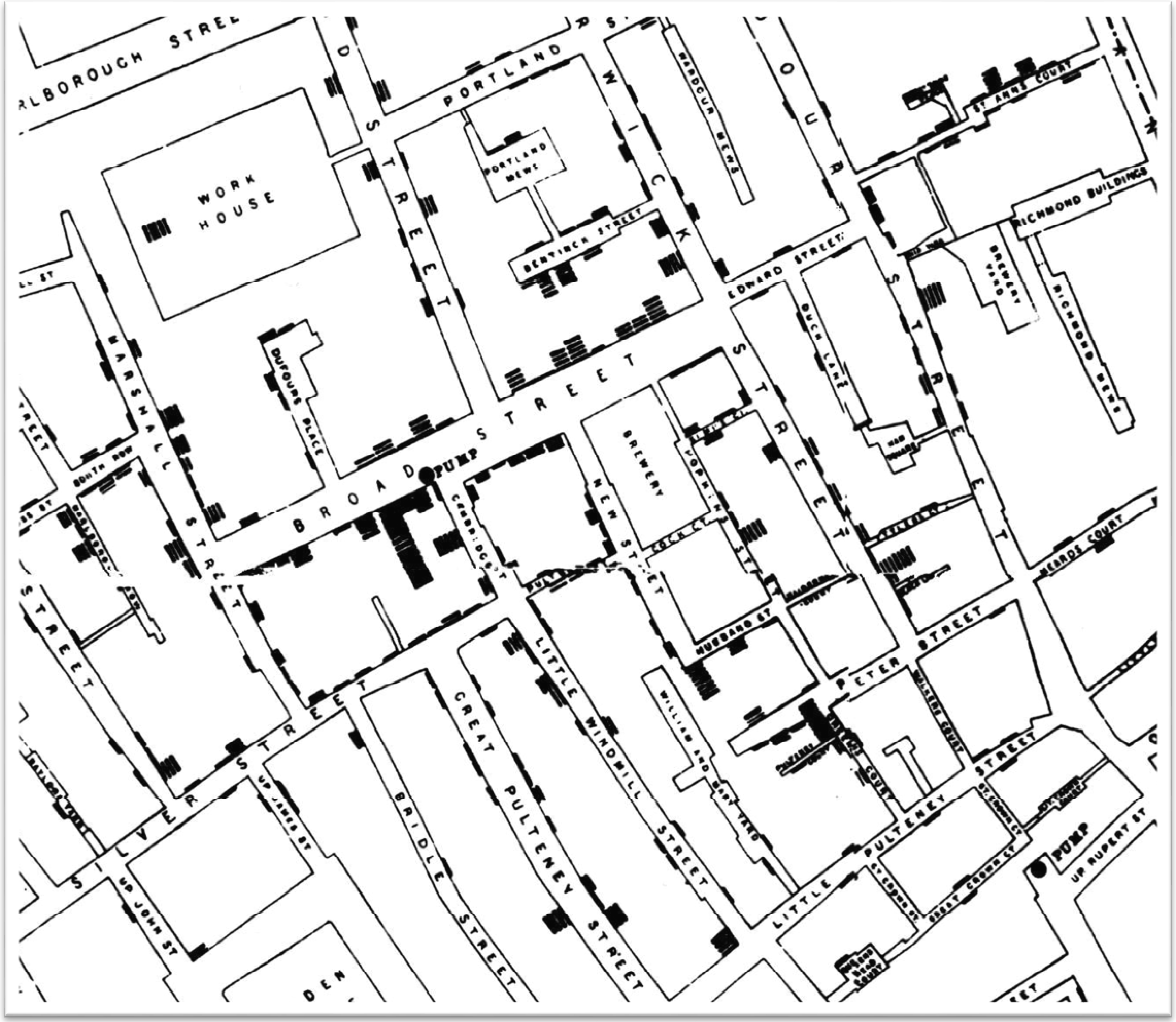
Data Modeling & Knowledge Generation

Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
@TWlyY29

What do you think is data modeling & knowledge generation



Snow, John (1854).

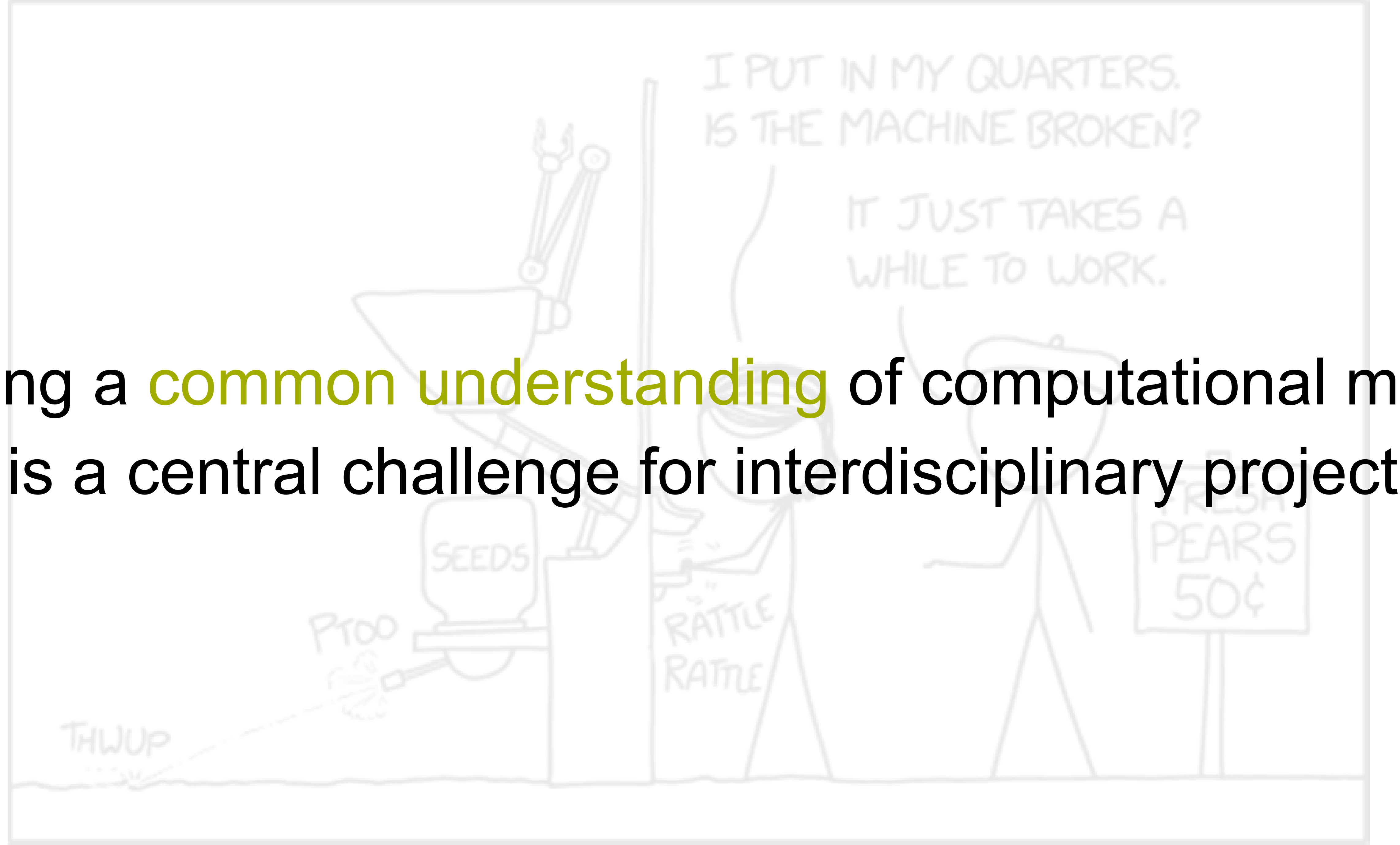
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niémen R.
422.000
6.000
22.000
400.000
Kowno
Wilna
60.000
33.000
Gloubokoe
Polotzk
Witebsk
175.000
145.000
Smolensk
Dorogobouge
Chjat
Moskowa R.
127.100
100.000
MOSCOU
100.000
100.000
Tarantino
Mojaisk
Malo-jarosewli
96.000
87.000
Wizma
55.000
37.000
24.000
Orscha
Mohilow
20.000
Bobr
50.000
30.000
Studienska
Bérézina R.
28.000
Minsk
Molodezno
12.000
Smorgoni
14.000
8.000
4.000
10.000
Niémen R.
Lieues communes de France (Carte de M.r de Fezensac)
0 5 10 15 20 25 50.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8.bre
Pluie 24 8.bre
– 9.° le 9 9.bre
– 21.° le 14 9.bre
– 11.°
– 20.° le 28 9.bre
– 24.° le 1.er X.bre
– 30.° le 6 X.bre
– 26.° le 7 X.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

Minard, Charles Joseph (1869).

Dude, that stuff is almost 200 years old…

Data Mining, Machine Learning, and Artificial Intelligence

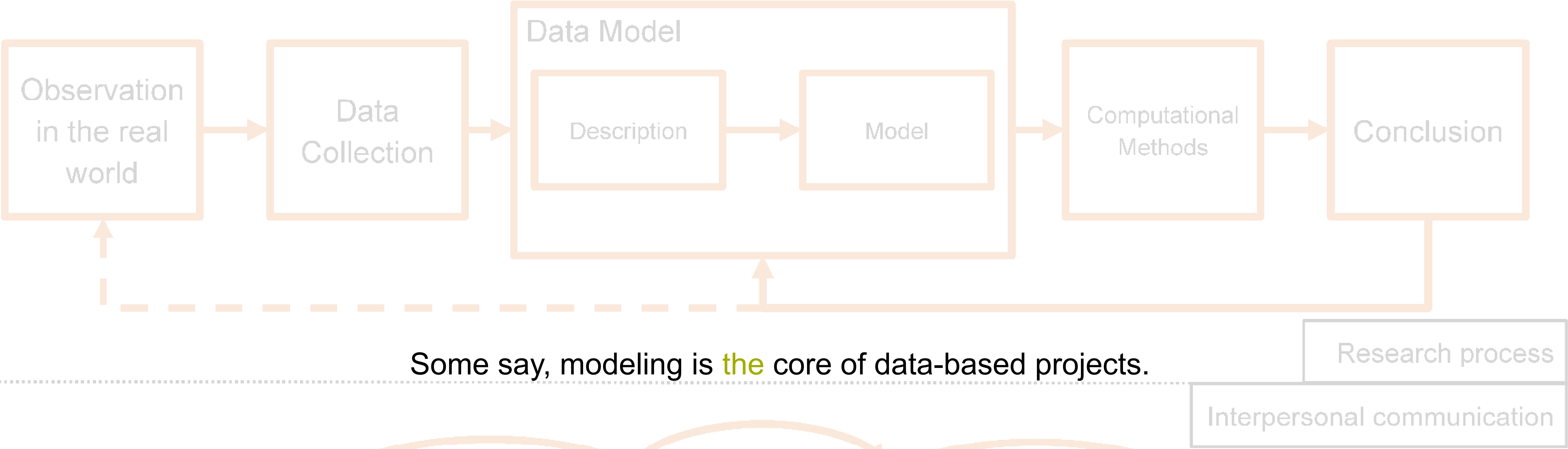
Why do you think you need data modeling & knowledge generation at all?


