Introducción a la plataforma Kaggle

Santiago Mota (@mota_santiago)

Kaggle

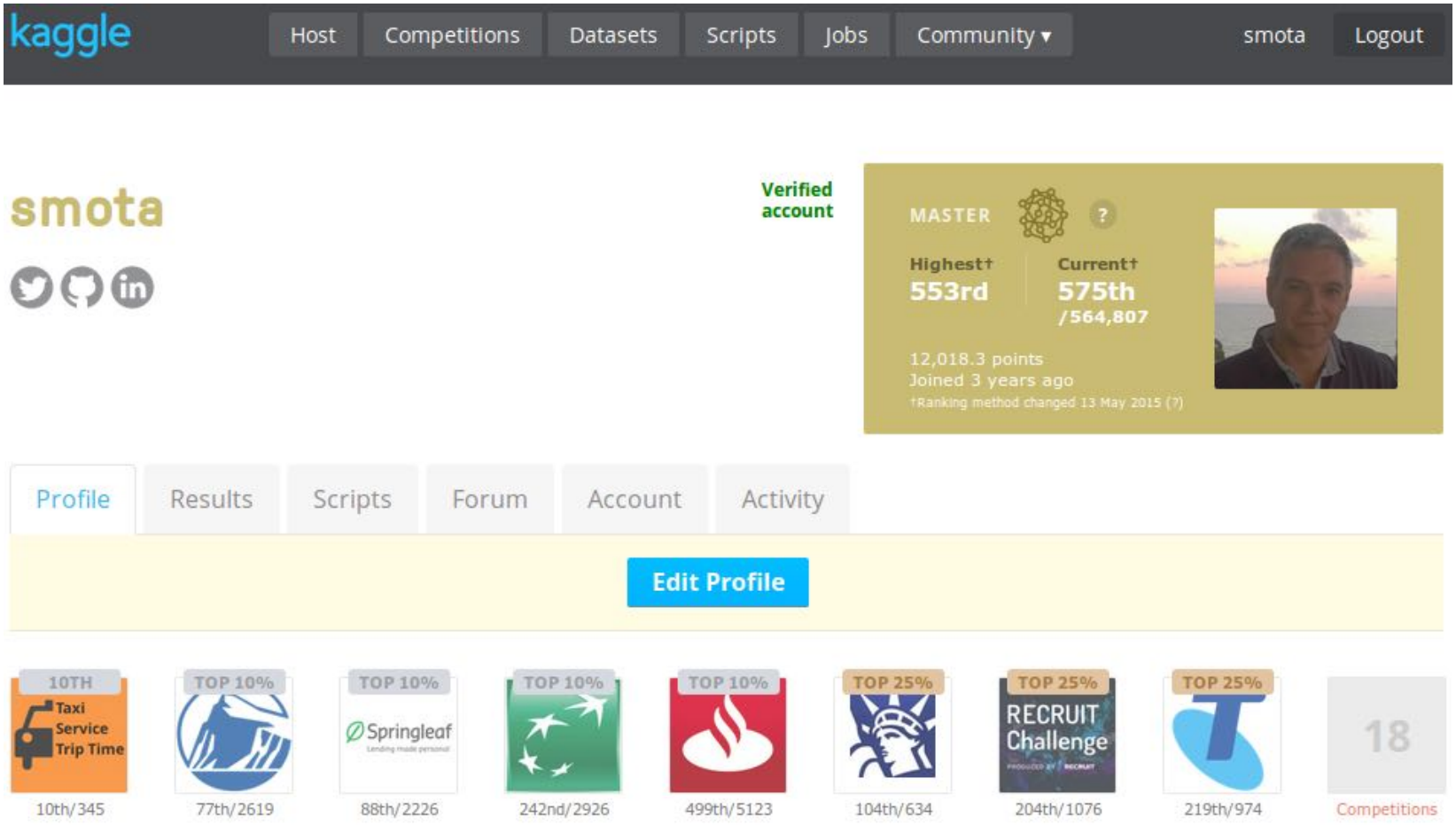
- Introducción. Jose A. Guerrero.
- Marchamo “de facto” para data science (primeros = TRABAJO).
- Mas de 500.000 usuarios en todo el mundo (creciendo).
- Zona de test para los algoritmos mas avazandos (xgboost).
- What has Kaggle learned from 2 million machine learning models?
- Lessons Learned from Running Hundreds of Kaggle Competitions

https://www.kaggle.com/

Netflix

- El concurso del millón de dólares.
- Mejorar el algoritmo de recomendación de las películas.
- Se desarrolló en distintas fases y se obligaba a publicar al final de cada fase.
- Colaboración en los foros.
- Ensamblado de soluciones.
- 1111111111 -> 1111111000 y 0001111111
- No se llegó a implementar.
- http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html
- http://www.research.att.com/articles/featured_stories/2010_01/2010_02_netflix_article.html

