Innovation Dataset for Machine Learning

One Dataset, Five ML Applications

Machine Learning Course

October 7, 2025

Five Questions About Innovation Success

Innovation Questions:

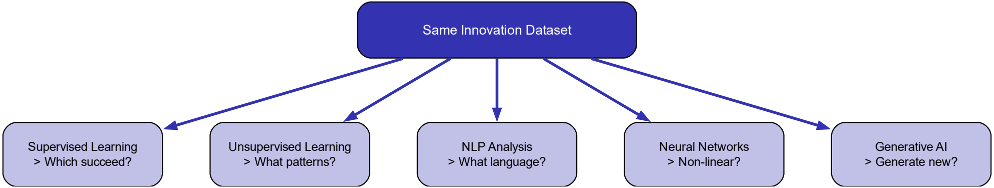
1. Which innovations will succeed?
2. What natural innovation archetypes exist?
3. What language patterns predict success?
4. Which features have non-linear relationships?
5. Can AI generate innovation pitches?

ML Solutions:

1. **Supervised Learning**
   Random Forest, Classification
2. **Unsupervised Learning**
   K-means Clustering
3. **NLP Analysis**
   BERT Embeddings, Sentiment
4. **Neural Networks**
   Deep Learning
5. **Generative AI**
   Text Generation

Course structure: Real business questions guide you to the right ML techniques

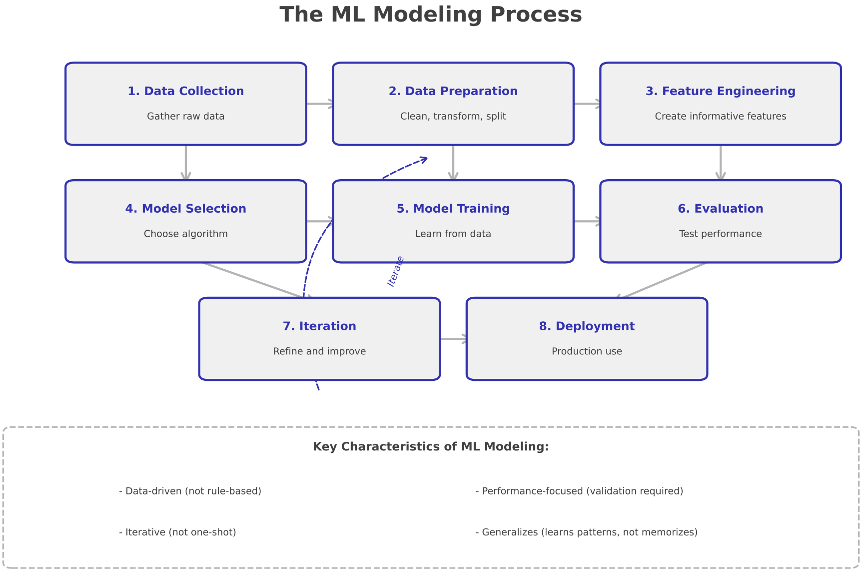
One Dataset, Five Complementary Insights



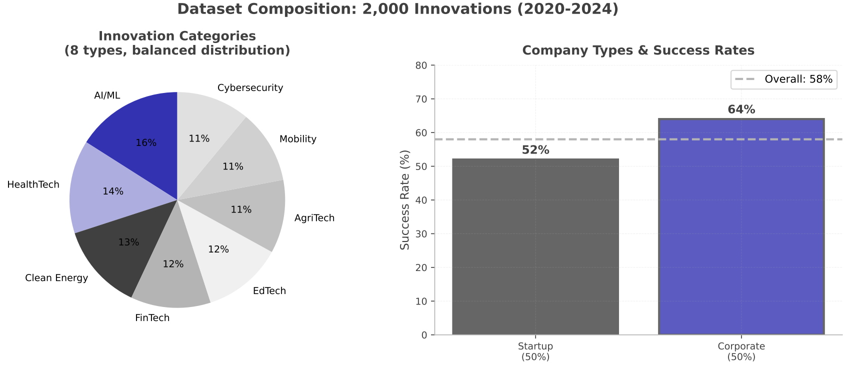
Key Insight: Each method reveals a different aspect of innovation success

