Q&A SESSION 2

NEWS

① WELCOME FEM21045 PEOPLE!

② PROF. DR. DICK VAN DIJK
KAREL DE WIT

Syllabus update
(soon...)

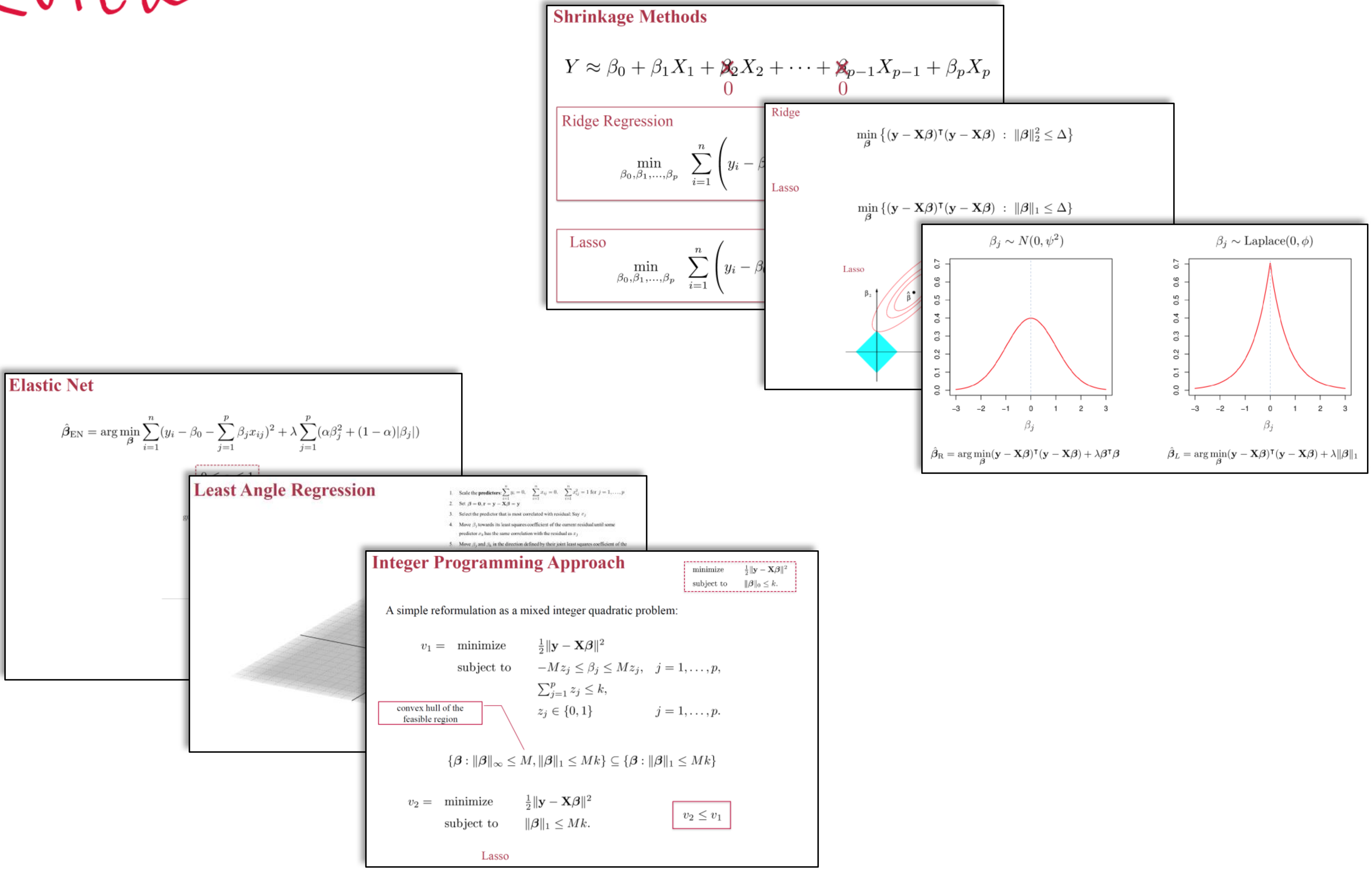
③ NEW OPTIONAL MATERIAL FOR FEM31002
(readings, application discussions, videos, and so on.)

