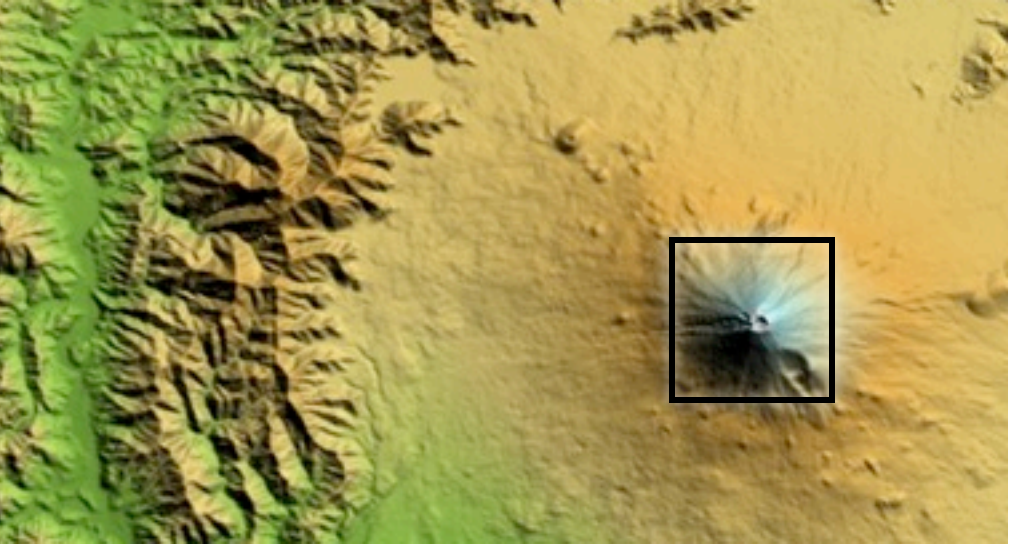


Subgroup Discovery

Universiteit Leiden

Exploratory Data Analysis

Universiteit Leiden

Exploratory Data Analysis

- Classification: model the dependence of the target on the remaining attributes.
  - problem: sometimes uses only some of the available dependencies, or classifier is a black-box.
  - for example: in decision trees, some attributes may not appear because of overshadowing.

Exploratory Data Analysis

- Classification: model the dependence of the target on the remaining attributes.
  - problem: sometimes uses only some of the available dependencies, or classifier is a black-box.
  - for example: in decision trees, some attributes may not appear because of overshadowing.
- Exploratory Data Analysis: understanding the effects of *all* attributes on the target.

Exploratory Data Analysis

- Classification: model the dependence of the target on the remaining attributes.
  - problem: sometimes uses only some of the available dependencies, or classifier is a black-box.
  - for example: in decision trees, some attributes may not appear because of overshadowing.
- Exploratory Data Analysis: understanding the effects of *all* attributes on the target.

Q: How can we use ideas from C4.5 to approach this task?

Exploratory Data Analysis

- Classification: model the dependence of the target on the remaining attributes.
  - problem: sometimes uses only some of the available dependencies, or classifier is a black-box.
  - for example: in decision trees, some attributes may not appear because of overshadowing.
- Exploratory Data Analysis: understanding the effects of *all* attributes on the target.

Q: How can we use ideas from C4.5 to approach this task?

A: Why not list the info gain of all attributes, and rank according to this?

Interactions between Attributes

- Single-attribute effects are not enough
- XOR problem is extreme example: 2 attributes with no info gain form a good model

- Apart from

  A=a, B=b, C=c, …
- consider also

  A=a∧B=b, A=a∧C=c, …, B=b∧C=c, …

  A=a∧B=b∧C=c, …

  …

Subgroup Discovery Task

"Find all subgroups within the inductive constraints that show a significant deviation in the distribution of the target attribute"

- Inductive constraints:
  - Minimum support
  - (Maximum support)
  - Minimum quality (Information gain, X^2, WRAcc)
  - Maximum complexity
  - …

Confusion Matrix

- A *confusion matrix* (or *contingency table*) describes the frequency of the four combinations of subgroup and target:
  - within subgroup, positive
  - within subgroup, negative
  - outside subgroup, positive

		target		
		T	**F**	
subgroup	**T**	**.42**	**.13**	**.55**
	F	**.12**	**.33**	
		.54		**1.0**

Universiteit Leiden

Confusion Matrix

- High numbers along the TT-FF diagonal means a *positive* correlation between subgroup and target
- High numbers along the TF-FT diagonal means a *negative* correlation between subgroup and target
- Target distribution on DB is fixed

		target		
		T	**F**	
subgroup	**T**	**.42**	**.13**	**.55**
	F	**.12**	**.33**	**.45**
		.54	**.46**	**1.0**

Universiteit Leiden

Confusion Matrix

- High numbers along the TT-FF diagonal means a *positive* correlation between subgroup and target
- High numbers along the TF-FT diagonal means a *negative* correlation between subgroup and target
- Target distribution on DB is fixed

target

subgroup	T	F	
T	.42	.13	.55
F	.12	.33	.45
	.54	.46	1.0

Universiteit Leiden

Confusion Matrix

- High numbers along the TT-FF diagonal means a *positive* correlation between subgroup and target
- High numbers along the TF-FT diagonal means a *negative* correlation between subgroup and target
- Target distribution on DB is fixed

		target		
		T	**F**	
subgroup	**T**	**.42**	**.13**	**.55**
	F	**.12**	**.33**	**.45**
		.54	**.46**	**1.0**

Universiteit Leiden

Confusion Matrix

- High numbers along the TT-FF diagonal means a *positive* correlation between subgroup and target
- High numbers along the TF-FT diagonal means a *negative* correlation between subgroup and target
- Target distribution on DB is fixed

target

subgroup	T	F	
T	.42	.13	.55
F	.12	.33	.45
	.54	.46	1.0

Universiteit Leiden

Confusion Matrix

- High numbers along the TT-FF diagonal means a *positive* correlation between subgroup and target
- High numbers along the TF-FT diagonal means a *negative* correlation between subgroup and target
- Target distribution on DB is fixed

		target		
		T	**F**	
subgroup	**T**	**.42**	**.13**	**.55**
	F	**.12**	**.33**	**.45**
		.54	**.46**	**1.0**

Universiteit Leiden

Quality Measures

A *quality measure* for subgroups summarizes the interestingness of its confusion matrix into a single number

WRAcc, weighted relative accuracy

- $WRAcc(S,T) = p(ST) - p(S) \cdot p(T)$
- between −.25 and .25, 0 means uninteresting
- Balance between coverage and unexpectedness

target

subgroup	T	F	
T	.42	.13	.55
F	.12	.33	
	.54		1.0

$WRAcc(S,T) = p(ST) - p(S) \cdot p(T)$
$= .42 - .297 = .123$

Universiteit Leiden

Quality Measures

- WRAcc: Weighted Relative Accuracy
- Information gain
- X^2
- Correlation Coefficient
- Laplace
- Jaccard
- Specificity

