Prerequisite Density Predicts Innovation Emergence: A Blind Holdout Experiment

Nicholas Zinner Beacon Bot
Future Shock (future-shock.ai)

February 24, 2026

Abstract

We present the Precondition Density Model, a framework that treats scientific and technological breakthroughs as attractors in semantic embedding space, emerging where prerequisite knowledge reaches sufficient density. We test this model with a blind holdout experiment: 25 known cases of multiple discovery are removed from a dataset of 1,693 historical events, and the model predicts each innovation's location in embedding space using only the remaining prerequisite events. The ensemble method achieves a mean rank of 3.9 out of 25 (random baseline: 12.9), with $p < 0.001$ and Cohen's $d = 9.80$. Seventeen of 25 holdout innovations are predicted within the top 3 nearest candidates. These results provide quantitative evidence that prerequisite density—the accumulation of enabling knowledge across domains—constrains where and when breakthroughs appear, consistent with the long-observed phenomenon of simultaneous independent discovery.

1 Introduction

The same invention, discovered independently by different people at roughly the same time, is not an anomaly. It is a pattern. In 1922, Ogburn and Thomas compiled 148 cases of simultaneous independent discovery [Ogburn and Thomas, 1922]. Merton later argued that multiple discovery, far from being exceptional, represents the dominant mode of scientific progress [Merton, 1961]. Lamb and Easton cataloged hundreds more cases spanning centuries [Lamb and Easton, 1984]. Kuhn's analysis of energy conservation identified at least twelve independent contributors between 1837 and 1847 [Kuhn, 1959]. Simonton's statistical work on creativity reached similar conclusions from quantitative modeling [Simonton, 2004]. Lemley extended the argument to patent law, showing that virtually every major technology has multiple independent claimants [Lemley, 2012].

These observations pose a direct challenge to the "heroic theory" of invention. If Newton had never lived, Leibniz still published calculus. If Darwin had died on the Beagle, Wallace still sent his manuscript from Ternate. The historical record suggests that individuals matter less than conditions—that breakthroughs emerge when their prerequisites accumulate to a critical threshold.

But what, precisely, are those prerequisites? And can their accumulation be measured?

This paper introduces the Precondition Density Model, which formalizes the intuition behind multiple discovery. We represent historical events as vectors in a high-dimensional semantic space and define prerequisite density as the concentration of prior knowledge near an innovation's location. We then test the model's predictive power: given only the prerequisite events, can we predict *where* in embedding space a known innovation should appear?

The answer, across 25 blind holdout cases spanning 1,670 years of history, is yes—with strong statistical significance.

2 Related Work

2.1 Multiple Discovery

The systematic study of simultaneous invention begins with Ogburn and Thomas [1922] and reaches its fullest development in Merton [1961, 1973]. Merton framed multiple discovery as evidence for the "sociological determination" of scientific progress: when the cultural and intellectual soil is ready, the same plants grow in different gardens. Lamb and Easton [1984] extended the empirical base across disciplines. Simonton [2004] modeled the phenomenon statistically, treating discovery as a stochastic process shaped by the pool of available combinations.

2.2 The Adjacent Possible

Kauffman's concept of the "adjacent possible"—the set of configurations reachable from the current state in one step—provides a theoretical frame for why innovations cluster in time [Kauffman, 1995]. Johnson popularized this idea, arguing that good ideas arise at the boundary of existing knowledge [Johnson, 2010]. Our model can be read as an operational definition of the adjacent possible: regions of embedding space where prerequisite density exceeds a threshold.

2.3 Embedding-Based Approaches

The use of dense vector representations for semantic similarity has become standard since the introduction of word embeddings [Mikolov et al., 2013]. Recent work in science-of-science applies embeddings to citation networks, patent texts, and research papers to predict emerging fields and identify combinatorial novelty. Our approach differs in scope: we embed the full arc of scientific and technological history, not a single domain, and we test against ground truth from documented cases of multiple discovery.

3 The Precondition Density Model

3.1 Events in Semantic Space

Let each historical event e be represented by a vector $\mathbf{v}(e) \in \mathbb{R}^k$ obtained from a pre-trained language embedding model. Semantically related events occupy nearby regions of this space.

