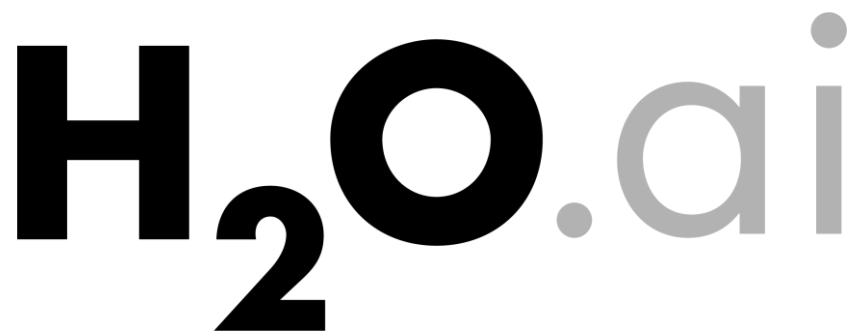


# From Kaggle to H<sub>2</sub>O

The true story of a civil engineer turned data geek



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

@matlabulous

SV Big Data Science at H2O.ai  
28<sup>th</sup> February, 2017

# About Me

- Civil (Water) Engineer

**2010 – 2015**

- Consultant (UK)

- Utilities

- Asset Management

- Constrained Optimization

- Industrial PhD (UK)

- Infrastructure Design Optimization

- Machine Learning +  
Water Engineering

- Discovered H<sub>2</sub>O in 2014

- Data Scientist

**2015**

- Virgin Media (UK)

- Domino Data Lab (Silicon Valley)

**2016 – Present**

- H<sub>2</sub>O.ai (Silicon Valley)

# Agenda

- My Data Science Journey
  - Life as a Water Engineer
  - Massive Open Online Course
  - Kaggle
  - New Skills
  - Side Projects
  - New Opportunities
  - Discovery of H<sub>2</sub>O & Domino
- To Kaggle, or not to Kaggle
  - Joy, Pain, Fear, Gain ...
  - ... and New Friends
- Life as a Data Scientist
- Using H<sub>2</sub>O for Kaggle
  - Rossmann Store Sales
  - Santander Products Recommendation
- Conclusions

# Life as a Water Engineer

# Joe the Outlier

Figure 1. Magic Quadrant for Data Science Platforms



# Massive Open Online Course (MOOC)

# My First MOOC Experience

- Introduction to AI (2011)
  - One of the first MOOCs
- Key messages from Sebastian Thrun:
  - “Just dive into it.”
  - “Get your hands dirty.”
- Met new friends
  - Decided to collaborate for fun
  - “How about Kaggle?”
  - “What is Kaggle?”

The image shows the landing page for the Stanford University's Introduction to Artificial Intelligence MOOC. At the top left is the Stanford Engineering logo. To the right, the text "Oct. 10 ~ DEC. 16, 2011" is displayed. In the center, there is a large image of a man's face with a metallic, robotic-looking mask over his eyes. To the right of the image, the text "INTRODUCTION TO Artificial Intelligence" is written in large, bold, blue letters. Below this, in red text, it says "In partnership with the Stanford University School of Engineering. You can join this online worldwide class this fall." At the bottom left, there is a portrait photo of Sebastian Thrun. To the right of the photo, a bio describes him as a Research Professor of Computer Science at Stanford University, a Google Fellow, and a member of the National Academy of Engineering and the German Academy of Sciences. It notes his work in robotics and machine learning. On the far right, there is a message indicating that signup is temporarily unavailable, along with social media links for Twitter and Facebook, and a note that over 135,000 people have signed up.



Sebastian  
Thrun

Sebastian Thrun is a Research Professor of Computer Science at Stanford University, a Google Fellow, a member of the National Academy of Engineering and the German Academy of Sciences. Thrun is best known for his research in robotics and machine learning. Fast Company Magazine selected him as the fifth most creative person in business, the UK Telegraph included him in their list of 100 living geniuses, and Popular Science included him in their list of Brilliant Ten. His self-driving car was

Signup is temporarily unavailable. Please check back in a few hours.

[Follow](#) [@aiclass](#)

Over 135,000 have signed up!

We're setting up the official registration page right now.

Stanford's [Introduction to Databases](#) and [Introduction to Machine Learning](#) are also available online this fall!

# About Kaggle

- World's biggest data mining competition platform
- Competition Types:
  - Featured (w/ Prize)
  - Recruitment
  - Playground
  - Beginner (101)

	<b>Data Science Bowl 2017</b> Can you improve lung cancer detection? <small>Featured · 2 months to go · Entered · 603 kernels</small>	\$1,000,000 1,256 teams
	<b>The Nature Conservancy Fisheries Monitoring</b> Can you detect and classify species of fish? <small>Featured · 2 months to go · 286 kernels</small>	\$150,000 1,646 teams
	<b>Google Cloud &amp; YouTube-8M Video Understanding Challenge</b> Can you produce the best video tag predictions? <small>Featured · 3 months to go · 44 kernels</small>	\$100,000 163 teams
	<b>Dstl Satellite Imagery Feature Detection</b> Can you train an eye in the sky? <small>Featured · 8 days to go · 158 kernels</small>	\$100,000 363 teams
	<b>Two Sigma Financial Modeling Challenge</b> Can you uncover predictive value in an uncertain world? <small>Featured · 2 days to go · 215 kernels</small>	\$100,000 2,061 teams
	<b>Two Sigma Connect: Rental Listing Inquiries</b> How much interest will a new rental listing on RentHop receive? <small>Recruitment · 2 months to go · 263 kernels</small>	Jobs 714 teams
	<b>Dogs vs. Cats Redux: Kernels Edition</b> Distinguish images of dogs from cats <small>Playground · 3 days to go · 250 kernels</small>	1,249 teams
	<b>Transfer Learning on Stack Exchange Tags</b> Predict tags from models trained on unrelated topics <small>Playground · 1 month to go · 112 kernels</small>	297 teams
	<b>March Machine Learning Mania 2017</b> Predict the 2017 NCAA Basketball Tournament <small>Playground · 16 days to go · 24 kernels</small>	Swag 243 teams
	<b>House Prices: Advanced Regression Techniques</b> Sold! How do home features add up to its price tag? <small>Playground · 2 days to go · Entered · 1,016 kernels</small>	4,766 teams
	<b>Leaf Classification</b> Can you see the random forest for the leaves? <small>Playground · 15 hours to go · 432 kernels</small>	1,587 teams
	<b>Digit Recognizer</b> Classify handwritten digits using the famous MNIST data <small>Getting Started · 3 years to go · Entered · 2,437 kernels</small>	1,422 teams
	<b>Titanic: Machine Learning from Disaster</b> Predict survival on the Titanic using Excel, Python, R & Random Forests <small>Getting Started · 3 years to go · 6,476 kernels</small>	5,943 teams

# My Very First Kaggle Experience

- Predict Bond Trade Price
  - No domain knowledge
  - Lots of numbers (I couldn't open the CSV in Excel)
- Regression Models
  - Random Forest
  - Support Vector Machine
  - Neural Networks
- Black Magic or Data Science?
  - Still, I wasn't so sure

Completed • \$17,500  
**Benchmark Bond Trade Price Challenge**  
Fri 27 Jan 2012 – Mon 30 Apr 2012 (4 years ago)

Team woobe & Me, Myself and AI Details

Vikram Jha Jo-fai Chow octonion leader ritesh

Mariahbarrio Mansi Sudip\_Jerry Sourangsu

Yousuf mohit Noureldin

# Teamwork

- Problems

- “Hey Joe, you are a nice guy.”
- “... but we can’t work together.”
- “Okay, wait ... why?
- “You love MATLAB so much.”
- “You even have a fan boy twitter handle!”