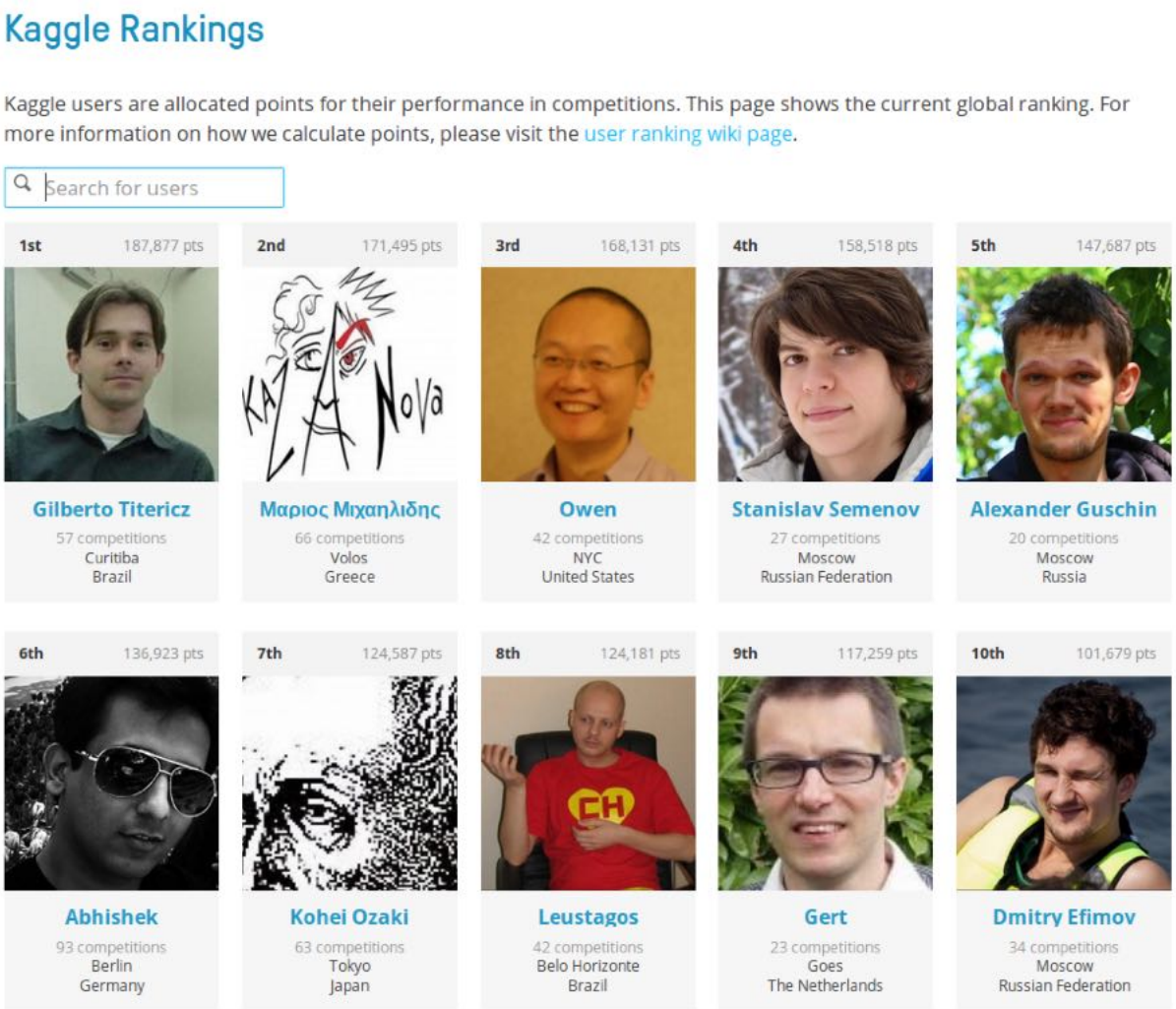
Perfil Kaggle



https://www.kaggle.com/santiagomota