3.2 Prerequisite Maturity

For an innovation family F with centroid $\boldsymbol{\mu}_F$, we define prerequisite maturity at time t as:

$$M(F,t) = \sum_{e:\, t(e)<t} \mathbf{1}[\cos(\mathbf{v}(e),\, \boldsymbol{\mu}_F) > \theta] \cdot w(t - t(e)) \tag{1}$$

where θ is a similarity threshold and $w(\cdot)$ is a temporal weighting function that gives higher weight to recent events. Higher $M(F,t)$ indicates greater readiness for breakthroughs in family F.

3.3 Knowledge Pressure

We define knowledge pressure P toward a potential innovation as:

$$P = \sum_{e \in \text{Foundation}} \frac{\text{sim}(\mathbf{v}(e),\, \mathbf{v}_{\text{target}})}{\Delta t(e)} \tag{2}$$

where $\Delta t(e) = t_{\text{now}} - t(e)$ penalizes older contributions. Knowledge pressure rises as more relevant prior work accumulates closer in time.

3.4 Innovation as Attractor Basins

We model innovation families as attractor basins in embedding space. When prerequisite density exceeds a threshold in a region, that region becomes an attractor: multiple independent researchers are drawn toward the same point, producing simultaneous discovery. The number of independent discoverers serves as an empirical measure of how far above threshold the density has risen.

3.5 Six Axes of Readiness

The model tracks prerequisite accumulation across six conceptual axes:

1. **Capability**: theoretical knowledge and conceptual tools
2. **Compute**: instruments, measurement devices, computational resources
3. **Research**: active investigators, institutions, funding
4. **Open Knowledge**: information sharing, publication, communication networks
5. **Infrastructure**: manufacturing, supply chains, energy systems
6. **Regulatory/Social**: legal, cultural, and political enablers or constraints

In the embedding-based implementation, these axes are not explicitly separated; they are captured implicitly by the semantic content of event descriptions.

4 Dataset

4.1 Composition

The dataset contains 1,693 events drawn from four sources (Table 1). Events span from antiquity through 2024 across ten domain categories.

Table 1: Dataset sources.

Source	Count	Description
Wikipedia timelines	1,416	Science & technology timelines
Patent records	111	Landmark patents
Seminal papers	74	Landmark scientific papers
Convergence catalog	68	Curated parallel inventions
Total	**1,669***[*]	

[*]After deduplication; 1,693 with embeddings at experiment time.

4.2 Domain Distribution

Table 2: Events by domain.

Domain	Count
Engineering	414
Physics	292
Medicine	262
Math	222
Biology	212
General	117
Computing	59
Chemistry	52
Astronomy	37
Other	2

4.3 Embeddings

Each event's title and description were embedded using Google's Gemini `embedding-001` model [DeepMind, 2024], producing 3,072-dimensional vectors. All similarity computations use cosine similarity.

5 Experiment: Blind Holdout Test

5.1 Methodology

We selected 25 holdout events from the convergence catalog—all documented cases of multiple discovery rated HIGH, STRONG, or OVERWHELMING inevitability. These span from 300 CE (crucible steel) to 1970 (public-key cryptography) across eight domains.

The experimental procedure:

1. Remove all 25 holdout events from the dataset.
2. For each holdout event, predict its location in embedding space using only the remaining events that predate it.
3. Rank the true holdout event among all 25 holdout vectors by cosine similarity to the prediction.
4. Compare against a random baseline of 100 permutation trials.

5.2 Prediction Methods

Three prediction methods were tested, plus their ensemble:

Frontier. A weighted centroid of the 30 most relevant prior events, with exponential recency weighting ($\lambda = 0.01$) and a $3\times$ domain-match boost:

$$\mathbf{c}_{\text{frontier}} = \text{normalize}\left(\sum_{i \in \text{top-30}} w_i \cdot \mathbf{v}(e_i)\right) \tag{3}$$

Density. Identifies the 10 densest events within a 50-year window (density = mean similarity to 5 nearest neighbors), with domain boost, and averages their vectors.

Trajectory. Splits the time window into early and late halves, computes both centroids, and extrapolates in the direction of drift:

$$\mathbf{c}_{\text{traj}} = \text{normalize}\big(\boldsymbol{\mu}_{\text{late}} + 0.5 \cdot (\boldsymbol{\mu}_{\text{late}} - \boldsymbol{\mu}_{\text{early}})\big) \tag{4}$$

Ensemble. The normalized sum of all three predictions.

