

TEKNOLOGI

Jagten på nålen i høstakken er vendt på hovedet

International Data Corporation (IDC) konstaterer, at da der i 2006 blev produceret 161 exabytes, var det lig med »3 mio. gange mere information end den, der er i alle bøger, som nogensinde er blevet skrevet«.

Clark Glymour, der er professor ved MIT, indså det allerede i 2004. I en artikel fra dengang sammenligner han videnskabens jagt på ny viden med at finde en nål i en høstak:

I gamle dage fiskede man et eller andet ud af høstakken og undersøgte det grundigt for at se, om der var hul i den ene ende. Hvis der var det, kunne man formode, at der var tale om en nål. Hvis ikke, begyndte man forfra med at rode. »Fortsæt til du finder nålen, eller til civilisationen når sit endeligt«.

Sådan lød devisen i gamle dage, ifølge Glymour. Den nye teknologi gav derimod forskerne mulighed for at »bevæge sig væk fra det gamle paradigme og omfavne automatisering, som svarer til at sætte ild til høstakken og puste asken væk for at blotlægge nålen«. (rtb)



Tegning: Philip Younert

Filosoffen: Vi kender vores begrænsninger

Mennesket håndterer informationsalderen fint, så længe det accepterer sin rolle i den, forklarer filosof Arno Victor Nielsen.

Den eneste grund til at fortvivle over og drukne i de enorme mængder information skulle således være, hvis man troede, at man kunne vide alt – at man skulle være encyklopædist.

Og det gør de færreste sig trods alt forhåbninger om. I stedet forstår mennesket, både menigmand og forsker, at så store mængder information kræver arbejdsdeling.

»Du har et lille hjørne, som du har styr på, og et sted på den anden side af kloden sidder der en, som har styr på et andet hjørne. Vi er afhængige af relationer for at håndtere informationsmængderne«, forklarer Arno Victor Nielsen.

Han beskriver computeren som et værktøj, der hjælper mennesket med at skabe både relationerne og overblik. (rtb)



Tegning: Philip Younert

Informationssamfundet drukner i ... data

Menneskeheden genererer enorme, ufattelige mængder information, og vi overgår os selv med længder hvert år. De store datamængder har et kæmpestort potentiale for videnskaben, men både forskere og menigmand risikerer også at drukne i dem.

RASMUS THIRUP BECK

En exabyte er lig en milliard gigabytes. I 2010 alene vil menneskeheden tilsammen frembringe 1.200 exabytes. En exabyte svarer til en dvd-film, der varer 50.000 år. Kan nogen forstå det?

Man kan prøve at sno sin hjernebark omkring de enorme tal på andre måder. Mellem 2000 og 2002 blev der ifølge forskere på Berkeley University på verdensplan produceret lige så meget information som i alle år før det. Det gentog sig i 2006 og vil snart gøre det igen.

Stadig for ubegribeligt? Man kan også tage fat i lidt mere konkrete eksempler på informationsekspllosionen:

I februar i år offentliggjorde YouTube, at der nu bliver uploadet 24 timers video til deres tjeneste hvert minut – i midten af 2007 var det 'blot' 6 timer i minutet. Brugere af Facebook har indtil videre lagt mere end 40 milliarder billeder ud på nettet.

Eller hvad med et eksempel fra videnskaben, som står for en stor del af dataproduktionen: Da et af de mest ambitiøse astronomiprojekter, The Sloan Digital Sky Survey (SDSS), blev sat i gang i 2000, indsamlede dets teleskop i løbet af blot et par uger mere data end i hele astronomiens historie.

Indtil videre har SDSS indsamlet 140 terabytes svarende til 140.000 gigabytes, og det nuværende teleskop arvtager vil kunne indsamle den mængde på bare fem dage.

Og lige præcis videnskabsfolk får stjerner øjnene, når de taler om potentialet i de enorme datamængder.

Lars Arge, der er professor i datalogi ved Aarhus Universitet og leder af Danmarks Grundforskningsfonds Center for Massive Data Algorithmics (center for massive dataalgoritmer), Madalgo, arbejder specifikt med at hjælpe forskere med at bearbejde datamængderne, og han har svært ved at skjule sin begejstring.

»Der ligger nogle utrolige muligheder gemt. Der er virkelig meget, man kan finde ud af. Der er usandsynlig meget data, og der må nødvendigvis være en hel masse opdagelser, som bare venter på at blive opdaget«, som han siger over telefonen fra Århus.

Det store men

Desværre for Lars Arge og hans forskerkolleger verden over ligger der også et voldsomt stort 'men' gemt i begejstringen – afsøret af ordene »man kan« og »der må«. Datamængderne bliver nemlig så store, at der ikke eksisterer software, der kan håndtere dem. Datamængden overstiger simpelthen regnekraften.

Datalogiprofessoren giver et eksempel fra Madalgo's eget arbejde. De seneste år er Danmark flere gange blevet topografisk opmålt med laser, hvorfor man nu har enorme modeller, der præcist kan gengive, hvordan landet bufter sig.

Hvor man for bare ti år siden kun havde målepunkter for hver hundrede meter, er der nu op til 40-50 punkter per kvadratmeter – altså en imponerende detaljerig-



dom, som i princippet meget præcist kan forudsige, hvad vej vand vil løbe ved voldsomme regnskyl, eller hvilke dele af landet der vil blive oversvømmet, når havene stiger.

»Men indtil videre har ingen kunnet bruge modellerne til noget, fordi de har været for store til, at softwaren kan håndtere dem«, konstaterer Lars Arge.

