Research Statement

Vineet Kumar, *Yale School of Management*

My overall research focus is at the intersection of technology and business & society.[1] Theory or structured knowledge is methodologically central to all my quantitative modeling. Given my background and training in engineering, economics and marketing, I am interested in two areas.

(A) *Digital business models:* My research focuses on distinct strategic long-run choices relevant to digital firms, a topic of interest to multiple fields. I study this area using methods based on microfoundations of agents' preferences, typically empirical structural models. These models yield economically interpretable estimates, enabling counterfactual evaluation of firm or regulatory policies, and exogenous changes in technologies.

(B) Building *theory-based machine learning (ML)* methods incorporating structured knowledge developed from first principles. These methods are also shown to be human interpretable, and provide representations satisfying desirable properties (e.g. monotonicity of demand curves).

(A) Digital Business Models

Digital firms have unique issues to consider, e.g. open versus closed source, indirect network effects, or freemium with perpetually free products, that are not seen with physical products. My research can broadly be themed as focusing on three different sources of connections: between *products*, *consumers* and *data*.

1) Connections across Products – Product Line

Product design is especially important to digital firms, which typically avoid traditional marketing approaches. I detail the connections between products, with a research overview and agenda provided in [P1].

Bundling Hardware and Software: Bundling is an especially flexible product strategy, and is commonly used across many markets. Past work focuses on how bundling can leverage the negative correlation in consumers' static valuations across products. However, very little is known about the dynamic effects of bundling. To evaluate the long-run impact, in [P4], I use a forward-looking model, where consumers first purchase hardware (consoles) followed by software (games). I find that bundles act by pulling demand forward, i.e. (lower valuations) consumers buy now rather than wait – a new mechanism, only possible to identify in a model with forward-looking consumers. Using this model, I am able to examine several strategic questions. First, I find that bundling increases not just the quantity of sales, but shifts the timing to occur earlier. This is especially important since having a larger installed base of hardware enables multiple software purchases earlier. Second, I find that pure bundling as a strategy is empirically dominated by mixed bundling. Third, a more positive correlation results in greater sales, relative to negative correlation, through

[1] For simplicity of exposition, this document is written in the first person singular, although most of my work is in collaboration with a wonderful set of co-authors.

the above mechanism. My results are in contrast to extant work that has considered a static context and found bundling to be more effective with negative correlation. I also show that bundling is more effective when indirect network effects are weaker, which was not known earlier. This work has implications to any context where dynamic or intertemporal factors are relevant and firms can bundle products. Methodologically, I connect the software and hardware market using an integrated structural model, where consumers form expectations over software value when they make hardware purchases. Prior work had instead used the number of available products as a proxy. I also develop a novel identification strategy for correlation in valuations based on the tying ratio, leveraging the feature that consumers purchase one hardware but many software units.

Related Methodological Studies: The above research also led me to investigate methods for estimating dynamic demand models. Examining the commonly used inclusive value approach, I show that it can lead to highly biased estimates of economically important quantities like elasticities and profits [P5]. Motivated by the need for models to flexibly accommodate dynamics with large state spaces, I develop a new method that can be estimated without needing to reduce the dimensionality [P2]. For a large class of problems (with terminating or renewal choice) using market-level data, I obtain consumer preferences with the simplicity of a linear regression. I formally prove identification and recover the evolution of unobservable product characteristics, enabling counterfactual analysis. I examine the monetary value of product features in the (early) market for smartphones, which has a large number of products, finding that Bluetooth and Wi-Fi had the biggest impact on Apple's sales, providing a competitive advantage, relative to other features.

