Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering

Artificial intelligence – VIMIAC16-EN, VIMIAC10

2024 Fall Semester

Dr. Gábor Hullám

Model evaluation

Decisions

- We want to make the right decision!
- But what is the right decision?
- Usually there is no perfect decision, only one is better and the other is worse!
- In order to compare two alternatives (decision methods), we need to have a scalar measure of the goodness of the decision
- First of all, we deal with the qualification and evaluation of decisions
- Basically, we present the methods on binary (good/bad, true/false, sick/healthy, innocent/guilty) decisions

Evaluation of binary decisions

Bivalent function => classification = binary decision

true state f(x) Fact	assumed state h(x) hypothesis,	decision
patient is **sick** f(x) = **I**	**modell detects it** h(x) = **I**	We treat it, we do it right **True Positive TP**
patient is **healthy** f(x) = **H**	**modell detects it** h(x) = **H**	We don’t treat it, we're doing the right thing now **True Negative TN**

Evaluation of binary decisions

true state f(x) Fact	assumed state h(x) hypothesis	decision
patient is **sick** f(x) = **I**	**We don't recognize it** h(x) = **H**	We don’t treat it and we don't do it right **False Negative: FN,** **Type 2 error**
patient is **healthy** f(x) = **H**	**We don't recognize it** h(x) = **I**	We are treating it unnecessarily, we are not doing well **False Positive: FP,** **Type 1 error**

Confusion matrix

patient is **sick** **We don't recognize it** **We don’t treat it, we don't do it well**

f(x) = **I** h(x) = **H** **False Negative: FN,**
Type 2 error

patient is **healthy** **We don't recognize it** **We are treating it unnecessarily, we are not doing well**

f(x) = **H** h(x) = **I** **False Positive: FP,**
Type 1 error

Facts - Reality

Decision - Model

	Sick	Healthy
Sick	TP	FP
Healthy	FN	TN

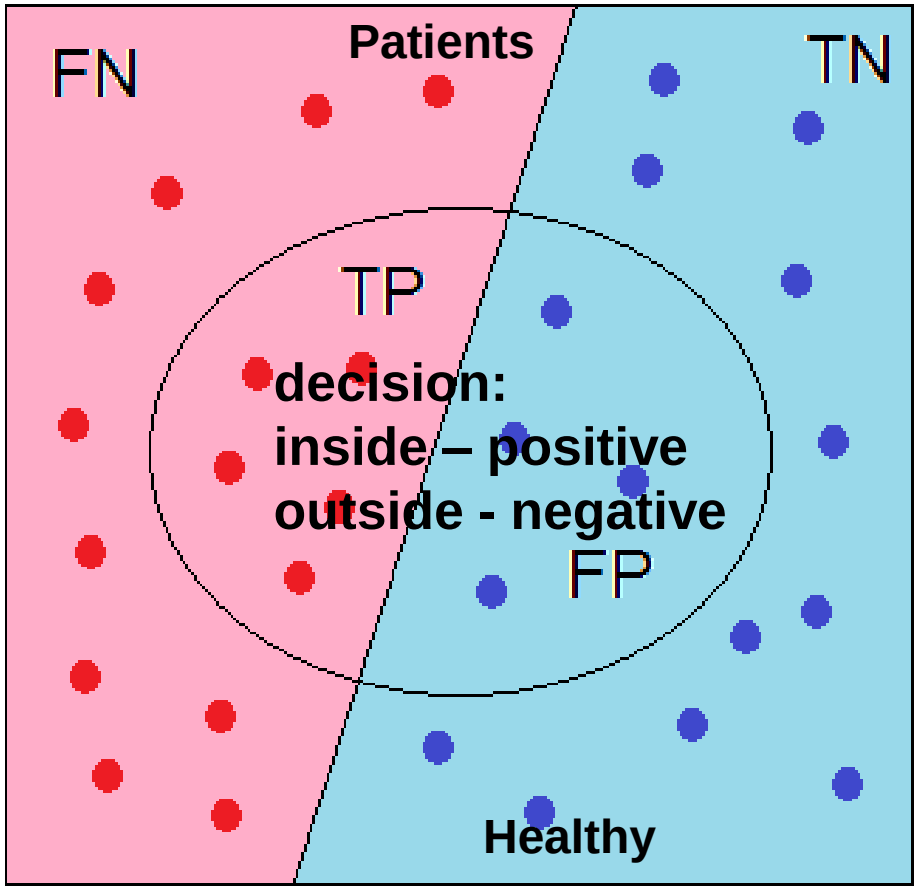
Performance metrics

true positive rate (**TPR**)
(recall, sensitivity)

TPR = TP/P = TP / (TP+FN)

true negative rate (**TNR**)
(specificity)

TNR = TN/N = TN / (TN+FP)



