

Kaggle Competitions, New Friends, New Skills and New Opportunities



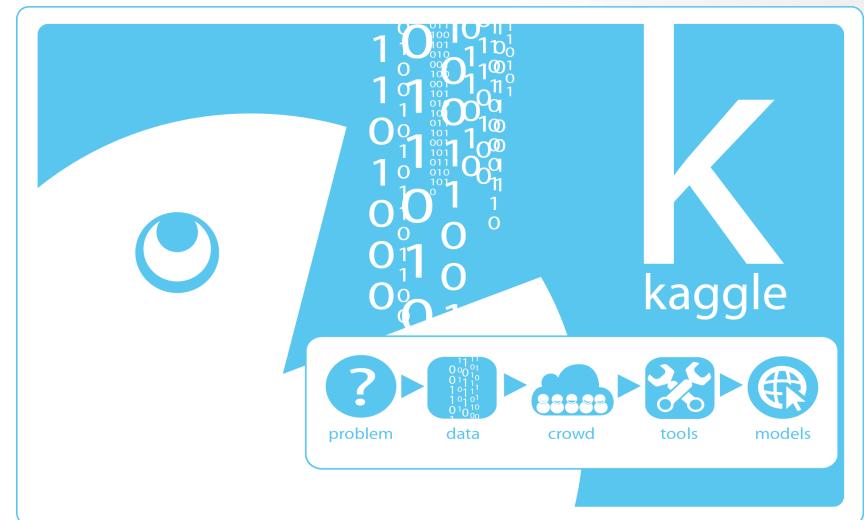
Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai
@matlabulus

About Me

- 2005 - 2015
- Water Engineer
 - Consultant for Utilities
 - EngD Research
- 2015 - Present
- Data Scientist
 - Virgin Media
 - Domino Data Lab
 - H2O.ai

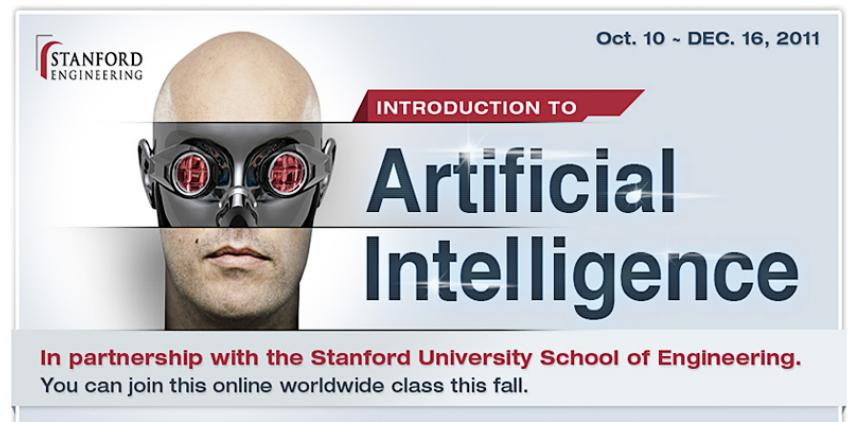
About This Talk

- What happened
 - Things I did since I started participating in Kaggle competitions.
 - New opportunities—results of new skills and friends.



First MOOC Experience

- One of the first Massive Open Online Courses.
 - Met some new friends.
 - Decided to collaborate for fun.
 - “How about Kaggle?”
 - “What is Kaggle?”



The banner features the Stanford Engineering logo at the top left. In the center is a graphic of a man's face with glowing red eyes, overlaid with a grid pattern. To the right of the graphic, the text "Oct. 10 ~ DEC. 16, 2011" is at the top, followed by "INTRODUCTION TO" in a red box, and "Artificial Intelligence" in large, bold, dark blue letters. Below this, a red banner states "In partnership with the Stanford University School of Engineering. You can join this online worldwide class this fall."

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.

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

[Follow @ailclass](#) 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!

First Kaggle Experience

- First time in my life
 - Supervised learning
 - Random Forest
 - Support Vector Machine
 - Neural Networks
 - Train, Validate & Predict.
 - “Is it black magic?”

