

# INF230

# Datahåndtering og analyse

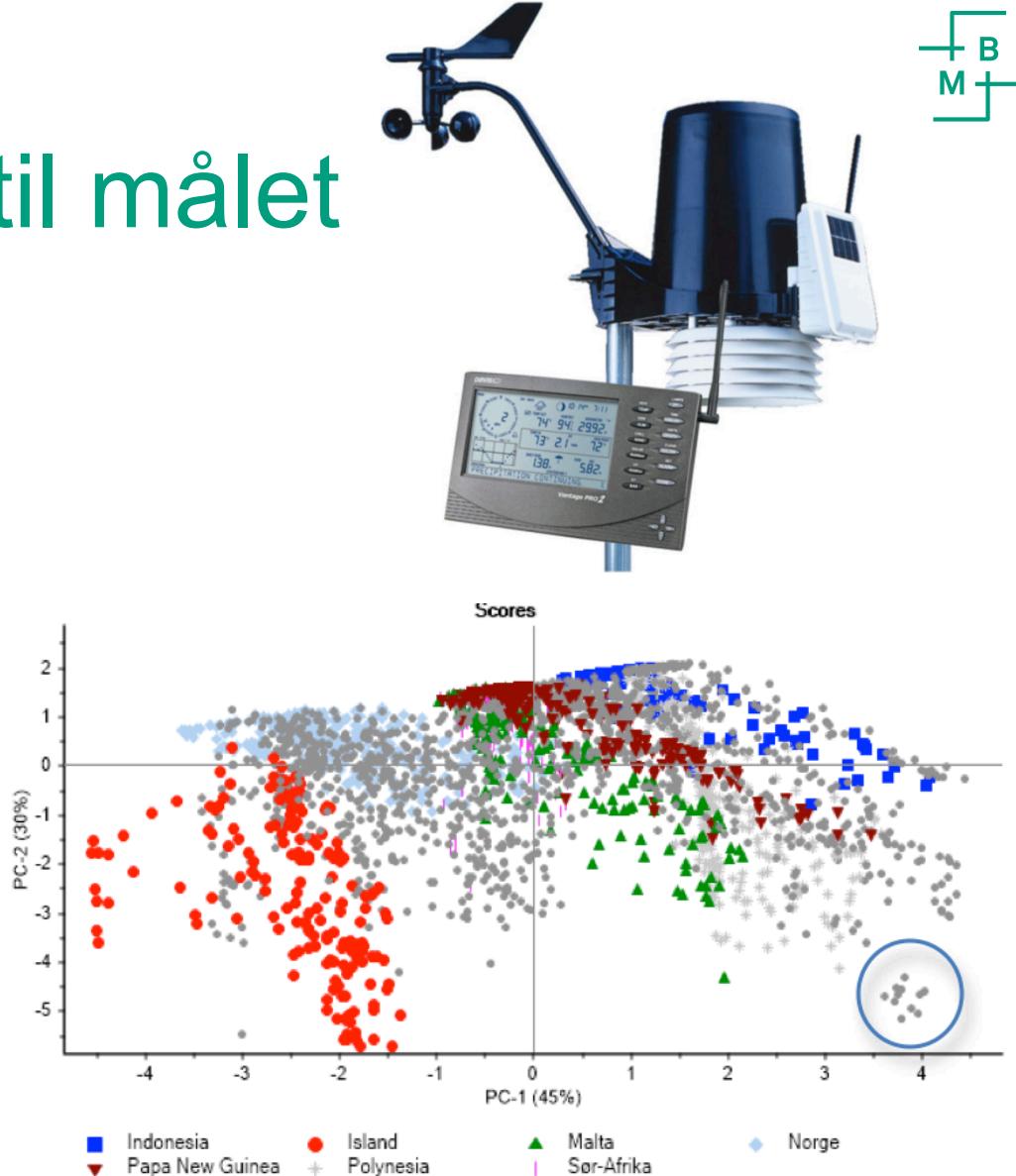
Introduksjon

Ingunn Burud

Aleksander Hykkerud

# INF230 fram til målet

- Lære databaser og datahåndtering
- Samle inn data fra dataloggere (Værstasjoner)
- Behandle data
- Visualisere
- Analysere



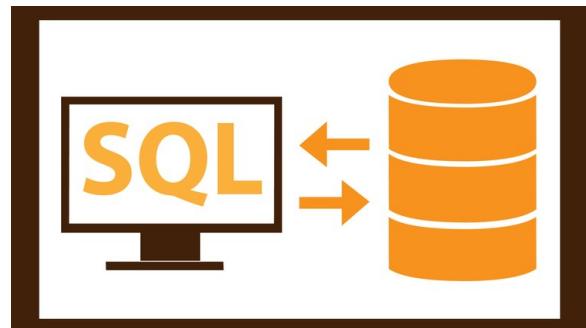
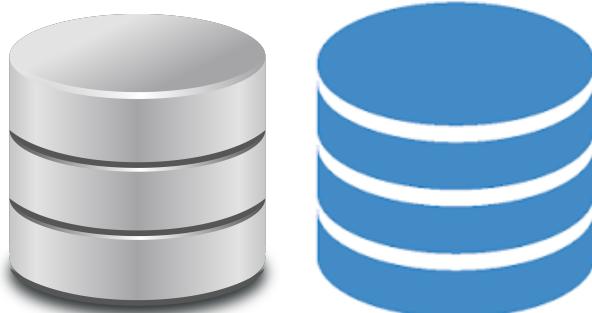
# Temaer vi skal behandle

- Datahåndtering og analyse
- Hovedtemaer
  - Databaser
  - Innsamling av data
  - Visualisering
  - Analyse av innsamlede data

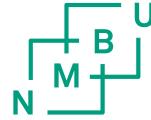
# Datahåndtering - Databaser



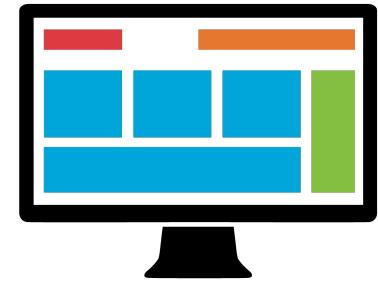
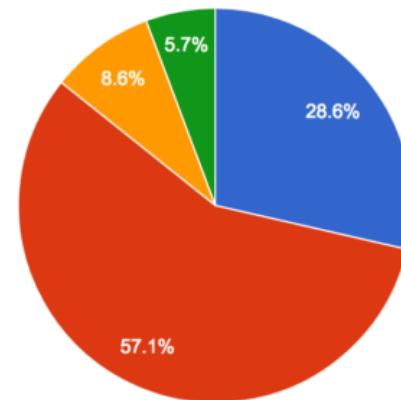
- Vi lærer grunnleggende databaseteori og hvordan vi kan benytte et databasesystem for å håndtere data
- Databaser består av tabeller og koblinger av tabeller
- Et spørrespråk (SQL) benyttes til å definere og hente data fra databasen
- Vi skal se på hvordan data i tabeller kan kobles



