1 Ce este Data Mining?

Iniţial "data mining" (extragerea de cunostinţe din date) a fost un termen din statistică însemnând suprautilizarea datelor pentru a deduce inferenţe invalide.

- Teorema lui Bonferroni ne avertizează că atunci când există prea multe concluzii posibile unele dintre acestea vor fi adevărate din motive pur statistice, fără o valabilitate fizică.
- Exemplu faimos: David Rhine, un "parapsiholog" de la Duke a testat în anii 1950 studenţii pentru "percepţie extrasenzorială" (PE) cerându-le să ghicească 10 cărţi de joc - roşii sau negre. A descoperit că 1/1000 din ei au ghicit toate cele 10 cărţi şi în loc să realizeze că este ceea ce trebuia aşteptat din ghicirea aleatoare i-a declarat ca având PE. Când i-a retestat a descoperit că nu sunt mai buni decât media. Concluzia sa: a spune oamenilor că au PE duce la pierderea acesteia!

Definiţia noastră: "descoperirea unor informaţii sumare utile despre date".

1.1 Aplicaţii

Câteva exemple de "succes":

1. Arbori de decizie construiţi din istoria împrumuturilor bancare pentru a produce algoritmi de decizie în vederea acordării împrumuturilor.
2. Şabloane privind comportamentul călătorilor folosite pentru a gestiona vânzarea locurilor cu reducere la cursele aeriene, a camerelor de hotel, etc.
3. "Scutece şi bere": observaţia că cei care cumpără scutece cumpără bere mai mult decât media a permis supermagazinelor să plaseze berea şi scutecele aproape unele de altele, ştiind că mulţi cumpărători vor circula între ele. Plasarea cipsurilor de cartofi între cele două a crescut vânzările la cele trei articole.
4. Skycat si Sloan Sky Survey: gruparea obiectelor de pe cer după nivelul lor de radiaţie în diferite benzi a permis astronomilor să delimiteze galaxii, stele apropiate şi alte feluri de obiecte cereşti.
5. Compararea genotipului persoanelor îndeplinind sau nu o anumită condiţie a permis descoperirea unei mulţimi de gene care împreună determină multe cazuri de diabet. Acest mod de extragere a cunostinţelor din date va deveni mult mai important în momentul construirii genomului uman.

1.1 Comunităţi implicate în data mining

Când data mining a fost recunoscută ca o unealtă puternică diferite comunitaţi au clamat prioritate asupra subiectului:

1. Statistica
2. Inteligenţa artificială (IA) unde este denumită "machine learning".
3. Cercetătorii din domeniul algoritmilor de grupare ("clustering algorithms")
4. Cercetătorii din domeniul vizualizării datelor ("data visualization")
5. Baze de date. Vom continua bineînţeles cu această abordare, concentrându-ne asupra provocărilor care apar atunci când cantitatea de date este mare iar calculele sunt complexe. Într-un anumit sens data mining poate fi văzută ca mulţimea algoritmilor pentru execuţia unor cereri foarte complexe asupra unor date care nu sunt în memoria centrală a calculatorului.

1.2. Etape în procesul de extragere a cunoştintelor din date

1. *Colectarea datelor*, e.g. din depozitele de date ("data warehouse"), parcurgerea webului.
2. *Curăţarea datelor*: eliminarea erorilor şi/sau a datelor incorecte, e.g. temperatura pacientului=125
3. *Extragerea proprietăţilor*: obţinerea doar a atributelor de interes ale datelor, e.g. "data achiziţiei" nu este probabil de interes în gruparea obiectelor cereşti în cazul Skycat.
4. *Extragerea şablonelor şi descoperirea*: Aceasta este etapa considerată adesea ca fiind "data mining" şi aici ne vom concentra eforturile.
5. Vizualizarea datelor.
6. Evaluarea rezultatelor; nu orice fapt descoperit este şi util ori adevărat!. O judecata este necesară înainte de a urma concluziile programelor de extragere de cunoştinţe din date.