

Data as Capital: Legal Ownership, Privacy, and Competitive Advantage in Digital Entrepreneurship

¹Dr. Vanadna Raghava, ²Mr. Sunny Seth, ³Dr. Sushil Kumar, ⁴Deepika Taparia

¹Professor, Jagan Institute of Management Studies

²Assistant Professor, Jagan Institute of Management Studies

³Professor, Jagannath University, Bahadurgarh

⁴Associate Professor, Jagannath University, Jaipur

Abstract:

Data is often called capital, an asset and even one of a firm's possessions. This paper proposes that all three explanations are wrong in relevant respects that impact the creation, funding and regulation of digital ventures. The four defining attributes of a capital asset non-rival, excludable, separable, transferable are missing from data: it is not excludable, it is not separably valuable, it is not transferable, it is not owned, but held as a set of adjacent and revocable entitlements. It also has no conventional asset has – that is, it has liability proportional to its size and sensitivity. Derived from this we create two constructs. The first is the returns of “contingent control rents” which occur not in the act of ownership but rather in the act of control, which exists only as long as consent, access and regulatory permission exist, all of which are subject to cancellation by those other than the owner. The second is the data holding problem where optimal holdings are interior to maximal holdings due to the saturation of the marginal predictive value and the increase of the marginal liability, which provides a self-interested, even if not only the ethical argument for data minimisation. We then go on to distinguish four types of data advantage and make the case that there are only two types that are defensible for entrants, and explore the now extensive evidence that the presence of consent-based privacy regimes has a salience effect which is more likely to favor incumbents than to be driven by compliance costs. There are seven propositions also provided, and an analytical protocol. No “made up” numbers are presented.

Keywords: data economy; data ownership; privacy regulation; competitive advantage; digital entrepreneurship; intangible assets; data minimisation; market concentration

1. Introduction

Data has gone from metaphor to assumption in planning the digital economy. Ventures are worth the data they'll be gathering, investors ask questions about proprietary data, and companies are encouraged to gather as much as possible on the grounds that it might come in very handy later. The vocabulary supports the assumption: Data is called an asset, it is owned, it is a moat, it is a bankable asset.

Not much of this is preserved for examination. There is no property right in data in most legal systems; facts are not owned, and what a firm has are a bundle of adjacent rights: the contract rights, the rights to trade secrets, whatever IP rights subsist in the structure but not content, or the de facto rights of access. (Determann, 2018; Purtova, 2015) Data is non-rival in the sense that one party's use does not affect the availability of the data to another party, and the economic implications of the non-rival use of data are significant and well understood (Jones & Tonetti, 2020). It's not separably valuable because it requires complementary capabilities and the decision context where it is used. And it isn't a line item in the accounts: generally, internally generated intangibles are not capitalised, and the principal asset of a digital venture is not a line item in that venture's balance sheet (Lev, 2001; Lev & Gu, 2016).

These are merely nitpicking criticisms. All have a direct impact on entrepreneurial practice. Where there is no title, a data position is considered a control position, and control positions can be lost in ways that don't involve a change in title, such as withdrawing consent, changing platform policy, regulatory reclassification, or a

counterparty leaving the platform that provides the access. However, if data is non-rival, then exclusivity is only based on the requirement to be secretive and to have control of access rather than on any legal exclusion, which is far less secure than ownership suggests. And, if the asset isn't in the accounts, then the financing of the firm, its valuation, and the internal capital allocation processes are all completed without a metric of the thing said to be at their core.

The paper then develops a property of data which has no analogue with conventional capital. A machine that is not being used is losing value due to depreciation. The ongoing exposure of a data stock is continuing breach risk based on volume of data and sensitivity, continuing regulatory obligations arising from data-holding, continuing deletion and access requirements based on number of data subjects, and ongoing litigation exposure that may materialise long after the decision to collect the data. The stock is at the same time an asset and a liability, and the best holding time is not the maximum but the interior. This gives rise to the data holding issue that was discussed in Section 4, and in turn to a case for minimisation that is not based on any moral principle firms that hold all the data are at some point, which is often reached early, eroding value on their own.

Four contributions follow. We make a very clear distinction when the capital analogy applies and when it doesn't apply, and we define contingent control rents as a correct characterization of what data actually produces. We deontologise the duality of liability and assets, and obtain the interior optimum. We identify four sources of data advantage and determine which are defensible to an entrant but not to an incumbent, and show that the one most frequently sought is the least defensible. We discuss the competition-structure impact of consent-based privacy regimes and why it is not the compliance cost, but salience-dependent consent, that is the actual mechanism at play, and therefore the two imply quite different remedies.

2. Is Data Capital? Four Tests

2.1 The four tests

Capital, in the sense used when firms speak of a data asset, has four properties: it is excludable, it is owned under a title that persists against the world, it is separably valuable and therefore transferable, and it depreciates in a manner that can be estimated. Table 1 tests data against each.

