BERKA, Petr. *Dobývání znalostí z databází*. Praha : Academia, 2003. 366 s. ISBN 80-200-1062-9.

Databáze jsou dnes samozřejmým prostředkem používaným pro uchovávání dat. Poslouží dobře knihovně pro ukládání údajů o knihách, čtenářích, výpůjčkách apod., stejně jako bance, pojišťovně, obchodnímu řetězci atd. Data je zapotřebí nejen ukládat ve vhodné struktuře, ale také se k nim později „dostat" (obvykle formou dotazu). Je přece nutné vědět, zda čtenář vrací knihy včas nebo zda klient pravidelně splácí poskytnutý úvěr. Časem databáze zachycují nejen „živý" odraz reality (aktivní výpůjčky, denní obraty zboží apod.), ale také historii (uzavřené pojistné události, splacené/nesplacené úvěry, odhalené pojistné podvody atd.). Taková data zdánlivě jen dokumentují, co se kdysi stalo. Ve skutečnosti mohou skrývat souvislosti, jejichž znalost může pomoci vyvarovat se předchozích chyb, nebo naopak těžit z osvědčených postupů. Komu by se např. chtělo půjčovat peníze někomu, o němž lze s velkou pravděpodobností říci, že je nevrátí? A přesně v tom tkví podstata nové disciplíny, která je anglicky označována jako *knowledge discovery in databases (KDD)*, což česky nejlépe vystihuje název recenzované knihy (tedy *dobývání znalostí z databází)* a tomu odpovídající zkratka DZD (u jiných autorů se lze rovněž setkat s označeními *dolování dat, dobývání dat, vytěžování dat* nebo anglickým termínem *data mining*).

Kniha je rozdělena do čtyř částí. V první – „Dobývání znalostí z databází" – jsou objasněny kořeny nové disciplíny, představeny typicky řešené úlohy a uvedeny aktuálně používané metodiky (5A, SEMMA, CRISP-DM). Druhá část – „Tři zdroje" – ukazuje ty principy z oblasti databázových systémů (EIS, OLAP, datové sklady), statistiky (kontingenční tabulky, regresní, diskriminační a shluková analýza) a strojového učení, které DZD ovlivnily nejvíce. Cílem třetí, nejobsáhlejší části – „Proces dobývání znalostí" – je charakteristika těch kroků procesu DZD, které lze uvést v obecné rovině nezávisle na konkrétní úloze a aplikaci. Jde o přípravu dat, modelování a vyhodnocování výsledků. Poslední, čtvrtá část – „Systémy a úlohy" – podává přehled o nabídce na současném trhu softwaru a současně ukazuje podrobnější příklad aplikace. V příloze je pak popsán jazyk PMML (Predictive Model Markup Language), který je aplikací jazyka XML a slouží pro reprezentaci modelů získaných v procesu DZD.

Kniha je vybavena doprovodným CD, jehož obsah je zpřístupněn jako sada lokálních webových stránek, k jeho použití proto postačuje čtenáři běžný webový prohlížeč. Obsah CD lze rozčlenit do čtyř částí: 1. software pro DZD, 2. sady testovacích dat pro různé typy úloh, 3. informace o výzkumných projektech a 4. uspořádaný soubor odkazů na zdroje dostupné prostřednictvím internetu (včetně plných textů některých zásadních monografií).

V čem lze spatřovat přínos knihy? Především je to první původní česká monografie na dané téma. Další příspěvek spočívá ve snaze o vyjasnění terminologie a zavedení vhodných českých ekvivalentů. Vždyť již samotné označení *knowledge discovery in databases* nebo *data mining* se překládají různě, různí autoři je používají jako synonyma, jiní je rozlišují podle použité metodiky atd. Za cenné lze považovat i to, že autor nepojal téma staticky – nejde o suchopárné čtení o různých teoretických přístupech (od relativně jednoduchých rozhodovacích stromů přes asociační pravidla až po neuronové sítě či evoluční algoritmy), ale výklad je osvětlován množstvím příkladů, ukázkami konkrétních systémů a aplikací. Kniha má otevřený konec, protože autor naznačuje nové směry, kterými se bude DZD ubírat. Zde stojí za zmínku především dobývání znalostí z webu (dobývání znalostí na základě obsahu, struktury nebo používání webu).

Kniha je čtivá, ale rozhodně se nejedná o oddechovou literaturu, protože ji nelze otevřít bez určité úrovně vstupních znalostí z oblasti matematiky, statistiky a informatiky. Pokud tak někdo učiní, daleko se nedostane. Na druhou stranu, znalosti na úrovni základních vysokoškolských kurzů nejsou příliš omezující a potenciální okruh čtenářů je proto velmi široký. Autor nezamýšlel vytvořit učebnici, nicméně lze očekávat, že kniha se dostane do seznamů povinné literatury řady vysokoškolských kurzů. Je pravděpodobné, že mnohé čtenáře kniha inspiruje k úvahám, jaké nevytěžené poklady se mohou skrývat v jejich databázích či datových skladech. Ne náhodou se říká, že historie je učitelkou života. Tato kniha a zejména její předmět, který rozpracovává, je toho vynikajícím důkazem.

Vilém Sklenák
Katedra informačního a znalostního inženýrství,
fakulta informatiky a statistiky,
Vysoká škola ekonomická
sklenak@vse.cz