Attack of the Android Zombies

A tour of the commercial web, digital ad fraud, big data and data science at comScore-Madison

Jeff Kline
jkline@comscore.com

Madison Big Data Meetup, October 2015

Overview

A talk with two parts.

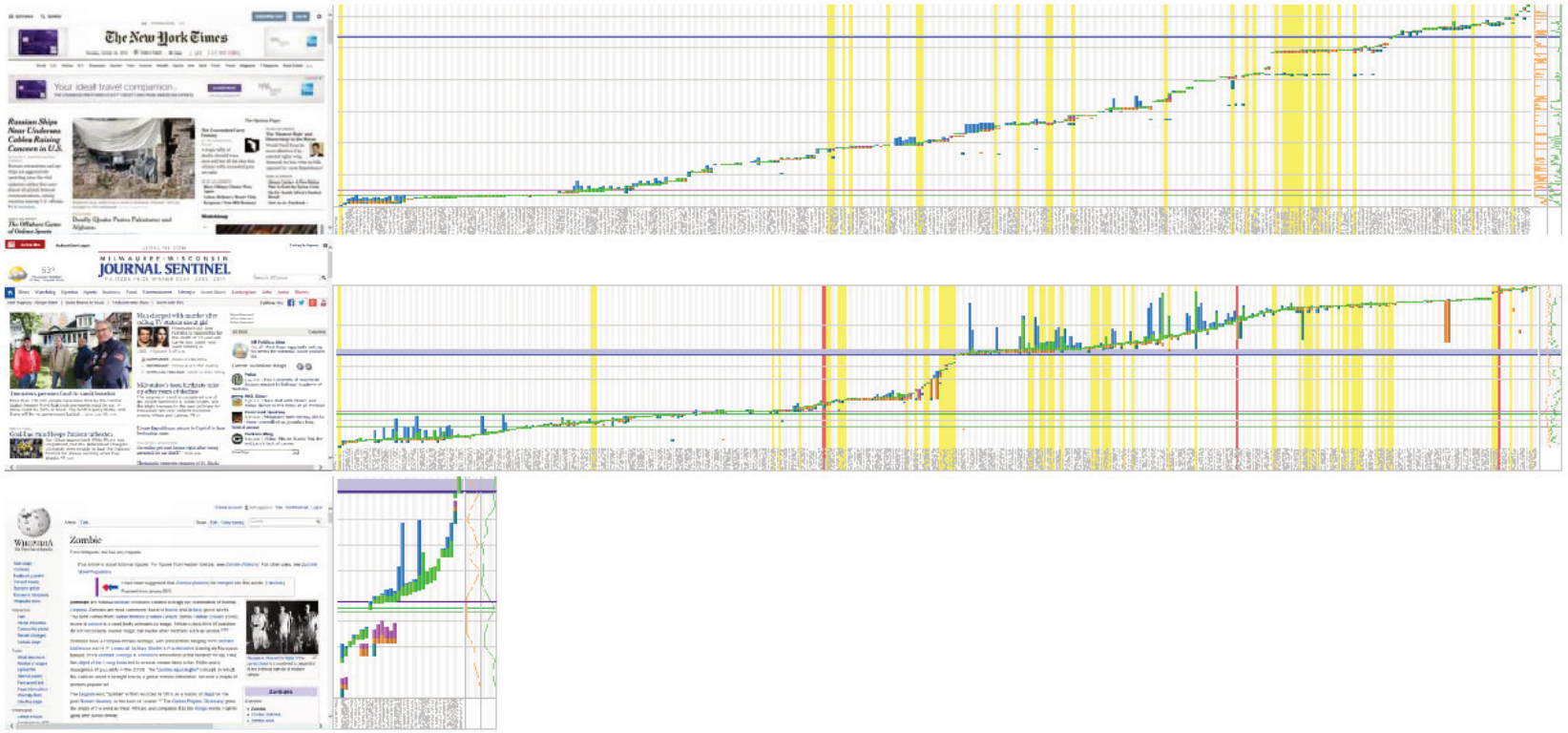
1. An introduction to comScore and the commercial web
   - A brief introduction to the online publishing world
   - It's complex and it evolves quickly
2. We saw Android zombies last summer. Sometimes data contradicts common sense. We must deal with this.

Introduction

Some information about me.

- Ph.D. in Math from UW Madison (Amos Ron)
- Background includes convex optimization, NMR spectroscopy, the LIGO Scientific Collaboration, and networking theory and practice.
- In 2013, Paul Barford contacted me about his new startup MdotLabs. Mdot was co-founded by Timur Yarnall.
- MdotLabs was acquired by comScore (August 2014).

