Feature balance of scale and scope of data in AI platform firms

Makoto Kimura
mkimura@nuis.ac.jp

Niigata University of International and Information Studies
Department of Business Administration, Faculty of Business and Informatics

Niigata University of International and Information Studies
3-1-1 Mizukino, Nishi-Ku, Niigata City, Niigata Prefecture, 950-2292, Japan.

Abstract

This study focuses on AI platform firms that, since the commercialization of deep learning with big data, have positioned and used the computational power of artificial intelligence (AI) as a core business function or mainstream product or service. This study argues for a cyclical structure that increases the scale and scope of data, enabling the exponential growth of AI platform firms. Therefore, we develop qualitative and dynamic models based on the scale and scope of data and investigate the mechanism of the exponential growth of AI platform firms. First, the simulation of AI platform firms was executed using a set of Julia packages, and the reproducibility of the execution results was verified using Vensim, a system dynamics development environment. Second, the sensitivity analysis of the dynamic model of AI platform firms was performed using the data network effect strength and the data sharing rate as parameters, and contour plots of the data boundary rate values as indexes of the scale and scope of the data were generated. Furthermore, through linear/nonlinear regression estimation that approximates the results of sensitivity analysis, we attempt to gain a qualitative and quantitative understanding of the feature balance between the scale and scope of data.

Keywords:
artificial intelligence, digital platforms, data network effects, scale and scope of data, Julia packages

Introduction

According to OpenAI, which launched an artificial intelligence (AI) service (ChatGPT) using large-scale language models in June 2020, "generative pre-trained transformers (GPTs) are general-purpose technologies" (Eloundou et al., 2023). General-purpose technologies can affect the entire economy, including printing, steam engines, and electricity, and are characterized by widespread diffusion, continuous improvement, and generation of complementary innovations (Gambardella and McGahan, 2010). Machine learning as a broad category is likely to be considered a general-purpose technology (Goldfarb et al., 2023). This study focuses on AI platform firms that positioned and used the computational power of AI as a core business function or mainstream product or service after 2016, when deep learning with big data became a practical application. AI platform firms develop and operate platforms with AI capabilities (Agrawal et al., 2022; Iansiti, 2021; Gregory et al., 2021) that can create value from the speed and accuracy of predictions based on collected user data. Such AI platform firms also serve as exponential growth organizations (Ismail et al., 2014) that can create a large order of magnitude (at least 10x) of value and impact compared to their competitors using new ways of operating organizations based on accelerating and evolving technologies. This study argues for a cyclical structure that increases the scale and scope of data, enabling the exponential growth of AI platform firms. Therefore, we develop qualitative and dynamic models based on the scale and scope of data and investigate the mechanism of the exponential growth of AI platform firms. Furthermore, through linear/nonlinear regression estimation that approximates the simulation results of the dynamic model of AI platform firms, we attempt to gain a qualitative and quantitative understanding of the feature balance between the scale and scope of data. The twentieth century has highlighted the importance of direct network effects (Katz and Shapiro, 1985, 1994): the more people remain connected to a network, the more valuable the network becomes. Stucke and Grunes (2016) indicate that data-driven firms working with big data are subject to multiple types of network effects: "(1) traditional network effects, including social networks such as Facebook; (2) network effects related to the scale of data; (3) network effects related to the scope of data; and (4) network effects where the scale and scope of data on one side of the market affect the other side (indirect network effects or cross-side network effects)." Meanwhile, Mayer-Schönberger and Ramge (2019) proposed the emergence of data capitalism and data-rich markets, where the driving factors of market behavior replace money with digital data. They presented data feedback effects as a new concept operating in data-rich

markets. In such markets, the combination of machine learning systems and matching algorithms in the processing of big data on product attributes and customer preferences improve experience value through customer decision support and transaction decision automation and enable product and service innovation (machine-based innovation) in addition to increasing customer data volume (data feedback effects). Iansiti (2021) also presents data quality, scale and scope, and uniqueness as key data value factors that firms can extract.

In this study, the perspective on scale and scope of data relates to other abstract concepts and terminologies (see Table 1). Clough and Wu (2022) indicated that data is a strategic resource within firms, and that firms should strategically decide whether to perform more value creation or value capture with data. Heimburg et al. (2023) also note that data sharing between competing platforms inhibits innovation. Conversely, Gregory et al. (2022), who proposed data network effects in AI-enabled platforms, emphasized the duality of value creation and value capture with data. Based on the arguments of these studies, the two categories of primary and secondary data, and the network effects of scale and scope of data in AI platform firms are addressed. This study focuses on closed data, such as customer data and trade secrets, as the scale of data for AI platform firms. Specifically, we treat the volume of primary data, which is the most important factor in AI platform firms, as the scale of data. Second, we focus on publicly-available open data or data shared with other firms as the scope of data for AI platform firms. I treated the size and type of secondary data to support the use of primary data in AI platform firms as the scope of data. The scale and scope of data in AI platform firms distinguished in this manner can be mapped to strategic dimensions of platform competition (Cennamo, 2021), which is the subject of previous research, and the concept of data network effects (Gregory et al., 2021, 2022), which are discussed in the next section.

Table 1. Concepts of scale and scope of data in this study

Cyclic logic of AI platform firm	Scale of data	Scope of data
Data type	Closed data	Open data (shared data)
Data Priority	Primary data	Secondary data
Key indicator of data	Data network effects strength	Data sharing rate
Model formulation	Adopters × Individual customer data	AI platform capabilities × Process innovation
Strategic dimension of platform	Platform identity	Platform size

competition (Cennamo, 2021)		
Main mechanism of data network effects (Gregory et al., 2022a)	User-centric design Data stewardship Platform legitimation	AI capability Data stewardship Platform legitimation

The remainder of this paper is organized as follows. First, we examine a qualitative model that visually illustrates how both the scale and scope of data in AI platform firms progress and how both prediction accuracy and platform value increase. Second, we develop a quantitative (dynamic) model based on the qualitative model of AI platform firms. From the set of parameters in this dynamic model, data network effects strength is selected as a key indicator of the scale of data, and data sharing rate is selected as a key indicator of the scope of data. The scale of data in this dynamic model is formulated as the product of the number of adopters (customers) of the AI platform company and the amount of data for each customer. The formulation of scope of data is a product of the AI platform capability and process innovation. The feature balance of the scale and scope of data for AI platform firms was examined through simulations using a dynamic model.

In this study, the data boundary rate is set as a measure of the feature balance of the scale and scope of data in AI platform firms. The data boundary rate is the proportion of the scope of data to the sum of the scale and scope of data. If the data boundary rate is greater than $\frac{1}{2}$, the scope of data is prioritized and the shared data constitute the majority of the total data volume of the AI platform firm. If the data boundary rate is lower than $\frac{1}{2}$, the scale of data is prioritized and the customer's proprietary data comprise the majority of the total data volume of the AI platform firm. We then simulate a dynamic model of the AI platform firms with the data network effects strength and data sharing rate as parameters, and create a contour map of the data boundary rate values from the calculated values. General linear and nonlinear regression equations are estimated to approximate these contours. Finally, by examining the properties of these regression equations, we attempt to gain a qualitative and quantitative understanding of the feature balance between the scale and scope of data for AI platform firms.

Previous studies related to this research

Data-enabled learning and data network effects

Hagiu and Wright (2020) presented the possibility that a virtuous cycle in which learning with customer-generated data (data-enabled learning) leads to customer benefits is a barrier to entry, which serves as a competitive advantage for existing firms. They also pointed out that if data-enabled learning improves products and services not only for individual users, but also for others, a data-enabled learning network effect will emerge and tend to strengthen traditional network effects. They also argued that data-enabled learning is possible even for firms that are unable to manifest network effects and that it can help reduce the churn rate of existing customers.

