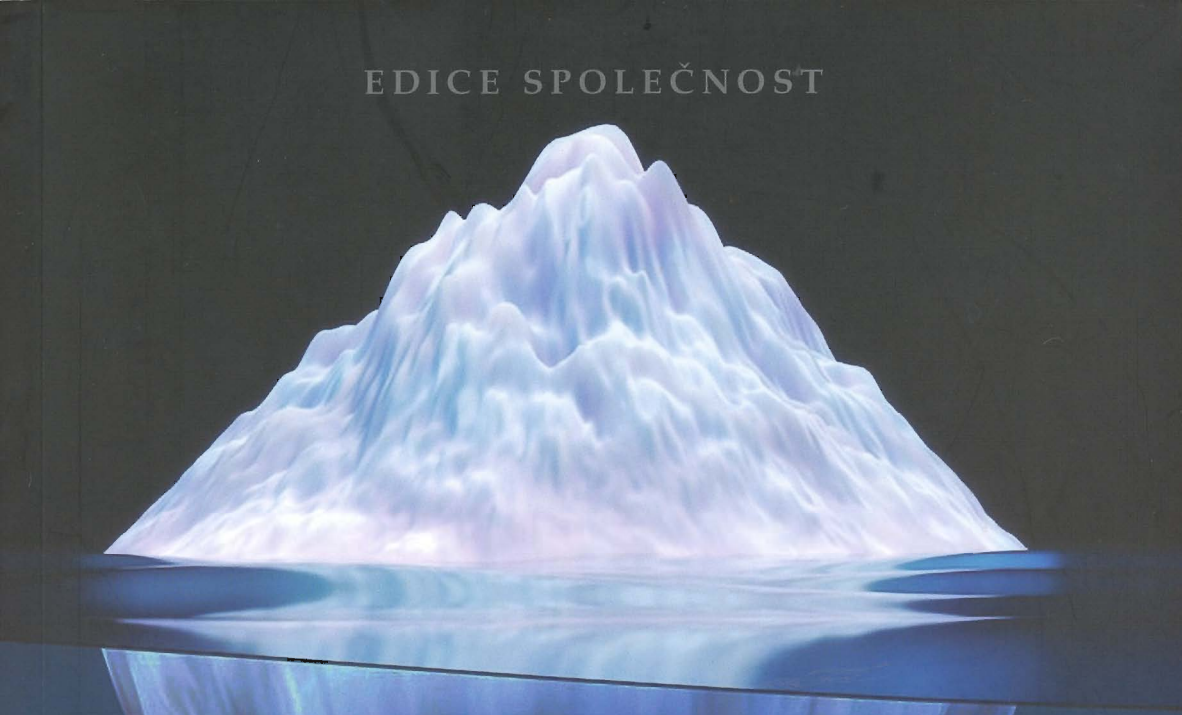
EDICE SPOLEČNOST



TEMNÁ DATA

PROČ ZÁLEŽÍ NA TOM, CO NEVÍME

• DAVID J. HAND •

academia

V éře velkých dat je snadné podlehnout dojmu, že máme všechny informace, které ke správnému rozhodování potřebujeme. Ve skutečnosti však nejsou data, která máme, nikdy úplná – obvykle jde o pouhou špičku ledovce. Stejně jako většina vesmíru sestává z temné hmoty, neviditelné a přesto přítomné, je infomační vesmír plný temných dat, která k vlastní škodě přehlížíme.
V *Temných datech* nás datový expert David Hand bere na fascinující a objevnou cestu po světě dat, která nevidíme. Kniha *Temná data* zkoumá množství způsobů, jimiž nás mohou chybějící data zaslepit před skutečností a vést nás k závěrům a skutkům, jež jsou pomýlené, nebezpečné, ba dokonce katastrofální. Na základě zkoumání celé řady skutečných případů, od zkázy raketoplánu *Challenger* až po složité finanční machinace, podává Hand ve své knize praktickou taxonomii různých druhů temných dat a situací, v nichž vznikají, abychom se je naučili rozpoznat a ovládat. Učí nás tak nejen jak se mít na pozoru před problémy, které představuje to, co nevíme, ale také jak využít temných dat ke své vlastní výhodě tím, že lépe porozumíme tomu, o čem rozhodujeme. A dnes se na základě dat rozhodujeme úplně všichni. Kniha *Temná data* nám ukazuje, jak se pokud možno nerozhodovat špatně.

DAVID J. HAND (*1950)
je britský statistik. Ve výzkumu se zaměřuje na multivariační statistiku, klasifikační metody, rozpoznání vzorů a na komputační statistiku. Kromě řady odborných publikací o statistice, finančnictví a datech je také autorem několika popularizačních prací jako *Statistika: Krátký úvod* či *Princip nepravděpodobnosti: Proč se dějí souhry náhod, zázraky i vzácné jevy.* V roce 1991 založil časopis Statistika a výpočetní technika (*Statistics and Computing*), který vloni oslavil 30 let na scéně. Za svou činnost publikační a pedagogickou dostal David J. Hand řadu ocenění a poct – v roce 2013 i Řád britského impéria (OBE).

Přeložil David Vichnar

Obsah

Temná data: svět kolem nás je utvářen tím, co nevidíme

Přízrak dat

Na úvod jednu anekdotu. Jednou jsem takhle kráčel po silnici a narazil na staršího pána, který za sebou uprostřed vozovky s odstupem zhruba patnácti metrů sypal hrstičky prachu. Ptám se ho, co to dělá.

„To je prášek proti slonům," on na to. „Nesnášejí ho, vyhýbají se mu jak čert kříži."

„Ale vždyť tady žádní sloni nejsou," namítl jsem.

„Přesně tak!" odvětil. „Účinkuje bezvadně."

A teď na mnohem vážnější notu. Spalničky každoročně zabijí téměř 100 000 lidí. Jeden z 500 nemocných zemře na komplikace, další utrpí nevratnou ztrátu sluchu či poškození mozku. Ve Spojených státech je to naštěstí vzácné onemocnění, například v roce 1999 bylo hlášeno pouhých 99 případů. Avšak v lednu 2019 se ve státě Washington objevilo tolik případů spalniček, že byl vyhlášen stav ohrožení, a dramatický nárůst případů zaznamenali také v další státech.[1] Podobný vzorec se vyskytl i jinde. Nemoc propukla také na Ukrajině, kde v polovině února 2019 zaznamenali 21 000 případů.[2] V Evropě bylo 25 863 případů v roce 2017, ale roku 2018 už přes 82 000.[3] Jen Rumunsko hlásilo od 1. ledna 2016 do konce března 2017 přes 4000 případů a 18 úmrtí na toto onemocnění.

Spalničky jsou choroba obzvlášť zákeřná, jelikož se šíří nepozorovaně a symptomy začnou být patrné až několik týdnů po nákaze. Umějí proklouznout obranou těla a než se nadějete, už jste nemocní.

