

KUNSTEN Å VITE

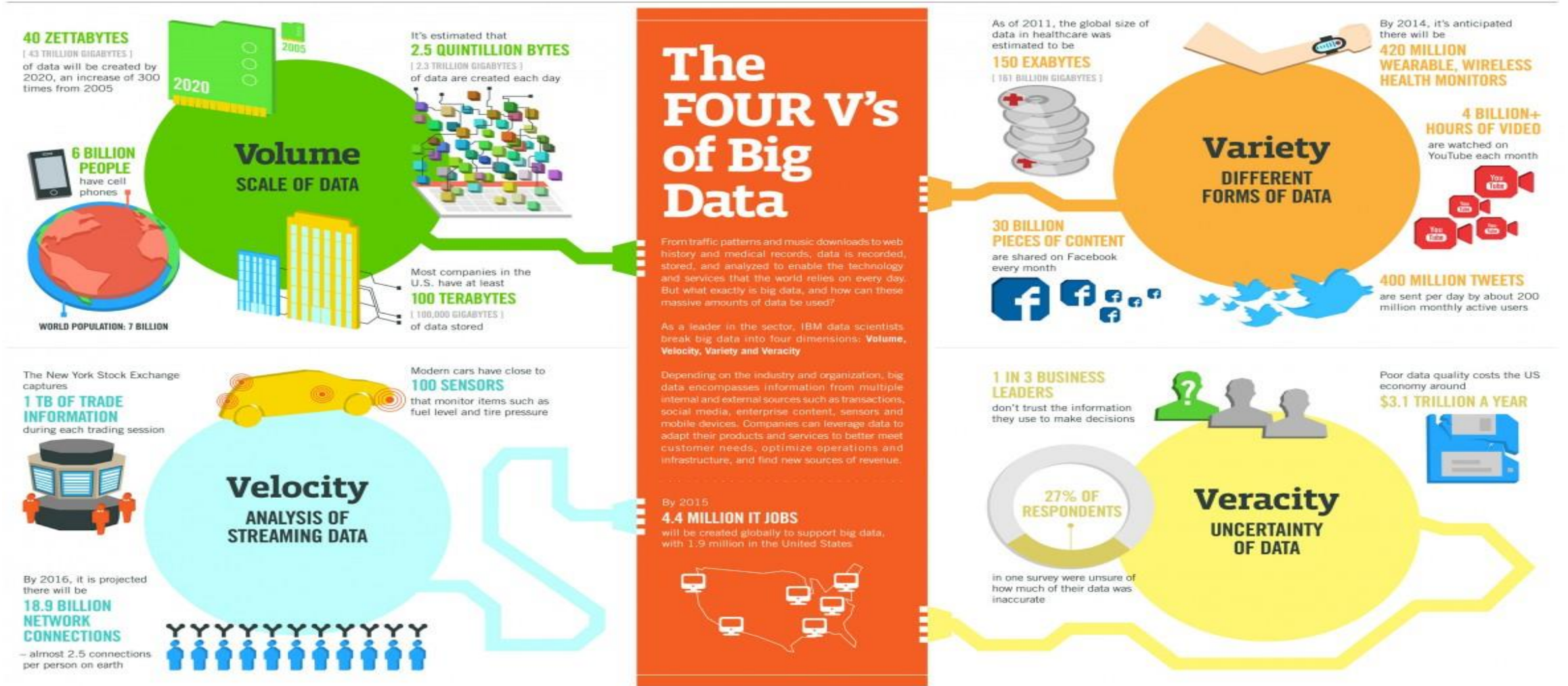
DATA SOM KILDE TIL INNSIKT

DFDS PEARL SEAWAYS

11. JUNI 2015

BERNT BREMDAL, PROF II - HIN, RÅDGIVER

HVA ALLE SNAKKER OM



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTec, QAS

IBM

Big Data - Bernt A. Bremdal

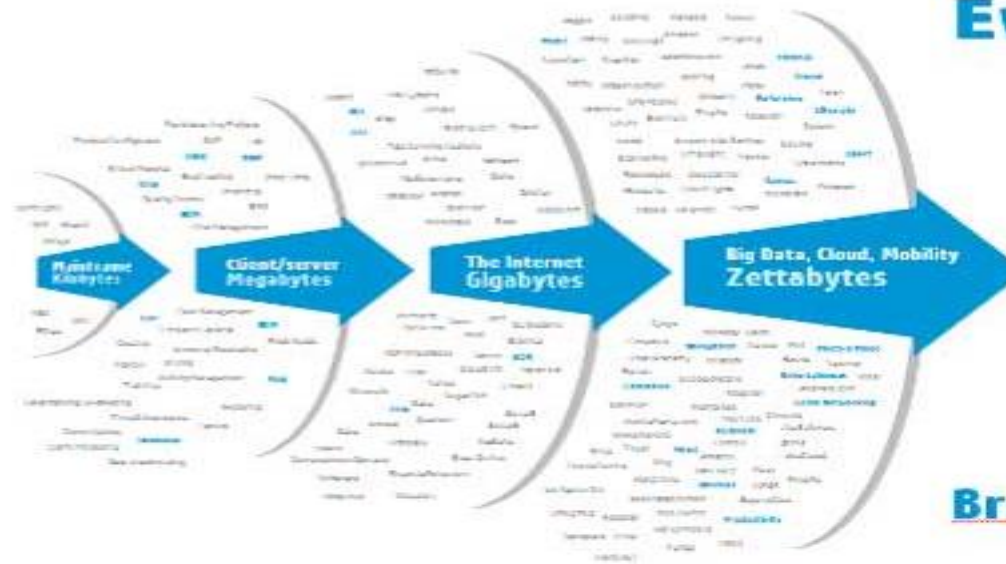
HVA ER EN ZETTABYTE?

- A zettabyte is 1,000 000,000,000,000,000 bytes, or units of information. Consider that a single byte equals one character of text. The 1,250 pages of Leo Tolstoy's *War and Peace* would fit into a zettabyte 323 trillion times.
- Or imagine that every person in the United States took a digital photo every second of every day for over a month. All of those photos put together would equal about one zettabyte.

..og så snakker vi brontobytes

VELOCITY

Information is changing the world



Every 60 seconds

- 204 million+ emails sent
- 100,000+ tweets
- 2 million+ Google searches
- \$275,000 spent online shopping
- 35,000 brand "Likes" on Facebook
- 38,000 new Tumblr blog posts
- 48 hours new video on YouTube
- 2,000 check-ins on Four Square

Brontobytes + Geopbytes

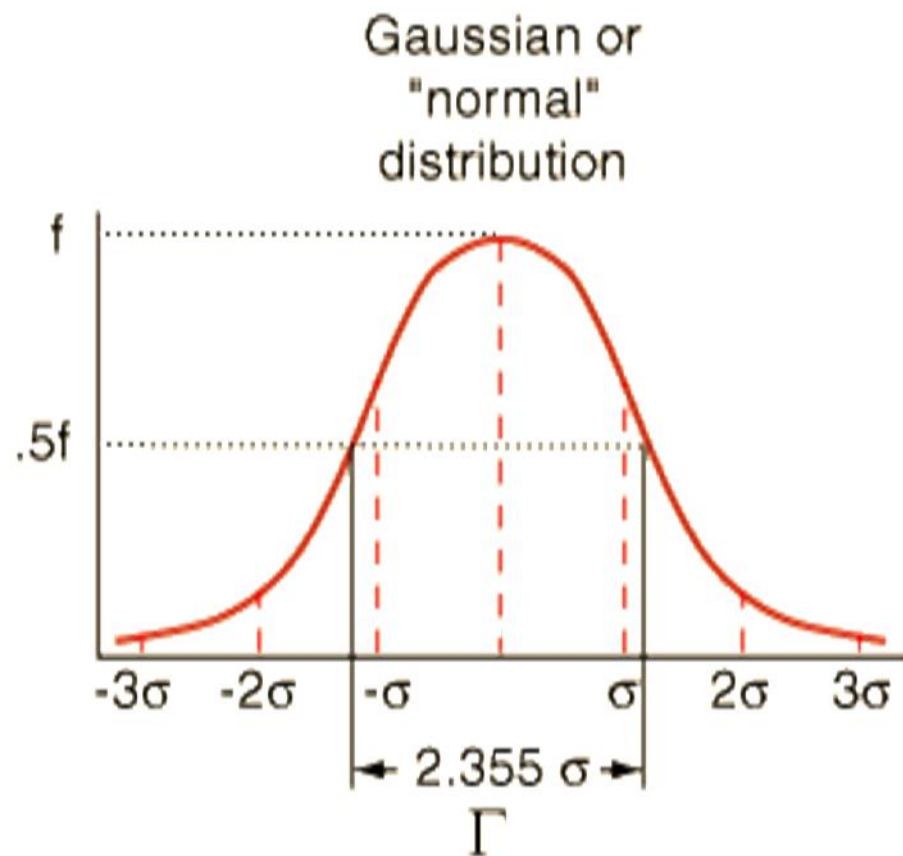
© Copyright 2013 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.



eWEEK

..Viktig å starte med det enkle og belage seg på større utfordringer senere

HVA GIR ØKT VOLUM OSS?



