STORE SKJEVHETER VED BRUK AV STORDATA? R

AUKE HUNNEMAN er førsteamanuensis ved Institutt for markedsføring på Handelshøyskolen BI. Han har en ph.d i økonomi fra Universitetet i Groningen. Han har publisert i flere forskningstidsskrift som Journal of Retailing og Journal of Business Research. Hans forskningsinteresser er innen detaljhandel, sosiale nettverk, markedsføringsmodeller og *marketing accountability*.

SAMMENDRAG

Skaper bruk av stordata store skjevheter? Til tross for at tilgangen av stordata har bidratt til store endringer i hvordan vi designer forretningsmodeller, og fortsatt vil være en viktig driver i denne utviklingen, har flere den siste tiden kommet opp med en del kritiske merknader. Målet med denne artikkelen er å nyansere de ofte over-optimistiske forventningene til de positive effektene av stordata som verktøy. Artikkelen identifiserer ulike mulige begrensninger ved bruken av stordata. Disse begrensningene er knyttet til utfordringer både med selve analyseverktøyet og de slutningsfeil som skjer som en følge av menneskelig fortolkning (og samspillet mellom disse to). Ved å anerkjenne de utfordringene som kan oppstå ved utarbeidelse av prediksjoner og forecasting, er det mitt mål å sikre at stordata ikke kun blir en variant av Keiserens nye klær.

Stordata er i vinden! Et enkelt Google-søk på ordene «Big Data» genererer i skrivende stund (per 26. januar 2018) over 88 millioner treff. Også i næringslivspressen er det stor entusiasme. McKinsey omtaler stordata som «den mest revolusjonerende muligheten innen markedsføring og salg siden internett ble allemannseie», og *Harvard Business Review* utpekte *data scientist* (dataingeniør) til det 21. århundrets mest sexy stillingstittel (Davenport & Patil, 2012; McKinsey Chief Marketing and Sales Officer Forum, 2013). Det later ikke til å være tvil om at stordata har snudd opp ned på måten vi driver forretninger på, og vil fortsette med å gjøre det. Samtidig har det også kommet noen innvendinger i den senere tid. Stordata er en global industri til en verdi av 130 milliarder dollar, men likevel viser det seg at over 73 prosent av stordataprosjektene er ulønnsomme (Wang, 2018). Eksperter advarer dessuten bedriftene mot å satse for mye på innsamling og lagring av data uten å vite hvordan man faktisk kan generere verdier (både for kundene og for bedriften selv) av disse dataene.

Jeg mener det er på sin plass å nyansere de tidvis altfor optimistiske forventningene til de store datasettenes fortreffelighet. Misforstå meg ikke: Jeg er ikke

motstander av å bruke store datasett, men jeg vil likevel oppfordre til noen forbehold iblandet en viss kritisk sans. Jeg har identifisert flere begrensninger ved bruken av stordata som jeg drøfter i denne artikkelen. Disse begrensningene er knyttet til utfordringer både med selve analyseverktøyet og de slutningsfeil som skjer som en følge av menneskelig fortolkning (og samspillet mellom disse to). Rekkefølgen disse begrensningene drøftes i, sier ikke nødvendigvis noe om problemenes relative størrelse.

DET SOM KAN MÅLES, KAN IKKE NØDVENDIGVIS TELLES

Vi mennesker foretrekker ofte det målbare fremfor det ikke målbare. Dette fenomenet kan vi kalle *systematisk kvantifiseringsskjevhet*. Mange bedrifter investerer i dag tungt i markedsføring på digitale plattformer som Facebook and Snapchat. Dette anses for å være attraktivt, dels fordi vi kan måle hvor mange inntrykk eller tomler disse annonsene genererer, og dermed kan evaluere hvor godt de «lykkes». Det faktum at det er enklere å måle effekten av digital annonsering enn for eksempel annonsering i tradisjonelle medier som radio og fjernsyn, gjør den imidlertid ikke nødvendigvis mer virkningsfull og dermed til en bedre investering. Inntrykk kan betraktes som en indikasjon på varemerkebevissthet, men det er et langt steg videre fra dette til faktisk salg. Det å skulle selge noe i et konkurranseutsatt forretningsmiljø som Facebook, er i realiteten mye vanskeligere enn i en del tradisjonelle medier (selv om du kan målrette budskapet mer direkte mot ditt publikum).

