

From data flow analysis to data security, privacy, responsible AI, and more

Shujun LI (李树钧)

Director, [Institute of Cyber Security for Society \(iCSS\)](#)

Professor of Cyber Security, School of Computing
University of Kent, UK

<http://www.hooklee.com/>

 @hooklee

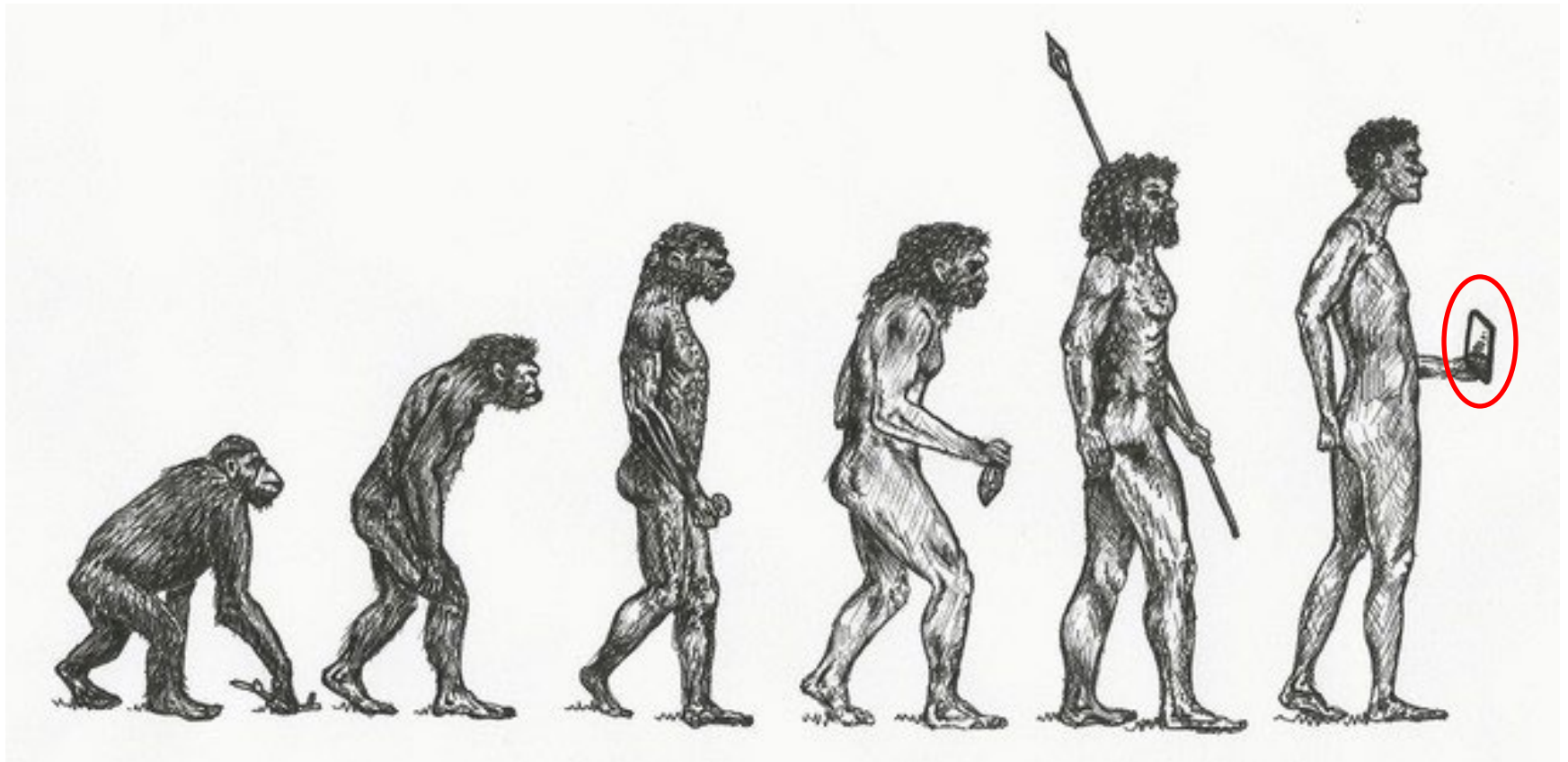
From data flow analysis to ...

The world and the age we are living in



The evolution of human

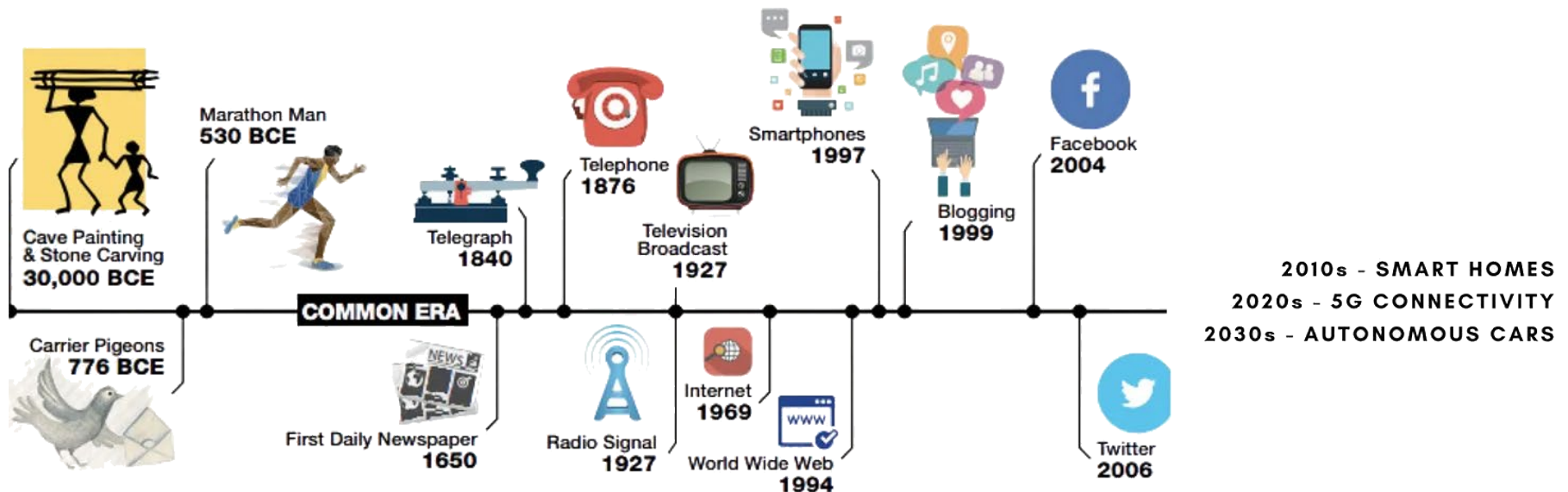
- From apes to modern humans



Source: <https://www.linkedin.com/pulse/timeline-human-prehistory-manjunath-r/>

The evolution of communication

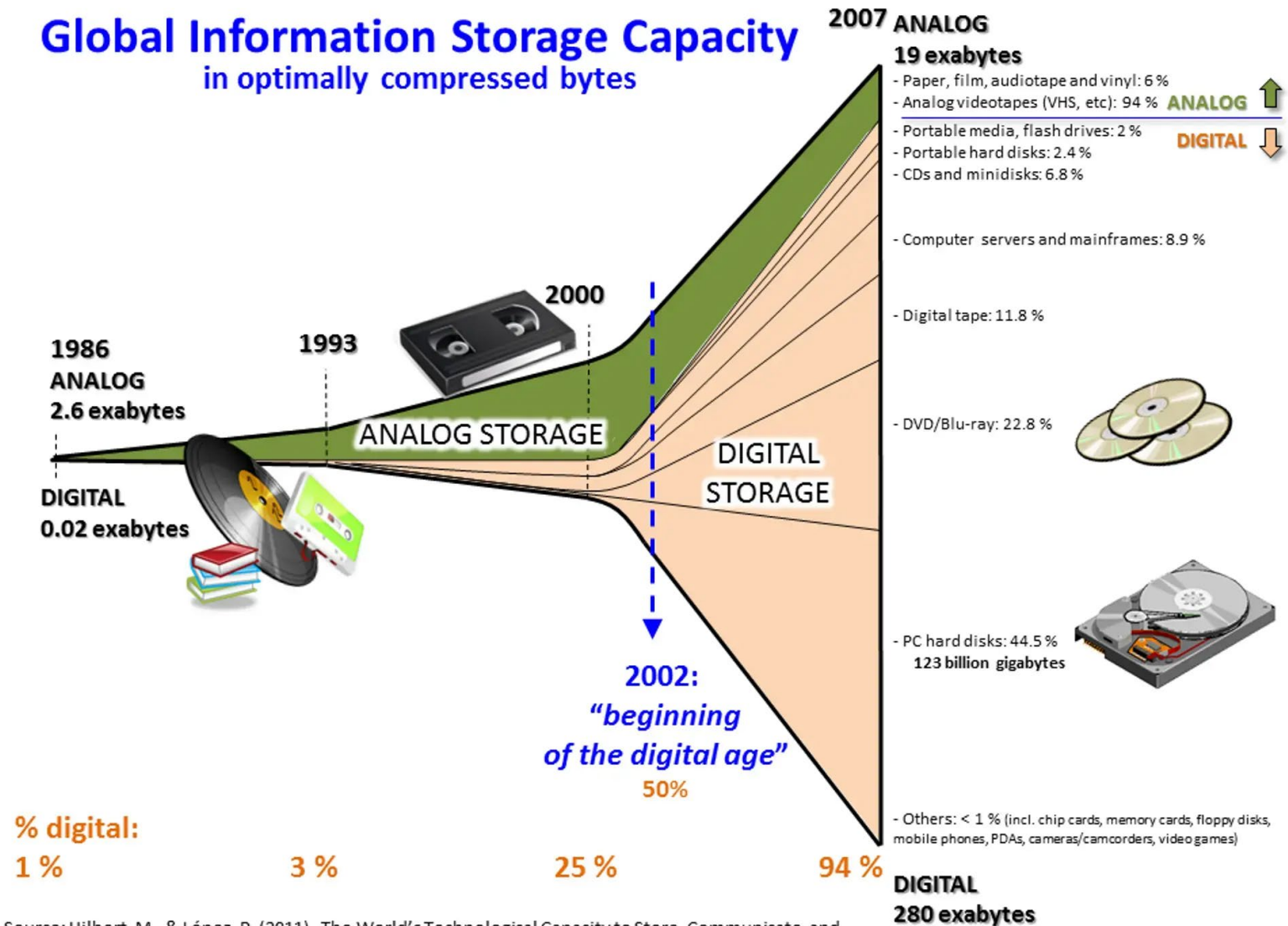
- From analogue to digital



Source: <https://medium.com/@thisisvibhuti/navigating-nuances-in-virtual-conversations-digital-body-language-263c8150f301>

The evolution of data

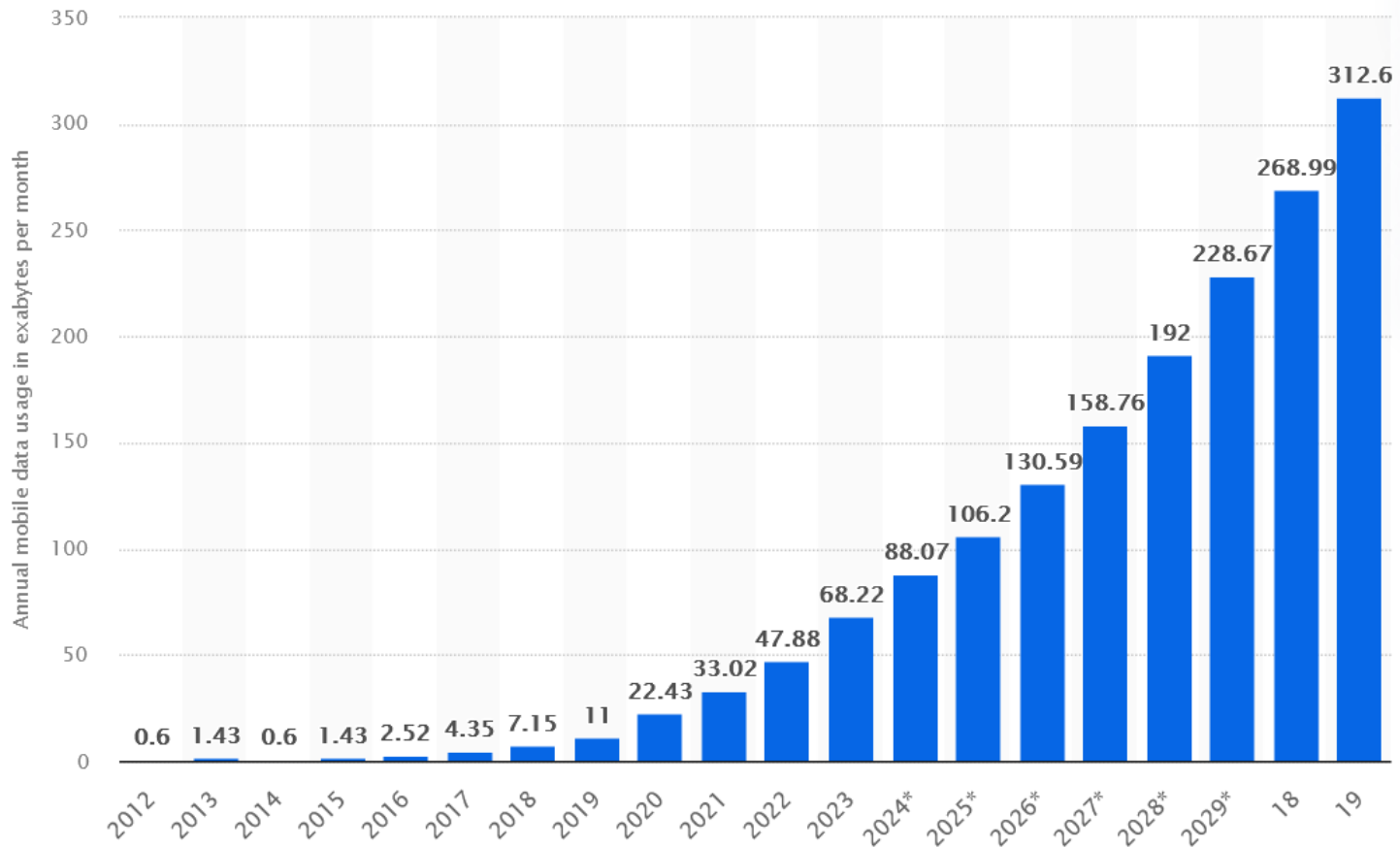
Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

The data economy era

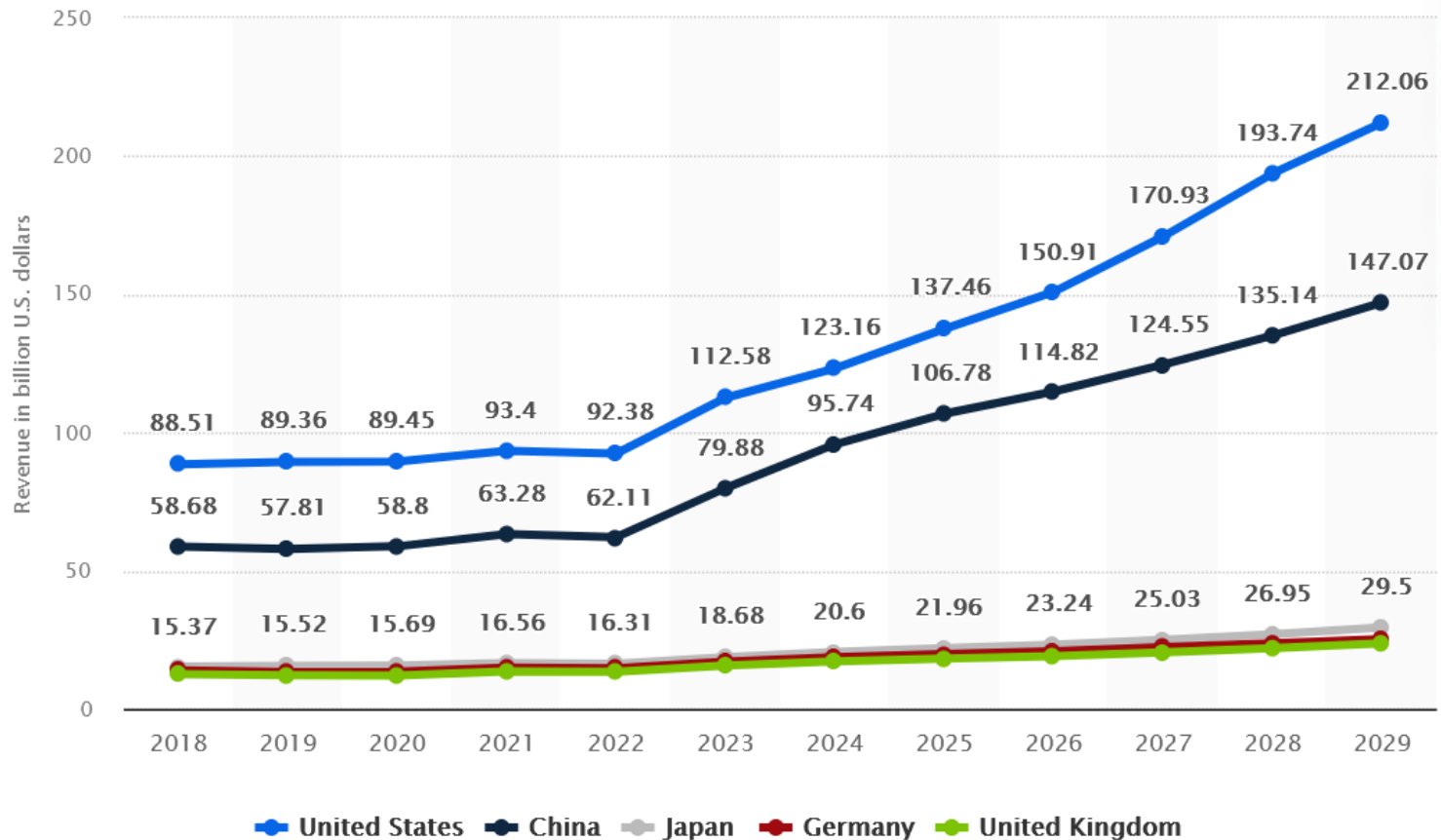
- Annual mobile data traffic worldwide 2012-29



Source: <https://www.statista.com/statistics/630107/annual-mobile-data-usage-vodafone-worldwide/>

