
edice
PRŮVODCE
GRADA
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 zvládání pandemie Covid-19


GRADA

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 RÁJI, s.r.o., Pardubice

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

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

Obsah

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 oblasti 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, revolucionalizovala 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 manažéři, 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. Mařík 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). Znamenitá 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 pročítáním 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 rozdílnostmi 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 zvládání 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 databázových 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	1021	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

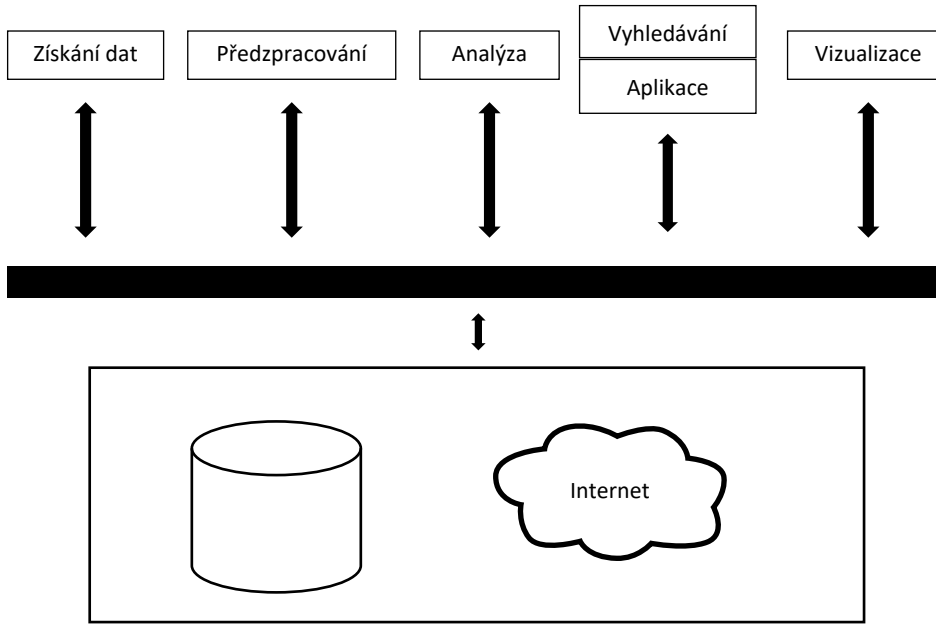
v soustavě SI. Nové binární předpony (kibi-, mebi-, gibi-, …) jsou definované také v normě ISO/IEC 80000.

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.

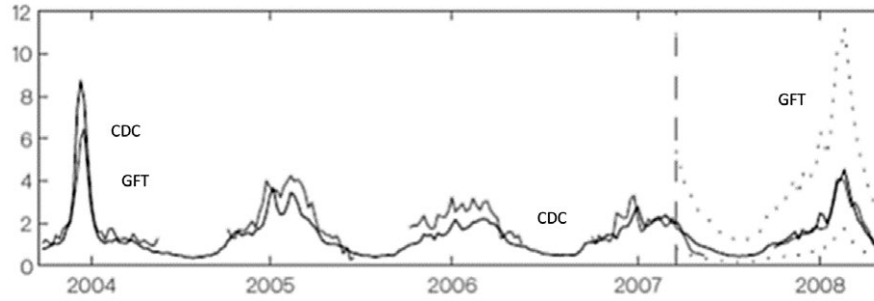


Obrázek 1.2: Zpracování Big dat

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.

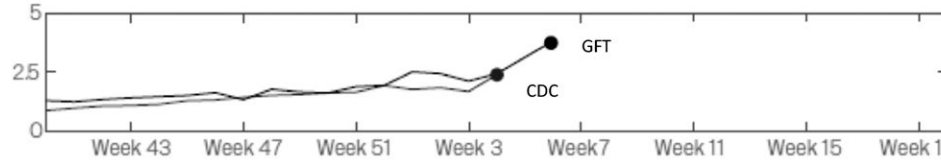
Uvedeme nyní jeden klasický příklad. Na možnosti uplatnění analýzy Big dat upozornil zdravotnickou veřejnost a zástupce politického života projekt internetové společnosti Google v souvislosti s možnostmi monitorování výskytu chřipkového onemocnění. Epidemie sezonní chřipky jsou velkým problémem systému 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 (Středoantlantská 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 potřeby 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 manažéra 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 revolucionalizoval 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 (nedoceněné) 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