Privacy Laws and Value of Personal Data

Simona Abis[1], Mehmet Canayaz[2], Ilja Kantorovitch[3], Roxana Mihet[4], Huan Tang[5]

NBER Economics of Privacy

April 1, 2022

[1]Columbia Business School. Contact: sa3518@gsb.columbia.edu
[2]Smeal College of Business, Penn State. Contact: mcanayaz@psu.edu
[3]CFI, The École Polytechnique Fédérale de Lausanne. Contact: ilja.kantorovitch@epfl.ch
[4]Swiss Finance Institute at HEC Lausanne. Contact: roxana.mihet@unil.ch
[5]London School of Economics. Contact: huan.ht.tang@gmail.com

‘Data is to this century, what oil was to the last one.’

Source: The Economist, 2017

Voice-generated data and consumer privacy

- **State-of-the-art human-computer interaction. Potential to create novel, personal, and valuable data**
  - Voice-enabled products: In the kitchen, in the car, on the phone. Always listening, always responding, never forgetting
- **Research says data can create value for business**
  - Data lowers cost of capital for some firms (Begenau et al. (2018)) enabling them to produce more efficiently and grow larger (Farboodi et al. (2019), and Hagiu and Wright (2020)). Firms that invest more in AI grow faster (Babina et al. (2021))
- **Is consumer privacy achievable? If yes, what are the business ramifications?**
  - Digital capital is accumulated in few "superstar" firms (Tambe et al. (2020))
  - If the collection and trade of voice data is restricted, which firms are affected?

! **We are the first to examine the impact of consumer privacy laws on conversational-AI firms.**

Heterogeneous impact of consumer privacy laws on firm dynamics

- **We study businesses that collect voice-generated data on U.S. consumers**
  - Scraped all Amazon Alexa Skills between January 2016 and February 2022.
  - Entry into conversational-AI space; consumer ratings; comments; all merged with financial and accounting information from CRSP/Compustat.
- **Exploit a "local" shock: The California Consumer Protection Act (CCPA)**
  - The CCPA makes it harder to acquire (buy, collect) data on consumers.
  - Heterogeneous effects on firms with and without previously collected consumer data.
- **We find**
  - Adverse effects on all firms.
  - Attenuated effects for firms with accumulated in-house data.
- **Build model of data generation and trade to rationalize findings**

Empirical Strategy: Explore staggered implementation of CCPA

$$Y_{i,j,t} = \alpha + \beta_1 \quad \text{CCPA Intro}_t \quad \text{In-House Data}_i + \beta_2 \quad \text{CCPA Effective}_t \quad \text{In-House Data}_i + \gamma_{i,j} + \phi_t + \epsilon_{i,j,t}$$

where

$$Y_{i,j,t} = \text{customer satisfaction on voice-AI product } j \text{ of firm } i \text{ at } t$$

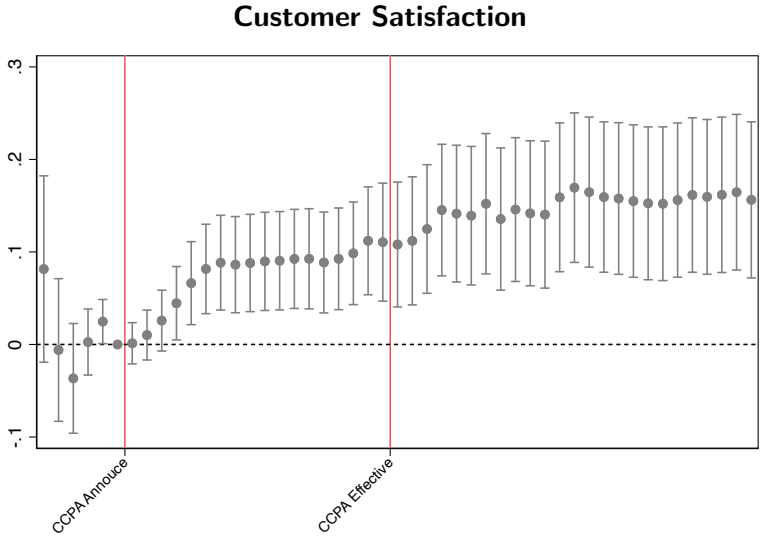
In-House Data_i = firm i's comments per product $>$ median value before the CCPA

