Big Data: Pristranskost zaradi kvantifikacije

Pristranskost zaradi kvantifikacije je dandanes prisotna vse okoli nas, vendar še toliko bolj pri ljudeh, ki se ukvarjajo s podatki. Zahodno družbo smo zgradili na podatkovnih analizah in računih, pri čemur prednjačimo vse od industrijske revolucije (dasiravno zadnja leta vedno manj). Skozi leta izobrazbe nam je s šolskim sistemom in načinom razmišljanja privzgojeno zaupanje v numerične podatke, naravoslovcem pa se na univerzah ta čut izpili kolikor se ga lahko. Zanimivo je, kako hitro se zatečemo k inventarju podatkov in statistik, ki smo si jih nabrali skozi življenje, ko moramo komu kaj dokazati, ali preprosto zmagati v debati.

Obsesijo s podatki smo pripeljali do te mere, da nihče več ne podvomi v nek podatkovni okvir, ko je ta dovolj velik. Velika podatkovja se zdijo zaupanja vredna, več kot imamo spremenljivk, več lahko opišemo in bolj smo lahko gotovi v rezultate naše analize, vendar slika ni tako črno-bela.

Dober primer velikega podatkovja so te, ki jih zbira Facebook za svoje analize. Ker naši računi na Facebooku dandanes generirajo toliko objav, da je nujno potrebno rangiranje objav, tako da bo uporabniku preživljanje časa na tem omrežju kar se da zabavno. Da bi Facebook to lahko dosegel pa je moral začeti razvijati algoritem, ki bo objave razporedil od te, ki bo uporabniku najbolj zanimiva, do te, ki mu bo najmanj. Facebook za ta namen zbira velika podatkovja o posameznem uporabniku, ki so sestavljena iz veliko različnih spremenljivk. V teh podatkovjih najdemo veliko spremenljivk, ki naj b opisovale uporabnika, vse od nabora njegovih prijateljev, kaj je v preteklosti všečkal, čemu sledi, koliko časa preživi na določeno objavi in kako reagira na določene objave, vendar pa to za dober opis zagotovo ne bolj dovolj. Ne glede na to koliko spremenljivk vključimo v model, ne bomo mogli natanko napovedati, kaj bo posamezniku všeč, saj ne morem trditi, da nekoga zares poznam, če popolnoma poznam njegov facebook ali pa tudi vse facebooke njogvih prijateljev. Tega se v zadnjih letih zavedajo tudi inženirji, ki stojijo za algoritmom, ki ureja naše časovnice, zato vse bolj vključujejo mnenje uporabnikov.

V tem primeru zagotovo gre za »big data«, saj facebook zbira ogromno spremenljivk o posamezniku in jih poskuša uporabiti za analizo, podatki pa niso nujno točno, oz. je lahko njihova interpretacija popolnoma zgrešena. Tak primer do »superhiderji«, gre za ljudi, ki uporabljajo gumb hide, ki skrije objavo kot gumb prebrano na mailu, čeprav facebook to interpretira kot zelo slab signal. Če facebook tega ne bi upošteval v svojem algoritmu, to privede v zalo slabo napoved, kaj želijo te uporabniki videti na svoji časovnici in kaj ne.

Drug zanimiv primer so pametna mesta in pametne stavbe, ki naj bi s pomočjo velikih podatkovji omogočila optimizacijo delovanja mest in omogočila boljše življenje vsem meščanom. Podatki se zbirajo preko senzorjev, ki merijo fizikalne pojave (temperaturo, odvodno vodo, veter, vlažnost …), prav tako pridobivajo podatke iz naprav meščanov, poročil o kriminalu ipd. S temi podatki bi lahko optimizirali določene aspekte mestnega življenja, kot so

uporaba podzemnih železnic, prometa in avtobusov. Takoj se postavi vprašanje, ali so rešitve postavljanje s pomočjo velikih podatkovji resnično najboljše, saj ni zagotovila, da taki podatki zares odražajo problematiko, ki jo rešujemo v celoti. Namesto, da zbiramo veliko spremenljivk, bi morali iti med meščane in sprožiti dialog ter poiskati rešitve na družbenem nivoju, kot so to storili s problemom hrupnega trga Plaça del Sol, kjer so meščani s pomočjo dialoga in participacije našli rešitve in sodelovanje z metnim uradom. Rešitve, ki so optimalne za vse udeležene bomo v velikem podatkovju hitro zgrešili, saj se odgovor skriva v kompleksnih družbenih odnosih.

Trendi nastanka pametnih mest nam govorijo, da bomo lahko s pomočjo zbiranja velike količine podatkov rešili težave, ki so nastale, kot je recimo gost promet na letališčih. S pomočjo podatkov, ki jih o letališčih zberemo, bi lahko optimizirali pristanke in vzlete letal, ter skrajšali razdaljo med posameznimi leti, kar bi direktno privedlo v večje kapacitete letališč in letalskih družb. Velika podatkovja nam govorijo, da bomo tako pozitivno vplivali na vse udeležence, vendar ob boljšem pogledu hitro ugotovimo, da take rešitve ne pomenijo na primer znižanja cen letov ali pa več zaposlitev. Rešitve na te probleme lahko najdemo samo v manjših podatkovjih ali »thick data«.

Tak primer je sodelovanje etnografinje Tricie Wang z Nokijo. S svojo raziskavo na majhnem vzorcu je odkrila, da se bo trend uporabe mobilnih telefonov pomaknil v smer pametnih telefonov, kar je svetovala Nokiji. Anketne analize omenjene družbe pa so kazale, da so pametni telefoni zgolj muha enodnevnica in da se uporabna ne bo pomaknila v to smer, kar se je kasneje izkazalo za zmotno prepričanje in je Nokio skoraj pahnilo v propad.

Velika podatkovja, statistiko in modele strojnega učenja zagotovo rabimo za učinkovito odločanje, vendar pa bi bilo zmotno postaviti naše analize zgolj na big data. Za popolno razumevanje sistema moramo iti med ljudi in jih opazovati. Poskusiti moramo razumeti sistem v celoti in ne zgolj pregledovati tabel s podatki. Zagotovo pa moramo biti zadržani ob uporabi trendi besede »data-driven« in se ob takih situacijah raje prevprašamo, ali se res odločamo pravilno. Veliko bolje bi bilo uporabljati kontekst besede »data-informed« ter končno odločitev preveriti s pomočjo drugih analiz, ne samo velikih podatkovji.

»Aha« momenti:

- Facebook like gumb je bil izumljen samo za namen kvantifikacije 😊
- Veliko spremenljivk ne pomeni nujno dobrega odraza sistema, ki ga raziskujemo.
- Malo kvalitetnih podatkov lahko veliko bolje opiše sistem, kot ga lahko veliko podatkov z veliko spremenljivkami.
- Rešitve, ki jih opisujejo velika podatkovja niso nujno optimalne.
- Analiza bo zgolj tako dobra kot je dober človek, ki je zbral podatke in človek, ki jih bo analiziral in interpretiral.
- Podatki sami niso nujno objektivni, saj jih zbirajo in interpretirajo ljudje, ki pa so venomer subjektivni.