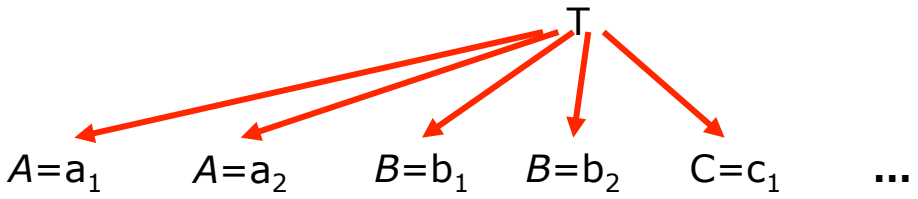
Universiteit Leiden

Subgroup Discovery as Search

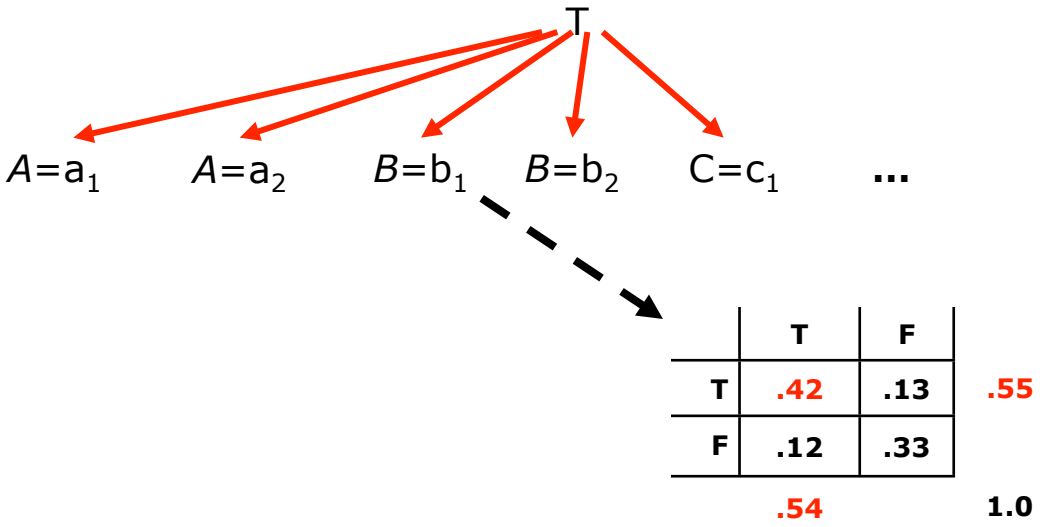
T

Universiteit Leiden

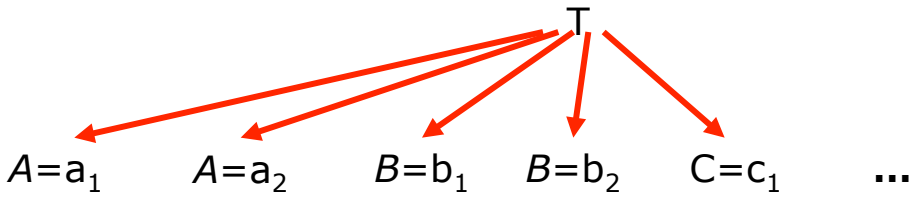
Subgroup Discovery as Search



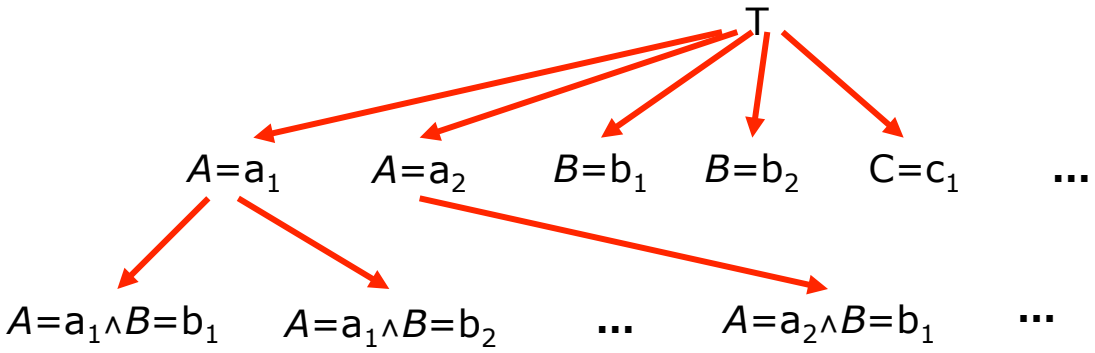
Subgroup Discovery as Search



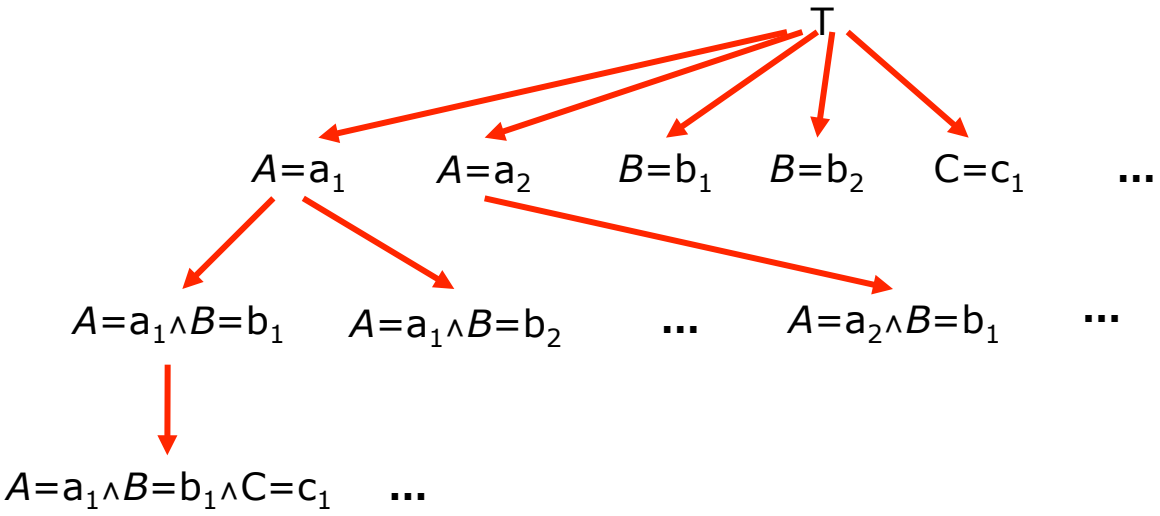
Subgroup Discovery as Search



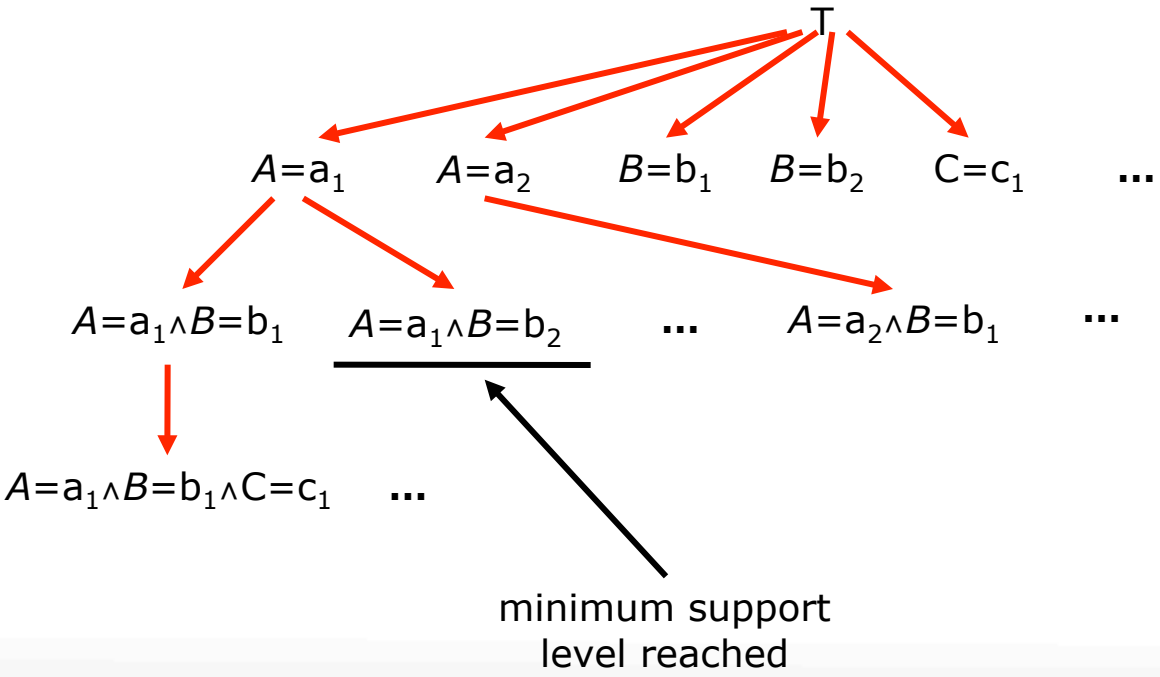
Subgroup Discovery as Search



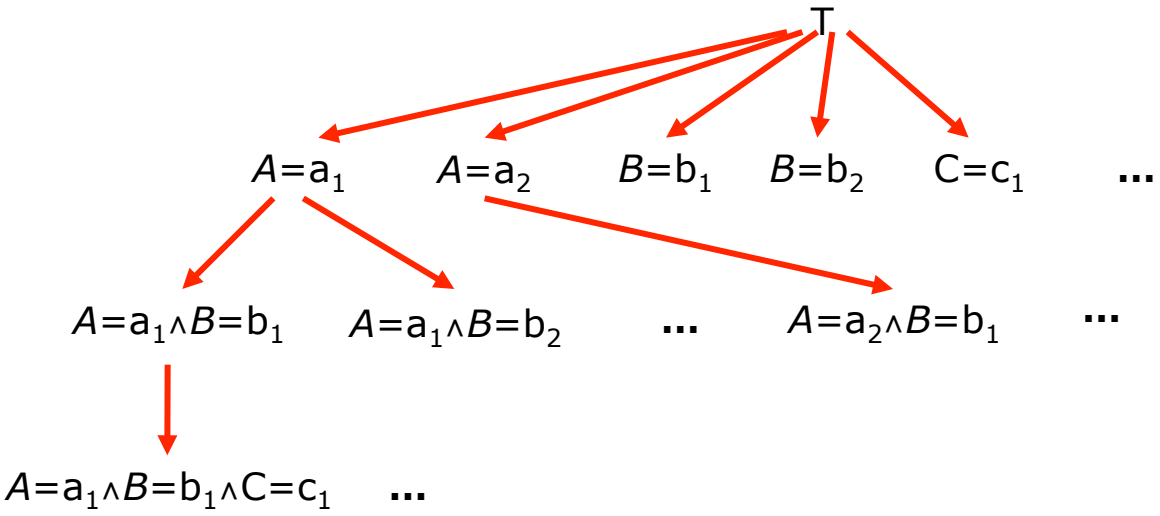
Subgroup Discovery as Search



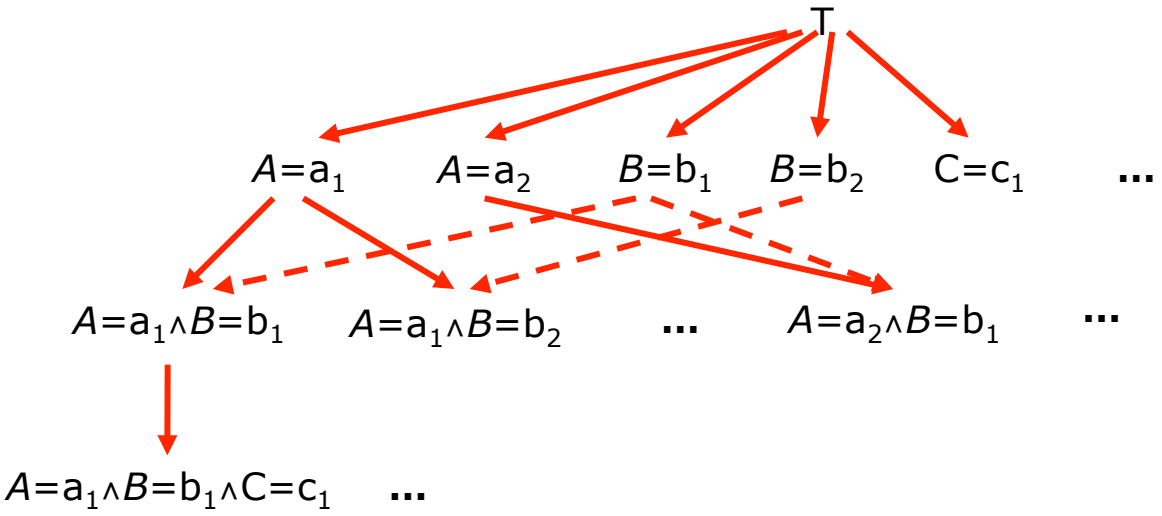
Subgroup Discovery as Search



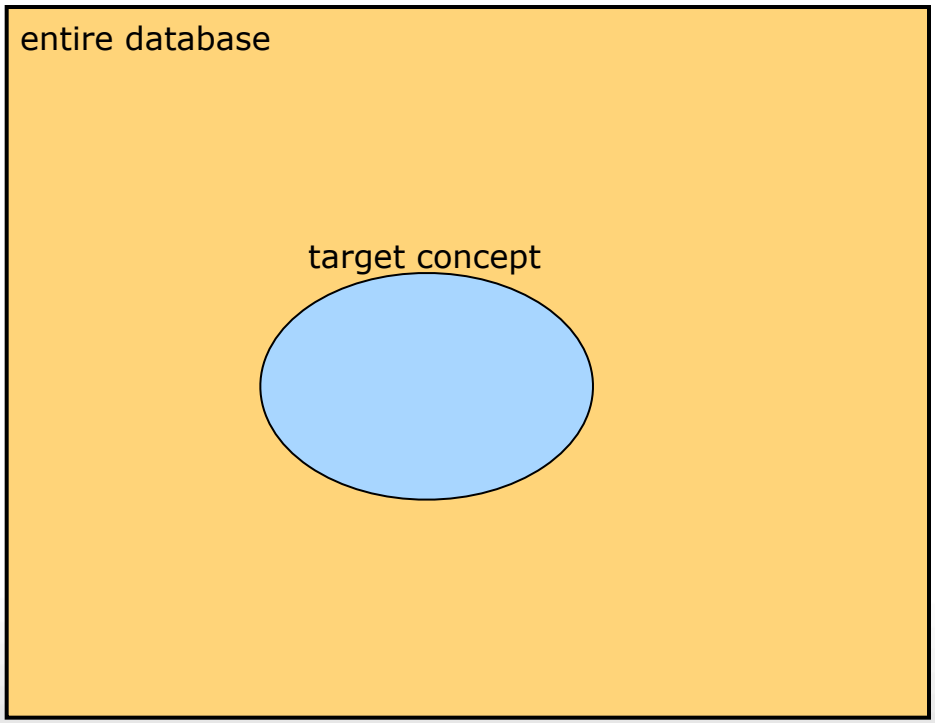
Subgroup Discovery as Search



Subgroup Discovery as Search

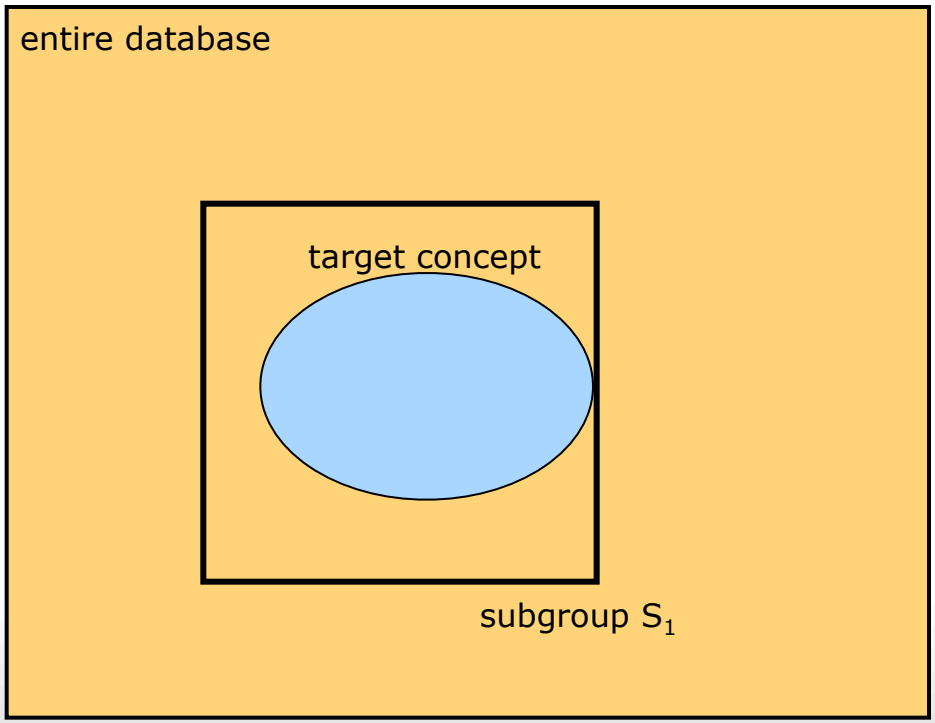


Refinements are (anti-)monotonic

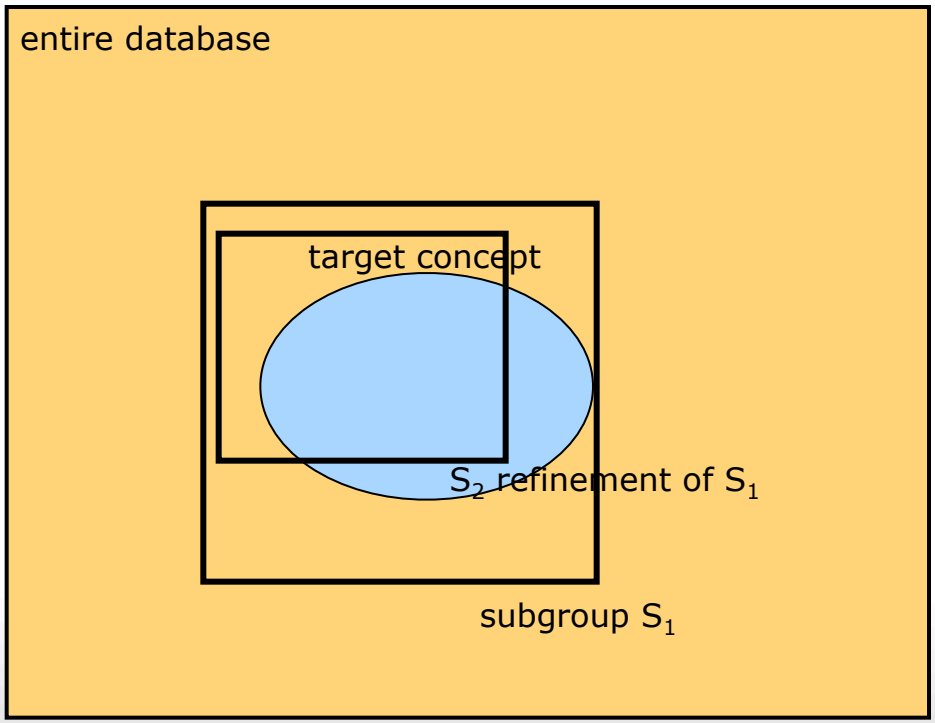


Universiteit Leiden

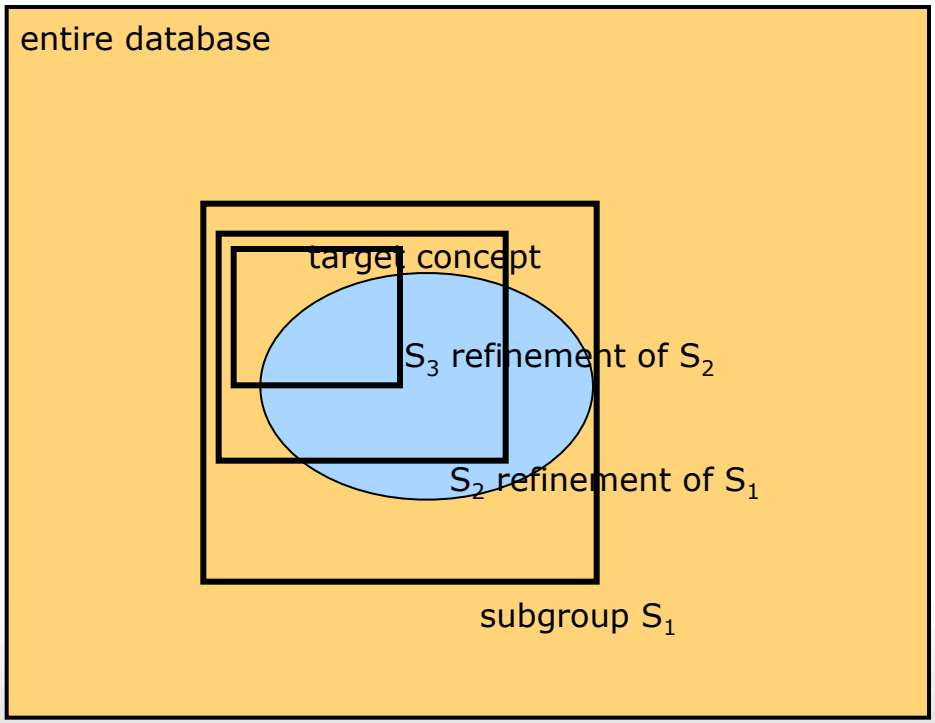
Refinements are (anti-)monotonic



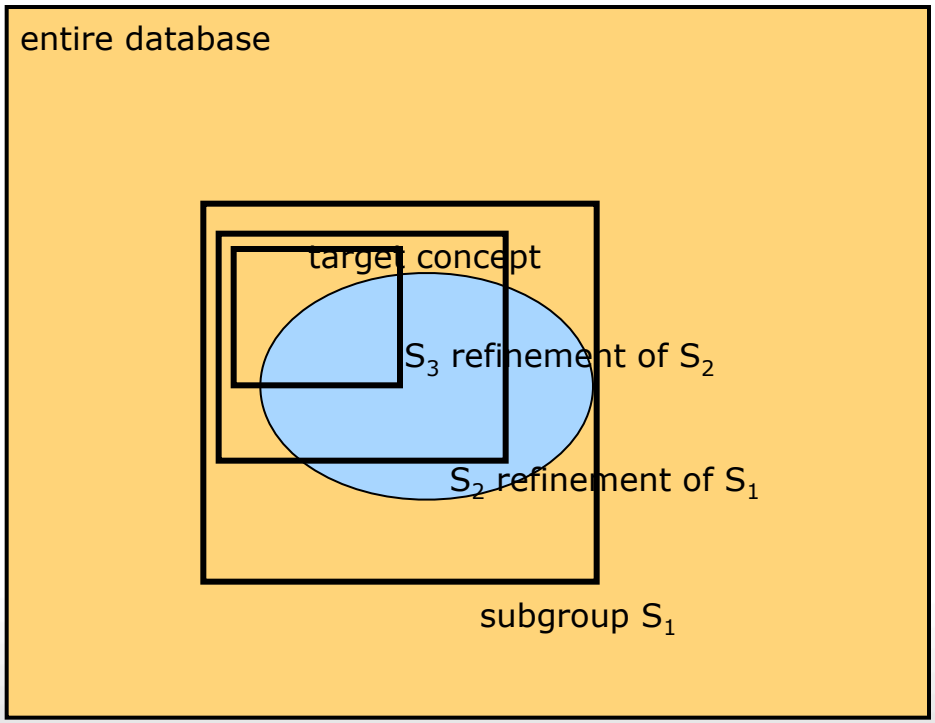
Refinements are (anti-)monotonic



Refinements are (anti-)monotonic

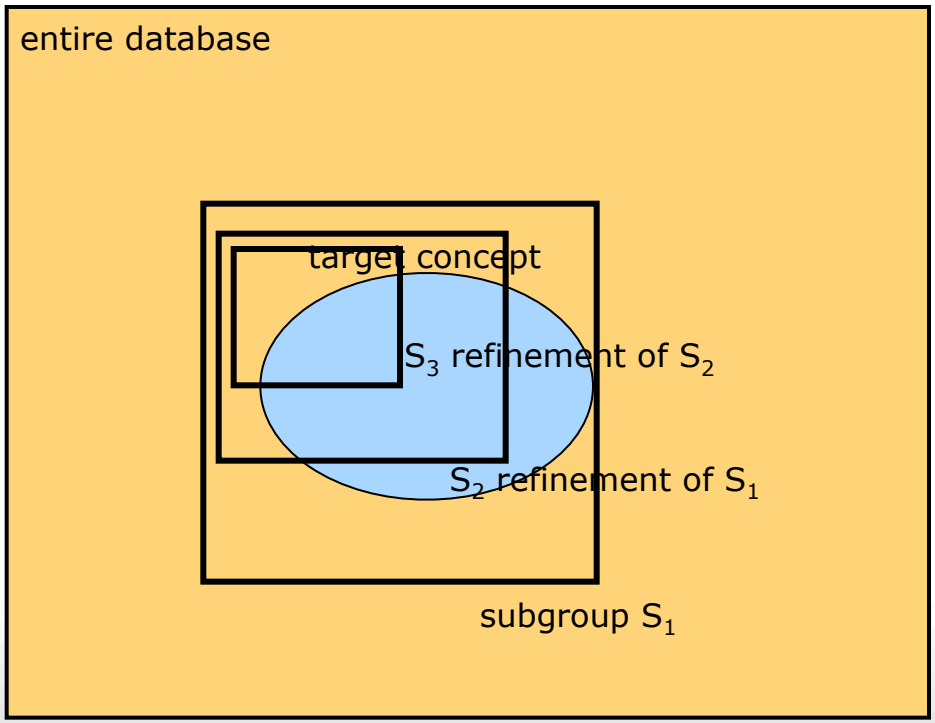


Refinements are (anti-)monotonic



Refinements are (anti-) monotonic in their support…

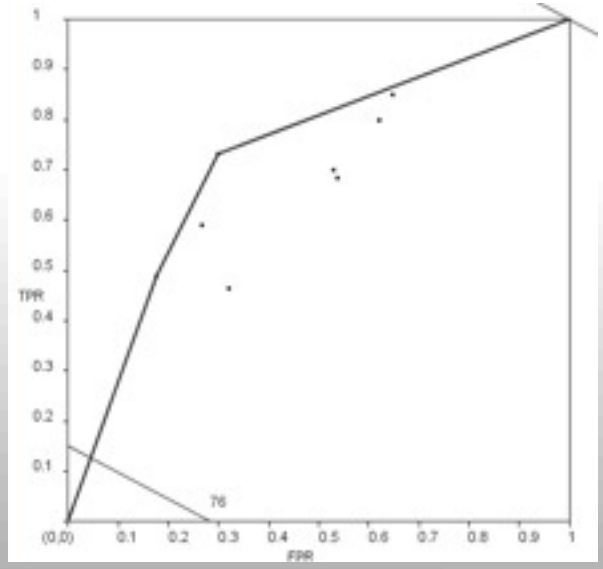
Refinements are (anti-)monotonic



Refinements are (anti-) monotonic in their support…

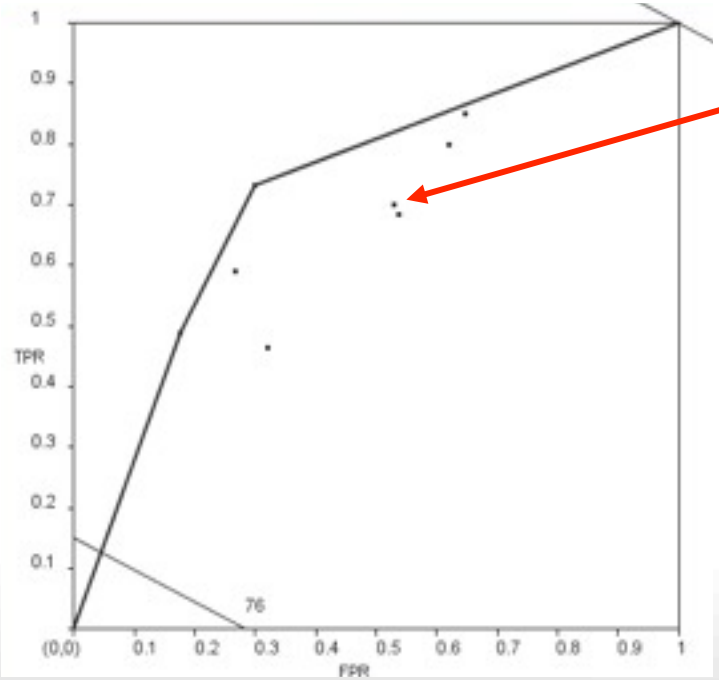
…but not in interestingness. This may go up or down.

