Kaggle Competitions, New Friends, New Skills and New Opportunities

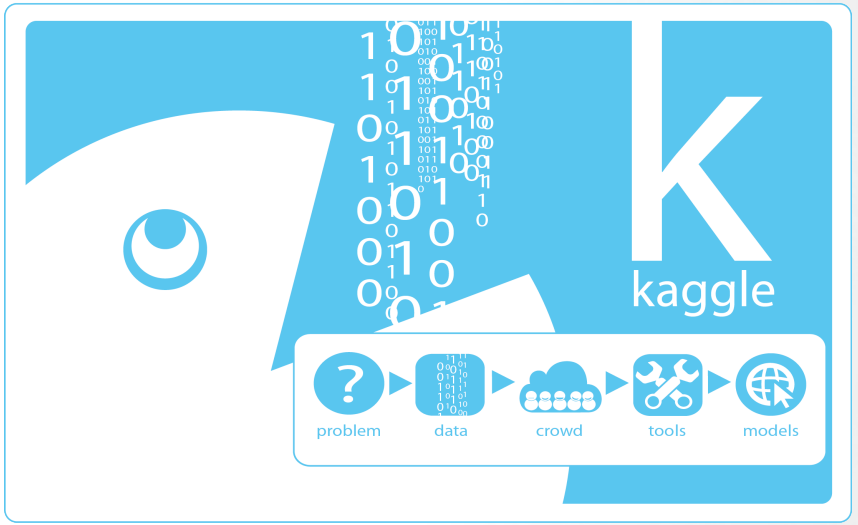
Jo-fai (Joe) Chow
Data Scientist
joe@h2o.ai
@matlabulus

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.

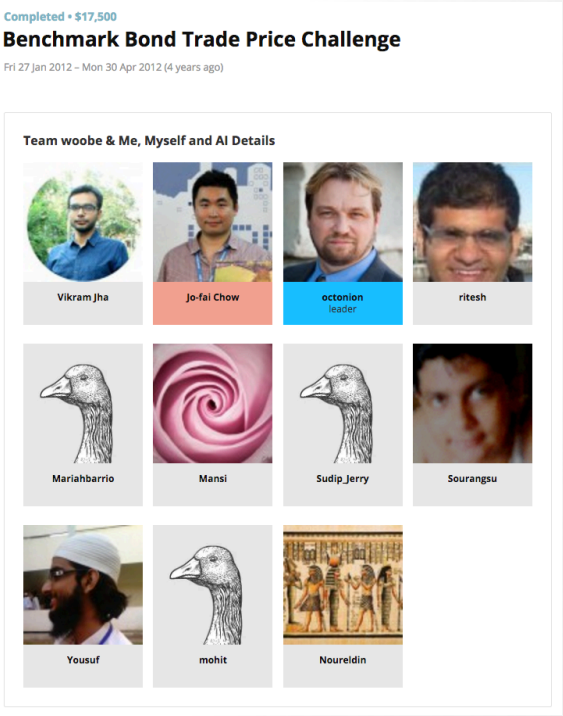


First MOOC Experience

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

First Kaggle Experience

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



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 ☹

Identifying Skills Gap

- That competition was a good wake up call.

