

# Klyngeanalyse

ISTx1003 Statistisk læring og Data Science

Stefanie Muff, Institutt for matematiske fag

November 12, 2021

# Anerkjennelse

Disse slides bygger på slides fra Mette Langaas, 2020.

Takk til Mette for at jeg fikk bruke noen av materialene.

# Plan for i dag (tema “Klyngeanalyse”)

- Hva er klyngeanalyse
- Læringsmål, pensum og læringsressurser
- Avstandsmål
- K-gjennomsnitt (“K-means”) klyngeanalyse
- Bruk av klyngeanalyse på et bilde (prosjektet fra i fjor)
- Hierarkisk klyngeanalyse
- Informasjon om prosjektet

## Eksempel 1: Genaktivitet

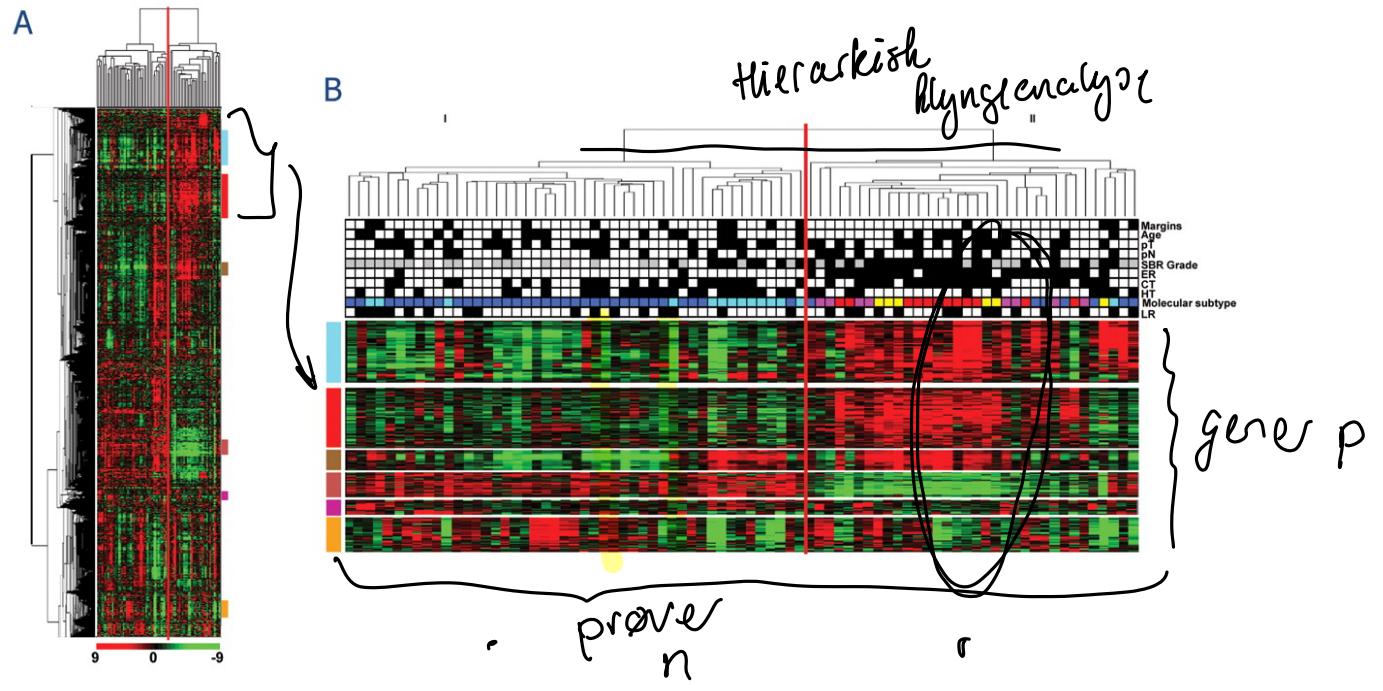
- $n = 81$  celleprøver fra kreftsvulster til ulike pasienter
- Genaktivitet for  $p = 12957$  gener

**Spørsmål:**

Hvilke celleprøver fra brystkreftpasienter ligner hverandre mest?

Kan vi finne ukjente klynger (av celleprøver) i dataene?

Dette kan hjelpe for å forutsi sannsynligheten for en tilbakefall.

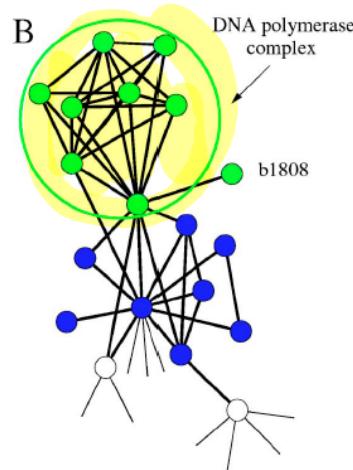
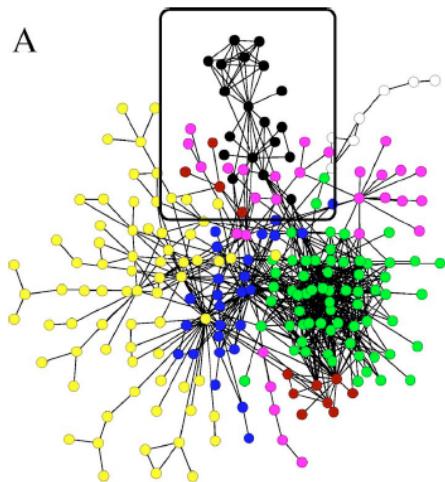


$$X = p \times n = \text{gener} \times \text{prøver} .$$

## Eksempel 2: Proteininteraksjonsnettwerk

Kan vi finne klynger med relatert funksjon?

MUFF, RAO, AND CAFLISCH



PHYSICAL REVIEW E 72, 056107 (2005)

FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) *LQ* clusterization of the black *Q* cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to *LQ* while the complete *Q* cluster is heterogeneous.