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	

First Kaggle Experience

- Problems
 - “Hey Joe, you are a nice guy but we can’t work together.”
 - “You love MATLAB so much. You even call yourself @matlabulous on twitter!”
 - “We prefer R/Python.”
- Results
 - I kept using MATLAB
 - Lone wolf
 - No collaboration ☹



87th/264

Identifying Skills Gap

- That competition was a good wake up call.
- Obvious skills gap:
 - Open-source Programming languages
 - Machine learning techniques
 - Collaboration
- Kind of related
 - Data visualisation
 - Handling large datasets
 - Explaining results

From MATLAB to R/Python

	MATLAB	Python	R
Neural Networks	✓	✓	✓
Random Forest	✓	✓	✓
SVM	✓	✓	✓
Other Machine Learning Libraries	Toolboxes (commercial + open source)	Scikit-learn and many more	CRAN, GitHub (A LOT!)
Data Visualisation	I wasn't good at it anyway ...	Matplotlib (plus a lot more since then)	ggplot2 (WOW!) (plus a lot more since then)

What can people do with R?



James Cheshire, UCL
[Link](#)



Paul Butler, Facebook
[Link](#)

Filling the Skills Gap

- More MOOC
 - Machine Learning
 - Andrew Ng (Coursera)
 - Data Analysis
 - Jeff Leek (Coursera)
 - R
 - Intro to Programming
 - Dave Evans (Udacity)
 - Python
- Things I also picked up:
 - Linux (Ubuntu)
 - Git
 - Cloud computing
 - HTML / CSS

Learning from other Kagglers

- Continuous learning
 - Kaggle's forums and blogs.
 - New tools and tricks.
 - Many things you cannot learn from school.
 - I am standing on the shoulders of many Kagglers.



123rd/634



21st/89



214th/699



87th/264



250th/3514

Side Project 1 – Crime Data Viz

Crime Data Visualisation

READY?

Continue to scroll down and modify the settings. Come back and click this when you are ready to render new plots.

[Update Graphs and Tables](#)

BASIC SETTINGS

Enter a Location of Interest: London Eye (Demo)

Examples: Oxford, Wembley Stadium, M16 ORA etc.

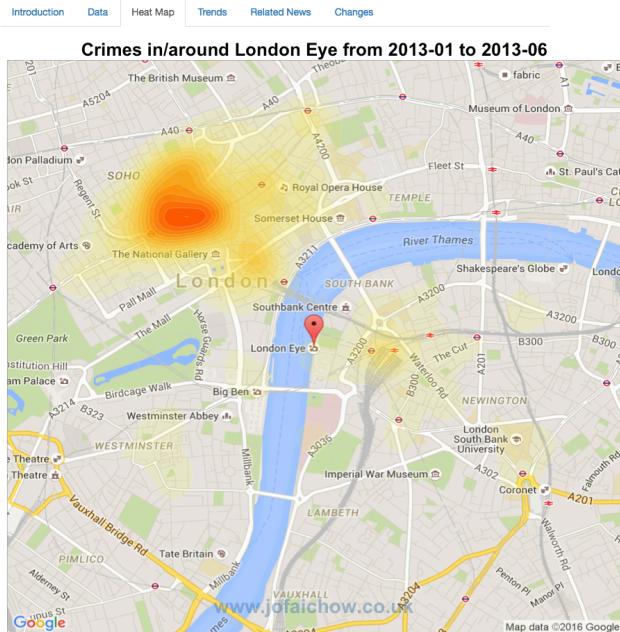
First Month of Data Collection: 2013-01

Length of Analysis (Months): 1 → 26

Note: data is available from Dec 2010 to Mar 2014. There is inconsistency in 2010-2011 records so I have omitted them for now. It takes longer to render the plots when you increase this number.

MAP SETTINGS

Choose Facet Type: none



MAP SETTINGS

Choose Facet Type: none

Choose Google Map Type: roadmap

High Resolution?

Black & White?