Nemoci lze však také předcházet. Před nebezpečím nákazy spalničkami nás ochrání prosté očkování. A národní imunizační programy jako ten, který se provádí ve Spojených státech, slaví obrovský úspěch – tak velký, že většina rodičů v zemích s těmito programy nikdy neviděla ani nezažila hrozné následky takových preventabilních onemocnění.

Když se rodičům doporučuje, aby nechávali své děti očkovat proti nemoci, kterou nikdy neviděli ani o ní neslyšeli, jíž se nenakazil ani nikdo

z jejich známých či sousedů a o níž Centrum pro kontrolu a prevenci nemocí prohlásilo, že se ve Spojených státech nevyskytuje, pak zcela přirozeně berou takové rady s notnou rezervou.

Očkovat proti něčemu, co neexistuje? To máte jako s tím práškem proti slonům.

Až na to, že na rozdíl od slonů v tomto případě rizika existují a jsou skutečná. Jde jen o to, že informace a data, která k rozhodování potřebují, tito rodiče nemají – a tak se riziko stává neviditelným.

Pro různé druhy chybějících dat používám obecný termín *temná data*. Temná data jsou nám skryta, což samo o sobě představuje riziko, že cosi nepochopíme, vyvodíme nesprávné závěry a budeme chybně rozhodovat. Zkrátka naše nevědomost znamená, že se nejspíš zmýlíme.

Termín „temná data" vznikl analogicky k pojmu „temná hmota" z fyziky. Zhruba 27 procent vesmíru sestává z této záhadné substance, která nereaguje na světlo ani na jiné elektromagnetické záření, takže není vidět. Jelikož je temná hmota neviditelná, neměli astronomové dlouho o její existenci ani ponětí. Pak se ale při pozorování pohybu galaxií přišlo na to, že vzdálenější hvězdy se nepohybují pomaleji než hvězdy blíž středu, což je v rozporu s očekáváním vycházejícím z našeho chápání gravitace. Rotační anomálii lze vysvětlit předpokladem, že galaxie mají víc hmoty, než kolik jí můžeme vidět teleskopy při pozorování hvězd a dalších objektů. A protože tuto „větší" hmotu nelze spatřit, nazývá se temnou hmotou. Odhaduje se dokonce, že naše domovská galaxie Mléčná dráha obsahuje zhruba desetkrát víc temné hmoty než hmoty běžné.

Temná data a temná hmota se chovají obdobně: taková data nevidíme, nemáme o nich žádné záznamy, a přesto mohou výrazně ovlivnit naše závěry, rozhodování a činy. A jak nám ukážou pozdější příklady, pokud si nejsme vědomi toho, že tam venku číhá cosi neznámého, může to mít katastrofální, ba dokonce fatální následky.

Cílem této knihy je zjistit, jak a proč temná data vznikají. Podíváme se na různé druhy temných dat a uvidíme, co je vyvolává. Dozvíme se, jaké prvořadé kroky můžeme učinit, abychom se vzniku temných dat vyvarovali. Uvidíme, co lze dělat při zjištění, že čelíme temným datům, která jsou nám skryta. Nakonec také poznáme, že když jsme dost chytří, dokážeme temná data někdy i využívat. Jakkoli se to může zdát zvláštní a paradoxní, naše vlastní neznalost a pohled na temná data mohou přispět k lepším rozhodnutím a činům. Prakticky vzato to znamená, že díky uvážlivému nakládání s neznámem můžeme vést zdravější život, vydělávat víc peněz a méně riskovat. To neznamená, že bychom měli informace skrývat před ostatními (i když, jak také uvidíme, úmyslně skrytá data jsou zcela běžným druhem temných dat). Jde o cosi mnohem důmyslnějšího a znamená to, že z toho dokáže profitovat kdokoli.

Temná data vznikají v rozmanitých tvarech a formách i z mnoha důvodů a tato kniha přichází s taxonomií takových důvodů, tedy *typů* temných dat, označených jako *TD-typ x*, tedy *„temná data – typ x"*. Celkem jich je 15. Moje taxonomie není dokonalá. Vzhledem k tomu, jak bohaté příčiny vznik temných dat má, by úplná taxonomie nejspíš byla zhola nemožná. Navíc jakýkoli jednotlivý příklad temných dat dokáže ilustrovat účinek víc než jednoho *TD-typu* současně – *TD-typy* mohou spolupracovat a dokonce se i spojit v nešťastnou synergii. Nicméně povědomí o těchto *TD-typech* a zkoumání příkladů, jak se temná data projevují, vás může dobře vybavit, abyste problémy zaznamenali a ochránili se před jejich nebezpečím. Na konci této kapitoly uvádím výčet *TD-typů* v pořadí zhruba podle podobnosti, podrobněji je popisuji v kapitole 10. V knize průběžně zmiňuji některé případy konkrétních *typů*. Záměrně to však nedělám vyčerpávajícím způsobem – to by působilo poněkud rušivě.

Abychom pokročili dál, vezměme si další příklad.

V medicíně se traumatem rozumí vážný úraz s možnými zásadními dlouhodobými následky. Jde o jednu z nejzávažnějších příčin „zkrácení života" předčasným úmrtím a postižením a o nejběžnější příčinu úmrtí pro věkovou skupinu do 40 let. Databáze Sítě pro sledování a výzkum úrazů (TARN) je největší databází zdravotních traumat v Evropě. Údaje o úrazech dostává z víc než 200 nemocnic, mezi nimiž je přes 93 % takových zařízení v Anglii a Walesu i nemocnice v Irsku, Nizozemsku a Švýcarsku. Jde zjevně o velmi bohatý zdroj dat pro studium prognóz a účinnosti intervencí při úrazech.

Dr. Evgeny Mirkes se spolu s kolegy z univerzity v britském Leicesteru na některá data z této databáze podíval.[4] Ze 169 559 případů, na které se zaměřili, mělo neznámý výsledek 19 289 úrazů. „Výsledkem" se ve výzkumu traumatu rozumí, zda pacient po úrazu přežil aspoň třicet dní. Takže třicetidenní přežití po úrazu nebylo známo u víc než 11 procent pacientů. Tento příklad dokládá běžnou podobu temných dat – náš *TD-typ 1: Data, o nichž víme, že chybějí*. Víme, že tito pacienti jistě nějaký výsledek měli – jen nevíme, jaký.

Možná si pomyslíte, že na tom nesejde – tak prostě analyzujme těch 146 270 pacientů, u nichž výsledek známe, a založme poznatky a prognózy na nich. Konec konců 146 270 je velká hodnota – v oblasti medicíny jsou to „velká data" –, takže můžeme s jistotou říci, že závěry založené na těchto datech budou správné.

Skutečně? Co když se těch chybějících 19 289 případů od ostatních významně liší? Jistě, odlišují se už tím, že mají neznámý výsledek, takže

je namístě předpokládat, že se mohou lišit i jinak. Tudíž jakákoli analýza oněch 146 270 pacientů se známým výsledkem může být vzhledem k celému vzorku úrazových pacientů zavádějící. Tím pádem mohou být kroky založené na takových analýzách chybné, mohou vést k mylné prognóze, předepsání nesprávných léků a k nevhodnému léčebnému režimu s neblahými, či dokonce fatálními následky pro pacienty.

