# Strojno učenje u službi laboratorija

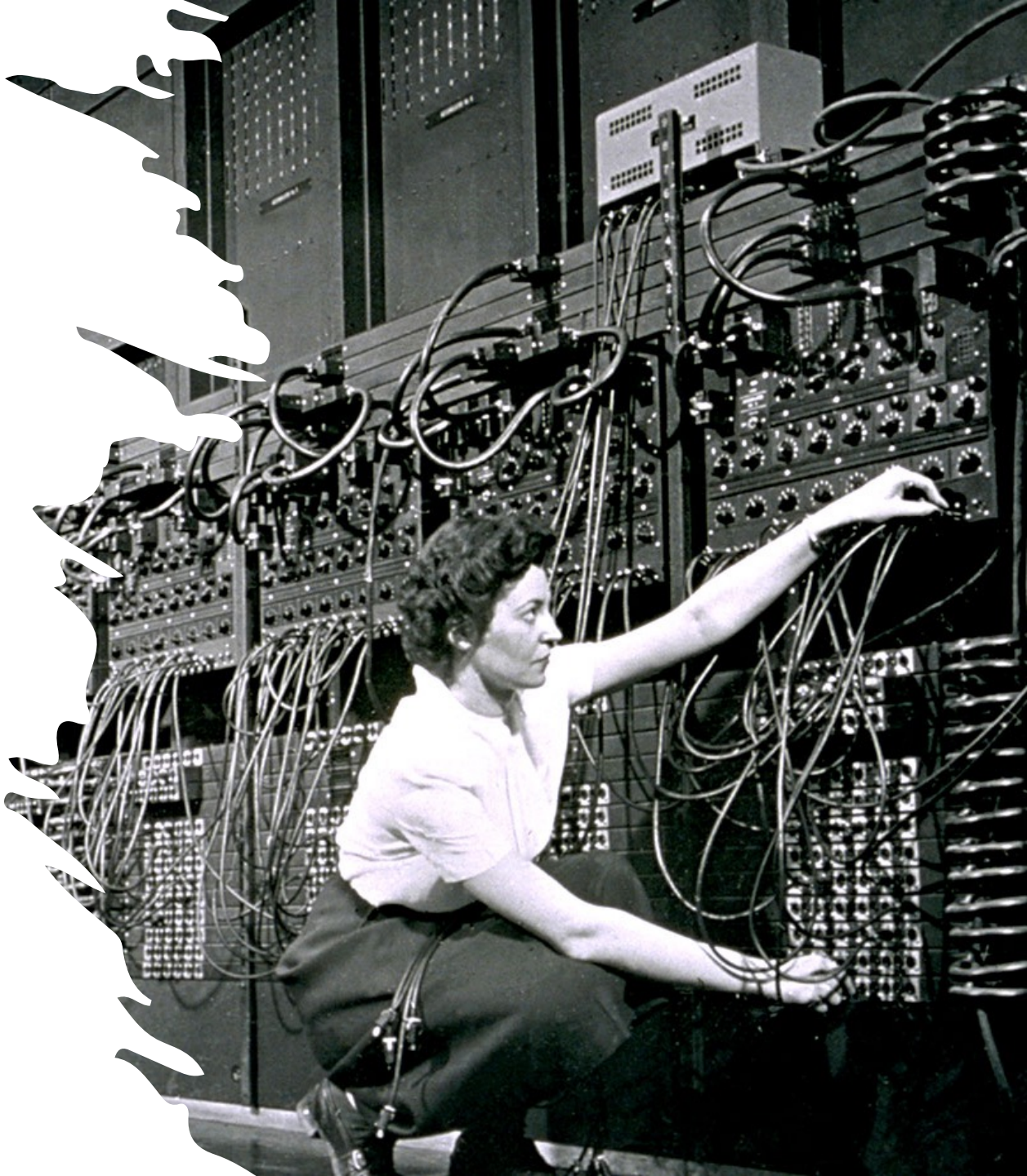
Zvonko Kostanjčar, Stjepan Begušić, Andro Merćep

Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva
Laboratorij za analitiku financija i rizika

Laboratory for Financial and Risk Analytics

# Strojno učenje

- Strojno učenje – računalo može učiti iz iskustva
- Algoritam strojnog učenja sastoji se od
  - Modela
  - Funkcije gubitka
  - Optimizacijskog postupka
- Usko je povezano s područjima
  - Statistika – zaključivanje o populaciji na temelju uzorka
    - često prisutne jake pretpostavke na populaciju
  - Umjetna inteligencija – sposobnost računala da izvodi zadaće koje uobičajeno pripisujemo inteligentnim bićima



# Povijest umjetne inteligencije

1940:
Prva računala i
računalni programi

1960:
Prvo računalo u
kontekstu AI-a (GPS)

