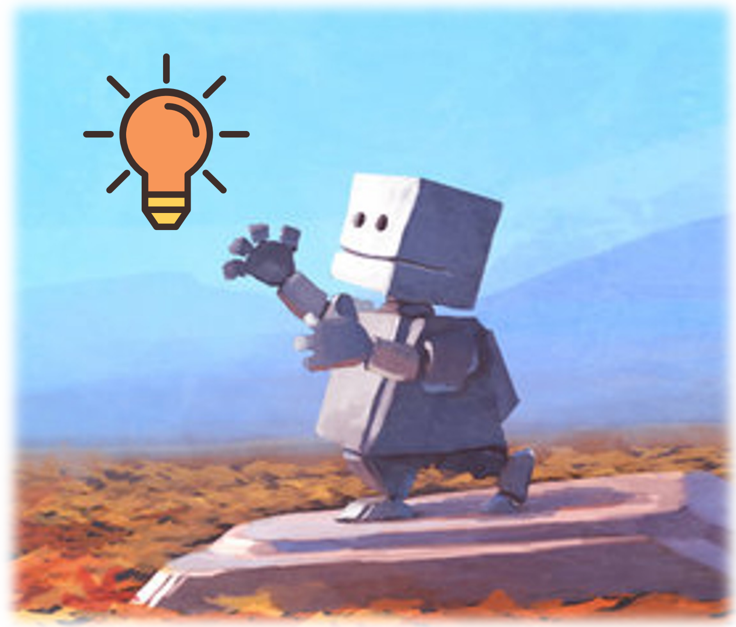
InnoVAE: Generative AI for Patents and Innovation



Paper at: tiny.cc/innovae

Zhaoqi Cheng	**Boston University**
Dokyun "DK" Lee	**Boston University**
Prasanna "Sonny" Tambe	**UPenn**

Patent Semantic-Based Innovation Vector Space Enables

- Better patent similarity comparison
- Characterization of firms' location in innovation-space
- Knowledge complementarity comparison between firms
- Etc

Which has implications for

- Acquisition & merger decisions (e.g., Makri et al. (2010), Cloodt et al (2006))
- Competition strategies (e.g., Ernst 2003)
- Evolution of industries (e.g., Helfat 1997)
- Evolution of Innovation (e.g., Ahuja et al 2008)
- Etc

Examples of Representation Methods for Innovation Corpora

- Token-based feature engineering
  - Kelly et al. (2021) operationalized the idea that a novel patent should resemble prior arts less, but future inventions more. Modified tfidf.
  - Gatchev et al. (2022) uses the emergence of new token in 10-K filing records as the indicator for novel innovation.
- Embedding-based models
  - Volkov et al. (2019) patent-to-patent similarity metric using document embedding.
- Topic modeling
  - Teodoridis, Lu and Furman (Working paper) applied Hierarchical Dirichlet Process on patent abstracts to map knowledge landscape.

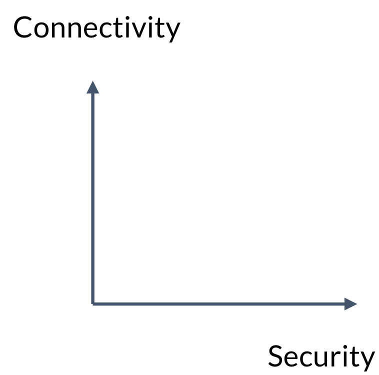
Challenges in Semantic-Based Patent Representation Methods

- Focused on solving one specific problem (e.g., similarity, breakthrough patent identification)
- Learned manifolds are not the focus (e.g., non-regularized, nonlinear)

Therefore, interpreting distance in these spaces are hard

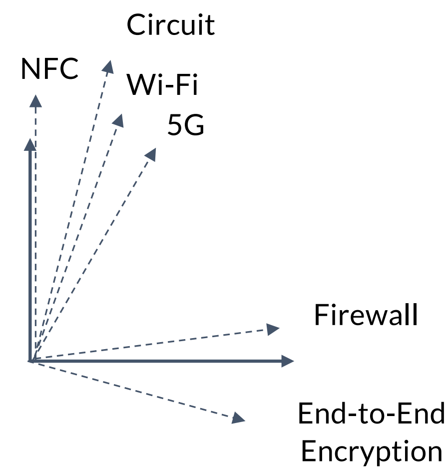
Challenges in Semantic-Based Patent Representation Methods

- Semantic orthogonality & dimension independence is not internalized in the model → **Interpretability Problem**



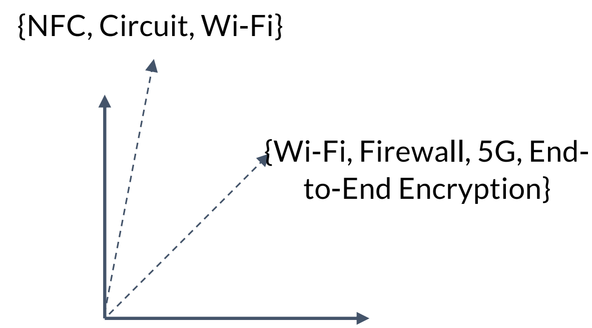
Human
Knowledge

- Human-labor Intensive
- Context-dependent



TF-IDF Or
Word Embedding

- Single word – too simplistic
- No relationship between a set of words



Topic Modeling

- Topic Diffusion
- Coherence problems
- Lack of Semantic Similarity Knowledge

Can we represent patents/innovation in a more interpretable vector space?

