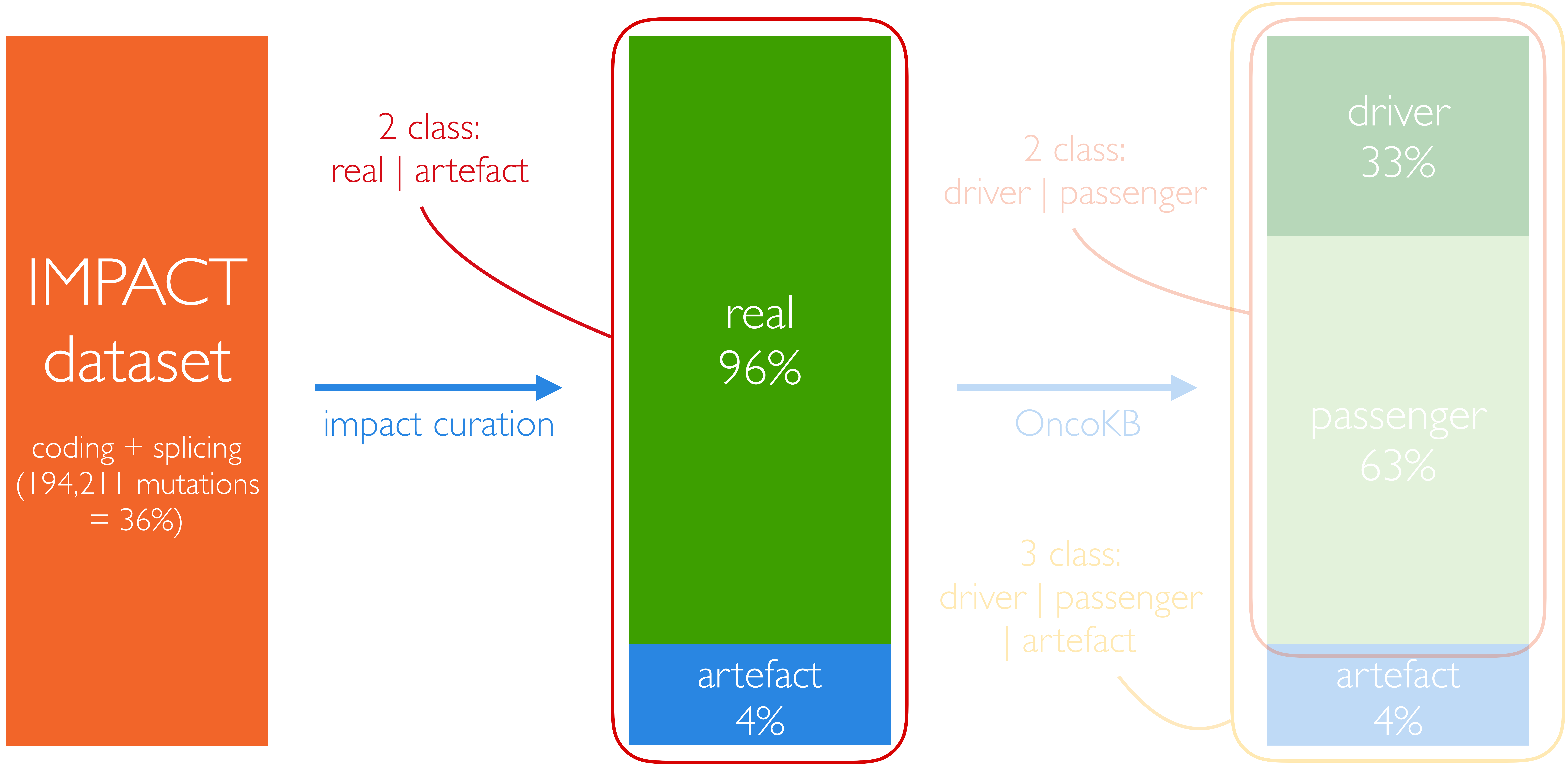
3 class:
driver | passenger | artefact

driver 33%

passenger 63%

artefact 4%

Two steps classification



IMPACT dataset

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

impact curation

2 class: real | artefact

real 96%

artefact 4%

OncoKB

2 class: driver | passenger

3 class: driver | passenger | artefact

driver 33%

passenger 63%

artefact 4%

The features used in our model

- Sequencing features (n = 11)

  Tumor VAF, tumor depth

- Genomic coordinates (n = 3)

  Chromosome, Hugo Symbol

- Control populations (n = 12)

  Population based
  GnomAD allele frequency

- Cancer populations (n = 4)

  COSMIC, OncoKB

- Normal control (n = 1)

  Frequency in normal control

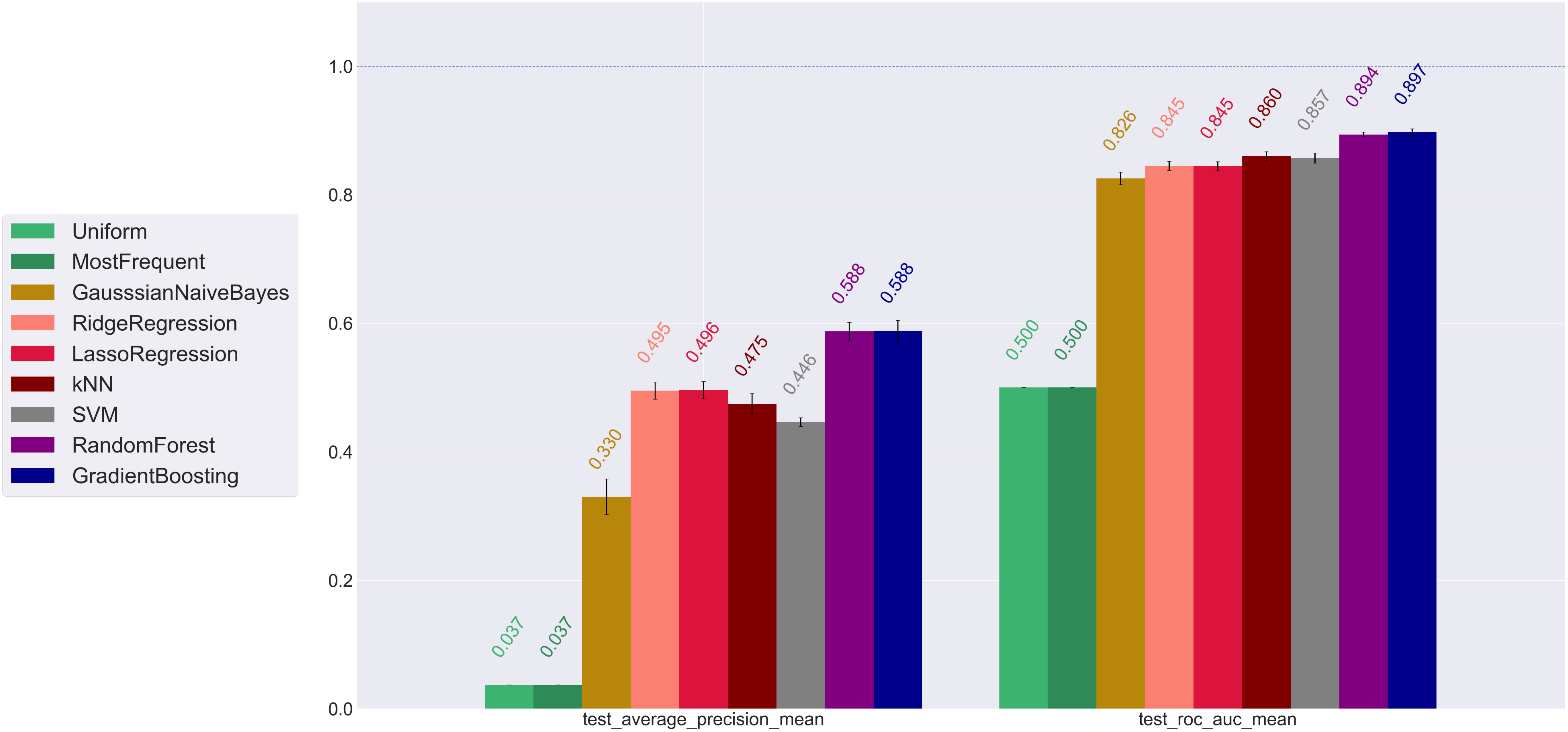
- Mutation consequence (n = 6)