~ 400 STUDENTS

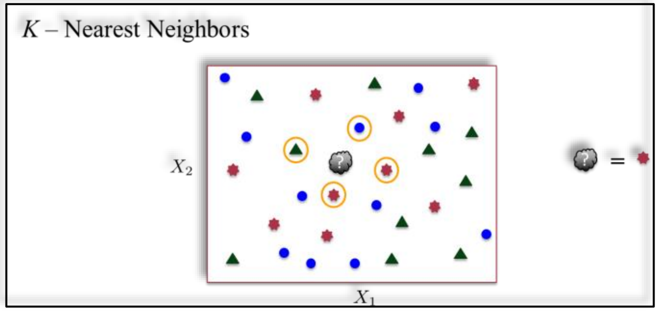
OUTLINE

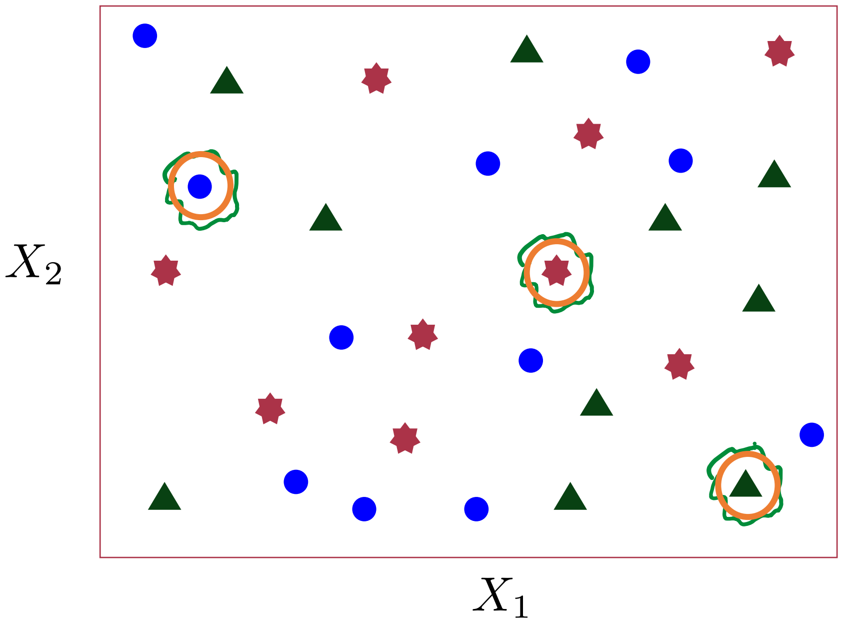
- OVERVIEW OF WEEK 2
- QUESTIONS FROM DISCUSSION FORUM 2
  - K-NN
  - K-FOLD CV
  - BIAS-VARIANCE
  - RIDGE-LASSO & CONSTRAINED OPTIMIZATION
  - INTEGER PROGRAMMING: CONVEX HULL
  - LEAST ANGLE REGRESSION
  - WHY REGULARIZATION?