V zájmu záměrně nereálného a extrémního příkladu předpokládejme, že všech 146 270 pacientů se známými výsledky přežilo a zotavilo se a že všech 19 289 pacientů s neznámými výsledky zemřelo do dvou dnů po přijetí do nemocnice. Kdybychom neznámé výsledky ignorovali, oprávněně dojdeme k závěru, že nebylo čeho se obávat, neboť všichni pacienti se po úrazu zotavili. Na základě toho bychom přestali léčit všechny přivezené úrazy v očekávání, že se zotaví zcela přirozeně. A pak bychom byli zděšení a rozhození z toho, že nám přes 11 procent pacientů umírá.

Než budu pokračovat, rád bych čtenáře o něčem ujistil. Můj extrémní příklad představuje nejhorší možný scénář – dost dobře můžeme očekávat, že tak špatné to ve skutečnosti není – a dr. Mirkes a jeho kolegové jsou odborníky na analýzu chybějících dat. Dobře si uvědomují nebezpečí, a aby takové problémy překonali, vyvíjejí k tomu statistické metody; dále v knize podobné metody popisuji. Z uvedeného příkladu bychom si však měli odnést, že *skutečnost nemusí být taková, jak vypadá*. Kdybychom si měli z celé knihy odnést jedno jediné, blížilo by se to tomuhle: i když je výhodou mít hodně dat – tedy „velká data" –, rozhodně to není všechno. A to, co nevíte, tedy data, která nemáte, může být pro pochopení toho, co se děje, ještě důležitější než data, která máte. V každém případě, jak ještě uvidíme, netýkají se problémy temných dat výhradně velkých dat: vznikají rovněž u malých datových souborů. Jsou všudypřítomné.

Moje fabule s databází TARN může být přehnaná, má však posloužit jako varování. Výsledky 19 289 pacientů množná nebyly přesně zaznamenány *proto*, že všichni do jednoho do 30 dnů zemřeli. Konec konců pokud byl výsledek založen na oslovení pacientů 30 dnů po přijetí, aby se zjistilo, jak na tom jsou, nikdo ze zemřelých už na otázky neodpověděl. Kdybychom si tuto možnost neuvědomili, nikdy bychom žádné zemřelé pacienty nezaznamenali.

Možná to zní trochu hloupě, ale k takovým situacím dochází v reálu poměrně často. Například model postavený tak, aby určil prognózu pacientů s konkrétní léčbou, bude založen na výsledcích pacientů, kteří tuto léčbu podstoupili už dříve. Co když ale neuplynula dostatečně dlouhá doba, aby všichni tito pacienti dosáhli nějakého výsledku? Pro takové pacienty by

případný výsledek nebyl znám. Model postavený pouze na těch, u nichž výsledek známe, by tedy mohl být zavádějící.

Podobný jev se vyskytuje v průzkumech, kde je zdrojem komplikací *chybějící odpověď*. Výzkumníci si obvykle sestavují úplné seznamy lidí, od nichž by rádi v ideálním případě získali odpověď, běžně se však stává, že ne všichni odpovědí. Pokud se ti, kteří odpověď poskytnou, nějak liší od těch, kteří neodpovědí, mají autoři průzkumu důvod k pochybnostem, zda jejich statistika pracuje s dobrým vzorkem populace. Konec konců kdyby nějaký časopis udělal mezi svými předplatiteli anketu s jedinou otázkou: „Účastníte se anket v časopisech?", nemohli bychom skutečnost, že sto procent respondentů uvedlo „ano", interpretovat tak, že na podobné ankety odpovídají úplně všichni předplatitelé.

Předchozí příklady jsou dokladem našeho prvního typu temných dat. Víme, že data o pacientech v registru TARN existují, třebaže tam nejsou zaznamenány všechny hodnoty. Víme, že lidé ze seznamu v anketě odpovědi měli, i když je třeba neuvedli. Obecně vzato tedy víme, že data mají své hodnoty; nevíme jen, jaké hodnoty to jsou.

Příklad odlišného druhu temných dat (*TD-typ 2: Data, o nichž nevíme, že chybějí*) vypadá následovně.

Spousta měst má problémy s výtluky v povrchu vozovek. Do drobných prasklin zatéká voda, která v zimě mrzne a praskliny ve vozovce se rozšiřují, navíc je poškozují pneumatiky aut. Vzniká začarovaný kruh, na jehož konci je výmol ve vozovce, který dokáže prorazit pneumatiku nebo poškodit nápravu. Město Boston se rozhodlo vyřešit tento problém pomocí moderní techniky. Dalo do oběhu aplikaci pro chytré telefony, která využívala vnitřní akcelerometr telefonu k detekci otřesu auta při přejezdu výmolu a poté prostřednictvím GPS automaticky předala informaci o poloze takové díry městským úřadům.

Úžasné! Teď budou silničáři přesně vědět, kam se mají vydat opravit výtluky.

A jsme zase u toho. Vypadá to jako elegantní a levné řešení reálného problému, založené na moderní technice pro analýzu dat – až na to, že vlastnictví automobilů a drahých chytrých telefonů se pravděpodobně bude soustřeďovat do oblastí s movitějšími obyvateli. Je tedy dost pravděpodobné, že výmoly v chudších regionech nebudou zaznamenány, tudíž nebude odeslána jejich poloha a v některých místech se opravy výtluků množná nikdy nedočkají. Místo toho, aby se problém s výtluky řešil systémově, může takový přístup prohloubit společenskou nerovnost. Situace se liší od příkladu s databází TARN, kde jsme věděli, že nám určitá data chybějí. Zde si toho vědomi nejsme.

Následuje další příklad tohoto druhu temných dat. Koncem října 2012 udeřil na východním pobřeží Spojených států amerických hurikán Sandy, zvaný též „superbouře Sandy“.[5] V té době to byl druhý nejničivější hurikán v amerických dějinách a nejmohutnější v Atlantiku zaznamenaný hurikán, který napáchal škody odhadem za 75 miliard dolarů a o život připravil víc než 200 lidí v osmi zemích. Sandy postihla 24 amerických států od Floridy k Maine až po Michigan a Wisconsin a kvůli výpadkům elektrické energie musely být uzavřeny finanční trhy. Nepřímo to vedlo i k nárůstu porodnosti o nějakých devět měsíců později.

A představovala také triumf novodobých médií. Fyzickou bouři hurikánu Sandy provázela bouře zpráv na Twitteru, popisujících, co se děje. Na Twitteru jde o to, že vám říká, co a kde se děje, když událost probíhá, stejně tak jako komu se to stalo. Tato platforma sociálních médií je způsob, jak držet krok s vývojem událostí v reálném čase. Právě to se stalo s hurikánem Sandy. V období od 27. října do 1. listopadu 2012 se ho týkalo přes 20 milionů tweetů. Klidně bychom si tedy mohli myslet, že jde o ideální materiál, z nějž lze získat neustále se měnící obrázek o bouři tak, jak se vyvíjí, se zjištěním nejvíc postižených oblastí, kam je potřeba vyslat okamžitou pomoc.

