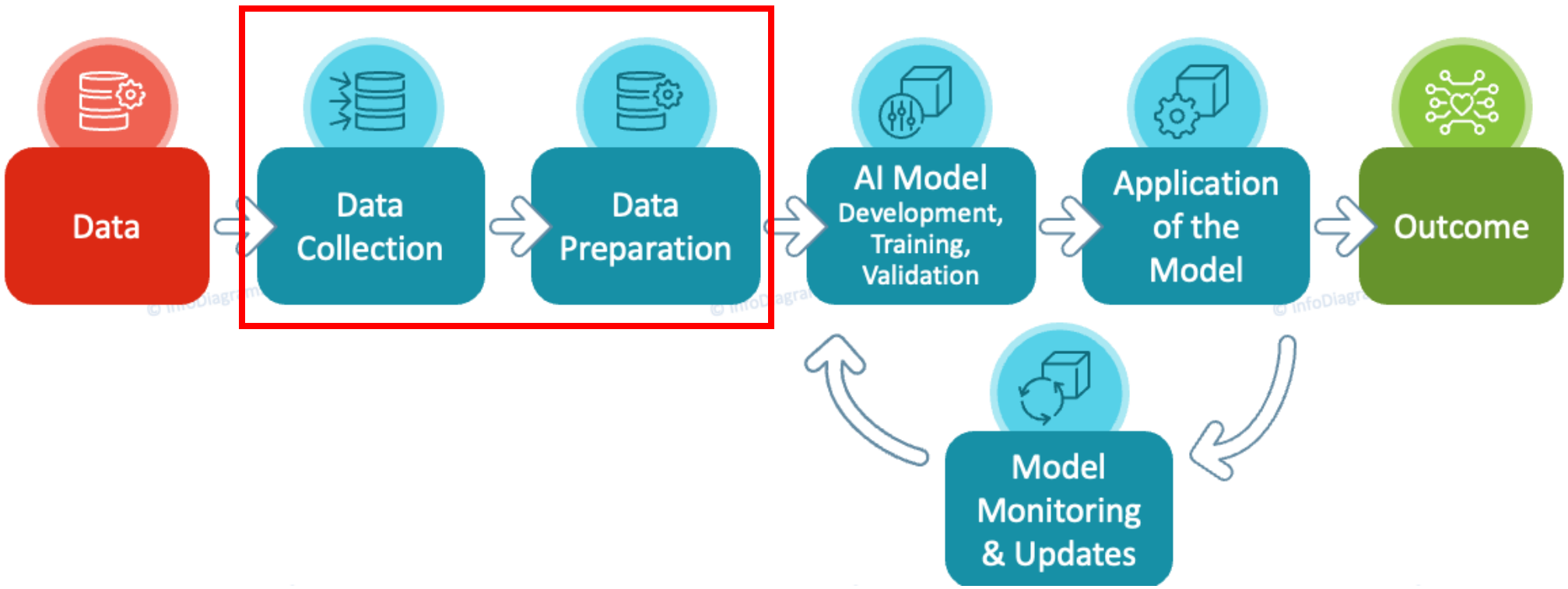
IMEN891K/IDSC501 Data Engineering

6. **Pre-processing III**: Reduction

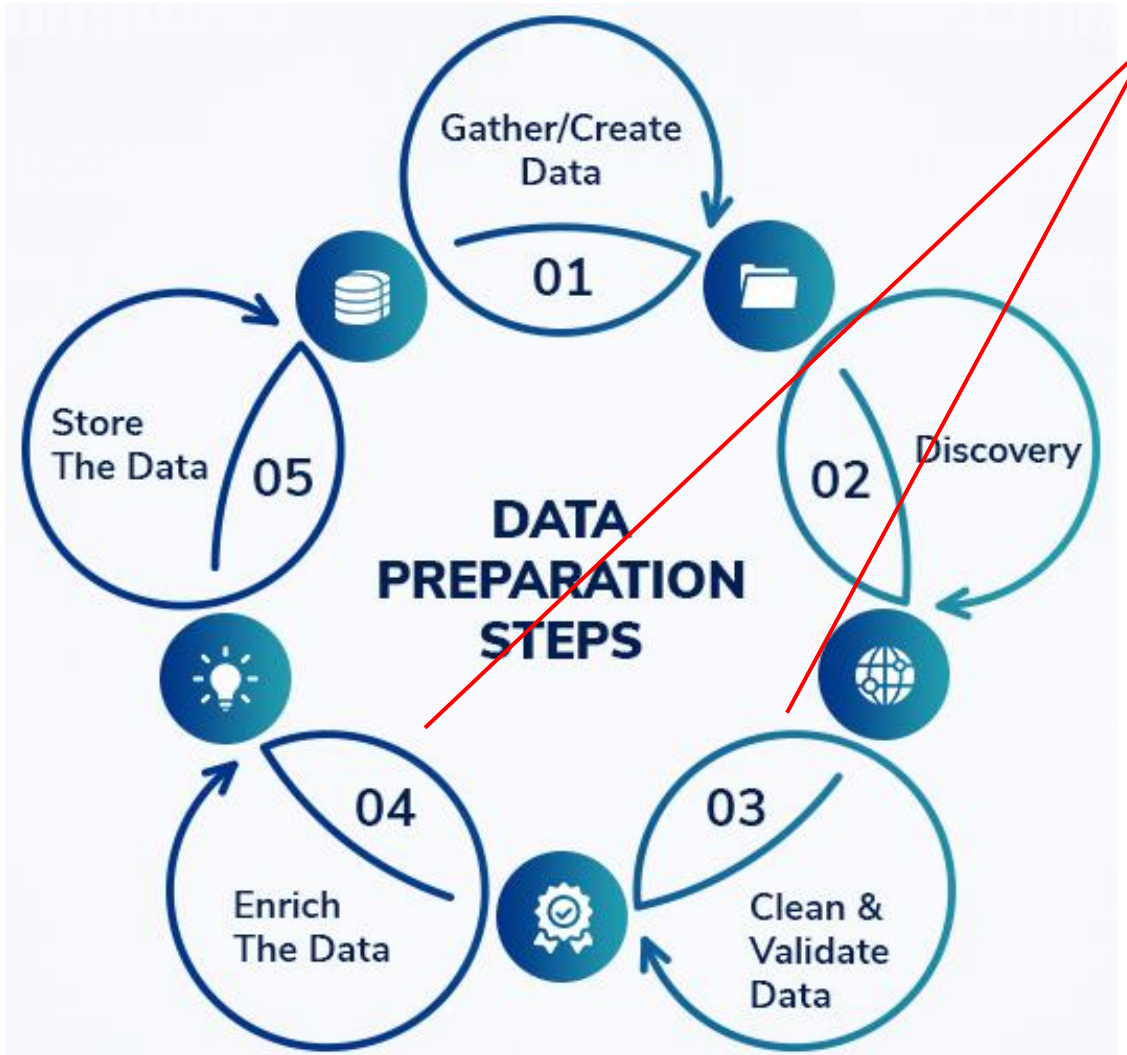
Fall 2025

Instructor: **Sangwoo Mo**

Recap: Course Overview

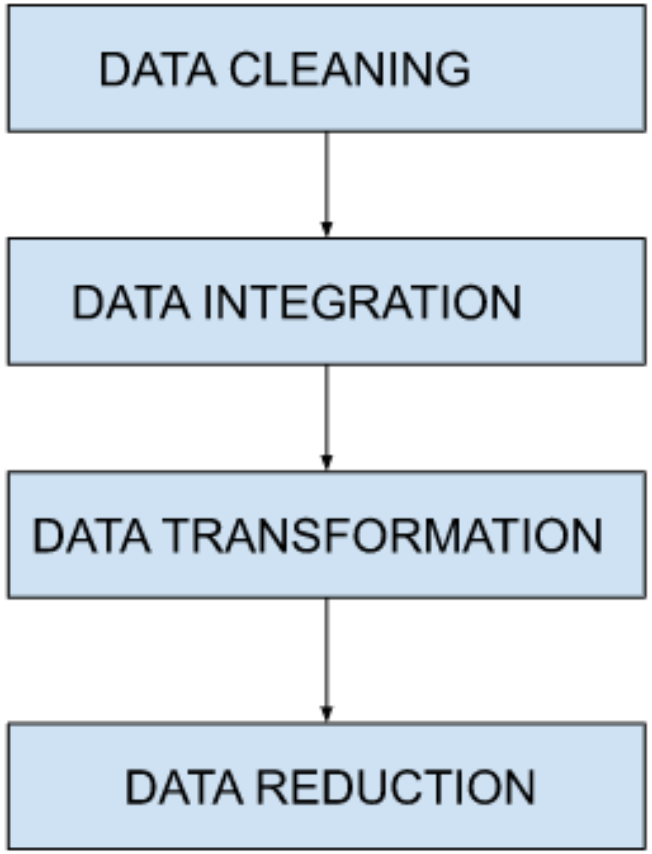


Recap: Course Overview

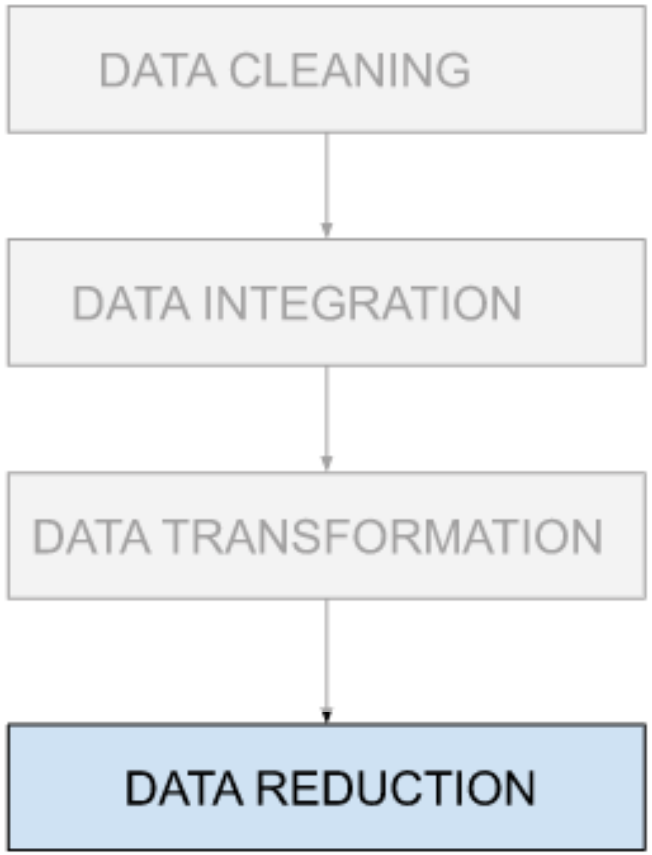


Lecture 4-6

Outline: Data Pre-processing



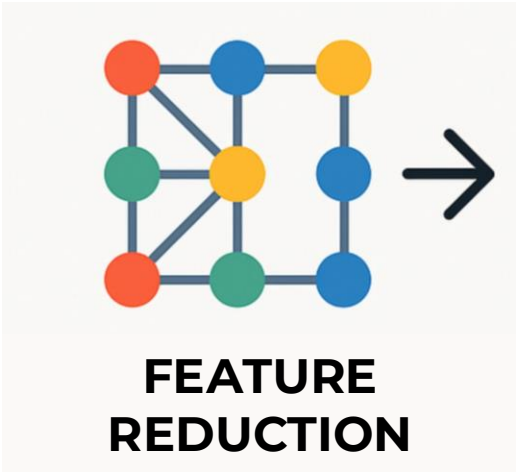
Outline: Data Pre-processing



Data Reduction



INSTANCE REDUCTION



FEATURE REDUCTION



DATA COMPRESSION

Instance Reduction



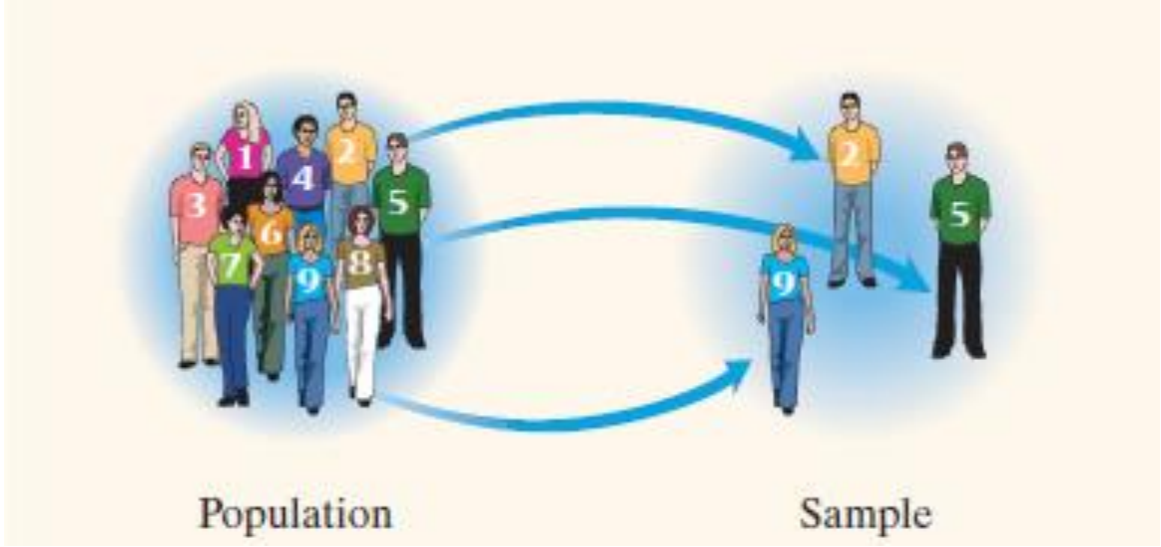
Sampling vs. Curation

Aspect	Data Sampling	Data Curation
Goal	Reduce computation, handle imbalance, ensure fairness	Improve data quality, consistency, and representativeness
Scope	Select a subset of the data	Refine and improve the entire dataset
Focus	Efficiency, smaller size, representative subset	Quality, diversity, removing noise or bias
Techniques	Random, Stratified, Cluster, Systematic sampling	Filtering, Cleaning, Balancing, Augmentation, Deduplication
Application	Model training with less cost, quick experiments	Reliable training/testing, trustworthy downstream tasks
In short	Choosing **which data to keep**	Improving **how good the data is**

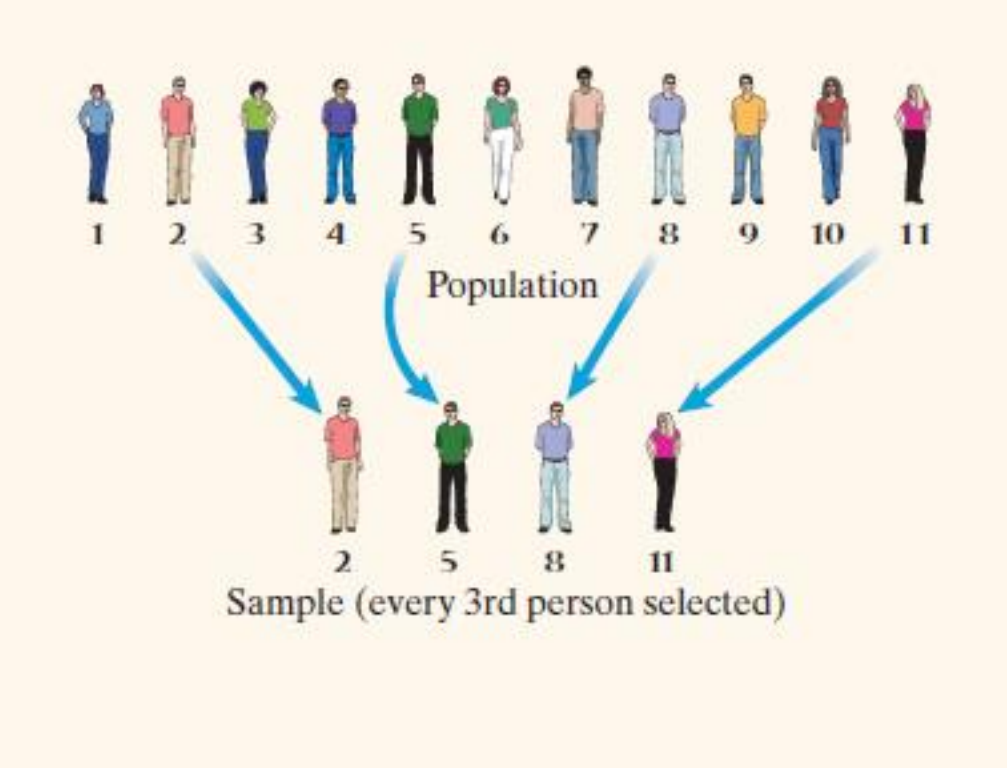
Data Sampling

Aspect	Data Sampling	Data Curation
Goal	Reduce computation, handle imbalance, ensure fairness	Improve data quality, consistency, and representativeness
Scope	Select a subset of the data	Refine and improve the entire dataset
Focus	Efficiency, smaller size, representative subset	Quality, diversity, removing noise or bias
Techniques	Random, Stratified, Cluster, Systematic sampling	Filtering, Cleaning, Balancing, Augmentation, Deduplication
Application	Model training with less cost, quick experiments	Reliable training/testing, trustworthy downstream tasks
In short	Choosing **which data to keep**	Improving **how good the data is**

