MSI Webinar: InnoVAE: Generative AI for Understanding Patents and Innovation

September 5, 2023 | Virtual | 12:00 pm – 12:30 pm EDT

Speaker:

Dokyun Lee – *Questrom Chair Associate Professor of Information Systems and Digital Business Fellow at Questrom School of Business, Boston University*

Overview:

Dokyun Lee (Boston University - Questrom School of Business) presents research he did with Zhaoqi Cheng (Boston University - Questrom School of Business) and Prasanna Tambe (Wharton School, U. Pennsylvania) on the ability of Generative AI to produce complex objects (images, documents, patents, portfolios, etc.). Lee notes that such models use large Internet data plus MADlib and human feedback to assemble knowledge and, though they do show some "reasoning capabilities," they are not appropriate for mapping and understanding. In terms of using generative AI for patents and innovation, Lee poses the research questions: Can AI represent innovation, and how is this useful? In response, Lee details how this process can be executed through a prototype called InnoVAE (variational autoencoder) to estimate disentangled representations of patents, using multimodal data, including structured and in-text data. He explores how to use this technology to combine patent information to highlight areas of current and potential innovation. Lee then provides examples of several applications: fusing patents (combinational creativity); firm-level exploration (ranking firms on innovation); identifying gaps in technology; identifying areas of competition; and ranking the exceptionality of patents by levels of creativity.

Takeaways:

Generative AI Overview – Strengths and Weaknesses

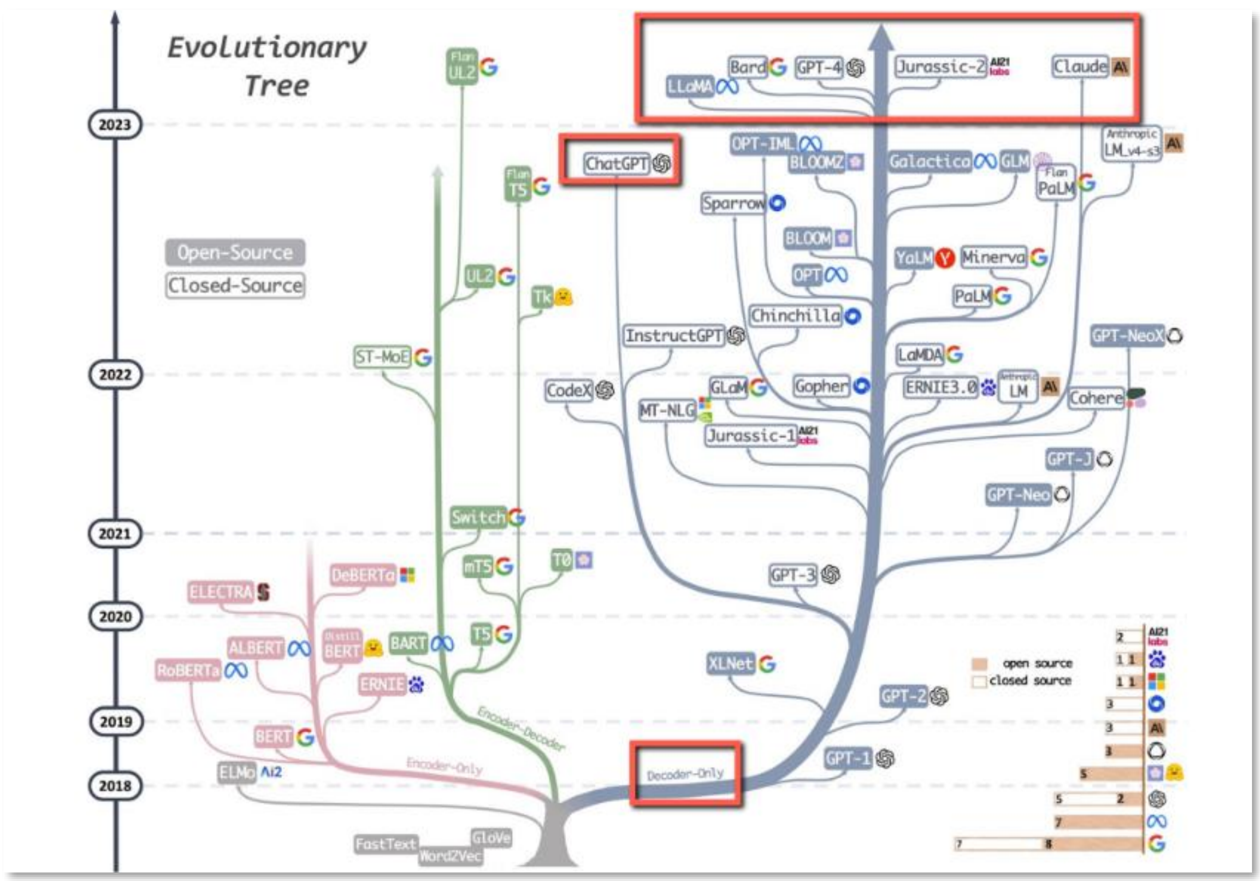
- Generative AI is a model data generative process. In short, Generative AI "generates things."
  - As opposed to discriminative AI which models the probability of Y given X, similar to regression and classification models, generative AI just outright models the probability of X or X and Y.

