BIG DATA
UTMANINGAR OCH MÖJLIGHETER

Surveyföreningen - Hantering av kvalitetsfrågor

Torsdagen den 23 februari 2017

Patrik Rydén

Institutionen för Matematik och Matematisk Statistik

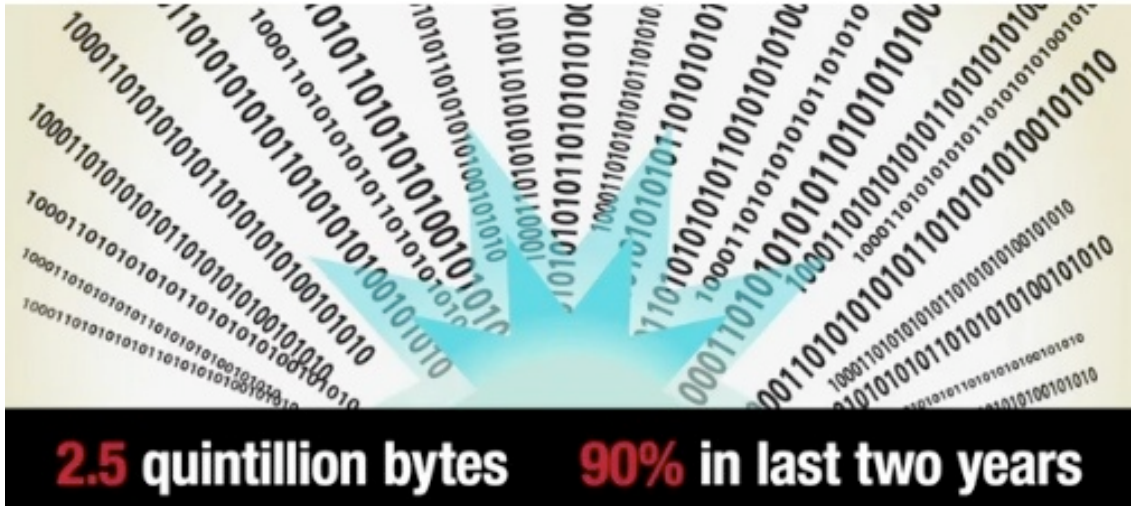
UMEÅ UNIVERSITET

What is all the fuss about Big Data?

UMEÅ UNIVERSITET

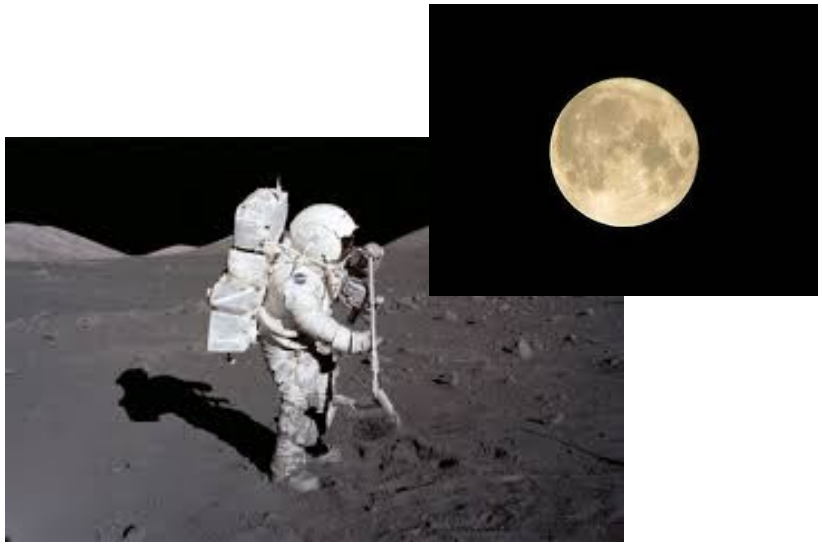
2012 genererade hela världen $2.5*10^{18}$ byte data

Ca. 90% av all världens data har genererats under de senaste 2 åren (IBM). ⟷
~ Mängden data som genereras fördubblas varje år.