Zoom Level (Recommended - 14): 12 → 14 → 16

DENSITY PLOT SETTINGS

Alpha Range: 0.1 → 0.4 → 1

Number of Bins: 5 → 15 → 50

Boundary Lines Width: 0.1 → 1

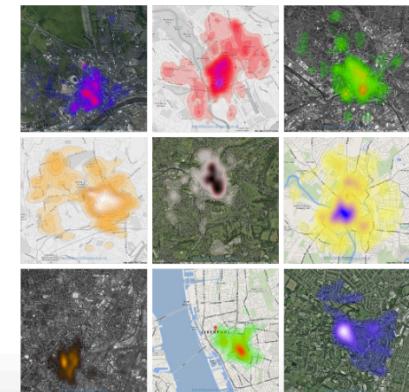
Boundary Lines Colour: grey95

Fill Gradient (Low): yellow

Fill Gradient (High): red

MISC. SETTINGS

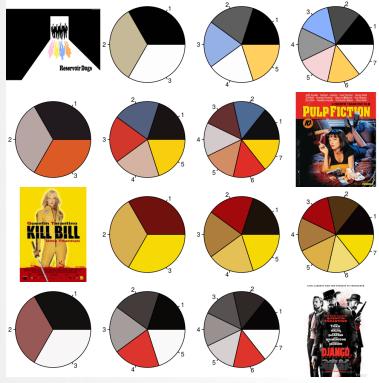
Use 'Blendiftbayes' Watermark?



```
shiny::runGitHub("rApps", "woobe",  
subdir = "crimemap")
```

Other Side Projects

- Colour Palette



- [github.com/
woobe/rPlotter](https://github.com/woobe/rPlotter)

- World Cup 2014 Correct Score Prediction

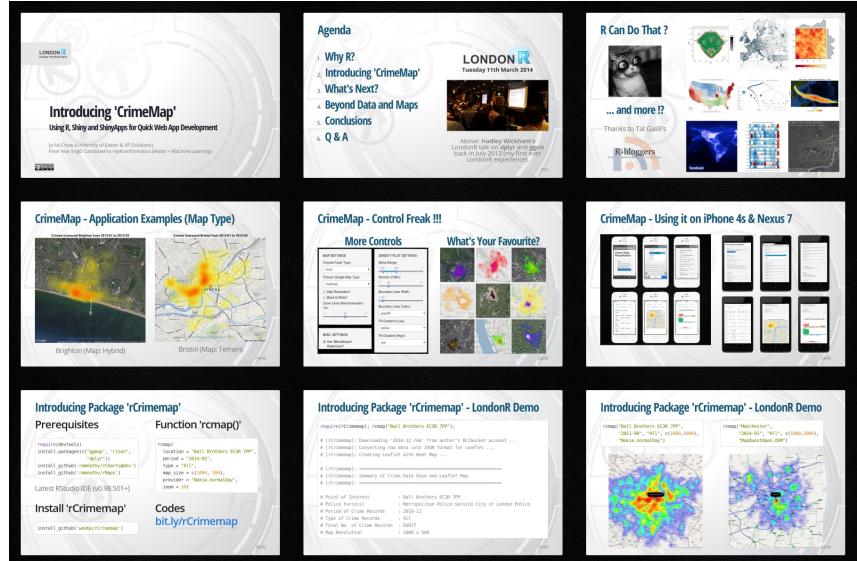
- ML vs. my friends
- 10 out of 64 (15.6%)
- Friends' avg. = 4 (6.3%)
- [github.com/
woobe/wc2014](https://github.com/woobe/wc2014)

Open Up Myself

- Before Kaggle/MOOC
 - I was drawing a circle around myself.
 - Fear of change.
 - Domain-specific problem solving.
- After Kaggle/MOOC
 - Data-driven approach.
 - Not a subject matter expert? No worries ☺
 - Free to try new tools, to learn and to create.