The commercial web is complex

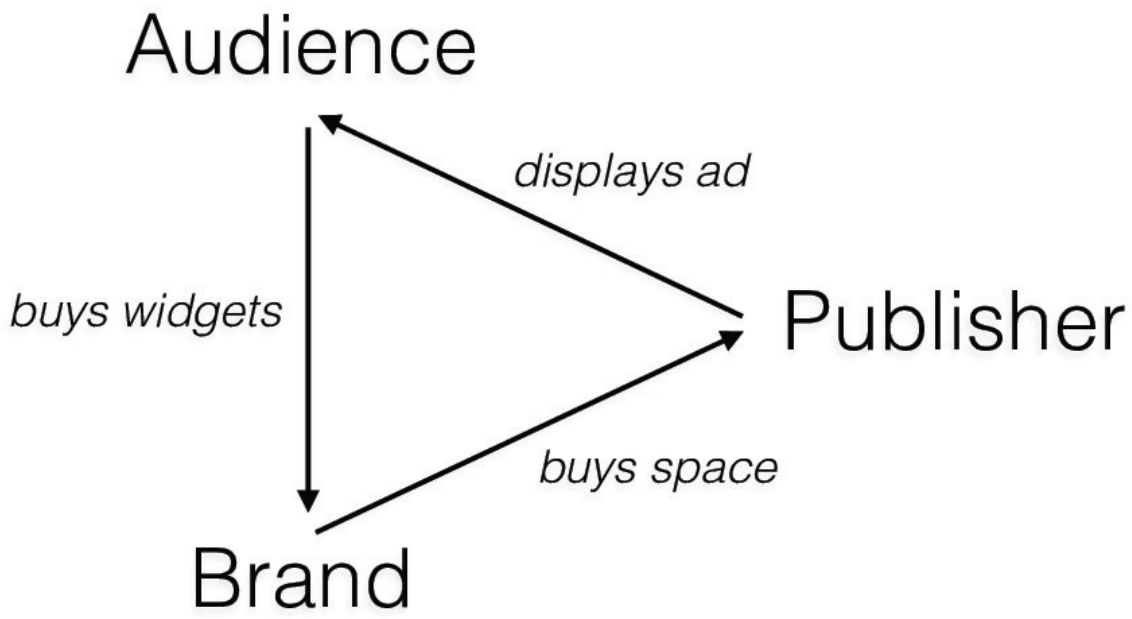


Source: http://www.webpagetest.org

comScore

A simplified model of publishing

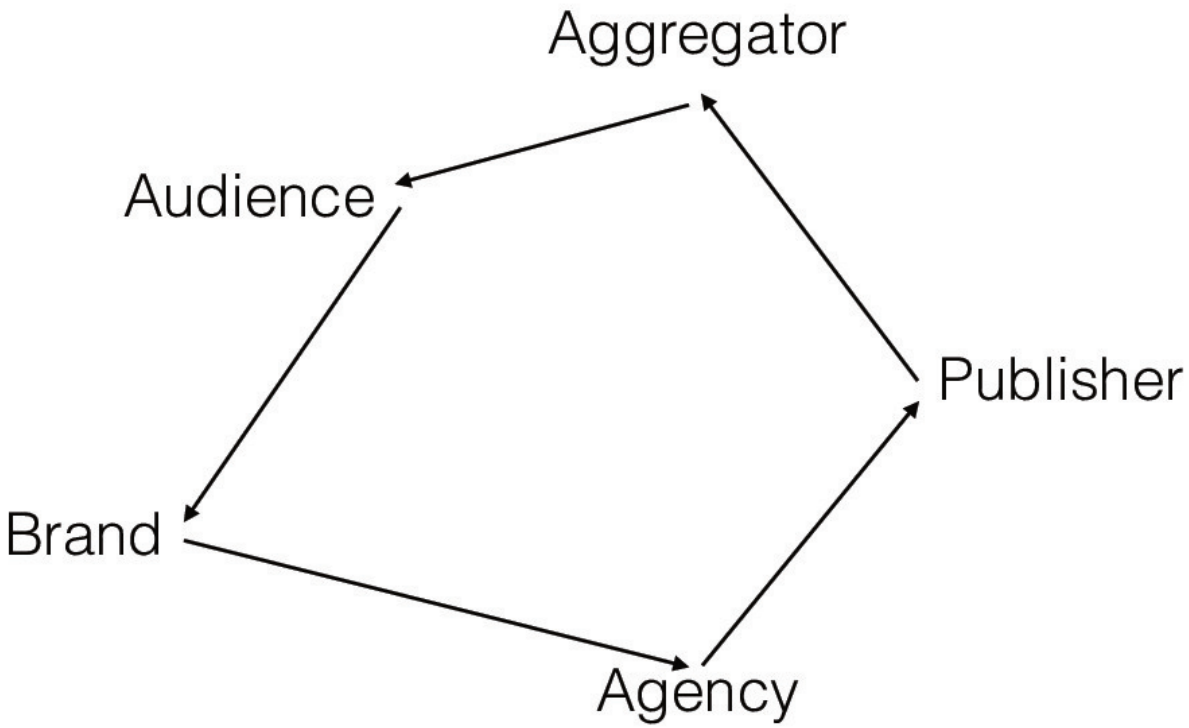
The virtuous cycle of advertising



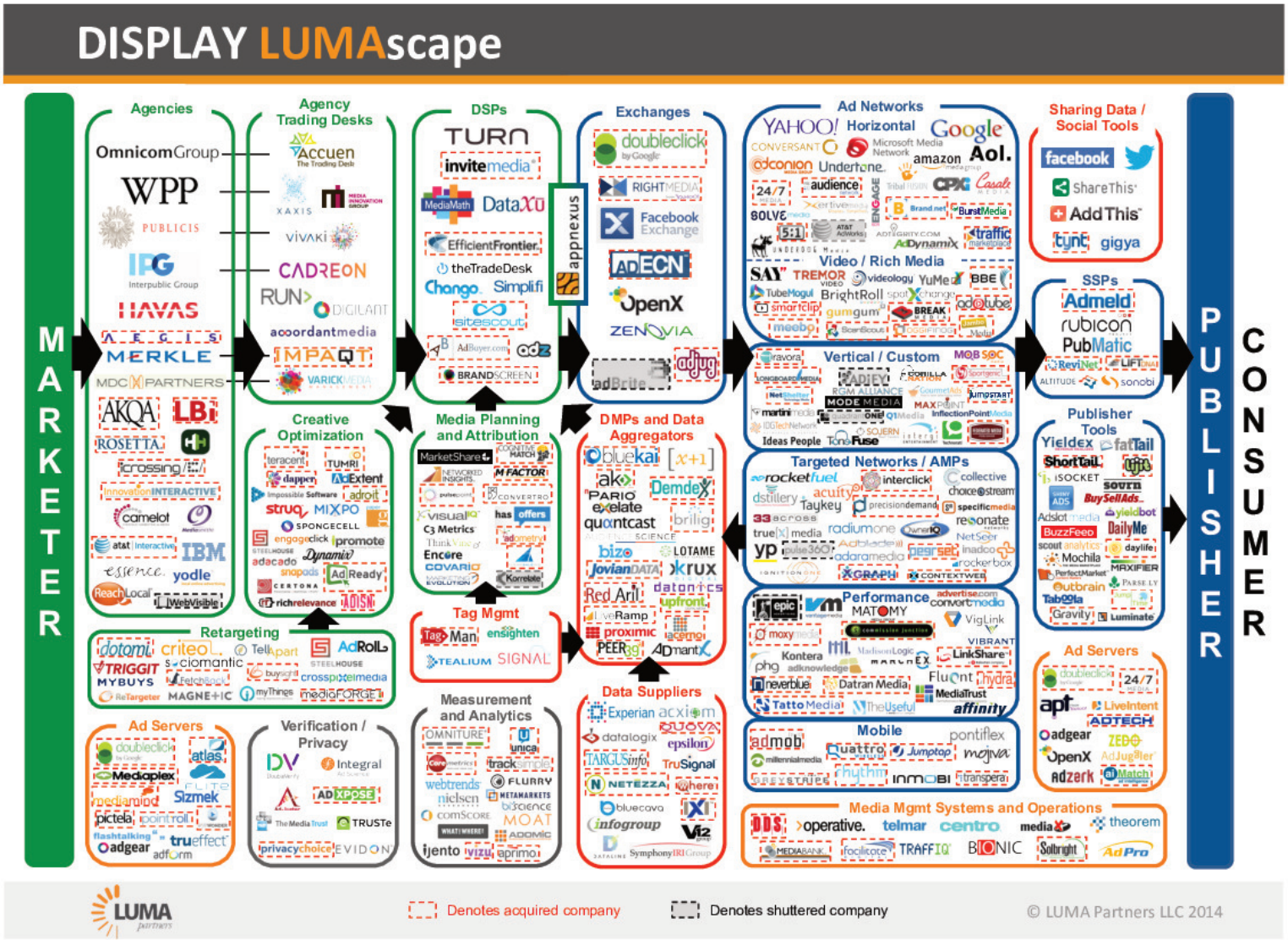

```
while True:
  brands pay publishers for space on their pages
  publisher produces content which attracts an audience
  the audience buys items advertised
```

