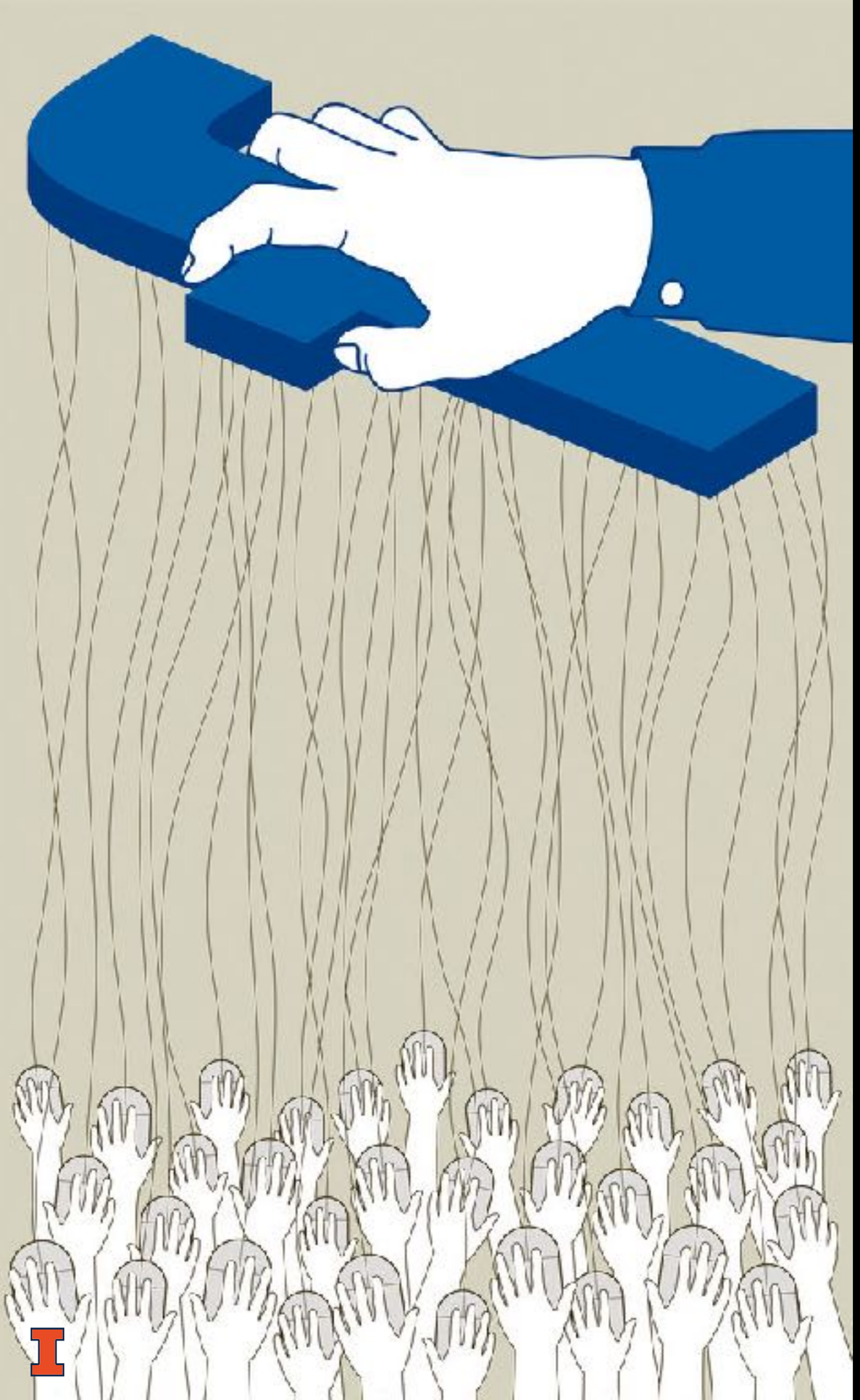


Privacy

Hari Sundaram
Associate Professor (CS, ADV)
hs1@illinois.edu

adapted from slides by Dipayan Ghosh

Introduction

Web search

Game Theory

Auctions

Data flows

Privacy

Text Ads

Display Ads

Recommender systems

Behavioral targeting

Emerging areas

Final Presentations

Data as Labor

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernandez, D., Lanier, J., and Weyl, E. G. (2018). Should we treat data as labor? moving beyond "free". AEA Papers and Proceedings, 108:38–42.

Posner, E. A. and Weyl, E. G. (2018). Radical Markets: Uprooting Capitalism and Democracy for a Just Society. Princeton University Press.

Should I annotate this photo on Facebook?

annotation: "**visiting Delhi with Vikram**"

is the exchange of services for labor fair?

portion of labor as a function of revenue

historic: 60-70%

what is it for Facebook?

~1%

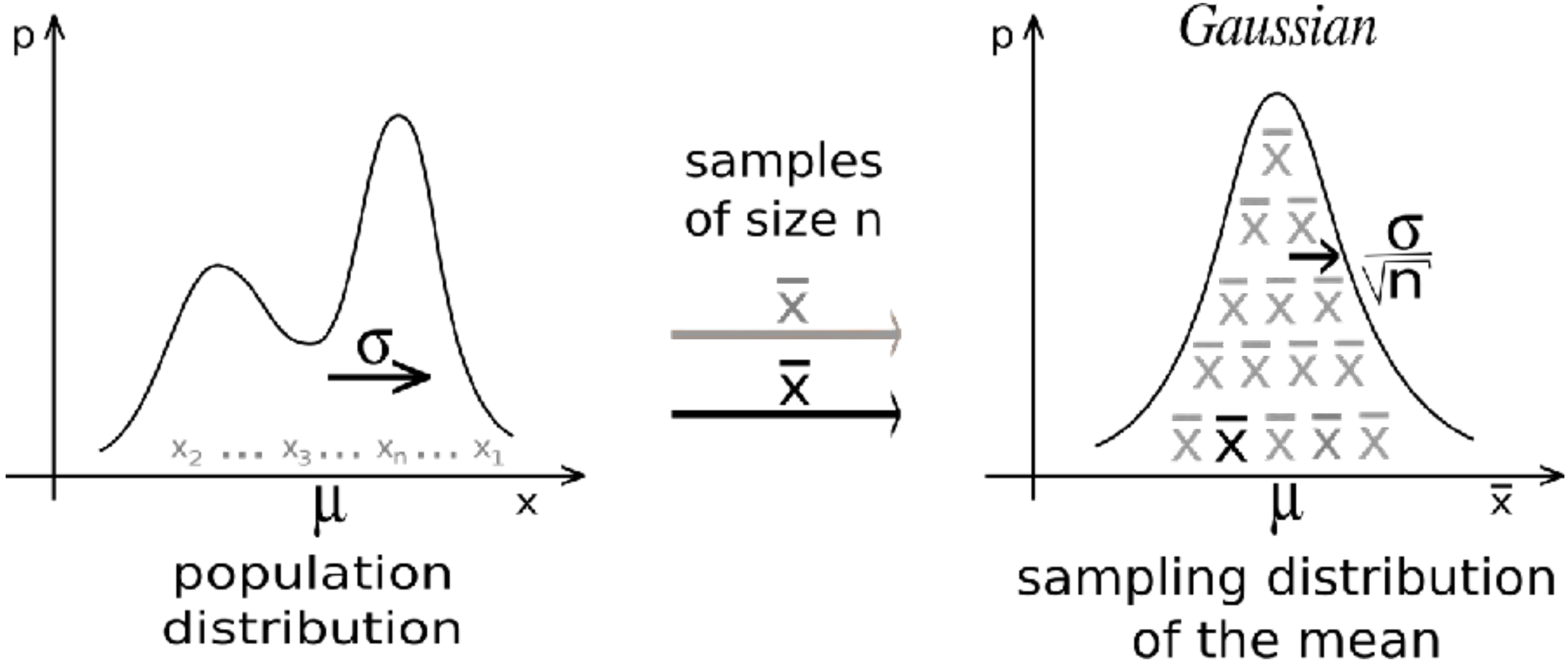
what is it for Walmart?

~60%

What is causing the difference?

what is the **marginal** value of data?

what is the difference to Facebook between revealing a number (say age), vs. annotating a photo?



source: https://en.wikipedia.org/wiki/Central_limit_theorem

With media objects, the marginal value can vary with time.

what has marginal value of a datum got to do with advertising revenues?

facebook needs classifiers
to sell categories and
contexts

{-1,1}

Let {X,Y} represent the training data, and our goal is to learn a function **h**: X → Y

sample complexity

optimal risk

$$E(h) \equiv \mathbb{E}_{\rho}[\ell(h(x), y)] \qquad E^{*}_{\mathcal{H}}(h) = \inf_{h \in \mathcal{H}} E(h)$$

a distribution

e.g.: a squared error loss

{-1,1}

Let {X,Y} represent the training data, and our goal is to learn a function **h**: X → Y

sample complexity

$$S_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \sim \rho^n$$

a distribution

$$h_n = ALG(S_n)$$

an algorithm to determine
h given n samples

{-1,1}

Let {X,Y} represent the training data, and our goal is to learn a function **h**: X → Y

ALG is consistent if:

empirical risk converges to optimal risk

empirical optimal

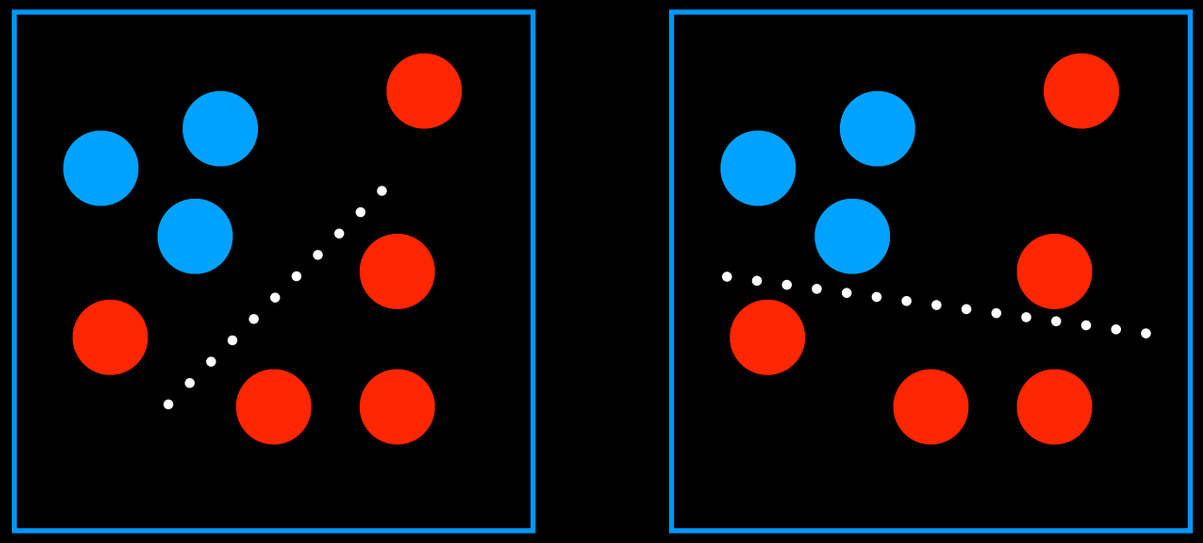
$$P[E(h_n) - E^*_{\mathscr{H}} \geq \epsilon] < \delta$$

