Data mining – kocepti i tehnike

- Udžbenik:
  - Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber
  - Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar
- Ocjenjivanje: kolokvijumi (35% + 35%), prezentacija (5%), projekat (10%), domaći zadaci (15%)

Pregled kursa

- Glava 1: Uvod
- Glava 2: Data warehousing i OLAP
- Glava 3: Pre-procesiranje podataka
- Glava 4: Arhitektura data mining sistema
- Glava 5: Opisivanje koncepata, karakterizacija i poređenje
- Glava 6: Asocijativna analiza
- Glava 7: Klasifikacija i regresija
- Glava 8: Klasterizacija
- Glava 9: Rudarenje složenih tipova podataka
- Glava 10: Primjene i trendovi razvoja data mining-a

Glava 1. Sadržaj

- Motivacija: zašto data mining?
- Šta je data mining?
- Data mining: koji tipovi podataka?
- Data mining tehnike
- Šta je znanje?
- Klasifikacija data mining sistema
- Osnovni koncepti data mining-a

Motivacija

- Eksplozija podataka: tehnologija baza podataka proizvela je velike količine podataka
- We are drowning in data, but starving for knowledge!
- Rješenje:
  - Data warehousing i OLAP
  - Rudarenje korisnog znanja (pravila, ograničenja itd.) iz podataka

Razvoj tehnologije baza podataka

- 1960: AOP
- 1970: Relacioni model i SUBP
- 1980: napredni modeli podataka i specijalizovani SUBP
- 1990: data mining, data warehousing

Šta je data mining?

- Data mining: pronalaženje korisnog znanja u velikim bazama podataka
- Knowledge Discovery in Databases (KDD)
- Šta nije data mining?
  - Procesiranje upita
  - Ekspertni sistemi
  - Sistemi za mašinsko učenje

Zašto data mining?

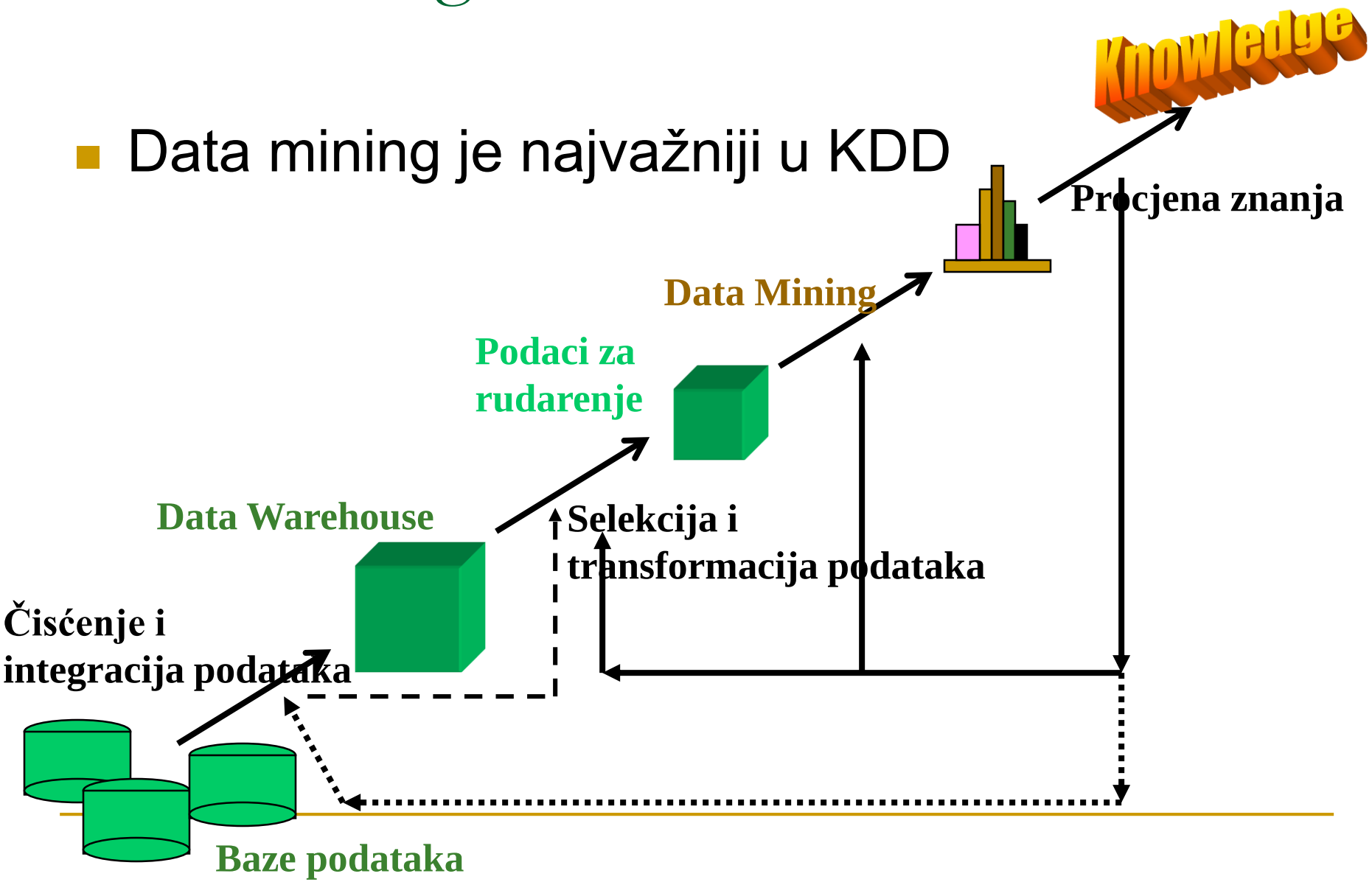
- Potencijalne primjene:
  - Upravljanje i analiza tržišta
  - Upravljanje i analiza rizika
  - Medicina
  - Nauka
  - Inžinjerstvo
  - ...

