

Získávání znalostí z dat

Úvodní informace



Ivana Burgetová

UPA 2022

Obsah

- Úvod
- Data pro získávání znalostí
- Základní typy dolovacích úloh
- Analytické projekty

Proč získávání znalostí z dat?

- Extrémní nárůst objemu dostupných dat:
 - TB – PB – EB – ...
 - Pokrok ve schopnosti sběru dat a jejich dostupnosti:
 - Databázové systémy, web, mobilní zařízení, automatizované nástroje pro sběr, ...
 - Hlavní zdroje obrovského množství dat
 - Podnikání: Web, e-obchod, transakce, burza, ...
 - Věda: Vzdálené monitorování, genomika, medicína, ...
 - Společnost: zprávy, e-maily, digitální fotoaparáty, ...
- Topíme se v datech, ale postrádáme znalosti!

Vývoj DB technologií a role dat

- ❑ 60. léta – počátek DB technologií, síťové a hierarchické databáze
- ❑ 70. léta – relační model dat, první osobní počítače
- ❑ 90. léta – web, datové sklady a OLAP, počátky získávání znalostí z dat, nerelační rozšíření SŘBD
- ❑ 21. století:
 - Webové aplikace, vznik NoSQL databází, cloud computing, sociální sítě, technologie pro distribuci digitálního obsahu.
 - Proudý dat, *big data*, *data science*

Proč získávání znalostí z dat

- Role dat uložených v databázích či jiných úložištích výrazně vzrostla

evidence → rozhodování řízené daty

- Data - významný zdroj potenciálně užitečných informací
- Data a schopnost extrahovat z nich užitečné informace = strategické aktivum (majetek) firmy.
- Získávání znalostí z dat je proces, který zahrnuje etapu přípravy/předzpracování.

Poznámky k terminologii

- Lze se setkat s pojmy:
 - Získávání znalostí z databází/Dolování z dat
(Knowledge Discovery in Databases/Data mining (KDD/DM)),
 - Data science,
 - Business Intelligence (BI).
- Vztah KDD/DM a Data science
 - KDD/DM populární (199x) vs. Data science (201x)
 - Přesné definice ani jednoznačný výklad neexistují
 - Zpravidla společné rysy: získávání znalostí z dat, multidisciplinární obor zahrnující mj. statistiku a strojové učení
 - Rozdíly různě, např. v [4]: data science se zabývá principy, DM algoritmy

Poznámky k terminologii

□ Business intelligence

(česky někdy *podniková intelligence*)

- Zpravidla metody, techniky a nástroje sloužící k **podpoře rozhodování** (dotazování, tvorba sestav, OLAP, pokročilé analýzy, vizualizace atd.).
- BI orientovaná spíše na analýzy/hodnocení týkající se minulosti a přítomnosti, Data science má více průzkumný charakter týkající se současnosti a rozhodování v budoucnu.

Pojem znalost

- **Pojem znalost** (v kontextu získávání z dat) - ***zajímavá informace***, která je:
 - **Netriviální** (nelze je získat např. pouhým SQL dotazem nebo logickým odvozováním)
 - **Skrytá** (modely a vzory, které nejsou na první pohled vidět)
 - **Dříve neznámá**
 - **Potenciálně užitečná** (má využití - pro rozhodování, potvrzení hypotézy)

Oblasti využití - příklady

- Analýza trhu & management
 - *Příklady:* Analýza nákupního košíku, cílená reklamní kampaň.
- Finanční analýza & analýza rizik
 - *Příklad:* Rozhodnutí o poskytnutí půjčky zákazníkovi.
- Detekce podvodů a neobvyklého chování
 - *Příklady:* Neobvyklé bankovní transakce → praní špinavých peněz, získání přístupu k cizímu účtu, pojištění aut, boj s terorismem.
- Získávání znalostí z textu a na webu
 - *Příklady:* Analýza e-mailů a detekce spamů, analýza sentimentu, analýza falešných zpráv.
- XXXinformatika
 - *Příklady:* Predikce rozpustnosti proteinu (bioinformatika), klasifikace typu hvězdy na základě vyzařovaného spektra (astroinformatika), klasifikace ploch dálkového průzkumu Země (geoinformatika).

