BIG DATA

VIKTOR MAYER-SCHÖNBERGER, KENNETH CUKIER

REVOLUCE, KTERÁ ZMĚNÍ ZPŮSOB, JAK ŽIJEME, PRACUJEME A MYSLÍME

computer press

Viktor Mayer-Schönberger
Kenneth Cukier

Big Data

Computer Press
Brno
2014

Big Data

Viktor Mayer-Schönberger, Kenneth Cukier

Překlad: Jakub Goner
Obálka: Martin Sodomka
Odpovědný redaktor: Martin Herodek
Technický redaktor: Jiří Matoušek

Objednávky knih:
http://knihy.cpress.cz
www.albatrosmedia.cz
eshop@albatrosmedia.cz
bezplatná linka 800 555 513
ISBN 978-80-251-4119-9

Vydalo nakladatelství Computer Press v Brně roku 2014 ve společnosti Albatros Media a. s. se sídlem Na Pankráci 30, Praha 4. Číslo publikace 18 403.

1. vydání

ALBATROS MEDIA a.s.

Obsah

Poděkování

Oba jsme měli velké štěstí, že jsme mohli spolupracovat s Lewisem M. Branscombem, jedním z prvních gigantů na poli informačních sítí a inovací, a učit se od něj. Svým intelektem, výřečností, energií, profesionalitou, vtipem a charakteristickou zvídavostí nás stále inspiruje. Jeho sympatické a moudré partnerce Connie Mullin se omlouváme, že jsme se neřídili jejím doporučením, abychom tuto knihu nazvali „Superdata".

Momin Malik byl díky svému výjimečnému intelektu a pracovitosti vynikajícím výzkumným asistentem. Bylo nám ctí, že nás zastupovali Lisa Adams a David Miller z agentury Garamond, kteří byli v každém ohledu prostě skvělí. Měli jsme také mimořádného redaktora Eamona Dolana. Patří ke vzácným stylistům, kteří mají téměř dokonalý smysl pro úpravy textu a zpřesnění autorských myšlenek, takže je výsledek mnohem lepší, než jsme mohli doufat. Naše díky si zaslouží všichni pracovníci nakladatelství Houghton Mifflin Harcourt, zejména Beth Burleigh Fuller a Ben Hyman. Děkujeme také Camille Smith za její odborné korektury. Jsme vděčni Jamesi Franshamovi z časopisu *The Economist* za jeho perfektní kontrolu uváděných faktů a bystrou kritiku rukopisu.

Děkujeme zejména všem praktikům na poli veledat, kteří věnovali čas tomu, aby nám vysvětlili svou práci. Byli to zejména Oren Etzioni, Cynthia Rudin, Carolyn McGregor a Mike Flowers.

• • •

Poděkování speciálně od Viktora: děkuji Philipu Evansovi, který vždy uvažuje dva kroky dopředu a vyjadřuje své myšlenky s přesností a výmluvností, za naše konverzace v průběhu více než deseti let.

Jsem vděčný také svému bývalému kolegovi Davidu Lazerovi, který patřil mezi první a významné akademické pracovníky v oboru veledat a s nímž jsem se mnohokrát radil.

Děkuji účastníkům Oxfordského dialogu o digitálních datech z roku 2011 (zaměřeného na veledata) a zejména jeho spolupředsedovi Fredu Catemu za mimořádně cenné diskuse.

Oxford Internet Institute, kde pracuji, mi nabídl dokonalé prostředí pro tvorbu této knihy, protože mnoho mých kolegů se podílí na výzkumu veledat. Nedokážu si představit lepší místo, kde bych mohl svou knihu napsat. S vděčností vzpomínám také na podporu Keble College, kde vyučuji. Bez této pomoci bych nezískal přístup k některým důležitým primárním zdrojům.

Na psaní knihy vždy nejvíce doplácí autorova rodina. Kromě mnoha hodin strávených před počítačovým monitorem v kanceláři se své ženě Birgit a malému Viktorovi musím omluvit také za mnoho dalších hodin, kdy jsem byl sice fyzicky přítomen, ale přemýšlel jsem o vlastních věcech. Slibuji, že se zkusím polepšit.

Poděkování speciálně od Kenna: za pomoc vděčím mnoha skvělým veledatovým odborníkům. Mimo jiné k nim patří Jeff Hammerbacher, Amr Awadallah, DJ Patil, Michael Driscoll, Michael Freed a mnoho dalších lidí, které jsem během let poznal ve společnosti Google (včetně Hala Variana, Jeremyho Ginsberga, Petera Norviga a Udi Manbera, neocenitelné byly také krátké rozhovory s Ericem Schmidtem a Larrym Pagem).

Můj rozhled obohatil Tim O'Reilly, opravdový vizionář internetového věku. Hodně mě naučil také Marc Benioff z firmy *Salesforce.com*. S neocenitelnými postřehy jako vždy přicházel Matthew Hindman. Zásadně mi pomohl James Guszcza ze společnosti Deloitte a také Geoff Hyatt, starý přítel a zkušený datový podnikatel. Zvláštní poděkování věnuji Pete Wardenovi, který je zároveň filozofem a veledatovým praktikem.

Své myšlenky a rady mi nabídlo mnoho přátel, k nimž patří John Turner, Angelika Wolf, Niko Waesche, Katia Verresen, David Wishart, Anna Petherick, Blaine Harden a Jessica Kowal. Jako zdroj inspirace k tématům této knihy musím uvést následující jména: Blaise Aguera y Arcas, Eric Horvitz, David Auerbach, Gil Elbaz, Tyler Bell, Andrew Wyckoff a mnoho dalších

v organizaci OECD, Stephen Brobst a tým společnosti Teradata, Anthony Goldbloom a Jeremy Howard z firmy Kaggle, Edd Dumbill, Roger Magoulas s týmem nakladatelství O'Reilly Media a Edward Lazowska. K významným osobnostem se řadí James Cortada. Díky si zaslouží také Ping Li ze společnosti Accel Partners a Roger Ehrenberg z IA Ventures.

Kolegové z časopisu *The Economist* mě zahrnuli úžasnými nápady a podporou. Vyzvednout musím zejména své redaktory Toma Standageho, Daniela Franklina, Johna Micklethwaita a Barbaru Beck, protože pracovali na zvláštním čísle „Data, Data Everywhere" (Všudypřítomná data), které posloužilo jako základ této knihy. Mým vzorem jsou tokijští kolegové Henry Tricks a Dominic Zeigler, protože neustále hledají nová témata a dokážou je krásně vyjádřit. Oliver Morton mi poskytoval svou typickou moudrost tehdy, když jsem ji nejvíce potřeboval.

Salzburský globální seminář v Rakousku nabídl dokonalou kombinaci idylické relaxace a intelektuálních podnětů, díky nimž jsem mohl snáze psát a přemýšlet. U kulatého stolu, který Aspen Institute pořádal v červenci 2011, zaznělo mnoho myšlenek, za něž děkuji účastníkům i organizátorovi Charliemu Firestonemu. Oceňuji také Teri Elniski za její ohromnou podporu.

Frances Cairncross, rektorka Exeter College v Oxfordu, mi zajistila poklidné místo, kde jsem mohl pobývat a čerpat tam povzbuzení. Naplňuje mě pokorou, že ve svých úvahách nad otázkami technologií a společnosti navazuji na její vlastní myšlenky, které formulovala již před patnácti lety ve své knize *The Death of Distance* (Smrt vzdálenosti). Když jsem byl mladým novinářem, tato práce mě velmi inspirovala. Každé ráno jsem si při chůzi přes univerzitní nádvoří s hrdostí uvědomoval, že smím nést pochodeň, kterou zažehla, i když v jejích rukou oheň zářil mnohem jasněji.

Svou nejhlubší úctu chci vyjádřit vlastní rodině, která to se mnou vydržela – přesněji řečeno snášela mou častou nepřítomnost. Díky si zaslouží moji rodiče, sestra a další příbuzní, ale největší porci svého vděku věnuji své ženě Heather a našim dětem Charlotte a Kazovi. Bez vaší podpory, povzbuzování a nápadů by tato kniha nikdy nemohla vzniknout.

Oba autoři jsou zavázáni mnoha lidem, se kterými o tématu veledat diskutovali, často i dlouho před tím, než se tento termín vůbec rozšířil. V tomto ohledu vyhrazujeme speciální poděkování účastníkům různých ročníků Rueschlikonské konference o informačních zásadách, na jejíž organizaci se Viktor podílel a kde Kenn působil jako zpravodaj. Adresáty našich díků jsou zejména Joseph Alhadeff, Bernard Benhamou, John Seely Brown, Herbert Burkert (který nás upozornil na komodora Mauryho), Peter Cullen, Ed Felten, Urs Gasser, Joi Ito, Jeff Jonas, Nicklas Lundblad, Douglas Merrill, Rick Murray, Cory Ondrejka a Paul Schwartz.

Viktor Mayer-Schönberger
Kenneth Cukier
Oxford/Londýn, srpen 2012

Kapitola 1

Současnost

V roce 2009 se objevil nový virus chřipky. Tento nový kmen označený H1N1, který kombinoval prvky virů způsobujících ptačí a prasečí chřipku, se rychle šířil. Po několika týdnech se pracovníci hygienických služeb po celém světě začali obávat příchodu nebezpečné pandemie. Někteří komentátoři varovali, že epidemie může nabýt rozsahu španělské chřipky z roku 1918, kdy se nakazilo půl miliardy lidí a zemřely desítky milionů. Proti novému viru navíc nebyla rychle dostupná žádná vakcína. Zdravotníci mohli jen doufat, že se jim postup nákazy podaří zpomalit. K tomu by však potřebovali vědět, kde už se nemoc vyskytla.

Agentura CDC (Centers for Disease Control and Prevention), která je součástí amerického ministerstva zdravotnictví, požádala lékaře, aby jí poskytovali informace o nových případech chřipky. Obrázek o pandemii, který mohla agentura vytvořit z těchto hlášení, byl však vždy o jeden či dva týdny zpožděný. Lidé se mohou obrátit na lékaře až po několika dnech potíží. Zpětný přenos informací nějaký čas trval a agentura CDC čísla vyhodnocovala jen jednou týdně. U rychle se šířící nemoci je dvoutýdenní zpoždění úplná věčnost. Kvůli neaktuálním údajům byla agentura v nejkritičtějších dnech úplně paralyzovaná.

Několik dní před tím, než se virus H1N1 dostal do novinových titulků, náhodou odborníci z internetového giganta Google publikovali ve vědeckém časopise *Nature* pozoruhodný článek. Většinu čtenářů sice příliš nezaujal, ale mezi pracovníky hygienických služeb a informatiky vyvolal značnou odezvu. Autoři vysvětlili, jak může společnost Google „předpovídat“ šíření zimní chřipky v USA, a to nikoli pouze na celostátní úrovni, ale i přesněji na úrovni oblastí, a dokonce jednotlivých států. Společnost to dokázala sledováním toho, co lidé vyhledávají v Internetu. Díky tomu, že Google zpracuje každý den více než tři miliardy vyhledávacích dotazů a všechny je ukládá, má ke zpracování spoustu dat.

Společnost shromáždila 50 milionů termínů, které Američané nejčastěji vyhledávají, a porovnala tento seznam s daty agentury CDC o šíření sezónní chřipky mezi lety 2003 a 2008. Cílem bylo identifikovat oblasti zasažené virem chřipky podle slov zadávaných do internetového vyhledávače. O podobnou analýzu internetových vyhledávacích termínů se pokusili i jiní, ale nikdo neměl k dispozici tolik dat, výpočetního výkonu ani statistických znalostí jako společnost Google.

Inženýři sice tipovali, že taková vyhledávání mají nejspíše poskytnout informace o chřipce, a nemocní budou tedy zadávat fráze typu „lék na kašel a horečku". To však nebylo podstatné: předem to nevěděli a navrhli takový systém, který žádné fráze neupřednostňoval. Jejich program hledal pouze korelace mezi frekvencí určitých vyhledávacích dotazů a šířením chřipky v čase a prostoru. Při testování vyhledávacích termínů celkem zpracovali neuvěřitelných 450 milionů různých matematických modelů a porovnávali přitom své předpovědi se skutečnými případy chřipky, jak je zachytila agentura CDC v letech 2007 a 2008. A opravdu narazili na zlatou žílu: jejich software našel kombinaci 45 vyhledávacích termínů, které po zahrnutí do jednotného matematického modelu vykazovaly silnou korelaci mezi svými předpověďmi a oficiálními celostátními čísly. Podobně jako agentura CDC dokázal tento model určit, kam se chřipka rozšířila. Na rozdíl od dat CDC však mohl tyto informace poskytovat téměř v reálném čase a nikoli o týden či dva týdny později.

Když se tedy v roce 2009 objevila epidemie viru H1N1, ukázalo se, že systém společnosti Google je užitečnějším a včasnějším indikátorem než vládní statistiky s jejich přirozenou oznamovací prodlevou. Hygienici ministerstva zdravotnictví mohli využívat cenné informace.

Metoda společnosti Google se překvapivě obejde bez distribuce tampónů k výtěru úst a nevyžaduje spolupráci lékařských ordinací. Místo toho je založena na „veledatech"* (Big Data), díky nimž může společnost inovativními

* Z důvodu snazší čitelnosti byla v textu knihy zvolena tato počeštěná varianta původního termínu Big Data. Pojem veledata není ustáleným ekvivalentem a častěji se setkáte s původním označením Big Data.

způsoby zpracovávat informace a získávat užitečné a cenné poznatky o zboží a službách. Až tedy udeří příští pandemie, bude mít svět k dispozici lepší nástroj, jak předpovídat její potenciální průběh a omezit jej.

