Introduction to Critical Data Studies

Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
@TWlyY29

Data Mining, Machine Learning, and Artificial Intelligence

> “The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that’s going to be a hugely important skill in the next decades.

Dr. Hal R. Varian,Google’s Chief Economist, 2009

> At first glance data are apparently before the fact: they are the starting point for what we know, who we are, and how we communicate. This shared sense of starting with data often leads to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself.

Gitelman, L., & Jackson, V. (2013). Introduction: Raw data is an oxymoron. Raw data is an oxymoron, 1-15.

What is data after all?

The term ‘data’

Based on the Latin term ‘dare’ = to give, ‘datum’ = something that has been given

Important written documents started with
“datum <timestamp> ...”
and became a datum
→ capturing something ephemeral

Data are characteristics associated to an individual, an organization, a location, etc.
→ objects of empirical research

> “Data are individual facts, statistics, or items of information, often numeric. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects […].

https://en.wikipedia.org/wiki/Data

> In computing, data […] is any sequence of one or more symbols. […] Data requires interpretation to become information.

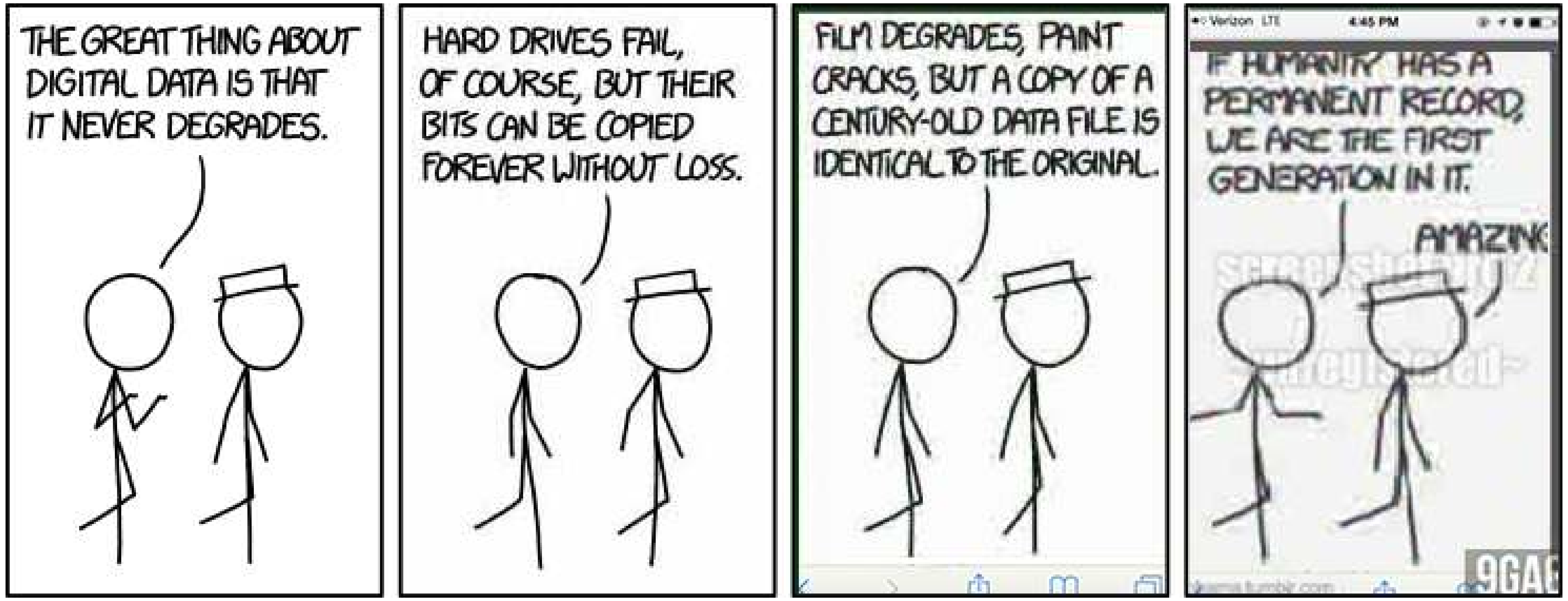
https://en.wikipedia.org/wiki/Data_(computing)

> Data are discrete, objective facts or observations, which are unorganized and unprocessed, and do not convey any specific meaning
>
> Data has no meaning or value because it is without context and interpretation

Rowley J. The wisdom hierarchy: Representations of the DIKW hierarchy. Journal of Information Science. 2007

Digital data

- Discrete (not continuous)
- Binary (0 and 1)
- Machine readable
- Replicable

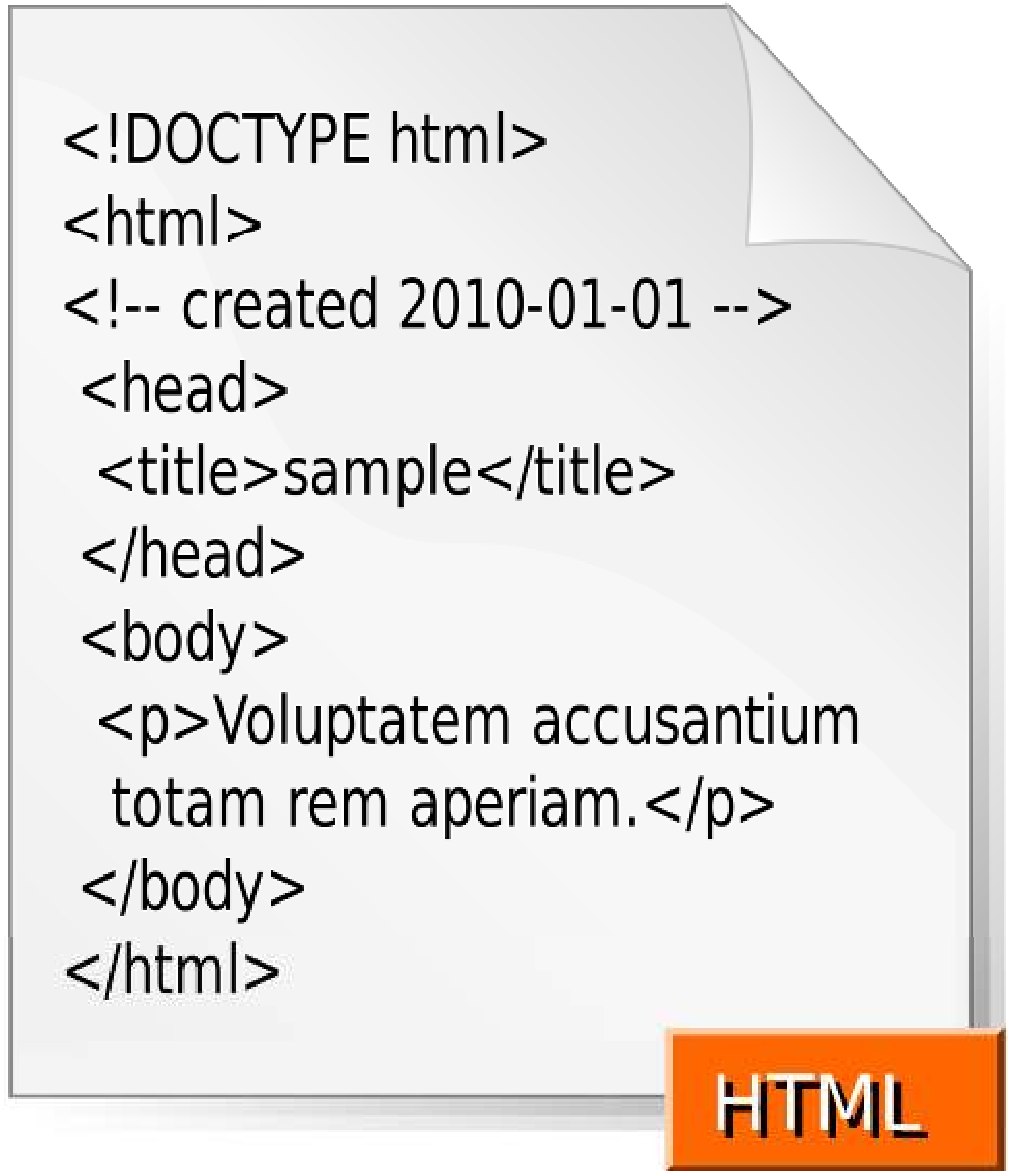

THE GREAT THING ABOUT DIGITAL DATA IS THAT IT NEVER DEGRADES.
HARD DRIVES FAIL, OF COURSE, BUT THEIR BITS CAN BE COPIED FOREVER WITHOUT LOSS.
FILM DEGRADES, PAINT CRACKS, BUT A COPY OF A CENTURY-OLD DATA FILE IS IDENTICAL TO THE ORIGINAL.
IF HUMANITY HAS A PERMANENT RECORD, WE ARE THE FIRST GENERATION IN IT.
AMAZING
9GAG

https://xkcd.com/1683/
https://commons.wikimedia.org/wiki/File:Standby_switch.jpg

