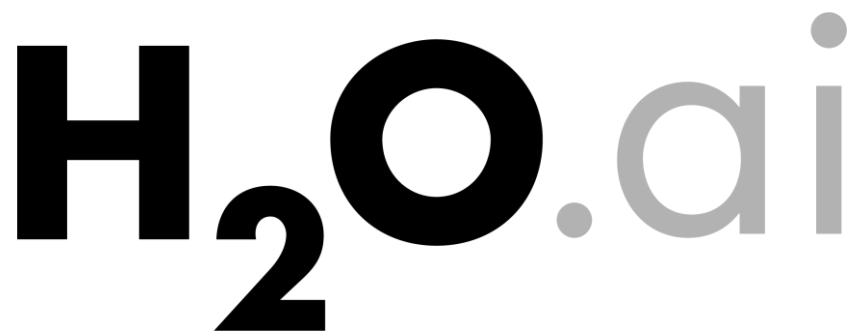


From Kaggle to H₂O

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 H₂O in 2014**

- Data Scientist
2015
 - Virgin Media (UK)
 - Domino Data Lab (Silicon Valley)
- 2016**
 - H₂O.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 H₂O & Domino
- To Kaggle, or not to Kaggle
 - Joy, Pain, Fear, Gain ...
 - ... and New Friends
- Life as a Data Scientist
- Using H₂O 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 a promotional graphic for a MOOC. At the top left is the Stanford Engineering logo. To the right is the text "Oct. 10 ~ DEC. 16, 2011". In the center is a photograph of a man's face, partially obscured by a metallic, futuristic-looking visor or mask. To the right of the image, the text "INTRODUCTION TO Artificial Intelligence" is displayed 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 right, there is a "Follow" button with the handle "@aiclass" and the text "Over 135,000 have signed up! We're setting up the official registration page right now."



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)

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

Jo-fai (Joe) Chow
@matlabulous

- Problems

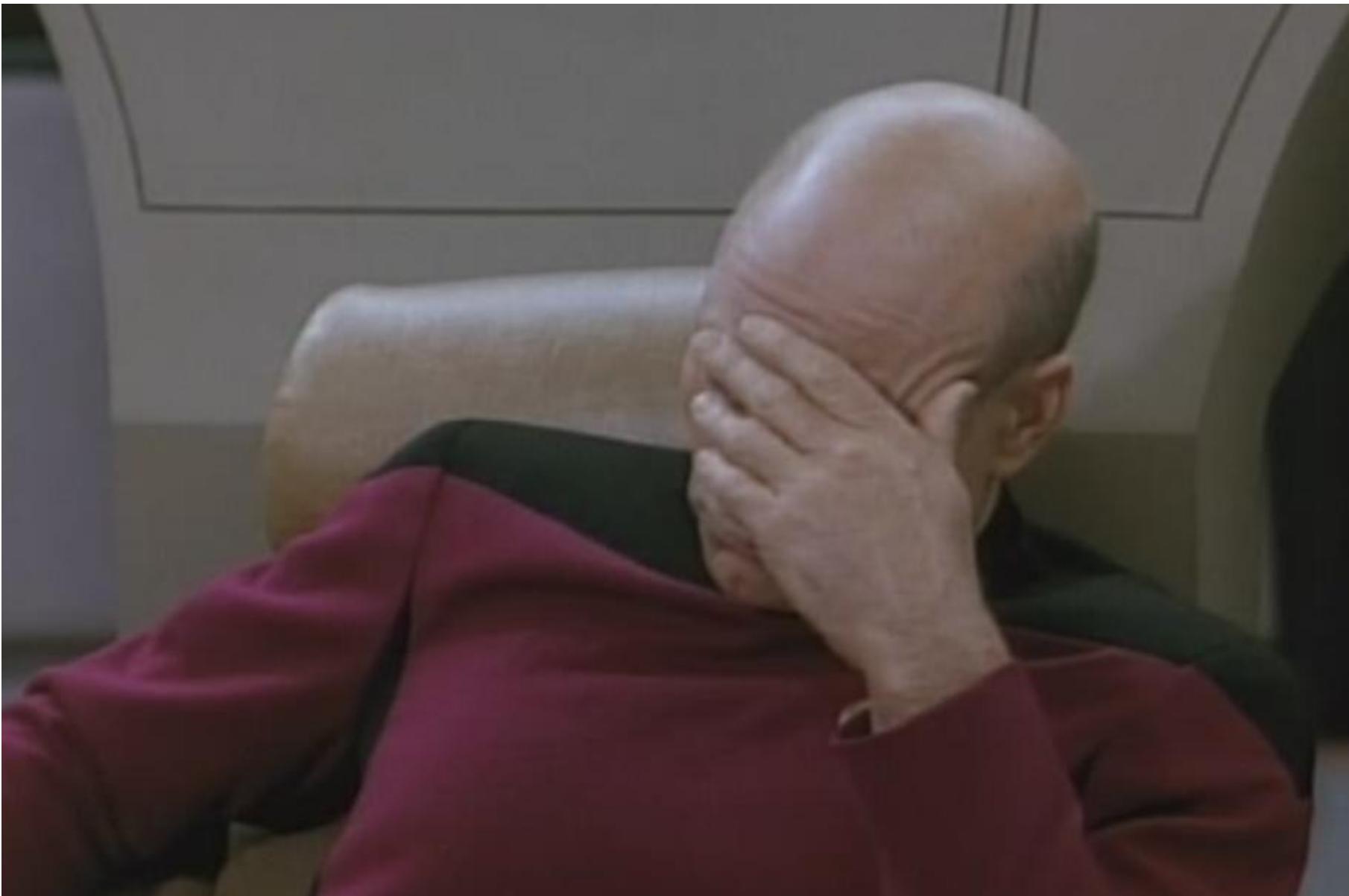
- “We prefer open source tools like R or Python.”
- “Wait ... you guys can use Octave”
- “Thanks, but no thanks ...”