1980:
Ekspertni sustavi

1990:
Inteligentni agenti,
problem-spec. AI

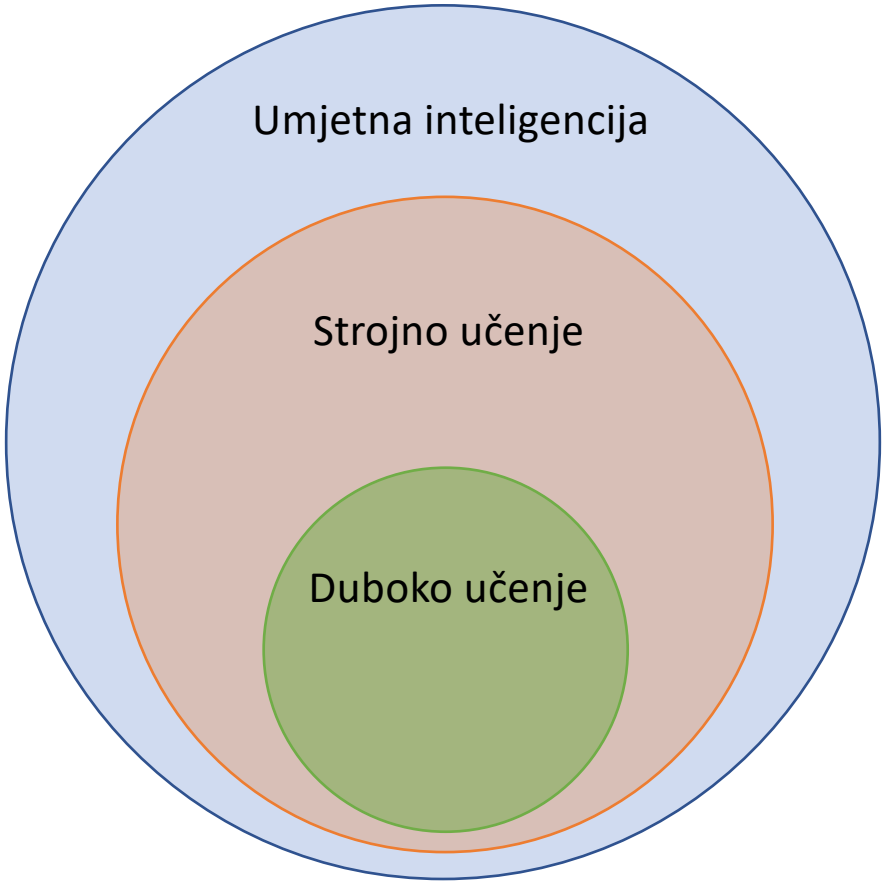
1956:
Umjetna inteligencija

1970:
Prva AI zima

Kraj 1980:
Druga AI zima

2010:
Duboko učenje

# Umjetna inteligencija – strojno učenje – duboko učenje



# Strojno učenje – preciznije

Strojno učenje – računalni program uči iz iskustva E (engl. *experience*) s obzirom na neku klasu zadataka T (engl. *task*) i mjere uspješnosti P (engl. *performance*) ako se njegova uspješnost u zadatcima T (mjerena mjerom P) popravlja s iskustvom E.

Primjer

- **zadatak T**: prepoznati multipli mijelom iz biokemijskih rezultata pacijenta
- **mjera P**: postotak točno klasificiranih pacijenata (ima, odnosno nema, multipli mijelom)
- **iskustvo E**: skup podataka biokemijskih rezultata pacijenata i njihovih dijagnoza (svaki pacijent je točno označen)

# Osnovni koncepti u strojnom učenju

- Iskustvo (E) – podaci
- Zadatak (T) – pristupi strojnom učenju
  - Nadzirano učenje
    - regresija
    - klasifikacija
  - Nenadzirano učenje
    - grupiranje podataka
    - redukcija dimenzionalnosti
    - procjena gustoće
  - Podržano učenje
- Mjera uspješnosti (P) – ovisi o zadatku

# Podaci

Slike

inese NORP market has the three CARDINAL most influential names of the retail and te
nd Tencent PERSON (collectively touted as BAT ORG ), and is betting big in the globa
. The three CARDINAL giants which are claimed to have a cut-throat competition with th
capital) are positioning themselves to become the 'future AI PERSON platforms'. The tri
countries and investing heavily in the U.S. GPE based AI GPE startups to leverage th
powerful initiatives and presence of these conglomerates, the market in APAC AI is for
CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE

orate on the geographical trends, North America LOC has procured more than 50%
and has been leading the regional landscape of AI GPE in the retail market. The U.S
gional trends with over 65% PERCENT of investments (including M&As, private equity, a
ence technology. Additionally, the region is a huge hub for startups in tandem with the
e ORG , IBM ORG , and Microsoft ORG .

