

The arguments and the vision for a [Personal] Data Internetnetwork (PDI)

Nikolaos Laoutaris

Research Professor
IMDEA Networks Institute



Members of IMDEA Networks' Scientific Council and researchers

Major Event (Highlighted Event)

IMDEA Networks Institute annually holds a **by-invitation-only thematic workshop** in Madrid. The workshop accompanies a meeting of our Scientific Council comprised of prominent researchers. In addition to talks by Scientific Council members and Faculty Members, the workshop includes a breakout session divided into three clusters: Networked Systems and Algorithms, Wireless Networking and Measurements, Privacy & Security. The workshop program aims to present participants' recent research results, discuss and identify priorities and challenges for the research agenda, as well as explore new paths toward making future networks a true commodity for the needs of society.

Agenda (June 8th)

09:00 – 09:20 | Coffee, Doors open

09:20 – 09:30 | Opening remarks

Albert Banchs and Joerg Widmer, IMDEA Networks Institute (Spain)

09:30 – 10:15 | Modeling Service Mixes in Access Links [Keynote]

Marco Ajmone, Politecnico di Torino and IMDEA Networks Institute (Italy and Spain)

Abstract: Access links, both wired and wireless, carry the traffic generated by a mix of services, such as video, audio, web, gaming, etc. Some services generate traffic with a constant or almost constant data rate (for example a voice conversation or the live streaming of a sport event) while other services can adapt their data rate to the available network resources (for example a file transfer or a web page download).

The prediction of the performance that can be obtained in presence of mixes of traffic of different nature is important for the definition of access control and scheduling algorithms and has been tackled by several researchers, mostly resorting to approximations, because exact solutions are not available. This talk shows that a simple quasi-optimal admission control scheme allows the simple derivation of exact results that highlight interesting, unexpected behaviors.

10:15 – 11:00 | IMDEA Networks: vision for the future

Nikolaos Laoutaris, IMDEA Networks Institute (Spain)

Our vision for the future is ...

... to build a [Personal] Data Internetwork

PDI
(aka PD-ICN)

an overlay network over the existing Internet for connecting data owners, data marketplaces, and data consumers in a fair, trustworthy, and democratic manner that copies the paradigm of the Internet

What do Information Centric Networks, Trusted Execution Environments, and Digital Watermarking have to do with Privacy, the Data Economy, and their future?*

Nikolaos Laoutaris
IMDEA Networks Institute, Spain
nikolaos.laoutaris@imdea.org

Costas Iordanou
Cyprus University of Technology
costas.iordanou@eecei.cut.ac.cy

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

What if instead of having to implement controversial user tracking techniques, Internet advertising & marketing companies asked explicitly to be granted access to user data by name and category, such as Alice→Mobility→05-11-2020? The technology for implementing this already exists, and is none other than the Information Centric Networks (ICN), developed for over a decade in the framework of Next Generation Internet (NGI) initiatives. Beyond named access to personal data, ICN's in-network storage capability can be used as a substrate for retrieving aggregated, anonymized data, or even for executing complex analytics within the network, with no personal data leaking outside. In this opinion article we discuss how ICNs combined with trusted execution environments and digital watermarking, can be combined to build a personal data overlay inter-network in which users will be able to control who gets access to their personal data, know where each copy of said data is, negotiate payments in exchange for data, and even claim ownership, and establish accountability for data leakages due to malfunctions or malice. Of course, coming up with concrete designs about how to achieve all the above will require a huge effort from a dedicated community willing to change how personal data are handled on the Internet. Our hope is that this opinion article can plant some initial seeds towards this direction.

CCS CONCEPTS

• **Networks** → **Network privacy and anonymity**; • **Information systems** → **Web services**; *Online advertising*; • **Computing methodologies** → **Distributed computing methodologies**; *Machine learning*; • **Computer systems organization** → **Distributed architectures**; • **Security and privacy** → *Cryptography*; **Access control**; **Privacy-preserving protocols**; *Database and storage security*; **Distributed systems security**; **Privacy protections**;

KEYWORDS

Information Centric Networking, Overlay Network, Personal Data Management System, Online Privacy

*Title inspired by Jaron Lanier's, "Who Owns the Future?" [42].

1 INTRODUCTION

The Internet and the Web were built decades ago to fulfill objectives and requirements far different from? those of today. Both have managed to exhibit tremendous evolvability and extendibility and have succeeded in supporting services and capabilities that could hardly be imagined in the 60s (Internet) or even the 90s (Web). This has mostly been achieved via *layering*, *standardization*, and *openness*. Every time that new applications (content distribution, video conferencing) or capabilities (broadband connectivity, mobility) become critically important, additional layers are introduced on the top or the bottom of the protocol stack to implement the necessary new functionality and requirements. Such layers are often standardized via the open and flexible standardization process of the Internet (IETF, RFCs, etc.), or are at least made unofficial de facto standards via the support of large corporations and groups. Many thousands of independent companies, public and private organizations, and government, collaborate in all the above activities.

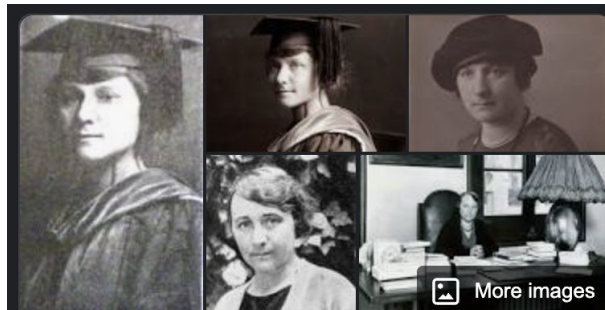
A landmark moment on the history of the Internet and the Web was the appearance of online advertising and marketing. This sector grew at tremendous rate, starting from a simple advertising banner for AT&T data services in Hot Wired in '94 [14], to becoming an entire industry that has overtaken broadcast and print advertising [49], and is currently funding a large part of the so-called free services of the Internet [12]. Of course, with online advertising, came user tracking and all the data protection and privacy problems that have challenged the Internet and the Web during their online advertising era [43].

2 TECHNOLOGICAL CONSEQUENCES

Like many other things, the Internet and the Web were not designed to support online advertising, especially advanced versions of it, such as targeted (or behavioral) advertising [9, 37], in which marketing offers are optimized on a per user basis based on complex recommendation and auctioning algorithms driven by detailed personal data collected online from millions of individuals.

2.1 Advertising sector

The digital advertising industry has built an immensely complex business ecosystem [47] and set of tools, protocols, and layers for empowering user tracking and advertising on a massive scale. These include third party tracking cookies and complex digital fingerprinting techniques [29], the Real Time Bidding (RTB) [7] and



Maria de Maeztu (María de Maeztu)

Spanish author

María de Maeztu Whitney was a Spanish educator, feminist, founder of the Residencia de Señoritas and the Lyceum Club in Madrid. She was sister of the writer, journalist and occasional diplomat, Ramiro de Maeztu and the painter Gustavo de Maeztu.

[Wikipedia](#)

Born: July 18, 1881, Vitoria-Gasteiz

Died: January 7, 1948, Mar del Plata, Argentina

Parents: Manuel de Maeztu Rodriguez

Education: Philipps-University Marburg

Siblings: Ramiro de Maeztu

Organization founded: Lyceum Club Femenino

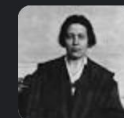
Nephew: Juan Manuel de Maeztu y Hill

People also search for

[View 5+ more](#)



Ramiro de Maeztu



Victoria Kent

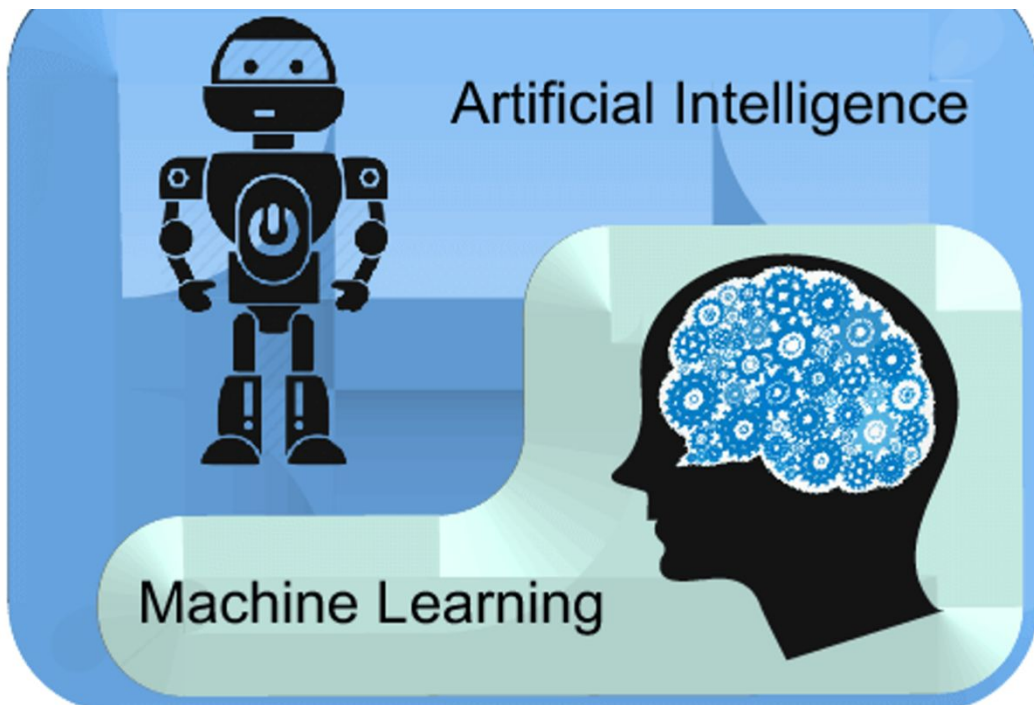


Clara Campoa...



Zenobia Camprubí

Seed 1: AI needs data



nature medicine ARTICLES
<https://doi.org/10.1038/A41591-022-01811-6>
 Check for updates

OPEN
Machine learning model to predict mental health crises from electronic health records

Roger Garriga^{1,2,3*}, Javier Mas^{1,3}, Semhar Abraha^{4,5}, Jon Nolan⁴, Oliver Harrison⁶, George Tadros^{4,6} and Aleksandar Matic^{1,2,3}

The timely identification of patients who are at risk of a mental health crisis can lead to improved outcomes and to the mitigation of burdens and costs. However, the high prevalence of mental health problems means that the manual review of complex patient records to make proactive care decisions is not feasible in practice. Therefore, we developed a machine learning model that uses electronic health records to continuously monitor patients for risk of a mental health crisis over a period of 28 days. The model achieves an area under the receiver operating characteristic curve of 0.797 and an area under the precision-recall curve of 0.159, predicting crises with a sensitivity of 58% at a specificity of 85%. A follow-up 6-month prospective study evaluated our algorithm's use in clinical practice and observed predictions to be clinically valuable in terms of either managing case-loads or mitigating the risk of crisis in 64% of cases. To our knowledge, this study is the first to continuously predict the risk of a wide range of mental health crises and to explore the added value of such predictions in clinical practice.

Nearly 1 billion people worldwide live with a mental disorder¹. With the global mental health emergency considerably exacerbated by the Coronavirus Disease 2019 pandemic, healthcare systems face a growing demand for mental health services coupled with a shortage of skilled personnel^{2,3}. In clinical practice, considerable demand arises from mental health crises—that is, situations in which patients can neither care for themselves nor function effectively in the community and situations in which patients may hurt themselves or others⁴. Timely treatment can prevent exacerbating the symptoms that lead to such crises and subsequent hospitalization⁵. However, patients are frequently already experiencing a mental health crisis when they access urgent care pathways as their primary entry point to a hospital or psychiatric facility. By this point, it is too late to apply preventative strategies, limiting the ability of psychiatric services to properly allocate their limited resources ahead of time. Therefore, identifying patients at risk of experiencing a crisis before its occurrence is central to improving patient outcomes and managing case-loads⁶.

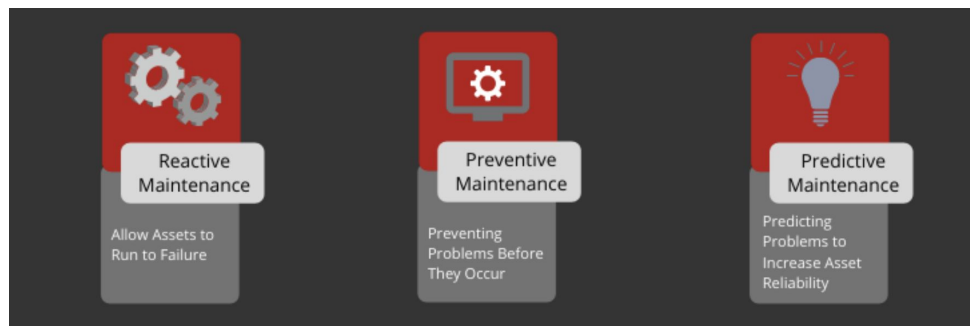
In busy clinical settings, the manual review of large quantities of data across many patients to make proactive care decisions is impractical, unsustainable and error-prone⁷. Thus, shifting such tasks to the automated analysis of electronic health records (EHRs) holds great promise to revolutionize health services by enabling large-scale continuous data review. Research has already demonstrated the feasibility of predicting critical events associated with a wide range of healthcare problems, including hypertension, diabetes, circulatory failure, hospital readmission and in-hospital death^{8–10}. However, the mental health literature is limited to predicting specific types of events—such as suicide, self-harm and first episode psychosis^{11–13}—rather than continuously predicting the breadth of mental health crises that require urgent care or hospitalization. Much remains unknown about the feasibility of querying machine learning models continuously to estimate the risk of an imminent mental health crisis. This would enable optimising healthcare staff allocation and preventing crisis onset. Furthermore, even a highly accurate predictive model does not guarantee improved mental health outcomes or long-term cost savings¹⁴; therefore, it remains unclear whether new predictive technologies could provide tools that are useful to mental healthcare practitioners¹⁵.

This research explores the feasibility of predicting any mental health crisis event, regardless of its cause or the underlying mental disorder, and we investigate whether such predictions can provide added value to clinical practice. The underpinning assumption is that there are historical patterns that predict future mental health crises and that such patterns can be identified in real-world EHR data, despite its sparseness, noise, errors and systematic bias¹⁶. To this end, we developed a mental crisis risk model by inputting EHR data collected over 7 years (2012–2018) from 17,122 patients into a machine learning algorithm. We evaluated how accurately the model continuously predicted the risk of a mental health crisis within the next 28 days from an arbitrary point in time, with a view to supporting dynamic care decisions in clinical practice. We also analyzed how the model's performance varied across a range of mental health disorders, across different ethnic, age and gender groups and across variations in data availability. Furthermore, we conducted a prospective cohort study to evaluate the crisis prediction algorithm in clinical practice from 26 November 2018 to 12 May 2019. The crisis predictions were delivered on a biweekly basis to four different groups of clinicians (in total, 60 clinicians attending 1,011 cases over 6 months), who evaluated whether and how such predictions helped them manage case-load priorities and mitigate the risk of crisis.