UMEÅ UNIVERSITET

2023



2075

Vi har en massa data, men vi har bara börjat!

UMEÅ UNIVERSITET

$2.5*10^{18}$ BYTE DATA 2012

DNA: 6 miljarder bokstäver.

Vi har ca. 37 trillioner ($37*10^{18}$) celler.

Eller
$2.2*10^{29}$ bokstäver finns i våra kroppar.

Med 2012 års hastighet så kommer det att ta 222 miljoner år att sekvensera en människa.

- Vi generar allt mer data.
- Vi kan lagra allt mer data.
- Vi får allt snabbare datorer som klarar av att hantera allt mer data.

Men vi har bara börjat!

Data är inte information och inte kunskap

≠

Genetiska data: Tarmflora hos barn med allergi och ”friska” barn

OTU	PRB001_1	PRB001_1	PRB001_1	PRB001_N	PRB039_1	PRB039_1	PRB039_N	PRB062_1	PRB062_1	PRB062_N	PRB083_1	PRB083_1	PRB083_N	PRB084_1	PRB084_1	PRB084_N	PRB097_1	PRB097_1	PRB097_N	PRB098_1	PRB098_N	PRB108_1	PRB108_N	PRB112_1	PRB112_1	PRB112_N	PRB114_1	PRB114_1	PRB114_1	PRB114_N	PRB120_1	PRB120_1	PRB120_N	PRB126_1	PRB126_1	PRB126_1	PRB126_N	PRB129_1	PRB129_1	PRB129_1	PRB129_N	PRB131_1	PRB131_1	PRB131_N	PRB151_1	PRB151_1	PRB151_1
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0
OTU97_00	2	0	0	10	0	0	5	0	0	0	0	0	0	0	0	1	0	0	7	0	3	0	124	0	0	1	0	0	0	9	0	0	20	0	0	0	0	0	0	0	3	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	2	0	0	0	0	0	0	0	0	8	2	147	2028	763	0	0	0	0	1	1	0	0	0	26	193	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	1	0	0	1	2	0	16	0	1	0	63	142	0	0	0	7	0	0	1	0	0	0	0	3	5	0	0	0	0	4	0	0	0	3	0	1	0	0	1	0	6	2	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	21	223	3	0	60	7	0	0	12	0	1000	0	4	141	35	0	0	1	5	3	2	0	1	116	36	2	1	0	0	0	0	20	0	307	1104	2	398	27	598	7	11	0	1	3	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	531	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	1
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	1	0	0	0	0	1	0	0	0	0	2	3	0	9	0	0	0	0	0	3	34	0	0	0	0	1	2	23	4	0	0	0	2	0	0	2	2	1	2	10	0	0	0	4	0	1	
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	96	0	0	0	0	15	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	6	52	36	2	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	31	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU97_00	0	0	0	0	0	0	0	0	0	0	0	0	0

Utmaningen är att extrahera information från data

+

=

Allergiska barn har vid tidig ålder (1-4 veckor) en tarmflora som är mindre bakteriellt diverse än hos friska barn.

Mål: Ta fram en modell som gör det möjligt att givet tarmdata bestämma risken för att ett barn ska få allergi.

Sannolikheten att barnet utvecklar allergi

Exempel med statistisk analys av ”big data”

- Volvo lastvagnar – Automatiserat system för alarm och rotorsaksanalys. **Klassificering/Regression**
- Allokering av ambulansresurser. **Simulering**
- Arabidopsis (backtrav) – Hur “kommunicerar” gener. **Korrelation och nätverksanalys**
- Nya undergrupper av njurcancer. **Klusteranalys**

FIQA

Final Inspection and Quality Analysis

AB Volvo
- Volvo GTO UMEÅ Cab Competence Center

Umeå University
- Department of Mathematics and Mathematical Statistics
- Department of Applied Physics and Electronics
- UMIT Research Lab