(Digital) Transformation: Transformation through technological change, linking the old product to the new, is a critically important but inherently risky and challenging process for firms. Yet, empirical structural studies of the impact of transformation are rare. In [P10], I examine the transformation for a firm moving a product based on an older technology (physical or slower service time) to one based on newer technology (digital or faster). Using a panel from a firm using the "Netflix" model with multiple versions (plans), I obtain consumer preferences for viewing content in physical form. Typically, quality or service improvements are expected to improve the value generated for all consumers, and therefore, the surplus available to the firm as revenue. However, I show that with service improvements, this logic may not hold. The reason is that such improvements also reduce the differentiation between the different product versions (plans). In the limit with zero service time, just one plan would be offered. Thus, the firm has fewer options to price discriminate, and facing heterogeneous consumers is not able to extract as much surplus. Indeed, in counterfactuals with reduced service time, I find the value created increases, but differentiation across versions decreases (even collapses). This novel mechanism shows how a transformation that enables more value to be created for all consumers results in lower profits and revenue for the firm. This new transformation risk is likely to exist in many transformations, especially digital. The implication is that firms need to determine how to evaluate these transformation risks and strategies to overcome them.

Methodologically, I develop a dynamic structural model with forward-looking consumers, incorporating a multiplicative error structure, rather than the typical discrete choice logit with additive errors. The model is appealing because it would never permit a dominated plan to be chosen, unlike the standard model, and allows me to obtain a rich specification of heterogeneity. I also construct a new strategy for identification of switching costs based on content enhancement, which increases the gains from switching, rather than price variation. This approach would be valuable in other contexts like digital app subscriptions, where price variation is typically absent.

Open Source: The puzzle in open source software is that free-riding can produce high quality products, and my model shows how this is sustained in equilibrium [P14]. I examine product strategy, where contributions to product features made either by developers or by any firm are a public good available to all competing firms (e.g. Linux, Android or Llama LLM). Developers signal their capabilities by making feature contributions to open source. Firms build on features by differentiating on a complementary dimension (usability), and a greater degree of open source contributions can enhance the differentiation value of usability. I show, in contrast to past research and industry leaders' beliefs, that allowing free-riding can result in increased product quality.[2]

Other Related Work: Freemium is the most popular digital business model (e.g. app stores, cloud storage, SaaS). I undertake a deep dive into designing freemium for a storage service [P16], where referrals are rewarded with an improvement to the free product. I evaluate how there is a tradeoff between growth and monetization, but if the firm is too generous, it can counterintuitively hurt both. I also examine the use of time as a versioning and monetization strategy (wait for free) in [P17]. I provide an overview of issues in Freemium for a general audience in [P11]. I also examine the case of a retailer with a physical and digital channel, and evaluate when it is optimal for the firm to match its own price across channels [P9].

2) Connections across Data – Linking Purchase and Usage

The next connection that I examine is the linkage between types of data, i.e. purchase data and usage data. In digital settings, usage data is more easily (even uniquely) available, and valuable in obtaining insights about consumer preferences. However, most studies in marketing and economics involve only purchase data. I connect purchase and usage data across several settings, including [P16, P10]. In [P3], I demonstrate how usage data is conceptually distinct, and is critically important for identification of demand, not just estimation. Specifically, I show that the nonparametric identification of the Willingness to Pay (WTP) distribution for subscriptions, without price variation is possible by leveraging usage data, but impossible without it. The key insight is that combining high-frequency usage data with purchase data allows for a conceptual leap in identification of the valuation (WTP) distribution, which was not thought to be possible without price variation. I crucially show that this result does not depend on any specific parametric functional form or distribution of shocks, which is challenging but valuable. I develop a model starting from the stream of usage utilities over time, and then connect the (expectations of) aggregated utility to the purchase process. I combine usage data with exogenous factors impacting usage to first estimate usage utility, then aggregate this stream, combined with purchase data to obtain the WTP

[2] Former Microsoft CEO Steve Ballmer described this free-riding thus: "Linux is a cancer that attaches itself in an intellectual property sense to everything it touches." (Link)

distribution. I can then conduct counterfactual analyses, such as product design. The framework is flexible in accommodating a large class of usage utility models, making it widely applicable.