Results
Prediction target. As our main goal was to develop a predictive tool that could help healthcare workers manage case-load priorities and pre-emptively intervene to mitigate the risk of crisis, we established the prediction target to align with the service-oriented approach to defining crisis—that is, the onset of severe symptoms that require substantial healthcare resources. Notwithstanding a wide range

¹Koia Health, Barcelona, Spain. ²Universitat Pompeu Fabra, Department of Information and Communication Technologies, Barcelona, Spain. ³Accanet, Barcelona, Spain. ⁴Birmingham and Solihull Mental Health NHS Foundation Trust, Birmingham, UK. ⁵University of Warwick, Warwick, UK. ⁶Aston Medical School, Aston University, Aston, UK. *e-mail: roger.garriga@upf.edu; javier.mas@koiahealth.com; garriga77@gmail.com; aleksandar.matic@koiahealth.com

NATURE MEDICINE | www.nature.com/naturemedicine



BUT

owners of knowledge \neq owners of data

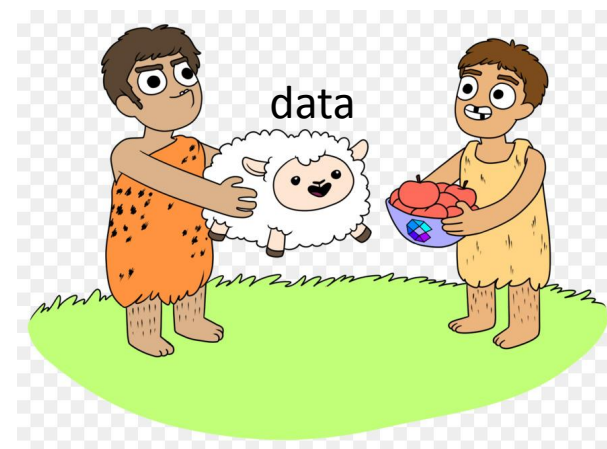
Seed 2: Most privacy problems of the web are due to its broken economics model



- Free data
- For free service
 - payment “in kind”
 - no cash

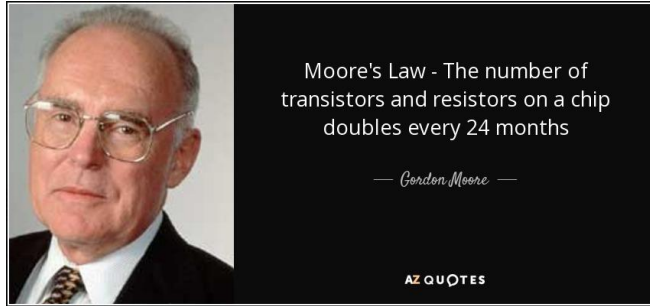
You

Online Service

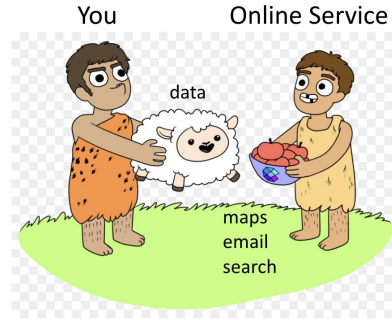


BAD for privacy!

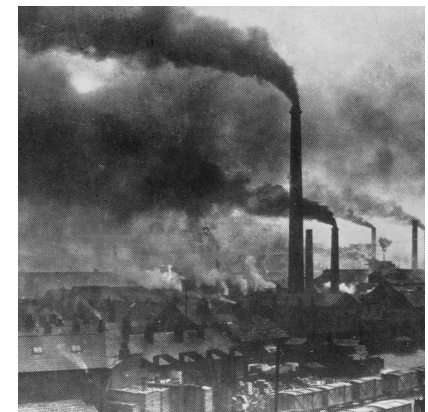
(tracking is cheap)



(no payments to users)



(collect ... everything)



BAD for sustainability!



Seed 3: Find a new exciting problem for networking

11th Annual IMDEA Networks Workshop
Networking Research: Present, Future and Beyond
2019 | Madrid, Spain



HOME PROGRAM ATTENDEES INVITED SPEAKERS WORKSHOP CHAIR

Home

11th Annual IMDEA Networks Workshop Networking Research: Present, Future and Beyond

In this workshop a group of renowned experts will deliver keynotes and participate in panels to assess where networking research stands today, and more important, where it is heading in the future. We will discuss developments in both traditional networking application domains such (SDN, 5G, and IoT) and new tools for them (Machine Learning, Blockchain) as well as new direction going beyond traditional networking (data protection, algorithmic discrimination & bias, news & (dis)information spread).

The workshop will also include invited talks covering the broader environment in which networking research exists, including business trends and development in telcos and ICT services in general, policy issues, as well as innovation and tech transfer.



WORKSHOP INFORMATION

PAST WORKSHOPS

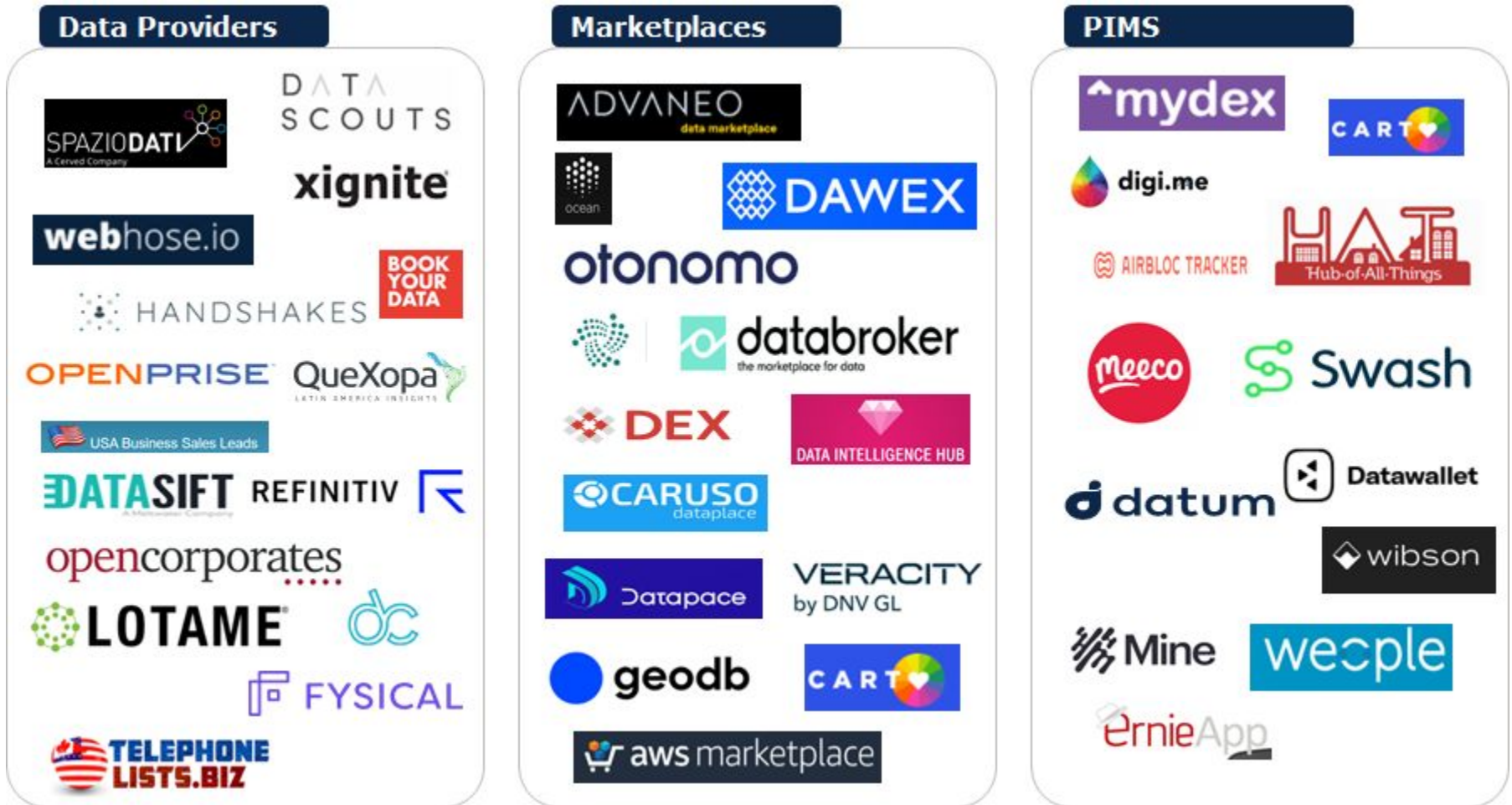
VENUE

LOCAL INFORMATION

CONTACT

Others of course have taken notice

Entities trading with data



Aggregate Data (B2B) for training ML algorithms

Personal Data

We checked more than 190 companies offering data products and services in order to understand how data is traded nowadays¹



¹S. Andrés Azcoitia, N. Laoutaris. “[A Survey of Data Marketplaces and Their Business Models](#)”, SIGMOD Record Sept’22

Data is an elusive “commodity”

- Data is the new oil
 - But oil cannot be copied/transmitted/processed at 0 cost
- Data vs. digital content
 - Authorship is not obvious
 - Top Gun 2 ⇒ Tom Cruise
 - But who’s the “author” of a list of URLs?
- It is not really a commodity
 - 1 litre of gasoline gives the same mileage
 - But the same data:
 - Vastly different value for different services
 - Even for the same service (different intensity of use)



Challenge 1: Pricing



A quiz

How valuable is this?



How about this?



A quiz

And this?

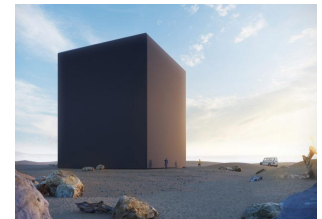
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States

Challenge 2: Federation



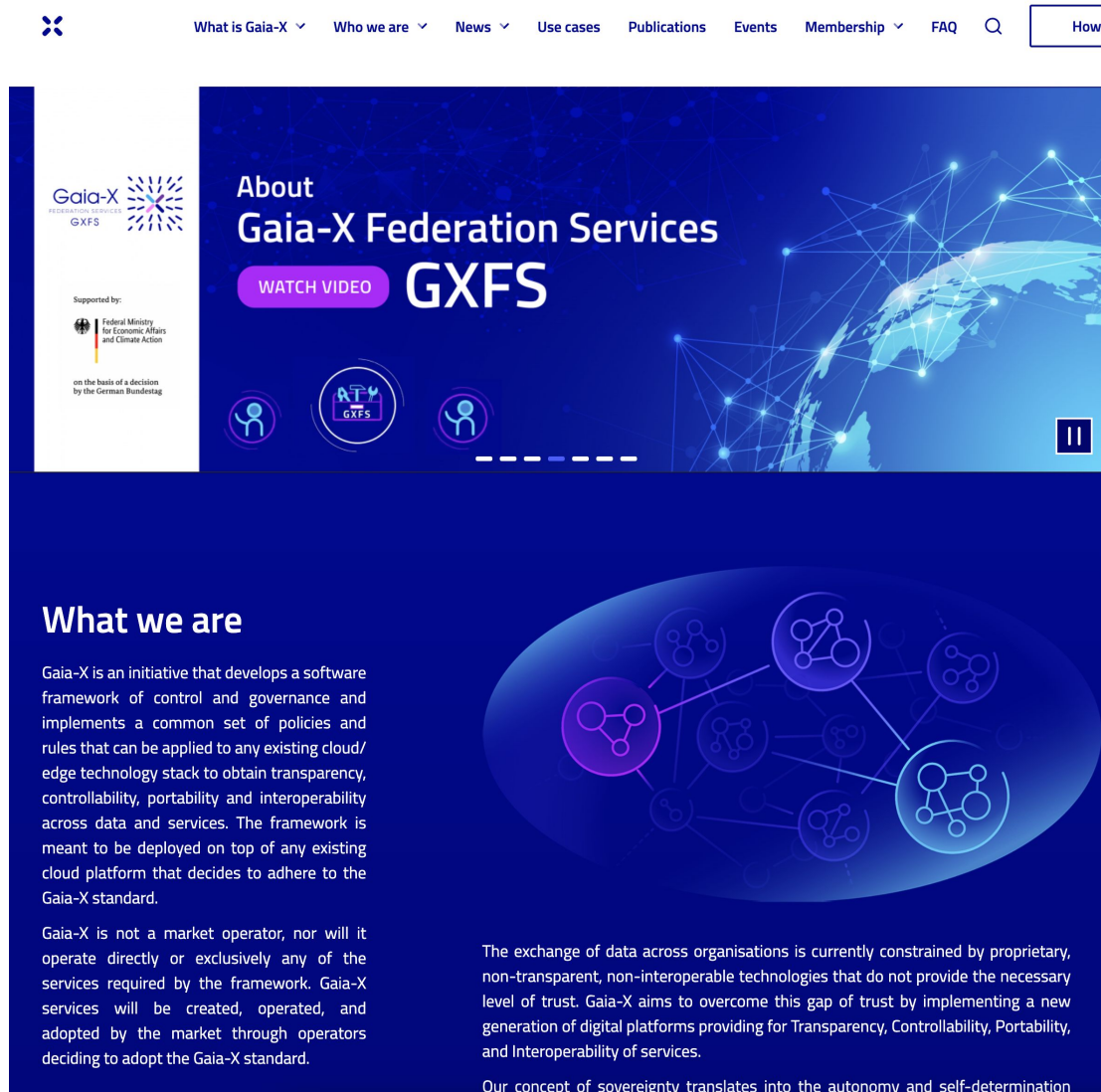
Fragmentation vs. Monopoly

We checked more than 190 companies offering data products and services in order to understand how data is traded



EU governments have taken notice

Industrial/standardisation activity



[What is Gaia-X](#) [Who we are](#) [News](#) [Use cases](#) [Publications](#) [Events](#) [Membership](#) [FAQ](#) [How](#)

About Gaia-X Federation Services GXFS

[WATCH VIDEO](#)

Supported by:
Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

What we are

Gaia-X is an initiative that develops a software framework of control and governance and implements a common set of policies and rules that can be applied to any existing cloud/edge technology stack to obtain transparency, controllability, portability and interoperability across data and services. The framework is meant to be deployed on top of any existing cloud platform that decides to adhere to the Gaia-X standard.

Gaia-X is not a market operator, nor will it operate directly or exclusively any of the services required by the framework. Gaia-X services will be created, operated, and adopted by the market through operators deciding to adopt the Gaia-X standard.

The exchange of data across organisations is currently constrained by proprietary, non-transparent, non-interoperable technologies that do not provide the necessary level of trust. Gaia-X aims to overcome this gap of trust by implementing a new generation of digital platforms providing for Transparency, Controllability, Portability, and Interoperability of services.

Our concept of sovereignty translates into the autonomy and self-determination

Industrial/standardisation activity

WHY WE MAKE USE ADOPT PUBLICATIONS | NEWS, BLOG, EVENTS

INTERNATIONAL DATA SPACES

The future of the data economy is here

What we do and why it matters

The International Data Spaces Association (IDSA) is on a mission to create the future of the global, digital economy with International Data Spaces (IDS), a secure, sovereign system of data sharing in which all participants can realize the full value of their data.

IDS enables new “smart services” and innovative business processes to work across companies and industries while ensuring that the self-determined control of data use (data sovereignty) remains in the hands of data providers.

