Kapitel 6:
At få computeren til at hjælpe os: Del II — datadrevne modeller

Henrik Kragh Sørensen Mikkel Willum Johansen

20. maj 2023

Indhold

Der vil givetvis være en række slåfejl, uklarheder og (forhåbentlig få) fejl. Informationer om slåfejl og forbedringsforslag modtages meget gerne.

6.1 Modeller med meget store mængder data

En anden type modeller, som særligt lægger op til brug af computere og datalogi, handler om at modellere fænomener ud fra meget store mængder data. Hvor der indtil midten af 1900-tallet var mangel på data, står vi i dag overfor en situation, hvor der er alt for meget data, til at vi kan behandle hvert stykke data udelukkende på dets egne præmisser. Derfor har vi brug for at *aggregere* data og behandle data med statistiske redskaber. Nogle sådanne statistiske metoder er klassiske i den forstand, at både metoderne og deres begrænsninger velkendte. Mange af disse begrænsninger handler dybest set om at have det rette kendskab til metoderne, så man kan anvende den bedst egnede metode i en given sammenhæng. Men der er også nogle mere principielle, videnskabsteoretiske vinkler, hvoraf vi vil behandle nogle udvalgte i det følgende.

6.2 Proxies: Hvad vil vi — og hvad kan vi — måle?

To af de allerførste overvejelser, man typisk gør sig, når man skal bygge en model, er: 'hvad skal modellen måle?' og 'hvordan vi måle det?'. Det er nemlig langt fra alle interessante oplysninger, vi kan måle direkte. I stedet anvender vi andre størrelser, som vi kalder for *proxies*, som det er muligt og evt. nemmere at måle, og som vi med mere eller mindre begrundelse anser for at være forbundet med det, vi i virkeligheden gerne ville have målt.

For eksempel kan det være, at vi gerne vil modellere klimaets udvikling igennem de sidste 10,000 år (dvs. siden starten af den nuværende mellemistid), og dertil har vi brug for at vide noget om atmosfærens temperatur. Men det er lettere sagt end gjort. For det første er der jo nogle uklarheder i spørgsmålet: Skal det være temperaturen hvert sekund de over disse mange år, temperaturen på denne dato igennem 10,000 år, eller fx middeltemperaturen over et års forløb.[1] Det sidste er nok både det mest realistiske og også teoretisk velbegrundet: Vi ved, at temperaturen varierer over årets gang, så det giver mening at aggregere på det niveau.

Men selv efter, at vi har afgrænset denne art af pragmatiske, men teoretiske valg, er vi endnu ikke i mål. Vi kan jo ikke rejse tilbage med et termometer, så selve temperaturmålingerne er nødt til at foregå *via proxy*. Dermed menes, at vi observerer en anden størrelse end den, vi faktisk er interesseret i. I eksemplet med tusinde år gamle temperaturer måler man fx forholdet mellem koncentrationer af to forskellige stabile iltisotoper i de dybe isboringer, man har foretaget i Grønland (Dahl-Jensen, 2009). For at koble iltisotoper og middeltemperaturer sammen kræver det ret omfattende teoretisk og empirisk begrundelser. Men det er nærmere et vilkår for empirisk og modelbaseret videnskab.

Nogle af disse oversættelser mellem *teoretiske* data og de *observerbare* data, som vi sætter i stedet, er både gamle, anerkendte og velbegrundede. Optimalt ville vi jo gerne vide, at vores proxy data er stærkt *korreleret* med vores teoretiske data; så ville observation af proxy jo tilsyneladende blot være en *indirekte* observation af teoretisk data. Men ofte kan vi ikke uden videre være sikre på, at de to størrelser faktisk er stærkt korrelerede, og selvom de måtte være det, er der indbyggede farer ved at lægge for meget vægt på korrelationer i store datamængder (se 6.6 nedenfor om korrelation).

Nogle *proxies* har en anden fordel ud over, at de er forbundne med det, vi ønsker at måle: De er blevet *kanoniske* i den forstand, at de har vist sig så bredt anvendelige og rimeligt målbare, at de er tæt på at blive opfattet som selvindlysende og uproblematiske. Faktisk kan de være medbestemmende for hele teorier og centrale elementer i *disciplinære matricer* for visse felter, fx inden for medicin. Når læger ønsker at *diagnosticere* sygdomme (fx udbredte folkesygdomme), er deres første *proxies* fx *BMI* (body mass index), blodtryk og oplysninger om arvelige sygdomme. De to første af disse *proxies* indgår i en kompleks vekselvirkning mellem lægelige ønsker og teknologiske muligheder: Jo mere *BMI* er blevet brugt diagnostisk, jo mere relevant er det, at kunne måle BMI og vurdere de resulterende data. Der er altså ikke blot tale om, at fx *BMI* er en kraftigt forsimplet *proxy*, idet den ser bort fra mange dimensioner i data, som vi har grund til at antager, har betydning. Men udviklingen af denne slags proxies er altså også resultat af en social og teknologisk udvikling, der *oversætter* mellem de teoretiske data og deres *proxies* (se nedenfor om modellers *performativitet*). Andre *proxies*, som her fx arvelige sygdomme, er ikke til at *måle* fra det enkelte individ, så i stedet forsøger læger og andre at opbygge databaser med genetisk

[1] Tilsvarende spørgsmål gælder naturligvis også den rumlige dimension: Taler vi om Danmark, om Grønland, om den nordlige halvkugle, eller en helt anden afgrænsning?

information eller skaffe elektronisk adgang til patientjournaler, således at også denne slags information kan indsamles, gemmes og bruges *diagnostisk*.

Som det præsenteres her, er proxies på flere måder sammenlignelige med de idealiseringer, der foregår i modelleringsprocessen: Vi sætter noget andet i stedet for det, vi egentlig er interesserede i, for overhovedet at muliggøre processen. I yderste tilfælde er en proxy en kontrafaktisk idealisering. Og ligesom andre idealiseringer, er valg og udvikling af proxies ikke værdi-neutrale. Faktorer som sædvane, tilgængelighed, og effektivitet er *pragmatiske* hensyn og værdier, som vi med rette kan anlægge. Deres effekt vil hyppigt have betydning for, *hvorvidt* modellen virker og er altså af epistemisk karakter. Men der kan også indgå mere udtalte værdier, hvis vi fx vælger proxies, der ser bort fra individers særinteresser (fx ved aggregering). I så fald kan valget af proxies siges at have indflydelse på, *hvordan og hvornår* modellen virker, og de kan derfor siges desforuden at have en etisk karakter. Som vi skal se nedenfor kan proxies endda have en *performativ* karakter, idet de kan påvirke den virkelighed, som de egentlig er tænkt som idealiseringer af.

