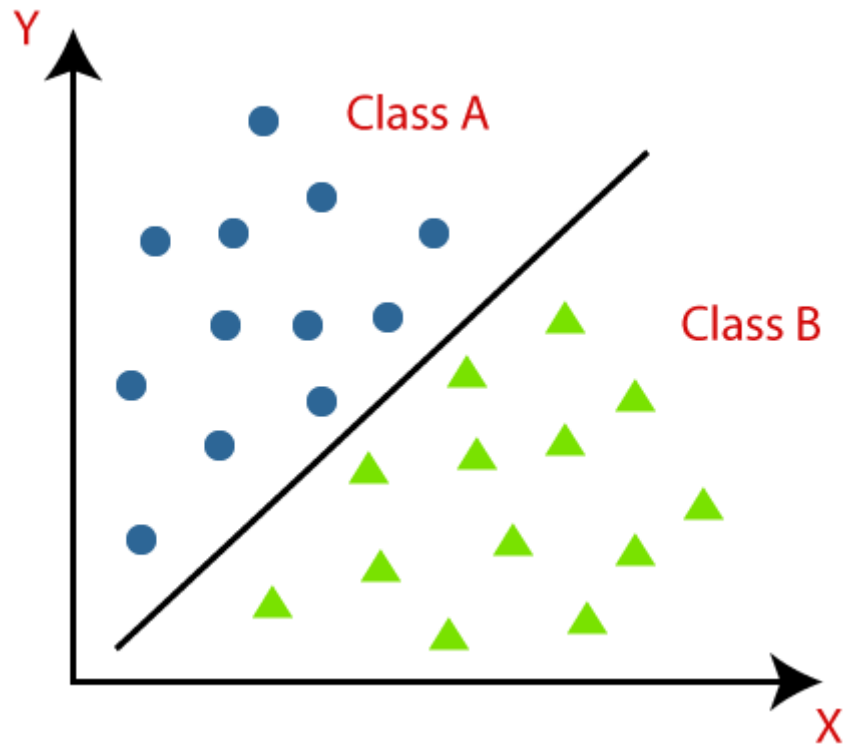
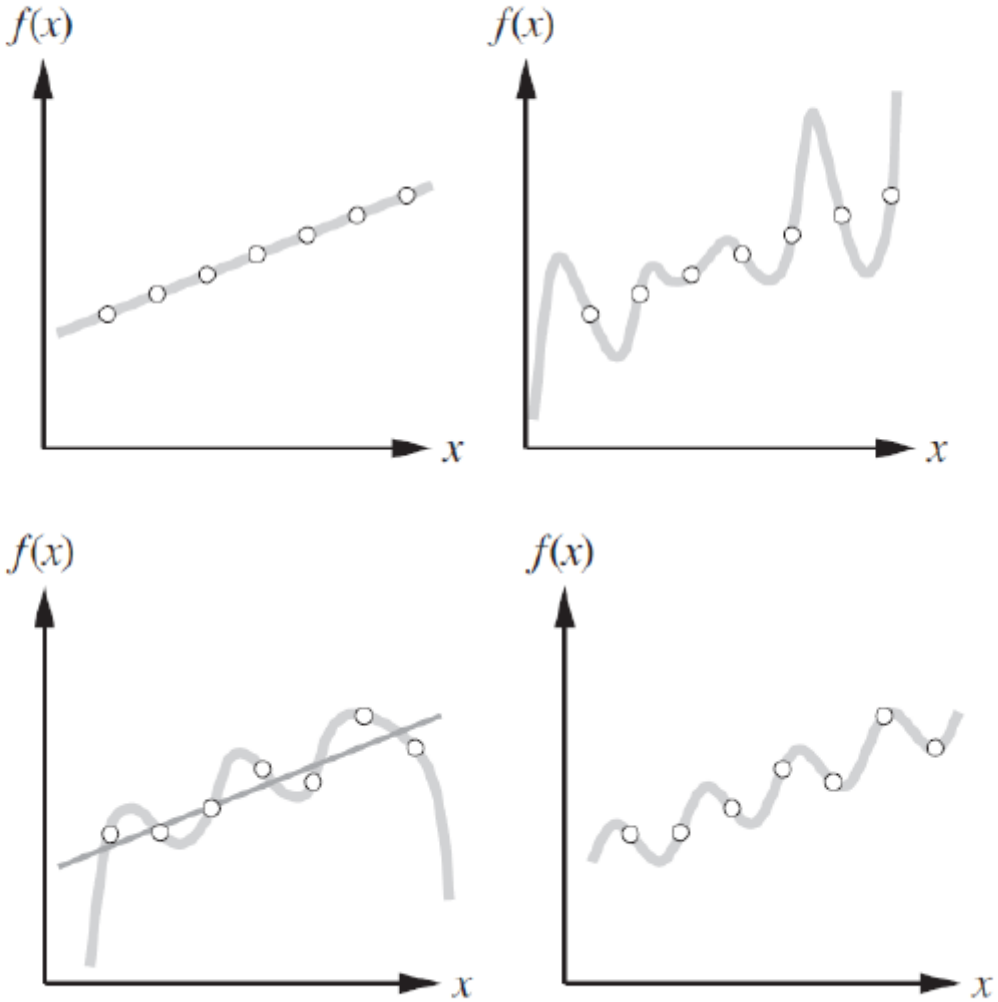
Mašinsko učenje 1

MAS Informatike – Nauka o podacima

Klasifikacija i regresija





Evaluacija naučenog

- Kako oceniti kvalitet obučenog modela?
- Koje podatke treba koristiti za procenu kvaliteta modela?
- Kako izmeriti sposobnost modela da daje tačna predviđanja?

Greška na trening skupu

- Greška koju model pravi nad podacima za učenje zove se greška učenja – **training error**.

- Ona nam govori o tome koliko je model dobro naučio trening podatke.

Evaluacija naučenog

- Ono što nas zanima jesu performanse modela na novim podacima, a ne dosadašnji učinak na starim podacima.
- Greška modela na skupu za obuku nije dobar pokazatelj budućeg učinka.
- Model je naučen iz podataka za obuku, pa je svaka procena performansi zasnovana na tim podacima suviše optimistična.

Training

Test

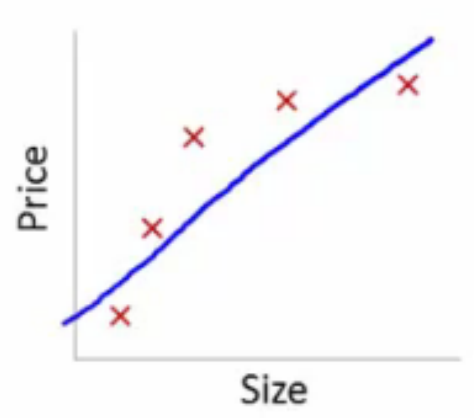
Greška na testnom skupu

- Greška koju model pravi nad testnim podacima zove se greška testiranja - **test error**.

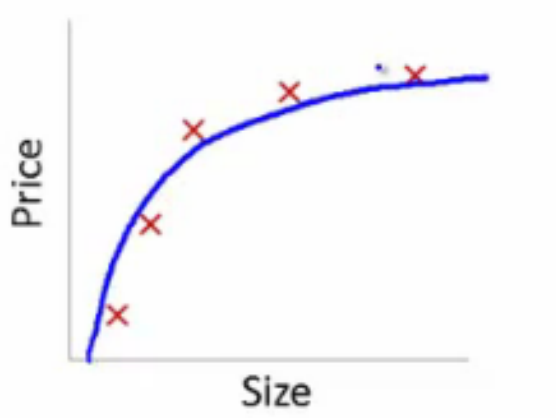
Training

Test

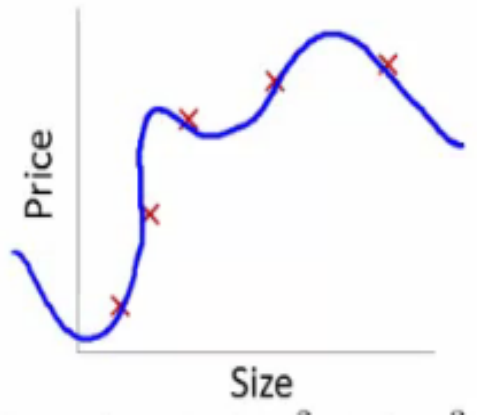
More precision ≠ Better results



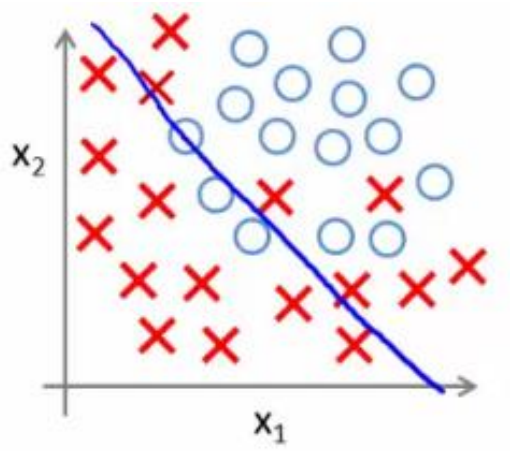
Underfit

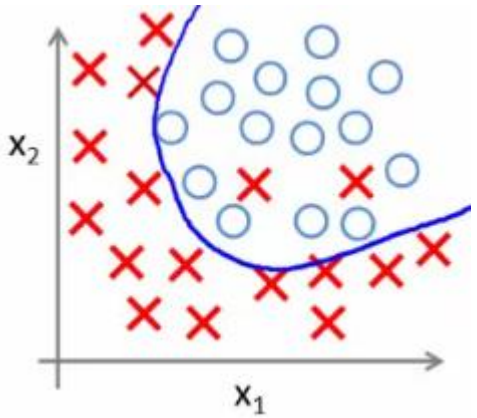


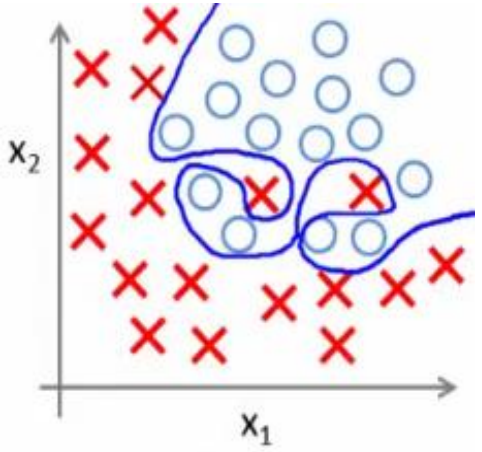
Just right!



