**Checkpoint 2**Eivind Johansen

**Level 1**

**What is a data warehouse? What is a data mart?**

I et datavarehus blir data fra ulike kilder samlet for å gi oss muligheten til å analysere på tvers av avdelinger i bedriften. Data blir kjørt igjennom en ETL prosess for å strukturere og rense tallene slik at vi får tall med samme format som kan aggregeres over. Det finnes 2 ulike fremgangsmåter. Top-down, der vi starter med å samle data i små data-marts først før vi samler alle data-martsene til et datavarehus og bottom-up der vi starter med å samle all data inn i et datavarehus før vi lager ulike data-marts som ser på mindre slicer av datavarehuset. Datavarehus har strukturert data som er optimalisert for komplekse spørringer (DQL). All data er knyttet til tid og ved oppdateringer blir ikke rader endret, men heller lagt til for å sørge for å beholde historikken i datavarehuset. I et datavarehus kan vi drille ned dataen ved hjelp av dimensjoner som f.eks. tid, lokasjon etc. Det vi får gruppert og aggregert ulike fakta som vi ønsker å se, som f.eks. salg i en butikk.

En data-mart er et subset av et datavarehus som ser på et fokusområde av datavarehuset.

**In the context of dimensional modelling: what is a fact, and what is a dimension?**

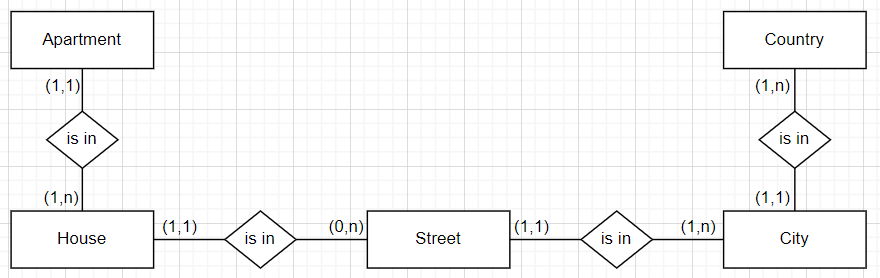
**Fact:** Er alle fakta/tall vi ønsker å aggregere. Dette kan være f.eks. salg, lønn, antall sengeplasser fylt på et sykehus eller lignende.   
**Dimension:** Dimensjoner er «kategorier» som vi ønsker å drille/gruppere facts’ene ned på. Dette kan være tid (år, måned, dag, time etc.), lokasjon (land, by, gate etc.), personer (artister, leger, idrettsutøvere) eller lignende.

**What are the advantages of normalization (using Normal Forms)? Are there any disadvantages?**Fordelene med å normalisere er at det er mye enklere og raskere å oppdatere databasen, det forhindrer redundans av data, forhindrer NULL-verdier i tabellen og gir oss en bedre integritet i dataen i databasen.

Men høy grad av normalisering øker kompleksiteten i databasen og kan derfor gjøre at spørringer går tregt.

**E-R modelling**

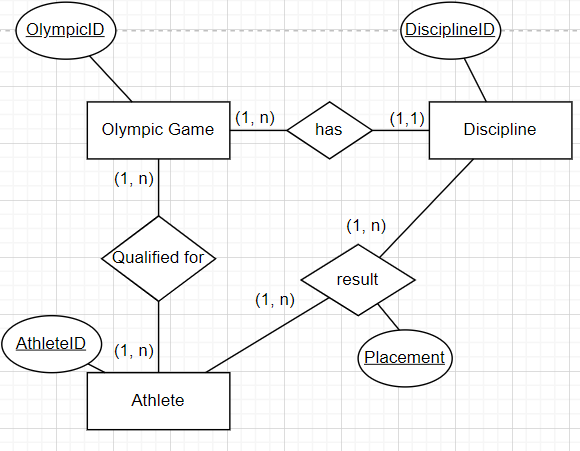
**Level 1: Short and sweet**



Vil ikke kalle noen av disse entitetene for en svak entitet da de alle kan stå for seg selv og må til for at vi skal kunne pin-pointe den eksakte leiligheten.

**Olympic Games**

Setter placement som en unik id i junction-tabellen mellom athlete og discipline som i kombinasjon med nøklene fra athlete og discipline vil sørge for at kun 1 gull, 1 sølv og 1 bronse eksisterer i hver disiplin.

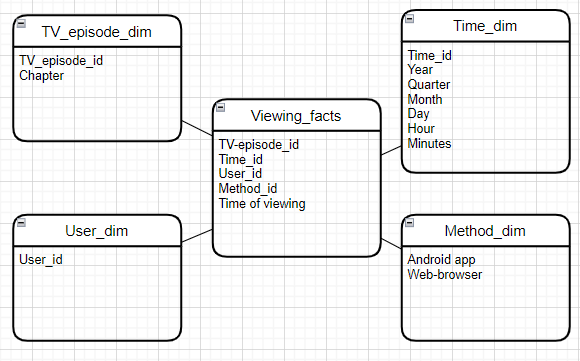


**Bank**

**Bank**(Code, Name, Addr)  
**Bank\_branch**(Branch\_no, Addr, Code)  
Code er FK fra Bank  
**Account**(Acct\_no, Code, Balance, Type)  
Code er FK fra Bank  
**Loan**(Loan\_no, Code, Amount, Type)  
Code er FK fra Bank  
**A\_C**(A\_C\_id, Acct\_no, Ssn)  
Acct\_no er FK fra Account, Ssn er FK fra Customer  
**L\_C**(L\_C\_id, Loan\_no, Ssn)  
Loan\_no er FK fra Loan, Ssn er FK fra Customer  
**Customer**(Ssn, Name, Phone, Addr)

**Dimensional modelling**

**Level 1: Netflix**



Number of viewings får vi igjennom å telle antall linjer i viewing\_facts.  
Method of viewings får vi igjennom å telle method\_id i viewing\_facts.

Tar med Time\_dim granulert ned på minutter selv om det står at chapter er fineste granularitet. Tolker det at de vil ha quarter gjør at de vil se en time dimension og for meg er det logisk å ha det ned på minutter i denne sammenhengen.