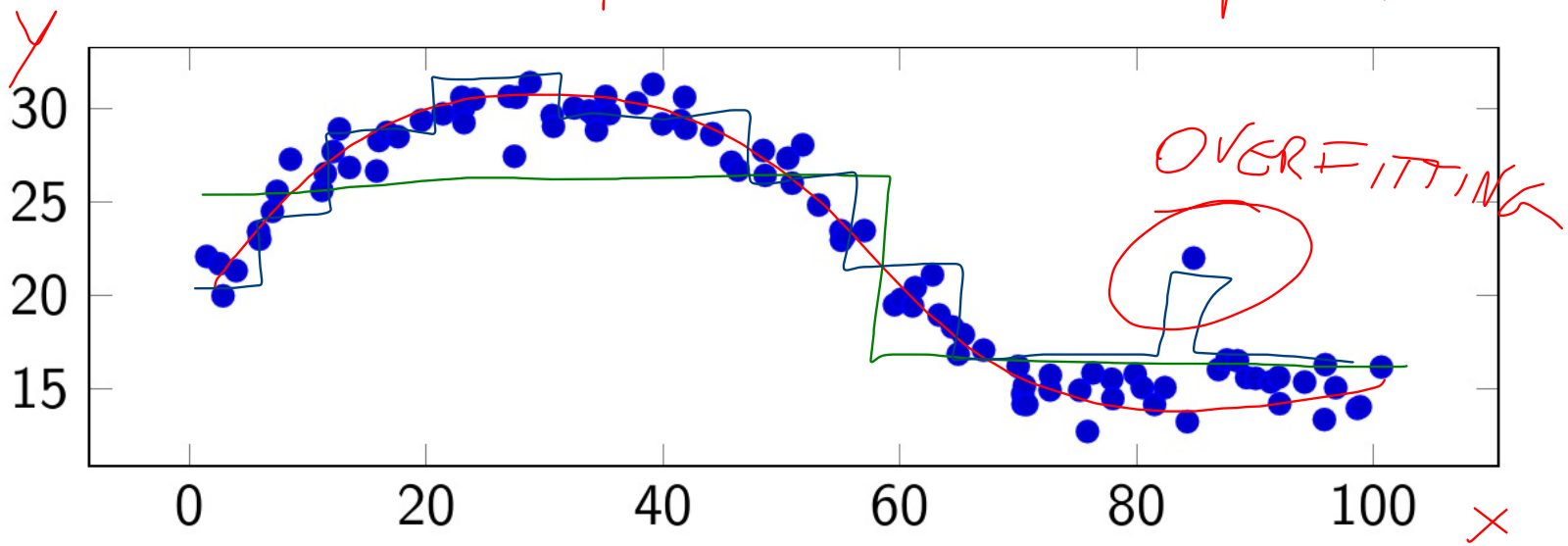
Subsection 2

Trees aggregation

Weakness of the tree



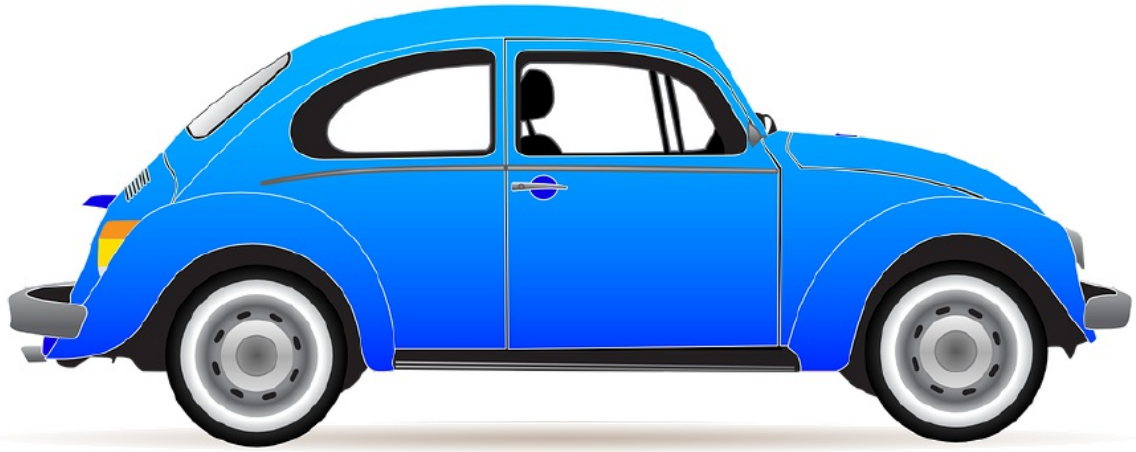
Small tree:

- low complexity
- will hardly fit the "curve" part
- *high bias, low variance*

Big tree:

- high complexity
- may overfit the noise on the right part
- *low bias, high variance*

The trees view



Small tree:

- "a car is something that moves"

Big tree:

- "a car is a made-in-Germany blue object with 4 wheels, 2 doors, chromed fenders, curved rear enclosing engine"

Big tree view

A big tree:

- has a detailed view of the learning data (high complexity)
- "trusts too much" the learning data (high variance)

What if we "combine" different big tree views and ignore details on which they disagree?

Wisdom of the crowds

What if we "combine" different big tree views and ignore details on which they disagree?

- many views
- independent views
- aggregation of views

$\approx$ *the wisdom of the crowds*: a collective opinion may be better than a single expert's opinion

Wisdom of the trees

- many views
- independent views
- aggregation of views

Wisdom of the trees

- many views
  - just use many trees
- independent views

- aggregation of views

Wisdom of the trees

- many views
  - just use many trees
- independent views

- aggregation of views
  - just average prediction (regression) or take most common prediction (classification)

Wisdom of the trees

- many views
  - just use many trees
- independent views
  - ??? learning is deterministic: same data $\Rightarrow$ same tree $\Rightarrow$ same view
- aggregation of views
  - just average prediction (regression) or take most common prediction (classification)

Independent views

Independent views $\equiv$ different points of view $\equiv$ *different* learning data

But we have only *one* learning data!