Péče o veřejné zdraví je pouze jednou z mnoha oblastí, kterou veledata zásadně ovlivňují. Velká data mění podobu celých podnikových odvětví. Dobrým příkladem je nákup letenek.

Roku 2003 se Oren Etzioni potřeboval dopravit ze Seattlu do Los Angeles na svatbu svého mladšího bratra. Několik měsíců před touto významnou událostí si přes Internet pořídil letenku v domnění, že čím dříve ji objedná, tím bude levnější. Během letu jej přemohla zvědavost a zeptal se svého souseda, kolik stála jeho letenka a kdy si ji objednal. Ukázalo se, že ten člověk zaplatil značně méně než Etzioni, ačkoli svou letenku koupil mnohem později. Znechucený Etzioni položil stejnou otázku několika dalším cestujícím. Většina z nich získala svou letenku výhodněji než on.

Většinu z nás by zlost nad touto ekonomickou nespravedlností přešla dříve, než bychom zavřeli své palubní stolky a narovnali opěradla svých sedaček. Etzioni však patří mezi nejpřednější informatiky ve Spojených státech. Svět se pro něj skládá z řady veledatových problémů, které dokáže vyřešit. V jejich řešení se zdokonaluje už od doby, kdy roku 1986 dokončil Harvardskou univerzitu jako její první absolvent magisterské informatiky.

Během svého působení na University of Washington založil několik firem zaměřených na veledata ještě dříve, než termín „big data“ vůbec vznikl. Podílel se na vývoji jednoho z prvních webových vyhledávačů MetaCrawler, který začal fungovat roku 1994 a později jej koupila společnost InfoSpace – v té době silný internetový hráč. Patřil mezi zakladatele firmy Netbot, která provozovala první významný webový porovnávač cen a kterou prodal portálu Excite. Jeho projekt ClearForest specializovaný na zjišťování významu textových dokumentů později převzala agentura Reuters.

Když letadlo přistálo, Etzioni už měl plán – umožní lidem zjišťovat, zda jsou ceny letenek nabízené na webu výhodné či nikoli. Sedadla v letadle jsou zboží: jednotlivá místa v rámci stejného letu se v zásadě neliší. Ceny však

výrazně kolísají v závislosti na spoustě faktorů, které jsou většinou známy jen samotným aerolinkám.

Etzioni usoudil, že příčiny či důvody těchto cenových rozdílů nemusí dešifrovat. Potřebuje jen předpovědět, zda zobrazená cena v budoucnosti pravděpodobně vzroste či poklesne. To sice není snadné, ale lze to zvládnout. Stačí k tomu „pouze" analyzovat všechny prodeje letenek na dané trase a prozkoumat zaplacené částky vzhledem k počtu dní, které zbývají do odletu.

Jestliže se průměrná cena za letenku obvykle snižuje, bylo by rozumné počkat a koupit letenku později. Pokud by se průměrná cena zpravidla zvyšovala, systém by doporučil pořídit letenku okamžitě za zobrazenou cenu. Etzioni si jinými slovy představoval dokonalejší verzi neformálního dotazování, které uskutečnil ve výšce deseti kilometrů. Bylo jasné, že se jedná o další masivní počítačový problém. Opět šlo ale o problém, jaký dokázal řešit. Pustil se tedy do práce.

Na základě vzorku 12 000 cenových údajů zjištěných sledováním cestovatelského webu v průběhu 41 dní vytvořil Etzioni prediktivní model, který svým simulovaným pasažérům ušetřil slušné peníze. Model nevysvětloval, *proč* se ceny mění, pouze ukazoval *jak*. Nezahrnoval tedy žádné proměnné, které se podílejí na stanovení ceny aeroliniemi, například počet zbývajících neprodaných míst, roční období či to, zda může letenku zlevnit nějaký kouzelný víkendový nocleh. Odhady modelu jsou založeny na zjištěných faktech: pravděpodobnostech, které vyplývají z dat o jiných letech. „Koupit, či nekoupit, toť otázka," hloubal Etzioni. Svůj výzkumný projekt nazval přiléhavě Hamlet.

Malý projekt se vyvinul do začínající firmy Farecast, kterou podpořil rizikový kapitál. Díky svým předpovědím, zda cena letenky pravděpodobně vzroste či klesne a o kolik, umožnil systém Farecast zákazníkům zvolit nejvhodnější čas nákupu. Vybavil je informacemi, k nimž do té doby neměli přístup. Systém Farecast transparentně přistupoval dokonce i ke svým vlastním předpovědím a uváděl také informace o jejich důvěryhodnosti.

Ke svému fungování systém potřeboval hodně dat. Aby zlepšil jeho výsledky, Etzioni dojednal přístup k jedné z databází s rezervacemi letenek. Na základě těchto informací mohl systém generovat předpovědi založené na všech sedadlech každého letu na většině tras systému osobní letecké dopravy v USA během jednoho roku. Farecast nyní ve svých odhadech vycházel z analýzy téměř 200 miliard záznamů o cenách letenek. Přitom dokázal zákazníkům ušetřit pořádný balík.

Etzioni se svými plavými vlasy, širokým úsměvem a andělskými rysy sotva vypadá jako člověk, který by připravil aerolinky o miliony dolarů potenciálního zisku. Ve skutečnosti se však rozhodl dokázat ještě více. V roce 2008 plánoval, že svou metodu nasadí na další zboží typu hotelových pokojů, koncertních vstupenek a ojetých aut: na všech trzích, kde se produkty příliš neliší, značně kolísají ceny a existuje dostatek dat. Než však mohl své plány rozvinout, ozvala se mu společnost Microsoft, která za službu Farecast nabídla nějakých 110 milionů dolarů a integrovala ji do svého vyhledávače Bing. V roce 2012 již systém poskytoval správné odhady v 75 procentech případů a v průměru každému cestovateli ušetřil 50 dolarů za letenku.

Farecast představuje typický příklad veledatové firmy a naznačuje, kam směřuje vývoj. Před pěti či deseti lety by Etzioni svou firmu nedokázal vybudovat. „Nebylo by to možné," říká. Výpočetní výkon a potřebná kapacita úložiště byly tehdy příliš drahé. Technologické změny sice představovaly kritický faktor, ale nenápadně se změnilo i něco důležitějšího. Nastal myšlenkový posun v tom, jak lze data využívat.

O datech se přestalo uvažovat jako o něčem, co je statické, zastarává a ztrácí svou užitečnost poté, kdy splnilo svůj účel – například po přistání letadla (nebo v případě společnosti Google po zpracování výsledků vyhledávacího dotazu). Data se místo toho změnila na surovinu podnikání, cenný ekonomický vstup, který umožňuje vytvářet nové ekonomické hodnoty. V praxi stačí správný myšlenkový přístup a data je možné chytře opakovaně používat jako zdroj inovací a nových služeb. Těm, kdo jsou otevření a ochotní a mají vhodné analytické nástroje, mohou data odhalit svá tajemství.

Data promlouvají

Plody informační společnosti nelze přehlédnout. Po kapsách nosíme mobilní telefony a v batozích notebooky a ve firmách pracují výkonné serverové systémy. Méně nápadné jsou však samotné informace. Půl století poté, co se počítače začaly běžně používat, došlo k takové akumulaci dat, že se děje něco nového a zvláštního. Kromě toho, že je svět zaplaven více informacemi než kdykoli dříve, navíc objem těchto informací narůstá stále rychleji. Změna měřítka způsobila změnu stavu. Kvantitativní změna se proměnila na kvalitativní. Vědy jako astronomie a genomika, které tuto datovou explozi zažily nejdříve (již v prvním desetiletí tohoto století), přišly s termínem „Big Data" (veledata). Tato koncepce se nyní šíří do všech oblastí lidské činnosti.

Veledata nemají žádnou přesnou definici. Původně se vycházelo z toho, že objem informací vzrostl natolik, že zkoumané množství dat se již nevejde do paměti počítačů, které je zpracovávají. Informatici proto museli změnit nástroje na analýzu dat. To je základem nových technologií zpracování, jako je MapReduce společnosti Google a její ekvivalent typu open source s názvem Hadoop, který vytvořila společnost Yahoo. Tyto nástroje umožňují zpracovávat mnohem větší rozsah dat než dříve. Důležité přitom je, že tato data není nutné ukládat do uspořádaných řádků klasických databázových tabulek. Začínají se objevovat také jiné technologie zpracování dat, které odstraňují dosavadní pevné hierarchie a homogenitu. Internetové firmy dokážou shromažďovat velké objemy dat a mají značný finanční zájem tato data analyzovat. Tyto firmy se proto staly průkopníky nejnovějších metod zpracování a předběhly přitom klasické společnosti z tohoto oboru, z nichž některé měly až desítky let zkušeností.

Celou problematiku, jak ji pojímáme i v této knize, můžeme přiblížit takto: veledata znamenají, že ve velkém měřítku lze provádět některé operace, které se v malém měřítku dělat nedají. Je například možné extrahovat nové poznatky nebo generovat novou hodnotu takovými způsoby, které mění trhy, organizace, vztahy mezi občany a úřady atd.

To je však pouze začátek. Éra veledat ovlivní náš životní styl a způsob, kterým interagujeme se světem. Zásadní změna spočívá v tom, že společnost

musí do jisté míry omezit svůj důraz na kauzalitu a spokojit se s jednoduchými korelacemi, kdy nevíme *proč*, ale pouze *co*. To představuje přelom oproti stovkám let zavedené praxe a narušuje to naši standardní představu o tom, jak přijímat rozhodnutí a chápat realitu.

Veledata symbolizují začátek rozsáhlé transformace. Podobně jako mnoho nových technologií se veledata nepochybně stanou obětí nechvalně známého módního cyklu: poté, co se bude tento termín objevovat na prvních stránkách časopisů a v proslovech na oborových konferencích, bude celý trend odmítnut a mnoho začínajících datových firem se dostane do problémů. Krajní nadšení i zavrhování se však zcela míjejí s významem toho, co se odehrává. Stejně jako dalekohled nám umožnil pochopit vesmír a díky mikroskopu jsme mohli poznat mikroby, nové technologie na shromažďování a analýzu velkých objemů dat nám pomohou porozumět našemu světu způsoby, jejichž význam teprve začínáme doceňovat. V této knize se nesnažíme veledata propagovat, ale pouze přispět k lepšímu povědomí o nich. Skutečná revoluce opět neprobíhá ve sféře počítačů, které data zpracovávají, ale v samotných datech a v tom, jak s nimi pracujeme.

Abychom si uvědomili, jak daleko již informační revoluce postoupila, zamysleme se nad trendy v různých společenských oblastech. Náš digitální vesmír se neustále rozšiřuje. Jako příklad můžeme uvést astronomii. Když roku 2000 začal projekt mapování oblohy SDSS (Sloan Digital Sky Survey), jeho teleskop v Novém Mexiku shromáždil v několika prvních týdnech více dat, než kolik jich bylo získáno za celou historii astronomie. V roce 2010 již archiv projektu obsahoval ohromujících 140 terabajtů informací. Obdobný projekt LSST (Large Synoptic Survey Telescope), jehož dalekohled v Chile by měl být uveden do provozu v roce 2016, však toto množství dat získá každých pět dní.

Takové astronomické hodnoty však nacházíme i na místech, která jsou nám bližší. Než vědci roku 2003 poprvé dekódovali lidský genom, trvalo jim sekvenování třech miliard párů bází lidské DNA deset let intenzivní práce. Nyní, o deset let později, dokáže takovou délku DNA sekvenovat jediná laboratoř za den. Ve světě financí zase jen na akciových trzích v USA mění

každý den své majitele asi sedm miliard akcií. Asi dvě třetiny tohoto objemu obchodů přitom zajišťují počítačové algoritmy založené na matematických modelech, které analyzují hory dat, aby dokázaly předpovědět zisky a pokud možno omezily rizika.

Se záplavou dat se musí vypořádat zejména internetové firmy. Google každý den zpracovává více než 24 petabajtů dat. Jedná se o tisíckrát více dat, než kolik jich obsahují všechny tištěné publikace v knihovně amerického Kongresu. Na Facebook, který ještě před deseti lety vůbec neexistoval, nahrávají uživatelé každou hodinu více než 10 milionů nových fotografií. Uživatelé Facebooku klepnou na tlačítko „líbí se" nebo vloží komentář přibližně třimiliardkrát denně a zanechávají tak digitální stopu, kterou může společnost podrobně analyzovat a zjišťovat tak uživatelské preference. Osm set milionů lidí, kteří měsíčně použijí službu YouTube společnosti Google, nahraje každou sekundu videosekvence v celkové délce jedné hodiny. Počet zpráv na Twitteru roste asi o 200 procent ročně a v roce 2012 již přesáhl 400 milionů tweetů denně.

Od vědeckého výzkumu po zdravotnictví, od bankovnictví po Internet se v nejrůznějších sektorech opakuje podobný příběh: objem dat na celém světě rychle roste a překonává nejen možnosti počítačů, ale také hranice naší představivosti.

