Databanken - les 1

joost.vennekens@kuleuven.be

http://www.cs.kuleuven.be/~joost/DN/

Onderwerp

- Gegevens
  - Hoe bijhouden?
  - Hoe gebruiken?
- Motivatie
  - Gegevens zijn overal!
  - Databanken zijn overal!

Biochemische testen

- Honderd verschillende gistsoorten
- Uittesten in diverse omstandigheden
- Op zoek naar aromaproductie



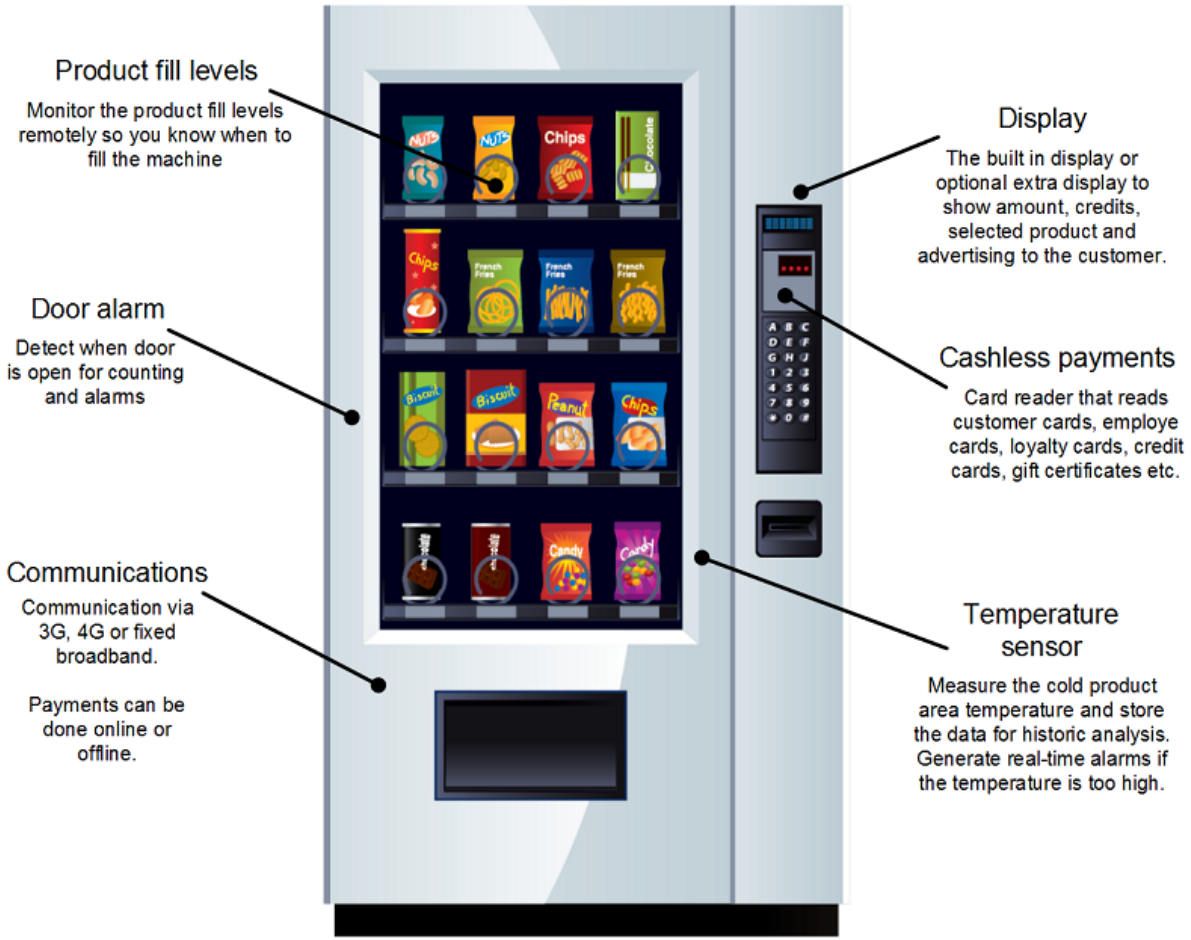
Productieplanning

- Grondstoffen nodig voor produkt
- Hoeveelheid in voorraad
- Bestellingen



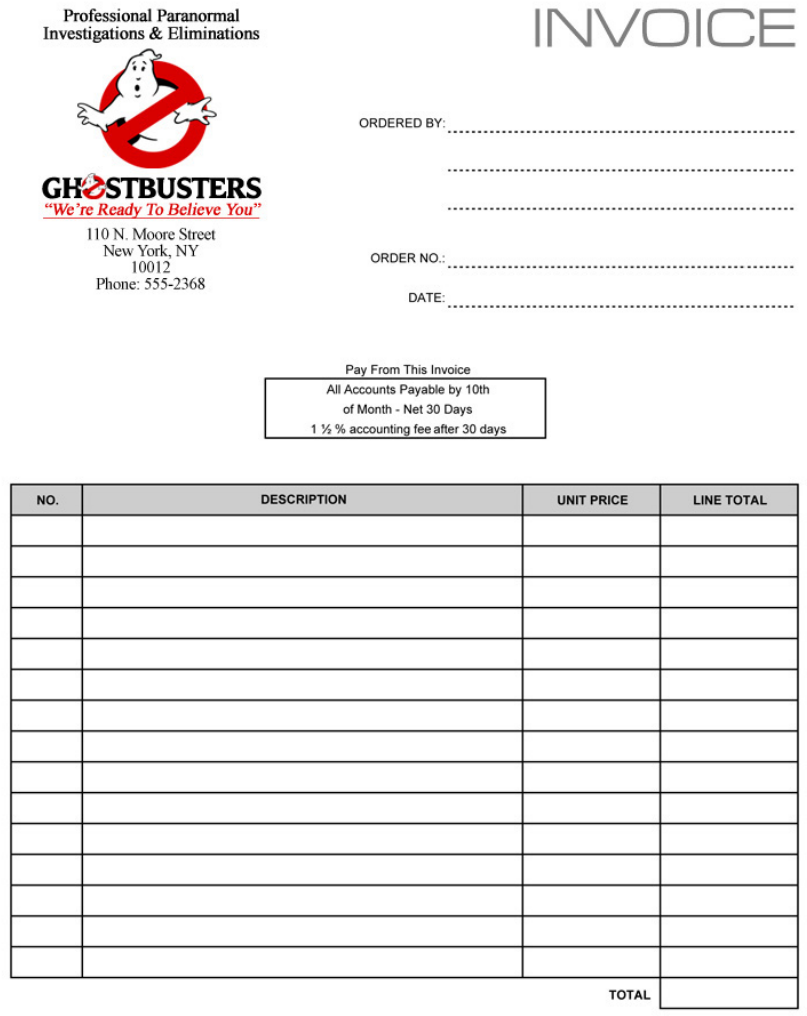
Smart Vending Machines

- Verkoopscijfers
- Omgevings-factoren
- Pro-actief voorraadbeheer



Klantenbestand

- Klantgegevens
- Facturen
- Betalingen



Professional Paranormal
Investigations & Eliminations

GHOSTBUSTERS
"We're Ready To Believe You"

110 N. Moore Street
New York, NY
10012
Phone: 555-2368

INVOICE

ORDERED BY: .....

.....

.....

ORDER NO.: .....

DATE: .....

Pay From This Invoice
All Accounts Payable by 10th
of Month - Net 30 Days
1 ½ % accounting fee after 30 days

NO.	DESCRIPTION	UNIT PRICE	LINE TOTAL
		TOTAL	

Hoe bijhouden?

- Moeilijk om specifieke gegevens terug te vinden
- Moeilijk om te onderhouden
- Risico op gegevensverlies

	A	B	C
1	Gist	t1-pH	t1-tmp
2	g132	0,799	17,6
3	g543	0,763	18,7
4	g322	0,87	16,5
5	g565	0,817	18,7
6	g544-2	0,81	17,4
7	g544-1	0,59	17,5



The Washington Post

Wonkblog

An alarming number of scientific papers contain Excel errors

By Christopher Ingraham August 26

A surprisingly high number of scientific papers in the field of genetics contain errors introduced by Microsoft Excel, according to an analysis recently published in the journal Genome Biology.

