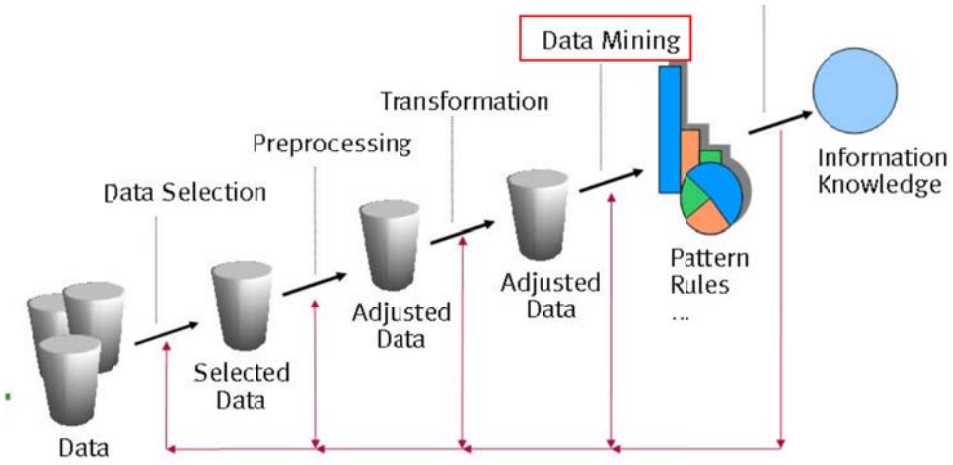
Rudarenje podataka - Data mining/pretraživanje podataka

Rudarenje (iskopavanje) podataka (eng. *Data mining: DM*) **je sortiranje, organiziranje ili grupisanje velikog broja podataka i izvlačenje relevantnih informacija.** Data mining je relativno novo polje u računarstvu . Pojednostavljeno **to je skup metoda i postupaka koje za cilj imaju otkrivanje zakonitosti** (ispitujući odnose, logičnost, pravilnost te uopšte bilo kakvu strukturu) **u masi podataka**. Rudarenje podrazumijeva organiziovanje (reorganizovanje) baza čišćenjem podataka kako bi se pristupilo znanju i sticanju istog na osnovu postojećih podataka u bazama.
DM see bavi procesiranjem i izdvajanjem šablona (*uzoraka*) u velikim kompletima (*setovima*) podataka kombinujući metode statistike, metode umjetne inteligencije i metode upravljanja bazama podataka
Sam termin mogli bismo objasniti kao proces pronalaženja korisnog znanja ili informacija, odnosno otkrivanje znanja iz velike količine podataka.

Pojam rudarenja često se poistovjećuje sa **2 različita procesa**: otkrivanje i predviđanje znanja.
Proces **otkrivanja** znanja implicira korisnikovo razumijevanje eksplicitnih informacija za koje je bitno da su u čitljivom obliku.
Predviđanje se odnosi na buduće događaje i u nekim pristupima može biti jasno predvidljivo, čitljivo i prozirno dok u drugim neprovidno.



Rudarenje podataka je prirodna evolucija tehnologije, a koja upotrebljava koncepte, metode i tehnike različitih disciplina kao što su baze podataka, statistika i umjetna inteligencija.
Ogromne **baze podataka bogate su podacima, ali i siromašne informacijama koje su skrivene u podacima**. Upravo je rudarenje podataka to koje pomaže otkriti važne informacije i znanje utkano u podatke, uvelike pridonoseći donošenju odluka, poslovanju i nauci.

