Machine Learning Automatizado

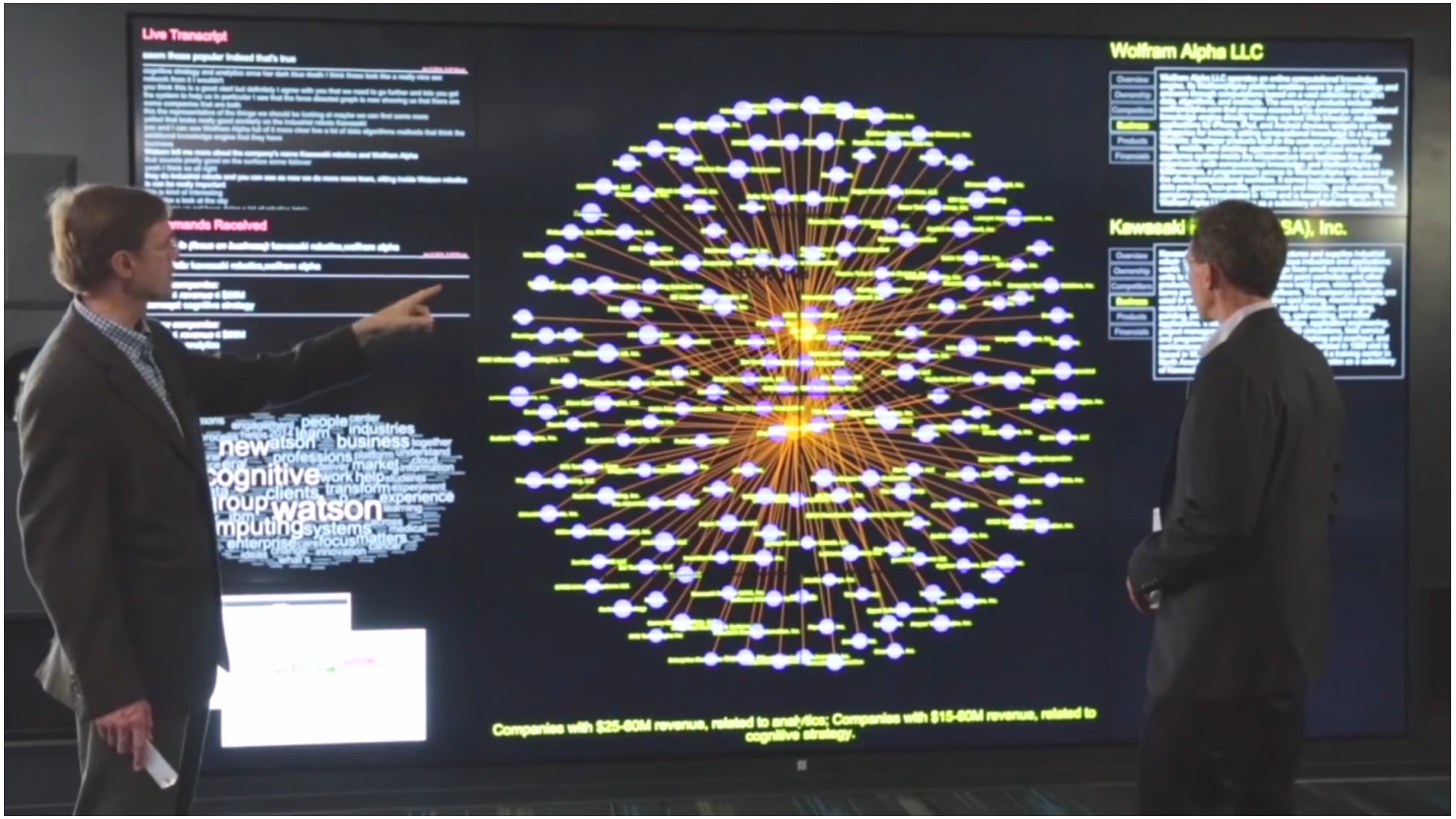
cc BY NC SA

Presentación

- Físico. MBA y Máster BI y BigData
- Consultor Freelance
- Kaggle Máster
- Socio Asociación de Usuarios de R de España
- santiago_mota@yahoo.es
- http://es.linkedin.com/in/santiagomota

CC BY NC SA

Dario Gil: Cognitive systems and the future of expertise TED (22/12/2014)



https://www.youtube.com/watch?v=0heqP8d6vtQ

CC BY NC SA

IBM anuncia Watson Analytics, un servicio de analítica cognitiva de negocio (19/09/2014)

- *"IBM Watson Analytics automatiza, además, algunos pasos del análisis, como la preparación de los datos, el análisis predictivo y la visualización… ".*
- *"Diálogo en lenguaje natural: el nuevo producto entiende el lenguaje natural, por lo que solo es necesario teclear lo que al usuario le gustaría ver...".*
- *"Analítica predictiva "guiada": el servicio es capaz de guiar al usuario en patrones y resultados de los datos en los que el usuario tradicionalmente no se fijaría".*

http://www.eleconomista.es/tecnologia/noticias/6085269/09/14/IBM-anuncia-Watson-Analytics-un-servicio-de-analitica-cognitiva-de-negocio.html

CC BY NC SA

Gartner. Data Science y ML Platforms

https://www.gartner.com/doc/reprints?id=1-4RQ3VEZ&ct=180223&st=sb

CC BY NC SA

Tools that Data Scientists actually use

https://thomaswdinsmore.com/2018/02/26/notes-on-gartners-2018-data-science-and-machine-learning-mq/?lipi=urn%3Ali%3Apage%3Ad_flagship3_feed%3BvtLwdqeRTBCwlam84Qf%2BOw%3D%3D

