Velká data jako alternativa výběrových šetření v kvantitativním sociálněvědním výzkumu[1]

Johana Chylíková, Sociologický ústav AV ČR, v.v.i.

Big Data as an Alternative to Surveys in Quantitative Social Research

EXTENDED ABSTRACT:

The goal of this article is to inform social scientists, especially those of a quantitative orientation, about the basic characteristics of Big Data and to present the opportunities and limitations of using such data in social research. The paper informs about three basic types of Big Data as they are distinguished in contemporary methodological literature, namely administrative data, transaction data and social network data, and exemplifies how they can be utilized by quantitative social research. According to many, questionnaire-based sample survey as the dominant method of quantitative social research has found itself in a crisis, especially as response rates have decreased in most developed countries and public confidence in opinion polling has declined.

The author presents the characteristics and specifics of Big Data compared to survey research – a method whose primary distinguishing characteristic is the capacity to quantify individual behaviour, social action and attitudes at the level of populations. In this context, the article draws attention to the differences between Big Data and survey data typically presented in scholarly literature, namely that datasets are not representative of known populations, the values of observed variables are systematically biased, there is a limited number of variables in Big Data sets, there is uncertainty about the meaning of observed values, and social environment has direct influence on

KEYWORDS:

big data
survey
data quality
epistemology
methodology

DOI LINK:

http://doi.org/10.13060/1214438X.2018.1.16.416

[1] Tento článek byl podpořen grantem „CSDA Český sociálněvědní datový archiv“ (MŠMT, kód: LM2015060) a také grantem „SERISS“ (EU Horizon 2020, kód 654221).

the behaviours captured by Big Data. Attention is also paid to such characteristics of Big Data that pose an obstacle to smooth integration of this type of data in the social scientific mainstream. First, the collection, processing and analysis of Big Data is extremely demanding in terms of programming skills, something social scientists typically do not have. Second, the availability of Big Data is limited as they are normally possessed by private corporations, some of which (Facebook, Google) have undoubtedly come to form data oligopolies – and their management is mostly unwilling to share their data with traditional academics.

Based on the above-mentioned specifics, differences and limitations, it is argued that Big Data currently do not have the potential of becoming a full-fledged source of social science data and replacing sample surveys as the dominant research method. Finally, the article draws attention to the specifics of different types of Big Data as they are primarily generated for purposes other than social research and result from specific situations framed by existing social relations – and it is from this perspective that Big Data should be viewed by social researchers.

V současném sociálněvědním akademickém výzkumu jsou primární kvantitativní metodou výběrová šetření, a to nejen v České republice, ale i v dalších zemích Evropy či v USA. Šetření generují velké množství dat pro sociálněvědní analýzy a široké využití v akademickém i komerčním výzkumu nacházejí zhruba od konce 2. světové války. V Česku tato metoda spíše vzkvétá, než aby upadala, o čemž svědčí zapojení českých akademiků a akademiček do většiny velkých evropských a světových komparativních projektů výběrových šetření (European Social Survey, International Social Survey Programme, Survey of Health, Aging and Retirement in Europe, EU SILC, European Values Study) či nedávný vznik velkého panelu domácností v rámci sociologického projektu Proměny české společnosti. Pokud mají být výběrová šetření užitečnou a spolehlivou výzkumnou metodou, musejí výzkumníci průběžně vyhodnocovat jejich účinnost – šetření by měla přinášet validní informace o chování lidí, jejich postojích, názorech a hodnotách a neměla by být jen záminkou pro akademickou hru s čísly odtrženou od empirické reality.

V posledních desetiletích se ovšem začínají objevovat problémy, v jejichž důsledku výběrová šetření ztrácejí prestiž [Gobo, Giamietro 2011]. Odborníci se shodují na tom, že největším problémem výběrových šetření je klesající návratnost, která se snižuje i navzdory úsilí investovanému do opakovaných oslovení a penězům na odměny pro respondenty [Savage, Burrows 2007; Couper 2013, Massey, Tourangeau 2013]. Populace je přezkoumaná a lidé se už šetření zúčastňovat nechtějí; jsou často zaneprázdněni, nejsou k dostižení a častými žádostmi o rozhovor jsou unaveni. Za snižující se návratnost nemůže jen akademický výzkum, ale především soukromé firmy a agentury, které ve velké míře zjišťují veřejné mínění či spotřebitelské chování, případně obchodníci, kteří prostřednictvím telefonických hovorů prodávají své zboží. V současnosti se v USA návratnost z akademických šetření, která využívají pravděpodobnostní výběr, pohybuje kolem 40 až 60 procent, přičemž od devadesátých let klesá výrazně rychleji než dříve [Massey,

Tourangeau 2013], v případě komerčních šetření je situace ještě výrazně horší; CATI šetření prestižní americké agentury Pew Research Center mají v posledních pěti letech návratnost pouhých 9 procent [Pew Research Center 2017]. Situaci v evropské vědě může dobře ilustrovat návratnost z šetření European Social Survey (ESS), které využívá pravděpodobnostní výběr z populace. V Německu je návratnost jedna z nejnižších; od roku 2010, tedy v posledních třech zveřejněných vlnách výzkumu, se pohybuje jen kolem 30 procent, přitom v předchozích vlnách, tedy mezi lety 2002 až 2008, činila návratnost kolem 50 procent (podobně nízkou návratnost mají německé pravděpodobnostní výběry i v šetření International Social Survey Programme (ISSP), v roce 2014 to bylo 35 procent [Jutz, Scholz 2016]). Ve Velké Británii mají šetření ESS dlouhodobě návratnost přes 50 procent, ale v poslední zveřejněné vlně z roku 2014 činila návratnost jen 44 procent. Česká republika si v ESS v porovnání s jinými zeměmi vede docela dobře: v posledních čtyřech vlnách šetření měla návratnost 70, resp. 68 procent [ESS 2017].

V pořadí druhým velkým problémem výběrových šetření je snižující se prestiž metody, jak mezi výzkumníky, tak u veřejnosti [Gobo, Giamietro 2011], přičemž v druhém případě je již možné mluvit o částečné ztrátě důvěry v průzkumy veřejného mínění. Velkou ránu důvěře veřejnosti utržily události roku 2016, a to hlasování o brexitu a volba amerického prezidenta. Oběma těmto lidovým hlasováním předcházelo velké množství průzkumů, které v situaci rovnoměrného rozložení podpory oběma soupeřícím stranám přinášely informace o těsném vítězství Hillary Clinton a odpůrců brexitu. V obou případech dopadly volby opačně. V obou případech však velký podíl na zklamání veřejnosti nesla i média, která o každém procentním bodu preferencí referovala jako o významném a reálném rozdílu a spoluvytvářela očekávání, která se nakonec nevyplnila. Nízká důvěra laické veřejnosti ve výběrová šetření může negativně ovlivnit budoucnost sociálněvědního výzkumu; nedůvěra se může projevit v dalším poklesu návratnosti nebo třeba v menší ochotě politiků financovat výzkum využívající výběrová šetření. V České republice zjišťovalo důvěru veřejnosti ve výběrová šetření Centrum pro výzkum veřejného mínění (CVVM), a to dosud pouze jednou, v březnu 2018. Výzkumům veřejného mínění podle CVVM rozhodně či spíše důvěřuje 49 procent respondentů, naopak 37 procent výběrového souboru spíše nebo rozhodně nedůvěřuje [Hanzlová 2018]. Z tohoto údaje se neodvažuji soudit, zda je v současnosti zjištěná důvěra české populace ve výzkumy veřejného mínění vysoká, či nízká, neboť nemám časové srovnání a informaci o tom, zda důvěra v posledních letech klesá, či nikoliv.

S přihlédnutím k výše uvedeným problémům výběrových šetření se nabízí otázka, zda nehledat lepší zdroj výzkumných dat. Nebylo by lepší využít technologického pokroku a poohlédnout se po nějaké jiné metodě? V akademickém prostředí se v uplynulých deseti letech začaly zvedat kritické hlasy, které právě k takovému poohlédnutí vybízejí. Savage a Burrows [2007] vyzývají k tomu, aby byla výběrová šetření ve výzkumu upozaděna a do středu zájmu nastoupila takzvaná velká data a sběr dat prostřednictvím internetu. Naznačují, že vědci a vědkyně, které se drží kritizované metody výběrových šetření, by se brzo mohli ocitnout v pozici toho, který „chvíli stál a již stojí opodál". Článek Savage a Burrowse vzbudil velkou pozornost a následovalo ho několik přímých reakcí, ať už pozitivních (např. [Webber 2009]), či kritických (např. [Crompton 2008;

Couper 2013]). Vyšly i další texty zabývající se velkými daty, ať už z epistemologického hlediska, které předpovídá změny, jež velká data přinesou do procesu vědeckého poznávání [Kitchin 2014; McFarland, Lewis, Goldberg 2016; Chandler 2015], či z metodologického hlediska, které velká data kriticky komentují jako omezený výzkumný nástroj sociálních věd [Grimmer 2018; Couper 2013; Boyd, Crawford 2012; Murphy et al 2011; Hargittai 2015].

Ačkoliv si velká data ještě neprorazila cestu do středního proudu sociologického výzkumu, lze očekávat, že se tak brzy může stát. Sociální vědci a vědkyně se opatrně seznamují s daty ze sociálních sítí a uvažují o možnostech, jak využívat i jiná velká data. Ovšem ještě před tím, než velká data vstoupí do sociologického výzkumu v plné síle, je nutné reflektovat jejich vlastnosti a identifikovat výhody a nevýhody, které využívání tohoto typu dat v sociálních vědách má. V další části tohoto článku přináším informace o tom, co jsou velká data, jaká je jejich specifická povaha, jak je lze využít pro výzkum a které jejich problematické charakteristiky ve výzkumu zohledňovat.