The data economy era

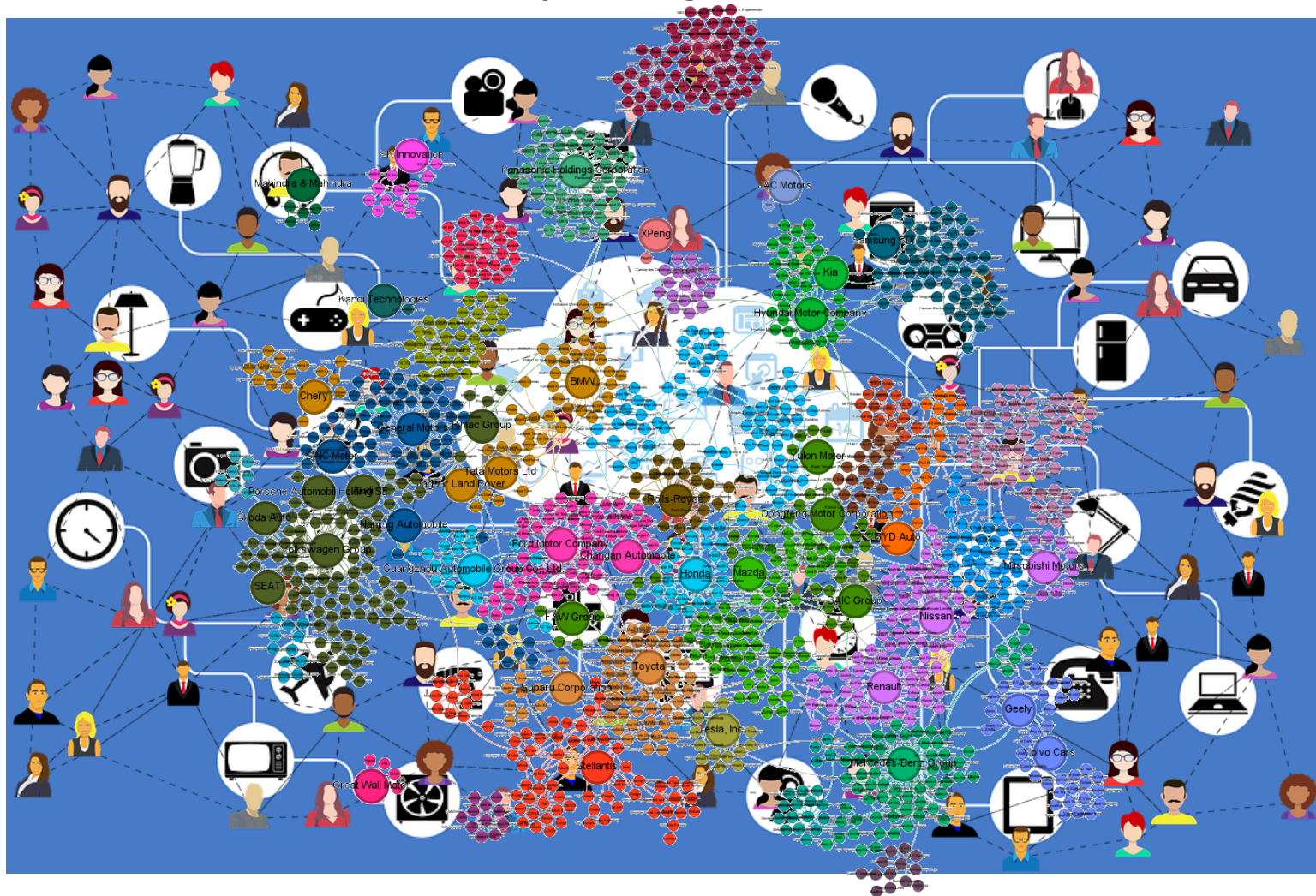
- Revenue of leading data centre markets 2018-29



Source: <https://www.statista.com/statistics/1370199/leading-data-center-markets-globally/>

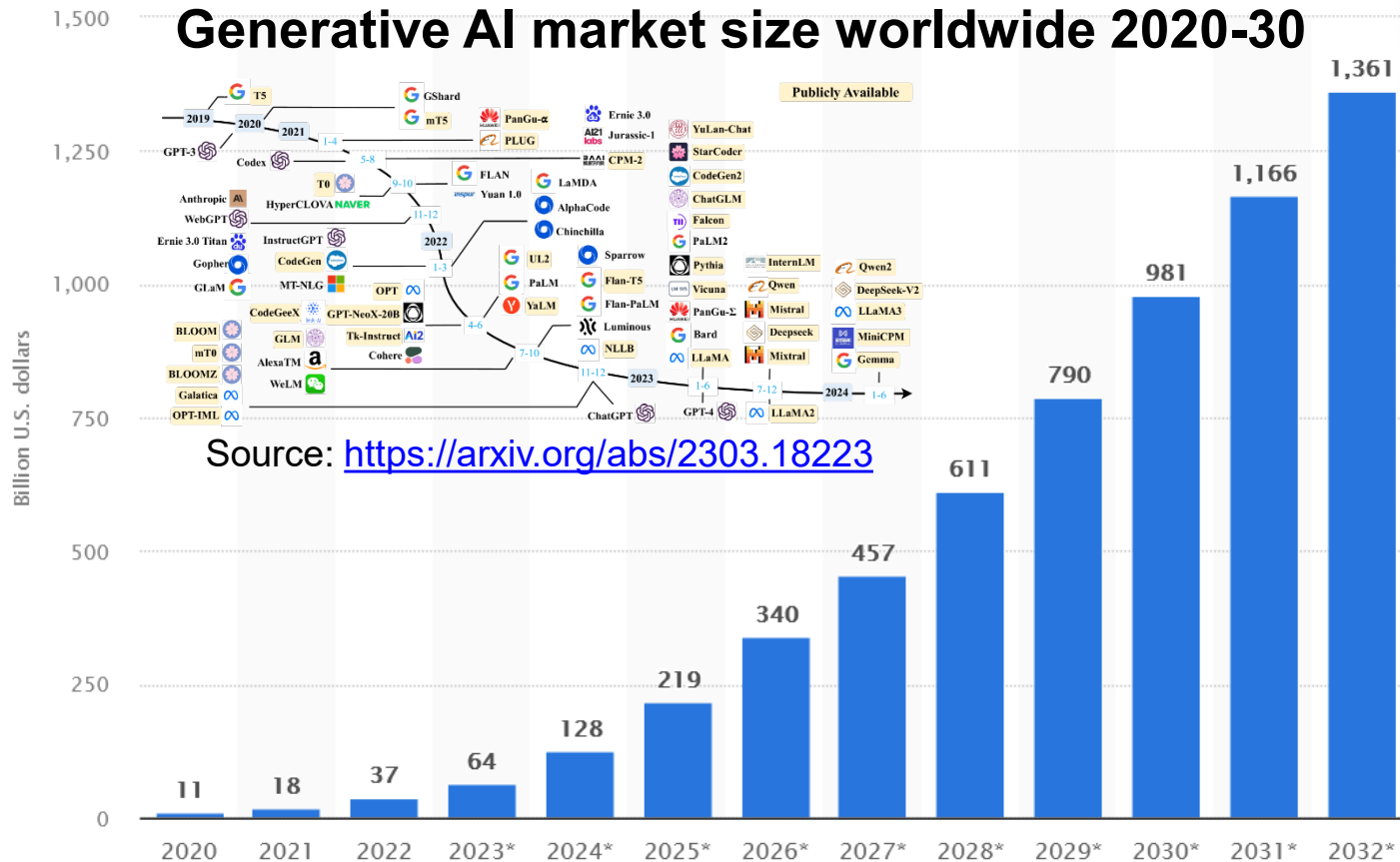
Who are the data generators?

- The Internet of Everything!



Who are the data generators?

- And the generative AI!



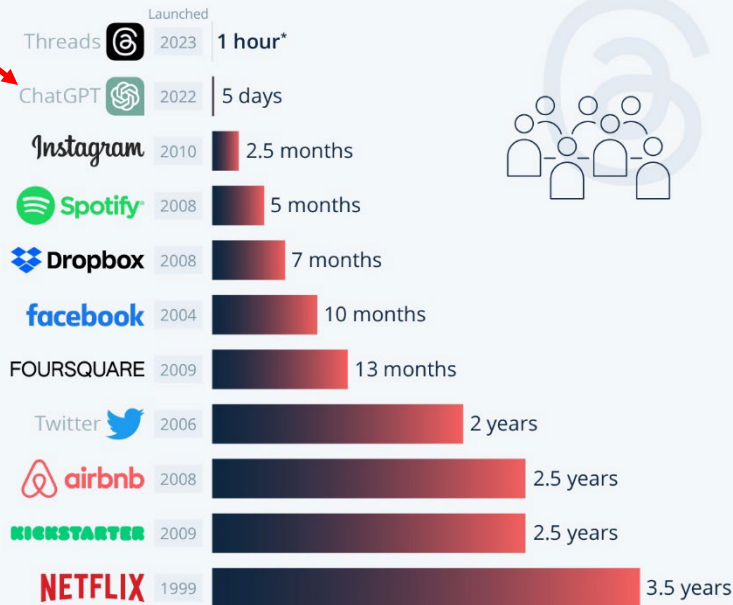
Source: <https://www.statista.com/statistics/1417151/generative-ai-revenue-worldwide/>

Who are the data generators?

- How fast has generative AI become popular?

Threads Shoots Past One Million User Mark at Lightning Speed

Time it took for selected online services to reach one million users



Refers to one million backers (Kickstarter), nights booked (Airbnb), downloads (Instagram/Foursquare)

* Two million signups in two hours

Source: Company announcements via Business Insider/LinkedIn

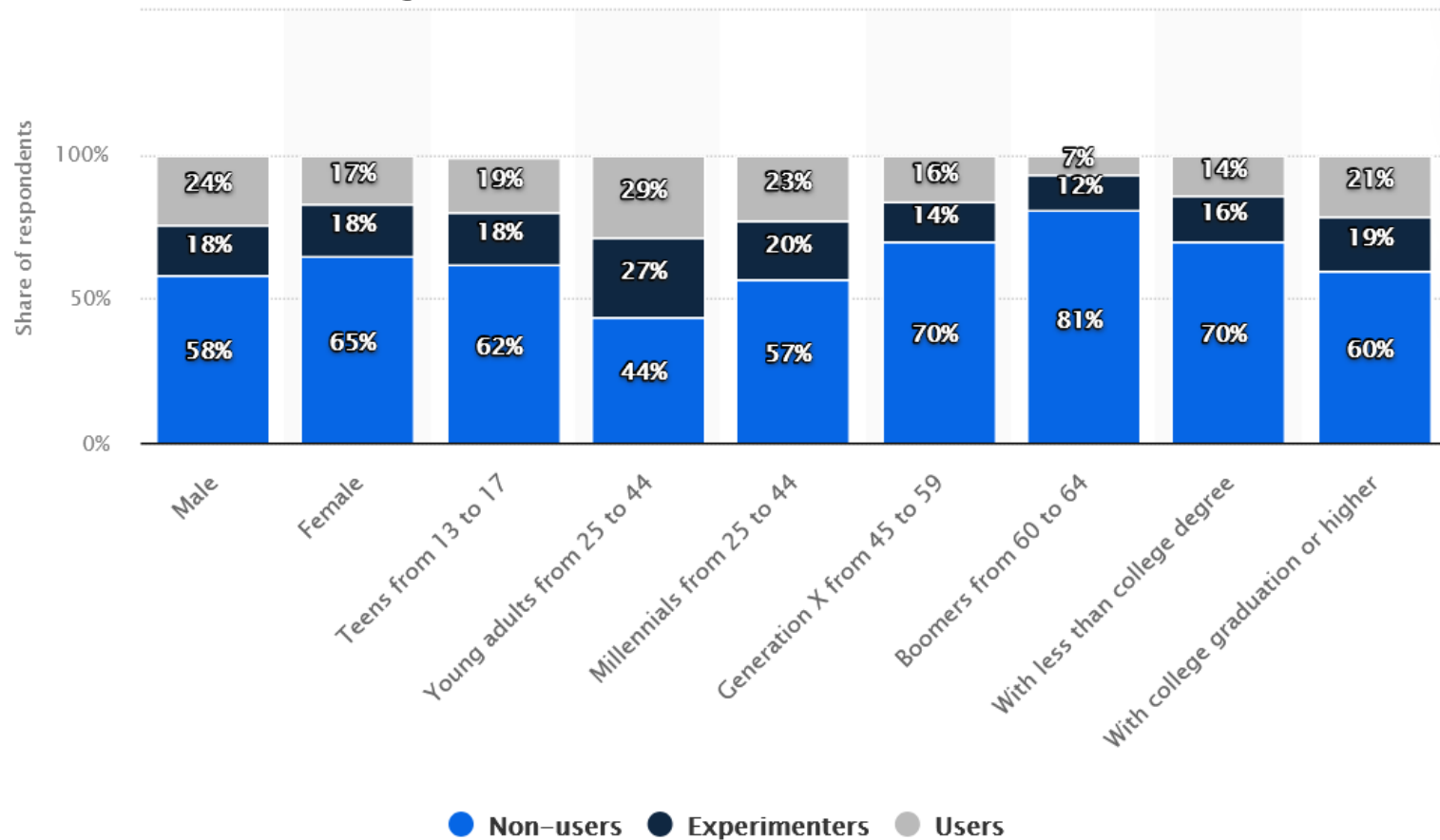


Source:

<https://www.statista.com/chart/29174/time-to-one-million-users/>

Who are the data generators?

- Global use of generative AI in 2023



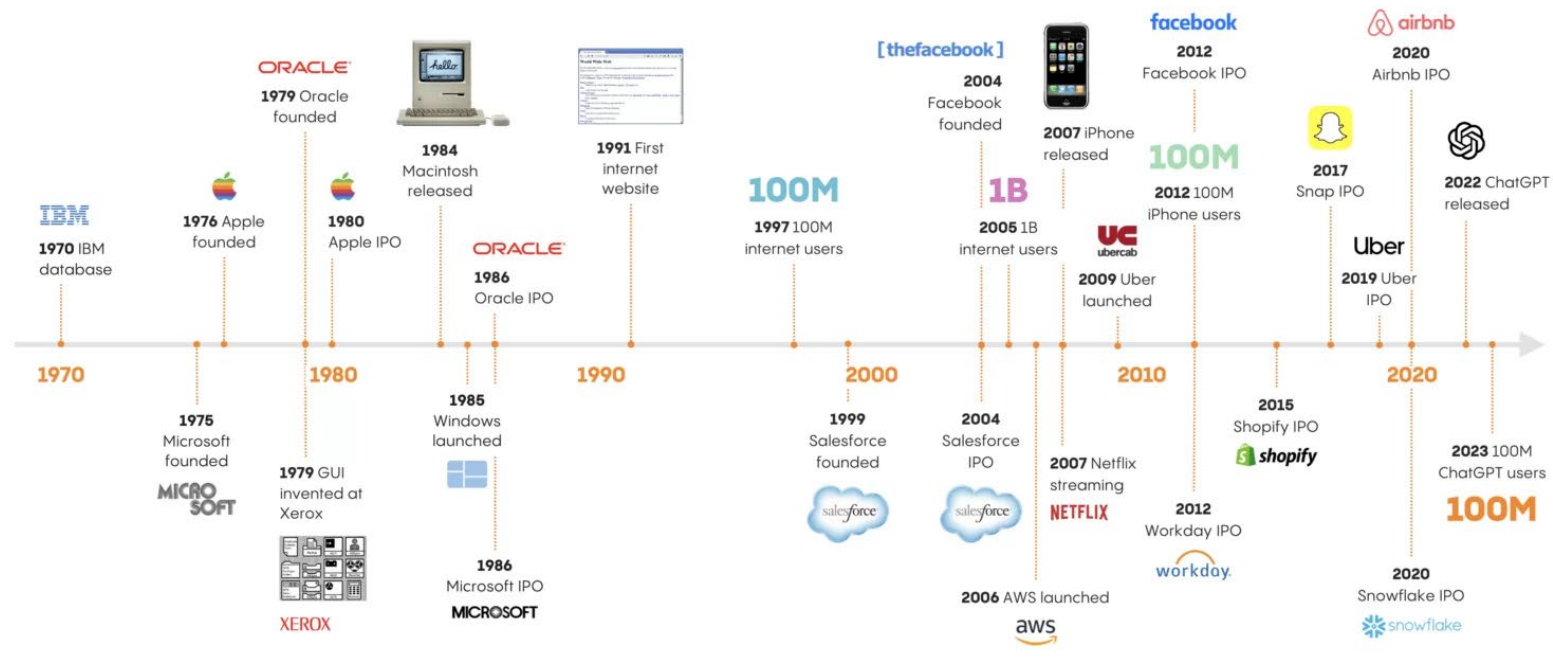
Source: <https://www.statista.com/statistics/1455933/generative-ai-use-worldwide-by-group/>

The evolution of innovation

- From databases (1970s) to generative AI (now)

A Brief History of Innovation

Key inflection points across computing, internet, cloud and mobile

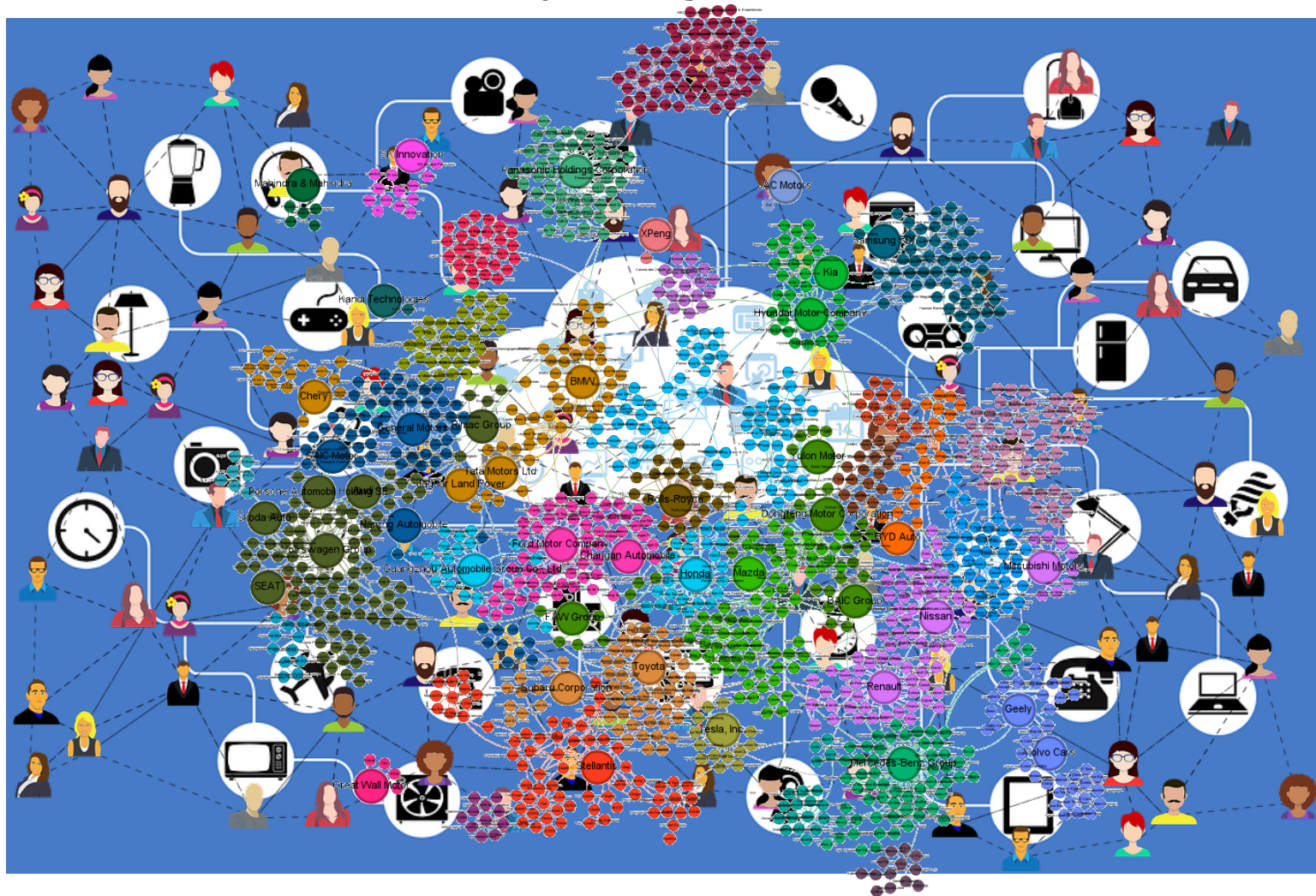


© 2023 Menlo Ventures

Source: <https://menlovc.com/perspective/generative-ai-lessons-from-prior-waves/>

Who are the data consumers?