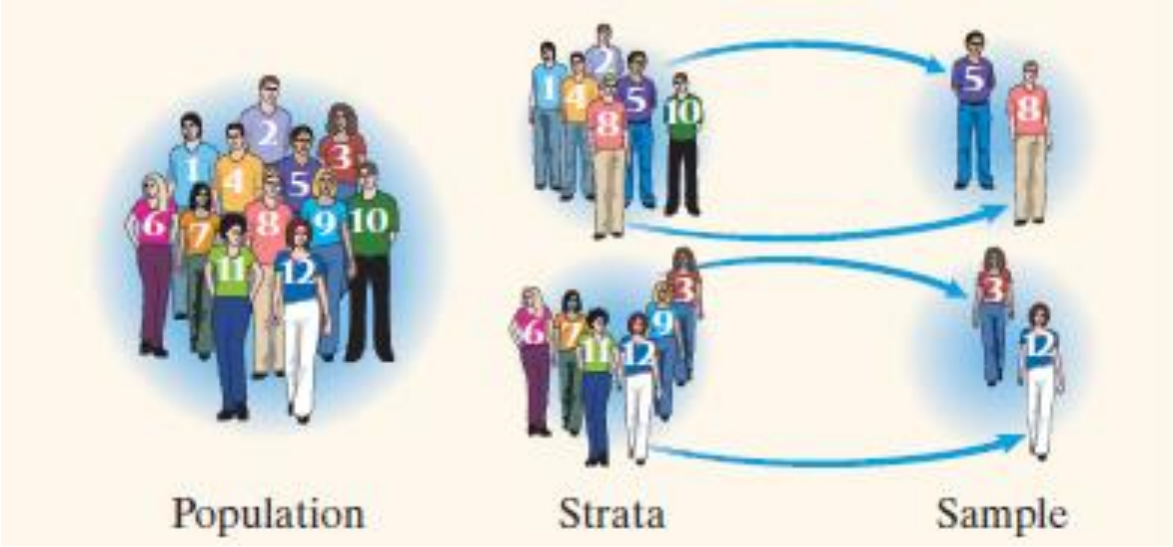
Random Sampling



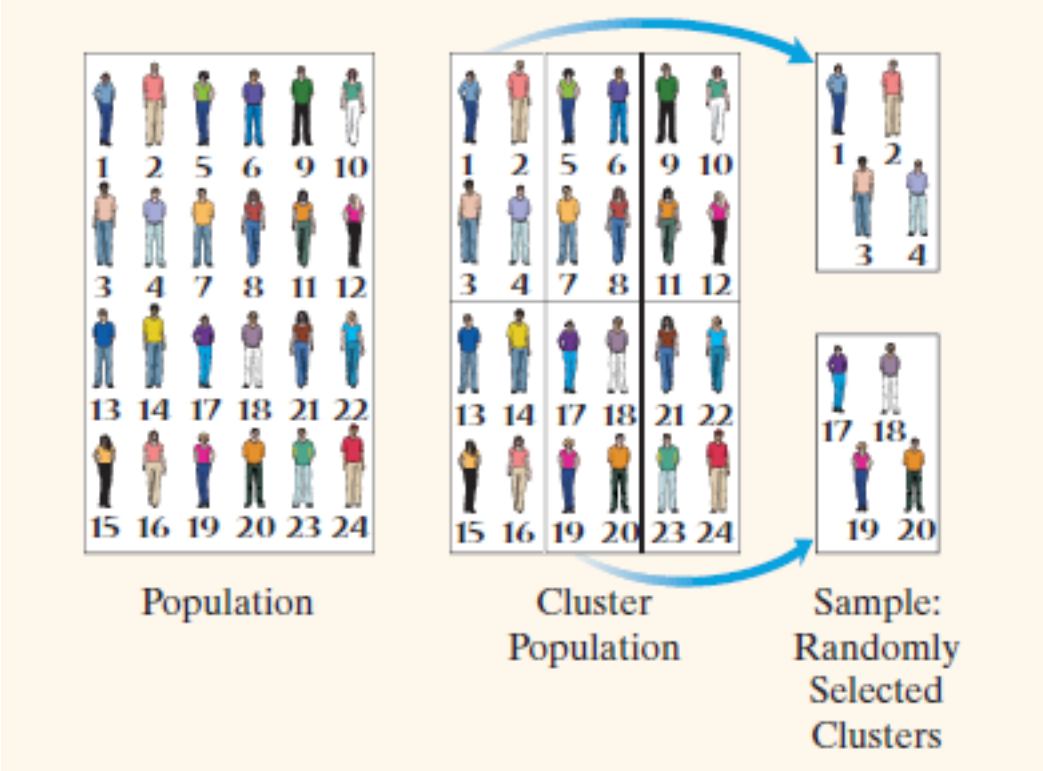
Systematic Sampling



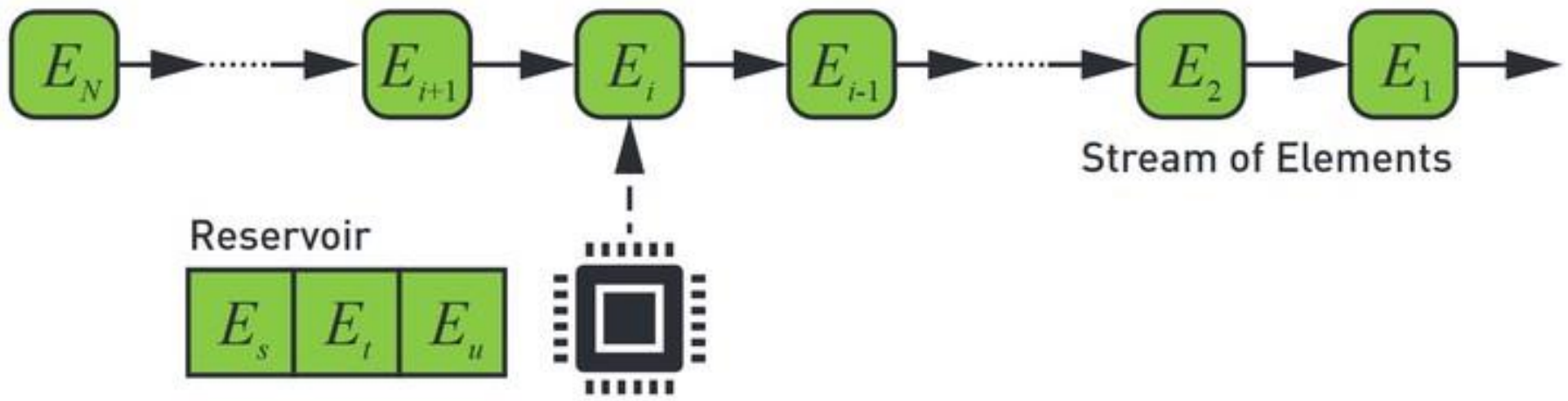
Stratified Sampling



Cluster Sampling



Reservoir Sampling

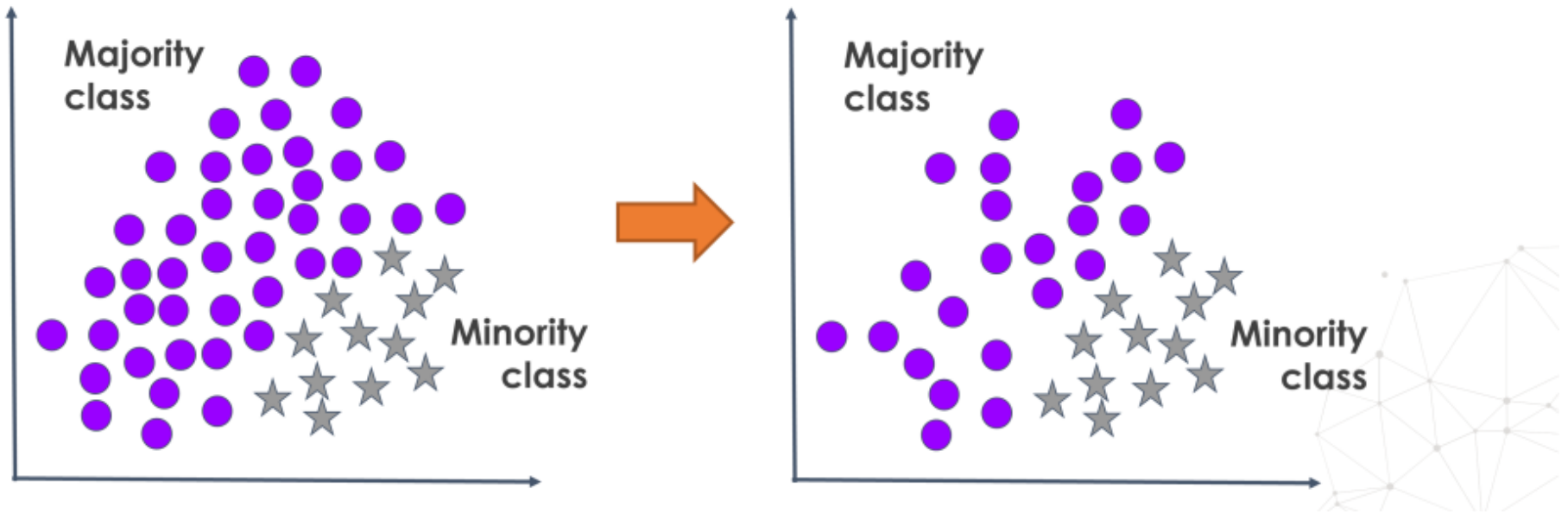


Application: Reduce Computation

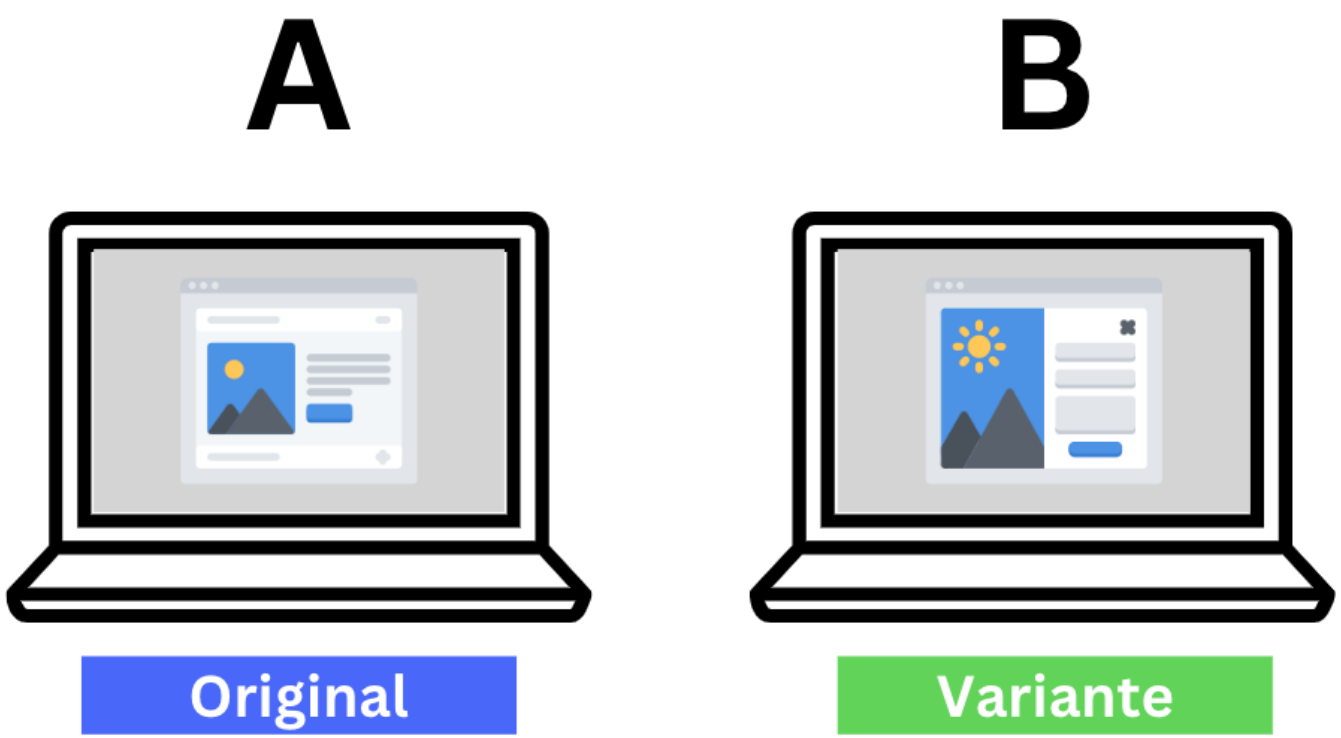
An estimator $\hat{\theta}$ for a parameter θ is called **unbiased** if its expected value equals the true parameter, for every sample size n. Formally:

$$E[\hat{\theta}] = \theta$$

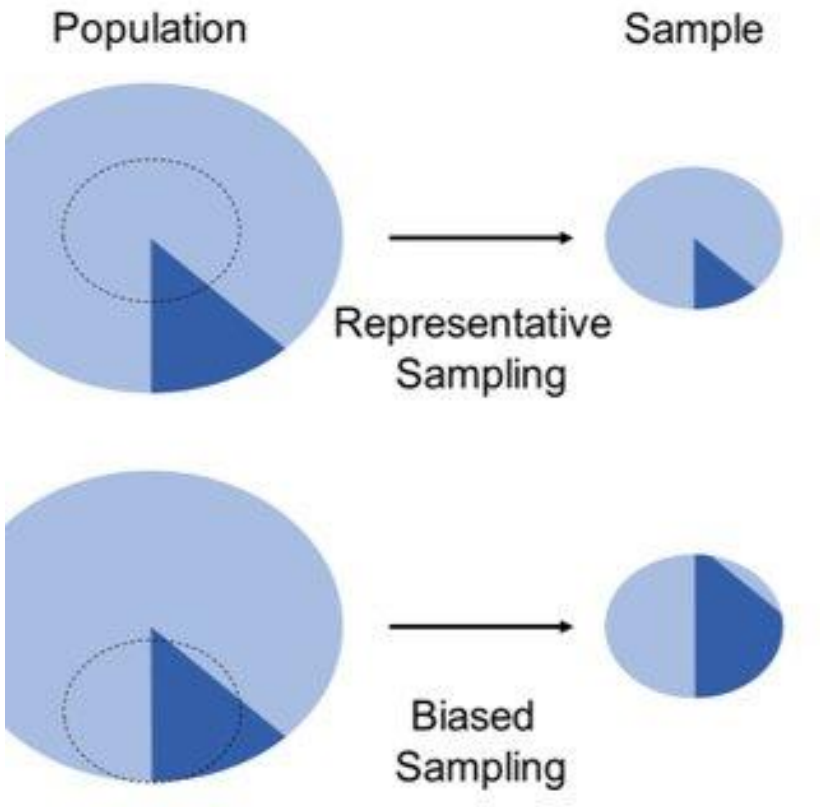
Application: Data Imbalance



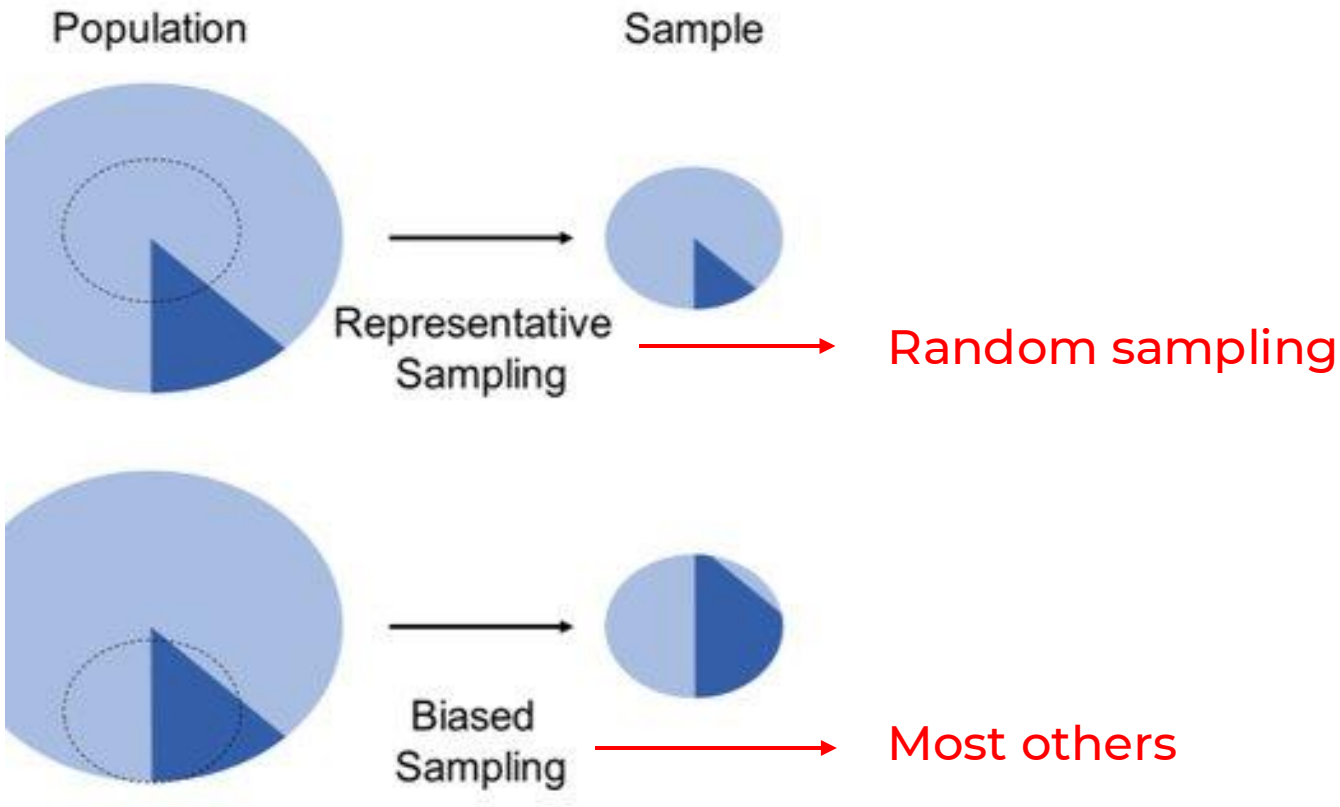
Application: A/B Test



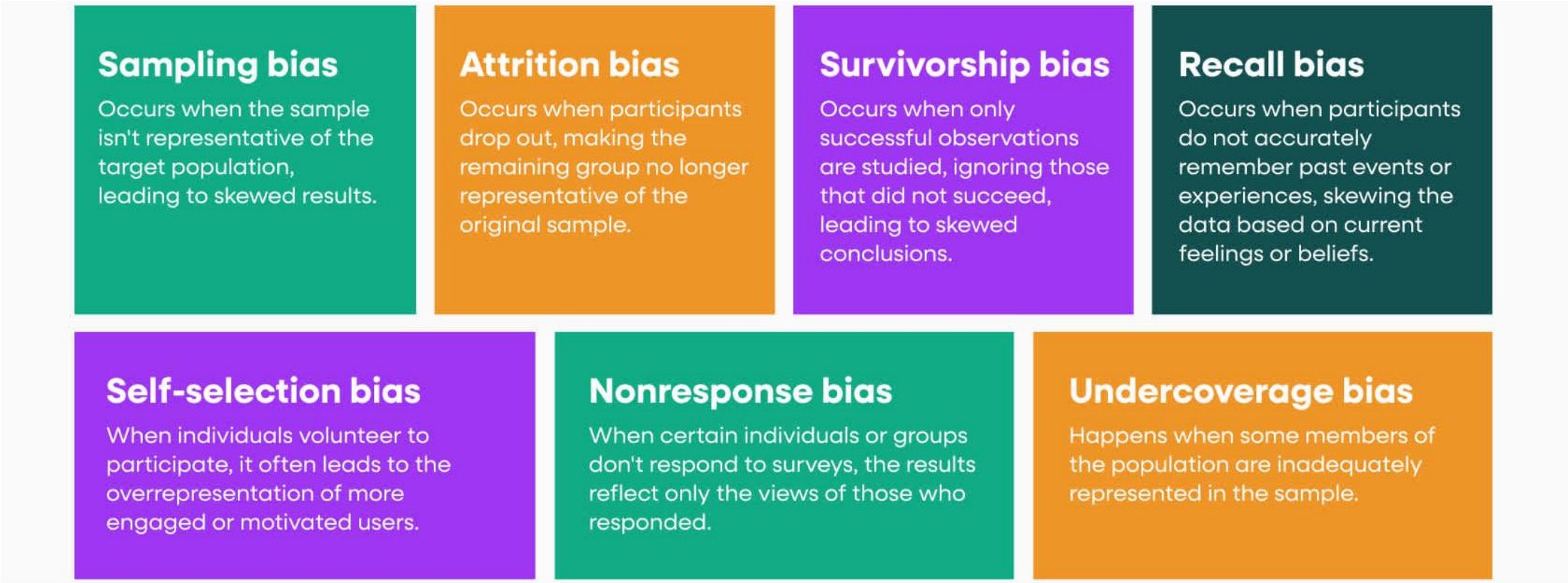
Caveat: Sampling Bias



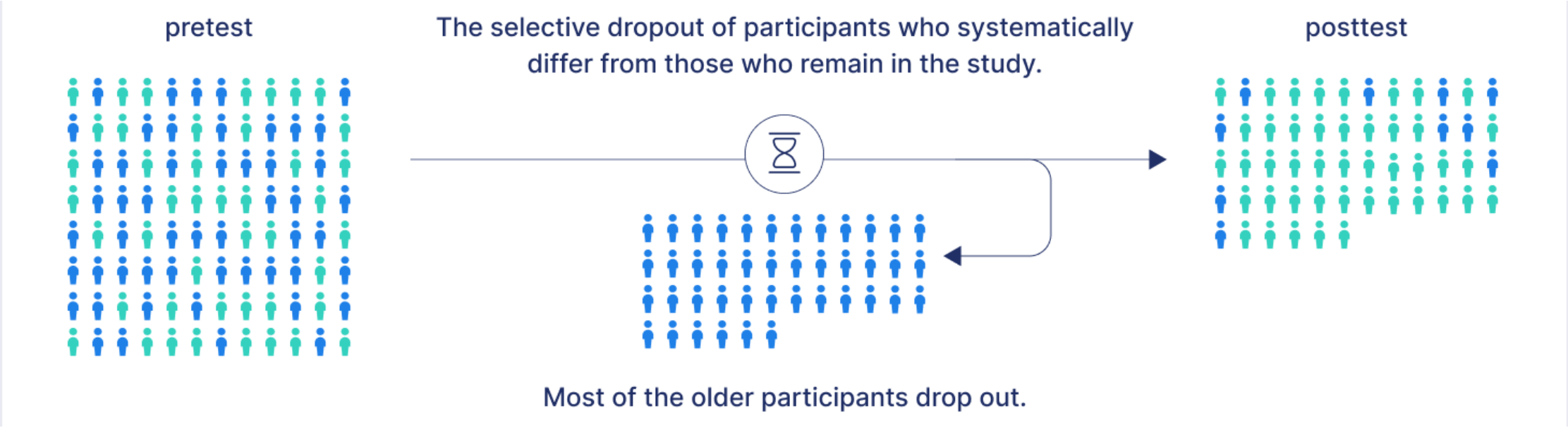
Caveat: Sampling Bias



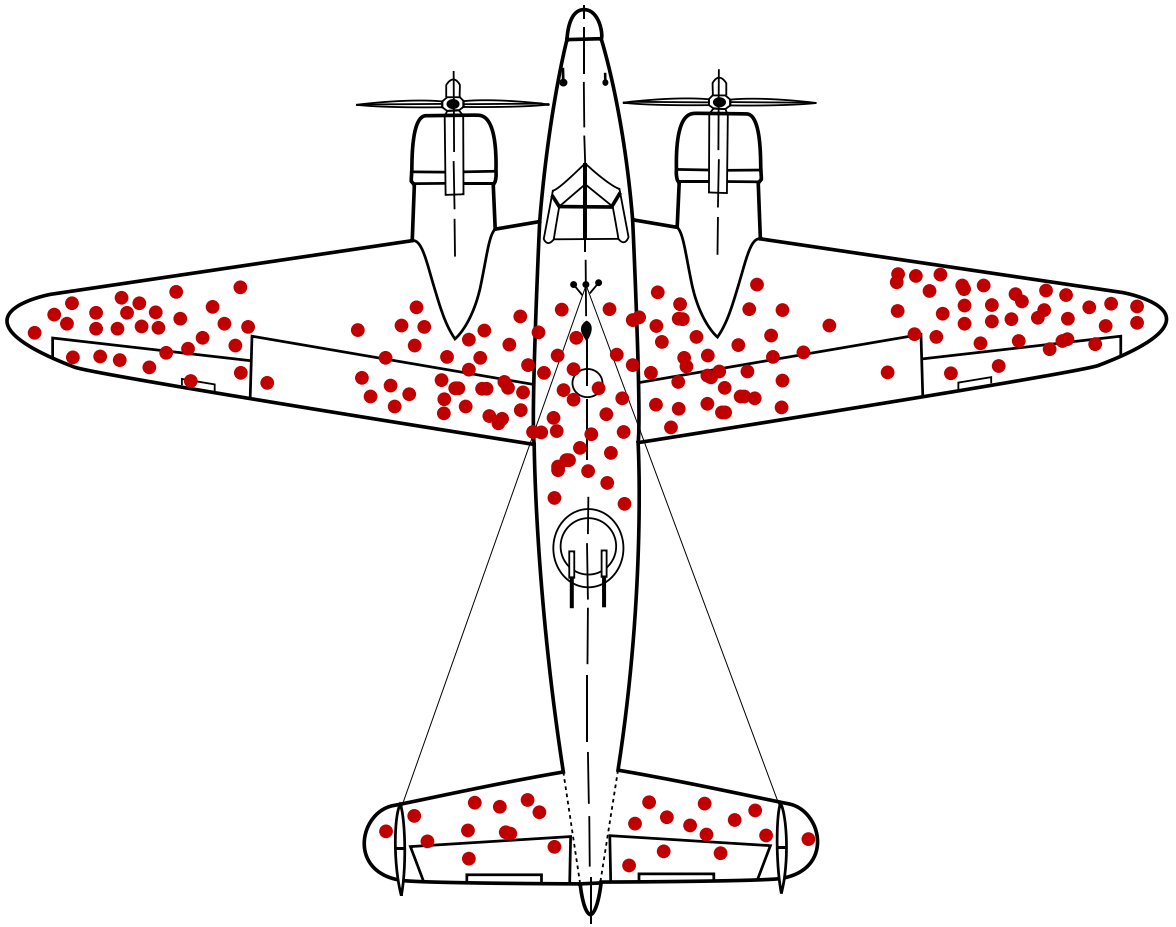
More Sampling Biases



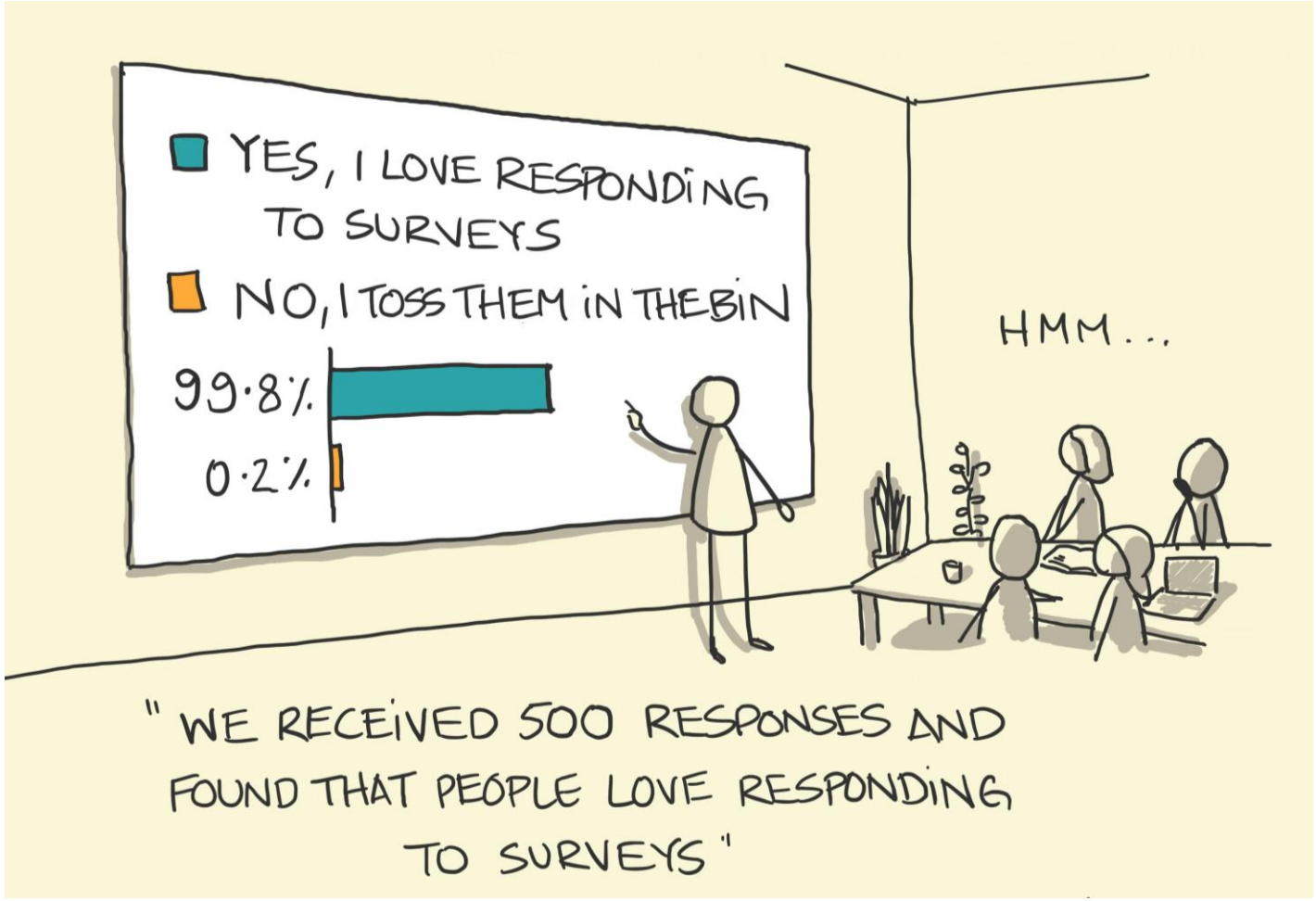
Attrition Bias



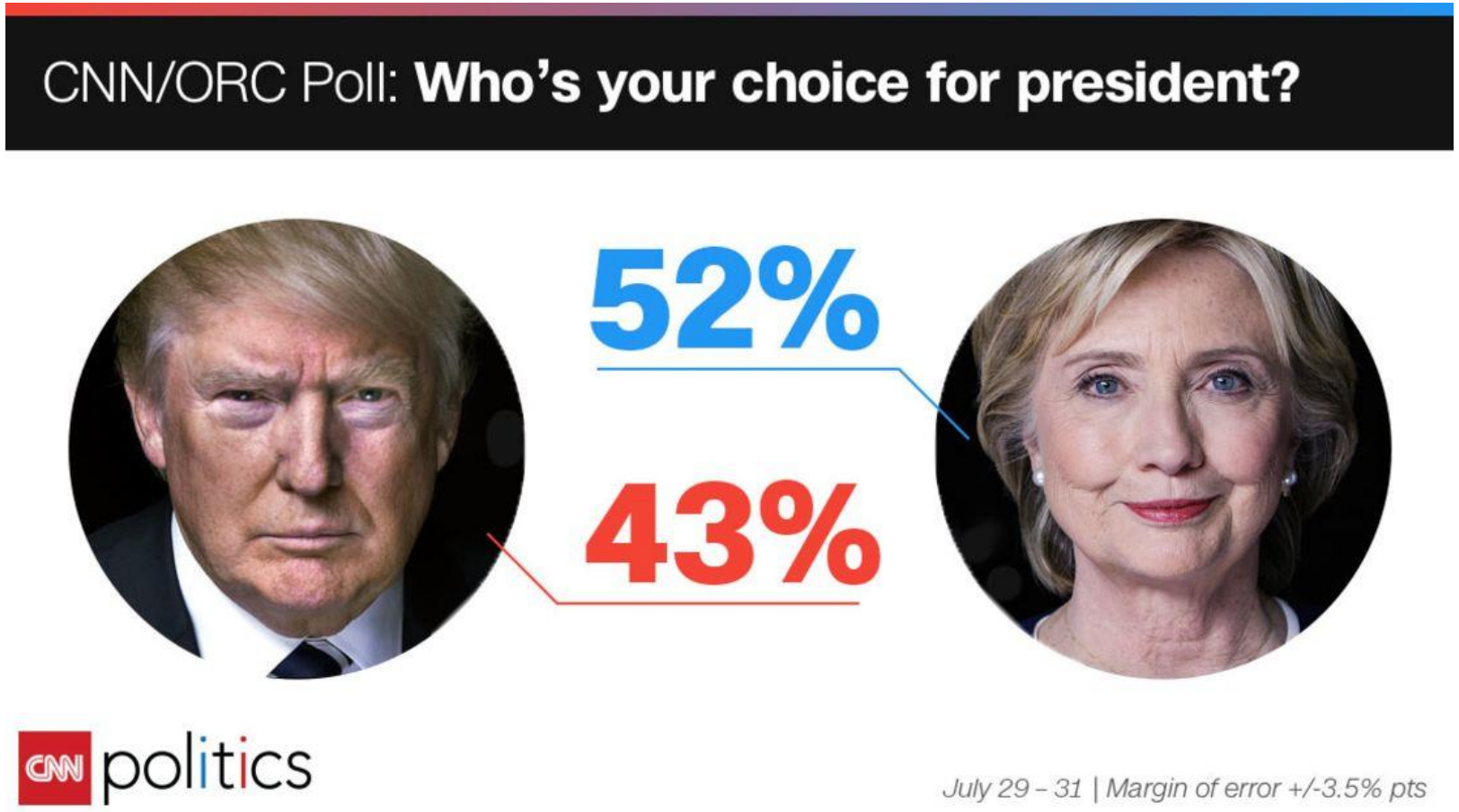
Survivorship Bias



Nonresponse Bias



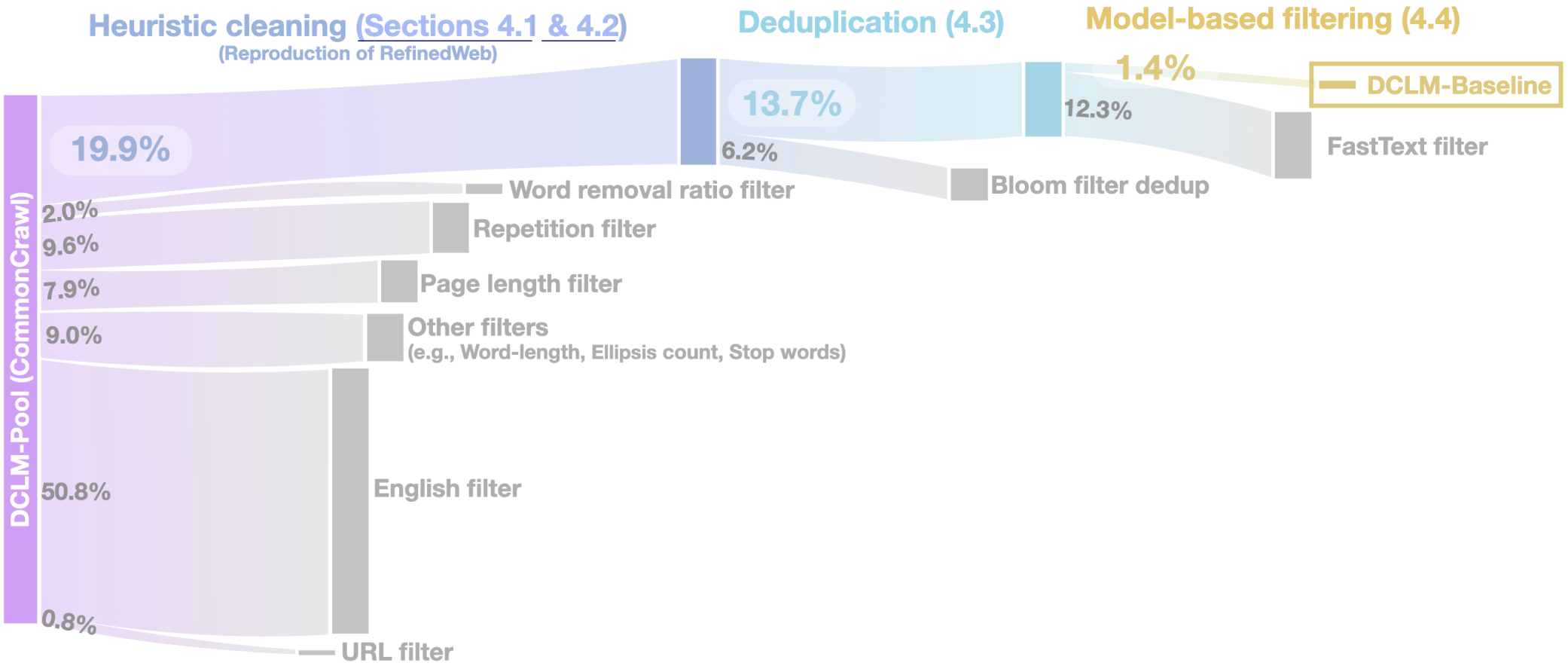
Case Study: Shy Trump



Data Curation

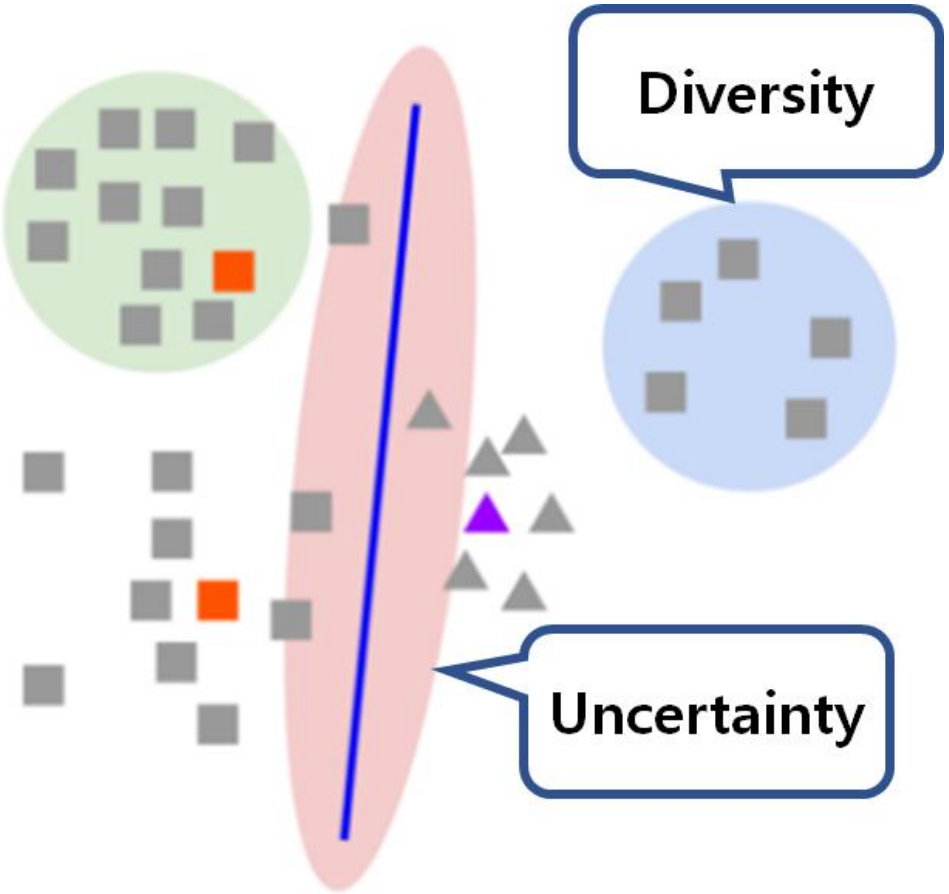
Aspect	Data Sampling	Data Curation