Gregory et al. (2021, p. 535) state that "a platform exhibits data network effects if, the more the platform learns from the data it collects on users, the more valuable the platform becomes to each user." According to them, the basic mechanism of the data network effect is that the platform's use of AI increases the scale of learning from the user data collected, thereby influencing the network through the platform and leading to an increase in user value. Specifically, user value from AI-enabled platforms is a function of the scale of AI-driven data learning and improvement. Their data network effects model was based on a positive direct relationship between the platform's AI capability and users' perceived platform value, which is a function of data stewardship, user-centered design, and platform legitimacy. Using this model, Gregory et al. argued that the relationship between platform quality, indirect network effects, and customer expectations in platform-based markets discussed by Zhu and Iansiti (2012) can also be explained using indirect and data network effect concepts. However, their model is similar in structure to a fishbone diagram, rather than the cyclical structure model used to explain network effects in the past. This can be a barrier to understanding data-network effects.

Dynamic model of data-driven firms

Prüfer and Schottmüller (2021) analyzed the equilibrium conditions of a dynamic model that considers the indirect network effects of data-driven user demand in R&D competition between two firms, and discussed monopoly avoidance through a firm monopoly and user data sharing. However, it did not consider differences in data types or a detailed study of simulation with parameter settings.

System dynamics (SD) is a system modeling and simulation technique proposed by Forrester (1961) that graphically models stock variables (accumulated variables), input and output flow variables acting on them, and feedback (cyclical action), among others, and simulation of time variation of variables (Sterman, 2000).

Ruutu et al. (2017) developed an SD model that includes platform development, indirect network effects, data network effects, and competitive efforts as explanatory variables and derived multiple scenarios of digital service platform competition through simulation. Their research interest differs from that of this study, which is the impact of user platform migration costs and user decision delays on accumulated data and platform financial resources. In addition, the SD model they developed only sets the presence or absence of data sharing among competing platforms as a policy flag and does not simulate a wide range of possibilities with the data sharing rate as a parameter. In this study, we support the definition formula for data network effects presented by Ruutu et al. and consider data network strength and data sharing rate constants as parameters for performing simulations and sensitivity analysis.

Qualitative model of AI platform firms

This section examines a qualitative model of AI platform firms using causal loop diagrams (Anderson and Johnson, 1997) that incorporate network effects related to the scale and scope of data. A causal loop diagram shows the behavior of a system in a circular structure, in which the relationships among the components of the system (among variables) are regarded as a chain of cause and effect, and the variables are connected to each other with arrows. This circular structure can be modeled as a combination of a virtuous circle of data-enabled learning on a platform (Hagiu and Wright, 2020) for network effects related to the scale of data and interface expansion (Gawer, 2021), an option for expanding platform boundaries, for network effects related to the scope of data. It also includes an AI algorithm update cycle, which is a software update in AI platform firms. Essentially, the qualitative model of AI platform firms can be expressed as a multiple cycle combining the virtuous cycles of data-enabled learning, data sharing through interface expansion, and AI algorithm update (see Fig. 1).

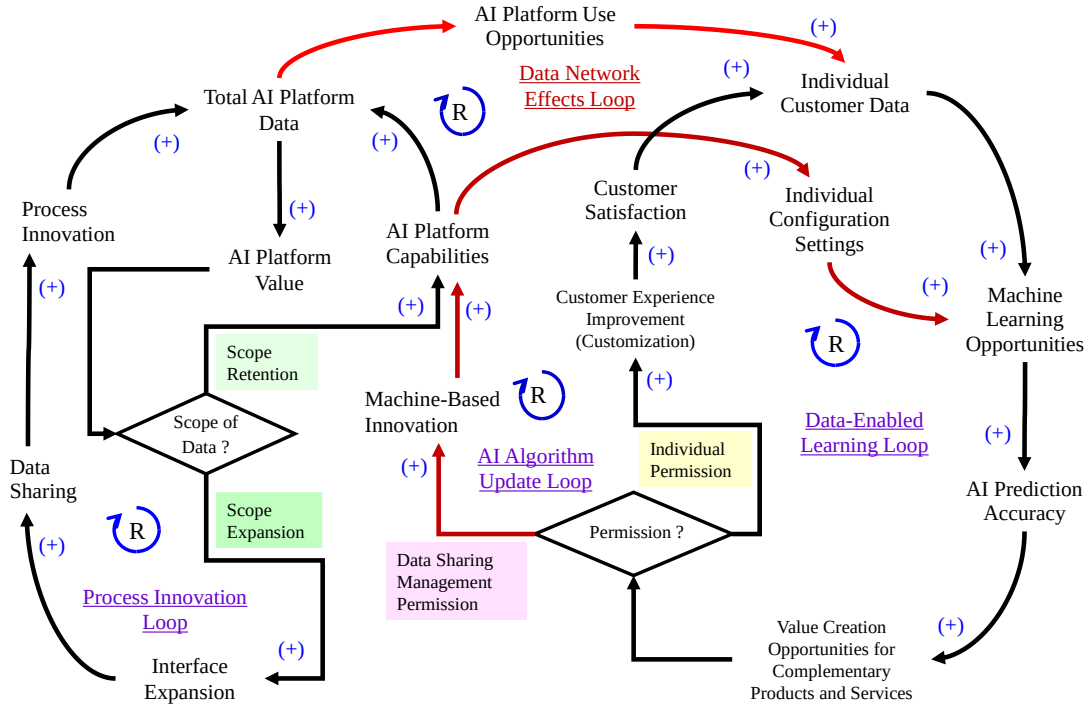

Fig.1 Qualitative model of AI platform firms

The data-enabled learning loop shows the cyclical logic of increasing the volume of individual customer data, opportunities for machine learning and value creation of complementary products and services through improved AI predictive accuracy. When customer data are licensed for individual customers only, customization improves customer experience under the terms of license. Customization improves customer satisfaction and increases the volume of individual customer data. By contrast, when customer data are licensed to other customers in addition to individual customers, the AI platform capability is improved through the realization of machine-based innovation under the license conditions of data sharing management.

AI platform firms can distinguish between machine-based innovation, which is the product of development and commercialization, and process innovation, which is the process adopted and commercialized by other firms. Not only the advancement of machine-based innovation through the expansion of the scale of data but also process innovation generated by the expansion of the scope of data may be the driving factor for the exponential growth of AI platform firms. In this qualitative model, we characterize the process innovation loop as a chain of process innovations realized through data sharing using interface expansion for expanding the scope of data. When AI platform firms decide to expand the boundaries of the AI platform, interface expansion and data sharing with other platforms and/or businesses will lead to process innovation. Owing to

process innovation, the total volume of data on AI platform firms increases further, enhancing AI platform customers' usage opportunities. An increase in customer usage opportunities in turn augments individual customer data. The process innovation loop expands the scope of data and process innovation in AI-platform firms. The AI algorithm update loop reinforces each of the above loops.

Dynamic model of AI platform firms

The growth and/or decline of AI platform firms appears to be a dynamic and complex phenomenon, resulting from multiple causal relationships. Simulations using a dynamic model are a natural choice as a theory-building method to explain such dynamic and longitudinal phenomena (Davis et al., 2007). SD is a dynamic modeling technique that allows graphical modeling and seamless simulation of multiple interacting processes, feedback loops, time delays, and other nonlinear phenomena (Sterman, 2000). Sterman proposed a formulation of expectation formation based on bounded rationality using the components of SD such as input flows, output flows, and the stock of accumulated differences between these flows. Using a modified formulation of expectation formation, this section develops a dynamic model based on a qualitative model of AI platform firms. Ruutu et al. (2017) developed an SD model of R&D and competition on digital service platforms that incorporated data network effects. In this paper, we utilize the formula for defining data network effects presented by Ruutu et al. to develop a dynamic model that can treat the constants of data network effect strength and data sharing rate as parameters.

Components of the dynamic model

The qualitative model of AI platform firms contains many components, whose causal and mechanistic formulations are difficult to identify. Examples include AI platform customer data accumulation, AI platform quality through machine learning opportunities for customer data, machine-based innovation for customer value creation, AI platform capabilities through machine-based innovation, and process innovation through data sharing. We formulate the dynamic changes in these variables as extrapolative expectations through first-order smoothing and link them to develop a dynamic model for AI platform firms.