Obsah

- Úvod
- Data pro získávání znalostí
- Základní typy dolovacích úloh
- Analytické projekty

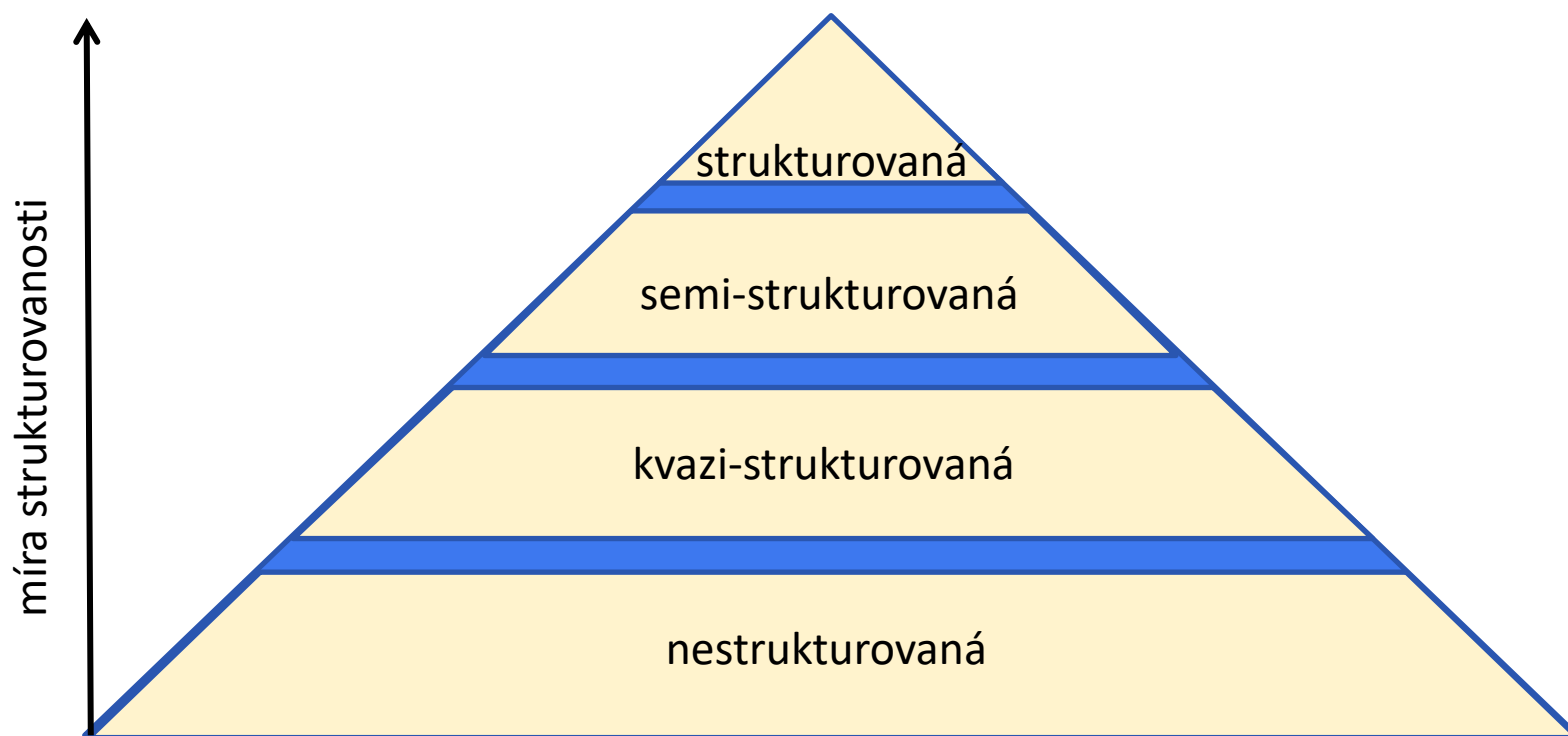
Klasifikace struktury dat

- **Strukturovaná** – lze rozpoznat určitou strukturu
 - („Plně“) **strukturovaná** – relační a transakční databáze, data z datových kostek, CSV atd.
Příklady: tabulka relační databáze, soubor obsahu nákupů v supermarketu.
 - Částečně (semi-)strukturovaná – textová data s rozeznatelným vzorem, který umožňuje syntaktickou analýzu (parsing).
Příklady: dokument v XML, HTML, JSON, BibTeX.
 - „kvazi“ **strukturovaná** – textová data s nepravidelnými formáty, která mohou obsahovat nekonzistence v hodnotách a formátech.
Příklad: záznam webovými stránkami (clickstream), např. FIT → Pro studenty → Studijní informace.

