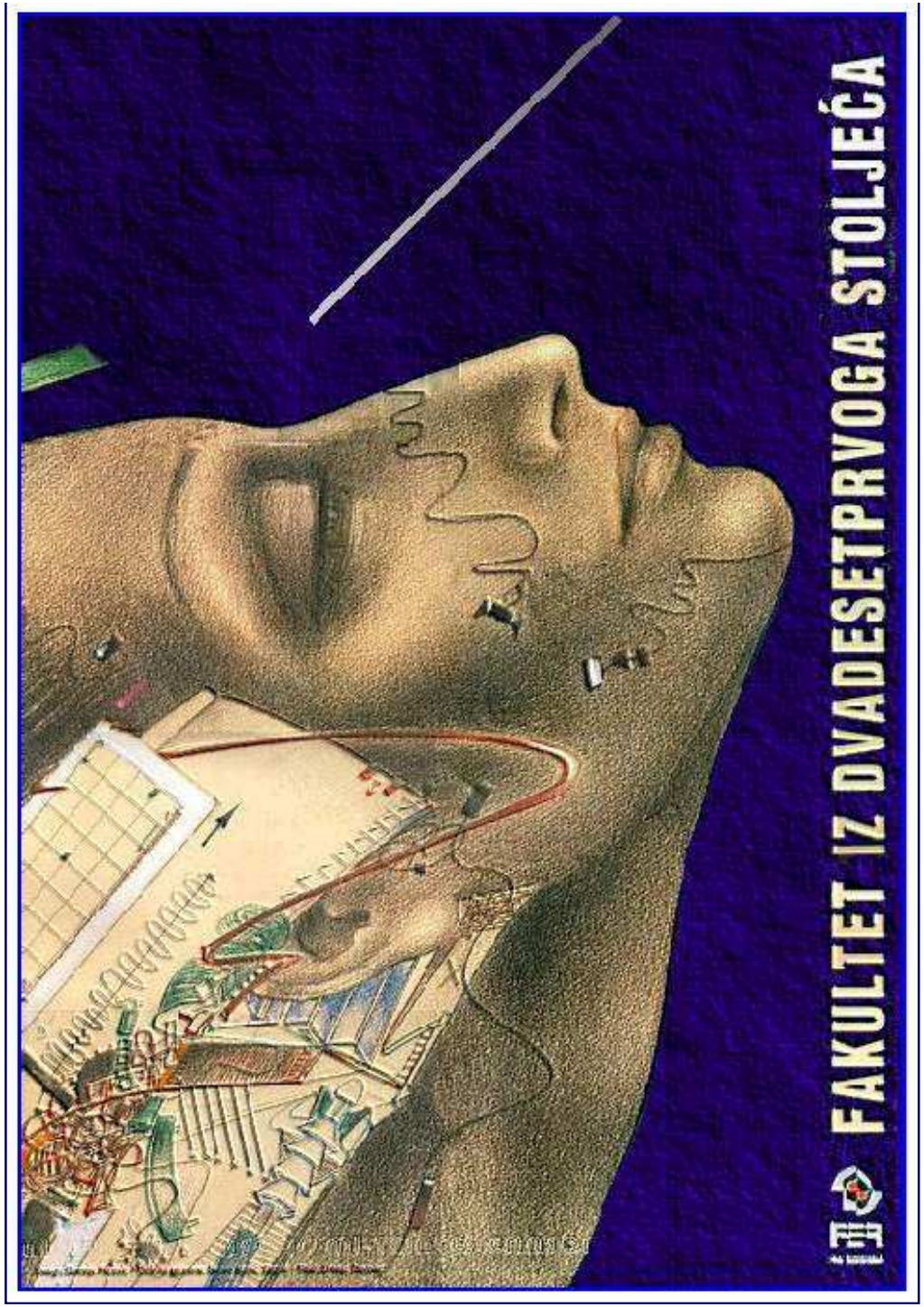
Napredni modeli i baze podataka

Predavanja

10.
Skladišta podataka
Dubinska analiza podataka
Poslovna inteligencija

Studeni 2008.



Sadržaj

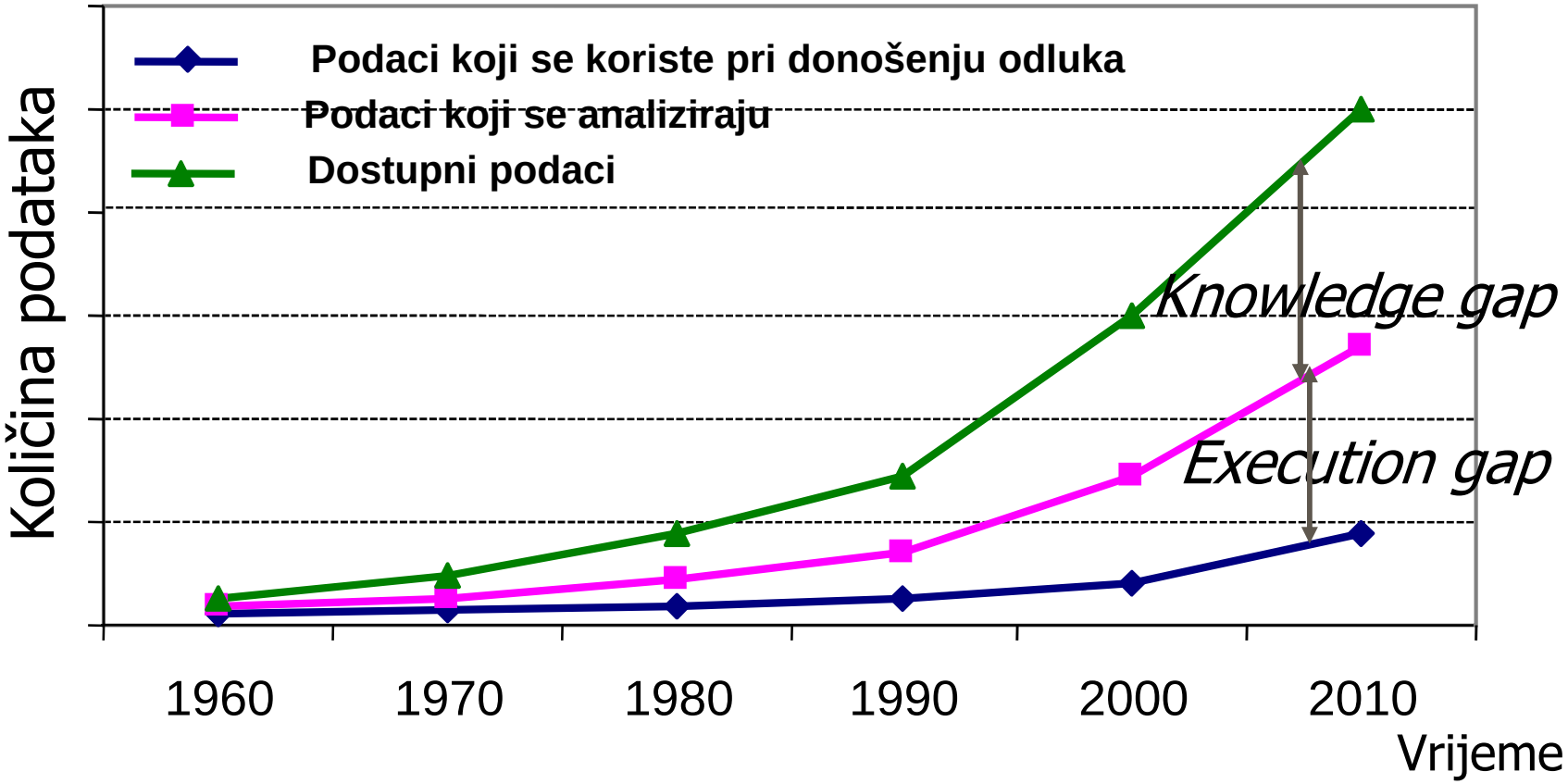
- Hijerarhija podataka, nedostatak relacijskog modela
- Skladišta podataka, dimenzijski model
- OLAP
- *Data mining*
- *Business Intelligence*

Hijerarhija podataka

Raskorak između postojećih podataka, informacija i znanja o poslovanju

- postoji sve više podataka i informacija, ali nema dovoljno vremena za njihovu analizu
- poduzeća su prenatrpana podacima, ali nema dovoljno informacija za donošenje odluka
- potrebno je oblikovati procese koji će prikupljati podatke i transformirati ih u informacije ili znanje

Jaz između prikupljenih i upotrijebljenih podataka



- *Knowledge gap* – analizira se samo dio podataka
- *Execution gap* – zbog nedostatka vremena samo dio analiziranih podataka se koristi

