BIG DATAS TITANIC?

Influenza-trackingtjenesten Google Flu Trends startede som en succes, men endte som en fiasko. Flu Trends viser de muligheder, der ligger i brugen af big data, men illustrerer samtidig de erkendelsesteoretiske faldgruber og etiske dilemmaer, som brug af store datasæt rummer.

Forfatterne

Mikkel Willum Johansen er lektor ved Sektionen for Videnskabsteori og Videnskabshistorie ved Institut for Naturfagenes Didaktik, Københavns Universitet. Hans forskning og undervisning ligger inden for videnskabsteori og etik med særligt fokus på matematikkens videnskabsteori.
mwj@ind.ku.dk

Henrik Kragh Sørensen er professor MSO samme sted. Hans forskning ligger inden for matematikkens og datalogiens videnskabshistorie og videnskabsteori.
henrik.kragh@ind.ku.dk

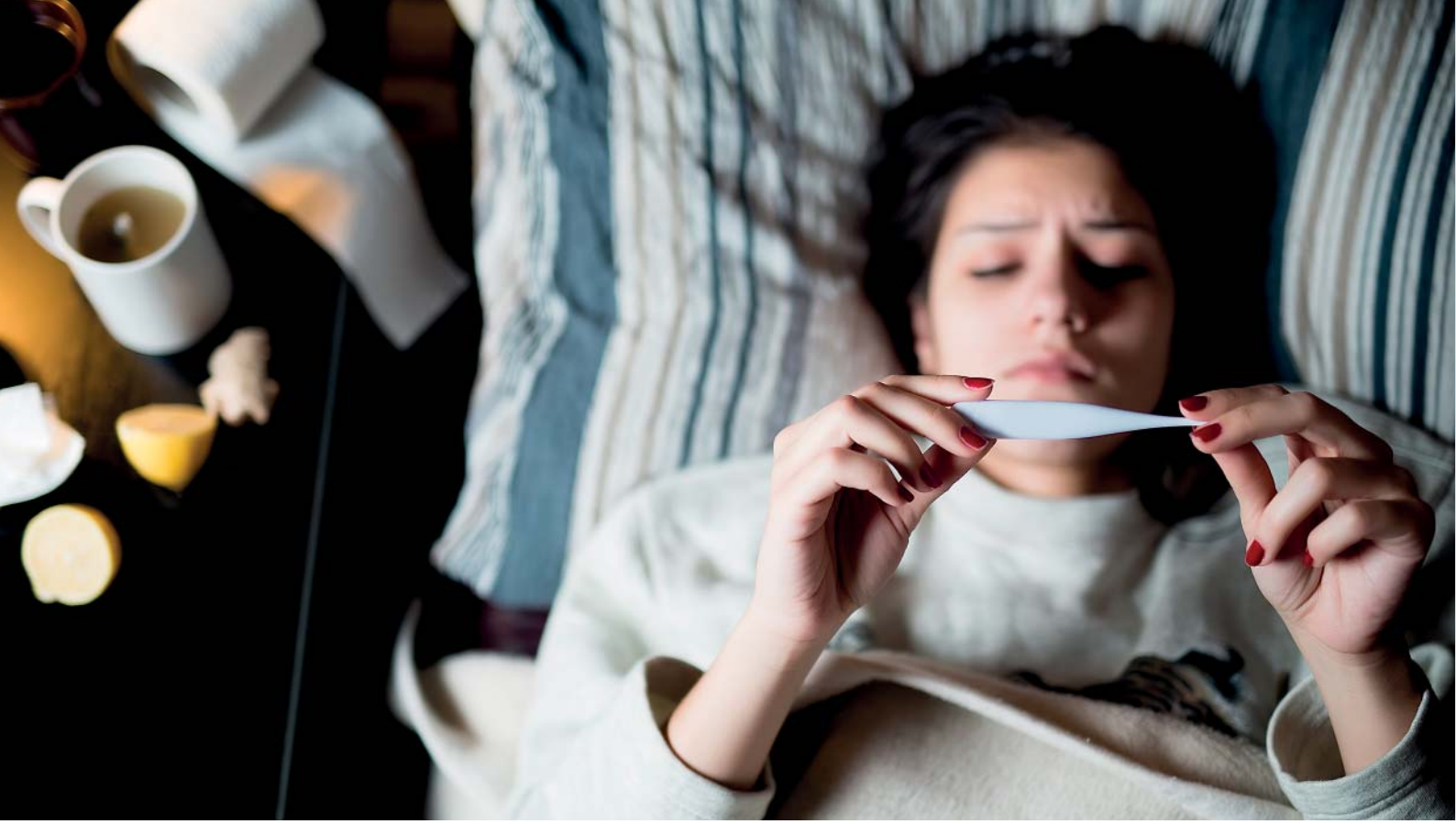
I 2008 lancerede forskere fra Google tjenesten *Flu Trends*, der ud fra folks googlesøgninger søgte at følge de årligt tilbagevendende influenzaepidemiers udbredelse. Influenza er en global dræber, der hvert år koster mellem 250.000 og 500.000 mennesker livet. Viden om, hvornår en epidemi rammer, gør det lettere at sætte ind med forebyggelse og behandling, og derfor vil et system, der effektivt kan følge en epidemi, potentielt kunne redde tusindvis af menneskeliv. Traditionelt har man fulgt influenzaepidemiers udbredelse ved hjælp af data, som de praktiserende læger indberetter. Desværre er der en vis forsinkelse på disse tal, da lægerne først skal indberette, hvor mange influenzatilfælde, de ser, og data herefter skal samles og offentliggøres. Med sine fuldautomatiske analyser af søgedata kunne Flu Trends derimod følge epidemiens udbredelse stort set i realtid, og tjenesten blev udråbt som en stor sejr for brugen af big data. Men i august 2015 lukkede Google Flu Trends efter flere spektakulære fejlforudsigelse, og tjenesten er ikke blevet åbnet igen siden. Så hvad skete der?

