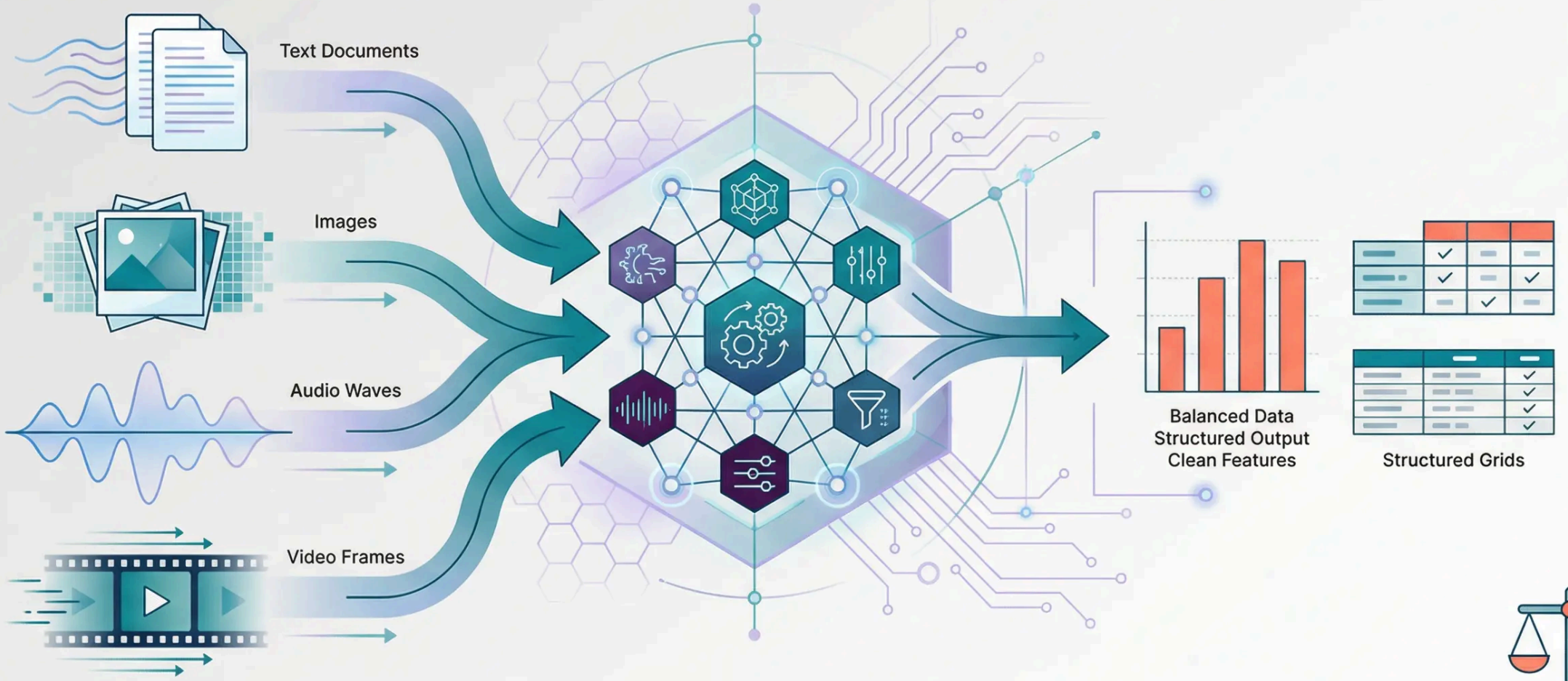
Handling Unstructured & Imbalanced Data

Advanced Data Preparation for Real-World AI Systems

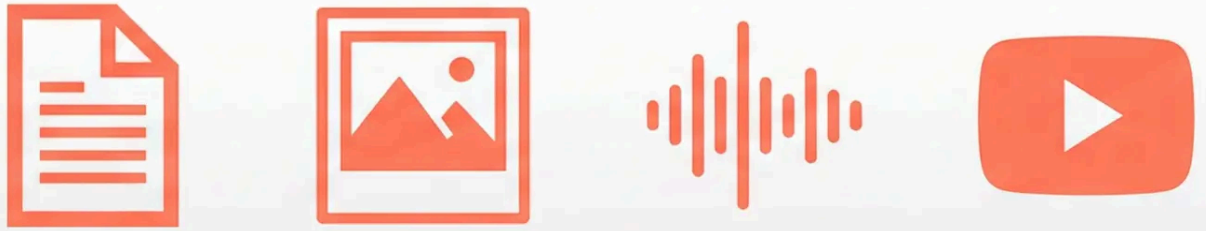


Why Unstructured Data Matters

85%

of enterprise data is unstructured

- Requires specialized AI processing pipelines
- Contains critical business insights
- Growing 57% from 2024 to 2026



Traditional databases cannot handle this complexity

What is Unstructured Data?