**Jo-fai (Joe) Chow**  
@matlabulous

- Problems

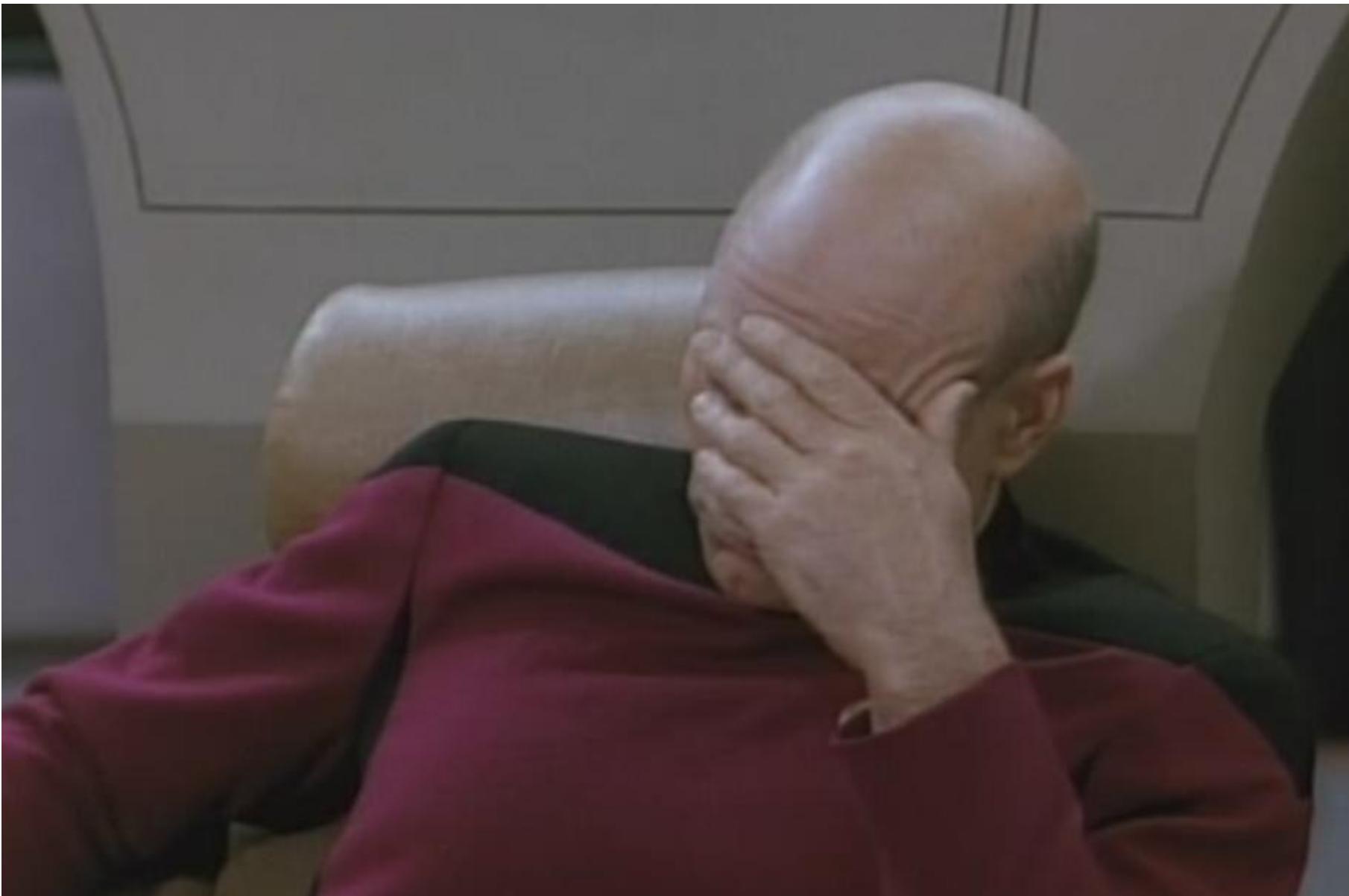
- “We prefer open source tools like R or Python.”
- “Wait ... you guys can use Octave”
- “Thanks, but no thanks ...”

- Solution

- I kept using MATLAB
- Lone wolf
- ZERO collaboration



87th/264



# Adapt or Die

If you can't change ~~the world~~ your friends, change yourself.

# Identifying Skill Gaps

- Obvious Skill Gaps
  - Open-source Programming Languages
  - Machine Learning Techniques
  - Big Data
  - Collaboration
- Kind of Related
  - Data Visualization
  - Explaining Results

## • Where to Start?

The screenshot shows the homepage of R-bloggers.com. At the top, there's a navigation bar with links for Home, About, RSS, add your blog!, Learn R, R jobs, and Contact us. The main header features the R-bloggers logo (an orange R with a blue feed icon) and the text "R news and tutorials contributed by (750) R bloggers". Below the header, a featured article is displayed with the title "Make your R simulation models 20 times faster", dated February 27, 2017, and written by Blueecology blog. The article includes a line graph showing population over time for three different solutions. To the right, there's a sidebar with a search bar, a "RECENT POPULAR POSTS" section listing articles like "Make your R simulation models 20 times faster" and "Reinforcement Learning in R", and a "MOST VISITED ARTICLES OF THE WEEK" section listing items 1 through 9. At the bottom, there's a "SPONSORS" section with the EARL logo and a "Call for abstracts" button.

<https://www.r-bloggers.com/>

# R Can Do That ?

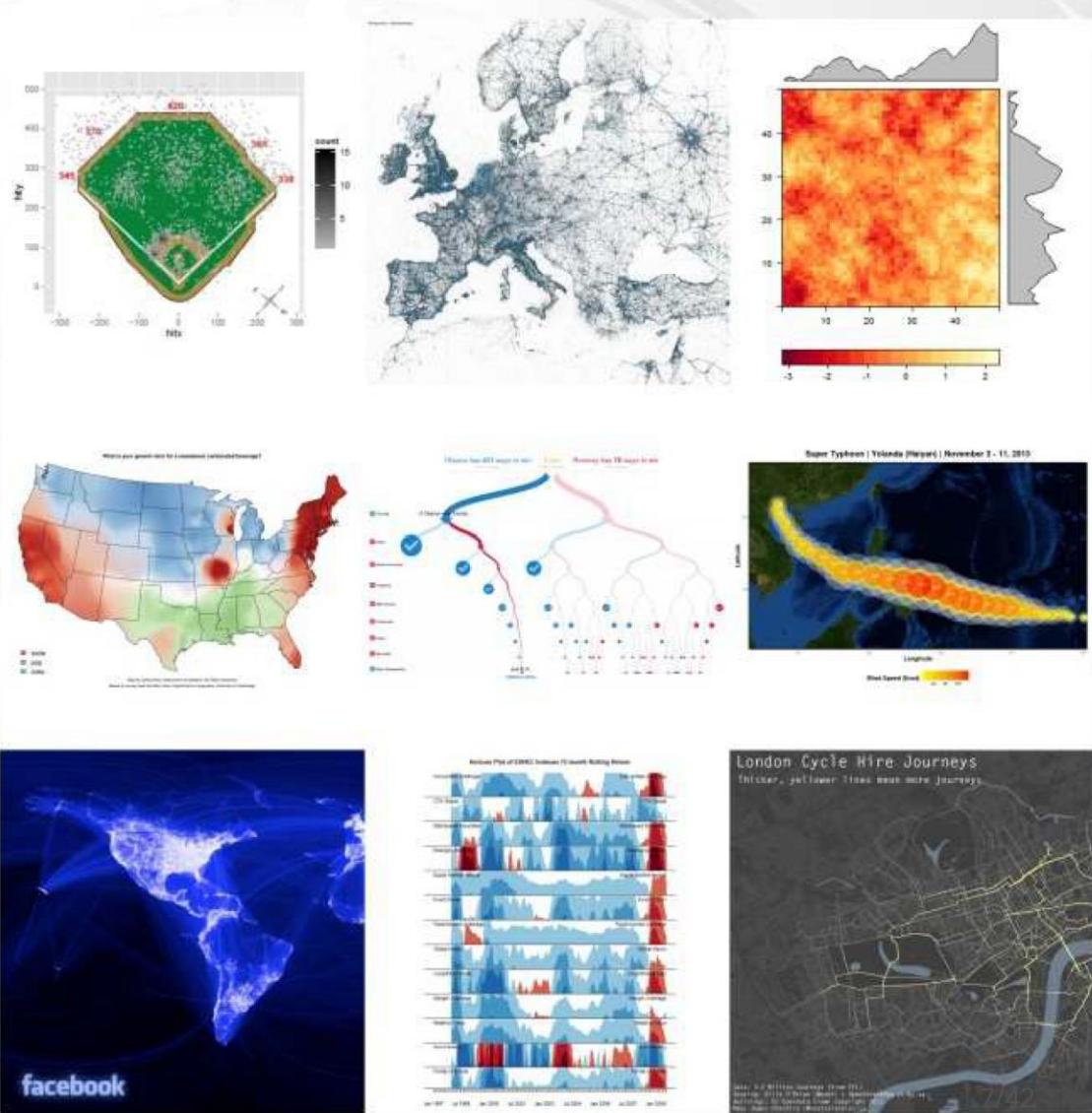


... and more !?

