



Kniha roku
podle
The Economist

HOST

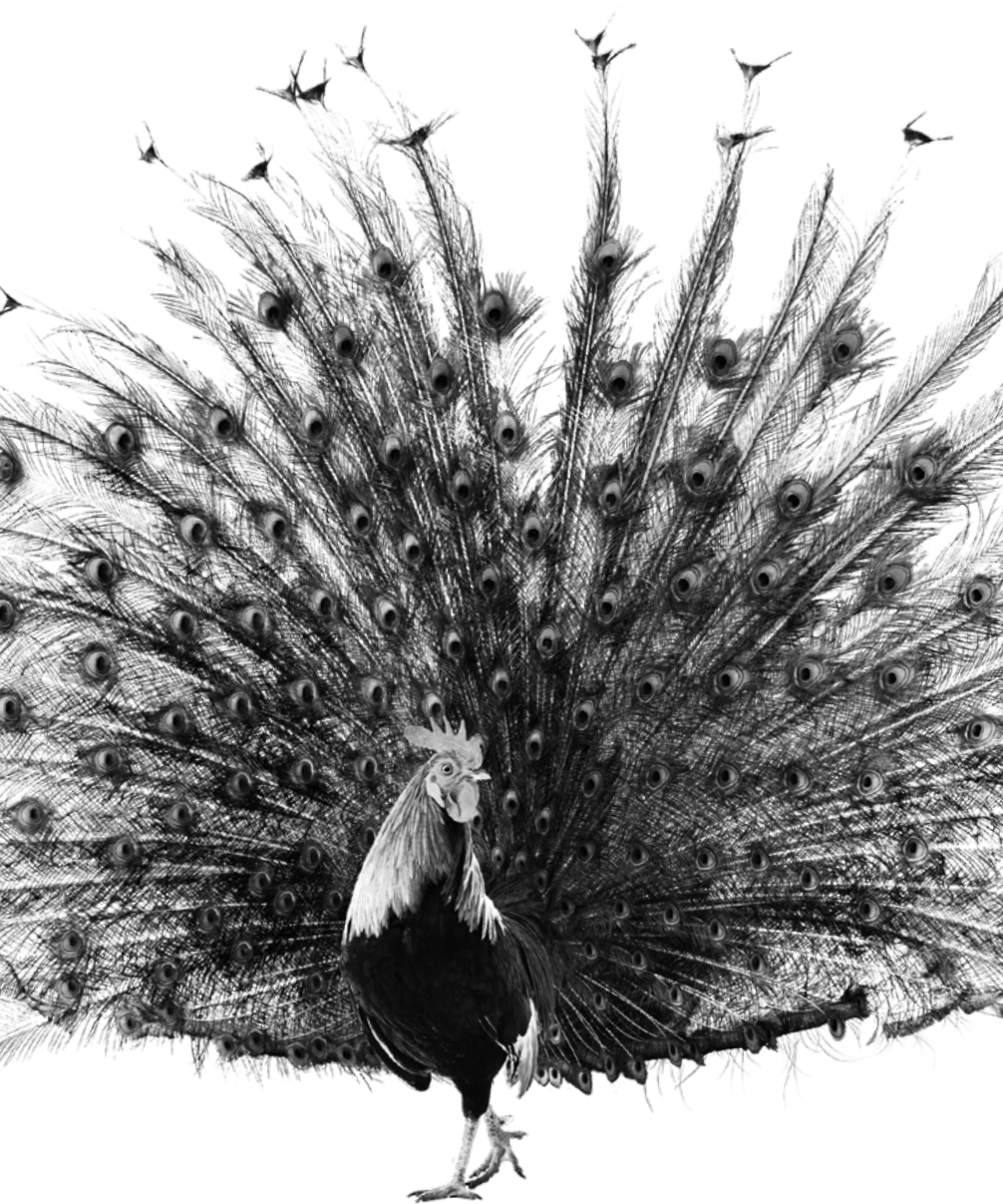
SETH STEPHENS-DAVIDOWITZ

VŠICHNI LŽOU

**VELKÁ DATA, CHYTRÉ ALGORITMY A JAK NÁM
INTERNET MŮŽE ŘÍCT, KDO OPRAVDU JSME**

Seth Stephens-Davidowitz

VŠICHNI LŽOU



SETH STEPHENS-DAVIDOWITZ

VŠICHNI LŽOU

**Velká data, chytré algoritmy a jak nám internet
může říct, kdo opravdu jsme**



Brno 2019

Everybody Lies

Copyright © 2017 by Seth Stephens-Davidowitz

Translation © Romana Hegedüsová, 2019

Czech edition © Host — vydavatelství, s. r. o., 2019

(elektronické vydání)

ISBN 978-80-7577-865-9 (PDF)

ISBN 978-80-7577-866-6 (ePUB)

ISBN 978-80-7577-867-3 (MobiPocket)

Mámě a tátovi

OBSAH

Předmluva Stevena Pinkera	11
Úvod: Náznak revoluce	15

VELKÁ A MALÁ DATA

1 Vaše chybná intuice	39
------------------------------	-----------

JAK MOCNÁ JSOU VELKÁ DATA

2 Měl Freud pravdu?	59
3 Co jsou a nejsou data	69
Těla jako data	76
Slova jako data	88
Fotografie jako data	109
4 Digitální sérum pravdy	116
Pravda o sexu	123

Pravda o nenávisti a předsudcích	138
Pravda o internetu	151
Pravda o zneužívání dětí a potratech	155
Pravda o vašich přátelích na Facebooku	161
Pravda o vašich zákaznících	164
Umíme zacházet s pravdou?	169
5 Je čas zaostřit	175
Co se doopravdy děje v našich městech a okresech?	182
Jak trávíme minuty a hodiny svého času	198
Naši dvojníci	206
Příběhy ukryté v datech	214
6 Svět je jedna velká laboratoř	216
Abeceda A/B testování	218
Kruté, ale poučné experimenty	230
 VELKÁ DATA: POZOR, KŘEHKÉ!	
7 Spousta dat, spousta braku? Co velká data nedokážou	253
Prokletí dimenzionality	256
Přehnaný důraz na vše měřitelné	262
8 Fůra dat, fůra problémů? Čemu bychom se měli vyhnout	267
Nebezpečí příliš mocných korporací	267
Nebezpečí příliš mocného státu	276
Závěr: Kolik lidí dočte knihu až do konce?	281
Poděkování	294
Poznámky	298

PŘEDMLUVA

Už od dob, kdy si filozofové pohrávali s myšlenkou vytvořit „cerebroskop“, bájný přístroj, který by dokázal promítnout myšlenky na plátno, se sociologové a psychologové snaží najít nástroje, s jejichž pomocí by bylo možné dobrat se podstaty fungování lidské mysli. Dříve jsem působil jako experimentální psycholog a za tu dobu se takových nástrojů vystřídalo na výsluní hned několik. Vyzkoušel jsem je všechny: ratingové stupnice, měření reakční doby, dilataci zornic, funkční neurozobrazování, dokonce práci s pacienty s epilepsií a implantovanými elektrodami, kteří byli šťastní, že si mohou jazykovým experimentem zkrátit dobu čekání na další záchvat.

Nicméně žádná z těchto metod neumožňuje volný přístup do lidské mysli. Problémem je obrovská ztráta dat. Lidské myšlenky představují složitou veličinu; Woody Allen prošel kurzem rychločtení a po přečtení *Vojny a míru* prohlásil: „Bylo to něco o Rusech.“ To si my dovolit nemůžeme. Jenže myšlenky se v celé své spleť multidimenzionální kráse velice těžko

vědecky analyzují. Jistě, když si člověk vylije srdce, dokážeme zachytit rozmanitost jeho proudu vědomí, jenže monolog nepředstavuje vhodná data pro testování hypotéz. Jestliže se na druhou stranu soustředíme pouze na hodnoty, které jsou snadno kvantifikovatelné, jako je například reakční doba na určitá slova nebo napětí pokožky v reakci na nějaké obrazy, snadno vytvoříme statistiky, zároveň však zredukujeme složitou strukturu lidského poznání na jediné číslo. I ta nejsložitější metoda neurovizualizace nám sice umí říct, jak vypadá myšlenka rozložená v 3D prostoru, ale nesdělí nám, co je jejím obsahem.

Vědce zabírající se lidskou myslí pronásledují nejen nutné kompromisy mezi zpracovatelností a bohatstvím myšlenek, ale navíc ještě takzvaný zákon malých čísel. Amos Tversky a Daniel Kahneman tak nazvali chybnou představu o tom, že rysy populace se projeví v jakémkoli jejím vzorku bez ohledu na jeho velikost. Dokonce i ti matematicky nejnadanější vědci se drží žalostně nesprávného tušení týkajícího se otázky, kolik je pro vědeckou studii potřeba zkoumaných subjektů, aby bylo možné její závěry — po odstranění všech náhodných podivností a výstřelků — zobecnit kupříkladu na všechny Američany, nemluvě o *homo sapiens*. Celé je to ještě ošemetnější, pokud není výběr takového vzorku náhodný, ale jde třeba o studenty druhých ročníků nalákané na peníze na pivo.

Tato kniha představuje zcela nový způsob, jak studovat lidskou mysl. Velká data z internetových vyhledávačů a další online data nejsou tak úplně to stejné jako zmíněný cerebroskop, ale Seth Stephens-Davidowitz dokázal, že skutečně nabízí bezprecedentní možnost, jak nahlédnout do lidské duše. V soukromí svých klávesnic se lidé přiznávají k těm nejpodivnějším věcem. Někdy proto, že mají nějaký význam v reálném životě (jako například na stránkách internetových seznamek nebo

při hledání profesionální rady), a někdy právě proto, že to žádný dopad na reálný život nemá: lidé tak odhazují svoje starosti nebo sdílejí svá přání, aniž by museli čelit vyděšeným reakcím nebo něčemu ještě horšímu. Ať tak či onak, tyto lidé nemačkají jen tlačítka na klávesnici. Oni zadávají biliony různě seřazených znaků, jimiž podrobně vysvětlují svoje myšlenky v celém jejich výbušném a kombinatorním rozsahu. A co je ještě lepší, veškeré digitální stopy zanechávají ve formě, ve které se snadno shromažďují i analyzují. Přicházejí ze všech oblastí života. Mohou se stát součástí nenápadných experimentů, ve kterých se obměňují podněty a odezva na ně se zaznamenává v reálném čase. A lidé svoje data naprosto bezelstně poskytují, a to v množství přímo gargantuovském.

