Databanking: Decoupling Data Custody from Data Analysis

D. Protopopescu*
University of Glasgow; databanking.org
(Dated: June 2026)

In the digital age, information can be easily copied, transferred, analysed and stored indefinitely, hence the control of each individual over personal information becomes crucial. We propose a simple and elegant solution to the privacy-versus-convenience dilemma, which simultaneously makes the individual the owner of their data and eventually lays the foundations for a new industry, Databanking. We discuss how every stakeholder wins from our proposal, situate it against existing proposals and partial implementations, and outline the challenges to implementing these ideas, requiring action from both regulators and entrepreneurs.

I. INTRODUCTION

Privacy concerns are nowadays one of the most pressing problems of our society. In an age when information can be easily copied, transferred, analysed and stored indefinitely, individuals and society at large have lost much of their control over personal information, becoming vulnerable to abuse and misuse.

Almost all information today is stored electronically and data processing is done by algorithms. Companies, big and small, from Alphabet and Meta to the corner stationery shop, are hungry for personal information and rush to obtain it in all sorts of ways. From an individual's point of view this is a disadvantage, because once given, control over what is done with the data is effectively lost. It is easy nowadays to store it indefinitely, sell it, or simply lose it in a data breach.

In the past, personal information was held by very few agents, mostly state-approved institutions, it was localised (e.g. on paper), access to it was restricted, and it was difficult to copy. Analysis of this data was done by humans, who have limited memory and limited interest. From a consumer's point of view this was an advantage, as it inherently reduced the spread of personal information.

Money can only be spent once, but data can be copied infinitely. The modern digital economy is built on an asymmetry that did not exist when most data-protection thinking was formed. When an individual hands personal information to a company, they do not transfer it the way they transfer money: they surrender effective control over an infinitely reproducible asset, one that can be sold repeatedly, merged with other datasets, moved across jurisdictions, retained for decades, and reused for purposes never anticipated by its source.

This distinction, rather than any single bad actor, is the root of the privacy problem of the digital age, and it is the asymmetry our proposal is designed to flip.

Importantly, we do not advocate withholding data from analysis altogether – the analysis of personal data carries genuine, often indisputable utility. Aggregated patient data underpins epidemiological research and clinical advances that would be impossible to obtain otherwise; purchase and browsing histories enable product recommendations, fraud detection, and credit assessment; large training datasets underlie the AI systems increasingly relied upon for translation, diagnosis, and decision support. The privacy problem we address in this paper is therefore better understood as a dilemma than as a simple prohibition: how to preserve the genuine benefits of data analysis without requiring individuals to surrender control over the underlying information in order to obtain them.

We propose that individuals should never hand over their raw personal data directly to service providers. Instead, personal information should be held by regulated, audited *Databanks*, trusted third parties that act as custodians and intermediaries, structurally forbidden from analysing or monetising the data they hold beyond regulated access fees. This is, in essence, the banking model applied to information rather than money: Databanks could be commercial, tiered-service companies, non-profits, or state-backed institutions, freely chosen and freely switched by the individual, exactly as one chooses and switches a bank account today.

Money and information differ in a fundamental economic respect. Money is rivalrous: spending it removes it from the spender's possession. Information is non-rivalrous: sharing it leaves the original intact while creating additional copies.

Traditional banking evolved to protect rivalrous assets. Databanking can be understood as an analogous institutional response to the protection of non-rivalrous assets.

Definition

A Databank is a regulated computational custodian of personal information that stores data on behalf of an Owner, enforces machine-readable access policies, executes approved third-party computations within controlled sandbox environments, and releases only policy-compliant outputs while preventing uncontrolled duplication of the underlying information.

*Electronic address: wp@databanking.org

A Databank is therefore not merely a storage provider. It is a regulated computation boundary separating data custody from data analysis.

II. RELATED WORK

The idea that personal data should be held by a neutral custodian rather than by the services that consume it has been proposed, in different forms, for over a decade, and several of its building blocks already exist independently. The contribution of this proposal lies in combining them under one regulatory model with a structural, legally enforced separation between custody and analysis.

Vendor Relationship Management. Doc Searls' ProjectVRM, launched in 2006, argued that many marketplace problems can only be solved from the customer side, by making the individual rather than the vendor the point of integration for their own data [1]. Databanking can be read as VRM's missing infrastructure layer: VRM describes the desired power relationship, while a Databank is the regulated institution that could hold the data and enforce it.

Self-sovereign identity and personal data stores. The Solid project, initiated by Tim Berners-Lee, gives individuals a personal data "pod" that applications must request permission to read rather than copy [2]. Personal Data Stores (PDS) and the MyData architecture pursue a similar goal of household- or individual-level data sovereignty, with the individual as account owner and explicit consent flows between data "sources" and "sinks" [3]. Self-sovereign identity (SSI) systems extend this to verifiable credentials that prove a claim about a person without exposing the underlying document, an idea directly related to the "bit-scarce result" principle discussed in Section IV C [4]. A recent critical review finds that current SSI implementations only partially satisfy GDPR principles, underscoring that technical decentralisation alone does not guarantee regulatory compliance, a gap a regulated Databank is intended to close [5].