  Protein effect, SIFT & PolyPhen class

Algorithm comparison

- 🟩 Uniform
- 🟩 MostFrequent
- 🟨 GausssianNaiveBayes
- 🟧 RidgeRegression
- 🟥 LassoRegression
- 🟥 kNN
- ⬜ SVM
- 🟪 RandomForest
- 🟦 GradientBoosting

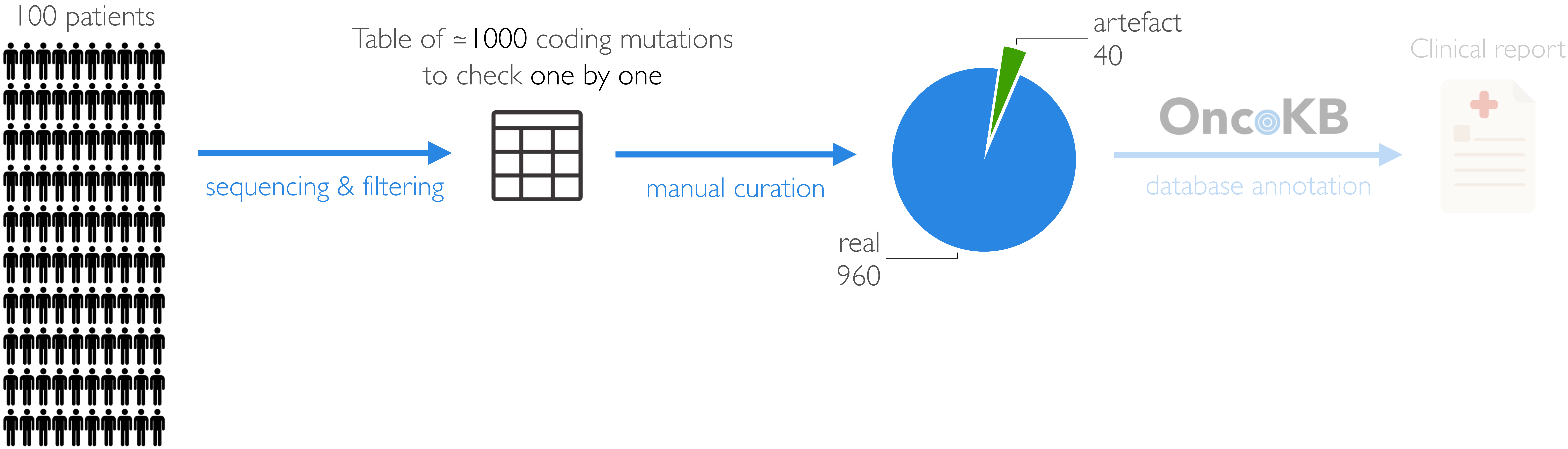
Algorithm comparison



Best algorithm probability output

artefact
real
density
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
predicted probability

