



ANCHORMEN
data activators

DATA FOUNDATION 2

Raoul Grouls



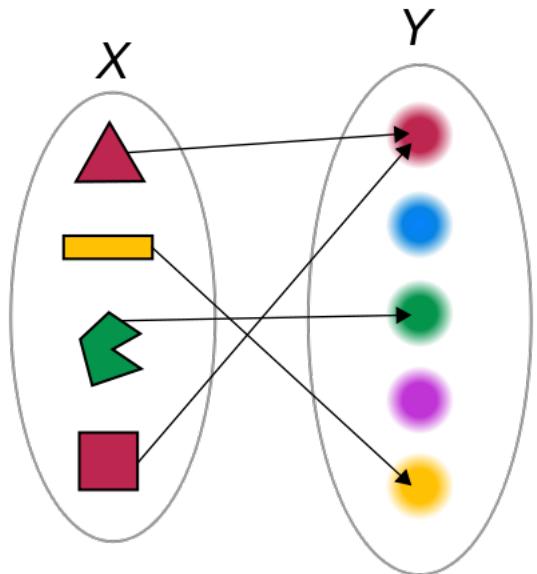
ANOMALIES

ANOMALIES

In het voorspellen worden fouten gemaakt.

We onderscheiden:

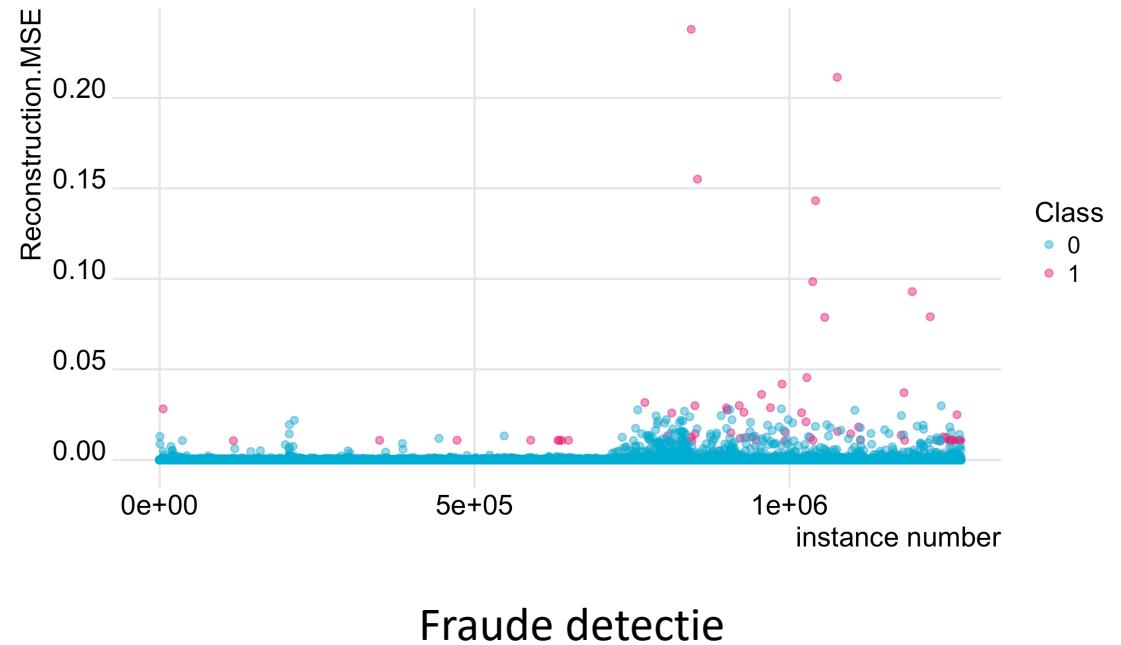
- Anomalies
 - Natuurlijke variantie
 - Noise
 - Bias
 - Variance



ANOMALIES

Anomalies zijn een breed concept dat ook de natuurlijke variatie binnen een populatie omvat.

Noise is type anomaly. Noise is een fout in de labels (Y) of in de meting van de features (X).



ANOMALIES

Als ik de lengte van een persoon opmeet, en als waarde 2.50m vind:

- Is dit altijd een anomalie? Waarom wel of niet?
- Is dit altijd noise? Wanneer wel of niet?