Independent views: idea! (**Bootstrap**)

Like in cross-fold, consider only a part of the data, but:

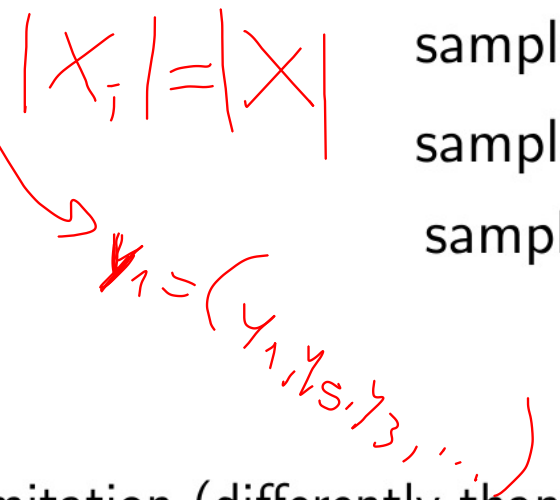
- instead of a subset
- a sample with repetitions

Independent views: idea! (**Bootstrap**)

Like in cross-fold, consider only a part of the data, but:

- instead of a subset
- a sample with repetitions

$$\mathbf{X} = (x_1^T\, x_2^T\, x_3^T\, x_4^T\, x_5^T) \qquad \text{original learning data}$$
$$\mathbf{X_1} = (x_1^T\, x_5^T\, x_3^T\, x_2^T\, x_5^T) \qquad \text{sample 1}$$
$$\mathbf{X_2} = (x_4^T\, x_2^T\, x_3^T\, x_1^T\, x_1^T) \qquad \text{sample 2}$$
$$\mathbf{X_i} = \ldots \qquad \text{sample } i$$



- (**y** omitted for brevity)
- learning data size is not a limitation (differently than with subset)

Tree **bagging**

When learning:

1. Repeat B times
   - 1.1 take a sample of the learning data
   - 1.2 learn a tree (unpruned)

WITH REPETITION

→ t_{MCE}

When predicting:

TREES → INDEPENDENT

1. Repeat B times
   - 1.1 get a prediction from ith learned tree
2. predict the average (or most common) prediction

↳ REGRESSION

→ CLASSIFICATION

For classification, other aggregations can be done: majority voting (most common) is the simplest

Using independent, possibly different classifiers together: *ensemble* of classifiers

How many trees?

B is a parameter:

- when there is a parameter, there is the problem of finding a good value
- remember $k_{\min}$, depth (**Q:** impact on?)

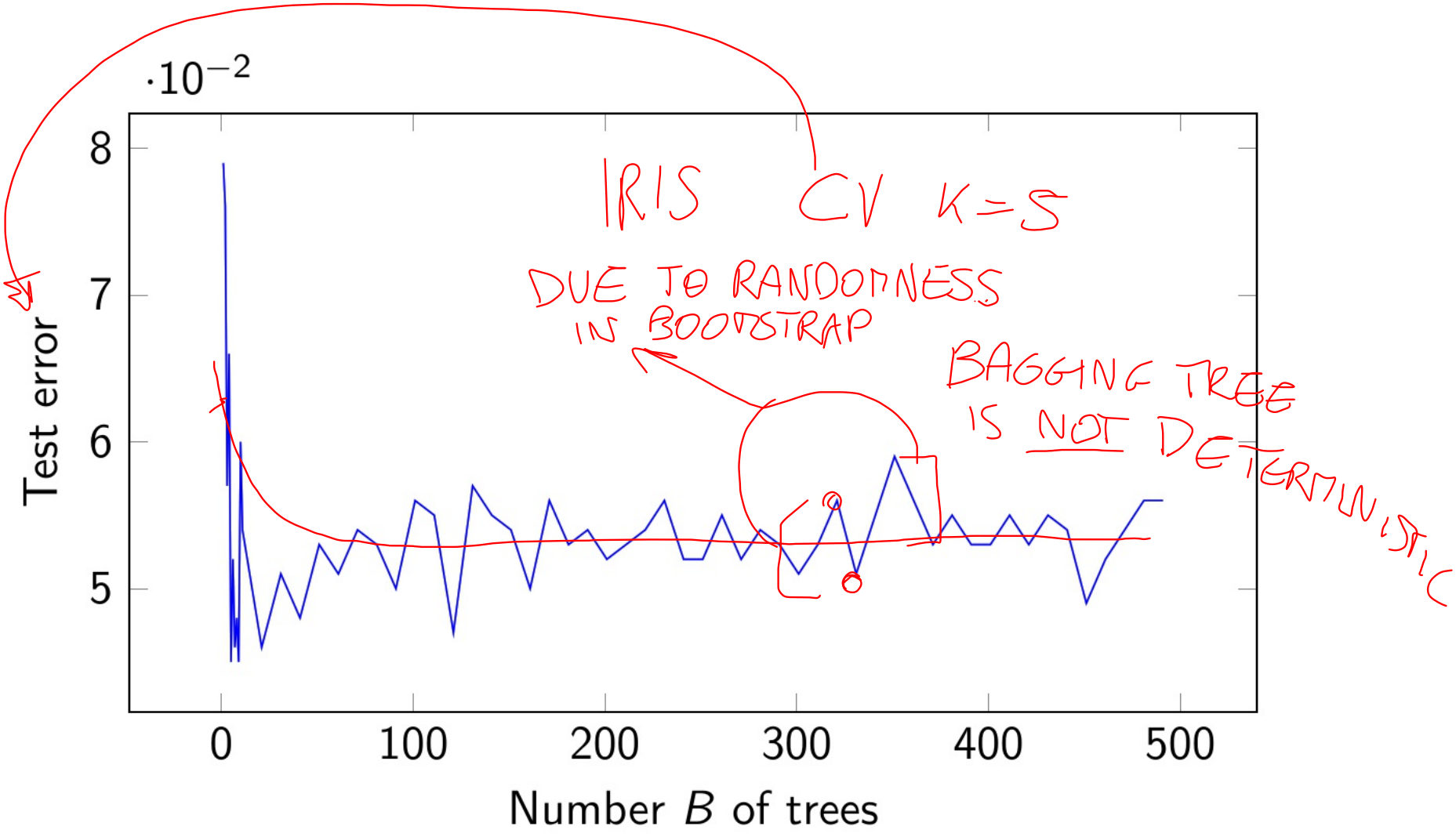
How many trees?

