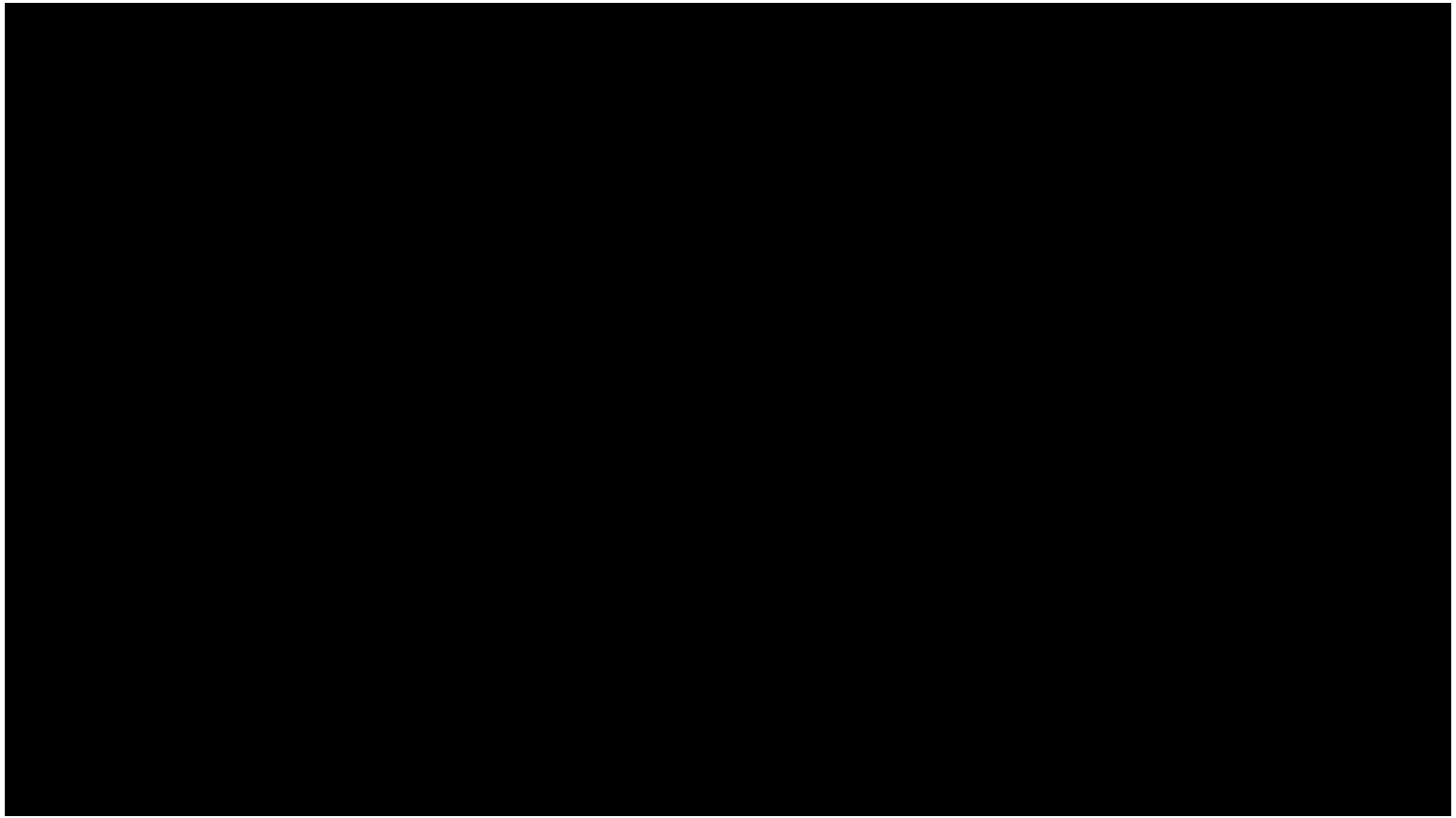
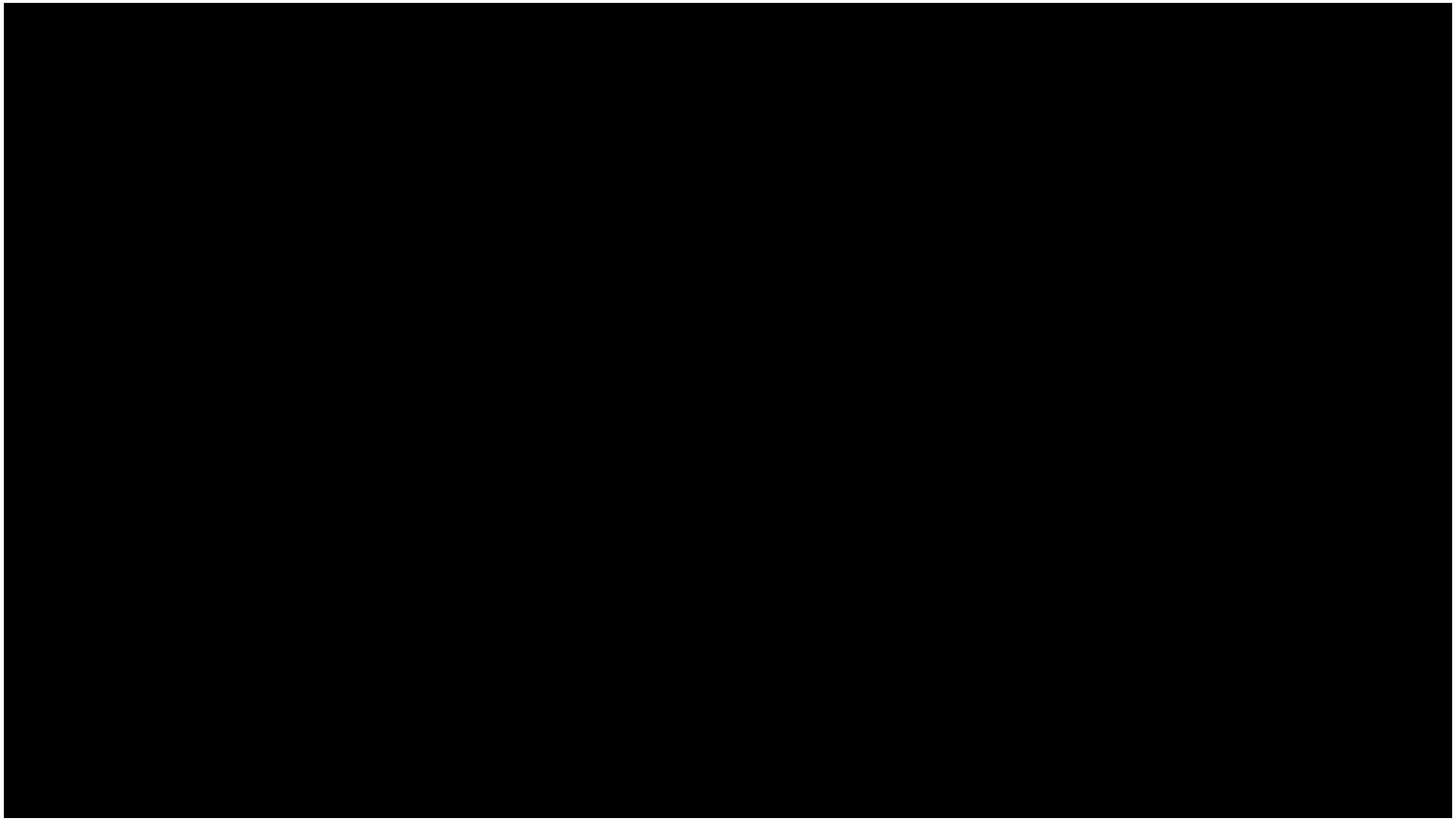
Data CLEANING