3) Connections across Consumers – Networks

Word-of-mouth and referral effects can impact business models. However, my research in this area is quite general, and broadly applicable to a wide class of network interventions beyond that, e.g. public health. In network interventions, highly-connected individuals are useful to leverage as seeds for interventions. I investigate privacy-sensitive methods for leveraging network structure to obtain higher-degree nodes in *unknown* networks [P15], e.g. for word-of-mouth. The literature focuses on obtaining the entire network structure (not privacy-friendly!), and also does not offer provable guarantees, which is quite challenging in networks. My research addresses both these issues. I demonstrate two strategies (ego-based and alter-based), based on the friendship paradox, that ask individuals to nominate one or a few random friends. These strategies have distinct mathematical properties, are simple to implement, and both offer provable guarantees of obtaining higher-degree individuals. I also discover a new network property called Inversity, that determines which strategy obtains more highly-connected seeds, based on network structure.

I empirically study whether using friendship paradox strategies can achieve greater product adoption in [P13]. Using a model of communication and adoption estimated with data on real networks, I evaluate counterfactuals using the above seeding strategies, which have not been empirically examined. I show that ego-based friendship seeding outperforms random seeding, and surprisingly, obtains higher adoption than even leader-based seeding. The upshot is that it is possible to improve product adoption by just using (randomly chosen) friends as seeds, which provides a valuable information-light approach. The results hold across a range of specifications and networks, demonstrating robustness and empirical value, and have implications for referral design. Methodologically, I provide a novel nonparametric identification strategy leveraging the adoption trajectory to identify the differential impact of leaders, even when all networks are seeded with leaders.

Overall, in contrast to the literature, I show that it is possible to leverage networks to impact interventions with theoretical guarantees for *any* network, without knowing the network, and is empirically shown to work better than using demographic attributes (e.g. leadership).

(B) Theory-based Machine Learning

My research focus in ML is based on integrating structured knowledge (theory) to develop new ML methods, which in turn, enable us to obtain insights about consumer responses. First, there is a growing set of important research questions where ML is required. Marketing in practice involves important elements that impact consumers that are based on unstructured data (e.g. text, visual design, music, videos). Yet, the vast majority of research has used primarily structured numerical data, since they are more available and have standardized methods to analyze them. Traditional non-ML quantitative methods are not great at capturing the nuances of unstructured data. Moreover, ML also makes it feasible to generate novel unstructured data like text or images.

Given their growing capabilities, ML models are being increasingly used in academia and industry. However, they are typically opaque black box models (e.g. ChatGPT), leading to significant problems. First, these models are highly complex (with billions of parameters) and humans don't know the interpretation of these parameters. We also don't know the data on which they are trained. Second, they are designed to be atheoretical and so do not have a true understanding of the consumer.[3] Third, they are typically not interpretable – we don't know *why* they work. They are prone to failure (e.g. hallucinations), and we cannot know *when* this will happen because we have little visibility.[4]

My focus is on solving these challenges, developing ML methods based on theory or structured knowledge.[5] My research is focused around three aspects: (a) develop methods to incorporate structured knowledge into ML models, (b) provide model and data transparency, and (c) improve explainability and interpretability along with performance. Rather than using commercially available black box models, I develop models from basic elements.

The sources of knowledge that I examine include ideas and concepts from fields like marketing and economics, but is not limited to these. What concepts and ideas can we bring to ML? How can we incorporate theory into complex deep learning models with unstructured data? How can we demonstrate impact? I detail these critical challenges across a variety of studies below.

Visual Characteristics: I aim to obtain and quantify interpretable visual characteristics of products and consumer preferences for them in [P18]. Visual appearance of products is important in many categories, yet very high-dimensional and therefore challenging to characterize and explain. Prior research either required human experts to pre-define the set of characteristics, or focused on obtaining characteristics without any interpretability. My theory-based method automatically discovers and quantifies visual characteristics without expert input. Theory plays a crucial role in the following ways. First, the objective function is designed to incentivize low-dimensional and orthogonal representations, based on the idea that the product designs satisfy that property. Second, I leverage the idea that products have a distinct look tied to the brand, with recognizable "visual signatures" like BMW cars or LV handbags. My method leverages brand and other characteristics to supervise disentanglement, in contrast to unsupervised ML methods. I find significantly higher performance *and* greater human interpretability. Importantly, we can automatically *generate* counterfactual visual designs without experts, which was not thought possible for products.