Then, the sample complexity of ALG is the **minimum value of N** for which this is true

If there is an N such that for all n ≥ N

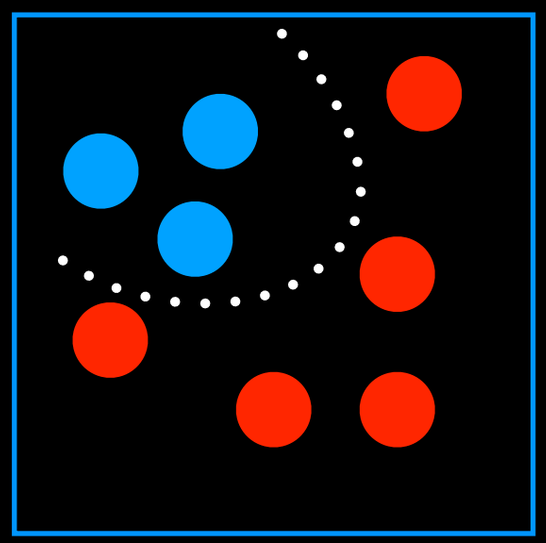
If there exists an algorithm for which N is finite, then $\mathscr{H}$ is learnable

lines



VC dimension

polynomials



binary functions

$$N = O\left(\frac{VC(\mathscr{H}) - \ln \delta}{\epsilon}\right)$$

Sample complexity depends on the VC dimension of the learner!

since Deep Neural Networks have **large** sample complexity, they need a lot of training examples!

how can we give users leverage?

data markets may be one answer

https://solid.inrupt.com

https://oceanprotocol.com

https://www.wired.com/story/i-sold-my-data-for-crypto/

Privacy and Advertising

Cambridge Analytica

The Cambridge Analytica Files

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

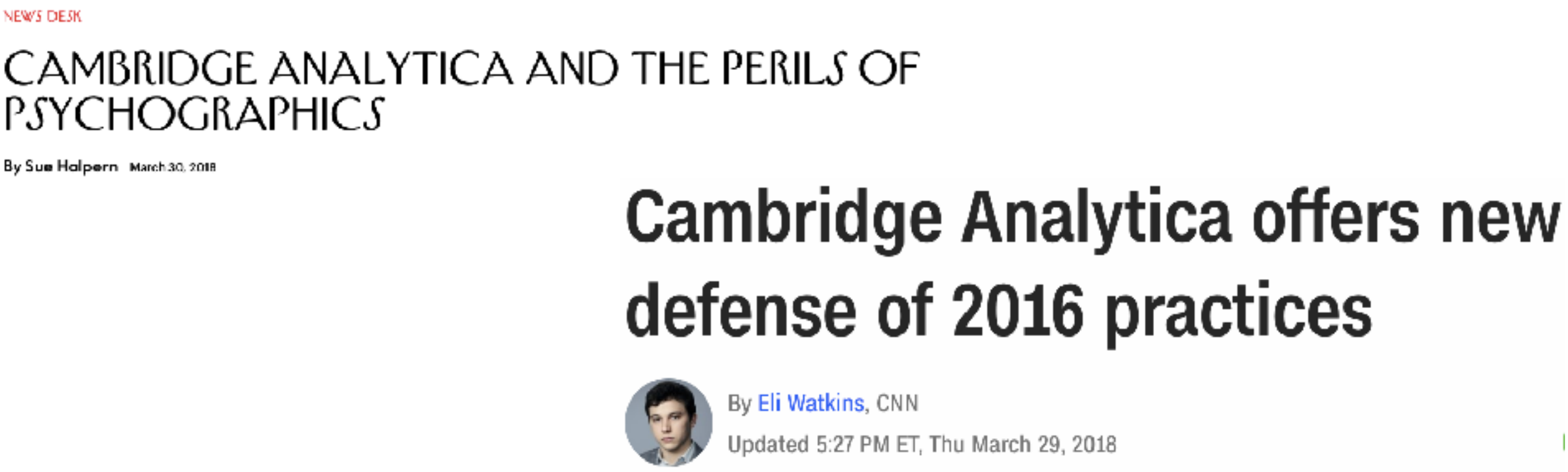
Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters

- ‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower
- Mark Zuckerberg breaks silence on Cambridge Analytica

Carole Cadwalladr *and* **Emma Graham-Harrison**

Sat 17 Mar 2018 18.03 EDT

Christopher Wylie


NEWS DESK

CAMBRIDGE ANALYTICA AND THE PERILS OF PSYCHOGRAPHICS

By Sue Halpern March 30, 2018

Cambridge Analytica offers new defense of 2016 practices

By Eli Watkins, CNN

Updated 5:27 PM ET, Thu March 29, 2018

Mark Zuckerberg on Facebook's hardest year, and what comes next

"We will dig through this hole, but it will take a few years."

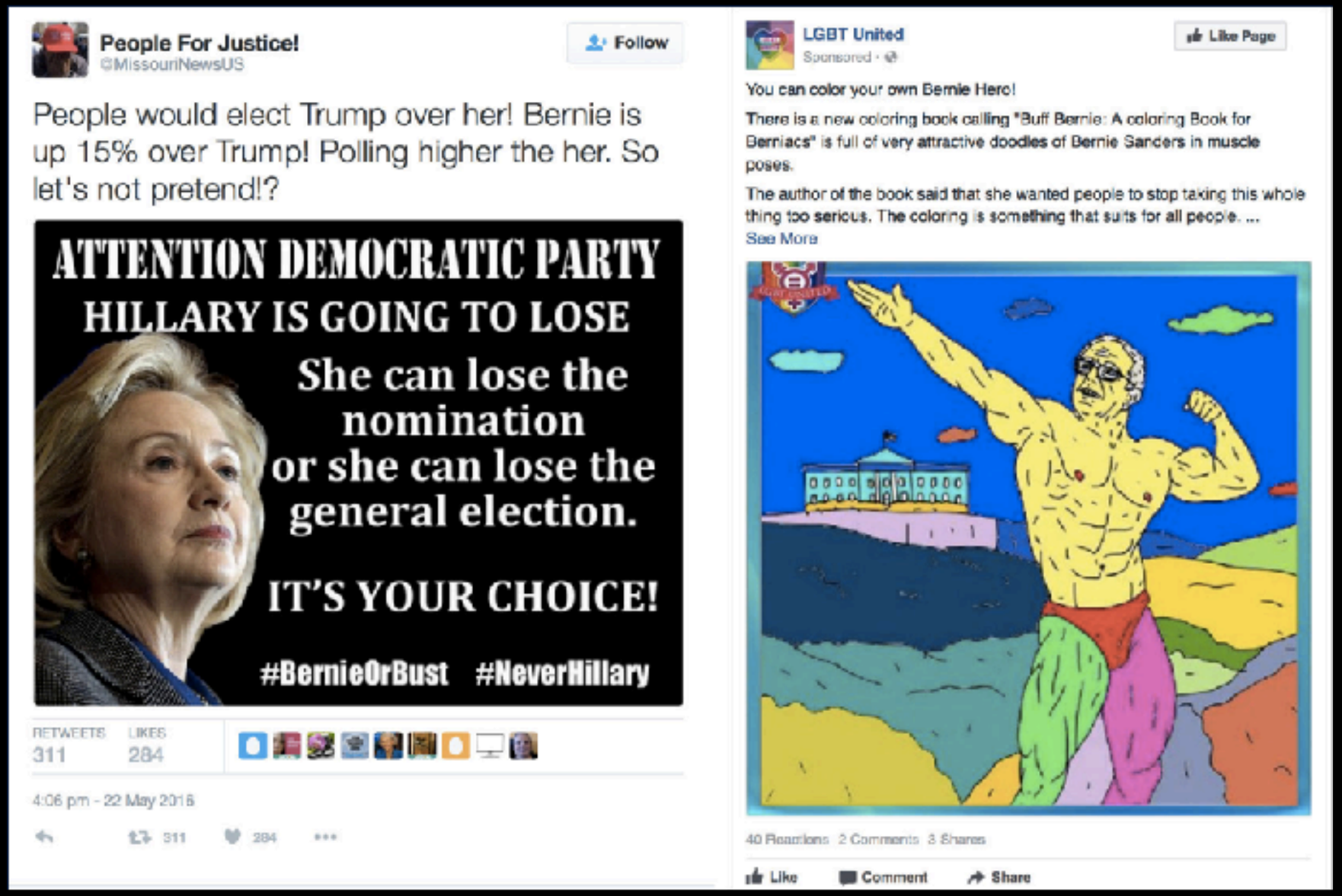
By Ezra Klein | @ezraklein | Apr 2, 2018, 6:00am EDT

Tim Cook says Facebook should have regulated itself, but it's too late for that now

"I think we're beyond that here."

By Peter Kafka | Mar 28, 2018, 12:15pm EDT

we saw many headline stories, among them that: it was really 87 million people's data, most of whom were American voters; that Cambridge was the culprit, through an intricate data-gathering exercise executed by Aleksandr Kogan; and that there would be new calls for expanded regulation of the tech sector.



We also saw vicious examples of propaganda pushed by the Russians on our favorite social media platforms. but after all the hearings and all the scrutiny of Facebook, its chief executive Mark Zuckerberg, Cambridge Analytica and all the rest, we've ended up right where we always were.