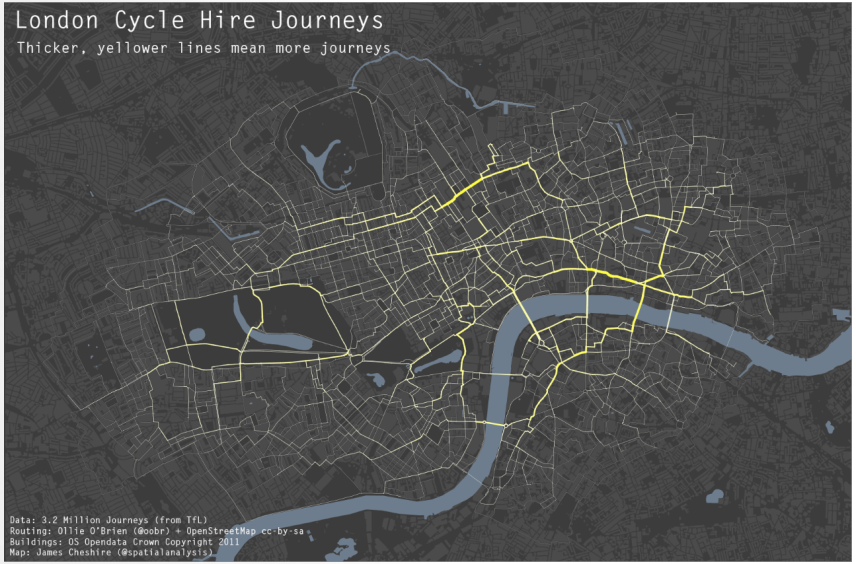
- Obvious skills gap:
  - Open-source Programming langauges
  - Machine learning techniques
  - Collaboration

- 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

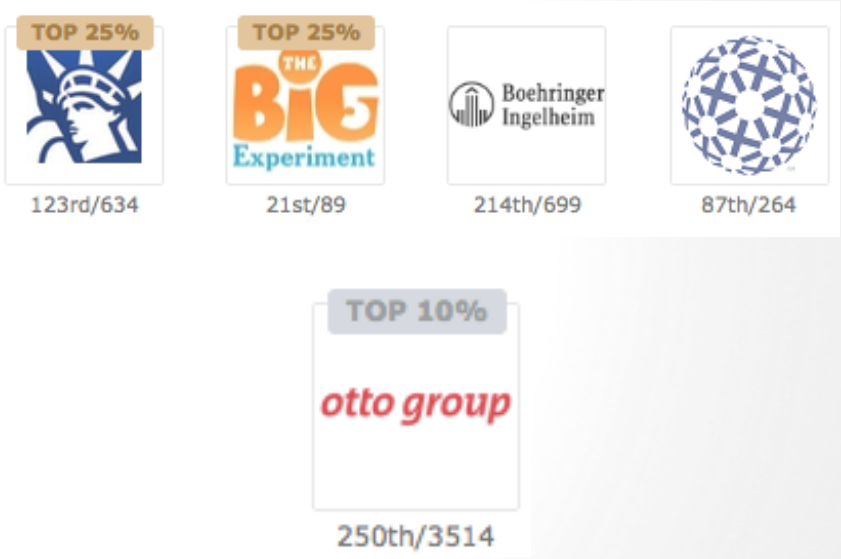
Paul Butler, Facebook
Link

Filling the Skills Gap

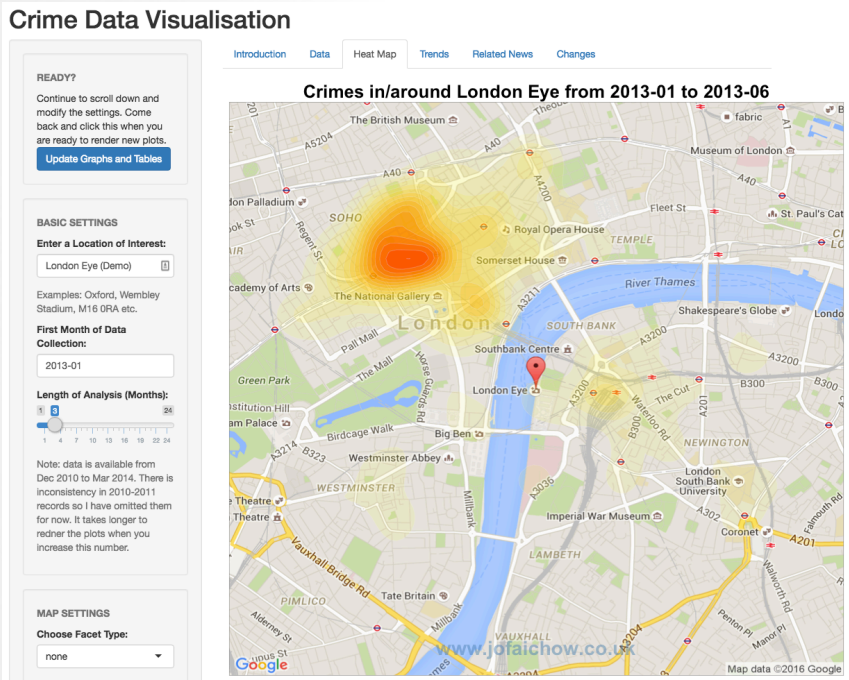
- More MOOC
  - Machine Learning
    - Andrew Ng (Coursera)
  - Data Analysis
    - Jeff Leek (Coursera)
    - R
  - Intro to Programming
    - Dave Evans (Udacity)
    - Python
- Things I also picked up:
  - Linux (Ubuntu)
  - Git
  - Cloud computing
  - HTML / CSS