MSI at the ARF
MARKETING SCIENCE INSTITUTE

Discriminative AI estimates $P(Y|X = x)$, regression/classification models

Generative AI estimates $P(X)$ or $P(X, Y)$

- A depiction of the Large Language Model Tree (Evolutionary Tree) demonstrates that **many mainstream LLMs are decoder-only models** that learn by predicting the next word(s).
  - These models **use large internet data plus MADlib and human feedback to assemble knowledge** and while they do show some "reasoning capabilities" **they are not appropriate for mapping and understanding**.



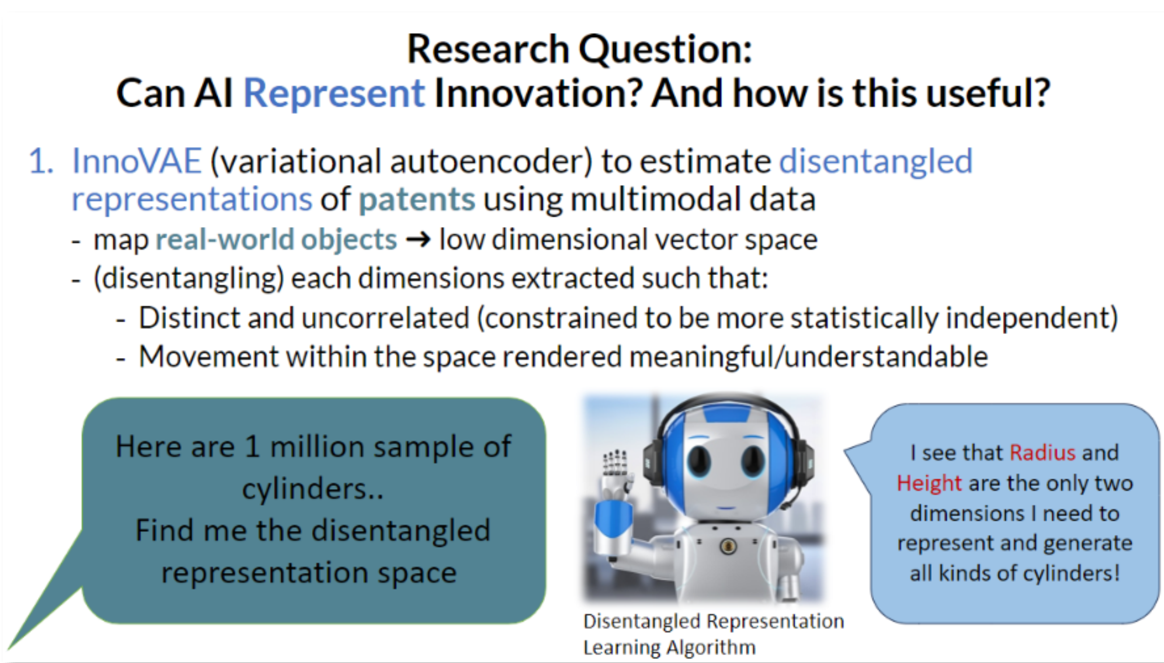
Generative AI's Powerful Capability Summarized