[MORE ABOUT US →](#)

133 MEMBERS

8 HUBS

22 COUNTRIES

A TRUSTWORTHY ARCHITECTURE FOR THE DATA ECONOMY

The IDS provides self-determined control between all imaginable data endpoints

Broker

IDS Connectors register the description of their data endpoints at an IDS Broker. This potential Data Consumers can look up available data sources and data in terms of their content, structure quality, actuality and other attributes.

App Stores

App Stores provide Data Apps, i.e. applications that can be deployed in IDS Connectors to execute tasks like transformation, aggregation or analytics on the data. Data Apps may be certified by IDS-approved certification bodies. App Stores can be provided by IDS members and must, themselves, be certified under IDS standards.

Data Provider

Data Providers give access to data under specific usage and price models. They are able to control the access to data and the usage of data at the Data Consumer.

Identity Provider

Identity Providers offer a range of services to create, maintain, manage and validate identity information of and for IDS participants, regardless under which of the above roles they consider to take part. Proven identity of all participants in the IDS represent an essential imperative to the IDS architecture.

Data - the economic asset in data driven business models

The key focus for any data-driven economy is on exploiting and sharing data from various data sources.

INTERNATIONAL DATA SPACES APPROACH



Endless Connectivity
Standard for data flows between all kinds of data endpoints



Trust between different security domains
Comprehensive and audit-proof security functions providing a maximum level of trust



Governance for the data economy
Usage control and enforcement for data flows and assignments of data

MISSION STATEMENT

Secure Data Exchange
It forms the basis for a variety of certifiable software solutions, smart services...



International Standards
IDS standard is the basic conditions and governance for a reference architecture and interfaces

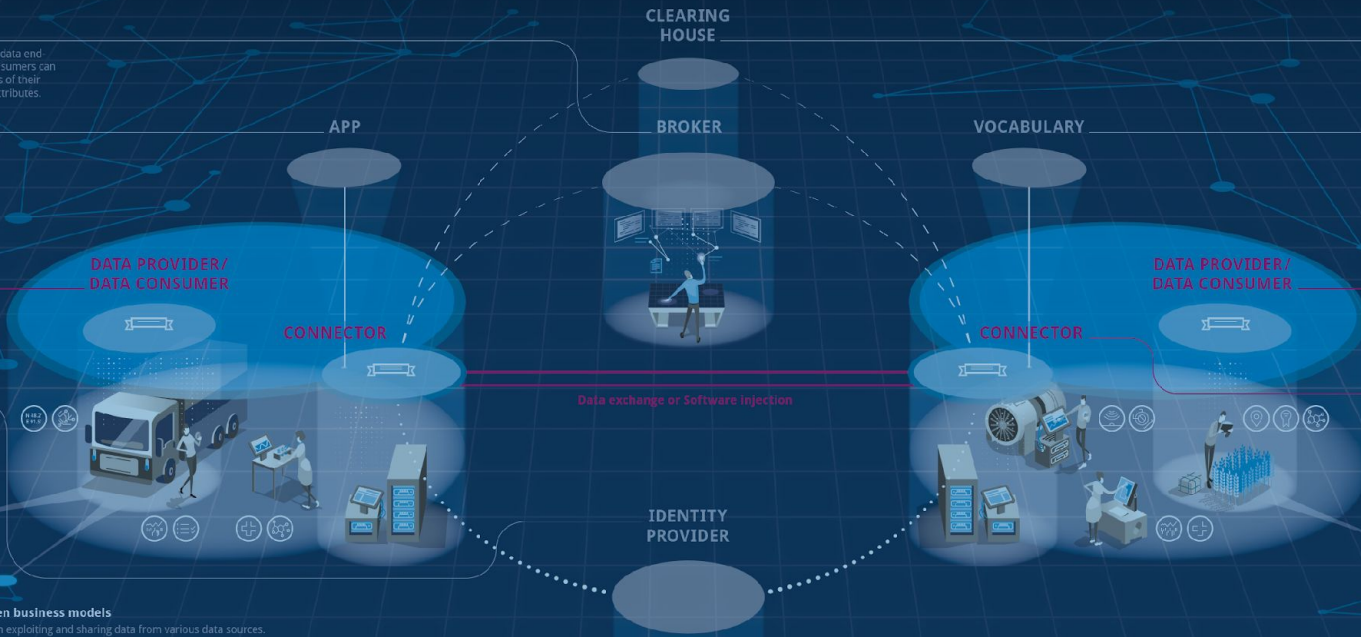
Business Models
Data Owners remain sovereign owners of their data at any time



Use Cases
This standard is actively developed and updated on the basis of use cases

DIGITAL IDENTITIES

A network of trusted entities in the data economy requires a mechanism for digital identities, that can reliably identify a participant and can provide more information on transaction partners. Additional information must be updated regularly and be provided in a trusted manner.



Clearing House

Intermediary providing clearing and settlement services for all financial and data exchange transactions within the IDS.

Vocabulary

Vocabulary Providers manage and offer vocabularies (ontologies, reference data models, metadata elements) which can be used to annotate and describe datasets. Vocabulary Providers provide (domain specific) vocabularies and their reference to the IDS Information Model, which is the basis for the description of data sources.

Data Consumer

Data Consumers can search for data and use the data of different data providers. Data Consumers are bound to the usage policy of the Data Provider.

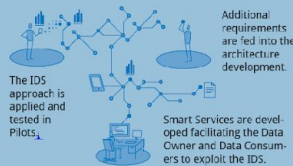
Connector

IDS Connectors provide standardized connectivity in the IDS Ecosystem. Connectors are responsible for connectivity and Usage Control. They allow the execution of trusted apps in an isolated Identity Provider environment.



USE CASES

Services and functionalities of the IDS are specified and validated in use cases.



Smart Services are developed facilitating the Data Owner and Data Consumers to exploit the IDS.

COMMUNITIES

Interest and user groups of same or similar domains with common challenges validate and proliferate the IDS approach, technology and eco-system. Based on their practical experience the IDS reference architecture and the eco-system around it are continuously developed. Thus, specific application scenarios for verticals are set up, implemented and systematically pushed forward, allowing participants to enhance existing or to launch new services.

- Medical Health
- Energy
- Materials
- Banking/Insurances
- Industrial
- Farm & Food
- Smart Cities
- Logistics

10 THINGS TO KNOW ABOUT

- Containerization, e.g. Docker
- WebServices, e.g. https, MQTT, REST, Multi Part Messages
- Enterprise Integration Patterns
- Software Engineering, e.g. tools like Maven, git
- Message Oriented Middleware
- Digital Identities and Digital Certificates, e.g. X509
- Semantic Data Descriptions, e.g. Resource Description Framework
- Requirements Engineering, Processes and tools, e.g. UML and BPMN
- Data Ecosystems
- Certification, e.g. IEC 62443, ISO 27001

CERTIFICATION APPROACH

The IDS Certification Body is appointed by the IDSA and regularly aligns with the IDSA to manage the certification process, defines the standardized evaluation procedures and supervises the actions of the Evaluation Facilities. An Evaluation Facility is contracted by an Applicant and is responsible for carrying out the detailed technical and organizational evaluation work during a certification.



CALL TO ACTION

Become a member in the International Data Spaces Association:



A case of ATM vs. IP?

The Sky Above The Clouds

A Berkeley View on the Future of Cloud Computing

Sarah Chasins, Alvin Cheung, Natacha Crooks, Ali Ghodsi, Ken Goldberg, Joseph E. Gonzalez, Joseph M. Hellerstein, Michael I. Jordan, Anthony D. Joseph, Michael W. Mahoney, Aditya Parameswaran, David Patterson, Raluca Ada Popa, Koushik Sen, Scott Shenker, Dawn Song, Ion Stoica

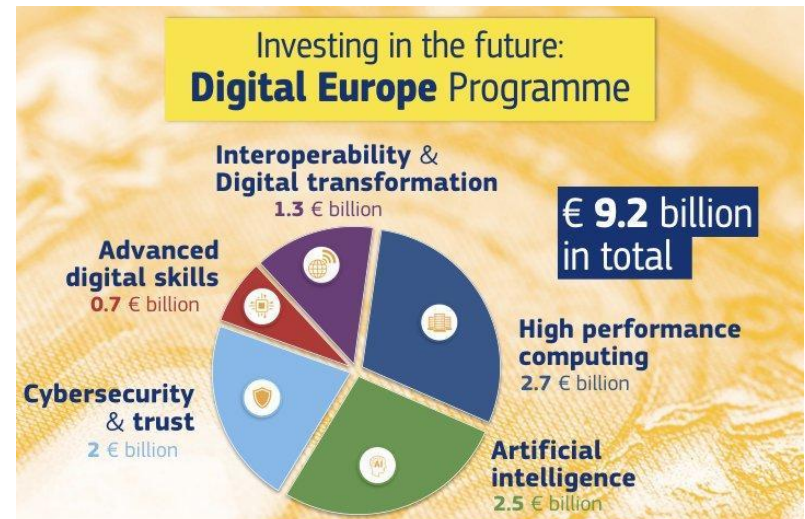
Executive Summary: *Technology ecosystems often undergo significant transformations as they mature. For example, telephony, the Internet, and PCs all started with a single provider, but in the United States each is now served by a competitive market that uses comprehensive and universal technology standards to provide compatibility. This white paper presents our view on how the cloud ecosystem, barely over fifteen years old, could evolve as it matures.*

Each of the early entrants to the cloud computing market offered their own proprietary interfaces. We predict that this market, with the appropriate intermediation, could evolve into one with a far greater emphasis on compatibility, allowing customers to easily shift workloads between clouds. However, the cloud's trajectory towards this more compatible and competitive market will be quite different from the examples cited above. In particular, we believe that a comprehensive compatibility standard supported by all clouds—essential in creating the markets in telephony, the Internet, and PCs—is neither likely to emerge nor necessary to allow customers to move workloads between clouds, and would in fact impede innovation. Instead, we argue that to achieve this goal of flexible workload placement, cloud computing will require intermediation, provided by systems we call intercloud brokers, so that individual customers do not have to make choices about which clouds to use for which workloads, but can instead rely on brokers to optimize their desired criteria (e.g., price, performance, and/or execution location). We believe that the competitive forces unleashed by the existence of effective intercloud brokers will create a thriving market of cloud services with many of those services being offered by more than one cloud, and this will be sufficient to significantly increase workload portability.

Lot's of regulatory activity



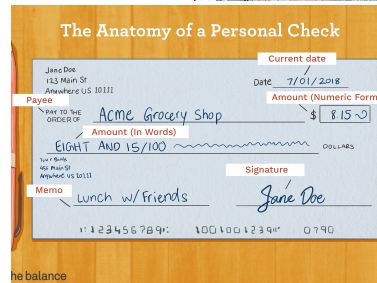
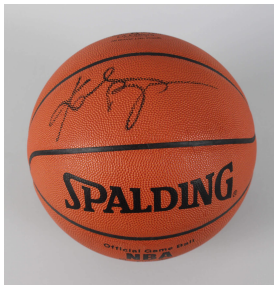
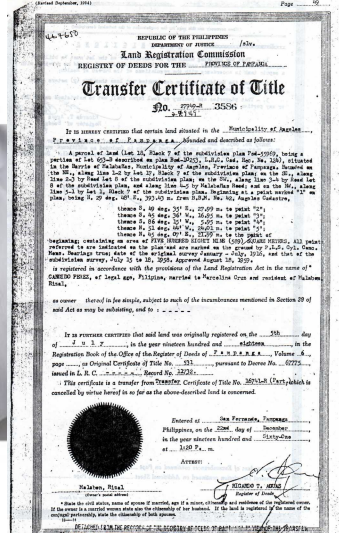
Lot's of funding opportunities



Challenge 3: Ownership



Physical ownership



Digital ownership



Microsoft
Office
Professional

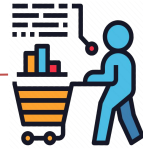




Dataset ownership: The Problem



Honest Seller



Honest Buyer



Malicious Buyer

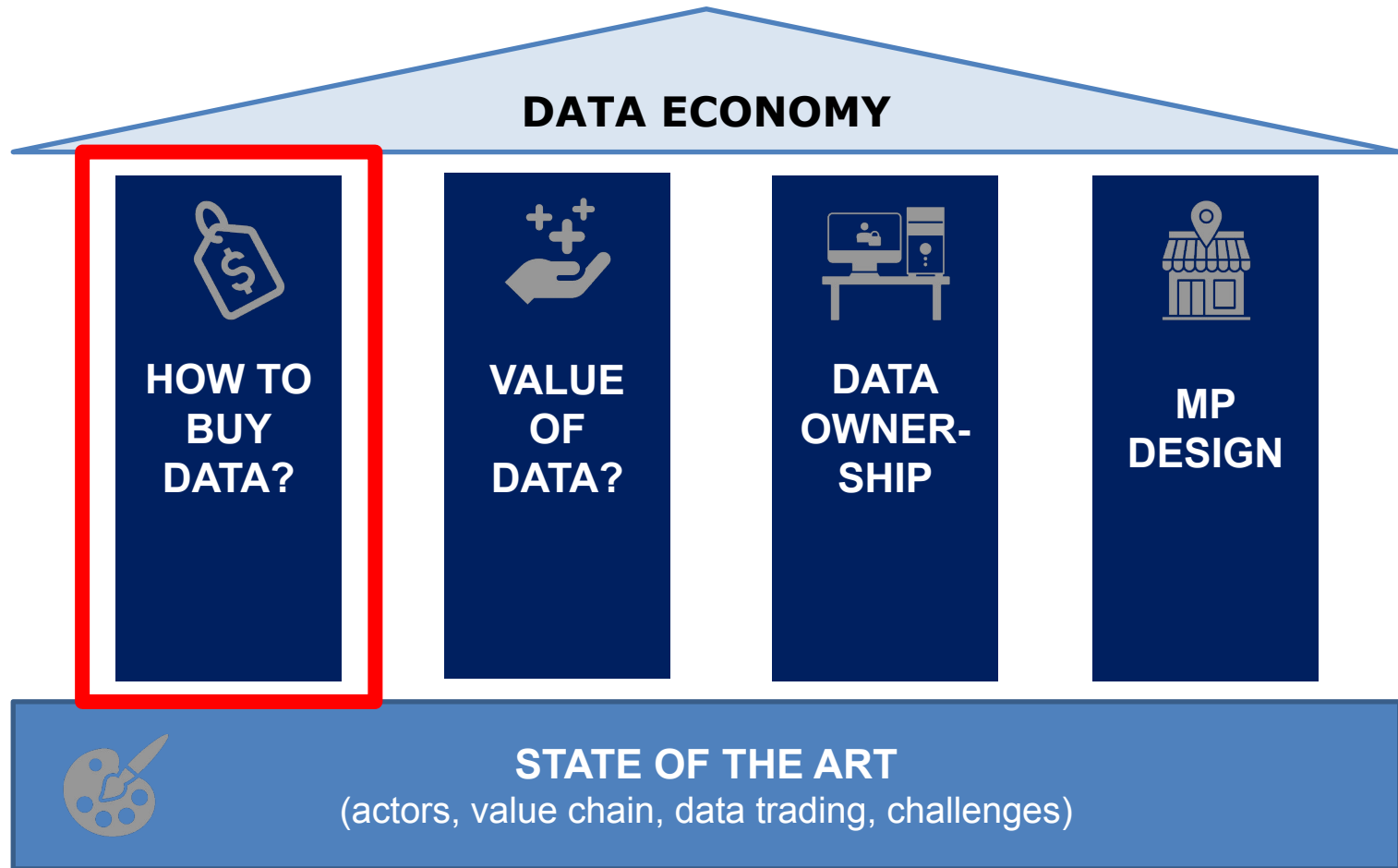


Dishonest Seller



- **The Malicious Buyer** distributes the data of the Honest Seller without her consent.
- Can re-appear as a **Dishonest Seller** (or pass it to somebody else for this purpose) and act like an Honest Seller.
- **How can the Honest Seller prove his original ownership of the data?**
- **How can the Honest Seller discover who the Malicious buyer was?**

What should we work on?