Kniha *Všichni lžou* přináší více než jen ověření koncepce. Objevy Stephense-Davidowitze znovu a znovu převrací moje vlastní předsudky a představy o této zemi a mých spoluobčanech vzhůru nohama. Kde se vzala nečekaná podpora Donalda Trumpa? Když se Ann Landersová v roce 1976 ptala svých čtenářů, zda litují toho, že mají děti, a k jejímu zděšení jí většina odpověděla, že ano, šlo o mylnou informaci způsobenou nereprezentativním vzorkem a jeho samovýběrem? Může internet za tak často zmiňovanou krizi před více než deseti lety, za vytvoření „informační bubliny“? Co je spouštěčem zločinů z nenávisti? Vyhledávají lidé vtipy, aby se rozveselili? A přestože o sobě rád tvrdím, že se mnou jen tak něco neotřese, to, co internet odhalil o lidské sexualitě, mě dost šokovalo. Mimo jiné šlo například o zjištění, že každý měsíc vyhledává jistý počet žen „pohlavní styk s plyšovými zvířátky“. Žádné měření reakční doby, dilatace zornic ani funkční neurozobrazování by nikdy nedokázaly tento fakt odhalit.

Kniha *Všichni lžou* je pro všechny. Stephens-Davidowitz s neutuchající zvědavostí a hřejivým humorem poukazuje na nové

způsoby zkoumání na poli společenských věd v jednadvacátém století. A pokud máme k dispozici tento nekonečně fascinující pohled do světa lidských obsesí, k čemu by nám byl cerebroskop?

Steven Pinker, 2017

ÚVOD

NÁZNAK REVOLUCE

Samozřejmě že prohraje, říkali.

Experti na předvolební průzkumy došli po republikánských primárnkách v roce 2016 k závěru, že Donald Trump nemá nejmenší šanci. Vždyť stihl pobouřit tolik menšinových skupin. Průzkumy volebních preferencí i jejich interpreti nás utvrzovali v tom, že takovou nehoráznost skutečně schvaluje jen málo Američanů.

Většina odborníků se v té době přikláněla k názoru, že Trump v prezidentských volbách prohraje. Příliš mnoho potenciálních voličů vyjádřilo nesouhlas s Trumpovým vystupováním i jeho názory.

Určité náznaky, že Trump vyhraje jak primární volby, tak volby prezidentské, ale přece jen existovaly — na internetu.

Jsem odborník na internetová data. Každý den sleduji digitální stopy, které po sobě lidé zanechávají na síti. Ze všech našich úhozů na klávesnici a kliknutí myši se snažím vyčíst a pochopit, co skutečně chceme, čím se opravdu zabýváme a kdo vlastně

jsme. Dovolte mi vysvětlit, jak jsem se k této neobvyklé profesi dostal.

Všechno to začalo — zdá se to tak hrozně dávno — v průběhu prezidentských voleb v roce 2008. V té době sociálními vědci již dlouho rezonovala jedna otázka: Jak velkou roli hrají ve Spojených státech rasové předsudky?

Barack Obama byl prvním afroamerickým kandidátem na prezidentský úřad, kterého kdy nominovala jedna z hlavních politických stran. Vyhrál, a docela hladce. Průzkumy veřejného mínění naznačovaly, že v tom, jak se Amerika rozhodovala, rasová otázka nehrála roli. Například agentura Gallup provedla řadu průzkumů veřejného mínění před Obamovým zvolením i po něm. A s čím přišla? Americkým voličům bylo do značné míry jedno, že Barack Obama je černoch.¹ Krátce po volbách se dva známí profesori z Kalifornské univerzity v Berkeley společně ponořili do dalších dat² získaných v jiných šetřeních a k jejich analýze použili ještě sofistikovanější techniky. Došli k obdobným závěrům.

Během Obamovy vlády se tyto výsledky staly součástí všeobecného mínění šířeného napříč masovými médii i převládajícím názorem v akademických kruzích. Zdroje, které média a sociální vědy přes osmdesát let využívají k lepšímu pochopení světa kolem nás, nám tvrdily, že když se Američané rozhodovali, zda má být Obama prezidentem, nebo ne, většině z nich nevadilo, že je černoch.

Zdálo se, že tato země, ač pošpiněná dlouhými léty otroctví i zákony Jima Crowa, konečně přestala soudit lidi podle barvy kůže. Zároveň to vypadalo, že rasismus v Americe mele z posledního. Někteří experti začali dokonce prohlašovat, že žijeme v postrasové společnosti.³

V roce 2012 jsem studoval magisterský obor ekonomie. V životě jsem byl úplně ztracený, bez sebemenšího zájmu o studium,

ovšem jinak sebejistý, až nafoukaný, a měl jsem pocit, že velice dobře rozumím tomu, jak v jednadvacátém století funguje svět, co si lidé myslí a co je zajímavé. A když jsem tehdy narazil na otázku předsudků, uvěřil jsem — na základě všeho, co jsem znal z psychologie a politologie —, že explicitní rasismus se týká pouze malého procenta Američanů, z nichž většina patří mezi konzervativní republikány žijící daleko na jihu.

A pak jsem objevil Google Trends.

Google Trends je nástroj, který byl bez většího zájmu spuštěn v roce 2009. Svým uživatelům umí ukázat, jak často jsou v různých částech světa a v různých časových obdobích vyhledávána určitá slova a fráze. Původně měl Google Trends sloužit jen pro zábavu, přátelé si třeba mohli zjišťovat, která celebrita je momentálně nejpopulárnější nebo co se právě nosí. Jedny z prvních verzí dokonce obsahovaly žertovné upozornění, že nejde o data, „která by lidé chtěli využít při psaní svých disertací“. To mě okamžitě motivovalo k tomu, abych s jejich využitím napsal tu svoji.*

V té době se vůbec zdálo, že informace nalezené na Googlu nelze považovat za „seriózní“ pro vědecký výzkum. Na rozdíl od průzkumů veřejného mínění nebyl vyhledávač Google vytvořen

* Z Google Trends pocházela většina mých dat. Tento nástroj ovšem umožňuje pouze srovnání relativních četností různých vyhledávání, absolutní počty konkrétních vyhledávání vám neukáže. Proto jsem čerpal informace také z Google AdWords, který přesně udává, jak často se provádí určité vyhledávání. Ve většině případů jsem také dokázal svá data specifikovat pomocí vlastního algoritmu založeného na Google Trends, který přesně popisují ve své disertační práci „Essays Using Google Data“ (volně přeloženo jako Eseje s využitím dat z Googlu — pozn. překl.) a v práci publikované v časopise *Journal of Public Economics* s názvem „The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data“ (volně přeloženo jako „Jak platí černošský kandidát za rasovou nesnášenlivost: Důkazy založené na datech z Googlu“ — pozn. překl.). Disertační práce, odkaz na článek i podrobný popis všech dat a kódování, které jsem v původním výzkumu použil a které zmiňuji v této knize, jsou dostupné na mých webových stránkách sethsd.com.

proto, abychom lépe porozuměli lidské duši. Google vyvinuli proto, aby se lidé poučili o tomto světě, nikoli proto, aby se vědci mohli dozvědět něco o lidech. Jenže se ukázalo, že stopy, které po sobě na internetu zanecháváme při své cestě za poznáním, toho o nás prozrazují víc než dost.

Jinými slovy, vyhledávání informací je samo o sobě jistou informací o člověku. Ukázalo se, že to, kdy a kde lidé hledají fakta, citace, vtipy, pomoc, informace o různých místech či věcech, nám prozrazuje, co si skutečně myslí, po čem touží, čeho se obávají a čím se doopravdy zabývají, a to v mnohem větší míře, než by kdy koho napadlo. Platí to zejména z toho důvodu, že někdy lidé na Googlu ani tak nevyhledávají, jako se mu svěrují: „Nesnáším svého šéfa.“ „Jsem namol.“ „Táta mi dal facku.“

Každodenní akt zadání nějakého slova či fráze do bílého obdélníkového okénka za sebou zanechá malou pravdivou stopu. Vynásobte ji milionem a vyjeví se vám zásadní skutečnosti. První slovo, které jsem vyhledával na Google Trends, bylo „Bůh“. Zjistil jsem, že nejvíc se toto slovo vyhledává v Alabamě, Mississippi a v Arkansasu — tedy v oblasti takzvaného biblického pásu. A hlavně v neděli. Na tom nebylo nic překvapivého, zajímavé však bylo, že data z vyhledávače dokázala odhalit tak jasný vzorec. Potom jsem zkusil zadat název newyorského basketbalového týmu „Knicks“. Ukázalo se, že tento termín je na Googlu nejvyhledávanější v New York City. Taký žádná záhada. Pak jsem naťukal svoje jméno. „Je nám líto,“ informoval mě Google Trends, „není k dispozici dostatečný počet vyhledávání.“ Zjistil jsem, že Google Trends vám dokáže poskytnout data jen v případech, kdy velké množství lidí vyhledávalo totéž.

Ovšem síla vyhledávače Google netkví v tom, že nám dokáže říct, že Bůh je populární v oblastech na jihu Spojených států, Knicks jsou oblíbení v New York City a já nejsem bůhvíjak populární nikde. Ke stejným výsledkům by došel jakýkoli jiný

průzkum. Síla dat z Googlu tkví v tom, že lidé tomuto obřímu vyhledávači svěří věci, které možná neřekli nikdy nikomu jinému.