Co jsou velká data?

Veřejnost i odborníci používají název velká data (z angl. Big data) běžně, v sociálních vědách se částečně ujal i alternativní název *organická data*, jehož autorem je významný sociálněvědní metodolog Robert Groves, tento termín se ale používá výjimečně [Couper 2013; Habermann, Kennedy, Lahiri 2016]. Velká data mohou být definována jako soubory dat, které jsou tak velké, že je nedokáže zpracovávat běžný statistický software [Snijders, Matzat, Reips 2012]. O jak velké soubory se tedy jedná? Může to být např. objem dat o velikosti 2,5 petabytů (peta = 10^{15}), která obsahují informace o jednom milionu nákupních transakcí amerického řetězce WalMart, jež byly zaznamenány během jedné hodiny. Nebo to mohou být dvě a půl miliardy komentářů, 2,7 miliardy lajků a 300 milionů fotek, které v roce 2012 nahráli uživatelé Facebooku za jeden den [Kitchin 2014]. Pro velká data není ale typická jen kvantita, ale i kvalita. Jsou vysoce granulovaná a detailní a v některých případech zachycují realitu v každém okamžiku („moment-to-moment" data), např. data z GPS, která obsahují údaje o poloze objektu v průběhu pohybu. Někteří dokonce tvrdí, že velká data jsou tak detailní, že v podstatě zachycují realitu jako takovou [Chandler 2015].

Pro popis toho, co velká data vlastně jsou, existuje několik definic. Nejpoužívanější z nich je definice třemi V, ve které každé V odkazuje k jedné typické vlastnosti velkých dat [Couper 2013; Chandler 2015; Japec et al. 2015; Daas, Roos, van de Ven, Neroni 2012]. Volume (objem) velkých dat překračuje kapacitu tradičních metod pro uchovávání a zpracování dat; velocity (rychlost) vyjadřuje, že velká data jsou získávána v reálném čase, tj. v okamžiku, kdy se událost skutečně děje, a variety (různorodost či variabilita) znamená, že velká data jsou syrová, neuspořádaná, nestrukturovaná a nepřipravená k analýze.

Další definice [Boyd, Crawford 2012] popisuje velká data jako „kulturní, technologický a vědecký fenomén", jehož existence je dána kombinací několika činitelů. Tím prvním je

technologie, konkrétně maximalizace výpočetní síly a algoritmické přesnosti získávat, analyzovat, spojovat a porovnávat velké soubory dat. Tím druhým je analýza, která umožňuje identifikovat ve velkých datových souborech struktury, což lze využít pro ekonomické, sociální a právní účely. Tím třetím je mytologie, konkrétně rozšíření přesvědčení, že velké datové soubory nabízejí vyšší formu poznání, jež přináší dosud neobjevené informace a jež „má auru pravdy, objektivity a přesnosti".

Velká data lze rozdělit na tři typy podle toho, z jakého zdroje vznikají:

1. Administrativní data vznikají pro účely veřejné či státní správy a disponují jimi zpravidla státní či veřejné instituce. Administrativními daty jsou záznamy ministerstev, státních úřadů nebo třeba data z elektronické evidence tržeb [Couper 2013].
2. Transakční data jsou generována při elektronických transakcích a můžou jimi být údaje o platbách z debetních/kreditních karet, informace z věrnostních karet obchodů, záznamy telefonních operátorů o hovorech a přenosech dat, údaje z internetového vyhledávače či historie internetového prohlížeče. Transakční data jsou „automatický vedlejší produkt transakcí a aktivit" [Couper 2013] a jejich primárním účelem je vyřízení transakce nebo uživatelského požadavku na službu. Transakční data bývají zpravidla využívána k marketingovým účelům a příjemci služeb firem si často vůbec nejsou vědomi toho, že data o jejich chování jsou ukládána a dále využívána.
3. Data ze sociálních médií či sítí, jako je Facebook, Instagram, Twitter, Tumblr a další, vznikají aktivitou uživatelů na síti, která představuje sdílení obsahu, jako jsou krátké texty, fotografie nebo internetové odkazy, a komunikaci s ostatními uživateli sítě [Murphy, Hill, Dean 2013].

Anglický název Big data odkazuje k fenoménu Velkého bratra (Big Brother) z Orwellova románu 1984 [Boyd, Crawford 2012]. Velká data podobně jako Velký bratr obsahují informace o všem, co člověk (uživatel v rámci jedné služby) dělá (placení kartou, prohlížení internetu, lajkování na Facebooku). Lidé, kteří disponují těmito daty, získávají velké množství leckdy intimních informací o velkém množství lidí. Podobnost s Velkým bratrem nutí k otázkám týkajícím se ochrany osobních údajů; tento problém je ale velice komplexní, nemá jednoznačné řešení a představuje jeden z největších praktických, právních a etických problémů práce s velkými daty [Habermann, Kennedy, Lahiri 2016; Couper 2013; Boyd, Crawford 2012; Savage, Burrows 2009; Kalyvas, Overly 2015; Murphy et al 2014].

Případné využití velkých dat v sociálněvědním výzkumu je specifické tím, že výzkumníci přímo nekontrolují jejich sběr [Japec et al. 2015]. Velká data jsou „rutinně získávána jako vedlejší produkt institucionálních transakcí" [Savage, Burrows 2007] a jejich sběr není nijak designován. Tinati a kolegové [Tinati, Halford, Carr, Pope 2014] tuto specifickou vlastnost popisují jako „sbírání dat v divočině" (in the wild), někdy se o velkých datech mluví jako o „digitálních stopách" (digital footprints), které za sebou zanechávají jednotlivci v každodenním offline i online životě [Chandler 2015; Golder, Macy 2014; McFarland, Lewis, Goldberg 2016].

Využití velkých dat ve výzkumu je různorodé a lze jej ilustrovat na reálných příkladech. Např. administrativní data mohou v metodologii posloužit ke zjištění validity dat

z výběrových šetření. Ansolabehere a Hersh [2012] srovnávali výpovědi respondentů v internetovém dotazníkovém šetření o účasti ve volbách s údaji z voličských registrů. Informace z registrů spojili s neanonymními daty ze šetření a porovnáním zjistili, že respondenti svou účast ve volbách nadhodnocovali. Data ze sociálních sítí lze využít např. ke studiu změn nálad uživatelů. Bollen a kolegové [2011] analyzovali vztah mezi náladou, kterou lidé vyjadřovali na sociální síti Twitter, a vývojem na burze. S pomocí tzv. mood tracking nástrojů měřili průměrnou náladu na Twitteru a zjistili, že zahrnutí této proměnné do modelu vývoje burzovního indexu Dow Jones dokáže odhady vývoje zpřesnit. Mediálně známým se stal experiment, který z vlastní iniciativy provedl Facebook v roce 2012 [Grifin 2014] a jehož cílem bylo zjistit, zda má tato sociální síť schopnost ovlivňovat náladu uživatelů. Do experimentu vstoupilo celkem 700 tisíc uživatelů, kteří byli rozděleni do dvou skupin; první skupině Facebook zobrazoval pozitivní mediální obsah, druhé negativní. Výsledkem bylo zjištění, že Facebook dokáže prostřednictvím zobrazovaného obsahu uživatele činit buď smutnějšími, nebo šťastnějšími. Rovněž Bond a kolegové [2012] manipulovali zobrazování obsahu několika skupinám uživatelů sociální sítě, aby zkoumali vliv zobrazovaného obsahu na volební chování. Podobné experimenty mohou provádět pouze výzkumníci a výzkumnice, kteří jsou přímo navázaní na provozovatele sociálních médií. Ostatním zbývá pouze sekundární využití dostupných dat, jako to např. udělali Kosinski a kolegové [Kosinski, Stillwell, Graepel 2013], kteří ukázali, že z facebookových lajků lze predikovat charakteristiky jednotlivých uživatelů Facebooku, jako je sexuální orientace, etnicita či politická orientace. Další výzkumníci využili transakční data z vyhledávače Google, aby se z nich pokusili odhadnout, zda se blíží chřipková epidemie [Lazer et al. 2014].

Velká data, především transakční data, kterými disponují velké firmy, nalézají obrovské využití v marketingu. Využívání transakčních dat je poměrně levné, protože jsou vedlejším produktem prodeje služeb či produktů, a tak nevznikají další náklady na sběr dat o zákaznících. Např. americký maloobchodní řetězec analyzoval data o nákupech za období dlouhé dvanáct let a nalezl v nich vztahy, které odhalily vzorce v nákupním chování zákazníků; některé výrobky zákazníci nakupovali společně s konkrétními jinými výrobky. Řetězec na základě vzorců upravil umístění zboží v regálech, aby zboží, které zákazníci kupují společně, bylo umístěno blízko sebe, a dosáhl tím 16procentního nárůstu tržeb [Kitchin 2014]. V podstatě stejný princip dlouhodobě využívá internetový obchod Amazon. Tento gigant mezi e-shopy uchovává informace o tom, co si jednotliví zákazníci v minulosti koupili, a při každé další návštěvě webu jim navrhuje, aby si koupili další zboží. Nabízí jim přesně to zboží, které si v minulosti koupili jiní lidé, kteří si mimo jiné už koupili to samé zboží, co zákazník, kterému je zobrazována nabídka. Marketingová strategie vyjadřovaná nedokončenou větou „Lidé, kteří si koupili stejnou věc jako vy, si rovněž koupili…“ se dlouhodobě vyplácí a převzali ji další webové obchody a stránky (např. IMDB, International Movie Database). Velká data k zajištění obchodního úspěchu použila i internetová televize Netflix, která produkuje vlastní televizní seriály. Při plánování natáčení seriálu House of Cards Netflix nejprve analyzoval ve své databázi preference svých předplatitelů, aby zjistil, které herce a režiséry mají nejvíc

v oblibě. Na základě tohoto průzkumu obsadil do hlavní role Kevina Spaceyho a režii svěřil Davidovi Fincherovi, neboť právě tito dva se ukázali jako neoblíbenější umělci předplatitelů Netflixu [Barnes 2013; Carr 2013].