Overfit



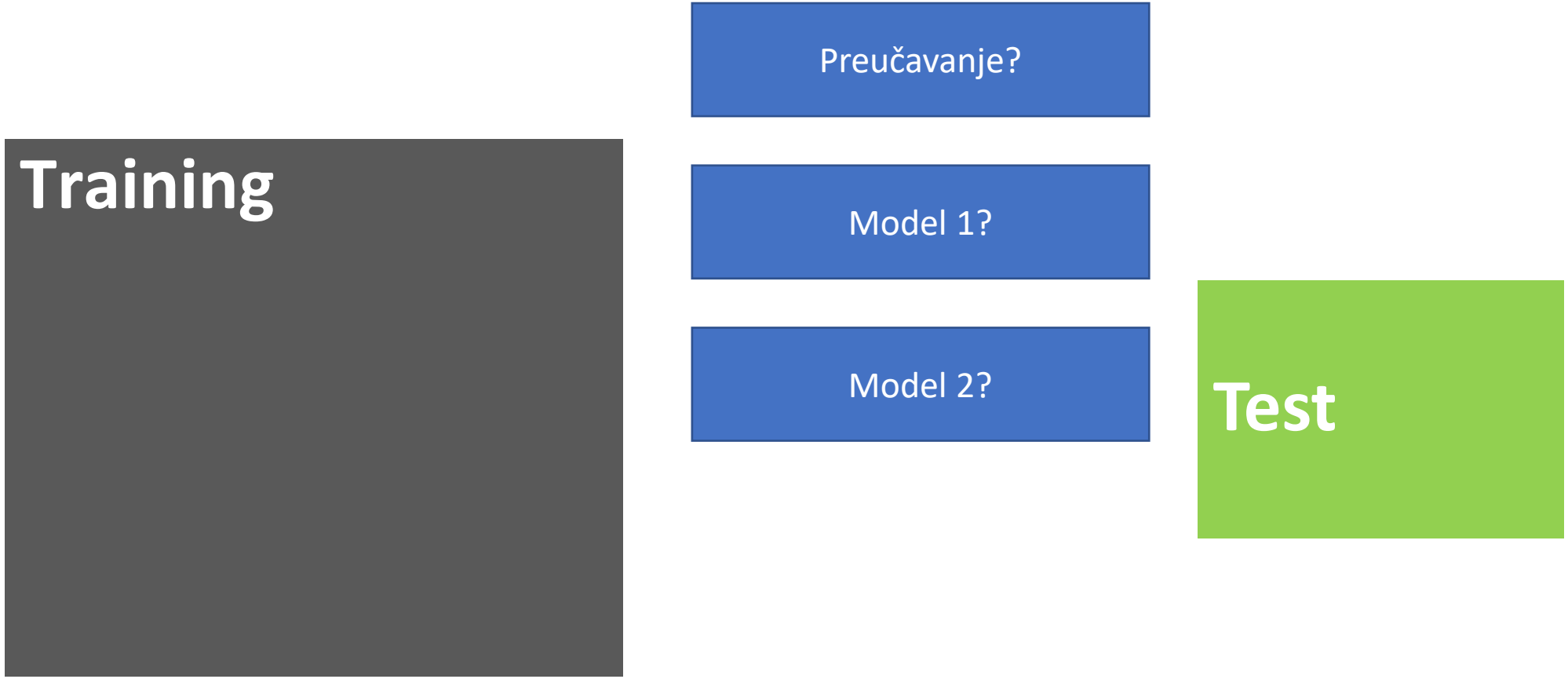


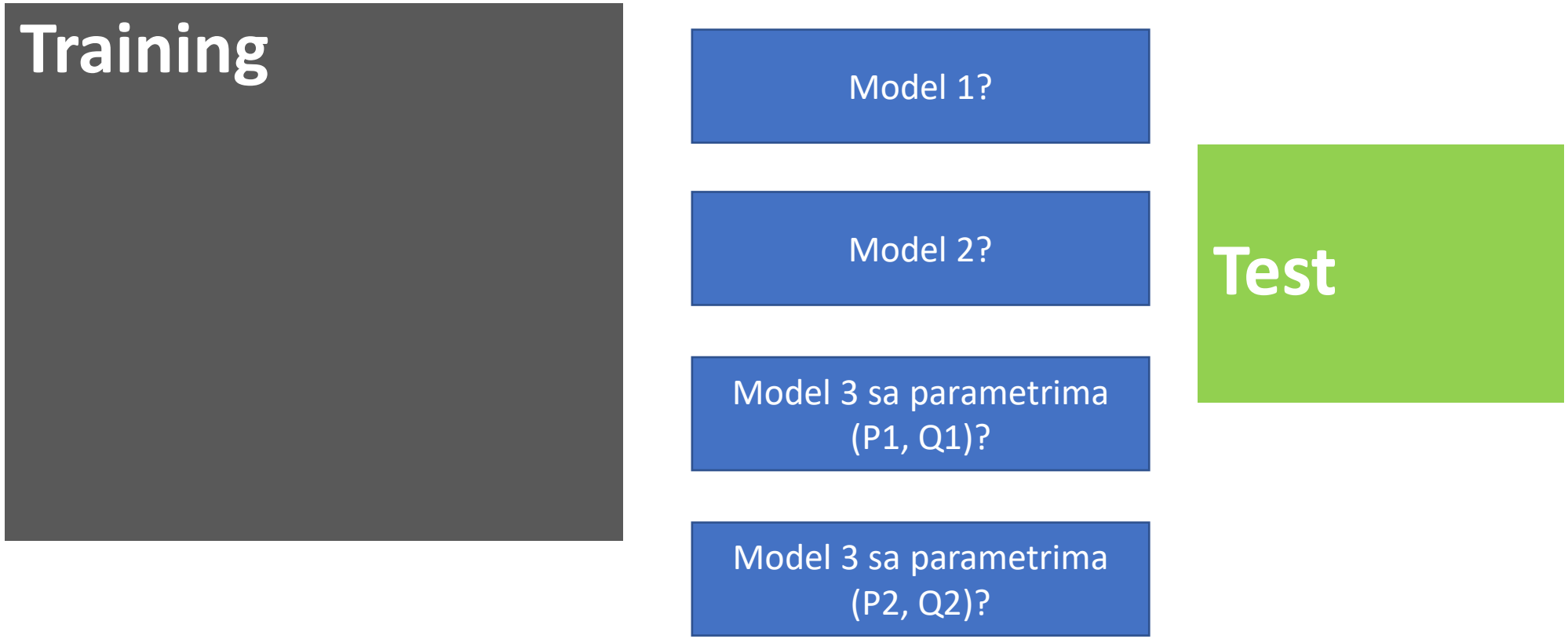


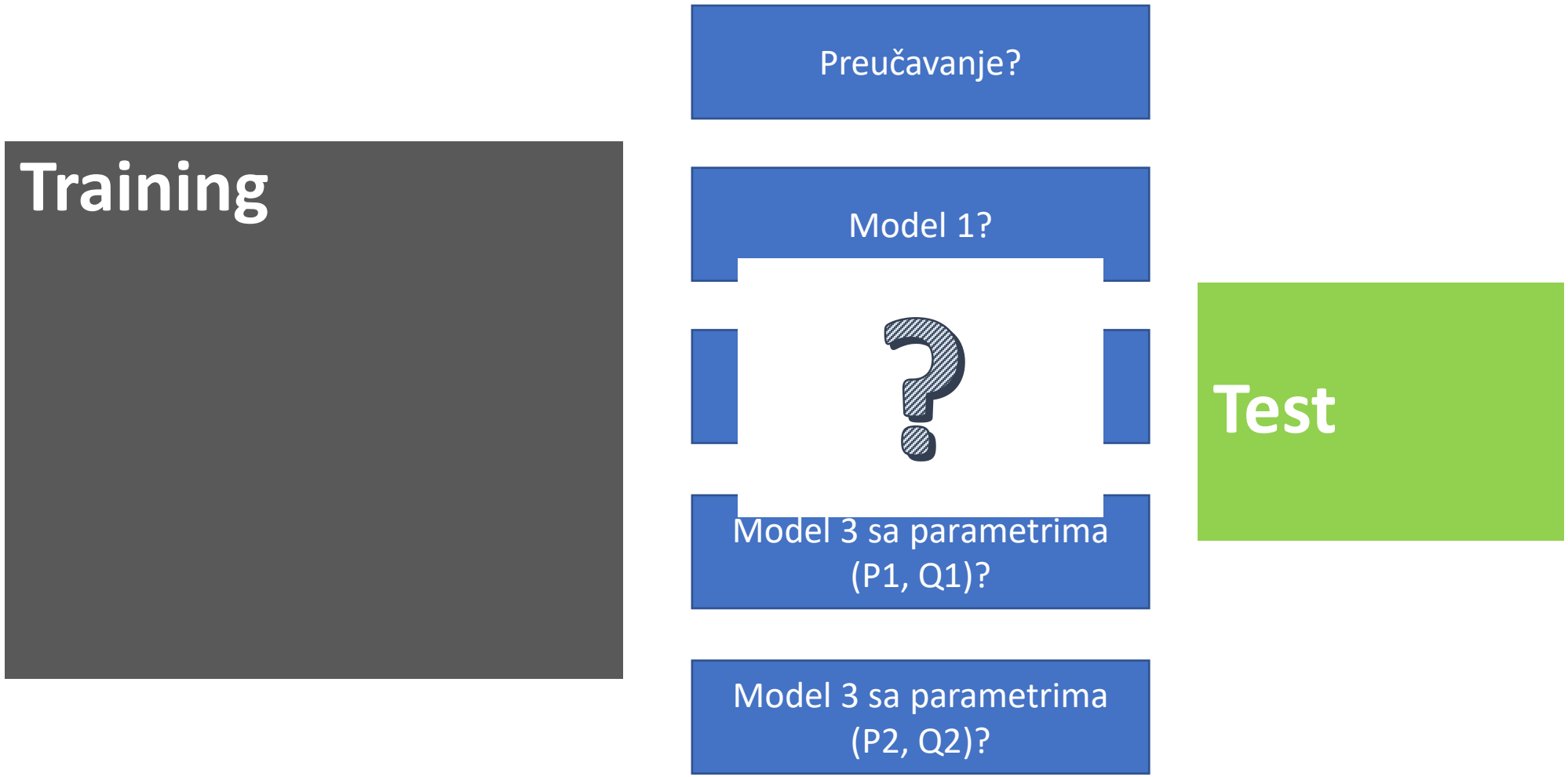
Preučavanje?