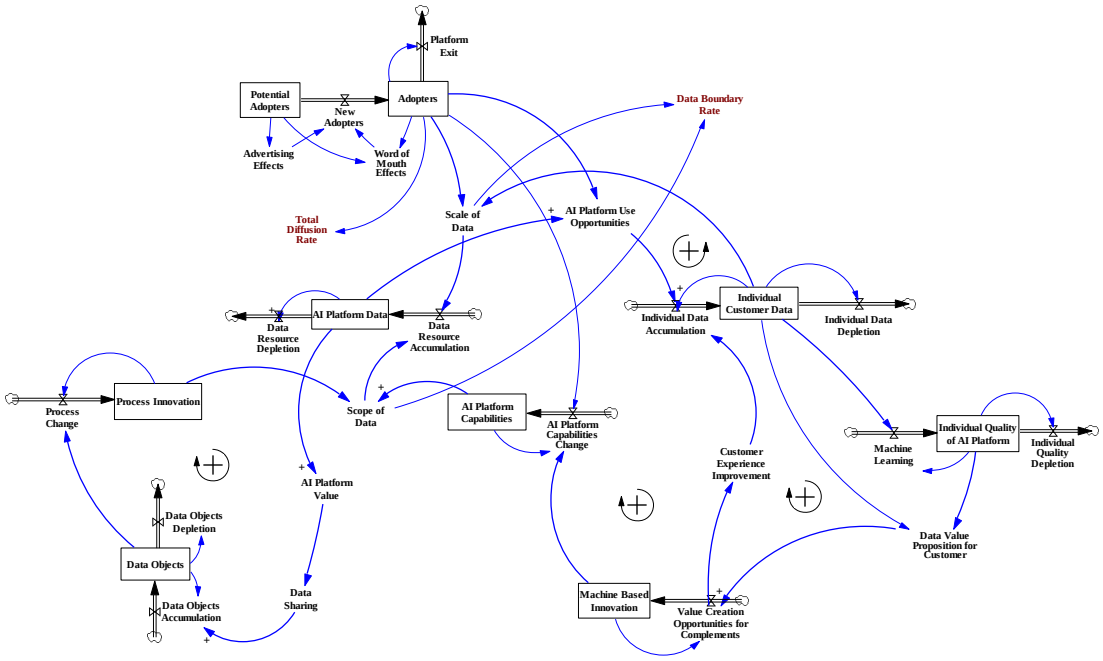

Fig.2 Dynamic model of AI platform firms

Dynamic changes in the scale of the data

We formulate the time series of the number of adopters owing to the diffusion of AI platforms based on the Bass model (Bass, 1969; Mahajan et al., 1990, 1995) (see Table 2). The Bass model uses external and internal influence coefficients that assume that potential adopters receive two types of communication: external influence through mass media and internal influence through word of mouth. The scale of the data is obtained as the product of the amount of individual customer data and the number of adopters.

Table 2. Scale of data

Name		Formula/parameter
Potential Adopters	$PA(t)$	$-\int NA(t)\,dt$
	$PA(0)$	TP
Adopters	$A(t)$	$\int [NA(t) - PE(t)]\,dt$
	$A(0)$	100
New Adopters	$NA(t)$	$AE(t) + WE(t)$
Platform Exit	$PE(t)$	$\frac{pef \times A(t) \times er}{art}$
		$If\ t < ast, 0$

Advertising Effects	$AE(t)$	$If\ t \geqq ast, \quad PA(t) \times aes$ $If\ t \geqq aet, 0$
Word of Mouth Effects	$WE(t)$	$\frac{cr \times PA(t) \times A(t)}{TP}$
Total Diffusion Rate	$TDR(t)$	$\frac{A(t)}{TP}$
Scale of Data	$SCAD(t)$	$A(t) \times ICD(t)$
Advertising Effectiveness	aes	0.0001
Advertising Start Time	ast	0
Adopters Response Time	art	2
Advertising End Time	aet	6
Exit Rate	er	0.01
Contact Rate	cr	0.7
Total Population	TP	1e+6
Platform Exit Flag	pef	1

Individual customer data learning loop

The equations defining the variables and constants included in the individual customer data learning loops are listed in Table 3. As opportunities to use an AI platform increase, the amount of individual customer data also increases, and the individual quality of the AI platform will improve through machine learning. The individual customer data value proposition leads to machine-based innovation through value creation opportunities for complements. Improving customer experience through value creation opportunities for complements will further increase the volume of individual customer data.

Table 3. Individual data enabled learning

Name		Formula/parameter
AI Platform Use Opportunities	$APU(t)$	$\frac{APD(t)}{A(t)}$
Individual Customer Data	$ICD(t)$	$\int [IDA(t) - IDD(t)]dt$
	$ICD(0)$	0
Individual Data Accumulation	$IDA(t)$	$max\left[\frac{APU(t) - ICD(t)}{idt} + CEI(t), CEI(t)\right]$

Individual Data Depletion	$IDD(t)$	$ICD(t) \times idr$
Individual Quality of AI Platform	$IQ(t)$	$\int [ML(t) - IQD(t)]\, dt$
	$IQ(0)$	0
Machine Learning	$ML(t)$	$max\left[\frac{ICD(t) \times ile - IQ(t)}{ilt}, 0\right]$
Individual Quality Depletion	$IQD(t)$	$IQ(t) \times iqdr$
Data Value Proposition for Customer	$DVP(t)$	$IQ(t) \times \left[1 + \frac{ICD(t)}{rdu}\right]^{dns}$
Machine-Based Innovation	$MBI(t)$	$\int VCO(t)\, dt$
	$MBI(0)$	0
Value Creation Opportunities for Complements	$VCO(t)$	$max\left[\frac{DVP(t) - MBI(t)}{vct}, 0\right]$
Customer Experience Improvement	$CEI(t)$	$VCO(t) \times rdi$
Data Network Effects Strength	dns	0.7
Individual Data Depletion Rate	idr	0.01
Individual Data Transformation Time	idt	1
Individual Learning Efficiency	ile	0.05
Individual Learning Time	ilt	1
Individual Quality Depletion Rate	$iqdr$	0.01
Reference Data per User	rdu	100
Reference Data for Improvement	rdi	10
Value Creation Time	vct	2

Dynamic change in the scope of data

The equations that define the variables and constants included in the process innovation

loop through data sharing are listed in Table 4. The AI platform capabilities increase through machine-based innovation. Data unions (objects) are accumulated by sharing data from the entire volume of AI platform data. Process innovation occurs through process changes using data objects. The scope of data expanded as a product of process innovation and AI platform capabilities. Furthermore, the total volume of AI platform data is expected to increase.

Table 4. Scope of data

Name		Formula/parameter
AI Platform Capabilities	$APC(t)$	$\int APCC(t)\,dt$
	$APC(0)$	1
AI Platform Capabilities Change	$APCC(t)$	$max\left[\frac{MBI(t)\times A(t)-APC(t)}{apt},0\right]$
AI Platform Data	$APD(t)$	$\int [DA(t)-DD(t)]\,dt$
	$APD(0)$	1
Data Resource Accumulation	$DA(t)$	$[SCAD(t)+SCOD(t)]\times dtr$
Data Resource Depletion	$DD(t)$	$APD(t)\times ddr$
Scope of Data	$SCOD(t)$	$APC(t)\times PI(t)$
AI Platform Value	$APV(t)$	$\left[1+\frac{APD(t)}{rdap}\right]^{dns}$
Data Sharing	$DS(t)$	$ief\times APV(t)\times dsr$
Data Objects	$DO(t)$	$\int [DOA(t)-DOP(t)]\,dt$
	$DO(0)$	1
Data Objects Accumulation	$DOA(t)$	$max\left[\frac{DS(t)-DO(t)}{dot},0\right]$
Data Objects Depletion	$DOP(t)$	$DO(t)\times dodr$
Process Innovation	$PI(t)$	$\int PC(t)\,dt$
	$PI(0)$	0
Process Change	$PC(t)$	$max\left[\frac{pif\times[DO(t)-PI(t)]}{pct},0\right]$
Data Boundary Rate	$DBR(t)$	$\frac{SCOD(t)}{SCAD(t)+SCOD(t)}$
AI Platform Coordination Time	apt	1