Vezměte si například sex (tomuto tématu se budu věnovat mnohem podrobněji dále v knize). Klasickým šetřením se nedá věřit, že přináší pravdivé informace o našem sexuálním životě. Analyzoval jsem data sesbíraná v rámci General Social Survey (Všeobecný sociální průzkum).⁴ Tento sociologický průzkum je považován za jeden z nejvlivnějších a nejautoritativnějších zdrojů informací o chování Američanů. Podle něj heterosexuální ženy přiznávají, že mají sex v průměru pětapadesátkrát do roka, z toho v šestnácti procentech případů s použitím kondomu. Když to sečtete, vyjde vám 1,1 miliardy použitých kondomů ročně. Ovšem heterosexuální muži tvrdí, že jich spotřebují 1,6 miliardy ročně. Tato čísla by ze své podstaty měla být stejná. Kdo tedy mluví pravdu? Muži, nebo ženy?

Jak se ukázalo, ani muži, ani ženy. Podle globální informační a analytické společnosti Nielsen, která sleduje chování spotřebitelů, se ročně prodá méně než šest set milionů kondomů.⁵ Takže lžou všichni; otázkou zůstává, jak moc.

Lhaní je ve skutečnosti velice rozšířené. Muži, kteří nikdy nebyli ženatí, prohlašují, že ročně spotřebují průměrně devětdvacet kondomů. Kdybychom to sečetli, vyšlo by nám víc, než je celkový počet kondomů prodaných v celých Spojených státech všem sezdaným i svobodným lidem dohromady. Sezdaní lidé pravděpodobně také přehánějí, když tvrdí, jak často mají sex. Průměrní ženatí muži mladší pětasedesáti let v průzkumech uvádějí, že mají sex jednou týdně. Jen jedno procento přiznává, že za celý předešlý rok neměli sex ani jednou. Vdané ženy udávají o něco nižší četnost souloží, ale ne o moc.

Data sebraná na Googlu nám přináší mnohem méně veselý — ovšem já tvrdím, že mnohem přesnější — obrázek o tom,

jak vypadá sexuální život sezdaných párů. Na Googlu je nejčastější stížností na sexuální život „žádný sexuální život“. Termín „manželství bez sexu“ je vyhledáván třikrát častěji než „nešťastné manželství“ a osmkrát častěji než „manželství bez lásky“. Dokonce i nesezdané dvojice si docela často stěžují na to, že spolu nespí. Vyhledávaný termín „vztah bez sexu“ je hned na druhém místě za termínem „zneužívání ve vztahu“. (Rád bych zdůraznil, že všechna uvedená data jsou anonymní. Google samozřejmě neshromažďuje a neprezentuje data o vyhledávání žádného konkrétního jedince.)

A data z Googlu přinesla také obrázek Ameriky, který se diametrálně lišil od oné postrasové utopie, kterou tak umně vykreslily průzkumy veřejného mínění. Pamatuji si, jak jsem na Google Trends poprvé napsal slovo „negr“. Jak jsem byl naivní. Skutečně jsem věřil, že to slovo je natolik toxické, že počet jeho vyhledávání bude nízký. Vedle jak ta jedle. Ve Spojených státech slovo „negr“ nebo jeho množné číslo „negři“ patří co do počtu vyhledávání do stejné skupiny jako slova „migréna“, „ekonom“ a jméno losangeleského basketbalového týmu „Lakers“. Říkal jsem si, že výsledky jsou možná zkreslené vyhledáváním textů rapových písní. Nebylo tomu tak. Slovo, které v těchto textech většinou najdete, je „nigga“ (jedna z hovorových verzí hanlivého „nigger — negr“ — pozn. překl.). Co tedy stálo za tím, že si Američané na Googlu vyhledávali slovo „negr“? V řadě případů šlo o hledání vtípů, které zesměšňují Afroameričany. Dokonce dvacet procent veškerých vyhledávání obsahujících slovo „negr“ obsahovalo také slovo „vtip“. Mezi další často vyhledávané fráze patřilo „blbí negři“ a „nesnáším negry“.

Podobných vyhledávání byly miliony ročně. Pěkná řádka Američanů v teple svých domovů pátrala na internetu po neskutečně rasistických informacích. A čím víc jsem hledal, tím horší to bylo.

Zatímco po Obamově prvním zvolení se většina komentářů soustředila na jeho chválu a historický dopad takového výsledku voleb, zhruba jedno ze sta vyhledávání jména „Obama“ na Googlu obsahovalo také „KKK“ nebo „negr/negři“. Možná se vám to nezdá tak moc, ale uvědomte si, že existují tisíce nerasistických důvodů, proč na Googlu hledat informace o tomto mladém outsiderovi s okouzující rodinou, který měl co nevidět obsadit nejmocnější pozici na světě. V den voleb se zdeseťinásobilo vyhledávání webu Stormfront,⁶ internetového fóra s neonacistickou tematikou, které je v USA překvapivě velice populární, a počet přihlašování na něm. V některých státech se termín „prezident negr“ vyhledával častěji než „první černošský prezident“.⁷

Tradičním zdrojům informací tato temná a nenávistná stránka věci unikla, ovšem z toho, co lidé vyhledávali na internetu, byla jasně patrná.

Uvedená vyhledávání neodpovídají společnosti, ve které hraje rasismus jen malou roli. V roce 2012 jsem o Donaldu J. Trumpovi věděl jen to, že je to byznysmen, který se objevuje v různých reality show. Stejně jako spousta dalších lidí jsem netušil, že z něj za čtyři roky bude vážný kandidát na prezidenta. Ale ta ohavná vyhledávání na Googlu už není tak těžké dát dohromady s úspěchem kandidáta, který svými útoky na imigranty, svojí zlostí a nenávistí tak úspěšně hraje na tu nejtemnější lidskou strunu.

Data z Googlu nám také prozradila, že většina z toho, co jsme si mysleli o lokalitách, kde bují rasismus, byl omyl. Průzkumy a veřejné mínění umístily moderní formu rasismu hlavně na jih a přisoudily ho z větší části republikánům. Ovšem mezi místa s největším počtem vyhledávání s rasistickou tematikou patřily i sever státu New York, západní Pensylvánie, východní

Ohio, průmyslové oblasti Michiganu a zemědělské části Illinois, stejně jako Západní Virginie, jižní Louisiana a stát Mississippi. Skutečná dělicí čára, jak ukázala data z Googlu, neležela mezi severem a jihem; ale šlo o východ versus západ. Západně od Mississippi už se s touto formou rasismu moc nepotkáte. A rozhodně nebyl omezený pouze na republikány. Ve skutečnosti nebylo vyhledávání s rasistickou tematikou o nic častější v oblastech s vyšším procentem republikánů než v oblastech s vyšším výskytem demokratů. Data vyhledávaná na Googlu pomohla překreslit mapu výskytu rasismu ve Spojených státech — a velice se lišila od toho, co byste očekávali. Republikáni na jihu se s větší pravděpodobností k rasismu přiznají, ovšem postoje demokratů ze severu jsou velice podobné, ač zůstávají nevyřčené.

Za čtyři roky tato mapa sehrála důležitou roli při analýze Trumpova politického úspěchu.

V roce 2012 jsem použil mapu rasismu, kterou jsem vyvinul pomocí dat z Googlu, a opětovně prozkoumal, jak důležitá byla barva Obamovy kůže. Data mluvila jasně. V oblastech s vyšším počtem vyhledávání termínů s rasistickou tematikou si Obama vedl podstatně hůř než o čtyři roky dříve John Kerry, demokratický prezidentský kandidát bílé pleti. Tuto souvislost nevysvětloval žádný jiný faktor týkající se těchto oblastí. Podstatná nebyla ani úroveň vzdělání, ani věk, ani příslušnost k církvi nebo zda byl volič majitelem zbraně či nikoli. Vyhledávání s rasistickou tematikou nepředpověděla volební nezdar žádnému jinému demokratickému kandidátovi. Pouze Obamovi.

A výsledky naznačily, že to skutečně mělo obrovský vliv. Obama ztratil na celostátní úrovni přibližně čtyři procentní body jen kvůli neskrývanému rasismu. Bylo to mnohem víc, než by se dalo očekávat na základě jakéhokoli průzkumu. Jistě, Barack Obama se stal prezidentem Spojených států dvakrát po sobě a pomohly mu v tom i podmínky příznivé pro Demokratickou

stranu. Musel ovšem překonat mnohem více překážek, než si uměl představit kdokoli spoléhající se na tradiční zdroje dat — a to byli skoro všichni. Rasistů bylo dost na to, aby v období nepřítliš nakloněném demokratům vyhráli primárky či zvrátili průběh všeobecných voleb.

Moji studii zpočátku odmítlo pět odborných časopisů.⁸ Řada oponentů (odpusťte mi mírné rozladění) tvrdila, že lze jen stěží uvěřit, že by tolik Američanů mohlo být tak krutými rasisty. To prostě nesedělo s tím, co se mezi lidmi povídá. Navíc jim údaje získané z Googlu připadaly jako dost bizarní soubor dat.

Poté co jsme byli svědky prezidentské inaugurace Donalda J. Trumpa, vypadají moje zjištění přece jen trochu důvěryhodněji.

Čím víc jsem se do této problematiky nořil, tím víc jsem zjišťoval, že Google skýtá spoustu informací, které běžné průzkumy nezachytí, a přitom mohou být velice užitečné pro toho, kdo se snaží pochopit volby (a samozřejmě mnoho dalších témat).

Najdete tam informace o tom, kdo doopravdy přijde volit. Více než polovina lidí, kteří nevolí, v průzkumech těsně předcházejících volbám tvrdí, že mají v úmyslu se k volbám dostavit, což zkresluje odhad volební účasti. Zatímco termíny jako „jak volit“ či „kde mohu volit“ vyhledávané na Googlu několik týdnů před volbami dokážou přesně předpovědět, ve kterých částech země lze očekávat vysokou volební účast.