The variant classifier performances



100 patients

sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation

artefact
40

real
960

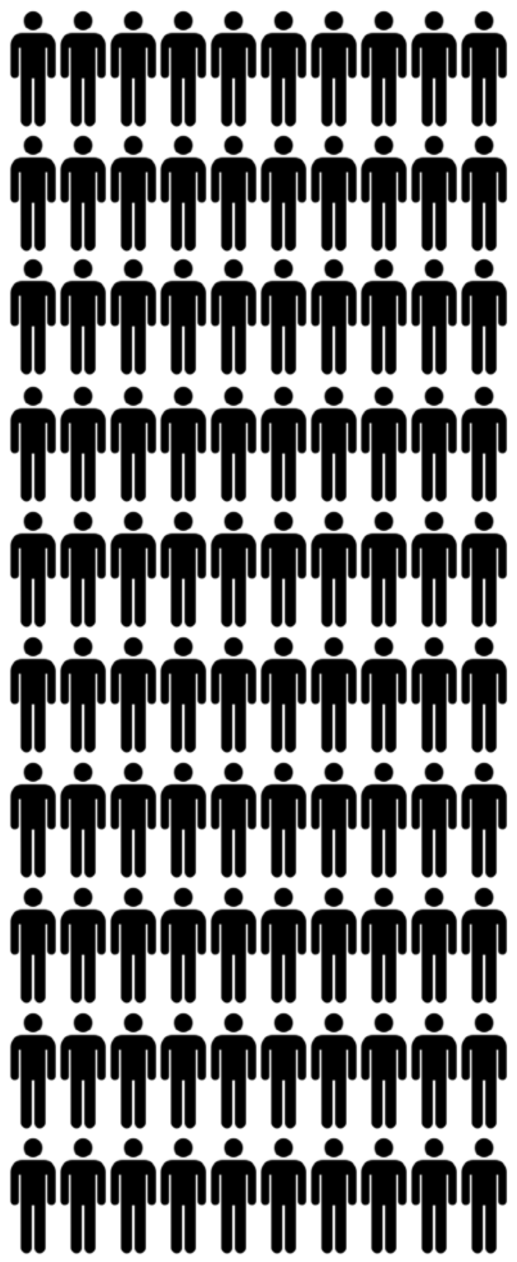
OncoKB

database annotation

Clinical report

The variant classifier performances

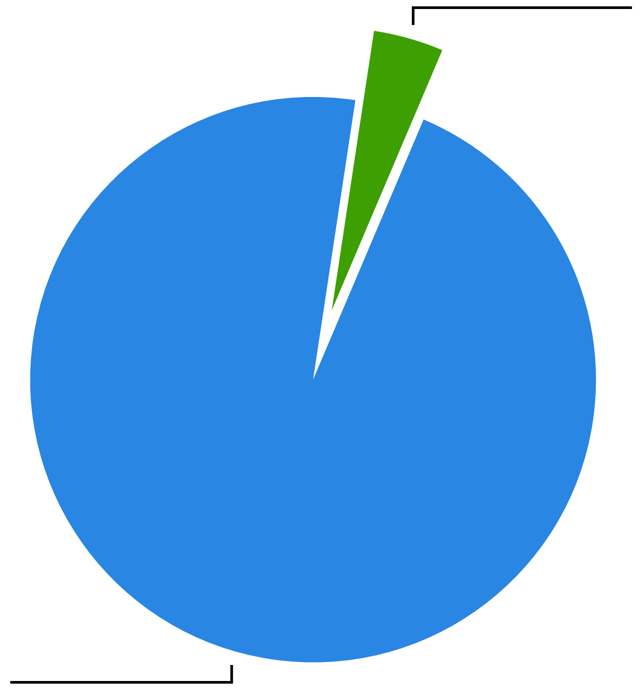
100 patients



sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation



artefact
40

real
960

OncoKB