Volvo Cars
- Department 81310 Manufacturing Engineering

Budget: 14.5 miljoner kr

Optimerad prehospitalsjukvård i Sverige

- Västerbottens Läns Landsting
- Region Norrbotten
- Region Västernorrland
- SOS-alarm
- Umeå Universitet

1 miljon larm/år

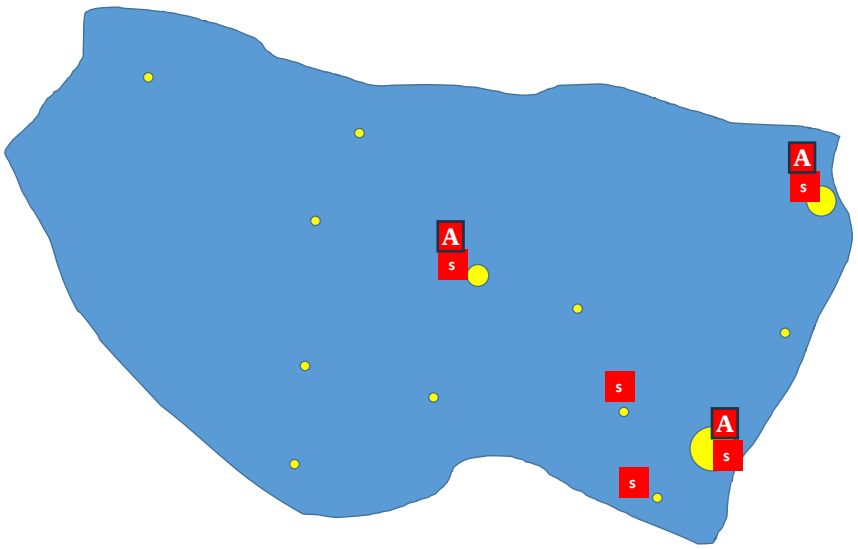
Mål: Utveckla ett verktyg för att placera ut resurser (ambulanser, stationer, mm) ”optimalt” inom en region.

Metod: Storskaliga datadrivna statistiska simuleringar.

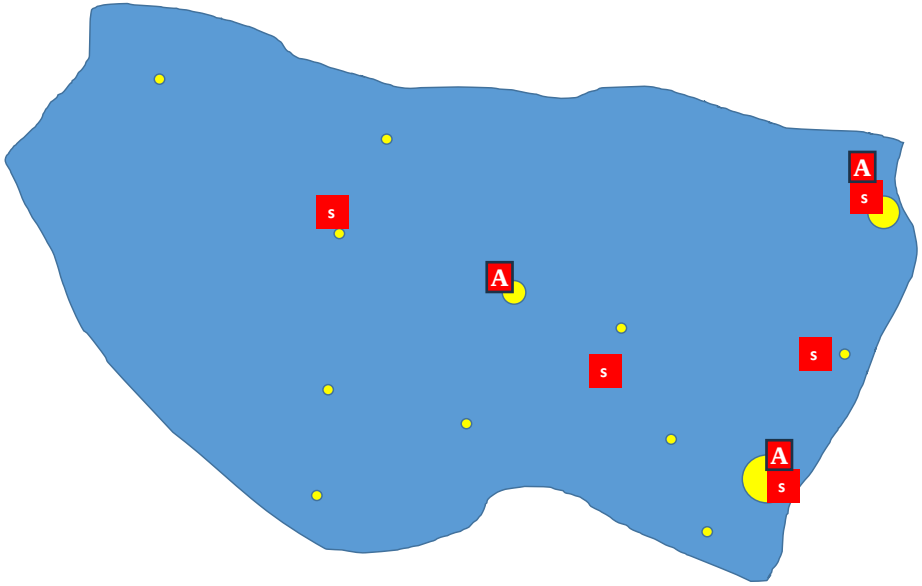
Principiellt exempel – Region med fyra tätorter (gul), tre akutsjukhus (A) och fem stationer (S), med 3 ambulanser/station.