Mnozí lidé se snažili kvantifikovat informace, které nás obklopují, a vypočítat, jak rychle jejich objem roste. Tyto pokusy přitom byly úspěšné v různé míře, protože měřily odlišné parametry. Jednu z podrobnějších studií provedl Martin Hilbert na Annenberg School for Communication and Journalism při Univerzitě Jižní Kalifornie. Jeho studie měla číselně vyhodnotit vše, co bylo vytvořeno, uloženo a sděleno. Týkala se nejen knih, obrazů, e-mailů, fotografií, hudby a videa (analogového a digitálního), ale také videoher, telefonních volání, dokonce systémů automobilové navigace a dopisů poslaných poštou. Hilbert do svých kalkulací zahrnul i vysílání typu televize a rádia, přičemž zohlednil sledovanost těchto médií.

Podle Hilbertova odhadu existovalo v roce 2007 více než 300 exabajtů uložených dat. Abychom dokázali pochopit, co toto číslo znamená, můžeme si je přiblížit následovně. Celovečerní film v digitální formě lze komprimovat

do souboru velikosti jednoho gigabajtu. Exabajt je jedna miliarda gigabajtů. Stručně řečeno – je to hodně dat. Stojí za pozornost, že v roce 2007 bylo pouze 7 procent dat v analogové formě (na papíře, v knihách, vyvolaných fotografiích atd.). Zbytek byl digitální. Není to však tak dlouho, kdy situace vypadala značně odlišně. O „informační revoluci" a „digitální éře" se sice mluví již od 60. let, ale teprve nyní se tyto ideje do jisté míry naplňují. Ještě nedávno – v roce 2000 – byla v digitální formě uložena pouze čtvrtina světových informací. Zbývající tři čtvrtiny byly na papíře, filmu, gramofonových deskách, magnetofonových kazetách apod.

Zkušené uživatele Internetu, kteří již léta nakupují knihy online, to sice může překvapit, ale digitální informace tehdy ještě netvořily většinu. (V roce 1986 mělo vlastně kolem 40 procent univerzálního výpočetního výkonu na světě formu kapesních kalkulátorů, které v té době poskytovaly větší výpočetní výkon než všechny tehdejší osobní počítače.) Avšak vzhledem k tomu, že objem digitálních dat se zvětšuje tak rychle – podle Hilberta se téměř zdvojnásobuje za pouhé tři roky – situace se rychle obrátila. Analogové informace oproti tomu téměř nepřibývají. V roce 2013 se tedy objem uložených informací na celém světě odhaduje na přibližně 1 200 exabajtů, z čehož nejsou digitální méně než dvě procenta.

Neexistuje vhodný způsob, jak si takový rozsah dat představit. Pokud by všechny tyto informace byly vytištěny v knihách, pokryly by celý povrch Spojených států amerických asi v 52 vrstvách. Kdybychom je vypálili na disky CD-ROM a naskládali na sebe, dosáhly by pětkrát ze Země na Měsíc. Když se ve třetím století př. n. l. egyptský vládce Ptolemaios II. pokusil shromáždit kopie všech napsaných knih, velkolepá Alexandrijská knihovna uchovávala kompletní světové znalosti. Digitální potopu, která dnes zaplavuje planetu, si můžeme přiblížit tak, jako bychom každému dnes žijícímu člověku poskytli 320krát tolik informací, než kolik jich nejspíše obsahovala Alexandrijská knihovna.

Svět se skutečně zrychluje. Objem uložených informací roste čtyřikrát rychleji než světová ekonomika, zatímco výpočetní výkon počítačů se

zvětšuje devětkrát rychleji. Lidé si pochopitelně stěžují, že jsou informacemi zahlceni. Se změnami se musí potýkat každý z nás.

Vraťme se znovu do historie a porovnejme současný datový příliv s předchozí informační revolucí, kterou přinesl Gutenbergův tiskařský stroj, vynalezený kolem roku 1439. Podle historičky Elizabeth Eisenstein bylo v padesáti letech od roku 1453 do roku 1503 vytištěno asi osm milionů knih. Udává se, že to je více knih, než kolik jich vytvořili všichni evropští písaři od založení Konstantinopole o nějakých 1 200 let dříve. Jinými slovy, zdvojnásobení uložených informací v Evropě si tehdy vyžádalo 50 let, zatímco dnes to trvá jen asi tři roky.

Co tento nárůst znamená? Peter Norvig, expert na umělou inteligenci ve společnosti Google, s oblibou používá analogii s obrazy. Nejdříve nás vybídne, abychom se zadívali na ikonické vyobrazení koně z jeskynních maleb ve francouzské jeskyni Lascaux, které vzniklo v paleolitu asi před 17 000 lety. Poté máme za úkol představit si fotografii koně – nebo ještě lépe rozmáchlé tahy Pablo Picassa, které jeskynní obrazy hodně připomínají. Mimochodem když Picassovi ukázali malby z Lascaux, poznamenal, že jsme od té doby nic nového nevymysleli.

Picasso měl sice pravdu, ale jen v jistém smyslu. Vraťme se nyní k té fotografii koně. Nakreslení obrazu koně kdysi zabralo hodně času, ale dnes můžeme pomocí fotografie vytvořit reprezentaci koně mnohem rychleji. Je to sice změna, ale nejspíše nikoli klíčová, protože v zásadě máme pořád totéž: obraz koně. Nyní nás však Norvig požádá, abychom myšlenkově zaznamenávali obraz koně a sérii obrazů začali přehrávat rychlostí 24 snímků za sekundu. Kvantitativní změna nyní přešla do stadia kvalitativní změny. Princip filmu a statické fotografie se zásadně liší. S veledaty je to obdobné: když změníme množství, mění se samotná podstata.

Nyní uvažujme analogii z oblasti nanotechnologie, kde se věci nezvětšují, ale neustále zmenšují. Princip nanotechnologie spočívá v tom, že když se dostaneme na molekulární úroveň, mohou se změnit fyzické vlastnosti. Když tyto nové vlastnosti zjistíme, dokážeme vyvíjet materiály s takovými vlastnostmi, jaké dosud nebylo možné vytvořit. V měřítku nanočástic

mohou například existovat ohebnější kovové materiály a pružná keramika. Naopak platí, že když zvětšíme rozsah dat, s nimiž pracujeme, můžeme získat výsledky, jakých bychom při analýze menších datových množin nedosáhli.

Omezení, se kterými žijeme a která považujeme za univerzální, jsou občas závislá na měřítku, v němž pracujeme. Naše třetí analogie opět pochází ze sféry vědy. Pro nás lidi je nejdůležitějším fyzikálním principem gravitační zákon: ovlivňuje vše, co děláme. V životě drobného hmyzu se však gravitace téměř neprojevuje. Zástupci některých druhů, jako jsou bruslařky, místo toho závisí na fyzikálním jevu povrchového napětí, které jim umožňuje pohybovat se po hladině rybníka, aniž by se potopili.

Stejně jako ve fyzice také u informací záleží na velikosti. Společnost Google proto dokáže mapovat výskyt chřipky stejně dobře jako oficiální data, která jsou založena na skutečných návštěvách pacientů u lékaře. Tuto schopnost získala rozborem stovek miliard vyhledávacích termínů a odpověď přitom může poskytnout téměř v reálném čase, mnohem rychleji než oficiální zdroje. Obdobně Etzioniho systém Farecast umí předpovědět kolísání cen letenek a přesunuje tak značný díl ekonomické moci do rukou spotřebitelů. Obě firmy k tomu však potřebují analyzovat stovky miliard datových bodů.

Tyto dva příklady ilustrují vědeckou a společenskou důležitost veledat a ukazují také, do jaké míry mohou veledata představovat zdroj ekonomické hodnoty. Naznačují dva způsoby, kterými svět veledat nevyhnutelně otřese základy všech oblastí od podnikání a výzkumu po zdravotní péči, veřejnou správu, vzdělávání, ekonomiku, humanitní obory a všechny další společenské sektory.

Éra veledat sice teprve začíná, ale už nyní na ně denně spoléháme. Spamové filtry jsou navrženy tak, aby se dokázaly automaticky přizpůsobovat novým typům nevyžádaných zpráv. Software není možné naprogramovat tak, aby standardně dokázal blokovat slovo „via6ra“ nebo jeho nesčetné varianty. Seznamovací weby dávají dohromady dvojice podle toho, nakolik jejich početné atributy korelují s atributy předchozích párů, které byly

úspěšné. Funkce automatických oprav ve smartphonech sleduje naše chování a podle toho, co píšeme, přidává do svého slovníku pro kontrolu pravopisu nová slova. Tyto aplikace však představují pouhý začátek. Od aut, která umí v pravý čas strhnout řízení nebo zabrzdit, po počítač Watson společnosti IBM, který dokáže porazit lidské účastníky televizní hry *Riskuj!*, uvedený přístup promění mnoho aspektů světa, v němž žijeme.

Smysl veledat spočívá v předpovědích. Lze je sice popsat v rámci informatické disciplíny zvané umělá inteligence a konkrétně se jimi zabývá oblast zvaná strojové učení, ale příslušnost k tomuto oboru je zavádějící. Veledata nemají „naučit" počítač, aby „myslel" jako lidé. Místo toho se na velké objemy dat aplikují matematické postupy, které umožňují určit pravděpodobnost: toho, zda e-mailová zpráva je spam, zda napsaná písmena „ajko" měla tvořit slovo „jako", případně zda ze směru a rychlosti člověka přecházejícího mimo přechod vyplývá, že se na druhou stranu ulice dostane včas, takže automaticky řízené auto může zpomalit jen mírně. Pro správné fungování všech těchto systémů je klíčové, že na svých vstupech dostávají hodně dat, na nichž poté zakládají své odhady. Systém jsou navíc vytvořeny tak, aby se v čase postupně samy zlepšovaly. V přicházejících datech přitom sledují nejlepší signály a vzory, kterých si mají všímat.

V budoucnu, které nastane dříve, než se nadějeme, bude počítačovými systémy doplněno nebo nahrazeno mnoho prvků našeho světa, které jsou v nyní výhradní doménou lidského úsudku. Nejde jen o řízení vozidel či hledání partnerů, ale i o složitější úlohy. Amazon přeci dokáže doporučit ideální knihu, Google umí určit nejvíce relevantní web, Facebook ví, co máme rádi, a LinkedIn odvozuje, koho známe. Stejné technologie se uplatní při diagnóze nemocí, doporučování vhodné léčby, a snad dokonce při identifikaci „zločinců" ještě před tím, než skutečně spáchají nějaký trestný čin. Stejně jako Internet radikálně proměnil svět, protože dal počítačům schopnost komunikace, tak i veledata transformují základní aspekty života. Dají mu totiž kvantitativní dimenzi, kterou nikdy předtím neměl.

Stále více, s chybami, bez vysvětlení

Veledata budou přinášet nové ekonomické hodnoty a inovace. V sázce je toho ale mnohem více. Nástup veledat ve třech ohledech mění způsob, jakým analyzujeme informace. Tyto změny nás nutí zásadně přehodnotit pohled na organizaci naší společnosti.

První posun popíšeme v druhé kapitole. V tomto novém světě umíme analyzovat mnohem více dat. V některých případech dokážeme zpracovat dokonce *všechna* data týkající se určitého jevu. Když se společnost potřebovala vypořádat s velkými čísly, již od devatenáctého století přitom spoléhala na vzorky. Nutnost vzorkování však představuje pozůstatek z období nedostatku informací – důsledek přirozených omezení při interakci s informacemi v analogové éře. Než se rozšířily vysoce výkonné digitální technologie, nepovažovali jsme vzorkování za umělý limit – obvykle jsme je brali jako něco samozřejmého. Když použijeme všechna data, můžeme pozorovat detaily, kterých bychom si při omezení na menší množství nikdy nemohli všimnout. Veledata nám poskytují mimořádně jasný pohled na hierarchickou strukturu: dílčí kategorie a tržní segmenty, které vzorky nedokážou postihnout.

Při sledování značně většího objemu dat také můžeme tlumit svou snahu o přesnost. To je druhý posun, kterým se budeme zabývat ve třetí kapitole. Něco za něco: když odstraníme chyby dané vzorkováním, můžeme připustit více chyb měření. Jestliže máme pouze omezenou schopnost měřit, kvantifikujeme pouze ta nejdůležitější fakta. V takové situaci je vhodné, abychom se snažili zjistit přesné hodnoty. Nemůžeme prodávat dobytek, když zájemci nedokážeme říci, zda stádo obsahuje 100 nebo jen 80 kusů. Všechny naše digitální nástroje až dosud předpokládaly přesné vstupy a výstupy: věřili jsme, že databázový modul načte záznamy, které dokonale odpovídají našemu dotazu, stejně jako tabulkový procesor poskytne odpovídající součet čísel ve sloupci.

Tento styl uvažování vycházel z prostředí „malých dat“: když jsme toho měřili jen málo, pak jsme museli těch nemnoho zjištěných čísel kvantifikovat co nejpřesněji. Z jistého pohledu je to samozřejmé: majitel malého obchodu může po skončení otevírací doby spočítat tržbu do poslední koruny, ale

v případě hrubého domácího produktu bychom se o to ani nepokusili, protože bychom nemohli uspět. Jak se zvětšuje měřítko, roste i počet nepřesností.

Přesnost vyžaduje, aby byla data pečlivě ošetřována. Takový přístup může fungovat u malých sad a v některých situacích je skutečně potřebný: v bance buď máme, nebo nemáme dost peněz, abychom mohli zadat platební příkaz. Když však ve světě veledat pracujeme s mnohem komplexnějšími datovými množinami, můžeme se rigidní snahy o přesnost do jisté míry zbavit.