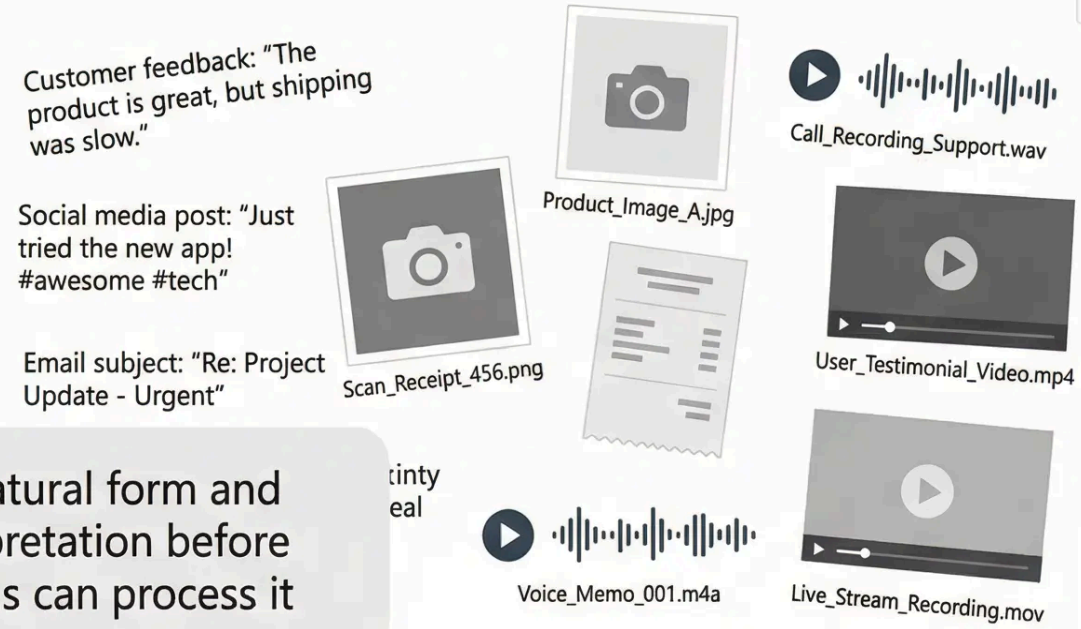
Structured Data

Fixed schemas & predefined formats

Customer ID	Name	Email Address	Purchase Date	Amount
1001	Jane Doe	jane.doe@email.com	2023-10-15	$150.00
1002	John Smith	john.smith@email.com	2023-10-16	$220.50
1003	Keae Willow	kann.smith@email.com	2023-10-27	$499.20
1004	...	...		
1005	...	...		

Unstructured Data

No predefined structure or schema



Data that exists in its natural form and requires semantic interpretation before machine learning models can process it

- **Cannot** fit into rows and columns
- Requires specialized processing pipelines
- Includes text, images, audio, and video

Types of Unstructured Data

Text

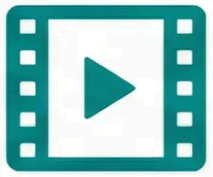
- Emails, reviews, documents
- Medical records, chat logs
- Social media posts

Images

- Photos, medical imaging
- Product scans, X-rays
- Surveillance footage

Audio

- Voice calls, podcasts
- Call center recordings
- Voice assistants



Video

- Security camera footage
- Interviews, live streams
- Training videos

Each modality requires specialized processing techniques

Challenges with Unstructured Data

High Dimensionality

Images and audio contain millions of dimensions, increasing computational costs

Noise & Redundancy

Raw data contains typos, slang, duplicates, and irrelevant content

Massive Storage

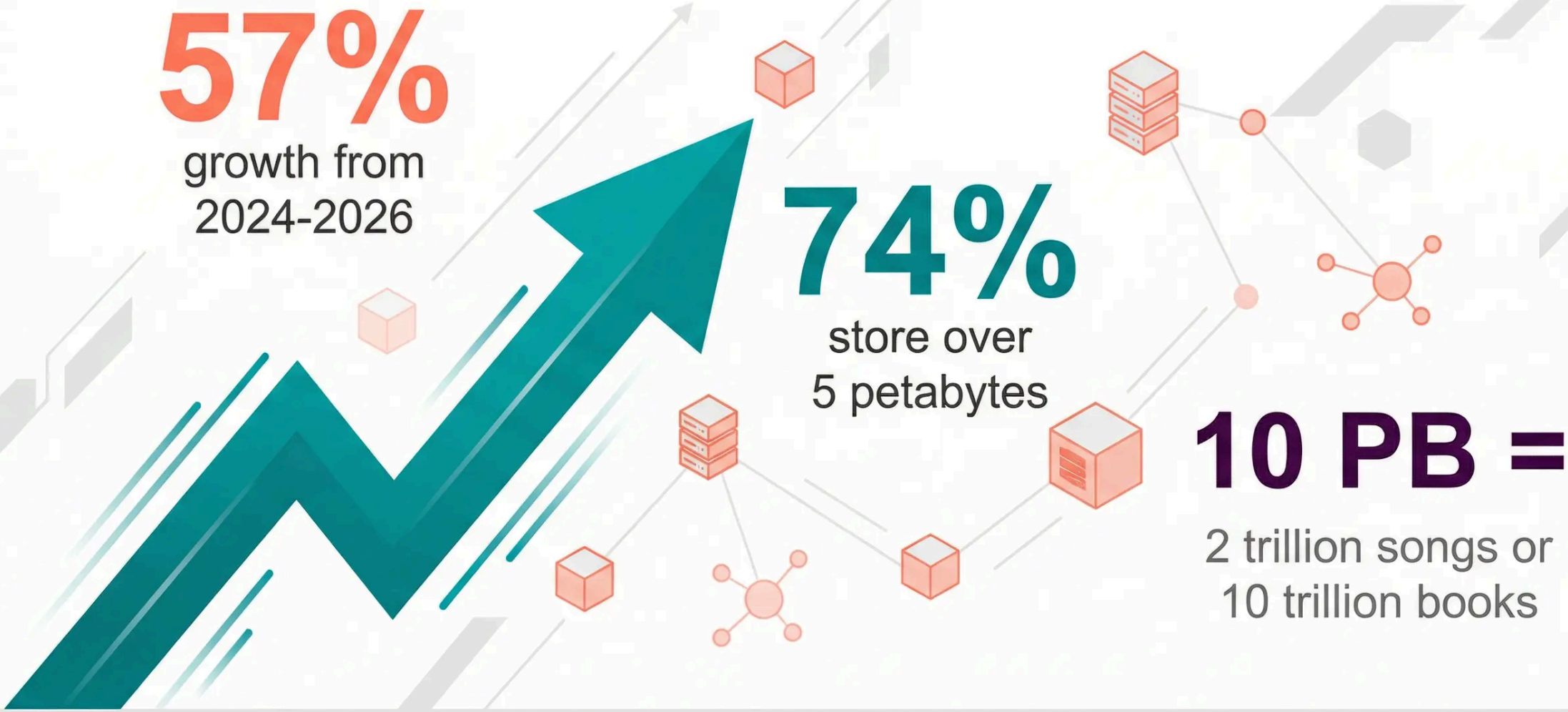
Exponential growth requires petabyte-scale infrastructure