comScore

A more realistic model of publishing

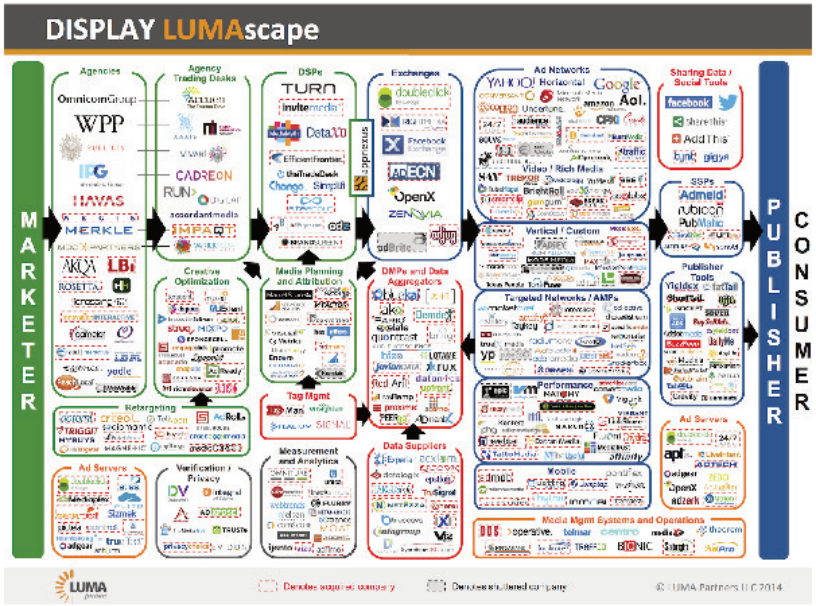


A map, c. 2014



comScore

The map was obsolete before it was finished



This map has symbols to indicate the churn.

Denotes acquired company

Denotes shuttered company

The commercial web changes quickly due to audience habits, social trends and technological innovation.

comScore

Background

What is comScore's role?

comScore acts as a trusted third-party to describe audiences for publishers and brands.

comScore also reports on other aspects of online audiences such as mobile device market share and App use.

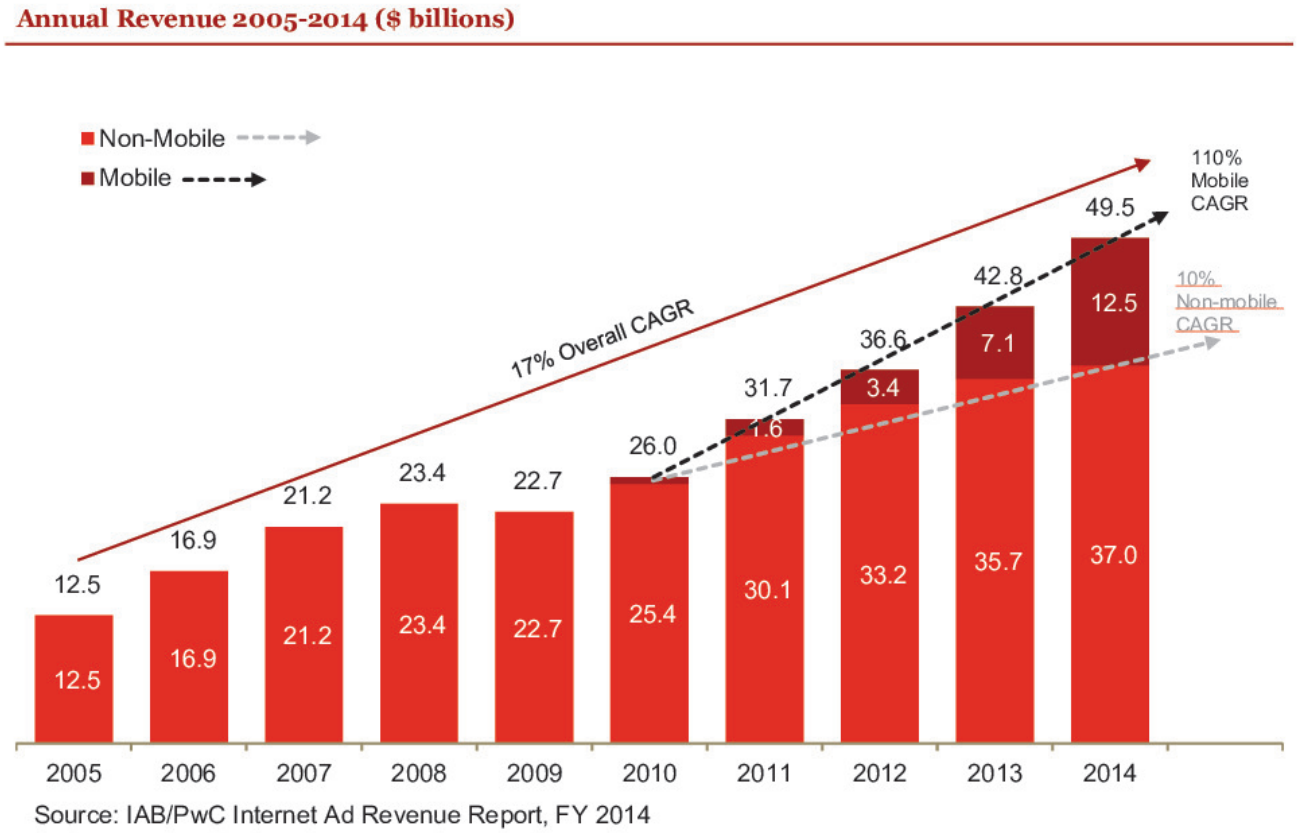
Publishers and brands partner with us, their audiences send us pixel requests.

Background of comScore Madison

Publishers fear clawbacks caused by invalid traffic.

Timur Yarnall and Prof. Paul Barford (University of Wisconsin, CS) co-founded MdotLabs in 2013 to act as a service that would offer insight into traffic health. It was acquired in 2014 by comScore.

