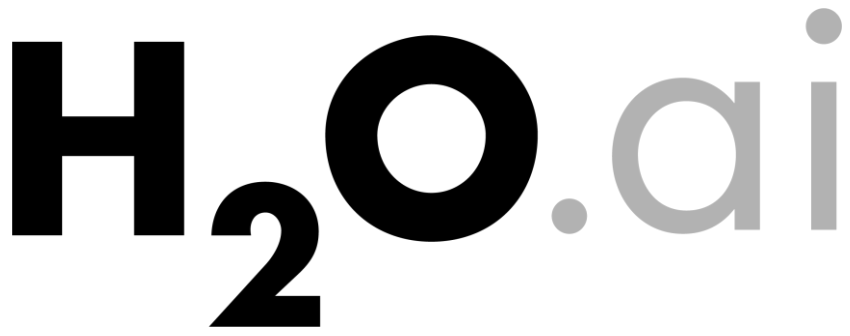
From Kaggle to H_2O

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

28th 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_2O in 2014

- Data Scientist

  2015

  - Virgin Media (UK)
  - Domino Data Lab (Silicon Valley)

  2016

  - H_2O.ai (Silicon Valley)

Agenda

- From Engineer to Data Scientist
  - Life as a Water Engineer
  - Massive Open Online Course
  - Kaggle
  - New Skills
  - Side Projects
  - New Opportunities
  - Discovery of H_2O & Domino
- To Kaggle, or not to Kaggle
  - Joy, Pain, Fear, Gain …
  - … and New Friends
- Life as a Data Scientist
- Using H_2O for Kaggle
  - Rossmann Store Sales
  - Santander Products Recommendation
- Conclusions

Life as a Water Engineer

Joe the Outlier

Joe (Water Engineer)

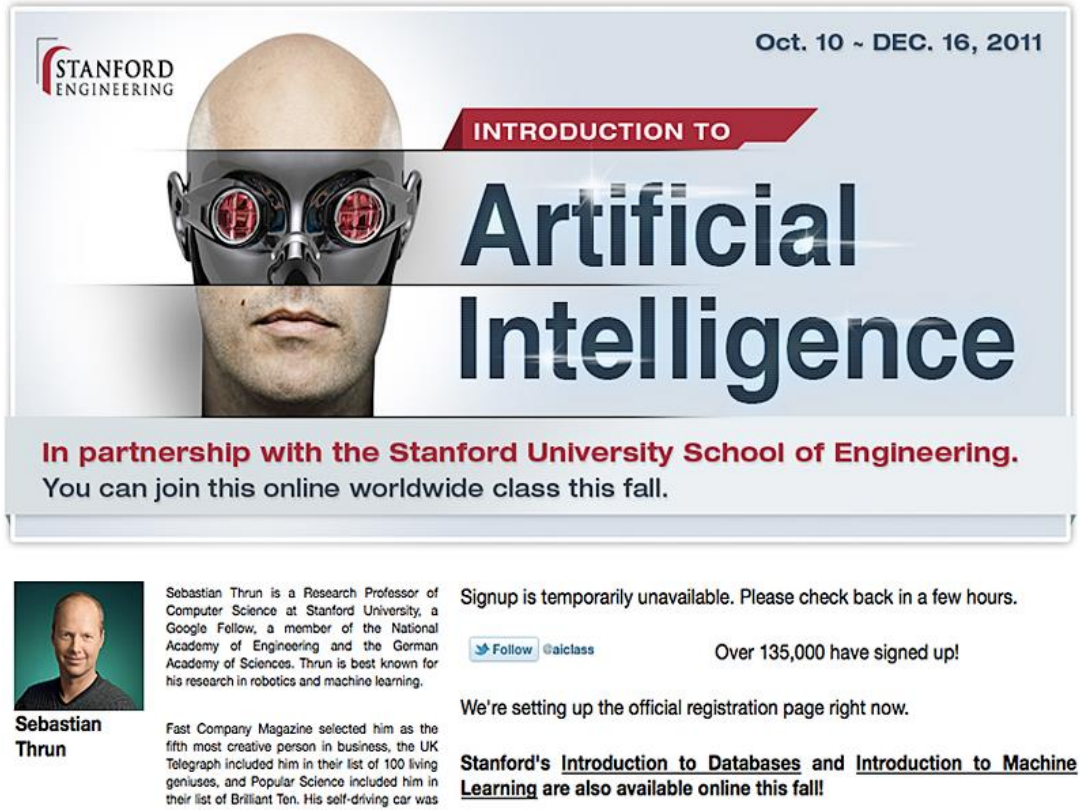
http://www.h2o.ai/gartner-magic-quadrant/

H2O.ai

Massive Open Online Course (MOOC)

My First MOOC Experience

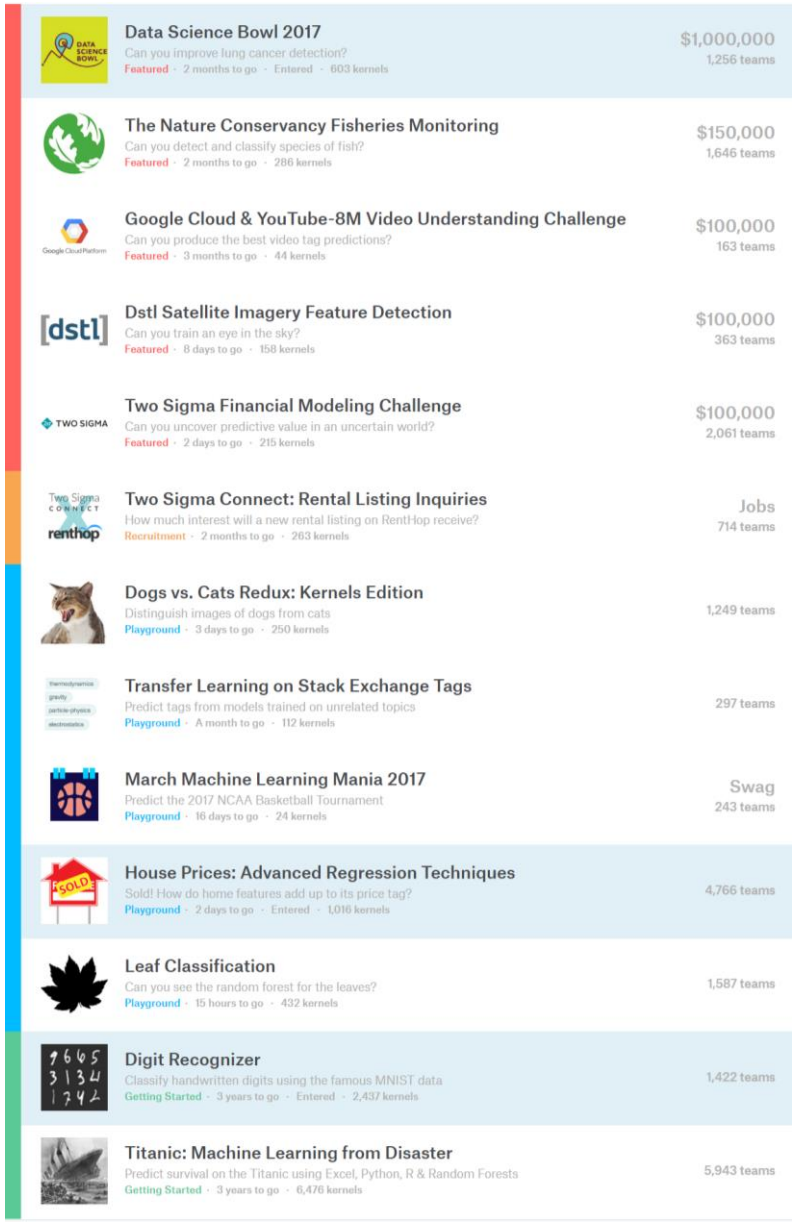
- Introduction to AI (2011)
  - One of the first MOOCs
- Key messages from Sebastian Thrun:
  - “Just dive into it.”
  - “Get your hands dirty.”
- Met new friends
  - Decided to collaborate for fun
  - “How about Kaggle?”
  - “What is Kaggle?”



