Kapitola 6

Hodnota

Koncem 90. let se z webu začínalo rychle stávat chaotické, nevlídné a nepřátelské místo. „Spamboti" zaplavovali e-mailové schránky a tapetovali webová fóra. Dvaadvacetiletý čerstvý absolvent Luis von Ahn však roku 2000 dostal nápad, jak tento problém vyřešit: stačí přinutit zájemce o registraci, aby dokázali, že jsou skutečně lidmi. Hledal tedy něco, co by bylo snadné pro lidi, ale těžké pro počítače.

Napadlo ho, že by se uživatelům během registrace daly zobrazovat deformované a těžko čitelné znaky. Lidé by je dokázali dešifrovat a zadat správný text během pár vteřin, ale počítače by si s tímto úkolem neporadily. Jeho metodu implementoval portál Yahoo a okamžitě tím postrach spambotů zlikvidoval. Von Ahn svůj výtvor nazval Captcha (ze slov Completely Automated Public Turing Test to Tell Computers and Humans Apart – plně automatizovaný veřejný Turingův test, který dokáže rozlišit počítače a lidi). O pět let později už uživatelé Internetu zadávali miliony testů Captcha denně.

Vývojem testů Captcha se von Ahn značně proslavil a po skončení svého doktorského studia díky nim získal učitelské místo na univerzitě Carnegie Mellon. Přispěly také k tomu, že ve svých 27 letech získal od MacArthurovy nadace jednu z prestižních cen pro „génie" ve výši půl milionu dolarů. Když si však uvědomil, že kvůli němu miliony lidí každého dne plýtvají svým časem na zadávání otravných zvlněných písmen – celá ta spousta informací se přitom vzápětí zahazovala – už si tak geniální nepřipadal.

Ve snaze najít způsob, jak tuto lidskou intelektuální kapacitu využít produktivněji, přišel s následníkem svého systému, který přiléhavě pojmenoval ReCaptcha. Místo zadávání náhodně generovaných písmen lidé opisovali dvě slova z projektu skenování textu, která počítačový program na optické rozpoznávání znaků nedokázal rozluštit. Účelem prvního slova je potvrdit, co zadali jiní uživatelé. Slouží tedy jako signál, že úkol skutečně řeší osoba a nikoli stroj. Poté následuje druhé slovo, se kterým program OCR potřebuje

pomoci. Aby byla zajištěna přesnost, systém stejné nejasné slovo předkládá v průměru pěti různým lidem. Teprve poté, když opakovaně dostane stejnou odpověď, předpokládá, že slovo je zapsáno správně. Data mají své primární použití – dokázat, že uživatelem je člověk – ale zároveň mají i sekundární účel: dešifrovat nejasná slova v digitalizovaných textech.

Když si uvědomíme, kolik by stálo na stejnou práci najmout placené pracovníky, je zřejmé, že systém poskytuje mimořádnou hodnotu. Při současné frekvenci 200 milionů testů ReCaptcha denně, z nichž každý trvá průměrně 10 sekund, se dohromady jedná o půl milionu hodin za každý den. Minimální mzda v USA roku 2012 činila 7,25 dolarů na hodinu. Pokud bychom si kontrolu nezřetelných slov, kterým počítačový program nerozumí, objednali na trhu, stálo by nás to asi 4 miliony dolarů denně neboli více než miliardu dolarů ročně. Systém, který von Ahn vymyslel, to však dokáže provést prakticky zdarma. Tento princip byl natolik cenný, že von Ahnovu technologii roku 2009 zakoupila společnost Google. Google ji nabízí k bezplatnému použití na libovolném webu. V současnosti tyto testy najdeme asi na 200 000 webech, k nimž patří i Facebook, Twitter a Craigslist.

Historie testů ReCaptcha ukazuje, jaký význam má opakované použití dat. Veledata způsobují, že se hodnota dat mění. V digitálním věku data přestávala sloužit jen k podpoře transakcí a často se měnila na samotný předmět obchodu. Ve veledatovém světě je znovu všechno jinak. Těžiště hodnoty dat se přenáší od jejich primárního použití k potenciálním budoucím aplikacím. Tento posun má zásadní důsledky. Podniky kvůli němu odlišným způsobem oceňují a zpřístupňují svá data. Umožňuje podnikům měnit své obchodní modely (nebo je k tomu může donutit). Ovlivňuje, jak firmy uvažují o datech a jejich použití.