De underliggende strukturene i store datamengder blir forsterket

Rene stokastiske elementer vil undertrykkes (de danner ikke noe varig mønster)

«Hvit støy» er bra som sammenligningsgrunnlag

MED STORE VOLUM FØLGER OGSÅ ØKTE ANTALL DIMENSJONER

	2000	2014
Mobile phone images	Sharp-J-SH04: 0,11 MPixels	NOKIA-Lumia 10201 EOS: 41 Mpixels
Bioinformatics, biomarkers	Leukemia= 10.000	Psoriasis II = 2.000.000
General dimensional level in data analytics	100 (ex. Insurance)	1.000.000 (ex. YouTube)



OK. Det er jo
relasjoner og
krysskoplinger
mellom data som er
interessante

KORRELASJON VS KAUSALITET

Et sett med butikkdata fra en bensinstasjon viser at mange som fyller bensin eller diesel også kjøper pølse, is og boller.

Man kan dermed hevde at det er viss «korrelasjon» mellom «kjøp av drivstoff» og «kjøp av andre varer»

- Men blir folk sultne av å fylle drivstoff?
- Stopper folk for å kjøpe mat, men ser at de også bør fylle opp tanken?

Årsak – effekt sammenhenger krever kunnskap om den verden vi analyserer og knyttes til tolkning.

Korrelasjon eller samvariasjon er kan etableres gjennom statistisk analyse, multi-variat analyse og maskinlæring – det kan en datamaskin gjøre.

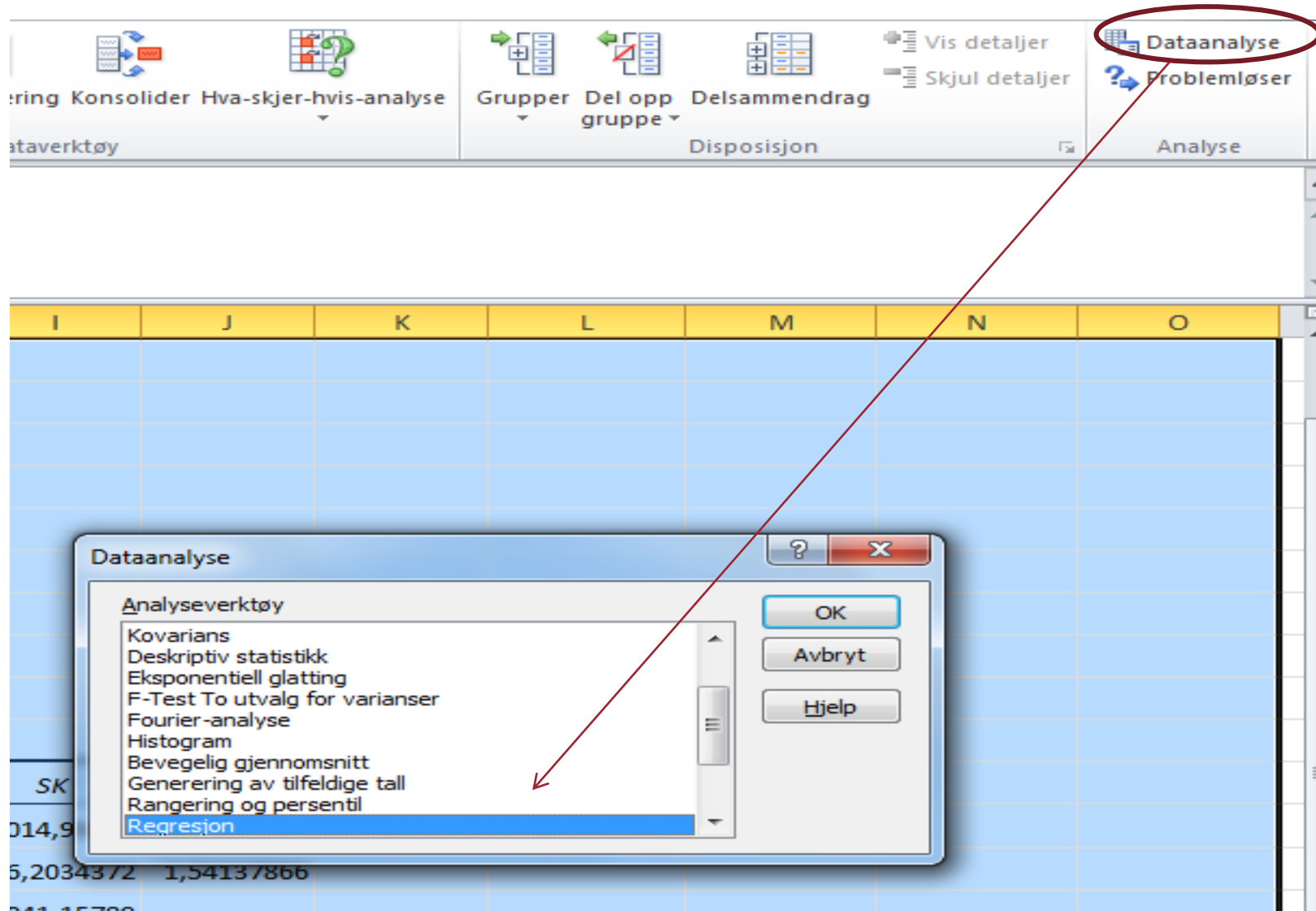
EXAMPLE: GAS STATION

Observation No.	Independent variable: Gas fillers (X)	Dependent variable: Buyers of other merchandise (y)
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82

Observationo	Independent variable: Gas fillers (X)	Dependent variable: Buyers of other merchandise (y)
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88
21	125	0
22	35	35

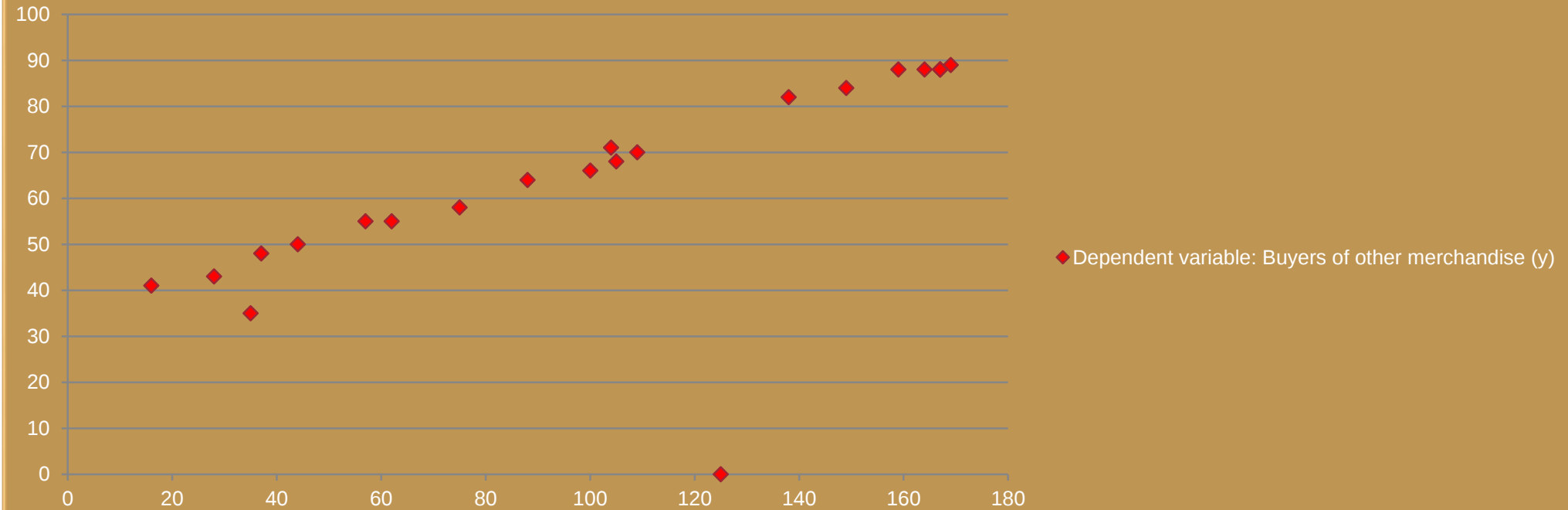
2 dimensjoner

EXCEL



SAMMENDRAG (UTDATA)								
Regresjonsstatistikk								
Multippel R	0,70280728							
R-kvadrat	0,49393807							
Justert R-kvadrat	0,46863498							
Standardfeil	16,5962904							
Observasjoner	22							
Variansanalyse								
	fg	SK	GK	F	Signifkans-F			
Regresjon	1	5376,76288	5376,76288	19,5208548	0,000264709			
Residualer	20	5508,73712	275,436856					
Totalt	21	10885,5						
	Koeffisienter	Standardfeil	t-Stat	P-verdi	Nederste 95%	Øverste 95%	Nedre 95,0%	Øverste 95,0%
Skjæringspunkt	33,1091367	7,93715956	4,17140874	0,00047119	16,55251202	49,66576145	16,55251202	49,66576145
X-variabel 1	0,30490022	0,06900941	4,41824115	0,00026471	0,160949109	0,448851325	0,160949109	0,448851325