Training

Test


Training
Preučavanje?
Model 1?
Model 2?
Test


Training
Preučavanje?
Model 1?
Model 2?
Model 3 sa parametrima (P1, Q1)?
Model 3 sa parametrima (P2, Q2)?
Test

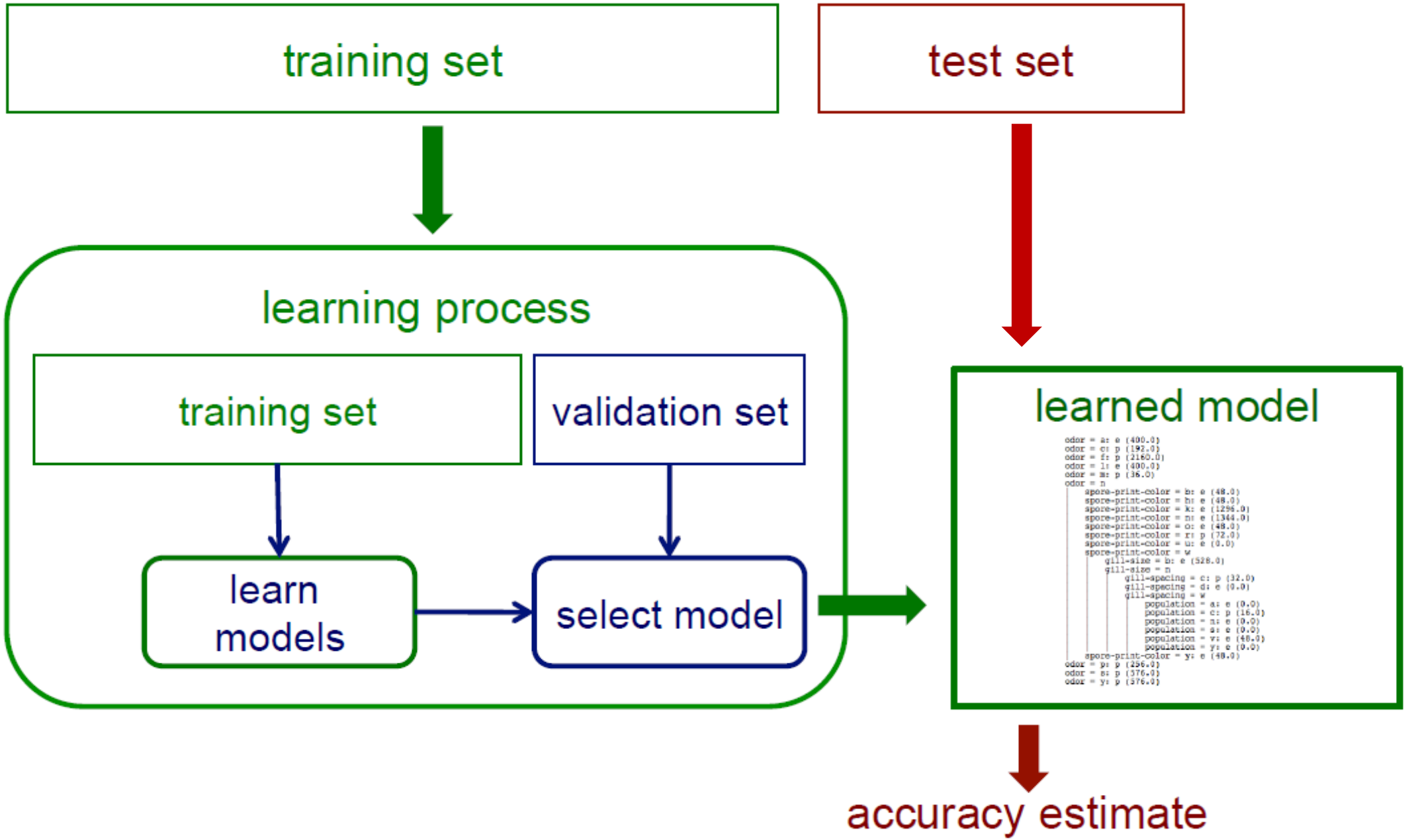

Training
Preučavanje?
Model 1?
?
Model 3 sa parametrima (P1, Q1)?
Model 3 sa parametrima (P2, Q2)?
Test


Training
Validation
Test

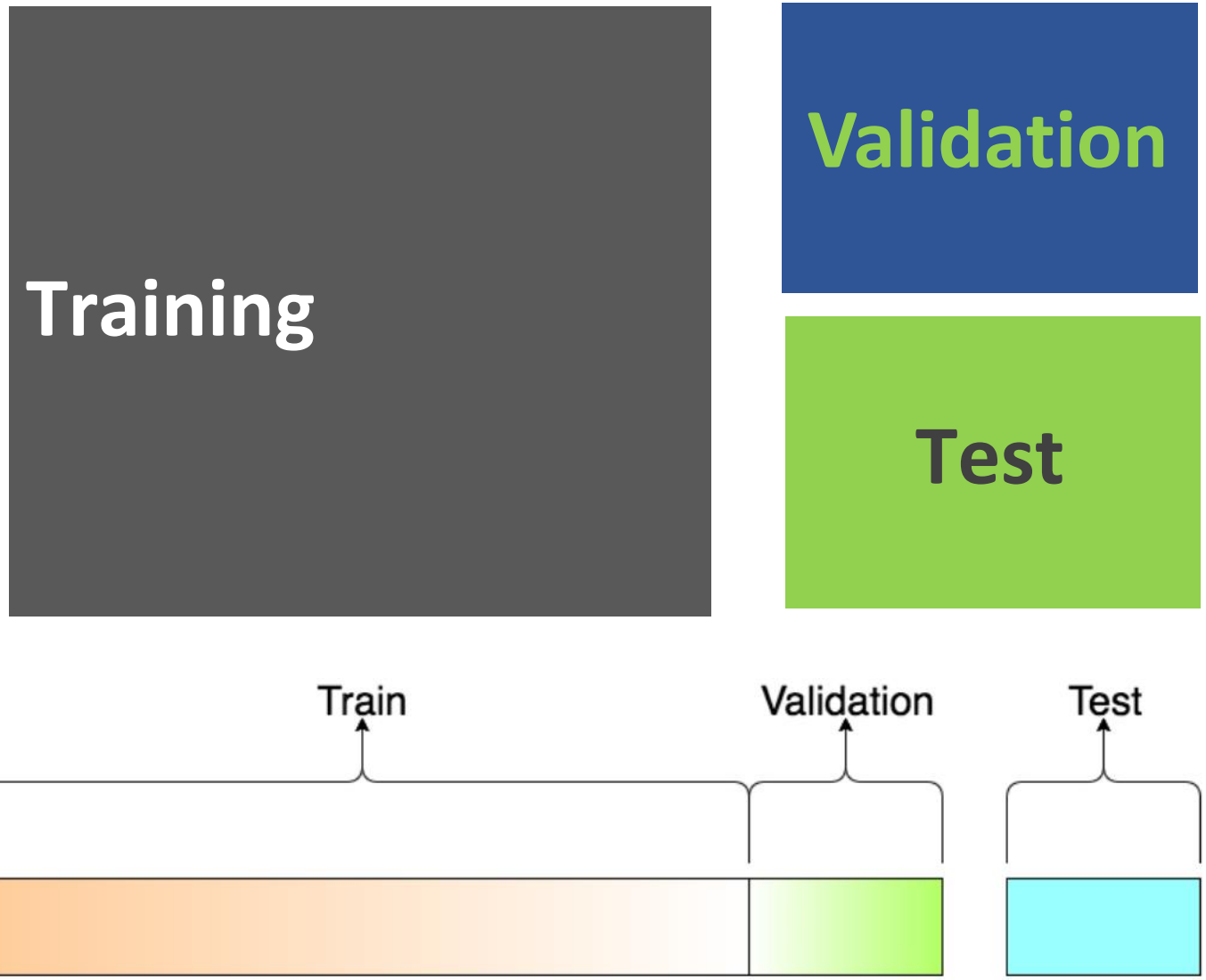
Greška na validacionom skupu

- Greška koju model pravi nad podacima za validaciju zove se greška validacije – **validation error**.