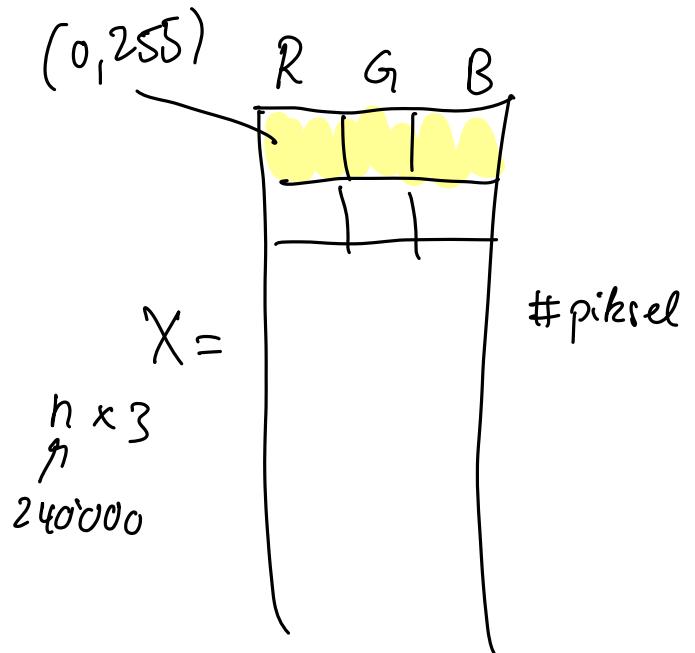
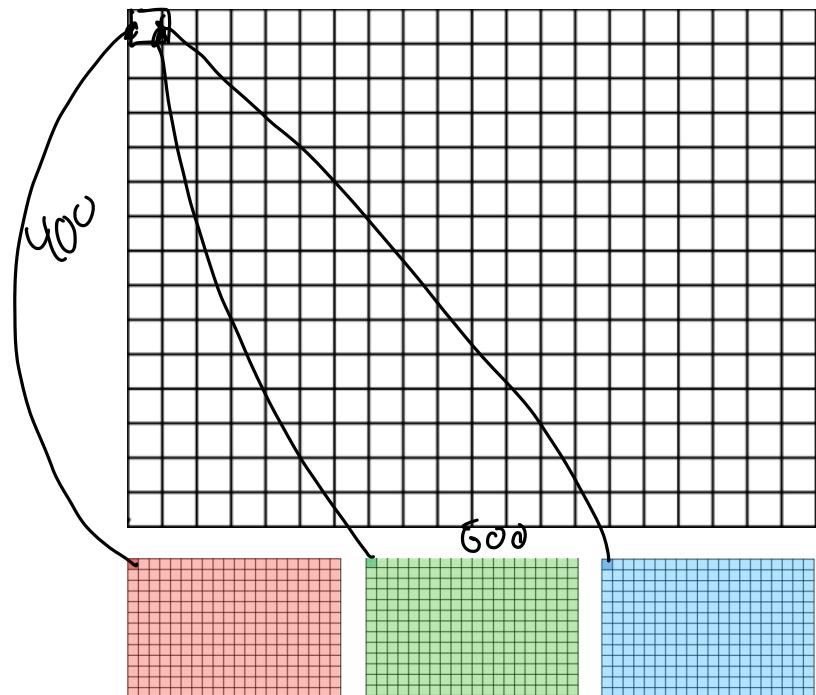
## Eksempel 3: Bildanalyse

Det var en prosjektoppgave i fjor.

**Mål:**

Å bruke klyngeanalyse til å fjerne detaljer og støy - ved å dele pikslene inn i to eller flere klynger.

Hver piksel har en farge som er definert som en blandig av rød, grønn og blå ( $x, y, z$ ):



Klyngeanalyse: Finn  $k$  "typiske farger" som representerer klynger (sentroider) og erstatt hver piksel med sentroidfargen.

Jeg var litt nysgjerrig...

Opprinnelig bilde



Jeg var litt nysgjerrig...

Opprinnelig bilde



Bilde i svart/hvitt



Jeg var litt nysgjerrig...

Opprinnelig bilde



Bilde i svart/hvitt



bilde med 8 klynger



# Læringsmål

- Forstå hvorfor det er interessant å gjøre klyngeanalyse
- Kjenne igjen situasjoner der klyngeanalyse vil være en aktuell metode å bruke
- Kjenne begrepene avstandsmål, koblingstype, dendrogram
- Forstå algoritmen for å utføre K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse *2 metoder du kan etter i dag!*
- Forstå hvordan klyngeanalyse utføres i Python
- Kunne besvare oppgave 3 av prosjektoppgaven på en god måte!

# Læringsressurser

Tema Klyngeanalyse:

- **Kompendium:** Klyngeanalyse (pdf og html, by Mette Langaas)
- **Korte videoer:** (by Mette Langaas)
  - Klyngeanalyse (8:43 min)
  - Hierarkisk klyngeanalyse (11:26 min)
  - K-gjennomsnitt-klyngeanalyse (8:38 min)
- Denne forelesningen
- **Disse slides** med notater



Som alltid se her:

<https://wiki.math.ntnu.no/istx1003/2021h/start>



# Klyngeanalyse – hva er det?

Vi har data

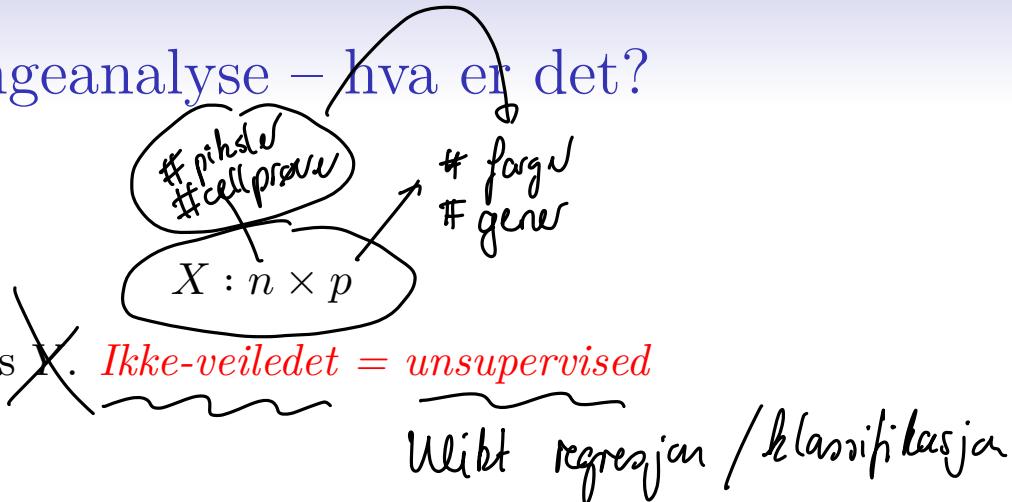
men ikke noen respons

Mål:

- Finn ukjente klynger i dataene.
- Observasjoner innen hver klynge er mer lik hverandre enn observasjoner fra ulike klynger.