Velká data jako data svého druhu

V roce 2008 vzbudil velkou pozornost článek internetového magazínu Wired, ve kterém jeho autor zvěstoval příchod nového vědeckého věku, kdy velká data dokážou přinést naprosto věrný, detailní a objektivní obraz reality [Anderson 2008]. Podle Andersona může věda konečně objektivně zachytit skutečnost v celé její šíři a hloubce, kterou výzkumníkům odhalí velká data. Na Andersonův komentář přímo reagovalo několik akademiček a akademiků [Kitchin 2014; McFarland, Lewis, Goldberg 2016; Boyd, Crawford 2012; Chandler 2015] a vytýkali mu především jeho představu o povaze velkých dat. Velká data podle kritiků nejsou obrazem reality, nemluví „sama za sebe" a nerodí objektivní, čisté vědění. I velká data vznikají v určitých situacích a za určitých podmínek, které jsou plně dány lidskými aktéry. Prostředí, ve kterém velká data vznikají, není sociální a kulturní vakuum a není to ani nově vytvořené „hřiště", jehož zákonitosti a pravidla nebyla ovlivněna člověkem. Reagující akademici proto volají po tom, aby velká data byla reflektována jako fenomén svého druhu, nikoliv jako objektivní obraz reality [Kitchin 2014; McFarland, Lewis, Goldberg 2016; Boyd, Crawford 2012].

Ve svém článku dále Anderson představuje radikální tvrzení, že díky velkým datům věda již nebude potřebovat teorii. V situaci, kdy „data mluví sama za sebe", jsou podle Andersona vědecké aproximace a modely zbytečné, protože poznání se bude samo vynořovat z velkých objemů dat. Deduktivnímu způsobu poznávání tak podle něj odzvonilo. Chandler [2015] ho v tomto kontextu doplňuje informací, že analytici, kteří pracují s velkými daty, nepřicházejí k datům s přesnými představami o tom, co v datech hledat, ale prvotní informace o zkoumaném fenoménu získávají nejprve z dat. Anderson a další, kteří vidí opuštění deduktivního způsobu poznávání jako velkou šanci pro rozvoj vědy, doufají, že se bude zvyšovat schopnost technologie zachycovat jevy, takže „svět začne mluvit sám za sebe bez zprostředkování omylným lidským interpretem" [Chandler 2015].

V kvantitativní sociologii je dedukce základním principem poznávání. Při práci s daty z výběrových šetření si výzkumníci přirozeně kladou otázky, na které hledají odpovědi deduktivním přístupem, a s tímto účelem přistupují k výzkumu od začátku až do konce. Induktivní přístupy neumožňují testovat hypotézy a vytvářet modely vztahů, což pro současnou praxi znamená obrovské omezení. Někteří autoři [Boyd, Crawford 2012; Barnes 2013] kritizují, že bez deduktivního přístupu nebudeme schopni studovat příčiny a následky sociálních jevů. Ovšem podle Savage a Burrowse [2007] to není problém; podle nich jsou sociální vědy ve studiu kauzality dlouhodobě neúspěšné a měly by se této ambice úplně vzdát. Naopak by se měly přeorientovat jen na popis sociálních jevů, k čemuž jsou podle nich velká data vhodnější než data z výběrových šetření.

Sociologie využívající velká data samozřejmě může používat induktivní analytický přístup, který opustil klasický princip. Z velkých dat lze např. získávat popisy společnosti či lidského chování. Příkladem sociálněvědní studie využívající indukci může být studie Webbera [2009], který z obrovského množství údajů o křestních jménech a příjmení odhadl informace o etnicitě a příslušnosti ke společenské třídě velkého množství jednotlivců žijících na daném území. Velká data tak využil k získání informace, kterou by jinak mohlo přinést jen velmi rozsáhlé výběrové šetření. Avšak navzdory možnému použití indukce není pravděpodobné, že se sociální vědy deduktivního přístupu v nejbližší budoucnosti vzdají. Dedukce bude nadále využívána pro analýzu dat z výběrových šetření a bude se snažit přežít i v oblasti analýzy velkých dat. Již v současné době využívají analytici a analytičky klasický statistickoanalytický přístup k analýze dat ze sociálních sítí, jako je např. obsahová analýza a popisná statistika k testování hypotéz o výskytu různých slov, klíčových slov a hashtagů, frekvencích sdílení různých obsahů či podobnosti jednotek v různých skupinách. Příklady takových prací nacházíme i v České republice (viz např. [Karaščáková 2013; Hrdina 2016; Vochocová, Mazák, Štětka 2016]). Deduktivní přístup se také jistě udrží v analýze dat, která vznikají propojením velkých dat s klasickým dotazníkovým šetřením, např. spojením dat z facebookového účtu s individuálními daty z neanonymního dotazníkového šetření. Tento druh analýzy může přinášet velmi zajímavé výsledky i navzdory omezením neanonymního dotazování [Kosinski, Stillwell Graepel 2013; Ansolabehere, Hersh 2012].

Reflexe povahy velkých dat pro využití ve výzkumné praxi

Ať už budeme velká data používat pro deduktivní, nebo induktivní poznávání, musíme je podrobit důkladnému rozboru z hlediska jejich povahy. Je potřeba se zabývat širšími souvislostmi používání velkých dat ve výzkumu, a to zejména ontologickou podstatou dat ze sociálních sítí a transakčních dat. V jakých podmínkách vznikají? Jak samotné sociální sítě a prostředky transakčních dat mění chování lidí? Jaké sociální skupiny přednostně vytvářejí velká data? V neposlední řadě je nutné reflektovat, jak ovlivňuje využití velkých dat samotný proces výzkumu; velká data vznikají bez dohledu výzkumníků a jejich samotný charakter svádí výzkumníky k určité perspektivě. Povaha těchto dat formuje design výzkumu a ovlivňuje výzkumníky v tom, jaké si kladou otázky a výzkumné cíle [Boyd, Crawford 2012]. Jako výzkumníci a výzkumnice využívající kvantitativní paradigmata musíme věnovat pozornost všem charakteristikám, ve kterých se velká data liší od dat z výběrových šetření, a zohlednit jejich netradiční vlastnosti ve statistické analýze a při interpretaci výsledků. Musíme se věnovat odchylkám od pravých hodnot zjišťovaných proměnných a dalším nepřesnostem, které snižují validitu závěrů výzkumu. Systematická analýza odchylek ve velkých datech však nebude jednoduchá, neboť tato data jsou výrazně komplexnější a komplikovanější než data z šetření. V následujícím textu se na využití velkých dat dívám optikou metodologie výběrových šetření a implicitně porovnávám jejich charakter s vlastnostmi, které mají data z výběrových šetření, jež nacházejí využití v klasické

statistické analýze. Postupně se zastavuji u několika zásadních problémů, které využívání velkých dat ve výzkumu doprovází. Jsou jimi náročnost analýzy velkých dat, jejich dostupnost, reprezentativita, vychýlení výběru jednotek, vliv sociálního okolí na hodnoty proměnných a další problematické jevy.

Náročnost analýzy velkých dat

V současnosti představuje překážku uvedení velkých dat do běžné výzkumné praxe skutečnost, že pro analytickou práci s nimi jsou potřeba programátorské schopnosti [Manovich 2011; Golder, Macy 2014; McFarland, Lewis, Goldberg 2016; Kitchin 2014]. Velká data je nutné umět sbírat, skladovat a analyzovat, spojovat je, případně je validizovat údaji z externích zdrojů. Dále je potřeba je uspořádat a připravit pro analýzu, neboť data sama o sobě nejsou strukturována tak, aby mohla být analyzována bez úpravy [Golder, Macy 2014]. Všechny uvedené operace jsou náročné na provedení a bez programátorského tréninku je nelze zvládnout. Sociální vědci a vědkyně s klasickým vzděláním v analýze dat v současnosti nemají potřebné znalosti k tomu, aby mohli velká data analyzovat, a proto se analýzou velkých dat zabývají především programátoři. Tomu napovídá i to, že většina dosud publikovaných studií využívajících velká data byla zpracována programátory a datovými vědci [Golder, Macy 2014; McFarland, Lewis, Goldberg 2016].

Angažmá IT expertů v hájemství sociálních věd je podle některých autorů problematické, neboť většina z nich nemá dostatečné sociálněvědní vzdělání. Jejich výzkum je proto často „redukcionistický, funkcionalistický a ignoruje vlivy kultury, politiky, veřejné politiky a kapitálu“ [Kitchin 2014]. Další autorky zase upozorňují na významnou genderovou disproporci v komunitě počítačových vědců, mezi kterými jsou nedostatečně zastoupeny ženy [Boyd, Crawford 2012]. Je známo, že gender výzkumníka nepřímo ovlivňuje tematické zaměření plánovaných studií i konkrétní výzkumné otázky a může mít vliv i na interpretace zjištění z výzkumu [Harding 2010; de Madariaga 2012.]. Silná převaha mužů mezi počítačovými experty, kteří analyzují velká data pro účely sociálněvědních analýz, tak může deformovat výzkumnou agendu směrem k tématům a problémům, které jako prioritní vidí především mužská část populace.

