The Scale of Data

Information Spillovers and Investment Allocation in the AI Era

Sylvia Chu and Laura L. Veldkamp

Unpublished manuscript

Sylvia Chu: Columbia University, IEOR Laura L. Veldkamp: Columbia Business School

A quiz before we start

A startup sells an AI inspection system to factories.

It is paid **every time a factory adopts it**.

Before selling, it pilots the system at one plant. Careful logging of the pilot (by line, by batch, by condition) is **costly**, *and buyers* **cannot check** *it.*

Will the startup keep careful notes?

A quiz before we start

A startup sells an AI inspection system to factories.

It is paid **every time a factory adopts it**.

Before selling, it pilots the system at one plant. Careful logging of the pilot (by line, by batch, by condition) is **costly**, *and buyers* **cannot check** *it.*

Will the startup keep careful notes?

No. And that is perfectly rational.

Why it is rational

- It will sell to every plant **whatever the pilot shows**.
- So learning from the pilot before selling is worth nothing to it.
- Careful notes cost money and earn nothing.

Why it is rational

- It will sell to every plant **whatever the pilot shows**.
- So learning from the pilot before selling is worth nothing to it.
- Careful notes cost money and earn nothing.

The paper in one sentence
Paying a data producer for *deployment* fixes **whether** data are produced, not **how well they are recorded**.

Why it is rational

- It will sell to every plant **whatever the pilot shows**.
- So learning from the pilot before selling is worth nothing to it.
- Careful notes cost money and earn nothing.

The paper in one sentence
Paying a data producer for *deployment* fixes **whether** data are produced, not **how well they are recorded**.

Customers get the system, without the evidence on *where it works*.

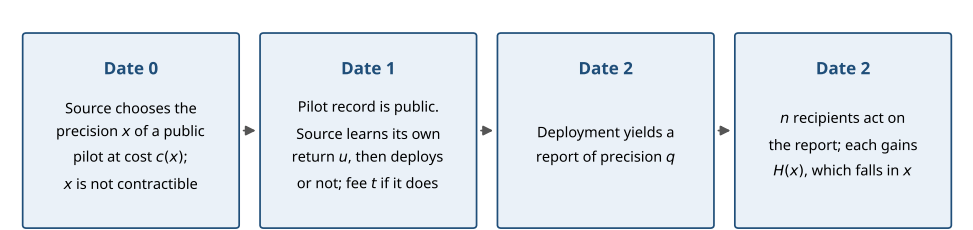
Why this matters

Data-economy models mostly treat data as a *quantity* that accumulates as a by-product of activity. Here recorded precision is a *costly choice*, and contracts pay for use, not for quality.

1. Data markets and platforms that pay per use may need to pay for **verified quality**. How good must verification be? (Prop. 3)
2. Logging and sharing standards matter *more*, not less, when payment is tied to deployment.
3. The shortfall **grows with the number of users**, where data should be most valuable. (Props. 1, 4)

All implications are conditional on the model, not evidence.

Core model (Gaussian)



- Source chooses precision x of a public record at cost $c(x)$; x is **not contractible**.
- Then decides whether to deploy; deployment yields a later report of precision q for n recipients.
- Recipient with quadratic loss values precision p at $-B/p$, so the report is worth $H(x) = \dfrac{Bq}{(p_0 + x)(p_0 + x + q)}$, **falling in** x: record and report are substitutes.
- Fee $t \in [0, n]$ is paid only on deployment.

Mechanism

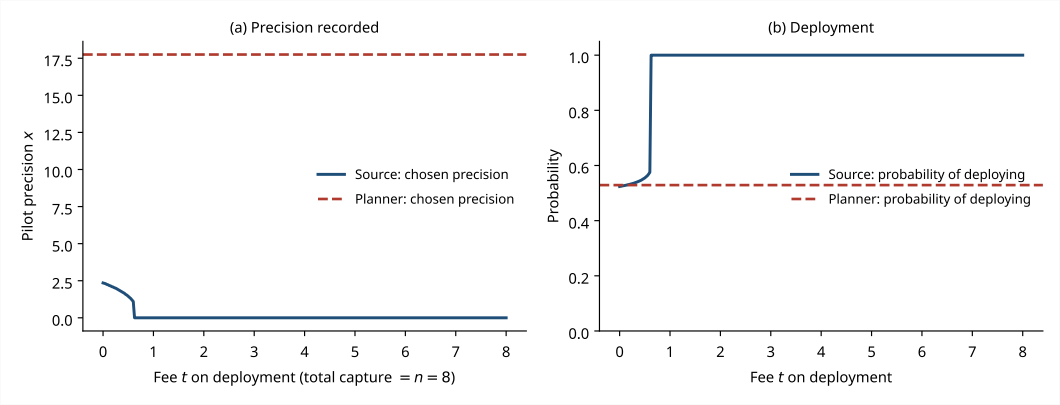
- A fee raises the return to deploying *whatever the record shows*.
- That removes the **option value of learning before deploying**.
- The record is then only a cost. The recipients' gain from precision, $\beta^2 n\,[w(x) - w(0)]$ with $w(x) = -B/(p_0 + x)$, is never paid.