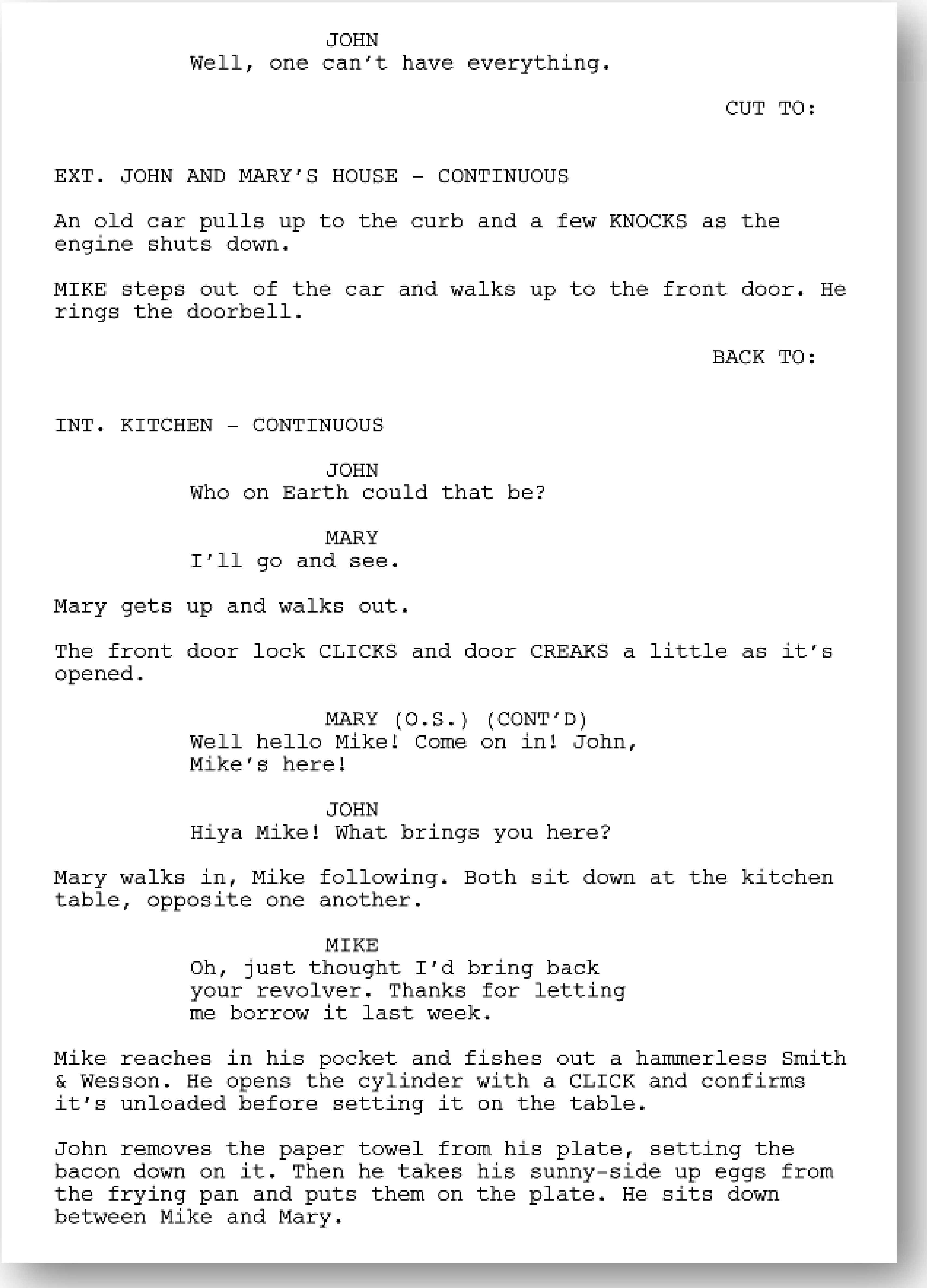
Form of data

- Highly structured: relational databases
- Semi-structured: XML, JSON, HTML
- Unstructured: plain text

Remember: this is the computers' point-of-view!



```
<!DOCTYPE html>
<html>
<!-- created 2010-01-01 -->
 <head>
  <title>sample</title>
 </head>
 <body>
  <p>Voluptatem accusantium
   totam rem aperiam.</p>
 </body>
</html>
```



```
                    JOHN
          Well, one can't have everything.

                                                 CUT TO:


EXT. JOHN AND MARY'S HOUSE - CONTINUOUS

An old car pulls up to the curb and a few KNOCKS as the
engine shuts down.

MIKE steps out of the car and walks up to the front door. He
rings the doorbell.

                                                BACK TO:


INT. KITCHEN - CONTINUOUS

                    JOHN
          Who on Earth could that be?

                    MARY
          I'll go and see.

Mary gets up and walks out.

The front door lock CLICKS and door CREAKS a little as it's
opened.

                    MARY (O.S.) (CONT'D)
          Well hello Mike! Come on in! John,
          Mike's here!

                    JOHN
          Hiya Mike! What brings you here?

Mary walks in, Mike following. Both sit down at the kitchen
table, opposite one another.

                    MIKE
          Oh, just thought I'd bring back
          your revolver. Thanks for letting
          me borrow it last week.

Mike reaches in his pocket and fishes out a hammerless Smith
& Wesson. He opens the cylinder with a CLICK and confirms
it's unloaded before setting it on the table.

John removes the paper towel from his plate, setting the
bacon down on it. Then he takes his sunny-side up eggs from
the frying pan and puts them on the plate. He sits down
between Mike and Mary.
```

https://commons.wikimedia.org/wiki/File:Screenplay_example.svg
https://commons.wikimedia.org/wiki/File:HTML.svg

Data = higher truth?

Data are *made* not given.

Data are worthless without an interpretive context or a *purpose*.

To become information, *knowledge* about purpose of data is *essential*.

Different information can be obtained from the same data.

From Digits to Knowledge

Principles of exchange market
KNOWLEDGE
1,00€ → 1,20$
INFORMATION
1,20
DATA
“0”, “1”, “2”, “,”
SIGNS

Connections to experience create knowledge

Semantics: assigning meaning to data

Syntax is used to bring signs into an order

Herrmann, R. (2012). Wissenspyramide. derwirtschaftsinformatiker.de. https://derwirtschaftsinformatiker.de/2012/09/12/it-management/wissenspyramide-wiki/

Where do you come in contact with all this?

Datafication

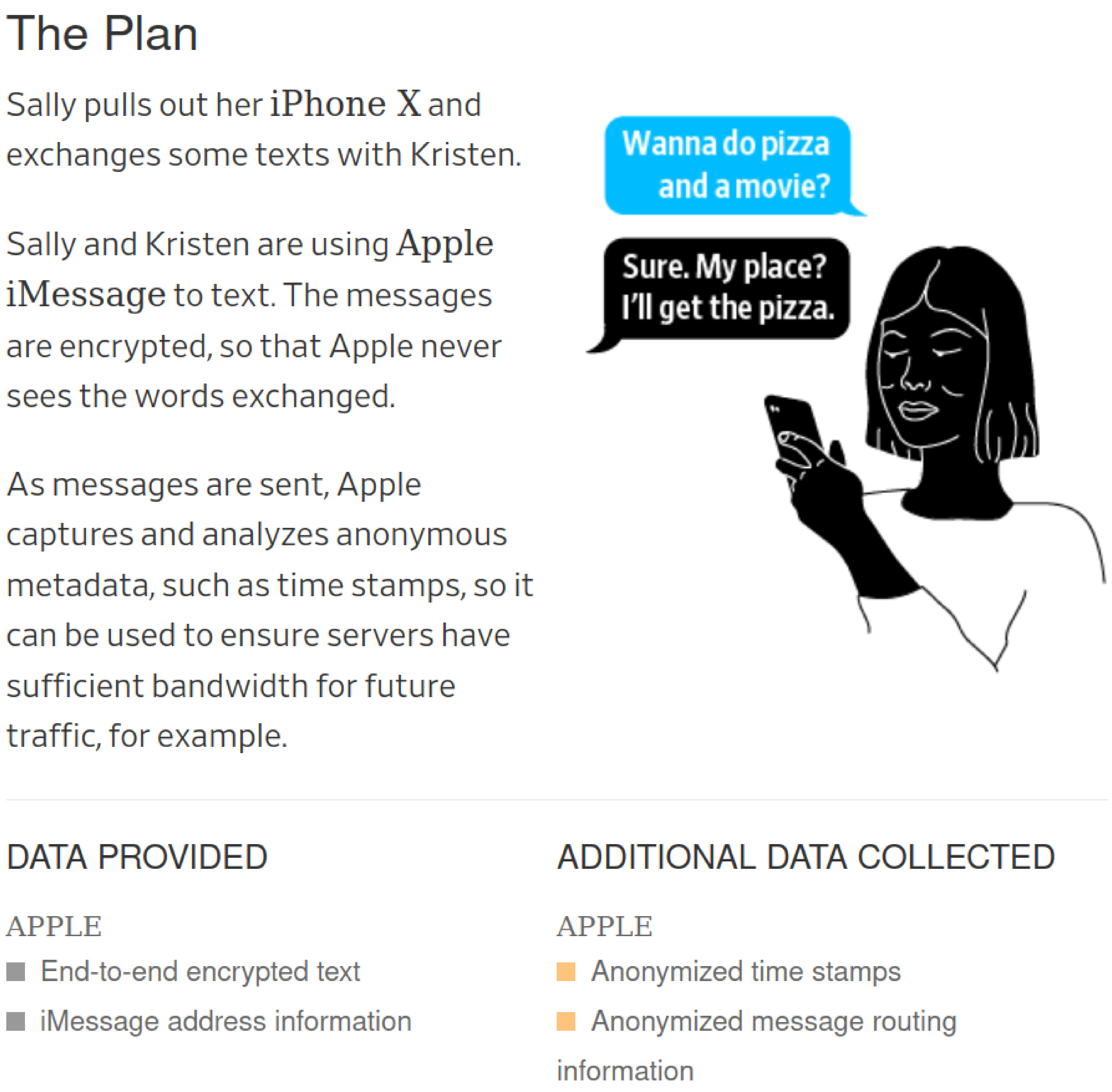
> Datafication is a modern technological trend turning many aspects of our life into computerized data and transforming this information into new forms of value.
>
> Wikipedia on “datafication”

Digitalization of our daily lifes &
Enriching human behavior with context information

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way
China is Using Facial Recognition Trash Bins to Make Sure People Recycle, Kezia Parkins,

Imagine “Sally” sets up a pizza-and-movie night with her friend “Kristen.” The Wall Street Journal reviewed privacy statements to assess just how much data could be unknowingly shared on top of the price of that pepperoni pie.

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/



The Plan

Sally pulls out her iPhone X and exchanges some texts with Kristen.

Sally and Kristen are using Apple iMessage to text. The messages are encrypted, so that Apple never sees the words exchanged.

As messages are sent, Apple captures and analyzes anonymous metadata, such as time stamps, so it can be used to ensure servers have sufficient bandwidth for future traffic, for example.

Wanna do pizza and a movie?

Sure. My place? I'll get the pizza.