Big data og influenza

Før vi rigtig kan forstå, hvordan Flu Trends kunne være så lovende og alligevel gå så galt, skal vi vide lidt om big data. Selve termen "big data" kom på mode i begyndelsen af dette årti. Der findes ikke nogen officiel definition af termen, men typisk forbinder man big data med søgning efter statistiske sammenhænge i store, men "beskidte" og rodede datasæt. Og det var præcis, hvad Google gjorde med Flu Trends: Man havde adgang til et enormt datasæt i form af tid, sted og indhold af alle søgninger, der nogensinde var lavet på Google. Ved at lede efter statistiske sammenhænge – korrelationer – mellem de 50 millioner mest almindelige søgetermer og antallet af influenzapatienter i de forskellige delstater i USA i årene 2003-2008 identificerede forskerne fra Google en række søgninger, der var statistisk korreleret med influenzaepidemiernes udbredelse. På den baggrund fandt de ud af, at de ved hjælp af de 45 søgetermer, der var bedst korreleret med influenza, kunne bygge en model, der passede på de historiske influenzatal. Denne proces kaldes at "træne" algoritmen og består altså i, at man tilpasser en model til historiske tal i håbet om, at modellen så vil kunne sige noget fornuftigt om fremtiden. Google har aldrig offentliggjort de termer, der indgik i modellen, men det kunne for eksempel være "hovedpine" eller "ledsmerter". Folk har selvfølgelig ikke nødvendigvis influenza, hver gang de søger på disse ord, og nogle af influenzapatienterne med hovedpine kommer måske til at skrive forkert og søger på "hodedpine" i stedet. I den forstand var Googles data beskidte og rodede, men fordi datasættet var stort nok, ville urenhederne forsvinde i mængden. Eller det var i det mindste forskernes håb. Og til at starte med gik det faktisk fint – i vinteren 2008/9 gav Flu Trends imponerende præcise tal for influenzasmitten i USA, og de leverede dem 14 dage hurtigere end sundhedsmyndighederne kunne.