B is a parameter:

- when there is a parameter, there is the problem of finding a good value
- remember $k_{\min}$, depth (**Q:** impact on?)
- it has been shown (experimentally) that
  - for "large" B, bagging is better than single tree
  - increasing B does not cause overfitting
  - (for us: default B is ok! "large" $\approx$ hundreds)

Q: how better? at which cost? COMPUTATIONAL EFFORT

Bagging: impact of B



Independent view: improvement

Despite being learned on different samples, bagging trees may be correlated, hence views are not very independent

- e.g., one variable is much more important than others for predicting (*strong predictor*)

Idea: force point of view differentiation by "hiding" variables

Random forest

When learning:

1. Repeat B times
   - 1.1 take a sample of the learning data
   - 1.2 consider only m on p independent variables
   - 1.3 learn a tree (unpruned)

BOOTSTRAP

PARAMETER

When predicting:

1. Repeat B times
   - 1.1 get a prediction from ith learned tree
2. predict the average (or most common) prediction

- (observations and) variables are **random**ly chosen...
- ...to learn a **forest** of trees

Q: are missing variables a problem? NO

Random forest: parameter m

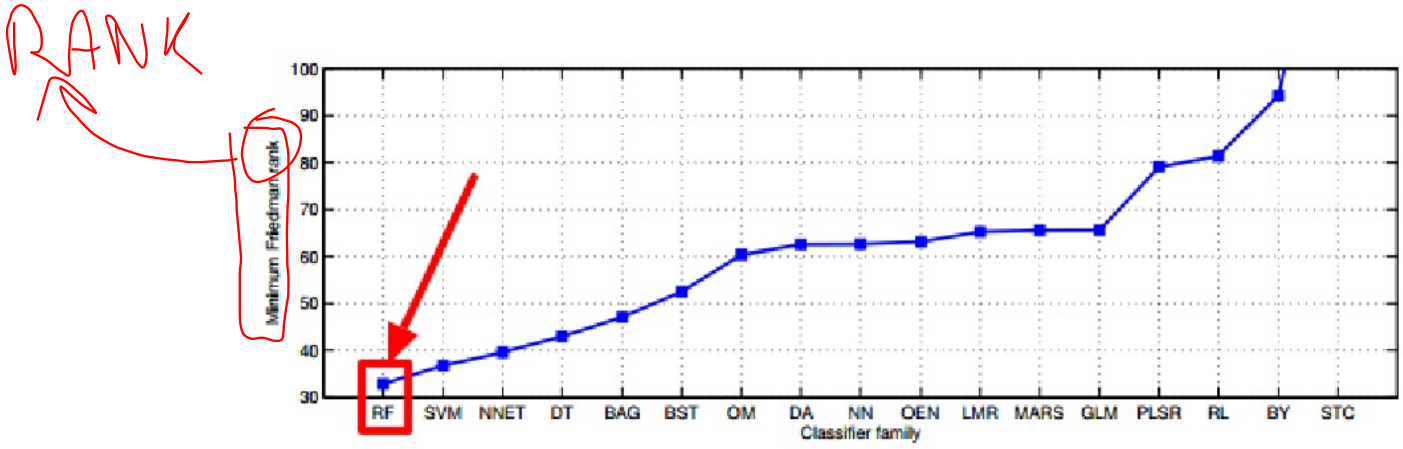
How to choose the value for m?

- $m = p \rightarrow$ bagging
- it has been shown (experimentally) that
  - m does not relate with overfitting
  - $m = \sqrt{p}$ is good for classification
  - $m = \frac{p}{3}$ is good for regression
  - (for us, default m is ok!)

Random forest

Experimentally shown: one of the "best" multi-purpose supervised classification methods

- Manuel Fernández-Delgado et al. "Do we need hundreds of classifiers to solve real world classification problems". In: *J. Mach. Learn. Res* 15.1 (2014), pp. 3133–3181



but...

No free lunch!

"Any two optimization algorithms are equivalent when their performance is averaged across all possible problems"

- David H Wolpert. "The lack of a priori distinctions between learning algorithms". In: *Neural computation* 8.7 (1996), pp. 1341–1390

Why free lunch?

- many restaurants, many items on menus, many possibly prices for each item: where to go to eat?
- no general answer
- but, if you are a vegan, or like pizza, then a best choice could exist

Q: problem? algorithm?

Observation sampling

When learning:

1. Repeat B times
   - 1.1 take a sample of the learning data
   - 1.2 consider only m on p independent variables (only for RF)
   - 1.3 learn a tree (unpruned)

Each learned tree uses only a portion of the observation in the learning data:

- for each observation, $\approx \frac{B}{3}$ trees did not considered it when learned

Observation sampling

When learning:

1. Repeat B times
   - 1.1 take a sample of the learning data
   - 1.2 consider only m on p independent variables (only for RF)
   - 1.3 learn a tree (unpruned)

Each learned tree uses only a portion of the observation in the learning data:

- for each observation, $\approx \frac{B}{3}$ trees did not considered it when learned
- those observation were *unseen* for those trees, like in cross-validation (**OOB = out-of-bag**)

Bonus 1: OOB error

- for unseen each observation there are $\frac{B}{3}$ predictions
- can "average" prediction among trees, observation and obtain an estimate of the testing error (OOB error)
  - like with cross-fold validation
  - for free!

