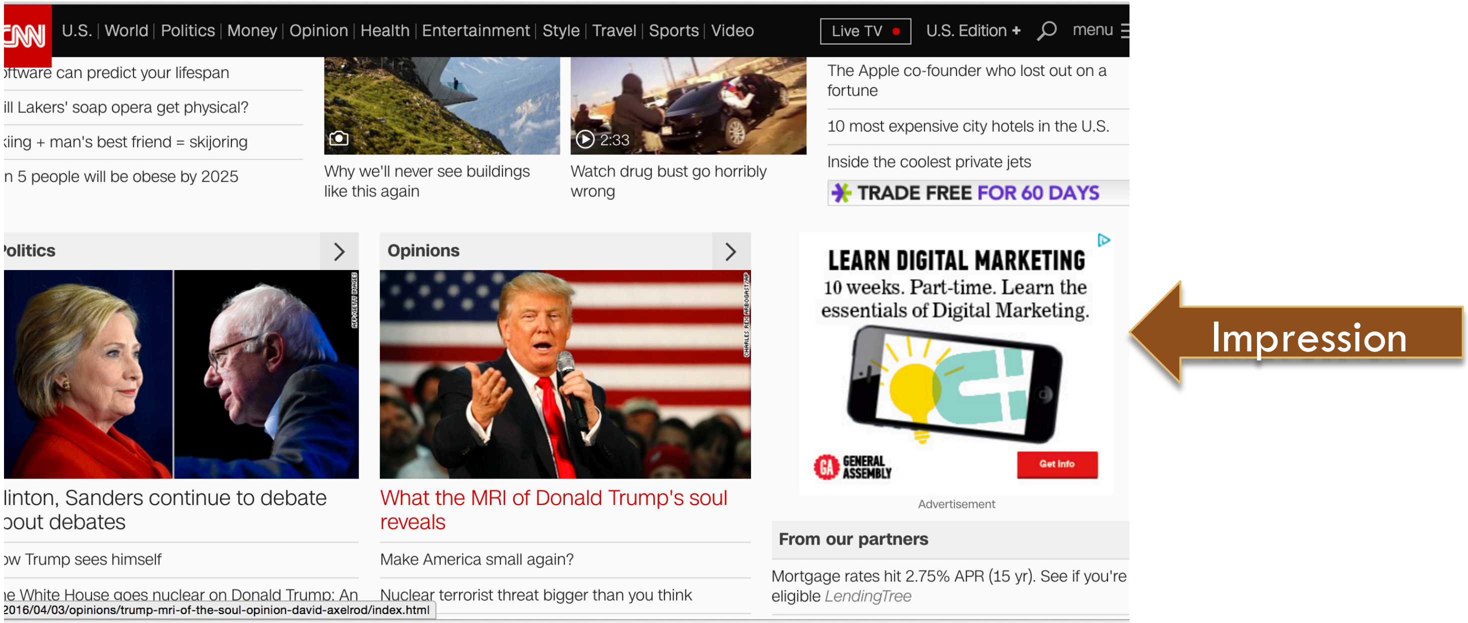
CONVERSION EVENT ATTRIBUTION BEHAVIOR IN PROGRAMMATIC DISPLAY ADVERTISING

SF-DAT-20 Final Project
Marlo Schneider

Hypothesis

- It is possible to interpret features that cause conversion events to be unattributed.
  - Attributed: The conversion event is tied to a banner ad impression that meets the conditions of the conversion pixel.
  - Non-Attributed: The conversion event is not tied to an impression or the impression does not meet the conditions.



Context

- Why does this matter?
  - Retention
  - Scale
  - Control
- User Level vs. Aggregate Causes
- Product Strategy and Communication

Data: Observations & Event Types

Request third party map for cookie sync

Anything defined as a conversion event by the advertiser is tracked as a unique event.