De mange dimensioners forbandelse

Flu Trends illustrerer nogle af de metodologiske udfordringer, man står med, når man bruger big data. Den første udfordring, som forskerne fra Google måtte håndtere, kan kaldes *de mange dimensioners forbandel-*



se. Forskerne ledte efter statistiske sammenhænge mellem de 50 millioner mest almindelige søgninger og hyppigheden af influenza, men når man holder så mange variable (50 millioner) op mod en enkelt anden variabel, er der risiko for, at nogle af de korrelationer, man finder, skyldes rene tilfældigheder. Forestil dig for eksempel, at du nummererer 50 millioner mønter og hver uge i et år slår plat eller krone med dem. Der er nu en god chance for, at nogle af mønterne vil være "heldige"' i den forstand, at de for det meste giver plat i uger, hvor antallet af influenzasmittede er i vækst, og krone på uger, hvor antallet af influenzasmittede aftager. Du kan med andre ord observere en korrelation mellem møntens adfærd og influenzasmitte, men det betyder naturligvis ikke, at du vil kunne bruge den heldige mønt til at forudsige antallet af influenzasmittede. Men hvordan kan man vide, om en søgeterm, der er korreleret til antallet af influenzasmittede, er en heldig mønt? Svaret er, at det kan man ikke (om end risikoen bliver mindre, hvis man har nok data). Hvis der er en statistisk sammenhæng mellem for eksempel søgetermen "sort fløjl" og influenza er det måske en tilfældighed, og måske en opdagelse af en hidtil ukendt og overraskende sammenhæng mellem influenza og hang til blødt, mørkt stof. I big datas historie er der eksempler både på heldige mønter og på opdagelsen af genuint nye og overraskende sammenhænge.