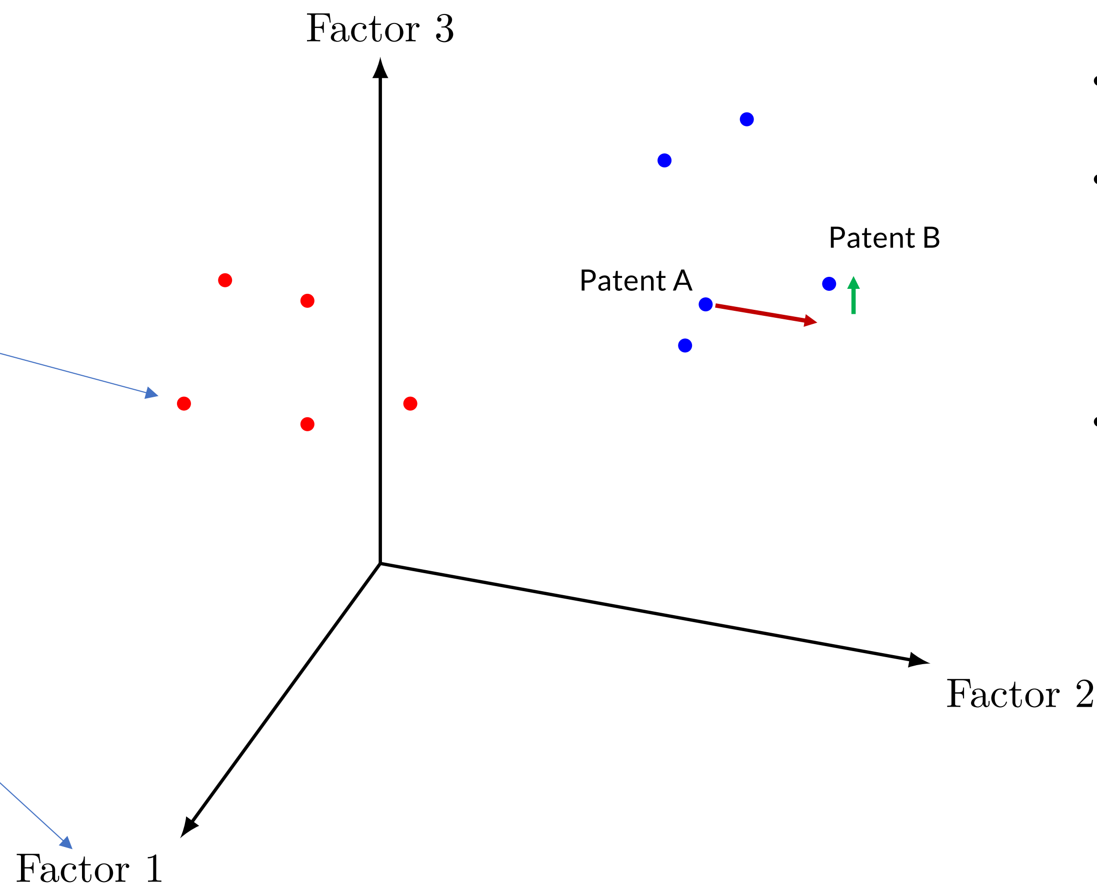
1. Prototype InnoVAE (variational autoencoder) to estimate disentangled representations of patents using structured/unstructured data
2. Representation Learning: map real-world objects ➜ low dimensional vector with preserved properties. i.e., similar patents are local to each other in this space.
3. Disentangling: each dimensions extracted such that…
   - Statistically more independent & semantically meaningful
   - Movement within the space rendered understandable
4. If successful, patents now reside in an interpretable vector space that characterize patents by their factors of innovation.

   e.g., Computing patents (G06) may reside in dimensions like “security”, “connectivity”

Contributions Overview

- **Innovation Space (IS)** - facilitate explorations into patents, innovation, and firms (providing distance and movement measures).
  - scalably construct economically interpretable measures that characterize a firm's IP portfolio from the text (+ structured) data of its patents over time
    - breakthrough innovation or not
    - volume of IP enclosed by a portfolio of patents
    - the density of patents at a point in Innovation Space.
- Firm-level characteristics engineered from **IS** are as predictive as the cumulative number of patents or forward citations predicting firm-level quality measures (Tobin's Q)

