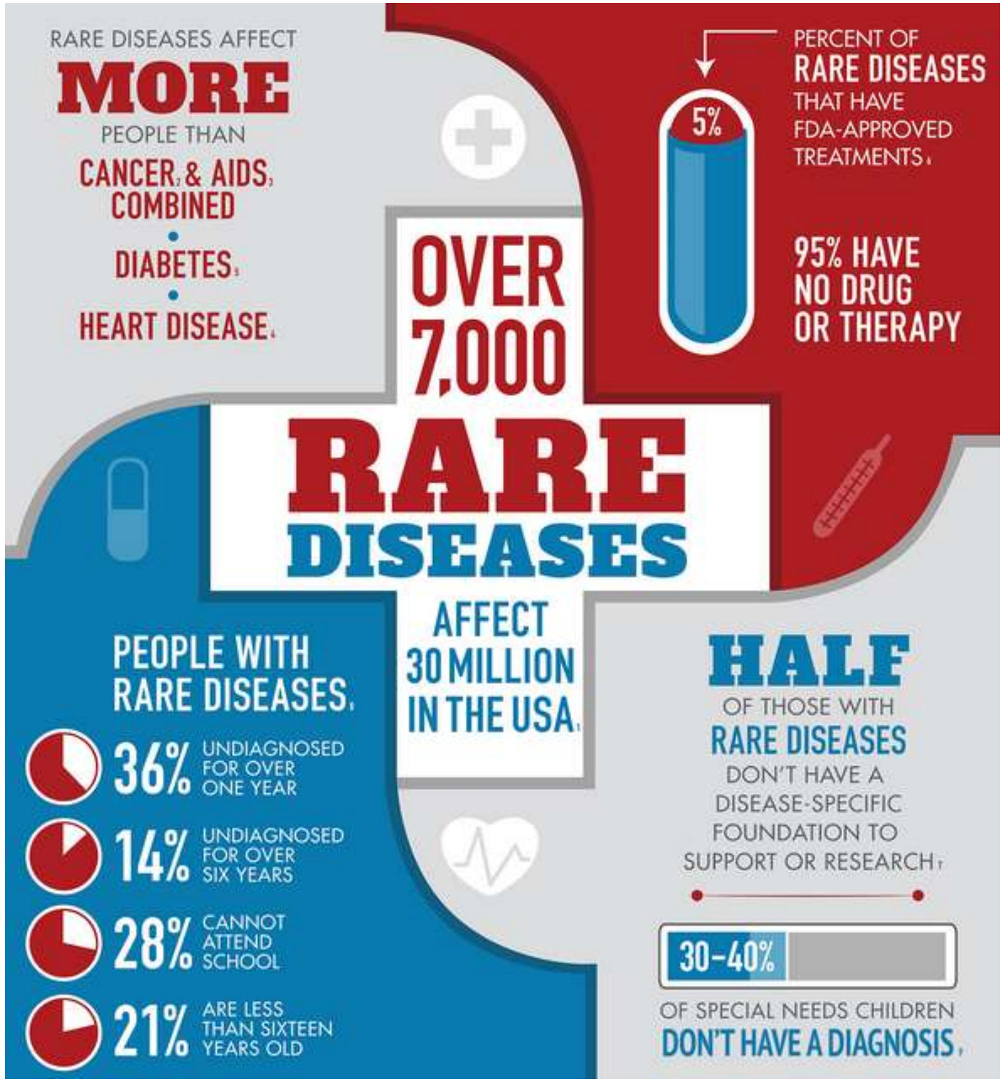
Precision-Recall (PR) vs Receiver Operating Characteristics (ROC) curves.

Which one to use if data is imbalanced?

dr. ir. Sultan K. Imangaliyev

Leiden, 22/01/26

Imbalanced data (1)



RARE DISEASES AFFECT **MORE** PEOPLE THAN CANCER & AIDS COMBINED • DIABETES • HEART DISEASE

PERCENT OF RARE DISEASES THAT HAVE FDA-APPROVED TREATMENTS

5%

95% HAVE NO DRUG OR THERAPY

OVER 7,000 **RARE DISEASES** AFFECT 30 MILLION IN THE USA

PEOPLE WITH RARE DISEASES

36% UNDIAGNOSED FOR OVER ONE YEAR

14% UNDIAGNOSED FOR OVER SIX YEARS

28% CANNOT ATTEND SCHOOL

21% ARE LESS THAN SIXTEEN YEARS OLD

HALF OF THOSE WITH RARE DISEASES DON'T HAVE A DISEASE-SPECIFIC FOUNDATION TO SUPPORT OR RESEARCH

30–40% OF SPECIAL NEEDS CHILDREN DON'T HAVE A DIAGNOSIS

http://blog.crowdmed.com/epidemic-no-one-talking-rare-disease-infographic/

Imbalanced data (2)

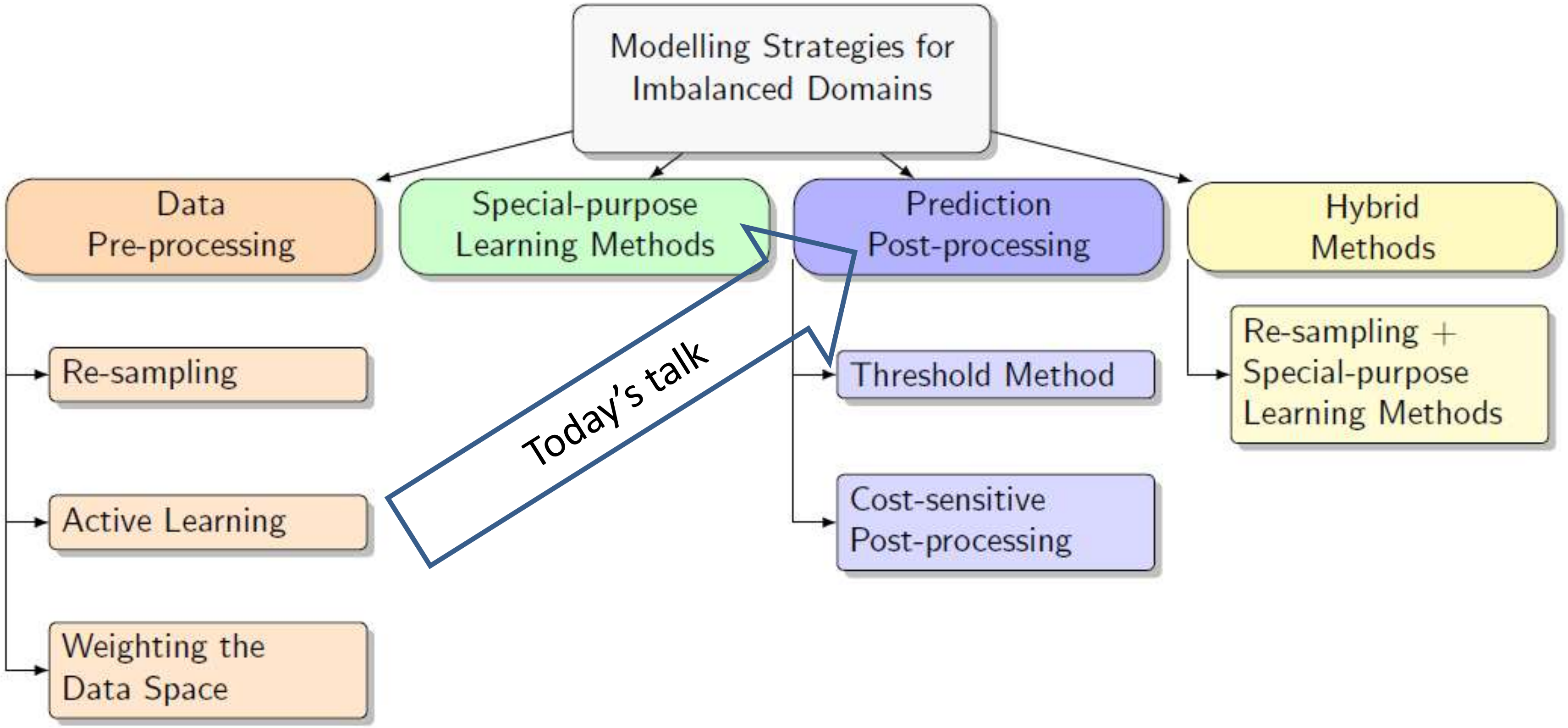
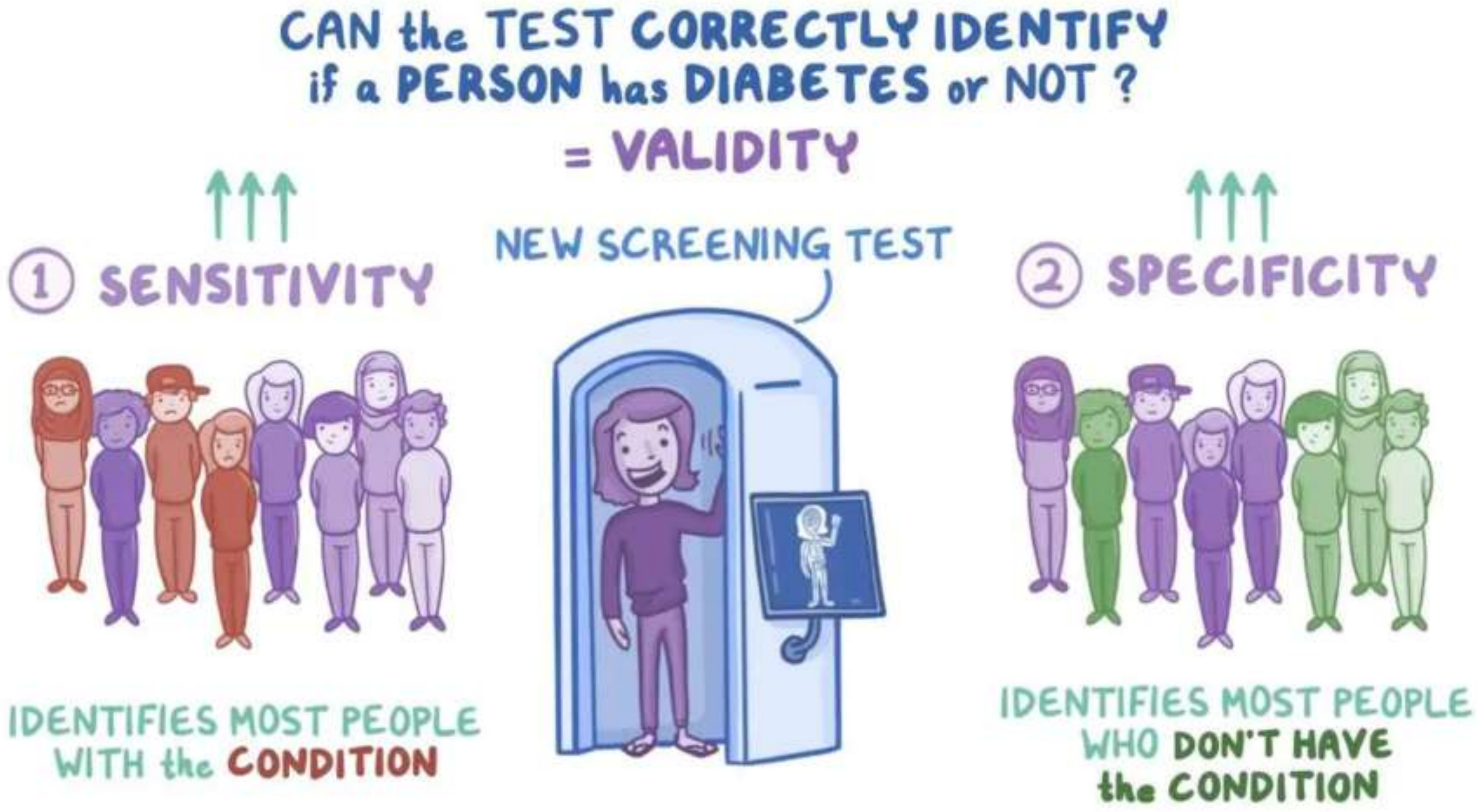