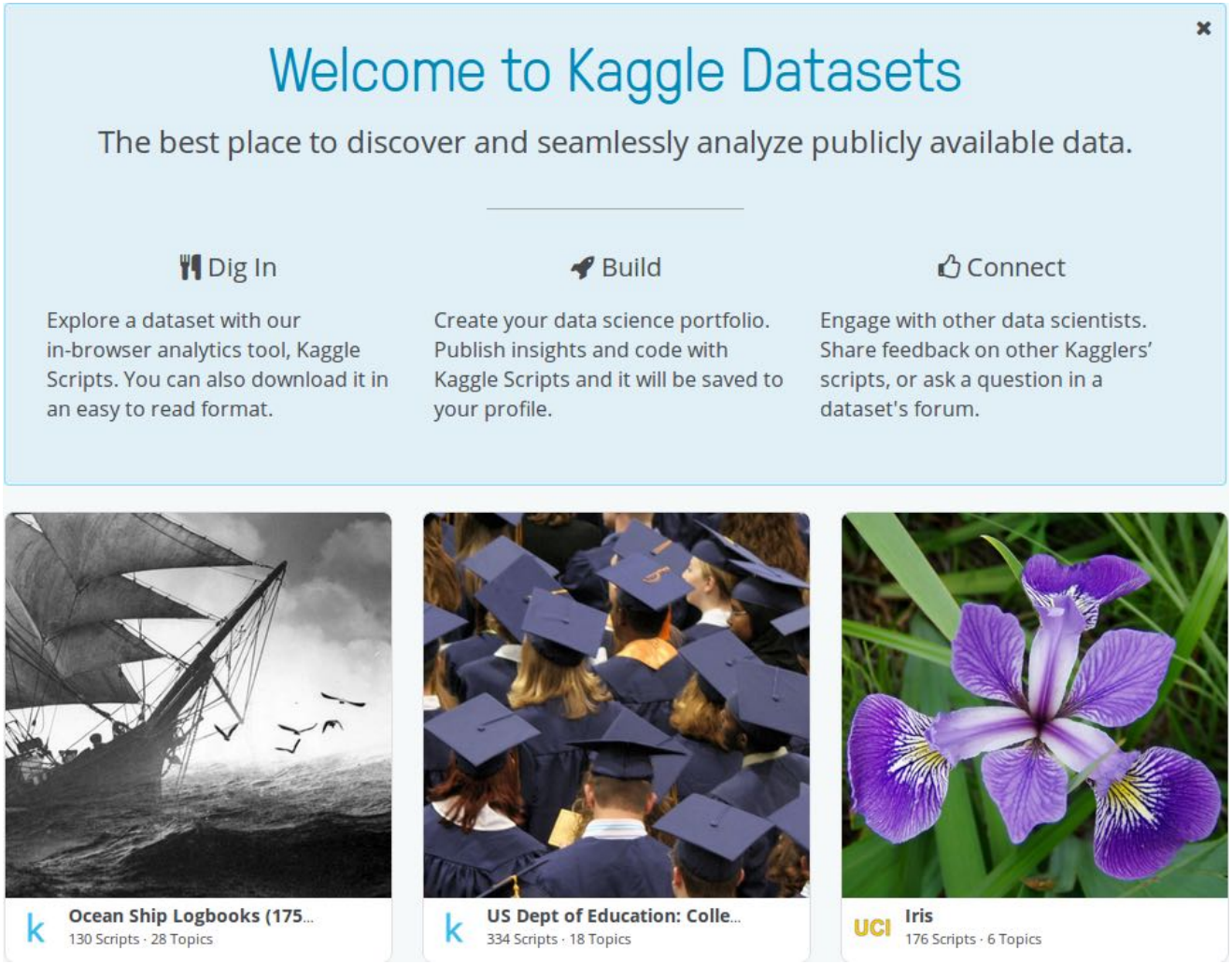
Santiago Mota (@mota_santiago)