OVERVIEW


Shrinkage Methods
$Y \approx \beta_0 + \beta_1 X_1 + \cancel{\beta_2} X_2 + \cdots + \cancel{\beta_{p-1}} X_{p-1} + \beta_p X_p$
0
0
Ridge Regression
Lasso
Ridge
$\min_{\boldsymbol\beta} \{(\mathbf{y}-\mathbf{X}\boldsymbol\beta)^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol\beta) \ : \ \|\boldsymbol\beta\|_2^2 \le \Delta\}$
Lasso
$\min_{\boldsymbol\beta} \{(\mathbf{y}-\mathbf{X}\boldsymbol\beta)^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol\beta) \ : \ \|\boldsymbol\beta\|_1 \le \Delta\}$
Lasso
$\beta_j \sim N(0, \psi^2)$
$\beta_j \sim \text{Laplace}(0, \phi)$
$\hat{\boldsymbol\beta}_R = \arg\min_{\boldsymbol\beta} (\mathbf{y}-\mathbf{X}\boldsymbol\beta)^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol\beta) + \lambda \boldsymbol\beta^\intercal\boldsymbol\beta$
$\hat{\boldsymbol\beta}_L = \arg\min_{\boldsymbol\beta} (\mathbf{y}-\mathbf{X}\boldsymbol\beta)^\intercal(\mathbf{y}-\mathbf{X}\boldsymbol\beta) + \lambda \|\boldsymbol\beta\|_1$
Elastic Net
$\hat{\boldsymbol\beta}_{EN} = \arg\min_{\boldsymbol\beta} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$
Least Angle Regression
Integer Programming Approach
A simple reformulation as a mixed integer quadratic problem:
$v_1 = \text{minimize} \ \tfrac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol\beta\|^2$
subject to $-Mz_j \le \beta_j \le Mz_j, \quad j = 1, \dots, p,$
$\sum_{j=1}^p z_j \le k,$
$z_j \in \{0, 1\} \quad j = 1, \dots, p.$
convex hull of the feasible region
$\{\boldsymbol\beta : \|\boldsymbol\beta\|_\infty \le M, \|\boldsymbol\beta\|_1 \le Mk\} \subseteq \{\boldsymbol\beta : \|\boldsymbol\beta\|_1 \le Mk\}$
$v_2 = \text{minimize} \ \tfrac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol\beta\|^2$
subject to $\|\boldsymbol\beta\|_1 \le Mk.$
$v_2 \le v_1$
Lasso

K-NN





K = 1

Training Error = 0 ?

Ties ?