Modelling Strategies for Imbalanced Domains
Data Pre-processing
Special-purpose Learning Methods
Prediction Post-processing
Hybrid Methods
Re-sampling
Active Learning
Weighting the Data Space
Threshold Method
Cost-sensitive Post-processing
Re-sampling + Special-purpose Learning Methods
Today's talk

Figure 7: Main modelling strategies for imbalanced domains.

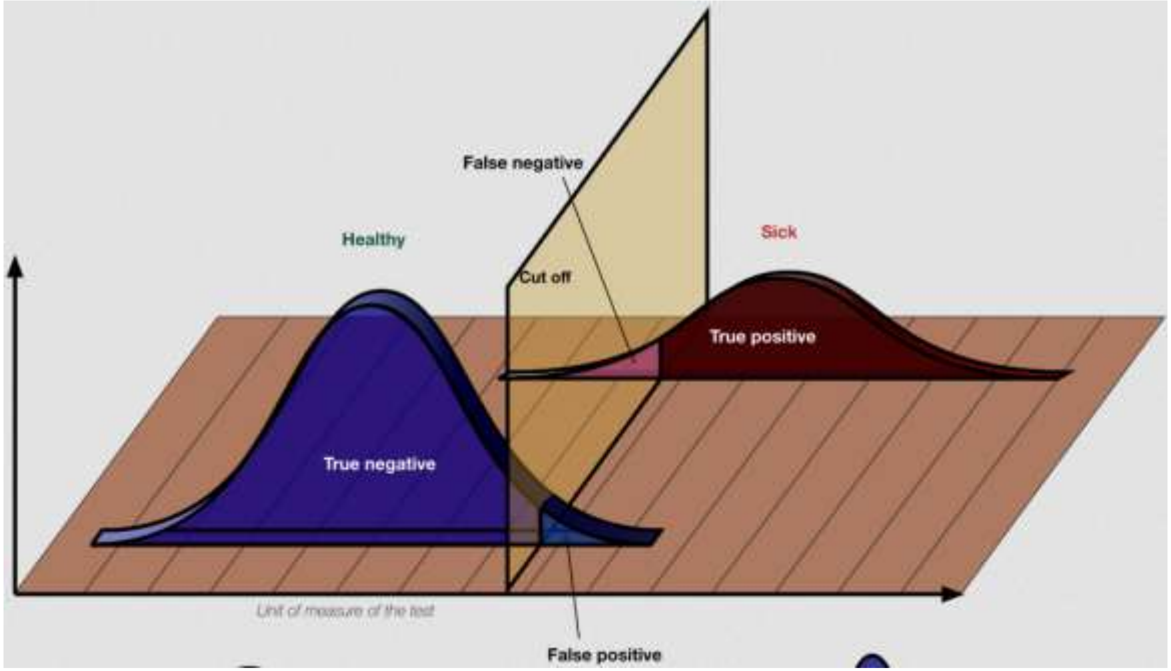
Branco, Paula, Luís Torgo, and Rita P. Ribeiro. "A survey of predictive modeling on imbalanced domains." *ACM Computing Surveys (CSUR)* 49.2 (2016): 1-50.

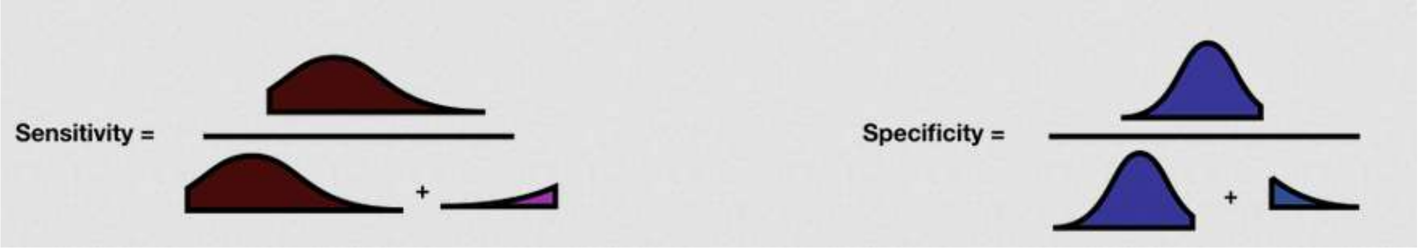
Sensitivity vs Specificity (1)


CAN the TEST CORRECTLY IDENTIFY
if a PERSON has DIABETES or NOT ?
= VALIDITY
↑↑↑
① SENSITIVITY
NEW SCREENING TEST
↑↑↑
② SPECIFICITY
IDENTIFIES MOST PEOPLE WITH the CONDITION
IDENTIFIES MOST PEOPLE WHO DON'T HAVE the CONDITION

https://www.osmosis.org/learn/Sensitivity_and_specificity?from=

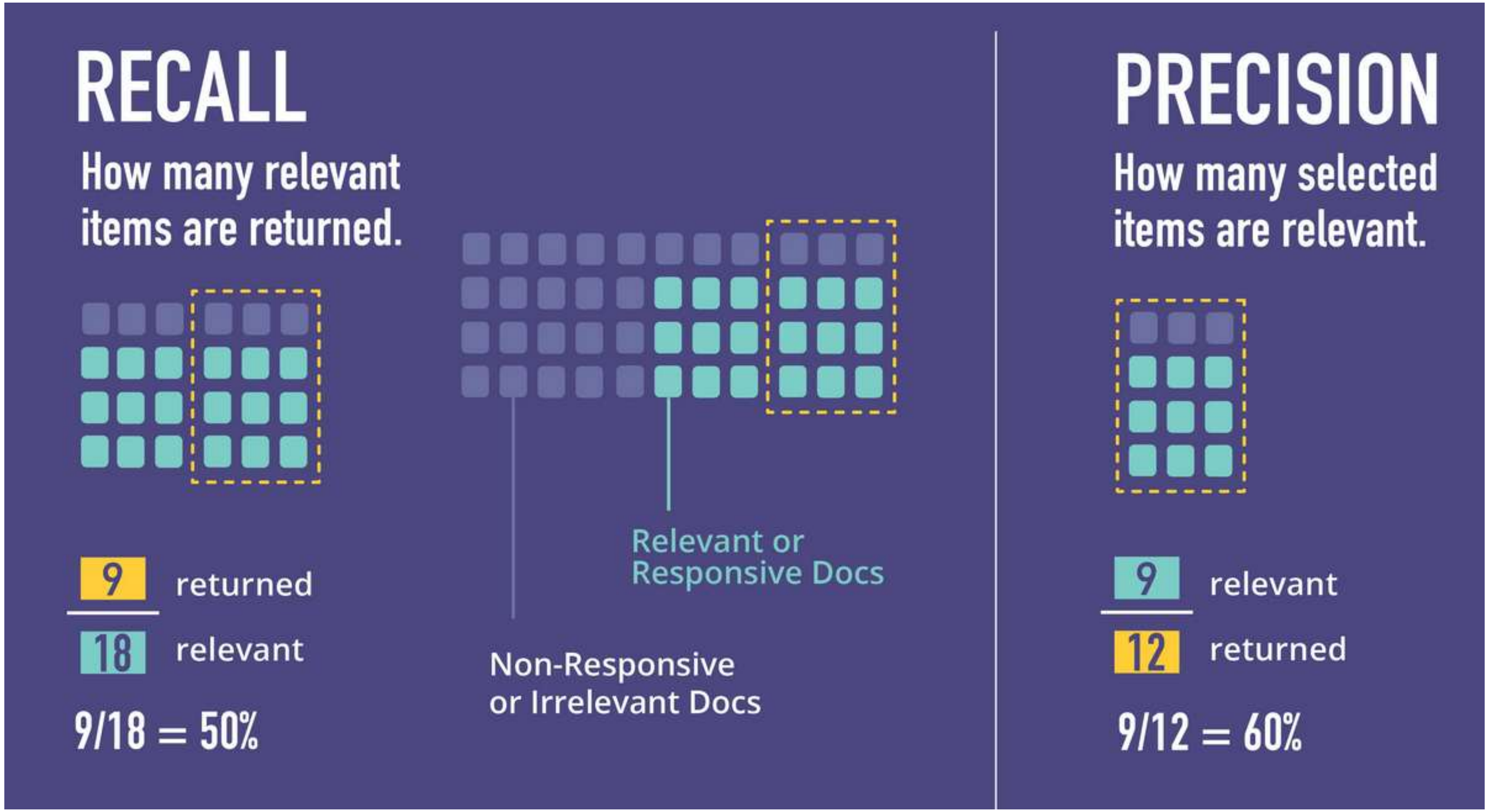
Sensitivity vs Specificity (2)