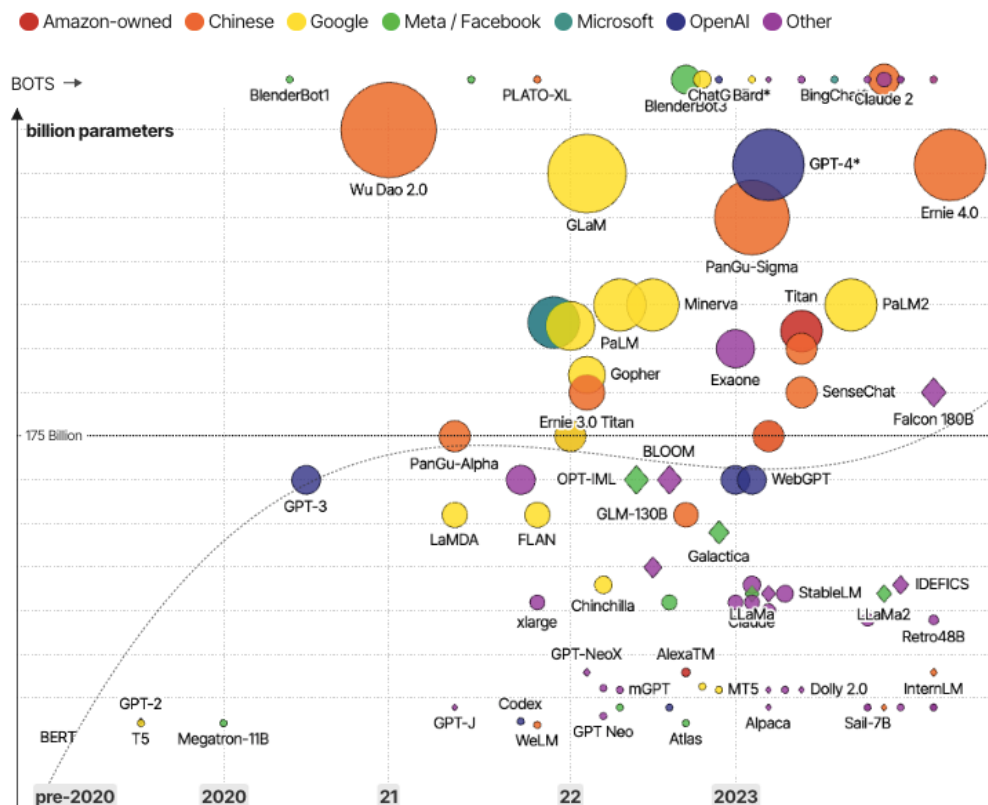
- The Internet of Everything!



Who are the data consumers?

- And AI systems (not just large ones)!

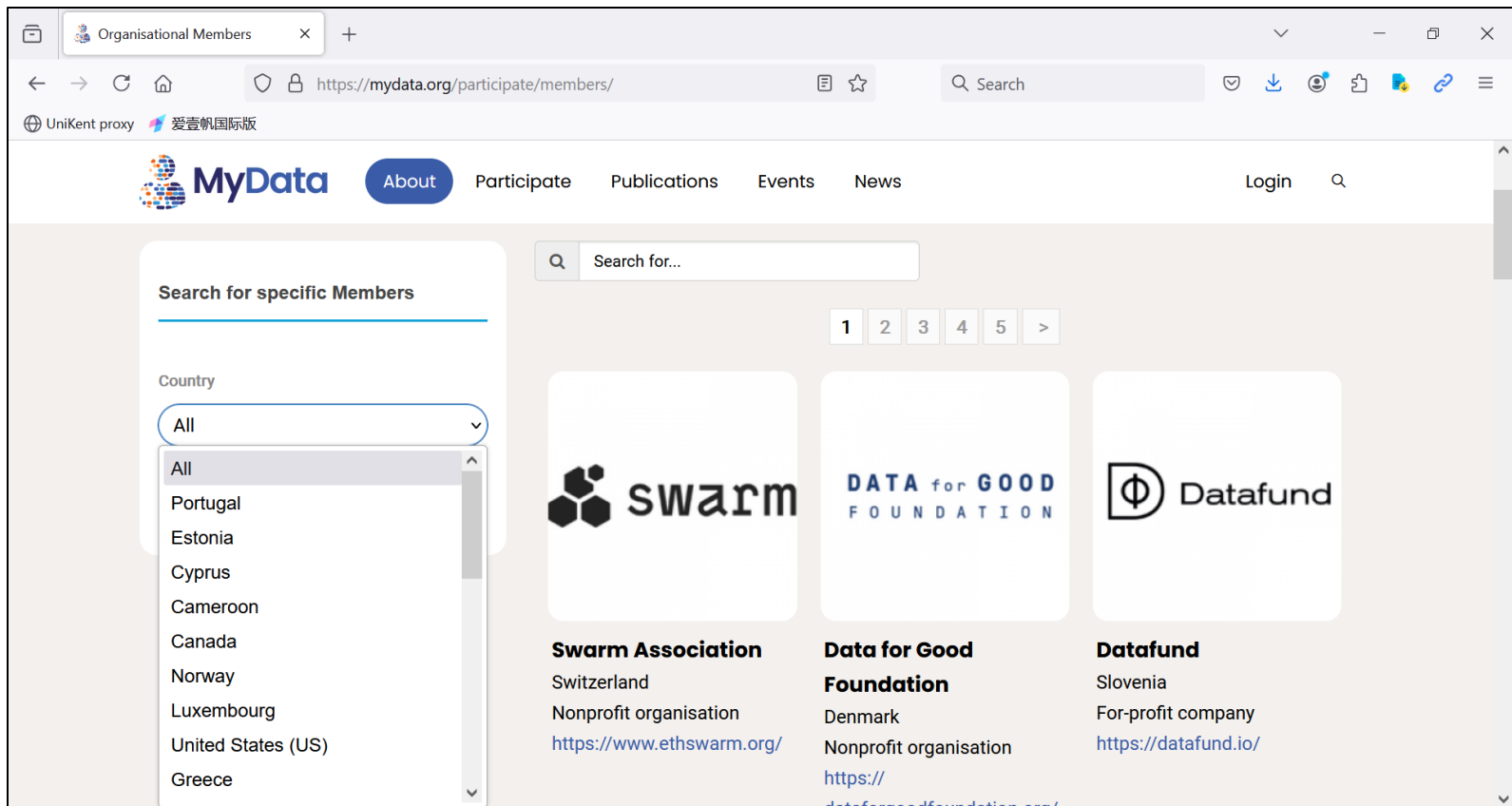
The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT



Source:
<https://www.linkedin.com/pulse/guide-large-language-models-how-work-yugank-aman-g1wif/>

What about data brokers?

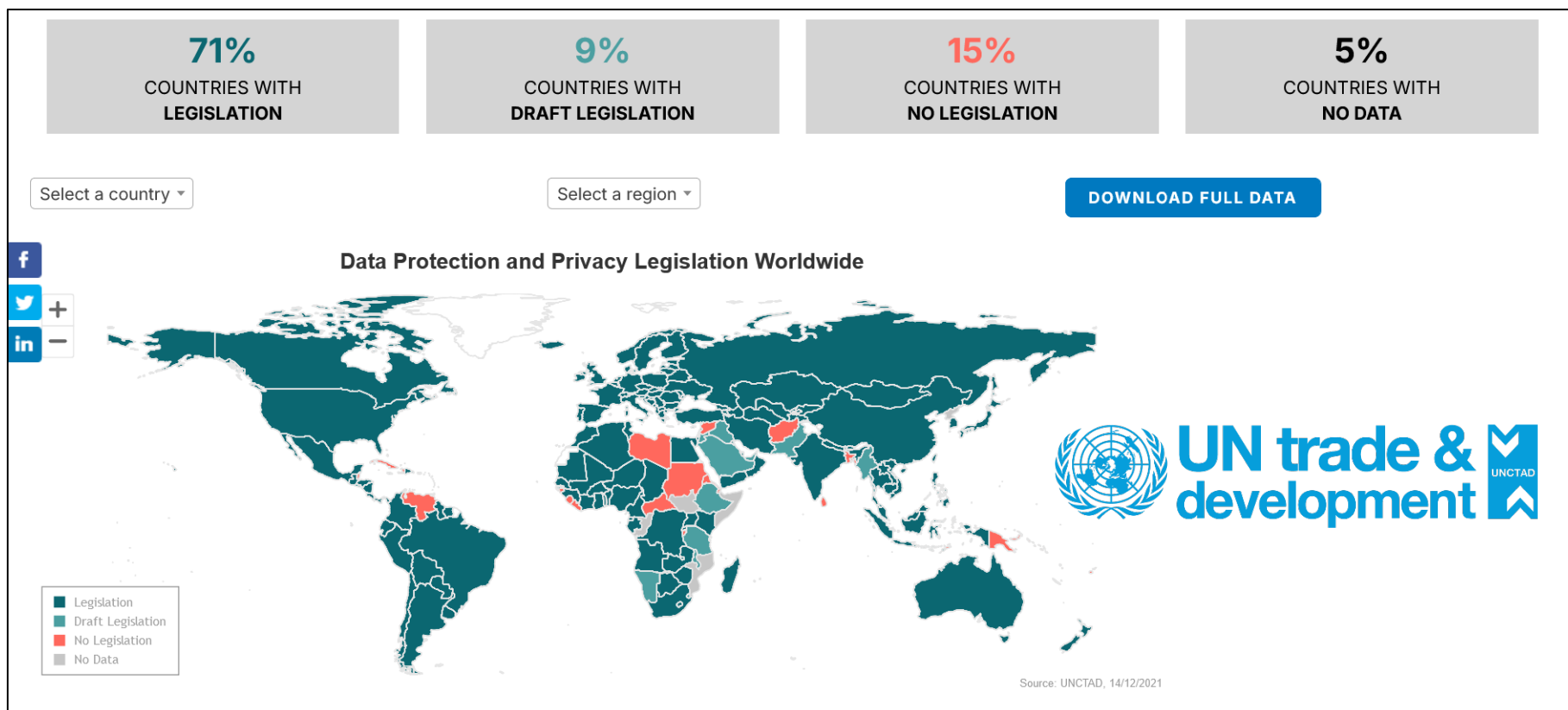
- It is a rapidly growing market!



Source: <https://mydata.org/participate/members/>

How are data regulated?

- More and more data-related laws and ...



Source: <https://unctad.org/page/data-protection-and-privacy-legislation-worldwide>

And data regulators?

- Data-related laws led to national data regulators.

The image shows three overlapping browser windows. The top window displays the European Data Protection Board (EDPB) website at https://www.edpb.europa.eu/about-edpb/about-edpb/members_en. The middle window shows the Information Commissioner's Office (ICO) website at <https://ico.org.uk>. The bottom window shows the Wikipedia article for the National Data Administration (NDA) at https://en.wikipedia.org/wiki/National_Data_Administration. The Wikipedia article includes a search bar, a table of contents, and a main text block stating: "The National Data Administration (NDA) is an administration under the National Development and Reform Commission (NDRC) of the State Council of the People's Republic of China." It also features a photo of the NDA building and an agency overview table.

Agency overview	
Formed	October 25, 2023; 12 months ago

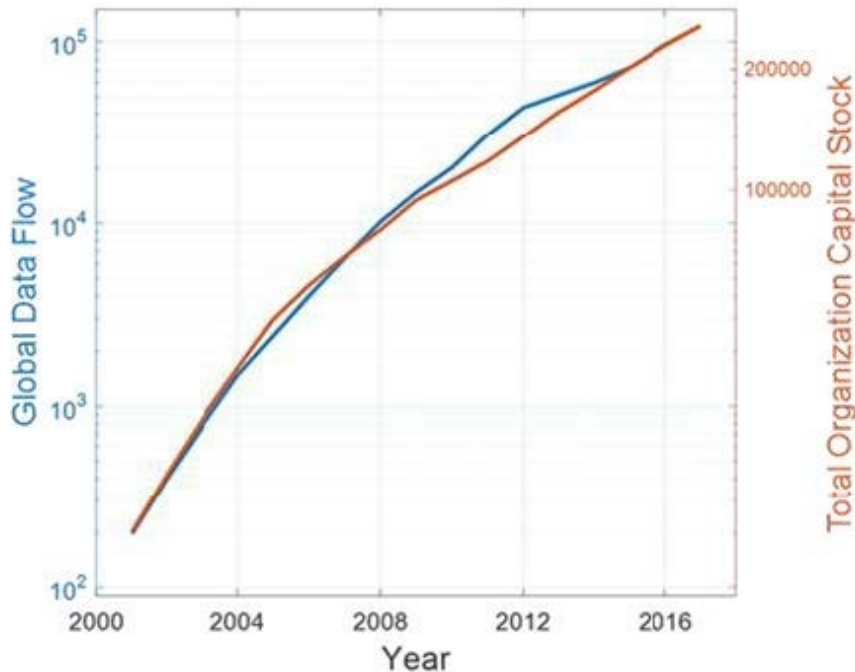
Sources: https://www.edpb.europa.eu/about-edpb/about-edpb/members_en;
<https://ico.org.uk/>; https://en.wikipedia.org/wiki/National_Data_Administration

Data are shared to be useful.



What is data sharing about?

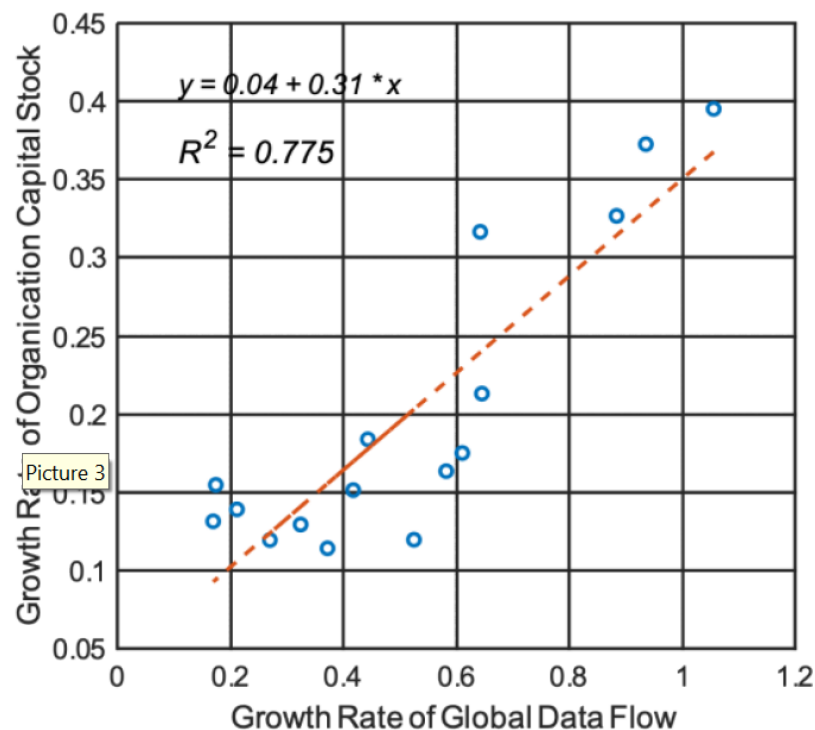
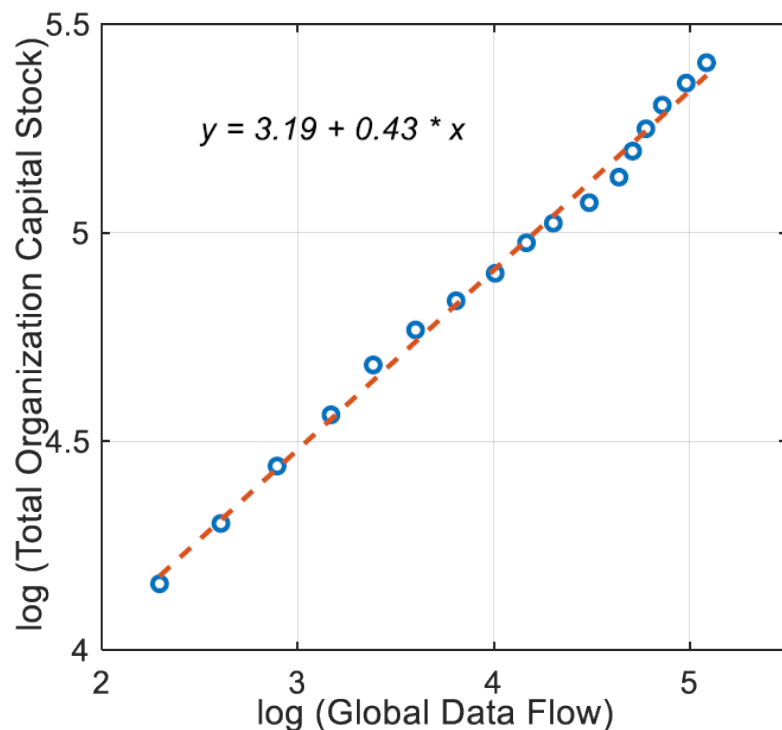
- Sharing = An entity A sends one or more data items to another entity B
- \Rightarrow One or more A-to-B data flows happen!



Source:
https://iariw.org/wp-content/uploads/2021/08/LiChi_paper.pdf

Data sharing makes data economy!