Complex Pipelines

Each data type requires specialized preprocessing workflows

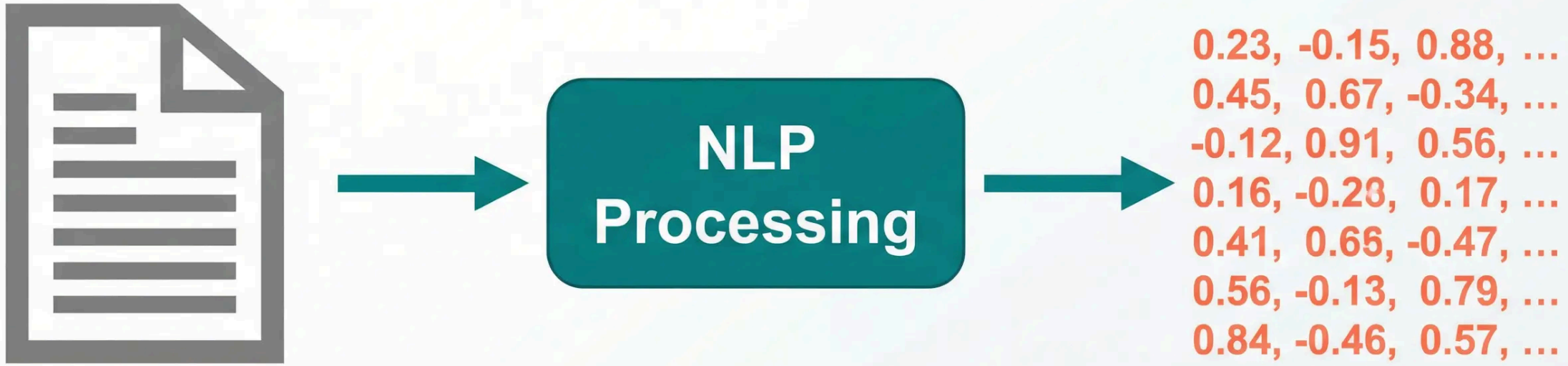
Traditional ETL tools fail under this complexity

The Unstructured Data Explosion



Enterprise data volumes are growing exponentially

Text Data Processing Overview

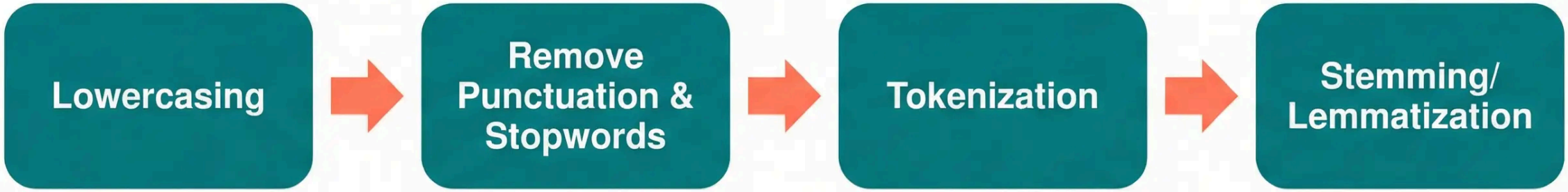


- Enables machine learning model consumption
- Critical for customer support and fraud detection
- Foundation for semantic understanding

Raw text → Numerical vectors

Text Preprocessing Pipeline

The RUNNING dogs are QUICKLY running!



run dog quick run

Standardized text ready for vectorization

Text Vectorization Techniques

Bag of Words (BoW)

Counts word frequencies, ignores grammar and order

Use case: Simple classification, spam detection

TF-IDF

Weights terms by importance across corpus

Use case: Information retrieval, document ranking

Word Embeddings

Maps words into dense vector spaces (Word2Vec, GloVe)

Use case: Semantic search, recommendations

LLM Embeddings

Context-aware, high-dimensional vectors

Use case: Advanced NLP, question answering

Evolution: Simple → Sophisticated

Advanced NLP with Transformers

BERT

- Bidirectional understanding
- Sentiment analysis
- Named Entity Recognition

RoBERTa

- Optimized BERT training
- Enhanced performance
- Document summarization

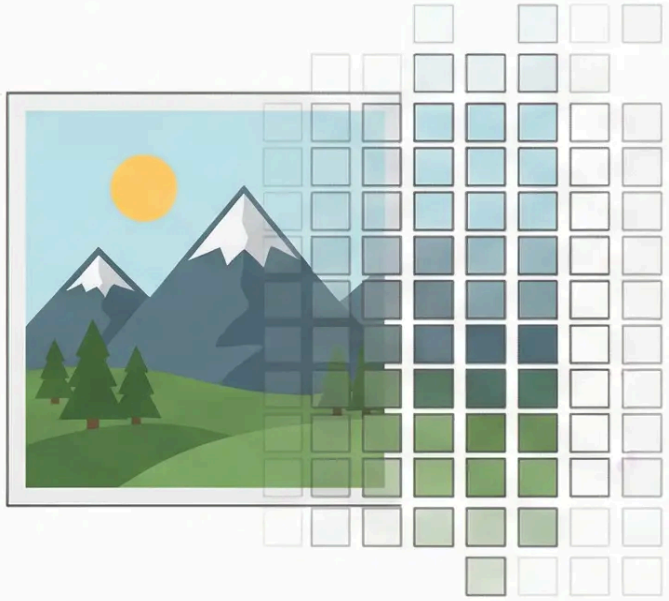
Llama

- Large language model
- Contextual processing
- Real-time intent classification

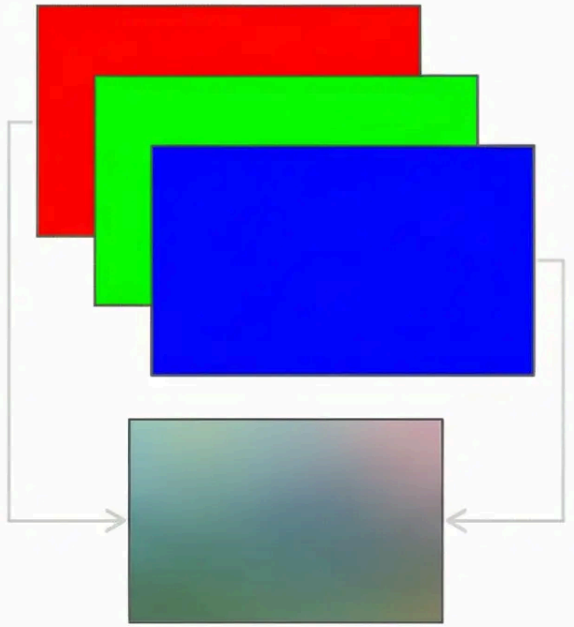
Transformers process entire sentences in parallel, capturing long-range dependencies and contextual relationships