Goal	Reduce computation, handle imbalance, ensure fairness	Improve data quality, consistency, and representativeness
Scope	Select a subset of the data	Refine and improve the entire dataset
Focus	Efficiency, smaller size, representative subset	Quality, diversity, removing noise or bias
Techniques	Random, Stratified, Cluster, Systematic sampling	Filtering, Cleaning, Balancing, Augmentation, Deduplication
Application	Model training with less cost, quick experiments	Reliable training/testing, trustworthy downstream tasks
In short	Choosing **which data to keep**	Improving **how good the data is**

Data Curation

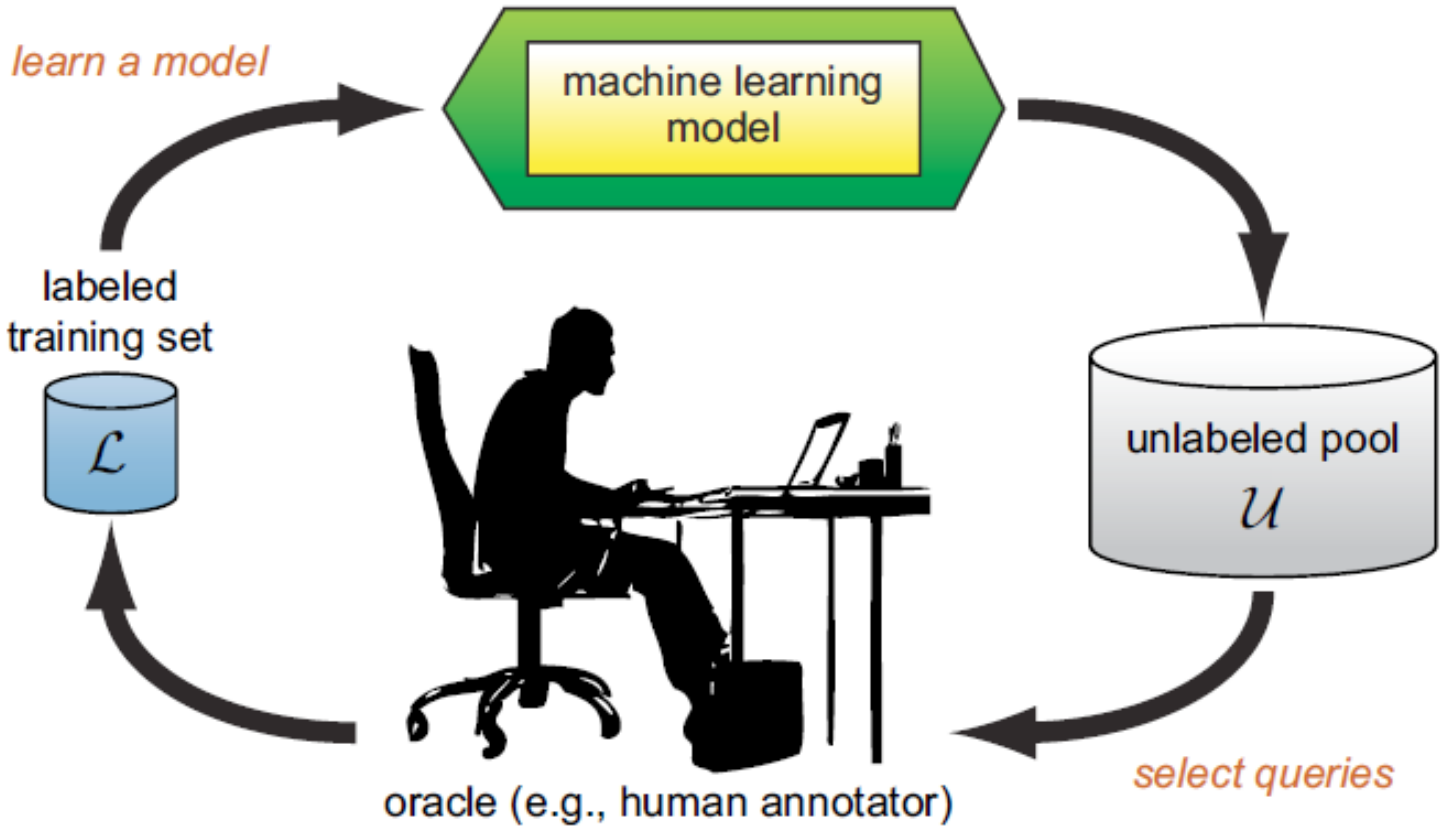


DataComp-LM: In search of the next generation of training sets for language models. arXiv, 2024.

