ML-workflow

Funkar modellen på ny data?

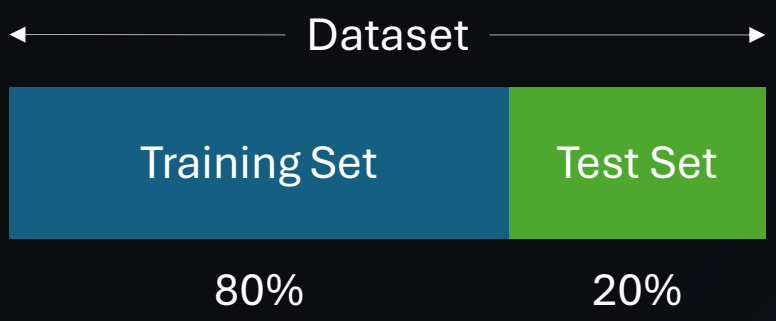
- Träning ≠ verklighet
- Målet: generalisering

Utvärdera inte på träningsdata

- Träningsresultat kan bli för bra
- Du riskerar att ”lura dig själv”
- Behöver data som modellen inte sett

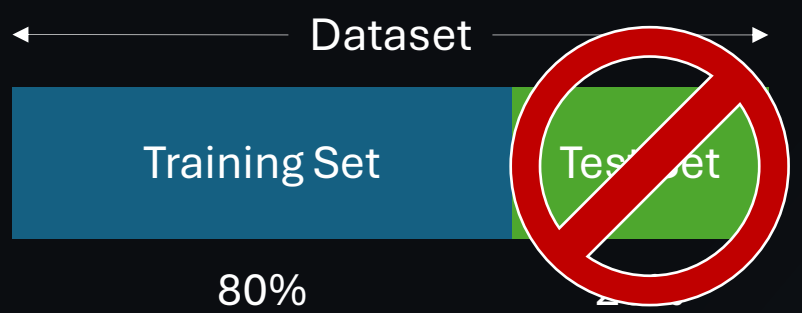
Train/Test split

- Train set: modellen lär sig
- Test set: slutlig kontroll
- Vanligt: 80/20



Testdata är ”helig”

- Test = bara en gång, på slutet
- Använd inte test för modellval
- Annars: ”data leakage”

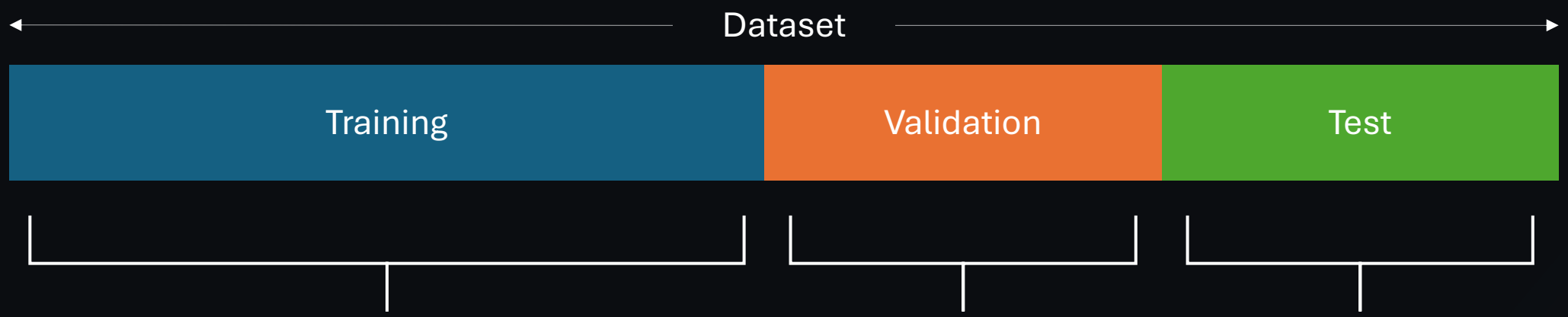


Problemet: vi vill testa flera modeller

- I praktiken testar vi ofta flera modeller:
  - Linjär/logistisk regression
  - Beslutsträd / Random Forest
  - SVM …
- Men… vi får inte välja med test

Train / Validation / Test

(Validation Set Metodiken)