Možná se tam skrývá také informace o tom, koho tito lidé budou volit. Je skutečně možné předpovědět, kterého kandidáta budou lidé volit, jen na základě toho, co vyhledávají na internetu? Je jasné, že nelze jednoduše sledovat, kteří kandidáti jsou nejčastěji vyhledáváni. Řada lidí si informace o svém kandidátovi hledá proto, že ho má ráda. Stejný počet lidí ho hledá, protože ho nesnáší. Společně s profesorem Stuartem Gabrielem,

přednášejícím v oboru financí na Kalifornské univerzitě v Los Angeles, jsme objevili překvapivý klíč k tomu, jak lidé plánují, komu dají ve volbách svůj hlas. Velké procento vyhledávání s tematikou prezidentských voleb obsahuje jména obou kandidátů. Během souboje mezi Donaldem Trumpem a Hillary Clintonovou v roce 2016 někteří lidé vyhledávali slovní spojení „volby Trump Clintonová“. Jiní hledali sestřihy z kandidátských debat a vyhledávali „debata Clintonová Trump“. V podstatě celých dvanáct procent vyhledávaných hesel obsahujících „Trump“ obsahovalo i jméno „Clintonová“. A naopak více než jedna čtvrtina vyhledávání obsahujících „Clintonová“ obsahovalo i jméno „Trump“.

Zjistili jsme, že tato zdánlivě neutrální vyhledávání nám ve skutečnosti mohou napovědět, komu dává vyhledávající přednost. Jak? Podle pořadí, v jakém uvádí jména obou kandidátů. Náš výzkum naznačuje, že ve vyhledávaném hesle obsahujícím obě jména člověk s mnohem větší pravděpodobností uvede na prvním místě kandidáta, kterého podporuje.

Ve trojích předcházejících prezidentských volbách v USA dostal nejvíc hlasů ten kandidát, který ve větším počtu vyhledávání figuroval na prvním místě. Ještě zajímavější bylo, že z pořadí, v jakém byli kandidáti uváděni ve vyhledáváních, bylo možné usuzovat, jakou cestou se konkrétní stát vydá.

Zdá se také, že pořadí, v jakém jsou kandidáti uvedeni při vyhledávání, může obsahovat i informace, které běžný průzkum nezachytí. Nate Silver, novinář a statistik, během prezidentských voleb mezi Obamou a republikánem Mittem Romneyem v roce 2012 naprosto mistrovsky odhadl výsledky ve všech padesáti státech. My jsme nicméně přišli ještě na to, že v těch státech, ve kterých lidé uváděli ve vyhledáváních Romneyho před Obamou, si Romney vedl mnohem lépe, než předpověděl Silver. Ve státech, kde byl nejčastěji uváděn Obama před Romneyem, na tom byl nakonec Obama lépe, než Silver očekával.

Tento ukazatel může obsahovat informace, které průzkumy veřejného mínění nezachytí. Buď proto, že voliči sami sobě lžou, nebo ve veřejných průzkumech neradi odhalují svoje skutečné preference. Kdyby tedy v roce 2012 tvrdili, že ještě nejsou rozhodnutí, koho budou volit, a přitom by soustavně vyhledávali hesla jako „volby Romney Obama“, „prezidentská debata Romney Obama“ či „průzkum preferencí Romney Obama“, možná ve skutečnosti po celou dobu plánovali dát svůj hlas Romneymu.

Předpověděl tedy Google Trumpovo vítězství? Čeká nás ještě hodně práce (a bude potřeba ještě mnohem více výzkumů), než se naučíme, jak data z Googlu co nejlépe využít k přesné predikci výsledků voleb. Jedná se o zcela novou vědeckou disciplínu a v tuto chvíli máme k dispozici data sesbíraná pouze z několikaletých předešlých voleb. Rozhodně netvrdím, že si můžeme dovolit jednou provždy zahodit průzkumy veřejného mínění jako nástroj k předpovídání volebních výsledků.

Ovšem v mnoha oblastech bylo možné zaregistrovat různá znamení, že si Trump nakonec povede mnohem lépe, než tyto průzkumy napovídaly.

Během všeobecných voleb existovaly náznaky, že voliči dají Trumpovi přednost. Černí Američané v předvolebních průzkumech tvrdili, že se k volbám dostaví ve velkém počtu, aby se postavili proti Trumpovi. Jenže počet vyhledávání informací o volbách byl v převážně černých oblastech velice nízký. Což naznačovalo, že v den voleb by nízká účast černého obyvatelstva poškodila Clintonovou.

Bylo možné zachytit i známky toho, že údajně stále nerozhodnutí voliči nakonec dají hlas Trumpovi. Společně s Gabrielem jsme zjistili, že se na Googlu vyhledávalo více hesel „Trump Clintonová“ než naopak, a to v klíčových státech amerického Středozápadu, kde Clintonová očekávala vítězství. A skutečně,

Trump za svoje zvolení vděčí faktu, že právě tam dostal výrazně vyšší počet hlasů.

Já však tvrdím, že nejzásadněji poukazoval na Trumpovo možné vítězství — a to počínaje primárkami — skrytý rasismus, který odkryla moje studie. Zkoumaná vyhledávání na Googlu odhalila u významné části amerického obyvatelstva velice temnou a nenávistnou stránku, která odborníkům celá léta unikala. Data jasně ukázala, že žijeme v úplně jiné společnosti, než jakou si představovali akademici a novináři na základě výsledků průzkumů veřejného mínění. Odhalila odpornou, děsivou a široce rozšířenou zlobu, která jen čekala na svého kandidáta, který by byl ochoten ji za všechny vyjádřit.

Lidé lžou často. Lžou sami sobě i ostatním. Při jednom z průzkumů v roce 2008 prohlásili, že už je rasové otázky nezajímají. O osm let později si za prezidenta zvolili Donalda J. Trumpa. Muže, který na Twitteru zveřejnil převzaté falešné obvinění, že za většinu vražd bílých Američanů jsou zodpovědní černoši; muže, který obhajoval svoje přívržence, když na jednom z jeho veřejných vystoupení hrubě napadli protestující členy hnutí Black Lives Matters (aktivistické hnutí upozorňující na násilí páchané na lidech černé pleti ze strany státních složek i samozvaných strážců zákona — pozn. překl.); a v neposlední řadě člověka, který váhal, zda odmítnout podporu od bývalého lídra Ku-Klux-Klanu. Stejný skrytý rasismus, který se obrátil proti Baracku Obamovi, pomohl Trumpovi k jeho zvolení.

Ještě na počátku primárních voleb vystoupil Nate Silver se svým slavným prohlášením, že Trump nemá nejmenší šanci. Jak primárky pokračovaly a na povrch začala vyplouvat Trumpova široká podpora, rozhodl se Silver znovu prozkoumat se sbíraná data a pokusit se pochopit, co se to děje. Jak je možné, že si Trump vede tak dobře?

Všiml si, že oblasti, ve kterých si Trump stojí nejlíp, tvoří zvláštní mapu. Trump měl největší podporu v různých částech průmyslového Středozápadu, na Severovýchodě a také na Jihu. Na první pohled hůř na tom byl na Západě USA. Silver se pustil do hledání možných proměnných, které by tuto mapu vysvětlily. Byla to nezaměstnanost? Příslušnost k církvi? Vlastnictví zbraní? Počty imigrantů? Byla to Obamova opozice?

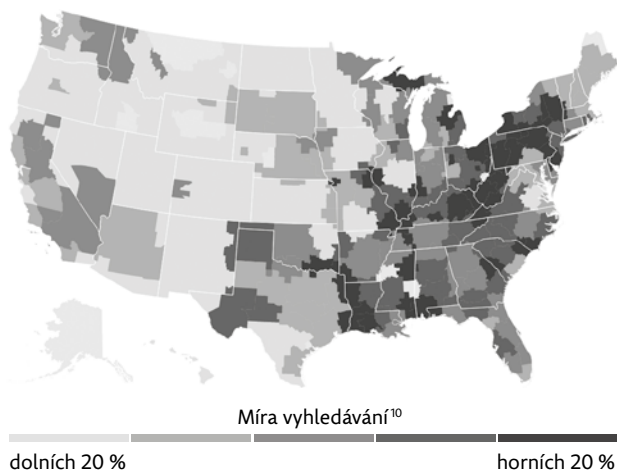
Silver zjistil, že jediným faktorem, který nejlépe koreluje s podporou Donalda Trumpa v republikánských primárkách, je měření, které jsem provedl o čtyři roky dříve.⁹

V oblastech, ve kterých byla podpora Trumpa nejsilnější, bylo zároveň nejčastěji vyhledáváno heslo „negr“.

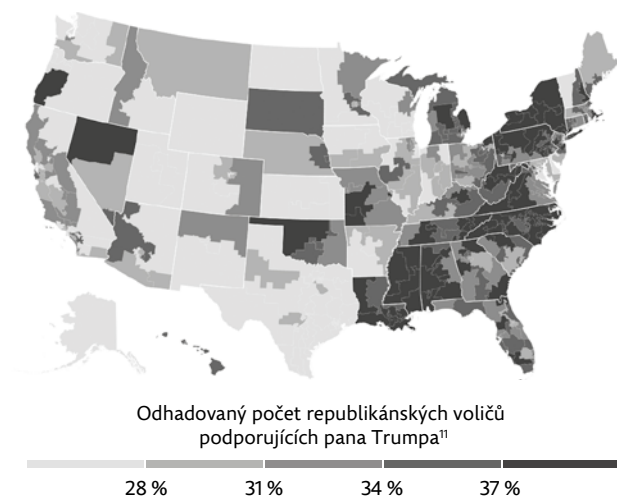
V průběhu posledních čtyř let jsem se analýzou dat z Googlu zabýval téměř každý den. A to včetně mé stáže ve společnosti Google, kam mě přijali poté, co se dozvěděli o mém výzkumu na téma rasismus. Stejnými daty se nadále zaobírám i v rámci datové novinářiny, které se věnuji coby publicista v deníku *The New York Times*. A odhaluji přitom stále něco nového v dalších oblastech. Duševní choroby, lidská sexualita, zneužívání dětí, potraty, reklama, víra, zdraví. Nejsou to právě lehká témata a soubor dat z Googlu, který ještě před několika desítkami let neexistoval, je všechny ukazuje z nové a překvapivé perspektivy. Ekonomové i sociologové neustále hledají nové zdroje dat, takže na rovinu: jsem přesvědčen, že vyhledávání na Googlu je tím nejzásadnějším zdrojem informací o lidské duši, jaký byl kdy k dispozici.

