



Odisee  
DE CO-HOGESCHOOL

# Data Science – week 3



Jens Baetens

## How to participate?



Click on the projected screen to start the question

 [Copy participation link](#)

wooclap

 100 % 

41 



Go to **wooclap.com** and use the code **SGEEOD**



Waar kan je data verzamelen?



api's en  
scraping

waar  
niet?



bestaand  
e  
datasets  
(statbel,  
...)

Databas  
e

statbel



interview  
s

data van  
overheid

github

files  
(csv's  
b.v.b.)



datafram  
e

Overheid  
sdatahan

formulier



**wooclap**



100 %



20



Welk keyword komt overeen met welk soort data?



Duur maar gericht



Primair

Goedkoper maar niet specifiek



Secundair

Tabelvormig



Gestructureerd

Figuren, audio, ...



Ongestructureerd

Beschrijving




Qualitatief

Numeriek



Quantitatief

Click on the projected screen to start the question

 You cannot vote anymore



Wat zijn de te volgen stappen in de data science lifecycle



✓ The correct combination was:

Business understanding

Data mining

Data cleaning

Data exploration

Feature engineering

Predictive modeling

Data vizualisation



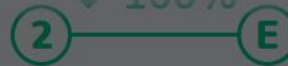
Click on the projected screen to start the question



Wat is het belangrijkste dat er gebeurt in elke stap van de Data Science lifecycle



Data Mining



Verzamelen van data

Data Cleaning



Oplossen van fouten in de data

Data Exploration



Zoeken naar verbanden in de data

Feature Engineering



Split de data af die je gaat gebruiken

Predictive modelling



Train een model voor de vraag te beantwoorden

Data Visualization



Rapporteer je resultaten

Click on the projected screen to start the question



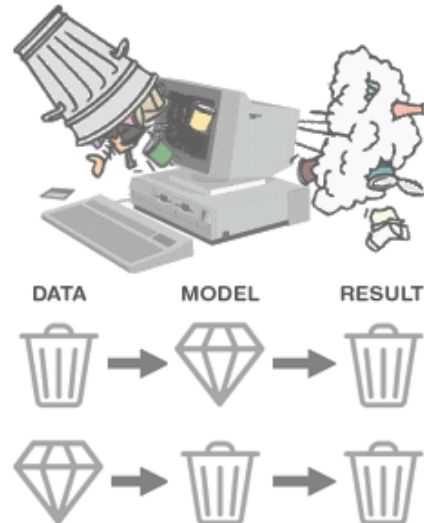
# Data Cleaning





# Garbage in = Garbage out

- ▣ Fouten in data leiden tot foute conclusies
- ▣ Hoe langer de fouten in de data aanwezig zijn
  - Hoe kostelijker / tijdrovender / complexer om het op te lossen



## Welke fouten kunnen er in data aanwezig zijn?

Go to **wooclap.com** and use the code **SGEEOD**



Welke fouten kunnen er aanwezig zijn in datasets



FOUT OMGEZETTE DATA VERKEERD FORMAAT  
TRUE/FALSE VS 0/1 MISMATCH INCORRECTE DATATYPE  
OUTLIERS  
TYPFOUTEN DUBBLICTIE DATUMS FOUTE RELATIES LEGE DATA  
VERTALINGSFOUT RIJS NIET REPRESENTATIEVE SCHRUFFOUTEN  
FOUT FORMAAT VERSCHILLENDE TALEN FORMATTERINGSFOUTEN  
ONTBREKENDE DATA

Click on the projected screen to start the question





## Welke fouten kunnen er in data aanwezig zijn?

- ▣ Ontbrekende data
- ▣ Duplicaten
- ▣ Foutieve waarden
- ▣ Onmogelijke waarden
- ▣ Verkeerde dataformaten / eenheden
- ▣ Onnodige data



## Hoe los je ontbrekende data op?

- ▣ Zoek de data op online of in andere datasets
- ▣ Kies een vaste waarde
- ▣ Bereken de waarde
  - Gemiddelde, minimum, maximum, gelijkaardige rij, ....
- ▣ Verwijder de bijhorende rij
- ▣ Is het altijd een probleem?