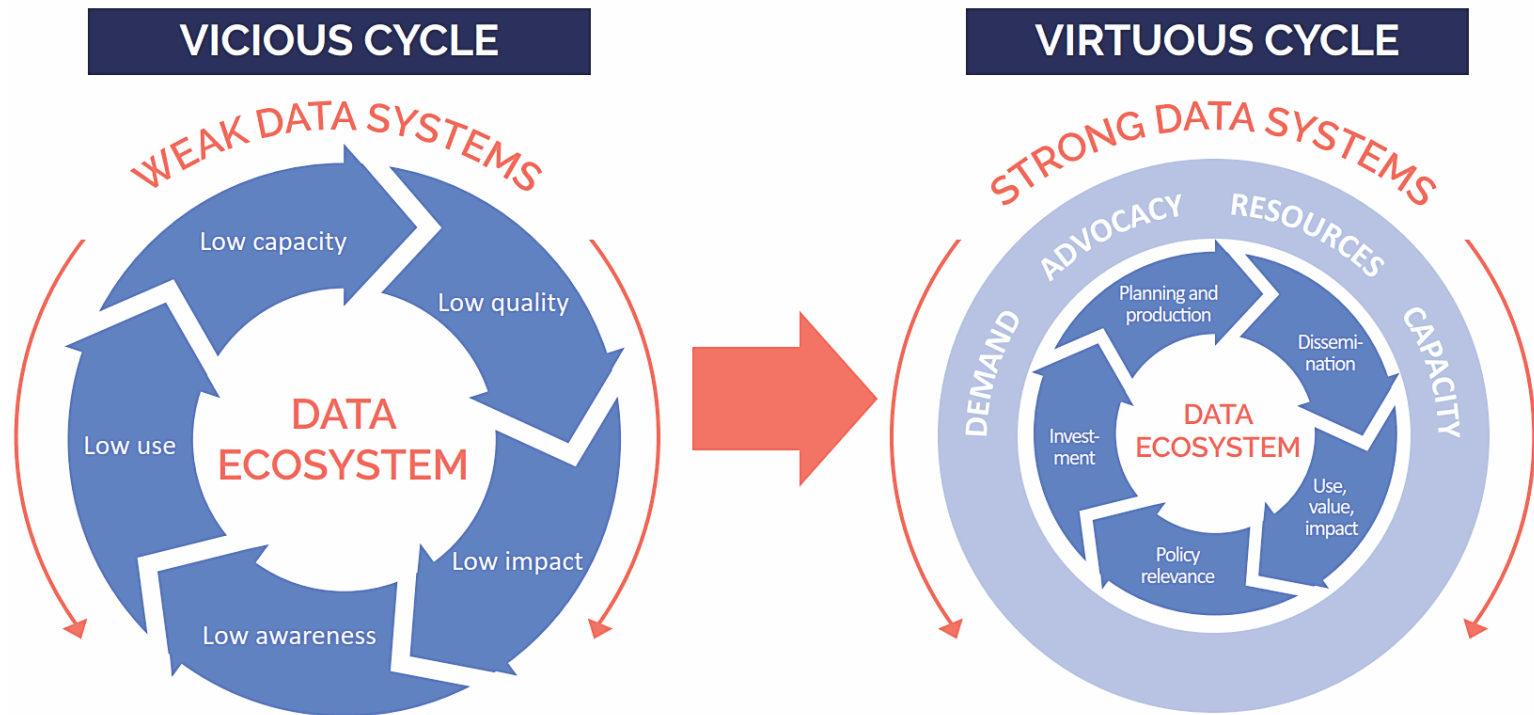
- The more data flows, the more opportunities for commercial growth!



Source: https://iariw.org/wp-content/uploads/2021/08/LiChi_paper.pdf

Data sharing makes data economy!

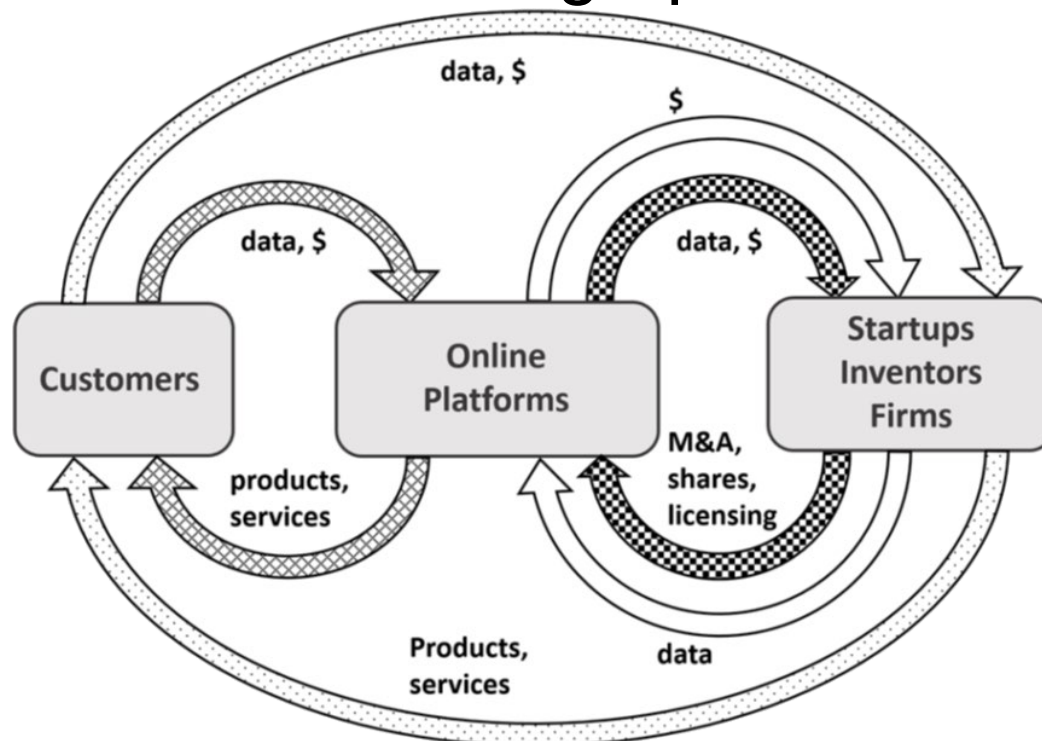
- The more data flows, the more opportunities for commercial growth!



Source: <https://opendatawatch.com/publications/navigating-the-politics-of-open-data/>

From data flow to data flow graphs

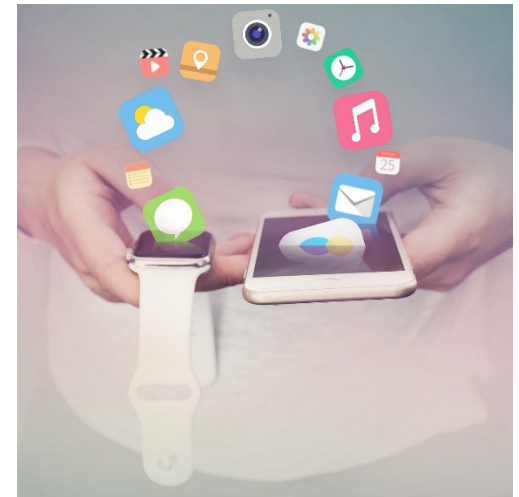
- Data flows within an ecosystem can be quantitatively and/or qualitatively modelled and visualised as a data flow graph.



Source: <https://www.escoe.ac.uk/publications/the-data-economy-market-size-and-global-trade/>

What data?

- Different formats and modalities
 - Texts, images, videos, audio/speech/music, 3D models, hypertext documents, source code, binary data, ...
- Different types
 - Personal and non-personal data
 - Public and private/confidential data
 - Physical and electronic data
 - Real and synthetic data
 - True and false information
 - Human- and machine-generated data
 - Data at rest, data in motion, and data in use
 - ...

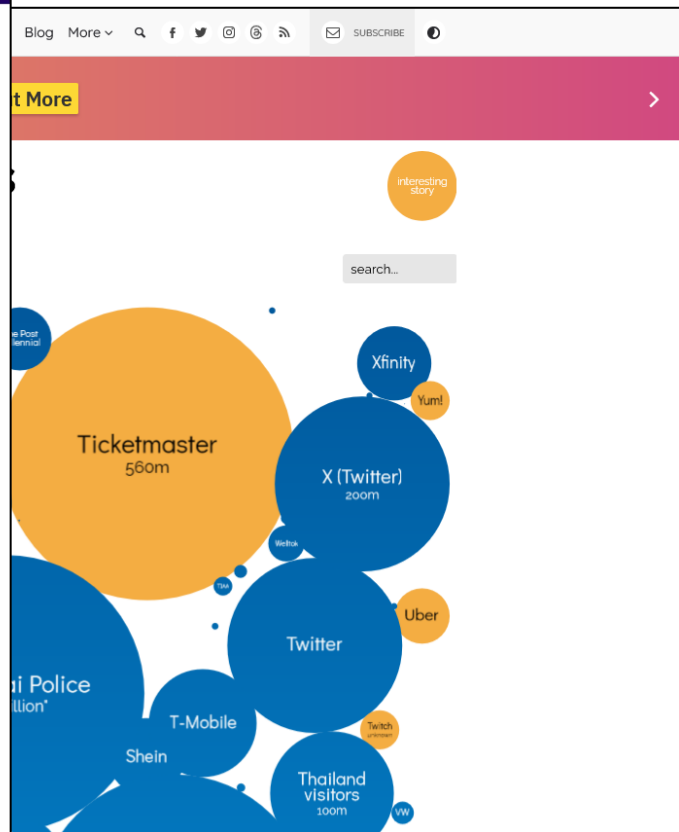
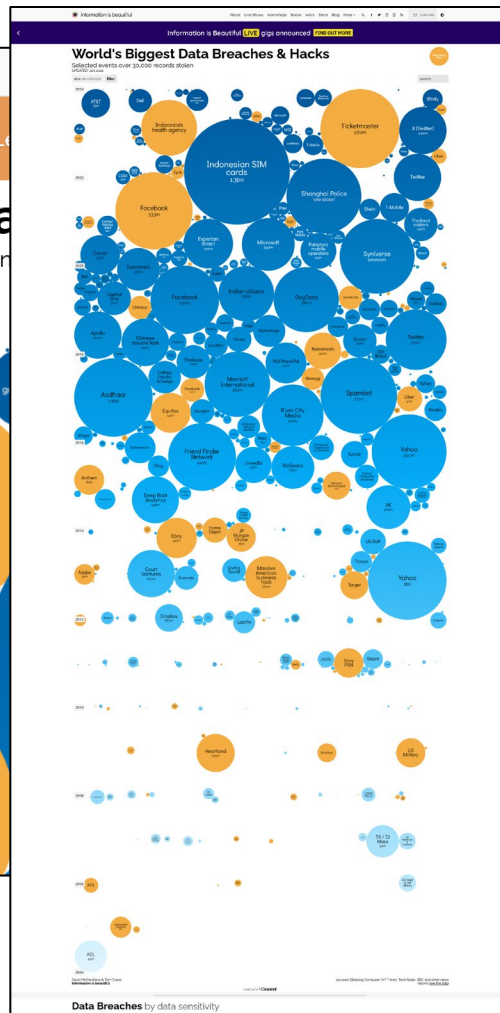
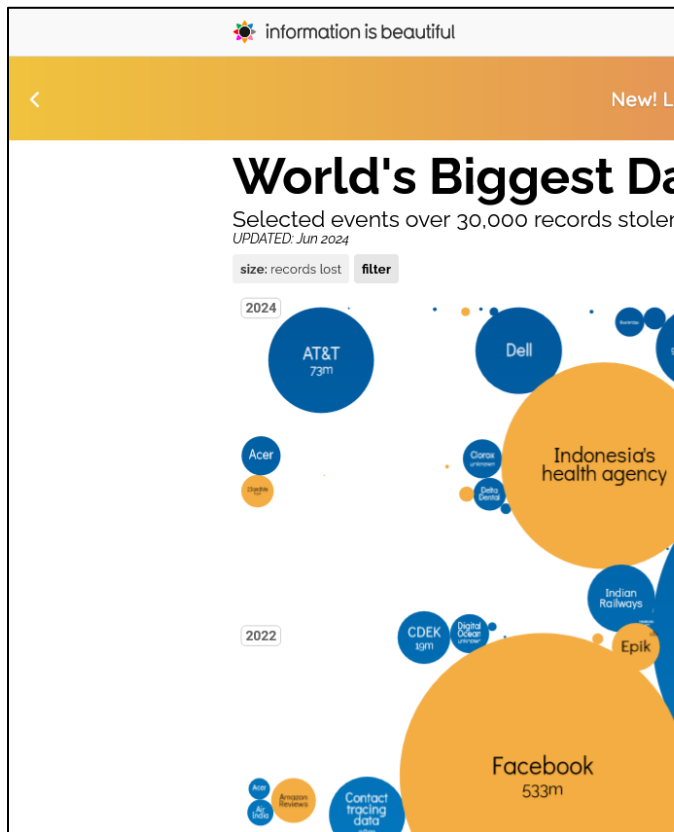


From data flow analysis to ...

Problems about big data



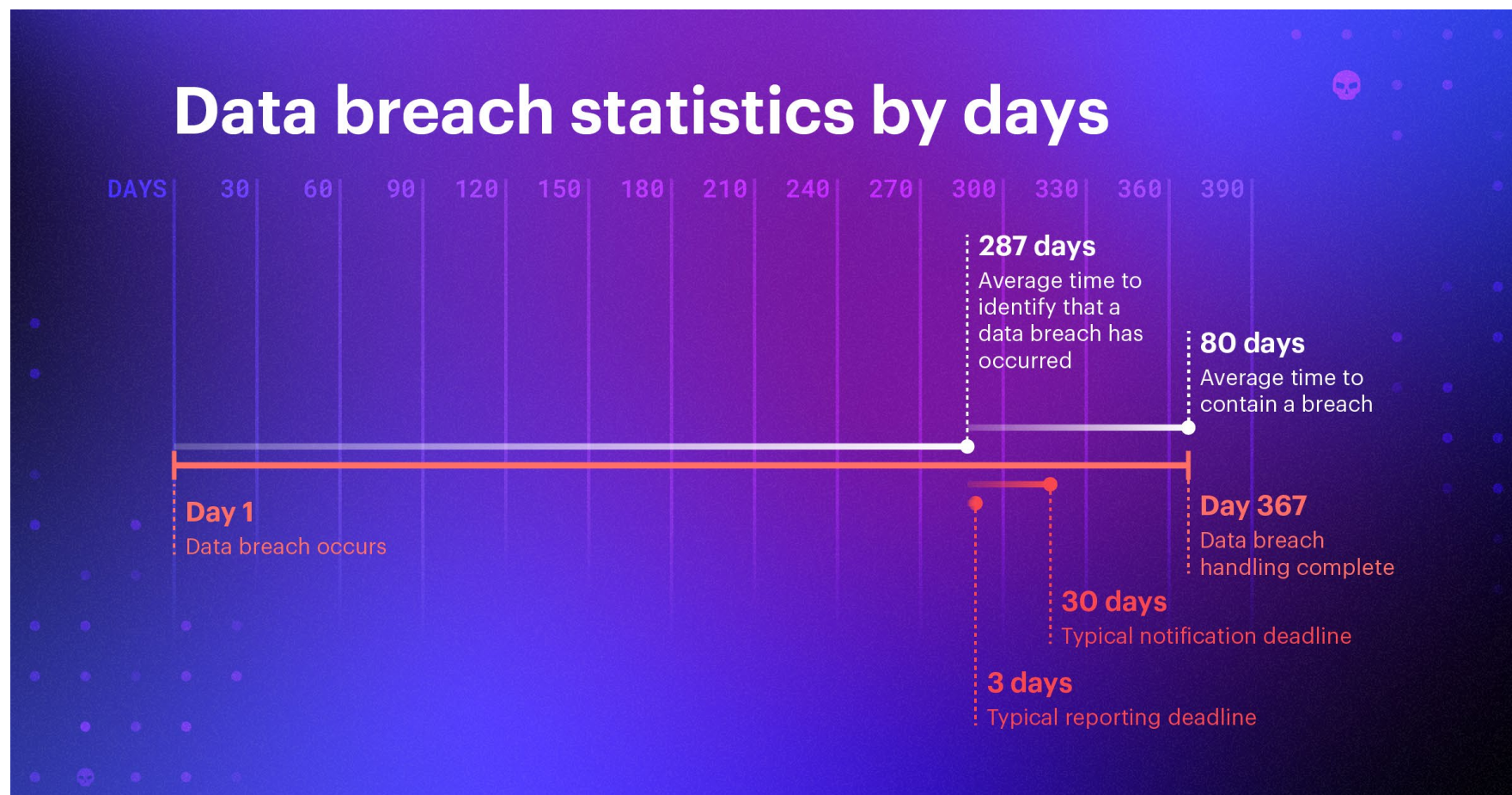
Big (centralised) data ⇒ Big data breaches



Source: <http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>

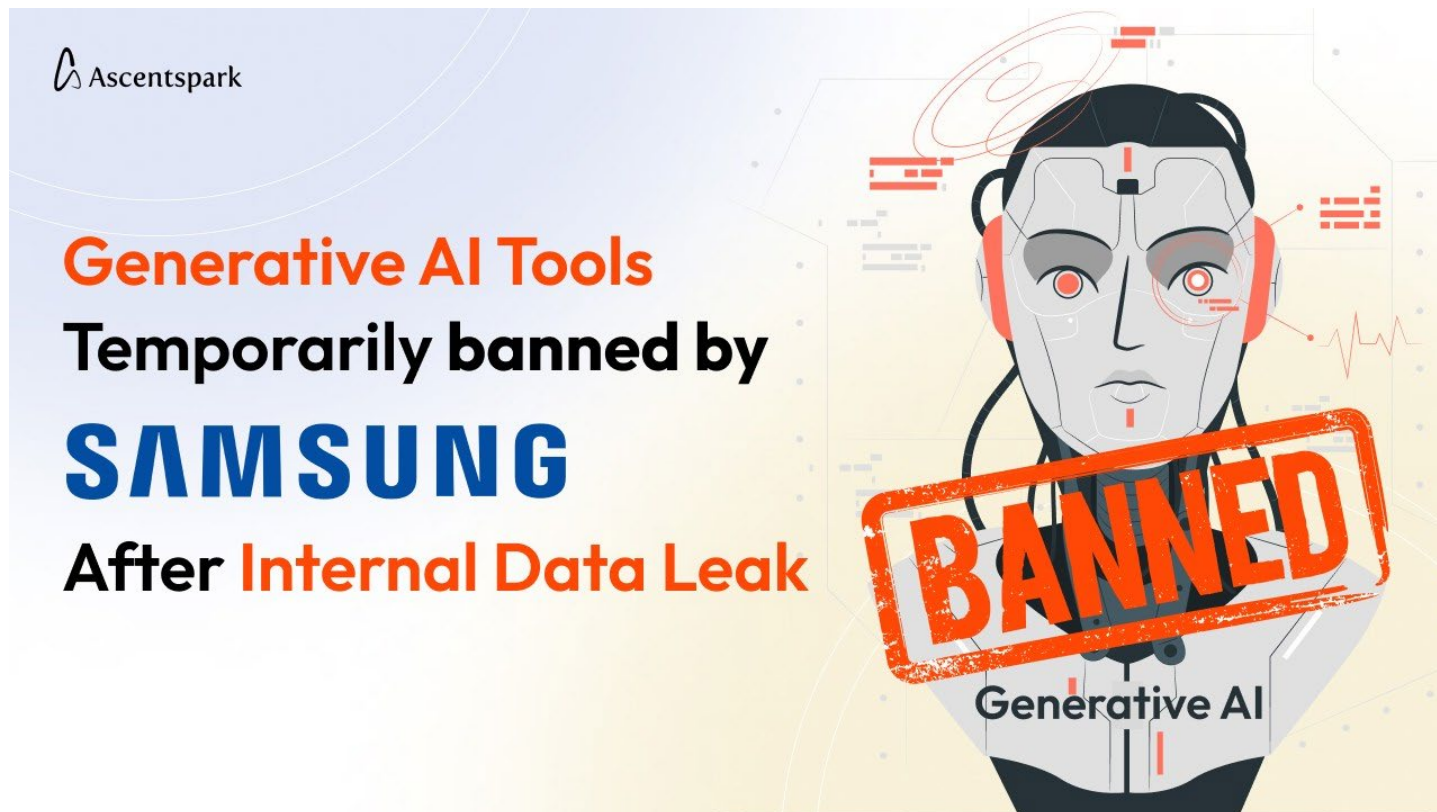
Data breach lifespan

- It can last long!