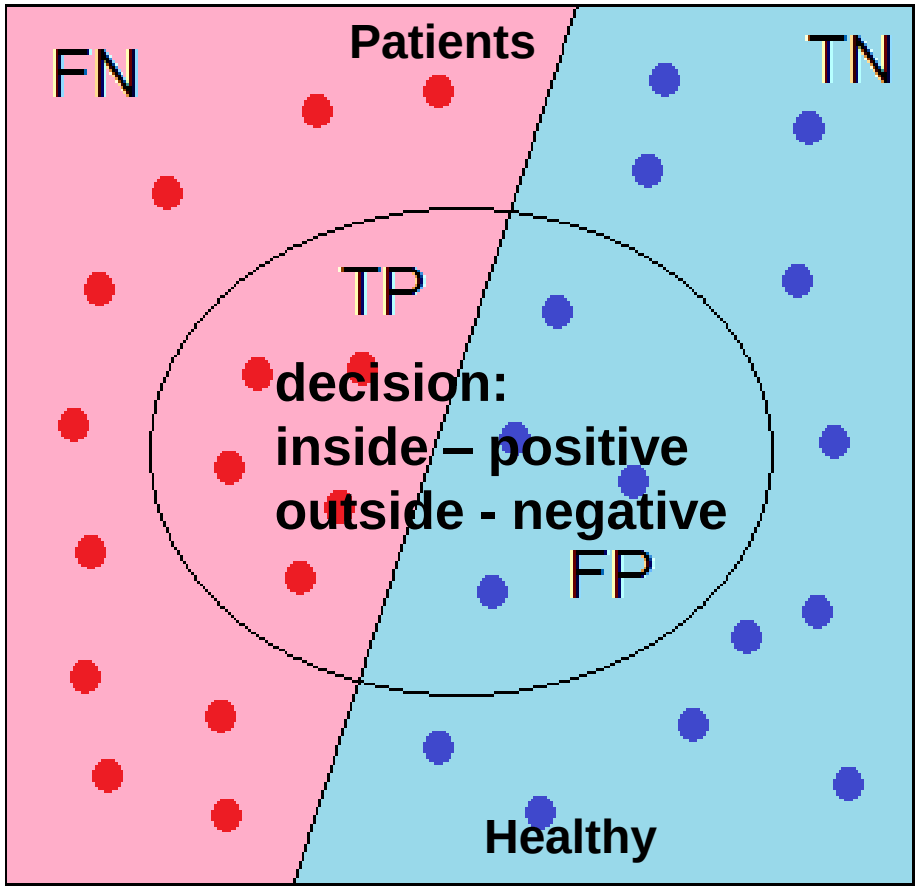
Performance metrics

true negative rate (**TNR**)
(specificity)

TNR = TN/N = TN / (TN+FP)

false positive rate (**FPR**)
(false alarm rate, fall-out)

FPR = FP/N = FP / (FP+TN)



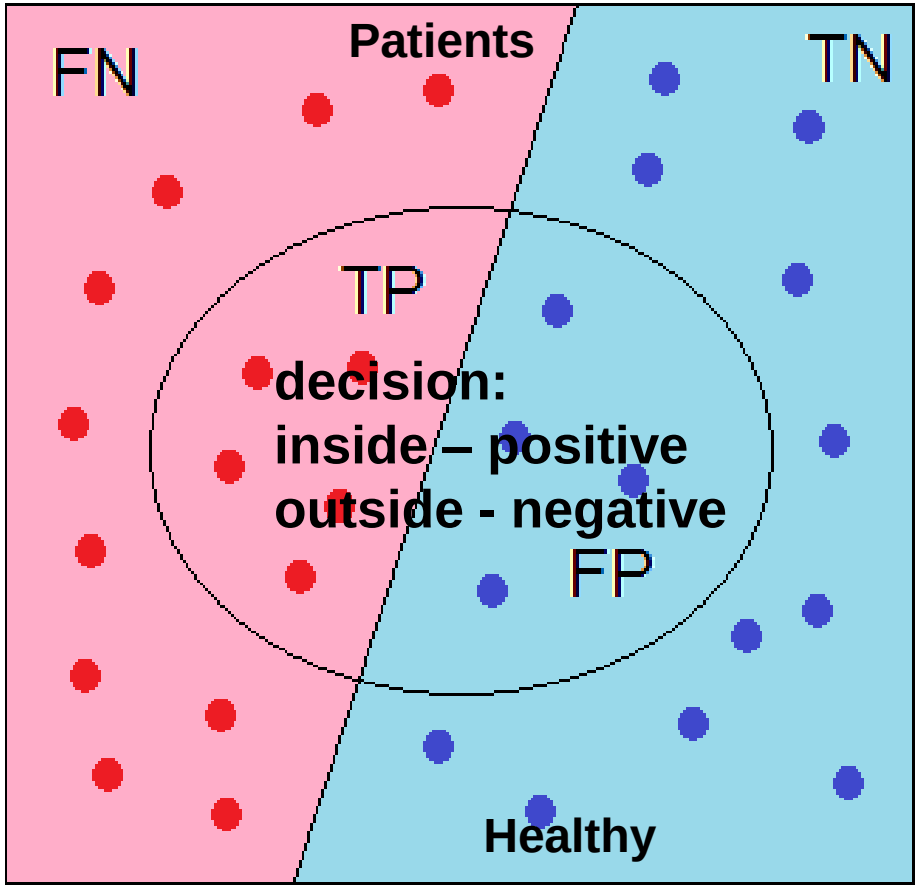
Performance metrics

Positive predictive value (**PPV**)
(precision)

PPV = TP/ (TP+FP)

Negative predictive value (**NPV**)

NPV = TN/ (TN+FN)