Na absenci sociálních vědců a vědkyň v oblasti velkých dat ukazuje i fakt, že v současné době nejsou k dispozici učebnice, příručky či manuály představující analytické metody práce s velkými daty. Významní výzkumníci a výzkumnice v oboru kvantitativní sociologie v uplynulých desítkách let publikovali množství nejrůznějších učebnic a instruktážních textů, které vysvětlují zásady práce s daty z výběrových šetření a principy mnoha typů statistické analýzy (např. [Babbie 2015; Bollen 2009; Saris, Gallhofer 2014; Tabachnick, Fidell, Osterlind 2001; Hox, Roberts 2011]), k analýze velkých dat však učebnice nejsou. Chybí také obecnější pojednání o tom, jaký typ úloh je možné s velkými daty realizovat či jaký druh výzkumných otázek lze s jejich využitím řešit. Dosud vycházejí především texty, které shrnují velká data jako fenomén, který může mít na sociálněvědní výzkum v budoucnosti velký vliv (např. [Boyd, Crawford 2012; Couper 2013; Murphy et al 2014; McFarland, Lewis, Goldberg 2016; Savage, Burrows 2007, Burrows, Savage 2014]).

Dostupnost velkých dat pro sociálněvědní výzkum

Omezená dostupnost velkých dat je považována za největší překážku, která stojí velkým datům v cestě do sociálněvědního mainstreamu [Haberman, Kennedy, Lahiri 2016; Couper 2013; Savage, Burrows 2007; Boyd, Crawford 2012; Manovich 2011; McFarland, Lewis, Goldberg 2016; Driscoll 2014]. Data ze sociálních sítí jsou omezena méně než transakční data, což se odráží i v jejich častějším používání v sociálněvědním výzkumu [Murphy et al. 2014]. Facebook či Twitter umožňují veřejnosti přístup do datového úložiště, ze kterého lze při znalosti programátorských postupů získat aktuálně dostupná data. I zde se však postupně utahují kohouty: informační giganti jako Facebook se v posledních letech stali datovými oligopoly, disponujícími obrovským množstvím informací, a uvědomují si výsadní postavení, které díky informacím získávají. Aby si toto postavení udrželi, poskytují veřejnosti, tedy i výzkumníkům, stále menší objemy dat [Boyd, Crawford 2012; Driscoll 2014; McFarland, Lewis, Goldberg 2016]. Vědcům je údajně k dispozici pouze 1 až 10 procent veřejných tweetů [Boyd, Crawford 2012; Driscoll 2014], množství a typ dat, které je možné získat z Facebooku, je rovněž regulováno. V oblasti analýzy dat ze sociálních sítí tak získávají rozhodující konkurenční výhodu vědci, kteří jsou u firem, jako je Facebook, *přímo zaměstnáni*. Nejenže jim jsou dostupná všechna existující data, ale získávají i možnost realizovat na sociální síti vlastní experimenty. Jejich vědecká práce tedy může nad prací akademiků v mnohém vyniknout [Manovich 2011; Savage, Burrows 2007].

Administrativní a transakční data jsou sociálním vědkyním a vědcům dostupná velice málo nebo vůbec; administrativními daty disponuje stát a jeho úřady a podléhají pravidlům ochrany soukromí, jejichž přísnost se liší v jednotlivých zemích. Transakční data mají plně ve své moci soukromé firmy, jako jsou banky, telefonní operátoři či provozovatelé internetových vyhledavačů a prohlížečů. Mnozí upozorňují na to, že se mocenská pozice soukromých firem oproti minulosti posiluje právě díky tomu, že velké firmy díky digitalizaci disponují obrovským množstvím informací o jednotlivcích i společnosti jako celku. S tím se na druhou stranu oslabuje pozice sociálních věd, které ztrácejí monopol na využívání zdrojů informací o společnosti, a v důsledku i monopol na popis a vysvětlování společenských jevů. Savage a Burrows [2007] pro tuto společenskou situaci používají termín převzatý od Thrifta [2005] *knowing capitalism*, který označuje kapitalistickou společnost, kde soukromé firmy mají prostředky k získávání vědění a disponují mocí toto vědění rekonstruovat a přetvářet ve svůj prospěch. V tomto novém společenském uspořádání ztrácejí svoji moc instituce, jako je stát a na něj navázaná akademická sféra.

Neschopnost získávat velká data pociťují sociální vědci jako obtíž, která může mít zásadně negativní vliv na budoucnost vědy a výzkumu. Metodolog Robert Groves vyzývá, aby věda v součinnosti se státem začala tento problém neprodleně řešit. Stát by měl akademické sféře nabídnout pomoc při vyjednávání s velkými firmami disponujícími velkými daty a iniciovat vznik institucí na pomezí akademické, resp. státní a soukromé sféry, které budou spolupracovat s vlastníky velkých dat a přesvědčovat je o nutnosti používání těchto dat ve výzkumu [Haberman, Kennedy, Lahiri 2016]. Čím konkrétně by mohli akademici velké firmy přesvědčit, aby jim data poskytovaly, však Groves neuvádí.

Reprezentativita za populaci

Výzkumníci, kteří běžně pracují s daty z výběrových šetření, kladou velký důraz na reprezentativitu, případně kvazireprezentativitu dat za populaci národního státu či za jinou, specifickou populaci (studenti, ženy na rodičovské dovolené atp.). Velká data, která nejsou reprezentativní za běžně zkoumané populace, jsou jim podezřelá [Savage, Burrows 2007]. Často vyjadřují námitku, že „velikost není všechno" a že obrovské objemy velkých dat nejsou schopny vyvážit to, že zahrnutí jedinci proporčně neodpovídají obecné (nebo jiné známé) populaci [Boyd, Crawford 2012; Couper 2013; Hargittai 2015; McFarland, Lewis, Goldberg 2016; Golder, Macy 2014].

Problém (ne)reprezentativity se týká dat ze sociálních médií i transakčních dat. Transakční data firem obsahují jen údaje o lidech, kterým daná firma dodává služby; jsou to tedy např. uživatelé platebních karet jedné banky či klienti konkrétního telefonního operátora. V marketingovém výzkumu pro účely firem není nereprezentativita problém, pokud jsou ale taková data využívána pro účely sociálněvědního výzkumu, musí být analýza doprovázena důrazným upozorněním, že výsledky nejsou zobecnitelné na populaci národního státu. Obdobně populace lidí využívajících sociální sítě není v současné době stejná jako obecná populace. Z analýz existujících profilů na Twitteru a na Facebooku vyplývá, že např. v USA jsou uživatelé sociálních sítí mladší, vzdělanější a bohatší [Golder, Macy 2014] a jednotlivci na síti se od obecné populace liší i v pohlaví a rase [Hargittai 2015]. Ani v Česku nezahrnují sociální sítě celou populaci; o demografickém složení profilů na Facebooku informuje např. Dočekal [2015].

Z hlediska sociologické metodologie lze porušení zásady reprezentativity u velkých dat považovat za problém. Existují však specifické případy, kdy je možné podmínku reprezentativity uvolnit a získat druh informace, kterou lze jinak zjistit jen obtížně. Ve výběrových šetřeních máme problém získávat data o běžných denních činnostech lidí; když používáme deníky, respondenti do nich často zapisují nepřesně, leccos opomenou, něco si třeba naopak vymyslí. Navíc tím respondenty velice zatěžujeme. Velká data nám mohou pomoci právě při získávání deníkového typu informací; jejich přesnost by mohla vyvážit nereprezantativitu. Data z platebních a věrnostních karet umožňují zjistit, v kterou denní dobu dělají lidé běžné nákupy nebo v jakém období nakupují speciální zboží, jako je třeba nábytek; data z vyhledávačů, prohlížečů nebo navigačních aplikací dodávají např. informace o tom, v kterou dobu lidé zhruba začínají pracovat či jakým způsobem a za jakým účelem se pohybují ve veřejném prostoru. Data z různých mobilních aplikací, které lidé využívají ke sportu, dietě nebo kontrole spánku a mnoha dalším činnostem, jsou rovněž zdrojem velice detailních informací o běžných činnostech a návycích.

Vychýlené soubory jednotek

Soubory dat ze sociálních sítí, ale i transakčních dat obsahují jednotky různého typu; mohou to být lidé, nákupy, platby kartou, tweety, fotografie, lajky, statusy a další jednotky informací

[Murphy et al. 2014]. Tyto soubory jsou vždy do jisté míry odchýlené od populace, na kterou by z dat šlo usuzovat (uživatelé Facebooku, zákazníci obchodu, klienti banky atd.), neboť ne všechny jednotky populace jsou stejně aktivní. Pro transakční data a data ze sociálních sítí platí, že obsahují jen informace pocházející od jednotlivců, kteří učinili nějakou akci. Do výběrů se dostávají jen lidé, kteří mají věrnostní kartu obchodu a používají ji, lidé, kteří použili hashtag, lidé, kteří zaplatili platební kartou, lidé, kteří sdílejí na Facebooku fotografie, případně hashtagy, které byly použity v daném období, fotografie, které byly zveřejněny v daném období na sociální síti, atp. Do výběru se naopak nedostanou ti, kteří sice nakupují, ale bez využívání věrnostních a platebních karet, nebo např. ti, kteří tweetují, ale nepoužívají hashtagy. Couper [2013] tuto vlastnost velkých dat vyjadřuje výstižnou větou „Big data tends to focus more on the 'haves' and less on the 'have-nots'". V důsledku tohoto fenoménu vypadává z analyzovaných souborů velkých dat podstatná část lidí, kterých se případný zkoumaný problém týká, a výzkumníci tak získávají hodnoty, které jsou odchýlené od pravé hodnoty platné pro zkoumanou populaci, tedy např. uživatele Facebooku, klientů spořitelny atp.