JENS BAETENS





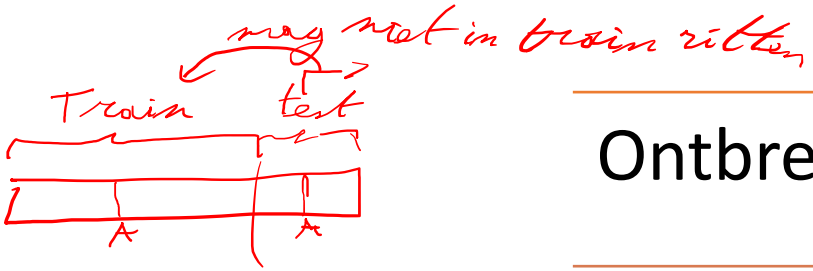
Kwaliteit van datasets is belangrijk

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
- Duplicaten → Data leakage
  - ↳ Problemen om model te evalueren
- Onmogelijke waarden → outliers
- Verkeerde dataformaten
  - ↳ lbs / kg
  - ↳ datums 13/3/0, 13:30, 1330
- Onnodige attributen verwijderen

naam oud nationaliteit
Jens 37 ?
Jens ? Belg

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

Age

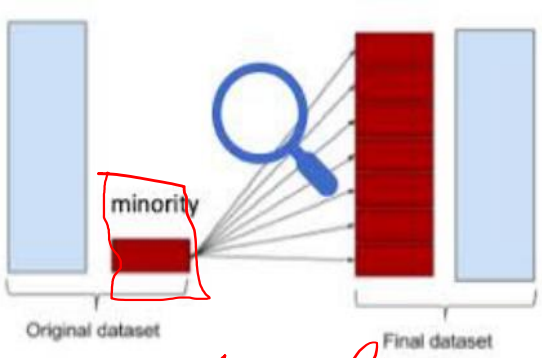
Jens 31

Jens 30

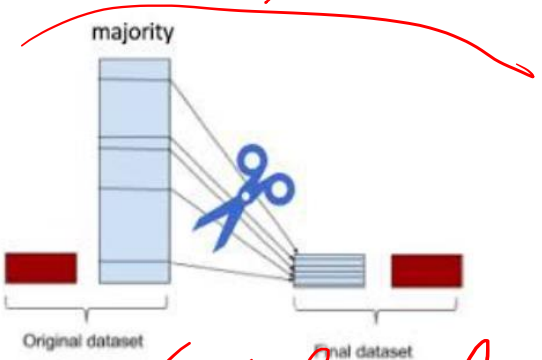
1. Verwijder exacte duplicaten
2. Gebruik duplicaten om ontbrekende informatie in te vullen
3. Wat met conflicterende data?