- 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?

The screenshot shows the homepage of R-bloggers.com. The header features the "R-bloggers" logo with a blue "R" icon and the text "R news and tutorials contributed by (750) R bloggers". Below the header is a navigation bar with links for Home, About, RSS, add your blog!, Learn R, R jobs, and Contact us. The main content area has a "WELCOME!" section with social media links (Twitter, Facebook, LinkedIn), a "Follow @rbloggers" button, and a "292499 readers BY FEEDBURNER" counter. A featured article titled "Make your R simulation models 20 times faster" by Blueecology blog is displayed, showing a line graph of population over time with three different solutions. To the right, there's a search bar, a "RECENT POPULAR POSTS" sidebar with links to articles like "Make your R simulation models 20 times faster" and "Reinforcement Learning in R", and a "MOST VISITED ARTICLES OF THE WEEK" sidebar with a list of top articles. 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 ?

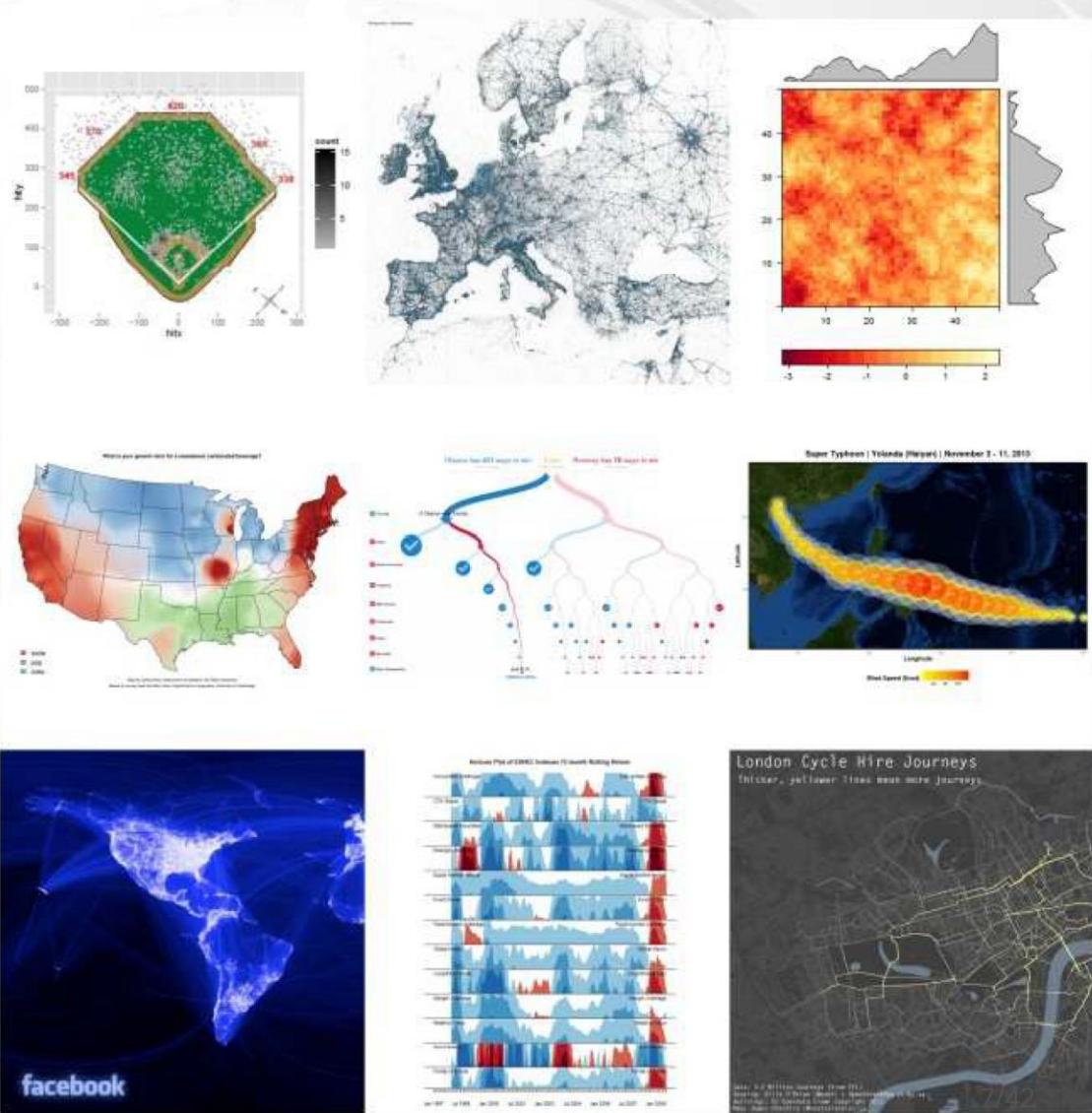


... and more !?

Thanks to Tal Galili's

R-bloggers

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



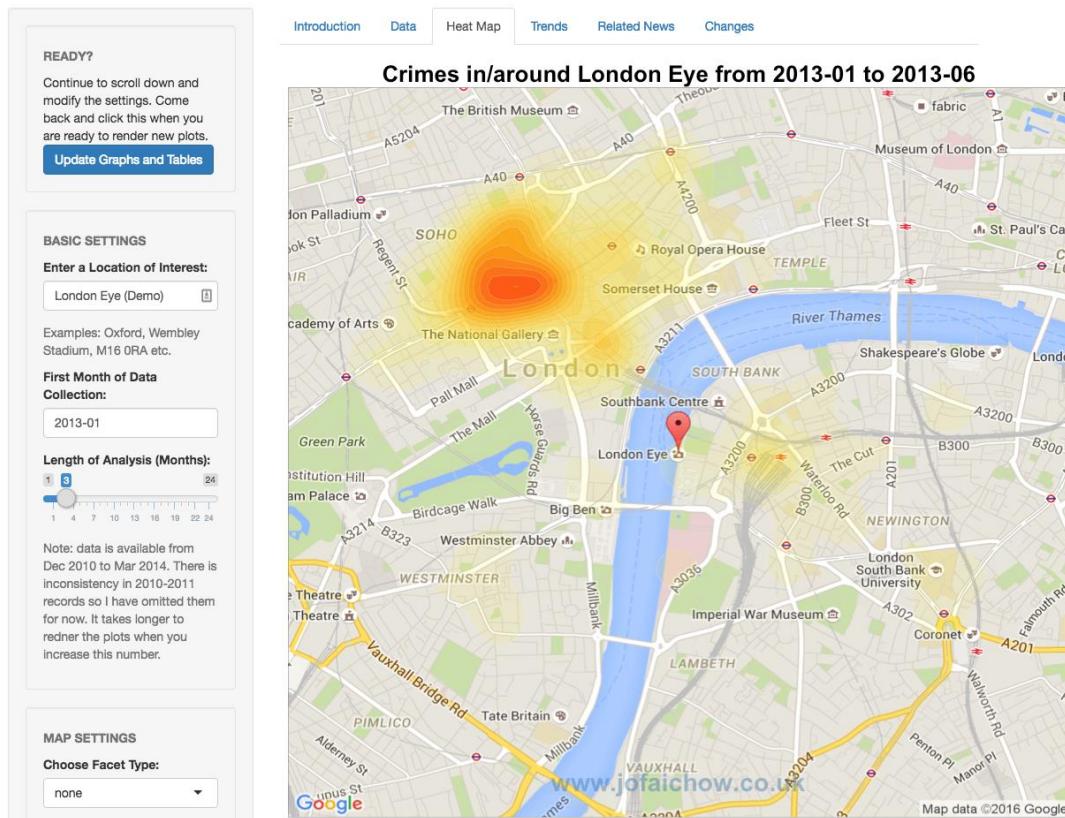
Learn

- 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

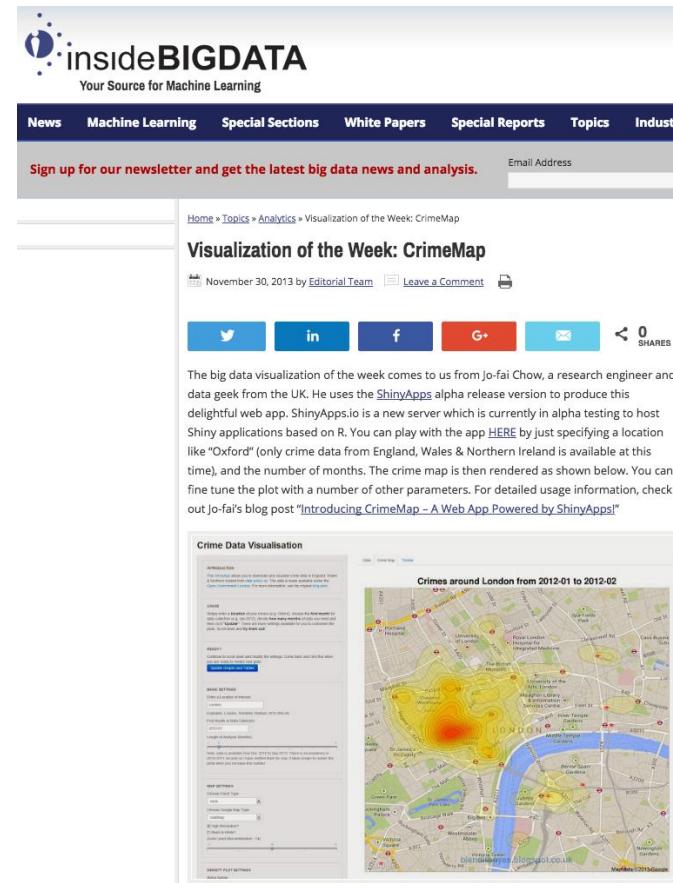
**Ubuntu is an ancient African word that means “I can’t configure Debian.”*

