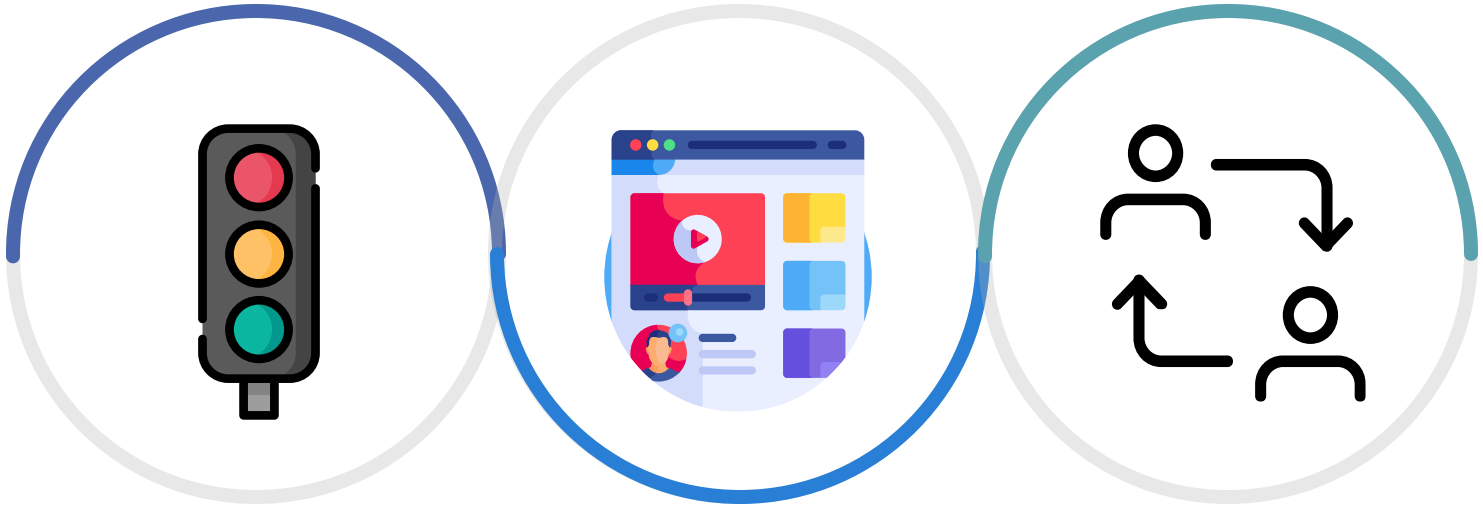
Channel NaN
Data Analytics Presentation

Winston Sun

Ellen Wang

Vincent Chen

Laurel Lu



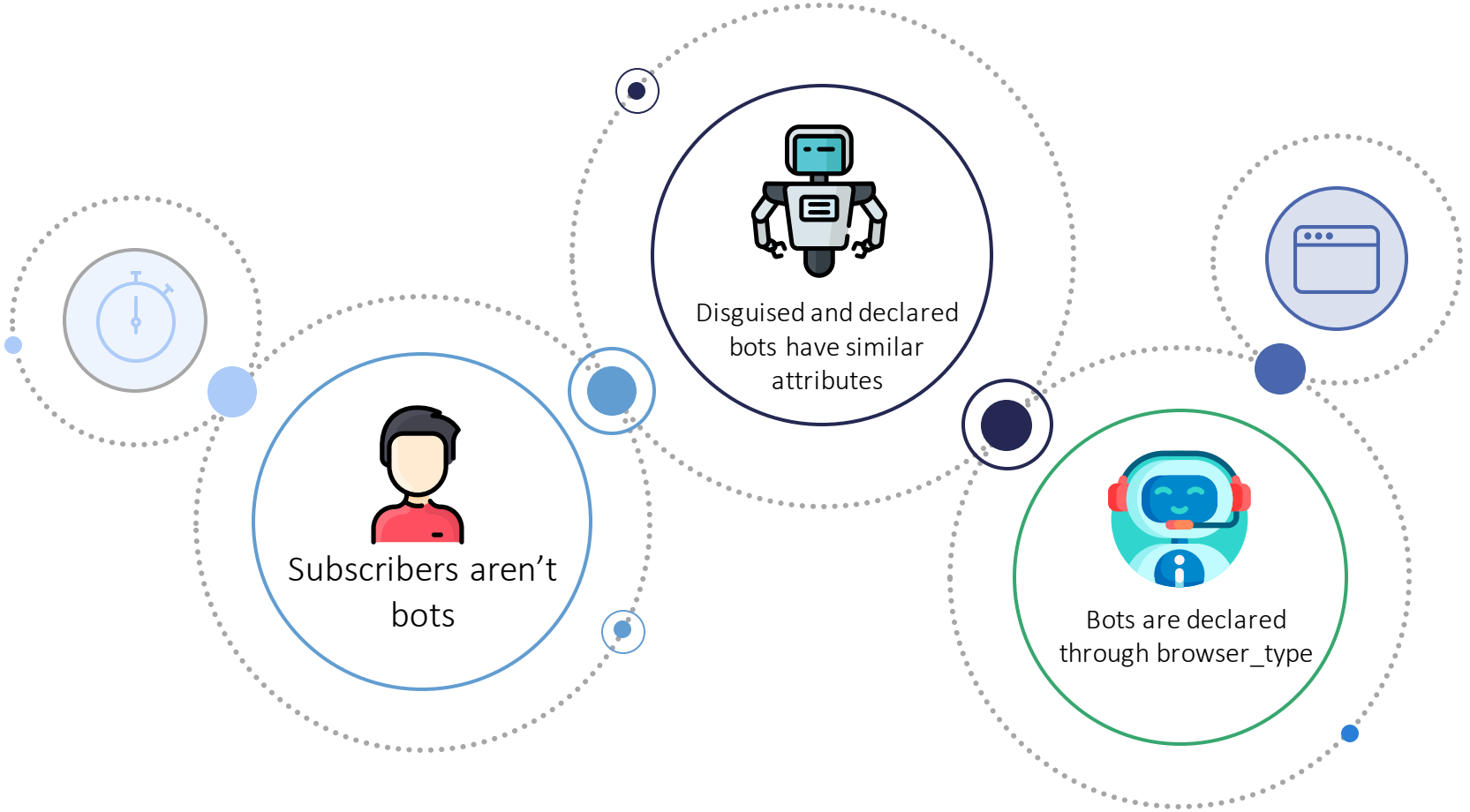
01 Disproportionately increase traffic on certain pages

02 Accessing large amounts of content very quickly