Ranking



https://www.kaggle.com/users

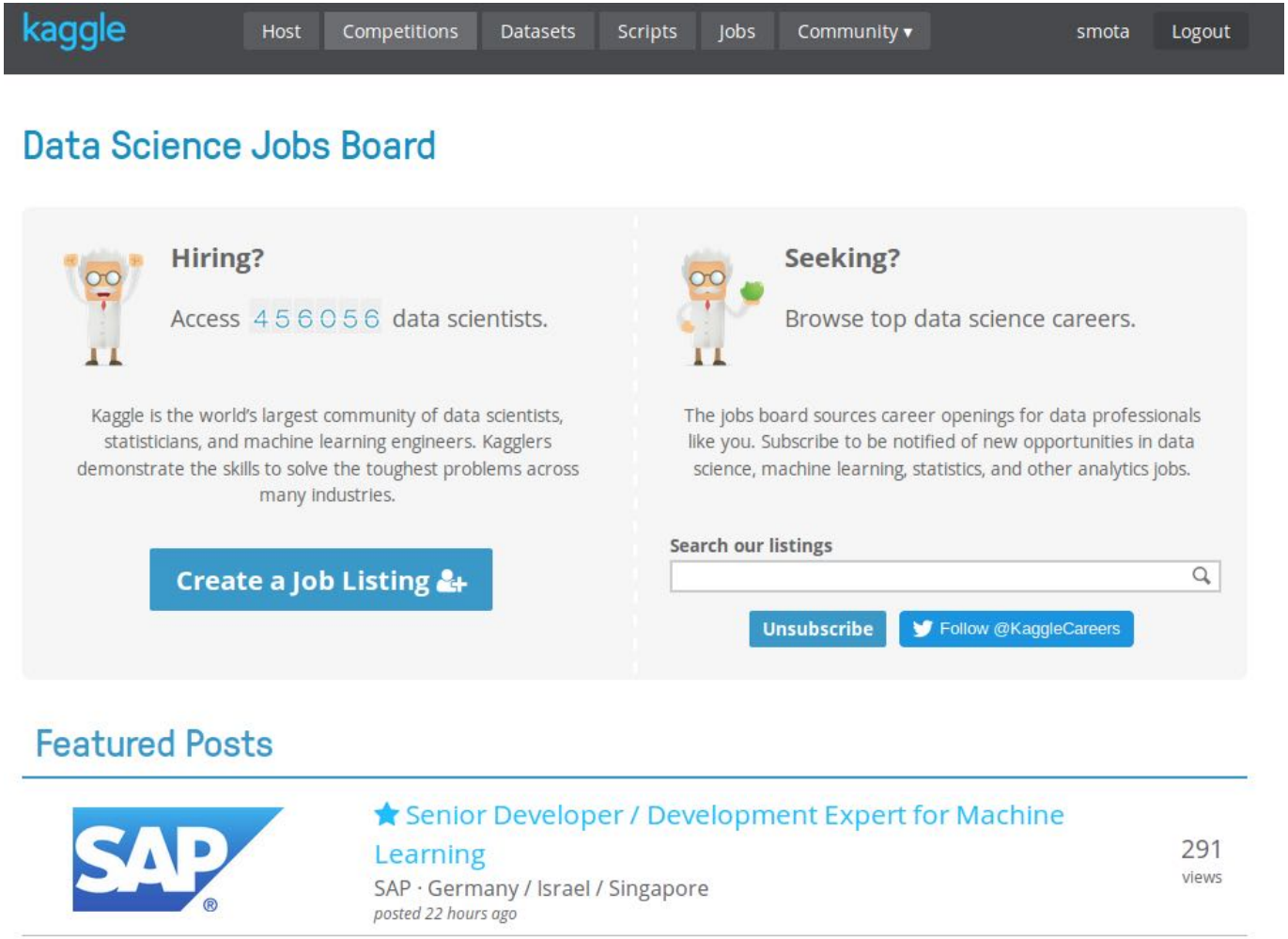
Datasets



https://www.kaggle.com/datasets

Santiago Mota (@mota_santiago)

Puestos de trabajo



https://www.kaggle.com/jobs

Ranking. Points. Tiers.

- Tres niveles (va a cambiar): Novice, Kaggle, Master.
- Los puntos “decaen”.
- Se tienen en cuenta los votos, número de participantes, si se forma parte de un equipo.
- Brainstorming Cambios.

https://www.kaggle.com/wiki/UserRankingAndTierSystem

Puntos

Points

Kaggle users are allocated points for their performance in competitions. The overall user rankings are shown at https://www.kaggle.com/users. These are recalculated at the end of every competition, once results have been finalized.

More points are earned for better results, with the maximum achievable points based on the number of total participants in the competitions, and a multiplier on the competition known as the "User Rank Multiplier". For certain competitions (e.g. Getting Started, or competitions with a public ground truth) the user rank multiplier of a competition is set to zero, meaning the competition will have no impact on users' points.

The current formula for each competition divides the points among the team members according to the square root, decays the points for lower finishes, adjusts for the number of teams that entered the competition, and decays the points as time elapses from the competition end. For each competition, the formula is:

$$\left[\frac{100000}{\sqrt{N_{\text{teammates}}}}\right]\left[\text{Rank}^{-0.75}\right]\left[\log_{10}\left(1+\log_{10}\left(N_{\text{teams}}\right)\right)\right]\left[e^{-t/500}\right]$$

Points are always calculated with time decay fixed at the time of the most recent competition deadline. Between competition deadlines points do not decay and ranks will not change.