- Ona nam govori o sposobnosti našeg modela da vrši generalizaciju.


training set
test set
learning process
training set
validation set
learn models
select model
learned model
odor = a: e (400.0)
odor = c: p (192.0)
odor = f: p (2160.0)
odor = l: e (400.0)
odor = m: p (36.0)
odor = n
| spore-print-color = b: e (48.0)
| spore-print-color = h: e (48.0)
| spore-print-color = k: e (1296.0)
| spore-print-color = n: e (1344.0)
| spore-print-color = o: e (48.0)
| spore-print-color = r: p (72.0)
| spore-print-color = u: e (0.0)
| spore-print-color = w
| | gill-size = b: e (528.0)
| | gill-size = n
| | | gill-spacing = c: p (32.0)
| | | gill-spacing = d: e (0.0)
| | | gill-spacing = w
| | | | population = a: e (0.0)
| | | | population = c: p (16.0)
| | | | population = n: e (0.0)
| | | | population = s: e (0.0)
| | | | population = v: e (48.0)
| | | | population = y: e (0.0)
| spore-print-color = y: e (48.0)
odor = p: p (256.0)
odor = s: p (576.0)
odor = y: p (576.0)
accuracy estimate

Trening skup – Validacioni skup – Testni skup



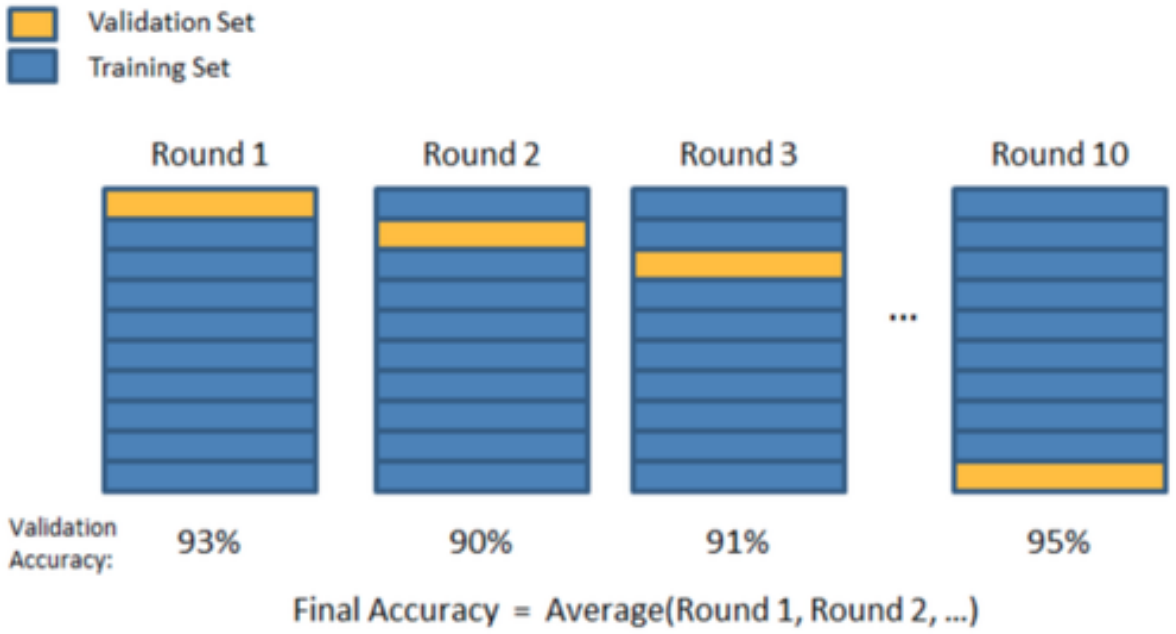
- Trening skup služi za treniranje modela.
- Validacioni skup služi za odabir konkretnog modela sa određenim parametarima.
- Testni skup se koristi za procenu performansi potpuno formiranog modela.

Dostupnost podataka za trening, validaciju i testiranje

- Ako je skup dostupnih podataka veliki, nema problema: uzmemo veliki uzorak i koristimo ga za obuku; zatim drugi, nezavisni uzorak za validaciju i treći uzorak različitih podataka za testiranje.

- Šta ako nemamo jako mnogo podataka?

k – unakrsna validacija

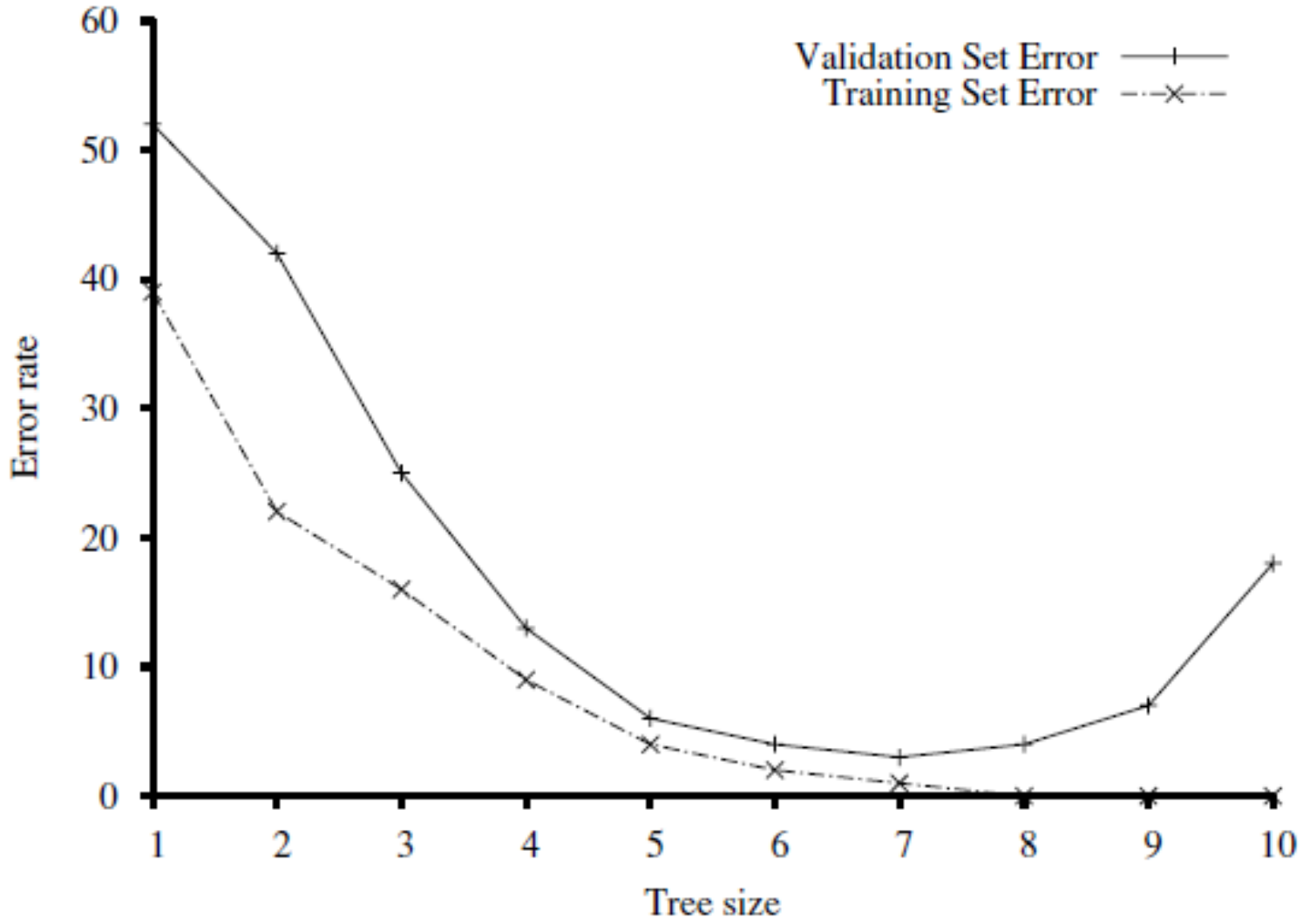


- Rešenje je unakrsna validacija.
- Početni skup podataka za obučavanje se deli na k delova ili podskupova.
  - Najčešće se uzima 10 podskupova.
- Potom se izvršava k tura učenja.
- U svakoj turi se $(k-1)/k$ primera koristi za obučavanje, a ostalih $1/k$ primera se koristi z izračunavanje stope greške.
- Na kraju se ukupna greška izračunava kao prosečna greška u prethodnih k slučajeva.

Primer rezultata dobijenih 10-fold validacijom

Fold	*Naive Bayes*	*Decision tree*	*Nearest neighbour*
1	0.6809	0.7524	0.7164
2	0.7017	0.8964	0.8883
3	0.7012	0.6803	0.8410
4	0.6913	0.9102	0.6825
5	0.6333	0.7758	0.7599
6	0.6415	0.8154	0.8479
7	0.7216	0.6224	0.7012
8	0.7214	0.7585	0.4959
9	0.6578	0.9380	0.9279
10	0.7865	0.7524	0.7455
avg	0.6937	0.7902	0.7606
stdev	0.0448	0.1014	0.1248

Odabir modela sa najboljim parametrima



„Ostavi-jednog-van“ unakrsna validacija

- **Leave-one-out cross-validation** je *n*-unakrsna validacija, gde je *n* broj primera u trening skupu.
- Jedan po jedan primer „ostavljamo sa strane“ a učimo model na svim ostalim primerima.
- Tačnost modela se ocenjuje na primeru koji je ostavljen.
- Greška modela se izračunava kao prosečna vrednost greške na svih *n* primera.