03 Inaccurately low conversion rates

The Problem

Web bots produce unusual levels of site activity which are inconsistent with consumer preferences



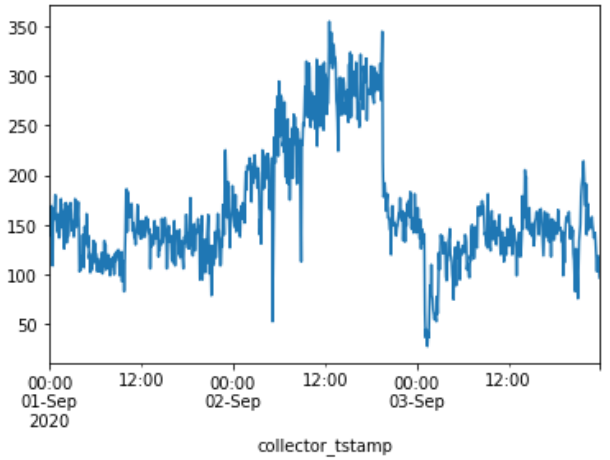
Assumptions

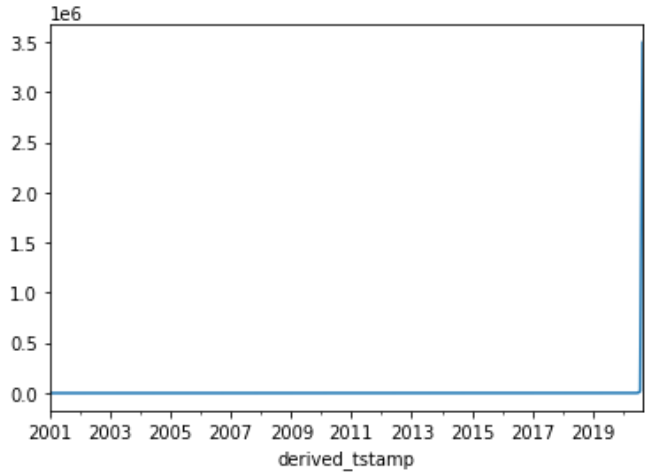
Selected features

Some users had very abnormal ‘time differences’