Visualization of Innovation Space from InnoVAE

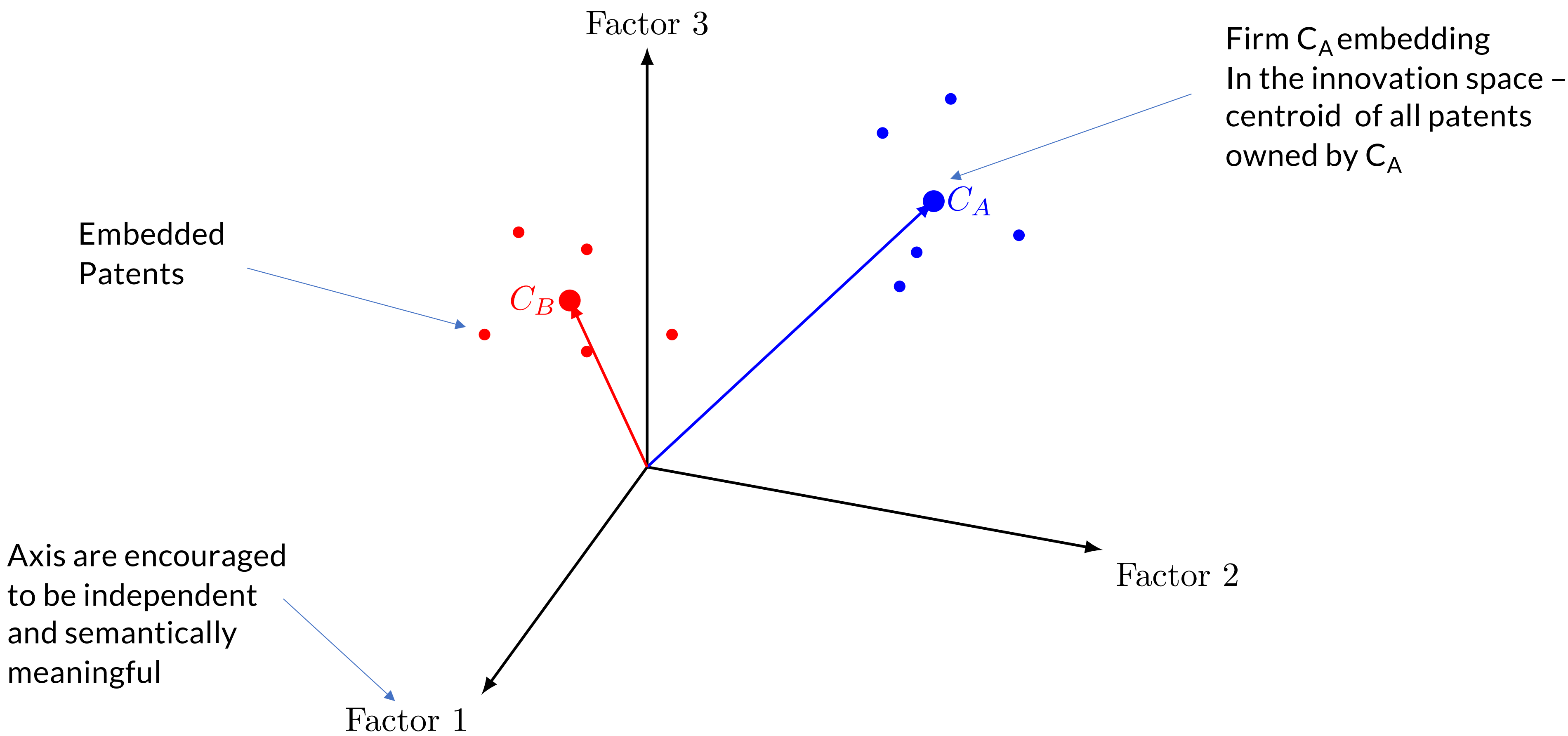


Embedded Patents

Axis are encouraged to be independent and semantically meaningful

- Factors are distinct & Data-driven
- Similar patents are near each other
- Directions **are meaningful.** Patent B has increased factors 2 and 3 compared to patent A (i.e., more exceptional)
- **Harder** with other visualization and mapping methods (e.g., PCA, t-SNE, Topic Models, traditional Autoencoders) – lacks **dimension independence & vector space regularization constraints**

Visualization of Innovation Space from InnoVAE

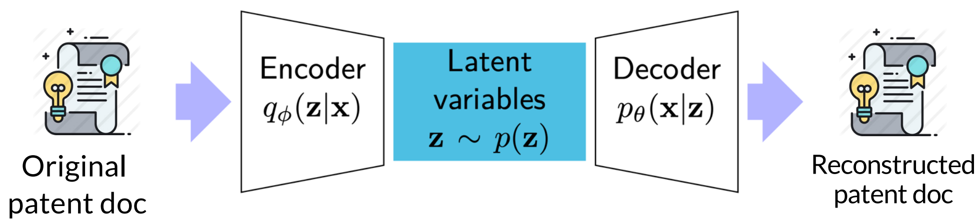


Good disentangled representation (Innovation Space) enables researchers to ask and explore:

1. What could you get if you combine patent A and B? (automate combinational creativity for abstracts/claims/etc)
2. How exceptional (unusual) is a patent (e.g., iPod related) with respect to specific technological factor (e.g., user-interface)?
3. What innovation factors inc/dec over time?
4. Rank and compare companies in innovation factor X
5. How do firms move in Innovation Space over time and how does that correlate to some performance?
6. What happens to innovation activity in specific technological region after event X (i.e., acquisition, mergers)
7. Etc…

Situate Patents in Innovation Space via Variational Autoencoder

- Represent patents as N dimensional vectors using "controllable" VAE
- Background: Variational autoencoder (VAE)