Buyer's problem

1. How useful is a dataset for me?

2. How does its value change when combined?

Automobiles

1996 FORD F250 XLT EXTENDED CAB 460. Auto, White, Excellent Condition, Runs Perfect, Only has two little rust spots. No body work. New Brakes, New Exhaust, Good Tires, Too Much To List. \$8700.00

1998 GMC YUKON 285,000 mi., \$1,800 414-254-0906

2001 HONDA CIVIC 182K, new a/c, alternator & tires. \$1,100 OBO. 262-534-5777.

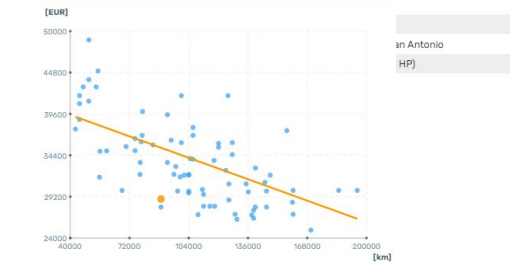


2015 Bmw X5 xDrive30d 28,900 EUR

BASIC INFORMATION

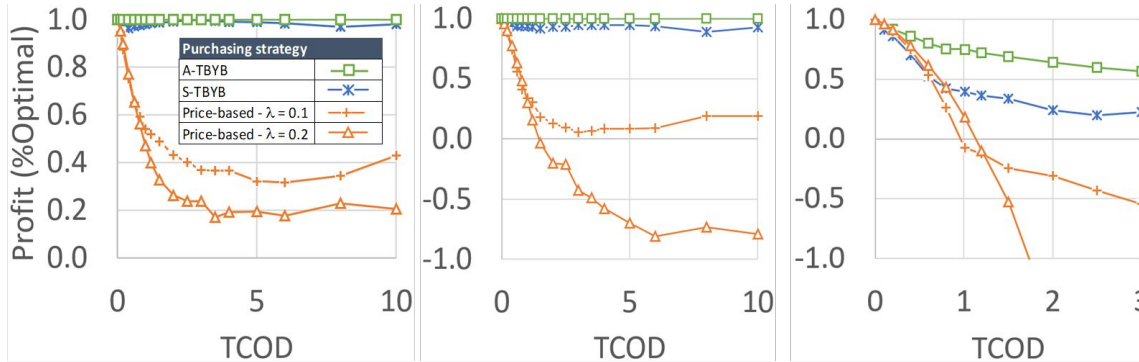
Price	28,900 EUR
Make	Bmw
Model	X5
Trim	xDrive30d
Mi	89,000 km
Year	2015

PRICE-MILEAGE DIAGRAM FOR BMW X5 2015





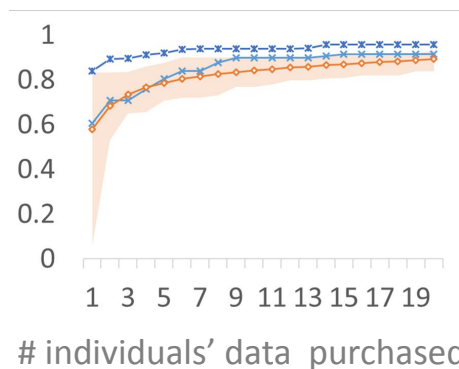
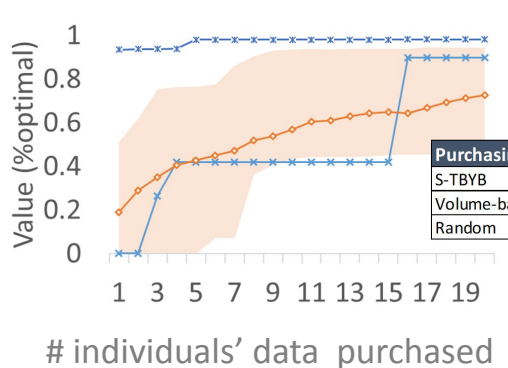
Buyer's problem



Just giving buyers' a hint of the accuracy of a dataset in their specific task significantly increases their profit in sourcing operations with data marketplaces, ...

Decreasing marginal utility

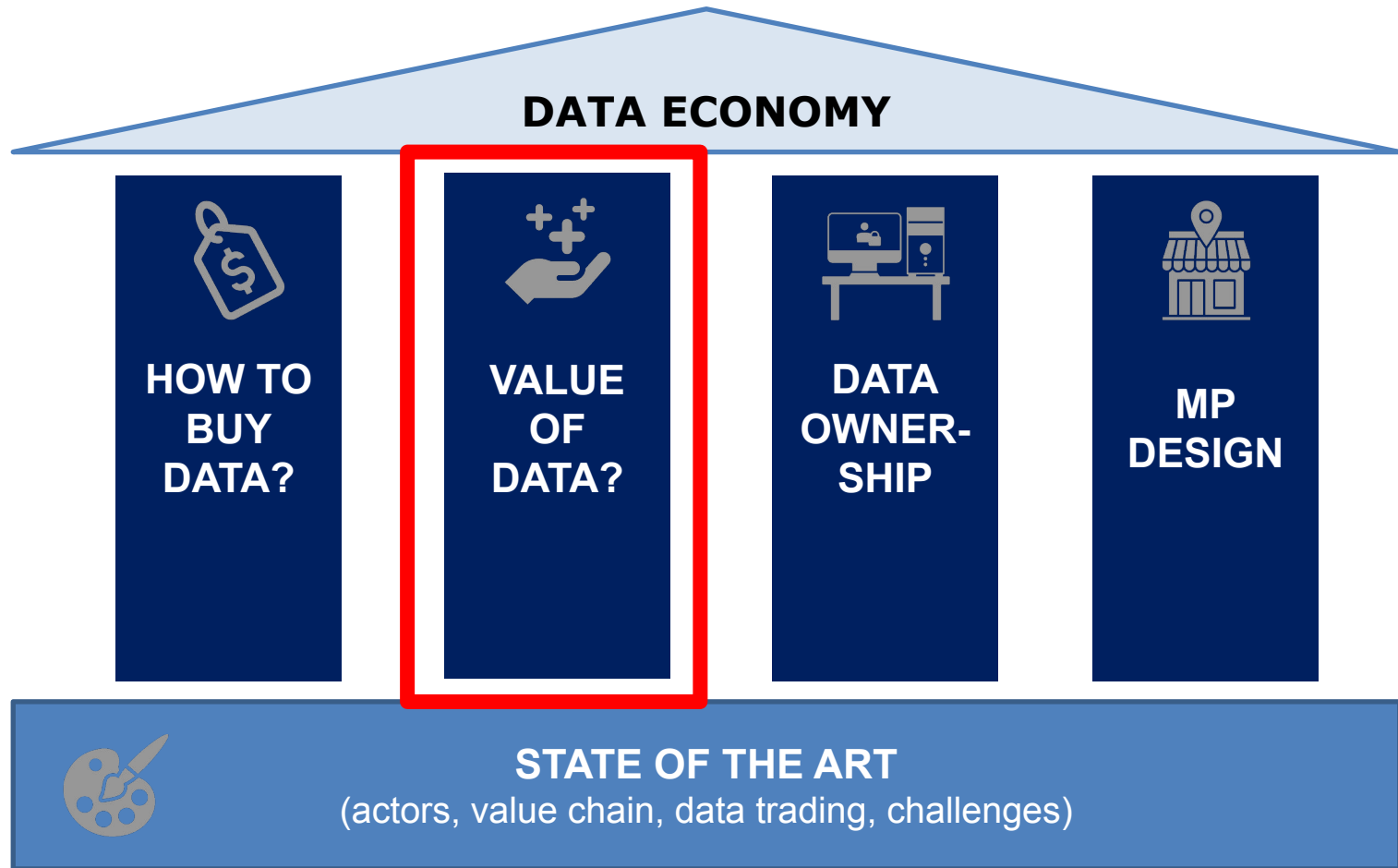
Increasing marginal utility



... and decreases the amount of information they need to buy to achieve their desired results, hence protecting privacy, as well

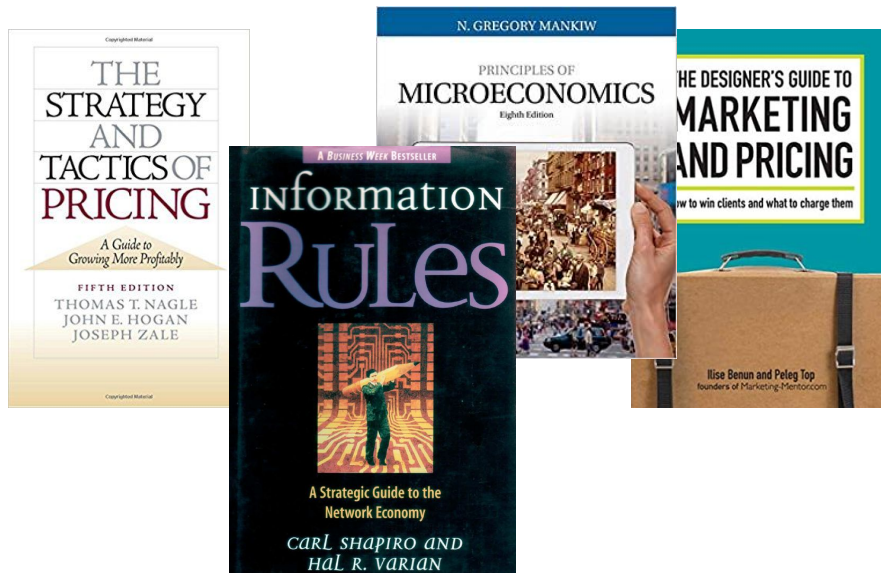
S. Andrés Azcoitia and N. Laoutaris. 2022. "[Try before you buy: a practical data purchasing algorithm for real-world data marketplaces](#)". In Proc. of the **1st ACM Workshop on the Data Economy (DE '22)**.

What should we work on?



Seller's problem

“Tech changes, economic laws do not” (Shapiro & Varian 98)...



Market study
Price-elasticity
Competition-level

Psychology
Value-based pricing
Dynamic pricing

... and pricing data brings additional problems

1. Value strongly depends on the buyer
2. Data maybe be easily replicated
3. Different Buyers ... different intensity of use

Uses it once



Repeated use

Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org
IMDEA Networks Institute
Leganés, Spain
Univ. Carlos III de Madrid
Leganés, Spain

Marius Paraschiv
marius.paraschiv@imdea.org
IMDEA Networks Institute
Leganés, Spain

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org
IMDEA Networks Institute
Leganés, Spain

Understanding the Price of Data in Commercial Data Marketplaces

1st Santiago Andrés Azcoitia
IMDEA Networks Institute
Universidad Carlos III de Madrid
Leganés, Spain
santiago.azcoitia@imdea.org

2nd Costas Iordanou
Cyprus University of Technology
Limassol, Cyprus
kostas.iordanou@cut.ac.cy

3rd Nikolaos Laoutaris
IMDEA Networks Institute
Leganés, Spain
nikolaos.laoutaris@imdea.org

ABSTRACT

Spatio-temporal information is used for driving a plethora of intelligent transportation, smart-city and crowd-sensing applications. Data is now a valuable production factor and data marketplaces have appeared to help individuals and enterprises bring it to market and the ever-growing demand. Such marketplaces are able to combine data from different sources to meet the requirements of different applications. In this paper we study the problem of estimating the relative value of spatio-temporal datasets combined in marketplaces for predicting transportation demand and travel time in metropolitan areas. Using large datasets of taxi rides from Chicago, Porto and New York we show that simplistic but popular approaches for estimating the relative value of data, such as splitting it equally among the data sources, more complex ones based on volume or the “leave-one-out” heuristic, are inaccurate. Instead, more complex notions of value from economics and game-theory, such as the Shapley value, need to be employed if one wishes to capture the complex effects of mixing different datasets on the accuracy of forecasting algorithms. This does not seem to be a coincidental observation related to a particular use case but rather a general trend across different use cases with different objective functions.

CCS CONCEPTS

• Information systems → Geographic information systems; • Human-centered computing → Ubiquitous and mobile computing.

KEYWORDS

Shapley Value, Value of Information, Spatio-Temporal Data, Data Marketplaces

ACM Reference Format:

Santiago Andrés Azcoitia, Marius Paraschiv, and Nikolaos Laoutaris. 2022. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)*, November 1–4, 2022, Seattle, WA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3557915.3561470>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9529-8/22/11.
<https://doi.org/10.1145/3557915.3561470>

1 INTRODUCTION

Data-driven decision making is bringing significant improvements to many sectors of the economy, including in several applications related to ubiquitous computing in the areas of transportation, mobility, and crowd-sensing. A solid body of research has studied matters of route optimization and city infrastructure planning [5, 14, 38], whereas companies are increasingly deploying and operating sophisticated systems for optimising their operations using live data. Such models and algorithms often require combining data from different sources and domains.

Data is by now considered a key production factor, comparable in importance to labour, capital, and infrastructure. Companies often need data from third parties, and for this they resort to data marketplaces. There are different types of marketplaces [8, 33, 34]. Personal Information Management Systems (aka PIMS, like Digi.me, Swash, or Meeco) allow individuals to sell their personal data, including their location, whereas general-purpose (AWS) and domain-specific marketplaces for spatio-temporal data allow companies to sell data to other companies, either integrated in already-existing services (HERE, CARTO, ESRI), as real-time streams (Streamr, IOTA, GeoDB), directly downloading datasets, or allowing access to them.

In almost all commercial marketplaces, pricing is left to sellers and buyers to agree. Sellers may set a fixed price, or let buyers bid for data [23], or even do a combination of the two. Such empirical pricing operates with minimal information, namely a high-level description of the dataset, including the number of data points it contains. The research community has already proposed different marketplace architectures to deal with AI/ML tasks, and industry-led initiatives aim to design trustworthy data spaces to share data [9, 17]. In a nutshell, the marketplace is able to train a model [2, 13] or to run code [1, 29] from potential buyers on data provided by sellers. It also ensures that data is accessed according to the terms agreed by both parties, that no data is leaked or replicated, that the intellectual property of buyers is protected, and that transactions and data usage are tracked and accounted.