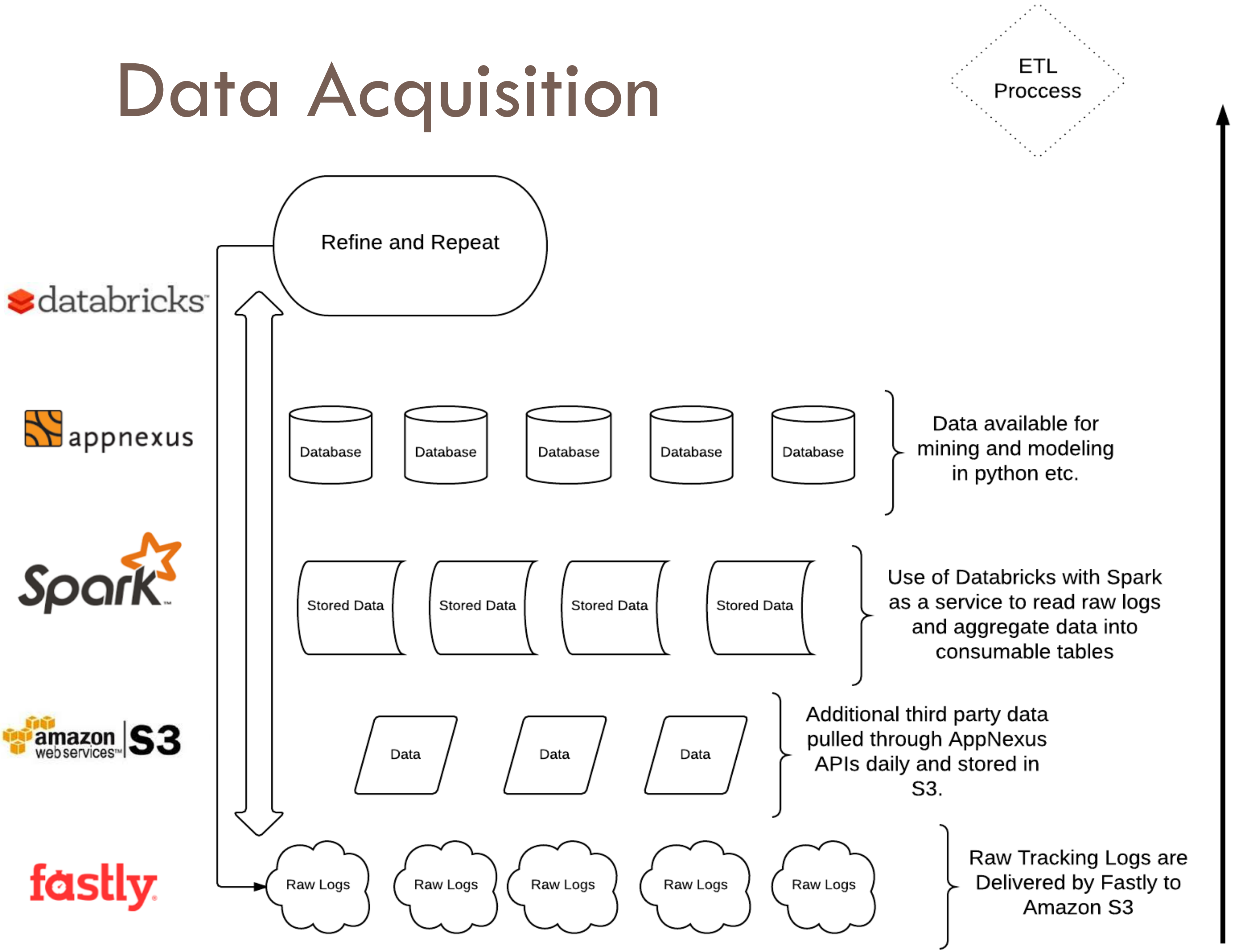
Every time a user visits an advertisers website we track each page load as an event.A unique identifier ("cookie") is stored in their browser.

While browsing external website, a banner impression is served. This is logged as an event in our tracking.

Data Challenges

- A workable data set required de-duping and joining raw logs together which track different event types.
  - Nothing is aggregated
  - No 'master fact table' at the user level
  - *This was much more difficult and time consuming than anticipated*
  - *Third party tracking stored raw events, actual usable features had to be built so decisions made along the way were critical.*
- Nature of cookies results in multiple many to many relationships that needed to be reconciled during each table join.
- Null values (NMAR) represent the discrepancy this project aimed to interpret.
- Many different quantitative and qualitative data types and highly variable observation counts for each user.
- Sequence of events is important i.e., Time delta of different TimeStamps.
  - Problems: window of time we were looking at
  - Number of events is highly variable, how do you organize this?

Data Acquisition



Databricks Notebook

```
    StructField("version",StringType,true),
    StructField("longitude",StringType,true),
    StructField("latitude",StringType,true),
    StructField("city",StringType,true),
    StructField("continent_code",StringType,true),
    StructField("country_code",StringType,true),
    StructField("country_code3",StringType,true),
    StructField("country_name",StringType,true),
    StructField("postal_code",StringType,true),
    StructField("Region",StringType,true),
    StructField("area_code",StringType,true),
    StructField("metro_code",StringType,true),
    StructField("z",StringType,true)
  ))

def safeFormatDatestamp(datestamp:String):org.joda.time.DateTime = {
  try{
    DateTimeFormat.forPattern("EEE, dd MMM yyyy HH:mm:ss z").parseDateTime(datestamp)
  }catch{
    case i:java.lang.IllegalArgumentException => new DateTime("1970-01-01")
  }
}

val getSessionId = udf{(uri:String) => parseURI(uri,"s").getOrElse("")}
val getVisitId = udf{(uri:String) => parseURI(uri,"visitid").getOrElse("")}
val getEventTime = udf{(uri:String) => parseURI(uri,"dt").getOrElse("")}
```