Hva skal vi bruke resultatene fra klyngeanalysen til?

- Bildet: Fjerne støy eller, spare lagringsplass
- Medisin: Finne subgrupper av en sykdom → relevant for behandling?



# Klyngeanalyse – hva er det?

**Generelt:** Finne *struktur* i dataene.

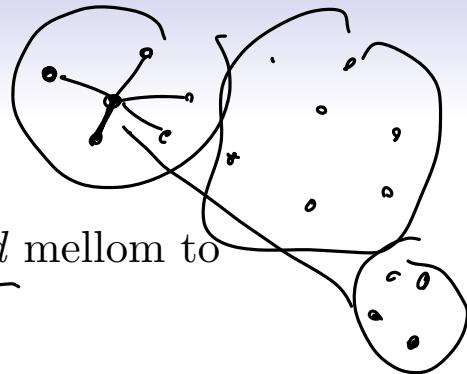


Kan vi stole på resultatene? Hvor robuste er de?

→ Fortsatt et forskningsområde!



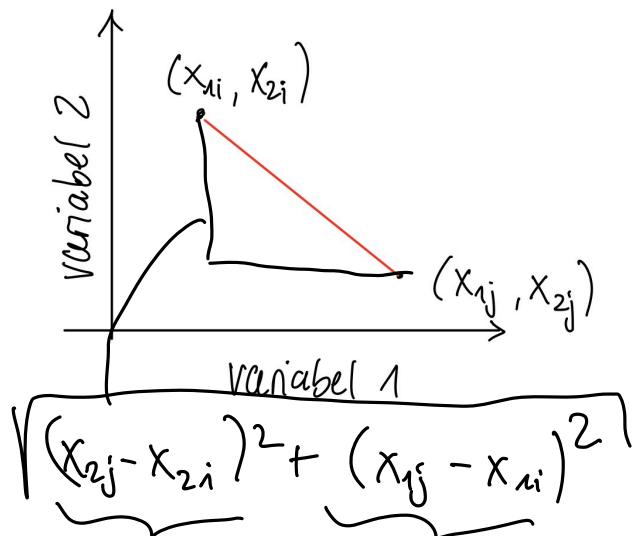
# Avstandsmål



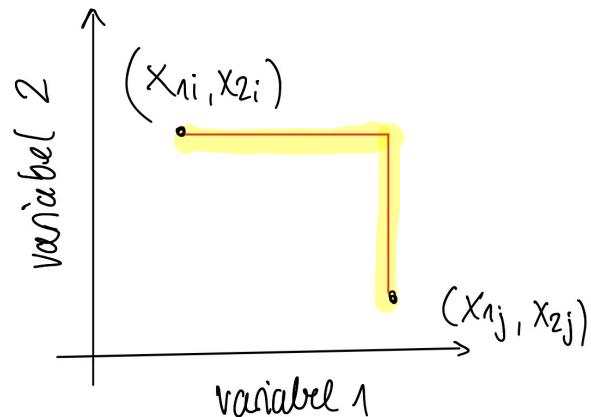
Før en klyngeanalyse må vi først definere en *avstand* mellom to datapoeng.

To populære avstandsmål:

Euklidisk



**City-block (=Manhattan)**



Euklidisk

*p dimensjoner  
= antall variable*

$$D_E(i, i') = \sqrt{\sum_{j=1}^p (x_{ji} - x_{ji'})^2}$$

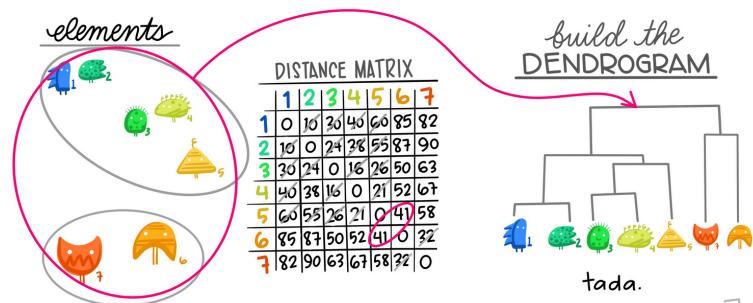
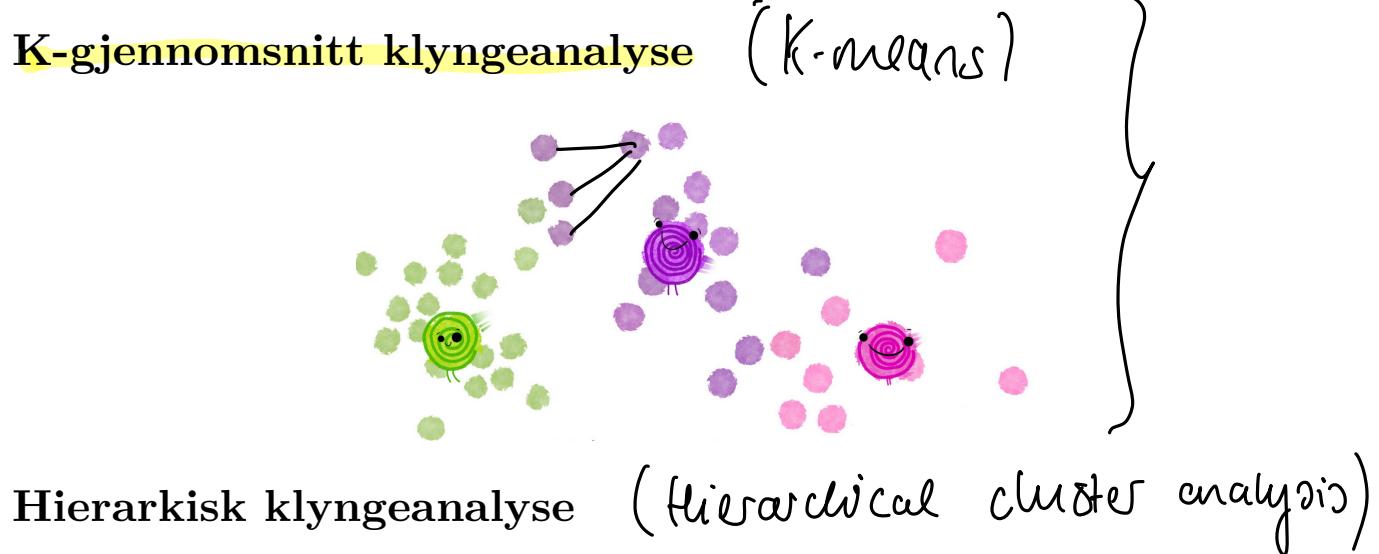
City-block (=Manhattan)