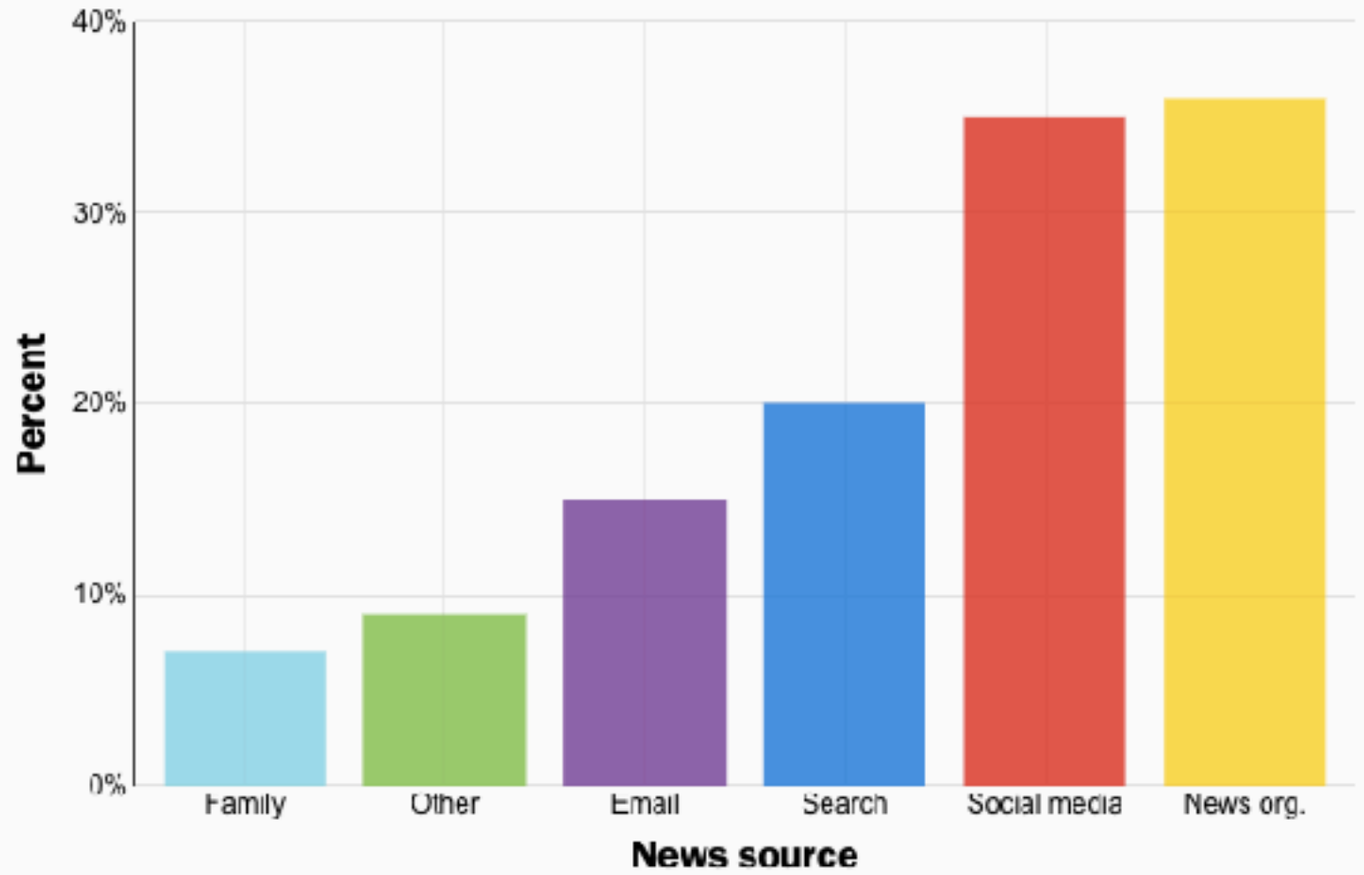
The **Infrastructure** to Spread Disinformation

&

The **Demand** for that Disinformation

why focus on social media at all?

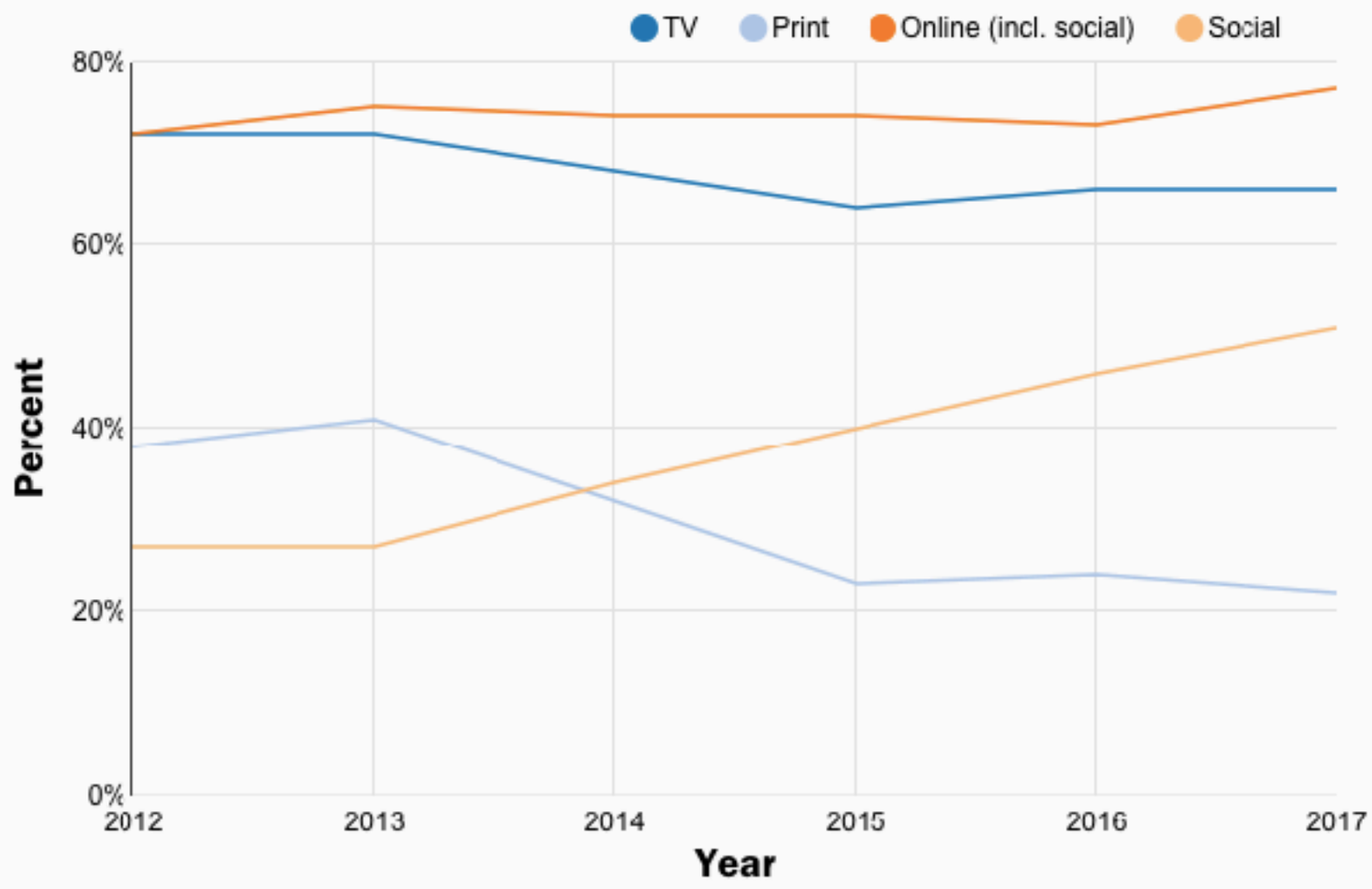
Figure 1: Where people get online news in the US, 2017



Source: Pew Research Center, "How Americans Encounter, Recall, and Act Upon Digital News," February 9, 2017.

BROOKINGS

Figure 2: Change in overall news sources, 2012-2017



Source: Nic Newman, "Digital News Sources," Reuters Institute for the Study of Journalism, 2017.

