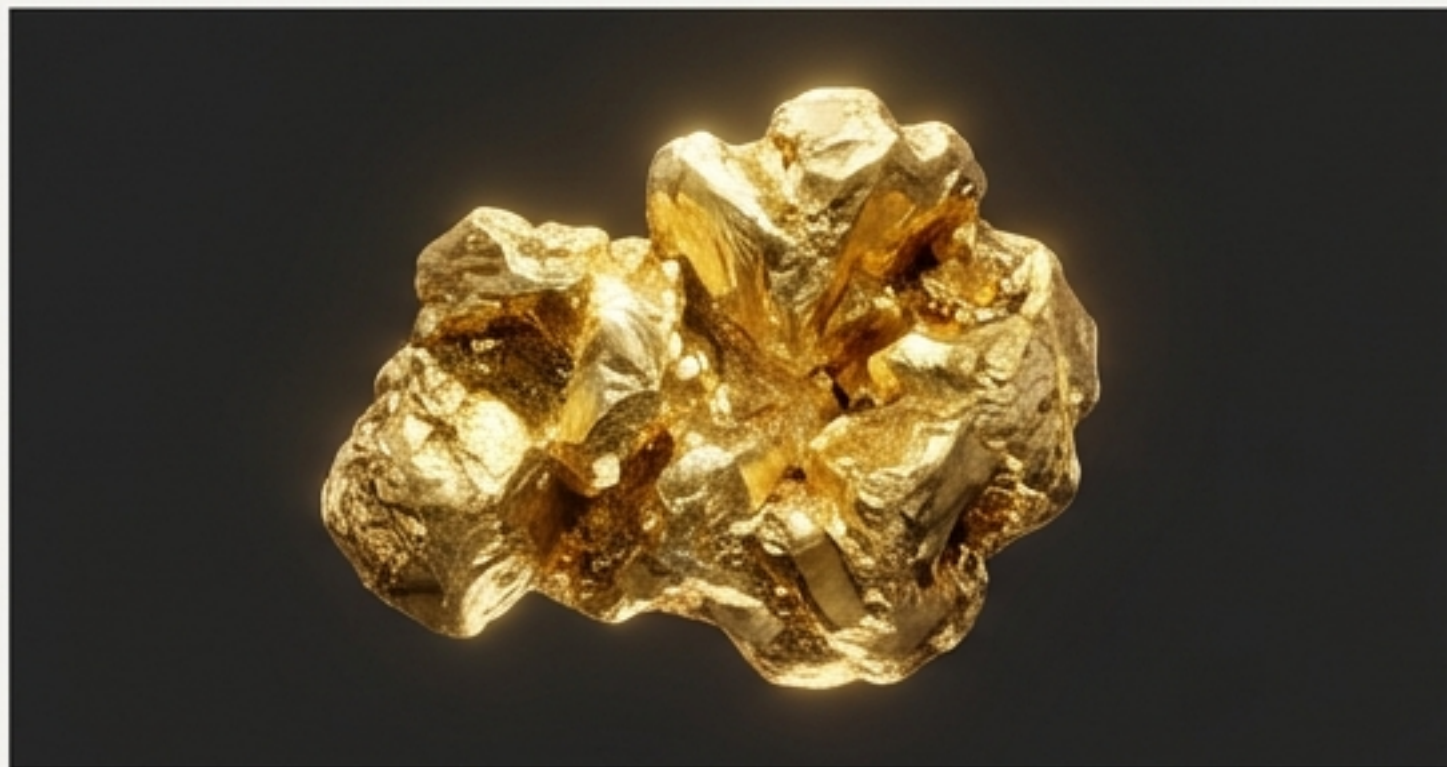




Podaci: Gorivo veštačke inteligencije

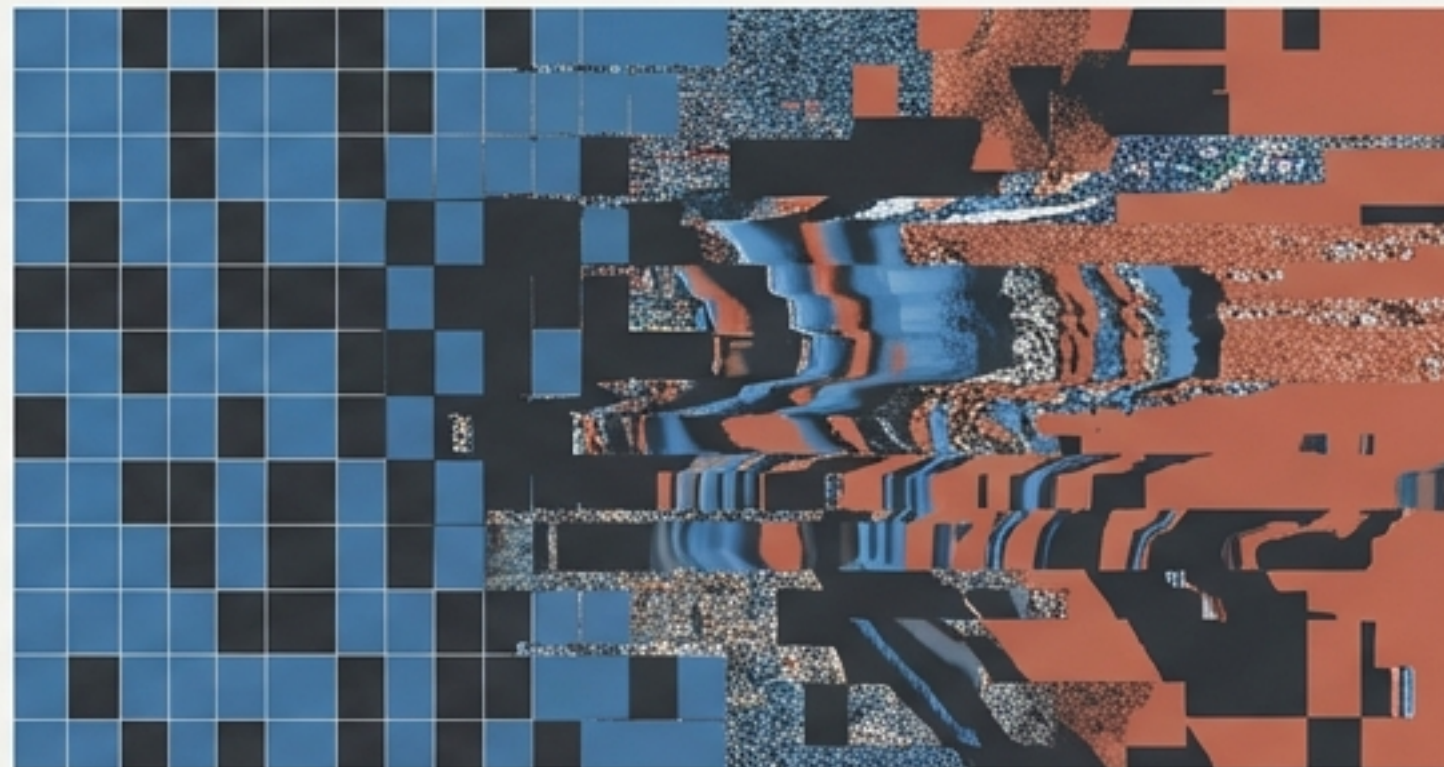
*Od sirovog resursa do
strateškog sredstva*

Dva lica podataka



„Podaci su novo zlato.“

Naglašava vrednost, potencijal i prilike koje podaci otvaraju za razumevanje i modeliranje sveta.



„Smeće unutra, smeće napolje.“

Upozorava na fundamentalnu zavisnost AI od kvaliteta podataka. Bez obzira na sofisticiranost algoritma, loši ulazni podaci će neizbežno proizvesti loše rezultate.

Izazovi stvarnog sveta: Zašto je rad sa podacima težak?



Kvalitet: Loše pripremljeni skupovi podataka **smanjuju** tačnost modela **za 30-50%**. Preciznost, pravovremenost i potpunost su ključni, ali retko dostižni.



Kvantitet: Mnogi problemi, posebno u domenima kao što je medicina, pate od nedostatka podataka ('data scarcity'), što dovodi do nepouzdatih modela ('under-fitting').