Table 1. Four tests of capital-ness applied to data

Property	What it requires	How data performs	Consequence for the firm	Nearest true analogue
Excludability	Use by one party prevents use by another	Fails; data is non-rival and can be copied at near-zero cost	Exclusivity rests on secrecy and access control, not on law	Trade secrets rather than physical plant
Title	An ownership right persisting against third parties	Fails; no general property right in data exists in most systems	Position is a bundle of contract, secrecy, and de facto control	A licence or an access relationship
Separable value and transferability	Value realisable independently of the holder	Largely fails; value is contingent on complements and context	Cannot be sold at anything approaching internal valuation	Firm-specific know-how
Estimable depreciation	Predictable decline in service potential	Partially holds, but rates vary by orders of magnitude across types	Aggregate data volume is a	A portfolio of assets with

Property	What it requires	How data performs	Consequence for the firm	Nearest true analogue
			meaningless accounting unit	heterogeneous lives
Positive carrying value	Holding the asset does not itself create liability	Fails; holding generates breach, regulatory, and deletion exposure	Optimal holdings are interior, not maximal	No conventional analogue

The last row is the unparalleled row. Conventional assets have positive valued costs (carrying costs), such as storage, upkeep, and insurance. A contingent liability can be a held data stock with magnitude dependent on the size and sensitivity of the data stock, and which may significantly exceed the informational value of the marginal records used to create it. The capital analogy ought to be enough to make it questionable, but that's enough for Section 4. Data doesn't have to be worthless because of the four failures. They refer to the value being of a different sort, and the language of ownership systematically misunderstands it. The thing a firm holds is a position of control with rent-producing properties while it is held. We refer to these contingent control rents as returns that come with access and secrecy, rather than with title, and returns that can be terminated by parties that are not necessarily the renter, such as data subjects with withdrawal rights, platforms changing the conditions for access, counterparties terminating supply relationships, or regulators reclassifying a processing activity. It is not academic when deciding between ownership and non-ownership. The owner of a legal right to expropriation has a legal right and the holder of contingent control does not generally have a legal right.

2.2 The valuation and accounting problem

The combination of two of the four failures is particularly significant in its own right and deserves to be considered separately since it has an impact on the financing of digital ventures. Data is not separately valuable, nor is it recognised in the accounts, and the reason for the latter is mainly the former. The valuation problem is a particular type of a general problem in information markets. Without inspection, a buyer cannot know the value of a dataset, and after inspection the buyer has substantially acquired the dataset (Arrow, 1962). Markets for information, then, depend on complex arrangements of sampling, reputation, warranty, and staged disclosure that increase transaction costs and limit the scope of transactions (Bergemann & Bonatti, 2019). Another challenge unique to data is that this value is determined by two factors: a buyer's value of complements to the data, and the actions of the buyer. Both of these are not fully understood by either party ahead of time. This is the classic issue with an asset whose returns rely on co-specialised skills (Teece, 1986) and it involves any specific valuation being made about a specific pairing, not the asset.

The accounting consequence is then analyzed. As per prevailing standards, intangibles can only be recognised if they are identifiable, controlled and likely to generate future economic benefits and reliably measured and internal generated intangibles are generally considered to be not reliable and are expensed as incurred. The data of a venture is therefore not recorded on its balance sheet and the expenditure that produced the data is recorded as cost. The traditional criticism of accounting's inability to account for intangible assets is particularly relevant here because it is not only an inconvenience to measure that these intangibles are not being recorded but that the entire class of firms is understated in terms of the intangible assets they have on their books (Lev, 2001; Lev & Gu, 2016). There are three practical effects: Financing becomes narrative because the balance sheet doesn't say anything about the thing the pitch is about. Data spending is skewed as an expense, as it must compete with capitalizing investments. At exit, the valuation difference between what the founder thinks that the data has value and what the acquirer is willing to pay for the data is not a negotiating issue, but an inherent effect of an asset that will only be of value when combined with other assets that the two parties value differently.

3. The Legal Architecture: Control Without Title

3.1 The rights bundle

Because there is no data property right, firms assemble functional equivalents from adjacent doctrines. Table 2 sets out the components, and the assembly is instructive: each element protects something other than the data itself, and the gaps between them are where data positions are actually lost.

Table 2. The bundle of entitlements assembled in place of data ownership

Entitlement	What it actually protects	How it is lost	Strength for a startup	Principal gap
Contract with the data source	A promise of continued supply or permitted use	Termination; renegotiation on renewal	Weak where the counterparty is larger	Binds only the counterparty, not the world
Trade secret	Secrecy of the compilation and its derivation	Disclosure; independent derivation; employee mobility	Moderate; requires demonstrable protection measures	Evaporates on disclosure, however caused
Database rights, where they exist	Substantial investment in obtaining or verifying contents	Jurisdictional absence; scope limits on generated data	Variable; unavailable in most jurisdictions	Protects investment in compilation, not the facts
Copyright in structure	Selection and arrangement, not the underlying facts	Independent reproduction of the facts	Weak for most operational datasets	Facts are unprotected by design
De facto control through access	Practical exclusivity while the channel persists	Platform policy change; API deprecation	Often the strongest element in practice	No legal recourse when the channel closes
Personal data entitlements	Nothing belonging to the firm; these run to the data subject	Not applicable; they are constraints, not assets	A liability, not a right	Frequently misunderstood as conferring a firm-side right

The last row should be emphasized as it is prone to confusion in practice. Personal data regimes establish rights of access, correction, portability, objection and erasure which are directed at the person and not the data holder. This sort of regime has a negative effect on a firm's data position, rather than a positive one. A business whose description of compliance posture claims that it is proof of a powerful data estate has turned the relationship upside down – the regime itself is the revocability that makes the position contingent.