OOB error

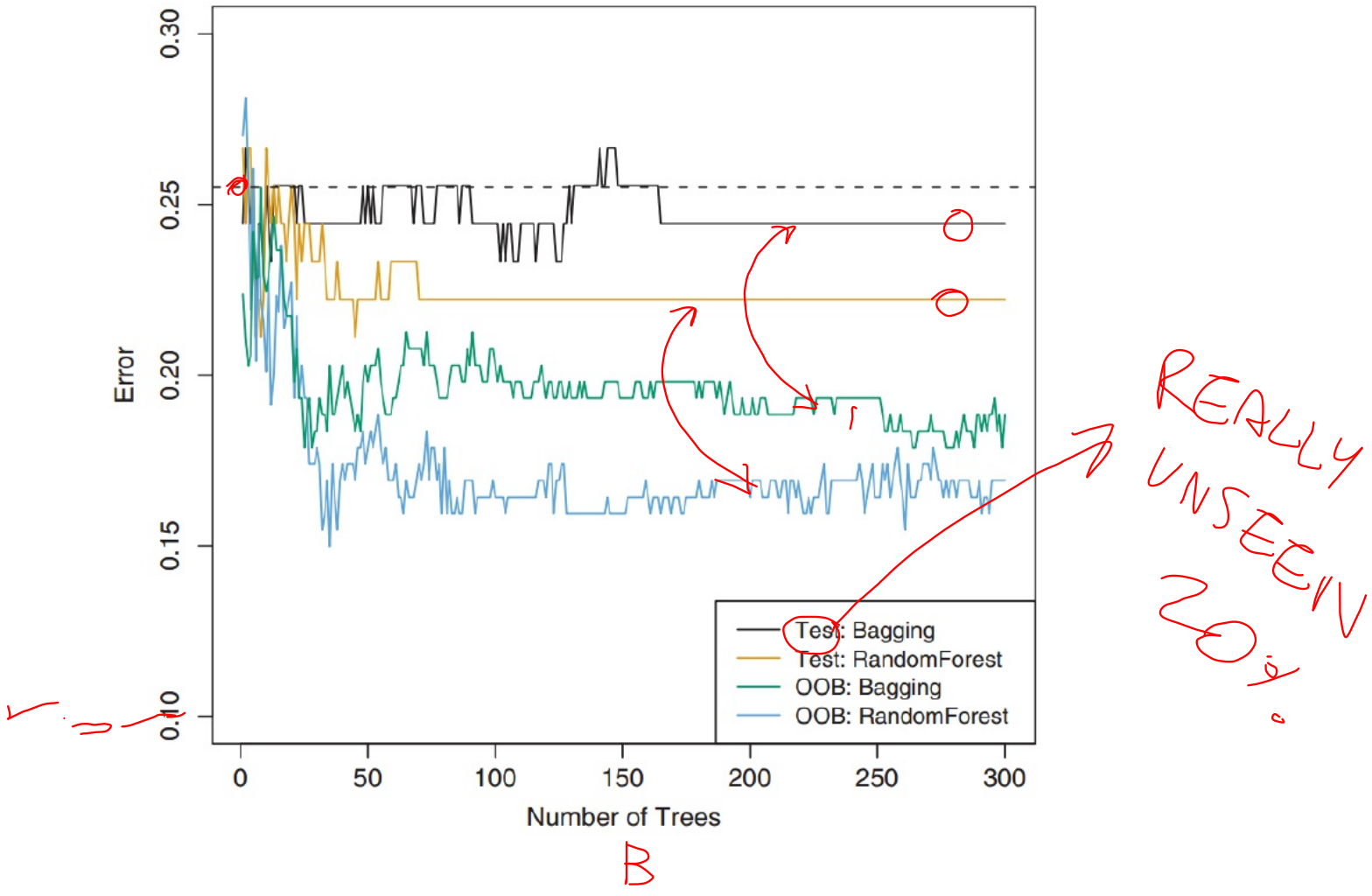


Image from An Introduction to Statistical Learning

Why estimating the test error?

Because the test data, in real world, is not available!

- will my ML solution work?

Bagging/RF and explicability

- Trees are easily understandable $\rightarrow$ explicability
- Hundreds of trees are not!