Data Acquisition Cont'd…

Cookie Sync:

|-- cs_spid: string (nullable = true)

|-- cs_apnx_id: string (nullable = true)

|-- cs_sessionid: string (nullable = true)

Conv Pixel:

|-- date: string (nullable = true)

|-- conv_spid: string (nullable = true)

|-- conv_sessionid: string (nullable = true)

|-- orderId: string (nullable = true)

|-- orderValue: string (nullable = true)

|-- browser: string (nullable = true)

|-- os: string (nullable = true)

|-- device: string (nullable = true)

Final Conv:

|-- date: string (nullable = true)

|-- conv_spid: string (nullable = true)

|-- conv_sessionid: string (nullable = true)

|-- orderId: string (nullable = true)

|-- orderValue: string (nullable = true)

|-- browser: string (nullable = true)

|-- os: string (nullable = true)

|-- device: string (nullable = true)

|-- cs_apnx_id: string (nullable = true)

Last Imp:

|-- imp_apnx_id: string (nullable = true)

|-- imp_time: string (nullable = true)

|-- adv_fr: string (nullable = true)

Imp:

|-- imp_time: string (nullable = true)

|-- imp_apnx_id: string (nullable = true)

|-- adv_fr: string (nullable = true)

|-- imp_auction_id: string (nullable = true)

|-- creative_id: string (nullable = true)

Attribution:

|-- att_apnx_id: string (nullable = true)

|-- att_order_id: string (nullable = true)

|-- post_click_or_post_view_revenue: string (nullable = true)

|-- att_auction_id: string (nullable = true)

|-- imp_conv_minute_diff: long (nullable = true)

|-- imp_time: string (nullable = true)

|-- adv_fr: string (nullable = true)

|-- creative_id: string (nullable = true)

Model: Logistic Regression

- 98.81% accuracy ☺…☹

```
from sklearn.cross_validation import cross_val_score
print(cross_val_score(lm,X,y,cv=10))
print(cross_val_score(lm,X,y,cv=10).mean())
```

```
[ 0.98389982  0.98568873  0.98389982  0.99283154  0.9874552   0.98387097
  0.98387097  0.99641577  0.98922801  0.994614   ]
0.988177482731
```

Coefficient Interpretation (After Tuning C)…

[(**-2.5053726581806166**, 'no_imp'),
(**-1.1355297446209049**, 'device_iPhone'),
(**-1.0400477808174831**, 'device_iPad'),
(-0.33167228295069495, 'imp_after'),
(-0.32495095070841951, 'imp_before'),
(0.2124694689533568, 'device_Gen_Smartphone'),
(0.36542793914625604, 'imp_freq_count'),
(1.3447032307203817, 'order_value'),
(**12.463610619185587**, 'imp_att'),
(**3594.0279907785866**, 'att_imp_fr')]

- An absence of impressions altogether had the highest negative impact on attribution
- Apple devices also had a negative impact (Safari cookie policies)
- The attributed impression frequency had the strongest positive effect, higher than the presence of an impression that was attributed (why?)

Insights and Next Steps

	attribution_bool	order_value	imp_freq_count	att_imp_fr	no_imp	imp_att	imp_before	imp_after
count	992	848.000000	992.000000	992.000000	992.000000	992.000000	992.000000	992.000000
mean	1	197.327347	26.197581	24.096774	0.009073	0.339718	0.364919	0.626008
std	0	235.508472	58.298782	56.563161	0.094865	0.473852	0.481650	0.484106
min	1	6.420000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1	68.700000	5.000000	2.000000	0.000000	0.000000	0.000000	0.000000
50%	1	131.525000	10.000000	7.000000	0.000000	0.000000	0.000000	1.000000
75%	1	229.667500	15.000000	19.000000	0.000000	1.000000	1.000000	1.000000
max	1	2280.360000	640.000000	639.000000	1.000000	1.000000	1.000000	1.000000

