

Big data

Věda o datech – základy a aplikace

Jan Hendl

- Úvod do disciplín spojených s Big daty
- Základy umělé inteligence
- Postupy analýzy dat metodami strojového učení
- Neuronové sítě
- Analýza textů a odhalování fake news
- Identifikace komunit v sociálních sítích
- Věda o datech a zvládnání pandemie Covid-19



Big data

Věda o datech – základy a aplikace

Jan Hendl

Grada Publishing

Věnuji památce prof. PhDr. Jana Průchy, DrSc., dr. h. c., zakladateli moderní české pedagogiky.

Jan Hendl

Big data

Věda o datech – základy a aplikace

Vydala Grada Publishing, a.s.
U Průhonu 22, Praha 7
obchod@grada.cz, www.grada.cz
tel.: +420 234 264 401
jako svou 8251. publikaci

Recenzenti:

Prof. Ing. Petr Berka, CSc.
Doc. MUDr. Pavel Kasal, CSc.

Odpovědná redaktorka Věra Slavíková
Sazba Jan Šístek
Počet stran 224
První vydání, Praha 2021
Vytiskla TISKÁRNA V RÁJLI, s.r.o., Pardubice

Vydání odborné knihy schválila Vědecká redakce nakladatelství Grada Publishing, a.s.

© Grada Publishing, a.s., 2021
Cover Design © Grada Publishing, a. s., 2021
Cover Photo © Depositphotos/klss777

Názvy produktů, firem apod. použité v knize mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

Upozornění pro čtenáře a uživatele této knihy

*Všechna práva vyhrazena. Žádná část této tištěné či elektronické knihy nesmí být reprodukována a šířena v papírové, elektronické či jiné podobě bez předchozího písemného souhlasu nakladatele. Neoprávněné užití této knihy bude **trestně stíháno**.*

ISBN 978-80-271-4479-2 (ePub)
ISBN 978-80-271-4478-5 (pdf)
ISBN 978-80-271-3031-3 (print)

Obsah

Slovo úvodem	9
--------------------	---

Předmluva	11
-----------------	----

1	Základy	13
1.1	Big data.....	13
1.2	Věda o datech	19
1.3	Umělá inteligence.....	24
1.4	Data mining.....	25
1.5	Strojové učení.....	26
1.6	Business Intelligence.....	27
1.7	Datové inženýrství.....	28
1.8	Průnik disciplín.....	28
1.9	Metodologie CRISP DM.....	30
1.10	Strategie používání dat	32
1.11	Výuka vědy o datech.....	34
1.12	Problematika COVID-19	36
1.13	Souhrn.....	38

2	Stručně o umělé inteligenci	40
2.1	Vybrané principy umělé inteligence	42
2.2	Expertní systémy.....	46
2.3	Pravidla v expertním systému.....	48
2.4	Souhrn.....	54

3	Uspořádání dat	55
3.1	Plochý soubor (Flat file).....	55
3.2	HTML a XML soubory.....	56
3.3	JSON formát.....	57
3.4	SQL databáze.....	58
3.5	NoSQL databáze	60
3.6	Hadoop, MapReduce, Spark, Mahout, HBase.....	61
3.7	Datový sklad.....	65
3.8	Jezero dat.....	66

3.9	Organizace dat a datová matice	66
3.10	Souhrn	68

4	Strojové učení – přehled	70
4.1	Principy algoritmů strojového učení	71
4.2	Typy proměnných a příprava dat	74
4.2.1	Volba proměnných	75
4.2.2	Chybějící hodnoty	75
4.3	Učení s učitelem	76
4.3.1	Lineární regrese a její modifikace	77
4.3.2	Logistická regrese	81
4.3.3	Složitější typy regresí (GLM a GAM)	81
4.3.4	K-nejbližších sousedů (regrese a klasifikace)	82
4.3.5	Lineární diskriminační analýza	82
4.3.6	Klasifikace podpůrnými vektory	83
4.3.7	Bayesova metoda	84
4.3.8	Neuronové sítě	84
4.3.9	Rozhodovací strom	85
4.4	Učení bez učitele	87
4.4.1	Shlukování	87
4.4.2	Redukce dimenzionality	90
4.4.3	Analýza asociací	91
4.4.4	Detekce anomálie	93
4.4.5	Autokódování	94
4.5	Učení posilováním	94
4.5.1	Q-učení	95
4.5.2	TD-učení	95
4.6	Algoritmy strojového učení podle podobnosti	95
4.7	Evaluace algoritmů strojového učení	97
4.7.1	Koeficienty kvality klasifikačních modelů	97
4.7.2	Ukazatele kvality regresních modelů	100
4.7.3	Interpreovatelnost prediktivních modelů	100
4.8	Souhrn	104

5	Metoda podpůrných vektorů	106
5.1	Algoritmus metody SVM	106
5.2	Klasifikace pomocí SVM do více tříd	109
5.3	Aplikace SVM pro COVID-19 data	109
5.4	Souhrn	112

6	Bayesova metoda	113
6.1	Modelová aplikace	114
6.2	Bayes a ohrožení COVID-19 pandemií	116
6.3	Souhrn	120

7	Umělé neuronové sítě	121
7.1	Popis jednoduchého perceptronu	121
7.2	Aktivační funkce	124
7.3	Vytvoření architektury neuronové sítě	125
7.4	Problémy procesu učení a interpretace	130
7.5	Typy neuronových sítí	131
7.6	Neuronové sítě a COVID-19	135
7.7	Souhrn	138
8	Rozhodovací stromy a lesy	139
8.1	Rozhodovací stromy – principy	139
8.2	Ensemble algoritmy	148
8.3	Stromy, lesy a COVID-19	150
8.4	Souhrn	153
9	Analýza sociálních sítí	154
9.1	Základy teorie grafů	154
9.2	Sociální síť	158
9.3	Shlukování uzlů – Louvainská metoda	158
9.4	Girvan-Newman algoritmus	159
9.5	PageRank algoritmus	160
9.6	Sociální sítě v pandemii COVID-19	163
9.7	Souhrn	164
10	Analýza textů	166
10.1	Základní definice	167
10.2	Analýza vztahů slov	168
10.3	Shlukování textů	169
10.4	Klasifikace textů	170
10.5	Shrnutí textů	172
10.6	Vyhledávání a analýza témat	173
10.7	Analýza sentimentu a názorů	176
10.8	Kombinovaná analýza	178
10.9	Zavádějící zprávy	180
10.10	Zavádějící texty o COVID-19	185
10.11	Souhrn	187