Pozdější analýza však ukázala, že největší množství tweetů o Sandy pocházelo z Manhattanu, naopak z oblastí jako Rockaway či Coney Island jich bylo jen poskrovnu. Znamenalo to snad, že Rockaway a Coney Island byly bouří postiženy méně? Je sice pravda, že metro a ulice Manhattanu byly zatopené, sotva ho lze ale prohlásit za nejhůř postiženou oblast, a to ani v samém New Yorku. Pravda je samozřejmě taková, že oblasti vykazovaly méně tweetů ne proto, že u nich neměla bouře takové následky, ale prostě proto, že tam bylo méně uživatelů Twitteru s menším počtem chytrých telefonů, ze kterých by mohli tweetovat.

Vlastně si opět dokážeme představit extrémní případ této situace. Kdyby Sandy nějaký region úplně smetla z povrchu, pochopitelně by odtamtud nepřišly žádné tweety. Na první pohled by se zdálo, že všichni tam jsou v pořádku. Vskutku temná data.

Podobně jako u prvního typu temných dat se s příklady tohoto druhého typu, u nějž nevíme, že nám něco chybí, setkáváme všude kolem. Vezměte si jen neodhalené podvody či neschopnost průzkumů mezi oběťmi trestných činů určit, že se nějaké vraždy vůbec staly.

Možná ve vás první dva druhy temných dat vyvolávají déjà vu. Bývalý americký ministr obrany Donald Rumsfeld je na slavném brífinku s novináři pěkně charakterizoval chytlavým sloganem, podle kterého „máme známé neznámé; tedy víme, že určité věci nevíme. Máme však také neznámé neznámé – to, co nevíme, že nevíme“.[6] Rumsfeld za to zamotané prohlášení schytal od novinářů značný výsměch, kritika však nebyla na místě. To, co řekl, naprosto dávalo smysl a bylo zcela jistě pravdivé.

Tyto dva typy jsou však pouhým začátkem. V další části zmiňuji některé z dalších druhů temných dat. Právě o nich a ještě o dalších popsaných dále tato kniha pojednává. Jak uvidíte, berou na sebe temná data mnoho podob. Dokud si neuvědomíme, že data mohou být neúplná, že něco zaznamenávat neznamená zaznamenávat úplně všechno, že postup měření může být nepřesný a že to, co měříme, nemusí být přesně tím, co změřit chceme, budeme dost možná mít jen velmi zavádějící dojem, co se děje. To, že nikdo zrovna neslyší padat strom v lese, ještě neznamená, že ten pád nevydal žádný zvuk.

Tak vy si myslíte, že máte všechna data?

Zákazník v supermarketu přijíždí k pokladně s plným vozíkem. Laserový skener načítá čárové kódy jednotlivých položek, pokladna pokaždé zapípá a sčítá celkovou útratu. Na konci toho všeho dostává zákazník účet a platí. Až na to, že tím to ještě nekončí. Data popisující zakoupené položky i jejich cenu putují do databáze, kde se archivují. Později nad nimi budou hloubat statistikové a datoví vědci, kteří z těchto údajů vytáhnou obrázek zákaznického chování od podrobností, jaké položky byly zakoupeny, které byly zakoupeny jedna kvůli druhé a jaký zákazník je zakoupil. Zde nám určitě žádná data neunikají, že? Data transakce se musí zachytit, chce-li supermarket vědět, kolik má zákazníkovi naúčtovat – pokud tedy nedojde k výpadku proudu nebo selhání paměťové buňky či k podvodu.

Zdá se dost zjevné, že zachycené údaje představují všechna data, která zde máme. Nejde přece jen o *některé* transakce či detaily pouze *některých* zakoupených položek. Jde o *všechny* transakce *všech* zákazníků spojené se *všemi* položkami v daném supermarketu. Je to přístup, kterému se občas pro jednoduchost říká „data = vše“.

Je tomu ale skutečně tak? Tato data přece říkají, co se stalo *minulý* týden nebo *minulý* měsíc. Je to užitečné, ale pokud supermarket provozujeme, patrně chceme zjistit spíš to, co se bude dít zítra, příští týden, příští měsíc. Ve skutečnosti chceme vědět, kdo, co a kdy si koupí a kolik si toho zákazníci koupí v budoucnu. Které zboží nám dojde, pokud toho nevyskládáme do regálů víc? Kterým značkám dají lidé při koupi přednost? Vlastně stojíme o data, která nikdo neměří. Zatemňující vliv času na data popisují temná data *TD-typu 7: Změny v čase*.

Kromě této komplikace bychom také třeba rádi věděli, jak by se lidé *byli zachovali*, kdybychom naskladnili jiné položky, pokud bychom je rozmístili do regálů jinak nebo kdybychom změnili otevírací dobu supermarketu. Těmto faktorům se říká *kontrafaktuály*, neboť neodpovídají skutečnosti – jde v nich o to, co by se stalo, kdyby se nestalo to, co se stalo. Kontrafaktuály tvoří *TD-typ 6: Data, která by mohla být*.

Netřeba říkat, že kontrafaktuály jsou důležité nejen pro manažery supermarketů. Jistě jste už někdy v minulosti užívali léky. Důvěřovali jste lékaři, který vám je předepsal, a předpokládali jste, že byly řádně testovány a byl u nich prokázán účinek ve smyslu zmírnění zdravotních obtíží. Jak byste se však cítili, kdybyste zjistili, že testovány nebyly? Že vůbec nebyla zjišťována data, zda tyto léky zdravotní stav zlepšují? A že je vlastně dost dobře možné, že naopak stav zhoršují? Nebo i když byly testovány a zjistilo se, že pomáhají, nebyly ty léky porovnávány s případem, kdy se nemoc ponechá neléčená, aby se ukázalo, jestli se léčený člověk zotaví rychleji než přirozeným ozdravným procesem? Nebo že léky nebyly porovnávány s jinými, aby se zjistilo, zda jsou účinnější než zavedené alternativy? V příkladu se sloním práškem by srovnání s nečinností rychle ukázalo, že *nedělat nic je stejně účinné jako držet slony z dosahu* tím, že rozsypáváme po zemi hrsti prášku. (A to by dál mohlo vést ke zjištění, že zde žádní sloni, které je třeba držet z dosahu, vlastně ani nejsou.)