Thanks to Tal Galili's

**R-bloggers**

R news and tutorials contributed by (452) R-bloggers



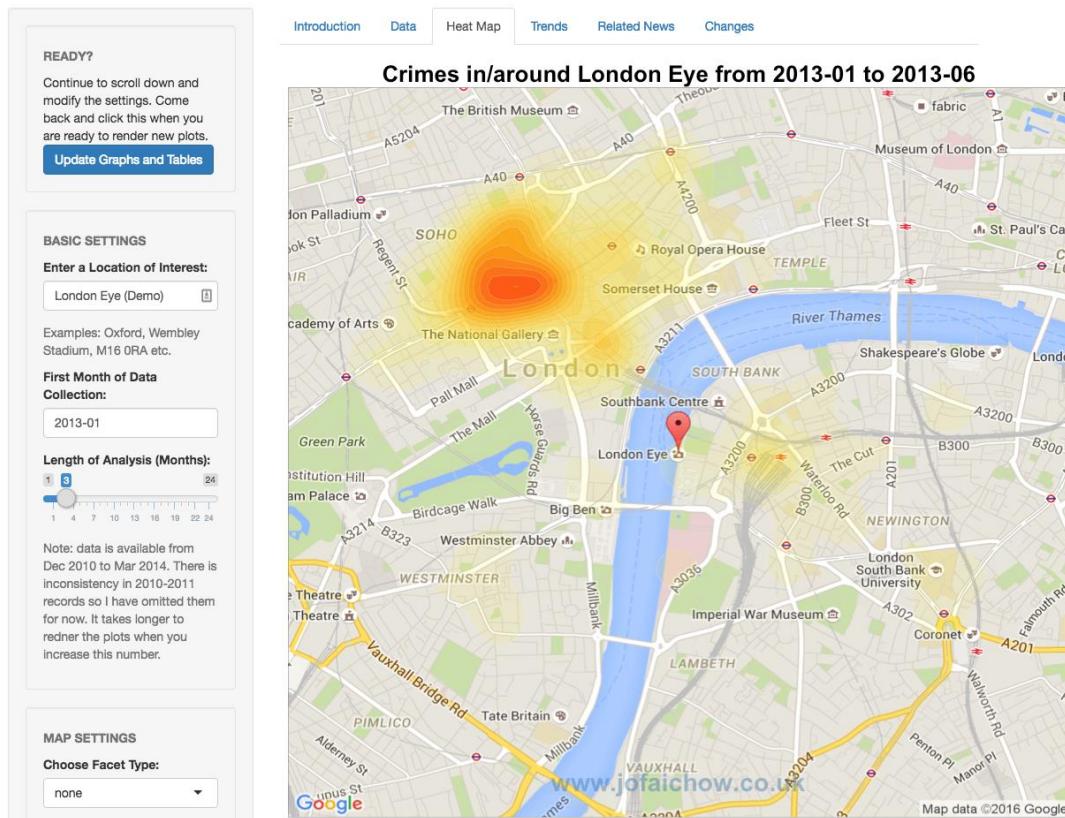
# Learn

- More MOOCs
  - Machine Learning
    - Andrew Ng (Coursera)
    - MATLAB / Octave
  - Data Analysis
    - Jeff Leek (Coursera)
    - R
  - Intro to Programming
    - Dave Evans (Udacity)
    - Python
- Kaggle Forums
  - Tricks you can't learn from schools/books
- Skills I also picked up
  - Linux – Ubuntu\*
  - Git (I mean Git with GUI)
  - Cloud
  - HTML / CSS

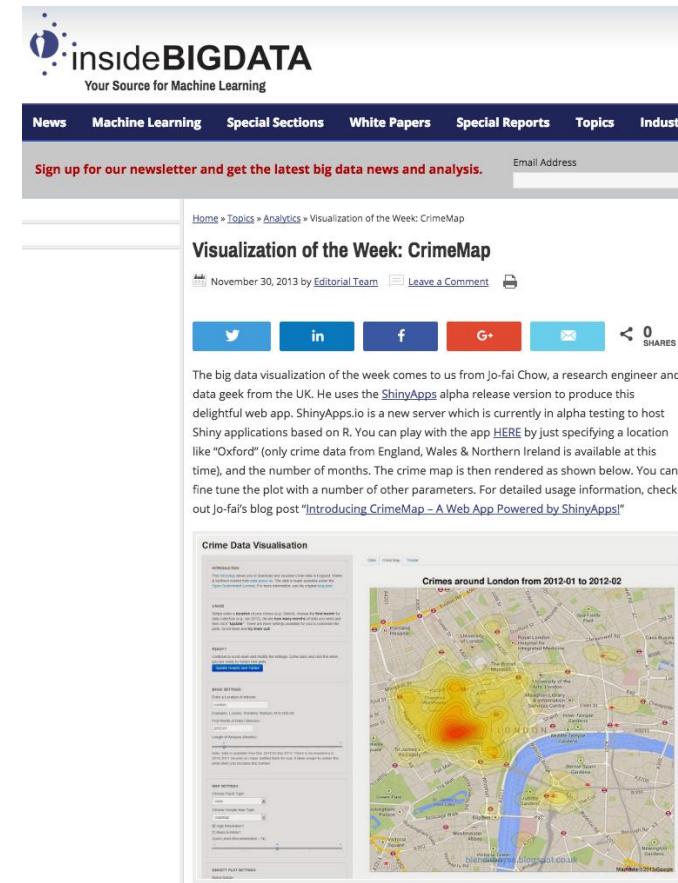
*\*Ubuntu is an ancient African word that means “I can’t configure Debian.”*

