Challenges in accurate prediction of rare events with penalized likelihood methods

Georg Heinze
Medical University of Vienna
CeMSIIS
with Angelika Geroldinger, Rok Blagus, Lara Lusa:
The PREMA consortium

Our aim:

- Predicting Rare Events More Accurately

- PREMA: A joint project of Vienna and Ljubljana

Rare events: examples

Medicine:	
• Side effects of treatment	1/1000s to fairly common
• Hospital-acquired infections	9.8/1000 patient days
• Epidemiologic studies of rare diseases	1/1000 to 1/200,000 pop
Engineering:	
• Rare failures of systems	0.1-1/year
Economy:	
• E-commerce click rates	1-2/1000 impressions
Political science:	
• Wars, election surprises, vetos	1/dozens to 1/1000s
…	

Logistic regression

- Classification procedures:

  main task is to predict whether

  an event occurs or not

- Logistic regression:

  explicitely estimates the event probability

  interpretability of model parameters

Logistic regression

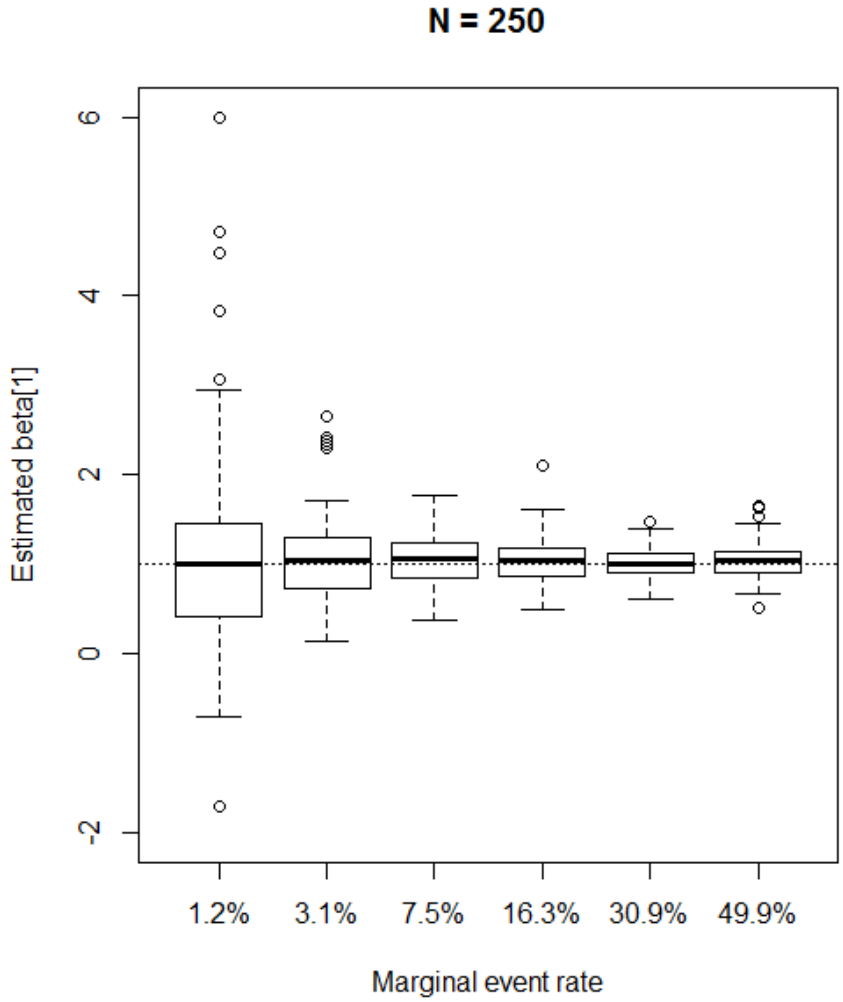
- $\Pr(Y = 1) = \pi = [1 + \exp(-X\beta)]^{-1}$