Bilden visar 2 allokeringar av stationerna.
Hur kan vi avgöra vilken allokering som är bäst?

En möjlig allokering av stationer



En alternativ allokering av stationer



Vad är en bra allokering?

Flera möjliga kriterier och ett flerdimensionellt problem, men vi måste välja ett endimensionellt kriterium, men vi kan ha bivillkor.

Exempel:

Kriterium: Att ha så kort *median responstid* som möjligt.
Responstid=Tid från larm tills ambulans på plats.

Bivillkor: *median responstid < c inom alla delregioner.*

SOS-alarmdata – tid och datum för larm, position, vem tog larmet, körtid till patient, behandlingstid, körtid till sjukhus, kliniska data, mm.

Trafikdata, meteorologiska data, demografiska data mm.

Idé 1 – Simulera utryckningar dynamiskt baserat på historiska larmdata

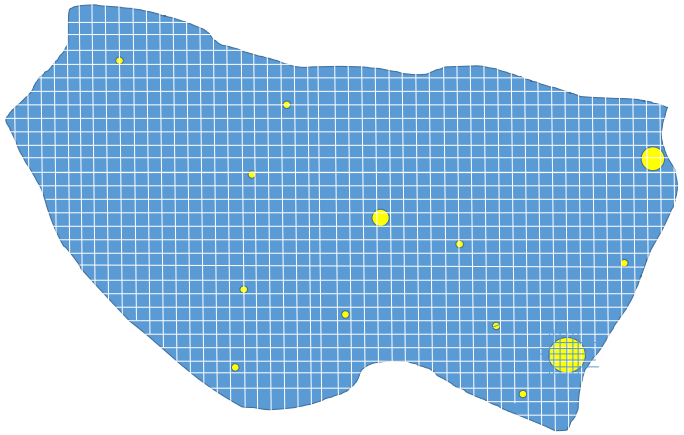
- *Position och tid för larm* (från historiska larmdata)
- *Bestämma vilken bil som får uppdraget* (**regel**, t.ex. närmaste)
- ***simulera körtid*** (**ger responstiden**)
- *Bestämma behandlingstid på plats* (historiska data)
- *Bestämma aktion* (leverans till sjukhus eller inte, historiska data)
- ***simulera transporttid*** till sjukhus, station eller alternativ plats.

Vi kan simulera data och beräkna t.ex. median responstider för olika allokeringar. Metoden kan användas för att t.ex. studera vad som händer när vi flyttar en station (Dorotea) eller lägger ned en akutavdelning (Sollefteå).

Idé 2 – Bygga en modell som kan simulera larmdata

Risken att få ett larm i ett *givet område och tidpunkt* beror av flera faktorer, t.ex.

- Hur många personer som det finns i området, samt deras demografi
- Tidpunkten (vilken vecka, vilken veckodag, vilken timme)
- Väderförhållanden



Dela in regionen i celler (storleken kan variera) och bygg en modell som skattar risken att få ett alarm i en cell vid en tidpunkt.

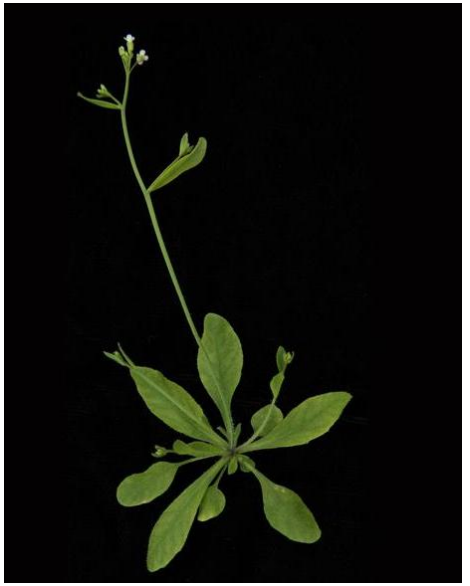
Skattningarna baseras på

- Historiska larmdata
- Demografiska data
- Meteorologiska data
- trafikdata

Svårt problem!