$$D_M(i, i') = \sum_{j=1}^p |x_{ji} - x_{ji'}|$$

Avstandsmål i mer enn 2 dimensioner: Enkelt å regne, men litt vanskelig å forestille seg.

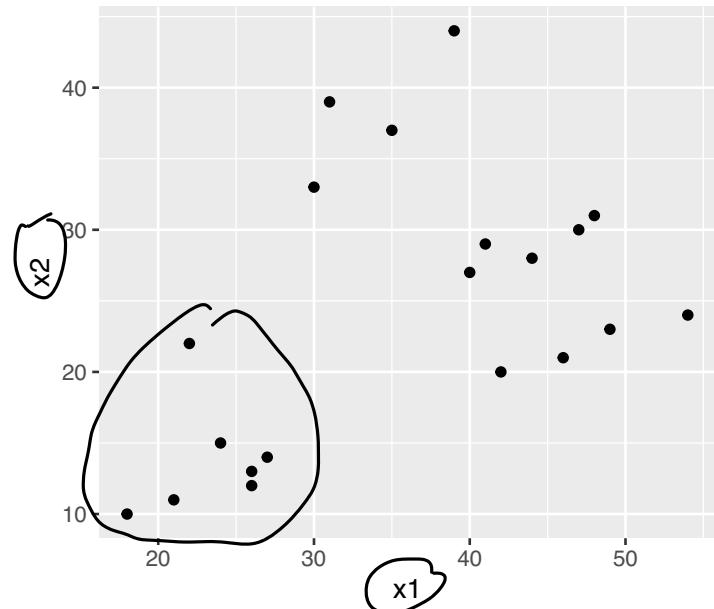
# Metoder for klyngeanalyse

Det finnes ganske mange metoder, men vi ser på to som er (mest?) populær:



## K-gjennomsnitt klyngeanalyse

- Finn  $K$  ukjente klynger i dataene.



→ unrealistisk  
med bare  $p=2$

- Alle observasjoner skal være medlem i akkurat én klynge.
- Variasjonen innen hver klynge skal være så liten som mulig.

matematiske formell'

## Variasjon innen en klynge $k$

- $K$  klynger  $C_1, \dots, C_k, \dots, C_K$ .
- Antall observasjoner i klynge  $k$ :  $|C_k|$ .
- Variasjon in klynge  $k$ :

$C_1, C_2, C_3 \leftarrow \{ \text{obs. i klynge } 3 \}$   
 $|C_k| = \text{antall obs. i klynge } k$

for klynge  $k$

$$\left\{ \frac{1}{|C_k|} \sum_{i, i' \in C_k} \underbrace{\sum_{j=1}^p}_{\text{summere over alle avstande}} (x_{ij} - x_{i'j})^2 \right.$$

kvadrert avstand  
 mellom obs.  $i, i'$

## Optimizeringsproblem

Vi vil minimere variasjon over alle klynger:

$$\sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

↓ summerer over  $K$  klynger

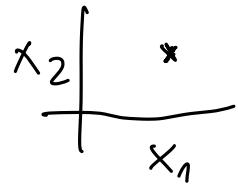
T.lks.  $n = 100$  obs.  
 $k = 3$  klynger

$$3^{100} \approx 5 \cdot 10^{45}$$

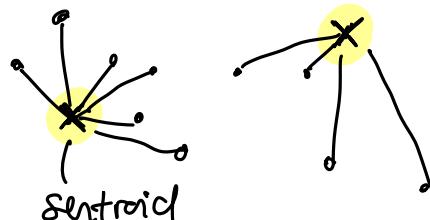
Nyttig sammenhang som er grunnlag for  $k$ -gjennomsnitt algoritme

$$\sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (\underbrace{x_{ij} - x_{i'j}}_2)^2$$

$$= \sum_{k=1}^K 2 \sum_{i \in C_k} \sum_{j=1}^p (\underbrace{x_{ij}}_{\text{point}} - \underbrace{\bar{x}_{kj}}_{\text{center}})^2,$$



med **klyngesentroide** i klynge  $k$ :  $\bar{x}_k = (\bar{x}_{k1}, \dots, \bar{x}_{kp})$ .



$$\rho = 2 (\bar{x}_{k_1}, \bar{x}_{k_2})$$

$k$  = naunet fil klynge

## K-gjennomsnitt algoritme

algoritmen kan ikke  
velge det for oss!