# Innsamling og Visualisering av Data

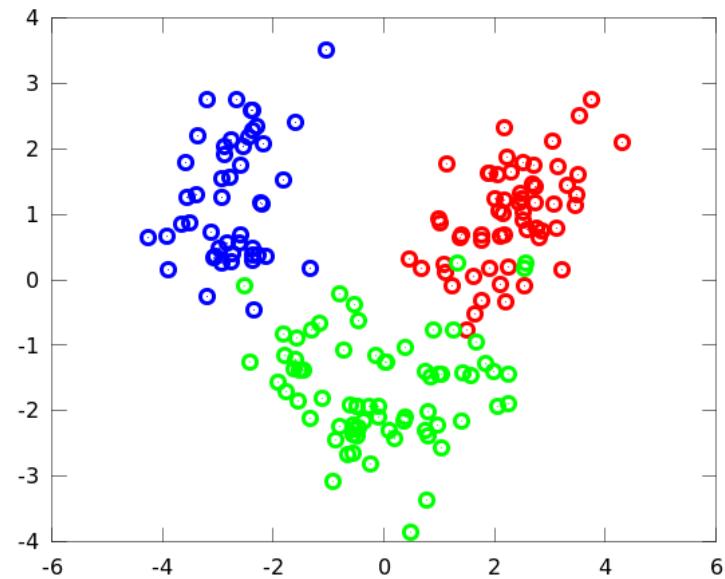


- Vi lærer grunnleggende metoder for å visualisere data hentet fra databaser
- Grunnleggende i denne delen er innsamling og datahåndtering av værstasjoner som finnes rundt om på kloden
- Vi ser på ulike typer visualiseringer
  - On-line datavisualisering
  - Ulike typer grafer
- Vi lager websider



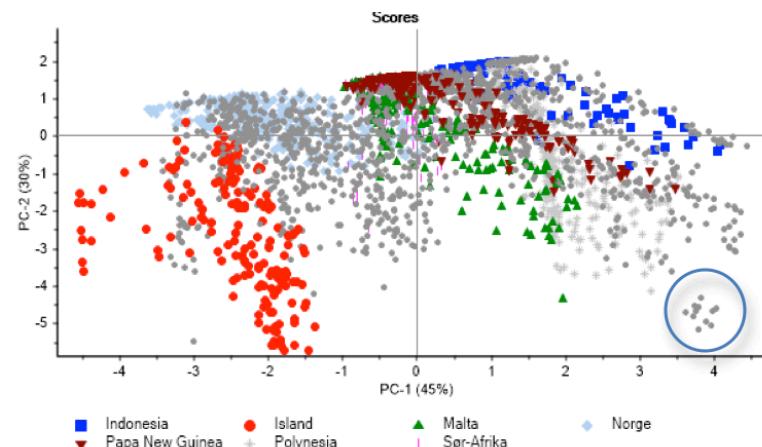
# Analyse - Data

- Vi lærer grunnleggende metoder for å analysere data hentet fra databaser
- Grunnleggende i denne delen er bruken av Multivariat statistikk
  - Prinsipal-komponent-analyse (PCA)
  - Klassifisering



# Automatisk innsamling av data

- Data fra loggere
- Værstasjoner
- GPS data
- Analyse og visualisering av værdata fra databaser



# Forelesninger

- Forelesningene holdes på onsdager kl 8:15-10
- Sentralt stoff vil gjennomgås
- Gjennomgangen er ikke en blåkopi av læreboka eller annet stoff som blir utlevert



# Plenums-gjennomgang

- Gjennomgang av stoff som er av felles interesse
  - Forslag til løsninger på øvingsoppgaver
  - Demonstrasjon av teknikker
  - Installasjon av programvare
  - Bruk av Python/Matlab mot databaser
- 
- Er det noe DU vil at vi skal gjennomgå så si ifra!!!

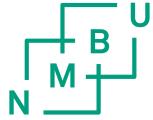
Mandager 8:15-10 ved behov, vil bli annonsert på Canvas

# Øvingene

- Øvinger er viktig. Du skal gjennomføre oppgavene fram til neste forelesning
- Øvinger på datasalen TF 02 mandager og tirsdager (grupper tildelt)
  - Mandag 10:15-12 (1)
  - Tirsdag 12:15-14 (2) og 14:15-16 (3)
  - Noen oppgaver bør løses i grupper. Kollokviegrupper meldes opp selv i Canvas
    - ca 3 i hver gruppe.
- MERK: Emnet er et “modningsemne”. Det er ikke mulig å jobbe etter “skippertakprinsippet”

# Obligatoriske oppgaver

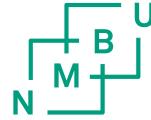
- 2 obligatoriske oppgaver
- Løsninger besvares ved opplasting til Canvas
- Alle må være godkjent før eksamen
- I tillegg skal en prosjektoppgave leveres mot slutten av kurset



# Lærebok og materiale

- Bjørn Kristoffersen: Databaser. 4. utgave
- Dokumenter på Canvas
- <http://dbsys.info> (bokas nettside)

# Programvare verktøykasse



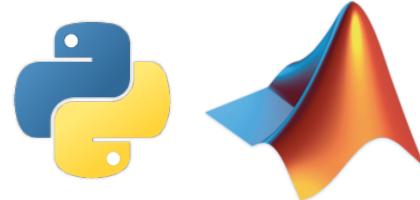
- BITNAMI
  - MySQL database
  - Apache Web server



- MySQL Workbench

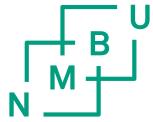


- Python/matlab



# Bruke egen datamaskin

- Det anbefales at du benytter din egen datamaskin
- All programvare kan installeres enkelt
- Uavhengig av NMBU nettverk



