Research Statement

Vineet Kumar, *Yale School of Management*

My overall research focus is at the intersection of technology + business & society.[1] Of the many areas where they intersect, my interests broadly fall into two broad areas.

(A) *Digital business models* have specific features relative to the markets for commonly studied goods, e.g. groceries. My research focuses on strategic long-run choices relevant to digital businesses. I study this area using methods based on microfoundations of agents' preferences, typically empirical structural models.

(B) Building *theory-based machine learning (ML)* methods incorporating structured knowledege (theory) developed from first principles to be human interpretable, and provide representations satisfying required properties (e.g. monotonicity). These methods also enable new business models (creating new sources of value).

(A) Digital Business Models

Digital business models have a number of distinct and important characteristics that I explore in my research. Within digital business models, my research can broadly be themed as focusing on three different sources of connections, between products, consumers and data.

1) Connections across Products

I study connections between products (e.g. versioning, bundling, and transformation from one offering to another). A research agenda overlapping with topics here in [P1].

Bundling Hardware and Software: In [P4], I examine the dynamic effects of bundling in digital platforms (hardware+software). Should a firm create mixed bundles (products+bundles), pure bundles, or no bundles? I investigate the dynamics of bundling using data on sales and product characteristics in markets for videogame consoles (hardware) and game titles (software). In dynamic settings with intertemporal tradeoffs, bundling is more effective with *positive correlation* of valuation across products. Bundles act by pulling demand forward, i.e. consumers buy now rather than wait, a new mechanism, and higher sales obtain with positive correlation. Prior research showed that bundling is more effective with negative correlation. I develop a novel identification strategy for correlation in valuations based on the tying ratio, leveraging the market feature that consumers purchase only one hardware unit, but potentially many software units. I show that bundling is more effective when indirect network effects are weaker.

Versioning with Free and Premium Products: Freemium is the most popular digital business model (e.g. app stores, cloud storage, SaaS). Few consumers (typically <5%) migrate to the paid version, so product design is of critical importance. I undertake a deep dive into designing freemium for a storage service, examining the growth-monetization tradeoff. In [P16], I evaluate product design

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

(value of free version) and referral incentives, which impact the value of the free product, and therefore, upgrades. In counterfactuals, I find that offering a greater referral bonus relative to the firm's bonus can increase growth but reduce monetization, while beyond a threshold, it can decrease both growth and monetization. I study how to optimally structure dynamic referral incentives.

I examine the strategic use of time as a versioning and monetization strategy (wait for free) in [P18]. This strategy is commonly employed by platforms publishing content (e.g. novels as serialized episodes), with consumers having complementary value for content across episodes. I leverage a natural experiment where wait time is reduced by the platform to study the impact on downstream consumer choices with a difference-in-differences framework. Existing consumers were found to increase paid consumption, and more new consumers start reading, increasing aggregate consumption. I show how a microfounded mechanism involving complementarity can explain these data patterns. Overall, this demonstrates the strategic value of temporal versioning. I provide an overview of issues in Freemium for a general audience in [P11].

Open Source: I examine the market for open source software [P14], where products made by competing firms share common elements. Open source contributions made either by developers or by any firm are available to all competing firms (e.g. Linux or Android). It is puzzling that encouraging free-riding can lead to high quality products; a my model with interconnected markets (developer and product) explains how this happens. Developers signal their capabilities by making contributions of features to the open source software (public good). Free-riding is sustained in equilibrium since firms can build on features to differentiate on another complementary dimension (usability); appropriability can even increase product quality. Broadly, this research provides insight into how these open source contributions impact competitive structure in the product market.

Digital Transformation: In [P10], I study digital transformation strategy for a firm moving from one offering to another, typically a product based on an older technology (physical) to one based on newer technology (digital). I examine pricing and product design in this empirical setting. Using a panel from the "Netflix" model, I obtain consumer preferences for viewing content in physical form, and evaluate optimally pricing the product line. In counterfactuals with improved operational performance (or service time), including digital distribution, I uncover novel mechanisms demonstrating how improved service time (better for all customers) could result in lower profits and even lower revenue for the firm under optimal pricing, a previously unappreciated transformation risk.

2) Connections across Individual Consumers

I investigate privacy-sensitive methods for leveraging network structure to obtain higher-degree nodes in a social network [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. I examine two strategies (ego-based and alter-based), leveraging the friendship paradox, asking individuals to nominate random friends. I show that these strategies have distinct mathematical properties, and also propose a new network property called Inversity, which determines which strategy works better based on network structure. The strategies are simple to implement and offer provable guarantees of obtaining higher-degree nodes in any network.