Methods are complementary perspectives, not competing alternatives - use multiple to build complete understanding

What ML Modeling Really Is



Dataset Overview: 6,000 Innovations Across 8 Categories



6,000 innovations, 57.3% success rate - realistic imbalance

Clean, complete data enables learning without technical debt - no missing values, no data wrangling

Dataset Details: innovations.csv

File: `innovations.csv`
Dimensions:

- 6,000 innovations
- 20 columns
- Years: 2020-2024

Innovation Categories (8):

- AI & Machine Learning (13%)
- HealthTech (13%)
- FinTech (13%)
- AgriTech (13%)
- Cybersecurity (12%)
- Mobility (13%)
- Clean Energy (12%)
- EdTech (11%)

Key Statistics:

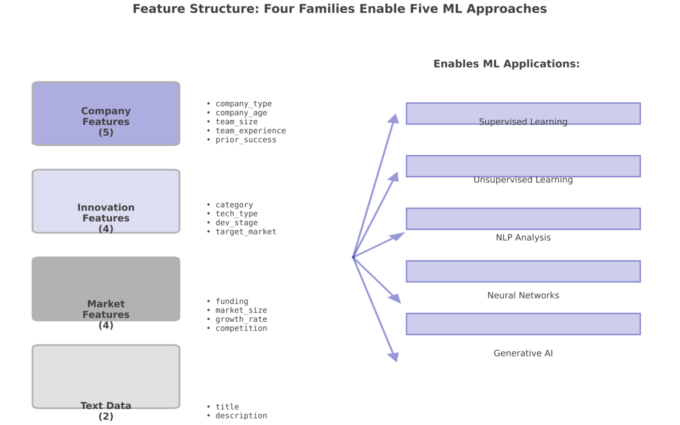
- Success rate: 57.3%
- Avg description: 319 chars
- Company types: Corporate (50%), Startup (50%)
- No missing values

Target Variables:

- `success`: Binary (0/1)
- `impact_score`: Continuous (1-10)

Why this matters:
Balanced categories, realistic imbalance, clean for learning

20 columns - complex enough to learn, simple enough to understand



Key Insight: Structured + Text features enable diverse analytical approaches

Company+Innovation+Market (Supervised/Unsupervised), Text (NLP), All Combined (Neural Nets), Generation (GenAI)

Dataset Structure: 20 Columns Explained

Company Features (5):

- `company_type`: Startup / Corporate
- `company_age_years`: 0-40
- `team_size`: 3-200
- `team_experience_avg_years`: 3-18
- `has_prior_success`: Binary

Innovation Features (4):

- `innovation_category`: 8 types
- `technology_type`: Software, Hardware, Biotech, Platform, Service
- `development_stage`: Prototype, MVP, Market-Ready, Scaling
- `target_market`: B2B, B2C, B2G

Market Features (4):

- `funding_raised_usd`: \$100K-\$100M
- `market_size_millions`: 51-5000
- `market_growth_rate`: 5-35%
- `competition_level`: Low, Medium, High

Text Data for NLP (2):

- `innovation_title`: Short title
- `innovation_description`: 50-100 words

How to use:
Mix and match features for different ML tasks

Rich feature set enabling multiple ML approaches - from supervised to generative AI

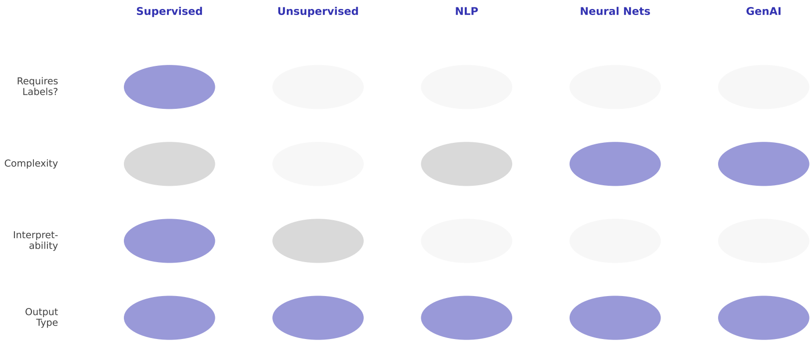
The 5 Innovation Applications: Overview

Application	Innovation Question	Method	Output
1. Supervised	Which innovations will succeed?	Random Forest, Logistic Regression	Success prediction
2. Unsupervised	What archetypes exist?	K-means clustering	4 innovation clusters
3. NLP	What language patterns predict success?	Hugging Face BERT embeddings	Semantic analysis
4. Neural Networks	Which features have non-linear relationships?	Feedforward NN vs RF	Pattern detection
5. GenAI	Can AI generate innovation pitches?	Text generation + scoring	Quality-evaluated pitches