DATA PROVIDED

APPLE

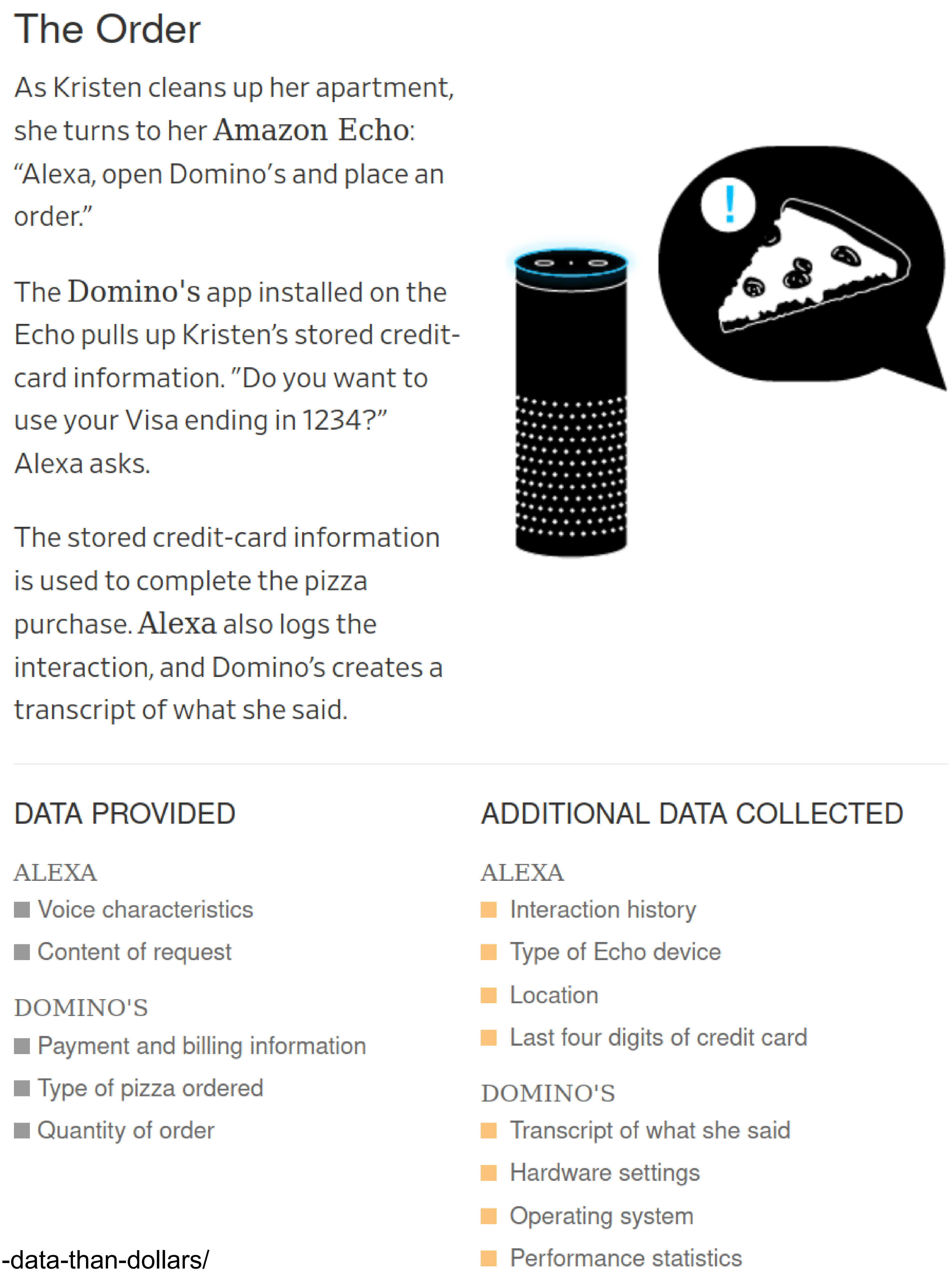
- End-to-end encrypted text
- iMessage address information

ADDITIONAL DATA COLLECTED

APPLE

- Anonymized time stamps
- Anonymized message routing information

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/



The Order

As Kristen cleans up her apartment, she turns to her Amazon Echo: “Alexa, open Domino’s and place an order.”

The Domino's app installed on the Echo pulls up Kristen’s stored credit-card information. ”Do you want to use your Visa ending in 1234?” Alexa asks.

The stored credit-card information is used to complete the pizza purchase. Alexa also logs the interaction, and Domino’s creates a transcript of what she said.

DATA PROVIDED

ALEXA

- Voice characteristics
- Content of request

DOMINO'S

- Payment and billing information
- Type of pizza ordered
- Quantity of order

ADDITIONAL DATA COLLECTED

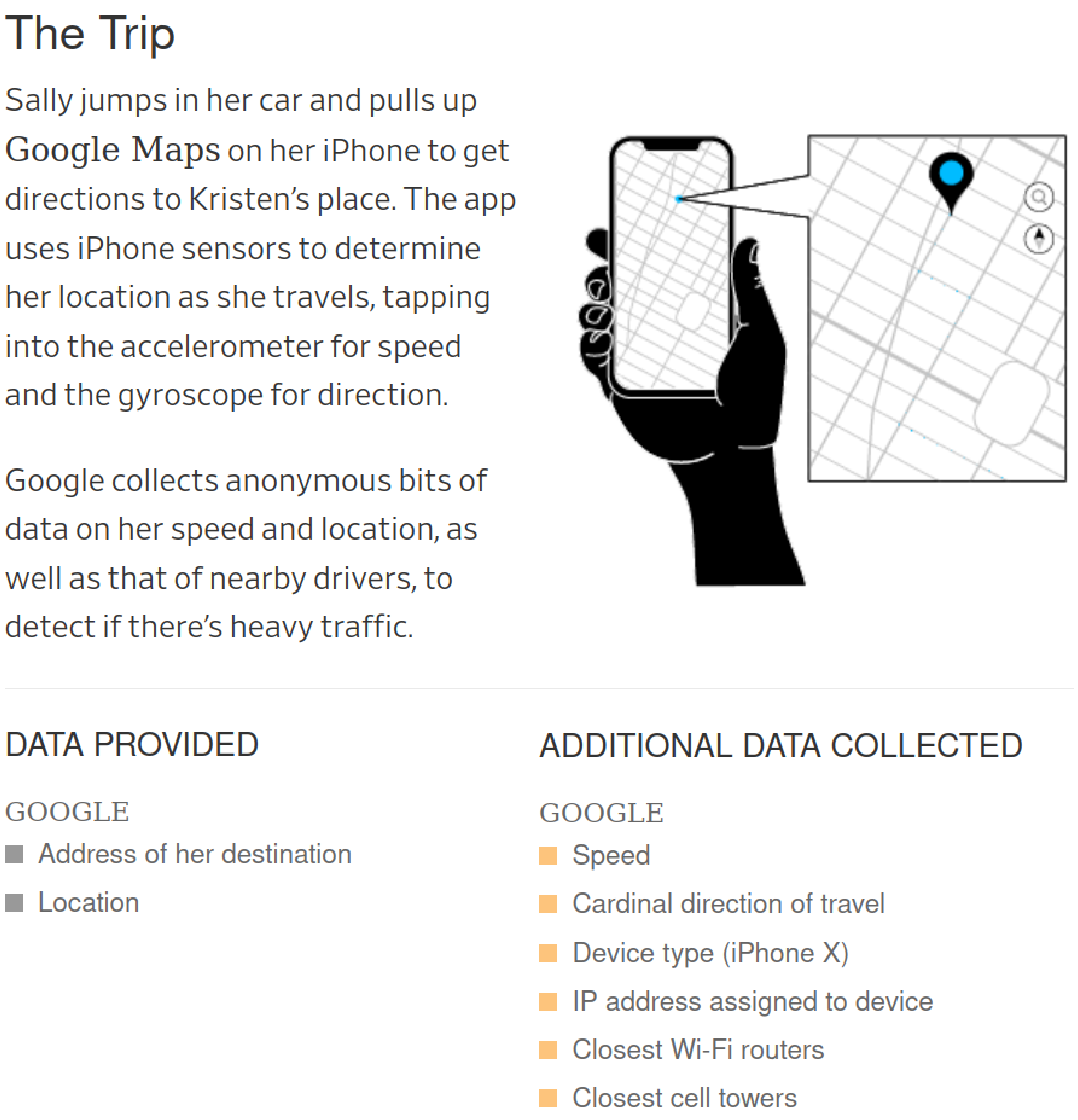
ALEXA

- Interaction history
- Type of Echo device
- Location
- Last four digits of credit card

DOMINO'S

- Transcript of what she said
- Hardware settings
- Operating system
- Performance statistics

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/



The Trip

Sally jumps in her car and pulls up Google Maps on her iPhone to get directions to Kristen's place. The app uses iPhone sensors to determine her location as she travels, tapping into the accelerometer for speed and the gyroscope for direction.

Google collects anonymous bits of data on her speed and location, as well as that of nearby drivers, to detect if there's heavy traffic.

