The Statistical Validation of Innovation Lens

Jonah Lynch

19 March 2025

Introduction

Researchers face unprecedented challenges. Information overload and the rapid pace of scientific advancement make it increasingly difficult to stay current in one's field. While specialization has become necessary, it often comes at the cost of interdisciplinary insights, which are crucial for groundbreaking discoveries. Innovation Lens (available in a public application at innovationlens.org) is a novel tool designed to address these challenges by optimizing access to information, fostering interdisciplinary exploration, and guiding researchers toward promising new topics.

The volume of scientific literature has grown exponentially, making it impossible for researchers to keep up with all developments in their fields. Specialization has emerged as a practical solution, but it risks creating silos that hinder innovation. Interdisciplinary research, though widely acknowledged as valuable, is often stymied by practical constraints such as time, funding, and the cognitive load of mastering multiple domains.

Innovation Lens is a tool that addresses these issues on multiple fronts. For individual researchers, it provides a curated, up-to-date view of their field, enabling them to identify emerging trends and prioritize their efforts. For interdisciplinary scholars, it highlights promising intersections between fields, such as bio-physics. For funding agencies, it offers a macro-level perspective on scientific research, facilitating informed decision-making about resource allocation.

Current tools like Elicit, Nomic Atlas, and Semantic Scholar have made significant strides in managing information overload. However, they are retrospective, requiring researchers to sift through vast amounts of data to generate new hypotheses. This process is not only time-consuming but also prone to gaps, as no literature review can be truly exhaustive. As a result, researchers often do not take all the relevant information into account, or they reinvent the wheel, duplicating efforts that could have been avoided with better knowledge tools.

Innovation Lens sets itself apart by offering two groundbreaking features. First, its proprietary algorithm identifies topics likely to generate high citation counts in the future, providing researchers with a data-driven roadmap for their work. Second, it reverse-engineers these topics into natural language, generating titles and abstracts for potential research articles. This dual functionality empowers researchers to focus their efforts on areas with the highest potential for impact.

How Innovation Lens Works

At its core, Innovation Lens relies on an algorithm that analyzes the distribution of scientific research over time. This analysis is translated into a high-dimensional vector space, where the algorithm predicts novel vectors likely to be valuable in the future. These vectors are then reverse-engineered into natural language, making them accessible to human researchers. While the specifics of the algorithm remain proprietary, we are committed to transparency regarding validation and ongoing efforts to increase the value of our tool. By focusing on predicting foundational moments in science, Innovation Lens hopes to guide researchers toward promising fields, regardless of their career stage or institutional affiliation.

The foundational data we use in the current iteration is the title and abstract of scientific articles on the arXiv and PubMed repositories. These text fields are embedded using a specialized language model which is optimized for representing scientific papers. 'Embeddings' are a mathematical representation of the content of the text which permit mathematical manipulation of that content in a high-dimensional 'latent space'. Our algorithm performs operations in latent space and outputs new vectors in the same space which point to the location of potential new topics. As we detail below, these operations have been shown to produce quality predictions.

