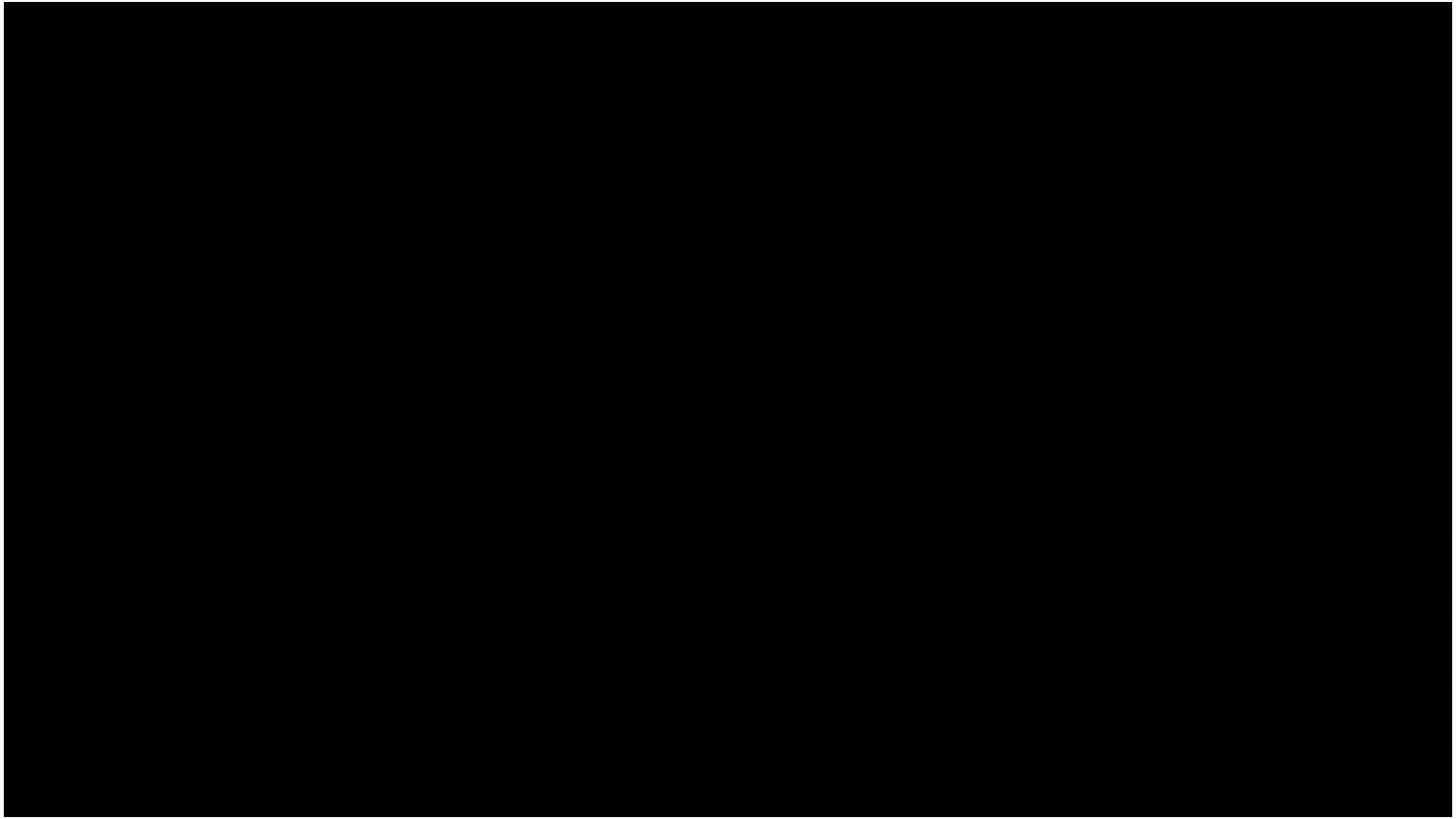
DATA CLEANING

JENS BAETENS

https://www.youtube.com/watch?v=Mp_o_qsyBhA



https://www.youtube.com/watch?v=-wibyFVm6yg

KWALITEIT VAN DATASETS IS BELANGRIJK

Garbage in = Garbage out

Hoe langer je fouten meesleept in je dataset, hoe kostelijker

- Zelfde als bij software development

Fouten in je data kan leiden tot een fout model

PROBLEMEN IN DE DATASETS

Ontbrekende data → missing NaN

Duplicaten

Onmogelijke waarden → leeftijd is negatief

Verkeerde dataformaten → datums: dd/mm/yy
yy-mm-dd
yy-maand-dd
...

Onnodige attributen verwijderen
↳ ID's
↳ Foreign keys

ONTBREKENDE DATA

Ontbrekende data:

- Zoek de data op online
- Bereken de waarde (gemiddelde, gelijkaardige rijen, …)
- Verwijder duplicaten (rij of kolom)

Niet altijd problematisch: gebrek aan data kan ook een bron van informatie zijn

DUPLICATE DATA

Verwijder exacte duplicaten

Gebruik duplicaten om ontbrekende informatie in te vullen

Hoe behandel je conflicterende data?