DATA PROVIDED

GOOGLE

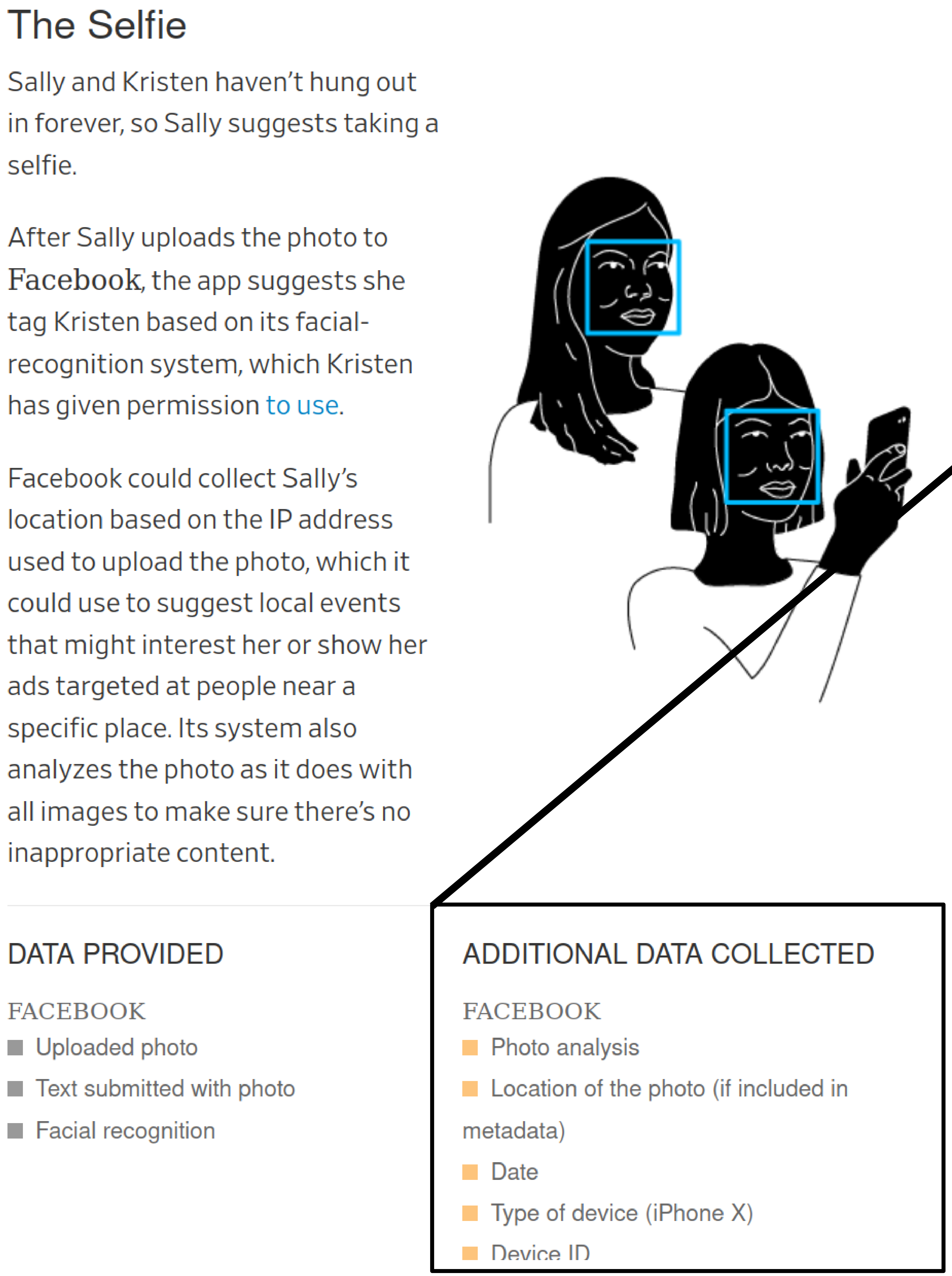
- Address of her destination
- Location

ADDITIONAL DATA COLLECTED

GOOGLE

- Speed
- Cardinal direction of travel
- Device type (iPhone X)
- IP address assigned to device
- Closest Wi-Fi routers
- Closest cell towers

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/



The Selfie

Sally and Kristen haven't hung out in forever, so Sally suggests taking a selfie.

After Sally uploads the photo to Facebook, the app suggests she tag Kristen based on its facial-recognition system, which Kristen has given permission to use.

Facebook could collect Sally's location based on the IP address used to upload the photo, which it could use to suggest local events that might interest her or show her ads targeted at people near a specific place. Its system also analyzes the photo as it does with all images to make sure there's no inappropriate content.

DATA PROVIDED

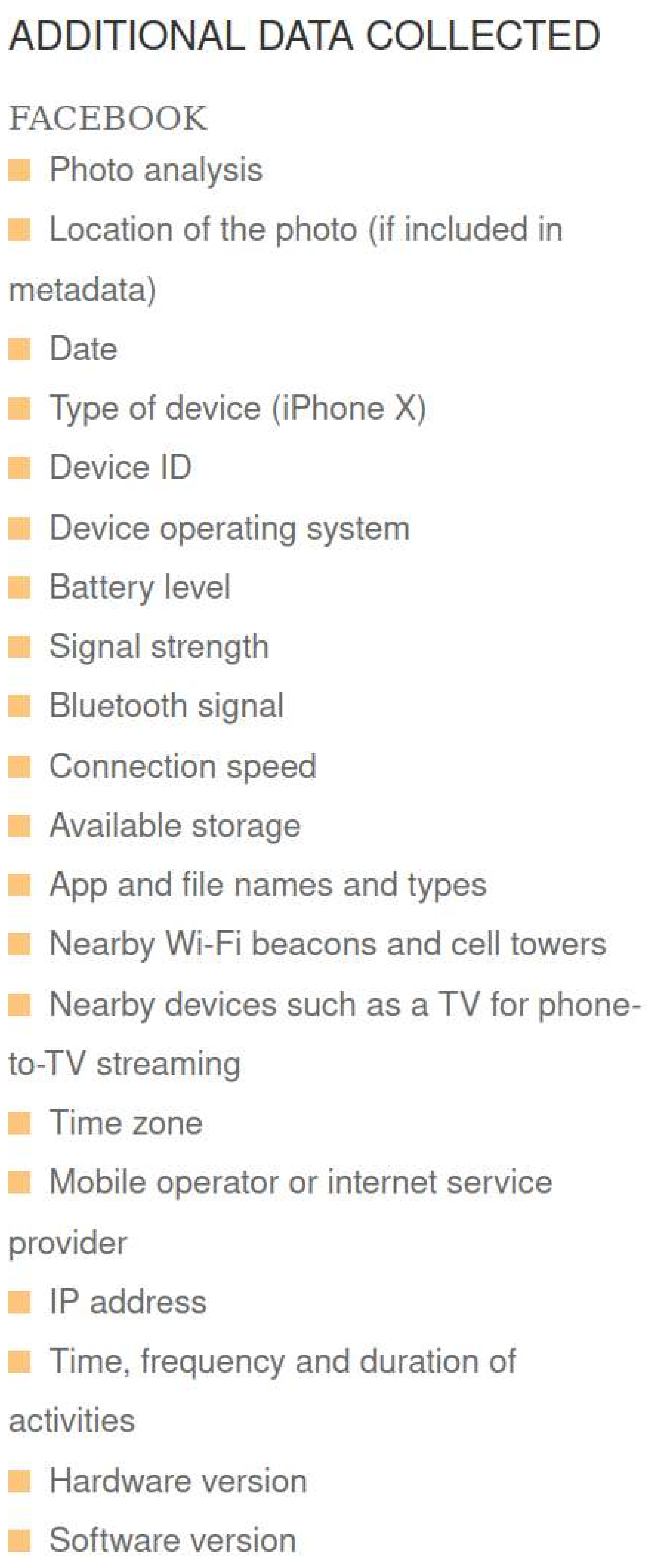
FACEBOOK

- Uploaded photo
- Text submitted with photo
- Facial recognition

ADDITIONAL DATA COLLECTED

FACEBOOK

- Photo analysis
- Location of the photo (if included in metadata)
- Date
- Type of device (iPhone X)
- Device ID

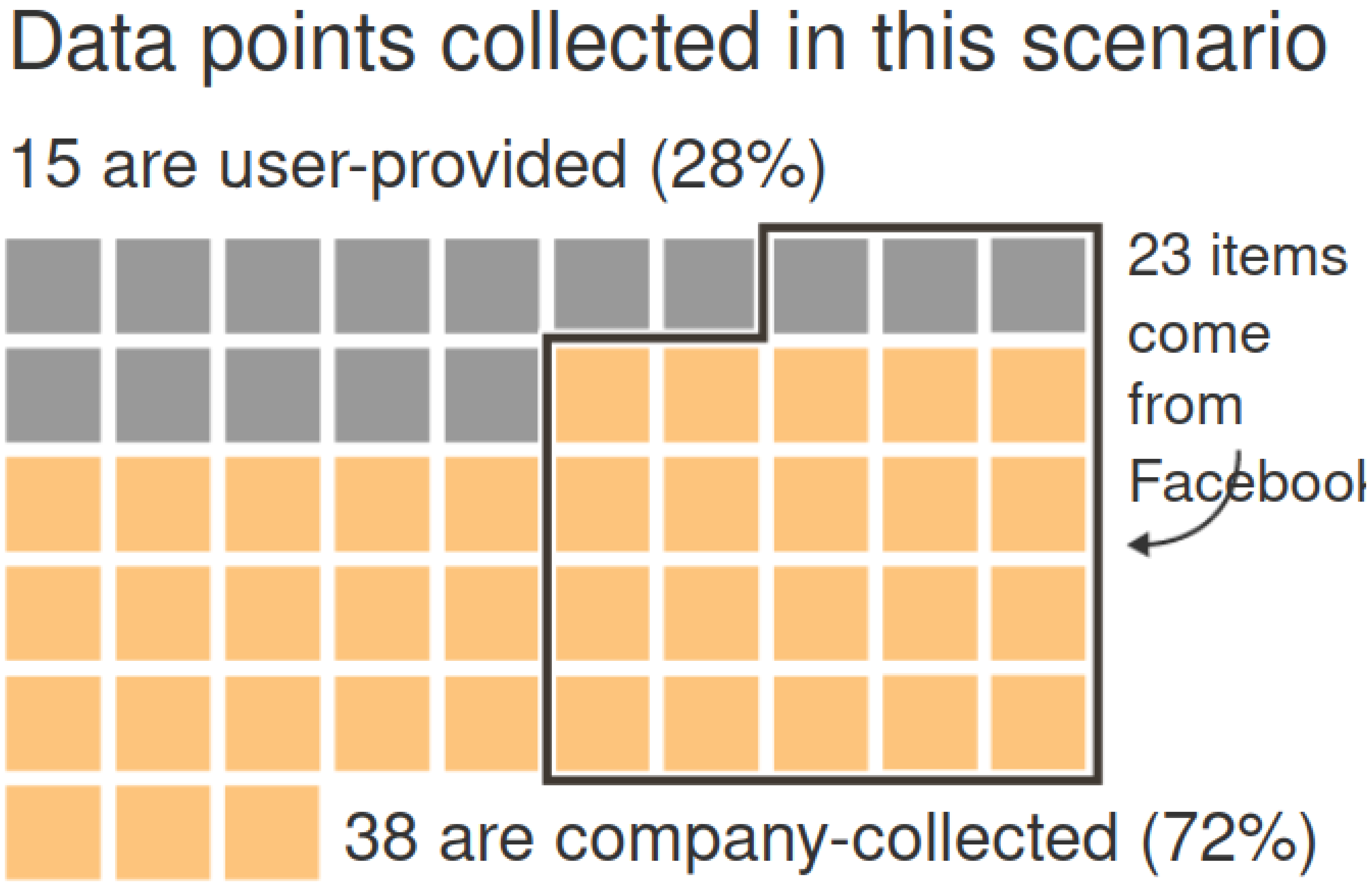


ADDITIONAL DATA COLLECTED

FACEBOOK

- Photo analysis
- Location of the photo (if included in metadata)
- Date
- Type of device (iPhone X)
- Device ID
- Device operating system
- Battery level
- Signal strength
- Bluetooth signal
- Connection speed
- Available storage
- App and file names and types
- Nearby Wi-Fi beacons and cell towers
- Nearby devices such as a TV for phone-to-TV streaming
- Time zone
- Mobile operator or internet service provider
- IP address
- Time, frequency and duration of activities
- Hardware version
- Software version