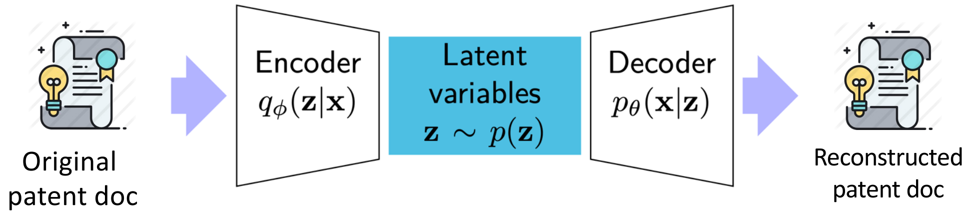
- Data-generating assumption
  Step 1: "humanity" sample a random (latent) vector $\vec{z} \sim p(\vec{z})$ as **innovation embedding**
  Step 2: "humanity" implement the concept vector $\vec{z}$ into **real-world instance $\vec{x}$**

- Customized objective function (for disentangling and self-supervision)

$$\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \underbrace{\mathbb{E}_{q_{\phi}(\mathbf{w},\mathbf{y}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{w},\mathbf{y})]}_{\text{Reconstruction loss}} - \underbrace{\beta_0\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{w}|\mathbf{x})\|p(\mathbf{w})\right) + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{x})\|p(\mathbf{y})\right)\right]}_{\text{Disentanglement-inducing term}} - \underbrace{\beta_1\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{x})\|p(\mathbf{y}|\mathbf{x})\right)\right]}_{\text{Supervision loss}}.$$

This approach can be thought of as…

- nonlinear factor analysis/dimension reduction algorithm (with benefits)



Differences and **benefits** are:

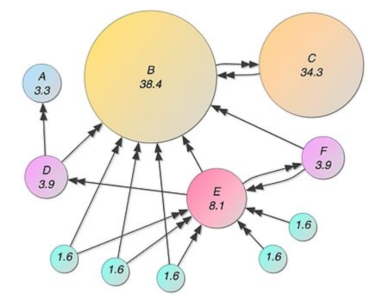
1) Multimodal – can incorporate structured/unstructured
2) Power of NN performance/framework – easier to add constraints
3) Controlled generation of any multi-modal entity
4) Supervision with known key variables (# of claims)
5) Internalized disentangled representation (interpretability and semantically more orthogonal axis)

Data Context

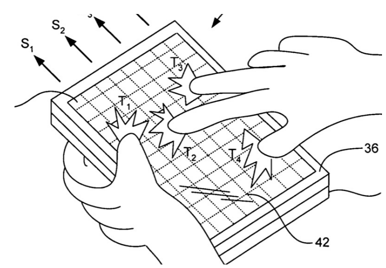
- US patent filed under “G06 “category
  - Patents on computing systems
  - 240K patent between 1980 to 2010

- Each datapoint contains:
  (1 textual feature)
  - Patent abstract

  (5 numeric features)
  - Wordcount of the abstract
  - Number of patent claims
  - Backward patent citations
  - Backward non-patent citations
  - Bibliometric originality[1]



PageRank
by Google
(US# 6285999)



Multipoint Touch
by Apple
(US# 7663607)

Examples of patents under G06 group

[1] Trajtenberg, M., Henderson, R. and Jaffe, A., 1997. University versus corporate patents: A window on the basicness of invention. Economics of Innovation and new technology, 5(1), pp.19-50

Innovation Factors Extracted

Latent dimension	Correlated phrases
Abstract wordcount	<EOS>, space, punctuation marks
Non-patent citing	java ^{TM}, garbage collection, probabilistic model
Broadcast	Broadcast, television, video, program, distribute, receiver, digital, broadcasting, distribution, connect
Ergonomics	skin, eye, body, face, person, surface, glass, position, gaze, say
Information Retrieval	delivery, ink, printer, receive, recording, transfer, scanner, handwriting, film, read
Hardware	signal, peripheral, interrupt, hardware, data, medium, storage, volumn, drive, array
Human-computer interaction	Interface, control, mouse, texture, color, polygon, body, reflection, render, virtual
Automation / Control	forecast, simulation, route, controller, driver, engine, configure, device, function, emulate
Finance / Transaction	payment, money, transfer, order, recipient, sender, merchant, payer, payee, request
Manufacture	panorama, motor, fan, mosaic, mainframe, radiographic, cool, tomogram, vehicle, duct
Connectivity	transmission, port, connect, ultrasound, remote, communication, transmit, magnetic, memory, allocate
Medicine	track, diagnostic, diagnose, surveillance, medical, imaging, device, image, recognition, motion
Security	check, authenticate, verify, malicious, authority, identification, protect, secure, signature, integrity
Document processing	document, image, extract, processing, design, attach, read, digital, notebook, deploy

Table 3: Selected supervised and unsupervised latent dimensions with their correlated phrases

Semantic structure of latent space (Patent Fusion)

$$\mathbf{z}^{(\text{fused})} = \mathbf{z}^{(a)} + \mathbf{z}^{(b)}$$

#5873080 Using multiple search engines to search multimedia data

$\mathbf{z}^{(a)}$

#7689506 System and method for rapid updating of credit information

$\mathbf{z}^{(b)}$

#5162638 Process for protection against fraudulent use of smart cards, and device for use of the process