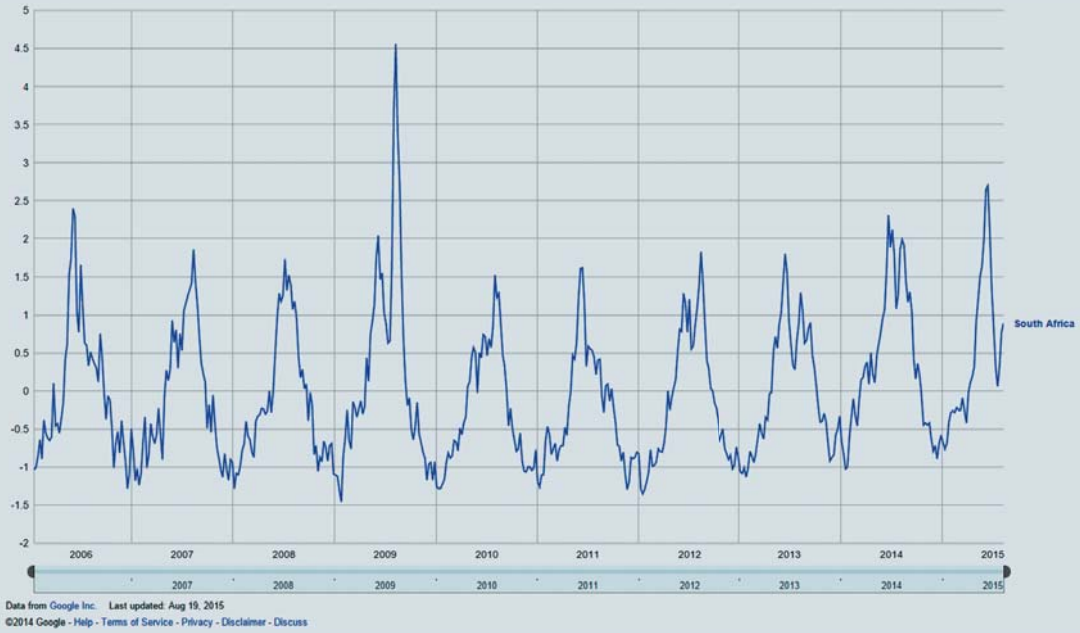
Problemet med sammenfaldende fænomener

En anden udfordring, der møder big data-analyse, er, at det fænomen, man er interesseret i, nogle gange kan falde sammen med andre fænomener; man kunne kalde det *problemet om sammenfaldende fænomener.* Influenza falder eksempelvis typisk, men ikke altid, sammen med vinter, mørke, kulde osv. Og hvordan kan man så vide, om den statistiske model forudsiger det ene fænomen eller det andet? Forskerne fra Google opdagede for eksempel, at "high school basketball" var i top 100 over søgetermer, der er bedst korreleret til influenza, men det skyldes formentlig et sammenfald mellem influenza- og basketballsæsonen, og de inkluderede ikke termen i deres model. Alligevel blev den første udgave af Flu Trends tilsyneladende ramt af problemet med sammenfaldende fænomener. I det datasæt, Google havde brugt til at træne Flu Trends, faldt influenzasæsonen pænt om vinteren, men i 2009 ramte influenzaen usædvanligt tidligt, og Flu Trends undervurderede markant antallet af smittede. Det tydede på, at en del af de søgetermer, der indgik i modellen, var korrelerede med vinter snarere end med influenza. En del af Flu Trends var altså ikke andet end en kompliceret metode til at holde øje med vinterens komme. Google justerede herefter modellen, men efter at have fungeret rimeligt et par år begyndte Flu Trends i 2011 systematisk at overvurdere antallet af smittede, og i 2013 fortalte den, at der var omkring dobbelt så mange influenzasmittede, som der reelt var.

Induktionsproblemet

Den tredje, og beslægtede, udfordring for big data-analyse er det såkaldte *induktionsproblem*, der påpeger, at det er usikkert at generalisere fra en stikprøve, da vi ikke kan vide om resten af verden opfører sig som den lille del, vi har set på. Så vidt vi ved, skyldes Flu Trends' overvurdering af antallet af influenzatilfælde netop induktionsproblemet i den variant, der siger, at fremtiden ikke nødvendigvis vil udforme sig på samme måde som fortiden. I naturvidenskaberne er denne del af induktionsproblemet mest teoretisk – det er forholdsvist fornuftigt at antage, at i det mind-

Et eksempel fra Google Flu Trends, der viser "influenza-søgeaktiviteten" i Sydafrika fra ca. 2006 til 2015. Aktiviteten vises som en afvigelse fra det normale (baseline).



ste de grundlæggende naturlove er stabile over tid – men når man beskæftiger sig med menneskelig adfærd, er problemet akut. For mennesker ændrer adfærd; herunder også søgeadfærd på Google. Vi ved ikke præcist hvorfor, men af en eller anden grund – måske på grund af en øget medieomtale af influenza, måske fordi Google ændrede sin søgealgoritme – begyndte folk at søge mere på de termer, Flu Trends associerede med influenzasmitte. Og det fik modellen til at tage fejl.

Big data og den videnskabelige metode

I en berømt og berygtet artikel fra 2008 luftede chefredaktøren af magasinet *Wired*, Chris Anderson, en vision om, at vi med big data helt ville slippe af med teori i videnskaben. I stedet for at lede efter årsager og forklaringer, sådan som videnskaben havde forsøgt i hvert fald siden Den videnskabelige Revolution i 1600-tallet, skulle ny videnskabelig viden produceres ved at lede efter statistiske sammenhænge i store datasæt:

»Dette er en verden, hvor store mængder data og anvendt matematik vil afskaffe et hvert andet redskab i værktøjskassen. Ud med enhver teori om menneskelig adfærd, fra lingvistik til sociologi. Glem alt om taksonomi, ontologi og psykologi. Hvem ved, hvorfor folk handler, som de gør? Pointen er, at de gør det, og vi kan følge og måle deres handlinger med hidtil ukendt præcision. Hvis vi bare har nok data, vil tallene tale for sig selv.«

Fra et videnskabsteoretisk og -historisk synspunkt rummer Andersons vision et ekko af den såkaldte positivisme, der dominerede videnskabsteorien i den første halvdel af 1900-tallet. Den grundlæggende idé i positivismen var, at videnskaben er en *induktiv* proces, hvor man gennem fordomsfrie og objektive analyser skulle søge lovmæssigheder i data og bagefter bekræfte hypoteserne eksperimentelt. Det kommer meget tæt på Andersons forestilling om, at tallene kan "tale for sig selv".