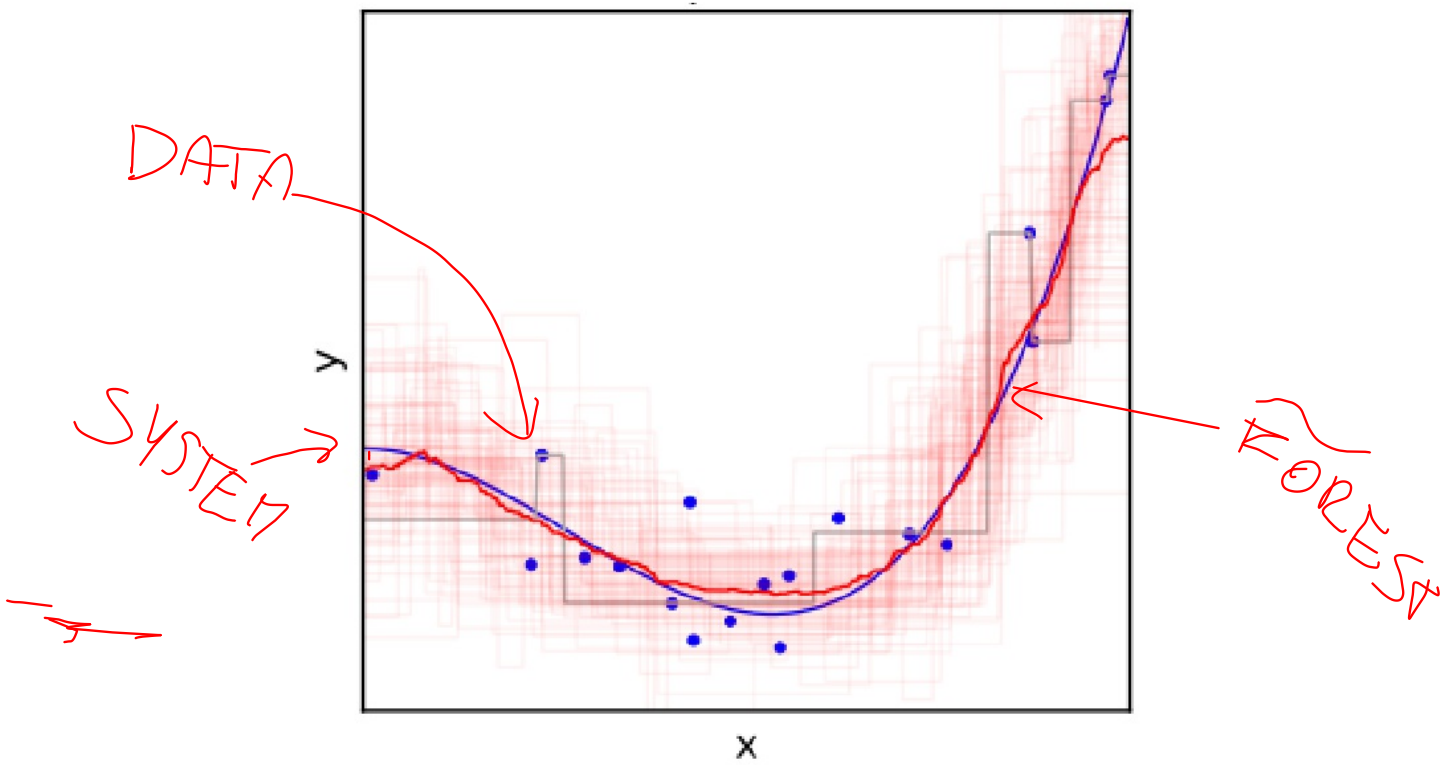


Image from F. Daolio

Bagging/RF and explicability: idea!

While learning:

1. for each tree, at each split
   - 1.1 keep note of the split variable
   - 1.2 keep note of RSS/Gini reduction
2. for each variable, sum reductions

The largest reduction, the more important the variable!

Bonus 2: variable importance

Instead of explicability based on tree shape:

- importance of variables based on RSS/Gini reduction

Nature of the prediction

Consider classification:

- tree $\rightarrow$ the class
- forest $\rightarrow$ the class, as resulting from a voting

Nature of the prediction

Consider classification:

- tree $\rightarrow$ the class
  - "virginica" is just "virginica"
- forest $\rightarrow$ the class, as resulting from a voting
  - "241 virginica, 170 versicolor, 89 setosa" is different than "478 virginica, 10 versicolor, 2 setosa"

Different **confidence** in the prediction

$\frac{241}{500} \approx 0.5$

$\frac{478}{500} \approx 0.95$

Bonus 3: confidence/tunability

Voting outcome:

- in classification, a measure of confidence of the decision
- in binary classification, voting threshold can be tuned to adjust bias towards one class (*sensitivity*)

Q: in regression?

Subsection 3

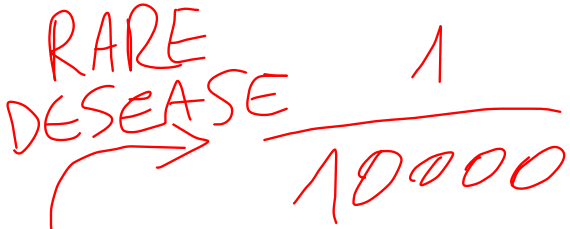
Binary classification

Binary classification

Binary classification:

- one of the most common classes of problems
- (comparative) evaluation is important!

Binary classification: evaluation



Consider the problem of classifying a person ('s data) as suffering or not suffering from a disease X.

Suppose we have "an accuracy of 99.99%". **Q:** is it good? NO

A SINGLE INDEX IS NOT ENOUGH

Binary classification: positives/negatives

Consider the problem of classifying a person ('s data) as suffering or not suffering from a disease X.

- **positive**: an observation of "suffering" class
- **negative**: an observation of "not suffering" class

In other problems, positive may mean a different thing: define it!

Effectiveness indexes: FPR, FNR

Given some labeled data and a classifier for the disease X problem, we can measure:

- the number of negative observations *wrongly* classified as positives: False Positives (**FP**)
- the number of positive observations *wrongly* classified as negatives: False Negatives (**FN**)

NUMBER OF ERRORS

THE LOWER, THE BETTER

To decouple FP, FN from data size:

"ALL NOT AFFECTED"

PREVIOUS SLIDE

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

P	1
N	9999
FP	0
FN	1

Accuracy and error rate

Relation of FPR, FNR with accuracy and error rate

$$\text{Accuracy} = 1 - \text{Error Rate}$$

$$\text{Error Rate} = \frac{\text{FN} + \text{FP}}{\text{P} + \text{N}}$$

ERRORS

OBS.