Enables high-accuracy NLP tasks

Image Data Processing Overview



- Images stored as 2D pixel arrays
- RGB channels: Red, Green, Blue (0-255)
- 224×224 image = 150,000+ data points



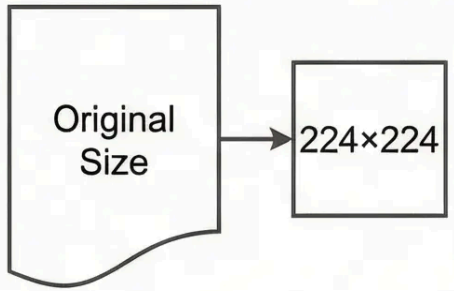
Raw images require preprocessing for ML model compatibility

Raw Image → Pixel Arrays → Preprocessed Data → ML Model Input

Image Preprocessing Techniques

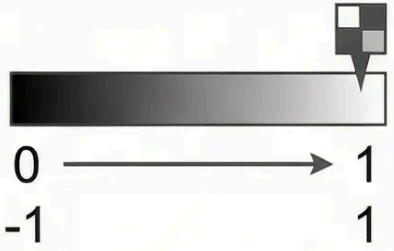
Resizing

Fixed dimensions (224×224)



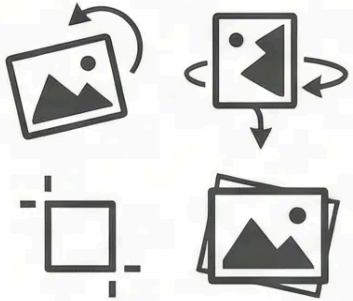
Normalization

Pixel values [0,1] or [-1,1]



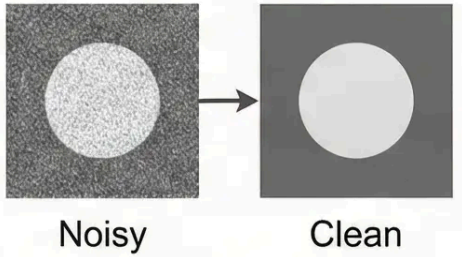
Data Augmentation

Rotation, flip, crop, jitter



Noise Reduction

Gaussian blur, filtering



Standardized preprocessing ensures consistent model input

Feature Extraction from Images

Traditional Handcrafted Features

- Edge detection (Sobel, Canny filters)
- Texture analysis (Local Binary Patterns)
- Shape contours and geometric features

Convolutional Neural Networks

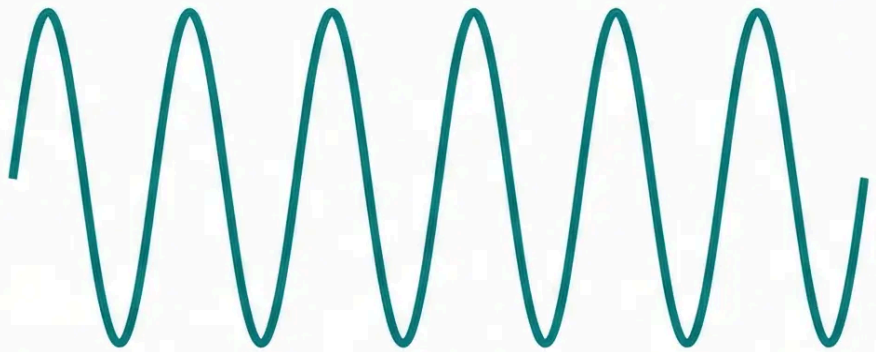
- Automatic hierarchical feature learning
- ResNet, EfficientNet architectures
- End-to-end training from raw pixels

Vision Transformers

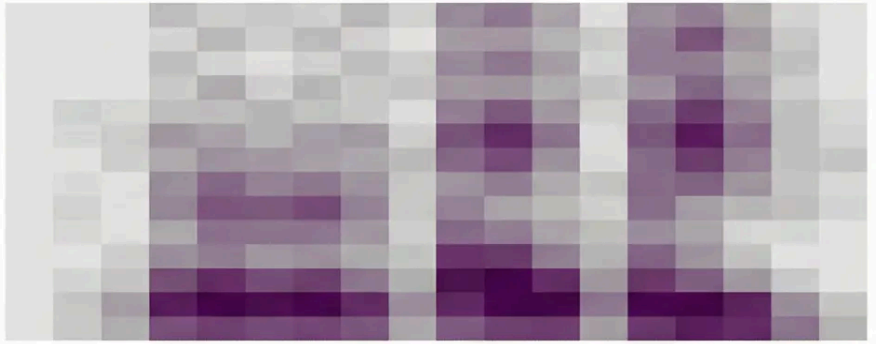
- Transformer architecture for images
- Global context and long-range dependencies

Evolution: Manual → Automatic → Global Understanding

Audio Data Processing Overview



Time-Series Waveforms



Frequency Representations

- Audio stored as amplitude variations over time
- Sampling rates: 16kHz or 44.1kHz typically used
- Spectrograms enable machine learning analysis

Raw waveforms are high-dimensional and temporally complex

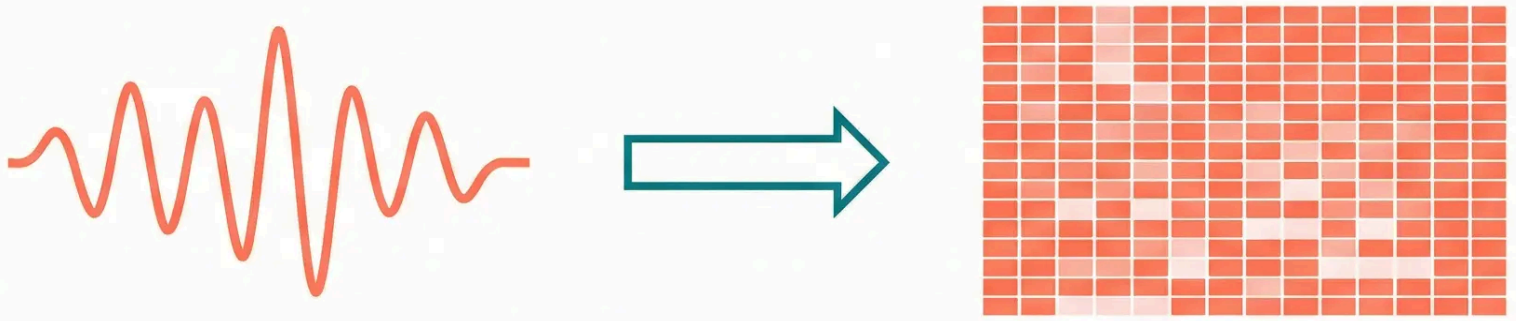
Audio Preprocessing and Features

Preprocessing Steps

- Noise removal and filtering
- Sampling rate standardization (16kHz)
- Silence trimming
- Audio segmentation