Informace byly v tržních transakcích vždy klíčové. Díky datům lze například zjišťovat ceny, které představují signál, kolik produkovat. Tento rozměr dat je obecně známý. Informace určitého typu se již dlouho obchodují na trhu. Příkladem je obsah knih, článků, hudebních a filmových výtvorů nebo finanční informace typu cen akcií. V uplynulých desetiletích k nim přibyly i osobní údaje. Specializovaní obchodníci s daty v USA (např. Acxiom,

Experian a Equifax) si nechávají bohatě platit za kompletní soubory osobních informací o stovkách milionů zákazníků. Platformy sociálních médií jako Facebook, Twitter a LinkedIn způsobily, že se bohaté zdroje osobních údajů, které jsou o nás již k dispozici, dále rozšiřují o naše osobní vztahy, názory, preference a podrobnosti o životním stylu.

Stručně řečeno: data sice měla svou cenu odedávna. Tato jejich charakteristika se však buď považovala za vedlejší aspekt role dat při řízení podniku, nebo byla omezena na relativně úzké kategorie, jako je duševní vlastnictví či osobní údaje. Oproti tomu v éře veledat budou za něco cenného považována *všechna* data sama o sobě.

Když říkáme „všechna data", myslíme tím i ty nejzákladnější, zdánlivě zcela obyčejné informační bity. Jako příklad můžeme uvést odečty z teplotního senzoru na továrním stroji. Nebo datový proud aktuálních informací se souřadnicemi GPS, měření akcelerometru a hladina paliva v dodávkovém vozidle – případně z flotily 60 000 takových vozidel. Případně si představme miliardy starých vyhledávacích dotazů nebo ceny téměř všech letenek u téměř každého leteckého dopravce v USA za několik posledních let.

Až dosud neexistoval jednoduchý způsob, jak taková data shromažďovat, uchovávat a analyzovat, což zásadně omezovalo možnosti, jak z těchto dat extrahovat jejich potenciální hodnotu. V případě slavného příkladu s výrobcem špendlíků, na kterém Adam Smith v 18. století vysvětloval výhody dělby práce, by museli teoretičtí pozorovatelé sledovat všechny dělníky nejen pro účely této konkrétní studie, ale nepřetržitě každý den, provádět podrobná měření a zaznamenávat výstupy na tvrdý papír pomocí pera z husího brku. Když klasičtí ekonomové uvažovali o výrobních faktorech (půdě, práci a kapitálu), shromažďováním dat se prakticky nezabývali. V dalších dvou stoletích sice náklady na sbírání a použití dat poklesly, ale až do celkem nedávné doby zůstávaly poměrně vysoké.

Naše éra se liší tím, že mnohá základní omezení sběru dat postupně přestala platit. Technologie dospěla do stavu, kdy lze rozsáhlé objemy informací mnohdy zachycovat a zaznamenávat velmi levně. Data je často možné sbírat pasivně bez vynaložení velkého úsilí, nebo dokonce bez vědomí těch, jejichž

činnost kvantifikujeme. Vzhledem k tomu, jak zásadně poklesly náklady na ukládání, můžeme se také mnohem častěji rozhodnout, že data nezahodíme, ale ponecháme si je. Všechny tyto faktory způsobují, že je k dispozici stále více dat, která jsou přitom levnější než dosud. V průběhu uplynulého půlstoletí náklady na digitální ukládání dat klesaly přibližně na polovinu každé dva roky, zatímco hustota úložišť vzrostla 50 milionkrát. Pro informační firmy, jako je Farecast či Google – kde na jednom konci digitální výrobní linky vstupují netříděná fakta a na druhém konci vycházejí zpracované informace – začínají data připomínat spíše nový výrobní zdroj nebo surovinu.

