SMARTINFER
Case Study

From Diagnosis to Impact: How Systematic Problem Decomposition Drove +23% Revenue Lift in E-Commerce Search

How a diagnostic-first methodology, explore–exploit experimentation, and demand–supply gap analysis transformed marketplace search at a major e-commerce retailer

Author	Role	Team
Anjan Goswami, Ph.D.	Director of Search Science	80+ engineers & scientists

1. The Starting Point

The e-commerce search engine at a major U.S. retailer served tens of millions of monthly search requests, generating billions of product impressions across a catalog of millions of products. The platform faced a stark concentration problem typical of large-scale e-commerce: a small fraction of items accounted for the vast majority of total search and browse revenue. The remaining catalog was effectively invisible to customers.

Conversion rates were low. The intuitive response—and the one most search teams default to—would have been to immediately begin building better ranking models. That would have been the wrong first move.

> *Key insight: Before building solutions, we needed to understand what was actually broken and why. The single most consequential decision in the entire effort was investing the first months in rigorous diagnosis rather than jumping to model building.*

2. The Diagnostic Framework

2.1 Building the Evaluation Infrastructure

We constructed a representative query sample of several thousand queries using traffic-based stratified sampling from six months of search data. The sample size was chosen to accommodate monthly feature evaluation needs while providing sufficient statistical power to identify major relevance issues.

Three independent channels identified poorly performing queries:

- **Crowdsourced relevance judgments:** Top results for each query were rated by multiple human raters on a graded relevance scale. Queries falling below an NDCG threshold were flagged as bad candidates —roughly corresponding to at least one top result rated as marginally relevant or worse.
- **Conversion analysis:** Queries in the bottom quintile for conversion rate were flagged independently of relevance ratings, capturing cases where results appeared relevant but failed to convert.
- **Direct user feedback:** Queries surfaced through user feedback channels, providing a third signal independent of both behavioral data and crowdsourced judgments.

2.2 The Critical Finding: Quantifying Root Causes

Approximately **one in four queries were confirmed as poorly performing**, representing a disproportionate share of total search traffic. But the breakthrough was not the bad query rate itself—it was the root cause decomposition.

Every confirmed bad query was classified across two independent taxonomies: business issue categories (assortment gaps, pricing mismatches, channel issues) and technical issue categories (concept detection, query normalization, ranking, attribute understanding). This dual classification was the pivotal analytical step.

The results fundamentally reframed the problem:

Root Cause Category	Approximate Share	Implication
Assortment gap (product not carried)	**~40%**	No technology fix possible
Not showing expected product	**~20%**	Retrieval or catalog integration gap
Query understanding failures	**~15%**	Concept detection, intent classification
Vocabulary mismatch	**~10%**	Spelling, synonyms, normalization
Ranking, pricing, and other issues	**~15%**	Ranking model, price range, diversity

> *Over 60% of all search failures were supply-side problems—assortment gaps and missing products—not technology problems. No ranking model improvement, however sophisticated, could fix queries for products the catalog simply did not carry.*

This decomposition became the project roadmap. Rather than building a single ML system to address “search quality,” we designed targeted interventions matched to each root cause category, routing supply-side problems to merchandising and assortment teams while focusing engineering effort on the genuine technology problems.

3. Three Intervention Streams

The diagnostic framework naturally decomposed the problem into three distinct intervention streams, each requiring different methods, different teams, and different success metrics.

Stream A: Demand–Supply Gap Analysis

The Problem

The largest root cause category—assortment gaps—could not be addressed by the search team alone. But we could provide the intelligence to make assortment decisions data-driven rather than intuition-driven. The deeper question was: across the entire query–product space, where does customer demand diverge from catalog supply, and by how much?

The Method

We formalized this as an information-theoretic problem. Customer search queries represent a probability distribution over latent topics—the **demand signal**. The product catalog, represented through item titles and descriptions, represents a separate distribution—the **supply signal**. The gap between these distributions is the marketplace inefficiency.

We constructed separate topic models using Latent Dirichlet Allocation on two corpora: the historical query log and the product catalog. We then measured the Kullback–Leibler divergence between the query topic distribution and the product topic distribution. The KL divergence decomposes per topic, producing a signed, ranked list of gaps:

- **Topics where demand >> supply** (high query mass, low product mass): unmet customer demand. Routed to assortment selection teams and buyers with the extracted topic keywords as sourcing criteria.
- **Topics where supply >> demand** (high product mass, low query mass): excess inventory with no matching customer interest. Keywords extracted and routed to the demand generation and SEM teams for targeted marketing campaigns.
- **Topics in approximate equilibrium:** supply roughly matches demand. Focus on ranking optimization within existing assortment.

