Statističko učenje

Izborni kolegij

Izv.prof.dr.sc. Danijel Grahovac, Dominik Mihalčić

Fakultet primijenjene matematike i informatike, Sveučilište J. J. Strossmayera u Osijeku

Uvod

Učenje iz podataka

- Posljednjih desetljeća dolazi do eksplozije količine prikupljenih podataka (90% svih podataka u svijetu prikupljeno u zadnje dvije godine – i trend se nastavlja)
- Učenje iz podataka postaje važna vještina za sve koji se susreću s podacima
- Učenje iz podataka obuhvaća:
  - prepoznavanje uzoraka i trendova
  - razumijevanje onoga *što podaci govore*
- Statističko (strojno) učenje – skup metoda za stvaranje novog i korisnog znanja (u obliku pravila, uzoraka i modela) na osnovu podataka

Statističko učenje

- Računalne mogućnosti i količina dostupnih podataka omogućavaju nezamisliva postignuća (npr. umjetna inteligencija)
- Područje učenja se velikim dijelom razvijalo u okviru računarstva kao strojno učenje (*machine learning*)
- U suštini se radi o statističkim modelima i metodama
- Iako izgledaju mistično, metode strojnog učenja nije teško objasniti uz predznanje statistike

Klasični pristup

- Regresijske metode učenja obuhvaćene su drugim kolegijima (Statistika, Multivarijantna analiza, Upravljanje kreditnim rizicima)
- Regresijske metode mogu uspješno rješavati probleme kao što su:
  - predviđanje vremena potrebnog za proizvodnju neke količine nekog proizvoda
  - procjena vrijednosti nekretnine na osnovu karakteristika
  - procjena rizika da klijent neće vratiti kredit na osnovu njegovih karakteristika
  - procjena rizika od srčanog udara na osnovu nalaza
- Za velike količine podataka i nelinearne veze, ove metode su često beskorisne