Shromažďujeme-li data, je pro nás obvykle evidentní, jakou mají okamžitou hodnotu. Pravděpodobně je totiž sbíráme za tímto konkrétním účelem. Obchody zaznamenávají data o prodeji, aby mohly správně zpracovat své účetnictví. Továrny sledují svou produkci, aby měly jistotu, že splňuje standardy kvality. Weby protokolují každé uživatelské klepnutí myší – někdy dokonce i pohyby myšího ukazatele – aby mohly analyzovat a optimalizovat obsah, který uživatelům nabízejí. Těmito primárními aplikacemi dat lze zdůvodnit, proč data sbíráme a zpracováváme. Když Amazon zaznamenává nejen knihy, které zákazníci kupují, ale také webové stránky, které si pouze prohlížejí, předem ví, že na základě těchto dat může nabízet personalizovaná doporučení. Podobně Facebook sleduje uživatelské „aktualizace stavu" a klepnutí na „líbí se", aby mohl zobrazit nejvhodnější reklamy, které zajišťují jeho příjmy.

Oproti materiálním předmětům – jídlu, které sníme, nebo svíci, která dohoří – hodnota dat při jejich použití neklesá. Data můžeme zpracovat znovu a znovu. Informace patří mezi zboží, kterému ekonomové říkají „nevýhradní": když je konzumuje jedna osoba, nebrání přitom v přístupu jiným. Informace se při používání neopotřebuje stejně jako materiální zboží. Amazon proto může svým zákazníkům nabízet doporučení pomocí dat z minulých transakcí – a může tato doporučení používat opakovaně, nejen pro zákazníka, který data generoval, ale také pro mnoho jiných.

Data lze tedy mnohokrát použít ke stejnému účelu, ale ještě důležitější je to, že mohou posloužit také k více různým účelům. Chceme-li porozumět

tomu, jaká bude cena informací v éře veledat, má tento aspekt značný význam. Některé případy, kdy se uplatnil potenciál starých dat, jsme již uvedli – například když Walmart prohledal svou databázi starých účtenek a nalezl lukrativní korelaci mezi hurikány a prodejem sucharů Pop-Tarts.

To vše naznačuje, že celková hodnota dat je mnohem větší než hodnota, jakou získáme při jejich prvním použití. Znamená to také, že firmy mohou svá data využít efektivně i tehdy, když první nebo každá následná aplikace těchto dat poskytne jen nevelké přínosy. Stačí jen data zpracovat mnohokrát.

„Hodnota možností" dat

Chceme-li získat představu o tom, jak může opakované zpracování dat ovlivnit jejich výslednou hodnotu, podívejme se na elektromobily. Jejich úspěch při nasazení v dopravě závisí na velkém počtu logistických faktorů, které vesměs souvisejí s dojezdem na baterie. Řidiči potřebují možnost rychlého a pohodlného dobití svých autobaterií a energetické společnosti zase musí zajistit, aby energie odčerpávaná těmito vozidly nenarušila stabilitu sítě. V současnosti máme celkem efektivně rozmístěné benzinové stanice, ale zatím nevíme, jaké budou požadavky elektromobilů na kapacitu a polohu dobíjecích stanic.

Je zřejmé, že nejde ani tak o problém infrastrukturní, jako spíše informační. A důležitou součástí řešení jsou veledata. Při testech roku 2012 spolupracovala společnost IBM s kalifornskou firmou Pacific Gas and Electric Company a výrobcem automobilů Honda na shromažďování velkého objemu informací, které měly odpovědět na základní otázky související s elektrickými automobily: kdy a kde budou čerpat energii a co to bude znamenat pro rozvodné sítě. Společnost IBM vyvinula propracovaný prediktivní model založený na mnoha vstupech: úrovni baterie v automobilu, poloze vozidla, denní době a dostupných zásuvkách v blízkých dobíjecích stanicích. Data propojila s aktuální spotřebou energie v síti a také s historickými vzorci spotřeby energie. Analýzou velkých objemů čerstvých i historických dat z více zdrojů dokázala společnost IBM určit optimální časy a místa, kde by řidiči mohli dobíjet své autobaterie. Zjistila také, na kterých místech je nejvhodnější postavit dobíjecí stanice. Systém musí nakonec zohledňovat i cenové

rozdíly v blízkých dobíjecích stanicích. Do kalkulací je potřeba zahrnout i předpovědi počasí: možná je slunečno a blízká solární elektrárna poskytuje hodně energie, ale předpověď hlásí deštivý týden, kdy solární panely nebudou tolik účinné.

