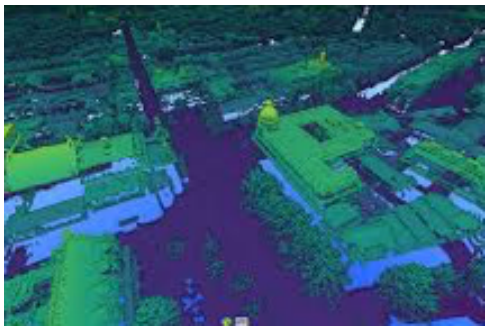
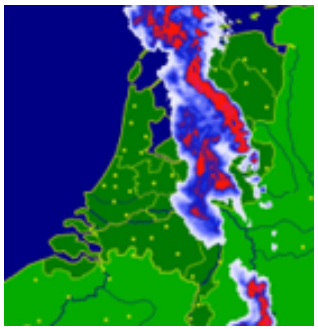



Case-study: Data science-technieken voor regenwateroverlast in stedelijk gebied

Jan N. van Rijn, Universiteit Leiden

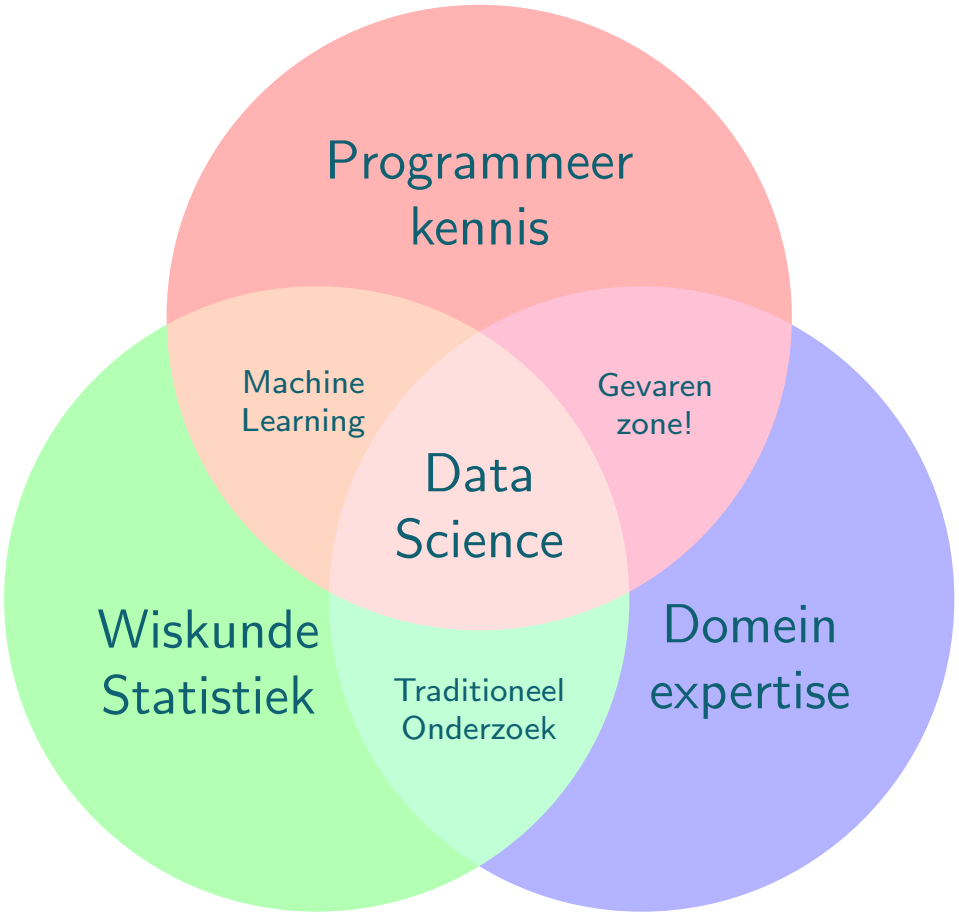
13 oktober 2022


Wicked Ale
LANDSHARK
Shiner BOCK
REDHOOK ESB ORIGINAL ALE
REDHOOK ESB
LONG HAMMER
Pyramid APRICOT ALE
Pyramid APRICOT ALE
ALASKAN
FULL MOON
Pyramid HEFEWEIZEN
WIDMER
FAT TIRE
RISING MOON
BLUE MOON
SAM ADAMS LIGHT
SAMUEL ADAMS
BECK'S
FAT TIRE

