Univerzita Karlova

Filosofická fakulta

Katedra Sociologie

Diplomová práce

Ing. Táňa Lančová

Big data v sociologii

Big data in sociology

2018

Vedoucí práce: Mgr. Petr Lupač, Phd

.

Poděkování

Děkuji vedoucímu práce Mgr. Petru Lupačovi, Ph.D. nejen za vylepšující postřehy odborného i stylistického rázu, ale především za suchý realismus, kterým mě vždy nekompromisně vracel k podstatě.

Za podnětné připomínky děkuji také doc. PhDr. Jiřímu Buriánkovi, CSc. a Mgr. Martinu Betinci, Ph.D., kteří práci velmi ochotně konzultovali a poskytli mi důležitou zpětnou vazbu.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Praha, 11.7.2018

Táňa Lančová

Abstrakt

Cílem práce je poskytnout ucelený pohled na problematiku Big dat v sociologii a reflektovat tak aktuální téma, které v tomto rozsahu doposud nebylo systematicky zpracováno.

Práce shrnuje přístupy k vymezení pojmu Big data, které nejlépe vystihují komplexnost tohoto fenoménu. Popisuje jednotlivé postoje, které sociologie vůči Big datům zaujímá. Identifikuje specifika Big dat a tím i důvody, proč dosud Big data nejsou v sociologii plně etablována. Přináší systematizovaný popis Big dat v členění dle jejich vlastníků a popisuje způsoby, kterými jsou Big data analyzována. Systematizuje a kriticky reflektuje výhrady, které jsou výzkumům využívajícím Big data adresovány, a přináší další.

Klíčová slova: Big data, analýza Big dat, metodologie

Abstract

The aim of this work is to provide a holistic view on Big Data in sociology and with this way to reflect the actual topic, which has not been systematically elaborated yet.

This theses summarizes approaches to Big Data specification, which provides insight into complexity of this phenomenon. It describes attitudes of contemporary sociology of Big Data. It identifies Big data specifics, which lead to reasons, why Big Data have not been fully accepted by sociology yet. It provides comprehensive description of Big Data sources sorted by the owners and brings an overview of methods for Big Data analysis. It sorts and reflects Critical Data Studies and brings new topics.

Key words: Big data, Big data analysis, methodology

Obsah

Seznam zkratek

AI	Artificial Inteligence
API	Application Programming Interface
CDS	Critical Data Studies
CPU	Central Processing Unit
ČVUT	České vysoké učení technické
EAI	Explainable Artificial Intellignce
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
IT (ICT)	Informační technologie (Inforrmační a Komunikační Technologie)
KNT	Kvantitativní
KVT	Kvalitativní
MU	Masarykova Univerzita
NN	Neuronová síť
PR	Public relations
UK	Univerzita Karlova
ÚOOÚ	Úřad na ochranu osobních údajů

1 Úvod

1.1 Fenomén Big dat

„Data sety jsou rozsáhlé, překračují kapacitu hlavní paměti, lokální paměti a dokonce i vzdálených/externích disků. Nazýváme to problém Big dat“ (Cox & Ellswort, 1997, str.1). Termín „Big data“ použili poprvé v technickém článku věnovaném problému vizualizace pracovníci laboratoří NASA Ames Research Center v roce 1997 pro označení souboru velikostí překračující 100 Gbytů, kdy data tehdejšími metodami nemohla být ani zpracována ani vizualizována.

Díky pokroku v oblasti informačních a komunikačních technologií (ICT) v posledních dvaceti letech působí dnes tato velikost úsměvně. Rozvoj ICT ale nejen že umožňuje zpracovávat stále větší množství dat, je i za jejich rostoucí produkci zodpovědný (web2.0[1], internet věcí[2], výkonnější a cenově dostupnější hardware apod.). Zároveň jde o jevy synergicky se podporující, neboť i velké množství dat posléze přináší nové výzvy (analytické, kapacitní apod.), které k rozvoji ICT přispívají. Pro představu o tempu růstu produkce dat lze uvést údaj z analýzy poradenské společnosti McKenzie (Henke, Bughin , Chui, Manyika, Saleh, Wiseman & Sethupathy, 2016), kde se uvádí, že objem celosvětově produkovaných dat se každé tři roky zdvojnásobí (str.22).

Obecné označení Big data jako zastřešující výraz pro data, jež jsou kvůli své velikosti náročné na zpracování, ale přetrvalo. Z oblasti informačních technologií do sociálně vědních oborů a také do komerčního sektoru začaly Big data pronikat teprve okolo let 2010- 2012, zato s nebývalou razancí.

Data Google trends ukazují, že do roku 2012 byla frekvence vyhledávání pojmu „Big data“ stabilně nízká, v roce 2012 přichází náhlý konstantní růst, který trvá stále. Komentář v NYT označuje toto období jako Age of Big data[3], podle Harvard business review je datový vědec nejvíc sexy povoláním dvacátého prvního století (Davenport,

[1] Ustálené označení pro web směřující k obsahu generovaném uživatelem (blogy, osobní stránky, sociální média..), někdy také nazýván web participativní.

[2] Ustálené spojení pro celý ekosystém bezdrátového propojení více identifikovatelných zařízení

[3] Lohr, S. The Age of Big data. Získáno 10.3.2018 z http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Patil, 2012), na on-line vyučovacích platformách EdX a Courserea od té doby stále roste počet kurzů věnovaných metodám Big data analýzy.

A stranou nezůstává ani sociologie. Vyhledávač v Informačních zdrojích Univerzity Karlovy[4] našel od roku 2000 celkem 1 313 publikací s klíčovými slovy „Big data" a „sociology", z toho téměř 78% jich bylo publikováno po roce 2012. Podobně i databáze Scopus dle analýzy di Bella, Leporattiho a Maggina (2016) ukazuje pouhé dva články s klíčovými slovy „Big data" v oblasti „social science" publikované do roku 2009. Dokonce až 97% článků bylo publikovaných po roce 2012. V éře nálepek je současné období i v ryze vědeckých článcích některými autory nazýváno „Era of Big data" (Boyd & Crawford, 2012).

Přesto však dosud nebyla problematika Big dat v sociologii systematicky zpracována. Ambicí této práce je tak poskytnout ucelený pohled na tuto problematiku.

1.2 Struktura a témata práce

První kapitolu je věnována představení, co se vlastně pod poněkud vágním pojmem Big data rozumí.

Analyzování dat bylo vždy doménou sociologie, která začátkem a v polovině minulého století stála buď v čele vývoje nových metod, nebo přinejmenším patřila mezi včasné osvojitele („early adopters"). Byť je nárůst článků označených jako „Big data" a „sociologie" působivý, je tomu skutečně tak i v souvislosti s Big daty? Etablovaly se už Big data v sociologii jako plnohodnotná součást empirického výzkumu a naopak pomáhá sociologie svým silným metodologickým zázemím kultivovat rozvoj analýz Big dat?

Na tyto otázky odpoví následující kapitola, která tématizuje a shrnuje diskuzi, která je okolo využití Big dat v sociologii vedena, a zkoumá zakotvení Big dat v české akademické sociologii.

[4] http://pez.cuni.cz/

Identifikací specifičnosti Big dat oproti ostatním ustáleným formám zisku informací v sociologii se zabývá kapitola následující, která pak také tyto určuje jako možné faktory brzdící rozvoj tohoto směru výzkumu v sociologii.

Další část práce je následně věnována přehledu dat, která spadají pod označení Big data. Data jsou systematizována podle vlastníků, jsou popsány způsoby jejich získání i nejzajímavější studie. Data jsou popisována i optikou jejich využitelnosti pro sociologických výzkum v ČR. Další kapitola pak laicky srozumitelnou formou přibližuje způsoby analýzy.

Rozsah informací z oblasti informačních technologií uváděných v těchto dvou kapitolách je sice v rámci sociologické diplomové práce nezvyklý, ale Big data bez technologického zázemí nedávají smysl. Kvůli různosti metod, technologií a programovacích jazyků tento přehled nemůže sloužit jako detailní uživatelská příručka, jejich účel je získat vhled do problematiky analýzy velkých datových souborů (a osvojit si leckdy náročnou terminologii).

Okouzlení z možností, popsaných v předchozích kapitolách, bude následně zrelativizováno v následující části popisující kritické výhrady vůči studiím založeným na Big datech v upřímné snaze shrnout všechna omezení a nástrahy, kterých si musí být sociologové při případné analýze Big dat vědomi a také poskytnout jisté kritické vodítko ke čtení analýz založených na Big datech. Zároveň tato kapitola může sloužit jako podklad pro hlubší metodologické rozpracování, neboť téměř všem těmto nedostatkům sociologie čelila a čelí i v rámci jiných metod zisku a analýzy dat.

Závěr práce pak je věnován shrnutí problematiky.

2 Big Data

Definice založená pouze na velikosti by v dnešní době neobstála, Big data se stala zastřešujícím pojmem pro široké spektrum dat i činností. Jak jsou tedy Big data vymezována?

2.1 Základní charakteristiky

Big data mohou být charakterizována přítomnosti 3V znaků. Jde o široce sdílenou a nejčastěji uváděnou definici Big Dat, používanou především v oblasti informačních technologií a v komerční sféře, bez určení autorství. Je to však definice **v**elmi **v**ágní a **v**ýplňková a lze se tak setkat s různými interpretacemi, co se pod tzv. V-slovy (v-word) skrývá (v textu naznačeno „/“ ve smyslu a zároveň, ale také nebo případně výhradně).

3V znaky Big dat odkazují k

1. velikosti (volume), tedy objemu zpracovávaných/analyzovaných/generovaných dat, typicky v terabytech[5]

2. rychlosti (velocity) přírůstků/přenosu/zpracování/zveřejňování dat

3. různorodosti typů dat, která jsou ukládána/analyzována/používána (variety), např. textové záznamy, video soubory, obrázky, audio. Případně je tento pojem chápán ryze ve smyslu struktury jako data strukturovaná, nestrukturovaná, semistrukturovaná.

Tato definice, případně její nejrůznější variace jako rozšíření o další V- slova (variability – proměnlivost, veracity případně value– věrohodnost – kvalita dat se může měnit), bývá stále využívaná především v článcích orientovaných na komerční využití, pro systematizaci a vědecké uchopení je však pro svou neurčitost nedostačující.

[5] 1 terabyte = 1 000 000 000 000 bajtů, případně ve staré notaci 1 099 511 627 776 bajtů

Porovnáním klasických souborů dat a Big dat souborů proto navrhl Kitchen (jak je citováno v Kitchen & McArdle, 2014) rozšíření tří hlavních V- charakteristik (velikost, rychlost, různorodost) na celkem sedm důležitých znaků. Doplnil

4.úplnost (exhaustivity; Big data obvykle pokrývají celou populaci využívající systém, ze kterého pochází, nikoli vzorek),

5.rozlišení a indexikalitu (resolution/indexicality, tedy Big data mají typicky pevnou a silnou, tzn. data nejsou agregována a obsahují identifikátory),

6.relationality (relationality, Big data typicky silnou, tzn. data z různých databází nebo pořízena v různý čas jsou porovnatelná díky shodným identifikátorům)

7. rozšířitelnost/škálovatelnost (extensiability/scalability, tedy schopnost rychlé proměny struktury sbíraných dat, Big data mají typicky vysoké).

Jako příklad důležitosti ostatních atributů uvádí například census, který sice velikostně Big datům může odpovídat, ale ostatní charakteristiky nesplňuje – „data jsou sbírána jednou za 10 let, odpovídá se na cca 30 otázek a jakmile je spuštěn, nelze otázky vyladit, přidat nebo odebrat“. (str.2)

Na tuto práci posléze Kitchen navázal spolu s McArdlem (2014) a identifikovali 26 hlavní možných zdrojů Big dat, které byly za Big data označeny v literatuře, v celkem sedmi doménách – mobilní komunikace, webové stránky, sociální média, sensory, kamery, transakční a administrativní data (těchto 26 typů datasetů bohužel není blíže specifikováno, jejich kategorie jsou jen pojmenovány, většinou bez příkladů, proto zde jejich přehled pro možnou desinterpretaci není uveden). Analyzovali vlastnosti všech datasetů v celkem deseti parametrech - velikost, rychlost vzniku, rychlost ukládání/zpracování/publikování, variabilita, úplnost, rozlišení, indexikalita, relationalita, rozšířitelnost a škálovatelnost – s výsledkem, že nelze určit jednoznačný profil Big dat ve všech těchto deseti charakteristikách.

Jako klíčové atributy Big dat proto z těchto deseti parametrů určili pouze dva hlavní - rychlost (ve všech ohledech, tedy rychlost vzniku a ukládání/ zpracování/ publikování) a úplnost, **„small data jsou pomalá a pocházejí ze vzorků, Big**

data jsou rychlá a n=všichni“[6] (str. 8). Velikost bývá často spíše důsledkem rychlosti a populace, nikoli rozhodujícím znakem. Poměrně ostře proto uvádějí, že „3V mem je skutečně falešný a zavádějící […] a je zodpovědný za zmatky ohledně definice hranic Big dat“ (str.9).

Je nutné upřesnit, že n=všichni odkazuje k možnosti získat data „celé populace v rámci systému“ (str. 7), tedy těch, kteří danou technologii používají/jsou v jejím zorném úhlu (jakkoli tato interpretace bude v sekci věnované critical data studies zrelativizována, protože samozřejmě vlivem technických chyb může dojít ke ztrátám záznamů).

Alternativní přístupy k definici Big dat kladou důraz na techniky použité k jejich ukládání/analýze. Jak podotýkají například Symons a Alvaro (2016, str.4), Big data často spíše než na velikost odkazují na soubor použitých statistických metod a výpočetních nástrojů analýzy, které tak tvoří hlavní distanční (rozdílové) charakteristiky. Právě těmto metodám se budeme věnovat v kapitole věnované metodám analýzy.

2.2 Rozkročenější přístup

Povědomí o Big datech je ale především v sociálních vědách poněkud volnější a umožňuje široký výklad. Například Burrows a Savage (2014) za jistý druh Big dat považují i on-line dotazník umístěný na stránkách BBC (str.3). A to díky inovativnosti (respondenti odpovídali interaktivně a po vyplnění jim byl odměnou grafický souhrn výsledku), neexistenci horního limitu počtu respondentů a nemožnosti kontrolovat jejich složení.[7]

Následující příklady pak nejvýstižněji ilustrují volné uchopení Big dat v sociálních vědách.

[6] Tučné zvýraznění není součástí citace, je použito pro zdůraznění významu definice, která dle mého názoru po pečlivém čtení materiálů věnujícím se problematice Big dat nejlépe vystihuje jejich podstatu.

[7] Pokud by byl sledován i počet návštěvníků stránek, kteří odešli bez vyplnění dotazníků, pak lze považovat za splněnou i podmínku n=všichni v rámci systému a on-line dotazník tak vyhovuje i definici Kitchena.

Rozšířený je přístup, kdy jsou za Big data pokládaná i data získaná tzv. "jinými prostředky" bez detailnější specifikace. „Big data jsou novým stylem dat" obcházejí problematiku přesnější definice McFarland et al. (2016, s.6).

Podobně spíše filosofickým příspěvkem k debatě o Big datech je i vymezení tohoto pojmu pomocí třinácti tzv. p-words tak, jak je rozpracovává na svém blogu věnovaném sociologii Lupton[8] (2015). Výhodou dle autorky je, že oproti v-words poukazují p-words na „sociokulturní rozměry", které sebou Big data přinášejí. Seznam p-words tak zahrnuje například charakteristiky jako „playful" (proces generování dat má v sobě lucidní charakter, stejně tak hravostí vynikají i datoví vědci při kreativních vizualizacích), „political" (Big data v sobě zahrnují otázky možnosti přístupu, mocenských výhod plynoucích z vlastnictví apod.), „personal" (Big data zahrnují detailní informace o chování, emocích, vztazích) atd. Velikost a rychlost z původní 3V definice jsou pak pouze skryty ve vlastnosti „perverse", odkazující k „dech beroucím příležitostem", ale i pocitu ohrožení ze ztráty kontroly nad neustálými přírůstky dat. Různorodost (poslední charakteristika 3V) je pak nahrazena synonymem „polymorphous".

Boyd a Crawford preferují také menší důraz na velikost dat, když uvádějí, že „Big data jsou méně o datech, která jsou velká, jako spíše o schopnosti hledat, agregovat a odkazovat (cross-reference) na velké datové soubory". Big data definují jako „kulturní, technologický a učený (scholar) fenomén, který je založený na vzájemné souhře technologie, analýzy a pověr." (2012, str. 2)

[8] Autorka je profesorkou sociologie věnující se problematice digitálních technologií.

3 Sociologie v éře Big dat

V literatuře lze vystopovat tři zřetelné tendence vymezení se sociologie vůči tomuto novému směru výzkumu. První můžeme nazvat rezervovaný, zahrnující široké spektrum projevů od ignorování nabízených možností až po kriticky odmítavé zhodnocení přínosu. Druhým přístupem je potom akceptace upozadění sociologie a přijmutí její role pouhého přispěvatele v rámci tzv. computer social science. A nakonec postoj třetí, který sice reflektuje současnou upozaděnost sociologie a její ostražitost vůči Big datům, ale chápe ji jen jako přechodnou a horuje pro vůdčí roli sociologie v rámci tohoto typu výzkumu.

3.1 Sociologie rezervovaná

Rezervovaný postoj vystihli už v roce 2007 M. Savage a R.Burrows (2007) v provokativně nazvaném článku „Přicházející krize empirické sociologie“, kde mimo jiné upozorňovali, že klasické ustálené sociologické metody sběru dat již nejsou dostačující pro výzkum společnosti, kterou označil jako znalostní kapitalismus. Sociologie, která ignoruje nabízející se velké množství takzvaných „transakčních dat“, vzniklých jako vedlejší produkt běžné digitální činnosti, ztrácí své výsadní postavení v sociálně vědním výzkumu. Článek se dle autorů (odkaz na statistiku uváděný v článku již bohužel není funkční, nicméně autoři sami v roce 2014 uvádějí téměř 300 citací na Google Scholar[9]; aktuální počet v květnu 2018 dosahuje 842) stal jedním z nejcitovanějších článků periodika Sociologie (Burrows & Savage 2014) a byť v něm pojem Big data ještě explicitně nezazněl, svým pojetím k němu směřoval. Článek je svým vyzněním burcující a pasivní postoj kritizuje.

Osborne, Rose a Savage (2008, str.532) uvádějí, že „profesionální sociologové nejsou jediní, kdo analyzují….[]…analytici soukromých firem, žurnalisté, dokumentaristé občas poskytují lepší sociologii než sociologové samotní“. M.Savage a R.Burrows (2014) k tomu přidávají v souvislosti s nárůstem fenoménu Big dat i „data scientist“, datové vědce, “kteří jsou schopni popsat sociální svět způsobem dosud nemožným“ (str.3).

[9] https://scholar.google.cz/

Naproti tomu Goldthorpe z opačné pozice roce 2016 za sociologii sumarizuje - jakkoli mohou být Big data užitečná pro znalostní kapitalismus, jejich „hodnota pro sociální vědy zůstává stále otevřenou otázkou“ (jak je citováno v Halford & Savage, 2017, str. 1132). Přehled výtek k metodám analýzy Big dat (zaznamenávají „jenom některé aktivity konkrétních lidí, kteří používají určitá zařízení a aplikace určené k záznamu specifických informací“; stejně tak kritizují „závislost analýzy Big dat na výpočetních metodách, zejména spekulativních dolování dat pomocí variant rozpoznávání a korelace“ atd.) shrnují Halford a Savage jednoznačně – analýza Big dat „nemá zásadní vliv na hlavní sociologické debaty a silný skepticismus zůstává“ (2017).

3.2 Sociologie jako součást computational social science

Paralelně s rezervovaným postojem probíhá rozvoj oboru nazývaného „computational social science“, instrumentálního a interdisciplinárního přístupu založeného na informačních technologiích, kde „sociální a behaviorální vědci, kognitivní vědci, teoretici AB (agent-based), počítačoví vědci, matematici a fyzikové kooperují bok po boku, aby přišli s inovativním a teorií podloženým modelem vybraných fenoménů. [tito vědci] silně věří, že nastala nová éra porozumění a funkce společnosti na různých úrovních“ (Lazer et al. citováno v Conte at al., 2012, str. 327).

Například Univerzita v Cambridge založila institut „Cambridge Big data“, který „spojuje dohromady výzkumníky všech šesti fakult univerzity, aby přijali výzvy nepředstavitelného počtu přístupných dat“. (Cambridge Big Data, n.d.)

Podobně interdisciplinaritu a „computational social science“ mezi jinými vyzdvihují i Chang et al. (2014), kdy explicitně dodávají, že „někdo, kdo má expertní znalosti sociologie a psychologie nemusí tušit, jaké možnosti poznání přináší strojové učení nad těmi samými daty“. (str.6)

Nadšení pro „computational social science“ lehce krotí Kitchen (2014), který obviňuje tyto vědce (stejně jako tzv. digital humanities vědce, tedy vědce zabývající se digitální stopu člověka) z post-pozitivismu a shrnuje své výhrady - „jedna věc je najít vzorec, druhá je najít vysvětlení. To vyžaduje sociální teorii a hluboké znalosti kontextu. Vzorec tak není zakončením, ale spíše začátkem nové analýzy“. (str.8) (pozn. ve smyslu měl by být). Je nutno i přidat lehkou korekci, neboť některé typy

analýz nepřinášejí ani ten vzorec, pouze výsledky (viz. black-box řešení, které představíme dále).

3.3 Sociologie jako hybná síla

Obavu z dočasného ústupu sociologie sice vyjadřují i McFarland et al. (2016), kteří se ptají, zda nyní v souvislosti s Big daty čelíme pouze změně paradigmatu nebo „kolonizaci, kdy některé oblasti sociologie a jejich tradice budou rozvráceny do „computer science" a tím skončeny" (str.13). Dle autorů žijeme v přelomové éře, kdy data a metody jejich analýzy neúměrně rostou a díky tomu zažíváme odklon od klasických výzkumných metod a paradigmat minulého století jako je výběrové šetření a metodologický individualismus. Kvůli náročnosti a novosti problematiky zatím spíše převažuje tzv. inženýrský způsob popisu sociálních faktů, tedy jen popis problematiky jak a proč funguje s důrazem na predikci, ale bez hlubšího objasnění, které je doménou sociálních věd. V delším časovém horizontu je to ale šance pro vznik tzv. „forenzní sociální vědy", vědy, která by kombinovala aplikovaný a teoretický přístup, „využívala teorii k deduktivnímu zkoumání dat a zároveň indukci k nalezení teorie, která vyzkoumané závěry vysvětlí nejlépe" (str. 30).