En annen ulempe med store datasett er at de ofte ikke tar hensyn til konteksten rundt den aktuelle atferden. Dermed kan stordata dekke over viktige, men delikate problemstillinger som vanskelig lar seg kvantifisere, så som kulturforskjeller, følelser og andre mer kvalitative sider ved faktisk atferd. Netflix oppdaget for eksempel at kundene deres liker «seriefråtsing» først etter at en netnograf hadde observert seerne i lengre tid. Etter dette valgte selskapet å avvikle arbeidet med å lage et bedre anbefalingssystem for filmene til brukerne sine. I stedet henvises seerne bare til neste episode i samme serie, i stedet for andre liknende serier.

STORE DATASETT, STOR STØY?

Nyheter inneholder mye støy; historiebøker generelt generer svært lite støy. Tilsvarende vil høyfrekvente data med mye støy gjøre det vanskeligere for beslutningstakerne å vite når de skal foreta seg noe, spesielt ved data som tilgjengeliggjøres nærmest fortløpende. Dette i kombinasjon med at vi ofte reagerer sterkere på tap enn på tilsvarende vinning (Kahnemann, 2013), gjør at vi har lett for å gripe inn i situasjoner hvor det kanskje ville vært bedre å vente. Spørsmålet er altså når man skal oppsøke lege, hvis «såret» fortsatt kan leges med tiden.

VI SER IKKE SKOGEN FOR BARE TRÆR I K-STORE DATASETT

Selv om vi antar at vi kan måle og kvantifisere mange relevante variabler, kan det likevel være vanskelig å få oversikten. Datasett som har svært mange variabler kan følgelig inneholde mange spuriøse korrelasjoner. En spuriøs korrelasjon oppstår i situasjonen der to eller flere variabler korrelerer med hverandre, men hvor det i realiteten ikke er noen logisk sammenheng mellom dem. Hvis du i et stort datasett forsøker å finne en *hvilken som helst* sammenheng mellom variablene, vil du sannsynligvis lykkes. Det er nettopp dette korrelasjonsbaserte algoritmer for maskinlæring er så gode til. Slike funn gir imidlertid åpenbart svært liten mening utover disse konkrete sammenfallene, i den forstand at vi ikke kan generalisere noe ut av dem. Dette er et vektig argument for å analysere data ut fra klare teorier eller begrepsapparater. Data må håndteres med utgangspunkt i en sammenhengende historie for å finne meningsfulle mønstre. Teorienes tid er ikke over (Anderson 2008) – store datasett krever etter min mening mer heller enn mindre teori.

Faktum er at det er en uoverensstemmelse mellom det Wedel og Kannan (2016) kaller *småstatistikk på stordata* og *storstatistikk på smådata*. Førstnevnte er en vanlig tilnærming innenfor praktisk markedsføringsaktivitet der aktørene bruker for eksempel instrumentpanel med ganske naive modeller for å sette seg raskt inn i enorme datamengder. Slike tilnærminger innebærer normalt en overforenkling av mekanismen for datagenerering og medfører dermed ofte store systematiske skjevheter. Akademikere, derimot, bruker ofte komplekse modeller på relativt små datasett, og det kan også være en utfordring. Så komplekse modeller fanger ofte opp hver eneste lille variasjon i tallmaterialet (også utvalgs- og målefeil), med overtilpasning som resultat (Silver, 2012). Med store datasett kan vi

unngå slike problemer, forutsatt at vi finner en god modell som representerer datagenereringsmekanismen. Større utvalg og større utregningskraft reduserer både variasjonen i datasettet og ressursene som kreves for å bearbeide dem, som igjen åpner for å estimere mer komplekse modeller som mer presist avdekker bakenforliggende mekanismer.

INTERNETT ER ET VINN-ELLER-FORSVINN-MARKED

Det skyldes to ting: Alt som selges på nettet, er skalerbart, det vil si at det er like lett å selge enkeltvis eller millionvis til lav marginalkostnad.