I PUT IN MY QUARTERS. IS THE MACHINE BROKEN?
IT JUST TAKES A WHILE TO WORK.
SEEDS
PTOO
THWUP
RATTLE RATTLE
FRESH PEARS 50¢

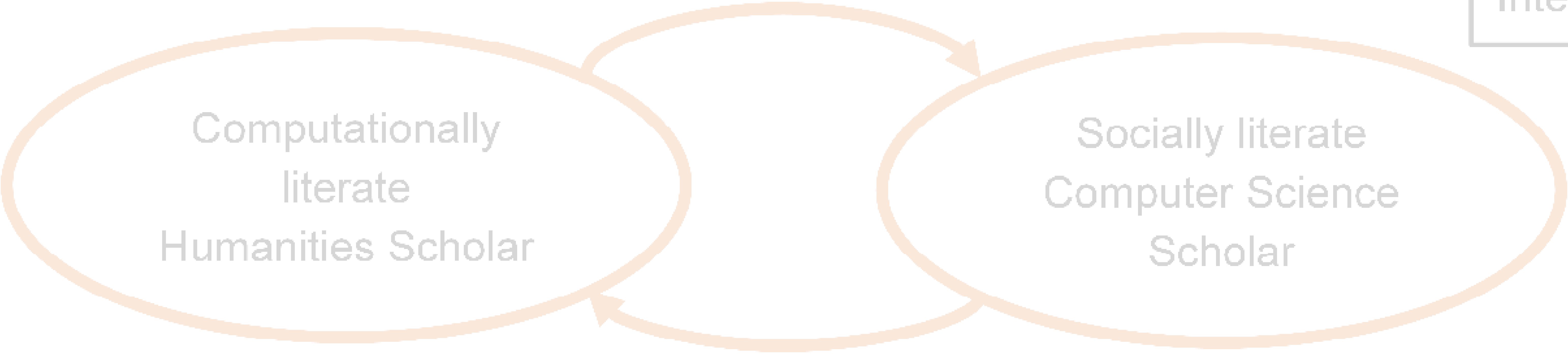
https://www.xkcd.com/2209/

Finding a common understanding of computational methods is a central challenge for interdisciplinary projects.

Expectations in Interdisciplinary Projects



Some say, modeling is the core of data-based projects.


Computationally literate Humanities Scholar
Socially literate Computer Science Scholar

Lazer, D., et al. (2009). Computational social science. Science, 323 (5915), 721–723.
McCarty, W. (2005). Humanities computing. Palgrave Macmillan.
Terras, M. (2012). Being the other: Interdisciplinary work in computational science and the humanities. Collaborative Research in the Digital Humanities. Farnham: Ashgate, 213-230.

> 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.

By the way, 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.

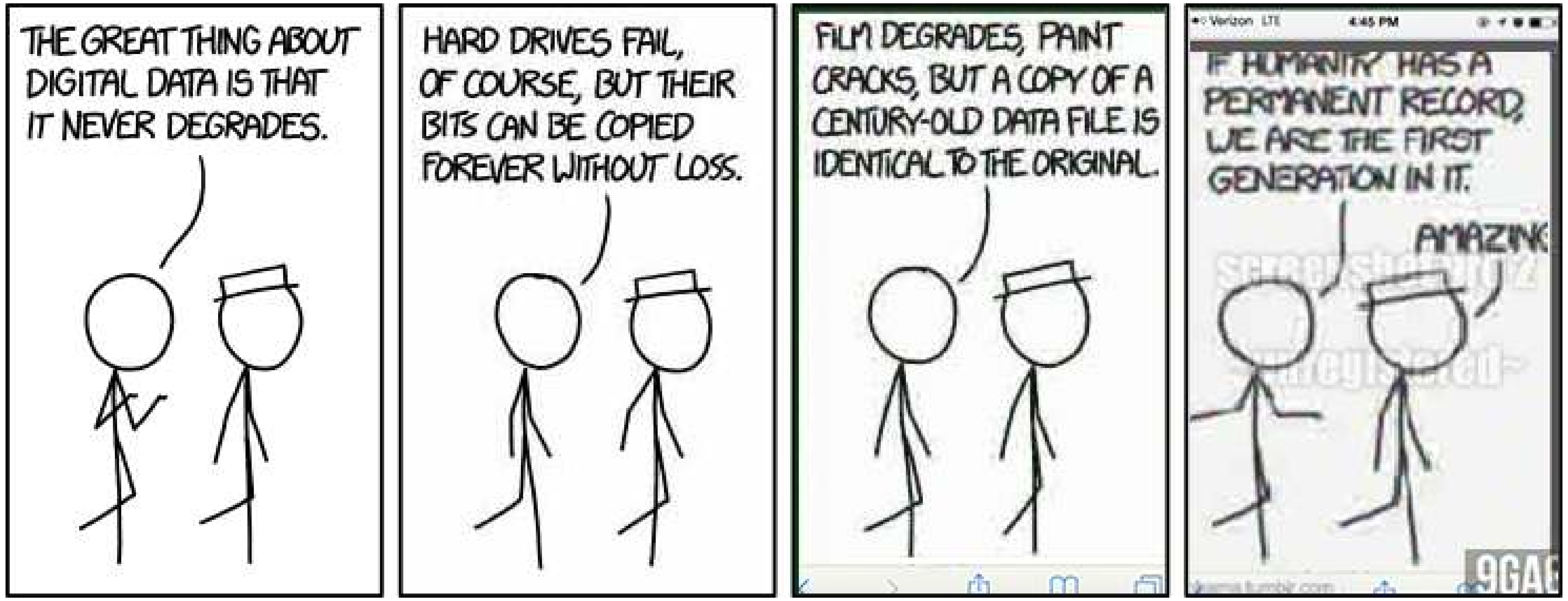
https://en.wikipedia.org/wiki/Data_(computing)

> 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

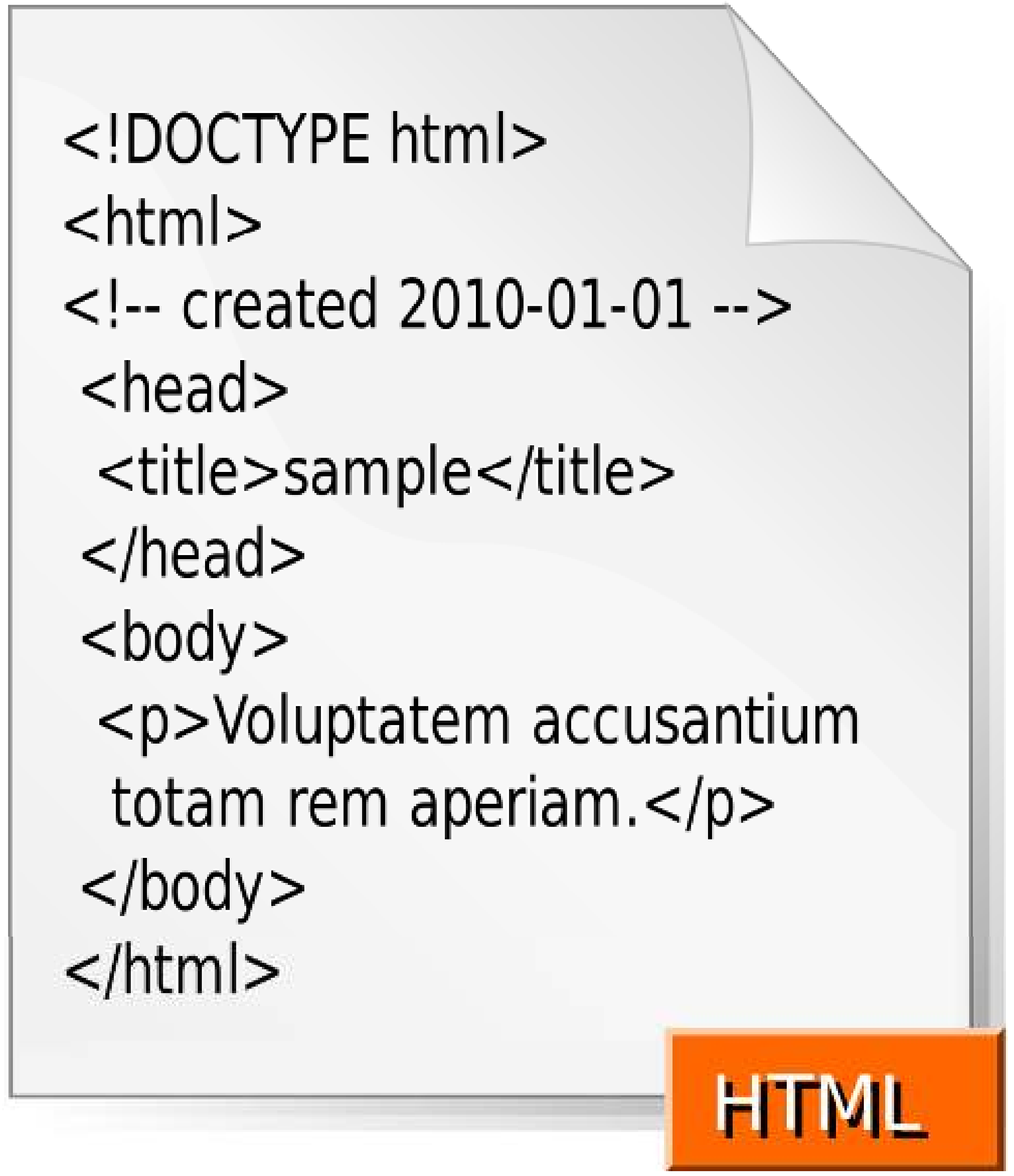

THE GREAT THING ABOUT DIGITAL DATA IS THAT IT NEVER DEGRADES.
HARD DRIVES FAIL, OF COURSE, BUT THEIR BITS CAN BE COPIED FOREVER WITHOUT LOSS.
FILM DEGRADES, PAINT CRACKS, BUT A COPY OF A CENTURY-OLD DATA FILE IS IDENTICAL TO THE ORIGINAL.
IF HUMANITY HAS A PERMANENT RECORD, WE ARE THE FIRST GENERATION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