# Side Project #1 – Crime Data Visualization

## Crime Data Visualisation

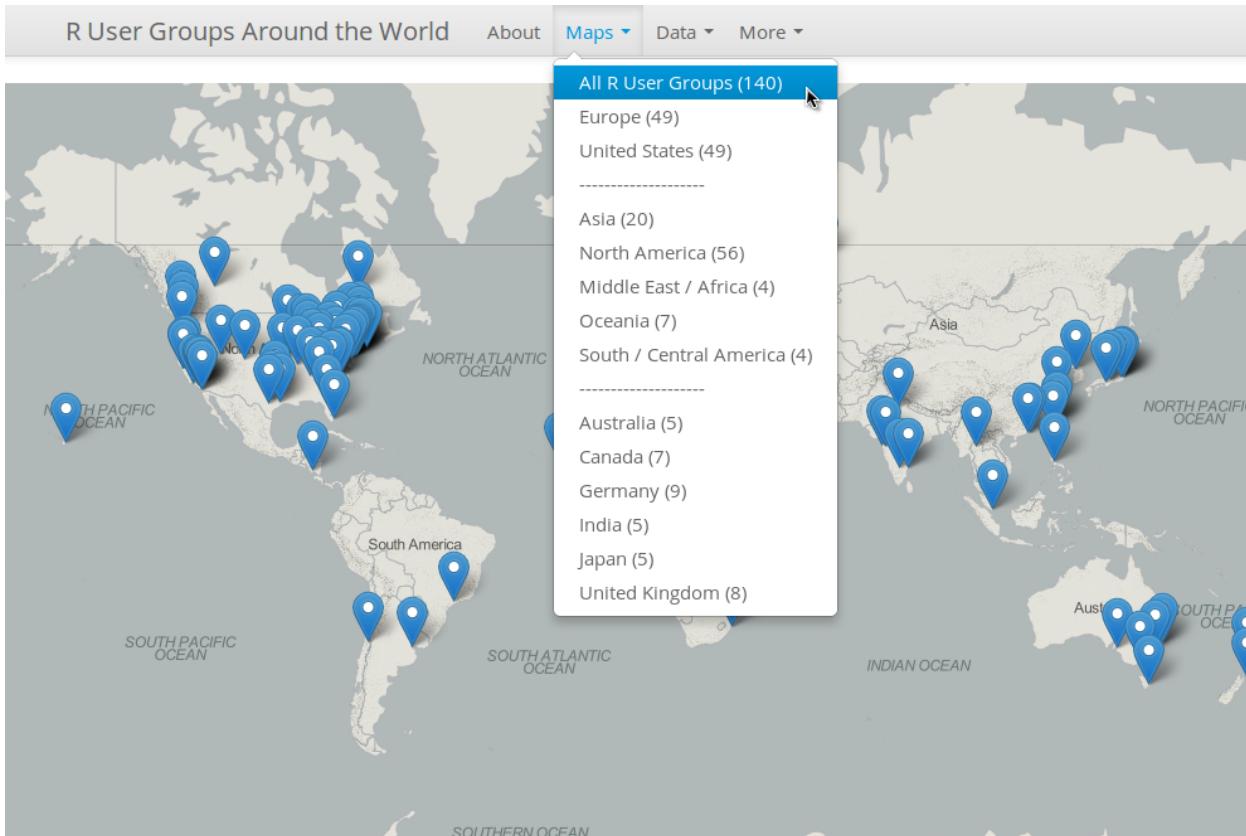


<https://github.com/woobe/rApps/tree/master/crimemap>

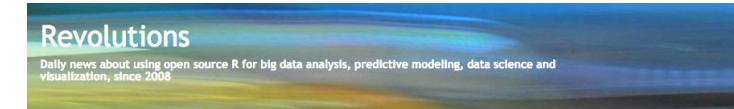


<http://insidebigdata.com/2013/11/30/visualization-week-crimemap/>

# Side Project #2 – Data Visualization Contest



<https://github.com/woobe/rugsmaps>



August 21, 2014

## Revolution Analytics' User Group Map Contest has a Winner

by Joseph Rickert

We are pleased to announce that [jo-fai Chow](#) is the winner of the Revolution Analytics contest. jo-fai's entry, which was implemented as a [Shiny project](#), may be viewed by clicking on the figure below.

### R User Groups Around the World

About Maps ▾ Data ▾ More ▾



jo-fai's work not only produced an aesthetically pleasing sequence of maps but also provides a superb example of a well-documented, small project developed on [Shiny](#) and [GitHub](#). The multiple maps are very nicely rendered, allow for zooming in and pulling back, and display information differently depending on the scale. A nice touch is the code to clean the data set. We wish to thank jo-fai for taking the trouble to craft an entry that exceeds the contest requirements by providing a roadmap for others to follow.

Information  
About this blog  
Comments Policy  
About Categories  
About the Authors  
R Community Calendar  
Local R User Group Directory

#### Search Revolutions Blog

Search Blog

Got comments or suggestions for the blog editor?  
Email [David Smith](#).

[Follow David on Twitter: @revodavid](#)  
[+David Smith](#)

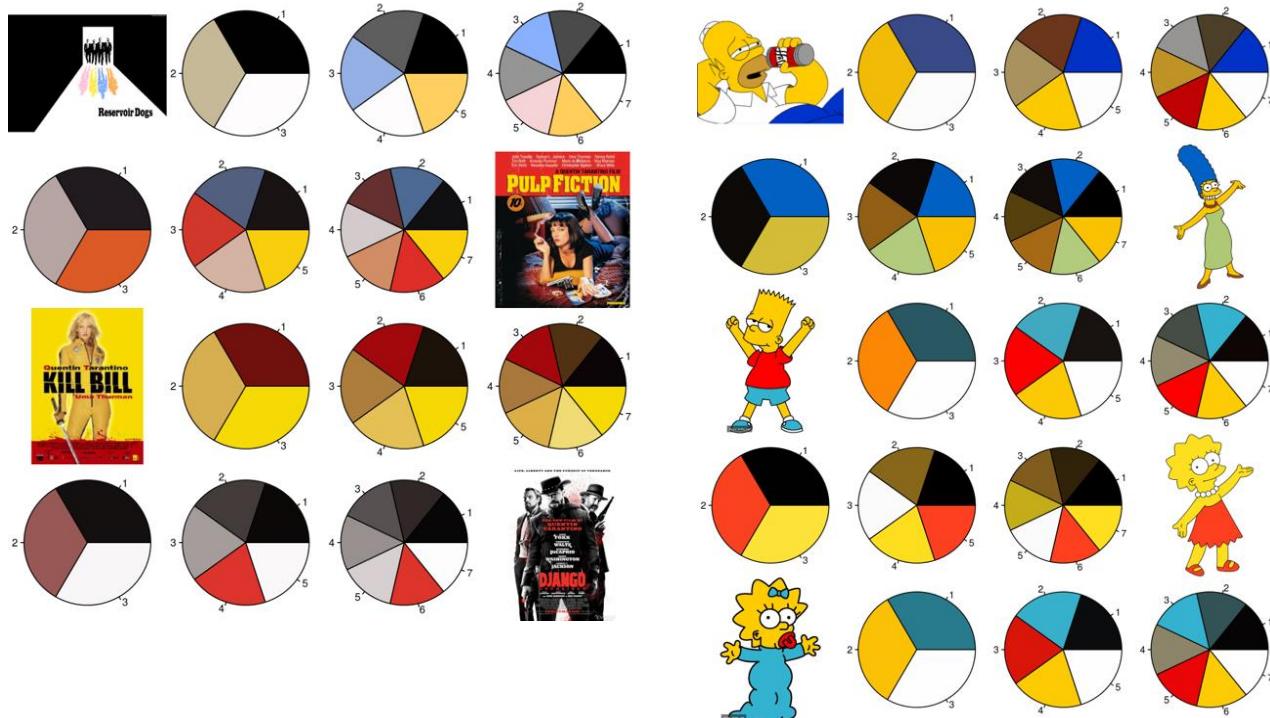
[Blogtrottr](#)

[Subscribe to this blog's feed](#)

#### Categories

academia  
advanced tips  
announcements  
applications  
beginner tips  
big data  
courses  
current events  
data science  
developer tips  
events  
finance  
government  
graphics  
high-performance computing  
life sciences  
Microsoft  
open source  
other industry  
packages  
popularity  
predictive analytics  
profiles  
R

# Side Project #3 – Color Extraction



<https://github.com/woobe/rPlotter>

#TheDress

**Revolutions**  
Daily news about using open source R for big data analysis, predictive modeling, data science and visualization, since 2008

March 04, 2015

**Color extraction with R**

Given all the attention the internet has given to the colors of this dress, I thought it would be interesting to look at the capabilities for extracting colors in R.

R has a number of packages for importing images in various file formats, including `jpeg`, `png`, `TIFF` and `PDF`. (`readbitmap` also works with all of them.) In each case, the image is a 3-dimensional array containing a 2D image layer for each of the color channels (for example red, green and blue for color images). You can then manipulate the array as ordinary data to extract color information. For example, Derek Jones has a nice blog post on how to do this to extract data from published heatmaps when the source data has been lost.

Photographs typically contain thousands or even millions of unique colors, but a very human question is: what are the major colors in the image? In other words, what is the image's palette? This is a difficult question to answer, but Russell Dinnage used R's k-means clustering capabilities to extract the 3 (or 4 or 5) most prominent colors from an image, while discarding all other perceptual shades of the same color and filtering out low-saturation background colors (like gray shadows). Without any supervision, his script can easily extract 6 colors from this tail of this beautiful peacock spider. In fact, his script generates five representative palettes:

I used a similar process to extract the 3 major colors from "that dress":

**Information**

- About this blog
- Contact Policy
- About Categories
- About the Authors
- R Community Calendar
- Local R User Group Directory

**Search Revolutions Blog**

Got comments or suggestions for the blog editor?  
Email [David Smith](#).