En als ik 1.75m vindt;

- Kan dit een anomalie zijn?
- En noise?

NOISE

Data = echte signaal + noise

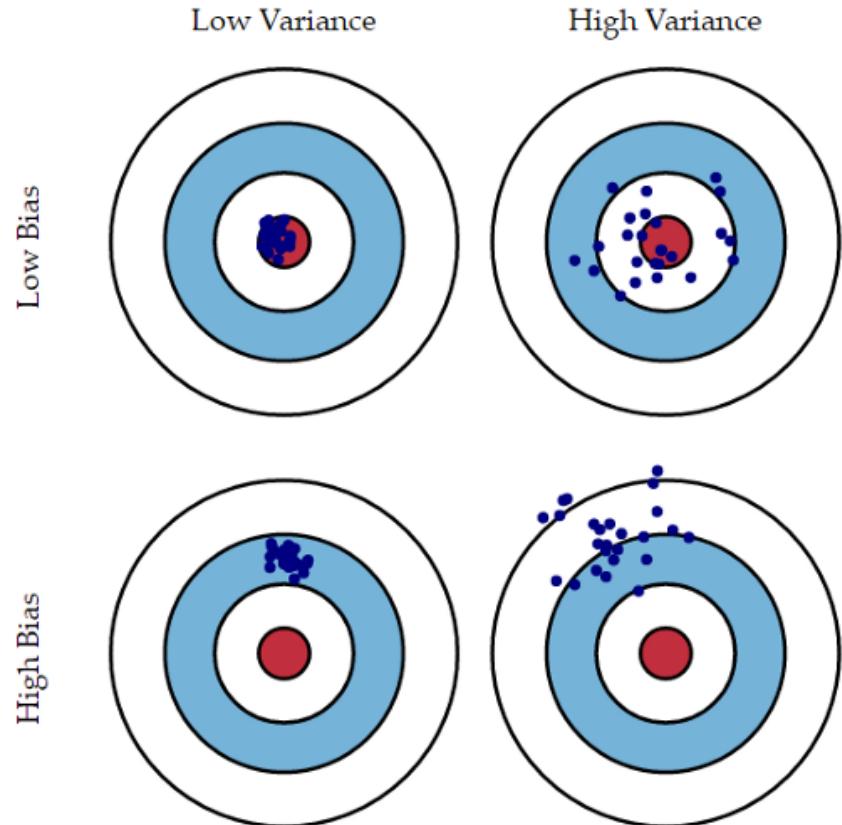
Noise is toegevoegde informatie zonder betekenis.

Is het mogelijk om meetfouten helemaal uit te sluiten? Waarom wel/niet?

Bestaat er altijd een foutmarge bij een meetresultaat? Kun je (tegen)voorbeelden geven?

Twee soorten:

- Random error (variance)
- Systematic error (bias)