- **Generative AI is able to learn to generate any complex objects** (e.g., multimodal business objects) and can include:
  - Images, documents, patents, jobs, firms, portfolios, consumers and the digital twin of anything, when provided enough data.

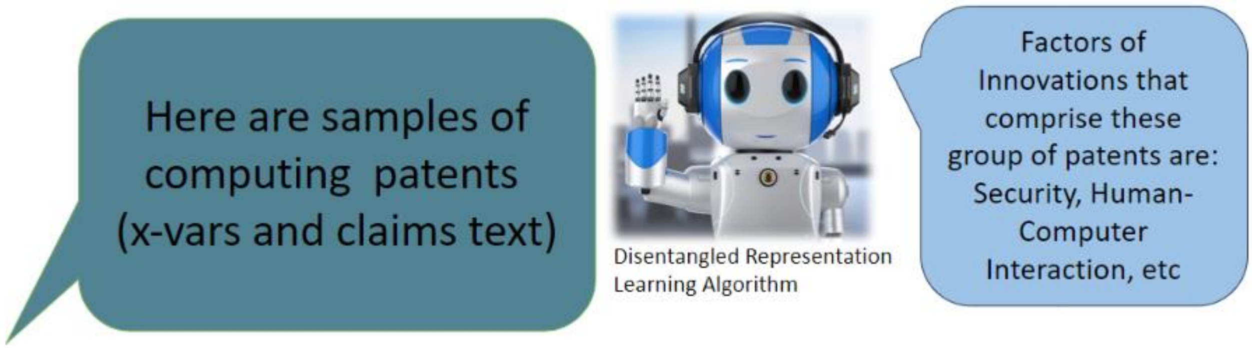
- In generating complex objects, **generative AI can learn object space and compositions in scale**.
    - After learning object space and compositions in scale, a **trained model can then map out the object space and provide deeper insights** enabling comparing and contrasting.
        - This can provide us the ability to extract insights and augment purposeful synthesis, **giving humans a say in how the AI model should "generate a new synthesis"** and to see if it will be successful, and do so **at incredible speed and scale.**

InnoVAE: Generative AI for Patents, Innovation and Firms

- Research Questions: **Can AI represent innovation, and how is this useful?**
    - **A prototype was created called InnoVae** (variational autoencoder) to **estimate disentangled representations of patents**, using multimodal data, including structured and in-text data.
        - **Representation learning** is taking a real-world object, in this case a patent, and embedding it, **finding a vector of representation while preserving certain characteristics** such as a local region and similar objects residing near each other.
        - **Disentangling** is the act of **putting a constraint on the space** so that each of the dimensions is extracted to be distinct and uncorrelated.
            - It is constrained to be statistically independent so that the movements within the space are rendered "meaningful and understandable."

