INFO221v12
IR I

Truls Pedersen
Institutt for informasjons- og medievitenskap
Universitetet i Bergen

Oversikt

- Viktige begreper
- Hva er IR?
- Hva er DBMS?
- Design
- Indeksering
- Forhåndsbehandling av tekst (databaser)
- Precision/recall

Viktige begreper

Hva er data? Hva er informasjon?

Viktige begreper

"Tidlige" forslag til definisjoner:

Data *"A representation of facts ... in a formalized manner capable of being communicated or manipulated by some process."*

Informasjon *"the meaning that a human ... extracts from data by means of known conventions of the representation used."*

Gould, 1971

Viktige begreper

"Tidlige" forslag til definisjoner:

Data *"A representation of facts ... in a formalized manner capable of being communicated or manipulated by some process."*

Informasjon *"the meaning that a human ... extracts from data by means of known conventions of the representation used."*

Gould, 1971

Data "Data are symbols inscribed in formalized patterns by human hands or instruments."

Informasjon "Information is the meaning someone assigns to data."

Denning, 2001

Viktige begreper

Den nyere definisjonen av *data* skiller seg fra den gamle ved å ikke kreve noen funksjon/egenskaper av data.

Informasjon må ikke være oppnådd vha. “kjente konvensjoner”.

Vanskelig å komme frem til endelig generell definisjon.

Viktige begreper

Vi trenger (generelt) en standard for hvordan vi representerer data.

Vi har sett at det finnes flere forskjellige måter å kode et bilde.

Uten å bli enig om hvordan hva kodes, betyr data svært lite:

```
01001000011000010110110001101100011011110000 1010
```

Representasjon

Hva data representerer er ofte implisitt kjent.

Når representasjonen *ikke* er kjent må vi informere om dette.

Denne informasjonen om dataene må kommuniseres til mottakeren for at dataene skal gi mening.

Denne informasjonen kommuniseres som *metadata*: data om data.

Representasjon

Data som representerer sammensetninger av ulike typer medier kalles *multimedia data*.

En ballettoppsetning, en illustrert ordbok og et *youtube*-klipp er eksempler.

Vi skal se på *digitale multimedier*.

Datasamlinger

1. En eller flere deltagere bidrar til å konstruere innholdet i én samling. (Dataene lagres på ett sted.)
2. Dataene kan så hentes ut igjen av en bruker.

Enklere enn “1, 2, 3”?

Datasamlinger

1. En eller flere deltagere bidrar til å konstruere innholdet i én samling. (Dataene lagres på ett sted.)
2. Dataene kan så hentes ut igjen av en bruker.

Enklere enn “1, 2, 3”?

Prosessen er gjerne kontinuerlig, dataene må lagres på en konsistent måte, det må være mulig å hente frem ønsket informasjon, det ene stedet alt lagres må være sikkert, ...

Datasamlinger

Data lagres vanligvis i *databaser*:

Database "a logically coherent collection of related data, representing some aspect of the real world, that is designed, build, and populated with data for some purpose."

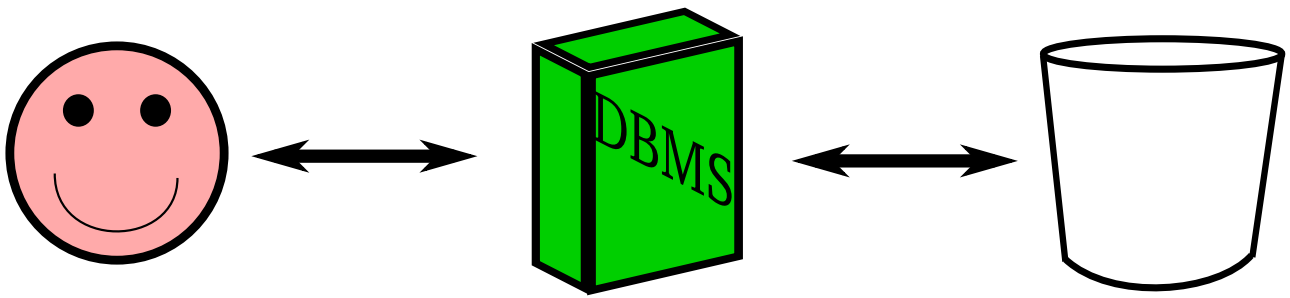
Vi krever at dataene i en database har en logisk sammenheng.

Dette setter ingen begrensninger på hva vi kan gjøre med en database, men krever at vi vet hva slags data som er i vår database.