- Start med å velge antall klynker  $K$ .
- Tilordne **hver observasjon til en klyng**
  - Mange muligheter
    - å tilfeldig velge ut  $K$  observasjoner og sette disse som klyngesentroider, og deretter tilordne de resterende observasjonene til klyngen med nærmeste klyngesentroide.
    - tilfeldig klynger
- **Repeter** (iterativt) *til ingen observasjoner endrer klyngemedlemskap:*
  1. For hver klyng regn ut klyngesentroiden
  2. Tilordne hver observasjon til klyngen til nærmeste klyngesentroide

## Illustrasjon av $K$ -gjennomsnitt algoritme ( $K = 3$ )

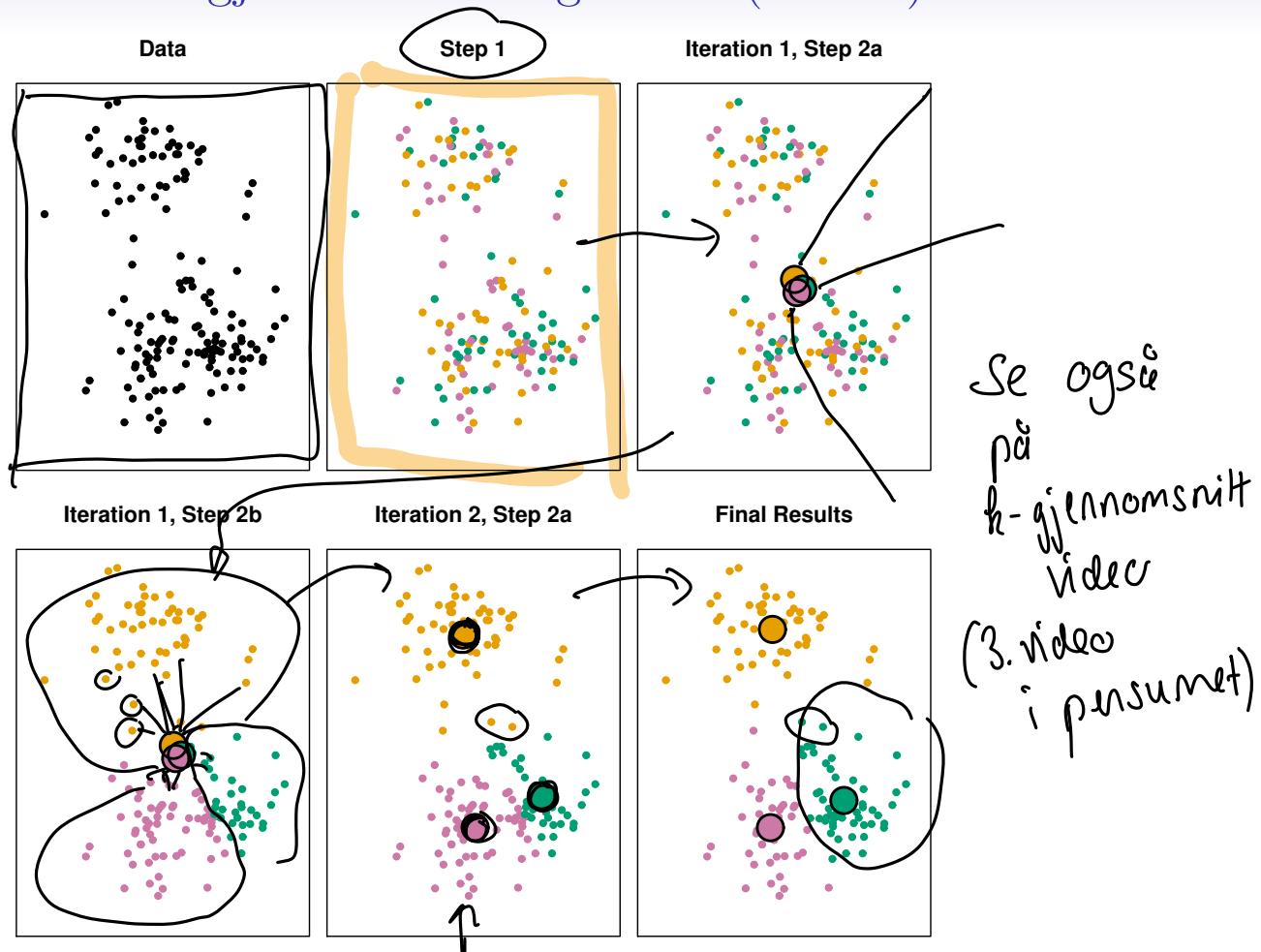


Fig. 10.6 fra "An Introduction to Statistical Learning with Applications in R", James et al 2013.

## Python kodechunk kmeans ( $K$ -gjennomsnitt-algoritmen)

- Package:

```
from sklearn.cluster import KMeans
```

- Steg 1: Antall klynger

```
antall_klynger = 10
```

hverdun velge?

Vet på forhand  
kjør med  
mange  $K$

- Steg 2: Initialiser k-means algoritmen

```
kmeans = KMeans(n_clusters = antall_klynger,  
random_state = 1)
```

seed

- Steg 3: Tilpass modellen

```
kmeans.fit(images)
```

dataset i prosjektet

- Sentroidene

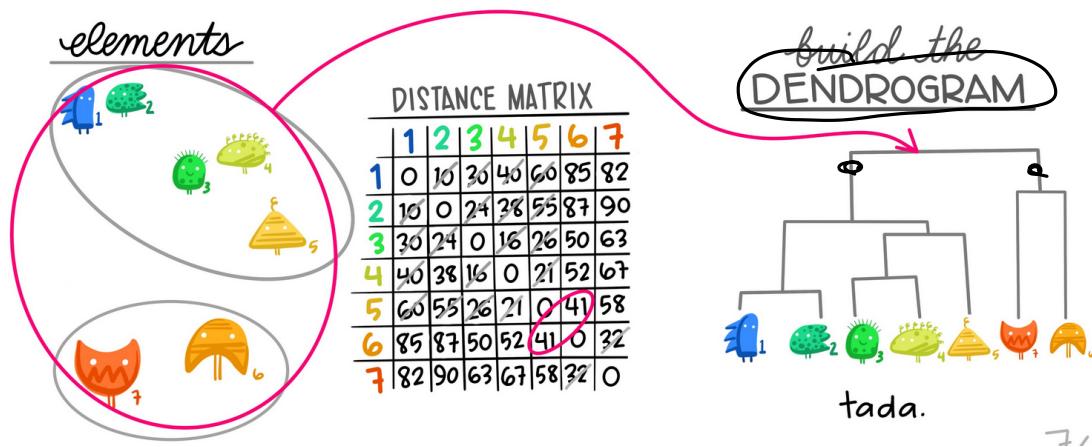
```
sentroider = kmeans.cluster_centers_
```

## Prosjektoppgaven

Vi kan se sammen på prosjektoppgaven.

|| Det var et problem  
med datasettet ...