- With firm-product FEs, we separate effects of the CCPA from contemporaneous shocks at firm & product levels
- With daily fixed effects, we eliminate the time trends

In-house data advantage after the adoption of consumer privacy rules

	Customer Satisfaction			
CCPA Announce x In-House Data	0.04	0.04	0.07**	0.06**
	(0.03)	(0.03)	(0.03)	(0.03)
CCPA Announce	-0.06**	...	...	...
	(0.03)			
CCPA Effective x In-House Data	0.11**	0.11**	0.14***	0.08*
	(0.04)	(0.04)	(0.04)	(0.05)
CCPA Effective	-0.15***	...	...	...
	(0.04)			
Product	Y	Y	Y	Y
Year-Month	N	Y	Y	Y
Product Category x Year-Month	N	N	Y	Y
Cohort x Year-Month FE	N	N	N	Y
N	216,250	216,250	216,179	216,179
R-sq	0.71	0.71	0.72	0.72

Dynamic diff-in-diff: superior ratings for firms with in-house data



Firm-level ramifications of in-house data advantage

	Tobin's Q	**CAPEX to Assets**	**Profit Margin**	**ROA**
	(1)	(2)	(3)	(4)
CCPA Intro x In-house Data	0.010	0.000	0.015**	0.001
	(0.13)	(0.07)	(2.17)	(0.63)
CCPA Effect x In-house Data	0.052**	0.003**	0.030***	0.004*
	(2.45)	(2.79)	(3.47)	(1.89)
Controls				
Log(BVA), Log(BVA) Sq., Log(Age)	Yes	No	No	No
CF2AT, Leverage, Tobin's Q	No	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Industry × Year-Quarter FE	Yes	Yes	Yes	Yes
Adj. R-squared	0.921	0.686	0.681	0.720
Observations	2,398	2,116	2,078	2,109

Effect heterogeneity

- We concentrate on the period after CCPA became effective in 2020
- We estimate a linear parametric conditional average treatment effect (CATE) model
- We analyze treatment effect heterogeneity across different dimensions
  - Lobbying, customer base, product diversification, customer ratings

$$
\begin{aligned}
Y_i =& \theta(X) \quad \text{In-House Data}_i + g(X, W) + \epsilon_i \\
& \theta_{i,j}(X) = \theta(X)^{0} \quad \text{coef}_{i,j} + \text{cate intercept}_{i,j}
\end{aligned}
$$

where Y_i = Average outcome variable of firm/product i after 2020

In-House Data$_i$ = firm i's comments per product $>$ median value before the CCPA

Firm-level outcomes

	Tobin's Q	**CAPEX to Assets**	**Profit Margin**	**ROA**
$\theta(X)$	(1)	(2)	(3)	(4)
Reviews	0.032***	0.01**	0.0849***	0.22***
	(11.905)	(1.990)	(11.332)	(4.835)
Products	0.057	-0.3**	-0.053	-0.93
	(1.079)	(-2.060)	(-0.336)	(-0.979)
Lobbied CCPA?	-0.099	-1.75	-2.08	-4.03
	(-0.144)	(-1.45)	(-1.473)	(-0.348)
Rating	-39.252*	-33.42	-21.7091	31.06
	(-1.887)	(-1.160)	(-0.521)	(0.133)
CATE Intercept	0.074	2.13**	-0.535	-1.65
	(0.117)	(2.040)	(-0.425)	(-0.228)

Model of data generation and trade

- Firms produce intermediate goods combining labor and knowledge.
- Production of Knowledge is a two-step process:
  1. Data gathering / acquisition.
  2. Data processing.
- Separation of data gathering and knowledge production as a realistic feature: Algorithms and Data are two necessary components for generating knowledge.
- Study effects on firms that are good in one respect, but not the other.
- **Firms trade data subject to an iceberg transportation cost.**

Standard Household and Firms

- Households have log utility and supply specialized labor inelastically.
- Final good firms combine intermediate goods in a CES-fashion

$$Y = \left(\int_0^1 Y_i^{\frac{\sigma - 1}{\sigma}} di \right)^{\frac{\sigma}{\sigma - 1}}. \tag{1}$$