Avšak zmíněný soubor dat není tím jediným nástrojem, který vzešel z internetu a díky němuž můžeme lépe pochopit svět kolem nás. Brzy mi došlo, že na internetu najdu i další digitální zlaté doly. Stáhl jsem si veškeré Wikipedie, pečlivě prošel profily na Facebooku i nacionalistické internetové fórum Stormfront.

MÍRA VYHLEDÁVÁNÍ VÝRAZŮ S RASISTICKOU TEMATIKOU



PODPORA DONALDA TRUMPA V REPUBLIKÁNSKÝCH PRIMÁRKÁCH



Pornhub, jedny z největších pornografických stránek na internetu, mi navíc poskytly kompletní informace o vyhledáváních a zhlédnutých videích anonymních uživatelů z celého světa. Jinými slovy, skočil jsem po hlavě do toho, čemu se dnes říká velká data. Navíc jsem vyzpovídal desítky akademiků, datových novinářů a podnikatelů, kteří tato zcela nová území také prozkoumávají. O řadě jejich studií se zmíním v této knize.

Ale ze všeho nejdřív se musím k něčemu přiznat. Nebudu se pokoušet o přesnou definici termínu velká data. Proč? Protože jde o velice neurčitý pojem. Jak velká jsou velká data? Spadá 18 462 pozorování ještě do kategorie malých dat, ale 18 463 už do velkých? Já dávám přednost vše zahrnujícímu pohledu na to, jaká data označit za velká. Sice většina dat, se kterými si hraji, pochází z internetu, ale беру v potaz i jiné zdroje. Žijeme v době, která nás zahlcuje množstvím dostupných informací různé kvality. Velké množství těchto dat plyne z Googlu a sociálních sítí. Některá z nich jsou výsledkem digitalizace zdrojů, které se doposud ukrývaly v zásuvkách a šanonech. Jiná pocházejí z průzkumů trhu. Určité studie zmíněné v této knize se velkými soubory dat vůbec nezaobíraly. Namísto toho aplikovaly nový a kreativní přístup k analýze dat, což je v době přesycenosti informacemi zcela zásadní.

Jak to, že mají velká data takovou moc? Představte si, kolik informací je denně všude na internetu roztroušeno. Umíme to dokonce vyjádřit i číselně. V průběhu jednoho průměrného dne na počátku jednadvacátého století umějí lidé vygenerovat dva a půl milionu bilionů bytů informací.¹²

A tyto byty nám slouží jako náповěda.

Je čtvrtek odpoledne a žena se nudí. Na Googlu vyhledává nějaké „dobré korektní vtipy“. Zkontroluje e-maily. Přihlásí se na Twitter. Na Googlu hledá „vtipy o negrech“.

Jednomu muži je smutno. Na Googlu hledá „příznaky deprese“ a „příběhy depresivních lidí“. Hraje solitaire.

Žena na Facebooku vidí oznámení o zasnoubení své kamarádky. Žena je svobodná. Svoji kamarádku na Facebooku zablokuje.

Muž si dá pauzu od vyhledávání informací o lize amerického fotbalu a rapu a položí vyhledávací otázku: „Jsou sny o tom, že líbám muže, normální?“

Žena na webových stránkách BuzzFeed klikne na odkaz na článek „Patnáct nejroztomilejších koček“.

Muž vidí stejný článek. Na jeho stránkách se však jmenuje „Patnáct nejrozkošnějších kočiček“. Na odkaz neklikne.

Žena hledá na Googlu: „Je můj syn génius?“

Muž hledá: „Jak přimět dceru, aby zhubla.“

Žena je na dovolené se svými šesti nejlepšími kamarádkami. Všechny šest neustále opakuje, jak se skvěle baví. Žena si na Googlu nenápadně postěžuje: „Bez manžela je mi smutno.“

Muž, manžel té ženy, je na dovolené se svými šesti nejlepšími kamarády. Na Googlu nenápadně hledá: „Jak poznám, že mě žena podvádí?“

V těchto datech se skrývají informace, které by jinak nikdo nikdy a nikomu nepřiznal. Pokud je všechny sesbíráme, uchováme v anonymitě tak, aby nebylo možné v nich rozeznat obavy, touhy či jednání konkrétních jedinců, a zpracujeme vědeckým způsobem, naskytne se nám zcela nový pohled na lidské bytosti — jejich chování, touhy i povahy. Ačkoli to může znít poněkud okázale, došel jsem k závěru, že nová data, kterých máme v naší digitální době čím dál víc, rozšíří radikálním způsobem naše chápání lidského druhu. Díky mikroskopu jsme také zjistili, že v jediné kapce vody je toho mnohem víc, než jsme si mysleli při pohledu pouhým okem. Teleskop nám ukázal, že noční obloha toho skrývá podstatně víc, než nás kdy napadlo

při jejím nočním pozorování. Totéž platí o nových digitálních datech. O lidské společnosti nám toho mohou říct mnohem víc, než si myslíme, že o ní už dávno víme. Jsou mikroskopem či teleskopem naší doby, protože nám otevírají cestu k důležitým, až revolučním poznatkům.

Podobná prohlášení s sebou ovšem nesou ještě jedno riziko. Nejen že zní příliš pompézně, ale také jsou módní. Spousta lidí se nějakým způsobem vyjadřuje k moci velkých dat. Jenže jim chybí důkazy.

A to je inspirací pro všechny, kteří jsou skeptičtí vůči termínu „velká data“. Řada z nich se odmítá honit za velkými soubory dat. „Tím nechci říct, že v tom, čemu říkáme Big Data, nejsou žádné informace,“ napsal esejista a statistik Nassim Taleb. „Je v nich spousta informací. Hlavní problém tkví v tom, že hledáme jehlu v čím dál větší kupce sena.“¹³

Jedním z hlavních cílů této knihy je tedy poskytnout důkazy o tom, co vše lze z velkých dat vyčíst, jak najít ty jehly ve stále větších a větších kupkách sena. Doufám, že uvádím dostatek příkladů, jakým způsobem velká data přinášejí nové poznatky o lidské psychice a chování tak, abyste i vy zahlédli obrysy něčeho skutečně revolučního.

No, možná si právě říkáte: „Přibrzdi, Sethe. Slibuješ nám tu revoluci. Básníš o velkých souborech dat. Ale zatím jsi všechny ty úžasné, pozoruhodné, skvělé a převratné informace využil pouze k tomu, abys nám sdělil dvě věci: v Americe žije spousta rasistů a lidé, hlavně muži, přehánějí, když tvrdí, jak často si užívají sexu.“

Uznávám, že i nová data často jen potvrdí to, co už je zřejmé. Pokud se vám tato zjištění zdála samozřejmá, počkejte, až se dostanete ke čtvrté kapitole. V ní přináším jasné a nevyvratitelné důkazy z vyhledávání na Googlu, že muži vyjadřují až neuvěřitelné obavy a nejistotu ohledně... velikosti svého penisu.

Myslím si, že má cenu přinášet důkazy o věcech, o kterých jste dosud jaksi předpokládali, že existují, ale chyběly vám důkazy. Něco tušit je jedna věc. Dokázat to je věc druhá. Pokud by však velká data měla jen potvrzovat naše tušení, nebyla by revoluční. Umějí toho naštěstí mnohem víc. Znovu a znovu díky nim zjišťuji, že to na světě chodí úplně naopak, než bych si kdy pomyslel. Zde je několik příkladů, které vás patrně překvapí.

Možná máte za to, že hlavním důvodem rasismu je ekonomická nejistota a zranitelnost. Možná se tím pádem přirozeně domníváte, že jakmile lidé začnou přicházet o práci, rasismus vzroste. Ve skutečnosti se však počty vyhledávání s rasistickou tematikou či členství v hnutí Stormfront s rostoucí nezaměstnaností nijak nezvyšují.¹⁴

Možná předpokládáte, že nejvíc lidí trpících úzkostí žije ve velkých městech s vysokou úrovní vzdělání. Velkoměstský neurotik je oblíbený stereotyp. Ovšem vyhledávání na Googlu týkající se úzkostí (například výrazů jako „příznaky úzkosti“ či „pomoc při úzkostech“) bývají vyšší v oblastech s nižší úrovní vzdělání, pod úrovní příjmového mediánu a v místech, kde větší část populace žije na venkově. Počty těchto vyhledávání jsou na venkově na severu státu New York vyšší než v New York City.

Možná si myslíte, že po teroristickém útoku, při němž přišly o život desítky či stovky lidí, bude automaticky následovat obrovská vlna strachu. A terorismus má ze své podstaty šířit děs a hrůzu. Prostudoval jsem hesla vyhledávaná na Googlu týkající se strachu.¹⁵ Testoval jsem, o kolik vzrostl jejich počet ve dnech, týdnech či měsících následujících po jakémkoli větším teroristickém útoku v Evropě či USA od roku 2004. Takže o kolik se v průměru navýšil počet takových vyhledávání? Nenavýšil se vůbec.

Možná byste řekli, že lidé hledají vtipy na internetu spíš tehdy, když jsou smutní.¹⁶ Řada velkých myslitelů už v minulosti prohlašovala, že humor je prostředek, jak se zbavit bolesti. Na

humor se také nahlíželo jako na způsob, jak se vyrovnat s frustrací, zármutkem či nevyhnutelným životním zklamáním. Slovy Charlieho Chaplina — „smích nás povzbuzuje, ulevuje od bolesti“.