6.3 Data er altid beskidt: Indsamling og kuratering af data

Som forklaret i kapitel 6a kan man inddele modeller i teori- og datadrevne modeller. Hvor teoridrevne modeller (herunder særligt matematiske modeller) er beskrevet i kapitel 6a, opstår der nogle særlige udfordringer for data-drevne modeller — og den mest fundamentale udfordring er netop *data*. For hvor kommer data fra, hvordan er det indsamlet og behandlet, og hvor repræsentativt er det? Det er problemer, der altid kan stilles til data, men når datamængderne bliver så store, at det ikke er menneskeligt muligt at undersøge kvaliteten af data *i hånden*, bliver problemet kun endnu større og mere akut.

Det kan være fristende at anlægge det synspunkt, at de data, som vi indsamler til vores modeller, bør være så *rå* som muligt for at give størst mulig *objektivitet*. Hvad der præcist menes med disse begreber kan være svært at forklare, men man kan måske forestille sig, at *rå data* er karakteriseret ved at være uafhængige af subjektive, menneskelige faktorer — *rene faktuelle oplysninger* om verden.

Men som vi allerede stiftede bekendtskab med i kritikken af logisk positivisme (se kapitel 1), er der principelle indvendinger imod *fordomsfrie* (teorifrie) observationer: Hvis vi bare skal *observere*, hvad skal vi så observere, hvordan skal vi gøre det, og hvilke delelementer er vigtige end andre.

I stedet for at opfatte og tilgå data som *rå*, giver det et mere passende nuanceret billede, hvis vi går til data som en indsamlingsprocess, hvor menneskelig interaktion er uundgåelig i form af data-design og (især) *kuratering*.

Ved begrebet *kuratering* forstår vi her den (oftest hermeneutiske) proces, hvorved genstandsfeltet afgrænses, data bliver målt (igennem proxies), data bliver annoteret ved *metadata*, data bliver filtreret (ikke kun for outliers, men ofte også med henblik på *stratificering* og andre systematiske forskelle i indsamlingsprocessen), data bliver holdt opdateret (hvad det end betyder i sammenhængen). Det er her vigtigt at pointere, at under opfattelsen af data som en *kurateringsproces*, er ingen af disse skridt i sig selv problematiske, så længe der opretholdes en form for *transparens* om dataprocessen (se nedenfor), og modellen vurderes under hensyntagen til denne *kuratering*.

6.4 Teori og model er underbestemt af data

Selv når vi har anerkendt, at vi ikke måler det, vi direkte ønsker men i stedet oftest (altid) måler via *proxies*, og at vi ikke måler på rå data alligevel, så udestår stadig udfordringer i form af implicitte eller (forhåbentlig) eksplicitte valg (Johansen og Sørensen, 2018): Hvilken model er egnet til at behandle hvilke data — og hvad stiller modelvalget af krav til data? Igen er disse udfordringer ikke udelukkende relevante for datadrevne modeller, men de antager en særlig form, når datamængderne er så store, at vi ikke kan være sikre på at kunne overskue dem på nogen direkte vis.

Alle disse valgprocesser er at sammenligne med de *dialektiske*, *pragmatiske* og kreative skridt i den venstre side af modelleringsprocessen illustreret i figur **??**. Den venstre side i modelleringsprocessen for teoridrevne modeller handler om *konceptualisering* og *idealisering* af et udsnit af virkeligheden til en teori, hvis kvalitet (skal) vurderes ud fra værdiladede kriterier som *forudsigelsespræcision*, *forklaringskraft* eller endnu mere subjektive værdier som *simplicitet* og *skønhed*.

Logisk positivisme 2.0

I begyndelsen af det 21. århundredes første årti, mens *big data* stadig var en helt ny og potentielt revolutionerende mulighed for videnskabelig forskning, blev der fremsat adskillige optimistiske, i dag at se som ret utopiske, bud på, hvordan den videnskabelige erkendelsesproces stod foran en stor transformation. En af de mest udtalte fortalere for den nye vision var Chris Anderson, som i en kort artikel i magasinet *Wired* annoncerede „The End of Theory“, hvori han sparkede liv i en ny form for logisk positivisme (Anderson, 2008). Han hævdede eksplicit heri, at ikke blot ville der ikke længere være behov for teorier i videnskaben, men også at hele 'den videnskabelige metode' var blevet overflødigt.

Anderson var ikke den første, til at annoncere den eksisterende, klassiske videnskabelige metode for død. Allerede tiår forinden havde Stephen Wolfram foreslået noget, der på visse punkter lignede. Men hans pointe var alligevel en anden: I stedet for empirisk verifikation foreslog han en *beregningsbaseret* videnskab, hvor computerberegninger kunne erstatte mange empiriske forsøg (Wolfram, 2002). Men Wolfram s argument forudsatte, at videnskabens teorier allerede var formulerede: Den var ikke en *heuristisk* men en *verifikationistisk* metode.

I forhold til Wolfram gik Anderson altså hele linen ud: De nye enorme datamængder kunne udgøre det for positivisternes *objektive og fordomsfrie observationer*, og anvendelsen af statistiske metoder i stor skala kunne udtømme mulighederne for teorier og dermed nå frem til *lovmæssigheder* ved en *induktiv* proces. Dermed var cyklussen komplet: Den logiske positivisme var blevet etableret igen — i en *in silico* version. Men som vi skal se nedenfor i afsnittet om korrelationer, møder denne nye form for automatiseret teoridannelse nogle uoverstigelige statistiske barrierer.

6.5 Machine-learning-modeller

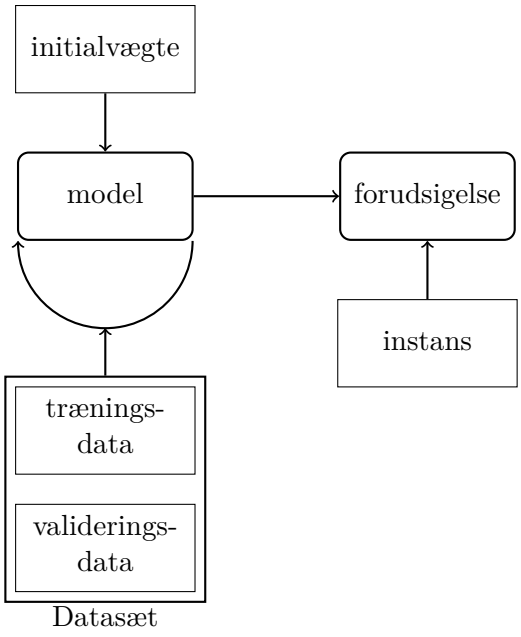
I det følgende vil vi beskrive machine-learning-modeller som *algoritmer*, selvom det er en diskussion i sig selv, hvorvidt og hvordan de opfylder sædvanlige definitioner af algoritmer, som fx den givet af Robin K. Hill og beskrevet i kapitel 3.

Machine-learning-modeller kommer i forskellige *familier* og til forskellige formål. Formålene kan være *klassifikation*, *rangordning*, *oversættelse* eller *genkendelse* og den præcise

anvendelse kan finde sted inden for forskellige områder, fx behandling af *naturligt sprog* (*NLP*, *natural language processing*) eller billedanalyse (*CV*, *computer vision*). Begge disse formål er for øvrigt gamle klassikere inden for forskning i og finansiering af kunstig intelligens: NLP i forsøget på at automatisere oversættelse, konkret oversættelser fra russisk til militære formål og behandling af visuel information i (autonome) robotter, igen med oplagt militær bevågenhed.