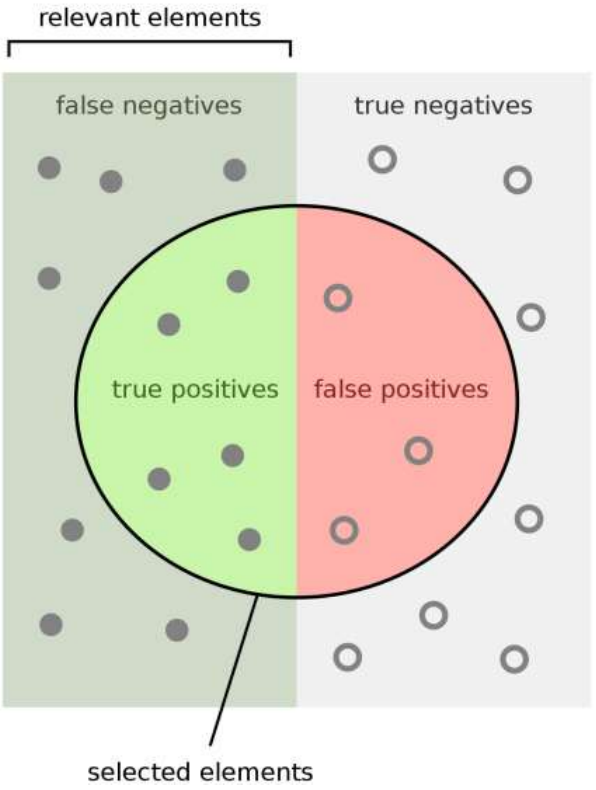
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

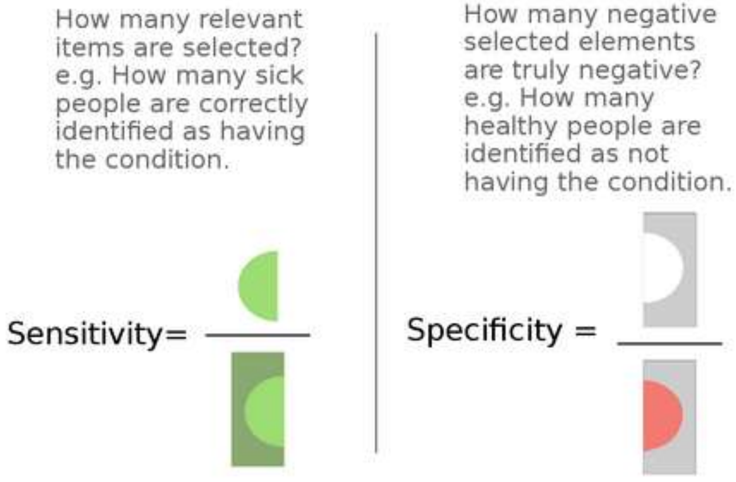
Precision vs Recall (1)

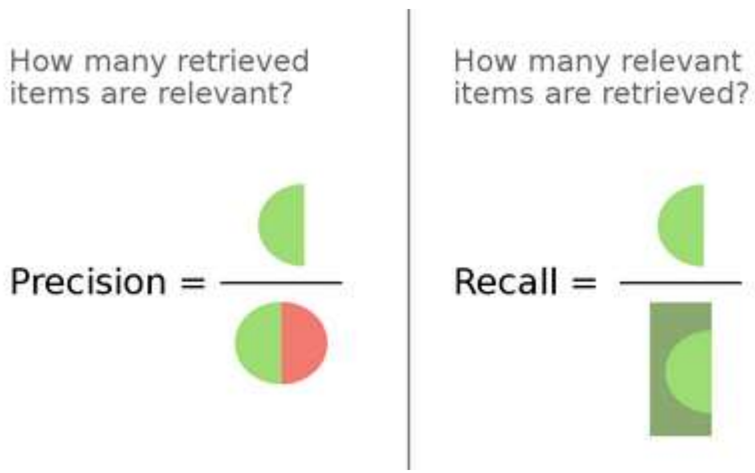

RECALL
How many relevant items are returned.
9 returned
18 relevant
9/18 = 50%
Relevant or Responsive Docs
Non-Responsive or Irrelevant Docs
PRECISION
How many selected items are relevant.
9 relevant
12 returned
9/12 = 60%

https://www.textiq.com/blog/how-to-measure-ai-recall-vs.-precision

Precision vs Recall (2)

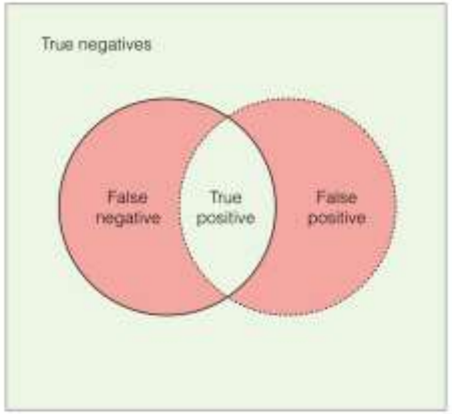



How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.
Sensitivity=
How many negative selected elements are truly negative? e.g. How many healthy people are identified as not having the condition.
Specificity =

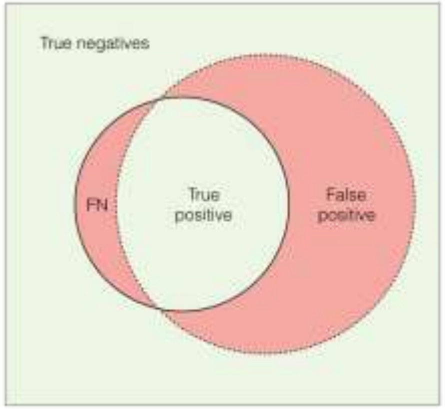

How many retrieved items are relevant?
Precision =
How many relevant items are retrieved?
Recall =

https://en.wikipedia.org/wiki/Sensitivity_and_specificity
https://en.wikipedia.org/wiki/Precision_and_recall

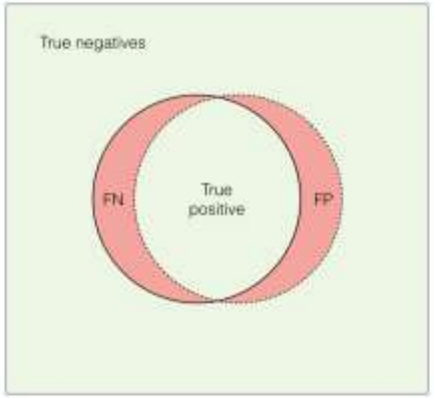
Precision vs Recall (3)