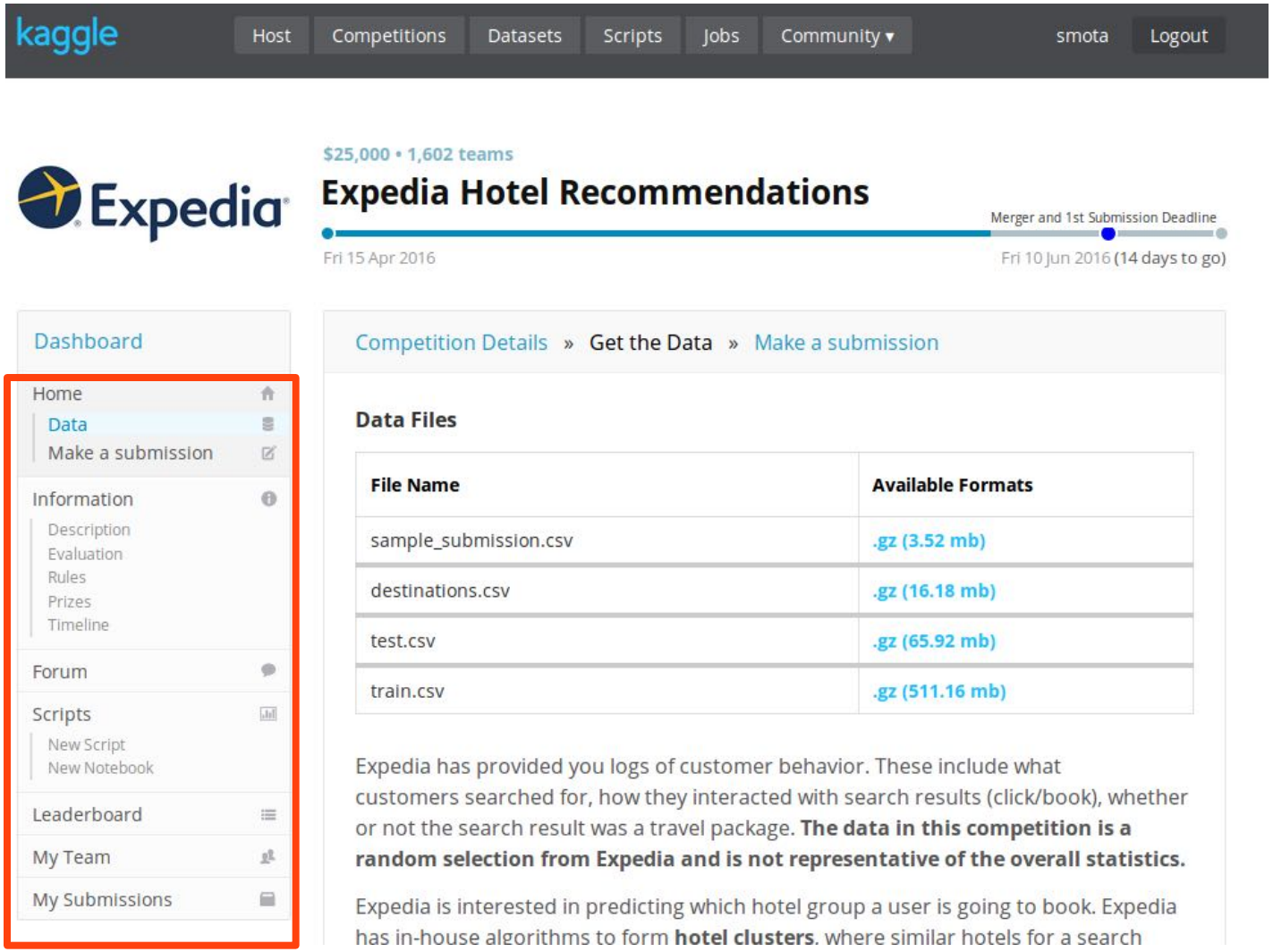
https://www.kaggle.com/wiki/UserRankingAndTierSystem

Masters

- Quedar entre los 10 primeros en una competición.
- Quedar en el primer 10% en otra.
- Da acceso a concursos especiales.
- Unos 1000.

https://www.kaggle.com/wiki/UserRankingAndTierSystem

Datos



https://www.kaggle.com/c/expedia-hotel-recommendations/data

Public/private leaderboard I

SalePrice	SquareFeet	Type	LotAcres	Beds	Baths
$88k	719	HOME	1.64	1	1
$164k	2017	APT		3	2
$72k	697	APT		1	1
$85k	948	HOME	1.02	2	3
$271k	3375	APT		3	4
$482k	3968	APT		4	4
$88k	790	APT		1	2
$128k	1341	HOME	0.66	3	3
$235k	2379	APT		3	3
$309k	2495	HOME	0.21	3	4
$163k	1356	APT		1	1
$375k	3361	HOME	1.64	3	4
$98k	1060	HOME	0.05	1	1
???	582	HOME	0.61	1	1
???	1640	APT		2	3
???	3546	HOME	0.4	4	4
???	903	APT		2	2
???	1096	HOME	0.04	3	4
???	1280	HOME	0.15	2	2
???	1139	APT		1	1

Predicted
$41k
$165k
$280k
$76k
$128k
$115k
$94k

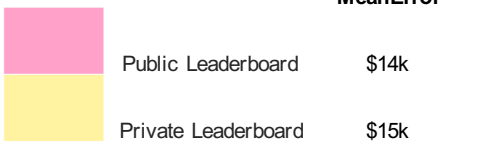
Training

Test

Submission

http://www.slideshare.net/b…eds-of-kaggle-competitions

Public/private leaderboard II



MeanError	
Public Leaderboard	$14k
Private Leaderboard	$15k

Delta	Predicted
-$9k	$41k
$20k	$165k
-$14k	$380k
-$6k	$76k
$13k	$128k
-$14k	$115k
-$12k	$94k