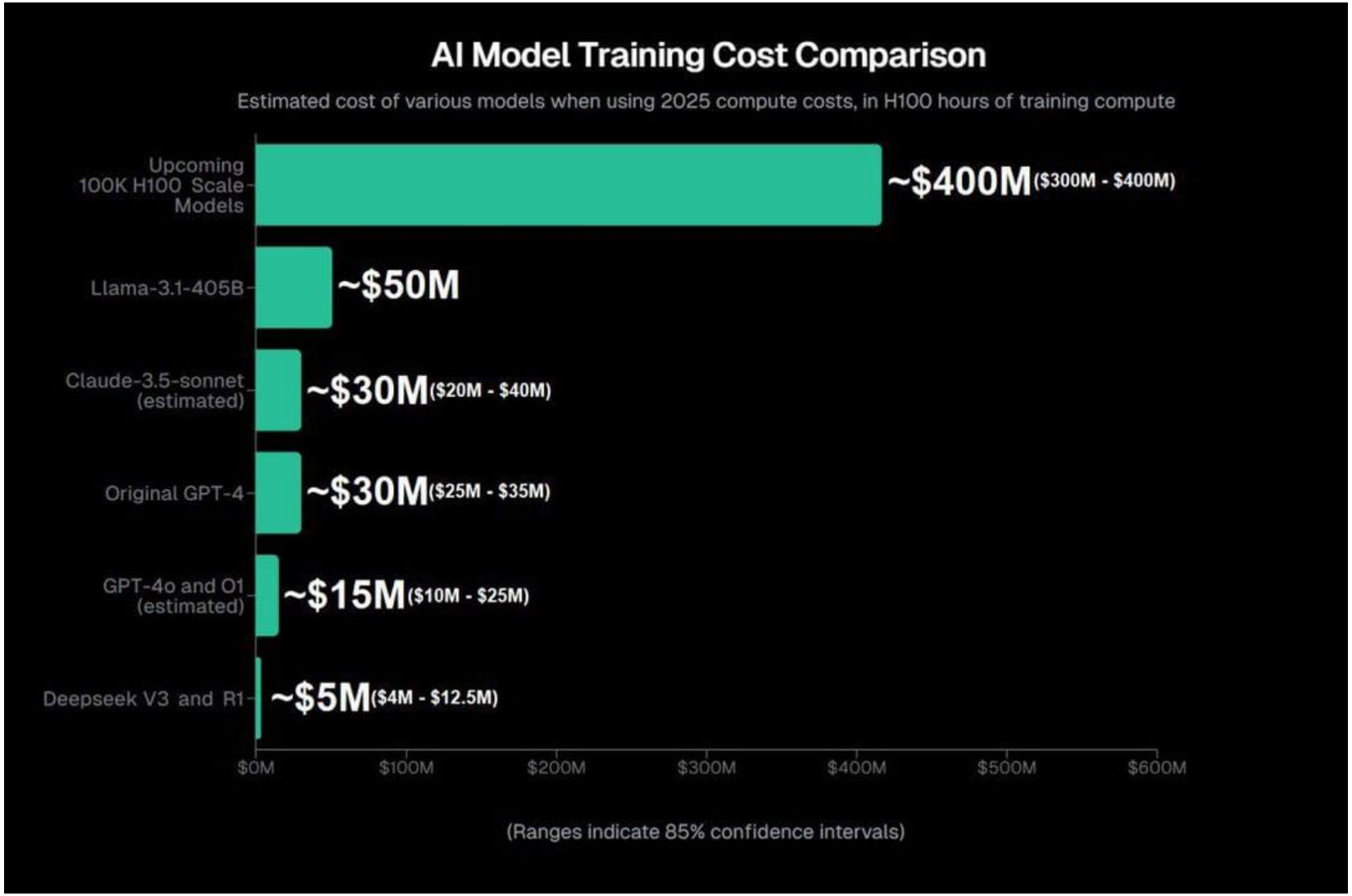
Quality & Diversity



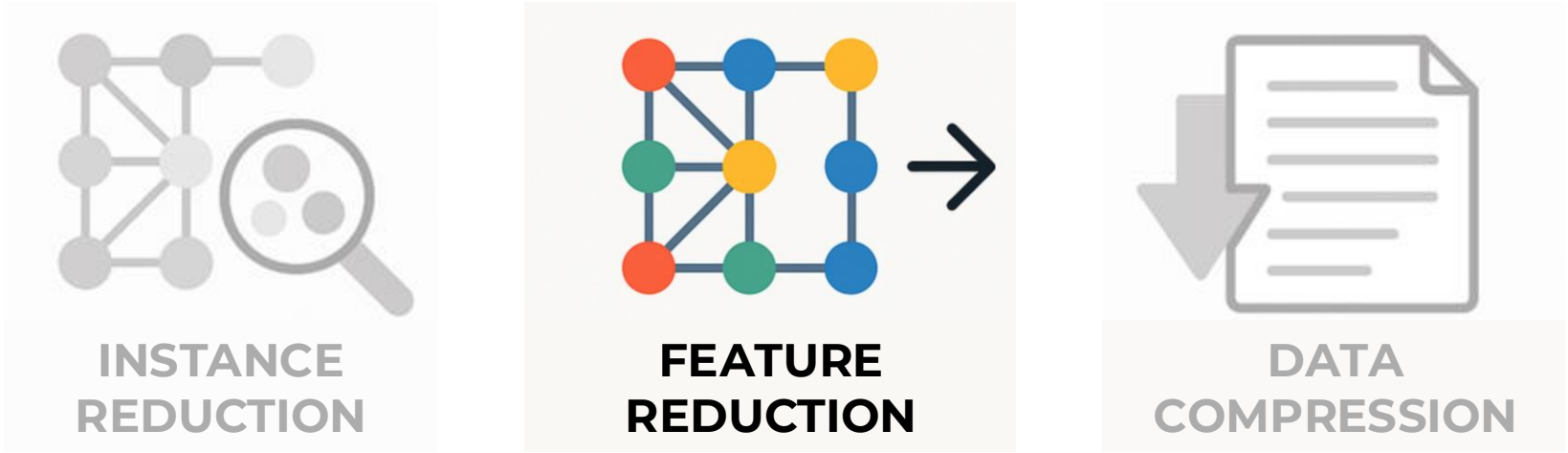
Active Learning



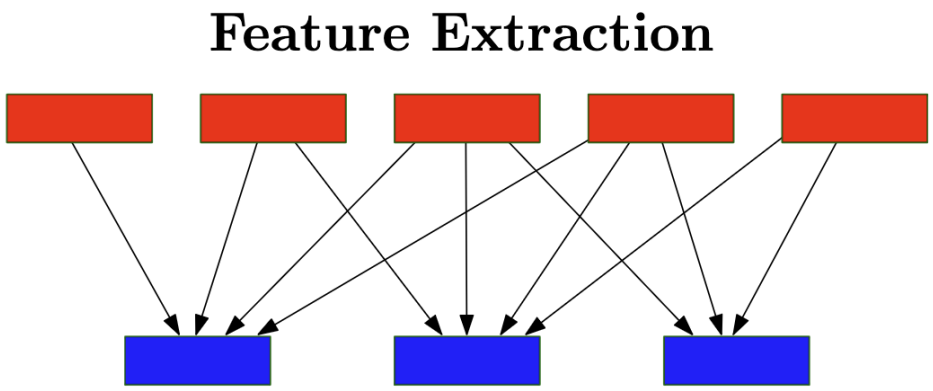
Case Study: DeepSeek

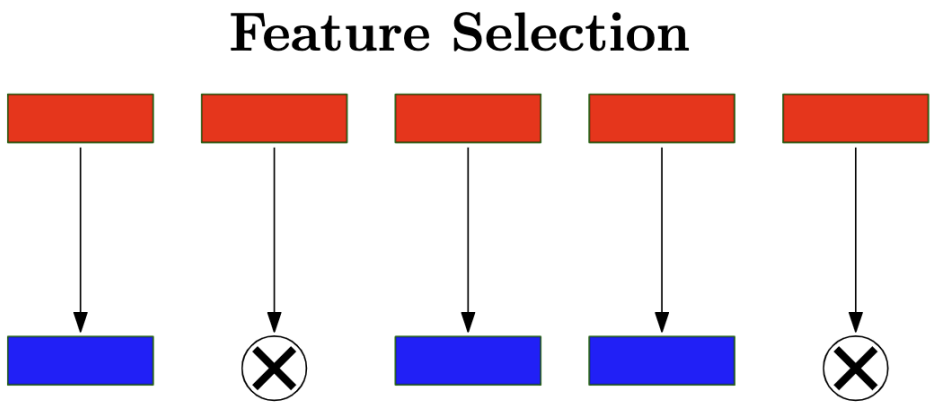


Feature Reduction

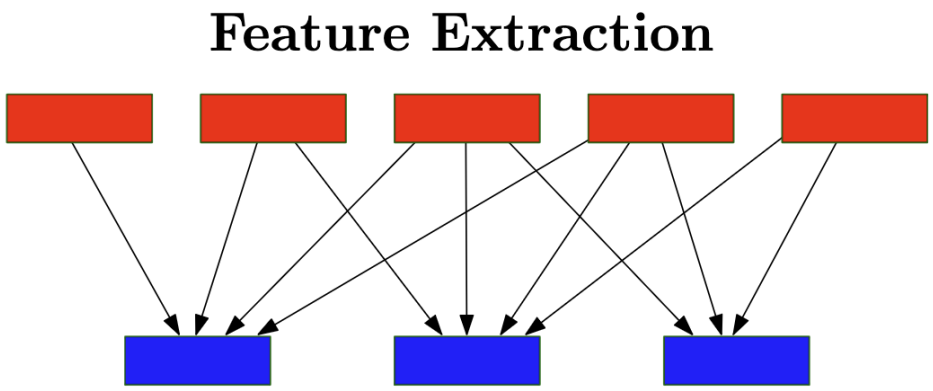


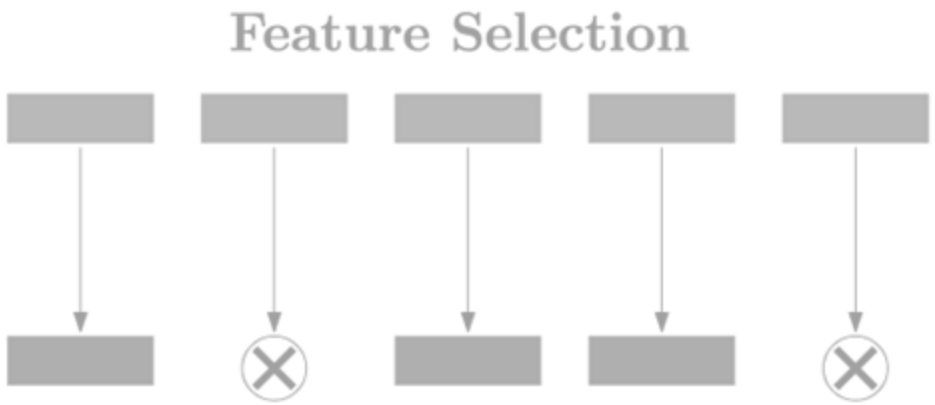
Extraction vs. Selection



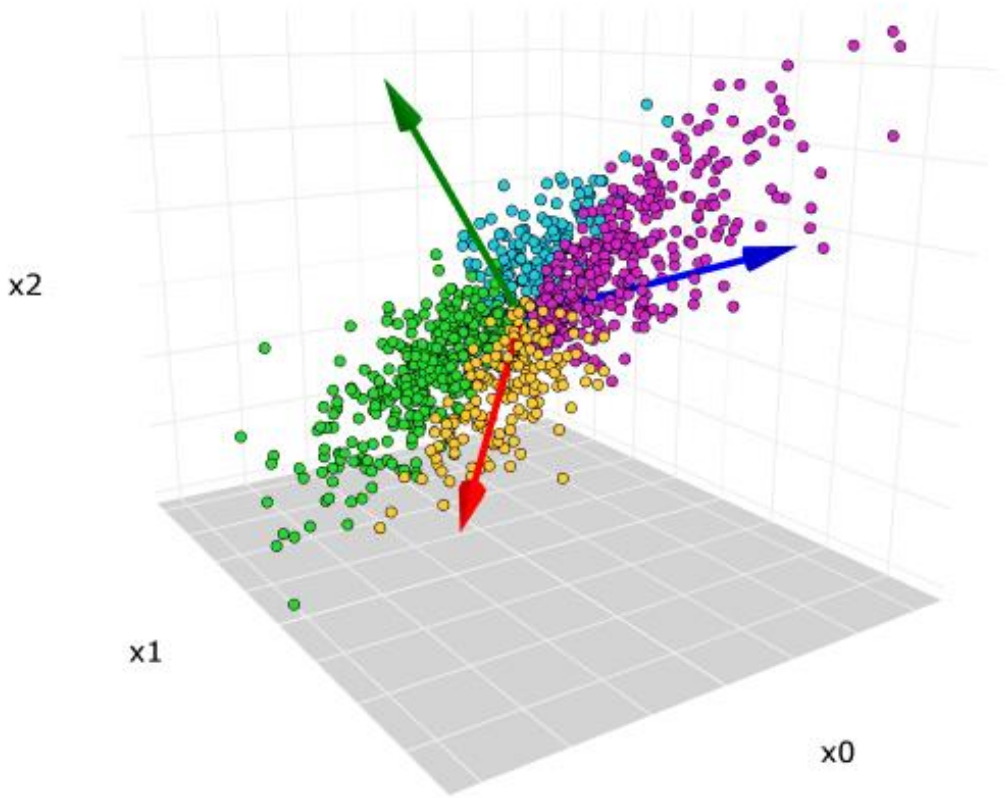


Feature Extraction

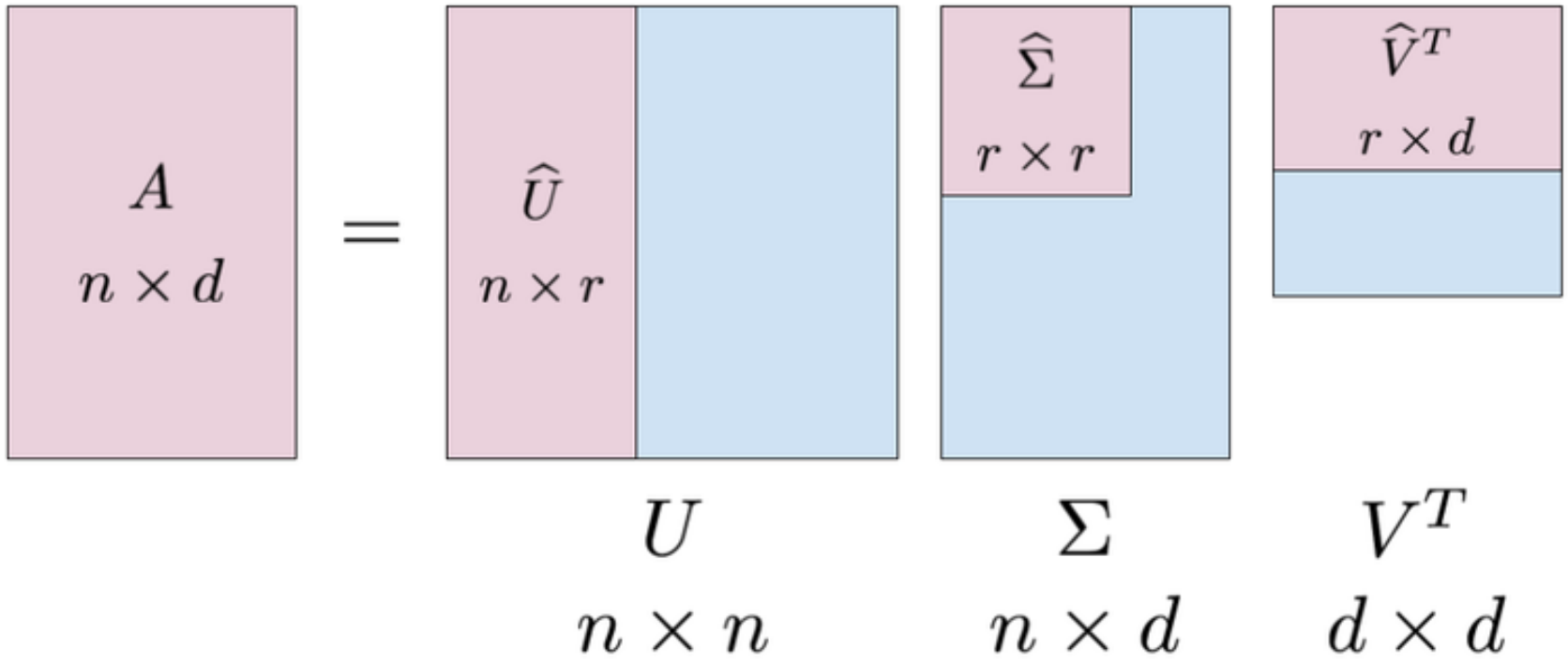




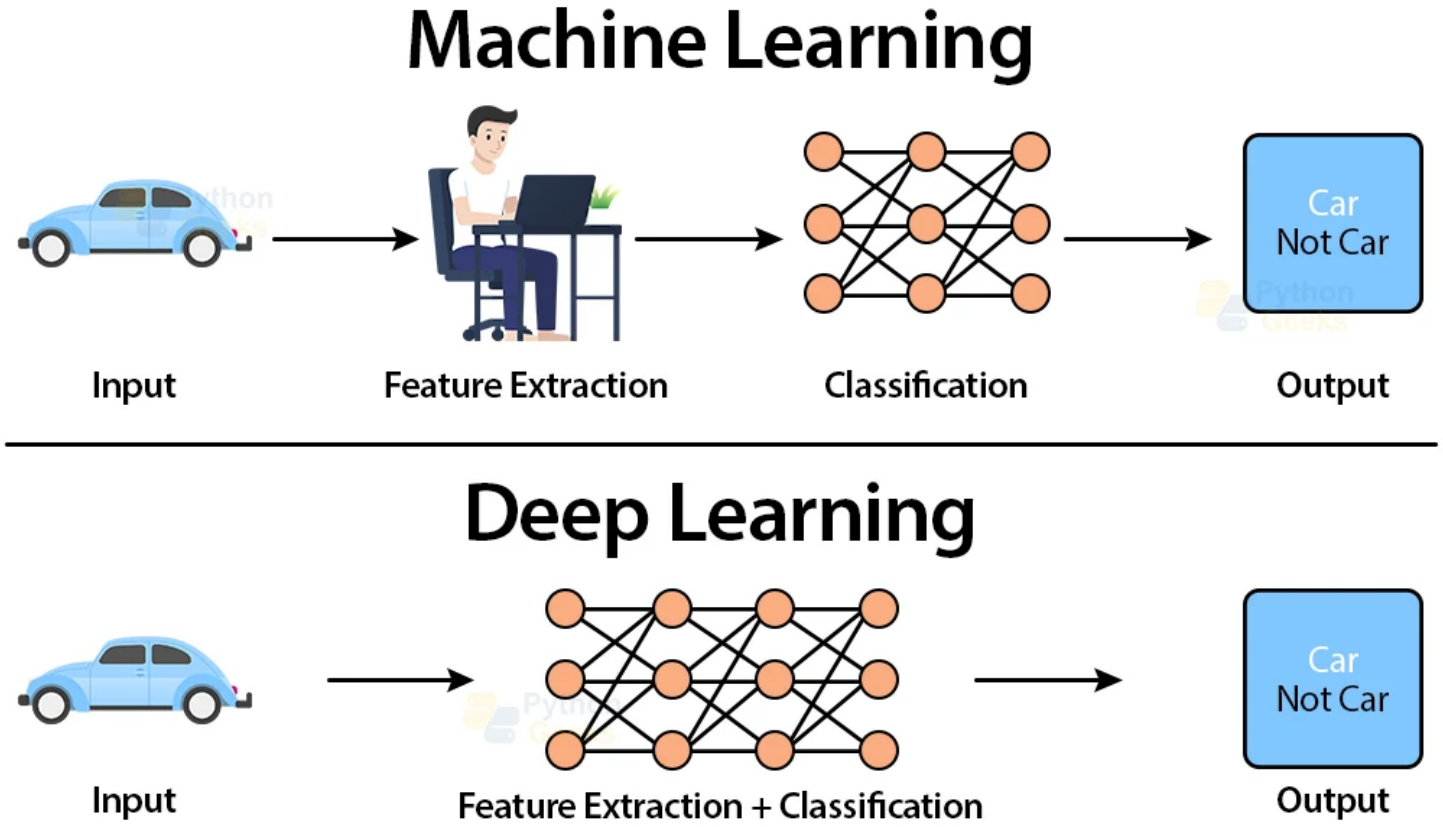
Recap: Principal Component Analysis (PCA)