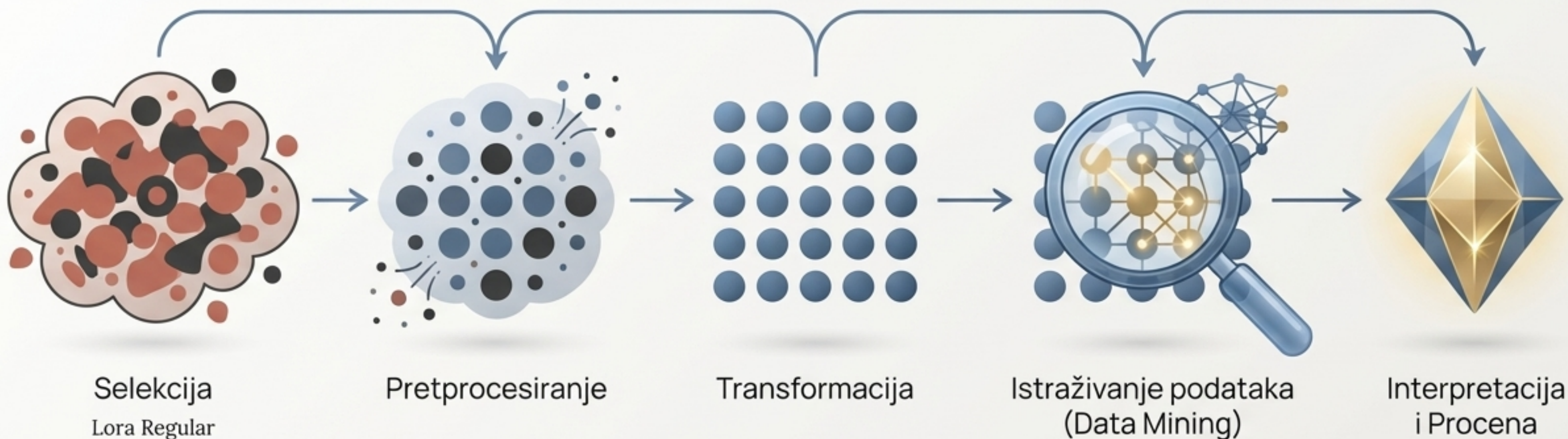


Cena: Prikupljanje, skladištenje, obeležavanje i obrada zahtevaju ogromne resurse: hardver, vreme i timove stručnjaka, što čini podatke skupim sredstvom.



Privatnost i regulativa: Osetljivi lični i poslovni podaci su zaštićeni zakonima (npr. GDPR). Njihovo korišćenje nameće stroga ograničenja i etičke dileme.

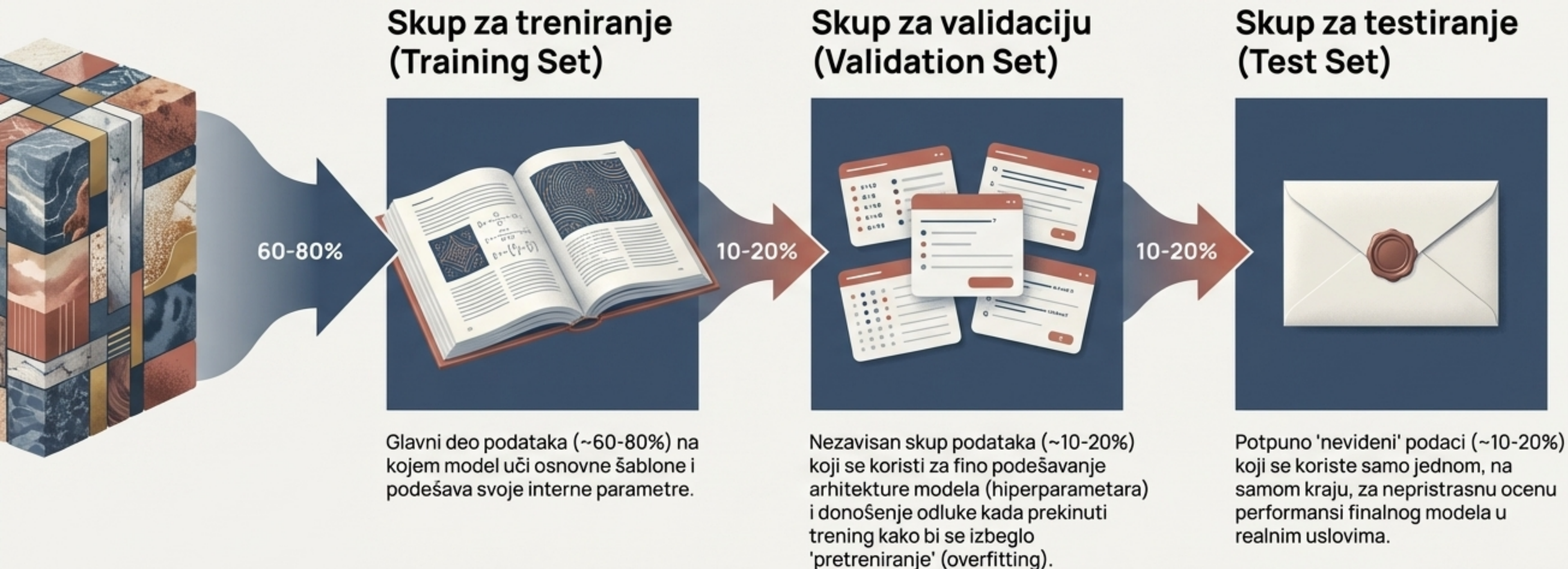
Ukroćivanje haosa: Proces otkrivanja znanja u podacima (KDD)