Data trusts and data cooperatives. The Open Data Institute has since 2018 explored *data trusts*: legal structures giving trustees a fiduciary duty to make data-access decisions on behalf of a group of people [6]. Hardjono and Pentland's *data cooperatives* similarly propose member-owned organisations with fiduciary obligations that manage and protect access to members' data [7, 8]. Both share Databanking's central premise – that someone other than the data subject *and* other than the data consumer should control access – but neither, to our knowledge, has been paired with the bit-scarce sandboxed computation we propose in Section IV A, nor with a statutory revenue-sharing mechanism of the kind discussed in Section V.

Data as labour and data dividends. Posner and Weyl's *Radical Markets* argues that data should be treated as labour rather than as capital freely appropriated by platforms, and that individuals should be compensated and able to organise collectively around it [9]. Jaron Lanier made an earlier, related argument for micropayments to the sources of data in *Who Owns the Future* [10]. California's 2019 "data dividend" proposal was a direct, if never fully realised, policy attempt at the same idea [11]. Databanking shares the goal of returning value to the individual, but proposes a different mechanism: a custodian collects access fees from data requestors and remits a regulated percentage of those fees to the individual, rather than the individual being paid directly for raw data, avoiding any incentive to sell data outright.

Regulatory precedent. The EU's General Data Protection Regulation (GDPR) created, in Article 20, a right to data portability: the ability to receive one's data in a structured, machine-readable format and transmit it to another controller. In practice this right has proved narrow: it excludes inferred data such as credit scores, does not apply to most processing grounds other than consent or contract, and creates no obligation for systems to be technically interoperable [12]. The EU Data Governance Act, binding since September 2023, goes further and creates a legal category of *data intermediation service providers*, who must be structurally and legally separate from any other business they operate and must remain neutral, with no interest in the data they intermediate beyond the intermediation service itself [13]. This is close in spirit to the structural separation we describe in Section III for an incumbent such as Meta spinning off a Databank subsidiary, and is to our knowledge the closest existing regulatory instrument to the Databank concept, though it does not yet mandate the sandboxed, bit-scarce-result architecture we propose, nor a statutory share of access revenue for the individual.

Privacy-preserving computation. Trusted execution environments (TEEs), such as hardware secure enclaves, and complementary techniques such as multi-party computation and federated learning, already allow a piece of code to be run against data it never directly exposes to the party that supplied the code [15]. These are the likely technical substrate for the in-Databank sandbox described in Section IV A: the regulatory model we propose does not depend on any one of them, but on the legal requirement that some such mechanism is used and audited. Zero-knowledge proofs, increasingly used for privacy-preserving age and identity verification, are a directly applicable instance of a bit-scarce result (e.g. "has age over 18") and illustrate both the promise and the limits of the approach: proving many narrow claims about the same attribute over time can itself leak information, a caution that should inform how Databanks rate-limit and log repeated queries [16].

Consumer-facing precedents. A handful of deployed, single-vendor features already implement narrow pieces of the bit-scarce principle without any regulatory mandate. Apple's *Sign in with Apple* and *Hide My Email* interpose a per-service forwarding address between a user and a website, so the recipient gets a working

contact channel rather than the user's real email; *iCloud Private Relay* similarly splits a user's browsing traffic across two relays so that, in Apple's own description, "no single party – not even Apple – can see both who you are and what sites you're visiting" [17]. Tokenised age-verification services such as Yoti return only the result of an age check (e.g. "over 18: yes") plus metadata on how and when it was performed, never the underlying date of birth or document [18]. Consumer credit checks offer a more ambivalent precedent: a *soft* inquiry, used for pre-qualification, returns a limited score or risk tier rather than the full file, but a *hard* inquiry, used for actual underwriting, still hands the lender the complete credit report held by the bureau, so the bureau acts as custodian for storage purposes only, not as an enforced sandbox [19]. All three cases show that narrowing what crosses an organisational boundary is *technically and commercially viable today*; none of them generalise the principle across categories of personal data, make it a statutory entitlement, or share resulting revenue with the individual, which is what Databanking proposes to do.

Taken together, the building blocks of Databanking already exist in fragments: a custody model (data trusts, cooperatives), a technical substrate (TEEs, ZKPs, SSI), an economic argument (data as labour), and an early regulatory instrument (the Data Governance Act). What is, to our knowledge, missing from the literature is the combination proposed here: a statutory, revenue-sharing custodial market with mandatory portability and a hard legal wall between custody and analysis, of the kind described in Sections III–V.

III. REGULATION

This is how we would envisage regulation for Databanks: only approved actors can store personal information (PI), with an individual's permission. PI includes not only information received directly from an individual, but any data, however derived, that relates to an identifiable individual, including profiles, scores, and other inferences generated by a Requestor. Storage and/or transfer of PI outside this framework is forbidden by law and subject to civil and criminal penalties.

Databanks are forbidden by law from analysing the data stored in user accounts. If Meta wished to become a Databank, for example, it would have to spin off a separate company, say Meta Databank, which holds PI accounts but is completely forbidden from mining them and is fully insulated from the rest of Meta Platforms, Inc., in essence, a new company under different ownership obligations. As discussed in Section II, this is close in spirit to the structural separation already required of data intermediation service providers under the EU Data Governance Act [13].