Follow David on Twitter: [@revodavid](#)  
[@David\\_Smith](#)

**Blognarr**

Subscribe to this blog's feed

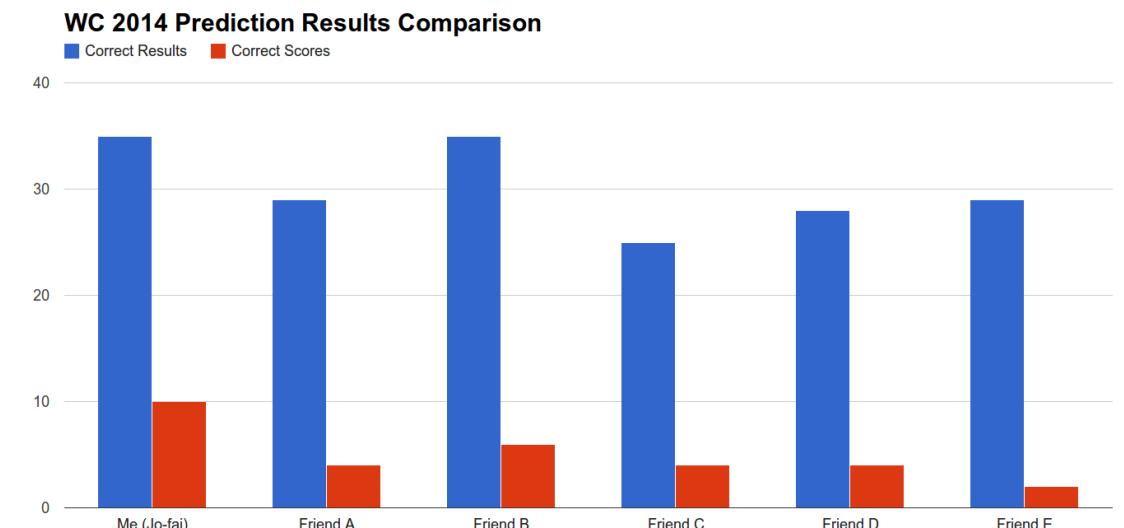
**Categories**

- academia
- advanced R tips
- announcements
- applications
- beginner tips
- big data
- courses
- current events
- data science
- devtools tips
- events
- Finance
- government
- graphics
- high-performance computing
- life sciences
- Machine Learning
- open source
- other industry
- packages
- popularity
- predictive analytics
- profiles
- R
- R User
- random
- reviews
- Revolution
- Rmedia
- roundups
- sports
- statistics
- user groups

<http://blog.revolutionanalytics.com/2015/03/color-extraction-with-r.html>

# Side Project #4 – World Cup 2014 Prediction

- Joe (Machine Learning) vs. Friends
  - Correct Results (WDL)
    - ML: 35 / 64 (55%)
    - Friends (Avg): 29 / 64 (46%)
  - Correct Score
    - ML: 10 / 64 (16%)
    - Friends (Avg): 4 / 64 (6%)



<https://github.com/woobe/wc2014>

# Open Up Myself



# New Opportunities

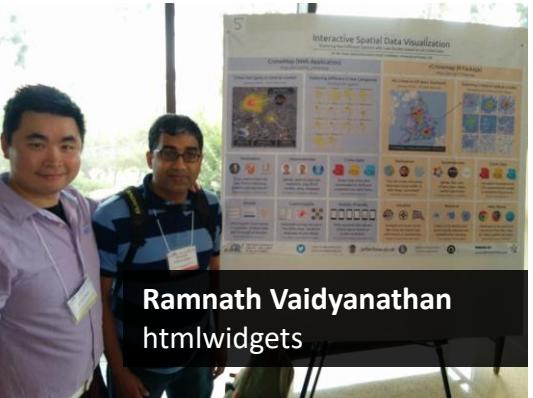
R Community, H2O & Domino Data Lab

# LondonR 2013 & useR! 2014

[http://www.jofaichow.co.uk/2014\\_03\\_11\\_LondonR/](http://www.jofaichow.co.uk/2014_03_11_LondonR/)

[https://github.com/woobe/useR\\_2014](https://github.com/woobe/useR_2014)

# useR! 2014



# Hardware tier

<https://blog.dominodatalab.com/using-r-h2o-and-domino-for-a-kaggle-competition/>

## How to use R, H2O, and Domino for a Kaggle competition

data science

R

✓ Free: 1 core, 1GB RAM  
Small: 2 cores, 8GB RAM



by [Nick Elprin](#) on September 19th, 2014

 SHARE

Search



*This is a guest post by [Jo-Fai Chow](#)*

The sample project (code and data) described below is [available on Domino](#).

If you're in a hurry, feel free to skip to:

- Tutorial 1: [Using Domino](#)
- Tutorial 2: [Using H2O to Predict Soil Properties](#)
- Tutorial 3: [Scaling up your analysis](#)

Introduction



# Dear Kaggle

Joy, Pain, Fear, Gain ... and New Friends ☺

# Kaggle – The Joy

kaggle

Customer Solutions Competitions Community ▾

\$8,000 • 413 teams

Africa Soil Property Prediction Challenge

Enter/Merge by

Wed 27 Aug 2014

Tue 21 Oct 2014 (40 days to go)

Dashboard Leaderboard - Africa Soil Property Prediction Challenge

This leaderboard is calculated on approximately 13% of the test data.  
The final results will be based on the other 87%, so the final standings may be different.

See someone using multiple accounts?  
[Let us know.](#)

#	Δ3d	Team Name * <small>in the money</small>	Score	Entries	Last Submission UTC
1	+13	Jo-fai Chow @ blenditbayes! + h2o.ai + Domino *	0.40406	23	Thu, 11 Sep 2014 21:12:59 (-14.5h)

Your Best Entry  
**Number One!**  
You jumped into first by improving your score by 0.00638.

# Kaggle – The Pain & The Fear

513 ▾ 449 Jo-fai Chow @ blenditbayes! + H2...



0.51401

133

2y



# Kaggle – The Gain

- New Skills
  - Exploratory Data Analysis
  - Machine Learning Algorithms
  - Feature Engineering
  - Model Stacking
  - Communication
- **THE FEAR OF OVERFITTING!**

- New Friends

- London Kaggle Meetup



**Mick**



**Yifan Xie**



**ZFTurbo**



**anokas**



# Life as a Data Scientist

# Toy (In-Class) vs. Kaggle vs. Real-World Data

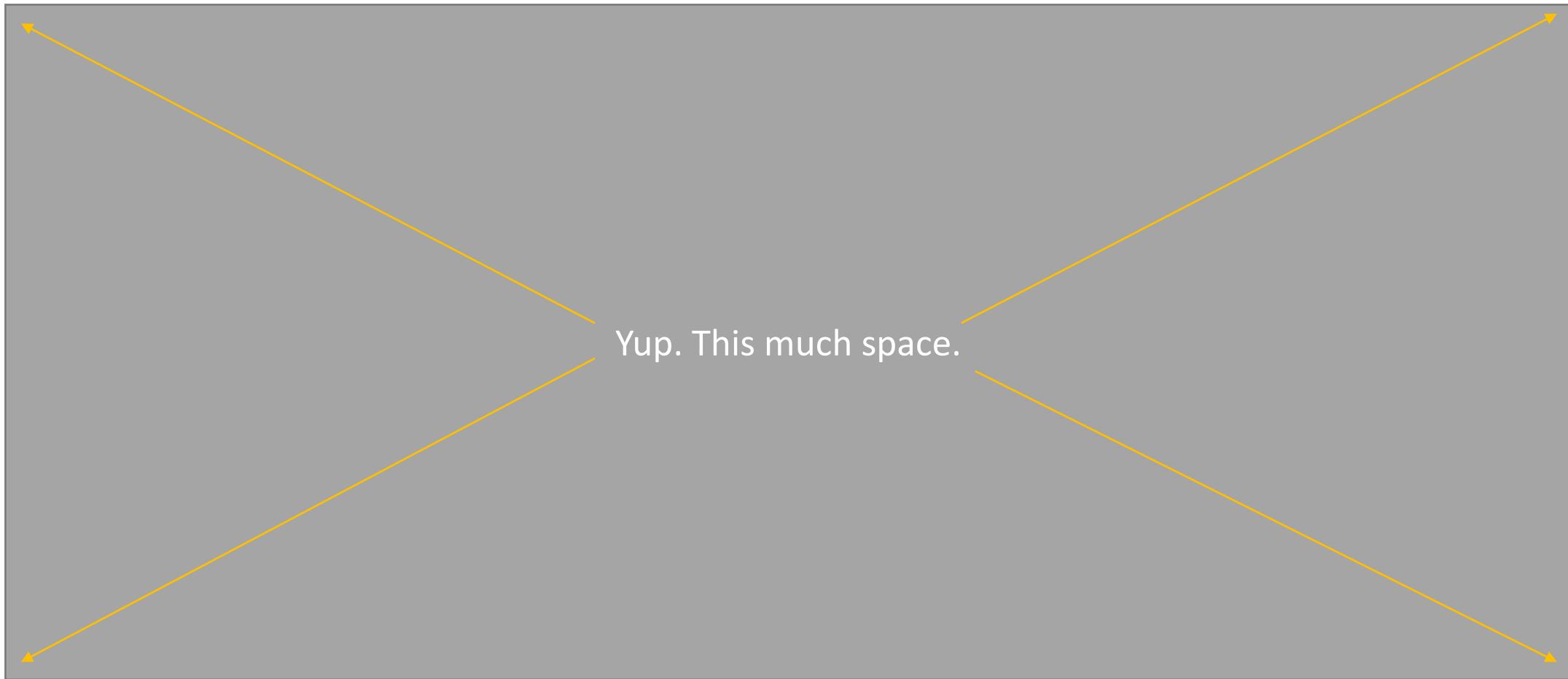


# Story Telling



AS THE YEARS WENT BY

# Story Telling with One Single Slide



# Using H<sub>2</sub>O for Kaggle



## Jo-fai Chow

Data Scientist at H2O.ai

London, England, United Kingdom

Joined 5 years ago · last seen in the past day



<http://www.jofaichow.co.uk/>



Competitions Expert

[Home](#)