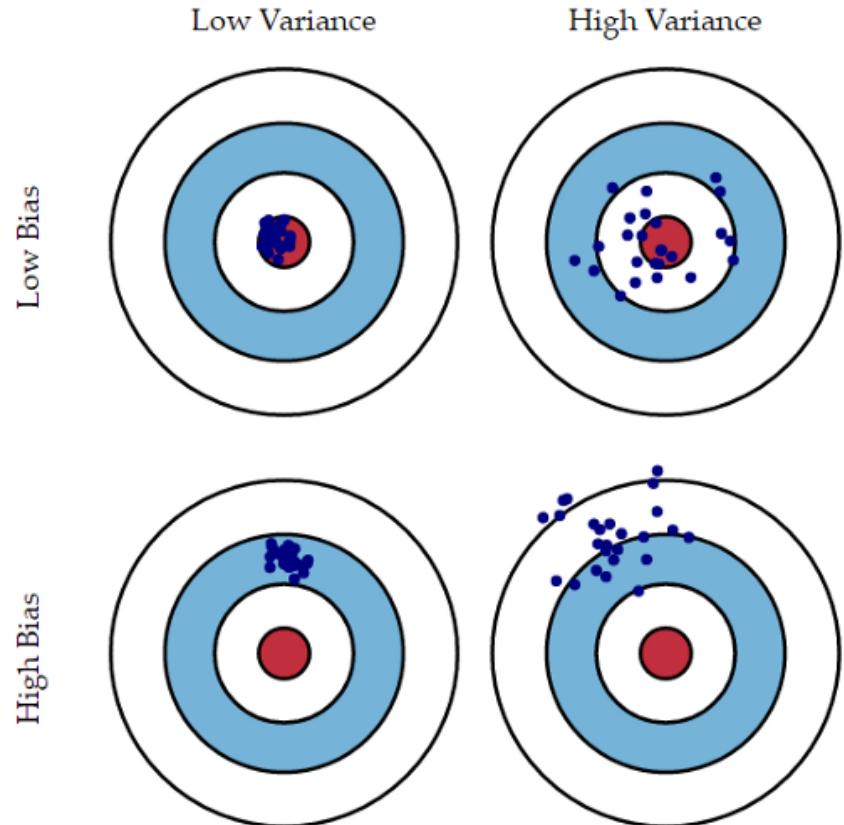


NOISE

Als we het over de error hebben:

Variance: de metingen zijn niet precies

Bias: de metingen hebben een structurele afwijking



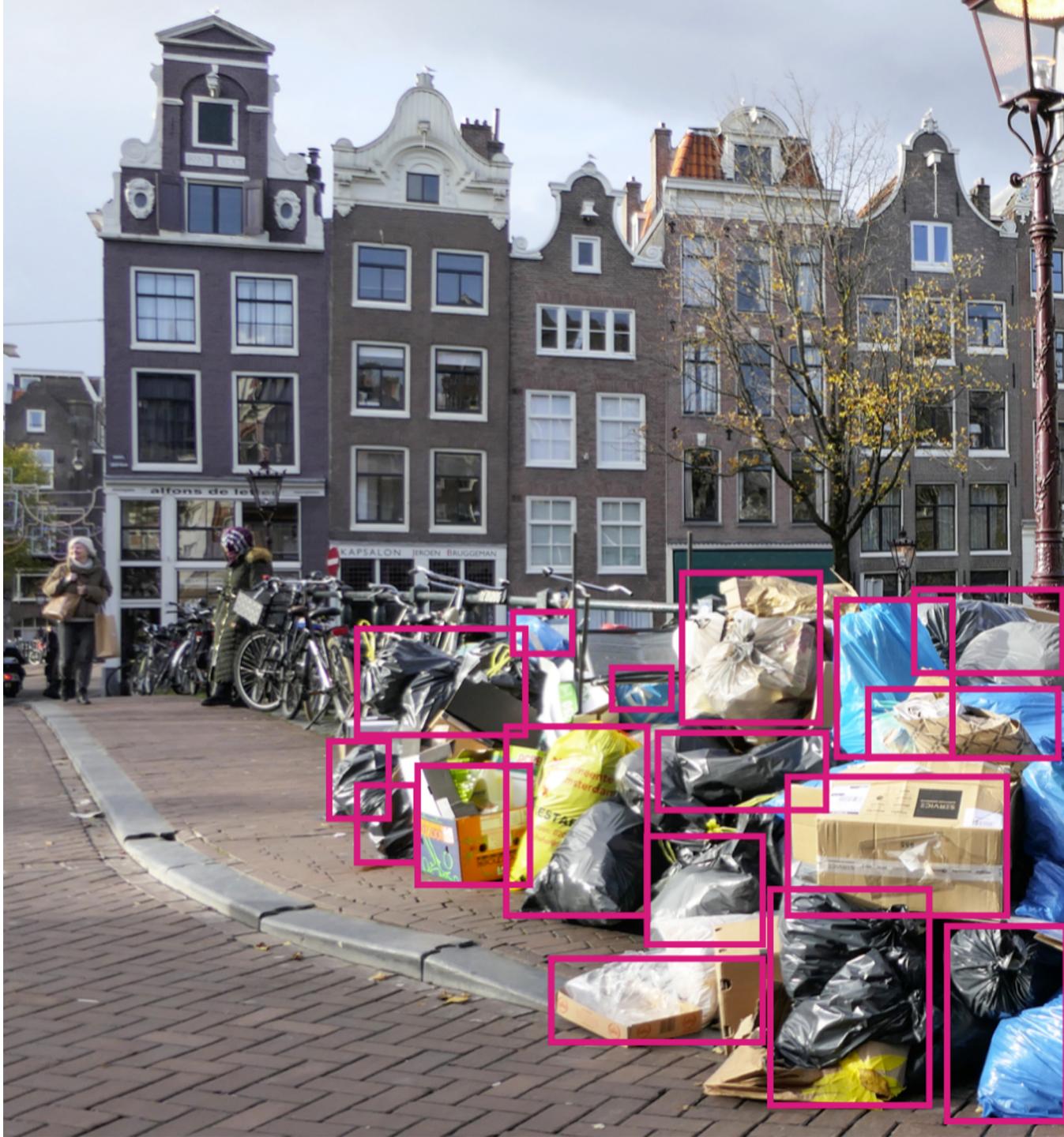
NOISE

Onderstaande zijn (minstens) prototypes in een testomgeving in Amsterdam:

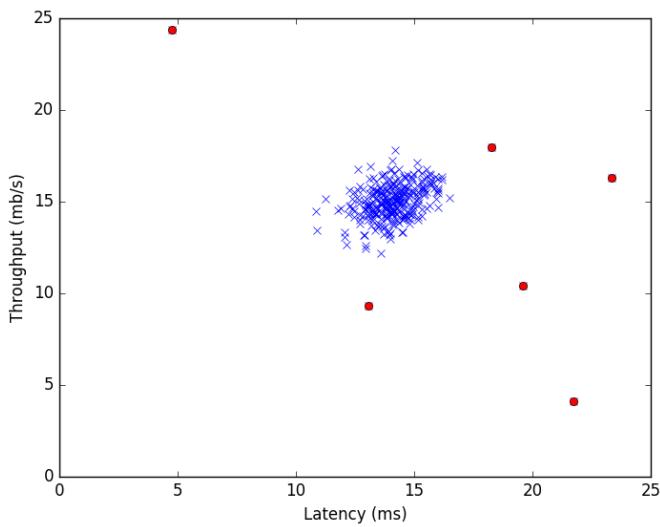
- Categorisering burgermeldingen op basis van woorden
- Kentekendetectie voor bepaling milieuzone
- Type afval herkenning vanaf vuilniswagens om afval te tellen
- Voorspellen parkeerdrukte
- Voorspellen stadsdrukte
- Voorspelling fraudekans wonen
- Tips op basis van persoonlijke voorkeuren voor culturele activiteiten via Facebook Messenger chatbot

Wat betekent bias in elk voorbeeld?

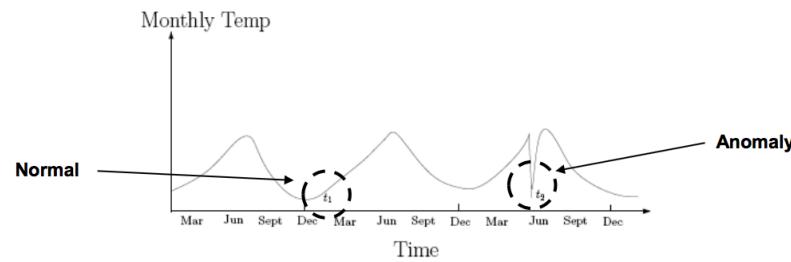
En variance?