Different questions demand different algorithms - five perspectives on innovation success

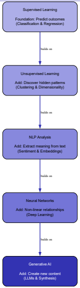
Method Comparison: Key Characteristics

Method Comparison: Key Characteristics



Each method has distinct strengths and constraints - choose based on your question and data

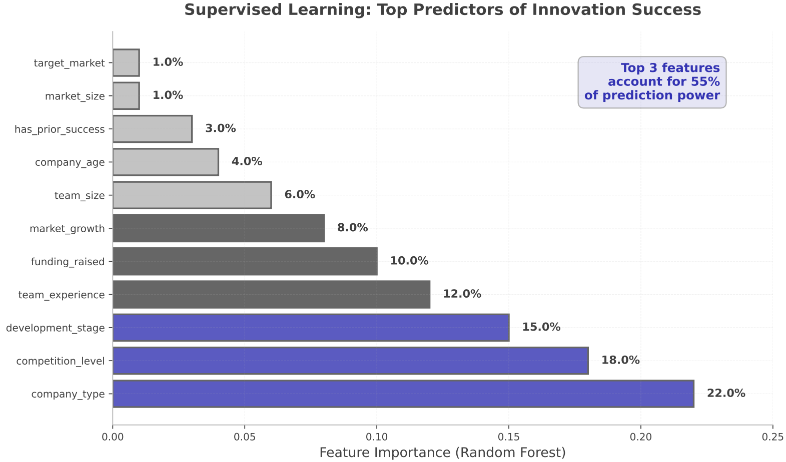
Learning Progression: Building Complexity



Each stage adds new capabilities while building on previous methods

Progressive learning: Start simple, add complexity as understanding deepens

Example: Supervised Learning Feature Importance



Potential Insight: Company characteristics may dominate prediction power

Feature importance analysis guides which variables to prioritize in modeling

Innovation Question: *Which innovations will succeed in the market?*

Theory & Methods:

- Classification (binary: success/failure)
- Feature importance analysis
- Model comparison: RF, Logistic Regression, XGBoost

Example Scenario:

`Given:` Corporate, 10 years old, team_size=50, low competition

`Potential prediction:` 60-85% success probability (model dependent)

Features Used:

- `company_type`, `company_age`
- `team_size`, `team_experience`
- `funding_raised_usd`, `competition_level`
- `development_stage`

Potential Outcomes:

- Accuracy: 60-75% (depends on data quality & features)
- Top predictors typically: company characteristics, competition
- Tree models often outperform linear models

When NOT to Use:

- Very small dataset (¡100 samples)
- Severe class imbalance (¿95:5) without handling
- Need real-time predictions (¡1ms)

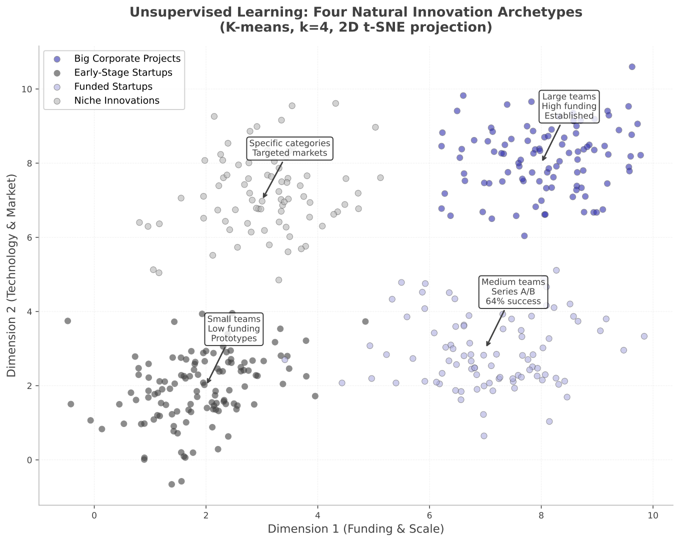
Exploration Goal:

Understand which company and market features correlate with innovation success

---

Supervised learning when you have labeled historical data - learn from past to predict future

Unsupervised Learning: Patterns Emerge Without Labels



Potential Discovery: Data may self-organize into natural archetypes

Discovery before prediction - let data reveal hidden structures

Unsupervised Learning: Discovering Innovation Archetypes

Innovation Question: *What natural innovation archetypes exist in our data?*

Theory & Method:

- Clustering (pattern discovery without labels)
- K-means (k=4), standardized features
- Dimensionality reduction (t-SNE for visualization)
- Anomaly detection capability

Example Approach:

`Features:` funding, team size, tech type, market
`Result:` 4 clusters representing different innovation profiles

Features Used:

- Technology profile
- Team characteristics
- Market features
- Funding levels

Potential Cluster Types:

1. **Big Corporate Projects**
   Large teams, high funding, established
2. **Early-Stage Startups**
   Small teams, low funding, prototypes
3. **Funded Startups**
   Medium teams, Series A/B funding
4. **Niche Innovations**
   Specific categories, targeted markets

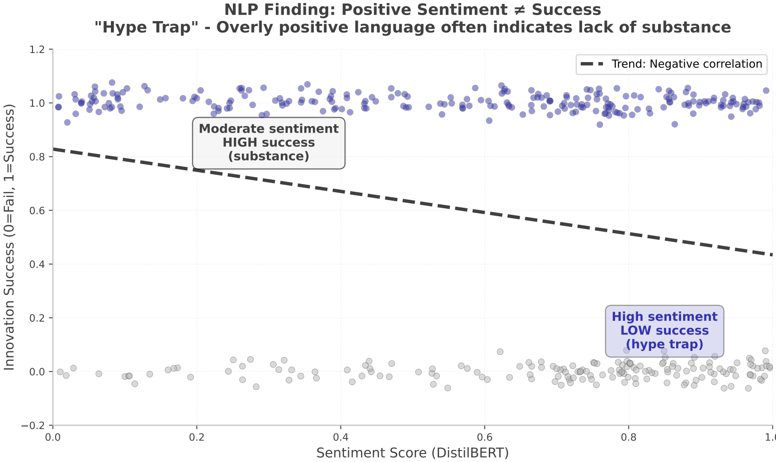
When NOT to Use:

- Known categories exist (use supervised)
- Need interpretability (k-means hard to explain)
- Require deterministic grouping

Exploration Goal: Discover natural groupings and compare success rates across clusters

Unsupervised when you don't know what patterns exist - exploratory data analysis

NLP Finding: Language Patterns and Success



Potential Pattern: Overly positive language may correlate with lower success

Language analysis can reveal unexpected correlations between communication style and outcomes

NLP Part 1: Sentiment Analysis with DistilBERT

Innovation Question: *What language patterns distinguish successful innovations?*

Approach 1: Sentiment Analysis

Theory:

- Transfer learning (pre-trained models)
- Pre-trained transformers (BERT family)
- Sentiment classification (POSITIVE/NEGATIVE)

Method:

- Hugging Face DistilBERT
- Model: `distilbert-base-uncased`
- Pipeline: sentiment-analysis
- Install: `pip install transformers torch`

Example Analysis:

`Input:` "Revolutionary AI-powered blockchain solution"
`DistilBERT:` POSITIVE (0.98 confidence)
`Exploration:` Does high positivity correlate with success?