I empirically study whether using friendship paradox strategies can achieve greater adoption in [P13]. Using a model of communication and adoption across social networks with counterfactual seeding strategies, I demonstrate that ego-based friendship seeding outperforms random seeding, and surprisingly, obtains higher adoption than leader-based seeding. A hybrid strategy combining the ego-based and leadership-based strategies performs marginally better. The results hold across a range of specifications and different networks, demonstrating robustness and empirical value.

Finally, I also look at the impact of user-to-user connections on referrals which forms an important strategic lever in the freemium business model of Dropbox.

3) Connections across Data

The third linkage that I examine is the linkage between types of data, i.e. purchase data and usage data. In digital settings, usage data is uniquely available and can be very valuable to understand and help provide insights about consumer preferences and design product & pricing strategy. However, most studies in marketing and economics only involve purchase data. I connect usage and purchase data across several settings, including [P16, P10]. In [P3], I demonstrate how usage data is conceptually distinct and is critically important for identification, not just estimation. Specifically, identification of the Willingness to Pay (WTP) distribution in subscription settings without price variation is possible by leveraging usage data. The key insight is that using high-frequency usage data in conjunction with purchase data allows for this conceptual advance in identification. I combine usage data with factors exogenously impacting usage to first estimate usage utility, then aggregate this stream to combine with purchase data to obtain the WTP distribution. This approach then allows us to conduct counterfactual analyses. The framework is flexible enough to accommodate a large class of usage utility models, making it widely applicable.

Methodological Overview and Contributions: Theory or structured knowledge is central to my microfounded models. This structural approach yields estimates with clear economic interpretations, evaluating the impact of firm or regulatory policy decisions. During the work on some of the substantive projects above, I investigated methods for dynamic demand models in technology markets. Examining the commonly used inclusive value approach, and showed that it could lead to highly biased estimates of economically important quantities like elasticity and profits [P5]. Motivated by the need for models to flexibly accommodate dynamics with large state spaces, I developed a new method [P2]. For a large class of problems (with terminating or renewal choice) using market-level data, it obtains consumer preferences with the computational complexity of a linear regression. The estimation procedure is accessible with a "recipe". Identification is proven formally and the method can be used for counterfactual analysis.

(B) Theory-based Machine Learning

My research focus in ML is based on integrating structured knowledge to develop better ML, which in turn, enables us to learn more about consumers and firms. I discuss why such an approach is needed, how I incorporate theoretical knowledge, and the benefits that accrue from doing so. There is a growing set of important research questions where ML is required. For instance, traditional non-ML quantitative methods are not great at capturing the nuances of unstructured data. ML also makes it feasible to generate novel unstructured data like text or images, like in [P7, P19]. With structured numerical data, reinforcement learning provides non-parametric methods with minimal assumptions that feature strong theoretical guarantees, as in [P20].

However, traditional ML models have typically been designed to be atheoretical and domain-independent. The same models (deep learning using CNNs) used for predicting breast cancer in imaging (in medicine) are used for predicting the presence of an exoplanet (in astronomy). What is the connection between medicine and astronomy? From a substantive viewpoint, not much![2]

My view is that developing better ML methods based on theory or structured knowledge has significant potential to advance the field in business (and more broadly, social science) applications.[3] My background, being fluent with both microfounded theory-based models and in ML, has enabled me to bring a unique perspective to research. The sources of knowledge that I examine include key ideas and concepts from fields like marketing and economics, but is not restricted to these. I show how such knowledge can be used across both structured and unstructured data.

Why do this? There are a few specific reasons. First, ML algorithms are typically evaluated based on their performance, typically maximizing accuracy or minimizing regret. I show that incorporating structured knowledge improves the performance of the algorithms substantially. Second, an equally important goal for me is to achieve human explainability and interpretability. ML methods have long been viewed as complex black box methods that use enormous amounts of data gathered from a variety of sources. We often don't have visibility either into the sources of data or the methods even for commonly-known foundation models, e.g. ChatGPT. Most are based on deep learning with neural networks featuring many interconnected layers with billions of parameters. The result is that no human can actually say what a specific parameter means or exactly how the model works, making them black box models. If we don't understand complex models, we don't understand their limitations, and where and how they could fail.[4] I expect that incorporating theory into ML provides transparency and improves stakeholder acceptance and trust in AI systems.