This concentrates personal information inside a small number of regulated entities, which is itself a legitimate objection. But it does not introduce concentration risk from nothing: it replaces today's uncontrolled replication risk – the same data copied across thousands of companies with uneven security practices – with a regulated custodial risk of the kind centuries of banking practice in operational security, auditing, and fraud detection are already well placed to manage.

IV. DATABANKING

Databanks can be non-profit, state-backed institutions, or commercial, audited, tiered-service companies. Customers, which we call Owners, can freely choose their Databank and service plan (by law including a free tier-0, no-frills service), choose what to store, and exercise granular access control. Owners can move their account to another Databank at any time. To make this portability real rather than nominal, the account format itself – the schema in which PI is stored, tagged, and permissioned – should be standardised across Databanks, e.g. via a regulator-maintained schema, so that moving an account between providers is as mechanical as a bank-to-bank switch today. Databanks would then compete on service rather than on lock-in, for example by offering different data-monetisation plans, tiered revenue shares, or bundled analytics to attract Owners.

An Owner (a person or a company, i.e. any legal entity) might hold a basic, "empty" account containing only a name and address, like a passport, proving only that the Owner exists. Or the Owner could store name, address, next of kin, financial data, purchase history, and medical history all in a single account, either as one simple all-in-one account or as a granular-access account with or without data aggregation restrictions. Alternatively, an Owner could split accounts across different Databanks, each holding only one category of personal information: a purchase-history account with one Databank, medical history with another, and so on.

Ownership does not imply unrestricted editing rights over an account's contents. Certain categories of data – credit history and medical history among them – are not directly editable by the Owner, since their evidentiary value depends on entries not being unilaterally alterable. Owners can nonetheless see this data in full and access an audit trail of how each entry was generated, so errors can be flagged and corrections requested through the Databank – a level of transparency current credit and medical records rarely offer.

In direct analogy with money, an Owner could in principle store all their data on a hard drive at home. The incentive to hold an account instead is that it conveniently allows granular access to be granted to Requestors (for example medical services, creditors, or customer loyalty schemes) when desired. Custody alone is not the service a Databank provides: authentication, guaranteed availability, portability, dispute resolution, regulatory compliance, statutory compensation handling, and key recovery

are functions a hard drive at home cannot offer, which is why Databanks more closely resemble financial institutions than storage providers.

Once regulation for this scheme is in place, Databanks are financially incentivised by a statutory percentage of the Requestors' access fees, and by monthly fees for tier-1 and higher accounts. Today, data collectors such as Meta, Google, and Amazon collect data and keep all the financial rewards from its analysis and sale, with Owners receiving nothing, or arguably negative outcomes such as behavioural targeting and loss of control.

Beyond single-query bit-scarce answers, Databanks could also offer analysers access to consolidated, anonymised, cross-sectional datasets drawn across many Owners' accounts – for example for academic research or market analysis – provided every contributing Owner has separately opted in to that specific use and is remunerated accordingly. This is a distinct revenue and consent mechanism from the per-query sandbox described in Section IV A: it trades some of the bit-scarce model's privacy guarantee for statistical usefulness, and should therefore require its own, stricter opt-in and anonymisation safeguards (e.g. k-anonymity or differential privacy) rather than being bundled into default account settings [24, 25].

A. Sandboxed computation inside the Databank

The core technical mechanism that could make the privacy-versus-convenience dilemma tractable is that an Owner's data never leaves the Databank that holds it. There is no separate third party brokering access: the Databank itself is both custodian and sandbox operator. When a request arrives, the Databank brings the specific subset of the Owner's data the query requires together with the Requestor's algorithm inside an ephemeral, internal container created for that query alone; only a bit-scarce result leaves the container, after which the container – and everything assembled inside it – is destroyed. Concretely:

1. Owners store their data with their chosen Databank, in a secure format, under access policies they control.

2. A Requestor submits an algorithm or query – not a request for raw data – to the Databank. The Databank checks the request against the Owner's standing permissions: admitting it, rejecting it outright and logging the denial, or, depending on product design, forwarding a one-off consent request to the Owner before proceeding.

3. If admitted, the Databank assembles, internally and ephemerally, the specific data subset the query requires together with the Requestor's algorithm inside a fresh container created for that query alone. Hardware trusted execution environments, multi-party computation, and zero-knowledge proofs are plausible technical substrates for this step [15, 16].

4. Only a bit-scarce result crosses the boundary back to the Requestor (Section IV C); the raw data never leaves the container, and the container is destroyed immediately afterward. Requestors pay the Databank for the service and the Owner receives a percentage. The Owner can see on its Databank statement all requests, with details on the data accessed, the algorithm's scope, and the actual result, and can ban certain Requestors.

Benefits of this approach include enhanced privacy (raw data is never shared), customisable access (granular, per-requestor permissions), security (a single, auditable point of encryption and access control rather than data scattered across many companies), and easier regulatory compliance (e.g. with GDPR or the California Consumer Privacy Act).