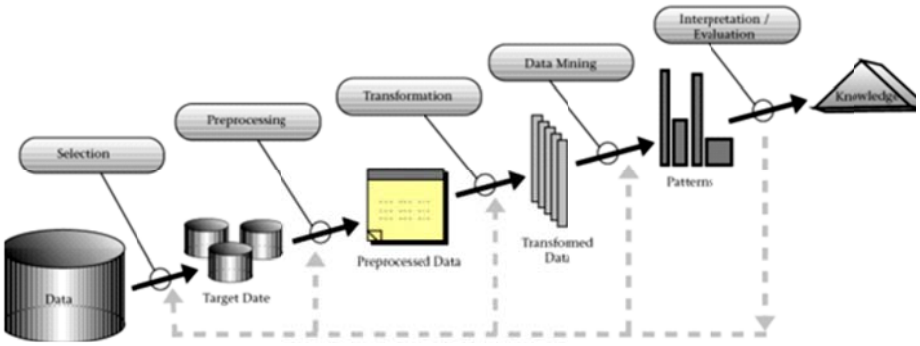
Što se samog naziva tiče, data mining, postoji još nekoliko naziva, *KDD* (Knowledge Development in a Database), *CRM* (Customer Relationship Management) ili *Database Intelligence*. Svi nazivi opisuju jednu stvar – korištenje svih mogućih alata kako bi se informacije dovele do optimuma i iskoristile na najbolji način.

Šta je BI i kako se KDP primjenjuje u BI?
Dvije su skraćenice termina koje često srećemo kod rudarenja podataka su KDP i BI.
Postupak **pronalaženja korisnog znanja** iz podataka u određenom području primjene naziva se proces sticanja znanja na osnovu podataka (Knowledge Discovery Processes–**KDP**).
Poslovna inteligencija (Business Intelligence- **BI**) je korištenje kolektivnog znanja organizacije sa ciljem postizanja konkurentske prednosti, odnosno to je proces prikupljanja raspoloživih internih i relevantnih eksternih podataka, te njihove konverzije u korisne informacije koje mogu pomoći poslovnim korisnicima pri donošenju odluka.
Rudarenje podrazumijeva organiziranje baza čišćenjem podataka kako bi se pristupilo znanju i stjecanju istog na temelju postojećih podataka u bazama.

Pojmovi DM (rudarenja) i KDP (proces otkrivanja znanja) mogu se smatrati sinonimima, jer i jedan i drugi podrazumjevaju obradu podatke iz baza podataka i njihovo pretvaranje u korisne informacije, koji se najčešće koriste za donošenje (pametnih, poslovnih) odluka (BI).

Faze pretraživanja podataka

Uobičajeno modeli procesa pretraživanja podataka opisuju tri faze:
- 1 pred-procesiranje,
- 2 pretraživanje (pretraga) podataka i
- 3 potvrda rezultata.



1. Pred-procesiranje

Pred-procesiranje je nužno da bi se mogli analizirati različiti setovi (kompleti) podataka prije same pretrage.
Prije nego što se algoritmi pretrage podataka mogu koristiti, mora se sastaviti set podataka koji će biti ciljani set. Budući da pretraživanje podataka može otkriti samo one uzorke, šablone koje već postoje, ciljani komplet podataka mora biti dovoljno velik da sadrži te šablone dok ostatak mora biti dovoljno sažet kako bi se mogao pronaći u odgovarajućem vremenu.
Zajednički izvor podataka se naziva *„skladište podataka"*: **Data warehouse**. Na kraju se brišu oni šabloni koji sadrže greške, koje ne dovode do traženog pojma ili dovode do podataka kojima nedostaju dijelovi.
Skladište podataka može biti bilo koje veličine i stepena kompleksnosti, ali glavno svojstvo kojim se odlikuje dobra kvaliteta skladišta jest brzina pristupa podacima (brzo ali i precizno izdvajanje i prikaz podataka) i mogućnost jednostavnog rukovanja podacima. Dobrim dizajnom skladišta podataka se povećava vrijednost baze podataka.

2. Pretraživanje podataka

Da bi podatke mogli pretraživati moramo definisati algoritam pretrage. **Da bi algoritam pretrage bio moguć same podatke treba urediti na odgovarajući način u proceu predprocesiranja.** Nakon toga potrebno je definisati konkretnu implementaciju algoritma i na osnovu nje realizovati pretragu. Samo pretraživanje podataka uključuje četiri vrste zadataka:

- **učenje pravila asocijacije**,
- **grupisanje**,
- **klasifikacija** i
- **regresija**.

