Limperg Instituut

Data Science in de opleiding voor accountants

Ferry Geertman
Nyenrode

VU

Limperg Instituut

Agenda

- Statistieken
- Leerdoelen data science
- Verschillende raamwerken
- Technieken
- Bias Variance Trade-off
- Responsible AI

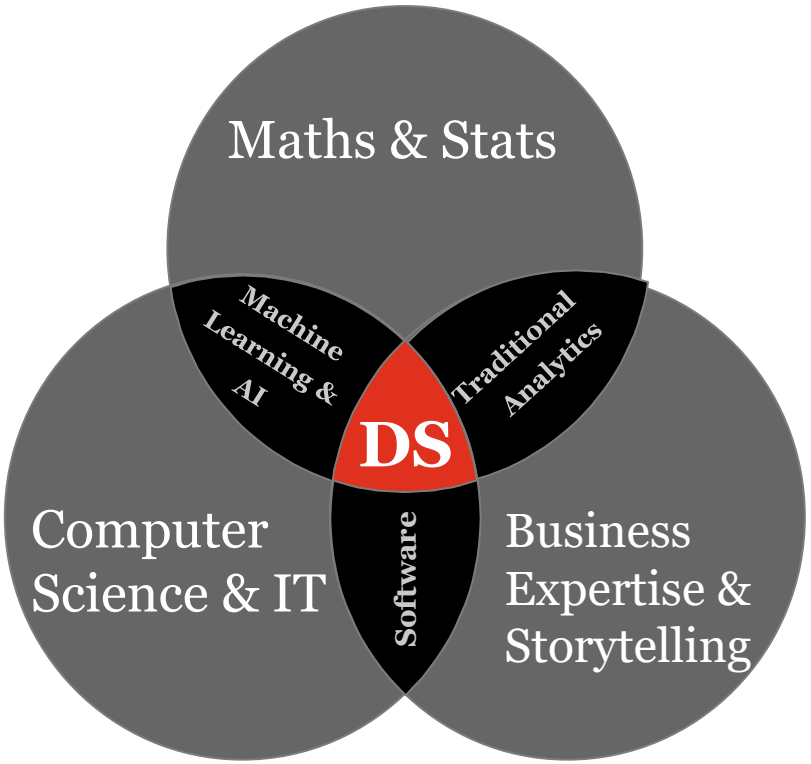
Limperg Instituut

Statistieken

- Vak Data Science gegeven sinds 2020 op Nyenrode
- Ruim 1.100 studenten
- Studielast – 4 ECTS (112 uur)
- Gemiddeld cijfer (1ste poging) 7,4

Limperg Instituut

Wat is Data Science?



Leerdoelen

- $y = f(x) + \varepsilon$
- Duiden van begrippen
- Kunnen beoordelen van kwaliteit van modellen/ algoritmen
- Zelf (eenvoudige) modellen in R bouwen