Challenges include the need to establish trust in the Databank operator itself, the risk of centralisation and single points of failure (which decentralised or federated Databank architectures could mitigate), the performance overhead of mediated access, and the difficulty of persuading incumbent companies, accustomed to direct data collection, to adopt the model voluntarily, which is why we argue regulation will be needed in order to drive the transition.

B. Reference architecture

At a conceptual level a Databank consists of seven functional components: (i) an identity service responsible for authentication and account recovery; (ii) an encrypted data vault containing the underlying personal information; (iii) a consent engine implementing user permissions; (iv) a sandbox manager responsible for executing approved third-party algorithms; (v) an output filter enforcing information-leakage limits; (vi) an audit ledger recording all queries and results; and (vii) a portability interface implementing the standardised account format discussed above, through which an Owner's account can be exported to, or imported from, another Databank. Figure 2 sketches how these components sit inside the Databank boundary, under regulatory oversight, between the Owner and a Requestor.

This architecture is intentionally designed around a separation between custody and analysis. Requestors do not obtain raw data. Instead, approved computations are brought to the data and executed within controlled environments whose outputs are constrained by policy, audited, and potentially rate-limited.

The same principle extends to AI training and inference: rather than exporting datasets to wherever a model is developed, an approved model can instead be brought to the data and run inside a Databank-controlled envi-

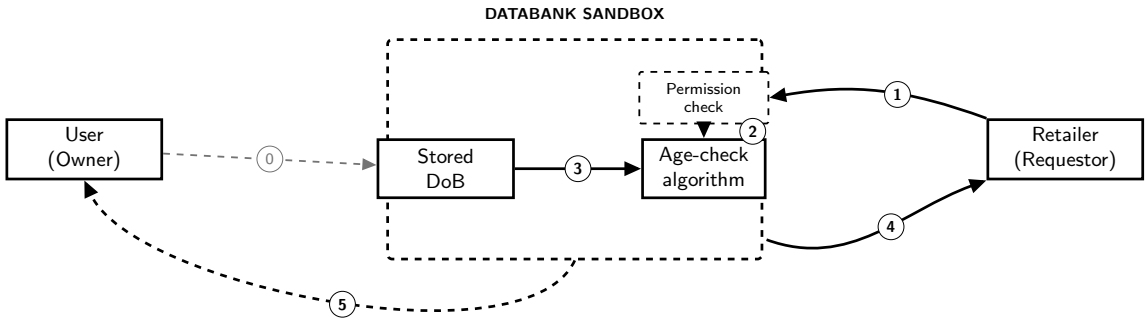

FIG. 1: A worked sandboxed query: a retailer checks whether a user is over 18 without ever receiving the user's date of birth. The user's date of birth is stored in their Databank account. For a given query: (1) the retailer submits an algorithm to the sandbox, not a request for data; (2) the Databank checks the request against the user's standing permissions, admitting it, rejecting it, or forwarding a one-off consent request to the user; (3) the admitted algorithm executes inside the sandbox, where it alone meets the stored record; (4) a single bit – "eligible: true" – crosses the boundary, never the underlying date of birth; (5) the transaction is logged on the user's statement and the user receives a statutory share of the access fee paid by the retailer.

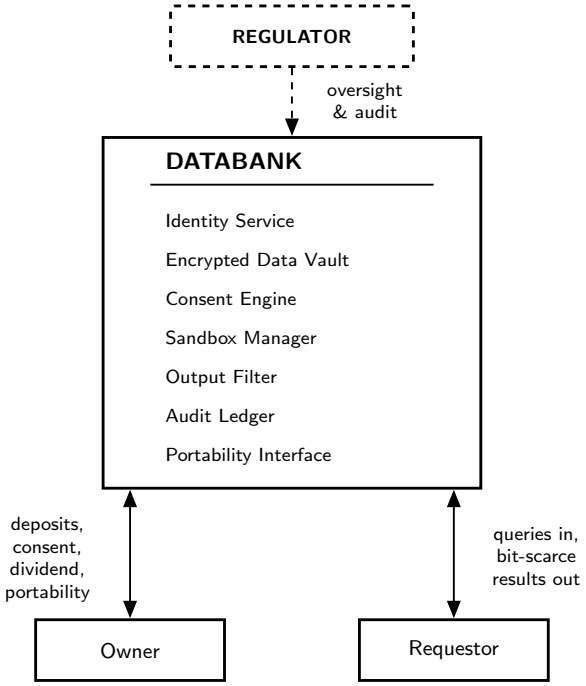

FIG. 2: Reference architecture of a Databank. A regulator audits the Databank's operation from outside; internally, the seven components listed above jointly enforce the custody/analysis separation; Owners deposit data and receive consent, dividend, and portability services, while Requestors submit queries and receive only bit-scarce results.

ronment, extending the custody/analysis separation to machine learning itself.

C. Bit-scarce results

We use the term *bit-scarce* to denote outputs intentionally constrained to reveal only a small amount of information relative to the underlying dataset. In simple cases, this may mean outputs of only a few bits, such as a yes/no eligibility result. More generally, the relevant constraint is not a fixed universal bit count, but an auditable leakage budget enforced across repeated queries. Without such a limit, a Requestor could encode parts of the PI within the result itself, defeating the purpose of the sandbox. This constraint, and its limits, parallel the information-leakage concerns already identified in the zero-knowledge-proof literature on repeated, narrow disclosures about the same attribute [16].