H2O.ai

About Kaggle

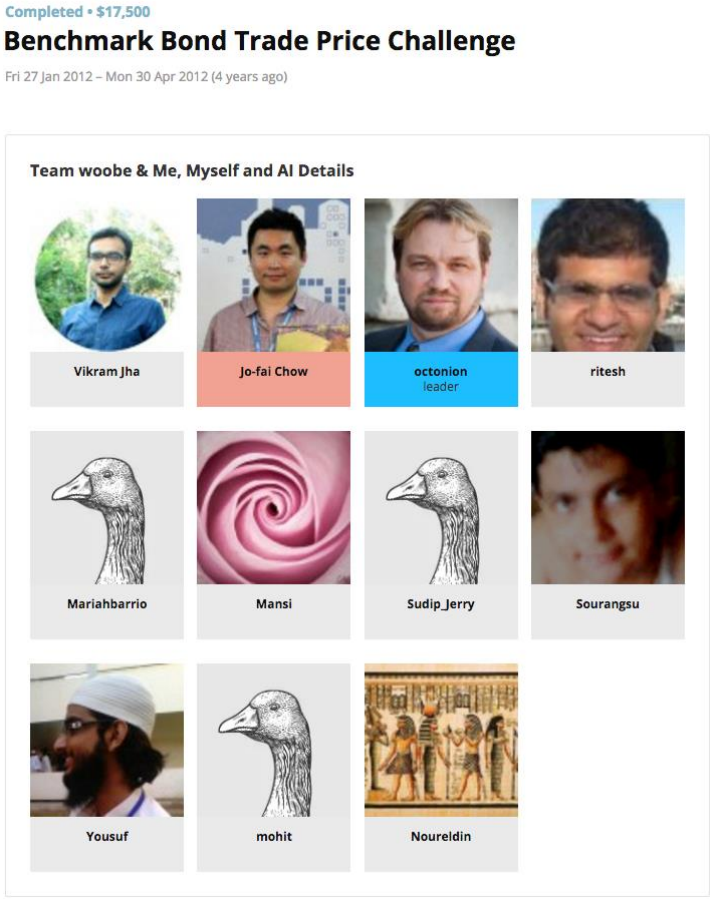
- World's biggest data mining competition platform
- Competition Types:
  - Featured (w/ Prize)
  - Recruitment
  - Playground
  - Beginner (101)



H2O.ai

My Very First Kaggle Experience

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



H2O.ai

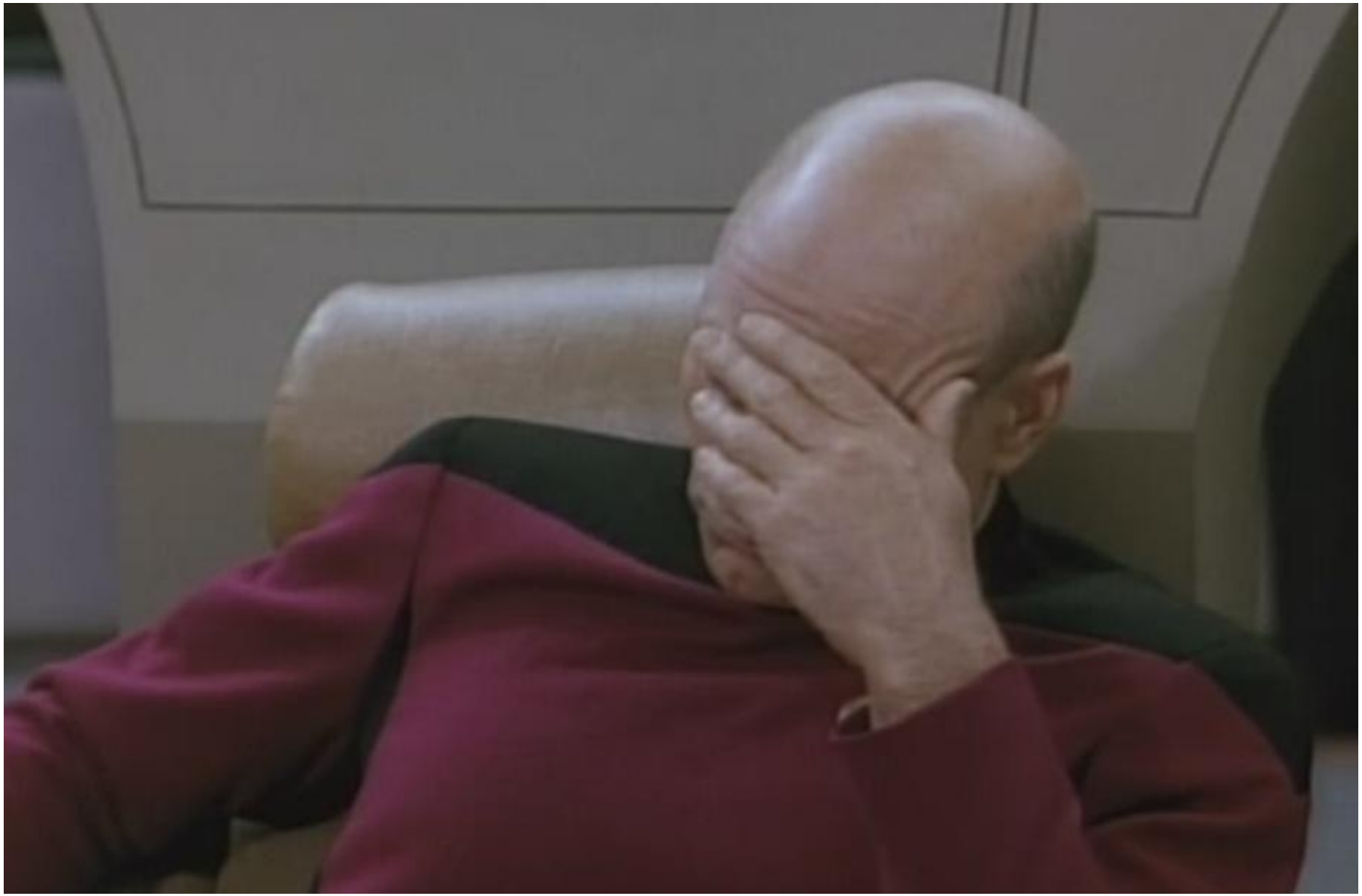
Teamwork

- Problems
  - “Hey Joe, you are a nice guy.”
  - “… but we can’t work together.”
  - “What … 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