Veledata jsou často neuspořádaná, jejich kvalita kolísá a hodnoty jsou distribuovány mezi nesčetné servery po celém světě. U veledat se často spokojujeme s představou o obecných trendech a nepotřebujeme daný jev vyčíslit do posledního centimetru, haléře či atomu. Přesnosti se nevzdáváme zcela, jenom jí nepřikládáme absolutní váhu. Co ztratíme na přesnosti na úrovni detailů, získáme při analýzách na makroskopické úrovni.

Tyto dva posuny vedou ke třetí změně, kterou vysvětlíme ve čtvrté kapitole: k opouštění klasického hledání kauzality. My lidé se ze své povahy snažíme najít příčiny, i když řetězec příčin a následků bývá často těžké sledovat a můžeme se přitom dostat na scestí. Ve světě veledat se oproti tomu nemusíme upínat na kauzalitu. Místo toho můžeme v datech vyhledávat vzory a korelace, které nám nabídnou dosud neznámé a neocenitelné poznatky. Korelace nám nemusí přesně říci, *proč* se něco děje, ale upozorní nás na to, *co* se děje.

A v mnoha situacích to úplně stačí. Pokud miliony elektronických zdravotních záznamů odhalí, že u onkologických pacientů, kteří užívají určitou kombinaci aspirinu a pomerančového džusu, dochází k ústupu nádoru, pak může být přesná příčina zlepšení zdravotního stavu méně důležitá než samotný fakt, že přežili. Podobně dokážeme-li ušetřit peníze, když známe nejlepší čas na nákup letenky, stačí to, i když pravidlům kolísání cen nerozumíme. Veledata odpovídají na dotaz *co*, nikoli *proč*. Pokaždé nepotřebujeme znát příčinu jevu, ale můžeme ponechat data, aby promluvila sama.

Před příchodem veledat jsme se při svých analýzách obvykle omezovali na testování malého počtu hypotéz, které jsme definovali ještě dříve, než jsme data začali sbírat. Když naslouchámе datům, můžeme najít vazby, které by

nás nikdy nenapadly. Některé investiční fondy proto předpovídají trend akciových trhů pomocí analýzy zpráv na Twitteru. Amazon a Netflix generují svá doporučení produktů z nespočtu uživatelských interakcí na svých webech. Služby Twitter, LinkedIn i Facebook mapují „sociální graf" vztahů svých uživatelů, aby se dozvěděly více o jejich preferencích.

Lidé samozřejmě analyzují data již tisíce let. Písmo ve starověké Mezopotámii vzniklo proto, že úředníci potřebovali efektivní nástroj k zápisu a zpracování informací. Již od biblických dob vládci organizují sčítání lidu, aby shromáždili velké množství dat o občanech svých zemí, a přibližně dvě stě let zase pojišťovací experti pořizují rozsáhlé záznamy týkající se rizik, aby jim dokázali porozumět – nebo se jim alespoň vyhnout.

V analogovém věku však bylo shromažďování a analyzování takových dat mimořádně drahé a časově náročné. Když se objevily nové otázky, často bylo potřeba data pořídit znovu a spustit analýzu od začátku.

Velký pokrok v oblasti správy dat přinesla nastupující digitalizace, kdy se analogové informace převádějí do podoby čitelné do počítače. Přitom lze tyto informace zároveň snáze a levněji ukládat a zpracovávat. Tato změna zvýšila efektivitu zásadním způsobem. Procesy shromažďování a analýzy informací, které kdysi trvaly léta, bylo možné nyní provést během několika dní nebo ještě rychleji. Jinak se toho však změnilo jen málo. Lidé, kteří data analyzovali, příliš často vězeli v analogovém paradigmatu a předpokládali, že datové množiny mají jediný účel, s nímž souvisí i jejich hodnota. Tento zavádějící předpoklad podporovaly i použité postupy. Digitalizace sice umožnila přechod k veledatům, ale samotná existence počítačů ještě k nástupu veledat nevedla.

Pro současné dění neexistuje vhodný termín, ale příslušné změny bychom mohli vyjádřit novotvarem *datafikace* (datafication). Jeho podstatu představíme v páté kapitole. Tento pojem se týká pořizování informací o všem, co se na světě odehrává – včetně toho, co jsme nikdy za zdroj informací nepovažovali, jako je poloha osoby, vibrace motoru či napětí mostu – a transformace těchto údajů do datového formátu, abychom je mohli kvantifikovat.

Díky tomu můžeme informace používat novými způsoby, jako při prediktivní analýze, kdy například na základě tepla či vibrací motoru detekujeme, že hrozí jeho porucha. Tímto způsobem můžeme odhalit implicitní a skrytou hodnotu informací.

Právě se odehrává hon za pokladem motivovaný cennými informacemi, které lze z dat získat, a skrytými hodnotami, které je možné uvolnit díky posunu od kauzality ke korelaci. Pokladů je však více. Každá jednotlivá sada dat v sobě pravděpodobně skrývá jistou zatím neobjevenou cenu a probíhá závod o to, kdo ji objeví a zpracuje jako první.

Jak ukážeme v šesté a sedmé kapitole, veledata mění povahu podnikání, trhů a společnosti. Ve dvacátém století došlo k přesunu hodnot od fyzické infrastruktury typu země a továren k nehmotným statkům, jako jsou ochranné známky a duševní vlastnictví. Tento trend se nyní rozšiřuje na data, která se stávají důležitým podnikovým aktivem, klíčovým ekonomickým vstupem a základem pro nové obchodní modely. Jedná se o palivo informační ekonomiky. Data se sice na podnikových rozvahách objevují jen málokdy, ale pravděpodobně je jen otázkou času, kdy se to změní.

Některé metody analýzy dat existují již dlouho, avšak v minulosti byly dostupné jen tajným službám, výzkumným ústavům a největším světovým společnostem. Společnosti Walmart a Capital One patřily mezi průkopníky uplatnění veledat v maloobchodě a bankovnictví a přitom změnily celá odvětví. Mnohé z těchto nástrojů jsou nyní k dispozici širšímu okruhu zájemců (ačkoli samotná data nikoli).

Největší otřes mohou představovat důsledky pro jednotlivce. Ve světě, který pracuje s pravděpodobností a korelací, již tolik nezáleží na expertních znalostech z konkrétní oblasti. Ve filmu *Moneyball* hledače baseballových talentů zastínili statistici, když se lépe než instinktivní dojmy osvědčila sofistikovaná analýza. Specialisté na určitou oblast tedy nepřijdou o práci, ale budou se muset potýkat s výstupy z analýzy veledat. Tento proces si vynutí změny tradičních představ o managementu, rozhodování, personalistice a vzdělávání.

Většina našich institucí měla ve vínku představu, že lidská rozhodnutí jsou založena na informacích, které jsou omezené, přesné a ze své povahy kauzální. Když jsou však data rozsáhlá, lze je zpracovávat rychle a toleruje se přitom nepřesnost, nastává zcela nová situace. Navíc vzhledem k mimořádnému objemu dat často rozhodnutí nedokážou přijímat lidé, ale jen počítače. Temnou stránkou veledat se budeme zabývat v osmé kapitole.

Lidská společnost tisíce let shromažďovala znalosti o lidském chování a možnostech, jak na ně dohlížet. Jak lze však regulovat algoritmus? Již v počátcích rozvoje počítačů si odborníci uvědomili, že by technologie mohly narušovat soukromí. Od té doby si společnost vyvinula sadu pravidel na ochranu osobních informací. Ve věku veledat však tyto zákony představují prakticky bezcennou Maginotovu linii. Lidé ochotně sdílejí informace online, což je základem příslušných služeb a nejde o slabinu, kterou by bylo nutné odstranit.

Nebezpečí pro nás jako jednotlivce se mezitím přestává týkat soukromí, ale spíše pravděpodobnosti: algoritmy budou předpovídat pravděpodobnost, že dostaneme infarkt (a budeme tedy platit vyšší zdravotní pojištění), přestaneme splácet hypotéku (a banka nám ji tedy raději nepřidělí) nebo spácháme zločin (a možná budeme zatčeni preventivně). Tento vývoj vede k etickým dilematům ohledně role svobodné vůle ve světě, kde o všem rozhodují data. Mělo by mít osobní rozhodování přednost před veledaty, i když statistiky naznačují opak? Stejně jako tiskařský stroj připravil půdu pro zákony garantující svobodu slova (ta dříve neexistovala, protože do té doby bylo příliš málo písemností, které by bylo potřeba chránit), tak éra veledat bude vyžadovat nová pravidla, která zabezpečí nedotknutelnost jednotlivce.

Způsob, kterým kontrolujeme data a manipulujeme s nimi, se v mnoha ohledech bude muset změnit. Vstupujeme do světa plného předpovědí založených na datech, kde často nebudeme schopni vysvětlit důvody pro svá rozhodnutí. Co to bude znamenat, jestliže lékař nedokáže zdůvodnit zdravotní zákrok, aniž by pacienta odkázal na černou skříňku? Právě tato situace nastává, když se lékař musí spoléhat na stanovení diagnózy pomocí veledat. Změní soudní systém svůj standard „pravděpodobné příčiny“ na

„pravděpodobnostní příčinu“ – a jaké budou v tom případě důsledky pro lidskou svobodu a důstojnost?

Doba veledat si žádá nové zásady, které navrhneme v deváté kapitole. Tyto zásady vycházejí z hodnot, které byly vyvinuty a kodifikovány pro svět malých dat. Nestačí však jen aktualizovat stará pravidla podle nových okolností, ale musíme si uvědomit, že potřebujeme úplně nové principy.

Společnost získá četné výhody, protože veledata budou důležitou součástí řešení naléhavých globálních problémů, jako je náprava klimatických změn, vymýcení nemocí a podpora efektivní státní správy a ekonomického rozvoje. Éra veledat však také vyžaduje, abychom se lépe připravili na změny, které nasazení nových technologií způsobí v našich institucích a v našem osobním životě.

Veledata představují důležitý postup v lidském úsilí kvantifikovat svět a porozumět mu. V současnosti se datafikuje většina jevů, které dosud nebylo možné měřit, ukládat, analyzovat a sdílet. Když místo jejich malých částí zpracujeme rozsáhlá data vcelku a upřednostníme více dat před vyšší přesností, otevřeme tím dveře novým metodám získávání poznatků. Společnost bude postupně opouštět svou dlouhodobě osvědčenou snahu o kauzalitu a v mnoha případech využije výhody korelace.

Ideál identifikace kauzálních mechanismů je iluze, která potvrzuje sama sebe a veledata ji rozbíjejí. Opět se dostáváme do historické slepé uličky, kde „Bůh je mrtev“. To znamená, že jistoty, kterým jsme věřili, se znovu mění. Ironií osudu je však tentokrát nahrazují lepší důkazy. Jaká role zůstává pro intuici, víru, nejistotu, jednání v rozporu s dostupnými informacemi a učení se ze zkušeností? Když se svět obrací od kauzality ke korelaci, jak se můžeme pragmaticky posunout dopředu, aniž bychom podkopali samotné základy společnosti, humanity a pokroku založeného na rozumu? Tato kniha se snaží vysvětlit, kde se nacházíme, vystopovat, jak jsme se tam dostali, a nabídnout aktuálně potřebného průvodce výhodami a riziky, která leží před námi.

Kapitola 2

Stále více

Význam veledat spočívá v tom, že umožňují uvnitř množin informací a mezi nimi pozorovat a pochopit vztahy, kterým jsme dosud plně nerozuměli. Podle Jeffa Jonase, specialisty na veledata ve společnosti IBM, musíme nechat data, aby „začala sama mluvit". Na jedné úrovni to může znít banálně. Lidé se při poznávání světa spoléhají na data již dlouho, ať už v neformálním smyslu nesčetných pozorování, která provádíme každý den, nebo (zejména v posledních několika stoletích) ve formálním smyslu kvantifikovaných jednotek, které lze zpracovávat pomocí výkonných algoritmů.

Digitální věk usnadnil a urychlil zpracování dat a umožnil během okamžiku spočítat miliony čísel. Smysl výroku o „mluvících datech" je však poněkud odlišný a širší. Jak jsme uvedli v první kapitole, veledata přinášejí tři zásadní posuny uvažování, které jsou spolu provázané, a proto se vzájemně posilují. První posun vychází ze schopnosti analyzovat rozsáhlé objemy dat o jistém předmětu, aniž bychom se museli spokojit s menšími podmnožinami. Druhý posun vyžaduje přijmout neuspořádanost dat z reálného světa a opustit důraz na přesnost. Třetí posun vede k většímu spoléhání na korelace, místo abychom se neustále snažili pochopit prchavou kauzalitu. V této kapitole se zaměříme na první z těchto posunů: použití všech dostupných dat namísto jejich malé výseče.

Přesné zpracování velkých datových souborů bývalo až dosud značně náročné. Větší část historie jsme pracovali jen s malou částí dat, protože jsme na jejich sbírání, organizaci, ukládání a analýzu měli k dispozici jen málo výkonné nástroje. Informace, na které jsme spoléhali, jsme redukovali na naprosté minimum, abychom je mohli snáze prozkoumat. Jednalo se vlastně o určitou formu nevědomé autocenzury: obtíže se zpracováním dat jsme považovali za nevyhnutelnou realitu, místo abychom je brali jako umělé limity kladené soudobou technologií, kterými ve skutečnosti byly. Dnešní technické možnosti jsou naprosto odlišné. Stále sice existují omezení

na množství dat, která můžeme zvládnout, ale tato omezení jsou mnohem volnější a časem se budou dále zmírňovat.