Data Depletion Rate	ddr	0.1
Data Transfer Rate	dtr	0.75
Reference Data Resource for AI Platform	$rdap$	10000
Data Sharing Rate	dsr	0.25
Data Objects Coordination Time	dot	1
Data Objects Depletion Rate	$dodr$	0.01
Process Coordination Time	pct	2
Interface Expansion Flag	ief	1

Calculation of an index of feature balance of AI platform firms

As an index of feature balance for AI platform firms, the data boundary rate is calculated as the proportion of the sum of the scale and scope of data to the scope of data. The definition equations are as follows:

$$\text{ScaleofData} = SCAD(t) = Adopters(t) \times IndividualCustomerData(t) \quad (1)$$

$$\text{ScopeofData} = SCOD(t) = AIPlatformCapabilities(t) \times ProcessInnovation(t) \quad (2)$$

$$\text{DataBoundaryRate(t)} = \frac{SCOD(t)}{SCAD(t)+SCOD(t)} \quad (3)$$

The scale of data is the total volume of individual customer data generated and used within AI platform firms. Increasing the scale of data enhances machine learning opportunities for AI platforms, improves the AI platform quality for individual customers, and improves customer experience. This variable can also be seen as an indicator of the competitive advantage of AI platform firms based on proprietary data resources. The scope of data is the total volume of data, comprising the product of AI platform capabilities and processes using data objects that can be shared between AI platform companies and other operators. This variable can also be viewed as an indicator of the diversity of processes through which AI platform firms can collaborate with other businesses. The data boundary rate values ranged from 0 to 1.
A data boundary rate greater than 0.5 can be interpreted as high and lower than 0.5 as low. When the data boundary rate is high, the scope of data is prioritized, and shared

data account for the majority of the AI platform firm's total data volume. If the data boundary rate is low, the scale of data is prioritized and customers' proprietary data account for the majority of the total data volume of the AI platform firm.

Simulation results

Julia code generation using generative AI

The simulation results of SD-based dynamic models are constrained by the analysis and output capabilities that are the specifications of SD tools, SD model development, and execution environments. The restriction of SD modeling to the specifications of SD tools may prevent the use of the most advanced analytical and visualization tools developed in the academic fields of AI, mathematics, and data science, where significant progress has been achieved in recent years. Meanwhile, converting SD models into other programming language codes requires significant learning time and trial and error. However, this effort can be largely avoided by using newly-available generative AI. In this paper, we used ChatGPT (GPT-4) (OpenAI, 2023) to convert the SD model of the AI platform firm into Julia code. The following conversion guidelines were set in the prompts (see Table 5):

Table 5. Prompts to convert Julia code from SD definition formula

Convert the system dynamics definition formula into Julia code that meets the following requirements #Set each parameter to a global variable. #Convert and organize all variable names without omitting them. #Rearrange the order of the Julia code arrangement to enable sequential computation. #Perform time integration on all variables set in INTEG(). #Use the DifferentialEquations package

In this section, we present simulation results using the latest analysis/visualization tools (a set of application packages released as open source) on a set of differential equations that are the building blocks of the dynamic model of AI platform firms generated by the Julia code through ChatGPT (GPT-4).

Initial values and reference model

For simulation using the dynamic model of the AI platform firm, the initial values were set to 1 million potential adopters (market size) and a maximum elapsed time (TMAX) of 24 months. Within this period, a dynamic model was used as a reference model with each parameter set to allow time variation to be observed within the range where the maximum time-integral variable (stock variable) does not overflow. In this reference model, the data network effects strength of the AI platform firm was fixed at 0.7. The reference model run results in an adoption rate of 200,000 (20% market penetration) in 10 months and over 800,000 (80% market penetration) in 15 months. The time-series change graphs of the stock variables in the reference model were compared by varying the percentage of available data sharing (DataSharingRate) and extending the interface of the AI platform firm to six different levels (0.1, 0.2, 0.4, 0.6, 0.8, 1.0) (see Fig. 3). The results of the reference model execution show that the higher the DataSharingRate, the more noticeable the exponential growth of the time-integrated variables (stock variables) from 15 months later. In particular, data objects shared with other platforms (DataObjects) and process innovations that use them (ProcessInnovation), owing to the expansion of the interfaces of the AI platform firm, show growth of the same magnitude. ProcessInnovation leads the growth, followed by exponential growth of other stock variables, AI platform data resources (AIPlatformData), and AI platform capabilities (AIPlatformCapabilities).