Limperg Instituut

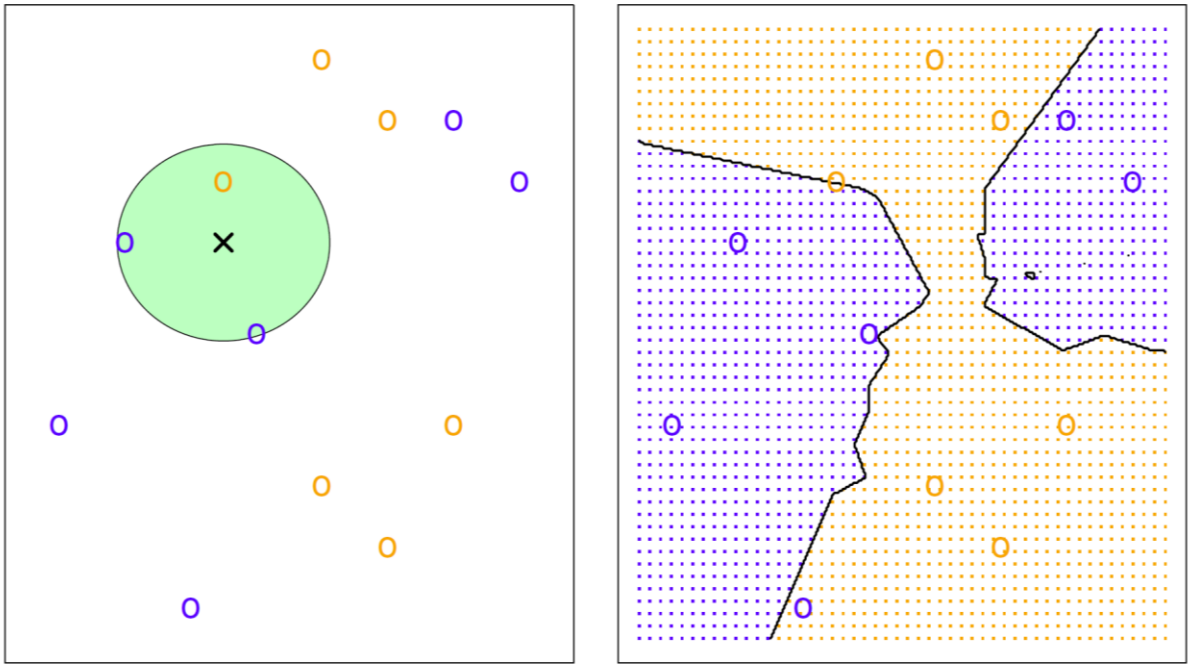
Data Science Lenzen

	Dimension	Low maturity	High maturity
1	Gartner	Descriptive	Prescriptive
2	Vs of big data	Small data	Big Data
3	Push-left	Reporting	Business partner
4	Wisdom pyramid	Deductive	Inductive

Limperg Instituut

Technieken

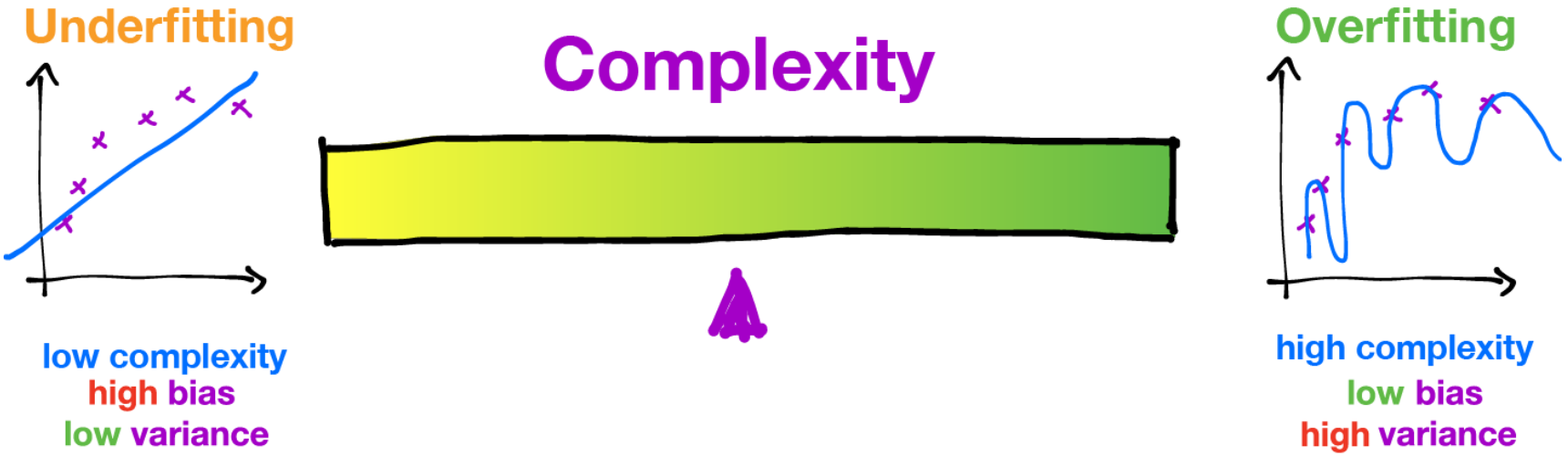
- Regressie (parametrisch)
- KNN (non-parametrisch)