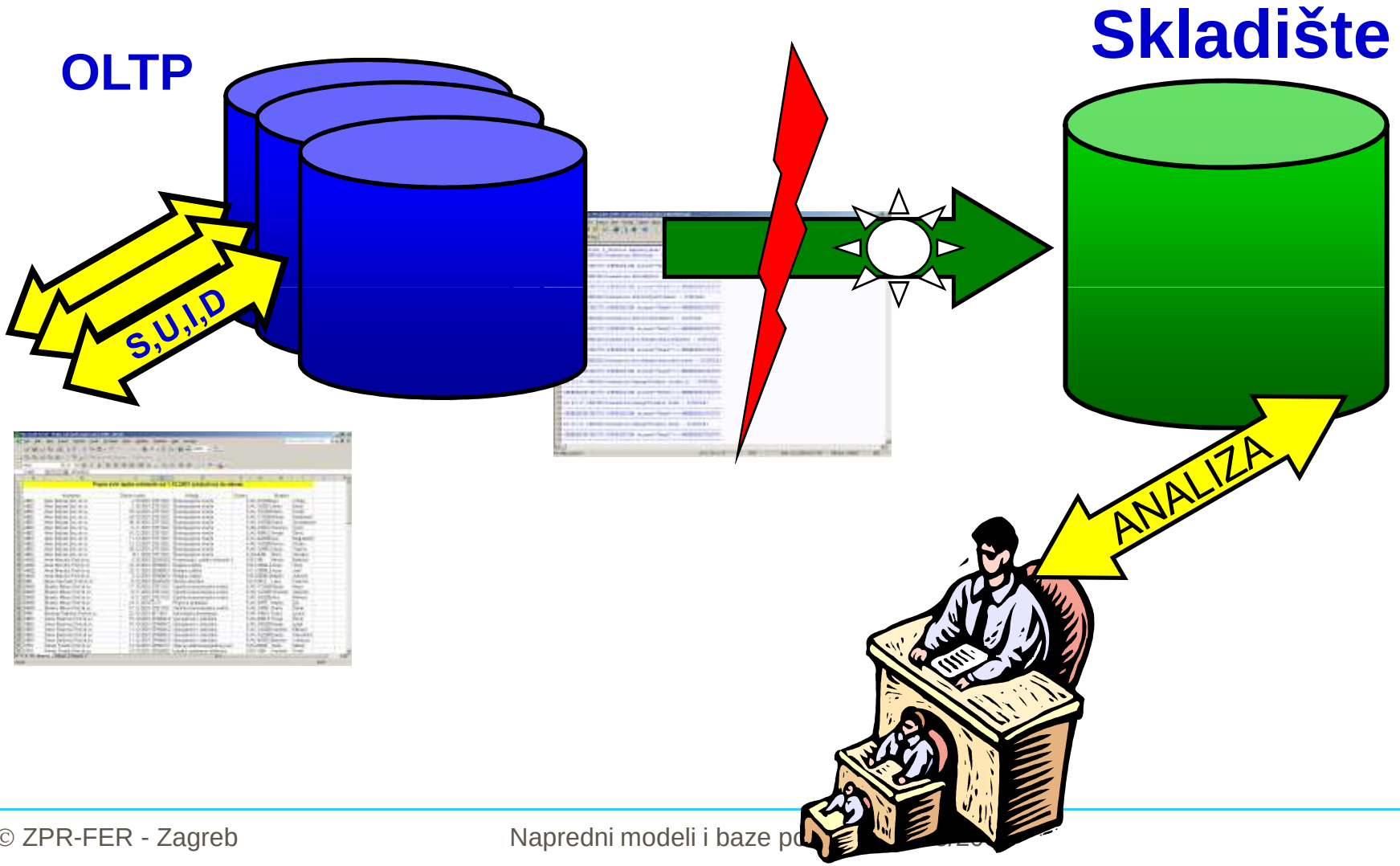
Relacijski model ne može pružiti bogate analitičke sposobnosti koje moderna poslovanja zahtijevaju

- Edgar F. Codd:

"*Attempting to force one technology or tool to satisfy a particular need for which another tool is more effective and efficient is like attempting to drive a screw into a wall with a hammer when a screwdriver is at hand: the screw may eventually enter the wall but at what cost?*"

Skladišta podataka

Osnovna ideja



Skladište podataka je …

- Skladište podataka jest subjektno orijentiran, integriran, postojan i vremenski različit skup podataka koji služi kao potpora odlučivanju. (B. Inmon)

- Kopija transakcijskih podataka specijalno strukturirana za upite i analize (R. Kimball)

- Jedinstven, kompletan i dosljedan repozitorij podataka pribavljen iz raznih izvora i predstavljen krajnjem korisniku na razumljiv način (B. Devlin)

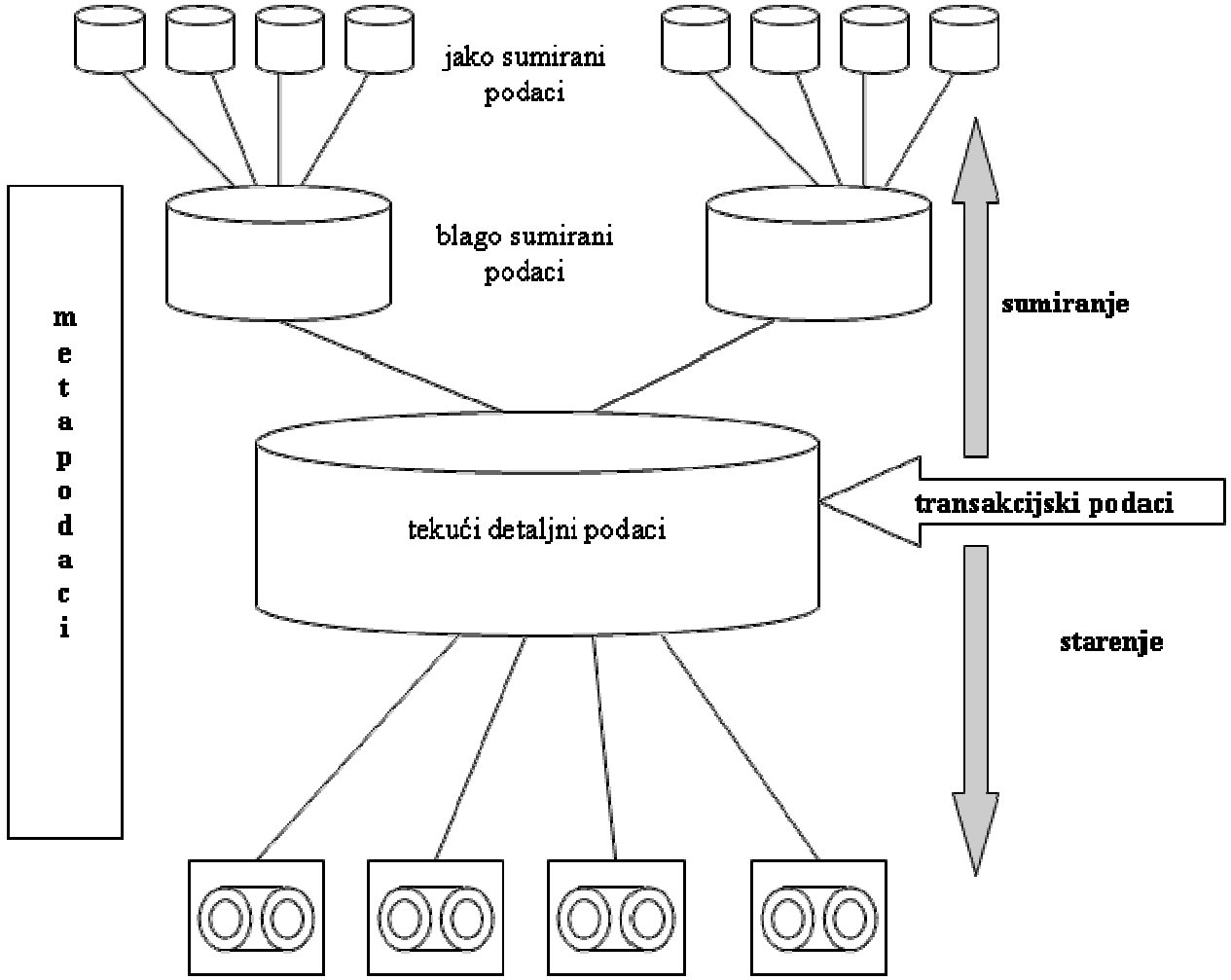
Razlike između transakcijskog sustava i skladišta podataka

Transakcijski sustav	Skladište podataka
Sadrži trenutne podatke	Sadrži povijesne podatke
Sadrži detaljne podatke	Sadrži detaljne i sumarne podatke
Podaci su promjenjivi	Podaci su postojani
Velika učestalost transakcija	Srednja i mala učestalost
Predvidljivi načini korištenja	Nepredvidljivi načini korištenja
Orijentiran ka dnevnim operacijama i vođenju poslovnog sustava	Orijentiran ka analizi podataka

Razlike između transakcijskog sustava i skladišta podataka (2)