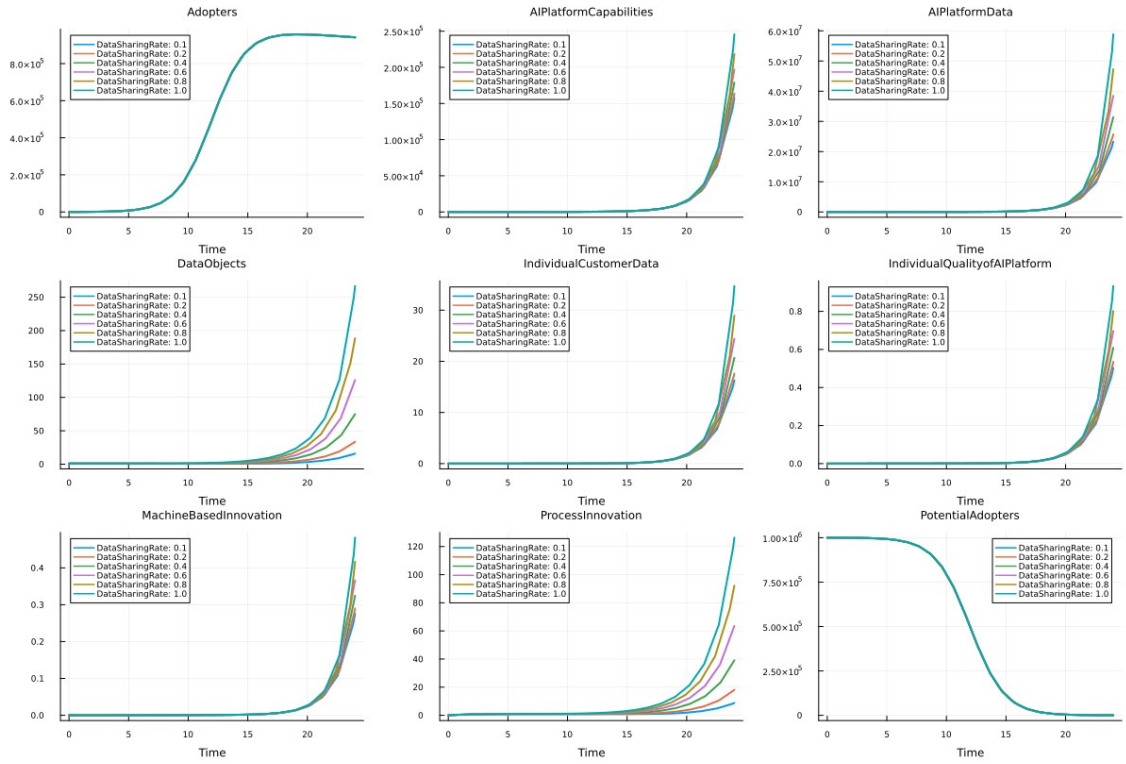

Fig. 3 Graph of time-series changes in stock variables

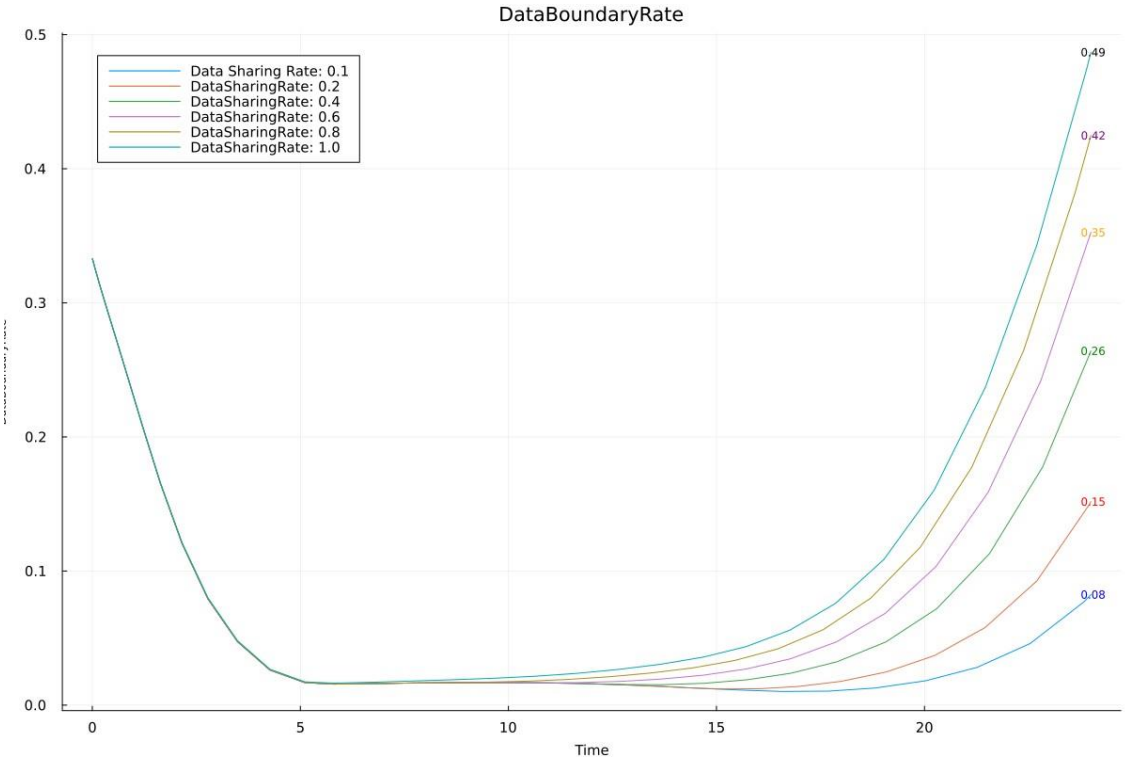

Fig. 4 Data boundary rate graph

Figure 4 shows a time-series graph of DataBoundarygRate, which is the proportion of the scope of the data to the sum of the scale and scope of the data. The DataBoundarygRate is 0.08 (8%) after the maximum elapsed time (TMAX) when the reference model is executed with DataSharingRate set to 0.1. When DataSharingRate was set to 1, DataBoundarygRate reached 0.49 (49%). DataBoundaryRate $\approx \frac{1}{2}$ can be interpreted as an equilibrium between the scale and scope of data in the AI platform firm.

Contour plots of variables at maximum elapsed time

Contour plots of the simulation results are useful for understanding the trend (change direction), density, and pattern of a variable. Changes in density can be either dense or relaxed. Tight contour intervals indicate abrupt changes in values. Conversely, the relaxed contour intervals indicate a gradual change of value. There may also be a pattern in the shape of the contour map. This pattern provides insights into the overall characteristics of simulation results. Contour plots for the stock variable values and the data boundary rate values at the maximum elapsed time (TMAX) are generated from

simulations that vary combinations of the AI platform firm parameters, data sharing rate ($0.1 \leq$ DataSharingRate ≤ 1) and data network effects strength ($0.1 \leq$ DataNetworkEffectsStrength ≤ 1) (see Fig. 5).

For each stock variable, a significant change (increase) in its value from $0.4 \leq$ DataSharingRate or greater and $0.7 \leq$ DataNetworkEffectsStrength is confirmed. The contour map of DataBoundaryRate values quantifies the degree of balance between the scale and scope of data in the AI platform firm. The blue areas in the contour map indicate that the scale of data is dominant, whereas the red areas indicate that the scope of data is dominant. The yellow–green line in the contour map is the data boundary ratio (0.5), which can be interpreted as the region of equilibrium between the scale and scope of data.

The change (increase) in the data boundary ratio becomes noticeable when the DataNetworkEffectsStrength is 0.6 or higher. For example, the data boundary rate is approximately 0.5 when the data network effects strength is 0.8 and data sharing ratio is 0.6, which nearly balances the size and scope of the data.

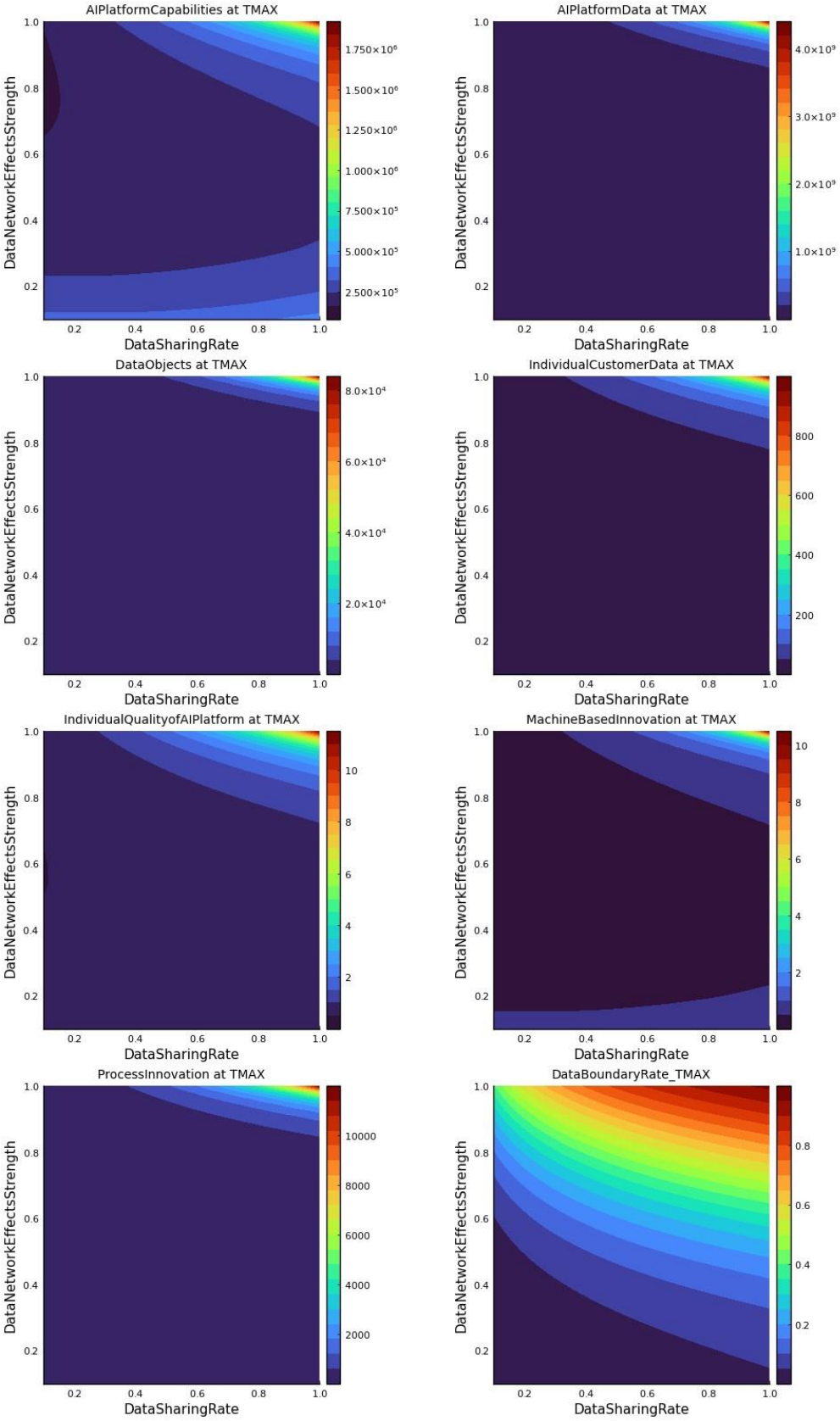

Fig. 5 Contour maps of variables at TMAX

General linear regression for data boundary rate

Regression analysis was performed using Julia's GLM, Statistics, and StatsBase packages to obtain the estimated data boundary rate (DataBoundaryRate) equations from the simulation. Four general linear regression models (linear, interaction, exponential, and power) were used.