Data balancing

Rijen die waarden bevatten die zelden voorkomen worden best niet verwijderd



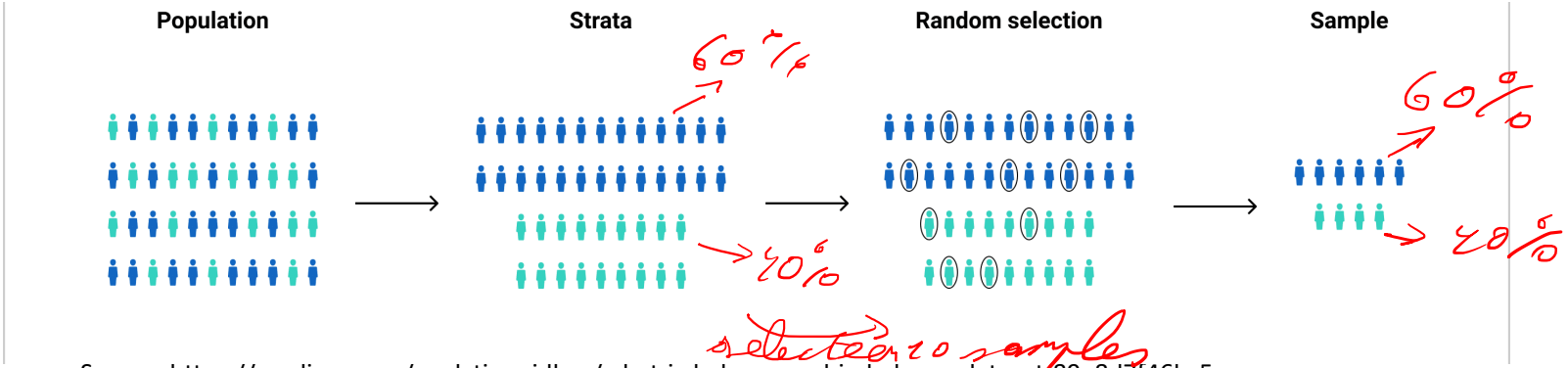
VS



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

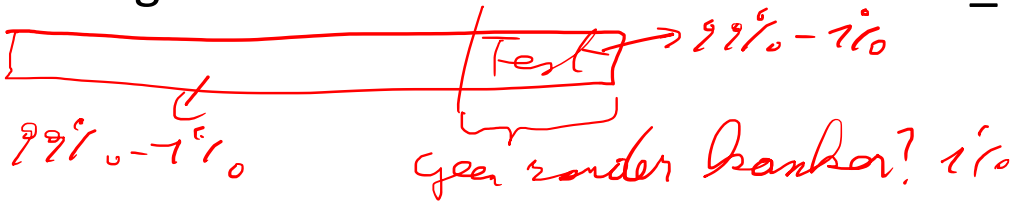
Data balancing – Stratified sampling

↳ kijken naar de verdeling



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

Stratification argument doet dit automatisch in train_test_split



Corrigeer data format

Typos

Verschillende waarden met dezelfde betekenis:

- 0/1 of True/False of … M/F → replace
- Naam van een stad in verschillende talen
- Straat en nummer in 1 veld ipv 2
- Datums: yyyy/mm/dd vs dd/mm/yyyy
  - altijd complex probleem
  - 10/12/2022
  - december of oktober?

One hot encoding vs ordinal

categorische data ↑ -> 3 klassen

Kleur
Rood
Geel
Blauw
Blauw
Geel

voor ML moet dit numeriek zijn

Rood -> 0
Geel -> 1
Blauw -> 2

Ordinal encoding
↳ kan met sommige replace

Kleur
0
1
2
2
1

one-hot encoding -> elke klasse apart kolom

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

kan problemen zijn bij veel klassen (>50)