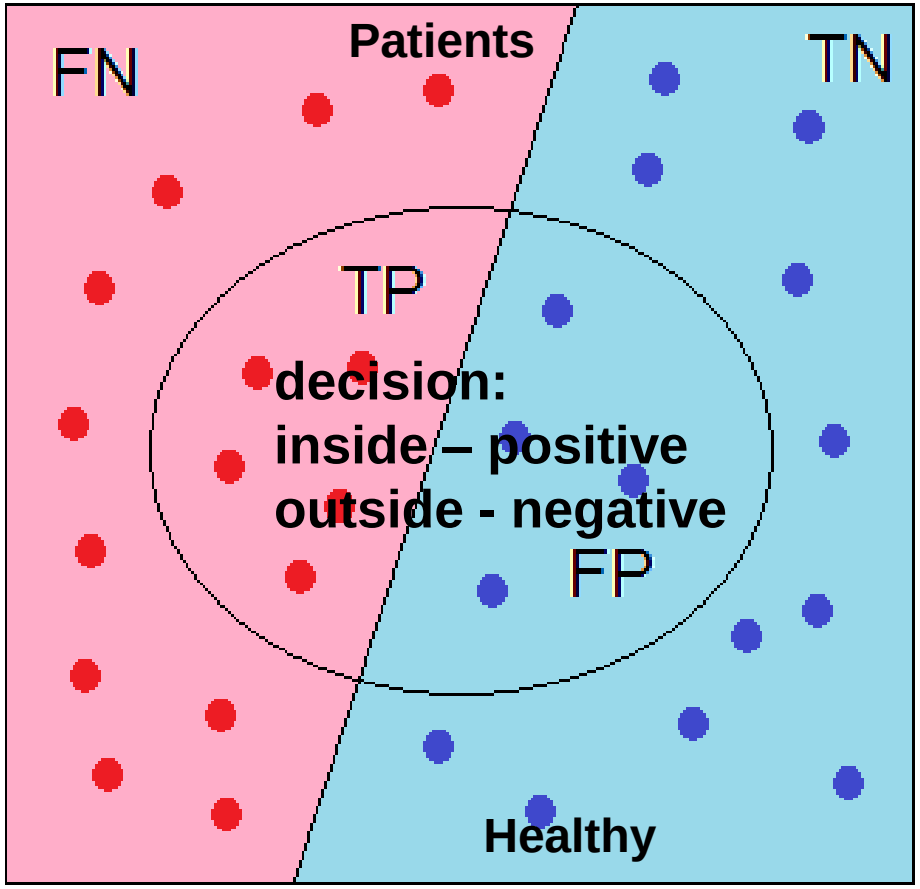


Performance metrics

Accuracy (ACC)

ACC = TP+TN/ (TP+FP+TN+FN)

= TP+TN/ (P+N)



Summary table

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

---

sensitivity or **true positive rate (TPR)**

eqv. with hit rate, recall

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

specificity (SPC) or **true negative rate (TNR)**

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

precision or **positive predictive value** (PPV)

$$PPV = \frac{TP}{TP + FP}$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN}$$

fall-out or **false positive rate (FPR)**

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - SPC$$

false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

miss rate or **false negative rate** (FNR)

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} = 1 - TPR$$

---

accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = \frac{2TP}{2TP + FP + FN}$$

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

Performance metrics - Example

In the case outlined in the figure:

Our decisions:

TP = 5, TN = 11,

FP = 4, FN = 10

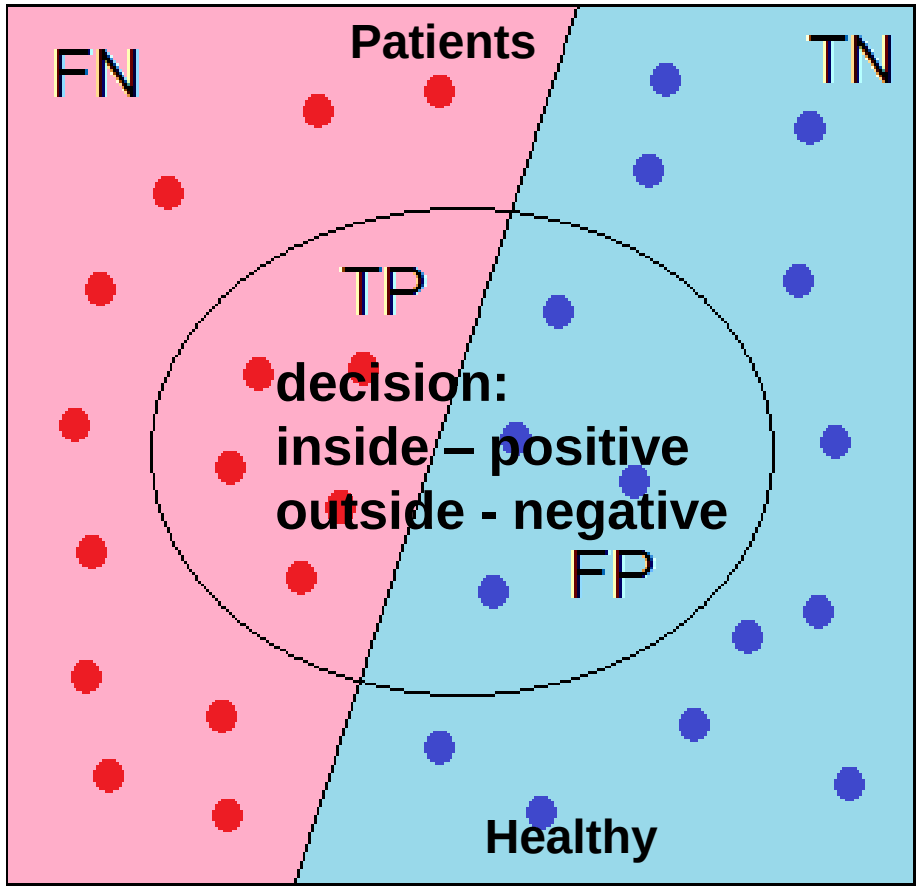
Facts: P = 15, N = 15

TPR = TP / (TP+FN) = 5/15 = 0.33

FPR = FP / (FP+TN) = 4/15 = 0.27

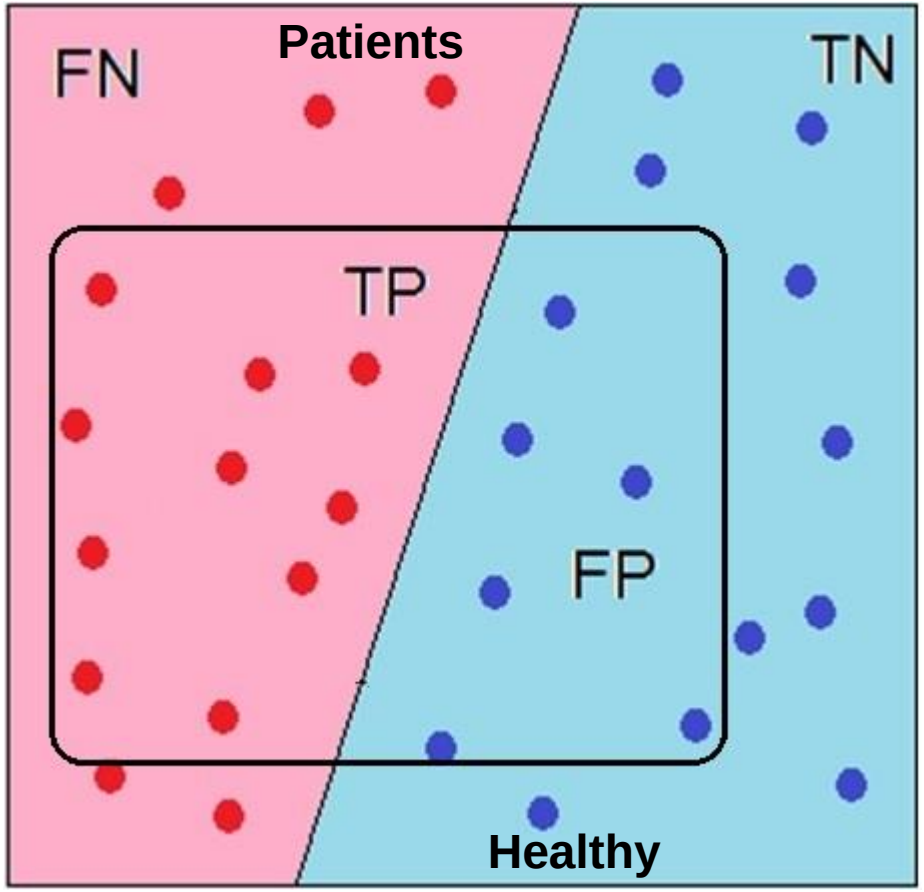