- Intermediate good firms combine knowledge and labor à la Cobb-Douglas

$$Y_i = l_i^{1-\alpha} K_i^{\alpha}. \tag{2}$$

Knowledge Production

- Firms generate knowledge combining labor from *data analysts* with a data bundle

$$K_i = \left(l_i^P\right)^{\gamma_i} D_i^{1-\gamma_i}. \tag{3}$$

- The data bundle is a CES-aggregate of *internal* and *external data*

$$D_i = \left(\left(D_i^I\right)^{\frac{\varepsilon-1}{\varepsilon}} + \xi\left(D_i^E\right)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{4}$$

  where $\varepsilon > 0$ is the constant elasticity of substitution and ξ 1 captures the relative importance of externally acquired data.

- Internal data is generated employing *data managers*

$$D_i^G = A_i^G \left(l_i^G\right)^{1-\phi}. \tag{5}$$

Data Trade

- Firms can decide to *share* data with other firms and earn p^D per unit shared.
- There are two main features of data trade:
  1. **Non-Rivalry:** Firms can keep a share ν *2* $[0,1]$ of all shared data.
  2. **Trade Frictions:** For any unit of data shipped, only a fraction $1 \quad \tau$ *2* $[0,1]$ arrives.
- Total internal data is

$$D_i^I = D_i^G \quad (1 \quad \nu) D_i^S, \tag{6}$$

- where D_i^S captures the number of firms with which firm i shares its data.

Firm Heterogeneity

- We consider firm heterogeneity in two dimensions:

  1. **Customer base:** Firms with more customers have an easier time gathering data:

  $$D_i^G = A_i^G \left(l_i^G\right)^{1-\phi}$$

  2. **Sophistication:** Firms that employ *machine learning* / *AI* analysis techniques analyse more data with less labor. However, these firms are also more reliant on access to data.

  $$K_i = \left(l_i^P\right)^{\gamma_i} D_i^{1-\gamma_i}$$

- Firms can be on a 2x2 grid:

	Large (A_H^G)	Small (A_L^G)
Sophisticated γ_S	Market Places	Tech-Startups
Unsophisticated γ_U	Data Vendor	Old Economy

Maximization Problem

- The full maximization problem can be written as

$$\max_{l_i, l_i^P, l_i^G, D_i^S, D_i^E} \quad Y^{\alpha_Y} K_i^{\alpha_K} l_i^{\alpha_L} \quad wl_i \quad w_P l_i^P \quad w_G l_i^G \quad \frac{p^D}{1 \quad \tau} D_i^E + p^D D_i^S \tag{7}$$

$$s.t. \quad D_i^S \; 2 \left[0, \frac{D_i^G}{1 \quad \nu}\right] \tag{8}$$

$$l_i, l_i^P, l_i^G, D_i^E \quad 0 \tag{9}$$

Comparative Statics τ

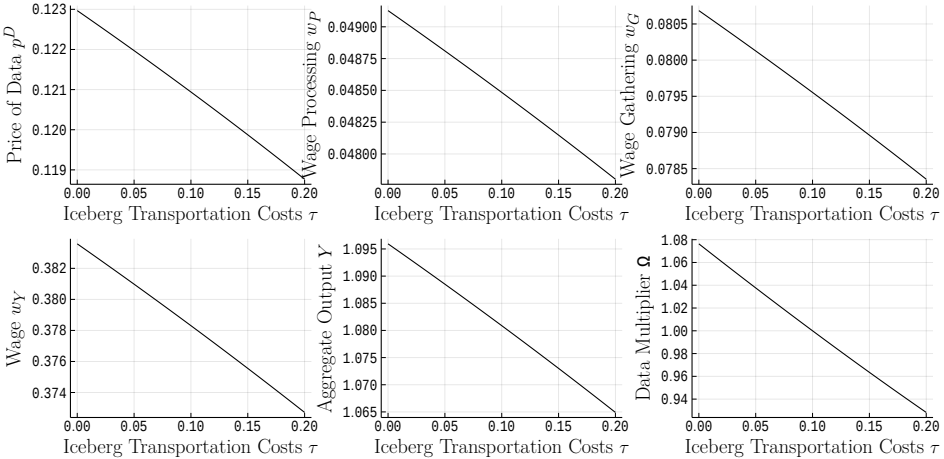


Figure: **Aggregate Responses to Higher** τ. An increase in the iceberg transportation cost τ leads to a fall in the price of data p^D, in wages w_Y, w_G and w_P, in lower aggregate output Y, and in the data multiplier Ω.