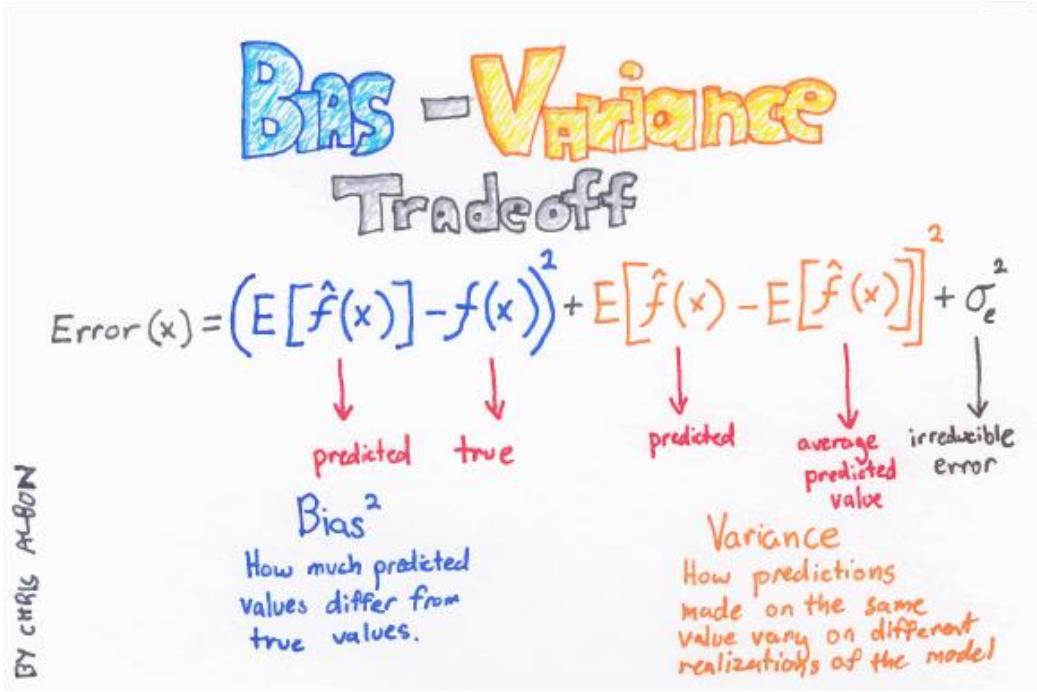
Beoordelen kwaliteit model

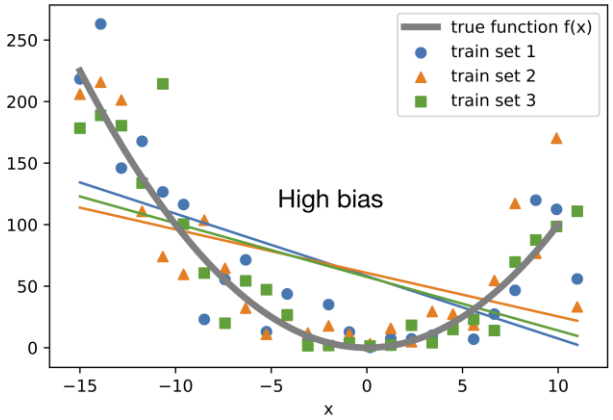
BIAS – VARIANCE TRADE-OFF

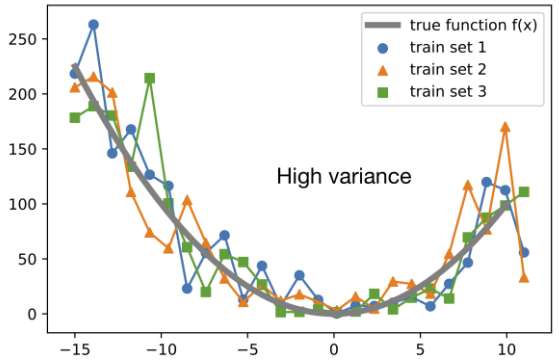
- Bias – fout geïntroduceerd door een complex probleem te benaderen door een simpel model
- Variance – Omvang waarmee de voorspelling gedaan met het model wijzigt als het model getraind zou zijn met andere data
- Trade-off – minder bias leidt tot meer variance en meer bias geeft minder variance
- Op zoek naar het optimum door de complexiteit (flexibiliteit) aan te passen