3.2 Why propertisation has been resisted

So far, proposals have been put forward for a data producer's right, which have mostly been ignored, and the rationale helps to get a handle on the issue. The establishment of exclusive rights to data would exacerbate the already existing information asymmetry, make it more difficult to reuse data and on which a lot of innovation

relies, and would be very difficult to define when it comes to data that is generated through an interaction between two parties (Drexler, 2017; Kerber, 2016; Zech, 2016). The economic case against is no less strident: data is non-rival, so exclusion is welfare-reducing compared to sharing, under a broad variety of circumstances, and the social value of a data set can considerably exceed its value to any particular data holder (Jones & Tonetti, 2020). Recent European instruments have thus shifted to access and sharing rather than to propertisation and have laid the conditions for data intermediation, and use rights, but not ownership. In practice for entrepreneurial companies, legal strategy should focus on the access relationship instead of the asset; for them, it is about who has access to the data, and when it can be terminated on notice, and what if it is terminated, what will happen to the data they gave.

4. The Liability-Asset Duality and the Data Holding Problem

4.1 The marginal condition

Think of the firm that is deciding on the quantity of data to store, as measured in volume weighted by scope and sensitivity. The incremental value of an extra record decreases and empirical studies of returns to data in production contexts have repeatedly demonstrated that the value of a longer dataset and a wider one is not particularly strong, with significant gains coming well before the very large datasets (Bajari et al., 2019; Varian, 2019). The marginal risk of an extra record, on the other hand, increases: the risk of breach is proportional to the number of records; the regulatory obligations to holding the records grow with the number; and the handling of rights requests is proportional to the number of data subjects. The result of the optimisation is shown in figure 1.

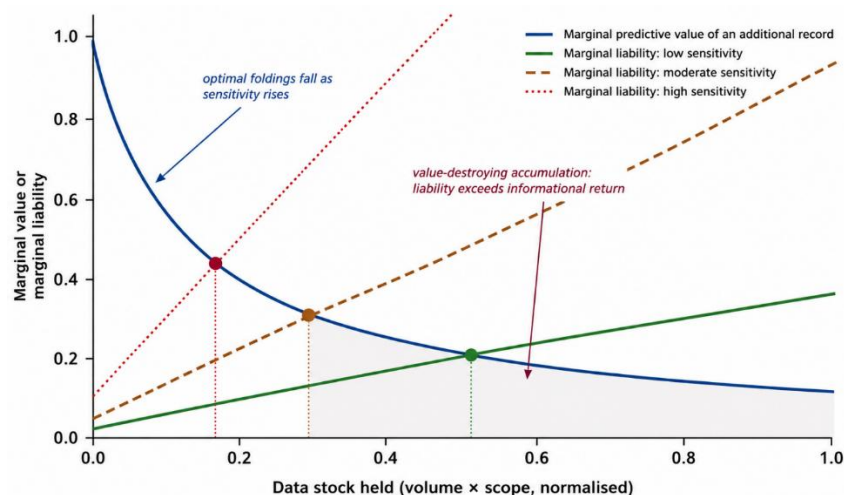


Figure 1. The difficulty of data holding. When the volume is increased, the interior optimum is to the left of the saturation point of the marginal value of the prediction, and the further out the point is, the greater the corresponding marginal liability. As sensitivity increases, the interior optimum moves increasingly to the left, to the saturation point of the marginal value of the prediction, and the further to the left the interior optimum lies, the greater the corresponding marginal liability. Values are derived from the declared parameters and represent the structure of the argument, but are not empirical estimates. There are three implications that come from this, none of which are the same as the ones that the collect-everything default takes for granted. The first is that the optimum is inwards and for sensitive categories, can be achieved very early on. The current trend of holding on to data forever for "some time in the future" is thus not a proactive approach, but a "who knows" liability on the balance sheet that is not priced, and an asset in the pitch book.

The second point is that when sensitivity increases, the optimal decreases, which is where the data firms that want the most the most information about the individual lie. This creates a systematic bias: you're right on value but wrong on net: that more valuable data should be collected more aggressively.

The third is that minimisation is typically said to be a regulatory requirement or ethical restriction, but on this analysis is simply a self-interested policy. A company that keeps only what it can prove is in use, discards it

according to measured decay and aggregates or de-identifies where the predictive value remains after the transformation is one that has not sacrificed value for compliance; it is operating at or closer to its optimal point. That the regulatory mandate and the self-interest policy are identical is a simple statement to make, since they are often described as being in opposition.