Competitions (40)

Kernels (6)

Discussion (24)

Datasets (0)

More

[Edit Profile](#)

### Competitions Expert



Current Rank

**818**  
of 54,344

Highest Rank

**181**



0



2



5

[Santander Product Recom...](#)

38<sup>th</sup>  
of 1787

[Rossmann Store Sales](#)

57<sup>th</sup>  
of 3303

[Personality Prediction Base...](#)

21<sup>st</sup>  
of 89

### Kernels Contributor



Unranked



0



0



0

[XGB\\_test\\_001](#)

a year ago

0  
votes

[H2O Starter GBM](#)

a year ago

0  
votes

[Testing](#)

a year ago

0  
votes

[Discussion Contributor](#)



Unranked



1



1



12

[H2O Deep Learning Start...](#)

2 years ago

21  
votes

[Best Ensemble References?](#)

2 years ago

7  
votes

[Leaderboard scores](#)

2 years ago

4  
votes

Bio

Edit

# Rossmann Store Sales

- Stuck at top 10% for a long time
- Final Breakthrough (Mickael)
  - Added external data – weather in different cities
  - 48 hours left
- Model Stacking (Joe)
  - H<sub>2</sub>O Deep Learning
  - Xgboost
  - Manual process (life before h2oEnsemble / Stacked Ensembles in H<sub>2</sub>O)



**ROSSMANN**

**Rossmann Store Sales**  
Forecast sales using store, promotion, and competitor data  
\$35,000 · 3,303 teams · a year ago

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [More](#) [My Submissions](#) [Submit Predictions](#)

**Description** [Evaluation](#) [Prizes](#) [Timeline](#)

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

In their first Kaggle competition, Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what's most important to them: their customers and their teams!

A row of nine small icons representing various product categories, including skincare (bottles), cleaning (spray bottle, sponge), and personal care (toothbrush, comb).

If you are interested in joining Rossmann at their headquarters near Hanover, Germany, please contact Mr. Frank König (Frank.Koenig {at} rossmann.de) Rossmann is currently recruiting data scientists at [senior](#) and [entry-level](#) positions.

**Rossmann Store Sales**

· a year ago · Top 2%

**57<sup>th</sup>**  
of 3303



# Santander Product Recommendation

- Predict new products that customers will add in the future
- Reframed as a Multiclass Classification (see next slide)
- Feature Engineering
  - Basic (Everyone)
  - Advanced (ZFTurbo, Yifan, Anokas)
  - Also see [Yifan's slides](#)
- Models
  - xgboost (ZFTurbo)
  - H<sub>2</sub>O GBM (Joe) – Single Best Model

**Santander Product Recommendation**  
Can you pair products with people?  
\$60,000 · 1,787 teams · 2 months ago

Overview Data Kernels Discussion Leaderboard More My Submissions Submit Predictions

Ready to make a downpayment on your first house? Or looking to leverage the equity in the home you have? To support needs for a range of financial decisions, Santander Bank offers a lending hand to their customers through personalized product recommendations.



Under their current system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. In their second competition, Santander is challenging Kagglers to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers.

With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

**Disclaimer:** This data set does not include any real Santander Spain's customer, and thus it is not representative of Spain's customer base.



## Santander Product Recommendation

Can you pair products with people?

\$60,000 · 1,787 teams · 2 months ago

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Leaderboard](#)[More](#)[My Submissions](#)[New Topic](#)**BreakfastPirate**

5th place

### When Less is More

posted in [Santander Product Recommendation](#) 3 months ago



89

This might not be the best approach, but it seems to be working okay, and it is fairly fast since it does not use much data.

I'm only training on about 37,000 accounts – only the accounts that added a new product in June 2015. I found that the distribution of products added varied a lot by month, and June seemed to be an unusual month. Since we are predicting June 2016, I trained on only June 2015.

I only used accounts that added a new product in June 2015. We are not trying to determine who will add a new product, we are only trying to predict which products they would add **if they did**. So I excluded all accounts that didn't add a product.

I collapsed the multiple, binary dependent variables into a single multi-value dependent variable. If an account added multiple products in June 2015, then I simply had that line of data in my training set multiple times – once for each product added. So I ended up with about 46,000 rows in my training set. I then used multiclass classification (multi:softprob) to give me the probability that each product will be added, and I selected the 7 with the highest scores (and that also were not products that were owned in the previous month since those are by definition not new).

[Options](#)

## kaggle-santander



Private

... Show Menu

## Backlog

Date &amp; time features

1

Encode categorical features

1

Create status feature

Logistic regression on each product

Add useful steps taken on previous winning solution to the backlog

Antigüedad is the same for the 1st 7 months

223,967 clients never used any of the product

Customer profile feature

1

Try xgboost on GPU

Add a card...

## To Do

Revisit H2O's GLRM



Feature 2: Create linear regression on all sparse data predicting target variable independently. We again need to do so CV way. Split data on N=10 part and create linear regression validation predictions. So 24 columns in feature file.

Generate frequency feature for categorical variables with high cardinality

K-mean features: generate K-mean cluster feature for customer - product ownership

total amount of time a product is owned by a customer

Previous-month product purchase feature

Add a card...

## Doing

Feature 1: Minimum and average cosine sim between products vector for 5 previous months with any record where given product was added by customer. Since we use target variable here, we need to create this feature CV way. eg split all data on (N=10) parts. Choose one part as "validation" for which we calculate feature all other 9 as "train". Then calculate cosine sim for each record in "validation" against each record from "train" where given product was added. Find average and minimum value. Repeat for all 9 other splits. So at the end we will have 24\*2 columns in feature file.

Multi-class h2o stacking (final push only)

1

Add a card...

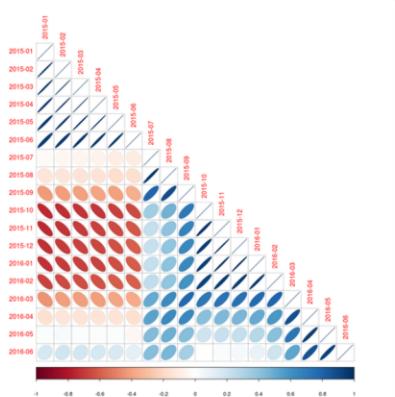
## Review

On a similar note, has anyone seen records = 3948607 fecha\_alta > 1999-07-14. So first contact with person happened in 1999-07-14, on 2015-01-25 is 192. He is with bank for last 16 years!!! Surprise!!! Any pointers?

#	fecha_alta	fecha_recibe_cuenta	edad	particularidades
1	2013-01-28	2013-01-28	192	3 MA SI PARTICULARES
2	2013-02-28	2013-02-28	192	3 MA SI PARTICULARES
3	2013-03-28	2013-03-28	192	3 MA SI PARTICULARES
4	2013-04-28	2013-04-28	192	3 MA SI PARTICULARES
5	2013-05-28	2013-05-28	192	3 MA SI PARTICULARES
6	2013-06-28	2013-06-28	192	3 MA SI PARTICULARES
7	2013-07-28	2013-07-28	192	3 MA SI PARTICULARES
8	2013-08-28	2013-08-28	192	3 MA SI PARTICULARES
9	2013-09-28	2013-09-28	192	3 MA SI PARTICULARES
10	2013-10-28	2013-10-28	192	3 MA SI PARTICULARES
11	2013-11-28	2013-11-28	192	3 MA SI PARTICULARES
12	2013-12-28	2013-12-28	192	3 MA SI PARTICULARES
13	2014-01-28	2014-01-28	192	3 MA SI PARTICULARES
14	2014-02-28	2014-02-28	192	3 MA SI PARTICULARES
15	2014-03-28	2014-03-28	192	3 MA SI PARTICULARES
16	2014-04-28	2014-04-28	192	3 MA SI PARTICULARES
17	2014-05-28	2014-05-28	192	3 MA SI PARTICULARES

Investigate the impact of surprisingly young customers

2 1



k-means & PCA for customer info (without product info)

6 1

Post-Prediction processing check list

0/3

Feature Engineering Check List

Add a card...