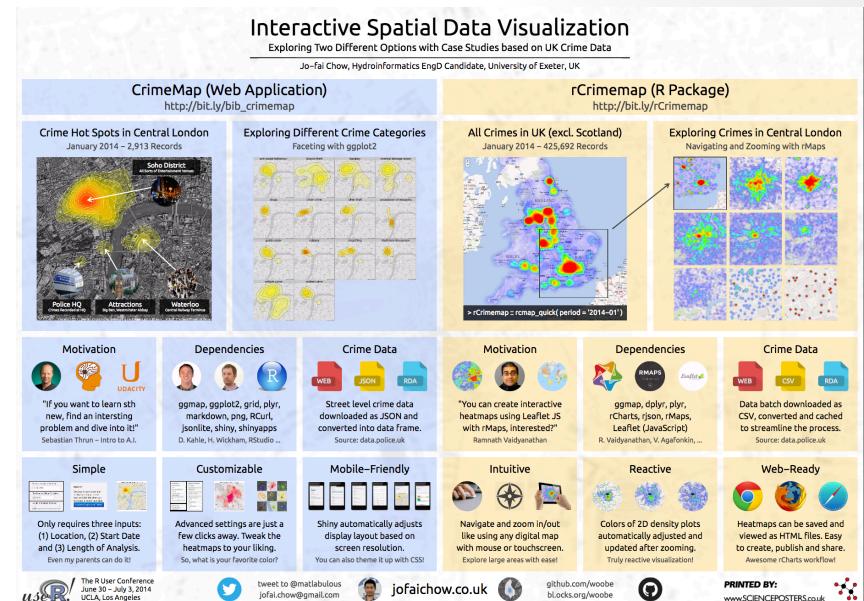
New Opportunities

- LondonR
 - First presentation outside water industry / academia.
 - Very positive feedback.
 - Led to other projects.
 - bit.ly
/londonr_crimemap

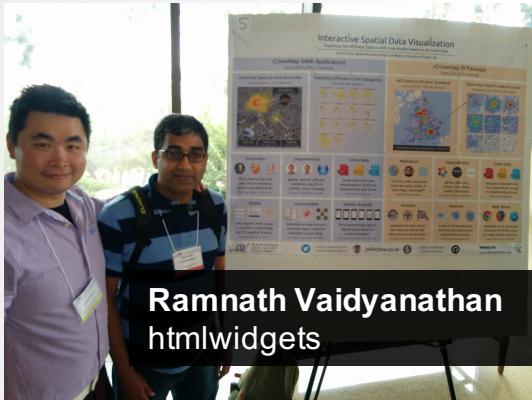


New Opportunities

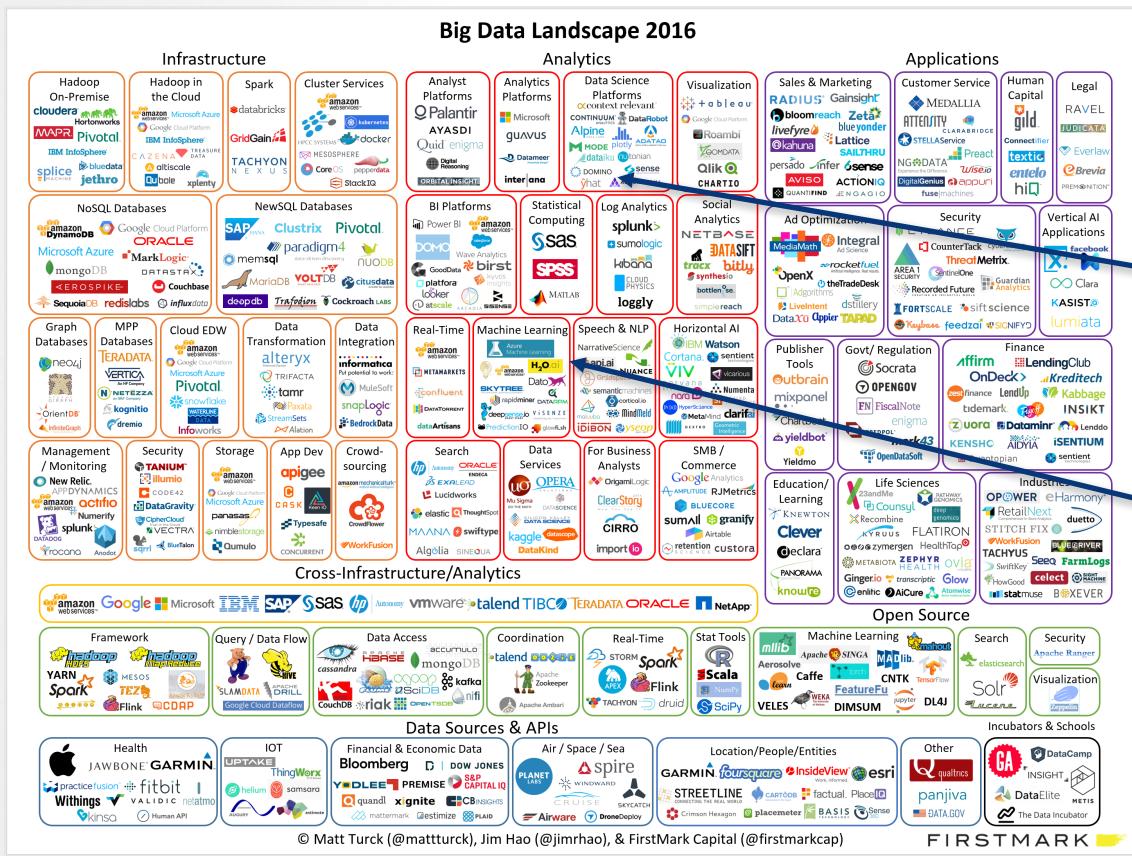
- useR! 2014 (UCLA)
 - Presented a poster.
 - Met new friends.
 - Life-changing event.
 - github.com/woobe/useR_2014