Annual revenues of online advertising over time

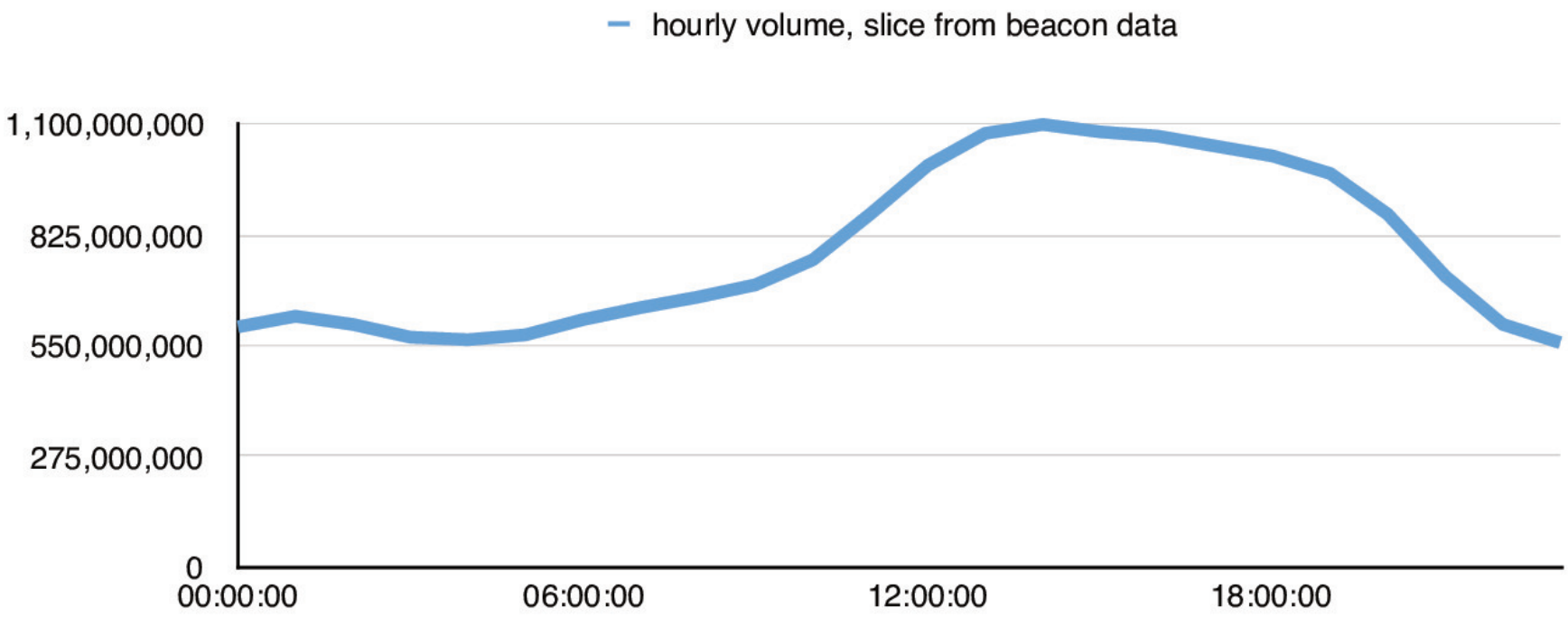


* Source for GDP growth: U.S. Bureau of Economic Analysis, "Table 1.1.5. Gross Domestic Product," (accessed March 31, 2015)

Source: IAB/PWC

comScore

Our Data



Each record counted in this timeseries represents an HTTP request.

comScore

About our data

- our primary data store manages about 60B records/day
- our live storage capacity is O(PB)
- most of this data is derived from HTTP events
- we also collect *panelist* data. The panel consists of about 3 million members. It is a unique data set that offers a clear view of client-side behavior.

About our software stack

- we have several Hadoop clusters
- we run Greenplum (SQL-based access with postgres UI)
- in-house developers build data visualization and access tools
- we also use Python (Numpy, Pylab), graphviz, and the classical command line tools such as awk, grep, wc and sort.