Poněkud poetičtější název pro nový typ vědy zvolili Halford a Savage (2017) – „symphonic social science". Zárodky takové vědy už v sociologii nyní existují, dle autorů jde například o knihu Bowling alone od R.Putnama[10]. „Symphonic social science" stojí na synergii dat, metod a teorie[11]. S Big daty má několik styčných bodů, především důraz na znovuvyužitelnost nalezených dat, korelace a vizualice, ale liší se důrazem na teorii, pečlivostí sběru dat, zaměřením na dlouhodobé trendy, odlišným přístupem ke korelacím a kauzalitám a umnějším využitím vizualizačních nástrojů. Právě implementace „symphonic social science" přístupu do klasické analýzy Big dat ji udělá důvěryhodnější a „sociologii dovolí prozkoumat možnosti Big dat analýzy pro sociologický průzkum bez kompromisu k závazkům kritického postoje k datům, metodologické rigoróznosti a teoretické interpretace". (str.1145) Podle autorů by

[10] Odlišný názor na knihu nicméně sumarizují Sedláčková a Šafr (2008) – „ve vědecké komunitě [*je*] Putnamova komunitaristická koncepce kritizována, vedle zjednodušené metodologie měření, zejména pro svou normativní zatíženost" (str.331)

sociologie přes všechny výhrady k Big datům měla v budoucnu „hrát centrální roli v rozvoji analýzy Big dat“ (str.1134).

Všechny postoje se tak shodují v jediném, sociologie dosud Big data plně nepřijala.

3.4 Akademická sociologie v ČR a Big data

Vztah české sociologie a Big dat bude zkoumán v kontextu Sociologického ústavu AV ČR (SOÚ AV), který jako „ústřední […] veřejná výzkumná organizace v oblasti sociologie“ (Nešpor, 2017) deklaruje „rozvíjení metodologie soudobého sociologického výzkumu“ (O ústavu, n.d.). A také v rámci vysokých škol, které vzděláváním budoucích specialistů oboru pomáhají formovat jeho podobu.

V přehledu projektů řešených v rámci Sociologického ústavu Akademie věd ČR (2018) nelze najít jediný projekt věnovaný problematice Big dat.[12] Na stránkách Českého sociálně vědního archivu (tedy části SOÚ) nalezneme ve výčtu zdrojů sociálněvědních dat jen na posledním, osmém místě uvedeny „specifické datové zdroje (např. administrativní data, internet)“ (Zdroje dat, n.d.). Dále se v rámci stránek také uvádí, že „společnost obecně prochází překotnou digitalizací a výsledky tohoto procesu jsou často využitelné jako zdroje dat pro účely výzkumu“. (Management dat, n.d.)[13] Big data (případně jejich jednotlivé reprezentace) detailněji rozpracovány nejsou ani v knize ústavu Cesty k datům (Krejčí & Leontyieva, 2012).

Explicitní zmínku o Big datech lze nalézt v rámci publikací Sociologického ústavu jen obtížně. Problematika není řešena v Sociologickém časopise, pouze v souvislosti

[12] Jen velmi málo pak Big data (bez explicitního označení) i využívá (například v rámci projektu zkoumajícího výběrového párování jsou využita „data ze seznamovacích serverů“, případně se využívají tzv. administrativní data státu). Několikrát ale oproti tomu jsou jako metody zisku dat výslovně uvedeny „panelový výzkum“, „hloubkové, ohniskové rozhovory“, „dotazníkové šetření“ apod.

[13] Podobně například pod heslem „techniky sběru informací“ v Sociologické encyklopedii dostupné na webových stránkách SOÚ AV ČR lze nalézt mezi hlavními technikami zisku sociologických informací přímé pouze pozorování, dotazník, rozhovor a analýzu dokumentů. Analýzu Big dat bez bližší specifikace si snad lze představit až v členění podle interakce s výzkumným objektem, kdy lze metody rozlišit takto - přímé pozorování, dotazování a sekundární analýza, tedy „[technika] postavená na použití a reinterpretaci zdrojů informací určených k jinému než danému výzkumnému účelu“ (Vodáková, 2017). Pojem sekundární analýza pak upřesňuje, že jsou „většinou v sociologii využívány prameny statistických dat nebo výsledky empirických výzkumů realizovaných dříve“, případně ji lze chápat jako „formu odhalování dalších latentních informací v již exploatovaném materiálu“. Jisté vodítko pak nabízí „snaha specifikovat sekundární analýzu jako techniku spojovanou s použitím náročnějších postupů, například vícerozměrné analýzy rozptylu“. (Buriánek, 2017). Jde však o zdigitalizovanou verzi z roku 1996, proto je ponecháno jen formou poznámky.

s GDPR[14] zmiňuje odborné periodikum Naše společnost, že „nové typy dat není možné redukovat pouze na data sociálních médii, ale je třeba vzít v potaz všechna big data“ (Dobrovolný & Kudrnáčová, 2017, str. 56).

Podobně zatím pouze omezeně se dle veřejně dostupných studijních plánů a anotací Big datům věnují vysoké školy ve výuce. Dle zjištění z veřejně přístupných webů jen Fakulta sociálních věd UK nabízí základy jazyka Python s velmi jemnými základy vytěžování dat z webu. Přehled studijních programů nicméně nemusí být plně vypovídající, leckde chybí detailní sylaby, Big data mohou být teoreticky přednášena v rámci obecného metodologického předmětu.

Osloveno bylo proto sedm kateder sociologie zajišťujících výuku sociologie jako oboru s prosbou o vyjádření formou odpovědí na čtyři krátké otázky týkající se výuky Big dat (a informačních technologií obecně)[15]. Je možná trochu symbolické, že ze sedmi odpověděly pouze tři, a to z prvních tří příček dle přehledu nejlepších fakult pro obor v žebříčku HN[16] (jakkoli metodika určení pořadí v průzkumu HN je silně znevěrohodňována[17]). Big data katedry jako směr v sociologii sice vnímají, systematicky zařazena do výuky ale zatím nejsou.

---

[14] Nařízení EU o ochraně osobních údajů

[15] Otázky ve znění:

- nabízíte nějaké dobrovolné/povinné kurzy věnované IT dovednostem (například výuka Pythonu, C++, MySQL, R,...) a pokud ano, můžete prosím specifikovat jaké?
- pokud tyto kurzy nabízíte, zaměřujete se v nich i na analýzu Big dat a možnosti jejich získání?
- pokud tyto kurzy nenabízíte, uvažujete výhledově o jejich vypsání? Pokud ne, můžete prosím stručně vysvětlit svoje stanovisko?
- podporujete analýzu Big dat i jinou formou než dedikovaným kurzem IT dovedností? Pokud ano, můžete to prosím specifikovat?(např. probíráte v rámci metod; máte odborníka, který je ochotný studentům poskytnout konzultace; spolupracujete s jinou katedrou apd.)

byly adresovány zástupcům kateder sociologie FF UK, FSV UK, FF ZCU, FSS MUNI, FF UPOL, FF UHK a FF OU spolu s vysvětlením, že jde o podklady do diplomové práce věnované Big datům (spolu se stručným nástinem témat této práce).

[16] Keményová, Z. Žebříček českých vysokých škol: Brno výrazně porazilo Univerzitu Karlovu. Získáno 6.6.2018 z https://archiv.ihned.cz/c1-63417720-nejlepsi-vysoke-skoly-v-cesku-special-hn

[17] Vinopal, J. V žebříčku Hospodářských novin chybí Katedra sociologie Filozofické fakulty UK. Získáno 6.6.2018 z https://ksoc.ff.cuni.cz/2016/02/01/v-zebricku-hospodarskych-novin-chybi-katedra-sociologie-filozoficke-fakulty-uk/

Fakulta sociálních studií MUNI problematiku Big dat ve svých kurzech zmiňuje, „ ale spíše se jedná upozorňování na tuto problematiku než o její systematickou výuku“. O přímém kurzu by uvažovali, ale nemají na něj odborníka, „této problematice se zatím nikdo na pracovišti nevěnuje, i když knihy o této problematice mapujeme a snažíme se je kupovat do knihovny“. (T. Katrňák, email ze 30.5.2018)

Podobně domovská Filosofická fakulta UK uznává, že „jeden povinně-volitelný kurz na toto téma by jistě odbyt, uplatnění a smysl našel….[]…naskytne-li se personální příležitost, rádi bychom téma do výuky zavedli“. Zatím „systematicky žádnou podporu neposkytujeme. Existenci takových dat na několika místech zmiňujeme a objeví-li se tip na open access data tohoto typu, studentům ho sdílíme na webu katedry v sekci k tomu určené.“ (E.Kyselá, email ze 7.6.2018)

Jediná Fakulta sociálních věd UK zvažuje, „že pro mgr. studenty bude takový kurz povinný a nebude učen externistou, ale členem katedry“, zatím nabízí volitelný kurz věnovaný analýze a zpracování Big dat vedený externím vyučujícím z firemního prostředí a „odkazují *[studenty]* na ústav nových médií na FF *[Filozofické fakultě]*, případně na odborné texty“. (P. Soukup, email z 29.5.2018)

Na webových stránkách kateder, které svým stanoviskem nepřispěly, nebyla nalezena vodítka opravňující k závěru, že se této problematice věnují[18].

Lze tedy sumarizovat, že ani v české akademické sociologii (zcela ve shodě se závěrem nad diskuzí vedenou okolo Big dat v sociologii světové) Big data nejsou dosud plně etablována.

Kapitolu lze tak uzavřít citací M. Savage (2010), který s politováním shrnuje, že zhruba od roku 2000 už o sobě „sociologové nemohou tvrdit, že jsou šampióny nových metod.[…] Přinejlepším jsou komentátoři tohoto vývoje, nesedí na sedadle řidiče“ (str.6)

[18] Tematické badatelské okruhy, profesní zájmy vyučujících, přítomnost pojmu „Big data“ apod.

4 Specifičnost Big dat

Nekomfortní pozici sociologie, ve které se vůči Big datům ocitá, pomůžou osvětlit i specifika Big dat a Big dat analýzy, oproti klasickým sociologickým metodám.

4.1 Paradigma Big dat

Pro objasnění lze vycházet z pojmu T.S.Kuhna, paradigma, tedy významnou částí vědecké komunity v daném čase a v dané disciplíně akceptovatelné způsoby zkoumání a výkladu světa; „model, ze kterého vycházejí vnitřně jednotné tradice vědeckého výzkumu, určitá struktura představ, hodnot a postupů“ (Parusníková, 2017). Pro pochopení významu paradigmatu Big dat[19] bude napřed shrnuta metodologická tradice sociologie.

4.1.1 Metodologická tradice sociologie

V rámci empirické sociologie lze rozlišit dvě zásadní paradigmata – kvantitativní a kvalitativní přístup, které jsou až poslední desetiletí spojovány do paradigmatu smíšeného přístupu.

Kvantitativní, podpírán (neo)pozitivistickou tradicí sociologie, je založen na předpokladu měřitelnosti a standardizace, „využívá náhodné výběry, experimenty a silně strukturovaný sběr dat pomocí testů, dotazníků nebo pozorování“. (Hendl, 2008, str.44), často se „zajímá o dokumentaci sociální trendů větších rozměrů“ (str.56). Salomon (str.56) jej považuje za přístup analytický, „snažící se rozumět několika málo proměnným“. Jeho ústředními charakteristikami je ověřování z teorie vzniklých hypotéz.

Naproti tomu kvalitativní, mající oporu v historicky mladším interpretativním sociologickém proudu, práci výzkumníka přirovnává „k činnosti detektiva“ (Hendl, str.48), který „konstruuje obraz, který získává kontury v průběhu sběru a poznání jeho částí“ (str.48). Salomon je považuje za přístup systemický, který „představuje

[19] Není ustáleným terminologickým názvem, v literatuře se používá např. „data-driven“ (Kitchen, 2014), nebo „Paradigmatický posun směrem k Computationl Social Science“ (Tinati et al., 2013). které ale zbytečně přitahují pozornost k dílčím aspektům.

pokus zachytit všechny proměnné v jejich interakci mezi sebou v rámci komplexního prostředí“. (str. 56) Obvykle také bývá provozován v delším časovém období.

Smíšený přístup, využívaný především v aplikovaném výzkumu, pak kombinuje oba přístupy v rámci jednoho výzkumu, kdy je použije buď sekvenčně, nebo simultánně (Hendl, 2008, str.273). V prvním případě se jedná například o výzkum, kde hypotézy, vzniklé v kvalitativním předvýzkumu jsou následně ověřeny v rámci výzkumu kvantitativního, v druhém pak třeba doplnění kvalitativního výsledku kvantitativními daty. (str.279) Je veden snahou eliminovat omezení plynoucí z příslušnosti k některému z dvou paradigmat, kdy „studenti byli vychováváni podle zaměření fakulty či pracoviště k odmítnutí toho či onoho paradigmatu“. (str.273)

V sociálních vědách „díky nástupu průzkumů a statistického modelování bylo jako převládající paradigma institualizováno testování hypotéz […] Generace vědců byly tudíž trénovány v pokládání otázek s ohledem na nulovou hypotézu […] výzkum má za úkol najít podporu pro předem připravenou statistickou hypotézu “ uvádí McFarland, Lewis & Goldberg (2016, str. 14). Tedy kvantitativní přístup, za využití „standardního schématu standardního empirického výzkumu: problém – hypotéza – operacionalizace – sběr dat – verifikace hypotéz – dedukce dalších hypotéz – sociotechnická doporučení“. (Petrusek, 2008, str.31).

Například v jedné ze základních učebnic sociologie, Giddens (2013) jako třetí bod výzkumného procesu uvádí - formulujte hypotézu. „Hypotéza vychází z poučeného předpokladu o zkoumaném dění […] a musí se formulovat takovým způsobem, aby ji shromážděný faktologický materiál buď prokázal, nebo vyvrátil“[20]. (str. 55) Jako hlavní metody sociologického výzkumu pak uvádí terénní výzkum, dotazníkové šetření, experimenty a výzkum založený na dokumentech (toto dále rozvádí jako sbírky, deníky, úřední dokumenty).

Oproti tomu například Hendl uvádí, že „kvalitativní výzkum postupně získal v sociálních vědách rovnocenné postavení“ (2008, str.47). Cílem další sekce bude

[20] Samozřejmě si uvědomujeme jistou zkratkovitost vyjádření a rozdělení hypotéz na východiskové, pracovní a statistické příslušné jednotlivým fázím výzkumu)

přiblížit Big data jako unikátní způsob překlenutí dichotomie kvalitativních i kvantitativních přístupů a to i nad rámec stávajících smíšených metod.

4.1.2 Big data paradigma

Big data „nepotřebují výzkumníkem vytvořenou hypotézu, aby vznikala“, jsou vytvářena mimoděk díky zařízením, které si lidé přirozeně vybrali, nikoli „vzniklým uměle kvůli výzkumu“ a eliminují tak „základní limitace klasických sociologických modeling – based analýz“ – dávají přístup k dosud nezdokumentovatelnému sociálnímu chování bez nutnosti výběru (str. 14). Big data tak „umožňují stavět teorie induktivně odspodu (myšleno od dat), nikoli teorii předpokládat, nasbírat data a na jejich základě ji potvrdit nebo vyvrátit“. (McFarland et al., 2016, str. 15)

Podobně vidí možnosti Big dat i Kitchen (2014), když popisuje tzv. data driven přístup, kdy data jsou používána pro tvorbu hypotéz, které potom mohou být vyhodnocovány na základě stávajících teorií. Data driven přístup tak „uznává úlohu konvenčních vědeckých pojmů a metod nad rámec pouhého rozpoznávání vzorů [v datech], ale jeho hypotézy jsou odvozeny ze samotných dat a nikoliv "jen" od teoretických principů.“

V tomto popise nelze nepostřehnout kvalitativní rysy, Big data ale ze své podstaty jsou také kvantitativní metodou.

Kvantitativní charakter je reprezentován

- využíváním klasických statistických metod a metod umělé inteligence
- kvantifikací, standardizací Big dat
- nutností využívat informační technologie
- počtem zkoumaných jednotek
- vírou, že poznání tímto způsobem je možné a účelné

Kvalitativní charakter

- Big data mohou být sbírána komplexně, a to jak v horizontálním a vertikálním směru, tak i časové rovině. Není typem výběrového šetření, data mohou být sbírána bez přesného zadání ve všech svých souvislostech (nejedná se tedy o

sběr partikulárních proměnných příslušných k hypotézám) a mohou obsahovat informace z delšího časového období

- Big data reálně existovala před sběrem, nebyla uměle vytvořena k analytickým účelům
- Big data nemusí být vyšetřována pro ověření platnosti hypotézy, ale mohou být arbitrárně zkoumána.
- některé metody jsou pouze průzkumné, bez predikativního potenciálu, například neuronové sítě v učení bez učitele. Obecně metody umělé inteligence neprodukují matematizovaný vztah, chybí tedy exaktní vysvětlitelnost.

Big data tak umožňují kombinovat KVL i KVN charakteristiky a to nikoli střídáním či doplňováním, ale unikátní syntézou, kdy v sobě obě tyto charakteristiky zároveň nesou jak data, tak metody. Vystihnout tento specifický rys lze označením Paradigma Big dat, neboť dosavadní používané názvy pro změnu paradigma vlivem Big dat „data-driven paradigma" (Kitchen, 2014) i „posun směrem k computational science" (Chan, Kauffman & Kwon, 2014) odkazují k dílčím částem a nevystihují jeho komplexnost.

Big data paradigma je ale obtížněji uchopitelné, neboť nutí vystoupit nad rámec dosavadních, v rámci sociologie dobře metodologicky zpracovaných postupů a technik. A klade také zvýšenou náročnost na dovednosti ve využívání informačních technologií.

4.2 Kompetence v oblasti informačních technologií

Big data jsou bytostně spjata s informačními technologiemi[21], které je umožňují získat, skladovat, analyzovat a díky kterým také vznikají. Je to ostatně jeden z přístupů k jejich vymezení (Symons & Alvaro, 2016).

Jako „data analysis divide" označuje L.Manovich (2012) propast mezi experty na analýzu dat a výzkumníky bez znalostí „computer science" (str. 461). Nastiňuje potenciální data (sociální média, API přístupy, volné datové archivy, apod) a vyvozuje - „možnosti jsou nekonečné – když umíte programovat, znáte datovou analýzu a jste

[21] Případně Informačních a komunikačních technologií (ICT), označení zdůrazňující důležitost přenosu informací, které začalo vytlačovat původní označení informační technologie (IT).

připraveni ptát se nové typy otázek“. Věří, že nakonec sociální vědci budou schopni pracovat sami a nezáviset na počítačových vědcích, „nicméně to vyžaduje velkou změnu v tom, jak jsou studenti v humanitních oborech vzděláváni“[22]. (str.473)

Jako ilustrativní příklad takovéhoto překlenutí propasti lze uvést diplomovou práci v oboru sociologie inovativně zkoumající komunikační struktury zpravodajského serveru (Pilnáček, 2016). Autor si data stáhl s využitím programovacího jazyku Python, pro analýzy využíval pokročilé programování v jazyce R založené na hluboké (jak ilustroval v textu, tedy nikoli povrchní, uživatelské) znalosti metod analýzy sociálních sítí. Autor byl kromě sociologie zároveň studentem teoretické informatiky[23].

Bylo by obtížné kvantifikovat přesnou úroveň ICT kompetenci sociologů, po dlouho dobu ale bylo v rámci sociologie postačující využívání založené na spíše uživatelském přístupu k aplikacím. Lze však jednoduše prozkoumat, jak jsou na zdolání „data analysis divide“ připravování studenti sociologie. Dle odpovědí tří ze sedmi poptávaných kateder[24], se lze na Filosofické fakultě UK ve volitelném jednosemestrálním předmětu seznámit s programovacím jazykem R (E.Kyselá, email ze 7.6.2018) a na Fakultě sociálních věd UK ve stejném rozsahu se základy jazyka Python (P. Soukup, email z 29.5.2018). Fakulta sociálních věd MUNI zaslala jen přehled vyučovaných statistických nástrojů (T. Katrňák, email z 30.5.2018), ostatní katedry neodpověděly a v rámci sylabů žádný takový předmět neuvádějí.

Trendu se tak zatím pokoušejí jít alespoň trochu naproti nesociologická pracoviště. Například Fakulta informatiky Masarykovy univerzity nabízí studijní bakalářský program Sociální informatika, obor, který „klade důraz na získání přehledu současných trendů sociální informatiky pro návrh a vývoj informačních systémů, rozvoj kritického myšlení a na získání schopností fundovaně skloubit oblasti informatiky a sociálních věd.“ [25] Dle nabídky předmětů si absolvent odnese nejen

---

[22] Ponecháváme citaci v původním znění, z vyznění článku je zřejmé, že autor myslí sociální i humanitní vědy.

[23] Mgr. Matouš Pilnáček. Získáno 10.6.2018 z https://www.soc.cas.cz/lide/matous-pilnacek

[24] Bylo popsáno v kapitole věnované Big datům v české akademické sociologii, otázky se týkaly i IT dovedností.

[25] https://www.muni.cz/bakalarske-a-magisterske-obory/17463-socialni-informatika#prijimaci-rizeni

znalosti sociologické teorie a metodologie výzkumu, ale i programování, algoritmizace, databází a matematiky.

Dalším příkladem progresivních oborů jsou studia zaměřená na nová média. Například v rámci Studia nových médii na Filosofické fakultě UK si studenti mohou zapsat předmět Informační věda: Reprezentace znalostí, kde se seznámí s výzkumy umělé inteligence, expertními systémy atd. Předmět Digital Humanities pro humanitní vědy se věnuje data miningu a studenti se mimo jiné seznámí se strojovým učením, strojovou analýzou atd.[26]

Lze samozřejmě namítnout, že v kreditovém systému vyučování je výuka otevřená a studenti si tak mohou zapsat jakýkoli předmět nejen v rámci fakulty ale i univerzity, iniciativa je tak ale nechána na jednotlivcích. Znalosti v oblasti ICT tak nejsou normou, ale spíše konkurenční výhodou. Systematická podpora širších kompetencí v oblasti informačních technologií ve výuce sociologie chybí a pro sociology tak může být zvládnutí problematiky obtížné. V éře Big dat je ale tato úroveň kompetence omezující.

Důsledkem tak je, že v oblasti Big dat analýz zatím dominují spíše technické obory, které mají k jejich zvládnutí patřičné „know-how". Nedisponují ale rozvinutým aparátem metodologickým ani sociálněvědními znalostmi, proto produkují analýzy „inženýrského typu" popsané v úvodu (McFarland et al, 2016) a sklouzávají k empiristickému paradigmatu (Kitchen, 2014), který bude rozebrán v části věnované Critical data studies.