V případě sociálních sítí existuje více možností uživatelské aktivity, než je jen existence profilu na síti. Jak velká je proporce lidí, kteří sociální médium aktivně používají? Sociální sítě jsou plné „posluchačů", tedy lidí, kteří především konzumují obsah ostatních uživatelů, ale sami vytvářejí velmi málo obsahu, případně se neprojevují vůbec [Boyd, Crawford 2012]. O těchto posluchačích tak výzkumníci nezískávají žádná data. Skutečností je i to, že lidé mají různou pravděpodobnost stát se aktivním uživatelem sítě [Murphy et al. 2014]; např. manuálně pracující mohou být v obsahu sociálních médií reprezentováni méně než lidé, kteří tráví naprostou většinu dne v kanceláři či doma u počítače. A nejde jen o uživatele, kteří se do souboru nedostanou, protože nejsou aktivní, ale i o uživatele, kteří sice aktivní jsou, ale de facto neexistují. Část profilů na sociálních sítích tvoří duplicitní účty a není ani známé množství falešných účtů, které jednotlivci zakládají z různých důvodů. Samostatný problém tvoří účty lidí, kteří jsou placeni za šíření propagandy. Tito lidé na sítích diskutují jako normální uživatelé, ovšem s účelem šířit propagandu [King, Pan, Roberts 2016; Murphy et al. 2014]. Příbuzným problémem je aktivita robotů, kteří vytvářejí obsah, jenž je na první pohled nerozeznatelný od obsahu, který vytváří skutečný uživatel [Boyd, Crawford 2012; McFarland, Lewis, Goldberg 2016; Bessi, Ferrara 2016; Murphy et al. 2014]. Roboty mohou být např. falešné twitterové účty, využívané pro účely politického boje s cílem ovlivnit diskurz na sociální síti. Existují sice metody, jak v datech identifikovat takové robotické účty, sociálněvědní výzkumníci je však běžně neovládají [Bessi, Ferrara 2016].

K problematice vychýlených souborů jednotek ze sociálních sítí se váže i fakt, že lidé používají konkrétní sociální sítě ke sdílení konkrétního typu obsahu [Hargittai 2015]. Např. Twitter lidé primárně používají k verbální psané komunikaci nebo k odkazování na jiný internetový obsah, naopak na Instagram dávají jenom fotky. Jednotliví uživatelé často neduplikují obsah na svých twitterových, facebookových a instagramových účtech, takže obsahy těchto médií se nepřekrývají, ale doplňují. Pokud analyzujeme obsah na sociálních sítích a nemáme spojená data ze všech profilů uživatelů na různých sociálních sítích, nemůžeme mít celkový přehled o obsahu, který

uživatelé sdílejí přes všechny své profily na sociálních sítích. Zkoumá-li např. analytik vztah mezi lajkováním na Facebooku a fotografiemi, které uživatel nahrává na svůj facebookový profil, mohou mu unikat podstatné informace, neboť významná část uživatelových fotek se možná nachází na Instagramu. V analýzách je proto nutné explicitně uvádět, že usuzování z dat z jednoho sociálního média se týká pouze a výhradně jen účtů na konkrétním zkoumaném médiu.

Množství proměnných ve velkých datech

Jeden z rozdílů mezi velkými daty a daty z výběrového šetření není jen v množství případů v jednom souboru, ale v i množství proměnných. „Velká data obsahují hodně případů a málo proměnných, zatímco data z šetření obsahují hodně proměnných, ale málo případů", píše ve svém článku zjednodušeně Couper [2013]. Nedostatkem je podle něj malé množství demografických proměnných či jejich úplná absence. Uvádí, že internetové prohlížeče, které generují transakční data, často nemají údaje o demografii uživatele, a tak musí základní demografické údaje odhadovat z vlastních dat o vyhledávání. Tyto odhady ovšem obsahují chyby; Google odhaduje pro účely svých výzkumů gender uživatelů, je jej však schopen identifikovat správně pouze v 75 procentech případů [Couper 2013]. Ani na sociálních sítích není identifikace genderu a dalších demografických proměnných zcela přesná; uživatelé o sobě často demografické informace neudávají, případně poskytují nepravdivé údaje. Savage a Burrows [2007] navrhují, aby sociální vědci tento problém řešili inspirací v marketingovém výzkumu, kde se prosadilo používání proměnné *místo bydliště*. Tato proměnná je údajně velmi silným prediktorem, schopným nahradit základní demografické proměnné, jako je třída, etnicita či vzdělání [Webber 2009]. Místo bydliště by prý mohlo suplovat množství demografických proměnných, se kterými sociální věda pracuje dnes; proti tomuto nápadu se však rychle zvedly kritické hlasy [Crompton 2008]. Zajímavou ukázku odvozování demografických proměnných přinesl Webber [2009], který ve velkých datech z údajů o křestním jménu a příjmení dokázal s vysokou pravděpodobností odhadnout etnicitu a společenskou třídu jedince. K problematice množství a charakteru proměnných ve velkých datech je třeba dodat, že vývoj v oblasti digitálních technologií jde kupředu velmi rychle a jednotlivci a firmy generující a zpracovávající velká data dosahují podstatného pokroku v tom, jak různorodé informace jsou schopni získávat. Zlepšují se v odhadování demografických proměnných a zvyšují množství proměnných, které dokážou v datech identifikovat.

Nejistota ohledně motivace k chování

Transakční data poskytují výhradně údaje o chování (např. platba kartou) a provedených akcích (např. vyhledávané heslo v Googlu), neříkají nám ale nic o motivacích k tomuto chování či o postojích a hodnotách, které člověka k chování či akci vedou. Data ze sociálních sítí dodávají o něco méně informací o chování, mohou ale lépe sloužit jako zdroj informací o postojích a hodnotách [Couper 2013]. I dat ze sociálních sítí se však často týkají pochybnosti o tom, co pozorované chování vlastně znamená. Na Twitteru např. nevíme, zda uživatel/ka retweetováním

vyjadřuje danému obsahu podporu, či zda se jedná o subversivní tweet, který může být sdílen z jiných důvodů, než je vyjádření souhlasu nebo podpory; typickým příkladem takového subverzivního tweetu je retweetování obsahu twitterového účtu kontroverzního politika či jiné veřejně známé osoby, které má upozornit na pozoruhodný či bizarní obsah. To samé platí o lajkování na Facebooku, ani u něj neznáme motivaci uživatele k lajkování.

Příkladem neznámé motivace k akci v transakčních datech je vyhledávání v internetovém vyhledavači. O tom, jak velkou roli hrají motivace k vyhledávání, se přesvědčili výzkumníci, kteří studovali rozšířenost užívání drogy zvané *Salvia Divinorum* [Murphy, Dean, Hill, Richards 2011]. Domnívali se, že ke zjištění rozšířenosti drogy je možné použít data z vyhledavače Google a z Twitteru, indikátorem rozšířenosti měla být podle jejich předpokladu četnost vyhledávání a tweetování o této látce. Z dat Googlu a Twitteru zjistili frekvence výskytu klíčového slova *Salvia Divinorum* v určitém období a porovnali je s daty projektu National Survey of Drug Use and Health, ve kterém respondenti vypovídali, zda tuto konkrétní látku užili, či nikoliv. Porovnání obou datových zdrojů ukázalo, že ve velkých datech byla četnost výskytu mnohem větší než v datech z šetření, což donutilo výzkumníky pátrat po příčinách. Zjistili, že bylinu ve zkoumaném období kouřila zpěvačka Miley Cyrus na oslavě svých narozenin a video zachycující tuto událost se dostalo na YouTube. Příčinou velkého výskytu drogy ve vyhledávání Googlu a na Twitteru bylo, že se lidé touto událostí bavili, nikoliv to, že by si drogu chtěli opatřit.

Vliv sociálního okolí

Stejně jako v offline životě jsou lidé, kteří za sebou nechávají digitální stopy, pod vlivem sociálního okolí. Na sociálních sítích se tento vliv může projevovat jako cenzura, kdy je administrátorem či majitelem účtu vymazáván obsah a blokováni uživatelé, nebo jako autocenzura, kdy sám uživatel omezuje svůj projev z důvodu obav z reakce sociálního okolí, případně za účelem budování vlastní image. Následkem cenzury i autocenzury je to, že některé jednotky pozorování, jako jsou komentáře, lajky, tweety či jiný typ obsahu, jsou buď vymazány cenzorem (cenzura), nebo vůbec nevzniknou (autocenzura).

Cenzura na sociálních sítích někdy probíhá ze strany provozovatele sítě z důvodu toho, že cenzurovaný obsah odporuje zásadám komunity (např. pornografie), jindy ji uplatňují majitelé profilů, jako jsou soukromé firmy, značky, celebrity nebo politici. Obsah na sociální síti je tak často filtrován z hlediska zájmů firem, politiků a dalších významných lidí, kteří nechtějí, aby jim uživatelé nevhodnými komentáři kazili budovanou image.