Submission

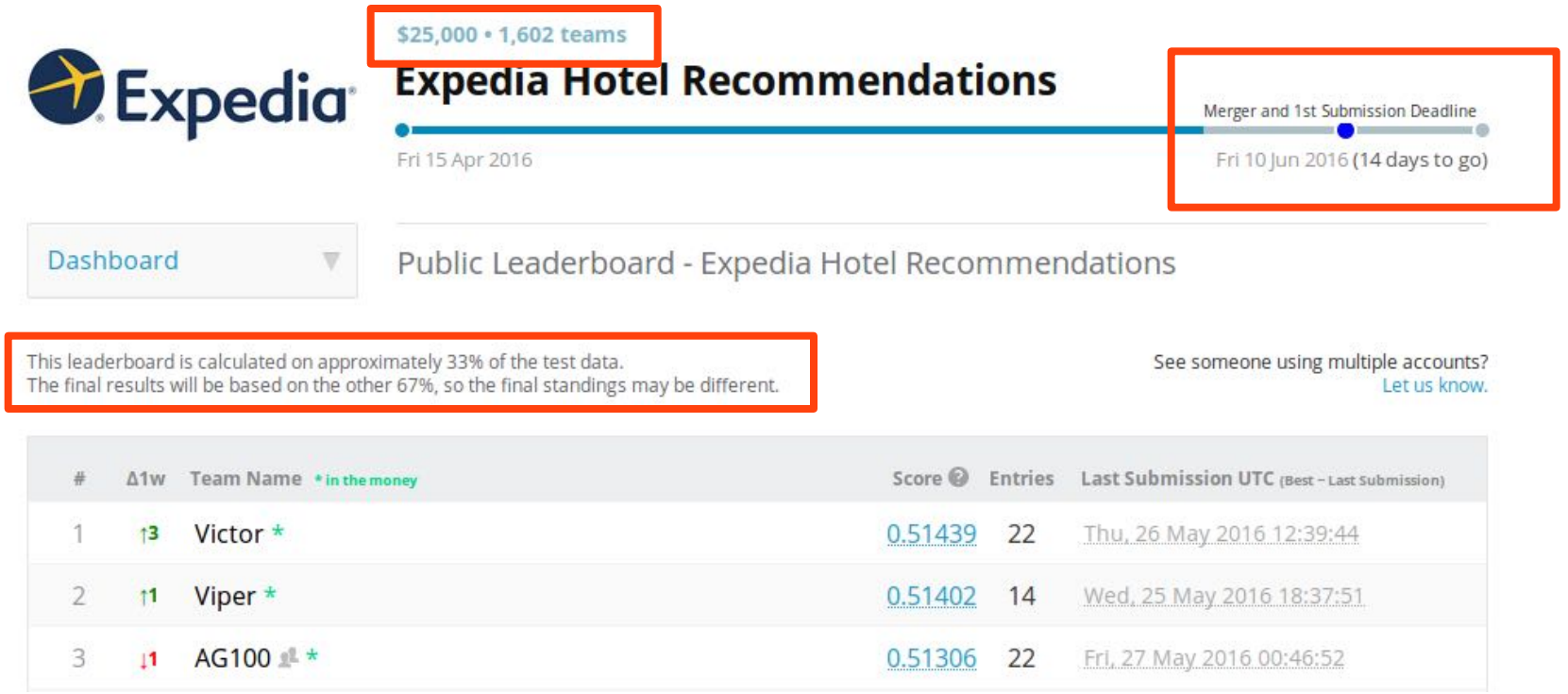
SalePrice	SquareFeet	Type	LotAcres	Beds	Baths
$88k	719	HOME	1.64	1	1
$164k	2017	APT		3	2
$72k	697	APT		1	1
$85k	948	HOME	1.02	2	3
$271k	3375	APT		3	4
$482k	3968	APT		4	4
$88k	790	APT		1	2
$128k	1341	HOME	0.66	3	3
$235k	2379	APT		3	3
$309k	2495	HOME	0.21	3	4
$163k	1356	APT		1	1
$375k	3361	HOME	1.64	3	4
$98k	1060	HOME	0.05	1	1
$50k	582	HOME	0.61	1	1
$145k	1640	APT		2	3
$394k	3546	HOME	0.4	4	4
$82k	903	APT		2	2
$105k	1096	HOME	0.04	3	4
$129k	1280	HOME	0.15	2	2
$106k	1139	APT		1	1