# Hierarkisk klyngeanalyse



Artwork by @allison\_horst

Fordel: Må ikke vælg K på forhånd!

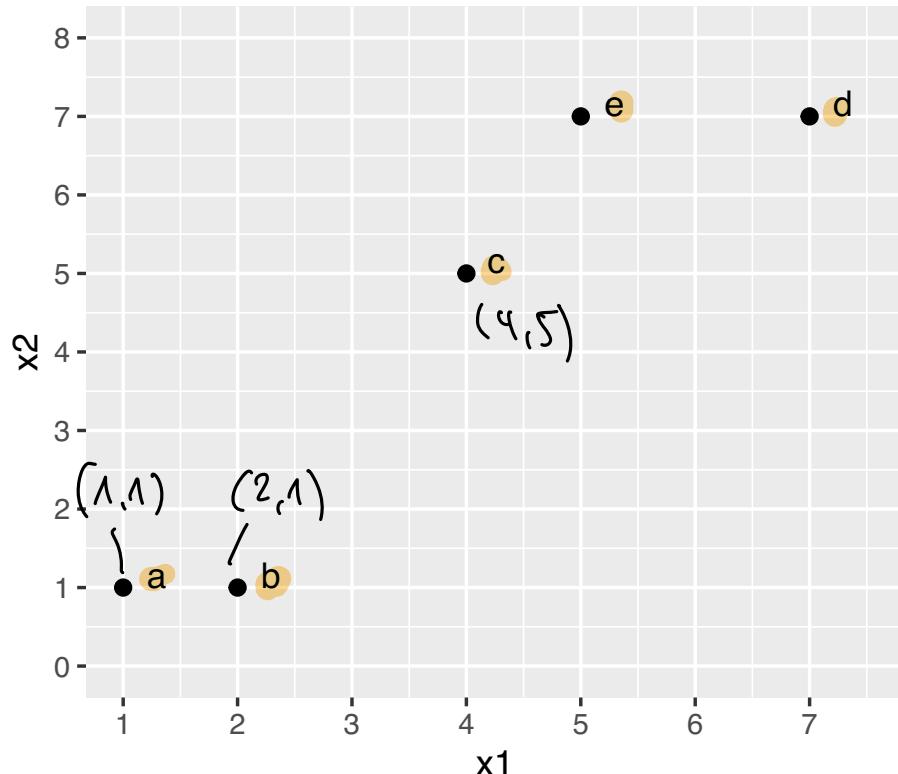
## Eksempel

$$n = 5, p = 2$$

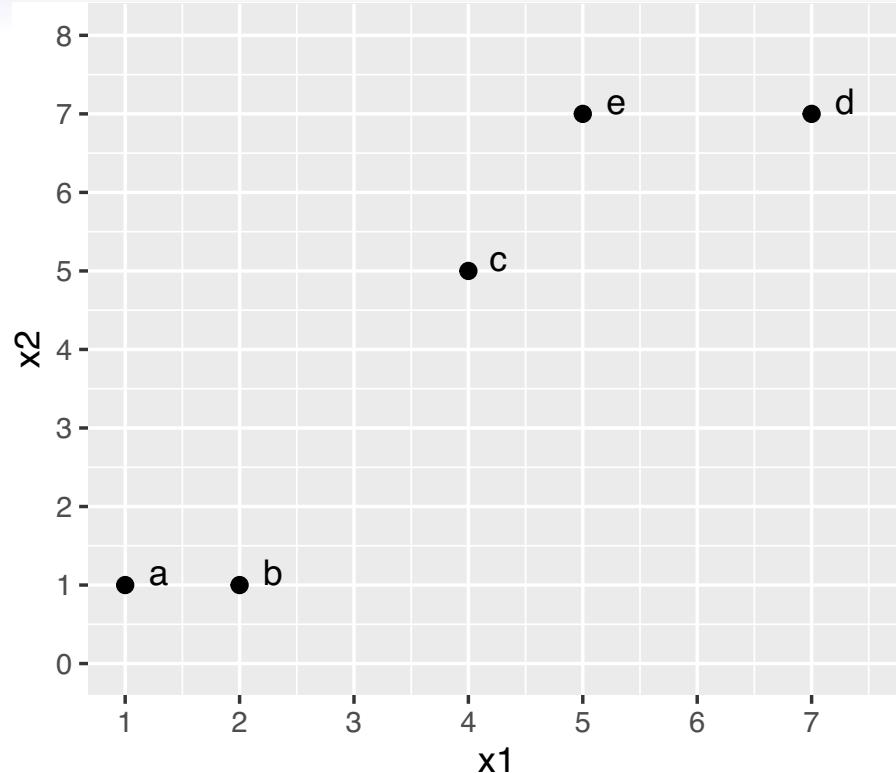
{

x1	x2	name
1	1	a
2	1	b
4	5	c
7	7	d
5	7	e

}

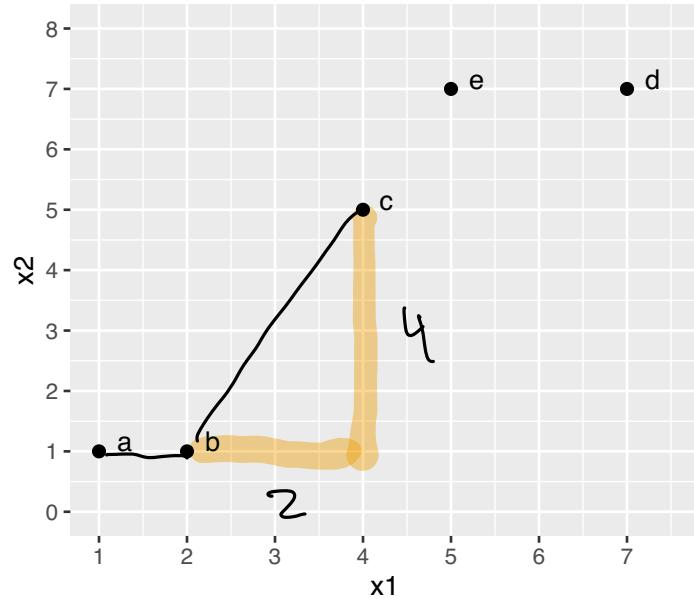


x1	x2	name
1	1	a
2	1	b
4	5	c
7	7	d
5	7	e



- 1) Velg avstandsmål. → Euklidisk avstand
- 2) Regn ut avstanden mellom alle par av observasjoner.
- 3) Plasser avstandene inn i en  $n \times n$  matrise.