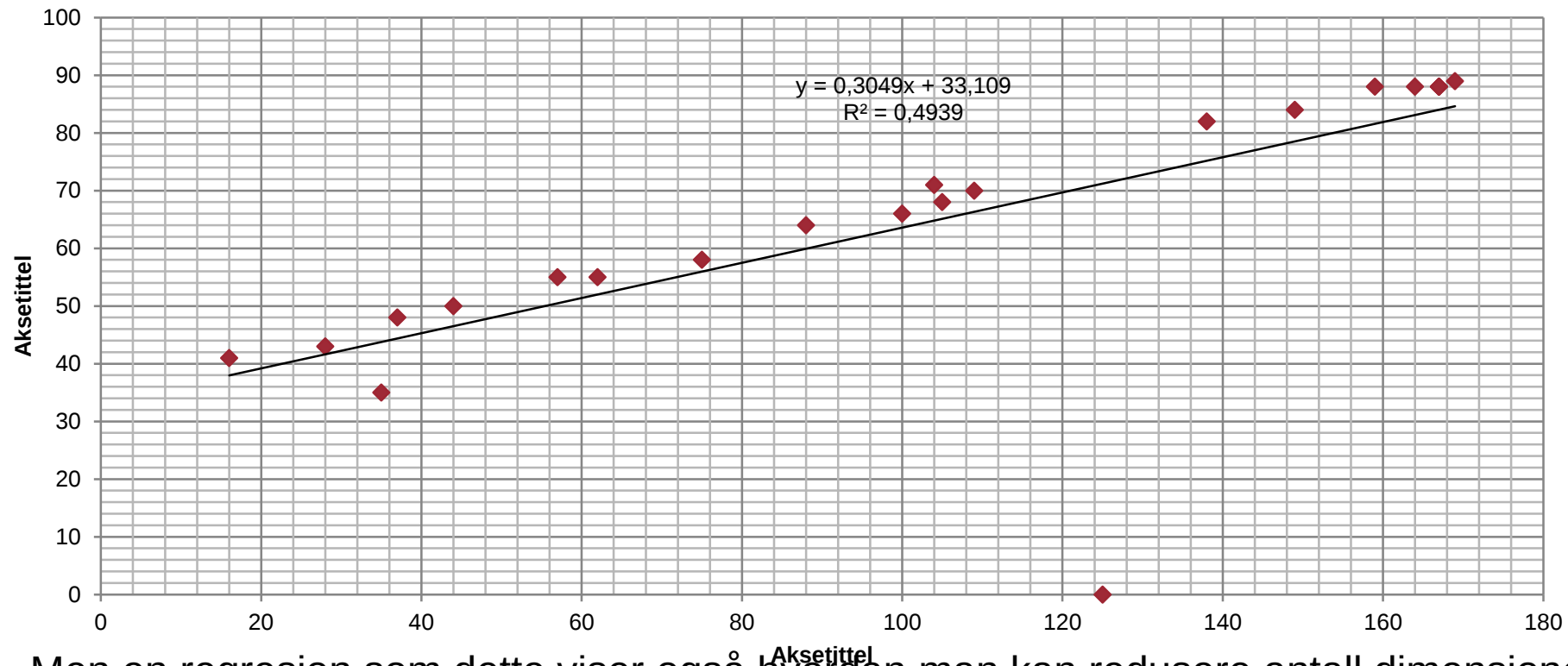
Dependent variable: Buyers of other merchandise (y)



NÅ HAR VI EN FORMEL OG VI KAN PREDIKERE

DERSOM 230 SJÅFØRER FYLLER DRIVSTOFF HVOR MANGE KAN VI FORVENTE KJØPER OGSÅ PØLSE, BOLLE ELLER IS?

$0,3049 \cdot 230 + 33,109$ gir ca. 103 Mao når 230 fyller drivstoff så vil 103 av disse handle mer



Men en regresjon som dette viser også hvordan man kan redusere antall dimensjoner. Vi har med dette tricket redusert problemet fra to dimensjoner til én.

MEN HVIS VI ØNSKET EN MER HØYOPPLØSLIG ANALYSE?

Drivstoff, Mat

bensin 98, bensin 95, diesel, pølse, bolle, kaffe, sjokolade, ukeblad, coca cola, vann, fanta, avis, baguette

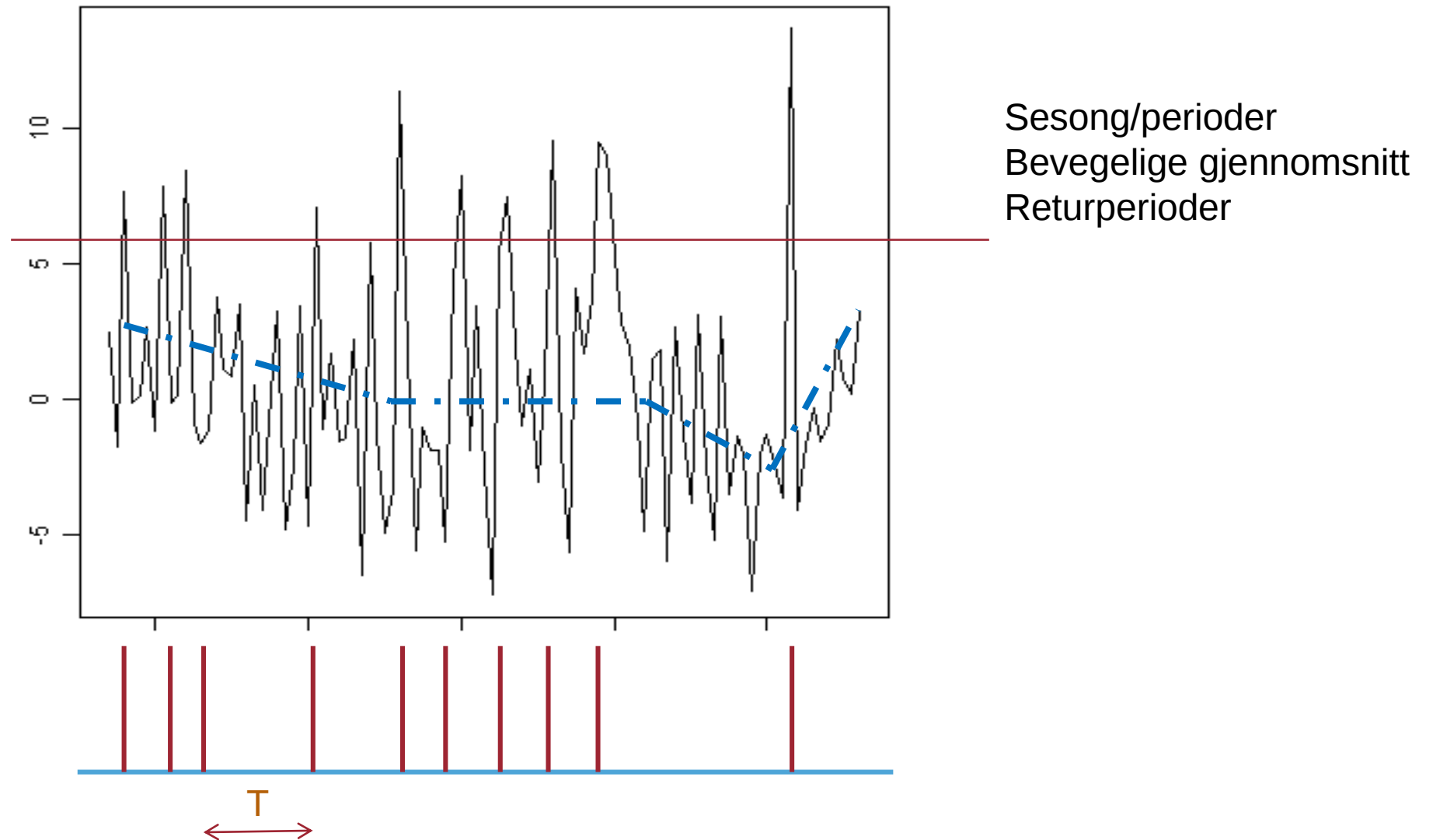
Måned, dag, time, bankkortnr, alder, bensin 98, bensin 95, diesel, pølse, bolle, kaffe, sjokolade, ukeblad, coca cola, vann, fanta, avis, baguette