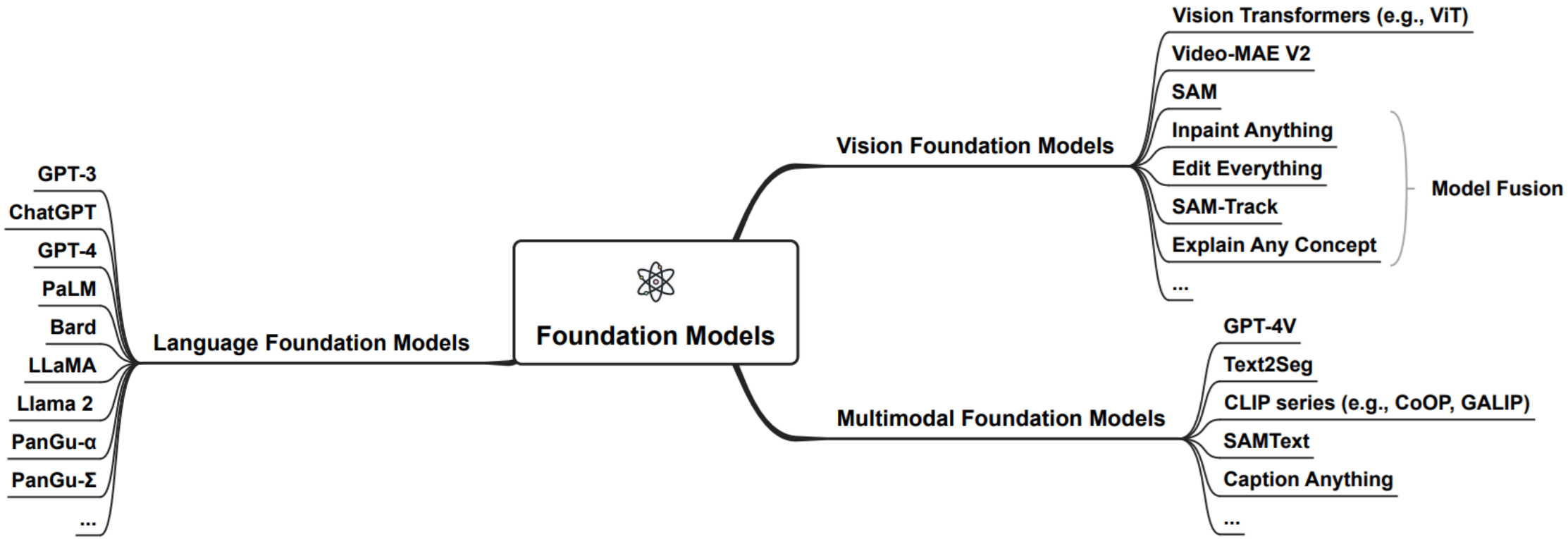
Recap: Principal Component Analysis (PCA)



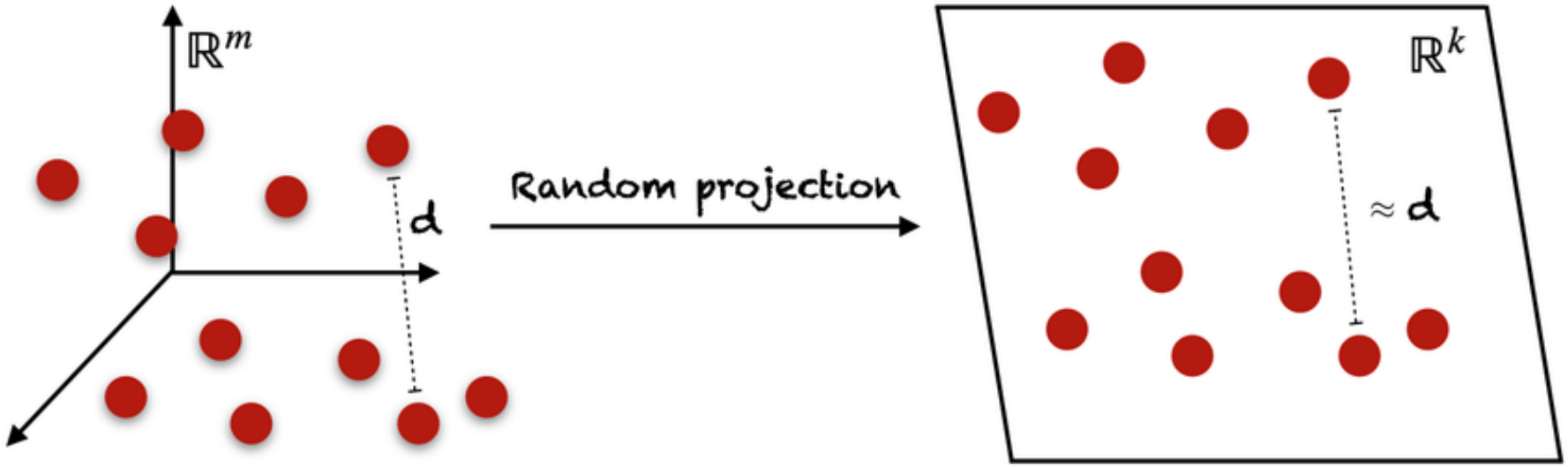
Recap: Deep Representation Learning



Recap: Foundation Models

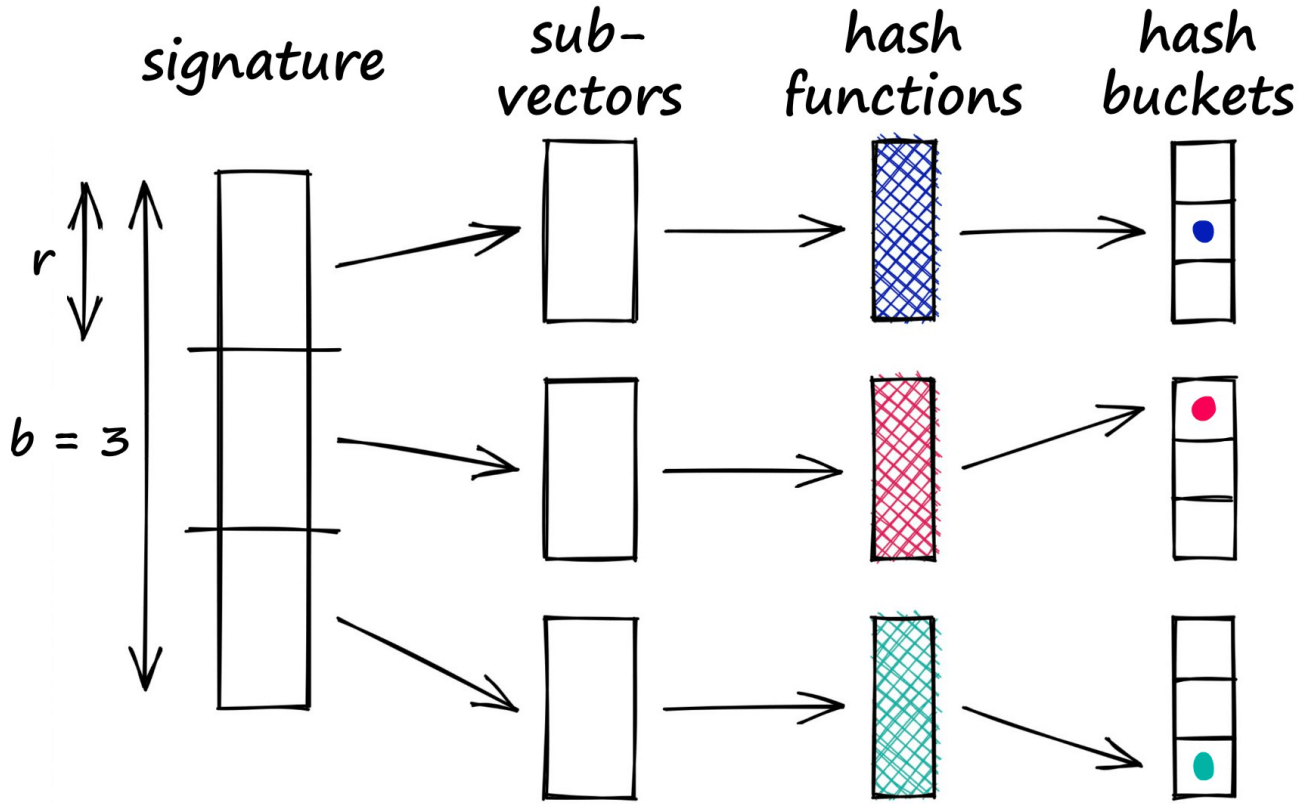


Johnson-Lindenstrauss Transform

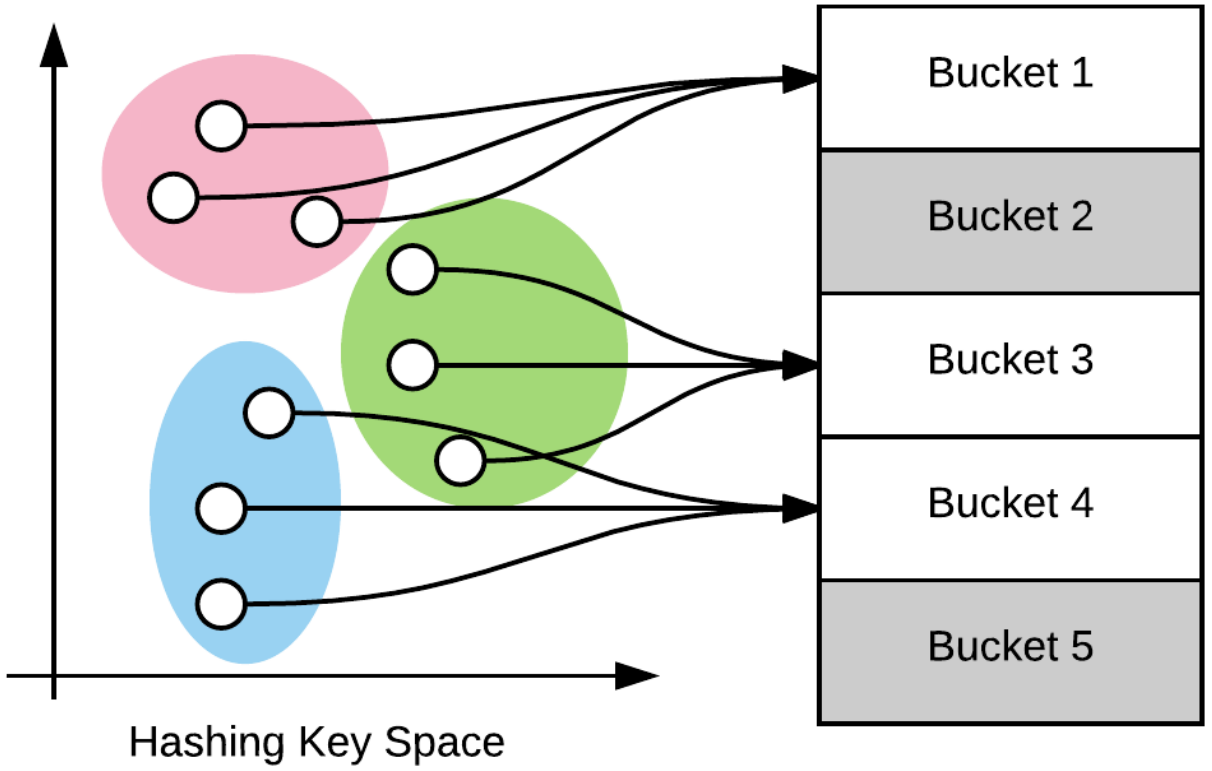


Random projection that preserves distances

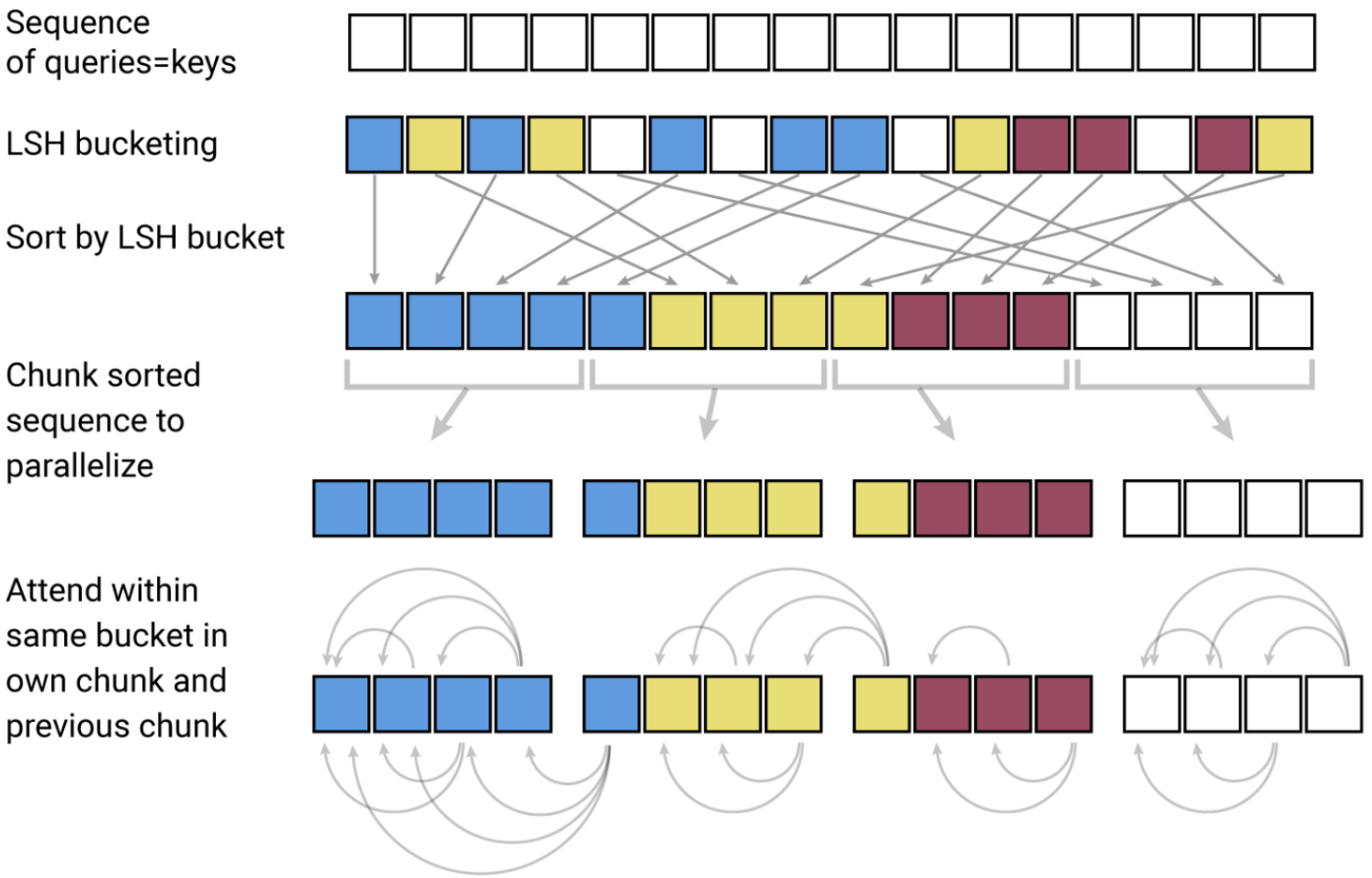
Hashing (Hash Code as a Feature)



