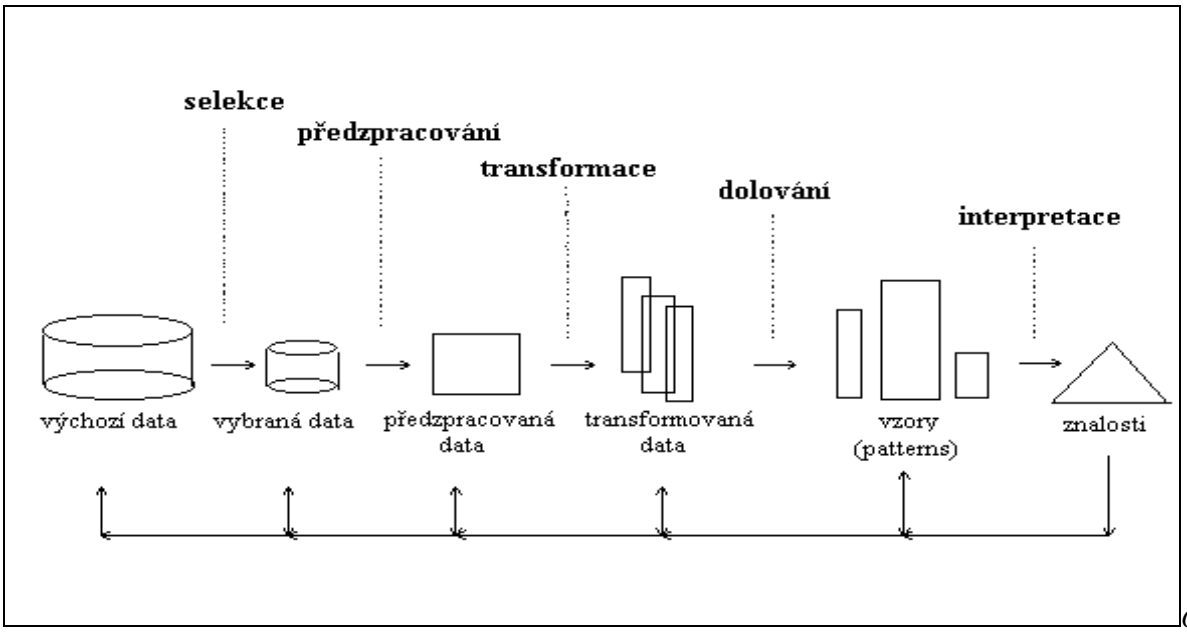
# 1. Dobývání znalostí z databází

O dobývání znalostí z databází (Knowledge Discovery in Databases, KDD) se začíná ve vědeckých kruzích mluvit počátkem 90. let. První impuls přišel z Ameriky, kde se na konferencích věnovaných umělé inteligenci (mezinárodní konference o umělé inteligenci IJCAI'89 nebo konference americké asociace umělé inteligence AAAI'91 a AAA'93) pořádaly první workshopy věnované této problematice. Nebyla to ale jen umělá inteligence (přesněji řečeno metody strojového učení), které stály u zrodu dobývání znalostí z databází. Databázové technologie představují osvědčený prostředek jak uchovávat rozsáhlá data a vyhledávat v nich informace, statistika představuje osvědčený prostředek, jak modelovat a analyzovat závislosti v datech. Po léta se tyto discipliny vyvíjely nezávisle, až přišla ta chvíle, kdy rozsah automaticky sbíraných dat začínal uživatelům přerůstat přes hlavu, a současně s tím vznikla potřeba tato data používat pro podporu (strategického) rozhodování ve firmách. Zájem finančně silných uživatelů o aplikace pak stimuloval ono propojení a dal vznik (a hlavně popularitu) dobývání znalostí z databází. Neustálý nárůst zájmu odborné komunity dokládá množství konferencí (americké konference KDD, asijské konference PAKDD, evropské konference PKDD), vznik odborných skupin (např. special interest group for KDD - SIGKDD při americké asociaci ACM) i vznik samostatných odborných časopisů (časopis Data Mining and Knowledge Discovery vydávaný nakladetelstvím Kluwer). Tématika dobývání znalostí si postupně našla cestu i do šířeji zaměřených počítačových časopisů. Dnes již není nic neobvyklého, že na pojmy knowledge discovery, data mining, nebo business intelligence[1] narazíme i v reklamách počítačových firem.

Dobývání znalostí z databází (KDD) lze definovat *jako netriviální extrakci implicitních, dříve neznámých a potenciálně užitečných informací z dat* [Fayyad a kol, 1996]. Zpočátku se pro tuto oblast razily nejrůznější názvy: information harvesting, data archeology, data destilery. Nakonec ale převládla hornická metafora; dobývání znalostí a dolování z dat (data mining). Po jistém období tápání se ustálilo i chápání KDD jako interaktivního a iterativního *procesu* tvořeného kroky selekce, předzpracování, transformace, vlastního „dolování“ (data mining) a interpretace (*OObr. 1*).



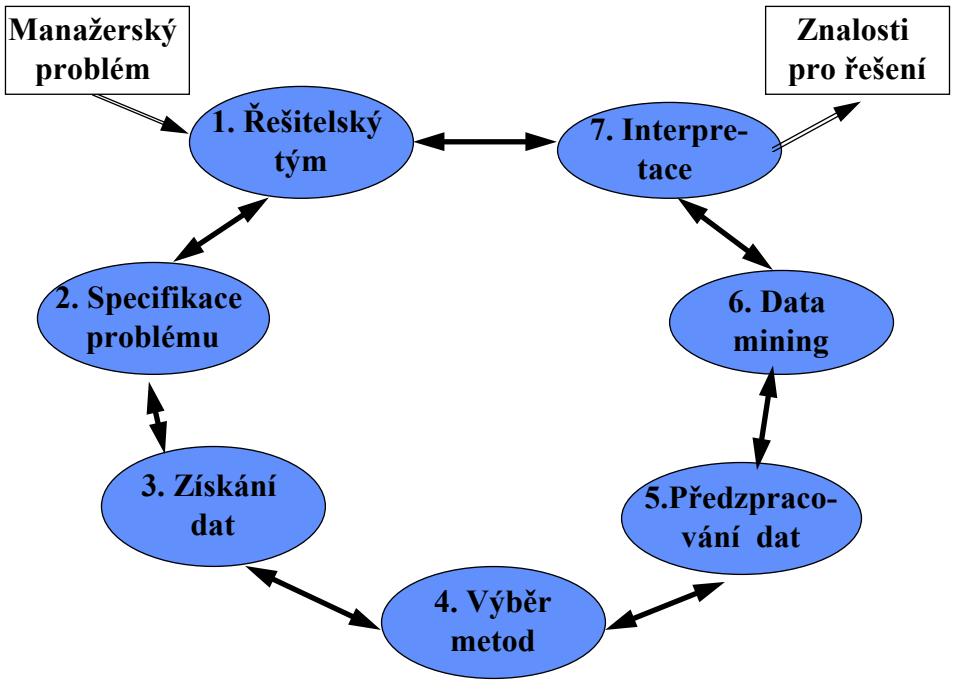

*O*

*Obr. 1 Proces dobývání znalostí z databází dle [Fayyad a kol, 1996].*

Na rozdíl od "prostého" použití statistických metod a metod strojového učení se v procesu dobývání

[1]Význam pojmu business intelligence je možno (s trochou nadsázky) interpretovat touto rovnicí: business intelligence = artificial intelligence + business

znalostí již klade důraz i na přípravu dat pro analýzu a na interpretaci výsledných znalostí. Při přípravě dat se obvykle z dat uložených ve složité struktuře např. datového skladu vytváří jedna tabulka, obsahující relevantní údaje (hodnoty atributů) o sledovaných objektech (např. klientech banky nebo zákaznících obchodního domu). Při interpretaci se nalezené znalosti[2] hodnotí z pohledu koncového uživatele.