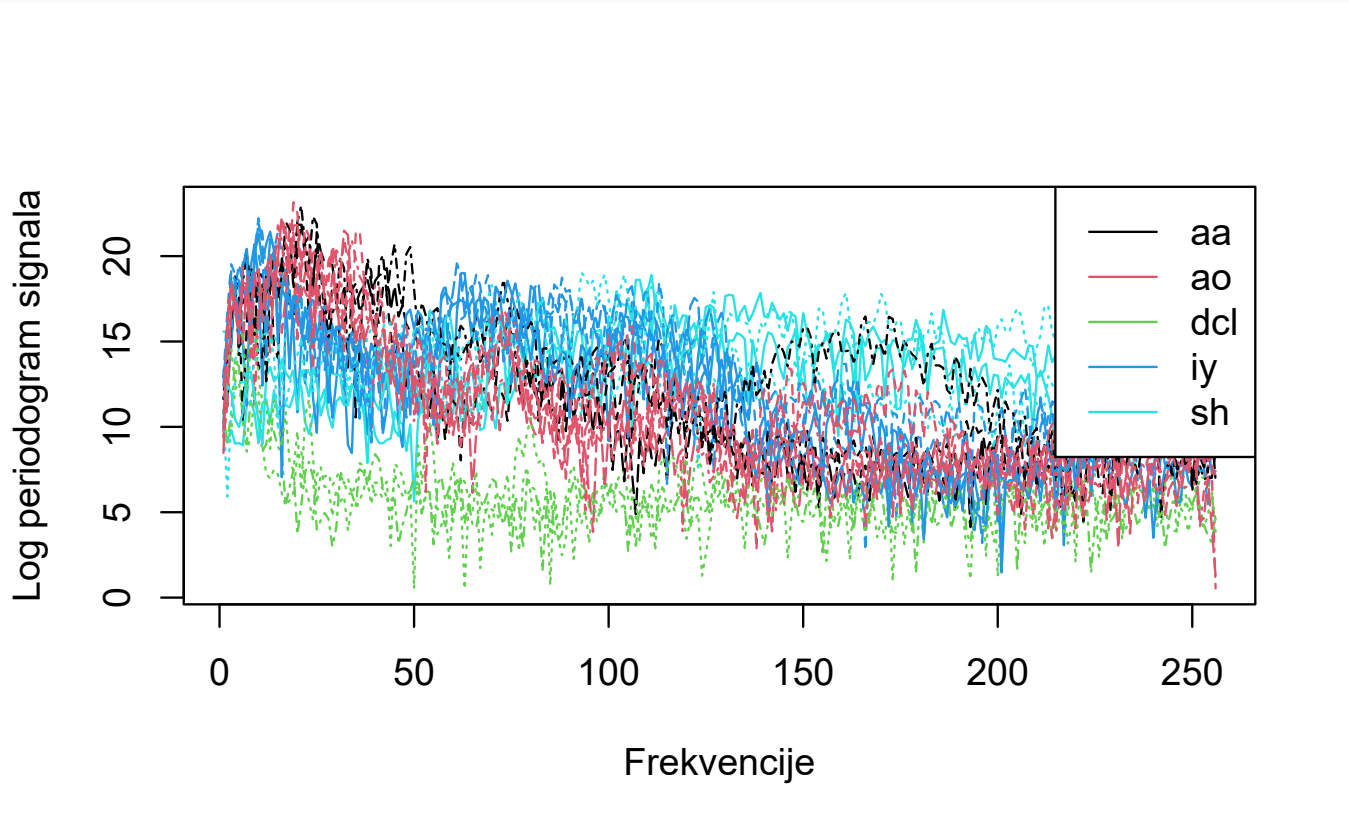
Moderne metode učenja

- U ovom kolegiju bavit ćemo se modernim metodama:
  - stabla odlučivanja
  - metoda potpornih vektora
  - **neuronske mreže** (posebno)

Primjeri problema

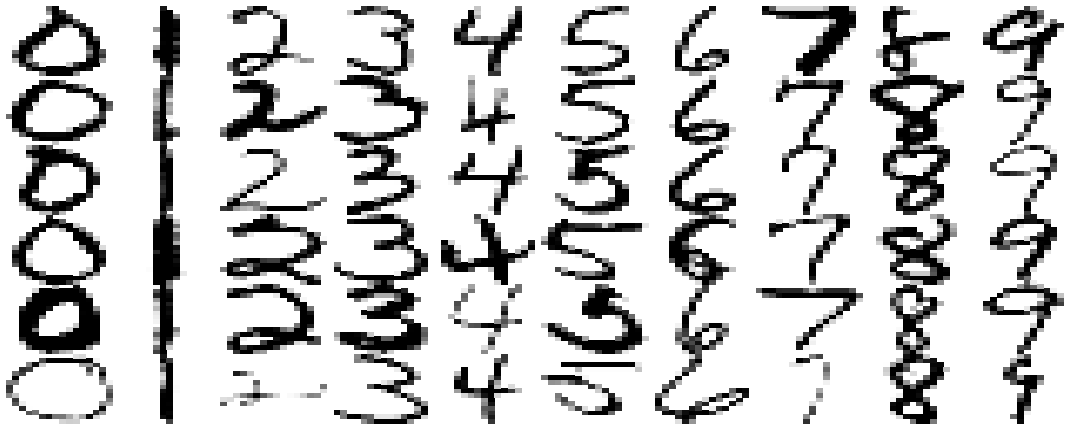
Prepoznavanje govora

- Na osnovu niza snimljenih glasova u digitalnom obliku za koje je poznato o kojem fonemu (jedinici govora) se radi, naučiti računalo da prepoznaje foneme i na novim glasovima



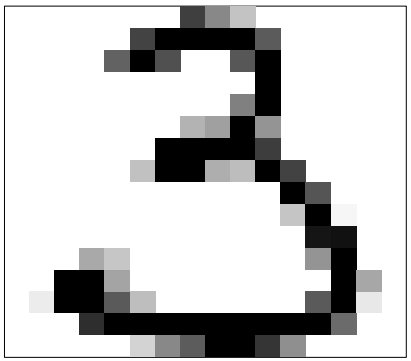
Prepoznavanje rukopisa

- Kako naučiti računalo da prepoznaje znamenke?
- **Problem**: treba nam funkcija koja će kao ulaz uzeti sliku znamenke i kao izlaz dati vrijednost 0,1,. . . ,9 (ili vjerojatnost svake znamenke)
- Primjer: MNIST baza podataka za učenje – slike znamenki s oznakom (*label*) o kojoj znamenki se radi