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

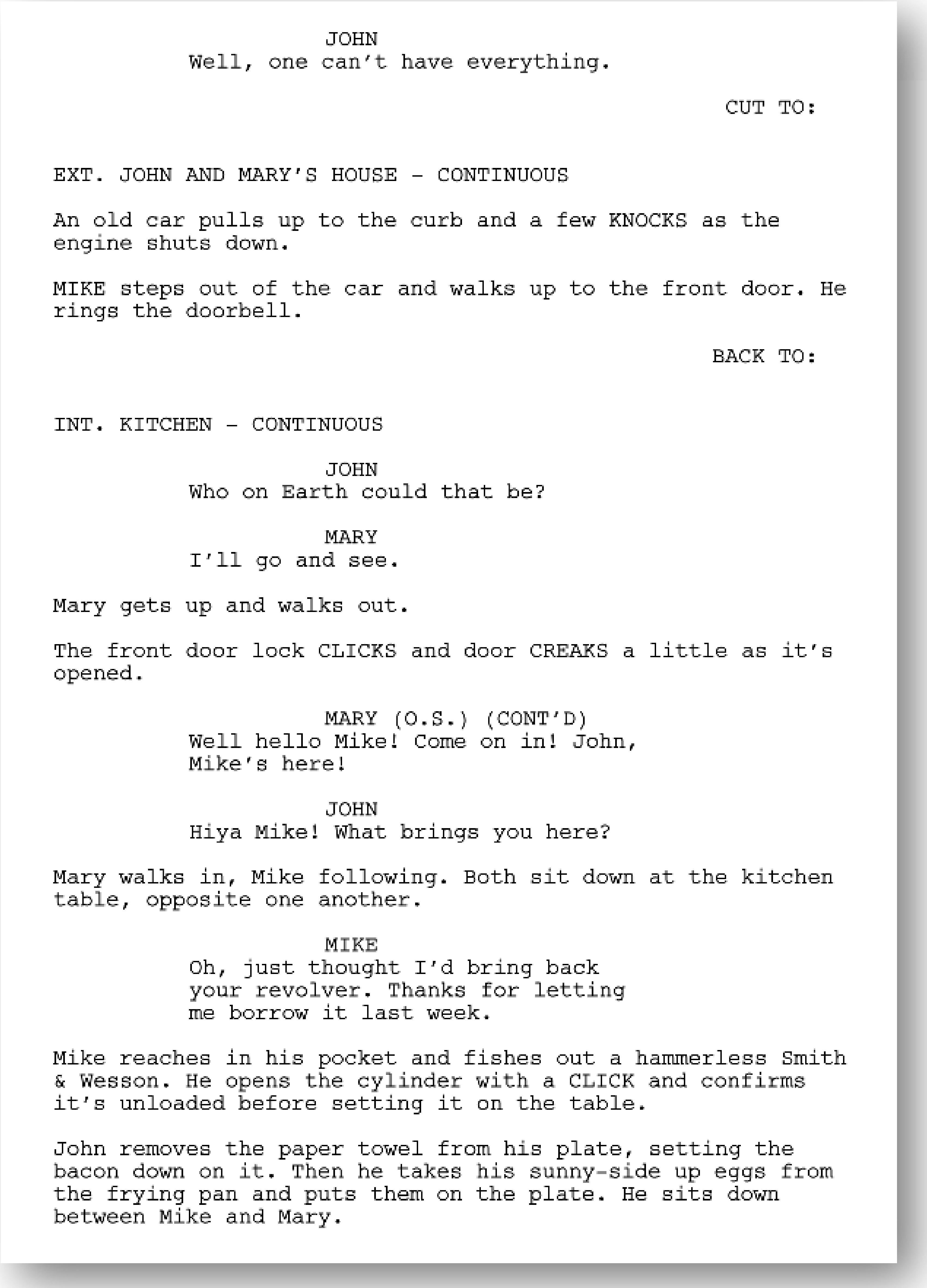
Form of data

- 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

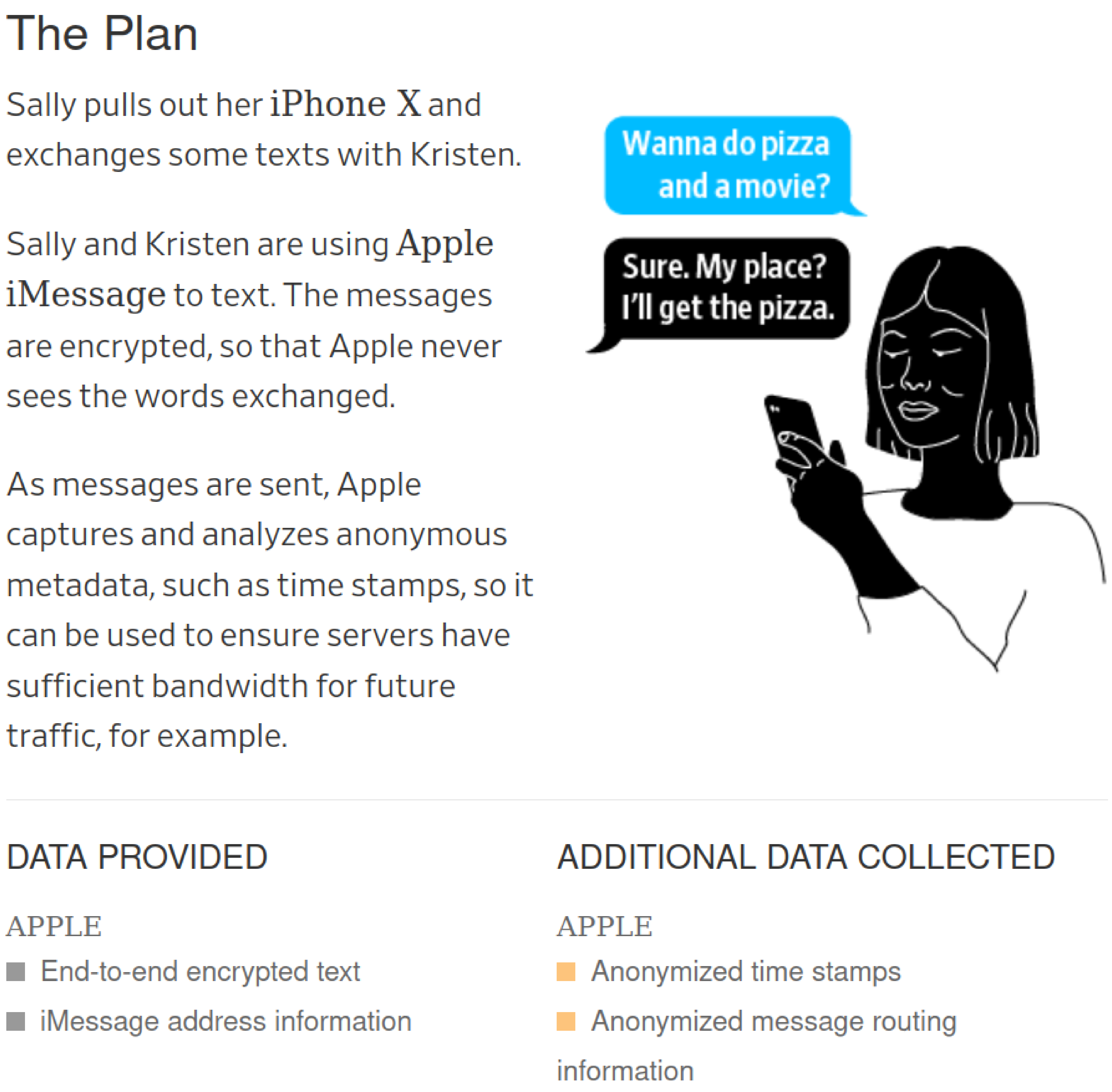
> Datafication is a modern technological trend turning many aspects of our life into computerized data and transforming this information into new forms of value.
>
> Wikipedia on “datafication”

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

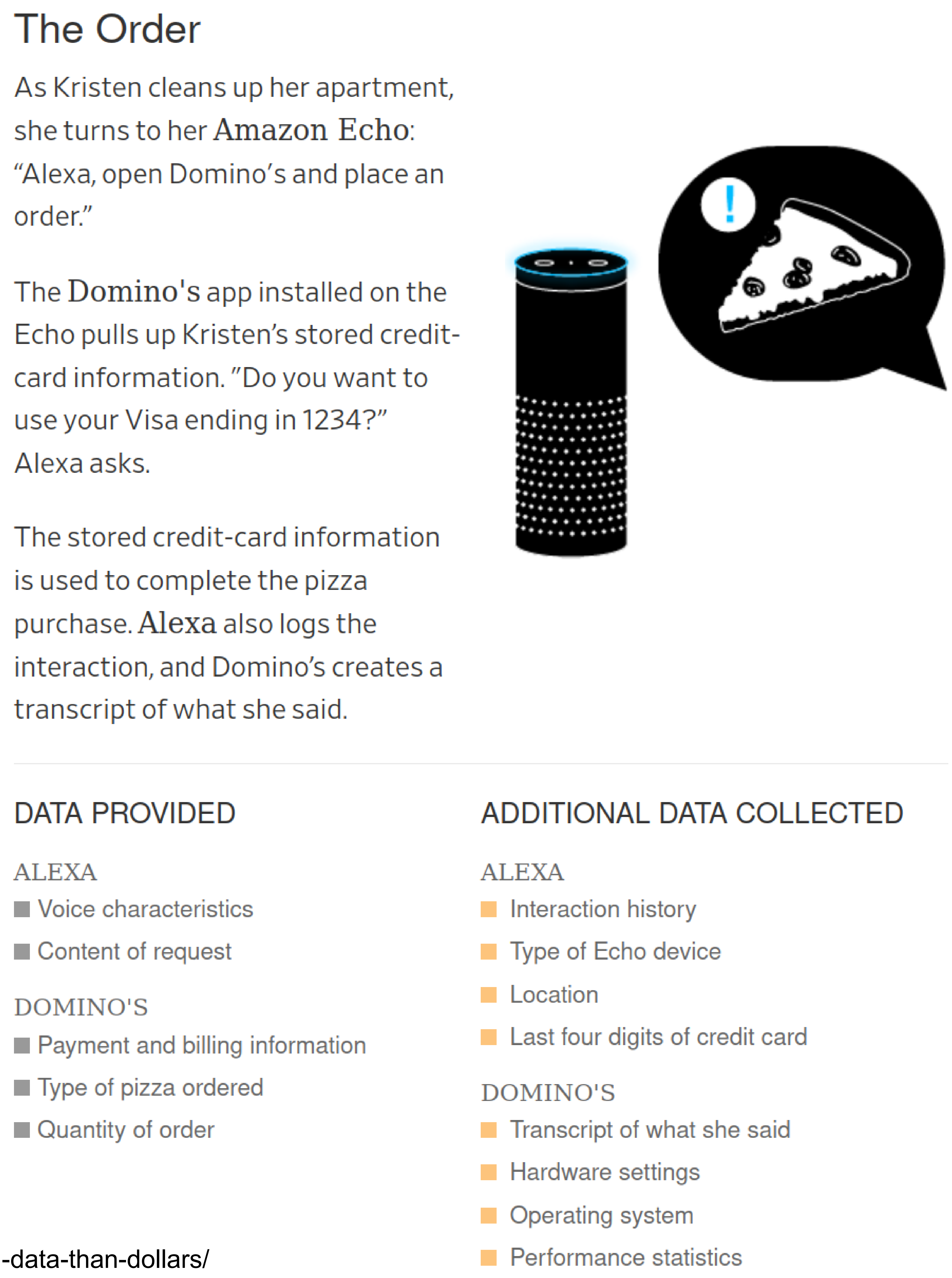
- ■ End-to-end encrypted text
- ■ iMessage address information