*Obr. 2 Manažerský pohled na proces dobývání znalostí z databází*

Zatímco schéma na OObr. 1 popisuje „technologický“ pohled na dobývání znalostí, Anand [Anand a kol., 1996] nabízí pohled manažerský (Obr. 2). Impulsem pro zahájení procesu dobývání znalostí je nějaký **reálný problém**. Cílem procesu dobývání znalostí je získání co nejvíce relevantních informací vhodných k řešení daného problému. Příkladem reálného problému je otázka nalezení skupin zákazníků obchodního domu nebo skupin klientů banky kterým by bylo možno nabídnout specielní služby. U zákazníků obchodního domu se může jednat o zjištění, že zákazník kupuje potravinářské zboží odpovídající jisté dietě, v případě klientů banky může jít o potenciální zájemce o hypoteční úvěr. Nalezené skupiny jsou interpretovány jako takzvané segmenty trhu v dané oblasti.

Prvním krokem při řešení problému je **vytvoření řešitelského týmu**. Jeho členy musí být *expert na řešenou problematiku*, *expert na data* - jak v organizaci tak případně i na externí data a *expert na metody KDD*. V případě rozsáhlejších problémů je obvyklé, že jednotliví experti mají k dispozici vlastní tým nebo alespoň využívají konzultací s dalšími experty

Prvním úkolem sestaveného týmu je **specifikace problému**, který je třeba řešit v souvislosti, z pohledu dobývání znalostí. U zákazníků obchodního domu nakupujících potravinářské zboží odpovídající jisté dietě je mimo jiné třeba specifikovat položky zboží odpovídající různým dietám. U skupin zákazníků nakupujících položku A a nenakupujících položku B je krom jiného třeba vytipovat vhodné skupiny položek, atd.

[2] Fayyad rozlišuje mezi znalostmi získanými jako výstup z kroku dolování (nazývá je vzory – patterns) a mezi znalostmi interpretovanými uživatelem. My toto rozlišení nebudeme provádět.

Po specifikaci problému je třeba **získat všechna dostupná data**, která mohou být použita pro řešení problému. Znamená to posoudit všechna dostupná data a zvážit, zda jsou relevantní k danému problému. Tento proces může vyvolat menší či větší přeformulováni problému. V některých případech je třeba pracovat i s daty, která jsou archivována po delší dobu ve formě datových souborů a ne v databázi, data jsou někdy dokonce uložena v několika různých systémech. Náročnost získání dat je nepřímo úměrná úrovni datové základny která je k dispozici.

V mnohých případech je vhodné uvažovat i *externí data* popisující prostředí ve kterém se analyzované děje odehrávají. V případě klientů banky i zákazníků obchodního domu je důležitou informací kalendářní období (např. vánoce, velikonoce, období dovolených letních a zimních, den kdy zákazníci dostávají výplatu, pondělí, úterý, ..., ). Na zákazníky bude mít jistě vliv i počasí, reklama probíhající ve sdělovacích prostředcích, v některých případech i politické události.

Cílem **výběru metody** je zvolit vhodné metody analýzy dat. V rámci dobývání znalostí z databází je používána řada typů metod analýzy dat, ve většině případů je k řešení konkrétní úlohy zapotřebí kombinovat více různých metod. Mezi používané typy metod patří např. klasifikační metody, různé klasické metody explorační analýzy dat, metody pro získávání asociačních pravidel, rozhodovací stromy, genetické algoritmy, Bayesovské sítě, neuronové sítě, hrubé množiny (rough sets), velmi používané jsou i metody vizualizace. Dá se také předpokládat vývoj dalších metod.

V rámci **předzpracování dat** se data získaná k řešení specifikovaného problému připravují data do formy vyžadované pro aplikaci vybraných metod. V řadě případů se může jednat o značně náročné výpočetní operace. Do této fáze se zahrnuje patří i odstranění odlehlých hodnot, případně doplnění chybějících hodnot.

Krok **data mining** zahrnuje aplikaci vybraných analytických metod pro vyhledávání zajímavých vztahů v datech. Obvykle jsou jednotlivé metody aplikovány vícekrát, hodnoty vstupních parametrů jednotlivých běhů závisí na výsledcích předchozích běhů. Zpravidla se nejedná o aplikace metod jenom jednoho typu, jednotlivé typy se kombinují na základě dílčích výsledků.

Cílem **interpretace** je nezbytné zpracování obvykle značného množství výsledků jednotlivých metod. Některé z těchto výsledků vyjadřují skutečnosti, které jsou z hlediska uživatele nezajímavé nebo samozřejmé. Některé výsledky je možno použít přímo, jiné je nutno vyjádřit způsobem srozumitelným pro uživatele. Jednotlivé výsledky je často vhodné uspořádat do analytické zprávy. Analytická zpráva však není jediným možným výstupem procesu dobývání znalostí. Výstupem může být i provedení vhodné akce jako například zapnutí monitorovacího programu.

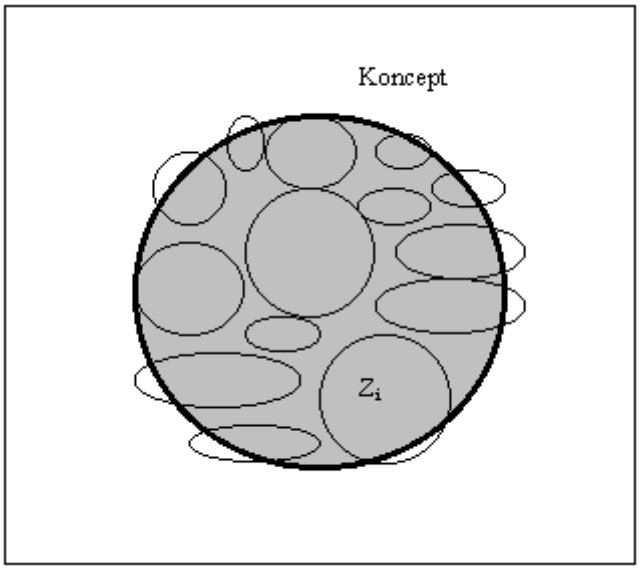
## 1.1 Úlohy

V případě dobývání znalostí z databází můžeme mluvit o různých typech úloh. Jsou to především [Klosgen, Zytkow, 1997] [3]:

- klasifikace/predikce,
- deskripce,
- hledání „nugetů“.