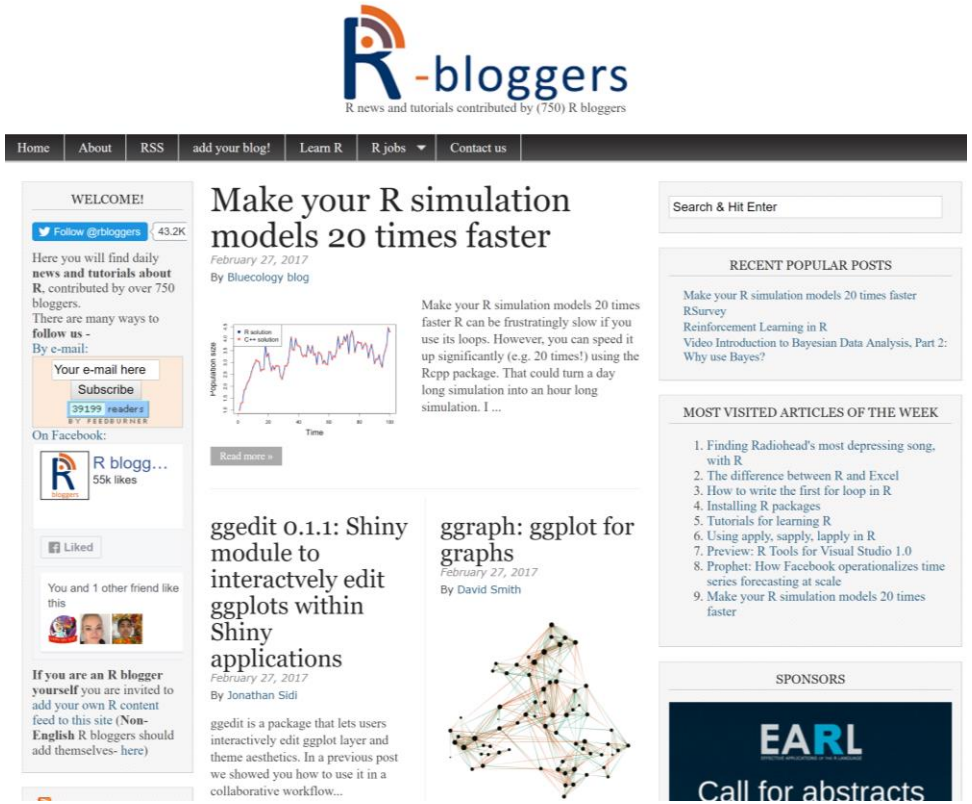
Adapt or Die

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?



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

H2O.ai

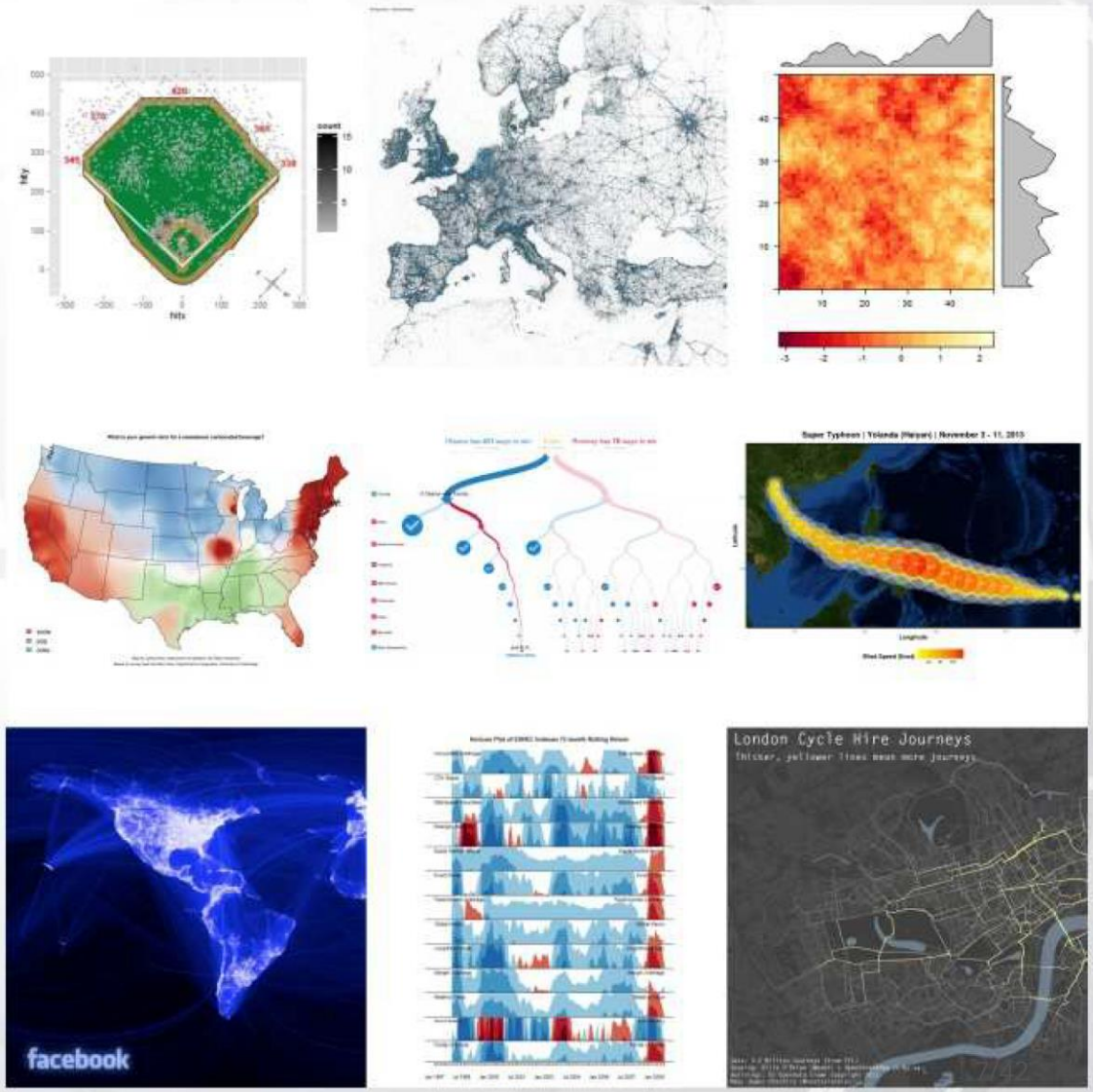
R Can Do That ?

... and more !?

Thanks to Tal Galili's

R-bloggers

R news and tutorials contributed by (452) R bloggers



http://www.jofaichow.co.uk/2014_03_11_LondonR/

Learn

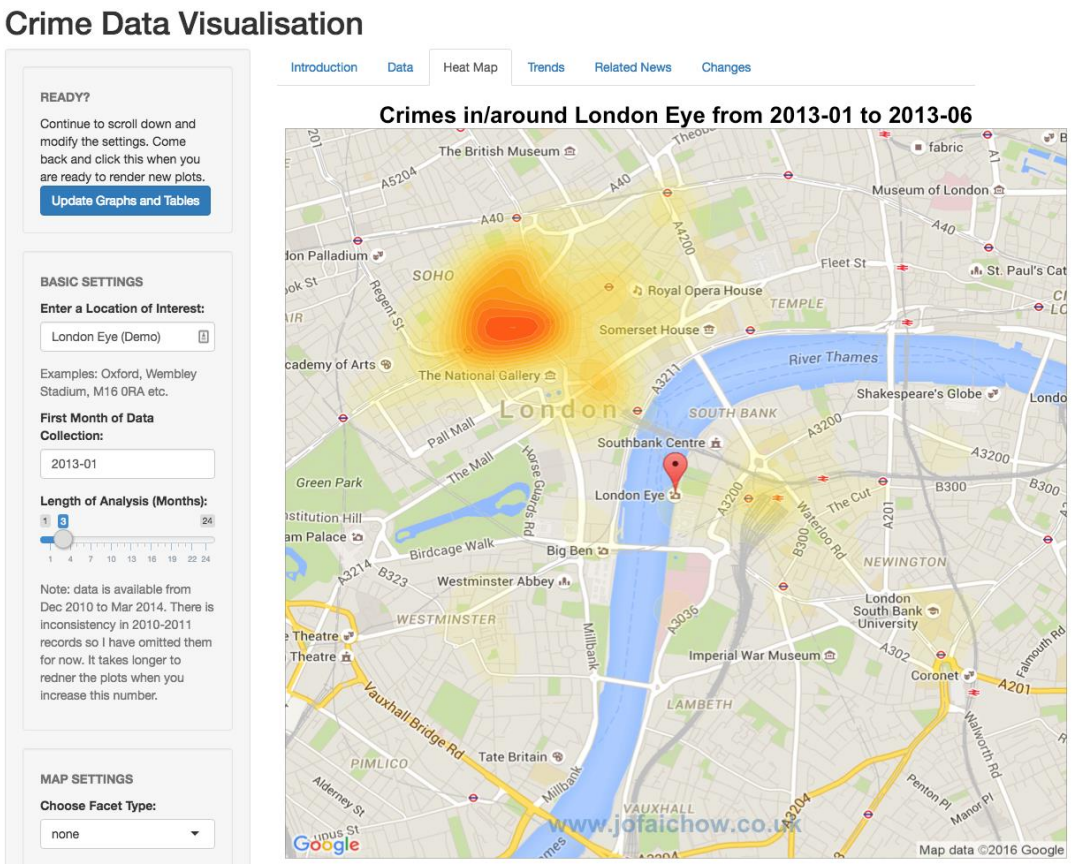
- More MOOCs
  - Machine Learning
    - Andrew Ng (Coursera)
    - MATLAB / Octave
  - Data Analysis
    - Jeff Leek (Coursera)
    - R
  - Intro to Programming
    - Dave Evans (Udacity)
    - Python