Do jisté míry jsme zatím stále plně nedocenili svou novou svobodu shromažďovat a používat velké datové sady. Převážná část našich zkušeností i struktura našich institucí vychází z předpokladu, že informace jsou dostupné jen omezeně. Mysleli jsme, že informací můžeme získat jen málo, a obvykle jsme se s tím také spokojili. Tento přístup posiloval sám sebe. Vyvinuli jsme dokonce propracované metody, jak použít co nejméně dat. Jedním z úkolů statistiky je přece potvrdit co nejvíce předpokladů pomocí minimálního objemu dat. Svůj sklon redukovat rozsah informací jsme v důsledku kodifikovali ve svých normách, procesech a podpůrných strukturách. Abychom získali představu o tom, co posun k veledatům znamená, začneme své povídání ohlédnutím.

Teprve nedávno získaly soukromé firmy a po nich i jednotlivci možnost shromažďovat a třídit informace v masivním měřítku. V minulosti se s tímto úkolem dokázaly vypořádat jen mocnější instituce, jako je církev a stát, které v mnoha společnostech splývaly. Nejstarší záznamy o zpracování dat pocházejí přibližně z roku 5000 př. n. l., kdy sumerští obchodníci začali zaznamenávat prodávané zboží na malých hliněných destičkách. Počítání ve větším měřítku však bylo doménou státu. Po tisíce let se vládcové snažili shromažďovat informace o obyvatelích svých zemí.

Vezměme si sčítání lidu. Uvádí se, že sčítání probíhala již ve starověkém Egyptě i Číně. Zmiňuje se o nich i Starý zákon a v Novém zákoně se uvádí, že sčítání nařízené císařem Augustem – „aby byl po celém světě proveden soupis lidu“ (L 2,1) – přivedlo Josefa a Marii do Betléma, kde se narodil Ježíš. Kniha Domesday Book z roku 1086, která patří mezi nejcennější britské památky, ve své době představovala dosud nevídaný podrobný soupis obyvatel Anglie, jejich pozemků a majetku. Královští pověřenci se rozptýlili po celé zemi a zjišťovali údaje, které měly být zapsány do knihy. Ta později získala svůj název „Domesday“ neboli „Doomsday“, protože sčítání mělo odkazovat na biblický Soudný den, kdy se všichni budou zodpovídat ze svých životů.

Sčítání lidu je nákladné a časově náročné. Král Vilém I., který sepsání knihy Domesday Book nařídil, se jejího dokončení nedožil. Nebýt této složité operace, musel by se však vládce obejít bez informací. Dokonce i s vynaložením hodně času a peněz byly informace jen přibližné, protože sčítací komisaři pravděpodobně nedokázali sečíst všechny lidi dokonale. Samotné slovo „census" (sčítání lidu) pochází z latinského výrazu „censere", který znamená „odhadovat".

Před více než 300 lety dostal britský prodavač John Graunt převratný nápad. Graunt chtěl zjistit, kolik lidí žilo v Londýně v době morové epidemie. Místo toho, aby počítal každého člověka, vymyslel postup – dnes bychom jej označili jako „statistiku" – který mu umožnil velikost populace *odvodit*. Jeho metoda sice nebyla příliš přesná, ale osvědčil se výchozí princip, že z malého vzorku lze extrapolovat užitečné znalosti o celé populaci. Důležité však je, jak se přitom postupuje. Graunt svůj vzorek pouze vynásobil.

Současníci jeho systém oceňovali, ačkoli později se zjistilo, že jeho čísla byla přibližně správná jen náhodou. Po celé generace se vzorkování provádělo velmi nespolehlivým způsobem. Při sčítáních lidu a podobných projektech typu veledat proto stále převládal přístup založený na hrubé síle, který usiloval o zpracování všech hodnot.

Vzhledem k tomu, že sčítání lidu byla tak složitá, nákladná a zdlouhavá, pořádala se jen zřídka. Ve starém Římě, který se od dávných dob mohl pochlubit stovkami tisíc obyvatel, probíhala sčítání jednou za pět let. Podle americké ústavy byla sčítání povinná každých deset let. Rostoucí populace mladého státu přitom čítala miliony osob. Koncem devatenáctého století se však ukázalo, že i tato velikost představuje problém. Rozsah dat narostl natolik, že je sčítací úřad přestal zvládat.

Než bylo dokončeno sčítání z roku 1880, uplynulo dlouhých osm let. Informace byly zastaralé ještě dříve, než byly zveřejněny. Ke všemu statistici odhadli, že zpracování dat ze sčítání v roce 1890 bude trvat celých 13 let, což bylo samozřejmě neúnosné, nehledě na to, že by to bylo v rozporu s Ústavou. Z údajů o populaci se vycházelo při rozdělování daňových výnosů

a stanovení počtu zástupců v Kongresu. Nestačilo tedy jen zjistit správné hodnoty, ale čísla musela být známa včas.

Potíže, se kterými se potýkali pracovníci sčítacího úřadu, se podobají těm, které zažívali vědci a ekonomové na začátku tohoto tisíciletí. Najednou jim došlo, že se topí v datech: pořizovaných informací bylo tolik, že naprosto zahltily nástroje určené ke svému zpracování, a bylo potřeba vyvinout nové metody. V roce 1880 byla situace natolik vážná, že se statistický úřad obrátil na amerického vynálezce Hermana Holleritha, aby při sčítání lidu roku 1890 uplatnil svou myšlenku děrných štítků a tabelačních strojů.

S vynaložením velkého úsilí se mu podařilo zkrátit čas tabelace z osmi let na necelý rok. Tento mimořádný úspěch představoval začátek automatizovaného zpracování dat (a položil základy společnosti, která později dostala název IBM). Pořizování a analyzování veledat touto metodou bylo však i nadále velmi drahé. Každý občan USA totiž musel vyplnit formulář a informace bylo nutné přenést na děrný štítek, který sloužil při tabulaci. Při takto nákladných metodách bylo těžko představitelné, že by sčítání lidu mohlo probíhat v kratších intervalech než jednou za deset let, ačkoli takové prodlevy mezi sčítáními byly u rychle rostoucího národa dosti nepraktické.

Objevovalo se tedy dilema: používat všechna data, nebo jen jejich malou část? Nejlepší samozřejmě je, když o libovolném měřeném objektu získáme veškerá data. Jenže v případech, kdy je měřítko příliš velké, to pokaždé nemusí být praktické. Jak ale zvolit vzorek? Někteří odborníci tvrdili, že nejvhodnější cestou vpřed bude účelný výběr vzorků, které budou reprezentovat datový celek. Roku 1934 však polský statistik Jerzy Neyman přesvědčivě dokázal, že takový přístup způsobuje velké chyby. Chceme-li se jim vyhnout, je klíčové, abychom se při výběru vzorku snažili o náhodnost.

Statistici zjistili, že přesnost vzorkování se zvyšuje hlavně díky náhodnému výběru, nikoli při zvětšování vzorku. Ačkoli to může znít překvapivě, náhodně zvolený vzorek 1 100 respondentů na otázku se dvěma možnostmi (ano-ne, s přibližně stejnou pravděpodobností) poměrně dobře reprezentuje celou populaci. V 19 z 20 případů leží v tříprocentním intervalu chyby bez ohledu na to, zda se celá populace počítá ve stovkách tisíc nebo ve stovkách

milionů. Matematické vysvětlení je sice komplikované, ale stručně lze říci, že od určité fáze, která nastává celkem záhy, se při rostoucím počtu pozorování neustále zmenšuje přírůstek nových informací, které nám každé další pozorování poskytuje.

Poznatek, že náhodnost je důležitější než velikost vzorku, byl poměrně překvapivý. Otevřel cestu pro nový přístup ke sbírání informací. Pomocí náhodných vzorků bylo možné shromažďovat data levně a přitom s velkou přesností extrapolovat výsledky na celek. Díky tomu mohly vlády organizovat malé verze sčítání lidu s náhodnými vzorky každý rok a nemusely se spokojit s jedním sčítáním každých deset let. A tuto možnost také začaly využívat. Americký sčítací úřad například kromě sčítání lidu jednou za deset let, kdy se snaží zaznamenat každého občana, každoročně provádí více než dvě stě ekonomických a demografických průzkumů založených na vzorkování. Vzorkování představovalo řešení problému s nárůstem informací v době, kdy se data sbírala a analyzovala velmi těžko.

Tato nová metoda se rychle rozšířila mimo sféru veřejného sektoru a sčítání lidu. Náhodné vzorkování v zásadě omezuje problémy s veledaty na lépe zvládnutelné problémy s daty. V podnicích se uplatňovalo při zajištění kvality výrobků, protože umožňovalo mnohem snáze a levněji zavádět zlepšení. Při komplexní kontrole kvality bylo původně potřeba prohlédnout každý produkt, který opouštěl výrobní pás. Nyní však postačoval náhodný vzorek testů sady produktů. Na podobném principu vznikly také zákaznické průzkumy v maloobchodě a volební průzkumy týkající se politiky. Nové metody ve velké míře transformovaly někdejší humanitní obory na společenské *vědy*.

Náhodné vzorkování dosáhlo mimořádných úspěchů a tvoří základ moderního měření v různém měřítku. Jedná se však pouze o pomůcku a náhražku optimální varianty, tj. shromáždění a analýzy celé datové sady. Vzorkování má ze své podstaty několik nedostatků. Jeho přesnost závisí na zajištění náhodnosti při výběru dat vzorku, ale této náhodnosti se dosahuje těžko. Systematické odchylky při sběru dat mohou vést k značným chybám výsledků zpracování extrapolovaných dat.

Problémy tohoto typu se projevují při volebních průzkumech, kdy tazatelé volají na pevné linky. Jak upozornil americký statistik Nate Silver, takový vzorek diskriminuje lidi, kteří používají jen mobilní telefony (a kteří jsou mladší a politicky liberálnější). Důsledkem byly nesprávné volební předpovědi. Před americkými prezidentskými volbami roku 2008, kdy proti sobě stáli Barack Obama a John McCain, významné volební agentury Gallup, Pew a ABC/Washington Post zjistily, že zahrnutí uživatelů mobilních telefonů změní výsledky o jedno až tři procenta, což byl vzhledem k vyrovnanému souboji poměrně velký rozdíl.

Ještě závažnější je to, že u náhodného vzorkování nelze snadno měnit měřítko a zahrnovat dílčí kategorie, protože při dělení výsledků na menší a menší podskupiny se zvyšuje pravděpodobnost chybných předpovědí. Není těžké pochopit, proč k tomu dochází. Předpokládejme, že náhodně vybereme tisíc lidí a dotazujeme se jich, pro koho budou hlasovat v nadcházejících volbách. Jestliže je náš vzorek dostatečně náhodný, máme šanci, že volební chování celé populace se nebude od názorů vzorku odchylovat více než o tři procenta. Jak ale postupovat, když nám přesnost plus minus tři procenta nestačí? Případně co udělat, chceme-li skupinu rozdělit na menší podskupiny podle pohlaví, místa bydliště nebo výše příjmu?

A jak lze tyto podskupiny zkombinovat a zaměřit se na specifický segment populace? Pokud máme celkový vzorek tisíce lidí, bude podskupina definovaná jako „majetné ženské voličky na Severozápadě" mnohem užší než sto osob. Budeme-li předpovídat volební úmysly *všech* dobře situovaných žen ze Severozápadu z pouhých několika desítek pozorování, dostaneme nepřesné výsledky i tehdy, když zajistíme téměř dokonalou náhodnost. Při nepatrných odchylkách v celkovém vzorku pak na úrovni podskupin dojde k tomu, že se chyby dále zvýrazní.

Užitečnost vzorkování se tedy rychle ztrácí, chceme-li přejít na nižší úroveň a podrobněji se zaměřit na nějakou zajímavou podkategorii dat. To, co funguje na globální úrovni, na úrovni detailů selhává. V tom vzorkování připomíná analogovou fotografickou zvětšeninu. Zdálky vypadá hezky, ale když se podíváme zblízka, uvidíme místo podrobností jen zrno.

Vzorkování je také nutné pečlivě naplánovat a realizovat. Obvykle není možné vzorku položit další otázky, které nebyly zvoleny již na začátku průzkumu. Jako náhražka je tedy vzorkování užitečné, ale vzhledem ke kompromisům, které přináší, se skutečně jedná jen o náhražku. Když je datová sada pouhým vzorkem, postrádá jistou rozšiřitelnost nebo pružnost. Oproti tomu kompletní data je možné opakovaně analyzovat zcela novým způsobem, který jsme při jejich původním shromažďování vůbec nepředpokládali.

Jako příklad můžeme uvést analýzu DNA. Náklady na sekvenování genomu jednotlivce v roce 2012 poklesly k tisíci dolarům, takže se přiblížilo uplatnění této metody na masovém trhu, kde lze využít úspory z měřítka. Díky tomu se rozvíjí nový obor podnikání, který nabízí sekvenování osobního genomu. Začínající firma ze Silicon Valley s názvem 23andMe od roku 2007 analyzuje lidskou DNA a účtuje za to pouze několik set dolarů. Svými metodami dokáže v individuálním genetickém kódu odhalit posloupnosti, které zvyšují náchylnost k některým nemocem, jako je rakovina prsu nebo srdeční choroby. Firma 23andMe navíc agreguje data kódu DNA svých zákazníků s informacemi o jejich zdravotním stavu a pokouší se tak získat nové informace, které by jinak nebylo možné zjistit.