## Hoe los je duplicate data op?

- ▣ Gebruik duplicaten om ontbrekende informatie in te vullen
- ▣ Verwijder de duplicaten zodat er slechts 1 entry overblijft
- ▣ Wat met conflicterende data?

## Foutieve dataformaten

- ▣ Typo's waardoor numerieke kolom als tekst gezien wordt
- ▣ Soms True/False soms 0/1
- ▣ Plaatsen in verschillende talen: Brussel, Brussels, Bruxelles
- ▣ Straat en nummer in 1 kolom ipv 2
- ▣ Datums yyyy/mm/dd vs dd/mm/yyyy

### PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27


THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:


02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. 27/2-13 2013.158904109

MMXIII-II-XXVII MMXIII <sup>LVII</sup>CCCLXV 1330300800

$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  2013 

10/1101/1101 02/27/20/13 



## Privacy requirements

- ▣ Zoek naar Personal Identifiable Information (PII)
- ▣ Deze velden moeten beter afgeschermd zijn dan andere
  - Data niet bruikbaar in het geval van hacking

## Privacy requirements – Data masking

# Examples of data masking techniques

PATIENT RECORD	Original production database		Development database with masked data	
	PATIENT NUMBER	113355	SCRAMBLING	100100
	NAME	John West	SHUFFLING	Peter Church
	ADDRESS	45 Broad Street	SUBSTITUTION	12 Johnson Square
	CITY/STATE/ZIP	Sunnyview, CA 90261		Rochdale, CA 91331
	SSN	778-62-8144		805-14-1893
	DOB	10/07/1972	VARIANCE	02/18/1975
	CREDIT CARD NUMBER	4145 1230 0000 0062	MASKING OUT	XXXX XXXX XXXX 0062
	FILE	mri_results.pdf	NULLIFYING	NULL



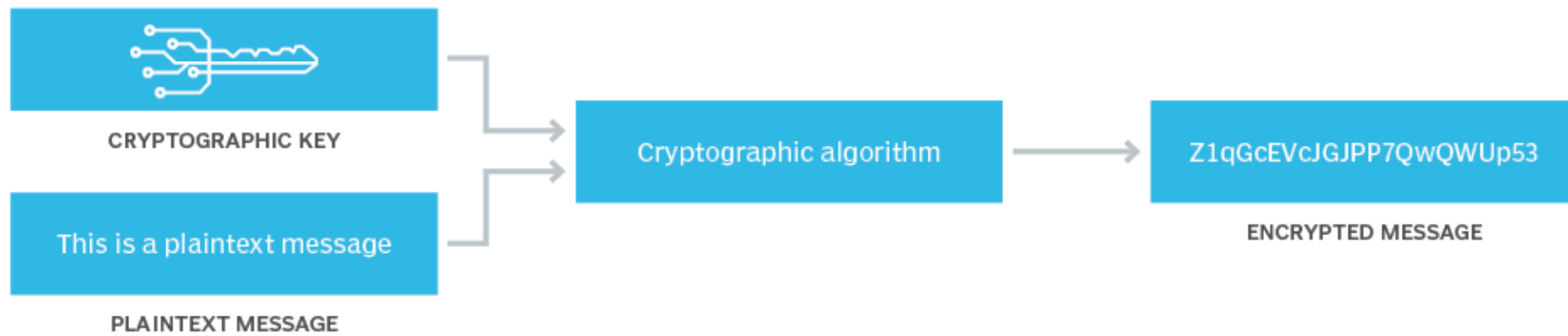


## Data masking best practices

- ▣ Localizeer en bescherm **alle** gevoelige data
- ▣ Hou ook rekening met niet-gestructureerde data
- ▣ Geef enkel toegang tot masked data aan vertrouwde personen
- ▣ Test je masking methoden