Subgroup Discovery and ROC space



ROC Space

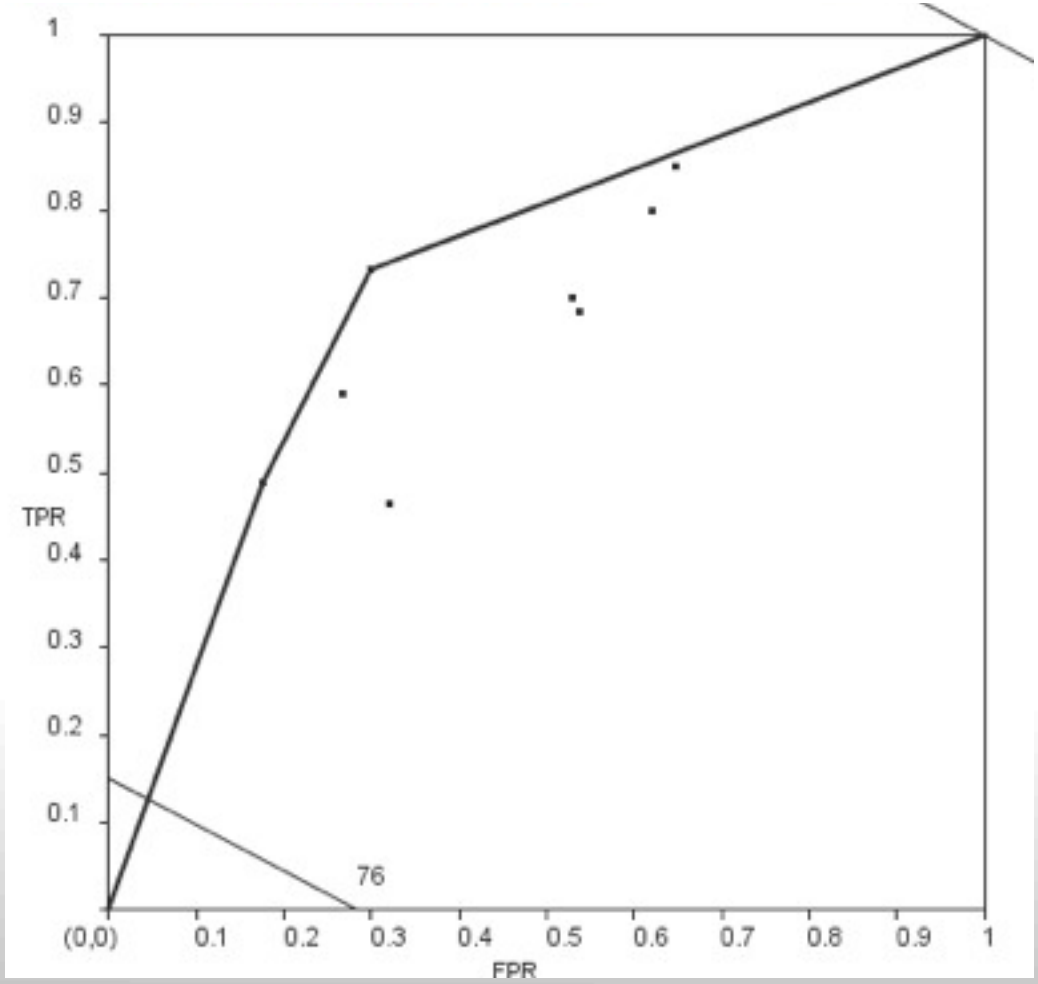
ROC = Receiver Operating Characteristics



Each subgroup forms a point in ROC space, in terms of its False Positive Rate, and True Positive Rate.

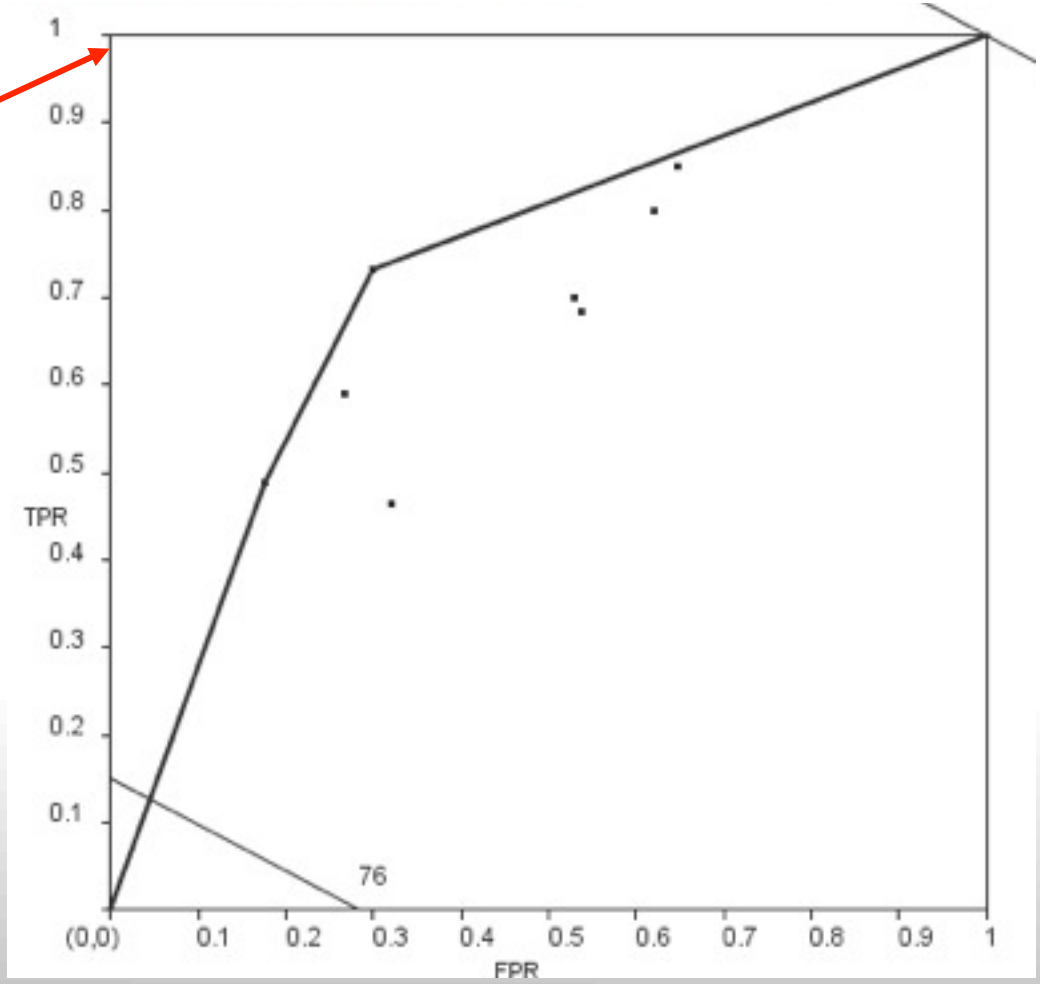
TPR = TP/Pos = TP/TP+FN (fraction of positive cases in the subgroup)
FPR = FP/Neg = FP/FP+TN (fraction of negative cases in the subgroup)