# 1

## Introduksjon

---

Bjørn Kristoffersens bok



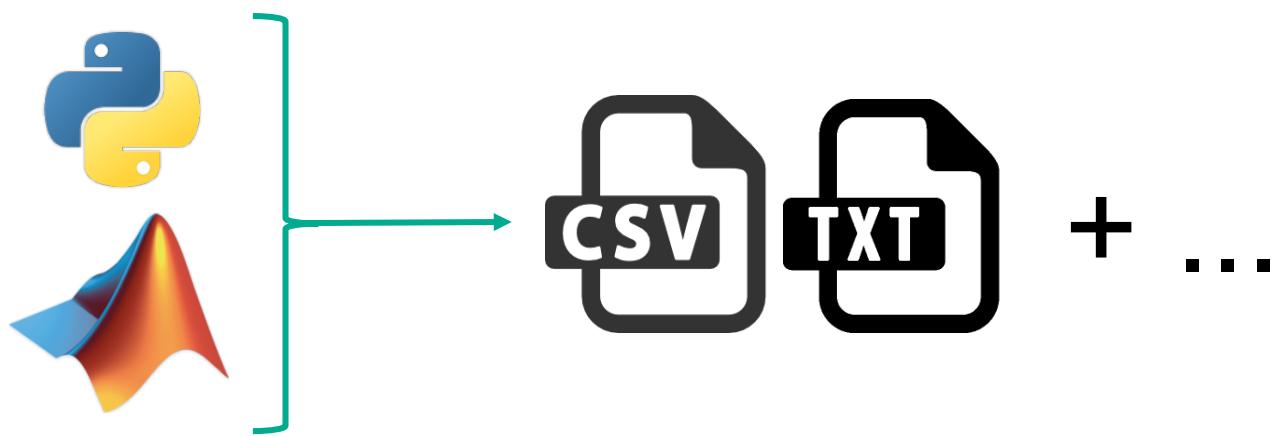
# Læringsmål

- Kjenne til viktige anvendelser av databasesystemer
- Kjenne til oppgavene et databasesystem har
- Kjenne til hvordan et databasesystem blir utviklet
- Forstå hovedprinsippene for digital representasjon av data

---

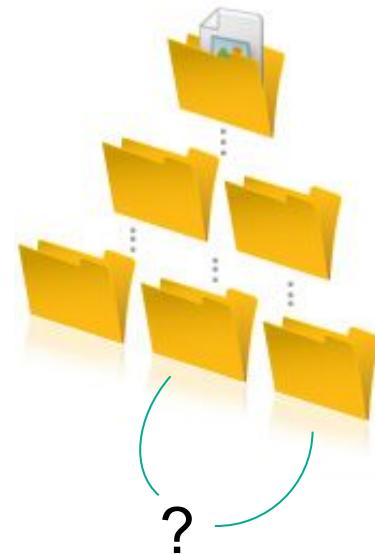
# Datalagring

- Kjente programmer for datalagring

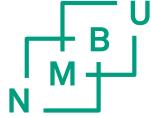


# Datalagring

- Uten databaser
  - Vanskelig og tidkrevende å finne og dele data
  - Lett å gjøre feil



# Datalagring



- Krav til datalagring
  - lagre store mengder **strukturerte** data over **lang tid** på en **sikker** måte
  - gjenfinne data **effektivt** og korrekt
  - betjene mange, **samtidige** brukere (deling av data)
  - kommunisere** med andre programsystemer
  - håndtere **feilsituasjoner** som diskkrasj og strømbrudd
- Databaser oppfyller alle disse kravene



# Det ligger ofte en database bak

The collage consists of three screenshots:

- Top Left:** A screenshot of a Microsoft Internet Explorer browser window showing the Komplett.no website. The page displays a search bar and a sidebar with categories like Datautstyr, Backup/Media, Barebones, Casemods, etc. Below the sidebar are product cards for a Western Digital Passport II 250GB drive and a D-Link DIR-615 Wireless N Router.
- Top Center:** A close-up photograph of a hand holding a USB flash drive. The drive is illuminated from within, showing a green glow and some internal components. The word "VISA" is printed on the side of the drive.
- Bottom Right:** A screenshot of the YR.no weather website. It features a search bar for location, a warning for difficult driving conditions in West-Fjords, and a weather forecast table for Oslo, Bergen, and Stavanger. To the right, there's a section titled "Studiestart" with a button for "INFORMASJON OM STUDIESTART". Below this are sections for "Studier", "Forskning", and "Om HSN". At the bottom, there's a map of Norway with weather icons and a table for Sihcejavri and Nedre Vats.

# Databasetabeller

- Data lagres i tabeller

**Rader**  **Kolonner** 


- På databasespråk
  - Kolonne = Attributt
  - Rad = Entitet

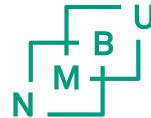
# Databasetabell eksempel



- Student tabell
  - Hver rad (entitet) er en **unik** student
  - Kolonnene er attributtene (egenskap) til hver student
  - Kan legge til restriksjoner (mer senere)

ElevNr	Fornavn	Etternavn	Fødselsdato	Kjønn
1	Ailin	Liane	09.10.2006	J
2	Gorm	Syrstad	13.05.2006	G
3	Ulf	Borgen	29.08.2006	G
4	Karina	Habbestad	02.12.2006	J
5	Anneli	Karlsen	19.06.2006	J

# Tabellen Ansatt



AnsattNr	Etternavn	Fornavn	AnsattDato	Stilling	Lønn
1	Veum	Varg	01.01.1992	Løpegutt	183 000.00
2	Stein	Trude	10.10.2000	DBA	270 700.00
3	Dudal	Inger-Lise	24.12.1988	Sekretær	299 000.00
4	Hansen	Hans	23.08.2006	Programmerer	325 000.00
5	Bjørnsen	Henrik	01.01.2000	Tekstforfatter	375 000.00
6	Gredelin	Sofie	18.05.1998	Underdirektør	625 850.00
7	Zimmermann	Robert	17.05.1995	Regnskapsfører	375 000.00
8	Nilsen				
11	Fosheim				
13	Lovløs				
16	Ibsen				
17	Fleksnes				
20	Felgen				
23	Karius	J.			
29	Wirkola	Gabriel	21.04.2009	Sekretær	255 000.00