CC BY NC SA

Proyecto de datos

id	superficie_sq_ft	tipo	parcela_acres	habitaciones	banos	precio_venta
1	719	Casa	1,64	1	1	88.000
2	2.017	Apartamento		3	2	164.000
3	697	Apartamento		1	1	72.000
4	948	Casa	1,02	2	3	85.000
5	3.375	Apartamento		3	4	271.000
6	3.968	Apartamento		4	4	482.000
7	790	Apartamento		1	2	88.000
8	1.341	Casa	0,66	3	3	128.000
9	2.379	Apartamento		3	3	235.000
10	2.495	Casa	0,21	3	4	309.000
11	1.356	Apartamento		1	1	163.000
12	3.361	Casa	1,64	3	4	375.000
13	1.060	Casa	0,05	1	1	98.000
14	582	Casa	0,61	1	1	50.000
15	1.640	Apartamento		2	3	145.000
16	3.546	Casa	0,40	4	4	394.000
17	903	Apartamento		2	2	82.000
18	1.096	Casa	0,40	3	4	105.000
19	1.280	Casa	0,15	2	2	129.000
20	1.139	Apartamento		1	1	106.000

cc BY NC SA

Proyecto de datos

id	superficie_sq_ft	tipo	parcela_acres	habitaciones	banos	precio_venta		
1	719	Casa	1,64	1	1	88.000		
2	2.017	Apartamento		3	2	164.000		
3	697	Apartamento		1	1	72.000		
4	948	Casa	1,02	2	3	85.000		
5	3.375	Apartamento		3	4	271.000		
6	3.968	Apartamento		4	4	482.000		
7	790	Apartamento		1	2	88.000		
8	1.341	Casa	0,66	3	3	128.000		
9	2.379	Apartamento		3	3	235.000		
10	2.495	Casa	0,21	3	4	309.000		
11	1.356	Apartamento		1	1	163.000		
12	3.361	Casa	1,64	3	4	375.000		
13	1.060	Casa	0,05	1	1	98.000	Prediccion	Error
14	582	Casa	0,61	1	1	50.000	41.000	-9.000
15	1.640	Apartamento		2	3	145.000	165.000	20.000
16	3.546	Casa	0,40	4	4	394.000	380.000	-14.000
17	903	Apartamento		2	2	82.000	76.000	-6.000
18	1.096	Casa	0,40	3	4	105.000	128.000	23.000
19	1.280	Casa	0,15	2	2	129.000	115.000	-14.000
20	1.139	Apartamento		1	1	106.000	94.000	-12.000

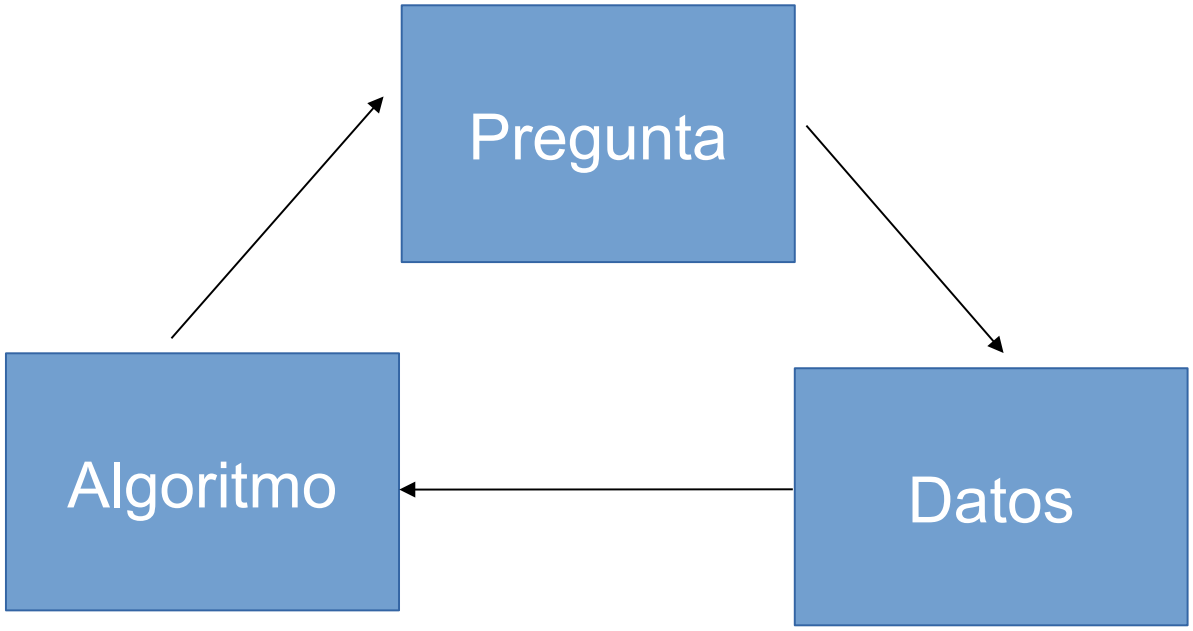
CC BY NC SA

Proyecto de datos

- Hay casas con mas baños que habitaciones
- División Casa / Apartamento
- Elementos vacíos, outliers (ELT)
- Nuevas columnas (feature engineering)
- Cross Validation
- Nuevos algoritmos
- De donde vienen los datos y, sobre todo: **Cuenta de resultados**

CC BY NC SA

Proyecto de datos



CC BY NC SA

Machine Learning automatizado

- Trifacta Wrangler (link)
- IBM Watson (link)
- Datarobot (link)
- Daitaku (link)
- Domino (link)
- Seldon (link)
- Alterix (link)
- H2O