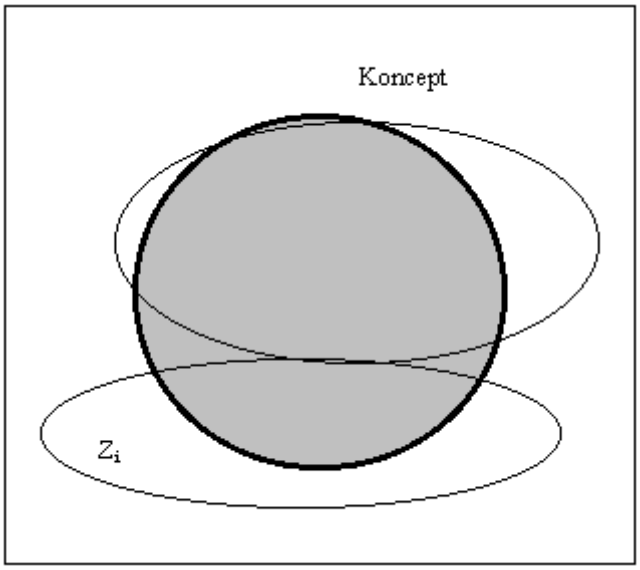
[3] Podrobnnější členění lze nalézt v [Chapman a kol, 2000]. Tvůrci metodiky CRISP-DM zde uvádějí úlohy deskripce dat a sumarizace, segmentace, deskripce konceptů, klasifikace, predikce a analýzy závislostí.

Při klasifikaci/predikci je cílem nalézt znalosti použitelné pro klasifikaci nových případů - zde požadujeme, aby získané znalosti co nejlépe odpovídaly danému konceptu; dáváme přednost přesnosti pokrytí na úkor jednoduchosti (připouštíme větší množství méně srozumitelných dílčích znalostí tak jak je to naznačeno na Obr. 3). Rozdíl mezi klasifikací a predikcí spočívá v tom, že u predikce hraje důležitou roli čas; ze starších hodnot nějaké veličiny se pokoušíme odhadnout její vývoj v budoucností (např. předpověď počasí nebo pohybu cen akcií).

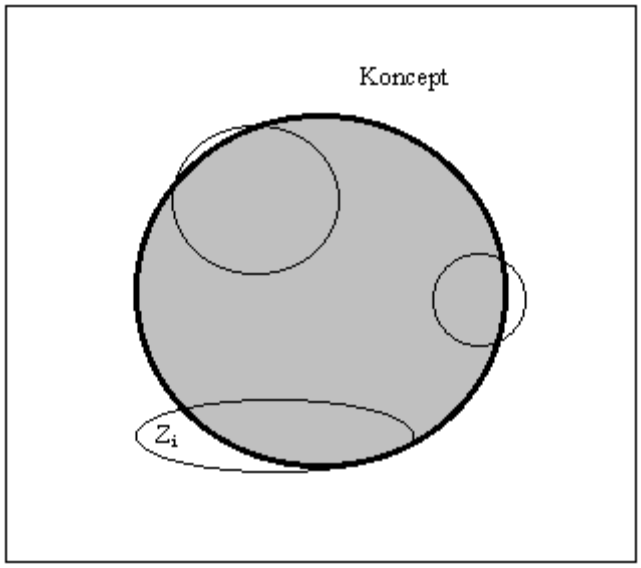



*Obr. 3 Klasifikace/Predikce*

Při deskripci (popisu) je cílem nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech. Požadujeme srozumitelné znalosti pokrývající daný koncept; dáváme tedy přednost menšímu množství méně přesných znalostí (viz. Obr. 4). Hledáme-li nugety, požadujeme zajímavé (nové, překvapivé) znalosti, které nemusí plně pokrývat daný koncept (Obr. 5).




*Obr. 4 Popis (deskripce)*




*Obr. 5 Nugety*

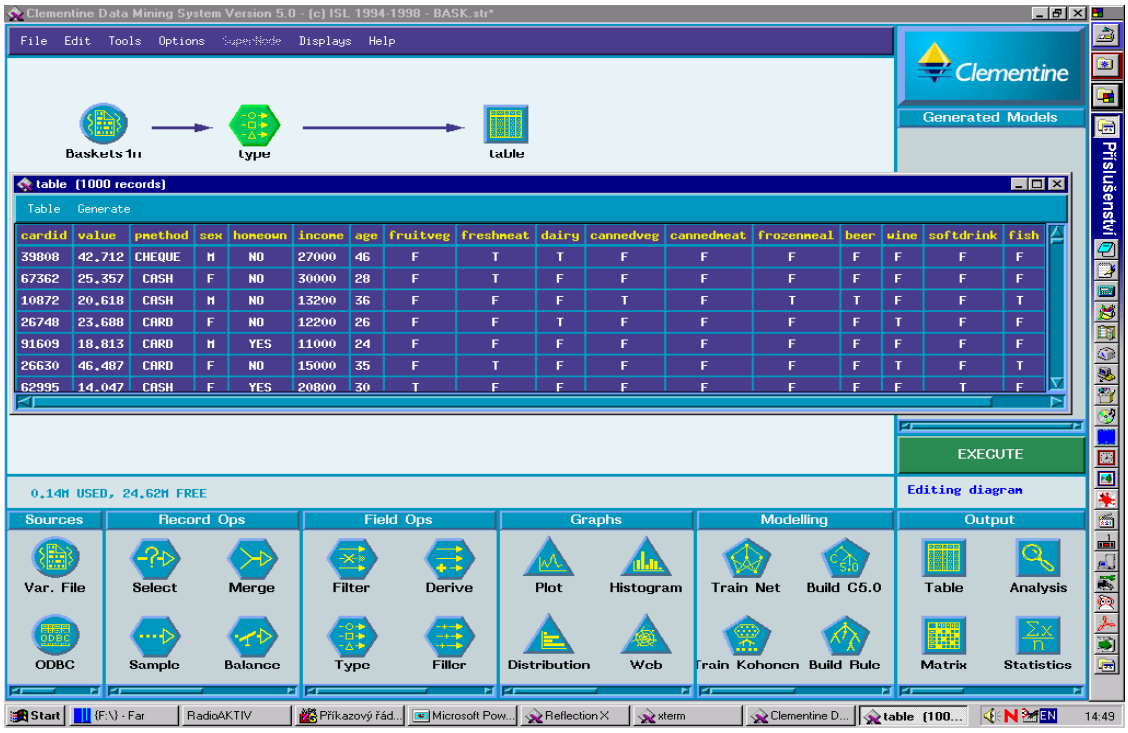
Úlohy dobývání znalostí lze nalézt v celé řadě aplikačních oblastí:

- Segmentace a klasifikace klientů banky (např. rozpoznání problémových nebo naopak vysoce bonitních klientů),
- Predikce vývoje kursů akcií,
- Predikce spotřeby elektrické energie,
- Analýza příčin poruch v telekomunikačních sítích,
- Analýza důvodů změny poskytovatele nějakých služeb (internet, mobilní telefony),
- Segmentace a klasifikace klientů pojišťovny,
- Určení příčin poruch automobilů,
- Rozbor databáze pacientů v nemocnici,
- Analýza nákupního košíku (Market Basket Analysis).