Selv om den grundlæggende idé bag positivismen kan have en intuitiv appel, blev positionen forladt i den anden halvdel af 1900-tallet, da videnskabsteoretikere som Karl R. Popper og Thomas S. Kuhn påviste en række fundamentale problemer i positivismens videnskabssyn. Specielt påpegede de, at observationer aldrig kan være teoriløse, og at positivismens induktive metode både er forbundet med usikkerhed og er en dårlig motor for videnskabelige fremskridt. Som vi så det, er big data stadig udfordret af positivismens problemer. Flu Trends blev hårdt ramt af induktionsproblemet, og forskerne fra Google gik ikke helt teoriløst til sagen. Hele ideen med Flu Trends byggede på en teori om, at folks søgeadfærd afspejler deres sundhedstilstand. Forskerne var godt klar over, at en søgning på for eksempel basketball formentlig kun havde en overfladisk sammenhæng med influenza, og desuden rummede modelleringsprocessen en række teoridrevne og pragmatiske valg, som spændte lige fra definitionen af influenzapatienter til valget af statistiske metoder. Flu Trends' udfordringer illustrerer, hvor sårbare rene statistiske metoder kan være. Store datasæt er ikke i sig selv et mirakelmiddel, der kan løse alle videnskabsteoretiske problemer og gøre videnskabelig opdagelse til en simpel og objektiv proces – hvilket de fleste brugere af big data heldigvis er klar over.

Når big data går godt

Flu Trends var måske nok en fiasko, men det betyder på ingen at den fagre nye verden, data-dreven

videnskab lover, er et fatamorgana. Big data har også haft store sejre og er på mange måder en frugtbar tilføjelse til den videnskabelige værktøjskasse. Big data-analyse er for eksempel blevet brugt med stor succes indenfor områder som medicinsk diagnostik, markedsføring og risikovurdering. Cancerdiagnostik, filmanbefalinger på Netflix og sikkerhedsfirmaers detektion af usædvanlig adfærd på virksomheders netværk benytter alle big data-analyse. Det mest kendte eksempel er dog formentlig Google Translate. Da man i 1950'erne først fik den ide at bruge computere til at oversætte tekst, troede man, at opgaven kunne løses, hvis man bare fodrede computeren med en ordbog og et sæt grammatiske regler. Men det viste sig overraskende nok, at den tilgang hurtigt kom til kort. Det store gennembrud kom først, da forskere fra IBI fandt på at bruge referater fra det Canadiske parlament som en big data-resurse. Canada har både engelsk og fransk som officielle sprog, og derfor fandtes referaterne på begge sprog. Med en ren statistisk analyse kunne forskerne bygge en model for, hvordan et givet engelsk ord hyppigst bliver oversat til fransk og omvendt. Google overtog ideen, men skruede kraftigt op for datamængden ved at inddrage alt det oversatte materiale, de overhovedet kunne finde – uanset kvalitet – lige fra brugsanvisninger til referater fra EU-parlamentet. Som brugere af Google Translate ved, er tjenestens oversættelser langt fra perfekte, men de er dog for det meste nogenlunde brugbare. Tjenesten giver dermed en god fornemmelse af, hvad man kan opnå med en ren statistisk analyse af beskidte og usikre data, når bare man har nok af dem.