# Avstandsmatrise (Euklidsk avstand)



$$\sqrt{2^2 + 4^2} = 4.5$$

name	a	b	c	d	e
a	0.0	1.0	5.0	8.5	7.2
b	1.0	0.0	4.5	7.8	6.7
c	5.0	4.5	0.0	3.6	2.2
d	8.5	7.8	3.6	0.0	2.0
e	7.2	6.7	2.2	2.0	0.0

Avstandsmatrise

$$\frac{5 \cdot 4}{2} = 10$$

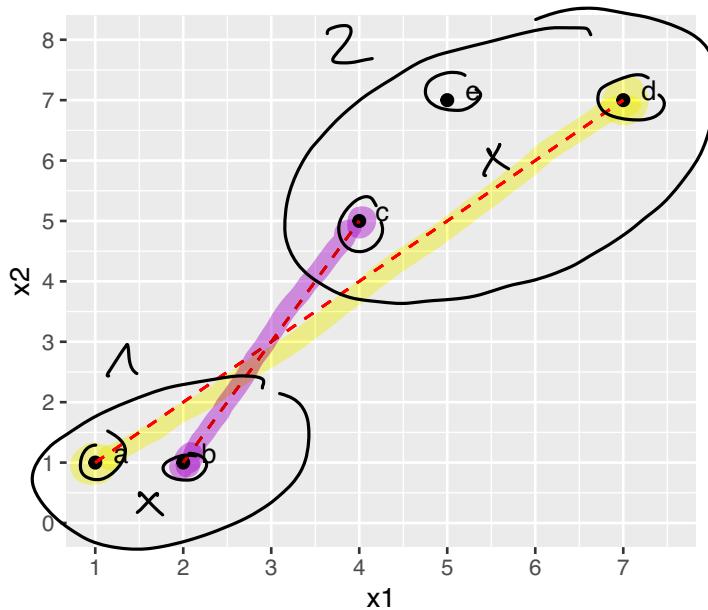
# Avstand mellom klynger?

Tre populære typer avstandsmål:



- **Singel kobling**: minimal avstand
- **Komplett kobling**: maksimal avstand
- **Gjennomsnittskobling**: gjennomsnittlig avstand

6 avstander  
a - c  
a - e  
a - d  
b - c  
b - e  
b - d  
f. eks



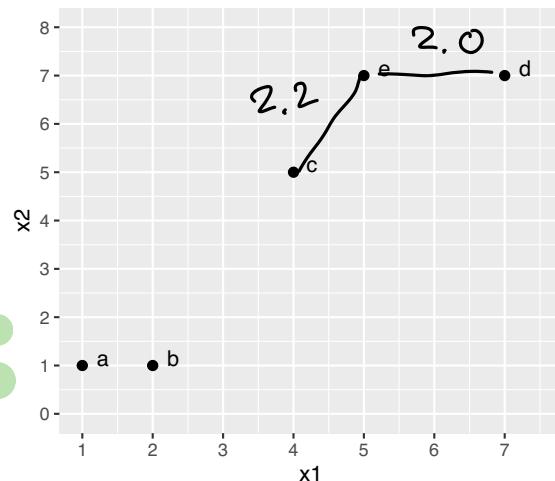
# Algoritme for hierarkisk klyngeanalyse - I

Før algoritmen starter man bestemme seg for

- hvilket avstandsmål skal brukes (f.eks: euklidisk, city block, korrelasjon,...)
- hvilken koblingstype skal brukes (f.eks: singel, komplett, gjennomsnitt, sentroide,...)

og regne ut avstandsmatrisen mellom alle observasjoner.

name	a	b	c	d	e
a	0.0	1.0	5.0	8.5	7.2
b	1.0	0.0	4.5	7.8	6.7
c	5.0	4.5	0.0	3.6	2.2
d	8.5	7.8	3.6	0.0	2.0
e	7.2	6.7	2.2	2.0	0.0



## Algoritme for hierarkisk klyngeanalyse – II

Behandle hver observasjon som om den var sin egen klynge (det er da  $n$  klynger).

### 1. Slå sammen

Finn de to klyngene som er nærmest hverandre og slå dem sammen til en klynge.

