Big data. Det er hamrende svært at forudsige nye opdagelser, finanskriser, imperiers fald og krige, men videnskabsfolkene vil så gerne. Måske kan nye algoritmer gennemskue menneskers adfærd. Måske er det hybris.

Fejl i krystalkuglen

Af GUNVER LYSTBÆK VESTERGÅRD

Valgresultater er relativt simple at forudsige – troede vi. For en række misvisende meningsmålinger, senest op til det amerikanske præsidentvalg, har vist, at metoden har alvorlige skavanker.

I flere år har statistikere derfor ledt efter nye moderne teknikker, der kan afløse de tidskrævende og usikre meningsmålinger baseret på telefonopkald. Én af dem er udviklet af Hernan Makse fra City University of New York. Sammen med kolleger kastede han sig ned i dyndet af big data på det sociale medie Twitter i et forsøg på at forudsige det amerikanske præsidentvalg ud fra folks tilkendegivelser i de små tweets.

Makses fremgangsmåde var et kolossalt nybrud og genialt udtænkt. Men den slog fejl. Hillary Clinton vandt ikke 55,5 procent af stemmerne, som Twitter-analysen havde spået.

Anekdoten fortælles i en artikel af videnskabsjournalisten John Bohannon i tidsskriftet Science, og han slutter med at konkludere, at de gode gamle meningsmålinger faktisk stadig er de mest pålidelige til at forudsige valgresultater – som vi også berettede i sidste uges »Falsificeret«. Bohannons artikel var en del af et nyligt temanummer af Science om videnskabens evne til at forudsige menneskers adfærd. Skal man opsummere temaet, lyder konklusionen, at vi håbede, at vi med mere data og stærkere algoritmer kunne finde forudsigelige mønstre i menneskehedens opførsel – som da astronomerne i Oldtiden gennemskuede himmellegemernes bevægelser og forudsagde sol- og måneformørkelser. Men det er ikke lykkedes endnu.

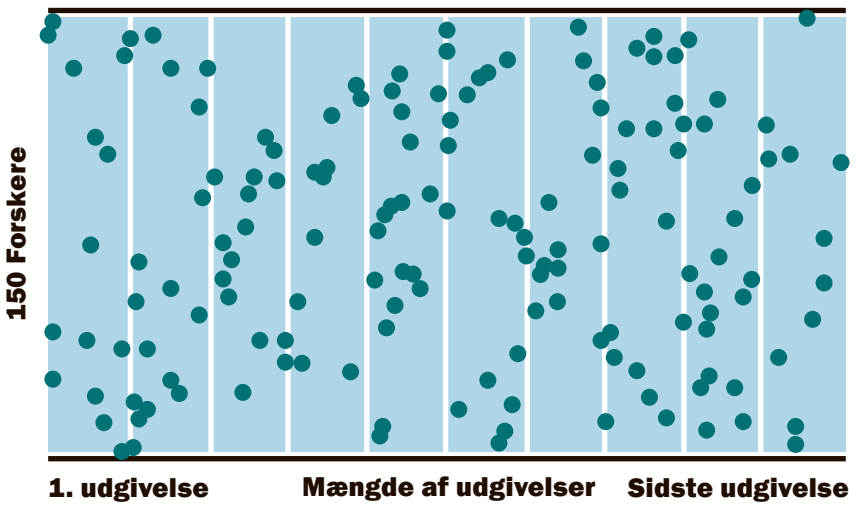
Sociologien har, siden Auguste Comte grundlagde den i 1800-tallet, drømt om at kunne forudsige samfundsudviklingen, og i takt med at den statistiske regnekraft steg i computeralderen, fulgte forventningerne med. Science fiction-forfatteren Isaac Asimov beskrev ligefrem en ny slags videnskab kaldet psykohistorie, der vil kunne forudsige imperiers storhed og fald.