Bias-Variance Trade-off

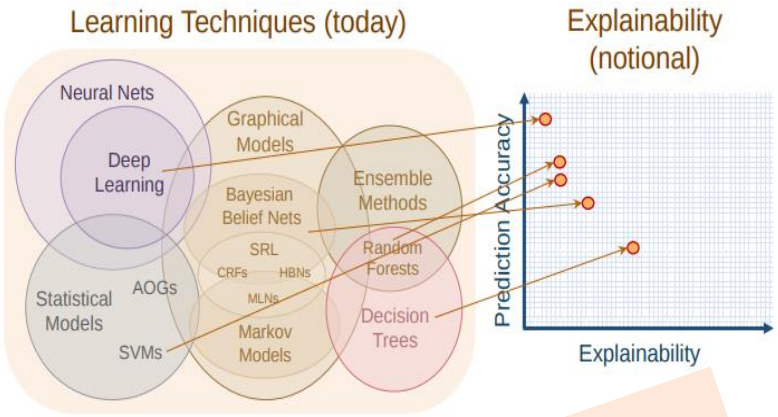






Limperg Instituut

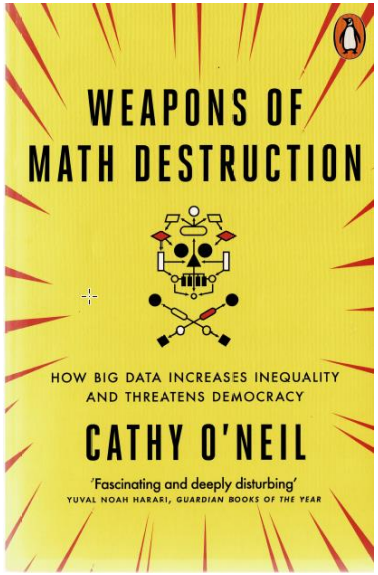
Responsible AI



OPINIE

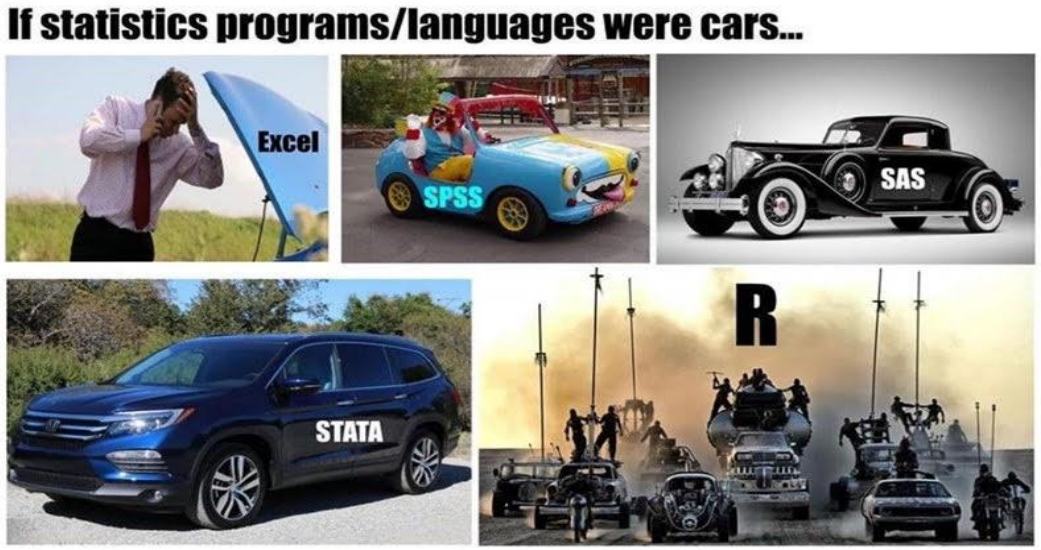
Data beweegapps van zorgverzekeraars mogen niet gekoppeld worden aan polisvoorwaarden

Opinie | Mark van Houdenhoven, bijzonder hoogleraar Economische bedrijfsvoering in de gezondheidszorg