TNR = TN / (FP+TN) = 11/15 = 0.73

ACC = (TP+TN)/(P+N) = 1/2 = 0.5

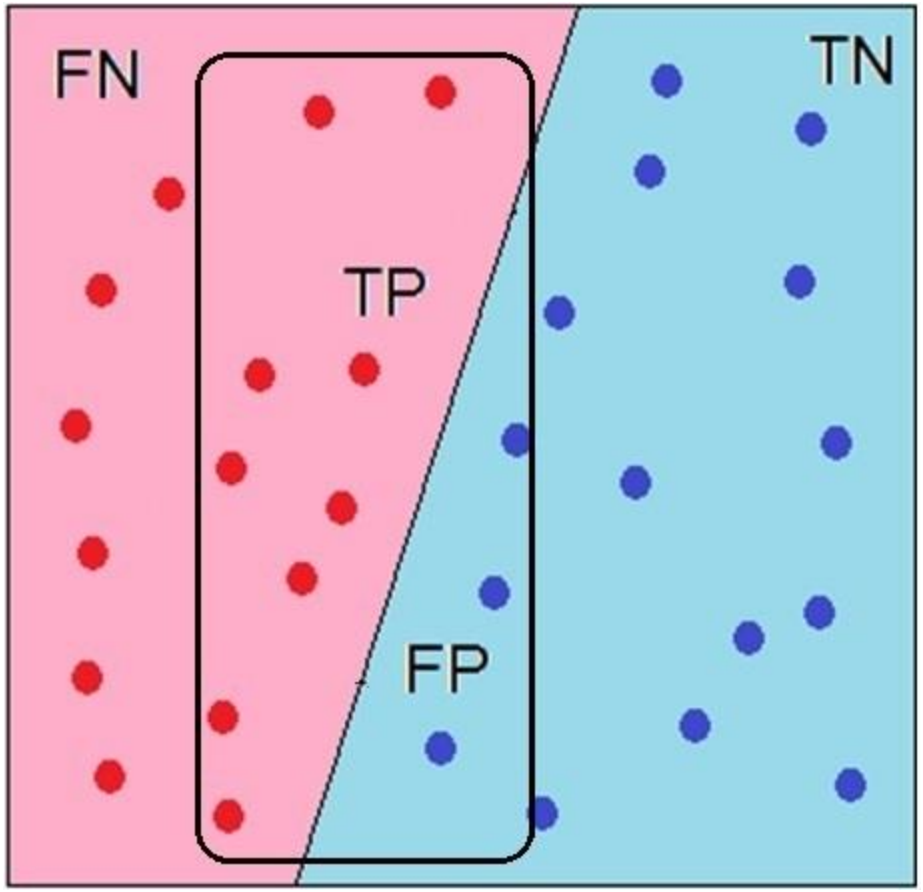


We tried it on a few other suggestions...

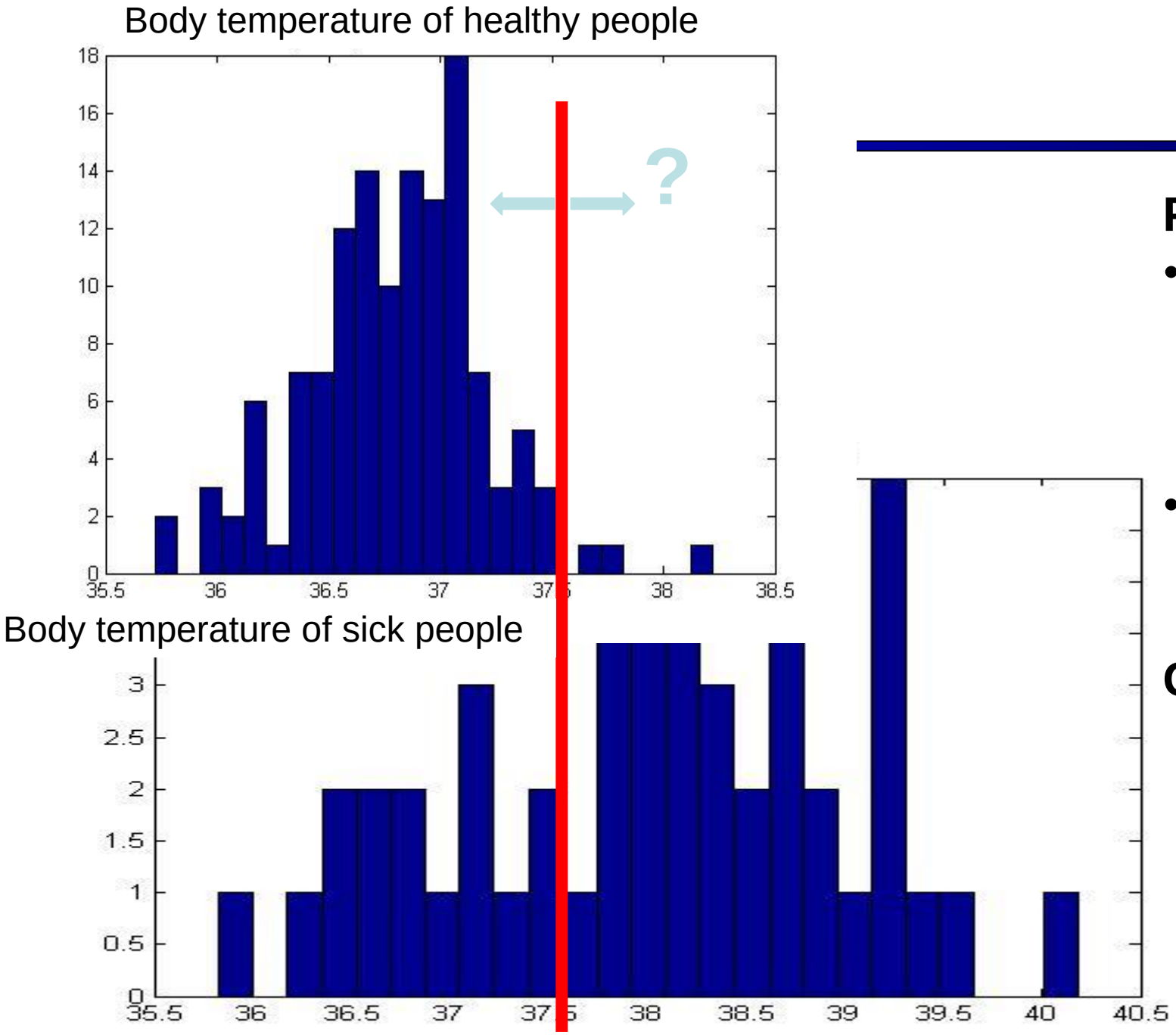
Which one is better?