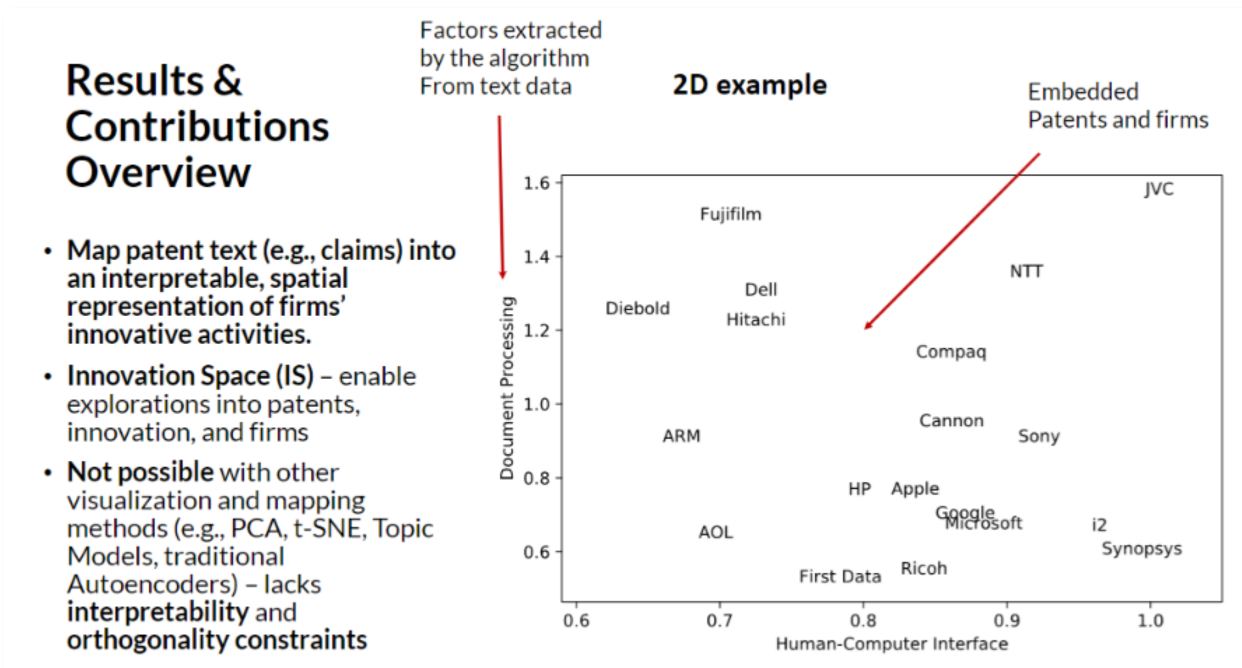


- **When applied to patent data, patents would now reside in our factors of innovation that comprise the set of patents.**
  - For example, **computing patents** may reside in dimensions like **security, human-computer interaction**, etc.



Results and Contributions Overview

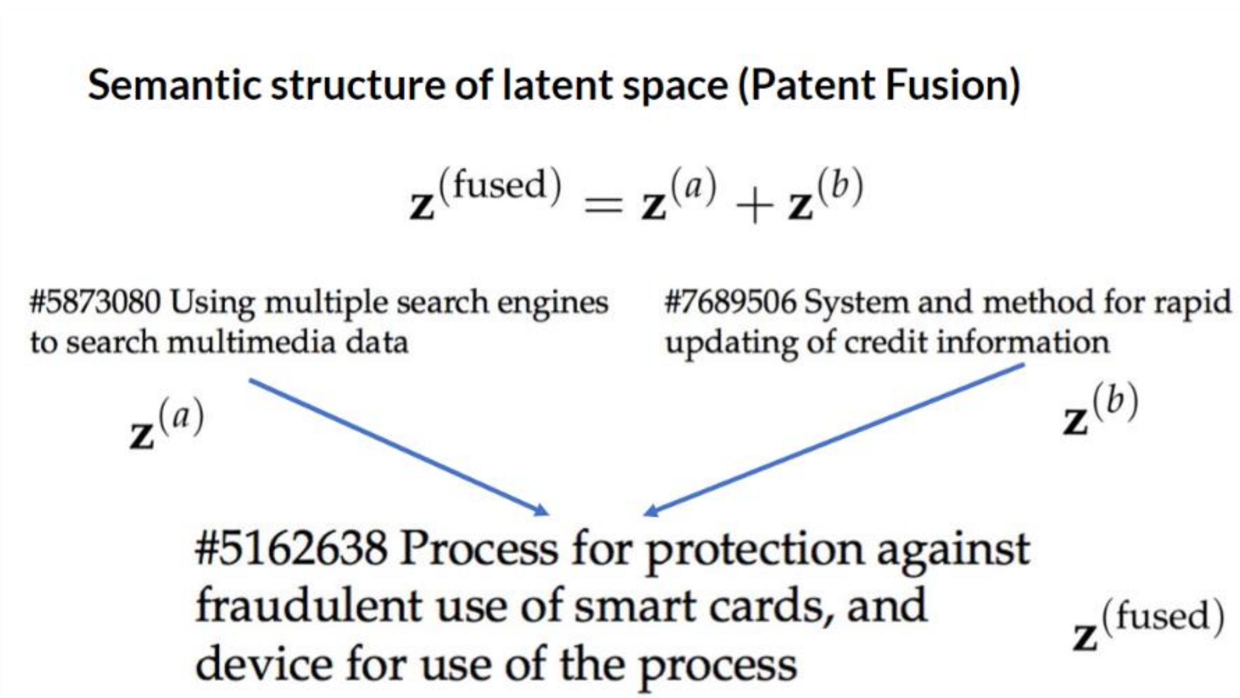
- Leveraging generational AI to the patent space **enables exploration into patents, innovations and firms** (Innovation space).