New Friends



About Domino and H2O



Data Science

Platforms
context relevant

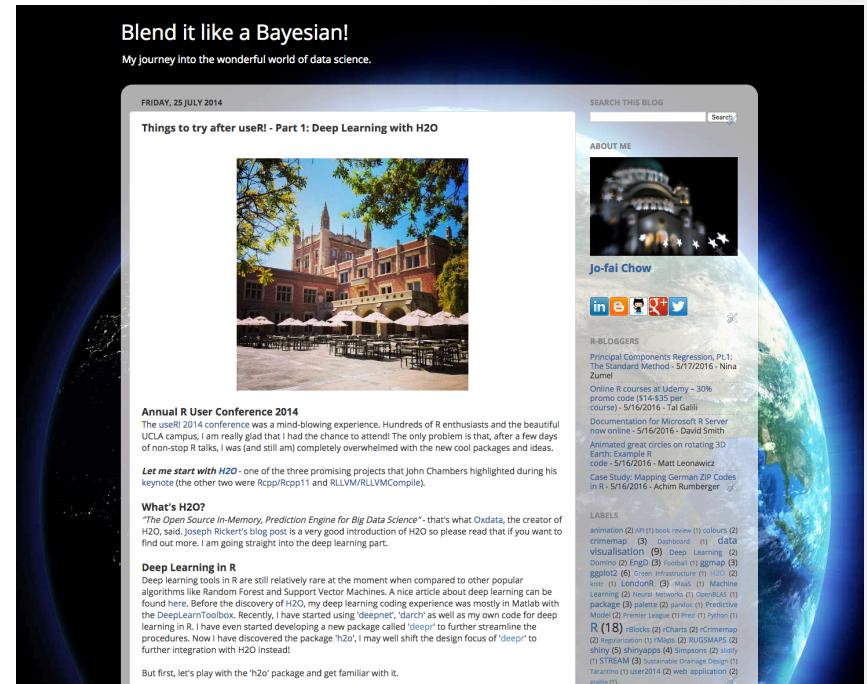


Machine Learning



More Opportunities

- First blog post about H2O
 - Things to try after useR!
 - Part1: Deep Learning with H2O



More Opportunities

- Blog post about Domino and H2O
 - I did it for fun. I did not have any expectation.
 - It helped attract customers to both Domino and H2O.



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

By Nick, on Sep 19 2014

share on: [Twitter](#) [Facebook](#) [Google+](#)

This is a guest post by Jo-Fai Chow

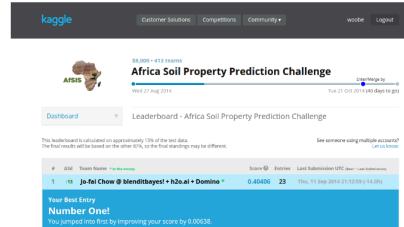
The sample project (code and data) described below is [available on Domino](#).

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

This blog post is the sequel to [TTTAR1 a.k.a. An Introduction to H2O Deep Learning](#). If the previous blog post was a brief intro, this post is a proper machine learning case study based on a recent [Kaggle competition](#): I am leveraging [R](#), [H2O](#) and [Domino](#) to compete (and do pretty well) in a real-world data mining contest.



Africa Soil Property Prediction Challenge

Score 0.40405 | Entries 23 | Last Submission UTC Mon, 15 Sep 2014 21:12:59 +14:00

Your Best Entry
Number One!

