FØLG PROJEKTET BIG DATA

Forskere vil samle data på alt fra SoMe til blodprøver: Integrationen er et kæmpe problem

Magnus Boye, @magnboye

https://pro.ing.dk/1920

24. jan 2019 02:19

Ved at sætte dansk registerdata sammen med data fra sociale medier, sensordata og løbeapps vil danske forskere kortlægge, hvordan vores sundhed bliver påvirket af miljøet.

I sidste måned åbnede Aarhus Universitet centeret BERTHA – Big Data Centre for Enviroment and Health – og foran sig har forskerholdet nu en data-udfordring af dimensioner, fortæller professor Clive Sabel, der skal lede centret.

»Mange af de ting, som dræber os nu, er sygdomme, man får som gammel, som fx hjertesygdomme og Alzheimers. Vi har ideer om, hvad der forårsager de sygdomme, men vi kun få beviser. Vi tror, at vores tilgang, kan give indsigt i sygdomme, der er svære at forstå, og hvor årsagerne er svære at identificere.«

For at levere den indsigt er der brug for ikke bare Big Data, men Rich Data.

»Fra mit synspunkt er Big Data bare den sidste catchphrase for noget, vi har arbejdet med i årtier, hvor vi har brugt algoritmer til at finde mønstre i store datasæt. Vi plejede ikke at kalde det Big Data, men det var Big Data,« siger Clive Sabel og fortsætter:

»Det, vi ser nu, er, at det ikke er det, at data er 'big', der gør dem interessante eller

udfordrende. Vi vil gerne bruge begrebet Rich Data i stedet.«

Rich Data-begrebet er coinet af Osmar Zaiane, Datalogi-professor ved Alberta University, som sidder i advisory boardet for BERTHA-projektet. Begrebet dækker over data, der er integreret fra mange meget forskellige datasæt.

»Du kan have en meget stor database og kalde det en big data, men for mig handler det om, at data kommer fra alle mulige kilder, fra hele landet, fra borgere og regeringer. Det kan være enten individuel data eller data om et område, det kan være målerdata, der kommer hvert minut.«

Data-selskaber kæmper også

Med Rich Data kommer potentialet for at få svar på nogle også ellers ubesvarlige hypoteser – som sammenhængen mellem en persons åndedrætsproblemer og luftkvaliteten under dennes løbetur.

Men det giver også udfordringer, fastslår Clive Sabel.

»Vi har et sæt meget komplekse problemer. Dels hvordan vi kan integrere data, og dels hvordan vi miner det for mønstre,« forklarer han.

Af gode grunde, kan forskningscenteret ikke samle al data på samme sted. Registerdata skal tilgås gennem Danmark Statistiks forskerservice, som sørger for at data, der kan bruges til at identificere en enkelt person, ikke hentes ned.

»At sammenkoble data, noget vi ejer og ikke ejer, noget der kommer hvert minut, noget der kommer hvert årti, er enormt udfordrende. Vi har talt med store data-selskaber, og de kæmper også med at integrere data over mange forskellige skalaer,« siger Clive Sabel, der meget gerne hører fra danskere, der har en ide om, hvordan man løser det.

»Dataintegration er et kæmpe stort problem,« konstaterer han.

Data fra citizen scientists

Forskningscentret skal udnytte Danmarks registerdata, men sammensætte det med data fra såkaldte citizen scientist – borgere, der frivilligt deler data gennem digitale platforme.

»Vi arbejder sammen med Garmin, og derigennem vil vi meget gerne have data fra danskere, der giver os lov til at bruge det. Det giver os data om fx løbe- og cykelture, der måske uploades dagligt,« fortæller Clive Sabel.