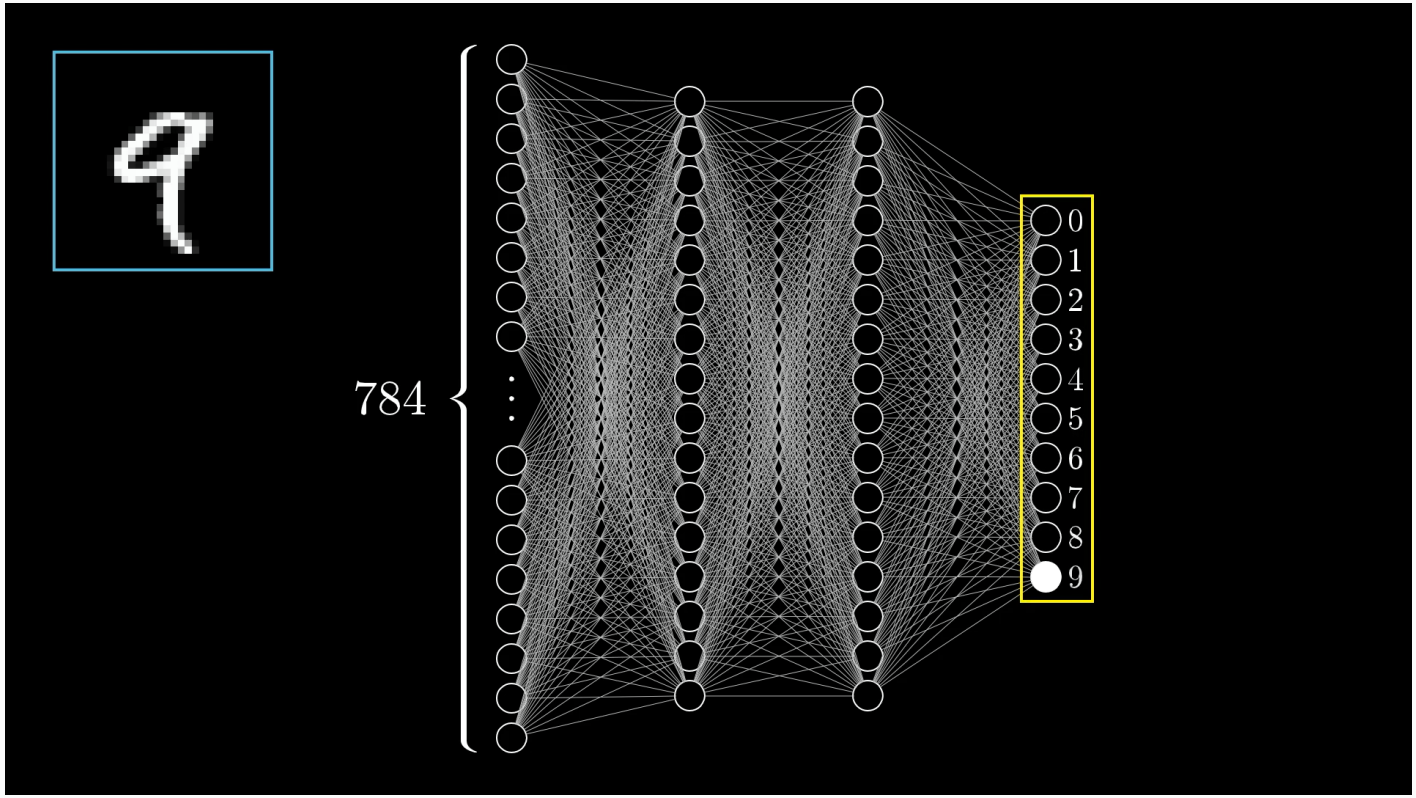


- Što je slika zapravo? Ovdje: matrica 16×16 piksela gdje je svaki piksel nijansa sive (od 0 do 256)

0	0	0	0	0	9	102	224	188	158	24	0	0	0	0	0
0	0	0	0	81	223	256	256	256	256	210	14	0	0	0	0
0	0	0	0	207	256	216	70	34	212	256	91	0	0	0	0
0	0	0	0	80	104	16	0	0	126	256	96	0	0	0	0
0	0	0	0	0	0	0	0	3	192	256	59	0	0	0	0
0	0	0	0	0	0	84	166	175	256	181	1	0	0	0	0
0	0	0	0	0	127	256	256	256	256	225	34	0	0	0	0
0	0	0	0	0	159	256	256	169	161	256	223	31	0	0	0
0	0	0	0	0	3	68	1	0	0	106	256	214	7	0	0
0	0	0	0	0	0	0	0	0	0	0	157	256	133	0	0
0	0	0	0	0	0	0	0	0	0	0	22	246	247	26	0
0	0	43	171	157	12	0	0	0	0	0	0	182	256	95	0
0	97	256	256	173	5	0	0	0	0	0	0	118	256	172	0
0	138	256	256	211	161	102	83	34	34	34	73	211	256	140	0
0	5	120	233	256	256	256	256	256	256	256	256	256	203	20	0
0	0	0	16	86	150	188	210	256	256	229	184	103	15	0	0