Transakcijski sustav	Skladište podataka
Potpora dnevnim, operativnim odlukama	Potpora strateškim odlukama
Poslužuje velik broj operativnih korisnika	Poslužuje manji broj korisnika obično pozicioniranih u upravljačkim strukturama poduzeća
Izuzetno važna raspoloživost	Manje važna raspoloživost
Težište na pohranjivanju podatka	Težište na dobavljanju informacija

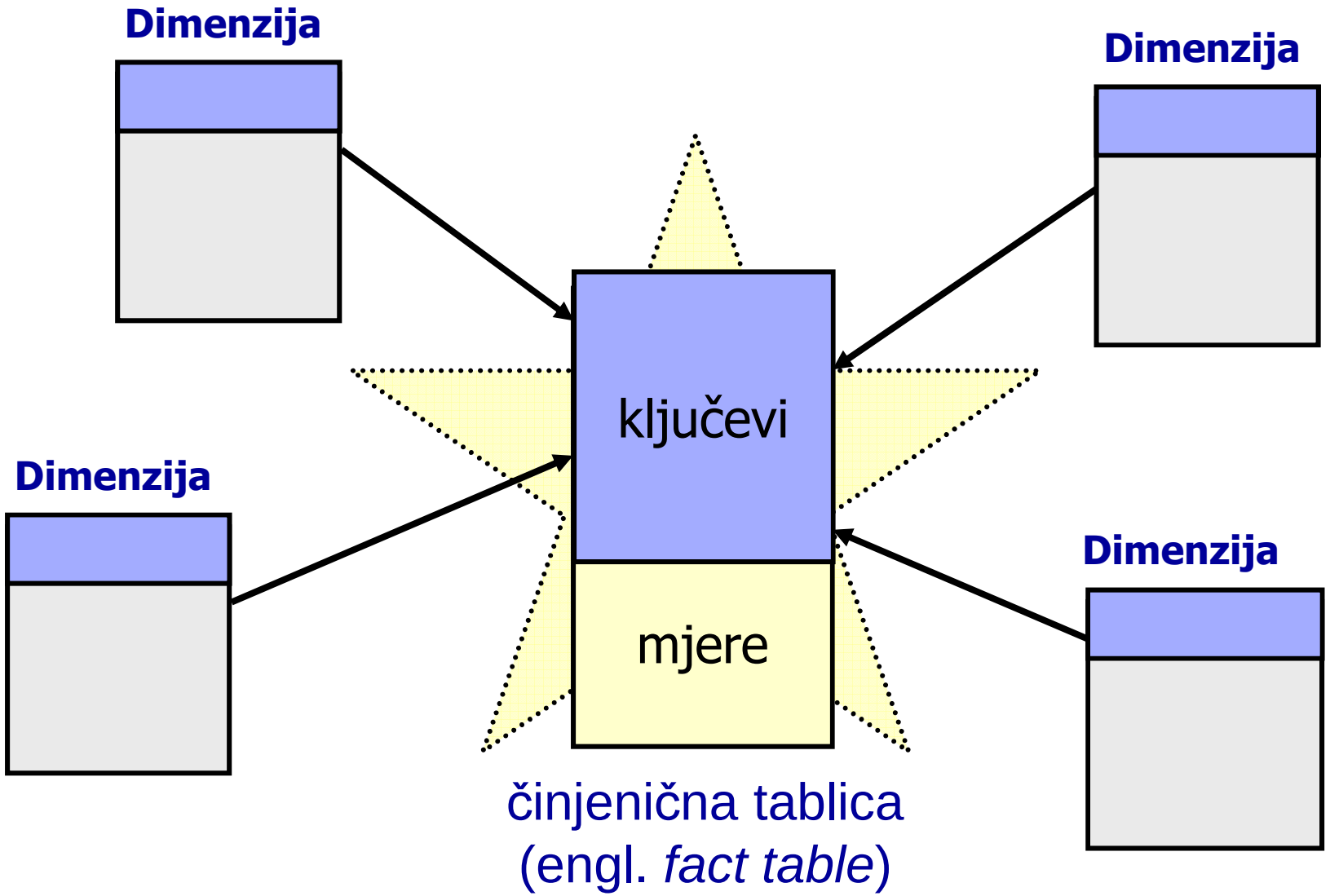
Značajna razlika između transakcijskog sustava i sustava skladišta podataka jest u zrnatosti (engl. *granularity*) pohranjenih podataka



Podaci su u skladištu podataka pohranjeni u dimenzijskom modelu

- Dimenzijski model predstavlja podatke u jednostavnom, intuitivnom, obliku koji dopušta vrlo učinkovit pregled
- Dimenzijski model nije normaliziran
- Zvjezdasti model (engl. *Star join*) i pahuljasti model (engl. *Snowflake*)

Zvjezdasti model



Svojstva činjeničnih i dimenzijskih tablica

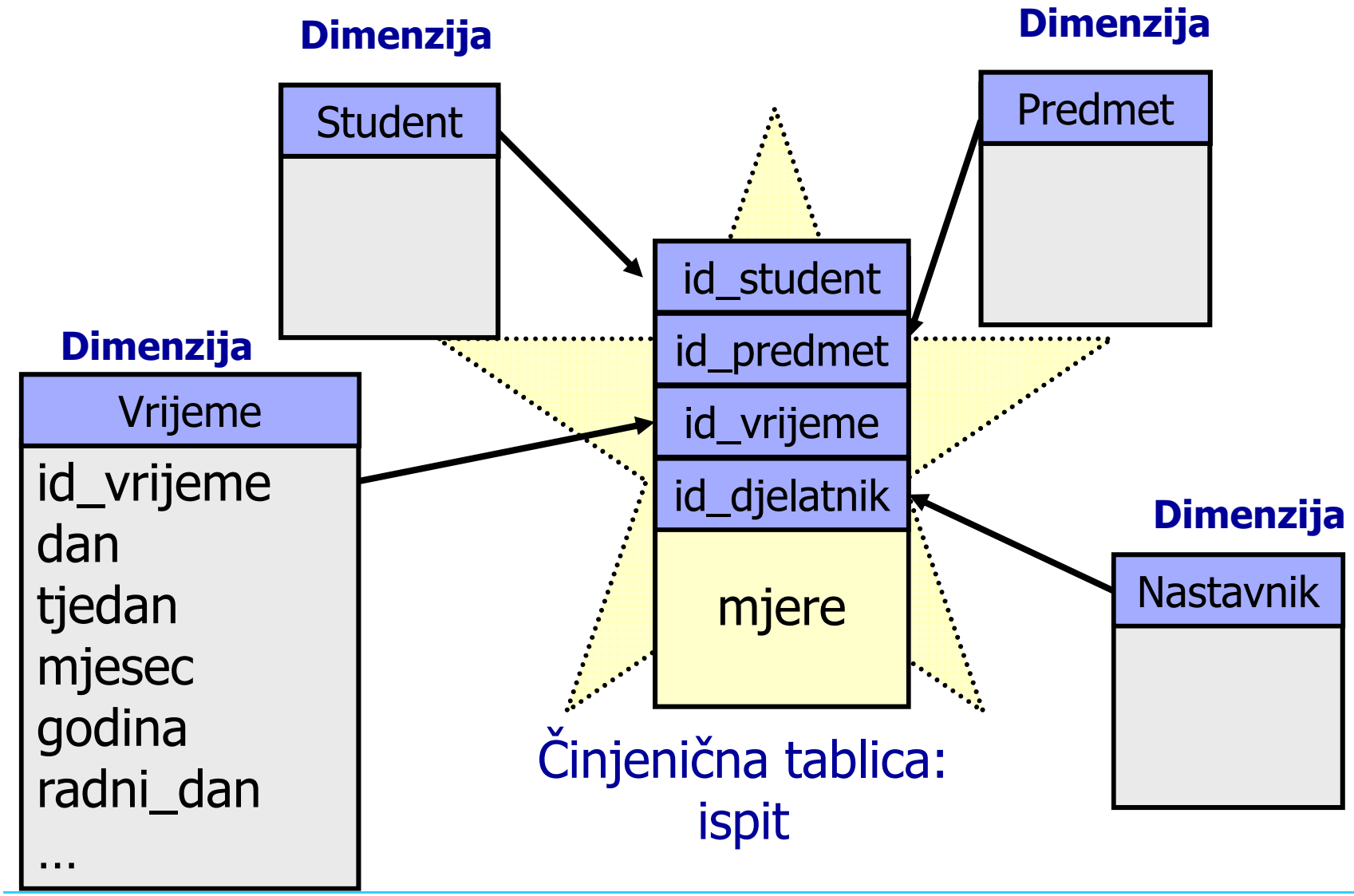
- Činjenična tablica se sastoji od dvije skupine numeričkih atributa:
  - (strani) ključevi dimenzijskih tablica
  - mjere – numerički atributi nad kojima se primjenjuju agregatne funkcije i koji daju "ocjenu" procesa koji se prati činjeničnom tablicom (npr. `cijena`, `količina`, `ocjenaIspit`, ...)
- Činjenična tablica je normalizirana (ili skoro normalizirana – ponekad ima izvedene atribute, npr. `cijena` i `cijenasPDVom`)
- Dimenzijske tablice predstavljaju subjekte (objekte) koji sudjeluju u procesu koji se prati (npr. nastavnik, student, predmet, ...)
- Dimenzijske tablice **nisu** normalizirane – tipično imaju povećani broj atributa