[26] https://is.cuni.cz/studium/predmety/index.php?do=ustav&dlpar=YToxOntzOjg6InByZWRtZXR5IjthOjE6e3M6Mzoic2tyIjtzOjQ6IjIwMTciO319&fak=11210&kod=21-UISKNM Získáno 10.3.2018

5 Zdroje Big dat pro sociologický výzkum

Představovaná data v této kapitole budou odpovídat definici, že za Big data je považováno vše, co vyniká rychlostí generování/ ukládání/ zpracování/ publikování a nejedná se o výběr (tedy typicky jsou zahrnuti všichni uživatelé v rámci systému) (Kitchen & McArdle, 2014).

Data budou systematicky členěna podle jejich vlastníků. Toto členění je zvoleno pro největší názornost a minimum překryvů v datech, jakkoli je zřejmé, že dojde-li na konkrétní příklady vlastníků, pak některé subjekty poskytují více druhů služeb/produktů, případně ji poskytují na více platformách, a spadají tak do více kategorií a stávají se tak vlastníky více druhů dat (například mobilní operátoři mají nejen data z provozu mobilní sítě, ale zároveň poskytují služby na webu a vlastní některé mobilní aplikace). Alternativním přístupem by bylo představování dat dle jejich typologie (text, video, audio atd.), které by bylo zvoleno v případě, že by k datům byly rovnou popisovány i metody analýzy.

Vlastníci samotní jsou z hlediska této diplomové práce marginálním tématem, jejím smyslem není varování před velkými informačními hegemony apod. Zároveň ale toto členění nenásilně poukazuje na další zásadní aspekt Big dat, totiž že jsou většinou někým vlastněna a přístup k nim tak zhusta závisí na dobré vůli vlastníků, případně je rovnou legislativně omezen. Popsané možné formy získání těchto dat (a jejich deklarovaná přístupnost) je tak nutno chápat spíše jako technický popis způsobu zisku. Legislativní omezení je diskutováno v kapitole věnované Critical data studies.

Možné zdroje dat budou představeny z více hledisek – o jaký typ dat se jedná, jaká je jejich vypovídající síla (tj. zachycené populace v ČR), jak je lze získat. U vybraných zdrojů dat bude uveden i přehled nejzajímavějších studií, jejich výsledky budou nicméně pouze nastíněny. Důvodem je snaha držet se meritu práce – tedy představit využitelnost dat, nikoli odpoutávat pozornost na samotné závěry.

Při přehledu možných metod zisku dat jsou uváděny (má-li to smysl) i externí, volně přístupné aplikace. Je vhodné na tomto místě upozornit na nutnost používat je ve shodě s jejich podmínkami užití (licenčními podmínkami), které je všechny v tuto chvíli opravňuje pro akademické a nekomerční účely (nicméně je to třeba vždy ověřit).

5.1 Data poskytovatelů mobilních služeb

Dle poslední Výroční zprávy ČTÚ (2017) v ČR v prosinci 2016 působili tři mobilní operátoři (s relativně vyrovnaným podílem na trhu) a 157 virtuálních (skutečně vykazujících aktivitu), kteří dohromady obhospodařují celkem 14,1 milionů SIM karet (z toho 7% připadá na virtuální). V ČR bylo v roce 2016 7,6 milionů SIM s internetem v mobilu (ČTÚ, 2017). Podle poslední zprávy ČSÚ (2017) na jednu domácnost připadá průměrně 2.7 přístrojů, přístup k mobilnímu telefonu má 98% domácností a na jednoho člena domácnosti staršího šesti let připadá průměrně 0.99 mobilu. Dá se tedy říct, že penetrace je téměř stoprocentní.

Přehled sbíraných dat přináší diplomová práce věnovaná právním aspektům retence, tedy zálohování a skladování dat (Jiřovský, 2015). Mobilní operátoři o každé aktivitě zaznamenávají tzv. CDR, call detail record, obsahující telefonní číslo volajícího i volaného, identifikátor IMSI, čas začátku hovoru, délku hovoru, typ (hlasová služba, SMS,…), identifikátory základových stanicí začátku a konce hovoru a mnoho dalšího.

V případě internetového připojení poskytovatelé mimo jiné uchovávají adresu MAC, datum a čas zahájení a ukončení připojení adresu IP a číslo portu. V případě využití mobilního paketu potom ještě identifikátory stanic začátku a konce přenosu, identifikátor telefonu.

Tato data jsou v surové formě uchovávána na základě Zákona a elektronické komunikaci po dobu, kdy je možná reklamace. Jednotlivé záznamy jsou samozřejmě pečlivě chráněny a mohou být případně zveřejněny pouze v důsledně anonymizované podobě. Uvolnění takto anonymizovaných dat, respektive přístup k nim, je zcela v režii vlastníků a data tak lze získat pouze dohodou.

Potenciální ochota poskytnout data k sociologickému výzkumu byla ověřena u třech hlavních operátorů. Z dotázaných společností se vyjádřily společnosti T-Mobile a Vodafone.

T-Mobile se teoreticky „spolupráci s VŠ nebrání, prakticky by ale záleželo na spoustě věcí [úroveň agregace, místo, kde by se analyzovalo apod. …] Záleželo by také na tom, k jakému účelu by ta data byla určena, aby nedošlo k překryvu s některými našimi

komerčními aktivitami“, nicméně po přesné specifikaci dat se „můžeme o tom pobavit, případně najít vhodný kompromis“. (Z.Stroblova, email z 8.6.2018).

Společnost Vodafone zpravidla data plošně neposkytuje, nicméně každou „jednotlivou žádost individuálně posuzujeme“. Pokud by se rozhodla data poskytnout, pak „šlo by o anonymní data s rozostřenou časovou známkou a polohou (např. zaokrouhleno na 30min a obec). Zároveň by šlo o omezené časové období.“ (I.Vejvodová, email z 11.6.2018)

V ČR zatím data komplexně zanalyzována nebyla. Převažuje zveřejňování dílčích výzkumů společností, například analýza vytíženosti stanic metra (02)[27] nebo dílčí studie společnosti T- Mobile shrnuté v PR prezentaci (Kovárník, Tůma, & Dvořák, 2014) - pohled na rozložení obyvatel Prahy ve dne a v noci, monitorování dopravy, studie mobility po povodních v Karlíně nebo analýza návštěvnosti Národního parku Šumava realizována se společností KPMG.

Data jsou také využívána pro vlastní komerční aktivity. Například společnost T-Mobile na svých stránkách nabízí využití vlastních Big dat, sestávajících z „anonymizovaných údajů o poloze zákazníků v síti, internetovém provozu, sledování televize a provozu aut k zlepšení jejich (zákaznického) businessu“. (Marek, 2017)

V Evropě obecně mobilní operátoři data o Evropských uživatelích k výzkumu příliš neuvolňují. Krásným příkladem využití mobilních dat pro vědecké účely je ale výzva společnosti Orange, Data for Development (D4D) Orange Senegal. Francouzský telefonní operátor Orange (původně France Telecom) vyzval v roce 2014 vědeckou komunitu, aby se podílela na otevřené kampani D4D- Senegal Challenge 2014, v rámci které nabídl k dispozici vybraná mobilní data uživatelů své pobočky v Senegalu za rok 2013. Podmínkou poskytnutí anonymizovaných dat bylo schválení návrhu výzkumu, který měl přispět k socioekonomickému rozvoji a „well-being“ Senegalské společnosti, a posléze publikování výzkumné zprávy a posteru v rámci sborníků přístupných po uzavření projektu na stránce D4D. (de Montjoye, Smoreda, Trinquart, Ziemlicki & Blondel, 2014)

[27] Pohl, O. (8.5.2014) Co o nás prozradí naše mobily operátorům? Získáno 10.3.2018 z https://mobilenet.cz/clanky/co-o-nas-prozradi-nase-mobily-operatorum-15676

Data byla poskytnuta ve třech souborech:

1. Hodinový souhrn aktivit pro jednotlivé antény (1666) za celý rok 2013.

2. Plná čtrnáctidenní aktivita náhodně vybraného vzorku uživatelů, změna vzorku vždy po čtrnácti dnech. Tento soubor poskytnul data o 300 000 uživatelích.

3. Plná aktivita během celého roku na náhodně vybraném vzorku 146 352 unikátních uživatelů, která byl uvedena na úrovni 123 územních celků.

K meta-analýze se nabízí celkem 52 studií využití mobilních Big dat v dokumentu Data for development – Challenge Senegal, Book of abstracts: Scientific Papers (2015) [28] . Rozsah představení výsledků se může zdát větší než u jiných následujících studií, ale to je dáno především počtem a mírně také využitelností v dalších kapitolách. Tyto studie jsou totiž zpracovány nad stejnými daty s podrobně pospanými vstupními materiály, s výhodou je tak lze použít v kapitole věnované Critical data studies (CDS), kde některé přístupy/aspekty podrobíme detailnějšímu (a často kritickému) pohledu.

Potenciální sociologická výtěžnost těchto dat bude také naznačena bez hlubší interpretace a nastínění Senegalských reálii, které nejsou předmětem této práce. Ilustruje jen, co se lze o společnosti dozvědět z dat (bez kritického zhodnocení, který bude součástí v kapitole věnované CDS).

Analýzou výzkumníci zjistili, že komunikace je vysoce symetrická, SMS a volání v jednom směru téměř vždy následuje v opačném a odehrává se především uvnitř regionu, případně s hlavním městem. Podobně nejvíce komunikace je vždy v rámci jedné antény. S rostoucí vzdáleností klesá počet hovorů, délka hovorů i počet SMS (ten nejvýrazněji).

Zvyklosti se mění v průběhu Ramadánu, kdy se v době mezi 22 – 06 počet hovorů zdvojnásobil a hovory také trvaly déle. Dle počtu hovorů lze také identifikovat další významné muslimské festivaly a svátky, případně politické události (lokální

[28] Více o projektu D4D Challenge Senegal lze nalézt na stránce projektu http://www.d4d.orange.com/en/Accueil

maximum v době útoku separatistů). Některé antény v zemědělských oblastech také vykazují významně jiné chování v období sklizně.

Volání mají obecně dva vrcholy, mezi 11-12 a 20-21, textové zprávy největší aktivitu v noci – od 20 do 02, s vrcholem ve 23. Den ve městě začíná přibližně okolo osmé a končí okolo jedenácté. Lidé si obecně více volají, než píší, zvýšená aktivita SMS byla zaznamenána v době letních prázdnin. Clusterovou analýzou lze rozlišit tři typy antén s odlišným chováním (vrcholy a typ komunikace) – městské, subměstské a venkovské.

Mobilní aktivita (se zohledněním penetrace) na úrovni departamentů (tedy čtyřiceti pěti územních celků) je vysoce korelovaná s oficiálními daty z censu (liší se na nižších úrovních) a také vysoce koreluje se spotřebou elektřiny a úrovní chudoby.

Cestování mezi regiony je úměrné počtu uskutečněných hovorů. Co se týče cestování, jen 40% uživatelů z dat cestuje do různých destinací, 35% má jednu pravidelnou destinaci, 15% dvě a 10% tři a více. Nejčastěji lidé spali v cca 50-ti kilometrové vzdálenosti (když se vzaly v potaz všechny cesty). Počet cest z regionu je úměrný hustotě obydlení, s výjimkou nejsevernějšího regionu.

Volají si především města v sousedství, s odchylkou míst postižených povodněmi a Dakaru. Uživatelé mimo hlavní město kontaktují méně čísel a méně často. Lidé v oblastech, kde významně dominuje volání nad textovými zprávami, jsou v kontaktu s menším počtem lidí (vztah s gramotností). Významné abnormality v provozu vykazují antény v zemědělských oblastech, což autoři přičítají nedostatku kontaktů a tedy sociálního kapitálu.

Přehled studií dat založených na mobilních datech nabízí také například článek Candia, Gonzalez, Wang, Schoenharl, Madey a Barabási (2008). Můžeme shrnout, že mobilní data mohou být ze sociologického pohledu využita ve studiu skupin, v urbanistické sociologii, sociologii každodennosti či pro průzkum šíření inovací.

5.2 Data poskytovatelů mobilních aplikací

Mobilní aplikace jsou programy uzpůsobené pro fungování na mobilním telefonu, běžících na jejich operačních systémech, tedy nejčastěji Android nebo iOS. Nejznámější platforma Google play pro operační systém Android nabízela v březnu

2018 2.8 milionů aplikací v celkem dvaceti kategoriích, zahrnujících hry, komunikační aplikace, vzdělávací, organizační atd. (stranou ponechme sociální sítě, kterým bude věnována další kapitola), Apple App store pro operační systém iOs pak 2.2 milionů[29].

Zatímco data skladována mobilními operátory podléhají zákonné regulaci, jejich vlastníci jsou známí, podléhají přísným regulím a tudíž přístup k jejich datům je kontrolovaný, mnohdy neznámí vlastnící/vývojáři mobilních aplikací spravují nikým nekontrolovanou rozsáhlou datovou základnu (pro úplnost jen stručně dodejme, že většinou zcela v souladu s uživateli pro forma odsouhlasenými podmínkami). Podle výzkumu sponzorovanému University of Carolina, například sedm z deseti Android aplikací posílá data dále[30].

V České republice používá chytré telefony 58% lidí (Studie Googlu) a z toho 62% vlastníků je využívá ke stahování aplikací (focus, Marketing a Social Research, 2016). Přes 25% uživatelů v ČR používá více než 10 aplikací a jen 5% uživatelů nepoužívá žádné, respektive pouze ze základního nastavení[31] (pro korektnost dodejme, že do výzkumu byly zařazeny i aplikace sociálních sítí).

Data z aplikací lze získat buď dohodou s vlastníky stávajících aplikací, nebo napsáním si aplikace vlastní, kterou lze distribuovat přes oficiální platformy dle příslušného operačního systému (stačí registrace, zaplacení poplatku a dodržení procesu uvolnění), případně umožnit stažení z webové stránky nebo ji lze odeslat jako přílohu mailu. Pro Android se standardně programuje v jazyce Java, ale existují i jednoduché tzv. hybridní aplikace vytvořené v jiném jazyku a tzv. „přeložené“. Pro iOS jde především o Objective-C, případně Swift.

Sběr dat z aplikací je využíván ponejvíce pro marketingové účely (telefonní čísla, emaily, cílení reklamy – prostorové nebo obsahové), ale i přímo pro zaměření obsahu

---

[29] Number of apps available in leading app stores as of March 2017. Získáno 10.4.2018 z https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/

[30] 7 in 10 smartphone apps share your data with third-party services. Získáno 10.4.2018 z http://theconversation.com/7-in-10-smartphone-apps-share-your-data-with-third-party-services-72404

[31] Nejrozšířenější aplikací u Čechů je FB Messenger. Získáno 10.4.2018 z https://www.mediaguru.cz/clanky/2017/04/nejrozsirenejsi-aplikaci-u-cechu-je-fb-messenger/

nebo služby. Posílány mohou být nejen data o uživateli a využívání aplikace, ale i data z dalších senzorů typu GPS lokátor (určí polohu mobilního zařízení), akcelerometr (měří změnu rychlosti pohybu mobilního zařízení), gyroskopický senzor (podle úhlové rychlosti měří změnu natočení mobilního zařízení), senzor přiblížení (určuje blízkost jiného objektu k mobilnímu zařízení), teplotní senzor atd. Kombinace jednotlivých dat přináší „obrovskou příležitost pochopit, jak uživatelé interagují se svými telefony, jaké mají vzorce mobility, sociální vztahy a osobní preference v rozmanitých kontextech“ (Cao & Lin, 2017)

Nejdále jsou ve využívání dat z mobilních aplikací v medicínském výzkumu. Existuje nepřeberné množství zdravotních aplikací, pomáhajících uživatelům zvládat nějaké zdravotní okolnosti (cukrovka, vysoký tlak, spánkové problémy, stravovací návyky), které zpětně od uživatelů dostávají tematická data, která jsou posléze analyzována. Například aplikace poskytující rady pro minimalizaci dopadů spánkového deficitu vlivem jet lag na základě spánkového režimu uživatele a spánkových pásem, ve kterých se vyskytuje, elegantně zpětně získává přesné, detailní a dobrovolné informace o spánkovém cyklu jednotlivých uživatelů. Tato aplikace byla dokonce napsána primárně za účelem zisku těchto dat, když si autor všiml zoufalství průzkumníků lovících před univerzitou potenciální placené respondenty dotazníku. Obdobně jsou pro medicínské účely hojně analyzována pohybová data.

V sociálních vědách byly aplikace využity například v Hong Kongu k analýze mobility a sociální segregace (Yip, Forrest, & Xian, 2016). Nejednalo se ale o sběr mimovolně produkovaných dat, byla vyžadována přímá spolupráce uživatele a to poměrně obtěžující formou po dobu sedmi dnů, což dokázalo splnit pouze 1.33% oslovených participantů a ve finále tak byly k dispozici údaje pouze od 71 uživatelů. Nicméně z výsledků vyplývá, že vyšší a středně příjmové skupiny a nízkopříjmové skupiny se drží většinou při sobě, mladí lidé většinu aktivit prožívají mimo domov a zatímco lidé nakupují především sami, zábava se jim pojí s návštěvami a přáteli[32].

[32] Práci byla do přehledu zařazena s vědomím všech metodologických výhrad i také s výhradou, že uvedený příklad jde lehce proti samotné logice Big dat (kde většinou produkce dat není primárním účelem vytvoření nástroje, data jsou produkována jako vedlejší produkt). Zařazena ale byla pro ilustrativní účely, na co by mohla být data mobilních aplikací použita (kdyby byla sbírána lépe).

Podobným příkladem úmyslně napsané aplikace pro výzkum byla aplikace sbírající logy (log je jakousi stopou zaznamenávající každou činnost, která byla na telefonu provedena) smartphonu (Rivron, Khan, Charneau, & Chrisment, 2016). Opět jako v předchozím případě se při výzkumu jednalo o kombinaci klasického sociologického dotazníkového šetření (věk, pohlaví, povolání atd.) a analýzu uživatelského chování z dat, což dokázalo splnit méně než 50% potenciálních uživatelů. Z výsledků mimo jiné vyplývá, že jaká je frekvence a množství pravidelně využívaných aplikací nebo jaké jsou rozdíly ve využívání telefonu podle pohlaví, pracovního zařazení. Jak shrnují autoři „využívání smatrphonu souvisí nejen s demografickými faktory, ale také s životním stylem, speciálně s kulturními zvyky a technologickými preferencemi" (str.8).

Sociální vědci si ale nemusí vytvářet aplikace sami, možností je samozřejmě i po dohodě využít data etablovaných provozovatelů, tak jako to při výzkumu homofilie, tedy tendence obklopovat se a spojovat se s podobnými lidmi, udělali Jeong-han a Da Young (2017) s daty mobilní seznamovací a radící aplikace Taxtet. Opět se jednalo o kombinaci analýzy dat a krátkého on-line dotazníkového šetření, výsledný vzorek obsahoval přes 7 000 objektů. Ze zjištění mimo jiné vyplývá, že signifikantní roli v rámci homofilie hrají demografické faktory typu věk, vzdělání apd., méně pak osobnostní charakteristiky typu introverze atd.

5.3 Data poskytovatelů internetových služeb a (nebo) obsahu

Obrovským množstvím dat disponují samozřejmě poskytovatelé internetových služeb (eshopy, on-line seznamky, vyhledávače atd.) a internetového obsahu (streamovací služby, on-line zpravodajské a zábavní weby, blogy, odborně nebo hobby zaměřené weby apd.). Na převahu těchto dat nad běžnými výzkumy poukázali už před desetiletím (a tedy dávno před vznikem současného analytického firemního boomu) ve svém ikonickém článku o přicházející krizi empirické sociologie Savage a Burrows (2007) povzdechem, že Amazon.com nemusí doporučovat další knihy ke koupi na základě určení sociální pozice zákazníka výběrovým šetřením (a nabídnutím mu knih, které si kupují lidé ve stejné skupině). Amazon.com má „mnohem mocnější nástroj. Oni *exaktně* vědí, jaké knihy jsou kupovány lidmi dělajícími každou jednotlivou objednávku" (str.9, zvýraznění zachováno z článku).

Ve firemních databázích může být uložena nejen jakákoli dokončená akce učiněná uživatelem, ale i každá činnost, kterou na internetové stránce udělal, kdy pomocí tzv. skriptů uložených v kódu stránky může být logován (tedy zaznamenáván) každý pohyb uživatele na stránce.

Akcí v této souvislosti rozumíme uskutečnění čehokoli, co poskytovatel nabízí, tedy kliknutí na článek/video v případě mediálních domů, uskutečnění objednávky v případě eshopu, kontaktáž v rámci seznamky apd. Činností pak například způsob, jakým je článek čtený (tzn. jak rychle a jestli se uživatel dostal na konec článku), pohyb na stránce se zbožím, posunutí v rámci sledování videa apd.

Podle výzkumu technologické společnosti TWDI jsou analýzy těchto dat poskytovateli služeb nejvíce využívány na lepší zaměření marketingu, detailnější pohled na business, segmantaci zákazníků, rozpoznání obchodních příležitostí a automatické rozhodovácí real-time procesy (Russom, 2011). Poskytovatelé obsahu pak využívají tato data především pro vyladění personalizačních algoritmů a samotného vyladění obsahu. V této souvislosti můžeme jmenovat například společnost Netflix, která je průkopníkem využívání datové analytiky. Podle výzkumu společnosti McKenzie (Henke et al., 2016) ji jen vylepšení doporučovacího algoritmu (tedy algoritmu, který uživateli nabídne obsah, který se mu bude pravděpodobně líbit) a jeho uvolnění globálním zákazníkům přineslo 1 miliardu dolarů ročních příjmů (str.21); obecně pak u firem investice do datové analytiky vedou k průměrnému zvýšení zisku o 6 až 8% (str.30).

Výzkumníci pak mohou z těchto dat čerpat třemi možnými způsoby – 1. přímou analýzou, pokud se se společností vlastnící data dohodnou, 2. analýzou agregovaných dat firmou zveřejněných nebo 3. z reportů, někdy společnostmi z PR důvodů produkovanými.

Typickým příkladem využití agregovaných dat poskytnutých provozovateli jsou analýzy nad daty internetového vyhledavače Googlu, využívající uživatelské rozhraní Google Trends, kde lze nalézt, jak často byl pokládán jednotlivý dotaz, s možností konkretizace času a místa (granularita poskytovaných informací se liší dle jednotlivých území), https://trends.google.cz/trends/. Nejkomplexnější poznatky získané analýzou Google Trends (spolu v kombinaci s ostatními statistikami

uvolněnými firmami jako například pornHub atd.) jsou shromážděny v populárně naučné knize „Everbody lies“ (Stephens-Davidowitz, 2017), kdy „[statistika] hledání na Googlu prezentovalo překvapivě odlišný obraz Ameriky“ (str.12). Zkoumány jsou sexuální zvyklosti, voličské preference, vztah k potratům a mnoho dalších témat. Výsledky jsou také dávány do kontextu s běžně přístupnými výsledky získanými klasickými metodami výzkumu (např. Gallupova institutu), protože „Google má mnoho informací, které v průzkumu veřejného mínění chybí a které mohou být užitečné v porozumění“. (str. 14)

Data Googlu jsou k dispozici i pro Českou republiku, podíl tohoto vyhledavače však není natolik dominantní jako angloamerickým zemích, v popularitě soutěží s vyhledávačem společnosti Seznam.cz (přesná statistika nicméně neexistuje, podle vícero nezávislých marketingových drobných průzkumů jsou jejich podíly víceméně vyrovnané). Podle reklamních údajů společnosti Seznam.cz každý den lidé položí 15 miliónů dotazů[33]. Přístup k analýze vyhledávání lze nalézt na stránce https://search.seznam.cz/stats/?search-service=1, kde po položení dotazu nabízí až roční historii s granularitou času (den, týden, měsíc) nebo platformy (PC, mobil, tablet, ostatní). K dispozici jsou i statistiky podobných dotazů s denním průměrem za poslední dva měsíce.