- **Good disentangled representation (Innovation Space) enables users to ask and explore:**
  - What could you get if you combine patents A and B (e.g. automate combinational creativity)?
  - How unusual is a patent (e.g., iPod interface) concerning a specific technological factor (e.g., user interface)?
  - What innovation factors increase/decrease over time?

- Rank companies in technological factor X (extracted from claims by the algorithm) .
- If I am Firm A, what innovation do I need to boost up to be more like Firm B?
- How do firms move in Innovation Space over time and how does that correlate to some performance?
- What happens to innovation activity in a specific technological region after event X (i.e., acquisition, mergers)?

- **An example where the data context is patents on computing systems** (approximately 240,00 patents from the 1980s-2010), this would include data from PageRank by Google and Multipoint Touch by Apple, which would yield the following innovation factors (Disentangled Axis):
  - Finance-transaction axis, automation and control axis, medicine, connectivity, information, retrieval, ergonomics, hyperconverged infrastructure (HCI) hardware, security, broadcasting, document processing, and manufacturing among others.
    - **Applied in a set of different groups of patents will yield different axes (results).**
- In an example, **two patents can be combined** (Patent Fusion): **A patent for using multiple search engines to search multimedia data** and another one **searching a patent for system and method for rapid updating of credit information**.
  - **We can ask the algorithm to generate a new patent (closest relevant patent).**



- **Firm-level patent exploration:** Firms can be characterized by the patents they own.
    - Based on technological patent factors that correspond to an axis in the innovation space that is extracted by the algorithm, **firms can be ranked by innovation factors.**

Innovation Factor & Top Ranking Firms (G06, 1980-2010)

Technological Factor i	Most Innovative Firms	Firm's Main Business Line	Innovation Index δ_i
Human-Computer Interaction	Nintendo	Video game	2.0781
	Pixar	Computer animation	1.6756
	Immerson	Haptic technology	1.6254
Automation / Control	Intertrust	Digital rights management	1.6756
	Silicon Motion	Hardware	1.6254
	Toyota	Automobile	1.2582
Finance / Transaction	VISA	Finance	2.3951
	CME	Exchange	1.3542
	Salesforce	Customer relationship management	1.2225
Connectivity	Wells Fargo	Finance	1.1004
	West Corp.	Telecommunication	1.0803
	CommVault	Data management	1.0668
Document processing	Fuji	Document solutions	1.4982
	NTT	Telecommunications	1.3436
	Dell	Computer products	1.2941

- With patent representation for each of the companies, **each of the firms can also be represented in the innovation space** by taking the centroid of all the patents that these firms own.
    - In addition, a firm representation in each year can be found and then the algorithm can generate a synthetic representative patent abstract.
        - **This can yield results in what innovation was occurring at a firm at a specific time.**

Firms & Their Synthetic **Representative Patent Abstract in year X**

A method comprising: receiving, by a computing device, a request to access a **social networking system**; determining, by the computing device, **a first set of social networking rules associated with the request to access the social networking system based on a user profile of the user**; determining, by the computing device, a second set of social networking rules associated with the request to access...