Svojstva činjeničnih i dimenzijskih tablica

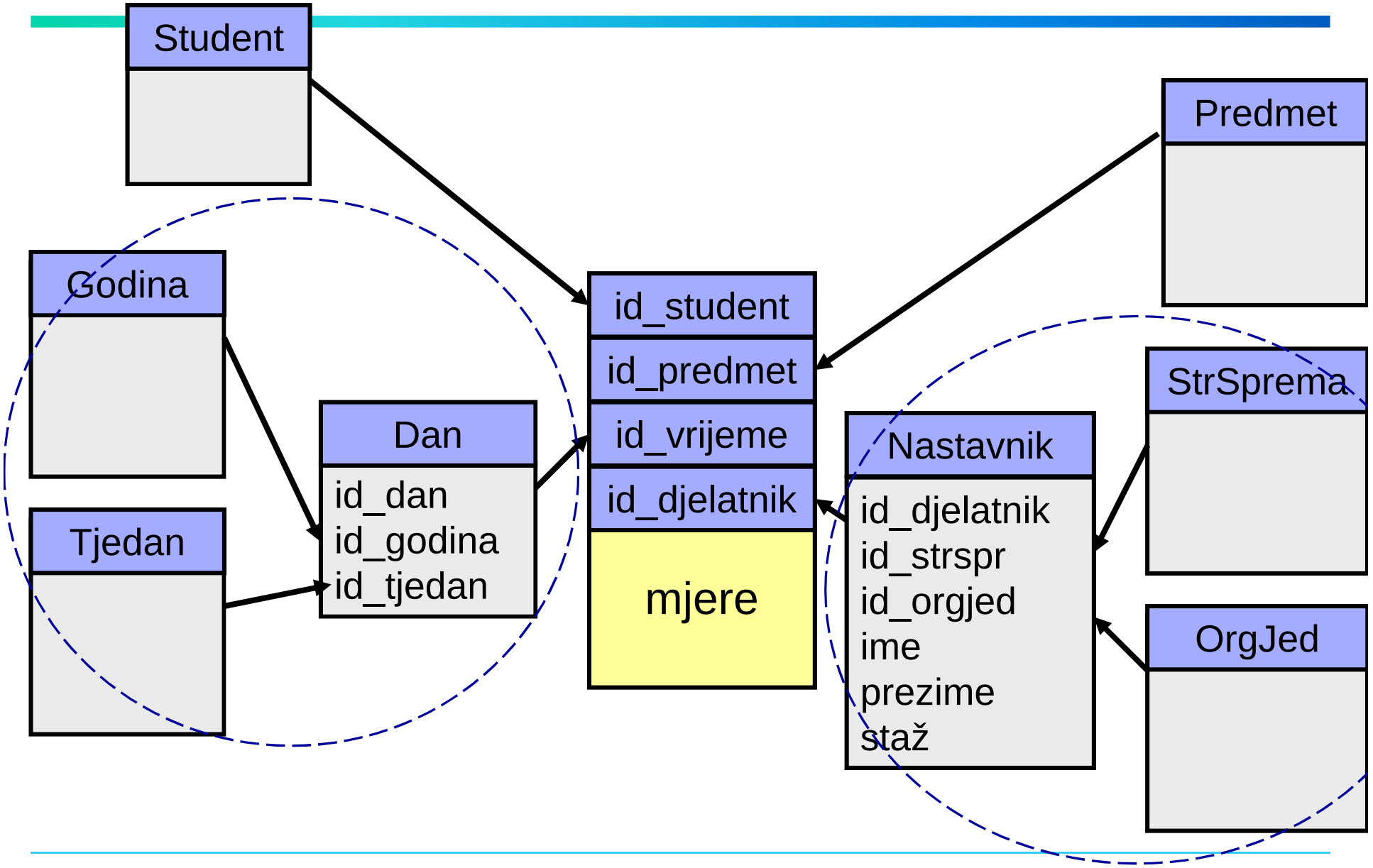
- Činjenične tablice:
  - velik broj zapisa (10^5, 10^6, 10^7, ...)
  - jedan redak (n-torka) ne zauzima puno prostora (normalizirana tablica, numerički atributi)

- Dimenzijske tablice:
  - mali broj zapisa ($<10^5$, većinom $<10^2$)
  - veliko zauzeće prostora po jednom retku (npr. 30-tak atributa, od toga 15-tak znakovni nizovi: 30*4 + 15 * 100 = 1620 byte)

Zvjezdasti model



Pahuljasti model



Prednosti dimenzijskog modela

- Predvidljiva, standardna struktura, omogućuje izradu standardnih alata za analizu
- Predvidljivi zvjezdasti spoj je otporan na neočekivana korisnička ponašanja
- Model se može lako proširiti (nove mjere, dimenzije, dimenzijski atributi)
- Postoji skup standardnih pristupa za obrađivanje uobičajenih situacija u poslovnom svijetu

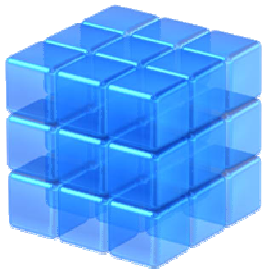
Najzahtjevniji proces pri skladištenju podataka jest ETL proces

- ETL (engl. *Extraction, Transformation and Loading*) jest postupak:
  - **Ekstrakcija podataka iz raznorodnih, najčešće transakcijskih, izvora podataka**
  - **Transformacija podataka, uključujući čišćenje, agregaciju i filtriranje**
  - **Učitavanje transformiranih i ujedinjenih podataka**

- Međuspremnik – engl. *staging area*

OLAP
(engl. *On-Line Analytical Processing*)

- Pristup analizi i izvještavanju koji omogućuje korisniku da lako i selektivno izdvaja i pregledava podatke sa različitih stajališta temeljeno na multidimenzijskoj strukturi podataka zvanoj kocka (engl. *cube*).
- FASMI (N.Pendse, R.Creeth, 1995.):
  - *Fast*
  - *Analysis*
  - *Shared*
  - *Multidimensional*
  - *Information*