A v souvislostech okomentované analýzy, tedy třetí způsob zisku informací, ze svých dat poskytuje společnost Seznam, c.z. na svém blogu https://blog.seznam.cz/technologie/vyhledavani/, kde lze zjistit například nejhledanější výrazy za jednotlivá období (nepřekvapivě diety po Novém roce a plesové šaty v únoru), nejsledovanější události apd. A také v rámci PR novinových rozhovorů, např. se lze například dozvědět, co v kterou denní dobu čtou Češi nejvíce na internetu.[34] Samotná datová základna Seznam.cz je přitom ohromující, uvědomíme-li si, že některou z jeho služeb využije měsíčně 9 z 10 Čechů a domovskou

[33]Seznam vyhledávání je hlavním zdrojem návštěvnosti většiny českých webů. Získáno 23.4.2018 https://www.seznam.cz/reklama/cz/obsahovy-web/sluzba-seznam-vyhledavani/

[34] Češi začínají na internetu den horoskopy a končí nakupováním. Získáno 18.3.2018 z https://relax.lidovky.cz/cesi-zacinaji-na-internetu-den-horoskopy-a-konci-nakupovanim-p5b-/zajimavosti.aspx?c=A170214_093342_ln-zajimavosti_ape

stránku měsíčně navštíví téměř 78% internetové populace ČR (SPIR NetMonitor, 2017).

5.4 Data webových stránek a on-line přístupných databází

Další možností zisku dat je tzv. scraping, tedy prohledávání a stahování obsahu z volně přístupných webových stránek (tedy bez jejího vlastnictví, pouze z přístupného obsahu).

Pro představu o počtu stránek k analýze uvádí poskytovatel internetové národní domény .cz ke konci května 2018 celkem 1 313 097 registrovaných domén[35], za kterými může existovat nespočet stránek. Tento počet však také není úplný, protože stránky nemusí končit jen národní doménou .cz, ale také některou z nadnárodních jako .org (původně značící nekomerční využití), .com (očekává se spíše komerční využití), .eu (doména v rámci EU), .info (informace) apd.

Stran počtu uživatelů internetu pak uvádí například studie ("Studie Google", 2017), že 91% lidí mladších 25-ti let využívá internet denně, v celé populaci se pak jedná o 65%.

Webové stránky jsou obdobou dokumentů vytvořených ve formě HTML, tedy HyperText Markup Language, značkovacího jazyka, který pomocí definovaných značek a příkazů obohacuje samotný text o strukturu a určuje způsob zobrazení. A je tak velmi snadné napsat jednoduchý skript v některém z programovacích jazyků, například Python (příslušné knihovny například Scrapy, BeautifulSoup), Java (příslušné knihovny například Jaunt, Jsoup) atd.

Případně existuje nespočet aplikací, volně ke stažení, které dokáží stáhnout a upravit obsah do formy vhodné k analýze. Například v okleštěnější verzi volně přístupný OutWit Hub (ke stažení http://www.outwit.com/), který dokáže ze zadané webové adresy stáhnout a logicky rozčlenit text stránky a nabízí jeho export do mnoha formátů k dalšímu zpracování (SQL,HTML, CSV, JSON apd.). Umí také pracovat s obrázky. Nebo RapidMiner (ke stažení https://rapidminer.com/), který na vyžádání

[35] https://www.nic.cz/ Získáno 28.5.2018.

umožňuje získat dvanáctiměsíční akademickou licenci pro studenty a profesory zdarma. RapidMiner posléze umožňuje nad daty provádět i sofistikované analýzy.

Využít lze také volně přístupné databáze, jako jsou například digitalizované přehledy publikovaných knih, Projekt Gutenberg nebo Google Books Ngram. Nad obojím existují prohledávací aplikace, případně lze přistupovat přímo pomocí programovacích jazyků.

Příkladem scrapingu webových stránek je například diplomová práce M.Pilnáčka[36] (2016). Data volně přístupných databází pak využili Chen a Yan (2016) pro půvabný přehled dějin sociologie. Srovnali počet publikací v čase (nejvíce literatury týkající se sociologie bylo publikováno v sedmdesátých letech, poté přišel ústup ve prospěch psychologie a ekonomie), zanalyzovali i frekvenci výskytu jmen dvanácti nejvlivnějších sociologů, kdy identifikovali několik zajímavých trendů (žádný sociolog od sedmdesátých let vlivem nepřekoná předchozí generaci), prozkoumali relevanci deseti nejvlivnějších sociologických teorií apod.

5.5 Data jiných zpracovatelů dat

Nejen provozovatelé on-line služeb mají rozsáhlou databázi informací, i ostatní poskytovatelé služeb a produktů sbírají obrovská množství dat. Ať už jsou to data vzniklá jako vedlejší produkt činnosti jejich zákazníků/klientů/návštěvníků apod., tedy tzv. transakční data, nebo jsou data samotná jejich hlavní obchodní náplní.

Transakčními daty jsou myšlena data, která vznikají jako otisk činnosti, někdy bývají označované také jako „by-pass“ produkt. Jde například o data poskytovatelů platebních karet, platebních terminálů, knihoven, poskytovatelů věrnostních kartiček apod.

Druhým typem soukromých dat jsou myšlena data, která jsou hlavním předmětem obchodní činnosti svých vlastníků. Jde například o data v rámci nestátních registrů, jako je například Registru dlužníků společnosti SOLUS apod, který systematicky sbírá od věřitelů data o jejich dlužnicích a poskytuje za úplatu společnostem poskytujícím půjčky informace o případném zápisu žadatelů v registru a míře jeho závažnosti. Pro

[36] Popsána byla v kapitole věnované Specifikům Big dat.

představu, „v Registru fyzických osob SOLUS bylo ke konci září 2017 zapsáno necelých 640 tisíc osob“.[37]

Pouze velmi omezené množství těchto dat jiných provozovatelů je v nějaké okleštěnější formě volně přístupné, většina je předmětem obchodního (nebo bankovního apod.) tajemství. Komunikovány jsou v rámci PR kampaní výsledky jejich vlastních analýz, jako například zprávy společnosti SOLUS (“V negativním registru SOLUS”, 2018) o procentu osob se záznamem v registru dlužníků v jednotlivých krajích (nejvíce jich je v kraji Ústeckém, nejméně pak v kraji Zlínském).

I přesto lze nalézt způsoby, jak se k těmto bezpochyby zajímavým datům v rámci akademického výzkumu dostat. Například Matematicko-fyzikální fakulta UK na základě smluvního výzkumu s Československou obchodní bankou analyzovala možnosti strojového učení nad anonymizovanými daty z transakcí na běžných účtech klientů. Jako další směr společného výzkumu pak vidí například propojení dat klientů s veřejně dostupnými daty (mapy kriminality, katastr nemovitostí atd.). (Šolcová, 2016) Fakulta informačních technologií ČVUT podobně spolupracuje s Komerční bankou, výsledkem je například diplomová práce na reálných klientských datech zkoumající vzorce chování firemních zákazníků. (Nenenko, 2017)

5.6 Data provozovatelů sociálních médií

Podle nejcitovanějšího článku týkajícího se sociálních médii může být sociální médium/síť[38] definována jako webová služba, která „1. umožňuje vytvořit si veřejný nebo semi-veřejný uživatelský profil v rámci služby, 2. umožňuje vytvořit si seznam uživatelů, se kterými jsme ve spojení a 3. sledovat jejich spojení a spojení ostatních

[37] Podíl dospělých občanů ČR zapsaných v registrech SOLUS je nejnižší od roku 2011. Získáno 21.5.2018 z https://www.solus.cz/zpravy/podil-dospelych-obcanu-cr-zapsanych-v-registrech-solus-je-nejnizsi-od-roku-2011/

[38] Vzhledem k postupné adaptaci pojmu sociální síť pro sociální média budeme oba pojmy v této části práce používat jako synonyma. Mluvíme-li ale v pozdějších kapitolách o metodě „analýza sociálních sítí“, myslíme tím metodu SNA, aplikovatelnou na všechny typy relačních dat, nikoli pouze data sociálních médii.

v rámci systému (Boyd a Ellison, 2007, str.211).Co ale činí sociální média jedinečnými je, že umožňují uživatelům expresivně vyjádřit jejich sociální síť, protože uživatelé primárně komunikují s lidmi, kteří už součástí jejich skutečné sociální sítě jsou (str.211). První sociální síť vyhovující těmto kritériím – Six Degrees.com, název evokující známý výzkum o šesti stupních volnosti sociálních kontaktů - vznikla v roce 1997, následována o dva roky později Live journal, Asian Avenue až po finální Twitter a všem přístupnou verzi Facebooku, kterými článek končí (str. 212). Mnoho v článku zmiňovaných sociálních sítí už v dnešní době neexistuje, v současnosti v euro-americkém okruhu patří mezi nejpopulárnější Facebook a Twitter.

Podle ČSÚ (2017) mělo koncem roku 2016 v ČR 41% lidí starších šestnácti let (tedy celkem 3.6 miliónu lidí) aktivní profil na sociálních sítích (a využili ho alespoň jednou v posledních třech měsících). Údaj je velmi věkově podmíněn, v populaci 16 – 24 tvořil tento podíl 95% oproti populaci 65+, kde byl pouze 15%.

Detailnější přehled o využívání sociálních sítí v ČR přináší agentura focus, Marketing a Media Research (2016). 46% z dospělé populace vlastní účet, 21% na dvou a více sociálních sítích (11% populace pak na třech a více). Nejpopulárnější sociální sítí je Facebook, kde má účet 42% z dospělé populace, následován s velkým odstupem YouTube[39] a Google+ (shodně 12%), Instagramem (8%) a Twitterem (4%). Facebook má nejen největší základnu účtů, ale patří také k nejnavštěvovanějším, 22% z dospělé populace ČR jej navštěvuje denně.

K odlišným údajům došla agentura Stem/mark při realizaci komerčního výzkumu AMI Digital index[40], podle které mělo v roce 2016 účet ne třech a více sítích dokonce 39% populace, v roce 2017 pak dokonce 69%. Všechny údaje byly zjištěny dotazníkovým šetřením (tedy nikoli analýzou uživatelských dat a jejich extrapolací).

Zkraje kapitoly také musí zaznít upozornění na problematiky falešných účtů. Údaje o předpokládaném počtu se i v rámci jednotlivých sítí natolik liší, že nemá smysl uvádět

[39] YouTube je skutečně díky chování uživatelů považováno za sociální médium.

[40] Crha, V. AMI Digital Index: Češi používají stále více sociálních médií, průměrný čas trávený na síti ale klesl. Získáno 8.4.2018 z http://www.amidigital.cz/digikydy/ami-digital-index-cesi-pouzivaji-stale-vice-socialnich-medii-prumerny-cas-traveny-na-siti-ale-klesl/

přesná procenta. Pro každou platformu lze nalézt doporučení, jak falešný účet rozeznat, založenou na triviálních indikátorech (chybějící fotka, fotka z databáze, apod) i sofistikovanějších modelech chování (obsah zveřejňovaných příspěvků, analýza přátel, návštěvnosti apod.)[41].

Pro pořadí představování dat byla zvolena jako ukazatel obliba sociálního média mezi výzkumníky daná přístupností, transparentností a důležitostí platformy.

5.6.1 Twitter

Sociální sít Twitter patří ze zřejmých důvodů k nejvytěžovanějším zdrojům dat, jde ze své podstaty o otevřenou platformu, kde každý příspěvek, tzv. tweet, zpráva o délce 280 znaků, je okamžikem svého publikování veřejně přístupná. A to nejen uživatelům sociální sítě, příspěvky mohou být na profilu autora sledovány i bez založení účtu. Uživatelé pak mají rozšířenější možnosti, mohou ostatní uživatele tzv. sledovat (tedy nové příspěvky vybraných uživatelů se jim zobrazují, stávají se tak tzv. „followery"), případně tweet sdílet (tedy tzv. „retweetovat"), okomentovat jej a vyjádřit mu podporu. Pro zpřehlednění komunikace se v síti vžilo používání tzv. hasthagů, #, kterými mohou uživatelé v rámci textu uvodit téma, ke kterému se daný tweet vztahuje.

Specifika a benefity této sítě pro sociálně vědní výzkum shrnuje nejkomplexněji na základě rešerší výzkumu Felt (2016). Twitter je viděn jako alternativní zdroj informací nekorigovaný neomezenou mocí editorů nebo obchodními zájmy vydavatelů (Lewis, str. 3) (jen pro korektnost doplňme, že i proud Twitterových zpráv podléhá výběrovému algoritmu, tedy představa naprosté nekorigovanosti není zcela přesná), umožňuje poklepávat na ducha doby internetu a jeho uživatelů a zachycuje diskurz obyčejného života (Zimmer a Proferes, str.3) a umožňuje dát hlas marginalizovaným skupinám (Papacharissi, str. 3). Je to zároveň sociální síť i informační proud okolo (Bruns a Burgess, str. 3).

Nejen kombinací těchto faktorů, ale i transparentním přístupem patří Twitter mezi nejpříjemněji analyzovatelné sociální sítě. Jak zaznělo v nadsázce od nejmenovaného

[41] Samozřejmě netvrdíme, že chybějící fotografie u uživatelského účtu znamená falešný profil, indikátory musí být vyhodnocovány v souvislostech.

účastníka konference Nových médii "běžte z Facebooku pryč, běžte na Twitter... Tam můžeme vaše data líp stahovat" (Šlerka, 2018).

Politika přístupu poskytování dat je ale stále restriktivnější, jen od ledna tohoto roku došlo k několika principiálním změnám, které nejen omezily rozsah volně poskytovaných dat, ale tím i defacto zamezily fungování některých standardních cest k jejich získání (některé balíčky nejsou v tuto chvíli aktuální, bude vysvětleno dále), proto budou jednotlivé způsoby zisku dat z Twitteru jen načrtnuty s odkazy na případné relevantní zdroje.

Data Twitteru lze získat několika způsoby – pomocí programování za využití tzv. API přístupu (application programing interface), s využitím již existujících analytických aplikací (leckdy volně přístupné aplikace vyvinuté v rámci univerzit) nebo se lze podívat na již agregovaná data, která Twitter dává k dispozici.

Twitter hlídá přístup pomocí tzv. autentikačních údajů, které musí být součástí přistupujícího programu. O tyto údaje se žádá v rámci klasického uživatelského účtu, kdy uživatel požádá o vytvoření tzv. aplikace, jejíž údaje pak využívá při každém svém přístupu k API Twitteru.

API přístupy jsou popsány na webových stránkách společnosti určených vývojářům, https://developer.twitter.com/en.html, s nejasnostmi se lze obrátit na k tomu určené diskuzní fórum, https://twittercommunity.com/, s podporou zaměstnanců Twitter. Pro programování lze využít některého ze standardních jazyků (např. Python), které pro integraci s Twitterem obsahují k tomu určené balíčky (např. v rámci Pythonu to jsou tweepy, searchtweet, python-twitter atd.). Mnoho funkčních a aktuálních příkladů lze nalézt na sdíleném uložišti tzv. open source software https://github.com/ . Vzhledem k dynamičnosti vývoje IT a závislosti na podmínkách jednostranně definovaných Twitterem, nelze samozřejmě funkčnost konkrétního programu (resp. balíčku) vždy v daný čas stoprocentně garantovat. Další možností programátorského přístupu je využití aplikace twurl, což je speciální druh tzv. curl aplikace přizpůsobené Twitteru.

Twitter nabízí v květnu 2018 více druhů API přístupů, od bezplatného poskytnutí dat 7 dní zpátky (bez garance úplnosti, s limity pro poskytnuté druhy informací i limity pro stažení per jednotlivý požadavek) po tzv. prémiový plný přístup k všem datům od

roku 2006. Prémiový přístup je samozřejmě placený a o jeho přiznání je třeba požádat. Toto rozdělení je letošní relativně čerstvou novinkou, která – soudě dle aktualizací, komentářů, atd. – si svou podobu stále hledá.

Důmyslnou možností nevyžadující programátorské zkušenosti je využití některé z volně přístupných aplikací. Například Netlytic (Gruzd, 2017) umožnuje stažení až 1000 Tweetu podle obsahu, hastagu nebo uživatele (samozřejmě s limitem podle práv uživatele, který musí být přihlášen svým Twitter účtem). Při přehledu Tweetu podle obsahu doplní nejen zprávu samotnou, ale datum publikace, počet Tweetu uživatele celkově i počet jeho sledujících. Přímo v aplikaci pak jdou nad souborem spustit jednodušší analýzy.

Jaká data tedy Twitter dával volně k dispozici (a nyní jsou všechny dostupná v prémiovém přístupu)?. Tweet object, User object, Twitter entities a Twitter extended entities, zahrnující údaje o uživateli a jeho účtu, obsahu Tweetu, časové údaje, geolokační údaje (přidané uživatelem nebo automaticky vygenerované) atd.

Z dominantních oblastí výzkumů Twitteru – žurnalismus, katastrofy, aktivismus a zdraví (Murthy citováno ve Felt, 2016), se sociologií nejvíce rezonuje aktivismus, kde k nejznámějším pracím známým i mimo odbornou veřejnost, patří bezesporu analýzy M. Castellse (2012), ukazující přímý vliv používání informačních technologii včetně Twitteru na sociální události a organizaci hnutí typu Arabské jaro či Occupy Wall Street.

Mimo tyto explanativní (tedy zpětně vysvětlující) analýzy zmiňme například pozoruhodnou analýzu vykonanou v rámci Massachusetts Institute of Technology (Kallus, 2014), predikující vznik, čas a místo vzniku událostí (typu velké protesty nebo cyber-aktivismus) nestátních aktérů na základě zpráv z mainstreamových médií, vládních zpráv a aktivity na sociálních sítích, především Twitteru.

Twitter může být využitý například k určení míry radikalismu, kdy kombinací indexu radikality z příspěvků a lokace byla vytvořena tzv. „teplotní mapa“ radikalismu per jednotlivé provincie v Indonésii (Mazumder, Das, Kim, Gokalp, Sen, & Davulcu, 2013).

Studii zaměřenou na dynamiku toku informací a roli aktérů v rámci komunikační sítě přináší Tinati, Halford, Carr, a Pope (2014). Analýzou hastagu věnovaného studentskému protestnímu hnutí ve Velké Británii definovali jednotlivé klíčové role (např. agregátor, který je sám víceméně bez vlivu může být důležitou spojnicí oddělených částí sítě) a proces sdílení (mimo jiné potvrdili koncept „preferenčního připojování“). Článek sám je kombinací metodologického pojednání a studie a vrátíme se k němu i v části Critical data studies.

Vyčerpávající přehled studií na pomezí psychologie založený na sociálních sítích, především Twitteru, přináší Tuna et al. (2016). Analýza dat Twitteru může kupříkladu identifikovat světové oblasti, které zajímají více lokální a které více globální témata (Zheng, str.15) , identifikovat korelující témata (Grimaudo, str. 16), ukázat, že se členy své skupiny používáme v komunikaci jiný jazyk než s nečleny apd.

Z českých výzkumů Twitteru jmenujme například diplomovou práci zkoumající inkluzivitu a exkluzivitu českých novinářů na této síti (Krsová, 2018), dokumentující mimo jiné „posunutí role novinářů v online prostředí do pozice ambasadorů značky sama sebe nebo média, pro které píší“ (str. 75).

5.6.2 Facebook

Facebook je semiuzavřenou sociální sítí, kde se uživatel stránky sám může rozhodnout, co bude sdílet veřejně a co pouze se svými přáteli. Uživatelem je buď fyzická osoba (stránka osobní) nebo stránka patří nějaké instituci/události/produktu/firmě. V rámci svých aktivit může uživatel komentovat, sdílet, označit náladovým emotikonem příspěvky přátel nebo veřejně publikované. Může také vybrané stránky sledovat, případně označit jako „to se mi líbí“, být členem nejrůznějších uzavřených/otevřených skupin. Dle vlastního uvážení může na FCB vyplnit osobní informace a nastavit si úroveň sdílení informací o svém uživatelském účtu.

Přesné údaje o počtu a chování uživatelů Facebook zveřejňuje ve výroční zprávě pouze globálně[42], regionální údaje za Českou republiku tak máme k dispozici pouze

[42] Facebook Reports Fourth Quarter and Full Year 2017 Results. Získáno 8.4.2018 z https://www.prnewswire.com/news-releases/facebook-reports-fourth-quarter-and-full-year-2017-results-300591468.html

z novinářských zpráv, podle kterých měl ve třetím čtvrtletí roku 2017 4.9 miliónu aktivních uživatelů měsíčně, denně pak průměrně 3.8 miliónu[43].

Představu o množství a struktuře uživatelů si ale lze udělat z rozhraní pro zadávání reklamy, kdy po upřesnění kritérii (lokace, platforma, jazyk atd) uvádí počet aktivních uživatelů, kterým bude reklama zobrazena. Z těchto údajů vidíme, že geograficky v České republice, v češtině je platforma Facebook nejsilněji a ve věkové kategorii 25-34 let (1.4 milionu uživatelů), ale ani nejméně zastoupená skupina 55+ není zanedbatelná (0.6 milionu uživatelů). Celkově pak Facebook předpokládá zobrazení reklamy 5.1 miliónu lidem (údaj k 8.4.2018).

A jaká data tedy Facebook o těchto uživatelích schraňuje a potenciálně jsou k dispozici k analýze? V průběhu psaní této práce pronikla na veřejnost kauza Cambridge analytika (CA). Byť při bližším průzkumu nic dramatického ve skutečnosti CA neprovozovala a problémem tak bylo spíše podvodné postoupení akademické licence ke komerčnímu využití (pro připomenutí bývalý zaměstnanec obvinil CA, že využila data uživatelů Facebooku k ovlivnění Amerických voleb), celá záležitost si kvůli zostřené a zjitřené pozornosti veřejnosti vyžádala proaktivní restriktivní opatření Facebooku. Byť množství poskytovaných informací Facebook omezuje průběžně, v důsledku CA aféry opět značně přiostřil, nejen do rozsahu, ale i jednoduchosti provedení.