Music and Emotion: Music is a major driver of user emotional response, yet greatly underexplored. In the music emotion research [P7], explaining why a listener feels a specific emotion when listening to music was a black box earlier, and this research changes that. I use ideas about consonance and dissonance of music, and how that connects to the listener's emotion using theory from multiple fields. Specifically, the knowledge is based on both: (a) the mathematics of sound waves and (b) psychology of human music perception. Here, theory is used as the basis for creating flexible and non-contiguous consonance filters, which is unique and novel, since almost all of ML uses contiguous filters for visual processing. Using these filters, we get a representation that enables

[3] For example, the same CNN deep net models used for marketing applications are also used in biology.

[4] Consider the algorithms for self-driving vehicles illustrating these points – only now are we beginning to get some idea about the inner workings. See The Hidden Autopilot Data That Reveals Why Teslas Crash (Wall Street Journal, 07/30/2024).

[5] Other researchers are trying to integrate theory from their academic fields into ML, e.g. in physics [O1].

explainability, so we can visualize how the features of music impact listener responses. The research also develops an application using emotional congruence (or contrast) as a a form of contextual targeting for ads, without using any consumer data, greatly improving privacy.

Learning Unknown Demand Curves: To learn an unknown demand curve by experimentation, I develop a reinforcement learning model with nonparametric multi-armed bandits (MAB) [P19]. The classic experiment (A/B) is inefficient since it explores all prices equally. MABs provide a more sophisticated approach ("learning while earning"), yet are atheoretical. Economic theory informs us that demand curves are downward sloping, but incorporating this knowledge into MABs is quite challenging.[6] I find that monotonicity adds two sources of value. First, the performance of the algorithm improves substantially, since it learns not just from each price (arm) experimented, but across arms (an informational externality). Second, incorporating a theoretical basis guarantees that the resulting demand curve is monotonic. This aspect is especially important when algorithms are used to make automated pricing decisions. Without monotonicity, we commonly get an upward sloping demand curve, resulting in unrealistically high prices and increasing the risk of failure.

Overall, all the ML-based research I have undertaken brings in the power of theory to enhance the capability of ML along several aspects: performance, interpretability / explainability, and providing representations satisfying desirable properties. In turn, these ML methods are useful to help us gain valuable insights into consumer and firm behavior, leading to a symbiotic process. All my ML research is transparent with the open source code publicly available for others to examine, critique and build on. I expect this approach to improve stakeholder trust and acceptance of AI systems.

Teaching

I have developed and taught elective courses at the masters level, and have also contributed to the executive MBA and executive education. I use a mix of lectures to help with learning principles, complemented with discussions featuring case studies by the Socratic method. In *Digital Strategy*, I explore digital business models (e.g. Dropbox), and evaluate models used by disruptors and complementors. I investigate the drivers and barriers of platform success, and examine digital transformation and emerging technologies. Given the connection to my research, I've incorporated exercises based on research into the course material. Digital transformation has been used in custom executive education programs at Yale, e.g. YGELP.

I recently developed and taught a masters-level elective course titled *Artificial Intelligence: Strategy & Marketing*. The objective is to help students understand the ideas, gain familiarity with the methods and their business applications. I introduce the primary ML methods (unsupervised, supervised, reinforcement and generative), followed by an examination of how organizations obtain value from AI, through case studies. I also demonstrate how fairness (and bias) can be and are often designed into algorithms, whether intentionally or not, and the resulting implications for all stakeholders. In my pedagogical experiments, I have found that assigning students to present (carefully selected) academic papers can be effective, with faculty guidance.