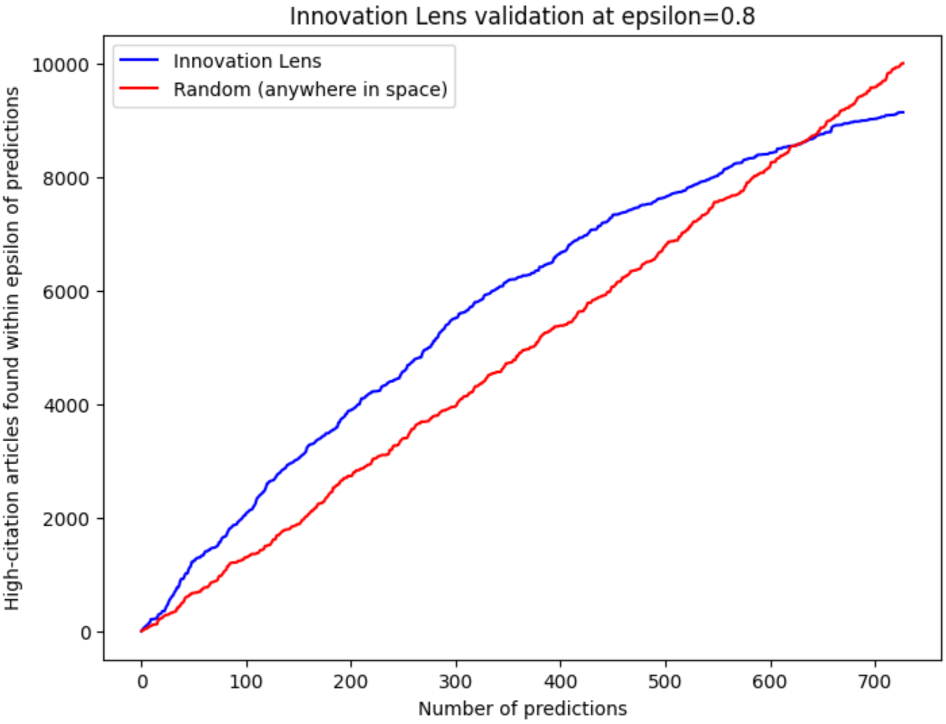
Once the text has been embedded and the algorithm has found promising vectors, a second language model is used to reverse-engineer a text representation of the predicted vectors. This model has been partially optimized for this process, and will be further improved in our ongoing research. A recursive approach is employed whereby the generated text is re-embedded and re-reverse-engineered multiple times in order to improve the precision of the final result. The process of text production in our case is different from most generative AI applications: text is not initially generated with a prompt. Rather, starting from a location in the latent space, the generative model interpolates a text that, if embedded, would be the given input vector. This output text thus includes information from the latent space which presumably has not been juxtaposed in quite this way before, or at least not within the model's training data.

Finally, two prompts to a commercially available LLM are made. The first is verbatim: 'The following is a draft of a new article proposal in a scientific field. Please rewrite it so it is expressed in complete sentences, and propose a title.' This prompt packages our previously generated text in a prettier format, without modifying its content. Finally, a last prompt is used to beat the LLM into submission and make it act as a simple classifier: 'You are a helpful assistant. Your specialty is to score proposed article titles and descriptions on a scale of 1-10, where 10 is best and 1 is worst. Answer with only a number, no words. If the text you receive is completely unsuitable as an article, or does not contain both title and abstract, return the number 0. Return ONLY an integer between 1-10 as specified. Any other response format will be treated as 0. Do NOT include explanations, markdown, or text formatting.' This prompt orders the results and returns them to the user in descending order of inferred usefulness.

Validation of the Algorithm

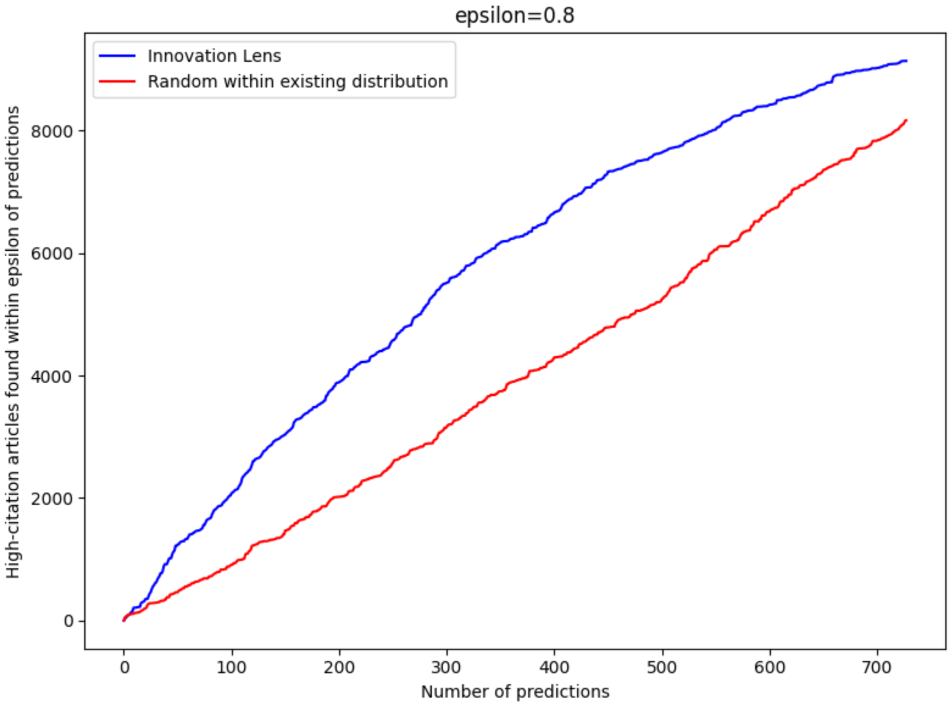
To validate our algorithm, we analyzed the computer science section of arXiv and compared our predictions against hundreds of thousands of existing articles. We considered articles that received more than 1000 citations as 'high value articles', and used them as the targets for our algorithm. We also established two baselines for comparison: first, random coverage of the entire domain space, and secondly, random coverage of the space defined by existing topics. Our results demonstrate that following our algorithm significantly outperforms both baselines, as illustrated by the lift curves below.