database annotation

Clinical report

variant
classifier

960 real

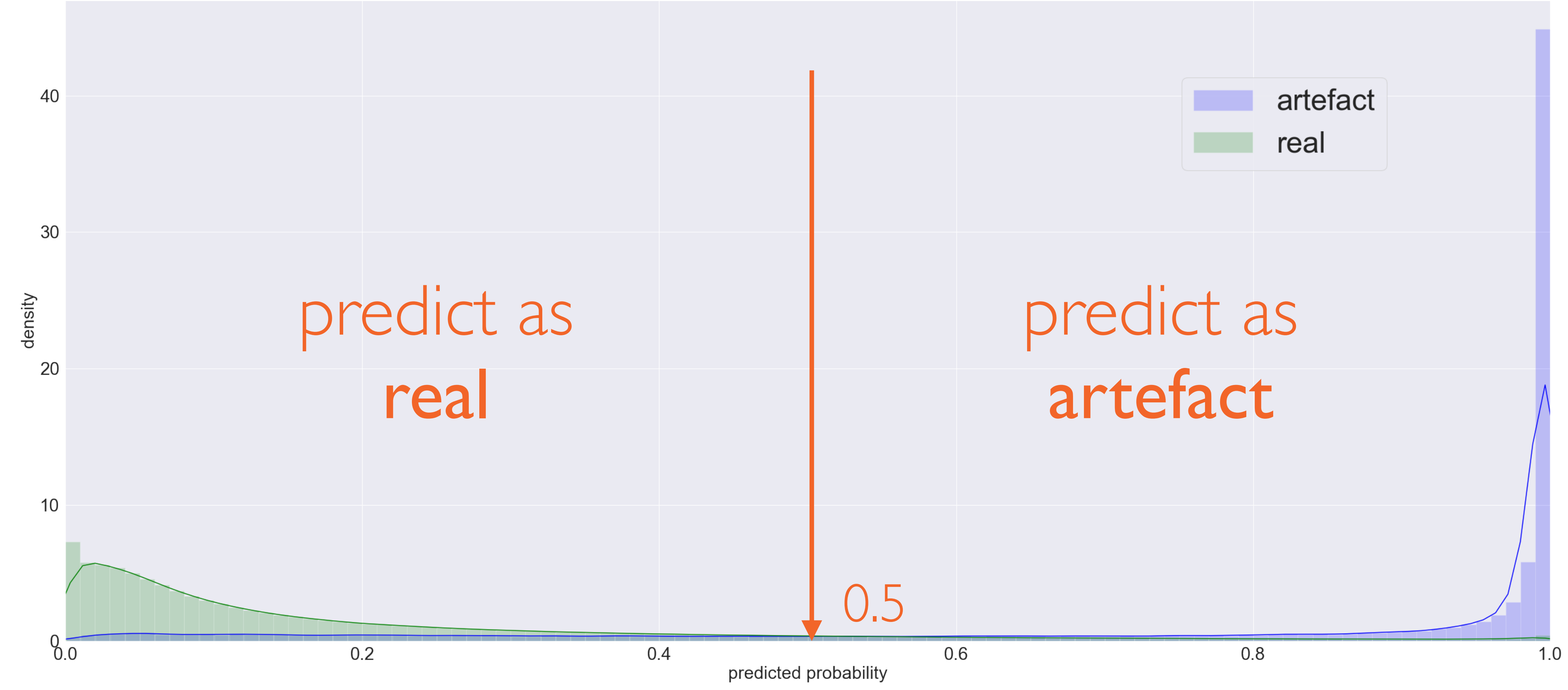
✅ 848 → real

❌ 112 → artefact

40 artefact

✅ 32 → artefact

❌ 8 → real

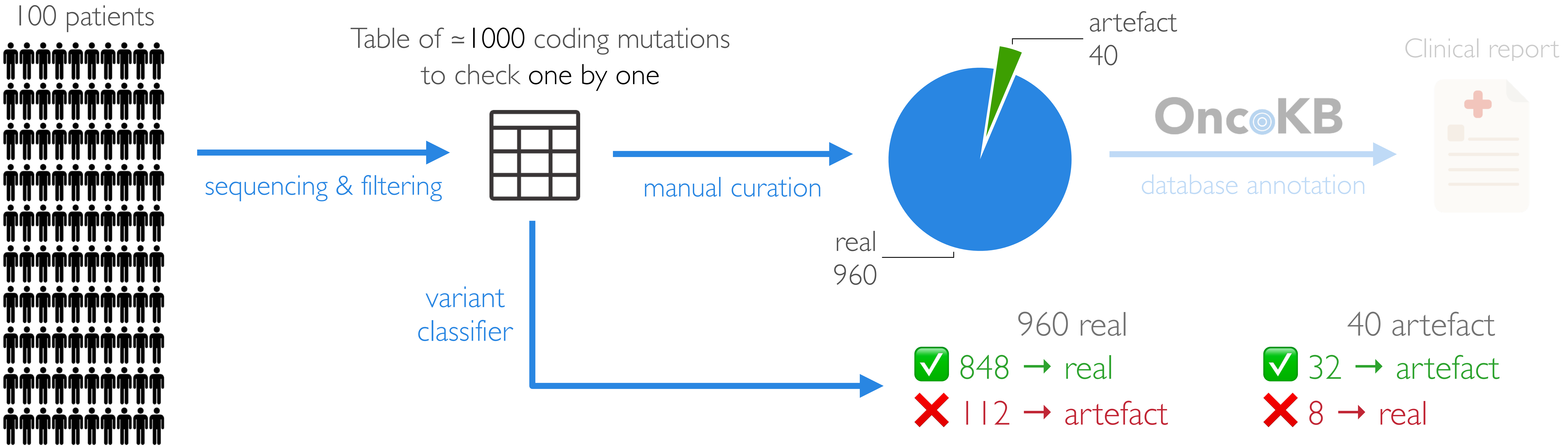


predict as
real

predict as
artefact

0.5

The variant classifier performances



100 patients

Table of ≃1000 coding mutations to check one by one

sequencing & filtering

manual curation