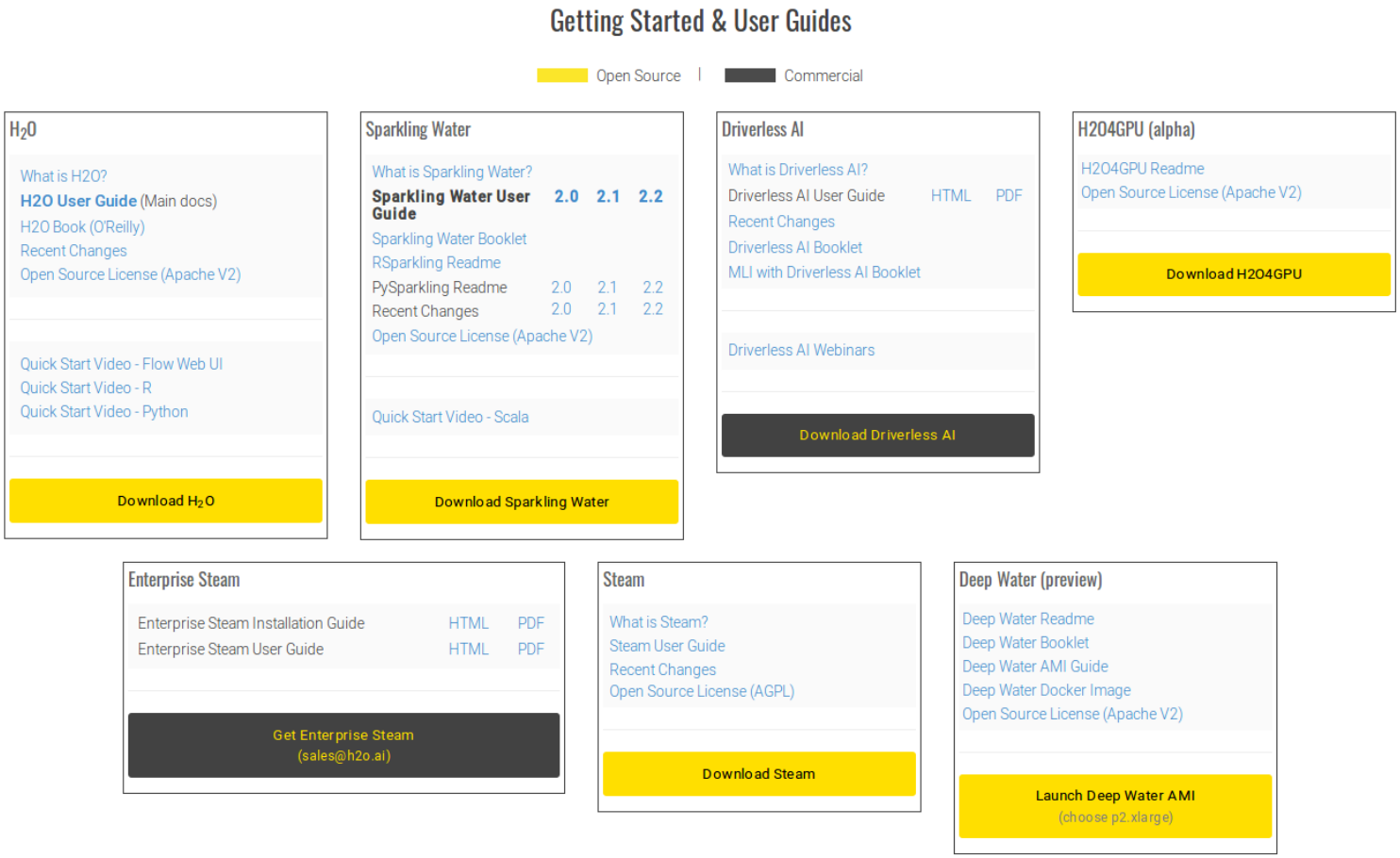
CC BY NC SA

H2O

- Creada en 2011 (inicialmente 0xdata)
- Noviembre 2015: +$20M (B) ya tenían $14M
- Noviembre 2017: +$40M (C) Total **$75M**. (Nvidia, Wells Fargo)
- Personas
- Kaggle