Table 3. Sources of negative carrying value in a held data stock

Source	Scales with	Timing of crystallisation	Insurable	Mitigation that preserves informational value
Breach exposure	Volume and sensitivity	Unpredictable; often years after collection	Partially; exclusions are broad	Aggregation; de-identification; segmentation of stores
Regulatory obligation attaching to holding	Number of data subjects	Continuous	No	Retention schedules tied to measured decay
Rights-request handling	Number of data subjects	Continuous, demand-driven	No	Automation; minimisation of subject count
Deletion and portability duties	Subject count and system complexity	On request or on schedule	No	Architectural design for deletability
Litigation and discovery exposure	Volume of retained records	On dispute	Partially	Defensible disposition policy
Acquisition diligence discount	Provenance quality and consent basis	At exit	No	Documented lawful basis and provenance records

The last row identifies a channel founders systematically underweight. At acquisition, a data stock with unclear provenance or a doubtful consent basis is not merely worth less; it can be a reason for the acquirer to exclude it from the transaction or to require its deletion, converting the venture's claimed principal asset into a remediation obligation at the moment of realisation.

4.2 Data externalities and the limits of individual optimisation

The statement above about the holding problem is from the standpoint of some firm, and there is no socially optimal solution to the problem. This is an externality peculiar to data: when people are correlated, one person's data contains information about other people who haven't provided it. You can have a lot of predictive power about the rest of the population and that's what a firm gets when it has data on a smaller subset of a population, and therefore, the individual who refuses has not protected themselves.

This has a very real impact on markets for data. Where the data of individuals is informative about other individuals, the price at which an individual is willing to give, is less than how much information is being lost to that individual; that is, data is systematically underpriced and over-collected relative to the social optimum even when it is consensually given (Acemoglu et al., 2022). The importance of this result is of interest because it challenges the standard regulatory reaction. In consent based regimes, the individual is the decision making entity, and in this context the externality means that no matter how well informed the individual is, his or her decision cannot be efficient. Under correlation, the differences between private and social values of data yield similar results from the economic studies of social data (Bergemann et al., 2022).

The critique of the privacy scholarship comes to a similar conclusion at a different point in the journey. Privacy self-management requires a vast number of decisions by the individual that they have no control over consent is thus a legitimating formality, not one of control (Solove, 2013); normative theories based on “contextual integrity” suggest that the right question is not whether consent was obtained, but whether there was compliance with the norms of the context where the information was created (Nissenbaum, 2010). In a more general context, the legal system of consent has been condemned as part of a legal frame that assumes the extractive practice it seeks to regulate (Cohen, 2019; Zuboff, 2019). There's no need to embrace the strongest of these arguments to embrace the narrower economic argument, that of course, is that individually optimal consent decisions don't add up to an efficient outcome in a regime that relies solely on them.

5. Where Data Advantage Actually Comes From

5.1 Four sources of advantage

If accumulated volume saturates and carries liability, the strategic question is not how much data a venture holds but what kind of position generates the rent. Table 4 distinguishes four sources and assesses each for defensibility, with particular attention to whether an entrant can hold it against an incumbent.

Table 4. Four sources of data advantage and their defensibility

Source	Mechanism	Why it may be defensible	Why it often is not	Defensibility for an entrant
Accumulated volume	More records improve model performance	Cost and time of replication	Returns saturate; rivals need only enough, not equal amounts	Low; the least defensible and most commonly pursued
Privileged access channel	Exclusive or preferential route to a data-generating process	Rivals cannot obtain the input at any price	Channel is usually contractual and therefore terminable	Moderate to high while the channel holds
Verified ground truth	Labelled outcomes that most parties never observe	Expensive to generate; often requires operating the process	Can be approximated by proxies in some domains	High where the venture operates the outcome-generating process
Data-enabled learning loop	Product improvement from use attracts use, which improves the product	Compounding and difficult to enter mid-cycle	Requires the loop to be fast and the improvement to be user-visible	High where present; frequently claimed and rarely real

Source	Mechanism	Why it may be defensible	Why it often is not	Defensibility for an entrant
Regulatory or consented exclusivity	A lawful basis rivals cannot obtain	Legally rather than technically protected	Consent is revocable and rarely exclusive	Low; contingent by construction

The practical difference between the first source and the rest is that. Research on when data is a source of competitive advantage has focused on the fact that it is not the number of units in the data store that matters, but the benefit that can be gained from using it, which has to be more for the next user than for the first one (Hagiu & Wright, 2020, 2023). The other side of the antitrust literature, however, is skeptical about the significance of data from the opposite perspective, suggesting that data is often non-rival, widely available, and quickly becomes outdated, so much of the competition-policy commentary concerning its role as a barrier to entry has been exaggerated (Lambrecht & Tucker, 2017; Tucker, 2019). A sceptical verdict or the worries voiced in the digital competition expert reviews (Cr  mer et al., 2019; Furman et al., 2019), but the analytical conclusion is essentially the same: volume is the weakest of all the positions available.

The underlying structure is provided in figure 2. Data types have a huge range of saturation and depreciation speeds; these two dimensions are not strongly coupled. The most accumulated and most unpleasant combination is undifferentiated behavioural logs which saturate early and decay rapidly. The best combination and the most expensive to produce is verified ground truth that saturates slowly and decays slowly. So the data strategy for a Venture is not about the volume of the cells it can fill in, but rather the portfolio question of which cells it can fill in.