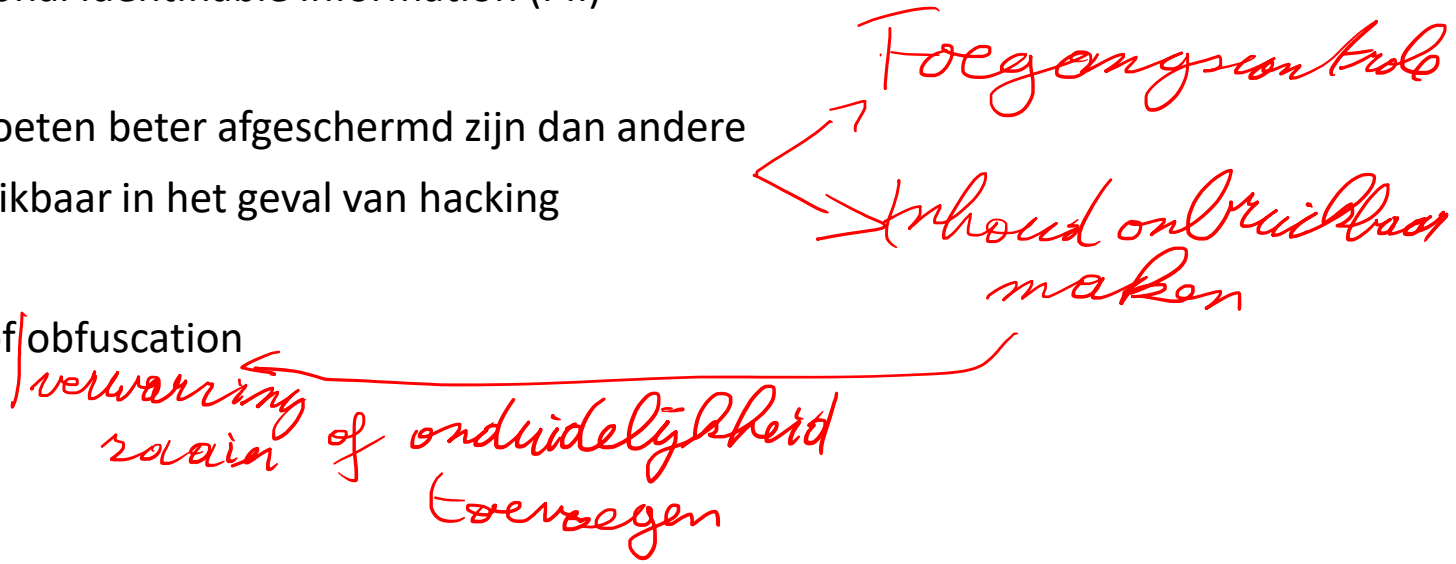
Privacy requirements

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



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

Vervangen

Wijzig PII door willekeurige waarden

Kan afhankelijk zijn van andere velden

Verwijderde data kan niet meer gebruikt worden in je model

Geeft aan dat er zaken gewijzigd zijn

→ moeilijker maar nog steeds detecteerbaar

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
Encryption: Indien de waarde nog moet gebruikt kunnen worden

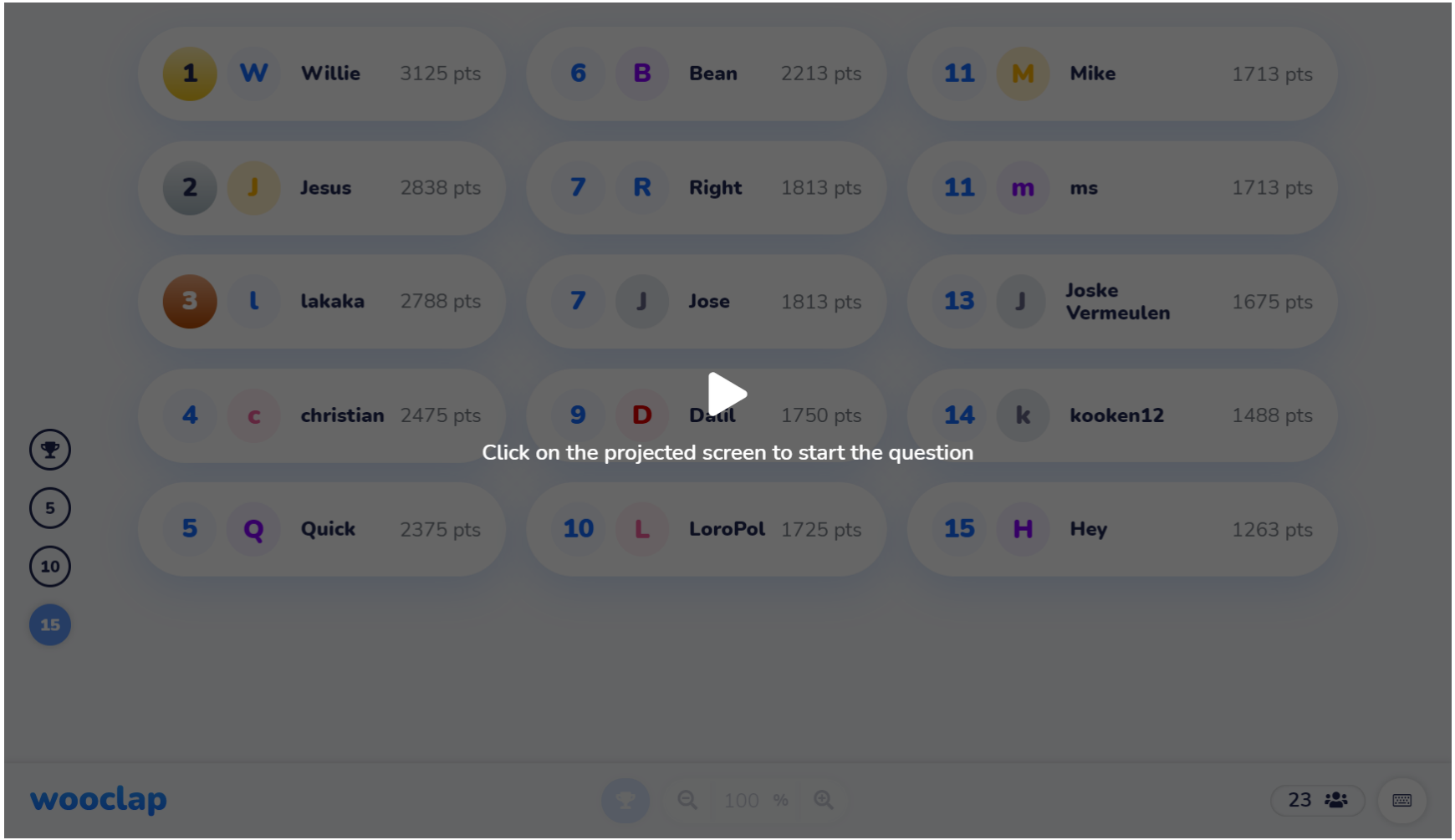
Nadelen:

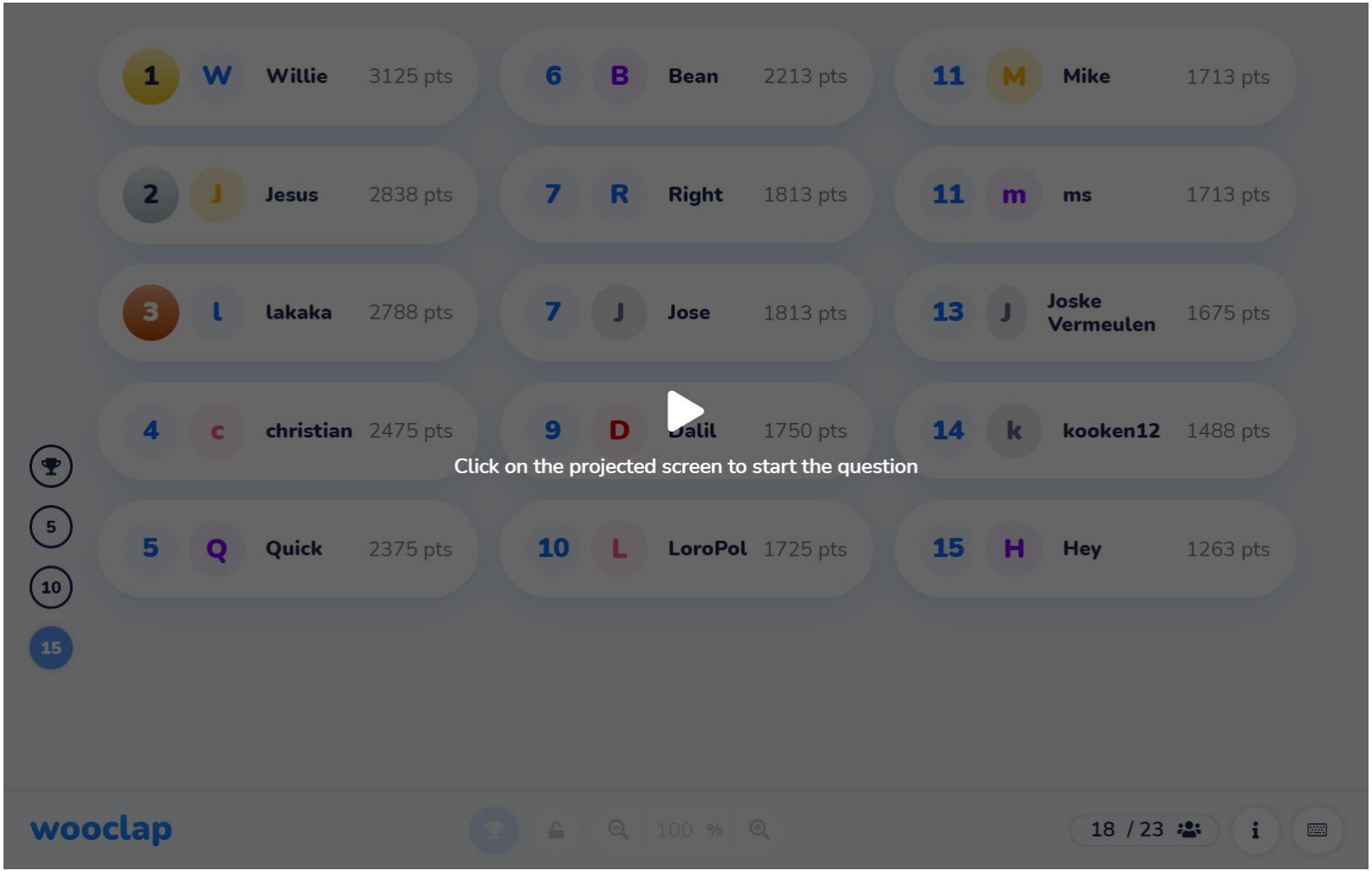
- Beheer van keys
- Encryptie en decryptie reken-intensief

Sleutel veilig → data | onleesbaar / vuilbaar

Samenvatting

- De volgende zaken vereisen je aandacht bij data cleaning
  - Datatypes van verschillende kolommen → astype / replace
  - NaN – waarden → isna / dropna / fillna
  - Outliers → volgens week
  - Foutieve dataformaten bij datums/tekst → to_datetime / replace
  - Gebruikte categorieën in categorieke kolommen
    - → unique / nunique / value_counts
    - → one-hot / ordinal
  - Persoonsgegevens
    - → deletion
      substitution
      variantie
      encryptie
      shuffling


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Dalil 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
23


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Dalil 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
18 / 23

1	W	Willie	3125 pts	6	B	Bean	2213 pts	11	M	Mike	1713 pts
2	J	Jesus	2838 pts	7	R	Right	1813 pts	11	m	ms	1713 pts
3	l	lakaka	2788 pts	7	J	Jose	1813 pts	13	J	Joske Vermeulen	1675 pts
4	c	christian	2475 pts	9	D	Dan	1750 pts	14	k	kooken12	1488 pts
5	Q	Quick	2375 pts	10	L	LoroPol	1725 pts	15	H	Hey	1263 pts