- Kaggle Forums
  - Tricks you can't learn from schools/books

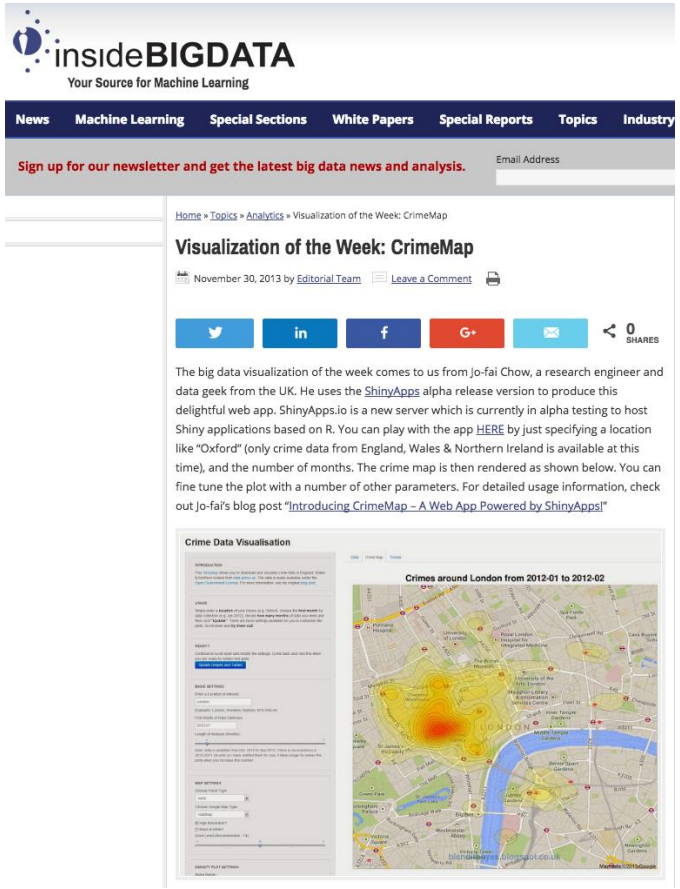
- Skills I also picked up
  - Linux – Ubuntu*
  - Git (I mean Git with GUI)
  - Cloud
  - HTML / CSS

**Ubuntu is an ancient African word that means "I can't configure Debian."*

Side Project #1 – Crime Data Visualization



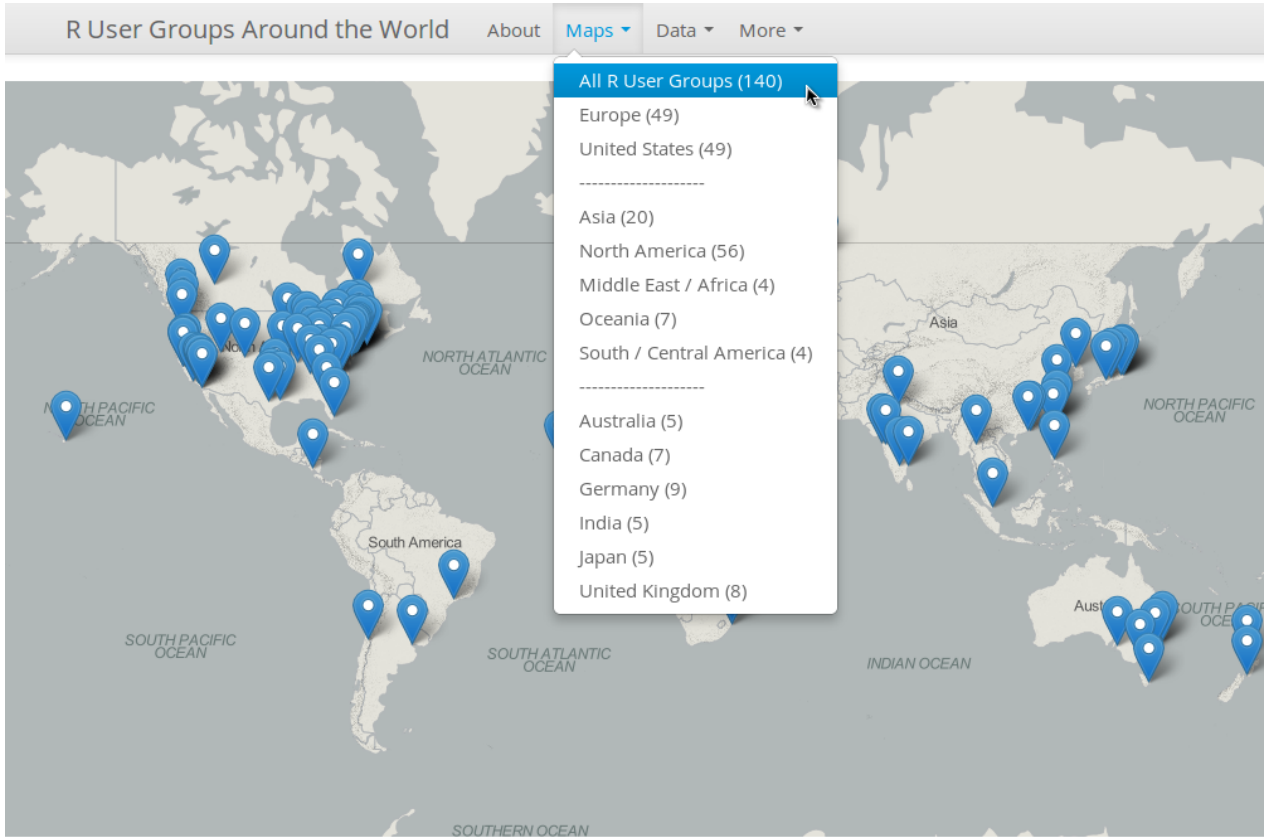
https://github.com/woobe/rApps/tree/master/crimemap