Med metoden kan vi simulera realistiska larmdata för framtida scenarier orsakade av t.ex. befolkningsökning, en åldrande befolkning och urbanisering.
Idé 1 + 2 gör det möjligt att planera verksamheten för framtida scenarier.

Hur kommunicerar gener i arabidopsis



Att förstå vilka gener som ”kommunicerar” med varandra är centralt för att förstå hur organismer fungerar. Vilket är viktigt för att utveckla t.ex.

- Nya läkemedel.
- Grönsaker som ”tål mörker” och kan transporteras lång tid.

Extern partner

Oliver Keech
(växtforskare Umu)

Data: genexpressionsdata från 887 arrayer (321 delförsök) med ca. 23000 gener/array (totalt ca. 20 miljoner observationer).

En stor del av alla publikt tillgängliga data från arabidopsis.

UMEÅ UNIVERSITET

Kempestiftelserna

Olika delförsök – olika förhållanden

Idé: Gener som kommunicerar med varandra borde vara ”starkt korrelerade”.

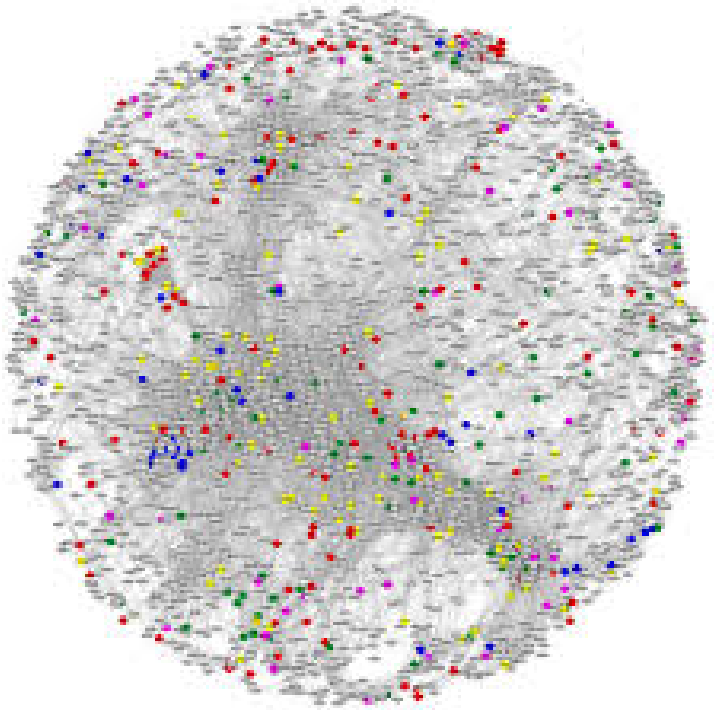
För vårt data har vi ett unikt partiellt facit så vi kan utvärdera våra nätverk.

UMEÅ UNIVERSITET

Kempestiftelserna

Korrelationsmatris

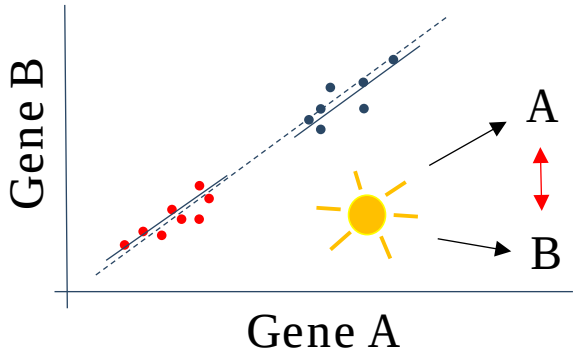
23000*23000
(ca. 250 miljoner obs.)



Hur ska vi beräkna korrelationen mellan två gener?