Tekst

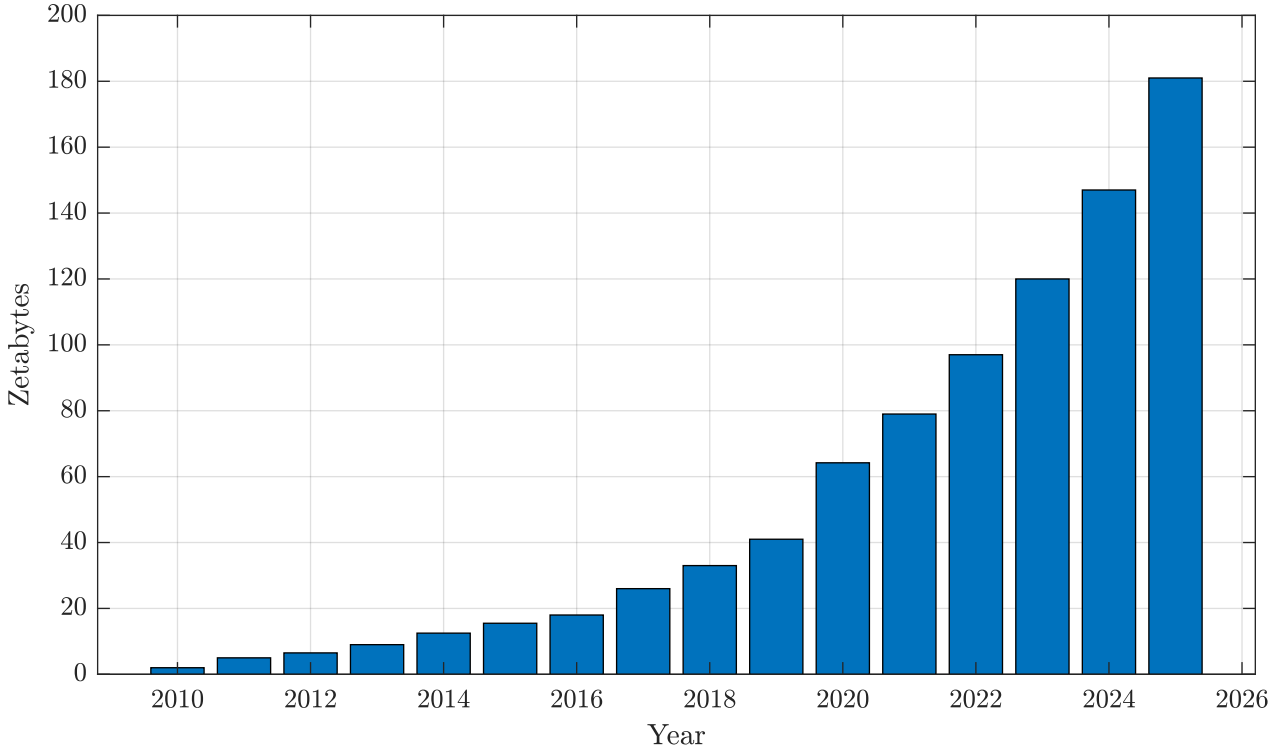

Signali – vremenski nizovi

| Age | Sex | BMI | Children | Smoker | Region | Expenses |
|---|---|---|---|---|---|---|
| 19 | female | 27.9 | 0 | yes | southwest | 16,884.92 |
| 28 | male | 33.0 | 3 | no | southeast | 4,449.46 |
| 32 | male | 28.9 | 0 | no | northwest | 3,866.86 |
| 46 | female | 33.4 | 1 | no | southeast | 8,240.59 |
| 37 | male | 29.8 | 2 | no | northeast | 6,406.41 |
| 25 | male | 26.2 | 0 | no | northeast | 2,721.32 |
| 23 | male | 34.4 | 0 | no | southwest | 1,826.84 |
| 27 | male | 42.1 | 0 | yes | southeast | 39,611.76 |
| 52 | female | 30.8 | 1 | no | northeast | 10,797.34 |
| 56 | male | 40.3 | 0 | no | southwest | 10,602.39 |
| 60 | female | 36.0 | 0 | no | northeast | 13,228.85 |
| 18 | male | 34.1 | 0 | no | southeast | 1,137.01 |
| 37 | male | 28.0 | 2 | no | northwest | 6,203.90 |
| 63 | female | 23.1 | 0 | no | northeast | 14,451.84 |
| 23 | male | 17.4 | 1 | no | northwest | 2,775.19 |
| 22 | male | 35.6 | 0 | yes | southwest | 35,585.58 |
| 19 | female | 28.6 | 5 | no | southwest | 4,687.80 |
| 28 | male | 36.4 | 1 | yes | southwest | 51,194.56 |
| 62 | female | 33.0 | 3 | no | northwest | 15,612.19 |
| 35 | male | 36.7 | 1 | yes | northeast | 39,774.28 |
| 24 | female | 26.6 | 0 | no | northeast | 3,046.06 |
| 41 | male | 21.8 | 1 | no | southeast | 6,272.48 |
| 38 | male | 37.1 | 1 | no | northeast | 6,079.67 |
| 18 | female | 38.7 | 2 | no | northeast | 3,393.36 |
| 60 | female | 24.5 | 0 | no | southeast | 12,629.90 |

Tablični podaci (numerički/kategorijski)

# Podaci

U zadnje dvije godine je stvoreno 90% podataka



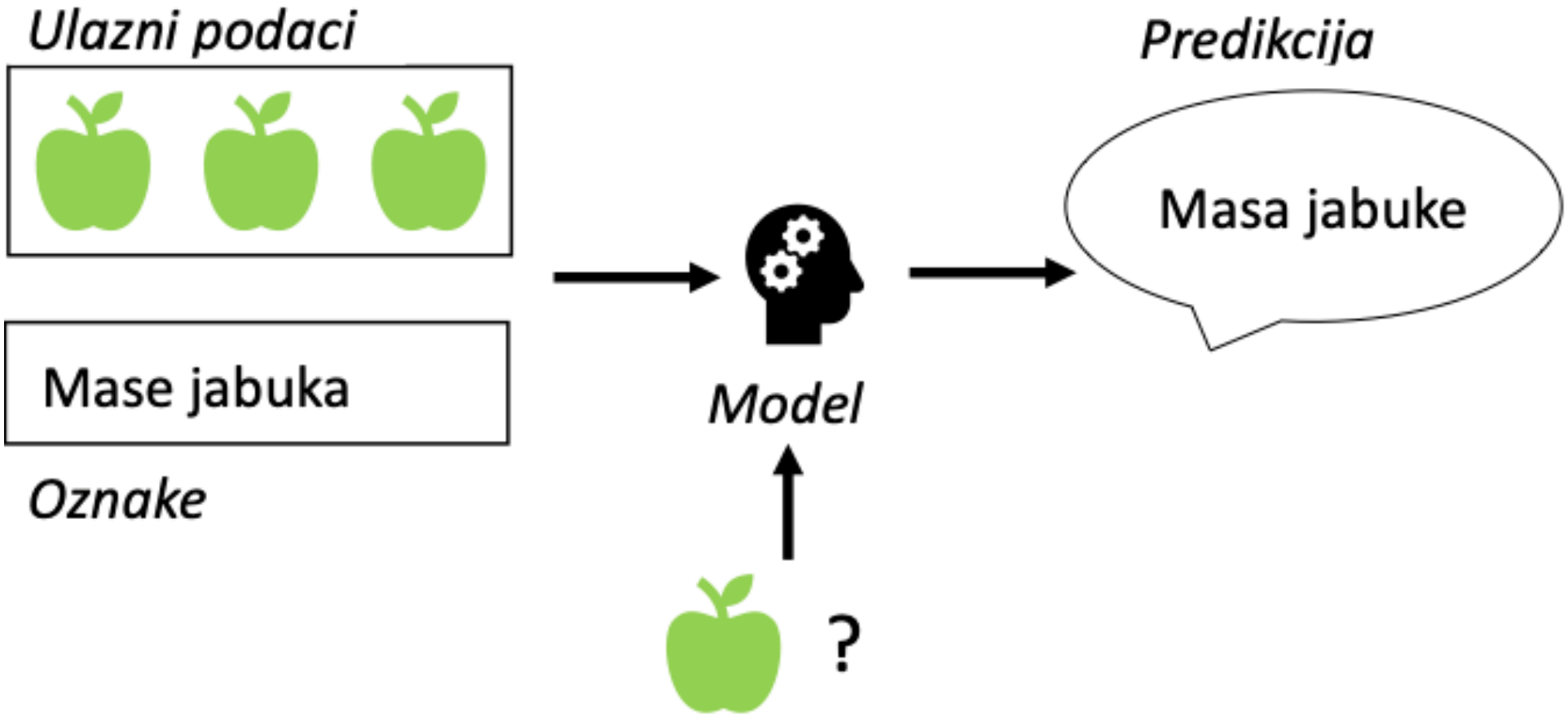

Količina podataka po godinama

# Podaci su nova “nafta”

# Podaci

- Uređivanje podataka i predobrada
  - Nedostajući podaci
  - Pogrešni podaci (pogotovo kod ručnog unosa)
- Izvođenje novih značajki
  - Sirovi podaci često nisu dovoljno dobri
  - Izgradnja varijabli / značajki koje su informativnije za danu primjenu
  - Često korisno domensko znanje
- Deskriptivna statistika