However, building in theoretical foundations is typically challenging. First, rather than using closed commercially available black-box models (e.g. ChatGPT), I develop models from the ground up from basic elements, with complete visibility and control. All of my ML research is transparent, and the open source code is publicly available for others to examine and build on. Second, there are significant challenges in trying to connect theory to deep learning models, both conceptually and in implementation, especially in complex deep learning models focused on predictive performance. However, this holds even in the case of reinforcement learning with structured data.

I quantify consumer preferences for visual characteristics in [P19]. Visual appearance is high-dimensional and hard to characterize and explain, without human input. I develop a theory-based

[2] This is a deliberate choice, because the focus of these ML methods is to be broadly applicable to the widest class of applications. The underlying idea is to develop better "pattern matching" algorithms, without being concerned much about structured knowledge that is relevant to any one domain.

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

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

algorithm to automatically discover and quantify visual characteristics of products. 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 is the idea that products often have a distinct look tied to the brand, with recognizable "visual signatures" like LV handbags or BMW cars. Brands typically have a consistent aesthetic, and consumers form expectations around this. The method extracts visual characteristics from product images using brand and other characteristics to supervise disentanglement. The method obtains significantly better performance *and* interpretability, and importantly, can *generate* counterfactual visual designs without human judgment.

In the music emotion research [P7], the ideas about consonance and dissonance of music, and how that connects to the listener's emotion is aided by using domain knowledge (theory) from multiple fields. Specifically, the knowledge is based on both: (a) the mathematics of sound waves and (b) psychology of human music perception. Explaining why a listener feels a specific emotion when listening to music was a black box. Here, theory is used as the basis for creating flexible and non-contiguous consonance filters, helping obtain a representation that enables explainability, so we can visualize the features of music impacting listener responses.

To learn an unknown demand curve by experimentation, I develop a reinforcement learning model with non-parametric multi-armed bandits (MAB) [P20]. Economic theory informs us that demand curves are downward sloping.[5] The classic experiment (A/B testing) benchmark is inefficient since it explores all prices equally, whereas MABs provide a more sophisticated approach ("learning while earning"), but including theory here is not quite obvious. Incorporating monotonicity adds two sources of value. First, it improves the performance of the algorithm substantially since the algorithm learns not just from each individual price (or arm) experimented, but across arms (an informational externality). Second, including theoretical guarantees that the resulting demand curve obtained is monotonic, especially important when algorithms are trusted and used in an automated way to make pricing decisions. Without this, we can get an upward sloping demand curve and the algorithm could choose an unrealistically high price, resulting in substantial risks.

Overall, all of the ML-based research I have undertaken brings in the power of structured knowledge (or theory) to enhance their capability and suitability for relevant applications. ML algorithms improve along several aspects: performance, interpretability / explainability, and providing representations satisfying desirable properties. In turn, these ML methods are used to help us gain valuable insights into consumer and firm behavior, leading to a symbiotic process.

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

Teaching

I have developed and taught elective courses at the masters and doctoral levels, and have also contributed to Executive MBA and Executive Education. I use a mix of lectures to help with learning principles, complemented with discussions featuring case studies using a socratic approach.

In *Digital Strategy*, I explore digital business models, such as 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 help students can understand the current state of the field and business applications and its future development. I provide an overview of of AI & ML, a historical view, and an understanding of the primary ML methods (supervised, supervised, reinforcement and generative). I explore how organizations obtain value from AI using a variety of case studies (retail, ridesharing platfoms, and medical). I also demonstrate how fairness (and bias) can be and are often designed into algorithms, whether intentionally or not. Regarding pedagogy, I've experimented with, and found that assigning students to present (carefully selected) academic papers can be effective, with faculty guidance. I'm impressed by how well students can navigate complex ideas, identify downstream implications, and effectively communicate this knowledge to their peers.

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. Intertemporal Price Discrimination for Serialized Media Products. *Working Paper*, 2024.

[P18] Peter S Lee, Vineet Kumar, and K Sudhir. Monetizing Serialized Content: How Wait for Free Impacts Paid and Free Consumption. *Working Paper*, 2024.

[P19] Ankit Sisodia, Alex Burnap, and Vineet Kumar. Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis. *Journal of Marketing Research (Forthcoming)*, 2024.

[P20] Ian Weaver, Vineet Kumar, and Lalit Jain. Nonparametric Bandits Leveraging Informational Externalities to Learn the Demand Curve. *Available at SSRN 4151019*, 2024.

Note: (*) papers chosen for distribution.

Other References

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