“Istraživanje podataka je poluautomatska ili automatska analiza velike količine podataka kako bi se izvukli prethodno nepoznati, zanimljivi šabloni.

Lora Italic

Sprečavanje grešaka: Kako mašine uče bez 'prepisivanja'



Pretreniranje (Overfitting): Stanje u kojem model nauči napamet šum i specifičnosti trening podataka, umesto generalizovanih pravila, što dovodi do loših performansi na novim podacima.

Katalizatori revolucije: Skupovi podataka koji su definisali jednu eru

Napredak u AI nije bio vođen samo algoritmima, već i dostupnošću javnih, visokokvalitetnih skupova podataka. Oni su postali 'zlatni standard' za testiranje, upoređivanje i unapređivanje modela, stvarajući zajednički jezik za istraživače širom sveta.



MNIST



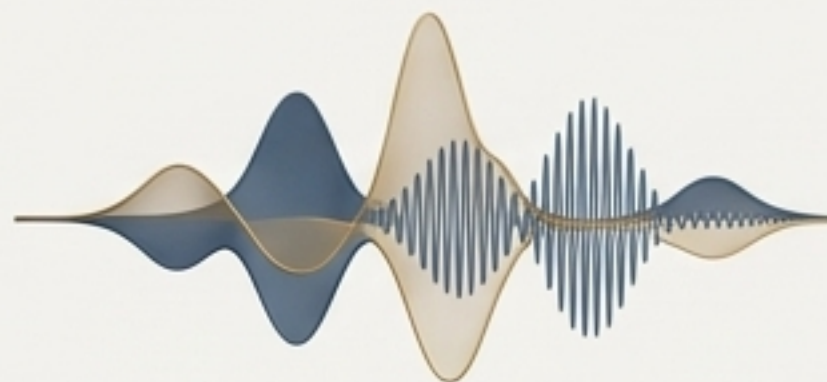
ImageNet



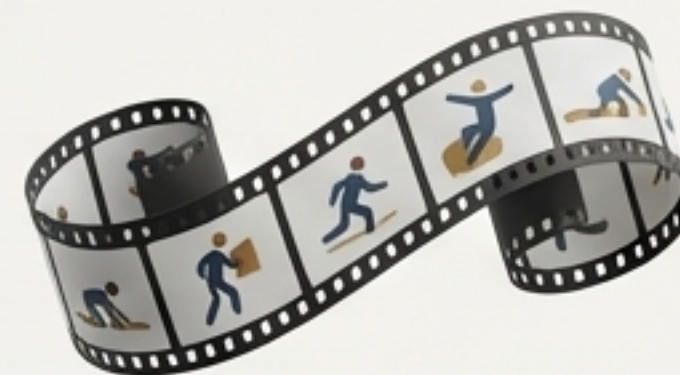
COCO



IMDB



AudioSet

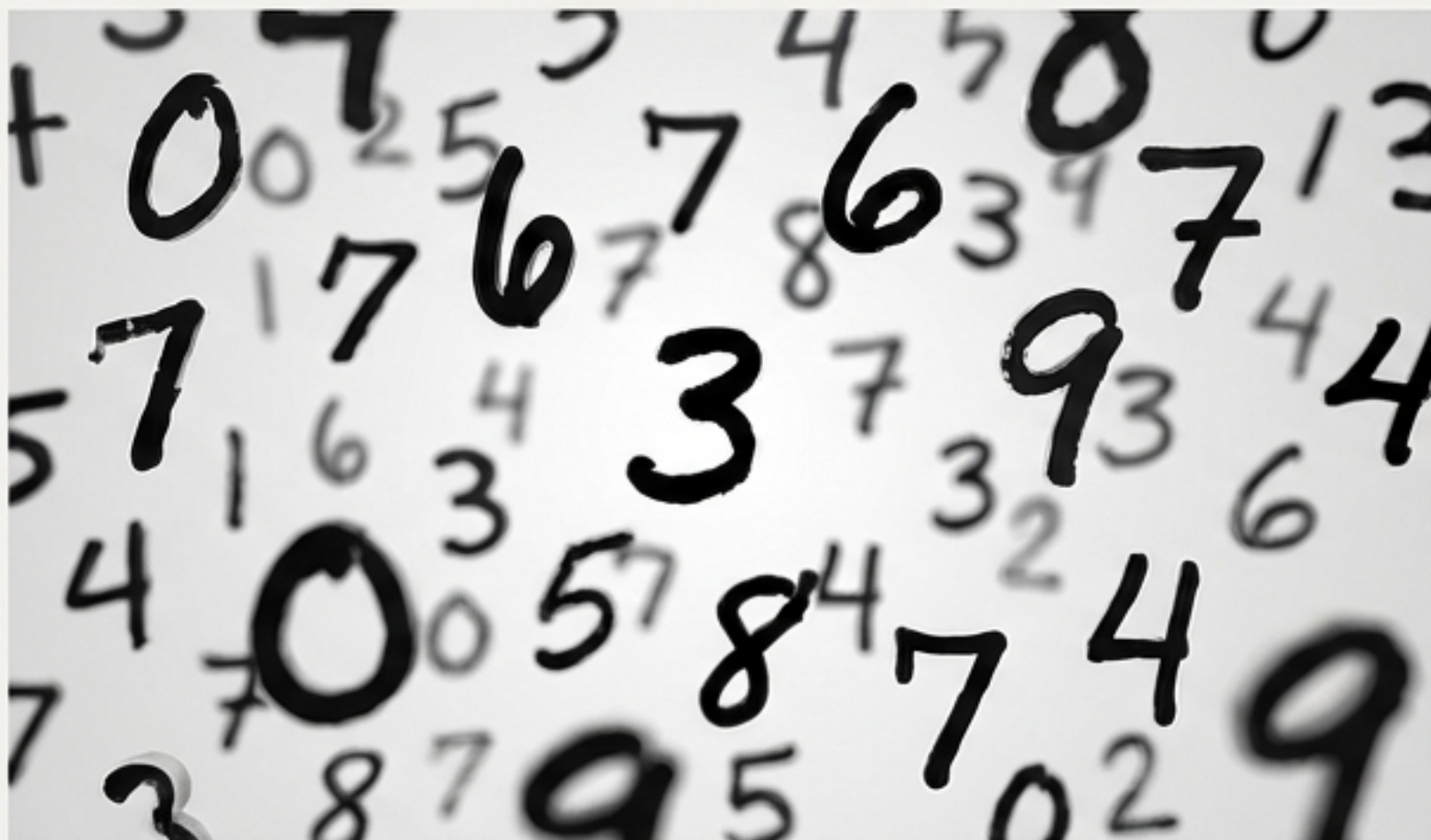


Moments in Time

Temelji Moderne AI: Ključni Ključni Skupovi Podataka

Dva pionirska resursa koja su definisala pravac razvoja računarskog vida.