Nicméně lidé na internetu vyhledávají vtipy nejméně v pondělí, které současně považují za nejméně šťastný den. Nejnižší počty vyhledávání zaznamenáváme i za pošmourných, deštivých dnů. A po událostech, jakou byl například teroristický útok během Bostonského maratonu v roce 2013, při kterém bomba zabila tři lidi a stovky jich zranila, tyto počty klesají ještě prudčeji. Lidé s větší pravděpodobností vyhledávají humor na internetu v dobách, kdy se jim daří dobře, ne naopak.

Stává se, že nový soubor dat odhalí chování, touhy či obavy, které by mě ani ve snu nenapadly. Do této kategorie spadají různé sexuální sklony a aktivity. Věděli jste například, že v Indii první místo v počtu internetových vyhledávání, která začínají slovy „můj muž chce...“, zaujímá fráze „můj muž chce, abych ho kojila“?¹⁷ Tento dotaz je mnohem běžnější v Indii než v jiných zemích. I pornografické výjevy, znázorňující ženy kojící muže, jsou v Indii a Bangladéši vyhledávány čtyřikrát častěji než v jakékoli jiné zemi na světě.¹⁸ Než se mi dostaly tyto údaje do rukou, něco takového by mě skutečně ani nenapadlo.

A pokračujeme. Zatímco skutečnost, že muži jsou posedlí velikostí svého penisu, nemusí být až takovým překvapením, největší nejistota žen, co se jejich těla týče, a způsob, jakým ji vyjadřují na Googlu, je opravdu překvapující. Na základě těchto nových dat je ženský ekvivalent mužských obav o velikost penisu — chvíle napětí — obava, zda není jejich vagina příliš citlivá. Vyhledávání žen vyjadřující obavy o vlastní genitálie jsou přinejmenším tak častá jako u mužů.¹⁹ A u žen jim vévodí obava o jejich odér a to, jak jej mohou vylepšit. To jsem dřív skutečně netušil.

Nová data někdy odhalí kulturní rozdíly, o kterých jsem dřív neměl ani zdání. Příklad za všechny: velice rozdílné způsoby, jakými muži na celém světě reagují na těhotenství své ženy. V Mexiku vyhledáváním s tematikou „těhotná manželka“ vévodí slovní spojení „frases de amor para mi esposa embarazada“ (láskyplná slova pro mou těhotnou ženu) a „poemas para mi esposa embarazada“ (básně pro mou těhotnou ženu).²⁰ Ve Spojených státech jsou nejčastěji vyhledávány věty jako „žena je těhotná, co teď“ či „manželka je těhotná, co mám dělat“.

Tato kniha ovšem přináší víc než jen sbírku zajímavostí nebo výsledků nárazových studií, i když i těch tu bude dost. Vzhledem k tomu, že tato metodologie práce s daty je natolik nová a v budoucnu bude nabírat na síle, představím několik nápadů, jak funguje a proč je tak průkopnická. Ukážeme si i slabá místa nakládání s velkými daty.

Někdy nebylo nadšení z nekonečných možností práce s velkými daty úplně na místě. Většina lidí, které datová revoluce okouzila, básní o tom, jak obrovské soubory lze využít. Tato posedlost velikostí souborů dat není nic nového. Dávno před Googlem, Amazonem a Facebookem, i dávno předtím, než se objevil termín „big data“, se v texaském Dallasu konala konference s názvem „Velké a komplexní soubory dat“. Jerry Friedman, profesor statistiky na Stanfordově univerzitě a můj kolega z dob, kdy jsem pracoval v Googlu, si na tuto konferenci z roku 1977 dobře pamatuje. Jeden proslulý statistik se postavil, aby promluvil. Vysvětloval, jak nasbíral neuvěřitelných, přímo závratných pět gigabytů dat. Další proslulý statistik se zdvihl ze sedadla, aby převzal slovo, a začal: „Kolega disponuje gigabyty. To nic není. Já mám terabyty.“ Jinými slovy, v každém proslovu byl kladen důraz hlavně na to, kolik informací dokážeme nahromáždit, a ne na to, co s těmi daty dál nebo na jaké otázky v nich budeme hledat odpověď. Friedman k tomu říká: „Tehdy

mi přišlo docela vtipné, že se všichni snažili udělat dojem hlavně velikostí svého souboru dat. A děje se to pořád.“²¹

Mnoho vědců sbírá ohromná množství dat, která nevypovídají o ničem zásadním. Dozvíme se z nich například, že v New Yorku je populární basketbalový tým Knicks. Odborníci z příliš mnoha oblastí se přímo topí v datech. Disponují spoustou terabytů dat, z nichž však umějí vytěžit jen velice málo důležitých informací. Myslím si, že se velikost souboru dat často přeceňuje. Dá se to snadno vysvětlit. Čím silnější jev, tím menší počet pozorování je potřeba k tomu, abychom si ho všimli. I horké plotýnky se stačí dotknout jen jednou, aby vám došlo, že je to nebezpečné. Kávu si můžete dát i tisíckrát, než budete moci říct, zda vás z ní bolívá hlava, nebo ne. Které z uvedených poučení je důležitější? Jednoznačně to s horkou plotýnkou, které se vlivem své intenzity projevilo tak rychle a s malým počtem pozorování.

Pravda je dokonce taková, že ty nejchytřejší společnosti využívající velká data snižují jejich objem. V Googlu se zásadní rozhodnutí dělají na základě malého vzorku všech jejich dat.²² K tomu, abyste objevili zásadní informaci, nepotřebujete vždycky tuny dat. Ale musí to být ta správná. Hlavní důvod, proč jsou vyhledávání na Googlu tak cenná, netkví v tom, že jich je tolik; jde o to, že jsou v nich lidé upřímní. Lidé lžou svým přátelům, milencům, lékařům, lžou v dotaznících, dokonce i sami sobě. Ale na Googlu sdílejí věci, za které by se jinak styděli, jako například to, že v jejich manželství chybí sex, mají psychické potíže, píšou o svojí nejistotě či nenávisti k černochem.

A co je nejdůležitější, chcete-li proniknout do velkých dat a pochopit, co nabízejí, musíte si klást ty správné otázky. Stejně jako nemůžete jen tak náhodně namířit teleskop někam na oblohu a očekávat, že pro vás objeví Pluto, nelze stáhnout hromadu dat a čekat, že vám sama vyjeví tajemství lidské duše. Musíte

hledat na slibných místech, třeba mezi indickými vyhledáváními na Googlu, která začínají slovy „můj manžel chce“.

V této knize si ukážeme, jak nejlépe využít velká data, a podrobně si vysvětlíme, co všechno nám to může přinést. A zároveň se dozvíte, co jsme přitom já i ostatní objevili. Mimo jiné:

- Kolik je mezi muži homosexuálů?
- Funguje reklama?
- Proč byl American Pharoah tak skvělý dostihový kůň?
- Přinášejí média zkreslené informace?
- Skutečně existují freudovská přerážení?
- Kdo podvádí při placení daní?
- Není jedno, na jakou vysokou školu jdete?
- Lze zvítězit nad akciovými trhy?
- Které místo je nejlepší k výchově dětí?
- Jak se z příběhu stane „virál“?
- O čem byste měli mluvit na prvním rande, jestliže chcete, aby nebylo poslední?

...a mnohem, mnohem víc.

Ale než se k tomu dostaneme, musíme probrat jednu základní otázku: K čemu nám vůbec data jsou? Při této příležitosti bych vám rád představil svoji babičku.

PRVNÍ ČÁST

VELKÁ A MALÁ DATA

VAŠE CHYBNÁ INTUICE

Když je vám třiatřicet a na posledních několik oslav Děkuvzdání jste dorazili sami, je velice pravděpodobné, že se začne probírat otázka výběru ideálního partnera. A všichni k tomu mají co říct.

„Seth potřebuje holku, co bude stejně trhlá jako on,“ domnívá se moje sestra.

„Neblázni! Potřebuje normální holku, která to bude vyvažovat,“ oponuje můj bratr.

„Seth není potrhlý,“ brání mě moje matka.

„No to mě podrž! Jistě že je Seth potrhlý,“ přidá se táta.

Zničehonic se tichým hlasem ozve moje nesmělá babička, která byla jinak v průběhu večere potichu. Hlučné agresivní newyorské hlasy umlknou a všechny zraky se upřou na tu malou starou paní s krátkými zlatými vlasy a stále patrnou stopou východoevropského přízvuku. „Sethe, ty potřebuješ milou holku. Ne moc hezkou. Hlavně chytrou. Co to umí s lidmi. Společenskou, ať spolu můžete něco podnikat. Co má smysl pro humor, protože ten ty máš taky.“

Jak je možné, že si rada této staré ženy okamžitě vynutila pozornost a respekt celé rodiny? Moje osmaosmdesátiletá babička toho v životě zažila víc než kdokoli jiný, kdo tehdy seděl u našeho stolu. Viděla spoustu manželství, funkčních i nefunkčních. A za ta léta vypožadovala řadu vlastností, které přispívají k úspěšnému vztahu. Tehdy u té slavnostní tabule měla moje babička přístup k největšímu souboru dat, a mohla tak nejlíp zodpovědět danou otázku. Moje babička byla Big Data.

V této knize bych rád demystifikoval analýzu dat. Ať se vám to líbí, nebo ne, data hrají v našem životě čím dál větší roli a jsou stále důležitější. Zpravodajské redakce mají v dnešní době celá oddělení, která se věnují práci s daty. Firmy zaměstnávají týmy, jejichž jediným úkolem je analyzovat data. Investoři sypou miliony dolarů do začínajících podniků, jen když mohou skladovat větší množství dat. I když se nikdy nenaučíte, jak provést regresní analýzu nebo jak vypočítat interval spolehlivosti, velkému množství dat se stejně nevyhnete — na stránkách této knihy, na pracovní schůzce, nebo když postáváte ve firemní kuchyňce a zaslechnete nějaký drb.

Mnoho lidí vývoj tímto směrem děsí. Z dat mají obavy a ve světě čísel se ztrácejí. Myslí si, že kvantitativní pohled na náš svět je pro pár vyvolených, logicky myslících génů, ale rozhodně nic pro ně. Jakmile narazí na nějaké cifry, ihned jsou připraveni obrátit list, ukončit pracovní schůzku nebo změnit téma hovoru.