Má to ale háček. Firma sekvenuje jen malou část osobního genetického kódu, konkrétně místa, o kterých je známo, že naznačují určitou genetickou vadu. Miliardy párů bází DNA přitom zůstávají nepovšimnuty. Firma 23andMe tedy dokáže najít odpovědi jen na otázky související s markery, kterými se zabývá. Pokaždé, kdy je objeven nový marker, je potřeba osobní DNA (přesněji řečeno její relevantní část) sekvenovat znovu. Práce s podmnožinou místo s celkem má své výhody i nevýhody: firma může hledanou informaci zjistit rychleji a levněji, ale nemůže odpovídat na otázky, o nichž předem neuvažovala.

Legendární ředitel společnosti Apple Steve Jobs zvolil při svém boji proti rakovině zcela odlišný přístup. Stal se jedním z prvních lidí na světě, kteří si nechali zjistit kompletní sekvenci své DNA, a kromě toho byla určena i sekvence DNA jeho nádoru. Za tuto analýzu zaplatil šestimístnou částku v dolarech – řádově stokrát více, než kolik si účtuje firma 23andMe. Za to však

místo vzorku – pouhé sady markerů – získal datový soubor, který obsahoval celý genetický kód.

Při výběru medikace pro běžné onkologické pacienty musí lékaři doufat, že je jejich DNA dostatečně podobná dědičnému materiálu pacientů, kteří se podíleli na úspěšných klinických testech daného léku. Tým lékařů Steva Jobse však mohl volit terapii podle toho, nakolik vyhovuje jeho konkrétní kombinaci genů. Kdykoli jedna léčba přestala kvůli mutaci nádoru účinkovat, lékaři mohli přejít na jinou účinnou látku – jak říkal Jobs: „Přeskočit na další list leknínu." „Buď budu jeden z prvních lidí, kteří dokážou nádoru tímto způsobem utéci, nebo budu jeden z posledních, kteří na něj zemřou," vtipkoval. Jeho předpověď se sice bohužel nesplnila, ale samotná metoda – získat všechna data a nikoli jen trochu – mu poskytla několik let života navíc.

Od části k celku

Vzorkování je produktem éry, která kladla omezení na zpracování informací. Lidé tehdy sice měřili svět, ale chyběly jim nástroje, aby získaná data analyzovali. Proto se zároveň jedná o pozůstatek této éry. Postupně mizí potíže, které souvisely s počítáním a tabulací. Senzory, přijímače GPS v mobilních telefonech, klepnutí na webové stránky i Twitter – všude se data sbírají automaticky. Počítače zároveň dokážou zpracovávat čísla stále rychleji.

Když můžeme shromažďovat velké objemy dat, princip vzorkování přestává dávat smysl. Technické nástroje na manipulaci s daty se již zásadně proměnily, ale naše metody a způsoby uvažování se přizpůsobují pomaleji.

Vzorkování je přitom vykoupeno nevýhodou, o které jsme dlouho věděli, ale ignorovali jsme ji. Dochází ke ztrátě detailů. V některých případech nemáme jinou možnost než vzorkovat. V mnoha oblastech však probíhá posun od sbírání části dat ke shromažďování co nejvíce dat, a je-li to možné, k měření všeho: *N=vše*.

Jak jsme již viděli, vztah *N=vše* znamená, že můžeme data procházet do velké hloubky. Vzorky to zdaleka tak dobře neumožňují. Za druhé připomeňme, že v naší ukázce vzorkování při volebním průzkumu jsme měli tříprocentní odchylku chyby pouze při extrapolaci na celou populaci.

V některých případech taková odchylka postačuje. Ztrácíme však detaily, jemnější strukturu a možnost podívat se podrobněji na určité podskupiny. Normální rozdělení je bohužel právě takové – normální. Skutečně zajímavá fakta se často nacházejí na místech, které vzorky nedokážou plně zachytit.

Systém Google Flu Trends se tedy nespoléhá na malý náhodný vzorek, ale místo toho používá miliardy internetových vyhledávacích dotazů z USA. Díky použití všech těchto dat namísto malého vzorku lze analýzu dostat na takovou úroveň, kdy umožňuje předpovídat šíření chřipky v určitém městě a nikoli jen v rámci státu či celé federace. Oren Etzioni z firmy Farecast zpočátku použil vzorek 12 000 datových bodů, který fungoval docela dobře. Když však přidal další data, kvalita předpovědí se zvýšila. Systém Farecast nakonec pracoval se záznamy o vnitrostátních letech na většině tras během celého roku. „Tato data jsou dočasná. Prostě je postupně shromažďujeme a přitom stále zpřesňujeme nalezené vzory," říká Etzioni.

Často je tedy možné, abychom se vyhnuli zkratce náhodného vzorkování, a místo toho můžeme usilovat o získání komplexnějších dat. Přitom potřebujeme dostatek výpočetní a úložné kapacity a špičkové nástroje, které dokážou všechna data analyzovat. Neobejdeme se také bez snadných a dostupných metod shromažďování dat. Každý z těchto úkolů v minulosti představoval nákladný problém. V nedávné době se však všechny uvedené prvky dramaticky zlevnily a zjednodušily. To, co bylo dosud doménou těch největších společností, je nyní dostupné většině z nás.

Na základě úplných dat můžeme vysledovat vazby a podrobnosti, které se jinak v záplavě informací ztrácejí. Detekce podvodů s platebními kartami je například založena na hledání anomálií, které lze přitom nejsnáze najít při analyzování všech dat namísto pouhého vzorku. Nejzajímavější jsou přitom mimořádné odchylky, které lze identifikovat jen porovnáním s masou normálních transakcí. Jedná se o problém veledat. Vzhledem k tomu, že trans s akce platebními kartami trvají krátce, musí analýza zpravidla také probíhat v reálném čase.

Firma Xoom se specializuje na mezinárodní převody peněz a podporují ji velcí hráči z oboru veledat. Analyzuje všechna data související s transakcemi,

které zpracovává. Roku 2011 systém zobrazil varování, protože zjistil poněkud nadprůměrný počet transakcí s kartami Discover Card, které byly uskutečněny ve státě New Jersey. „Systém zachytil vzor tam, kde žádný vzor neměl existovat," vysvětluje John Kunze, ředitel společnosti Xoom. Samy o sobě vypadaly jednotlivé transakce legitimně. Ukázalo se však, že za nimi stojí skupina podvodníků. Takovou anomálii bylo možné zaznamenat jen díky analýze všech dat – při vzorkování by se mohla ztratit.

Úkol zpracovat všechna data nemusí být mimořádně náročný. Veledata pokaždé nemají velký absolutní rozsah, ačkoli často tomu tak je. Systém Google Flu Trends ladí své předpovědi na základě stovek milionů testovacích matematických modelů, které pracují s miliardami datových bodů. Úplná sekvence lidského genomu čítá tři miliardy párů bází. Uvedené příklady však patří mezi veledata nikoli vzhledem k samotnému absolutnímu počtu datových bodů, tj. velikosti datové množiny. Do kategorie veledat se řadí proto, že místo použití zástupného náhodného vzorku jak systém Flu Trends, tak lékaři Steva Jobse pracovali s co nejkompletnější datovou množinou, jaká byla dostupná.

Princip *N=vše* nemusí nutně znamenat mnoho dat. Dobrým příkladem je studie, která nalezla manipulaci s výsledky zápasů v japonském národním sportu sumo. Podezření na prodané zápasy pronásledují tento císařský sport odedávna a všichni zúčastnění je vždy důrazně odmítali. Steven Levitt, ekonom na Chicagské univerzitě, zjišťoval korupci pomocí záznamů o utkáních v uplynulých více než deseti letech. Přitom analyzoval všechny výsledky z této doby. V působivém odborném článku, který vyšel v časopise *American Economic Review* a později v rámci autorovy knihy *Freakonomics* (česky jako „Špekonomie aneb Freakonomics"), spolu se svými kolegy ukázal, jak je užitečné prozkoumat všechna tato data.

Analyzovali 11 let zápasů sumo s více než 64 000 utkáními a hledali v nich anomálie. A narazili na poklad. K manipulaci s výsledky skutečně docházelo, ale nikoli tam, kde to většina lidí čekala. Místo utkání o titul šampiona, která sice zmanipulovaná být mohla, ale také nemusela, se v datech ukázalo, že se něco výstředního děje při přehlížených zápasech

na konci turnajů. Zdánlivě téměř o nic nejde, protože zápasníci již nemají naději za zisk titulu.

Sumo je však zvláštní v tom, že zápasníci potřebují vyhrát většinu zápasů v turnajích s 15 utkáními, aby si udrželi svou kategorii a příjem. To někdy vede k soubojům, kde oba protivníci mají odlišnou motivaci a zápasník se skóre 7:7 stojí proti soupeři se skóre 8:6 nebo lepším. Prvnímu sportovci na výsledku záleží hodně a pro druhého neznamená téměř nic. Jak ukázala číselná analýza, v takových případech je velmi pravděpodobné, že vyhraje zápasník, který výhru potřebuje.

Co když ten, který potřebuje vyhrát, bojuje odhodlaněji? Možná. Data však naznačují, že se děje také něco jiného. Zápasníci, kterým na výsledku více záleží, vyhrávají asi o 25 procent častěji než obvykle. Tak velký rozdíl lze těžko připsat samotnému adrenalinu. Z další analýzy dat vyplynulo, že hned při následujícím utkání stejných dvou soupeřů bylo oproti jejich budoucím soubojům mnohem pravděpodobnější, že vyhraje ten, který v posledním zápase prohrál. První vítězství tedy vypadá jako „dárek" jednoho konkurenta druhému, protože v domáckém prostředí sportu sumo se všichni dobře znají.

Tyto informace byly k dispozici odjakživa. Všichni je měli před očima. Při náhodném vzorkování výsledků zápasů by se však tyto souvislosti mohly ztratit. Autoři sice vystačili se základními statistickými vzorci, ale zpočátku nevěděli, co hledají. Nedokázali by proto správný vzorek vybrat. Místo vzorkování jej našli pomocí mnohem větší datové sady – při analýze výsledků všech zápasů. Výzkum založený na veledatech hodně připomíná rybolov: zpočátku nevíme, zda vůbec něco chytíme, ani *co* můžeme ulovit.

Datová sada nemusí zabírat několik terabajtů. U studie ze sportu sumo obsahovala celá množina méně bitů než současná digitální fotografie. Stejně jako u jiných analýz veledat se však autoři nespokojili s typickým náhodným vzorkem. Když mluvíme o veledatech, máme spíše než absolutní „velikost" na mysli velikost relativní: vzhledem k celkovému objemu dat.

Náhodné vzorkování po mnoho let představovalo výhodnou zkratku. V analogové době umožňovalo řešit problémy, které vyžadovaly analýzu rozsáhlých dat. Avšak obdobně jako při konverzi digitálního obrázku nebo

hudební skladby na menší soubor také při vzorkování dochází ke ztrátě informací. Úplná (nebo téměř kompletní) datová množina poskytuje mnohem větší svobodu zkoumat, pozorovat data z různých úhlů nebo se podrobněji zaměřit na vybraná hlediska.

Jako vhodnou analogii můžeme zmínit fotoaparát Lytro, který nezachycuje jedinou obrazovou rovinu jako konvenční přístroje, ale zaznamenává paprsky z celého světelného pole – přibližně 11 milionů. Teprve později se fotograf může rozhodnout, na který obrazový prvek v digitálním souboru zaostří. Není nutné ostřit hned při fotografování, protože díky zaznamenání kompletních světelných informací to lze provést až dodatečně. Vzhledem k tomu, že digitální fotografie obsahuje paprsky z celého světelného pole, více se blíží úplným datům. V důsledku toho jsou obrazové informace později lépe využitelné než v případě obyčejných snímků, kde musí fotograf zaostřený objekt vybrat ještě před stisknutím spouště.

Obdobně veledata jsou založena na kompletních informacích, nebo přinejmenším na jejich co největším množství. Dovolují nám proto pozorovat podrobnosti či zkoumat nové možnosti analýzy bez rizika rozostření. Nové hypotézy můžeme testovat na mnoha úrovních detailů. Díky této kvalitě veledat lze odhalit manipulaci s výsledky v zápasech sumo, sledovat šíření viru chřipky podle oblastí a bojovat s nádory léčbou, která je zacílena na konkrétní sekvenci pacientovy DNA. Při zpracování veledat lze snížit úroveň nejasnosti na minimum.

Někdy se samozřejmě bez použití kompletních dat obejdeme. Pořád žijeme ve světě, kde musíme vystačit s omezenými prostředky. Ve stále větším počtu případů však analýza všech dostupných dat dává smysl a lze ji provést i tam, kde to dosud nebylo možné.