Velkou nevýhodou získávání dat z Facebooku je také jeho netransparentnost. Není zcela jasné, proč požadovaná data nevrátil, případně zda je vrátil všechna. Například čerstvě omezil počet získaných příspěvků v rámci dotazu na „600 publikovaných příspěvků za rok“[44], ale zcela chybí vysvětlení, podle jaké logiky je vybírá (chronologicky?, nejvíce komentované?, nejvíce sdílené? apod.)

K datům se přistupuje pomocí Graph API. Prvním krokem je opět vytvoření aplikace na rozhraní FCB https://developers.facebook.com/apps/ , kde se vygenerují přístupové údaje (ID aplikace a security data). Nejjednodušším způsobem lze o data

[43] Hušková, L. V Česku vyrostl počet denních uživatelů Facebooku na 3,8 milionů. Získáno 8.4.2018 z https://newsfeed.cz/v-cesku-vyrostl-pocet-dennich-uzivatelu-facebooku-na-38-milionu/

[44] Graph API version Získáno 1.6.2018 z https://developers.facebook.com/docs/graph-api/reference/v3.0/page/feed

požádat v Graph API exploreru, kde po vygenerování bezpečnostního tokenu lze jednoduchou dotazovou syntaxí získat některá data veřejně přístupných stránek (například datum a publikování příspěvků, komentáře, atd.). Případně lze napsat skript v některém z příslušných jazyků, např. v jazyce Python pomocí balíčku facebook-sdk, v R pomocí Rfacebook apod.

Využít lze také některou z volně dostupných FCB aplikací, například Nettviz (Rieder, 2013)[45], která po zadání identifikátoru FCB stránky nabízí stažení informací do příjemného tsv formátu. Umožní stáhnout informace o jednotlivých příspěvcích, konkrétně datum publikace, typ (obrázek, text), počet „like", komentářů a sdílení, případně souhrny na denní bázi. Kvůli novým pravidlům však umožňuje získat jen informace o veřejně sdílených příspěvcích, které se mohou pro jednotlivé uživatele aplikace lišit (je-li nebo ne fanoušek stránky). Uživatelsky velmi příjemné jsou i výstupy aplikace Netlytic (Gruzd, 2017), představené už v sekci Twitter, která stahuje veškeré příspěvky, komentáře, včetně obsahu, data publikování a počtu „like". Omezena je opět počtem 1000 stáhnutelných jednotlivých akcí (tedy přidání příspěvku nebo komentáře).

Díky dlouhodobé existenci Facebooku, jeho popularitě mezi uživateli a původně vstřícnému přístupu k získávání dat existuje množství dobře známých prací věnujících se problematice využívání dat. Uvést tak stačí jen jednu z nejvíce kontroverzních a nejvýznamnějších. Je ji bezpochyby výzkum Kramera, Guilloryho a Hancocka (2014), kdy na náhodně vybraném vzorku uživatelů manipulací zpráv zobrazovaných na feedu (potlačování pozitivních v prospěch negativních a naopak) zkoumali, zda funguje tzv. emocionální nákaza (funguje, jakkoli účinek je malý, ale vzhledem k rozsahu sociálních sítí „mohou i malé efekty mít rozsáhlé agregované důsledky", str.5).

5.6.3 Ostatní sociální sítě

Sociálních sítí je ale celá řada, Linked In, Youtube, Google +, Instagram apod. Jedná se většinou o součást rodiny produktů velkých korporací a poskytování dat se řídí

[45] Informace obsažené v dokumentu však nejsou aktuální, autor ji na stránkách nicméně uvádí jako doporučované ozdrojování aplikace.

politikou vlastníků, jak je Microsoft, Google, Facebook. Každá z nich má své dedikované API, které umožňuje programátorský přístup k datům popsaný výše.

Samozřejmě lze využít i aplikace, například částečný přístup k datům nabízí i Netlytic (Gruzd, 2017). Umožňuje stáhnout až 15 000 komentářů k vybranému videu na Youtube, obsahující označení autora komentáře, datum publikování a text komentáře. Umožní také získání až 2500 příspěvků z Instagramu, včetně adresy příspěvku, autora, data publikace a popisu. Vybrat si lze příspěvky nejen podle tematického obsahu, ale i podle geografické oblasti publikování.

Data nejsou natolik analyzována, ale například výzkumem Youtube se zjistilo, že jen malá část uživatelů chce vyjádřit své přesvědčení veřejně (Kohli et al., str. 15).

5.7 Data státu a obcí

Data státu bývají nazývány taktéž administrativní data (Kitchen, 2014). Jde o data průběžně sbírána a aktualizována státem, jako například registry nezaměstnanosti, registr insolvence, matriční záznamy apod. Stejně tak zde mohou být zařazena i data, zveřejňována městy v rámci iniciativy otevíraní dat. Program zaštiťován Ministerstvem vnitra, přináší na webové stránce https://data.gov.cz/ soubor všech volně přístupných dat, včetně tutoriálu pro poskytovatele i potenciální uživatele dat (specifikují, jaké jsou formáty, kde lze o další data požádat, co je k dispozici apod.).

5.8 Data vlastníků kamerových systémů

Rozsáhlými daty disponují také provozovatelé kamerových systémů, tedy orgány státu (typicky například městská policie na základě zákona o polici, data pak v takovém případě spravuje obec) ale i jiné právní (zaměstnavatelé, společenství vlastníků apod.) a fyzické subjekty (vlastníci rodinných domů apod.). Kamerové systémy ukládající záznam podléhaly do 25.5.2018 zákonu o ochraně osobních údajů[46] (od tohoto data pak podléhají Obecným nařízením o ochraně osobních údajů GDPR) a měly být hlášeny Úřadu na ochranu osobních údajů, spolu se specifikací

[46] K tomuto datu byl nahrazen Obecným nařízení o ochraně osobních údajů (GDPR), která již registraci neukládá. Ostatní je obdobné. Zdroj: K provozování kamerových systémů. (2.5.2018) Získáno z https://www.uoou.cz/k-nbsp-provozovani-kamerovych-systemu/d-29535/p1=1099

účelu pořízení. Omezená na „nezbytně nutnou dobu“ byla také možnost skladování těchto dat. (ÚOOÚ, 2012)

Oficiální počet kamer monitorujících veřejný prostor v ČR nelze z otevřených zdrojů zjistit, pro představu jen Prahu monitoruje 924 „kamerových stanovišť“ Městského kamerového systému, který umožňuje „řešit specifické problémy v oblasti zajištění veřejného pořádku a pouliční kriminality“. Využíván je také k monitorování dopravy (Městský kamerový systém, 2015).

Ve všech zemích EU[47] jsou kamerové systémy regulovány přísnou legislativou, žádnými právními omezeními se naopak nezatěžuje například Čínská lidová republika (Gan, 2018). Sice odtud nepřicházejí teoretické sociologické studie založené na datech kamerových systémů, zato lze vidět použití těchto dat v praxi, v jednom z nejhorších příkladů sociálního inženýrství, chystaném projektu Social Credit System[48] (Botsman, 2017).

5.9 Data vlastníků senzorů

Sensorem je obecně myšleno zařízení, které je schopno měřit (a hodnotu posílat dále) nějakou veličinu. Typů senzorů tak dle měřených veličin existuje celá řada, rozděleny by mohly být na monitorující biologické hodnoty organismů, dále pak senzory monitorující přírodu (senzory povětrnostních podmínek, teploty apod.) a v neposlední řadě senzory monitorující vlastnosti živých organismů i neživých věcí (například senzory pohybu automobilu, apod.)

Data sensorů jsou hojně využívána ve zdravotnictví (například podle predikce společnosti IBM mělo být v roce 2014 přes 420 miliónů bezdrátových nositelných zdravotních monitorů[49],) pro budování konceptu „chytrého města“ eliminujícího

---

[47] Dlouho dobu bylo právo na soukromí v rámci EU podobné (Braunová, 2009), kvůli nebezpečí terorismu a snaze zbránit jim, ale v některých zemích dochází k rozvolňování (například VB). Otázkou je i skutečné dodržování, protože data z kamerového systému mají sloužit nejen k monitorování situace a rozeznávání obličejů, ale na základě detekce neobvyklého chování s využitím umělé inteligence mají útoky předvídat.

[48] Systém sociálního ohodnocení obyvatel, běžící dosud v dobrovolném módu, povinný od roku 2020. Systém je designovaný na odměňování (a tedy podporu) žádoucího chování (například zvýhodněním půjček apod.) a trestání (a tedy potlačování) chování nežádoucího (například zpomalováním internetu, zákazem cestování apod.)

[49] The Four V's of Big Data. Získáno 10.5.2018 z http://www.ibmbigdatahub.com/infographic/four-vs-big-data

dopravní zácpy, zlepšujícího veřejnou dopravu a veřejné služby vůbec. Množství těchto dat neustále poroste, především v souvislosti s rozvojem tzv. internetu věcí, tedy bezdrátového propojení více identifikovatelných zařízení.

6 Metody analýzy Big Dat a jejich specifika

6.1 Přehled

Při představování možných zdrojů Big dat v minulé kapitole vyhovující definici, že jsou „rychlá a n= všichni“ (Kitchen & McArdle, 2014), byly popsány jednotlivé metody jejich získání. Už při čtení muselo být pravděpodobně zřejmé, že některé datové soubory (například Tweety vybrané dle konkrétního obsahu v definovaném časovém období) se svou velikostí nikterak nevymykají v sociologii běžně analyzovaným souborům a nekladou zvláštní požadavky na metody a nástroje užité při analýze.

Big data v tomto širokém pojetí v sociologii spíš jen rozšiřují možnosti (například možnost cílených experimentů) a více využívají klasické, ale dosud spíše okrajovější analytické metody (například síťové modelování).

Například McFarland et al. (2016) rozeznávají čtyři základní směry výzkumu Big dat, kterými jsou informace obecně (tedy nejen v sociálních vědách, ale i ve firmách) z dat získávány

1. výpočetní lingvistika („computational linguistics“) – tedy přístup vzniklý prolnutím cílů lingvistiky, umělé inteligence a kognitivních věd, který pomáhá v rozvoji sociolingvistiky a analýzám obrovského množství textu.

2. síťové vědy („network science“) – tedy přístup spojující dohromady poznatky počítačových vědců, fyziků a sociální vědců zabývajících se sítěmi zabývající se především sociálními médii

3. strojové učení („machine learning“) – tedy spíše inženýrský přístup pro objevení vzorců v datech

4. experimentální metody – přístup zkoumající, jaké změny v sociálním prostředí vyvolají požadovaný účinek

Toto rozdělení sice pojmově lehce nešťastně plete metody (výpočetní lingvistika, síťové vědy), nástroje (strojové učení) a design výzkumu (experimentální metody), ale

názorně dokumentuje tezi o Big datech jako spíše o rozšířených možnostech – zisku, kontroly, analýzy.

Podobně vidí benefit Big dat i Chang, Kauffman, & Kwon (2014), kdy dle autorů Big data umožňují vystoupit nad obvyklé třírohové dilema výzkumu, kdy tři perspektivy - tedy obecnost nalezených závěrů („generality“), kontrola nad měřením a jeho přesnost („control“) a kontext („realism“) - v různých typech výzkumů vzájemně soupeří o dominanci, kterou může vždy jedna získat pouze na úkor ostatních (tedy kontrola na vrub obecnosti apod., str 4).

Protože klasické způsoby analýzy (včetně okrajovějších) jsou běžně vyučovány v povinných nebo volitelných kurzech v rámci studia sociologie, tato kapitola se zaměří především na analýzu datových souborů, které skutečně vynikají jedním z rozdílových rysů Big dat, velikostí. Jak ale definovat hranici?

6.2 Vymezení (a omezení) kapitoly

Echort (2018) uvádí, že zásadní otázkou v každé fázi výzkumu Big dat je zodpovědět si, zda si při jednotlivých krocích poradíme s konvenčními analytickými nástroji, nebo musíme využít metody jiné. A právě tato otázka nám v každé fázi ohraničí hlavní předmět kapitoly, skutečně velké soubory Big dat. Pro odlišení budeme tento typ dat v této kapitole označovat jako Big Data.

Dále by pro upřesnění mělo explicitně zaznít, že některé pasáže kapitoly popisují perspektivu uživatele PC. Mohutné zázemí komerčních firem umožňuje sofistikovanější přístup, který je ale podporován několika nezávislými IT odděleními (oddělení informačních technologií) a tato práce si neklade ambice plně popsat jejich řešení. Někdy je ale přibližuje z edukativních důvodů, neboť terminologie a metody popsané dále jsou univerzální.

6.3 Jak si poradit s velikostí

Existuje několik osvědčených klasických způsobů, jak si poradit s Big daty v závislosti na zvoleném analytickém nástroji (předpokládejme v tuto chvíli již optimalizovaná data, techniky optimalizace jsou řešeny dále). Můžeme zvětšit výpočetní kapacitu přechodem na šedesáti čtyř bitový procesor (což nám teoreticky umožní práci se souborem o více než dvou miliardách řádků), přidat externí paměťové jednotky, řešit

úlohu postupně na podvýběrech (pokud to umožňuje algoritmus analýzy) nebo přidat CPU (centrální procesorovou jednotku). Přidání CPU se obecně nazývá „parallelizing“ a umožňuje souběžný běh výpočtů. Nejjednodušší formou je například počítač s tzv. vícejádrovým procesorem a využití některé z funkcí, které umožní části výpočtu běžet zároveň.

Pokud všechny tyto metody selžou, jedná se beze vší pochybnost o Big Data vyžadující sofistikovanější verzi paralelního zpracování.

MapReducing je nástroj/přístup, který dovoluje paralelní zpracování úloh na více systémech, přičemž zajišťuje veškerou datovou komunikaci a distribucí úkoly mezi systémy (Map) a výslednou syntézu informací (Reducing). Tato logika může být implementována ve více programech/nástrojích, například Python i R. Je hlavních principem ekosystému Apache Hadoop, který se dle diskuzních fór stal téměř synonymem pro řešení problému Big Dat.

Apache Hadoop[50] je volně přístupná rodina softweru, umožňující skladování Big Dat i výpočty nad nimi. Vyžaduje nicméně programátorské zkušenosti s Javou s její obtížnější syntaxí (programovací jazyk), případně využití překladače z některého z dalších jazyků (například Jython pro Python). Neumožňuje implementaci některých algoritmů strojového učení a díky své architektuře je spíše vhodný pro dávkové zpracování. Za jeho rozšířením nicméně stojí fakt, že může být využit jako uložiště a nad ním mohou fungovat aplikace vytvořené v některém z dalších programovacích jazyků jako je Python nebo R.

Velmi výrazně se prosazující alternativou paralelního zpracování k Apache Hadoop eliminující její nedostatky je ApacheSpark. Díky odlišné logice implementace tzv. generalizovanému počítání („generalized computation“) pracuje efektivněji s pamětí (načítá si data do mezipaměti a proto při opakovaných operacích nad daty nemusí tato opětovně číst z pevného disku), což umožňuje rychlejší zpracování. Umožňuje také integraci s programy Python, R, Java, SLQ, HiveSQL apod. Slouží ale jen

[50] What is the difference in idea, design and code, between Apache Spark and Apache Hadoop? Získáno 10.6.2018 z https://www.quora.com/What-is-the-difference-in-idea-design-and-code-between-Apache-Spark-and-Apache-Hadoop

k výpočtům nad daty a nezahrnuje uložení. Umožňuje ale integraci s uložištěm Hadoop.

Jinou strategii k zvýšení výpočetního výkonu je využití tzv. GPU („General Process Unit“), tedy grafického procesoru. Tento velmi hrubě řečeno představuje analogii klasického CPU (tedy hlavního procesoru počítače), ale GPU byly vyvinuty pro grafické zpracování, tudíž jsou designovány primárně pro práci s vektory a maticemi. Dokážou tak data zpracovávat řádově rychleji (a tento rozdíl téměř exponenciálně roste s rostoucí velikostí dat[51]). Programovacím jazykem je CUDA, což patří k drobným nevýhodám, protože jde o jednoduchý skriptovací jazyk vyžadujícím tím pádem mnohem sofistikovanější způsob programování (paradox je pouze zdánlivý).

6.4 Postup při práci s daty

Fayyad (jak je citováno v Tsai, Lai, Chao, & Vasilakos, 2015) shrnuje jednotlivé etapy zisku informací z dat do tří stádií

1. vstupní fáze - zahrnující zisk dat, selekci, přípravu („preprocessing“) a transformaci

2. analýza dat – tedy „data mining“ (vytěžování informací z dat)

3. výstupní fáze – skládající se z vyhodnocení a intepretace

6.4.1 Vstupní fáze

Cílem vstupní fáze je získat data optimalizovaná vzhledem k analýzám, které na nich mají být provedeny.

Techniky zisku dat s ohledem na jejich jednotlivé typy byly popsány v předchozí kapitole, nezodpovězena ale zůstala otázka skladování. V případě Big dat může jít o libovolný formát pro uložení na pracovní stanici (PC), Big Data, pohybující se obvykle v terabytech, vyžadují sofistikovanější řešení. Formátově může jít o databázi uloženou v externím uložišti, případně na více uložištích. Tato uložiště mohou být privátního charakteru (tzn. například jen přidání více zařízení s pamětí propojených vnitřní sítí)

[51] Is GPU computing suitable for big data analytics? Získáno 10.6. 2018 z https://www.quora.com/Is-GPU-computing-suitable-for-big-data-analytics

nebo tzv. cloudu. Cloudem je obecně myšlena virtuální služba poskytovaná přes internet, ve formě PaaS (platforma jako služba, různé typy databází relačních i neralečních databází jako SQL, NoSQL apod.) nebo IaaS (infrastruktura jako služba, Apache Hadoop, Microsoft Azure apod.). S externím datovým uložištěm je potom komunikováno skrz privátní/veřejnou síť za pomocí adekvátně zvoleného programovacího jazyku.

Například ve firemním prostředí je jedním z typů cloudových řešení tzv. Data lake, typ ukládání firemních dat v jejich původním formátu, kdy v jednom datovém uložišti spolu mohou koexistovat strukturované databázové formáty, semistrukturované data (CSV, logy, JSON,..),nestrukturovaná data (email, PDF,..) i binární data (audio, video,...). Příkladem je například Apache Hadoop, Azure Data Lake nebo Amazon S3. Smyslem je uchovat vše pro možné budoucí použití. Nad tímto "jezerem" je potom aplikována průřezová logika, dovolující další práci s daty. Firmy mohou mít toto řešení provozované na svých vlastních serverech (tzv. inhouse řešení), nebo umístěné na dedikovaných cloudech. (alternativou skladování firemních informací jsou tzv. Data warehousy, řešení, které vyžaduje data strukturalizovaná, založená na relačních databázích, už předpřipravená pro reporty.)

Součástí vstupní fáze je spojování různých nestrukturalizovaných dat z různých zdrojů a převod na jednotný formát, který většina dalších analýz vyžaduje, stejně tak čištění dat od různých nekonzistencí.

A v neposlední řadě zde patří úprava dat tak, aby následná analýza proběhla v co nejkratším možném čase, případně s co nejmenšími nároky na operační systém. Tyto úpravy mohou jít dvěma směry, buď zmenšením počtu případů (omezením na partikulární podmnožinu) nebo zmenšením rozsahu při zachování počtu případu, tedy dimenzionální redukcí. Lze například vyřadit proměnné, které pro budoucí analýzu nejsou podstatné, případně změnit datový formát jednotlivých polí na úspornější (například pole typu string zabere více paměti než pole typu číslo, tato transformace je vhodná u polí obsahujících datum). (Tsai et al., 2015)

Je osvědčenou metodou vybrat si malý vzorek dat, na kterém si lze data tzv. osahat a prozkoumat a vyzkoušet. Aplikace na celá Big Data potom obvykle vyžaduje modifikaci klasických algoritmů i přístupu.

Dle potvrzení z více zdrojů se zkušeností z praxe s Big Daty, zabere přípravná fáze obvykle 80-90% celého analytického procesu.

6.4.2 Analýza dat

Pod analýzou Big Dat si lze představit rozšíření klasických algoritmů, případně lze využít technik strojového učení, které si představíme v samostatné podkapitole – Analýza dat- machine learning.

Přehled možností, jak využít klasické algoritmy, přináší s odkazy na dosavadní provedené studie Tsai et al. (2015).

1. shlukovací algoritmy – klasické shlukovací analýzy představují pro Big Data velkou výzvu, neboť vyžadují, aby byla data v jednotném formátu a v jednom datovém uložišti. V rámci Big Dat lze využít řešení CloudVista pro paralelní zpracování nebo GPU. Jako příklad užití CloudVisty nicméně uvádějí spočítání výsledků národního censu s 25 milióny případů.

2. klasifikační algoritmy – opět je jedním z řešení využití paralelního zpracování, případně autoři uvádějí „quantum based vector machine“ algoritmus, který ale chápeme spíše příklad „machine learning“ přístupu (viz dále). Obecně jsou ale velmi slibné naděje na využití algoritmu založených na kvantových výpočtech, „jakmile bude hardware v rámci kvantového počítání dostatečně připravený“.

3.frequent pattern miming (algoritmus pro určení vzorců z dat) – opět nepřekvapivě dominují přístupy využití paralelního zpracování, konkrétně MapReduce přístup.

6.4.3 Výstupní fáze

Interpretování a zobrazování výsledků Big dat se vesměs neliší od klasických interpretací, vizualizační metody bývají běžnou součástí softwarových nástrojů, případně existují samostatně. V rámci komunity zabývající se analýzou Big Dat je oblíbeným vizualizačním nástrojem například Tableau nebo QlikView, v rámci jazyků programovacím jazyků pak jednotlivé příslušné balíčky (např. ggplot2 pro R, Bokeh pro Python apod.)

Odlišná může být jen míra zisku informací. Specifickým doprovodným efektem analytických metod využívaných u Big Dat je speciální druh nejistoty, který nemá

analogii v klasických analýzách. Burrell (2016) rozlišuje tři základní nejistoty, se kterými se lze při analýze obecně setkat a to v důsledku:

1. „secret code“ - představuje nejistotu ve formě tajemství, kdy je účelem zachování kompetitivní výhody; tuto nejistotu lze eliminovat zpřístupněním kódu nezávislým skrutátorům.

2. „complexity“ -znamená nejistotu ve smyslu porozumění kódu, což je vysoce expertní znalost; možností obrany je v tomto případě rozšíření technického vzdělání.

3. „black-box“ - přináší fundamentální nejistotu v důsledku využití strojového učení.

Fundamentální nejistota plyne z toho, že v rámci těchto metod není k dispozici přesná informace o průběhu analýzy, k dispozici je pouze výsledek, jde o takzvaná „black-box“ řešení. Nedokáže se například zjistit, proč systém rozhodl o zařazení konkrétního případu v rámci klasifikační analýzy, tak jak rozhodl. Jde o specifickou vlastnost takzvaných machine learning metod (metod strojového učení).