Já se však v oblasti analýzy dat pohybuji už deset let a měl jsem to štěstí, že jsem se mohl setkat s řadou špiček v tomto oboru. A jedna z nejdůležitějších věcí, které jsem se naučil, je následující: správné vědecké zpracování dat je mnohem snazší, než si lidé myslí. Ty nejlepší metodologie jsou dokonce překvapivě intuitivní.¹

Co dělá z vědeckého zpracování dat intuitivní záležitost? Základem je objevit nějaký opakující se vzorec a dokázat předpověď, jak jedna proměnná ovlivní druhou. Takové věci lidé dělají neustále.

Jen si vzpomeňte, jak mi babička radila ohledně mých vztahů. Dokázala zpracovat a využít obrovskou databázi vztahů, kterou si v průběhu svého téměř století trvajícího života nahrála její mozek. Byly to příběhy, které slyšela od své rodiny, přátel a známých. Svoji analýzu omezila na vzorek vztahů, ve kterých figuroval muž se stejnými vlastnostmi, jaké mám já: citlivá povaha, sklon k izolaci, smysl pro humor. Dále se zaměřila na klíčové povahové rysy jeho partnerky: jestli je hodná, jak je chytrá a pohledná. Zjistila, zda a jak tyto klíčové vlastnosti partnerky korelují s klíčovými atributy vztahu, tedy jde-li o fungující vztah. A nakonec podala svoji zprávu o výsledcích analýzy. Jinak řečeno, objevila ve svých datech opakující se vzorec a předpověděla, jak jedna proměnná ovlivní druhou. Moje babička je zkrátka odborník na zpracování dat.

I vy jste takovými odborníky. Když jste byli malí, všimli jste si, že kdykoli zapláčete, přitáhnou tím matčinu pozornost. To je zpracování dat. Jakmile jste dospěli, všimli jste si, že když si příliš často stěžujete, lidé se vám začnou vyhýbat. I to je zpracování dat. Když se vám začnou všichni vyhýbat, jste z toho smutní. Když jste smutní, nejste příliš přátelští. Když jste míň přátelští, lidé se vám vyhýbají ještě víc. Data. Data. Data.

Vzhledem k tomu, že práce s daty je tak přirozená věc, došlo mi, že ty nejlepší studie zpracovávající velká data může pochopit každý inteligentní člověk. Pokud nějaké studii nerozumíte, problém je pravděpodobně v ní, ne ve vás.

Chcete důkaz, že kvalitní analýza dat může být velice intuitivní? Nedávno jsem narazil na studii, která patří možná mezi

ty nejdůležitější, jaké byly za posledních pár let provedeny. Zároveň je to jedna z nejintuitivnějších studií, které jsem kdy viděl. Rád bych, abyste nemysleli jen na to, o jak důležitou studii se jedná, ale na to, jak je přirozená, až „babičkovská“.

Studii provedl tým vědců z Kolumbijské univerzity a ze společnosti Microsoft. Jejich cílem bylo zjistit, jaké symptomy mohou provázet rakovinu slinivky.² Tato choroba má velice nízkou míru přežití, jen pouhá tři procenta pacientů přežijí pět let. Včasně odhalení choroby šanci pacientů zdvojnásobuje.

Jaká byla výzkumná metoda? Vědci nashromáždili údaje od desítek tisíc anonymních uživatelů Bingu, internetového vyhledávací firmy Microsoft. Uživatelé, kterým byla v nedávné době stanovena diagnóza rakoviny slinivky, určovali podle jasných vyhledávacích výrazů jako „právě mi diagnostikovali rakovinu slinivky“ nebo „mám rakovinu slinivky, co mohu očekávat“.

Tým vědců se dále zaměřil na vyhledávání zdravotních příznaků. Porovnali malý počet uživatelů, kteří o něco později přišli s tím, že jim byla diagnostikována rakovina slinivky, s těmi, kteří byli zdraví. Jinými slovy, jaké příznaky předpověděly, že o několik týdnů později uživatel internetu uvede danou diagnózu?

Výsledky byly pozoruhodné. Vyhledávání informací o bolestech zad současně se žlutým nádechem kůže prokázalo, že jde o příznaky rakoviny slinivky. Hledání informací o pouhé bolesti zad se ukázala jako nepravděpodobná známka onemocnění. Obdobně tomu bylo u vyhledávání informací o zažívacích potížích v kombinaci s bolestmi břicha, což byla známka onemocnění rakovinou slinivky, zatímco pouhé potíže s trávením znamenaly, že člověk touto chorobou pravděpodobně netrpí. Vědci dokázali identifikovat pět až patnáct procent případů s téměř nulovým počtem falešně pozitivních výsledků. Ano, možná vám to nezní jako příliš vysoké číslo. Jenže když

onemocníte rakovinou slinivky, i desetiprocentní možnost, že se vaše šance na přežití zdvojnásobí, je jako dar z nebes.

Laikovi by možná kompletní práce s veškerými detaily studie úplně nedávala smysl. Je v ní spousta odborných pojmů, jako například Kolmogorovův—Smirnovův test, jejichž smysl, musím přiznat, jsem už zapomněl. (Tímto testem se zjišťuje, zda teoretický model odpovídá nasbíraným datům.)

Nicméně si všimněte, jak je tato studie ve své podstatě přirozená a intuitivní. Výzkumný tým prostudoval široké spektrum lékařských případů a pokusil se propojit příznaky s konkrétním onemocněním. Víte, kdo ještě používá stejnou metodologii, když chce zjistit, zda je někdo nemocný, nebo ne? Manželé a manželky, mámy a tátové, zdravotní sestry a lékaři. Na základě svých zkušeností a znalostí se snaží najít spojitost mezi horečkou, bolestí hlavy, rýmou či bolestmi břicha a různými chorobami. Abych to shrnul, tým vědců z Kolumbijské univerzity a Microsoftu přišel s průlomovou studií, v rámci níž využil přirozenou a jasně uchopitelnou metodologii, jakou k určení diagnózy používá každý.

Tady bych se ovšem na chvíli zastavil. Pokud je metodologie vědeckého zkoumání dat často tak intuitivní a přirozená, jak tvrdím, vyvstává zásadní otázka týkající se významu velkých dat. Jsme-li všichni od přírody tak trochu odborníky na analýzu dat a je-li práce s daty tak intuitivní, na co nám jsou počítače a statistický software? K čemu potřebujeme Kolmogorovův—Smirnovův test? Nemůžeme se prostě jen držet vlastní intuice? Nemůžeme na to jít stejně jako moje babička, jako každá zdravotní sestra či lékař?

To nás přivádí k otázce, která se po vydání bestselleru Malcolma Gladwella *Mžik: Jak myslet bez přemýšlení* stala ještě palčivější. Kniha vychvaluje kouzlo rozhodování na základě instinktu. Gladwell přináší příběhy lidí, kteří pouze na základě

svých vnitřních pocitů dokážou určit, zda je nějaké umělecké dílo padělek, říct, zda tenista udělá chybu ještě předtím, než odpálí první míč, nebo kolik je zákazník ochoten zaplatit. Hrdinové knihy neprovádějí regresní analýzy; nevypočítávají intervaly spolehlivosti ani se nezatěžují Kolmogorovovým—Smirnovovým testem. A přesto umějí předpovídat s pozoruhodnou přesností. Spousta lidí intuitivně podporovala Gladwellovu obhajobu intuice, jelikož sami důvěřují svému instinktu i svým pocitům. Fanoušci knihy *Mžik* jsou možná nadšení moudrostí mé babičky, která umí poradit ohledně vztahů, aniž by si musela vypomáhat počítačem. Stejní fanoušci jsou možná o něco méně nadšení studiemi, které uvádím v této knize, protože ty se o práci s počítačem opírají. Máme-li velká data považovat za revoluci (spíš ta počítačově zpracovaná než babiččin typ dat), musejí dokázat, že mají větší sílu než naše intuice. A ta, jak pokazal Gladwell, může také vést k pozoruhodným výsledkům.

Studie Kolumbijské univerzity a Microsoftu nabízí jasnou ukázkou precizní vědecké práce s daty zpracovanými počítačovou analýzou, ze které se dozvíme něco, co naše intuice sama o sobě nemůže nikdy zachytit. Zároveň slouží jako příklad, ve kterém hraje důležitou roli objem dat. Intuice se někdy nemá čeho chytit, protože nemáme dostatek potřebných zkušeností. Je dost nepravděpodobné, že jste se vy nebo vaši přátelé či ostatní členové rodiny v životě setkali s tolika případy rakoviny slinivky, abyste dokázali rozeznat rozdíl mezi poruchou trávení s následnou bolestí břicha coby příznakem diagnózy a obyčejným pokaženým zažíváním. Je jasné, že čím větší je soubor dat z Bingu, tím víc nenápadných vzorců v načasování jednotlivých příznaků vědci zachytí, a nezáleží na tom, o jakou nemoc se jedná. Tyto vzorce mohou být tak nepatrné, že je přehlédnou i lékaři.