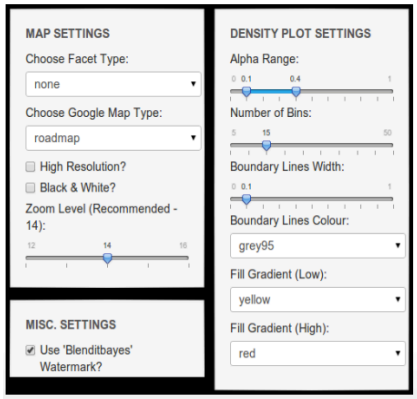
Learning from other Kagglers

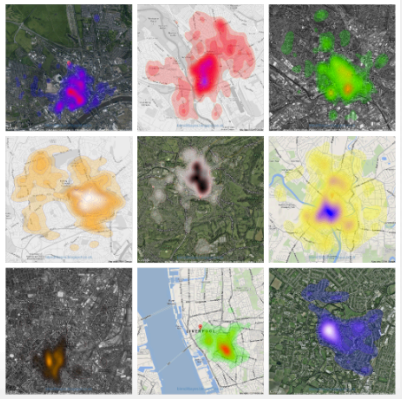
- Continuous learning
  - Kaggle's forums and blogs.
  - New tools and tricks.
  - Many things you cannot learn from school.
  - I am standing on the shoulders of many Kagglers.



Side Project 1 – Crime Data Viz



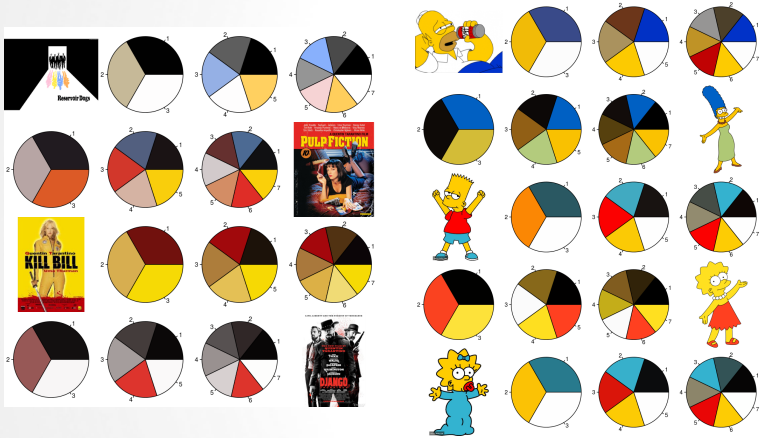




```
shiny::runGitHub("rApps", "woobe",
                 subdir = "crimemap")
```