Potencijalne primjene

- Upravljanje i analiza tržišta
  - Kreiranje marketinških kampanja, analiza potrošačkih navika itd.
  - Izvor podataka mogu da budu baze transakcija velikih supermarketa
  - Rezultati analize
    - Grupe klijenata koji imaju zajedničke potrošačke navike, godišnje prihode itd.
    - Zavisnost prodaje jednih proizvoda od drugih

Data mining i KDD

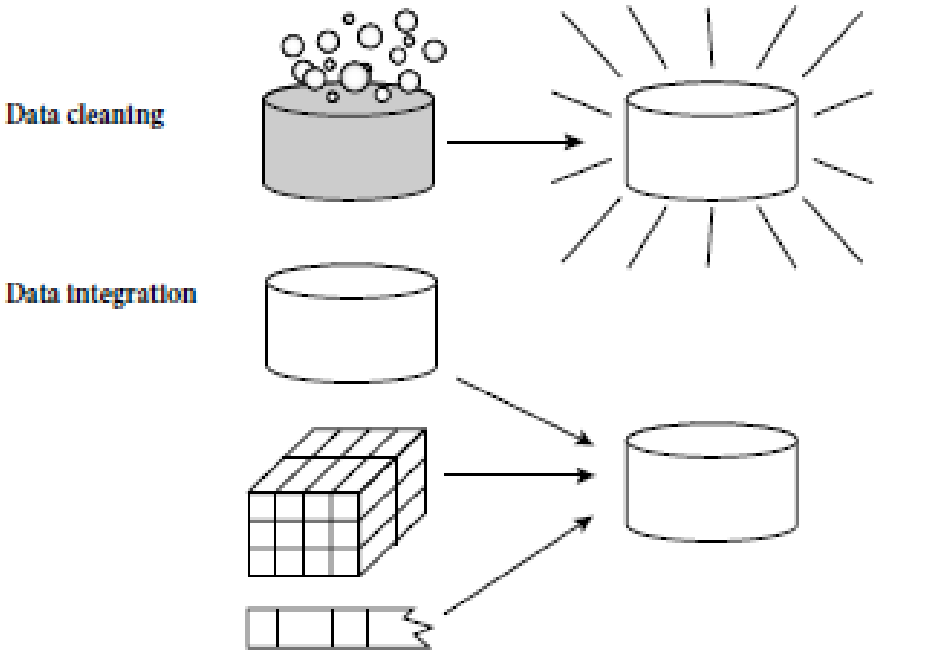
- Data mining je najvažniji u KDD



Knowledge discovery in databases

- Čišćenje podataka je eliminacija šuma, nekonzistentnosti u podacima itd.
- Integracija podataka je objedinjavanje podataka iz više izvora
- Transformacija je prevođenje podataka u formu pogodnu za primjenu data mining algoritama (npr. normalizacija, agregacija)
- Selekcija je generisanje podataka koji su potrebni za analizu