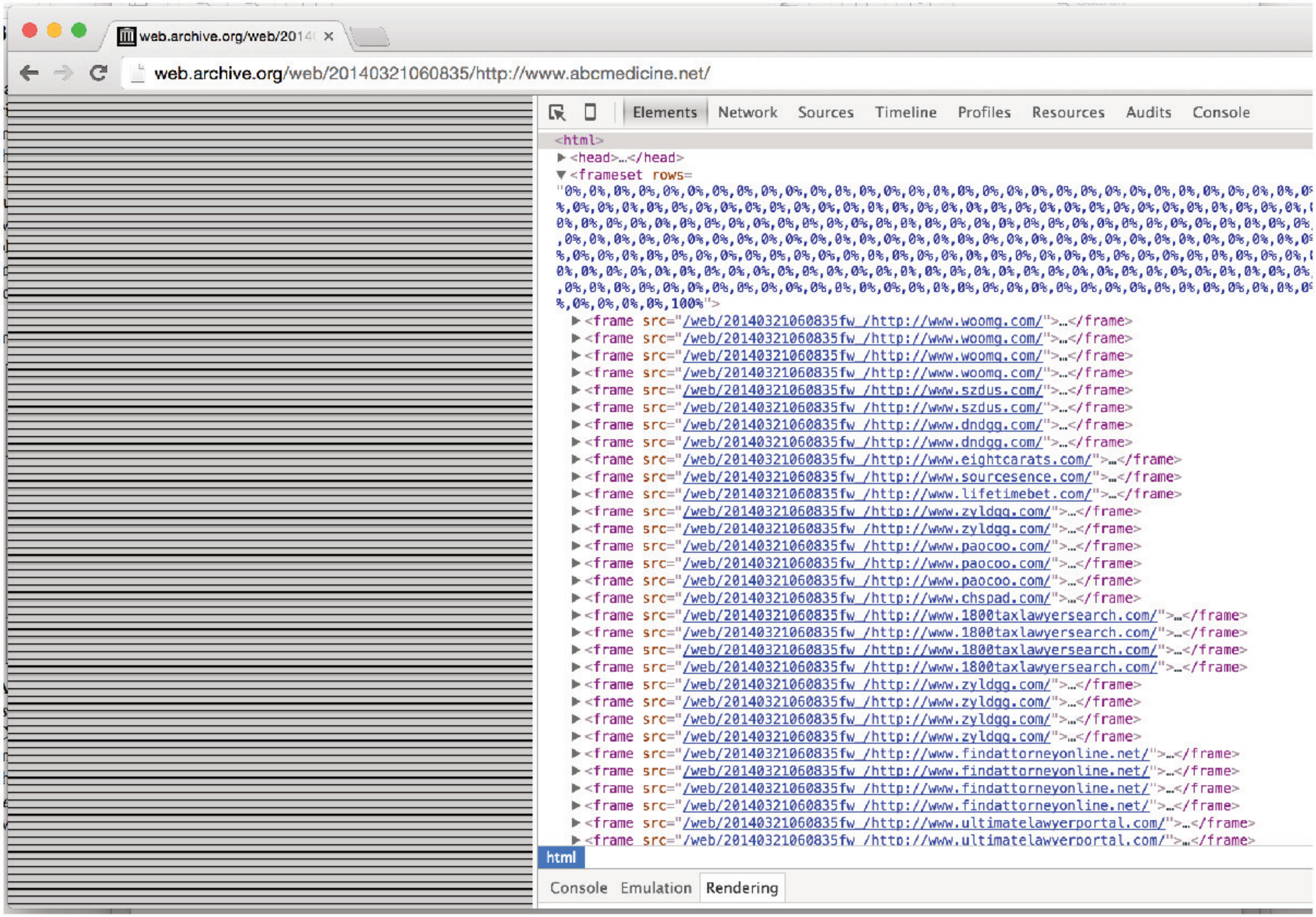
Commercial web traffic is really messy

Measurements contain the results of

- human error (bad tagging, wrong data type, apps, publishers, ads, mobile, etc)
- technical problems (misconfiguration, hardware failure)
- benign activity (developer testing, crawlers, "reviewing services")
- maliciousness (malware, DOS attacks, profit)

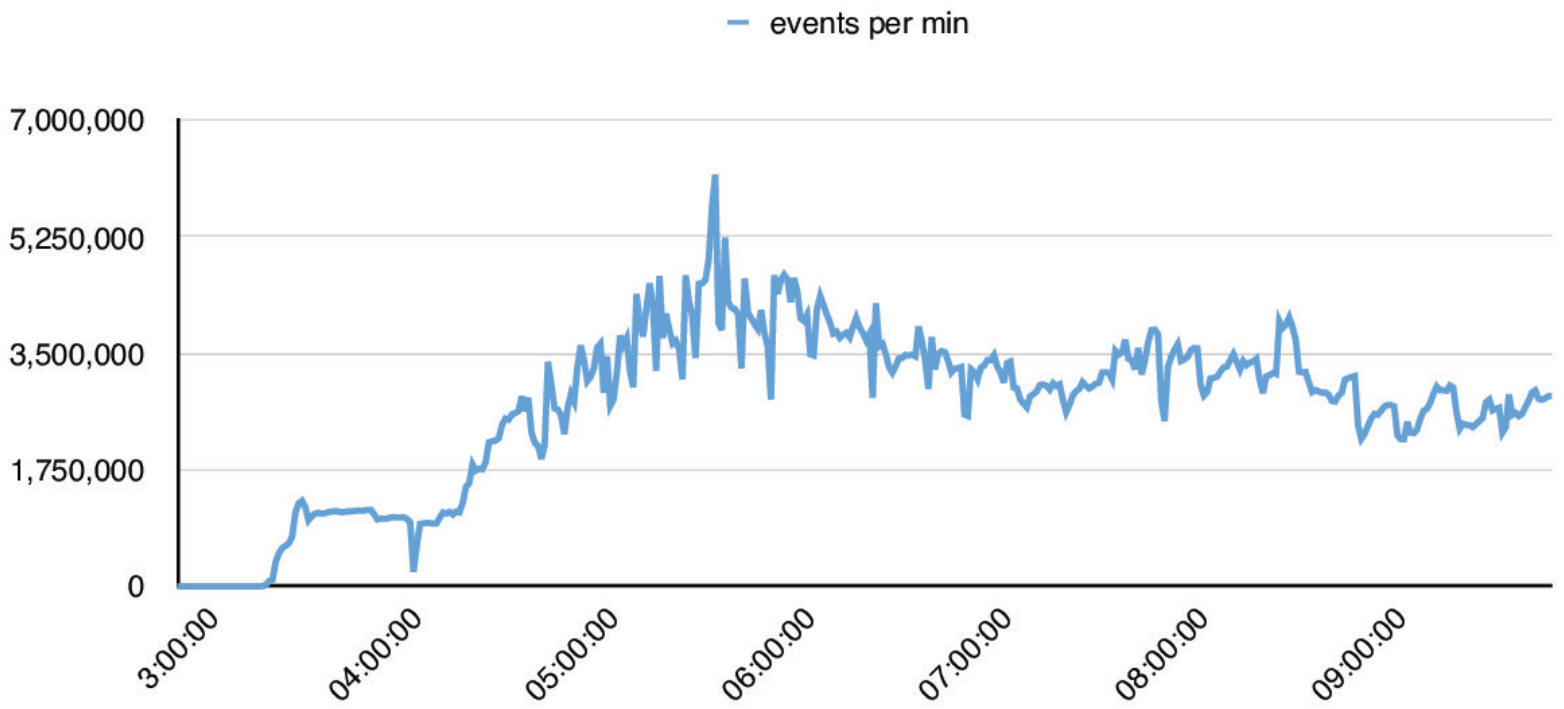
Let's see some examples!

A malicious example



Don't go here with a browser that you like.

A client's error resulted in this DOS attack

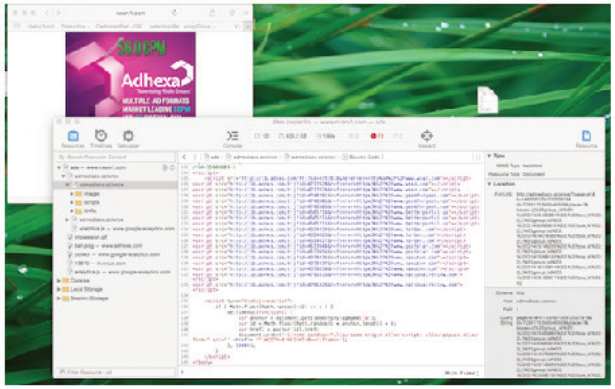


Above is shown the initial hours of a DOS attack caused by a deployment error. This event peaked at several billion records per day. This was deemed invalid traffic and was excluded from our reporting.