Q: Error Rate $\stackrel{?}{=} \frac{\text{FPR}+\text{FNR}}{2}$

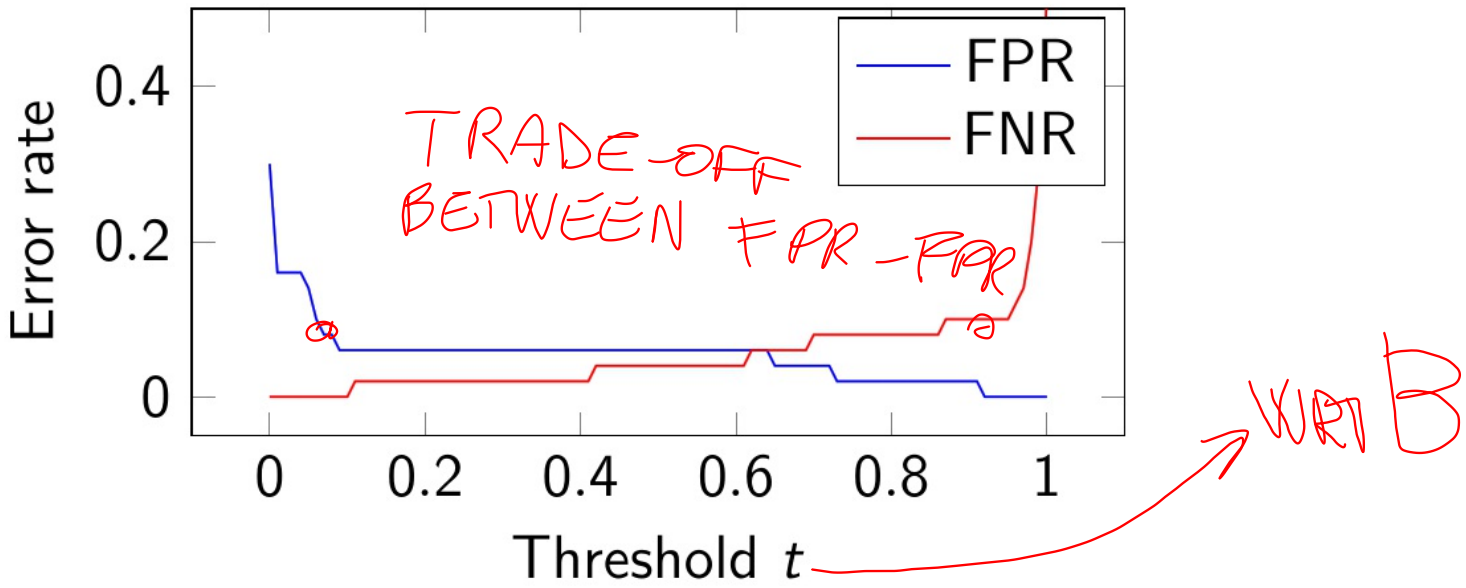
NO

YES IF N=P

FPR, FNR and sensitivity

- Suppose FPR $= 0.06$, FNR $= 0.04$ with threshold set to 0.5 (default for RF)
- One could be interested in "limiting" the FNR $\rightarrow$ change the threshold

Experimentally:



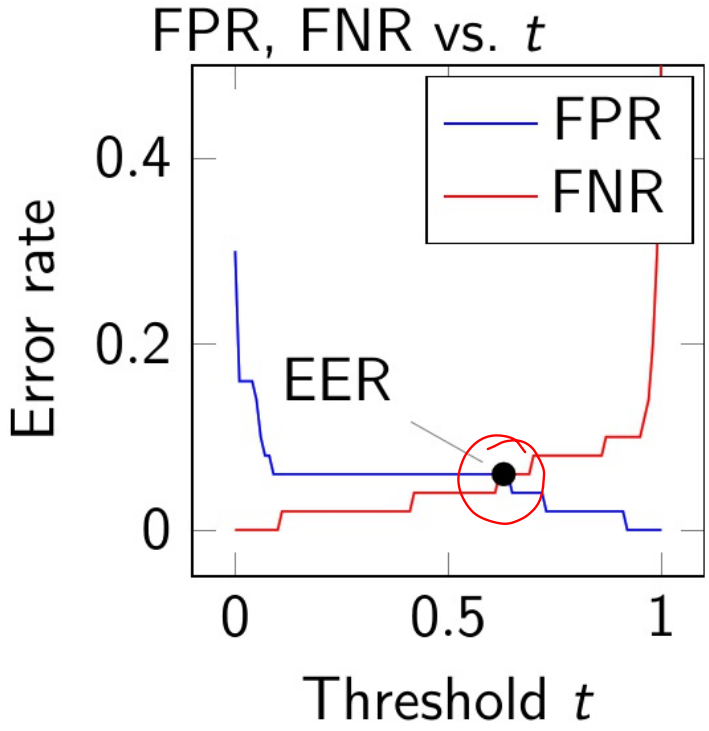
Comparing classifiers with FPR, FNR

- Classifier A: FPR $= 0.06$, FNR $= 0.04$
- Classifier B: FPR $= 0.10$, FNR $= 0.01$

Which one is the better?

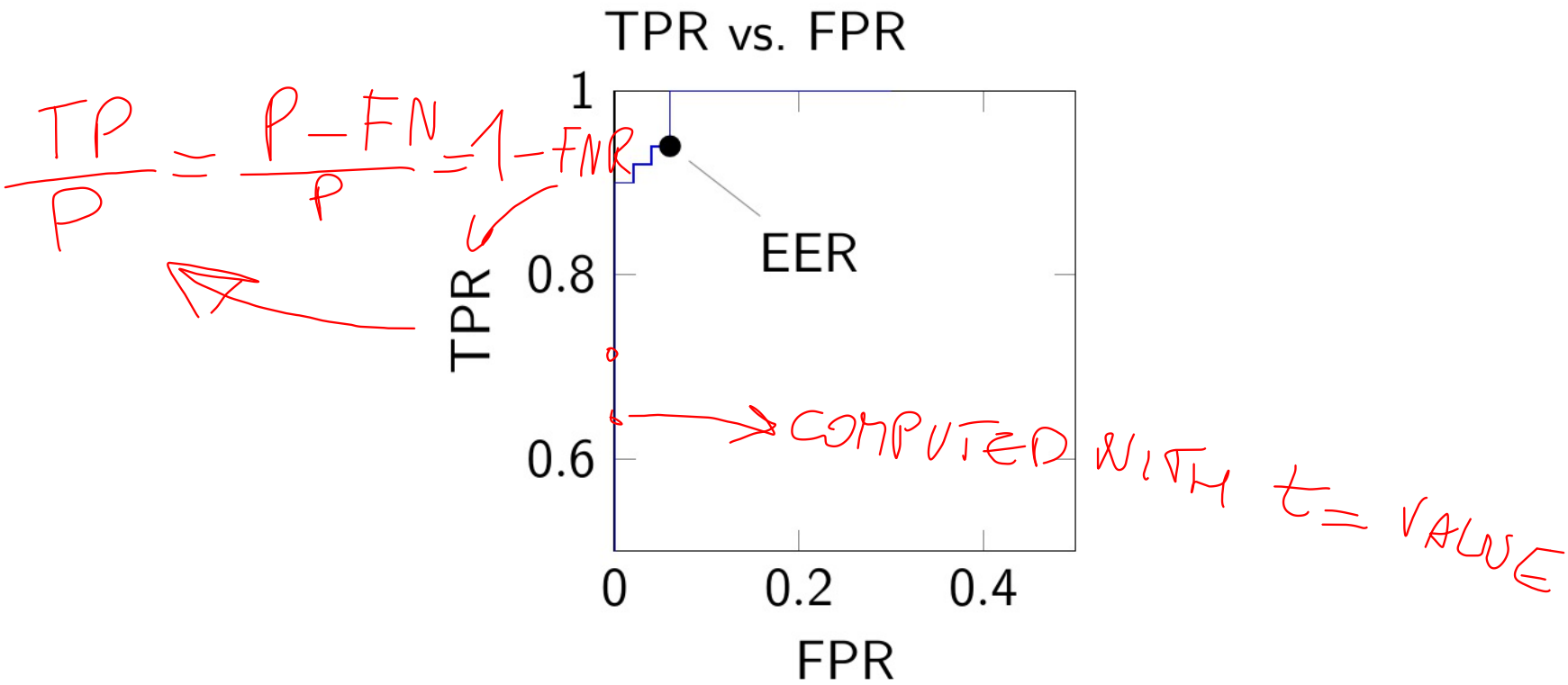
We'd like to have one single index $\rightarrow$ EER, AUC