Linear and interaction models account for simple additive and interaction effects, respectively. Exponential and power models account for multiplier effects. AIC (Akaike information criterion), mean square error (MSE), and R2 (coefficient of determination) were used as evaluation indices to indicate the predictive performance, error, and goodness-of-fit of each regression model (see Table 6).

AIC is an evaluation index of the balance between goodness-of-fit and complexity of the model; the smaller the value, the better the model (Akaike, 1974). From the AIC, the interaction model has a better fit to the data, while MSE is a measure of the prediction error of the model; the smaller the value, the better the prediction accuracy of the model.

However, the prediction accuracy and data fit of the interaction model are almost identical to those of the power model. In the interaction model, DataSharingRate alone has a negative impact. The interaction between DataSharingRate and DataNetwork EffectsStrength has a positive impact.

The power model is exponentially transformed into the following estimation equation (DataSharingRate is abbreviated as DSR and DataNetworkEffectsStrength is abbreviated as DNES):

$$DataBoundaryRate = e^{-0.073423} \times DSR^{0.74899} \times DNES^{1.78237} \quad (4)$$

In the power model, both DataSharingRate and DataNetworkEffects variables have positive exponential effects. In particular, changes in the DataNetworkEffectsStrength affect the Data Boundary Rate.

Table 6. Linear Regression Models for DataBoundaryRate(TMAX)

	Linear Model	Interaction Model	Exponential Model	Power Model
(Intercept) Coef.	0.366***	-0.139***	-5.229***	0.073***
	(0.006)	(0.009)	(0.012)	(0.014)
DataSharingRate	0.295***	-0.116***	1.651***	
	(0.007)	(0.014)	(0.014)	
DataNetworkEffectsStrength	0.838***	0.426***	4.259***	

	(0.007)	(0.014)	(0.014)	
Interaction Term		0.749***		
		(0.024)		
log(DataSharingRate)|				0.749***
				(0.011)
log(DataNetworkEffectsStrength)				1.782***
				(0.011)
AIC	-4508.66	-5355.02	-1252.29	1822.26
MSE	0.00961	0.00685	0.0122	0.00659
R-Squared	0.8523	0.8948	0.8126	0.8987

Standard errors are indicated in parentheses. * p<0.05, ** p<0.01, *** p<0.001

Nonlinear regression for data boundary rate

We also attempt to estimate the DataBoundaryRate using a nonlinear regression model. For this purpose, a nonlinear regression model was formulated by parameter optimization with nonlinear least squares fitting, using the curvefit function from the LsqFit package of the Julia code.

The nonlinear regression model for DataBoundaryRate shows dependence on the data sharing rate, the data network effects strength and its square, the quadratic term, and interaction term. Similar to the general linear model, AIC, MSE, and R2 were used as evaluation indices to assess the goodness-of-fit and predictive accuracy of the nonlinear regression model (see Table 7).

The nonlinear regression model has a high AIC but the lowest MSE and fits the data well (the coefficient of determination is closest to 1).

In the nonlinear regression model, the first-order term in DataSharingRate is positive, and the second-order term is negative. Conversely, DataNetworkEffectsStrength had a negative first-order term and a positive second-order term. The absolute values of these coefficients are larger for the quadratic terms. The interaction term is positive and its coefficient values are comparable to those of the interaction model. Essentially, when DataSharingRate is microvalued, it has a positive impact on DataBoundaryRate. However, as DataSharingRate increases, the positive effect decreases and reverses to a negative effect.

When the value of DataNetworkEffectsStrength is minor, it negatively affects DataBoundaryRate. As the DataNetworkEffectsStrength increases, the negative impact decreases and reverses to a positive impact. The interaction between DataSharingRate and DataNetworkEffects Strength also has a positive impact on DataBoundaryRate.

The nonlinear regression model for DataBoundaryRate described above is nonlinear (1), which may output negative values.

Nonlinear model (2) is a nonlinear regression model that forcibly replaces the negative value of DataBoundaryRate with a zero value when a negative value is obtained, and the following estimation equation:

$$DataBoundaryRate = max(0, 0.10484 + 0.10632 \times DSR - 0.95293 \times DNES - 0.20241 \times DSR^2 + 1.25398 \times DNES^2 + 0.74867 \times DSR \times DNES) \quad (5)$$

Table 7 Nonlinear Regression Models for DataBoundaryRate(TMAX)

	Interaction Model	Power Model	Nonlinear Model(1)	Nonlinear Model(2)
(Intercept) Coef.	-0.139***	0.073***	0.1048	0.1048
	(0.009)	(0.014)		
DataSharingRate	-0.116***		0.1048	0.1048
	(0.014)			
DataNetworkEffectsStrength	0.426***		-0.9529	-0.9529
	(0.014)			
Interaction Term	0.749***		0.7487	0.7487
	(0.024)			
log(DataSharingRate)|		0.749***		
		(0.011)		
log(DataNetworkEffectsStrength)		1.782***		
		(0.011)		
$DataSharingRate^2$			-0.2024	-0.2024
$DataNetworkEffectsStrength^2$			1.254	1.254
AIC	-5355.02	1822.26	-3.266	-3.4479
MSE	0.00685	0.00659	0.00048	0.00044
R-Squared	0.8948	0.8987	0.99256	0.99321

Standard errors are indicated in parentheses. * p<0.05, ** p<0.01, *** p<0.001

Comparison of contour plots by general linear regression models

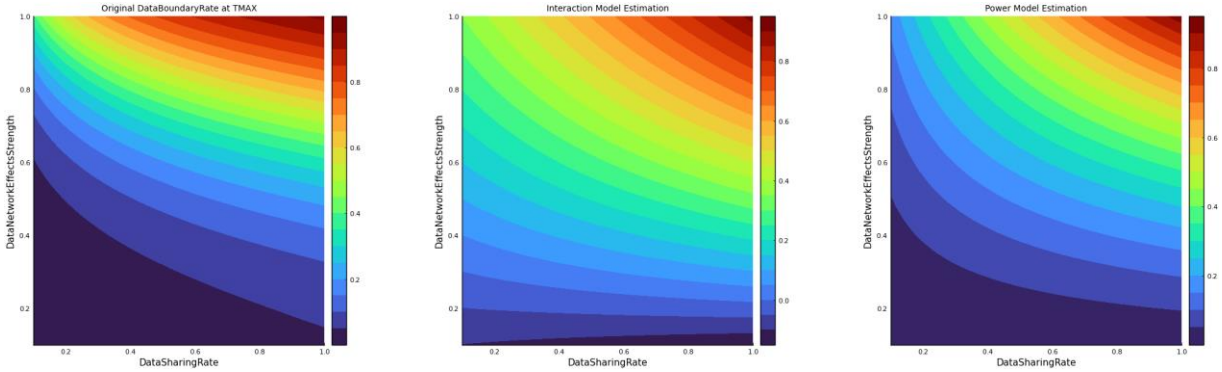

Fig. 6 Contour plots of DataBoundaryRate by the interaction/power models

In this section, we use simulation results and the general linear regression model to compare the maximum elapsed time (TMAX) contour plots of AI platform firms (see Fig. 6) and the data boundary rate (DataBoundaryRate) contour plots of AI platform firms. First, we compare the contour plots of DataBoundaryRate using the interaction and power models, which are highly accurate as general linear regression models. The MSE and coefficient of determination, which are regression estimation indices, indicate that the difference in accuracy between the two models is minimal. The first (left side) contour plot is based on the calculated DataBoundaryRate at TMAX obtained from the simulation. The second (middle) contour plot is based on the predictions of the interaction model, and the third is based on the power model predictions. However, as can be seen from this comparison of contour plots, in the case of the interaction model prediction, for DataNetworkEffectsStrength ≤ 0.2, negative influences based on negative coefficients are at work, resulting in negative values of DataBoundaryRate.
In the power model, only positive influences based on positive coefficients act and only positive values of DataBoundaryRate are obtained.
In the power model, the data sharing effects indicated by DataSharingRate and data network effects indicated by DataNetworkEffectsStrength are multiplicative, suggesting that the transition in DataBoundaryRate is nonlinear.