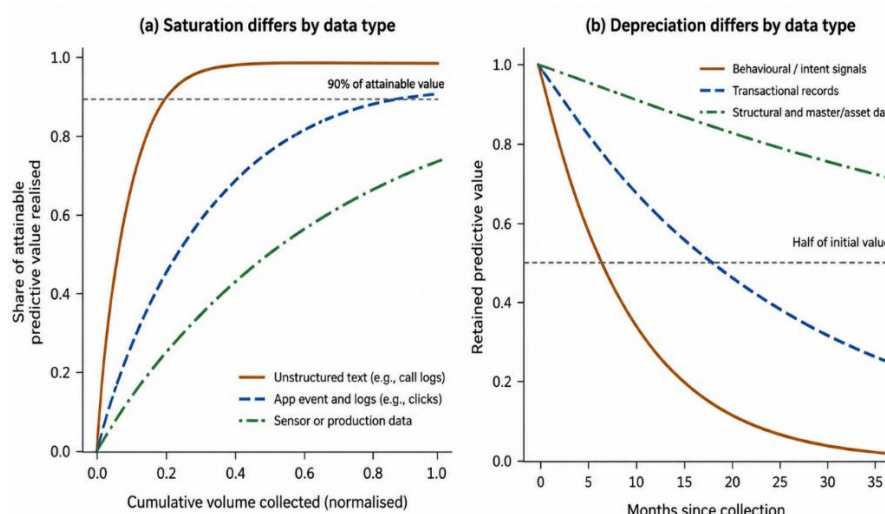


Figure 2. Panel (a) saturation of predictive value by data type. Panel (b) depreciation of predictive value over time by data type. All values are generated from declared parameters and illustrate the argument's structure only; they are not empirical estimates.

A corollary concerns the cold-start position of entrants. A venture without a stock cannot compete on accumulation and should not try, because the incumbent's stock is both larger and, on the saturation argument, mostly redundant. The viable entrant positions are those in the lower rows of Table 4: obtaining an access channel the incumbent lacks, operating a process that generates ground truth nobody else observes, or building a loop that compounds from a small base. Each of these is a design decision made early and is difficult to retrofit, which is why data strategy is properly a founding question rather than a scaling one.

5.2 Partial substitutes and their limits

Three developments are often touted as easing the above constraints, but each does so in a specific way and to a limited degree, and each requires a precise formulation of the easing:

Records created synthetically, but without any association to any real individuals. Where it does make sense, it takes care of the liability part of Table 3 directly because a synthetic stock has a lot less liability. Its upper bound is information: a generator trained with real data will be able to generate the same structure, but will not be able to generate information that is not in the training data; synthetic data can be used to augment the training and test sets, and for sharing, and is much less valuable in the case of very rare events or the long tail. It also doesn't completely remove exposure, as the generators can store and reveal training records.

In federated and distributed learning, computation is pushed to the data, not the data to the computation, which means a venture can learn from data which it doesn't store. This takes care of the holding problem beautifully: liability of Table 3 is connected to holding and thus it turns the question of the strategy from accumulating to accessing, as argued in Section 5. It does not impose any theoretical boundaries; it needs the full cooperation of parties who have the data, which makes it a return to bargaining power; it doesn't prevent any inference about individuals from the model that results.

The collective action problem is addressed by data intermediaries, trusts and pooled arrangements that attempt to combine data supplies under governance that individual suppliers can trust. The economic argument is robust because of the non-rivalry of benefits: pooling realised the social value that bilateral exclusion missed out on (Jones & Tonetti, 2020), and recent regulatory frameworks have shifted to create conditions for intermediation instead of ownership. The restriction is that intermediation only assists entrants when data that is pooled is not something the incumbent already has internally, as would be the least incentive for suppliers to pool.

The thing that's common is: each one of the substitutes relaxes one constraint, and not the other ones. None of these are defensible by accumulated volume, and all of them work better to deal with the liability side of the duality than to the value side.

6. Privacy Regimes as Competitive Structure Policy

Privacy regulation is a policy for individual rights and works, in large measure, as a competition policy. The theoretical case was presented before the current regimes came into existence: privacy restrictions increase the cost of acquiring data through arm's-length transactions as opposed to an existing relationship, and that benefits firms that already have a relationship and consumer recognition (Campbell et al., 2015). It is not primarily a compliance cost mechanism. It is consent. The structure is outlined in Figure 3. Consent rates are salience dependent, with people more likely to consent when the firm they are consenting to is familiar and has a prior relationship with them, and less likely to consent or ignore the request when the firm is not familiar. If both firms provide the same number of records at the same cost, while the accessible data ratio becomes bigger for a recognised firm than for an unrecognised firm, the ratio of accessible data will increase when the default changes from collection to consent. The result is a concentration effect created by user activity not regulatory burden, one not covered by the small-firm relief mechanisms typically provided by regulatory design.