A team of Australian researchers analyzed nearly 3,600 genetics papers published in a number of leading scientific journals — like Nature, Science and PLoS One. As is common practice in the field, these papers all came with supplementary files containing lists of genes used in the research.

The Australian researchers found that roughly 1 in 5 of these papers included errors in their gene lists that were due to Excel automatically converting gene names to things like calendar dates or random numbers.

You see, genes are often referred to in scientific literature by symbols — essentially shortened versions of full gene names. The gene "Septin 2" is typically shortened as SEPT2. "Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase" gets mercifully shortened to MARCH1.

But when you type these shortened gene names into Excel, the program automatically assumes they refer to dates — Sept. 2 and March 1, respectively. If you type SEPT2 into a default Excel cell, it magically becomes "2-Sep." It's stored by the program as the date 9/2/2016.

Even worse, there's no easy way to undo this automatic formatting once it has happened. Edit -> Undo simply deletes everything in the cell. You can try to convert the formatting from "General," the default, to "Text," which you might expect to change it back to the original characters you enter. But instead, changing the formatting to "Text" makes the cell contents appear as 42615 — Excel's internal numeric code referring to the date 9/2/2016.

Even more troubling, the researchers note that there's no way to permanently disable automatic date formatting within Excel. Researchers still have to remember to manually format columns to "Text" *before* you type anything in new Excel sheets — every. single. time.

But even the genetics researchers among us are only human, and they sometimes forget to do this. Hence, you end up with 20 percent of these genetics papers containing preventable errors introduced by Excel.

Databank

- Veilige opslag
- Risico op fouten vermijden
- Vlot & krachtig gegevens opvragen

Doelstelling

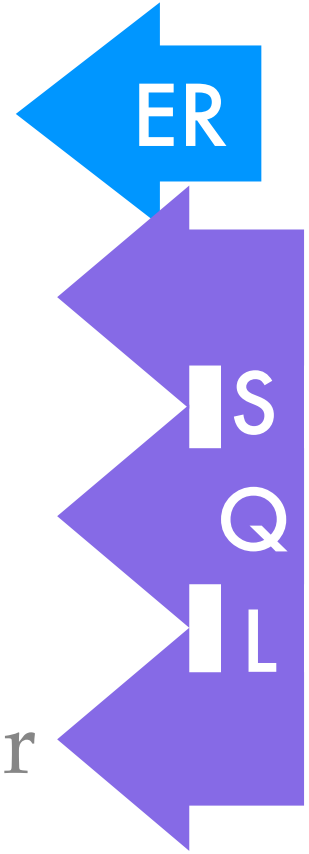
- Databank maken en gebruiken
  - Niet alleen relevant voor ITers
- Achterliggende concepten
  - Communicatie: domeinexpert <> IT
    Terminologie, aandachtspunten
  - Wat is mogelijk, moeilijk, gemakkelijk?

Evaluatie

- 75% examen, gesloten boek
  - 1/3 theorie (mondeling)
  - 2/3 oefeningen (schriftelijk)
- 25% permanente evaluatie in practica

Een DB: hoe doe je dat?

- Installeer een DBMS
- Ontwerp een databank
- Maak een databank
- Steek gegevens in databank
- Haal er weer uit wanneer nodig





Entiteit

individueel object

Relatie

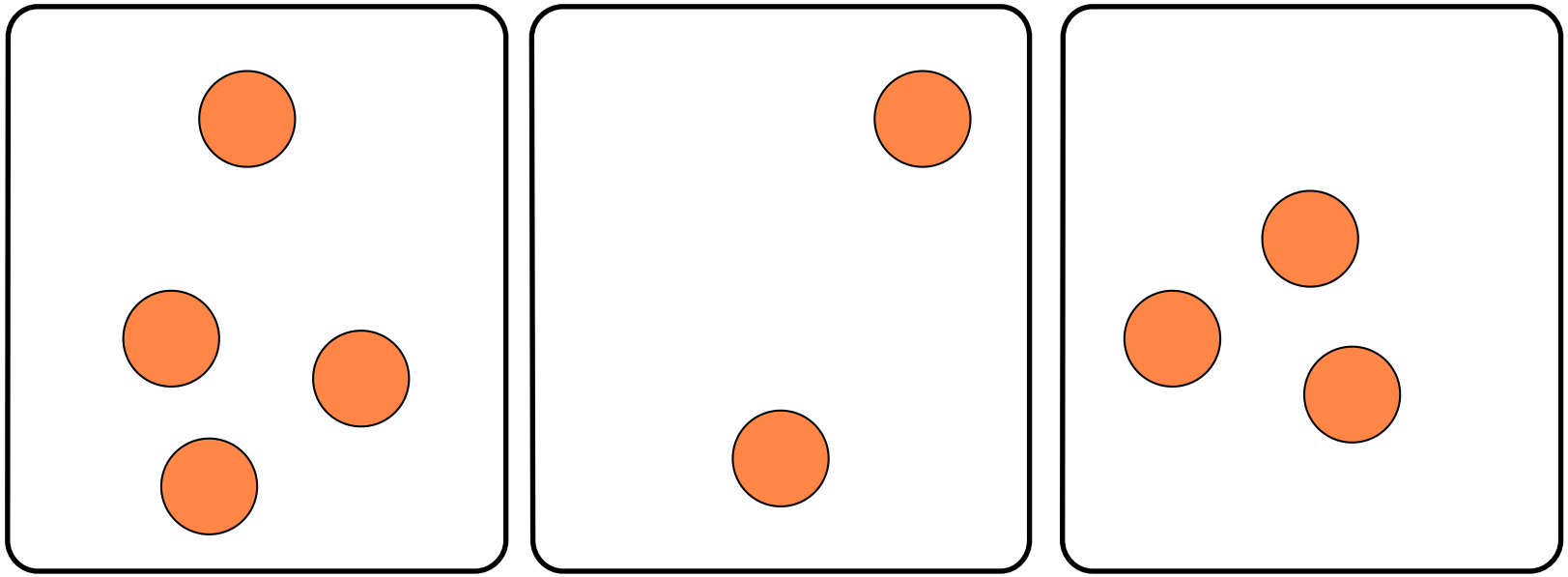
verband tussen entiteiten

Attribuut

eigenschap van entiteit/relatie

Entiteiten

- Elke entiteit is individueel iets
- Entiteiten kunnen worden opgedeeld in soorten: **entiteitverzameling**



- Alle entiteiten in een EV hebben zelfde attr.