ROC Space Properties

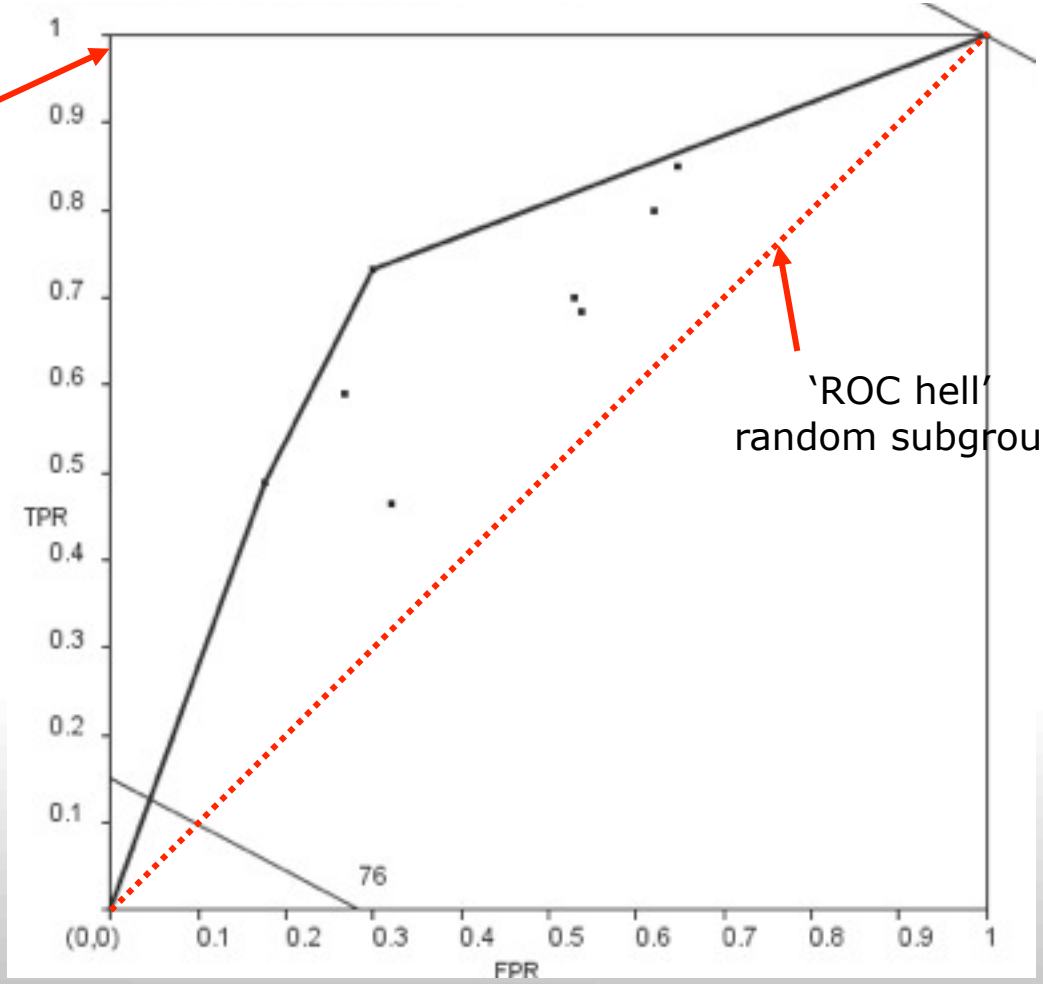


ROC Space Properties

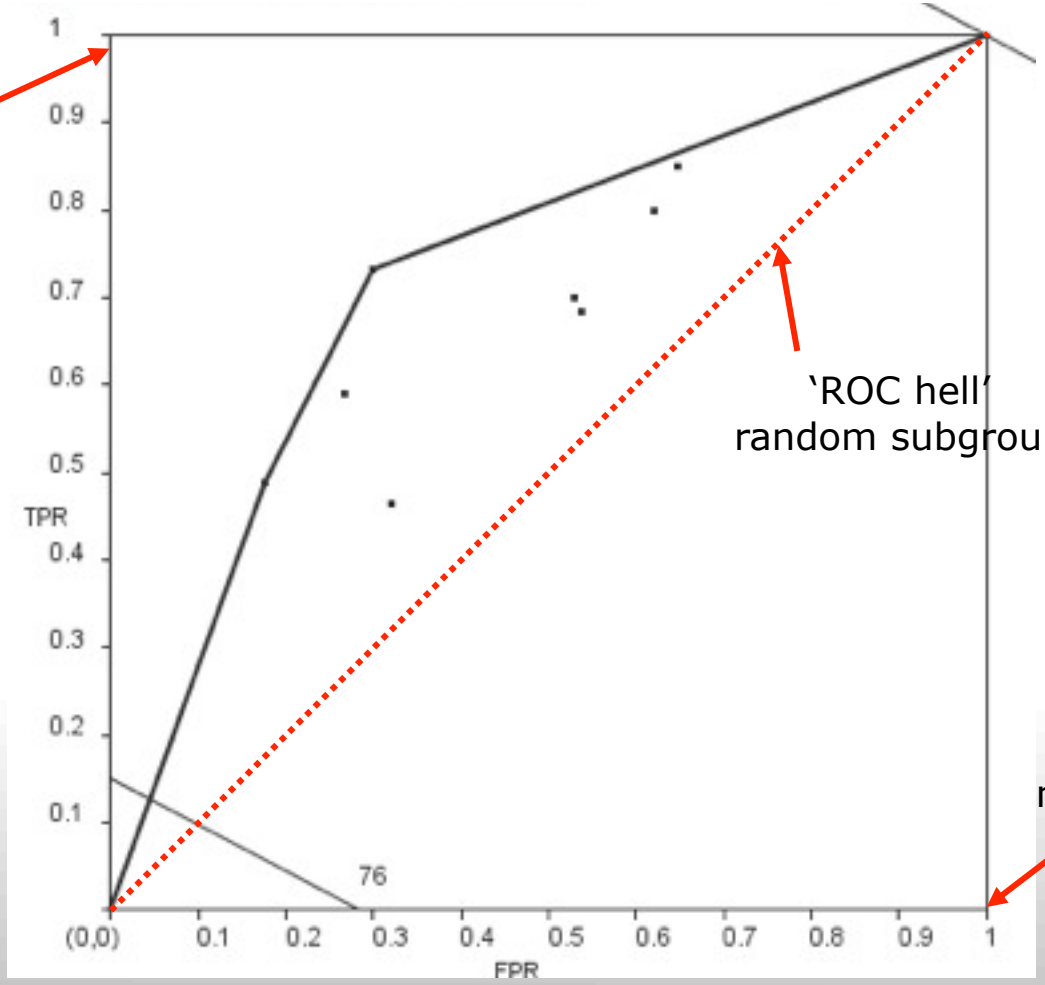
'ROC heaven'
perfect subgroup



ROC Space Properties



ROC Space Properties

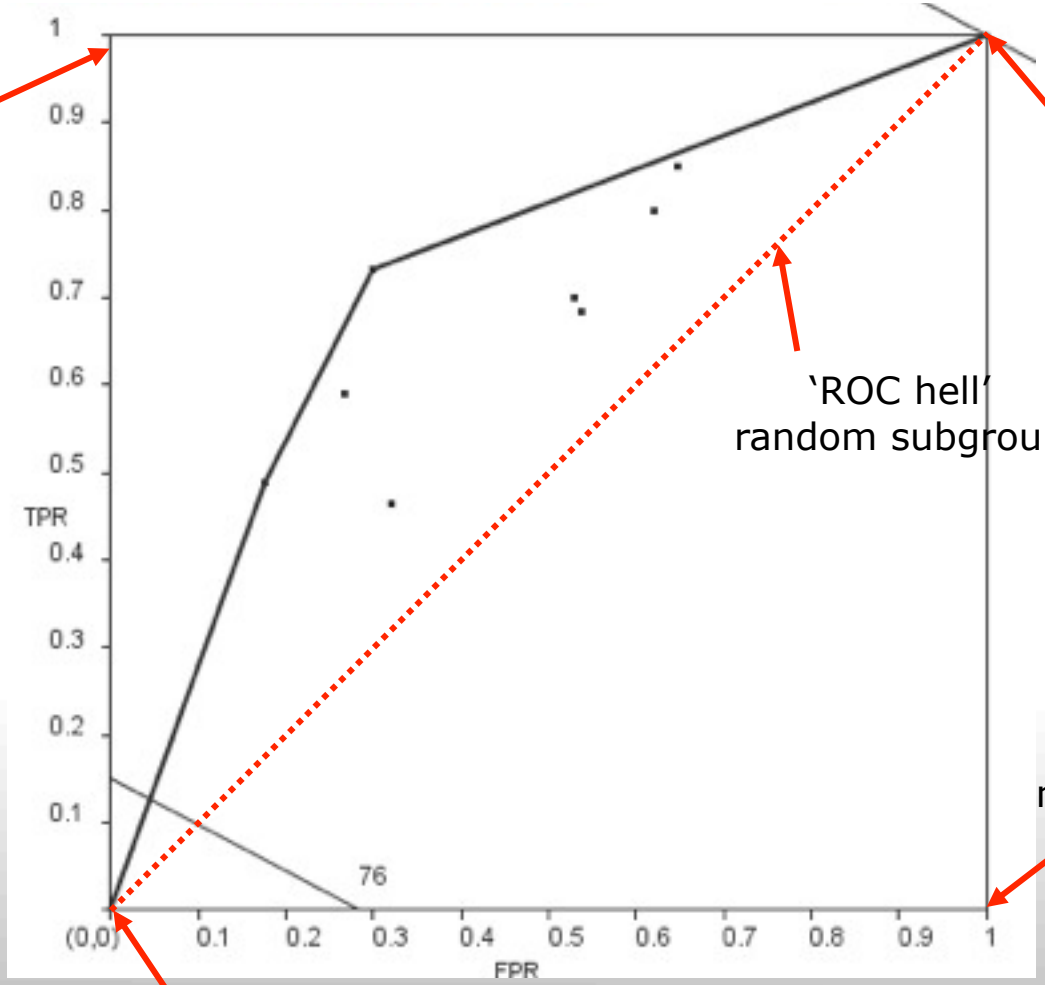


'ROC heaven'
perfect subgroup

'ROC hell'
random subgroup

perfect
negative subgroup

ROC Space Properties



'ROC heaven'
perfect subgroup

entire database

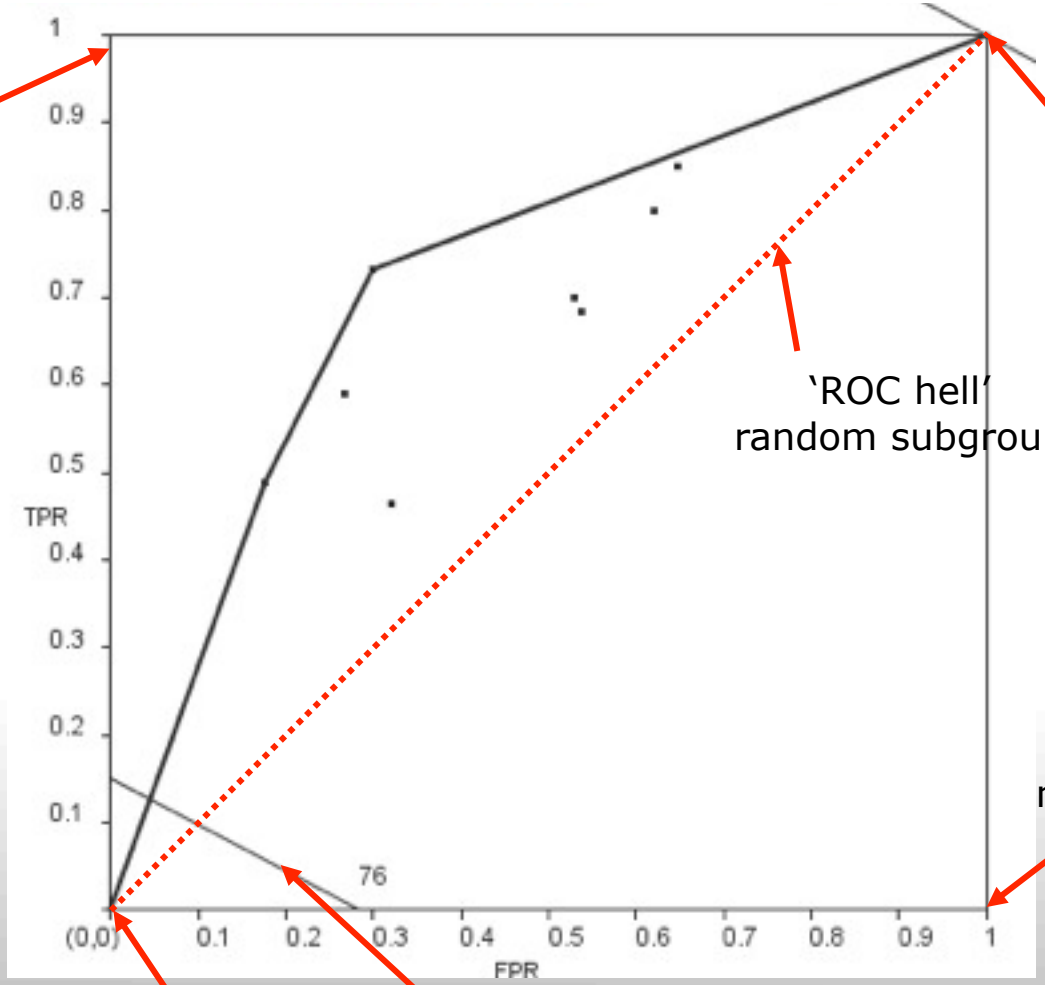
'ROC hell'
random subgroup

perfect
negative subgroup

empty
subgroup

Universiteit Leiden

ROC Space Properties



'ROC heaven'
perfect subgroup

entire database

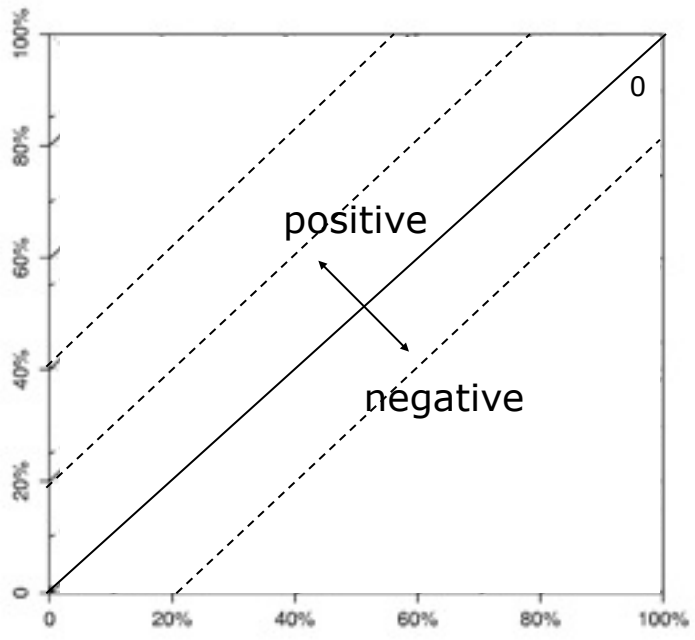
'ROC hell'
random subgroup

perfect
negative subgroup

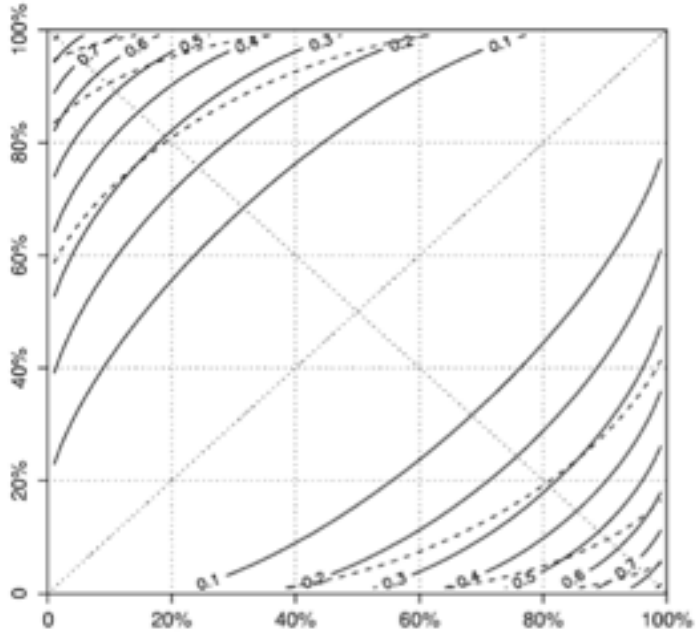
empty
subgroup

minimum support
threshold

Measures in ROC Space



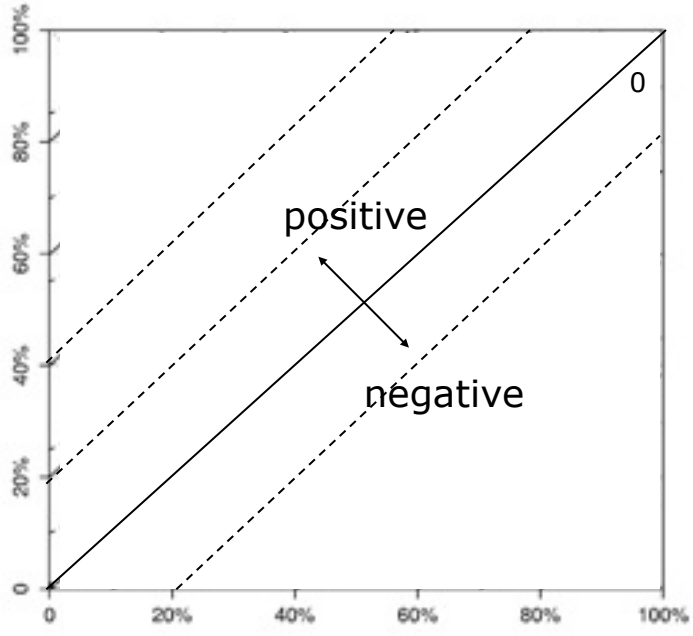
WRAcc



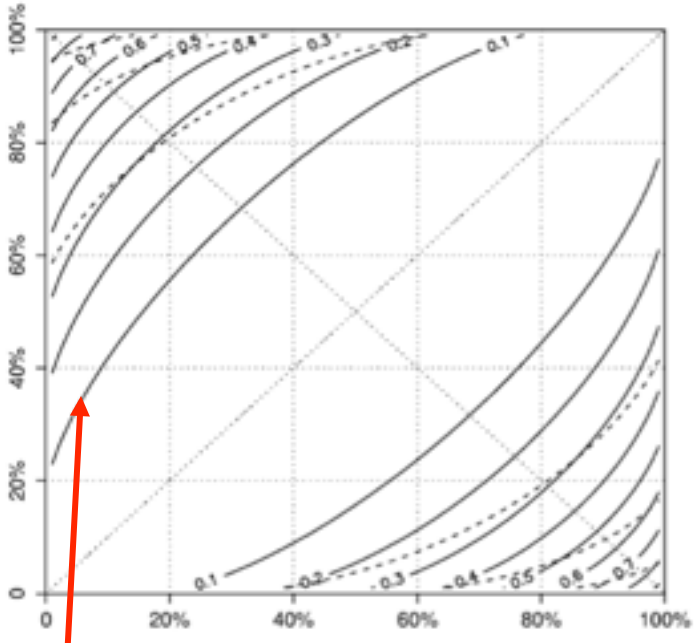
Information Gain

source: Flach & Fürnkranz

Measures in ROC Space



WRAcc



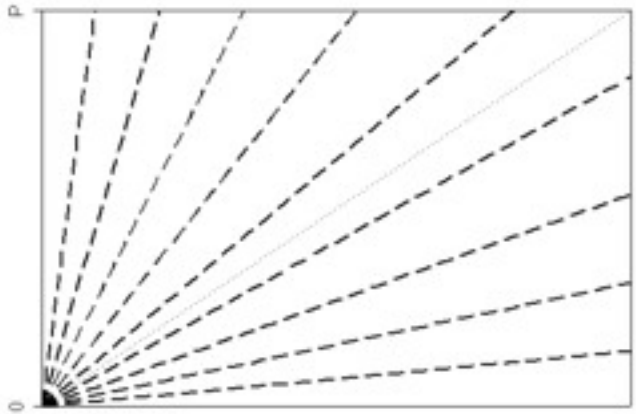
Information Gain

source: Flach & Fürnkranz

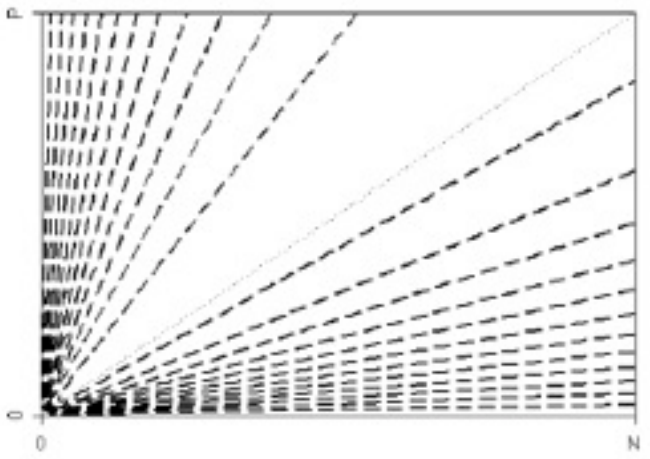
isometric

Other Measures

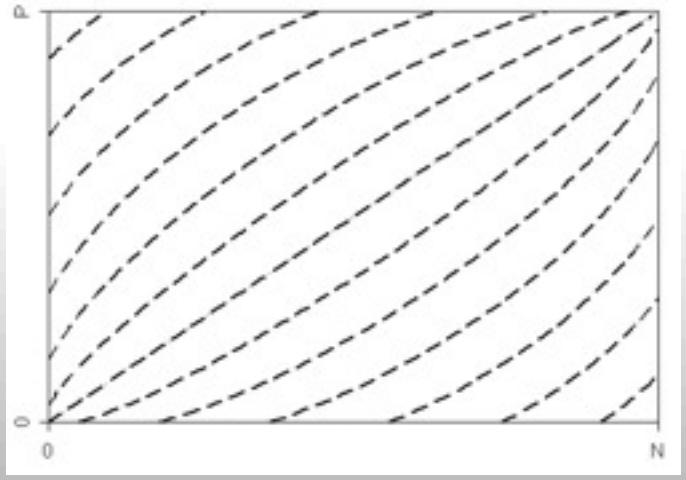
Precision



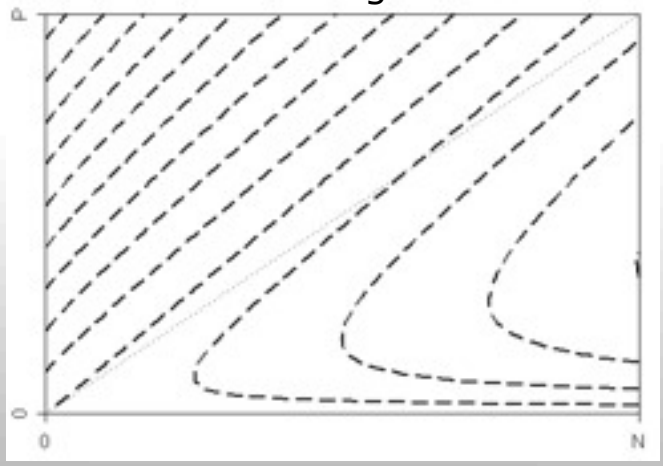