Source: <https://www.varonis.com/blog/data-breach-statistics>

- Generative AI Tools Temporarily Banned by Samsung After Internal Data Leak (2023)



Source: <https://www.linkedin.com/pulse/generative-ai-tools-temporarily-banned-samsung-after/>

Be informed about data breaches

Home Notify me Domain search Who's been pwned Passwords API About Donate

';--have i been pwned?

Check if your email address is in a data breach

email address pwned?

Using Have I Been Pwned is subject to the terms of use

Generate secure, unique passwords for every account

824 pwned websites 14,176,839,979 pwned accounts 115,796 pastes

Largest breaches

Count	Breach Name	Logo
772,904,991	Collection #1 accounts	Collection #1
763,117,241	Verifications.io accounts	Verifications.io
711,477,622	Onliner Spambot accounts	Onliner Spambot
622,161,052	Data Enrichment Exposure From PDL Customer accounts	Data Enrichment
593,427,119	Exploit.In accounts	Exploit.In
509,458,528	Facebook accounts	Facebook
457,962,538	Anti Public Combo List accounts	Anti Public Combo List
393,430,309	River City Media Spam List accounts	River City Media
361,468,000	Combolists Posted To Telegram	Combolists

Recently a

Count	Breach Name	Logo
1,374,344	TN	TN
3,118,964	Vir	Vir
1,772,620	Str	Str
6,342	Th	Th
304,337	dig	digDirect
134,336	Fair vote Canada accounts	Fair vote Canada
898,681	AlpineReplay accounts	AlpineReplay
31,081,179	Internet Archive accounts	Internet Archive
1,910,261	Muah.AI accounts	Muah.AI

Source:

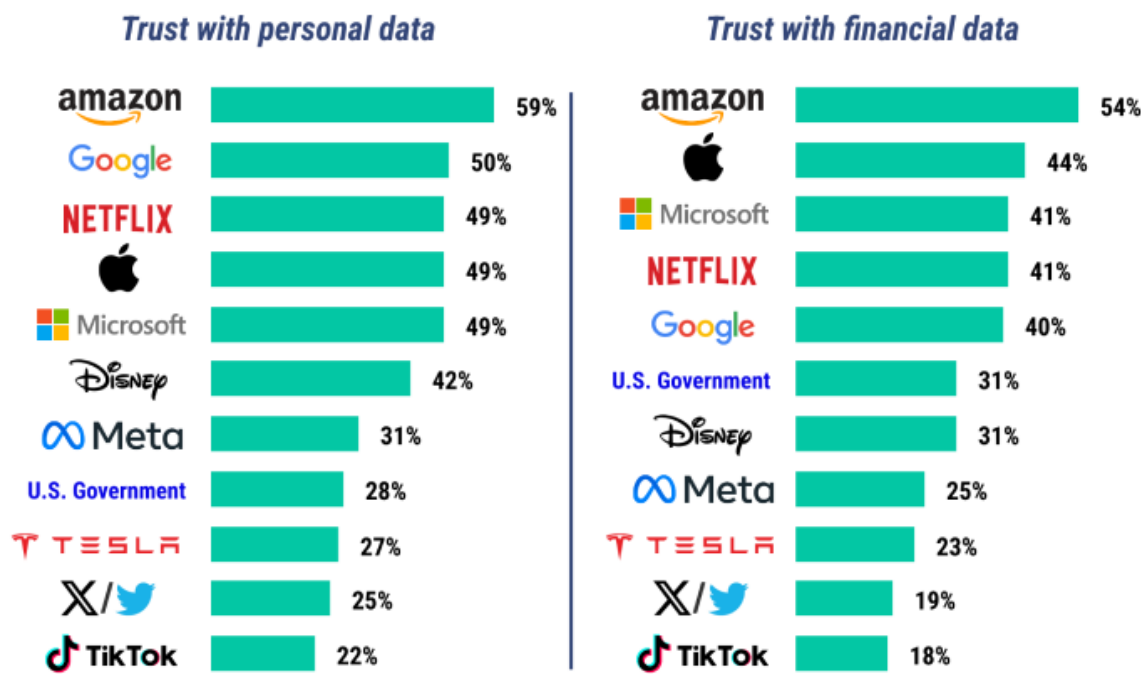
<https://haveibeenpwned.com/>

Big data ⇒ Big privacy concerns

- Big data consumers lead to big privacy concerns.

Do You Trust Tech Companies With Your Data?

Which companies are people most and least confident can keep their personal and financial data safe? We found how many people trust different companies when it comes to data security.



Based on a survey of 1,000 U.S. adults.

Big data \Rightarrow Big privacy concerns

- UGC leads to many privacy problems!



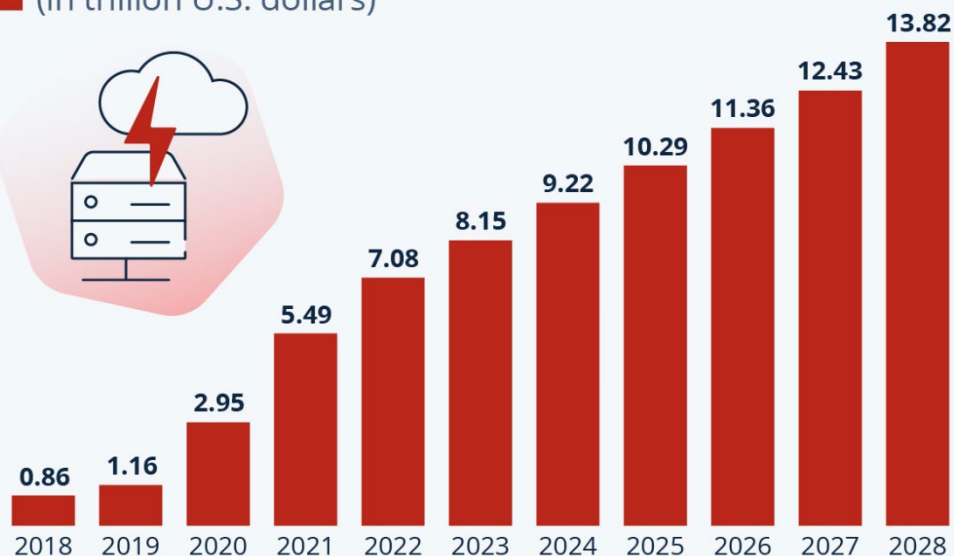
Source: <https://youtu.be/F7pYHN9iC9I>

Big data \Rightarrow Big cybercrime cases

- Cyber criminals have access to more data!

Cybercrime Expected To Skyrocket

Estimated annual cost of cybercrime worldwide
(in trillion U.S. dollars)



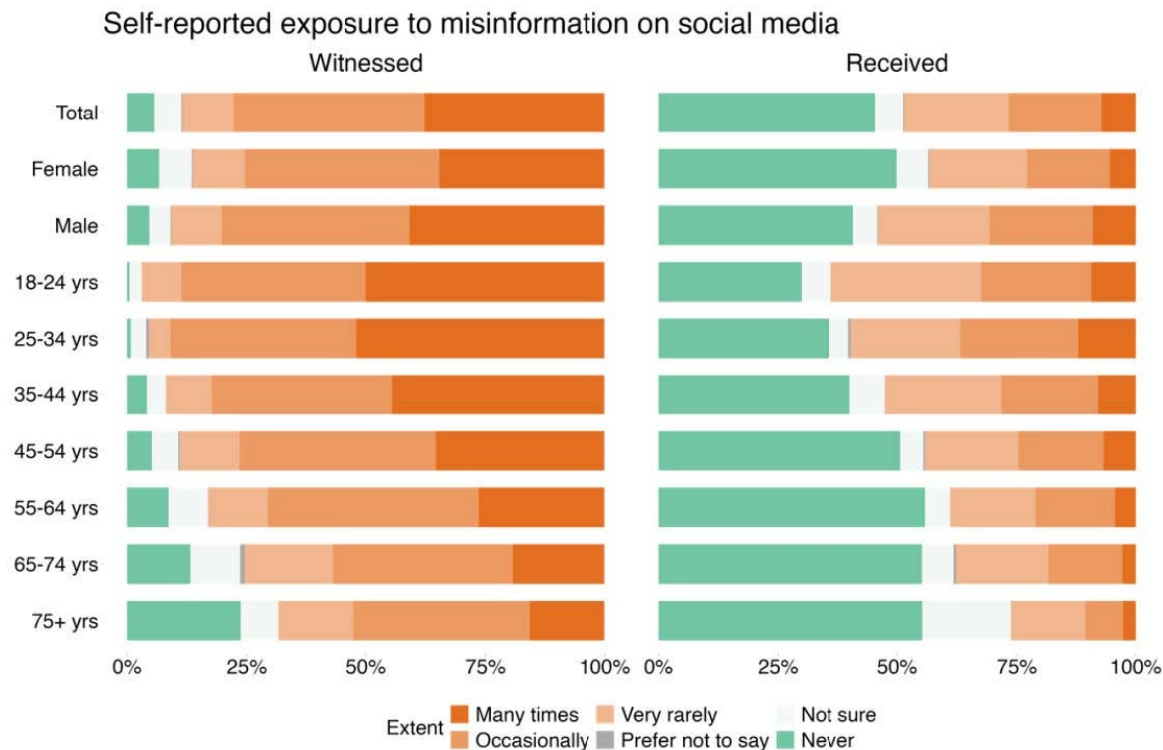
As of Sep. 2023. Data shown is using current exchange rates.

Source: Statista Market Insights

Source:
<https://www.statista.com/chart/28878/expected-cost-of-cybercrime-until-2027/>

Too much misinformation!

- Self-reported exposure to misinformation on social media (1,993 UK participants of a 2024 survey)



Source: <https://www.turing.ac.uk/news/publications/how-do-people-protect-themselves-against-online-misinformation>

Too much harmful online data!

- Harmful content leads to online safety concerns.

Online Safety for Kids

27% of 7- to 17-year-olds have come into contact with harmful content online. 

Despite **93%** of parents discussing online safety, only **49%** of children aged **12 to 15** claim to have had these conversations.

1 in 3 children have received links leading to malicious sites. 

40% of children in grades 4 to 8 admit to interacting with strangers online. **15%** have tried to meet strangers.

Talk. Protect. Educate.
Keep the conversation going for a safer online world.

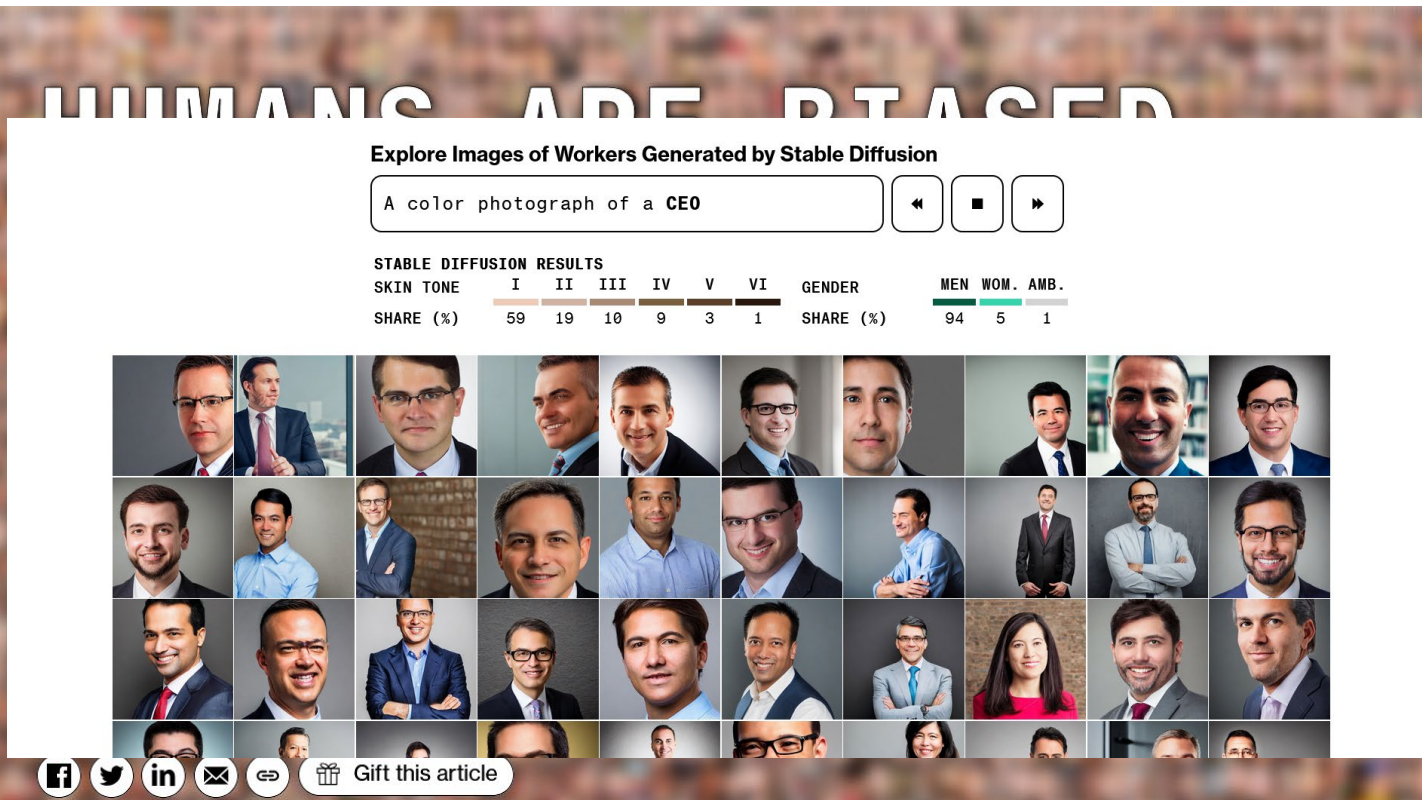
Sources: Zipdo, Salesforce 

Source:

<https://www.vpnmentor.com/blog/the-ultimate-parent-guide-for-child-internet/>

With AI, things can get worse!

- Biased, discriminative, irresponsible and unethical AI systems lead to many concerns.



- A simple lack of knowledge and understanding of the data ecosystem leads to missed (known and unknown) opportunities for many good things!
 - Better commercial opportunities
 - More personalised services
 - Better protection of people and other assets
 - Improved regulations, policies, and guidelines
 - Saved operational costs and more efficient of resources
 - ...



There is also a research gap!

- “Big data” vs “data flow” research: 163k vs <9k

The image displays two screenshots of Google Scholar search results. The top screenshot shows the search results for "big data", with approximately 163,000 results. The bottom screenshot shows the search results for "data flow", with approximately 8,990 results. Both screenshots show a list of articles with titles, authors, and publication dates.

Search 1: "big data"

- Articles: About 163,000 results (0.10 sec)
- Any time: Since 2024, Since 2023, Since 2020, Custom range...
- Sort by relevance, Sort by date
- Any type: Review articles
- include patents, include citations, Create alert
- Articles:
 - Big data: A review** [PDF] uccs.edu
 - S Sagiroglu, D Sinanc - 2013 international conference on ... 2013 - ieeexplore.ieee.org

Search 2: "data flow"

- Articles: About 8,990 results (0.40 sec)
- Any time: Since 2024, Since 2023, Since 2020, Custom range...
- Sort by relevance, Sort by date
- Any type: Review articles
- include patents, include citations, Create alert
- Articles:
 - [book] **Data flow analysis: theory and practice** [PDF] academia.edu
 - U Khedker, A Sanyal, B Sathe - 2017 - taylorfrancis.com
 - ... This book provides a detailed treatment of **data flow** analysis. Although we explain it in the ... model of **data flow** equations to represent the specification of **data flow** analysis. These ...
 - Data flow languages** [PDF] ieee.org
 - WB Ackerman - 1979 International Workshop on Managing ..., 1979 - ieeexplore.ieee.org
 - ... **Data flow** computers also have the goal of taking advantage of parallelism. As will be seen below, the parallelism in a **data flow** ... Like the other forms of parallel computer, **data flow** ...
 - Synchronous data flow** [PDF] ieee.org
 - EA Lee, DG Messerschmitt - Proceedings of the IEEE, 1987 - ieeexplore.ieee.org
 - ... **Data flow** is a natural paradigm for describing DSP applications for concurrent implementation on parallel hardware. **Data flow** ... usually associated with **data flow** evaporates. Multiple ...
 - Data flow analysis in software reliability** [PDF] acm.org
 - LD Fosdick, LJ Osterweil - ACM Computing Surveys (CSUR), 1976 - dl.acm.org
 - ... Our primary goal in using **data flow** analysis is the detection of **data flow** anomalies. The examples above hardly require very sophisticated techniques for their detection. However, it can ...

Most data flow research is about computational technologies rather than understanding different real-world data ecosystems and helping end users.

From data flow analysis to ...

A case study: leisure travel



Travel-related apps

- How many such apps do you use?



What does “travel-related” mean?

- You use not just “travel-related” apps while travelling, don’t you?



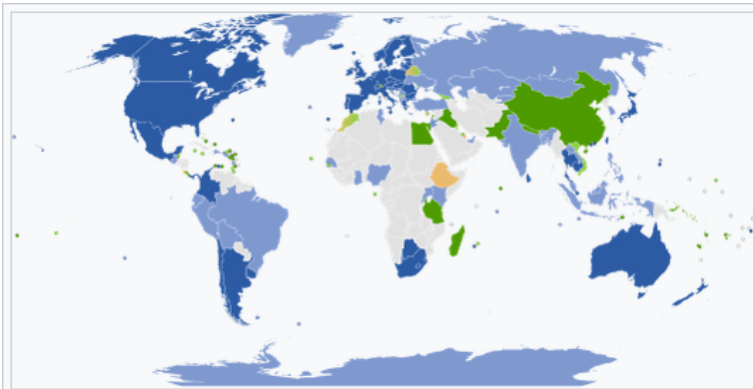
And it is not just about you!

- Others record your travel-related data, too!