Side Project #1 – Crime Data Visualization

Crime Data Visualisation

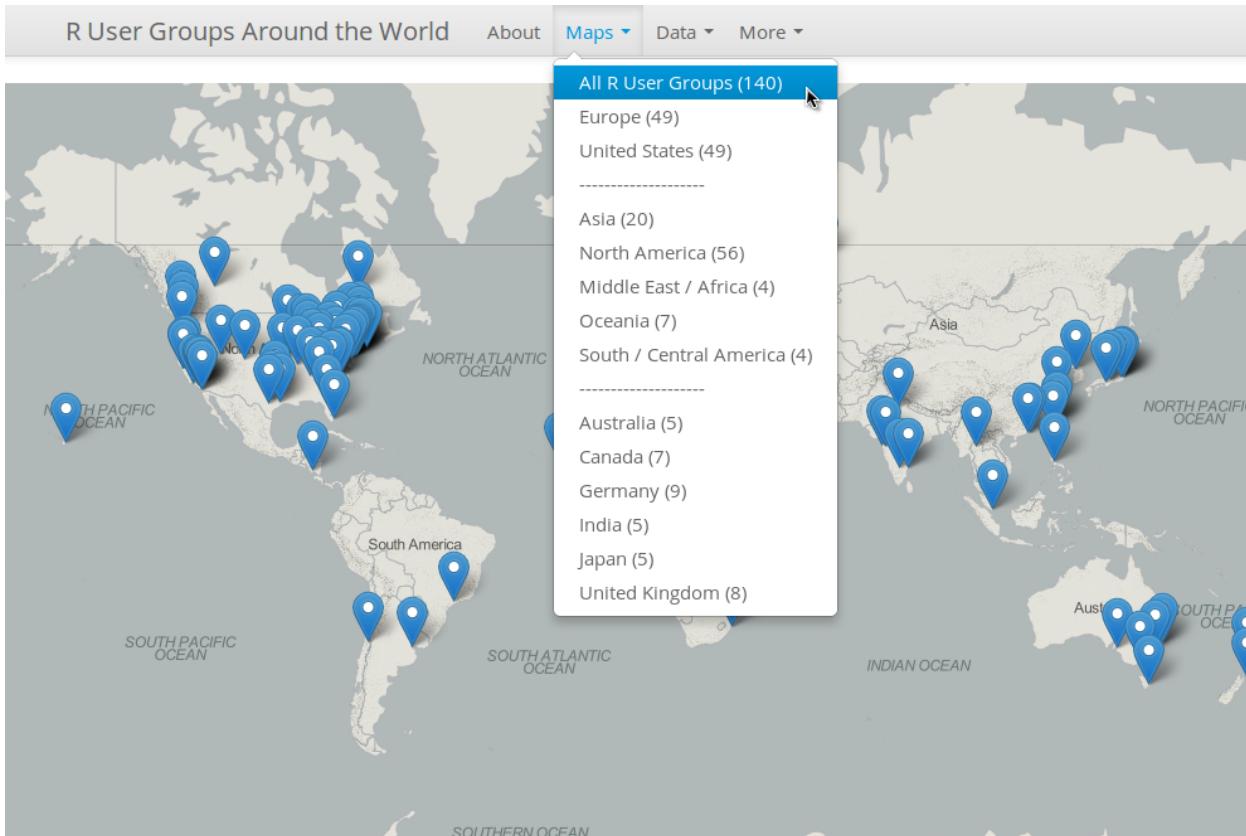


<https://github.com/woobe/rApps/tree/master/crimemap>

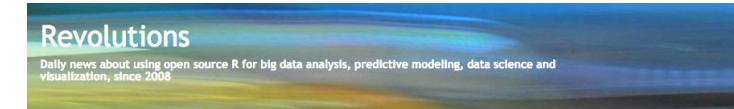


<http://insidebigdata.com/2013/11/30/visualization-week-crimemap/>

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](#)