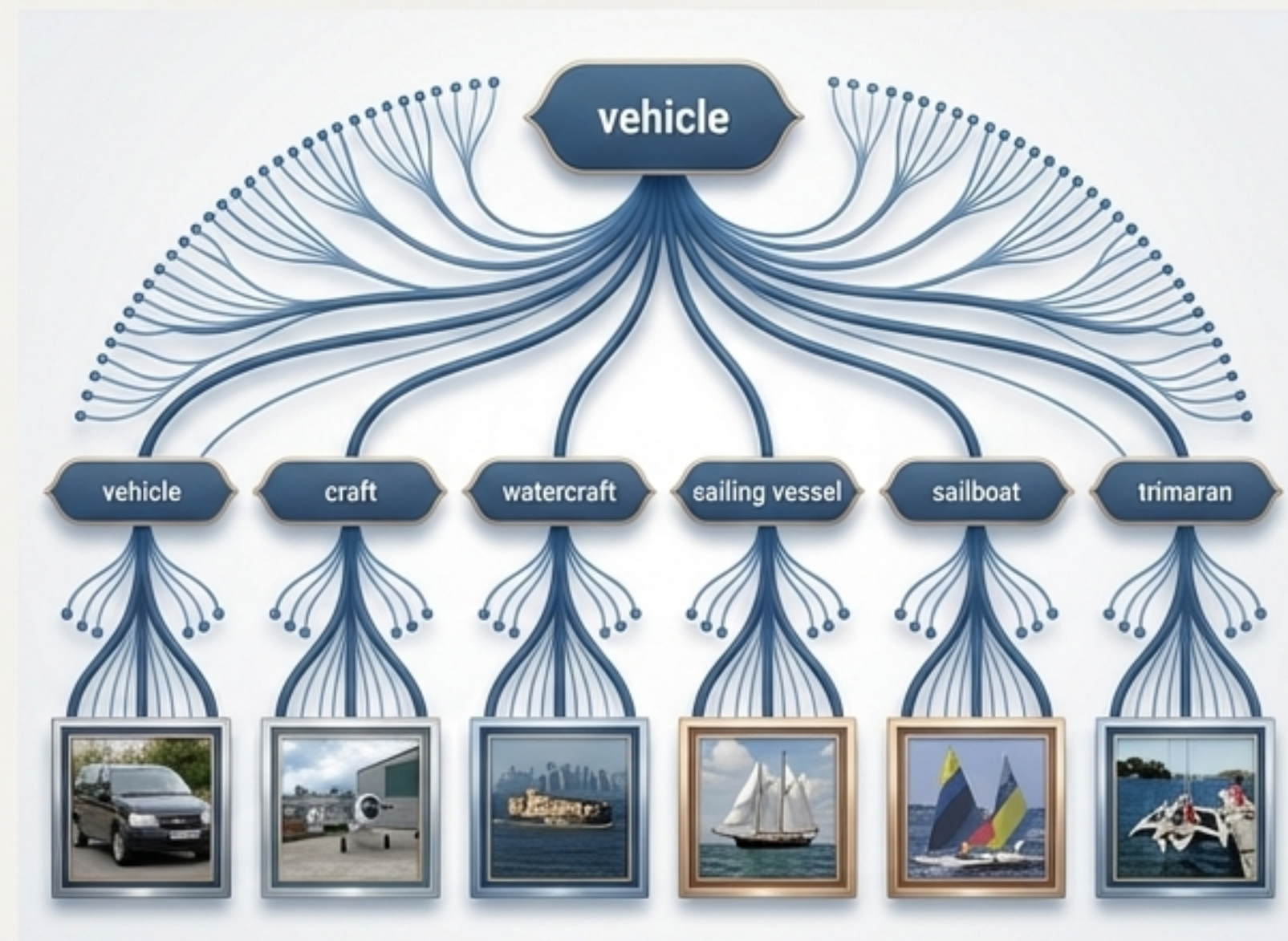
MNIST: "Hello, World!" za računarski vid



Jedan od najpopularnijih setova, sastavljen od 70.000 crno-belih slika rukom pisanih cifara (28x28 piksela). Njegova jednostavnost i dostupnost omogućile su brzu iteraciju i testiranje ranih konvolucionih neuronskih mreža.

60.000 za trening | 10.000 za test | Razvijen 1998.