11	Programovací jazyky, systémy a nástroje	
	zpracování dat	189
11.1	Programovací jazyky	189
11.1.1	Python	189
11.1.2	R	190
11.1.3	Julia	193
11.2	Prostředky pro scraping	193
11.3	Programové systémy pro vytváření modelů	195
11.4	Oblasti datového inženýrství	198
11.5	Souhrn	199
	Závěr	201
	Slovníček anglických termínů	203
	Literatura	216
	Rejstřík	221

Slovo úvodem

Otevíráte knihu významného českého statistika orientovaného především na biomedicínu prof. RNDr. Jana Hendla. Kniha je zaměřena na problematiku označovanou obvykle ve světě anglickým termínem „Big data“. Analýza velkých skupin dat není v informatice a statistice tématem novým. Je to však obor, který se stále dynamicky rozvíjí a publikace vyžadují aktualizaci zejména ve specializovaných oblastech, které se dotýkají jednotlivých částí informatiky a dalších vědních oborů. Právě této specializované problematice se podstatná část knihy věnuje.

Kniha prof. Hendla je z tohoto hlediska velmi aktuální. Příklady jsou s ohledem na zaměření autora v poslední době převážně z medicíny. Příklady k tématu pandemie COVID 19 jsou tím nejaktuálnějším, co si lze dnes představit. Právě v té době se ukázalo, jak významná je analýza dat pro predikci vývoje pandemie i sledování adekvátnosti přijatých opatření.

V mnoha kapitolách druhé části knihy se prof. Hendl věnuje interdisciplinárním tématům např. strojovému učení, podpoře rozhodování, umělé inteligenci, analýze textů a sociálním sítím. Především to jsou oblasti, kde Big data mají velké uplatnění.

Celková koncepce knihy je tedy multidisciplinární a velká data představují jen sjednocující prvek a kapitoly se dotýkají téměř všech oblastí biomedicínské informatiky a mnoha oblastí jiných oborů. To je dáno autorovou pracovní zkušeností statistika v mnoha oborech – v Institutu pro doškolování lékařů a farmaceutů, na Fakultě tělesné výchovy UK, Fakultě sociálních věd UK a na lékařských fakultách UK. Kniha tedy není rozhodně určena jen statistikům a informatikům, ale i lékařům, pracovníkům v humanitních oborech, přírodních vědách, tělovýchově a v nelékařských zdravotnických oborech. Nejvíce budou z knihy profitovat studenti doktorských studijních oborů všech uvedených fakult a dále jistě i studenti informatiky, technických, ekonomických a matematických oborů.

Kniha má logickou strukturu a zdaleka nemusí být čtena jako učebnice od začátku do konce. Lze využít i rejstřík a obsah nebo prostudovat jen úvodní dvě kapitoly a pak číst jen skupiny kapitol podle oboru zájmu.

prof. MUDr. Štěpán Svačina, DrSc., MBA
předseda České lékařské společnosti Jana Evangelisty Purkyně

Předmluva

Sherlock Holmes řekl „Velká chyba spočívá v tom, že teoretizujeme bez dat“. Sherlock Holmes by jistě rád žil v 21. století, protože svět je dnes přímo zaplaven velkými objemy dat. Nová věda, věda o datech (data science), která se jimi zabývá, revolucionizovala obchod, celou vědu a ovlivnila společnost. Věda o datech bere ohled na škálu technologických vymožeností, jako jsou smartphony, digitální propojení pomocí radiových vln, sociální sítě a nové počítačové vybavení.

V každém okamžiku a na mnoha místech se generuje velké množství elektronických stop a dat. Přístup k „Big datům“ vytvořil příležitost využívat počítačové a statistické přístupy k proměně hrubých dat do aktivních znalostí, které mohou podporovat řešení různých aplikačních úloh. To je zvláště pravda při optimalizaci provádění rozhodnutí v oblastech, jako je medicína, bezpečnost, výuka, věda a obchod. Stejně jako mikroskop umožňuje vidět věci v „mikrosvětě“ a obří dalekohledy v dalekém vesmíru, různé prostředky analýzy „Big dat“ nám mohou rozšířit schopnost zachytit užitečnou, ale mnohdy skrytou informaci schovanou v datech.

Věda o datech je typicky zaměřena na analýzu komplexních a velkých objemů dat. Tento obor zahrnuje také aplikaci principů sběru dat, jejich uchování a organizace, integrace, komunikace, statistiku a etiku. Lze předpokládat, že všichni studenti a akademici budou v budoucnu znalosti této vědy využívat. Pracovníci s těmito znalostmi a dovednostmi budou činní vlastně v každé oblasti, přičemž budou mít na starosti operační systémy, přípravu dat k analýze, koordinaci analýzy dat, vizualizaci informací a podporu rozhodování pomocí dat v tom, že objeví v datech vztahy a konfigurace. Dovednosti v práci s daty využijí manažeři, úředníci, novináři, umělci, právníci, učitelé, sociologové a ostatní pracovníci, kteří potřebují schopnost porozumět datům a používat je.

V knize se popisují nejdůležitější procedury pro analyzování velkých množství dat s cílem získat poznatky, které pomáhají uživatelům provádět rozhodnutí v mnoha oblastech lidské činnosti. Také se věnuje prostor výkladu analýzy textů, určitým aspektům analýzy sociálních sítí a organizace dat. V textu se objevují často slova „stroj“, „strojní“ nebo „inženýrství“. Čtenář s humanitním nebo společenskovědním zaměřením by měl uznat oprávněnost jejich použití a potlačit svoji případnou iritaci a jistou nedůvěru k tomu, co tato slova signalizují. V USA a v dalších zemích si lidé příslušných oborů váží a uznávají jejich rozhodující přínos k národnímu bohatství i jejich široké uplatnění (i mimo techniku a obchod).