V. REGULATORY FRAMEWORK

A regulatory framework for Databanks should focus on ensuring data privacy and preventing the misuse of user data:

1. **Prohibition of data analysis.** Regulations should clearly prohibit Databanks from analysing user data for their own benefit or for any purpose not explicitly authorised by users, keeping their focus on secure storage and management. The same prohibition extends to Requestors, who must not retain derived data, profiles, or inferences about an individual outside the Databank that holds the underlying PI. PI should be standardised, e.g. via an XML or JSON schema published and periodically revised by regulators.

2. **Revenue model restrictions.** Databank revenue should come only from user subscriptions and third-party access fees, discouraging Databanks from ex-

ploiting user data for profit and aligning their incentives with secure, privacy-preserving service.

3. **User income from data access.** Users should be entitled to a statutory share of the income generated whenever third parties access their data, encouraging participation and giving users a direct stake in the monetisation of their own data, the "data dividend" argument made by Lanier and by Posner and Weyl [9, 10], implemented here as a custodial fee share rather than a direct data sale.

4. **Granular access controls.** Databanks must give users the ability to set granular access controls and permissions, determining who can access their data and for what purpose. Beyond fully bespoke settings, Databanks should also offer predefined template access schemes (e.g. Maximum Privacy, Research Friendly, Healthcare Friendly, or Revenue Maximising), so that Owners who do not wish to configure permissions item by item can opt in or out with a single choice, rather than facing the same granular form for every category of data.

5. **Transparency and accountability.** Databanks should provide transparency reports and undergo regular audits, building trust and ensuring compliance.

6. **Data portability.** Users must be able to transfer their data between Databanks easily, promoting competition and choice. This is a stronger, mandatory version of the right already established, in narrower form, by GDPR Article 20 [12].

7. **Security requirements.** Databanks must meet strict security requirements (robust encryption, secure storage, strong authentication), to protect against breaches and unauthorised access.

Databanking does not eliminate lawful government access to personal data; it changes the access path. Today a warrant or equivalent legal instrument is typically served directly on the company holding the data. Under Databanking it would instead be served on the Databank, through a regulated access mechanism rather than an ad hoc one: substituting Government → company database with Government → regulated access mechanism → Databank. Because every Databank already logs every access to a single audit ledger (Section IV A), lawful requests become more accountable and more transparent than today's fragmented landscape of per-company compliance, not less.

Regulation along these lines would require companies such as Alphabet or Meta either to delete all existing user PI or to spin off a separate, insulated entity to hold it, much as the EU Data Governance Act already requires of data intermediation service providers, though without (yet) the statutory revenue share or sandboxed bit-scarce architecture proposed here [13]. Individuals could hold several Databank accounts, for different categories of data, or a single one, and could legally transfer all data held by one Databank to another of their choosing – just as one can transfer money from one bank to another today.

VI. A NEW INDUSTRY

A system of trusted Databanks managing user data securely and privacy-preservingly could give rise to a new industry, with several implications: new business opportunities in data management and privacy-as-a-service; value-added services such as in-sandbox analytics or privacy-preserving machine learning; increased consumer trust in online services; job creation in data security, privacy engineering, and compliance; broader economic participation, as users share directly in the value generated by their data; and pressure toward regulatory harmonisation across jurisdictions, comparable to what is beginning to happen around data intermediaries under the EU Data Governance Act [13].

Adoption would likely be gradual: early movers position themselves as privacy-conscious; competitive pressure and growing consumer awareness push wider adoption; collaboration among Databanks, service providers, and regulators establishes interoperable standards; and legislative change formalises and accelerates the transition. The combination of demand-side preference and regulatory pressure mirrors how data portability and intermediary regulation have themselves evolved over the past decade, from a niche idea (e.g. in ProjectVRM) to binding law (the Data Governance Act) [1, 13].

The scale of this opportunity is suggested, though certainly not proven, by a simple comparison of growth rates. Industry forecasts put the compound annual growth rate of global data creation at roughly 20–25% over the early 2020s, with the total volume of data created worldwide projected to triple again within a few years [26]. Broad money supply (M2) in major economies, by contrast, has historically grown at a much lower compound rate, on the order of 6–8% per year outside crisis-driven monetary expansions [27]. We do not claim that a Databanking industry would simply inherit data's growth rate, or that it would outgrow banking specifically; data and money are different kinds of asset, traded under different constraints, and the comparison above is illustrative rather than a forecast. What it does suggest, as a hypothesis rather than a prediction, is that if even a small, regulated share of the value created by querying data – rather than money – were captured by a Databanking industry, the prospective growth of that industry could be unprecedented among custodial industries, simply because the underlying asset it would custody is growing several times faster than money ever has.

Data also resembles money in one further respect, while exceeding it in another: just as deposited money earns interest, deposited and queried data generates fur-

ther data (derived insights, model outputs, aggregated statistics), through analysis and synthesis, each of which can itself become a fresh, bit-scarce, queryable asset inside the same Databank. Unlike compound interest on money, this compounding is not bounded by a fixed rate set by a central bank, since the number of distinct analyses that can be run over a growing body of data could grow faster than the data itself.