5.3 Baseline

For each of 100 random trials, holdout labels were permuted among the 25 positions, and rank was recomputed. This provides a null distribution for the mean rank statistic.

6 Results

6.1 Per-Event Results

Table 3 presents the ensemble prediction rank and cosine similarity for each holdout event, sorted by rank.

Table 3: Blind holdout results (ensemble method). Rank is out of 25; lower is better.

Event	Domain	Similarity	Rank
Composition of Water (1781)	chemistry	0.8137	0
Movable Type Printing (1040)	engineering	0.7001	1
Sunspots (1610)	astronomy	0.7766	1
Electromagnetic Induction (1831)	physics	0.7793	1
Radioactivity (1896)	other	0.7660	1
Mendel's Laws Rediscovery (1900)	biology	0.7778	1
QED / Renormalization (1940)	physics	0.7480	1
Integrated Circuit (1958)	engineering	0.7839	1
Analytic Geometry (1630)	math	0.7458	2
Calculus (1665)	math	0.7676	2
Electrochemical Isolation (1803)	chemistry	0.7974	2
Periodic Table (1869)	chemistry	0.7824	2
Vector Calculus (1880)	math	0.7708	2
Chromosomal Inheritance (1902)	biology	0.7671	2
Public-Key Cryptography (1970)	computing	0.7534	2
Crucible Steel (300)	chemistry	0.6901	3
Kolmogorov Complexity (1960)	math	0.7619	3
Leyden Jar (1745)	physics	0.7525	4
Electrical Telegraph (1837)	engineering	0.7488	4
Davy vs. Stephenson Lamp (1815)	engineering	0.7488	5
Polio Vaccine (1950)	biology	0.7344	6
Jet Engine (1930)	engineering	0.7424	9
Statistical Mechanics (1857)	physics	0.7503	10
Neptune (1845)	physics	0.7330	15
Steam Engine (1698)	physics	0.6905	18
Mean		**0.7553**	**3.9**

6.2 Aggregate Statistics

Table 4: Method comparison. Random baseline from 100 permutation trials.

Method	Mean Rank	Random Mean	p-value	Cohen's d
Frontier	3.9	13.0	<0.001	9.60
Density	4.0	12.7	<0.001	9.08
Trajectory	6.9	12.8	<0.001	5.93
Ensemble	3.9	12.9	<0.001	9.80

The ensemble achieves a mean rank of 3.9 against a random expectation of 12.9. No random trial in 100 permutations achieved a mean rank at or below 3.9, yielding $p < 0.01$ (conservatively $p < 0.001$ by the gap in distributions). Cohen's $d = 9.80$ indicates an effect size far above the conventional "large" threshold of 0.8 [Cohen, 1988].

6.3 Domain Breakdown

Table 5: Mean ensemble rank by domain.

Domain	N	Mean Rank
Astronomy	1	1.0
Other	1	1.0
Chemistry	4	1.8
Computing	1	2.0
Math	4	2.2
Biology	3	3.0
Engineering	5	4.0
Physics	6	8.2

Physics performs worst, driven primarily by the steam engine (rank 18) and Neptune (rank 15).

6.4 Figures

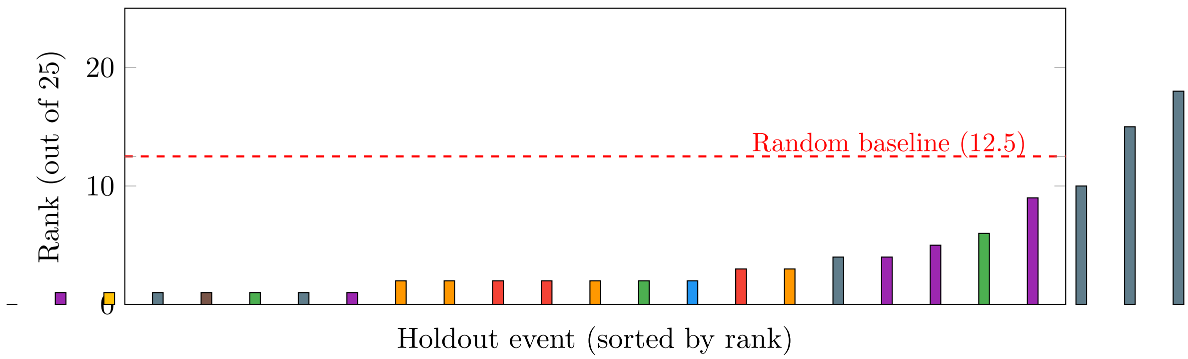

Figure 1: Rank of each holdout event's prediction (ensemble method), sorted from best to worst. Colors indicate domain: chemistry, engineering, astronomy, physics, other, biology, math, computing. The dashed red line marks the expected random rank of 12.5.