TP = 10, TN = 9, FP = 6, FN = 5



TP = 9, TN = 12, FP = 3, FN = 6

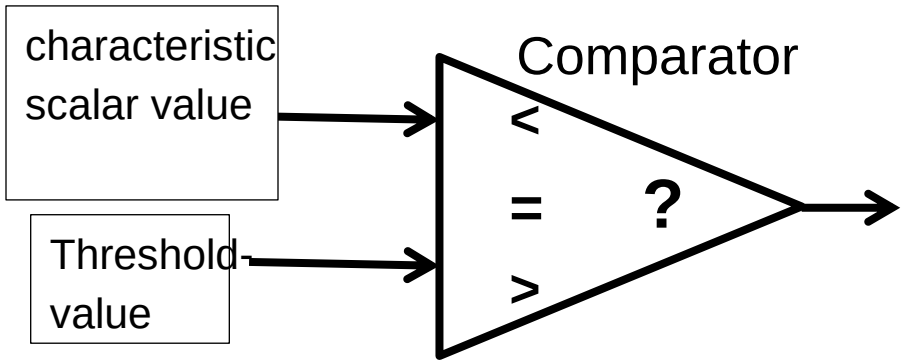


Problem:

- We have to make a decision, based on the value of a numerical parameter, about what kind of case we are facing
- e.g. based on body temperature, whether you are healthy or sick?

Goal: figuring out what is (optimal) threshold value for the test

Decision modell

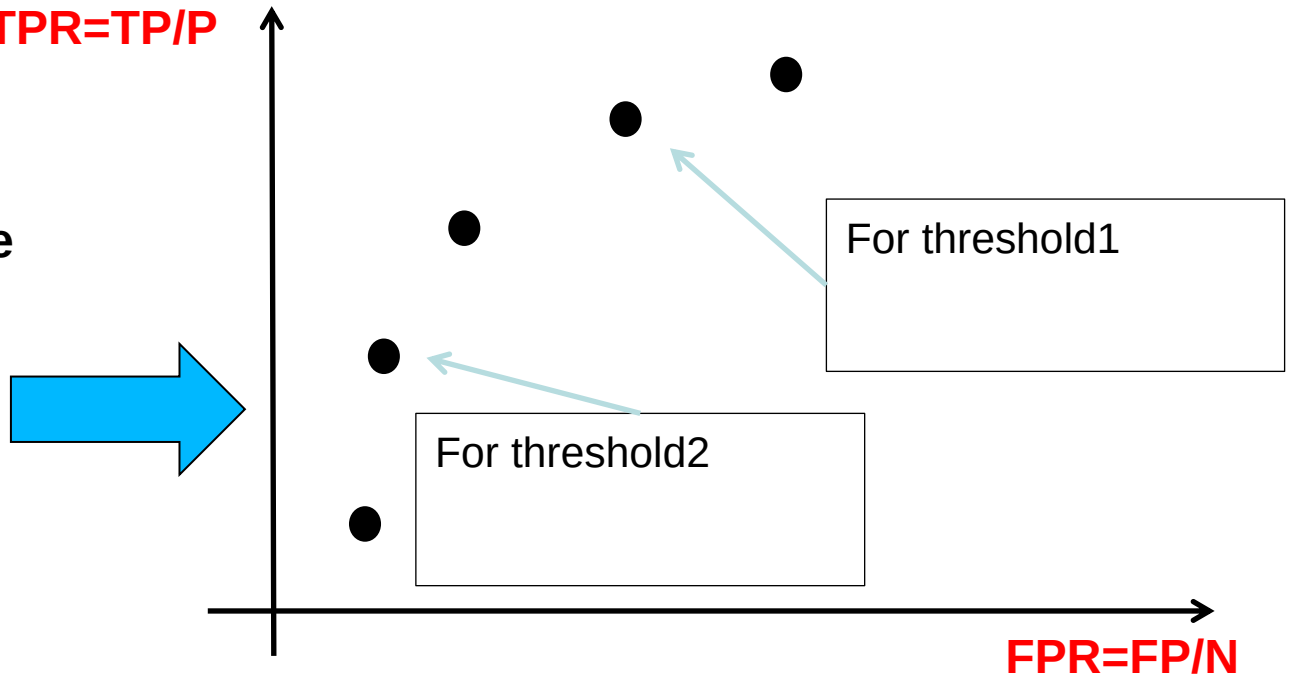


For example (it could also be the other way around):
if > then output=1
We take it as a positive case
if <=, then output= 0
We take it as a negative case