Locally Sensitive Hashing (LSH)

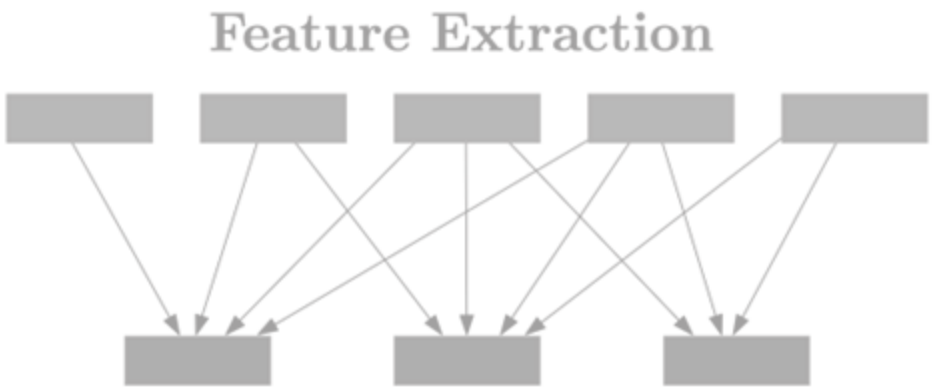


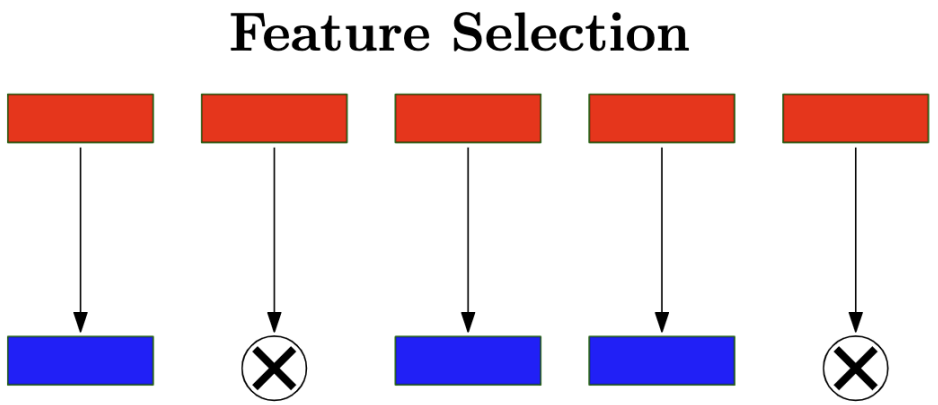
Case Study: Efficient Search via LSH



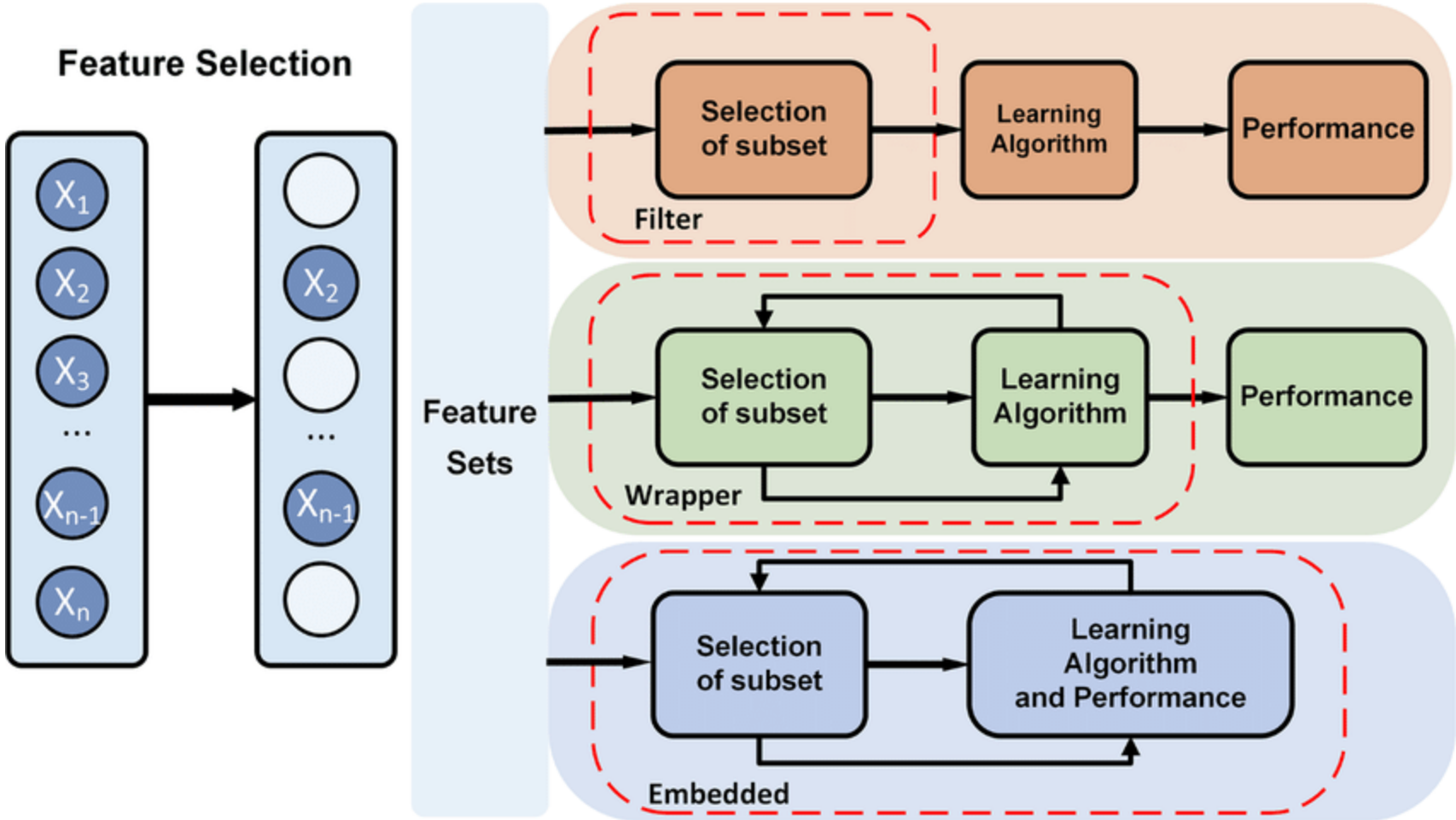
Reformer: The Efficient Transformer. ICLR, 2020.

Feature Selection

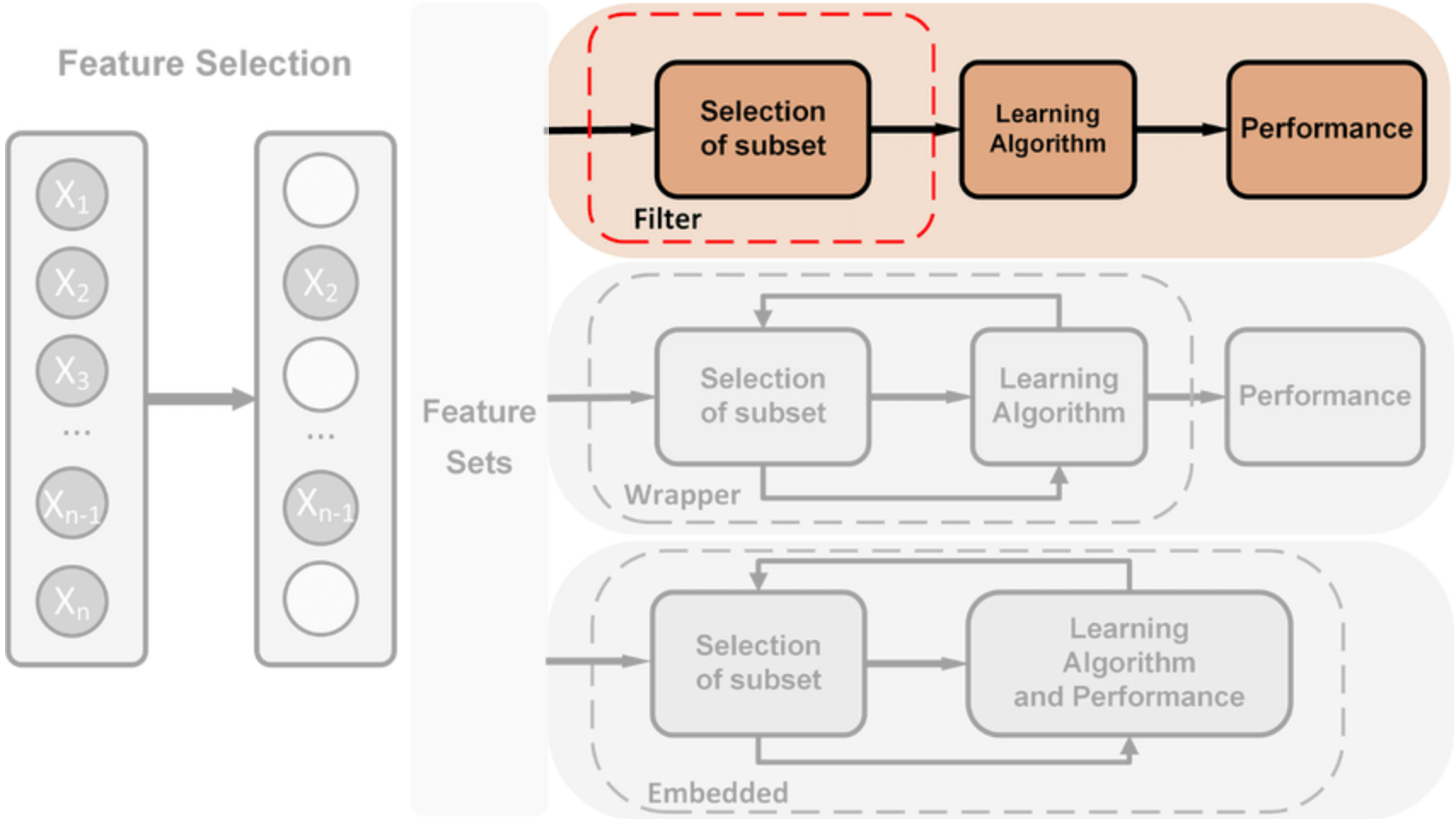




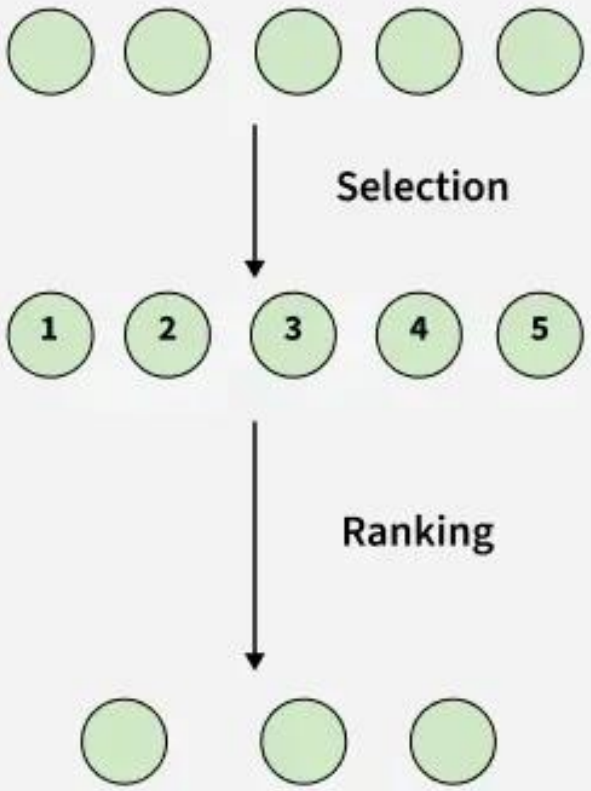
Feature Selection



Filter Methods



Filter Methods

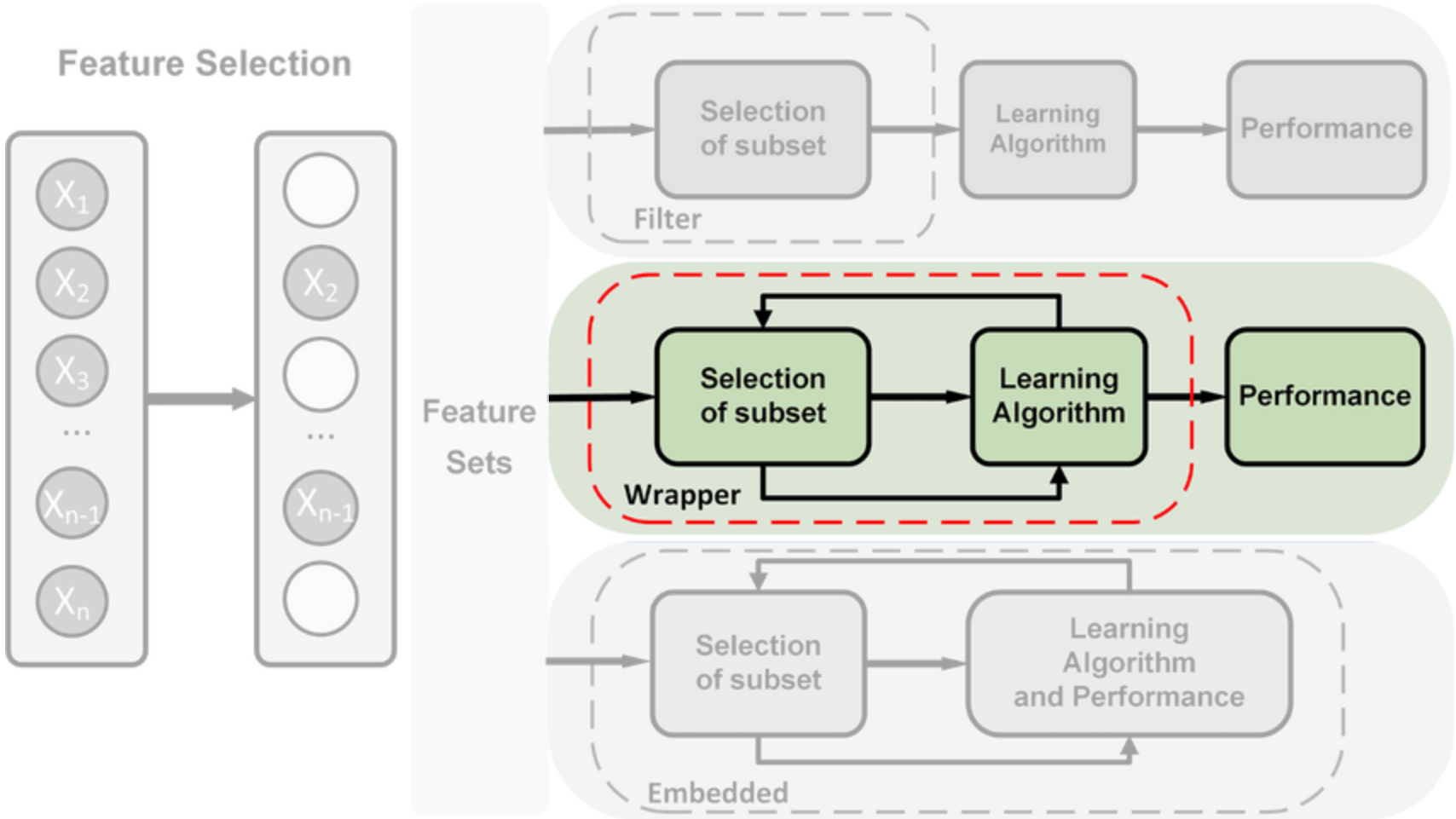


Filter Method

(e.g., variance threshold)