Universiteit Leiden

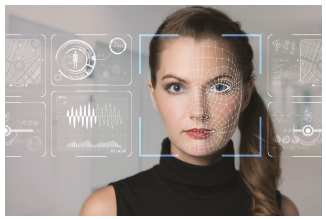
‘Mensen Centraal in Kunstmatige Intelligentie’

- Machine Learning modellen zijn ontworpen om fouten te maken
- Kunstmatige Intelligentie om menselijk besluitvorming te ondersteunen
- Niet bedoeld om mensen of banen te vervangen
- Goede data nodig
- Gebruik deze technieken met gezond verstand


Programmeer kennis
Machine Learning
Gevaren zone!
Data Science
Wiskunde Statistiek
Traditioneel Onderzoek
Domein expertise

We zijn omringd door data...

- Gezichtsherkenning: Facebook, smart phones
- Spraakherkenning (Siri / Google talk)
- Vertaling (Google translate, Skype auto-translate)
- Product/song/video aanbevelingen in online winkels
- Spam filter
- Zondag met Lubach: Fabeltjes fuik en Deep Fakes
- ...



Universiteit Leiden

Data

sepal length	sepal width	petal length	petal width	class
5.1	3.7	1.5	0.4	setosa
5.1	3.3	1.7	0.5	setosa
6.2	2.8	4.8	1.8	virginica
6.9	3.1	5.4	2.1	virginica
6.1	3.0	4.6	1.4	versicolor
4.7	3.2	1.3	0.2	setosa
4.4	3.2	1.3	0.2	setosa
6.5	2.8	4.6	1.5	versicolor
6.8	3.0	5.5	2.1	virginica
6.3	3.4	5.6	2.4	virginica
5.1	3.5	1.4	0.3	setosa
6.3	3.3	6.0	2.5	virginica
5.0	3.2	1.2	0.2	setosa
5.1	3.8	1.9	0.4	setosa
5.7	4.4	1.5	0.4	setosa

Data

Data types

- Numeriek (in de Iris: petal width, petal length, . . .)
- Categorisch (e.g., het weer: regenachtig, storm, zon)
- Ordinaal (niet numeriek, maar met volgorde)
- Cyclisch (e.g.: seizoenen)

Meestal: Numeriek

Data

Typische datasets

- Aandelen – stijgen of dalen
- Iris – plant classificatie
- Koning en Toren tegen Koning Pion – Schaak posities
- Boston housing – voorspel huizen prijs
- Diabetes – voorspel of een persoon diabetes heeft
- MNIST – handschrift herkenning
- CIFAR – object herkenning