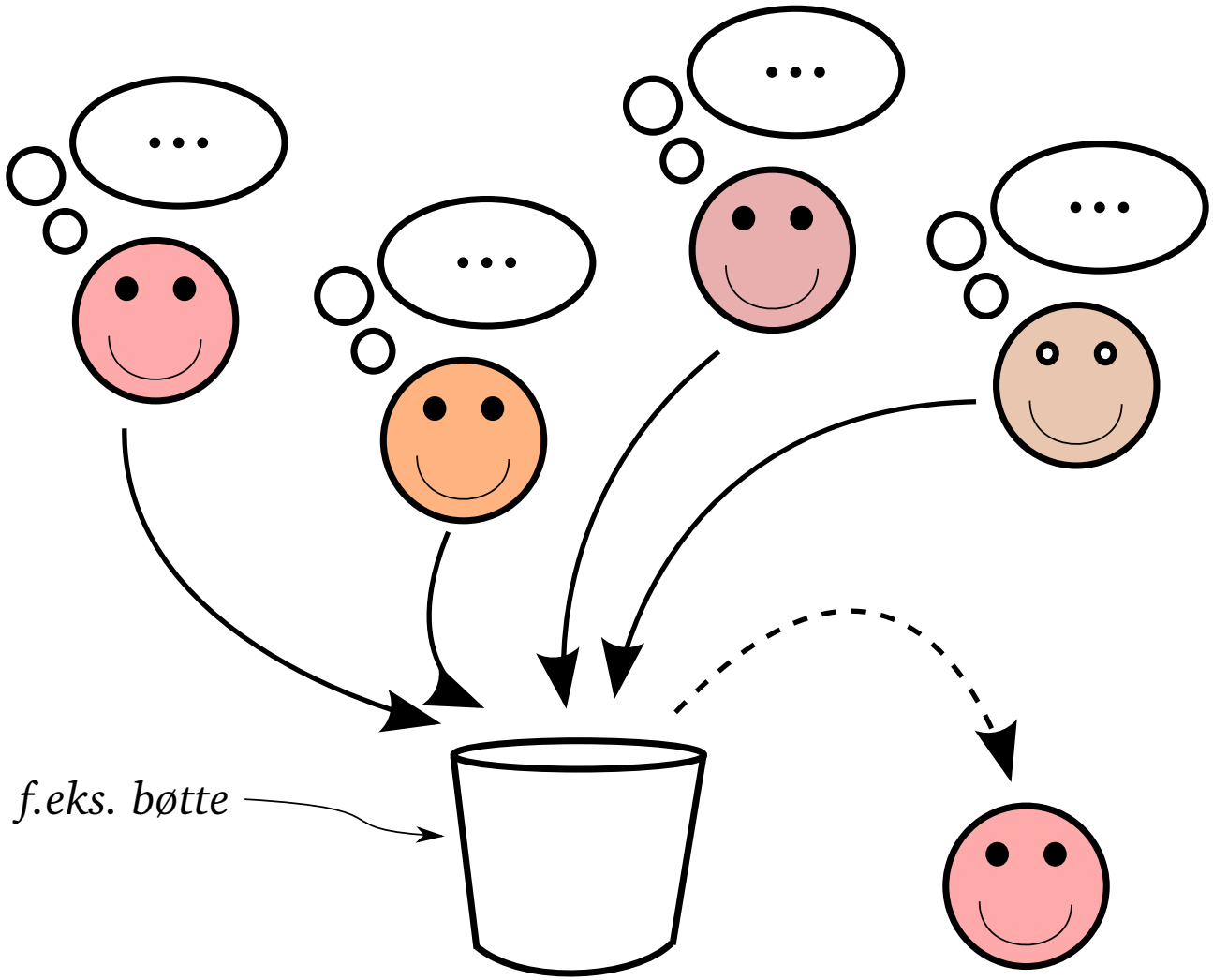
Datasamlinger

Bokhyllen min er en database.

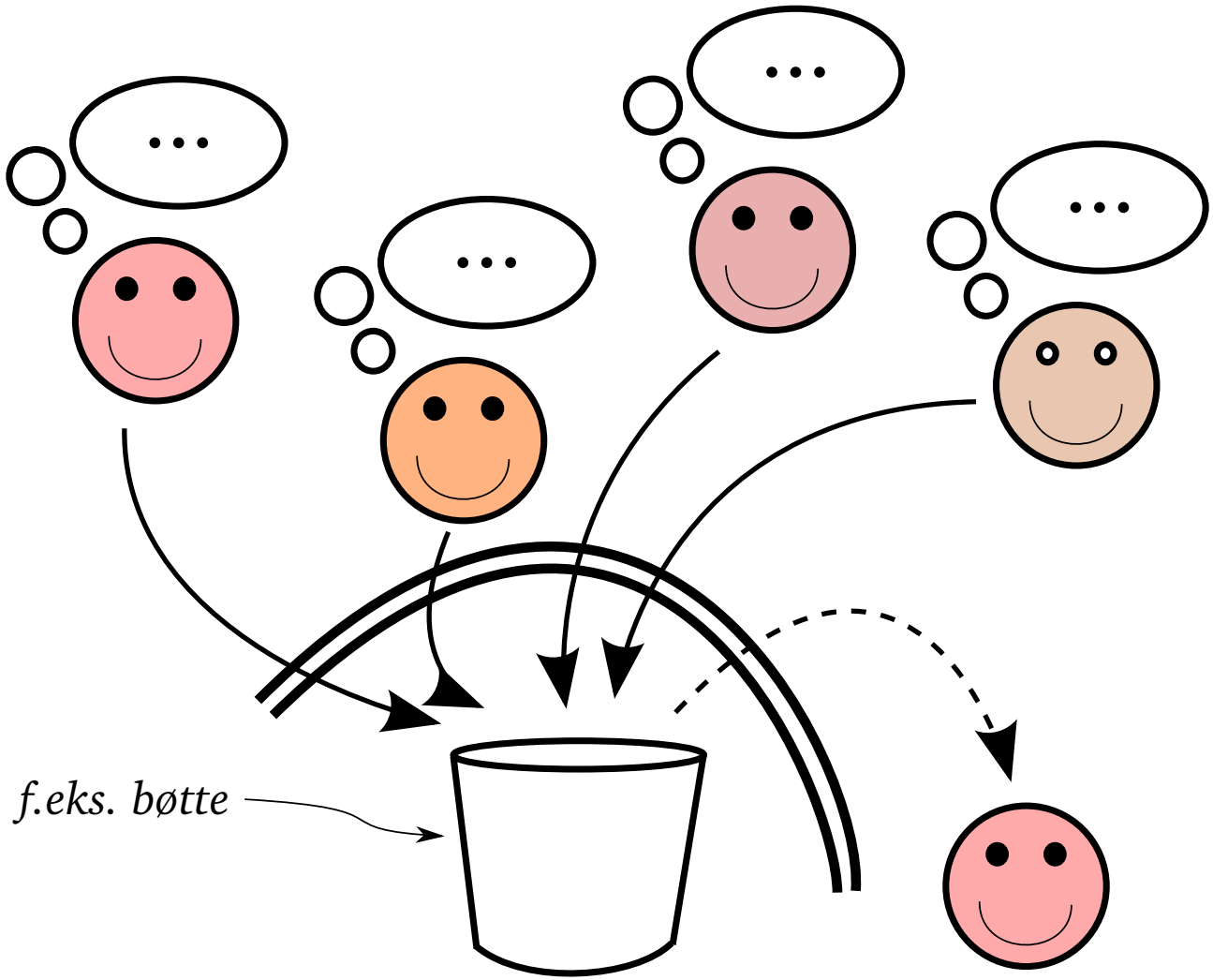
... men vi fokuserer fortsatt på de digitale databasene.