Ostavi-jednog-van unakrsna validacija: prednosti

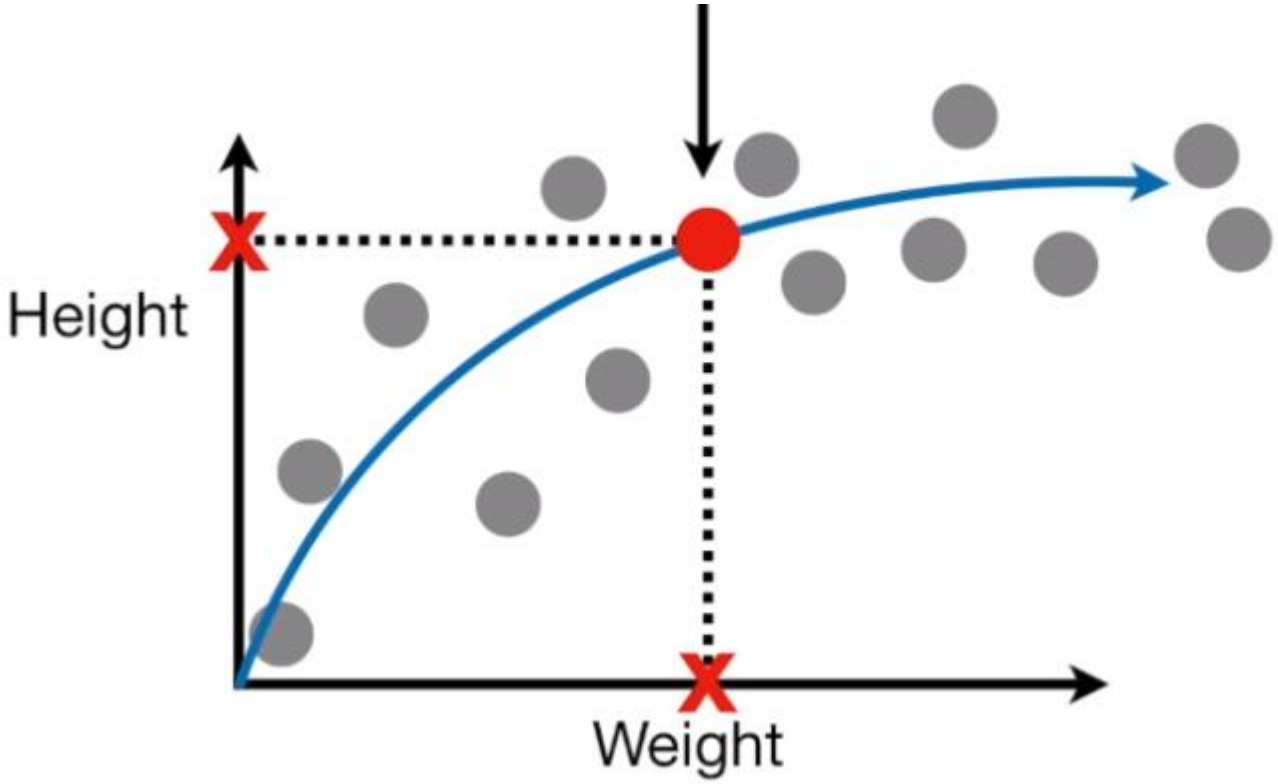
- Najveća moguća količina podataka se koristi za obuku, što verovatno povećava šansu da je klasifikator tačan.
- „Ostavi-jednog-van“ pruža šansu da se iz malog skupa podataka istisne maksimum i dobije se što preciznija procena.
- Postupak je deterministički: nije uključeno slučajno uzorkovanje.
- Nema smisla ponavljati ga 10 puta ili ga uopšte ponavljati: svaki put će se dobiti isto.

Ostavi-jednog-van unakrsna validacija: mane

- Visoki troškovi izračunavanja, jer se celokupni postupak učenja mora izvesti n puta, a to je obično prilično neizvodljivo za velike skupove podataka.

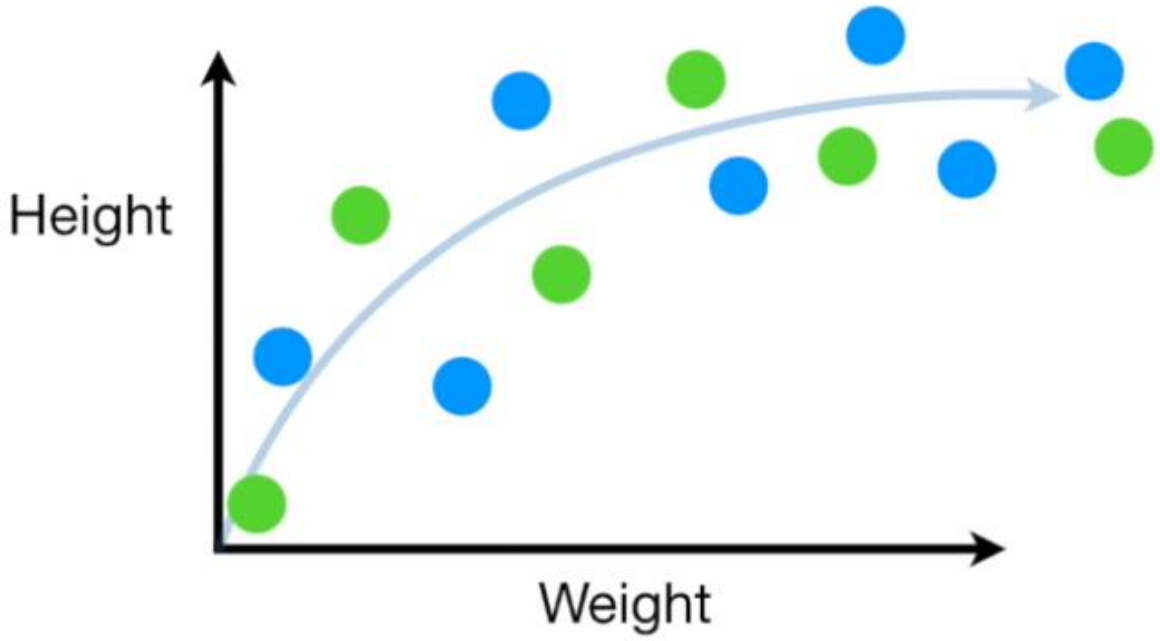
- Nestratifikovani uzorak. Stratifikacija uključuje unošenje tačnog udela primera iz svake klase u validacioni skup, a to je nemoguće kada validacioni skup sadrži samo jedan primer.

Primer



- Da bi smo na osnovu težine mogli da predvidimo visinu, najbolje je da znamo matematičku formulu.

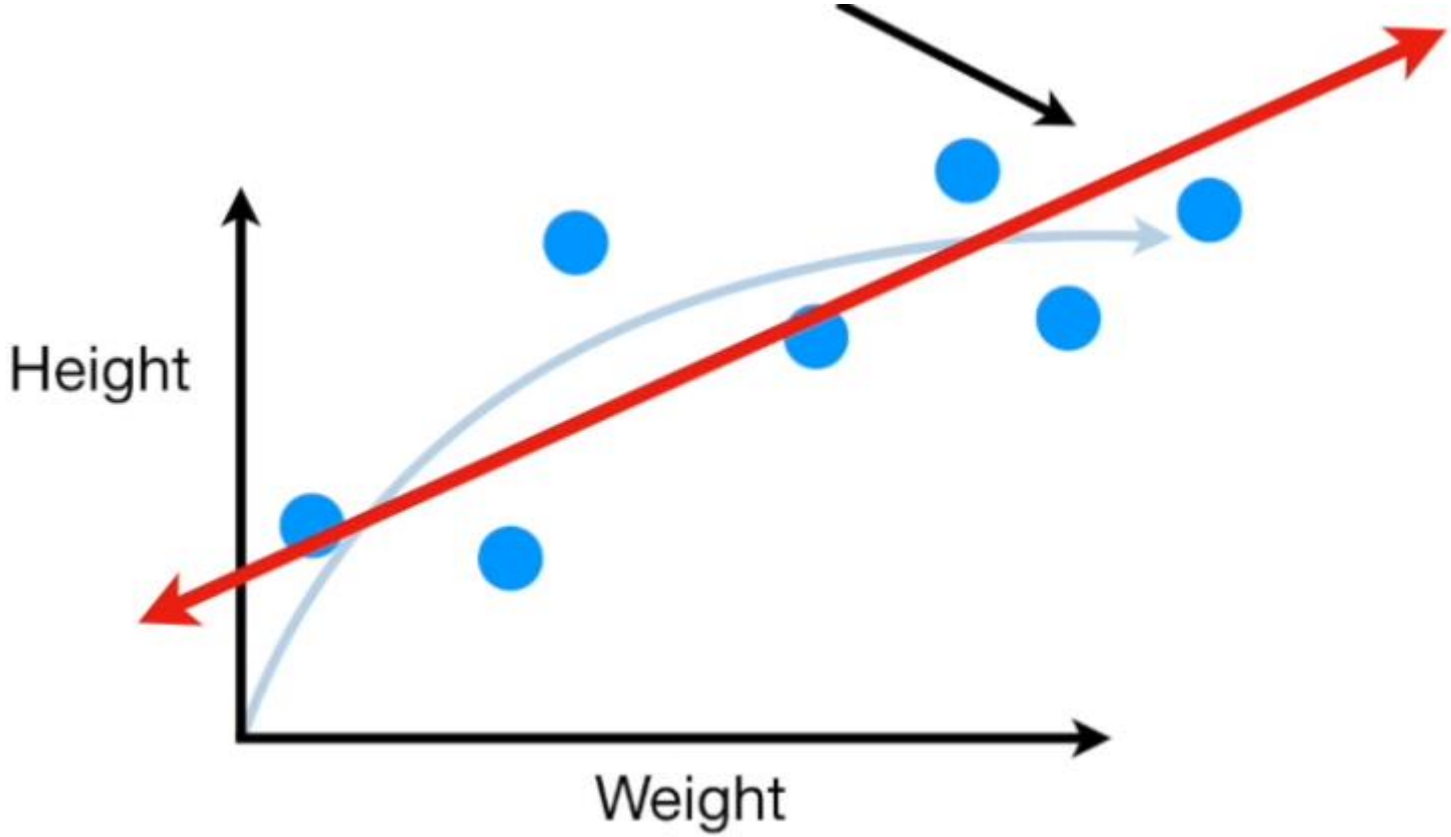
Primer



- Pošto je ne znamo, koristimo mašinsko učenje.
- Skup primera koji imamo delimo na:
  - **Skup za treniranje**
  - **Skup za validaciju**