Data

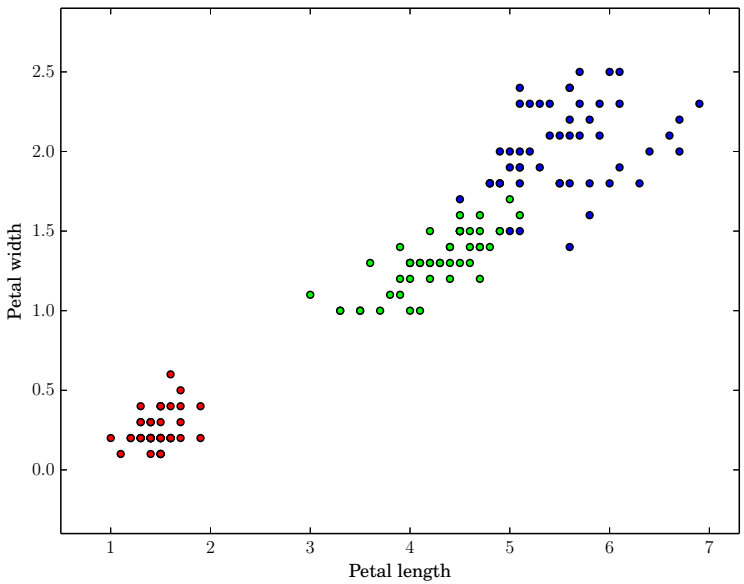
Uitdagingen

- Gebrek aan ‘labels’
- Missende getallen
- Verschillende typen (numeriek, categorisch)
- Menselijke fouten, bias (ruis)
- ‘Sparsity’
- ‘Big’ data
- ontoereikende attributen (tegensprekende data)
- ‘Sampling bias’

Data

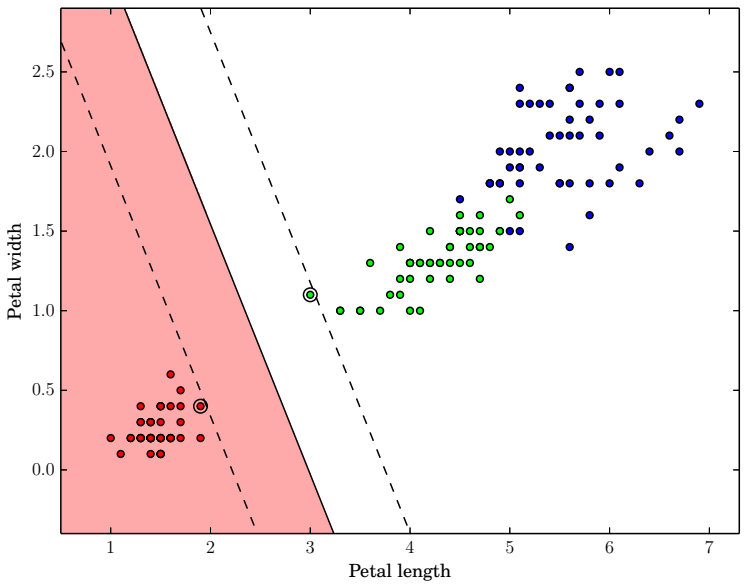
- Complexe vormen
  - text, plaatjes, geluid, relationele databases, logische statements
- in veel gevallen om te vormen tot tabulaire data
- operaties: ‘convolutions’, ‘bag of words’, lineaire combinaties, ‘queries’