Contributions: Our work looks at this open problem for the case of spatio-temporal data. In particular, we study how to compute the relative value of different spatio-temporal datasets used in i) forecasting future demand for a service across space and time, and ii) forecasting the travel time between two points A and B in a metropolitan area. Companies already offering service in overlapping areas can, for example, pool together their data to increase the accuracy of forecasting and its coverage. Improved forecasting can be used by the same companies to improve operations, such as dispatching vehicles, providing consumers with better information,

Abstract—A large number of Data Marketplaces (DMs) have appeared in the last few years to help owners monetize their data, and data buyers optimize their marketing campaigns, train their ML models, and facilitate other data-driven decision processes. In this paper, we present a first of its kind measurement study of the growing DM ecosystem, focused on understanding which features of data are actually driving their prices in the market. We show that data products listed in commercial DMs may cost from few to hundreds of thousands of US dollars. We analyze the prices of different categories of data and show that products about telecommunications, manufacturing, automotive, and gaming command the highest prices. We also develop classifiers for comparing data products across different DMs, as well as a regression analysis for revealing features that correlate with data product prices of specific categories, such as update rate or history for financial data, and volume and geographical scope for marketing data.

Index Terms—Data economy, data marketplaces, measurement, data pricing

I. INTRODUCTION

Data-driven decision making powered by Machine Learning (ML) algorithms is changing how the society and the economy work and is having a profound positive impact on our daily life. A McKinsey report predicted that data-driven decision-making could reach US\$2.5 trillion globally by 2025 [30], whereas a recent market study within the scope of the European Data Strategy estimates a size of 827 billion euro for the EU27 [14]. ML is driving up the demand for data in what has been called the fourth industrial revolution.

To satisfy this demand, several data marketplaces (DMs) have appeared in the last few years. DMs are mediation platforms that aim to connect data providers (acting as sellers) to data consumers (acting as potential buyers), and to manage data transactions between them. This ecosystem includes open data repositories [28], [33], general-purpose [2], [7], [18], [19], [21], and specialized or niche DMs targeting specific industries, such as automotive [13], [50], financial [8], [55], marketing [41], [42], and logistics [65], to name a few.

An issue of paramount importance is that of *data pricing*. Some marketplaces leave it to sellers to set a price for their

Our research has been supported by MLEDGE project (REGAGE22e0052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR, and by the European Union's HORIZON project DataBr-X (101070069).

data products. Many of them do not list prices of their products, but leave it to buyers and sellers to agree on a price after a negotiation. Due to the elusive nature of the traded “commodity”, pricing is a very complex matter, even more than in the case of material goods [53]. Unlike oil, to which it is often compared [17], data can be copied / transmitted / processed with close to zero cost. Even the use of the term commodity is a gross oversimplification of what data is. Notice that whereas two liters of gasoline yield a similar mileage on two similar cars under similar driving styles, nothing of this sort applies to data since 1) two datasets of equal volume may carry vastly different amounts of usable information, 2) the same information may have tremendously different value for Service A than for Service B, and 3) even if the per usage value of two services is the same, Service A may use the data 1,000 times more intensely than Service B leading to extremely different produced benefits. Some authors compared data to labor, too [6]. However, unlike labor, data is non-rivalrous meaning that its supply is not affected by its consumption, and thus selling data for a Service A does not prevent a provider from selling (a copy of) the same data for a Service B.

The research community at the intersection between computer science and economics has studied several aspects of data pricing. Still its elusive nature, and the complex business models under which it is made available makes it very hard to prescribe a price for data. Ultimately it is the market that decides and sets prices via complex mechanisms and feedback loops that are hard to capture. Despite some other works trying to measure the price of personal data of individuals [12], [43], [51], there is no systematic measurement study about the price of data products traded in commercial data marketplaces.

Our Contributions: In this paper we present what is, to the best of our knowledge, the first systematic measurement study of marketplaces for B2B data products. This ecosystem, despite being quite vibrant commercially, remains completely unknown to the scientific community. Very basic questions such as “What is the range of prices of data traded in modern DMs?”, “Which categories and types of data products command the highest prices?”, “Which are the features, if any, that correlate with the most expensive data products?” appear to have no answer and evade most meaningful speculations.

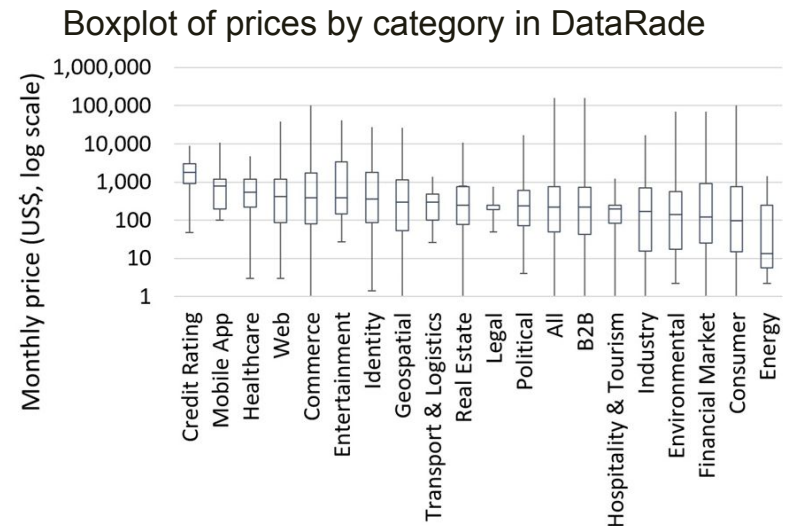
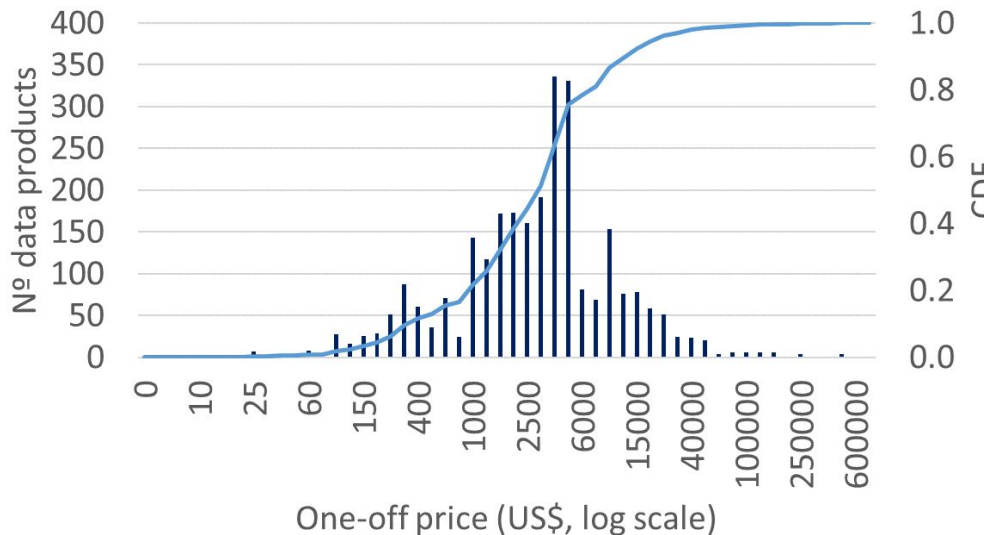
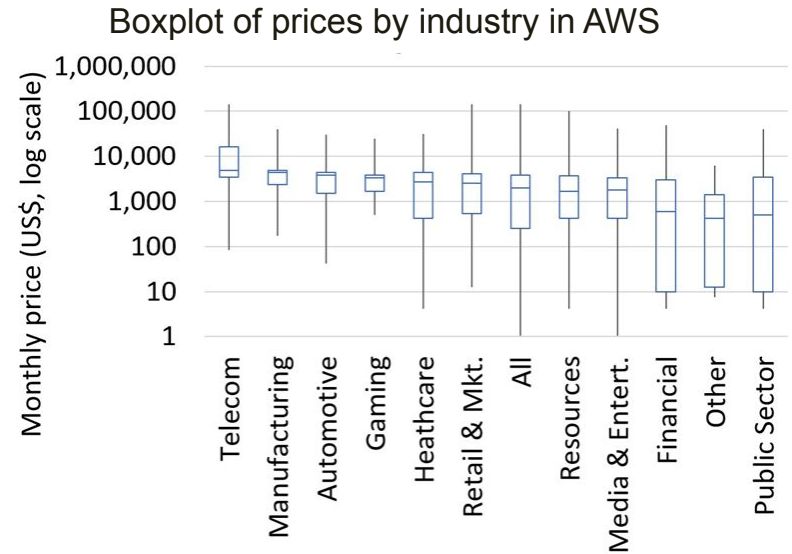
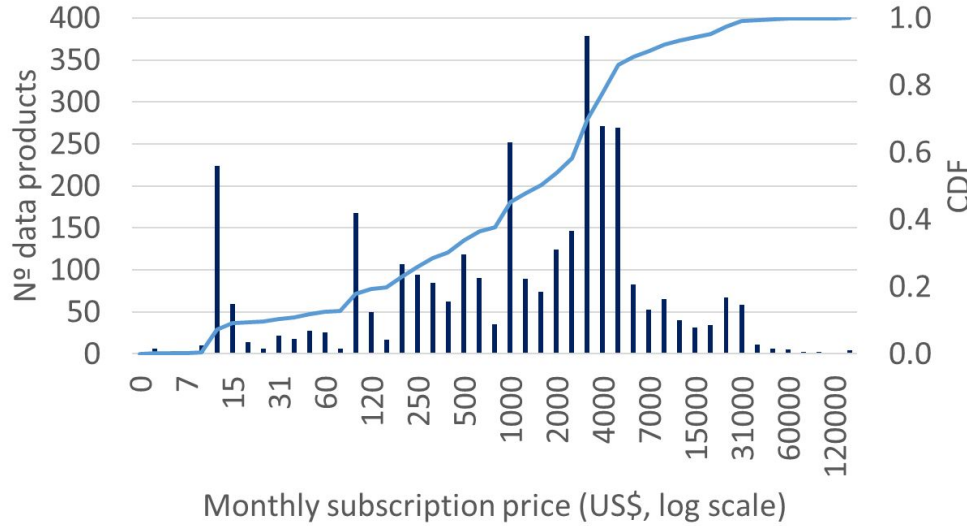
S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. 2022.

“Computing the relative value of spatio-temporal data in data marketplaces”. In Proc. of **ACM SIGSPATIAL'22**.

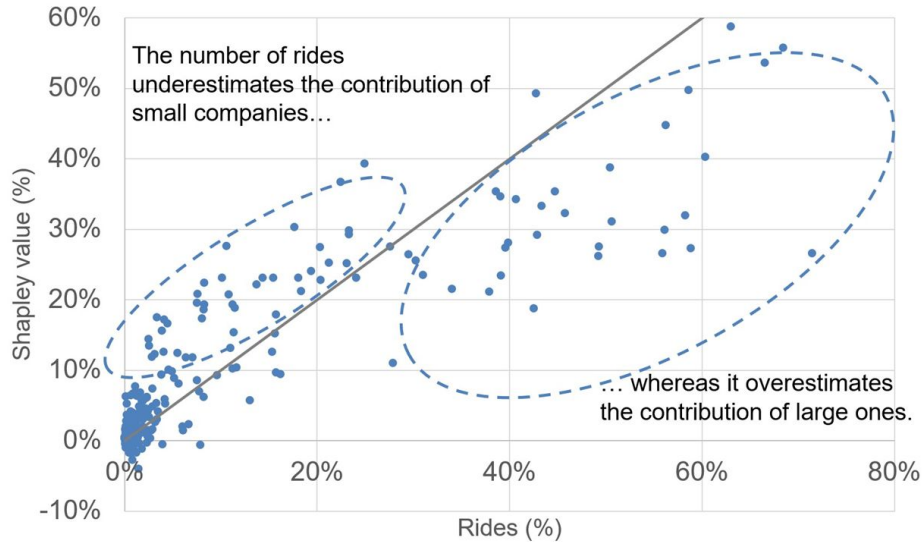
S. Andrés Azcoitia, C. Iordanou and Nikolaos Laoutaris.

2023. “Understanding the Price of Data in Commercial Data Marketplaces”. Accepted for publication in **IEEE ICDE'23**.

We found a wide range of prices for different data products, which depend on the category or industry, as well



... and that the price and the value of data is not necessarily proportional to its volume, but depends on 'what' kind of data it is and other factors

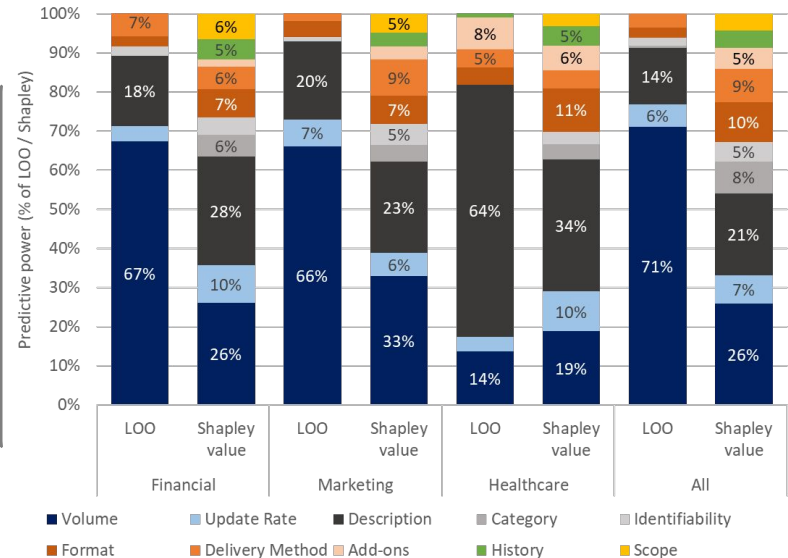


Shapley seems to be a "necessary evil" to capture the importance of data to a given ML task which simpler heuristics based on volume and LOO fail to approximate.