Click on the projected screen to start the question

5

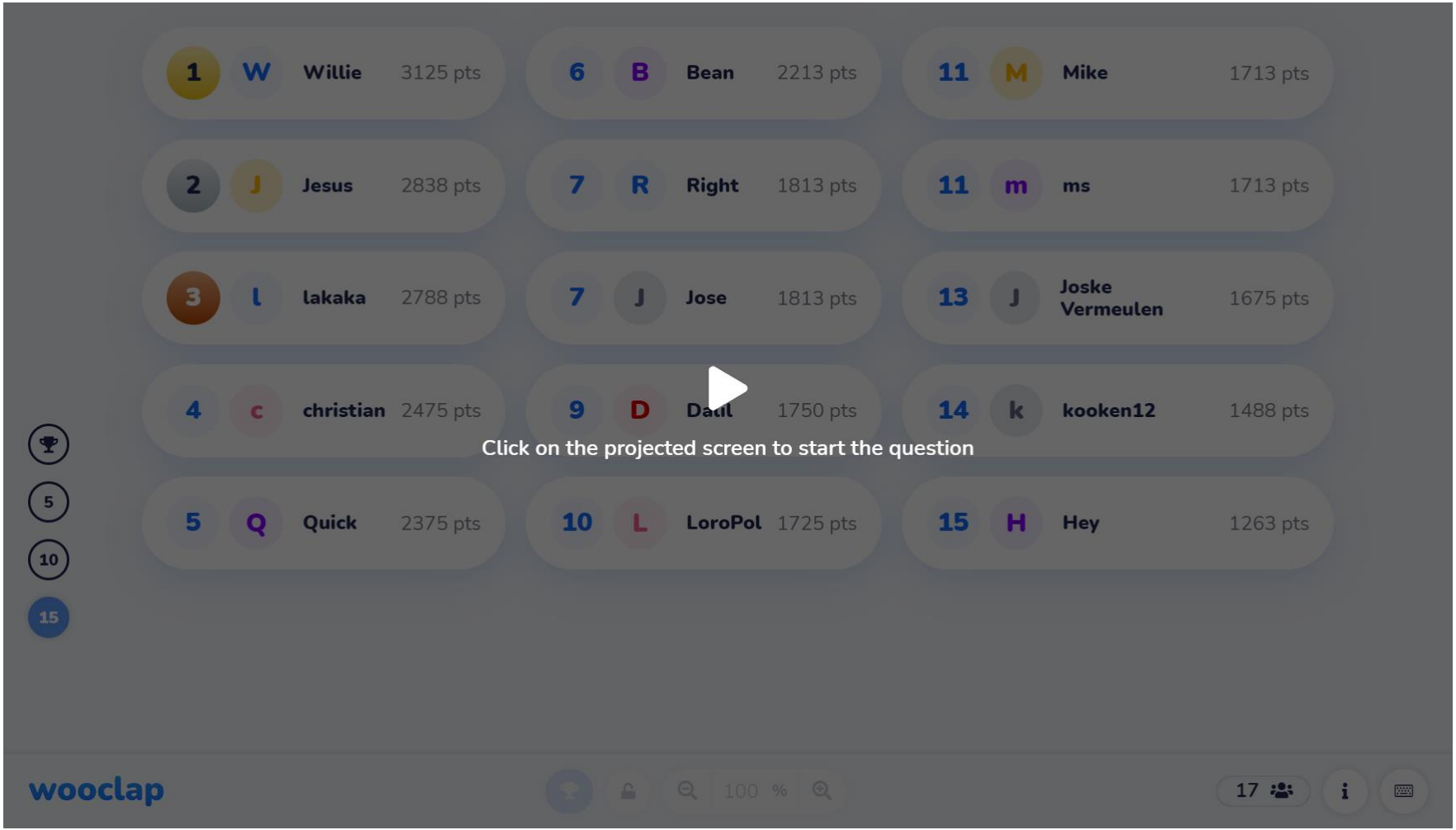
10

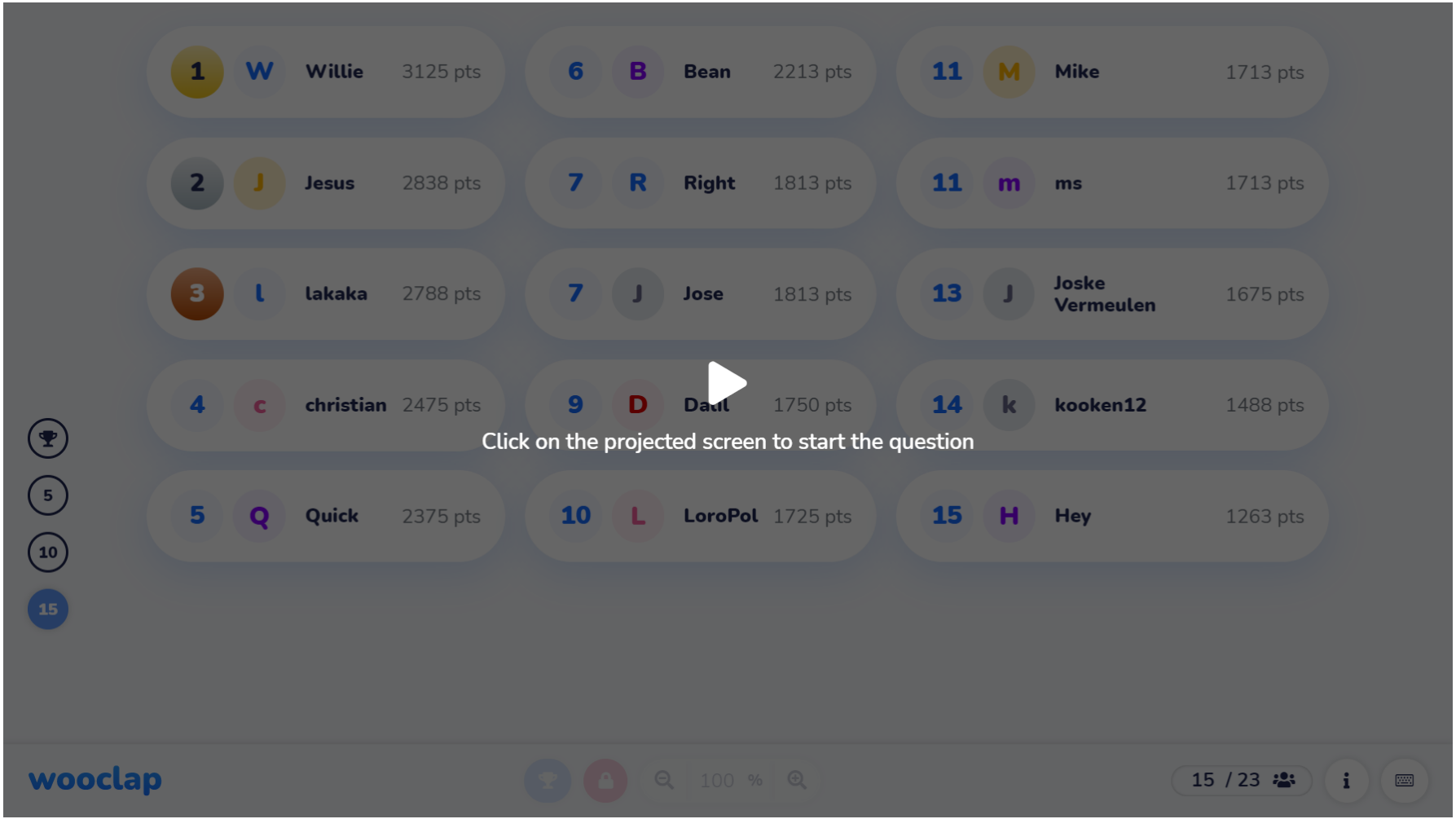
15

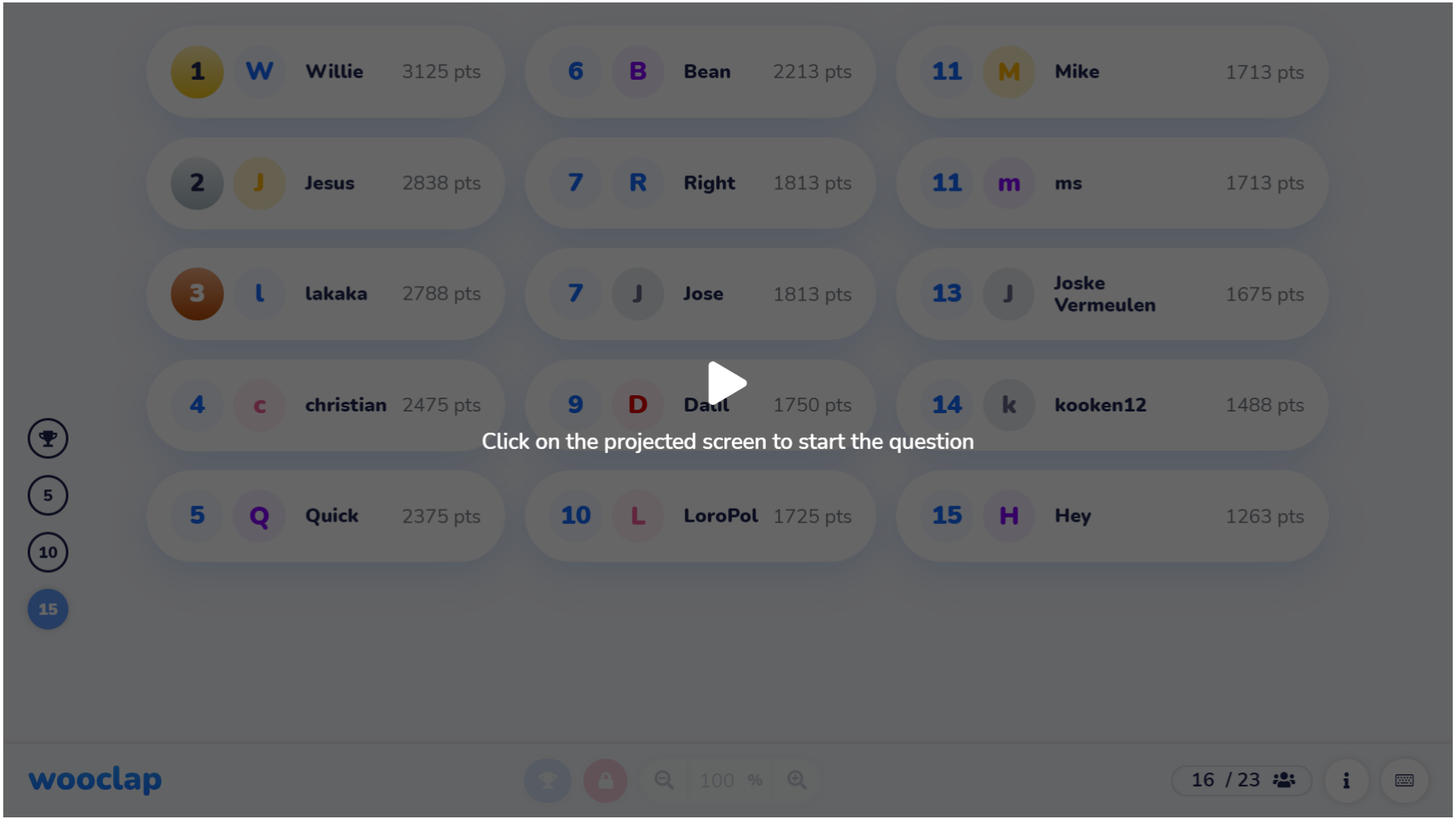
wooclap

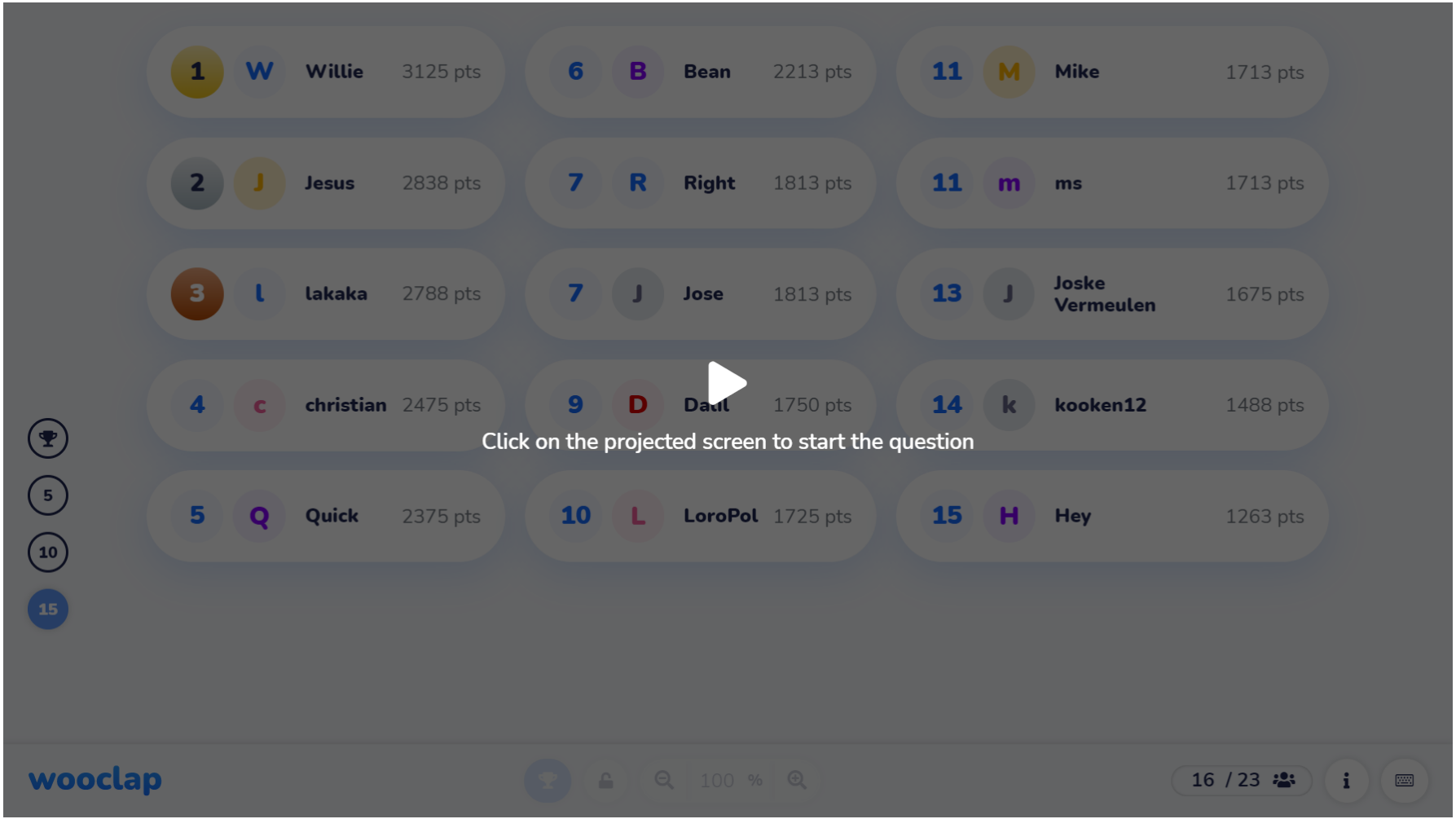
100 %

19 / 23


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Daan 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
17


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Daul 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
15 / 23


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Daul 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
16 / 23


1 W Willie 3125 pts
6 B Bean 2213 pts
11 M Mike 1713 pts
2 J Jesus 2838 pts
7 R Right 1813 pts
11 m ms 1713 pts
3 l lakaka 2788 pts
7 J Jose 1813 pts
13 J Joske Vermeulen 1675 pts
4 c christian 2475 pts
9 D Daul 1750 pts
14 k kooken12 1488 pts
Click on the projected screen to start the question
5 Q Quick 2375 pts
10 L LoroPol 1725 pts
15 H Hey 1263 pts
5
10
15
wooclap
100 %
16 / 23

1	W	Willie	3125 pts	6	B	Bean	2213 pts	11	M	Mike	1713 pts
2	J	Jesus	2838 pts	7	R	Right	1813 pts	11	m	ms	1713 pts
3	l	lakaka	2788 pts	7	J	Jose	1813 pts	13	J	Joske Vermeulen	1675 pts
4	c	christian	2475 pts	9	D	Dalil	1750 pts	14	k	kooken12	1488 pts
5	Q	Quick	2375 pts	10	L	LoroPol	1725 pts	15	H	Hey	1263 pts

5
10
15