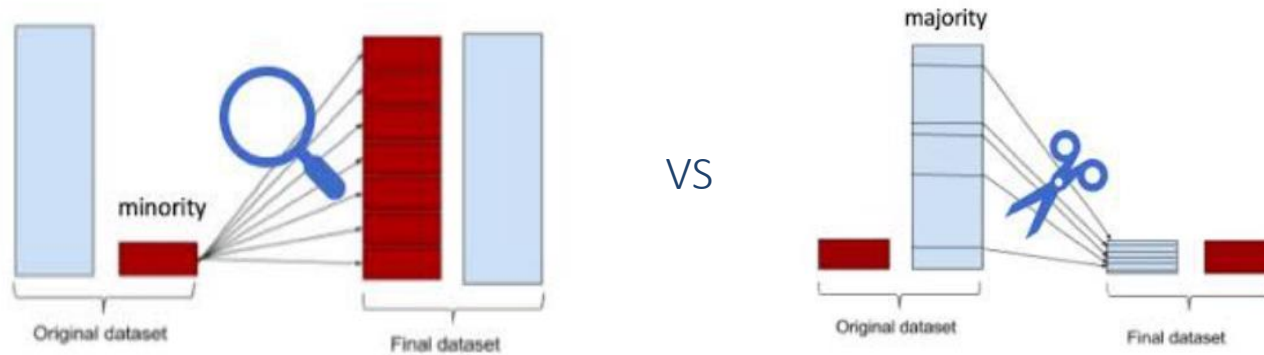
## Privacy requirements - Encryptie

### Encryption operation



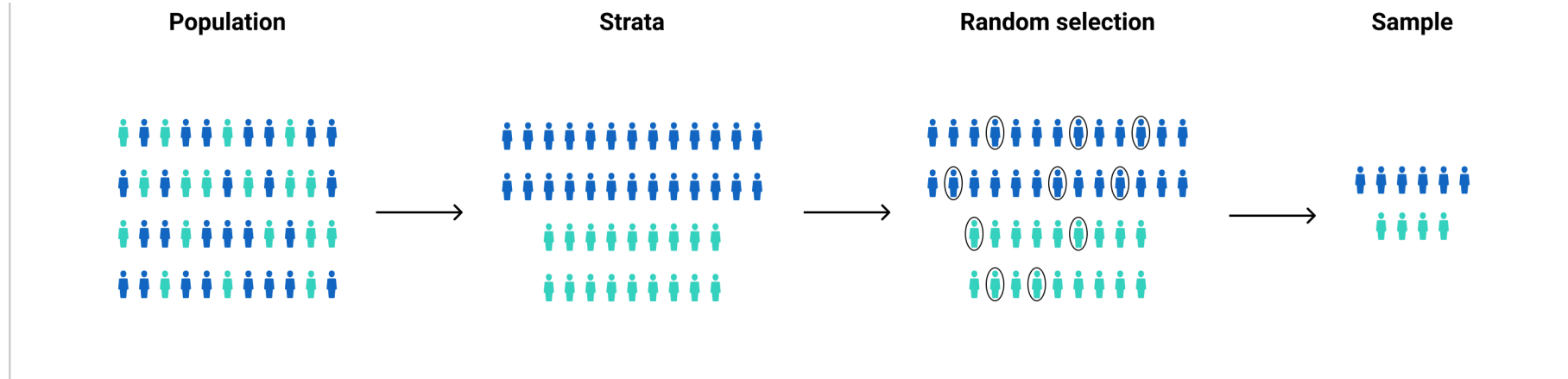
# Data balancing

- ▣ Waarden in een rij komen zelden voor
  - Best niet verwijderen



Source: <https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>

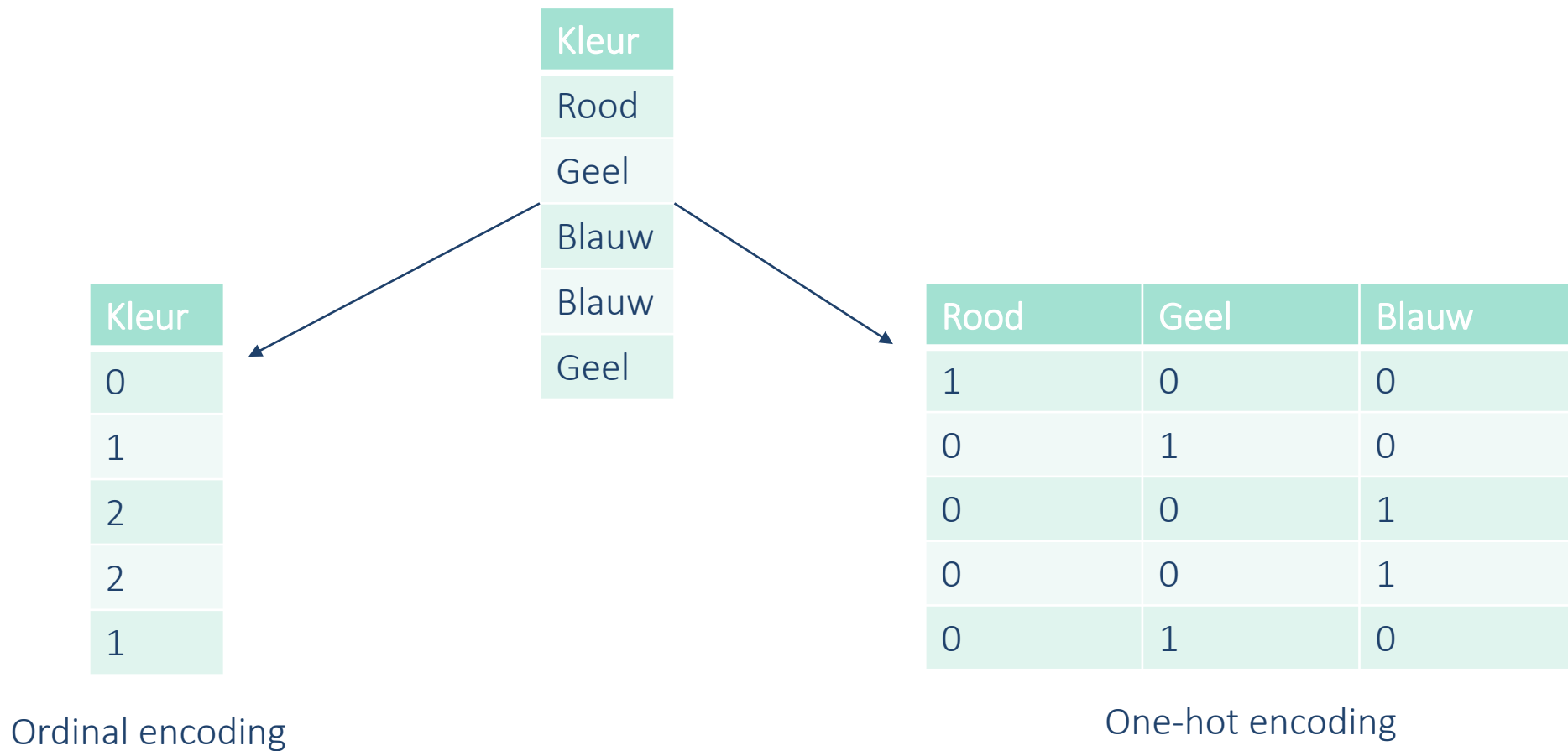
# Stratified sampling



Source: <https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>

- ▣ Gebeurt vaak automatisch bij ML-technieken

# Encodings





# Data Quality Issues and Solutions



# Importance of Data Quality

Using EPL as an example



## Samenvatting

- ▣ Bij data cleaning, hou rekening met:
  - Datatypes van verschillende kolommen
  - NaN, Null waarden
  - Outliers
  - Foutieve dataformaten bij datums, tekst
  - Hoe categorieke data bijgehouden wordt
  - Persoonsgegevens