Klasifikace struktury dat

□ **Nestrukturovaná** – nemají žádnou strukturu.

Příklady: textové dokumenty (prostý text), PDF dokumenty, obrázky, video.



Typické zdroje dat

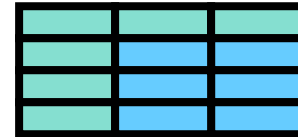
□ Úložiště „statických“ dat

■ „tradiční“

- relační databáze – tabulky
- databáze transakčních dat – (IDtransakce, položky)
- datové sklady – vícedimenzionální kostka
- soubory strukturovaných dat

■ „pokročilá“

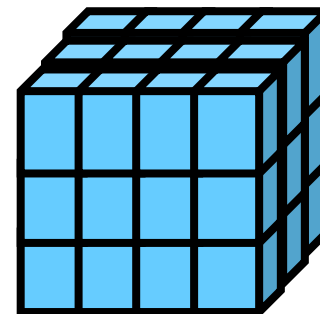
- nerelační a NoSQL databáze, soubory nestrukturovaných dat



tabulka

ID transakce	ID zboží
T100	10, 12, 15
T200	10, 18
...	...

transakční databáze



multidimenzionální kostka

Typické zdroje dat

- Web a sociální sítě
- Proudý dat - potenciálně nekonečný → data dostupná jen dočasně.
Příklady: data ze senzorů, dohledové systémy, bankovní transakce.
- Velmi rozsáhlá data (Big Data) – obecně směs dat ze zdrojů výše, zpravidla charakterizovaná několika „V“:
 - objem (Volume) – dat je hodně
 - rychlost vzniku (Velocity) – vznikají rychle
 - rozličnost (Variety) – různé zdroje, typy, formáty atd.
 - věrohodnost (Veracity) – nejistá data vlivem nekonzistence, neúplnosti, nejasnosti apod.

Obsah

- Úvod
- Data pro získávání znalostí
- Základní typy dolovacích úloh
- Analytické projekty

Prediktivní analýza

- V terminologii strojového učení (ML) učení s učitelem (*supervised learning*).
- **Podstata:** predikce třídy/konceptu daného objektu (**klasifikace**) nebo predikce/odhad hodnoty nějaké neznámé vlastnosti objektu (**regrese**).
- **Výsledek/znalost:** model, který popisuje a rozlišuje třídy/koncepty (u klasifikace) nebo predikuje hodnoty *spojité* veličiny.
 - **Příklad:** Klasifikace zákazníků banky z hlediska schopnosti splácet úvěr na základě historických dat.
 - **Příklad:** Predikce hodnoty akcie na základě předchozího vývoje.

Prediktivní analýza

□ Klasifikační metody:

- Rozhodovací strom (ID3, C4.5, CART), náhodný les
- Bayesovská klasifikace
- Neuronové sítě
- SVM
- Rozhodovací pravidla
- k-NN

□ Regresní analýza:

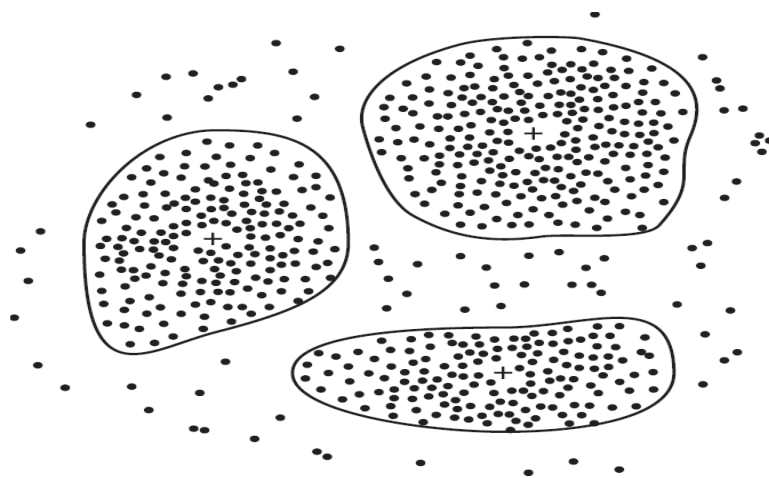
- Využití různých regresních funkcí
- Neuronové sítě
- SVM
- k-NN