Becoming a Data Scientist

Data Scientist

Virgin Media

June 2015 – January 2016 (8 months) | Hook

Main duties and responsibilities:

- Data integration (MSSQL, IBM Netezza and SSIS)
- Reproducible technical analysis using RMarkdown
- Interactive data visualisation using R and D3
- Report automation using scheduled R scripts
- Combining outputs from above into apps for non-technical users (R-Shiny and QlikView)

Projects:

- Home phone usage (pattern detection and time-series forecast)
- Customer lifetime value and churn analysis (identifying and prioritising investment opportunities)
- Fraud Investigation (anomaly detection, alert system and web application)



Data Science Evangelist

Domino Data Lab

May 2015 – January 2016 (9 months) | Remote

Promoting Domino product via writing blog posts/tutorials and giving talks at meetups/conferences.

Projects:

- How to use R, H2O, and Domino for a Kaggle competition (blog)
- How to turn your predictive models into APIs Using Domino (blog)
- Deploying your Predictive Models as a Service via Domino (LondonR talk)
- How to avoid overfitting for Rossmann store sales prediction on Kaggle (blog - coming up)



A screenshot of the Domino Data Lab interface. It shows two main windows: one for exploring datasets and another for building machine learning models. The interface is clean with a white background and blue accents. A search bar is visible at the top of both windows.

How to use R, H2O, and Domino for a Kaggle...

A screenshot of a terminal window demonstrating the command-line interface for Domino. The user is shown running commands to build an R model and then exposing it as a REST API. The terminal output includes URLs and API documentation snippets.

Turning your R (or Python) models into APIs

London Kagglers Assemble

- London Kaggle Meetup
 - Sep 2015
 - I met my Kaggle buddy
Mickael Le Gal
 - He is a product data
scientist at Tictrac



London Kagglers Assemble

- Rossmann Store Sales
 - We got stuck at top 10% for a long period.
 - Mickael had a breakthrough in feature engineering with 48 hours to go.
 - I re-trained all models and completed model stacking just a few hours before the deadline (thanks to Domino Data Lab).
 - Top 2% finish (our best result so far).



Joining H2O.ai

Data Scientist

H2O.ai

February 2016 – Present (4 months) | Mountain View, California / Europe (Remote)



- Build machine learning models with data using R, Python and H2O.
- Prepare customer data and pipeline for improved analysis.
- Derive insights from data for users and customers to make decisions.
- Work with applications and data science team to prototype demonstrative smart applications.
- Build community in the geography by blogging and presenting at meetups/conferences.
- Liaise with customers to expand use of H2O beyond initial footprint.
- Demonstrate the value of H2O vs other possible industry standard systems like SAS, SPSS or other closed systems for users trying to make a choice.

A screenshot of a presentation slide. The title is "Using H2O Random Grid Search for Hyper-parameters Optimization". Below the title is the H2O.ai logo. To the right of the logo is a small blue icon of a person's head with a brain-like pattern. At the bottom of the slide, there is a grey footer bar with the text "Hyper-parameters Optimization".

Using H2O Random Grid Search for
Hyper-parameters Optimization

H2O.ai

Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai

Hyper-parameters Optimization

Summary of Benefits

- Direct
 - Identify data science skills gap.
 - Learn quickly from the community.
 - Expand your network.
 - Prepare yourself for real-life data challenges.
- Indirect
 - You also learn non-ML skills along the way.
 - You learn to build small data products (e.g. graph, web app, REST API) and help others gain insight.

Strata Hadoop London Talks

- “Model stacking and parameters tuning.”
 - Wednesday 1 June
 - 7pm
 - LondonR at Strata
 - Free event
- “Generalised Low Rank Models.”
 - Thursday 2 June
 - 2pm
 - Strata Hadoop main conference event

Big Thank You

- London Kaggle Meetup Organisers
- Mickael
- Marios
- You (Kagglers)
- Prof. Dragan Savic (supervisor at uni. of Exeter)
- Mango Solutions
- Virgin Media
- Domino Data Lab
- H2O.ai

Any Questions?

- Contact
 - joe@h2o.ai
 - @matlabulous
 - github.com/woobe
- Links (All Slides)
 - github.com/h2oai/h2o-meetups
- H2O in London
 - Coming soon!
 - Meetups
 - Office
 - We're hiring!
 - www.h2o.ai/careers