Pojasnimo malo zadatke:
Učenje pravila asocijacije- potraga za vezom između varijabli. **Asocijacija** se bavi pitanjem koje se stvari dešavaju istovremeno.
Na primjer, supermarket može odrediti koji se proizvodi često kupuju zajedno te iskoristi tu informaciju za marketniške svrhe.Primjer jednostavne analiza kupovne: *uz čips ide pivo.*
Vjerovtno najćešće eksploatisani **primjer je *Pelene i pivo.*** Opštepoznata metoda u Data Miningu **potrošačke košarice** gdje se gleda koji se proizvodi često kupuju zajedno. Analizom podataka možete otkriti

da se često uz pjenu za brijanje kupuju i britvice što je očigledno. Međutim, mogu se otkritii neke **skrivene veze** poput primjera piva i pelena. Naime, prije otprilike 10 godina, Teradata (jedan od DM pionira) je vršeći analize podataka jednog svog klijenta utvrdila da se u večernjim satima često zajedno kupuju pivo i pelene.

Razmislite i metodm analize i asocijacije rješite i pronađite korist od ovih podataka.

Ili ipak evo rješenja (doduše malo sitnijim fontom)

Pelene kupuju očevi petkom poslijepodne, i uz njih obično kupe pivo za gledanje utakmice preko vikenda!
Korist od ovog otkrića je sa aspekta marketinga očigledna reklame za pelene i pivo idu zajedno. U praksi se navodi da je lanac supermarketa povećao prodaju i zaradu tako što je police sa pelenama primako vitrini sa pivom i petkompelene profavao bez popusta. (procjenjujući da da se očevi više razumiju u pivo, nego u cijene pelena).

Grupisanje- otkrivanje grupa i struktura u podacima koje su na neki način slične, **bez da se koriste već poznate strukture u podacima.**

Klasifikacija- uopštavanje **poznate strukture** kako bi se ona mogla primjeniti **na nove podatke**. Klasifikacija se bavi svrstavanjem objekata u neku od predefinisanih kategorija.
Metoda najbližeg suseda je tehnika koja se takođe koristi za klasifikaciju podataka. Za razliku od ostalih tehnika, ne postoji proces učenja kako bi se kreirao model. Podaci koji se koriste za učenje u stvari jesu model. **Kada se pojavi novi podatak, algoritam analizira sve podatke u bazi kako bi našao podgrupu slučajeva koji najbolje odgovaraju tom slučaju i na osnovu toga je u stanju da predvidi ishod**.
Primjer *klasifikacije je razvrstavanje potražioca kredita u nisko, srednje ili visoko rizičnu skupinu.Ono što će se desiti ispod haube je da će Data Mining algoritam proći kroz bazu bivših korisnika kredita te utvrditi koje to karakteristike imaju recimo, korisnici koji nisu uredno vraćali kredit. Pomoću tih karakteristika banka može tražitelja kredita svrstati u neku od kategorija, te tražiti većeili manje osiguranje povrata sredstava*.
Ili: *Na primjer, neki program elektronske pošte može pokušati klasificirati neku elektronsku poštu kao legitimnu ili kao bezvrijednu elektronsku poštu: spam.*
Kao posebna vrsta klasifikacije uzima klastering. **Klastering** je grupisanje podataka u klase. Princip: maksimizacija sličnosti unutar klastera i minimizacija sličnosti van klastera. **Klasteriranje** se takođe bavi svrstavanjem objekata u kategorije, samo ovdje te kategorije nisu unaprijed definisane, što problem čini većim. Primjer primjene te metode je razvrstavanje kupaca u kategorije prema kojima se onda mogu definisati različite marketinške strategije. Kupci su različiti, različitih ukusa, uvjerenja, stila kupovine i, što je najvažnije, različite profitabilnosti. Zato kupce treba i različito tretirati!