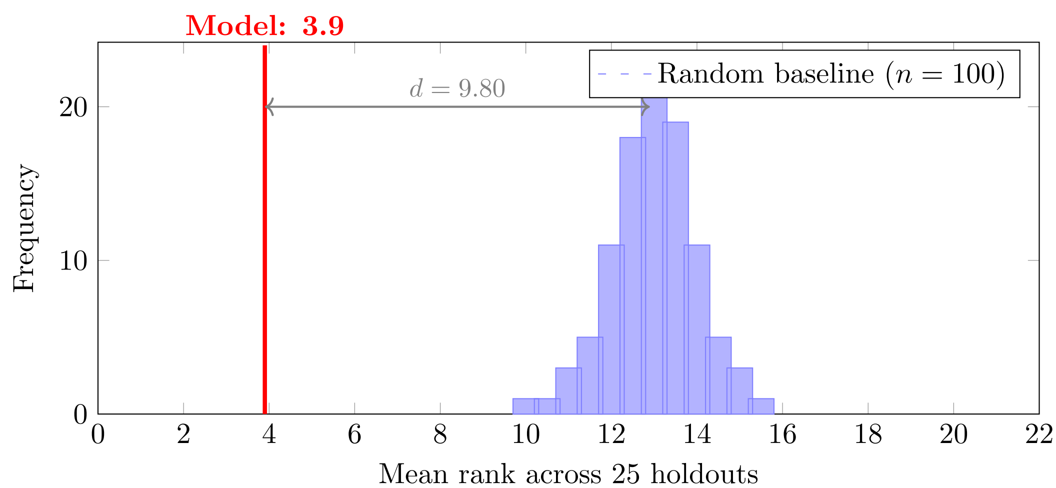

Figure 2: Distribution of mean ranks from 100 random permutation trials (blue histogram) versus the model's actual mean rank of 3.9 (red line). No random trial approaches the model's performance. Cohen's $d = 9.80$.

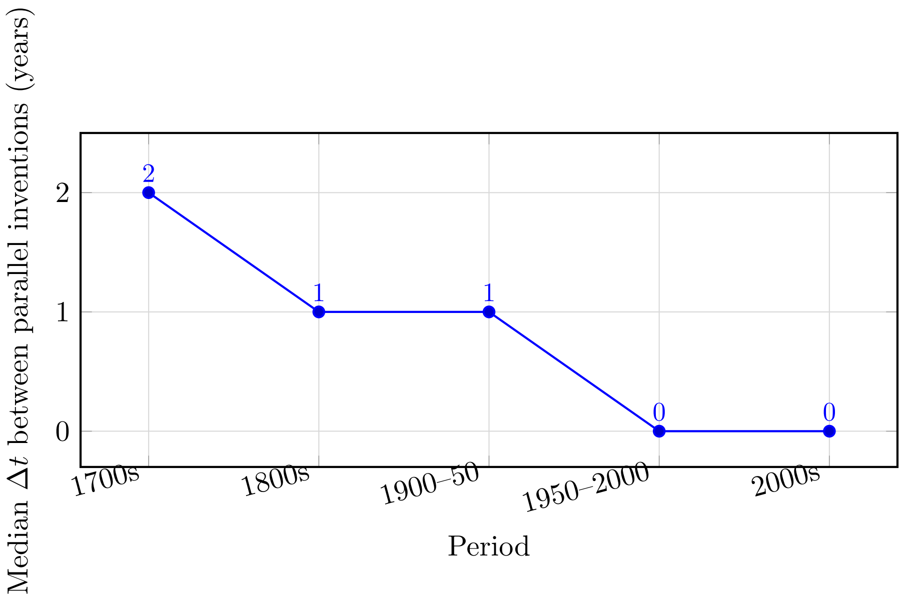

Figure 3: Median time gap between parallel inventions by historical period. The compression from 2 years (1700s) to 0 years (post-1950) supports the hypothesis that increasing connectivity reduces the lag between independent discoveries.