6.5 Analýza dat – machine learning

Machine learning (strojové učení, používanější je ale spíše zkratka ML) je jeden z typů tzv. umělé inteligence („artificial intelligence“, AI). Z hlediska oboru se pohybuje na rozhraní statistiky a „computer science“. Algoritmy v tomto pojetí nejsou programovány jako série přesných příkazů, ale díky využití statistických metod dokáží vyhodnocovat data samy, bez explicitního zadání. „Místo, aby využívaly algoritmy na vysoké úrovni k vyřešení problémů v explicitní, imperativní logice, aplikuje [ML] jednoduché algoritmy, aby zjistily vzory implicitně v datech obsažené“.[52]

Jednou z největších výhod ML je přístupu je možnost tzv. učení bez učitele, tedy unsupervised learning, kdy cílem algoritmu je „prozkoumat data a sám najít nějakou vnitřní strukturu“.[53]

Pro korektní upřesnění by mělo být zdůrazněno, že představené metody nejsou metody určené výhradně pro analýzu Big Dat, jde o tradiční metody Data minigu

[52] How are big data and machine learning related? Získáno 10.6.2018 z https://www.quora.com/How-are-big-data-and-machine-learning-related

[53] https://www.sas.com/cs_cz/insights/analytics/machine-learning.html

(tedy získávání informací z dat) a lze je aplikovat i na menší soubory. Ale právě velikost Big Dat umožnila jejich skokový rozvoj, především tam, kde se uplatňuje učení, které nejlépe funguje právě na velkých datových souborech.

6.5.1 Neuronové sítě

Neuronová síť (NN) je druh algoritmu, který je volně inspirován stavbou mozku (odtud název), kde spolu přes několik různých vazeb (synapsí) komunikují neurony (Fonseca & Cabral, 2017).

Neuronová sít se skládá ze tří typů vrstev („layer“),

1.vrstva vstupní, která nese vstupní informaci

2. vrstva(y) skrytá(é) („hidden“), kde probíhají logické operace. V této úrovni může být řazeno několik vrstev za sebou.

3. výstupní vrstva, které nese výsledek

Všechny neurony v jednotlivé vrstvě jsou spojeny se všemi neurony ve vrstvě následující. Neuronové sítě jsou pak charakterizovány rozložením (tedy počtem vrstev a neuronů v nich) a váhami, které značí preferenci jednotlivé synapse vůči ostatním. Právě tyto váhy jsou na začátku nastaveny náhodně a s každým průchodem prvku v síti se vyhodnocuje chybovost a váhy se samy iteračně upravují. Průchodu prvků se známým výsledkem se říká „trénování sítě“ a schopnosti opravovat si váhy pak „učení se“ (v uvozovkách, aby upozornily na přenesenost významu tohoto slova nikoli jeho doslovnou interpretaci tak, jak je chápána například v psychologii).

NN jsou typicky trénovány na tvz. trénovacím souboru. Nejrozšířenějším přístupem je potom rozdělení dat na tzv. trénovací množinu (tady se algoritmus učí), testovací množinu (kde se testuje kvalita nastavení a struktury) a validační množinu (pro výsledné ověření kvality modelu). Rozdělení se dat se může lišit, používá se například 50/25/25, případně 70/15/15. Jako nejlepší je potom „vybrán model, který nemá velké výkyvy na jednotlivých množinách“ (Ulrych & Jurczyk, 2014). Záleží ale samozřejmě i na počtu vstupních neuronů, například pro dosažení dobrého výsledku v rámci klasifikace by měl počet prvků v tréninkové množině třikrát přesáhnout počet vstupních vlastností (ergo neuronů ve vstupní vrstvě).

Výše popsaný model trénování platí pro takzvané „supervised learnig“ (učení se učitelem), kdy síť je trénovaná na známých výsledcích. Variantou je ale také tzv. „unsupervised learning“ (tedy učení bez učitele), kdy si síť sama koriguje nastavení podle jiných veličin než chyby, například podle typického představitele v rámci klasifikačních úloh, minimalizací post pravděpodobnosti apd. Tento typ sítě se ale nehodí na všechny úlohy, například predikativního charakteru.

Mezi největší přednosti NN patří schopnost „učit se“ a generalizovat. „Tedy zapamatovat si kombinace, které vedly k požadovanému výstupu […] na základě zkušeností odhadovat nový výsledek […] správně zareagovat i na vstupy, které nebyly součástí trénovacích dat, a vyvodit z nich obecné závěry o datech“. (Ulrych & Jurczyk, 2014)

Pro představu například v rámci klasifikačních NN by tak první vrstvu tvořily vlastnosti a poslední jednotlivé podskupiny. Pro ilustraci si načrtněme hypotetický analyzátor volebních preferencí, kdy prvek na vstupu – voliče - si potom můžeme představit jako vektor socioekonomických vlastností <pohlaví, věk, bydliště, …>, kde každá vlastnost je reprezentovaná jedním neuronem ve vstupní vrstvě. Po proběhnutí sítě volič zkolabuje do jednoho ze stavů volič strany xy, což si můžeme představit jako vektor tvořený nulami a jednou jedničkou, kdy každý potenciální výsledek „volič strany xy“ je opět reprezentován neuronem, tentokrát ale ve výstupní vrstvě.

U Big Dat se i v rámci NN využívá paralelního zpracování, kdy může být například rozdělena trénovací množina do více setů a výsledné váhy pak vzniknou zkombinováním jednotlivých výsledků. Při trénování se také využívá GPU, který dokáže významně snížit čas.

Význačným podtypem neuronových sítí jsou takzvané „Deep neural networks“, tedy neuronové sítě obsahující více skrytých vrstev, typicky 10-20. Čím jich je víc, tím větší je pravděpodobnost, nalezení „charakteristik […] ale také delší čas výpočtu a těžší trénování“ (Heller, 2017).

V rámci neuronových sítí se typicky řeší úkoly typu klasifikace, regrese, shlukovací analýza. Velkou výhodou NN je možnost aplikace na nelineární problémy. Na zvládnutí na skutečně dobré úrovni jsou však komplikované, nejen kvůli množství vstupních parametrů, ale především kvůli riziku tzv. přeučení, tedy nutnosti

vystihnout moment, kdy by další učení modelu už nevedlo k lepším výsledkům, ale naopak je zhoršilo.

Algoritmus neuronových sítí lze naprogramovat v mnoha již dříve představených jazycích, například v Python (pomocí balíčku numpy a pandas), v R (balíček neuralnet) apd.

6.5.2 Ostatní metody

Další v praxi využívaných metod ML je například „Support vector learning“, strojový algoritmus využívaný pro klasifikaci, regresi a nalezení odlehlých hodnot, také aplikovatelný na nelineární problémy. Zjednodušeně řešeno například při klasifikaci „vyzvedne“ odlišná data, proloží je diferencující nadrovinou (nadrovina je matematický pojem pro prostor o jednu dimenzi menší než původní prostor, tedy k rovině je nadrovina přímka, k třírozměrnému prostoru rovina apod.) a opět vrátí data zpátky. Cílem nejjednodušších SVM je právě nalezení této diferencující nadroviny. Podle odborníků z praxe však tento přístup vyniká i oproti NN značnou složitostí. Opět jsou ale aplikovatelná skrze jazyky Python nebo R.

Lze využít nástroje ve vědě běžně vyučované/používané?

6.6 Použití klasických nástrojů

Pro přehled využití klasických analytických nástrojů jsme byly vybrány nejpoužívanější nástroje ve vědeckých článcích na platformě Google Scholar do roku 2017 (Muenchen, získáno 6.5. 2018) :

1. SPSS statistic - suverénní vítěz s více než osmdesáti pěti tisíci články, s výrazným vrcholem v roce 2009, od té doby jeho použití stále strmě klesá

2. R - více než dvacet pět tisíc článků se stále rostoucí tendencí

3. SAS - téměř dvacet pět tisíc článků, pozvolný pokles od roku 2010

Tyto nástroje jsou také nejpoužívanějšími nástroji mezi datovými vědci podle průzkumu z roku 2015, jen v odlišném pořadí (R, SPSS Statistics, SAS), v podobném výzkumu „self-service“ analytických nástrojů bylo jen SAS nahrazeno Excelem.

(jedním z důvodů masivního využívání SPSS a SAS může být i jejich rozšíření na univerzitách)

6.6.1 IBM SPSS Statistics

IBM SPSS Statistics je zpoplatněný analytický nástroj společnosti IBM, nicméně s možností za mírný poplatek oproti klasické ceně využívat tzv. studentskou licenci. Podle výrobce zvládne 32-bitová verze až 2 miliardy případů, 64-bitová je potom limitovaná jen výkonem počítače[54]. Tyto predikce jsou ale velmi teoretické, závisí na volné operační paměti počítače a při velkých velikostech souboru výkon dramaticky klesá. Podle pročtených diskuzí bývá více zatěžující počet proměnných (tedy počet sloupců), což ale lze označit za známý důsledek tzv. prokletí dimenzionality („curse of dimensionality“, více dimenzí je početně náročnější).

V posledních verzích SPSS Statistics jde nicméně aplikace Big Datům vstříc. Aktuální verze 25 [55] přináší integraci s programy Python i R a v rámci nabízených analýz inzeruje i neuronové sítě. Data také mohou být čerpána z externích relačních databází.

A v rámci rodiny analytických nástrojů IBM SPSS jsou možnosti ještě širší[56]. SPSS Analytic Server umožňuje spojení s uložištěm Hadoop, na kterém může běžet SPSS Modeler, umožňující MapReduce.

6.6.2 SAS

SAS (Statistical analysis system)[57] je statistický nástroj i programovací jazyk, komerční produkt nabízející zdarma studentskou licenci. Umožňuje propojení s uložištěm Hadoop a s využitím MapReduce a databázových jazyků Hive a Pig. Pro oblast Big dat pořádá SAS certifikované kurzy (pro svou platformu).

---

[54] http://www-01.ibm.com/support/docview.wss?uid=swg21476061

[55] IBM SPSS Statistics. Získáno 10.6.2018 z https://www.ibm.com/products/spss-statistics

[56] Apply SPSS analytics technology to big data. Získáno 10.6.2018 z https://www.ibm.com/developerworks/library/bd-spss/index.html

[57] https://en.wikipedia.org/wiki/SAS_(software)

6.6.3 R

R je volně přístupná softwarová platforma a programovací jazyk, primárně vyvinutý pro využívání ve statistice. Jde o takzvaný open source běžící pod GNU General Public licencí, což znamená, že jeho kód je volně přístupný pro užití i další vývoj. A právě tato volnost v rozvoji je jeho poněkud slabší stránkou. R je ekosystém mnoha balíčku/knihoven, které si uživatel stahuje dle aktuální potřeby. Zdrojem pro stažení může programátorská platforma GitHub, k dispozici jsou balíčky komerčních firem (Microsoft, Oracle), dedikovaná platforma pro rozvoj R CRAN[58] (k 10.6.2018 obsahovala 12 619 balíčků) a další. R je velmi mocný analytický nástroj, ale jeho slabinu lze nejlépe vystihnout- „je jako polostrukturovaná knihovna s obrovským množstvím knih, kde chybí jasná katalogizace“ (Smirnova, Ivanescu, Bai & Crainiceanu, 2018, str.28). Pro Big Data je ale z klasických nástrojů nejlepší alternativou, neboť „pokud existuje nějaká technika [ve smyslu zisku informací], je pravděpodobně implementována v některém z R balíčku“ (Heller, 2017).

Velmi nahrubo lze říct, že standardně lze v R analyzovat data do miliónu záznamů. Analýza datových souborů do miliardy záznamů „vyžaduje dodatečné úsilí. A analýza dat nad miliardu záznamů už vyžaduje použití MapReduce algoritmu, který může být vytvořený v R a navázaný na Hadoop apod.“ (Wijffels, jak je citováno v Bracht, O., získáno 10.6.2018). Ale tyto odhady jsou spíše ilustrativní, vždy záleží především na velikosti výpočtů, které se ukládají do paměti (proto například některé typy analýz mohou být dělány na řádově větších souborech než jiné).

Speciálně pro analýzu Big dat jsou významné dva balíčky knihoven, oba lze označit za API (tedy application programming interface, tedy jakýsi vstupní bod umožňující vytvoření programu) k příslušným platformám - RHadoop a SparkR. RHadoop je designovaný pro platformu Apache Hadoop, umožňuje nakládání se soubory v uložišti HDSF, vytvoření MapReduce apod. RSpark je verze pro Apache Spark.

Uživatelsky příjemnou vstupní branou do R je tzv. RStudio, tedy aplikace umožňující programování v R bez nutnosti psaní kódu v příkazovém řádku, s kontrolou syntaxe, možností přímého spuštění a ukázání výsledku.

[58] Contributed packages. Získáno 10.6.2018 z https://cran.r-project.org/

6.7 Python

Python byl zmiňován na několika místech této diplomové práce, v rámci kapitoly věnované metodám analýzy Big Dat si proto zaslouží speciální místo. Python je volně přístupný programovací jazyk, který dle diskuzních fór soutěží s R o první místo v rámci analýzy. Například v souboru padesáti dvou analýz výzvy Orange D4D-Senegal Challenge 2014 (2015), devatenáct týmů použilo Python a sedmnáct R (s velkým rozdílem následuje Matlab se šesti použitími).

Výhodou Pythonu oproti R je významně méně balíčků (existuje pouze několik hlavních), větší komplexnost ve smyslu celého procesu (tedy zisku dat, analýzy, možnost vytvoření aplikace), umožňuje lépe vytvořit deep (hluboké) neuronové sítě. Jeho filozofií je založená na „psaní čistého kódu, který může být jednoduše čtený každým programátorem“, zatímco R je primárně statistický nástroj preferován statistiky. Naopak nevýhodou Pythonu oproti R je slabší grafické zpracování a R také pokrývá větší rozsah statistických technik, „od psychometrie ke genetice po finance“ (Paruchuri, získáno 10.6.2018).

Pro psaní kódu v Pythonu lze s výhodou využít tzv. Jupyter Notebook, jakousi obdobu notepadu, která umožňuje logické členění, detekci chyb a spuštění programu.

Nyní tedy máme přehled o zisku i metodách analýzy Big dat (Dat), jaké jsou ale nástrahy výzkumu, čeho je třeba se v jeho rámci vyvarovat, případně na co brát při čtení závěrů založených na Big datech zřetel?

7 Critical data studies

S lehkým časovým odstupem po masivním nástupu fenoménu Big dat se začal rozvíjet obor, který tento fenomén reflektuje s kritickým nadhledem ve snaze pomoci kultivovat tento nový, rychle se rozvíjející a dosud patřičně vědecky neukotvený směr výzkumu. Pro tyto kritické reflexe, které jsou samy jistou nekoncepčností a ad hoc přístupem poznamenány, se vžil pojem „critical data studies“ (dále CDS). Jak píší Illiadis a Russo (2016), „CDS se dosud ukazuje jako volný propletenec rámců, návrhů, otázek a manifestů.“ (str. 3). Přesto lze v CDS studiích vystopovat jistá dominantní témata, na které by se měl brát v rámci Big dat zřetel.

Některé výhrady, byť jsou primárně adresovány Big datům, lze samozřejmě vztáhnout i na klasické sociologické výzkumy a netýkají se tak pouze Big dat, jen mohou být v Big datech zastřenější (viz například poslední bod, apriorní důvěra). Pokud jsou v klasických sociologických přístupech zisku dat ještě vyhraněnější, explicitně to bude zdůrazněno.

7.1 Problém relevance dat

Jak upozorňují McFarland et al. (2016) Big data jsou vždy připravená („found“) data, která jsou zatížena chybami (str.6). Například „každý dataset je spjat s platformou svého vzniku a sociální aktivitou (Facebook přátelství, Twitter vyjádření)“. Upozorňují především na problematiku sociálních sítí, kdy „dataset je zatížen selekcí uživatelů“, tedy jde o lidi využívající moderní technologie a případně i osobnostními charakteristikami typu extraverze. Na druhou stranu argument zkreslení vlivem využívání moderních technologií má samozřejmě globálně smysl, ale omezí-li se výzkum na vhodnou podskupinu, případně zohledníme-li její specifika, svou argumentační váhu už lehce ztrácí.

Netýká se to ale jen sociálních sítí, podobnou výhradu lze mít také k datům získaným z mobilních aplikací, kdy je selekce dokonce dvojí – musí vlastnit smart phone a aktivně si aplikaci nainstalovat/využívat ji. A toto je samozřejmě třeba v studiích vždy pečlivě zohlednit.

Podobně při důkladném prozkoumání studií Orange D4D- Senegal Challenge 2014 (2015), představenou v kapitole o datech mobilních operátorů, překvapí, jak málo jich byť i zmínilo otázku validity (potažmo reprezentatitvity) svých leckdy silných závěrů. Penetrace mobilními telefony v roce 2013 byla sice 92.93%, ale podíl společnosti Orange na trhu pouze mezi 56 – 62%. Do studie se také dostala pouze data uživatelů, kteří byli aktivní více než 75% daného časového období. Také se nabízí otázka, zda se chování uživatelů Orange neliší od ostatních, respektive zda operátorovi Orange nedávají přednost různé etnické/sociální/náboženské skupiny (k tomu jsem relevantní podklady nenalezla).

Jakkoli musí být problematice relevance dat v Big datech věnována pozornost, nejedná se o výhradní problém Big dat a v klasických sociologických přístupech je zastoupen také. Jak se liší lidé, kteří jsou ochotni zodpovídat na ulici dotazy včetně přibližné výšky platu nebo lidé, kteří přeruší rozdělanou práci, aby zodpověděli několik dotazů po telefonu? Groves (jak je citováno v Krejčí, 2008) tuto problematiku ve výběrových šetřeních systematizuje jako „chyby chybějících pozorování“, vzniklé v důsledku chyb v „pokrytí populace“, „výpadku návratnosti“ a „výběru“. Přesto si sociologie našla způsoby, jak s těmito chybami pracovat a výběrové šetření patří mezi pilíře zisku dat.

Do této sekce také spadá problematika falešných účtů na sociálních sítích (jak byla diskutována v části věnované zisku dat), ale i obecně na internetu.

7.2 Problém s kontextem

Z jiného důvodu relativizuje tento typ výzkumů L.Manovich (2012): „Neznamená to, že nelze využít […seznam sítí…], jen musíme myslet na to, že tato data nejsou průhledným oknem do lidských představ, zájmů, motivů a ideí. Vhodnější je tak podle Manoviche představit si je „jako rozhraní, jimiž se lidé představují světu.“ (str.466) Upozorňuje především na to, že aktivita na sítích představuje pouze výseč skutečného života a může být ovlivněna budováním si vlastního obrazu, kdy lidé „pečlivě formují svou prezentaci pro druhé“ (str.465)

Exaktně tuto tušenou formu zkreslení potvrdili Squicciariniová a Griffin (2014), kdy v on-line výzkumu mezi vysokoškolskými studenty[59] zkoumali, co vede uživatele sociálních sítí ke zkreslení/zadržení poskytnutí informací týkajících se věku, výdělku, práce, atd. Z výsledku vyplývá, že to nejsou obavy ze ztráty soukromí, ale snaha vykreslit „úspěšný sociální charakter". Uživatelé ale nejsou jen vedení snahou vybudovat si patřičnou sociální image, důležití jsou pro ně okolí a přátelé, kteří mají velký vliv na strategii zkreslení, kterou uživatelé zvolí.

Dalton, Taylor a Thatcher (2016) v této souvislosti dokonce nastolují principiální otázku „epistemiologického a ontologického skoku (leap) mezi jedinci produkujícími Big data a reprezentací těchto jedinců daty" (str.4)

Podobně zdůrazňují i Tinati et al. (2014), že data nevznikají samostatně, ale „jsou vždy sociotechnicky konstruovaná", vznikají na lidských výtvorech, jejich přijetím a přizpůsobením se jim (str. 12).

Obecně lze samozřejmě souhlasit, ale Big data vznikají „přirozeně", tedy pouze v důsledku lidského chování a jednání, tedy nevědomého i uvědomělého vnějšího projevu (použije-li se toto zjednodušení pojmů vymezovaných různě jednotlivými sociology i psychology). Pokud by měla být rozporována možnost jejich pochopení a výkladu průzkumníkem, pak se lze dostat až na fenomenologickou úroveň radikální intersubjektivity ve smyslu Husserla, zpochybňující možnost poznání obecně. (Ďurďovič, 2018)

Big data nevznikají primárně pro výzkum a právě proto obsahují oproti třeba dotazníkovému šetření (ustálené sociologické metodě zisku dat) o jednu interpretační rovinu méně a čelí tak méně riziku sociální intersubjektivity i vlivu zkreslení podmínkami. Dotazovaný v dotazníkovém šetření musí pochopit, na co je tázán, co přesně znamená odpověď a vědomě se rozhodnout (nebo nevědomě určit), jak upřímně a zodpovědně odpoví. Leckdy vhodnou odpověď do škály naleznou společnými silami s méně zkušeným tazatelem. Navíc může být ovlivněn situačními podmínkami šetření, jako je osobnost tazatele („způsob čtení otázek, intonace hlasu, způsob vystupování, vzhled" (Krejčí, 2008, str. 28)), místo šetření apod. Ovlivněn

[59] Nabízí se samozřejmě otázka ve smyslu předchozí podkapitoly o specifičnosti této skupiny vůči ostatním uživatelům sociálních sítí.

může být i samotnou konstrukcí dotazníku, zahrnující posloupnost otázek, znění apod. Dále muže zahrnout psychické faktory jako je tendence ke konformitě, vliv psychického naladění apod. Petrusek (1993) některé z výše uvedených nazývá „zamlčenými předpoklady“, neboť dotazníkové šetření je „ryze pozitivistické, přesněji behavioristické provenience“. (str.117) A přesto si klasická sociologie našla cesty, jak tyto vlivy v rámci možností eliminovat (například práce Sarise a Gallhofera z roku 2007) a dotazníkové šetření je nejpoužívanější metoda zisku dat.

Big data naproti tomu vznikají „přirozeně“, ve smyslu popsaném výše. Samozřejmě při vzniku čelí situačním faktorům, jsou ovlivněny psychickými vlastnostmi jedince a předpokládají jeho vlastní interpretaci světa, ale ty jsou pouze projevem přirozeného vnějšího světa v koexistenci s vnitřní konstituce jedince. Světa neovlivněného, nepředloženého a nezkonstruovaného externím výzkumníkem.