# Pristupi strojnom učenju – nadzirano učenje – regresija



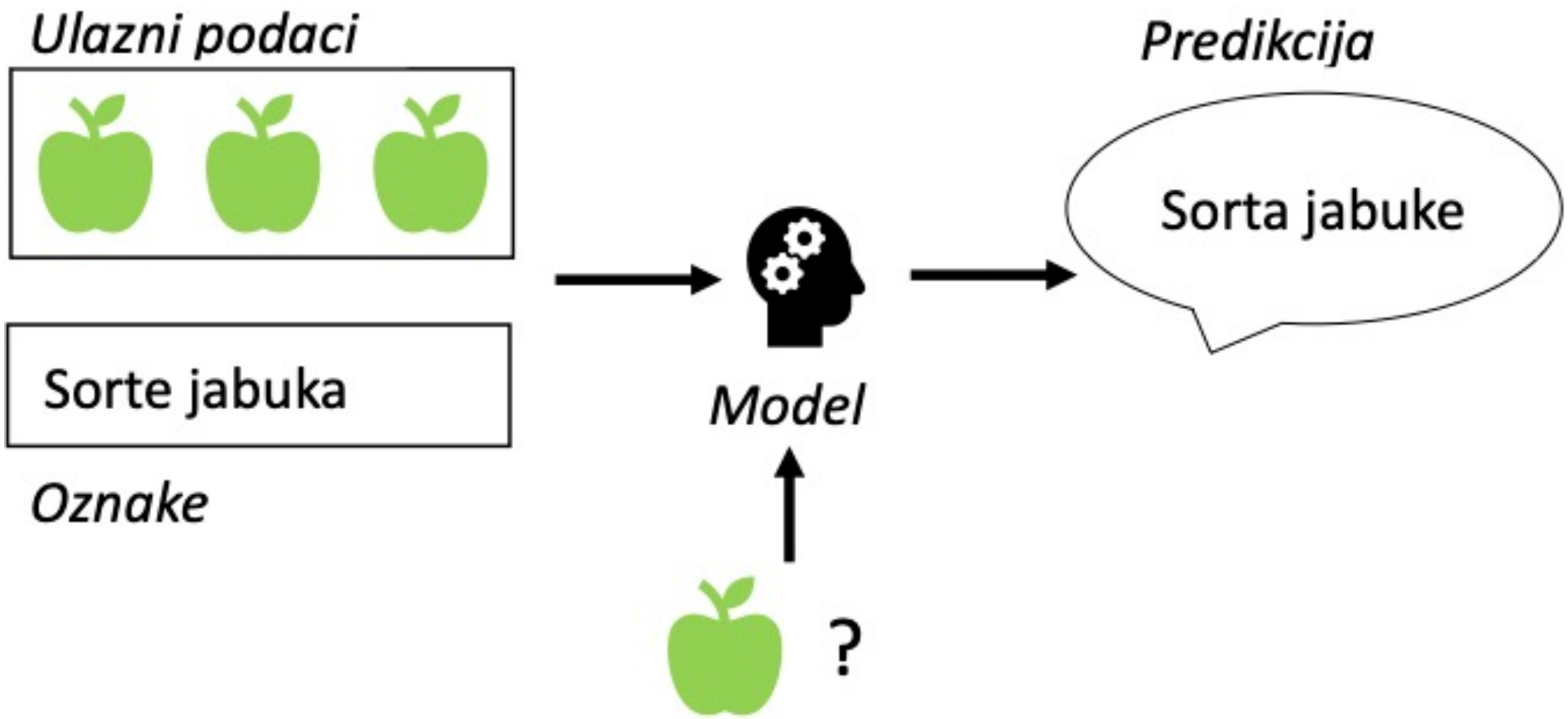
# Pristupi strojnom učenju – nadzirano učenje - regresija

- Podaci – označeni
  - $x_n$ - vektor značajki (regresori)
  - $y_n$ - ciljna vrijednost (zavisna varijabla, odziv, …), metrička, često kontinuirana varijabla

- Zadatak: korištenjem skupa za učenje s $N$ opservacija (ulaznih vektora $x_n$ uz pripadajuće oznake $y_n$) predvidjeti ciljnu vrijednost $y$ za novu ulaznu vrijednost $x$
  - Konkretno, tražimo funkciju $F$ (model), tako da je $y = F(x)$

- Mjera uspješnosti – ima ih više, često se koriste:
  - $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2$
  - $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$
  - $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \overline{y})}$

# Pristupi strojnom učenju – nadzirano učenje – klasifikacija



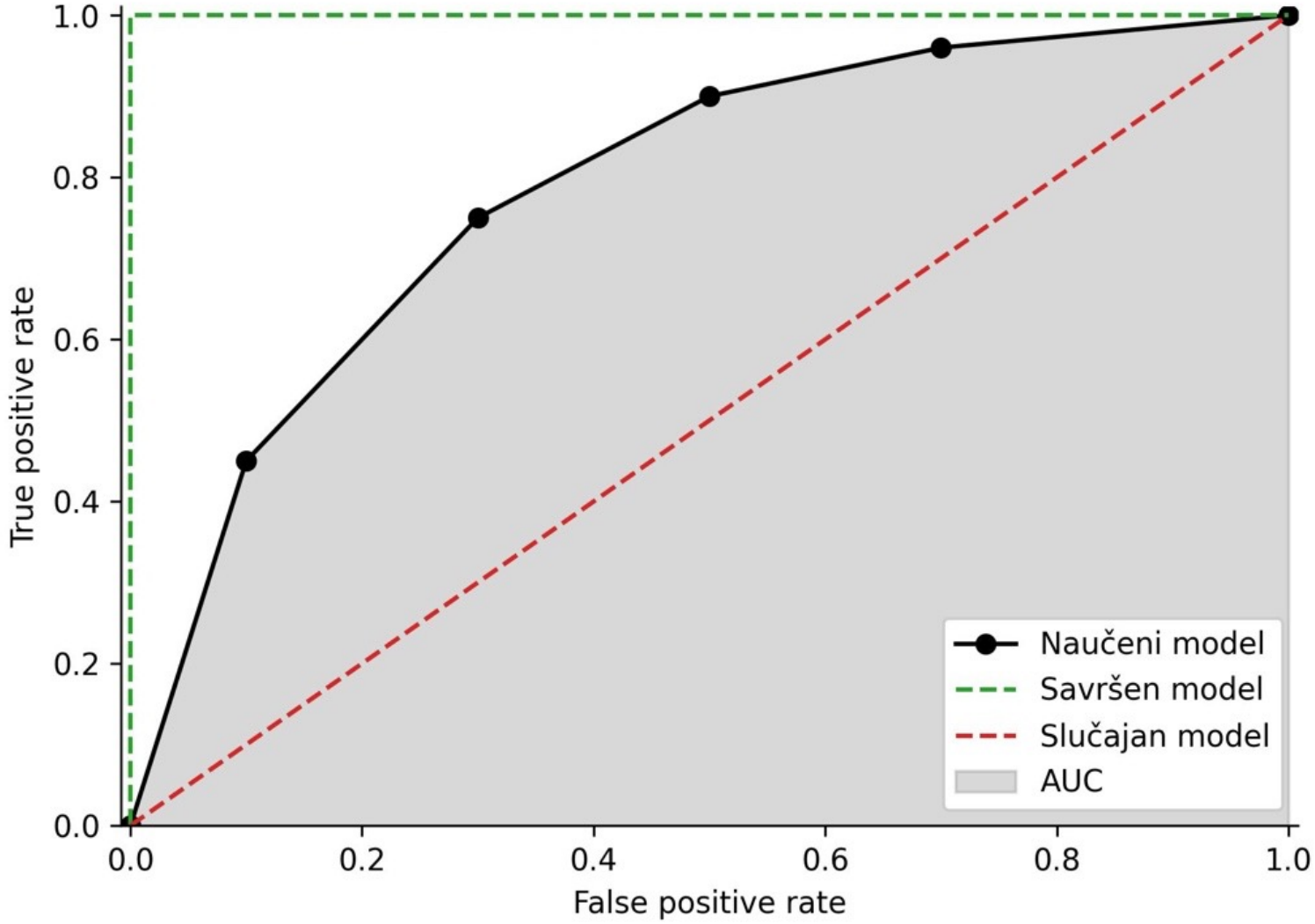
# Pristupi strojnom učenju – nadzirano učenje – klasifikacija