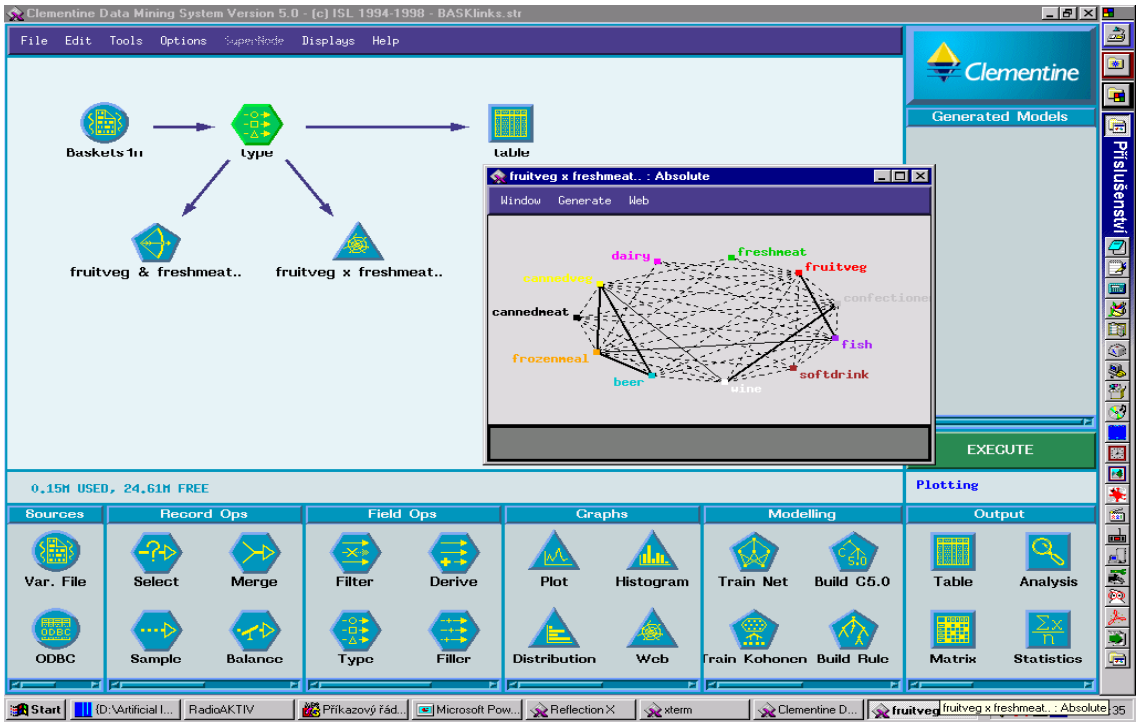
Podrobněji se zde podívejme na poslední z nich. Při analýze nákupního košíku se vychází z dat, která shromažďují různé řetězce supermarketů (u nás např. Delvita nebo Meinl). Data (alespoň podle následujícího příkladu[4] tvoří jednak charakteristiky zákazníků (pohlaví, vlastnictví domu, příjem a věk), jednak údaje o jednotlivých nákupech (způsob placení, částka, zakoupený typ zboží). Data jsou již předzpracována do podoby relační tabulky. Co záznam, to jeden zákazník, typy zboží jsou pevně dány – uvádí se, zda byl konkrétní výrobek zakoupen nebo ne (Obr. 6). V takovýchto datech můžeme například hledat souvislosti mezi jednotlivými typy zboží; bude nás zajímat, zda existují skupiny produktů, které si zákazníci kupují současně (např. pivo a párek). Obr. 7 ukazuje, že se velmi často objevuje v nákupním košíku současně např. pivo, zmrazené maso a konzervovaná zelenina, nebo ryby, ovoce a zelenina[5].

[4] Příklad je převzat z ukázkového příkladu v systému *Clementine*. O tomto systému podrobněji v příslušné kapitole.

[5] Systém *Clementine* zde použil grafický způsob prezentace těchto asociací; hrany v grafu odpovídají vazbám mezi produkty.