Other Side Projects

- Colour Palette



  - 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

Open Up Myself

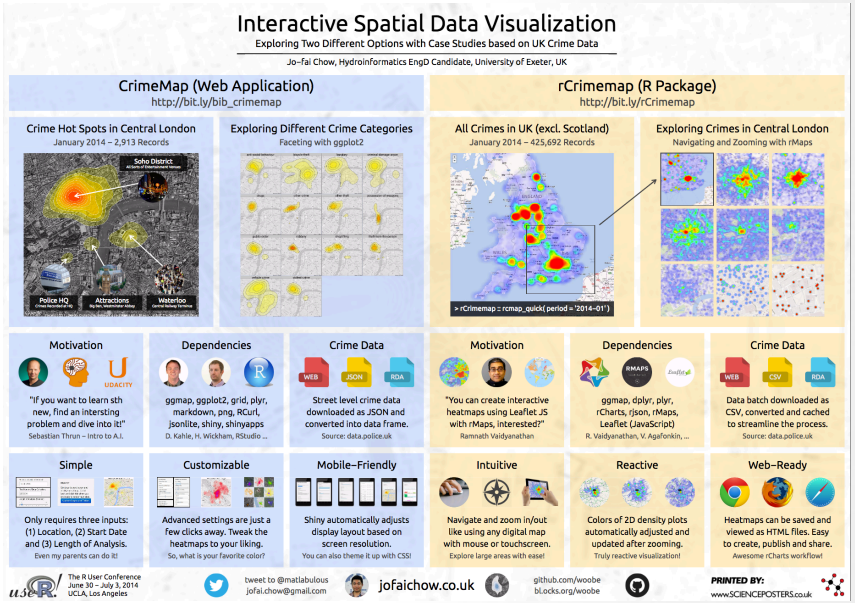
- Before Kaggle/MOOC
  - I was drawing a circle around myself.
  - Fear of change.
  - Domain-specific problem solving.
- After Kaggle/MOOC
  - Data-driven approach.
  - Not a subject matter expert? No worries ☺
  - Free to try new tools, to learn and to create.

New Opportunities

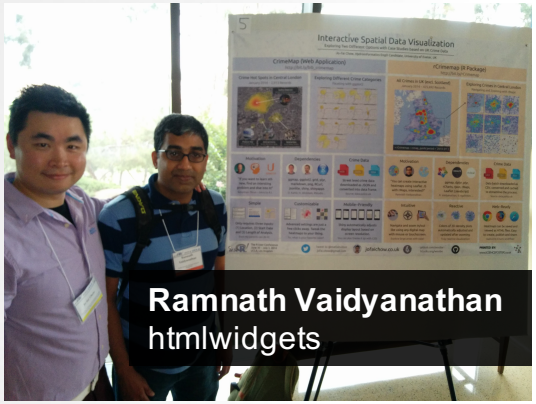
- LondonR
  - First presentation outside water industry / academia.
  - Very positive feedback.
  - Led to other projects.
  - bit.ly /londonr_crimemap