Comparative Statics, Firm-level τ

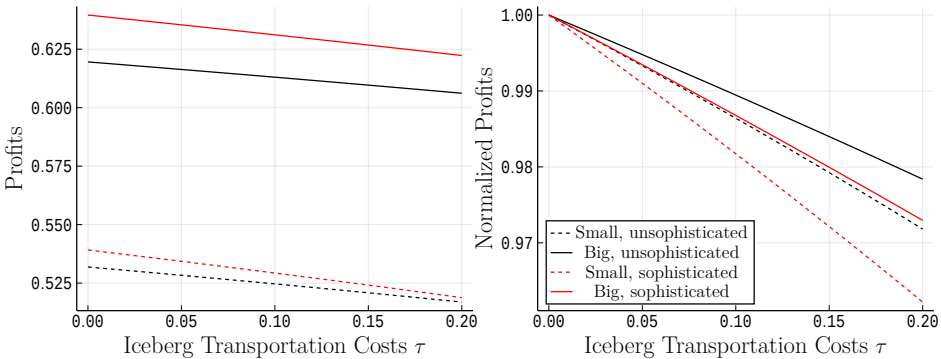


Figure: **Firm Valuations and Higher τ**: All firms become less profitable, but small, sophisticated firms are hit the hardest.

Results

- The effects of privacy regulation are *heterogeneous* and depend on firm-characteristics.
- Restrictions to the trade of data harm especially
  1. Sophisticated firms that are more reliant on data
  2. Small firms that cannot easily generate data
- **Implications for competition:** Adverse effects on small, sophisticated firms may deter entry and entrench large incumbents.
- Restrictions to the generation of data harm especially sophisticated firms.
  - **No anti-competitive effects!**

Conclusion

- We scrape the universe of Amazon Alexa Skills (Reviews, Ratings, Product Descriptions,...) and combine with CRSP / Compustat.
- Empirically, we find that the effect on all firms is negative, but less negative on firms that have collected data before the introduction of CCPA.
- Theoretically, we focus on two dimensions of heterogeneity
  1. **Size of Customer Base:** Generate data internally cheaply.
  2. **Sophistication:** Analyze large amounts of data, but also more reliant on data.
- Small, sophisticated firms are hit hardest from data sharing restrictions.
  - Anti-competitive effects are absent when data generation is limited.

Data Set: Hand-collect Info on Businesses Relying on Consumer Data

- New, hand-collected data on conversational-AI firms using web-scraping on Amazon Alexa Skills
- Skills are like applications for Alexa; They allow customers to use their voices to:
  - check the news; listen to music, play games
  - shop for goods or services
  - get personal recommendations
  - schedule transportation, etc.
- For each Skill:
  - firm identifying information
  - date and number of consumer ratings
  - the date of said rating
  - the average rating of said Skill over time

Data Set: Amazon Alexa Has Large Market Share

Figure: Alexa Sales ($ Million) Are Highest

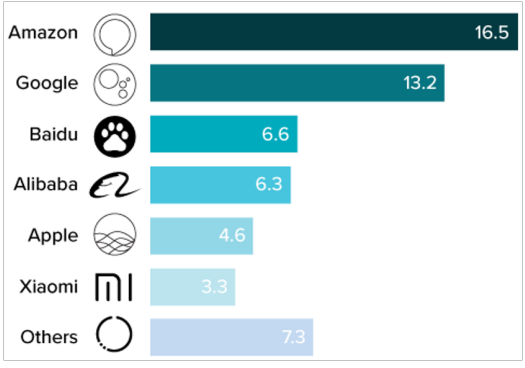
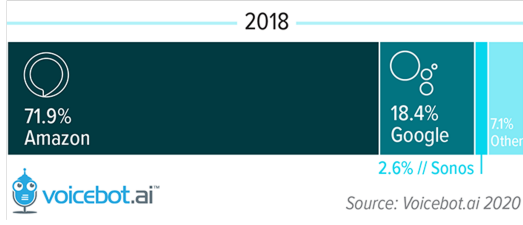