## Done

Basic data prep in R



Removing highly correlated product lag features

1

Test and set up large scale features selection on Domino

1

Include May-16 for training and use kfold CV for early stopping

1

Deal with outliers

3 0/1

Try H2O GBM + DRF + DNN Ensemble and evaluate local validation set

1

Try H2O Deep Learning, RF, Deep

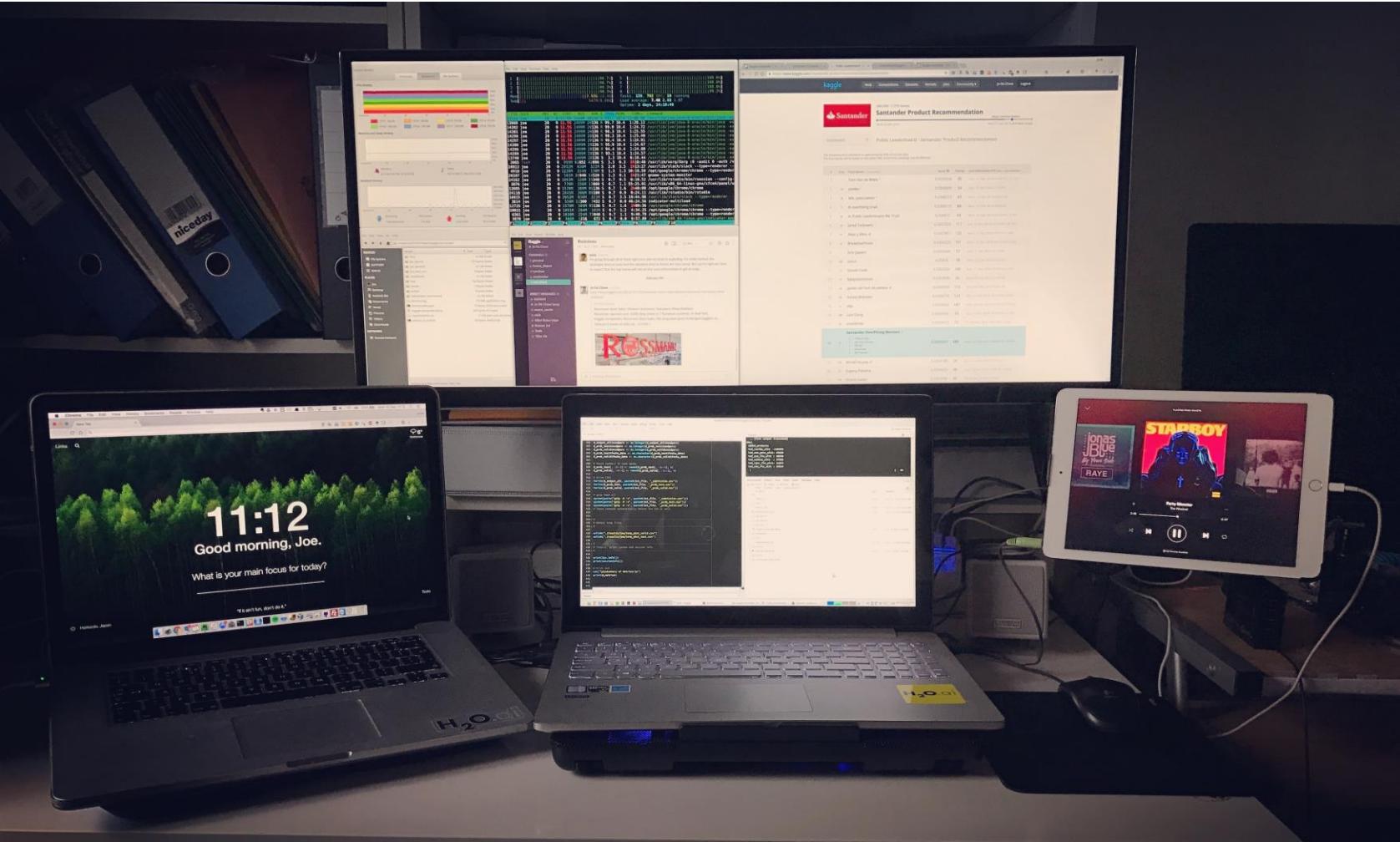
Add a card...

## Submissions

Add a card...