Porovnávání podle podobnosti

- Identifikace **podobných** objektů.
 - K danému objektu hledáme podobné objekty
 - *Příklad:* Nalezení uživatelů e-shopu podobných z pohledu nakupovaných produktů – **doporučování produktů**.
- Metody:
 - k-NN

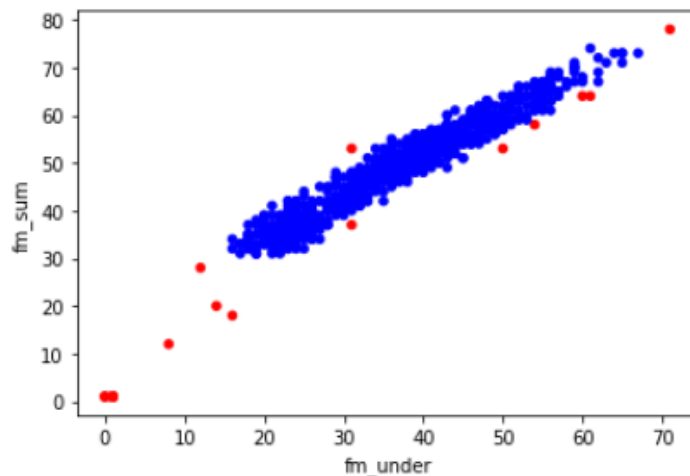
Shlukování

- V terminologii strojového učení bez učitele (*unsupervised learning*).
- **Podstata:** nalezení skupin/shluků **podobných** objektů. Ty představují předem neznámé třídy.
- **Výsledek/znalost:** kolekce shluků s co největší podobností uvnitř shluků a co nejmenší podobností mezi shluky.
 - **Příklad:** Nalezení shluků zákazníků pro cílenou reklamu.
- **Metody:**
 - K-means
 - Hierarchické shlukování
 - DBSCAN
 - EM



Analýza/detekce odlehlých objektů

- ❑ **Odlehlá hodnota** (angl. outlier): datový objekt, který se vymyká z hlediska podobnosti nebo chování ostatním.
- ❑ **Podstata:** hledání/detekce (online) odlehlých objektů v datech.
- ❑ **Výsledek:** seznam odlehlých objektů, resp. signál odlehlosti (při detekci) pro daný objekt.
 - *Příklad:* Detekce podvodů, analýza odlehlých trajektorií pohybujících se objektů, zjišťování odlehlých hodnot.
- ❑ **Metody:**
 - LOF
 - Isolation forest
 - OC-NN
 - OC-SVM
 - DBSCAN



Seskupování podle současnosti výskytu

- ❑ **Podstata:** hledání objektů/vztahů v datech, které se často vyskytují někde společně.
- ❑ **Výsledek/znalost:** seznam množin objektů/vztahů vyskytujících se dostatečně často společně – frekventované množiny (vzory).
 - *Příklad:* Nalezení položek nákupu, které zákazníci nakupují často společně – analýza nákupního košíku (transakce).
- ❑ Z frekventovaných vzorů lze odvodit **asociační pravidla** tvaru $A \Rightarrow B$ – *když se vyskytne A, pak se často vyskytne i B* – vztah k $P(B | A)$.
- ❑ Sílu pravidla vyjadřují metriky – podpora a spolehlivost
 - *Příklad:* Jaké zboží (B, obecně množina) si často kupují zákazníci společně s daným zbožím A, obecně množina)? – doporučení.
Například asociačního pravidlo: Fotoaparát $\rightarrow$ SD karta [0.5%, 75%]
- ❑ Metody: algoritmus Apriori