Compared to random coverage of the entire 'space' represented by articles in computer science, our algorithm outperforms baseline over a broad range:



At extreme quantities of predictions, randomness outperforms the algorithm. This can be interpreted to mean that given unlimited resources, science should be conducted scattershot in every possible direction at once, in order to maximize the production of high quality results. But in all more reasonable regimes subject to limited resources, our algorithm is a better predictive tool.

What about alternative baselines? One might expect random coverage of the space to perform worse than the approach usually chosen by early-career scientists, who tend to choose topics within the existing tradition of their field. However, our results show that such a conservative approach is even *less* likely to produce high-citation articles, and the Innovation Lens algorithm outperforms this baseline at all scales:



Further testing shows that these results are robust to perturbation and are not dependent on random initialization seeds. Interestingly, we have also discovered qualitative differences in the predictive value of points in different ranges of the algorithm: predictions in the middle of the range produced by the algorithm are more valuable than those at the beginning. Our research is ongoing to characterize and fine-tune these features of our algorithm, as well as apply and validate it in fields besides computer science.

A Spatial Metaphor for Discovery

A spatial metaphor can help illustrate the algorithm's function. Just as American explorers could not immediately discover gold in California, but first had to traverse the Great Plains, climb the Rocky Mountains, and cross broad deserts, big advances in scientific discovery require intermediate steps. Random exploration may once in a while yield valuable results, but it is inefficient. And prudently staying within established boundaries is unlikely to produce novel breakthroughs.

Innovation Lens identifies an intermediate distance between existing research and potential future discoveries, which guides researchers to confidently step a reasonable distance outside of their tradition, toward their goal.

