How does a team of Data Scientists go about researching?
Engine Room
From zero to first prototype (2017)
....and revisited version (2018)

Scalability
What corners to cut to make it works in production?

Still under development
...no finish line
Global Service & Support

800+ Employees
19 Countries
360 Developers
250 Client Services
8 Data Points of Presence

Americas Activation Hub
Global Competency Center & EMEA Hub
APAC Activation Hub
the stage
mission
External Deterministic Providers

YouGov
YouGov®
(Deterministic)

> 700K events/day

VERTICA
(Adform bid responses)

Cookie Time Location OS Device

1.5B events/day

~2.5 Months of data
7M events
Filtering
Remove cookies with multiple clusters

74K Profiles

80%
Somewhat small, but enough non-singletons (users with multiple cookies)

20%
Singletons Non-singletons
Training
2.236,222 events

Validation
4,776,974 events

More of it...

300M

300M
engine room
Cluster cookies into user profiles.
Cross-Device Graph

Graph cutting algorithms
- Min-Cut
- Markov Random Fields
- Community Detection
- Graph Clustering...

Production Scale

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>B</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>events</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nodes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edges</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Scale...what to do??

Divide Graph ...by location
(Pre-Clustering)

(Cookie, IP)

1 2 3 n-1 n

(Connected Components Algorithm)

...by location

Ignored
Cookie association
Cookie association by Pair-wise classification

Create cookie pairs as classification data set...with ground truth

Positive

A B  A C  A E
C D  A D  B E
B C  C E
B D  D E

Negative
Crafting Features
for each cookie pair...

Observations

<table>
<thead>
<tr>
<th>ID</th>
<th>cookie_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>log_time</td>
</tr>
<tr>
<td>Location</td>
<td>ip_v4</td>
</tr>
<tr>
<td></td>
<td>country_id</td>
</tr>
<tr>
<td></td>
<td>region_id</td>
</tr>
<tr>
<td></td>
<td>city_id</td>
</tr>
<tr>
<td></td>
<td>zip_code_id</td>
</tr>
<tr>
<td>Device</td>
<td>device_type_id</td>
</tr>
<tr>
<td></td>
<td>os_id</td>
</tr>
<tr>
<td></td>
<td>browser_id</td>
</tr>
<tr>
<td></td>
<td>browser_language_codes</td>
</tr>
<tr>
<td></td>
<td>mobile_app_id</td>
</tr>
</tbody>
</table>

Aggregation

Pageviews

0 12 24 hours

0 31 days

set a
set b
Crafting Features for each cookie pair...

**Observations**
- **ID**
  - cookie_id
- **Time**
  - log_time
- **Location**
  - ip_v4
  - country_id
  - region_id
  - city_id
  - zip_code_id
- **Device**
  - device_type_id
  - os_id
  - browser_id
  - browser_language_codes
  - mobile_app_id

**Aggregation**
- Pageviews over 0-24 hours
- Pageviews over 0-31 days

**Similarity**
- Cosine Similarity
- Correlation Similarity
- Hellinger Distance
- Max Overlap

**Features**
- Jaccard Index

Set A: set of pageviews for cookie A
Set B: set of pageviews for cookie B
Crafting Features for each cookie pair...

Observations
- ID
- Time: log_time
- Location:
  - ip_v4
  - country_id
  - region_id
  - city_id
  - zip_code_id
- Device:
  - device_type_id
  - os_id
  - browser_id
  - browser_language_codes
  - mobile_app_id

Features
- Numerical
- Vectorial
- Sets

Classifier
- Gradient Boosted Decision Trees
- 0.78

Vectorial Sets
- 26 dim. feature vector

Observations A B

(http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html)
Boosted Decision Trees

XGBoost

(https://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html)
Classification

Clustering

Pruning

Degrees of edges

95% percentile

(Classified Components Algorithm)
Pre-Clustering

Classification

Clustering

Pruning

Full Graph

Components

Sub-Graph

Clusters

Users
performance
Profile-cookie space

Precision / Recall

Precision = \frac{TP}{TP + FP}

Recall = \frac{TP}{TP + FN}

F1-Score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}

(True Positives)

(True Negatives)

(False Positives)

(False Negatives)

Estimated

Class A

Class B

(Harmonic Mean)
Performance Graphs

(Preliminary YouGov data set)

Dasgupta et al. “Overcoming Browser Cookie Churn with Clustering” 2012

Pre-clusters almost there...