Figure: Alexa Market Share in US Is High



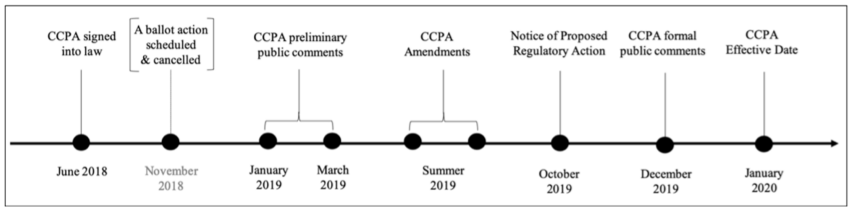
Summary Statistics

	N	Mean	Median	Std	P5	P95
Panel A: Voice-AI Products						
Panel A.1: All Firms						
Customer Satisfaction	335,661	3.47	3.50	1.01	1.60	5.00
Customer Reviews	335,661	28.60	6.00	59.00	1.00	174.00
Average Customer Reviews (firm level)	216,250	4.20	1.62	10.97	0.00	15.33
Panel A.2: Public Firms						
Customer Satisfaction	18,861	3.31	3.30	0.94	1.80	5.00
Customer Reviews	18,861	320.08	5.00	2501.63	0.00	662.00
Average Customer Reviews (firm level)	15,456	7.90	1.17	14.74	0.00	19.96

Data Set: CCPA as Shock to Data Use

- The California Consumer Privacy Act (CCPA)
  - A broad based law protecting information that identifies California residents
  - All companies that serve California residents, neglecting whether based in CA
  - Firms must have at least $25 million in annual revenue
  - Firms must have data on $> 50,000$ people or collect $1/2$ revenues from sale of data
  - Covers any info that could reasonably be linked (directly/indirectly) to a person:

▸ Details



Results consistent with prior research on GDPR

- **Empirically:** Peukert et al. (2021) and Campbell et al. (2015) show that data regulation such as GDPR creates barriers to entry and hurts competition.
  - Our innovation: Privacy regulation hurts competition by offering an advantage to firms with in-house data.

- **Theoretically:** Eeckhout and Veldkamp (2021) warn that big firms are using data to reallocate production to the goods consumers want most. As large firms already have this data, it is easier for them, relative to young, small firms, to tailor their products to consumers' preferences.
  - Our innovation: Show that it is firms with in-house data that benefit the most from data privacy regulations.

Bibliography

Babina, T., A. Fedyk, A. X. He, and J. Hodson (2021): "Artificial Intelligence, Firm Growth, and Industry Concentration," Tech. rep.

Begenau, J., M. Farboodi, and L. Veldkamp (2018): "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, 97, 71–87.

Campbell, J., A. Goldfarb, and C. Tucker (2015): "Privacy Regulation and Market Structure," *Journal of Economics & Management Strategy*, 24, 47–73.

Eeckhout, J. and L. Veldkamp (2021): "Data and Market Power," Columbia university wp.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019): "Big Data and Firm Dynamics," *AEA Papers and Proceedings*, 109, 38–42.

Hagiu, A. and J. Wright (2020): "Data-enabled learning, network effects and competitive advantage," Nus wp.

Peukert, C., S. Bechtold, M. Batikas, and T. Kretschmer (2021): "European Privacy Law and Global Markets for Data," Tech. rep.

Tambe, P., L. M. Hitt, D. Rock, and E. Brynjolfsson (2020): "Data-enabled learning, network effects and competitive advantage," Nber wp 28285.

Appendix: What Data Does CCPA Cover?

