



Data Preparation

Ténis na Bélgica

Grupo 4:

ALLAN KARDEC, N° 103380, TURMA CDB1
DIOGO FREITAS, N°104841, TURMA CDB1
JOÃO BOTAS, N°104782, TURMA CDB1
RICARDO ÂNGELO, N° 104826, TURMA CDB1

Data Preparation

3ª fase do CRISP-DM



Algumas
alterações
efetuadas

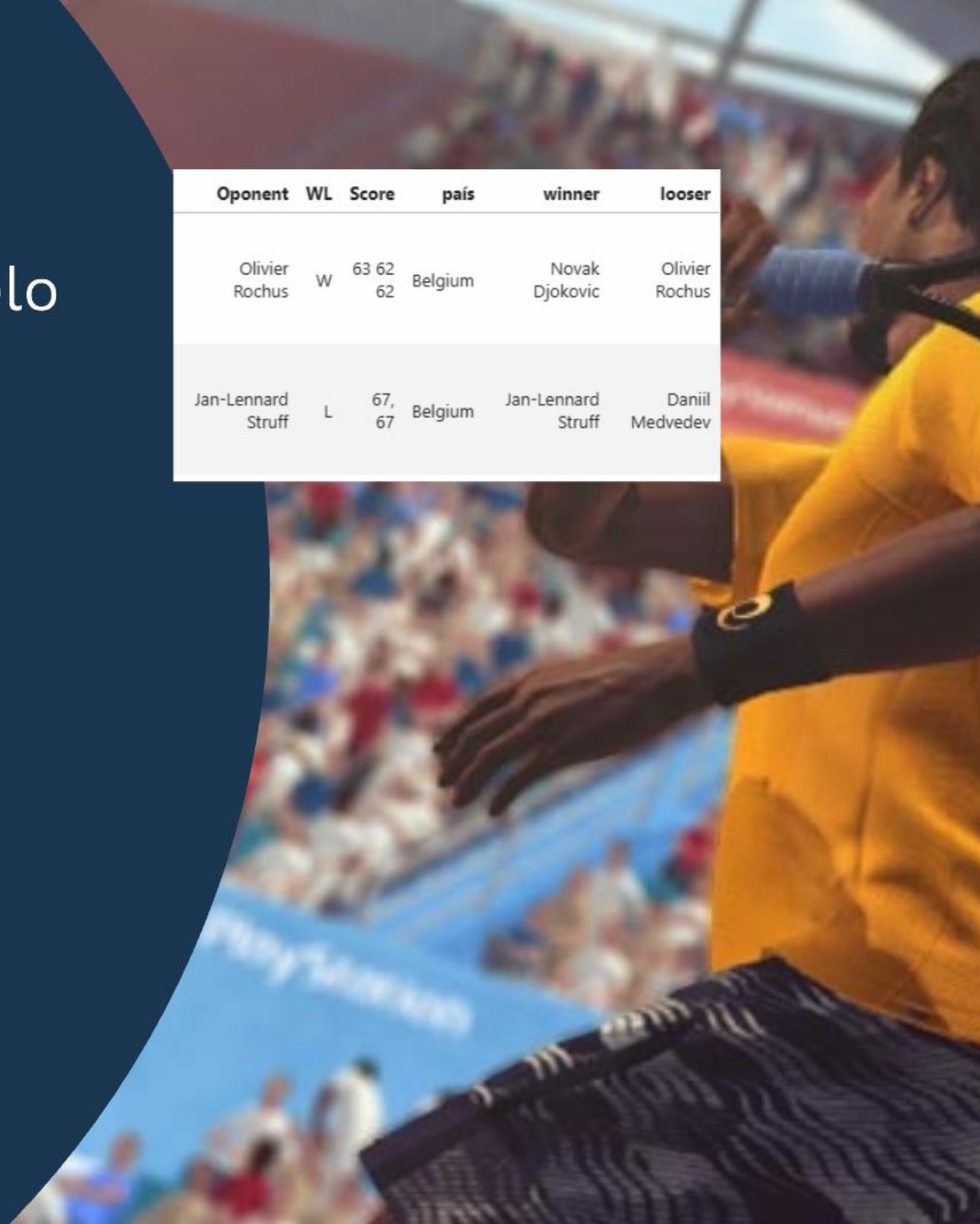
Tratamento
de variáveis
existentes

Novas
features-
Feature
Engineering

Conclusão e
Planeamento

Algumas alterações efetuadas

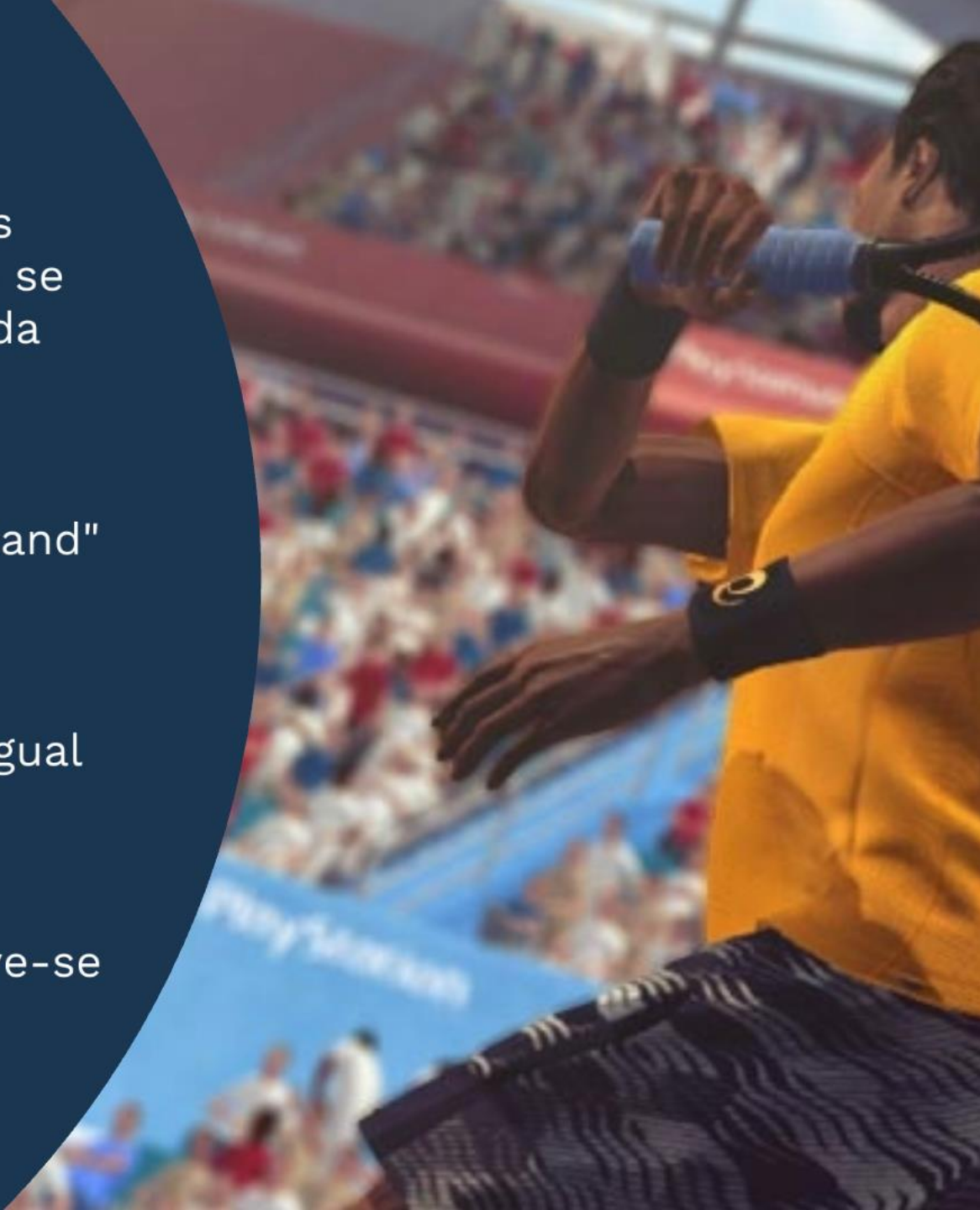
- Remoção de jogos espelhados pelo vencedor e perdedor do jogo (deixando os "W")
- Retificação dos gráficos, sem os jogos repetidos
- WebScrapping
 - Sites utilizados
 - Variáveis atualizadas
 - Novas variáveis



Oponent	WL	Score	país	winner	looser
Olivier Rochus	W	63 62 62	Belgium	Novak Djokovic	Olivier Rochus
Jan-Lennard Struff	L	67, 67	Belgium	Jan-Lennard Struff	Daniil Medvedev

Tratamento de variáveis existentes

- Height:
Os valores nulos foram substituídos pelos novos encontrados (os restantes que não se tinha os dados, foram corrigidos através da imputação)
- Hand:
Foi separado as "front-hand" das "back-hand"
- Prize
Preço uniformizado para euros (no Game Round "Round Robin" foi colocado prize igual a 0)
- Número de sets (nova variável)
Foi colocado os espaços corretos e obteve-se o número de sets através do número de espaços + 1 (nos casos do **W/O** foram colocados 0 sets)



Novas features - Feature Engineering

- Variáveis Dummy
 - Se o jogador é belga ou não
 - Ser destro ou canhoto
 - Se o GameRound possui a palavra "Final" no nome (são as rondas mais importantes)
 - Tipo de piso (3 possíveis apenas)
- Data do Torneio-> Apenas ano em que ocorreu o torneio
- Idade atual dos jogadores, através da sua data de nascimento
- Etc...

Para a próxima semana...

- Aplicar o que falta do Feature Engineering
- Ver novas features relevantes
- Pensar em modelos

Tentativas de sites para web scrapping:

- ATP Tour (<https://www.atptour.com/>)
- Tennis.com (<https://www.tennis.com/>)
- Tennis Explorer (<https://www.tennisexplorer.com/>)