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

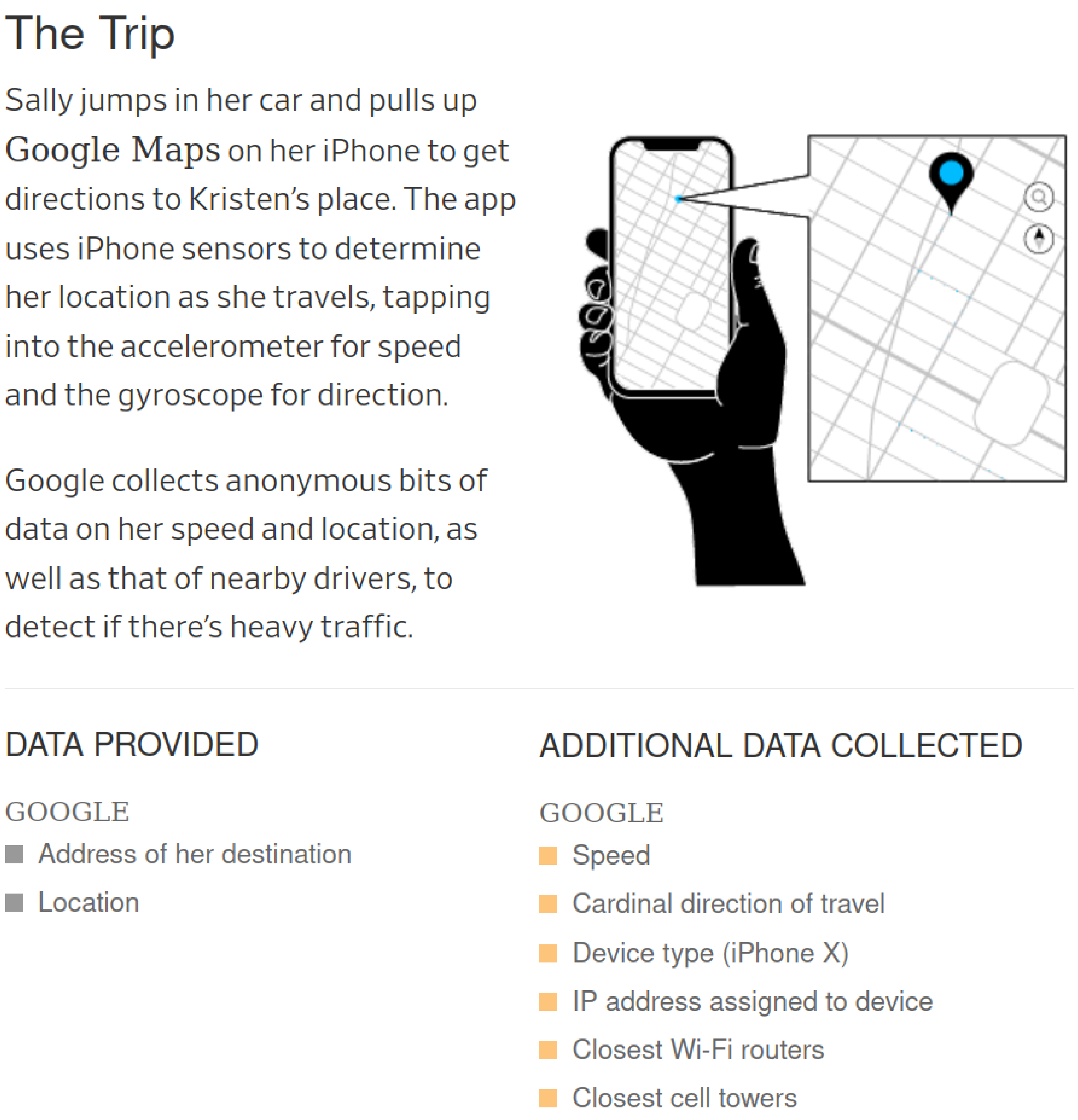
ALEXA

- 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

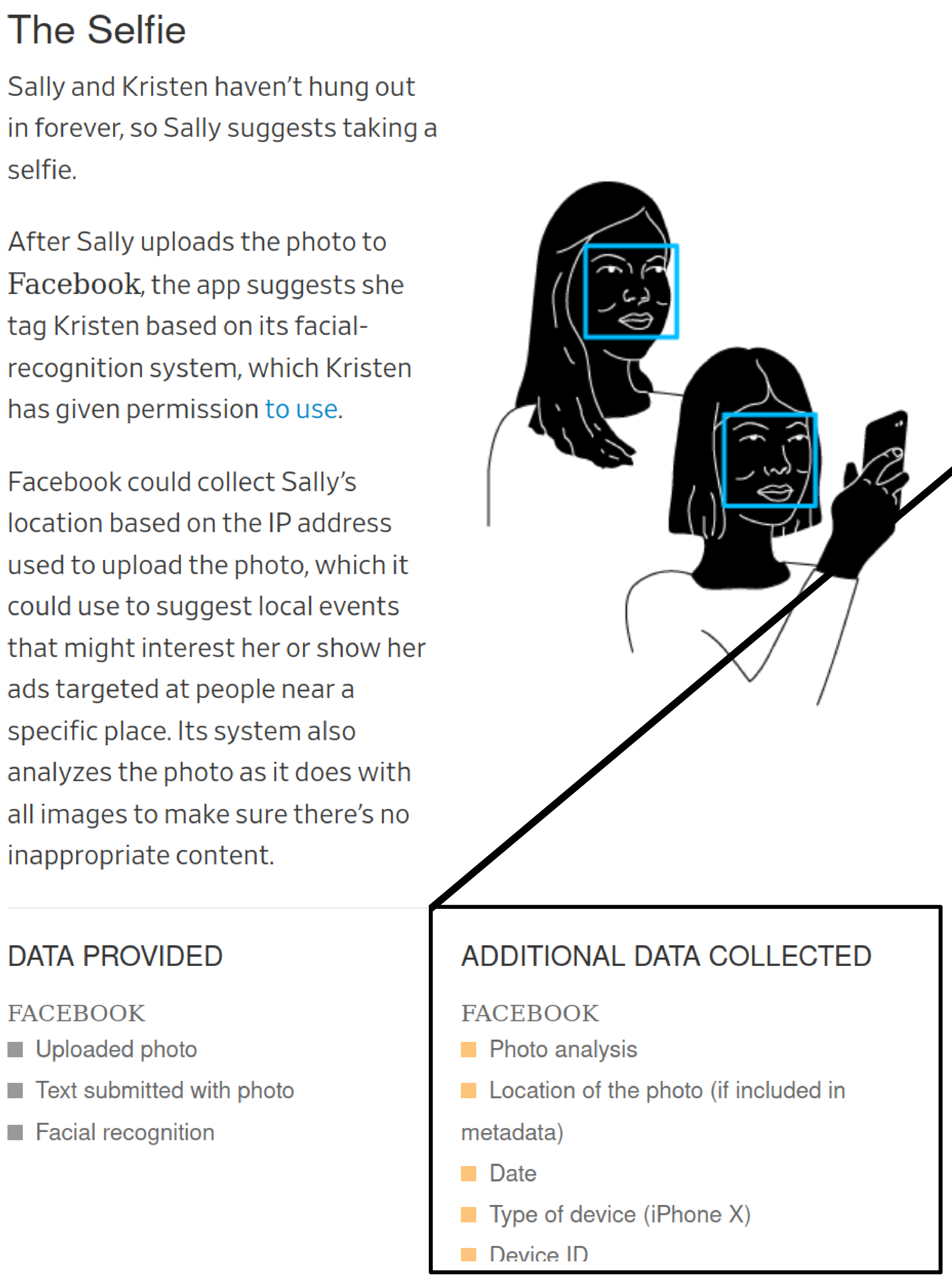
- Address of her destination
- Location

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

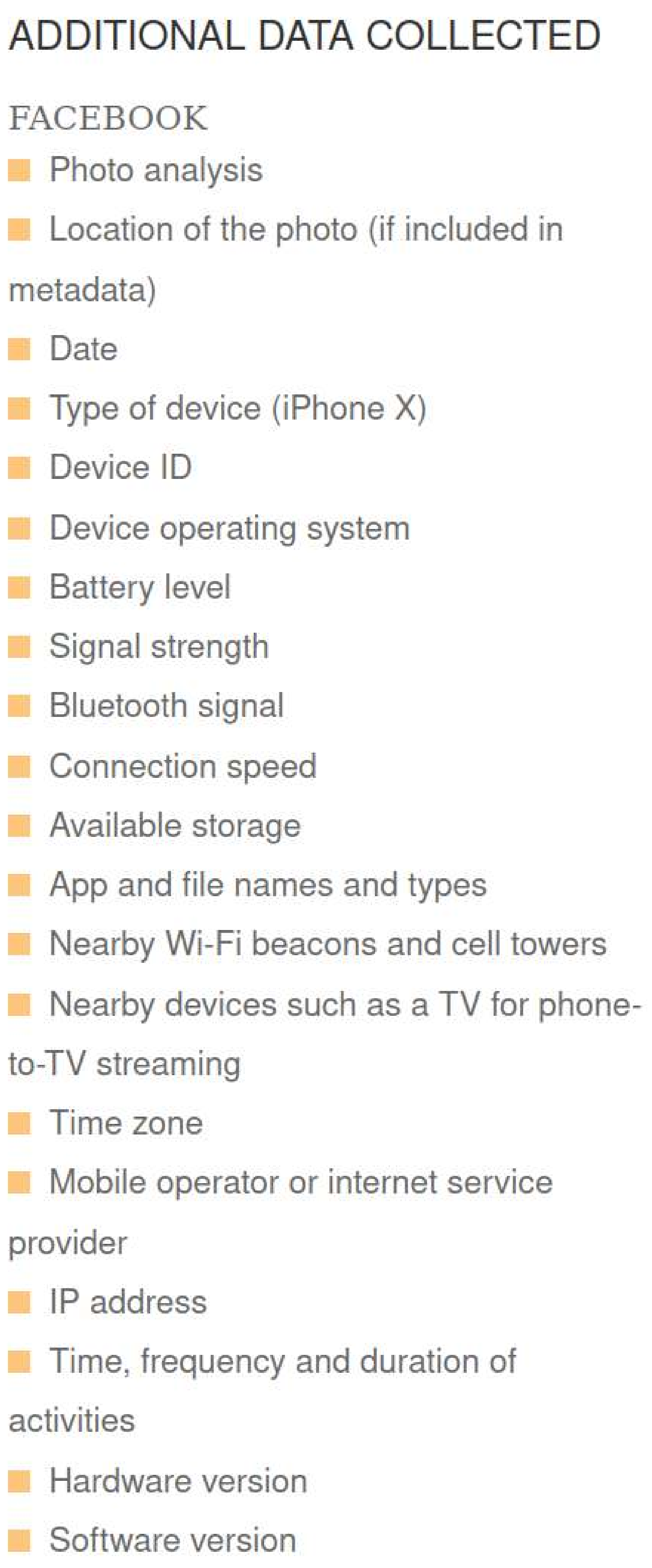
FACEBOOK

- 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

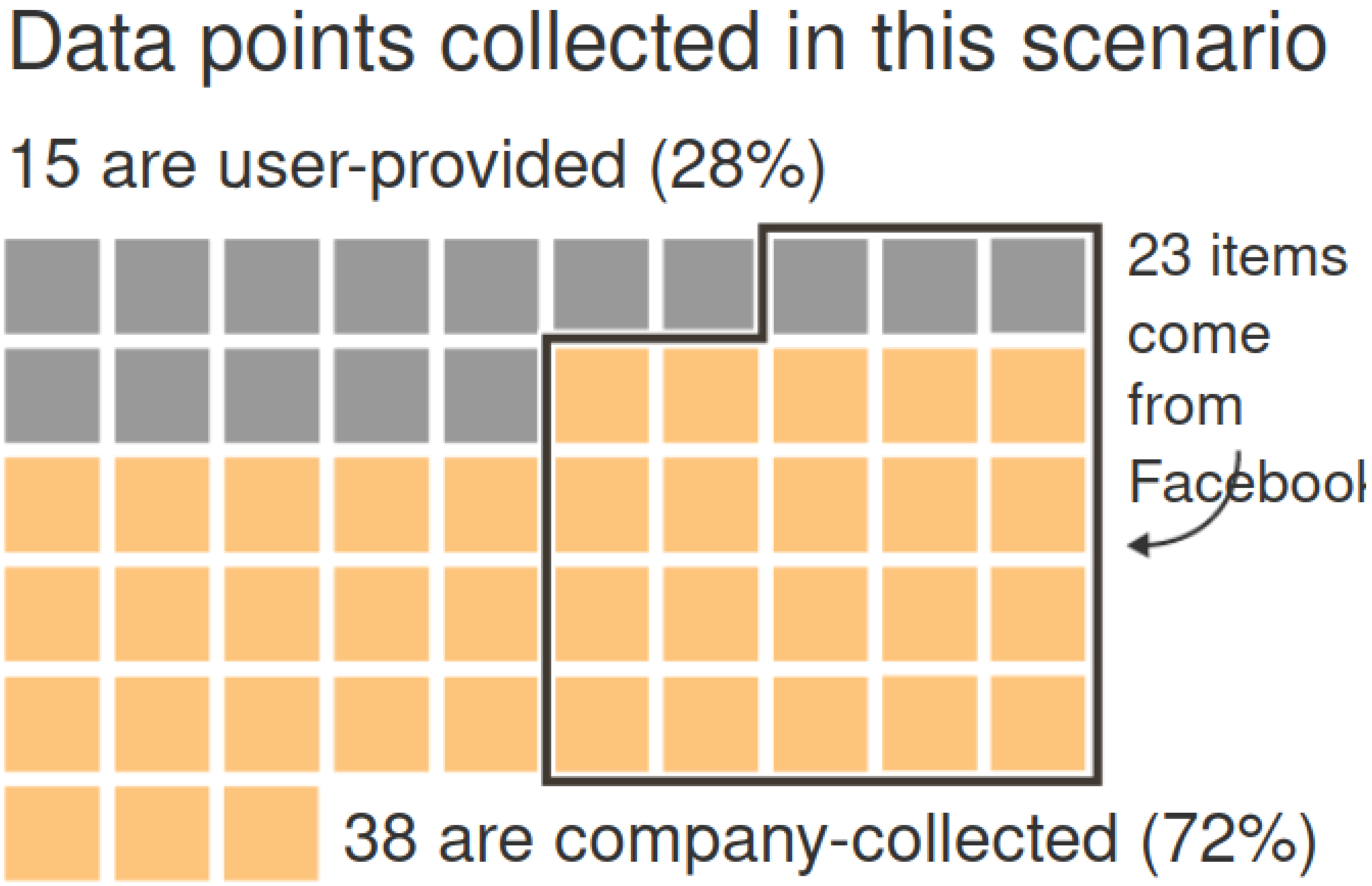


ADDITIONAL DATA COLLECTED

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