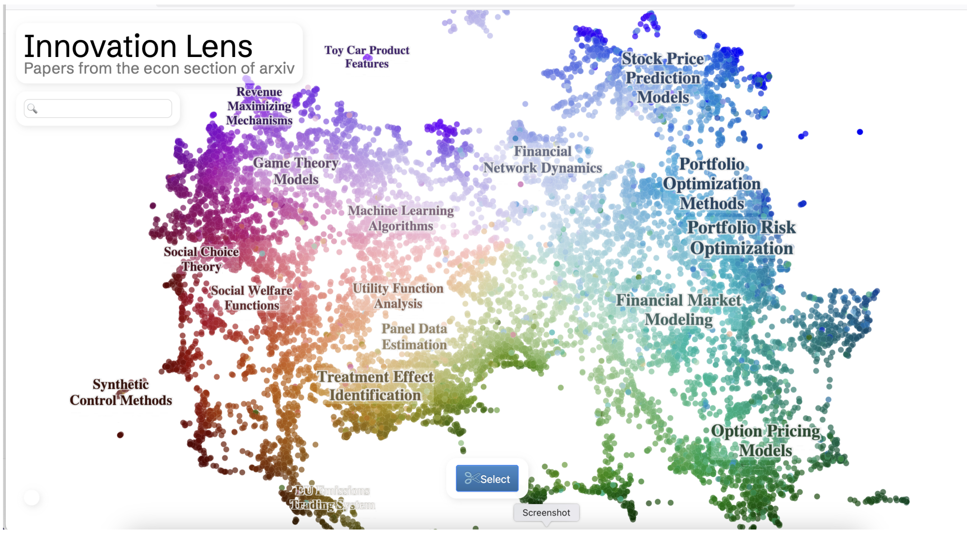
Reverse Engineering Novel Ideas

Finding a spatial description of where highly valuable discoveries might be made is an important step, but it is not sufficient to guide research. The second key component of Innovation Lens is its ability to reverse-engineer the algorithmically predicted vectors into human-readable terms. This is achieved using specialized large language models, which encode vast amounts of human knowledge. While these models excel at retrieving existing information, they can also hint at novel ideas by juxtaposing existing concepts. By generating hypothetical abstracts or problem statements, Innovation Lens helps researchers focus on areas ripe for discovery.

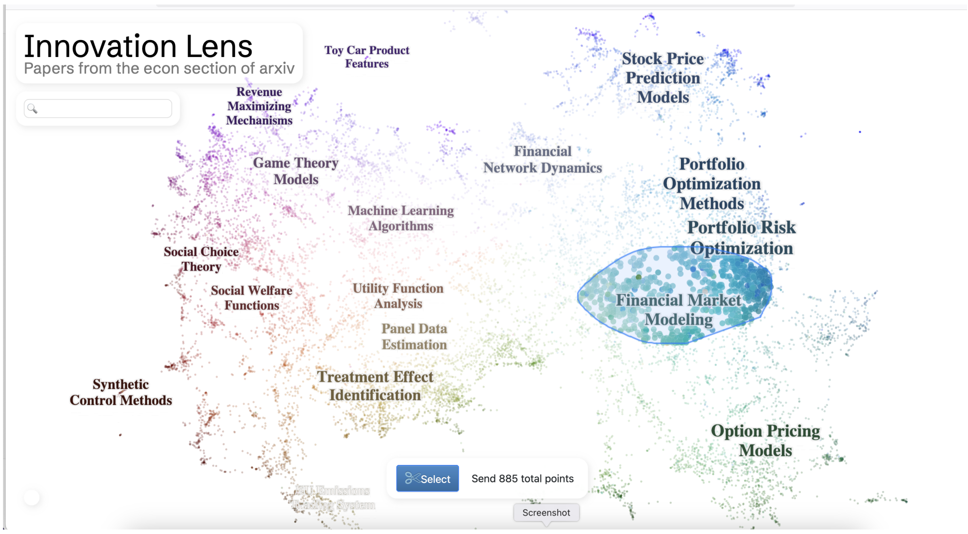
How to Use innovationlens.org

Our algorithm has been deployed on a consumer-facing application at innovationlens.org. We want our research to have a direct positive impact on the work of researchers and funding agencies. For some use cases, the technical limitations in deploying a responsive web application would be too limiting, but for individual researchers the online version is powerful enough to be useful.

The web application is designed to simplify a complex process as much as possible. First, the user is invited to define a repository of reference (either ArXiv or PubMed), and then a main topic (like Biology or Physics), and finally a list of keywords which can be freely chosen. These steps bring the total quantity of data down to a level that can be managed over a web connection and on a browser. Once the user has made these decisions, a data map made by UMAP embedding the chosen articles is loaded:

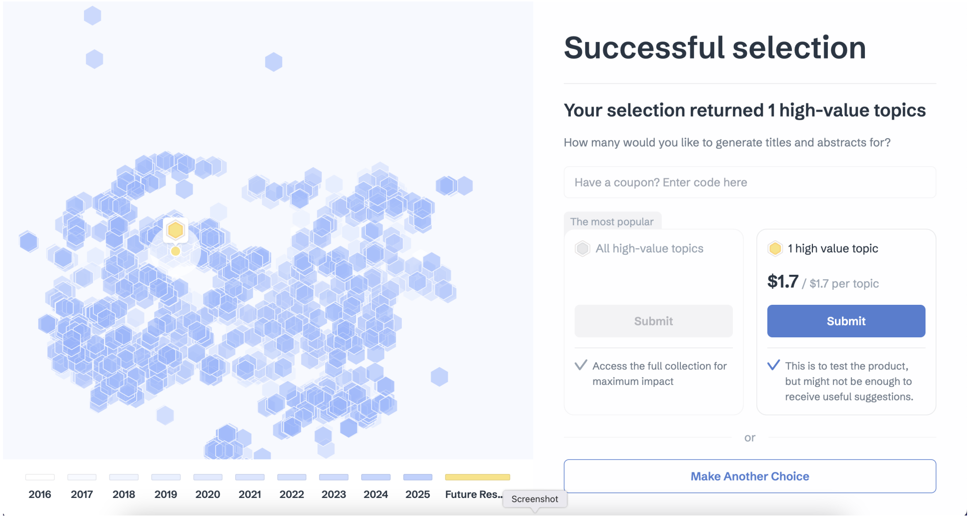

Innovation Lens
Papers from the econ section of arxiv
Toy Car Product Features
Revenue Maximizing Mechanisms
Stock Price Prediction Models
Game Theory Models
Financial Network Dynamics
Portfolio Optimization Methods
Machine Learning Algorithms
Portfolio Risk Optimization
Social Choice Theory
Social Welfare Functions
Utility Function Analysis
Financial Market Modeling
Panel Data Estimation
Synthetic Control Methods
Treatment Effect Identification
Option Pricing Models
Select
EU Emissions Trading System
Screenshot

This map permits panning and zooming to see details of the scientific articles represented. On hover, article titles are displayed, and by zooming the user can see more and of the brief summaries come into view as an orientation in space. Once the user finds a topical area of interest, they circle it with the lasso tool:


Innovation Lens
Papers from the econ section of arxiv
Toy Car Product Features
Revenue Maximizing Mechanisms
Stock Price Prediction Models
Game Theory Models
Financial Network Dynamics
Portfolio Optimization Methods
Machine Learning Algorithms
Portfolio Risk Optimization
Social Choice Theory
Social Welfare Functions
Utility Function Analysis
Financial Market Modeling
Panel Data Estimation
Synthetic Control Methods
Treatment Effect Identification
Option Pricing Models
Select
Send 885 total points
EU Emissions Trading System
Screenshot

By then clicking on ‘Send’, this region of the map is submitted to the algorithm. Calculation usually takes a half a minute or so, depending on the number of articles selected by the user. Graphically, this procedure occurs only with a relatively small selection of recent articles, but in the backend a much deeper dataset is used in the algorithm.

Once calculation is complete, the user is presented with a summary screen like this:



Successful selection

Your selection returned 1 high-value topics

How many would you like to generate titles and abstracts for?

Have a coupon? Enter code here

If any high-value topics have been found in the chosen region, their number and context is displayed to the user, along with the option to purchase the reverse-engineered text summary of what each high-value topic might be. This process requires intensive computation on GPUs, for a length of time depending on the number of high-value topics to be calculated. Once they are all complete, they are sent by email to the user.

Future Directions

Looking ahead, we plan to expand the scope of Innovation Lens to include additional fields beyond those contained in the most easily accessible public archives, ArXiv and PubMed. We are also exploring ways to incorporate more nuanced metrics of research impact, such as researcher authoritativeness, citation graph, and societal relevance and real-world applications. By continuously refining our algorithm and expanding its capabilities, we aim to make Innovation Lens an indispensable tool for researchers worldwide.

Innovation Lens represents a significant advancement in research tools, addressing information overload, interdisciplinary exploration, and the generation of novel hypotheses. By leveraging advanced algorithms and large language models, it empowers researchers to navigate the complexities of modern science and make meaningful contributions to their fields. As the scientific landscape continues to evolve, tools like Innovation Lens will play an increasingly vital role in fostering innovation and discovery.