Autocenzura na sociálních sítích je důsledkem existence společenských norem a přímého kontaktu s ostatními uživateli sítě. Uvnitř sociálních sítí existuje velké množství menších sítí uživatelů, pro které jsou typické sdílené postoje a hodnoty. V těchto menších sítích panují mocenská pravidla, která jsou zčásti dána reálnými mocenskými silami ve společnosti, vtělenými v charakteristiky, jako je pohlaví, vzdělání či etnicita, zčásti nově vytvářená v síti aktivitou uživatelů. Uživatelé těmto mocenským pravidlům přizpůsobují svoje chování, postupem času se učí, co si mohou a nemohou dovolit, a sdílejí obsah, který je více či méně v souladu s pravidly

subpopulace. Jejich komentáře, lajky, sdílené fotografie a statusy tedy vznikají pod přímým vlivem ostatních uživatelů sítě. Cílem přizpůsobení se pravidlům v síti je minimalizace sankcí za chování, postoje či názory, se kterými většina v subpopulaci nesouhlasí. Sankcemi na síti zpravidla bývají negativní, či dokonce agresivní komentáře ostatních uživatelů či zablokování přístupu na profil, případně zablokování účtu. Autocenzura v důsledku může vést i k tomu, že některé skupiny lidí se na některých místech sítí vůbec nevyjadřují, např. proto, že patří k diskriminované skupině, která je pro diskutující většinu snadným cílem útoku. Představovat si sociální sítě jako svobodné virtuální fórum, které přináší skutečné a objektivní informace o tom, co si lidé myslí, je liché.

Vliv sociálního okolí může aktivitu uživatelů nejen potlačovat, ale také ji podporovat či moderovat. Takový tlak může vytvořit např. vysoká popularita příspěvku na sociální síti vyjádřená množstvím lajků. Sama popularita příspěvku může ovlivnit postoj uživatele k zobrazovanému obsahu; nejenže se např. na Facebooku vysoce populární obsah zobrazuje více lidem, čímž může získat více lajků, ale i samotný fakt, že je příspěvek populární, může v uživatelích vzbudit sympatii k zobrazovanému obsahu, kterou uživatel vyjádří lajkem. Aktivita ostatních uživatelů v síti hraje velkou roli v diskuzích, kde uživatelé přímo reagují jeden na druhého a většina komentářů je přímo závislá na ostatních komentářích.

Právě přímý vliv sociálního okolí je činitelem, který je zodpovědný za jeden z hlavních rozdílů mezi daty z výběrových šetření a daty ze sociálních sítí. Data z výběrových šetření obsahují jednotky, které jsou na sobě vzájemně nezávislé v tom smyslu, že nejsou přímo ovlivněny anticipovanými reakcemi ostatních lidí. V momentě dotazování je přítomno mnohem méně činitelů, které ovlivňují prezentovaný postoj či názor, a panuje v něm často výrazně příjemnější klima než např. v diskuzi na Facebooku. Ve výzkumném rozhovoru se respondent do značné míry vyjadřuje nezávisle na mínění ostatních lidí, neboť neví, jak odpovídají ostatní respondenti ve výběru, a navíc se neobává, že prezentovaný názor vzbudí hlasitou a leckdy nevybíravou kritiku. Na sociálních sítích je situace zcela jiná; vliv sociálního okolí je zde přímý a reakce uživatelů jsou závislé na reakcích ostatních uživatelů, neboť obsahují inklinaci k chování skupiny, ovlivnění ostatními uživateli či obavy z projevení vlastního názoru. Nezávislost pozorování v datech z výběrových šetření je kvalita, které nemohou data ze sociálních sítí konkurovat.

Vliv sociálního okolí hraje svoji roli i v transakčních datech, zde má však spíše charakter odchylky vzniklé v důsledku sociální desirability než přímého ovlivnění ostatními lidmi. Z transakčních dat mohou vypadávat údaje reflektující chování, jež lidé považují za společensky nežádoucí, jako je např. nákup alkoholu či erotických pomůcek. Nákupy tohoto zboží mohou být výrazně častěji placeny v hotovosti, nikoliv platební kartou či bankovním převodem [Couper 2013]. Pokročilejší uživatelé internetu mohou rovněž maskovat svůj pohyb v online prostoru, pokud mají pocit, že se dopouštějí sociálně nevhodného chování. Dalším příkladem vlivu sociální desirability v transakčních datech může být selektivní používání mobilních aplikací zaměřených na fitness a zdravou výživu, které svým uživatelům poskytují informace o kalorických hodnotách jídel, množství spotřebované energie při určitých typech pohybu atd. (např. www.kalorickétabulky.cz). Uživatelé těchto aplikací si zde vedou dietní deníky a provozovatelé

aplikace získávají informace, které lze využít pro marketingové účely. Na data z těchto aplikací však nelze plně spoléhat, neboť uživatelé nemusí být v reportování o své dietě zcela upřímní.

Online vs. offline život

Někteří autoři upozorňují, že uživatelé se na sociálních sítích chovají jinak než v offline životě [Boyd, Crawford 2012; Golder, Macy 2014]. V prostředí, kde se většinou komunikuje psaným textem a kde existuje jistá míra anonymity, se u lidí mohou projevit vlastnosti, které se v offline prostředí neuplatňují, např. agresivita. Podle Goldera a Macyho [2014] si však uživatelé do virtuálního prostředí přenášejí především své skutečné charakterové vlastnosti a z chování na síti se dá o psychologických a sociálních vlastnostech jednotlivce usuzovat leccos. Např. člověk, který vyhledává ve svém životě úspěch a obdiv ostatních, se tak bude chovat i na sociální síti.

Sociální sítě poskytují relativně velkou kontrolu nad sebeprezentací, čímž lidem umožňuji budovat si image způsobem, který byl před vznikem internetu nemožný [Pospíšilová 2016]. Sociální média mají oproti reálnému světu tu výhodu, že na pomyslné goffmanovské jeviště sociálního jednání se většinou může dostávat jen to, co o sobě sám uživatel zveřejní (pomineme-li extrémní případy, kdy uživatelé zveřejňují citlivý obsah týkající se jiných uživatelů). Kontroly nad zveřejňovaným obsahem využívá v menší či větší míře každý uživatel sociálních sítí a buduje tak vlastní obraz, který o sobě chce šířit na veřejnosti [Couper 2013; Manovich 2011]. V tom uživatelům pomáhají i funkce sociálních sítí, které umožňují skrýt před veřejností obsah, který vznikl v emotivní chvíli či v důsledku snížené pozornosti. V jazyce výběrových šetření se jedná o jakýsi haló efekt, kdy si jsou jednotlivci vědomi toho, že jsou sledováni, a přizpůsobují tomu své chování; nezveřejňují fotografie, kde jim to nesluší, a vědomě sdílejí jen informace, o kterých se domnívají, že je v očích těch, na kterých jim záleží, nepoškodí.

Chování lidí na síti dále ovlivňují uživatelé, kteří jsou na síti a často i v offline světě silnými hráči. Těmito silnými hráči jsou facebokové či instagramové účty firem a značek, které ovlivňují uživatele přímo prostřednictvím svých účtů nebo zprostředkovaně přes celebrity a bloggery, kteří dané značky propagují. Uživatelé jsou prostřednictvím účtů vyzýváni, aby ke svým fotografiím přidávali konkrétní hashtagy či se fotili s daným produktem, např. za účelem účasti v soutěži. Často jsou také nabádáni, aby komentovali určitý obsah, a je jim sděleno, o čem konkrétně mají v komentáři psát; mohou být rovněž požádáni, aby lajkovali určitý obsah či ho sdíleli na svých profilech. Některé velké značky jako Apple nebo Starbucks se již dostaly do tak silné pozice, že uživatelé používají jejich produkty jako statusový symbol a sdílejí fotografie těchto produktů na sociálních sítích sami od sebe, aniž by je k tomu musel kdokoliv vyzývat.

Kromě výše uvedených specialit online světa, které je nutné mít na paměti, když používáme data ze sociálních sítí k výzkumu, je potřeba vzít v úvahu i fakt, že analýza dat ze sociálních sítí může přinášet zkreslený obraz sociálních vztahů mezi lidmi. V prostředí sociálních sítí může docházet ke vzniku vzájemných, leckdy silných vazeb mezi lidmi, kteří se ve skutečném světě nikdy nepotkali. Ze silné vazby identifikované na sociální síti tak nelze se stoprocentní jistotou usuzovat na silnou vazbu v reálném světě, ze které vyplývají sociálním vědám známé benefity,

jako je např. pomoc při získání zaměstnání. Nevalidní informace o sociálních vazbách se však netýkají jen sociálních médií, matoucí mohou být i údaje z transakčních dat. Např. GPS data z mobilních telefonů lze použít k rekonstrukci vzájemné prostorové blízkosti s ostatními lidmi, nemělo by z nich ale být automaticky usuzováno na kvalitu a sílu vztahu mezi lidmi, kteří spolu tráví hodně času. Chybu, která může plynout z neuváženého používání GPS dat, lze ilustrovat na příkladu: GPS zařízení zaznamená, že majitelka sledovacího zařízení tráví dlouhodobě devět hodin denně s kolegy, se kterými sedí v kanceláři, a jen dvě hodiny týdně se svou matkou. Pokud bychom usuzovali na sílu mezilidských vztahů z času, který spolu lidé tráví, dopustili bychom se v tomto případě chyby. Žena sice s kolegy tráví mnohem více času, její skutečná vazba k nim je však výrazně slabší než vazba k vlastní matce [Ruppert, Law, Savage 2013]. Ruppert a kolegové [2013] upozorňují, že než začneme používat nové technologie, jako je sběr velkých dat prostřednictvím chytrých telefonů, je nutné analyzovat a reflektovat novou sociální situaci samu o sobě a „konceptuálně uchopit specifické vlastnosti digitálních přístrojů a dat, která generují".