Hva er IR?



Hva er IR?



Hva er IR?

La oss anta at vi allerede har konstruert en database og at denne inneholder data.

Hvordan henter vi ut informasjon vi leter etter?

Hva er IR?

La oss anta at vi allerede har konstruert en database og at denne inneholder data.

Hvordan henter vi ut informasjon vi leter etter?

Det kommer naturligvis an på hva slags data databasen inneholder.

- Hvis man søker etter en student og vi kjenner studentnummeret kan vi søke etter et *eksakt treff*.
- Hvis man søker etter en artikken som omhandler en gitt ting må vi kanskje ha ganske kompliserte mekanismer.

Hva er IR?

Det finnes ikke én endelig løsning for informasjonsgjenfinning, systemene må generelt tilpasses et domene.

Eksempler på spesielt tilpassede systemer finnes for

- medisinske journaler
- rettsdokumenter
- telefonkataloger
- osv...

Hva er IR?

Det finnes ikke én endelig løsning for informasjonsgjenfinning, systemene må generelt tilpasses et domene.

Eksempler på spesielt tilpassede systemer finnes for

- medisinske journaler
- rettsdokumenter
- telefonkataloger
- osv...

Hva med bilder og lyd?

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

Magritte

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?
- Vi kan forsøke å beskrive bildet, f.eks “maleri av pipe”

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?
- Vi kan forsøke å beskrive bildet, f.eks “maleri av pipe”
- Siden vi allerede har et eksemplar kan vi søke etter bilder som er like (men har høyere oppløsning)

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?
- Vi kan forsøke å beskrive bildet, f.eks “maleri av pipe”
- Siden vi allerede har et eksemplar kan vi søke etter bilder som er like (men har høyere oppløsning)
- Et eksempel sammensatt av standardelementer

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?
- Vi kan forsøke å beskrive bildet, f.eks “maleri av pipe”
- Siden vi allerede har et eksemplar kan vi søke etter bilder som er like (men har høyere oppløsning)
- Et eksempel sammensatt av standardelementer
- En frihåndsskisse av bildet vi søker etter

Hva er IR?

Hvordan skal vi gå frem hvis vi ønsker å finne en bedre versjon av bildet vi så tidligere?

- Vi kan forsøke å finne alle bilder malt av René Magritte eller skal vi finne bilder malt av René Francois Ghislain Magritte?
- Vi kan forsøke å beskrive bildet, f.eks “maleri av pipe”
- Siden vi allerede har et eksemplar kan vi søke etter bilder som er like (men har høyere oppløsning)
- Et eksempel sammensatt av standardelementer
- En frihåndsskisse av bildet vi søker etter

Hva med lyd?

Hvordan fungerer IR?

Noen av disse søkemetodene vil kreve en del forhåndsbehandling av data.

Vi skal se på en del forhåndsbehandling for tekst i dag

Hvordan fungerer IR?

Noen av disse søkemetodene vil kreve en del forhåndsbehandling av data.

Vi skal se på en del forhåndsbehandling for tekst i dag, for bilder:

- bilder malt av René Francois Ghislain Magritte? (manuell) karakterisering

Hvordan fungerer IR?

Noen av disse søkemetodene vil kreve en del forhåndsbehandling av data.

Vi skal se på en del forhåndsbehandling for tekst i dag, for bilder:

- bilder malt av René Francois Ghislain Magritte? (manuell) karakterisering
- beskrive bildet, f.eks “maleri av pipe” objektgjennfinning

Hvordan fungerer IR?

