THESSALONIKI

digital analytics

MEETUP

The Kaggle experience

From a digital analysts' perspective

HELLO!

I ‘m Alexandros

Independent Consultant
“Data analytics for digital growth”

@alpapag

alex-papageo.com

So how does it all work ?

Kaggle started as a platform for ML competitions

–now it’s much more than that

Active competitions today

15 Active Competitions

Two Sigma: Using News to Predict Stock Movements

Use news analytics to predict stock price performance

Featured · Kernels Competition · 6 months to go · news agencies, time series, finance, money

$100,000
2,902 teams

LANL Earthquake Prediction

Can you predict upcoming laboratory earthquakes?

Research · 4 months to go · earth sciences, physics, signal processing

$50,000
680 teams

Elo Merchant Category Recommendation

Help understand customer loyalty

Featured · a month to go · regression, tabular data, banking

$50,000
3,059 teams

Google Analytics

@googleanalytics

Can you predict revenue per customer using @googleanalytics' demo account with the Google Merchandise Store? @kaggle wants to know. #measure

The data source

Google

official merchandise store

Woven Winter Scarf - Green

GGL1486

£18.00

Bobble winter hat

GGL1484

£25.00

Winter Collection Mini Gift Box

GGL1482

~~£18.00~~ **£9.60**

Not everything went as expected...

THE INNOVATION MACHINE Writer

HOME A.I. & DATA SCIENCE INNOVATION TECHNOLOGY & SOCIETY STARTUP & PRODUCT IDEAS SELECTED PATENTS

Kernels, Clicks and Boosted Trees: Highlights from the 1st Google Analytics Kaggle Competition

Alexandros Papageorgiou
Dec 12, 2018 · 8 min read

It was no doubt an interesting encounter, with Google Analytics meeting:

Kaggle (the machine learning competition platform)

Rstudio (the cash prize sponsor)

and Big Query (the data host).

The purpose ? to organise a machine learning competition- **the first of its kind** having Google Analytics data as its raw material.

Image by Orestis Papageorgiou

The event (in its first phase, i.e. before the redesign due the data leakage, more on this later) attracted **more than 3,500 teams**, making it one of the most popular competitions hosted in the platform's 8 year history.

As mentioned earlier the reception of the competition was exceptionally good considering the high level of participation and the initial excitement. But it turned out not to be a walk in the park neither for the participants nor for the organisers.

Photo: sumologic.com/blog

Right after the initial launch there were a couple of rounds with issues raised from the teams regarding the consistency in the data and metrics definitions provided by the organisers. Shortly after, **the news broke out that a data leakage had taken place.**

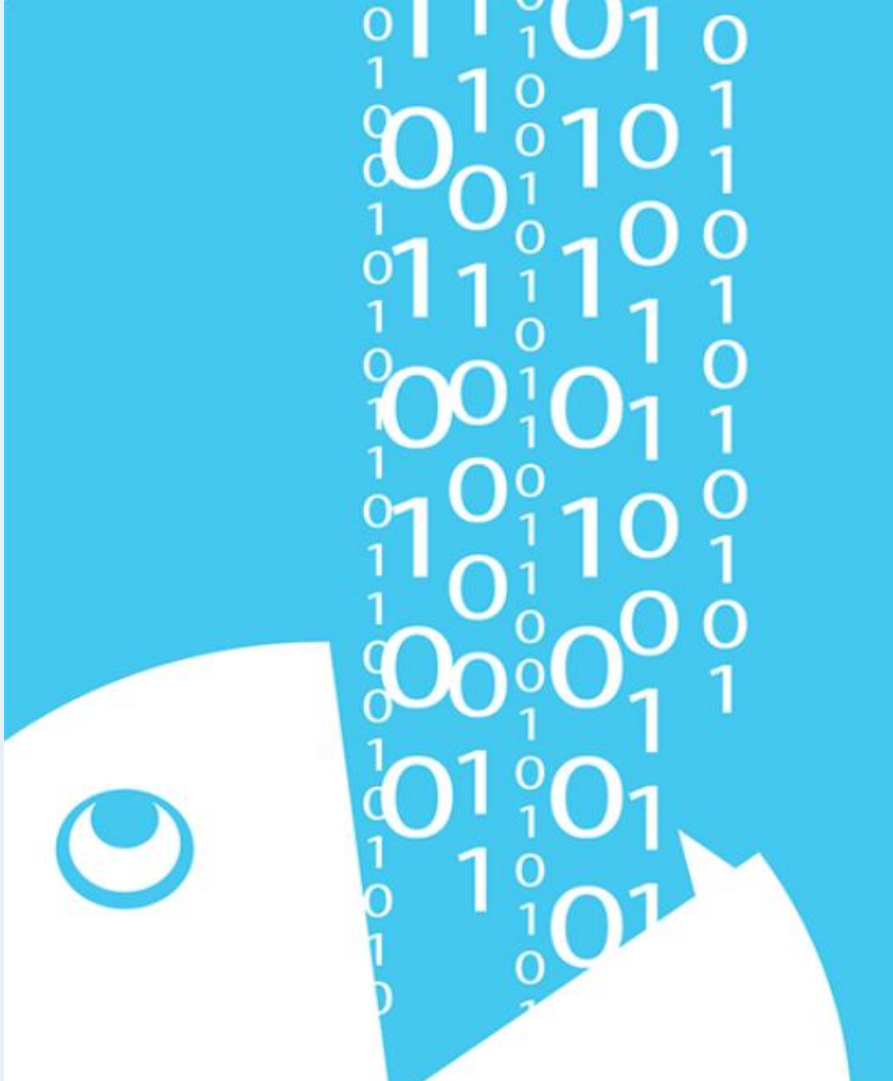
A data leakage can take several forms but in this instance it was linked to the publicly available Google Demo account (this alternative source of truth for the G store traffic left doors open to mine the answers).