BROOKINGS

Let’s examine infrastructure

traditional advertising

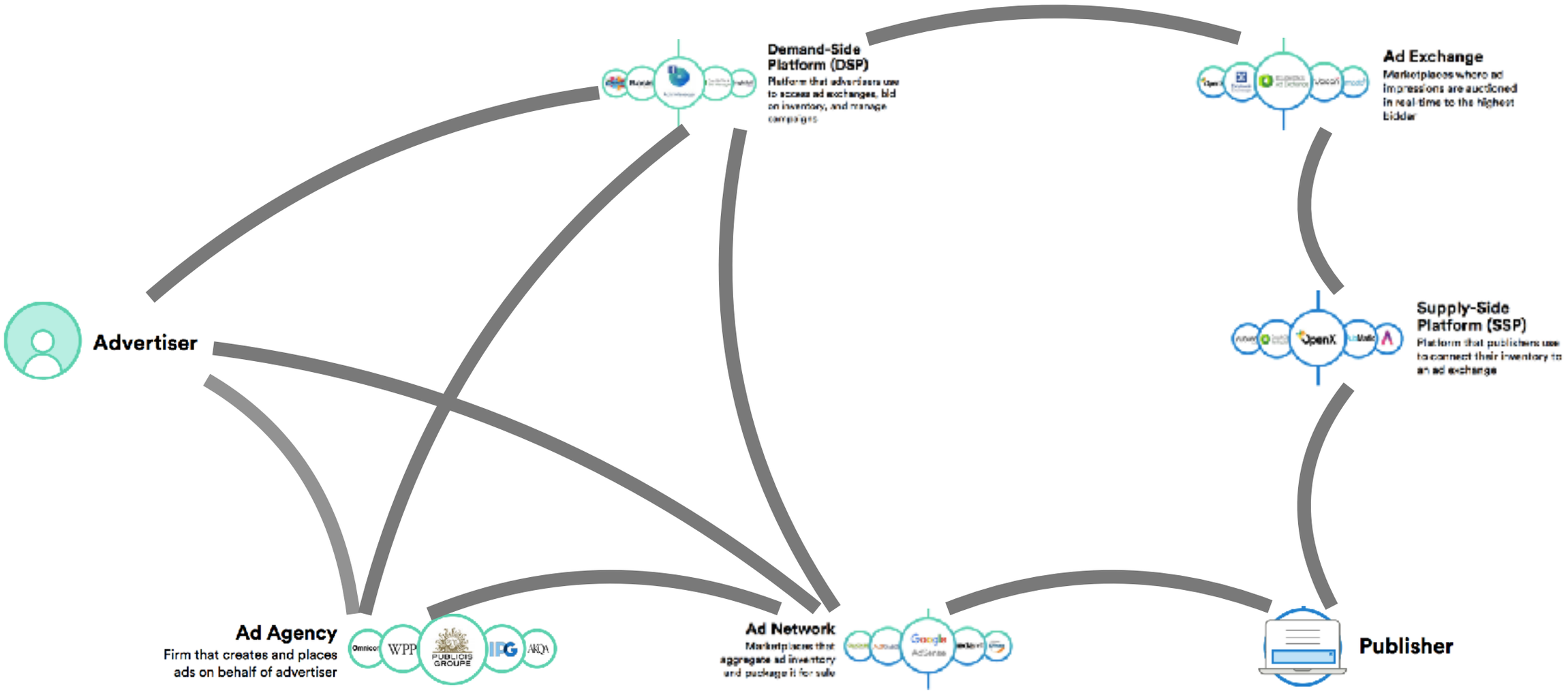
Advertiser

Ad Agency

Firm that creates and places ads on behalf of advertiser

Publisher

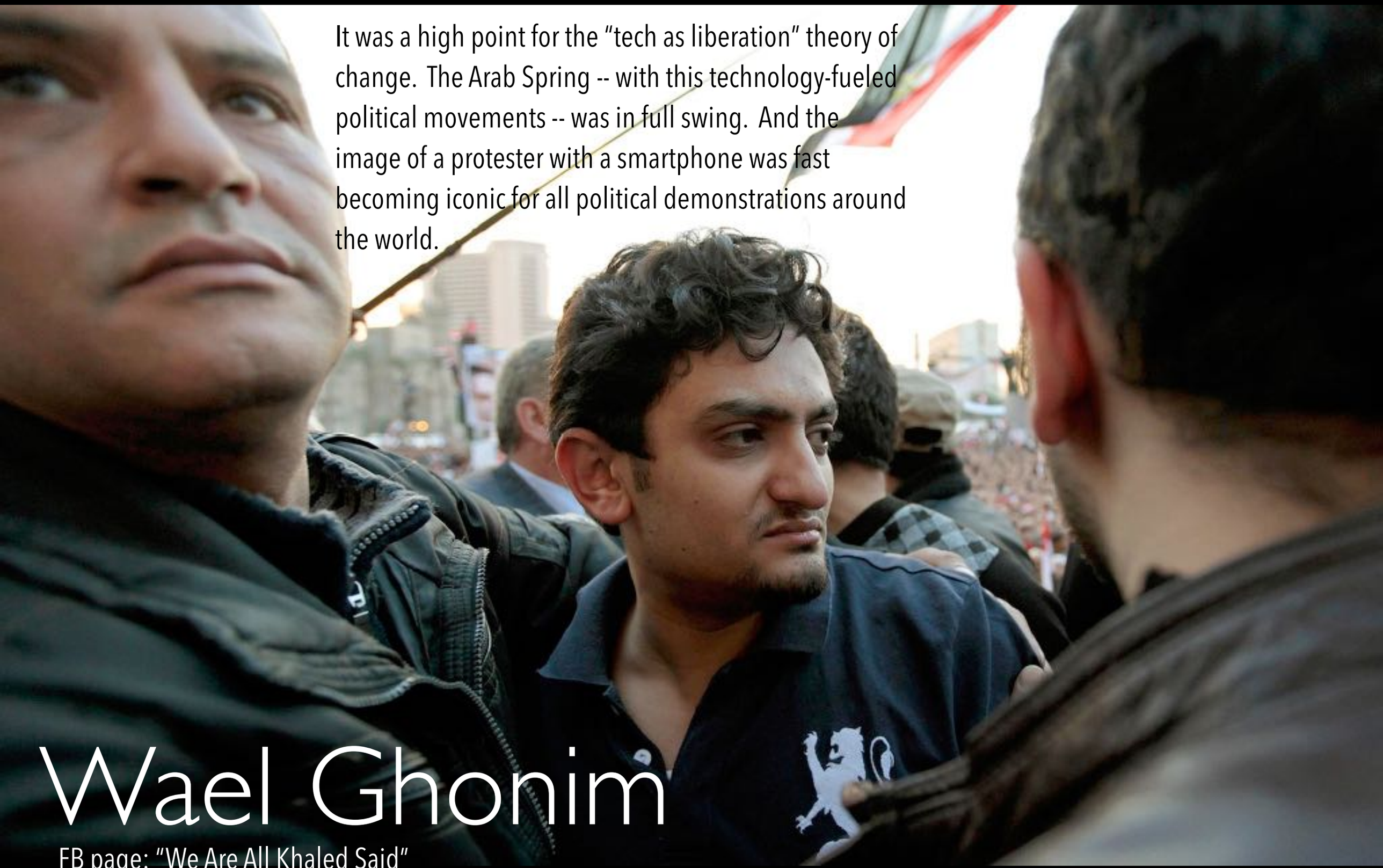
the new digital advertising ecosystem



What these entities all collectively do is to gather up troves of personal data and feed it to each other for money and information, so that they can figure out how to most effectively show you an ad that is likely to engage you on Facebook or Twitter or Google or YouTube or Snapchat.

2011

what was it like at the beginning of this decade?



It was a high point for the "tech as liberation" theory of change. The Arab Spring -- with this technology-fueled political movements -- was in full swing. And the image of a protester with a smartphone was fast becoming iconic for all political demonstrations around the world.

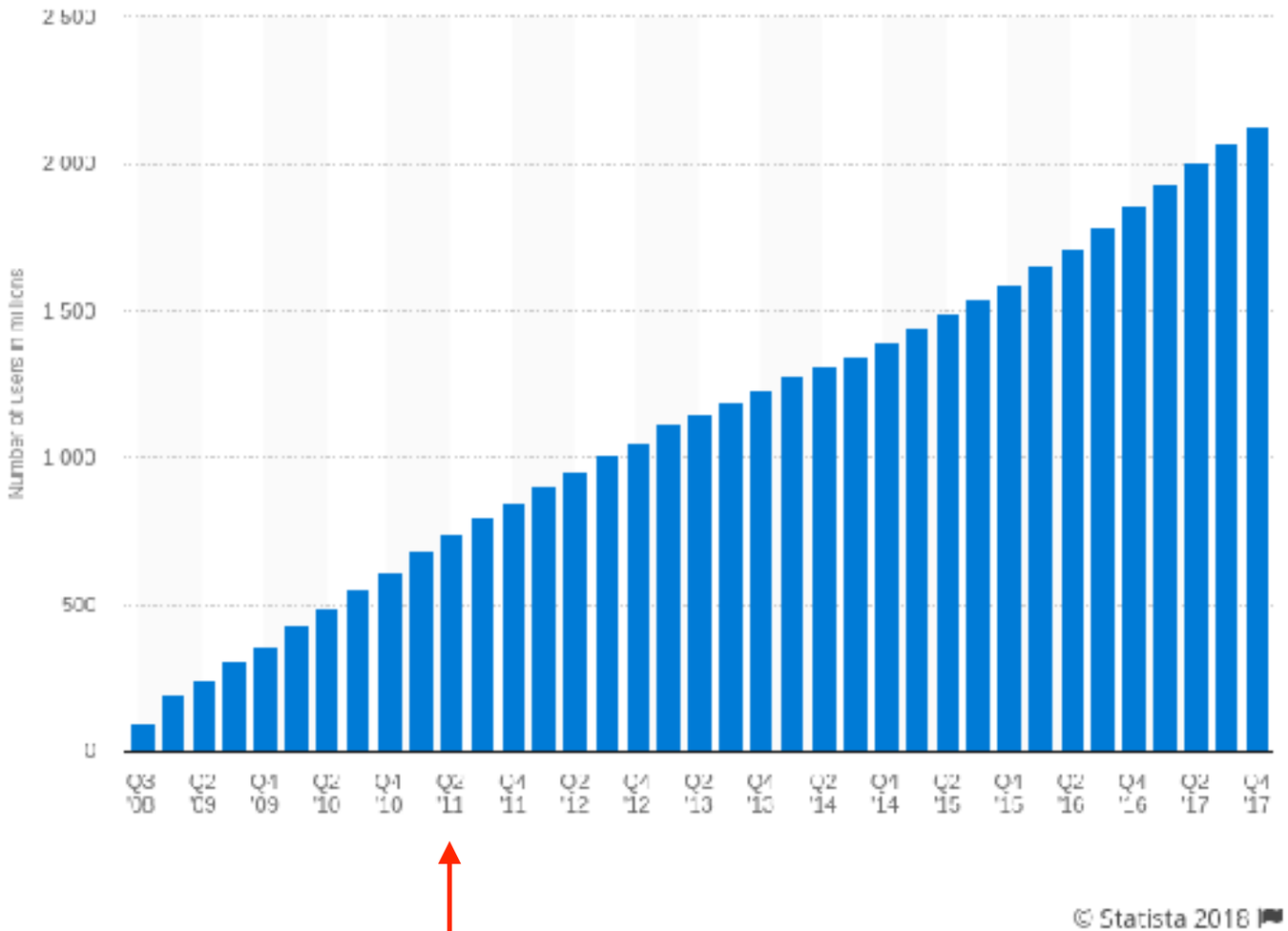
Wael Ghonim

FB page: "We Are All Khaled Said"


Jonathan Rashad © 2011

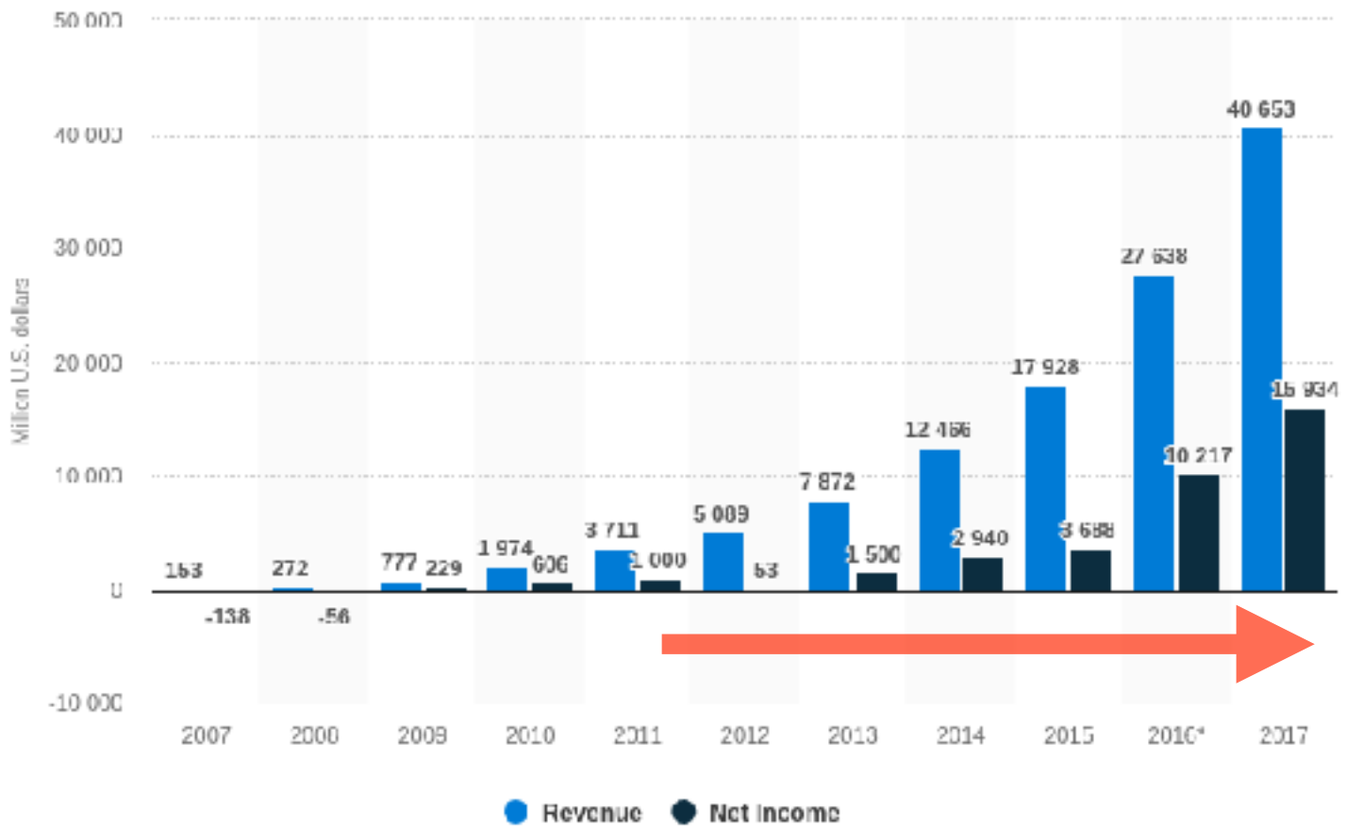
what was facebook's role?