- Podaci – označeni
  - $x_n$ - vektor značajki (regresori)
  - $y_n$ - oznaka klase, kategorijska varijabla (često samo dvije vrijednosti – binarna klasifikacija)

- Zadatak: korištenjem skupa za učenje s $N$ opservacija (ulaznih vektora $x_n$ uz pripadajuće oznake $y_n$) predvidjeti ciljnu klasu $y$ za novu ulaznu vrijednost $x$
  - Kod binarne klasifikacije $y \in \{0,1\}$, tražimo funkciju $F$ (model), najčešće tako da je $\mathrm{P}(\mathrm{y} = 1) = F(x)$

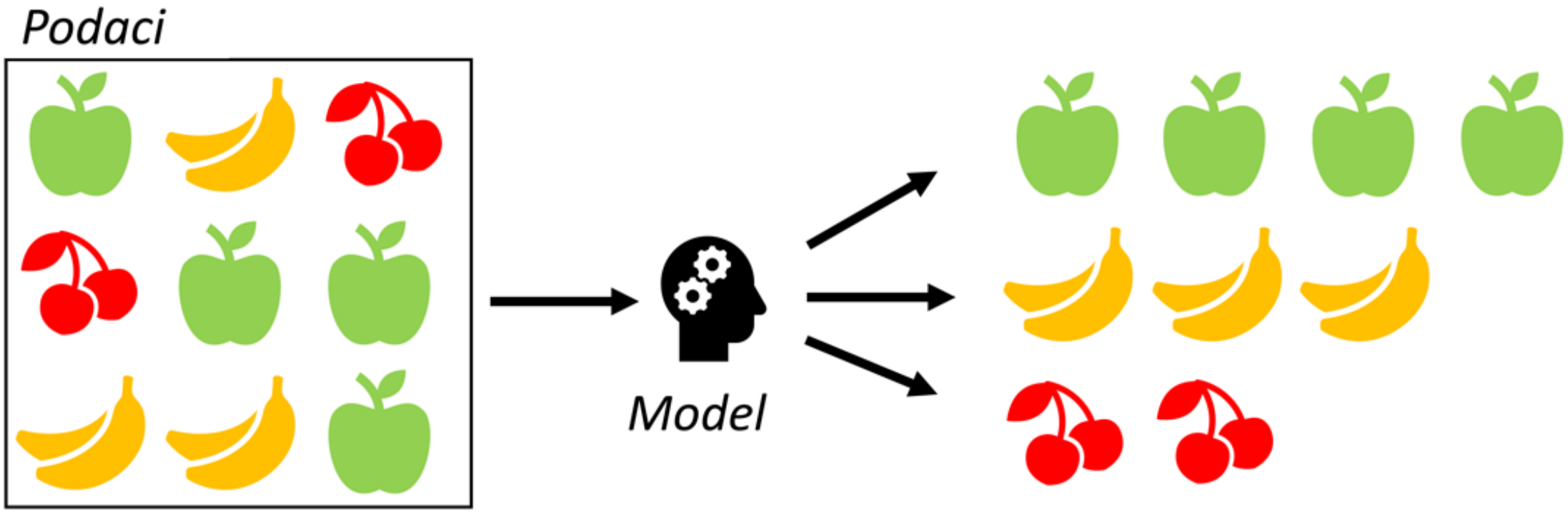
- Mjera uspješnosti – ima ih više, poput:
  - $Acc = \frac{TP + TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$
  - $TPR = \frac{TP}{TP+FN}$
  - $FNR = 1 - TPR = \frac{FN}{TP+FN}$
  - $TNR = \frac{TN}{TN+FP}$
  - $FPR = 1 - TNR = \frac{FP}{TN+FP}$

# Pristupi strojnom učenju – nadzirano učenje – klasifikacija




Primjer ROC krivulje – površina ispod krivulje (AUC) označena je sivom bojom. AUC savršenog modela iznosi 1.0, odnosno 0.5 za slučajan model

# Pristupi strojnom učenju – nenadzirano učenje – grupiranje podataka



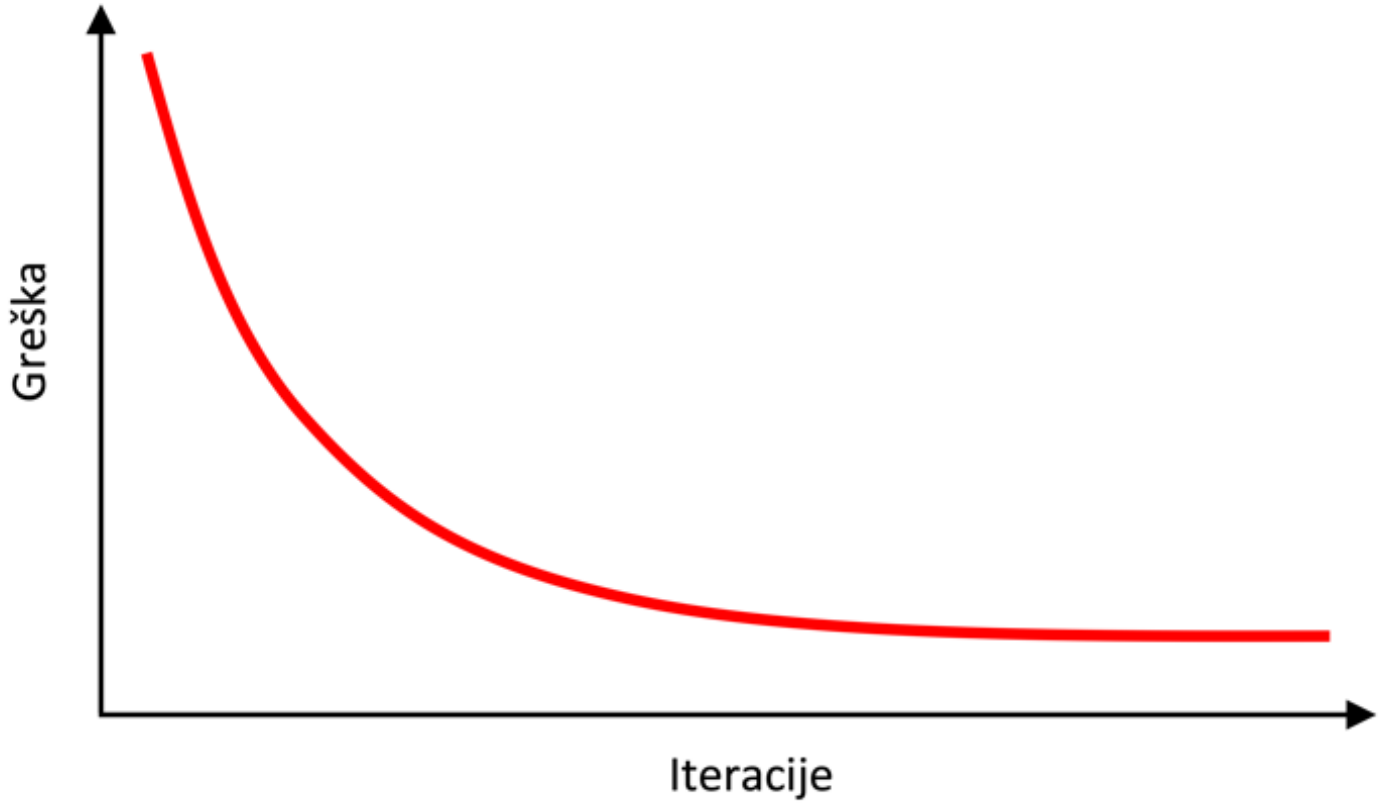
# Pristupi strojnom učenju – nenadzirano učenje