Training

Test

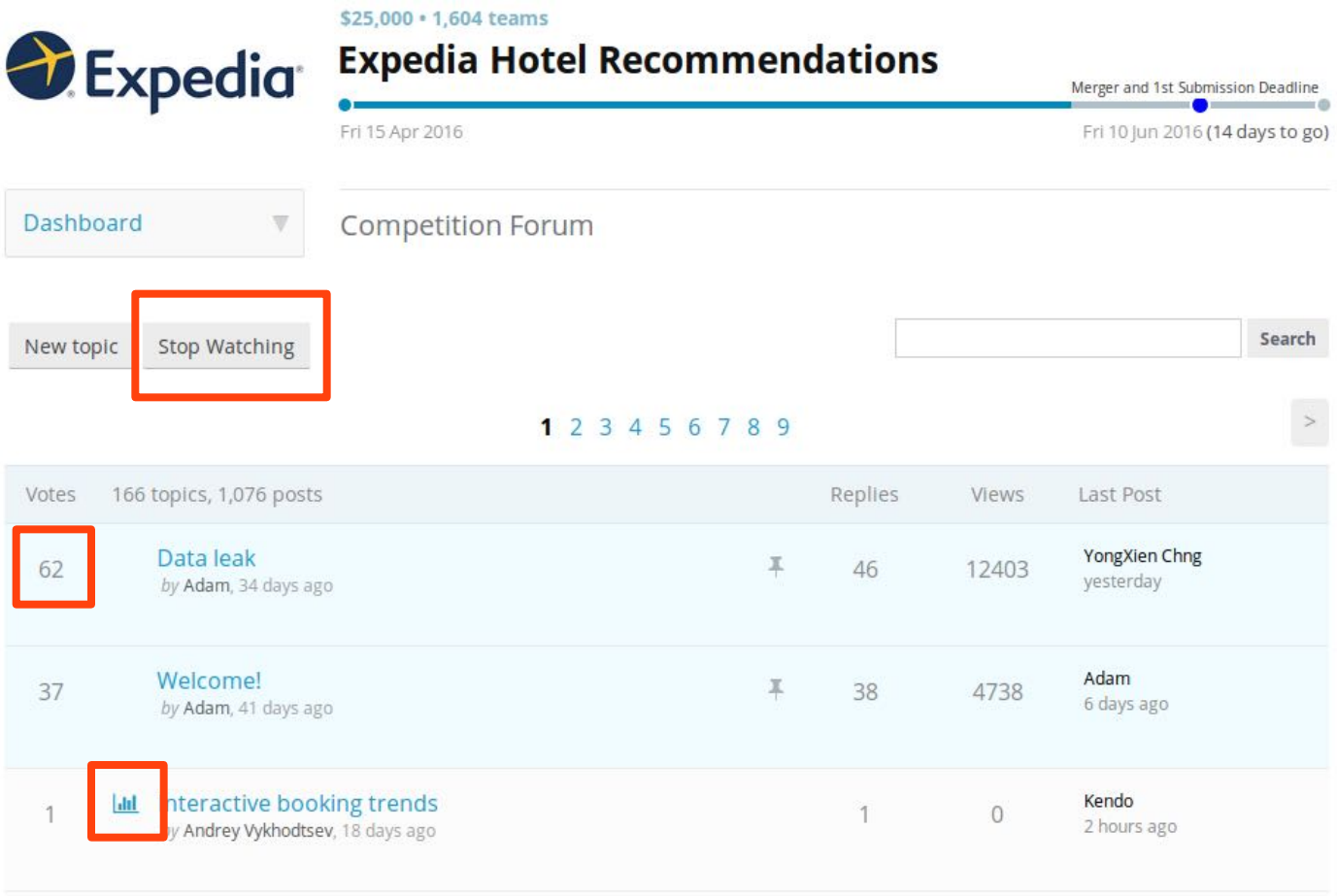
http://www.slideshare.net/b ds-of-kaggle-competitions