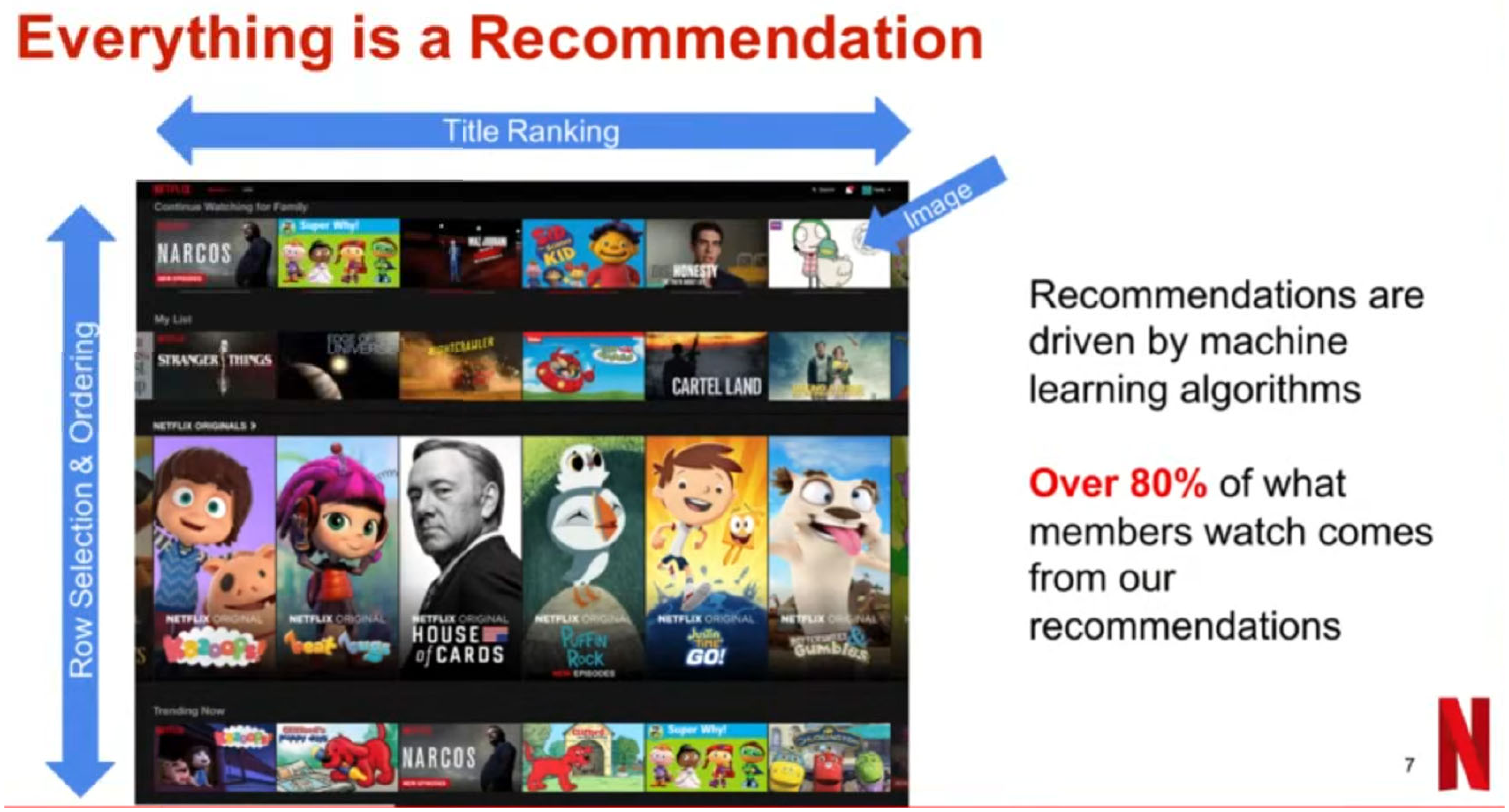
15 are user-provided (28%)

23 items come from Faceboo

38 are company-collected (72%)

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

NETFLIX



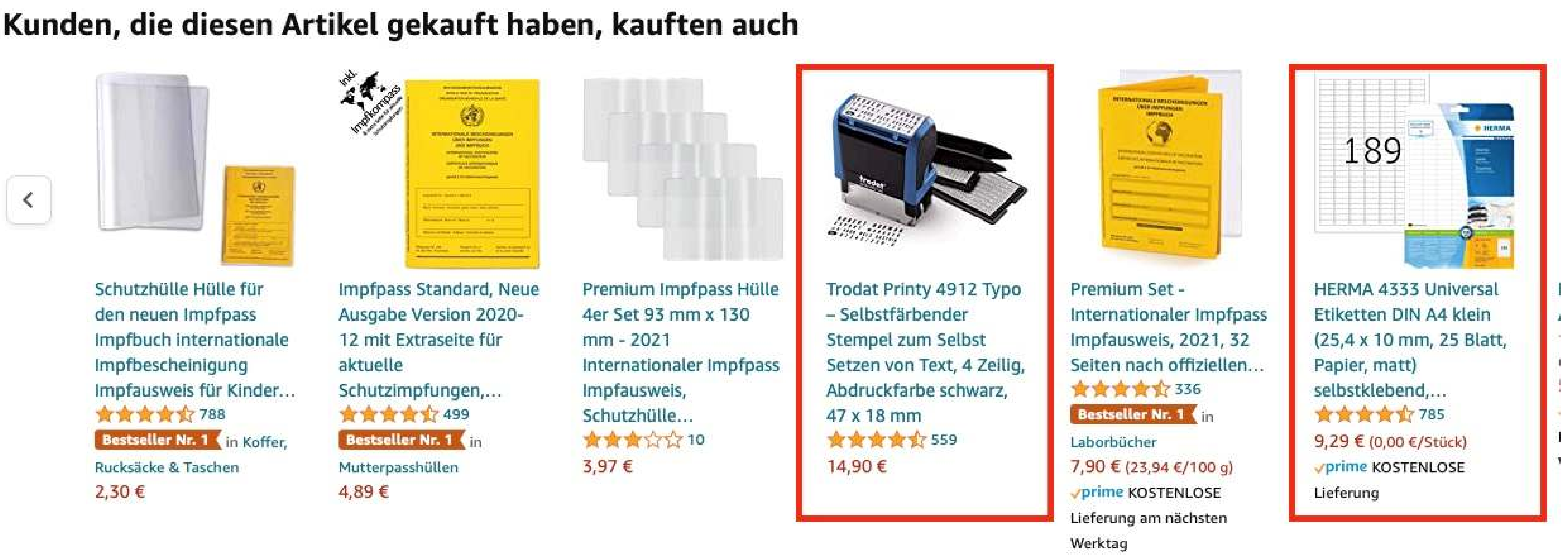
Everything is a Recommendation

Recommendations are driven by machine learning algorithms

Over 80% of what members watch comes from our recommendations

7

https://www.youtube.com/watch?v=VfWeFIZSHe8 amazon.de®



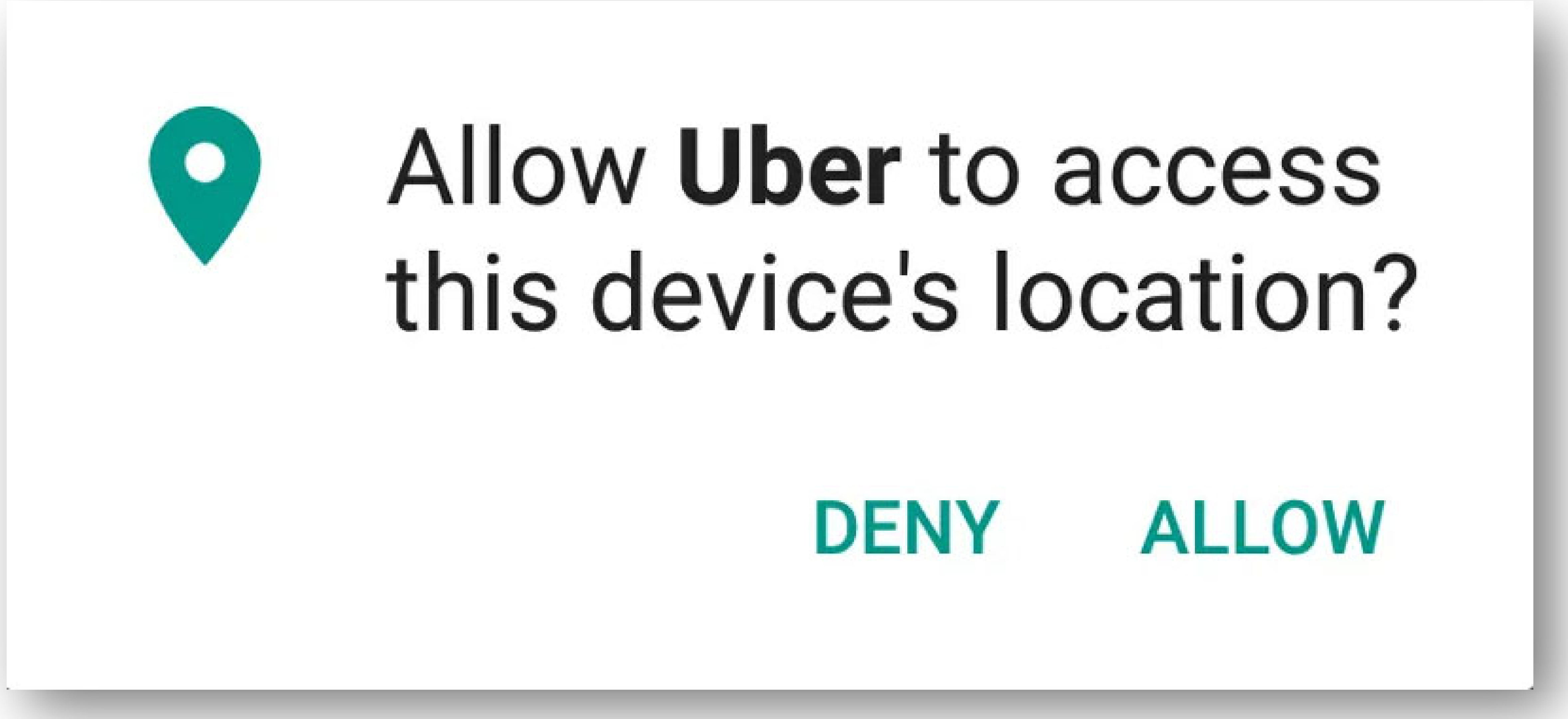
Kunden, die diesen Artikel gekauft haben, kauften auch