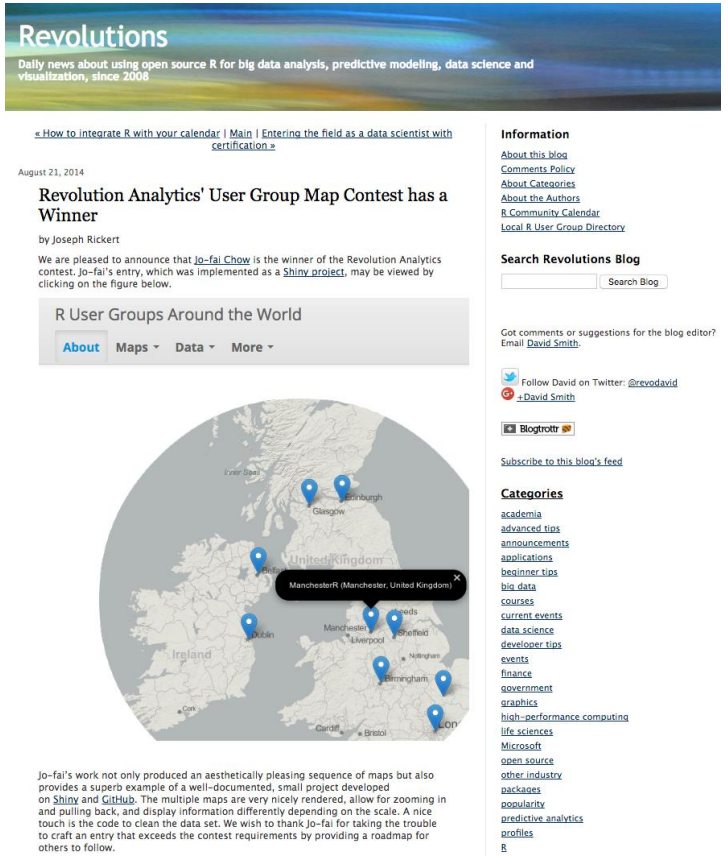


http://insidebigdata.com/2013/11/30/visualization-week-crimemap/

H2O.ai

Side Project #2 – Data Visualization Contest



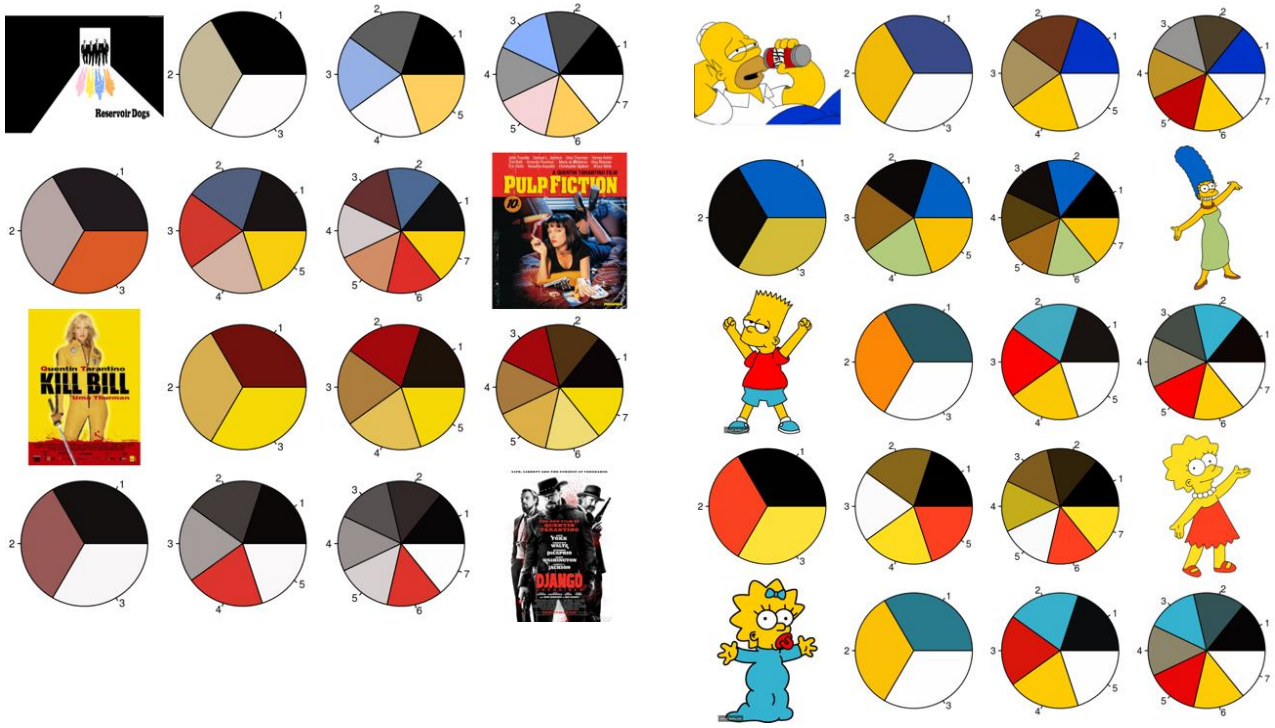


https://github.com/woobe/rugsmaps

http://blog.revolutionanalytics.com/2014/08/winner-for-revolution-analytics-user-group-map-contest.html