Regresija- pokušava se pronaći funkcija koja modelira podatke sa najmanjom geškom. Regresioni model je statistički model koji matematičkim formulama, uz određene pretpostavke najbolje opisuje kvantitativnu zavisnost između varijacija posmatranih pojava u realnosti.

3. Potvrda i provjera rezultata

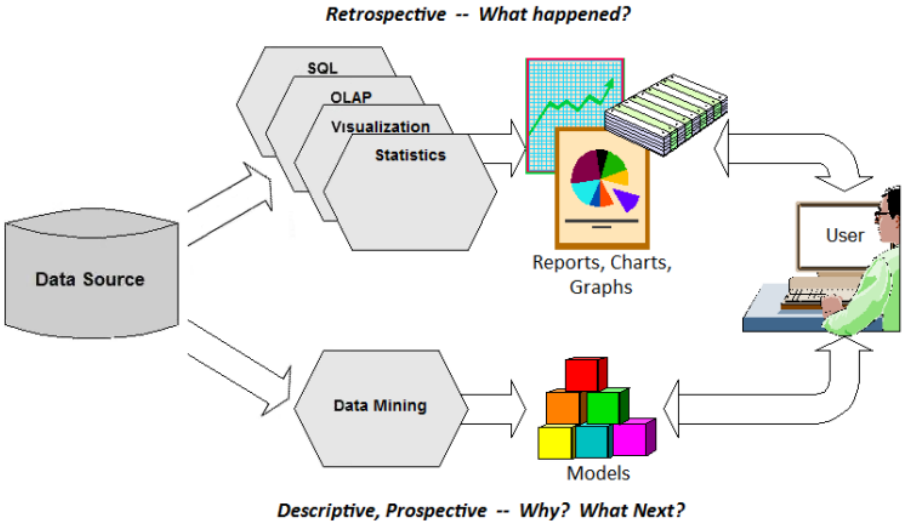
Potvrda rezultata- konačni korak uključuje provjeru i potvrdu uzoraka proizašlih iz algoritama pretrage podataka u većem setu podataka. Nisu svi uzorci nađeni algoritmima pretrage podataka obavezno i nužno dobri.
Naime, često algoritmi pretrage podataka pronađu uzorke prisutne u probnom setu podataka, koji nisu prisutni u opštem setu podataka. Kako bi se ovaj problem riješio, koristi se ***test kompleta (seta)*** podataka algoritmu nepoznatih od ranije pretrage podataka. Tako se naučeni uzorci primjenjuju u ovom testu a dobiveni rezultat se uspređuje sa željenim rezultatom.
Na primjer, algoritam pretrage podataka koji pokušava prepoznati spam od legitimne elektronske pošte će se testirati na probnom setu elektronske pošte. Naučeni uzorci će se primjeniti na testni set elektronske pošte, koji nije algoritmu od ranije poznat. Preciznost tih uzoraka se tada može vidjeti po broju točno klasificirane elektronske pošte. Ako naučeni uzorci ne zadovoljavaju željene standarde, tada je nužno napraviti ponovnu procjenu i promjeniti pred-proces te pretragu podataka.
Ukoliko naučeni uzorci zadovoljavaju željene standarde, tada je zadnji korak interpretacija naučenih uzoraka i njihovo pretvaranje u znanje.

4. Čišćenje podataka

Kao **četvrtu fazu** može se izdvojiti ažuriranje podatakae koje se najčešće svodi na prečišćavanje i provjeru ispravnosti podataka. Uopšteno ona pripada trećoj fazi.
Kvaliteta baze podataka ima direktan utjecaj na pretraživanje podataka iz baze.
Većinom se čak i nakon **čišćenja baze** podataka ona sastoji od 20%-50% zastarjelih podataka, podataka s greškama, neupotpunjenih podataka, duplikata ili iz drugih razloga nekorisnih podataka. Da bi baza podataka bila najbolja što može biti, može se koristiti stohastički model sudjelovanja, koji rangira svaki pojedini podatak prema stvarnoj održivosti podatka, te briše ili ažurira podatke zavisno o njihovoj održivosti.