CC BY NC SA

Oferta de H2O



http://docs.h2o.ai/?_ga=2.107667714.1485748875.1520325919-538902739.1512117166

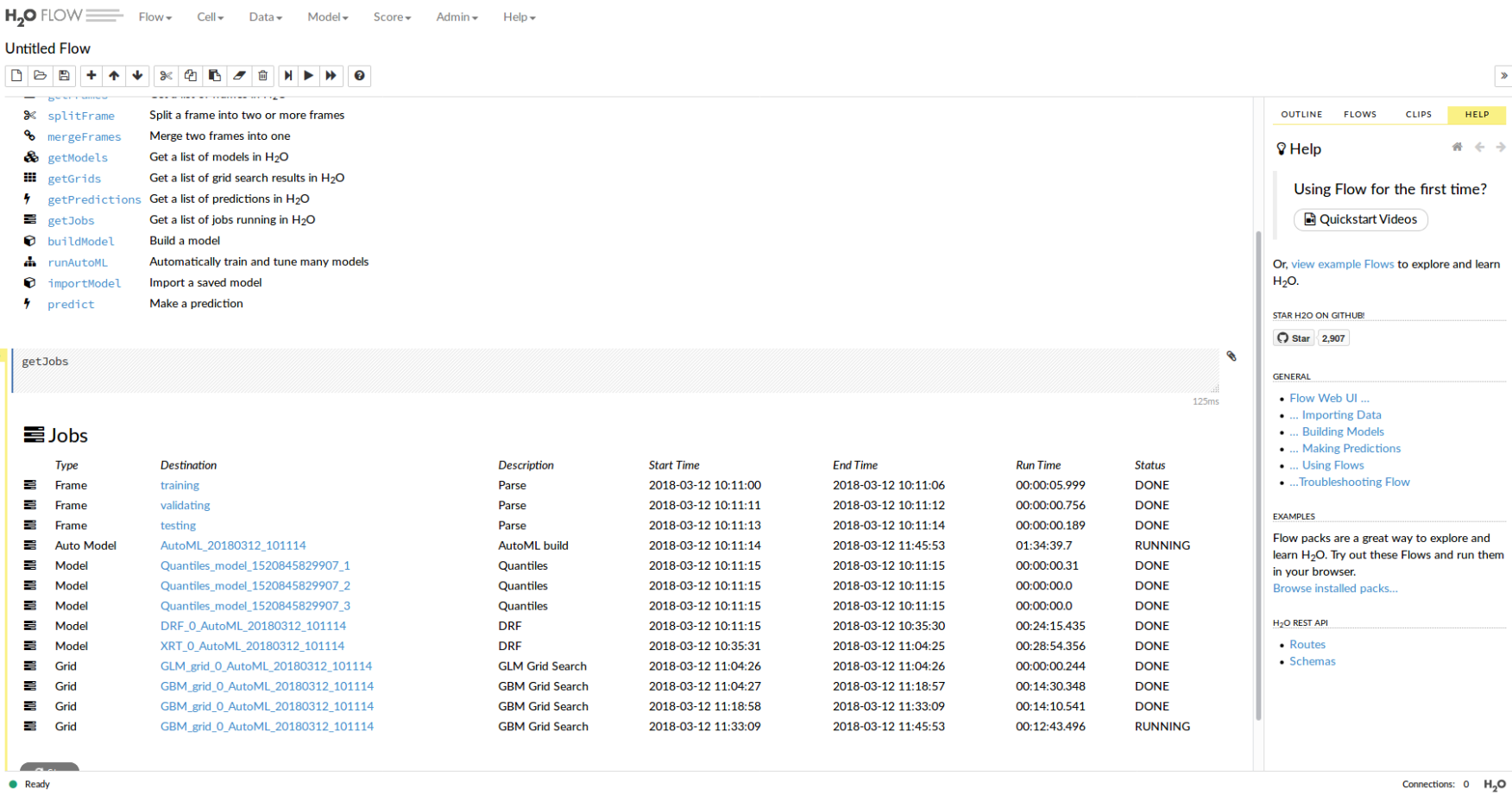
CC BY NC SA

H2O

- Basada en java
- Facilidades para escalar
- Paralización. Para R, substituto data.table
- Maquina local, cluster o en cloud
- Funciona como API, pero tiene navegador
- Acceso desde R o Python
- Pagina (link), blog (link) y para iniciarse (link y link)

CC BY NC SA

Localhost H2O



CC BY NC SA

Instalación desde R

DOWNLOAD AND RUN | **INSTALL IN R** | INSTALL IN PYTHON | INSTALL ON HADOOP | USE FROM MAVEN

Use H_2O directly from R

Copy and paste these commands into R one line at a time:

```
# The following two commands remove any previously installed H2O packages for R.
if ("package:h2o" %in% search()) { detach("package:h2o", unload=TRUE) }
if ("h2o" %in% rownames(installed.packages())) { remove.packages("h2o") }

# Next, we download packages that H2O depends on.
pkgs <- c("RCurl","jsonlite")
for (pkg in pkgs) {
if (! (pkg %in% rownames(installed.packages()))) { install.packages(pkg) }
}

# Now we download, install and initialize the H2O package for R.
install.packages("h2o", type="source", repos="http://h2o-release.s3.amazonaws.com/h2o/rel-wolpert
/4/R")

# Finally, let's load H2O and start up an H2O cluster
library(h2o)
h2o.init()
```

http://h2o-release.s3.amazonaws.com/h2o/rel-wolpert/4/index.html

CC BY NC SA

Conectar dos servidores

CLOUD STATUS

HEALTHY CONSENSUS LOCKED

Version	Started	Nodes (Used / All)
3.18.0.2	3 minutes ago	2 / 2

NODES

	Name	Ping	Cores	Load	My CPU %	Sys CPU %	GFLOPS	Memory Bandwidth	Data (Used/Total)	Data (% Cached)	GC (Free / Total / Max)
✓	192.168.1.68:55555	a few seconds ago	16	0.032	-1	-1	13.799	11.54 GB / s	- / NaN undefined	NaN%	12.84 GB / NaN undefined / 13.33 GB
✓	192.168.1.148:55555	a few seconds ago	4	0.510	-1	-1	12.229	17.03 GB / s	- / NaN undefined	NaN%	6.95 GB / NaN undefined / 6.97 GB
✓	TOTAL	-	20	0.542	-	-	26.028	28.57 GB / s	- / NaN undefined	NaN%	19.78 GB / NaN undefined / 20.30 GB