comScore

Placement laundering

Browser and source view

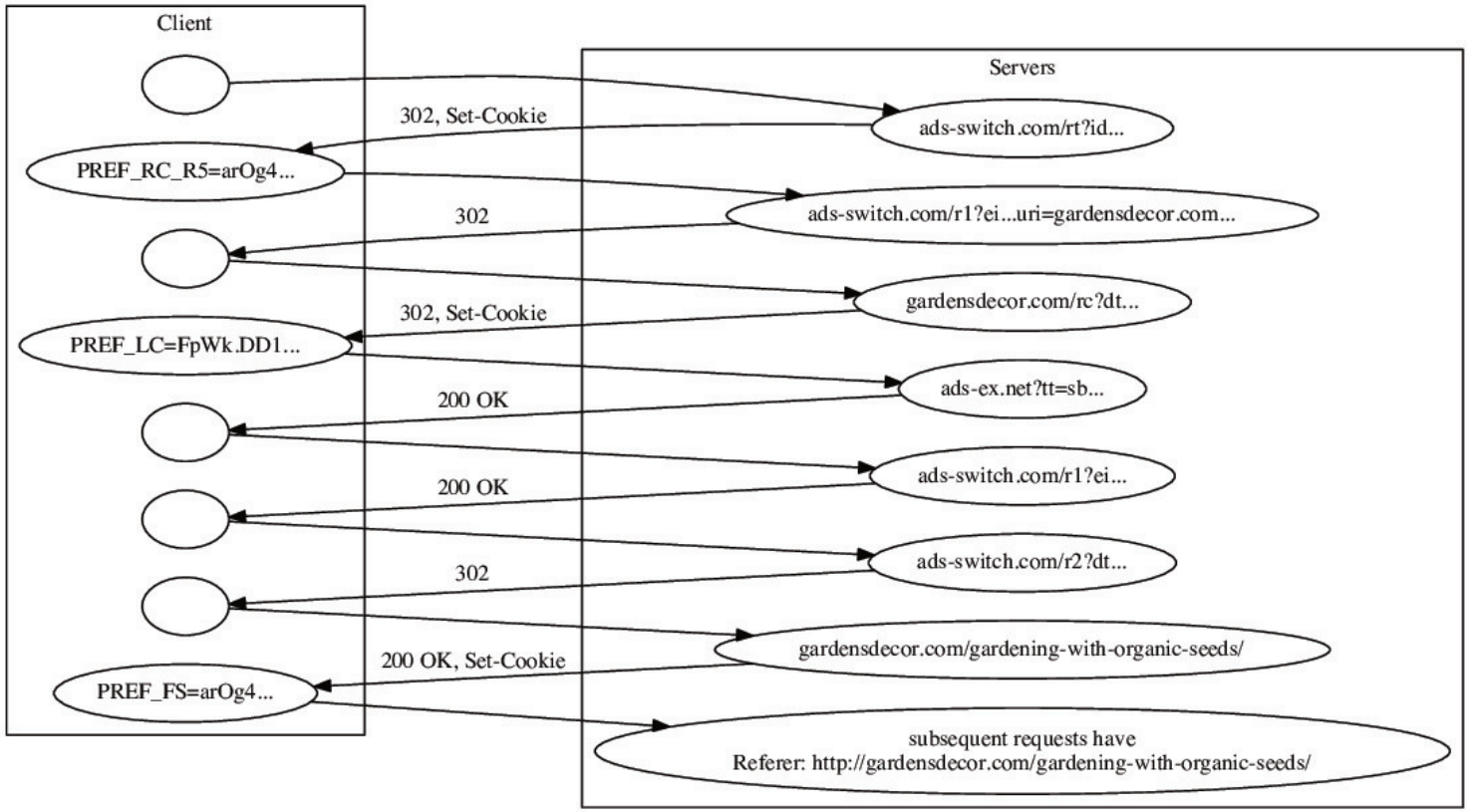


Zoomed in view of source

```
<script src="http://ib.adnxs.com/ttj?id=4565715&referrer=http%3A%2F%2Fwww.betterportal.us"></sc
<script src="http://ib.adnxs.com/ttj?id=4565716&referrer=http%3A%2F%2Fwww.betterportal.us"></sc
<script src="http://ib.adnxs.com/ttj?id=4324424&referrer=http%3A%2F%2Fwww.forbes.com"></script>
<script src="http://ib.adnxs.com/ttj?id=4324426&referrer=http%3A%2F%2Fwww.forbes.com"></script>
<script src="http://ib.adnxs.com/ttj?id=4324427&referrer=http%3A%2F%2Fwww.forbes.com"></script>
<script src="http://ib.adnxs.com/ttj?id=4613944&referrer=http%3A%2F%2Fwww.poststar.com"></scrip
<script src="http://ib.adnxs.com/ttj?id=4613956&referrer=http%3A%2F%2Fwww.poststar.com"></scrip
<script src="http://ib.adnxs.com/ttj?id=4378436&referrer=http%3A%2F%2Fwww.rakuten.com"></script
<script src="http://ib.adnxs.com/ttj?id=4378437&referrer=http%3A%2F%2Fwww.rakuten.com"></script
<script src="http://ib.adnxs.com/ttj?id=4378439&referrer=http%3A%2F%2Fwww.rakuten.com"></script
<script src="http://ib.adnxs.com/ttj?id=4796316&referrer=http%3A%2F%2Fwww.nationalreview.com">
```

Placement laundering

Schemes can be highly sophisticated.



comScore

Industry Standards

Guidelines describe in technical detail what is a page view, what is an impression, etc. Oversight organizations include the Media Rating Council and Interactive Advertising Bureau.

But sometimes the data make no sense. There's no rulebook for this.

It is the data scientist's role to develop principled methodologies for dealing with this.

It is now time to talk about zombies.

How to spot a zombie

This is the zombie user agent:

```
Mozilla/5.0 (Android; U; Android 2.1; en-us;) \
        AppleWebKit/525.10 (KHTML, like Gecko)
```

About the user agent

The user agent is part of the HTTP specification. Its intent is to facilitate traffic measurement. User agents take many forms. There are more traditions here than rules. Common user agents:

```
Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 \
              (KHTML, like Gecko) Chrome/41.0.2228.0 \
              Safari/537.36

Mozilla/5.0 (Windows NT 10.0) AppleWebKit/537.36 \
              (KHTML, like Gecko) Chrome/39.0.2171.71 \
              Safari/537.36 Edge/12.0

Dalvik/1.6.0 (Linux; U; Android 4.3; R8007 Build/JLS36C)
```