Primer

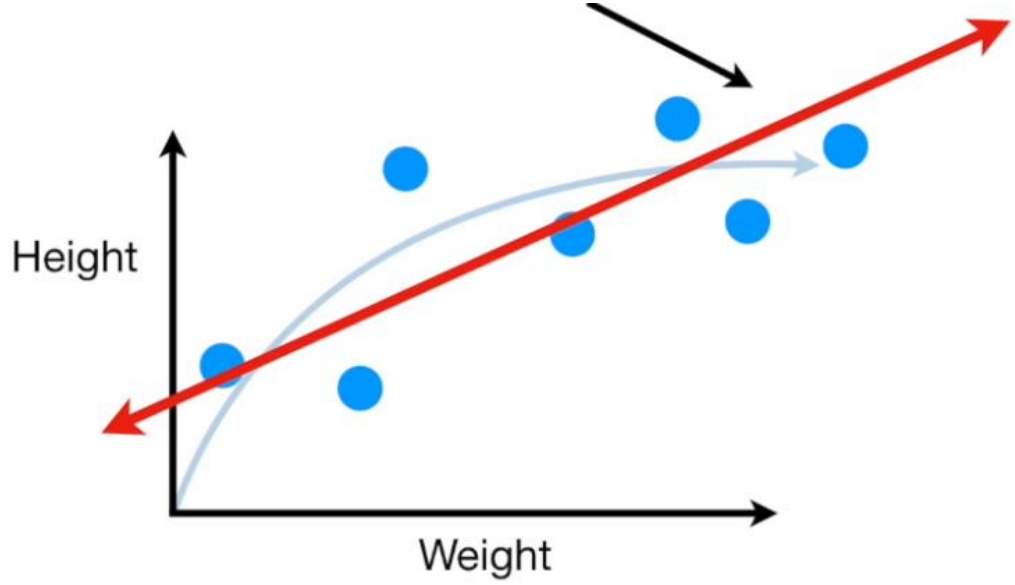
- Vrsta modela mašinskog učenja koja fituje trening skup pravom linijom - linearna regresija



Primer

- Svaki pokušaj da pravu liniju prilagodimo ovakvom trening skupu bi bio neuspešan.

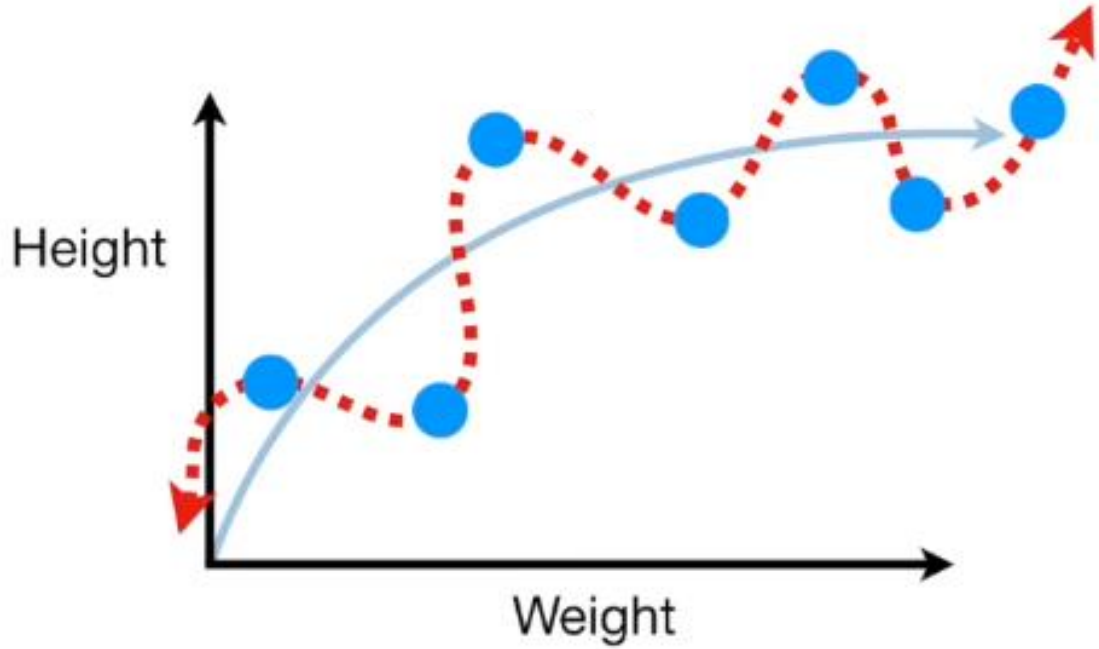
- Prava linija nema fleksibilnost neophodnu da se predstavi prava veza između ulaza i izlaza.



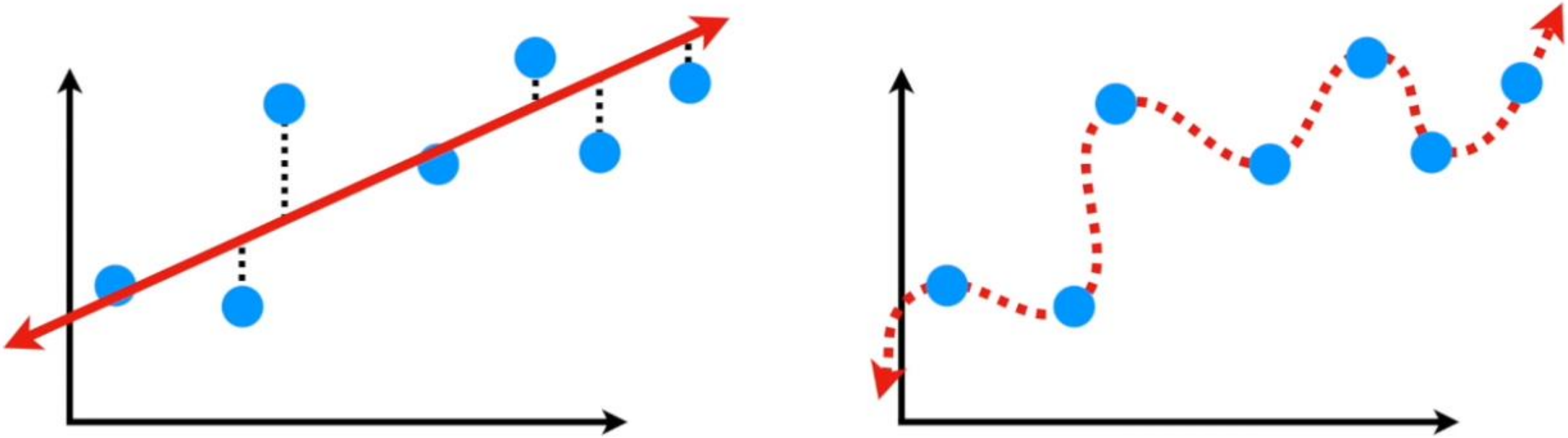
Primer

- Neka druga vrsta mašinskog učenja fituje trening skup krivudavom linijom.

- Linija potpuno tačno oslikava ulaz-izlaz relaciju trening skupa i ima nizak bias.



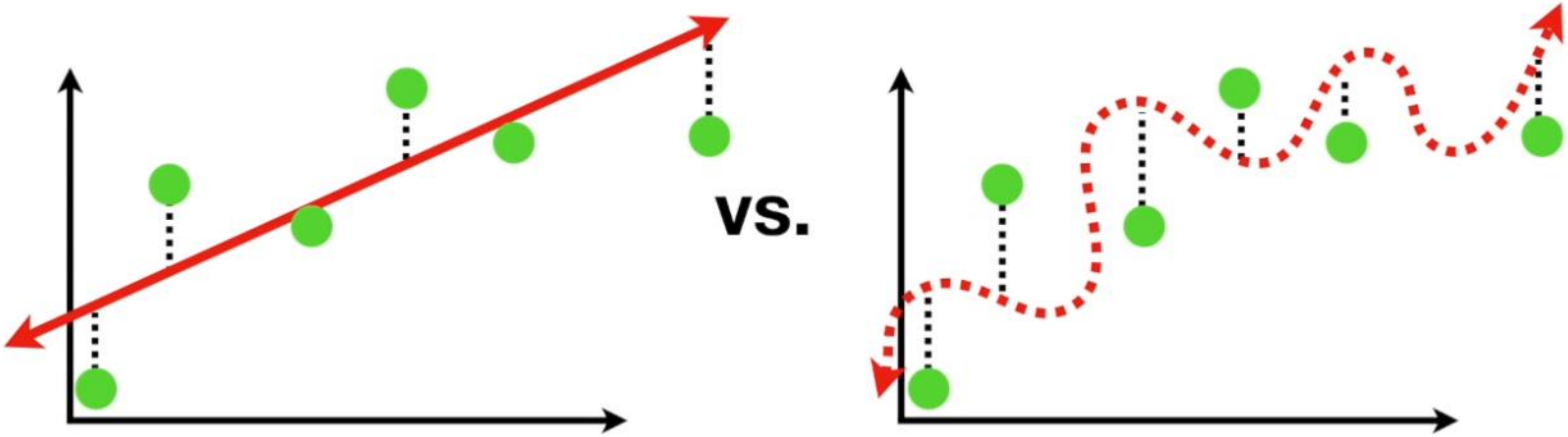
Primer



- Metodom sume kvadrata rastojanja tačaka trening skupa od tačaka modela dobijenog učenjem može se zaključiti da krivudava linija ima sumu kvadrata 0, i da perfektno odgovara trening skupu

Primer

- Ako istu metodu primenimo na skup tačaka za testiranje, krivudava linija se pokazuje kao lošije rešenje.
- Model koji ne ispoljava grešku na trening podacima, a pravi veliku grešku na validacionom skupu je preprilagođen trening skupu (preučen).