VU

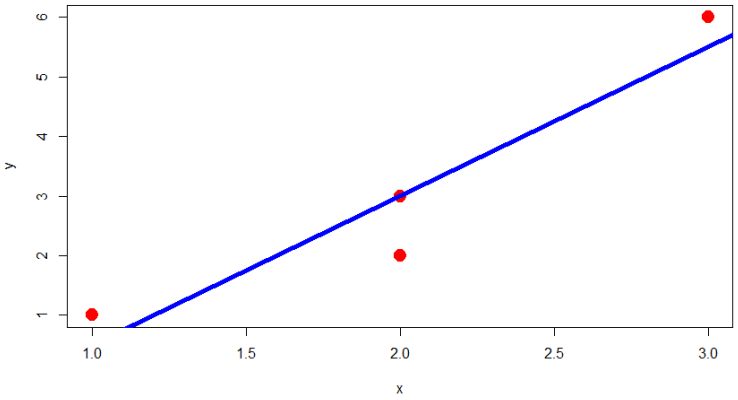
Gebruik van tooling




```
> x=c(1,2,2,3)
> y=c(1,2,3,6)
> x
[1] 1 2 2 3
> y
[1] 1 2 3 6
> mean(x)
[1] 2
> mean(y)
[1] 3
> plot(x,y,col="red",pch=19,cex=2)
> linreg <- lm(y~x)
> abline(linreg,col="blue",lwd=5)
> summary(linreg)
Call: lm(formula = y ~ x)
Residuals:          1          2          3          4
                5.00e-01 -1.00e+00 1.11e-16 5.00e-01
Coefficients:
            Estimate    Std. Error      t value   Pr(>|t|)
(Intercept) -2.0000          1.2990     -1.540    0.2635
x            2.5000          0.6124      4.082    0.0551 .
```

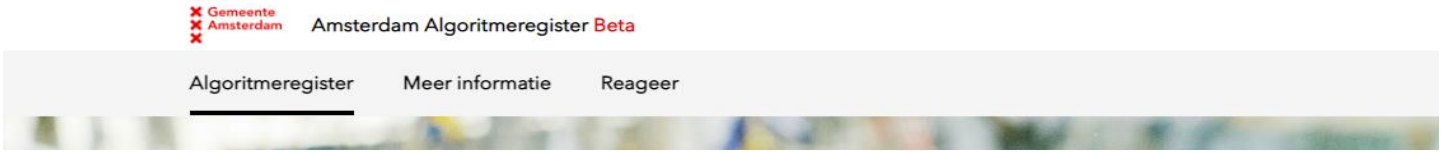
```
> fit[5,100,]
Error in fit[5, 100, ] : subscript out of bounds
```

```
> library(Hmisc)
Error: package or namespace load failed for 'Hmisc' in loadNamespace(j <- i[[1L]], c(lib.loc,
 .libPaths()), versionCheck = vI[[j]]):
 there is no package called 'xfun'
> describe
Error: object 'describe' not found
> describe(COPD)
Error in describe(COPD) : could not find function "describe"
```



Limperg Instituut

De Praktijk



Een algoritme kan echter zo goed zijn in het vinden van patronen, dat het direct uitsluiten van gevoelige gegevens niet voldoende is. We hebben daarom ook uitgezocht of de niet-gevoelige gegevens die het algoritme verwerkt indirect toch tot ongewenste verschillen in behandeling leiden. Het zou bijvoorbeeld kunnen dat in bepaalde wijken veel mensen wonen met een bepaalde nationaliteit, of dat een bepaalde groep gemiddeld genomen grotere gezinnen heeft. Als het algoritme dan gebruik kan maken van een gegeven als postcode of gezinsgrootte, kan het indirect alsnog onderscheid maken tussen bepaalde groepen, simpelweg door onderscheid te maken tussen wijken/gezinsgrootte. Zo kan een groep alsnog benadeeld worden door het algoritme, ook al is de groep niet expliciet bekend bij het algoritme.

We hebben ervoor gekozen om hier nader onderzoek naar te doen. Het zogenaamd onderzoek naar onbewuste vooringenomenheid (bias) uitgevoerd en de resultaten hiervan zijn te vinden op:

Blog post 1:

Analyzing Bias in Machine Learning: a step-by-step approach (amsterdamintelligence.com)

Er kan worden geconcludeerd dat het algoritme vrij is van een ongerechtvaardigde vooringenomenheid en daarmee geschikt is om een pilot mee te starten.

Gebruik van algoritmes binnen gemeente Amsterdam

Maak kennis met de diensten van gemeente Amsterdam waarbij algoritmes gebruikt worden.

VU