**1 rad = 1 ansatt**

AnsattNr = 4

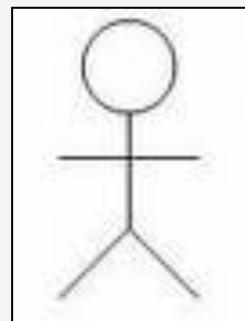
Etternavn = Hansen

Fornavn = Hans

AnsattDato = 23.08.2006

Stilling = Programmerer

Lønn = 325 000



# Databaser



- Vi har ofte flere tabeller
  - Kan grupperes i databaser (disk icon)

Studenter


Eksamensresultater


Rom


Bygg




Studentoversikt



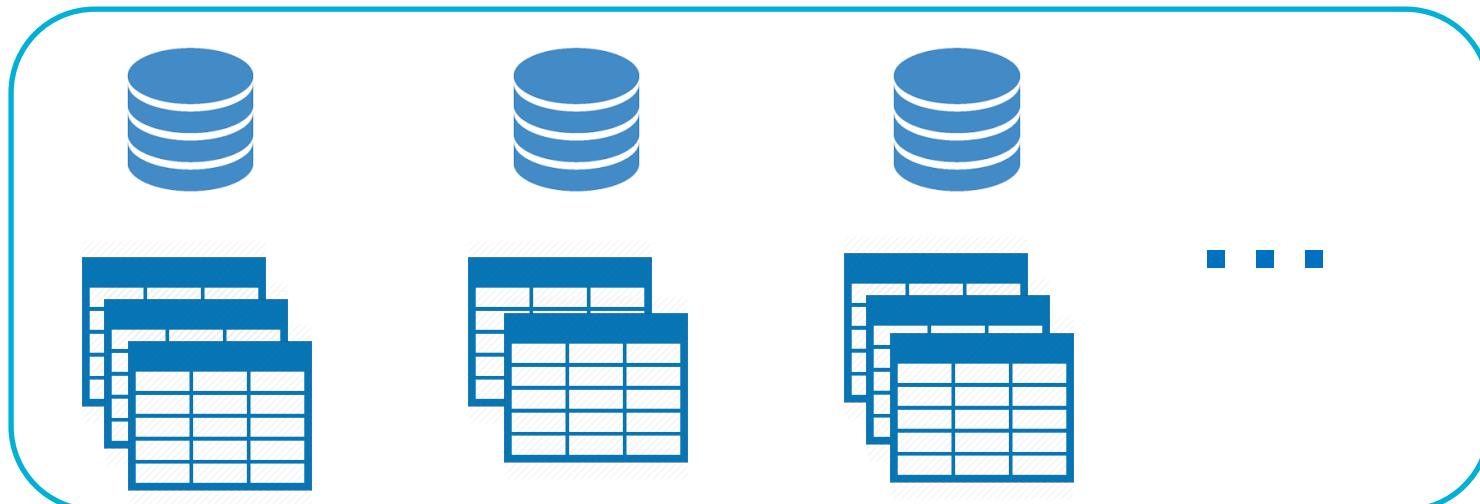
Eiendommer

# Databasesystem



- Database = **en logisk samling av data**
- Eksempel: En høgskole har flere databaser
  - Hver database kan inneholde mange **tabeller**
- Et databasesystem kan inneholde mange databaser

Databaser



# Flere eksempler

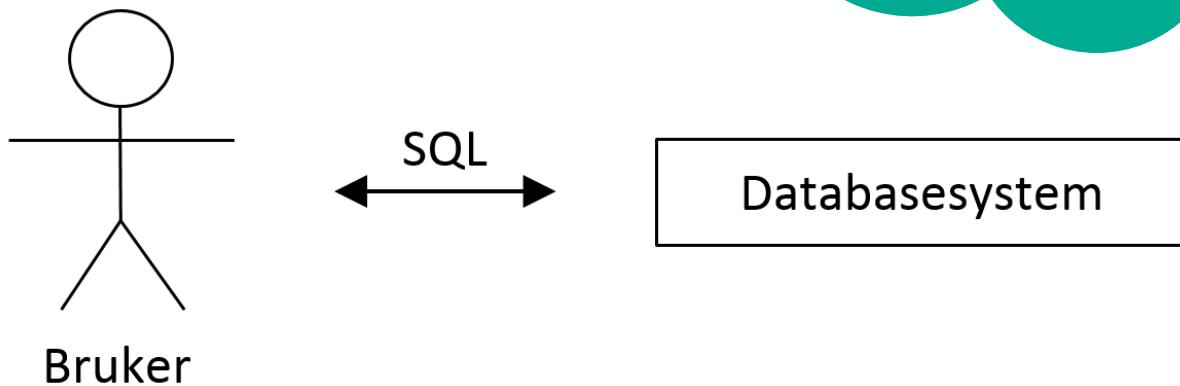
- **Handel**: Varer, kunder, bestillinger, leveranser, ...
- **Bibliotek**: Bøker, låntakere, lån, ...
- **Bank**: Kunder, kontoer, lån, innskudd, uttak, overføringer, ...
- **Sykehus**: Pasienter, journaler, medisiner, ansatte, turnus, ...
- **Kart**: Eiendommer, bygninger, veier, rørsystemer, ...
- **Kino**: Filmer, forestillinger, reservasjoner, ...
- **Forskning**: Spørreundersøkelse, respondenter, svar, ...
  
- Alle (?) virksomheter bruker databaser
- Mange systemer må være i drift **24/7** (virksomhetskritiske)
- Mange databaser er en del av et større **informasjonssystem**
  - Eksempel: Regnskapssystem, timeplansystem

# SQL-spørninger

Kommunikasjon med databaser

```
SELECT Fornavn, Etternavn  
FROM Elev  
WHERE Kjønn = 'J'  
ORDER BY Etternavn
```

SQL ligner litt på et programmeringspråk



# Koblinger mellom tabeller

- Det er ofte behov for koblinger mellom tabeller
- Eksempel
  - Et bygg har ofte flere rom

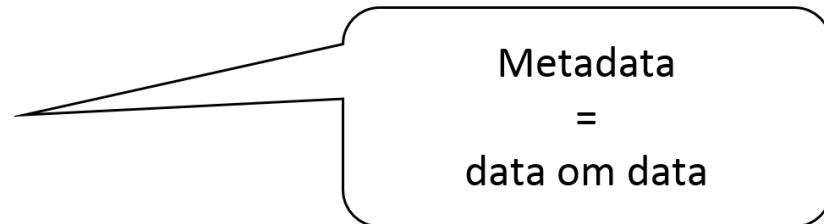
Rom


Bygg


# Metadata

- Metadata er «data om data».
- Metadata lagres også i databasen på tabellform, og kan avleses med SQL.