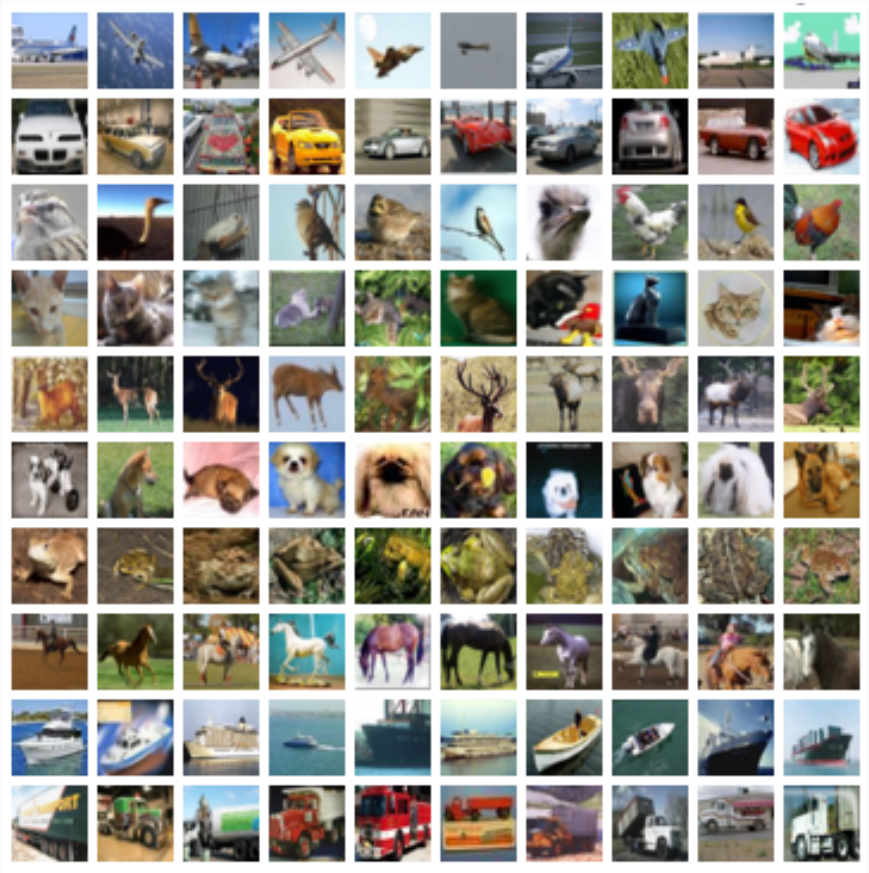
- Primjer arhitekture neuronske mreže za prepoznavanje znamenki



- Neuronske mreže odlično mogu rješavati ovakve probleme
- Lako je postići točnost i 98% a uz naprednije arhitekture i 99.75% (gotovo bolje od čovjeka)