Zaměřuji se na popis vědy o datech jako samostatné disciplíny a zdůrazňuji její význam. Jiný pohled považuje vědu o datech za část umělé inteligence. Text obsahuje v přehledu popi-

sy strategií, postupů a schémat algoritmů. Uvádím několik příkladů aplikace vědy o datech při hledání řešení problémů spojených s pandemií COVID-19.

Při zpracování látky jsem se opíral především o knihu P. Berky (Berka 2003) a práce J. Maříka a jeho spolupracovníků (Mařík 1–6). Také jsem využil texty českých autorů J. Antocha (1988), J. Klaschky a E. Kotrče (2004), J. Tučkové (2009) a J. Raucha (2013). Znameníť je slovenská knížka M. Tereka, A. Horníkové a V. Labudové (2010). O hlubokém učení a programování v jazyku Python podrobně pojednává v českém překladu F. Chollet (2019). Filosofické pozadí Big dat rozebírá Kitchin (2014). Základy informatiky česky hezky vysvětluje Kernighan (2019). Můj přístup k látce byl ovlivněn pročitáním internetových stránek skupiny nadšených autorů a autorek kolem KDnuggets.

Závěrem bych rád poděkoval oběma recenzentům prof. Ing. Petru Berkovi, CSc. a doc. MUDr. Pavlu Kasalovi CSc. za přečtení textu a hojné připomínky.

Jan Hendl

Omlouvám se za případné chyby a opomenutí v textu a uvítám návrhy čtenářů na jeho vylepšení.

Jan.Hendl@ruk.cuni.cz

1 Základy

Nejdříve probereme obsah pojmu Big data. Dále uvedeme přehled jednotlivých částí vědy o datech (data science). Vědu o datech považujeme za významnou, a tak aktuální, že si zaslouží pozornost všech zájemců o informatiku a statistickou analýzu dat. Do vědy o datech zahrnujeme oblasti, se kterými se dnes běžně setkáváme, jestliže uvažujeme rozmanité aplikace analýzy velkých množin dat, umělé inteligence, strojového učení, Business Intelligence a Data Engineering. Metody umělé inteligence mají zvláštní postavení v souvislosti s vědou o datech, někdy se dokonce probírají zcela odděleně. Strojové učení však propojuje umělou inteligenci s analýzou dat, protože strojové učení vychází z dat.

1.1 Big data

Současný svět se rychle mění a je prosycen daty (Press 2014, McKinsey 2018). Dříve změny nebyly tak rychlé. Pokud se například definoval způsob uložení dat, pak se používal mnoho let. K zpracování a organizování dat se používaly výhradně relační databáze s propracovanou strukturou. Nyní se nová data nedrží zavedených struktur. Data přicházejí v mnoha rozdílných formátech. Změna těchto formátů do jednotného formátu není přitom zcela žádoucí, což je způsobeno rozdílností aktuálních předmětných aplikací.

Svět je prostoupen mnoha typy možností získání dat, které sahají od sensorů až po mobilní telefony, do databází se dostávají informace o dopravě, pohybu lidí a podmínkách v jiných částech města. Tyto informace mohou být ve vizuální, zvukové nebo textové podobě. Internet věcí (IoT, Internet of things) je také datově bohatá novinka. Chytrá domácnost je místo, kde si různé domácí spotřebiče vyměňují informace. Takové aplikace vyžadují zpracování velkých objemů dat okamžitě v reálném čase.

Provedení obchodní transakce potřebuje software pro inteligentní rozhodování. Obchodní scénáře se mění velmi rychle vlivem elektronického přenosu dat. Jedna událost ve světě obchodu může generovat celou kaskádu událostí v jiných oblastech obchodu. Takové situace vyžadují rychlý sběr dat a jejich zpracování.

Propojení internetem vede k virtuální společnosti, v které jedinec na druhém konci světa má vlastně stejný profil jako kolega ve vedlejší kanceláři. Sociální média jako Twitter, Facebook, Instagram a další platformy poskytují každému jejich členu spojení nutné k výměně informací a interakci. Lidé spolu mohou na dálku mluvit o počasí, politice, výzkumu, rodině atd. Tyto

média jsou naneštěstí používány i pro nekalé aktivity. Každý okamžik si miliony lidí vyměňují ohromné množství informací.

Pokroky v medicíně vedou k tomu, že se uplatňuje personalizovaná medicína, upravená terapie pro daného jedince. To vyžaduje monitorování medicínských proměnných, jejichž hodnoty ovlivňují další průběh terapie. Taková data mají podobu rentgenových snímků, průběhu srdeční činnosti nebo teploty. Tato data se mohou dále využít na populační úrovni, v epidemiologii, ve výzkumu.

Výzkum v biologických vědách, epidemiologii a v dalších oblastech využívají data v daném okamžiku nebo shromážděná za delší dobu. Zpracování dat o novém viru a jeho šíření vyžaduje rychlou reakci. Data jsou v různých formátech. Vizualizace tvaru proteinů je důležitá pro biology, aby pochopili jejich vztah k životu.

Klasifikaci Big dat můžeme provést podle mnoha dimenzí. Klasifikace různých perspektiv pohledu na Big data bere v úvahu zdroje dat, formáty obsahu, uložení dat, předzpracování dat a způsob zpracování. Schéma možných perspektiv ukazuje následující seznam:

- Zdroje dat (internet, stroje, sensory, transakce, IoT).
- Formát (data strukturovaná, semistrukturovaná, nestrukturovaná).
- Uložení a organizování (orientace na dokumenty, orientace na sloupce, grafově založené, podle klíče).
- Předzpracování (čištění, normalizace, transformace).
- Zpracování (dávkové, v reálném čase).

Základní definiční charakteristiky Big dat jsou jejich velikost, formáty a změna.