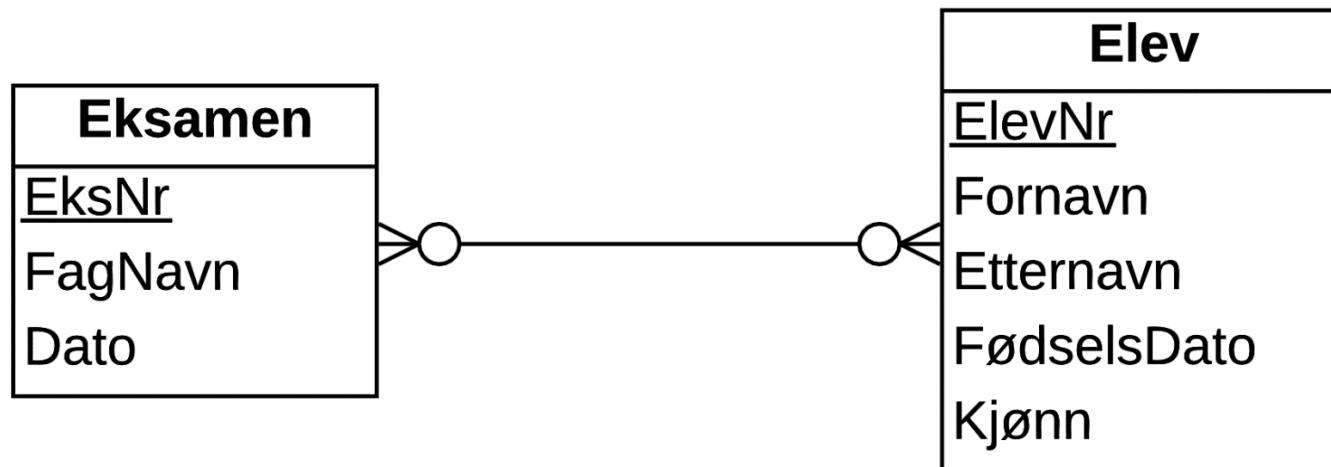
Tabell	Antall Rader
Kunde	507
Vare	1305
Ordre	5729
...	...



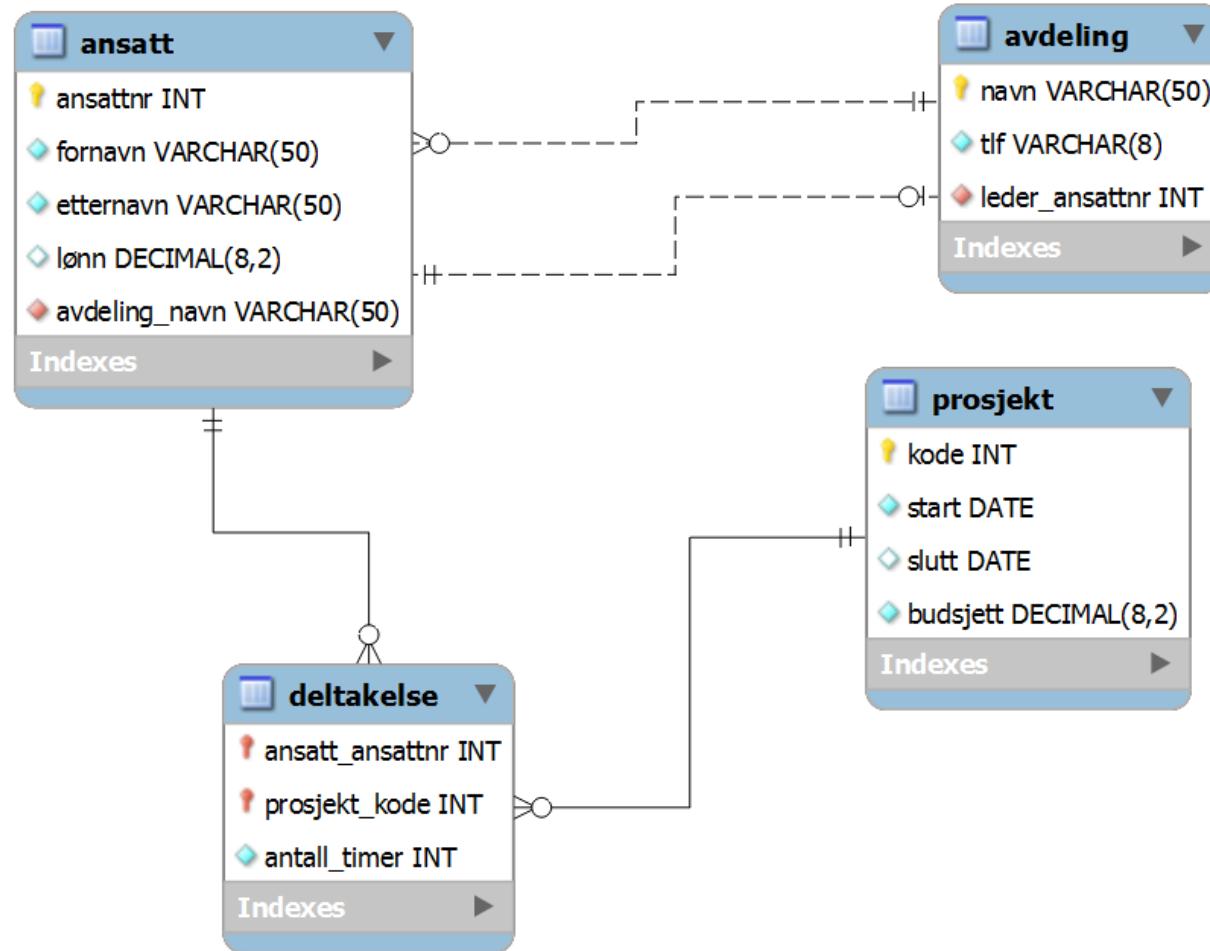
Metadata  
= data om data

# Databaser modellering

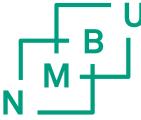
- Systemutvikling er komplisert og krever planlegging.
- Nyttig å lage modeller.
- En databasemodell er en **arkitekttegning** av systemet.
  - Kan generere SQL koder til å lage databaser og tabeller fra modellen