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/



Data points collected in this scenario

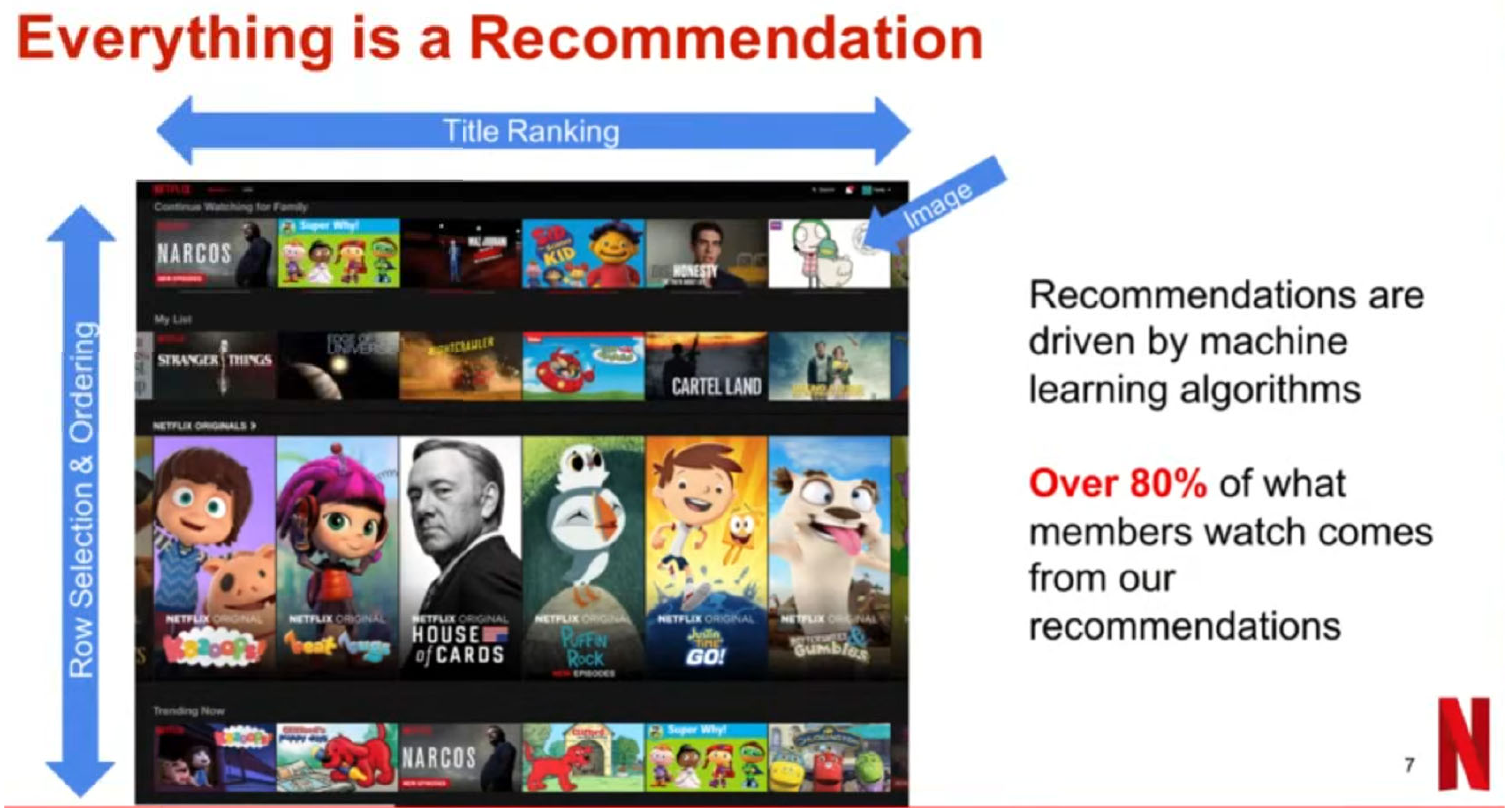
15 are user-provided (28%)

23 items come from Facebook

38 are company-collected (72%)

https://www.wsj.com/graphics/how-pizza-night-can-cost-more-in-data-than-dollars/

NETFLIX



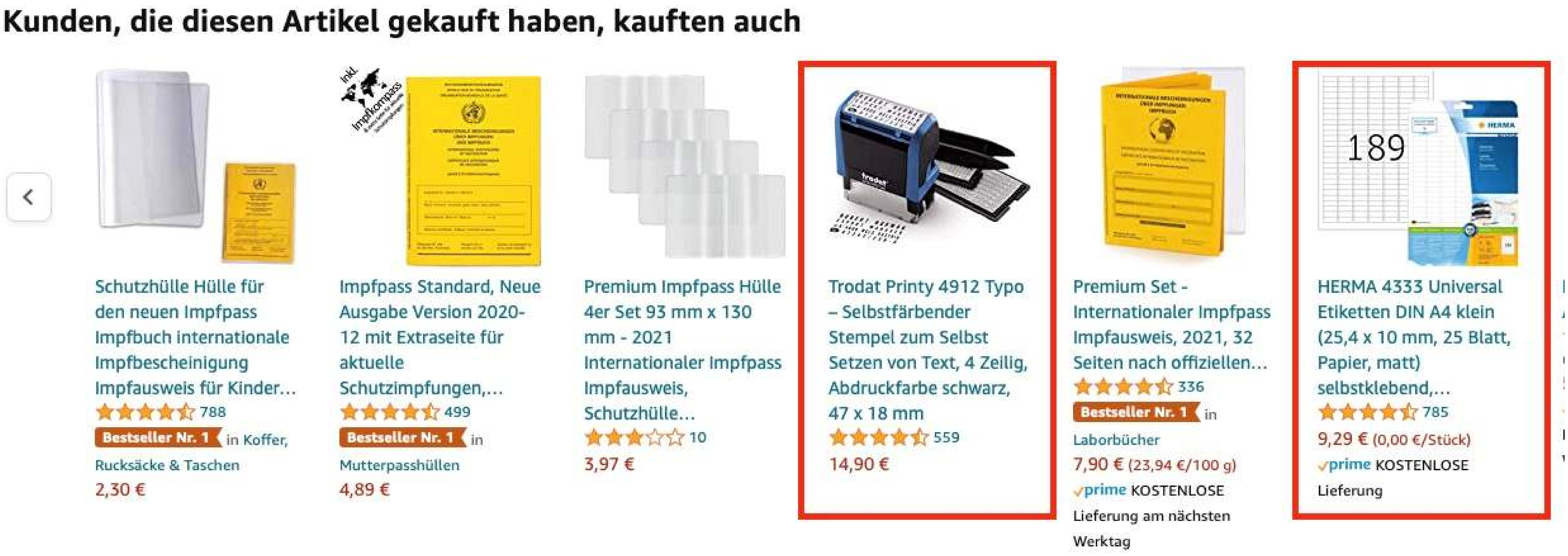
Everything is a Recommendation

Recommendations are driven by machine learning algorithms

Over 80% of what members watch comes from our recommendations

7

https://www.youtube.com/watch?v=VfWeFlZSHe8 amazon.de®



Kunden, die diesen Artikel gekauft haben, kauften auch

Schutzhülle Hülle für den neuen Impfpass Impfbuch internationale Impfbescheinigung Impfausweis für Kinder… ★★★★☆ 788 Bestseller Nr. 1 in Koffer, Rucksäcke & Taschen 2,30 €

Impfpass Standard, Neue Ausgabe Version 2020-12 mit Extraseite für aktuelle Schutzimpfungen,… ★★★★☆ 499 Bestseller Nr. 1 in Mutterpasshüllen 4,89 €

Premium Impfpass Hülle 4er Set 93 mm x 130 mm - 2021 Internationaler Impfpass Impfausweis, Schutzhülle… ★★★☆☆ 10 3,97 €

Trodat Printy 4912 Typo – Selbstfärbender Stempel zum Selbst Setzen von Text, 4 Zeilig, Abdruckfarbe schwarz, 47 x 18 mm ★★★★☆ 559 14,90 €

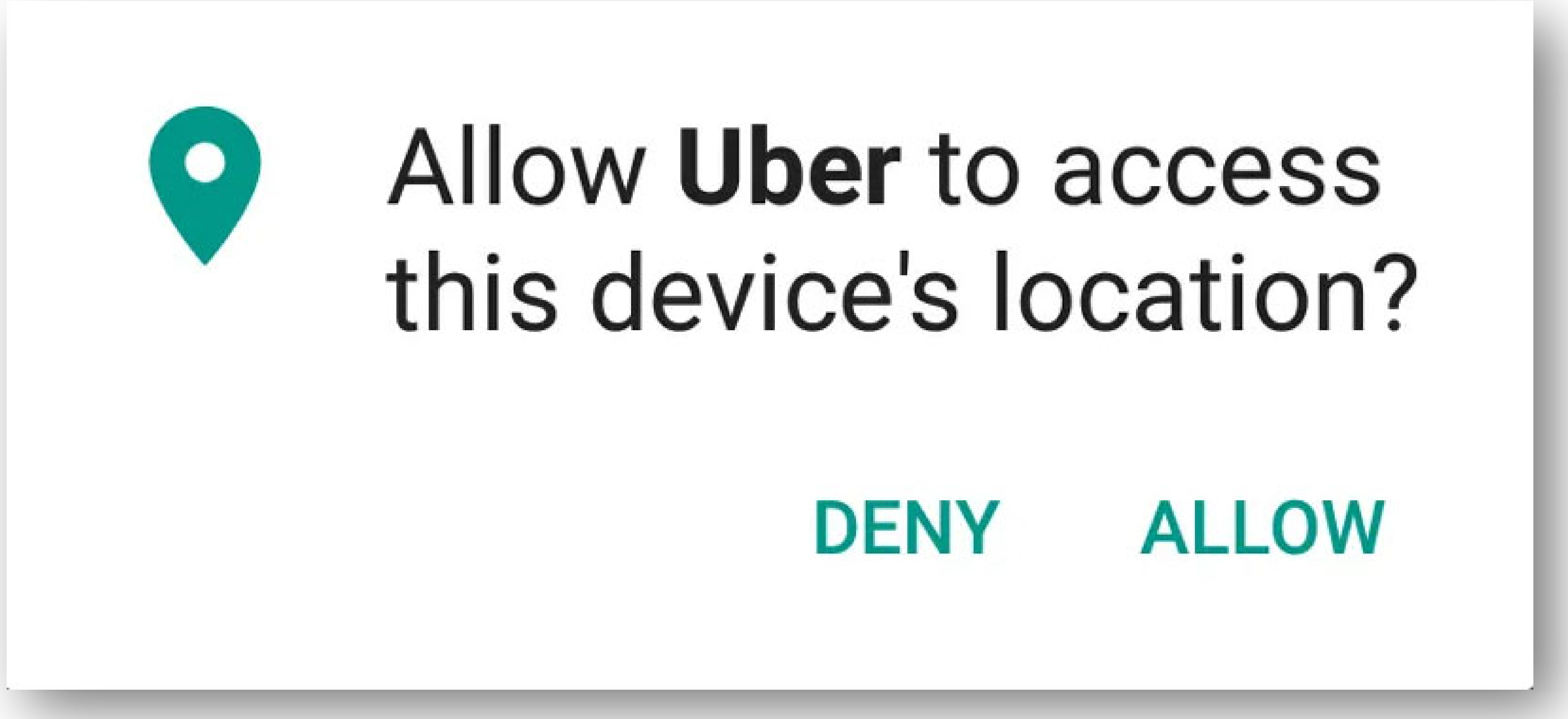
Premium Set - Internationaler Impfpass Impfausweis, 2021, 32 Seiten nach offiziellen… ★★★★☆ 336 Bestseller Nr. 1 in Laborbücher 7,90 € (23,94 €/100 g) ✓prime KOSTENLOSE Lieferung am nächsten Werktag