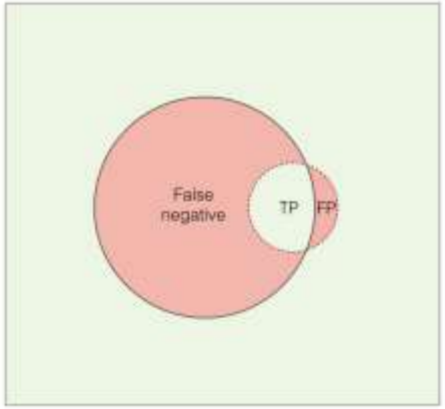
Low recall, low precision



High recall, low precision



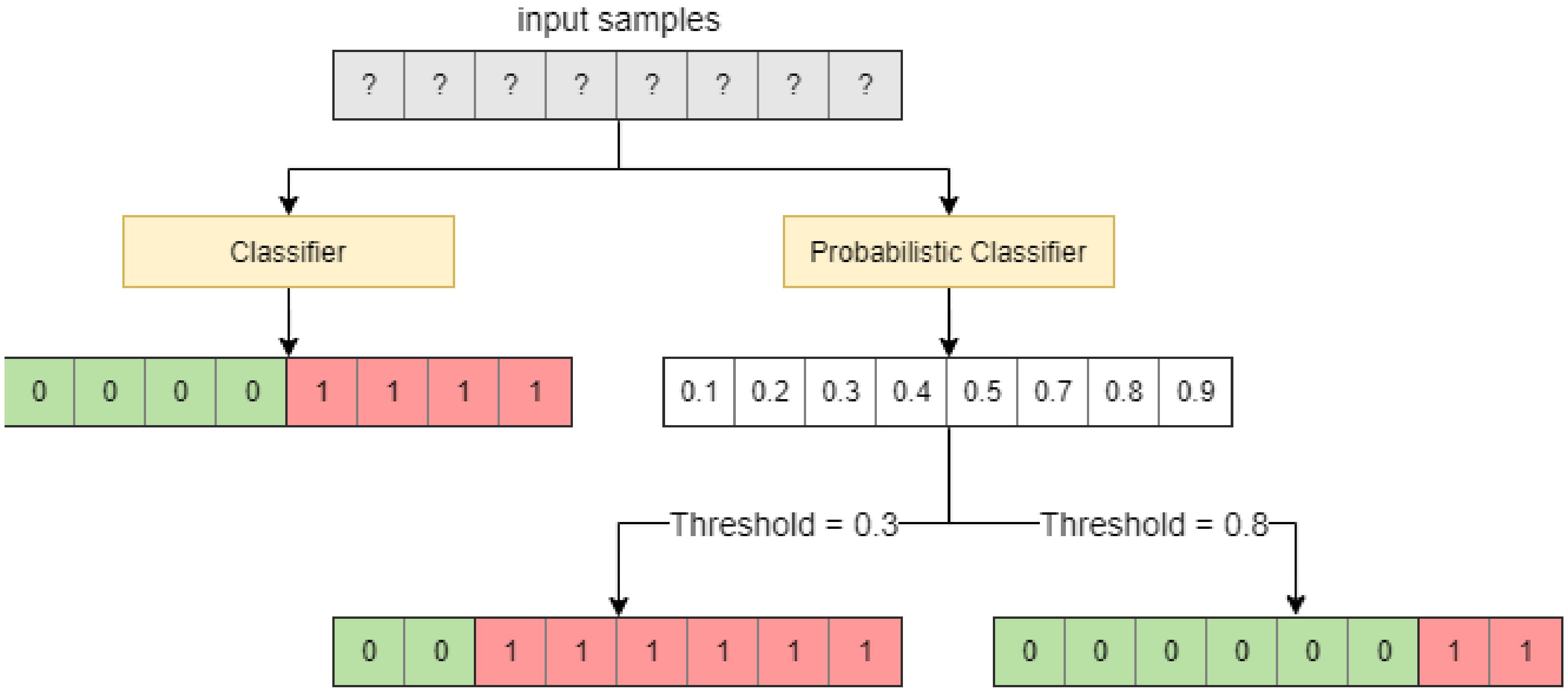
High recall, high precision



Low recall, high precision

https://medium.com/@klintcho/explaining-precision-and-recall-c770eb9c69e9

Probabilistic classifiers



input samples

?	?	?	?	?	?	?	?

Classifier

0	0	0	0	1	1	1	1

Probabilistic Classifier

0.1	0.2	0.3	0.4	0.5	0.7	0.8	0.9

Threshold = 0.3

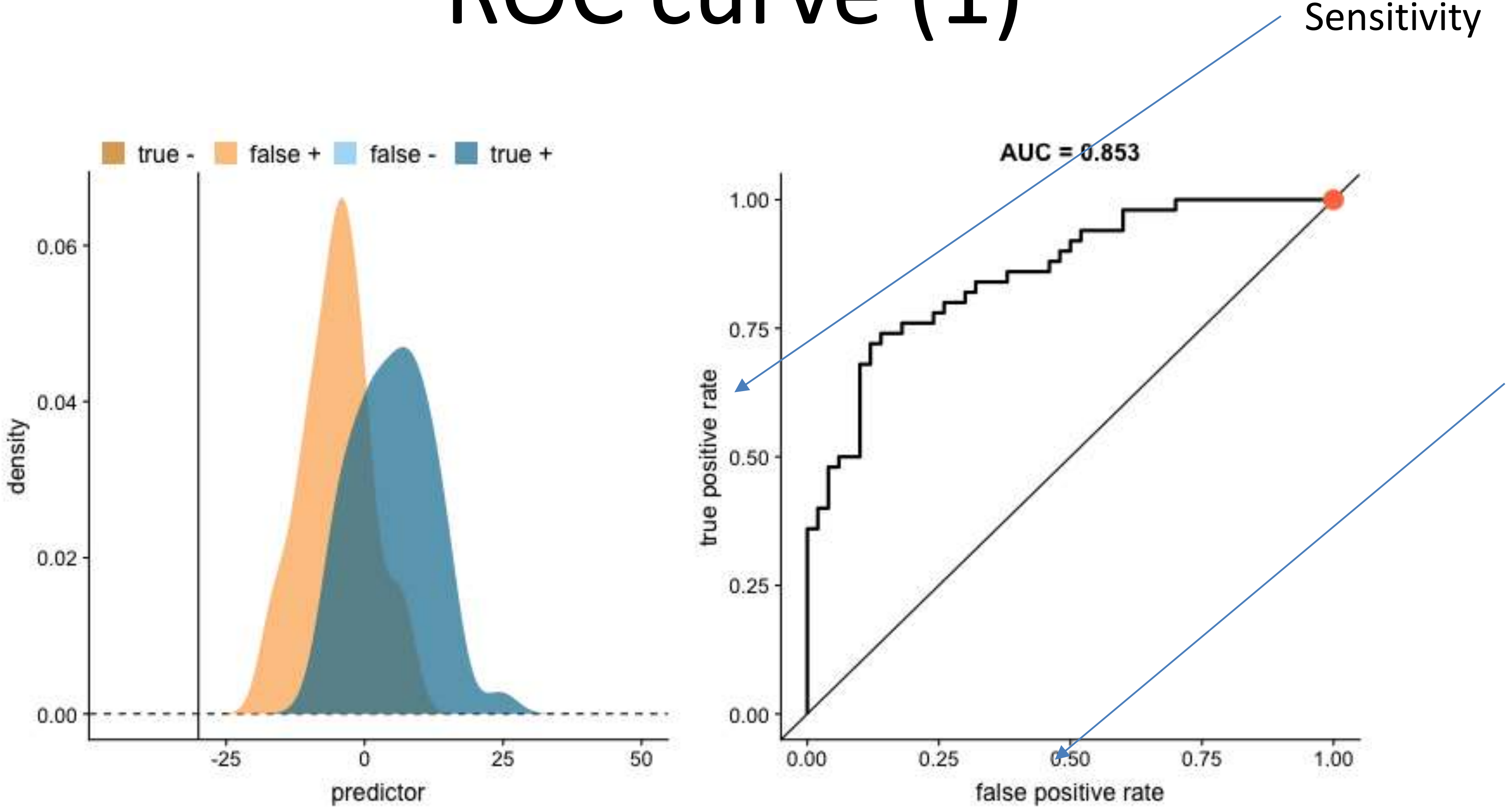
0	0	1	1	1	1	1	1

Threshold = 0.8

0	0	0	0	0	0	1	1

https://medium.com/knowledge-engineering-seminar/evaluating-probabilistic-classifier-roc-and-pr-g-curves-6647c3379d60

ROC curve (1)