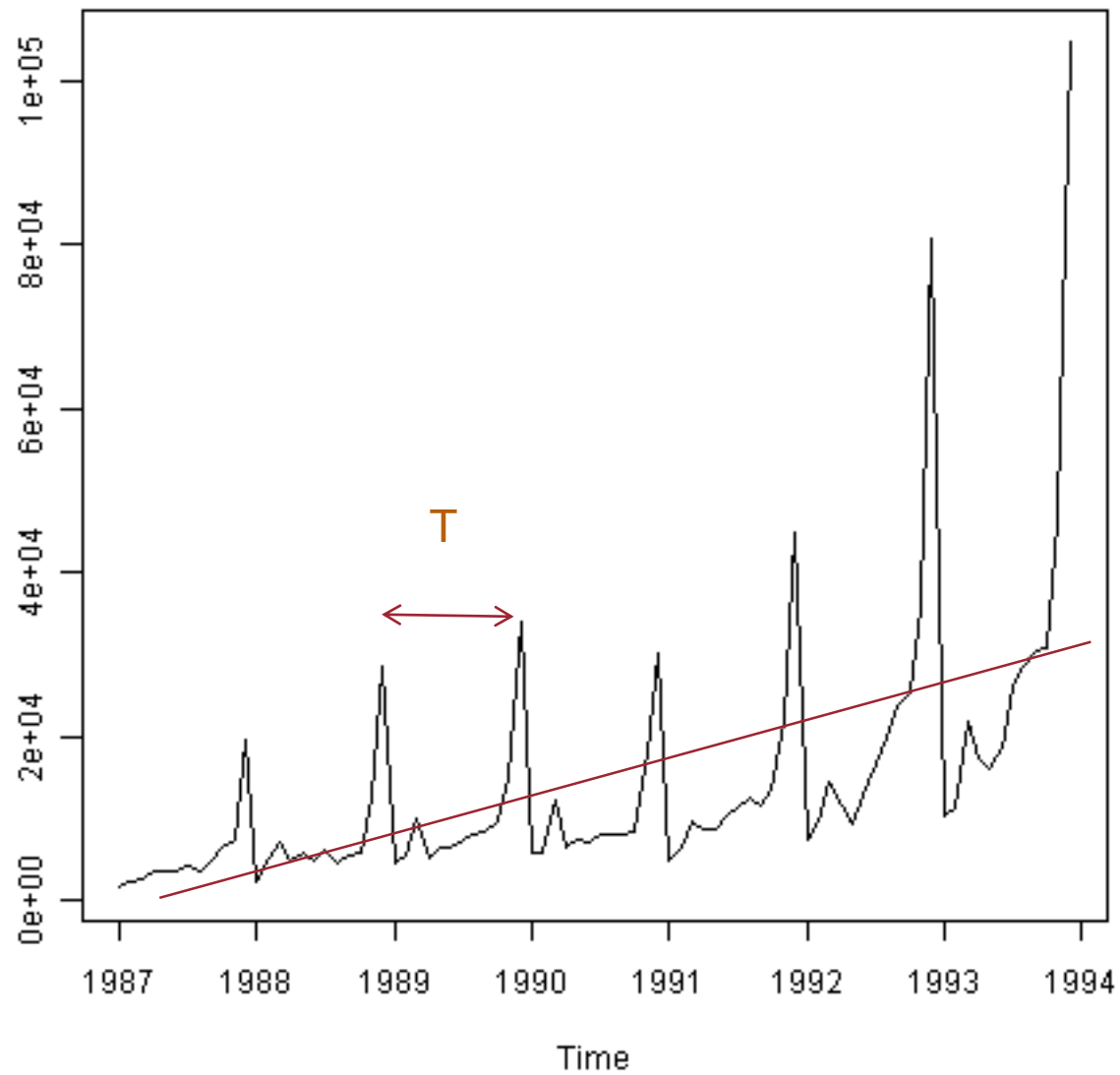
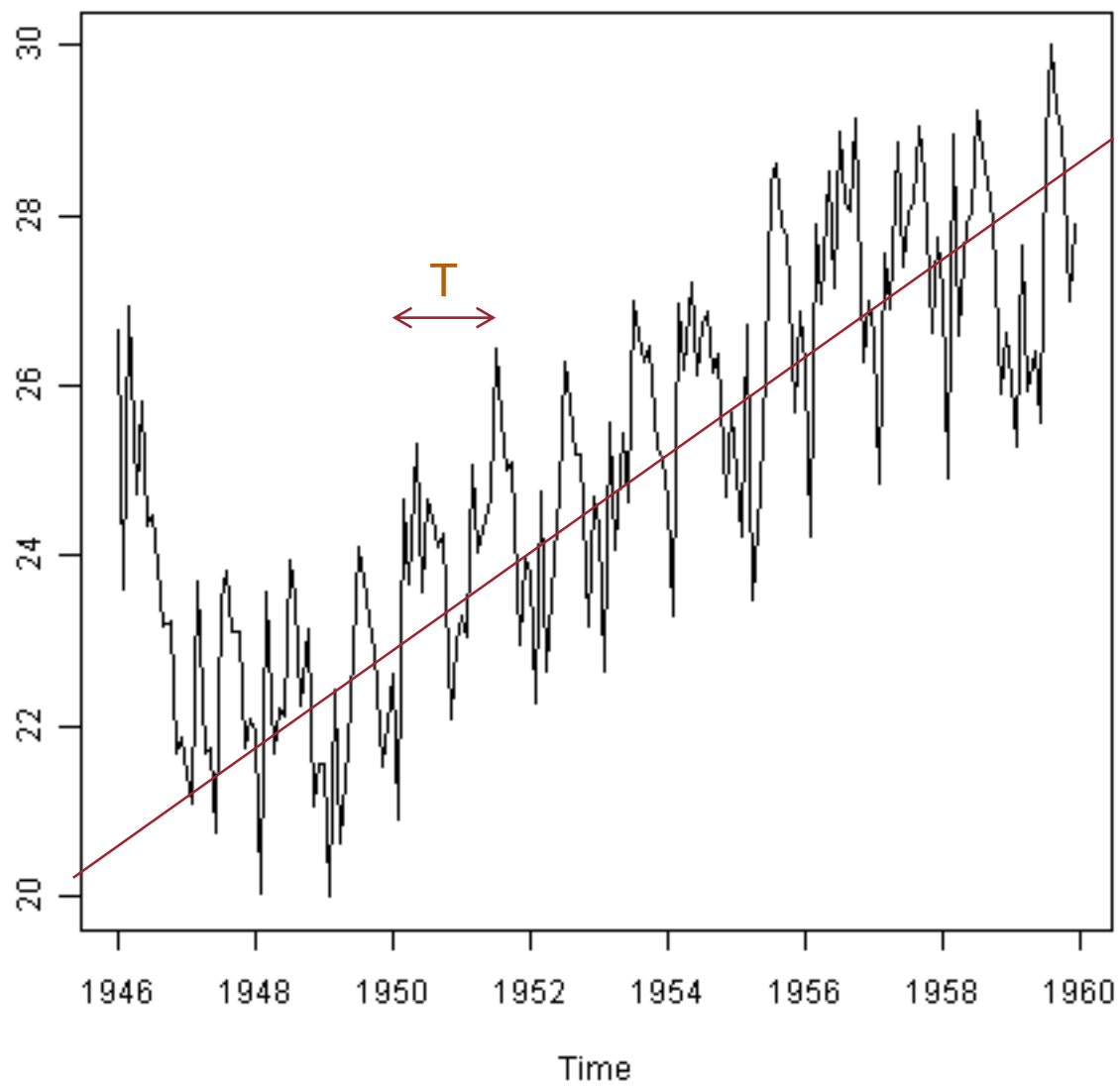
NÅR DET BEGYNNER Å BLI MANGE DIMENSJONER OG STORE MENGDER DATA....