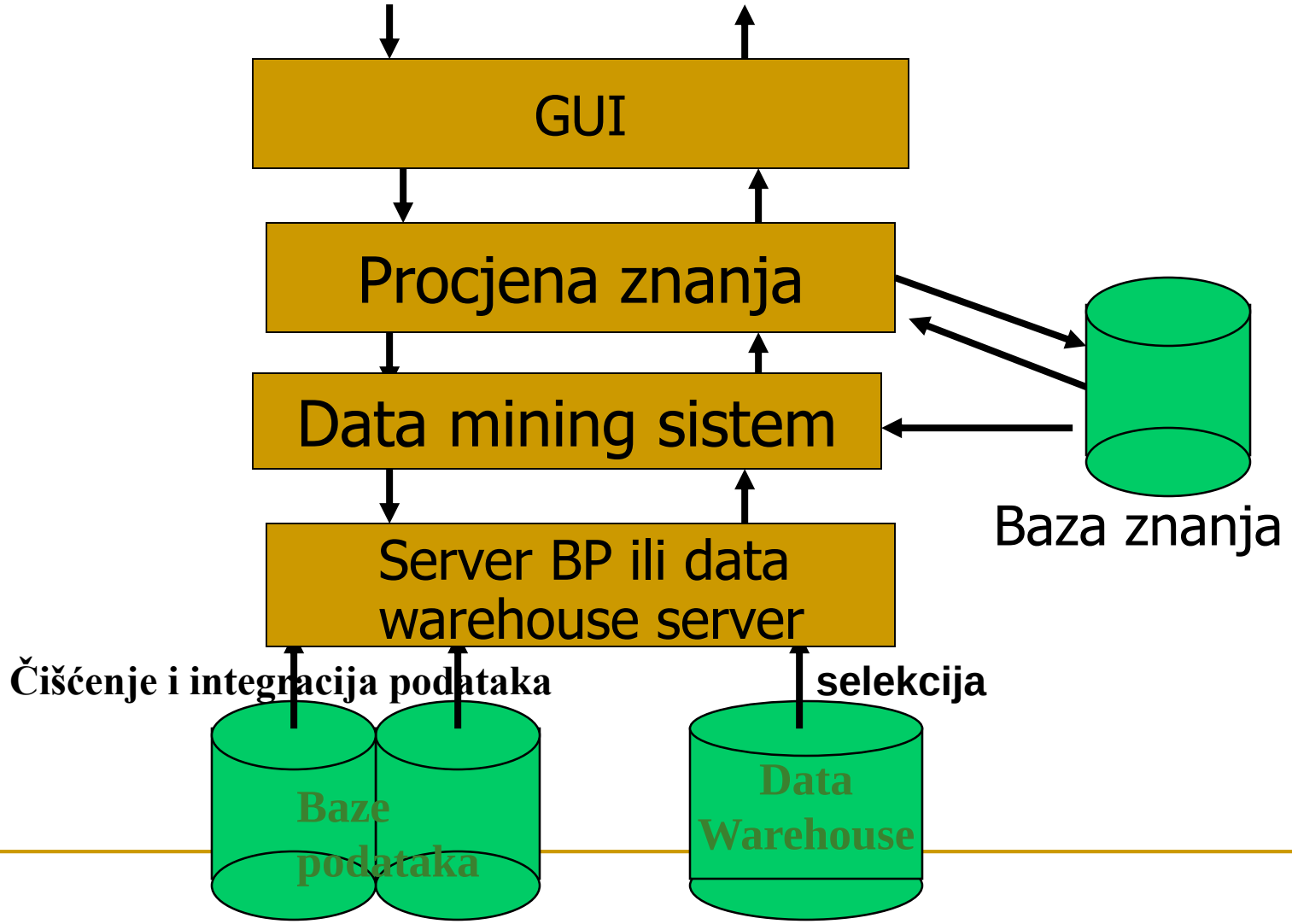
Preprocesiranje podataka



Knowledge discovery in databases (2)

- Data mining je izbor i primjena algoritama da bi se prepoznali sakriveni šabloni u podacima
- Procjena otkrivenih znanja na osnovu utvrđene mjere identifikuje korisne šablone
- Prezentacija znanja je grafičko predstavljanje otkrivenog znanja

Arhitektura data mining sistema



Data mining: koji tipovi podataka?

- Relacione baze podataka
- Data warehouse sistemi
- Transakcione baze podataka
- Napredne baze podataka
  - Objektno-orijentisane i objektno-relacione baze podataka
  - Prostorne baze podataka
  - Vremenske baze podataka
  - Multimedijalne baze podataka
  - WWW

Data mining tehnike (1)

- Opisivanje klasa/koncepata
  - Karakterizacija ciljne klase/koncepta
  - Poređenje ciljne klase sa suprotavljenom klasom ili klasama
- Asocijativna analiza
  - godine(X, “20..29”) && prihodi(X, “20..29K”) → kupuje(X, “PC”) [support = 2%, confidence = 60%]