Big data aplikace se týkají především zvládnutí velkých objemů dat, ale také mixu různých typů dat (různorodost dat) a toho, jakou roli při jejich vzniku hraje čas (rychlost změny). Tyto charakteristiky Big dat se obvykle v anglické literatuře nazývají 3V (Volume, Variety, Velocity). V tabulce 1.1 uvádíme jejich vymezení. Někdy se ještě přidávají další vlastnosti „V“ jako hodnověrnost (Veracity) a hodnota (Value).

Tabulka 1.1: Základní vlastnosti 3V Big dat: objem, různorodost, rychlost

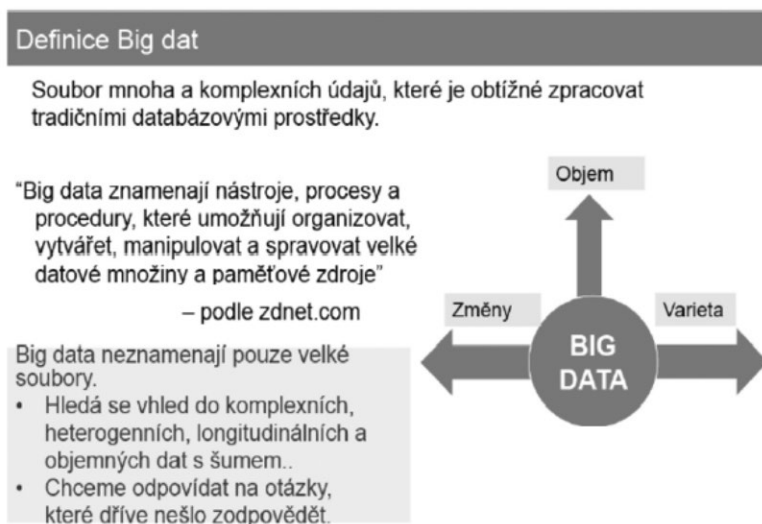
Aspekt	Charakteristika
Objem (Volume)	Hlavní aspekt, v posledních letech se nesmírně zvětšilo množství generovaných dat. Nepředstavuje však hlavní obtíže.
Různorodost (Variety)	Mnoho rozličných formátů dat, od strukturovaných dat po nestrukturovaná data.
Rychlost (Velocity)	Rychlost změny dat. Zvyšuje se množství dat, které se musí rychle uložit a hned zpracovat.

Objevují se různé definice pojmu Big data. Opírají se obvykle o uvedené tři vlastnosti. Jedna z přijímaných definic říká, že Big data jsou data s různorodým formátem, o velkém objemu a rychle se měnící, což v souhrnu způsobuje, že je nelze spravovat pomocí konvenčních datábázových prostředků. Vazbu mezi vlastnostmi Big dat a technologickými změnami ukazuje tabulka 1.2.

Tabulka 1.2: Technologie a Big data

Aspekt	Možnosti a technologie
Objem	Virtualizace ukládání do cloudů v datových centrech, množství prvků připojených k internetu.
Různorodost	Existuje potřeba analyzovat i nestrukturovaná dat. Uplatnění databází NoSQL.
Rychlost	Miliony připojených chytrých telefonů a senzorů zvyšují objem i rychlost změn.

Obrázek 1.1 názorně přibližuje definici pojmu Big data.



Obrázek 1.1: Vymezení pojmu Big data

Jednotky pro vyjádření objemů informací jsou v tabulce 1.3, přičemž o Big datech mluvíme od velikostí mnoha gigabytů.

Tabulka 1.3: Přehled násobných jednotek paměti (v nové úpravě)

Jednotka	Značka	B	kB	KiB	MB	MiB	GB	GiB	TB	TiB
Kilobajt	kB	1000	1	~0,9766						
Kibibajt	KiB	1024	1,024	1						
Megabajt	MB	1 000 000	1000	~976,6	1	~0,9537				
Mebibajt	MiB	1 048 576	~1048,6	1024	1,049	1				
Gigabajt	GB	10 ⁹	1 000 000	976 562,5	1000	953,7	1	~0,9313		
Gibibajt	GiB	~1,074×10 ⁹	~1 073 742	1 048 576	~1073,7	1024	1,074	1		
Terabajt	TB	10 ¹²	10 ⁹	~0,9766×10 ⁹	1 000 000	~953 574,3	1000	931,3	1	~0,9095
Tebibajt	TiB	~1,1×10 ¹²	~1,1×10 ⁹	~1,074×10 ⁹	~1 099 512	1 048 576	~1099,5	1024	~1,1	1

V prosinci 1998 IEC (úřad pro normy Evropské unie v oblasti elektrotechniky) vytvořil dodatek k normě IEC 60027-2, ve kterém zavedl pro počítačové jednotky nový systém označování násobků. V tomto systému bylo pro původní „velké kilo“ = 1024 B navrženo označení kibibajt a značka KiB, zatímco jednotka kilobajt (se značkou kB) označuje 1000 B, tak, jak je obvyklé

v soustavě SI. Nové binární předpony (kibi-, mebi-, gibi-, ...) jsou definované také v normě ISO/IEC 80000.

V přehledu ještě uvádíme jednotky objemu informací, které se používají v oblasti Big data:

Kilo- znamená 1,000 bytů; kilobyte je tisíc bytů.

Mega- znamená 1,000,000 bytů; megabyte je milion bytů.

Giga- znamená 1,000,000,000 bytů; gigabyte je miliarda bytů.

Tera- znamená 1,000,000,000,000 bytů; terabyte je trilion bytů.

Peta- znamená 1,000,000,000,000,000 bytů; petabyte je 1,000 terabytů.