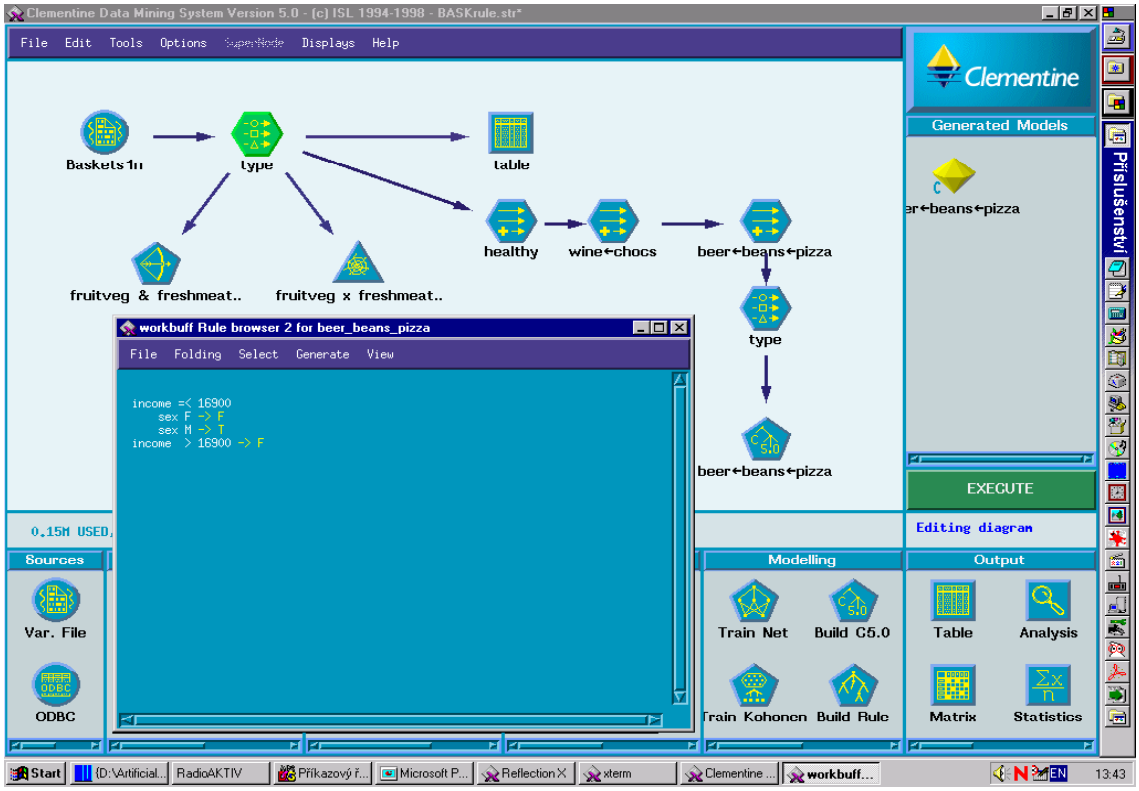



*Obr. 6 Vstupní data pro analýzu nákupního košíku*




*Obr. 7 Asociace mezi typy zboží*

Samozřejmě, že „zdravou“ a „nezdravou“ výživu nekupují titíž zákazníci. Může nás tedy zajímat, čím se tyto skupiny zákazníků vyznačují. Takové znalosti je možno získat např. pomocí rozhodovacích stromů. Obr. 8 ukazuje, že pizzu, pivo a fazole nakupují muži s nižším příjmem.




*Obr. 8 Konzumenti nezdravé výživy*

## 1.2 Metodiky

S postupem doby začaly vznikat metodiky, které si kladou za cíl poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí. Tyto metodiky umožňují sdílet a přenášet zkušenosti z úspěšných projektů Za některými metodikami stojí producenti programových systémů (metodika „*5A*“ firmy SPSS nebo metodika *SEMMA* firmy SAS), jiné vznikají ve spolupráci výzkumných a komerčních institucí jako „softwarově nezávislé“ (*CRISP-DM*).

### 1.2.1 Metodika „5A“

Metodika „*5A*“ nabízí firma SPSS jako svůj pohled na proces dobývání znalostí. Název metodiky je akronymem pro jednotlivé prováděné kroky [5A, 2000]:

- *Assess* – posouzení potřeb projektu,
- *Access* – shromáždění potřebných dat,
- *Analyze* – provedení analýz,
- *Akt* – přeměna znalostí na akční znalosti,
- *Automate* – převedení výsledků analýzy do praxe.

Žádná data nemají význam, jestliže jsou oddělena od kontextu Prvním krokem v analytickém procesu je tedy stanovení kontextu - cílů, strategií a procesů. Materiál SPSS [5A, 2000] k tomu říká:

- Určete data, jejichž sběr, pořízení a skladování je nutné zajistit pro provedení takových analýz, které chcete realizovat.
- Připravte se na své projekty a obory, v nichž rozhodujete - jejich porozuměním zabezpečte ty analytické nástroje, které potřebujete.
- Vzdělávejte a trénujte všechny lidi, kteří myslí analyticky a používají efektivně software jako součást přemýšlení nad problémy a analýzu dat jako imanentní složku rozhodovacího procesu.

Druhým krokem v metodologii *5A* je sběr a příprava dat. Je třeba získat vhodné soubory z podnikových datových skladů, datových bází, odkazových systémů a jiných interních zdrojů. Lze využít i data týkající se daného problému, která jsou nabízena veřejně (oficiální statistiky, rezortní data, demografické a psychografické údaje apod.). Data lze rovněž získat vlastními průzkumy nebo od výzkumné firmy.

Třetím krokem je používání různých analytických postupů k tomu, aby byly nalezeny odpovědi na otázky stanovené v prvním kroku. V tomto kroku se data přeměňují na informace a znalosti. Firma SPSS doporučuje širokou škálu nástrojů pro zkoumání a porozumění datům počínaje deskriptivní statistikou, přes metodu OLAP až po metody strojového učení (rozhodovací stromy, neuronové sítě). Doporučení je zřejmé: „*Použijte více metod a porovnejte jejich výsledky a vhodnost, abyste získali nejlepší řešení a navíc rychle a jednoduše“*.

Čtvrtý krok procesu obsahuje doporučení, řadu dodatečných otázek a následné rozhodnutí. Znalosti nalezené v předcházejícím kroku se zde mění na znalosti akční. Nalezené výsledky by měly být předkládány v jasné a srozumitelné podobě.

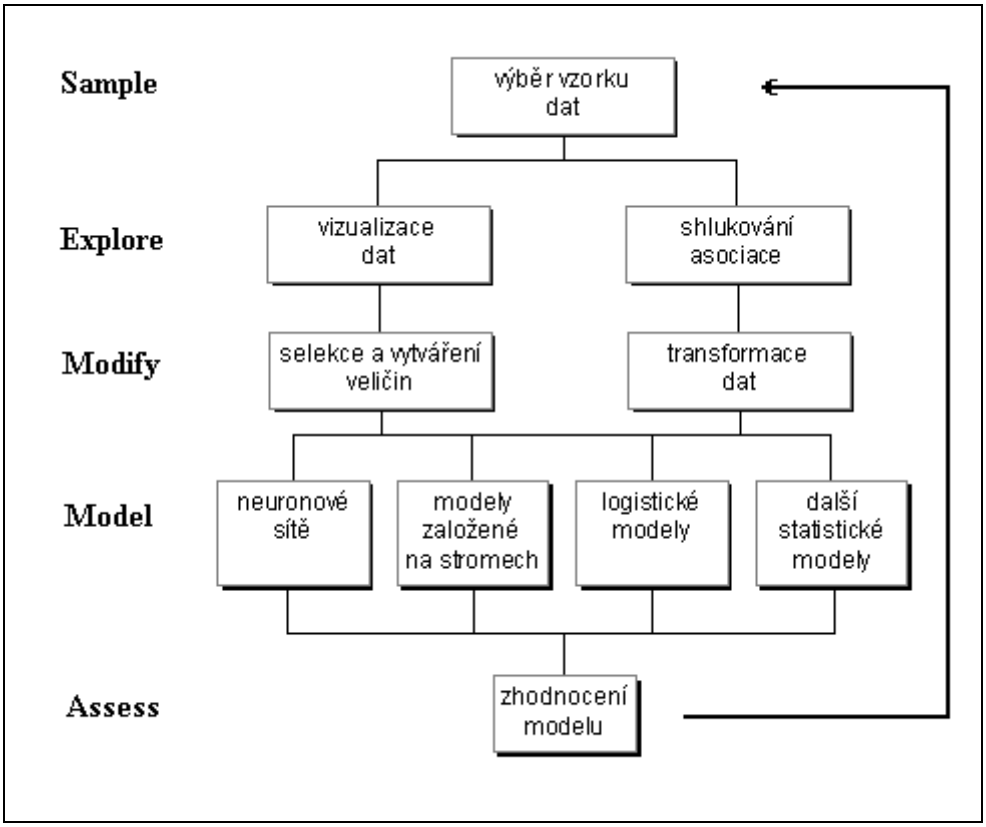
Pátým krokem je převedení výsledků analýzy do praxe. Tento krok obsahuje všechny činnosti, kterými lze zajistit aplikaci učiněných rozhodnutí. Sem patří např. vytvoření praktického rozhraní k rozvinutí nalezených modelů do takového formátu, který je snadný pro užívání a porozumění v běžné a opakované praxi organizace a monitorování výsledků (a důsledků) prováděných rozhodnutí. Další z doporučení zní „*Automatizujte své analýzy tak, aby opakující se úlohy nezabíraly čas a abyste mohli snadno aktualizovat své modely s tím, jak přicházejí nové výsledky*“.