artefact 40

real 960

OncoKB

database annotation

Clinical report

variant classifier

960 real

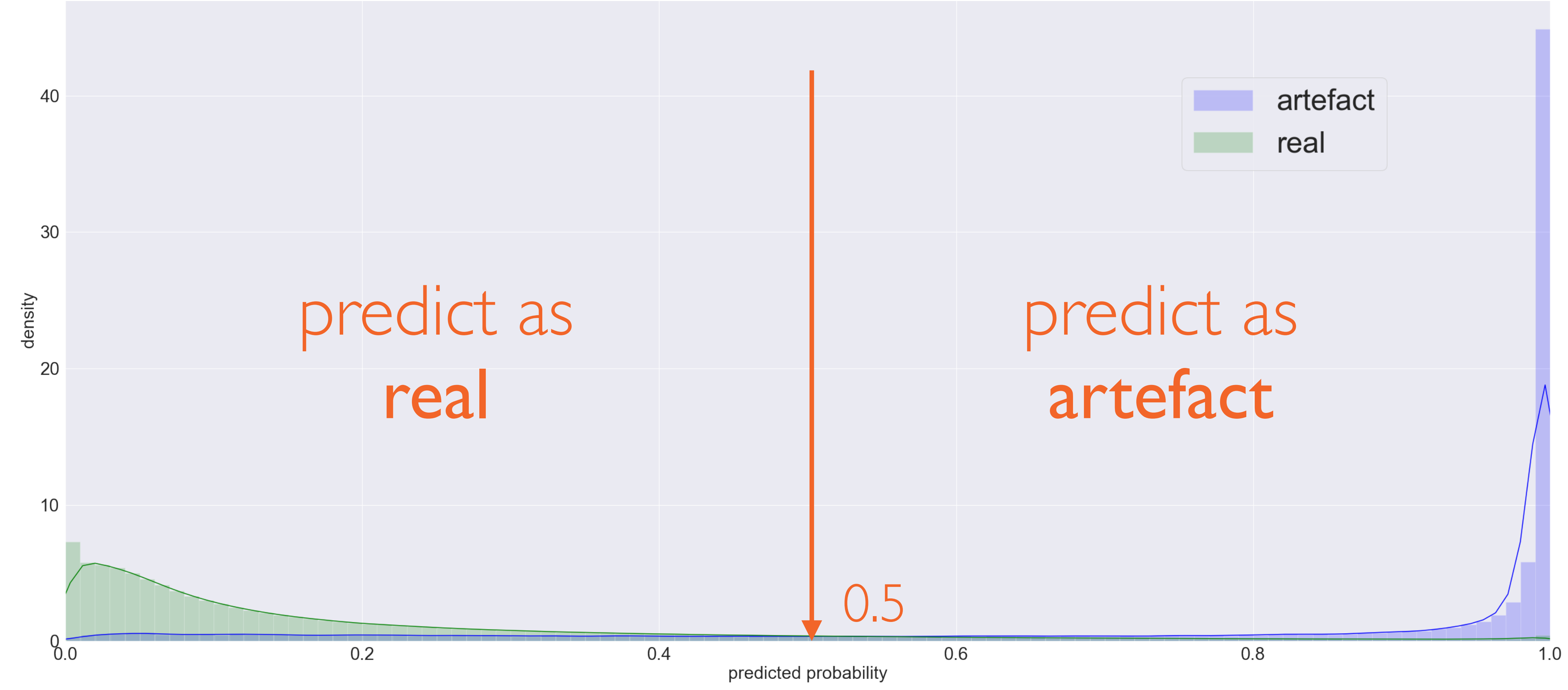
✅ 848 → real

❌ 112 → artefact

40 artefact

✅ 32 → artefact

❌ 8 → real



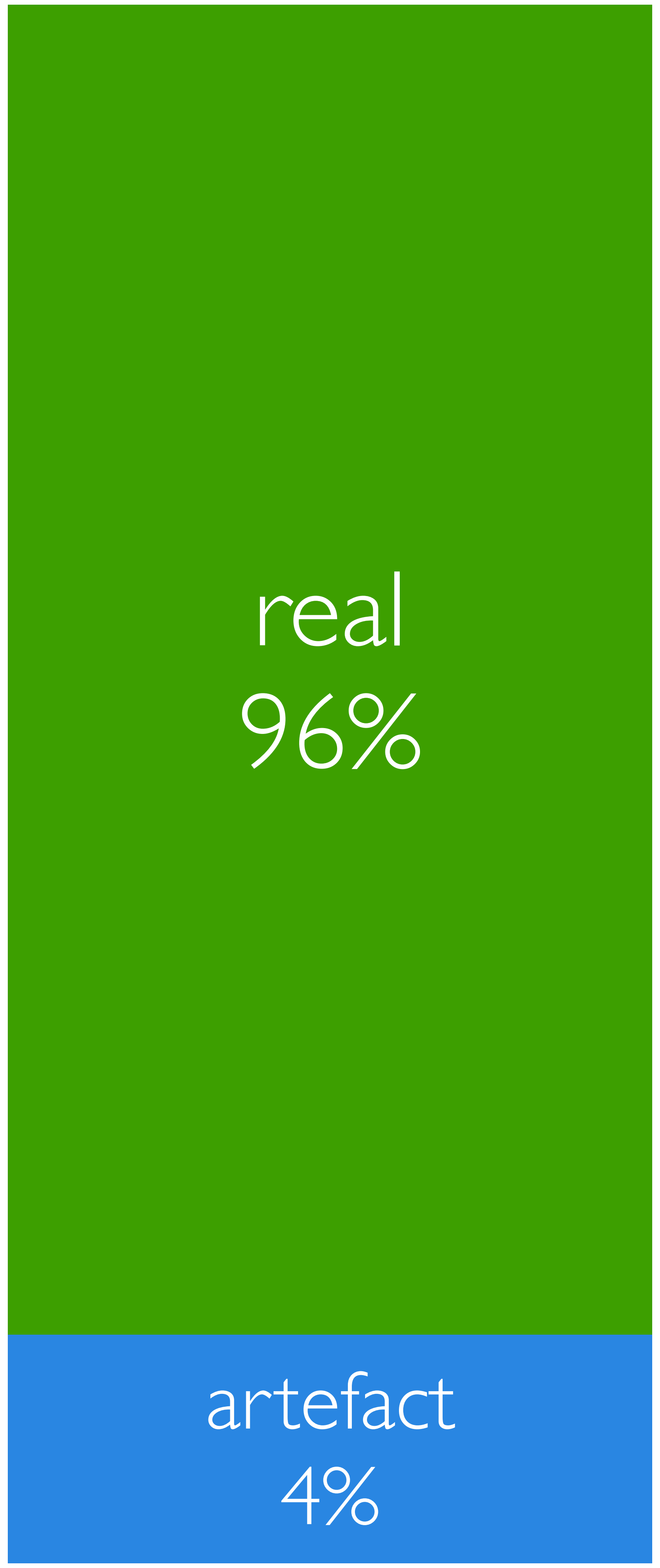
❌ 8/40 artefacts considered as real

📝 8x less work

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

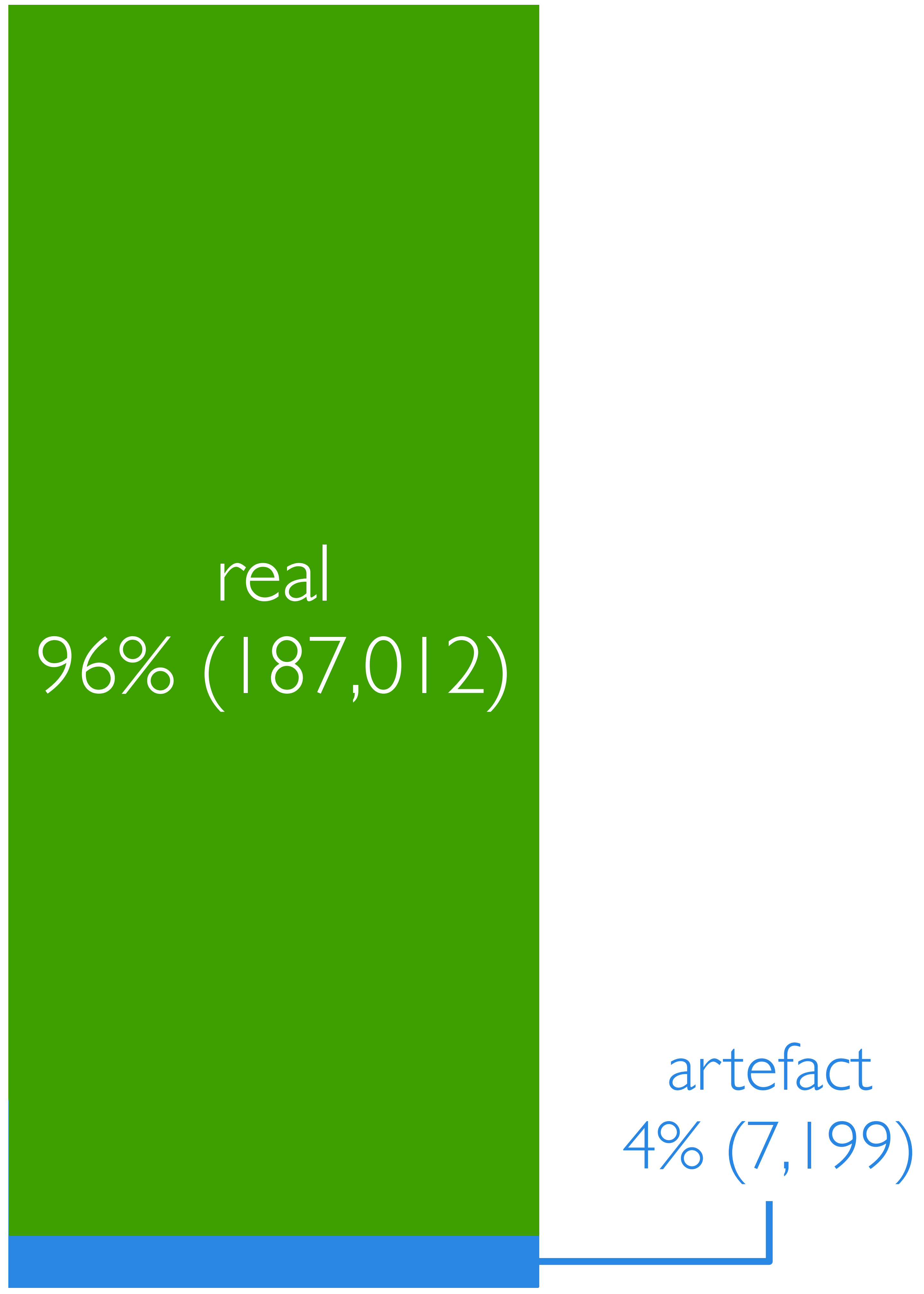