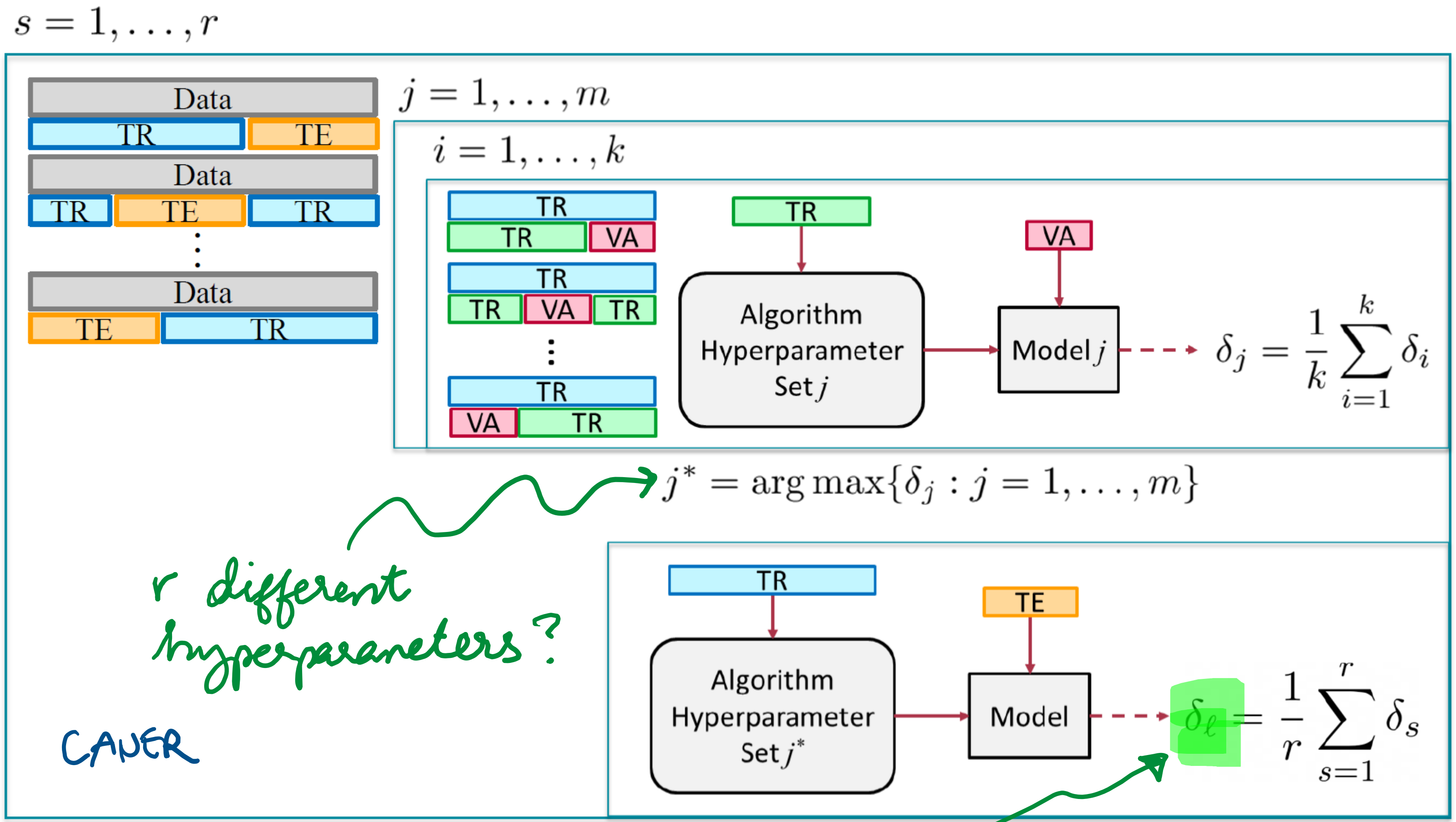
random assignment

(next) odd number for K

LIEVE
- CANER
- MARTEN

WAYNE
- MARTEN

K-FOLD CV



r different hyperparameters?