Gini index



Correlation coefficient

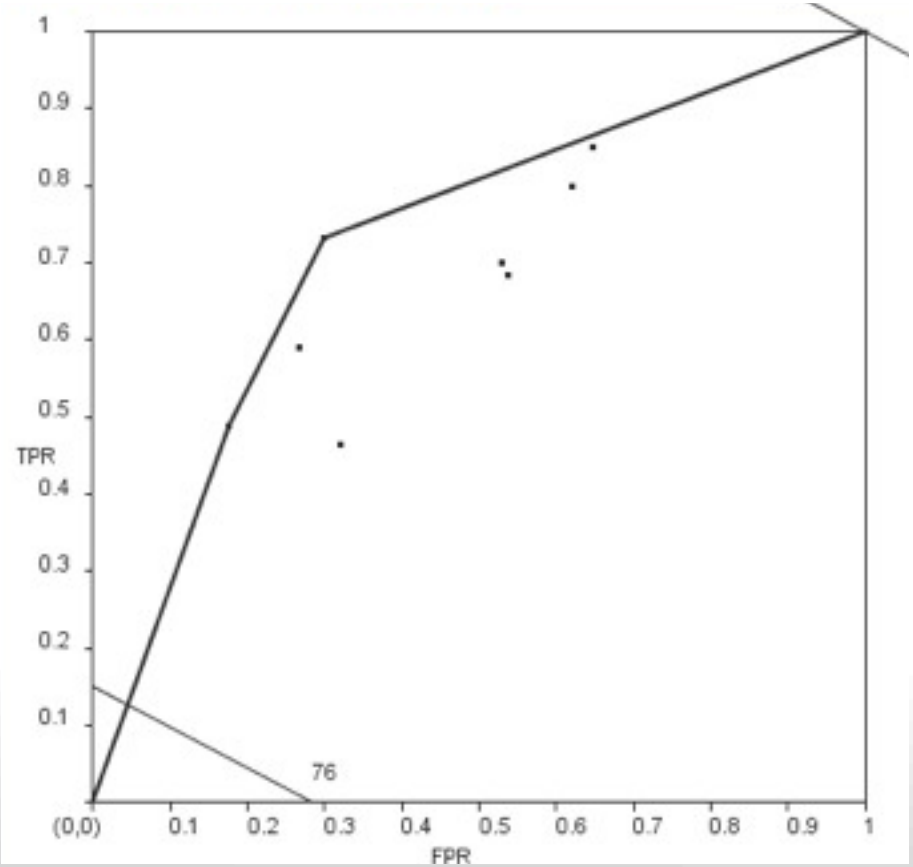


Foil gain



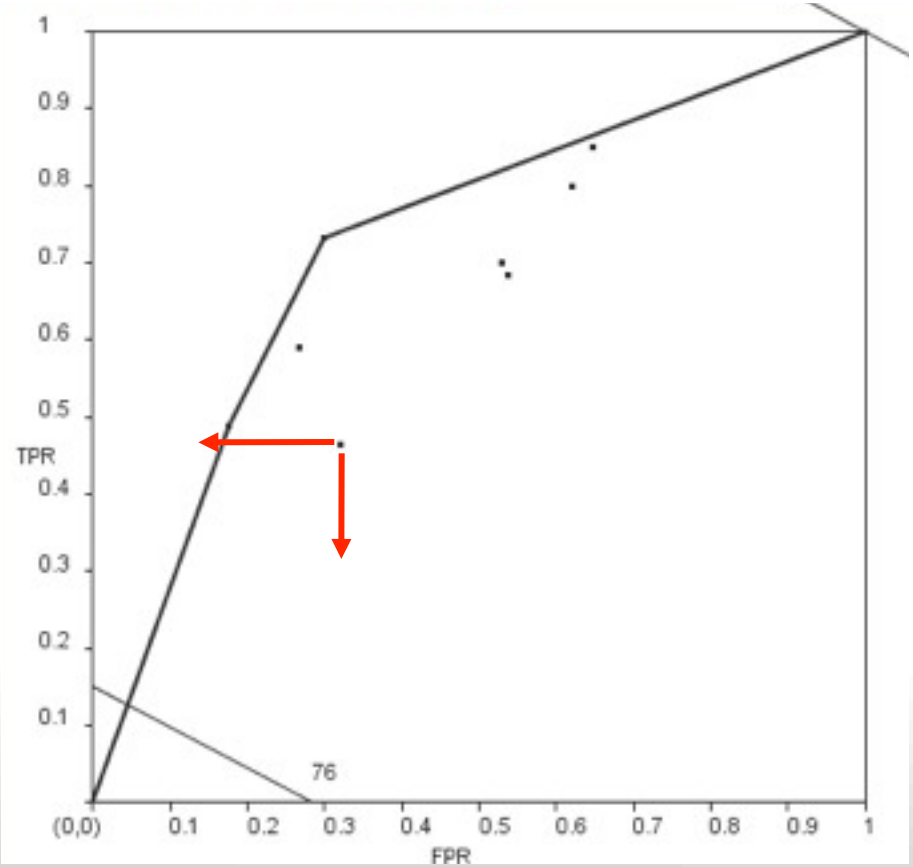
Universiteit Leiden

Refinements in ROC Space



Universiteit Leiden

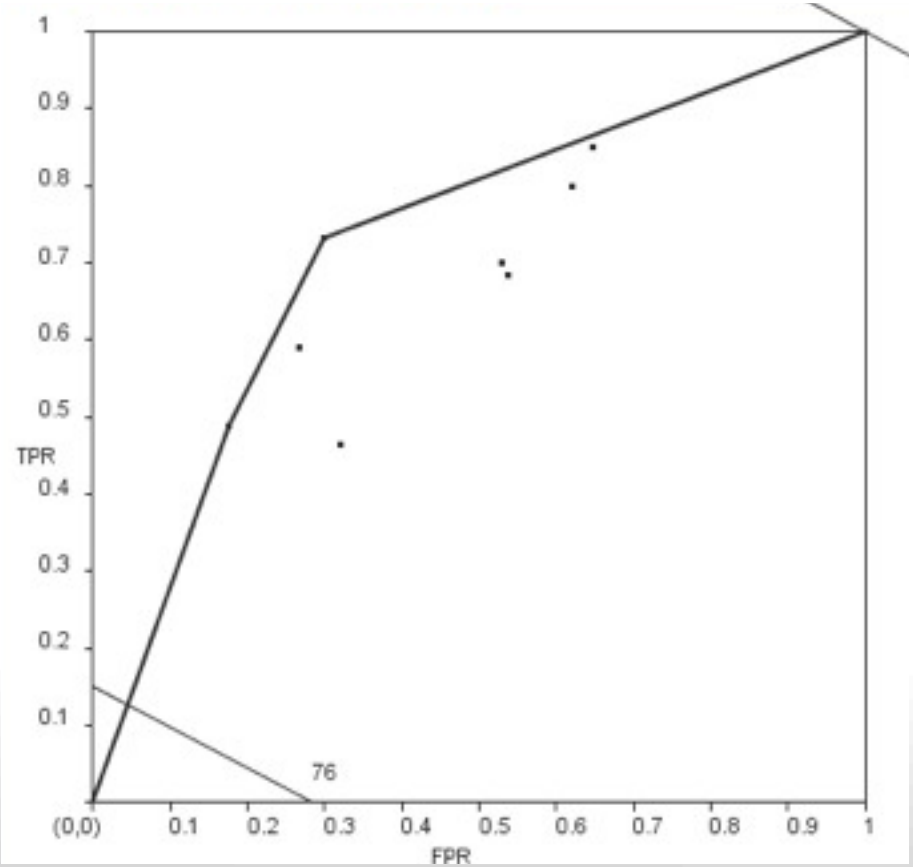
Refinements in ROC Space



Refinements of S will reduce the FPR and TPR, so will appear to the left and below S.

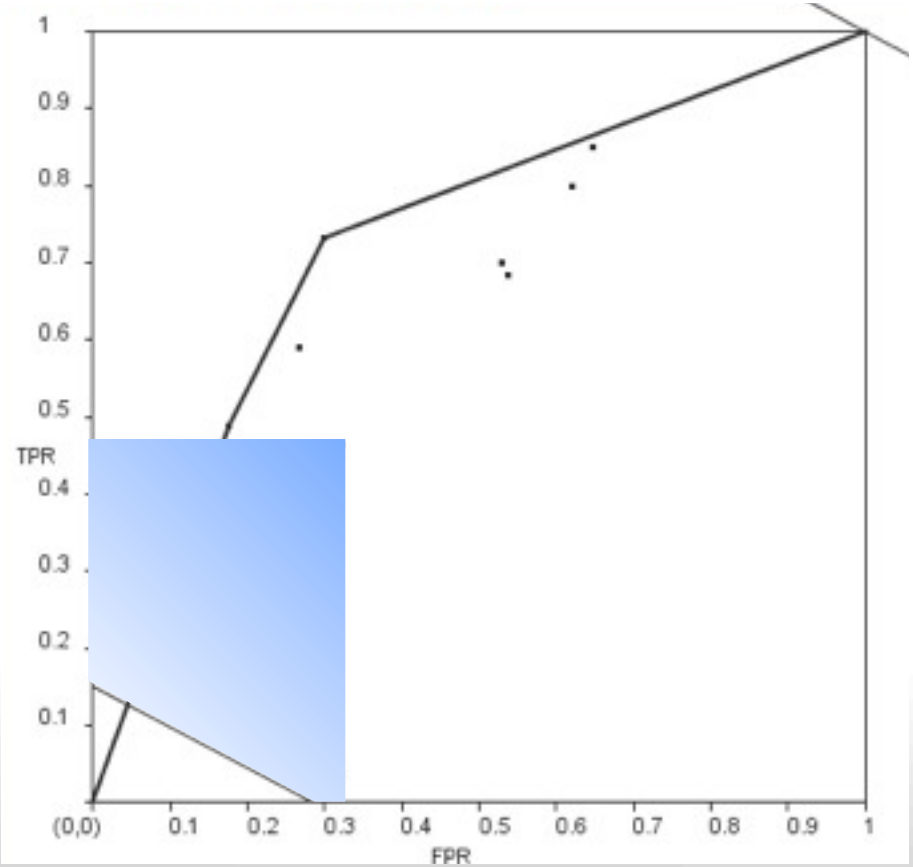
Universiteit Leiden

Refinements in ROC Space



Refinements of S will reduce the FPR and TPR, so will appear to the left and below S.

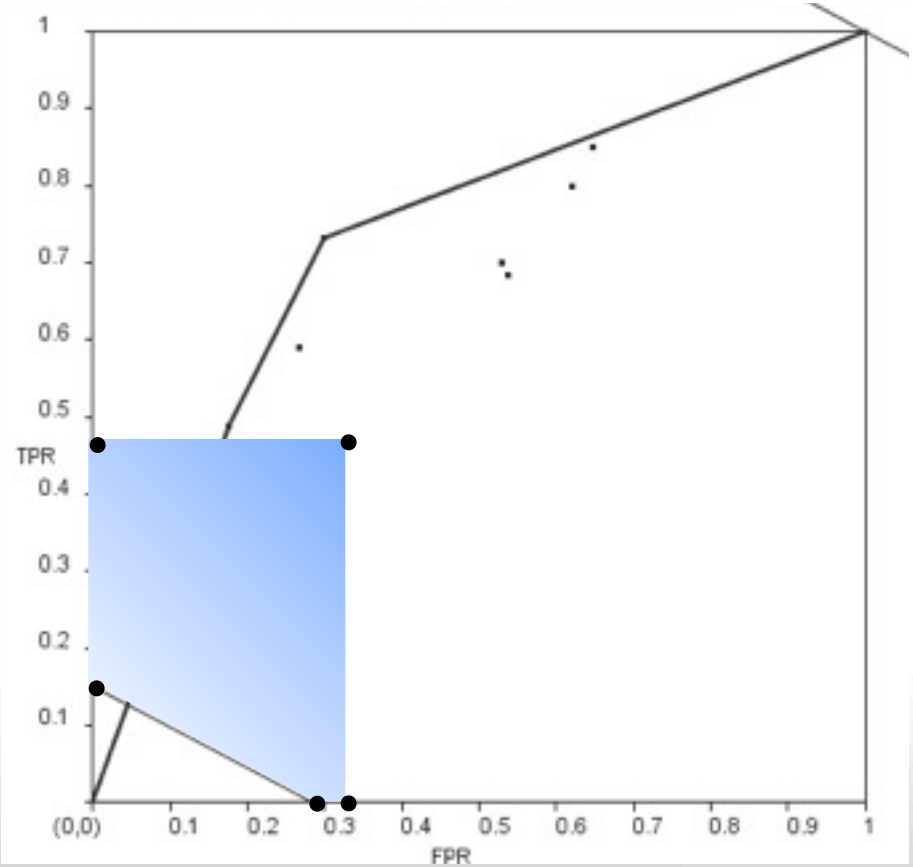
Refinements in ROC Space



Refinements of S will reduce the FPR and TPR, so will appear to the left and below S.

Blue polygon represents possible refinements of S. With a convex measure, *f* is bounded by measure of corners.

Refinements in ROC Space



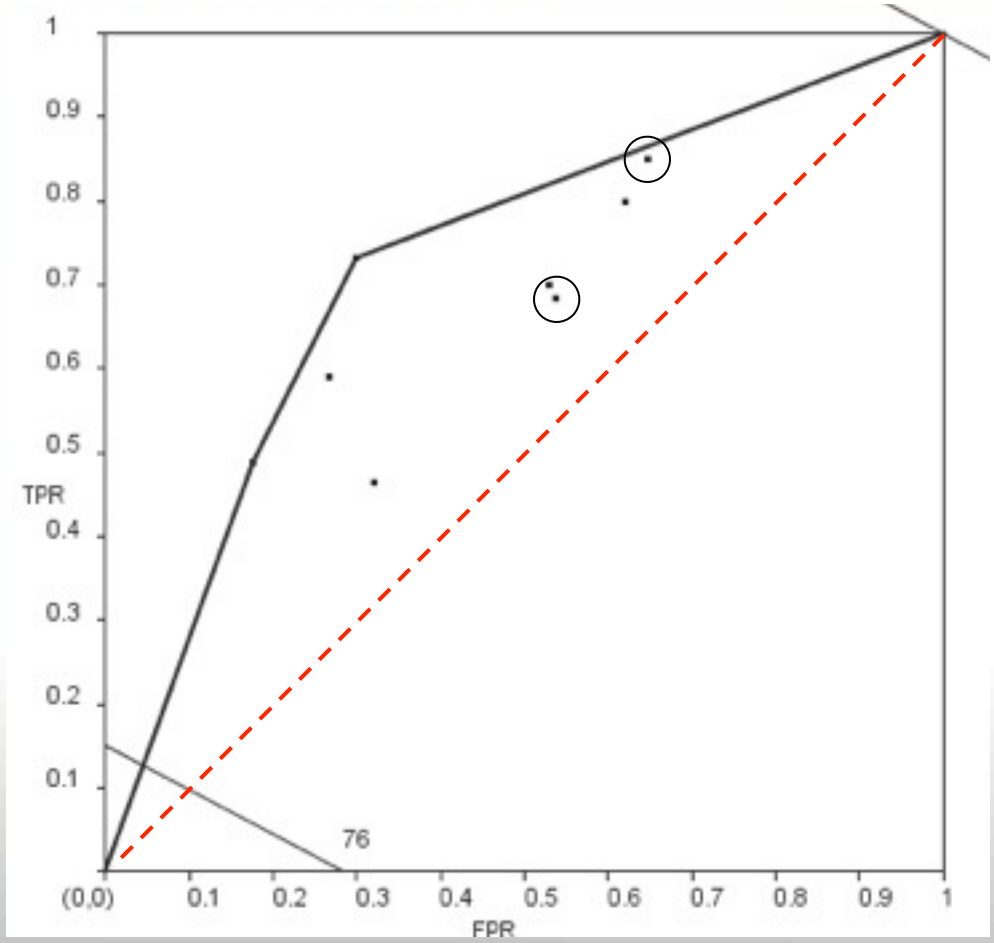
Refinements of S will reduce the FPR and TPR, so will appear to the left and below S.

Blue polygon represents possible refinements of S. With a convex measure, *f* is bounded by measure of corners.

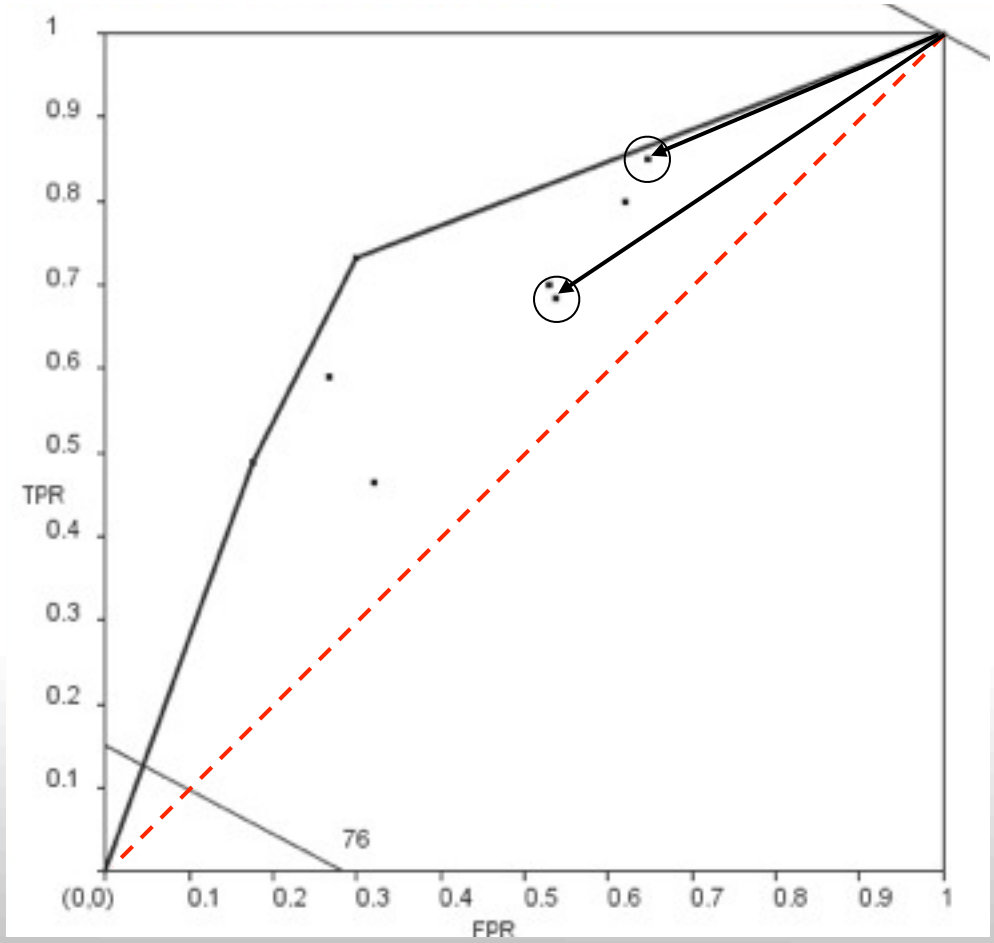
If corners are not above minimum quality or current best (top *k*?), prune search space below S.