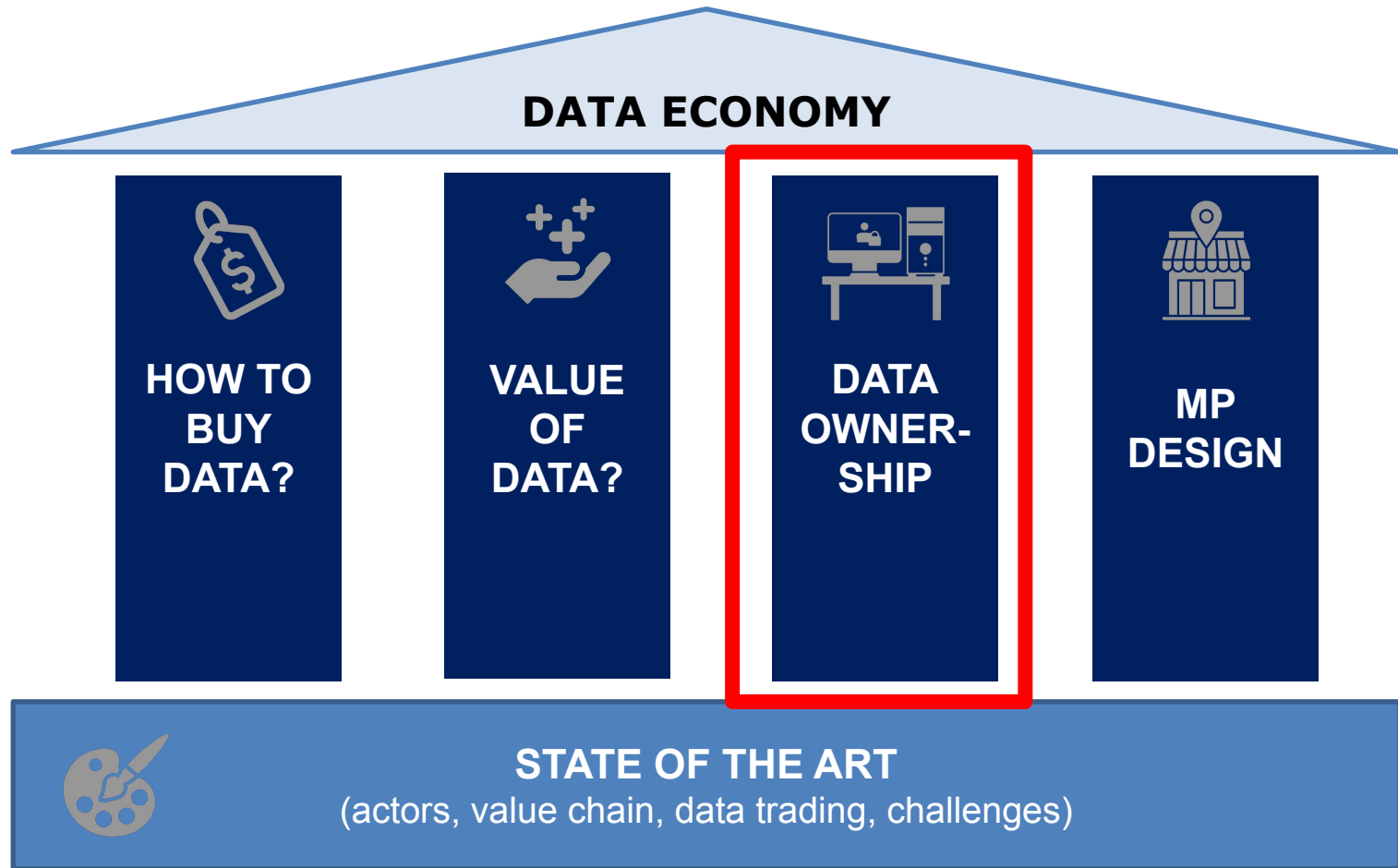
We EVEN found context-specific heuristics measuring valuable inherent features of data, such as its averageness or its diversity, which do better approximate Shapley values

We scraped 10 DMs and added information from other 30 providers to based on +200k products from +2k providers, and +4k price references.

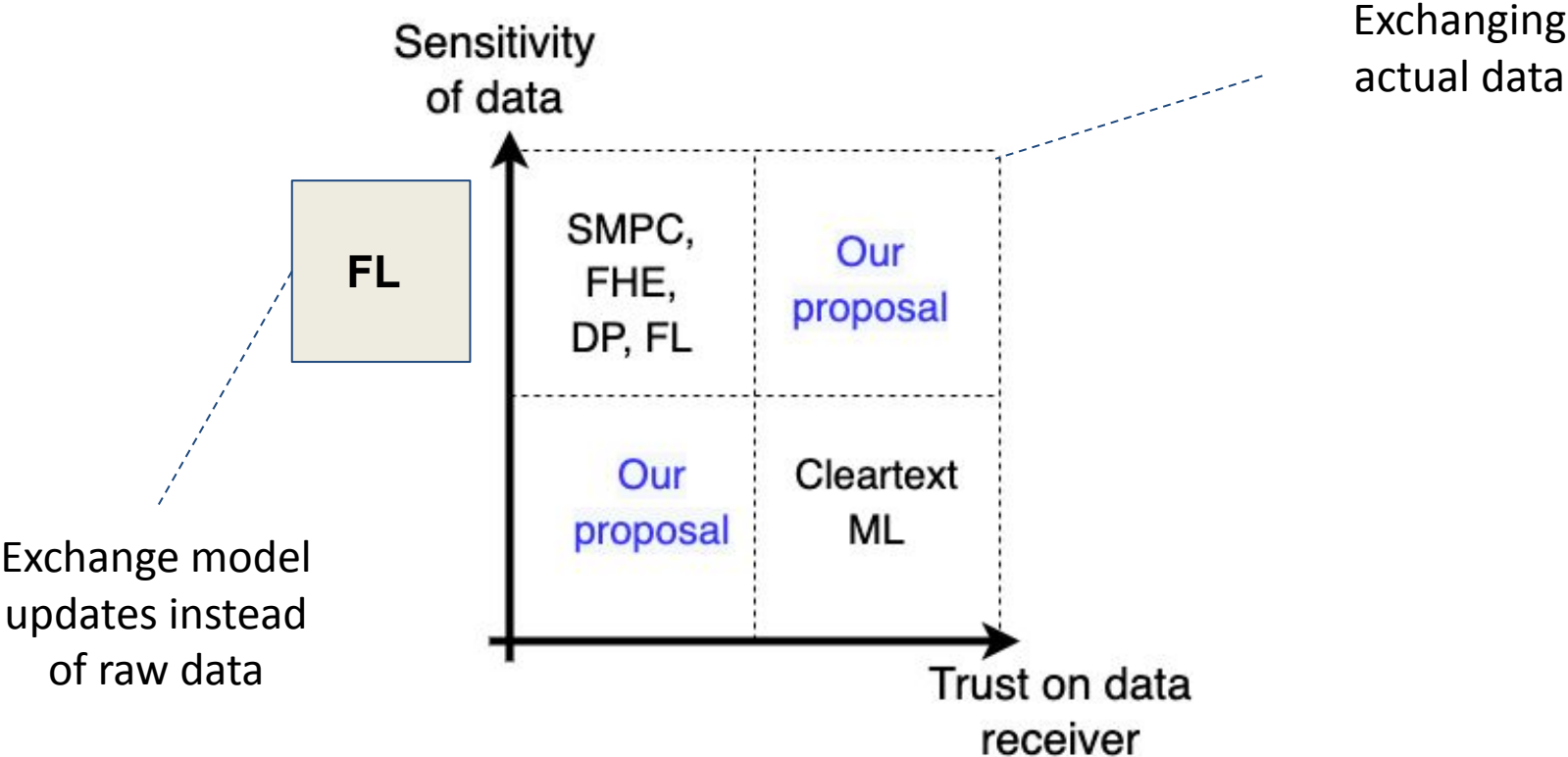
To understand the features driving the price of data , we managed to fit the prices of commercial products using regression models on top of their metadata features. We found that it is both 'volume' and specific characteristics of data products usually captured by their 'descriptions' that determine 66% of data prices



What should we work on?



Data ownership technology plane



Data ownership-Watermarking

- Watermarking is a process of : adding some distortion in a data content by using a secret (w/o destroying data utility).
- Resilient to **certain attacks** (e.g., guess, destroy)
- **Remove** watermark → Data utility **loss**

- **Traitor detection:** by generating a unique watermark per buyer (**fingerprinting**).

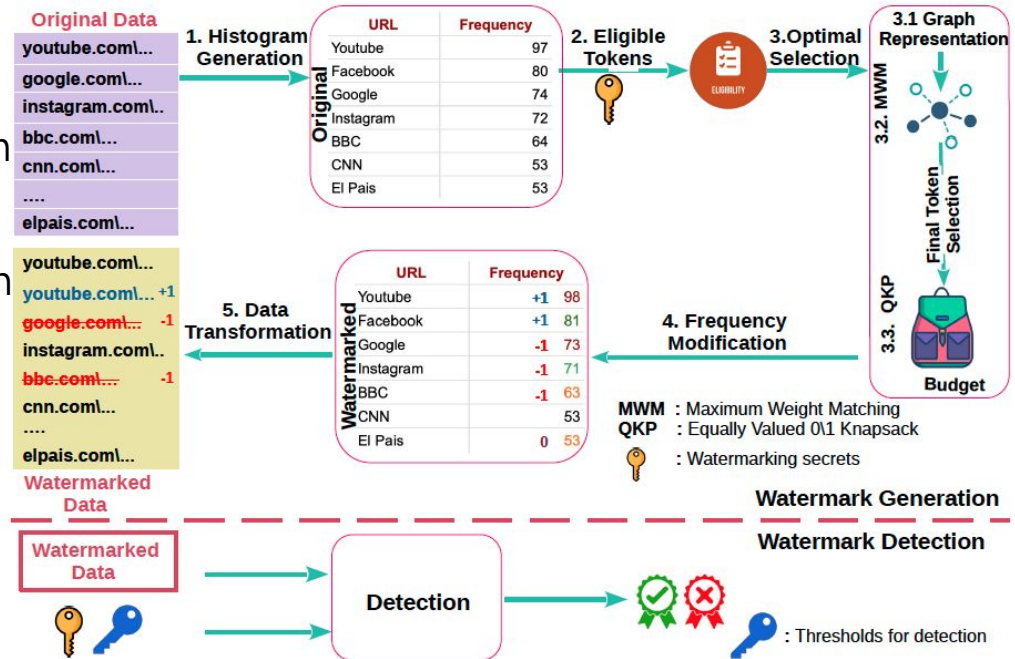
- Watermarking various types of assets:
 - Audio, images, video, text,...
 - Databases
 - ML models



Frequency-based (FreqY) Watermarking

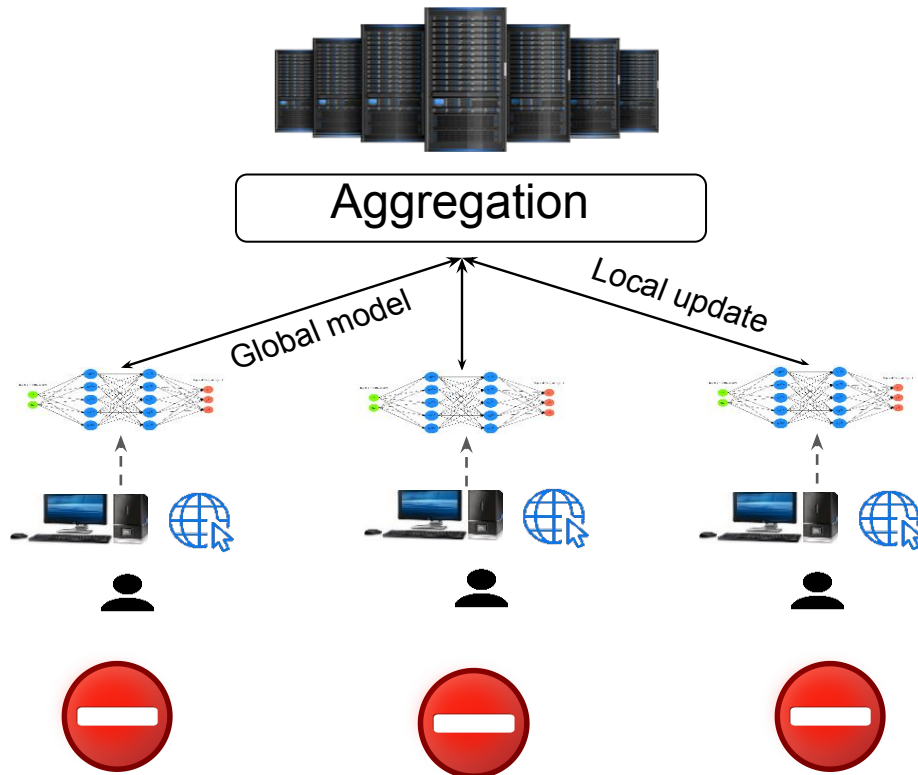
How would it work?

- **Dataset:** a dataset with a power-law distribution of its frequency/histogram (e.g., websites visited).
- Modify the frequencies by using a high entropy secret.
- Create a (new) histogram satisfying :
 - Preservation of **ranking**/order
 - **Budget:** minimum similarity to the original histogram (e.g., cosine)
- **How?** Create a relation between (some) pairs based on the secret.



Opt. WM: Choose max # of pairs s.t. the budget (min similarity required) is not exceeded

Federated Learning (FL)



- Up-to-date: continuously learn from real-time web data gathered by users

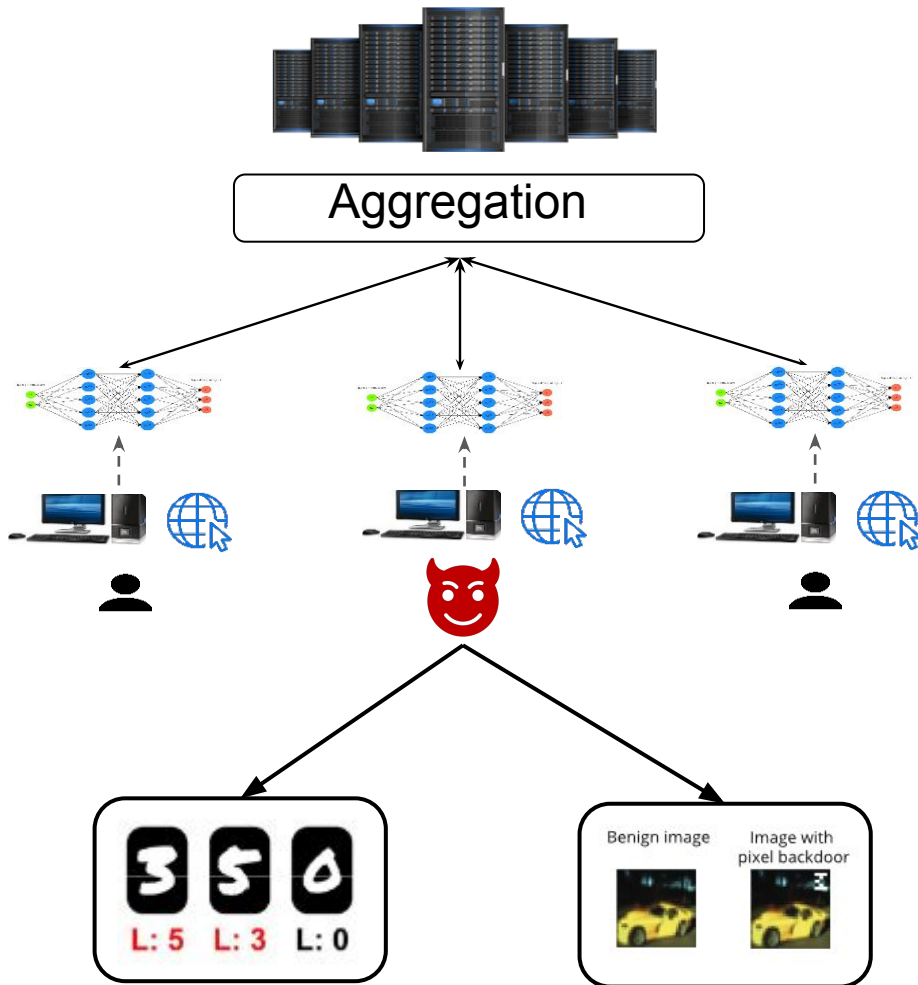


- Being distributed with privacy: users train the classifier locally using personal data