Hvis vi skal indkredse typerne eller familierne af *ML-modeller*, så er der forkellige niveauer, man kan anlægge til at skelne. Machine-*learning* handler jo, som navnet siger, om, at algoritmen *lærer* og derved udvikler sig. Man kan forsøge at anlægge dette perspektiv til en beskrivelse af forskellige typer læring: *ikke-superviseret læring* er former for statistiske analyser, der forsøger at uddrage information fra data *uden* et niveau af menneskeligt verificeret basisviden. For eksempel kan man lave *clustering algoritmer*, der inddeler et datasæt bestående af n-dimensionale vektorer i grupper, hvis elementer ligger tæt på hinanden og langt fra de andre grupper. Vi siger, at *datasættets størrelse* er *antallet* af vektorer i datasættet, mens vi bruger betegnelsen *datasættets dimension* til at angive *længden* af hver vektor. Hvad den underliggende mening af disse vektorer måtte være er ikke vigtigt for klikedannelsen, så derfor er der — efter at datasættet er dannet — ikke yderligere brug for menneskelig overvågning af algoritmens læring. Men for at kunne anvende denne slags algoritmer i praksis kan man være nødt til at lægge forskellige betingelser ned over deres proces for at de kan bruges i anvendelser og fortolkning. For eksempel er antallet af kliker en størrelse, som er vigtig for cluster-algoritmer, men hvor det er umuligt at give en *meningsfuld* værdi uden at der er mennesker involveret på dette trin i processen.

Der findes også en meget stor og vigtig klasse af algoritmer baseret på menneskelig involvering i træningsprocessen, mere specifikt at algoritmen er trænet ud fra data, der er beriget med en form for mening. Denne familie af algoritmer betegnes *superviseret læring*. Pointen med denne type modeller er, at de ud fra de tidligere *labels* givet til data i træningssættet kan forudsige labels for data, som modellen ikke hidtil har mødt. En skematisk illustration af, hvordan en ML-model af denne type kan trænes og bruges til forudsigelser er givet i Figur 1. Disse modeller består af et meget stort antal *vægte* (tal mellem 0 og 1), som *trænes* ud fra data i træningsdata, således at modellens forudsigelser af labels for data fra valideringsdata bliver bedst muligt. Denne forudsigelse består i at beregne virkningen af vægtene på en given instans, og den resulterer typisk i et tal mellem 0 og 1, som ofte kan fortolkes som en form for sandsynlighed. Træningsprocessen kan eventuelt tage udgangspunkt i et sæt af *fortrænede vægte*, som er beregnet på forhånd til en bestemt anvendelse, for eksempel billedklassifikation. Modellerne kan også trænes videre ved at tilføje nye data til trænings- og valideringssæt og dermed øge *størrelsen af datasættet*.



Figur 1: Skematisk fremstilling af træning og fortolkning ved superviserede ML-modeller (se også en yderligere udfoldet version i Figur 7).

Eksempel: En binær classifier

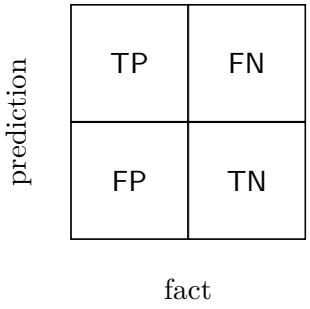
For at illustrere nogle epistemologiske aspekter ved superviserede ML-modeller, kan vi tage udgangspunkt i en af de simplete typer modeller overhovedet, nemlig *binær classifier*. Dette er en type *superviseret* model, som skal klassificere data i *to* disjunkte klasser A og B — det kan fx være, at data er træfrugter, og modellens opgave er at adskille æbler (A) og pærer (B). Ingen frugt kan være både et æble og en pære, så klasserne er disjunkte (ikke-overlappende, $A \cap B = \emptyset$), og alle data, som meningsfuldt kan gives til modellen er elementer i $A \cup B$.

Hvis denne klassifier er baseret på CV modeller, vil den altså tage udgangspunkt i et billede af en frugt og forsøge at placere billedet i den klasse, der passer bedst. Som træningsdata vil modellen have en række billeder af frugter, som alle er annoteret med en label, der angiver, om billedet viser et æble eller en pære. Træningen består i, at justere modellens vægte således, at modellens forudsigelser på data fra valideringssættet er mest muligt korrekte forstået som at den label, som modellen forudsiger stemmer overens med den label, som billedet er annoteret med på forhånd.

Der er nu fire tilfælde at betragte (se Figur 2), nemlig følgende, hvis vi betragte æbler som positive og pærer som negative (tænk på det, som at modellen skal fjerne de uønskede pærer fra et samlebånd med æbler og pærer).

modellens forudsigelse	annoteret label	tilfældets navn
æble	æble	Korrekt positiv (TP, true positive)
æble	pære	Forkert positiv (FP, false positive)
pære	æble	Forkert negativ (FN, false negative)
pære	pære	Korrekt negativ (FN, false negative)

Figur 2: Mulige resultater af en binær classifier. Man kalder sommetider falske positiver for *Type-I fejl* og falske negativer for *Type-II fejl*.



Figur 3: Udkommet af en binær opdeling kan illustreres som sande og falske positiver og negativer i form af en *confusion matrix*.

Korrekte positive og korrekte negative forudsigelser dækker altså over, at modellens forudsigelse stemmer overens med den angivne label. Hvis modellen fx skal forudsige om en patient har en dødelig sygdom, som kan behandles, er både sande positiver (modellen siger korrekt at patienten har sygdommen, og vi kan behandle den) eller falske positiver (modellen siger korrekt, at patienten er rask, og vi kan lade være med at bekymre patienten yderligere) selvsagt langt at foretrække. Men modellen kan også tage fejl på to forskellige måder: Den kan klassificere et æble som en pære eller omvendt. Disse to tilfælde måske fremstå som symmetriske, men er det sjældent i anvendelser. Hvis modellen fra før for eksempel forkert forudsiger, at patienten er rask kommer vi til at undlade at behandle patienten, som så dør af sygdommen. Omvendt hvis modellen forkert forudsiger, at patienten har sygdommen, så går vi i gang med at behandle patienten, men finder ud af, at der ikke er nogen sygdom at behandle. På den måde fører den første type fejl (FN) til patientens død, hvorimod den anden type fejl (FP) kun fører til en del bekymring og en unødig operation. Selv hvis vi optimerer træningen af modellen til at minimere risikoen for falske negativer, vil beslutninger på grundlag af machine-learning-modeller *altid* foregå på et *ufuldstændigt grundlag*, hvilket vil blive behandlet nedenfor i forbindelse med *modellers forudsigelser* og *ekspertroller*.