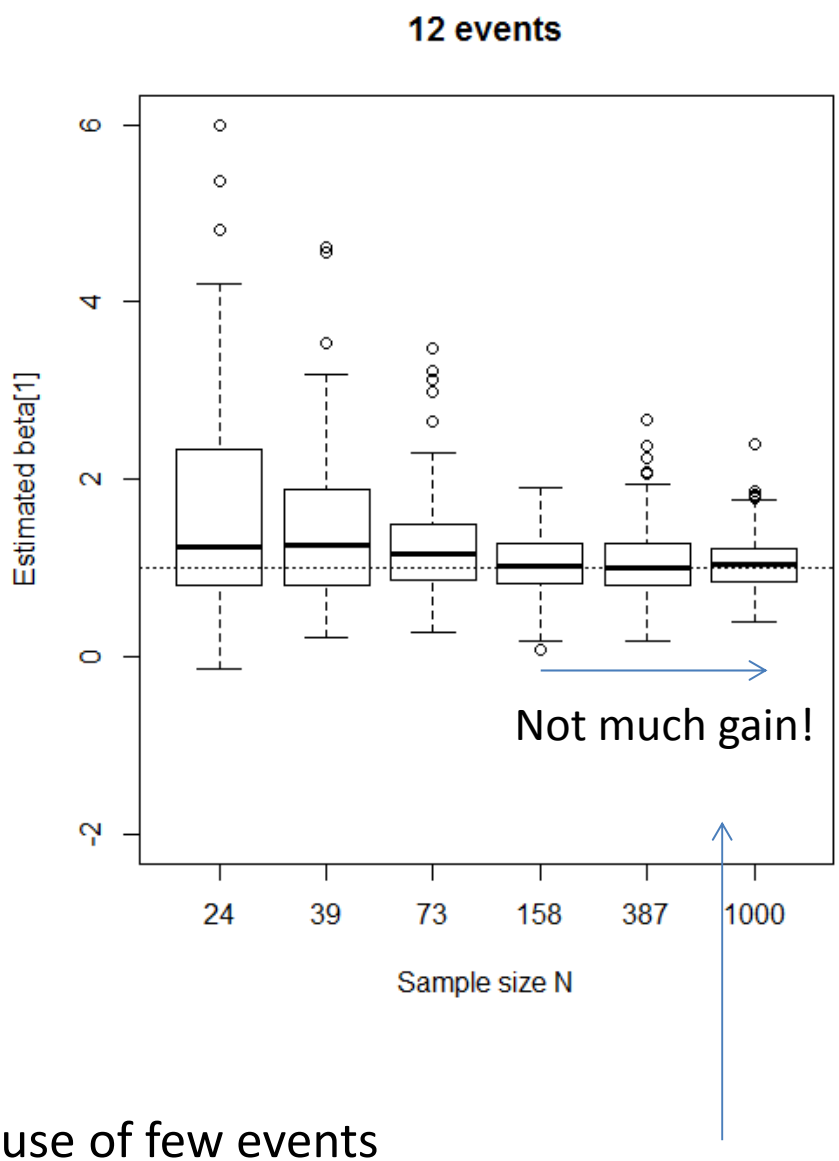
- Leads to odds ratio interpretation of $\exp(\beta)$:
- $\exp(\beta) = \frac{\Pr(Y = 1|X = x_0 + 1)/\Pr(Y=0|X=x_0+1)}{\Pr(Y = 1|X = x_0)/\Pr(Y=0|X=x_0)}$

- Likelihood: $\mathrm{L}(\beta|X) = \prod_{i=1}^{n} \hat{\pi}_i{}^{y_i}(1 - \hat{\pi}_i)^{1-y_i}$
- Its n^{th} root: Probability of correct prediction

- How well can we estimate β if events ($y_i = 1$) are rare?

Rare event problems…





Logistic regression with 5 variables:

- estimates are unstable (large MSE) because of few events
- removing some ‚non-events' does not affect precision

Focus on rare events

- $\mathrm{L}(\beta|X) = \prod \hat{\pi}_i{}^{y_i}(1-\hat{\pi}_i)^{1-y_i}$
- Suppose there are many subjects with $y_i = 0$
- Likelihood is then dominated by $(1-\hat{\pi}_i)^{1-y_i}$

- We will still have a ‚good average prediction'
- ‚Bad' predictions for $y_i = 1$ will be outbalanced

- Likelihood and many other conventional measures to assess model performance are dominated by non-events

- In our project, we will consider penalized likelihood methods as they improve on instability issues
- It is not yet well understood how they behave in rare event situations!