Greska modela na test skupu

- $E = Bias^2 + Šum + Varijansa$

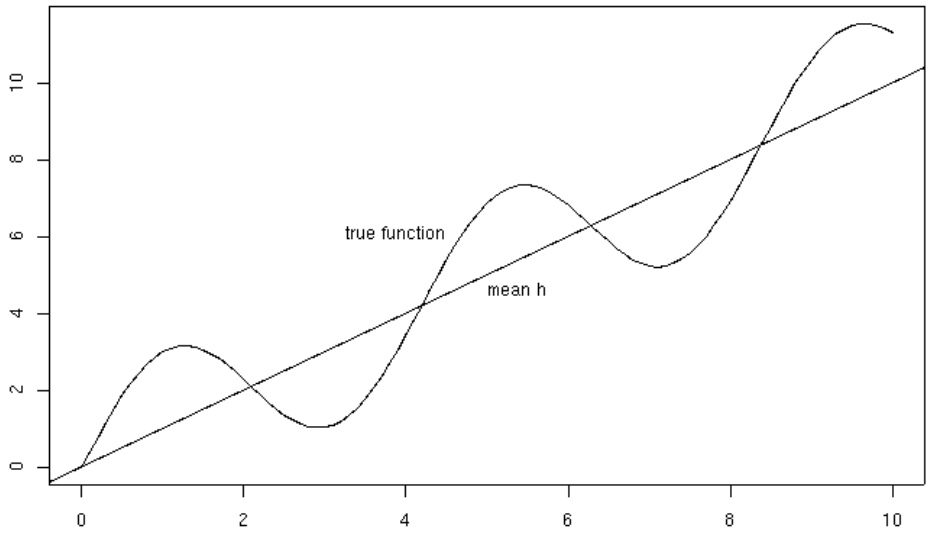
Bias

- **Bias** se odnosi na grešku modela koja nastaje kao posledica uprošćavanja pretpostavki ili pogrešnih pretpostavki koje model pravi u odnosu na podatke za učenje kako bi što lakše naučio ciljnu funkciju.

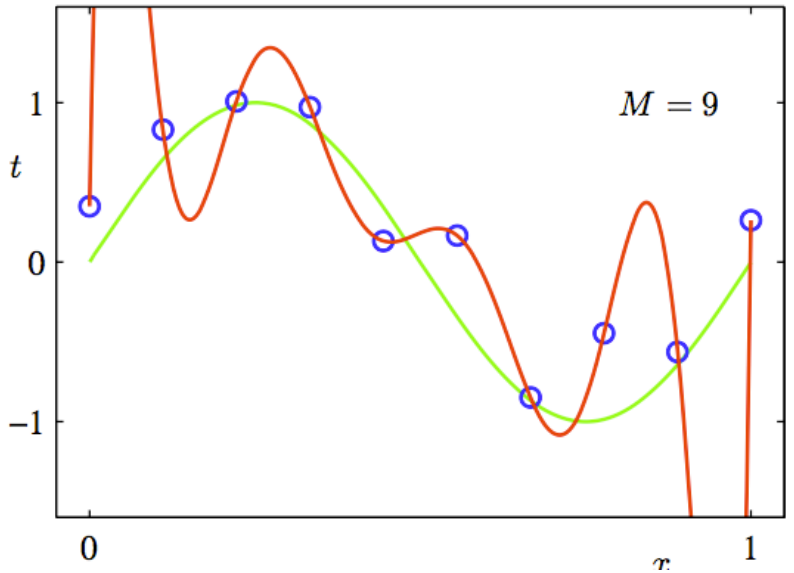
- **Visok bias** sugeriše da algoritam ne uočava bitne trendove među podacima.

- Visok bias se može umanjiti prelaskom na kompleksniji model, dodavanjem atributa ili boosting-om.

High Bias: Model is not powerful enough to learn function. Learns a less-complicated function (*underfitting*).



Low Bias: Model is too powerful; learns a too-complicated function (*overfitting*).



Varijansa

- Varijansa oslikava preprilagođenost modela trening skupu.
- Meri odstupanje predviđanja našeg modela od očekivanog predviđanja.
- Visoka varijansa se može umanjiti odabirom manje složenog modela, povećanjem trening skupa ili bagging-om.

Šum

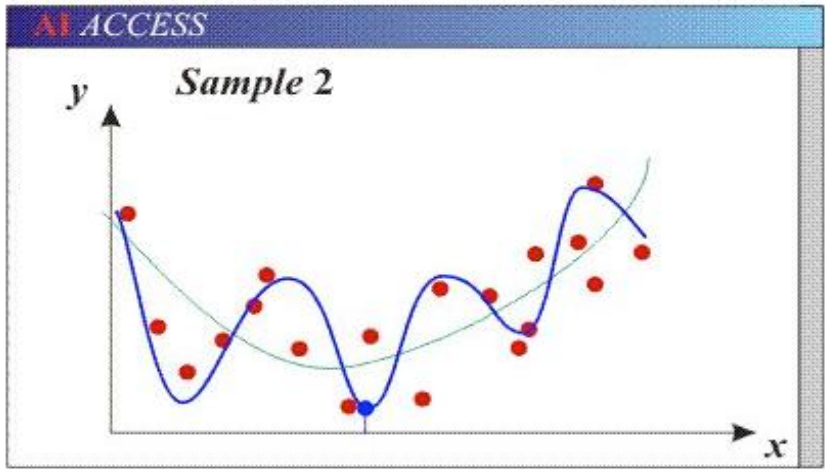
- Šum meri dvosmislenosti koje proističu iz raspodele podataka.

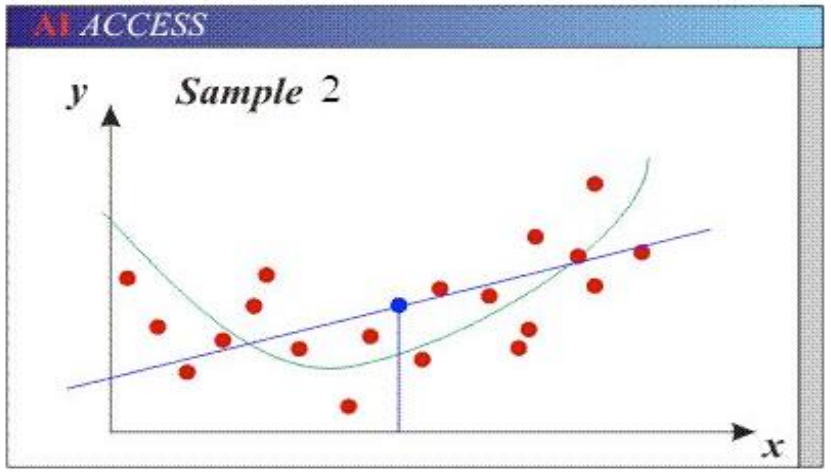
- Šum nije odlika modela, već podataka i ne može se smanjiti intervencijom nad modelom, već nad podacima.

Bias-variance tradeoff

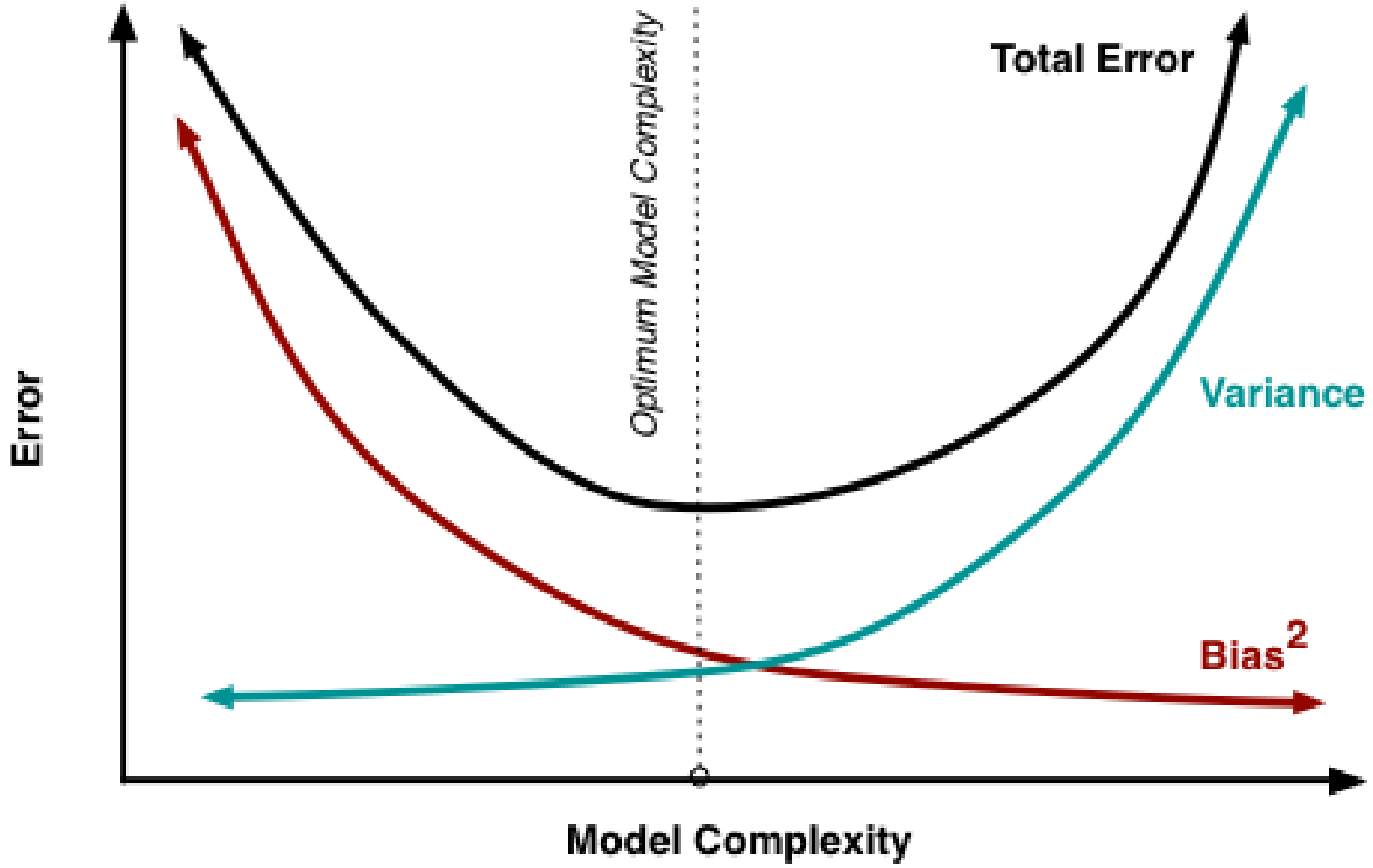
- Modeli sa mnogo parametara se dobro prilagođavaju podacima (nizak bias), ali su osetljivi na odabir trening skupa (visoka varijansa)
- Greška usled overfittinga.