Equal Error Rate (EER)



EER: the FPR at the value of t for which FPR = FNR

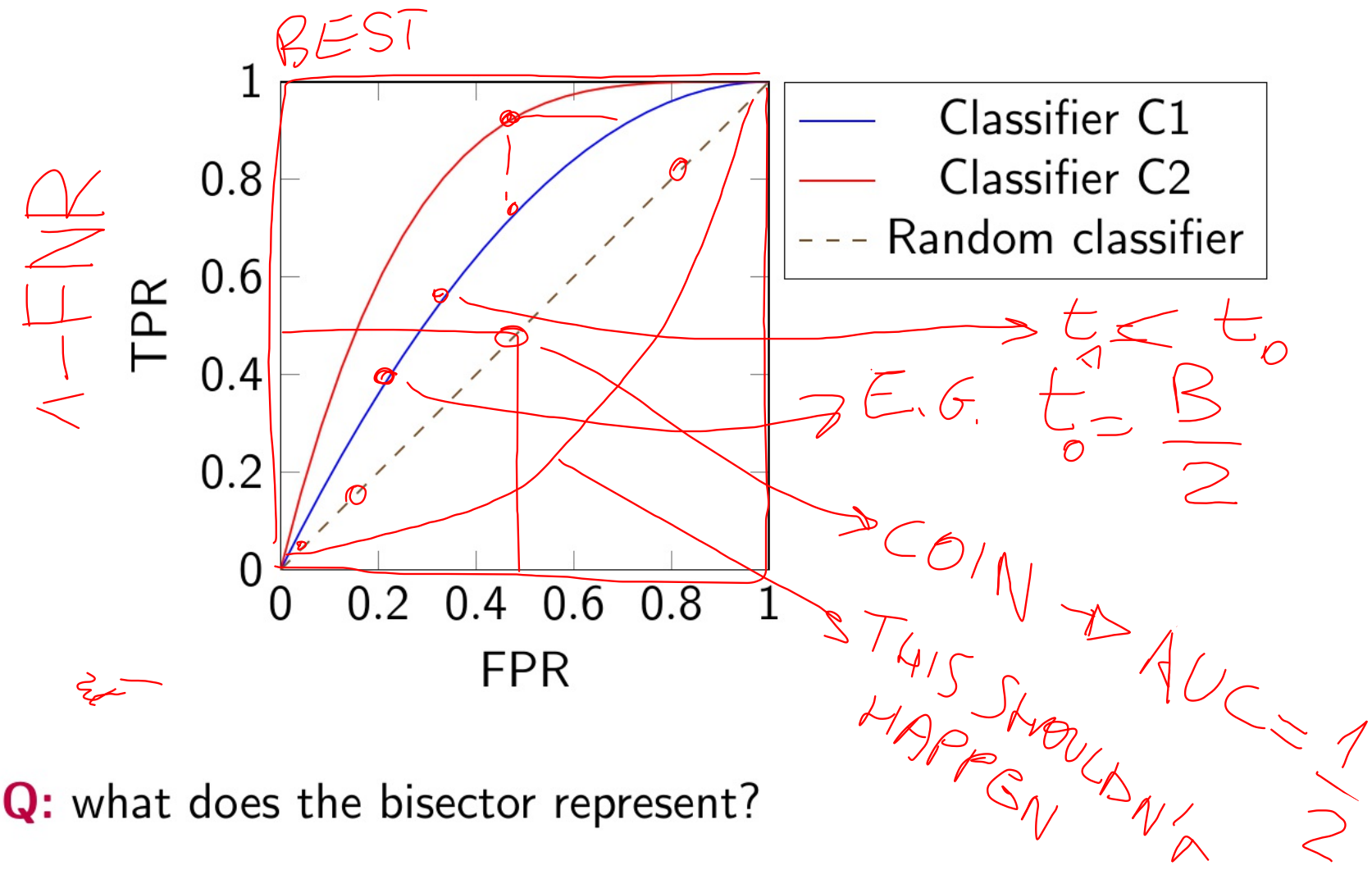
AUC: Area Under the Curve



AUC: the area under the TPR vs. FPR curve, plotted for different values of threshold t