I tillegg er det flere kontaktflater mellom forbrukerne på nettet, med blant annet den effekten at «de rike blir rikere». Det forklarer hvorfor bestselgerlistene til Amazon (og Spotify) domineres av noen få titler (sanger) som lastes ned millioner av ganger. Dette gjør at man får en lang hale i fordelingen av salget (Anderson 2009), der noen få produkter står for mesteparten, mens mange andre (i forlengelsen av fordelingen, den lange halen) selger mye mindre. Ett resultat av dette er at Google kontrollerer nesten 90 prosent av all søkebasert annonsering på nettet, Facebook nesten 80 prosent av mobil sosial trafikk, og Amazon rundt 75 prosent av salget av e-bøker (Kolbert, 2018). Denne forsterkende effekten i form av sosial innflytelse gjør det vanskeligere å forutsi hvor godt et produkt vil klare seg i markedet. Et eksempel: Salganik, Dodds og Watts (2005) gjorde et forsøk der de ba folk om å evaluere kvaliteten på flere nye sanger. De spurte også hvem som ville laste ned de aktuelle sangene for å høre på dem igjen senere. Det viste seg at vurderingen av kvalitet ga en pekepinn om hvor godt produktene ville lykkes, altså antall nedlastinger. Så langt alt vel. Men hvis deltakerne fikk vite hvor mange andre som hadde lastet ned de ulike sangene, snudde de kappen med vinden, og da var det ikke lenger noen klar sammenheng mellom produktkvalitet og hvor godt produktet klarte seg i markedet. Av dette kan vi lære at når markeder betraktes som komplekse systemer der forbrukerne kan påvirke hverandre, er det ikke nødvendigvis et én-til-én-forhold mellom produktkvalitet og suksess. Hvor godt produktet vil lykkes, er faktisk nesten umulig å forutsi i slike omgivelser. Derfor er det vanskelig å forutsi hvor vi vil finne den neste Mona Lisa eller Harry Potter-bok.

En annen viktig lærdom vi kan trekke av dette, er at vi trenger svært store utvalg for å kunne trekke slutninger på områder der fordelingen har lang hale. Dette gjelder både tverrsnitts- og tidsrekkedata. Hvis utvalget ditt for eksempel ikke inneholder den siste Harry Potter-boken, kan du komme til å trekke fullstendig urealistiske slutninger om hva det er som driver boksalget. Slik du også ville gi loddsalget en annen vurdering hvis du ikke tok med vinneren. Eller hvis du, slik flyplasslitteraturen gjør, utelukker mislykkede forretningsfolk og kun tar med de vellykkede når du gir råd om hvordan man blir rik. Tilsvarende må vi anta at datautvalg som strekker seg over lengre tid, inneholder mer av de sjeldne fenomenene (f.eks. diller og påfunn) som vi ganske enkelt ikke kan overse i dagens verden, hvor alt henger sammen med alt. Så selv om vi fortsatt gjerne vil prøve å spå om fremtiden, trenger vi derfor ekstremt store utvalg: Det største fjellet er ikke nødvendigvis det største fjellet i utvalget ditt.

PROBLEMET MED FRIHETSGRADEN

La oss se litt på alle sporene som folk legger igjen etter seg når de forteller oss hvordan de har det, oppfører seg og samspiller i omgang med produkter og tjenester. Det viser seg å være svært vanskelig å utlede av disse sporene nøyaktig hva det er som har ført til dem (eller med andre ord å utvikle en gangbar teori ut fra dem). De samme sporene kan ha blitt lagt igjen på så mange måter at vi ikke uten videre kan si at «det må være dette som har skjedd». Taleb (2010) sammenlikner det med det med hvordan en vanndam på gulvet er et resultat av en isbit som stadig smelter. Isbiter i alle mulige former kan lage nøyaktig den samme vanndammen (den trenger ikke en gang nødvendigvis være resultat av en isbit). Poenget er at samme konsekvens kan være et resultat av mange forskjellige foregående begivenheter. Det forklarer hvorfor ulike mennesker tolke samme hendelsesforløp på helt forskjellig måte. Dette, kombinert med menneskets tendens til *narrative feilslutninger*, altså til å sette sammen fakta gjennom narrativer eller fortellinger, forklarer også hvorfor folk i perioder med informasjonsflom og altfor mye data har lettere for å ty til konspirasjonsteorier (Harari, 2015).