Leaderboard



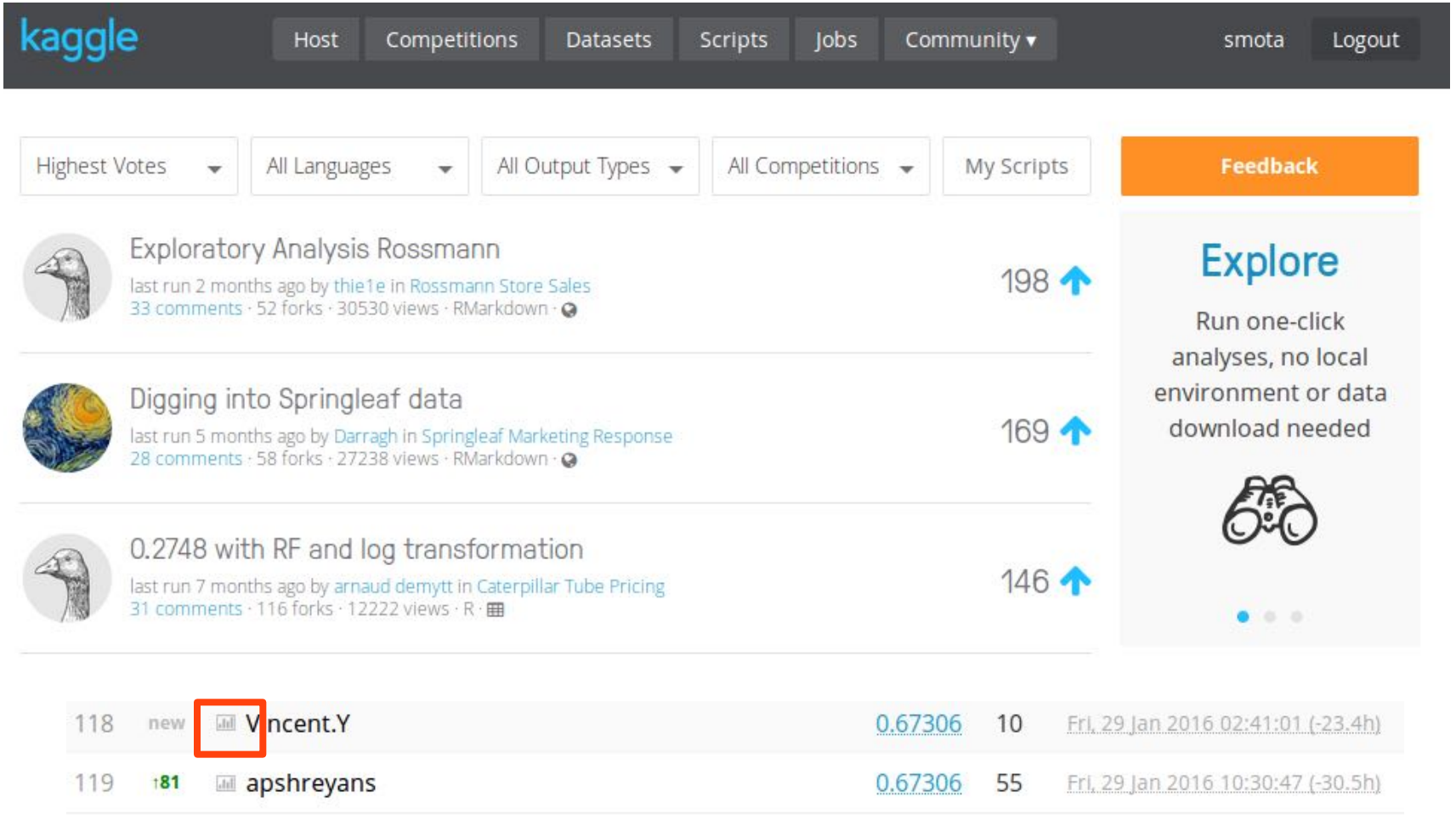
https://www.kaggle.com/c/expedia-hotel-recommendations/leaderboard

Forum



https://www.kaggle.com/c/expedia-hotel-recommendations/forums

Scripts



https://www.kaggle.com/scripts?sortBy=votes

Al empezar el concurso

- Tipo de concurso. ¿Alguno anterior?
- Cantidad de datos.
- Métrica de evaluación (library(Metrics)) y (General).
- Fechas límite.
- Partición public/private leaderboard.
- Subscribirse al foro.
- Buscar en Github.
- Leer las condiciones.
- Reproducción de la solución final.

https://www.kaggle.com/wiki/WinningModelDocumentationTemplate

Estrategias

- Formación de equipos (límites).
- Scripts.
- Número de submissions al día.
- ¿Sobre cuantos modelos se hará la evaluación final?
- Elección de los modelos para el private leaderboard.
- Gestión de tiempos (dedicación).
- ¿Me fio del public leaderboard (overfitting)?
- Foro durante el concurso y al finalizar (huevos de pascua).

https://www.kaggle.com/wiki/UserRankingAndTierSystem

Varios

- Data leakege.
- Cuentas anónimas / imagen.
- Preguntar en el foro (puntos).
- Confirmación por SMS.
- Seed (xgboost).
- Titanic.
- Digit recognizer.
- 50% python, 40% R, 10% otros.
- Metodología de trabajo (inversion)

¿Cuanto cuesta?

- En Kaggle el coste (premios incluidos) es de unos 100.000$.
- Se incluye la preparación, seguimiento y análisis.
- Opciones gratuitas:
- Como profesor, en ese caso se limita y los alumnos y no se da asistencia (Kaggle Inclass).
- Con un proyecto que les parezca interesante.

Otras plataformas

- CrowdAnalytics
- DrivenData
- Devpost
- Innocentive
- TunedIT
- Enlaces a competiciones en Kdnuggets

Concursos presenciales

- Fin de semana vs extensos en el tiempo.
- Dotación económica.
- En equipo (casi siempre).
- Uso de otras “soft-skills”.
- Mas valor de la idea/presentación vs. datos/algoritmo.
- Limitaciones: Tiempo, datos, presentación.

Gracias

Datos de contacto:

Santiago Mota Herce

Teléfono: 670702852
Twitter: @mota_santiago
E-mail: santiago_mota@yahoo.es
LinkedIn: https://es.linkedin.com/in/santiagomota