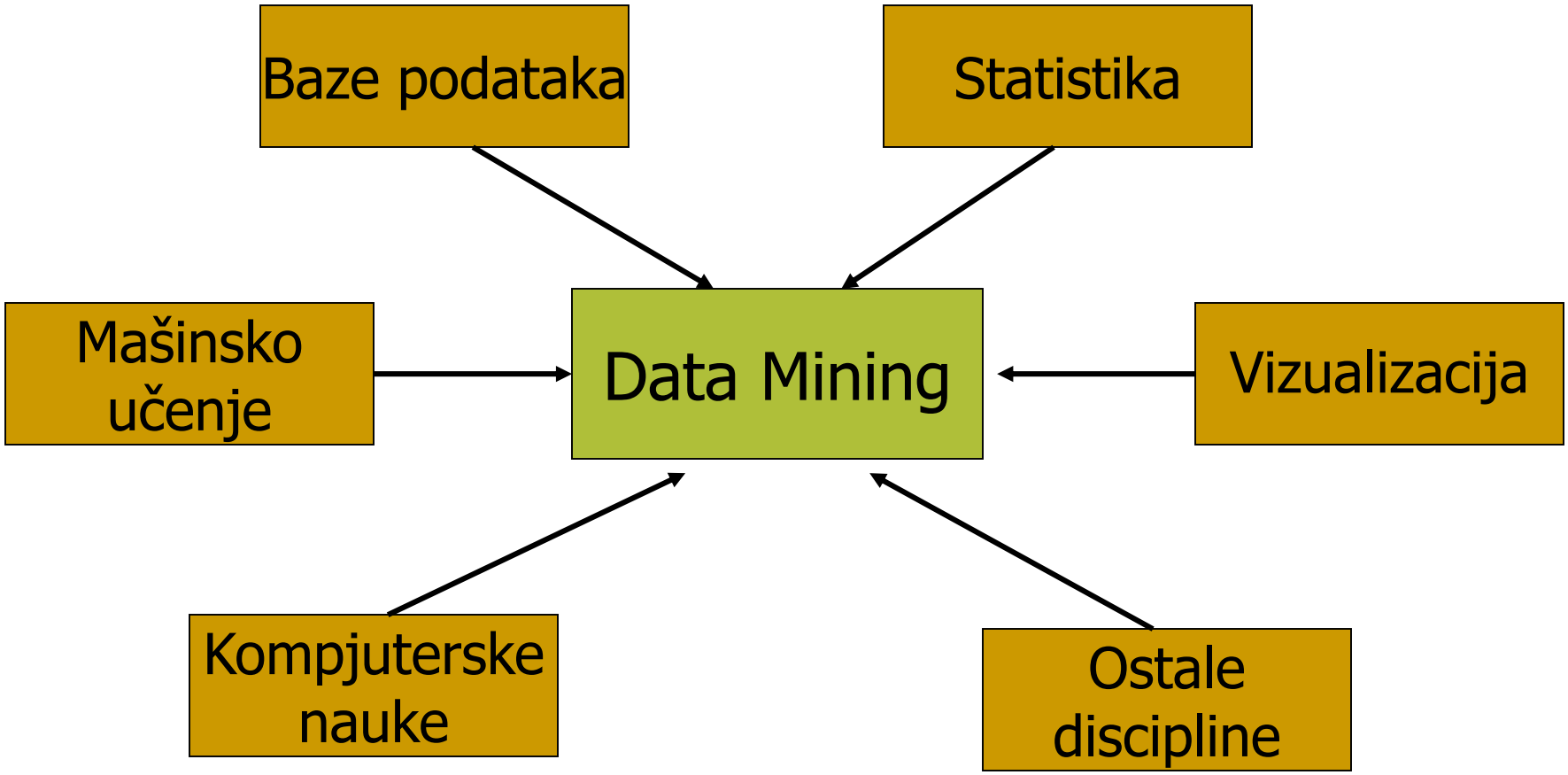
Data mining tehnike (2)

- Klasifikacija i regresija
  - Konstrukcija modela koji opisuju različite klase
  - Prezentacija modela: drveta odlučivanja, neuronske mreže itd.
  - Regresija: predviđa se brojna vrijednost
- Klasterizacija
  - Grupisanje podataka u klase
  - Princip: maksimizacija sličnosti unutar klastera i minimizacija sličnosti van klastera

Šta je korisno znanje?

- Pravilo je korisno ako je lako razumljivo, validno na novim ili testnim podacima sa određenom tačnošću, prethodno nepoznato i potencijalno upotrebljivo.
- Objektivne i subjektivne mjere korisnosti
- Kompletnost: generisati sva pravila
- Optimizacija: generisati samo korisna pravila

Data mining, interdisciplinarnost



Klasifikacija data mining sistema

- Na osnovu funkcionalnosti:
  - Deskriptivni
  - Prediktivni
- Na osnovu
  - Tipa baze podataka
  - Tipa znanja koji se otkriva
  - Upotrijebljenih tehnologija
  - Domena primjene