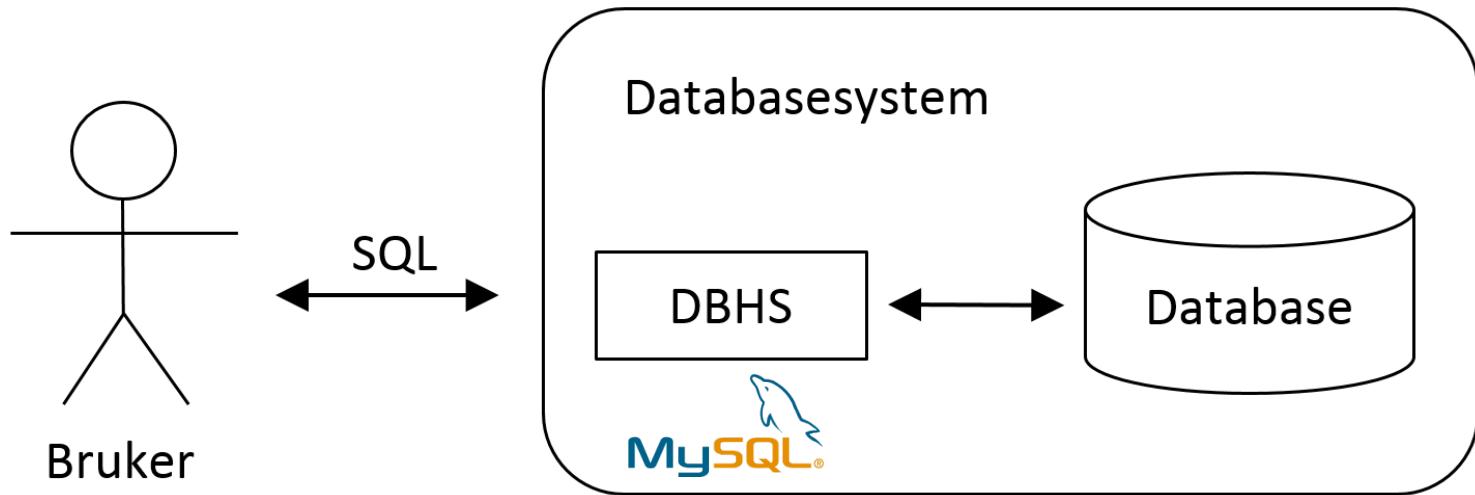
# Eksempel fra MySQL Workbench



# Databasesystem = DBHS + database

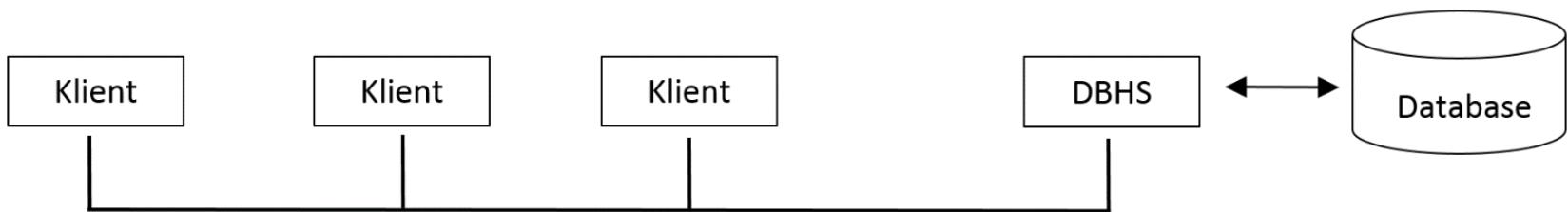


- DBHS = DataBaseHåndteringsSystem
- DBMS = Data Base Management System
- Eksempler: MySQL, Access, Oracle, PostgreSQL, SQL Server, ...
- SQL = språk for å jobbe med databaser (via DBHS)



# Databaseklienter i et nettverk

- En database er en **delt** ressurs.
- Flere klienter (brukere) kan være koblet opp til et databasehåndteringssystem via et nettverk.



# Databaseapplikasjoner

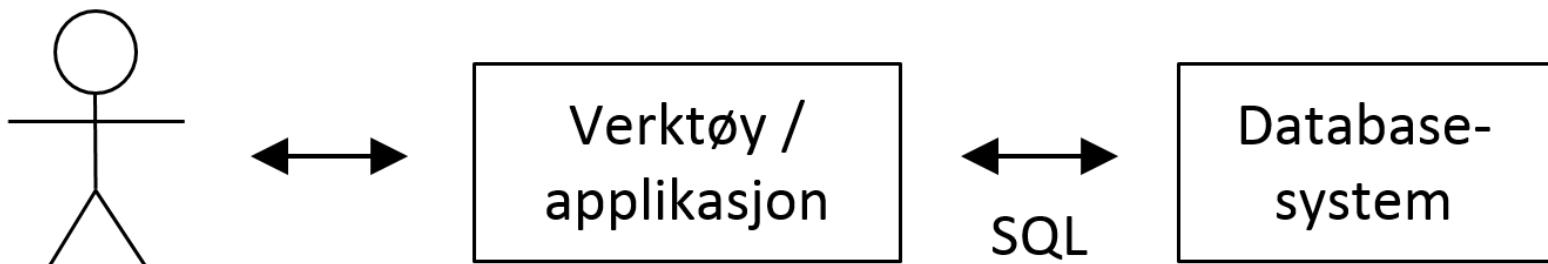
- Vanlig å jobbe med databaser via en applikasjon

The screenshot shows a Windows application window titled "Ordre". Inside, there's a form for an order with fields for OrderNr (20505), SendtDato (25.08.2015), OrdreDato (20.08.2015), BetaltDato (14.09.2015), and a customer section with fields for Kunde (5022), Name (Torgrim), and Address (Østbø). Below this is a table titled "Produkter på ordre" listing products with columns for VNr, Betegnelse, PrisPrEnhet, Antall, and Beløp.

Overlaid on the bottom right of the application window is a weather forecast from yr.no. It features a search bar with placeholder text "Søk blant 7 millionar værvarsel i Noreg og hele verda:" and a "VÆRSØK" button. Below the search bar, it says "Skriv stadnamn, f.eks. Stavanger, Røst eller Beijing." A red warning box displays "OBS-varsel: Vanskelige kjøreforhold i Vest-Finnmark fredag." with a warning icon. The forecast table shows weather details for Oslo, Bergen, and Stavanger across the days of the week (Fredag, Lørdag, Søndag) with icons and temperatures.