You cannot vote anymore



Een gebalanceerde dataset is



1

een dataset met evenveel rijen als kolommen

24%

5



2

heeft ongeveer evenveel rijen van elke klasse/ categorie

57%

12



3

een dataset waar er geen onbruikende data is

19%

4



4

een dataset zonder persoonlijke gegevens

0%

0



Click on the projected screen to start the question

wooclap



100 %



57% correct

21 / 41



Go to **wooclap.com** and use the code **SGEEOD**



Welke technieken kunnen gebruikt worden om een niet-gebalanceerde ...



1

Invullen ontbrekende data

18%

3



5

Toevoegen van variantie

6%

1



2

Undersampling

35%

6



6

Oversampling

41%

7



3

Data Shuffling

12%

2



7

Schalen van data

29%

5



4

Stratified Sampling

88%

15



8

Encoderen van categorieke data

0%

0



Click on the projected screen to start the question



Go to **wooclap.com** and use the code **SGEEOD**



Welke opties zijn er om tegemoet te komen aan privacy requirements in ee...



Substitut  
ion

Shuffling

encryptie

Encriptie  
van data

masking



variantie  
toevoege  
n

masking

masking

Masking

variance

scrambli  
ng

masking

encryptio  
n

Data  
Shuffle



**wooclap**



100 %



21



Click on the projected screen to start the question

Go to **wooclap.com** and use the code **SGEEOD**



Welke functie kan gebruikt worden om het aantal NaN waarden te ...



1

fillna

0%

0



2

drop

0%

0



3

isna

100%

2



4

dropna

0%

0



5

info

100%

2



wooclap



100 %



2 / 41



Go to **wooclap.com** and use the code **SGEEOD**



Welke zin is correct?



1

fillna kan voor elke kolom een aparte waarde invullen

0%

0



2

fillna kan enkel NaN waarden invullen in het hele dataframe door ze te vervangen door dezelfde waarden

0%

0



3

fillna kan enkel categorieke data invullen

0%

0



4

fillna kan enkel kolommen met numerieke data invullen

0%

0



Click on the projected screen to start the question

wooclap



100 %



0 / 41



Go to **wooclap.com** and use the code **SGEEOD**



Waarom zou je One-Hot encoding gebruiken ipv Ordinal Encoding? ...



1

Om je data te comprimeren

0%

0



2

Om het modelleren te verbeteren

0%

0



3

Om fouten in de data te detecteren

0%

0



4

Om meerdere klassen toe te kennen

0%

0



Click on the projected screen to start the question

wooclap



100 %



0 / 41



Go to **wooclap.com** and use the code **SGEEOD**



Data shuffling kan je doen door de volgende functie te gebruiken



1

groupby

0%

0



2

join

0%

0



3

sample

0%

0



4

replace

0%

0



5

reset\_index

0%

0



Click on the projected screen to start the question



wooclap



100 %



0 / 41





## Belangrijke termen

- ▣ Primaire data
- ▣ Secundaire data
- ▣ Quantitatieve data
- ▣ Qualitatieve data
- ▣ Gestructureerde data
- ▣ Niet-gestructureerde data





## Data cleaning tutorial

### ▣ Ga naar:

- <https://www.kaggle.com/learn/data-cleaning>
- Volg de tutorial
  
- De informatie in de tutorials is te kennen leerstof en helpt bij het maken van de oefeningen



## EDA oefening

- ▣ Opgave: <https://classroom.github.com/a/UAH3I74x>
- ▣ Maak de oefening individueel tegen volgende week
- ▣ De deadline is 5 november om 23u59