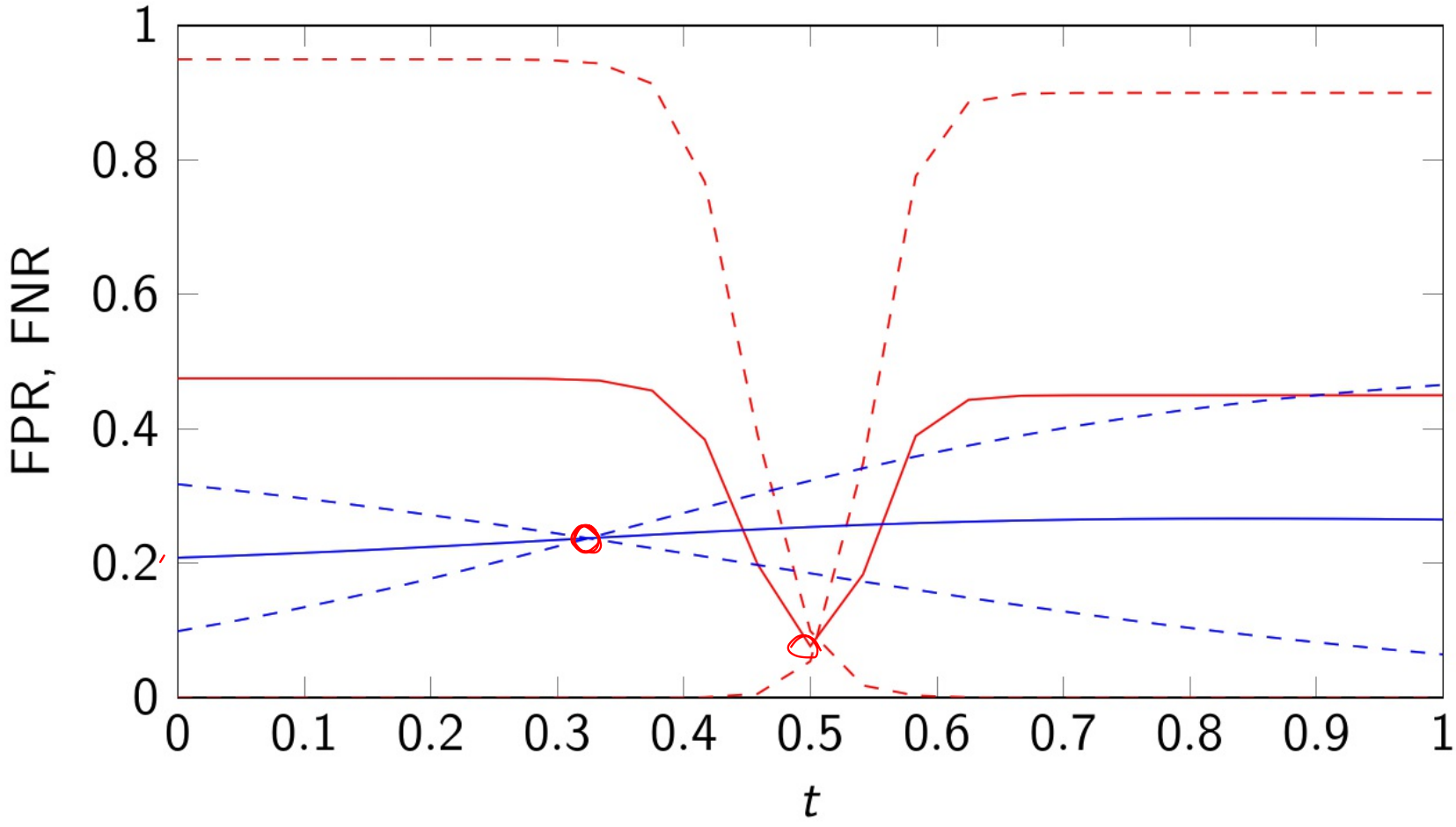
- the curve is called the *Receiver operating characteristic* (ROC)

ROC and comparison



Q: what does the bisector represent?

Other issues: robustness w.r.t. the threshold



"Same" with other parameters

Other issues: robustness w.r.t. random components

Consider A vs. B, AUC measured with cross-fold validation:

- A: $0.85, 0.73, 0.91, \cdots \rightarrow \mu = 0.83, \sigma = 0.15$
- B: $0.81, 0.78, 0.79, \cdots \rightarrow \mu = 0.81, \sigma = 0.03$

Can we say that A is better than B? (for effectiveness only)

In general, other sources of performance variability:

- random seed
- subclass of problem class (e.g., image recognition of dogs, cats, ...)

Comparing techniques

Technique A, B; different index (e.g., AUC) values:

- A $\rightarrow (x_a^1, x_a^2, \dots) \rightarrow$ random variable X_a
- B $\rightarrow (x_b^1, x_b^2, \dots) \rightarrow$ random variable X_b

Do X_a, X_b follow different distributions?

- yes: A and B are different (concerning the AUC)
- no: difference in μ_a, μ_b might be due to randomness $\rightarrow$ A, B are not *significantly* different

Statistical significance in a nutshell

Just the way of thinking:

1. State a set of assumptions (the *null hypothesis* H_0), e.g.:
   - X_a, X_b are normally distributed and independent
   - $\bar{x}_a = \bar{x}_b$ (or $\bar{x}_a \geq \bar{x}_b$)
   - any other assumption in the *statistical model*

Statistical significance in a nutshell

Just the way of thinking:

1. State a set of assumptions (the *null hypothesis* H_0), e.g.:
   - X_a, X_b are normally distributed and independent
   - $\bar{x}_a = \bar{x}_b$ (or $\bar{x}_a \geq \bar{x}_b$)
   - any other assumption in the *statistical model*
2. Perform a statistical test, appropriate choice depending on many factors, e.g.:
   - Wilcoxon test (many versions)
   - Friedman (many versions)
   - ...

Statistical significance in a nutshell

Just the way of thinking:

1. State a set of assumptions (the *null hypothesis* H_0), e.g.:
   - X_a, X_b are normally distributed and independent
   - $\bar{x}_a = \bar{x}_b$ (or $\bar{x}_a \geq \bar{x}_b$)
   - any other assumption in the *statistical model*
2. Perform a statistical test, appropriate choice depending on many factors, e.g.:
   - Wilcoxon test (many versions)
   - Friedman (many versions)
   - ...
3. ...which outputs a p-value $\in [0, 1]$
   - 0 is "good", 1 is "bad"

p-value: meaning

0 is "good", 1 is "bad"
The p-value is the degree to which the data conform to the pattern predicted by the null hypothesis

- p-value $= P(x_a^1, x_a^2, \ldots, x_b^1, x_b^2, \ldots | H_0)$

If p-value is low:

- we've been very (un)lucky in having observed $x_a^1, x_a^2, \ldots, x_b^1, x_b^2, \ldots$
- "maybe" because H_0 is not true

p-value: meaning

0 is "good", 1 is "bad"
The p-value is the degree to which the data conform to the pattern predicted by the null hypothesis

- p-value $= P(x_a^1, x_a^2, \ldots, x_b^1, x_b^2, \ldots | H_0)$

If p-value is low:

- we've been very (un)lucky in having observed $x_a^1, x_a^2, \ldots, x_b^1, x_b^2, \ldots$
- "maybe" because H_0 is not true
  - **Warning!** Any part of H_0, not necessarily the $\bar{x}_a = \bar{x}_b$ part!

Statistical significance

Things are much more complex than this...

Some interesting papers:

- Joaquín Derrac et al. "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms". In: *Swarm and Evolutionary Computation* 1.1 (2011), pp. 3–18
- Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. "How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments". In: *arXiv preprint arXiv:1806.08295* (2018)
- Sander Greenland et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations". In: *European journal of epidemiology* 31.4 (2016), pp. 337–350