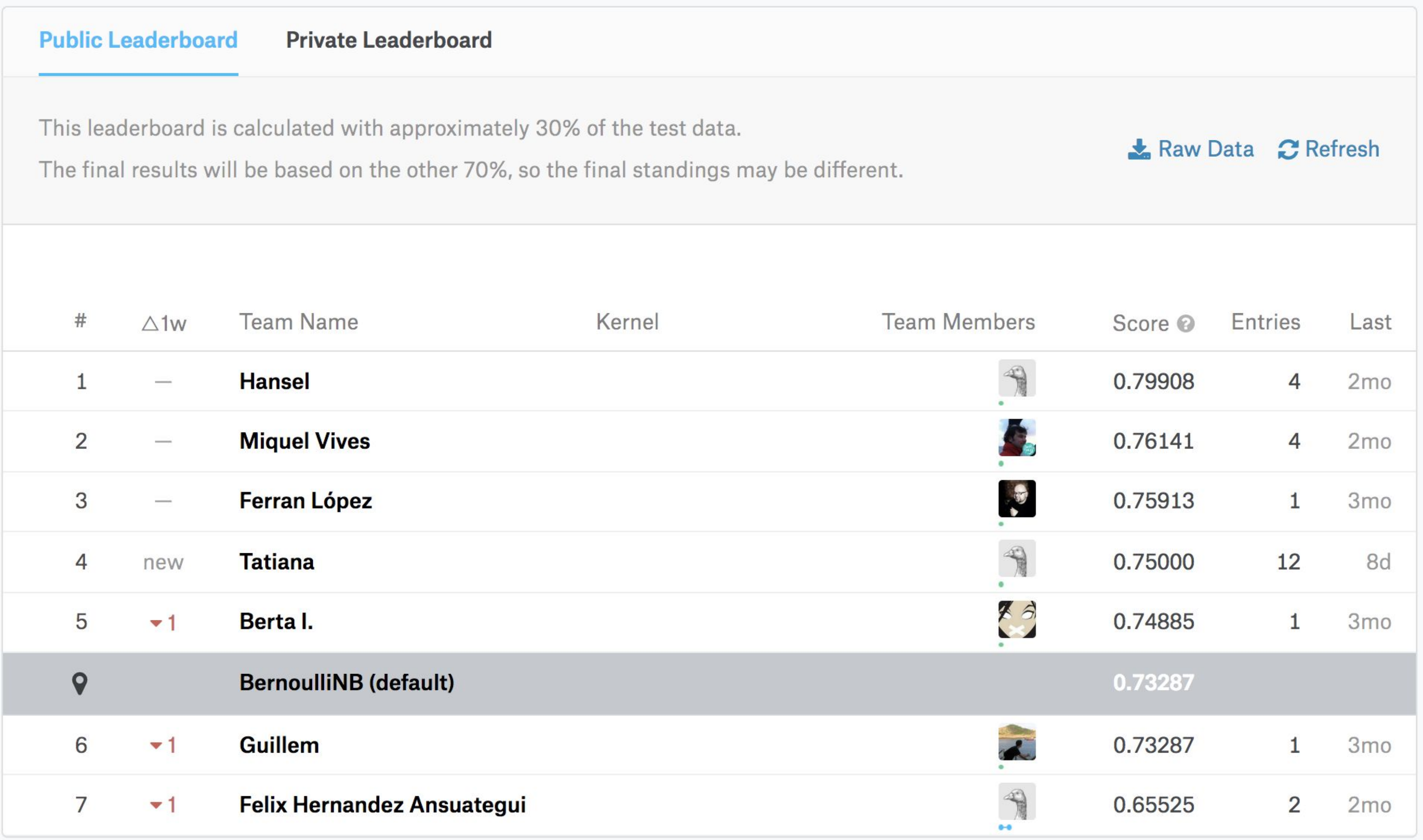
Capstone Project

Postgraduate Course on Data Science & Big Data

UNIVERSITAT DE
BARCELONA

Recap

- Baseline project
- Data engineering
- Kaggle



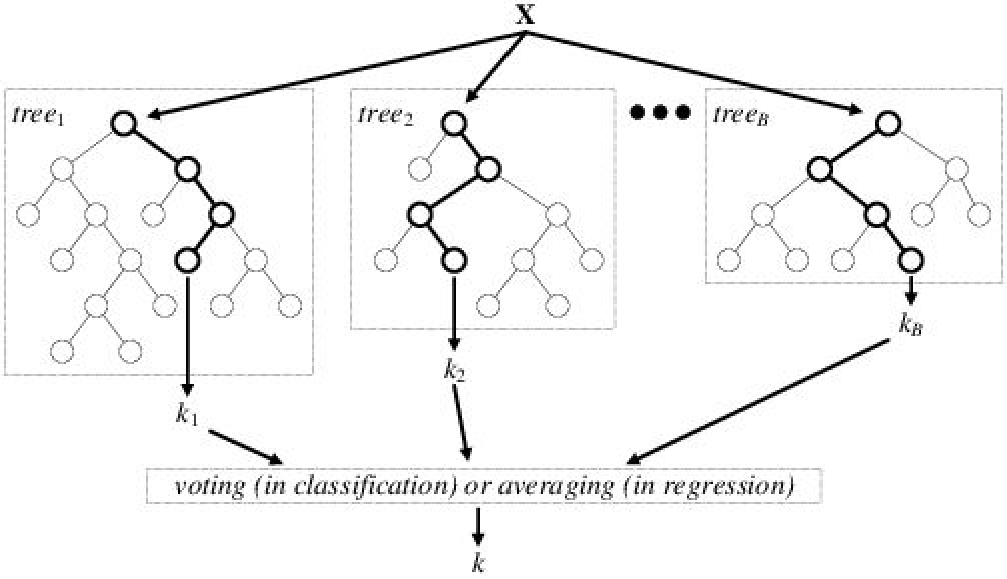
This leaderboard is calculated with approximately 30% of the test data.

The final results will be based on the other 70%, so the final standings may be different.