- The law includes detailed disclosure requirements, provides individuals with extensive rights to control how their personal information is used, imposes statutory fines and creates a private right of action.
- 'Personal information' defined much more broadly than any other U.S. privacy law
  - real name, physical address
  - biometric information I
  - address, online identifier
  - licence number, passport number, race
  - records of purchasing history or tendencies, internet browsing and search history
  - geolocation data, audio data
  - employment, or education data
  - as well as inferences drawn from these

Back

Maximization Problem

I The full maximization problem can be written as

$$\max_{l_i, l_i^P, l_i^G, D_i^S, D_i^E} \quad Y^{\alpha_Y} K_i^{\alpha_K} l_i^{\alpha_L} \quad wl_i \quad w_P l_i^P \quad w_G l_i^G \quad \frac{p^D}{1 \quad \tau} D_i^E + p^D D_i^S \tag{10}$$

$$s.t. \quad K_i = \left(l_i^P\right)^{\gamma_i} D_i^{1 \quad \gamma_i} \tag{11}$$

$$D_i = \left(\left(D_i^I\right)^{\frac{\varepsilon \quad 1}{\varepsilon}} + \xi \left(D_i^E\right)^{\frac{\varepsilon \quad 1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon \quad 1}} \tag{12}$$

$$D_i^I = D_i^G \quad (1 \quad \nu) D_i^S \tag{13}$$

$$D_i^G = A_i^G \left(l_i^G\right)^{1 \quad \phi} \tag{14}$$

$$D_i^S \; 2 \left[0, \frac{D_i^G}{1 \quad \nu}\right] \tag{15}$$

$$l_i, l_i^P, l_i^G, D_i^E \quad 0 \tag{16}$$

First-Order Conditions

- Denoting firm revenue as $\Pi_i = P_i Y_i$ leads to the FOCs:

$$(l_i) : \alpha_L \frac{\Pi_i}{l_i} = w \tag{17}$$

$$\left(l_i^G\right) : \alpha_K \left(1 \quad \gamma\right) \left(1 \quad \phi\right) \frac{D_i^G}{l_i^G} \frac{\Pi_i}{D_i} \left(\frac{D_i}{D_i^I}\right)^{\frac{1}{\varepsilon}} = w_G \tag{18}$$

$$\left(l_i^P\right) : \alpha_K \gamma_i \frac{\Pi_i}{l_i^P} = w_P \tag{19}$$

$$\left(D_i^S\right) : \alpha_K \left(1 \quad \gamma_i\right) \left(1 \quad \nu\right) \frac{\Pi_i}{D_i} \left(\frac{D_i}{D_i^I}\right)^{\frac{1}{\varepsilon}} = p^D \quad \lambda_i \tag{20}$$

$$\left(D_i^E\right) : \alpha_K \left(1 \quad \gamma_i\right) \xi \frac{\Pi_i}{D_i} \left(\frac{D_i}{D_i^E}\right)^{\frac{1}{\varepsilon}} = \frac{p^D}{1 \quad \tau} \tag{21}$$

Market-Clearing Conditions

- **Labor Markets**

$$\int_0^1 l_i di = \int_0^1 l_i^P di = \int_0^1 l_i^G di = 1 \tag{22}$$

- **Data Market**

$$\int D_i^S di = \int D_i^E di + \tau \int D_i^S di, \tag{23}$$

- Market-clearing in the good's market holds by Walras' Law:

$$w + w_P + w_G + \int_0^1 \pi_i di = Y. \tag{24}$$

where π_i are firm profits.

Data-Multiplier

- Due to the non-rivalry in the use of data, there is a difference between data *produced* and data *used*.
- The difference between both concepts is captured in the *Data Multiplier*