Mechanism

- A fee raises the return to deploying *whatever the record shows*.
- That removes the **option value of learning before deploying**.
- The record is then only a cost. The recipients' gain from precision, $\beta^2 n\,[w(x) - w(0)]$ with $w(x) = -B/(p_0 + x)$, is never paid.

Proposition 1 (payment fixes deployment, not recording)
Let $s_0 \geq 0$. The source's chosen precision is nonincreasing in the fee t, and is zero for a large enough fee. The planner's precision is at least $\hat{x}(n)$, which for $c(x) = kx$ is $\sqrt{n\beta^2 B/k} - p_0 - q$, growing like $\sqrt{n}$.

One illustration ($p_0 = B = q = 1$, $\beta = 0.95$, $s_0 = 0.05$, $c = 0.02x$)



As the fee rises the source deploys more often and records **less**; it stops recording at a fee of about 0.63. Illustration, not an estimate.

No fee on deployment alone repairs both margins

Proposition 2
Suppose precision is not contractible and the planner's precision is positive. No fee $t \in [0, n]$ contingent on deployment induces the planner's precision *and* deployment rule.

- $t = n$: deployment is efficient given x, but x is strictly too low.
- $t < n$: deployment is too rare on a set of positive probability, or the source records nothing.

$n = 8$ recipients (one parameter set)	Social value	Precision
Fee 0	-1.60	–
Fee n (full payment)	-3.56	0
Planner	-0.34	17.75

Only differences are meaningful (Table 2 of the paper).

Leakage, resale, and what is not solved

Proposition 4 (leakage)
If the report can be passed on at no cost, so the source collects at most one recipient's gain, the planner's precision exceeds the source's by at least $\hat{x}(n) - x_S(0)$, which is **unbounded in** n.

- **Commit to one sale**: source can collect up to n recipients' gain (Prop. 1 at $t = n$).
- **Free resale among holders**: capture falls toward one recipient's gain.
- The two cases **bracket** the outcome. A resale game is **not solved**; recipients competing for a scarce slot are **not modeled**.

"Why is precision hidden?" A matter of degree

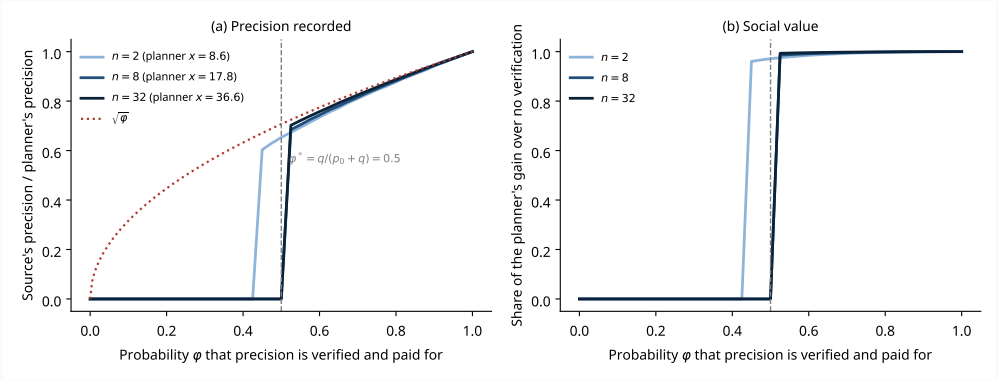
With probability φ recipients **can verify** x and pay what it is worth to them:

$$V^{\varphi}(x) = V_S(x; n) + \varphi\,\beta^2 n\,[w(x) - w(0)].$$

$\varphi = 0$ is Proposition 1 at full capture; $\varphi = 1$ is the planner.

Proposition 3 (partial verification)
(i) Source's precision is nondecreasing in φ. (ii) For $\varphi < 1$ it is strictly below the planner's. (iii) With $c(x) = kx$ and $\varphi^* = q/(p_0 + q)$: for large n, $\varphi < \varphi^* \Rightarrow$ the source records **nothing**; $\varphi > \varphi^* \Rightarrow$ it records something.

How good must verification be?



Parameters of Table 1, $n = 8$: $\varphi^* = 0.5$; at $\varphi = 0.6$ the source records 13.2 against the planner's 17.75, and social value is -0.37 against -0.34. The $\sqrt{\varphi}$ ratio is a heuristic checked numerically, not a theorem.

Second ingredient: the form of the record

Pooled write-up
"The combination cut defects sharply on some lines and not at all on others."

Linked write-up
"Large gains on short-run, high-mix lines; none on continuous lines."

Same outcomes. A high-mix plant learns something only from the second.

Which data to collect = whether outcomes stay attached to the *type* a reader can act on.
(Stylized vignette, not a documented case.)