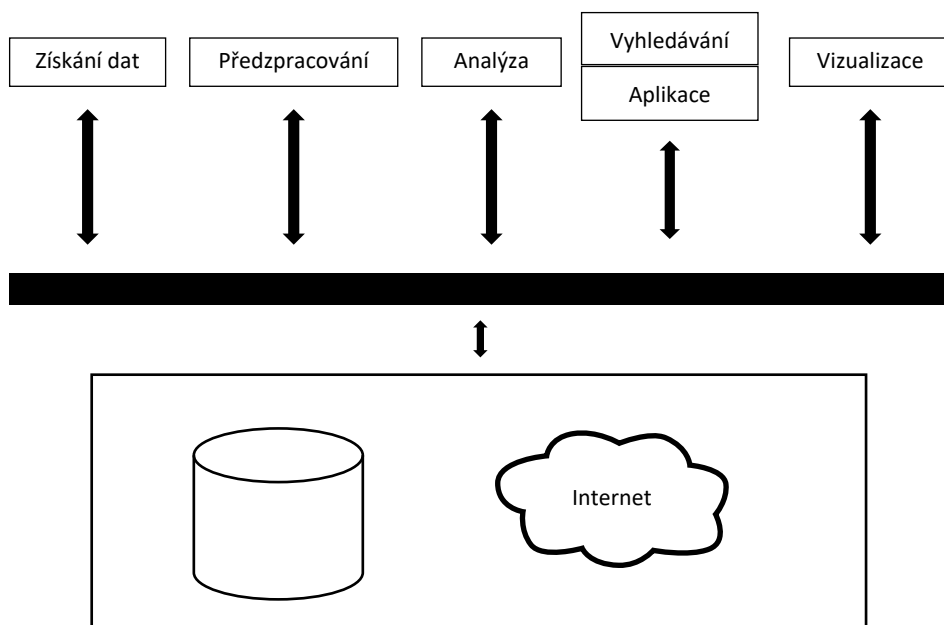
Exa- znamená 1,000,000,000,000,000,000 bytů; exabyte je 1,000 petabytů.

Zetta- znamená 1,000,000,000,000,000,000,000 bytů; zettabyte je 1,000 exabytů.

Yotta- znamená 1,000,000,000,000,000,000,000,000 bytů; yottabyte je 1,000 zettabytů.

Objem v současnosti vznikajících dat je ilustrován tvrzením, že mezi vznikem civilizace a rokem 2003 bylo vytvořeno 5 exabytů informací, nyní tolik vytvoříme každé dva dny (Kitchin 2014).

Základní schéma procesu zpracování Big dat obsahuje několik kroků: získání/hledání dat, předzpracování dat, analýza dat a zpracování dat, aplikace, vizualizace dat. Obrázek 1.2 s rámcem zpracování Big dat zjednodušeně ukazuje fungování celého systému.

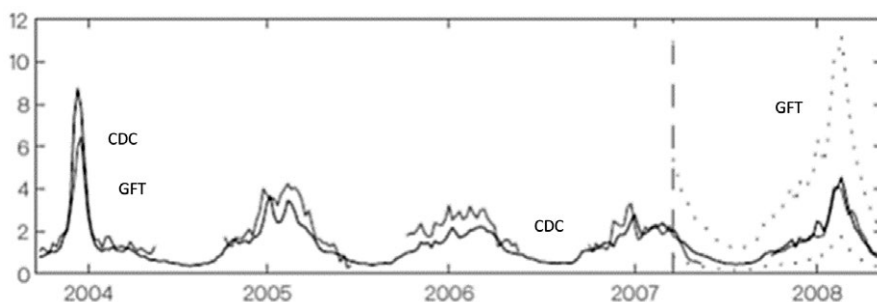


Obrázek 1.2: Zpracování Big dat

Zdroje Big dat jsme zmínili, předzpracování zahrnuje odstranění nepožadovaných nebo duplicitních dat, v procesu předzpracování je v relačních databázích důležitá normalizace k zabránění redundance dále jde o čištění a předběžnou úpravu dat. Uživatelé z různých oblastí se snaží tato data využít pomocí různých druhů analytik. Analytické techniky zahrnují strojové učení, data mining, statistické metody využití paralelních algoritmů pro rychlé zpracování a uplatnění modelu při optimalizaci rozhodování. Vizualizace představuje důležitý krok, protože výsledky předchozího zpracování je často výhodné zprostředkovat ve vhodné obrazové podobě.

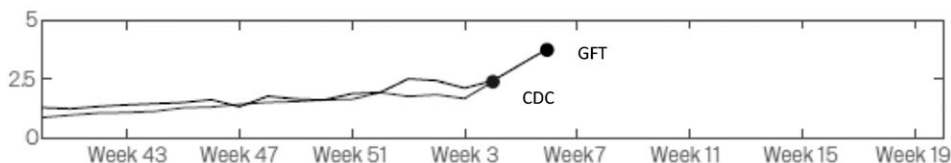
Uvedeme nyní jeden klasický příklad. Na možnosti uplatnění analýzy Big dat upozornil zdravotnickou veřejnost a zástupce politického života projekt internetové společnosti Google v souvislosti s možnostmi monitorování výskytu chřipkového onemocnění. Epidemie sezónní chřipky jsou velkým problémem systému veřejného zdravotnictví. Včasná detekce aktivity onemocnění vede k rychlejší reakci úřadů a může redukovat dopad jak sezónní, tak pandemické chřipky. Jeden ze způsobů včasné detekce představuje monitorování chování uživatelů při vyhledávání informací na internetu. Google společnost navrhla metodu analýzy velkého počtu Google dotazů s cílem monitorovat nemoci podobné chřipce. Některé dotazy jsou silně korelované s pravděpodobností návštěvy lékaře, při kterých pacient prezentuje symptomy chřipky. To umožnilo navrhnout algoritmus pro odhad šíření chřipky v jednotlivých oblastech USA.

Příslušná služba se označuje zkratkou GFT (Google Flu Trends). GFT porovnává své predikce s historickou základní úrovní chřipkové aktivity pro danou oblast a pak rozhoduje, zda se jedná u aktuální aktivity o minimální, nízkou, střední, vysokou nebo intenzivní aktivitu. Získané odhady podle firmy Google velmi dobře korelují s konvenčními epidemiologickými daty (CDC, Centers for Disease Control and Prevention), jak na národní, tak na oblastní úrovni.



Obrázek 1.3: GFT modelová data a CDC data o chřipkové aktivitě

Obrázek 1.3 ukazuje GFT data v oblasti (Středoatlantská oblast v USA) a CDC data s dosaženou korelací 0,96. Také je uveden 95% pás predikce (přerušované křivky).