Universiteit Leiden

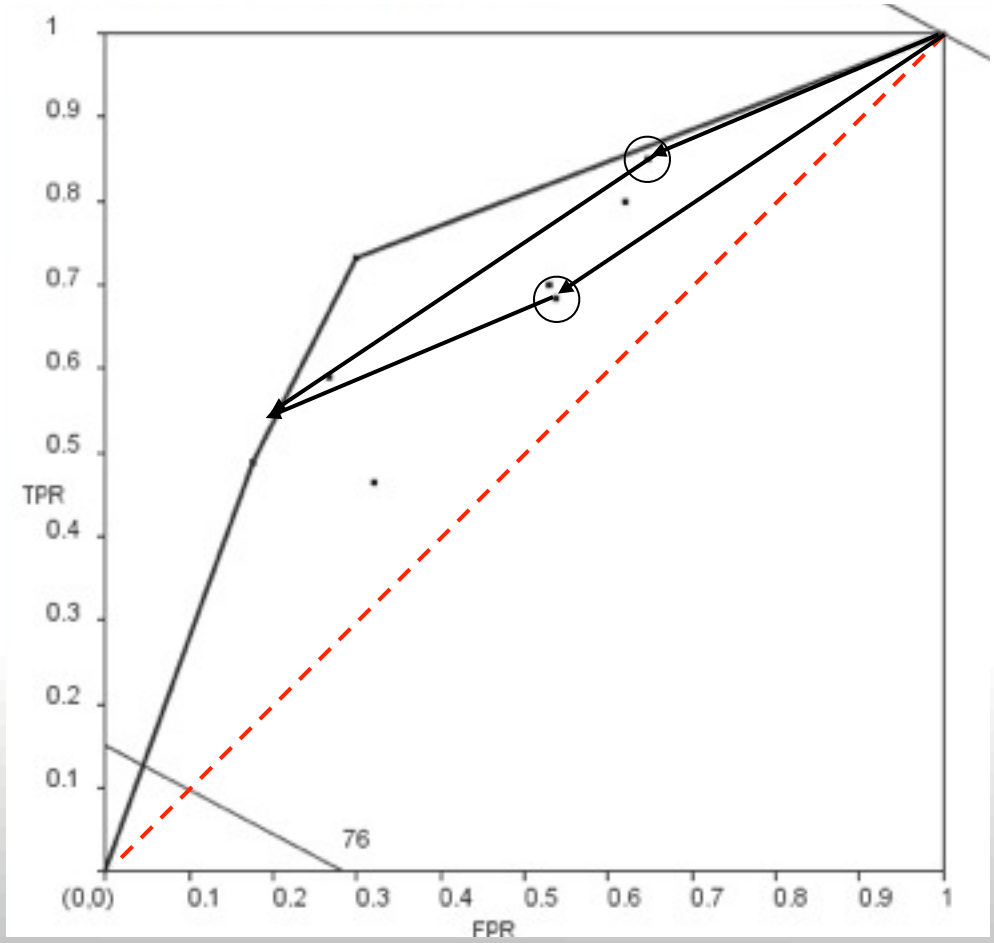
Combining Two Subgroups



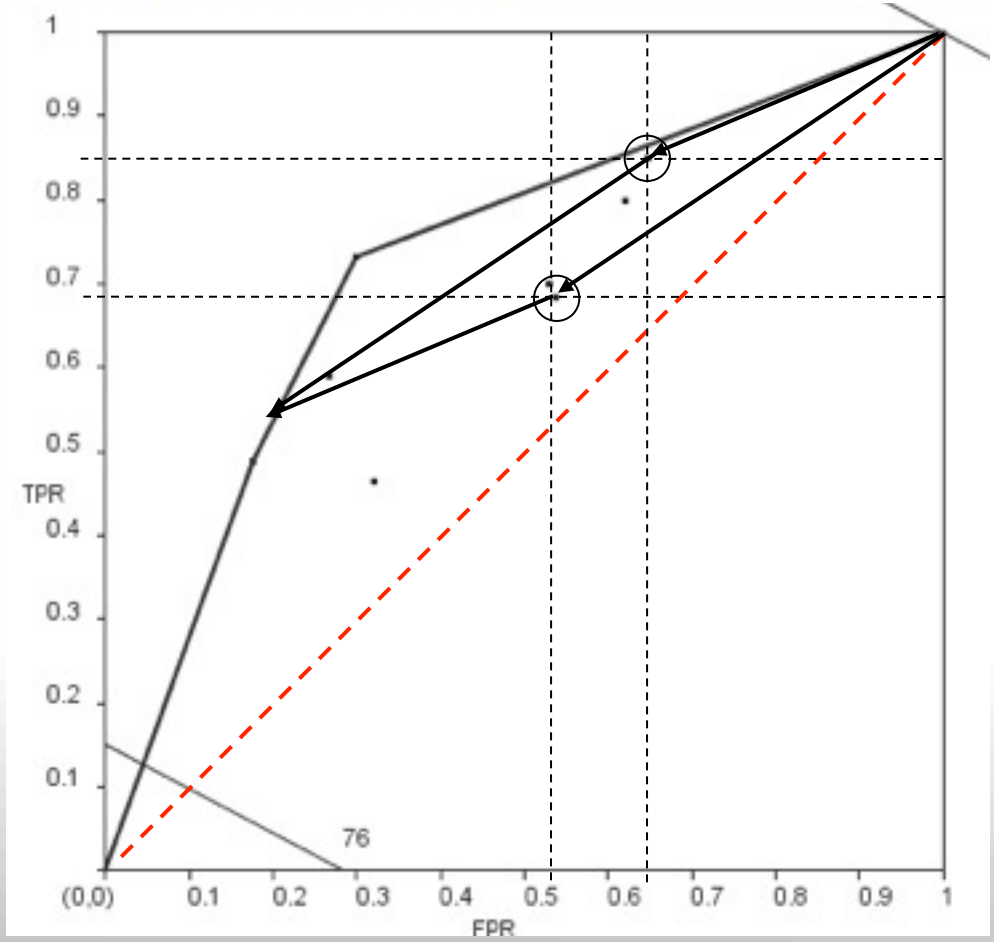
Combining Two Subgroups



Combining Two Subgroups

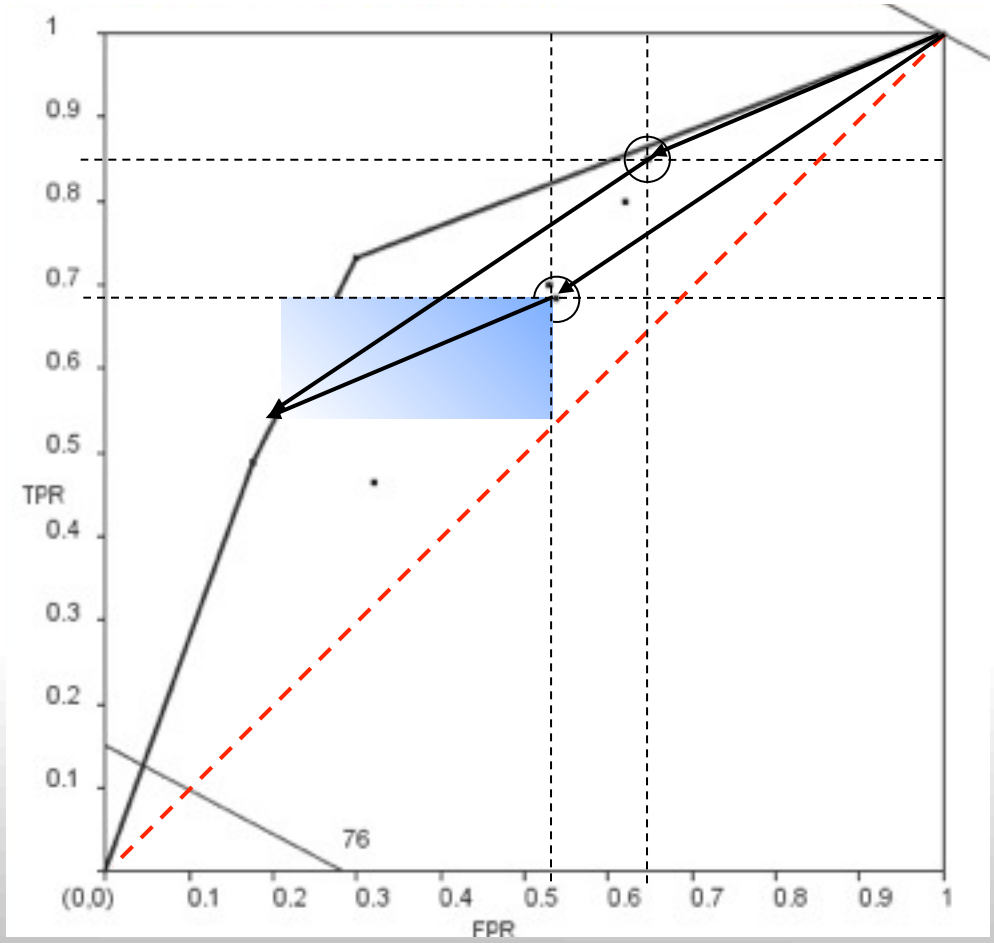


Combining Two Subgroups



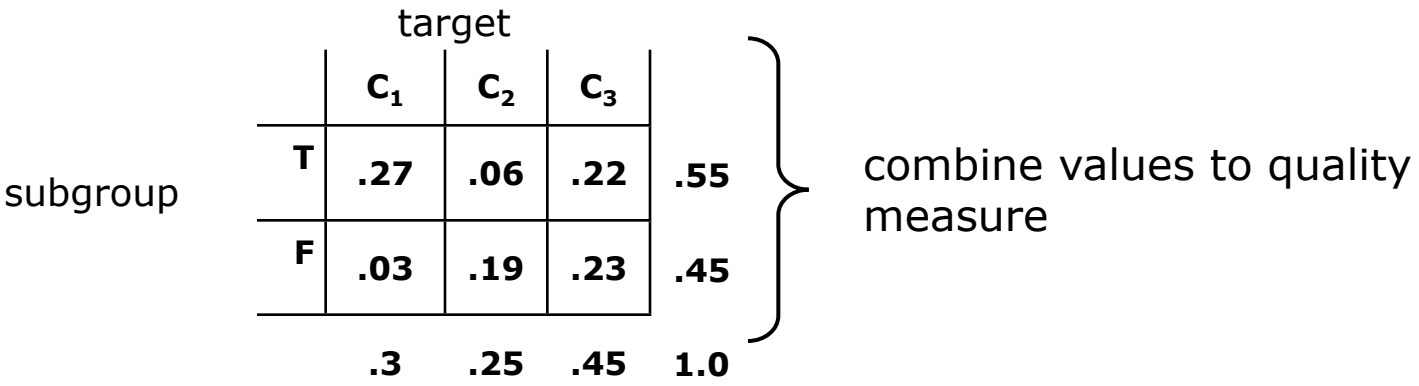
Universiteit Leiden

Combining Two Subgroups



Multi-class problems

- Generalising to problems with more than 2 classes is fairly staightforward:



Multi-class problems

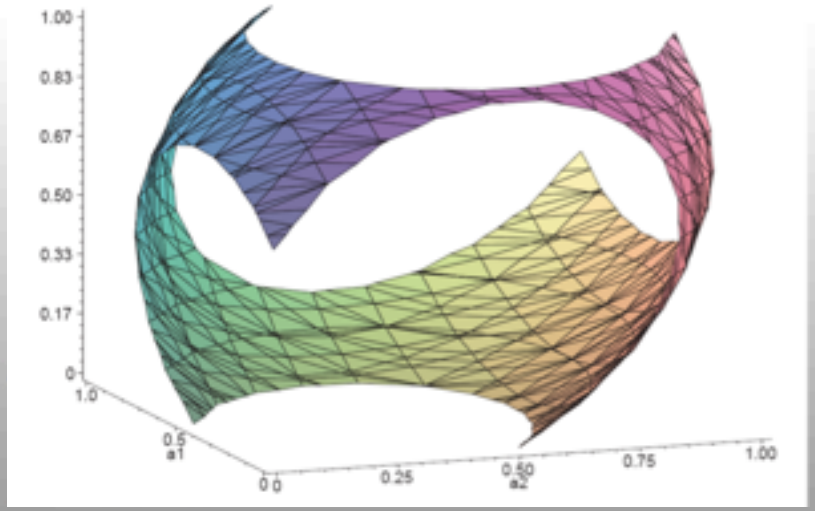
- Generalising to problems with more than 2 classes is fairly staightforward:

target

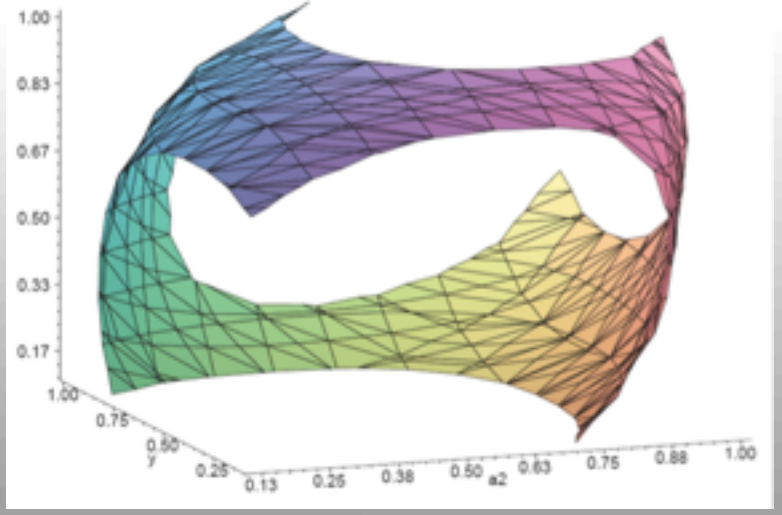
subgroup	C_1	C_2	C_3	
T	.27	.06	.22	.55
F	.03	.19	.23	.45
	.3	.25	.45	1.0