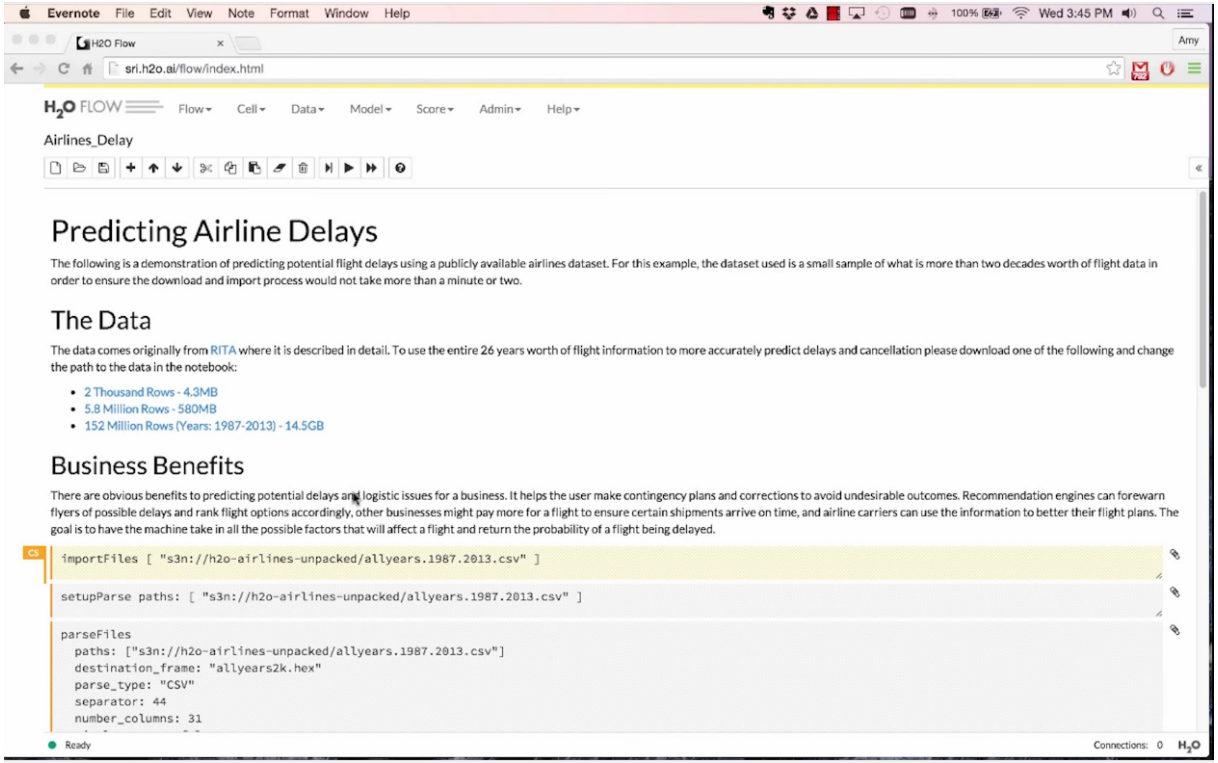
Refresh

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/faq/tunneling.html

CC BY NC SA

Analizar retrasos en vuelos con H2O

- Video (link), en flow (link), en R (link) y datos (link)



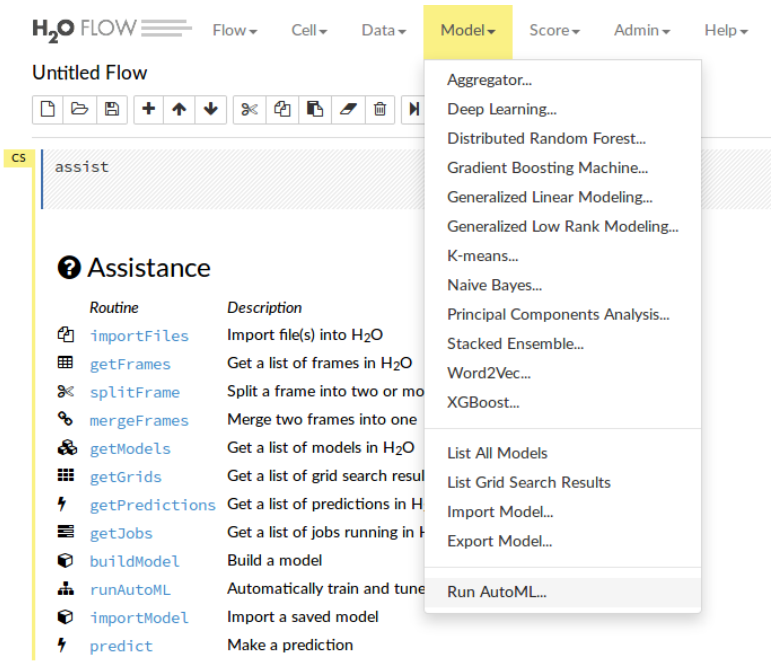
http://university.h2o.ai/data-science-101/lesson2.html

CC BY NC SA

AutoML

- Sólo hay que darle el dataset, target y tiempo
- Básicamente hace un stacking de modelos

```
tiempo_inicio <- Sys.time()
automl_models_h2o <- h2o.automl(
    x                 = x,
    y                 = y,
    training_frame    = training_h2o,
    validation_frame  = validating_h2o,
    # leaderboard_frame = test_h2o,
    max_runtime_secs  = 6000, # 180
    stopping_metric   = "AUTO")
print(Sys.time()-tiempo_inicio)
```



CC BY NC SA

Demo. Crímenes en L.A.