Big data og privatlivet

Big data har også haft andre og mindre glorværdige succeser. Ved at lave en big data-analyse af kundernes købemønstre lykkedes det for eksempel i 2002 for den amerikanske supermarkedsgigant Target at lave en model, der kunne afgøre om en kunde var gravid. Tilsvarende var kernen i den såkaldte Cambridge Analytica-skandale, der rullede i det tidlige forår 2018, en big data-model, der kunne bestemme folks personlighedstype på baggrund af deres adfærd på Facebook. Ingen af de to modeller havde til formål at skabe viden, man ikke kunne opnå sikrere med andre metoder – hvis en kvinde vil vide, om hun er gravid, gør hun klogt i at købe en graviditetstest frem for at køre sine gamle supermarkedsboner gennem Targets model. De to modeller havde derimod til formål at udlede private og følsomme oplysninger fra oplysninger, vi er mere villige til at give fra os, og det rejser nogle åbenlyse etiske spørgsmål. Man kan argumentere for, at vi har en grundlæggende ret til selv at bestemme, hvem vi deler hvilke oplysninger om os selv med. Hvis det er tilfældet, krænker big data-analyser af Target-typen vores autonomi, idet de kortslutter processen, hvor vi beslutter, hvem vi deler hvad med; at give et supermarked lov til at se, hvad man putter i indkøbskurven, er ikke det samme som at give det lov til at finde ud af, om man er gravid. Og hvor mange ville egentlig åbne en Facebookkonto, hvis man som en del af tilmeldingsproceduren skulle udfylde en personlighedstest?

Big data-etik

Udover spørgsmålet om autonomi og privatliv har navnlig spørgsmål om profilering og algoritmisk bias vakt etisk bekymring i forhold til brugen af big data. Profilering består groft sagt i, at man bliver sat i bås med en gruppe af andre mennesker udelukkende ud fra overfladiske kendetegn. Hele big datas grundlæggende metode, hvor man netop fokuserer på overfladedata frem for underlæggende årsager og motiver, lægger op til profilering. Som et eksempel undersøgte tre amerikanske økonomer i 2016 sammenhængen mellem de begrundelser, folk gav, når de søgte om et lån, og hvorvidt de faktisk betalte lånet tilbage. Det viste sig, at der var markant lavere sandsynlighed for at lånerne i undersøgelsen betalte deres lån, hvis de havde brugt ord som "gud", "jeg lover" og "hospital" i deres ansøgning. Det er klart, at viden af den slags er interessant for banker, men hvis din bank afviser din låneansøgning udelukkende ud fra en statistisk analyse at dit sprog, er du blevet profileret. Og det kan være etisk problematisk, at man som individ bliver bedømt ud fra en adfærd, som en gruppe, man mere eller mindre tilfældigt tilhører, har.

Algoritmisk bias er et nært forbundet fænomen. Som vi så ovenfor havde Flu Trends svært ved at skelne influenza- og vintersæsonen, fordi de to fænomener forekom samtidigt i det oprindelige træningssæt. Det var et ret uskyldigt eksempel, men hvad nu hvis de to fænomener, der var blevet sammenblandet, var noget mere følsomt som etnicitet og kriminalitet? I navnlig det amerikanske retssystem støtter man sig i stigende grad til statistisk trænede algoritmer, når man skal afgøre, hvorvidt fanger skal prøveløslades. Fangerne skal udfylde et spørgeskema, og på baggrund af, hvordan det er gået fanger, der tidligere har udfyldt samme skema, kan man udregne en score for, hvor sandsynligt det er, at en person vil begå ny kriminalitet. Metoden er imidlertid blevet beskyldt for konsekvent at give sorte en højre kriminalitetsscore end folk af andre hudfarver. Man må ikke dømme folk alene ud fra deres hudfarve – det er racisme – men noget tyder altså på, at algoritmen i dette tilfælde til dels har lært at gætte, om folk er sorte eller ej, præcis som Flu Trends til dels havde lært at gætte på, om vinteren nærmede sig.

Vi risikerer naturligvis altid at blive mødt af fordomme, men når fordommene bygges ind i modeller, der fremstår som neutrale og objektive, kan de være sværere at adressere. Når man bruger big data er der derfor ikke bare gode metodologiske grunde til at være opmærksom på sammenblandingen mellem overfladisk relaterede fænomener. Der kan også være gode etiske grunde til det. ■

Cases til videnskabsteori

I en serie af artikler vil undervisere i Fagets Videnskabsteori præsentere læserne for videnskabsteoretiske aspekter af alle de naturvidenskabelige gymnasiefag. Vi tager udgangspunkt i cases, som vi på Institut for Naturfagenes Didaktik bruger i vores undervisning i videnskabsteori på bacheloruddannelser ved SCIENCE på Københavns Universitet.