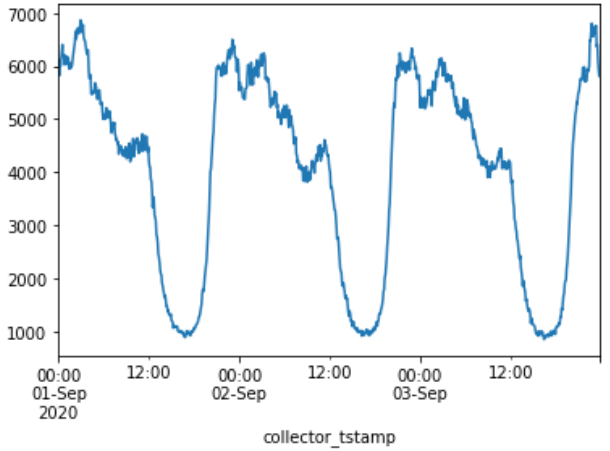
- (collector_tstamp - derived_tstamp)

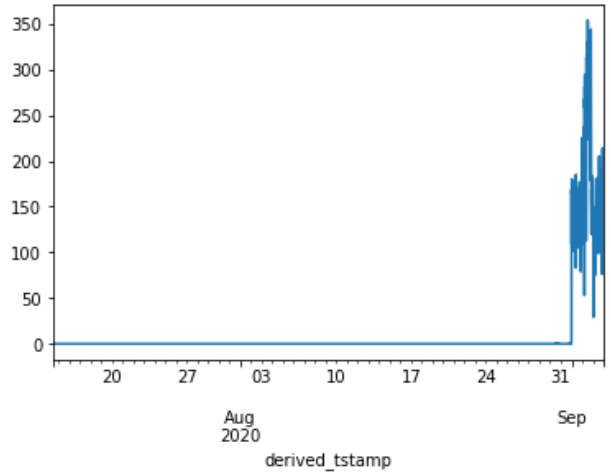
Collector & Derived Timestamp of Non Declared Bots





Collector & Derived Timestamp of Declared Bots





Selected features

Bots visit a lot more pages which humans do not

- INDEX Pages vs. Actual articles

```
data_bot.article_type.value_counts()
```

```
INDEX              107681
ARTICLE             33838
HOMEPAGE             1256
VIDEO                 413
FEATUREARTICLE        337
LIVE ARTICLE          308
PHOTOGALLERY          154
ERROR                 117
SEARCH                 31
SUBSCRIPTION           14
BUNDLES                 8
PUZZLE                  2
SOUNDSLIDE              1
BESPOKE                 1
Name: article_type, dtype: int64
```

```
non_data_bot.article_type.value_counts()
```

```
HOMEPAGE                2020762
ARTICLE                 1096528
LIVE ARTICLE             192463
INDEX                    115118
VIDEO                     31285
FEATUREARTICLE            29747
BESPOKE                   10056
BUNDLES                    2742
SEARCH                     2416
PUZZLE                     2128
ERROR                      1509
PHOTOGALLERY               1265
LOGIN                      1096
SUBSCRIPTION                777
SUBSCRIPTIONS               564
CONCIERGE                   510
PROFILE DETAILS             458
SUBSCRIPTION DETAILS        350
SUBSCRIPTION FUNNEL         327
BILLING INFORMATION         229
SIGNUPS                     198
PAYMENT METHOD              193
LOGIN DETAILS               173
```

```
data_bot.page_urlpath.value_counts()
```

```
/coronavirus-pandemic                                                                                                  2455
/national/queensland                                                                                                   2332
/world/europe                                                                                                          2319
/politics                                                                                                              2316
/politics/western-australia                                                                                            2311
                                                                                                                       ...
/business/workplace/bull-rider-claims-workers-compensation-after-severe-injury-at-rodeo-20170612-gwpdis.html             1
/national/western-australia/como-farmer-jack-s-development-saga-twist-as-councillor-pushes-fresh-traffic-count-20190723-p529wx.html    1
/business/small-business/employee-or-contractor-the-grey-area-that-can-turn-nasty-20180627-p4zo39.html                 1
/world/nato-weathers-criticism-after-bombing-gaddafi-family-20110501-1e38f.html                                        1
/entertainment/art-and-design/eminently-outrageous-20030203-gdg7kl.html                                                1
Name: page_urlpath, Length: 29559, dtype: int64
```

```
non_data_bot.page_urlpath.value_counts()
```

```
/
2024732
/national/coronavirus-updates-live-victorian-covid-19-cases-in-decline-as-health-alerts-issued-for-multiple-sydney-venues-australian-death-toll-stands-at-652-20200831-p
55r0l.html      53615
/national/coronavirus-updates-live-victorian-cases-continue-to-decline-as-australia-officially-enters-recession-australian-death-toll-stands-at-663-20200902-p55rre.html
51899
/national/coronavirus-updates-live-victorian-state-of-emergency-extended-two-sydney-schools-close-hotel-quarantine-inquiry-resumes-20200902-p55rip.html
49850
/national/nsw/transport-blueprint-reveals-two-new-metro-lines-for-sydney-s-east-20200901-p55rds.html
26050

...
/lifestyle/fashion/why-were-obsessed-with-what-women-on-tv-wear-20160315-gnjpfi.html
1
/hsc-study-guide-2020&clkt=11
1
/business/companies/woolworths-and-caltex-in-marriage-of-convenience-20180705-p4zpme.html
1
/entertainment/movies/im-strong-im-ready-to-take-my-clothes-off-emily-browning-20110601-1ffl9.html
1
/entertainment/abc2-screens-best-tv-show-ever-20090901-f5yq.html
1
Name: page_urlpath, Length: 66428, dtype: int64
```