Tasks:

- Classify sentiment of `innovation_description`
- Extract sentiment labels
- Calculate confidence scores

Potential Findings:

- Relationship between sentiment and success
- "Hype trap" hypothesis: overly positive language
- Confidence scores vs value proposition clarity

When NOT to Use:

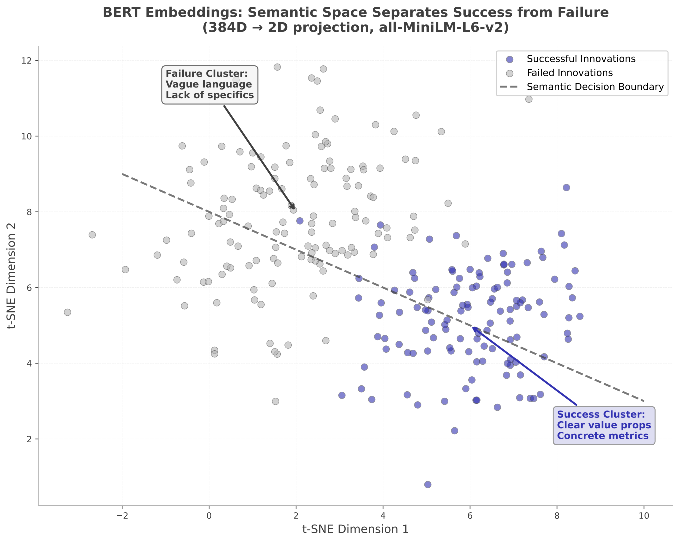
- Sarcasm or irony in text
- Domain-specific sentiment (medical, legal)
- Multilingual requirements

Exploration Goal: Test if pre-trained models capture domain-specific language patterns

---

Pre-trained models transfer knowledge, not domain expertise - fine-tuning often needed

BERT Embeddings: Semantic Space Analysis



Potential Pattern: 384 dimensions may encode success-related semantic features

Semantic understanding beyond keywords - embeddings can reveal hidden language structures

NLP Part 2: BERT Embeddings for Semantic Analysis

Approach 2: Sentence Embeddings (Main Focus)

Theory:

- Semantic representation (meaning as vectors)
- 384-dimensional vectors
- Capture meaning beyond keywords

Method:

- Sentence-transformers library
- Model: `all-MiniLM-L6-v2`
- Generate embeddings for all descriptions
- Install: `pip install sentence-transformers`

Three Use Cases:

1. **Semantic Similarity**: Find similar innovations (cosine similarity)
2. **Thematic Clustering**: Discover topic groups (k-means on embeddings)
3. **Prediction**: Use embeddings as features

Example:

`Embedding[0:5]:` [0.23, -0.15, 0.41, 0.08, -0.32]
`Cosine similarity:` 0.87 (highly similar)

Combined Approach Potential:
BERT embeddings (384 dims) + structured features (5 dims) = 389 total dimensions

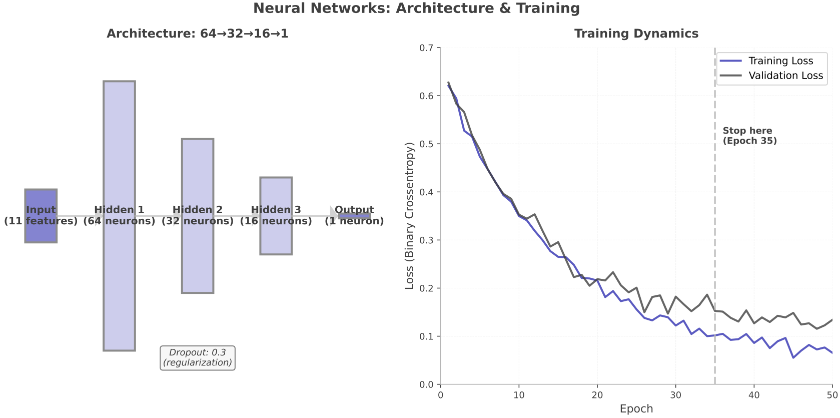
Potential Performance:

- Structured only: 60-70%
- Embeddings only: 65-75%
- **Combined: possibly 70-80%** (best case)

When NOT to Use:

- Need interpretability (384 dims hard to explain)
- Very short text (¡10 words)
- Non-English text (use multilingual models)

Exploration Goal:
Test whether combining semantic understanding (language) with business fundamentals (structured data) improves predictions



Key Learning: Validation curve behavior indicates when to stop training

---

Architecture matters, but monitoring validation loss prevents overfitting

Neural Networks: Deep Learning vs Classical ML

Innovation Question: *Which innovation features have non-linear relationships with success?*

Theory:

- Deep learning vs classical ML
- Feedforward neural networks
- Regularization techniques (dropout)
- Training dynamics (epochs, batches)

Example Architecture:

- Input: 11 features (structured)
- Hidden: 64 $\rightarrow$ 32 $\rightarrow$ 16 neurons
- Output: 1 (sigmoid activation)
- Dropout: 0.3 (regularization)

Training Setup:

- Optimizer: Adam
- Loss: Binary crossentropy
- Epochs: 50
- Batch size: 32

Example Monitoring:

```
Epoch 45: train_loss=0.08, val_loss=0.35
```

Potential Performance:

Method	Potential Accuracy
Logistic Regression	60-70%
Random Forest	65-75%
Neural Network	**65-75%**

Learning Points:

- Training curves (train vs validation)
- Overfitting detection (diverging curves)
- When NNs add value (non-linear patterns)

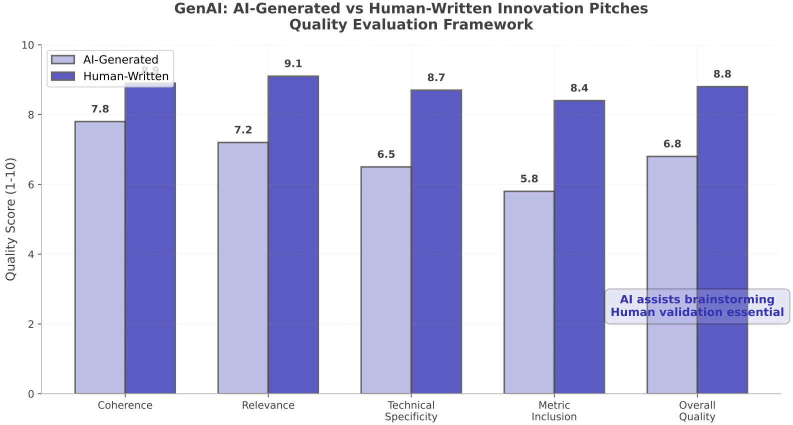
When NOT to Use:

- Small dataset (¡1,000 samples)
- Need interpretability (black box)
- Limited compute resources

Exploration Goal:

Compare neural network vs classical ML on same data to understand when complexity adds value

GenAI: AI-Generated vs Human-Written Quality



Exploration: Can AI generate plausible innovation descriptions?

GenAI assists, doesn't replace judgment - human validation essential for quality

GenAI: Generating and Evaluating Innovation Pitches

Innovation Question: *Can AI generate realistic innovation pitches?*

Theory:

- Text generation (sequence-to-sequence)
- Prompt engineering (instruction design)
- Quality evaluation (multi-metric)
- Model comparison (GPT vs Claude vs template)

Method (Current):

- Template-based generation
- Variable substitution
- Quality scoring framework (1-10 scale)

Future Extension:

- Real LLM APIs (GPT-4, Claude)
- Temperature variations (0.3-0.9)
- Model comparison experiments
- Cost-performance analysis

Example:

`Prompt:` "Generate HealthTech innovation"
`AI Output:` "AI diagnostic platform..."

Exploration Tasks:

1. Generate innovation descriptions
2. Evaluate quality (coherence, relevance, creativity)
3. Predict success of generated innovations
4. Compare AI vs human-written text

Quality Metrics:

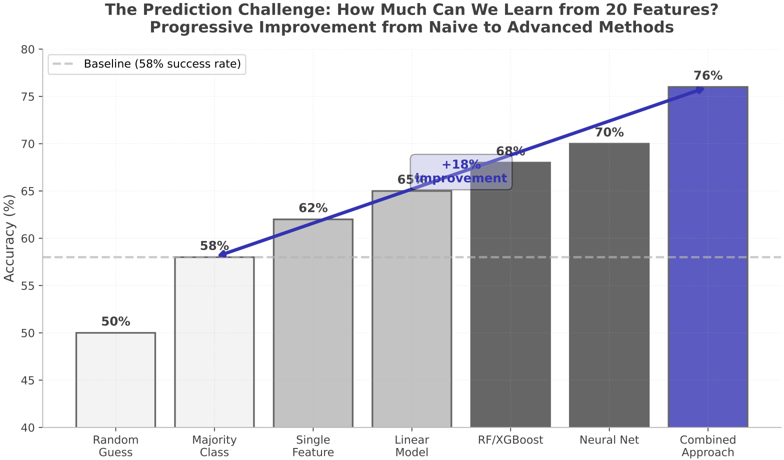
- Coherence (1-10): Logical flow
- Relevance (1-10): Domain appropriateness
- Technical specificity: Concrete details
- Metric inclusion: Quantified claims