Intuice nám obvykle umí dobře posloužit, když jde o běžné věci, se kterými se můžeme v životě potkat, ale často nebyvá

zcela přesná. K zaostření obrazu potřebujeme data. Vezměte si například vliv počasí na svou náladu. Jistě byste si tipli, že když je venku deset stupňů pod nulou, vládne pochmurnější nálada, než když je dvacet stupňů nad nulou. A měli byste pravdu. Co byste však možná neodhadli, je to, jak velký vliv může uvedený rozdíl v teplotách mít. Hledal jsem korelace mezi vyhledáváními na Googlu na téma deprese v různých oblastech USA a mnoha dalšími různými faktory, jako je ekonomická situace, úroveň vzdělání či chození do kostela. Chladné podnebí všechny ostatní faktory „převálcovalo“. ³ V oblastech s teplým klimatem, jaké vládne například v havajském Honolulu, se během zimních měsíců objevuje o čtyřicet procent méně vyhledávání na téma deprese než v chladných oblastech, jako je například Chicago. Do jaké míry na to má vliv počasí? Optimistický odhad účinnosti antidepressiv naznačuje, že i ty nejlepší preparáty dokážou snížit výskyt depresí jen o asi dvacet procent. Podle čísel z Googlu by jako léčba zimních depresí bylo dvakrát účinnější přestěhovat se z Chicaga do Honolulu než brát léky.*

Někdy se může naše intuice šeredně mýlit, pokud ji neusměrníme pečlivou počítačovou analýzou. V takovém případě nás mohou zaslepit vlastní zkušenosti a předsudky. Skutečně. Přestože moje babička dokáže využít své zkušenosti se vztahy nasbírané za desítky let a poradit mi líp než zbytek rodiny, některé její názory na to, díky čemu vztah dlouho vydrží, jsou přinejmenším pochybné. Několikrát mi například zdůraznila, jak je důležité mít s partnerkou společné přátele. Babička věří, že v jejím manželství to byl jeden z klíčových faktorů úspěchu: většinu vlahých večerů strávila se svým mužem, mým dědečkem, v zahradních křeslech na trávníku za domem a klábosili s přáteli ze sousedství.

* Úplné odhalení: krátce po dokončení této studie jsem se přestěhoval z Kalifornie do New Yorku. Zjistit pomocí dat, co máte správně udělat, je docela snadné. Skutečně to provést je mnohem těžší.

Avšak zpracovaná data naznačují, že její teorie je chybná (i když mě mrzí, že tímto házím svoji babičku přes palubu). Tým počítačových vědců nedávno provedl analýzu největšího souboru dat o lidských vztazích, jaký byl kdy nashromážděn: Facebook.⁴ Sledovali velké množství dvojic, jejichž status byl po nějakou dobu „ve vztahu“. Některým tento status vydržel, jiní ho změnili na „nezadaný/á“. Vědci zjistili, že společná základní skupina přátel je silným ukazatelem toho, že vztah nevydrží. Možná že trávit každý večer se svým partnerem a stejnou skupinou přátel není zase tak dobrý nápad; různé okruhy známých by mohly vztah posílit.

Jak jistě víte, naše vlastní intuice nás může docela překvapit, když vynecháme počítače a necháme se jí vést. Stejně tak se může hodně mylit. Babička se nechala chytit do jedné z kognitivních pastí: všichni máme sklon přehánět důležitost vlastních zkušeností. Řeceno vědeckým jazykem, vážíme hodnotu svých dat, ale jednomu souboru dat přisuzujeme větší váhu než ostatním: sobě samým.

Babička se natolik zaměřila na večerní povídání s dědečkem a jejich přáteli, že jí příklady jiných párů ani nepřišly na mysl. Nevzala v potaz svého švagra a jeho ženu, kteří také většinu večerů strávili povídáním v malém a sevřeném kruhu přátel, také se často hádali, a nakonec se rozvedli. Nevzala plně v potaz ani mé rodiče, tedy svou dceru a zetě. Máma s tátou trávili večery každý po svém. Otec chodíval se svými přáteli do jazzového klubu nebo na baseball, máma do restaurací a divadla se svými kamarádkami; a přesto jsou stále šťastně sezdání.

Když se spoléháme na svoji intuici a vnitřní pocit, může se stát, že náš pohled na věc překrouť obyčejná lidská fascinace vším dramatickým. Máme sklon přeceňovat všechno, co je vzrušující, a tudíž zapamatovatelné. Například v dotazníkových šetřeních lidé neustále označují tornáda za častější příčinu smrti

než astma.⁵ Ve skutečnosti je astma sedmdesátkrát častější příčinou úmrtí než tornáda.⁶ Jenže úmrtí na astma nijak nevyčnívá — a proto nebývá obsahem senzačních novinových titulků. Zato tornáda ano.

Jinými slovy, jakmile se spoléháme jen na to, co jsme zaslechli nebo sami zažili, ve svých závěrech ohledně světa kolem nás se často pleteme. Metodologie vědeckého zpracování dat bývá intuitivní, ovšem výsledky právě naopak. Vědecké zpracování dat využívá přirozený proces nacházení vzorců a pátrání po jejich smyslu. Poté ho nakopne steroidy, díky čemuž se ukáže, že to na světě chodí úplně jinak, než jsme si mysleli. To se stalo, když jsem se zabýval prediktory úspěchu v basketbalu.

Když jsem byl malý kluk, měl jsem jen jediný sen: chtěl jsem vyrůst a stát se ekonomem a statistikem. Ne, žertuji. Zoufale jsem toužil stát se profesionálním hráčem basketbalu a následovat tak Patricka Ewinga, svého hrdinu, který byl pivotem all-star týmu a hrál za newyorské Knicks.⁷

Někdy mám pocit, že uvnitř každého vědce pracujícího s daty je malý kluk, který se snaží přijít na to, proč se mu nesplnily jeho dětské sny. Nikoho by tedy nemělo překvapit, že jsem se nedávno zabýval tím, co je potřeba udělat, abyste se dostali do NBA. Výsledky zkoumání byly překvapivé. Vlastně jen znovu potvrdily, jak může kvalitní zpracování dat změnit váš pohled na svět a jak zcela neintuitivní mohou čísla být.

Hledal jsem odpověď na jednu konkrétní otázku: Je pravděpodobnější, že se dostanete do NBA, když pocházíte z chudých poměrů, nebo ze střední vrstvy?

Většina lidí by asi hádala první možnost. Převládá názor, že když dítě vyrůstá ve složitých podmínkách, například v neúplné rodině nebo třeba s mladistvou matkou, pomáhá to v něm

vypěstovat chuť se prosadit, což je k dosažení vrcholu v tak silně konkurenčním prostředí zcela nezbytné.

Stejný názor vyjádřil v rozhovoru pro časopis *Sports Illustrated* i William Ellerbe, středoškolský basketbalový trenér z Filadelfie: „Děti z bohatších předměstí hrávají pro samotnou radost ze hry. Kdežto pro děti z chudších čtvrtí je basketbal otázkou života a smrti.“⁸ Já jsem bohužel vyrůstal na předměstí New Jersey a měl jsem oba rodiče. Zatímco nejlepší hráč mé generace LeBron James se narodil v ohyjském Akronu šestnáctileté matce a do chudých poměrů.

Dotazníkové šetření na internetu skutečně ukázalo, že si většina Američanů myslí totéž co já a trenér Ellerbe, tedy že převážná část hráčů NBA vyrůstala v chudobě.⁹

Je však toto všeobecné mínění správné?

Pojďme se podívat na data. Žádný souhrnný soubor dat týkající se socioekonomické situace hráčů NBA neexistuje. Když si však zahrajeme na detektivy a shromáždíme údaje z mnoha jiných zdrojů — například basketball-reference.com, ancestry.com, sčítání lidu a dalších —, můžeme se dopátrat toho, jaké rodinné zázemí nejčastěji vede do NBA. Jak si jistě sami povšimnete, tato studie využívá řady různých zdrojů dat, velké i malé soubory, z internetu i tradičních zdrojů. Ať už jsou některé nové digitální zdroje dat jakkoli vzrušující, správný vědec se neobráť zádý ke starým dobrým zdrojům, pokud mohou přinést něco zajímavého. Nejlepší způsob, jak se dobrat odpovědi na otázku, je kombinace všech dostupných poznatků.

První podstatnou informací je místo narození každého hráče. Nashromáždil jsem údaje o všech černých i bílých mužích, kteří se narodili v osmdesátých letech ve všech okresech USA. Dále jsem se podíval na to, kolik z nich se dostalo do NBA. Tyto údaje jsem porovnal s průměrným příjmem domácnosti v daném okrese. Hlídal jsem i demografické údaje o etnickém rozdělení,

protože černoši mají čtyřicetkrát větší šanci dostat se do NBA než běloši (to je ovšem téma, které by vydalo na celou knihu).

Data uvádějí, že podstatně lepší šanci dostat se do NBA má muž z některého bohatšího okresu. Například u mladého černocho, který se narodil v jednom z nejbohatších okresů USA, existuje dvakrát vyšší pravděpodobnost, že se dostane do NBA, než u černocho narozeného v chudé oblasti. U bílého chlapce představuje výhoda vyrůstání v bohatším okrese v porovnání s těmi nejchudšími okresy asi šedesát procent.

Tyto informace naznačují — navzdory všeobecnému mínění —, že v NBA je hráčů pocházejících z chudých poměrů spíše menšina. Nicméně uvedené údaje nejsou dokonalé, jelikož v řadě bohatých okresů v USA, jako je například New York County (Manhattan), najdeme i chudinské čtvrti, jako je Harlem. Takže zatím je stále možné, že složité dětství vám může pomoci ke kariéře v NBA. Potřebujeme další vodítka, více dat.

Bylo tedy potřeba zapátrat v rodinné historii jednotlivých hráčů. Potřebné informace jsem hledal v novinových článcích a na sociálních sítích. Tato metoda byla časově velice náročná, proto jsem analýzu omezil na sto afroamerických hráčů NBA, kteří se narodili v osmdesátých letech¹⁰ a získali nejvíc bodů. Ve srovnání s průměrným americkým černochem byla u hvězd NBA o třicet procent menší pravděpodobnost, že se narodili mladistvé nebo neprovdané matce. Jinými slovy, rodinné zázemí těch nejlepších hráčů NBA naznačuje, že pohodlí a dostatek jsou při snaze o dosažení úspěchu velkou výhodou.

Ale ani informace o narození omezeného počtu hráčů NBA z hlediska okresů a jejich rodinného zázemí nám nedávají dokonalý obrázek o tom, jaké bylo dětství všech hráčů. Tudíž jsem stále nebyl přesvědčen, že úplné rodiny ze střední vrstvy produkují více hráčů NBA než chudé rodiny s jedním rodičem. Čím více údajů k této otázce získáme, tím lépe.