$\mathbf{z}^{(\text{fused})}$

Innovation Index

- Given patent x, innovation Index i : absolute difference between posterior mean and prior mean at dimension i

$$\delta_i = \left| \mathbb{E}[z_i | x] - \mathbb{E}[z] \right|$$

- High index: The patent is exceptional in this technological factor
- Low Index: The patent is average in this factor

Innovation Factor & Top Ranking Firms

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

Sanity Check Validation:
On Predictive Signal Strength of Latent Representation

Predictive power tested on:

- Tobin's Q
- Kogan Value
- Etc

Against

- Topic model
- TF-IDF
- Embedding approaches

Our latent dimension representation useful for simple downstream task

Additional Validation

In theory, disentangled representation should have better generalization, interpretability, and performance in downstream task (Bengio et al 2013)

Shown to be true in many tasks in CS such as prediction tasks, reinforcement learning, visual reasoning, QNA, etc.

(Higgins et al., 2017b; 2018b; Achille et al., 2018; Steenbrugge et al., 2018; Nair et al., 2018; Laversanne-Finot et al., 2018; van Steenkiste et al., 2019; Locatello et al., 2019……… The list goes on and on and on)

We adopted a “downstream” task relevant to patents (open to ideas).

Call out to the community: having a set of agreed upon downstream task (Common Task Framework) would be highly useful

Need Common Dataset, Task (prediction/regression), and Metrics

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013), “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, 35 (8), 1798–1828.

Common Task Framework

Common Task Framework (1980's)

Under CTF we have the following ingredients

(a) A **publicly available training dataset** involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.

(b) A set of **enrolled competitors** whose **common task** is to **infer** a class **prediction rule from the training data**.

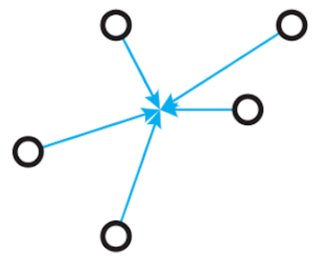
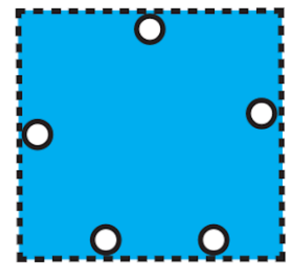
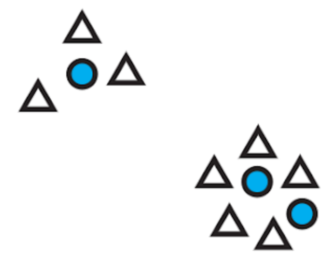
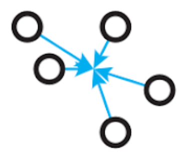
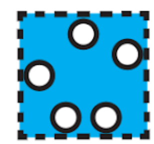
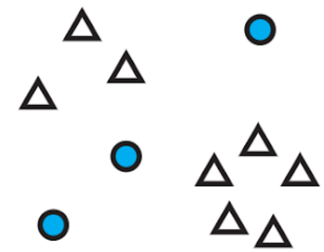
(c) A **scoring referee**, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score achieved by the submitted rule.

See Mark Liberman's description (Liberman, 2009).

Donoho, David. "50 years of data science." Journal of Computational and Graphical Statistics 26.4 (2017): 745-766.

Mark Liberman "Reproducible Research and the Common Task Method" 2015

Downstream Application: Innovation Space for Predicting Tobin's Q

	Total Squared Distance	**Bounding Box Volume**	**Average Saturation**
Higher			
Lower			
Geometric Meaning	Diversity of a firm's technology stock	Potential of technology synergy within an organization	Propensity for a firm to situate its invention in a saturated space
Mathematical Definition	$W(C_k) = \sum_{z_i \in C_k} \lVert z_i - \mu_k \rVert_2^2$	$V(C_k) = \prod_{d=1}^{n} \left(\overline{z}_{(d)} - \underline{z}_{(d)} \right)$	$D(C_k) = \sum_{z_i \in C_k} \rho(z_i) / \lvert C_k \rvert$
Related Literature	Technological Diversification (Miller, 2006; Leten et al., 2007)	Technology Synergy (Song and Parry, 1996; Park et al., 2013)	Patent Thickets (Von Graevenitz et al., 2011; Egan and Teece, 2015)