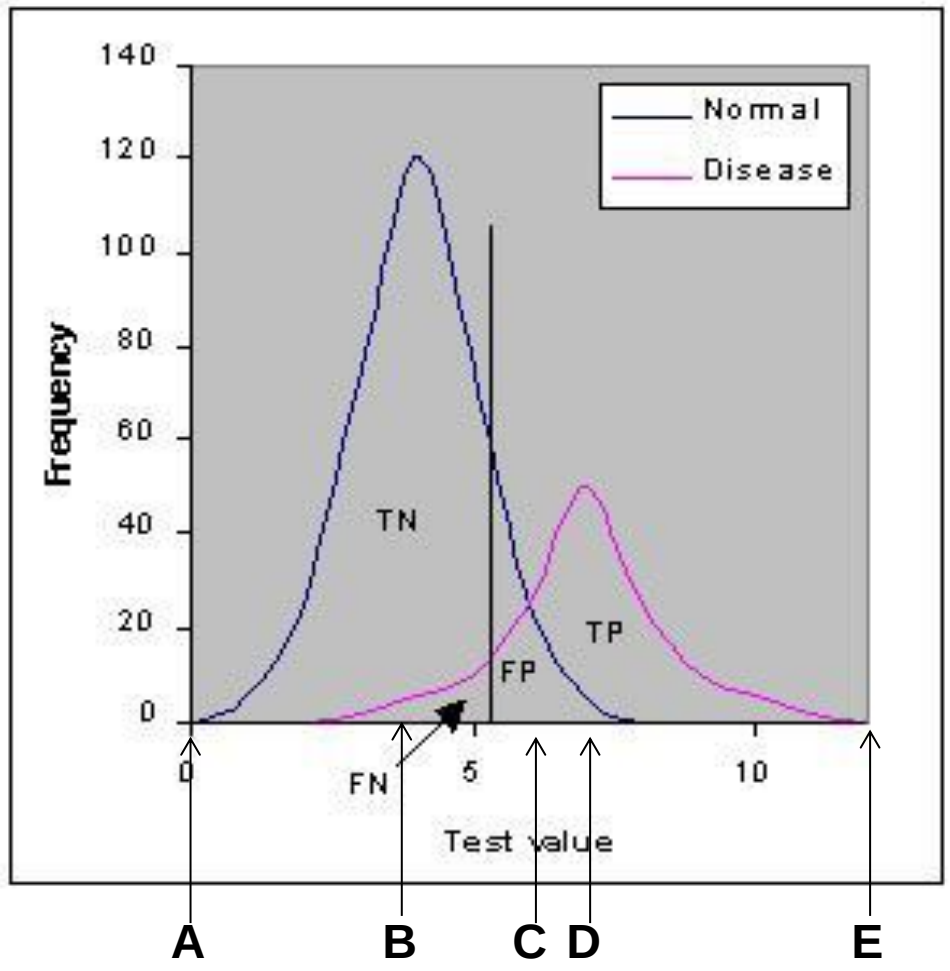
Example:
If the measured body temperature is greater than 38°C, then fever (disease), if less than or equal to it, then no fever (healthy).

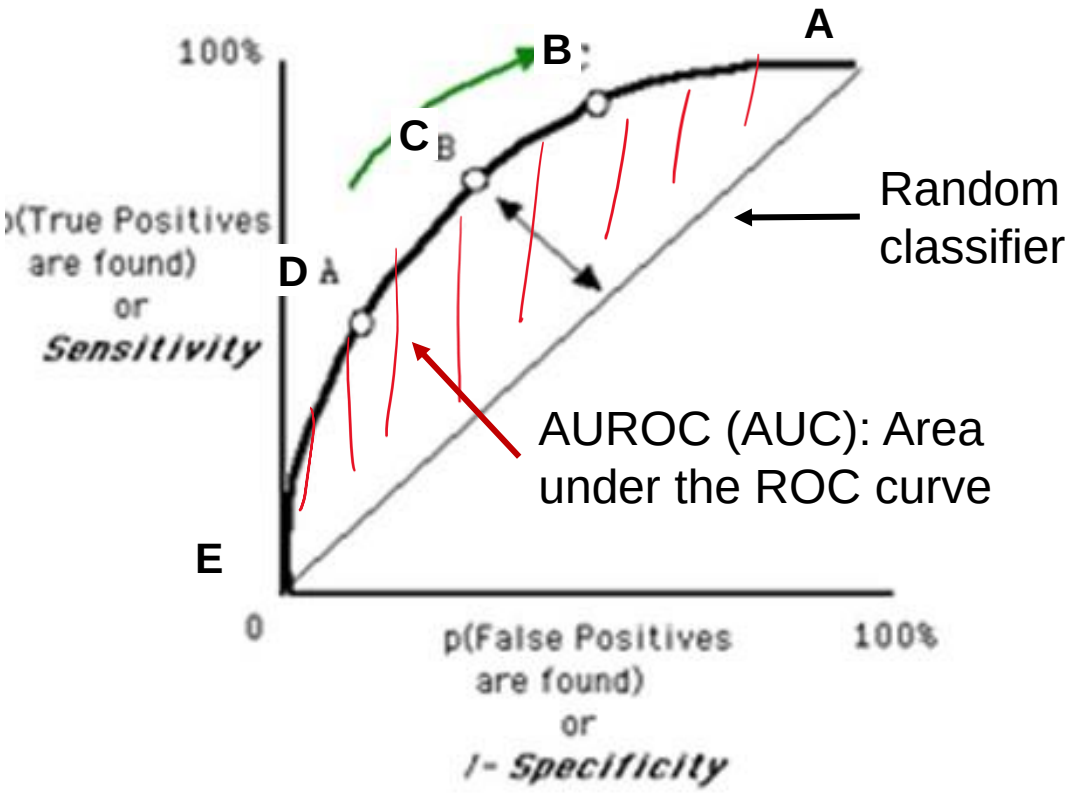
Model evaluation at different "working points"”

This is the **ROC curve** (Receiver Operating Characteristic curve)



ROC: Receiver Operating Characteristic





Above the threshold the patient is considered sick, below it is considered healthy