Når vi skal udtale os om modellens kvalitet, gør vi det ofte ved at anvende en metrik, der måler, hvor godt modellens forudsigelser på valideringsdata svarer til de kendte labels. Der er flere forskellige metrikker, som man kan anvende, og de indfanger alle forskellige betydninger af, hvad der vægtes som modellens kvalitet (se Figur 4).

$$\mathsf{Acc} = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{TP} + \mathsf{FP} + \mathsf{TN} + \mathsf{FN}} \qquad \text{(Accuracy)}$$

$$\mathsf{Pre} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}} \qquad \text{(Precision)}$$

$$\mathsf{Spec} = \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}} \qquad \text{(Specificity)}$$

$$\mathsf{Rec} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}} \qquad \text{(Recall/sensitiviy)}$$

$$\mathsf{F1} = 2 \times \frac{\mathsf{Pre} \times \mathsf{Rec}}{\mathsf{Pre} + \mathsf{Rec}} \qquad (1)$$

$$= \frac{\mathsf{TP}}{\mathsf{TP} + \frac{1}{2}(\mathsf{FP} + \mathsf{FN})} \qquad \text{(F1)}$$

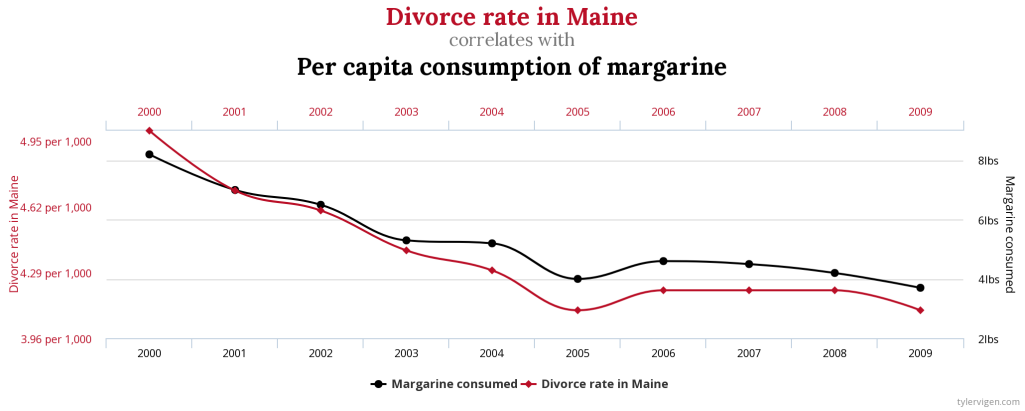
Figur 4: Forskellige metrikker som kan finde anvendelse for at måle kvaliteten af en binær classifier.

6.6 Skilsmisser og margarine: Korrelationer og årsager

En vigtig viden om ML-modeller er, at de virker ved *korrelationer*. To hændelser siges uformelt at være korrelerede, hvis de *følges ad*, således at når den ene af dem for eksempel vokser, så gør den anden det også. Nogle gange er der faktisk tale om en *kausal sammenhæng*, idet man kan finde ud af, at den ene hændelse er en direkte *årsag* til den anden, som har form af årsagens *virkning*. Effekten er her direkte, hvis den har form af, at årsagen *trigger* virkningen uden variation. Man kan måske tale om, at relationen nærmest er *deterministisk*: virkningen er forudbestemt af årsagen, hvorfor årsagen tidsligt må ligge forud for virkningen.

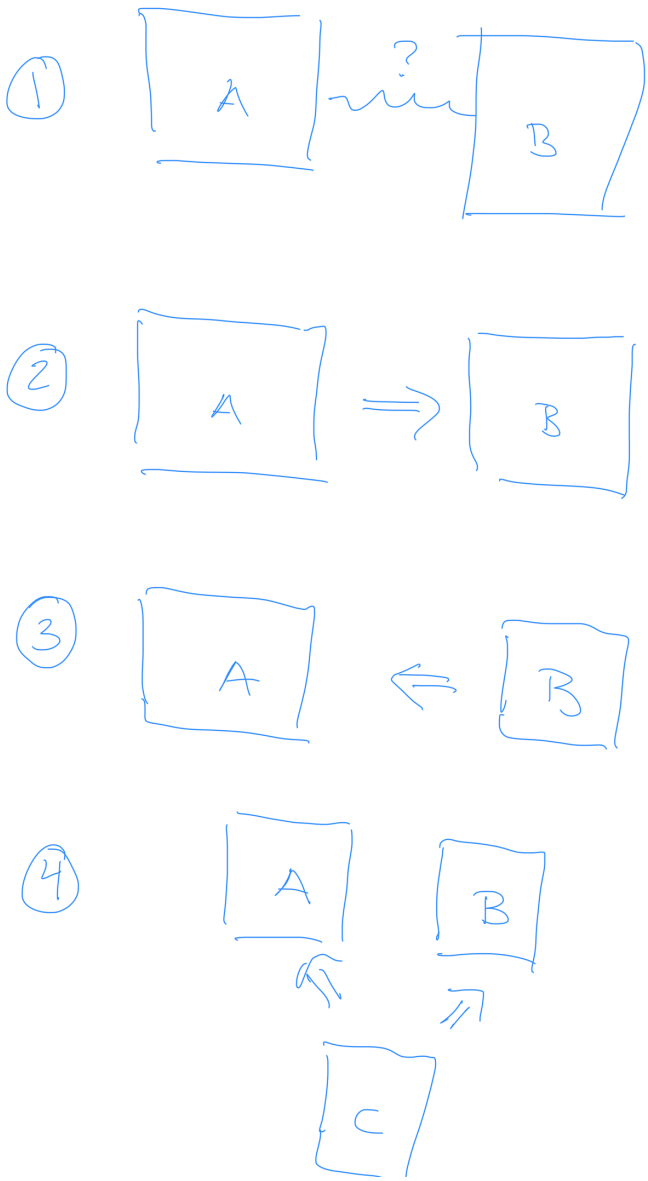
Men når ML-modeller foretager forudsigelser, så er der netop tale om korrelationer og ikke årsag-virkning-sammenhænge: Modellen er trænet på datasæt, hvor en label er tilknyttet hvert datapunkt, så visse variationer i datapunktet vil medføre en forandring i den tilhørende label. Men der er ikke tale om en kausal virkning, idet vi (typisk) ikke kan på en række direkte virkninger, men i stedet ser sammenhængende opførsel.

For at illustrere forskellen mellem korrelation og kausalitet, kan man betragte nogle af de såkaldte *spurious correlations* (se Figur 5).



Figur 5: En korrelation, der næppe er en kausal sammenhæng — en *tilfældig korrelation* (*spurious correlation*), se `https://www.tylervigen.com/spurious-correlations`. Se også Ringgaard (2017) for yderligere forklaring og danske eksempler på tilfældige korrelationer.

Selvom to variable fx følges ad over en periode, behøver der langtfra at bestå en kausal sammenhæng imellem dem. Der kan fx være tale om, at begge variable er relaterede til en tredje variabel (*fælles faktor* eller *confounding factor*) eller at det tilsyneladende parallelle forløb er en ren tilfældighed (se Figur 6).