Obrázek 1.4: Podrobné srovnání predikcí chřipkové aktivity v letech 2007–2008

Obrázek 1.4 ukazuje podrobné srovnání predikcí chřipkové aktivity v letech 2007–2008. V 5. týdnu byl detekován strmý nárůst signálu, který byl později potvrzen daty z CDC institutu.

Prvotní motivace pro GFT spočívala v úsilí o včasnou identifikaci aktivity nemoci a rychlou reakci. V jedné zprávě se dokazovalo, že GFT signál detekoval zvýšený výskyt chřipky až o 10 dní dříve než tuto skutečnost ohlásila služba CDC.

Google signál GFT byl podle společnosti Google příkladem „kolektivní inteligence“, které je možné využít k identifikaci trendů a k výpočtu predikcí. Chování lidí na internetu ukazuje jejich vůli a potřebu bez omezení.

Poznamenejme, že později po podrobném přezkoumání jinými vědci byly optimistické názory společnosti Google na užitečnost signalizace epidemie zvolenou metodou značně korigovány (viz např. Google Flu 2013).

Dále seznámíme čtenáře se dvěma odlišnými příklady využití Big dat. První příklad má vztah ke světu sportu, druhý se týká výzkumu kultury.

V roce 2011 získal prestižní cenu film Moneyball s Bradem Pittem v hlavní roli. Film byl natočen podle románové předlohy Michaela Lewise z roku 2003. Román i film jsou historií hlavního manažera baseballového týmu Oakland Athletics (Kalifornie) Billyho Beanea (B. B.), který využil statistickou analýzu dat k tomu, aby zjistil, které dovednosti hráče přispívají významně k výhře v zápasu. Jeho přístup revolucionizoval způsob, jak jsou hodnoceni hráči. V soutěžích baseballu jde totiž o velké peníze. Nový manažér (B. B.) měl tým vytáhnout z bídy, protože prohrával jedno utkání za utkáním. Uplatněním metod data science získal tipy na nové (nedocenené) hráče a s poměrně malým rozpočtem (40 milionů dolarů) sestavil tým, který běžně vyhrával s týmy s mnohem větším rozpočtem (100+ milionů dolarů). Beane analyzoval záznamy o stovkách hráčů a identifikoval statistiky, které byly vysoce prediktivní pro úspěch hráče ve hře. Tyto statistiky se lišily od čísel, které využívali tradičně scouti v té době. Věřil své teorii a nápadům i přes nesouhlasný postoj zaměstnavatele. Jeho metody se rychle rozšířily v ekonomice a obchodu. Úspěch B. B. rezonuje u datových výzkumníků. Vypráví o výhodách přístupu, jestliže data science se stane částí „DNA“ organizace. Osvětluje, jak velké myšlenky z Big dat se mohou promítnout do obchodního zisku.



Baseball byl vždy také hrou čísel a statistiky. Díky explozi dat v současné době a existenci softwaru a počítačů bylo možné využít obecné myšlenky regresní analýzy a datové analytiky. Význam Moneyballu spočíval ve faktu, že poprvé informoval veřejnost, jak je možné využít Big data ve sportu. Platí to jistě pro americké sporty jako košíková, kopaná, kriket a hokej. V současné době má většina sportovních jednot v USA sekci datových výzkumníků. Získané modely pomáhají manažérům přesněji evaluovat hodnotu hráče a minimalizovat riziko, které vzniká rozhodováním pomocí intuice. Sportovní analytika má budoucnost a stále hledá nové cesty (B. J. Alamar: Sports analytics 2013).

Příklad analýzy velkých souborů dat v rámci historického výzkumu je popsán na Wikipedii. Týká se dokumentů shromážděných v rámci projektu Seshat: History Databank nekomerční

organizací Evolution Institute. V databázi jsou data (miliony starobylých textů), které pomáhají testovat různé historické hypotézy. V tomto případě šlo o téze německo-švýcarského filosofa K. Jasperse týkající se Axiální doby (800–1000 let před Kristem). Ty předpokládají, že v Axiální (Osové) době došlo ve velkých kulturách Číny, Řecka, Indie, Iránu a Izraele k vývoji, který položil základy modernity. Teze tvrdí, že velké civilizace přijaly v té době morální universalismus a objevily transcendenci. Lidé v nich začali jinak přemýšlet o světě: ne že funguje, ale proč funguje. Vzniklo myšlení druhého řádu o tom, jak lidé uvažují o universu a osudu člověka. Dokonce se spekuluje o tom, zda se dnes nenalézáme v nové Axiální době.

Koncept Axiální doby je známý a vlivný, existují však pochybnosti, protože schází historická a sociologická přesvědčivá evidence. Koncept je stále kontroverzní. Seshat umožnil zkoumat toto téma. Výsledky analýzy Big dat podkopaly tezi o Axiální době jako celkem jasně časově a geograficky lokalizovaném procesu. Autoři (Mullins, Hoyer et al 2018) prokazují, že podobný kulturní pohyb se odehrával na různých místech stovky i tisíce let předtím. Tito autoři ukazují cestu, jak lze postupovat při historicko-sociologickém výzkumu ideologických a institucionálních základů společností. Výsledky studie popsali v knize o 500 stránkách s názvem Seshat History of the Axial Age. Kniha se zaměřuje na historický výzkum pomocí Big dat, jeho výsledky a použitou metodologii. Data umožnily porovnání vývoje společností vzdálených od sebe v čase i prostoru analytickými metodami určenými pro Big data. Široce diskutované výsledky studie znamenají významný doplněk izolovaného historického výzkumu.