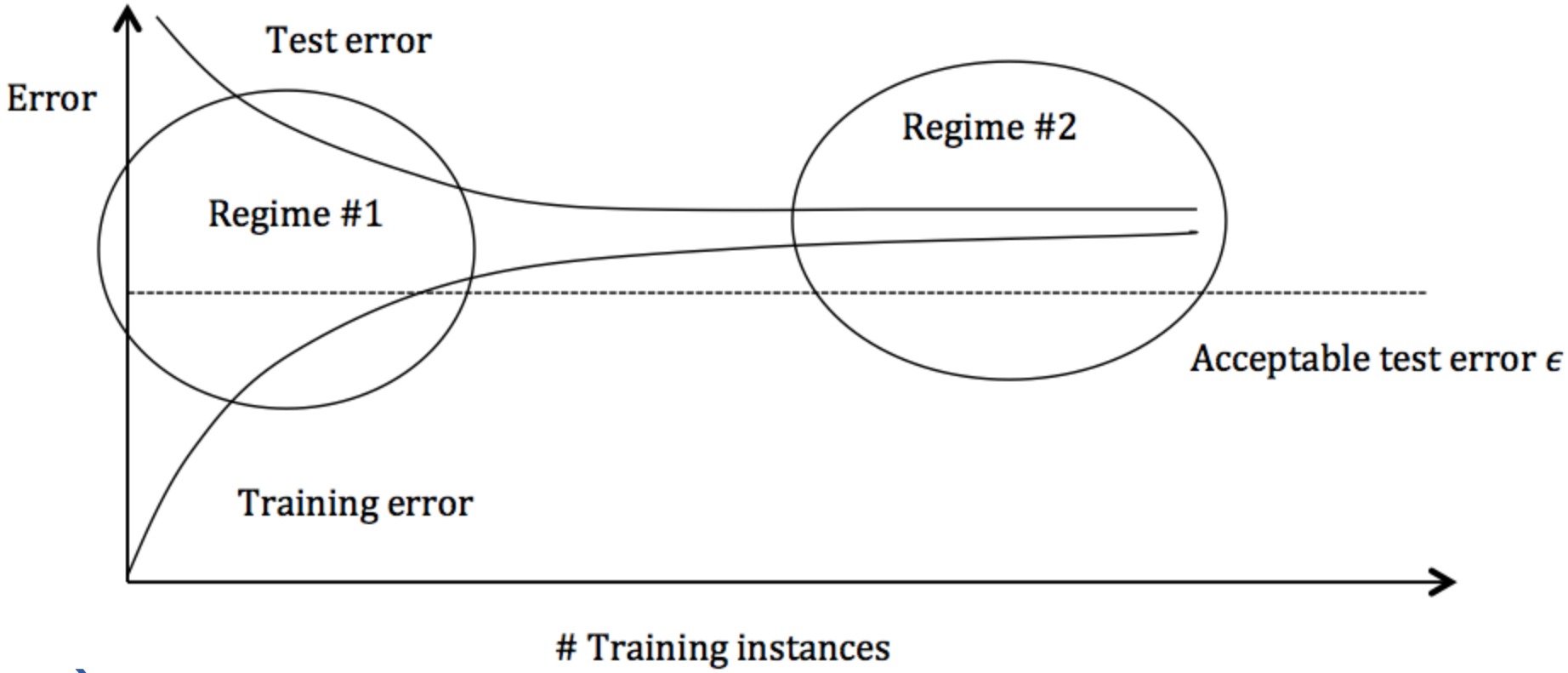
- Modeli sa manje parametara se lošije prilagođavaju podacima (visok bias), ali su nepromenljivi nad različitim trening skupovima (niska varijansa).
- Greška usled underfitting-a.





Kada smanjujemo varijansu bias raste i obrnuto. Potrebno je uspostaviti balans između ove dve veličine

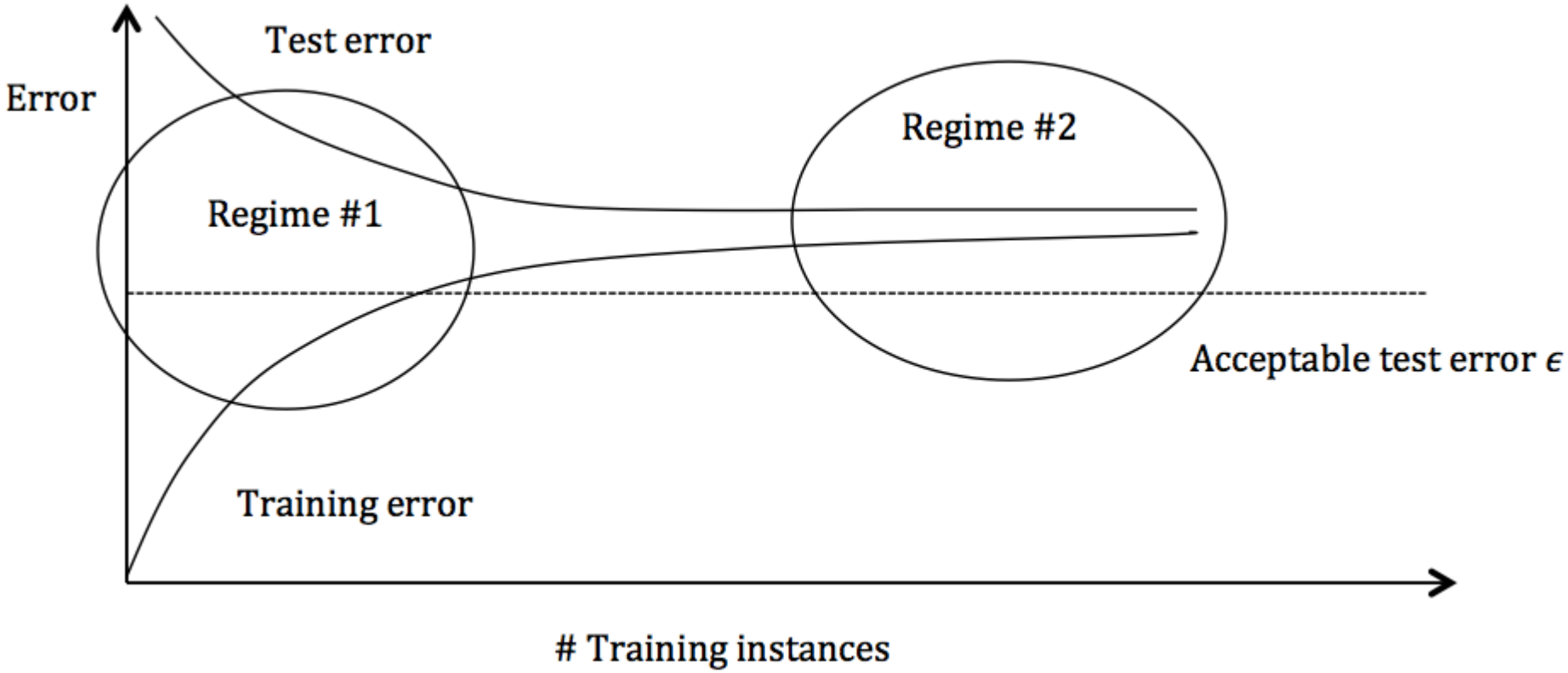




Regime 1 (High Variance)

Symptoms:
1.Training error is much lower than test error
2.Training error is lower than єє
3.Test error is above єє

Remedies:
•Add more training data
•Reduce model complexity - complex models are prone to high variance
•Bagging



Regime 2 (High Bias)

Unlike the first regime, the second regime indicates high bias: the model being used is not robust enough to produce an accurate prediction.

Symptoms:

1.Training error is higher than ϵ

Remedies:

- Use more complex model (e.g. use non-linear models)
- Add features
- Boosting