HERMA 4333 Universal Etiketten DIN A4 klein (25,4 x 10 mm, 25 Blatt, Papier, matt) selbstklebend,… ★★★★☆ 785 9,29 € (0,00 €/Stück) ✓prime KOSTENLOSE Lieferung

https://twitter.com/balkonschlaefer/status/1395656386839814147

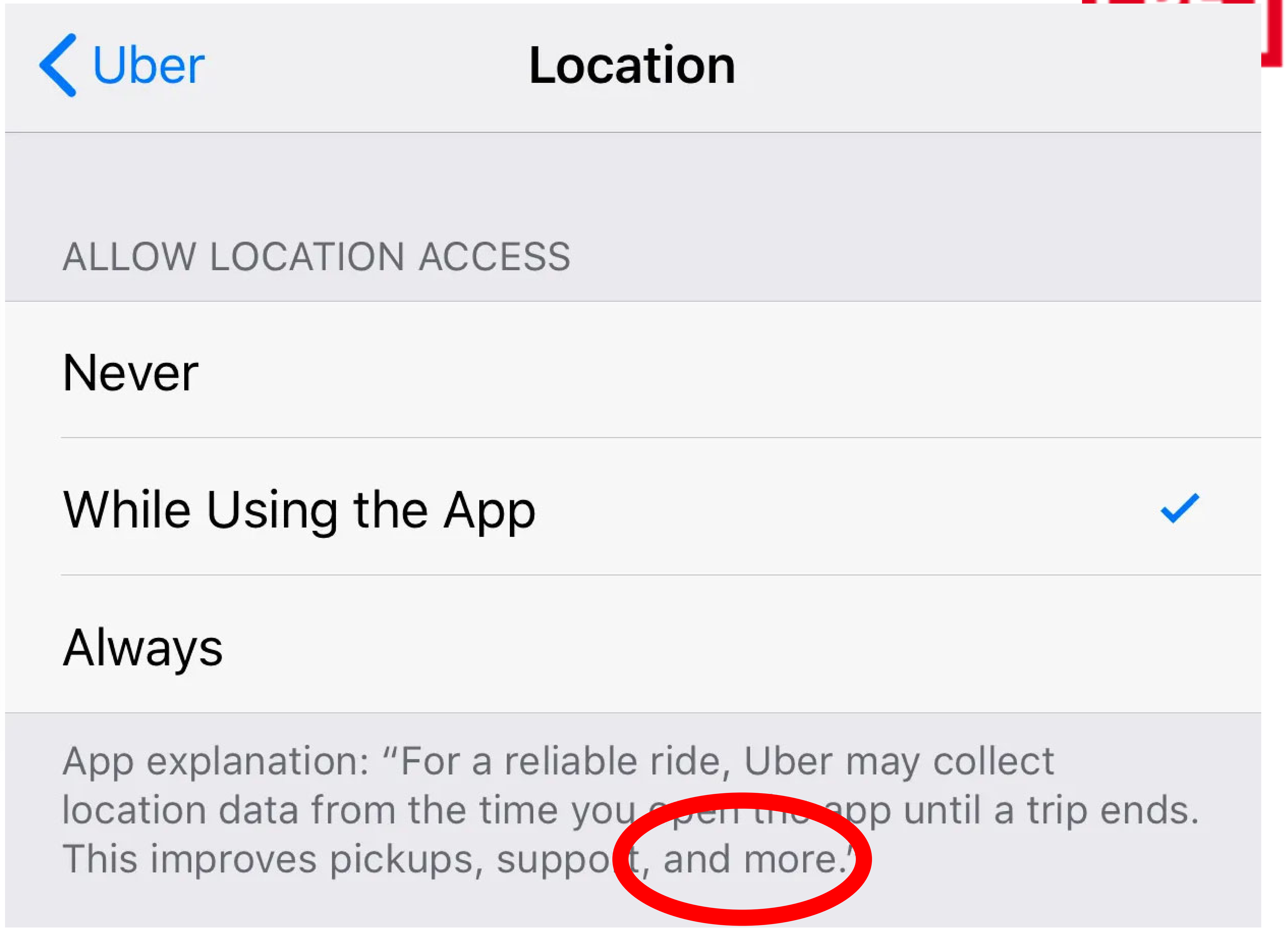
How do you sustain a business model in which users don’t pay for your service?
Senator, we run ads
NEWS SPECIAL REPORT
LIVE
FACEBOOK CEO MARK ZUCKERBERG TESTIFIES BEFORE SENATE
NBC NEWS

https://www.youtube.com/watch?v=n2H8wx1aBiQ



Allow **Uber** to access this device's location?

DENY ALLOW



< Uber **Location**

ALLOW LOCATION ACCESS

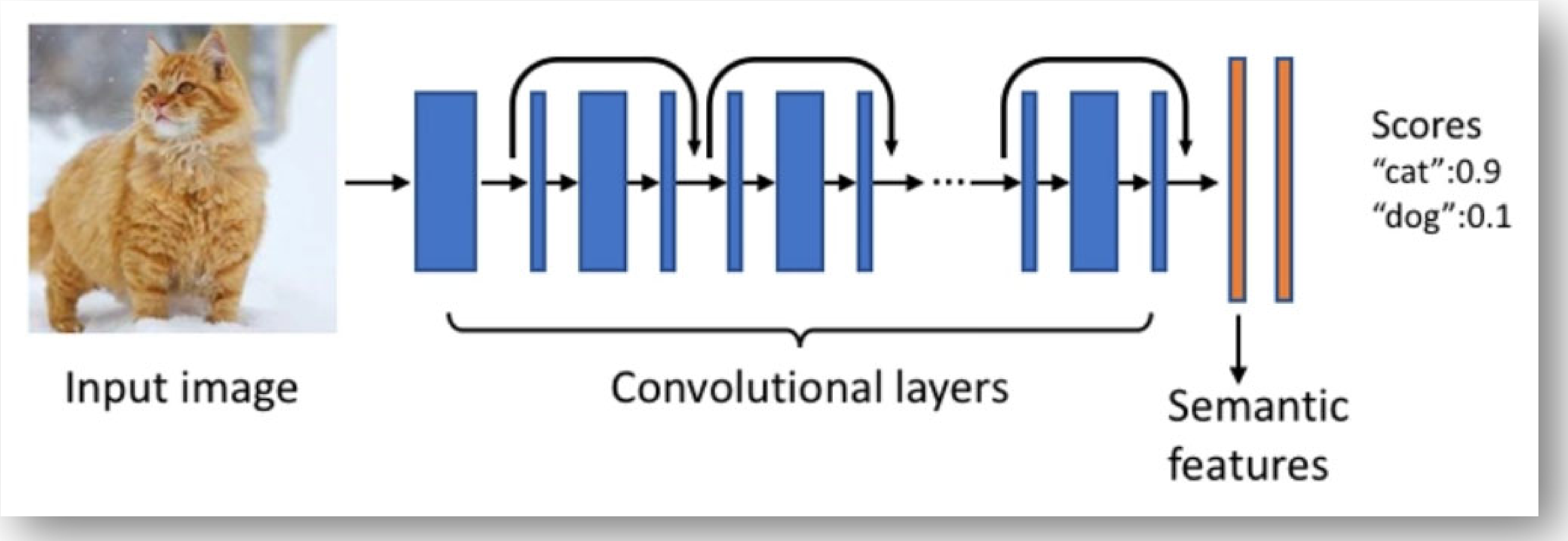
Never

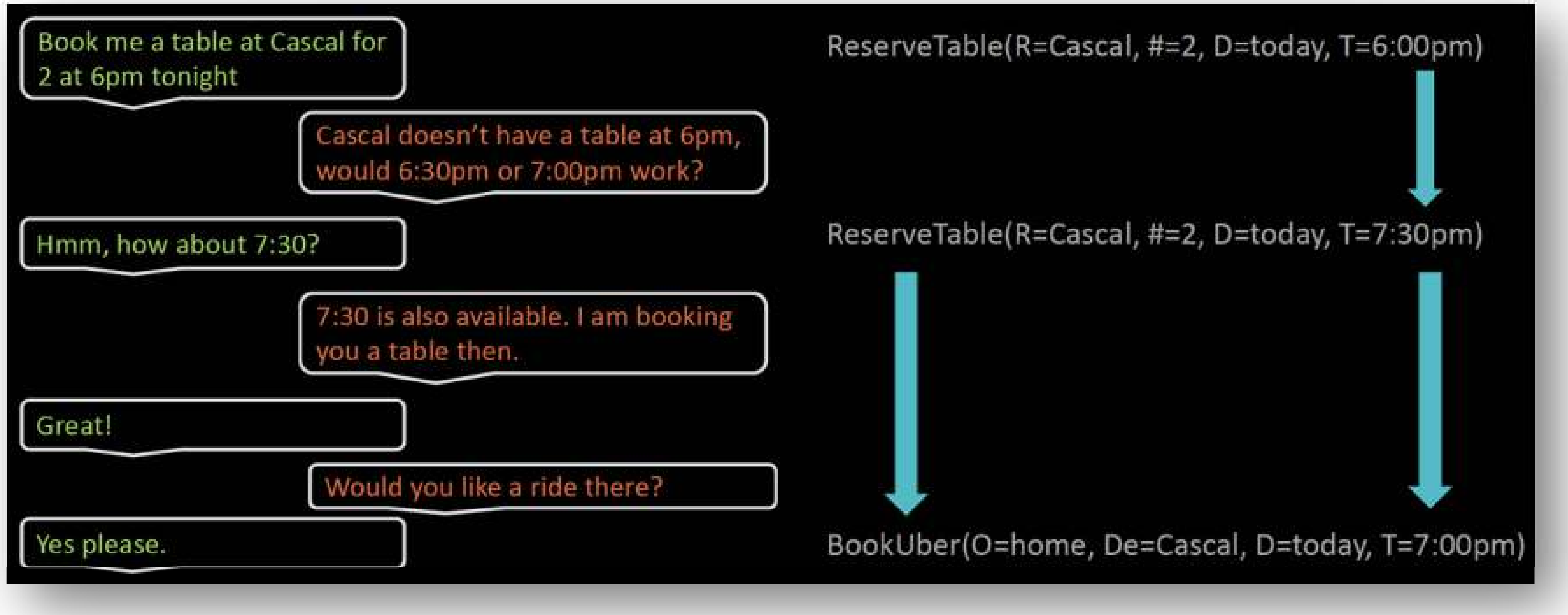
While Using the App ✓

Always

App explanation: “For a reliable ride, Uber may collect location data from the time you open the app until a trip ends. This improves pickups, support, and more.”

https://www.wired.com/story/app-permissions/





Book me a table at Cascal for 2 at 6pm tonight

ReserveTable(R=Cascal, #=2, D=today, T=6:00pm)

Cascal doesn't have a table at 6pm, would 6:30pm or 7:00pm work?