Sekvenční vzory

- Obdoba frekventovaných množin/vzorů aplikovaná na sekvence.
- **Podstata:** hledání podsekvencí vyskytujících se často v nějaké množině sekvencí.
- **Výsledek:** seznam podsekvencí/sekvenčních vzorů vyskytujících se v sekvencích dostatečně často.
 - *Příklad:* Jaké jsou časté posloupnosti nákupů našich zákazníků → koupí A, později B a C a následně D – **doporučování**.
 - *Příklad:* Jaké vzory se vyskytují často v dané databázi bio-sekvencí (např. DNA nebo proteinů)?
- Metody: GSP (vlastnost Apriori), SPADE/SPAM, LAPIN

Profilování/charakterizace

- ❑ **Podstata:** hledání vlastností, resp. chování určité třídy/skupiny objektů, které je charakterizují.
- ❑ **Výsledek/znalost:** profil/charakterizace objektů dané třídy.
 - **Příklad:** Kteří zákazníci utratí v daném obchodě více než 100.000 Kč ročně? → věk 40-50 let, zaměstnaný, plat větší než 30.000
- ❑ Variantou může být **diskriminace** – hledání vlastností, ve kterých se **objekty dvou tříd** nejvíce liší.
 - **Příklad:** Čím se liší zákazníci, kteří nakupují v daném obchodě pravidelně (více než 2x za měsíc) od těch, kteří nakupují zřídka (méně než 3x ročně)? → první skupina: 20 – 40 let, VŠ vzdělání × druhá skupina: junioři a senioři, bez VŠ vzdělání.

Obsah

- Úvod
- Data pro získávání znalostí
- Základní typy dolovacích úloh
- Analytické projekty

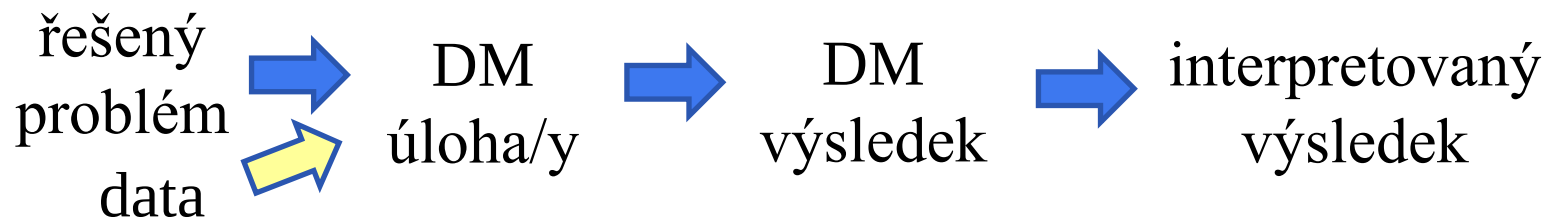
Analytické projekty

□ Vstupy:

- byznys problém rozhodování na základě dat (DDD – Data-Driven Decision making) nebo hypotéza k ověření na základě dat
- data

□ Výstupy:

- využitelná znalost – model dat, vzory v datech, profily atd.
- kvantitativní ohodnocení kvality výsledku