### 1.2.2 Metodika SEMMA

*Enterprise Miner*, softwarový produkt firmy SAS, vychází z vlastní metodiky pro dobývání znalostí z databází. Název *SEMMA* opět charakterizuje jednotlivé prováděné kroky:

- *Sample* (vybrání vhodných objektů),
- *Explore* (vizuální explorace a redukce dat),
- *Modify* (seskupování objektů a hodnot atributů, datové transformace),
- *Model* (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování),
- *Assess* (porovnání modelů a interpretace).

Důraz se klade na snadnou interpretaci výstupů ve formě srozumitelné obchodnímu uživateli.



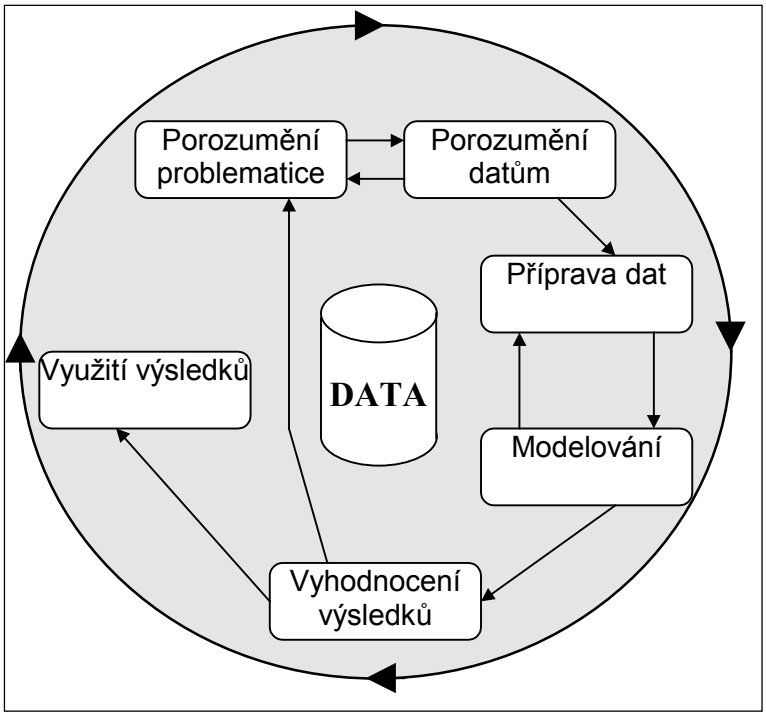

*Obr. 9 Metodika SEMMA*

### 1.2.3 CRoss-Industry Standard Process for Data Mining

Metodika *CRISP-DM* (CRoss-Industry Standard Process for Data Mining) vznikla v rámci Evropského výzkumného projektu. Cílem projektu bylo navrhnout univerzální postup (tzv. standardní model procesu dobývání znalostí z databází), který bude použitelný v nejrůznějších komerčních aplikacích [Chapman a kol, 2000]. Vytvoření takovéto metodikye umožní řešit rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady. Kromě návrhu standardního postupu má *CRISP-DM* nabízet „průvodce“ potenciálními problémy a řešeními, které se mohou vyskytnout v reálných aplikacích.

Na projektu spolupracovaly firmy NCR (přední dodavatel datových skladů), DaimlerChrysler, ISL (tvůrce systému Clementine) a OHRA (velká holandská pojišťovna). Všechny tyto firmy mají bohaté zkušenosti s reálnými úlohami dobývání znalostí z databází.

Životní cyklus projektu dobývání znalostí je podle metodiky *CRISP-DM* tvořen šesti fázemi (Obr. 10). Pořadí jednotlivých fází není pevně dáno. Výsledek dosažený v jedné fázi ovlivňuje volbu kroků následujících, často je třeba se k některým krokům a fázím vracet. Vnější kruh na obrázku symbolizuje cyklickou povahu procesu dobývání znalostí z databází jako takovou.




*Obr. 10 Metodiky CRISP-DM*

Jednotlivé fáze metodiky *CRISP-DM* bude ilustrovat poměrně realistický příklad úlohy dobývání znalostí z dat v bance XY [Berka, 1999]. Nejprve několik slov k pozadí úlohy:

> *Banka XY je zaměřena na drobné klienty kterým vede účty, poskytuje půjčky apod. Pod rostoucím tlakem konkurence chce tato banka zlepšit své služby. Management banky má jen velmi vágní představu, co je možno od metod dobývání znalostí očekávat. Doufá ale, že mu tyto nové metody umožní lépe pochopit klienty a tak například cíleněji nabízet své produkty, nebo rozlišovat mezi různými skupinami klientů (bonitní resp. problémoví).*

### *Porozumění problematice (Business Understanding)*

Tato úvodní fáze je zaměřena na pochopení cílů úlohy a požadavků na řešení formulovaných z manažerského hlediska. Manažerská formulace musí být následně převedena do zadání úlohy pro dobývání znalostí z databází.