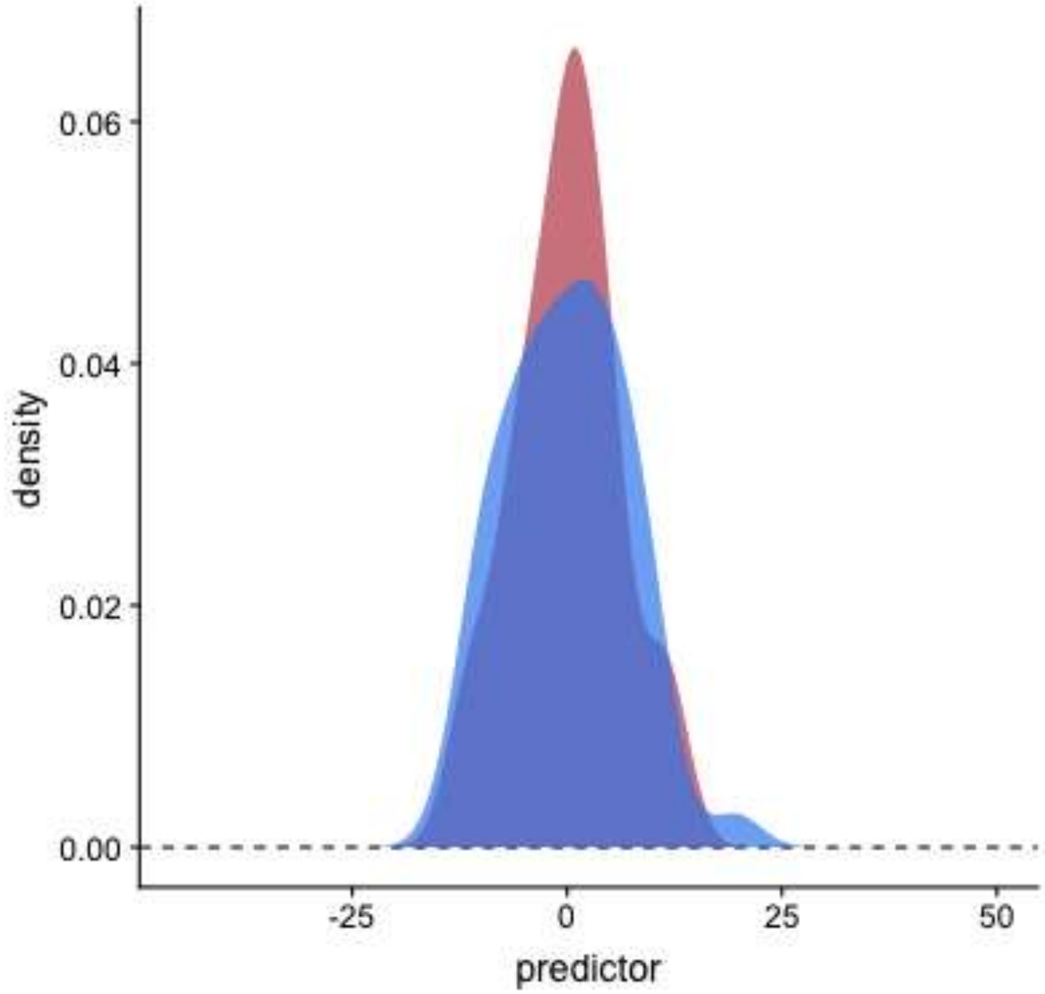
Sensitivity

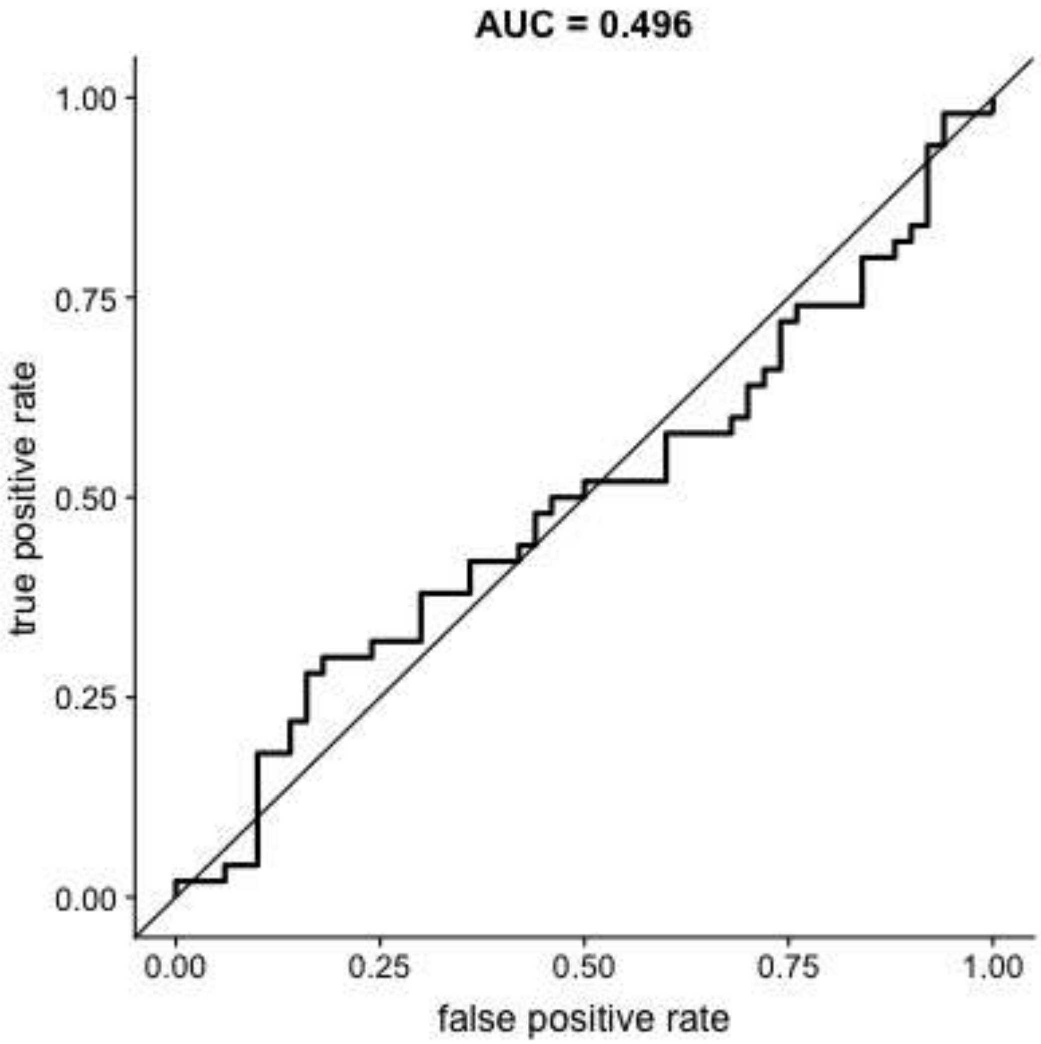
1 - Specificity

AUCROC can be interpreted as the probability that the scores given by a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

ROC curve (2)





https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

PR curve

The PR curves plots the following parameters:

Precision = TP/(TP+FP)

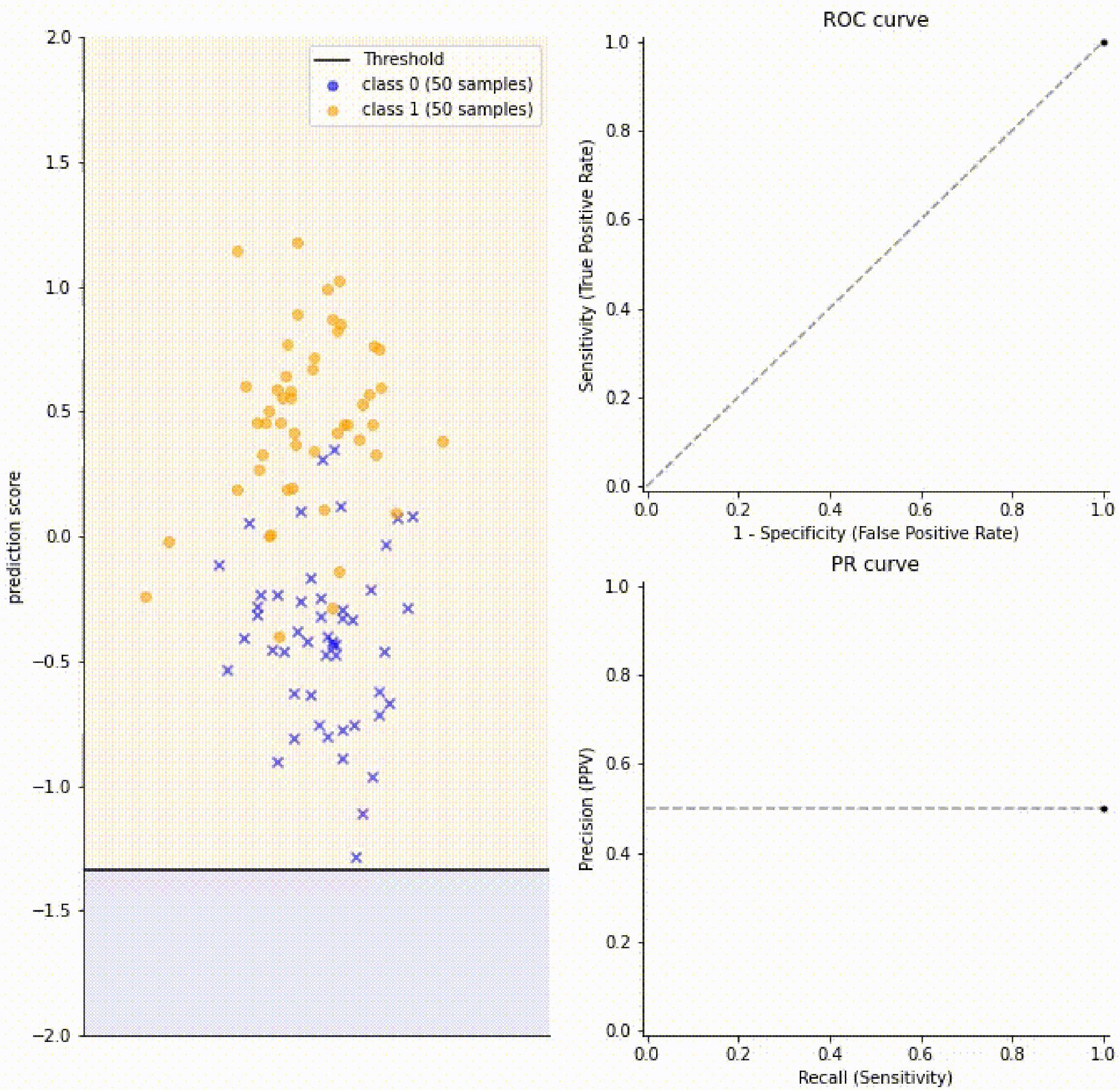
Recall = TP/(TP+FN)

Notice how True Negatives (TN) are absent from the equation?

PR curves are useful when positive examples are rare.

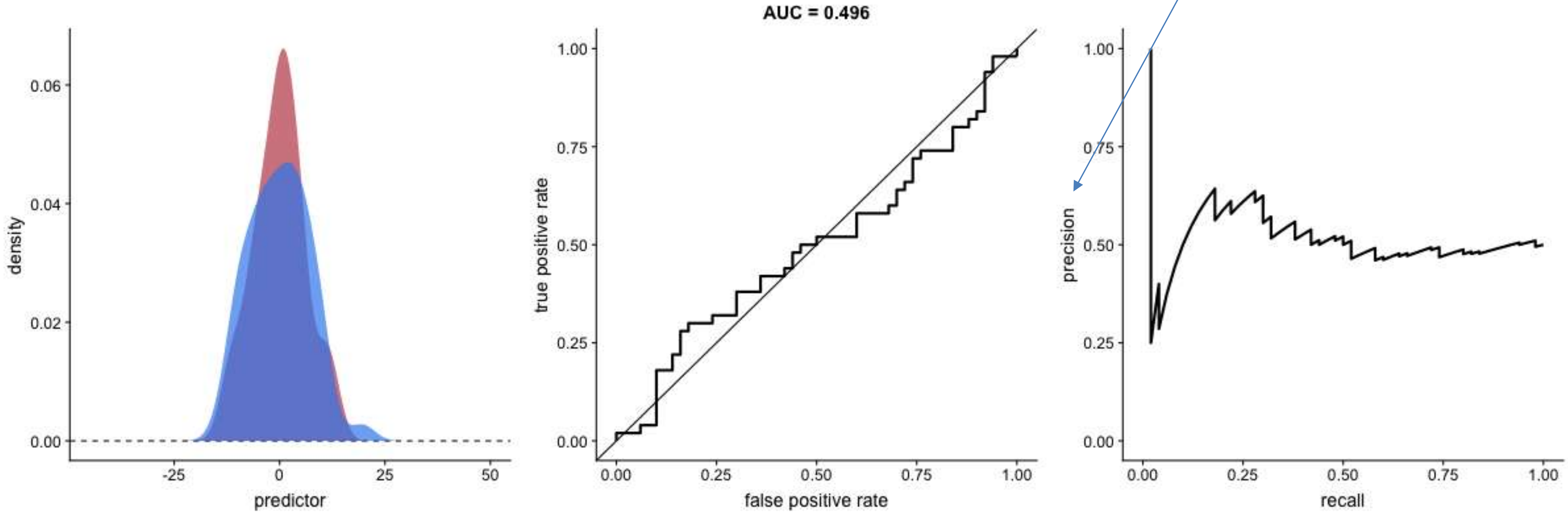
https://stats.stackexchange.com/questions/398199/why-is-a-pr-curve-considered-better-than-an-roc-curve-for-imbalanced-datasets