Problém kontextu ale nabývá i jiných podob. Například v jedné ze studií Orange D4D-Senegal Challenge 2014 (2015) věnované urbanismu z mobilních dat vyvozují, že Senegalská města vykazují rozvinutý pracovní trh, protože lidé nemusejí příliš cestovat za prací (str. 243). Měřeno porovnáním zachycení mobilní aktivity v denních a nočních hodinách, která se odehrává převážně na stejných místech. Na první pohled logický závěr a tudíž krásný výsledek, dokud si ale neuvědomíme (jak je poznamenáno v jiné nezávislé studii jiných výzkumníků), že Senegal v té době trápila 48 % nezaměstnanost, 62.5% obyvatel Senegalu bylo v době studie mladších 24-ti let a panovala mezi nimi v městech 40-ti % nezaměstnanost (str. 292). Skutečně je tak malý rozdíl v místě denní a noční aktivity důsledkem rozvinutého pracovního trhu?

Dalton et al. (2016) obecně vztahují problém kontextu k propasti Big dat (Big data divide, pojem prvně užitý Andrejevicem, jak je citováno v článku). Upozorňují tím, že data jsou vždy produkována v jisté lokaci a čase s konkrétními zvyklostmi ve využívání digitálních technologií a přístupem k nim. Výsledky průzkumu by tak s jistou nadsázkou měly místo „my“ používat „ti, které jsem schopen rozpoznat z mojí strany propasti Big dat“. (str.3) Problém je ilustrován na vyhodnocení Big dat produkovaných využíváním mobilního telefonu v Londýně (bohatá metropole s technologicky vyspělou infrastrukturou) a Mauretánii (opak). Nelze než souhlasit, ale opět se jen obtížně zbavit dojmu, že tento problém je univerzální pro všechny

sociálně vědní výzkumy, viz například dobře známá problematika mezinárodních výzkumných projektů.

7.3 Problém paradigmatu

Široce diskutovanou otázkou v CDS je problematika změny přístupu k výzkumu, kterou Big data mohou přinést, případně kterému pokušení musí v této souvislosti odolat. Pro formování debaty o této změně se využívá pojem změny paradigma, jak jsme jej nastínili v kapitole věnované metodologické tradici sociologie.

Kitchen (2014) v souvislosti s Big daty upozorňuje, že se „vytváří zásadně odlišná epistemologie; že probíhá přechod na nové paradigma“ a rozlišuje dvě podoby, kterých toto paradigma nabývá - empirické nebo poháněno daty (data – driven[60]).

Empirikové, „vyznávající heslo smrt teorii“ (heslo vycházející z ikonického článku publikovaného v roce 2008 v internetovém magazínu Wired[61]) chápou Big data jako nástroj k odhalení vzorců bez teoretických analýz a výběrových šetření, „data a korelační vzorce jsou dostatečné pro vědecký pokrok, modus operandi je čistě induktivní povahy“(str.4).

Empirický přístup je ale dle Kitchena chybný, protože je založen na mylných předpokladech - 1. soubor Big dat nikdy není úplný, 2. ani induktivní nálezy nevznikají ve vědeckém vákuu a jsou formovány předchozími nálezy a teoriemi, 3. data vždy vyžadují interpretaci, zjištěné korelace mohou být náhodné bez kauzality, 4. data musí být vykládána v kontextu a se specifickými znalostmi oboru (ne pouze datovými znalostmi), jinak jde interpretaci anemickou a nevhodnou.

Jako varující uvádí práci fyziků při modelování sociálních a prostorových procesů, kterou označuje za empirickou, „úmyslně ignorující několik staletí společenskovědních znalostí. […] Výsledkem je analýza měst, která je redukcionistická, funkcionalizovaná a ignoruje účinky kultury, politiky, řízení.“ (str.5)

[60] Data driven přístup byl představen v kapitole.

[61] Anderson, Ch. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Získáno 12.5.2018 z https://www.wired.com/2008/06/pb-theory/

A velké množství dalších ilustrativních příkladů poskytuje i Orange D4D- Senegal Challenge 2014 (2015), především ve studiích týkajících se chudoby, denních zvyklostí ve využívání mobilních telefonů apod. Ve studiích bohužel není zastoupena jediná katedra sociologie (jsou-li katedry zmíněny) nebo alespoň sociologická instituce. Zato jsou hodně zmíněny instituce typu Computer science and engineering, Department of Electrical and Computer Engineering, Department of Computer Science, Institut of theroretical physic atd. Přitom data telekomunikačních operátorů „vysoce převyšují to, co mohou posbírat akademičtí sociologové [*klasickými cestami*]" (Savage a Borrows, 2008, str.5).

Podobné výtky k těmto typům analýz, které označují výstižně za inženýrské, adresují i McFarland et al. (2016).

7.4 Problém přílišné závislosti na výpočetní technice

Z jiného úhlu pohledu přistupují k Big datům Symons a Alvarado (2016). Kritizují samotné CDS z přílišné omezenosti a kladení důrazu na nesprávné elementy, které jsou podobné i klasickým výzkumům, kdy přiznání, že výzkum je ovlivněný lidským činitelem je sice „rozumné, ale relativně triviální filozofické tvrzení". (str 5)

Doporučují zaměřit pozornost na jiný aspekt Big dat a to softwarovou podstatu výzkumu („software intensive"). Nazývá se tak stav, kdy software rozhodujícím způsobem ovlivňuje samotnou podstatu jako celku. Vhled do Big dat by nebyl možný bez užití výpočetních metod za použití softwaru, který do výzkumu přináší tzv. „epistemickou nejistotu". Autoři v této souvislosti nezmiňují pouze fundamentální nejistotu plynoucí z užití některých metod, jak jsem byla popsána v dřívější části práce (tuto nazývají problémem „black box"), ale míní tím i například využití už dále nepodporovaného kódu („legacy code") nebo vznik náhodných vzorců během simulace („weak emergence"), které mohou být nesprávně vyhodnoceny a nadhodnoceny.

V prvním případě („legacy code") přiznávají, že podobný problém může nastat i v klasických výzkumech, ale díky menší komplexnosti to nemusí být tak palčivé. Druhý případ („weak emergence") popisuje vznik „nezamýšlených/ nenaprogramovaných/ nečekaných vzorců chování", které se vyskytují pouze v běžící simulaci a jsou pouze produktem její dynamiky.

S epistemickou nejistotou úzce souvisí i složité testování průchodu programu („path complexity“), kdy program obsahuje tolik různých podmínek (a tedy variant průchodu), že není možné použít statistické metody k detekci chyb.

Autoři nepřinášejí řešení, oba tyto aspekty ale adresují k pozornosti CDS. Je to skutečně důležité téma, které pravděpodobně pro svou technologickou obtížnost v CDS skutečně nebývá často zmiňováno.

V oboru informačních technologií je ale otázka testovatelnosti a správných postupů při využívání softwaru už dlouho rozvíjeným oborem s poměrně dobře zpracovanými jednoduchými „best practise“ doporučeními i obsáhlými studiemi pro komplexnější systémy. Například pokud jsou ve výzkumu používány vlastní zdrojové kódy, pak by měly být součástí výstupů (nebo alespoň dostupné na vyžádání), stejně tak alespoň stručně popsáno testování apod.

Některé výše popsané problémy by tak šly jejich aplikací přinejmenším zmírnit.

Fundamentální nejistota je komplikovanějším problémem. Ale díky vzrůstajícímu využívání metod umělé inteligence (AI) v citlivých oborech (zdravotnictví, finančnictví apod.) a tedy rostoucí závislostí na jejich výsledcích, probíhá výzkum tzv. explainable AI (EAI). Tedy vysvětlitelné umělé inteligence, která by s udržením všech výhod AI „umožňovala lidským uživatelům [*ji*] rozumět, adekvátně věřit a efektivně řídit“[62] (Grunning, 2017, str. 7). Tato nová EAI by tak eliminovala problém fundamentální nejistoty.

7.5 Problém apriorní důvěry

A v neposlední řadě je to otázka důvěryhodnosti. Big data budí jakési zdání nezpochybnitelné vědeckosti, tvrdých dat zpracovaných sofistikovanými metodami, kde není prostor pro zkreslení a kde výsledné závěry jsou dokonalou interpretací reality. Boy a Crawford (2012) s jistou mírou nadsázky považují „mytologii“ za jeden ze tří určujících rysů Big dat. Míní tím přesvědčení, že „velké datové soubory nabízejí

[62] Jako ilustrativní příklad uvádějí současnou AI, která při rozpoznávání neznámého obrázku odpoví „Je to s pravděpodobností 0.93 kočka“. Vyvíjená EAI by měla poskytnout odpověď ve stylu „Je to kočka, protože má srst, fousy a drápy. Má tyto znaky: [určující znaky neznámého obrázku porovnané s obdobnými znaky obrázku vzoru kočky]“.

vyšší formu inteligence a poznání, které mohou generovat poznatky, které byly dříve nemožné, s aurou pravdy, objektivity a přesnosti.“ (str. 3)

Ale samozřejmě to není pravdou, mimo všechny výhrady zmíněné výše, do hry vstupuje i v případě Big dat známý průvodce sociologických výzkumů – operacionalizace a konceptualizace.

Například v mnoha výzkumech studií Orange D4D- Senegal Challenge 2014 (2015) bylo rozhodujícím vstupem určení domovské destinace vlastníka mobilního telefonu. A kolik studií, tolik přístupů. Domov je tam, kde buď:

a) strávili nejvíce kalendářních dní v roce (šlo určit u 90% případů, u zbylých bylo vybráno náhodně);

b) tam, kde uskutečnili svůj poslední denní hovor, den byl definován jako 5-16:59,

c) kde nejčastěji trávili noc, tedy období mezi 22:00 – 06:00;

d) odkud největší počet dnů volali;

e) uživatelé volají minimálně každý druhý den mezi 20-24 hodinou a minimálně v 95-ti% případů.

Není obtížné představit si objekt, který bydlí v městě D (domov), pracuje v městě P (práce), přičemž se domů vrací přes město C (cesta) a občas pracovně dojíždí do města E (externí), kde přespí. Z domu a s domovem příliš netelefonuje, telefon využívá primárně v práci, jen cestou domů pravidelně okolo města Z telefonicky ověří, zda tam nemá provést nákup, a v případě přespání mimo domov v detašované práci pak zavolá v noci rodině. Jako jeho domov bude vyhodnocen ve studii používající definici b) město s obchodem C, ve studii používající definici c) pak místo jeho externí práce E a ve studii používající definici d) pak místo, kde pracuje P. Studie využívající definici e) jej nezachytí, definice a) pravděpodobně vybere opět místo práce P.

Bylo by bezpochyby zajímavé přepočítat všechny výsledky jednotnou metodikou.

A bez nadhledu a jistého smíření se s nejednoznačností se samozřejmě neobejde ani samotný sběr dat. Například údaje poskytovatelů obsahu (čtenost/počet zhlédnutí

atd.) sice mohou být na první pohled považována za tvrdší a průkaznější data než data odvozená z uživatelských průzkumů, ale za jednoznačně tvrdá data je považovat nelze (použijeme-li výstižné pojmenování „tvrdá data“, byť je toto pojmenování „výrazem odborného slangu“ viz Buriánek (2017)).. Jeden počítač může být sdílen několika členy domácnosti a naopak jeden uživatel může využívat více zařízení (domácí PC, pracovní PC, tablet, smartphone atd.), jedno video může sledovat společně několik uživatelů (např. kolegů v práci) apod. (jak je řešeno například v Lupač, 2015).

Otázka důvěry nabývá i velkého významu v souvislosti s tím, jak je s výsledky získanými pomocí Big dat nakládáno. Předmětem CDS se tak také stává způsob, jakým jsou Big data využívána k podpoře politických rozhodnutí a řízení, v oblasti sociální politiky apd. a zda k těmto rozhodnutím mohou poskytnout dostatečný mandát (Illiadis & Russo, 2016)

Například výsledky studií Orange D4D- Senegal Challenge 2014 (2015) jsou koncipovány jako podklady ke zlepšení života v Senegalu, přinášejí návody na zlepšení dopravní infrastruktury, doporučení pro opatření v případě epidemií včetně predikcí jejich průběhu, tipy pro efektivní výstavbu rozvodné sítě či nových zdravotnických zařízení, apod. Vzhledem k výtkám zmíněným v celé kapitole (sporná reprezenativita, nedostatečné zahrnutí všech faktorů, chybějící teoretický rámec apod.) je skutečně otázkou, zda jsou podkladem dostatečně validním.

7.6 Problémy nad rámec dosavadních CDS

7.6.1 Problém věrohodnosti a zabezpečení dat

Dalším důležitým bodem, který dosud není v CDS rozpracován, ale ve výzkumech založených na Big datech by mu měla být věnována pozornost, je i problém věrohodnosti a zabezpečení dat. Jen obtížně si lze představit, že by někdo dokázal cíleně ovlivnit výsledky obsáhlého dotazníkového šetření, pro jednotlivé drobné manipulace (typu falšování dotazníků konkrétním tazatelem) má pak sociologie metody, jak je odhalit (např. faktorovou analýzou).

Big data však přinášejí v tomto směru novou výzvu. S digitálními daty může být lehce a komplexně manipulováno. Vlastníci dat nemusí dodat úplný soubor dat, mohou jej

jednoduše modifikovat (ať už chybou nebo záměrně), stejně tak může dojít k cíleným záškodným úpravám bez jejich vědomí, data mohou být změněna cíleným útokem.

Problematiku lze rozčlenit pomocí tří základních pojmů[63] – důvěrnost, integrita a autentizace. Důvěrnost se vztahuje k „zabezpečení dat před neautorizovaným přístupem“, integrita značí „zabezpečení dat před jakoukoliv změnou v procesu jejich sběru, přenosu a uchování“ (a lze doplnit i zpracovávání). Autentizace pak odkazuje k „zabezpečení, že data pochází ze známého zdroje, že jsou autentická“. (Pomaizlová, 2016, str. 14)

Důležitost dobře zvládnutého datového managementu do sociologie postupně proniká, jak ukazuje publikace Českého sociálněvědního datového archivu SOÚ AK (Krejčí, 2014). V rámci výzkumů založených na Big datech by ale speciálně tématům kyberbezpečnosti a „best IT practicies“, měla být věnovat obzvlášť velká pozornost. Prvním myslíme soubor bezpečnostních opatření zamezujících cílené nevyžádané manipulaci s daty (například užití firewallů, pouze dedikovaného hardwaru a ověřeného softwaru apod.). Druhým pak soubor doporučení uplatňovaných v rámci IT projektů. Například pokud jsou data získána od vlastníků, pak součástí dodávky dat by měly být i zdrojové kódy použité k jejich získání včetně záznamu o otestování apod.

7.6.2 Právní a etické aspekty

Legální otázky zpracování dat při sociálněvědních výzkumech jsou například řešeny na webové stránce Českého sociálněvědního archivu (“Etické a legální otázky”). Jde ale o verzi neaktualizovanou (odkazující na publikaci z roku 2012 (Krejčí & Leontiyeva, 2012)) a koncipovanou primárně pro klasické metody zisku dat.

Problematika Big dat (stejně tak jako nové legislativní úpravy) dosud není zpracována. Není to ani ambicí této práce, která tuto problematiku jen nastiňuje.

S daty by mělo být obecně nakládáno ve shodě s *Obecným nařízením o ochraně osobních údajů 2016* (EU), které k 25.5.2018 nahradilo *Zákon č. 101/2000 Sb na*

[63] Systematizace použita v oblasti práva v kontextu „Bezpečnostní výzev IoT“ (Internet věcí)

ochranu osobních údajů a o změně některých zákonů[64] (ČR). Dohledem nad ochranou osobních údajů byl a nadále je pověřen Úřad na ochranu osobních údajů.

Jak například k problematice dat poskytovaných vlastníky v roce 2016 uváděl v článku věnovaném Big datům[65] jeho tehdejší předseda I.Němec - „čistě teoreticky k záměrům prodávat[66] soubory dat mohu říci, [...] že zákon zabývající se osobními údaji manipulaci s daty dost striktně reguluje. Bezpochyby budeme tuto oblast hodně pečlivě sledovat." („Big data", 2015) Obecně pak ÚOOÚ uvádí Big data jako problematiku, „které se bude muset mnohem víc věnovat" (Iniciativa Průmysl 4.0, 2016).

Zákon o ochraně osobních údajů důsledně rozlišoval mezi osobními („jakákoliv informace týkající se určeného nebo určitelného subjektu údajů"), citlivými (národnostní, členství v odborech, ... ale i biometrické údaje) a anonymizovanými údaji a pro každý typ přesně specifikoval, jak lze s nimi nakládat. Dle stanoviska ÚOOÚ č. 2/2006 (ÚOOÚ, 2013, str. 4) bylo „i při zpracování osobních údajů pro vědecké účely nutné dodržovat veškeré povinnosti stanovené zákonem o ochraně osobních údajů" (s výjimkou doby archivace a získání souhlasu).

Obecné nařízení o ochraně osobních údajů 2016 (EU), známé pod zkratkou GDPR, osobní data definuje velmi podobně (čl. 4, odst 1)), jen je detailněji vypočítává (například lokační údaje, síťový identifikátor apod.)[67]. Oproti původnímu zákonu také explicitně uvádí, že přestože data nesmí být zpracovávána v rozporu se svým účelem, „další zpracování [..] pro účely vědeckého či historického výzkumu nebo pro statistické účely se podle čl. 89 odst. 1 nepovažuje za neslučitelné s původními účely" (čl. 5, odst. 1, b)). Nelze na tomto místě detailně analyzovat dopady GDPR a provádět srovnání s původními předpisy stran samotného zpracování dat, jen ještě speciálně

---

[64] Ve znění účinném od 1. července 2017. Zákon bude novelizován a bude specifikovat především postavení ÚOOÚ, a také dílčí otázky, které byly ponechány vnitrostátním úpravám.

[65] Článek byl původně publikován v Parlamentních listech, je ale uváděn na webových stránkách ÚOOÚ.

[66] Vyjádření se vztahuje k tehdy avizované aktivitě telekomunikačního operátora využít svá data ke komerčním aktivitám. Není ale důvod se domnívat, že v případě bezplatného postoupení dat by to bylo jiné.

[67] ÚOOÚ však argumentuje, že výčet nikdy nemůže být plný a že síťový identifikátor byl osobním údajem i před platností GDPR, jen nebyl v zákoně explicitně jmenován. (viz „Desatero omylů", bod 3. Získáno z https://www.uoou.cz/desatero-omylu/ds-4818/archiv=0)

upozorníme na čl. 89[68] odst 1., který pro vědecký výzkum ukládá dodržování zásad minimalizace údajů.

Ovšem nejen GDPR, v souvislosti s daty získanými z internetu je na portálu ePravo diskutována i problematika možné kolize s autorským zákonem, který „poskytuje zvláštní ochranu tzv. pořizovateli databáze". Databází je myšlen „soubor nezávislých děl, údajů nebo jiných prvků, systematicky nebo metodicky uspořádaných a individuálně přístupných elektronickými nebo jinými prostředky". Bez uděleného souhlasu pořizovatele pak nesmí být data vytěžována, dokonce ani jejich malá část.[69] Podobně je v článku uváděno, že zisk dat z internetu může omezovat i „zákon o některých službách informační společnosti, či jiné předpisy obsahující ochranu tohoto typu dat [*myšlena ochrana osobních údajů*]", případně být v kolizi s pravidly hospodářské soutěže. (Nielsen & Strakatý, 2016)

Explicitně lze také říct, že pokud má nějaká služba definovány podmínky užití (terms of service), pak by data z ní měla být získávána a zpracovávána v souladu s těmito podmínkami. Týká se to i sociálních médií, které je mají obsáhle definovány a ve výzkumech by tak na ně měl být brán zřetel. Byť je třeba možnost jejich data specifickým způsobem využít a v minulosti tak učiněno bylo, současné podmínky užití mohou být striktnější a je třeba to vždy před výzkumem ověřit.

Nejasná je i otázka využití dat z kamerových systémů, kde se obecně uvádělo dokonce několik možných kolizí s lidskými právy (právo na soukromí, právo na volný pohyb a právo na ochranu osobních údajů. ((Braunová, 2009)) V případě dat generovaných internetem věcí pak bude třeba dokonce „smluvně upravit mezi účastníky jednotlivých systémů, kdo tato data vlastní". (Pomaizlová 2016)

---

[68] Čl 89 odst. 1 pak stran vědeckého výzkumu uvádí že: „Zpracování pro účely archivace ve veřejném zájmu, pro účely vědeckého či historického výzkumu nebo pro statistické účely podléhá v souladu s tímto nařízením vhodným zárukám práv a svobod subjektu údajů. Tyto záruky zajistí, aby byla zavedena technická a organizační opatření, zejména s cílem zajistit dodržování zásady minimalizace údajů. Tato opatření mohou zahrnovat pseudonymizaci za podmínky, že lze tímto způsobem splnit sledované účely. Pokud mohou být sledované účely splněny dalším zpracováním, které neumožňuje nebo které přestane umožňovat identifikaci subjektů údajů, musí být tyto účely splněny tímto způsobem." *Obecné nařízení o ochraně osobních údajů 2016*

[69] Krejčí a Leontiyeva (2012) řeší problematiku autorského práva primárně vzhledem k databázím vytvořeným v rámci výzkumných institucí a uvádějí, že „ochraně podléhá dílo, nikoli fakta v něm uvedená" (str. 35).

S právními otázkami se úzce vážou otázky etické. A to nejen ve vztahu k práci s daty, ale i s celkovým designem výzkumu. Sociologie měla dosud spíše omezené možnosti využití experimentu, to se ale především rozvojem internetu změnilo. Jako příklad uveďme práci experimentálně potvrzující platnost Matoušova efektu, kdy výzkumníci po jistou počáteční fázi na čtyřech typově různých platformách aktivně ovlivňovali úspěšnost náhodně vybraných entit, aby je nakonec porovnali s úspěchem entit, kde tato intervence neproběhla[70]. Aktivně a nepřiznaně tak ovlivnili realitu (o to více, že platnost efektu skutečně prokázali), aniž by o tom ostatní uživatelé platforem věděli. V rámci Big dat je podobných experimentů představitelná celá řada.

Dále je například etickou otázkou, zda je nutné ve výzkumné zprávě citovat celé pasáže z obsahu generovaného uživateli, jak bývá rozšířenou praxí ve výzkumech vycházejících z těchto typů dat (například příspěvky uživatelů na sociálních sítích, z diskuzí, blogů apod). Pokud je citovaná pasáž delší (nějak specifická, časově či jinak označená), pak její autor může být jednoznačně rozpoznán. Nemusí se cítit komfortně s pozicí výzkumného subjektu, spojením s tématem výzkumu, případně i s výzkumnými závěry. Pokud je například citace uváděna jako typický příklad zástupce nějaké skupiny (Pilnáček, 2016)[71], nebylo by pro ilustraci vhodnější použít pro skupinu typická slova, slovní vazby apod? Někdy jsou tyto citace uváděny dokonce s uživatelským jménem (!), jako například v Tinati et al (2014).

Vzhledem k obsáhlosti, komplexnosti a citlivosti právní i etické problematiky je tato část jen jejím nastíněním a vyžaduje systematické rozpracování uchopené právní autoritou. V případě etických zásad pak i širokou oborovou diskuzi.