#	△1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	—	**Hansel**			0.79908	4	2mo
2	—	**Miquel Vives**			0.76141	4	2mo
3	—	**Ferran López**			0.75913	1	3mo
4	new	**Tatiana**			0.75000	12	8d
5	▼1	**Berta I.**			0.74885	1	3mo
📍		**BernoulliNB (default)**			0.73287		
6	▼1	**Guillem**			0.73287	1	3mo
7	▼1	**Felix Hernandez Ansuategui**			0.65525	2	2mo

What’s next (I)

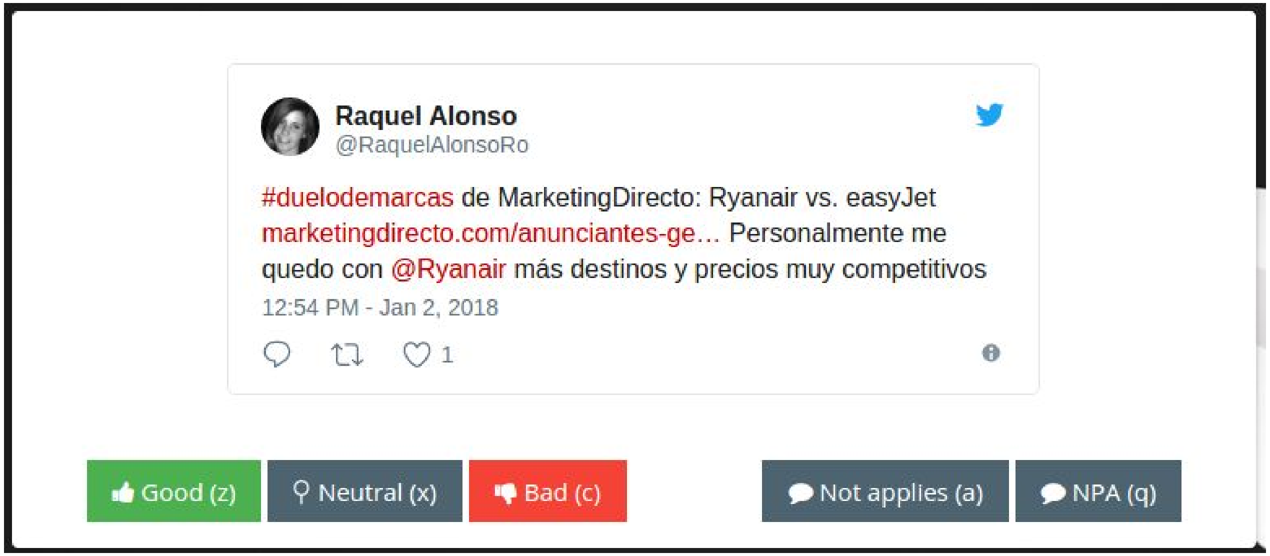
- Working with models
  - Multinomial Naive Bayes, Random forest, Vector Machines...



Verikas et al.