Vrátíme-li se k pojmu „data = vše", pak představa, že bychom mohli mít „všechna" data, je v ostatních souvislostech *naprosto jasně* nesmyslná. Vezměte si svou tělesnou hmotnost. Tu lze změřit docela snadno – stačí skočit na váhu v koupelně. Když však měření zopakujete, klidně i vzápětí, můžete zaznamenat trochu jiný výsledek, zejména pokud měříte na dekagramy či gramy. Všechna fyzikální měření s sebou nesou potenciální nepřesnosti, které jsou výsledkem chyby měření či náhodných fluktuací, vznikajících z nepatrných změn okolností (*DT-typ 10: Chyby a nepřesnosti měření*). Aby se tomuto problému vyhnuli, provádějí vědci zkoumající rozsah nějaké veličiny – řekněme rychlosti světla nebo elektrického náboje elektronu – víc měření, která pak zprůměrují. Můžou provádět deset měření nebo také sto. Zjevně však nemohou provést „všechna" měření. V tomto kontextu nic takového jako „vše" neexistuje.

Další druh temných dat dokládá jízda londýnskými červenými autobusy: každý ví, že jsou velice často přeplněné cestujícími. A přesto data ukazují, že obsazenost průměrného autobusu činí pouhých 17 lidí. Jak vysvětlit tento zdánlivý rozpor? Manipuluje snad někdo s údaji?

Trocha zamyšlení ukáže, že odpověď prostě zní, že víc lidí jede v autobusu tehdy, když je plný – právě to je význam slova „plný". Důsledkem je, že víc lidí zaznamená plný autobus. V opačném extrému nebude mít prázdný autobus nikoho, kdo by oznámil, že je prázdný. (Řidiče samozřejmě ponechávám stranou.) Tento příklad popisuje *DT-typ 3: Když vybíráme jen některé případy*. Navíc může být tento druh temných dat dokonce nutným důsledkem sběru dat, v tom případě ilustruje *DT-typ 4: Vlastní výběr*. Následují dva moje oblíbené příklady, opačné extrémy co do významnosti.

Prvním je kresba, na níž si muž prohlíží jednu z takových těch velkých map před vlakovými nádražími. Uprostřed mapy je červený puntík s nápisem „Tady stojíte". „Jak," pomyslí si muž, „to jen mohou vědět?" Věděli to proto, že si uvědomili, že *každý*, kdo se na červený puntík dívá, musí nutně stát před tou značkou. Šlo o důsledně vybraný vzorek, který *nutně* vynechal kohokoli stojícího jinde.

Jde o to, že temná data lze sbírat pouze tehdy, je-li někdo nebo něco – například měřicí nástroj –, co je shromažďuje. Druhou extrémní ukázku popisuje *antropický princip*, který v podstatě říká, že vesmír je nutně takový, jaký je, jinak bychom tu totiž nebyli, abychom si toho všimli. Data z odlišných vesmírů neshromažďujeme proto, že bychom tam nemohli existovat a data sbírat. To znamená, že veškeré učiněné závěry jsou nutně omezeny naším vesmírem (typem): stejně jako v případě výtluků na silnicích se může dít spousta dalších věcí, o nichž nevíme.

To pro vědu představuje důležité ponaučení. Vaše teorie může dokonale odpovídat vašim datům, ale ta budou mít svá omezení. Nemusí se vztahovat k velmi vysokým teplotám, k dlouhým časovým intervalům či k obrovským vzdálenostem. A pokud provedete extrapolaci za hranice shromážděných dat, může se stát, že se vám teorie zhroutí. Ekonomické teorie založené na datech sesbíraných za příznivých podmínek mohou dramaticky selhat během recesí a Newtonovy zákony fungují dobře, pokud je nevztáhneme na drobné předměty či vysoké rychlosti a další extrémní okolnosti. V tom tkví podstata *TD-typu 15: Extrapolace za hranicemi dat*.

Mám tričko s kresbou z webového komiksu *xkcd*, na níž spolu hovoří dvě postavičky. Jedna říká: „Vždycky jsem si myslel, že korelace značí kauzalitu." V dalším obrázku dodává: „Pak jsem chodil do kurzu statistiky. Teď už si to nemyslím." Druhá postava to uzavře: „Zřejmě tedy kurz pomohl." Na to první: „No, možná."[7]

Korelace prostě znamená, že dvě věci se proměňují společně: například pozitivní korelace znamená, že když je něco velké, to druhé je také velké, a když je jedno malé, je malé i to druhé. To je něco jiného než kauzalita. Jedno totiž *zapříčiňuje* to druhé tehdy, pokud změna v prvním způsobí změnu ve druhém. A potíž tkví v tom, že dvě věci se mohou společně měnit, aniž by změny v jedné způsobovaly změny ve druhé. Například

pohled na prvňáčky ve škole ukazuje, že děti s větší slovní zásobou bývají v průměru vyššího vzrůstu. Asi byste však nevěřili tomu, že rodiče toužící po vyšších potomcích by si měli najímat domácí učitele na rozšiřování slovní zásoby. Pravděpodobnější je zde přítomnost nezjištěných temných dat, jakéhosi třetího faktoru, který tu souvislost vysvětluje – třeba věk dětí. Když postavička na *xkcd* tričku říká „no, možná", vyjadřuje tím názor, že změnu jejího poznání možná způsobilo absolvování kurzu statistiky, ale také to mohlo být něco jiného. Ukážeme si několik zjevných příkladů situace, kterou vystihuje *TD-typ 5: Když chybí to podstatné.*

Už jsem zmínil několik druhů temných dat. Ale je jich víc. Cílem této knihy je odhalit je, ukázat, jak je poznat, jak zaznamenat jejich důsledky a jak se vypořádat s problémy, které způsobují – a dokonce i to, jak je využít. Jejich seznam je uveden na konci této kapitoly a souhrn najdete v kapitole 10.

Nic se nestalo, tak jsme to ignorovali

Poslední uvedený příklad dokládá, že temná data mohou mít katastrofální následky a nejsou problémem výhradně velkých datových souborů.

Před více než třiceti lety, 28. ledna 1986, došlo v jednom z pomocných vzletových raketových motorů raketoplánu *Challenger* k silnému výbuchu, který ho 73 sekund po vzletu ve výšce devět mil roztrhal na kusy. Kabina s posádkou pokračovala po své trajektorii a než se zřítila do Atlantiku, dosáhla výšky přes 19 kilometrů. Zahynulo všech sedm členů posádky, sestávající z pěti astronautů a dvou letových specialistů.

Pozdější prezidentská vyšetřovací komise zjistila, že střední řídící pracovníci NASA porušili bezpečnostní pravidla ukládající, že data mají být předávána ve sledu příkazů. Přičítalo se to ekonomickým tlakům, bylo velmi důležité dodržet datum vzletu: to už se totiž posunulo z 22. na 23. ledna, pak na 25. ledna a poté na 26. ledna. Jelikož teplotní předpověď uváděla na ten den nepřijatelně nízkou teplotu, byl start raketoplánu opět odložen – na 27. ledna. Už běžel odpočet, když kontrolky ohlásily, že zámek poklopu nebyl řádně uzavřen. Než závadu odstranili, foukal už příliš silný vítr, a tak byl start opět odsunut.