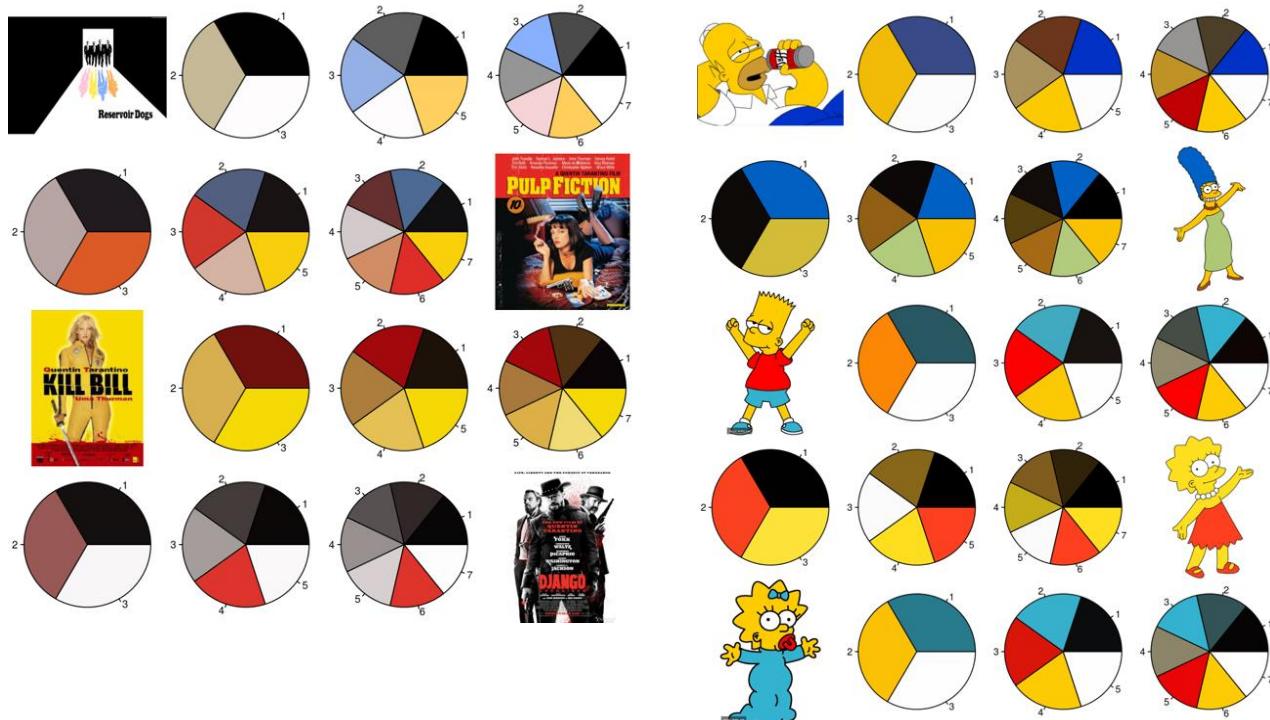
[Subscribe to this blog's feed](#)

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

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

Side Project #3 – Colors 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

[» Plenty Graphs with Domino's New R Notebook](#) | [Main](#) | [R User Group Activity](#)

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 `gd`. (`TIFF` is the `readTIFF` package which will do all of this.) 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 extraction 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 perceptually similar 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":

The color palette for the dress is approximately: 1. Dark Blue (top), 2. Light Blue (middle), 3. White (bottom).

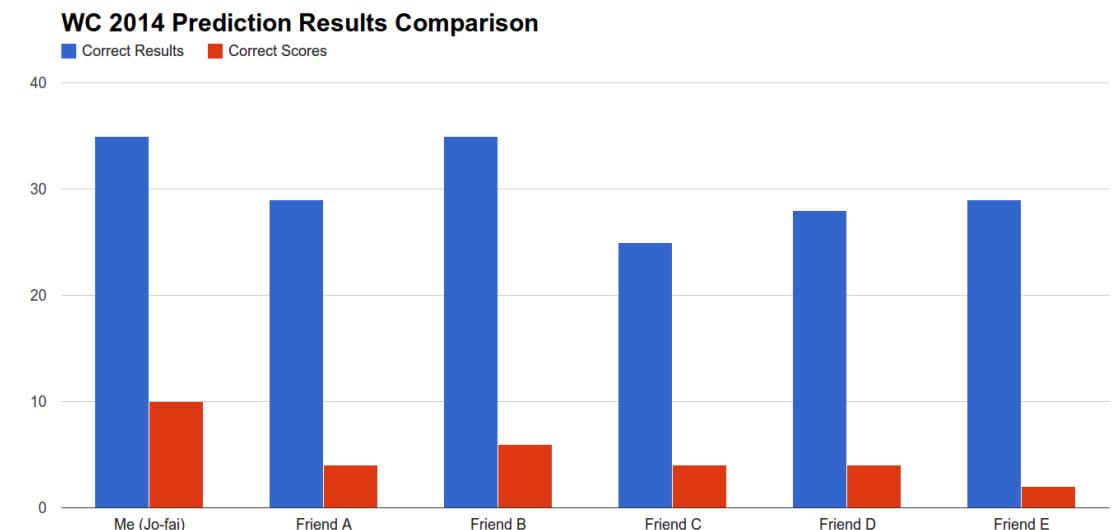
<http://blog.revolutionanalytics.com/2015/03/color-extraction-with-r.html>