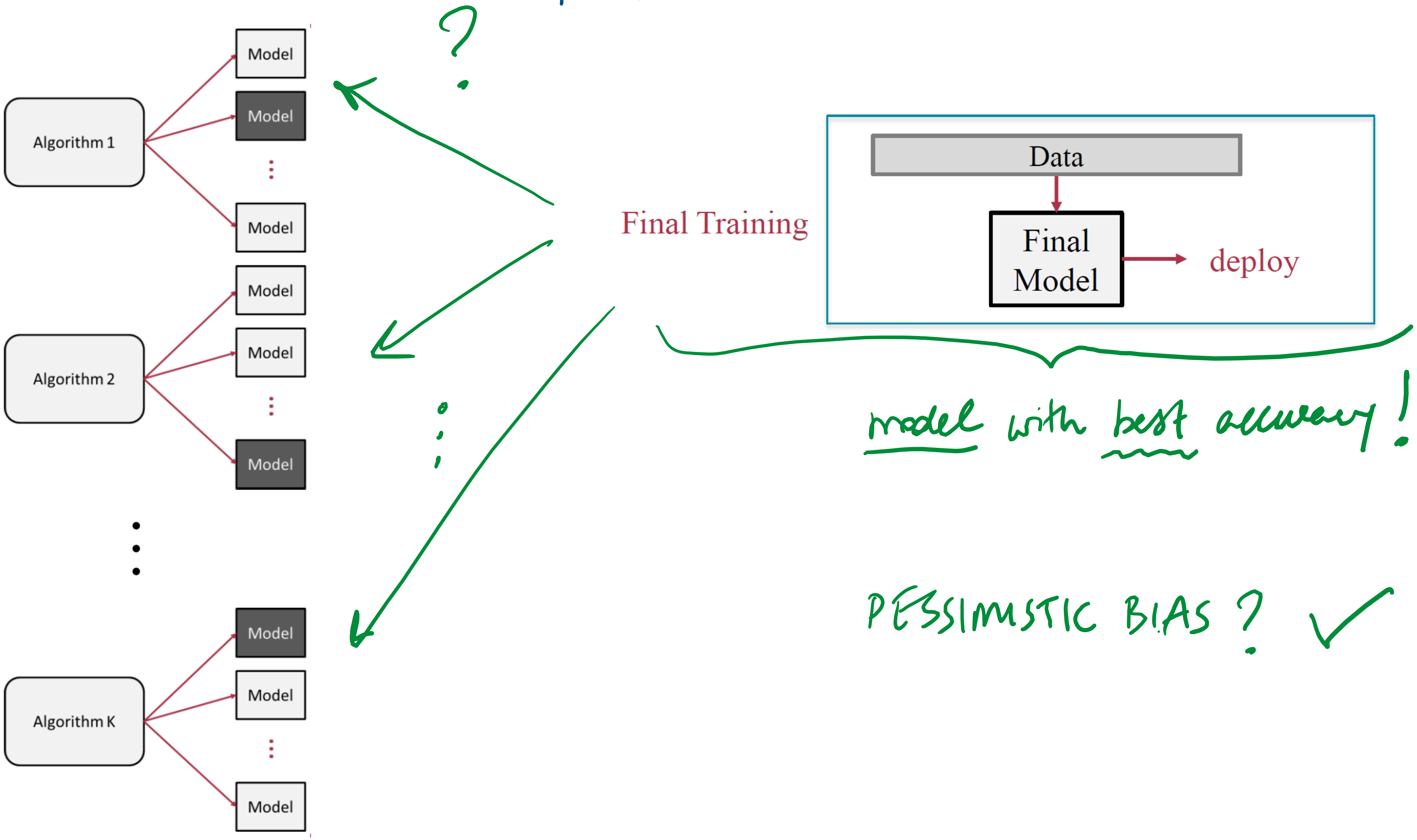
CANER

performance measure for algorithm ℓ

⊕ TUNING

K-FOLD CV

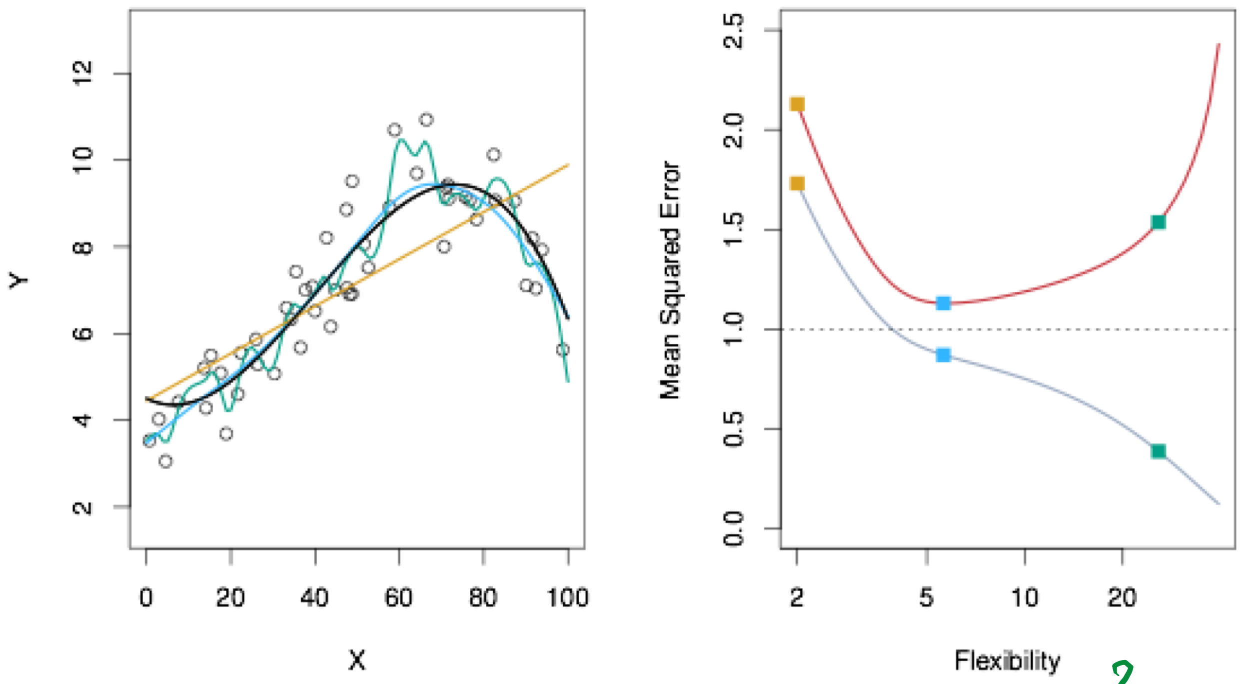
KOEN



model with best accuracy!

PESSIMISTIC BIAS ? ✓

BIAS_VARIANCE



?