Attributen

Samengesteld ⟷ Enkelvoudig

adres schoenmaat

Meerwaardig ⟷ Enkelwaardig

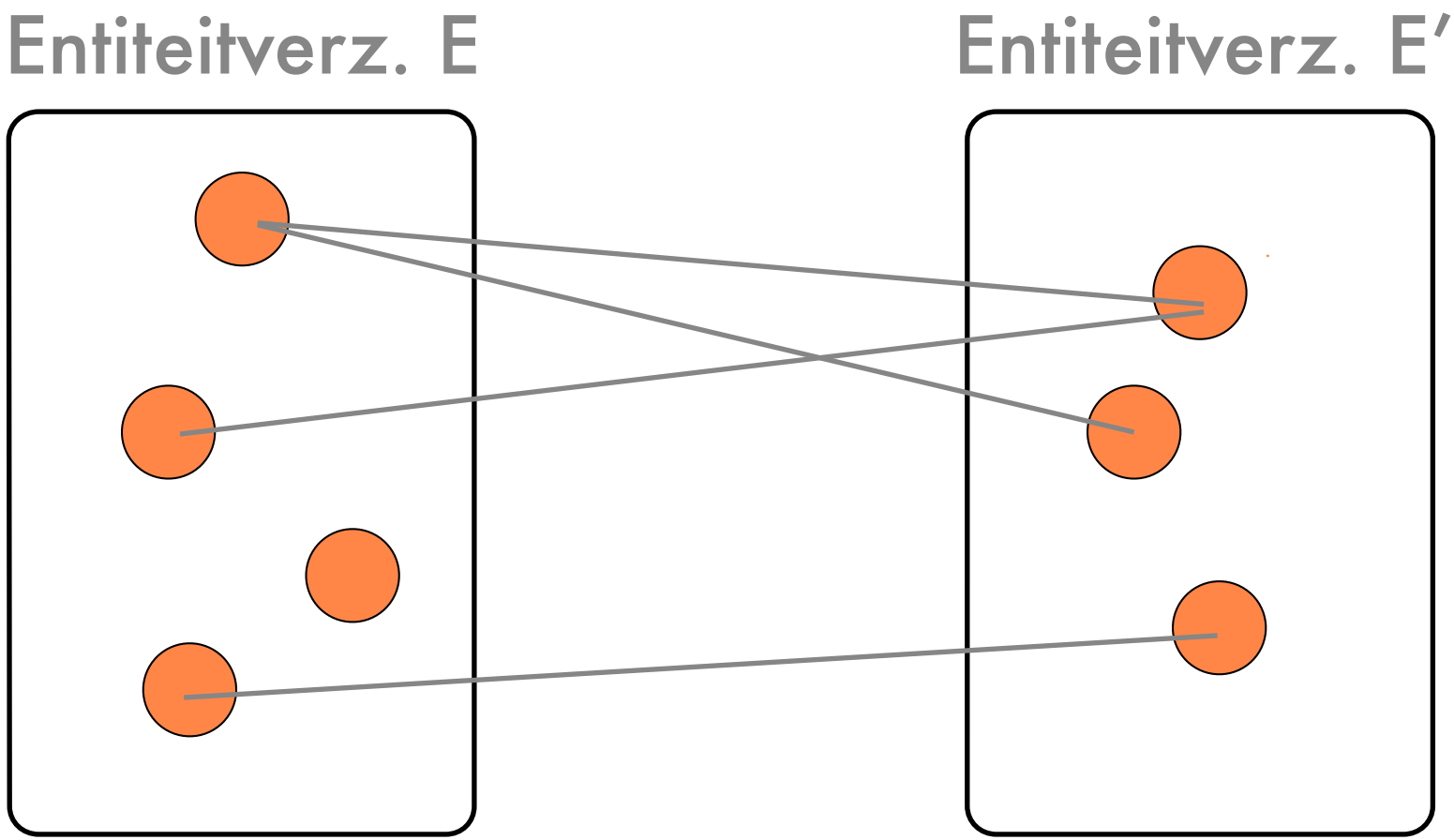
Attributen

- Afleidbaar
  - Attribuut -> Attribuut

    Geboortedatum -> leeftijd
  - Relaties -> Attribuut

    Examenresultaten -> studie-efficiëntie
- Soms is "**null**" toegelaten als waarde