The per-topic decomposition made the gap actionable: each topic came with its constituent keywords, providing the exact vocabulary for demand generation campaigns or buyer search criteria. The gap metric correlated with revenue, validating that topics with high KL divergence corresponded to genuine business opportunity.

This work was published at **The Web Conference 2019**: Goswami, A., Mohapatra, P., and Zhai, C. “Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models.” WWW ’19 Companion, pp. 348–353.

Why This Matters

Traditional assortment planning relies on historical sales data, which only captures transactions that actually happened. It is silent about demand that was never fulfilled. The topic model approach identifies gaps before they manifest as lost revenue—a forward-looking capability that sales-based methods cannot provide. In a marketplace where internal business units each operated semi-independently—selecting and submitting their own inventory—the gap analysis gave buyers a data-driven view of unmet customer demand that was previously invisible to them.

Stream B: Discoverability and Explore–Exploit Experimentation

The Problem

The product catalog exhibited an extreme power-law distribution: a small fraction of items generated the vast majority of revenue. ML ranking models trained on behavioral signals (clicks, purchases) created a **feedback loop**: items that rank well accumulate more behavioral data, which trains the model to rank them higher, which generates more data. Items never shown never generate signal. The system converges on a locally optimal assortment that may be globally suboptimal.

This is a causal inference problem: an item's low observed demand is confounded by low exposure. Observational data cannot distinguish "customers don't want this" from "customers never saw this."

The Architecture

We designed and implemented a **generic explore–exploit framework** as a core component of the search infrastructure. The architecture allocated a controlled fraction of traffic to exploration and the remainder to exploitation. A feasibility analysis confirmed that the exploration budget was sufficient to make all catalog products minimally discoverable without materially impacting revenue on the exploit side.

The explore ranker used an Item Swapper mechanism: for each query, relevant low-impression items were swapped into discoverable positions in place of high-impression items that had already accumulated sufficient behavioral signal. The exploit ranker continued optimizing based on the enriched behavioral data.

Exploration Strategies and Results

We tested multiple exploration strategies through rigorous A/B experimentation. The results revealed that the method of exploration matters enormously:

Strategy	Outcome
Random relevant items	Negative impact on both conversion and revenue per visit. Naive exploration destroyed value.
Inventory planning signal	Modest positive lift in conversion and revenue. Using supply-side signals as exploration priors improved over random.
Base/variant boosting	Strong conversion lift in applicable categories (e.g., apparel) by increasing product diversity in results.
Competitive intelligence (items popular on competing platforms)	**Significant conversion lifts (5–8%+) across multiple product categories. The highest-performing exploration strategy by a wide margin.**

> *Naive exploration destroyed value. Informed exploration using external quality signals produced significant conversion lifts. The prior matters enormously—what you choose to explore determines whether exploration creates or destroys value.*

The Competitive Intelligence Strategy

The highest-performing strategy used **review data from competing platforms as an exogenous quality signal**. The logic: if an identical item appears in a competitor's top results with strong reviews, and the same item on our platform has low exposure and no review data, the asymmetry represents an information gap, not a demand gap. The product has proven demand elsewhere; our system simply hasn't discovered it yet.

This is an application of the multi-armed bandit framework with informed priors. Rather than exploring uniformly, we allocated exploration budget to items where external evidence suggested the highest expected value of information acquisition.

Demand Forecasting

To complement the exploration framework, we built a demand forecasting model for items lacking behavioral history. The model predicted potential revenue as a function of price, predicted conversion, and predicted impressions. Conversion prediction used ensemble tree models trained on item attributes (price, category, relative price within category, intent match score, out-of-stock rate, review signals, and semantic features). Impression prediction used time-series models on query traffic. This allowed principled prioritization of which items to explore first.