Noen av disse søkemetodene vil kreve en del forhåndsbehandling av data.

Vi skal se på en del forhåndsbehandling for tekst i dag, for bilder:

- bilder malt av René Francois Ghislain Magritte? (manuell) karakterisering
- beskrive bildet, f.eks “maleri av pipe” objektgjennfinning
- bilder som er like/ligner eller frihåndsskisse av bildet grafisk “likhet”

Hva er D(B)MS?

Det som ofte er tilfelle når disse systemene utvikles er at forskjellige avdelinger utvikler systemer som passer *seg selv*.

Siden systemene da ofte er ganske enkle å utvikle kan det fort bli mange av dem.

Det er et stort problem for bedrifter å holde styr på alle databasesystemene sine.

Tenk for eksempel på oppkjøp/fusjon.

DBMS

Hvor bra er et gitt databehandlings system?
Hvilke kvalitetskrav setter vi til systemet?

- Hvor fort får vi resultatene vi søker etter?
- Hvor mye plass trenger vi?
- Hvor relevante resultater får vi når vi bruker systemet?

DBMS

Hvor bra er et gitt databehandlings system?
Hvilke kvalitetskrav setter vi til systemet?

- Hvor fort får vi resultatene vi søker etter?
- Hvor mye plass trenger vi?
- Hvor relevante resultater får vi når vi bruker systemet?

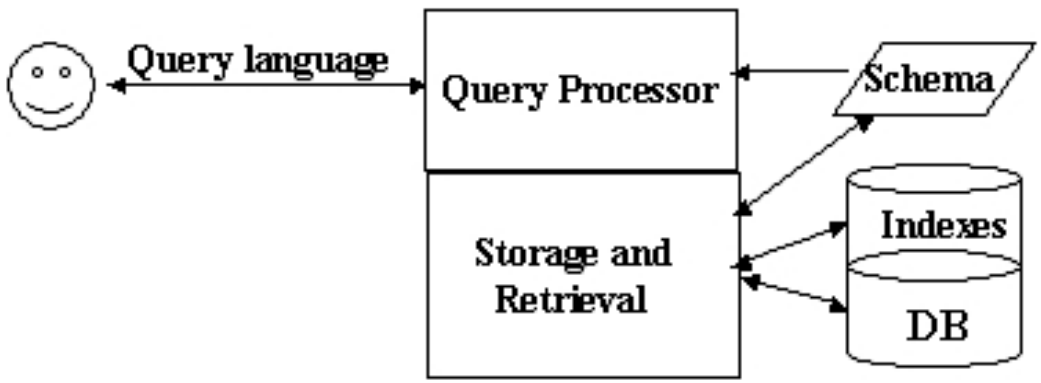
For dette bruker vi ofte følgende teknikker:

- Indeksgenerering
- Kompresjon
- Brukergrensesnitt
- Diverse algoritmer (f.eks. likhetstesting)

DBMS

Når vi søker i en database:

1. Bruker gir input
2. Tolkeren tolker input i forholdet som databasestrukturen
3. Vi søker etter de dokumenter som passer søket v.h.a. eksakte søk, likhetstester, indekser, ...
4. Presenterer resultatet for brukeren



DBMS

DBMS vi designer må ta hånd om alle disse elementene.

Hvordan spesifiserer brukeren hva hun leter etter?

- Fritekstsøk som google?
- Fylle ut utvalgte felter, som på biblioteket?
- Profesjonelt spørrespråk som SQL?

Hvordan presenterer vi resultatet? Alt på en side?

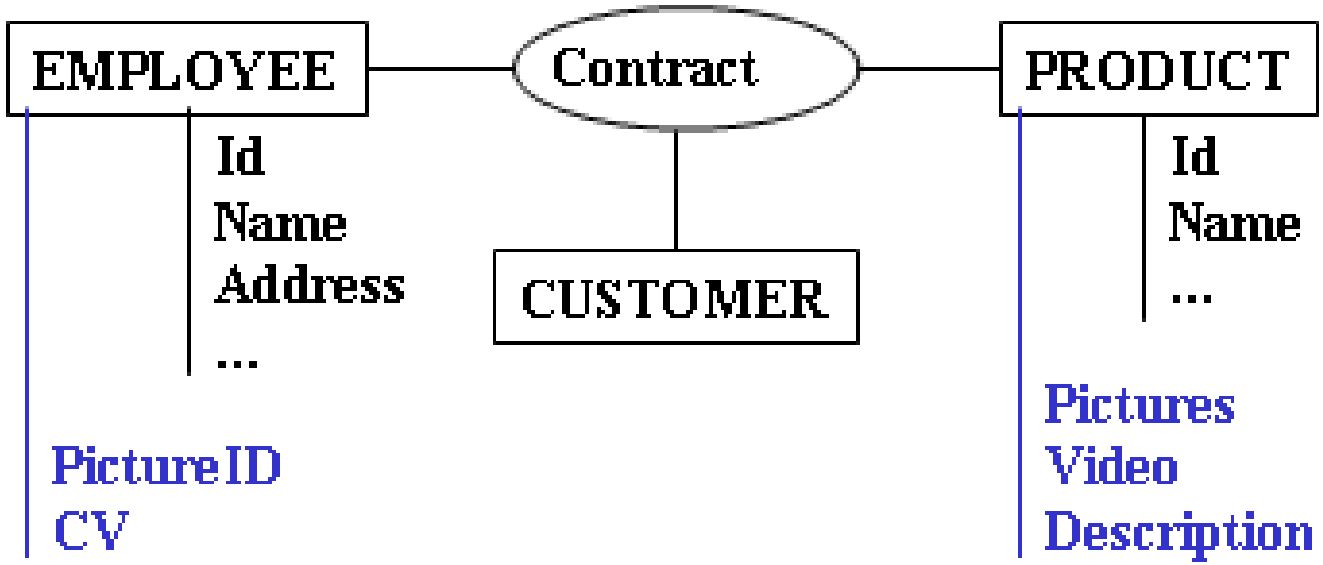
Tolker tolkeren input på en fornuftig måte? Samsvarer det med intuitiv bruk?