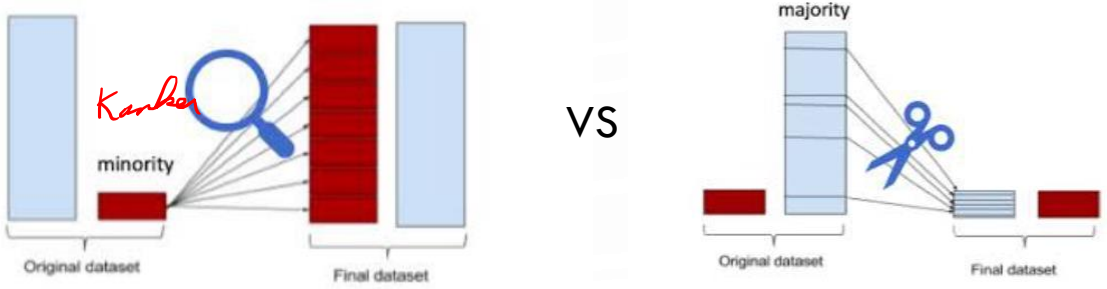
- Zelfde naam maar ander adres bijvoorbeeld

→ rij volledig matchen
→ # kolommen "
→ ID is voldoende

⇒ welke data
houje over

DATA BALANCING

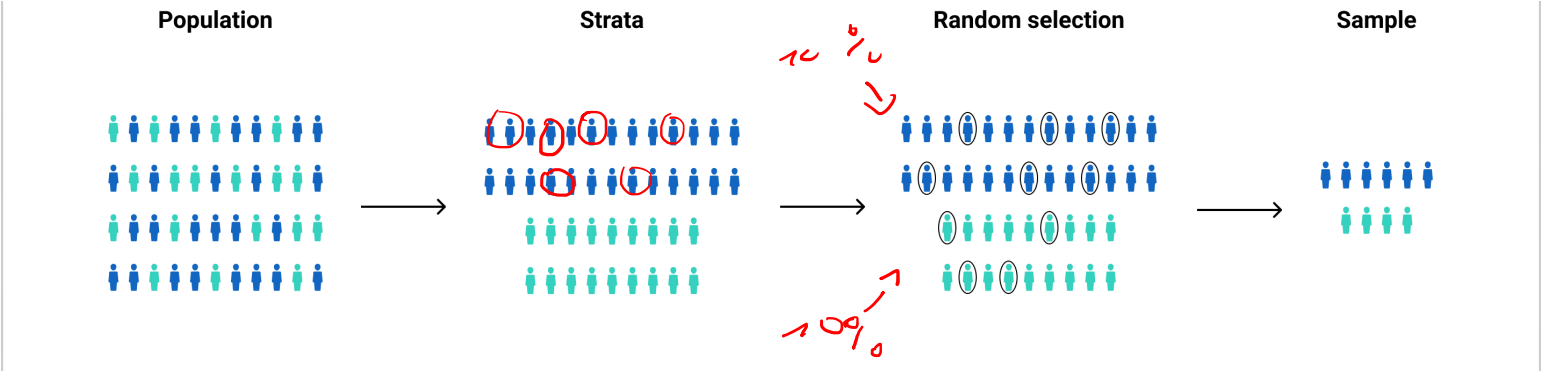
Rijen die waarden bevatten die zelden voorkomen worden best niet verwijderd



Stratification argument doet dit automatisch in train_test_split

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

DATA BALANCING – STRATIFIED SAMPLING



Stratification argument doet dit automatisch in train_test_split

Verdeling blijft behouden

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

CORRIGEER DATA FORMAT

Typos

Verschillende waarden met dezelfde betekenis:

- 0/1 of True/False of …
- Naam van een stad in verschillende talen
- Straat en nummer in 1 veld ipv 2
- Datums: yyyy/mm/dd vs dd/mm/yyyy

Brussel
Brussels
Bruxelles

Vertalen!

ONE HOT ENCODING VS ORDINAL

Kleur
Rood
Geel
Blauw
Blauw
Geel

Kleur
0
1
2
2
1

Rood	Geel	Blauw
1	0	0
0	1	0
0	0	1
0	0	1
0	1	0

PRIVACY REQUIREMENTS

Zoek naar Personal Identifiable Information (PII)

Deze velden moeten beter afgeschermd zijn dan andere

- Data niet bruikbaar in het geval van hacking

Data masking of obfuscation

op de data zelf

Big Data

→ data access control

DATA MASKING – DELETION

Verwijder de gevoelige informatie

Zeer simplistische aanpak

Verwijderde data kan niet meer gebruikt worden in je model

Geeft aan dat er zaken gewijzigd zijn

DATA MASKING - SUBSTITUTION

Wijzig PII door willekeurige zelf gekozen waarden

- Kan afhankelijk zijn van andere velden
- Vb: wijzig naam maar wel afhankelijk van het opgegeven geslacht

Shuffling

- Lijst met mogelijke waarden is de kolom zelf

DATA MASKING – VARIANTIE TOEVOEGEN

Voeg ruis toe aan de data

Numerieke waarden

- Tot 10% geeft nog bruikbare data voor prijzen / salarissen

Datums

- Afhankelijk van de toepassing
- tot 120 dagen variantie behoudt de verdeling in de kolom

DATA MASKING – ENCRYPTION

Encrypteer de gevoelige kolommen

Hashing: Indien enkel een identifier moet hebben → 1-way / original way

Encryption: Indien de waarde nog moet gebruikt kunnen worden

Nadelen:

- Beheer van keys → gestolen key → alles kan gelezen worden
- Encryptie en decryptie reken-intensief