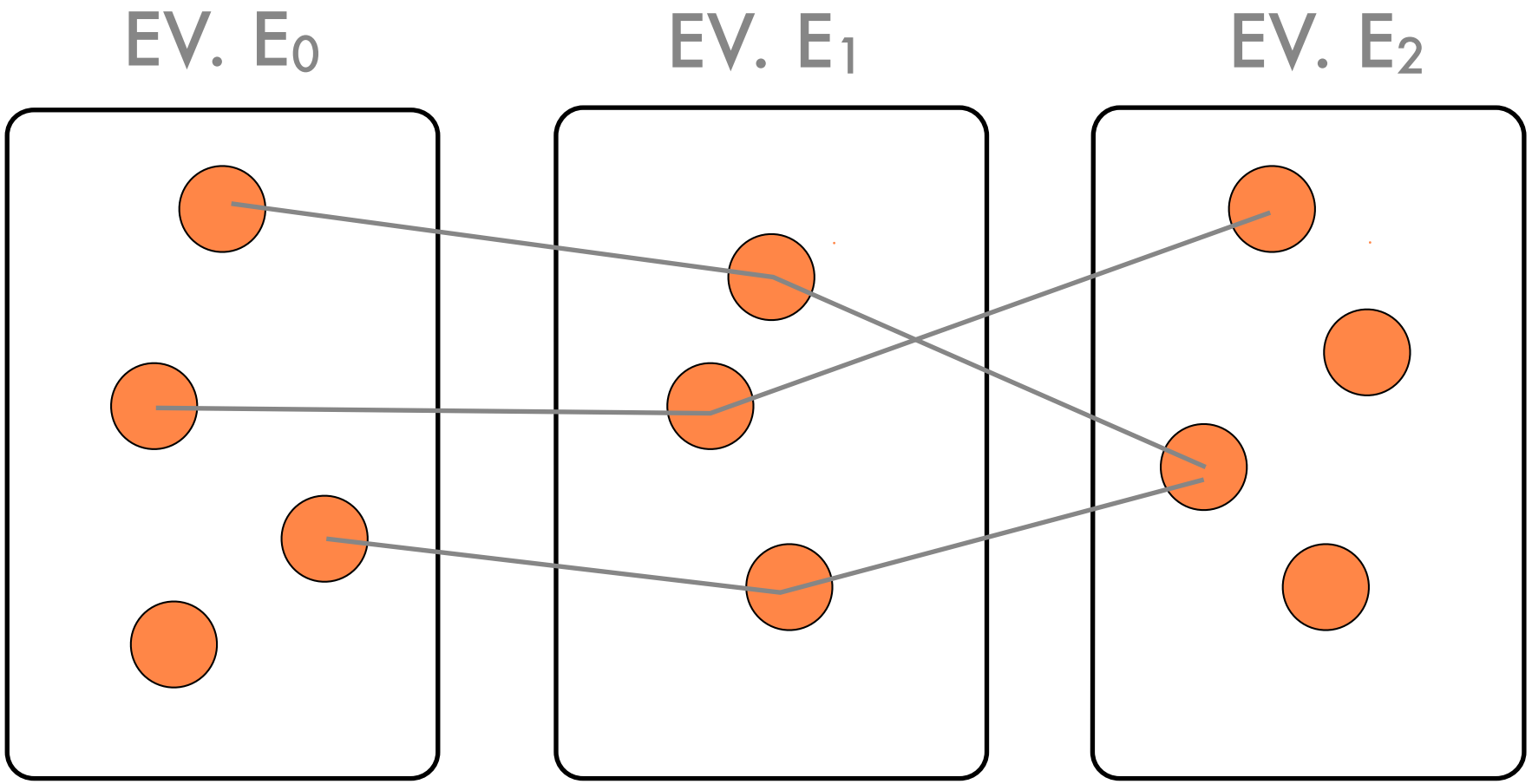
Relaties



$$R \subseteq \{(e, e') \mid e \in E, e' \in E\}$$

Cartesisch produkt E x E′

Relaties

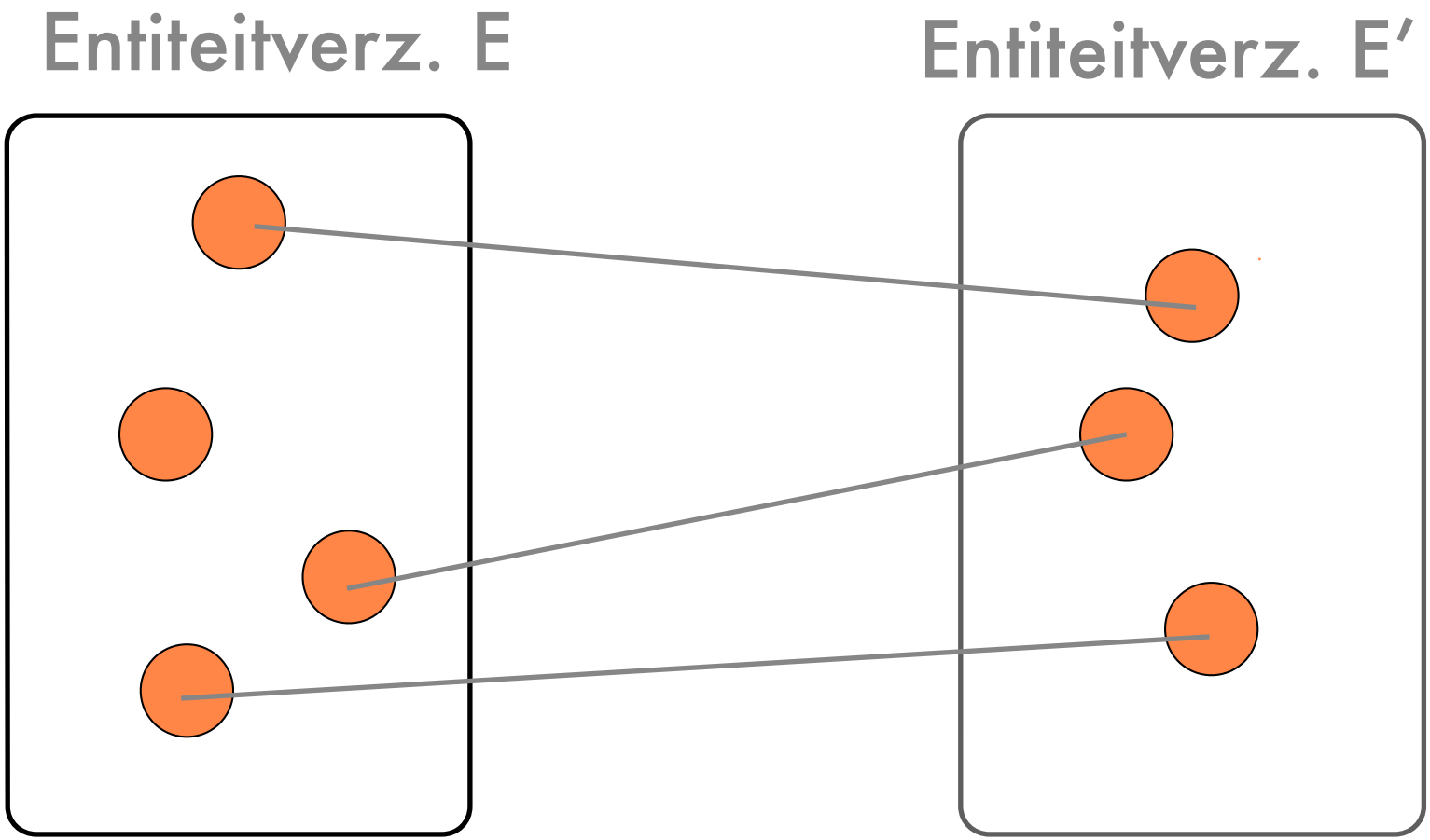


$$R \subseteq \{(e_0, \ldots, e_n) \mid e_0 \in E_0, \ldots, e_n \in E_n\}$$

Integriteitsbeperkingen

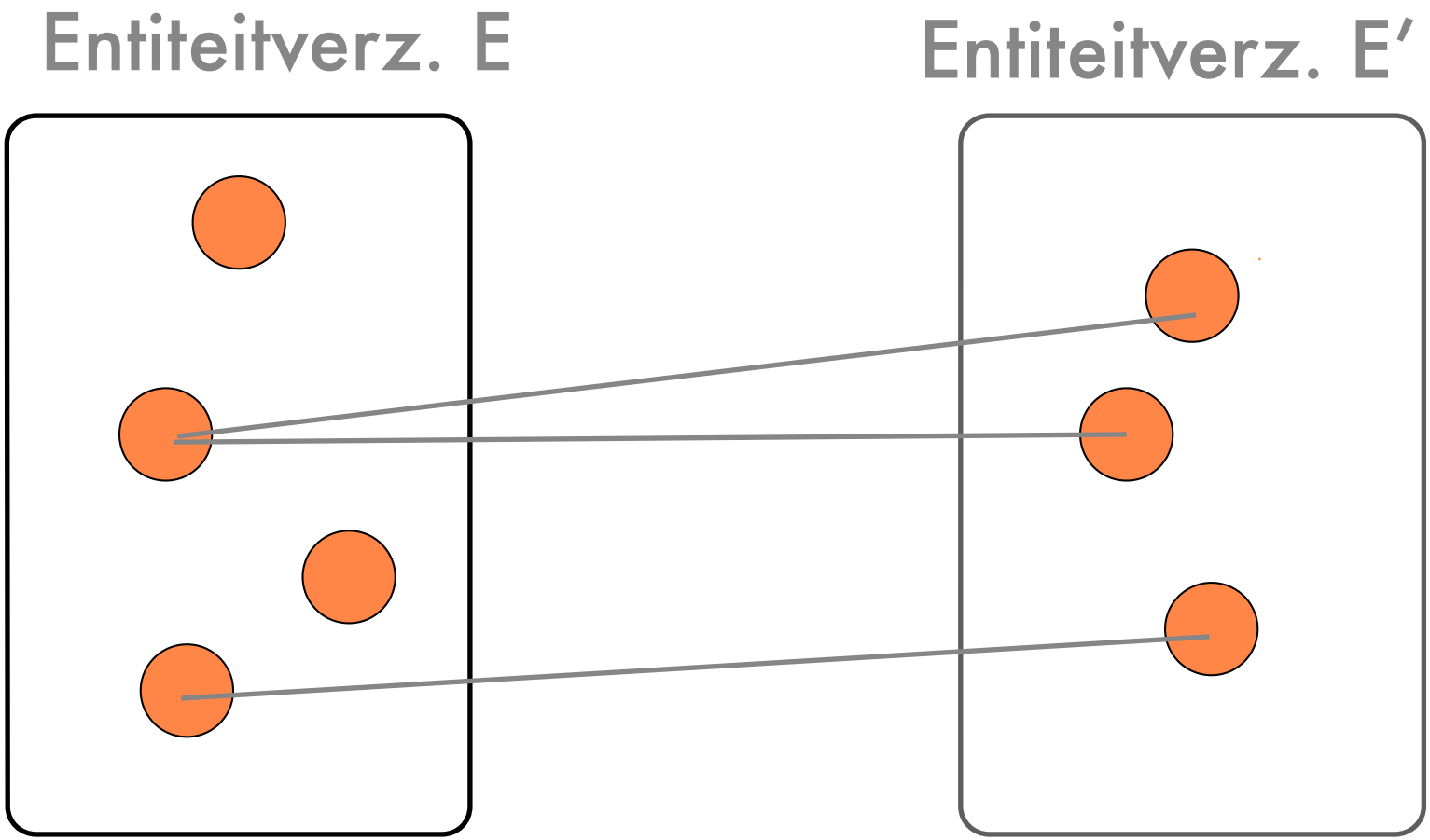
- Voorwaarden waaraan gegevens moeten voldoen
- Bestaansbeperking
  - Voorwaarde op een attribuut
    (bv. 30 <= schoenmaat <= 50)
- Mapping cardinaliteit
  - Voorwaarde op relaties
- Bestaansafhankelijkheid
- Sleutelbeperkingen