“Hvor like” må to bilder være for å komme med i resultatmengden?

Vi må velge dette utifra behovene vi prøver å tilfredsstille.

Design

I designprosessen må vi gjøre klart hvordan data skal lagres.



ER diagram for en database.

Et viktig spørsmål vi må ta stilling til her er *hvor* lagres mediaelementene?

Design

Vi kan lagre alle videoene i et filsystem og lagre peker til denne filen i databasen.

Dette fører til at databasen blir mindre. Eksterne programmer har lett tilgang til alle videoene.

Alternativt kan vi lagre videodataene i en kolonne i databasen.

Boken fraråder oss mot å gjøre dette, men nyere database teknikker gjør dette greit.

Design

Vi kan lagre alle videoene i et filsystem og lagre peker til denne filen i databasen.

Dette fører til at databasen blir mindre. Eksterne programmer har lett tilgang til alle videoene.

Alternativt kan vi lagre videodataene i en kolonne i databasen.

Boken fraråder oss mot å gjøre dette, men nyere database teknikker gjør dette greit.

Hvordan søker vi så etter en “sånn-og-sånn” video?

Spørring (veldig kjapt)

Q1: *List the titles of database texts written by Joan Nordbotten.*

```
SELECT D.Title
FROM PERSON P, AUTHOR A, DOCUMENT D
WHERE P.Name ='Joan Nordbotten'
AND P.Id = A.PId
AND A.DId = D.Id
AND D.Title LIKE '%database%';
```

Result: a list of titles

Spørring (veldig kjapt)

Q1: *List the titles of database texts written by Joan Nordbotten.*

```
SELECT D.Title
FROM PERSON P, AUTHOR A, DOCUMENT D
WHERE P.Name ='Joan Nordbotten'
AND P.Id = A.PId
AND A.DId = D.Id
AND D.Title LIKE '%database%';
```

Result: a list of titles

Hva med `D.Body LIKE '%database%';` ?
Eller videoer som er mer enn 70 minutter?

Det kan bli svært tidkrevende å søke gjennom all teksten for hvert søk. Vi konstruerer indekser!

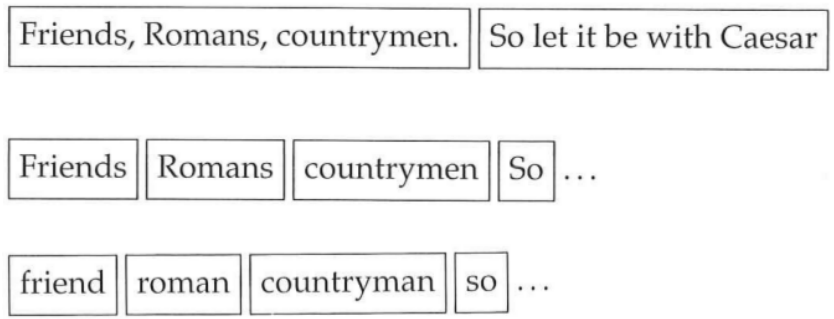
Indeksering

For tekstdokumenter vil indeksering grovt gjøres som følger:
For hvert dokument d:

1. Finn ordene teksten består av (tokens).
2. Normaliser disse.
3. Fjern uønskede ord.
4. For hver token/term t som blir igjen enten:
   - 4.1 la dokumentet referere til termen: $d \rightarrow t$
   - 4.2 la termen referere til dokumentet: $t \rightarrow d$ (*invertert*)

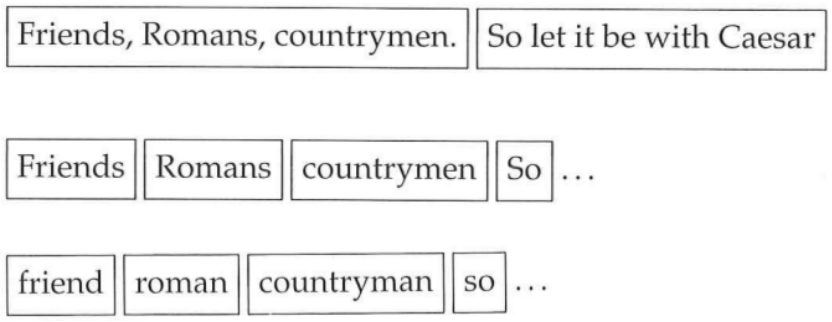
Indeksering

Finner ordene i teksten og “vasker” dem (steg 1-3):