V noci 27. ledna se odehrála tříhodinová telekonference mezi Morton Thiokol, společností, která pomocné raketové motory vyrobila, vedením NASA v Marshallově centru kosmických letů a lidmi z Kennedyho vesmírného střediska. Larry Wear z Marshallova centra požádal firmu Morton Thiokol, aby ověřila možný vliv nízkých teplot na raketové motory s tuhým palivem. Tým Morton Thiokol upozornil, že za nízkých teplot může dojít k tuhnutí O-kroužků.

O-kroužky, gumovité těsnění o průměru zhruba sedm milimetrů, bylo umístěno po obvodu stěny spoje mezi každým ze čtyř segmentů raketového motoru. Pomocné vzletové motory byly přes 45 metrů vysoké a po obvodu měřily 11,5 metru. Při podmínkách vzletu se mezera cca 0,01 cm, kterou O-kroužky obvykle utěsnily, rozevřela na 0,15 cm – tedy na patnáct setin centimetru. A tato větší mezera zůstala při startu otevřená po dobu zhruba šest desetin sekundy.

Roberta Ebelinga z Morton Thiokol znepokojovalo, že za nízkých teplot působí tuhnutí O-kroužků ztrátu jejich schopnosti vytvářet mezi segmenty účinné těsnění, když se mezery o zmíněný jeden a půl milimetru na 0,6 sekundy roztáhnou. Robert Lund, viceprezident společnosti Morton Thiokol, během telekonference uvedl, že teplota u O-kroužků nesmí při startu klesnout pod předchozí nejnižší vzletovou teplotu 11,5 °C. Následovala rozsáhlá, místy vzrušená debata, a to jak v rámci telekonference, tak i offline při hovorech o přestávkách. Nakonec společnost Morton Thiokol své stanovisko přehodnotila a start doporučila.

Přesně 58,79 sekundy po vzletu vyšlehl plamen poblíž posledního spoje z pravého raketového motoru na tuhé palivo. Rychle se rozhořel v ohnivý jazyk, který poškodil závěsy spojující raketový motor s vnější palivovou nádrží. Motor se natočil, čímž zasáhl nejprve křídlo a potom vnější nádrž raketoplánu. Ohnivý jazyk následně zasáhl vnější nádrž s vodíkovým a kyslíkovým palivem. V čase 64,66 sekundy došlo k porušení povrchu nádrže a o devět sekund později pohltila *Challenger* ohnivá koule, která ho rozlámala na několik kusů.[8]

Jedno si musíme zapamatovat, totiž že lety do vesmíru jsou především rizikem. Žádná mise, a to ani za těch nejlepších možných okolností, není zcela bezpečným podnikem: na nulu riziko prostě srazit nelze. A k tomu ještě se vždycky dostanou do hry protichůdné požadavky.

V tomto případě, podobně jako v jakémkoli podobném incidentu, je navíc představa „příčiny" dost komplikovaná. Bylo jí snad porušení bezpečnostních pravidel, přílišný tlak vyvíjený na vedoucí pracovníky kvůli ekonomickým zájmům, další důsledky škrtů v rozpočtu, nebo snad mediální tlak poté, co byl vzlet předchozího raketoplánu *Columbia* sedmkrát odložen a každý odklad byl provázen novinářským výsměchem? Zde je jako příklad přepis slov moderátora Dana Rathera ve večerních zprávách v pondělí 27. ledna, tedy po čtyřech odkladech startu *Challengeru*: „Další drahý odklad tolik očekávaného startu raketoplánu. Tentokrát se vina svaluje na špatný poklop a na krapet špatného počasí na čistém nebi." Nebo

to byl důsledek politického tlaku? U tohoto startu byly ve hře výrazně větší zájmy než u těch předchozích, protože raketoplán měl na palubě „obyčejného člověka", učitelku Christu McAuliffeovou, a na večer 28. ledna byl plánován prezidentův projev o stavu Unie.

V takové situaci obvykle hraje roli víc faktorů. A složité a nejasné vazby mají nečekané důsledky. Ovšem v tomto případě se navíc objevil další faktor: temná data.

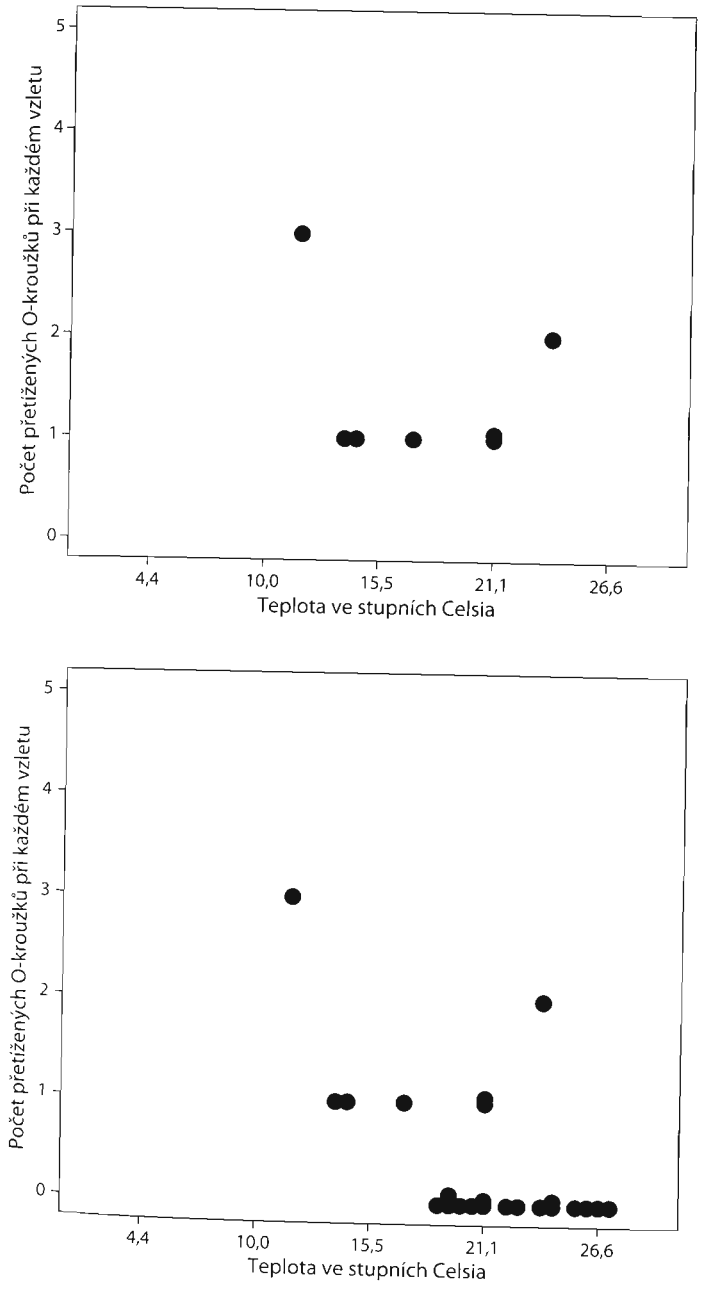
Komise vedená bývalým ministrem zahraničí Williamem Rogersem po neštěstí upozornila na skutečnost, že do programu probíraného při telekonferenci nebyly zahrnuty lety, které nevykazovaly nepřiměřenou zátěž O-kroužků (temná data *TD-typ 3: Když vybíráme jen některé případy*, ale také *TD-typ 2: Data, o nichž nevíme, že chybějí*). Ve zprávě komise na straně 146 stálo: „Vedoucí pracovníci posuzovali teplotní vliv u těch letů, při nichž byly u O-kroužků zjištěny termické problémy – avšak nikoli četnost jejich výskytu u všech letů."[9] A v tom to vězí: *analýza nevzala v potaz data z některých letů*. Moje předchozí příklady ukazují, jaké druhy problémů mohou nastat, vynecháme-li některá data.