FB's monthly active users, 2008-17



not quite the behemoth it is today

FB's annual revenues 2007-17



And it wasn't yet a money printing machine. It was "merely" a fabulously wealthy company. 2011 was the first time FB broke $1B in net profit. In 2017–it earned nearly $16B in profit. That is extraordinary growth that indicates the potency of the technologies and services the company rolled out during these years.

the growth was entirely due to advertising

true across the industry

FB's changes:

Power Editor – June 2011

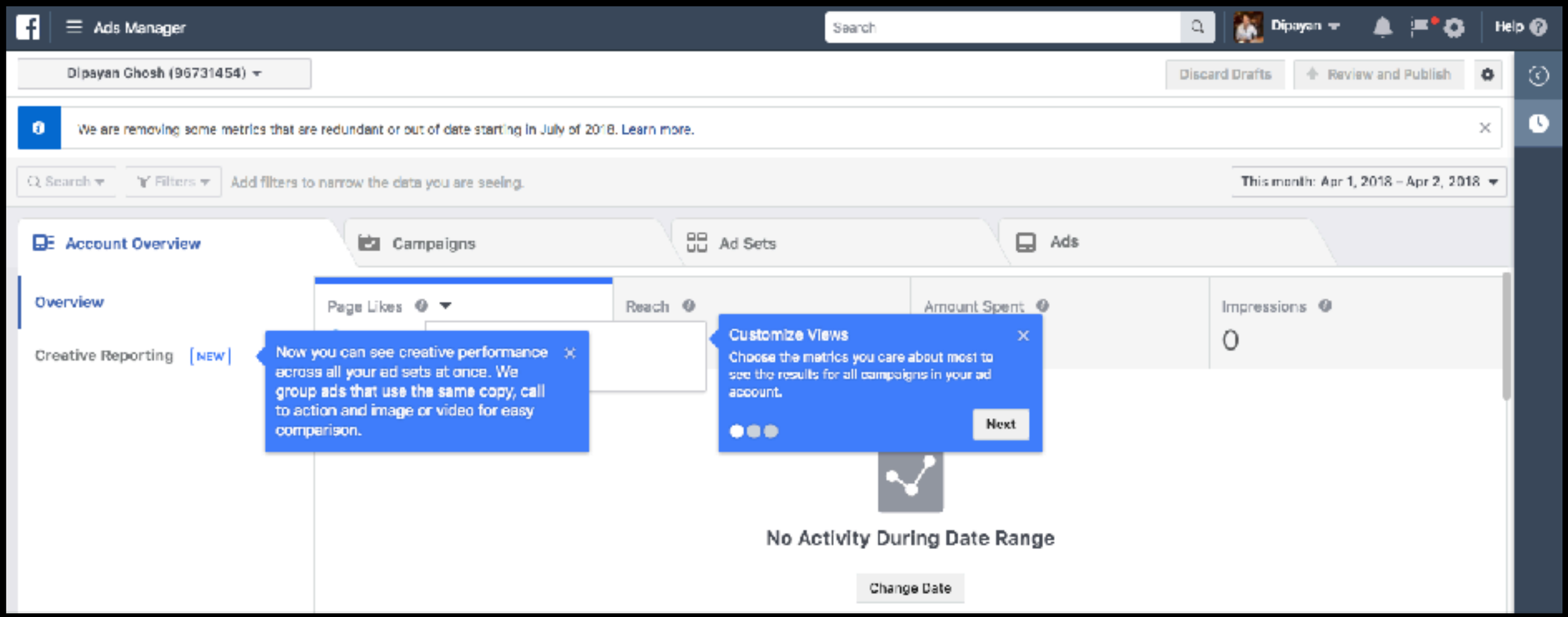