When NOT to Use:

- Need factual accuracy (hallucination risk)
- Regulated content (legal, medical)
- Require creative originality

Exploration Goal: Test GenAI for brainstorming innovation concepts with quality evaluation

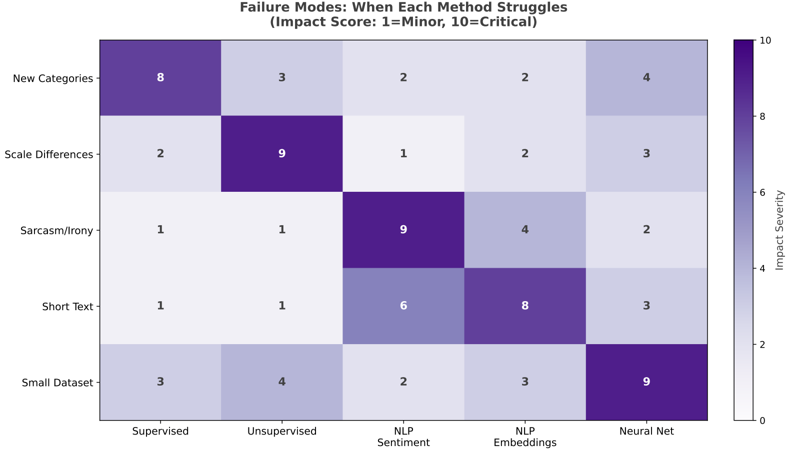
The Prediction Challenge: Progressive Learning



Exploration: How do different features and methods compare?

Feature selection and method choice both impact prediction quality

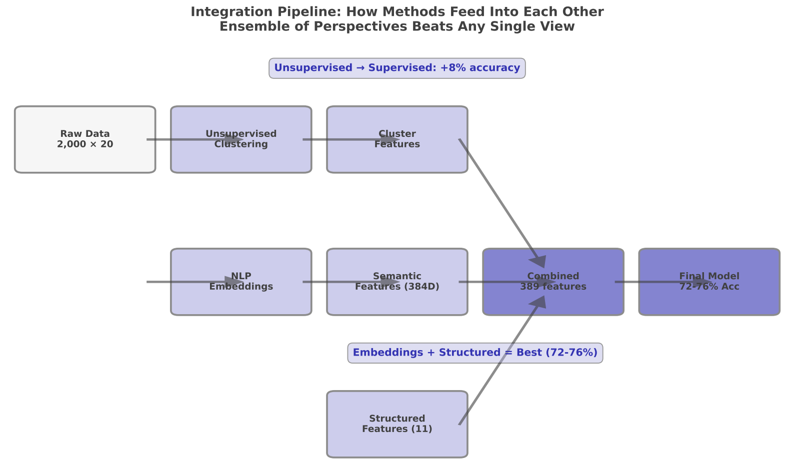
When Methods Fail: Know Your Weaknesses



Key Insight: Each method has critical failure scenarios

Know your method's weaknesses - judgment separates practitioners from script-runners

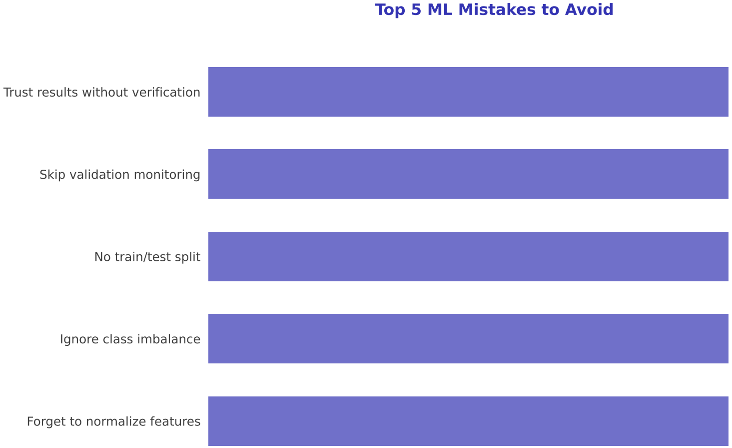
Integration Solution: Methods Feed Into Each Other



