NORSK SOSIOLOGISK TIDSSKRIFT
Norwegian Journal of Sociology
BOKANMELDELSE

UNIVERSITETSFORLAGET

Årgang 3, nr. 3-2019, s. 231–234
ISSN online: 2535-2512
https://doi.org/10.18261/issn.2535-2512-2019-03-06

Bokanmeldelse

Matthew Salganik
Bit by bit
Princeton: Princeton University Press, 2017

ISBN: 9781400888184

Torkild Hovde Lyngstad
Professor
Universitetet i Oslo
t.h.lyngstad@sosgeo.uio.no

Åpent tilgjengelig på http://bitbybitbook.com

Dan Ariely sa i 2013 at «Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.» Vi har snakket lenge om «Big data», og nå har også «algoritmer» sklidd inn i hverdagsvokabularet. Hvilke konsekvenser har veksten i digitale spor og datatilgang for samfunnsforskning? Har det hele fislet ut, som Ariely påsto, eller er samfunnsforskningens vilkår radikalt forandret?

Du trenger ikke lure på hva Matthew Salganik mener om disse spørsmålene. Den digitale tidsalderen er her, den vokser og den endrer rammevilkårene for samfunnsforskning. Omtrent slik begynner i hvert fall *Bit by bit*. Salganik er professor og ekspert på nettverksanalyse og analytisk sosiologi ved Princeton University. Han er blant de fremste innen feltet som på engelsk kalles «computational social science» og kan oversettes til norsk som *data-intensiv samfunnsvitenskap*. Boken som har vært til «open review» ligger fritt tilgjengelig på internett, selv om den også finnes i papirutgave.

Hvordan endrer så digitaliseringen og veksten i digitale spor samfunnsforskeres rammevilkår? Hva kan vi gjøre nå som vi ikke kunne gjøre tidligere? *Bit by bit* gir gjennomgående gode eksempler på hvordan vi nå kan utnytte Big data og teknologi for å flytte forskningsfronten. Det handler ifølge Salganik om fire hovedendringer.

Vi kan observere samfunnet og dets deltakere på nye måter. I den «analoge» tidsalderen var det svært dyrt å samle inn data om menneskers holdninger og atferd. Følgelig ble det ikke gjort særlig ofte eller i stort omfang. Alle som har erfaringer med surveydesign, kjenner

til avveiinger av nytten av et ekstra spørsmålsbatteri mot kostnadene batteriet vil medføre. I vår digitale tidsalder er derimot data overalt: Det samles inn digitale spor om milliarder av mennesker, deres atferd og preferanser. Disse dataene kan samles inn, bearbeides og analyseres. Navnet har slike data fått fordi de er biproduktet av menneskers handlinger i hverdagen, de er sporene av vår virksomhet som dyrespor i naturen.

Hovedutfordringen er ikke lenger at vi mangler data. Det er snarere å få tilgang til, bearbeide og i praksis analysere store datamengder som allerede finnes «der ute». Hva kjennetegner så Big data?

Salganik setter ti merkelapper på Big data. De er 1. Store (og ofte upraktisk store). 2. Alltid i endring, siden data samles inn kontinuerlig. 3. Areaktive: observasjon endrer i seg selv sannsynligvis ikke atferd. 4. Ukomplette: de mangler sannsynligvis informasjon du trenger. 5. Utilgjengelige, fordi de kontrolleres av private eller offentlige aktører som i utgangspunktet ikke driver forskning. 6. Ikke-representative, siden det ikke er tilfeldig hvem som legger igjen digitale spor. 7. I drift: de endres med endringer i systemene som samler dem inn. 8. Formet av algoritmer: Big data er formet av systemene som samler dem inn. 9. Skitne, og krever betydelig bearbeiding. 10. Sensitive, og kommer med sine særegne forskningsetiske utfordringer.

Selv om de klassiske eksemplene på Big data er digitale spor fra sosiale medier og annen internettbruk, gjør Salganik oss også oppmerksom på flere andre typer. Store bedrifter og organisasjoner har for eksempel kunderegistre som kan tjene som forskningsdata, og stater og offentlige byråkratier har administrative data som kan tjene samme formål. (Det siste er selvfølgelig intet nytt for skandinaviske forskere som i tiår har brukt registerdata i banebrytende statistiske analyser, men for den jevne samfunnsforsker er det offentliges dataarkiver nye datakilder.)

Samfunnsforskere kan stille spørsmål på nye måter. Mye samfunnsvitenskapelig virksomhet handler om å stille mennesker spørsmål i ulike spørreundersøkelser. Svarprosentene i slike undersøkelser er nå rekordlave, men kostnadene ved datainnsamlingen er fortsatt høye. Man kunne trodd at surveyforskeren vil dø ut, og den representative spørreundersøkelsen med henne. Digitaliseringen er en mulig redning: Internett gjør at datainnsamling kan bli lettere, f.eks. gjennom bruk av online surveys. Slike data vil ofte ha svært skjeve, ikke-representative utvalg, men utviklingen i metoder for å justere for utvalgsskjevheter og mulighetene for koblinger til andre digitale datakilder, gjør at slike skjevheter ikke er like dødelige som de har vært.