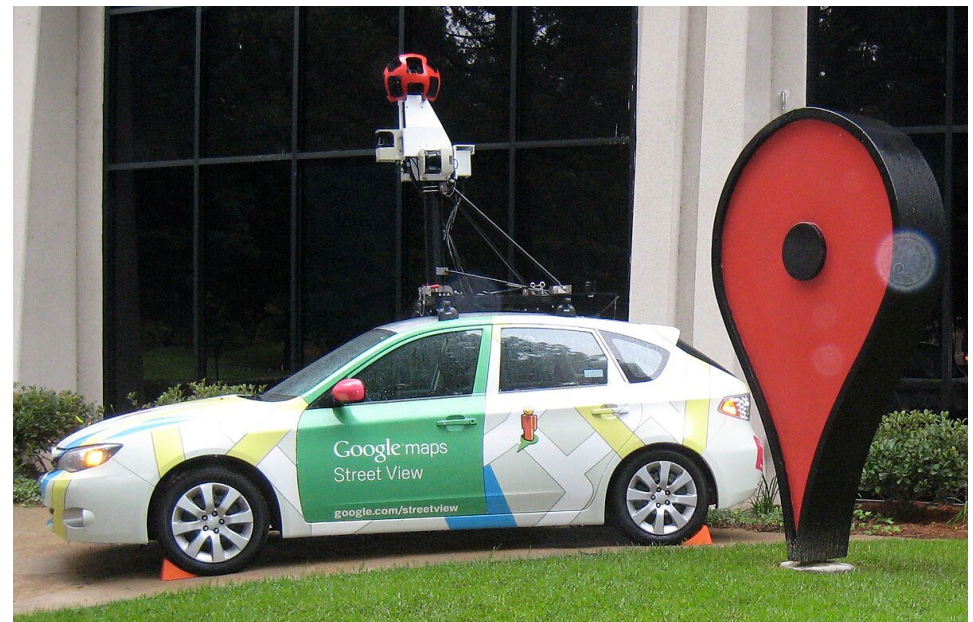


And it is not just about you!

- Others record your travel-related data, too!



- Countries and dependencies with mostly full coverage
- Countries and dependencies with partial coverage
- Countries and dependencies with official coverage planned
- Countries and dependencies with unofficial coverage planned
- Countries and dependencies with views of selected businesses and/or tourist attractions only
- Countries and dependencies with views of third party images of streets and/or landmarks
- Countries and dependencies without current or planned coverage

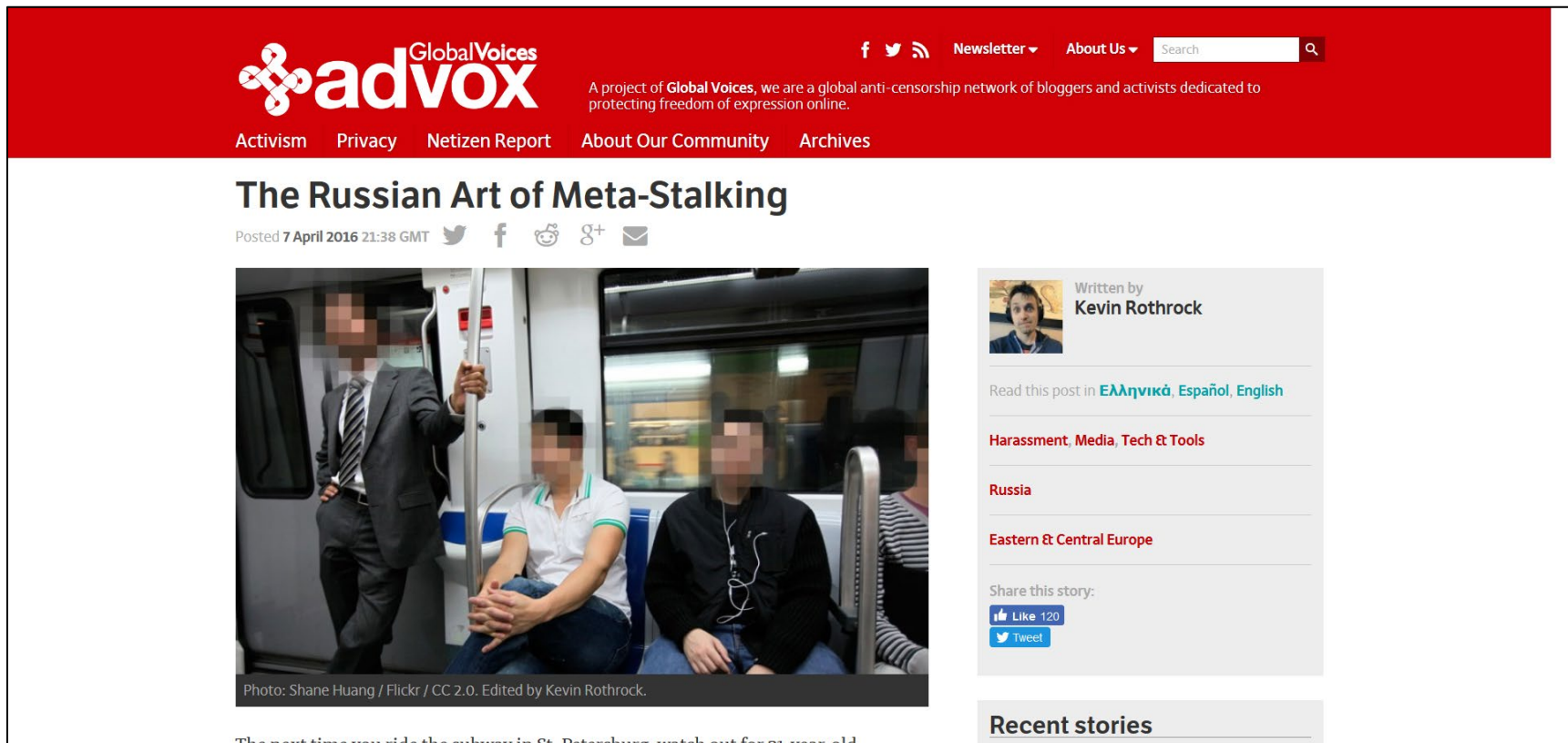


Source:

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

And it is not just about you!

- Others record your travel-related data, too!



The screenshot shows the Advox website interface. At the top, there is a red navigation bar with the Advox logo (a stylized 'a' made of four dots) and the text 'GlobalVoices advox'. To the right of the logo are social media icons for Facebook, Twitter, and RSS, followed by 'Newsletter', 'About Us', and a search bar. Below the navigation bar, the article title 'The Russian Art of Meta-Stalking' is displayed in large black font. Underneath the title, it says 'Posted 7 April 2016 21:38 GMT' and includes social sharing icons for Twitter, Facebook, Reddit, Google+, and Email. The main content area features a photograph of a subway car interior. A man in a dark suit and tie is standing and holding a vertical pole, looking towards the camera. Two other people are seated in the foreground, their faces blurred. The photo is credited to 'Photo: Shane Huang / Flickr / CC 2.0. Edited by Kevin Rothrock.' To the right of the photo, there is a sidebar with a profile picture of Kevin Rothrock, the author, and the text 'Written by Kevin Rothrock'. Below this, there are links to read the post in 'Ελληνικά, Español, English', and tags for 'Harassment, Media, Tech & Tools', 'Russia', and 'Eastern & Central Europe'. At the bottom of the sidebar, there are social sharing buttons for 'Like 120' and 'Tweet'. Below the sidebar, there is a section titled 'Recent stories'.

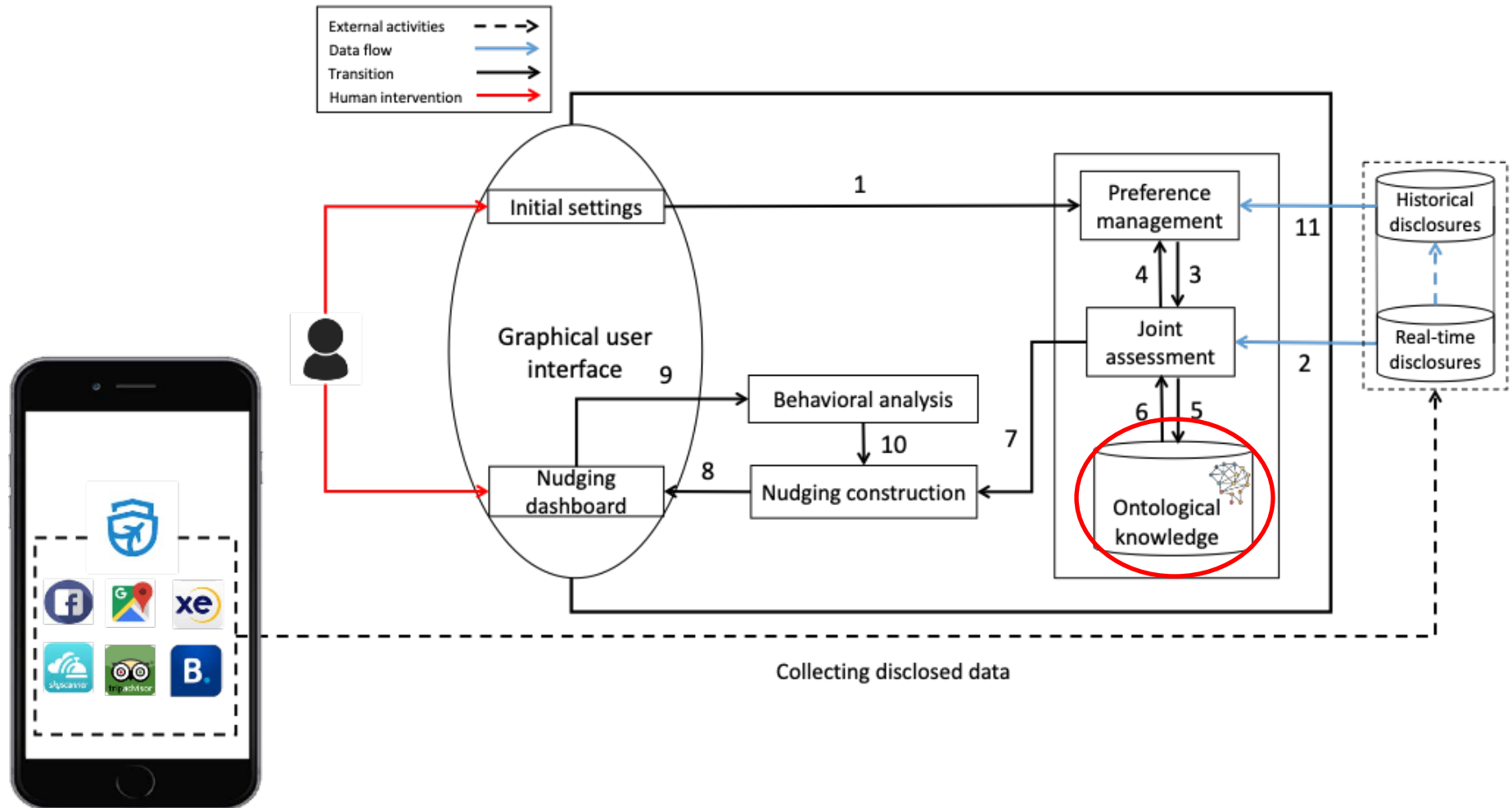
Source: <https://advox.globalvoices.org/2016/04/07/the-russian-art-of-meta-stalking/>

- Title: PRiVacy-aware personal data management and Value Enhancement for Leisure Travellers (**PRiVELT**)
- Funder:  Engineering and Physical Sciences Research Council
- Call: Trust, Identity, Privacy and Security in the Digital Economy 2.0 (2018)
- Budget: £~1.4m
- Duration: 10/2018 – 06/2023 (57 months)
- Website: <https://privelt.ac.uk/>

(Part of) The (former) project team

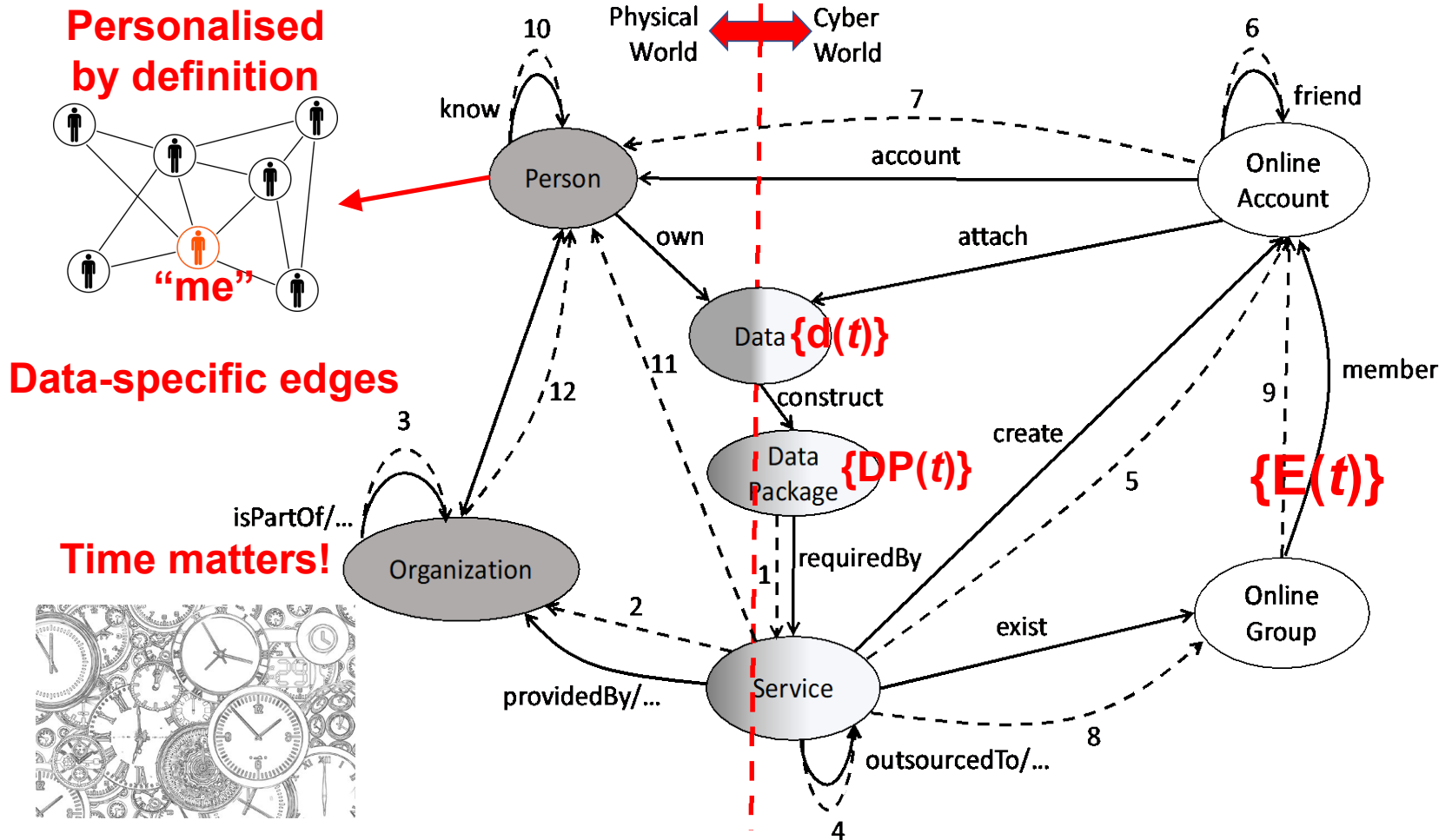


The vision: user-centric, server-less, from privacy awareness to nudging



Yang Lu, Shujun Li, Athina Ioannou and Iis Tussyadiah (2019) [From Data Disclosure to Privacy Nudges: A Privacy-aware and User-centric Personal Data Management Framework](#). In *Proc. DependSys 2019*, Springer. doi:10.1007/978-981-15-1304-6_21

Data sharing (flow) ontology



Yang Lu and Shujun Li (2020) [From Data Flows to Privacy Issues: A User-Centric Semantic Model for Representing and Discovering Privacy Issues](#). In *Proc. HICSS 2020*, University of Hawai'i at Mānoa. doi: 10.24251/HICSS.2021.651

Is a data flow graph complex?

- Number of nodes: **large**
 - “Me”: the “centre” / owner of the graph
 - All data item and data packages about “me”
 - All people your data can flow to (could be **anyone**)
 - All physical and online services you data can flow to
 - All organizations your fata can flow to
- Number of edges: **huge**
 - Relationships between different types of nodes
 - Often more than one edge between any two nodes

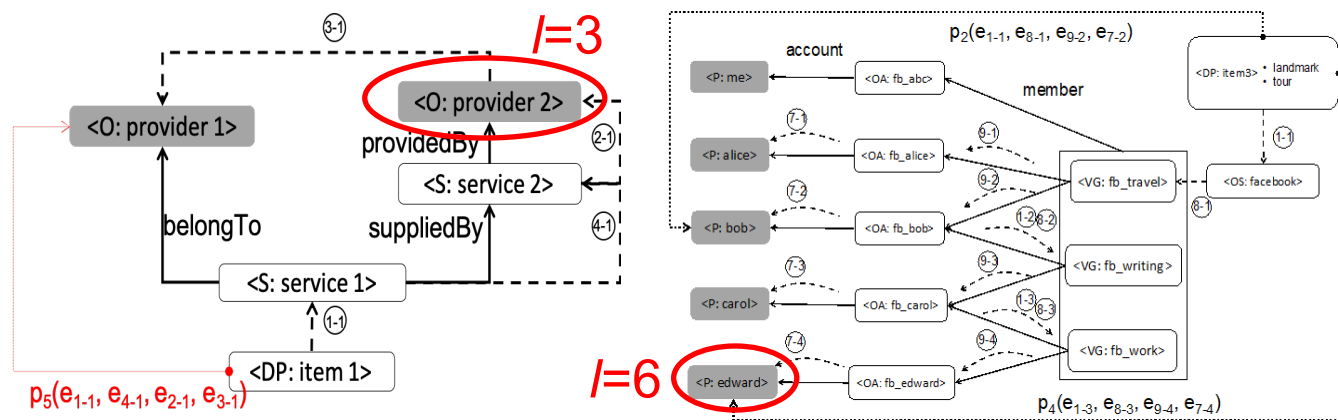
- Out-degree (of “me” node), given a time window
 - The amount of data shared
- Average nodal degree (of a data consumer node)
 - The average amount of “my” data disclosed to that data consumer
- Node / Link connectivity (of the whole graph)
 - The number of “essential” data consumers / data sharing activities
- Centrality metrics (of data consumer nodes)
 - For identifying major (potentially “hidden”) data consumers
- The longest path(s) originating from “me”
 - For identifying the most “hidden” data consumer(s)
- Network type (of the whole graph)
 - Small-world network, scale-free network or something else?
- ...

“Topological” privacy issues

- A specific privacy issue with a specific data item or a data package corresponds to a **data flow path**.
- A specific privacy issue with more than one data items and/or data packages corresponds to **a set of data flow paths**.
- A specific **type** of privacy issues of one or more data item / package type(s) is **a set of sets of data flow paths**.
- All of them can be **described** and **potentially detected** via their **common topological features**.