V rámci výzkumu využívajícího Big data lze tak jen v tuto chvíli doporučit legálnost konkrétního způsobu práce s daty pečlivě ověřit. Bude-li to pak spolu s případným osvětlením aplikování etických principů zahrnuto ve výstupu z výzkumu, přispěje to k lepšímu povědomí o těchto aspektech a také jejich kultivaci.

---

[70] Entitami byly projekty v rámci crow-fundingových kampaní, uživatelské recenze, on-line petice a diskuzní příspěvky. Autoři sami si byli vědomi etických výhrad a věnovali tomuto problému speciální přílohu.

[71] Autor uvedení celého, v tomto případě diskuzního příspěvku, zvažoval, přiklání se ale k názoru, že přispěvatelé vědí, že s veřejně publikovanými příspěvky může být i jinak nakládáno. (M.Pilnáček, osobní rozhovor)

8 Závěr: sociologie dvacátého prvního století

Stále rostoucí produkce dat, zdokonalující se hardwarové a softwarové zázemí k jejich zpracování a rozvíjející se metody jejich analýzy jsou realitou jednadvacátého století.

Oblast Big dat je kvůli své technologické náročnosti zatím převážně domestikována technickými obory, které tím vstupují na pole témat, která byla dosud řešena sociálními vědami, včetně sociologie.

Tento technický přístup však postrádá její erudici, především stran kritického pohledu, rozvinuté metodologie a teoretického ukotvení. Big data analýzy jsou tak podrobovány mnoha oprávněným výtkám, které se souhrnně označují jako Critical data studies. Zpravidla však nenabízí systematizované řešení.

Ve světle těchto kritik zaniká jeden z největších přínosů Big dat, tedy unikátní propojenost kvantitativních a kvalitativních hodnot. Big data tak nejen znamenají nová data a sofistikované analytické metody, ale i nový přístup, který může být označován jako Big data paradigma („data driven“).

Pro jeho správné uchopení ale stále chybí metodologické rozpracování. Podobně problémy, které jsou aktuálně v Big datech rozeznávány, jsou velmi podobné těm, s kterými sociologie dokáže pracovat v rámci svých ustálených postupů zisku informací. Bylo by tedy vhodné napnout úsilí sociologie tímto směrem. Disponuje totiž nejen patřičnou erudicí, ale jako obor už zvládla překonat nástrahy pozitivismu. Zapojení sociologie by tak mohlo přispět k definování správné podoby Big dat paradigmatu.

Větší participaci sociologie však brání technologická náročnost, která je s Big daty neodmyslitelně svázána. Zvýšení kompetencí v oblasti informačních technologií by sociologům umožnilo nejen nahlédnout a využít nabízející se možnosti, ale i osvojit si zásady v oboru informačních technologií používané (tedy jakousi metodologii informačních technologií). V rámci výuky sociologie je tak žádoucí nasměrovat úsilí i tímto směrem tak, jak to v tuto chvíli dokáží nové obory.

Pro sociologii novým prvkem je také vlastnictví dat, kdy některá nemusí být volně dostupná. Odpovědí by měla být schopnost sociologů akademického výzkumu

přinášet přidanou hodnotu a navazovat partnerství s vlastníky dat tak, jak to v tuto chvíli činní technické fakulty. Nezodpovězené v rámci Big dat zůstávají i právní a etické otázky výzkumu, které budou teprve vyjasňovány.

Sociologie by měla výzvu Big dat uchopit, neboť jen to jí umožní zůstat plnohodnotnou vědou jednadvacátého století, využívající všechny dostupné prostředky k rozvoji poznání.

9 Seznam zdrojů

Big data ropou blízké budoucnosti. (11.5.2015). Získáno z https://www.uoou.cz/assets/File.ashx?id_org=200144&id_dokumenty=15244

Botsman, R. (21.10.2017). Big data meets Big Brother as China moves to rate its citizens. *Wired* . Získáno z http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

Boyd, D. & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Journal Information, Communication & Society, v. 15. 662 – 679,* DOI: 10.1080/1369118X.2012.678878

Boyd, D. M., & Ellison, N.B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication 13,* DOI:10.1111/j.1083-6101.2007.00393.x

Bracht, O. (2013, 27.11.). *Five ways to handle Big Data in R*. Získáno z https://www.r-bloggers.com/five-ways-to-handle-big-data-in-r/

Braunová, V. (2009). *Kamerové sledování veřejných prostranství a institucí*. Získáno z http://www.mvcr.cz/clanek/kamerove-sledovani-verejnych-prostranstvi-a-instituci.aspx

Buriánek, J. (11.12.2017). Analýza sekundární. v *Sociologická encyklopedie*. Získáno z https://encyklopedie.soc.cas.cz/w/Anal%C3%BDza_sekund%C3%A1rn%C3%AD

Buriánek, J. (11.12.2017). Data měkká a tvrdá. v *Sociologická encyklopedie*. Získáno z https://encyklopedie.soc.cas.cz/w/Data_m%C4%9Bkk%C3%A1_a_tvrd%C3%A1

Burrel, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society* [online]. Dostupné z http://journals.sagepub.com/doi/10.1177/2053951715622512

Burrows, R. & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society* [online]. Dostupné z http://journals.sagepub.com/doi/10.1177/2053951714540280

Cambridge Big Data. (n.d.). *Cambridge Big Data* Získáno z https://www.bigdata.cam.ac.uk/

Candia, J., Gonzalez, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *The Journal of Physics A: Mathematical and Theoretical.* Dostupné z 10.1088/1751-8113/41/22/224015

Cao, H. & Lin, M. (2017). Mining smartphone data for app usage prediction and recommendations: A survey. *Pervasive and Mobile Computing*. 37. 1-22. DOI: 10.1016/j.pmcj.2017.01.007.

Castells, M. (2015). *Networks of Outrage and Hope: Social Movements in the Internet Age, 2nd Edition*. Cambridge (UK): Polity

Conte,R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., ..., Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics,* v 214. Dostupné z https://doi.org/10.1140/epjst/e2012-01697-8

Cox, M., & Ellsworth, D. (1997). Application-Controlled Demand Paging for Out-of-Core Visualization. Dostupné z https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

Český statistický úřad (2017). *Informační společnost v číslech 2017.* Dostupné z https://www.czso.cz/documents/10180/46014808/061004-17_S.pdf/b9a0a83e-7a6f-4613-b1df-33fe8b5d1a8e?version=1.1

Dalton, C. M., Taylor, L. & Thatcher, J. (2016). Critical Data Studies: A dialog on data and space. *Big Data & Society* [online]. Dostupné z https://doi.org/ 10.1177/2053951716648346

Davenport, Thomas H., and D. J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90, no. 10 (October 2012): 70–76.

Data for development – Challenge Senegal, Book of abstracts: Scientific Papers (2015). Dostupné z http://www.d4d.orange.com/fr/content/download/43452/406501/version/1/file/D4DChallengeSenegal_Book_of_Abstracts_Scientific_Papers.pdf

di Bella, E., Leporatti, L., & Maggino, F. (2016). Big Data and Social Indicators: Actual Trends and New Perspectives. Social Indicators Research. Dostupné z https://doi.org/10.1007/s11205-016-1495-y

Dobrovolný, V. & Kudrnáčová, M. (2017) Expertní seminář organizace CESSDA: Legální a etický rámec pro užití, znovuužití a archivaci nových typů dat. *Naše*

společnost. Dostupné z https://cvvm.soc.cas.cz/media/com_form2content/documents/c3/a4495/f28/CESSDA%20Expert%20Seminar.pdf

Ďurďovič, M.(2017). Simmelův příspěvek k teorii sociální intersubjektivity. [učební materiál]

Etické a legální otázky managementu dat (n.d.) Získáno 20.6. 2018 z http://archiv.soc.cas.cz/eticke-legalni-otazky-managementu-dat

Fonseca, A. & Cabral,B. (2017). Prototyping a GPGPU Neural Network for Deep-Learning Big Data. *Big data research*. DOI: 10.1016/j.bdr.2017.01.005

Felt. M (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society* [online]. Dostupné z https://doi.org/10.1177/2053951716645828

focus, Marketing & Social Research (2016). *Uživatelé sociálních sítí v ČR*. Dostupné z http://www.focus-agency.cz/aktuality/uzivatele-socialnich-siti-v-cr

Gan,N. (8.1.2018) China's security chief calls for greater use of AI to predict terrorism, social unrest. *South China Morning Post* . Získáno z https://www.scmp.com/news/china/policies-politics/article/2112203/china-security-chief-calls-greater-use-ai-predict

Giddens, A. (2013). *Sociologie*. (Knotková Čapková, B., překlad). Praha: Argo. (původní vydání z roku 2009)

Grunning, D. (2017). Explainable Artificial Intelligence (XAI). [Power point]. Získáno z http://explainablesystems.comp.nus.edu.sg/wp-content/uploads/2018/03/XAI%20for%20IUI%202018.pdf

Gruzd, A. (2017). Netlytic: Software for Automated Text and Social Network Analysis. [Software] Available at http://Netlytic.org

Henke, N., Bughin , J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (prosinec 2016). The age of analytics: competing in a data-driven world. Získáno z https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

Halford, S., & Savage, M. (2017). Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research. *Sociology*, Vol 51, Issue 6, pp. 1132 – 1148. https://doi.org/10.1177/0038038517698639

Heller, M. (7.8.2017). *10 hot data analytics trends — and 5 going cold*. Získáno z https://www.cio.com/article/3213189/analytics/10-hot-data-analytics-trends-and-5-going-cold.html

Hendl, J. (2008). *Kvantitativní výzkum. Základní teorie, metody a aplikace.(vydání 2)* Praha:Portál, s.r.o.

Illiadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society* [online]. Dostupné z https://doi.org/10.1177/2053951716674238

Iniciativa Průmysl 4.0. (4.5.2016). Získáno z https://www.uoou.cz/vismo/zobraz_dok.asp?id_org=200144&id_ktg=3840&n=iniciativa%2Dprumysl%2D4%2D0&p1=1271

Chang, R.M., Kauffman, R.J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems,* 63, 67-80. Dostupné z https://doi.org/10.1016/j.dss.2013.08.008

Chen, Y., & Yan, F. (2016). Centuries of sociology in millions of books. *The Sociological Rewiev*. Dostupné z https://doi.org/10.1111/1467-954X.12399

Jeong-han, K., & Da Young. Ch. (2017). Homophily in an Anonymous Online Community: Sociodemographic Versus Personality Traits. *Cyberpsychology, Behavior, and Social Networking*. Dostupné z http://doi.org/10.1089/cyber.2016.0227

Jiřovský, L. (2015). Data retention – ukládání provozních a lokalizačních údajů (Diplomová práce). Repozitář závěrečných prací UK.

Kallus, N (2014). Predicting Crowd Behavior with Big Public Data. Dostupné z http://dx.doi.org/10.1145/2567948.2579233.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society* [online]. Dostupné z https://doi.org/ 10.1177/2053951714528481

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* [online]. Dostupné z https://doi.org/10.1177/2053951716631130

Kovárník, L., Tůma, Z., & Dvořák, J. (2014). Velká data a jejich praktické využití [Power point]. Získáno z www.t-press.cz/cs/files/get?file=tk-big-data.pdf

Kramer A, Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National*

Academy of Sciences 111 (24): 8788–8790. Dostupné z http://www.pnas.org/content/pnas/111/24/8788.full.pdf

Krejčí, J. (2008). *Kvalita sociálněvědních výběrových šetření*. Praha: Sociologické nakladatelství

Krejčí, J. & Leontiyeva, Y. (2012) Cesta k datům. Zdroje a management sociálněvědních dat v ČR. Praha: Sociologické nakladatelství

Krejčí, J. (2014) *Introduction to the Management of Social Survey Data*. Praha: Sociologický ústav AV ČR,

Krsová, L. (2016). Čeští novináři na Twitteru: Analýza sociálních interakcí českého mediálního prostoru. [Diplomová práce]. Získáno z Repozitáře diplomových prací UK https://is.cuni.cz/webapps/zzp/detail/166329/

Lupač, P. (2015). *Za hranice digitální propasti*. Praha: Sociologické nakladatelství

Lupton, D. (11.5.2015) The thirteen Ps of big data. The Sociological Life. [blog] Získáno 10.5.2018 z https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/

Mazumder, A., Das, A., Kim, N., Gokalp, S., Sen, A., & Davulcu, H. (2013). Spatio-temporal Signal Recovery from Political Tweets in Indonesia. 2013 International Conference on Social Computing, 280-287. Dostupné z https://www.researchgate.net/publication/237011838_Spatio-Temporal_Signal_Recovery_from_Political_Tweets_in_Indonesia

Manovich, L. (2012). Trending: The promises and the challenges of Big social data. In Gold,M.K. (Eds.) *Debates in the digital humanities* (460-475). Minneapolis. University of Minnesota Press

Management dat: koncepce správy dat a výzkumný projekt (n.d.) Získáno z http://archiv.soc.cas.cz/management-dat-koncepce-spravy-dat-vyzkumny-projekt

Marek, F. (24.7.2017) *Big data zdokonalí vaše analýzy*. Získáno z https://www.t-mobile.cz/podnikatele-firmy/blog/big-data-zdokonali-vase-analyzy

McFarland, D., Lewis, K., & Goldberg, A. (2016): Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *American Sociologist* [online]. 2016, 47(1), 12-35. DOI: 10.1007/s12108-015-9291-8.

Městský kamerový systém (11.2.2015). Získáno z http://www.praha.eu/jnp/cz/o_meste/magistrat/odbory/oddeleni_krizoveho_managementu/mestsky_kamerovy_system.html

de Montjoye, Y., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V.D. (2014): D4D-Senegal: The Second Mobile Phone Data for Development Challenge. Dostupné z arXiv:1407.4885v2

Muenchen, R.A. *The Popularity of Data Science Software* (aktualizace 19.6.2017). Získáno 6.6.2018 z http://r4stats.com/articles/popularity/

Nenenko, E. (2017). Analysis of bank data [Diplomová práce]. Získáno z Repozitáře ČVUT https://dspace.cvut.cz/handle/10467/69569

Nešpor, Z. R. (2017) Státní a veřejné výzkumné instituce v *Slovník institucionálního zázemí české sociologie*. Získáno z https://encyklopedie.soc.cas.cz/w/Kategorie:St%C3%A1tn%C3%AD_a_ve%C5%99ejn%C3%A9_v%C3%BDzkumn%C3%A9_instituce

Nielsen, T. & Strakatý, T. (21.10.2016) Sběr dat na internetu není bez rizika. Získáno z https://www.epravo.cz/top/clanky/sber-dat-na-internetu-neni-bez-rizika-103413.html

O ústavu (n.d.) Získáno z https://www.soc.cas.cz/o-ustavu

Obecné nařízení o ochraně osobních údajů 2016 (CZ) (EU). Získáno z https://eur-lex.europa.eu/legal-content/CS/TXT/HTML/?uri=CELEX:32016R0679&from=CS

Osborne, T., Rose, N., Savage, M. (2008) Editors' Introduction Reinscribing British sociology: some critical reflections. *Sociological Review*. DOI: 10.1111/j.1467-954X.2008.00803.x

Paruchuri, V. (16.12.2016) Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Pandas, to a large extent, in data analysis projects? [první odpověď] Získáno 10.6.2018 z https://www.quora.com/Which-is-better-for-data-analysis-R-or-Python-Is-R-still-a-better-data-analysis-language-than-Python-Has-anyone-else-used-Python-with-Pandas-to-a-large-extent-in-data-analysis-projects

Parusníková, Z. (11.12.2017). Paradigma. v *Sociologická encyklopedie*. Získáno z https://encyklopedie.soc.cas.cz/w/Paradigma

Petrusek, M. (1993). *Teorie a metoda v moderní sociologii.* Praha: Karolinum

Petrusek, M. (2008). Neopozitivistická sociologie: Sociologie jako výzkum. In Šubrt, J. (2008) *Soudobá sociologie II (Teorie sociálního jednání a sociální struktury)* (1, 9-34). Praha: Karolinum

Pilnáček, M. (2016). Struktury veřejné komunikace na zpravodajském serveru. [Diplomová práce]. Získáno z Repozitáře diplomových prací UK https://dspace.cuni.cz/bitstream/handle/20.500.11956/74070/DPTX_2014_1_11210_0_407792_0_161933.pdf?sequence=1&isAllowed=y

Pomaizlová, K (2016). Právní aspekty automatizace a digitalizace [Power point]. *Konference Práce 4.0: Revoluce začala*. Získáno z http://www.pmf-hr.com/wp-content/uploads/2016/11/Pomaizlov%C3%A1-Karin.pdf

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. *WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference* (pp. 346-355). New York: ACM.

Rivron, V., Khan, M., Charneau, S., & Chrisment, I. (2016). Exploring Smartphone Application Usage Logs with Declared Sociological Information. SocialCom 2016 : The 9th IEEE International Conference on Social Computing and Networking. Dostupné z https://hal.inria.fr/hal-01378795/document

Russom, F.(2011). TDWI best practices report: Big data analytics. Získáno z https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

Saris,W.,E., & Gallhofer, I. N. (2007). Design, Evaluation, and Analysis of Questionnaires for Survey Research. New Jersey: John Wiley & Sons.

Savage, M. (2010). Unpicking sociology's misfortunes. *The British Journal of Sociology*. Dostupné z https://doi.org/10.1111/j.1468-4446.2010.01333.x

Savage, M. & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology (Sociology and its Public Face(s))*

Sedláčková, M. & Šafr, J. (2008) Koncept sociální kohezedůvěry s sociálního kapitálu v sociologii in Šubrt, J. (eds.) *Soudobé sociologie II (teorie sociálního jednání a sociální struktury)*. (2008, 309-353). Praha:Karolinum

Smirnova,E., Ivanescu, A., Bai, J., & Crainiceanu, C.M. (2018). A practical guide to big data. Získáno z https://doi.org/10.1016/j.spl.2018.02.014

SPIR NetMonitor (2017). Výzkum sociodemografie návštěvníků internetu v České Republice. Získáno 20.5.2018 z https://1.im.cz/r2/onas/socio/cz/2017/06/seznam_cz/2017_06_seznam.cz_seznam.cz_Homepage_PC.pdf

SOÚ AK ČR (2018). Výroční zpráva o činnosti a hospodaření Sociologického ústavu AV ČR, v.v.i., za rok 2017. Dostupné z https://www.soc.cas.cz/sites/default/files/soubory/vyrocni_zprava_za_rok_2017.pdf

Squicciarini, A., Griffin, C. (2014). Why and How to Deceive: Game and Complexity Results with Sociological Evidence. *Social Network Analysis and Mining* 4: 161. https://doi.org/10.1007/s13278-014-0161-0

Stephens-Davidowitz, S. (2017). Everybody Lies : Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are. New York, NY: Bloomsbury Publishing Plc

Studie Googlu: Počet Čechů připojených k internetu ze třech a více zařízení se od roku 2012 zdesetinásobil. (2017) Získáno 13.3.2018 z http://googlepresscz.blogspot.cz/2017/02/studie-googlu-pocet-cechu-pripojenych-k.html

Symons, J., & Alvarado, R. (2016). Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society* [online]. Doi: 10.1177/2053951716664747.

Šlerka, J. (3.3.2018). "Běžte z Facebooku pryč, běžte na Twitter... Tam můžeme data vaše data líp stahovat." není nad dobrou radu od odborníků:-)) #NMI18 [Facebook status]. Získáno z https://www.facebook.com/josef.slerka.

Šolcová, K. (9.4.2016). *Matfyz a ČSOB úspěšně dokončily pilotní smluvní výzkum*. Získáno z https://www.matfyz.cz/clanky/614-matfyz-a-csob-uspesne-dokoncily-pilotni-smluvni-vyzkum

Tinati, R., Halford, S., Carr, L., & Pope, C. (2014) Big Data: Methodological Challenges and Approaches for Sociological Analysis. *Sociology* . Vol 48, Issue 4, pp. 663 – 681. Dostupné z https://doi.org/10.1177/0038038513511561

Tsai, Ch., Lai, Ch., Chao, H., & Vasilakos, A. (2015). Big data analytics: a survey. *Journal of Big Data*. Dostupné z https://link.springer.com/article/10.1186/s40537-015-0030-3

Tuna, T., Akbas, E., Aksoy, A., Canbaz, M.A., Karabiyik, U., Gonen, B., & Aygun, R. (2016). User characterization for online social networks. *Social Network Analysis and Mining*. 6: 104. Dostupné z https://doi.org/10.1007/s13278-016-0412-3

Ulrych, M. & Jurczyk, T. (2014). Neuronové sítě a jejich využití. *IT Systém*. Dostupné z https://www.systemonline.cz/clanky/neuronove-site-a-jejich-vyuziti-1.htm

Úřad na ochranu osobních údajů (2012). *Provozování kamerových systémů.* Získáno z https://www.uoou.cz/files/metodika_provozovani_kamerovych_systemu.pdf

Úřad na ochranu osobních údajů (2013). *Stanovisko č.2/2006. Zpracování osobních údajů v rámci vědy*. Získáno z https://www.uoou.cz/assets/File.ashx?id_org=200144&id_dokumenty=22293

V negativním registru SOLUS má meziročně dluh o 15 tisíc občanů méně, zejména díky nové legislativě. (18.1.2018). Získáno z http://www.sid.cz/en/tiskove-zpravy-sid-a-sdruzeni-solus/18-1-2017-v-negativnim-registru-solus-ma-mezirocne-dluh-o-15-tisic-obcanu-mene-zejmena-diky-nove-legislative

Van de Rijt, A., Kang, S. M, Restivo, M. & Patil, A. (2014) *Field Experiments of Success-Breeds-Success Dynamics*. Proceedings of the National Academy of Sciences no. 19 (2014): 6934–39.

Vodáková, A. (11.12.2017). Techniky sběru informací. V *Sociologická encyklopedie*. Získáno z https://encyklopedie.soc.cas.cz/w/Techniky_sb%C4%9Bru_informac%C3%AD

Výroční zpráva ČTU za rok 2016 (2017). Dostupné z https://www.ctu.cz/vyrocni-zpravy-rok-2016

Yip, N.M., Forrest, R. & Xian, S. (2016): Exploring segregation and mobilities: Application of an activity tracking app on mobile phone. *Cities*. Vol(59), str. 156-163. Dostupné z http://dx.doi.org/10.1016/j.cities.2016.02.003

Zákon č. 101/2000 Sb., o ochraně osobních údajů a o změně některých zákonů, ve znění účinném od 1. července 2017. (ČR) Získáno z https://www.uoou.cz/zakon-c-101-2000-sb-o-ochrane-osobnich-udaju-a-o-zmene-nekterych-zakonu-ve-zneni-ucinnem-od-1-cervence-2017/ds-3109/archiv=0&p1=1271