18

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>

Open Up Myself



New Opportunities

R Community, H2O & Domino Data Lab

LondonR 2013 & useR! 2014

Introducing 'CrimeMap'
Using Shiny and ShinyApps for Quick Web App Development

Agenda

- 1 Why R?
- 2 Introducing 'CrimeMap'
- 3 What's Next?
- 4 Beyond Data and Maps
- 5 Conclusions
- 6 Q & A

R Can Do That ?

CrimeMap - Application Examples (Map Type)

CrimeMap - Control Freak !!!

What's Your Favourite?

Introducing Package 'rCrimemap'

Function 'rcmap()'

```
require(rCrimemap); install.packages("dplyr", "json")
install.github("ramnathv/rCrimemap")
install.github("ramnathv/rmaps")
```

Latest RStudio IDE (v0.98.501+)

Install 'rCrimemap'

```
install.github("wooble/rCrimemap")
```

Introducing Package 'Crimemap' - LondonR Demo

```
require(Crimemap); rmap("Ball Brothers ECR 199P");
# [rCrimemap] Location = "Ball Brothers ECR 199P"
# [rCrimemap] period = "2010-12-01", "2011-01-01"
# [rCrimemap] type = "ALL", "Burglary", "Theft", "Assault", "Sexual Offense", "Robbery", "Murder", "Kidnapping", "Arson", "Homicide", "Suspicious Occurrence", "Fraud", "Vandalism", "Shoplifting", "Petrol Theft", "Other Crime", "Holding for Trial", "Police Incident"
# [rCrimemap] provider = "Nokia.northeast", zoom = 10
# [rCrimemap] 
```

Codes

<http://bit.ly/rCrimemap>

Introducing Package 'Crimemap' - LondonR Demo

```
rcmap("Ball Brothers ECR 199P",
      "2010-12-01", "2011-01-01", c(1000, 1000),
      "Nokia.northeast")
```

**rcmap("Manchester",
 "2014-01-01", "2014-01-01", c(1000, 1000),
 "ReverbOpen-JP")**

Interactive Spatial Data Visualization

Exploring Two Different Options with Case Studies based on UK Crime Data
Jo-fai Chow, Hydroinformatics EngD Candidate, University of Exeter, UK

CrimeMap (Web Application) http://bit.ly/bib_crimemap	rCrimemap (R Package) http://bit.ly/rCrimemap
Crime Hot Spots in Central London January 2014 – 2,913 Records	All Crimes in UK (excl. Scotland) January 2014 – 425,692 Records
Motivation "If you want to learn sth new, find an interesting problem and dive into it!" Sebastian Thrun – Intro to A.I.	Motivation "You can create interactive heatmaps using Leaflet JS with rMaps, interested?" Ramnath Vaidyanathan
Dependencies ggmap, ggplot2, grid, plyr, markdown, png, RCurl, jsonlite, shiny, shinyapps, D. Kahle, H. Wickham, RStudio ...	Dependencies ggmap, dplyr, plyr, rCharts, json, rMaps, Leaflet JavaScript, Ramnath Vaidyanathan, V. Agafonkin, ...
Crime Data Street level crime data downloaded as JSON and converted into data frame. Source: data.police.uk	Crime Data Data batch downloaded as CSV, converted and cached to streamline the process. Source: data.police.uk
Simple Only requires three inputs: (1) Location, (2) Start Date and (3) Length of Analysis. Even my parents can do it!	Customizable Advanced settings are just a few clicks away. Tweak the heatmaps to your liking. So, what is your favorite color?
Mobile-Friendly Shiny automatically adjusts display layout based on screen resolution. You can also theme it up with CSS!	Intuitive Navigate and zoom in/out like using any digital map with mouse or touchscreen. Explore large areas with ease!
Reactive Colors of 2D density plots automatically adjusted and updated after zooming. Truly reactive visualization!	Web-Ready Heatmaps can be saved and viewed as HTML files. Easy to create, publish and share. Awesome Charts workflow!