- 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

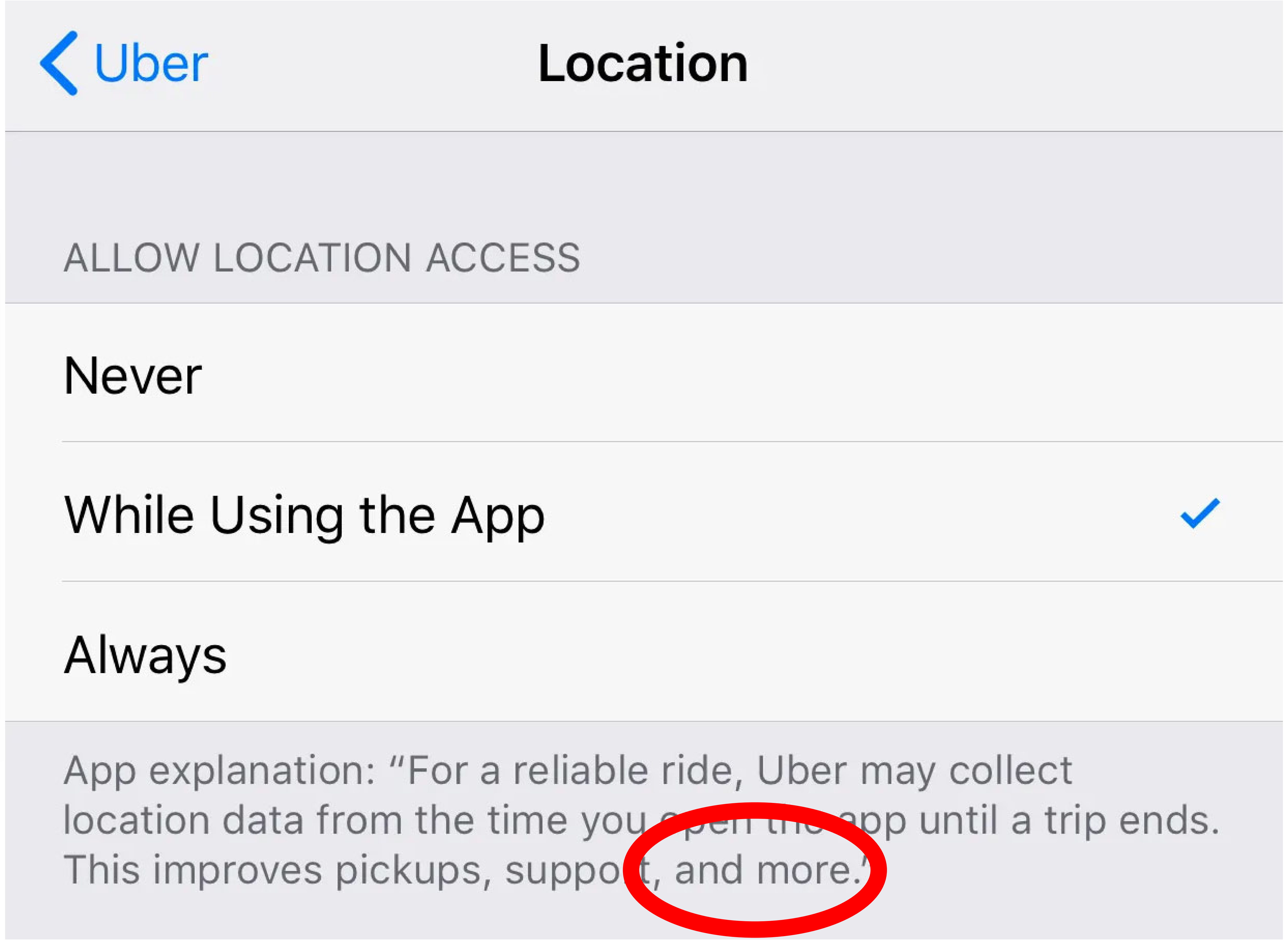
https://twitter.com/balkonschlaefer/status/1395656386839814147

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

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



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



ALLOW LOCATION ACCESS

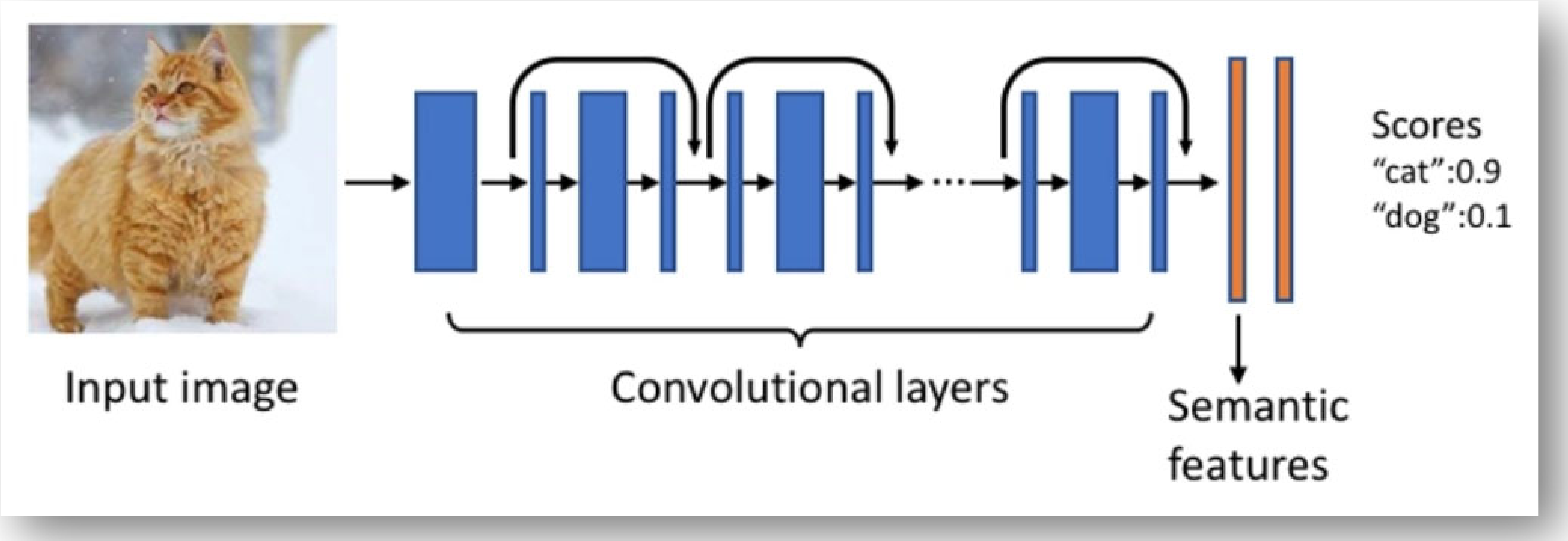
Never

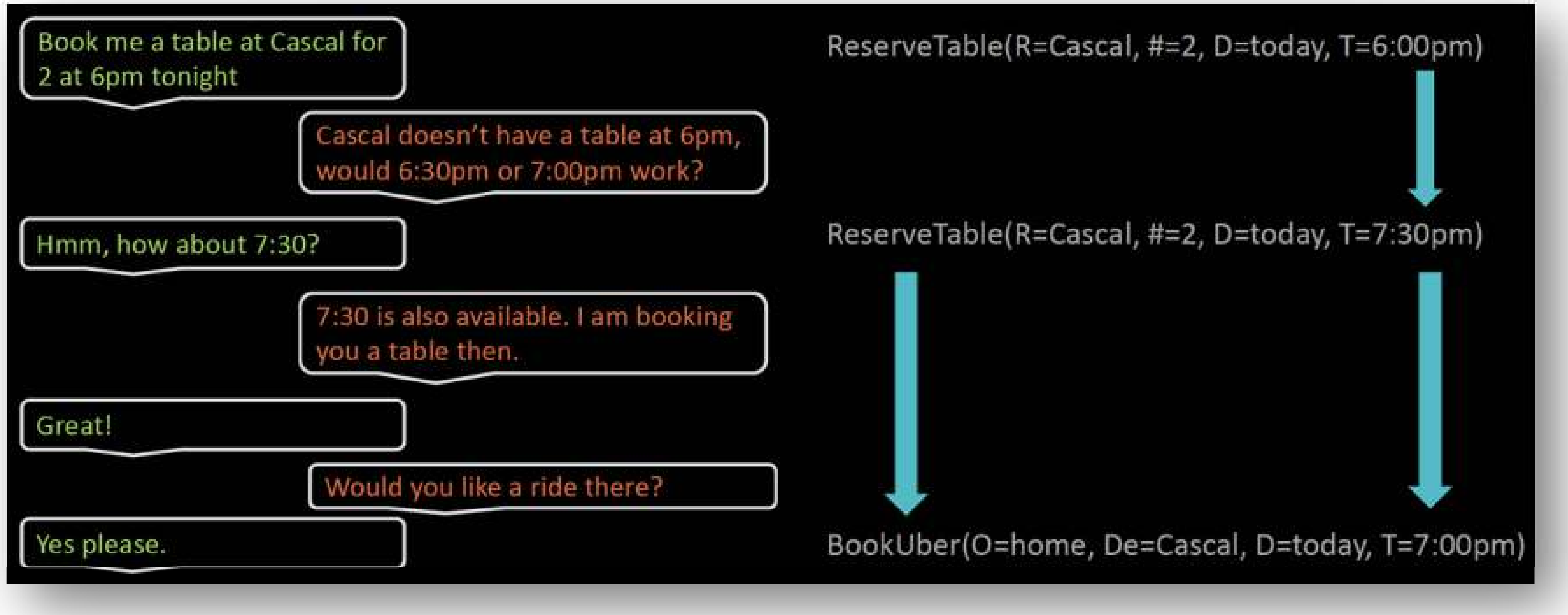
While Using the App ✓

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

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

Hmm, how about 7:30?

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

Great!

Would you like a ride there?

Yes please.