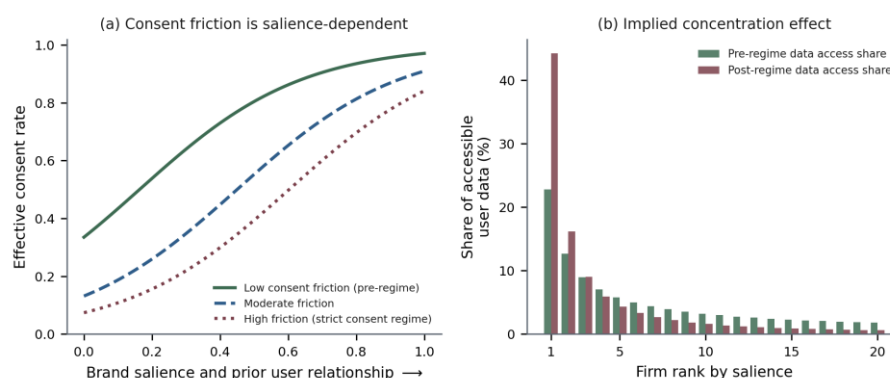


Figure 3. Panel (a) effective consent rate as a function of brand salience under three levels of consent friction. Panel (b) implied distribution of accessible user data across firms before and after a strict consent regime. All values are generated from declared parameters and illustrate the mechanism only; they are not empirical estimates.

There is a large empirical literature to this now. The European regime has been studied and found to lead to more concentration in the advertising tech vendor market, as larger vendors gained share (Johnson et al., 2023), a decrease in third-party services on websites, and a concentration of those that remain, as well as impacts on the data brokerage market consistent with more concentration (Aridor et al., 2023). More specific evidence specific to entrepreneurship is found in studies of venture investment, which show that the regime drove down financing levels for affected technology ventures after it was implemented (Jia et al., 2021), and studies of application markets, which find decreased number of new applications entering the market (Janssen et al., 2022). Of course, more comprehensive welfare assessments arrive at a mixed conclusion: there are intended welfare gains and unintended welfare losses (Goldberg et al., 2024).

Table 5 separates the mechanisms, because they imply different remedies and are routinely conflated in policy debate.

Table 5. Mechanisms by which privacy regimes affect competitive structure

Mechanism	How it operates	Who is advantaged	Addressed by small-firm relief?	Appropriate remedy
Consent salience	Recognised firms obtain consent at higher rates	Incumbents with direct user relationships	No; it is a user-behaviour effect	Neutral consent interfaces; limits on defaults favouring first parties
Compliance capability	Obligations require organisational structures, not only spending	Firms with existing compliance functions	Partially	Shared infrastructure; templates; graduated obligations
First-party advantage	Restrictions bind third-party data flows more than first-party collection	Firms with owned distribution and login relationships	No	Symmetric treatment of first- and third-party processing
Legal uncertainty	Ambiguous scope deters entry more than exit	Established firms with legal capacity	Partially	Guidance; worked examples; binding opinions
Enforcement salience	Regulators pursue visible firms, and small firms self-regulate against a distant threat	Ambiguous; can cut either way	Not applicable	Proportionate and predictable enforcement policy
Portability underuse	Rights intended to counteract lock-in are rarely exercised	Incumbents, by default	No	Machine-usable, continuous portability rather than one-off export

The last line indicates the best example of an instrument that is meant to inhibit concentration but has succeeded in doing nothing of the sort. Portability rights are in a number of regimes and are hardly used, but in part because being able to export once, in an unspecified format, does not allow for switching. Delegated portability, mechanical portability, and continuous portability would be a materially different instrument and it is the reform which this section points to which is most targeted at the mechanism.

There are two cautions against overstatement. First, the concentration results are disputed on scale and on any linkage to the privacy regime other than the one discovered in this study; many platform-level restrictions happened at the same time, and it is hard to tell how much impact a privacy regime has compared to the other changes that were going on. Second, and more importantly, if the result of a privacy regime is to increase concentration, it doesn't mean that the privacy regime should be avoided. It is an argument that the competitive consequences are real, foreseeable, and can be addressed through design, and that a regime that fails to address the competitive consequences will attain its rights goal and create a market structure that will reduce the practical exercise of the rights that the regime provides, because a concentrated market will give individuals less options to exercise the rights the regime provides.

7. Propositions and Analytical Protocol

Table 6 states seven propositions with testing strategies. Several are testable with data already held by firms, registries, or regulators.

Table 6. Propositions and testing strategy

Proposition	Statement	Testing strategy	Parameter of interest
P1	Marginal predictive value saturates well below the volumes firms retain, so retained stock exceeds the informational optimum in most ventures.	Learning curves estimated on held-out subsamples across firms	Saturation point [k1] relative to retained volume
P2	Optimal holdings decline in sensitivity, so the value-maximising stock is smallest for the most commercially attractive categories.	Joint estimation of value and expected liability by category	Elasticity [e2]
P3	Data positions are lost through channel and consent events rather than through competitor replication.	Case coding of observed position losses by cause	Cause distribution [d3]
P4	Advantage from a learning loop persists, whereas advantage from accumulated volume dissipates as rivals reach saturation.	Persistence analysis by claimed advantage source	Decay contrast [b4]
P5	Consent rates rise with brand salience, producing an access differential independent of compliance expenditure.	Field or panel measurement of consent by firm recognition	Salience gradient [g5]

Proposition	Statement	Testing strategy	Parameter of interest
P6	Ventures whose data provenance is documented realise higher acquisition values conditional on data volume.	Transaction analysis with provenance coding	Coefficient [b6]
P7	Portability rights are exercised at rates too low to affect switching, so their concentration-mitigating effect is negligible.	Exercise rates and subsequent switching behaviour	Exercise rate [r7]; switching elasticity [s7]