Uncommon user agents

```
Hello World
```

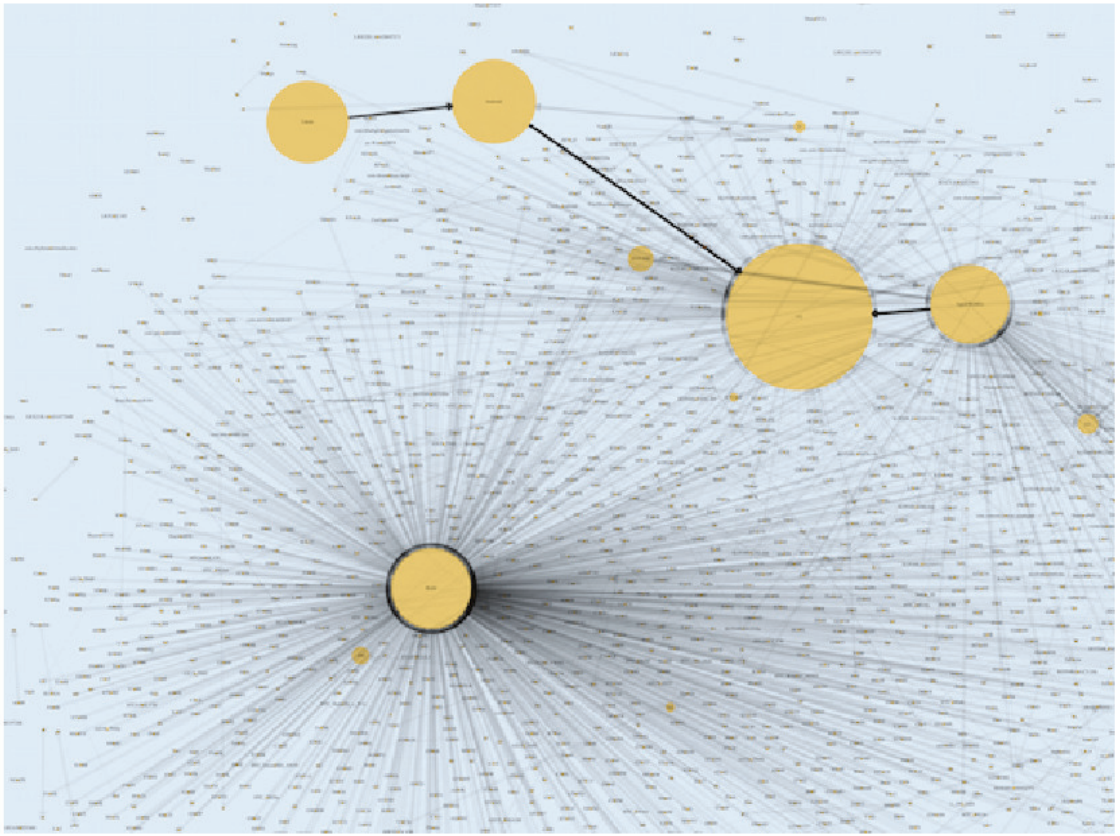
```
Mozilla/4.0...\(((|))))))^\[>\]>\>\\))))...
```

φ

It is common to see things like this.

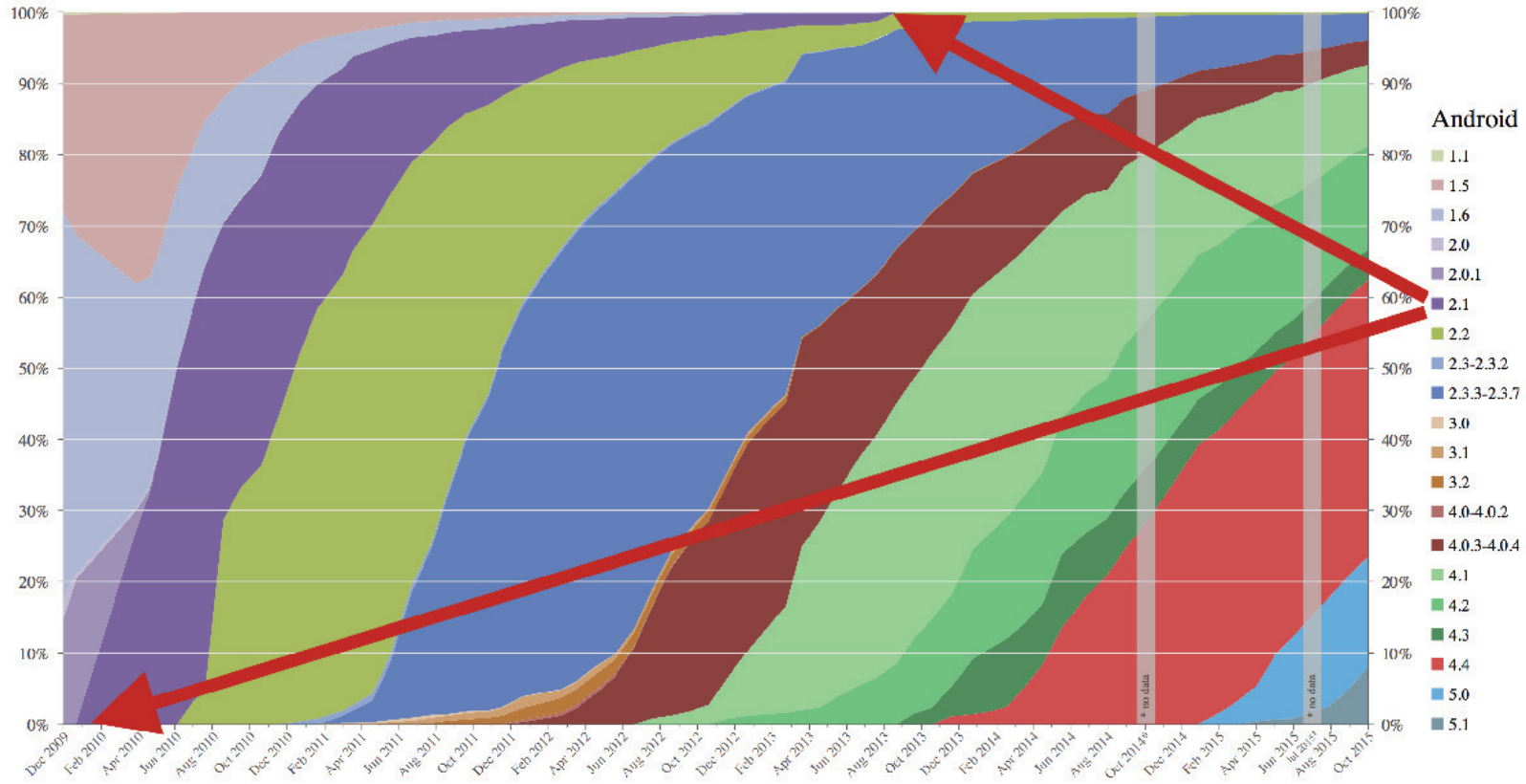
The space of user agents

The space of user agents is complex. Following is a visualization of relatively popular mobile UA's.