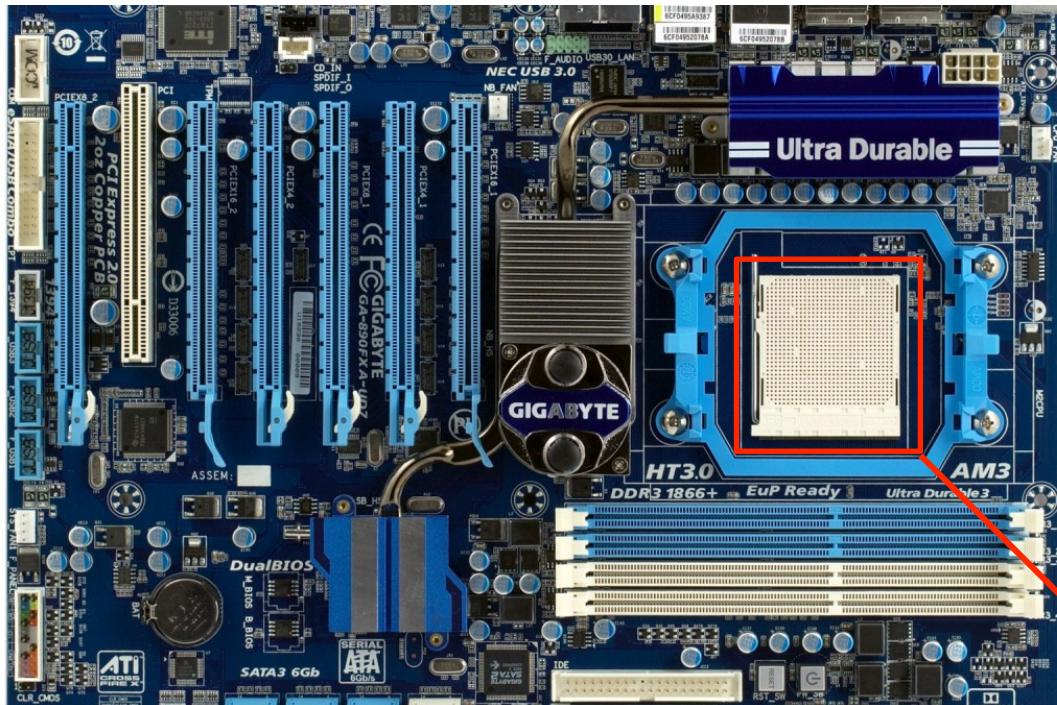
# Databasebrukere og verktøy

- Applikasjonen eller verktøyet danner et ekstra «**lag**» mellom brukeren og databasen.
- Applikasjonen/verktøyet kommuniserer med databasen (ved hjelp av SQL).



# Fra datamaskin til databaser

- Det er nyttig å vite om oppbygningen til datamaskinen for å få en bedre forståelse av databaser



Lagring

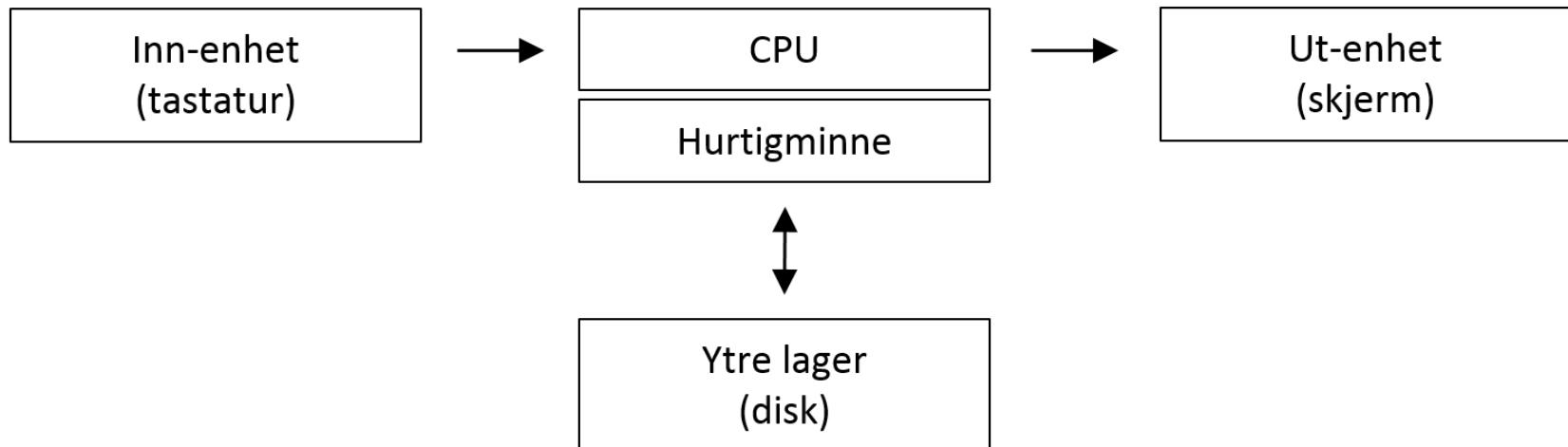


Prosessor (CPU)