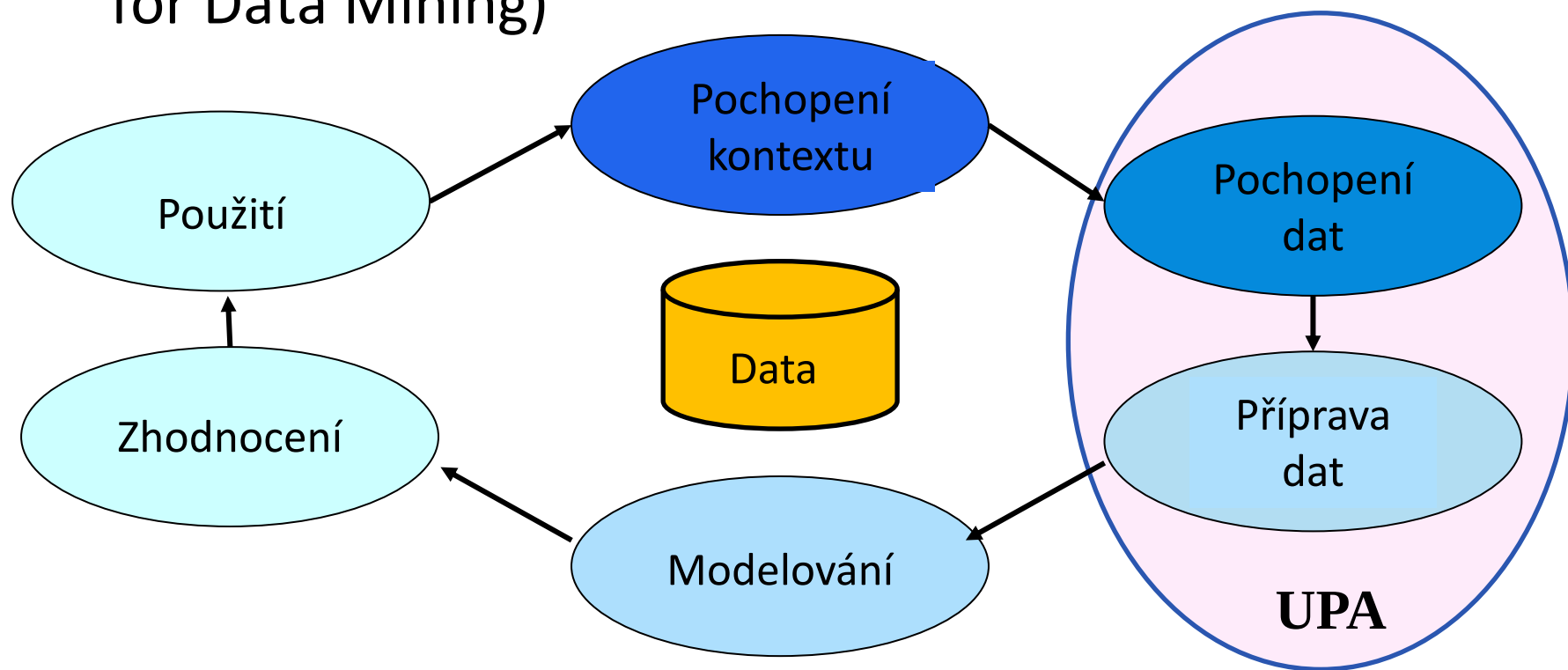


□ Analytické vs. SW projekty

- zaměření na data a aplikace existujících algoritmů pro řešení dílčích úloh, ne na algoritmy zpracování dat
- klíčová role doménových analytiků a doménových expertů

Životní cyklus analytického projektu

- ❑ Řada modelů zahrnujících podobné kroky/fáze
- ❑ Model **CRISP-DM** (CRoss-Industry Standard Process for Data Mining)



Kroky/fáze modelu CRISP-DM

- ❑ **Pochopení kontextu** – pochopení problematiky, určení cílů projektu a kritérií úspěšnosti, zhodnocení zdrojů vč. dostupných dat, definice analytického cíle (úloh(y) získávání znalostí, plánování projektu, ...
- ❑ **Pochopení dat** – rozpracování informace k dostupným/potřebným datům, případný sběr dat, získání charakteristik dat (explorační analýza), zhodnocení kvality dat.
- ❑ **Příprava dat** – výběr, čištění, integrace, úpravy datové sady, transformace dat.
- ❑ **Modelování** – výběr modelovací techniky, příprava způsobu ohodnocení kvality modelu, vytvoření modelu, ohodnocení modelu.
- ❑ **Zhodnocení** – interpretace a zhodnocení výsledku z hlediska cílů projektu, stanovení dalšího postupu.
- ❑ **Použití** – plán využití, monitorování a údržby

Literatura (1)

1. Foote, K. D.: **A Brief History of Database Management**. Dostupné na <https://www.dataversity.net/brief-history-database-management/>.
2. Meier A., Kaufmann M. (2019) **Postrelational Databases**. In: SQL & NoSQL Databases. Springer Vieweg, Wiesbaden 2019 (online). Dostupné na https://link.springer.com/chapter/10.1007/978-3-658-24549-8_6.
3. Han, J., Kamber, M.: **Data Mining: Concepts and Techniques**. Third Edition. Elsevier Inc., 2012, ISBN 978-0-12-381479-1, s. 1 – 37.

Literatura (2)

4. Provost, F., Fawcett, T.: **Data Science for Business**. O'Reilly Media Inc., 2013. s. 1 – 33.
5. EMC Education Services: **Data Science and Big Data Analytics**, Wiley. 2015. s. 1 - 30.