Growth of this industry, however, would not necessarily mean growth of the data economy's physical footprint; in aggregate, it could mean the opposite. Most new industries add infrastructure; a Databanking industry would grow largely by eliminating redundant infrastructure: records that today exist as hundreds of near-identical copies – the same identity record held by each of the 600-plus registered data brokers a single deletion request must reach, the same medical history replicated across multiple providers – could instead exist once, under custody, and be queried in place. Since a substantial share of the global datasphere consists of replicated rather than original data [26], and every copy carries storage, energy, governance, and security overhead, the Databanking sector's growth would represent, in aggregate resource terms, a contraction of the datasphere it replaces.

VII. CASE STUDIES

The following case studies illustrate how the Databanking model could apply across a range of everyday interactions involving personal data. The list is illustrative, not exhaustive: it is meant to demonstrate the breadth of the model's applicability, not to delimit it, and further use cases will likely emerge once the underlying regulatory and technical infrastructure is in place.

A. Credit-worthiness

A creditor needs only a yes/no answer – does this individual meet the lending criteria for a particular product – not the underlying financial history. This is a direct application of the bit-scarce result described in Section IV C. Existing credit bureaus already occupy something close to this custodial role (they hold the financial history on the individual's behalf), but operate as data sellers rather than enforced sandboxes, and a hard inquiry generally gives the lender substantially richer access to the credit file than a simple yes/no eligibility result (Section II). Under a mandatory Databanking regime, such bureaus would face the same choice as any other incumbent custodian: retool into a regulated Databank – enforcing the sandbox boundary, sharing query revenue with the individual, and supporting portability – or relinquish the custodial role to one, retaining only a scoring or analytics business built on top of sandboxed queries.

B. Authentication

An individual can prove a fact about themselves – age, residency, citizenship, professional qualification – without revealing the underlying document, using zero-knowledge proofs or comparable selective-disclosure credentials, as already explored in the self-sovereign identity literature [4, 16].

C. Health data

Large-scale, cross-sectional health data has produced genuine public benefit: population health surveillance, epidemiological research, and clinical studies all depend on aggregated patient data, and none of that value is in question here. The problem is not that such studies happen, but how individuals are currently made to participate in them: in most existing systems, almost every hospital, clinic, insurer, pharmacy, and research programme an individual deals with typically requires its own copy of the relevant medical history. So the same sensitive information ends up duplicated, indefinitely retained, and exposed to a breach, across an ever-growing number of unrelated organisations, each with its own security posture and incentives. Denmark is a notable exception: its national health portal aggregates access to information from multiple underlying health-data sources, and its Shared Medication Record [14] gives citizens and treating professionals access to a shared medication view, which is close in spirit to the query-don't-copy principle proposed here.

This duplication, where it persists, is also a direct economic cost, separate from the privacy one: each copy must independently be secured, backed up, audited, migrated between systems, and eventually archived or deleted, so a meaningful share of health-IT spending goes toward maintaining many partially overlapping copies of the same record rather than toward care itself. A Databank replaces copying with querying – providers consult a record when they need it rather than each holding their own replica – the same principle that lets a bank balance be checked without copying the account. A single longitudinal record, enriched by every provider an individual sees over a lifetime rather than fragmented across them, is also clinically more valuable than the isolated snapshots typical of today's system, since medical information tends to gain rather than lose usefulness over time.

Individuals should instead retain full ownership of their health data in a single account, with medical providers, insurers, and researchers accessing it only through the sandbox, under explicit, auditable authorisation, and revocable at any time. A hospital establishing care would query the record it needs rather than receive a copy of it; a researcher running a population study would, as discussed above, access consolidated, anonymised, cross-sectional data with the same opt-in and remuneration safeguards as any other aggregate analysis, rather than

the broad, indefinite consent typically obtained once and relied on for any future study. The aggregate research benefit is preserved; what disappears is the need to hand the same raw, identifiable record to every entity that might one day want to query a piece of it.

The standard objection to any scheme that tightens individual control over health data is the public-health emergency: an epidemic cannot wait for millions of individual opt-ins. A Databanking regime can accommodate this without abandoning custodial control by default, through a predefined, legislated emergency-access mechanism under which broader, time-limited access becomes available, but only under conditions fixed in advance, logged, and independently reviewable after the fact. The existence of this exception should constrain its own scope, not become the default architecture for ordinary, non-emergency care. In short: the problem addressed here is not that medical information is used, but that it is copied; this is the same distinction, applied to health data, that motivates the rest of this proposal.

D. Online browsing and advertising

Shopping history and browsing interests could, with permission, be made accessible across sellers in real time or offline, for example for relevant recommendations, while advertisers purchase access to audience-matching queries rather than to raw user profiles – a model already anticipated, in less regulated form, by the VRM and data cooperative literature [1, 7].

A Databanking regime would also require today's cross-site tracking infrastructure to be replaced or sharply constrained under Section III. A redesigned browser would, with the user's opt-in, authenticate with a token issued by the user's own Databank and route navigation data to that Databank account rather than to third-party trackers. Sites wishing to use this browsing history for personalised offers would then query it directly from the user's Databank, for a fee shared with the user, rather than collecting and retaining it themselves. This relocates the existing ad-tech tracking layer onto the same custody/analysis boundary as the rest of the proposal, rather than introducing a new mechanism.