Main challenges

Imbalanced dataset



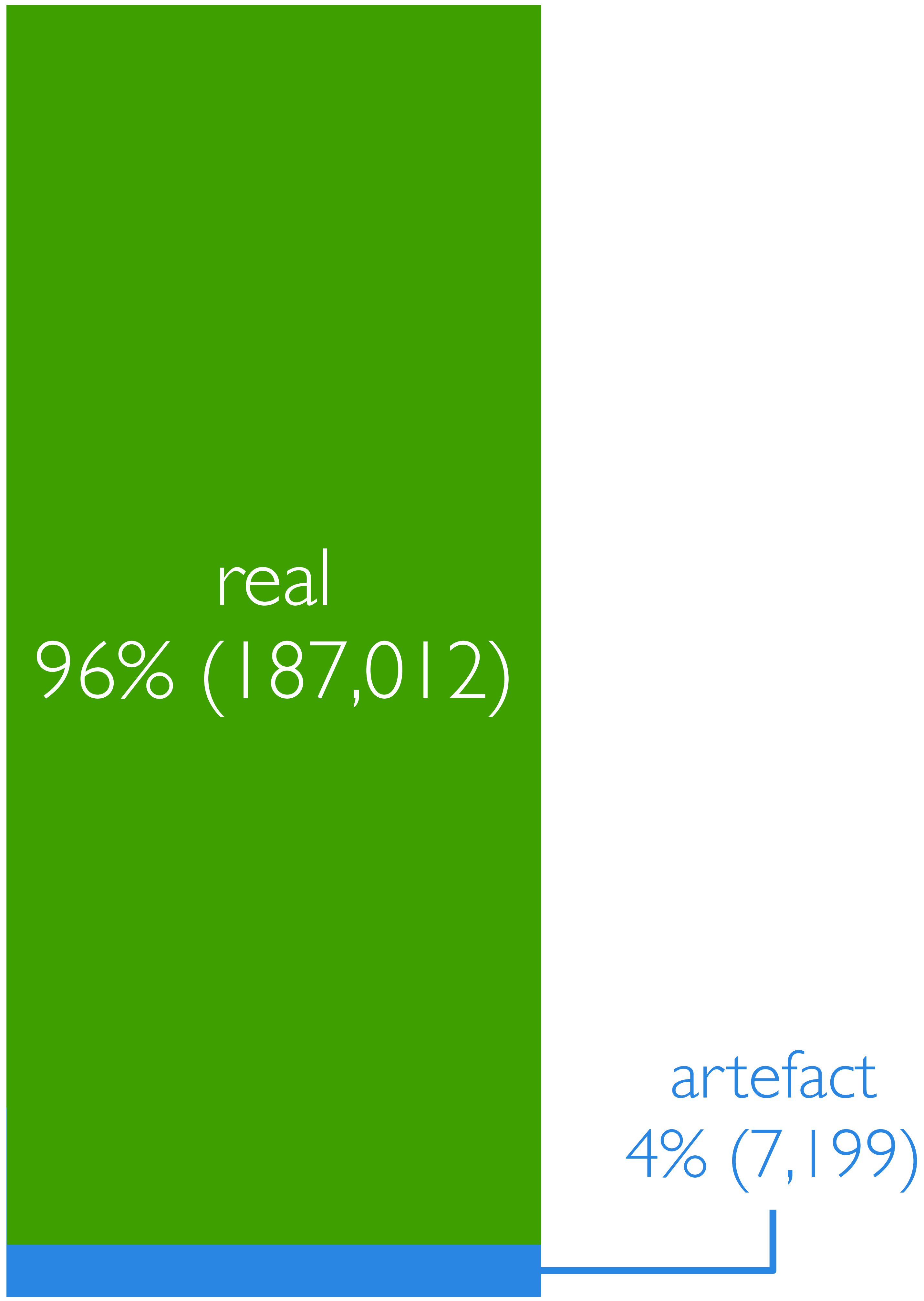
Main challenges

Imbalanced dataset

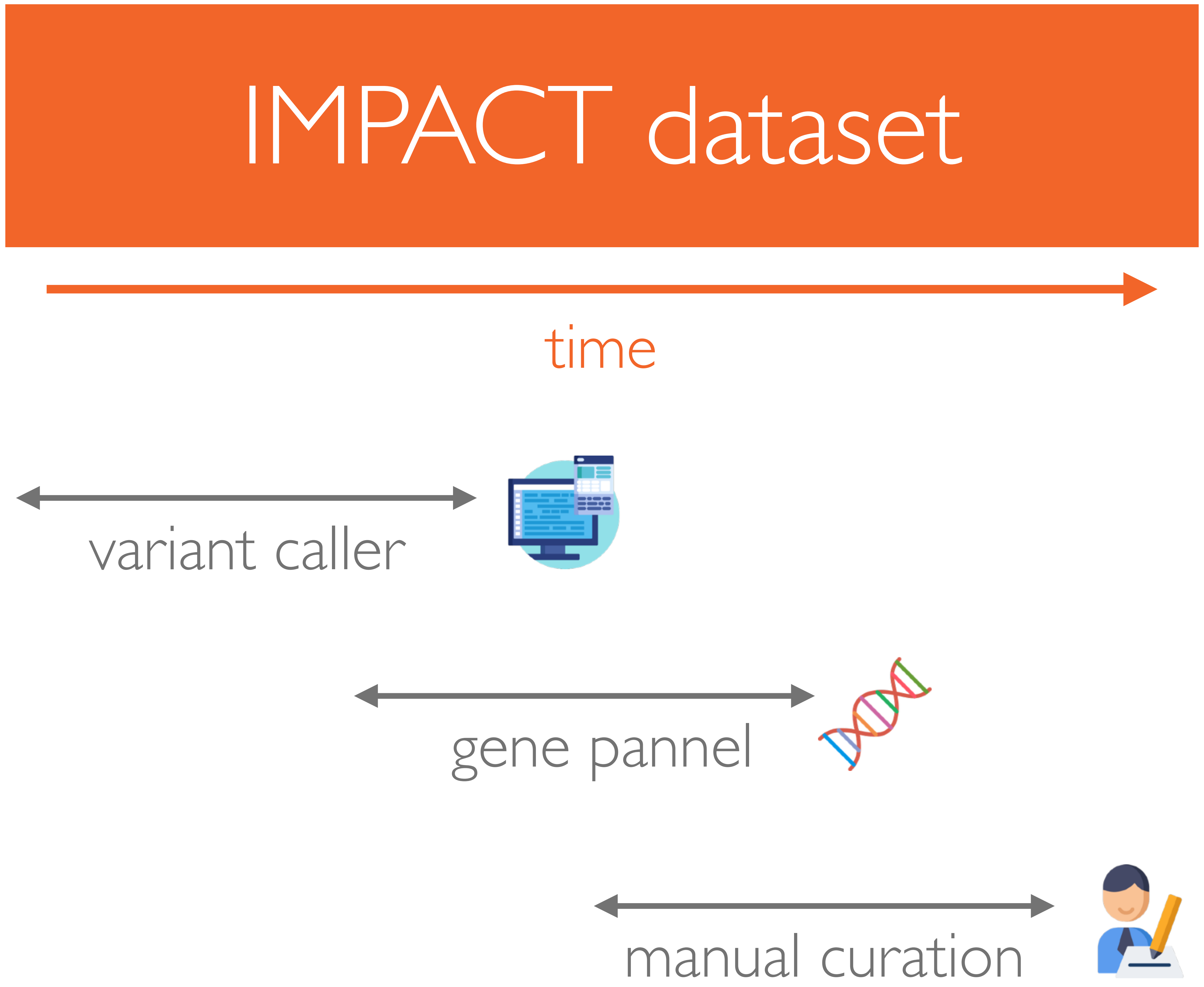


Main challenges

Imbalanced dataset



Evolution over time



IMPACT dataset

time

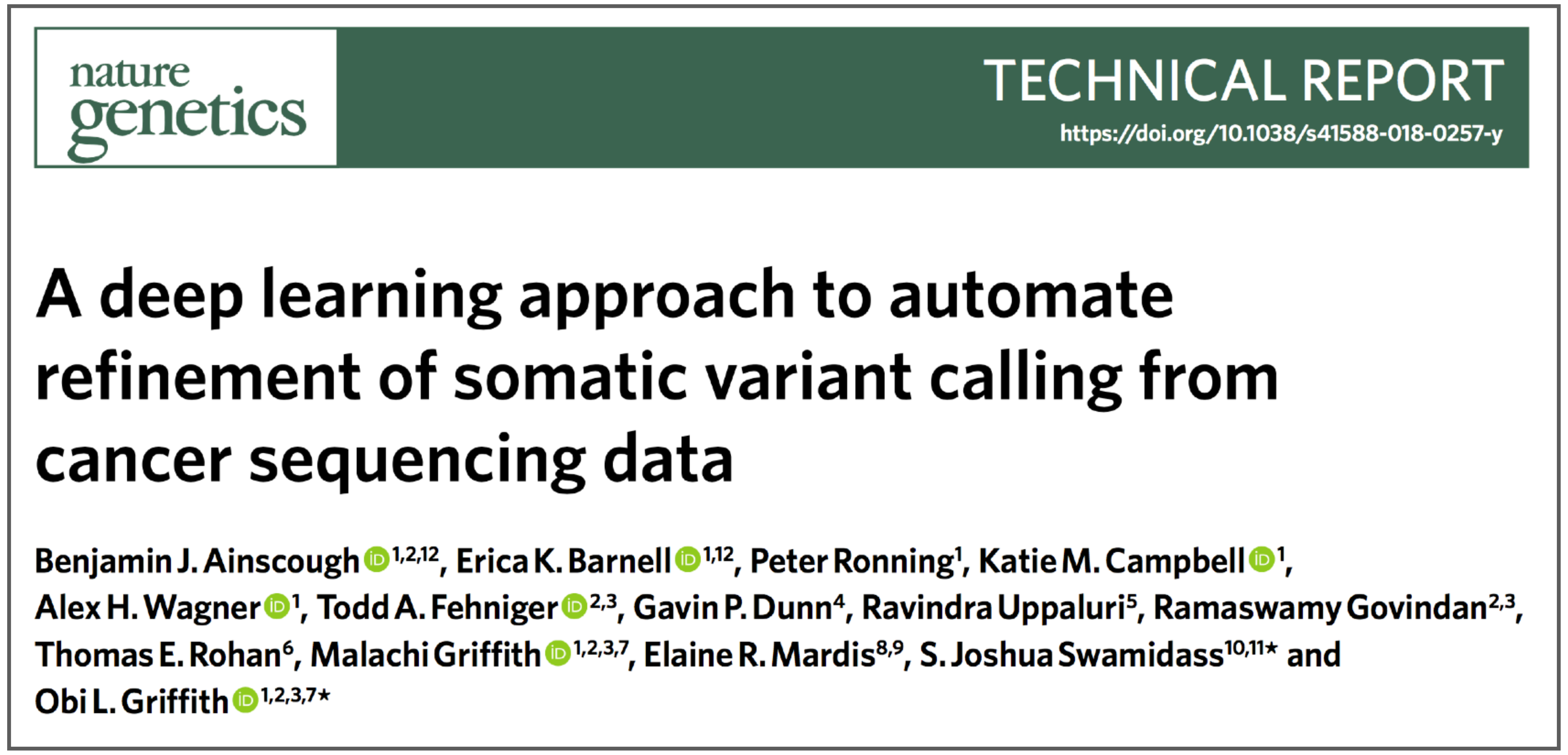
variant caller

gene pannel

manual curation

Next steps

Method comparison



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*]