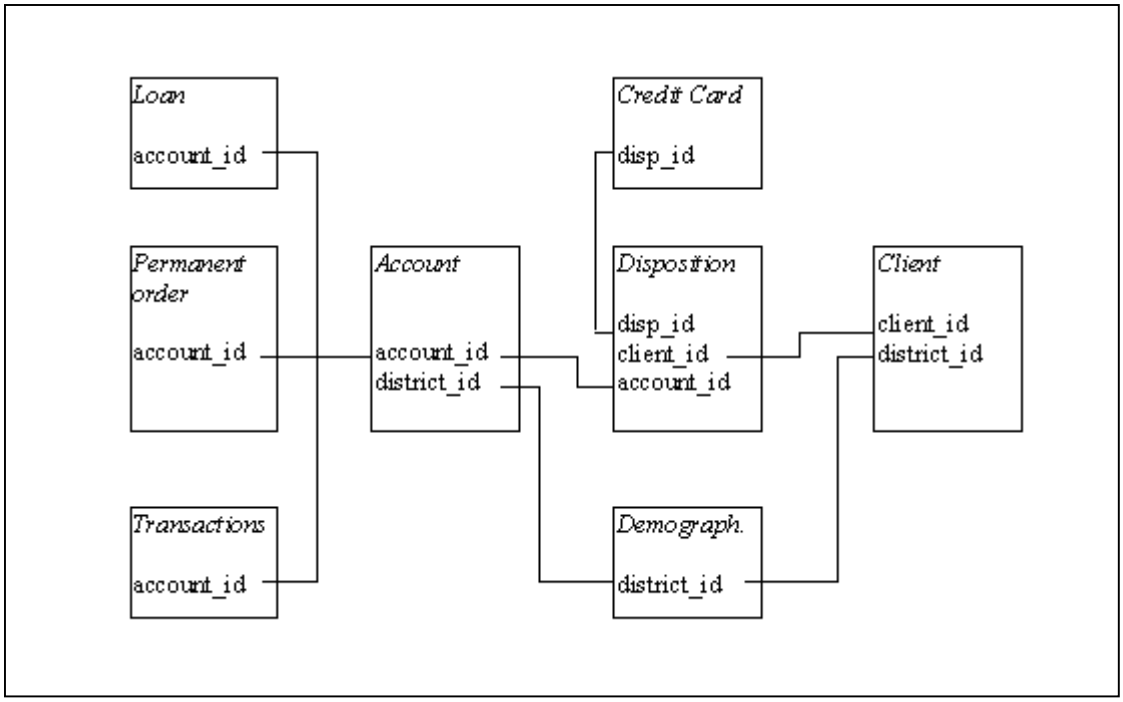
> *Manažerský problém, ke kterému jsou pomocí metod KDD hledány informace, může být formulován (téměř) bez vazby na informace získávané pomocí metod KDD z dostupných dat. Příkladem může být snaha nabídnout uložení části peněz na zvláštní účet s delší výpovědní lhůtou reklamou vhodně zacílenou na v tomto směru nadějnou skupinu klientů (i potenciálních) banky. Pro KDD to znamená nalézt takovou charakteristiku klientů, které zajišťují, že ve skupině klientů s touto charakteristikou bude velká část klientů mít stále dostatečně vysoký zůstatek na účtu. V tomto případě je zadání pro KDD formulováno relativně přesně, přesto je však třeba počítat s možností přeformulování nebo upřesnění manažerského problému na základě provedených analýz. Jinou možnou úlohou je otázka včasného rozpoznání klientů, kteří představují rizikovou skupinu z hlediska splácení poskytnutého úvěru.*

V této fázi se rovněž provádí inventura zdrojů (datových výpočetních i lidských), hodnotí se možná rizika, náklady a přínos použití metod KDD a stanovuje se předběžný plán prací.

### *Porozumění datům (Data Understanding)*

Fáze porozumění datům začíná prvotním sběrem dat. Následují činnosti které umožní získat základní představu o datech, která jsou k dispozici (posouzení kvality dat, první „vhled“ do dat, vytipování zajímavých podmnožin záznamů v databázi…). Obvykle se zjišťují různé deskriptivní charakteristiky dat (četnosti hodnot různých atributů, průměrné hodnoty, minima, maxima apod.), s výhodou se využívají i různé vizualizační techniky.

*Data sledovaná bankou XY mají podobu několika navzájem propojených tabulek (Obr. 11). Základní tabulkou je tabulka Account (účty). S každým účtem může disponovat nějaký klient (tabulka Client). K jednomu účtu může mít přístup více klientů, jeden klient může mít zřízeno více účtů; tato skutečnost je zachycena v tabulce Disposition, která přiřazuje klienty k účtům. Klientovi, který disponuje nějakým účtem, může být k tomuto účtu vydána kreditní karta (tabulka Credit Card). Nejdůležitější údaje o účtech jsou údaje o prováděných operacích, to je zachyceno v tabulce Transactions (transakce). Na některých účtech mohou být zřízeny trvalé platební příkazy (tabulka Permanent order), na základě některých účtů banka poskytuje úvěr (tabulka Loan). Banka poskytla pro analýzu jen určitý (relativně malý) vzorek těchto dat; první představu o podobě dat tedy bylo možno získat relativně jednoduchými nástroji (Access, Excel). Bylo konstatováno, že některé údaje v tabulce transakce (např. konstantní symbol) mají mnoho chybějících hodnot.*




*Obr. 11 Data banky XY*

### *Příprava dat (Data Preparation)*

Příprava dat zahrnuje činnosti, které vedou k vytvoření datového souboru, který bude zpracováván jednotlivými analytickými metodami. Tato data by tedy měla

- obsahovat údaje relevantní k dané úloze,
- mít podobu, která je vyžadována vlastními analytickými algoritmy.

Příprava dat tedy zahrnuje selekce dat, čištění dat, transformace dat, vytváření dat, integrování dat a formátování dat. Tato fáze je obvykle nejpracnější částí řešení celé úlohy. Jednotlivé úkony jsou obvykle prováděny opakovaně, v nejrůznějším pořadí.

*Vzhledem k tomu, že data o klientech a jejich účtech jsou uložena v několika tabulkách navzájem spojených relacemi 1:n, n:1 a n:m, velkou část předzpracování představovalo vytvoření jediné tabulky obsahující údaje vybrané z více tabulek. Příslušné operace vedoucí k tomuto cíli tedy zahrnují agregování hodnot odpovídajících jednomu klientovi. Dalšími operacemi bylo např. vypočtení průměrných měsíčních zůstatků, průměrných měsíčních příjmů, převod rodného čísla klienta na jeho věk a pohlaví, apod.*

## *Modelování (Modeling)*

V této fázi jsou nasazeny analytické metody (algoritmy pro dobývání znalostí). Obvykle existuje řada různých metod pro řešení dané úlohy, je tedy třeba vybrat ty nejvhodnější (doporučuje se použít více různých metod a jejich výsledky kombinovat) a vhodně nastavit jejich parametry. Jde tedy opět o iterativní činnost (opakovaná aplikace algoritmů s různými parametry), navíc, použití analytických algoritmů může vést k potřebě modifikovat data a tedy k návratu k datovým transformacím z předcházející fáze.

*Pro hledání zajímavých skupin klientů je možno použít metody shlukování nebo asociační pravidla. Pro rozpoznání rizikových klientů z hlediska půjček jsou (vzhledem k tomu, že jedna z tabulek obsahuje informace o průběhu splácení) vhodné např. algoritmy pro tvorbu rozhodovacích stromů nebo rozhodovacích pravidel. Tyto metody je vhodné kombinovat, např. shluková analýza může rozdělit klienty do skupin a rozhodovací strom pak umožní jednotlivé skupiny charakterizovat dostatečně srozumitelným způsobem.*

Součástí této fáze je rovněž ověřování nalezených znalostí z pohledu metod dobývání znalostí. To může představovat např. *testování* klasifikačních znalostí na nezávislých datech.

*Znalosti „deskriptivní“ (charakteristiky skupiny klientů „zajímavých“ z hlediska připravovaného produktu) byly předloženy expertům z banky. Znalosti klasifikační (umožňující „rozpoznat“ klienty, kteří nesplácejí úvěr) byly testovány na novém vzorku dat.*

## *Vyhodnocení výsledků (Evaluation)*

V této fázi jsme se dopracovali do stavu, kdy jsme nalezli znalosti, které se zdají být v pořádku z hlediska metod dobývání znalostí. Dosažené výsledky je ale ještě třeba vyhodnotit z pohledu manažerů, zda byly splněny cíle formulované při zadání úlohy.