kaggle-santander		joe_gbm_038																					
File Edit View Insert Format Data Tools Add-ons Help		Last edit was on 20 December 2016																					
		Formatting Tools																					
fx		joe_gbm_038																					
1	run	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S			
2		data source	preprocessing	training	validation	lag	algo	ntrees / epochs	learn_rate	row_samp	col_samp	max depth	pred noise bw	run time	early stopping	mlogloss	mean per class error	overall class error	Public LB	Public LB (20/21 fix)			
40	v8 start from raw data, only with Joe's features (with flag_closed_acc) = NEW BASELINE MODEL!!!																						
41	joe_gbm_032	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	672	0.05	0.9	0.5	5	0.00	2h33m	3/100/mpcrr	0.9154	0.9302	0.5373	0.5656	0.3443	0.3557	N/A	N/A
42	joe_gbm_033	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	693	0.05	0.8	0.8	5	0.00	2h37m	3/100/mpcrr	0.9138	0.9375	0.5246	0.5702	0.3437	0.3569	N/A	N/A
43	joe_gbm_034	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	350	0.05	0.9	0.5	5	0.00	48m	10/10/mpcrr	0.9107	0.9310	0.4886	0.5699	0.3432	0.3556	N/A	N/A
44	joe_gbm_035	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	360	0.05	0.8	0.8	5	0.00	52m	10/10/mpcrr	0.9103	0.9354	0.5195	0.5708	0.3432	0.3568	N/A	N/A
45	joe_gbm_036	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	205	0.05	0.9	0.5	5	0.00	38m	5/10/mpcrr	0.9190	0.9326	0.5408	0.5700	0.3449	0.3558	N/A	N/A
46	joe_gbm_037	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	170	0.05	0.8	0.8	5	0.00	38m	5/10/mpcrr	0.9281	0.9394	0.5415	0.5726	0.3469	0.3579	N/A	N/A
47	joe_gbm_038	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	350	0.05	0.9	0.5	5	0.00	55m	5/20/mpcrr	0.9107	0.9310	0.4886	0.5699	0.3432	0.3556	0.0303787	N/A
48	joe_gbm_039	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	360	0.05	0.8	0.8	5	0.00	1h1m	5/20/mpcrr	0.9103	0.9354	0.5195	0.5708	0.3432	0.3568	N/A	N/A
49	v9 = v8 + two features from Deep Autoencoder (recon_err and flag_outlier)																						
50	joe_gbm_040	Raw	data_prep_v009	Jun15-Apr16	May-2016	5	H2O_GBM	435	0.05	0.9	0.5	5	0.00	1h3m	5/20/mpcrr	0.9109	0.9359	0.5241	0.5636	0.3432	0.3571	N/A	N/A
51	joe_gbm_041	Raw	data_prep_v009	Jun15-Apr16	May-2016	5	H2O_GBM	305	0.05	0.9	0.5	5	0.00	48m	5/20/mpcrr	0.9125	0.9383	0.5296	0.5709	0.3454	0.3570	N/A	N/A
52	joe_gbm_042	Raw	data_prep_v009	Jun15-Apr16	May-2016	5	H2O_GBM	355	0.05	0.9	0.5	5	0.00	53m	5/20/mpcrr	0.9129	0.9376	0.4848	0.5692	0.3457	0.3556	N/A	N/A
53	v8 + remove high corr																						
54	joe_gbm_043 - cutoff 0.95	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	390	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9274	0.9529	0.5712	0.6027	0.3481	0.3615	N/A	N/A
55	joe_gbm_044 - cutoff 0.90	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	400	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9578	0.9645	0.5975	0.6250	0.3578	0.3651	N/A	N/A
56	joe_gbm_045 - cutoff 0.85	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	335	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9818	0.9832	0.6075	0.6323	0.3649	0.3718	N/A	N/A
57	joe_gbm_046 - cutoff 0.80	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	370	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9827	0.9821	0.6104	0.6361	0.3651	0.3712	N/A	N/A
58	joe_gbm_047 - cutoff 0.75	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	350	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	1.0079	1.0039	0.6196	0.6330	0.3736	0.3774	N/A	N/A
59	joe_gbm_048 - cutoff 0.70	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	390	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	1.0720	1.0726	0.6346	0.6441	0.3958	0.4033	N/A	N/A
60	joe_gbm_049 - cutoff 0.95	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	385	0.05	0.8	0.8	5	0.00	59m	5/20/mpcrr	0.9266	0.9536	0.5738	0.6026	0.3480	0.3627	N/A	N/A
61	joe_gbm_050 - cutoff 0.90	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	375	0.05	0.8	0.8	5	0.00	56m	5/20/mpcrr	0.9575	0.9688	0.5968	0.6257	0.3580	0.3651	N/A	N/A
62	joe_gbm_051 - cutoff 0.85	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	335	0.05	0.8	0.8	5	0.00	50m	5/20/mpcrr	0.9807	0.9886	0.6072	0.6328	0.3643	0.3726	N/A	N/A
63	joe_gbm_052 - cutoff 0.80	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	340	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	0.9827	0.9868	0.6047	0.6314	0.3650	0.3717	N/A	N/A
64	joe_gbm_053 - cutoff 0.75	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	350	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	1.0056	1.0129	0.6066	0.6413	0.3731	0.3800	N/A	N/A
65	joe_gbm_054 - cutoff 0.70	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	385	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	1.0715	1.0773	0.6305	0.6446	0.3953	0.4050	N/A	N/A
66	joe_gbm_055 - cutoff 0.975	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	275	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	0.9157	0.9404	0.5352	0.5753	0.3443	0.3580	N/A	N/A
67	joe_gbm_056 - cutoff 0.99	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	325	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	0.9101	0.9339	0.5270	0.5712	0.3430	0.3575	N/A	N/A
68	joe_gbm_057 - cutoff 0.99	Raw	data_prep_v008	Jun15-Apr16	May-2016	5	H2O_GBM	375	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9116	0.9314	0.5325	0.5658	0.3434	0.3557	N/A	N/A
69	v10 = v8 + ZFTurbo's feat_group_ids_for_purchased_products_detailed																						
70	joe_gbm_058	Raw	data_prep_v010	Jun15-Apr16	May-2016	5	H2O_GBM	300	0.05	0.9	0.5	5	0.00	51m	5/20/mpcrr	0.9036	0.9322	0.5060	0.5595	0.3411	0.3559	N/A	0.0304047
71	joe_gbm_059_rerun	Raw	data_prep_v010	Jun15-Apr16	May-2016	5	H2O_GBM	370	0.05	0.8	0.8	5	0.00	1h3m	5/20/mpcrr	0.9024	0.9350	0.4666	0.5623	0.3409	0.3553	N/A	ensemble 00
72	joe_gbm_060	Raw	data_prep_v010	Jun15-Apr16	May-2016	5	H2O_GBM	200	0.05	0.9	0.5	5	0.00	41m	5/20/mpcrr	0.9088	0.9347	0.5072	0.5678	0.3406	0.3535	N/A	N/A
73	joe_gbm_061	Raw	data_prep_v010	Jun15-Apr16	May-2016	5	H2O_GBM	200	0.05	0.8	0.8	5	0.00	44m	5/20/mpcrr	0.9167	0.9384	0.5181	0.5658	0.3413	0.3553	N/A	N/A
74	v11 = v8 with 6-month lag																						
75	joe_gbm_062	Raw	data_prep_v011	Jun15-Apr16	May-2016	6	H2O_GBM	360	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	0.9019	0.9251	0.5279	0.5622	0.3406	0.3525	N/A	N/A
76	joe_gbm_063	Raw	data_prep_v011	Jun15-Apr16	May-2016	6	H2O_GBM	340	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	0.8997	0.9302	0.5217	0.5662	0.3398	0.3531	N/A	N/A
77	v12 = v8 with 7-month lag																						
78	joe_gbm_064	Raw	data_prep_v012	Jun15-Apr16	May-2016	7	H2O_GBM	325	0.05	0.9	0.5	5	0.00	1h	5/20/mpcrr	1.0161	0.9949	0.5582	0.5967	0.3815	0.3817	N/A	N/A
79	joe_gbm_065	Raw	data_prep_v012	Jun15-Apr16	May-2016	7	H2O_GBM	305	0.05	0.8	0.8	5	0.00	1h	5/20/mpcrr	1.0141	1.0307	0.5509	0.6089	0.3811	0.3925	N/A	N/A
80	v13 = v8 starting from May2015																						
81	joe_gbm_066	Raw	data_prep_v013	May15-Apr16	May-2016	5	H2O_GBM	420	0.05	0.9	0.5	5	0.00		5/20/mpcrr	0.9165	0.9468	0.5382	0.5384	0.3445	0.3560	N/A	0.0303708
82	joe_gbm_067	Raw	data_prep_v013	May15-Apr16	May-2016	5	H2O_GBM	370	0.05	0.8	0.8	5	0.00		5/20/mpcrr	0.9145	0.9576	0.4881	0.5501	0.3439	0.3588	N/A	ensemble 00
83	v14 = v8 starting from Apr2015																						
h2o_gbm_results		joe_stacking		stacking_grid_ref																			

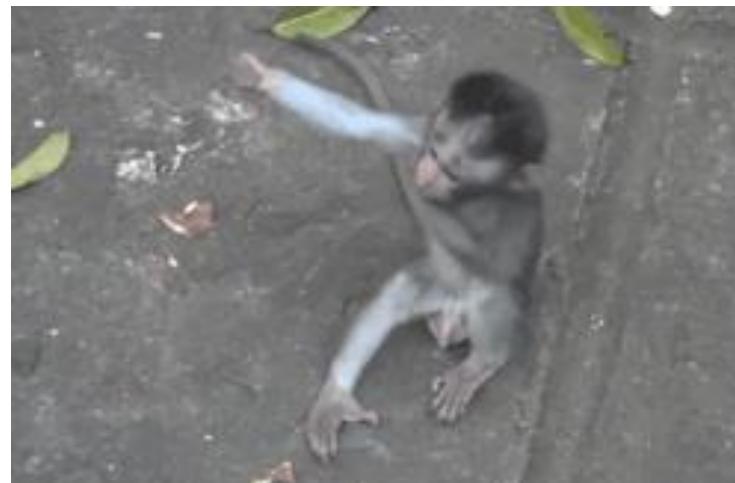
# When I Kaggle ...



# Conclusions

# To Kaggle, or not to Kaggle?

# New Skills, New Friends & New Opportunities



Giphy is your friend when you don't have enough time for bullet points.

# Differences between Kaggle & Data Science



**littleboat** 14:40

1. real world data is messier usually, so it takes a bit time to collect and clean to make them "kaggle like" dataset. I think kaggle admins are actually doing pretty good with every datasets except that sometimes they didn't catch leak. 2. it is normally not easy to find a good evaluation function for real world problem. So instead of all in for one specific metric (accuracy, logloss, etc.), it is more often that you just want to optimize the revenue for the company which is much harder to define in some cases. 3. normally when you want to build a product you will have way more constraints (predicting time, training time and time to build a scalable infrastructure) than kagglng while you don't have to ensemble a lot of models for the extra 0.02% improvement.

Quote from [Littleboat](#)'s AMA on Kagglenoobs Slack Channel

# Thanks!

- People who have helped me along the way
  - Kaggle Friends
  - H<sub>2</sub>O.ai
  - Domino Data Lab
  - Mango Solutions
- Slides
  - [bit.ly/h2o\\_meetups](http://bit.ly/h2o_meetups)
- Contact
  - [joe@h2o.ai](mailto:joe@h2o.ai)
  - [@matlabulous](https://twitter.com/matlabulous)
  - [github.com/woobe](https://github.com/woobe)



“Complexity is your enemy. Any fool can make something complicated. It is hard to make something simple.”

**H<sub>2</sub>O.ai**

Making Machine Learning  
Accessible to Everyone

*Photo credit: Virgin Media*