New Opportunities

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

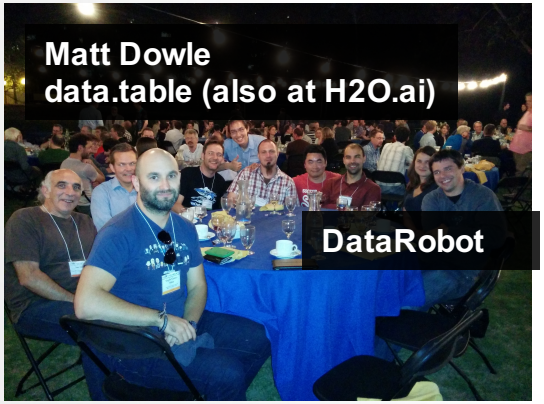


New Friends


Interactive Spatial Data Visualization
Ramnath Vaidyanathan
htmlwidgets


RStudio

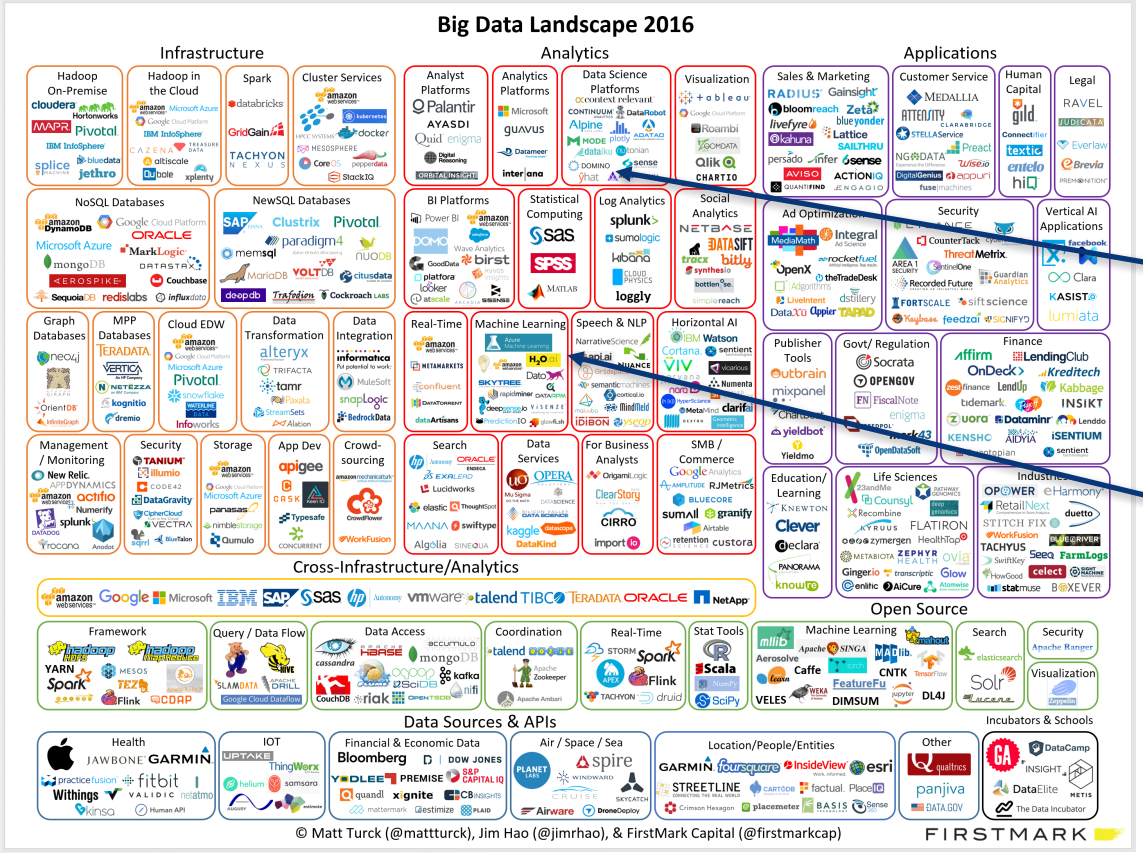

rOpenSci


Matt Dowle
data.table (also at H2O.ai)
DataRobot


Nick @
DominoData
Lab

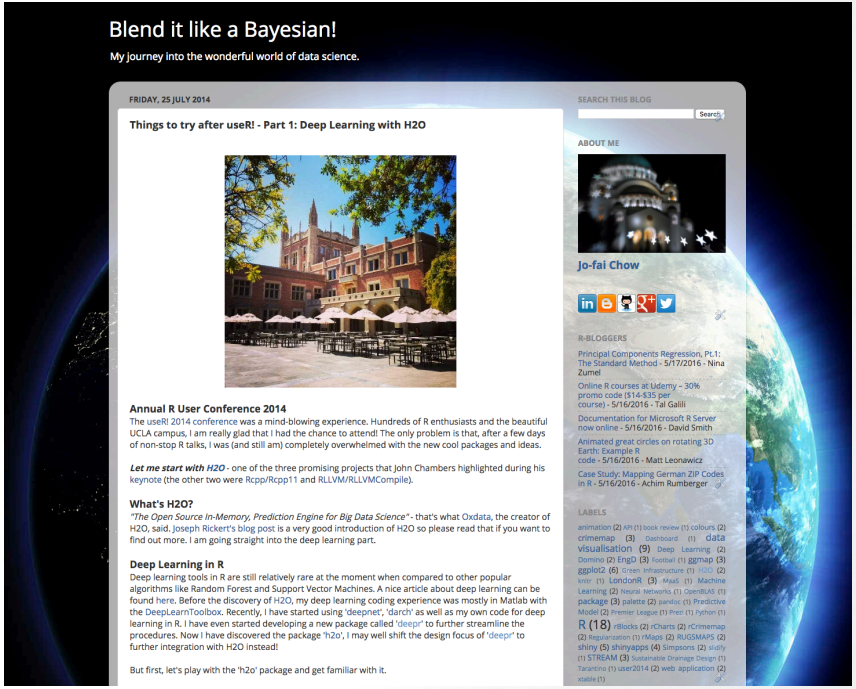

H2O.ai &
John Chambers!

About Domino and H2O



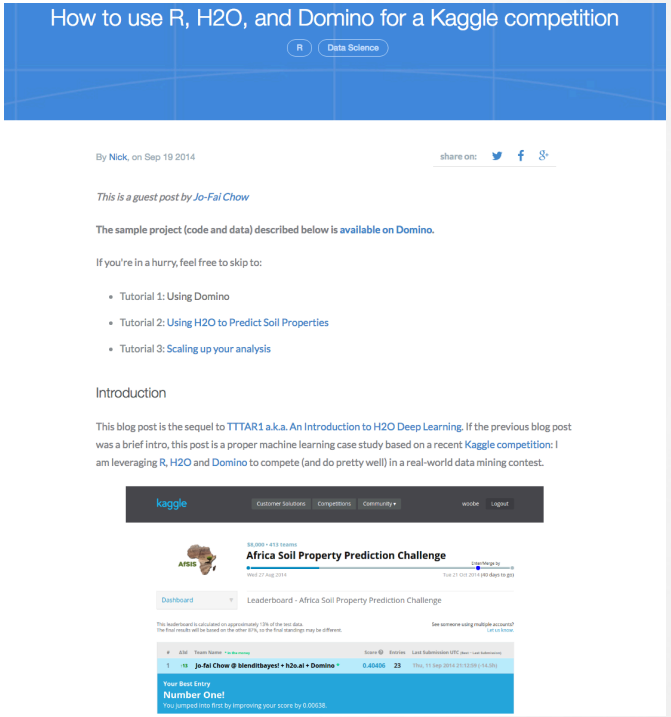
More Opportunities

- 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.



Becoming a Data Scientist

Data Scientist
Virgin Media
June 2015 – January 2016 (8 months) | Hook

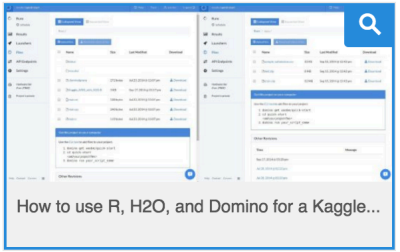
Main duties and responsibilities:
- Data integration (MSSQL, IBM Netezza and SSIS)
- Reproducible technical analysis using RMarkdown
- Interactive data visualisation using R and D3
- Report automation using scheduled R scripts
- Combining outputs from above into apps for non-technical users (R-Shiny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

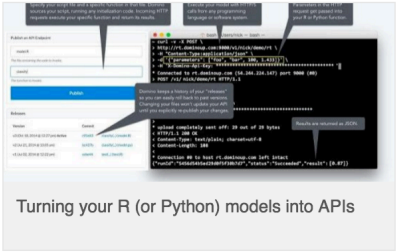
Projects:
- Home phone usage (pattern detection and time-series forecast)
- Customer lifetime value and churn analysis (identifying and prioritising investment opportunities)
- Fraud Investigation (anomaly detection, alert system and web application)

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)


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

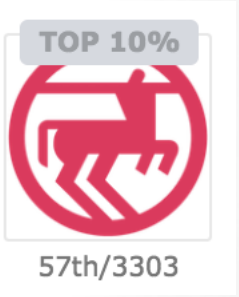

Turning your R (or Python) models into APIs

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.

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