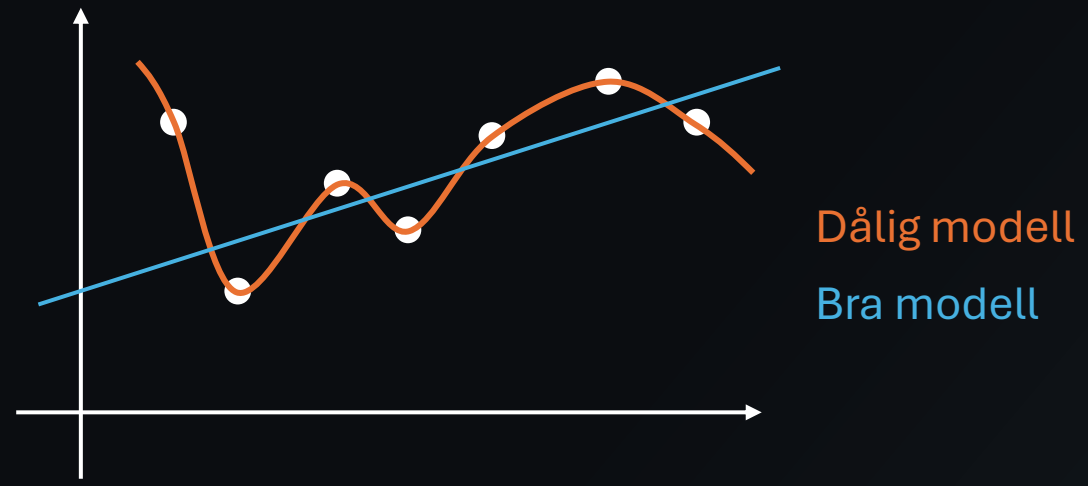
Skapa t.ex. 5 olika modeller

Utvärdera modellerna här och välj den bästa

Testa den valda modellen här och få ett estimate av “generalization error”. Innan du gör det, träna om den valda modellen på train + validation datan.

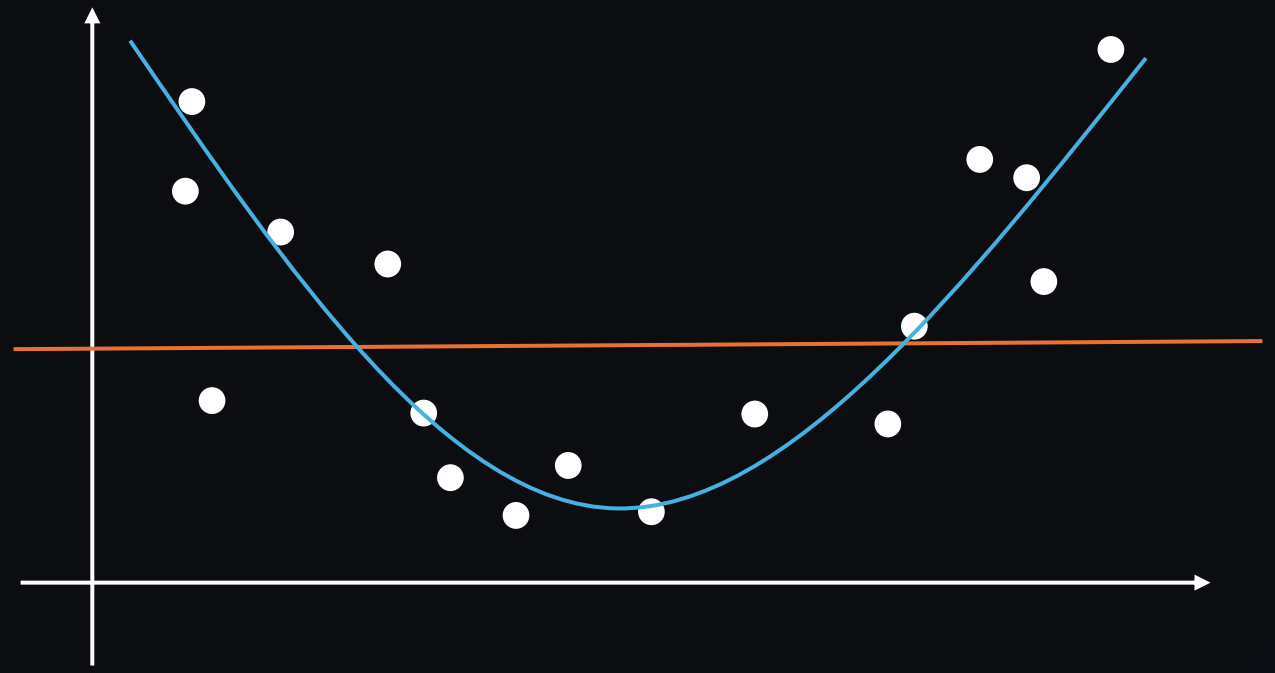
Overfitting (Överanpassning)