Comparing curves, balanced simulated data



https://towardsdatascience.com/on-roc-and-precision-recall-curves-c23e9b63820c

PR curve, balanced data

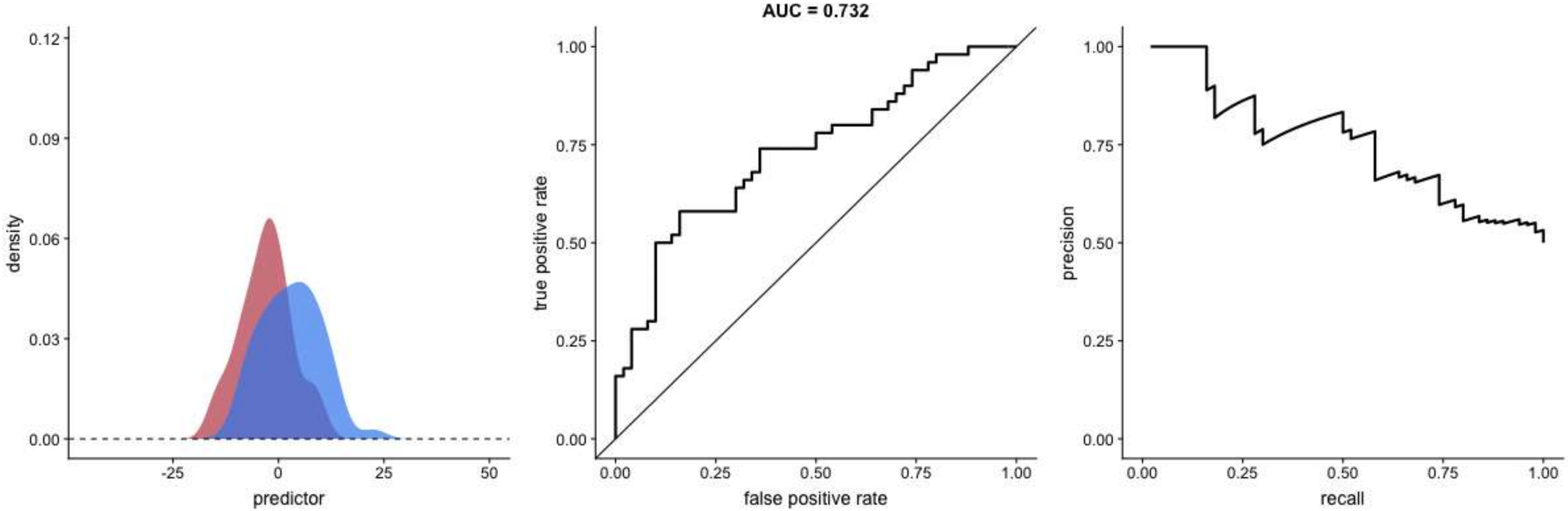


PPV = Positive Predictive Value

Sensitivity

https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

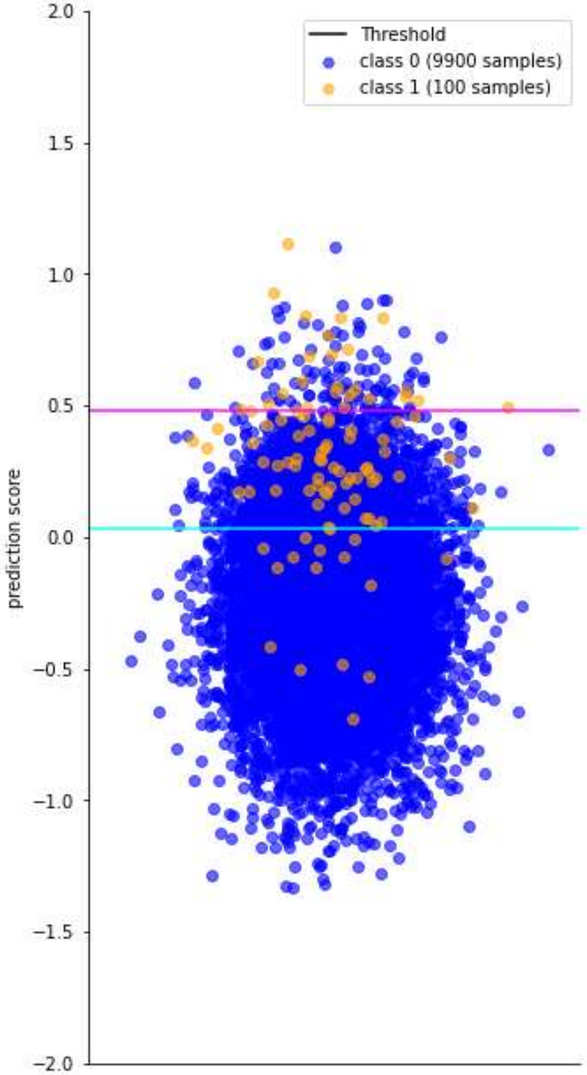
Misleading AUC values

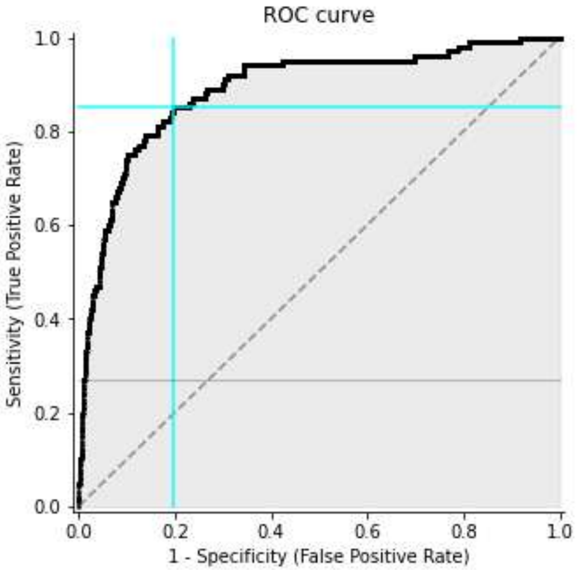


https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

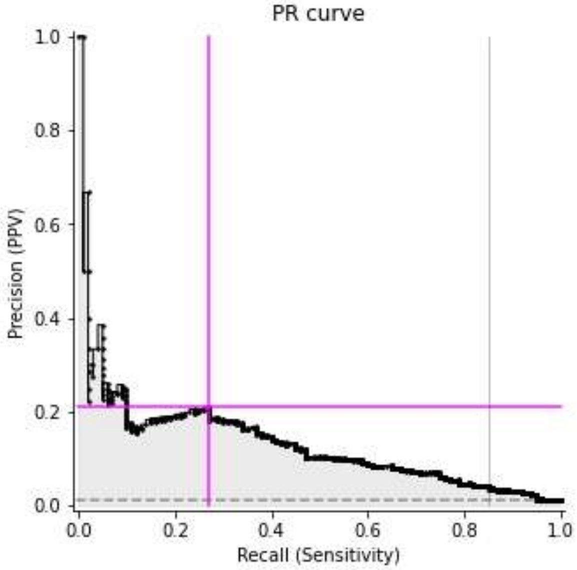
Comparing curves, imbalanced simulated data

optimal threshold for ROC (purple) and PR (green).





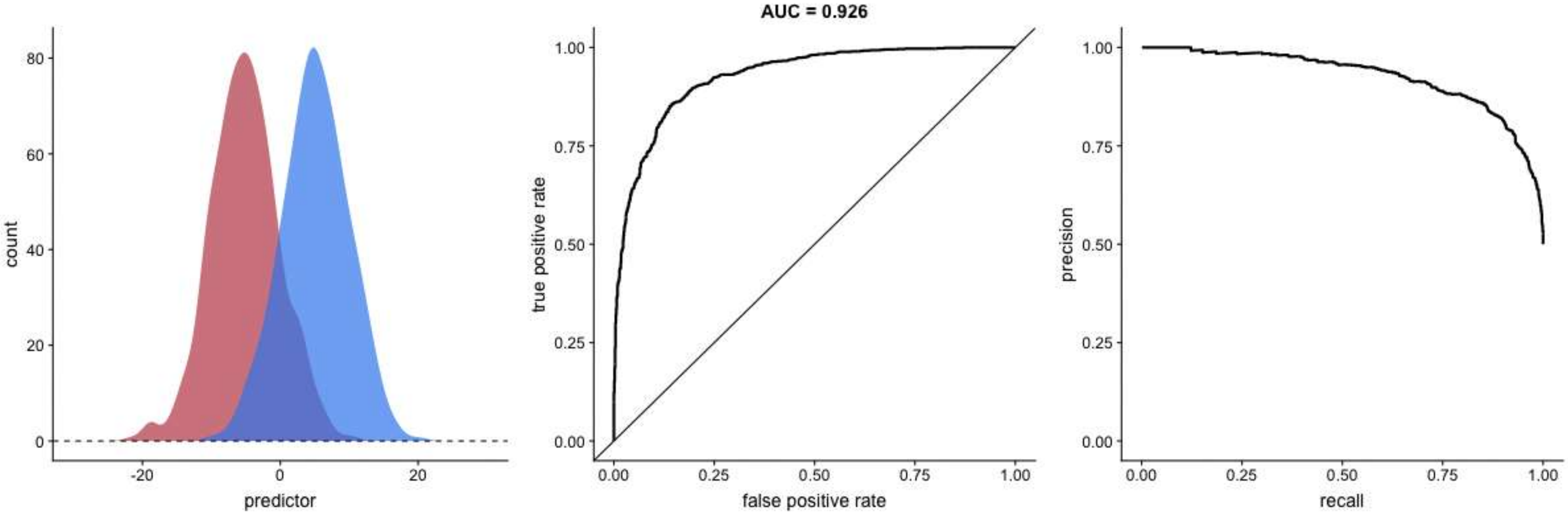
ROC curve with its optimal threshold (green) and the PR optimal threshold (grey)