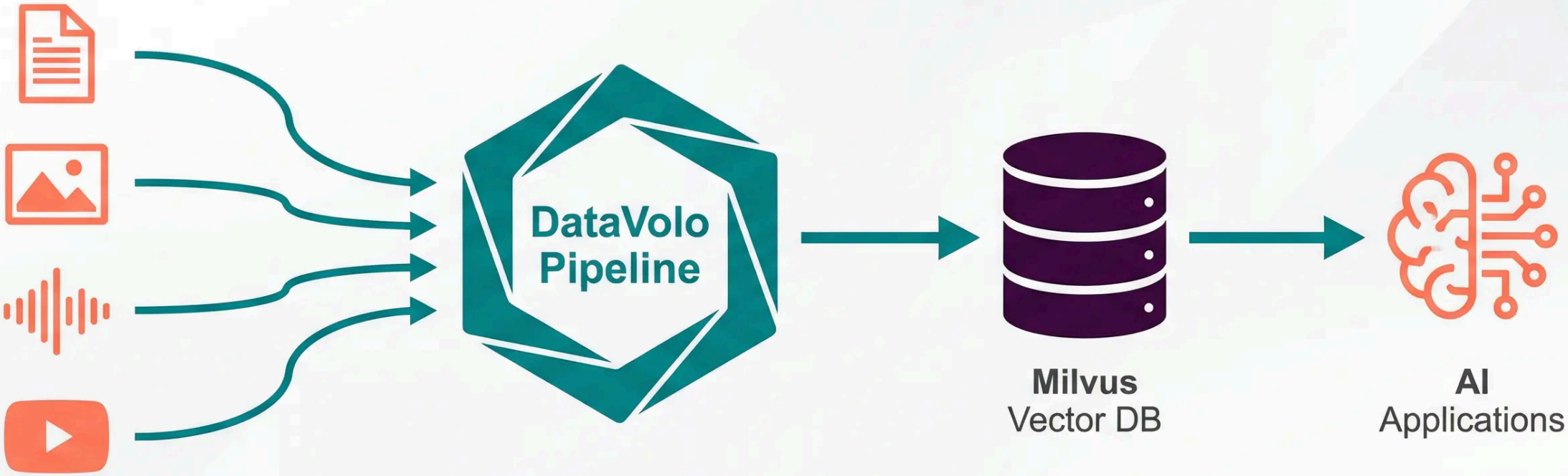
Feature Extraction

- Spectrograms (time-frequency analysis)
- MFCCs (Mel-Frequency Cepstral Coefficients)
- Pitch, energy, and duration features



Raw audio → Processed signal → Feature vectors

Unified Multimodal Pipelines

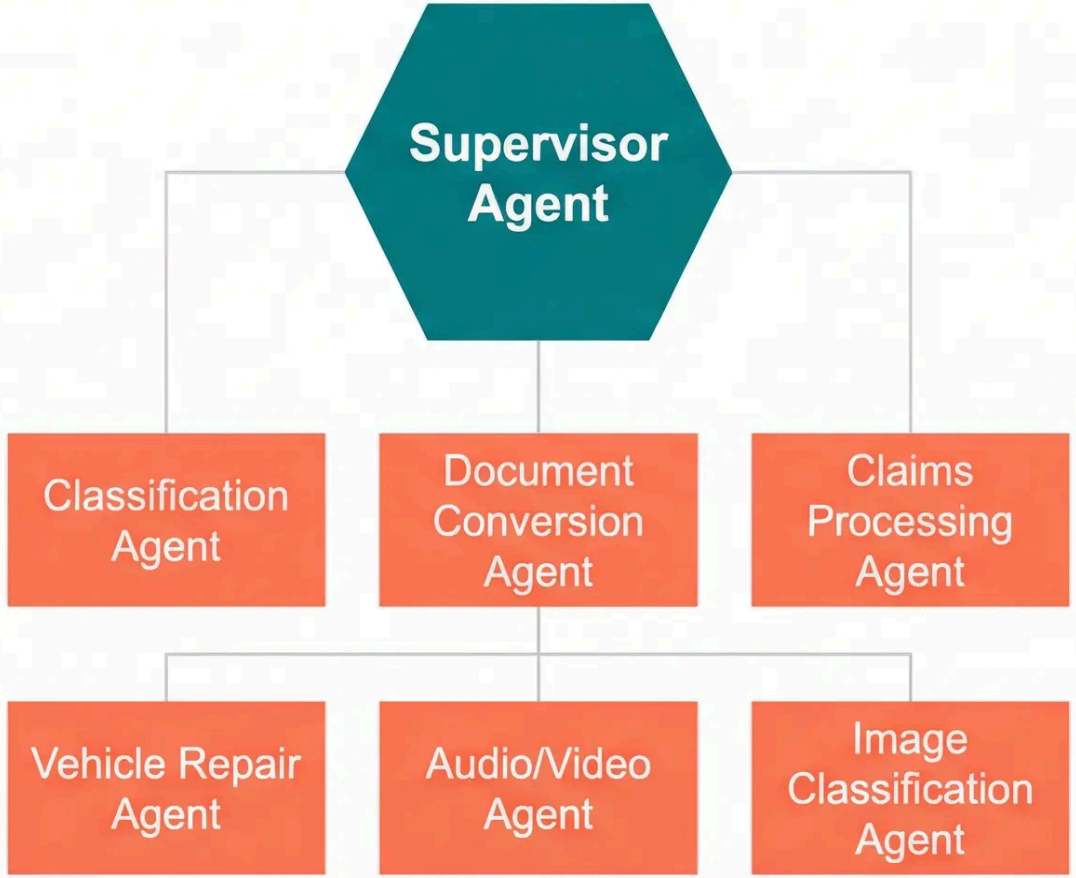


- Real-time ingestion across AWS, GCP, Azure
- Automated document parsing with layout detection
- Vector storage enabling fast similarity search