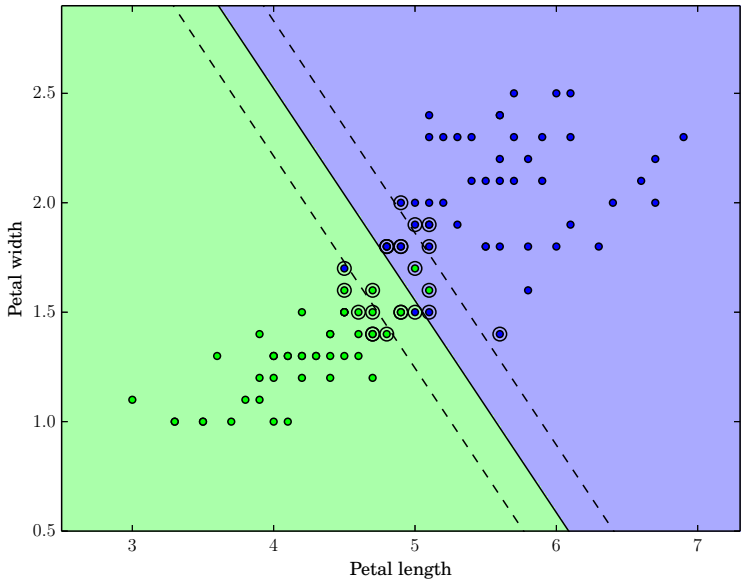
Data



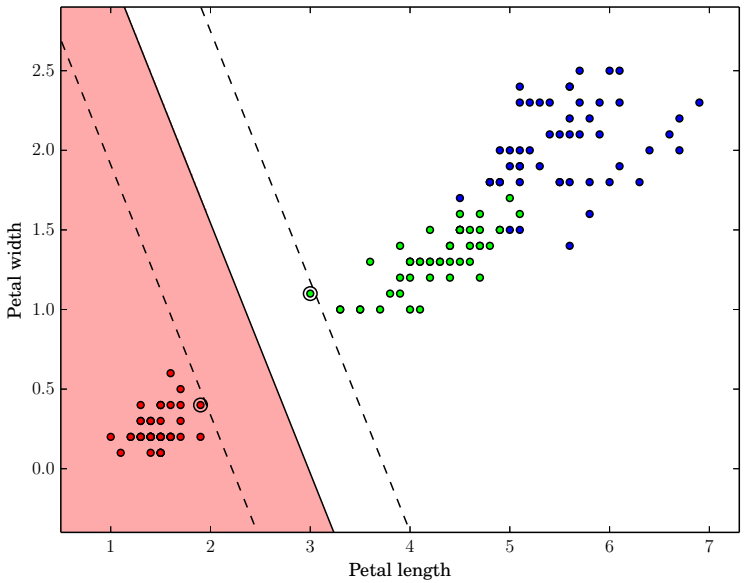
- We kunnen slechts 2 dimensies plotten
- Elke kleur representeert een klasse
- De ‘Setosa’ (rood) bloem is makkelijk te onderscheiden van de rest

Modellen


Petal width
Petal length
0.0
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7


Petal width
Petal length
0.5
1.0
1.5
2.0
2.5
3
4
5
6
7

Modellen



- Modellen vertellen ons iets over de data
- Geven een mate waarmee we nieuwe (onbekende) gevallen kunnen voorspellen
- Soms weten modellen hoe zeker ze zijn over hun voorspelling
- Simpel en interpreteerbaar vs. Complex, krachtig en lastig te begrijpen

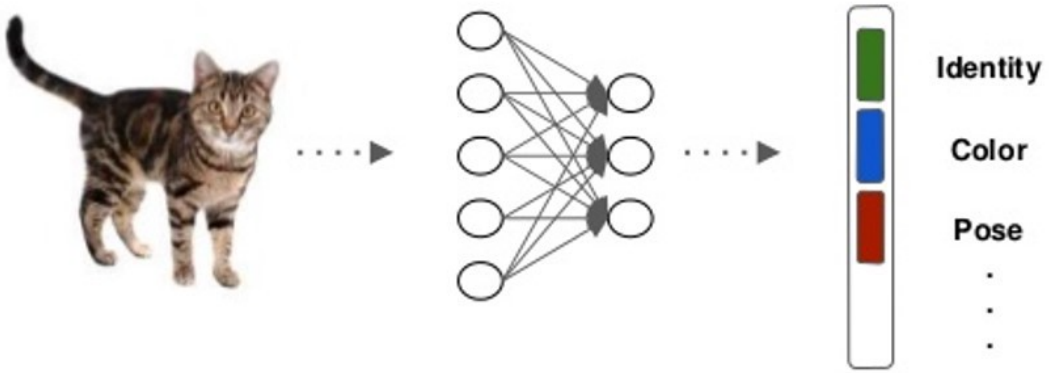
Models

Traditionele Machine Learning

- Menselijke attributen
- Domein kennis, 'trial and error'
- Kleine hoeveelheden data zijn voldoende

'Representation Learning'

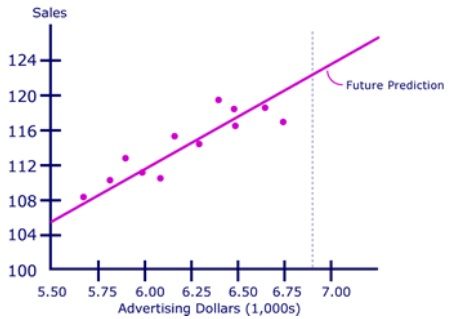
- 'Deep Learning', 'Neural Networks'
- Weinig (voor)kennis nodig over de data
- Veel gelabelde data nodig