PRINTED BY:
www.SCIENCEPOSTERS.co.uk

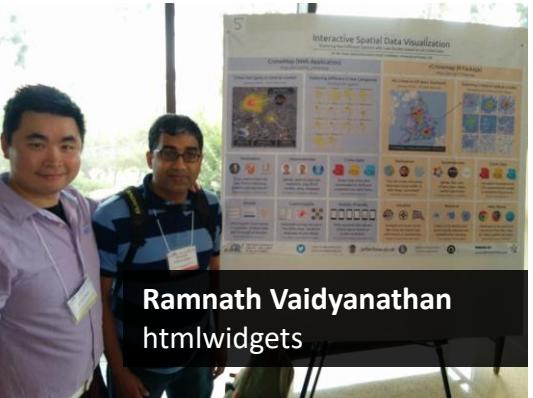
The 20th Conference July 30 – July 3, 2014 UCLA, Los Angeles

tweet to @matlabulous jo.fai.chow@gmail.com

jofaichow.co.uk

github.com/woobleblocks.org/wooble

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

The screenshot shows a Kaggle competition page for the "Africa Soil Property Prediction Challenge". The page features a dark header with the Kaggle logo and navigation links for Customer Solutions, Competitions, and Community. Below the header, there's a banner for the competition with a map of Africa and the text "\$8,000 • 413 teams". The main content area displays the challenge details: "Wed 27 Aug 2014" to "Tue 21 Oct 2014 (40 days to go)". A progress bar indicates the current date. The "Leaderboard - Africa Soil Property Prediction Challenge" section shows a table with one entry:

#	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)

A blue banner at the bottom of the leaderboards section reads "Your Best Entry" and "Number One!". It also states: "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 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

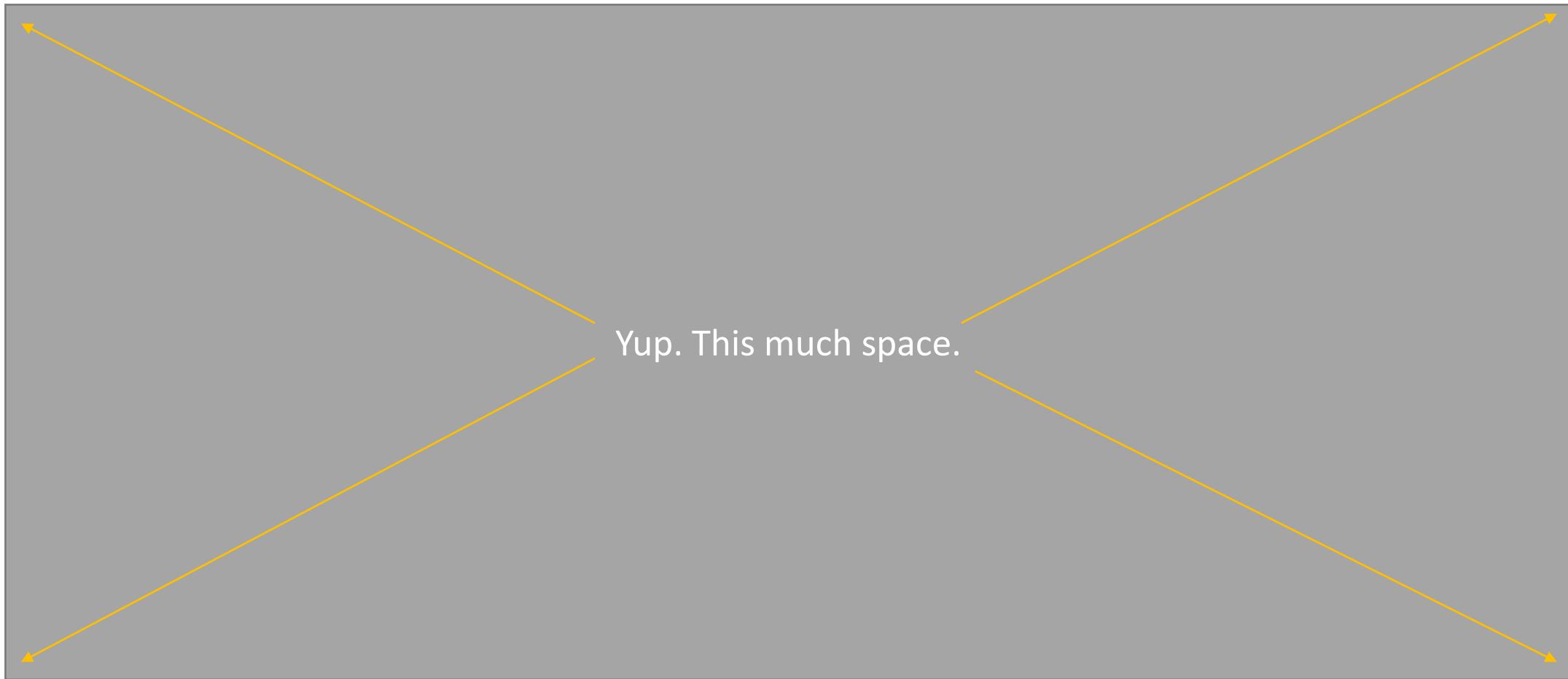


Story Telling



AS THE YEARS WENT BY

Story Telling with One Single Slide



Using H₂O 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...](#)

38th
of 1787

[Rossmann Store Sales](#)

57th
of 3303

[Personality Prediction Base...](#)

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)
 - H₂O Deep Learning
 - Xgboost
 - Manual process (life before h2oEnsemble / Stacked Ensembles in H₂O)

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](#)

Overview

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.