There are four cautions or warnings concerning measurements. Predictive value should be evaluated not for model fit statistics but against a decision-relevant criterion one that is relevant to the decision that the model is used to inform. Liability is not a point, instead it is a distribution and any estimate must have clear assumptions made about the probability of breach and the severity of the regulation; these assumptions should be reported, not embedded in the estimate. Comparisons of consent rates between different firms are difficult to interpret, because they are confounded with salience, an issue that can be addressed either through experimental manipulation or careful matching between firms on the dimensions of salience. Concentration literature is complicated by the possibility of regime changes that are in place at the same time as the concentration, and requires either variation across jurisdictions in the timing of regime changes, or variation across firms in exposure to the regime.

8. Discussion

8.1 Implications for founders and investors

Four implications follow. The first is that a venture should be able to tell what one of the four sources in Table 4 it rests on and that a response citing volume should be considered by both the founder and the investor to be a statement of no advantage. Second, data holdings should be treated as a portfolio of data, with clearly stated retention policies associated with measured decay, both as a matter of the value maximising policy of Figure 1 and because it is the position that comes out of an acquisition diligence process. The third is that legal action should be taken on the access channel, rather than on the asset: notice periods, change of control provisions and post termination retention rights in Supply Contracts preserve the position of the asset producing the rent, while attempts to assert ownership of data protect an asset that doesn't exist. The fourth is that the provision of the documentation of the provenance is not a cost of compliance, but a value add; as explained in the last row of Table 3.

8.2 Implications for policy

The analysis provides three policy observations. The idea of propertising data would be a bad idea for reasons that have emerged from the economic and legal discourses, while the trend towards access and sharing is more sound. Competitive consequence should be considered when designing privacy regimes, not once they are already in place, and the mechanism of salience outlined in Section 6 needs instruments that will result in an interface neutral approach, and that address the imbalance between first and third party processing, both of which are not “small-firm relief” measures. But if portability is to make a difference to switching, then it also needs to be redesigned around continuous machine usable access, the problem nominally addressed to it.

Another comment relates to jurisdictions other than those from which the majority of this evidence originates. Where a detailed data protection legal framework is a recent development, where localisation requirements can be seen as linking processing to territory and where there are limited digital public infrastructure systems which handle a large proportion of transactions, the former access-channel analysis of Section 5 becomes more relevant than the latter accumulation analysis, where access to the channel is administrative rather than commercial. In the

more mature enforcement regimes, the definitional breadth of the personal data is less relevant because it is offset by the fact that no one has to be prosecuted in enforcement. However, with less mature enforcement, a broad definition of personal data is more important because it leads to an increase in the number of ventures that are formally in scope but practically unadvised (Purtova, 2018). Consent salience is not, however, a competitive-structure question in such contexts, but rather a governance question of infrastructure – rather than personal privacy – and who is admitted to the channels of participation.

8.3 What would change the argument

The argument is refutable, and it is good to say what would refute the argument. The holding problem of Section 4 would lose its practical bite if it were discovered that there was no upper limit to the value of marginal predictive value and that, therefore, the additional records that a firm keeps would continue to show material improvement at the quantities that the firm is able to hold. The evidence to date suggests the opposite, but is based on a restricted class of prediction tasks, and it may not apply to situations where the value comes from rare events, especially generative applications, where previous studies on tabular prediction did not account for returns to scale over the range of factors considered.

If this is determined as being the case from data (i.e. lost data is mainly due to competitors copying it instead of to channel/consent events), the focus was incorrect in Section 2 and a more ownership framing of it is justified. If advantage from accumulated volume turns out to carry over for as long as advantage from learning loops, then the rank in Table 4 would go away, and the strategic recommendations that follow would be incorrect. If, but only if, the consent rates are found to be unrelated to the salience of the brand after controlling for interface design, then the mechanism proposed in Section 6 would be rejected and the concentration findings interpreted as being due to the compliance cost alone, thereby calling for small-firm-compliance-cost-relief measures rather than the interface and symmetry remedies proposed here.

All of these are estimable with the existing data infrastructure, and the last is the most consequential, as the two candidate mechanisms suggest remedies that do not overlap. A policy debate that does not differentiate them will pick up instruments intuitively and the instruments which address compliance cost are the very ones which a salience mechanism would make ineffective.

8.4 Limitations

There are three restrictions that must be clearly expressed. The propositions are not tested, and the figures are not meant to estimate any quantity but to show the structure of an argument; this is a conceptual paper. Second, the legal characterisation is based primarily on European and Anglo-American doctrine, and the position may vary at least significantly in jurisdictions that have sui generis data rights, sectoral legislation and/or data localisation obligations that impact access-channel analysis; legal positions in this area are also evolving rapidly and should be confirmed before relied upon. Third, the empirical literature on privacy regimes and concentration is active and controversial, and the summary presented in Section 6 does not quote magnitudes; readers are urged to refer directly to the cited studies, and recall the reasons for difficulty in identification presented in Section 7.