Een-op-Een



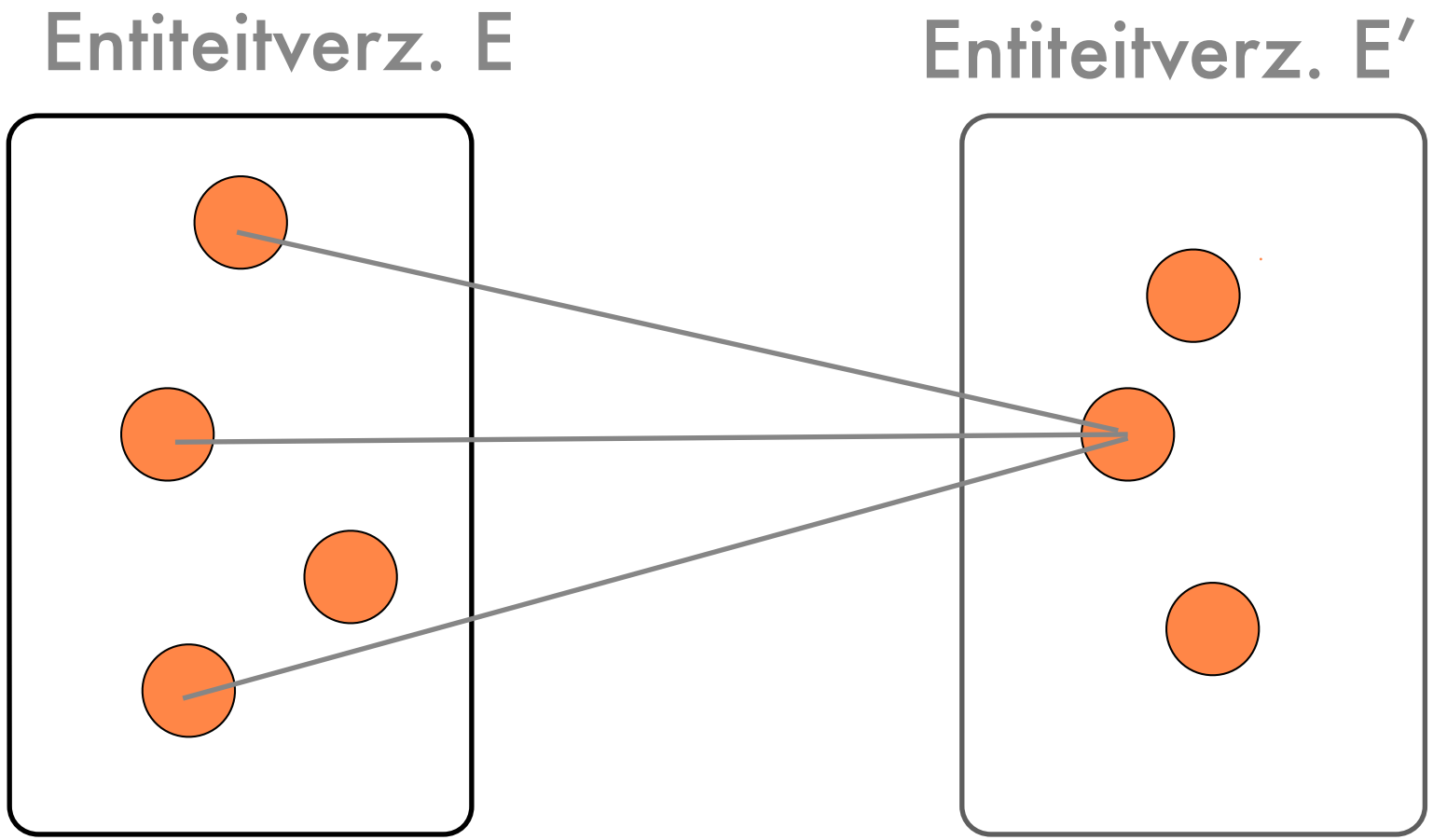
man is getrouwd met vrouw

Een-op-Veel



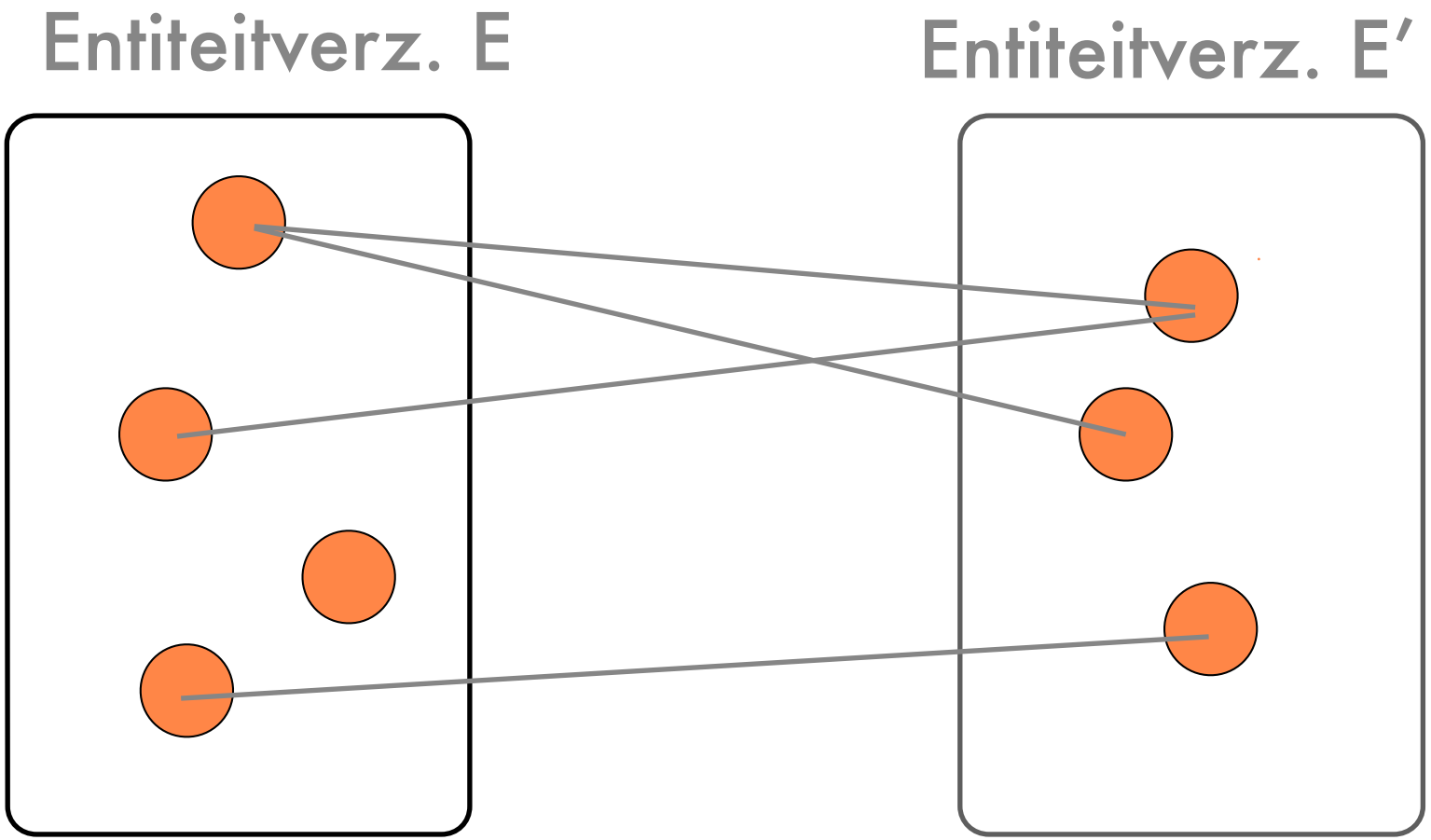
huis bevat kamer

Veel-op-Een

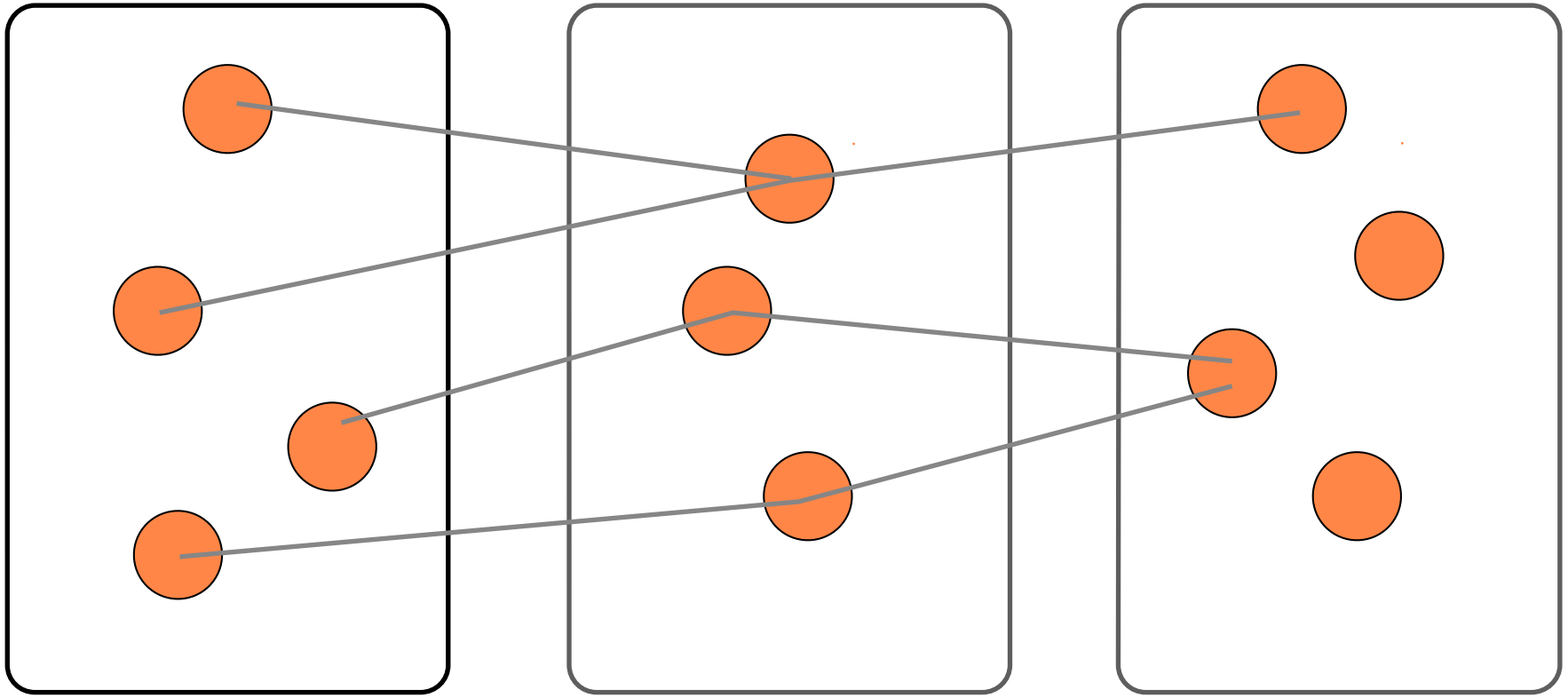


persoon stemt voor partij

Veel-op-Veel



Veel-op-Een-op-Een



kind heeft man en vrouw als ouder

Bestaansafhankelijkheid

- Een huis bevat veel kamers
- Elke kamer moet in een huis liggen
  - Als huis verdwijnt, verdwijnen alle kamers mee
  - Een kamer kan verdwijnen zonder dat het huis verdwijnt

ondergeschikte entiteit

dominante entiteit

Sleutels

- Elke entiteit is afzonderlijk object
- We hebben een manier nodig om entiteit uniek te identificeren
  - Supersleutel: verzameling van attributen die entiteit uniek bepalen
  - Kanditaat sleutel: zo kleine mogelijke supersleutel
  - Primaire sleutel: keuze uit de kandidaatsleutels