Den vision har man ikke opgivet, og flere konfliktforskere forsøger i dag at varsle krige. Det beskriver statskundskabsprofessorerne Lars-Erik Cederman fra ETH Zürich og Nils B. Weidmann fra tyske Universität Konstanz i en af tema-artiklerne.

»Der er nogle, der håber, at videnskabsfolkene heltemodigt kommer verden til undsætning og redder os fra krige og grusomheder ved at forudsige dem. Det er da også rigtigt, at der for eksempel er større risiko for voldelige konflikter i fattige, udemokratiske lande med en ekskluderet etnisk gruppe end i rige, demokratiske lande med stor lighed. Men når det kommer til at forudsige tidspunktet og stedet for et krigsudbrud, er vi ikke bedre stillet, end når seismologer forsøger at forudsige et jordskælv,« siger Lars-Erik Cederman til Weekendavisen.

Han mener, at der i litteraturen er flere eksempler på hybris, hvor forskere har troet, at kriges og for den sags skyld andre samfundstendensers indre logik kunne brydes ned til fysiske lovmæssigheder. De tager ikke højde for »historiske uheld«, som Cederman kalder anomalier som Sovjetunionens sammenbrud og valget af Donald Trump, der suger spådomskraften ud af eksisterende prognoser.

Sovjetunionens sammenbrud (her Jeltsin på tanken) er eksempel på en historisk begivenhed, der ikke kunne forudsiges. FOTO: REUTERS/SCANPIX



Hvis man læser grafikken vandret, ses højdepunkterne i de enkelte forskeres publikationshistorie sorteret kronologisk fra første til sidste udgivelse. De blå prikker markerer således hver forskers mest citerede publikation. Det spredte mønster viser, at det ikke er muligt at forudsige, hvornår i en forskerkarriere de mest indflydelsesrige publikationer typisk bliver udgivet. KILDE: ROBERTA SINATRA ET AL./SCIENCE

»Selvom begrænsede forudsigelser er mulige, er drømmen om en krystalkugle en utopi,« siger han.

UTOPI eller ej så forskes der på højtryk i at forudsige alt fra genindlæggelser på hospitalerne over kriminalitet og forbrugeradfærd på nettet til økonomiske kriser.

Favoritteknikken er såkaldt *supervised machine learning* forkortet SML. Den går ud på at lære et softwareprogram at forudsige fremtiden ved at finde og genkende mønstre i gigantiske mængder arkiveret data. Lidt som når et barn lærer et sprog ved at lytte til de voksne. Programmet bruger ikke modeller eller teorier til forudsigelserne, så enhver computerkyndig person kan bruge SML. Firmaet Kaggle afholder for eksempel konkurrencer online, hvor verdens statistikere konkurrerer om at kunne forudsige, hvilke patienter der har lungekræft, hvem der vinder en basketball-turnering, og hvilke typer af passagerer der overlevede Titanic.

Økonomiprofessor Susan Athey fra Stanford University i USA er SML-ekspert, og i Science-temaet fortæller hun om teknikkens fremmarch. SML er ret god til at forudsige simpel adfærd som hjemmesidetrafik og dødelighed blandt patientgrupper, men i hendes øjne opstår problemet, når man ud fra forudsigelserne prøver at sige noget om årsag og virkning, for her er maskinen blank.

»Du kan nemt forudsige med SML, hvor mange mennesker der vil besøge en bestemt hjemmeside på en given dag. Ved at se på historiske data kan du finde ud af, hvor mange der køber en bestemt type produkter ti dage før jul, når det sner, og juleaften er en tirsdag,« fortæller Susan Athey over telefonen og fortsætter:

»Men vi vil jo gerne gå videre end forudsigelserne og handle på dem. Det er bare meget sværere. Som patient vil man gerne vide, om man skal tage en bestemt slags medicin, der dog har mange bivirkninger, for at undgå at dø. Forudsigelsen siger, at mange, der tager medicinen, dør, men den siger ikke noget om, hvorvidt det skyldes bivirkningerne. Det kan jo i stedet være, fordi mange af patienterne er svækkede i forvejen. Sat på spidsen vil maskinen fortælle dig, at du slet ikke skal tage på hospitalet, for der dør rigtig mange mennesker.«

Et andet eksempel, Athey bruger, er sammenhængen mellem hotelpriser online og antal gæster. Jo højere priser, jo flere gæster, lyder forudsigelsen, men et hotel kan ikke hæve prisen og forvente flere gæster. Der er altså en korrelation mellem pris og belægning, men ikke en kausalitet fra pris til antal gæster.