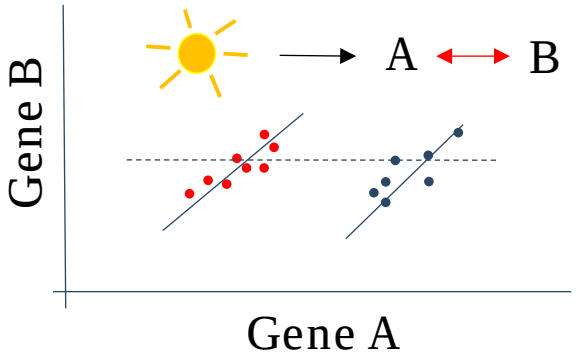
Kommunikation mellan gen x och y om cor(x,y)>c.

a



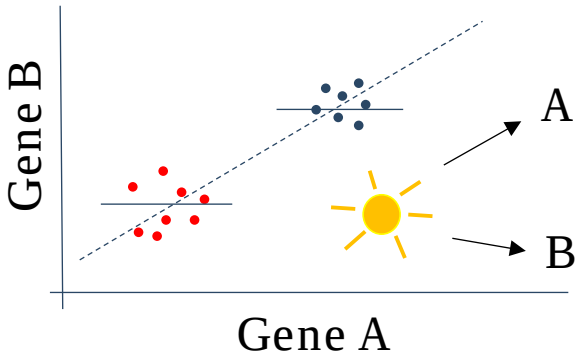
Sant positiv

b



Falsk negativ

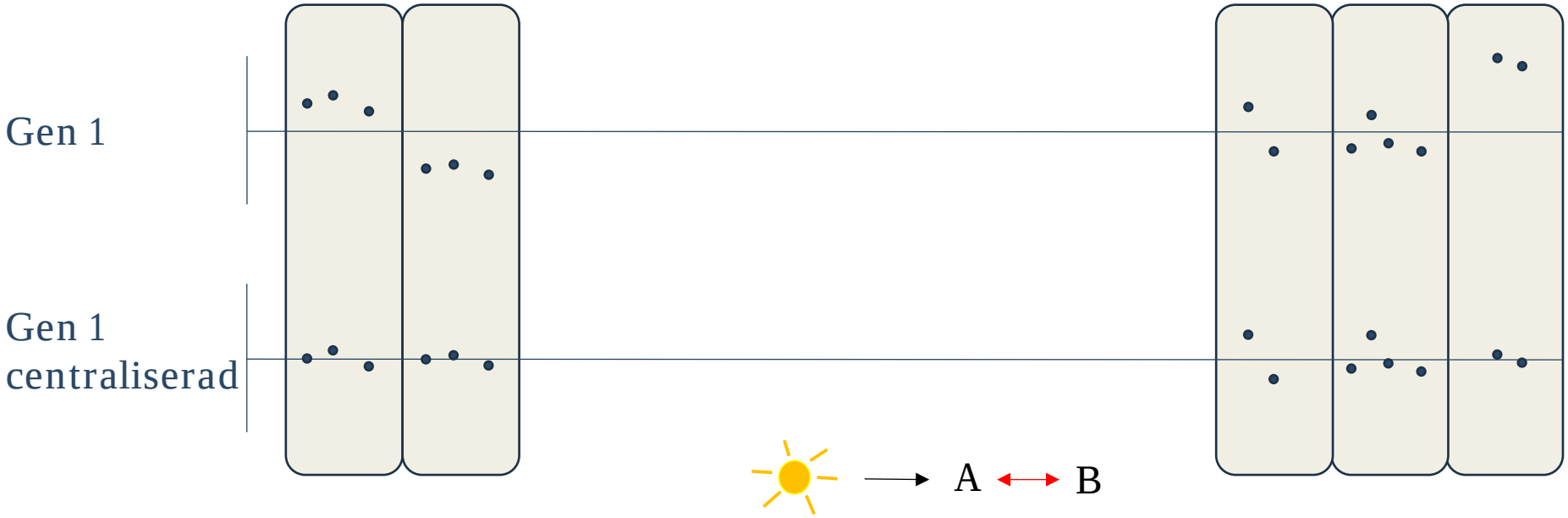
c

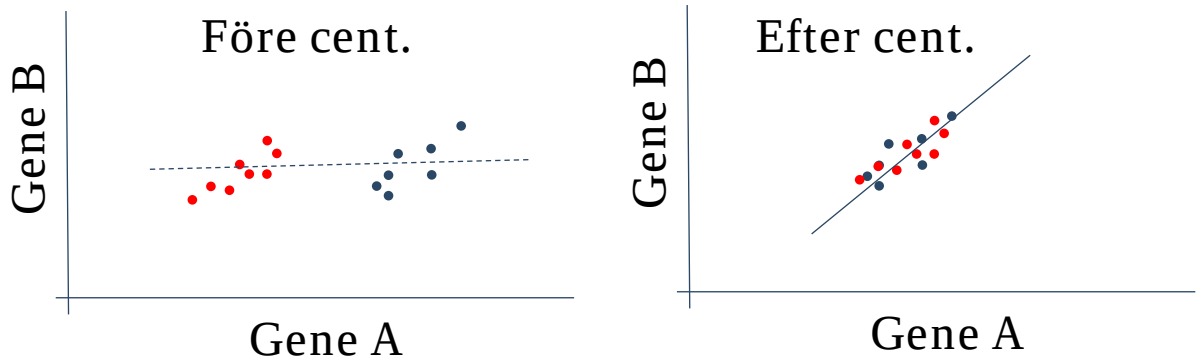


Falsk positiv

Röd: mörker
Blå: ljus
↔ sann kommunikation