Sleutels - vbn

- Politieke partij: naam, afkorting, strekking, ledenaantal
- Huis: aantal kamers, bewoonbare oppervlakte, straat, nummer, gemeente
- Hond: naam, kleur, ras, geboortejaar

Zwakke entiteiten

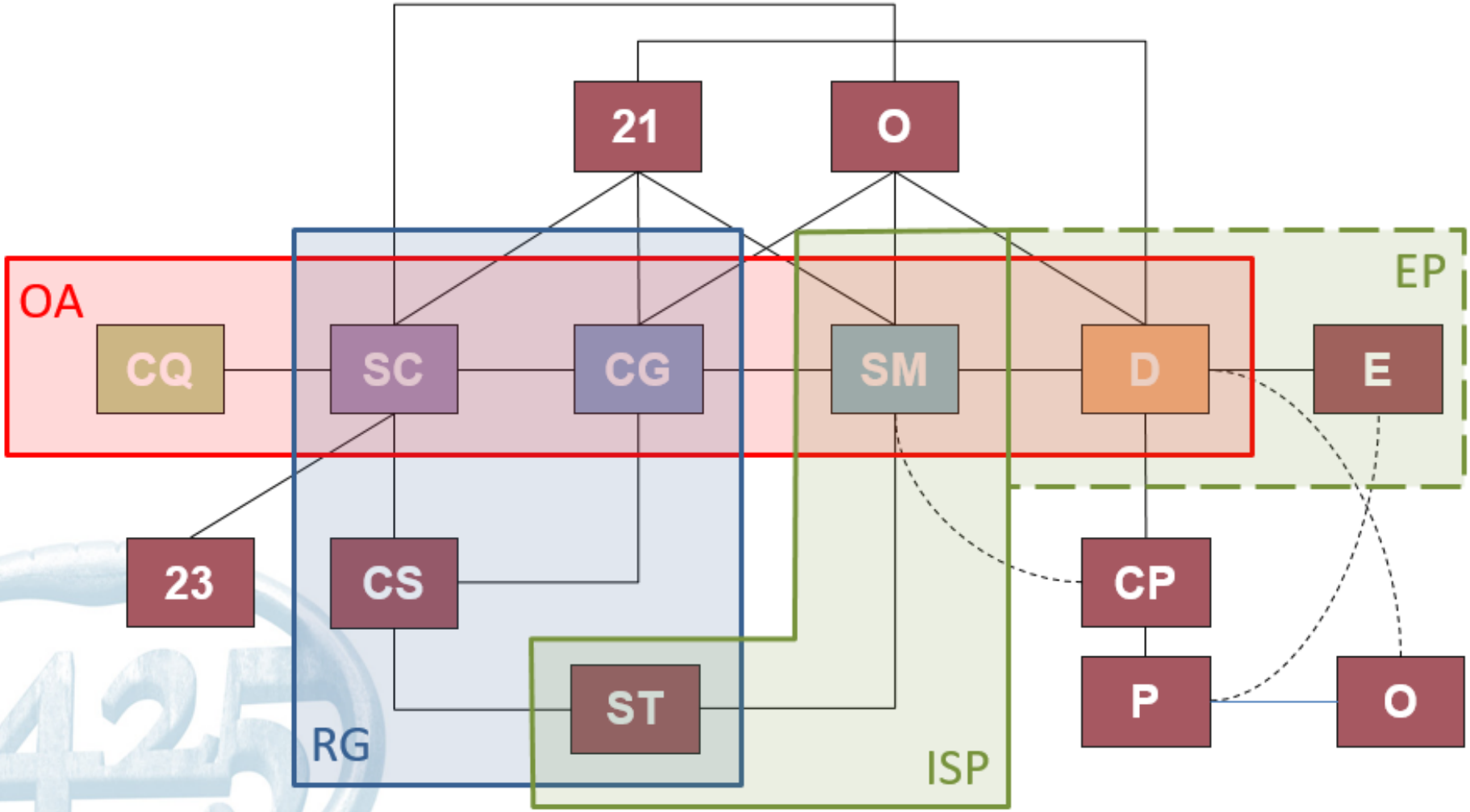
- Sterke entiteit heeft een primaire sleutel
- Zwakke entiteit heeft zelf geen primaire sleutel
  - Maar kan toch geïdentificeerd worden door relatie met sterke entiteit
  - Discriminator: attributen die onderscheid toelaten
- Bv: Eigenaar + naam identificeert hond
- Sterk -> dominant; zwak -> ondergeschikt

ER Diagramma

Stap 1

- Welke gegevens moeten in DB?
  - Welke entiteiten?
  - Welke relaties tussen die entiteiten?
  - Welke attributen hebben die entiteiten/ relaties?
  - Welke integriteitsbeperkingen?
- Diagramma volgens **conventies**: Overzichtelijk en éénduidig