Jednou z oblastí, kterými paradigma *N=vše* otřáslo nejvýrazněji, jsou společenské vědy. Ztratily svůj monopol na vysvětlování empirických společenských dat, protože analýza veledat bere práci zkušeným specialistům na průzkumy. Disciplíny společenských věd se ve značné míře spoléhaly na vzorkovací studie a dotazníky. Když se však data shromažďují pasivně a lidé přitom dělají to, co by dělali stejně, odpadá riziko zkreslení, které dříve k vzorkování

a dotazování patřilo. Nyní dokážeme sbírat informace, k nimž jsme se dříve nedostali, ať už se jedná o vztahy prozrazené díky mobilním telefonátům nebo nálady zjištěné z tweetů. Důležitější je, že mizí potřeba vzorkování.

Albert-László Barabási, který patří k uznávaným světovým autoritám v oboru teorie sítí, chtěl studovat interakce mezi lidmi v měřítku celé populace. Spolu se svými kolegy tedy analyzoval anonymní protokoly o mobilních telefonátech od telekomunikačního operátora, jehož služby využívá asi pětina populace neidentifikované evropské země. Jednalo se o všechny protokoly za čtyřměsíční období. Šlo o první analýzu sítě na úrovni celé společnosti, kdy datová sada odpovídala koncepci *N=vše*. Při zpracování tak velkého objemu volání mezi miliony lidí v průběhu delšího časového intervalu se podařilo získat nové poznatky, které by pravděpodobně nebylo možné zjistit jiným způsobem.

Za pozornost stojí fakt, který je v rozporu s dřívějšími menšími studiemi. Tým přišel na to, že pokud jsou ze sítě odstraněni lidé, kteří mají hodně vazeb ve své komunitě, zbývající sociální síť sice degraduje, ale neselže. Pokud na druhou stranu ze sítě vypadnou osoby, které mají vazby mimo svou bezprostřední komunitu, sociální síť se náhle rozpadne, jako by se zhroutila její struktura. Tento důležitý výsledek byl poněkud nečekaný. Koho by napadlo, že lidé s mnoha blízkými přáteli jsou pro stabilitu síťové struktury mnohem méně důležití než ti, kteří se znají se vzdálenějšími lidmi? Uvedené zjištění naznačuje, že pro skupinu a společnost jako celek má velký význam rozmanitost.

Statistické vzorkování obvykle považujeme za neměnný základ, něco jako geometrické poučky nebo gravitační zákon. Celý princip je přitom starý sotva sto let a vznikl při řešení konkrétního problému v konkrétním časovém období pod vlivem určitých technologických omezení. Tato omezení přitom již ve stejném rozsahu neexistují. Když v éře veledat volíme náhodný vzorek, chováme se tak, jako bychom v éře automobilu sahali po koňském postroji. V jistých kontextech můžeme vzorkování stále použít, ale již to nemusí být – a ani nebude – převládající způsob, kterým analyzujeme velké datové množiny. Stále častěji budeme chtít obsáhnout vše.

Kapitola 3

S chybami

Ve stále více situacích lze použít všechna dostupná data. Tato možnost ale není zadarmo. Když zvětšujeme objem dat, otevíráme tím dveře nepřesnosti. Chybné hodnoty a poškozené bity se samozřejmě do databází dostávaly odjakživa. Vždy jsme je však považovali za problém a snažili jsme se jich zbavit – mimo jiné i proto, že jsme mohli. V žádném případě jsme je však nechtěli pokládat za něco nevyhnutelného a smířit se s nimi. V tom spočívá jedna ze zásadních změn při přechodu od malých dat k velkým.

Ve světě malých dat bylo přirozené a logické omezovat chyby a snažit se o vysokou kvalitu dat. Vzhledem k tomu, že shromážděných informací bylo poměrně málo, dávali jsme si záležet na tom, aby získané údaje byly co nejblíže skutečnosti. Celé generace vědců optimalizovaly své přístroje tak, aby byla jejich měření stále přesnější – ať už šlo o určování pozic vesmírných těles nebo velikost objektů v hledáčku mikroskopu. Ve světě vzorkování měla posedlost přesností ještě větší význam. Při analýze omezeného počtu datových bodů se chyby mohou zvýrazňovat, což potenciálně snižuje přesnost celkových výsledků.

Historicky vzato byly největší lidské úspěchy při dobývání světa založeny na jeho měření. Úsilí o přesnost má své kořeny v Evropě v polovině 13. století, kdy astronomové a jiní učenci začali stále důkladněji kvantifikovat čas a prostor – slovy historika Alfreda Crosbyho „měřit realitu".

Vycházeli přitom z předpokladu, že dokážou-li jistý jev změřit, mohou mu porozumět. Později se měření stalo základem vědecké metody pozorování a vysvětlení, kdy vědci dokázali kvantifikovat, zaznamenávat a prezentovat reprodukovatelné výsledky. „Měřit znamená vědět," prohlásil lord Kelvin. Měření poskytovalo základ autority. „Vědění je moc," tvrdil Francis Bacon. Ve stejné době matematici a pozdější pojistní technici a účetní vyvíjeli metody, které umožnily přesně sbírat, zaznamenávat a udržovat data.

Ve Francii 19. století, která v té době stála na čele světového vědeckého pokroku, vznikl systém přesně definovaných měrných jednotek, pomocí nichž bylo možné popisovat čas, prostor a další fyzikální veličiny. Tento standard postupně začaly přebírat i jiné země. Vývoj dospěl až k vytvoření mezinárodně uznávaných etalonů, na které se odvolávaly mezistátní smlouvy. Éra měření tím dospěla ke svému vrcholu. O pouhé půlstoletí později – ve 20. letech 20. století – však sen o úplném a dokonalém měření navždy pohřbily objevy kvantové mechaniky. Mimo poměrně úzký okruh fyziků však inženýři a vědci i nadále vycházeli z překonané koncepce bezchybného měření. Ta ve světě podnikání svůj vliv dokonce rozšiřovala, protože exaktní obory matematiky a statistiky se začaly prosazovat ve všech oblastech obchodu.

V mnoha situacích, které dnes řešíme, však můžeme připuštěním nepřesnosti či chybovosti získat a nikoli ztratit. Něco za něco: když zmírníme své nároky na přípustný počet chyb, můžeme shromáždit mnohem více dat. Neplatí jen „raději více dat než méně“, ale někdy dokonce platí, že „raději více horších dat než méně lepších“.

Musíme se vypořádat s několika druhy chybovosti. Termín „chybovost“ může popisovat prostý fakt, že při rostoucím počtu datových bodů jsou chyby častější. Jestliže tedy napětí mostu měříme tisíckrát častěji, roste i pravděpodobnost chyb. Chybovost však můžeme zvýšit i kombinací různých typů informací z odlišných zdrojů, jejichž struktura si někdy přesně neodpovídá. Když budeme například charakterizovat telefonické stížnosti v zákaznickém centru pomocí softwaru na rozpoznávání řeči a porovnáme tato data s tím, jak dlouho operátorům trvá vyřízení příslušných hovorů, můžeme získat nedokonalou, ale užitečnou představu o situaci. Pojem chybovost označuje také inkonzistenci formátování, kvůli níž je potřeba data před zpracováním „vyčistit“. Jak upozorňuje expert na veledata DJ Patil, existuje spousta způsobů, jak můžeme označit společnost IBM – například I.B.M., IBM, International Business Machines nebo „Velká Modrá“. Chybovost se může projevit také během získávání či zpracování dat, protože přitom data transformujeme a měníme je na něco jiného. Příkladem je třeba analýza nálad ve zprávách na

Twitteru při předpovídání kasovních příjmů hollywoodských filmů. Samotná chybovost je chaotická.

Předpokládejme, že potřebujeme měřit teplotu na vinici. Máme-li pouze jeden teplotní senzor pro celý pozemek, potřebujeme, aby byl přesný a fungoval nepřetržitě. Žádnou chybovost si nemůžeme dovolit. Jestliže oproti tomu vybavíme senzorem každou rostlinu, můžeme použít levnější a jednodušší senzory (za předpokladu, že nezpůsobují systematickou odchylku měření). Některé ze senzorů budou pravděpodobně občas oznamovat nesprávné hodnoty, takže získáme méně přesný – tj. „chybovější" – datovou sadu, než jakou by zajistil jediný přesný senzor. Libovolné konkrétní měření sice nemusí být správné, ale ze souhrnu mnoha údajů dostaneme podrobnější obrázek. Datová sada obsahuje více datových bodů, takže nám poskytne mnohem cennější výsledky, jejichž hodnota pravděpodobně převáží nad nižší přesností.

Předpokládejme nyní, že zvýšíme frekvenci odečtu senzorů. Pokud provádíme jedno měření za minutu, můžeme si být poměrně jisti, že data budou přicházet v nepřerušené číselné řadě. Nastavíme-li však místo toho deset nebo sto měření za sekundu, pravděpodobnost, že celá sekvence bude souvislá, kvůli tomu klesne. Jak informace putují po síti, může se některý záznam zpozdit a dorazit mimo pořadí nebo se v záplavě dat může jednoduše ztratit. Informace budou poněkud méně přesné, ale díky jejich velkému objemu stojí za to, abychom z nároků na úplnost slevili.

V prvním případě jsme obětovali přesnost jednotlivých datových bodů, abychom jich mohli měřit více, a za to jsme zjistili podrobnosti, které bychom jinak nemohli pozorovat. V druhém případě jsme se vzdali přesnosti, abychom mohli zvýšit frekvenci, a díky tomu jsme mohli pozorovat změny, kterých bychom si jinak nevšimli. Je sice pravda, že pokud investujeme dostatek prostředků, můžeme chyby překonat – nakonec na burze v New Yorku probíhá každou sekundu 30 000 obchodů, na jejichž správném pořadí hodně záleží – ale v mnoha případech je vhodnější chyby tolerovat, než se jim snažit předcházet.

Můžeme například připustit určitou chybovost jako cenu za větší rozsah dat. Technologická konzultační firma Forrester to vysvětluje takto: „Dvě

a dvě se někdy rovná 3,9 a tento výsledek je docela dobrý.“ Data samozřejmě nemohou být zcela chybná, ale chceme-li zjistit obecný trend, můžeme nároky na přesnost trochu snížit. Při zpracování veledat se čísla mění z přesných hodnot spíše na pravděpodobnostní údaje. Na tuto změnu není snadné si zvyknout a přináší nové problémy, kterými se budeme zabývat v další části této knihy. Prozatím však stačí poznamenat, že při zvyšování měřítka se často musíme smířit s chybovostí.

Podobný posun můžeme sledovat v tom, jak v informatice vzrůstá význam většího množství dat oproti jiným zlepšením. Každý ví, jak se v posledních letech zvýšil výpočetní výkon. Jeho nárůst předpovídá Moorův zákon, který tvrdí, že počet tranzistorů na čipu se zdvojnásobuje přibližně každé dva roky. Díky tomuto neustálému zdokonalování se počítače zrychlují a paměť je stále dostupnější. Méně je však známo, že se také zvýšil výkon algoritmů, které zajišťují funkci mnoha počítačových systémů – v mnoha oblastech šlo přitom o významnější pokrok, než jakého dosáhly procesory při svém vývoji podle Moorova zákona. Mnohé výhody, které společnost získává díky veledatům, však nejsou způsobeny ani tak rychlejšími čipy nebo lepšími algoritmy, ale spíše tím, že je k dispozici více dat.

Šachové algoritmy se například v posledních několika dekádách změnily jen málo, protože pravidla hry jsou dobře známa a pevně omezena. Šachové programy v současnosti hrají lépe než v minulosti zčásti proto, že lépe zvládají koncovku. A to je dáno prostě tím, že systémy mají k dispozici více dat. Koncovky, kde na šachovnici zbývá šest a méně kamenů, byly v praxi analyzovány kompletně a všechny možné tahy (*N=vše*) jsou reprezentovány v masivní tabulce, která v nekomprimovaném stavu zabírá více než jeden terabajt. Šachové počítače tak mohou koncovku odehrát bez chyby. Žádný člověk nikdy tento systém nedokáže přehrát.

To, do jaké míry více dat překonává lepší algoritmy, lze přesvědčivě ukázat v oblasti zpracování přirozeného jazyka. Tento obor se zabývá tím, jak se počítače učí analyzovat slova, která používáme v každodenní konverzaci. Kolem roku 2000 hledali výzkumníci ve společnosti Microsoft Michele Banko a Eric Brill způsob, jak zdokonalit kontrolu gramatiky, která je součástí

firemního programu Word. Váhali nad tím, zda by bylo vhodnější věnovat úsilí na zlepšování stávajících algoritmů, hledání nových metod nebo doplňování pokročilejších funkcí. Než se pustili jednou z těchto cest, rozhodli se vyzkoušet, co se stane, když stávajícím metodám poskytnou mnohem více dat. Většina algoritmů strojového učení byla postavena na textových korpusech, které zahrnovaly nejvýše milion slov. Banko a Brill vzali čtyři běžné algoritmy a postupně jim nabídli až o tři řády více dat: nejdříve 10 milionů, pak 100 milionů a nakonec miliardu slov.

Výsledky je zaskočily. Jak přicházela další a další data, výkon všech čtyřech typů algoritmů se dramaticky zlepšoval. Jednoduchý algoritmus, který na základě půl milionu slov poskytoval nejhorší výsledky, dokonce po zpracování miliardy slov překonal všechny ostatní. Jeho relativní přesnost se ze 75 procent zvýšila nad 95 procent. Naopak algoritmus, který byl optimální při minimálním množství dat, se u větších datových objemů ukázal jako nejhorší, ačkoli se stejně jako zbývající algoritmy hodně zlepšil a z relativní přesnosti asi 86 procent se dostal k přibližně 94procentní přesnosti. „Naše výsledky naznačují, že může být vhodné přehodnotit poměr časových a finančních investic do vývoje algoritmů a do vývoje korpusů," napsali Banko a Brill ve svém odborném článku na toto téma.