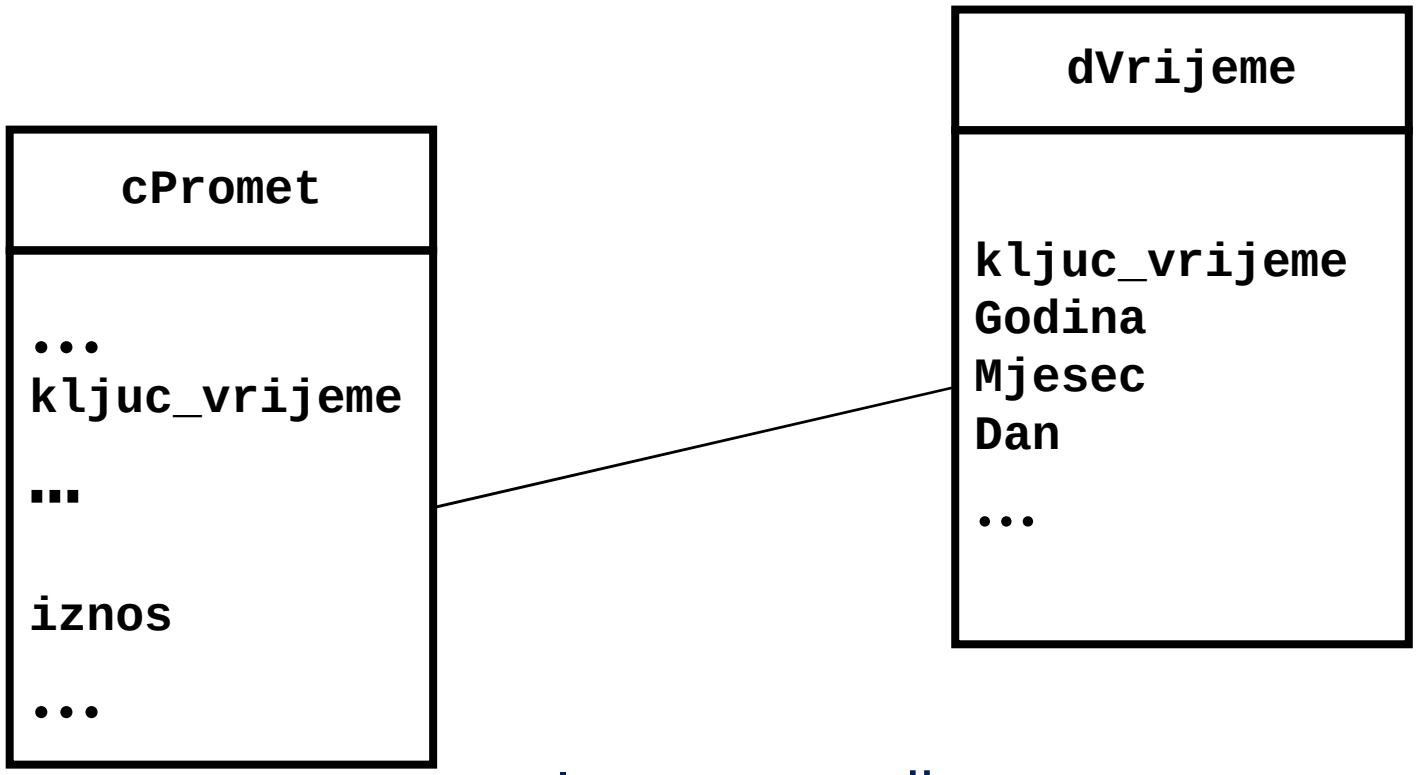
Analysis, Shared

- *Analysis* (analiza) znači da se sustav mora nositi s bilo kojom poslovnom logikom i statističkom analizom koja je bitna korisniku, pri čemu ne smije biti presložena za krajnjeg korisnika (npr. ne smije se očekivati od krajnjeg korisnika da zna programirati u nekom 4GL jeziku kako bi ostvario neki izračun).
- *Shared* (dijeljeni) znači da sustav ostvaruje sve sigurnosne zahtjeve za tajnošću podataka (po mogućnosti na razini ćelije) i, ako je omogućeno upisivanje u skladište, odgovarajuće zaključavanje podataka

Multidimensional, Information

- Sustav mora omogućiti multidimenzijski konceptualni pogled na podatke, uključujući punu potporu za (višestruke) hijerarhije
- Nije bitno (uvjetno govoreći) koja se tehnologija upravljanja i pohranjivanja podataka koristi.
- Information - svi potrebni podaci i izvedene informacije potrebne, gdjegod se nalazile i kolikogod te informacije bile bitne za aplikaciju. Zanimljivo je koliko ulaznih podataka sustav može podnijeti, a ne koliko gigabajta troši na pohranjivanje podataka

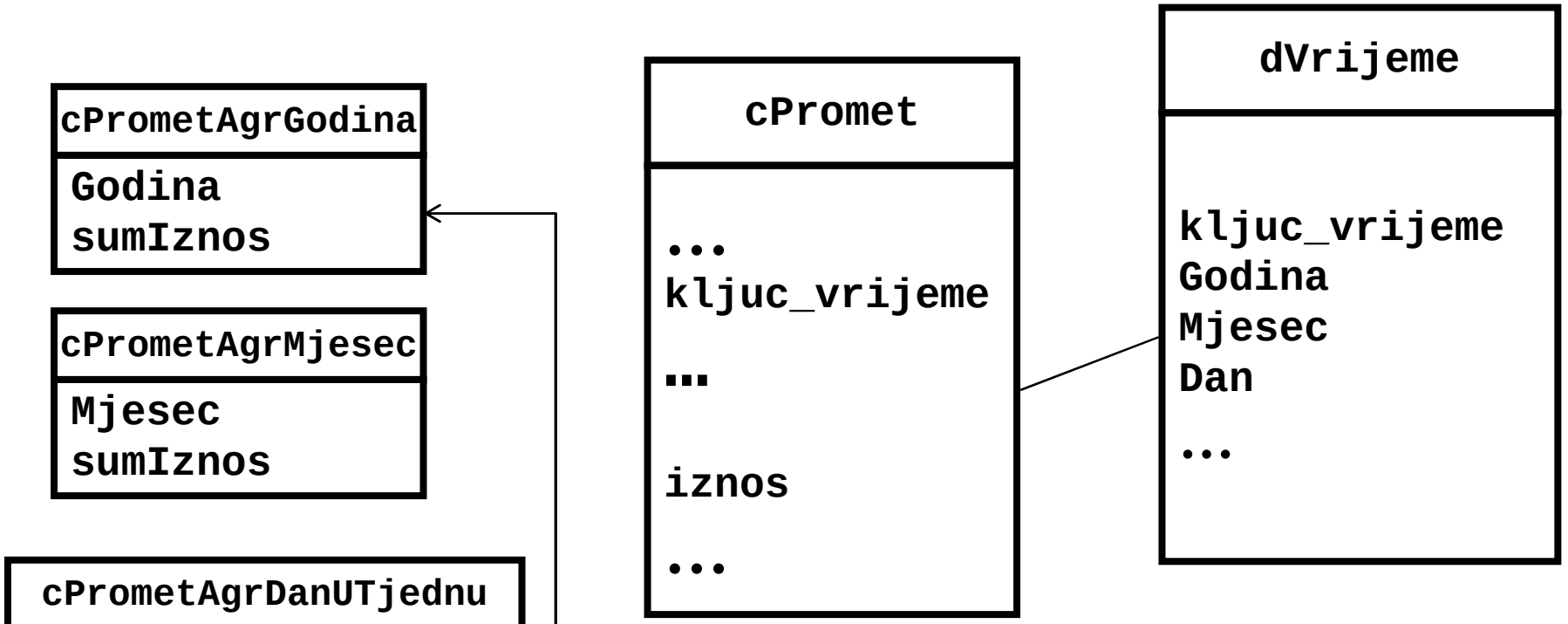
Kako dobiti na brzini?



Iznos po godinama:

```
SELECT Godina, sum(iznos)
FROM cPromet, dVrijeme …
```

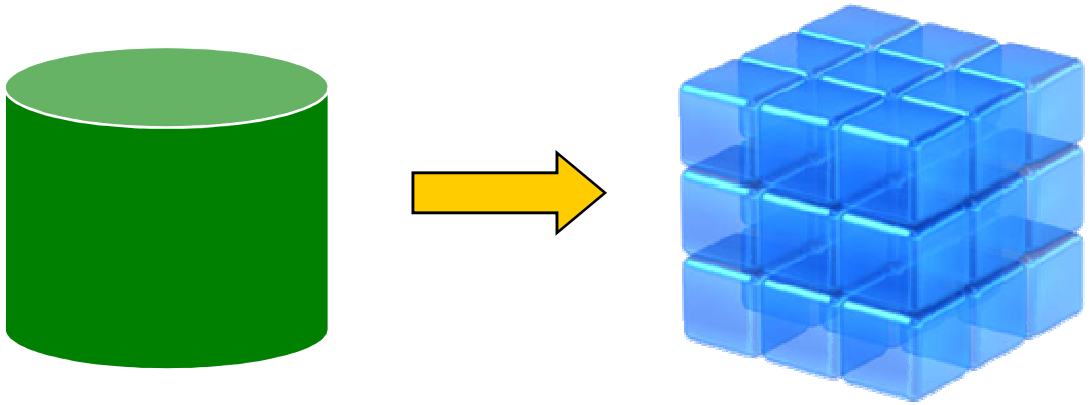
Agregacija je temeljna tehnika kojom OLAP sustavi ostvaruju brzinu