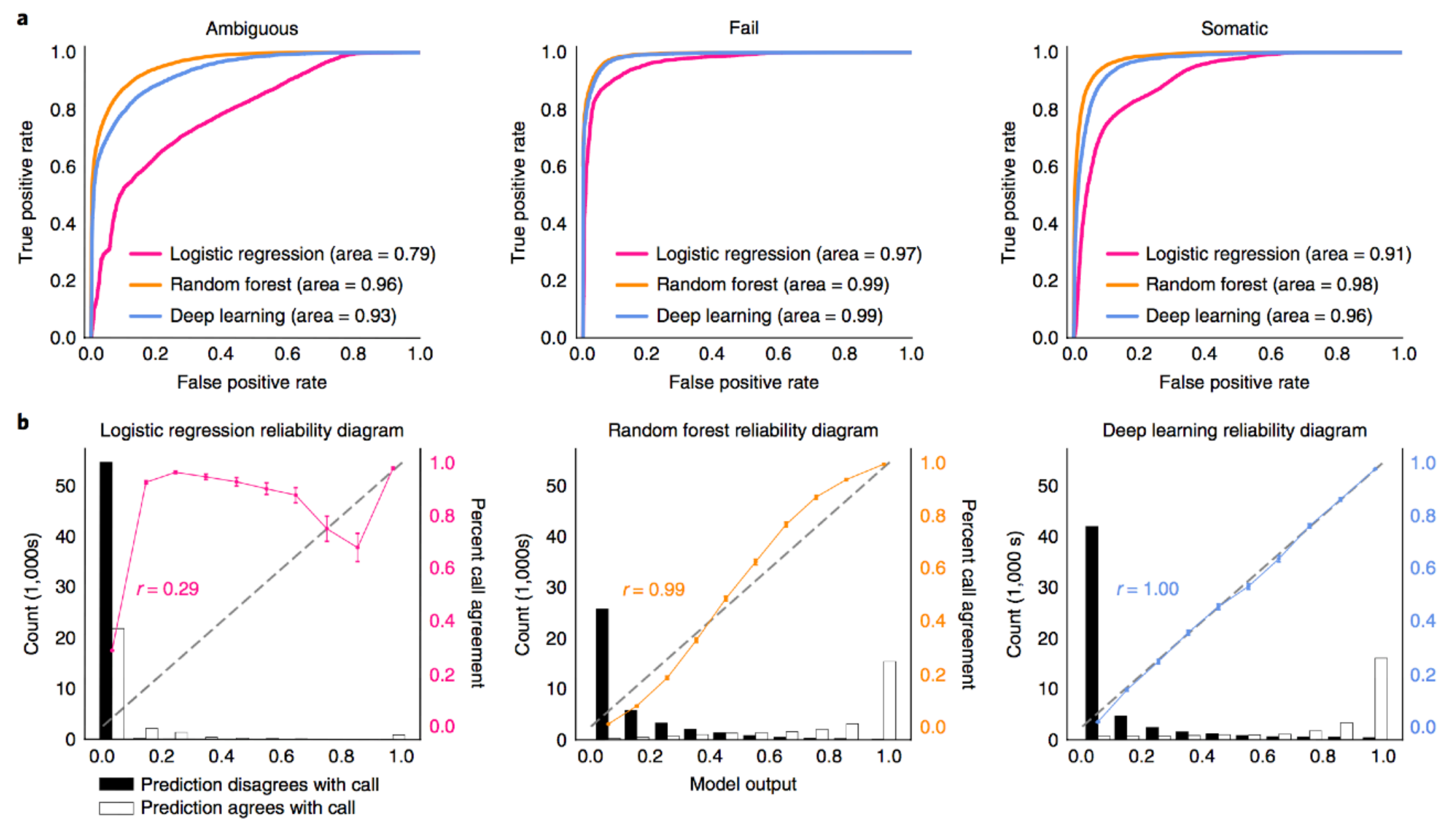
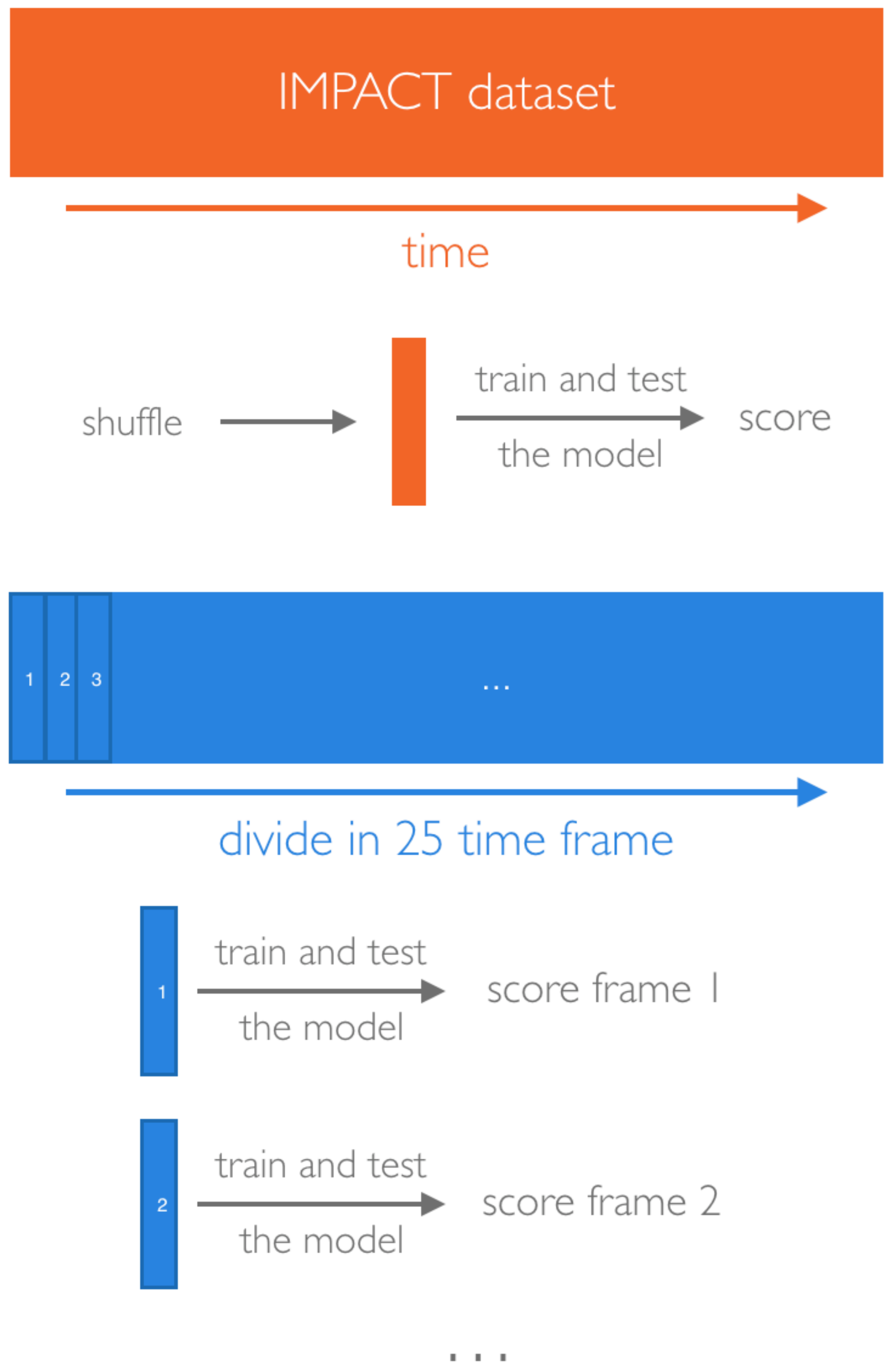
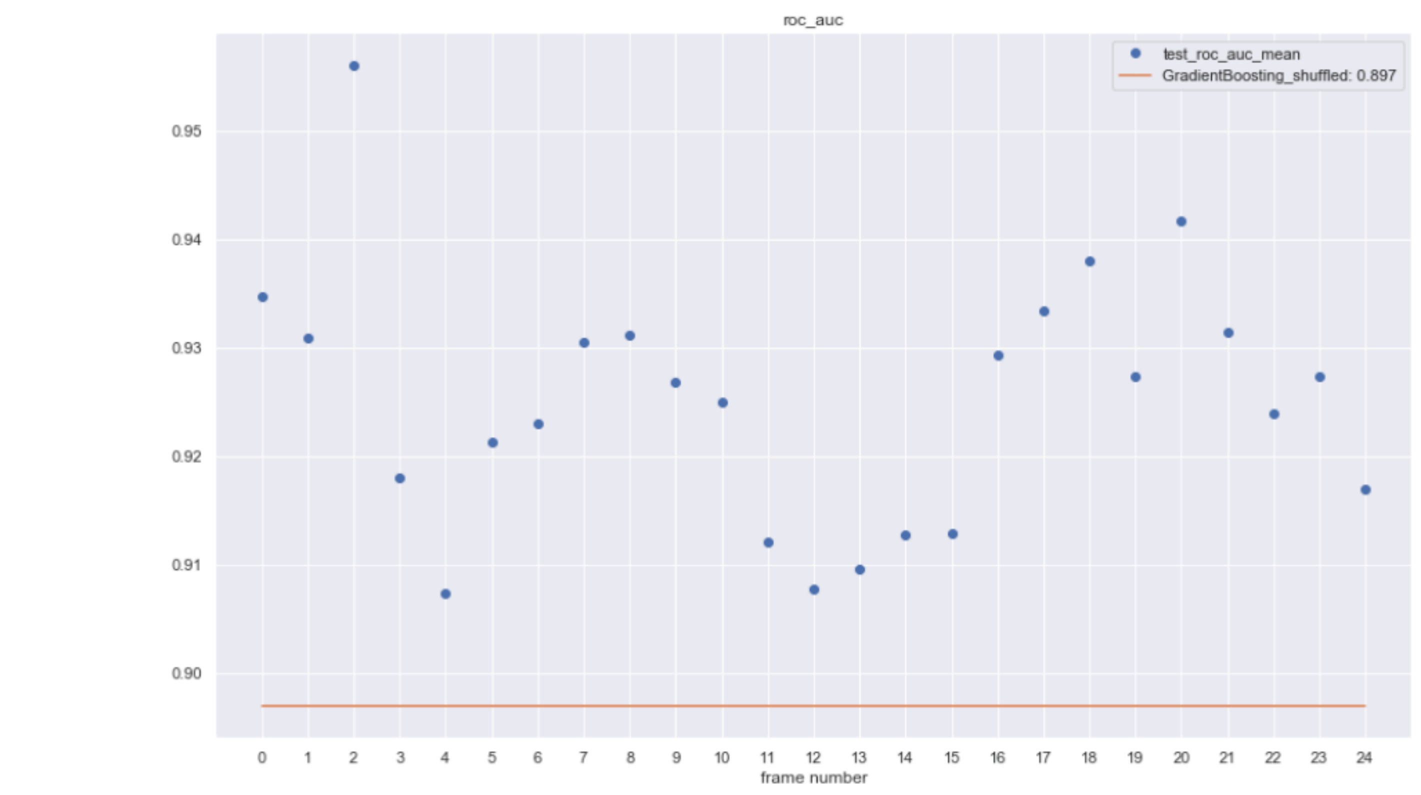
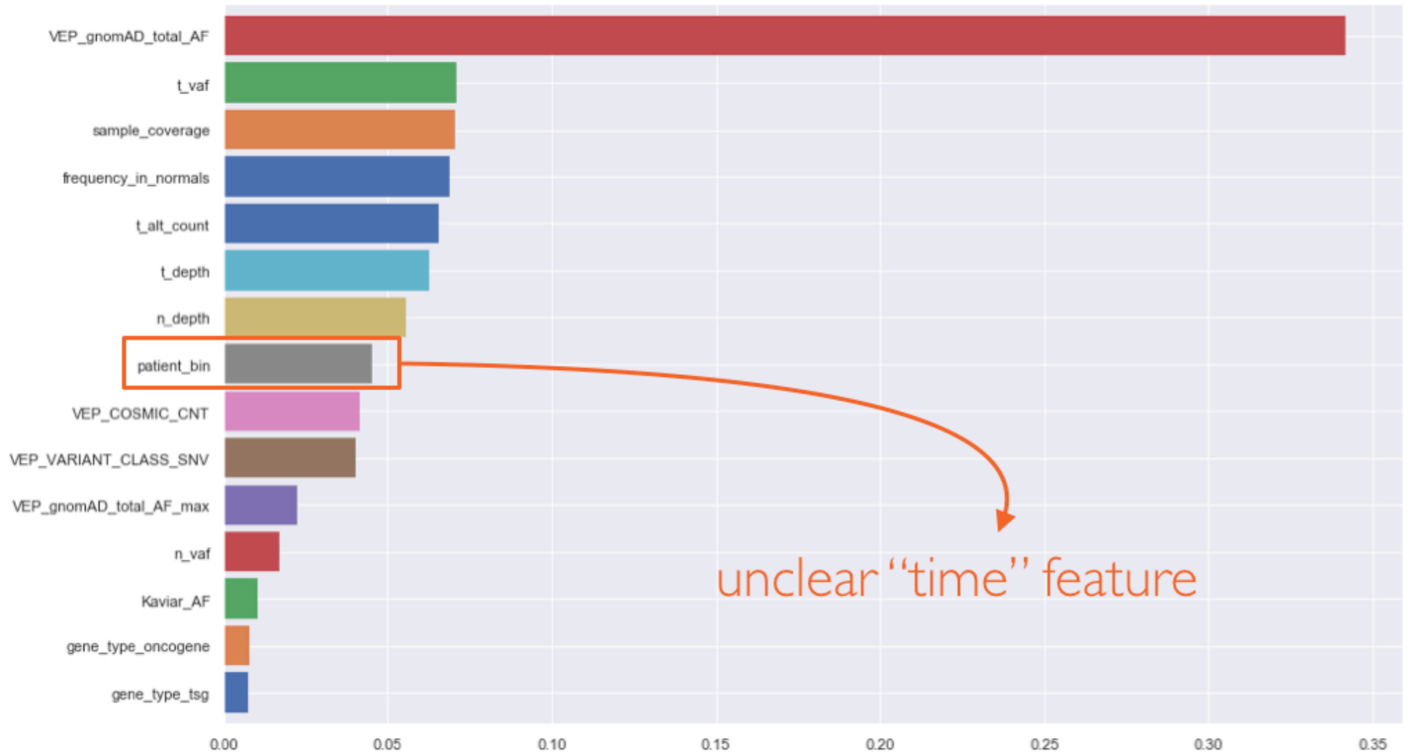