- Podaci – nisu označeni
  - $x_n$ - vektor značajki (regresori)

- Zadatak: korištenjem skupa za učenje s $N$ opservacija pronaći uzorke ili strukture u podacima
  - Grupiranje – identificirati koliko različitih grupa ima u podacima i koji primjeri se nalaze u kojim grupama
  - Redukcija dimenzionalnosti – pronaći reprezentaciju podataka u novom nižedimenzionalnom prostoru
  - Procjena funkcije gustoće vjerojatnosti – procjeniti vjerojatnost pojavljivanja pojedinog primjera

- Mjera uspješnosti – ovise o zadatku

# Prilagodba modela podacima – funkcija gubitka




Greška modela kroz iteracije treniranja

# Prilagodba modela podacima

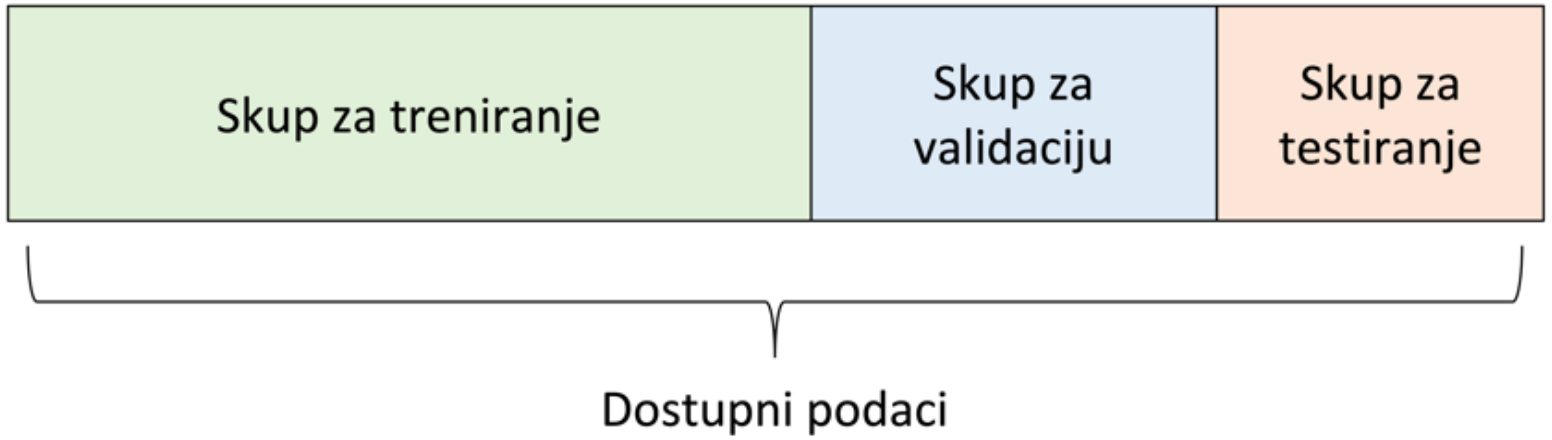
- Kapacitet modela – razina detalja koju mogu modelirati u podacima

- Cilj treniranja
  - ~~Dobiti najbolje moguće performance na skupu za treniranje (razvojni skup)~~
  - Model radi dobro na novim neviđenim primjerima – dobro generalizira

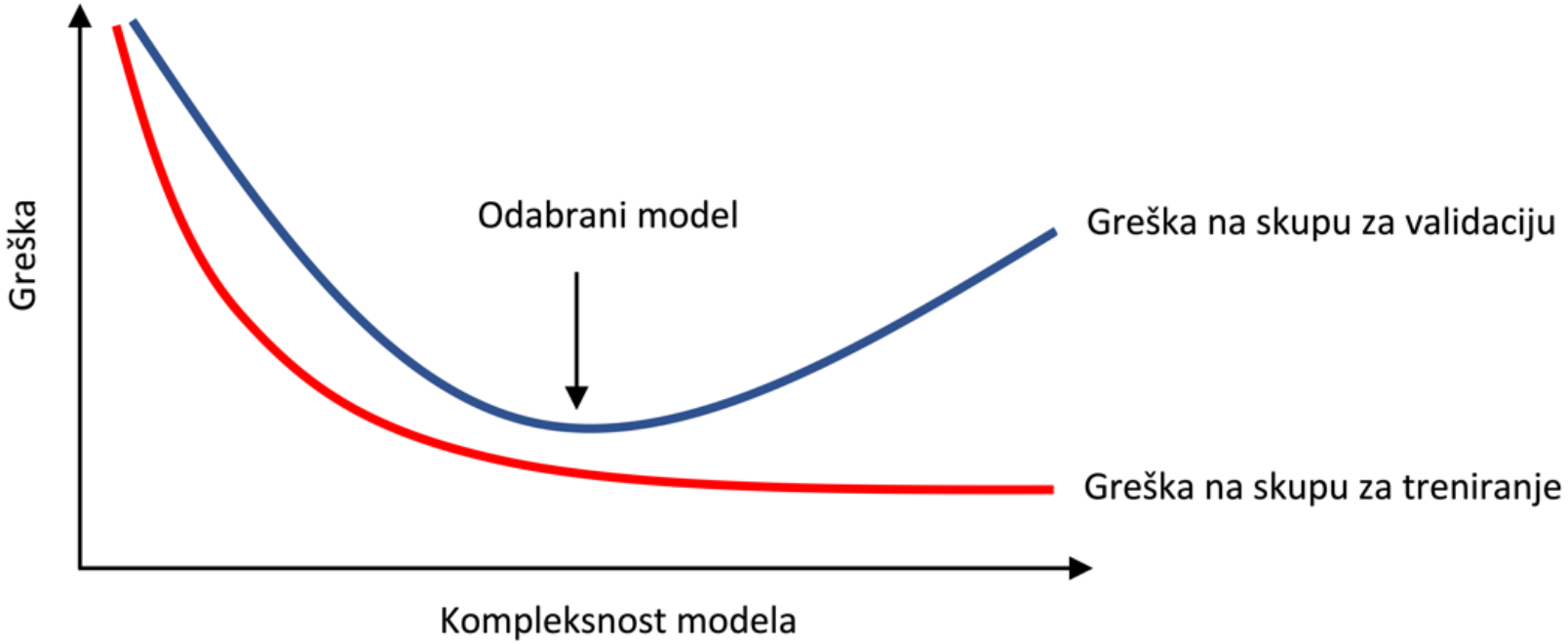
- Prema generalizacijskim performancama razlikujemo:
  - Podnaučenost – situacija kada je model prejednostavan da bi objasnio i naučio neke složene zavisnosti u podacima
  - Prenaučenost - situacija kada je model presložen s obzirom na problem i naučio je neke zavisnosti koje u podacima zapravo ne postoje
  - U oba slučaja loša generalizacija