Selected features

- Declared bots always had **the same domain_userid**
- Bots always had **NaN os_name and os_family**
- Bots always had a **NaN browser version**
- Defined bots were **never subscribers**

```
1 bots_data = pd.DataFrame(data.query("br_family == 'Robot/Spider'"))
2 bots_data.br_version.value_counts
```

```
<bound method IndexOpsMixin.value_counts of 22575      NaN
22576      NaN
22577      NaN
22578      NaN
22930      NaN
          ...
```

```
1 bots_data.member_type.value_counts()
```

```
Non-Subscriber    144163
Name: member_type, dtype: int64
```

```
1 not_bots_data.member_type.value_counts()
```

```
Non-Subscriber    2474647
Subscriber        1040596
Name: member_type, dtype: int64
```

```
77.0.3833.0              1
68.10.0                  1
Name: br_version, Length: 954, dtype: int64
```

Classifier Model

Coefficients and Features of Logistic Regression:

domain_userid: -9.29

br_version: -12.88

member_type: 5.63

article_type_index: -0.55

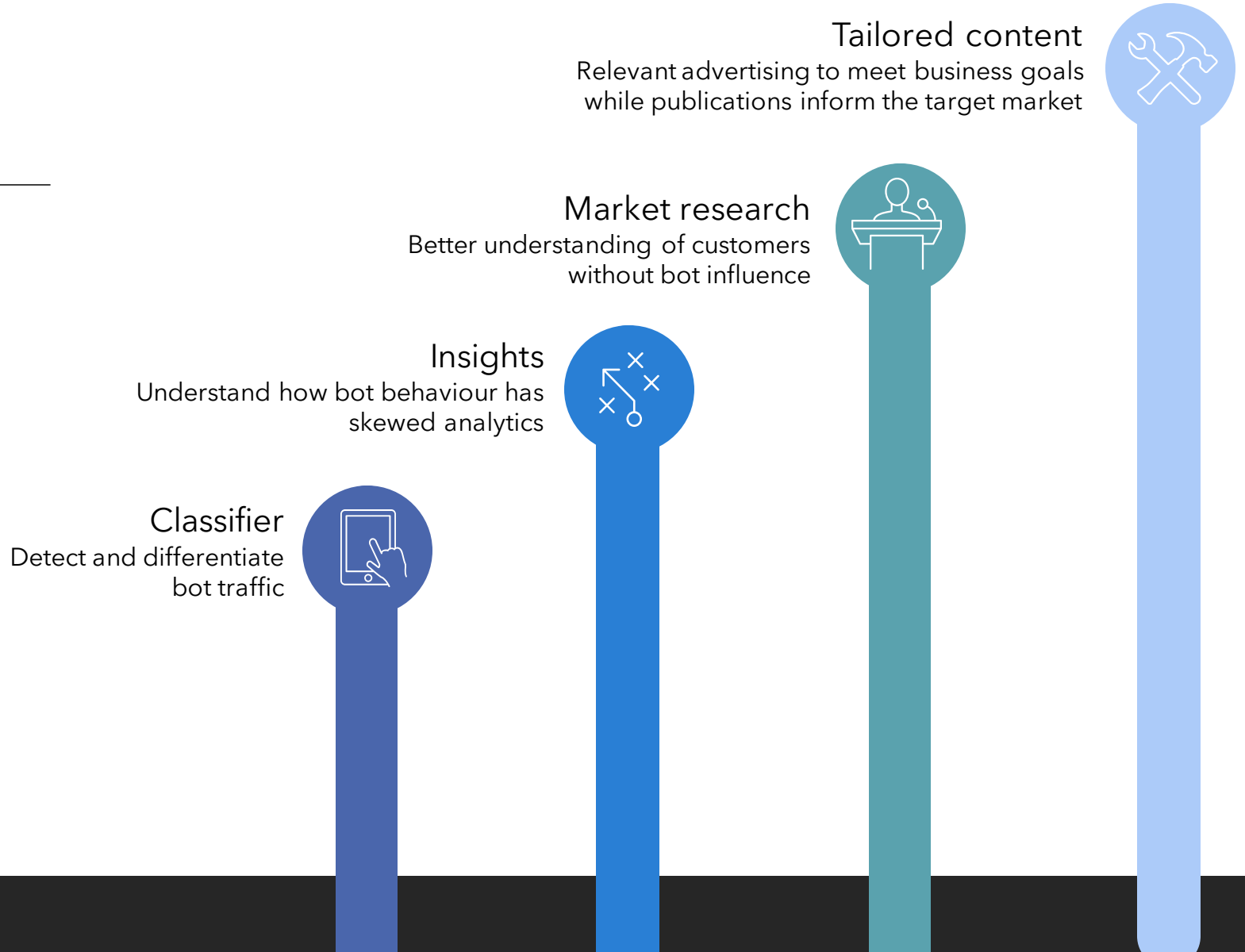
article_type_homepage: 3.74

```
#test_data copy
test_data = test_df.copy()
test_data.drop(columns=["page_urlscheme", "page_urlhost", "page_urlport","refr_urlscheme", "refr_urlhost", "refr_urlport", "refr
test_data["br_version"] = test_data["br_version"].fillna(-1)
test_data["page_view_id"] = test_data["page_view_id"].fillna(0)