- Multi-variabel statistikk
 - Multi-variable regressions
 - Principal Component Analysis
 - Singular Value Decomposition
 - ARMA/ARIMA
 - M.m.
- Operasjonsanalyse
 - Dynamisk programmering
 - Stokastisk programmering
 - Lineær programmering
 - M.m
- Maskinlæring
 - Nevrale nett
 - Genetiske algoritmer
 - ACO/Sverm intelligens
 - Kollaborativ filtrering
 - Reinforcement learning
 - M.m

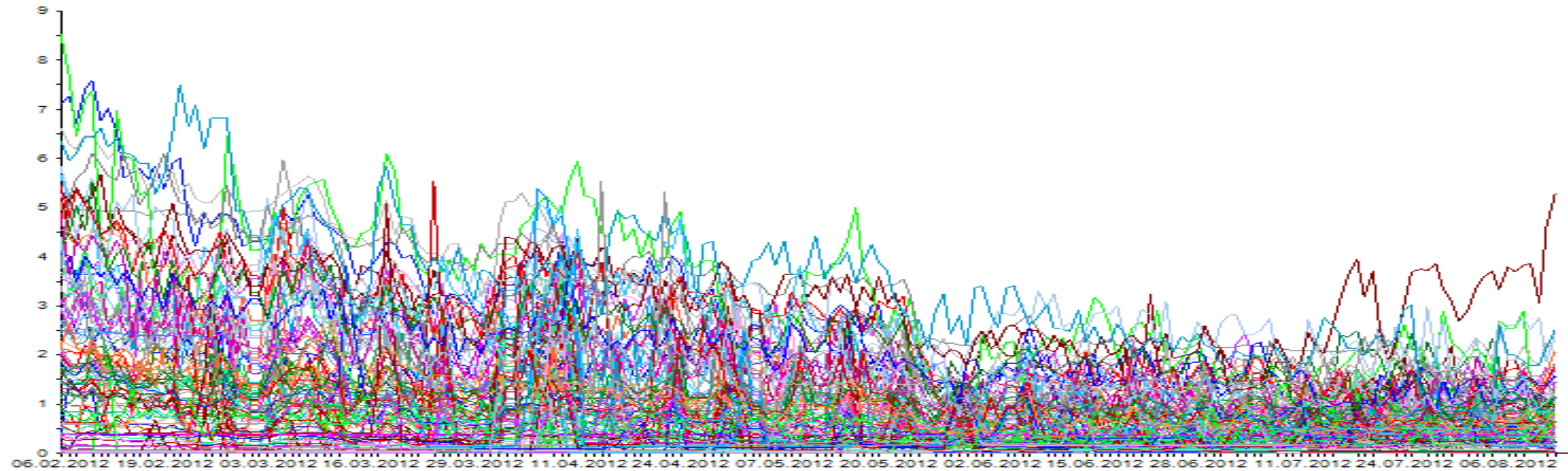
Ikke glem data warehousing/standard data mining

TIDSSERIER OG BRUK AV HISTORISKE DATA

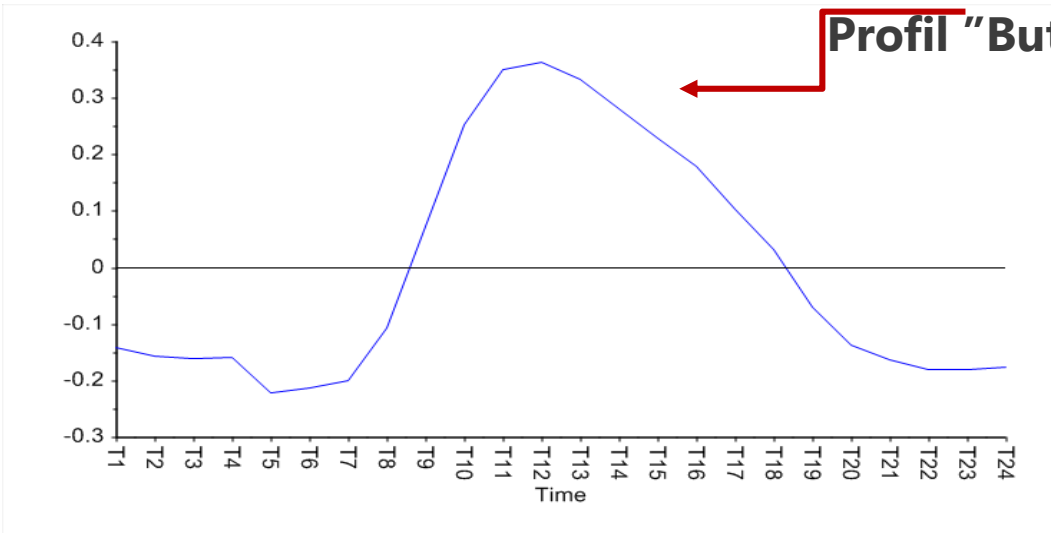
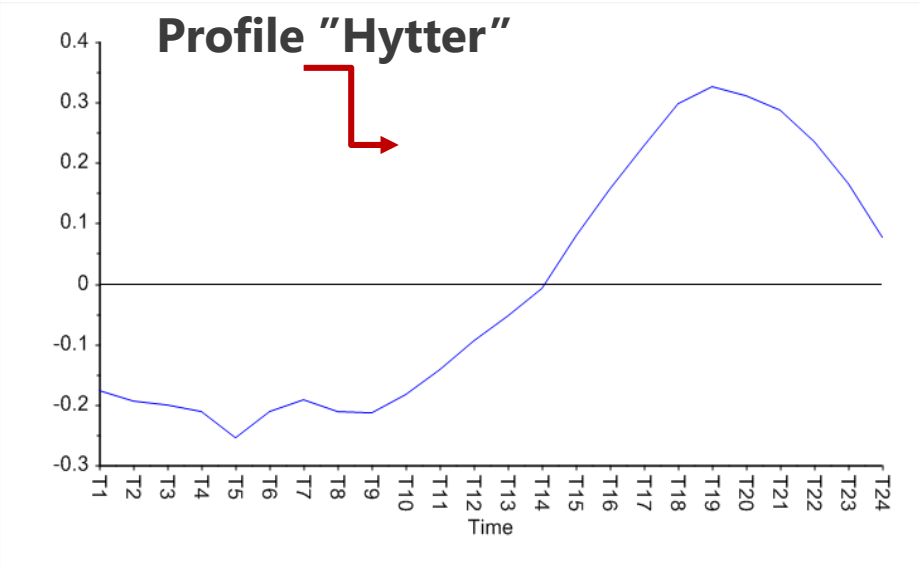
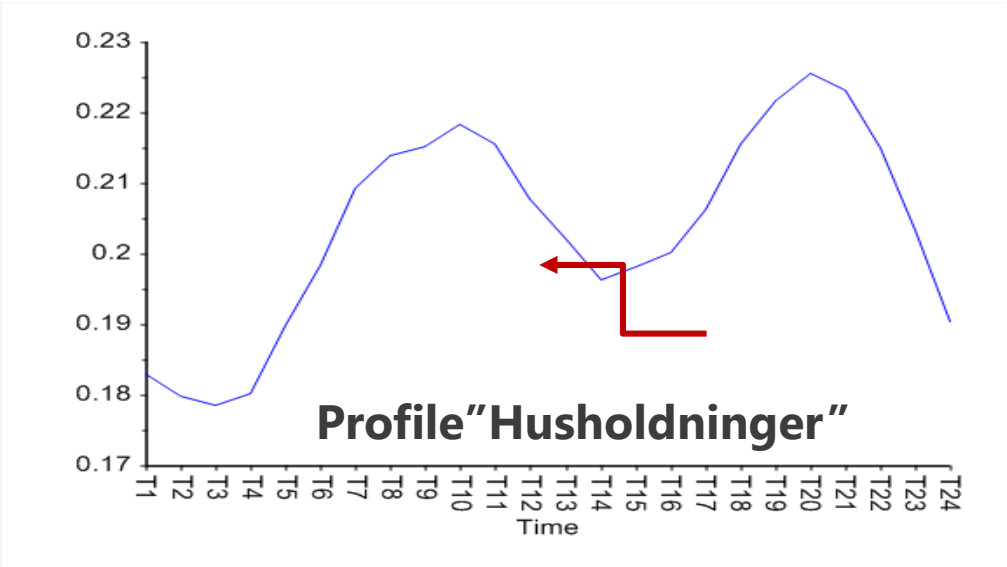




I VIRKELIGHETEN LIGNER OPPGAVENE MER PÅ DETTE



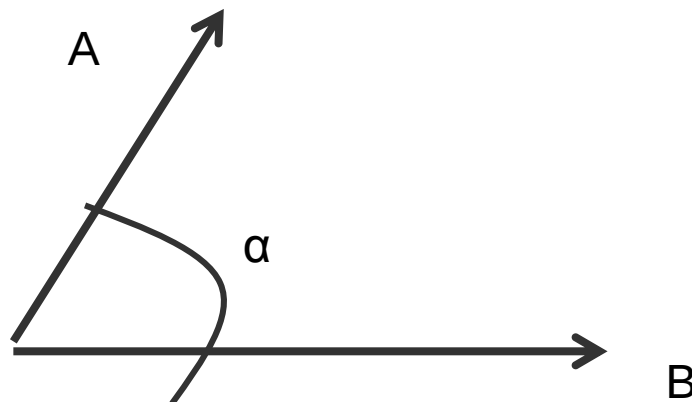
Måling av energiforbruk



SAMMENLIKNING AV KJØPSPROFILER

Kjøpshistorikk A: #(3 2 1 1 2 3 4 5 6 7 8 6 5 4 3 3 4 5 6 7 9 6 4 3)

Kjøpshistorikk B: #(1 1 1 1 1 1 2 3 7 8 9 9 9 8 9 9 6 5 4 3 2 2 1 1)



$$\cos \alpha = \frac{A \cdot B}{|A||B|}$$

ANBEFALINGER AV VIDEO, BØKER ETC

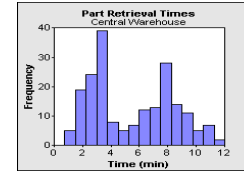
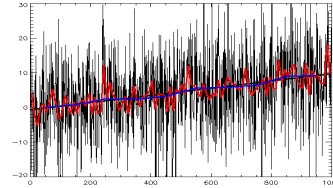
	Age	Sex	Interstellar	Lost River	The Gambler	Whiplash	John Wick	The theory of everything
Mary	23	F	3	3	5	3	3	1
Peter	43	M	4	5	3	4	2	3
John	34	M	5	6	4	6	3	2
Thomas	24	M	6	6	6	6	4	4
Susan	25	F	4	4	5	4	3	3
Jill	44	F	1	2	1	3	3	3
Tom	32	M	2	3	??	??	??	??