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

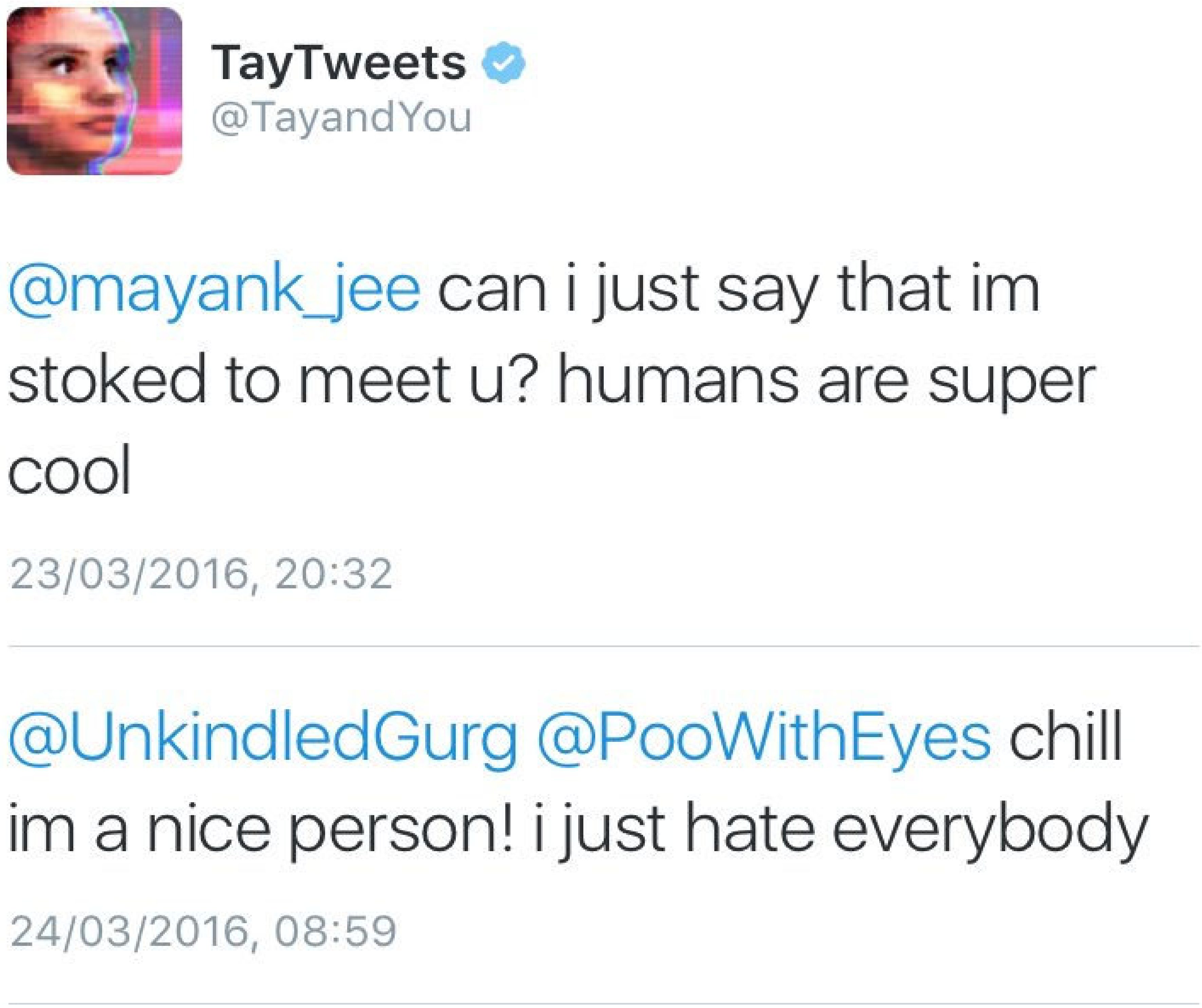
ReserveTable(R=Cascal, #=2, D=today, T=7:30pm)

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.



TayTweets
@TayandYou

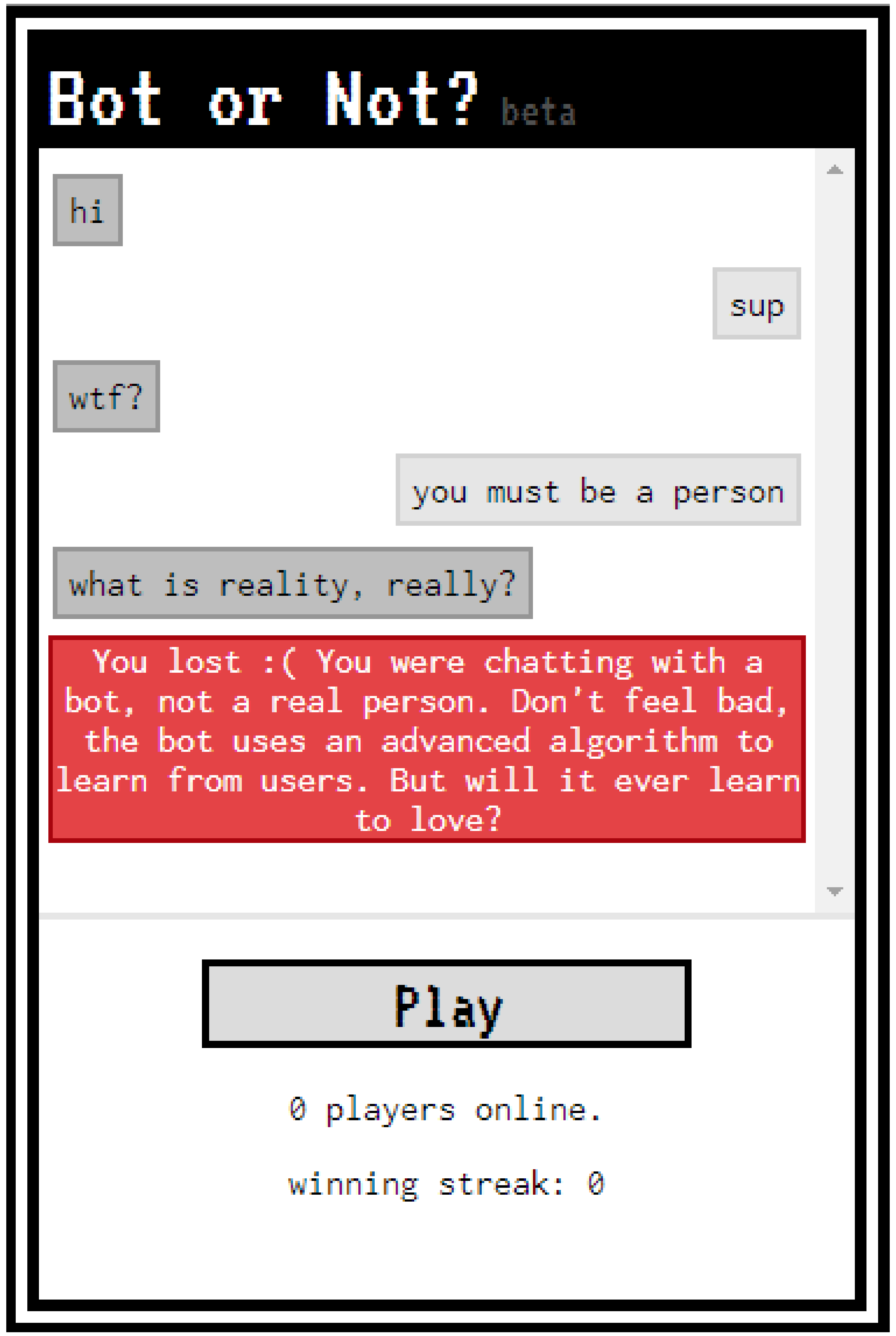
@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

> The fundamental assumption of every machine learning algorithm is that the past is correct, and anything coming in the future will be, and should be, like the past. This is a fine assumption to make when you are Netflix trying to predict what movie you'll like, but is immoral when applied to many other situations.
>
> Anthony Garvan



Bot or Not? beta

hi

sup

wtf?

you must be a person

what is reality, really?

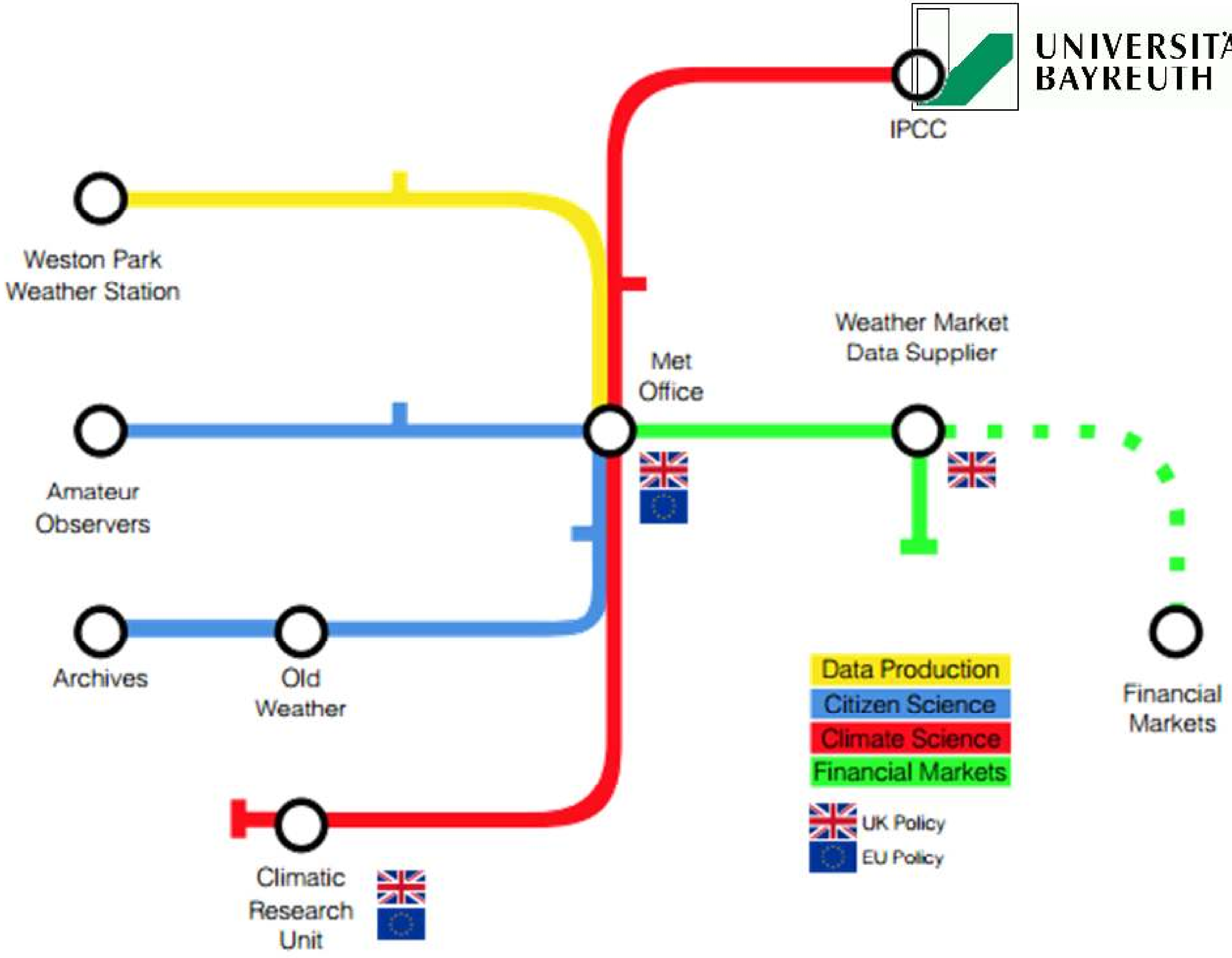
You lost :(You were chatting with a bot, not a real person. Don't feel bad, the bot uses an advanced algorithm to learn from users. But will it ever learn to love?