Stream C: Technical Ranking and Query Understanding

For the roughly 35–40% of bad queries that were genuine technology problems, we built a comprehensive ML ranking pipeline addressing each root cause identified in the diagnostic:

- **Concept detection:** ML models to correctly identify the primary concept in multi-token queries, reducing misclassification of user intent. This was the single largest technical issue category.
- **Synonym matching:** Expanded the query–item vocabulary mapping to handle cases where customer terminology diverged from product descriptions.
- **Attribute understanding:** Built attribute extraction to distinguish which tokens in a query represent attributes versus core concepts, preventing irrelevant items from being retrieved on attribute-heavy queries.
- **Query classification:** Improved category classification to prevent boosting irrelevant items in the wrong product category.
- **Baseline ranking:** Re-tuned field weights and implemented phrasal matching to improve the base text-match ranking.
- **Result diversification:** Implemented category diversity and price-aware re-ranking to prevent single-category or narrow-price-range domination of results for broad queries. The theoretical foundation for this work drew on Zhai's risk minimization framework for diversification in information retrieval (SIGIR 2009), which models diversification as a decision-theoretic problem balancing redundancy reduction, subtopic coverage, and active exploration.
- **Base/variant resolution:** Detected and demoted falsely defined marketplace base items (actually variants), increasing product diversity in results.

Each component was validated through A/B testing before production deployment, following a quarterly cycle of measurement and prioritization against the original diagnostic sample.

4. Results and Business Impact

Metric	Result
Search-attributed revenue lift	**+23%**
Sales conversion improvement	**+17%**
Bad query rate	**Reduced significantly over successive quarters**
Exploration-driven discovery	**5–8%+ conversion lifts across multiple categories**

The results came from the combination of all three streams. The diagnostic-first approach ensured that effort was allocated in proportion to impact: the largest share of bad queries (assortment gaps) received the largest investment through the demand–supply gap system and buyer routing, rather than disproportionate investment in ranking improvements that could only address the smaller technical share.

5. Organizational Leadership

The search science organization comprised **80+ engineers and scientists**, structured into specialized groups aligned with the diagnostic framework:

- **Evaluation team:** Owned the quarterly diagnostic process—query sampling, crowdsourced evaluation, root cause classification. Provided the measurement infrastructure that all prioritization depended on.
- **Relevance science team:** Technical deep-dive on poorly performing queries, identifying specific infrastructure limitations. Built the ML ranking components.

- **Discoverability and analytics team:** Built the explore–exploit framework, demand forecasting, competitive intelligence pipeline, and the analytics dashboard. The team included multiple PhDs in online learning, operations research, optimization, graph processing, and NLP.
- **Cross-functional coordination:** Business issue categories were routed to merchandising, supply chain, and marketplace operations teams—requiring Director-level relationships across organizational boundaries to drive action on non-technical root causes.

6. Publications

The work produced peer-reviewed publications at top venues:

- *Goswami, A., Mohapatra, P., and Zhai, C. (2019). "Quantifying and Visualizing the Demand and Supply Gap from E-commerce Search Data using Topic Models."* Companion Proceedings of The Web Conference (WWW '19), pp. 348–353.
- *Goswami, A. et al. (2015). "Controlled Experiments for Decision-making in E-commerce Search."* IEEE International Conference on Big Data.
- *Goswami, A. et al. (2018). "Towards Optimization of E-Commerce Search and Discovery."* SIGIR Workshop on E-Commerce (ECOM).

7. Generalizable Principles

This effort offers several principles that transfer to any marketplace or discovery system:

Diagnose before building. The single highest-ROI investment was the diagnostic framework, not any individual model. Understanding that the majority of search failures were business problems—not technology problems—redirected engineering effort toward the right interventions. Most search teams skip this step and spend years optimizing ranking for queries that fail for non-technical reasons.

Separate business problems from technology problems. The root cause taxonomy forced a distinction that prevented the most common failure mode in applied science: building sophisticated solutions to the wrong problem. An ML model cannot fix an assortment gap. A marketing campaign cannot fix a broken synonym dictionary. Matching interventions to root causes sounds obvious but is rarely practiced systematically.

Informed exploration creates value; naive exploration destroys it. The expected value of information depends entirely on the prior. Random exploration degraded revenue. Exploration guided by competitive intelligence priors produced significant lifts. Science teams must own the exploration budget and invest it where external signals predict the highest information gain.

Search logs reveal demand independently of transaction history. The demand–supply gap framework uses query data to identify unmet demand before it shows up as lost revenue. This is a forward-looking capability that sales-based planning cannot provide. It generalizes to any two-sided market where discovery happens through search—job marketplaces, content platforms, B2B procurement.

Training data is a product of past decisions. In any ML-driven marketplace, behavioral training data reflects the system's past ranking decisions. If you only learn from what you've already surfaced, you converge on a local optimum. Breaking this feedback loop requires deliberate, structured investment in information acquisition—exploration as a science program, not an afterthought.