PR curve with its optimal threshold (purple) and the ROC optimal threshold (grey)

https://towardsdatascience.com/on-roc-and-precision-recall-curves-c23e9b63820c

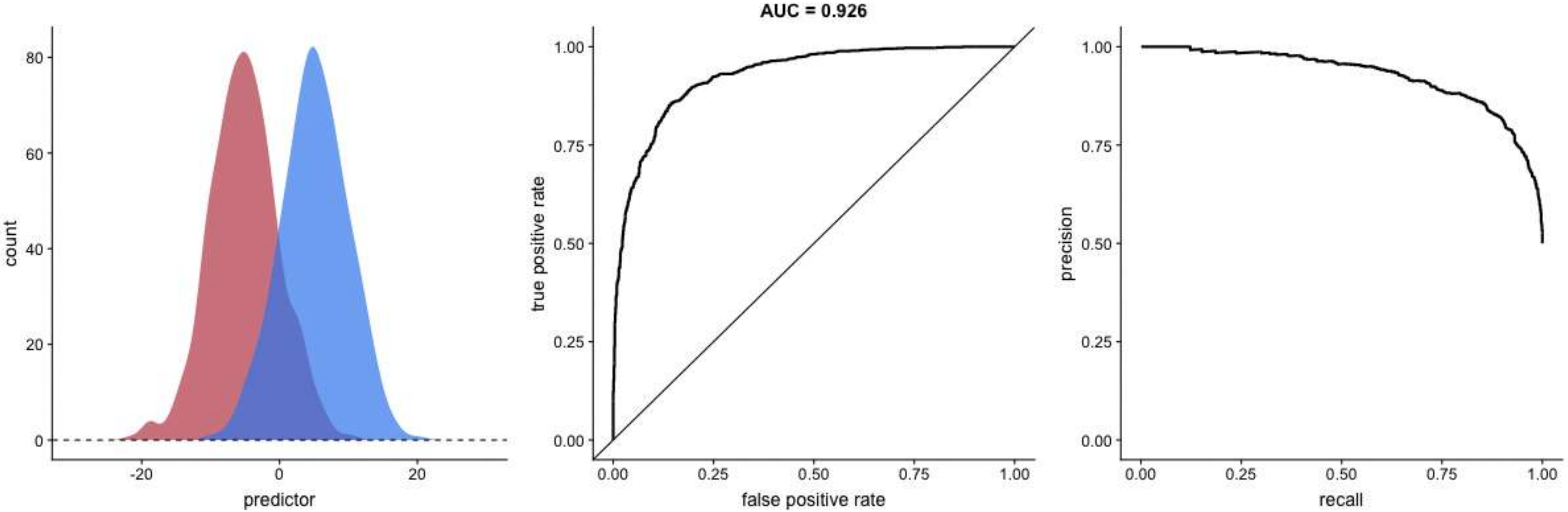
PR curve, imbalanced data (*positive examples are rare*)



… precision and recall make it possible to assess the performance of a classifier on the minority class

https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

PR curve, imbalanced data (*negative examples are rare*)



https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

Practical tip (simplified)

- When to use PR AUC or AUROC?
  - When two classes are equally important
    - AUROC if the goal is to perform equally well on both classes.
      - Image classification between cats & dogs; the performance on cats is equally important on dogs.
  - When minority class is more important
    - PR AUC if the focus of the model is to identify correctly as many positive samples as possible.
      - Spam detectors; regular emails are not of interest at all — they overshadow the number of positives.

https://sinyi-chou.github.io/classification-pr-curve/
https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

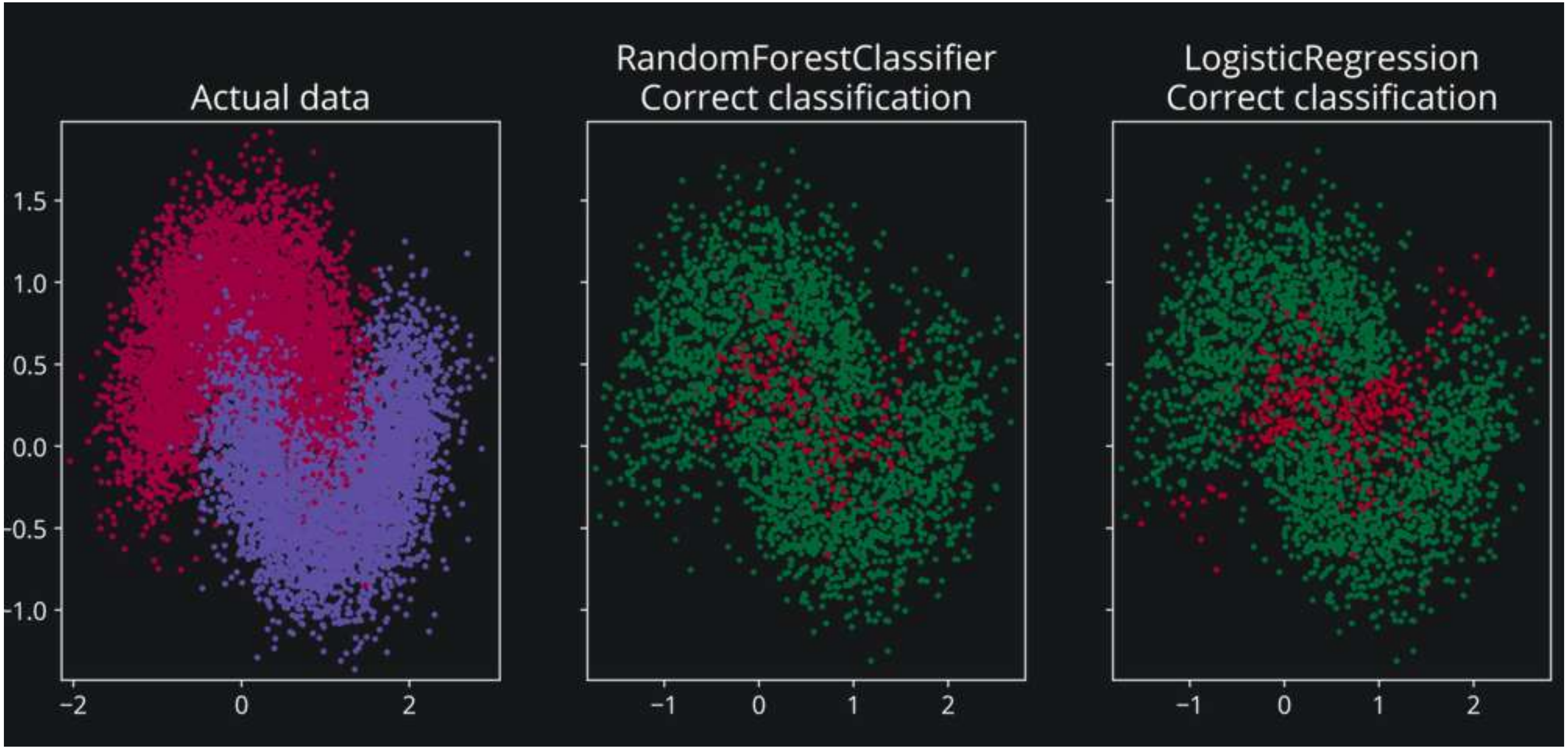
Words of caution

- ROC curves have common advantages
  - Universal baselines
    - A random classifier is represented by a major diagonal
  - Linear interpolation
    - Any two points on an ROC curve can be linearly interpolated
  - Interpretable area:
    - AUROC can be used to calculate an expected accuracy of the model:

- PR curves have pitfalls
  - No universal baseline
    - Performance of a random classifier depends on the prevalence in the data set
  - No linear interpolation
  - Uninterpretable area
    - Area under curve, commonly calculated by a trapezodial rule that performs linear interpolation, might be an overly optimistic measure

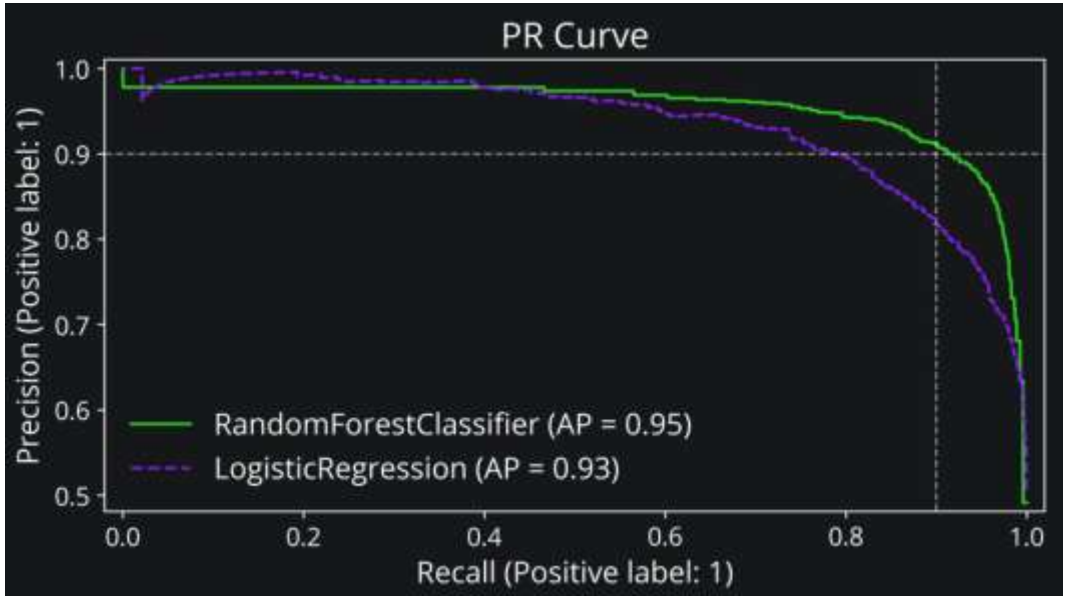
https://medium.com/knowledge-engineering-seminar/evaluating-probabilistic-classifier-roc-and-pr-g-curves-6647c3379d60