(continuous) parameter of a smoothing spline

VIVI

≈ power of a polynomial in curve fitting

BIAS_VARIANCE

K-FOLD CV

Test Error

k ↑	Variance ↑	Bias ↓

$k \uparrow$ test size ↓ train size ↑

KOEN

OPTIMISTIC / PESSIMISTIC BIAS ?

Underestimates the test error rate ⇒ optimistic bias

LOOCV ($k=n$) gives almost the unbiased estimator of the test error.

RIDGE-LASSO & CONSTRAINED OPTIMIZATION

? LOURENS

Ridge

(a) $$\min_{\boldsymbol{\beta}} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ : \ \|\boldsymbol{\beta}\|_2^2 \leq \Delta\}$$

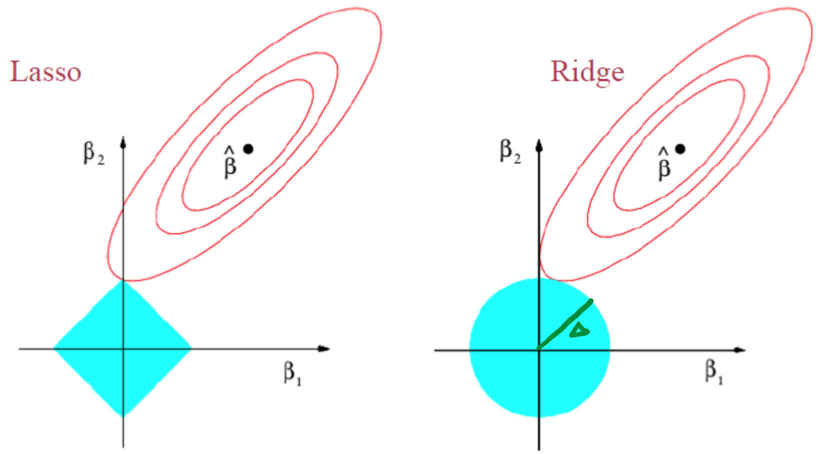
Lasso

(b) $$\min_{\boldsymbol{\beta}} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ : \ \|\boldsymbol{\beta}\|_1 \leq \Delta\}$$

fix λ

RIDGE(λ) → $\hat{\beta}_R$

LASSO(λ) → $\hat{\beta}_L$



Least Squares Solution: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$

$\exists\, \Delta > 0$ such that
- solving (a) gives $\hat{\beta}_R$
- solving (b) gives $\hat{\beta}_L$

INTEGER PROGRAMMING: CONVEX HULL

$$
\begin{aligned}
v_1 = \quad & \text{minimize} && \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\
& \text{subject to} && -Mz_j \le \beta_j \le Mz_j, \quad j = 1, \ldots, p, \\
& && \textstyle\sum_{j=1}^{p} z_j \le k, \\
& && z_j \in \{0, 1\} \qquad j = 1, \ldots, p.
\end{aligned}
$$

$z_j = 1 \Rightarrow |\beta_j| \le M$

at most k-many z_j can be 1

convex hull of the feasible region

LOURENS ?

$$\{\boldsymbol{\beta} : \underbrace{\|\boldsymbol{\beta}\|_\infty \le M}, \underbrace{\|\boldsymbol{\beta}\|_1 \le Mk}\} \subseteq \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \le Mk\}$$

$\max\{|\beta_1|, \ldots, |\beta_p|\} \le M$

$\sum_{j=1}^{p} |\beta_j| \le Mk$

LEAST ANGLE REGRESSION

CANER

β_1, β_2 ?