Figur 6: Figuren viser forskellige måder, hvorpå to størrelser kan være korrelerede: 1. Tilfældig korrelation, 2. A er årsag til B, 3. B er årsag til A, 4. A og B har en fælles årsag. Der er mange flere muligheder for relationerne, men disse fire alene viser, at man ikke fra korrelation mellem A og B kan udlede en kausal relation mellem A og B. Punkterne 2 og 3 alene viser, at eftersom kausalitet har en retning, og korrelationer ikke har, kan vi ikke engang afgøre, *hvad* der evt. er en årsag, og hvad der er en virkning.

Sorte kasser for enhver

Vi bruger betegnelsen *black box* til at beskrive dele af et system, som *for nogle* relevante aktører forbliver lukket, uigennemskueligt og typisk af en størrelse, der overstiger, hvad et individ kan overskue. Denne definition spejler beskrivelser af, hvad der sker, når den individuelle autonome erkendelse bliver umuliggjort af forskellige slags kompleksitet (se fx Sørensen og Andersen, 2018). Man kan altså skelne imellem forskellige gruppers behov for indsigt (dannelse) i *machine-learning-modellers* virkemåde og epistemiske status, og vi kan i hvert fald identificere tre forskellige grupper: forbrugere, anvendere og skabere af *machine-learning-modeller*.

Fra *forbrugerens perspektiv* vil det for fx forslagsalgoritmer fra Netflix eller Facebook være relevant at vide, at anbefalingerne er baseret på, at det enkelte individ placeres i en *referencegruppe* med andre individer i den nærmeste klike baseret på modellens forudsigelser. Man får altså de samme anbefalinger som *andre ligesom en selv* også får, men hvor den nærmere beskrivelse af referencegruppen ikke er til at give, da den typisk er baseret på meget store og højdimensionale datasæt, som typisk forbliver en *black box* for forbrugeren. Resten af modellens opbygning, træning og tekniske funktion forbliver *black-boxet* for forbrugeren.

Anlægger man derimod et perspektiv, som tager udgangspunkt i dem, der skal *anvende machine-learning-modeller* fx i beslutningsprocesser eller til rådgivning, hører der yderligere elementer til en dannet forståelse. Foruden at vide, at forudsigelser er baseret på *referencegrupper*, vil anvenderen også skulle vide, at disse referencegrupper er baseret på fortidig data, og at *machine-learning-modeller* derfor har sit eget *induktionsproblem*: Den kan, ligesom en papegøje, kun gentage det, den allerede er trænet på. For anvenderen har dette særlig betydning, når biases i datasættet, som er baseret på fortidige forhold, propageres ind i fremtiden. Men hvordan datasæt er *kurateret* vil ofte stadig forblive en *black box* for anvenderen, ligesom selve modellens interne funktion også vil være det.

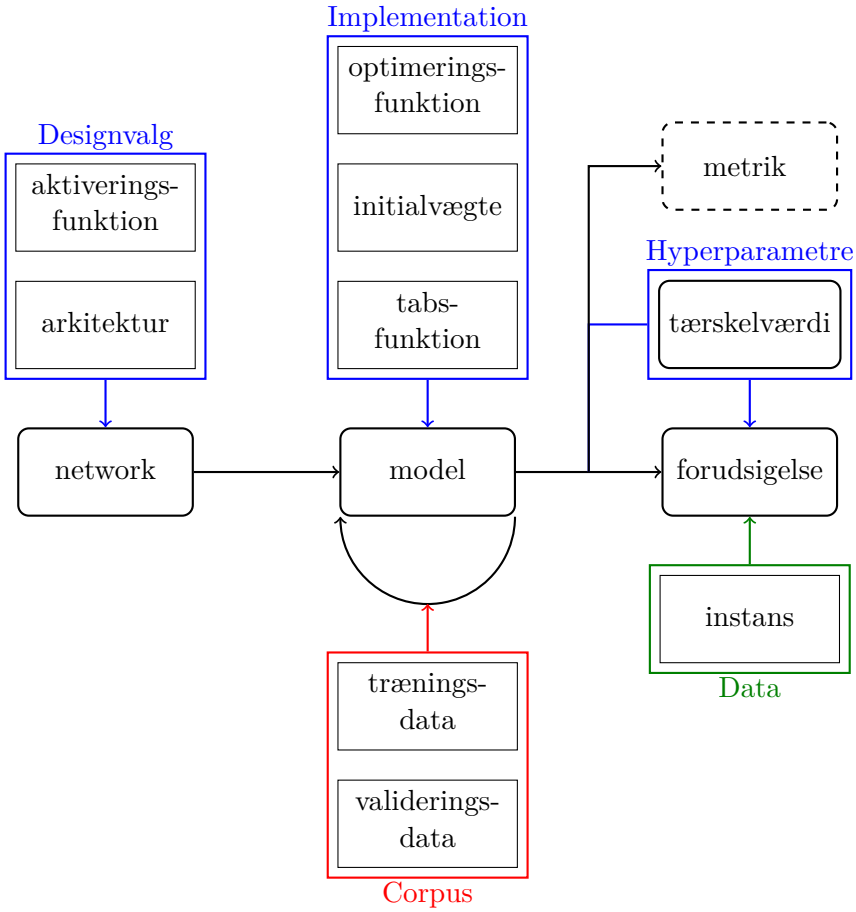
Endelig, og måske mest relevant i denne kontekst, skal vi behandle spørgsmålet om nødvendig indsigt ud fra perspektivet, som tilhører en, der *skaber machine-learning-modeller*. Den overordnede skematiske beskrivelse af en superviseret *machine-learning-model* i Figur 1 dækker over en række elementer, der er af central betydning for udfordringerne ved at forklare *machine-learning-modeller* ud fra et *skaber-perspektiv*. I Figur 7 er denne figur udvidet med yderligere grupper af *designvalg*, *implemtationsvalg* og *hyperparametre*, alle indrammet i blåt. Figurens element om *datasæt* er indrammet i grønt, hvilket angiver, at indsigt i dette område fra et *skaber-perspektiv* vil blive taget op igen i forbindelse med diskussioner om etisk og professionelt ansvar.

Når man bygger og udvikler *machine-learning-modeller*, er der først en række *designvalg*, der skal træffes. Det drejer sig blandt andet om valg af det underliggende netværks arkitektur og antallet af neuroner, og det drejer sig om at fastsætte, hvornår neuronerne aktiveres. Disse valg er *pragmatiske* i den forstand, at de ikke behøver at være *absolut optimale*, men blot skal være *gode nok* til, at modellen kan udføre den funktion, den er bygget til, på en tilfredsstillende vis. Men i praksis er de også hyppigt formet af hensyn til *sædvane* og tidligere *erfaring*. På den måde ligner disse valg de valg, som *ingeniører* træffer, når de skal udvikle nye teknologier, idet dog en omfattende inddragelse af *sædvane* og *tavs viden* vil placere konstruktionen af *machine-learning-modeller* som skridtet *før* egentlig *ingeniørvidenskab*.