Fig. 1 | Deep learning and random forest models achieved very high manual review classification performance during tenfold cross-validation.

Uniform IMPACT processing?



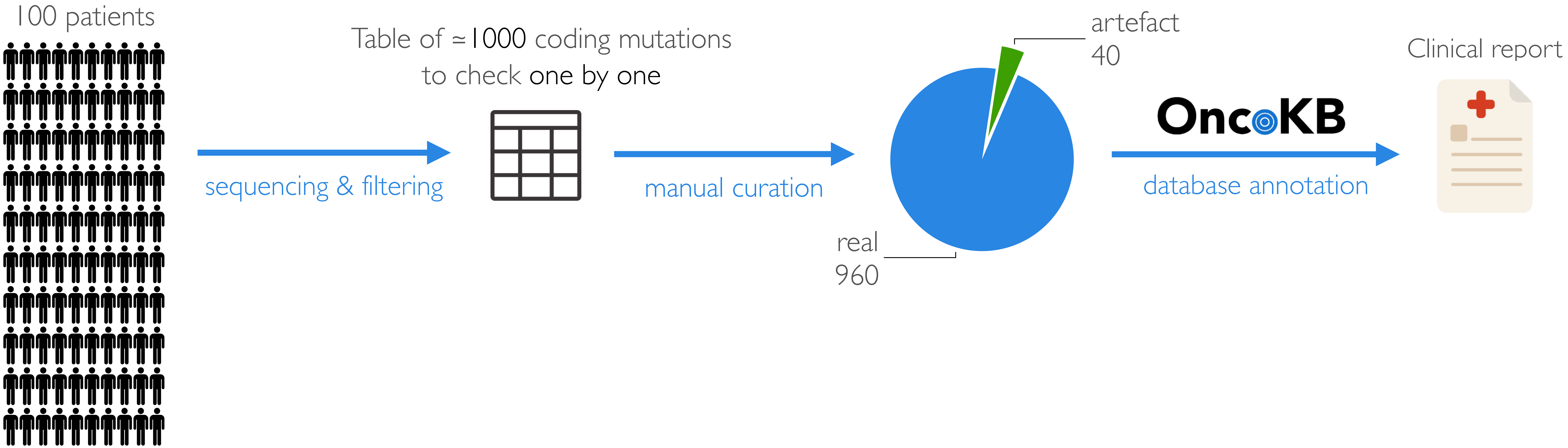




→ Uniform variant callers across time & panels

→ Enable detailed feature extraction | Technical & Flags

Final goal: a two-steps web-based classifier



100 patients

sequencing & filtering

Table of ≃1000 coding mutations
to check one by one

manual curation

artefact
40

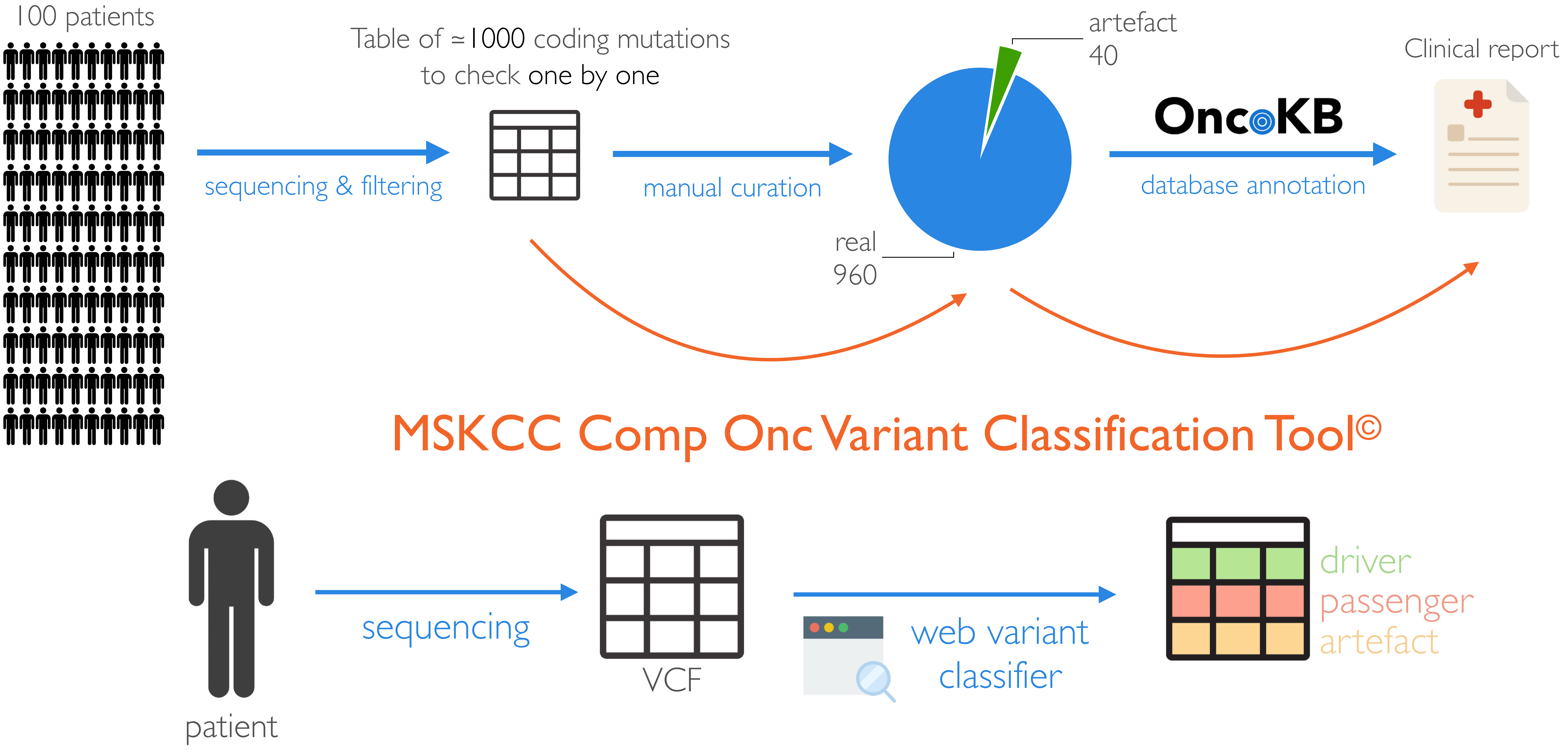
real
960

OncoKB

database annotation

Clinical report

Final goal: a two-steps web-based classifier



MSKCC Comp Onc Variant Classification Tool©