Taken

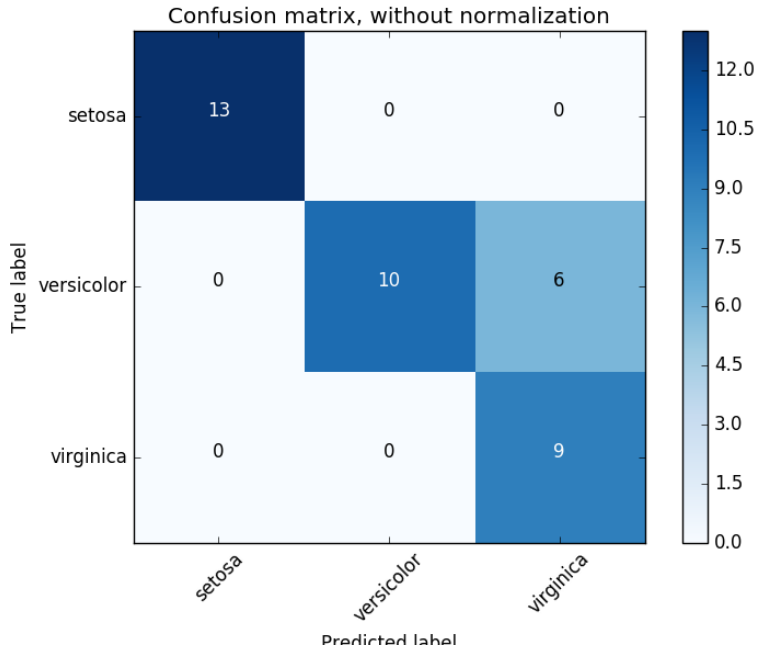
- Classificatie: voorspel een klasse
  - Tot welke familie behoord een plant?
  - Welk karakter heeft een gebruiker geschreven?
  - Is deze e-mail spam?
- Regressie: voorspel een nummer
  - Wat is de waarde van een huis?
  - Hoe oud is de persoon op dit plaatje?
  - Hoeveel zal een bepaald aandeel stijgen of dalen in waarde?

- Veel andere taken:
  - 'Subgroup Discovery'
  - 'Exceptional Model Mining'
  - 'Clustering'
  - 'Pattern recognition'



Evalueren wat we hebben geleerd

- Meetbare waarde hoe een model het doet
- Verschillende maatstaven
  - 'Accuracy' (percentage correcte voorspellingen)
  - Confusion Matrix
  - Precision (per class)
  - Recall (per class)
  - F-measure (per class)
  - Area under the ROC curve
- Een gemene zaak ..