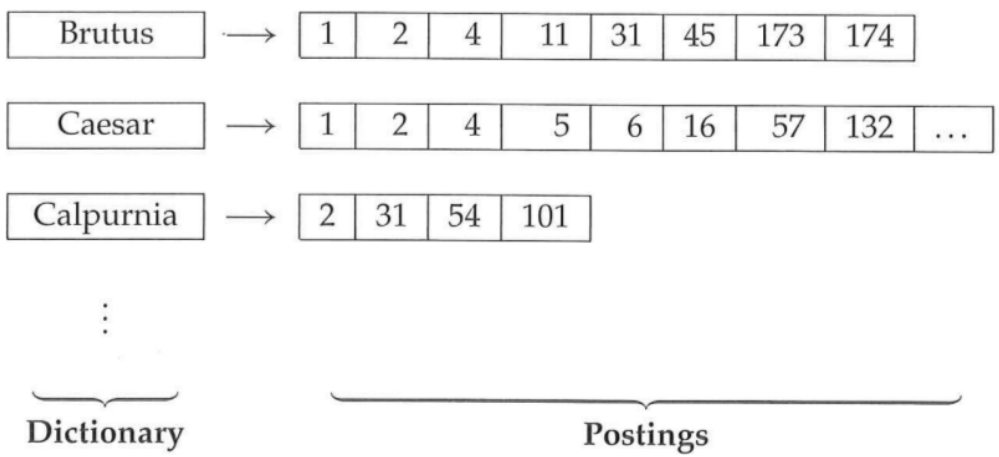


Indeksering

Finner ordene i teksten og “vasker” dem (steg 1-3):



Peker til dokumentene ordene kom fra: (steg 4.2):



Indeksering

Gitt noen dokumenter (horisontalt) og alle interesante termer de inneholder (vertikalt), kan vi visualisere samlingen som en matrise

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Indeksering

Gitt noen dokumenter (horisontalt) og alle interesante termer de inneholder (vertikalt), kan vi visualisere samlingen som en matrise

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Hvilke ord forekommer i dokument 2?

Indeksering

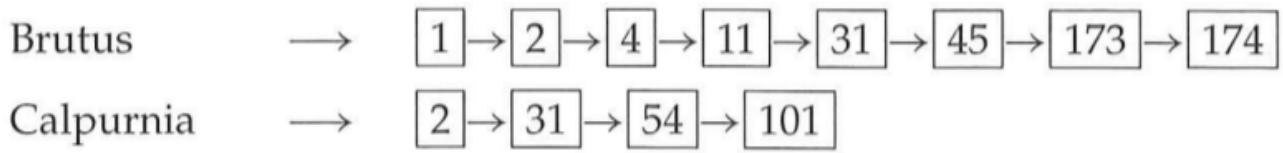
Gitt noen dokumenter (horisontalt) og alle interesante termer de inneholder (vertikalt), kan vi visualisere samlingen som en matrise

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Hvilke dokumenter inneholder ord 5 (‘Cleopatra’)?

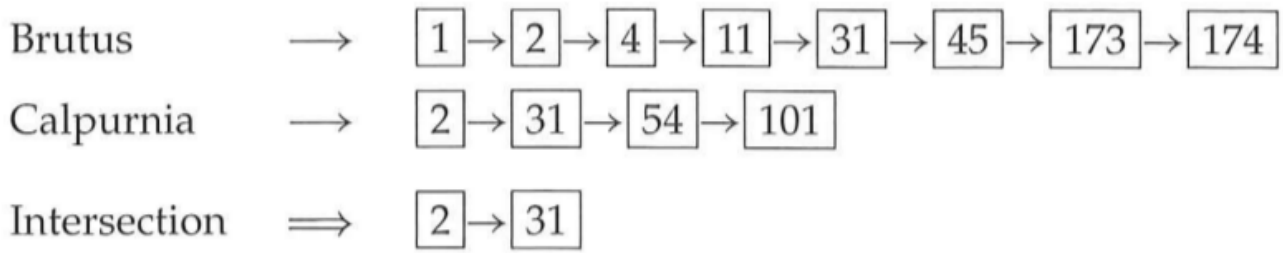
Søking

Nå kan vi enkelt sile ut de dokumentene som inneholder f.eks. ‘Brutus’, eller ‘Calpurnia’

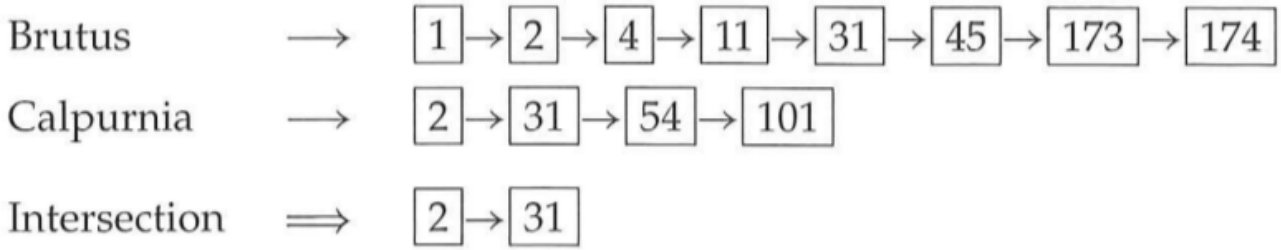


Søking

Nå kan vi enkelt sile ut de dokumentene som inneholder f.eks. ‘Brutus’, eller ‘Calpurnia’
... og finne de doumentene som inneholder *begge*.