Vereenvoudigd datamodel

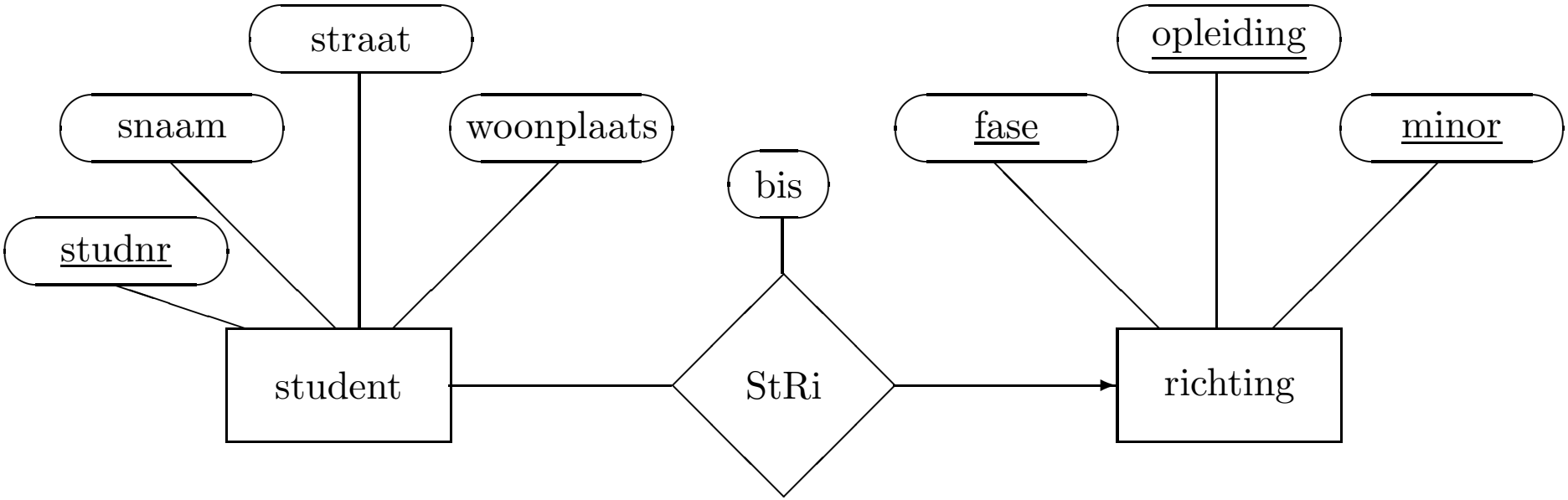


21 POC - O Faculteit - CQ diploma - SC Opleiding - CG Modulegroep - SM OPO - D OLA
E Evenement - 23 Akkoord - CS Studie - ST Student - CP Central Person - P Persoon

ER-diagramma

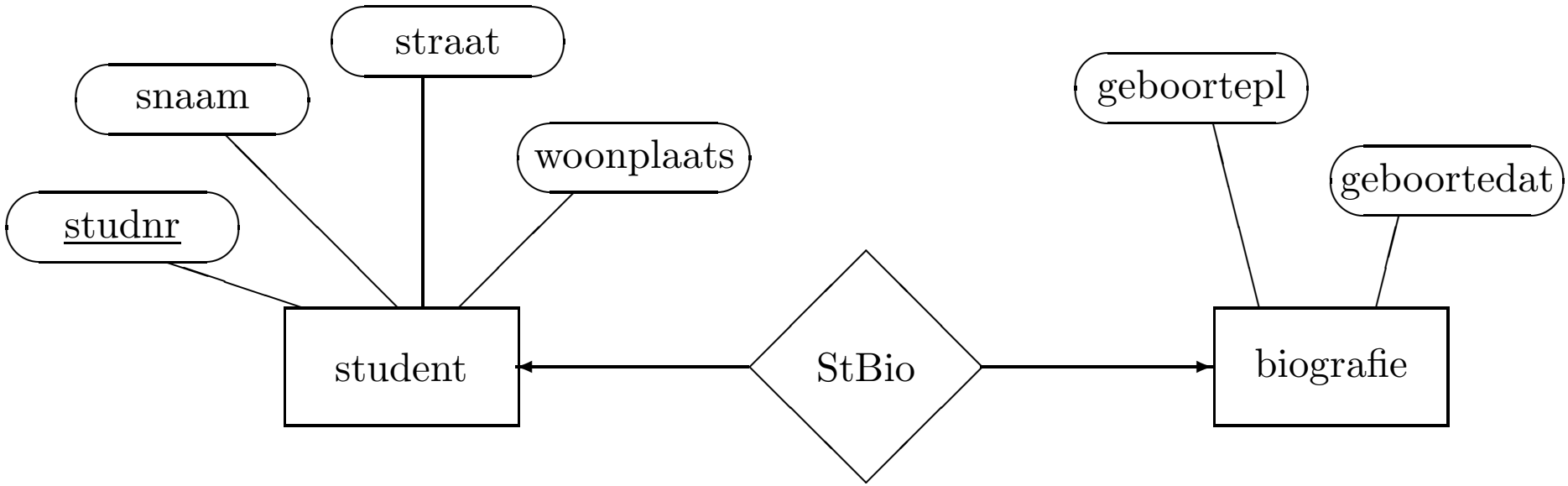
- Grafische voorstelling van
  - Entiteitverzamelingen (rechthoek)
  - Attributen (ovaal)
  - Relaties (ruit)