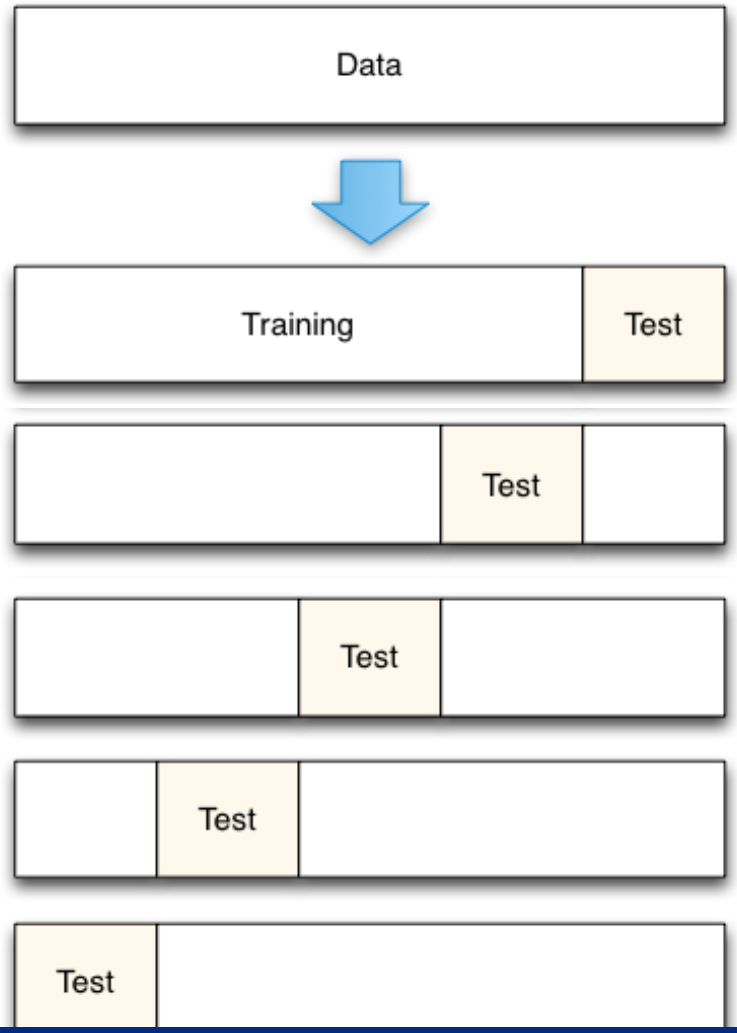
Generaliseerbaarheid

Machine Learning gaat over **Generaliseerbaarheid**.

- Wat gebeurt er als we een model toepassen op train data
- Hoe doet een model het op data die we nog niet gezien hebben
- We gebruiken hiervoor een holdout set

K-fold Cross-validation

- Model wordt getest op verschillende data
- Haalt een bias weg
- Bouwen verschillende keren



Universiteit Leiden

Welke bebouwde gebieden lopen een groot risico op regenwater overlast?


Tijdpoort

Gebaseerd op de Theses

- "Using machine learning to predict rainwater damage in urban areas'' (2021) door Christie Bavelaar
- "Lowering the Resolution of Damage Data in a Model that Predicts Rainwater Damage" (2022) door Teun de Mast
- Begeleiding: Ton Beenen (STOWA, RioNED), Mitra Baratchi, Jan N. van Rijn

Motivatie

- Zware buien zijn zeldzaam
  - We willen kennis opgedaan in bepaalde gebieden toepassen in andere gebieden
- In kaart brengen risico gebieden
- Pro-actief risico management
- Risico gradatie (hoeveel regenwater kan een bepaald gebied aan)
- Variabele premies voor verzekeraars, validatie verzekerings claims
- Aanvulling op rekenmodellen
- ...

Factoren en Gevolgen

Factoren
Regenval
Toestroom
Riool capaciteit
Bevolkingsdichtheid
Bebouwing
Constructie eigenschappen
Gevolgen
Regenwateroverlast

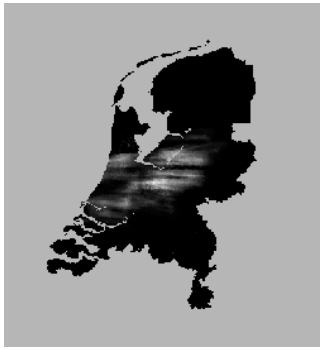
Factoren en Gevolgen

Factoren	
Regenval	KNMI Radardata
Toestroom	AHN hoogtekaart
Riool capaciteit	
Bevolkingsdichtheid	
Bebouwing	basisregistratie gebouwen en terreinen
Constructie eigenschappen	
Gevolgen	
Regenwateroverlast	

Factoren en Gevolgen

Factoren	
Regenval	KNMI Radardata
Toestroom	AHN hoogtekaart
Riool capaciteit	
Bevolkingsdichtheid	
Bebouwing	basisregistratie gebouwen en terreinen
Constructie eigenschappen	
Gevolgen	
Regenwateroverlast	Twitter berichten
	P2000 berichten
	Verzekerings claims