5. Prikaz rezultata

Po završenoj analizi informacija, svi rezultati se prikazuju na pregledan način, najčešće u vidu tabela ili dijagrama koji mogu biti dvodimenzionalni ili trodimenzionalni.
Programi čak omogućavaju korisniku da mijenja neku od promenljivih, a da se efekat njene promene prikazuje u realnom vremenu na dijagramu. Na taj način korisniku su mnogo jasnije uzročno-posljedične veze između elemenata sistema i izuzetno je jednostavno isprobati različite mogućnosti, to jest dati odgovor na čuveno pitanje **ŠBBKBB** („šta bi bilo kad bi bilo?"- what happened).



Tehnike DM

Osnovne tehnike za rudarenje podataka su:

- statističke metode,
- genetički algoritmi,
- neuronske mreže,
- stabla -grafovi odlučivanja,
- umjetna inteligencija,
- asocijacijska pravila, itd.

Alitička statistika je jezgra svih ovih metoda za otkrivanje znanja. Statistika je osnov tzv. **ako-onda pravila**.
Iz statističke perspektive, rudarenje podataka se može opisati kao računarski automatizirana istraživačka analiza podataka iz (obično) velikih i složenih baza podataka.

Ciljevi DM

Dva osnovna cilja DM projekata se mogu svrstati u dvije kategorije:

- Predviđanje
- Deskripcija

Predviđanjem se pokušava iz postojećih podataka prognozirati buduće vrijednosti varijabli (npr. prodaje), dok se **deskripcijom** nastoje pronaći uzorci u podacima čijim se interpretiranjem može objasniti ponašanje čitavog sistema.

Primjeri rudarenja podataka

U poslovanju Data mining se najviše koristi na području marketinga, koji je usmjeren sve više pojedinačnom kupcu –**upravljanje odnosima sa kupcima (Customer Relationship Management-CRM**) koje je usmjeren ostvaranju, održavanju ili poboljšavanju odnosa sa kupcima. Cilj ove pojedinačne usmjerenosti na kupca jeste pridobijanje novih kupaca i zadržavanje starih kupaca.
CRM pokušava uvidjeti želje i potrebe kupaca, razumetinjihovo ponašanje i predvideti buduće ponašanje.
Tipični primeri upotrebe Data Minig-a su još:

- **Bankarstvo** –Predviđanje nivoa loših plasmana, utvrđivanje rizika kod kreditnih kartica, predviđanja zarade od novih klijenata;
- **Osiguranje** –Predviđanje nivoa odštetnih zahteva, sprečavanje prevara;
- **Trgovina** –projekcije prodaje, sprječavanje krađa i prevara, utvrđivanje plana snabdevanja maloprodaja, određivanje optimalnih zaliha;
- **Policija** –Praćenje šema zločina, predviđanje kriminalnog ponašanja pojedinaca, lociranje zločinaca;
- **medicinske ustanove** – za predviđanje uspješnosti operacija, medicinskih testova, ili lijekova

Rudarenje podataka se koristi još i u:

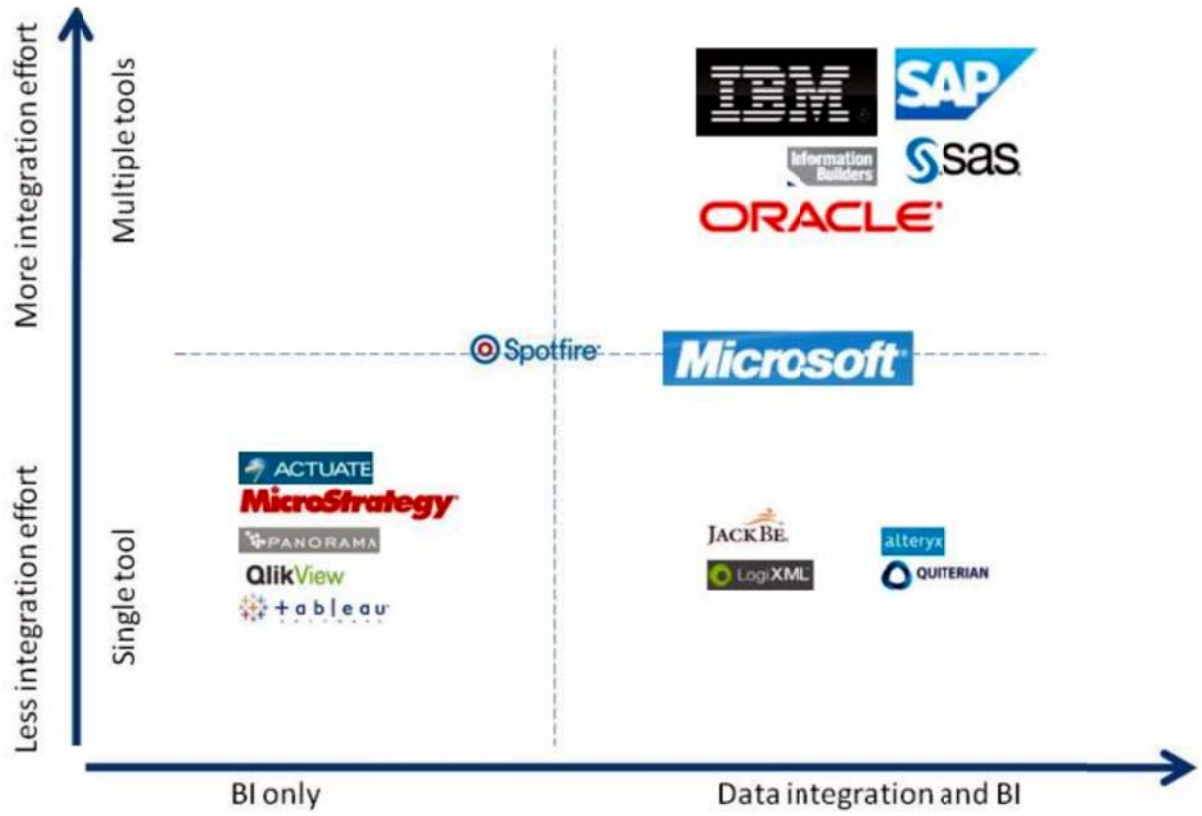
- **Politici**: rudarenje je metoda kojom je U.S. Army uspjela identifikovati vođu napada na Twin Towers, 11.9.2001.; a tom se metodom također koriste CIA i Canadian Security Intelligence Service
- **Igrama**: već od 60-ih godina u nekim kombinatornim igrama poput šaha
- **Poslovanju uopšte**: pomaže u bržem donošenju poslovnih odluka zbog kontaktiranja samo onih klijenata za koje postoji visoka vjerojatnost da će odgovoriti