Systém přebírá informace generované k jednomu účelu a znovu je používá pro další účel – jinak řečeno, data přecházejí od primární k sekundární aplikaci. Díky tomu jejich hodnota časem roste. Indikátor úrovně baterie v elektromobilu informuje řidiče, kdy je potřeba dobíjet. Data o zatížení rozvodné sítě sbírá samotný dodavatel elektřiny, aby mohl zajistit stabilitu sítě. To jsou primární aplikace. Obě sady dat však nacházejí sekundární použití a získávají novou hodnotu, když je aplikujeme k dosažení zcela jiného cíle: abychom mohli zjistit, kdy a kde dobít baterie a kde postavit dobíjecí stanice pro elektromobily. Do výpočtů zahrnujeme i pomocné informace, jako je poloha automobilu a historická spotřeba v síti. Společnost IBM přitom data nezpracovává pouze jednou, ale opakovaně, protože svůj profil energetické spotřeby elektromobilu a jeho zátěže pro rozvodnou síť neustále aktualizuje.

Skutečná hodnota dat připomíná ledovou kru na hladině oceánu. Na první pohled vidíme jen její malou část, protože většina hmoty je skryta pod hladinou. Inovativní firmy, které si to uvědomují, mohou tuto skrytou hodnotu vytěžit a potenciálně získat značné výhody. Stručně řečeno, hodnotu dat musíme posuzovat s ohledem na všechny možné způsoby, kterými je lze v budoucnu uplatnit, nikoli pouze na to, co s nimi děláme v současnosti. Viděli jsme to na mnoha příkladech, které jsme dosud uvedli. Systém Farecast dokáže díky analýze dat z dříve prodaných letenek předpovídat jejich budoucí ceny. Společnost Google znovu použila staré vyhledávací dotazy, aby mohla sledovat výskyt chřipky. Maury přepsal data ze starých námořních deníků a díky nim zmapoval oceánské proudy.

Podniky ani zbytek společnosti však význam opakovaného použití dat zatím plně nedoceňují. Jen málokteří manažeři z firmy Con Edison v New Yorku by si asi dokázali představit, že sto let staré mapy kabelů a záznamy o údržbě mohou nějak pomoci při prevenci budoucích nehod. Aby se hodnota těchto dat mohla projevit, musela přijít nová generace statistiků

s moderní sadou metod a nástrojů. Dokonce i mnohé internetové a technologické firmy si až donedávna neuvědomovaly, jak cenné výsledky může opakované zpracování dat poskytnout.

Není od věci si data představit způsobem, jakým fyzikové pohlížejí na energii. Mluví o „uložené" neboli „potenciální" energii, kterou objekt obsahuje, ale která se nijak neprojevuje. Jako příklad lze uvést stlačenou pružinu nebo balón ležící na vrcholu kopce. Energie těchto objektů zůstává skrytá (potenciální), dokud ji neuvolníme – například tím, že pustíme konec pružiny nebo strčíme do balónu, aby se začal kutálet. Energie těchto objektů se nyní mění na „kinetickou", protože se pohybují a svou silou působí na jiné fyzické objekty. Po svém primárním použití hodnota dat nemizí, ale zůstává nevyužita. Data si ponechávají svůj potenciál jako stlačená pružina nebo balón na kopci, dokud je nevyužijeme novým způsobem a znovu jejich sílu neuvolníme. Ve věku veledat konečně máme k dispozici vše potřebné – správný myšlenkový přístup, znalosti i nástroje – abychom mohli utajenou hodnotu dat využít.