Závěr: Budoucnost velkých dat, internetu a výběrových šetření

V současné době se zdá, že se budoucnost sociálněvědního výzkumu bez velkých dat a dalších nástrojů využívajících internet neobejde. Velká data lze již dnes považovat za potenciálně převratný fenomén, který může do výzkumu přinést mnoho inovací. Někteří nevítají velká data s optimismem a domnívají se, že obrovské nestrukturované datové soubory nemají pro sociální vědy velké využití a že obsahují především šum. Analýza těchto dat bude údajně přinášet především nevalidní výsledky [Barnes 2013]. Pro část výzkumníků však velká data představují výzvu a vzácný zdroj informací o lidských společnostech, který může sociálním vědám pomoci dále se vyvíjet. Americký metodolog Robert Groves vidí zapojení velkých dat do výzkumu jako nevyhnutelné; podle něj bude v následujících deseti letech potřeba začít propojovat data z výběrových šetření s velkými daty a přizpůsobit této nové praxi designy výběrových šetření [Habermann, Kennedy, Lahiri 2016].

Kromě odchylek, chyb, nepřesností a zkreslení ve velkých datech, které jsem představila v tomto textu, existují další chyby, které jsou důsledkem procesu získávání dat ze zdroje, editování a spojování s dalšími datovými zdroji [Japec at al. 2015]. Velké množství odchylek a chyb ve velkých datech zmiňuje jako vysoce problematický aspekt nových dat Robert Groves, podle kterého se však s většinou těchto chyb bude možné vypořádat prostřednictvím již známých modelů pro odhadování chyb měření v datech. Z hlediska nových příležitostí, které umožňují velká data, oceňuje dlouhodobou snahu metodologů v oblasti výzkumu chyb měření v datech z šetření, jež přinesla řadu postupů, které bude možné využít právě v analýze validity velkých dat [Habermann, Kennedy, Lahiri 2016]. Publikace Japecové a kolegů [2015] poskytuje základní rámec pro identifikaci chyb ve velkých datech, inspirovaný Grovesovým konceptem celkové chyby šetření (Total Survey Error).

Velká data mohou být navzdory všem chybám a nedostatkům pro kvantitativní sociálněvědní výzkum v mnohém přínosná. V první řadě mohou sociálním vědám umožnit získávat druh informací, ke kterým vědci dosud neměli přístup nebo je mohli získávat jen s vysokými náklady a s velkým množstvím chyb [Savage, Burrows 2007]. McFarland, Lewis a Goldberg [2016] zdůrazňují, že ideální vytěžení více zdrojů velkých dat umožní v budoucnosti získávat informace na úrovni sociálních systémů, takže bude možné zjišťovat v dříve nemyslitelném rozsahu, jak se vzájemně ovlivňují a proplétají ekonomická, sociální a kulturní sféra společnosti. Tyto nové znalosti podle nich způsobí významný posun v sociologické, politologické a další teorii a zasadí se o rozvoj společenských věd; umožní rovněž zefektivnění veřejných politik. Podle Goldera a Macyho [2014] najdou data ze sociálních médií a transakční data, jako jsou záznamy z telefonních rozhovorů či e-maily, velké využití ve výzkumu sociálních sítí a interakcí a pomohou významně přispět k oživení těchto oblastí výzkumu. Velká data rovněž nepochybně rozšíří agendu metodologie kvantitativních sociálních věd, neboť bude potřeba najít řešení pro spojování různých druhů dat a identifikaci různých chyb v těchto datech.

Dalším potenciálním přínosem velkých dat v sociálních vědách může být jejich vliv na změnu organizace vědecké práce a větší důraz na týmovou spolupráci. Tradičně existuje v sociologii a v příbuzných oborech poměrně malá dělba práce. Je běžné, že jeden výzkumník či výzkumnice obstará veškeré úkoly v procesu zpracování jedné studie od stanovení výzkumných cílů přes rešerši literatury a statistickou analýzu až po grafickou úpravu obrázků v textu publikovaného článku. Malá či žádná dělba práce se odráží v množství autorů publikovaných studií; autorství v sociálních vědách se většinou připisuje jednomu či dvěma autorům, vícečetné autorství je výjimkou [McFarland, Lewis, Goldberg 2016]. Aby se velká data mohla stát běžným datovým zdrojem sociálních věd, bude nutné stavět výzkumné týmy, jejichž členové budou zodpovědní za různé druhy analýz. Již teď jsou nároky na schopnosti úspěšných výzkumníků poměrně vysoké – ideální kvantitativní sociolog či socioložka by měl/a mít nadprůměrné nadání v humanitní, ale i technické oblasti – a přibude-li mezi vyžadované schopnosti ještě zvládnutí pokročilého programování, nebude lehké takové superženy a supermuže pro sociálněvědní akademický výzkum získat. Pro úspěšné zvládnutí velkých dat v sociálních vědách proto bude zcela nezbytná týmová spolupráce lidí s různými specializacemi. Měly by začít vznikat heterogenní týmy, ve kterých budou kromě sociologů, politoložek, metodologů a datových analyteček i datoví odborníci, zodpovědní za programování databází a analýzu velkých dat. Na tuto poptávku by měly zareagovat i vzdělávací instituce; vysokoškolské fakulty sociálních a humanitních věd by mohly otevřít studijní specializace zaměřené na zpracování a analýzu nových druhů digitálních dat, které budou školit odborníky vzdělané v základech sociálních věd i práci s velkými daty. Bez týmové spolupráce mezi různými specializacemi a školení datových specialistů znalých sociálních věd podle mého názoru k rozšíření velkých dat v sociálních vědách dojde jen těžko.

Třetím možným pozitivem rozšíření analýzy velkých dat by mohlo být osvěžení přezkoumané populace, unavené častými žádostmi o vyplnění dotazníku či účast ve výzkumném rozhovoru. Pokud by soukromé firmy a agentury pro výzkum trhu a veřejného mínění začaly ve větší míře využívat velká data, snížila by se poptávka po respondentech. Pro zjišťování

spotřebitelského chování jsou velká data stejně či možná více vhodná než data z šetření, leckdy mohou být i levnější [Japec et al. 2015], takže pokud k nim budou mít agentury a firmy přístup, sníží se množství realizovaných výběrových šetření. Lidé by dostávali méně nabídek k účasti ve výzkumném rozhovoru, což by se mohlo pozitivně promítnout do návratnosti akademických výběrových šetření, a tím do kvality dat ze šetření. Pokud nás v budoucnosti čeká tento scénář, nebudou nakonec velká data něčím, co výběrová šetření zničí, ale naopak fenoménem, který zvýší jejich kvalitu a důvěryhodnost.

Velmi těžko se odhaduje, jak bude vypadat sociálněvědní kvantitativní výzkum v budoucnosti, ve které zřejmě bude hrát technologie stejnou nebo ještě větší roli než v současnosti. Kvantitativní vědci a vědkyně dosud ještě nevědí, jak kombinovat tradiční zdroje dat s velkými daty, a nemají ani jasnou představu o tom, k čemu a jak velká data využívat. Výhled navíc zastírá fakt, že internet se velice rychle mění; mění se nejsilnější internetové firmy (kdo si dnes vzpomene na vyhledávač Yahoo či síť MySpace?), mění se rovněž prostředí a pravidla sociálních sítí, jako je Facebook, a mění se populace, která využívá nové technologie. Toto dynamické prostředí má za následek, že každé pojednání o velkých datech rychle zestárne. V tuto chvíli lze s určitostí říct jen to, že v současné době nemohou velká data v sociálněvědním výzkumu nahradit výběrová šetření, neboť jejich povaha neumožňuje zodpovídat podstatnou část výzkumných otázek. Sociální vědy by si tedy měly výběrová šetření ještě ponechat.

LITERATURA:

Anderson, Chris. 2008. „The end of theory: The data deluge makes the scientific method obsolete." Wired Magazine.

Ansolabehere, Stephen, Eitan Hersh. 2012. „Validation: What big data reveal about survey misreporting and the real electorate." *Political Analysis* 20 (4): 437–459. http://dx.doi.org/10.1093/pan/mps023

Babbie, Earl R. 2015. *The practice of social research*. Boston: Nelson Education,

Barnes, Trevor J. 2013. „Big data, little history." *Dialogues in Human Geography* 3 (3): 297–302.

Burrows, Roger, Mike Savage. 2014. „After the crisis? Big Data and the methodological challenges of empirical sociology." *Big Data & Society* 1 (1). http://dx.doi.org/10.1177/2053951714540280

Bollen, Johan, Huina Mao, Xiao-Jun Zeng. 2011. „Twitter mood predicts the stock market." *Journal of Computational Science* 2 (1): 1–8. http://dx.doi.org/10.1016/j.jocs.2010.12.007

Bollen, Kenneth A. 2009. *Structural equations with latent variables*. Hoboken: Wiley.

Bessi, Alessandro, Emilio Ferrara. 2016. „Social bots distort the 2016 US Presidential election online discussion." *First Monday*, *21* (11).