2012

Meta

2007

A method comprising: receiving a **request to create a multi-dimensional representation of a user interface**, the request comprising a plurality of input data; determining whether the multi-dimensional representation is a valid representation of the user interface …

2011

a method for providing a user interface for a computer system have a plurality of computer system, each of which are capable of communicating with a computer system…

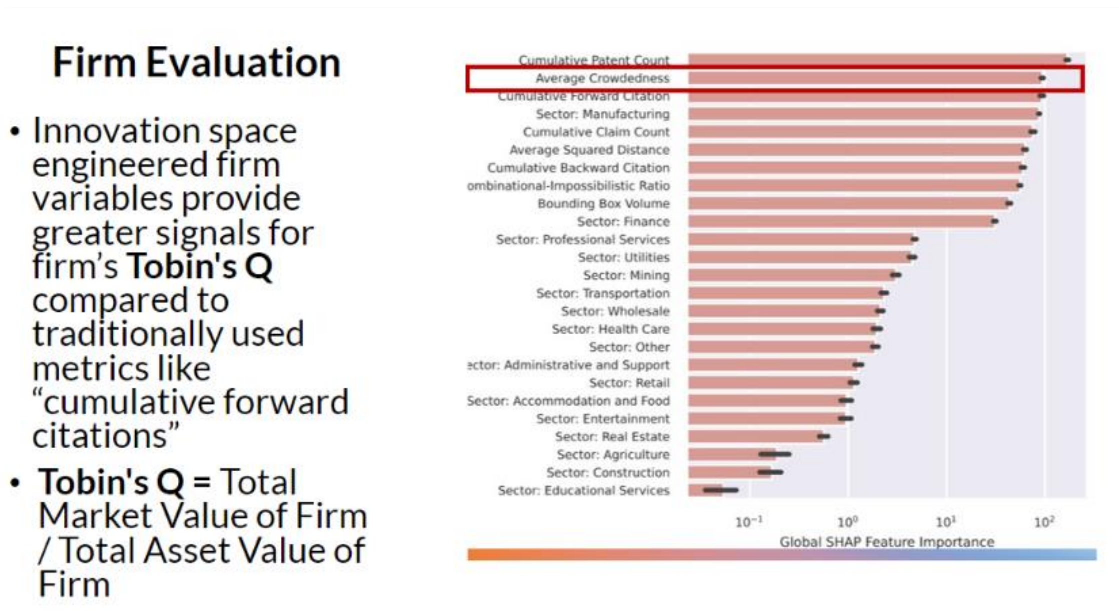
- The innovation space can enable users to **access factors of innovation to understand technology gaps** between incumbent and entrant firms (i.e. consider entering a new market).
    - This can yield differences in each firm's innovation factors.

Table 8: Technology gaps between incumbent and entrant firms

Incumbent	Entrant	Factors of Technology Gap
Toyota	Google	Automation (0.599), Manufacture (0.315), Security (0.155)
	Apple	Automation (0.526), Manufacture (0.264), Connectivity (0.200)
Honda	Google	Information retrieval (0.183), Security (0.171), Ergonomics (0.160)
	Apple	Ergonomics (0.213), Connectivity (1.89), Information retrieval (0.126)
Ford	Google	Security (0.233), Hardware (0.227), Automation (0.179)
	Apple	Connectivity (0.203), Security(0.173), Ergonomics (0.149)