- Basada en estos posts (link1 y link2)
- Con datos de opendata de Los Ángeles (link). Hay que bajarlos
- Los datos necesitan de tratamiento previo
- Necesitaría mas ETL y mas feature engineering
- Página de github (link)

CC BY NC SA

DriverlessAI

- Licencia
- Coste (precio anual + equipos)
- Docker
- Vídeo

CC BY NC SA

DriverlessAI. Requerimientos

- 64G de RAM
- GPU con CUDA (Pascal o Volta)
- Docker (o Nvidia docker)
- Cloud, Server, Desktop
- Linux, Mac, Windows 10

http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/installing.html

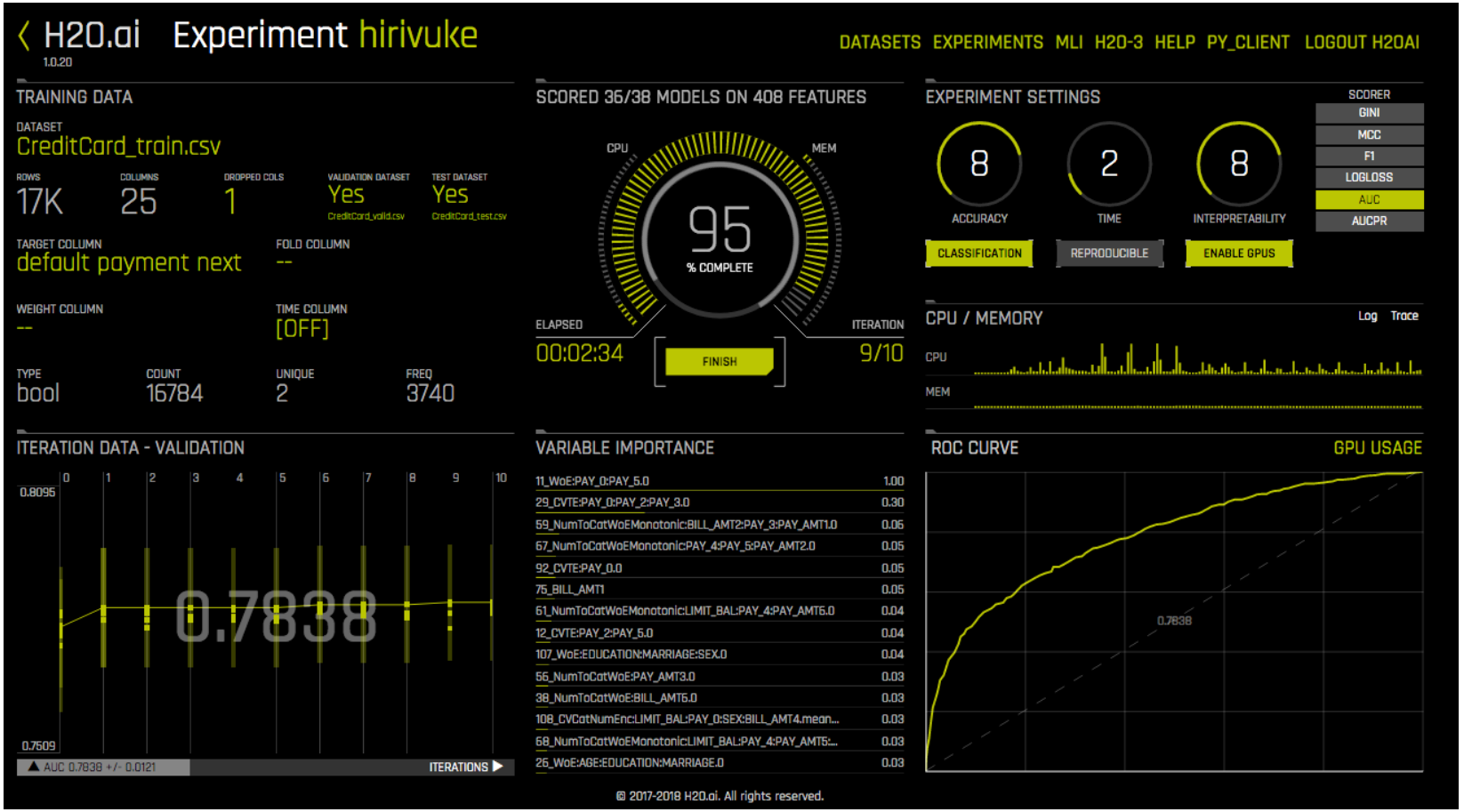
CC BY NC SA

DriverlessAI

http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/launching.html

CC BY NC SA

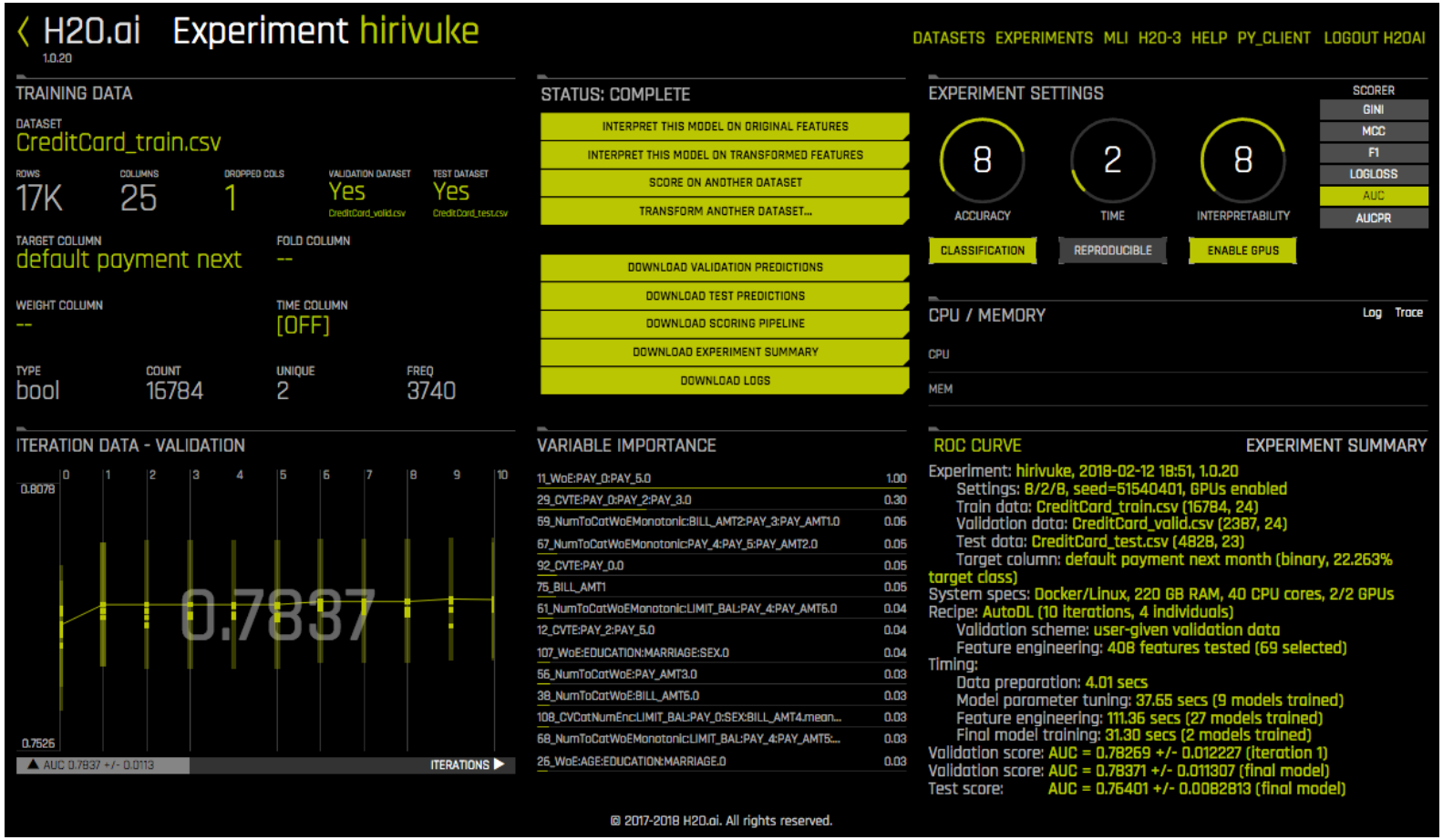
DriverlessAI



http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/launching.html