LIKE BARN LEKER BEST

Internett har gjort det mulig for oss å komme i kontakt med likesinnede til en lav kostnad. Denne mekanismen,

på engelsk kalt *homophily*, kan skape såkalte ekkokamre og i ytterste konsekvens segregering. I dagens verden, hvor alt henger sammen med alt, kan vi (bevisst eller ubevisst) utsette oss utelukkende for holdninger og meninger som tilsvarer og harmonerer med våre egne. Resultatet er at vi ofte søker etter informasjon som bekrefter våre egne overbevisninger. Det sier seg selv at slik *systematisk bekreftelsesskjevhet* undergraver evnen til nytenkning, men den gjør det også vanskeligere å oppnå kontakt med annerledestenkende. Det ironiske er at digitale plattformer gjør det enkelt for bedriftene å rette budskapet sitt direkte mot individuelle kunder, men det er det ikke nødvendigvis behov for hvis folk oppfører seg stadig mer likt hverandre og har liten «rekkevidde» utover sine egne grupperinger av likesinnede. Det er allerede lenge siden Granovetter (1977) appellerte til å bruke det han kalte «styrken i de svake bånd». Verdien av tilfeldige møter med fremmede (de svake båndene) er at slike møter kan tilføre ny informasjon som ikke er direkte tilgjengelig for deg i det nettverket du befinner deg i akkurat nå. Hvis du for eksempel er jobbsøkende, kan det være nyttig å ty til slike svake bånd. De vil gi deg et konkurransefortrinn fremfor andre som ikke har direkte tilgang til den nye informasjonen som svake bånd kan tilføre.

Kort oppsummert kan vi si at optimismen rundt store datasett kan ha nådd toppen av trendkurven i Gartners såkalte *Hype Cycle* (Gartner, 2018). I denne artikkelen har jeg forsøkt å gi leserne et realistisk perspektiv på utfordringene ved å skulle hente mening ut av datamettede omgivelser. Vi må være bevisst på snubletrådene ved bruk av prognoser og spådommer, og sikre at store datasett ikke blir bare nok et sett med Keiserens nye klær. M

Forfatteren takker Henrik Jensen (ved Handelshøyskolen BI) for konstruktive innspill til en tidligere versjon av denne artikkelen.

SITERTE KILDER

Anderson, C. (2009). *The longer long tail: How endless choice is creating unlimited demand*. London: Random House Business Books.

Anderson, C. (16.7.2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*. Hentet fra https://www.wired.com/2008/06/pb-theory (hentedato 29.1.2018).

Davenport, T.H., & D.J. Patil (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(10), 70–76.

Gartner (2018). Research Methodologies: Gartner Hype Cycle. Hentet fra https://www.gartner.com/technology/research/methodologies/hype-cycle.jsp (hentedato 5.2.2018).

Granovetter, M. (1977). The strength of weak ties. *Social Networks, 78*(6), 1360–1380.

Harari, Y.N. (2015). *Sapiens: A brief history of humankind*. New York: HarperCollins.

Kahnemann, D. (2013). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Kolbert, E. (28.8.2017). Who owns the Internet? What Big Tech's monopoly powers mean for our culture. *The New Yorker*. Hentet fra https://www.newyorker.com/magazine/2017/08/28/who-owns-the-internet (hentedato 5.2.2018).

McKinsey Chief Marketing and Sales Officer Forum (2013). *Marketing & Sales Big Data, Analytics, and the Future of Marketing & Sales*. McKinsey & Company.

Sagalnik, M.J., P.S. Dodds, og D.J. Watts (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5762), 854–856.

Silver, N. (2012). *The signal and the noise: The art and science of prediction*. London: Penguin Books.

Taleb, N.N. (2010). *The black swan: The impact of the highly improbable*. London: Penguin Books.

Wang, T. (2018). *The human insights missing from big data* [video-fil]. Hentet fra https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data (hentedato 30.1.2018).

Wedel, M., & P.K. Kannan (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121.