1.2 Věda o datech

V roce 1962 John Tukey (Brillinger 2002) popsal oblast, kterou nazval „analýza dat“. Jeho popis se podobal moderní vědě o datech. V roce 1985 použil C. F. Jeff Wu spojení „Data Science“ jako alternativní název pro statistiku (Press 2013). V roce 1996 se stala International Federation of Classification Societies první konferencí, která tematizovala úlohy vědy o datech. Definice se však teprve tvořila. V roce 1997 opět C. F. Jeff Wu navrhl, že věda o datech přispěje k opravě nepřesného stereotypu, který ztotožňuje statistiku s evidováním a pouhým popisem dat. V roce 1998 Chikio Hayashi doporučil vědu o datech jako nový interdisciplinární koncept se třemi aspekty: způsob návrhu, získání a analýzy dat. V 90. letech se věda o datech začala spojovat s hledáním zajímavých konfigurací v datech, což zahrnovalo „data mining“ a „objevování znalostí“ (knowledge discovery). Vymezení moderní koncepce vědy o datech jako nové disciplíny se připisuje W. S. Clevelandovi. V roce 2001 doporučil, že aplikovanou statistiku je nutné propojit s technologickou oblastí (IT). V roce 2003 byl založen The Journal of Data Science.

Vědu o datech (data science) pojímáme jako interdisciplinární oblast, která využívá vědecké metody, procesy a algoritmy k extrahování znalostí a získání vhledu do daného problému pomocí mnoha strukturovaných a nestrukturovaných dat. Věda o datech je v úzkém vztahu k dolování dat (data mining), k strojovému učení (ML, Machine Learning) a k velkým objemům dat (Big data). Pracuje se však také s „malými daty“, tedy daty, které pocházejí ze standardních šetření a experimentů (tab. 1.4).

Věda o datech v praktických aplikacích znamená systematický a strategický přístup k využití dat pro řešení určitých problémů. Problém, který se řeší, je praktického charakteru. Nejde tedy o ryze akademické řešení. Datový vědec v instituci se tolik nezajímá o zobecňující řešení problému, spíše usiluje o zlepšení toho, jak instituce dosahuje svých cílů.

Tabulka 1.4: Rozdíly mezi „malými“ daty a Big daty

Vlastnost	Malá data	Big data
Cíle	Zodpovězení specifické otázky	Cíle a otázky se pružně mění
Umístění	V jedné instituci, obvykle na jednom počítači	V elektronickém prostoru, na mnoha počítačích
Struktura, obsah	Strukturovaná data, tabulka, z jedné oblasti	Strukturovaná i nestrukturovaná data (dokumenty, obrázky, nahrávky, čísla), mnoho disciplín
Příprava	Uživatel připraví data pro vlastní jedno použití	Různé zdroje, odlišné skupiny lidí data získávají, připravují a analyzují
Trvání	S koncem projektu data po krátké době mizí	Data přetrvávají dlouhou dobu, přemísťují se na jiná místa
Opakovatelnost	Projekty jsou reprodukovatelné, pracuje se s novými daty	Není obvykle možné zopakovat sběr dat, doufáme, že data jsou v pořádku
Analýza	Všechna data se podrobí analýze, obvykle pouze jednou	Nevíme, která data využijeme, data se extrahují, normalizují, transformují, vizualizují, interpretují a znovu analyzují

Věda o datech znamená prolnutí navzájem se doplňujících disciplín, lidé orientovaní v této oblasti jsou datoví výzkumníci, „data scientists“. Takto označení odborníci by měli zvládnout široké pole znalostí, které jsou vystaveny dynamickým změnám, protože svět kolem „data science“ se rychle mění. Daný problém z oblasti data science se obvykle zpracovává v týmu, pro jedince je totiž obtížné spojit všechny dovednostní požadavky na dostatečné úrovni. Tým se vyznačuje různorodostí akademického původu svých členů. Skládá se z matematiků, fyziků a přírodovědců, doktorů, inženýrů a informatiků. V závislosti na charakteru úlohy v něm také nalézají uplatnění ekonomové, sociologové, psychologové a zástupci humanitních oborů. Předpokládá se však, že každý má základní znalosti metodiky analýzy úlohy a matematiky.

Věda o datech se realizuje v praxi lidmi, kteří se zajímají o otázky jiných profesí a o jiné oblasti znalostí. Mezi tyto oblasti patří informatika, statistické metody a metody analýzy dat, obznanenost s aplikační oblastí:

Informatika: Základ informatiky zahrnuje umění získávat a organizovat data z různých zdrojů (např. internet, lékařské přístroje). Jestliže chceme analyzovat data, měli bychom znát techniky ukládání dat a způsoby manipulace s nimi. Strukturovaná data jsou obvykle uložena v relačních databázích. Proto je nutná znalost SQL jazyka. Jiná data mají nestrukturovanou podobu, příkladem jsou textové dokumenty nebo produkty zobrazovací techniky. Taková data jsou uložena ve speciálních databázích s jinou logikou.

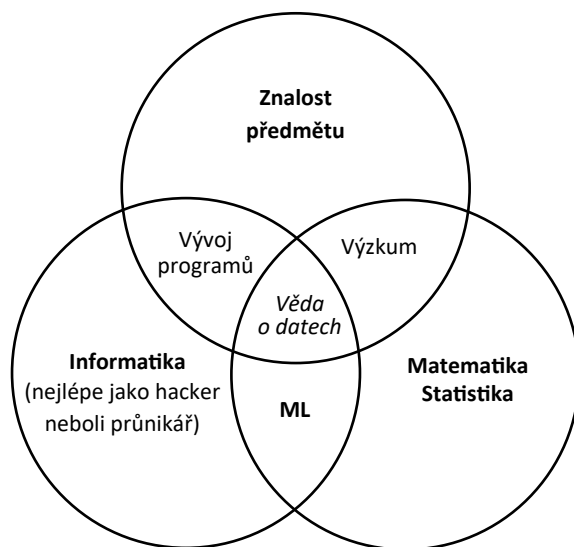
Data musíme z databází nějakým způsobem získat, takže potřebujeme ovládat rozhraní pro ovládání takových databází, kde hrají velkou roli například práva k přístupu, síťové protokoly a problémy kódování. Z informatiky jsou vhodné základní znalosti algoritmizace pomocí programovacích jazyků (Java, Python, R, Julia a další). Platí, že programovací jazyk musí být vhodný pro danou úlohu. Vyžaduje obvykle aplikaci knihoven podprogramů a dalších užitečných programovacích prostředků.