Attacks in FL

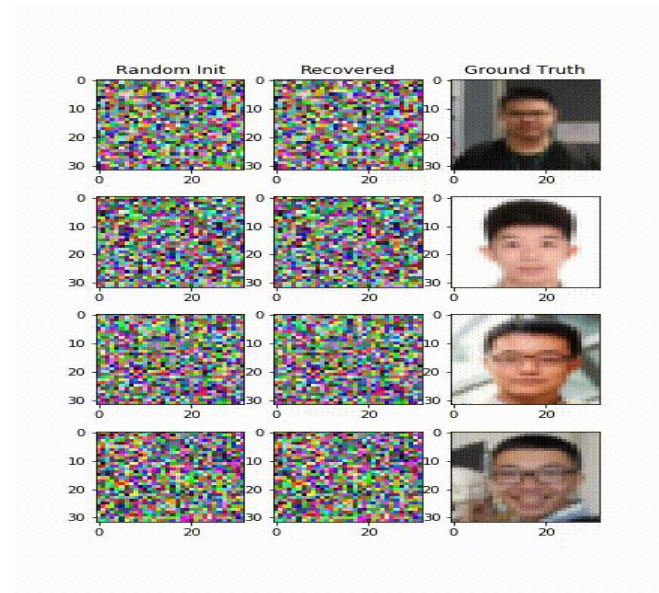
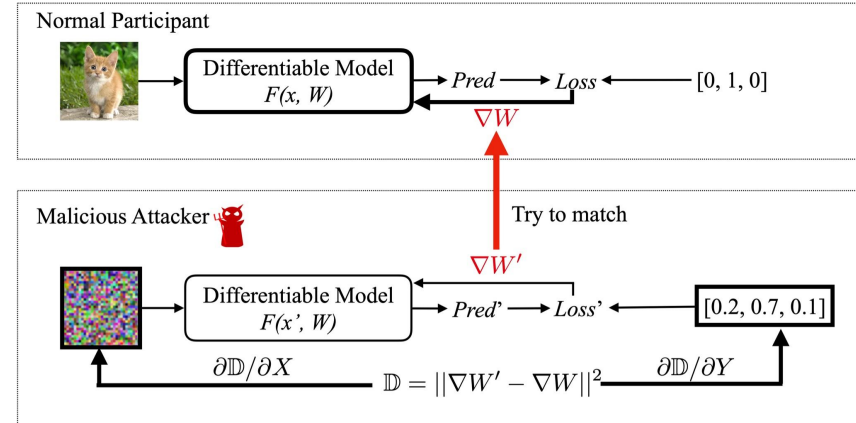
Poisoning attacks



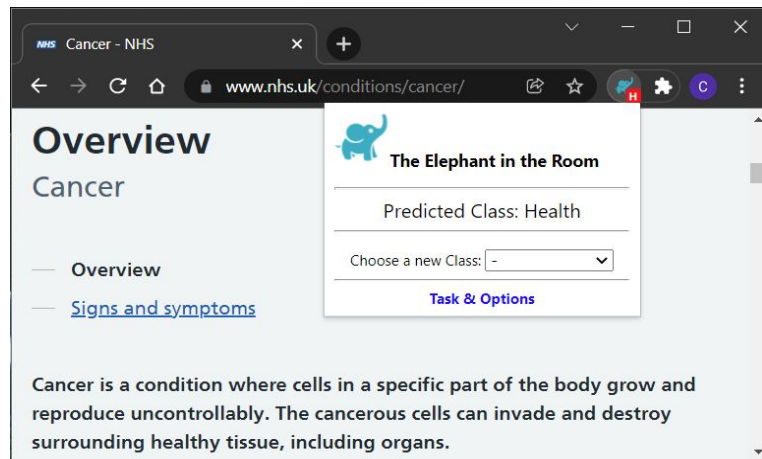
label- flipping attack

backdoor attack

Privacy attacks



FL Classifier for Sensitive content



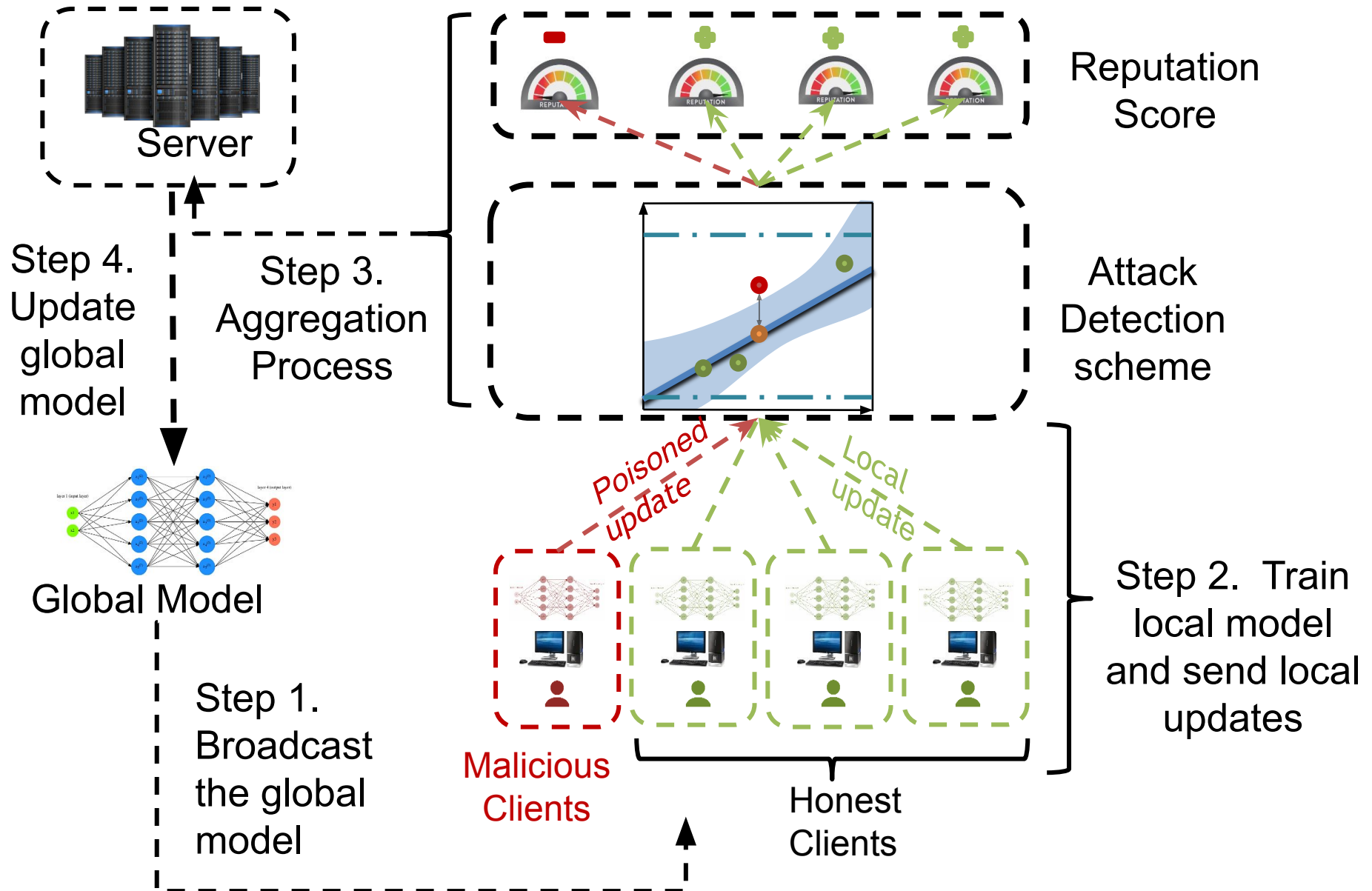
A FL classifier capable of detecting URLs containing sensitive content in real-time

when you visit a webpage, it pops up which category this webpage falls in



T. Chu, A. Garcia-Recuero, C. Iordanou, G. Smaragdakis, N. Laouraris. "[Securing Federated Sensitive Topic Classification against Poisoning Attacks](#)". In Proc. of **NDSS'23**.

Reputation-based FL Aggregation



Evaluation

Theorem

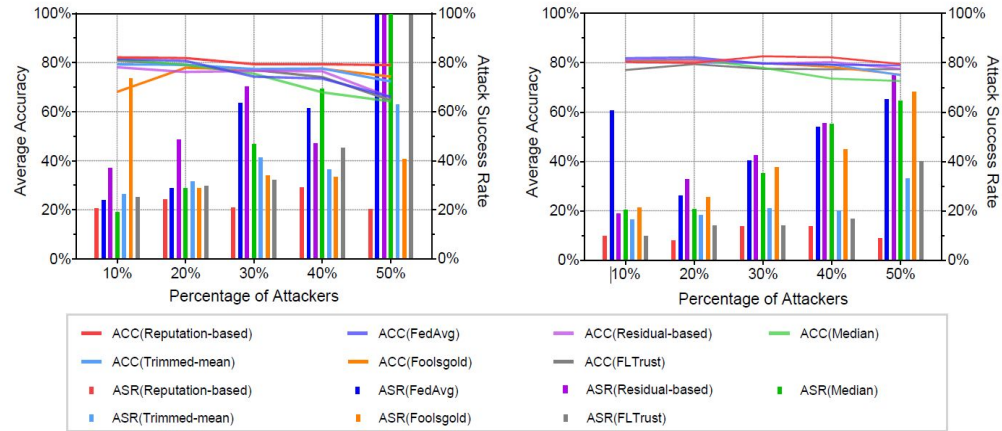
Under Assumptions, $\exists \epsilon > 0$ that:

$$\sqrt{\frac{d \log(1 + \hat{Q}MLD)}{M(1-p)}} + c \frac{G_w}{\sqrt{\hat{Q}}} + p \leq \frac{1}{2} - \epsilon \quad (1)$$

After t rounds, Our Algorithm converges with probability at least

$$1 - \xi \in \left[1 - \frac{4d}{(1 + \hat{Q}MLv)^d}, 1 \right) \text{ as}$$

$$\|w^t - w^*\|_2 \leq (1 - Lr)^t \|w^0 - w^*\|_2 + \frac{\sqrt{N}}{L} \Delta_1 + \frac{1}{L} \Delta_2 \quad (2)$$



The screenshot shows a web browser with multiple tabs. The active tab is 'The Elephant in the Room', which displays a page titled 'Your Task' with a list of URLs. The page content includes:

The Experiment of Identifying Sensitive URLs

In our opinion, being tracked/spied) when visiting web pages that contain sensitive and sexual preference, is the "Elephant in the Room" of privacy. Several data prot in Europe, safeguard online content that contains sensitive data.

In our [recent article](#) S. Matic, C. Jordanou, G. Smaragdakis, N. Laourtaris, "Ident Scale," ACM IMC'20. [pdf], we showed that such spying is taking place on hundres are currently developing technologies to warn users when such tracking is taking f for YOUR help.

In this experiment, we will be showing you URLs from the internet and asking you non-sensitive from your perspective. Below, you will find detailed instructions URLs. We expect that the experiment will take less than 10 minutes and upon can safely uninstall the addon if you do not wish to keep it.

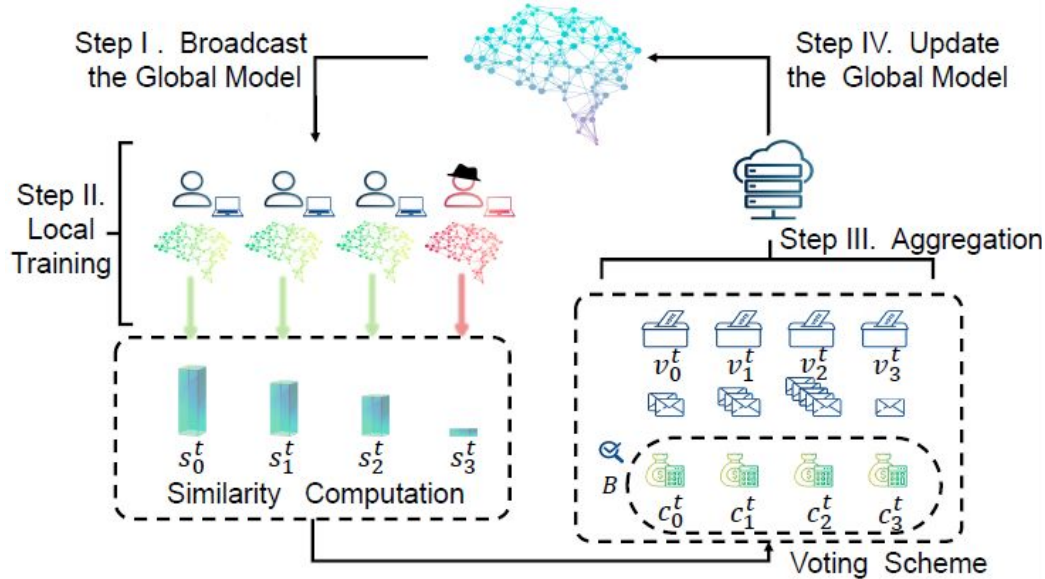
In order to help you to understand what sensitive content is from a legal point of v of sensitive information provided by the current General Data Protection Regulation EU countries.

ARTICLE 9 EU GDPR: "Processing of special categories of personal data"
Processing of personal data revealing racial or ethnic origin, political opinio beliefs, or trade union membership, and the processing of genetic data, biom uniquely identifying a natural person, data concerning health or data concerning sexual orientation shall be prohibited

The right sidebar shows a 'Main Tags' section with categories like 'Archeology and Paleontology', 'Biology & Health', 'Chemistry', 'Computer Science', 'Computing', 'Electronics', 'Energy', 'Materials', 'Mathematics', 'Medical Science', 'Nanotechnology', 'News', 'Other Science', 'Physics', 'Planet and Environment', 'Plants & Animals', 'Robotics', 'Space & Astrophysics', 'Technology', and 'Unusual'. A 'Predicted Class: Health' is also visible.

The bottom part of the browser shows a research paper titled '2019 Novel Coronavirus' with a 3D model of the virus and CT scan images of a patient's lungs. The text discusses the ultrastructural morphology of the virus and its potential transmission routes.