Søking




```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Indeksering

Bilder består ikke av lett identifiserbare deler som dokumenter gjør.

Nøkkelord og andre søkedata må oppdrives.

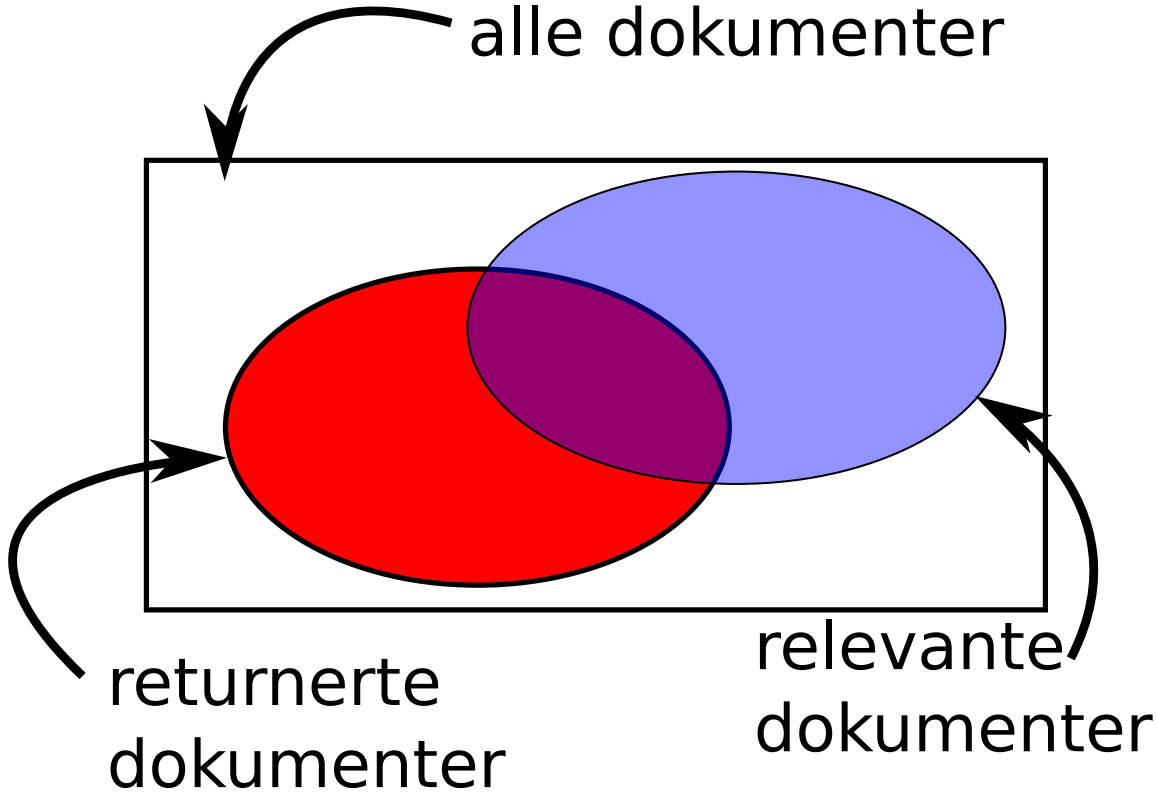
Disse får vi hovedsakelig fra:

Metadata samler automatisk, genererer manuelt, får *crowd*'en til å hjelpe oss, ...

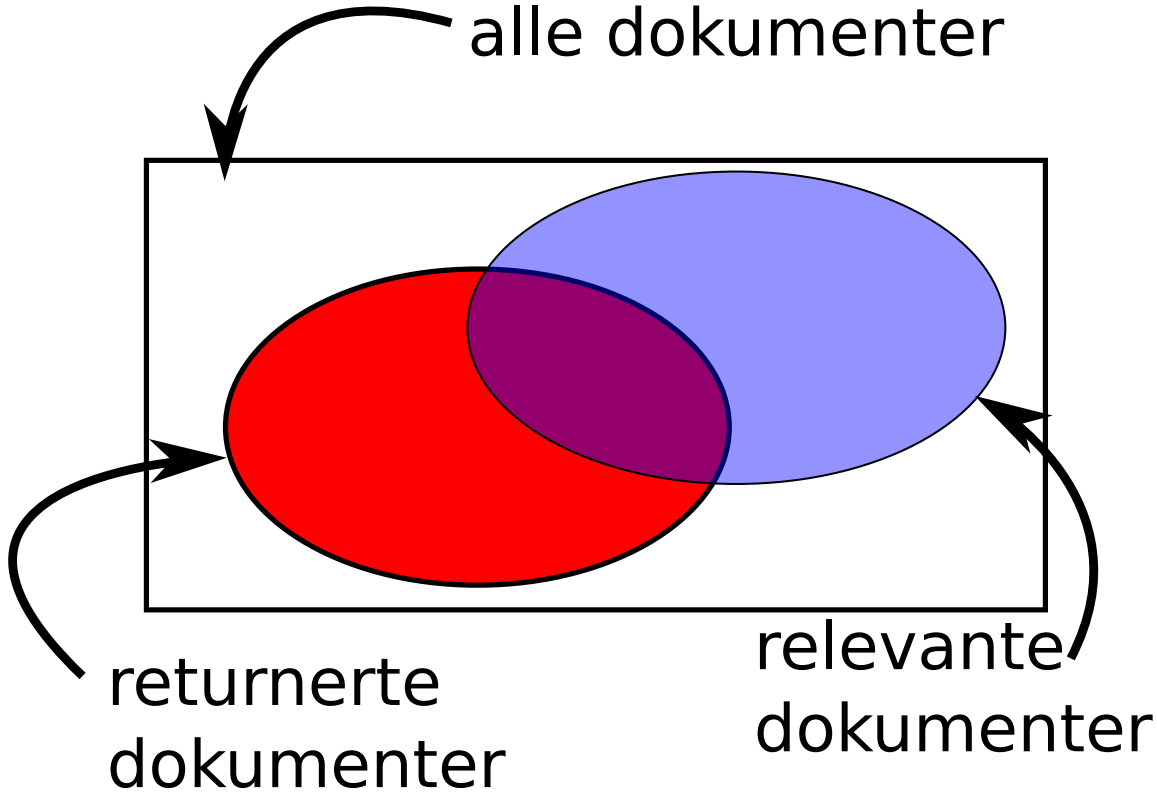
Enkle egenskaper farge, mønster, "skisse-poeng", ...