Ads in the News Feed – January 2012

Custom Audiences – September 2012

Lookalike Audiences – March 2013

Partner Categories – April 2013

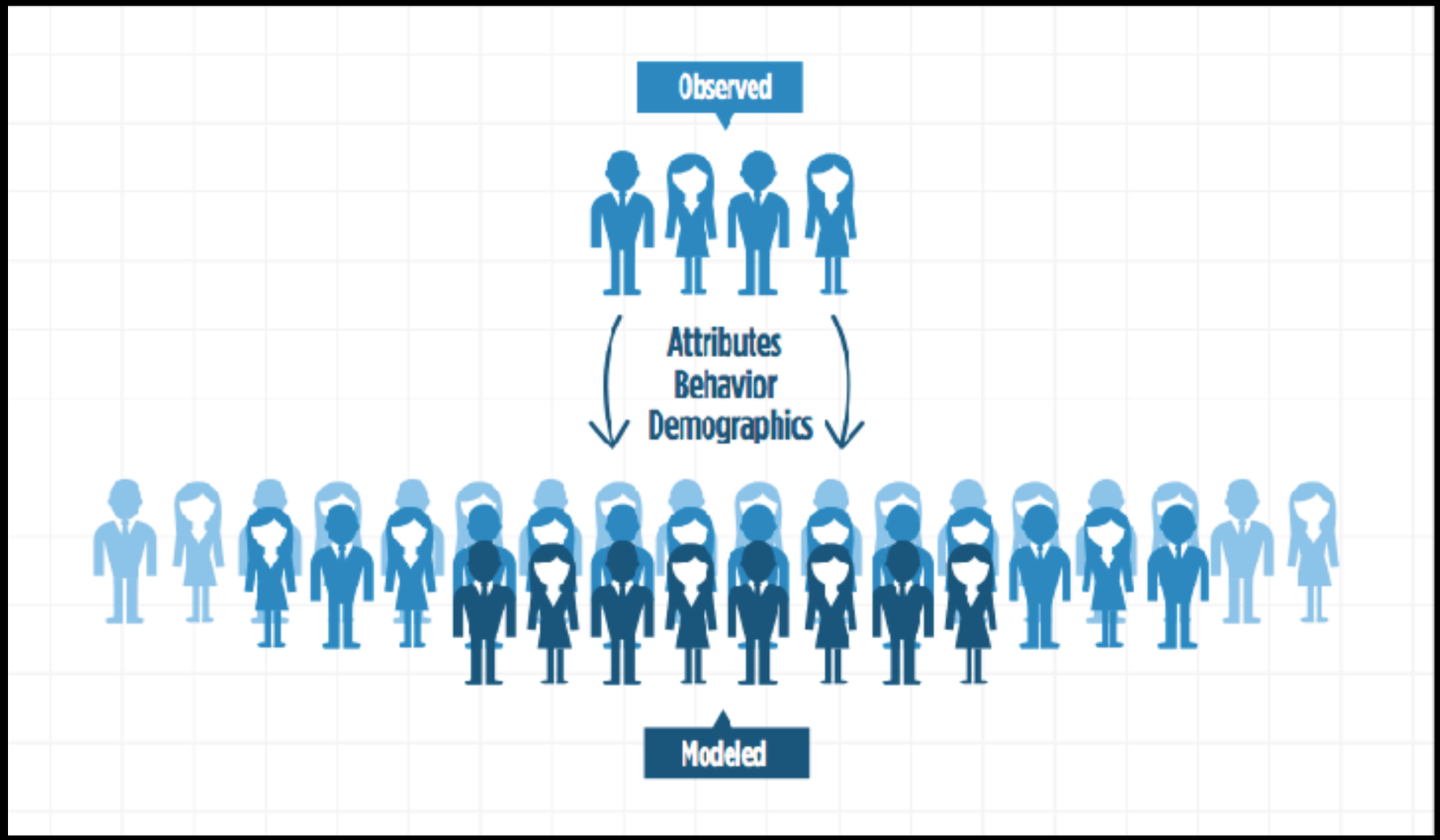
Power editor



Custom audiences

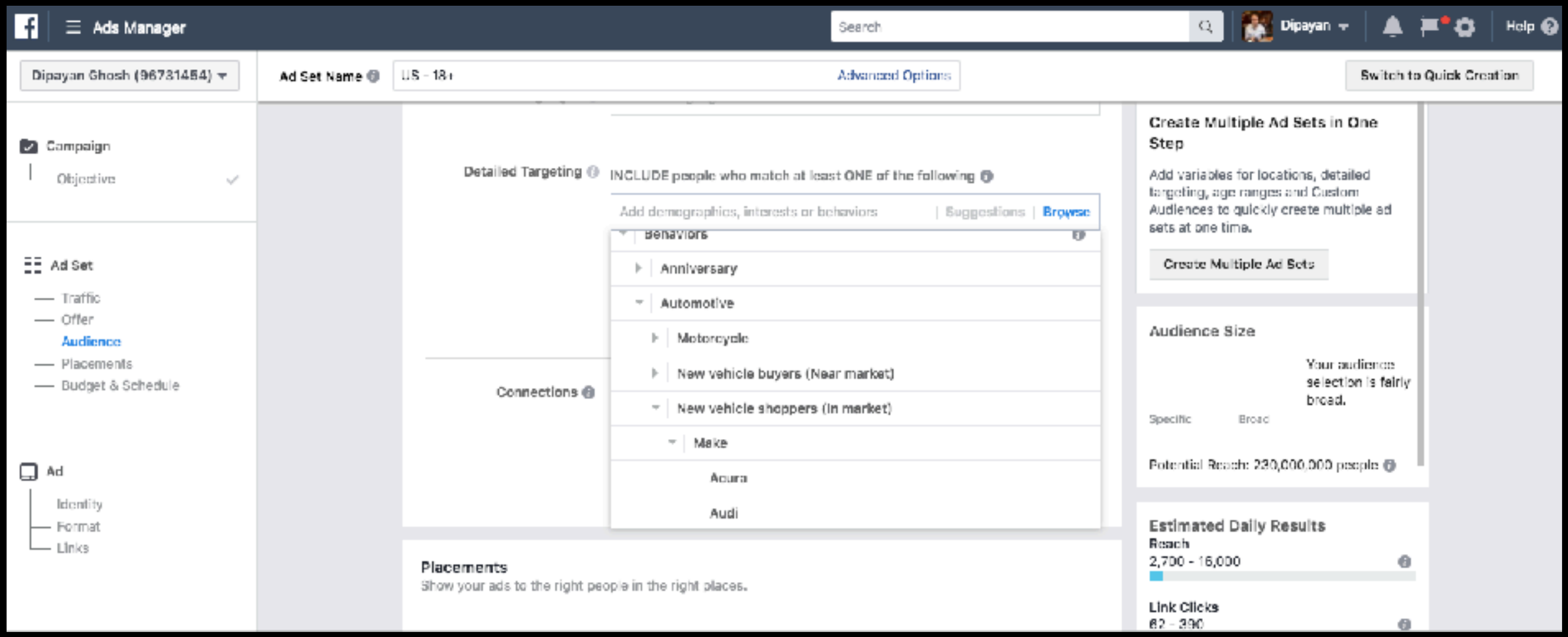
source: stream SEO

Lookalike audiences

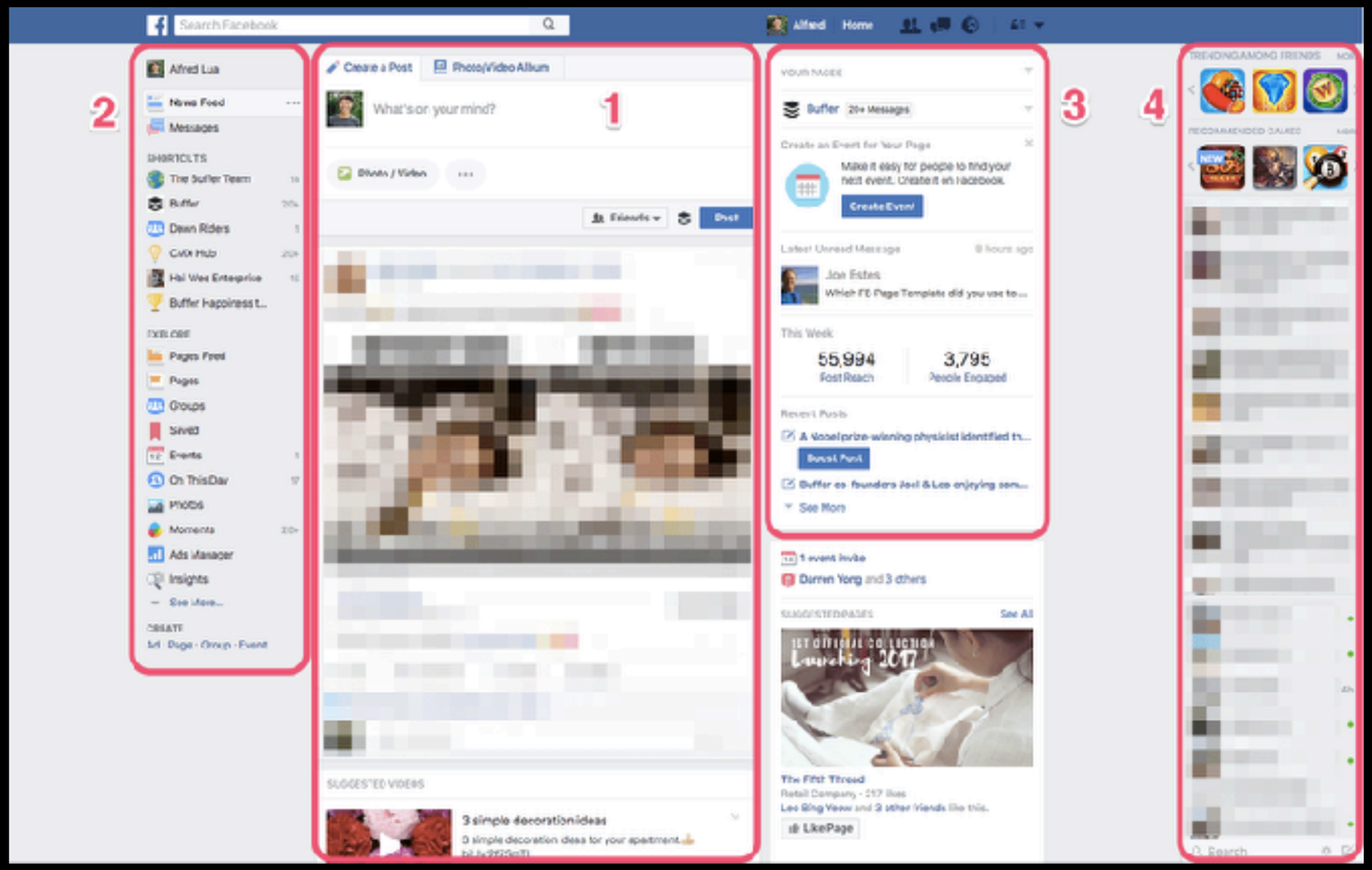


source: Hitwise

Partner Categories



Ads in the newsfeed



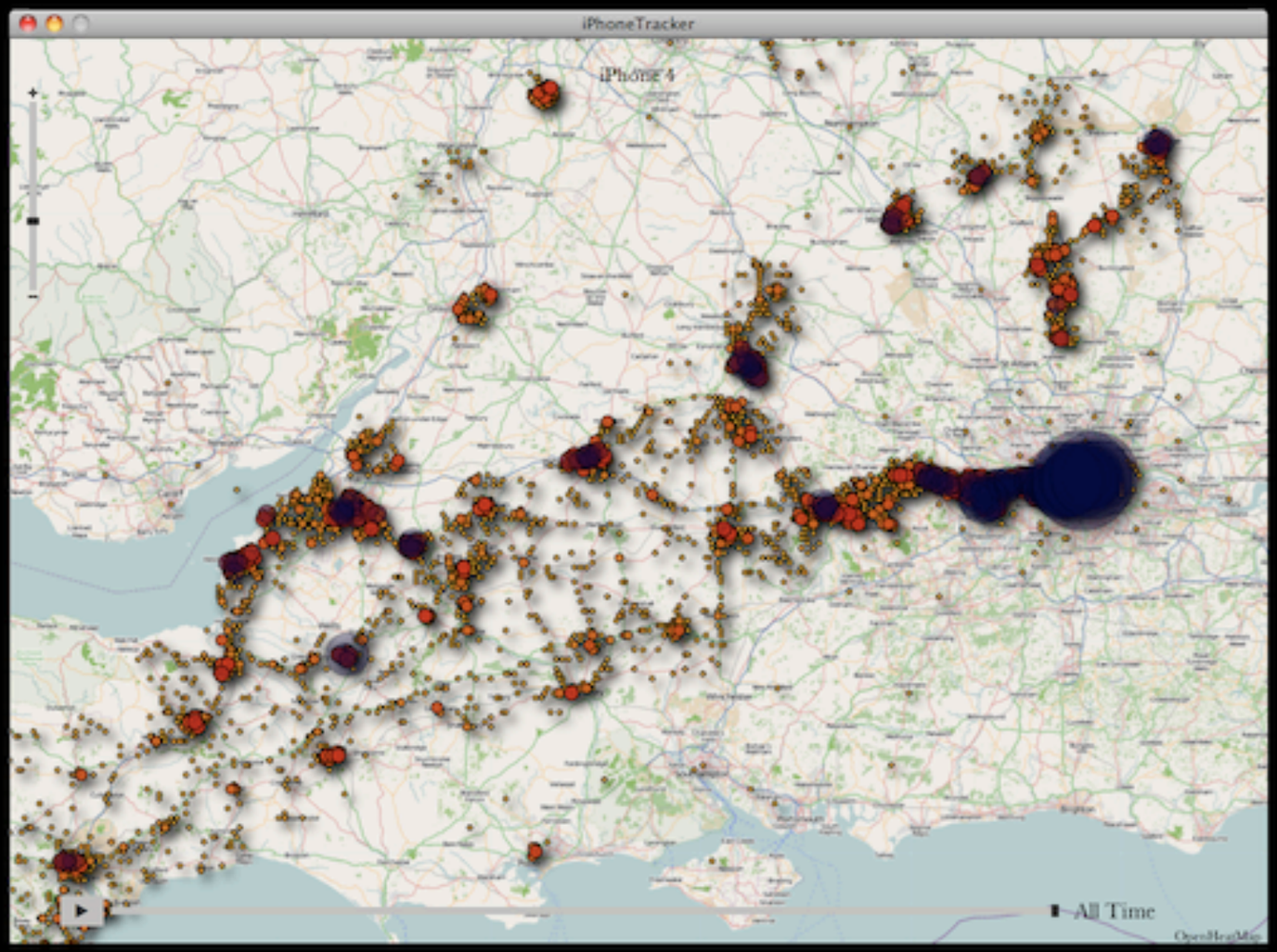
source: Buffer Social

The commercial interests of disinformation actors and internet platforms are —in some ways— aligned.