Iznos po godinama:

```
SELECT Godina, sum(iznos)
FROM cPromet, dVrijeme …
```

ROLAP, MOLAP, HOLAP

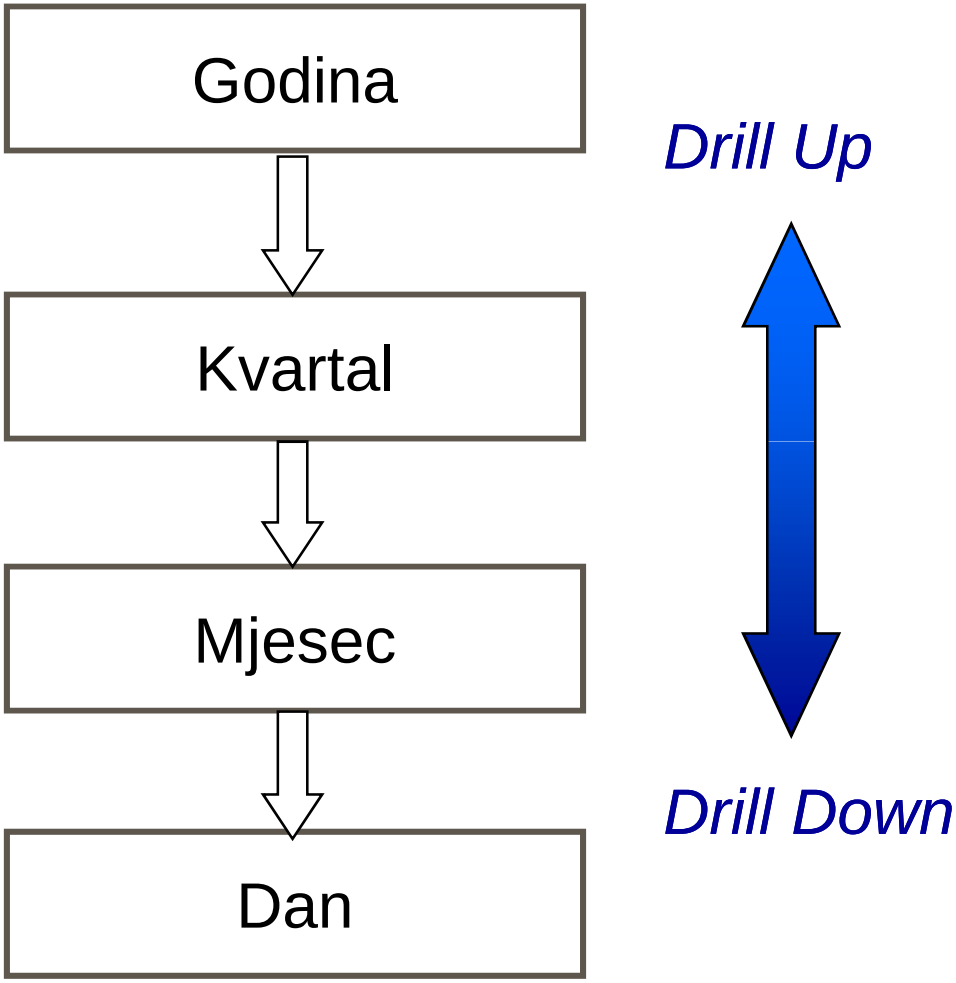


Tri modela:

1. ROLAP (***Relational OLAP***) – agregati i podaci su u relacijskoj bazi
2. MOLAP (***Multidimensional OLAP***) – agregati i podaci su u OLAP serveru
3. HOLAP (***Hybrid OLAP***) –agregati su u OLAP serveru, podaci u relacijskoj bazi

OLAP sustav mora omogućiti multidimenzijski konceptualni pogled na podatke, uključujući punu potporu za hijerarhije

1998				3,38
1999				3,51
2000	1	ožujak	1.3.2000	4,00
			13.3.2000	3,94
			14.3.2000	4,06
			15.3.2000	3,94
			16.3.2000	4,18
			17.3.2000	4,13
			20.3.2000	4,00
			21.3.2000	4,67
			22.3.2000	4,50
			23.3.2000	4,50
			24.3.2000	5,00
			27.3.2000	5,00
			30.3.2000	5,00
			6.3.2000	4,75
			7.3.2000	4,50
			9.3.2000	5,00
		ožujak Total *		4,16
		siječanj		3,71
		veljača		3,72
	1 Total *			3,74
	2			3,62
	3			3,14
	4			3,32
2000 Total *				3,52



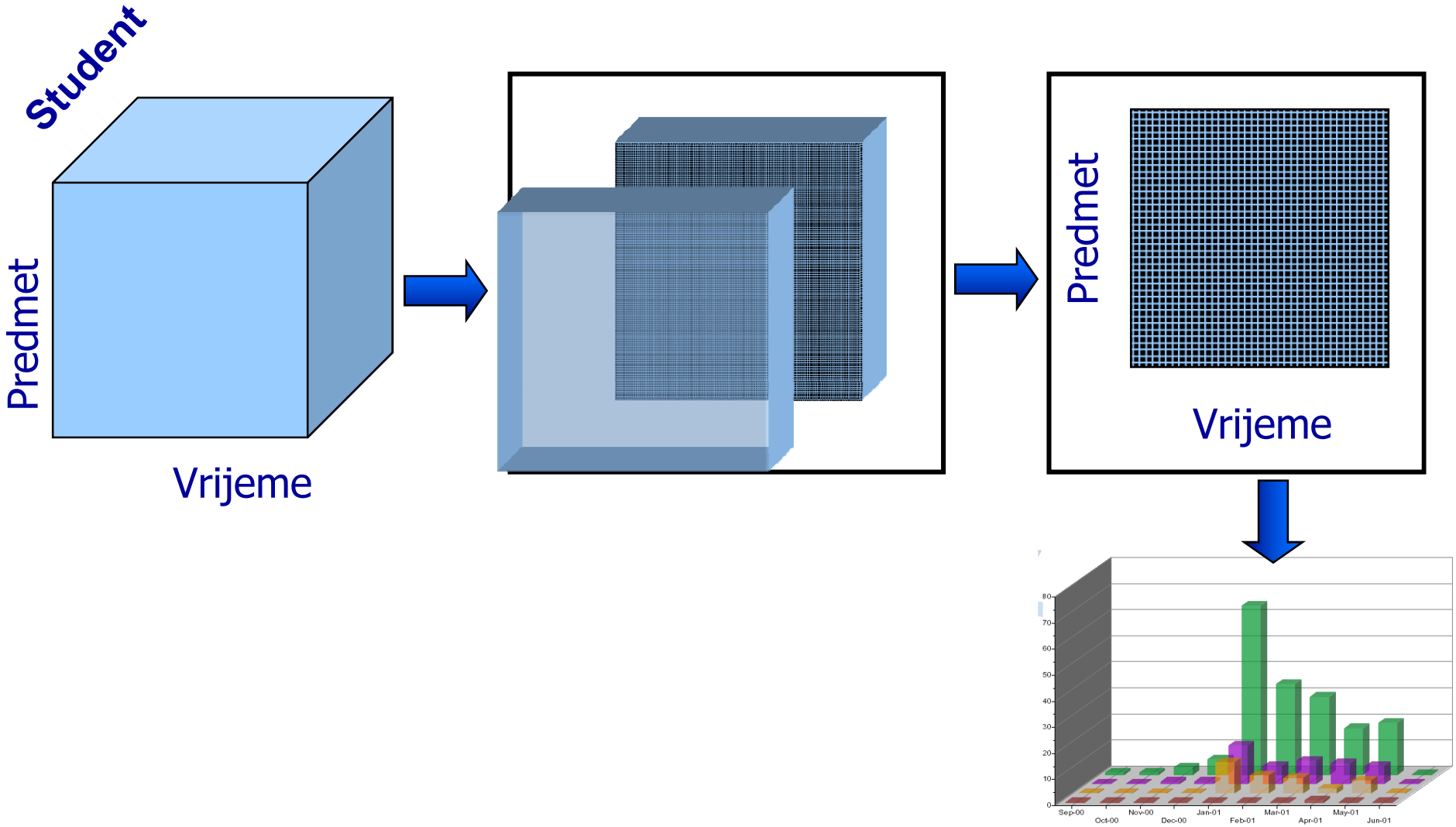
OLAP sustav:
hijerarhijski pregled i *crossjoin*

		F	M
		Profit	Profit
Drink	CA	4.277,12	4.263,85
	OR	3.601,07	3.699,87
	WA	6.828,08	6.688,99
	Bellingham	128,12	117,86
	Bremerton	1.291,72	1.469,36
	Seattle	1.516,20	1.107,71
	Spokane	1.374,13	1.209,38
	Tacoma	1.829,69	1.850,87
	Walla Walla	106,24	106,30
	Yakima	581,98	827,51
Food	CA	34.519,91	34.692,91
	OR	30.041,63	31.555,59
	WA	57.539,19	57.415,64
	Bellingham	1.001,45	1.033,34
	Bremerton	10.921,87	12.210,04
	Seattle	12.442,67	10.683,56
	Spokane	11.246,38	10.411,18
	Tacoma	16.323,86	16.173,90
	Walla Walla	971,19	1.143,21
	Yakima	4.631,77	5.760,42
Non-Consumable	CA	8.662,15	9.221,48
	OR	8.318,24	8.288,18
	WA	14.661,36	15.335,65
	Bellingham	262,85	299,00
	Bremerton	2.830,95	3.050,39
	Seattle	3.157,43	2.779,70
	Spokane	2.719,97	2.877,94
	Tacoma	4.241,42	4.464,94
	Walla Walla	232,21	266,50
	Yakima	1.216,54	1.597,19

OLAP sustav:
hijerarhijski pregled i *crossjoin*

		Makedonija		Njemačka		Republika Hrvatska		Slovenija	
		Broj ispita	Prosjek oc...	Broj ispita	Prosjek oc...	Broj ispita	Prosjek oc...	Broj ispita	Prosjek oc...
M	1997					365	3,298		
	veljača (2)					60	3,525		
	ožujak (3)					10	3,375		
	travanj (4)					34	3,176		
	svibanj (5)					5	5,000		
	lipanj (6)					48	3,739		
	srpanj (7)					72	3,070		
	rujan (9)					51	2,980		
	listopad (10)					39	2,853		
	studeni (11)					3	3,000		
	prosinac (12)					43	3,512		
	1998					610	3,490		
	1999			6	3,600	1.176	3,518		
	2000	4	2,750	7	3,800	2.353	3,439	1	
	2001	5	3,000	5	3,500	5.133	3,105	4	2,500
Ž	1997					682	2,992		
	siječanj (1)					8	4,000		
	23.1.1997 ,četvrtak					1	3,000		
	28.1.1997 ,utorak					1	2,000		
	29.1.1997 ,srijeda					3	5,000		
	31.1.1997 ,petak					3	4,000		
	veljača (2)					104	2,960		
	ožujak (3)					21	3,238		
	travanj (4)					58	2,948		
	svibanj (5)					29	4,724		
	lipanj (6)					82	3,512		
	srpanj (7)					124	2,758		
	rujan (9)					104	2,515		
	listopad (10)					86	2,720		
	studeni (11)					2	3,000		
	prosinac (12)					64	2,917		
	1998					1.031	2,975		
	1999					1.867	3,249		
	2000					3.585	3,318		
	2001					6.965	3,231	7	2,750
	2004					3	4,500		

OLAP sustav: *slice & dice* način pregledavanja



Dubinska analiza
(engl. *Data Mining*)

Neke definicije dubinske analize

- netrivijalni postupak identifikacije valjanih, novih, potencijalno korisnih i prije svega razumljivih načela u podacima (**Fayyad**)

- proces otkrivanja i izlučivanja znanja iz velikih baza podataka (**Zornes**)

Zašto danas?