What’s next (II)

- Data labeling
  - Air Europa, Spanair, Ryanair, Iberia, Norwegian Airlines

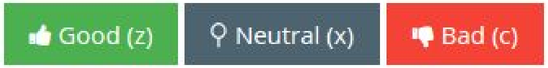


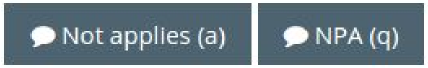
Raquel Alonso
@RaquelAlonsoRo

#duelodemarcas de MarketingDirecto: Ryanair vs. easyJet marketingdirecto.com/anunciantes-ge... Personalmente me quedo con @Ryanair más destinos y precios muy competitivos

12:54 PM - Jan 2, 2018

http://34.207.47.25:5000/

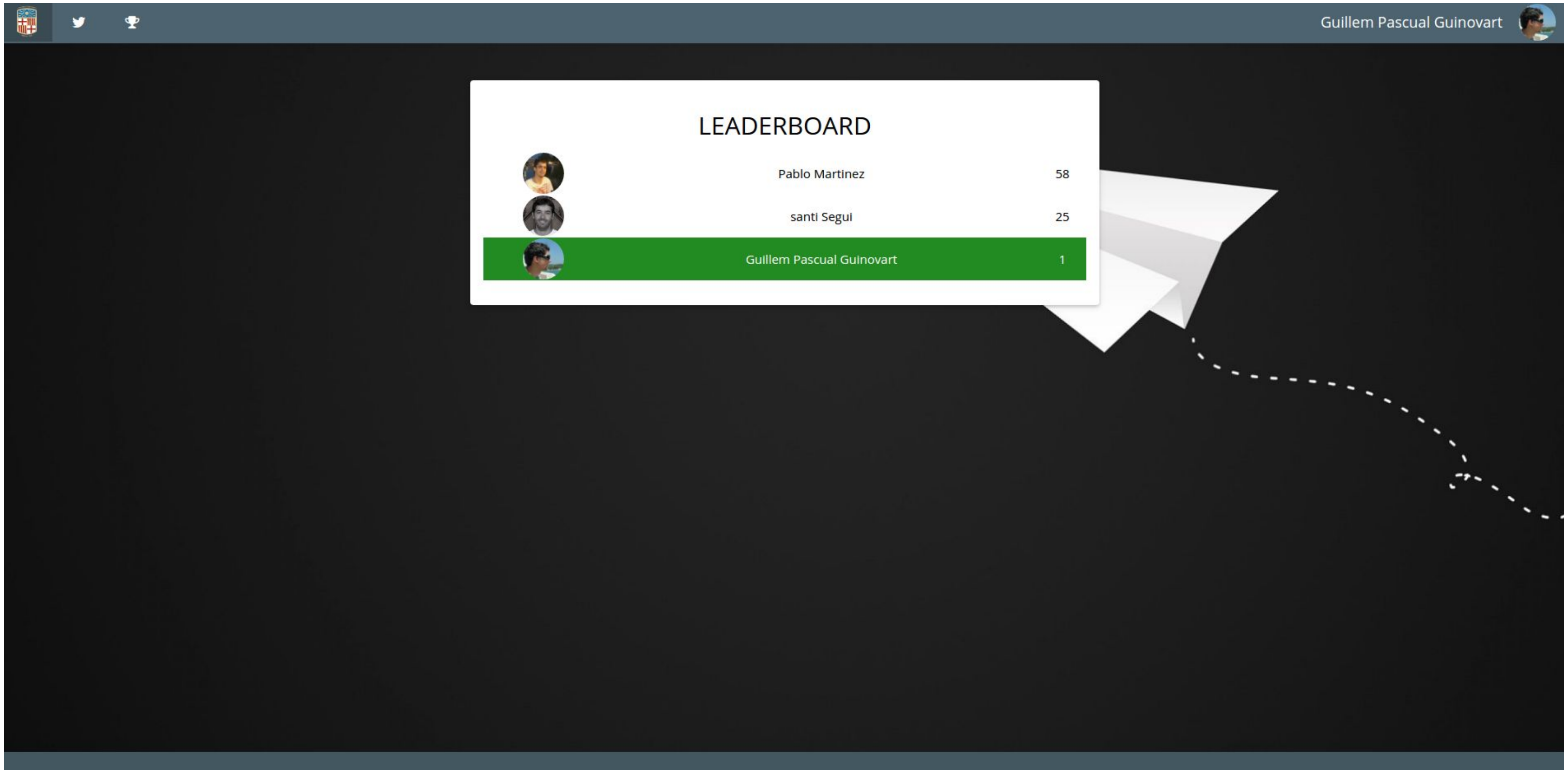
- Dataset:
  - ~20,000 Tweets
  - 1,000 labels / student
  - ~ 5 labels / minute
  - ~ 3h 30’ / student
- Options:
  - Basic: 

  - Advanced: 

***Not Applies** removes the tweet from the dataset.*

***NPA** (Not a Personal Account) Removes all tweets of the account from the dataset.*

http://34.207.47.25:5000/



LEADERBOARD

	Pablo Martinez	58
	santi Segui	25
	Guillem Pascual Guinovart	1

http://34.207.47.25:5000/

Labeling due date
1st of April

Everyone is expected to have done some labeling of the data

Next session
17th of April

Each team must at least have tested 1 model and uploaded the solution to Kaggle

Following Sessions

17th of April

24th of May

Final Delivery:

3rd of July

NOW?

Time for questions, coding and modeling!