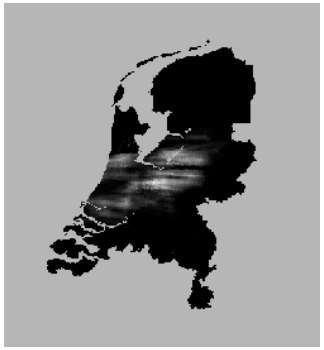
Data



KNMI Radardata

- 1000x1000 meter
- Kwantiteit: hoog
- KNMI radarvakken

Data



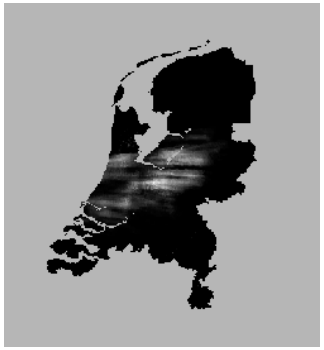
KNMI Radardata

- 1000x1000 meter
- Kwantiteit: hoog
- KNMI radarvakken

AHN2 Hoogtekaart

- 5x5 meter
- Kwantiteit: hoog
- longitude
  lattitude

Data

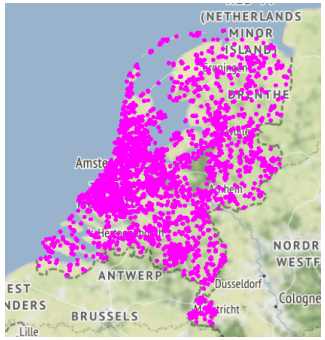


KNMI Radardata

- 1000x1000 meter
- Kwantiteit: hoog
- KNMI radarvakken

AHN2 Hoogtekaart

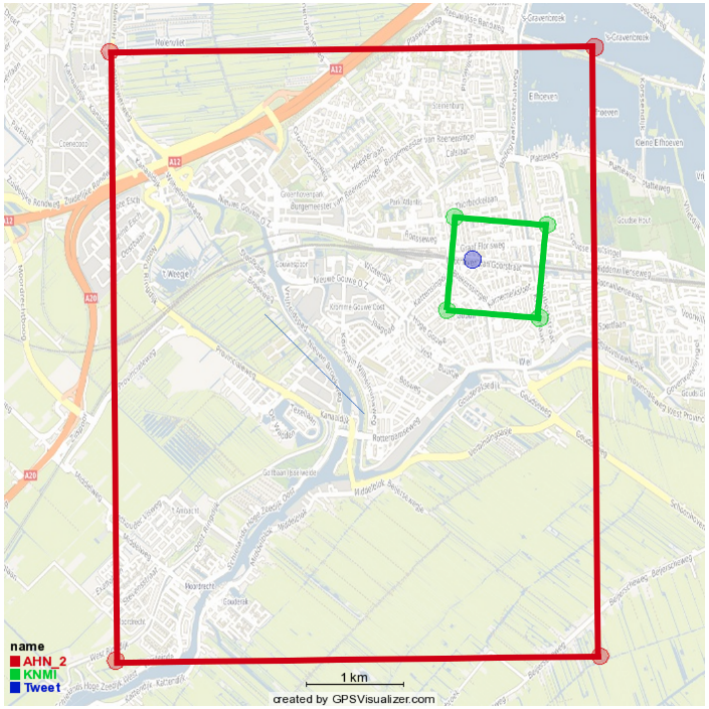
- 5x5 meter
- Kwantiteit: hoog
- longitude lattitude