Hodnota dat je nakonec dána tím, kolik můžeme získat, když data nasadíme všemi možnými způsoby. Tyto zdánlivě neomezené potenciální aplikace jsou jako možnosti neboli volby. Hodnota dat se rovná součtu těchto voleb: dá se říci, že je to jakási „hodnota možností" dat. Když jsme v minulosti dosáhli svého hlavního cíle, často jsme data považovali za něco, co již splnilo svůj účel, a co můžeme vymazat nebo alespoň odsunout do archivu. Zdánlivě jsme totiž z dat extrahovali jejich klíčovou hodnotu. Ve věku veledat připomínají data spíše kouzelný diamantový důl, kde se po vytěžení objevují stále nové a nové drahé kameny. Hodnotu možností dat lze uvolnit třemi účinnými postupy: pomocí základního opakovaného použití, sloučením datových množin a nalezením kombinací typu „dva za cenu jednoho".

Opakované použití dat

Klasický příklad inovativního opakovaného použití dat představují vyhledávací termíny. Na první pohled se zdá, že když splní svůj primární účel, jsou tyto informace bezcenné. Při chvilkové interakci mezi uživatelem

a vyhledávačem vznikl seznam webů a inzerátů. Tento seznam splnil určitou funkci, která byla specifická pro daný okamžik. Staré vyhledávací dotazy však mohou mít mimořádnou cenu. Společnosti jako Hitwise, což je firma specializovaná na měření webového provozu, kterou vlastní zpracovatel dat Experian, umožňují klientům dolovat provoz vyhledávačů, aby mohli získat informace o spotřebitelských preferencích. Marketingoví pracovníci mohou pomocí služeb Hitwise odhadnout, zda bude nadcházející jaro v kurzu růžová barva, nebo zda se do módy vrátí černá. Vyhledávač Google uživatelům nabízí volně dostupnou verzi svého analytického nástroje vyhledávacích termínů. Ve spolupráci s druhou největší španělskou bankou BBVA spustil službu ekonomických předpovědí, která se zaměřuje na turistický sektor, a prodává také aktuální ekonomické indikátory založené na vyhledávaných datech. Bank of England pomocí vyhledávacích dotazů týkajících se nemovitostí získává lepší přehled o tom, zda ceny bydlení rostou či klesají.

Společnosti, které si neuvědomily význam opakovaného použití dat, často značně prodělaly. Například společnost Amazon na začátku svého působení uzavřela smlouvu s poskytovatelem internetového připojení AOL, že bude zajišťovat technologickou podporu jeho internetového obchodu. Většina lidí to považovala za obyčejnou dohodu o outsourcingu. Bývalý šéf vývojového oddělení společnosti Amazon Andreas Weigend však vysvětluje, že firmu ve skutečnosti zajímal hlavně přístup k datům o tom, co uživatelé služby AOL hledají a nakupují, aby mohla zlepšit výsledky svého modulu doporučování zboží. Společnost AOL na to ke své smůle nikdy nepřišla. Hodnotu dat posuzovala pouze z hlediska jejich primárního účelu – prodeje. Společnost Amazon však chytře dokázala získat značné výhody díky tomu, že data použila sekundárním způsobem.

Jiný příklad: společnost Google se pustila do oblasti rozpoznávání řeči a nabídla službu GOOG-411 pro místní vyhledávání, která fungovala mezi lety 2007 a 2010. Vyhledávací gigant tehdy neměl vlastní technologii rozpoznávání řeči, takže potřeboval získat příslušnou licenci. Podepsal smlouvu se zástupci úspěšné firmy Nuance, kteří byli nadšeni, že získali tak významného klienta. Firma Nuance však neměla ani páru o veledatech: smlouva

nijak neřešila, komu zůstanou záznamy o převodu hlasu na text, a Google si je ponechal. Analýzou dat mohl vyhodnotit pravděpodobnost, že daný digitalizovaný úsek řeči odpovídá určitému slovu. To má klíčový význam pro zlepšování technologie na rozpoznávání řeči nebo tvorbu úplně nových služeb. Firma Nuance se ovšem v té době zajímala o poskytování softwarových licencí a nikoli o analýzu nějakých dat. Ihned poté, co si uvědomila svou chybu, začala uzavírat dohody s mobilními operátory a výrobci mobilních telefonů, kteří měli zájem o její službu rozpoznávání řeči, aby tak mohla data shromažďovat sama.