# Prinsippskisse av en datamaskin

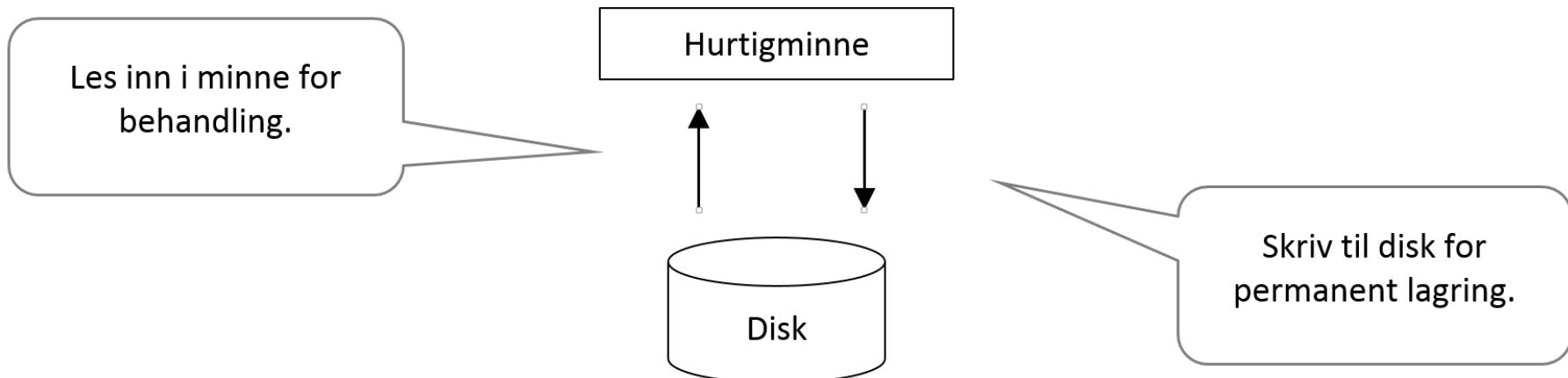
N M B U

- Fysisk er en datamaskin komplisert.
- Det er nyttig å lage seg et forenklet bilde (modell).

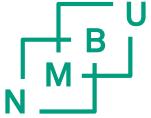


# Bruk av disk og hurtigminne

- Hurtigminnet er mye raskere enn disken.
- Data i hurtigminnet går tapt når maskinen skrus av.



# Oppbygning av minne/disk



Adresse	Verdi
A1	22
A2	58
A3	12
A4	232
A5	11
A7	43
.	.
.	.
.	.

- Minnet/Disk er en lang liste
- Hver adresse har plass til 1 byte
- 1 byte består av 8 bits
- 1 bit er enten 1 eller 0

0 0 0 0 0 0 0 = 0

1 1 1 1 1 1 1 1 = 255

# Representer tall



- 1 byte kan **tolkes** som heltallene [0..255]
- 2 bytes kan tolkes som heltallene [0..65 535]
- Kan også representer negative tall, men da reduseres absolutt maks til halvparten [-32 768..+32 767]
- Ethvert **desimaltall** kan representeres som to heltall:
  - Tallet 486.229 kan skrives som  $0.486229 \times 10^3$ .
  - Lar seg representere ved heltallene 486229 og 3.
  - Samme teknikk kan brukes på alle desimaltall.

# Representere tekst

- Et **tegnsett** tilordner et tall til hvert symbol.
  - Bokstaver, siffer, spesialtegn kan representeres som tall.
  - Med 2 byter kan vi representere 65 535 symboler.
  - En **tekststrenge** er en sekvens av tegn.
  - ASCII og **UTF-8** er to eksempler på tegnsett.

Tegn	Kode	Tegn	Kode
A	65	æ	145
Z	90	Æ	146
a	97	!	33
z	122	=	61
0	60	?	63
9	71	@	64

# ASCII

# Tall til tekst eksempel

Adresse	Verdi
A1	72
A2	97
A3	108
A4	108
A5	111

H

a

|

|

O



# Digital tekst-data

- En vanlig standard er UTF-8
  - Kan representere 1,112,064 forskjellige tegn
- Det er viktig å benytte samme standard i databasen som det er i teksten som skal importeres.
- Det er alltid feil valg av tegntabell som forårsaker ØÆÅ problemer
- Dersom norske tegn blir vist feil kan ISO 8859-1 benyttes

# Fra 0 og 1 til databasetabeller

- Hver celle i en databasetabell inneholder en **verdi**:
  - Heltall
  - Desimaltall
  - Tekststreng
  - Sannhetsverdi (true/false)
  - Dato/klokkeslett
- Vi har sett hvordan slike verdier kan representeres som heltall, som igjen kan representeres som 0 og 1.
- En **rad** i en databasetabell kan lagres som en sekvens av verdier.
- Og **databasetabeller** kan lagres rad for rad...

# Datatyper

- Velg riktig datatype for attributtene(kolonnene)
  - Data tar mindre plass
  - Raskere søk i databasen

**INT (Heltall)**    **Varchar (tekst)**                      **Date**                      **Char (tekst)**

ElevNr	Fornavn	Etternavn	Fødselsdato	Kjønn
1	Ailin	Liane	09.10.2006	J
2	Gorm	Syrstad	13.05.2006	G
3	Ulf	Borgen	29.08.2006	G
4	Karina	Habbestad	02.12.2006	J
5	Anneli	Karlsen	19.06.2006	J

# En databasetabl

<b>AnsNr</b>	<b>Fornavn</b>	<b>Etternavn</b>
11	Varg	Veum
12	Ada	Lovløs
13	Jon	Nymann

# Mulig fysisk organisering

1024	11
1025	v
1026	A
1027	r
1028	g
...	...
1045	v
1046	e
1047	u
1048	m
...	...
1085	12
...	...

Rad 1 starter på  
adresse 1024.

Rad 2 starter på  
adresse 1085.

# "Null" merker

- Legg merke til at det mangler noen verdier i tabellen Prosjekt. Vi kaller dette for **nullmerker**.
- Oppstår fordi:
  - Vi har glemt å registrere data.
  - Vi kjenner ikke til den korrekte verdien.
  - Det gir ikke mening å registrere data.
- Nullmerker er *ikke* verdier.
- Nullmerker kan skape problemer!

OBS: Uttal "Null" på engelsk

---

# Måleenheter for datamengder

Måleenhet	Verdi	IT-bruk	Binær måleenhet	Verdi
kilobyte (kB)	$10^3$	$2^{10}$	kibibyte (KiB)	$2^{10}$
megabyte (MB)	$10^6$	$2^{20}$	mebibyte (MiB)	$2^{20}$
gigabyte (GB)	$10^9$	$2^{30}$	gibibyte (GiB)	$2^{30}$
terabyte (TB)	$10^{12}$	$2^{40}$	tebibyte (TiB)	$2^{40}$
petabyte (PB)	$10^{15}$	$2^{50}$	pebibyte (PiB)	$2^{50}$
exabyte (EB)	$10^{18}$	$2^{60}$	exbibyte (EiB)	$2^{60}$
zettabyte (ZB)	$10^{21}$	$2^{70}$	zebibyte (ZiB)	$2^{70}$
yottabyte (YB)	$10^{24}$	$2^{80}$	yobibyte (YiB)	$2^{80}$

# Øvinger

- Oppgave 1
  - Gjør deg kjent med programvare og grunnleggende database (se Canvas)
    - Installer programvare på din egen datamaskin
  - Plenumsgjennomgangen på mandag 5 februar