Innovation Space Engineered Variables Provide Signals for Firm Valuation

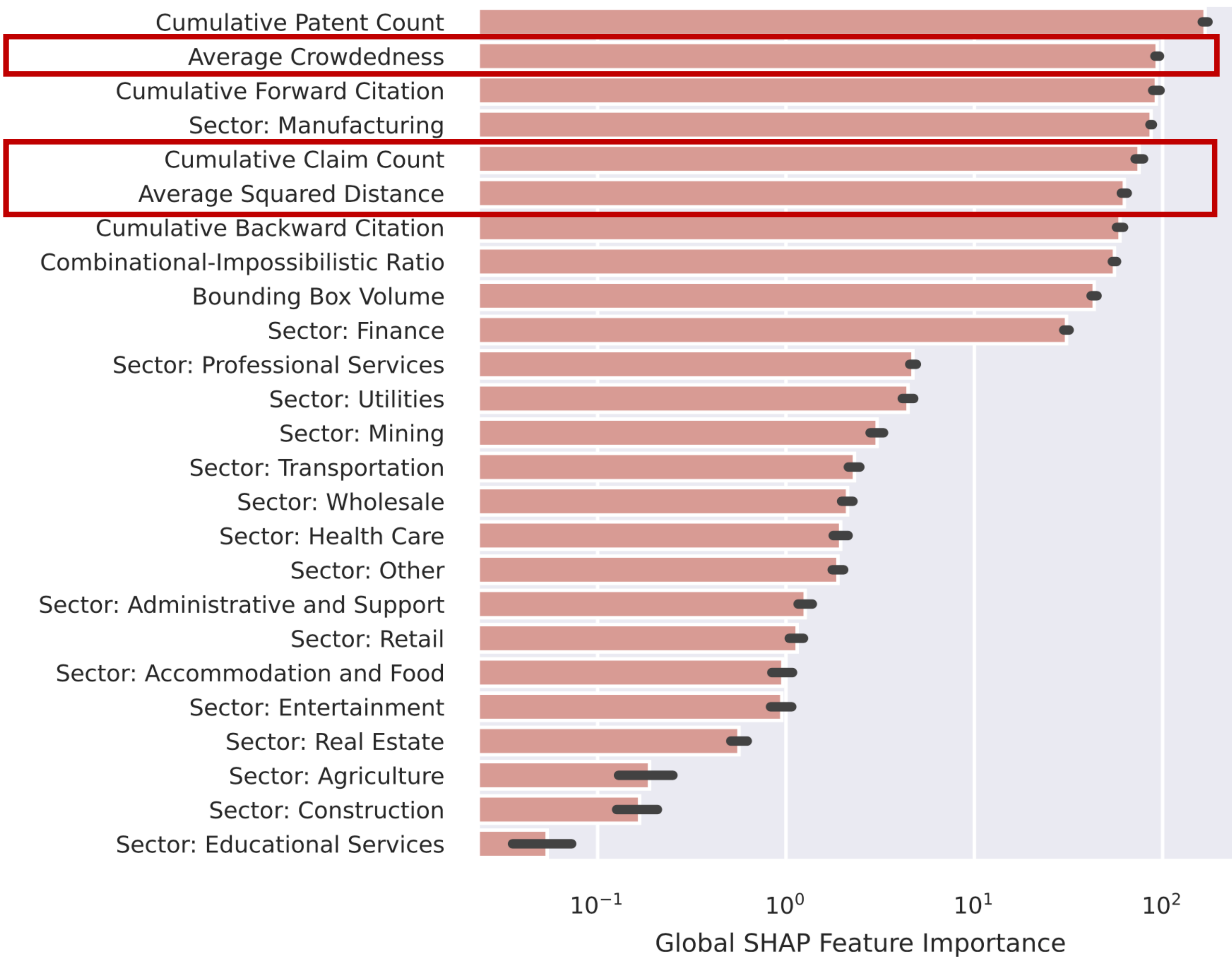
Tobin's Q = XGBoost(X)
SHAP (Fitted XGBoost)

- **Tobin's Q =**
Total Market Value of Firm / Total Asset Value of Firm

- **Total Squared Distance**
Total distance between firm's patents and patent centroid

- **Average Crowdedness**
Density of nearby patents

- **Combinational-Impossibilistic Ratio**



Descriptive analyses of Innovation Space on AI dataset

- USPTO AI Patent Dataset
  - Provides label on whether the invention is related to a specific AI technology (e.g., NLP, ML, AI hardware, 8 in total)
  - Manually curated by experts, generalized by prediction models

- Task
  - Using 120,000 AI-related patents
  - Predict AI-labels using **ONLY128-dimensional InnoVAE factors** (newly trained using claims)

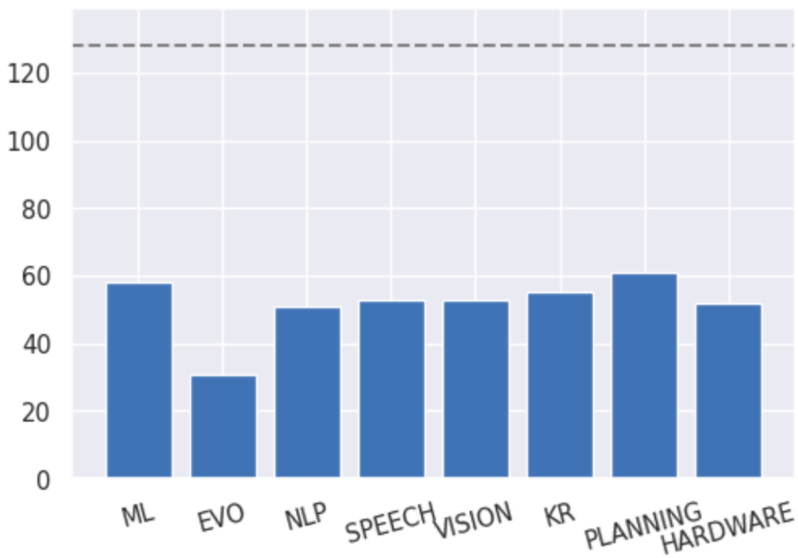
- Model
  - Break the task into 8 independent prediction tasks
  - Logistic Regression with L1 norm for feature selection