CC BY NC SA

DriverlessAI



https://www.youtube.com/watch?time_continue=43&v=KkvWX3FD7y

CC BY NC SA

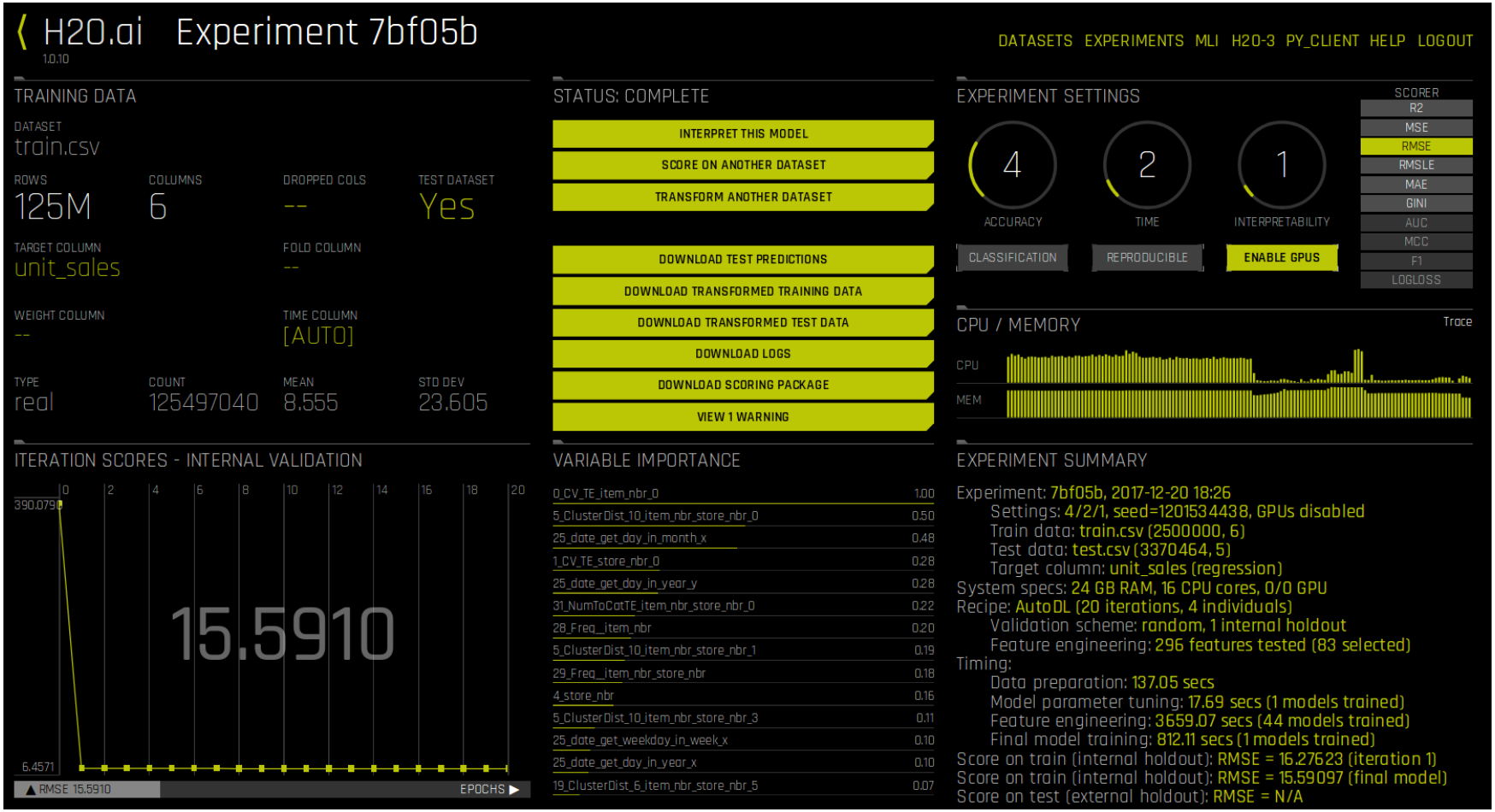
DriverlessAI. Prueba en Kaggle Favorita

- Concurso Kaggle Favorita (link)
- Estación de trabajo Z800. 16 cores. 24G RAM sin GPU
- Posición final: 126 de 1675 (medalla de bronce)
- Métrica: Normalized Weighted Root Mean Squared Logarithmic Error
- Mis mejores resultados: 0,520 (combinado) y 0,521 con un modelo LGBM.
- Resultado del ganador: 0,509
- Mejor resultado DriverlessAI: 1,240 (posición 1.131)

https://www.kaggle.com/c/favorita-grocery-sales-forecasting/leaderboard

CC BY NC SA

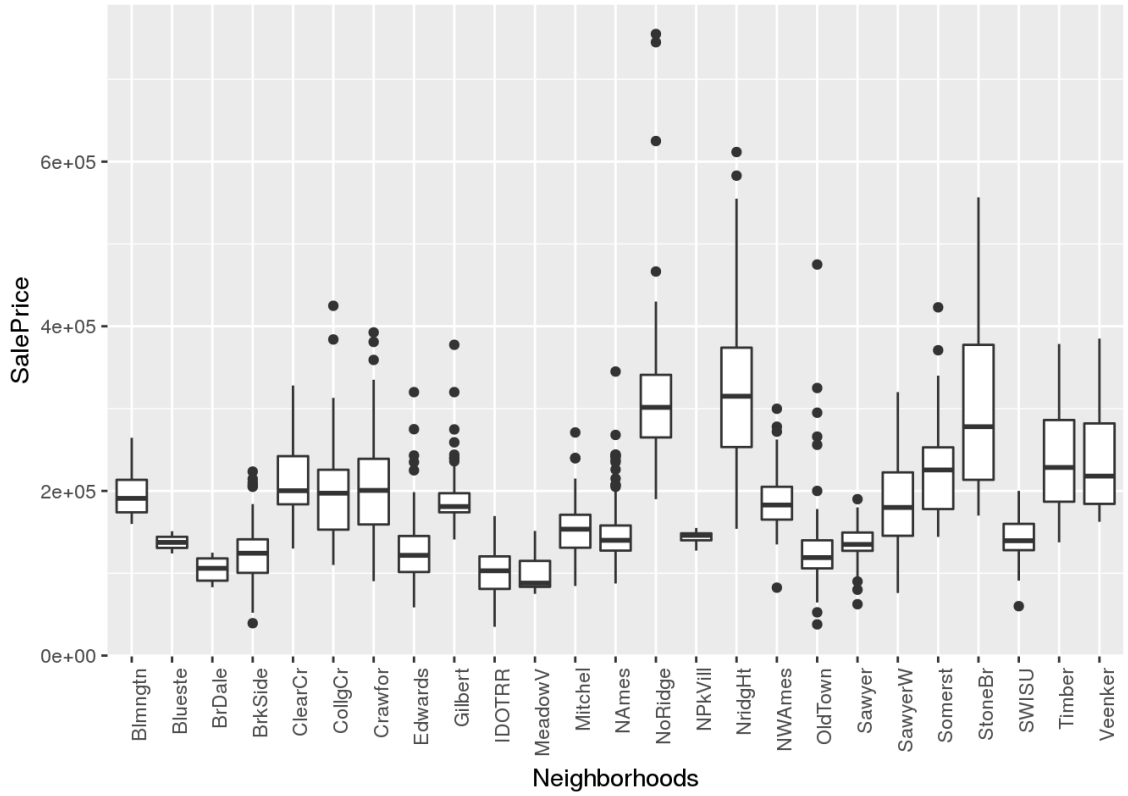
Una solución: Pred028. 1,264. 1:20:00



CC BY NC SA

Concurso precios de alquileres (Kaggle)

- Página del concurso. Tutoriales. Kernels y Foro



CC BY NC SA

Conclusiones

- En muy poco tiempo (¿este año?) vamos a tener herramientas comerciales de Machine Learning Automatizado como DriverlessAI
- Inicialmente su uso tendrá sentido en determinados escenarios
- La herramienta H2O (gratuita) tiene mucho sentido, por su capacidad de escalar, sus distintas interfaces y sus posibilidades de paralelización
- AutoML tiene aún mucho camino que recorrer

CC BY NC SA

Meetup de R de Madrid

- Página de meetup (link)
- Página del grupo con vídeos y presentaciones (link)

http://madrid.r-es.org/

CC BY NC SA

GRACIAS

Datos de contacto:

Santiago Mota Herce

E-mail: santiago_mota@yahoo.es
Github: https://github.com/santiagomota
LinkedIn: https://es.linkedin.com/in/santiagomota

CC BY NC SA