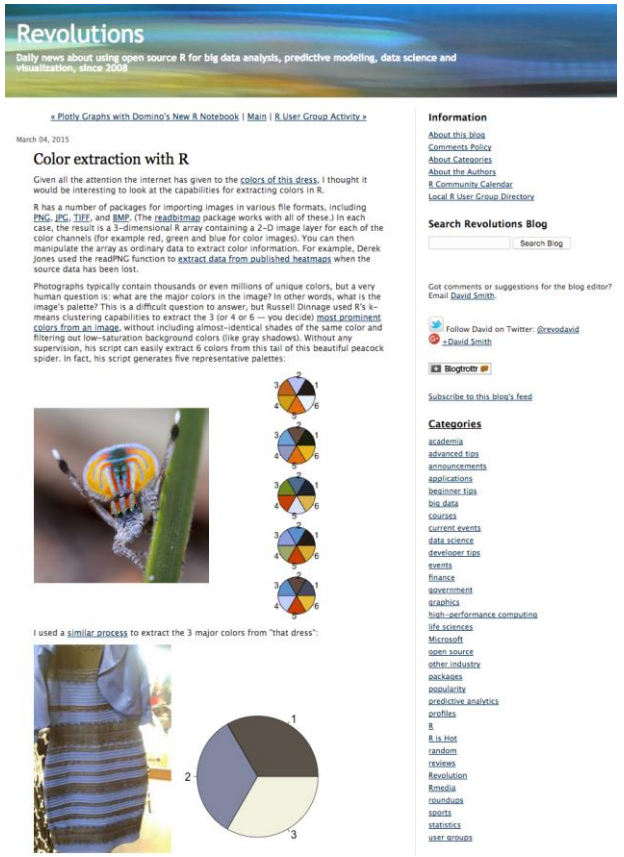
H2O.ai

Side Project #3 – Colors Extraction



https://github.com/woobe/rPlotter

#TheDress

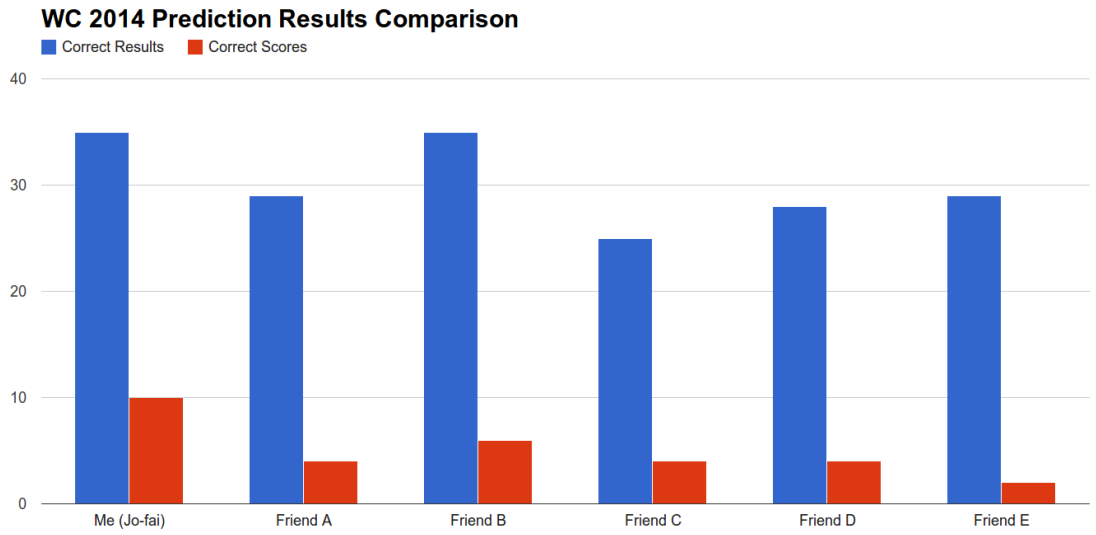


http://blog.revolutionanalytics.com/2015/03/color-extraction-with-r.html

H2O.ai

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

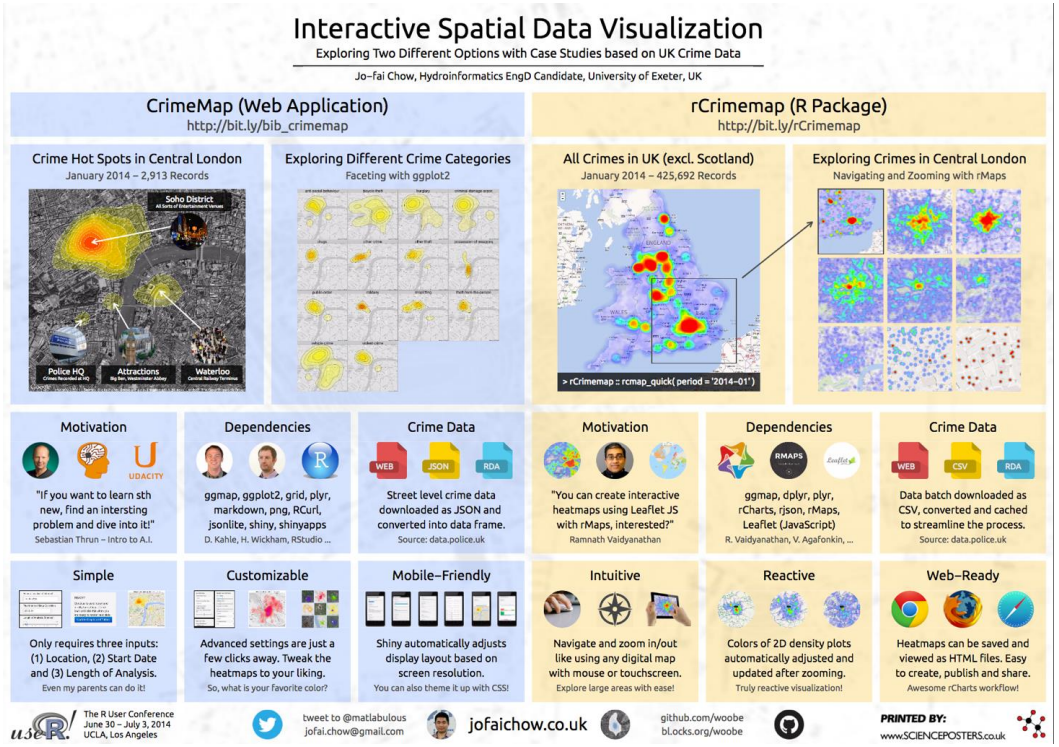
H2O.ai

Open Up Myself

New Opportunities

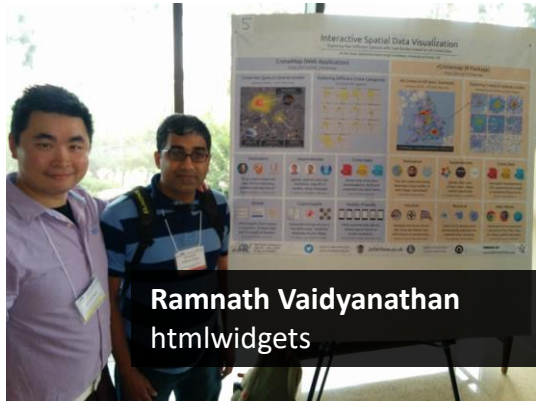
R Community, H2O & Domino Data Lab

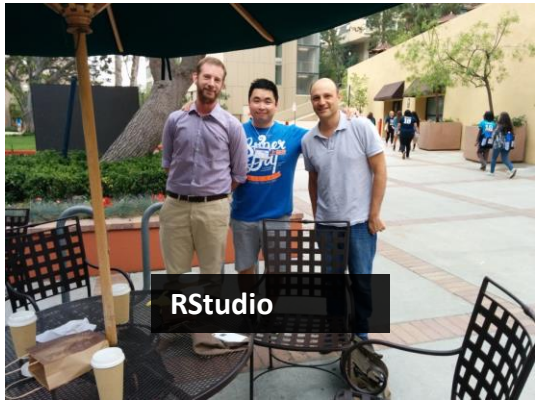
LondonR 2013 & useR! 2014

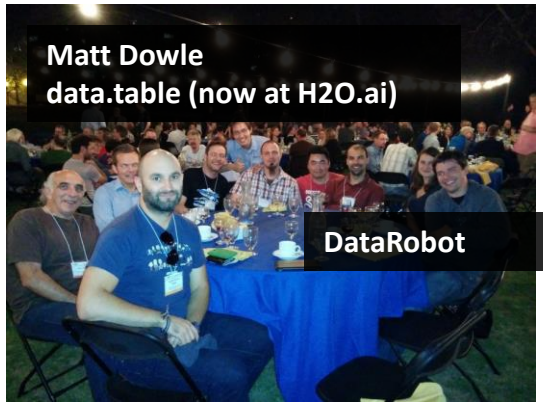


H2O.ai

useR! 2014







H2O.ai

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

by Nick Elprin on September 19th, 2014

SHARE

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 * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
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.

H2O.ai

Kaggle – The Pain & The Fear

513	▼ 449	Jo-fai Chow @ blenditbayes! + H2...		0.51401	133	2y

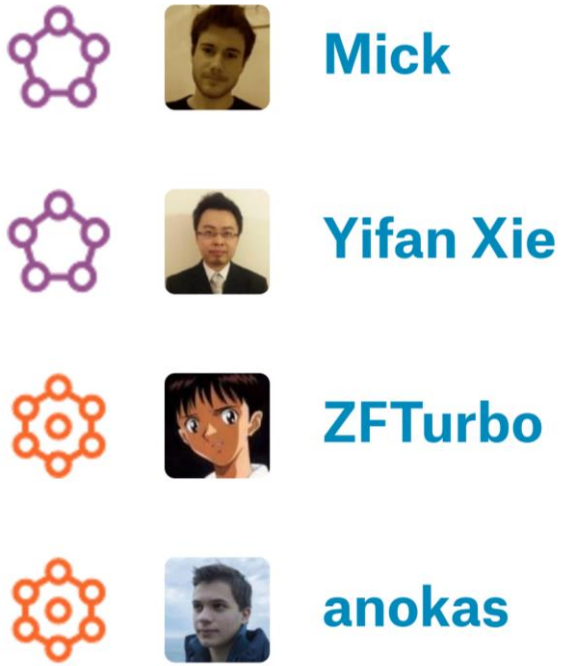
H2O.ai

Kaggle – The Gain

- New Skills
  - Exploratory Data Analysis
  - Machine algorithms
  - Feature engineering
  - Model stacking
  - Communication