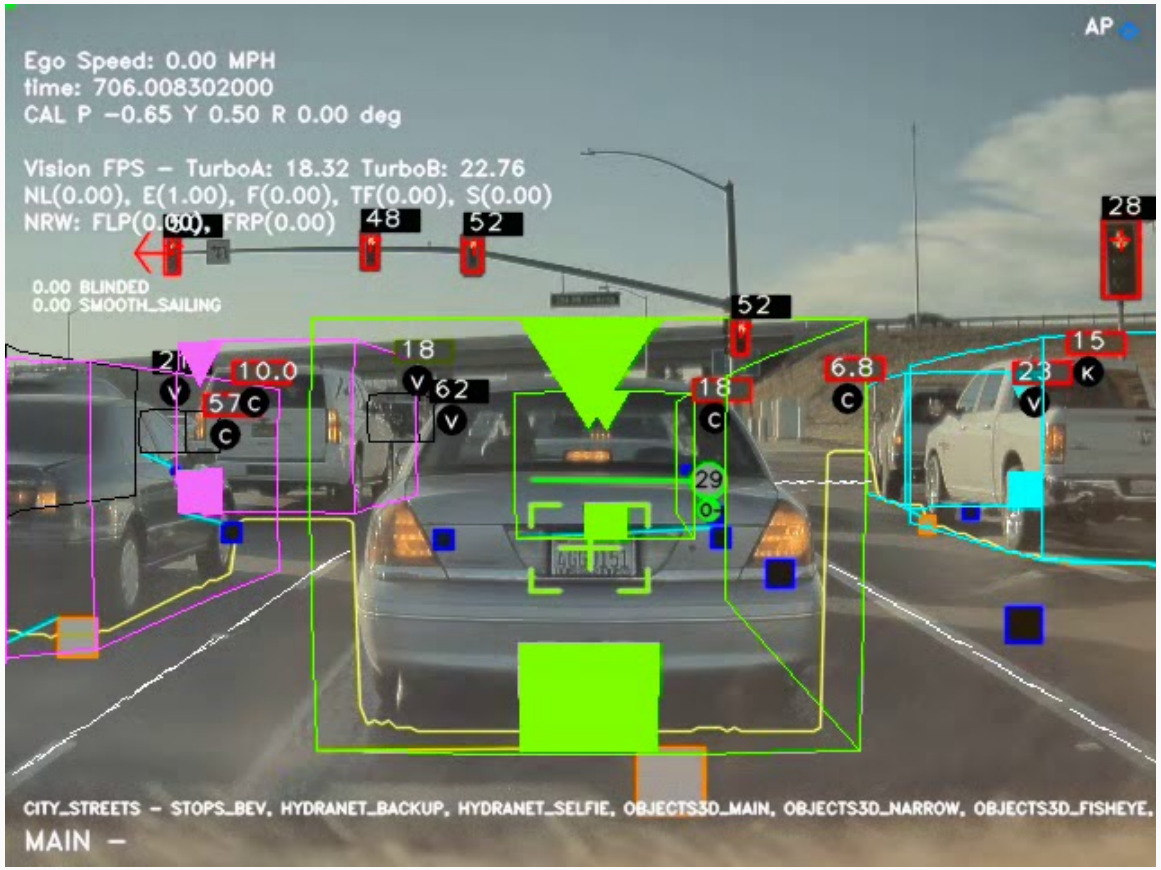
Drugi primjeri

- Prepoznavanje objekata na slici
  - primjerice, je li na slici avion, automobil, mačka, pas. . .
  - podaci se sastoje od puno slika za koje je označeno što predstavljaju



- Prepoznavanje lica
  - primjerice za autorizaciju ili kameru
- Prepoznavanje spama
  - podaci su emailovi za koje je označeno jesu li spam ili ne
- Razgovorni agenti (*chatbot*)
  - na osnovu baze pitanja i odgovora treba naučiti računalo da razgovara s korisnikom što sličnije čovjeku

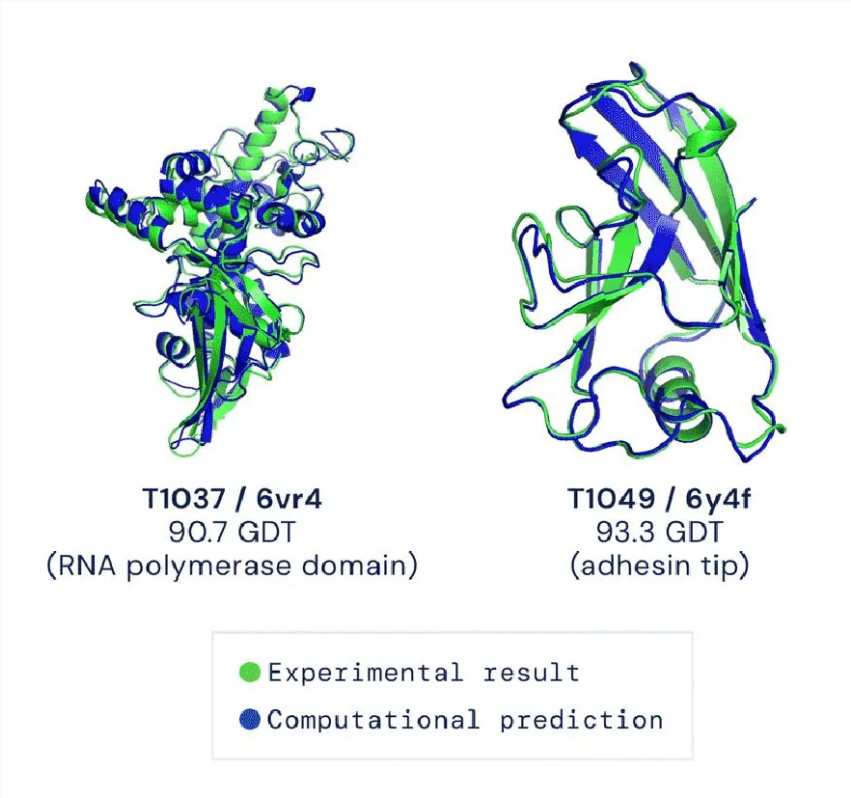
- Prepoznavanje objekata iz videa
  - primjerice za autonomne automobile