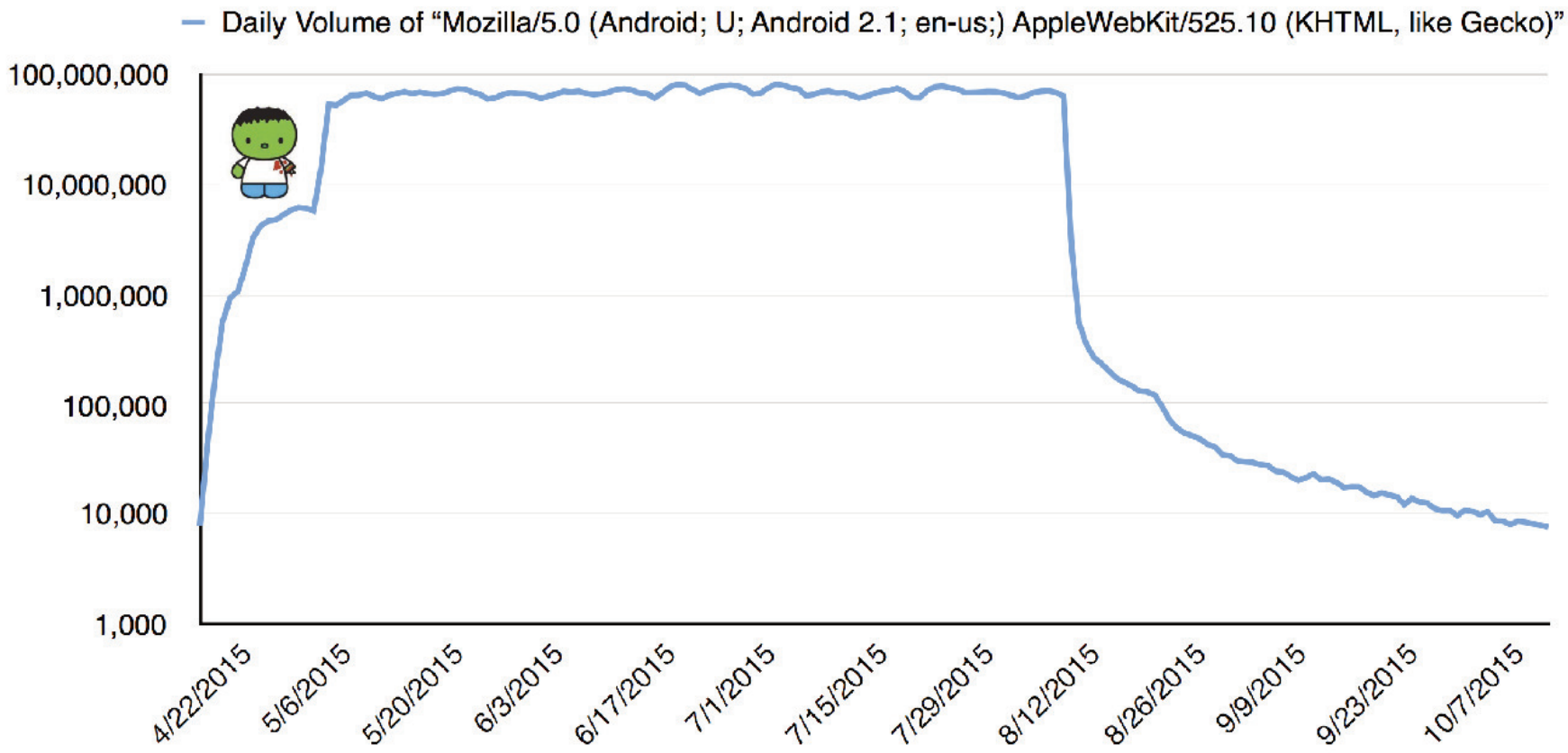
A little Android history

Android 2.1 died in 2013. Wikipedia charts their demise.



But. . .

comScore



This was first observed by C. Shuck. He wrote of "...billions of hits ...from a UA that should have gone the way of the Dodo bird..."

comScore

Zombies

What is a data scientist to do?

It is the data scientist's role to apply principled methodologies to situations like this.

Conjectures and refutations

1/4

C This is not real. It isn't on Wikipedia. It is a telemetry error.

R We cross-validated primary data with other data assets.

- independent source code lineage
- independant developer team and
- it ran on independent infrastructure

We saw the same thing. The zombies are real.

Conjectures and refutations

2/4

C This is malicous activity (e.g. a botnet or traffic generation, say)

R
- r_0 Not a DOS-style attack, the volume is large but not large enough to cause anybody harm. It isn't vandalism.
- r_1 Not artifical traffic generation. It had the characteristics of real organic human brains consuming content.
- r_2 No clear way for anybody to profit from this.

It does not appear malicious.

Conjectures and refutations

3/4

C This is a mobile carrier (ISP) interference. This has precident. Two examples include Verizon X-UIDH tracking and cookie manipulation.

R This traffic is not isloated to mobile ISP's. The user agent also appears on residential ISPs and other non-mobile ISPs. This is not ISP interference.

Conjectures and refutations

4/4

C This is a misconfiguration pushed as part of a major ISP, device manufacturer or app upgrade

R Pending. It could be refuted in a number of ways. For example, if the trend cut into iPhone or desktop traffic or if the daily volume were highly volitile or exhibited traits consistent with invalid traffic.

Conjectures and refutations

In support of the "upgrade" conjecture

- we observed cookie transitions
  -  →  (April)
  -  →  (August)
- the rise in zombie traffic was sudden and widespread. This is consistent with an upgrade pushed by a single agent with wide reach.
- likewise for the drop in zombie volume
- the volume shifts appeared limited to a small number of device manufacturers
- the upgrade appears limited to one application.

Zombie Postmortem

Postmortem on a zombie?

Our belief is that the zombie mob was mostly harmless.

Other factors are consistent with this theory

- Android update model is decentralized
- The mobile OS maintenance lifetime is short

Zombie Postmortem

Postmortem on a zombie?

Explicitly formulating conjectures and refutations was a useful exercise.

Progress occurs when you successfully apply the data you have in hand to a refutation statement.

Final thought

We addressed questions common to the experimental sciences, namely

- is the instrument working?
- what is really happening?

The scientific method is invaluable in resolving these questions.

This is *data science*.

comScore Madison is hiring!

comScore Madison is lead by chief scientist Paul Barford.

Thank you.

jkline@comscore.com