Vi kan eksperimentere i stort, på internett. Eksperimentell samfunnsforskning har vært begrenset til små studier. Både laboratorieeksperimenter og felteksperimenter har vært vanskelige å gjennomføre i virkelig stor skala. Digitaliseringen sprenger grensene for eksperimentell datainnhenting. Et enkelt eksempel er forskningsprosjektet GEMM, som har gjennomført et stort felteksperiment i arbeidsmarkedet i en rekke europeiske land – ved hjelp av helt eller delvis automatiserte systemer for sending av jobbsøknader. Andre eksempler er eksperimenter på sosiale medier, hvor eksperimentgrupper blir eksponert for en bestemt type innhold i stor skala.

Vi kan samarbeide i stor skala. Internett muliggjør samarbeid i stor skala. Salganik bruker Wikipedia som et eksempel. Wikipedia ble ikke mulig på grunn av ny kunnskap, men på grunn av en ny type samarbeid. En fritt redigerbar, hyperlenket dokumentsamling

kalles nå for en Wiki. Slikt storskala samarbeid er en ny mulighet også for samfunnsforskning.

Storskala digitalt samarbeid kan ta flere former. En mulighet er å dele på enkle oppgaver. For eksempel kan koding av ulike typer data nå gjøres på nett av mange personer samtidig, så lenge kodingen ikke krever ekspertkunnskap.

Om man selv ikke har noen tydelig løsning på et problem, kan man nå – ved hjelp av elektronisk kommunikasjon – lett invitere andre forskere til å foreslå løsninger. Et slikt samarbeid er the Fragile Families Challenge.[1] Dette var en konkurranse hvor gruppen bak spørreundersøkelsen Fragile Families oppga en overordnet problemstilling og gav tilgang til dataene sine til deltakerne. Deltakerne, som kom fra hele verden og et bredt spekter av disipliner, leverte så løsningsforslag som bidrag i en konkurranse.

Distribuert datainnsamling er en tredje form for samarbeid. I prosjekter hvor tidligere en hær av forskningsassistenter har utført relativt enkle innsamlingsoppgaver, kan disse nå «outsources» til bidragsytere over nett. Mobilapper har på kort tid blitt en viktig innsamlingsmetode, men mange andre kan brukes. Det finnes mange naturvitenskapelige eksempler på slik distribuert innsamling av data, men stadig flere sosiologiske prosjekter bruker slike systemer. Et hjemlig eksempel er Willy Pedersens prosjekt MittBlikk, hvor ungdommer sender bilder fra sine mobiltelefoner som så analyseres av forskere.

Utover disse hovedtemaene dekker *Bit by bit* også en rekke forskningsetiske aspekter ved digitale spor og bruken av slike, og boken har et kort kapittel om hva fremtiden kan bringe.

Hva dekker den så ikke? Salganik avgrenser tematikken boken tydelig i første kapittel. Boken gir ikke noen innføring i hvordan man praktisk arbeider med Big data, storskala samarbeid eller andre anvendelser. Det er klokt, fordi feltet endres raskt. En bok kan bli utdatert mens korrekturen leses. Heldigvis har dataintensiv samfunnsvitenskap arvet den sterke delingskulturen som finnes i den digitale sfæren.[2] Det er derfor svært mye å lære av andre som deler sine prosjekter, og det finnes mye læremateriell fra kurs og møter, f.eks. på GitHub. Et sted å begynne kan være websidene til den internasjonale sommerskolen i dataintensiv samfunnsvitenskap.[3]

Ifølge forfatteren har boken to typer lesere. Den ene typen er samfunnsvitere som ønsker å bedrive mer «data science» og bruke digitaliseringen til sin fordel. Den andre typen er «data scientists», personer med kunnskaper om programmering og anvendt statistikk, men uten samfunnsvitenskapelig bakgrunn. De sistnevnte arbeider i privat sektor og driver med analyser og forskningslignende arbeid. Dette gjelder ikke bare internettgigantene Facebook og Google, men også norske selskaper. Et enkelt søk på *finn.no* vil vise leseren at det er kamp om de beste «data scientists» i det norske arbeidsmarkedet.

Hvordan passer så sosiologer, og mer spesifikt norske sosiologer, inn i dette? Etter som datarevolusjonen ruller videre, vil sosiologer og andre tradisjonelle samfunnsvitere få

1. Se http://www.fragilefamilieschallenge.org/ for ytterligere informasjon.
2. Mange programutviklere og «hackere» deler sine prosjekter fritt i databaser, som for eksempel GitHub. Etter hvert som kravene til replikasjon av forskningsfunn øker, vil også flere samfunnsvitere måtte ta skrittet over i denne delingskulturen og dele sine programmer og data med andre.
3. https://compsocialscience.github.io/summer-institute/2017/#schedule

mange konkurrenter som vil mestre de tekniske og praktiske sidene ved digital samfunnsvitenskap, selv om de har en svakere teoribase. Om sosiologer skal forsvare sitt terreng i møte med slik konkurranse, må de være klar over og villige til å utnytte digitale data. I så måte er *Bit by bit* et utmerket utgangspunkt.