Vil Tom like «The theory of everything»?

Kjøpshistorikken til andre kunder kan benyttes til å forutsi hva en ukjent kunde kommer til å kjøpe!!!

- Sammenlign person og person
- Sammenlign sammenhenger mellom person og vare

TAMPERING WITH THE FUTURE - PREEMPTION



History of the homegenious group



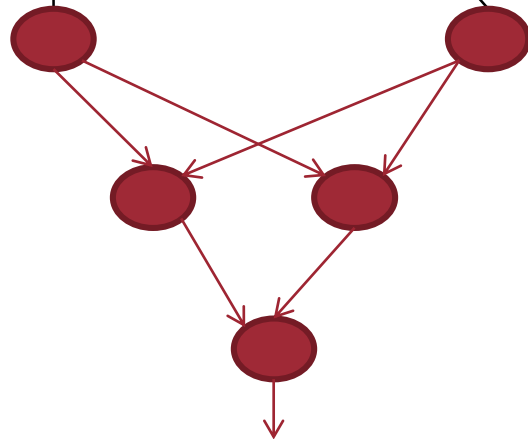
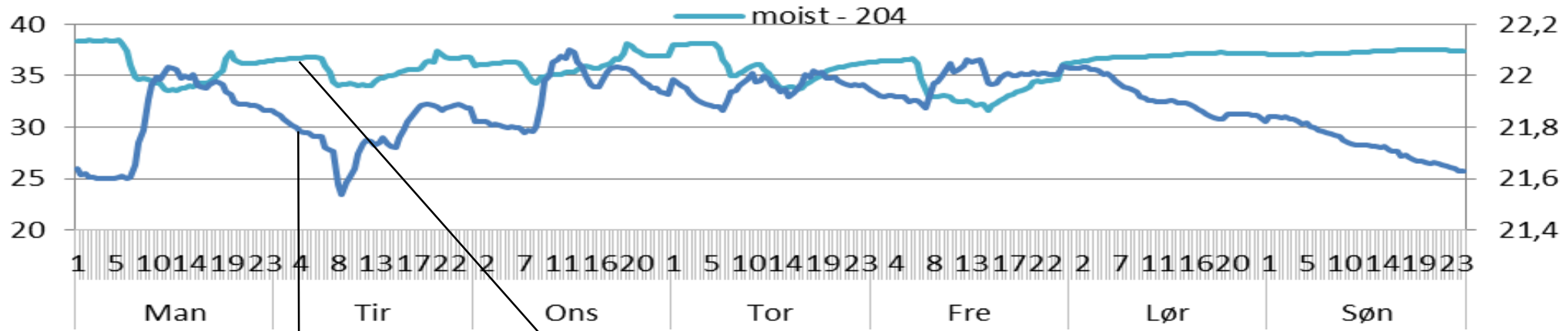
The new person arrives

- This person is very likely to have cancer in 15 years
- This person is very likely to be divorced in 5 years
- This woman is pregnant
- This person is a tacit opponent of our politics and is therefore not loyal