Hmm, how about 7:30?

ReserveTable(R=Cascal, #=2, D=today, T=7:30pm)

7:30 is also available. I am booking you a table then.

Great!

Would you like a ride there?

Yes please.

BookUber(O=home, De=Cascal, D=today, T=7:00pm)

https://en.wikipedia.org/wiki/Self-driving_car
https://www.amazon.science/blog/new-alexa-research-on-task-oriented-dialogue-systems
https://engineering.fb.com/2017/02/02/ml-applications/building-scalable-systems-to-understand-content/

> When Google set out to scan the pages of millions of books, it not only digitized the pages but it also datafied the text so that letters, words and paragraphs could be read and indexed and searched. An estimated 130 million unique books have been published since the invention of the printing press, estimate the authors. As of 2012, Google had scanned over 20 million titles, more than 15 percent of the world’s books. This data has multiple uses, only one of which is actually reading a book. For example, the project allows scholars to discover when certain words or phrases are used for the first time. The Google project has also been used to facilitate the accuracy of Google’s language translation algorithms. Other key sectors where datafication is changing our world is the datafication of location through GPS and cell phone signals, and the datafication of relationships, i.e. Facebook’s one billion users and 100 billion “friendships.”

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

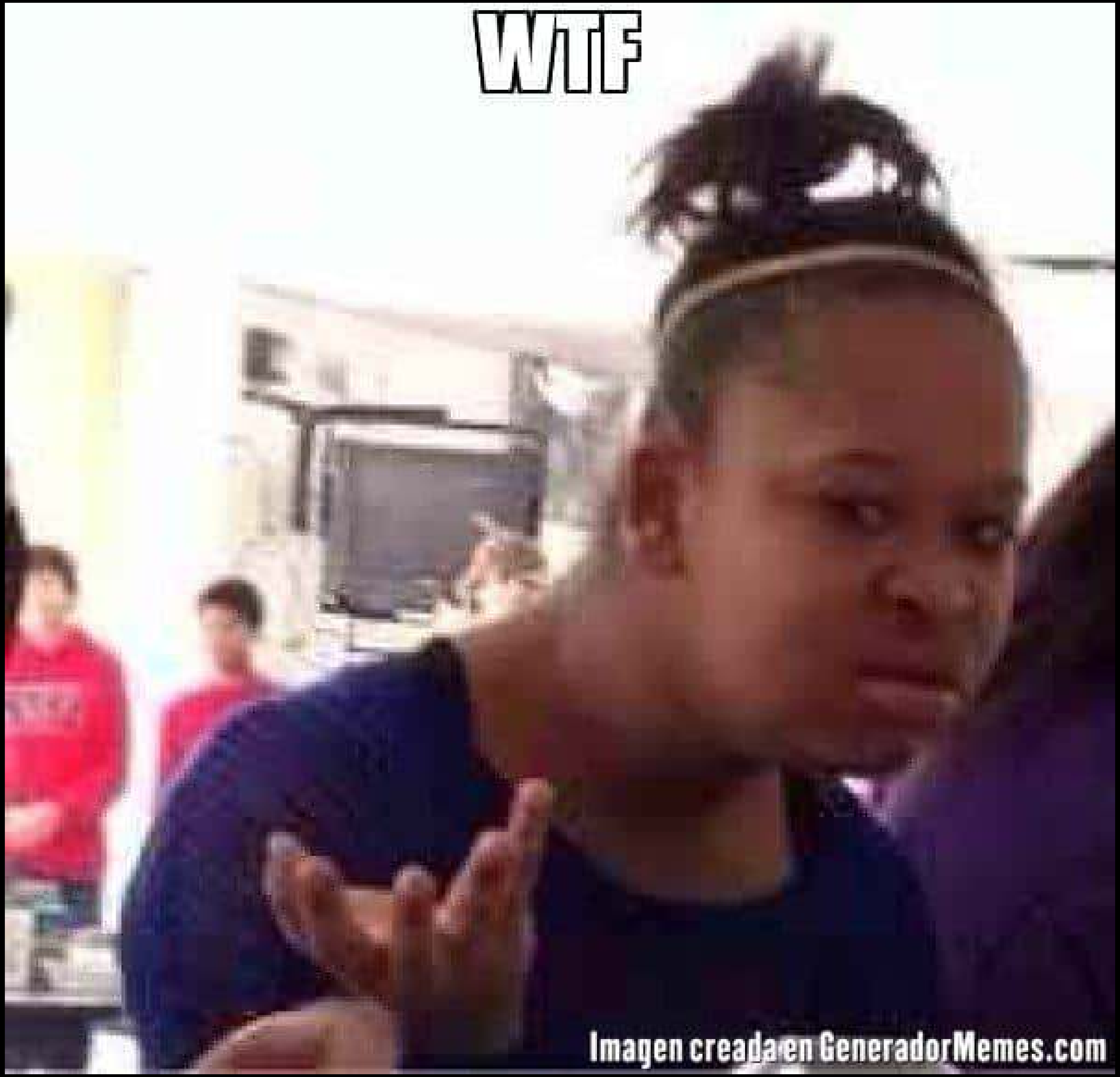
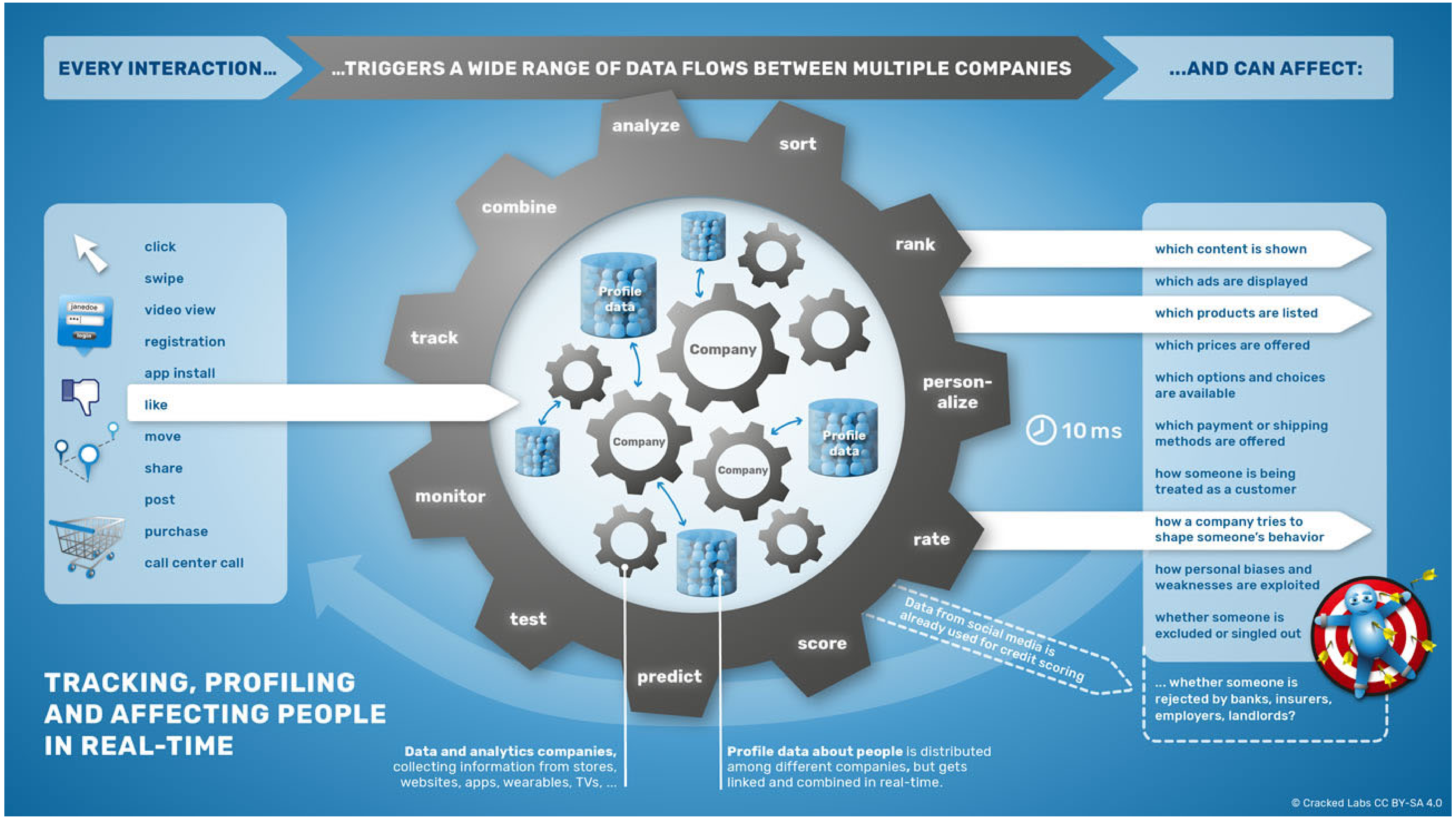


Imagen creada en GeneradorMemes.com

So…what‘s the problem?

- Tracking has become almost inevitable
- No realistic options to avoid it
- no informational self-determination
- Intransparency of the process
- High level of intrusion into privacy
- "context collapse" = convergence of actually separate social circles
- Already taking over too many (uncontrolled) coordination, orientation and order functions

What can become data?



