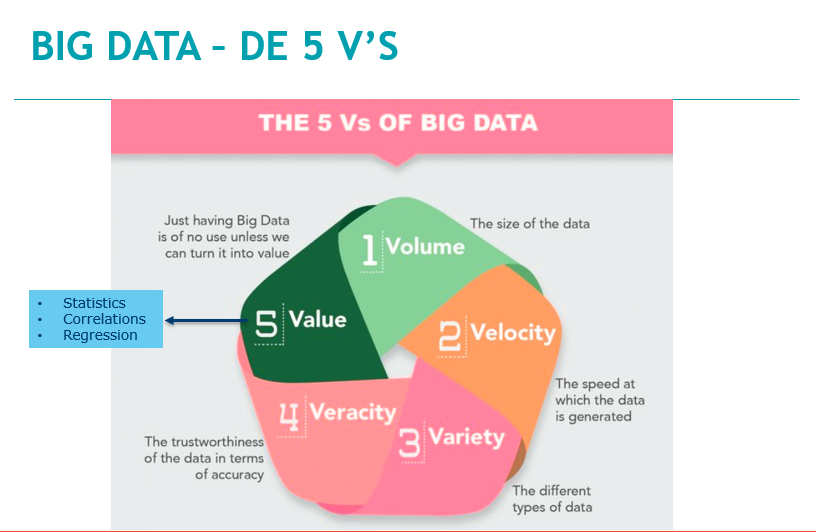
Big Data

# Wat is Big Data?

Eén of meer datasets die te groot zijn om met reguliere databasemanagementsystemen onderhouden te worden. Niet alleen de opslag van deze hoeveelheden is een uitdaging.   
Ook het analyseren van deze data speelt een steeds grotere rol.

* Bevat veel informatie voor verschillende doeleinden, zoals marketing, wetenschappelijk onderzoek,...

# De 5 V’s!



# Basisconcepten statistiek

* Een populatie  
  = de verzameling van eenheden die we willen bestuderen
  + Voorbeelden:
    - Alle inwoners van België
    - Alle stemgerechtigden in België
    - Alle personen die een bepaald merk van mobiele telefoon hebben gekocht
    - Alle ongelukken die gedurende september op de E34 zijn gebeurd
* Een steekproef   
  = een deelverzameling van de eenheden van de populatie
  + Voorbeeld:
    - Je wil de politieke keuze kennen van de ThomasMore- student. Hiervoor onderzoek je 100 studenten. Hoe stel je de steekproef samen?
    - Je wil de politieke keuze vergelijken van ITF, EM, ENT en BOUW. Hoe stel je je steekproef samen?
  + Belangrijk is dat er bij de steekproef is dat de populatie representatief is. Er moeten van elke deelgroep evenveel studenten zijn. (Niet alleen ITF)
  + Als **alle deelgroepen** van de populatie **even sterk vertegenwoordigd** zijn in de steekproef, dan noemen de steekproef representatief voor de volledige populatie.
* Kwantitatieve gegevens zijn gegevens waarmee gerekend kan worden.
  + Kwantitatieve gegevens zijn **continu** als alle waarden in een bepaald interval mogelijk zijn.
  + Kwantitatieve gegevens zijn **discreet** als slechts een bepaald aantal waarden mogelijk zijn. De tussenliggende waarden hebben geen betekenis.
* Kwalitatieve gegevens zijn gegevens waarmee je niet kunt rekenen.

Met welke soort gegevens hebben we hier te maken?

(a) bloedgroep -> Kwalitatief

(b) temperatuur -> Kwantitatief continu

(c) rendement (hoog-laag-normaal) kwalitatief

(d) geslacht -> Kwalitatief

(e) examencijfers -> kwantitatief discreet

(f) godsdienst -> Kwalitatief

(g) Salarisschaal (HEI19, HEI18, HEI17, ...) -> Kwantitatief Continu

(h) het aantal producten dat niet voldoet aan de kwaliteitscriteria -> Kwantitatief Discreet

(i) de kleur van een kledingsstuk -> Kwalitatief

(j) de mate van tevredenheid -> Score 0-10 (Kwantitatief Discreet), Kwalitatief (gemiddeld, zeer tevreden)

(k) het merk van een auto -> Kwalitatief

(l) de levensduur van een wasmachine -> Kwantitatief Continu

Opiniepeilers houden regelmatig enquêtes om de populariteit van de huidige president in een republiek te peilen. Stel dat er morgen een enquête gehouden wordt waarin 2000 personen gevraagd zal worden of de president het goed of slecht doet. De 2000 personen worden gekozen door random telefoonnummers te kiezen, en de vragen worden per telefoon gesteld.