This is why it was such a massive privacy breach / when combined with its political value in particular

Underlying this entire infrastructure is highly sensitive personal data.

behavior tracking (location)



app location data requests

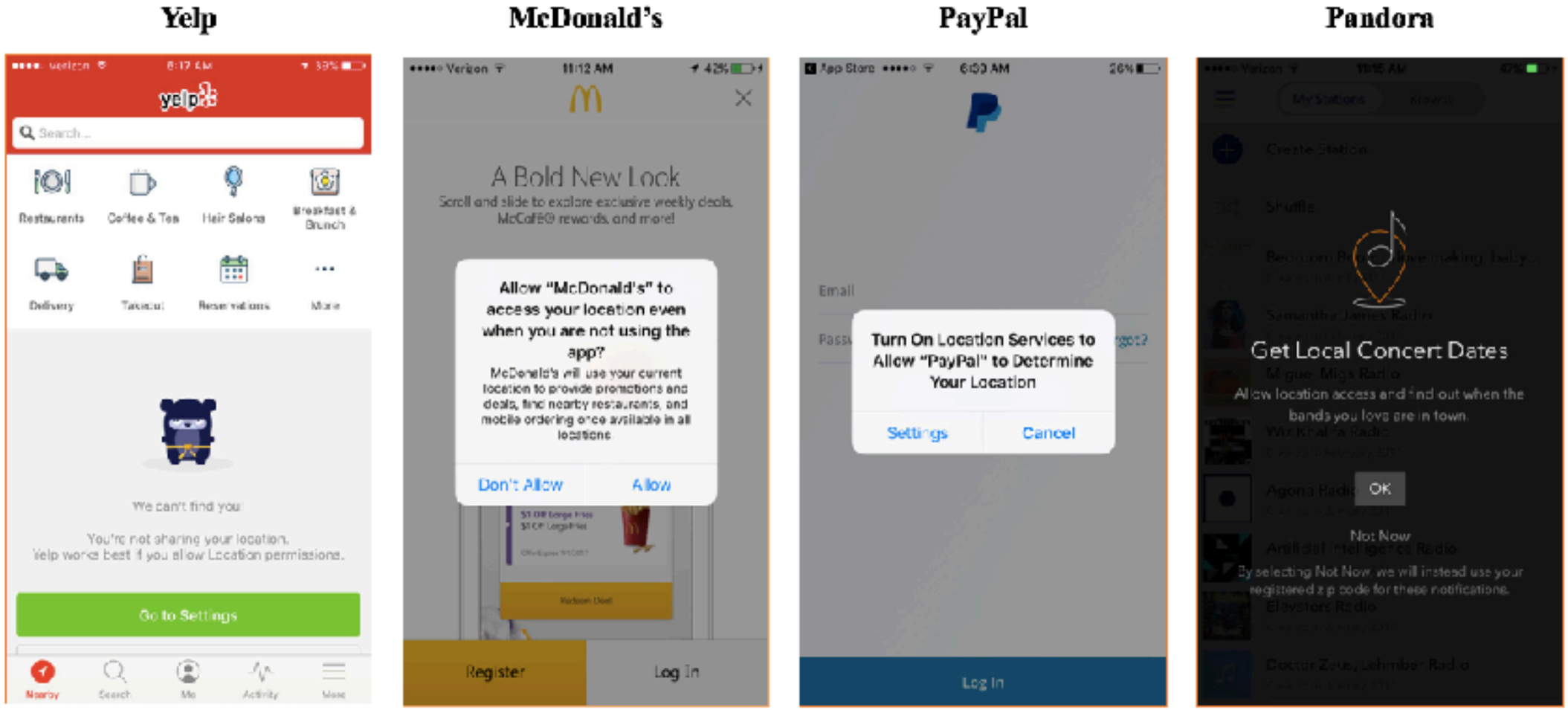


Figure: Location data requests in mobile applications offered by Yelp, McDonald's, PayPal, and Pandora.

The Cambridge-Facebook data was especially sensitive because it included **Facebook** user ids.

some applications of AI

automated ad creation

source: PSFK

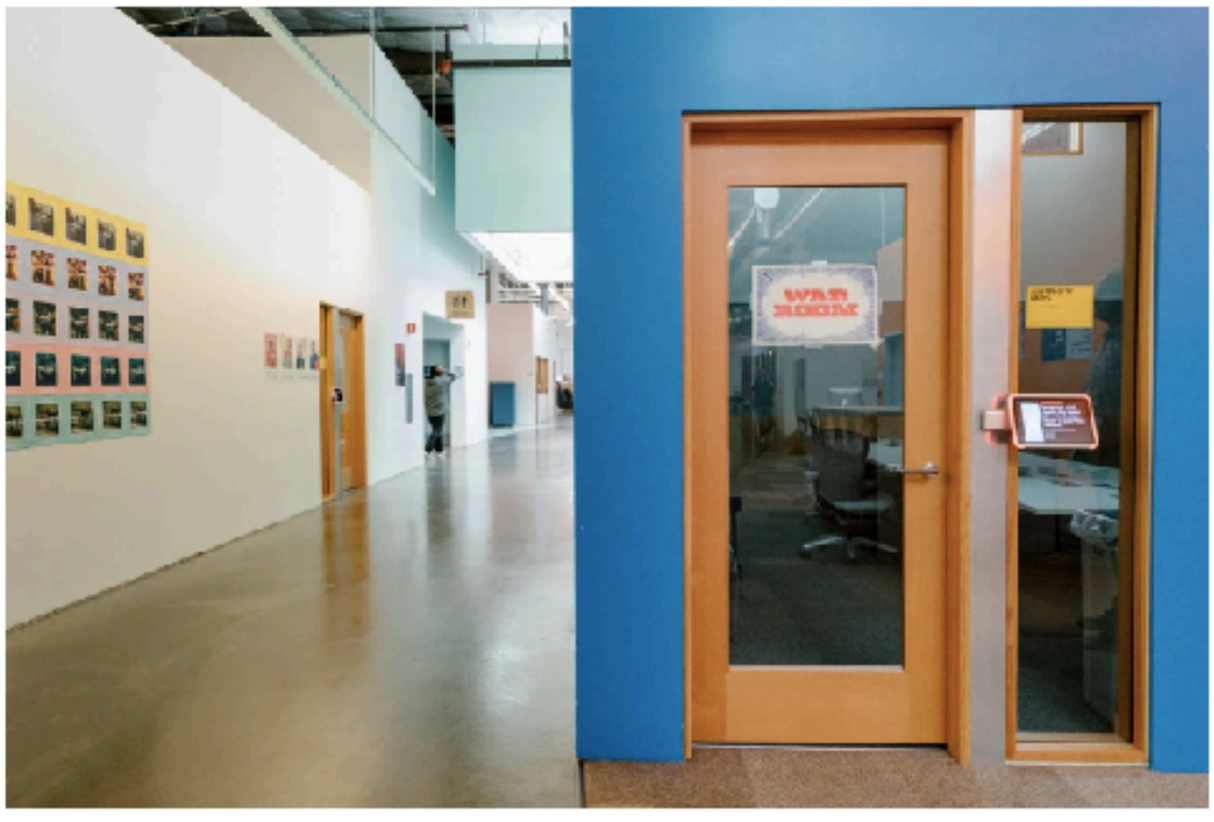
contingency based ad targeting

how do we fix this?

better detection

The New York Times

Inside Facebook's Election 'War Room'



The War Room at Facebook's headquarters in Menlo Park, Calif. As of next week, it will be the company's hub for safeguarding elections around the world. Jason Henry for The New York Times

By Sheera Frenkel and Mike Isaac

Sept. 19, 2018

76

a long-term change requires addressing the business model

The market structure that delivered us to perdition is not very old–and changing its course should not be a monumental effort.