Penalized likelihood regression

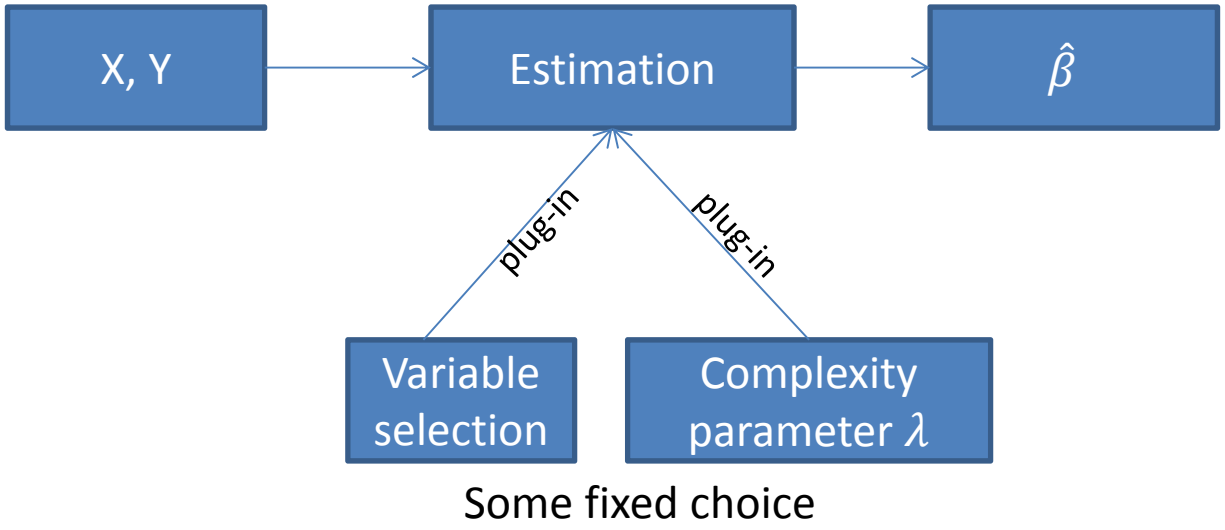
- $\log L^*(\beta) = \log L(\beta) + A(\beta)$

- Imposes priors on model coefficients, e.g.
- $A(\beta) = -\lambda \sum \beta^2$ (ridge: normal prior)
- $A(\beta) = -\lambda \sum |\beta|$ (LASSO: double exponential)
- $A(\beta) = \frac{1}{2} \log \det(I(\beta))$ (Firth-type: Jeffreys prior)

- To avoid extreme estimates and stabilize variance (ridge)
- To perform variable selection (LASSO)
- To correct small-sample bias in β (Firth-type)