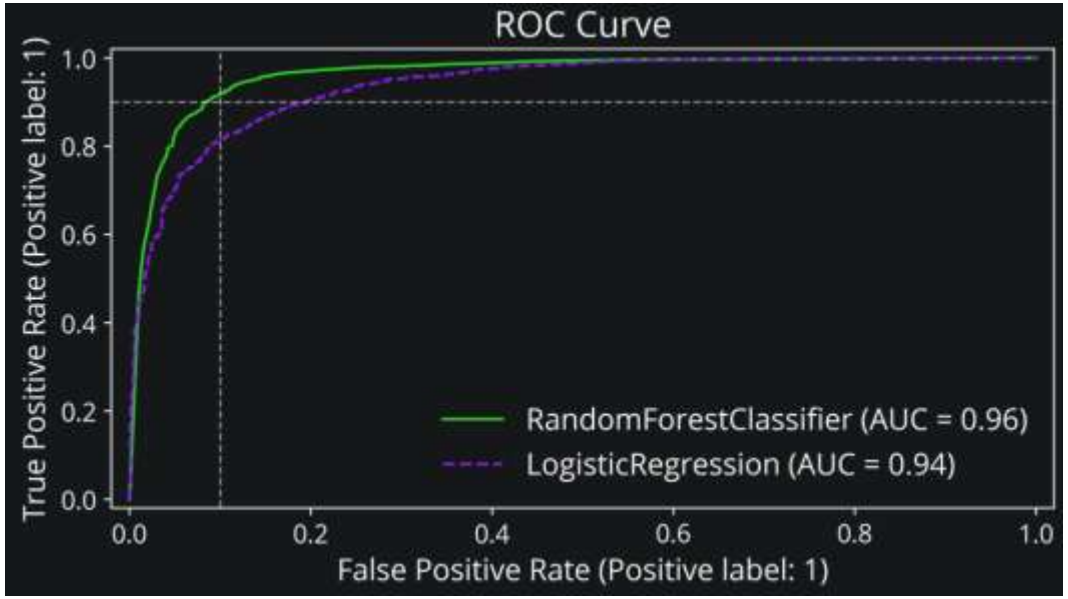
Comparing classifiers, balanced simulated data (1)



https://cosmiccoding.com.au/tutorials/pr_vs_roc_curves

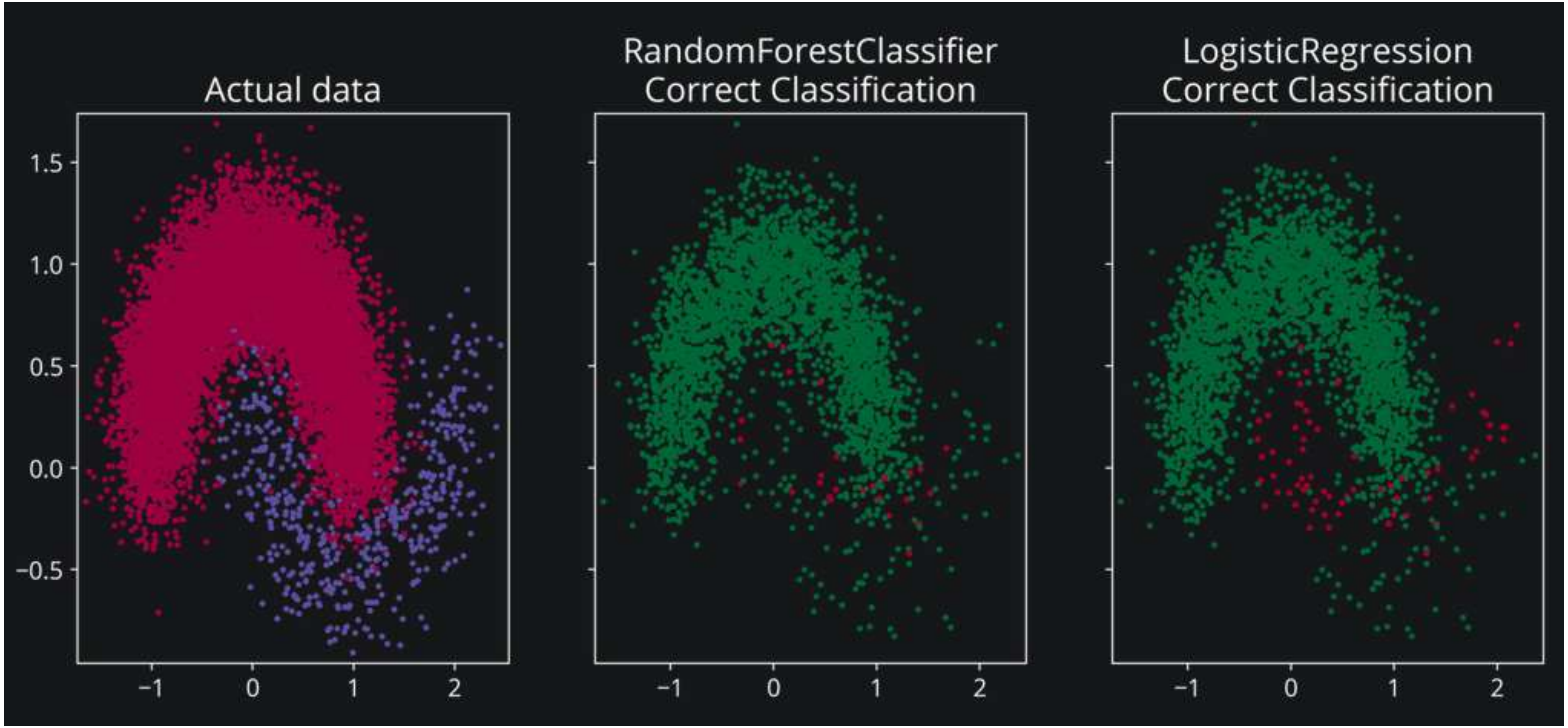
Comparing classifiers, balanced simulated data (2)





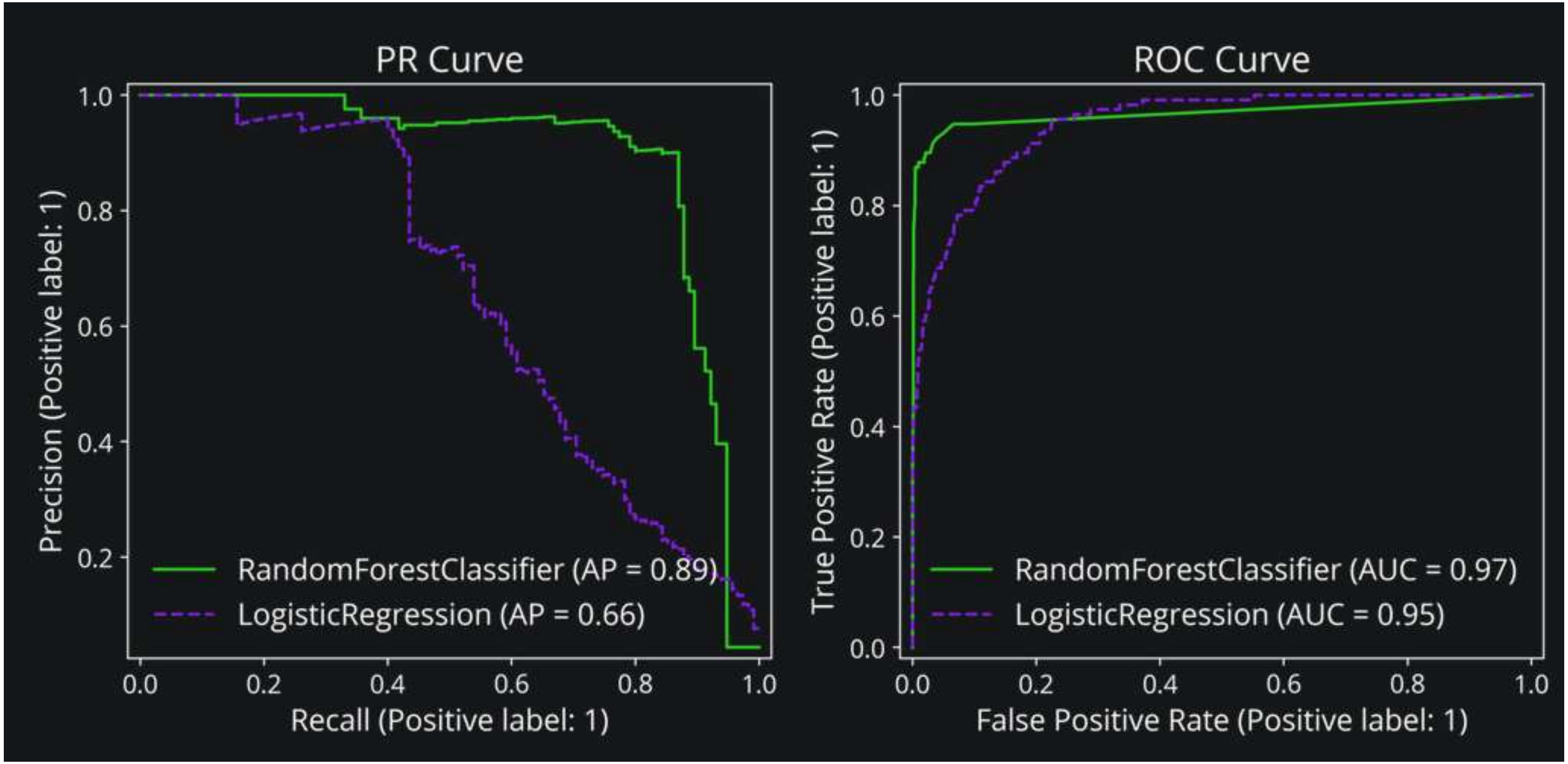
https://cosmiccoding.com.au/tutorials/pr_vs_roc_curves

Comparing classifiers, imbalanced simulated data (1)



https://cosmiccoding.com.au/tutorials/pr_vs_roc_curves

Comparing classifiers, imbalanced simulated data (2)



https://cosmiccoding.com.au/tutorials/pr_vs_roc_curves

Conclusions

- There is no universal way of treating imbalanced data
  - Choices are application-specific

- In case of balanced data AUROC and AP are comparable

- For imbalanced data
  - Both classes are important – use AUROC
  - Minority positive class is more important – use PR curve

- Extremely high imbalance?
  - Consider switching to anomaly detection methods

Q&A