- veća snaga i brzina računala
- smanjena cijena pohrane podataka
- izgrađena skladišta podataka (60%-90% posla)

Primjer: kreditna sposobnost osobe

Početni skup podataka	Godine	Prihod	God. staža	OK
	41	9000	8	D
	32	4000	5	D
	26	2500	2	N

Testni skup podataka	Godine	Prihod	God. staža	OK	Model
	39	9000	4	D	N
	29	4000	5	D	D

Neke primjene (1)

- **Financijske aplikacije**
  - Da li je ovaj klijent kreditno sposoban?
  - Tko bi mogao proglasiti bankrot?
- **Proizvodne aplikacije**
  - Gdje se može uštediti u materijalu?
- **Medicinske aplikacije**
  - Koji su pacijenti kandidati za odustajanje od terapije?
  - Koji je pacijent kandidat za operaciju?

Neke primjene (2)

- **Farmaceutske aplikacije**
  - Povezivanje kemijske strukture s kemijskim svojstvima
  - Povezivanje gena i bolesti
- **Osiguravajuća društva**
  - Otkrivanje prijevara
- **E-trgovina**
  - Što ponuditi kao novi proizvod?
- **Održavanje**
  - Što se kvari i zašto?

Problemi pri realizaciji

- Veličina skupa podataka
  - velik broj redaka
  - velik broj atributa
- Podaci s prisutnim šumom
- Podaci kojima nedostaju vrijednosti pojedinih atributa (ponekad je i to vrijedan podatak!)
- Prepoznavanje atributa (i)relevantnih za analizu

OLAP i dubinska analiza

- **OLAP**
  - analiza podataka kako bi se potvrdile ili opovrgle pretpostavke
  - opisni model; ne omogućuje predviđanje
  - deduktivni pristup – korisnik predloži hipotezu i zatim je pokuša potkrijepiti podacima, odnosno nizom upita
- **Dubinsko istraživanje podataka**
  - pronalaženje novih veza među podacima ili predviđanje novog
  - opisni model koji omogućuje predviđanje
  - induktivni pristup
- **OLAP i dubinska analiza se nadopunjuju**:
  - OLAP – upoznavanje s podacima
  - DM – otkrivanje uzoraka
  - OLAP – analiza implikacija uporabe otkrivenih modela

Neki osnovni modeli

- **Regresija (numeričko predviđanje)**
  - Linearna regresija $f(x) = ax + b$
  - Logistička regresija $f(x) = 1 / (1 + e^{-x})$
- **Razvrstavanje** (kategoričke ili diskretne vrijednosti) (engl. *classification*)
  - Stabla odlučivanja
  - Genetski algoritmi
  - Neuronske mreže
  - K-najbližih susjeda
- **Grupiranje** (engl. *clustering*)
- **Asocijativna pravila** (engl. *association rules*)

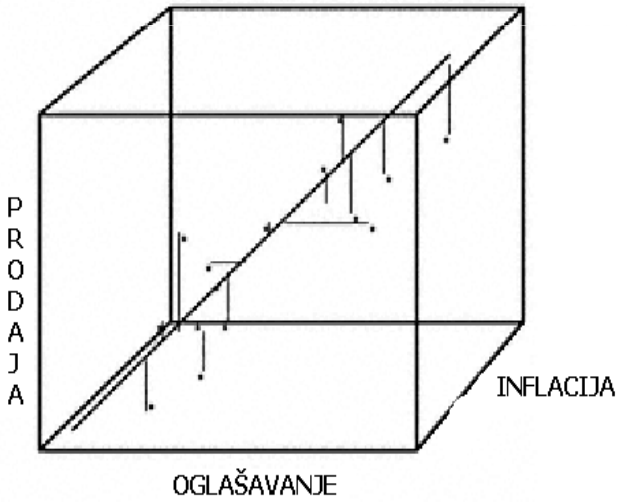
Linearna regresija

- Linearna regresija je statistička tehnika koja kvantificira odnos između dvije kontinuirane varijable:
  - zavisna varijabla koju pokušavamo predvidjeti (npr. prodaja)
  - nezavisna varijabla, prediktor (npr. oglašavanje)

- Minimiziranje kvadratne pogreške

Prodaja = 415.60 + 7.90* Oglašavanje + 12781* Inflacija

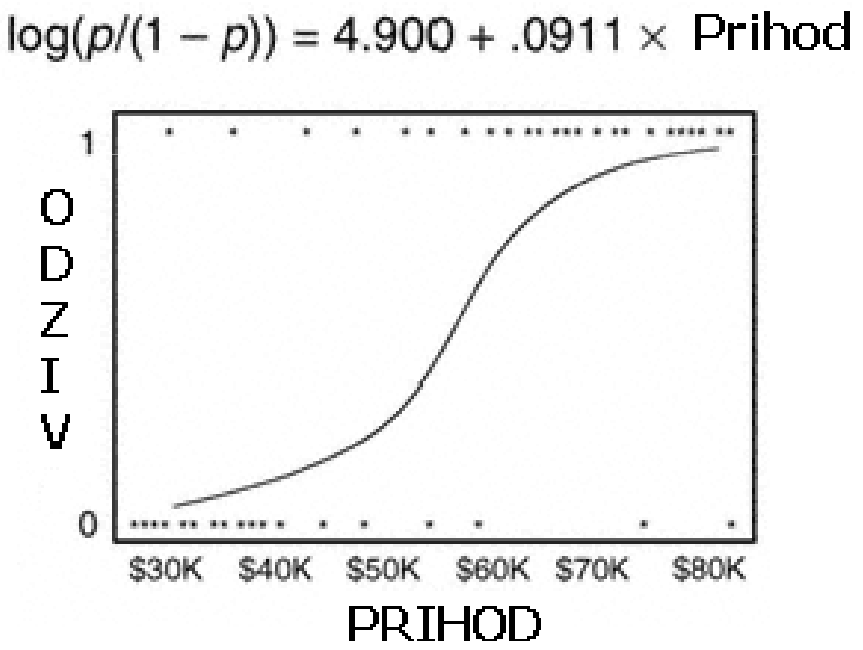
Oglašavanje	Inflacija	Prodaja
$120	3.4%	$1,503
$160	3.3%	$1,755
$205	3.6%	$2,971
$210	3.5%	$1,682
$225	3.4%	$3,497
$230	3.3%	$1,998
$290	3.2%	$4,528
$315	3.3%	$2,937
$375	3.3%	$3,622
$390	3.4%	$4,402
$440	3.2%	$3,844
$475	3.1%	$4,470
$490	3.2%	$5,492
$550	3.2%	$4,398