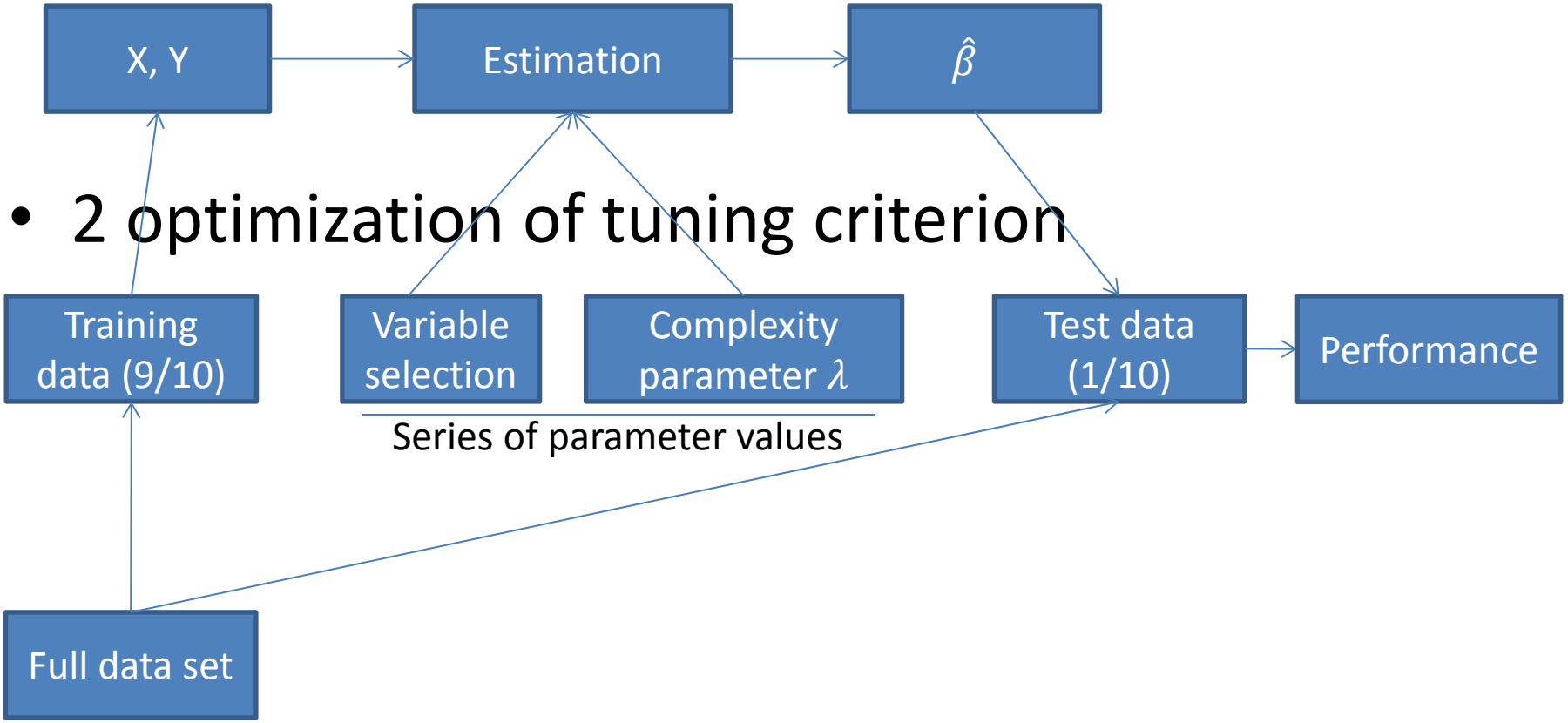
Layers of predictive model building

- 1 coefficient estimation module



Layers of predictive model building

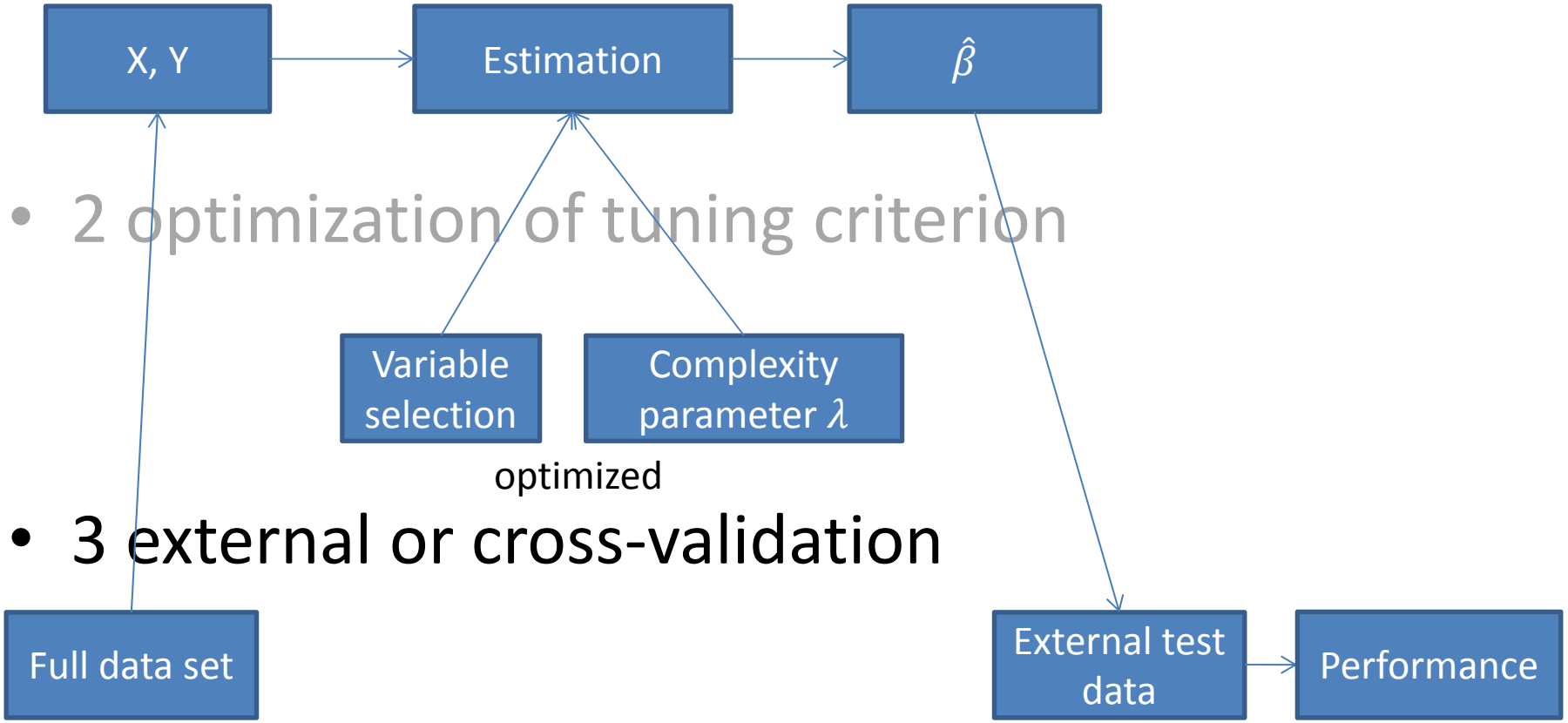