→ påverkan av ljus
----- samband med ”vanlig” korrelation

Lösning – centralisera inom alla delförsök



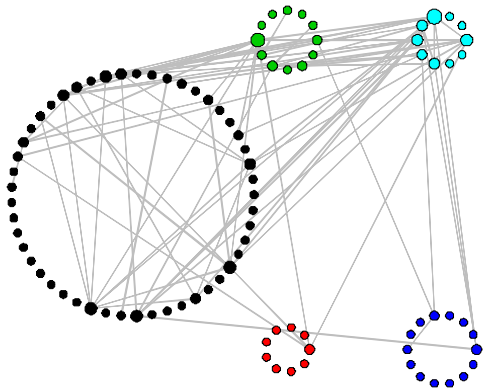


UMEÅ UNIVERSITET

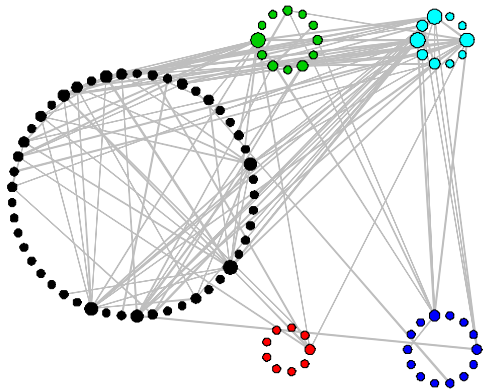
Kempestiftelserna

Känd hubb där vi vet att det finns mycket kommunikation – ju fler linjer ju bättre skattning!