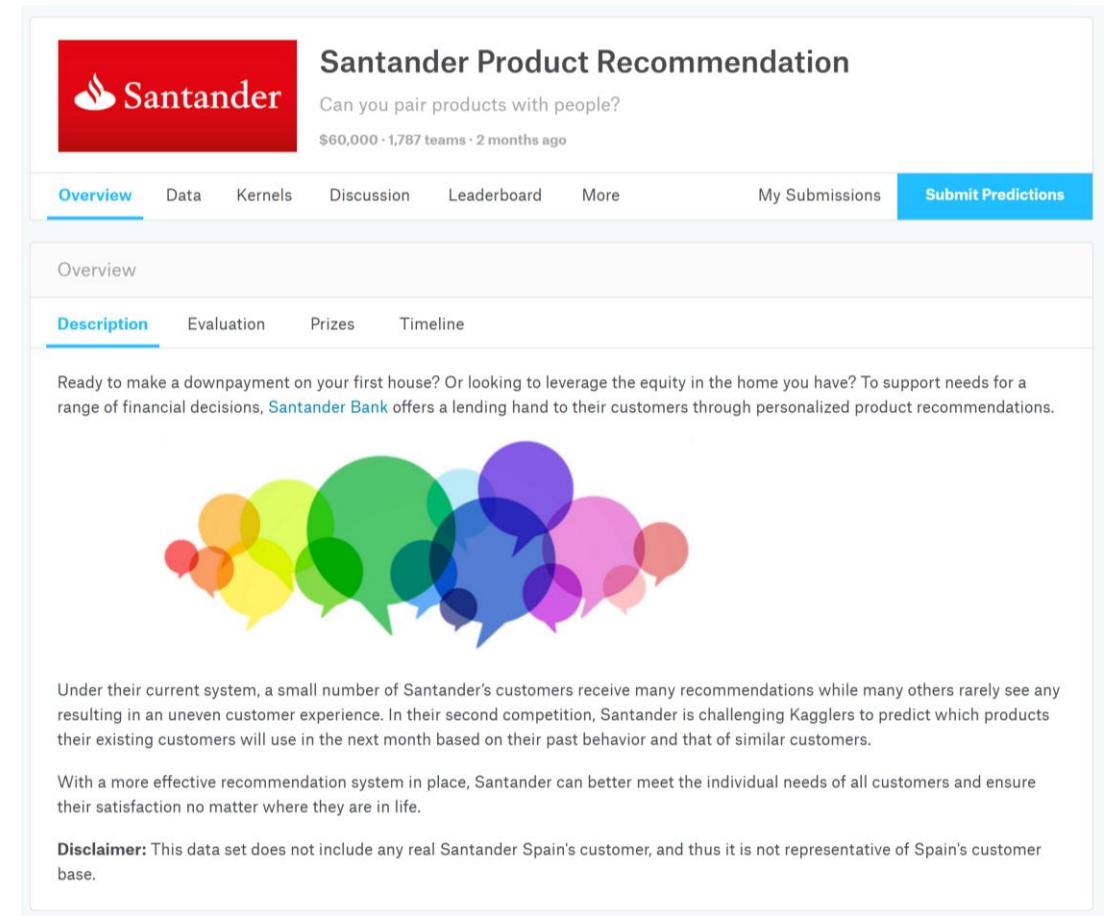
In their first Kaggle competition, Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction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



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.

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)
 - H₂O GBM (Joe) – Single Best
 - Simple averaging only this time



The screenshot shows the Kaggle competition page for the Santander Product Recommendation challenge. At the top, there's a red header with the Santander logo. Below it, the title "Santander Product Recommendation" and a subtitle "Can you pair products with people?". It also shows statistics: "\$60,000 · 1,787 teams · 2 months ago". A navigation bar includes "Overview" (which is underlined), "Data", "Kernels", "Discussion", "Leaderboard", "More", "My Submissions", and a blue "Submit Predictions" button. The main content area has tabs for "Description" (underlined), "Evaluation", "Prizes", and "Timeline". The "Description" tab contains text about the competition: "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." Below this is a graphic of colorful speech bubbles. The "Evaluation" tab contains text: "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." The "Prizes" tab contains text: "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." The "Disclaimer" at the bottom states: "Disclaimer: This data set does not include any real Santander Spain's customer, and thus it is not representative of Spain's customer base."

Conclusions

New Skills, New Friends & New Opportunities



Giphy is your friend when you don't enough have 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 kaggle while you don't have to ensemble a lot of models for the extra 0.02% improvement.

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

Thanks!

- People who have helped me along the way
 - Kaggle Friends
 - H₂O.ai
 - Domino Data Lab
 - Mango Solutions
- Slides
 - bit.ly/h2o_meetups
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe



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

H₂O.ai

Making Machine Learning
Accessible to Everyone

Photo credit: Virgin Media