Takže čím více, tím lépe. A více dat někdy vyhrává nad větší inteligencí. Jak je to tedy s chybovostí? Několik let poté, co Banko a Brill zasypali algoritmy spoustou dat, výzkumníci ve společnosti Google uvažovali podobným způsobem, jen v ještě větším měřítku. Místo miliardy slov testovali algoritmy s tisíckrát větším objemem. Společnost Google přitom nezkoušela vyvinout modul na kontrolu gramatiky, ale snažila se rozlousknout ještě tvrdší oříšek: překlad z jednoho jazyka do jiného.

O takzvaných strojových překladech snili počítačoví vizionáři již od počátků svého oboru ve 40. letech 20. století, kdy se výpočetní stroje skládaly z elektronek a zabíraly celé místnosti. Úsilí o strojový překlad získalo na významu zejména během studené války, kdy Spojené státy zachycovaly značné objemy textů a zvukových nahrávek v ruštině, ale neměly dostatek pracovníků, kteří by tyto materiály rychle přeložili.

Informatici nejdříve zkusili zkombinovat gramatická pravidla a dvojjazyčný slovník. Počítač IBM přeložil roku 1954 do angličtiny šedesát ruských frází. Stačilo mu k tomu přitom 250 dvojic slov ve slovníku a šest gramatických pravidel. Výsledky vypadaly velmi slibně. Pomocí děrných štítků byla do počítače IBM 701 vložena věta „*Mi pyeryedayem mislyi posryedstvom ryechyi*" a počítač odpověděl: „We transmit thoughts by means of speech" (Předáváme myšlenky pomocí řeči.). Šedesát vět bylo „hladce přeloženo", psalo se v tiskové zprávě IBM vydané při této příležitosti. Šéf výzkumného programu Leon Dostert z Georgetownské univerzity předpovídal, že strojový překlad bude „hotovou věcí" během „pěti, možná tří následujících let".

Ukázalo se však, že počáteční úspěch byl velice klamný. Roku 1966 musela komise expertů z oboru strojového překladu přiznat porážku. Problém byl obtížnější, než zpočátku připouštěli. Mají-li počítače překládat, nestačí, aby se naučily pravidla, ale musí se naučit také výjimky. Překlad nespočívá jen v zapamatování a vybavování slov. Je při něm také potřeba vybírat správná slova z mnoha variant. Znamená „*bonjour*" skutečně „dobré ráno"? Nebo spíše „dobrý den", „nazdar" či „ahoj"? Odpověď zní: záleží na kontextu…

Koncem 80. let dostali výzkumníci ve společnosti IBM novátorský nápad. Místo toho, aby se snažili počítači předat explicitní gramatická pravidla spolu se slovníkem, rozhodli se ponechat počítač, aby pomocí statistické pravděpodobnosti určil nejvhodnější slovo nebo frázi pro slovo či frázi jiného jazyka. Projekt Candide společnosti IBM v 90. letech vycházel z deseti let záznamů z jednání kanadského parlamentu, které byly publikovány ve francouzštině a angličtině. Jednalo se asi o tři miliony párů vět. Vzhledem k tomu, že se jednalo o oficiální dokumenty, měly překlady mimořádně vysokou kvalitu. A podle dobových standardů šlo o značné množství dat. Statistický strojový překlad, jak se tomuto postupu začalo říkat, chytře převedl problém překladu na jeden rozsáhlý matematický problém. Zdálo se přitom, že to funguje. Počítačové překlady se náhle výrazně zlepšily. Po prvním úspěchu této nové koncepce však společnost IBM přes značné investice dosahovala již jen malých pokroků. Nakonec se rozhodla projekt ukončit.

Za necelých deset let – roku 2006 – se však do překladů pustila společnost Google, protože to zapadalo do její vize „uspořádat světové informace a zajistit, aby byly univerzálně přístupné a užitečné“. Místo úhledně přeložených stránek textu ve dvou jazycích si společnost Google posloužila větším, ale zato mnohem chaotičtějším datovým souborem: celým globálním Internetem doplněným o další zdroje. Její systém při svém učení hltal všechny překlady, které dokázal najít. Procházel firemní weby s více jazykovými mutacemi, odpovídající překlady oficiálních dokumentů a zprávy mezivládních organizací, jako je OSN a Evropská unie. Do projektu společnosti Google se dostaly i překlady knih z jejího projektu skenování knih. Jak vysvětluje vedoucí týmu Google Translate Franz Josef Och, který patří mezi uznávané autority v oboru: tam, kde systém Candide pracoval se třemi miliony pečlivě přeložených vět, systém Google využíval miliardy stran překladů značně kolísavé kvality. Jeho korpus s tisícem miliard slov obsahoval 95 miliard anglických vět, i když pochybné úrovně.

Navzdory svým neuspořádaným vstupům funguje služba Google nejlépe ze všech. Její překlady jsou přesnější než u jiných systémů (i když zůstávají značně nedokonalé). A je mnohem širší než její konkurenti. V polovině roku 2012 její datová základna zahrnuje více než 60 jazyků. Dokáže dokonce simultánně překládat podle hlasového vstupu ve 14 jazycích. Vzhledem k tomu, že jazyky považuje prostě za chaotická data, která umožňují vyhodnocovat pravděpodobnosti, může dokonce překládat mezi jazyky, kde pro vývoj systému existuje velmi málo přímých překladů – například z hindštiny do katalánštiny. Takové jazyky propojuje pomocí angličtiny. Popsaný přístup je mnohem pružnější než jiné přístupy, protože dokáže doplňovat a vyřazovat slova tak, jak přicházejí do módy či naopak ustupují.

Překladový systém společnosti Google neposkytuje přijatelné výsledky proto, že má chytřejší algoritmy. Funguje dobře proto, že jej jeho tvůrci podobně jako Banko a Brill z Microsoftu nakrmili více daty – a nejen vysoce kvalitními. Díky tomu, že společnost Google připustila chybovost svých vstupů, mohla zpracovat datovou sadu, která byla *deset tisíckrát* větší než v případě systému IBM Candide. Korpus s tisícem miliard slov, který

společnost Google vytvořila roku 2006, byl sestaven ze změti internetového obsahu – jakýchsi „divokých dat". Na základě tohoto korpusu se systém naučil vypočítat například pravděpodobnost toho, že jisté anglické slovo následuje po jiném. Tento korpus ani zdaleka nepřipomínal svého prapředka, slavný Brownův korpus z 60. let, který sestával z jednoho milionu anglických slov. Díky větší datové sadě bylo možné dosáhnout při zpracování přirozeného jazyka významného pokroku, který následně umožnil rozvoj systémů rozpoznávání hlasu a počítačového překladu. „Jednoduché modely se spoustou dat dosahují lepších výsledků než propracovanější modely založené na menším množství dat," napsal spolu se svými kolegy expert na umělou inteligenci ve společnosti Google Peter Norvig v článku s názvem „The Unreasonable Effectiveness of Data" (Nesmyslná efektivita dat).

Norvig se svými spoluautory vysvětluje, že klíčem je právě chaotičnost: „V jistém smyslu tento korpus představuje oproti Brownovu korpusu krok zpět: pochází z nefiltrovaných webových stránek, a obsahuje proto neúplné věty, překlepy, gramatické chyby a všechny možné další vady. Neobsahuje pečlivě zkontrolované uvozovky, které by vyznačovaly mluvenou řeč. Nad těmito nevýhodami však převažuje fakt, že je milionkrát větší než Brownův korpus."

Raději více horších dat než méně lepších

Analytici, kteří byli zvyklí na klasické vzorkování a celý život se snažili chybám předcházet a odstraňovat je, si na nepořádek v datech zvykají jen těžko. Při sbírání vzorků usilují o co nejnižší frekvenci chyb, a než oznámí své výsledky, testují, zda jejich vzorky nejsou zkreslené. Používají více strategií redukce chyb: mimo jiné kontrolují, zda vzorky sbírali speciálně školení experti podle přesného protokolu. Implementace takových strategií se prodražuje i v případě omezeného počtu datových bodů a u veledat jsou téměř neproveditelné. Kromě toho, že by byly příliš nákladné, také v tomto měřítku prakticky nelze dodržet přísné standardy sběru dat. Dokonce ani vyloučení lidského faktoru by problém nevyřešilo.

Při přechodu do světa veledat musíme změnit své uvažování ohledně výhod přesnosti. Pokud bychom chtěli aplikovat konvenční přístup k měření na digitální a propojený svět 21. století, jednalo by se o zásadní nedorozumění. Jak jsme již uvedli výše, posedlost přesností je pozůstatkem analogové doby s nedostatkem informací. Když bylo dat málo, každý datový bod byl kriticky důležitý. Věnovali jsme proto značnou péči tomu, aby žádný bod celou analýzu nezkreslil.

V současnosti již takový hlad po informacích neznáme. Když pracujeme se stále podrobnějšími datovými množinami, které nepostihují jen malou výseč zkoumaného jevu, ale jeho podstatnou část nebo celek, nemusíme se již tolik starat o to, zda celou analýzu nezkreslí jednotlivé datové body. Místo toho, abychom se s vynaložením stále větších nákladů pokoušeli vymýtit každý nepřesný bit, při výpočtech zohledňujeme i chybovost.

Příkladem je nasazení senzorů v továrnách. V celé rafinerii Cherry Point společnosti BP v Blaine ve státě Washington jsou nainstalovány bezdrátové senzory, které tvoří neviditelnou síť a produkují velké objemy dat v reálném čase. Měřené hodnoty mohou být pochopitelně zkresleny prostředím, kde panují vysoké teploty a fungují elektrická zařízení. Tyto nedostatky jsou však vyváženy značným množstvím informací, které drátové i bezdrátové senzory poskytují. Pouhé zvýšení frekvence měření a počtu senzorů, kde probíhá měření, může poskytnout značné výhody. Díky tomu, že se společnost BP rozhodla měřit zatížení potrubí nepřetržitě a nikoli pouze v určitých intervalech, zjistila, že některé druhy ropy jsou korozivnější než jiné. Z menší datové sady by tento poznatek nezískala a příslušný problém by tedy nemohla řešit.

Když je dat značně více a jedná se o data nového typu, v některých případech již neusilujeme o přesnost. Stačí nám, že dokážeme sledovat obecný trend. Přecházíme-li na vyšší měřítko, mění se nejen očekávání na přesnost, ale také praktické možnosti k dosažení přesnosti. Na první pohled to sice může vypadat paradoxně, ale pokud data považujeme za něco nedokonalého a nepřesného, můžeme odvozovat lepší předpovědi, a tedy i lépe porozumět okolnímu světu.

Sluší se poznamenat, že chybovost nepostihuje jen veledata. Spíše se jedná o produkt nedokonalosti nástrojů, kterými měříme, zaznamenáváme a analyzujeme informace. Pokud bychom dokázali vyvinout dokonalé technologie, problém s nepřesností by zmizel. Dokonalost však zůstává jen ideálem. Chybovost proto musíme brát jako realitu, kterou je nutné zvládat. A pravděpodobně se s ní budeme potýkat ještě dlouho. Často se ekonomicky nevyplatí s vynaložením velkého úsilí zvyšovat přesnost, protože mnohem výhodnější je získat řádově větší objem dat. Podobně jako statistici minulé éry přestali sbírat větší vzorky a dali přednost vyšší náhodnosti, můžeme se dnes výměnou za více dat smířit s trochou nepřesnosti.

Zajímavým příkladem je projekt BPP (Billion Prices Project). Americký úřad pracovních statistik každý měsíc publikuje index spotřebitelských cen, který slouží k výpočtu inflace. Tato hodnota má klíčový význam pro investory i podniky. Při rozhodování o zvýšení či snížení úrokových sazeb ji bere v úvahu i americká centrální banka. Podniky podle inflace zvyšují mzdy. Federální vláda na základě inflace valorizuje platby sociálního zabezpečení a upravuje vyplácené úroky některých dluhopisů.

Aby mohl úřad pracovních statistik tuto hodnotu určit, zaměstnává stovky pracovníků, kteří volají, faxují a navštěvují obchody a kanceláře v 90 městech po celém území USA a podávají hlášení o přibližně 80 000 cenách nejrůznějších produktů: od rajčat až po jízdné v taxislužbě. Tato činnost stojí asi 250 milionů dolarů ročně. Daňoví poplatníci za ty peníze dostávají data, která jsou úhledná, čistá a uspořádaná. Již při zveřejnění jsou však tato čísla několik týdnů stará. Jak se ukázalo během finanční krize v roce 2008, zpoždění několik týdnů může být až příliš dlouhé. Řídicí pracovníci potřebují mít čísla o inflaci k dispozici rychleji, aby dokázali lépe reagovat. Konvenční metody orientované na vzorkování a přesné výpočty cen však potřebné rychlosti nedokážou dosáhnout.

Dva ekonomové z Massachusettského technologického institutu, Alberto Cavallo a Roberto Rigobon, přišli s veledatovou alternativou a vydali se přitom mnohem chaotičtější cestou. Pomocí softwaru, který procházel webové stránky, shromáždili půl milionu cen produktů, které se v USA každého dne