E. Online shopping and shipping

On marketplaces such as eBay or Amazon, the buyer's address is handed today to every seller, large or small, each of whom could in principle retain it, resell it, or expose it in a future data breach.

Under Databanking, the buyer's Databank issues an opaque delivery token in place of the buyer's address. The seller receives only this token and passes it to the carrier; the carrier alone resolves the token to an actual delivery address, through a secure channel with the buyer's Databank, so the buyer's address never reaches the seller. Existing payment intermediaries such as PayPal already demonstrate a partial version of this pattern for payment details; Databanking extends the same principle to the delivery address itself.

F. Taxis and ride hailing

A taxi or ride-hailing operator today retains, often for years, a complete record of a rider's identity, address, and movements. Location history is among the most sensitive categories of personal data – it reveals home, workplace, medical appointments, and personal associations – and its concentration inside operators has already been abused: Uber's internal "God View" tool, which displayed riders' real-time locations to staff, prompted an investigation and settlement by the New York Attorney General [20].

Under Databanking, a booking generates a token that the operator passes to the driver, and that materialises into a pair of addresses $[A, B]$ on the driver's dashboard. Once the ride is completed and paid, the addresses are deleted; the full ride record is written to the rider's Databank. The operator retains only the date, the payment received, and a link token. Unlike the shipping case above, where the seller never sees the address at all, the driver necessarily sees the pickup and destination for the duration of the ride; the mechanism here is ephemerality and retention-minimisation rather than complete concealment – an illustration that different use cases call for different token semantics.

The link token preserves the residual legitimate uses of the deleted data through bit-scarce queries (Section IV C). An auditor can ask "did this ride take place?"; the operator itself can ask, under the rider's standing authorisation, "did this rider travel from A to B within the last two weeks?" in order to offer, say, a discounted return ride to the airport. Both questions warrant a binary answer, not the rider's full ride history.

Retention obligations point the other way in some jurisdictions: Transport for London, for example, requires private hire operators to keep booking records for twelve months [21]. The link-token arrangement can satisfy the audit purpose such rules serve – each retained record remains verifiable against the rider's Databank – under the regulatory retooling described in Section III.

Finally, tokens – here and in the delivery case above – need not be indefinitely valid: an expiration date can be encoded into the token itself, so that neither the token nor the link it establishes outlives its purpose.

G. Data brokers and the right to erasure

Deletion requests today carry a paradox: to ask a company to erase one's personal data, one must first hand over that very data – enough of it, at least, for the company to locate the matching record. Commercial

removal services inherit rather than solve the problem, since they themselves become custodians of their subscribers' personal information. California's Delete Act platform (DROP), which since January 2026 lets a resident issue a single deletion request to all registered data brokers [22], goes further: the deletion lists brokers retrieve carry the requester's identifiers in hashed form. Deterministic hashes of low-entropy identifiers such as email addresses or phone numbers remain enumerable, however – hashing candidate identifiers suffices to test membership – and the match itself runs inside each broker's own systems, unobserved. A deletion list, even hashed, is a testable register of precisely the privacy-conscious.

Under Databanking, an Owner can broadcast an erasure request without this exposure. Each company receives a deliberately lossy fingerprint of the Owner's identifying data – a salted locality-sensitive hash, say – sufficient only to pre-filter candidate records, together with a link token. For each candidate, the company asks the Databank "does this record match this user?" and receives a single bit; on a match, it deletes the record. The company never obtains the data itself: its records meet the Owner's data only inside the sandbox. Established primitives for private matching – private set intersection and privacy-preserving record linkage [23] – already provide the machinery.

Two caveats apply. A dishonest company could attempt to use match queries as an oracle for locating individuals; the same discipline that governs every sandbox query – authorisation, logging, per-query fees, and the bit budget of Section IV C – bounds what such probing can extract, and match queries are accepted only for candidate records that the sandbox verifies against the fingerprint, not for arbitrary identities. And the Databank can confirm the match but cannot verify that deletion actually occurred; that gap is closed by legislation and penalties, as under the Delete Act, not by the protocol – the position today's removal services are in as well.

This use case doubles as a transition mechanism: opening a Databank account would naturally be followed by a broadcast request to erase one's data everywhere else, seeding the migration from the current regime of dispersed copies to the custodial one proposed here.

VIII. SUMMARY

We have proposed a mutually beneficial solution to the privacy-versus-convenience dilemma that currently dominates data-privacy discussions. This approach builds on, and goes further than, GDPR-style regulation of data use, and is closely related to – while extending – existing partial implementations: vendor relationship management, data trusts and cooperatives, self-sovereign identity and personal data stores, and the EU's nascent regime for data intermediaries. Our proposed model incentivises the creation of Databanks by offering financial rewards, grants Owners complete control over their Personal Information, and adds two ingredients we believe are still missing from the existing literature and regulation: a hard legal wall between custody and analysis, and bit-scarce sandboxed computation as the only channel through which third parties may learn anything about an individual's data. In doing so, this proposal aims to empower users while fostering a privacy-centric industry, striking a balance between privacy and usability in the digital age.