Statistické metody a metody analýzy dat: Analýza dat využívá matematické a statistické dovednosti ke konverzi dat do návodů k jednání. Děje se tak pomocí matematických modelů. Matematika se skládá z teoretické matematiky a aplikované matematiky, která nám pomáhá řešit nějaký reálný problém. Rozsah aplikované matematiky je široký a věda o datech se vyznačuje oportunismem při volbě přístupů. Mnoho příslušných prostředků a programových

knihoven je volně dostupných na internetu. Kdo však chce provést regresní analýzu nebo řešit nějaký klasifikační problém pomocí neuronových sítí, shlukovacích metod nebo jiné techniky, měl by znát základy matematiky (maticového počtu, pravděpodobnosti a konečné matematiky), metod statistiky a strojového učení (což zahrnuje i hluboké učení), aby se předešlo jejich špatnému použití. Metody statistiky se v této oblasti opírají o mnohorozměrné přístupy, které využívají data o mnoha proměnných. S obecně platnými základními koncepty statistiky se již setkáváme v jednorozměrné a dvourozměrné statistice. K podstatným odchylkám v mnohorozměrné statistice nedochází. Přibývají však nové koncepty svázané specificky s vědou o datech. Kreativní aplikace komplexní matematiky je základní vlastností vědy o datech.

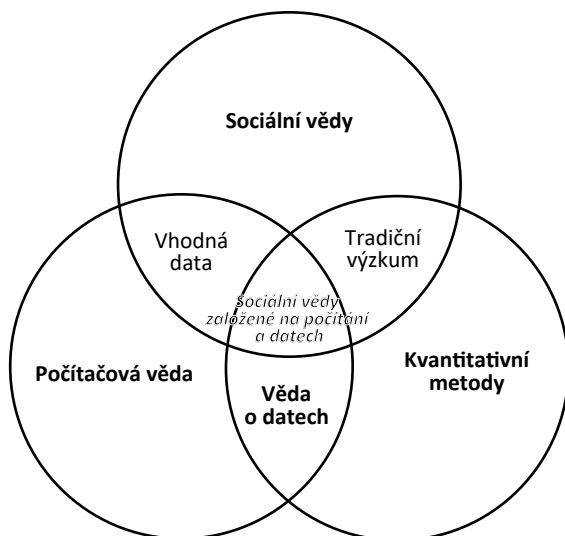
Znalosti z oblasti aplikace: Expertní znalosti neboli doménové znalosti jsou podstatnou částí aplikované vědy o datech. Aby členové týmu reagovali adekvátně na požadavky praxe, musí být obeznámeni s jejími základy a příslušnými odbornými termíny, což zaručuje schopnost dorozumění. Každý, kdo analyzuje problém, musí znát kontext problému a možné požadavky na řešení. Uvažují se oblasti aplikací, které jsou v tomto smyslu dosti odlišné. Například obchod se liší od medicínského prostředí, dopravy nebo sportu. Datový výzkumník se musí připravit na to, že bude řešit problémy v nejrůznějších oblastech. Manažeři v těchto oblastech se rozhodují v prostředí, které se mění a vyznačuje se velkou nejistotou. Zkušenosti a intuice v rozhodování hrají důležitou roli, ale mozková kapacita lidí je omezená. Rozhodování může být ovlivněno mnoha faktory, které nemáme pod kontrolou. V takových okamžicích jsou důležitá dostupná data a počítač. Teoretické znalosti aplikované matematiky (třeba strojového učení) a informatiky mohou být značné, ale bez znalosti reality mohou vyvinutá řešení způsobit více škody než užitku. Věda o datech je o věcech a lidech v reálném světě, ne o číslech a algoritmech.

Oblast data science znázorňuje schematicky známý Vennův graf teoretika Drew Conwaye (obr. 1.5, 2010). Tento jednoduchý graf pomáhá pochopit podstatu a rozsah znalostí vědy o datech.



Obrázek 1.5: Vennův graf disciplín vědy o datech (data science)

Conway (<http://drewconway.com/>) definuje tři kompetence, které by měl datový výzkumník nebo tým orientovaný na vědu o datech zvládnout. Graf na obrázku 1.5 pojímá vědu o da-



Obrázek 1.6: Vennův graf dovedností pro vědu o datech

tech jako interdisciplinární aktivitu, která má tři dimenze: znalost oblasti aplikace, matematiku a statistiku, informatiku. Datový výzkumník je někdo, kdo rozumí problémům aplikační oblasti, přičemž se snaží využívat matematické a statistické prostředky. Také rozumí informatice a své představy o řešení problému dokáže vyjádřit počítačovým kódem.

Může se namítnout, že graf na obrázku 1.5 neobsahuje měkké dovednosti (umění komunikace, týmová spolupráce atd.). Tyto dovednosti se předpokládají automaticky jako všude jinde u pracovníků v týmu.

Také je nutné poznamenat, že jedinci zvládající všechny zmíněné kompetence do hloubky jsou mýtickými postavami, které neexistují ve skutečném světě. Většina datových vědců začíná s kvalifikací matematika nebo informatika a postupně se dovzdělává. V oblasti data science pracuje mnoho jedinců, kteří mají společenskovědní nebo humanitní vzdělání. Pro pracovníky v oblasti sociálních věd vytvořil obdobné schéma zkušeností a znalostí Ulrich Matter (obr. 1.6, podle KDnuggets-David Taylor).

Tabulka 1.5: Příklady úloh a odpovídající základní zkušenost

Úloha	Primární zkušenosti
Příprava serveru pro uložení velkých dat a sběr v reálném čase a z mnoha různých míst	INFORMATIKA
Vysvětlení přítomnosti vychýlených hodnot	ORIENTO VANOST
Vysvětlení, do jaké míry lze závěry z analýzy zobecnit	STATISTIKA
Rozhodnutí, jaká je vhodná vizualizace pro laiky	STATISTIKA
Rozhodnutí, jaká data se mají propojit	ORIENTO VANOST
Automatizace spojování dat	INFORMATIKA
Detekce signálu v datech	STATISTIKA
Redukce počtu proměnných	STATISTIKA
Příprava dlouhodobého ukládání dat (organizace verzí dat)	INFORMATIKA