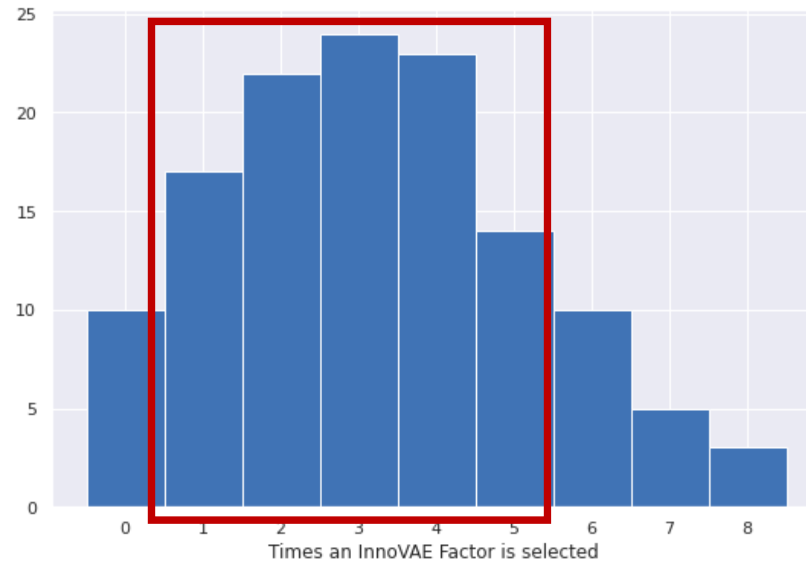
Descriptive analyses of Innovation Space on AI dataset

- Only a subset of InnoVAE factors is needed

Only 40% Innovation Factors needed for < 1% performance loss:
AUC ~0.73. NOTE: We do NOT use any text or any attribute. We just use coordinates of our new space.