Zpráva pokračovala: „Při takovém srovnání [tedy užívajícím jen omezený soubor dat] není na výskytu ‚přetížení' O-kroužků během startu v teplotním rozmezí 11,7 až 23,9 stupně Celsia nic zvláštního," což znamená: mezi teplotou a počtem O-kroužků vykazujících nepřiměřenou zátěž není žádný zjevný vztah. Avšak „jakmile vezmeme v potaz zkušenosti ze všech letů včetně těch ‚normálních', tedy bez eroze či překonání žáruvzdorného tmelu, naskytne se nám zcela odlišné srovnání"; tj. zahrneme-li všechna data, naskytne se nám jiný obrázek. U letů při vyšších teplotách bylo vlastně mnohem pravděpodobnější, že žádné problémy nenastanou, a právě v tom spočívala temná data, která se nedostala do hry. Platí-li však, že čím vyšší je teplota, tím je menší pravděpodobnost problému, pak také naopak platí, že čím nižší je teplota, tím vyšší je pravděpodobnost, že k problému dojde. A teplota vzduchu byla tehdy předpovězena na pouhých –0,6 °C.

Tato část zprávy uzavírala: „Vezmeme-li v potaz všechny dosud uskutečněné starty raketoplánů, dojdeme k závěru, že pravděpodobnost přílišné zátěže O-kroužků se začne *blížit jistotě* při teplotě spojů nižší než 18,3 °C." (moje kurziva).

Situaci znázorňují dva grafy na obrázku 1. Horní obrázek 1a) ukazuje graf diskutovaný při telekonferenci. Jde o výsek počtu přetížených O-kroužků u každého startu za určité vzletové teploty ve stupních Celsia. Například při nejnižší předchozí vzletové teplotě, tedy při 11,7 °C, došlo k enormnímu zatížení tří O-kroužků a při nejvyšší vzletové teplotě, tedy

Obr. 1 a) Data posuzovaná během telekonference před vzletem *Challengeru*; b) úplná data



23,9 °C, byla nadměrná zátěž pozorována u dvou O-kroužků. Mezi vzletovou teplotou a počtem přetížených O-kroužků není zjevný vztah.

Přidáme-li však chybějící data, jež ukazují vzlety, u nichž nenastalo přetížení O-kroužků, dostaneme obrázek 1b). Závislost už je naprosto zřejmá. Ve skutečnosti platí, že při *všech* vzletech za teploty pod 18,3 °C došlo k nějaké nepřiměřené zátěži O-kroužků, oproti tomu při vyšších teplotách došlo k témuž jen ve čtyřech případech z 21. Obrázek ukazuje, že čím nižší je teplota, tím vyšší je nebezpečí. A co je horší, předpověď vzletové teplota byla výrazně nižší než kdykoli v minulosti (*TD-typ 15: Extrapolace za hranicemi dat*).

Chybějící data hrají klíčovou roli pro pochopení toho, co se děje.

Uvedený případ má zajímavý důsledek. Třebaže trvalo dlouhé měsíce, než oficiální zpráva dospěla ke svým závěrům, už v den katastrofy klesla hodnota akcií společnosti Morton Thiokol o 11,86 procenta. Na základě pohybu hodnoty akcií Morton Thiokol před incidentem lze říci, že změny byť jen o čtyři procenta byly vzácností. Ceny akcií dalších společností podílejících se na konstrukci nosné rakety také klesly, ovšem mnohem méně. Jako by trh věděl, co havárii způsobilo. Takže opět temná data?

Síla temných dat

Tento poslední příklad ukazuje, jak strašlivá spoušť může nastat, nedbáme-li temných dat. Zdá se, že představují skutečnou hrozbu. Reálně však celkový obraz až tak bezútěšný není. Ukazuje se, že pochopení temných dat můžeme v džungli dat využít ve svůj prospěch. Dosáhnout toho můžeme několika způsoby, jež popisuji ve druhé části knihy. Tady je jeden z nich.

V kapitole 2 se věnuji takzvaným randomizovaným kontrolovaným studiím a v kapitole 9 se k nim vracím z jiného úhlu pohledu. V kontextu lékařské vědy představuje nejjednodušší takovou studii srovnávání dvou léčivých přípravků tak, že jeden podáváme jedné skupině lidí a jiný té druhé. To však s sebou nese určité riziko. Pokud výzkumníci vědí, který z přípravků byl podáván které skupině, může tento poznatek výzkum ovlivnit – výzkumníci totiž mohou být v pokušení přistupovat k jedné skupině pečlivěji než k té druhé. Je-li studie zacílena například na srovnání neověřeného nového preparátu se standardně používaným lékem, mohou badatelé sledovat možné nežádoucí účinky u první skupiny důkladněji (třeba i podvědomě), případně jí věnovat větší pozornost při měření možných výsledků. Abychom se této možné tendenci vyhnuli, je nasazení léčivých přípravků v takových studiích

před výzkumníky skryto (*TD-typ 13: Záměrně zatemněná data*). Používá se k tomu termín *zaslepená studie*, který svědčí o tom, že jde o temná data.

Dalším dobře známým způsobem prospěšného využití temných dat je výběrové šetření. Můžeme například chtít znát názory lidí ve městě nebo zákazníků kupujících produkty určitého výrobce, ale náklady spojené s oslovením všech takových lidí jsou zřejmě neúměrně vysoké. Zcela jistě by to navíc stálo spoustu času, během nějž by se názory mohly měnit. Namísto toho, abychom se ptali všech, se tedy můžeme zeptat jen některých z nich. Názory ostatních lidí, které neoslovíme, pak budou temnými daty. Taková strategie se může zdát velmi riskantní – přesně se to podobá příkladu s TARN. Ovšem ukazuje se, že prozíravým výběrem respondentů lze získat přesné a spolehlivé odpovědi – rychleji a levněji, než kdybychom se pokusili dotazovat všech.

Třetí možností, jak využít temná data ku prospěchu, je takzvané vyhlazování dat. Jak uvidíme v kapitole 9, jde o ekvivalent k odhalování nezjištěných a nezjistitelných druhů temných dat (*TD-typ 14: Smyšlená a umělá data*) a umožňuje přesnější odhady a lepší předpovědi.