Denne informasjonen legger vi inn i databasen ved siden av selve bildedata.

Hvor lang er en meter?

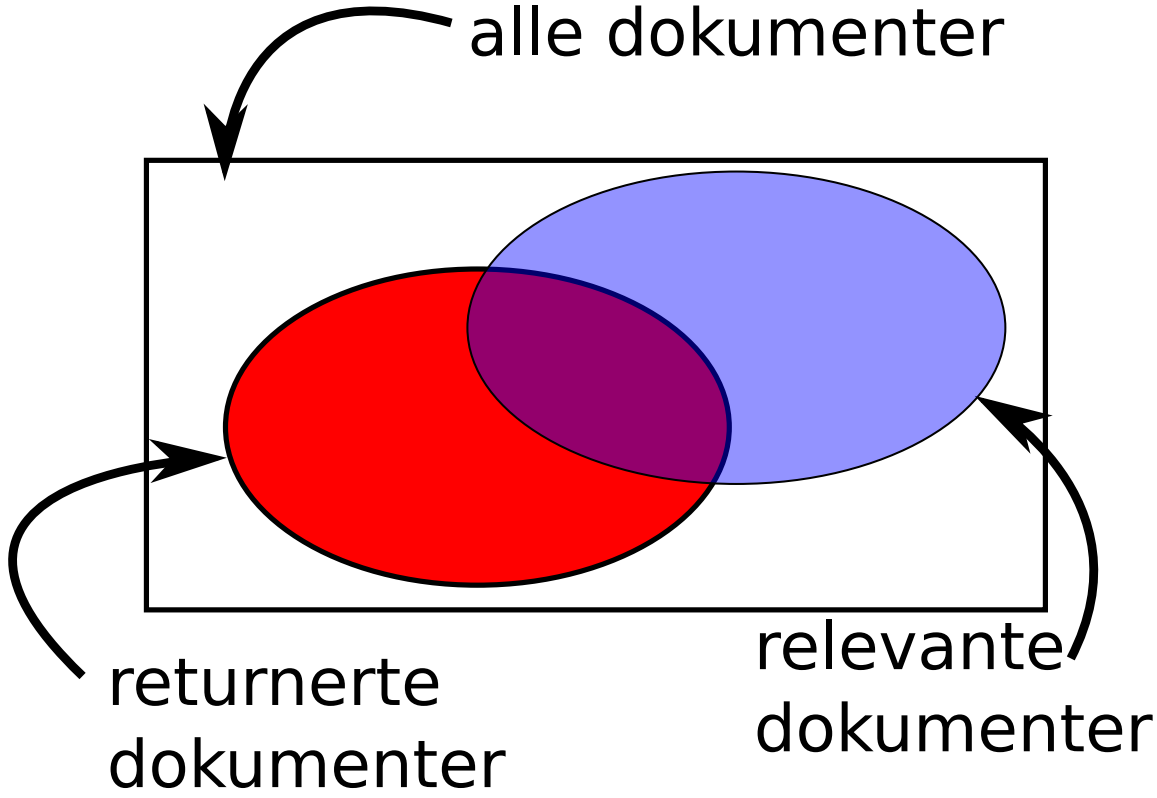


Hvor lang er en meter?



- ***Precision:*** hvor stor andel av de returnerte dokumentene er relevante?

Hvor lang er en meter?



- ***Precision:*** hvor stor andel av de returnerte dokumentene er relevante?
- ***Recall:*** hvor stor andel av de relevante dokumentene ble returnert?

Anbefalte oppgaver

Manning et. al. (IR2):

- Ex. 1.1, 1.2, 1.3, 1.8
- Ex. 2.1, 2.2