combine values to quality measure

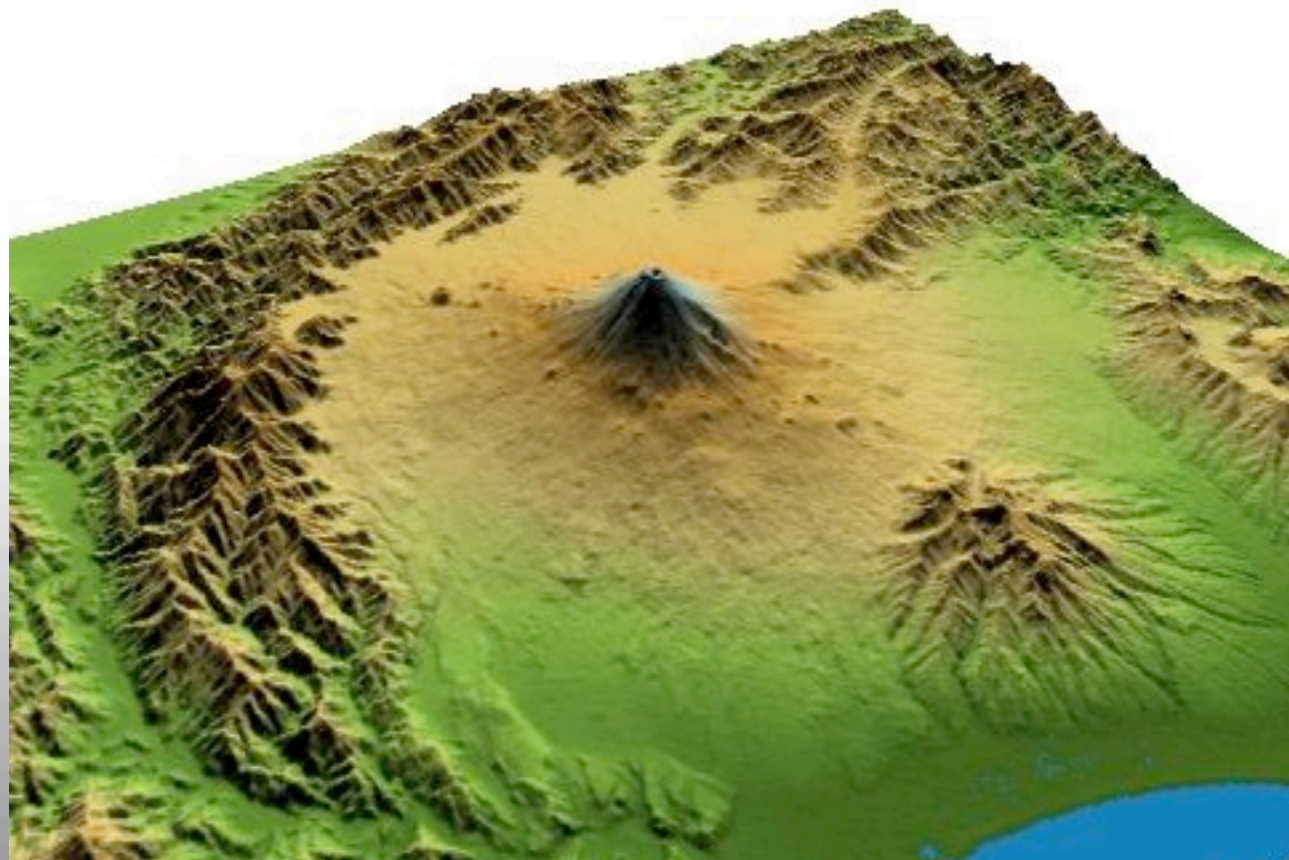
X^2



Information gain

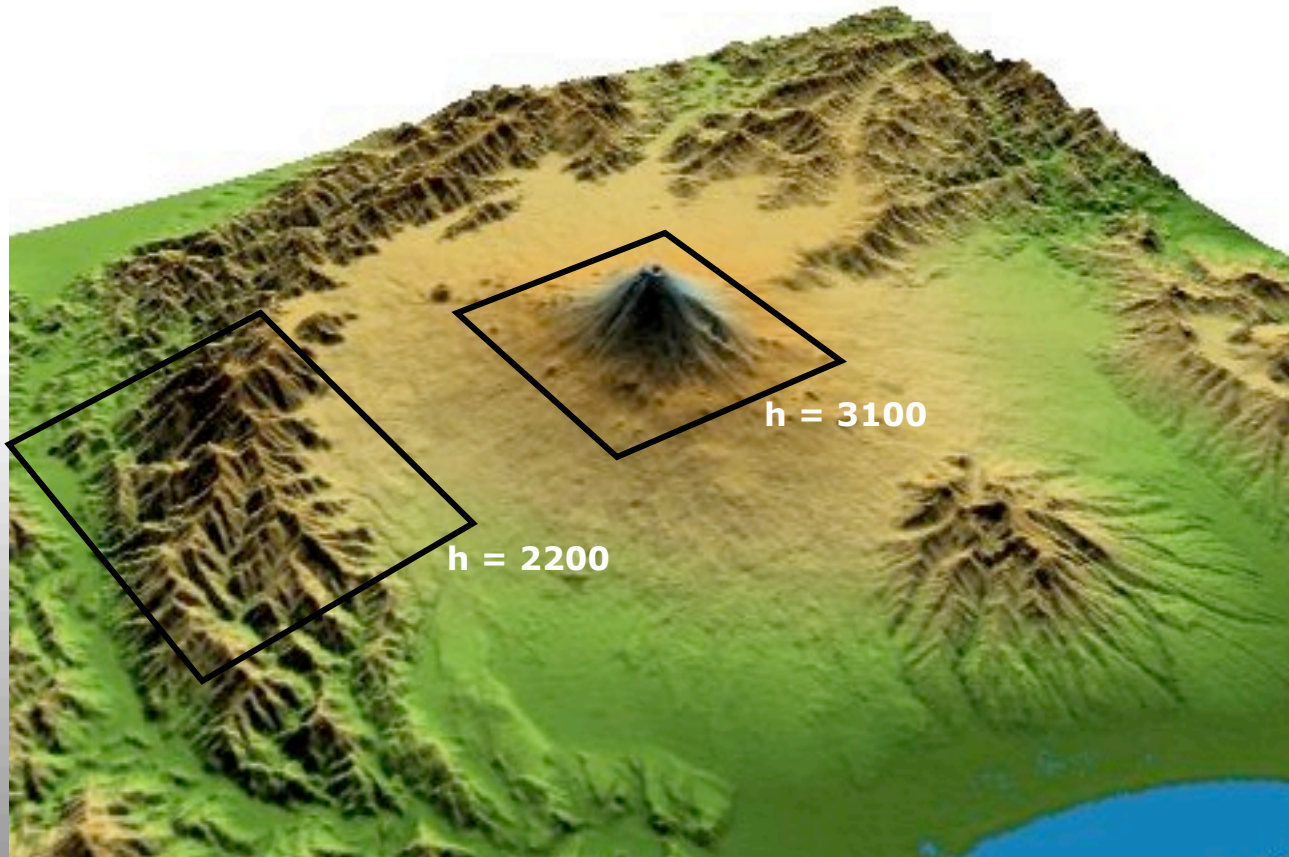


Numeric Subgroup Discovery



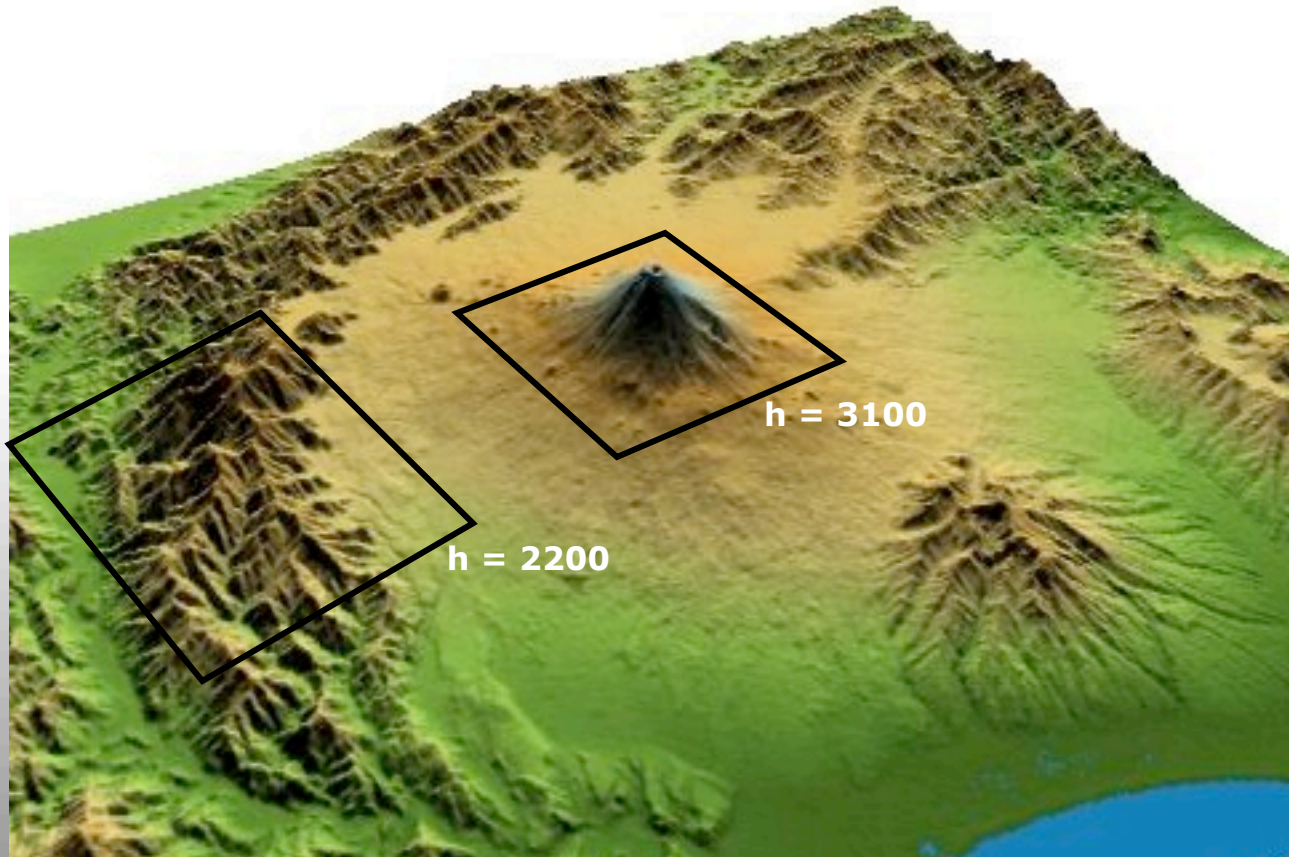
Universiteit Leiden

Numeric Subgroup Discovery



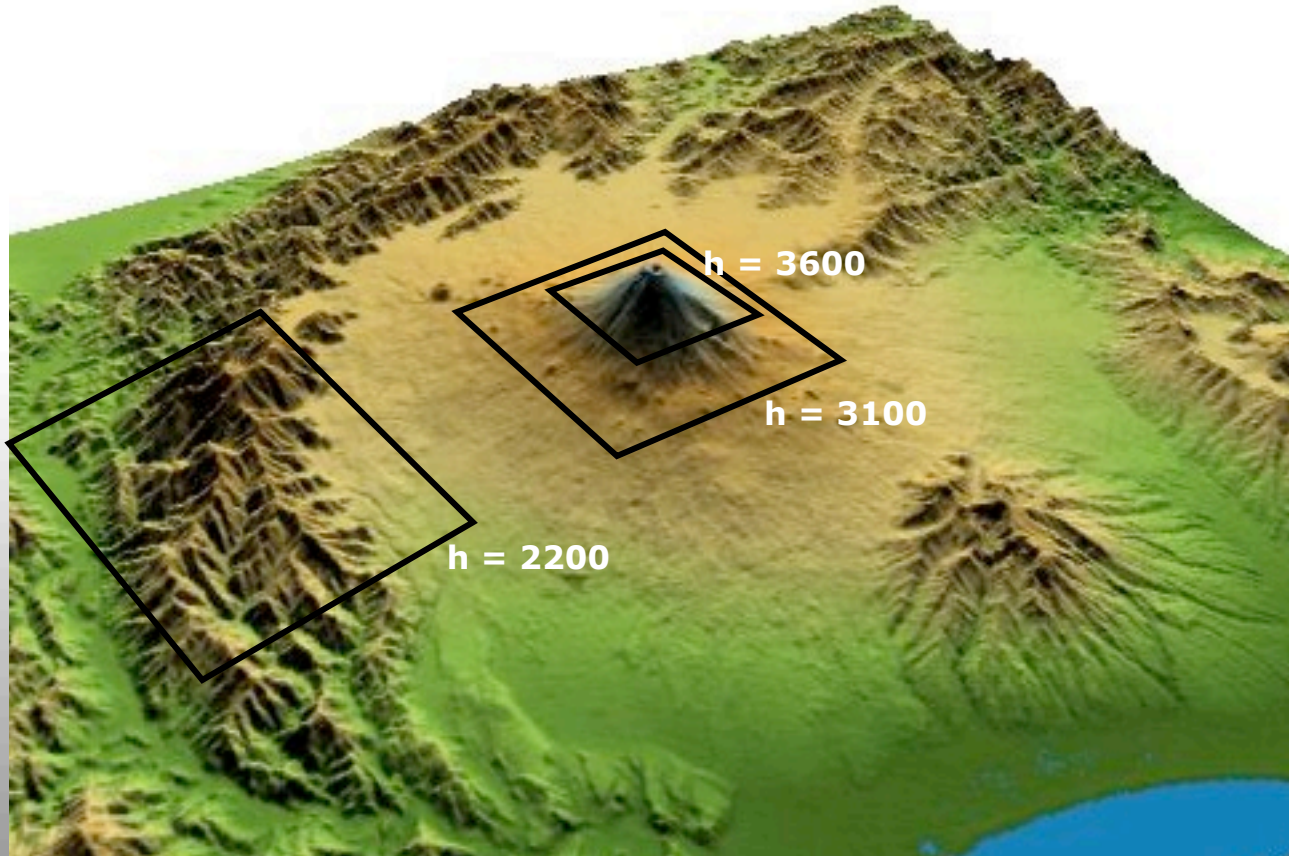
Numeric Subgroup Discovery

- Target is numeric: find subgroups with significantly higher or lower average value



Numeric Subgroup Discovery

- Target is numeric: find subgroups with significantly higher or lower average value
- Trade-off between size of subgroup and average target value



Universiteit Leiden

Quiz 1

Universiteit Leiden

Quiz 1

Q: Assume you have found a subgroup with a positive WRAcc (or infoGain). Can any refinement of this subgroup be negative?

Quiz 1

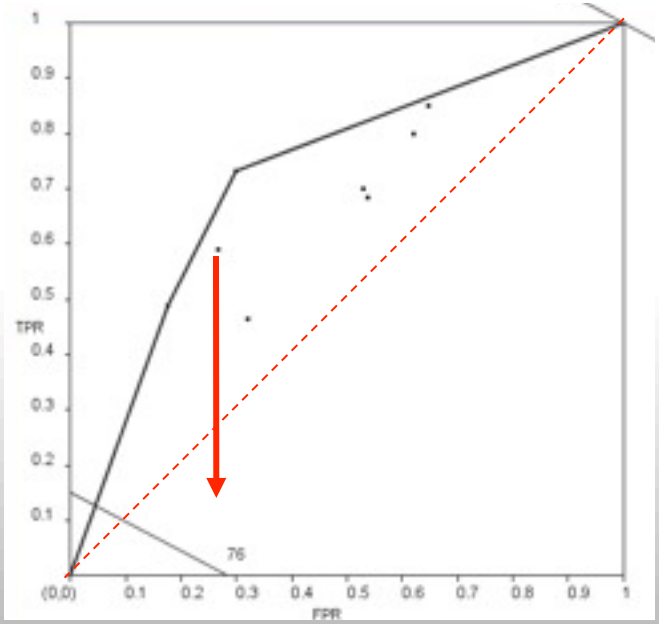
Q: Assume you have found a subgroup with a positive WRAcc (or infoGain). Can any refinement of this subgroup be negative?

A: Yes.

Quiz 1

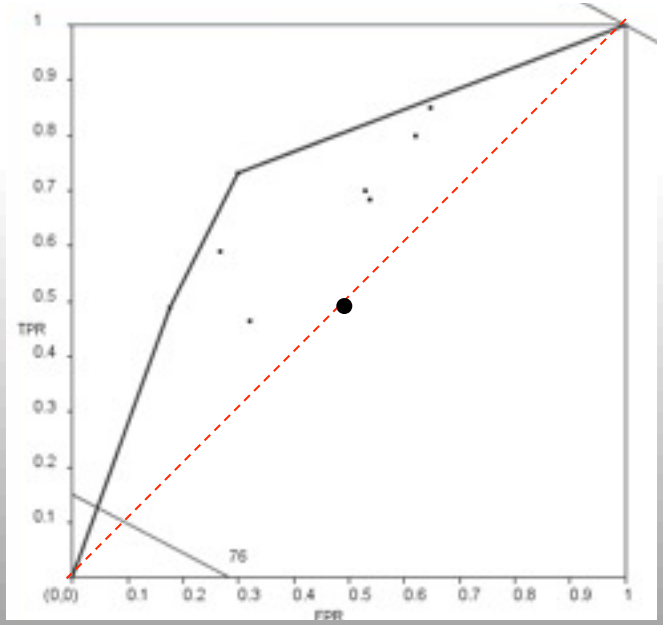
Q: Assume you have found a subgroup with a positive WRAcc (or infoGain). Can any refinement of this subgroup be negative?

A: Yes.



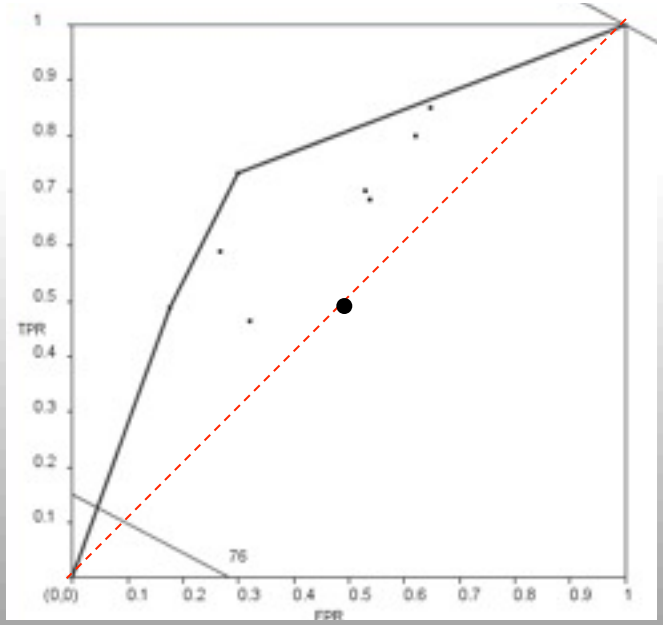
Universiteit Leiden

Quiz 2



Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

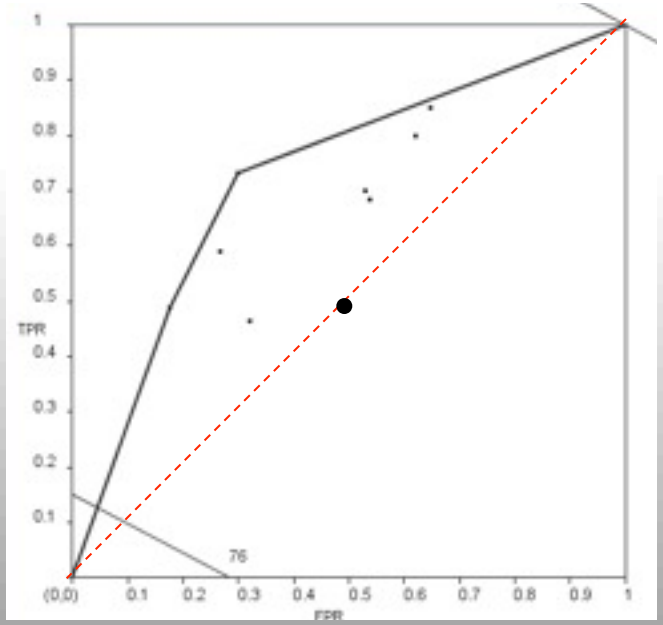


Universiteit Leiden

Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

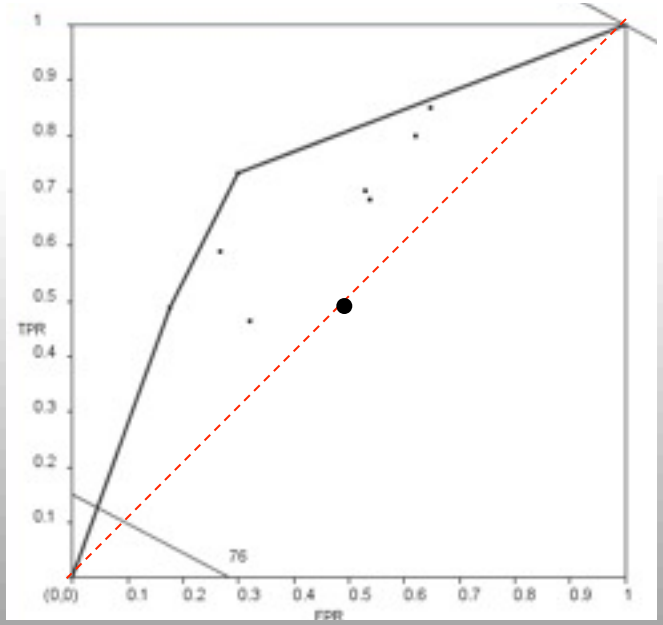


Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

Think of the XOR problem. A ∧ B is either completely positive or negative.