Powered by Apache NiFi

Production-Ready Multi-Agent Systems

- Modular & Scalable Architecture
- Independent Agent Updates
- Real-time Processing



AWS Integration

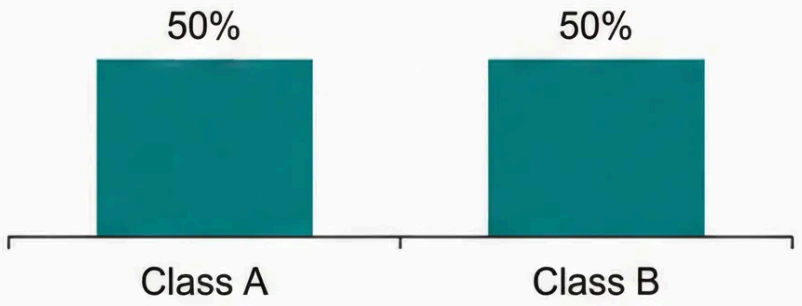
- Event-driven workflows
- Cloud-native deployment
- Metadata-rich data lakes

Each agent specializes in specific data types and domain tasks

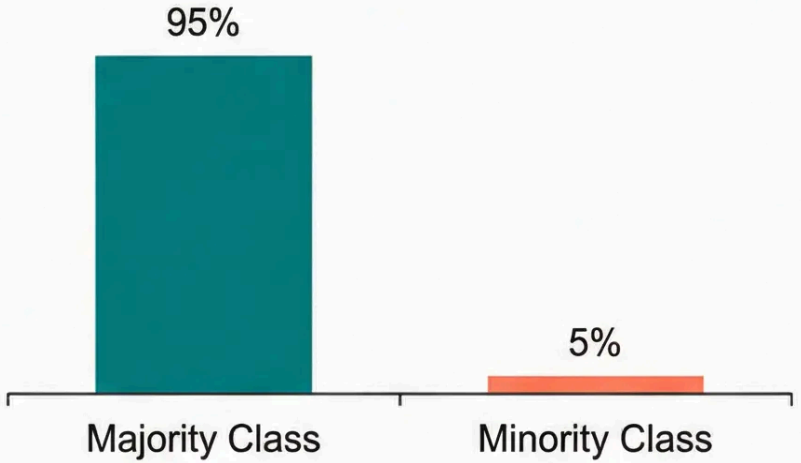
What is Imbalanced Data?

Class distribution where one class significantly outnumbers others

Balanced Dataset



Imbalanced Dataset



Common in real-world scenarios:

Fraud Detection: 0.17% fraudulent transactions

Medical Diagnosis: 0.06% disease cases

Real-World Impact of Imbalanced Data

Medicare Fraud Detection

Only 0.06% fraud rate

6 fraudulent cases per 10,000 transactions

Healthcare Cyberattacks

624 attacks in 2023

Double from 304 attacks in 2022

Detection failures can result in financial losses and life-threatening consequences

The Danger of Misleading Metrics

99.94%

Accuracy

0%

Fraud Cases Detected

- Medicare fraud: only 0.06% of cases
- Model predicts ‘normal’ for everything
- Silent failure in critical applications

High accuracy ≠ Good performance

Cybersecurity Crisis in Healthcare

624 cyberattacks in 2023

***More than double** the 304 attacks in 2022*

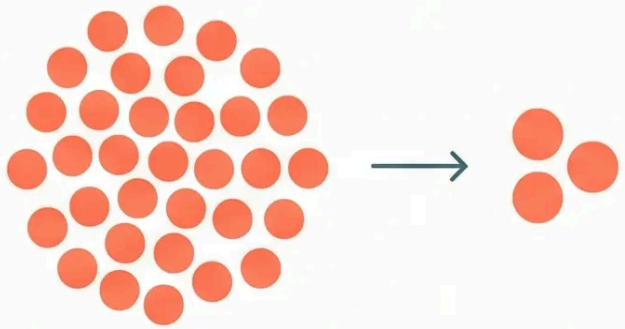
- AI-powered threats like FraudGPT
- Available for just $200/month
- Indistinguishable from legitimate communications

Düsseldorf Hospital Case: Cyberattack disrupted emergency care, contributing to preventable patient death