(a) Wat is de relevante populatie? 2000 mensen

(b) Wat is de relevante variabele? Is deze kwantitatief of kwalitatief? kwalitatief

(c) Wat is de steekproef? 2000 mensen die een telefoonnummer hebben

(d) Is dit een goede manier om een steekproef te trekken? Nee niet iedereen heeft een telefoon

# Centrummaten

**Gemiddelde (µ):** Alle getallen optellen en delen door het aantal getallen.

Bv: 10 10 9 7 6 7 5 8 8 7 10 7 7 9 (110 totaal; 14 getallen; 110/14 = 7,8)

Het gemiddelde van deze reeks gegevens is: ... 7,8

**Mediaan (Me):** Het middelste getal van alle waardes. (of het gemiddelde van de 2 middelste getallen)

Bv: 10 10 9 7 6 7 5 8 8 7 10 7 7 9

Gerangschikt krijgen we

5 6 7 7 7 7 7 8 8 9 9 10 10 10 -> Mediaan is 7,5

**Modus (Mo):** De modus Mo is de waarde die het vaakst voorkomt in de gegevensverzameling.

# Spreidingsmaten

**Het** **bereik**: Het verschil tussen het grootste en het kleinste gegeven.

Geen goede maat, houd alleen rekening met de uiterste waarde

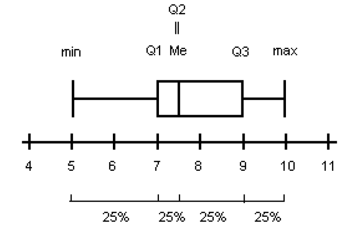
**Interkwartielafstand**: Het verschil tussen het derde en het eerste kwartiel. Extreem hoge of lage getallen hebben geen invloed op de interkwartielafstand.

De gegevens verdelen in 4 gelijke delen.

Q2: alle gegevens verdelen in 2 delen

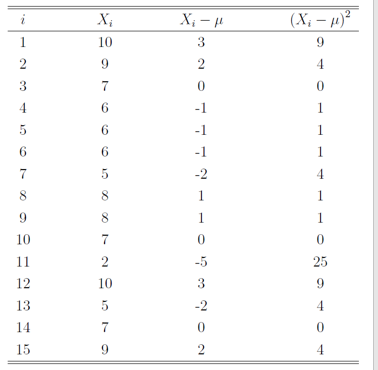
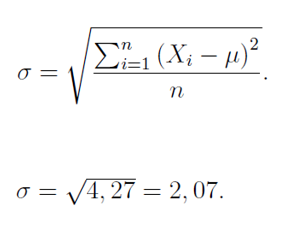
Q1 + Q3: Beide helften nog eens verdelen in 2 delen.

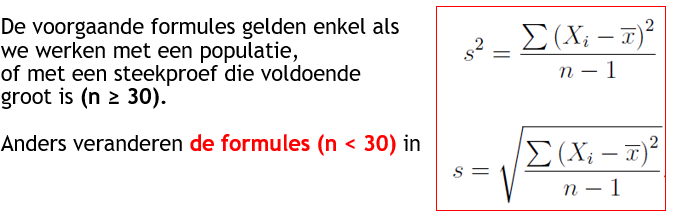
(Q0: minimale waarde, Q4: maximale waarde)



**Deviaties:** Het verschil van het gegeven en het gemiddelde van alle waarden.

**Standaarddeviatie :** Kwadraten van de deviaties (geen negatieve getallen). Daarvan het gemiddelde berekenen (**variantie**). Vierkantswortel van de variantie is standaarddeviatie.



# Correlaties

**Een positieve correlatie:** (een vermeerdering van de ene grootheid heeft een vermeerdering van de andere grootheid tot gevolg)

**Een negatieve correlatie:** (een vermeerdering van de ene grootheid heeft een vermindering van de andere grootheid tot gevolg)

**Correlatiecoëfficiënt:** De mate van correlatie tussen twee grootheden kan echter ook worden uitgedrukt in een getal.

In de grensgevallen R = -1 en R = +1 is er sprake van volledige correlatie.

Bij R = -1 is dat volledige negatieve correlatie en

bij R = +1 volledige positieve correlatie.

Als er geen sprake is van enige correlatie, dan geldt R = 0.

**Covariantie:** Een maat voor de spreiding van twee gekoppelde variabelen.

Als Cov(x, y) een positief getal is, dan is er sprake van positieve correlatie   
Als Cov(x, y) een negatief getal is, dan is er sprake van negatieve correlatie.

* + Bereken **het gemiddelde µ*x* van de *x*-waarden** en **het gemiddelde µ*y* van de *y*-waarden**.
  + Bereken **voor elk getal *xi* de deviatie *dxi = xi -* µ*x*** en bereken **voor elk getal *yi* de deviatie *dyi = yi -* µ*y***.
  + Bereken de **producten van de deviaties, dus (*xi -* µ*x* )(*yi -* µ*y* ).**
  + Bereken het **gemiddelde van die producten**.