- 1 coefficient estimation module



- 2 optimization of tuning criterion

Layers of predictive model building

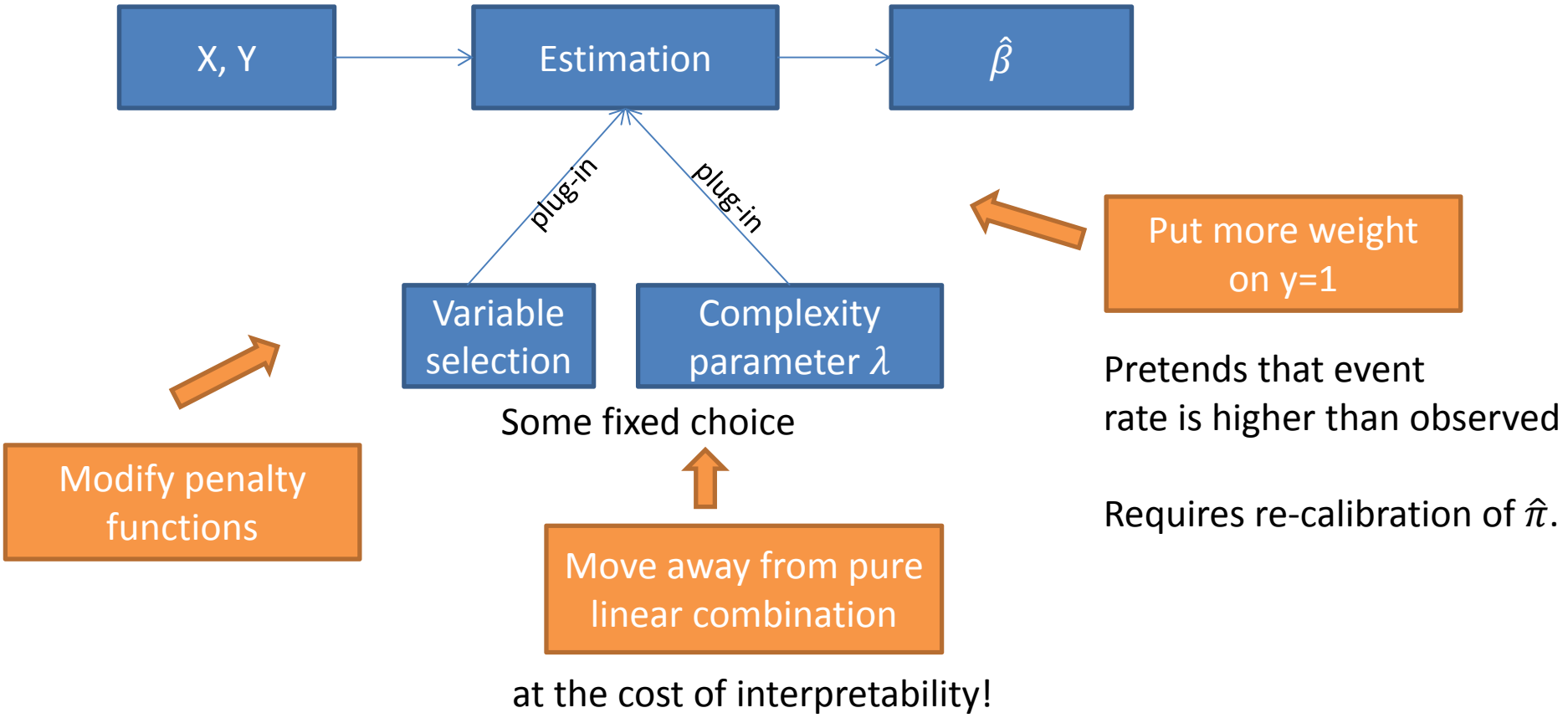
- 1 coefficient estimation module