Comparison of contour plots by general linear and nonlinear regression models

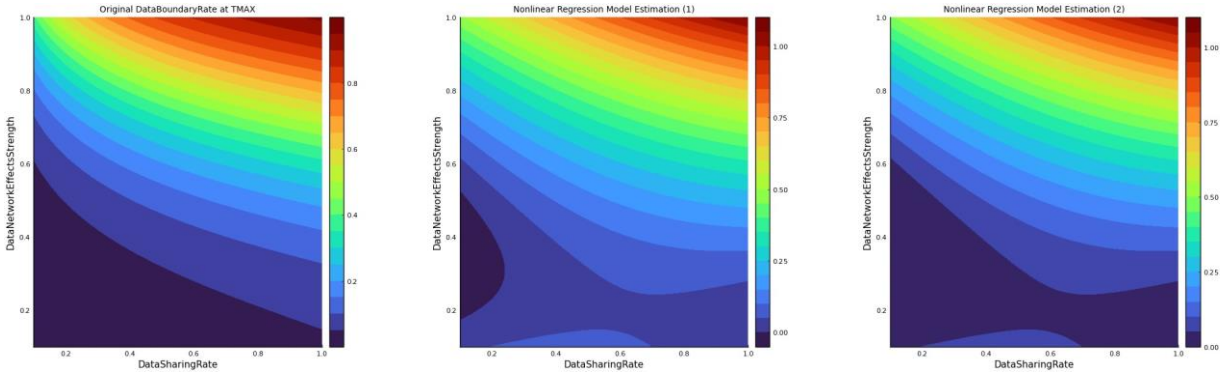

Fig. 7 Contour plots of DataBoundaryRate by nonlinear models

Contour plots of the DataBoundaryRate simulation results, nonlinear model (1) predictions, and nonlinear model (2) predictions were generated (see Fig. 7). Comparing the contour plots of nonlinear model (1) and nonlinear model (2), in nonlinear model (1), when DateSharingRate is lower than 0.2, the data boundary rate is negative in the range $0.2 \leq$ DataNetworkEffectsStrength ≤ 0.5. Since negative rates are not realistic, nonlinear regression model (2) is adopted as the nonlinear regression model.

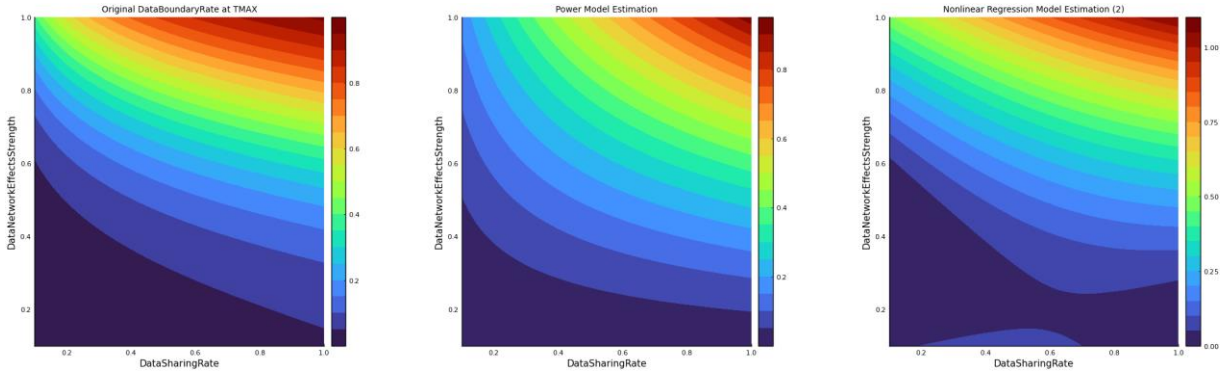

Fig. 8 Contour plots of DataBoundaryRate by the power/nonlinear models

The first (left) contour plot in Fig. 8 shows the values calculated from the simulation results, the second (middle) contour plot shows the predictions of the power model, and the third (right) contour plot shows the predictions of the nonlinear model (2). Based on the evaluation of the predictions by both models (see Tables 6 and 7) and the comparison of the contour patterns, nonlinear model (2) was selected as the estimating equation for the DataSharingRate.

Three areas of data boundary rates

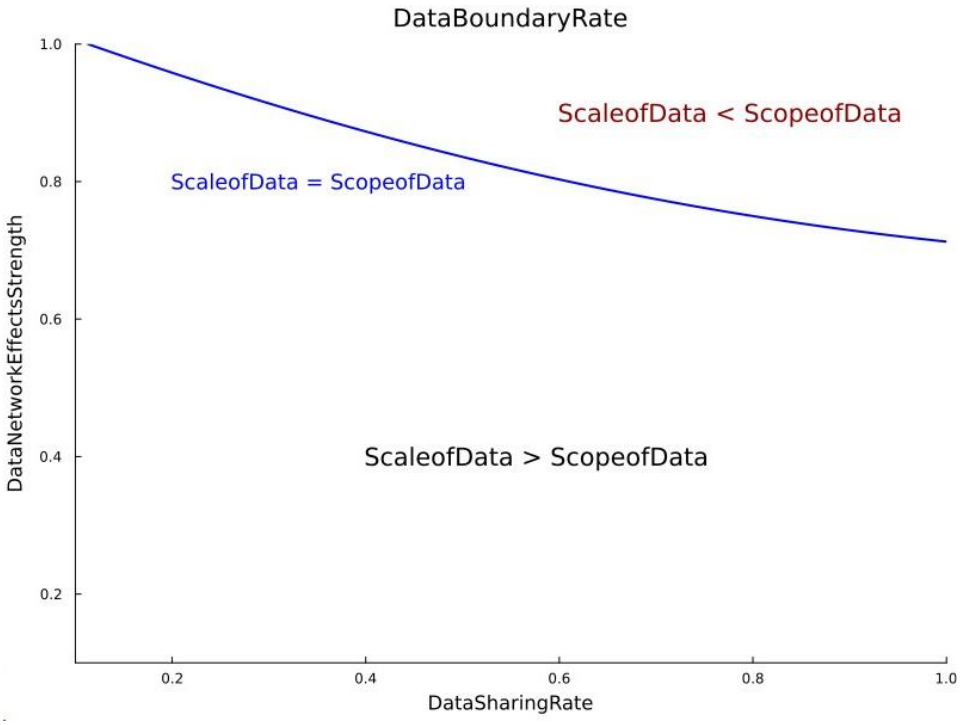

Fig. 9 Three areas within the DataBoundaryRate contour map

The data boundary rate (DataBoundaryRate) was set as an indicator of the feature balance between the size and scope of data in AI platform firms, and nonlinear model (2) was used as its estimating equation. Nonlinear model (2) was used to draw the three areas of the data boundary rate (Fig. 9). The data boundary rate is the proportion of the scope of the data to the sum of the scale and scope of data, and its value ranges from 0 to 1: Fig. 9 shows the three areas of the DataBoundarygRate contour map created using the nonlinear model (2). The areas of data boundary rate can be categorized into three areas based on the scale and scope of data perspectives: the scale of data-dominant area, the scale and scope equilibrium area, and the scope of data-dominant area. Within the contour map obtained on the plane generated from the value ranges of the DataNetworkEffectsStrength and DataSharingRate, the scale of data-dominant area is the largest, accounting for approximately 80% of the total area. The equilibrium area of scale and scope of data was obtained as an equilibrium line from the contour map obtained in this study. The scope of data-dominant area is the remaining area divided by the scope and scale of data equilibrium line, which tends to have a larger DataNetworkEffectsStrength. On the scope and scale of data equilibrium line, the maximum value of DataNetworkEffectsStrength (1.0) corresponds to the minimum value of DataSharingRate (0.10). When DataSharingRate reaches its maximum value (1.0), DataNetworkEffectStrength is 7.125.

Basically, DataNetworkEffectStrength is characterized by a linear decrease as

DataSharingRate increases.