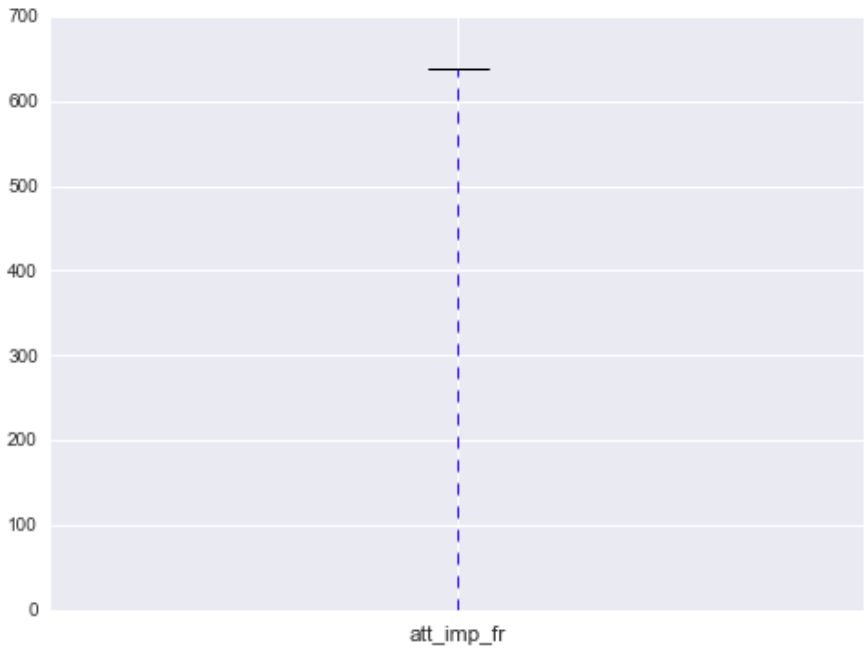
Median for Impression Frequency is only 10.

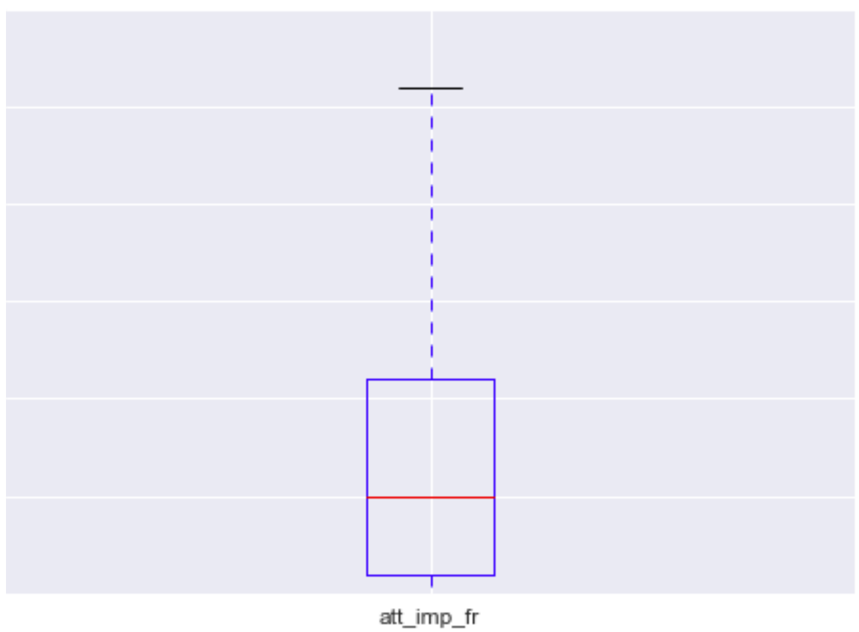
25% of attributed conversions have been exposed to 2 or less in their lifetime. 75% of these users is 19 or less.

There are some significant outliers (i.e., people who don't know what a cookie is and therefore don't clear them). Therefore, when running reports on the aggregate optimal frequency is going to be skewed.

Insights Cont'd…

Data set did have a very interesting distribution of **impression frequency** at the user level. This is important in order to re-define a better experiment.





New Hypothesis:
It is possible to predict whether a unique UID will be served an impression.

Next Steps…

Additional Features:

- Browser & OS
- Impression events (over the lifetime with TimeStamps)
- Page load activity over time
- Extend lookback window
- Cookie birthday
- Frequency and duration of page load sessions
- Day of Week data
- Page load to impression time delta
- Count of unique first and third party UIDs associated with a 'user'
- What segments users belong to

Short Term:

Set up auto-segmentation process that bids very high for the first impression, then removes user into a blocked segment for 7 days to test theory in real time.

Long Term:

Additional features would answer less obvious questions. i.e.,

- Is there a problem with our audience segmentation?
- Bidder logic?
- Impression tracking?
- Javascript errors in specific environments?
- Problematic distribution of impressions over time?

Broader Implications

- We’re tracking a lot of valuable information but do not have ETL processes in place so that it is accessible and actionable.
- We need to define the features, everything is raw and features don’t exist!
- Validating the right set of features and using this workflow will result in a recommendation for automating the process so that it’s scalable.
- Ad operations/Campaign Management teams need to work closer with Engineering and Data Science to provide subject mater expertise to guide feature development.

Thank You!