There was a pause and **it took a few weeks for Kaggle to relaunch** the competition, with an updated dataset and prediction targets.

https://goo.gl/e4ZCBn

My 4 ways to make the most of Kaggle

*Remember it's not just about competing!



1

Datasets

15K datasets to explore

Datasets

Documentation | New Dataset

Public | Your Datasets | **Favorites**

Sort by: Hotness

9 Datasets | Sizes | File types | Licenses | Tags | Search datasets

2018 Kaggle ML & DS Survey Challenge
Explore the 2018 Kaggle ML & Data Science Survey for $28,000 in cash prizes
Kaggle updated 3 months ago

783

survey anal...
CSV | 3.9 MB | CC4
256 | 17 | 299k

Google Analytics Sample
Google Analytics Sample (BigQuery)
Google BigQuery (Continuous updates)

87

web sites
marketing
marketing ...
bigquery
BigQuery | 6.3 GB | CC0
17 | 4 | 23k

Trending YouTube Video Statistics
Daily statistics for trending YouTube videos
Mitchell J updated 2 months ago (Version 114)

798

languages
popular cul...
statistics
+ 2 more...
CSV | 199.1 MB | CC0
93 | 21 | 166k

Avocado Prices
Historical data on avocado prices and sales volume in multiple US markets
Justin Kiggins updated 8 months ago (Version 1)

355

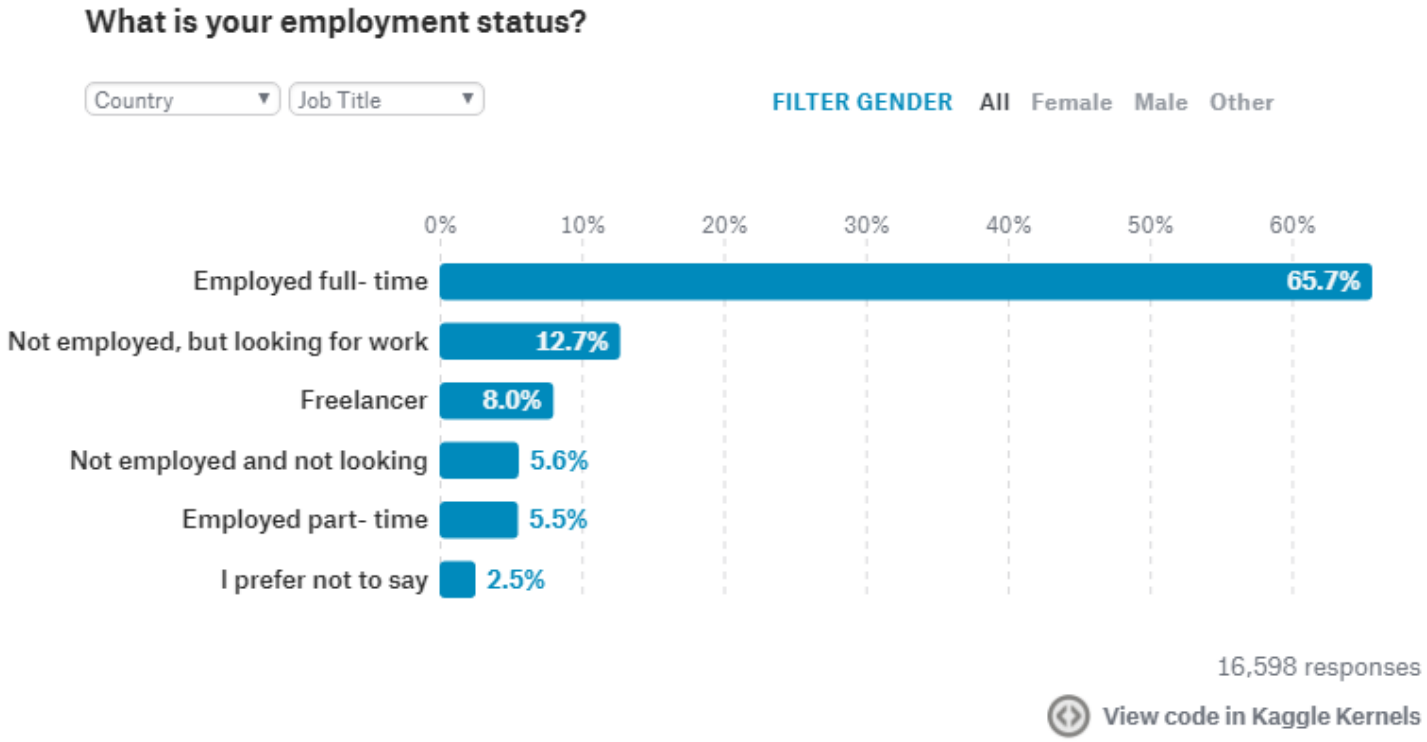
food and dr...
CSV | 628.7 KB | ODbL
98 | 6 | 85k