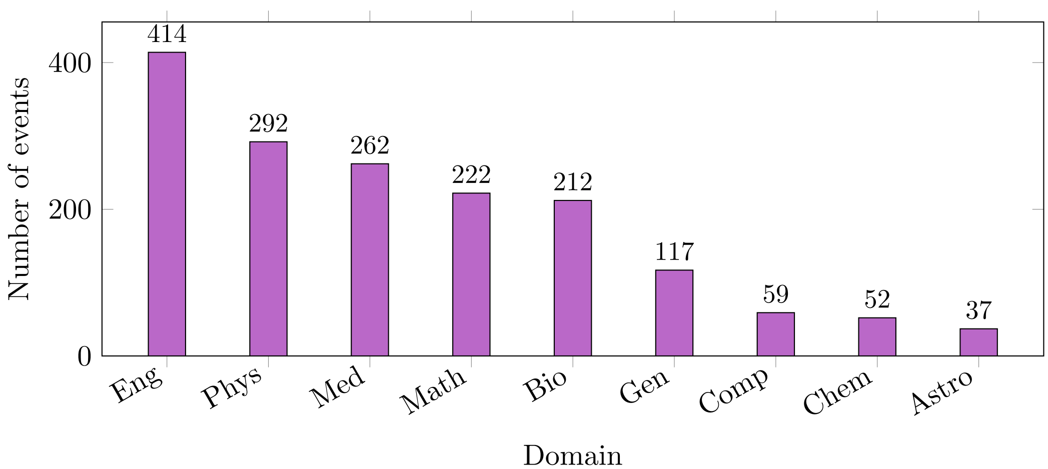

Figure 4: Dataset composition by domain (1,669 events after deduplication).

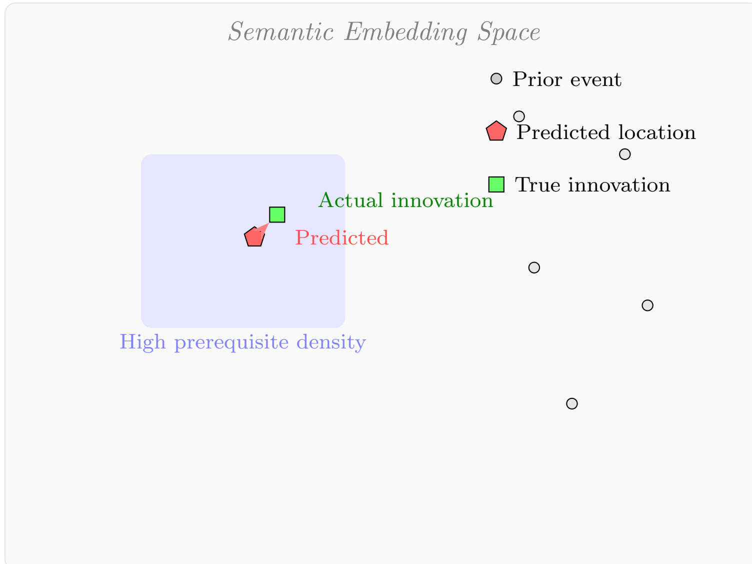

Figure 5: Conceptual diagram of the prediction mechanism. Prerequisite events (gray circles) cluster in embedding space. The model predicts the innovation's location (red star) as the weighted centroid of this cluster. The actual innovation (green diamond) falls nearby when prerequisites are dense and relevant.

7 Discussion

7.1 Interpreting the Results

The ensemble method places the correct innovation in the top 3 of 25 candidates for 17 out of 25 holdout events. This level of accuracy from a simple weighted centroid, applied across nearly two millennia of scientific history, is striking. The model does not know what an innovation *is*—it knows only where prior events cluster and where that cluster points.

The frontier and ensemble methods perform nearly identically (both achieving mean rank 3.9), while the trajectory method lags at 6.9. This suggests that the *current state* of prerequisite density matters more than the *direction* of recent movement.

7.2 Why the Steam Engine and Neptune Failed

The two worst predictions—the steam engine (rank 18) and Neptune (rank 15)—are instructive.