Bijvoorbeeld



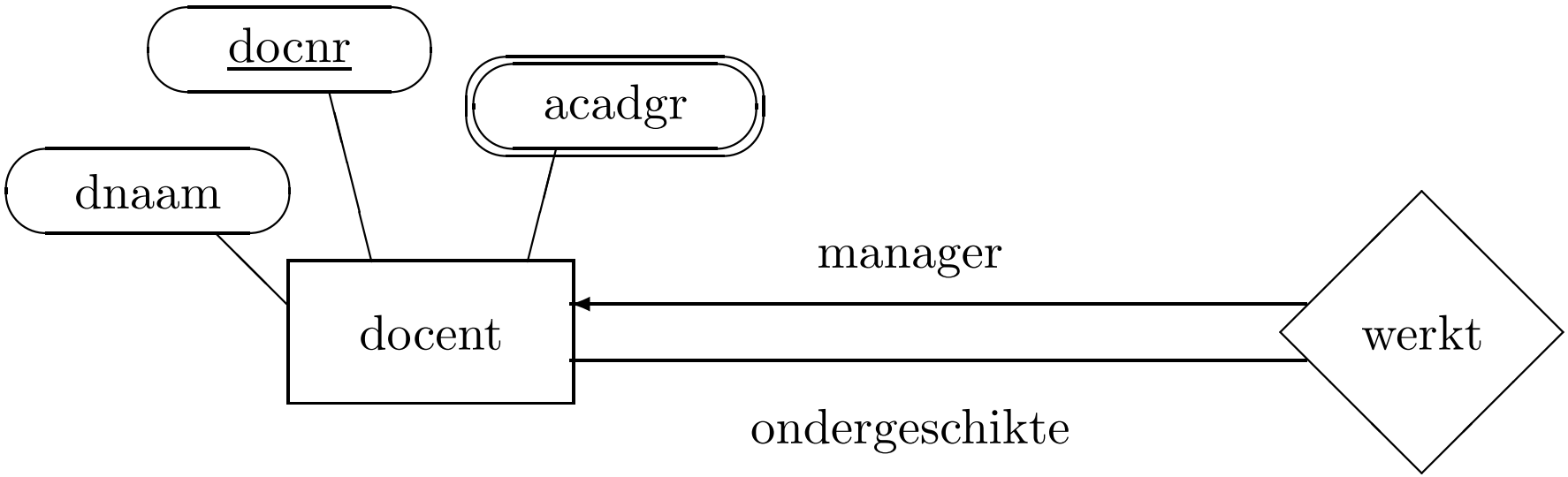
Veel-op-een

Bijvoorbeeld

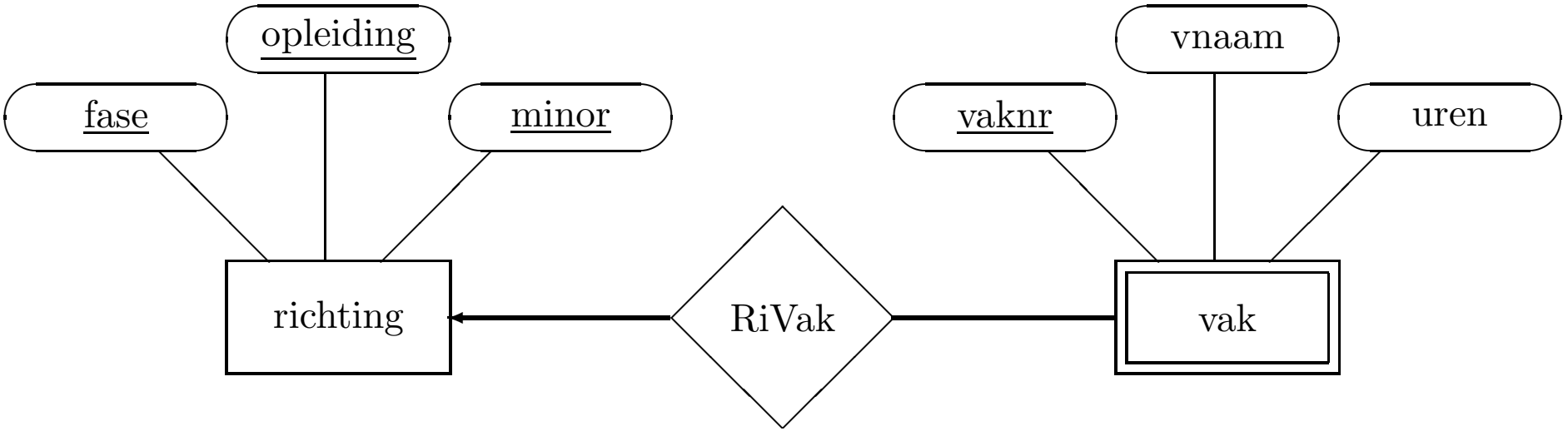


Een-op-een

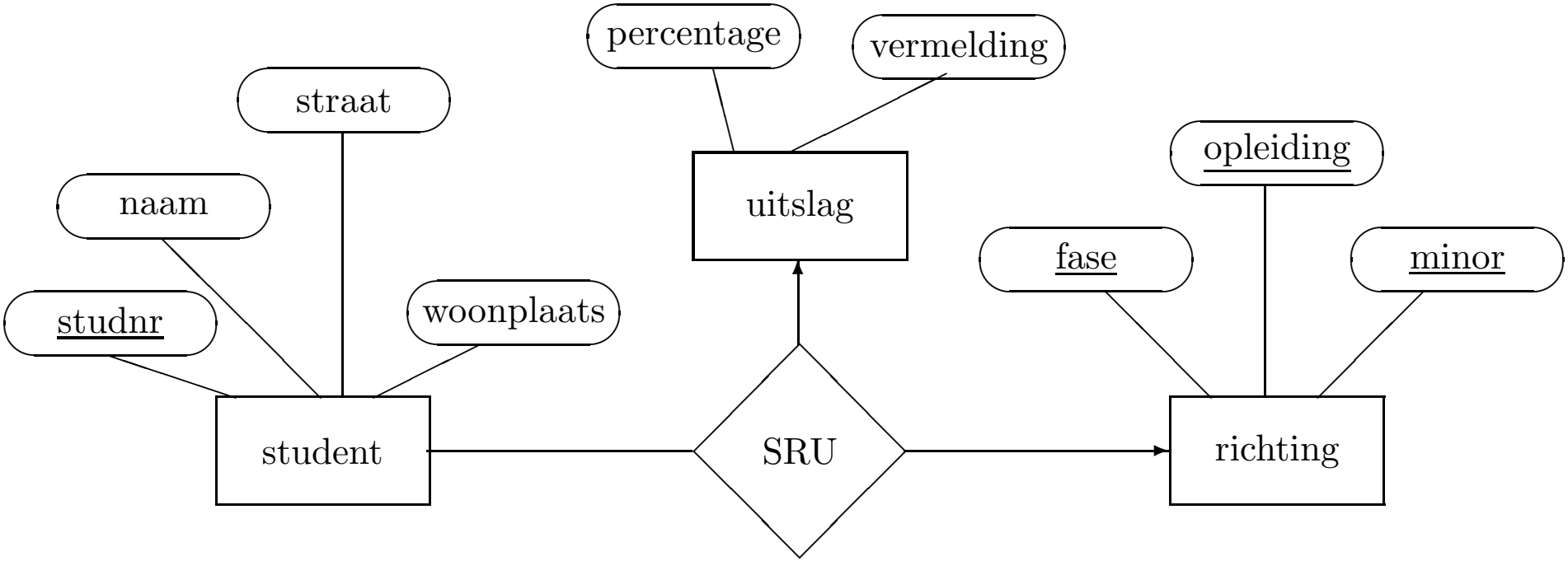
Rollen & meerwaardige attributen



Zwakke entiteit



Niet binaire relatie



Huiswerk

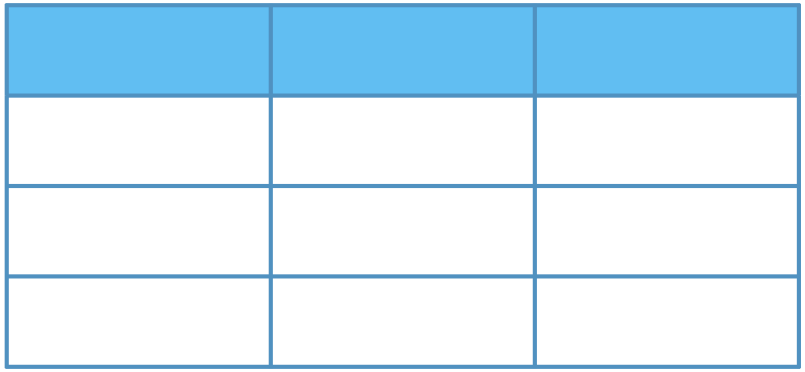
- Maak ER diagramma
- Geef af bij 1e labo
- (En doe uw best!)

Van ER naar DB?

- ER-diagramma beschrijft structuur van gegevens
- Bedoeling is om gegevens zelf op te slaan
  - Manier waarop wordt afgeleid van ER-diagramma

Databank



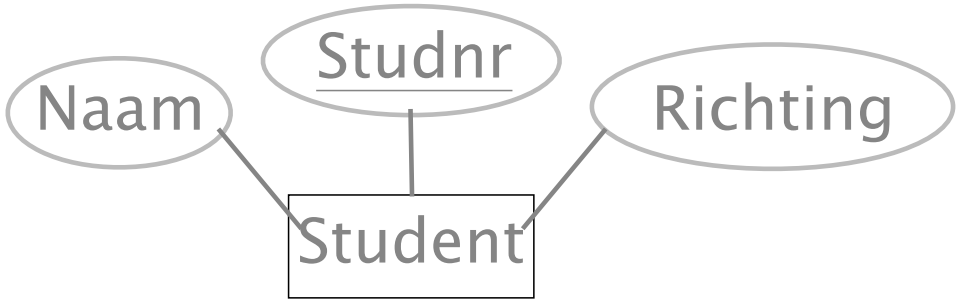






Van ER naar DB?

Student		
Studnr	Naam	Woonplaats
s00001	Jef	Mechelen
s00002	An	Duffel
…	…	…



Tabel maken

CREATE TABLE *naam* (*kolommen*, *integriteitsbeperkingen*);

Kolom: <naam> <datatype>

Student		
Studnr	Naam	Woonplaats
...	...	...

```
CREATE TABLE Student
  ( Studnr INTEGER,
    Naam CHARACTER(30),
   Woonplaats CHAR(40)      );
```

Data types

- SMALLINT: “kleine” gehele getallen (-32767 en 32767)
- INTEGER: grote gehele getallen
- DECIMAL(*x*,*y*): kommagetal met *x* cijfers waarvan *y* na de komma
- CHARACTER(*x*): woord van *x* letters
- DATE: datum

Tabel maken

CREATE TABLE *naam* (*kolommen*, *integriteitsbeperkingen*);

Kolom: <naam> <datatype>

Student		
Studnr	Naam	Woonplaats
...	...	...

```
CREATE  TABLE Student
  ( Studnr INTEGER PRIMARY KEY,
    Naam CHARACTER(30) NOT NULL,
    Woonplaats CHAR(40)       );
Of: ... , PRIMARY KEY (Studnr) );
```