Discussion and conclusions

In this study, the scale of data in AI platform firms focuses on closed data, such as customer data and trade secrets. Essentially, the scale of data is treated as the amount of primary data, which is the most important data in AI platform firms. The scope of data in AI platform firms focuses on publicly-available open data or data shared with other firms. Specifically, we treat the volume of secondary data and the types that support the use of primary data in AI platform firms as the scope of data. From the perspective of the aforementioned scale and scope of data, this paper develops qualitative and quantitative models (dynamic model) of AI platform firms. This implies that the qualitative model of AI platform firms is based on three virtuous cycles through data network effects: autonomous data-enabled learning over time, the level of AI algorithm updates, and the expansion of business areas with process innovation from data sharing are interdependent and reinforcing, suggesting that they lead to the exponential growth of AI platform firms. Through simulated results using a dynamic model of AI platform firms, we quantitatively confirmed the exponential growth driven by the scale and scope of data.

Feature balance of scale and scope of data as an indicator

In this study, we set the data boundary rate (DataBoundaryRate) as an indicator of the feature balance of the scale and scope of data of AI platform firms. The indicator value is the proportion of the scope of data to the sum of the scale and scope of data.
Using the simulation results of the AI platform firms, a contour map of the DataBoundaryRate was generated on a plane with features of the DataNetworkEffectsStrength and DataSharingRate set along the two axes.
General linear and nonlinear regression equations approximating DataBoundaryRate values in this plane were estimated. By examining the properties of these regression equations, we investigated the feature balance between the scale and scope of data in AI platform firms. We focus on the case of DataBoundaryRate=1/2 as the domain where the scale and scope of data in AI platform firms grow in equilibrium. This equilibrium can be interpreted as the volume of primary data that is closed within the firm and the volume of secondary data (open data and data shared with other firms) that supports the use of primary data being equal in size, and this equilibrium growth increases the total data volume of AI platform firms. In the equilibrium area (on the line) of the scale and

scope of data, a linear relationship was observed, where the data network effects strength decreased as the data sharing rate increased. This equilibrium between the scale and scope of data in AI platform firms can be interpreted in terms of the strategic dimension of platform competition proposed by Cennamo (2021), which corresponds to platform identity and platform size. This suggests that the volume of primary data and the predictions and improvements based on it are considered the identity of AI platform firms, and that the total volume of data, including process innovation using shared data, is considered the size of AI platform firms (see Table 1). Essentially, this feature balance can be used as a key performance indicator for AI platform firms' business policy formulation, that is, strategic decision making.

Study limitations and future directions

This study has both theoretical and empirical limitations. First, from a theoretical perspective, the study examines AI platform firms based on only the scale and scope of data. We present a qualitative model that focuses only on data network effects as a mechanism acting on the exponential growth of AI platform firms. It does not consider the impacts of direct or cross-side network effects.

Moreover, the qualitative model has a multiple causal loop structure; however, the causal relationships between each component are based on logical reasoning, and their validity has not been verified. Based on this qualitative model, we develop a dynamic model for AI platform firms. The structure of this dynamic model and the set of parameter values are logical possibilities, but the validity of the simulation results has not yet been verified. The simulation results suggest only the possibility of exponential growth of AI platform firms based on the feature balance of scale and scope of data.

The future research agenda includes the interpretation and evaluation of the case study analysis using the qualitative model of AI platform firms as a framework, the calibration of the parameters of the dynamic model based on the quantitative data obtained from the case study, and the validation of the simulation results.

Furthermore, theoretical extensions of the qualitative and dynamic models of AI platform firms that consider the interaction of data network effects and other network effects and study the patterns of balancing characteristics in the scale and scope of data through simulations are also important research topics.

Acknowledgements

In this study, ChatGPT (GPT-4) (OpenAI, 2023) was used to generate the Julia code for

the simulation and result output. The authors of this paper are solely responsible for the accuracy of facts and quotations, mathematics, logic, commonsense reasoning, and originality of this paper.

References

Agrawal A, Gans J, Goldfarb A. 2022. Power and Prediction: The Disruptive Economics of Artificial Intelligence. *Harvard Business Review Press*.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723.

Anderson V, Johnson L. 1997. *Systems Thinking Basics: From Concepts to Causal Loops.* Pegasus Communications, California.

Bass FM. 1969. A New Product Growth for Model Consumer Durables. *Management Science* **15**(5): 215–227.

Cennamo C. 2021. Competing in Digital Markets: A Platform-Based Perspective. *Academy of Management Perspectives* **35**(2): 265–291.

Clough, D. R., Wu, A. 2022. Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems. *Academy of Management Review* **47**(1): 184–189.

Davis JP, Eisenhardt KM, Bingham CB. 2007. Developing Theory Through Simulation Methods. *Academy of Management Review* **32**(2): 480–499.

Forrester JW. 1961. Industrial Dynamics, M.I.T. Press, Cambridge, MA.

Gambardella A, McGahan, AM. 2010. Business-Model Innovation: General Purpose Technologies and their Implications for Industry Structure. *Long Range Planning* **43**(2–3): 262-271.

Gawer A. 2021. Digital platforms' boundaries: The interplay of firm scope, platform sides, and digital interfaces. *Long Range Planning* **54**(5): 102045.

Goldfarb A, Taska B, Teodoridis F. 2023. Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings. *Research Policy* **52**(1): 104653.

Gregory RW, Henfridsson O, Kaganer E, Kyriakou H. 2021. The Role of Artificial Intelligence and Data Network Effects for Creating User Value. *Academy of Management Review* **46**(3): 534–551.

Gregory RW, Henfridsson O, Kaganer E, Kyriakou H. 2022. Data Network Effects: Key Conditions, Shared Data, and the Data Value Duality. *Academy of Management Review*, **47**(1): 189–192.

Hagiu A, Wright J. 2020. When Data Creates Competitive Advantage. *Harvard Business Review* **98**(1): 94–101.

Heimburg VS, Julian, Wiesche M. 2023. The Future of Digital Platform Design - The

Case of the EU Platform Regulation Discourse, *ECIS 2023*, Vol. Research Papers. Norway.

Iansiti M. 2021. The Value of Data and Its Impact on Competition, 1–19. Harvard Business School.

Ismail S, Malone MS, van Geest Y, Diamandis PH. 2014. *Exponential Organizations: Why new organizations are ten times better, faster, and cheaper than yours (and what to do about it)*. Diversion Books, New York City.

Katz ML, Shapiro C. 1985. Network Externalities, Competition, and Compatibility. *The American Economic Review* **75**(3): 424–440.

Katz ML, Shapiro C. 1994. Systems Competition and Network Effects. *Journal of Economic Perspectives* **8**(2): 93–115.

Kimura M. 2022. A Cyclical Model of the Data Network Effects: Deepening Data Learning and Expanding Boundaries of AI-enabled Platforms. *Journal of the Japan Society for Management Information* **31**(2): 59–76.

Mahajan V, Muller, E., Bass FM. 1990. New Product Diffusion Models in Marketing: A Review and Directions for Research. *Journal of Marketing* **54**(1): 1–26.

Mahajan V, Muller E, Bass FM. 1995. Diffusion of New Products: Empirical Generalizations and Managerial Uses. *Marketing Science* **14**(3_supplement): G79-G88.

Mayer-Schönberger, V., & Ramge, T. 2018. *Reinventing Capitalism in the Age of Big Data*. Basic Books, New York City.

OpenAI. 2022. Introducing ChatGPT, Vol. 2022.

Prüfer J, Schottmüller C. 2021. Competing with Big Data. *The Journal of Industrial Economics* **69**(4): 967–1008.

Ruutu S, Casey T, Kotovirta V. 2017. Development and competition of digital service platforms: A system dynamics approach. *Technological Forecasting and Social Change* **117**: 119–130.

Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw-Hill School Education Group.

Stucke ME., Grunes AP. 2016. *Big Data and Competition Policy*. Oxford University Press, Oxford, UK.

Eloundou T., Manning S., Mishkin P., Rock D. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. OpenAI Working Paper.

Zhu F., Iansiti M. 2012. Entry into platform-basedmarkets. *Strategic Management Journal* **33**: 88–106.