- 2 optimization of tuning criterion
- 3 external or cross-validation

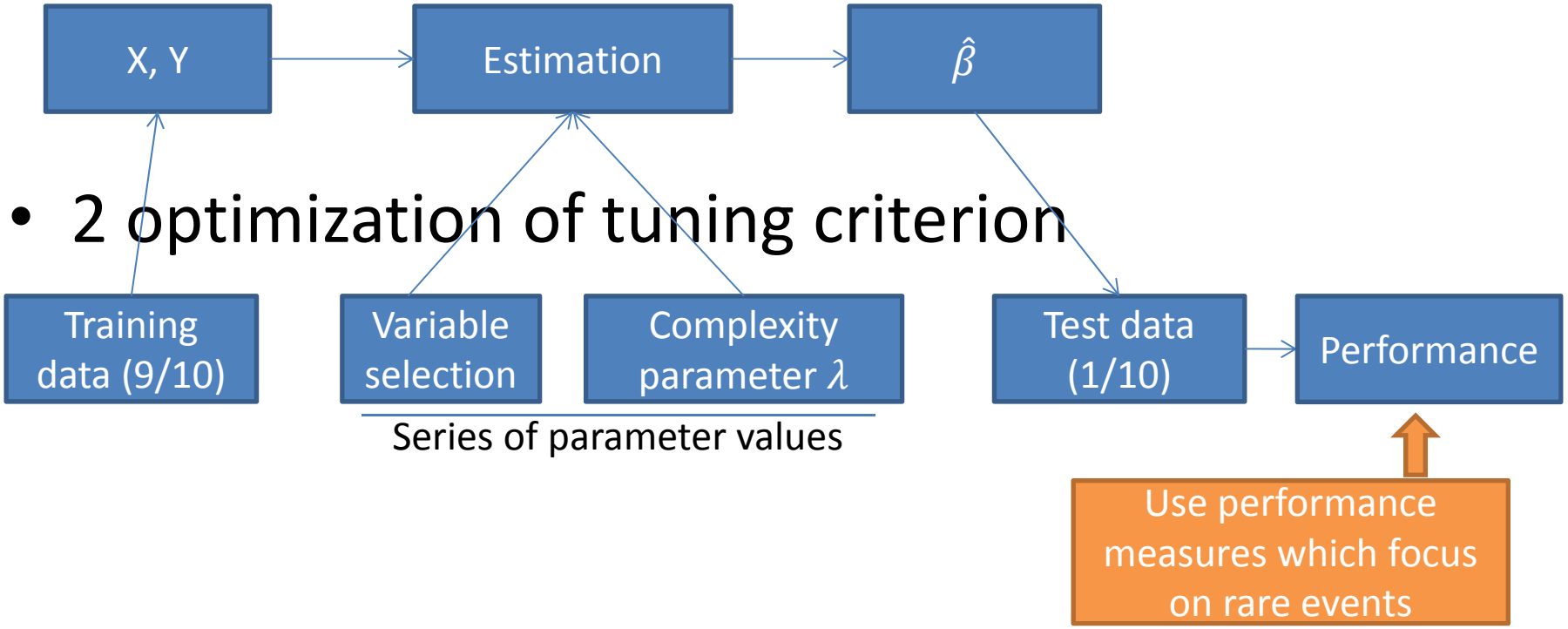
Modifications for rare events (?)

- 1 coefficient estimation module



Modifications for rare events (?)