Resampling Techniques Overview

Undersampling

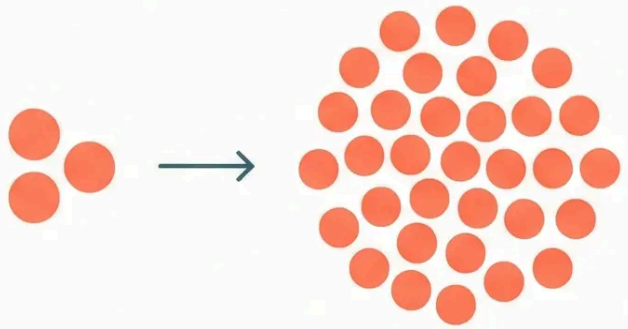
Reduces majority class samples



Risk: Information loss

Oversampling

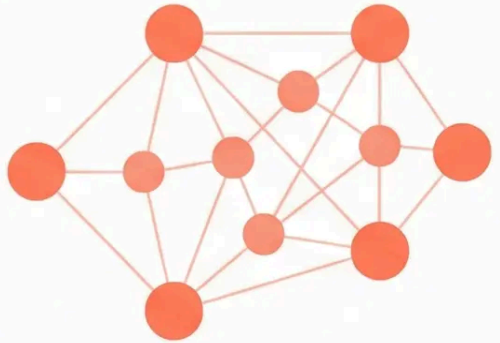
Duplicates minority class samples



Risk: Overfitting

Synthetic Generation

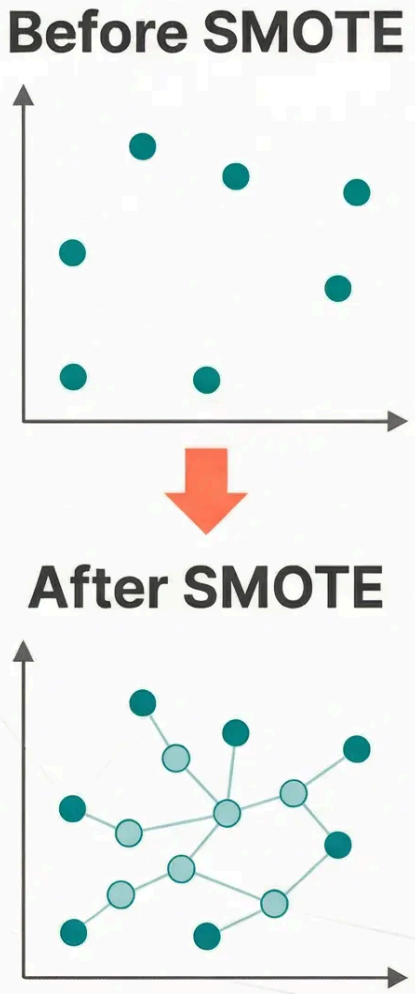
Creates new artificial samples



Best: Balanced approach

Goal: Balance class distributions for better model performance

SMOTE and Advanced Variants



- Creates synthetic samples by interpolation
- Avoids simple duplication of data
- Improves decision boundaries

Technique	Accuracy	Precision	Recall
SMOTE	91.30%	49.90%	91.00%
SMOTE-ENN	92.10%	51.20%	92.50%

SMOTE-ENN combines oversampling with **noise removal**

Cost-Sensitive Learning

Assigns higher misclassification costs to minority class errors

Incorporates business costs into learning objectives

Traditional Learning

All errors = equal cost

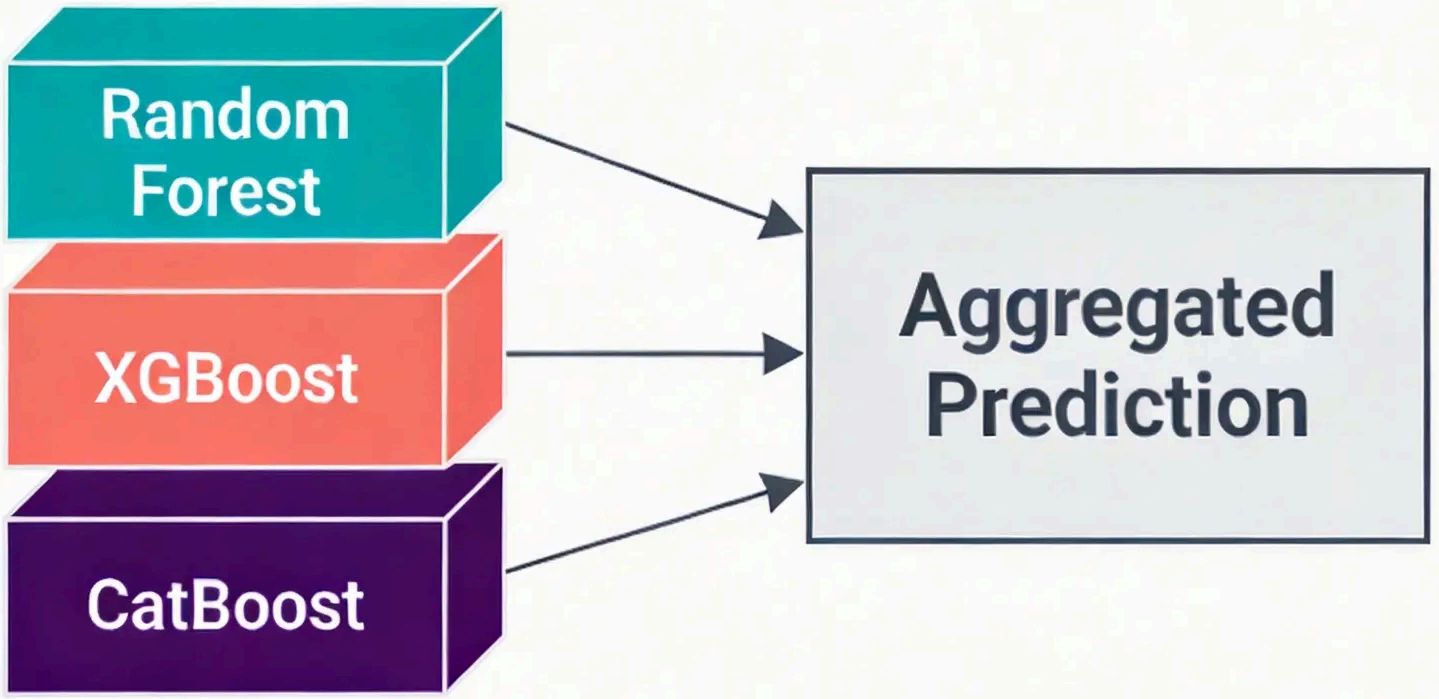
Cost-Sensitive Learning

Minority errors = higher cost

Example: Fraud Detection
Missing fraud >> False alarm
Cost ratio: 100:1

Implementation: class_weight parameter in scikit-learn

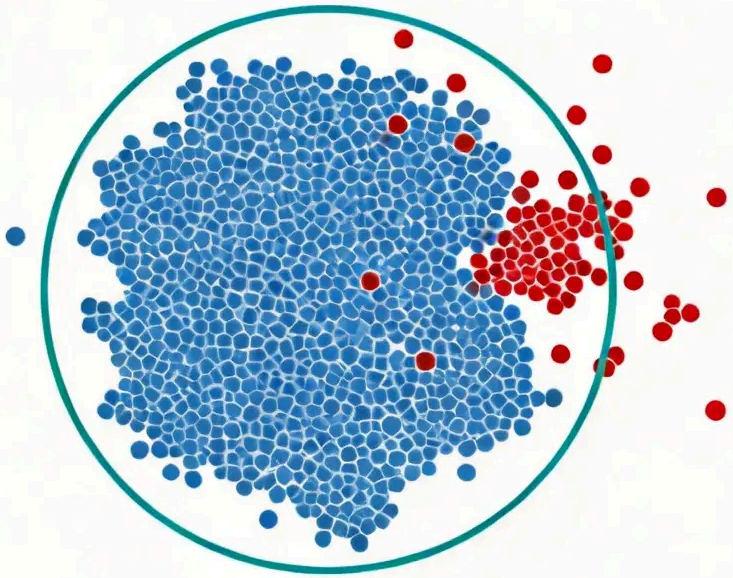
Ensemble Methods for Imbalanced Data



- Combines multiple weak learners
- Reduces variance and bias
- Handles class imbalance naturally