# Prilagodba modela podacima – funkcija gubitka

- Kako bi poboljšali generalizaciju
  - Podatke dijelimo u više disjunktnih podskupova

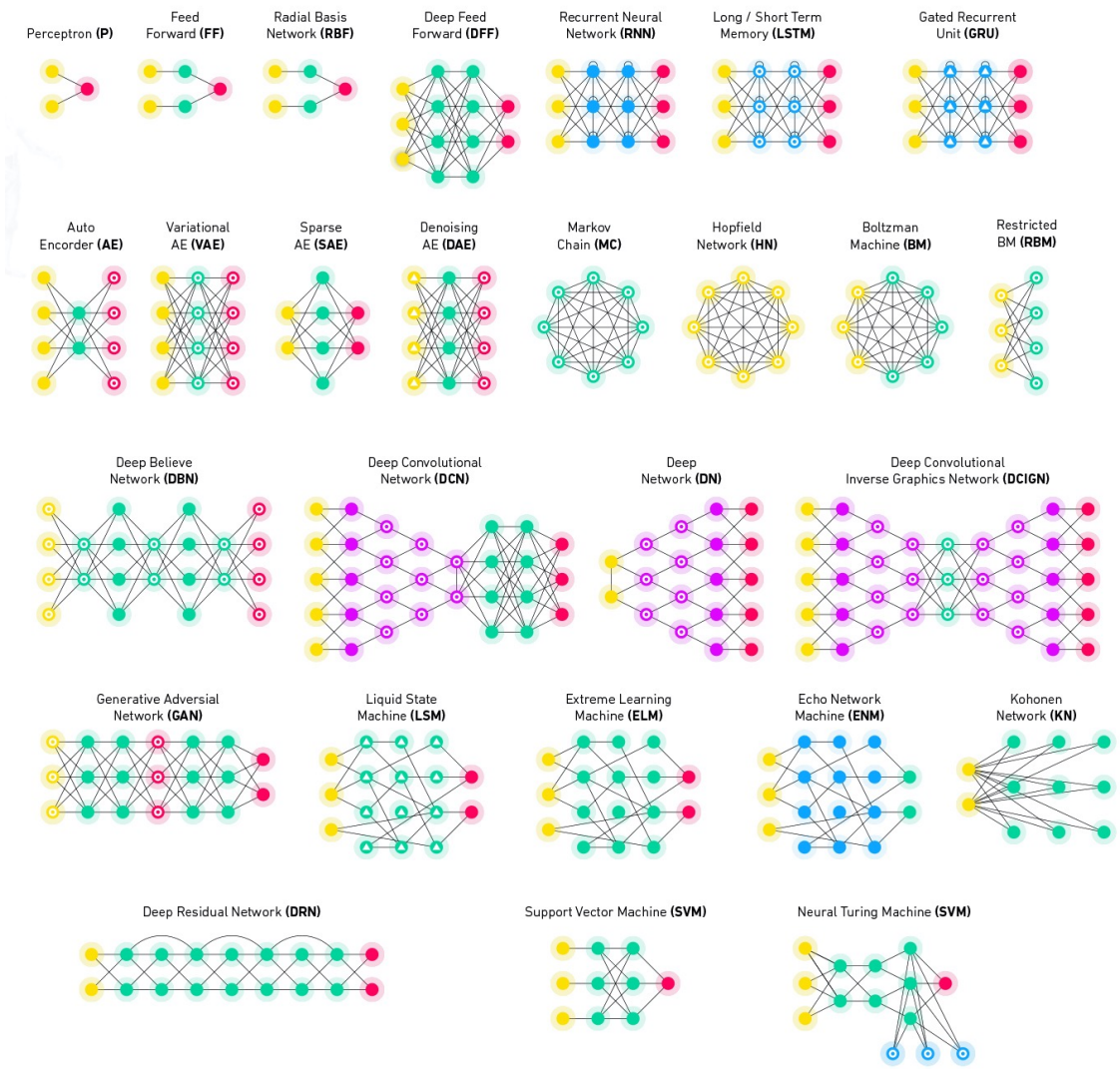



- Greška modela na skupovima za treniranje i validaciju

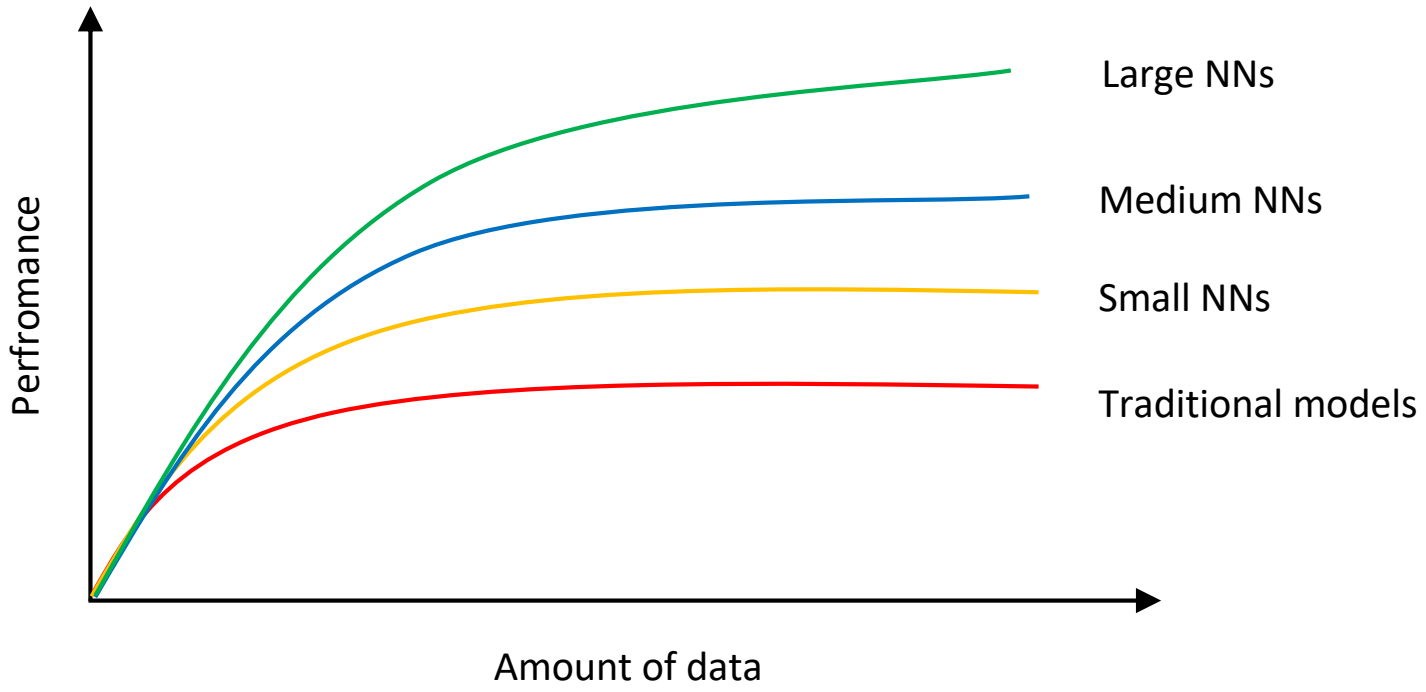


# Modeli strojnog učenja

- Linearna regresija
- Logistička regresija
- Algoritam *k*-sredina
- Stroj potpornih vektora
- Stabla odluke
- Slučajne šume
- XGB
- (Duboke) neuronske mreže
- …



# Modeli strojnog učenja s obzirom na kapacitet i količinu podataka



# Primjer – regresija (članak Topcu i suradnici)

- Problem: procjena osmolalnosti mokraće
- Podaci za treniranje: 300 uzoraka pacijenata kojima je prethodno napravljena analiza mokraće
  - Prikupljeni ulazni podatci su: osmolalnost, provodljivost, specifična težina te vrijednosti proteina, glukoze, albumina, kreatinina i pH određenih test trakom
- Odabir značajki – četiri skupine značajki
  - A. Provodljivost,
  - B. Provodljivost, specifična težina,
  - C. Provodljivost, standardni parametri analize urina (glukoza, pH, proteini, specifična težina),
  - D. Provodljivost, prošireni parametri analize urina (glukoza, pH, proteini, specifična težina, albumin, kreatinin).

# Primjer – regresija (članak Topcu i suradnici)

| Grupa značajki | Metoda | Mjere performansi | | | |
|---|---|---|---|---|---|
| | | Skup podataka | $R^2$ | MAE | RMSE |
| Grupa A (Provodljivost) | AutoML GLM | Treniranje | 0.60 | 102 | 140 |
| | | Testiranje | 0.66 | 88 | 124 |
| Grupa B (Provodljivost, specifična težina) | AutoML GBM | Treniranje | 0.90 | 43 | 71 |