Boyd, Danah, Kate Crawford. 2012. „Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.“ *Information, communication & society* 15, no. 5 (2012): 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam DI Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. „A 61-million-person experiment in social influence and political mobilization.“ *Nature* 489 (7415): 295–298. http://dx.doi.org/10.1038/nature11421

Carr, David. 2013. „Giving Viewers What They Want.“ The New York Times. online: http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html

Chandler, David. 2015. „A world without causation: Big data and the coming of age of posthumanism.“ *Millennium-Journal of International Studies*, *43* (3): 833–851. http://dx.doi.org/10.1177/0305829815576817

Couper, Mick P. 2013. „Is the sky falling? New technology, changing media, and the future of surveys.“ *Survey Research Methods* 7 (3): 145–156.

Crompton, Rosemary. 2008. „Forty years of sociology: Some comments.“ *Sociology*, 42 (6): 1218–1227. http://dx.doi.org/10.1177/0038038508096942

Daas, Piet, Marko Roos, Mark van de Ven, Joyce Neroni. 2012. *Twitter as a potential source for official statistics in the Netherlands*. CBS Discussion Paper.

Dočekal, Daniel. 2015. Velký pohled Facebooku na české uživatele: co mají nejvíce v oblibě? Lupa online: http://www.lupa.cz/clanky/velky-pohled-facebooku-na-ceske-uzivatele-co-maji-nejvice-v-oblibe/

de Madariaga, Inés Sanchéz. 2012. *Structural change in research institutions: Enhancing excellence, gender equality and efficiency in research and innovation. Report of the Expert Group on Structural Change. European Commission.*online: https://ec.europa.eu/research/science-society/document_library/pdf_06/structural-changes-final-report_en.pdf

Driscoll, Kevin, Shawn Walker. 2014. „Big data, big questions. Working within a black box: Transparency in the collection and production of big twitter data.“ *International Journal of Communication* 8 (2014): 20.

European Social Survey 2017. webová stránka http://www.europeansocialsurvey.org/

Gobo, Giamietro. 2011. „Back to Likert. Towards the conversational survey“ in *Innovation in Social Research methods* edited by Malcolm Williams, W. Paul Vogt. Sage.

Golder, Scott A, Michael W. Macy. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*: 129–152. http://dx.doi.org/10.1146/annurev-soc-071913-043145

Grimmer, Justin. 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48* (1): 80–83. http://dx.doi.org/10.1017/S1049096514001784

Grifin, Andrew. 2014. Facebook manipulated users' moods in secret experiment. Independent. Online: http://www.independent.co.uk/life-style/gadgets-and-tech/facebook-manipulated-users-moods-in-secret-experiment-9571004.html

Habermann, Hermann, Courtney Kennedy, Partha Lahiri. 2016. „A Conversation with Robert Groves". *Statistical Science* 31(4): 128-137. http://dx.doi.org/10.1214/16-STS594

Hanzlová, Radka. 2018. „Tisková zpráva: Důvěra k vybraným institucím veřejného života – březen 2018". Centrum pro výzkum veřejného mínění. https://cvvm.soc.cas.cz/cz/tiskove-zpravy/politicke/politicke-ostatni/4584-duvera-k-vybranym-institucim-verejneho-zivota-brezen-2018

Harding, Sandra. 2010. „Feminism, science and the anti-Enlightenment critiques" in Women, Knowledge and Reality: Explorations in Feminist Philosophy, eds A. Garry & M. Pearsall, Unwin Hyman, Boston, MA, pp. 298–320.

Hargittai, Eszter. 2015. „Is bigger always better? Potential biases of big data derived from social network sites." *The ANNALS of the American Academy of Political and Social Science*, *659 (1):* 63–76. http://dx.doi.org/10.1177/0002716215570866

Hrdina, Matouš. 2016. „Identity, Activism and Hatred: Hate Speech against Migrants on Facebook in the Czech Republic in 2015." *Naše společnost* 14 (1): 38–47. http://dx.doi.org/10.13060/1214438X.2016.1.14.265

Hox, Joop, J. K. Roberts (eds.). 2011. Handbook of advanced multilevel analysis. London: Psychology Press.

Japec, Lilli, et al. 2015. "Big data in survey research: AAPOR task force report." Public Opinion Quarterly 79 (4): 839–880. https://doi.org/10.1093/poq/nfv039

Jutz, R., E. Scholz. 2016. International Social Survey Programme: ISSP 2014 - Citizenship II. GESIS

Kalyvas, James, R., Michael R. Overly. 2015. Big data business and legal guide. CRC Press

Karaščáková, Zuzana. 2013. Prvá priama voľba prezidenta v Českej republike na Twitteri. Naše společnost 2 (11): 41–52.

Kitchin, Rob. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1):1-12. http://dx.doi.org/10.1177/2053951714528481

Kosinski, Michal, David Stillwell, Thore Graepel. 2013. „Private traits and attributes are predictable from digital records of human behavior." *Proceedings of the National Academy of Sciences* 110 (15): 5802–5805. http://dx.doi.org/10.1073/pnas.1218772110

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2016. „How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American Political Science Review* 111 (3): 484–501. http://dx.doi.org/10.1017/S0003055417000144

Lazer, David, Ryan Kennedy, Gary King, Alessandro Vespignani. 2014. „The parable of Google Flu: traps in big data analysis." *Science* 343 (6176): 1203–1205. http://dx.doi.org/10.1126/science.1248506

Manovich, Lev. 2011. „Trending: The promises and the challenges of big social data.“ *Debates in the digital humanities* 2: 460–475. http://dx.doi.org/10.5749/minnesota/9780816677948.003.0047

Massey, Douglas S, Roger Tourangeau. 2013. „Where do we go from here? Nonresponse and social measurement.“ *The ANNALS of the American Academy of Political and Social Science 645 (1): 222*–236.

McFarland, Daniel A., Kevin Lewis, Amir Goldberg. 2016. „Sociology in the era of big data: The ascent of forensic social science.“ *The American Sociologist* 47 (1): 12–35. http://dx.doi.org/10.1007/s12108-015-9291-8

Murphy, Joe, Elizabeth Dean, Craig A. Hill, Ashley Richards. 2011. „Social media, new technologies, and the future of health survey research“ in *National Center for Health Statistics (Ed.), Proceedings of the 10th conference on health survey research methods (pp. 231*–241).

Murphy, Joe, Craig A. Hill, and Elizabeth Dean. 2013. “Social Media, Sociality, and Survey Research.” in *Social Media, Sociality, and Survey Research*, edited by Craig A. Hill, Elizabeth Dean, and Joe Murphy, 1–34. Hoboken, NJ: John Wiley and Sons.

Murphy, Joe, Michael W. Link, Jennifer H. Childs, Casey L. Tesfaye, Elizabeth Dean, Michael Stern, Paul Harwood. 2014. Social Media in Public Opinion Research Executive Summary of the Aapor Task Force on Emerging Technologies in Public Opinion Research. *Public Opinion Quarterly*, *78*(4), 788–794. http://dx.doi.org/10.1093/poq/nfu053

Pew Research Center. 2017. What Low Response Rates Mean for Telephone Surveys. Online http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys Citováno 11. 9. 2017.

Pospíšilová, Marie. 2016. *Facebooková (ne) závislost: Identita, interakce a uživatelská kariéra na Facebooku*. Praha: Karolinum Press.

Ruppert, Evelyn, John Law, Mike Savage. 2013. „Reassembling social science methods: The challenge of digital devices.“ *Theory, culture & society*, *30* (4): 22–46. http://dx.doi.org/10.1177/0263276413484941

Saris, Willem. E., Gallhofer, Irmtraud N. 2014. *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken: John Wiley & Sons.

Savage, Mike, Roger Burrows. 2007. „The coming crisis of empirical sociology.“ *Sociology* 41 (5): 885-899. http://dx.doi.org/10.1177/0038038507080443

Savage, Mike, Roger Burrows. 2009. „Some further reflections on the coming crisis of empirical sociology.“ *Sociology*, *43* (4): 762–772. http://dx.doi.org/10.1177/0038038509105420

Snijders, Chris, Uwe Matzat, Ulf-Dietrich Reips. 2012. „Big Data: big gaps of knowledge in the field of internet science.“ *International Journal of Internet Science* 7 (1): 1–5.

Tabachnick, Barbara G., Linda S. Fidell, Steven J. Osterlind. 2001. Using multivariate statistics.

Tinati, Ramine, Susan Halford, Leslie Carr, Catherine Pope. 2014. „Big data: methodological challenges and approaches for sociological analysis.“ *Sociology* 48 (4): 663–681. http://dx.doi.org/10.1177/0038038513511561

Thrift, Nigel. 2005. Knowing capitalism. London: Sage.

Vochocová, Lenka, Jaromír Mazák, Václav Štětka. 2016. „Nic pro holky?: Genderové nerovnosti v politické participaci na sociálních sítích.“ *Gender, rovné příležitosti, výzkum (Gender, Equal Opportunities, Research)* 17 (2): 64–75.

Webber, Richard. 2009. Response to The Coming Crisis of Empirical Sociology: An Outline of the Research Potential of Administrative and Transactional Data. *Sociology*, *43* (1): 169–178. http://dx.doi.org/10.1177/0038038508099104

O AUTORECH

▼

Johana Chylíková pracuje v Českém socialněvědním datovém archivu Sociologického ústavu AV ČR. Doktorské studium absolvovala na Fakultě sociálních věd UK. Zabývá se metodologií sociálněvědního výzkumu, zejména chybami měření v datech ze sociálněvědních šetření.
Lze ji kontaktovat na adrese johana.chylikova@soc.cas.cz.