Dalšímu užití temných dat se věnuje kapitola 9, kde uvidíme, že mají často exotické názvy. Některá z nich se hojně užívají v oblastech, jako jsou strojové učení a umělá inteligence.

Jsou všude kolem nás

Viděli jsme, že temná data jsou všudypřítomná. Mohou se vynořit úplně kdekoli a patrně největším nebezpečím s nimi spojeným je to, že už z podstaty věci si nemůžeme být jisti, že tu s námi *nejsou*. Znamená to, že se musíme mít neustále na pozoru a klást si otázku: *Co nám uniká?*

Nevšimneme si spousty podvodů, protože policie chytí jen packaly, zatímco skutečně dobří zločinci unikají bez povšimnutí? Bernie Madoff založil svou investiční společnost Bernard L. Madoff Investment Securities LLC v roce 1960, zatčen byl však až v roce 2008 a odsouzen (na 150 let vězení) o rok později, když mu bylo už 71 let – málem mu to prošlo.

Nevšímáme si mnoha potenciálně léčitelných pacientů jenom proto, že závažnější případy jsou víc vidět, zatímco ty méně vážné nevykazují tolik příznaků?

Jsou sociální sítě spravované moderními sociálními médii nebezpečné už jen proto, že se v nich odráží to, co už víme a v co věříme – a nikam nás neposouvají, jelikož nám neukazují fakta ani události mimo naši komfortní zónu?

A co je ještě horší: popisy, které si lidé vybírají ke zveřejnění na sociálních sítích, v nás mohou vyvolat falešný dojem, jak báječný život vedou všichni ostatní, a kontrast s naším vlastním životem plným překážek nás může stáhnout do deprese.

Máme tendenci považovat data za číselná. Data však nutně nejsou jen čísla. A to znamená, že ani temná data nemusejí být číselná. Uvedu příklad, v němž klíčovou chybějící informaci představuje jediné písmenko.

Účastníci výprav do Arktidy v letech 1852, 1857 a 1875 byli vybaveni zásobou Allsopp's Arctic Ale, piva s mimořádně nízkým bodem mrznutí, které jim zvlášť pro tu příležitost uvařil pivovarník Samuel Allsopp. Toto pivo ochutnal v roce 1889 Alfred Barnard, který je popsal jako „pěkně hnědé, s vinnou až ořechovou chutí a chutné právě jako v den, kdy ho uvařili... Vzhledem k vysokému obsahu nekvašeného výtažku, který se v něm uchoval, je záhodno pokládat je za velmi hodnotnou a výživnou potravinu".[10] Právě takové pivo člověku na výpravách do Arktidy jistě dobře poslouží.

Jedna lahev z várky v roce 1852 se objevila v roce 2007 k prodeji na eBay, s vyvolávací cenou 299 dolarů. Aspoň takový byl cíl. Stalo se totiž, že prodejce, který onu lahev vlastnil celých 50 let, napsal název piva chybně – Allsop jen s jedním p. Následkem toho se položka neukázala v hledáčku většiny nadšenců pro stará piva, takže přišly jen dva příhozy. Na vítězné 304 dolary přihodil 25letý Daniel P. Woodul. Ve snaze zjistit skutečnou cenu lahve ji Woodul obratem dal znovu k prodeji na eBay, tentokrát však již se správným názvem. Přišlo 157 nabídek, ta vítězná ve výši 503 300 dolarů!

Takže na onom chybějícím *p* očividně záleželo, když jeho cena obnášela nějaký ten půlmilion dolarů.* Tento příklad ukazuje, že chybějící informace mohou mít dalekosáhlé následky. Jak vlastně ještě uvidíme, půl milionu dolarů je naprostý pakatel v porovnání s tím, k jakým ztrátám vedla chybějící data v jiných situacích. Chybějící data mohou skutečně zničit životy, přivést společnosti ke krachu a mohou (jako v případě katastrofy raketoplánu *Challenger*) dokonce vést k úmrtím. Na chybějících datech prostě záleží.

V případě Allsoppova arktického piva stačilo jen být trochu pozornější a problém by nenastal. Třebaže nedbalost je jistě běžnou příčinou temných dat, existuje ještě celá řada dalších. Smutnou skutečností je, že temná data

* Vlastně se ukázalo, že vítězná nabídka byla myšlena jako vtip a výherce ji neměl v úmyslu zaplatit. I tak si však Woodul bezpochyby přišel na pěknou sumičku: soukromý sběratel ze Skotska nedávno za jednu lahev z expedice v roce 1875 zaplatil 3300 liber (zhruba 4300 dolarů).

mohou být temná z široké palety rozmanitých důvodů, jak ještě v knize uvidíme.

Je lákavé považovat temná data za synonymum údajů, které mohly být zaznamenány, ale z nějakého důvodu nebyly. Zcela jistě jde o nejpatrnější druh temných dat. Mezi temná data bezesporu patří chybějící platové stupně v průzkumu, v němž někteří lidé odmítli prozradit, kolik berou, ale také k nim patří platové stupně těch, kteří nepracují, a tudíž ani nemají co prozrazovat. Skutečné hodnoty jsou zatemňovány chybami v měření, souhrny dat (například průměry) zase skrývají podrobnosti a nesprávné definice navíc zkreslují výsledky toho, co chcete zkoumat. Ještě obecněji lze za temná data pokládat jakékoli neznámé charakteristiky populace (statistikové je často označují jako *parametry*).

Jelikož množství možných příčin temných dat je v podstatě neomezené, může se nám nesmírně hodit (abychom se vyhnuli chybám a přešlapům), víme-li aspoň, na jaký *druh* si máme dávat pozor. A k tomu slouží v knize popisované *TD-typy*. Nejde o základní příčiny (jako neschopnost započítat konečné výsledky u těch pacientů, kteří se výzkumu zúčastnili jen krátce), poskytují však obecnější taxonomii (jako rozlišování mezi daty, u nichž víme, že nám chybějí, a daty, u kterých nevíme, že chybějí). Povědomí o těchto *TD-typech* může pomoci při ochraně před chybami, omyly a katastrofami pramenícími z neznalosti toho, co nevíte. *TD-typy* uvedené v této knize a shrnuté v kapitole 10 jsou následující:

TD-typ 1: Data, o nichž víme, že chybějí
TD-typ 2: Data, o nichž nevíme, že chybějí
TD-typ 3: Když vybíráme jen některé případy
TD-typ 4: Vlastní výběr
TD-typ 5: Když chybí to podstatné
TD-typ 6: Data, která by mohla být
TD-typ 7: Změny v čase
TD-typ 8: Definice dat
TD-typ 9: Souhrny dat
TD-typ 10: Chyby a nepřesnosti měření
TD-typ 11: Zpětná vazba a gaming
TD-typ 12: Informační asymetrie
TD-typ 13: Záměrně zatemněná data
TD-typ 14: Smyšlená a umělá data
TD-typ 15: Extrapolace za hranicemi dat