9. Conclusion

Data isn't capital, and the ramifications of claiming that's the case are real, not semantic. A lack of a title means a lack of control over what a venture contains as others can take it back. There is no exclusivity in terms of laws, but only of secrecy and access. No separable value exists and the asset doesn't get added to the accounts. The value-maximising holding of interior is the only thing that firms call an asset that actually creates liability as per size, unlike other assets, making the maximising of value from being minimising an advantage of not doing so from a business perspective.

All that is left after stripping away the metaphor is more serviceable than anything it promised. Yes, Ventures earn rents from data, but not from ownership, and not from data that is not specifiable: access channel competes with others but can't be granted; ground truth that others don't observe; loop that compounds from a small base. These can be built up, and they can be distinguished from the buildup strategy that most ventures try to engage in

and the saturation evidence shows that most of it is not used up. The regulatory environment is meanwhile affecting the competitive structure, in a fashion which was not intended by the regulator, and one which its small-firm provisions will not impact. The seven propositions and the analytical protocol proposed here are designed to be testable (and thus refutable) assertions rather than assertions.

References

- [1] Acemoglu, D., Makhdoui, A., Malekian, A., & Ozdaglar, A. (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4), 218-256.
- [2] Aridor, G., Che, Y.-K., & Salz, T. (2023). The effect of privacy regulation on the data industry: Empirical evidence from GDPR. *RAND Journal of Economics*, 54(4), 695-730.
- [3] Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity* (pp. 609-626). Princeton University Press.
- [4] Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. *AEA Papers and Proceedings*, 109, 33-37.
- [5] Bergemann, D., & Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11, 85-107.
- [6] Bergemann, D., Bonatti, A., & Gan, T. (2022). The economics of social data. *RAND Journal of Economics*, 53(2), 263-296.
- [7] Campbell, J., Goldfarb, A., & Tucker, C. (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy*, 24(1), 47-73.
- [8] Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press.
- [9] Crémer, J., de Montjoye, Y.-A., & Schweitzer, H. (2019). *Competition policy for the digital era*. European Commission.
- [10] Determann, L. (2018). No one owns data. *Hastings Law Journal*, 70(1), 1-44.
- [11] Drexler, J. (2017). Designing competitive markets for industrial data: Between proprietisation and access. *Journal of Intellectual Property, Information Technology and E-Commerce Law*, 8(4), 257-292.
- [12] Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). *Unlocking digital competition: Report of the Digital Competition Expert Panel*. HM Treasury.
- [13] Goldberg, S. G., Johnson, G. A., & Shriver, S. K. (2024). Regulating privacy online: An economic evaluation of the GDPR. *American Economic Journal: Economic Policy*, 16(1), 325-358.
- [14] Goldfarb, A., & Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1), 3-43.
- [15] Hagiu, A., & Wright, J. (2020). When data creates competitive advantage. *Harvard Business Review*, 98(1), 94-101.
- [16] Hagiu, A., & Wright, J. (2023). Data-enabled learning, network effects, and competitive advantage. *RAND Journal of Economics*, 54(4), 638-667.
- [17] Janssen, R., Kesler, R., Kummer, M. E., & Waldfogel, J. (2022). *GDPR and the lost generation of innovative apps* (NBER Working Paper No. 30028). National Bureau of Economic Research.
- [18] Jia, J., Jin, G. Z., & Wagman, L. (2021). The short-run effects of the General Data Protection Regulation on technology venture investment. *Marketing Science*, 40(4), 661-684.
- [19] Johnson, G. A., Shriver, S. K., & Goldberg, S. G. (2023). Privacy and market concentration: Intended and unintended consequences of the GDPR. *Management Science*, 69(10), 5695-5721.

- [20] Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9), 2819-2858.
- [21] Kerber, W. (2016). A new (intellectual) property right for non-personal data? An economic analysis. *Gewerblicher Rechtsschutz und Urheberrecht Internationaler Teil*, 65(11), 989-998.
- [22] Lambrecht, A., & Tucker, C. E. (2017). Can big data protect a firm from competition? *Competition Policy International Antitrust Chronicle*, January.
- [23] Lev, B. (2001). *Intangibles: Management, measurement, and reporting*. Brookings Institution Press.
- [24] Lev, B., & Gu, F. (2016). *The end of accounting and the path forward for investors and managers*. Wiley.
- [25] Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
- [26] Peukert, C., Bechtold, S., Batikas, M., & Kretschmer, T. (2022). Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science*, 41(4), 746-768.
- [27] Purtova, N. (2015). The illusion of personal data as no one's property. *Law, Innovation and Technology*, 7(1), 83-111.
- [28] Purtova, N. (2018). The law of everything: Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, 10(1), 40-81.
- [29] Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review*, 126(7), 1880-1903.
- [30] Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, 15(6), 285-305.
- [31] Tucker, C. (2019). Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, 54(4), 683-694.
- [32] Varian, H. R. (2019). Artificial intelligence, economics, and industrial organization. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 399-419). University of Chicago Press.
- [33] Zech, H. (2016). A legal framework for a data economy in the European digital single market: Rights to use data. *Journal of Intellectual Property Law & Practice*, 11(6), 460-470.
- [34] Zuboff, S. (2019). *The age of surveillance capitalism*. Public Affairs.