*Některé nalezené skupiny klientů experty nepřekvapily, vědělo se o nich a banka se připravovala je oslovit dopisem. Jiné (rovněž bonitní skupiny) byly shledány zajímavými, ale budou ještě podrobeny dalšímu zkoumání. Výsledky testování klasifikačních znalostí ukázaly, že systém byl příliš „přísný“, tedy správně rozpoznával klienty rizikové, ale v určitých případech (obzvláště u vyšších půjček) za rizikové označil i klienty bonitní.Bylo tedy rozhodnuto, že na všech pobočkách banky bude využíván program, který bude rozhodovat o úvěrech do určité částky.*

Na závěr této fáze by mělo být přijato rozhodnutí o způsobu využití výsledků.

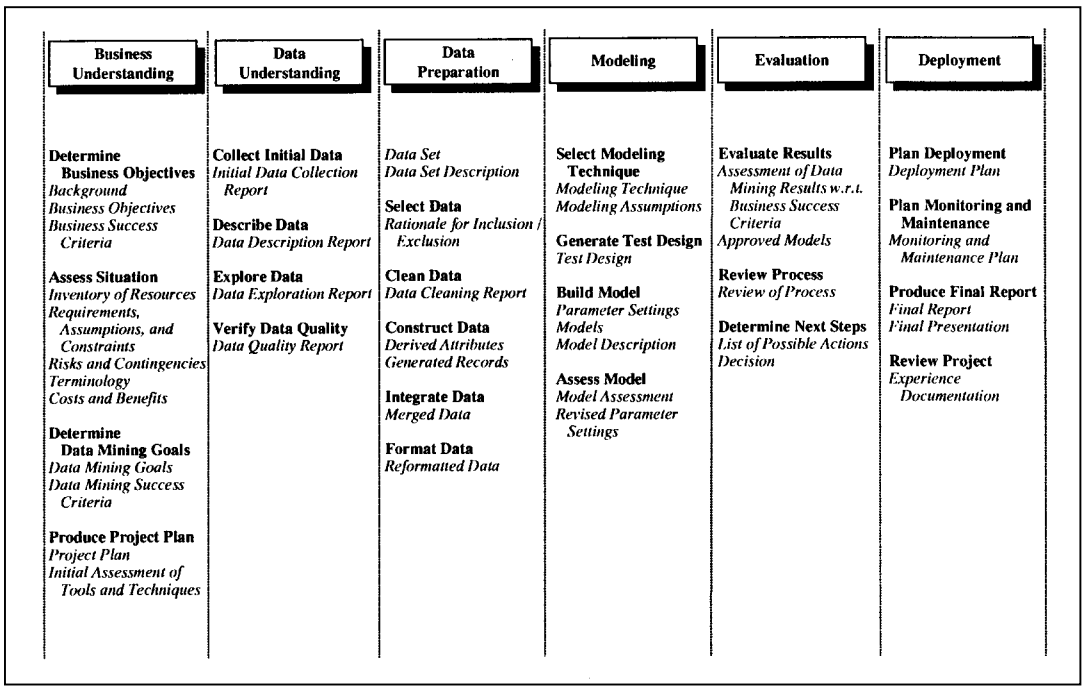
### *Využití výsledků (Deployment)*

Vytvořením vhodného modelu řešení úlohy obecně nekončí. Dokonce i v případě, že řešenou úlohou byl „pouze“ popis dat, získané znalosti je třeba upravit do podoby použitelné pro zákazníka (manažera - zadavatele úlohy). Podle typu úlohy tedy využití (nasazení) výsledků může na jedné straně znamenat prosté sepsání závěrečné zprávy, na straně druhé pak zavedení (hardwarové, softwarové, organizační) systému pro automatickou klasifikaci nových případů.

Ve většině případů je to zákazník a nikoliv analytik, kdo provádí kroky vedoucí k využívání výsledků analýzy. Proto je důležité, aby pochopil, co je nezbytné učinit pro to, aby mohly být dosažené výsledky efektivně využívány.

> *Systém pro rozhodování o půjčkách bude nasazen ve dvou fázích. V první fázi bude systém nasazen jen na vybraných pobočkách, po tomto poloprovozním ověření pak bude nasazen všude. Toto rozhodnutí vyžaduje:*
> - *implementaci klasifikačního algoritmu v uživatelsky přátelské podobě,*
> - *přípravu uživatelského manuálu*
> - *instalaci programu na všech pobočkách banky*
> - *zaškolení uživatelů*
> - *změnu metodiky poskytování úvěrů a tomu odpovídající změnu vnitřních předpisů banky*

Celkové schéma jednotlivých kroků metodiky CRISP-DM ukazuje Obr. 12. Jednotlivé kroky procesu dobývání znalostí jsou různě časově náročné a mají i různou důležitost pro úspěšné vyřešení dané úlohy. Praktici v oboru uvádějí[6], že nejdůležitější je fáze porozumění problému (80% významu, 20% času) a časově nejnáročnější je fáze přípravy dat (80% času, 20% významu). Překvapivě málo práce zaberou vlastní analýzy (5% času, 5% významu).




*Obr. 12 Úlohy v metodologii CRISP-DM [Chapman a kol. 2002]*

[6] Následující údaje jsou převzaty z vystoupení J.U. Kietze na konferenci PKDD 2000.

## Literatura:


[Anand a kol, 1996] Anand, S. a kol.: Towards Real-World Data Mining. In: Practical Aspects of Knowledge Management, Schweizer Informatiker Gesellschaft, Basel, 1996.

[Berka, 1999] Berka,P.: Guide to Financial Data Set. In: Workshop notes on Discovery Challenge, PKDD'99, 1999.

[Fayyad a kol, 1996] Fayyad,U. – Piatetsky-Shapiro,G. – Smyth,P. – Uthurusamy,R. eds.: Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, 1996, ISBN 0-262-56097-6.

[Chapman a kol, 2000] Chapman,P. – Clinton,J. – Kerber,R. – Khabaza,T. – Reinartz,T. – Shearer,C. – Wirth,R.: CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc. 2000.

[Klosgen, Zytkow, 1997] Klosgen,W. - Zytkow,J.: Knowledge Discovery and Data Mining. Tutorial Notes. PKDD'97. Trondheim.

[Rauch, 1999] Rauch,J.: Získávání znalostí z databází. Podklady pro posluchače semináře IISEM 491 v letním semestru 1998/99, VŠE, 1999.

[SEMMA,2002] SASS: SEMMA. Internet, http://www.sas.com/products/miner/semma.html, 2002

[5A, 2000] SPSS: Metoda 5A. Internet, http://www.spss.cz/datamin_jak.html, 2000.