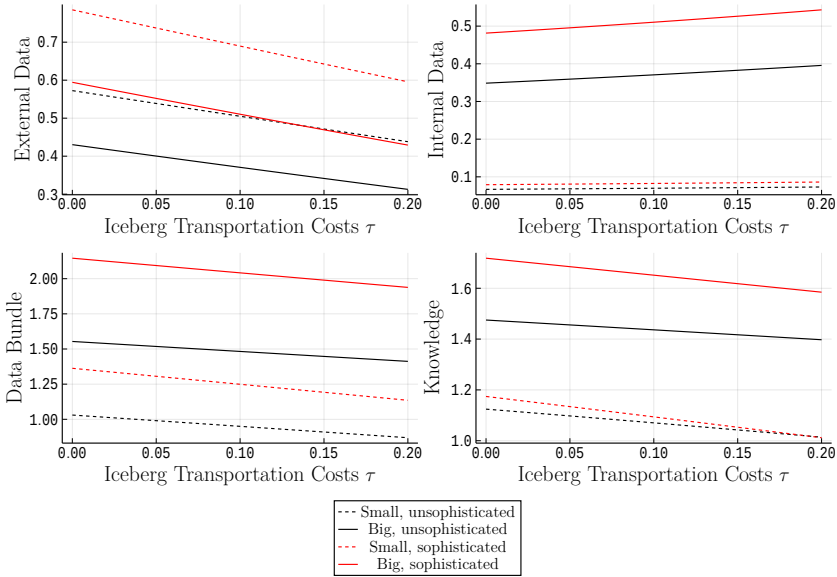
$$\Omega = \frac{D^U}{D^G} = 1 \quad (\tau \quad \nu)\, \lambda(\tau, \nu). \tag{25}$$

where

$$D^U = D^G \quad (\tau \quad \nu)\, D^S \tag{26}$$

$$\lambda(\tau, \nu) = D^S / D^G \tag{27}$$

Comparative Statics Data, Firm-level τ



Comparative Statics A_G

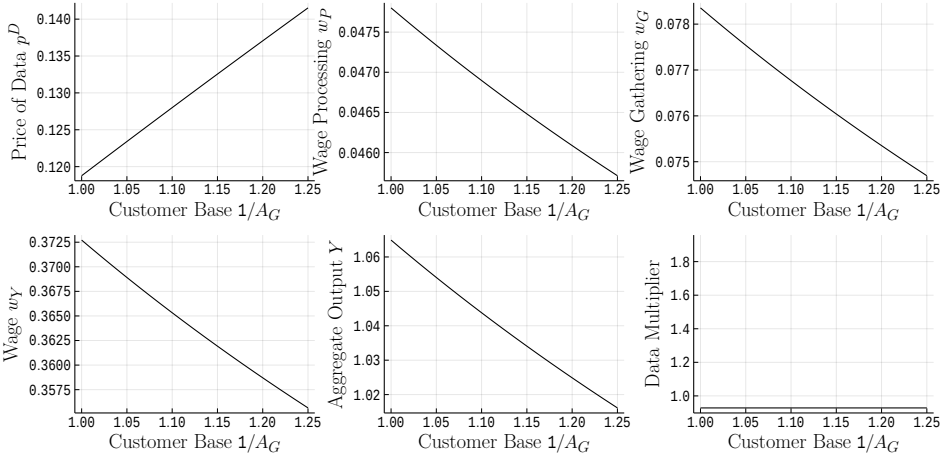


Figure: **Aggregate Responses to Lower A_G.** Less data generation increases the price of data and makes labor less productive. Data multiplier Ω is unaffected.

Comparative Statics, Firm-level A_G

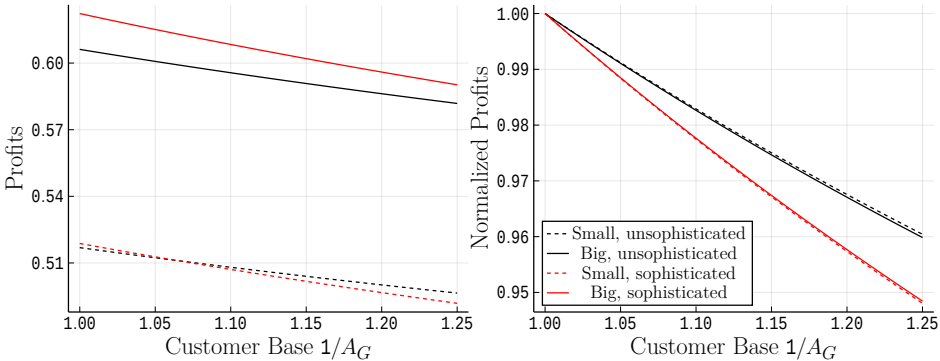


Figure: **Firm Valuations and Lower A_G**: All firms become less profitable, but both small and large sophisticated firms are hit the hardest.

Comparative Statics Data, Firm-level A_G