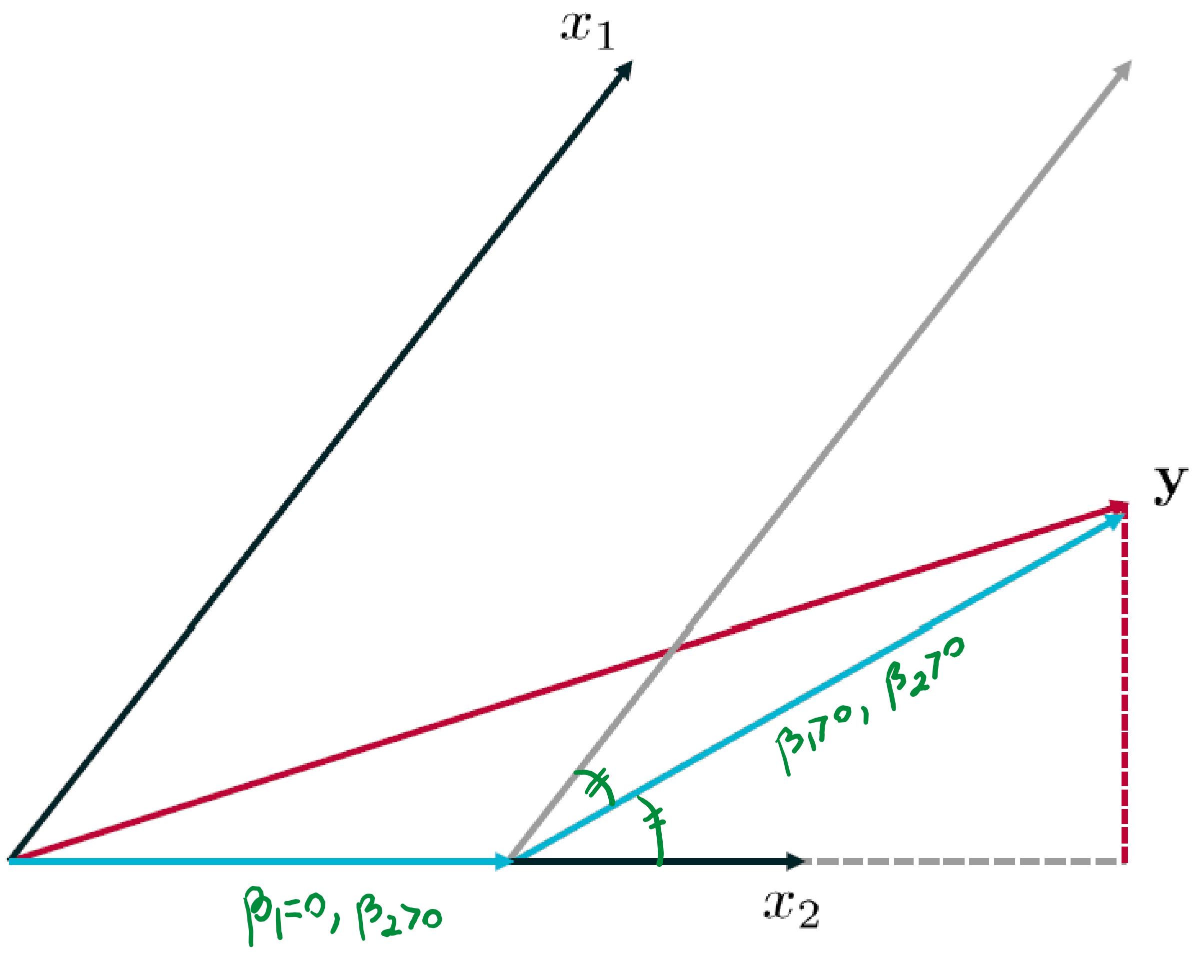
$\min\{n-1, p\}$ steps

?

1. Scale the **predictors**: $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0$
2. Set $\boldsymbol{\beta} = \mathbf{0}, \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$

Columns are centered

$\Rightarrow$ row rank of $X = n-1$



WHY REGULARIZATION ?

How about ? { linear regression + significance tests for parameters

MARTEN

OLS

→ Collinearity

→ $p > n$

Regularization:

- Dimension reduction (interpretability)
- Avoiding overfitting (e.g. NNs)

Forward Selection

→ Many tests

→ Problems with sequential testing

Sequential Selection Procedures and False Discovery Rate Control

Max Grazier G'Sell
Department of Statistics, Carnegie Mellon University, Pittsburgh, USA.
Stefan Wager
Department of Statistics, Stanford University, Stanford, USA.
Alexandra Chouldechova
Heinz College, Carnegie Mellon University, Pittsburgh, USA.
Robert Tibshirani
Departments of Health Research & Policy, and Statistics, Stanford University, Stanford, USA.

link

TILL NEXT WEEK!

* Assignment 2 (Due date: 14 Sep.)
* Videos for lecture 3 (↓ 12 pm)