Noget tilsvarende gør sig delvist gældende for de valg, som træffes under selve implementationen. Men implementationsvalgene er typisk noget, man lettere kan eksperimentere med at justere for bedre performance. På den måde synes disse valg mere at ligne empirisk

udledte dimensioneringer, som foretages af *ingeniører*. En sidste række valg handler om modellens *hyperparametre*, som er parametre, der blandt andet bestemmer træningsprocessen og de tærskler, som forudsigelserne er underlagt. Disse parametre er (eller er ved at blive det) empiriske, idet man enten kan træne modellen med mange forskellige parametre, ligesom en *scenarie-beregning*, eller man kan forsøge at få en slags meta-modeller til at forudsige passende *hyperparametre* for en given model.

Alle disse valg er med andre ord nogle, som skaberen af modellen skal forholde sig mere eller mindre eksplicit til. Når der er tale om pragmatiske valg kan den nærmere betydning ofte være uden for udviklerens rækkevidde, men ganske tit vil de kunne opsøges, og disse elementer er altså ikke *black-boxet* i samme grad, som tilfældet var fra de andre gruppers perspektiver. Til gengæld vil valg, der er truffet på grundlag af *sædvane* og *tavs viden* være epistemiske *sorte kasser*, idet den enkelte udvikler her deler fælles men tavs viden.



Figur 7: Skematisk fremstilling af *machine-learning-model* fra et skaber-perspektiv.

6.7 FFF: forklaring, fortolkning og forsvar af modeller

Det står i den internationale *GDPR-forordning*, at enhver der er blevet udsat for en beslutning truffet helt eller delvist på grundlag af en automatiseret procedure (en algoritme) har krav på en *forklaring*. Dette kan synes at være et rimeligt demokratisk krav, men det viser sig i både teori og praksis at være virkelig svært at realisere.

For at komme nærmere på problemerne med at forklare *machine-learning-modeller*, vil vi tage udgangspunkt i begrebet *forståelse*, som i filosofisk forstand er meget nært knyttet til begrebet *forklaring*. Men ved at fokusere på individers eller gruppers *forståelse* bliver det klarere, hvilke roller *sorte kasser* spiller, idet de netop er forhindringer for *forståelsen*.

Hvad er en forklaring overhovedet?

Som vi allerede er stødt på i kapitel 6a, er et af de mulige formål med modeller, at vi ønsker at bruge dem til at *forklare* egenskaber ved det genstandsfelt, som modellen er bygget over. Men hvad skal vi forstå ved en *forklaring*? Der findes mange forskellige filosofiske tilgange til dette ret centrale spørgsmål (se fx Woodward, 2011), men her er udvalgt to af de klassiske forklaringsmodeller, som er særligt vigtige for diskussionerne af forklaringer af modeller med store datamængder: *forklaringer fra lovmæssigheder* og *forklaringer fra (mekanisk) kausalitet*.

Som et første forsøg på indkredse betydningen af en forklaring kan vi skelne mellem videnskabelige udsagn og *forklaringer* af selv samme videnskabelige udsagn. En simpel første begrebsafgrænsning kunne være, at selve udsagnet hævder, *at* noget er tilfældet, hvorimod forklaringen hævder, *hvorfor* noget er tilfældet. Denne tilgang skelner mellem *hvad-spørgsmål* ('what questions') og *hvorfor-spørgsmål* ('why questions').

Hvor videnskabelige udsagn henter belæg i form af evidens, som for eksempel kan være empirisk, må forklaringer hente deres belæg et andet stedet fra. Men præcist *hvor* vi skal finde dette belæg afhænger i høj grad af hvilke andre filosofiske antagelser, vi har gjort os.

Hvis man fx går ud fra et synspunkt som *logisk positivisme*, så vil førende logiske positivister typisk hævde, at forklaring opstår som svar på et *hvorfor-spørgsmål*, og det som kan forklares er fænomener, der falder under en af de generelle lovmæssigheder, som er udledt ved induktion ud fra de fordomsfrie observationer (se Figur **??** i kapitel 1). En forklaring på, hvorfor Saturns bane er ellipseformet ville således være, at ud fra Keplers love er *alle* planeters baner elliptiske, og Saturn er en planet. På den måde bliver forklaringer ud fra dette synspunkt til deduktioner.

Denne ide om hvad en forklaring består i, blev fremført klarest af Carl Gustav Hempel (1905–1997) i 1940'erne, og den har siden udgjort den position, som de fleste andre forslag på at beskrive, hvad det vil sige at forklare en videnskabelig sammenhæng, er gået ud fra at forfine. Man kalder ofte denne form for forklaringsbegreber for *deduktiv-nomologiske* (DN) forklaringsteorier, idet forklaringen består af en udledning (*deduktion*) ud fra lovmæssigheder (som filosoffer betegner *nomologiske*). Vi kan sammenfatte positionen som at forklaringer har form af svar på hvorfor-spørgsmål af formen, at påstanden følger deduktivt fra lovmæssigheder.

Siden Hempels forklaringsmodel er der formuleret en del andre, og særligt forklaringsmodellen baseret på såkaldt *mekanistisk kausalitet* udgør et interessant alternativ (se evt. Ross og Woodward, 2023). Den indfanger en anden intuition omkring forklaringer, denne gang baseret på intuitioner om fysiske hændelser, nemlig at det, der skal forklares, er en kausal virkning af en årsag. Denne tilgang indfører dermed et tidsligt element i forklaringen, idet årsagen må forekomme før virkningen og være en direkte (mekanisk) *trigger* af denne, ligesom en billardkugle, der rammer en anden kugle, er årsag til, at den anden kugle bevæger sig.

En tredje forklaringsmodel forsøger at indfange *forklaring* ved hjælp af intuitioner om *statistisk relevans* — denne slags forklaringer kaldes derfor *SR-forklaringer*. Som det formuleres i Woodward (2011) er en egenskab C for en population A statistisk relevant for en

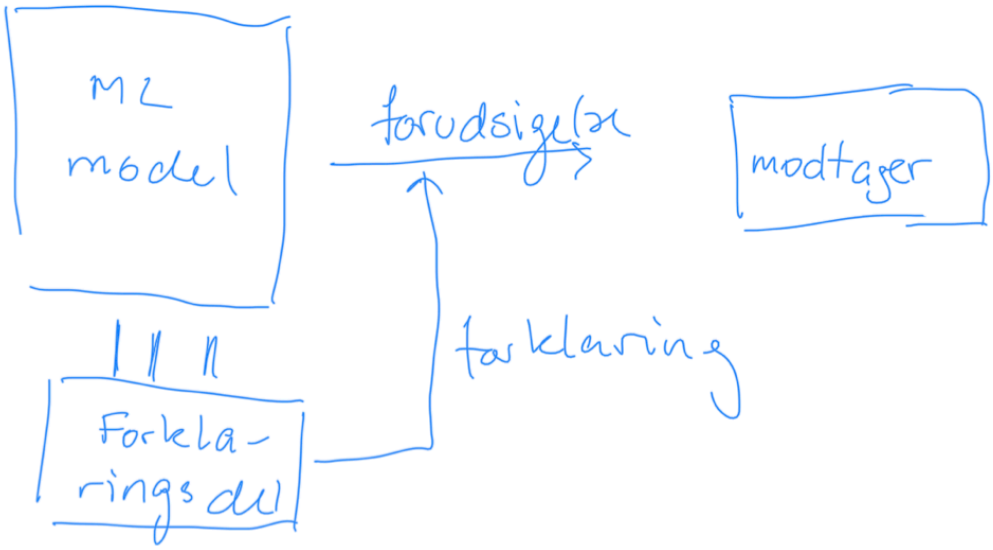
anden egenskab B, hvis $P(B|A.C) \neq P(B|A)$, dvs. hvis den betingede sandsynlighed for B givet A forandres af antagelsen af, at C også indtræffer. *SR-forklaringsmodellen*, fremført af Wesley Salmon (1925–2001), siger da, at statistisk relevante forhold (som C) er *forklarende* for de forhold, som de er statistisk relevante for (i vores tilfælde B).