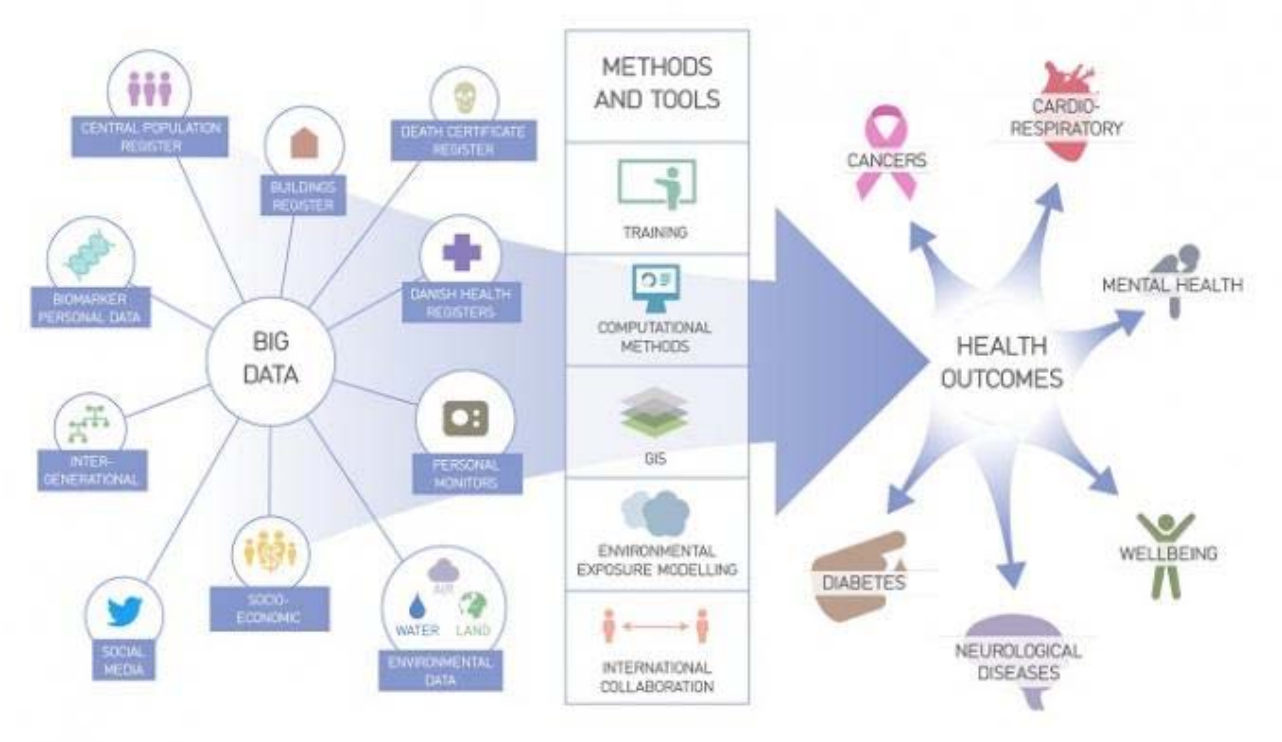
»Samtidig har vi sensordata om luftforurening, som opdateres hvert minut. Andre steder – som CPR-registret – bliver data kun opdateret, når nogen flytter.«

Dertil kommer blodprøver fra bloddonorer. Typisk bliver de samme personer ved med at donere i årevis hver 6. måned og giver på den måde indsigt i miljøet omkring donoren på de pågældende tidspunkter. Den data alene sammenligner Clive Sabel med iskerneboringer i Arktis, der kan vise forurening i luften fra over 1.000 år siden.

Samtidig arbejder forskerne på at få SoMe-data fra frivillige.

»Med deres tilladelse kan vi bruge text-mining og undersøge sammenhænge mellem, hvor folk er, og hvilket humør de er i. Hvis mange siger de samme ting om et sted, kan vi se det som et punkt, der giver folk glæde, i modsætning til en motorvej, hvor folk måske tweeter om bilkøer og så videre,« siger Clive Sabel og fortsætter:

»Alt i alt har vi en enorm variation af rum og tid i vores data, og det gør det meget udfordrende.«



En visualisering af den overordnede målsætning og arbejdsmetode for BERTA-projektet.
Illustration: BERTHA

Komplekse hypoteser

Den anden del af udfordringen for BERTHA-projektet handler om datamining.

»Hvordan kan vi udvikle algoritmer, der kan se efter mønstre i det her ekstremt komplekse datasæt,« spørger Clive Sabel retorisk.

»Kan vores data vise, hvordan dårlig luftkvalitet påvirker åndedrætssystemet hos folk, der løber. Eller om folks mentale helbred ændrer sig, når de er i grønne miljøer som i skoven eller på stranden. Det er meget vanskeligt at besvare hypoteser som dem,« siger Clive Sabel, der blandt andet håber at få fat i en ph.d-studerende eller en forsker, der kan udvikle spatial analytics som en del af BERTHA-projektet.

Analysen skal sikre fortrollighed og integritet og foregå på tværs af et distribueret netværk. Derudover har BERTHA-holdet ikke lagt sig fast på, hvordan det teknisk skal foregå.

»Vi er stadig meget tidligt i projektet. Vi er på det punkt, hvor vi overvejer en række forskellige dataløsninger,« siger Clive Sabel og tilføjer:

»Vi kan nok ikke gøre, som man som datalog ønsker, hvor vi renser alting, indtil det er perfekt. Dertil kommer data til at være alt for rodet.«

BIG DATA

Magnus Boye
Redaktør på DataTech. Magnus er uddannet i Journalistik og Informatik fra Roskilde Universitet og har tidligere dækket it-ledelse og forretnings-it for Version2 og Ingeniøren.