- Overfitting: bra på train, dålig på ny data
- Typiskt tecken:
  - Lågt fel på train
  - Högt fel på validation/test
- Ofta för komplex modell



Underfitting (Underanpassning)

- Underfitting: modellen är för enkel
- Högt fel på train **och** validation/test
- Modellen missar grundmönstret



Balansen

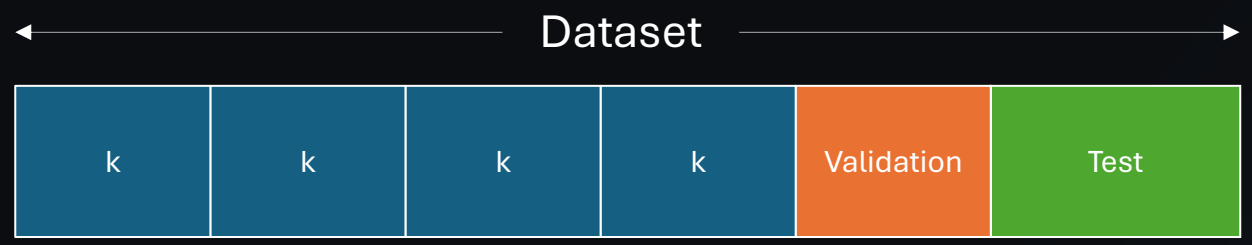
- Målet: lagom komplexitet
- För enkel → underfitting
- För komplex → overfitting
- Vi hittar balans genom validering/CV

Validation-split kan vara osäker

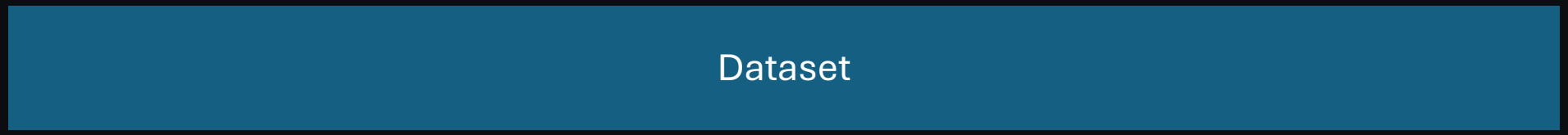
- En split kan ge ”tur/otur”
- Resultat kan variera beroende på hur du delar
- Vi vill ha mer robust uppskattning

K-fold Cross Validation

- Dela train i k delar (folds)
- Träna k gånger:
  - K-1 folds = träning
  - 1 fold = validering
- Ta medelvärde av resultaten