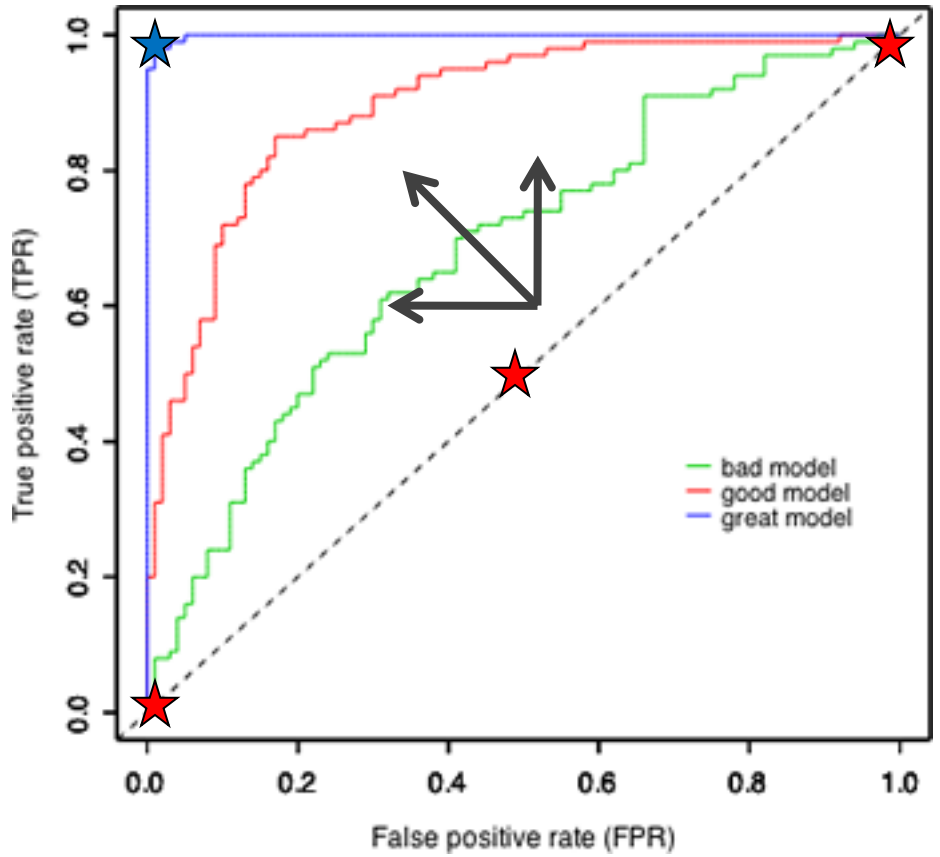
AUC – compact performance indicator

true positive rate (**TPR**)
TPR = TP/P = TP / (TP+FN)

false positive rate (**FPR**)
FPR = FP/N = FP / (FP+TN)

ROC: **R**eceiver **O**perating **C**haracteristic (curve)

AUC: **A**rea **U**nder RO**C**urve



Mistakes can have different costs

Type 2 error **false negative (FN)**

– we think it's healthy (H0) but it's sick (T1)

"We think friend is enemy”

cost: C_{01}

Type 1 error **false positive (FP)**

– we think he is sick (H1) but healthy (T0)

– we think it is an enemy, but it is a friend

cost: C_{10}

Most of the time, false negatives are worse: $C_{01} >> C_{10}$

Making the right decisions also comes at a cost

true negative(TN)

– we think it is healthy (H0), indeed it is (T0)
"We think it is a friend, and really it is

cost: C_{00} We need to do an investigation
– medical screening, program test, military radar

true positive (TP)

– we think it is sick (H1), it really is (T1)
"We think it's an enemy, and it really is

cost: C_{11} we need to do an investigation + treatment
– medical screening, program test, military radar
+ handling, anti-aircraft missile, bug fix

So, four types of costs influence the decision:

C_{00}, C_{11}, C_{01}, C_{10}

K1 K2

$$C_{10} \cdot P_0 \cdot N \cdot P(z_k|T0) + C_{11} \cdot P_1 \cdot N \cdot P(z_k|T1) < C_{00} \cdot P_0 \cdot N \cdot P(z_k|T0) + C_{01} \cdot P_1 \cdot N \cdot P(z_k|T1)$$

$$(C_{10} - C_{00}) \cdot P_0 \cdot N \cdot P(z_k|T0) < (C_{01} - C_{11}) \cdot P_1 \cdot N \cdot P(z_k|T1)$$

$$(C_{10} - C_{00}) \cdot P_0 \cdot P(z_k|T0) < (C_{01} - C_{11}) \cdot P_1 \cdot P(z_k|T1)$$

- If K1<K2 then choose positive for all cases with zk
- If K1>K2 then choose negative for all cases with zk

Decision based on costs

- **Our decision depends to a large extent on our good estimation of each cost**
- **If C01 is large, that's what we're afraid of, then something else will develop than**
- **like if we are afraid of C10 and take it as big.**

Examples:

Law: Are we more afraid that an innocent person will be convicted, or that a criminal will get away with it (or commit it again later)?

Security question: are we afraid that a terrorist will somehow sneak in, or are we afraid of wrongfully accusing someone?

Healthcare: are we more afraid of not noticing someone's illness or of needlessly scaring them and ordering them back for examination?