Forklaring af modeller

Hvis vi søger at anlægge Hempels model om forklaringer ud fra deduktioner fra lovmæssigheder til at forklare *black boxes* i anvendelsen af machine learning, så løber vi ind i en variant et klassisk problem ved den logiske positivisme (se også Kapitel 1): Hvordan kan vi opstille og tilgå de lovmæssigheder, hvis de er opstillet af en stor og kompliceret *machine-learning-model* ud fra korrelationer i vores eksisterende datasæt. Vi kan sagtens bruge modellen til at lave forudsigelser (svarende til deduktioner hos Hempel), men at formulere en *forklaring* af en forudsigelse som „det følger af modellen“ uden at være i stand til at åbne den sorte kasse, som omgiver modellen, er ikke nogen god forklaring, selv ud fra det positivistiske synspunkt, da det næppe kvalificerer som et svar på et hvorfor-spørgsmål, men i stedet henviser til modellen som *blind autoritet*.

Men hvorfor ikke lade en computer forsyne forklaringer af *machine-learning-modeller*, der er uigennemskuelige og uigennemtrængelige *for mennesker*? Man har forsøgt sig med forskellige moduler, som enten kan bygges ind i selve den oprindelige *machine-learning-model* eller kan kobles efter den (se Figur 8). Men alle disse forsøg er (hidtil) stødt ind i de samme principielle problemer, som er forbundet med at give en *forklaring*: De er typisk baseret på *visualiseringer*, som giver gode måder at *fortolke* modellens forudsigelser (se næste afsnit), men ikke er formulerede i hverken lovmæssigheder eller årsagssammenhænge. Der findes dog også lovende forsøg på at bygge *selvforklarende modeller*, fx såkaldte *bayesiske machine-learning-modeller*, der er en slags generalisering af *beslutningstræer*, som de blev brugt i de første ekspertsystemer (se også Kapitel 7). Forskellen fra de overskuelige, klassiske *beslutningstræer* er imidlertid at også i denne form for *machine-learning-modeller* bliver selve *mængden* af skridt i en *forklaring* uoverskueligt stort.

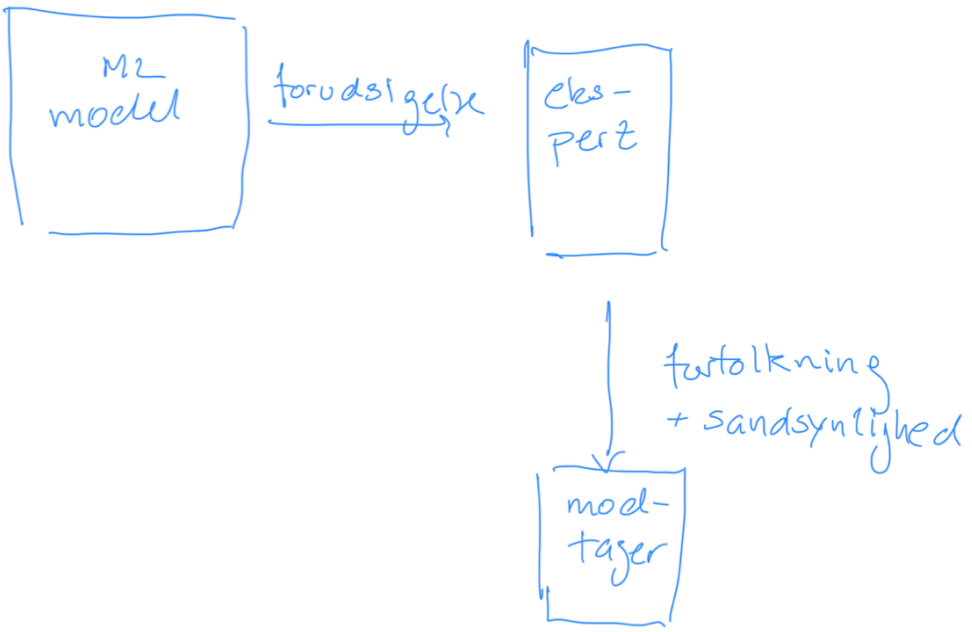
Når en *machine-learning-model* bliver uoverskuelig, fx fordi antallet af vægte bliver enormt stort, kan man forsøge at bygge andre typer forklaringsmodeller ind i tekniske løsninger (se Figur 8). Hidtil har man dog ikke opnået forklaringer af *machine-learning-modeller*, der levet op til klassiske intuitioner om forklaringsmodeller, hverken i form af lovmæssigheder, statistisk relevans eller mekanistisk kausalitet.



Figur 8: Skitse af en *machine-learning-model*, hvortil er tilføjet en *forklaringsenhed*, enten internt i modellen eller som en teknisk tilføjelse.

Fortolkning af modeller

I stedet for at søge forklaringer *inden i* videnskabelige teorier eller metoder, kan man også anlægge et *kommunikativt* perspektiv på forklaringer: *Nogen* forklarer *noget* til *nogen*. Med denne tilgang vil *forklaringen* næppe leve op til de strenge forhåbninger diskuteret ovenfor, så for at adskille denne proces fra Hempel s forklaringsmodel og kausale forklaringer, betegner vi den her som *formidling af modeller* og deres forudsigelser.



Figur 9: En fortolkning af en models forudsigelser kan anskues som en kommunikationsproces mellem modellen, en ekspert-fortolker og en modtager.

Dette kommunikationsbaserede syn på *fortolkning* af *machine-learning-modeller* involverer tre *aktør*aktører: selve modellen, som producerer forudsigelser, en ekspert, som har tilstrækkelig viden om og forståelse af modellen til at kunne omsætte forudsigelserne til fortolkninger, og en modtager. I forhold til det syn på *forklaringer* af *machine-learning-modeller*, som blev præsenteret ovenfor, er hele processen altså tænkt som en kommunikation og der indskudt en (menneskelig) *ekspert*. At processen er en kommunikation betyder altså blandt andet, at information og spørgsmål kan flyde begge veje.