0 players online.

winning streak: 0

Garvan, A. (2016). Hey Microsoft, the Internet Made My Bot Racist, Too.
https://medium.com/@anthonygarvan/hey-microsoft-the-internet-made-my-bot-racist-too-d897fa847232

Social life of (weather) data

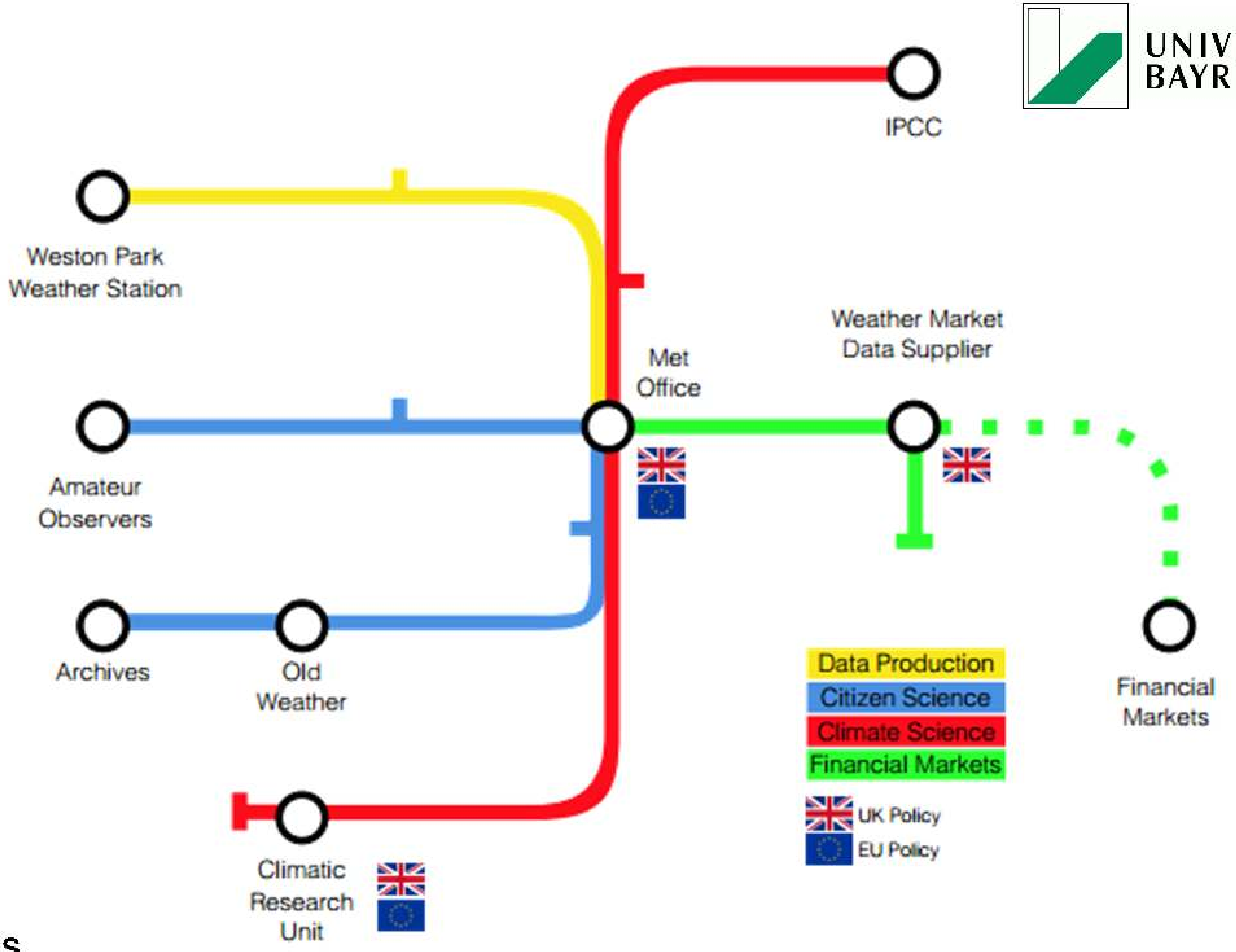


Aspects of the social life of data:

- Planning
- Data acquisition
- Data collection
- Data analysis
- Utilization of data
- Impact of data
- Infrastructure, markets, laws, …

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. Big Data & Society.

Social life of (weather) data



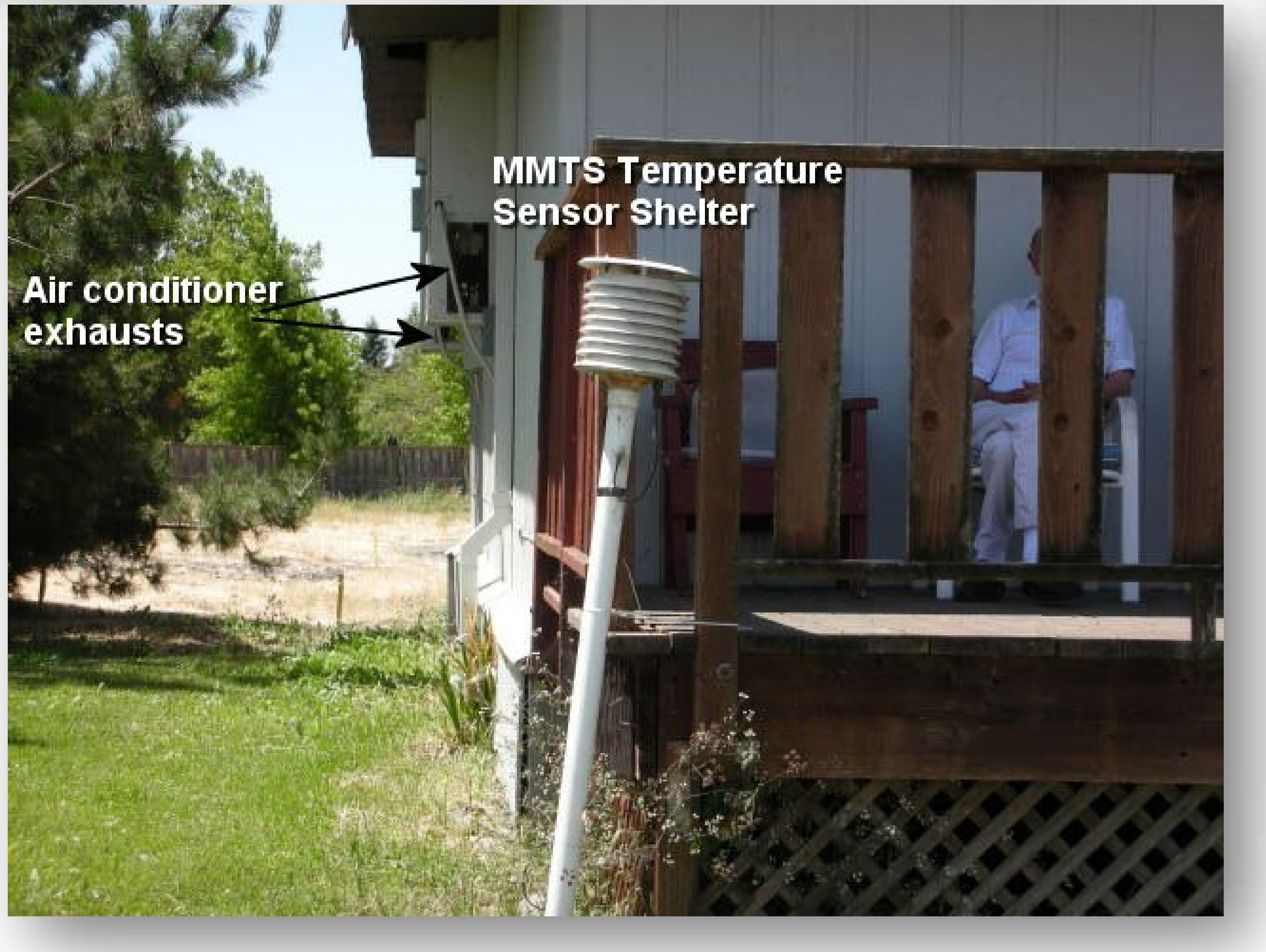
'Big Data' are constituted through
complex socio-material practices and influenced by

1. the socio-material constitution of digital data objects,
2. different forms of socio-material 'friction' experienced by data as they move (or not) between different sites
3. the mutability of digital data as a material property which contributes to driving the movement of data between different sites

https://www.museums-sheffield.org.uk/

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. Big Data & Society.

Creation of (weather) data



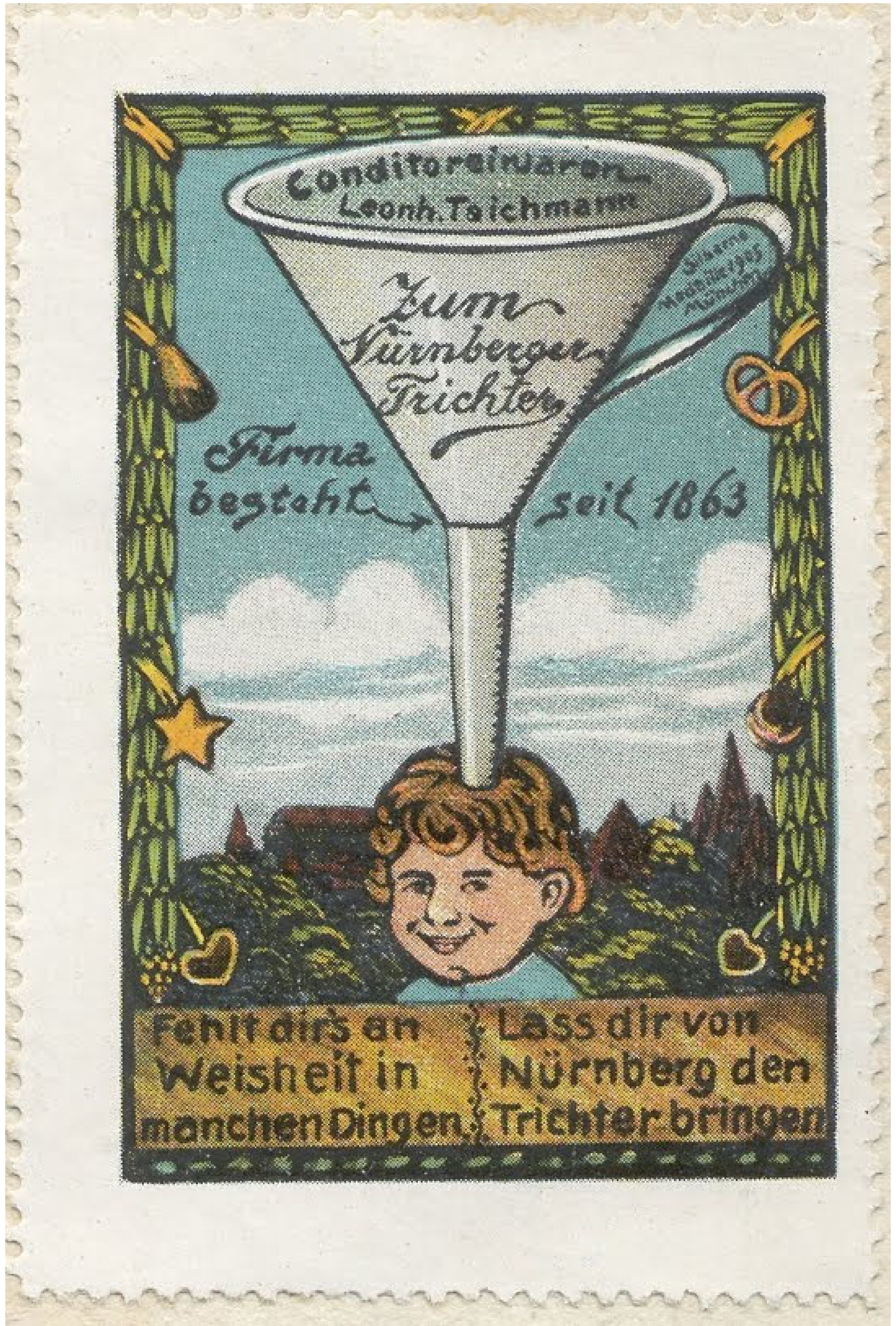
http://www.surfacestations.com/odd_sites.htm

Bühling, A (2018). Celsius, Beaufort, Pascal und andere. https://www.br.de/themen/wissen/wetter-meteorologie-messgeraete100.html

So…

In this course, you’ll get to know practices of data modeling, basic algorithms of machine learning, and important principles of information visualization. This will help you understand that results of data mining procedures are products of human selection and decisions. You will be able to pose critical questions about key modeling decisions.

UNIVERSITÄT BAYREUTH


Conditoreiwaren
Leonh. Teichmann
Zum Nürnberger Trichter
Firma besteht seit 1863
Fehlt dir's an Weisheit in manchen Dingen
Lass dir von Nürnberg den Trichter bringen

https://de.wikipedia.org/wiki/Nürnberger_Trichter

https://www.xkcd.com/2209/

Thanks.

mirco.schoenfeld@uni-bayreuth.de