ImageNet: Iskra koja je zapalila revoluciju dubokog učenja



Masivna baza od ~14 miliona slika organizovanih u hijerarhiju od preko 21.000 kategorija ("synsets"). Godišnje takmičenje (ILSVRC) na ovom setu je direktno dovelo do proboja modela poput AlexNet-a i pokrenulo modernu eru dubokog učenja.

COCO: Razumevanje objekata u složenom kontekstu



Skup od 330.000 slika dizajniran za napredne zadatke: detekciju objekata, segmentaciju (iscrtavanje kontura) i automatsko generisanje opisa slika. Sadrži 1.5 miliona instanci objekata iz 80 kategorija.

IMDB: Dešifrovanje ljudskih osećanja u tekstu



Sadrži 50.000 filmskih recenzija na engleskom jeziku, podeljenih na 25.000 pozitivnih i 25.000 negativnih. Ključan resurs za razvoj modela za analizu sentimenata, što je fundamentalni zadatak u obradi prirodnog jezika (NLP).

Sledeća granica: Kada stvarni podaci nisu dovoljni

Šta ako su nam potrebni podaci o retkim događajima koji se skoro nikada ne dešavaju?

Šta ako je prikupljanje podataka fizički nemoguće ili previše opasno?

Šta ako zakoni o privatnosti u potpunosti zabranjuju korišćenje originalnih podataka?

REŠENJE: SINTETIČKI PODACI

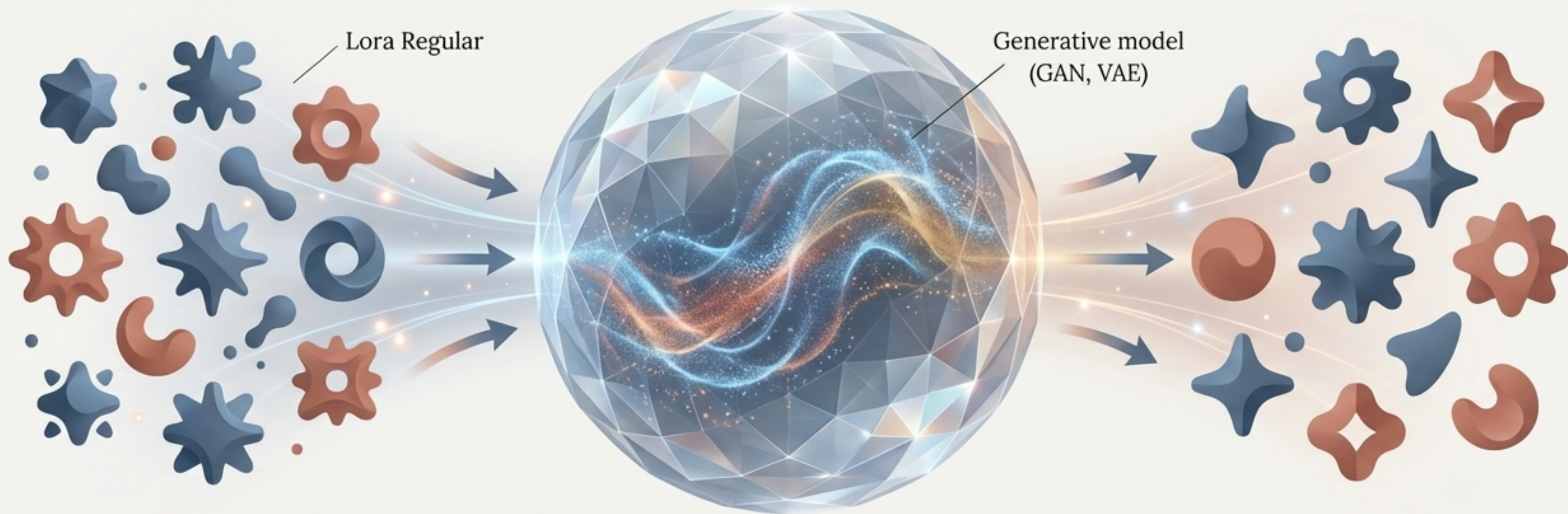
Veštački generisane informacije kreirane pomoću računarskih algoritama, dizajnirane da oponašaju statistička svojstva stvarnih podataka, ali bez direktne veze sa stvarnim događajima ili pojedincima.