EVERY INTERACTION… …TRIGGERS A WIDE RANGE OF DATA FLOWS BETWEEN MULTIPLE COMPANIES …AND CAN AFFECT:

- which content is shown
- which ads are displayed
- which products are listed
- which prices are offered
- which options and choices are available
- which payment or shipping methods are offered
- how someone is being treated as a customer
- how a company tries to shape someone's behavior
- how personal biases and weaknesses are exploited
- whether someone is excluded or singled out
- … whether someone is rejected by banks, insurers, employers, landlords?

Data from social media is already used for credit scoring

TRACKING, PROFILING AND AFFECTING PEOPLE IN REAL-TIME

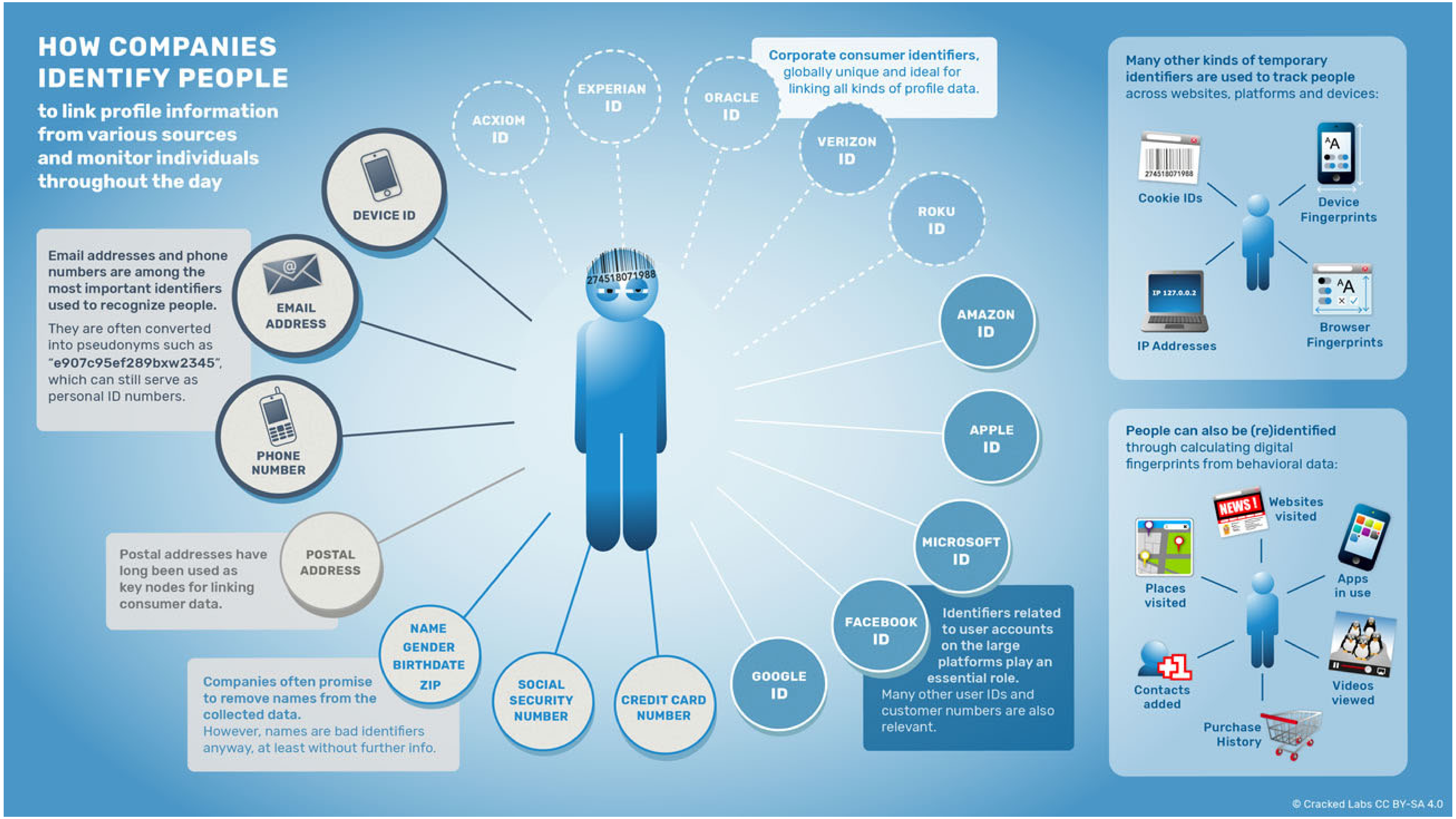
Data and analytics companies, collecting information from stores, websites, apps, wearables, TVs, …

Profile data about people is distributed among different companies, but gets linked and combined in real-time.

© Cracked Labs CC BY-SA 4.0

https://crackedlabs.org/en/corporate-surveillance

What can become data?



HOW COMPANIES IDENTIFY PEOPLE

to link profile information from various sources and monitor individuals throughout the day

Email addresses and phone numbers are among the most important identifiers used to recognize people.

They are often converted into pseudonyms such as "e907c95ef289bxw2345", which can still serve as personal ID numbers.

Postal addresses have long been used as key nodes for linking consumer data.

Companies often promise to remove names from the collected data. However, names are bad identifiers anyway, at least without further info.

Corporate consumer identifiers, globally unique and ideal for linking all kinds of profile data.

Identifiers related to user accounts on the large platforms play an essential role. Many other user IDs and customer numbers are also relevant.

Many other kinds of temporary identifiers are used to track people across websites, platforms and devices:

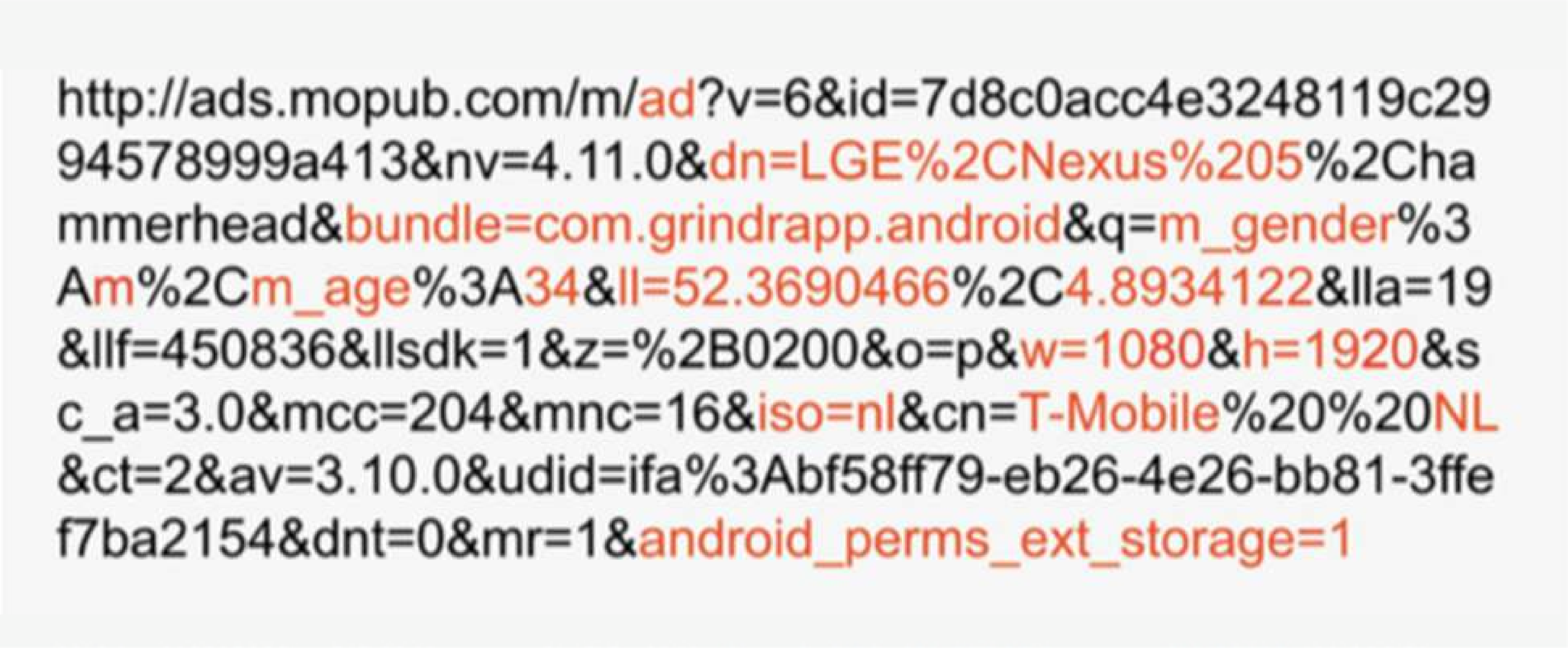
People can also be (re)identified through calculating digital fingerprints from behavioral data:

© Cracked Labs CC BY-SA 4.0

https://crackedlabs.org/en/corporate-surveillance

Data Flow

- Data transfer across different locations (and limits of jurisdiction).
- Contents of the data transfers
- In which processes is the dataflow involved?
- Who is behind the various nodes in the network?



```
http://ads.mopub.com/m/ad?v=6&id=7d8c0acc4e3248119c29
94578999a413&nv=4.11.0&dn=LGE%2CNexus%205%2Cha
mmerhead&bundle=com.grindrapp.android&q=m_gender%3
Am%2Cm_age%3A34&ll=52.3690466%2C4.8934122&lla=19
&llf=450836&llsdk=1&z=%2B0200&o=p&w=1080&h=1920&s
c_a=3.0&mcc=204&mnc=16&iso=nl&cn=T-Mobile%20%20NL
&ct=2&av=3.10.0&udid=ifa%3Abf58ff79-eb26-4e26-bb81-3ffe
f7ba2154&dnt=0&mr=1&android_perms_ext_storage=1
```

Fig. 3: An encoded MoPub URL with unencrypted HTTP ad request parameters and values (device name, bundle ID, gender, age, lat long, screen width, height, language, carrier network, permissions).

Dieter et al 2018 https://www001.zimt.uni-siegen.de/ojs/index.php/wps1187/article/view/29/31

Stimuli

- How do your observations differ from what the vendors or media say? (e.g. privacy policy)
- Think about what your research means in terms of data protection, personal law but also informatics and data science.
- Where should the responsibilities lie?
- Who should regulate it?
- How can you react as a user?

NOW AND THEN, I ANNOUNCE "I KNOW YOU'RE LISTENING" TO EMPTY ROOMS.
IF I'M WRONG, NO ONE KNOWS.
AND IF I'M RIGHT, MAYBE I JUST FREAKED THE HELL OUT OF SOME SECRET ORGANIZATION.

https://xkcd.com/525/

Thanks.

mirco.schoenfeld@uni-bayreuth.de