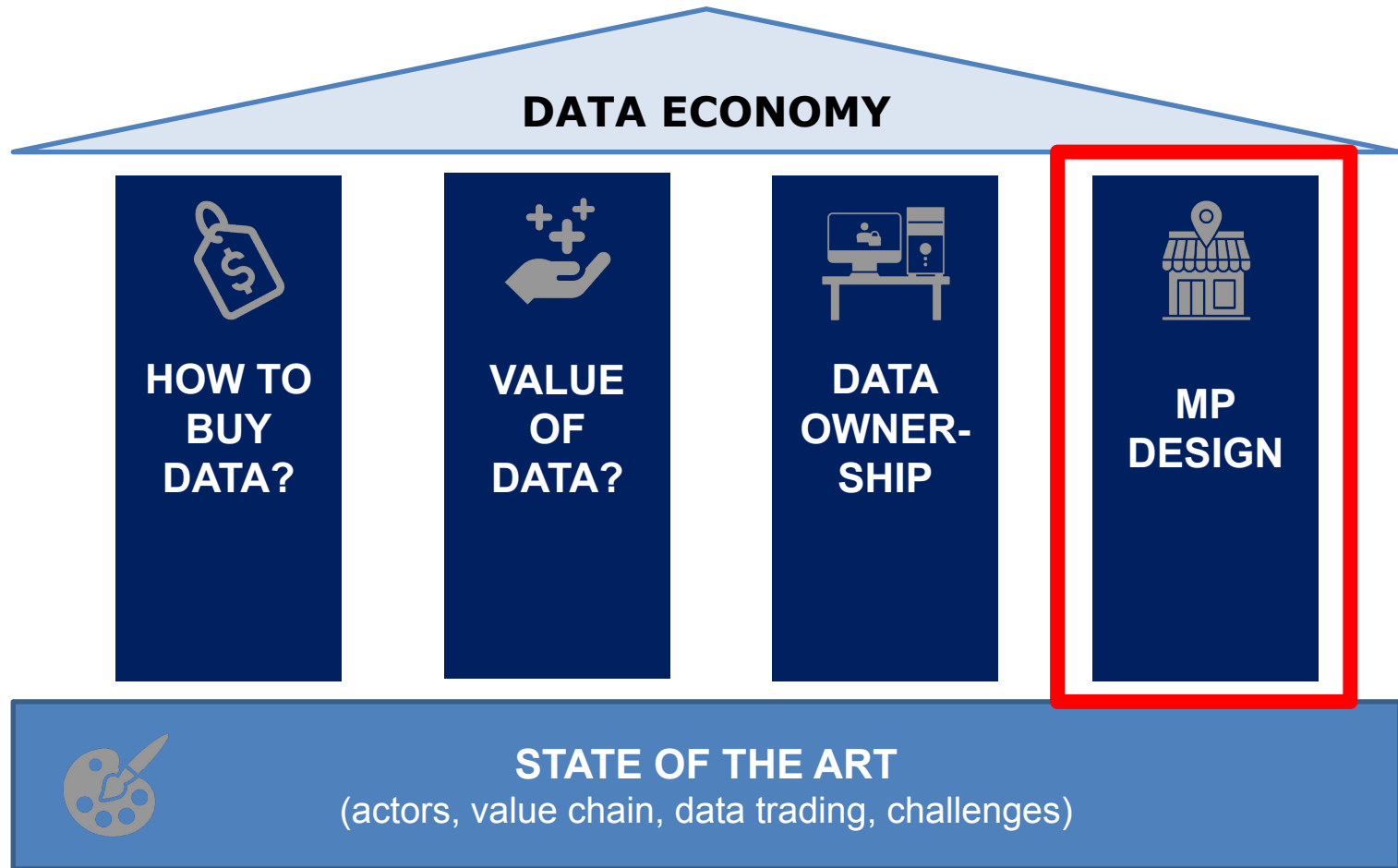
Leveraging Quadratic Voting in FL



Incorporate quadratic voting (QV) into an FL setting to provide a **truthful mechanism** in order to

- Fight **poisoning attacks** and extend it to support **unequal voting budgets** with the **reputation** score of each peer
- **Extend** it to other **Byzantine-robust FL** frameworks as a general design principle

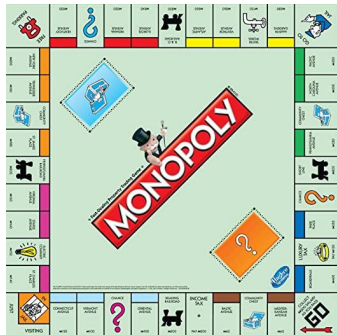
What should we work on?



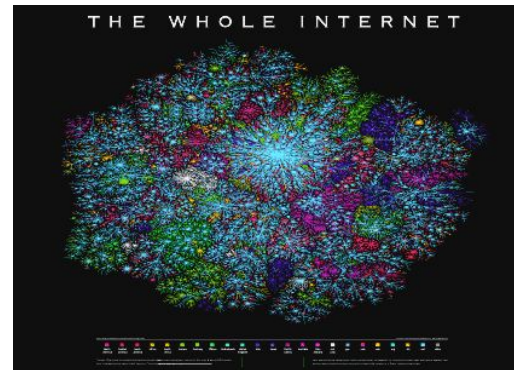


Data Marketplace design

Existing approaches



What we seek instead

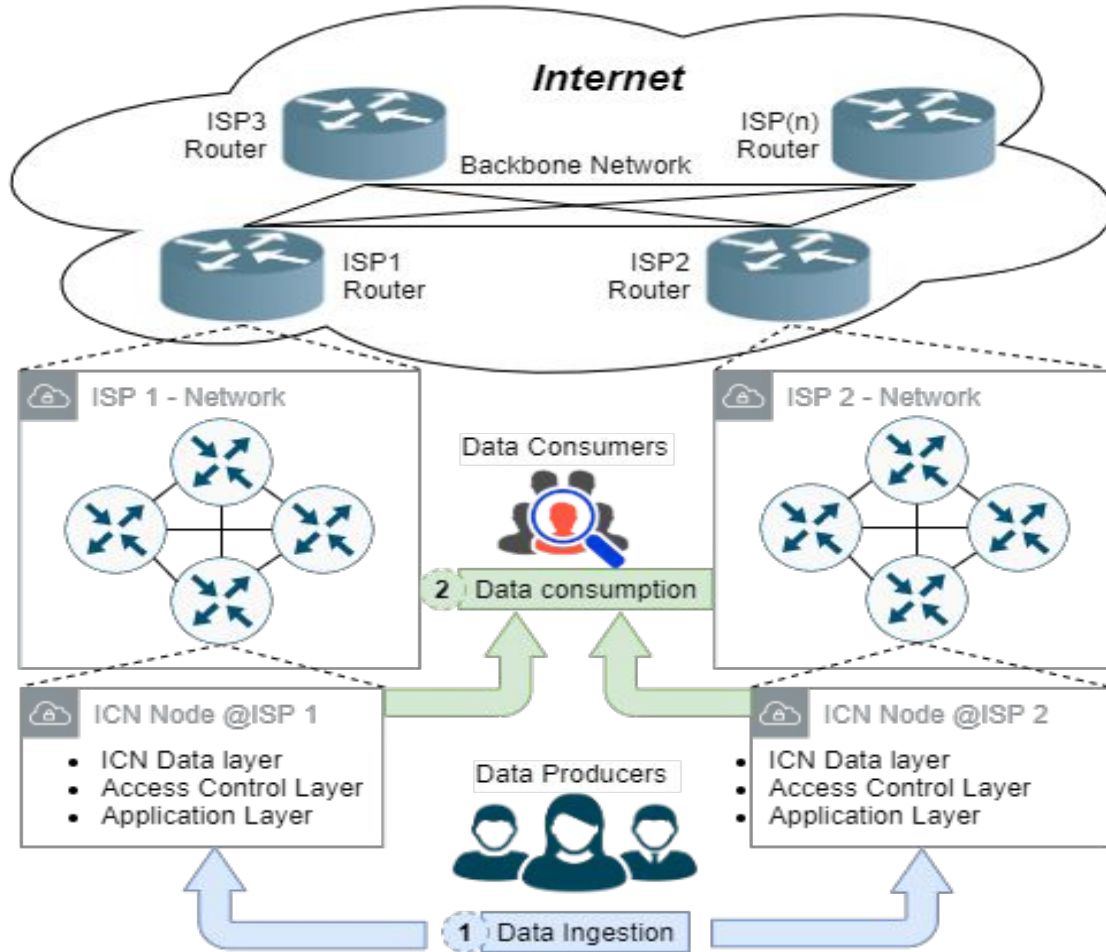


- Open
 - Distributed
 - Democratic
 - Easy to bootstrap
- ... a standard?

[Personal] Data Internetwork

PDI

A distributed Data Inter-Network

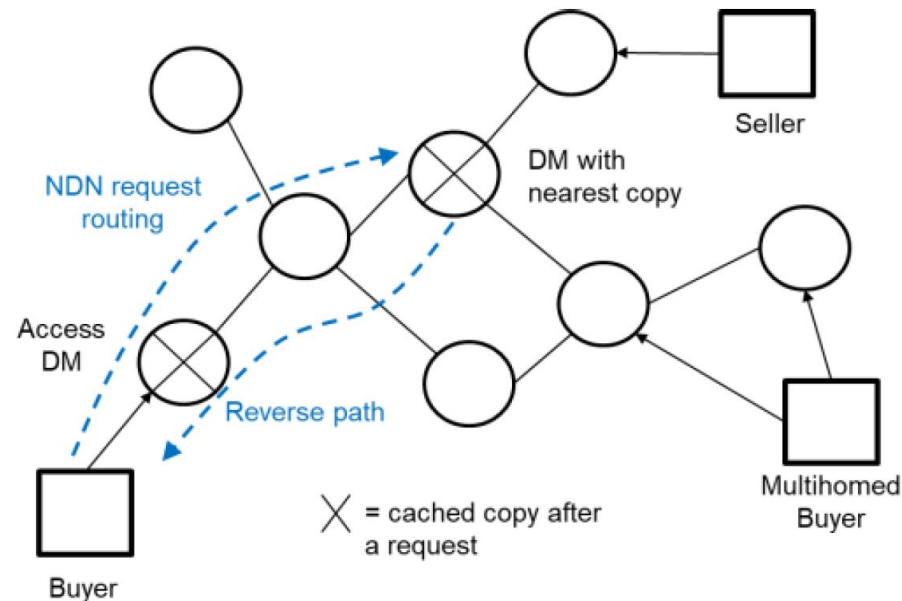


Backbone Personal Data Router

Access Personal Data Router

N. Laoutaris and C. Iordanou, "What do Information Centric Networks, Trusted Execution Environments, and Digital Watermarking have to do with Privacy, the Data Economy, and their future?," ACM CCR, Vol. 51, No. 1, Jan. 2021.

Named Data Networking (NDN) and Data Marketplace Federation



Named Data Networking (NDN) and Data Marketplace Federation

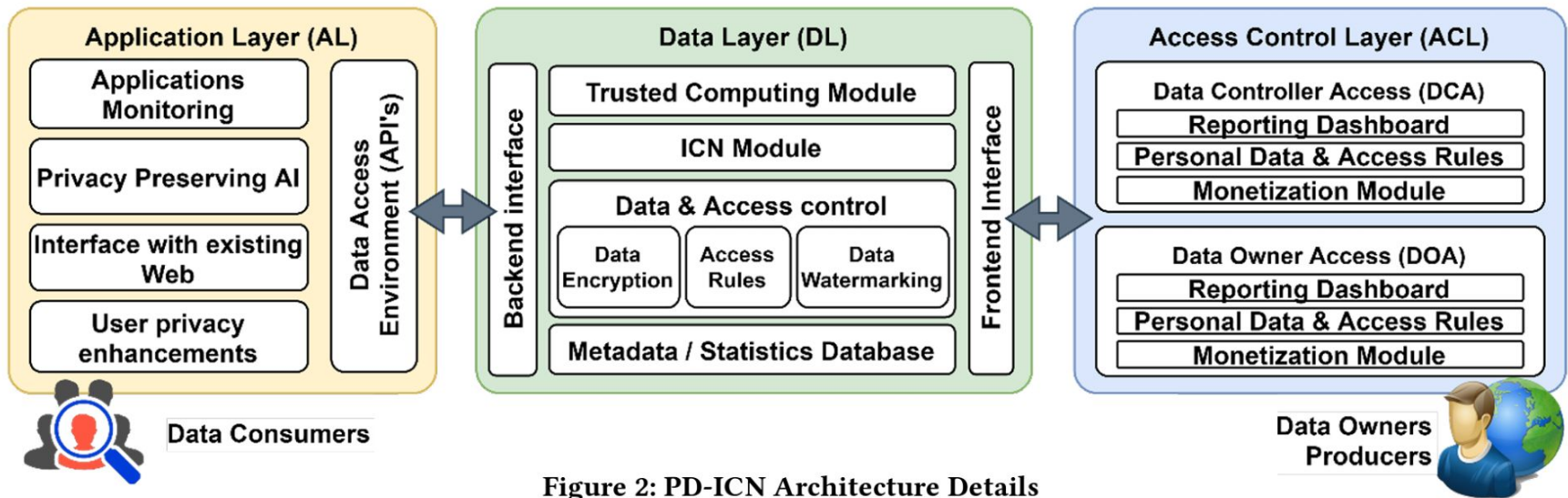


Figure 2: PD-ICN Architecture Details

1st ACM Data Economy Workshop

KEYNOTE SPEAKERS



Data-driven decision making powered by Machine Learning (ML) algorithms is changing how the society and the economy work and is having a profound positive impact on our daily life. With the exception of very large companies that have both the data and the skills to develop powerful ML-driven services, the large majority of provably possible ML services, from e-health, to transportation and predictive maintenance, to name just a few, still remain at the idea or prototype level for the simple reason that data, the skills to manipulate them, and the business models to bring them to market, seldom co-exist under the same roof. Data has to somehow meet with the ML and business skills that can unleash its full power for the society and economy. This has given rise to a highly dynamic sector around the Data Economy, involving Data Providers/Controllers, data Intermediaries, often-times in the form of Data Marketplaces or Personal Information Management Systems for end-users to control and even monetise their personal data. Despite its huge potential and observed initial growth, the Data Economy is still at its nascent phase and, therefore, faces a yet uncertain future and a series of existential challenges. Such challenges include a broad range of technical matters across multiple disciplines of Computer Science including networks and distributed systems, security and privacy, machine learning and human computer interaction. The mission of the ACM DE workshop will be to bring together all the CS skills required for helping the Data Economy liftoff by addressing a range of technical challenges including, but are not limited, to the ones below:



Bruce Pon

Founder of Ocean Protocol



Damian Boeselager

Member of the European Parliament



Carme Artigas

Secretary of State for Digitization and Artificial Intelligence, Government of Spain

2nd ACM Data Economy Workshop



Data-driven decision making powered by Machine Learning (ML) algorithms is changing how the society and the economy work and is having a profound positive impact on our daily life. With the exception of very large companies that have both the data and the skills to develop powerful ML-driven services, the large majority of provably possible ML services, from e-health, to transportation and predictive maintenance, to name just a few, still remain at the idea or prototype level for the simple reason that data, the skills to manipulate them, and the business models to bring them to market, seldom co-exist under the same roof. The value of data comes from its contextualisation and combination with other data. Indeed, this can give way to many new services and products. Furthermore, data has to somehow meet with the ML and business skills that can unleash its full power for the society and economy. This has given rise to a highly dynamic sector around the Data Economy, involving Data Providers/Controllers, data Intermediaries, often-times in the form of Data Marketplaces or Personal Information Management Systems for end-users to control and even monetise their personal data. Despite its huge potential and observed initial growth, the Data Economy is still at its nascent phase and, therefore, faces a yet uncertain future and a series of existential challenges. Such challenges include a broad range of technical matters across multiple disciplines of Computer Science including databases, machine learning, distributed systems, security and privacy, and human computer interaction. The mission of the Data Economy workshop will be to bring together all the CS skills required for helping the Data Economy.



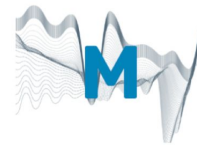
WORKSHOP CHAIRS

- Nikolaos Laoutaris (IMDEA, Spain)
- Georgia Koutrika (Athena Research Center, Greece)

Thank you!



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101070069.



MLEDGE

Cloud and Edge Machine Learning

IMDEA Networks is the beneficiary of this project

[More info](#)

This project (REGAGE22e00052829516) has been funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR.



GOBIERNO
DE ESPAÑA

VICEPRESIDENCIA
PRIMERA DEL GOBIERNO

MINISTERIO
DE ASUNTOS ECONÓMICOS
Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO
DE TELECOMUNICACIONES
E INFRAESTRUCTURAS DIGITALES



Financiado por la union europea
NextGenerationUE