Logistička regresija

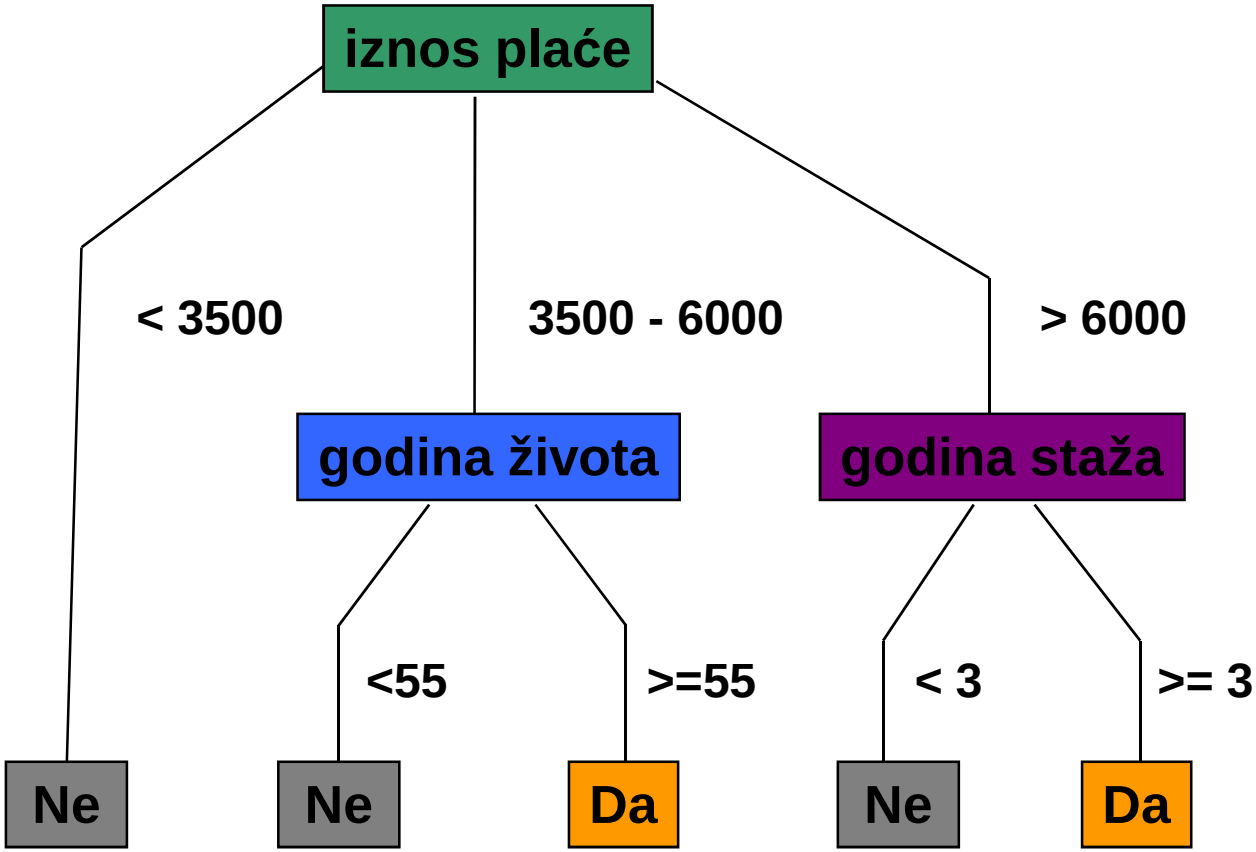
- Često se koristi u medicini, društvenim znanostima, marketingu
- Zavisna varijabla nije kontinuirana već diskretna ili kategorička
- p je uvijek u intervalu [0, 1]

- p – vjerojatnost događaja
- omjer vjerojatnosti: p/(1-p)
- log (p/(1-p)) je linearna funkcija prediktora
- -> p = $1/(1+e^{-lf(x)})$



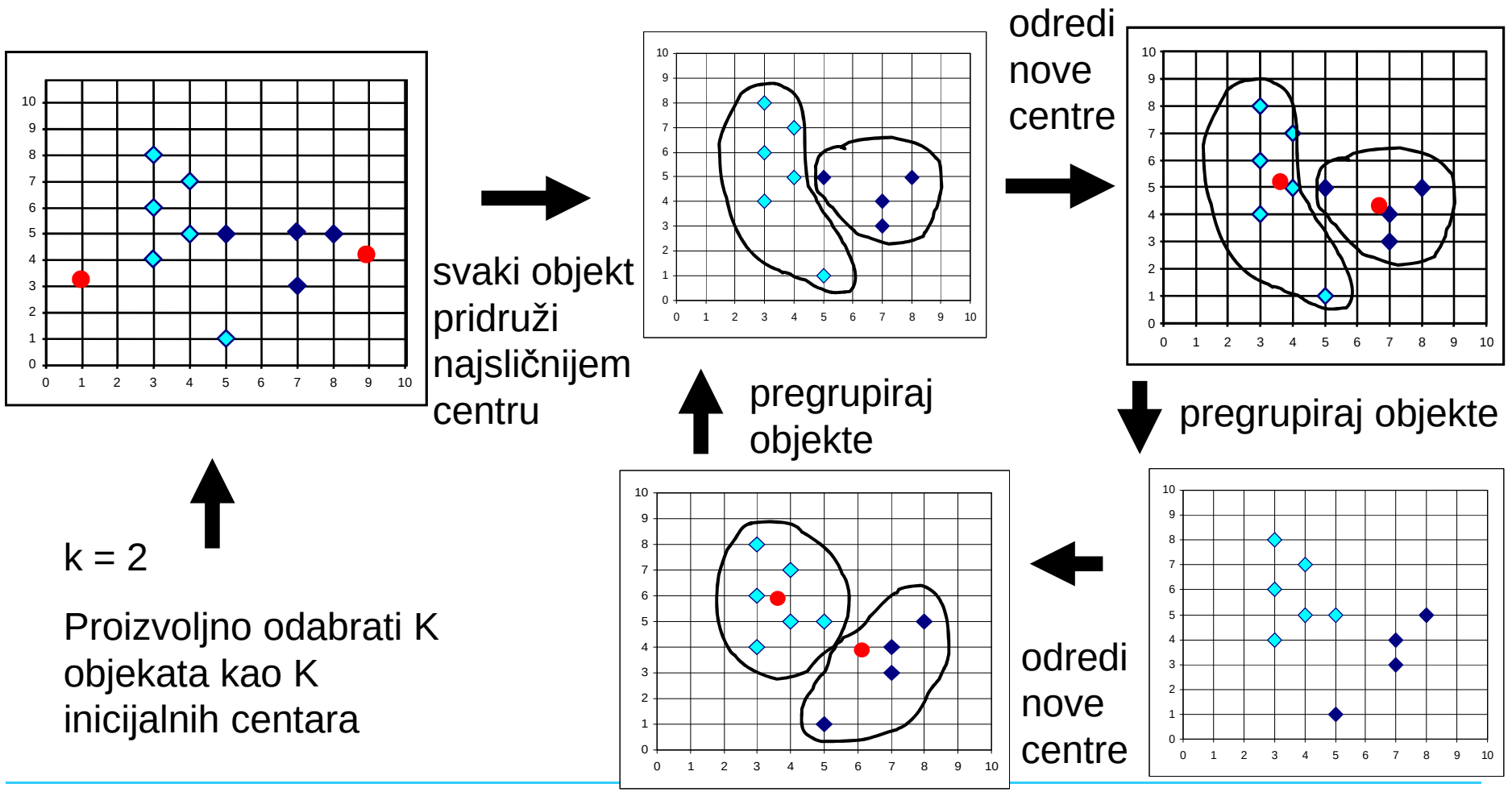
Razvrstavanje - izlazna vrijednost se predviđa na temelju ulaznih varijabli

- Stablo odlučivanja za određivanje kreditne sposobnosti:



Grupiranje

- **Primjer algoritma za grupiranje oko k centara:**



Asocijativna pravila

- asocijativno pravilo: A → B
- koristi se u transakcijskim sustavima
  - tipično: analiza potrošačke košarice (engl. *market-basket analysis*)
- Primjer:

  Kupuju (X, “pivo”) → *Kupuje* (X, “pelene”)

$$support = \frac{N_{A,B}}{N} = P(A\&B)$$

$$confidence = \frac{N_{A,B}}{N_A} = P(B|A)$$

Npr.
- 1000 transakcija
- 50 kupljenih pakiranja piva
- 80 kupljenih pakiranja pelena
- 20 kupljenih piva i pelena
- s = 20/1000 = 2%
- c = 20/50 = 40%

Mitovi dubinske analize

- Daje odgovore na nepostavljena pitanja
- Uklanja potrebu za razumijevanjem poslovanja
- Uklanja potrebu za skupljanjem kvalitetnih podataka
- Nije potrebno posjedovati dobre analitičke sposobnosti

"Teoremi"

- George Box
  - *All models are wrong, but some are useful*
  - *Statisticians, like artists, have the bad habit of falling in love with their models*
- Model nije bolji od podataka
- Twyman’s law
  - *If it looks interesting, it’s probably wrong.*
- De Veaux’s corollary
  - *If it’s not wrong, than it’s probably obvious*

Poslovna inteligencija
(engl. Business Intelligence)

Business Intelligence

- Tipična organizacija analizira samo 10% podataka koje prikupi
- BI jest način kako iskoristiti preostalih 90%
- Upravljanje poslovnim informacijama
- Pretvaranje podataka u informacije
- BI nije proizvod, to je koncept

Business Intelligence

- BI je široka kategorija aplikacija i tehnologija za skupljanje, pohranjivanje, analiziranje i dijeljenje informacija u svrhu donošenja boljih poslovnih odluka
- BI aplikacije uključuju:
  - sustave za podršku odlučivanju
  - upite i izvještavanje
  - OLAP
  - statistička analiza
  - predviđanje
  - dubinska analiza

Pitanja