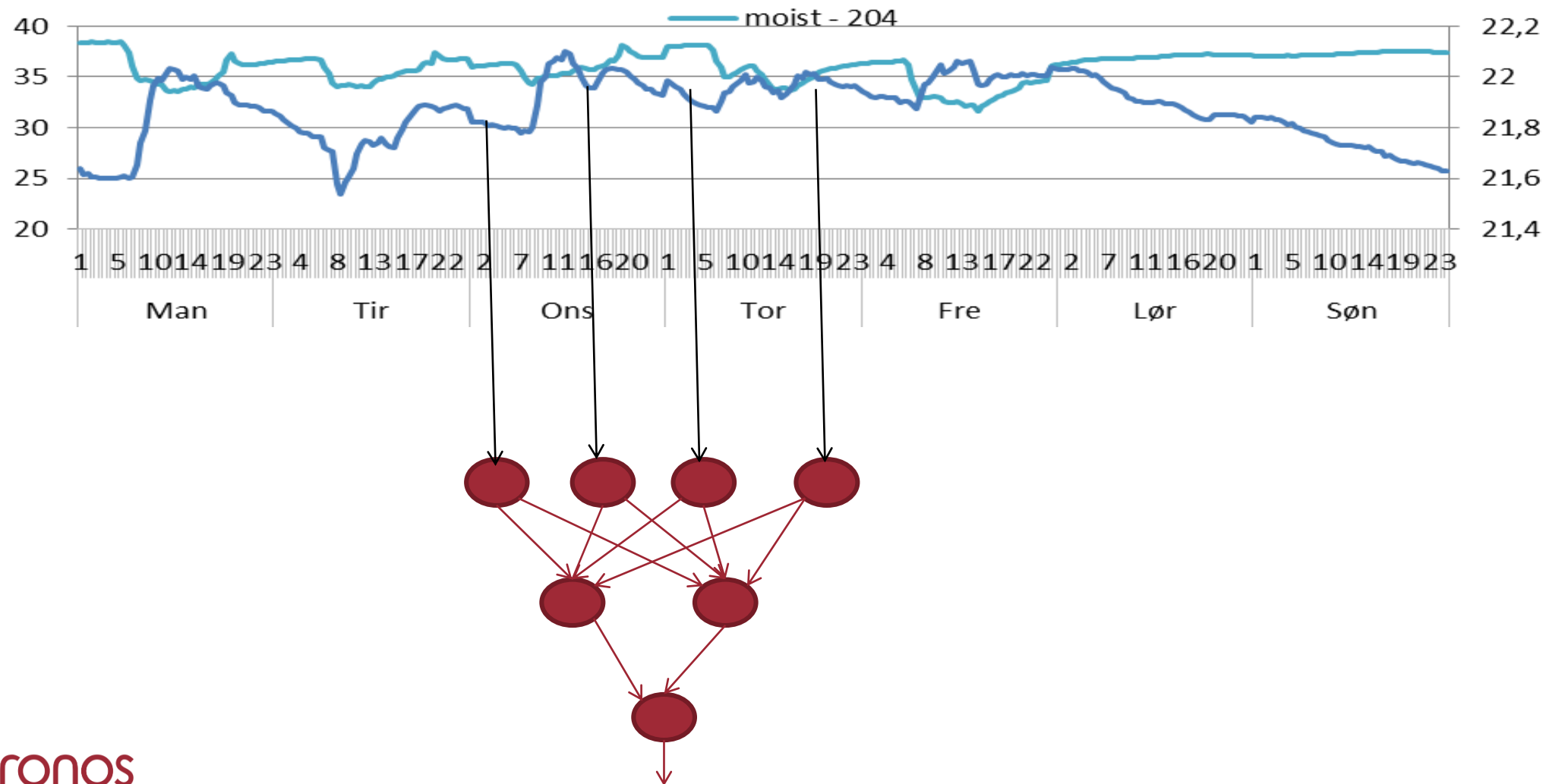
METODER SOM GJØR (NESTEN) ALT SELV

MASKINLÆRINGSMETODER F.EKS. NEVRALE NETT

FUSING AND ANALYZING TIME SERIES FOR CLASSIFICATION AND DETECTION



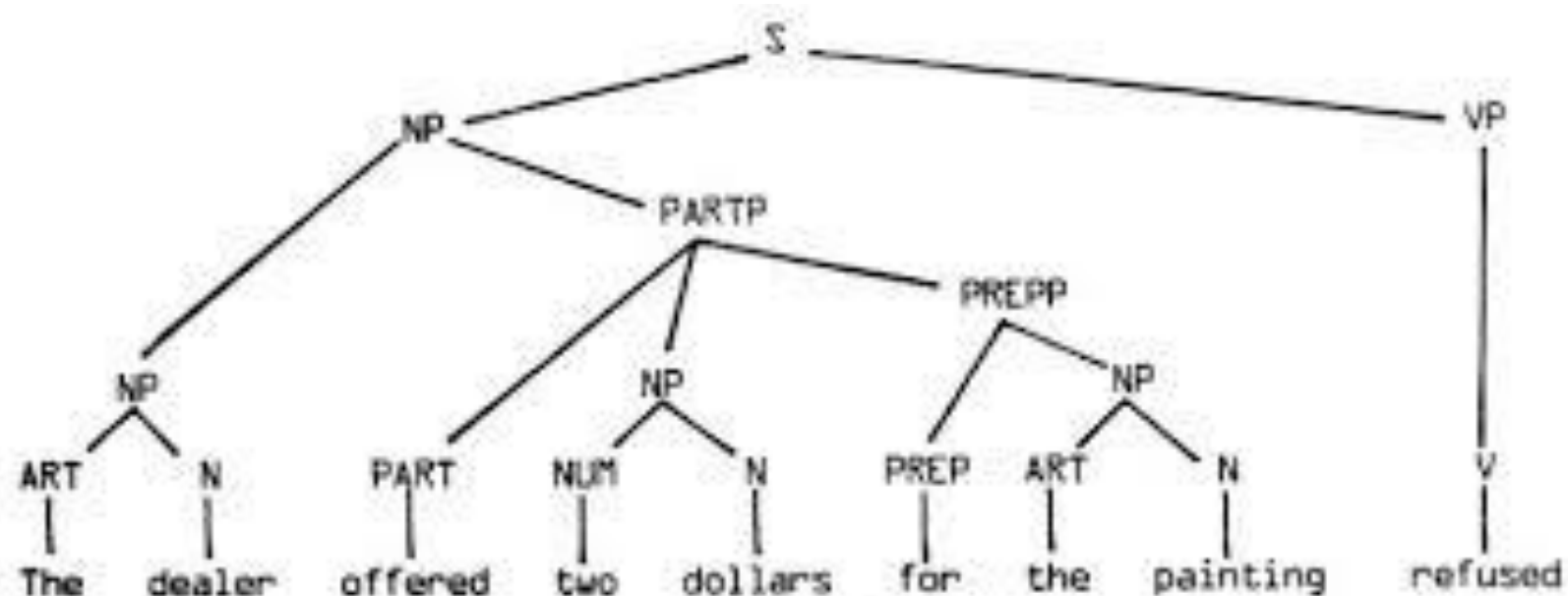
PREDIKSJON



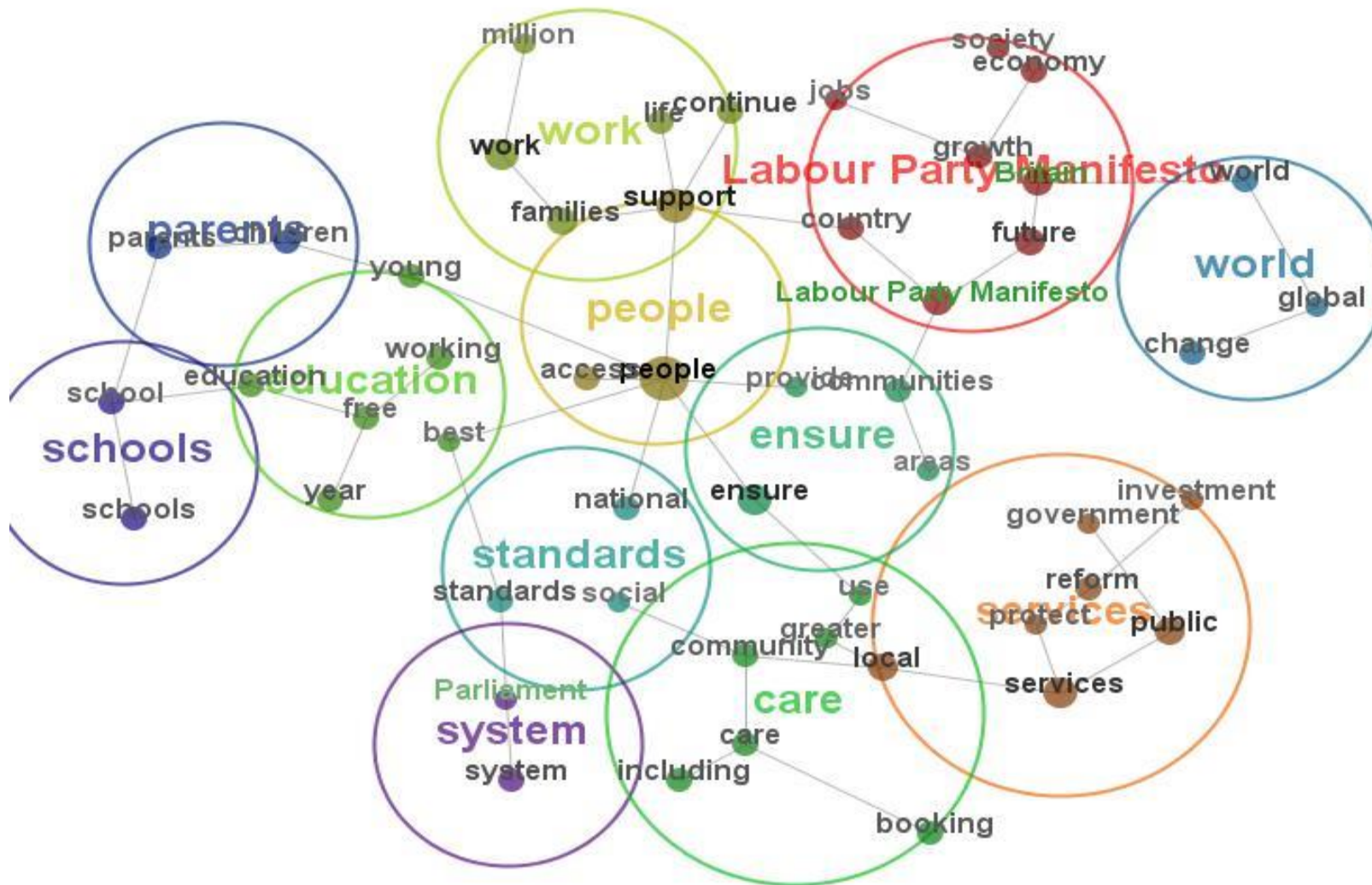
TEKSTANALYSE

- Strukturelle/syntaktiske metoder
- Semantiske metoder
- Statistiske metoder

STRUKTURELL – SVÆRT GRAMATIKKORIENTERT



SEMANTISK ANALYSE



The creation of ontologies and semantic networks

STATISTISK ANALYSE

Word Frequency

- Comparing the frequency of unigrams (single words), bigrams (two words) and trigrams (three words)

Rank	Counts	Word	Rank	Counts	Word	Rank	Counts	Word	Rank	Counts	Word
1	521411	/	31	22401	unplay	61	18254	second	81	96	tail
2	341367	/	32	21907	/	70	18084	public	82	95	affixing
3	236460	see	33	21601	death	71	18001	unplay	83	94	knife
4	232873	it	34	21114	parties	72	17999	will	84	93	zoological
5	226459	growth	35	21007	Italy	73	17930	also	85	92	home
6	171756	corruption	36	20757	children	74	17816	pink	86	91	area
7	161970	marion	37	20384	from	75	17646	/	87	90	air
8	162069	later	38	19331	phases	76	17639	map	88	89	hole
9	159886	no	39	19054	despite	77	17512	old	89	88	blot
10	156706	/	40	18754	phases	78	17400	growth	90	87	blot
11	151291	/	41	18671	also	79	17369	substantial	91	86	more
12	150619	/	42	17869	structure	80	17219	radio	92	85	gold
13	122846	/	43	17870	you	81	16995	consequence	93	84	soon
14	106626	corruption	44	17801	level	82	16762	also	94	83	folk
15	106017	locality	45	16756	being	83	16716	gas	95	82	ing
16	104647	lateral	46	14582	achievement	84	16519	books	96	81	corruption
17	103796	/	47	14084	no	85	16378	breeding	97	80	sanction
18	103123	from	48	14086	affairs	86	16223	discovery	98	79	will not
19	100960	new	49	13977	sign	87	16116	radio	99	78	is
20	17964	lateral	50	13541	/	88	15994	for	100	77	surface
21	100172	/	51	13460	another	89	15919	19-month-old			
22	100060	strong	52	13384	consequence	90	15834	radio			
23	111614	animals	53	12984	growth	91	15727	discovery			
24	106116	downside	54	11719	world	92	15616	one			
25	105177	/	55	11660	discovery	93	15572	good			
26	100790	behaviour	56	11569	more	94	15479	radio			
27	177164	but	57	11217	will	95	15386	parties			
28	104121	animal	58	11160	the	96	15311	household			
29	104172	it	59	10986	also	97	15211	discovery			
30	100960	death	60	10871	also	98	15164	radio			
31	100001	also	61	10779	parties	99	15011	also			
32	100001	also	62	10607	new	100	14911	discovery			
33	100001	also	63	10401	also						
34	100001	also	64	10206	but						

CABS 2014 - Kyoto, Japan

Google for  

[Nettet](#) [Bilder](#) [Google Maps](#) [Videoer](#) [Nyheter](#) [Mer ▾](#) [Innstillinger for søket](#)

Omtrent 25 270 000 000 resultater (0,37 sekunder)

I nyhetene



Herregården har kostet 240 mill - selges for en brøkdel
VG - For 9 timer siden
LONDON (VG) For tre år siden var prisen på den fornemme herregården i Skottland på 180 ...