test_data["article_type"] = test_data["article_type"].fillna("ERROR")
test_data['domain_userid'] = np.where(test_data['domain_userid'] == "84896d88-79b6-4a07-92a4-82a6352fa98d", 1, 0)
test_data.drop("domain_sessionid", axis=1, inplace=True)
test_data.drop("user_ipaddress", axis=1, inplace=True)
test_data.drop("useragent", axis=1, inplace=True)

#os_name
test_data['os_name'] = np.where(test_data['os_name'] == "unknown", 1, 0)
#os_family
test_data['os_family'] = np.where(test_data['os_family'] == "unknown", 1, 0)
#br_version
test_data['br_version'] = np.where(test_data['br_version'] == -1, 1, 0)
#page_urlpath
test_data['path_len'] = test_data['page_urlpath'].apply(lambda x: len(x))
test_data['path_len_article'] = np.where(test_data['path_len'] > 40, 1, 0)
#member_type
test_data['member_type'] = np.where(test_data['member_type']=="Subscriber", 1, 0)
#article_type
test_data['article_type_index'] = np.where(test_data['article_type']=="INDEX", 1, 0)
test_data['article_type_article'] = np.where(test_data['article_type']=="ARTICLE", 1, 0)
test_data['article_type_homepage'] = np.where(test_data['article_type']=="HOMEPAGE", 1, 0)
test_data.drop(columns=["derived_tstamp", "collector_tstamp", "page_urlpath", "page_view_id", "article_type", "path_len_article"
final_test = test_data
```

Impacts

- Advertising
- Subscriber conversion



QnA session