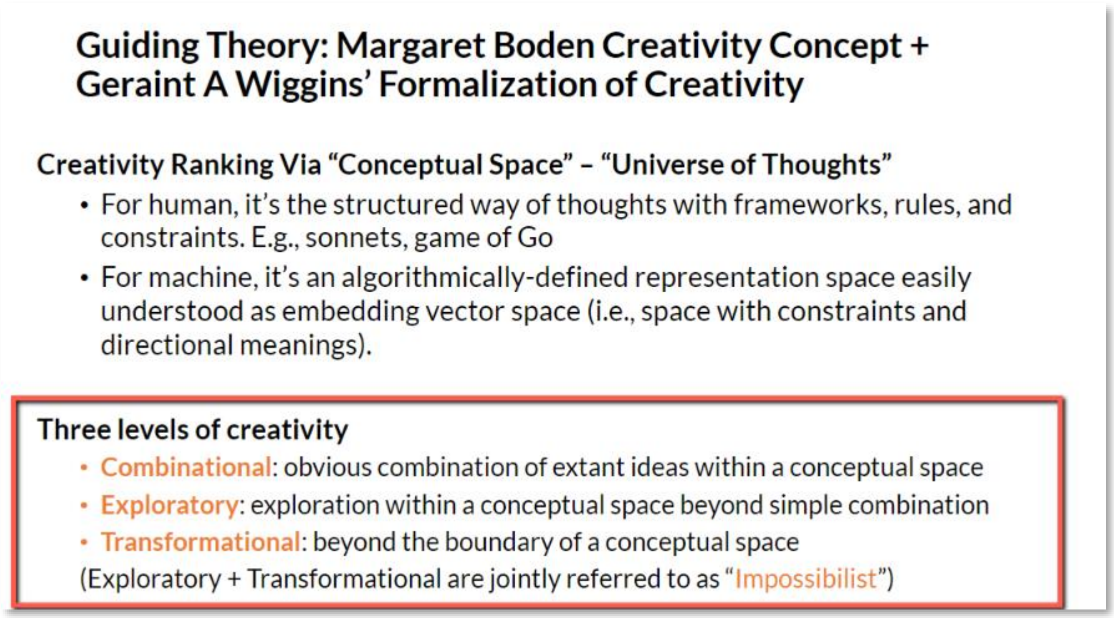
(a) Automobile market

- Gathering all patents owned by a firm can yield **how crowded (competitive) a particular innovation space is.**