[Motvind i vente for syklistene](#)
Aftenbladet.no - For 14 timer siden

[Grønt lys for nytt Valle Hovin](#)
Dagens Næringsliv - For 12 timer siden

[Flere nyheter for for](#)

For | Define For at Dictionary.com

[dictionary.reference.com/browse/for ▾](#) [Oversett denne siden](#)
intended to belong to, or be used in connection with: equipment **for** the army; . a closet **for** dishes. 3. suiting the purposes or needs of: medicine **for** the aged. 4. **For** the sake of - **For** a change - **For** real - **For** keeps

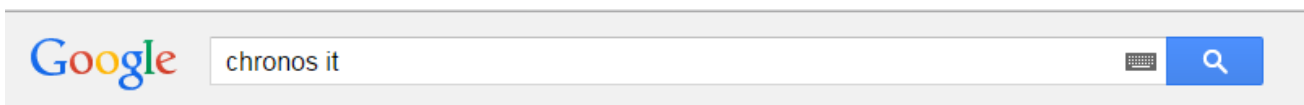
Værvarsel time for time for Oslo – yr.no

[www.yr.no/sted/Norge/Oslo/Oslo/Oslo/time_for_time.html ▾](#)
40+ elementer - Værvarsel **for** 9,3 millioner steder i verden!

Tid	Temp	Vind
onsdag kl 23	14°	Svak vind, 3 m/s fra vest-nordvest.
torsdag kl 0	13°	Svak vind, 2 m/s fra vest-nordvest.

Sverige: 13-åringer mistenkt for drap - Sverige - VG

[www.vg.no/nyheter/utenriks/sverige/sverige-13...for-drap/.../23467991/ ▾](#)
For 45 timer siden - To 13 år gamle gutter mistenkes for å ha drept en 66 år gammel



[Nettet](#) [Bilder](#) [Videoer](#) [Google Maps](#) [Nyheter](#) [Mer ▾](#) [Innstillinger for søket](#)

Omtrent 2 420 000 resultater (0,48 sekunder)

Chronos

www.chronosit.com/ ▾

Chronos leverer konsulenttjenester, løsninger og drift innen avansert informasjonsbehandling, datavitenskap, bigdata og databaser. Leveranseområder:.. Du har besøkt denne siden 3 ganger. Siste besøk: 29.10.14

Kontakt oss

Takk for at du tar kontakt med oss.
Vi vil ta kontakt så snart som ...

Bli med på laget

Trives du med å arbeide med BI, ...
Er du den rette - bli med på ...

Om oss

Chronos AS - Om oss - Homepage
for Chronos AS, a norwegian IT ...

Seminarer

Chronos arrangerer seminarer om
datavitenskap, big data og ...

Teamet

Bigdata, Datasience, maskinlæring,
statistikk ...

English

Chronos is a senior staffed IT-
company. We offer advisory ...

[Flere resultater fra chronosit.com »](#)

Chronos I.T. | LinkedIn

<https://www.linkedin.com/company/chronos-i-t-> ▾ [Oversett denne siden](#)

Learn about working at Chronos I.T.. Join LinkedIn today for free. See who you know at Chronos I.T., leverage your professional network, and get hired.

Cronos e-business integrator | Cronos

www.cronos.be/ ▾ [Oversett denne siden](#)

Focuses on Customer Relationship Management (CRM) and electronic business, building and supporting state-of-the art Frontoffice and business-to-business ...
[Company Profile](#) - [Contact Us](#) - [Route](#) - [Events](#)

EN LITEN KALKULATORJOBB

$$\textit{Relativ dokument frekvens} = \frac{2\,420\,000}{25\,270\,000\,000} = 0,0000957657$$

- Snur det hele på hodet og lager en logaritme (sorry!) = 4,0187900125
- Hvis vi finner et dokument med 4 forekomster av «chronos it» får vi en rangeringsvekt for dokumentet på 6,44
- Dersom vi hadde funnet et dokument som hadde 4 forekomster av «Obama» får vi en rangeringsvekt for dokumentet på mindre enn halvparten, nemlig 3,3 (siden det er relativt flere dokumenter som har ordet Obama i seg).
- Desto høyere vekt desto større tematisk innflytelse har ordet i dokumentet.
 - Mao dokumentet omhandler f.eks. begrepet «Chronos»

LEKSE

- **Analyser alle møtereferat i bedriften**
- **Hent et sett ord fra visjonsdokumentet: Strategi, verdier, mål etc.**
- **Hent et sett ord fra HMS planen: Sikkerhet, brann, avvik etc.**
- **Hent et sett ord fra økonomisk drift: Timer, inntekter, bunnlinje, topplinje**
- **Hva er den relative dokumentfrekvensen?**
- **Hvor mange møter har virkelig tatt for seg HMS?**
- **Hvilket møte var viktigst i forhold til visjons og misjonsarbeidet?**

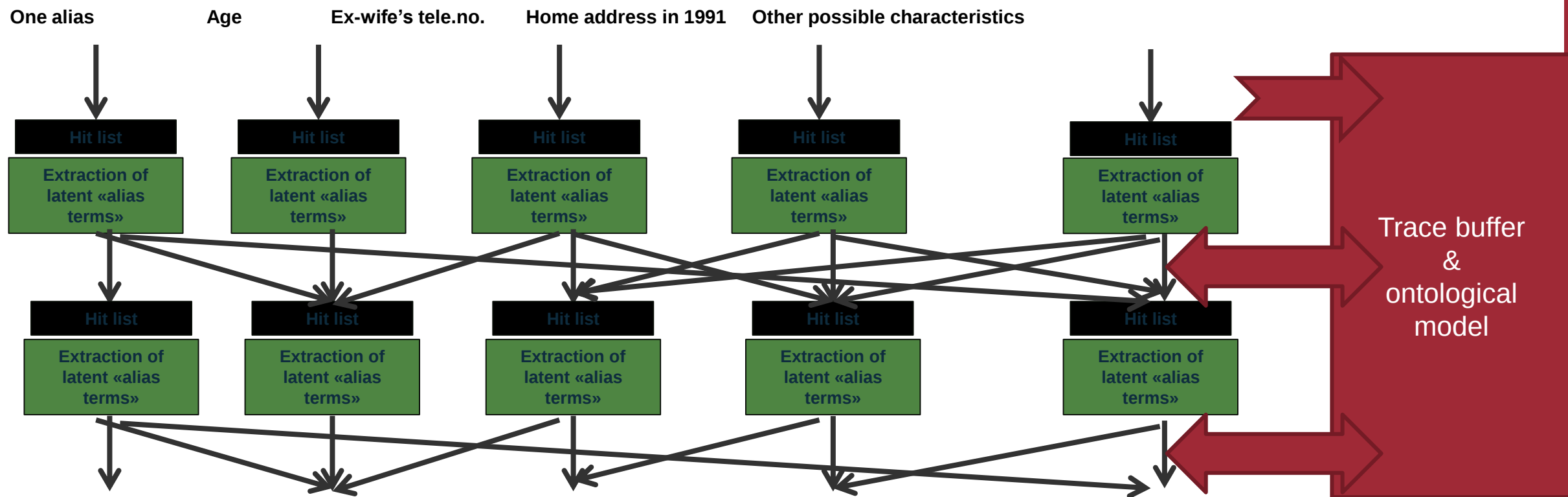
GJØR TILSVARENDE MED PRESSEOMTALER

- Finn fram en presseomtale som du virkelig liker. Kall den «Flott»
- Finn fram en presseomtale som du absolutt misliker. Kall den «Fæl»
- Luk ut alle ord som er felles for «Flott» og «Fæl»
- Gå gjennom alle presseomtaler
- Sammenlign alle ord i hver av dem med «Flott» og «Fæl». Bruk cosinus regelen.
- Hvor mange ligner på «Flott» og «Fæl»?
- Hvilken av de pressomtalenene som du har sett er verst og best?

EN LITEN HISTORIE TIL SLUTT



General fragments of an identity:



Continue until ontology is filled

TAKK OG VELKOMMEN TIL INNLEDENDE BIG DATA ANALYSEKURS I HØST

SLUTT

BIGDATA

PEARL DFDS SEAWAYS

+47 90 06 11 73 | BERNT.BREMDAL@CHRONOSIT.NO

DATO: 11.06.2015

WWW.CHRONOSIT.NO