[6] There are exceptions (e.g. Veblen goods).

References

[P1] Steve Berry, Ahmed Khwaja, Vineet Kumar, Andres Musalem, Kenneth C Wilbur, Greg Allenby, Bharat Anand, Pradeep Chintagunta, W Michael Hanemann, Przemek Jeziorski, et al. Structural models of complementary choices. *Marketing Letters*, 25:245–256, 2014.

[P2] Cheng Chou, Tim Derdenger, and Vineet Kumar. Linear estimation of aggregate dynamic discrete demand for durable goods: Overcoming the curse of dimensionality. *Marketing Science*, 38(5):888–909, 2019.

[P3] **(*)** Cheng Chou and Vineet Kumar. Estimating Demand for Subscription Products: Identification of Willingness to Pay Without Price Variation. *Marketing Science*, 2024.

[P4] **(*)** Timothy Derdenger and Vineet Kumar. The Dynamic Effects of Bundling as a Product Strategy. *Marketing Science*, 32(6):827–859, 2013.

[P5] Timothy Derdenger and Vineet Kumar. Estimating dynamic discrete choice models with aggregate data: Properties of the inclusive value approximation. *Quantitative Marketing and Economics*, 17(4):359–384, 2019.

[P6] Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K Vishnoi. Fairness for AUC via Feature Augmentation. *arXiv preprint arXiv:2111.12823*, 2021.

[P7] **(*)** Hortense Fong, Vineet Kumar, and K Sudhir. A Theory-based Interpretable Deep Learning Architecture for Music Emotion. *Marketing Science (Forthcoming)*, 2024.

[P8] Soheil Ghili, Vineet Kumar, and Fei Teng. Spatial Distribution of Access to Service: Theory and Evidence from Ridesharing. *Available at SSRN 4915262*, 2023.

[P9] Pavel Kireyev, Vineet Kumar, and Elie Ofek. Match your own price? self-matching as a retailer's multichannel pricing strategy. *Marketing Science*, 36(6):908–930, 2017.

[P10] **(*)** Vineet Kumar and Yacheng Sun. Designing pricing strategy for operational and technological transformation. *Management Science*, 66(6):2706–2734, 2020.

[P11] Vineet Kumar. Making 'freemium' work. *Harvard Business Review*, 92(5):27–29, 2014.

[P12] Vineet Kumar and Kannan Srinivasan. Predicting Customer Value using Clumpiness – Commentary. *Marketing Science*, Mar-Apr, 2015.

[P13] Vineet Kumar and K Sudhir. Can Friends Seed More Buzz? *Management Science (Accepted)*, 2024.

[P14] Vineet Kumar, Brett R Gordon, and Kannan Srinivasan. Competitive Strategy for Open Source Software. *Marketing Science*, 30(6):1066–1078, 2011.

[P15] Vineet Kumar, David Krackhardt, and Scott Feld. On the friendship paradox and inversity: A network property with applications to privacy-sensitive network interventions. *Proceedings of the National Academy of Sciences*, 121(30), 2024.

[P16] Clarence Lee, Vineet Kumar, and Sunil Gupta. Designing Freemium: Strategic Balancing of Growth and Monetization. *Available at SSRN 2767135*, 2019.

[P17] Peter S Lee, Vineet Kumar, and K Sudhir. Monetizing Serialized Content: How Wait for Free Impacts Paid and Free Consumption. *Working Paper*, 2024.

[P18] **(*)** Ankit Sisodia, Alex Burnap, and Vineet Kumar. Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis. *Journal of Marketing Research (Forthcoming)*, 2024.

[P19] Ian Weaver, Vineet Kumar, and Lalit Jain. Nonparametric Bandits Leveraging Informational Externalities to Learn the Demand Curve. *Marketing Science (Minor Revision)*, 2024.

Note: (*) papers chosen for distribution.

Other References

[O1] George Em Karniadakis et al. Physics-informed machine learning. *Nature Reviews Physics.* 3(6). 2021