Den slags »naive fejlslutninger« er Athey stødt på mange af. Hun fortæller dog også, at den nyeste forskning inden for SML netop går mod at kunne udlede kausaliteter, og firmaer som Microsoft, Facebook og Google har opbygget interdisciplinære forskergrupper til formålet.

FLERE af temaartiklerne nævner problemet med komplet uforudsete hændelser – som de »historiske uheld« omtalt tidligere. De findes inden for alle felter og kaldes også for »sorte svaner«. Det er den uanmeldte og ukendte onkel fra Amerika, der pludselig dukker op til brylluppet, og hvordan forudser man det?

Et af de steder, hvor den slags begivenheder optræder hyppigst, er netop inden for videnskab. Tidsskrifter, universiteter og fonde prøver ihærdigt at udregne, hvor det næste store gennembrud sker ved hjælp af millioner af databidder om fortidens opdagelser, men indtil videre kan de lige så godt slå med en terning.

»Lige nu gennemsøger vi store datasæt for at forstå, hvor forudsigelser i det hele taget er mulige, og hvor grænsen til det uforudsigelige er. Citationer er for eksempel velstuderede, men lige nu kan vi ikke bruge dem til at sige noget om fremtiden,« siger Daniel B. Larremore fra Santa Fe Institute i USA. Han og to medforfattere har i Science-temanummeret skrevet om videnskabens evne til at forudsige sig selv.

Nye undersøgelser af forudsigelseskraften i videnskaben har blandt andet vist, at den gamle påstand om, at gennembrud kommer fra unge forskere, er forkert. En gennemgang af 10.000 forskeres publikationer har vist, at det er umuligt at forudsige, om den første eller den sidste publikation fra en forsker bliver den mest citerede.

Der er også nye studier af såkaldte »torneroser«, der har vist, at »nogle artikler har et atypisk citationsmønster, hvor de efter udgivelsen kun citeres få gange, men pludselig nogle år senere citeres helt enormt. Det er nemt at genkende dem, efter at de er blevet berømte, men ikke når de bliver udgivet,« forklarer Larremore.

Det største problem for videnskaben selv er dog, at mens vi gerne vil forudsige krige og finanskriser for at *undgå* dem, tilbeder vi nye videnskabelige gennembrud som penicillinen. Derfor må evnen til at forudsige gennembrud ikke forhindre dem, og det kan blive vanskeligt, for forudsigelserne vil ikke tage højde for torneroser og sorte svaner, og hvis fondenes bevillinger følger prognoserne, bliver videnskaben for pæn og stillestående.

»Forskere vil kun udforske ideer med lav risiko og ikke forfølge de dristige, for fondene leverer maden i videnskabens økosystem, og forskerne går derhen, hvor maden er,« siger Larremore.

I artiklen opstiller forfatterne en skala med opdagelser, der går fra »uventet« til »forventet«. Mest uventet var opdagelser som penicillinen og den kosmiske mikrobølgebaggrundsstråling, mens opdagelsen af Higgs-bosonen og tyngdebølger hører til nogle af de mest forventede opdagelser. Pointen er, at uanset genstandsfeltet er visse typer adfærd forudsigelige, mens andre typer er gemt væk i en sociologisk sort boks, som alverdens supercomputere stadig ikke kan lirke op.

Én forudsigelse, som undertegnede godt tør komme med, er dog, at forskningen i forudsigelser vil fortsætte.

LAYOUT: BENTE BRUUN KORREKTUR: FLEMMING GERTZ