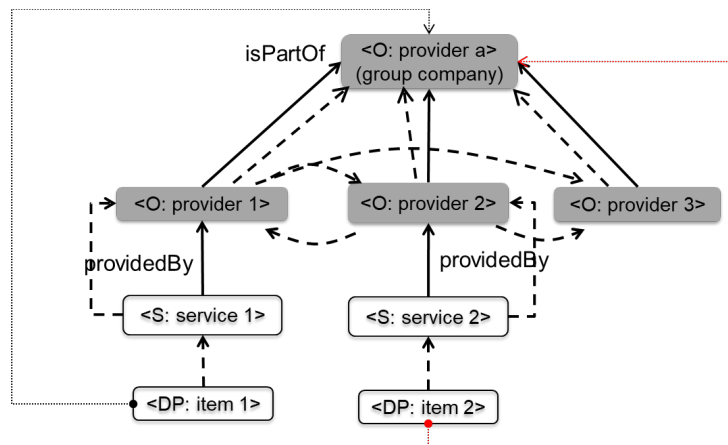
“Topological” privacy issue #1

- Data shared with (potentially) **unknown consumers**
 - **Hypothesis**: the longer the data flow path length between “me” and a data consumer node is, the more likely the user is unaware of the data consumer
 - **Risk assessment**: $r=f(l)$, where l is the path length
 - **Visualization**: show a ranked list of all potential unknown data consumers with decreasing values of r
 - **Detection (naïve method)**: $r > r_t \Rightarrow$ issue an alert



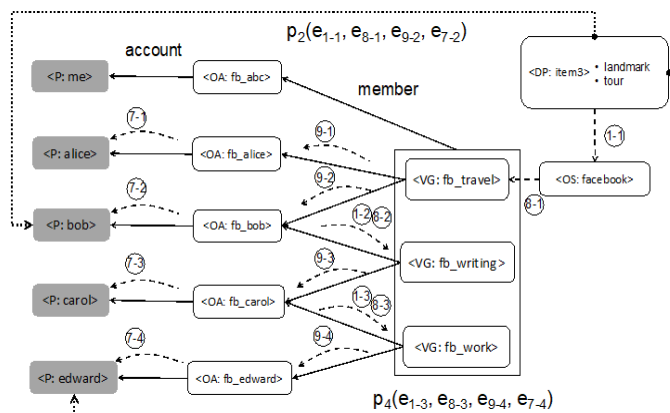
“Topological” privacy issue #2

- Indirect (= potentially unknown) **data aggregator**
 - **Hypothesis**: given a tree whose root node is a data consumer, the taller the tree is, the more likely the root node is an unknown (super) data aggregator
 - **Risk assessment**: $r=f(h)$, where h is the tree's height
 - **Visualization**: show a ranked list of all potential data aggregators with decreasing values of r
 - **Detection (naïve method)**: $r > r_t \Rightarrow$ issue an alert



“Topological” privacy issue #3

- Data shared with **too many consumers**
 - **Hypothesis**: given a tree whose root is a data node, the bigger the tree is, the more likely the data has been over-shared too much
 - **Risk assessment**: $r=f(n)$, where n is the total number of nodes in the tree minus 1 (the root node)
 - **Visualization**: show the whole tree
 - **Detection (naïve method)**: $r > r_t \Rightarrow$ issue an alert



Automatic reasoning is possible!

DL query:
Query (class expression)
Service_Provider **that** access **some** (Data **that** has **some** Sensitive)

Execute Add to ontology

Query results

Direct superclasses (1 of 1)
● Service_Provider

Instances (11 of 11)

- ◆ Agoda
- ◆ Booking.com
- ◆ GoToGate
- ◆ Kayak
- ◆ OpenTable
- ◆ Princline.com
- ◆ Rentalcars.com
- ◆ TravelJigsaw
- ◆ flygresor.se
- ◆ mytrip.com
- ◆ supersaver

DL query:
Query (class expression)
Person **that** access **some** (Data_Package **that** has **some** Location) **and** access **some** (Data_Package **that** has **some** Event)

Execute Add to ontology

Query results

Instances (1 of 1)
◆ edward

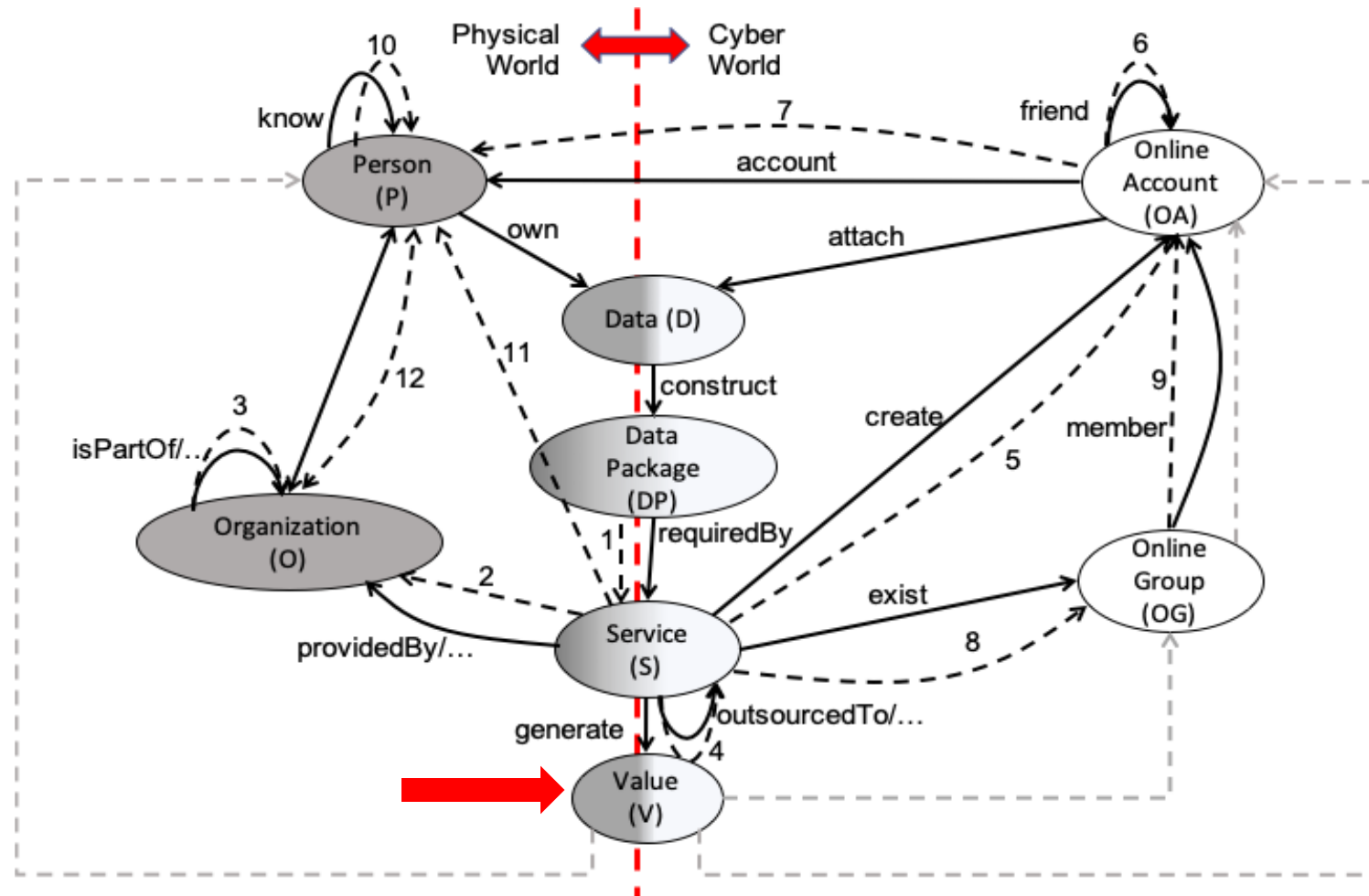
DL query:
Query (class expression)
Data_Package **that** has **some** Entertainment **that** (flowTo **some** Work)

Execute Add to ontology

Query results

Instances (1 of 1)
◆ item3

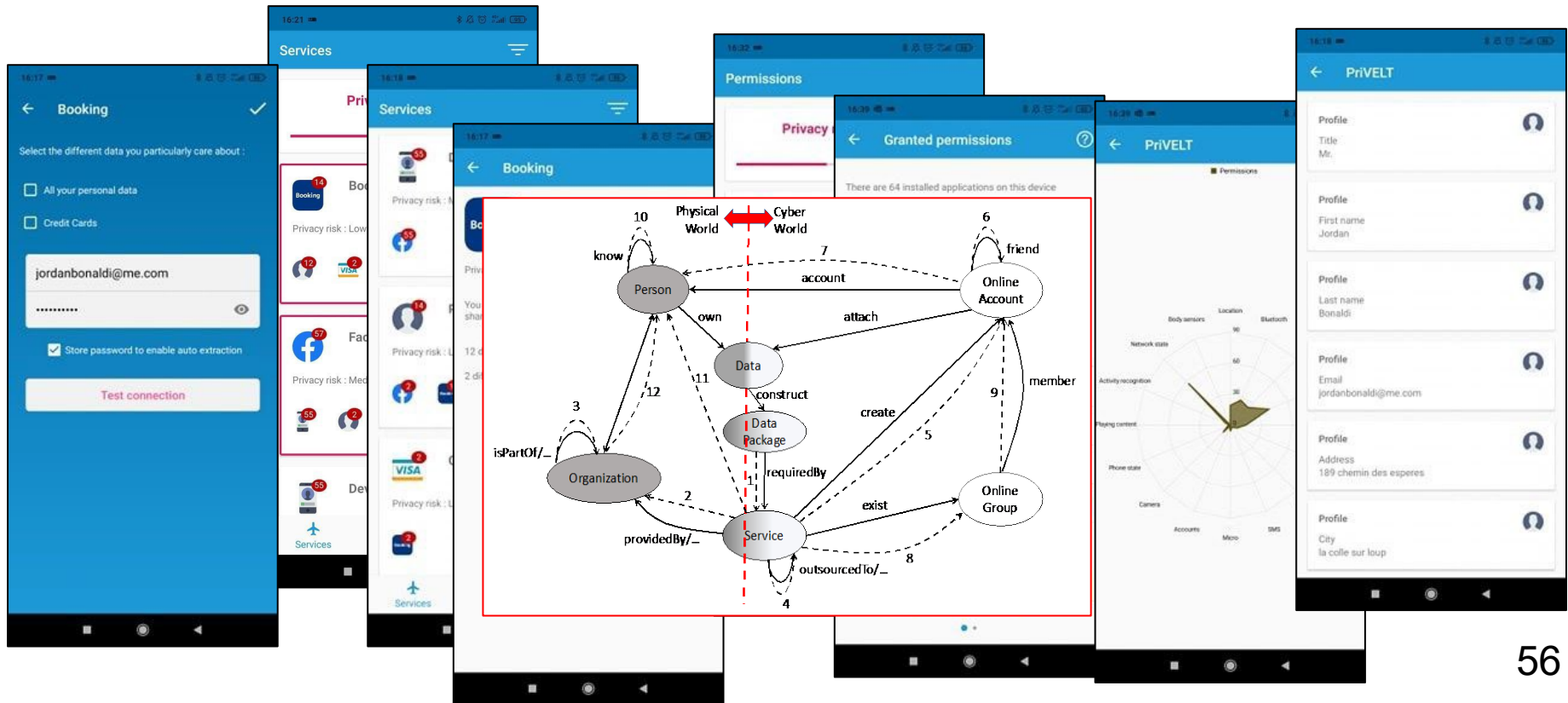
Adding returned values



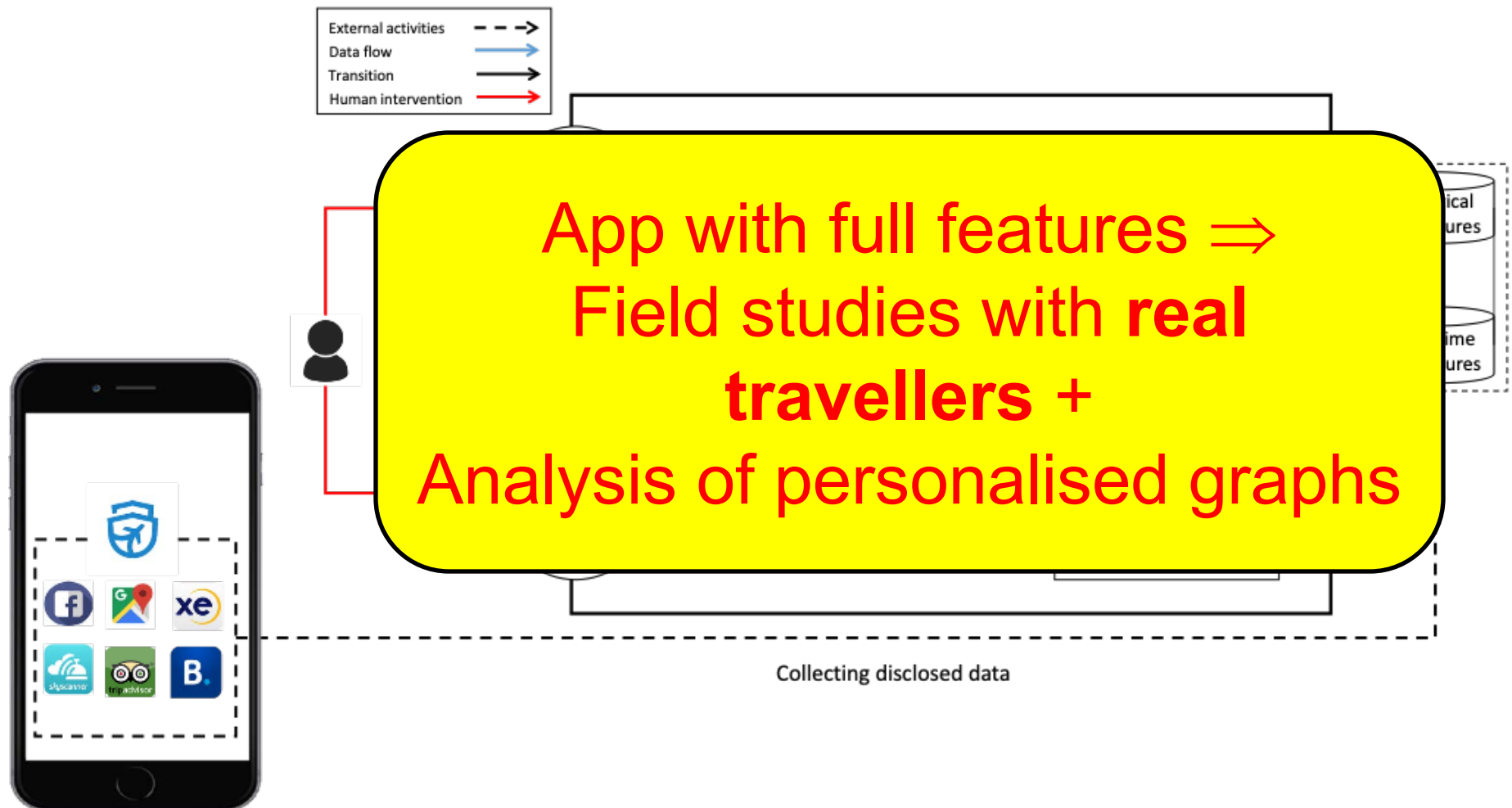
Yang Lu and Shujun Li (2022) [From Data Flows to Privacy-Benefit Trade-offs: A User-Centric Semantic Model](#). *Security and Privacy*, 5(4):e225, 24 pages, [John Wiley & Sons, Inc.](#) doi: 10.1002/spy2.225

Personalised data flow graphs

- User-centric and service-independent tools are needed to build “my” data flow graph.
- \Rightarrow We have been developing an Android app.



The vision: user-centric, server-less, from privacy awareness to nudging



Yang Lu, Shujun Li, Athina Ioannou and Iis Tussyadiah (2019) [From Data Disclosure to Privacy Nudges: A Privacy-aware and User-centric Personal Data Management Framework](#). In *Proc. DependSys 2019*, Springer. doi:10.1007/978-981-15-1304-6_21

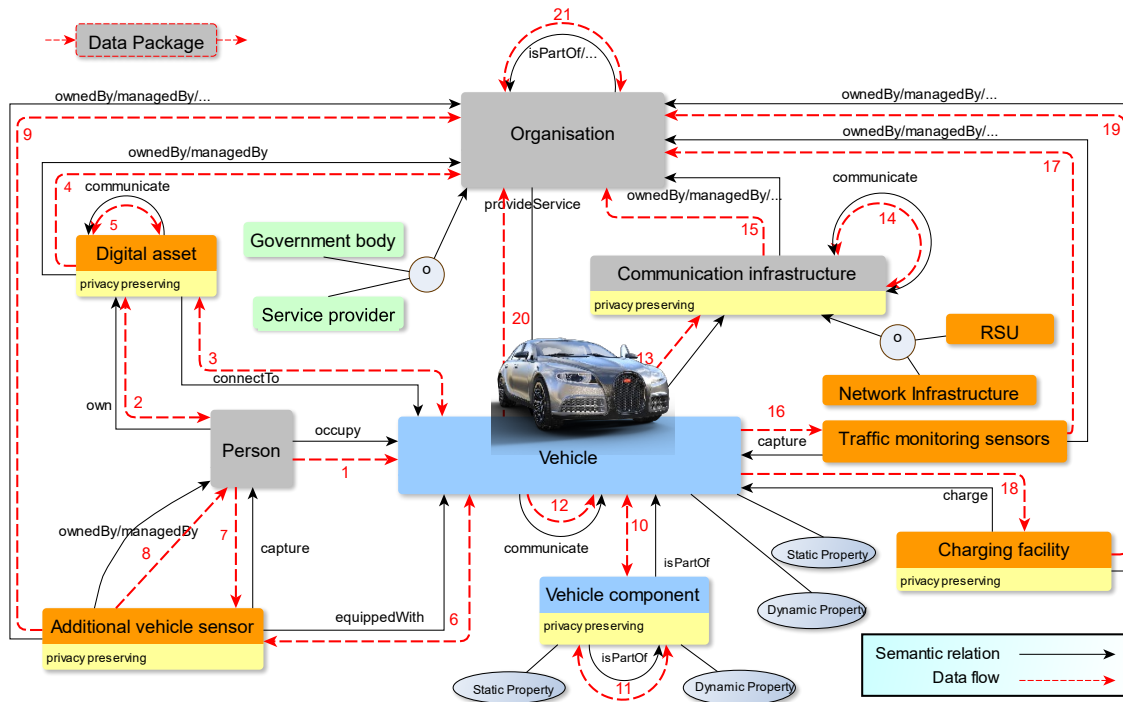
From data flow analysis to ...