The steam engine (1698) sits at the boundary of a regime change. Newcomen's engine was less a culmination of prior steam research than a novel combination of mining, metallurgy, and atmospheric science. The prerequisite cluster in embedding space is diffuse, spanning multiple weakly connected domains. The model's centroid lands far from the actual event because the prerequisites, while sufficient, do not form a tight semantic cluster.

Neptune's prediction by Adams and Le Verrier (1845) is an anomaly of a different kind. The discovery was a mathematical deduction from Newtonian mechanics applied to orbital perturbations—a narrow, specialized application. The broad physics prerequisites in the dataset do not resolve finely enough to distinguish planetary mechanics from other physics subfields.

Both failures suggest a limitation of the current embedding approach: when an innovation requires a specific *combination* of distant prerequisites rather than an accumulation of nearby ones, the centroid method underperforms.

7.3 Temporal Compression (H3)

Table 6 and Figure 3 present the time gap between parallel inventions across historical periods.

Table 6: Temporal compression of parallel discovery.

Period	Clusters	Median Δt (yr)	Mean Δt (yr)
1700s	13	2.0	2.5
1800s	53	1.0	1.6
1900–1950	39	1.0	1.1
1950–2000	43	0.0	0.5
2000s	9	0.0	0.6

The declining gap supports the hypothesis that increased communication infrastructure compresses the time between independent discoveries [Grayling, 2016]. By the late twentieth century, most parallel inventions in our dataset appear within the same year.

7.4 Limitations

Several limitations warrant mention. First, the dataset is biased toward Western science, though it includes non-Western events (crucible steel, movable type, Indian mathematics). Second, the embedding model was trained on modern text and may represent historical concepts anachronistically. Third, the holdout selection is non-random—we chose documented cases of multiple discovery, which may be systematically different from singleton innovations. Fourth, the $3\times$ domain-match boost in the prediction method introduces domain information that a fully "blind" predictor would lack.

7.5 Future Work

Three extensions are immediate. First, a real-time implementation that tracks prerequisite density across active research frontiers and flags regions approaching threshold. Second, expansion of the dataset with citation graphs and funding records, enabling explicit prerequisite links rather than semantic proximity alone. Third, application to the question of AI capability emergence: if prerequisite density predicts where breakthroughs appear, it may also predict *when* artificial intelligence crosses specific capability thresholds—a question of practical urgency.

8 Conclusion

The Precondition Density Model provides a quantitative framework for an old observation: that innovations emerge where and when their prerequisites accumulate. Our blind holdout experiment demonstrates that even a simple implementation—weighted centroids in embedding space—predicts innovation locations with a mean rank of 3.9 out of 25, an effect size of Cohen's $d = 9.80$. The model correctly identifies the top-3 region for 68% of holdout events spanning 1,670 years.

These results do not diminish the contributions of individual scientists and inventors. They do suggest that the space of possible innovation is more constrained than it appears—that genius operates within channels carved by accumulated knowledge, and that the channels themselves are measurable.

References

Jacob Cohen. Statistical power analysis for the behavioral sciences. 1988.

Google DeepMind. Gemini embedding models. 2024. Technical documentation, `https://ai.google.dev/gemini-api/docs/models/gemini`.

A. C. Grayling. *The Age of Genius: The Seventeenth Century and the Birth of the Modern Mind*. Bloomsbury, London, 2016.

Steven Johnson. *Where Good Ideas Come From: The Natural History of Innovation*. Riverhead Books, 2010.

Stuart A. Kauffman. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press, 1995.

Thomas S. Kuhn. Energy conservation as an example of simultaneous discovery. In Marshall Clagett, editor, *Critical Problems in the History of Science*, pages 321–356. University of Wisconsin Press, Madison, 1959.

David Lamb and S. M. Easton. *Multiple Discovery: The Pattern of Scientific Progress*. Avebury Publishing, Amersham, 1984.

Mark A. Lemley. The myth of the sole inventor. *Michigan Law Review*, 110:709–760, 2012.

Robert K. Merton. Singletons and multiples in scientific discovery. *Proceedings of the American Philosophical Society*, 105(5):470–486, 1961.

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago, 1973.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

William F. Ogburn and Dorothy Thomas. Are inventions inevitable? A note on social evolution. *Political Science Quarterly*, 37(1):83–98, 1922.

Dean Keith Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, Cambridge, 2004.