Poznati uspješni primjeri

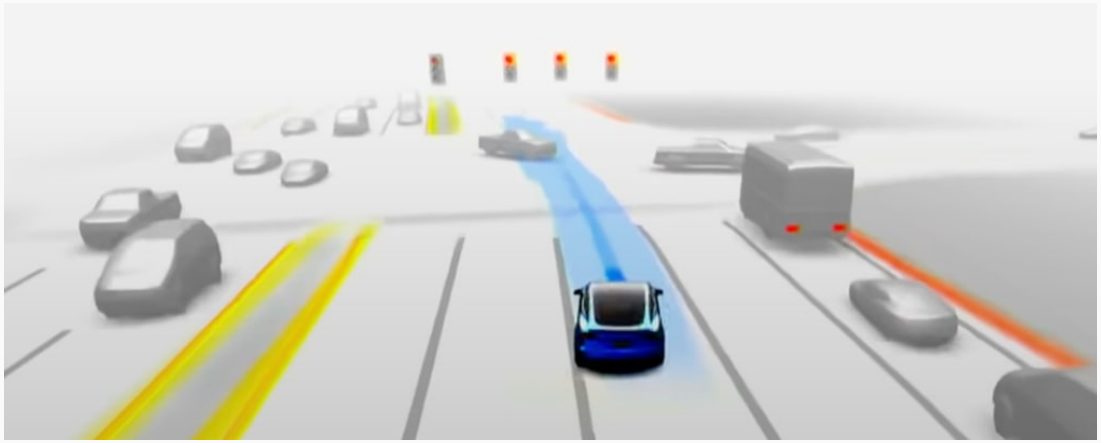
- DeepMind AlphaGo pobijedio svjetskog prvaka u igri Go
  - teže od šaha (prosječan broj mogućih sljedećih poteza 250 u odnosu na 35 u šahu)
  - koristi učenje neuronskom mrežom

- Predviđanje 3D proteinskih struktura iz njihovih proteinskih sljedova
  - ključno za razumijevanje biološke funkcije proteina (kako utjecati na protein ili ga izmijeniti)
  - DeepMind AlphaFold model dubokog učenja 2020. godine smanjio vrijeme predviđanje s oko 5 godina na nekoliko sekundi



- ChatGPT – razgovorni agent (chatbot) koji razgovara, odgovara na pitanja, piše tekst i programski kod
  - radi se o funkciji koja na osnovu ulaznih riječi (pitanje) predviđa najvjerojatnije sljedeće riječi (odgovor)
  - temeljen na velikom jezičnom modelu učenom na 375 milijardi riječi
  - 175 milijardi parametara
  - Trošak samo treniranja se procjenjuje na 10-ak milijuna dolara

- Tesla Autopilot i Full Self Driving
  - ADAS (napredni sustavi za podršku vozaču) u potpunosti se oslanjaju na kamere i naučenu neuronsku mrežu
  - Tesla Model Y dobio je najveću ocjenu na Euro NCAP testiranju ikad u kategoriji sigurnosnih sustava
  - FSD može voziti samostalno uz sve manje potrebnih intervencija



Zašto upisati ovaj kolegij?

Zašto upisati ovaj kolegij?

- Razumijevanje pojmove kao što su neuronske mreže i umjetna inteligencija postaje pitanje opće kulture (ovaj kolegij će u budućnosti sigurno biti obvezan)
- Želite znati kako funkcionira ChatGPT, autonomno vozilo, AI fotografiranje na mobitelu, otključavanje mobitela licem...
- Želite naučiti nove metode s velikim mogućnostima
- Uz predznanje statistike i vjerojatnosti, nije teško razumjeti nove metode
- Naglasak na samostalnom praktičnom radu
- Svaki *data scientist* je nezamisliv bez poznavanja tehnika strojnog učenja

Izvedba

- Kroz dva sata predavanja i tri sata seminara naglasak će biti na praktičnim primjerima i problemima
- Nastava u praktikumu
- Samostalan rad - kroz domaće zadaće i seminare