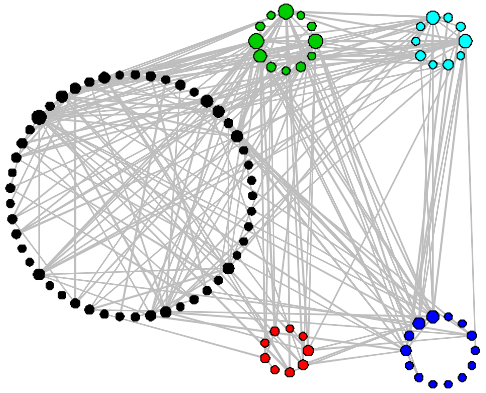
Pearson correlation



Partial correlation

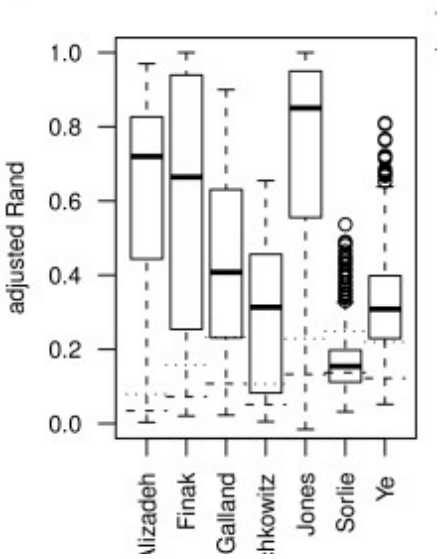


Trial centralized Pearson correlation



Directed Cluster Analysis (DCA) - en ny metod för att hitta nya undergrupper av cancer

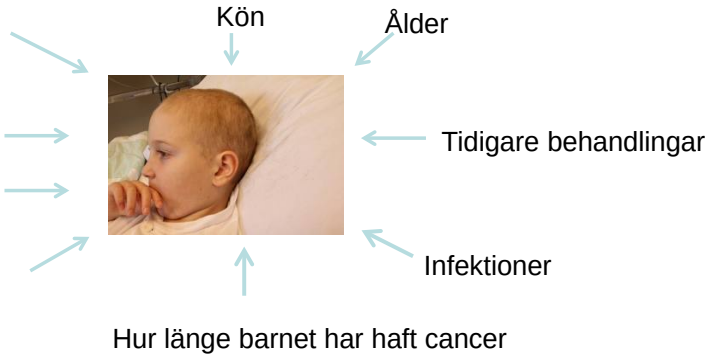
- En cancer kan ha flera okända underklasser, t.ex. klass A och B.
- Intressant att identifiera dessa och undersöka hur effektiva olika behandlingsformer är för respektive klass.
- Hos varje patient observerar vi ett stort antal variabler, t.ex. genuttrycket och metyleringsgrad hos alla våra gener, totalt nästan 1 miljon variabler per patient.
- Vi vill använda klusteranalys för att gruppera individerna så att patienter med klass A cancer hamnar i en grupp och patienter med klass B cancer hamnar i en annan grupp. ***Ett svårt problem!***



Primär faktor: Typ av cancer

Varför misslyckas vi?

- Genexpressionsdata (25 000 gener) från 7 cancerförsök – alla med två klasser (som vi inte låtsas om, men använder för utvärdering)
- Alla data analyserades med 2780 olika ”klusteranalysmetoder” och utvärderades med adjusted Rand (1 = perfekt, 0 som slumpen).



Kempestiftelserna

Vetenskapsrådet

UMEÅ UNIVERSITET

Big Data är på riktigt!

Utmaningen är inte att generera data utan att förädla data till information.

Mer resurser behöver läggas på dataanalys.

Statistiker har rätt bakgrund, men fler statistiker behövs och statistiker behöver bli bättre på att arbeta inom tillämpningar med big data.

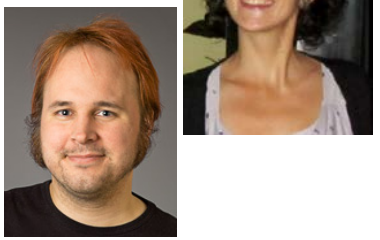
- *Utbildningen av statistiker behöver modifieras.*
- *Mer fokus på analys av storskaliga data.*
- *Mer tvärvetenskapligt angreppssätt.*

Det är en spännande och kul tid att vara statistiker!

Statistiker är uppskattade och efterfrågade!

Vi får vara med på alla möjliga projekt!

Vi får vara med om allt möjligt!



Tack för visat intresse!

Patrik.ryden@umu.se

UMEÅ UNIVERSITET