CatBoost achieves 83.36% accuracy on imbalanced datasets

One-Class Classification



Blue: Normal Data
Red: Detected Anomalies

- Models majority class characteristics
- Identifies deviations as anomalies
- Ideal for extreme imbalance (1:100+ ratios)
- No need for minority class examples

Fraud Detection

Intrusion Detection

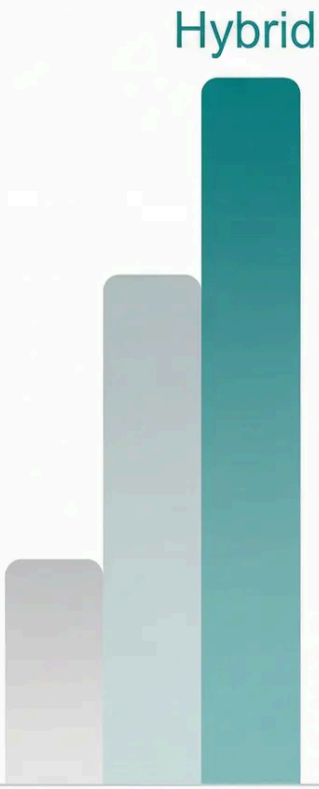
Industrial Monitoring

Best for: Learning 'normal' behavior patterns

Hybrid Approach Performance

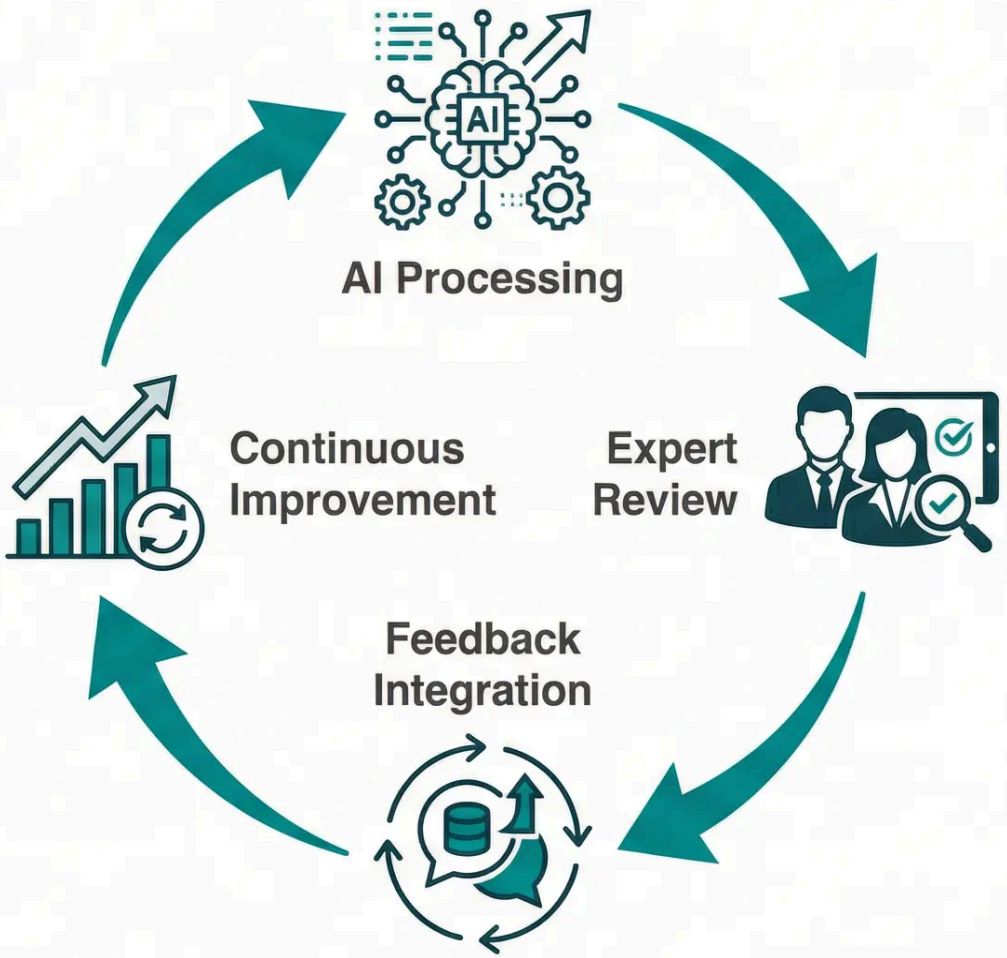
Approach	Accuracy	Log Loss	Performance
Undersampling	72.68%	0.73	Baseline
Oversampling	82.86%	0.48	Good
XGBoost	83.05%	–	Good
SMOTE + CatBoost	**83.36%**	**0.46**	**Best**

Hybrid approaches combining data-level and algorithm-level techniques achieve superior performance



Human-in-the-Loop Validation

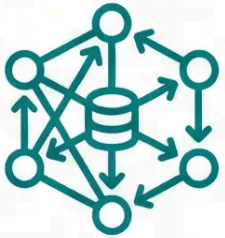
- Domain experts validate extracted metadata
- Corrections mapped to specific issue types
- Automated resolvers reduce manual intervention



Critical for high-stakes domains: Insurance, Healthcare, Finance

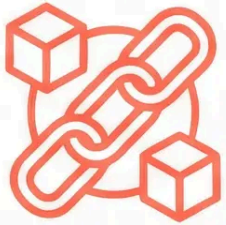
Ensures accuracy while progressively reducing human dependency

Future Directions



Federated Learning

- Train models without sharing raw data
- Google & Apple successfully deployed
- Ideal for healthcare & finance sectors



Blockchain for Data Integrity

- Immutable data provenance tracking
- Dynamic consent management
- Automated compliance via smart contracts

Privacy-Preserving AI Ecosystems

Key Takeaways

- Most AI data is unstructured and imbalanced
- Multimodal preprocessing is essential for real-world AI
- Advanced resampling techniques handle data imbalance
- Intelligent data management leads to trustworthy AI

The future belongs to those who turn data chaos into actionable intelligence