- 1 coefficient estimation module



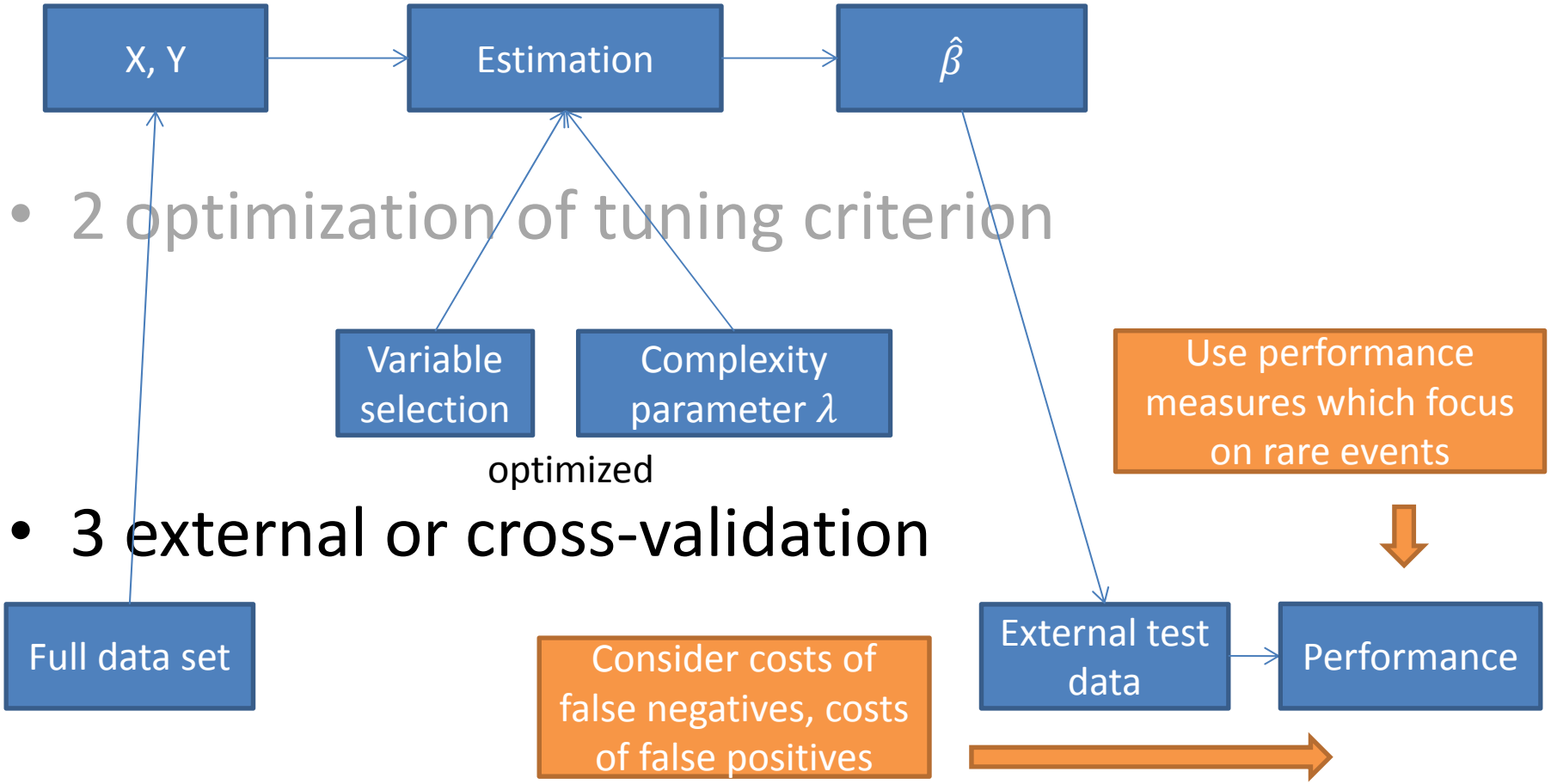
- 2 optimization of tuning criterion

Redefinition of:

Deviance, AUROC, misclassification error, …

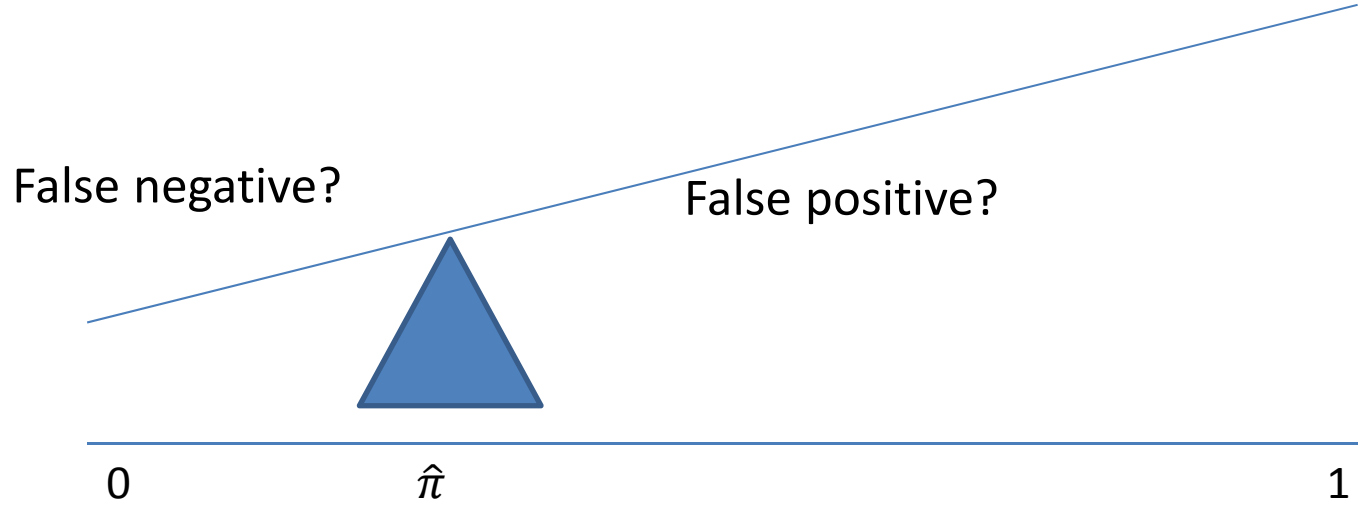
Modifications for rare events (?)