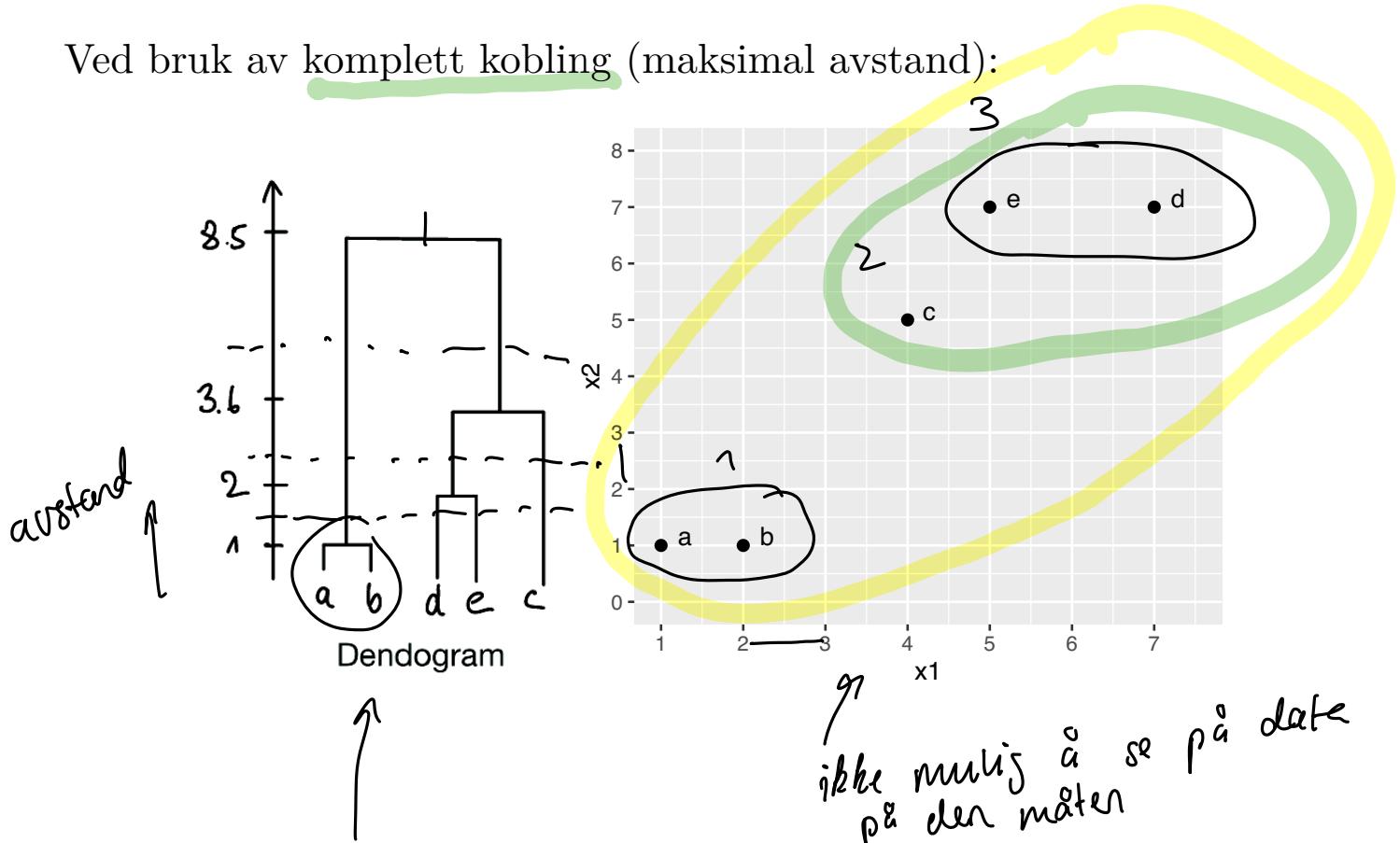
### 2. Beregn avstander

Beregn nye parvise avstander mellom alle klynger ved bruk av valgt avstandsmål og koblingstype.

Repeterer til alle observasjonene er i samme klynge.

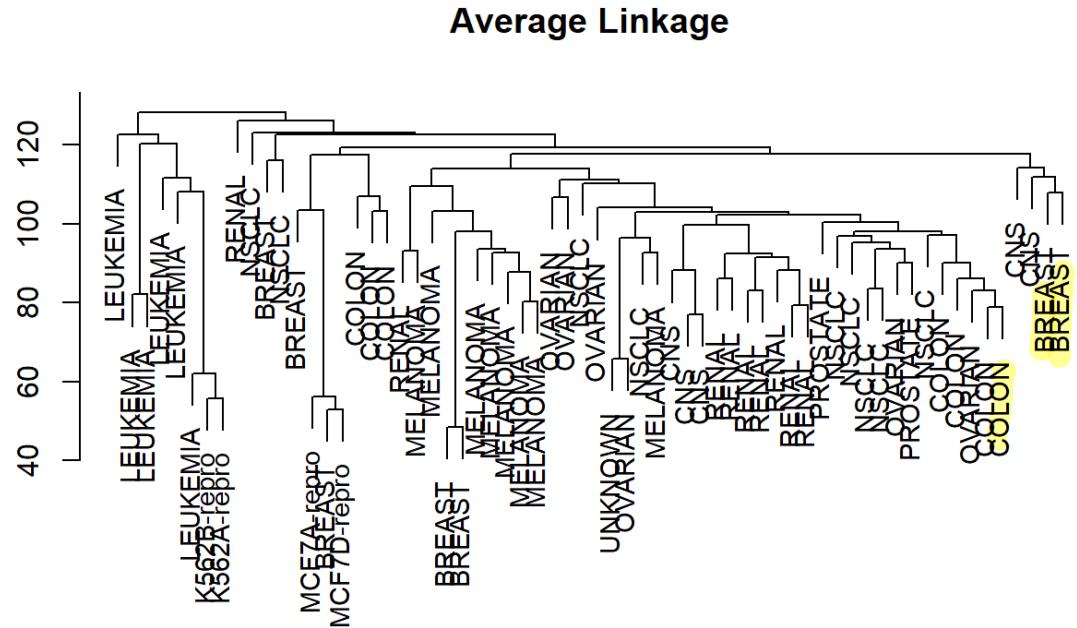
## Resultat fra vår eksempel (dendogram)

Ved bruk av komplett kobling (maksimal avstand):



## Eksempel: Genaktivitet

- $n=64$  celleprøver fra kreftsvulster til ulike pasienter.
  - genaktivitet for  $p = 6830$  gener.



## Prosjektoppgaven 3

- Hva er hovedforskjellene mellom K-gjennomsnitt-klyngeanalyse og hierarkisk klyngeanalyse?
- Hva er parameteren/parametrene på K-gjennomsnitt? På hierarkisk klyngeanalyse?  
Hilfsmiddel:
- Hvorfor har vi ikke brukt trenigns-, validerings- og testsett her?

## Videre de neste to ukene

- Hvis dere ikke har gjort det: Se på de korte videoene for hvert tema.
- Jobb med prosjektoppgaven. Husk at frist for innlevering av prosjektet til Inspera er **mandag 29.november kl 09.00**.
- Vi har **6** timer digital veiledning via Wherby ([www.wherby.com](http://www.wherby.com)) begge uker. Se her:  
<https://wiki.math.ntnu.no/istx1003/2021h/start>
- Husk også mattelab forumet – men bare for korte spørsmål.  
Lange spørsmål fungerer best med direkt interaksjon. {