Potential Benefit: Ensemble of perspectives may beat any single view

Unsupervised clusters can become supervised features — Embeddings + Structured may improve results

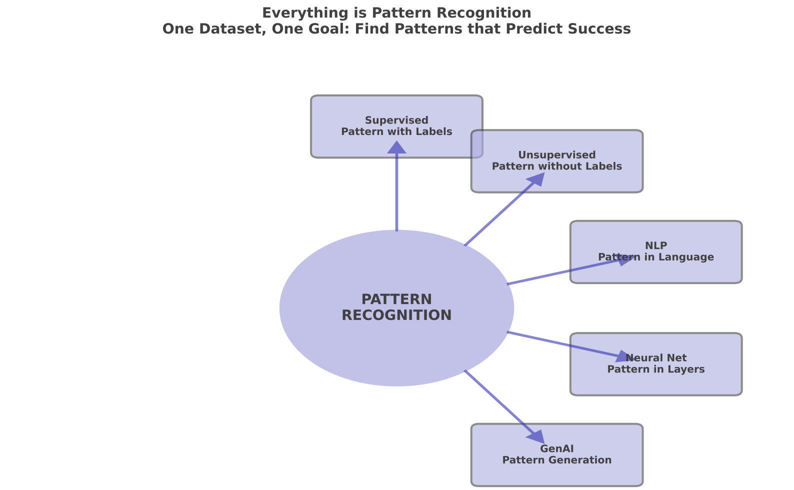
Common Pitfalls: Learn from These Mistakes



Key Insight: Top 5 mistakes - normalize, balance, split, monitor, verify

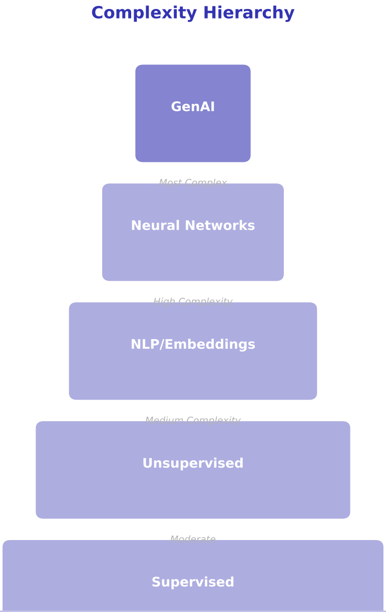
Learn from failures to accelerate success - prevention beats debugging

Everything is Pattern Recognition

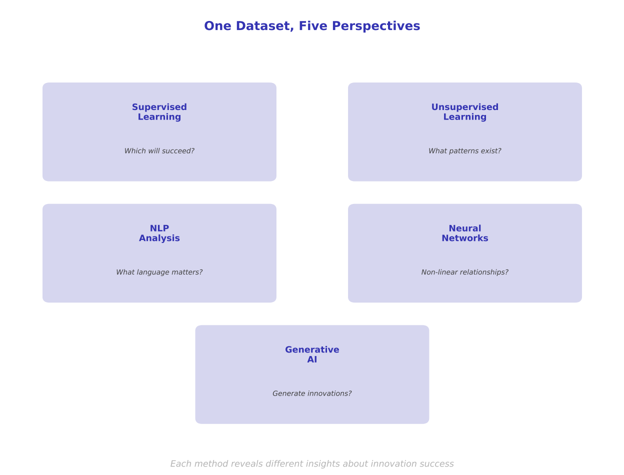


Key Insight: One dataset, one goal - find patterns that predict success

The Hierarchy of Approaches: Simple $\rightarrow$ Complex



Summary: One Dataset, Five Perspectives



Questions to Explore:

- **Supervised**: Which features predict success?
- **Unsupervised**: What natural groupings exist?
- **NLP**: What language patterns matter?