  - **THE FEAR OF OVERFITTING!**

- New Friends
  - London Kaggle Meetup



H2O.ai

Life as a Data Scientist

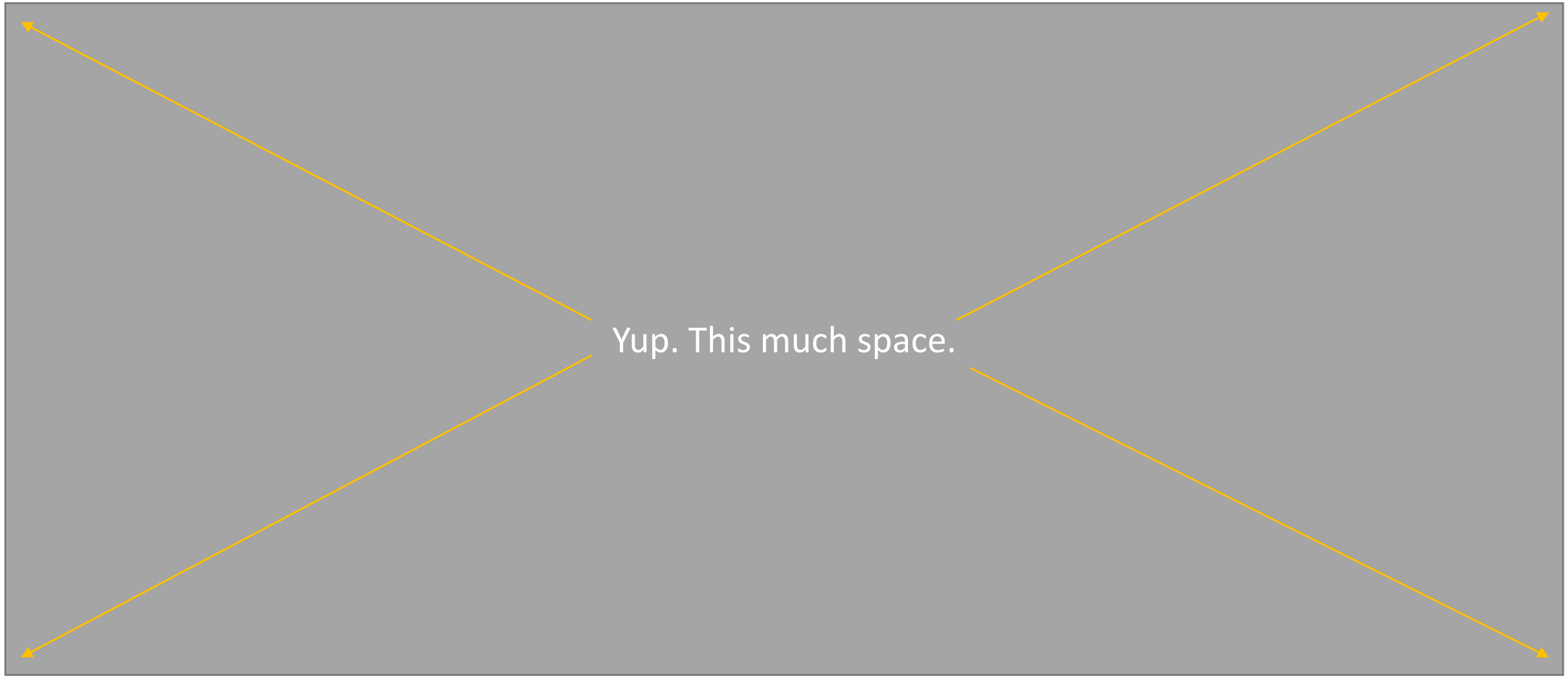
Toy (In-Class) vs. Kaggle vs. Real-World Data

H2O.ai

Story Telling

H2O.ai

Story Telling with One Single Slide



H2O.ai

Using H_2O 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

Competitions Expert

Current Rank	Highest Rank
818 of 54,344	181

0	2	5

Santander Product Recom... · 2 months ago · Top 3%	38th of 1787
Rossmann Store Sales · a year ago · Top 2%	57th of 3303
Personality Prediction Base... · 5 years ago · Top 24%	21st 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

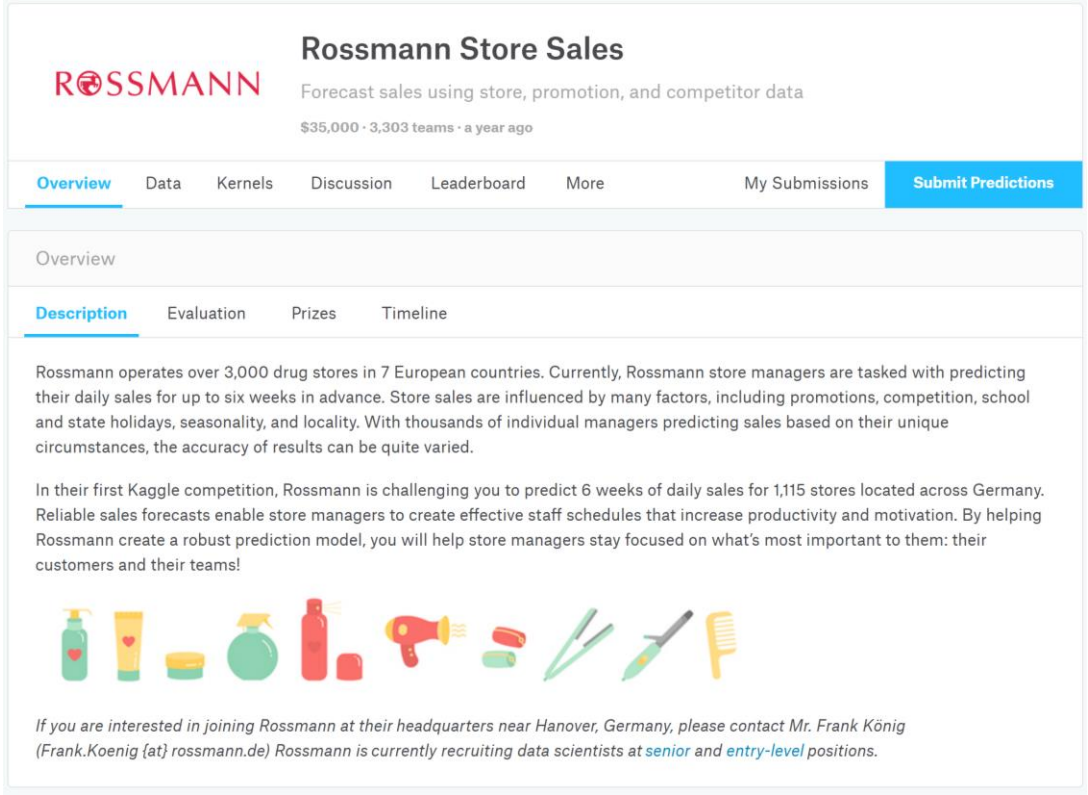
H2O Deep Learning Starte... · 2 years ago	21 votes
Best Ensemble References? · 2 years ago	7 votes
Leaderboard scores · 2 years ago	4 votes