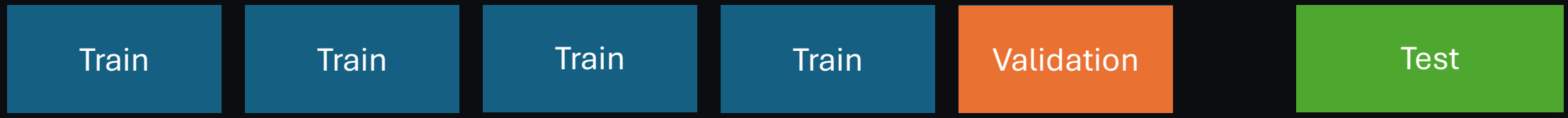


K-fold Cross Validation

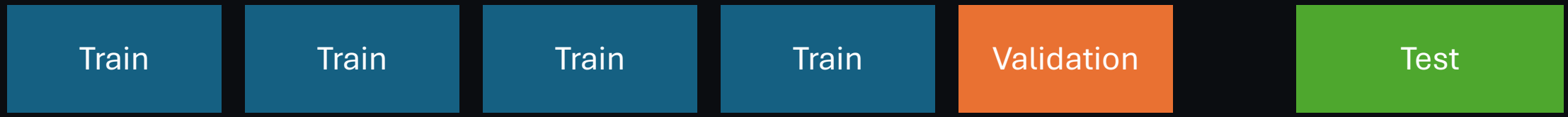


K-fold Cross Validation

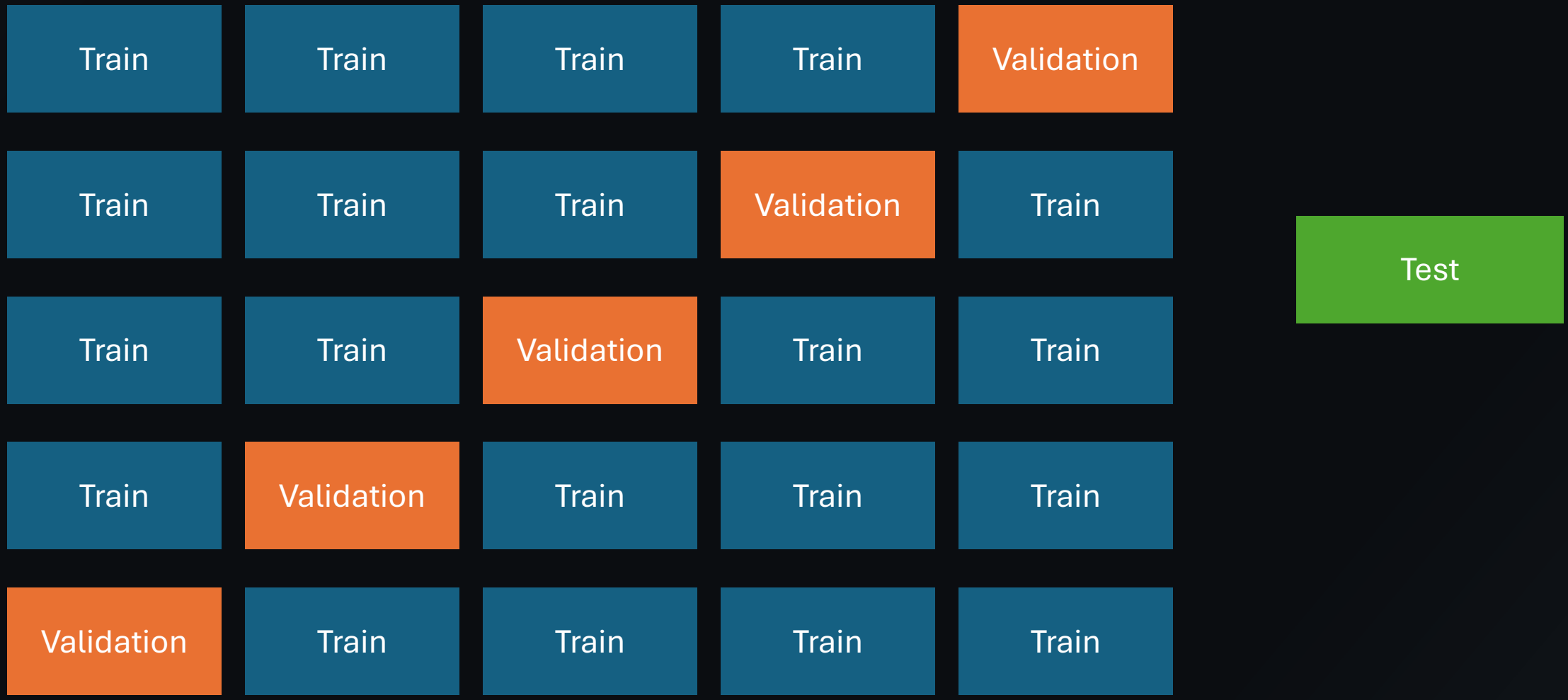
K-fold Cross Validation



K-fold Cross Validation



K-fold Cross Validation



Vanliga val

- Vanligast: k = 5 eller k = 10
- Fördel: stabilare skattning
- Nackdel: tar längre tid (träna k gånger)

Helheten

Rekommenderat workflow

1. Split Train/Test
2. På Train: K-fold CV för modellval/hyperparametrar
3. Träna om bästa modellen på hela Train
4. Utvärdera en gång på Test

Data-utmaningar

- För lite data
- Dålig datakvalitet
- Otydliga definitioner
- Icke-representativ data
- Irrelevanta features

ML-workflow

Joakim Lindh