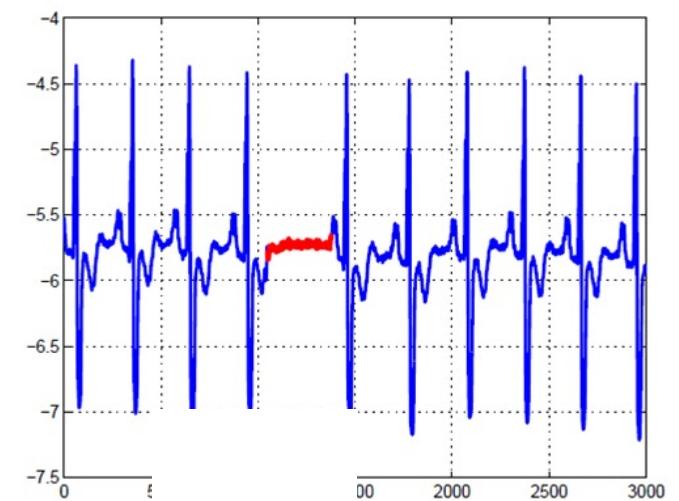
TYPES OF ANOMALIES



point



contextual

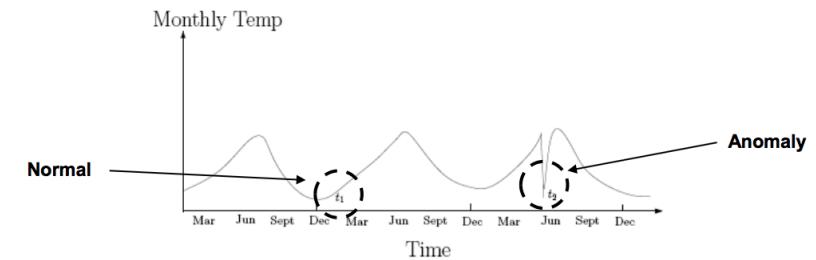


collective

TYPES OF ANOMALIES

Een anomaly kan binnen de variance vallen, en daarmee niet opvallen als point-anomaly.

Maar binnen de ‘context’ van de meetingen eromheen, kunnen dezelfde waarden op bepaalde momenten toch een anomaly zijn.

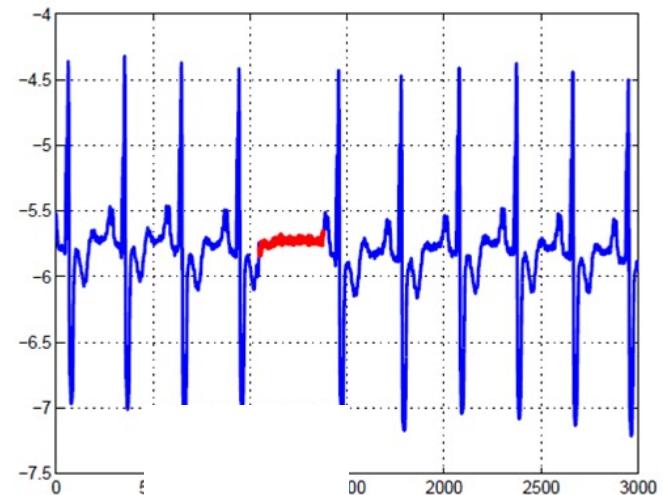


contextual

TYPES OF ANOMALIES

Sommige metingen wijken niet af wanneer we enkel naar de directe context van metingen kijken.

Kijkend naar *alle* metingen, valt het toch op dat de groep rode punten (wat *een verzameling van punten* is, anders dan bij een punt-anomaly of bij een context-anomaly) langer duurt dan in de rest van de observatie.



collective

CONFUSION MATRIX

CONFUSION MATRIX

De “verwarring” die ontstaat bij het maken van verkeerde voorspellingen kan worden weergeven in een confusion matrix.

FP is een type I error
FN is een type II error

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

CONFUSION MATRIX

Type I Error



False positive

Type II Error



False negative

CONFUSION MATRIX

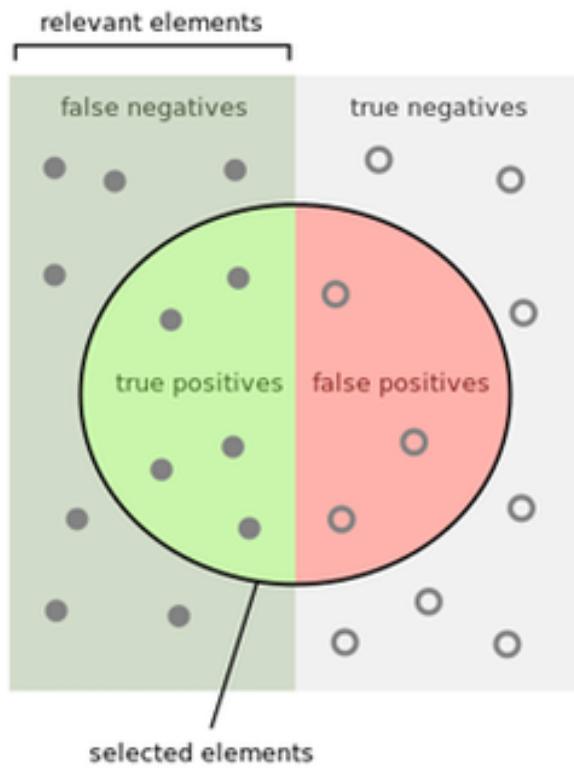
AirBnB heeft software die background checks doet via analyse van o.a. social media om in kaart te brengen hoe betrouwbaar je persoonlijkheid is (o.a. narcisme, psychopathie, etc).