Fakt, že při opakovaném použití lze z dat vyždímat další hodnotu, je dobrou zprávou pro organizace, které shromažďují nebo kontrolují velké objemy dat, ale momentálně je samy příliš nevyužívají. Může se jednat například o tradiční podniky, které fungují převážně offline. V některých případech mohou vlastnit nevytěžené informační poklady. Některé společnosti shromáždily data, použily je jednou (pokud vůbec) a poté je vzhledem k nízkým nákladům na úložiště pouze skladovaly v takzvaných „datových hrobech", jak datoví experti označují místa, kde takové staré informace spočívají.

Internetové a technologické firmy jsou na čele úsilí o zpracování datové záplavy, protože sbírají spoustu dat jen díky tomu, že jsou online, a při analýze těchto dat mají oproti zbytku oboru náskok. Prospěch však mohou mít všechny společnosti. Konzultanti ve společnosti McKinsey & Company zmiňují firmu z oboru logistiky (její název nesdělují), která si všimla, že během doručování zásilek získává mimořádný objem informací o dodávkách produktů po celém světě. Tato logistická firma vycítila příležitost a zřídila speciální divizi, aby mohla agregovaná data prodávat ve formě obchodních a ekonomických předpovědí. Jinými slovy vytvořila offline verzi podnikání společnosti Google se starými vyhledávacími dotazy. Případně můžeme zmínit společnost SWIFT, která provozuje globální mezibankovní systém pro převody peněz. Zjistila, že platby korelují s globální ekonomickou aktivitou. Společnost tedy nabízí předpovědi HDP založené na datech o finančních převodech, které procházejí její sítí.

Některé firmy mohou díky své pozici v řetězci zpracování informací shromažďovat velké množství dat, ačkoli je momentálně příliš nepotřebují nebo je nedokážou efektivně zpracovávat. Například operátoři mobilních sítí zaznamenávají informace o poloze svých zákazníků, aby mohli směrovat jejich hovory. Tyto společnosti mají pro uvedená data jen omezené technické uplatnění. Stejná data však najednou získávají na hodnotě, když je opakovaně používají firmy distribuující personalizované reklamy a zvýhodněné nabídky, které jsou závislé na poloze. Hodnota někdy nepochází z jednotlivých datových bodů, ale z toho, co ukazuje jejich souhrn. Firmy působící v oblasti geolokace jako AirSage či Sense Networks, s nimiž jsme se setkali v předchozí kapitole, tedy mohou prodávat informace o tom, kde se shromažďují lidé v pátek večer nebo jak pomalu se vlečou auta v dopravní zácpě. Pomocí těchto celkových informací lze určit hodnotu určité nemovitosti nebo reklamy na konkrétním billboardu.

Dokonce i ty nejbanálnější informace mají zvláštní cenu, pokud je aplikujeme správným způsobem. Vraťme se k mobilním operátorům: mají záznamy o tom, kde a kdy se telefony připojují k základnovým stanicím, včetně údajů o síle signálu. Operátoři pomocí těchto dat již odedávna vylepšují výkon svých sítí a určují, kde je vhodné přidat nový vysílač nebo posílit infrastrukturu. Stejná data však mají také mnoho dalších potenciálních aplikací. Výrobci mobilních přístrojů se z nich mohou dozvědět, co ovlivňuje sílu signálu, a mohou tak například zvýšit kvalitu příjmu svých výrobků. Mobilní operátoři jsou při prodeji těchto informací tradičně opatrní, protože se obávají, že by mohli porušit předpisy na ochranu osobních údajů. Svůj postoj však postupně zmírňují spolu s tím, jak se zhoršuje jejich finanční situace a vidí, že data by mohla sloužit jako další zdroj jejich příjmu. Velká španělská společnost Telefónica, která působí i v dalších zemích, dokonce roku 2012 vytvořila samostatnou firmu s názvem Telefónica Dynamic Insights, která se specializuje na prodej anonymních a agregovaných dat o poloze uživatelů mobilní sítě obchodníkům a dalším zájemcům.