Kako mašine 'izmišljaju' podatke: Uloga generativnih modela

ULAZ: Stvarni podaci

UČENJE: Model distribucije

IZLAZ: Sintetički podaci



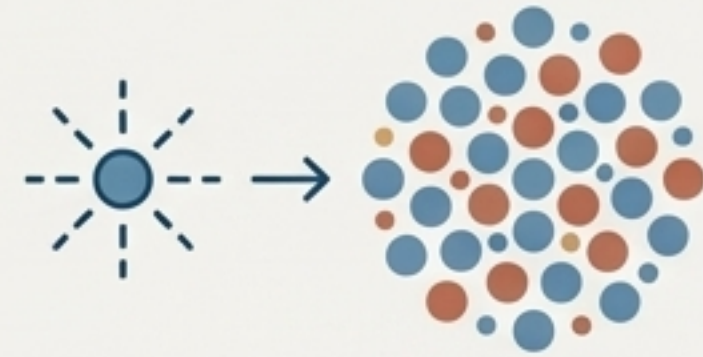
- Generativne suparničke mreže (GANs)
- Varijacioni autoenkoderi (VAEs)
- Difuzioni modeli

Moć sintetičkih podataka: Rešavanje ključnih izazova



Zaštita privatnosti

Omogućava deljenje i analizu podataka u osetljivim domenima (zdravstvo, finansije) generisanjem anonimizovanih skupova koji zadržavaju statističku korisnost bez otkrivanja identiteta.



Povećanje količine podataka

Rešava problem retkih događaja (npr. prevare, retke bolesti) generisanjem dodatnih primera za balansiranje trening skupova i poboljšanje robusnosti modela.



Bezbedna simulacija

Omogućava trening i testiranje sistema u ekstremnim ili opasnim uslovima (npr. autonomna vozila u scenarijima sudara) bez realnog rizika.



Smanjenje pristrasnosti

Omogućava kreiranje idealno balansiranih skupova podataka za trening, pomažući u borbi protiv algoritamske diskriminacije nasleđene iz nefer stvarnih podataka.

Transformacija industrija: Primena sintetičkih podataka



Sa velikom moći dolazi i nova odgovornost

Generativni modeli nisu objektivni. Oni uče iz stvarnog sveta i mogu ne samo da reprodukuju, već i da pojačaju postojeće društvene pristrasnosti i nepravde koje su prisutne u originalnim podacima.

Pristrasnost (Bias)

Ako model uči iz istorijskih podataka gde su određene grupe bile diskriminisane, da li će sintetički podaci perpetuirati tu diskriminaciju?

Pravičnost (Fairness)

Kako osigurati da AI sistemi trenirani na sintetičkim podacima donose pravične odluke za sve demografske grupe?

Transparentnost

Da li imamo obavezu da jasno naznačimo kada se koriste sintetički podaci i koja su ograničenja modela koji ih je generisao?

Upravljanje rizicima: Privatnost i zloupotreba

1. Curenje privatnosti

- Napadi poput "Membership Inference" mogu potencijalno otkriti da li je podatak određene osobe bio deo originalnog trening skupa, čak i ako se radi sa sintetičkim izlazom.
- **Diferencijalna privatnost:** Formalna matematička garancija koja osigurava da dodavanje ili uklanjanje podataka jedne osobe iz skupa za učenje ima zanemarljiv statistički uticaj na izlaz modela, čime se štiti individualna privatnost.



2. Maliciozna upotreba

- Ista tehnologija koja generiše korisne podatke može se zloupotrebiti za kreiranje sofisticiranih dezinformacija ('deepfakes'), lažnih vesti i drugog štetnog sadržaja (AI-Generated Content - AIGC).



Od pasivnog resursa do aktivnog alata: Redefinisanje uloge podataka

Era veštačke inteligencije transformiše podatke. Oni više nisu samo pasivan resurs koji se sakuplja i analizira, već postaju **aktivno**, dizajnersko sredstvo. Kroz sintetičke podatke, mi sada možemo da ih kreiramo, oblikujemo i optimizujemo za rešavanje najtežih problema.

Ova evolucija zahteva promenu načina razmišljanja – **od puke potrošnje podataka ka svesnom i odgovornom stvaranju podataka.**