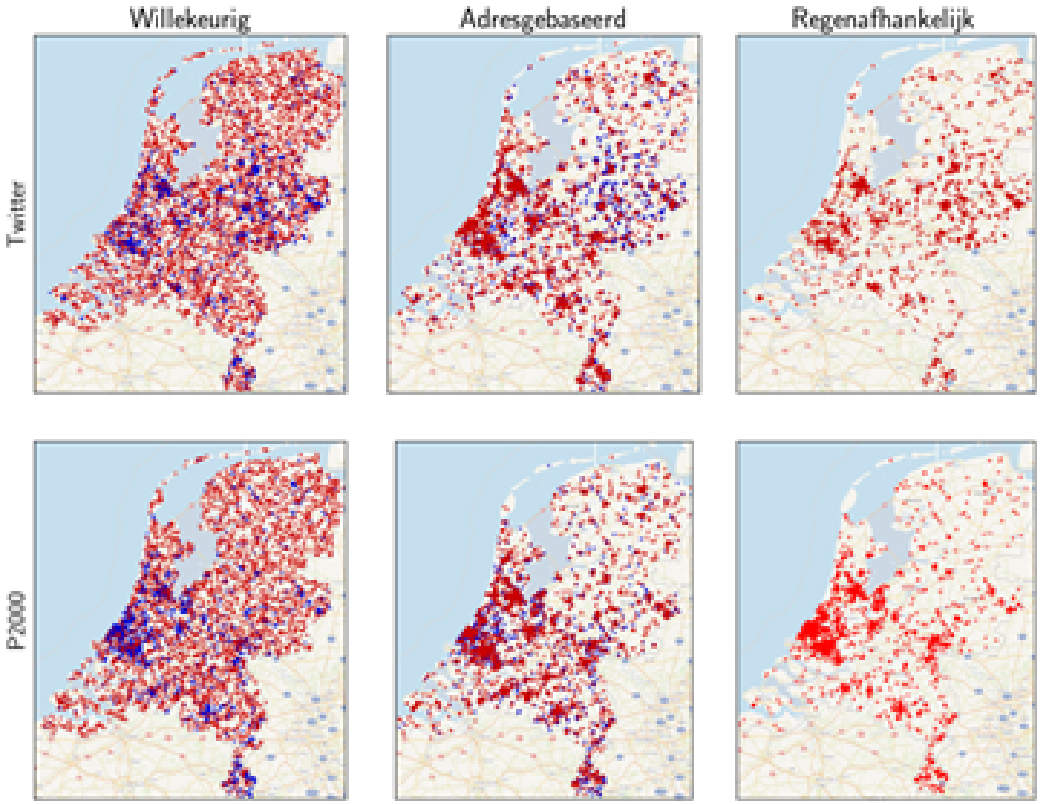
P2000 berichten

- GPS precisie
- Kwantiteit: redelijk
- longitude lattitude
- Grove benadering van werkelijkheid

Verschillende schalen



Verschillende settings



Resultaten

Gebruik de data om model te bouwen, dat voorspelt of een gebied bij bepaalde hoeveelheid regen risico loopt.

Gebruik dit model vervolgens op data die we hebben achter gehouden, en meet hoe goed het model werkt.

Resultaten

Gebruik de data om model te bouwen, dat voorspelt of een gebied bij bepaalde hoeveelheid regen risico loopt.

Gebruik dit model vervolgens op data die we hebben achter gehouden, en meet hoe goed het model werkt.

Experiment 10x uitgevoerd, gemiddelde gepresenteerd.

Hoevaak voorspellen we correct of er een P2000 bericht is geweest?	63,6%
Als we overlast voorspellen, hoe vaak is er dan ook een P2000 bericht geweest?	64,1%
Hoeveel procent van de P2000 berichten kunnen we identificeren?	62,8%

Volgende stappen

- Verzekeringsdata (Verbond Van Verzekeraars)
- Hogere resolutie data (AHN3)
- Bui kenmerken (resolutie in tijd)
- Meer precieze neerslagdata (100m radar metingen)
- Automated Machine Learning
- Extra data bronnen

Universiteit Leiden

Data Science voor waterbeheer

1. Waterbalans bebouwd gebied inzichtelijk maken
2. Ontdekken van trends in de gevolgen van regenval
3. 'Voorspellend onderhoud' op rioleringen, dijken, polders en waterwegen

   ...
4. Capaciteit management polders, rioleringen, rivieren
5. Effecten van droogte
6. Simulaties van fysische modellen

Data Science voor waterbeheer

1. Waterbalans bebouwd gebied inzichtelijk maken
2. Ontdekken van trends in de gevolgen van regenval
3. 'Voorspellend onderhoud' op rioleringen, dijken, polders en waterwegen
   ...
4. Capaciteit management polders, rioleringen, rivieren
5. Effecten van droogte
6. Simulaties van fysische modellen

Data verzamelen kost veel tijd en moeite, maar de verzamelde data is van grote waarde wanneer deze correct wordt toegepast!