Acknowledgments

This paper was first conceived in 2013. The availability of modern generative AI systems, particularly Claude and ChatGPT, proved instrumental in bringing the project to completion by assisting with literature review, figure preparation, and the organisation of the material into its present form.

[1] D. Searls, *ProjectVRM*, Berkman Klein Center for Internet & Society, Harvard University (2006–). `projectvrm.org`
[2] Solid Project, *Solid: Your data, your choice.* `solidproject.org`
[3] Personal Data Stores (PDS): A Review, *Sensors* **23**, 1477 (2023). `https://www.mdpi.com/1424-8220/23/3/1477`
[4] F. Schardong and R. Custódio, "Self-Sovereign Identity: A Systematic Review, Mapping and Taxonomy," arXiv:2108.08338 (2021); Sensors 22(15), 5641 (2022).
[5] On the Compliance of Self-Sovereign Identity with GDPR Principles: A Critical Review, arXiv:2409.03624 (2024).
[6] Open Data Institute, *Defining a 'data trust'* and *Data trusts in 2020.* `theodi.org`
[7] T. Hardjono and A. Pentland, Data Cooperatives: Towards a Foundation for Decentralized Personal Data Management, arXiv:1905.08819 (2019).
[8] Data Cooperatives: Democratic Models for Ethical Data Stewardship, arXiv:2504.10058 (2025).
[9] E. A. Posner and E. G. Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society* (Princeton University Press, 2018).
[10] J. Lanier, *Who Owns the Future?* (Simon & Schuster, 2013).
[11] Governor G. Newsom, California State of the State Address (Feb. 2019); reported in, e.g., CNN Business, "Companies use your data to make money. California thinks you should get paid" (Feb. 13, 2019).
[12] Regulation (EU) 2016/679 (GDPR), Article 20; see also discussion in *International Data Privacy Law* **9**, 173 (2019).

[13] Regulation (EU) 2022/868 (Data Governance Act), in force since September 2023. https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[14] Healthcare Denmark, *Digital Infrastructure*; see also Danish Shared Medication Record, in *J. Med. Internet Res.* (2023), describing the federated, query-based architecture of *sundhed.dk*. https://healthcaredenmark.dk/national-strongholds/digitalisation/digital-infrastructure/

[15] Enabling Privacy-Preserving, Compute- and Data-Intensive Computing using Heterogeneous Trusted Execution Environment, arXiv:1904.04782 (2019).

[16] A Scalable, Privacy-Preserving Decentralized Identity and Verifiable Data Sharing Framework based on Zero-Knowledge Proofs, arXiv:2510.09715 (2025); see also Brave, "The limits of zero-knowledge for age-verification" (2025).

[17] Apple Inc., *About iCloud Private Relay* and *iCloud Private Relay & Privacy*. https://support.apple.com/en-us/102602

[18] Yoti, *Age Verification Privacy Policy* (age tokens). https://www.yoti.com/privacy/age-verification/

[19] TransUnion, *Hard vs Soft Inquiries: Different Credit Checks*; see also Experian, *What Is a Soft Inquiry?* https://www.transunion.com/blog/credit-advice/the-difference-between-hard-and-soft-credit-inquiries

[20] New York State Office of the Attorney General, "A.G. Schneiderman Announces Settlement with Uber to Enhance Rider Privacy" (Jan. 2016). https://ag.ny.gov/press-release/2016/ag-schneiderman-announces-settlement-uber-enhance-rider-privacy

[21] Transport for London, *Private Hire Operator Register and Booking Records Guidance* (booking records to be retained for 12 months from acceptance). https://content.tfl.gov.uk/phv-operator-register-and-booking-records-guidance.pdf

[22] California Senate Bill 362 (Delete Act, 2023); California Privacy Protection Agency, *Delete Request and Opt-out Platform (DROP)*, live since January 2026. https://privacy.ca.gov/drop/

[23] D. Vatsalan, P. Christen, and V. S. Verykios, A Taxonomy of Privacy-Preserving Record Linkage Techniques, *Information Systems* **38**, 946 (2013).

[24] L. Sweeney, k-Anonymity: A Model for Protecting Privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**, 557 (2002).

[25] C. Dwork, Differential Privacy, in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, Lecture Notes in Computer Science vol. 4052, pp. 1–12 (Springer, 2006); see also C. Dwork, F. McSherry, K. Nissim, and A. Smith, Calibrating Noise to Sensitivity in Private Data Analysis, in *Proceedings of the 3rd Theory of Cryptography Conference (TCC 2006)*, pp. 265–284.

[26] IDC, *Global DataSphere* forecasts; see also IDC/Seagate, The Digitisation of the World: From Edge to Core (2018), projecting global data creation, capture, and replication from 45 zettabytes in 2019 to 175 zettabytes by 2025.

[27] World Bank, Broad money growth (annual %), World Development Indicators, indicator FM.LBL.BMNY.ZG; Federal Reserve Bank of St. Louis, FRED series M2SL, M2 Money Stock. The 6–8% figure cited in the text is an approximate author calculation from broad-money growth series for major economies, excluding crisis-period monetary expansions.