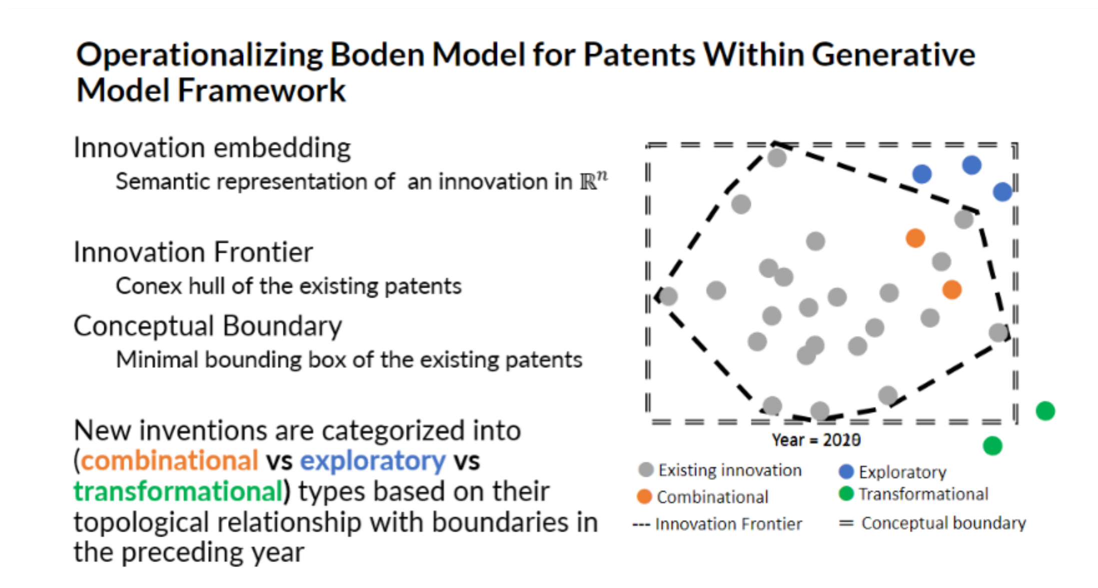


- **Patent level exploration:** Ranking the overall exceptionality of patents.
    - **Following a specific set of rules and frameworks** within this conceptual space can allow for the instantiation and **generation of a new "creative artifact."**
    - **In terms of machines, this is an algorithmically defined space** understood as "embedding vector space," pointing back to constraint and directional meanings (What InnoVAE has estimated).
    - In this conceptual space, there are **three levels of creativity:**

- **Combinational is the easiest form of creativity.** It is a **combination of extant ideas within a conceptual space**. So any sort of startup ideas that are "things like Google of this, Apple of that Uber, of this, Airbnb of that, is a combinational idea."
- **Exploratory** is the instantiating of a **new point in the conceptual space** without the need for **combining existing things.**
- **Transformational** is the highest form of creativity which is essentially **defining a new conceptual space.**



- In terms of **operationalizing this in our innovation space**, for instance, there is a representation of patents in a vector space innovation space in the year 2019. These are existing innovations. Then boundaries are defined.



Sources:

InnoVAE: Generative AI for Understanding Patents and Innovation.
Source: Cheng, Z., Lee, D., & Tambe, P. (2022). SSRN.

A lack of interpretability limits the use of common unsupervised learning techniques (e.g., PCA, t-SNE) in contexts where they are meant to augment managerial decision-making. We develop a generative deep learning model based on a Variational AutoEncoder ("InnoVAE") that converts unstructured patent text into an interpretable, spatial representation of innovation ("Innovation Space"). After validating the internal consistency of the model, we apply it to three decades of computing system patents to show that our approach can be used to construct economically interpretable measures—at scale—that characterize a firm's IP portfolio from the text of its patents, such as whether a patent is a breakthrough innovation, the volume of intellectual property enclosed by a portfolio of patents, or the density of patents at a point in Innovation Space. We show that for explaining innovation outcomes, these interpretable, engineered features have explanatory power that augments and often surpasses the structured patent variables that have informed the very large and influential literature on patents and innovation. Our findings illustrate the potential of using generative methods on unstructured data to guide managerial decision-making [Abstract from the author(s)].

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.
Source: Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). arXiv.

This paper presents a comprehensive and practical guide for practitioners and end-users working with Large Language Models (LLMs) in their downstream natural language processing (NLP) tasks. We provide discussions and insights into the usage of LLMs from the perspectives of models, data, and downstream tasks. Firstly, we offer an introduction and brief summary of current GPT- and BERT-style LLMs. Then, we discuss the influence of pre-training data, training data, and test data. Most importantly, we provide a detailed discussion about the use and non-use cases of large language models for various natural language processing tasks, such as knowledge-intensive tasks, traditional natural language understanding tasks, natural language generation tasks, emergent abilities, and considerations for specific tasks. We present various use cases and non-use cases to illustrate the practical applications and limitations of LLMs in real-world scenarios. We also try to understand the importance of data and the specific challenges associated with each NLP task. Furthermore, we explore the impact of spurious biases on LLMs and delve into other essential considerations, such as efficiency, cost, and latency, to ensure a comprehensive understanding of deploying LLMs in practice. This comprehensive guide aims to provide researchers and practitioners with valuable insights and best practices for working with LLMs, thereby enabling the successful implementation of these models in a wide range of NLP tasks [Abstract from the author(s)].