- Most InnoVAE factors tend to be signal-worthy for few AI subfields

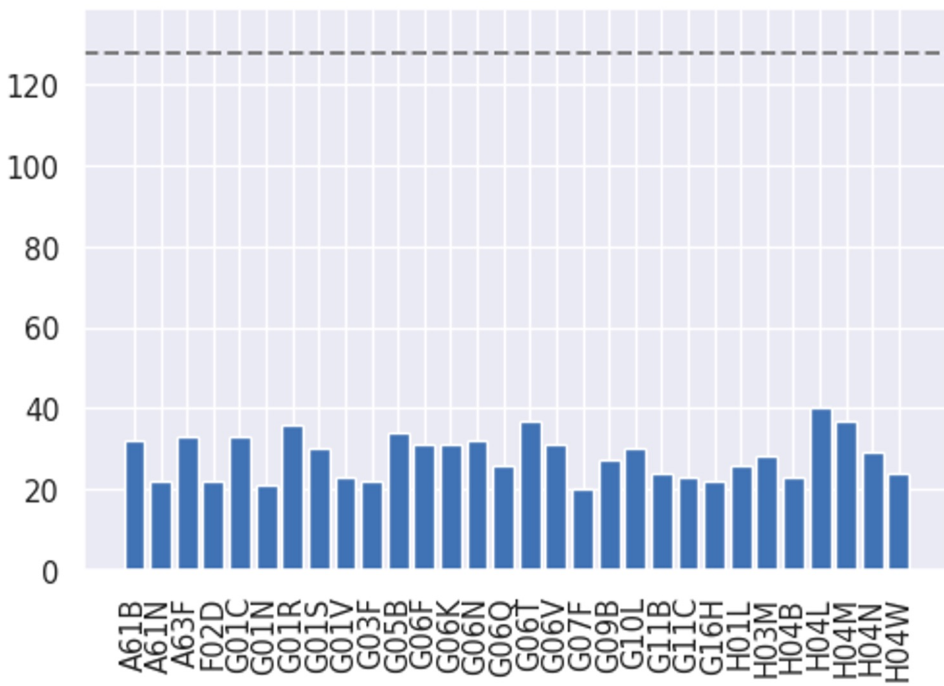


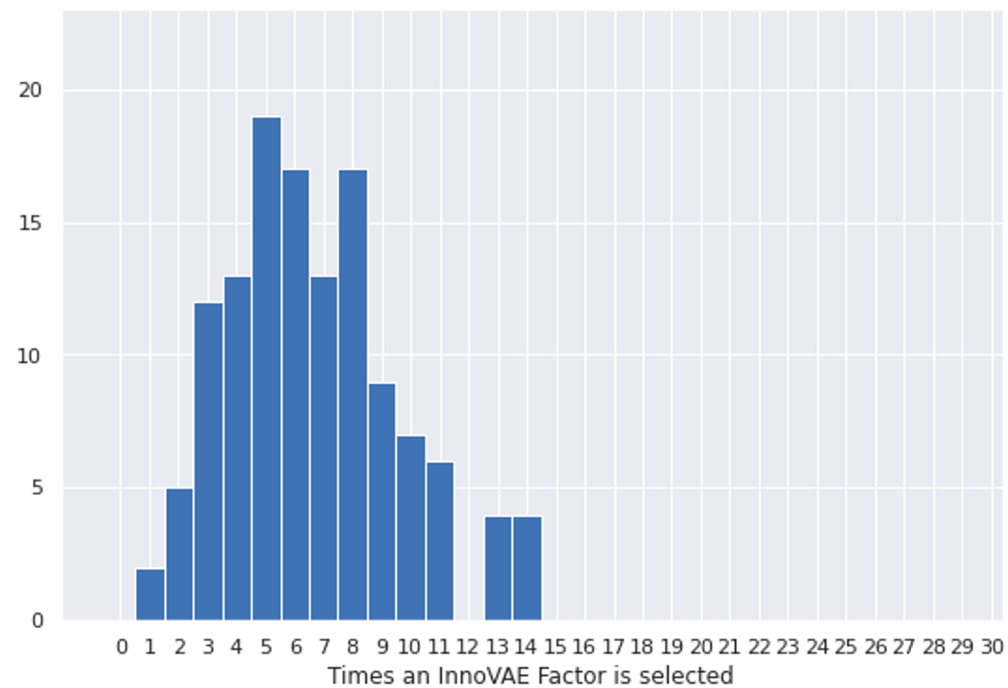


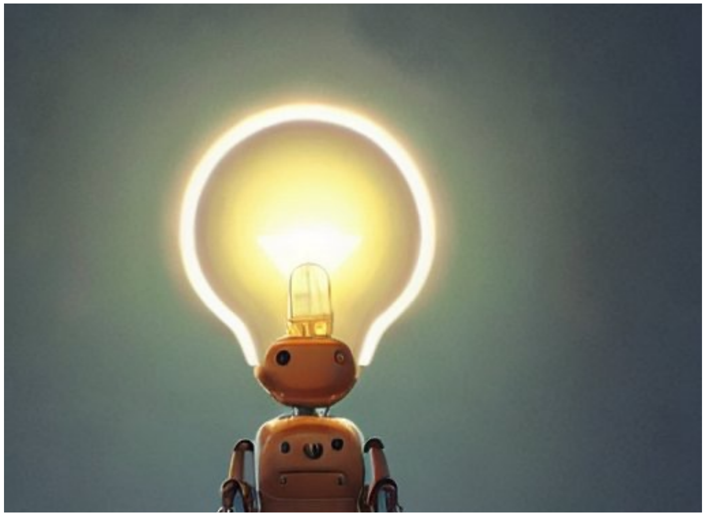
100 out of 128 InnoVAE factors are in the box

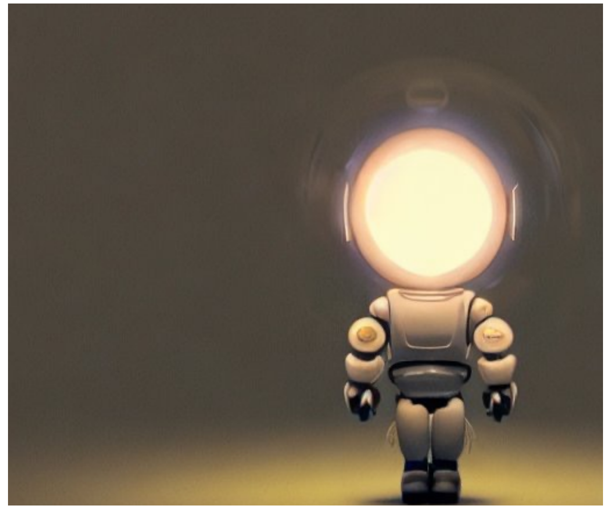
Descriptive analyses of Innovation Space on AI dataset

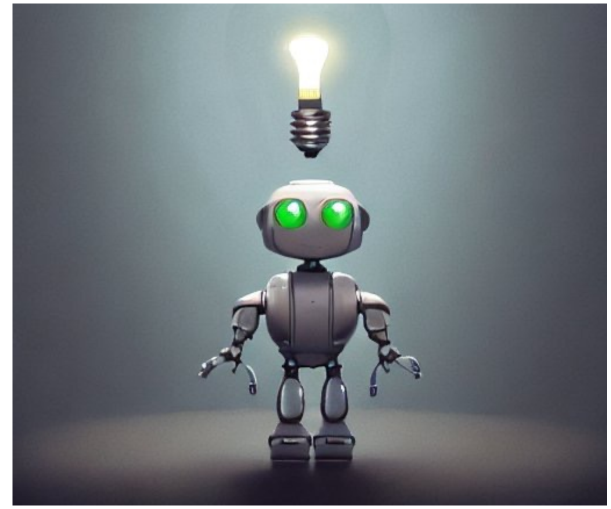
- Validate on patents sampled from 30 AI-related **CPC subclasses**











Thank you!
Feedbacks and Comments are Appreciated

Manuscript at
Tiny.cc/innovae