Failed attempts...:-)
Performance Graphs

Precision/Recall for all experiments

(Preliminary YouGov data set…)

Legend:
- pre-cluster
- ip_usage
- ipos
- lifetime_pageviews
- pageviews
- timegap
- timegap_exp
- timegap_overlap
- x_pairwise
- Target
Performance Graphs

(Preliminary YouGov data set...)

91% 32% 47%

Precision  Recall  F1-score

(Preliminary YouGov data set...)

Precision/Recall for all experiments
Performance Graphs

Pruning

(Preliminary YouGov data set...)

Precision 93%
Recall 43%
F1-score 59%
2018 data...
1.5B events/day

Location

⏲

Cookie

Device

OS

2%

81K Profiles

Singletons

Non-singletons

More dense and lots of non-singletons
(users with multiple cookies)

98%

Filtering

Remove cookies with multiple clusters

14 Days of data

>12M events

(Deterministic)

(Deterministic)

(kafka)

>440K events/day

1.5B events/day

(Adform bid responses)

(Users with multiple cookies)
More of it...

Training:
- 2018-10-26
- 7 days
- Polen
- 5M events

Testing:
- 2018-11-04 to 2018-11-05
- 7 days
- Polen
- 7M events
Performance Graphs

Cookie-Matches data

Precision/Recall for fresh 2018 data

Precision: 94%
Recall: 79%
F1-score: 86%

(Cookie-Matches data)
production + scalability
Scheduled tasks

Training
- XD Deterministic Data
- Pre-Processing
- Pre-Clustering
- Train Classifier
- Model parameters
  - S3
  - Kafka
- Model ready

Predicting
- Pre-Processing
- Pre-Clustering
- Predict Cross Device ID's
  - Cross-Device Cookies Ready

Serving
- XD Data
- S3
- Kafka
- Serve...

Technologies
- Scala
- Parquet
- Spark
- Apache
- Kafka
- Python
Data

2018-11-01
2 days

2018-11-03

2018-11-04

16GB

2018-11-04
12 days

2018-11-13

96GB

Predicting

XD Data

Pre-Processing

Pre-Clustering

clusters >1000

Model parameters

Predict Cross Device ID’s

>48 hours

Cross-Device Cookies Ready

Problems

Large pre-clusters

One IP to rule them all....

Too many features

APACHE Spark
Pre-clusters <1000 gives <500K cookie pairs

Pre-clusters >1000 are ignored

One IP to rule them all....

Large pre-clusters

Scale of pre-cluster sizes (N choose K curve)
Ablation Study
Performance vs. scalability

Too many features

What feature to cut?

Performance Graphs
Feature importance ranking

Precision 94%  Recall 79%  F1-score 86%
(Cookie-Matches data)

Too many features

What feature to cut?
challenges
One IP to rule them all.

Large pre-clusters

IP Obscurity

176.22.252.???

Cookie

IP

Noisy pre-clusters
Anti-tracking mechanisms
The cat and mouse game

GDPR
Limited data set
Fake user agent
IP Obscurity
References

[1] Tran et al.  
“Classification and Learning-to-rank Approaches for Cross-Device Matching”  
CIKM Cup 2016

“Overcoming Browser Cookie Churn with Clustering”  
CIKM Cup 2016

“Fast unfolding of communities in large networks”  
2008

“Internet Device Graphs”  
KDD 2017
thank you

https://site.adform.com/careers/open-positions/