- 1 coefficient estimation module
- 2 optimization of tuning criterion
- 3 external or cross-validation



Cost of a false prediction

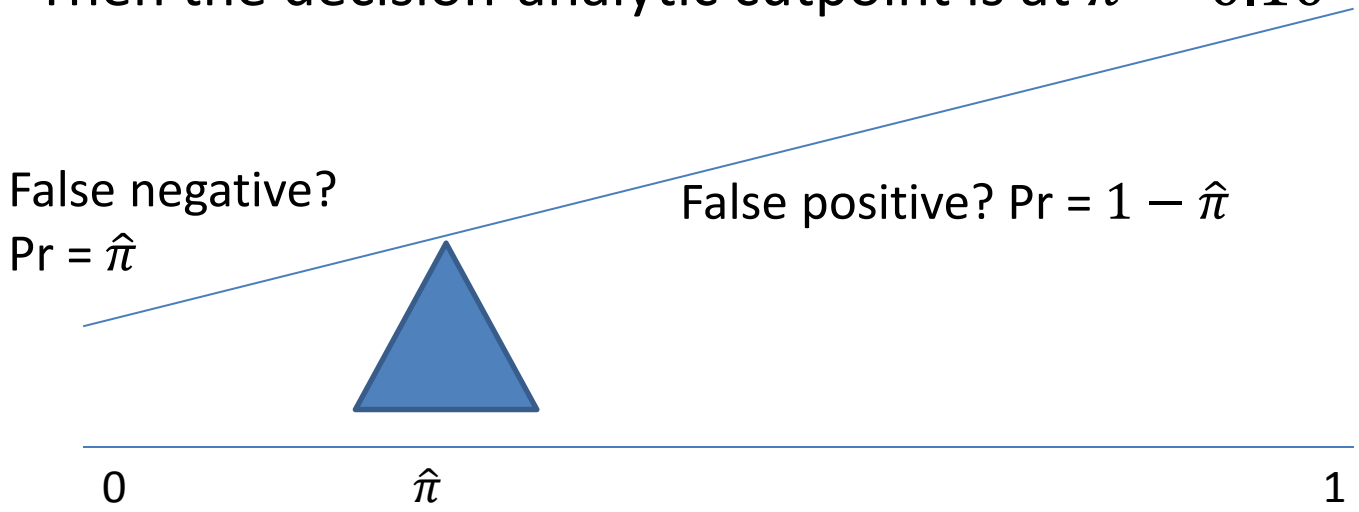
- A decision-maker‘s individual cut-point depends on the costs for…



Vickers, Med Dec Making 2006

Costs of a false prediction

- Suppose the ‚costs' of a false negative (eg, no treatment of a patient) are 9 times the ‚costs' of a false positive (unnecessary treatment)
- We take into account 9 unnecessary treatments for saving 1 patient
- Then the decision-analytic cutpoint is at $\pi = 0.10$



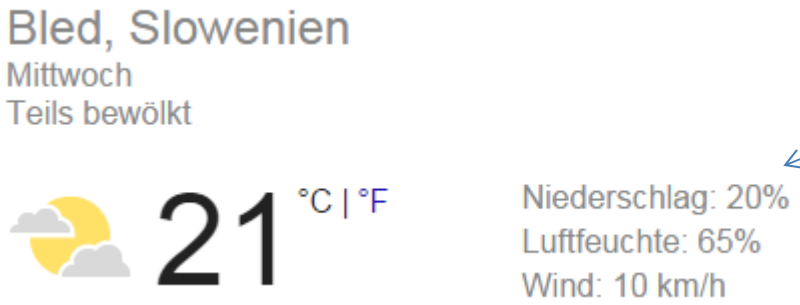
Vickers, Med Dec Making 2006

From the decision-analytic viewpoint…

- The decision-analytic cutpoint depends on the risk of missing an event and the costs of a false alarm
- It may be very different from 0.5 or from the marginal event rate
- Not the same cutpoint for all ‚users‘ of a model!

From the decision-analytic viewpoint…

- Weather forecast for tomorrow:



Probability of rainfall

‚Costs' of unnecessarily carrying (and probably loosing) my umbrella,
‚Risk' of getting wet

- If my personal cutpoint for taking the umbrella with me is 20% (*getting wet is 4 times worse than loosing my umbrella*),
- then I am not interested whether the prediction is 50% or 70%, but I need high accuracy for a prediction around 20%.

Conclusions

- Estimation/Tuning/Validation criteria should reflect our use of the model:
- Do we need high accuracy in the full range of predictions?

- What are the costs of a missed event?

getting wet	←→	loosing my umbrella
financial crisis	←→	monetary actions
heavy refugee movement	←→	border control
stroke	←→	drug side effects

- What are the costs of a false alarm?

- → Consequences of predictions are important

Still…