Wat betekent hier een FP? En een FN?

Welke fout is voordeeliger voor AirBnB?
En voor de verhuurder?
En voor de huurder?

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

PRECISION VS RECALL



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

PRECISION VS RECALL

Recall wordt ook wel sensitivity genoemd.

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

Er is helaas geen standaard die zegt of actual/predicted in de kolommen/rijen moeten (zoals hier rechts), of andersom.

Sommige paketten draaien dit weer om.

	Actual positive	Actual negative
Predicted positive	Precision	
Predicted negative		

	Actual positive	Actual negative
Predicted positive	recall	
Predicted negative		

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. (TP+FN)

Strategie: schieten met hagel

Je selecteert heel veel gevallen, bijvoorbeeld wel 90% (TP+FP), zodat je alle TP gevallen vindt maar ten koste van heel veel FP, en heel weinig FN.

In welke gevallen zou je zo iets willen doen?

	Actual positive	Actual negative
Predicted positive	Precision	
Predicted negative		

	Actual positive	Actual negative
Predicted positive	recall	
Predicted negative		

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. ($TP+FN=100$)

Strategie: schieten met hagel

Je selecteert bijna 90% ($TP+FP=890$), zodat je alle TP (90) gevallen vindt maar ten koste van heel veel FP (800), en heel weinig FN (10).

Heb je nu een hoge of lage precision?
En een hoge of lage recall?

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

	Actual positive	Actual negative
Predicted positive	90	800
Predicted negative	10	100

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. ($TP+FN=100$)

Strategie: schieten met hagel

Je selecteert bijna 90% ($TP+FP=890$), zodat je alle TP (90) gevallen vindt maar ten koste van heel veel FP (800), en heel weinig FN (10).

Heb je nu een hoge of lage precision?
En een hoge of lage recall?

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{90}{90+800} = 0,10$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{90}{90+10} = 0,9$$

	Actual positive	Actual negative
Predicted positive	90	800
Predicted negative	10	100

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. (TP+FN)

Strategie: scherpschutter

Je selecteert maar 1% ipv 10%, maar als je iemand selecteert maak je eigenlijk nooit een fout.

In welke gevallen zou je zo iets willen doen?

	Actual positive	Actual negative
Predicted positive	Precision	
Predicted negative		

	Actual positive	Actual negative
Predicted positive	recall	
Predicted negative		

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. (TP+FN=100)

Strategie: scherpschutter

Je selecteert maar 1% (TP+FP=11) ipv 10%, maar als je iemand selecteert maak je eigenlijk nooit een fout (FP=1).

Heb je nu een hoge of lage precision?
En een hoge of lage recall?

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

	Actual positive	Actual negative
Predicted positive	10	1
Predicted negative	90	899

PRECISION VS RECALL

Stel, 10% van de gevallen die je test zou positief gelabeld moeten worden. ($TP+FN=100$)

Strategie: scherpschutter

Je selecteert maar 1% ($TP+FP=11$) ipv 10%, maar als je iemand selecteert maak je eigenlijk nooit een fout ($FP=1$).

Heb je nu een hoge of lage precision?
En een hoge of lage recall?

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{10}{11} = 0,91$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{10}{10+90} = 0,1$$

	Actual positive	Actual negative
Predicted positive	10	1
Predicted negative	90	899

PRECISION VS RECALL

Het soort fout dat je maakt heeft gevolgen voor de performance. Elk type fout heeft een andere consequentie.

Welk type fout zwaarder weegt, hangt af van de context! Dit is vaak een ethische afweging die de opdrachtgever moet maken (en niet de programmeur).

Conclusie 1: je **wilt dus niet perse de hoogste performance**. Soms is de ene soort fout erger dan de andere soort fout!

Conclusie 2: als opdrachtgever moet je richting geven aan de verhouding precision-recall.