Bio

Edit

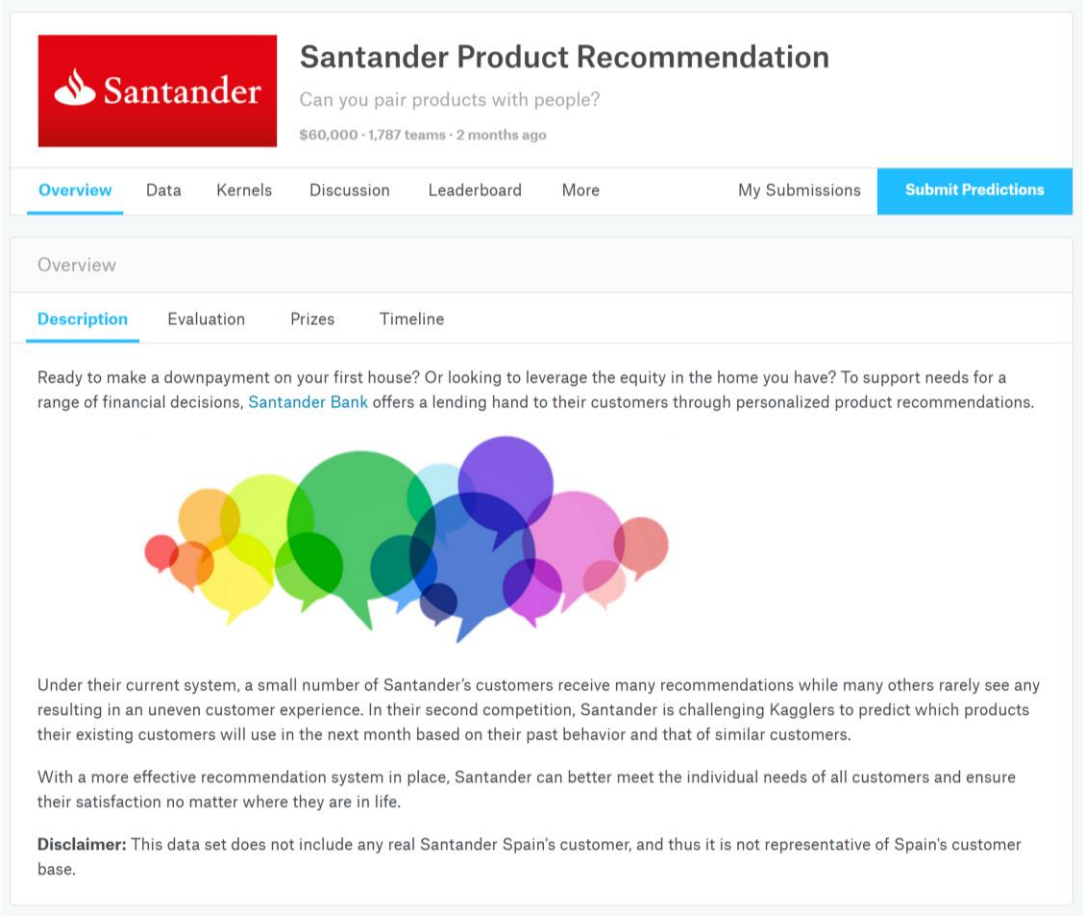
Rossmann Store Sales

- 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_2O Deep Learning
  - Xgboost
  - Manual process (life before h2oEnsemble / Stacked Ensembles in H_2O)



Santander Product Recommendation

- Reframed as a Multiclass Classification problem
- Feature Engineering
  - Basic (Everyone)
  - Advanced (ZFTurbo, Yifan, Anokas)
  - Also see Yifan's slides
- Models
  - xgboost (ZFTurbo)
  - H_2O GBM (Joe) – Single Best
  - Simple averaging only this time



kaggle-santander

File Edit View Insert Format Data Tools Add-ons Help Last edit was on 20 December 2016

woobe1208@gmail.com

Comments Share

Calibri 10

joe_gbm_038

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	run	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)
2																train	valid	train	valid	train	valid		
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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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_purchaised_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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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/mpcerr	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 | variable_importance | h2o_gbm_grid | features_selection | fea_eval_stage_1 | joe_data_prep | features | old_test_ignore_this

Sum: May-2017

Explore

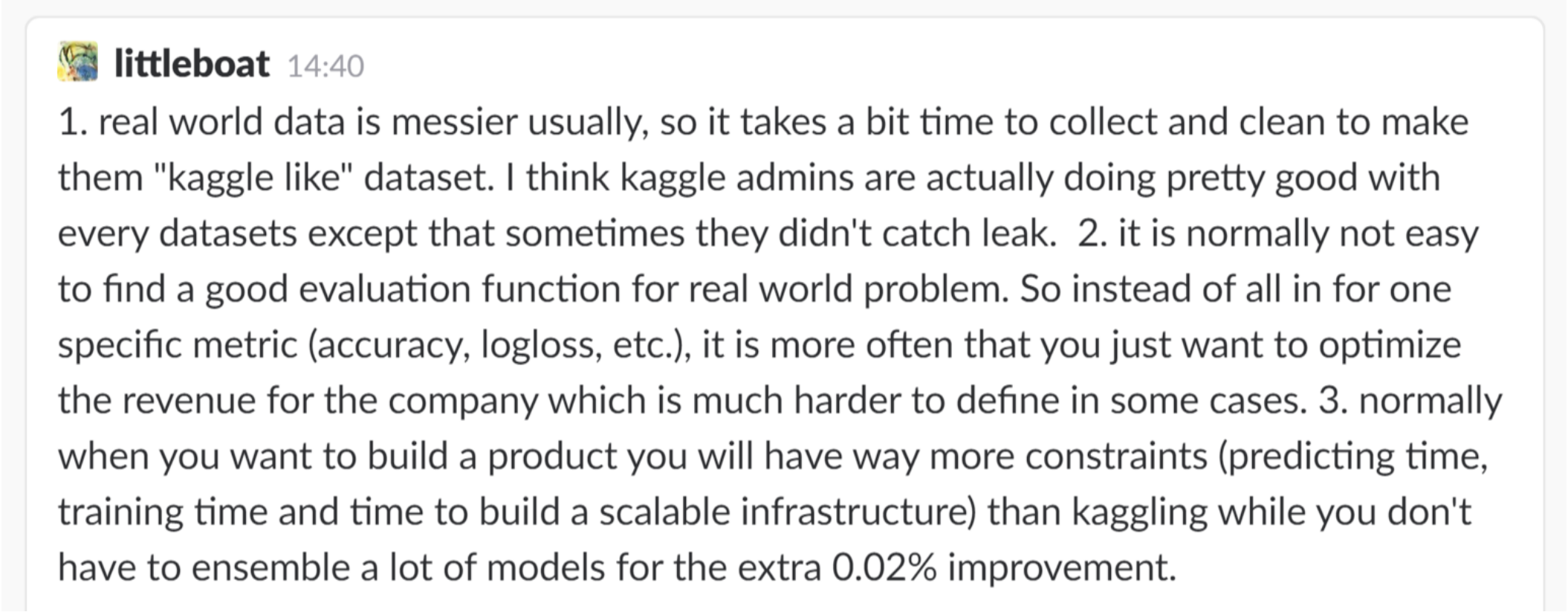
Conclusions

New Skills, New Friends & New Opportunities

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

H2O.ai

Differences between Kaggle & Data Science



From Littleboat's AMA on Kagglenoobs Slack Channel

Thanks!

- People who have helped me along the way
  - Kaggle Friends
  - H_2O.ai
  - Domino Data Lab
  - Mango Solutions
- Slides
  - bit.ly/h2o_meetups
- Contact
  - joe@h2o.ai
  - @matlabulous
  - github.com/woobe

H_2O.ai

Making Machine Learning Accessible to Everyone

Photo credit: Virgin Media