| | | Testiranje | 0.83 | 56 | 87 |
| Grupa C (Provodljivost, standardna analiza urina) | AutoML GLM | Treniranje | 0.81 | 53 | 97 |
| | | Testiranje | 0.79 | 57 | 99 |
| Grupa D (Provodljivost, proširena analiza urina) | AutoML GBM | Treniranje | 1.00 | 10 | 14 |
| | | Testiranje | 0.83 | 54 | 88 |
| Biokemijski izračun | Formula | Treniranje | 0.88 | 51 | 83 |
| | | Testiranje | 0.70 | 65 | 120 |
| Izračun proizvođača | Formula proizvođača | Treniranje | 0.60 | 109 | 157 |
| | | Testiranje | 0.67 | 102 | 140 |

# Primjer – klasifikacija (članak Zhou i suradnici)

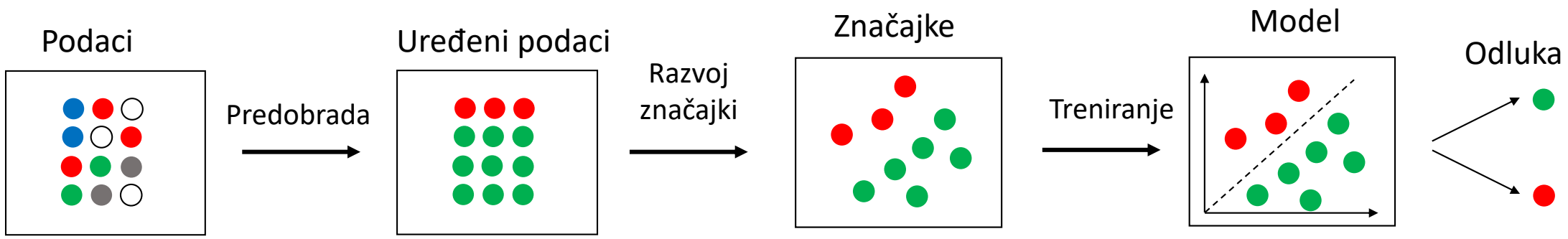
- Problem: detekcija zamijenjenih uzoraka u kliničkom laboratoriju
- Podaci za treniranje: 500,000 rezultata hematoloških analiza iz dva različita laboratorija
  - Ulazne značajke – 22 parametra kompletne krvne slike
- Klasifikacijski problem
  - “1” – zamijenjeni uzorak pacijenta
  - “0” – točno klasificirani uzorak pacijenta

# Primjer – klasifikacija (članak Zhou i suradnici)

| Model | Osjetljivost | Specifičnost | Točnost | AUC |
|---|---|---|---|---|
| Duboka mreža vjerovanja | 0.9295 | 0.9325 | 0.9310 | 0.9773 |
| k-najbližih susjeda | 0.9309 | 0.8878 | 0.9094 | 0.9455 |
| Stroj potpornih vektora | 0.9261 | 0.9196 | 0.9229 | 0.9678 |
| Slučajna šuma | 0.9117 | 0.9285 | 0.9201 | 0.9689 |
| Logistička regresija | 0.9247 | 0.8972 | 0.9110 | 0.9698 |
| Naivan Bayesov klasifikator | 0.9164 | 0.8322 | 0.8743 | 0.9509 |

# Zaključak

**Tradicionalni pristup**




**Duboko učenje**