Discrete model: what we get

- **Type correspondence has investment value** even when the distribution of returns is known (Prop. 5): a linked report is worth more than a pooled one with the same outcomes.
- **Deployment is never more frequent than the planner's**, and can be too rare (Prop. 6). Holds for any recipient decision problem (Prop. 14).
- **Paying for the later report aligns deployment, not the pilot** (Prop. 8). With priced reports the source captures only part of the gain and excludes some recipients (Prop. 9).
- **Scale**: the pilot-choice shortfall grows in proportion to the number of recipients in the example (Props. 10, 11).

The sign can reverse

- Gaussian case: value of information is concave, so pilot and report are always *substitutes*; payment lowers recording.
- Discrete case: the sign of the effect of payment on the informative pilot is the sign of the **curvature of the recipient's value of information** (Prop. 12).
- Complements arise when the recipient's decision threshold is far from the prior and signals are weak: then a fee on the report *encourages* the informative pilot.

Numerical check: substitutes fail in 12 of 48 combinations, all with $p^* \geq 0.7$.

Where this sits

- **Data as a stock or input**: Farboodi–Veldkamp; Farboodi–Koh–Xia; Teng. Quality of the record is not chosen against payment.
- **Data markets and externalities**: Bergemann–Bonatti–Smolin; Bergemann–Bonatti–Gan; Ichihashi; Acemoglu et al. Suppliers choose whether to sell, not how precisely to record.
- **Closest on costly precision**: Ozerturk (2004); Liu–Ma–Veldkamp (2025); Colombo–Femminis–Pavan (2026). Payment tied to use rather than to quality; public record; fee on deployment only.

Teng's full text was not available (abstract only). I found no paper with a public record, a fee on deployment and recipients who use the record; the search may have missed one.

What is proved, what is not

Proved

- Props. 1, 2, 4 (Gaussian core, $s_0 \geq 0$)
- Prop. 3 (threshold: linear cost, large n)
- Discrete-model results under stated assumptions

Not done / numerical only

- Competition for a scarce slot; resale game
- Precision hidden by a probability, not derived from a contract
- $\sqrt{\varphi}$ ratio; tables, figures; 3-type Monte Carlo
- No empirics. No peer has read it yet.

Where I would love your input

1. What is the most natural micro-foundation for unobservable precision?
2. Is there a clean competition or resale setting I should solve first?
3. Is the pay-per-deployment story the right motivating example?
4. Which closest papers have I missed?

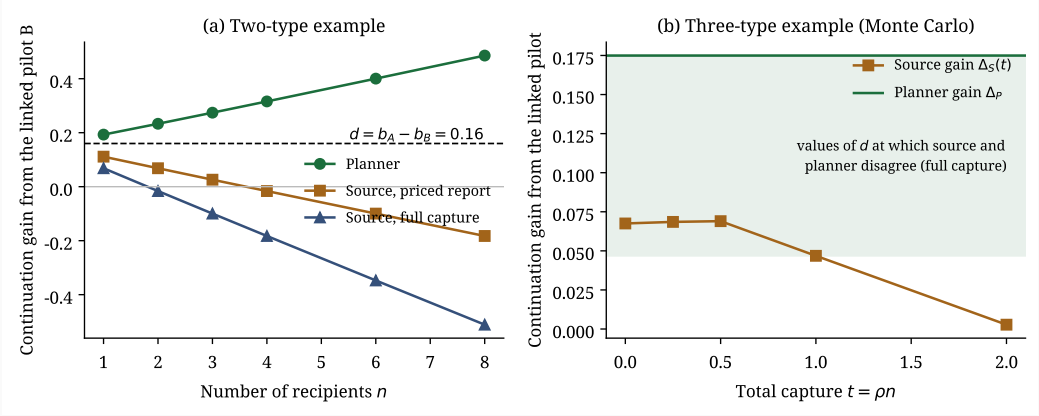
Thank you.

Backup: proof idea for Proposition 3(iii)

$V^{\varphi}(x) - V^{\varphi}(0) = -kx + n\beta^2 h(x) + \beta R(x)$ with $h(x) = H(x) - H(0) + \varphi[w(x) - w(0)]$, $R \geq 0$ the option-value term.

- $h(0) = 0$, $h'(x) = B[(p_0 + x + q)^{-2} - (1 - \varphi)(p_0 + x)^{-2}]$ changes sign at most once, from $-$ to $+$.
- $h(\infty) = \frac{B}{p_0}(\varphi - \varphi^*)$.
- $\varphi < \varphi^*$: $h < 0$ on $(0, \infty)$; R is bounded, so for large n the unique maximizer is 0.
- $\varphi > \varphi^*$: at $x = \sqrt{n}$, $n\beta^2 h \to \infty$ faster than kx, so 0 is not a maximizer.

Backup: the discrete model's scale example



Continuation gains from the linked pilot against the number of recipients; two-type and three-type examples, one parameter set each.