2

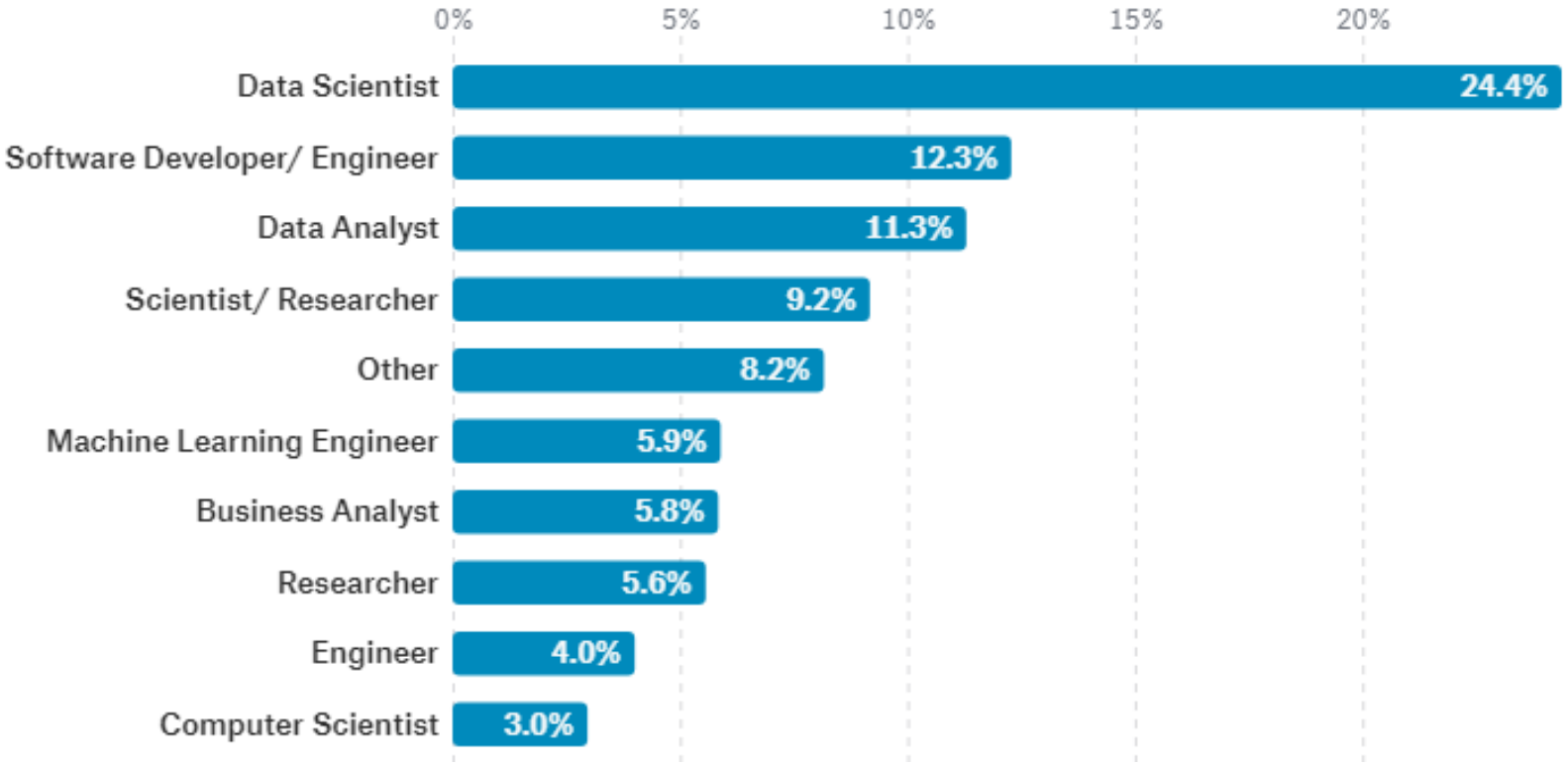
Community

of 2.5M Kagglers to interact with

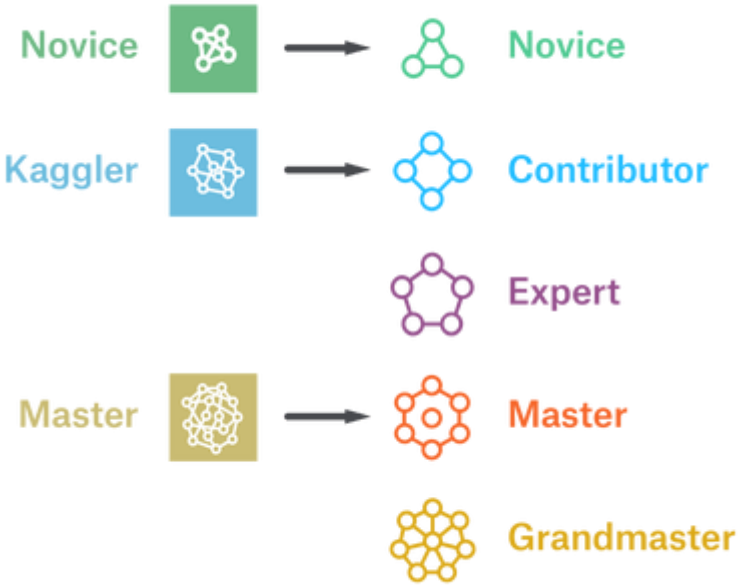
Kagglers are mostly professionals



The top 10 titles



Kaggle progression ladder



3 Compute Power

on the Cloud

Kaggle kernels: a powerful “laptop” on the cloud

- 9 hours execution time
- 5 Gigabytes of auto-saved disk space
- 16 Gigabytes of temporary disk space

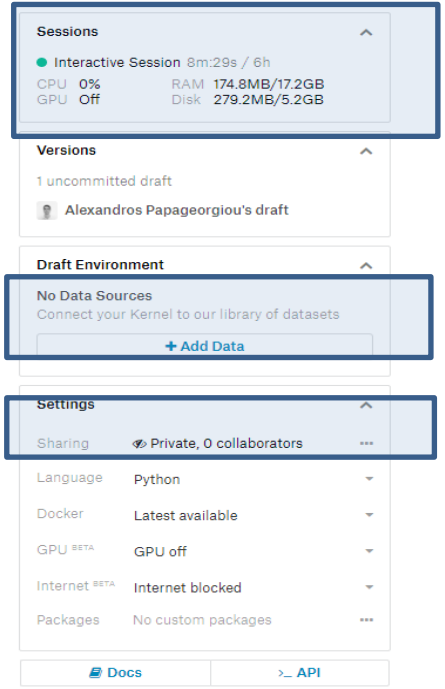
- CPU Specifications
  - 4 CPU cores
  - 17 Gigabytes of RAM
- GPU Specifications
  - 2 CPU cores
  - 14 Gigabytes of RAM

As of 2018: Private kernels possible too

Kaggle for private projects

[ ]:
```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files

import os
print(os.listdir("../input"))

# Any results you write to the current directory are saved as output.
```

[ ]:
```
print("hello digital analytics meetup!")
```



4 Professional

Development

Getting started with Kaggle

- Explore the datasets
- Start a private project
- Join a learning competition
- Join a regular competition

Remember though...

Kaggle provides for you both the data and the question to answer

Life outside Kaggle might not be like this

Take aways

- 15K datasets waiting to be explored
- 2.5M Kagglers to learn from and network with
- Powerful free resources for data science on the cloud
- Can be part of your professional development plan

Happy Kaggling

THANK YOU!

www.alex-papageo.com/blog

@alpapag

Alexandros Papageorgiou

Useful Kaggle Resources

- kaggle.com/datasets
- kaggle.com/discussion
- kaggle.com/kernels
- kaggle.com/jobs
- kaggle.com/learn-forum
- kaggle.com/c/ga-customer-revenue-prediction