Addictive services

Competitive Markets

Uninhibited data collection

Digital rights

Development of opaque algorithms

Transparency

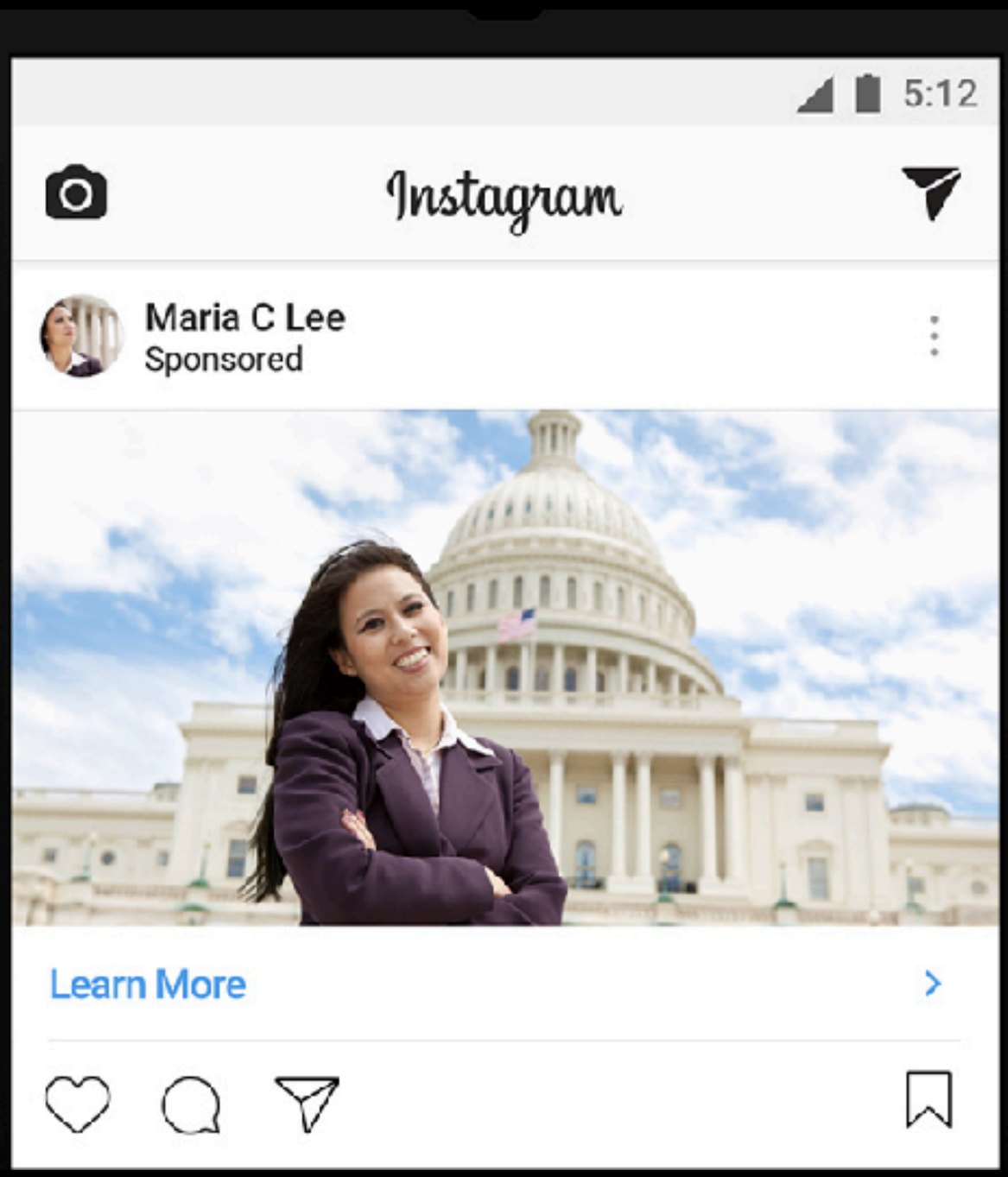
lets talk about political ads

we need ad transparency

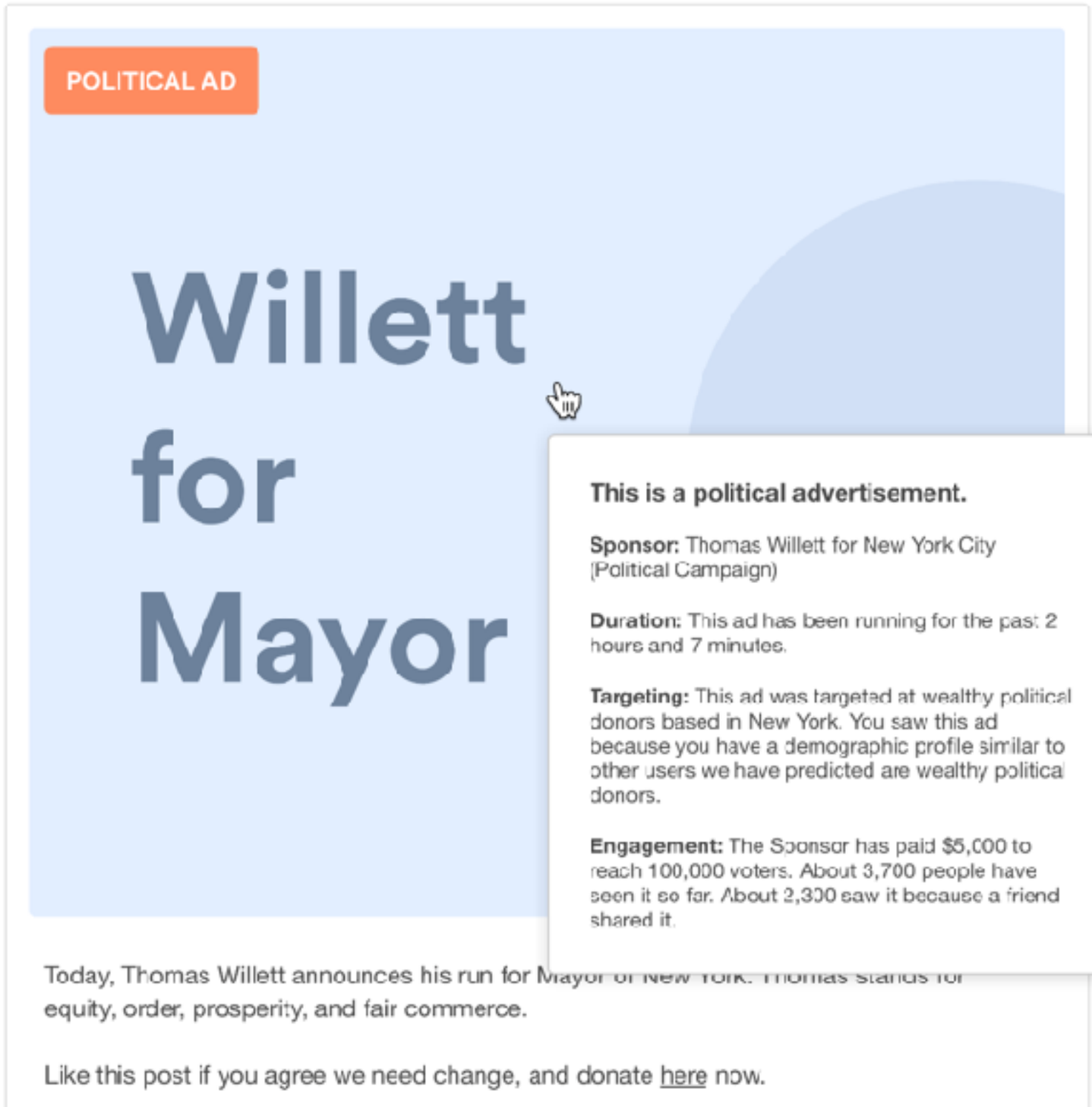

Bernie
FOR PRESIDENT
A FUTURE TO BELIEVE IN
APPROVED BY BERNIE SANDERS. PAID FOR BY BERNIE 2016

notice


5:12
Parents + Teachers
Sponsored · Paid for by Parents + Teachers
Join us in urging community leaders and school board members to increase funding for our schools. Tell them to support Prop 3 because our kids come first!
Paid for by Parents + Tea...

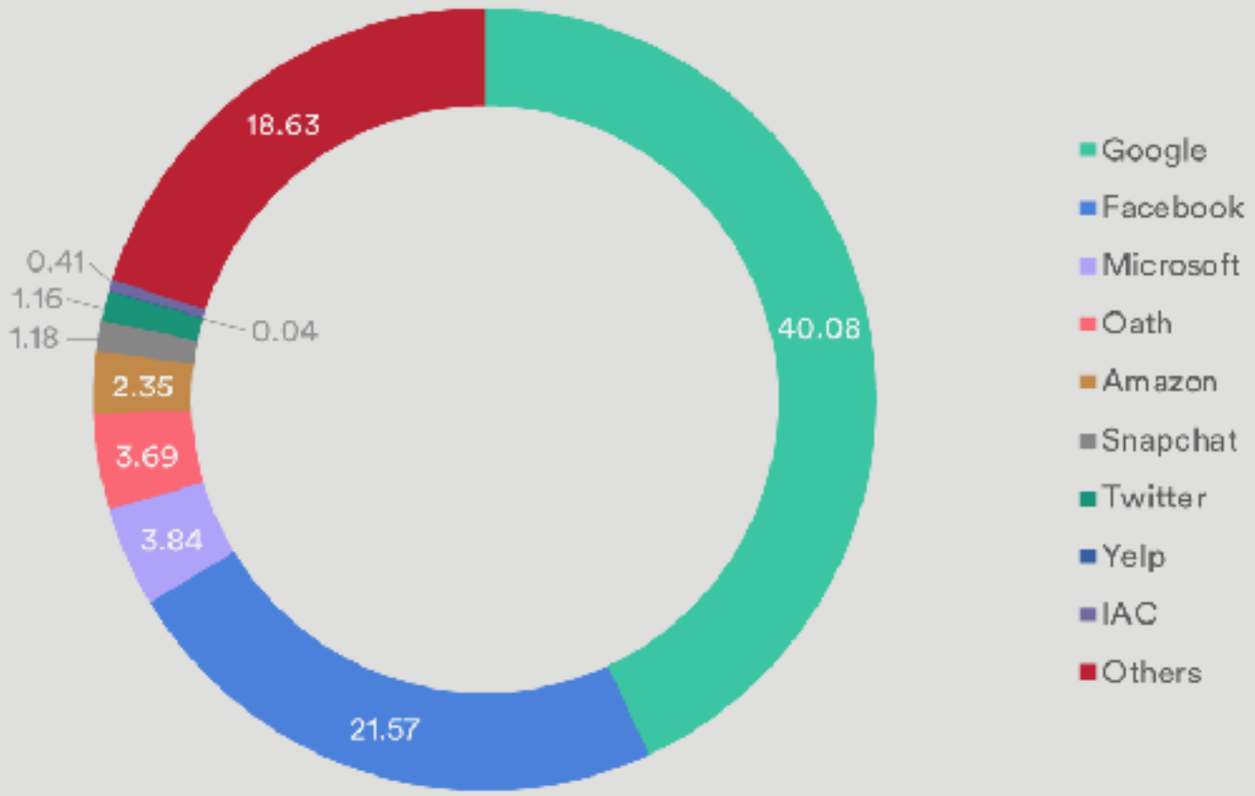

5:12
Instagram
Maria C Lee
Sponsored
Learn More

what we need


POLITICAL AD
Willett
for
Mayor
This is a political advertisement.
Sponsor: Thomas Willett for New York City (Political Campaign)
Duration: This ad has been running for the past: 2 hours and 7 minutes.
Targeting: This ad was targeted at wealthy political donors based in New York. You saw this ad because you have a demographic profile similar to other users we have predicted are wealthy political donors.
Engagement: The Sponsor has paid $5,000 to reach 100,000 voters. About 3,700 people have seen it so far. About 2,300 saw it because a friend shared it.
Today, Thomas Willett announces his run for Mayor of New York. Thomas stands for equity, order, prosperity, and fair commerce.
Like this post if you agree we need change, and donate here now.

Is the market competitive? are the tech giants harming innovation?

Share of Digital Ad Market Held by Major Tech Companies in Billions of USD



NEW AMERICA

Apple		facebook		amazon		Google	
Acquisition	Price	Acquisition	Price	Acquisition	Price	Acquisition	Price
Beats Electronics	$3,000,000,000	WhatsApp	$19,000,000,000	Whole Foods Market	$13,700,000,000	Motorola Mobility	$12,500,000,000
NeXT	$404,000,000	Oculus VR	$2,000,000,000	Zappos	$1,200,000,000	Nest Labs	$3,200,000,000
Anobit	$390,000,000	Instagram	$1,000,000,000	Pillpack	$1,000,000,000	DoubleClick	$3,100,000,000
AuthenTec	$356,000,000	LiveRail	$400m-500,000,000	Ring	$1,000,000,000	YouTube	$1,650,000,000
PrimeSense	$345,000,000	Face.com	$100,000,000	Twitch	$970,000,000	HTC properties	$1,100,000,000
P.A. Semi	$278,000,000	Atlas Solutions	<$100,000,000	Kiva Systems	$775,000,000	Waze	$966,000,000
Quattro Wireless	$275,000,000	Parse	$85,000,000	Souq.com	$580,000,000	AdMob	$750,000,000
C3 Technologies	$267,000,000	Snaptu	$70,000,000	Quidsi	$545,000,000	ITA Software	$676,000,000
Turi	$200,000,000	Pebbles	$60,000,000	Elemental Technologies	$500,000,000	Postini	$625,000,000
Lattice Data	$200,000,000	FriendFeed	$47,500,000	Annapurna Labs	$370,000,000	DeepMind Technologies	$625,000,000


FEDERAL • TRADE • COMMISSION • UNITED • STATES • OF • AMERICA •
MCMXV


DEPARTMENT OF JUSTICE
ANTITRUST DIVISION


State aid:
Ireland - Apple
Commission européenne
Margrethe Vestager
GDPR / Competition
AP

Digital rights, based on public interest commitments to privacy and security

Require meaningful consent for collection and use

Offer a right to access

Offer a right to deletion / Right to be forgotten

Require consent before processing

Require strong security practices

Introduction

Web search

Game Theory

Auctions

Data flows

Privacy

Text Ads

Display Ads

Recommender systems

Behavioral targeting

Emerging areas

Final Presentations