I praksis vil en af ekspertens vigtigste (og sværeste) opgaver hyppigt være at udlægge sandsynligheder på en tilgængelig vis for en almindelig modtager. Den slags udfordringer er ikke nye, og der er fx i lægeverdenen udviklet forskellige redskaber til at inddrage patienten i vigtige beslutninger, der er baseret på sandsynligheder. En af de mest brugte metoder er at inddrage *visualisering* i form af kurver eller kort, der fx angiver 'høj', 'middel' eller 'lav' risiko for at oversætte procentsatser og tekniske formuleringer til et mindre formelt medium.

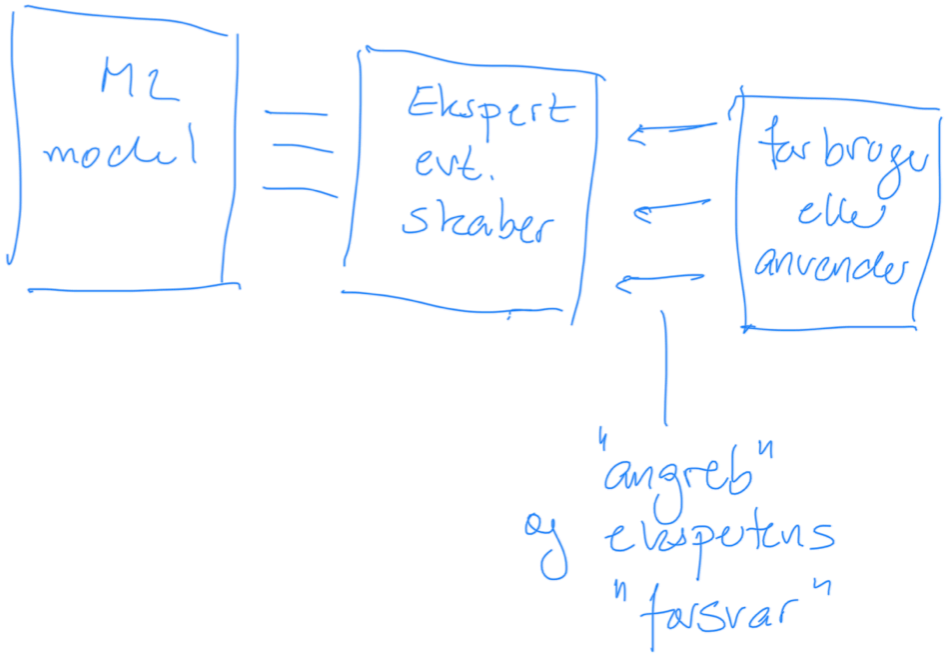
Tidligere var disse former for scenarier typisk bygget op omkring *proxies*, der er lettere at måle og for hvilke, der ofte er en traditionel forståelse også hos modtageren. En vigtig del af fortolkning af sandsynligheder og inddragelse i valgprocesser for modtageren er muligheden for at påvirke, hvorvidt modellens forudsigelser også kommer til at holde stik. Derfor kan nye, nærmest interaktive *scenarie-beregninger* næsten virke som en form for personlig coach. I stedet for at få at vide, hvordan forudsigelserne ville se ud for andre i nogle af mine nærmeste referenceklasser (typisk udvalgt efter traditionelle *proxies*), kan jeg nu interaktivt eksperimentere med forskellige parametre i modellen. Men for at disse eksperimenter giver faglig mening og for at udlægge modellens forudsigelser og begrænsninger, er ekspertens rolle vigtig. Eksperten kommer dermed til at kombinere *domæneviden* (fx inden for lægevidenskab) med tilstrækkelig indsigt i modellen til at bygge en form for bro mellem modtageren og den for vedkommende uigennemsigtige model.

Forsvar af modeller

Som et tredje og sidste forsøg på at antyde, hvordan vi kan komme til bedre forståelse af *machine-learning-modeller*, kan man betragte en anden form for kommunikation mellem en model, en ekspert og en eller flere modtagere. Hvor dialogen i fortolkningen af modeller fandt sted mellem modellen og modtageren med eksperten som mediator, er ideen i dette tilfælde, at eksperten skal kunne *forsvare* modellen imod angreb fra modtagerne (se Figur 10). Modtagernes angreb kan være kritiske spørgsmål eller påpegning af uhensigtsmæssige forhold omkring modellens forudsigelser. Hvis eksperten på troværdig vis kan *forsvare* modellen, kan vi være tilbøjelige til at tilskrive eksperten *forståelse* af modellen.

Men hvori består så sådanne forsvar? Det er ikke tilstrækkeligt, at eksperten kan give en teknisk forklaring ved at opremse faktiske forhold omkring modellens konstruktion. Der må også skulle finde et vist stykke tilpasning til modtagernes angreb sted, men vigtigere er det nok, at ekspertens forsvar har en tilstrækkelig *holistisk* tilgang til at kunne forklare modellens *opførsel* og *funktion*. Herunder vil det også være relevant at kunne forsvare modellens kvalitet og kvalificere dette på individniveau. Det vil være oplagt, at eksperten også skal kunne forsvare de valg, som er truffet i modellens konstruktion (se Figur 7).

Forsvaret af en model ligner altså på flere måder det, som vi ovenfor har kaldt fortolkning af modellens forudsigelser. Men alligevel adskiller de to tilgange sig: Det at kunne forsvare en model kræver en del mere indsigt fra eksperten — tilgengæld er hun 'kun' forpligtet på at kunne forsvare modellen mod de angreb, der faktisk forekommer.



Figur 10: Som forsvar af en model skal eksperten kunne svare på kritiske spørgsmål og indvendinger mod modellens opbygning og funktion.

Litteratur

Anderson, Chris (23. jun. 2008). „The End of Theory. The Data Deluge Makes the Scientific Method Obsolete". *Wired*, bd. 16, nr. 7, s. 108–109.

Dahl-Jensen, Dorthe (okt. 2009). „Istiden sluttede ekstremt hurtigt". *Kvant*, s. 3–8.

Johansen, Mikkel Willum og Henrik Kragh Sørensen (2018). „Modelvalg og ansvar". *Aktuel Naturvidenskab*, nr. 1, s. 36–39.

Ringgaard, Anne (31. maj 2017). „Korrelation eller kausalitet. Hvornår er der en årsagssammenhæng?" *Videnskab.dk*.

Ross, Lauren og James Woodward (2023). „Causal Approaches to Scientific Explanation". I: *The Stanford Encyclopedia of Philosophy*. Red. af Edward N. Zalta og Uri Nodelman. Spring 2023. Metaphysics Research Lab, Stanford University.

Sørensen, Henrik Kragh og Line Edslev Andersen (mar. 2018). *Mapping Social Aspects of Mathematical Knowledge Production*. RePoSS: Research Publications on Science Studies 43. Aarhus: Centre for Science Studies, University of Aarhus.

Wolfram, Stephen (2002). *A New Kind of Science*. Champagne (IL): Wolfram Media.

Woodward, James (2011). „Scientific Explanation". I: *The Stanford Encyclopedia of Philosophy*. Red. af Edward N. Zalta. Winter 2011.