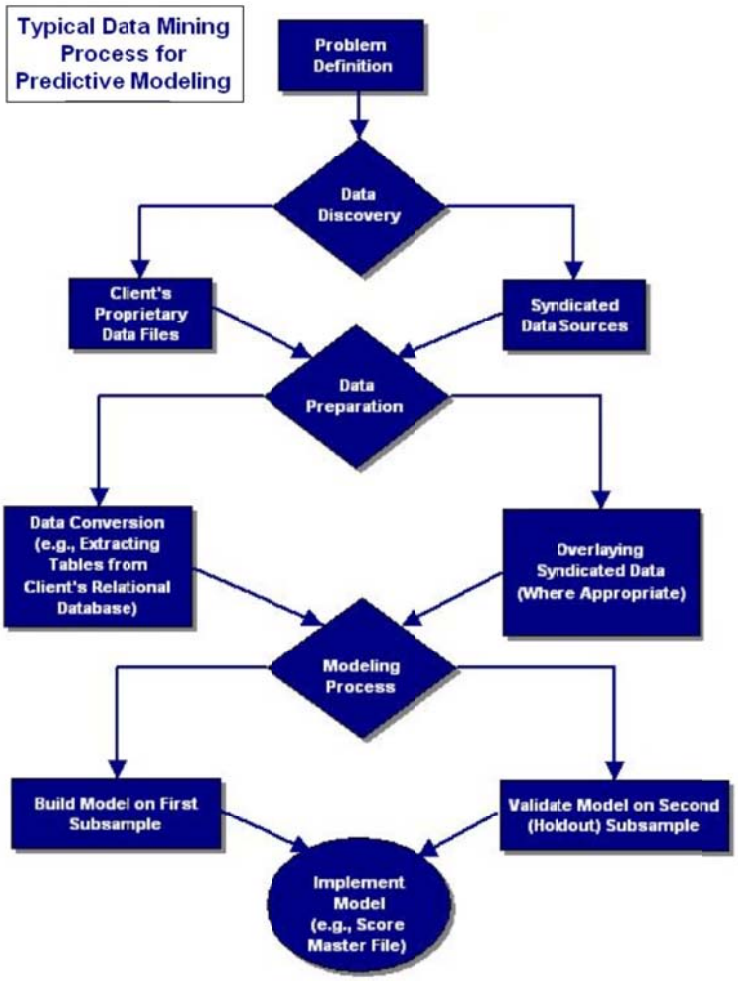
Alati za DM

Softver za *data mining* je pristupačan kako za velike mejnfrejm sisteme tako i za samostalne PC platforme. Cijena sistema varira od nekoliko hiljada dolara pa do nekoliko miliona dolara za najveće sisteme.
Dva osnovna uslova za izbor odgovarajuće platforme jesu veličina baze podataka i kompleksnost upita. Velika baza podataka sa sobom povlači veliki broj podataka koji treba skladištiti i održavati i samim tim zahteva moćniji sistem. Kompleksnost upita i njihov veliki broj takođe povećavaju potrebu za procesorskom moći. Ubrzavanje rešavanja upita može se postići indeksiranjem podataka. Takođe, paralelno procesiranje značajno ubrzava rad s velikim bazama podataka.
Daćemo neke primjere alata gdje se koristi DM:

- statistički softverski paketi (npr. SAS, Statistika)
- matematički softverski paketi (npr. MathLab, Matematica)
- alati uključeni u skladištenje podataka (OLAP[1]) ili sistem za upravljanje bazom podataka (npr. Microsoft SQL Server Business Intelligence, ali i besplatni **MySQL, pa čak i Exel**)
- specijalizirani alati za općenite ili poslovne primjene (npr. DataMiner, IntelliMiner, i sl.)

[1] On-Line Analytical Processing, se uobičajeno prevodi kao „online analitička obrada“ , skraćenica OLAP podrazumijeva kategoriju aplikacija i tehnologije namijenjenu za skupljanje, upravljanje, obradu i prezentaciju multidimenzijskih podataka namijenjenih analizama za potrebe upravljanja.


More integration effort
Multiple tools
Single tool
Less integration effort
BI only
Data integration and BI
IBM
SAP
Information Builders
SAS
ORACLE
Spotfire
Microsoft
ACTUATE
MicroStrategy
PANORAMA
QlikView
tableau
JACKBE
LogiXML
alteryx
QUITERIAN


Typical Data Mining Process for Predictive Modeling
Problem Definition
Data Discovery
Client's Proprietary Data Files
Syndicated Data Sources
Data Preparation
Data Conversion (e.g., Extracting Tables from Client's Relational Database)
Overlaying Syndicated Data (Where Appropriate)
Modeling Process
Build Model on First Subsample
Validate Model on Second (Holdout) Subsample
Implement Model (e.g., Score Master File)