Other case studies



Data flows around a car

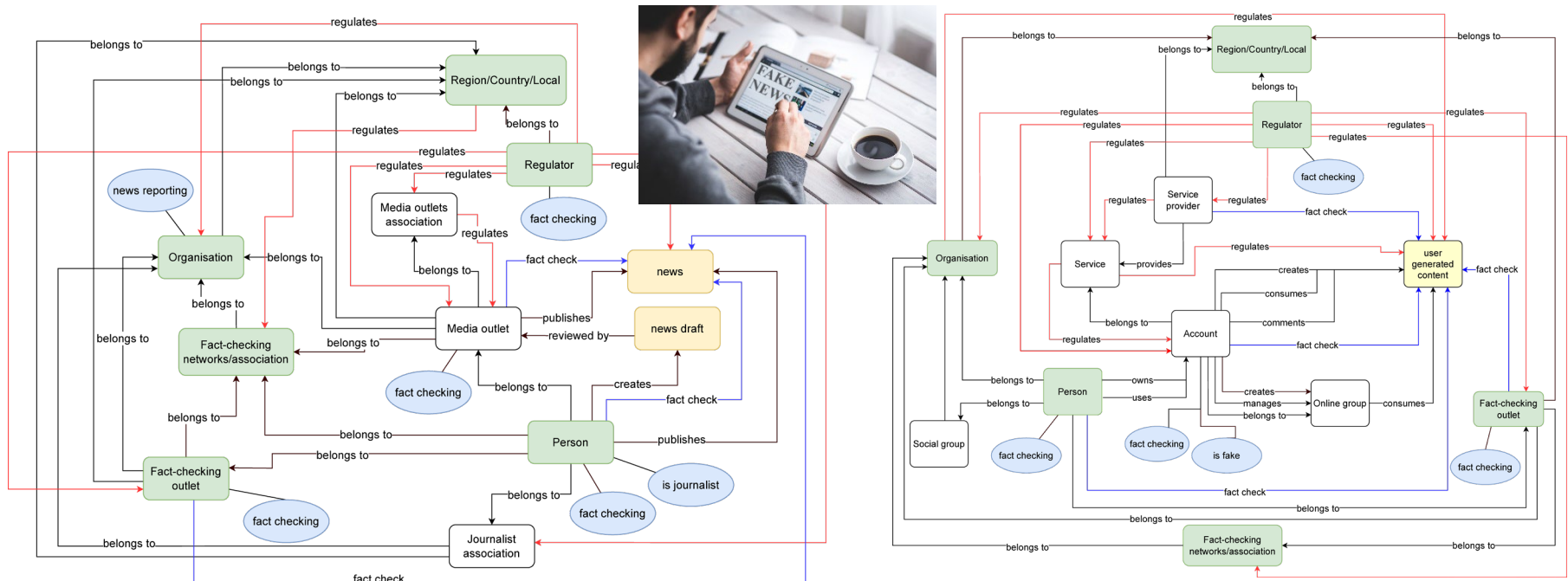
- Results from a research project funded by [Honda Research Institute Europe \(HRI-EU\)](#) in Germany.



 **OpenAI**
GPT-4
The use of LLMs was explored.

Haiyue Yuan, Ali Raza, Nikolay Matyunin, Jibesh Patra and Shujun Li (2024) [A Graph-Based Model for Vehicle-Centric Data Sharing Ecosystem](#). *Proceedings of ITSC 2024*, doi: 10.48550/arXiv.2410.22897

- For modelling the fact-checking ecosystem



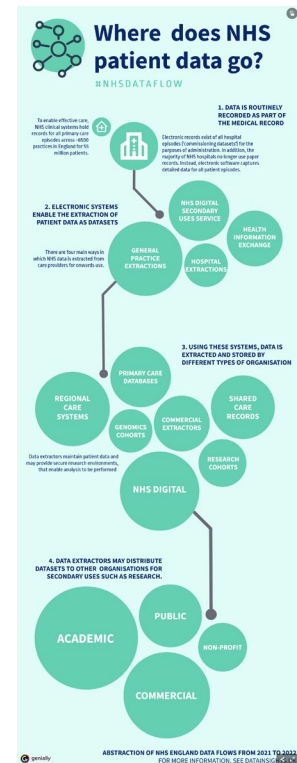
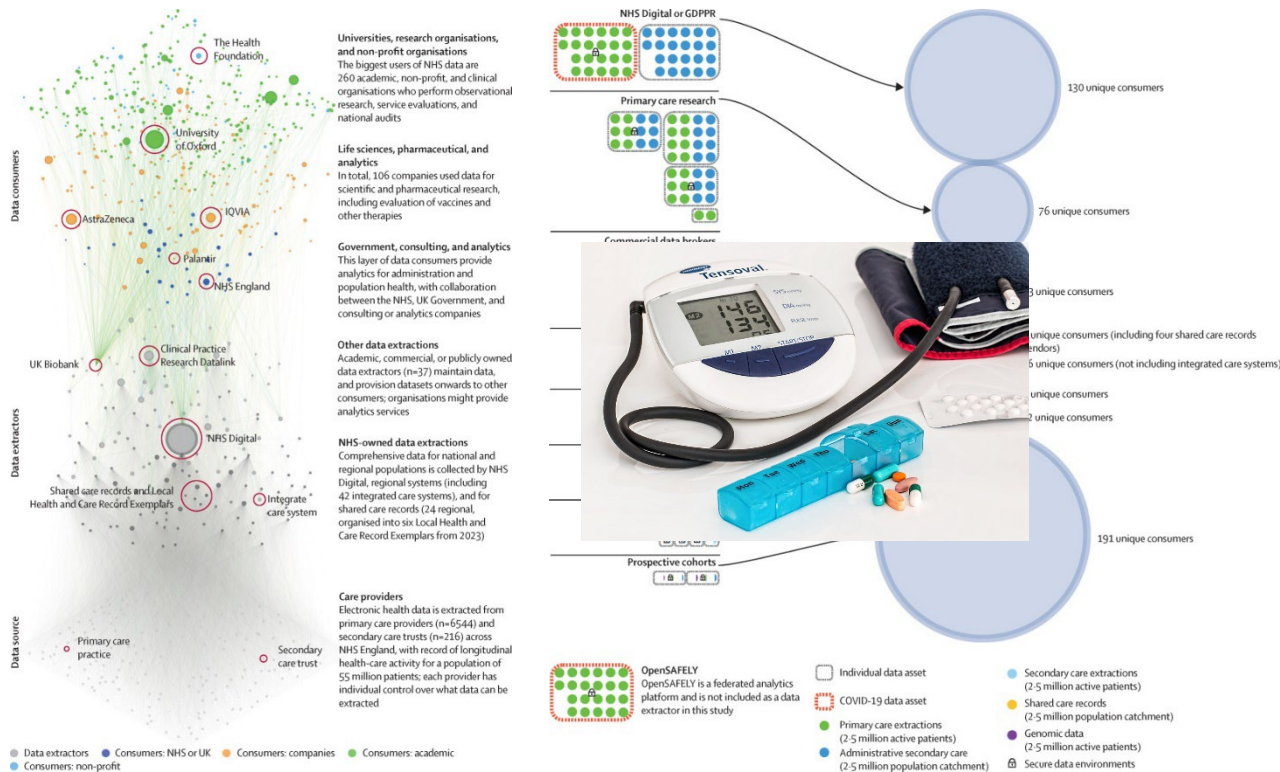
Haiyue Yuan, Enes Altuncu, Shujun Li and Can Baskent (2022) [Graphical Models of False Information and Fact Checking Ecosystems](#). Online preprint, arXiv:2210.04541

[cs.CR], doi: 10.48550/arXiv.2208.11582

(A substantially updated version will be uploaded soon.)

Data flows around health data

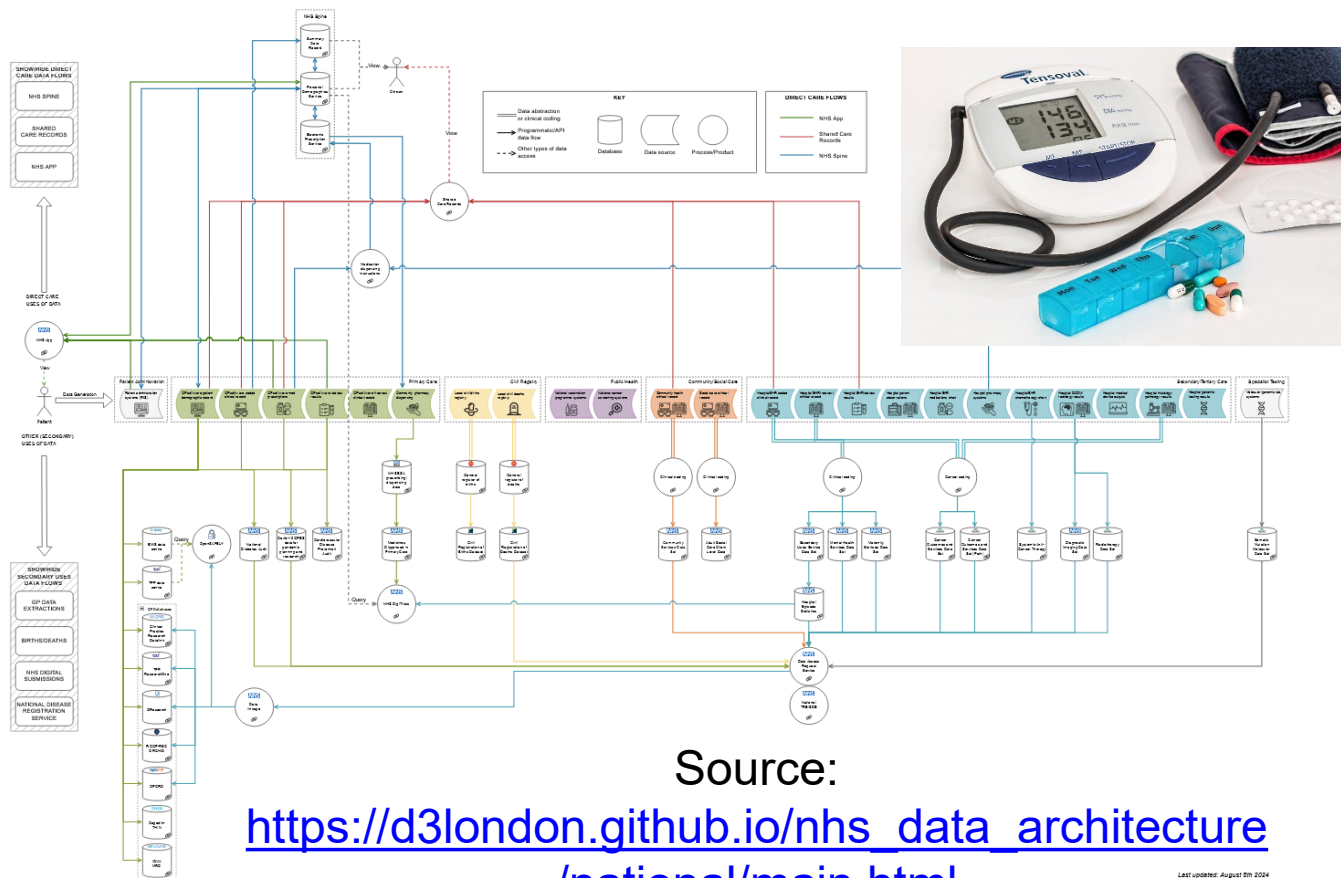
- We are starting a new research collaboration with Dr Joe Zhang of NHS and others on this topic.



Joe Zhang et al. (2023) [Mapping and evaluating national data flows: transparency, privacy, and guiding infrastructural transformation](#). *The Lancet Digital Health*, 5(10):e737-e748, doi: 10.1016/S2589-7500(23)00157-7

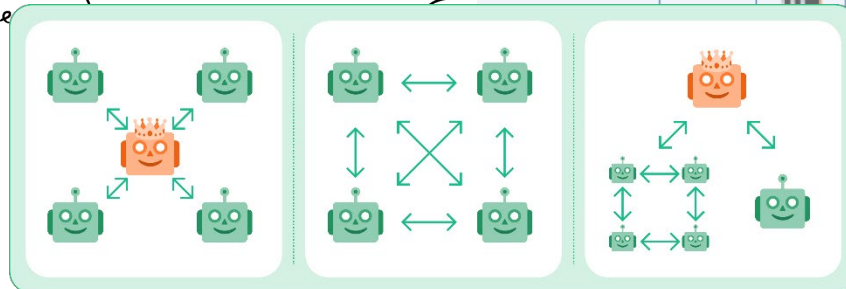
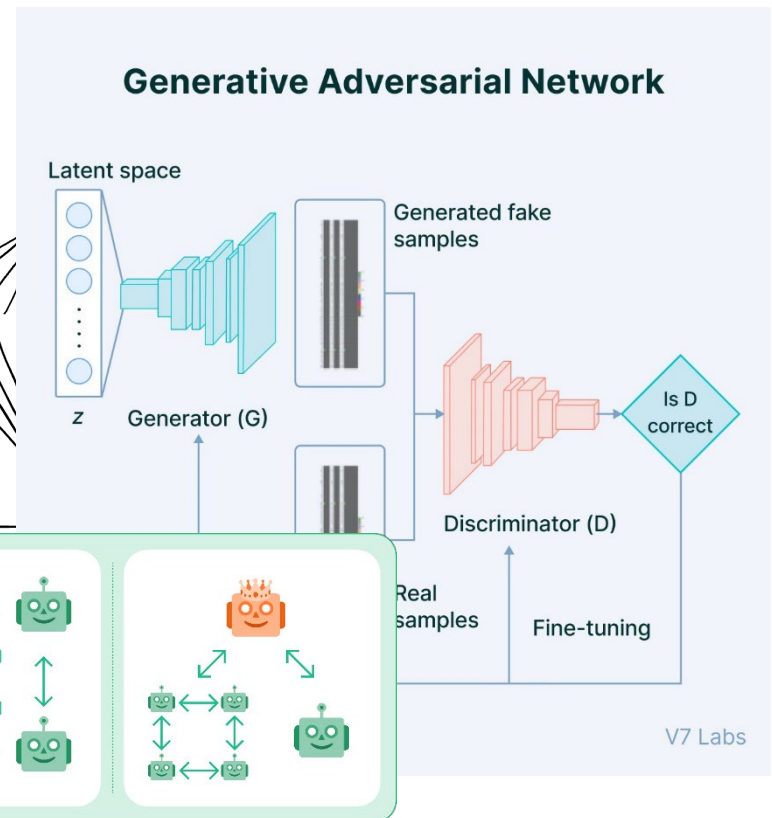
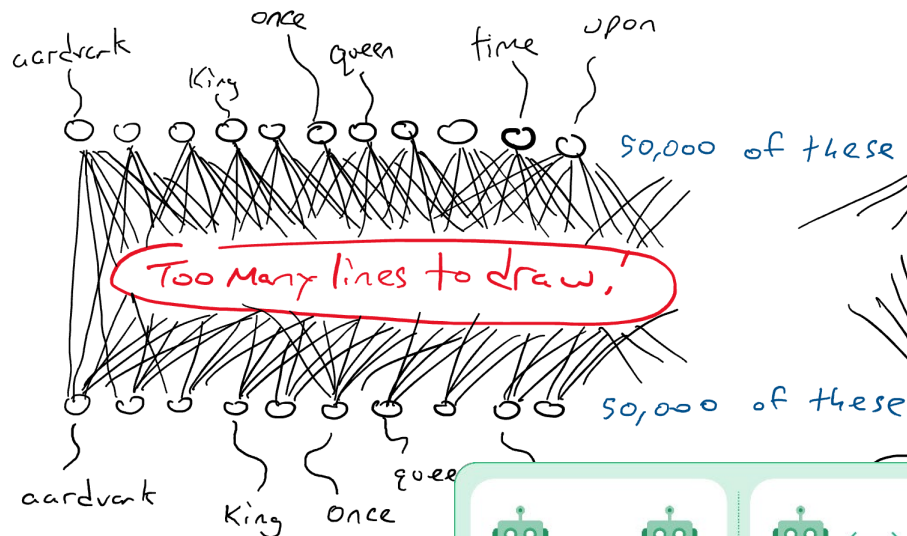
Data flows around health data

- We are starting a new research collaboration with Dr Joe Zhang of NHS and others on this topic.



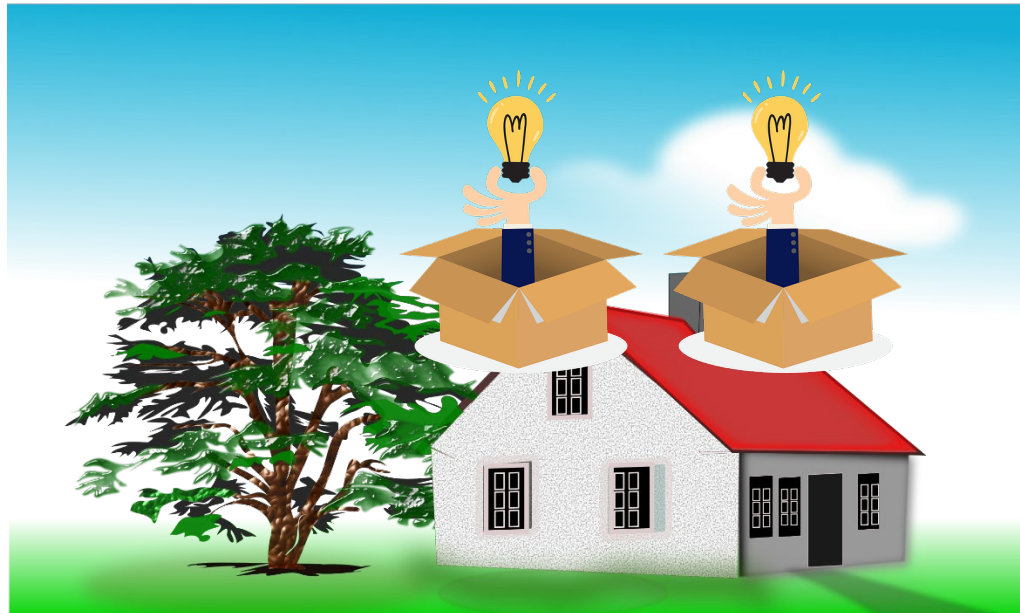
Data flows vs AI models/systems

- More responsible/transparent/trustworthy and safer AI require a better understanding of data flows within/between AI models.



From data flow analysis to ...

Take-Home Messages



What you can bring home...

- Many data-related problems **in all domains** can be studied using **data flow graphs**.
- Personal data flow graphs can be **personalised** around a node called “me”.
 - **User-centric** tools are needed to engage users.
- There are essential research questions across **multiple disciplines**.
 - Computer Science, Engineering, Psychology, Law, Business, Economics, Sociology, Ethics, Media and Communication, Education, domain-specific disciplines (e.g. Health and Tourism), ...
- We call more people (researchers, innovators, developers, policy makers, users and others) to **join us** on data flow related research, innovation, education and discussions.

From data flow analysis to ...



Shujun LI (李树钧)

<http://www.hooklee.com/>