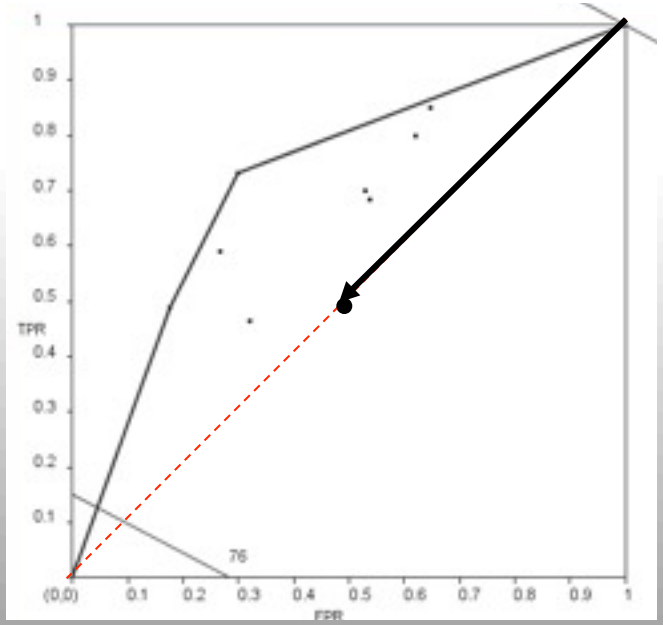


Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

Think of the XOR problem. A ∧ B is either completely positive or negative.

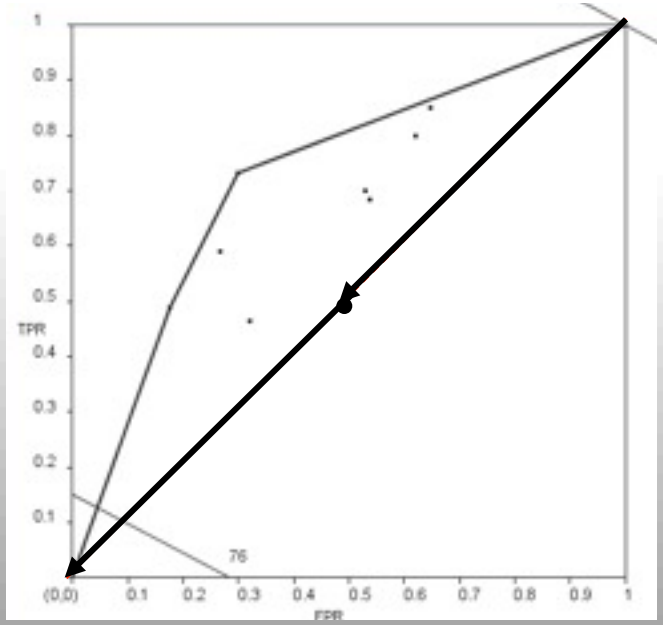


Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

Think of the XOR problem. A ∧ B is either completely positive or negative.

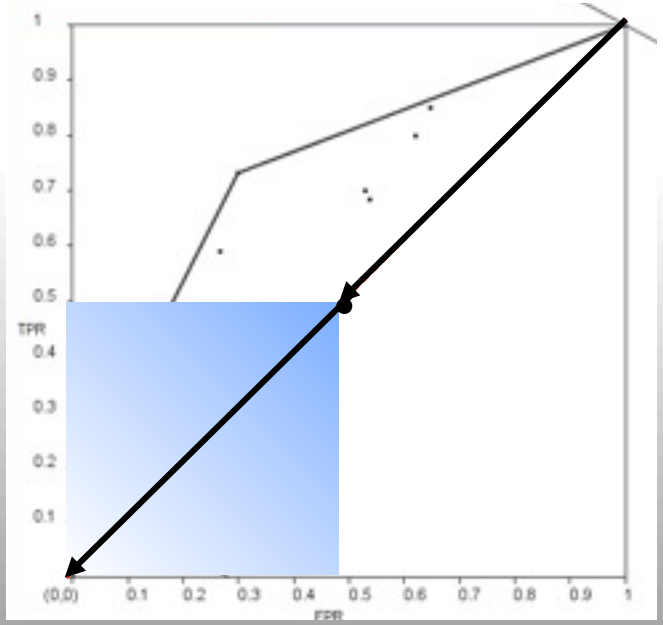


Quiz 2

Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

Think of the XOR problem. A ∧ B is either completely positive or negative.

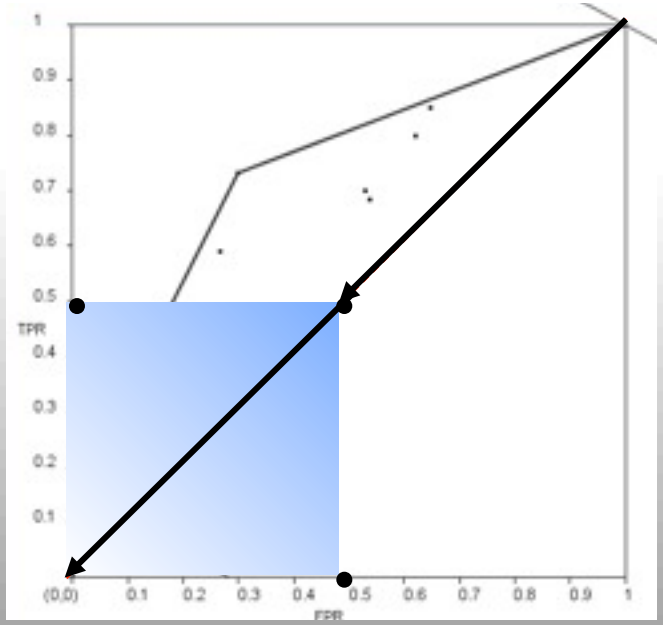


Quiz 2

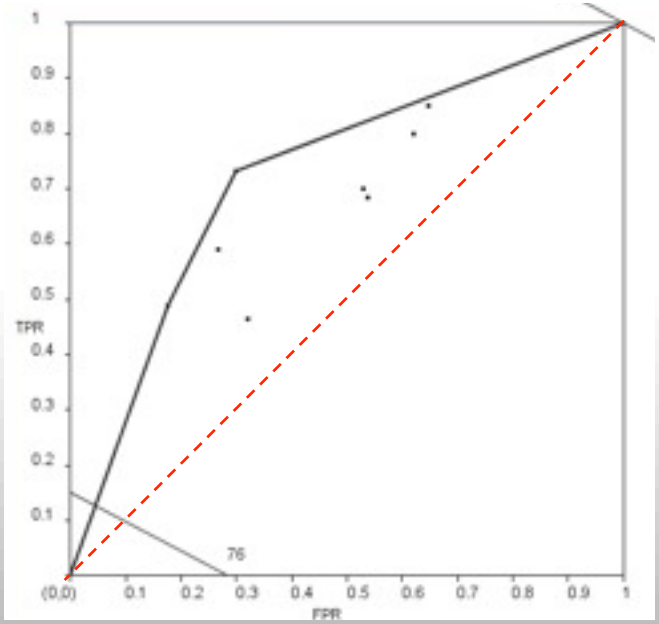
Q: Assume both A and B are uninteresting subgroups. Can the subgroup A ∧ B be an interesting subgroup?

A: Yes.

Think of the XOR problem. A ∧ B is either completely positive or negative.

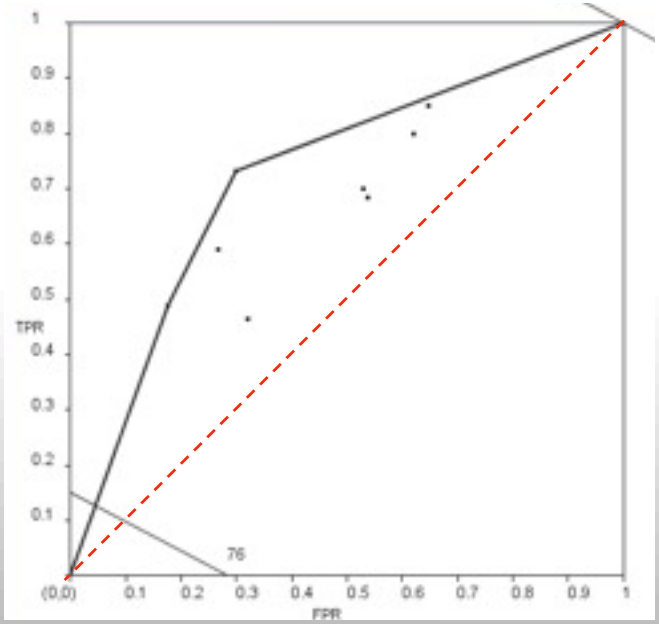


Quiz 3



Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

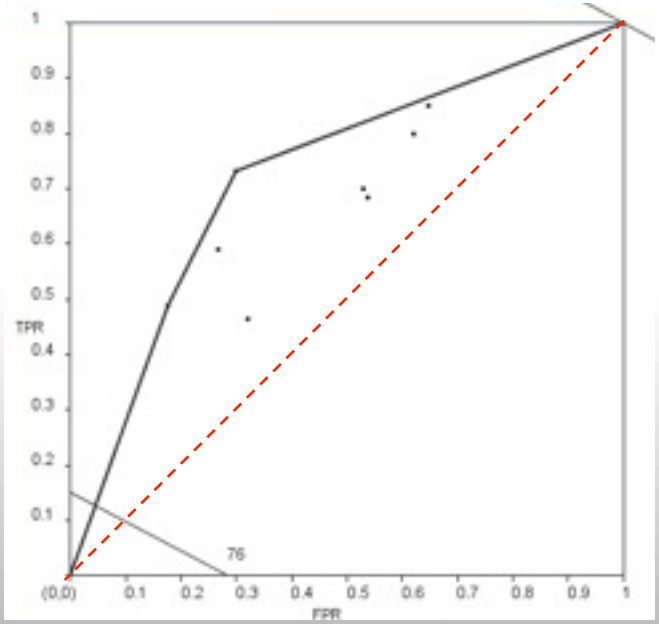


Universiteit Leiden

Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

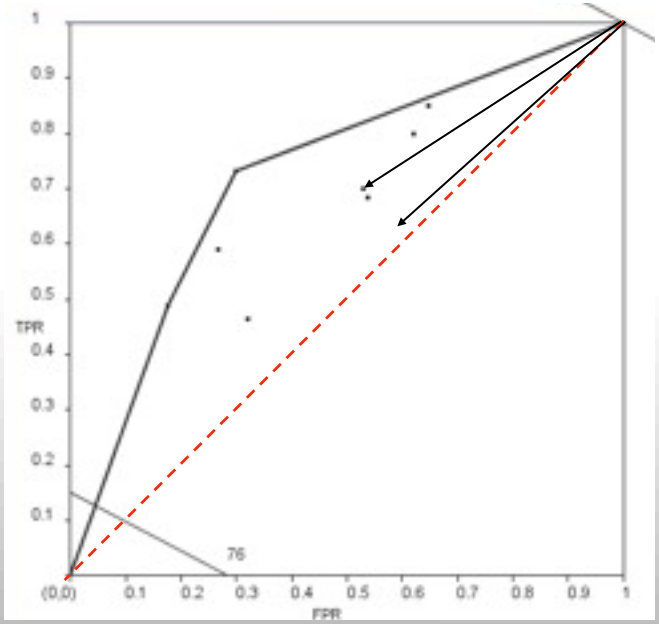
A: Yes.



Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

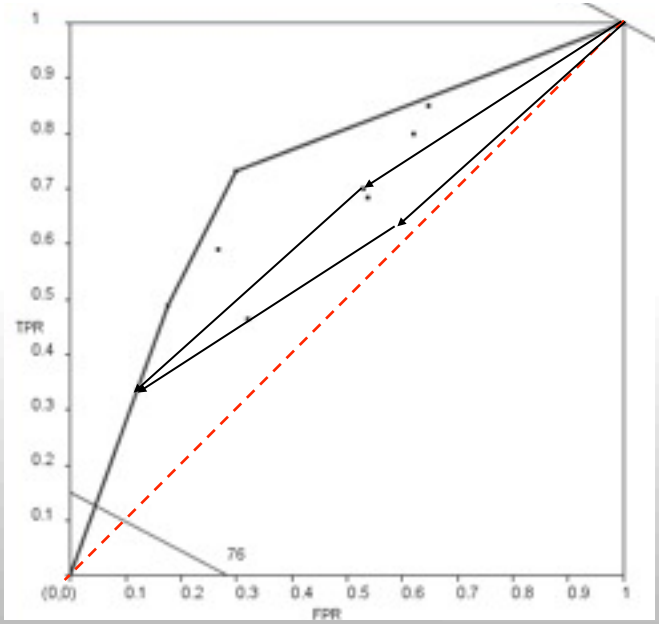
A: Yes.



Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

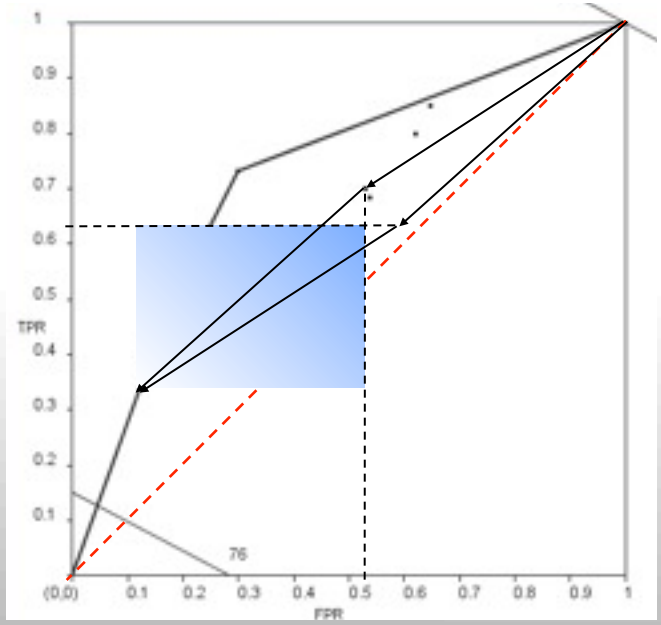
A: Yes.



Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

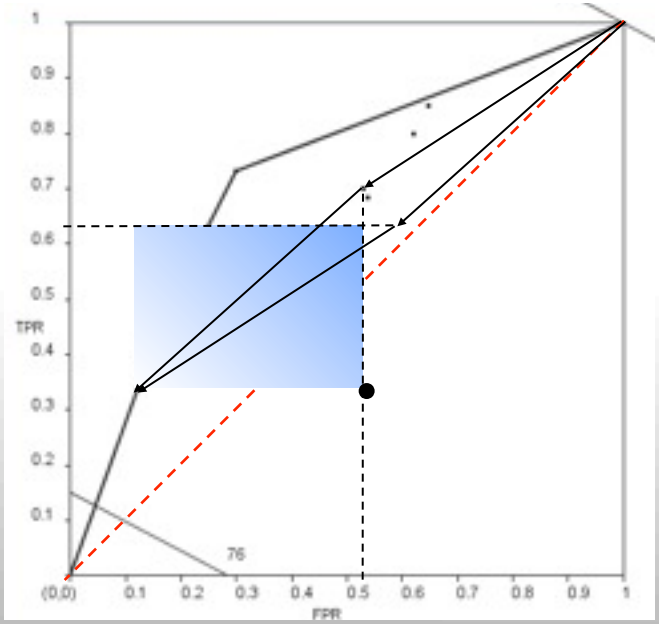
A: Yes.



Quiz 3

Q: Can the combination of two positive subgroups ever produce a negative subgroup?

A: Yes.



Universiteit Leiden