Selv når datamængden er så overskuelig, at softwaren kan klare den, har kompleksiteten haft en konsekvens, at forskerne inden for nogle forskningsområ-

der, især fysikken, er blevet afhængige af dataloger for at kunne komme til at se de mønstre, de leder efter. Eller som Thore Husfeldt, der er docent i datalogi med speciale i algoritmer og datastrukturer ved Lund Universitet og lektor i det samme på IT-Universitetet i København, siger: »Der er gledet en linse af algoritmer ned mellem forskerne og deres materiale«.

Fænomenet har selvfølgelig også fået navn. Amerikanske videnskabsfolk og dataloger refererer således tørt til de tyn-

gende informationsmængder som 'big data' – altså 'store data'. Den anerkendte datalog Joe Hellerman fra Berkeley er gået et skridt videre og har kaldt den vilde udvikling for »dataens industrielle revolution«.

Og Lars Arge viger ikke tilbage fra de amerikanske kolleger. Ligesom mange andre mener han, at de store datamængder har ført til et decideret paradigmeskifte med hensyn til, hvordan forskere arbejder.

»Før brugte de langt størstedelen af deres tid på at indsamle data, som de så analyserede. I dag bruger de fleste det meste af deres tid på at analysere – data trækker de 'bare' ud af de store databaser«, forklarer han.

Servere på flere kvadratkilometer

Med den enorme dataproduktion kommer også et stadig stigende behov for datalagring. Den individuelle computer-bruger kender det selv. For fem år siden kunne de fleste nøjes med en harddisk på 50-100 gigabytes, mens få computere i dag bliver solgt med mindre end 200 gigabytes – og så er et stadig voksende antal computerbrugere oven i købet nødt til at supplere med eksterne harddiske i varierende omfang.

Går man et niveau op, begynder lagringsomfanget at blive lige så overvældende som datamængderne. Apple nærmer sig færdiggørelsen af en enorm såkaldt serverpark i North Carolina. Byggeprisen er en milliard dollar, og de endeløse rækker af servere, der skal være opbakning for firmaets nye Cloud-teknologi – som kort sagt går ud på, at brugerne lagrer deres ting online i stedet for på deres egen computer – kommer til at fylde godt 46.000 kvadratkilometer. Google (der-

ENERGISLUGERE. Computerserverne bliver større og større i takt med den stigende informationsmængde. Og de sluger en masse energi, fordi de skal køles. Billedet er fra det internationale forskningscenter Cern i Schweiz.

Arkivfoto: Martial Trezzini, Keystone, AP

under YouTube), Facebook, Microsoft, IBM og en lang række andre firmaer har eller er i gang med at bygge tilsvarende digitale lagerparker. Statsinstitutioner viger ikke tilbage. Ifølge en række amerikanske medier er efterretningstjenesten NSA således ved at bygge en serverpark på næsten 100.000 kvadratkilometer i Utah.

En temmelig overraskende konsekvens af de kæmpemæssige serverparker er, at lagring af menneskeheden store mængder data er blevet en klimamæssig udfordring – fordi store servere kræver tilsvarende store køleelementer. Servereksperter vurderer således, at CO₂-udledningen fra servernedkøling i 2020 vil overgå CO₂-udledningen fra flytransport.

Tallet overrasker overhovedet ikke Lars Arge:

»Det er et problem, og det er noget, som der er ved at komme fokus på i forskerverdenen. Man konkurrerer f.eks. i, hvem der kan sortere data med brug af mindst energi og ikke som før kun hurtigst«.

Nogle lagringsfirmaer kigger mod nord efter en løsning. Island er således ved at køre sig selv i stilling som et center for den globale datalagring, fordi de for det første har en varm undergrund til at skabe bæredygtig energi og for det andet har et koldt klima. Lagringsfirmaet Verne Global er som det første gået i gang med at bygge en næsten komplet bæredygtig serverpark – i første omgang dog kun på sølle 4.200 kvadratkilometer.

viden@pol.dk

Se, hvor mange bit information, der er dannet alene i 2010. Tallet er baseret på verdens største datalagringsvirksomhed, EMC's, beregninger: http://www.emc.com/digital_universe/ Ved deadline fredag var vi oppe på over 223 exabyte.

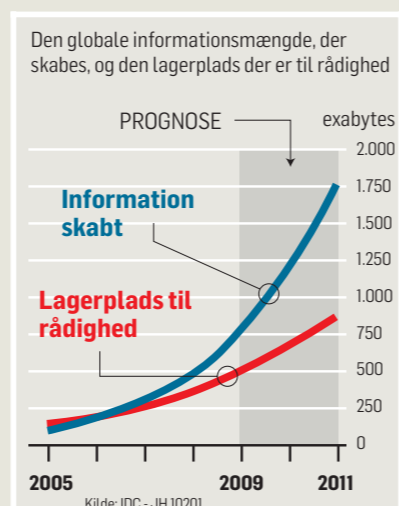
Informationsoverload

2007 var et banebrydende år for den globale produktion af data. For første gang i verdenshistorien overgik informationsmængden nemlig den samlede lagringskapacitet – og siden er forskellen på de to blevet stadig større.

Den seneste opgørelse fra The International Data Corporation i 2009 forudsiger således, at informationsmængden her ved indgangen til 2010 ville være cirka dobbelt så stor som lagringskapaciteten.

Al den genererede information er naturligvis ikke nødvendig at gemme. Live net-tv og netradio, som ikke bliver optaget, overvågningskameraer, som bliver overspillet efter et stykke tid, billeder, som brugerne tynder ud i og så videre, kan verden godt leve uden.

Men kløften mellem dannelse og lagring af data vil ifølge et voksende kor af it-eksperter med tiden blive et pro-



blem for de mennesker, der har ansvaret for, hvad vi skal gemme for eftertiden – og hvad skal vi slette. Medmindre lagringskapaciteten indhenter dataproduktionen, vil de simpelthen skulle vælge stadig mere fra.