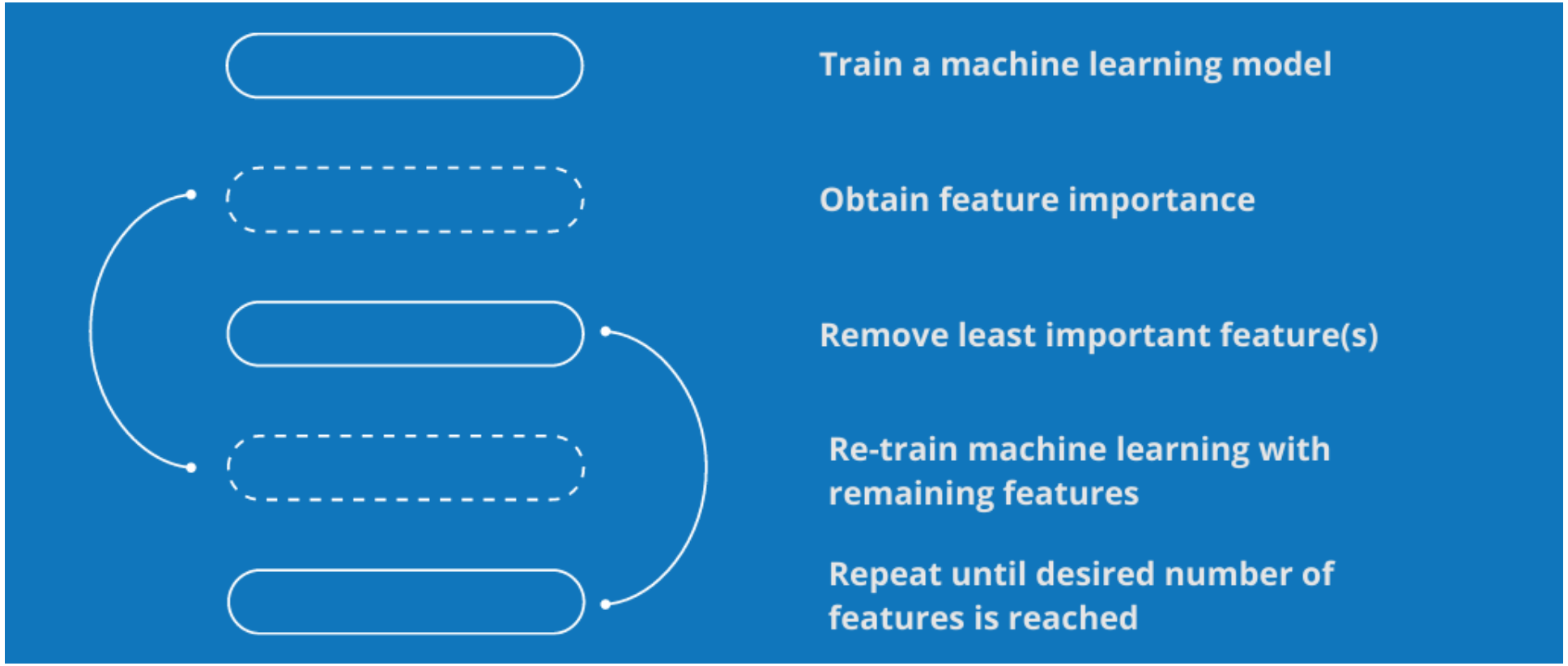
1. Rank features by statistical score
2. Select top-ranked features
3. Feed into model

Wrapper Methods

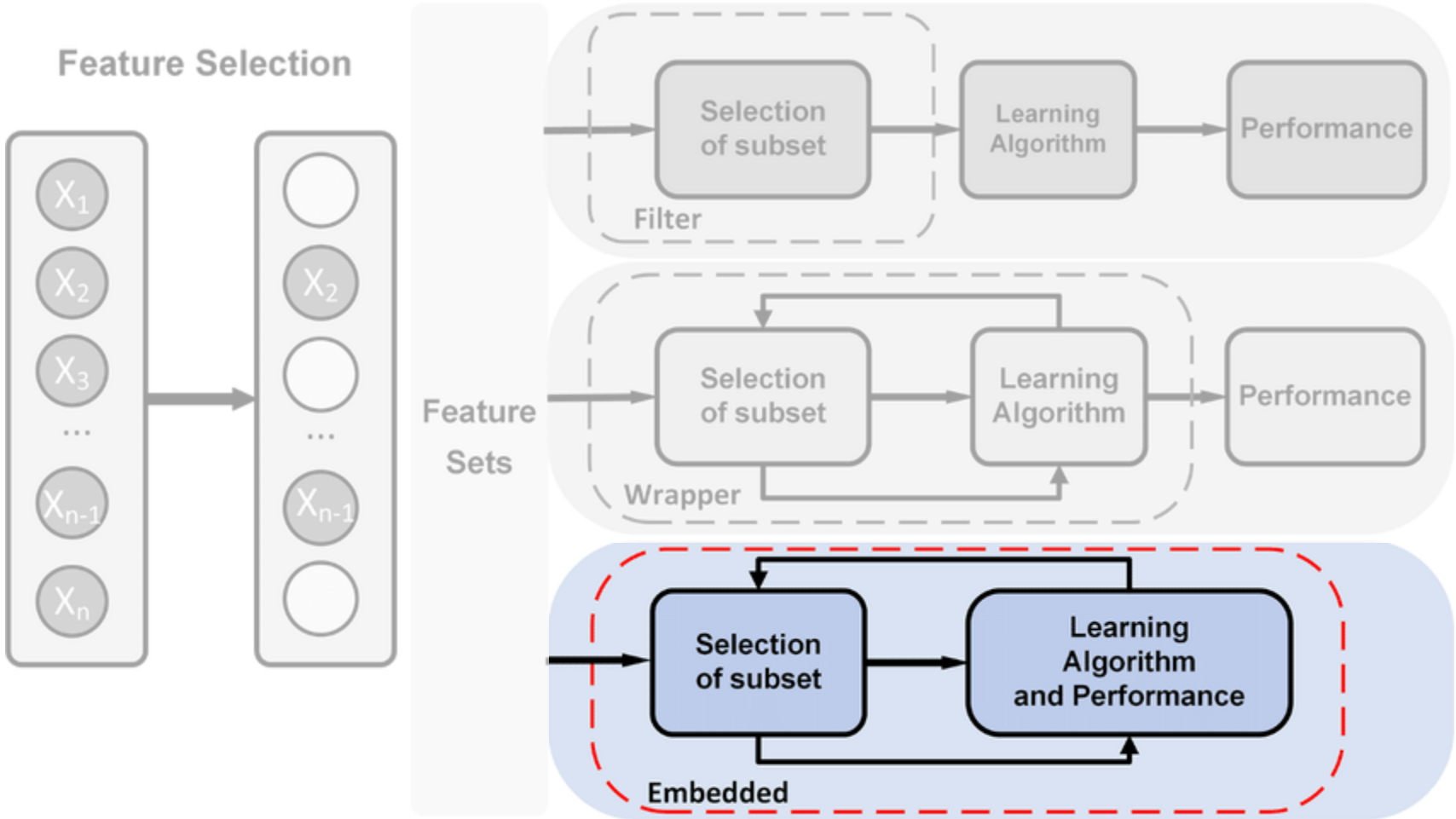


Wrapper Methods

- Forward selection
- Backward elimination
- Recursive Feature Elimination (RFE)

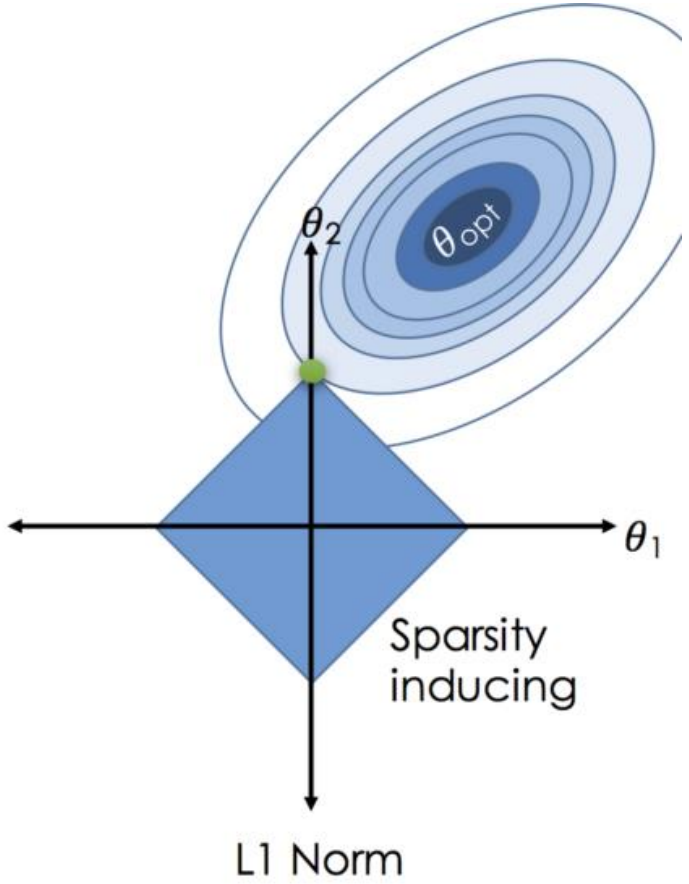


Embedded Methods



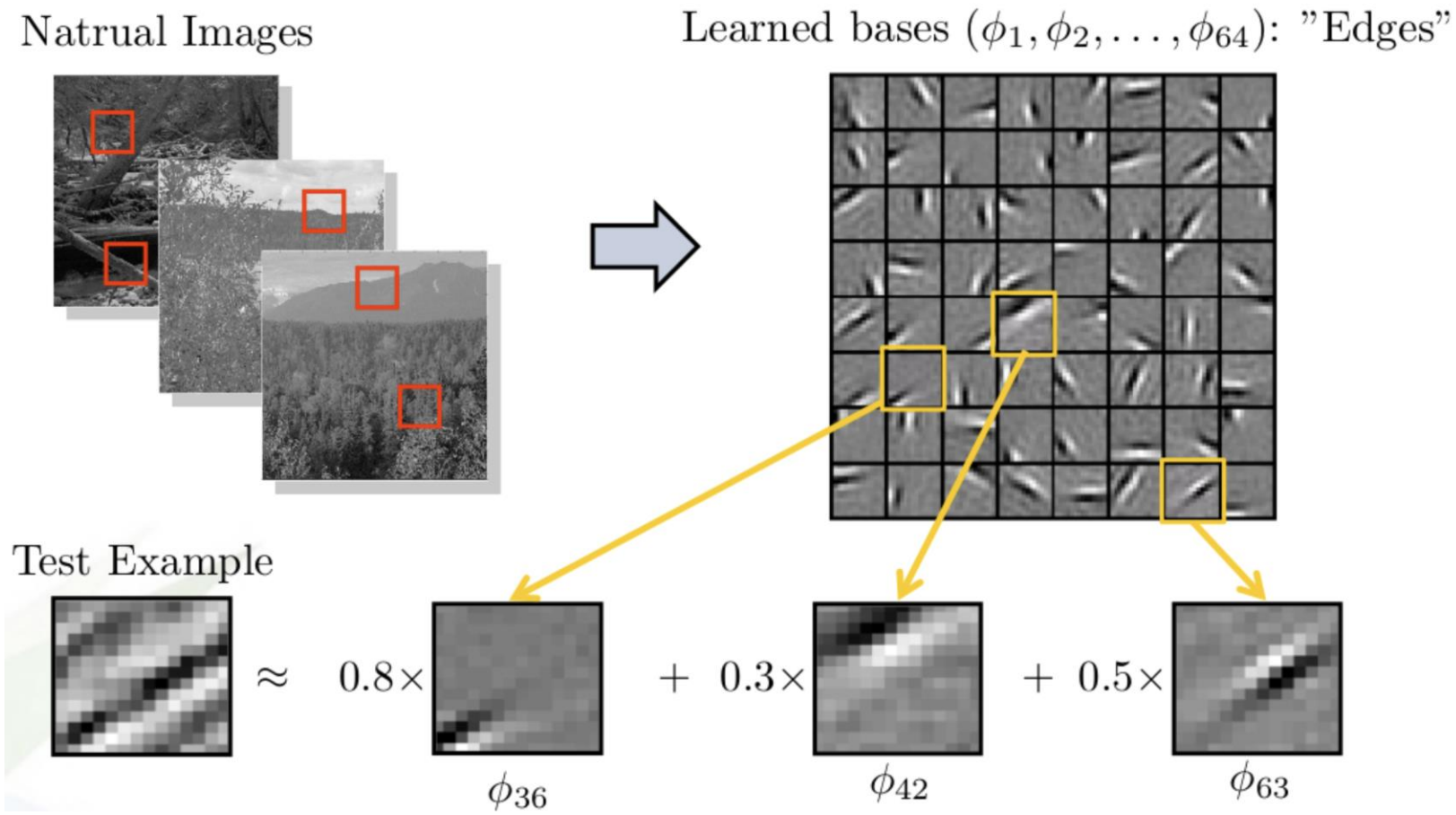
Embedded Methods

Sparsity Regularization (e.g., LASSO)

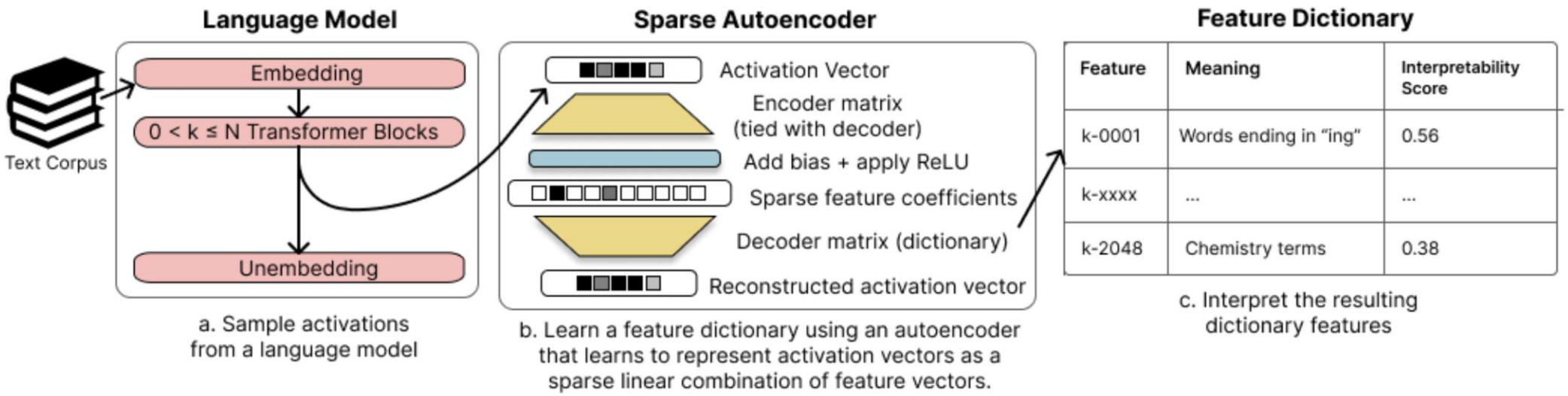


Embedded Methods

Sparse Coding (autoencoder w/ sparsity)



Sparsity → Interpretability

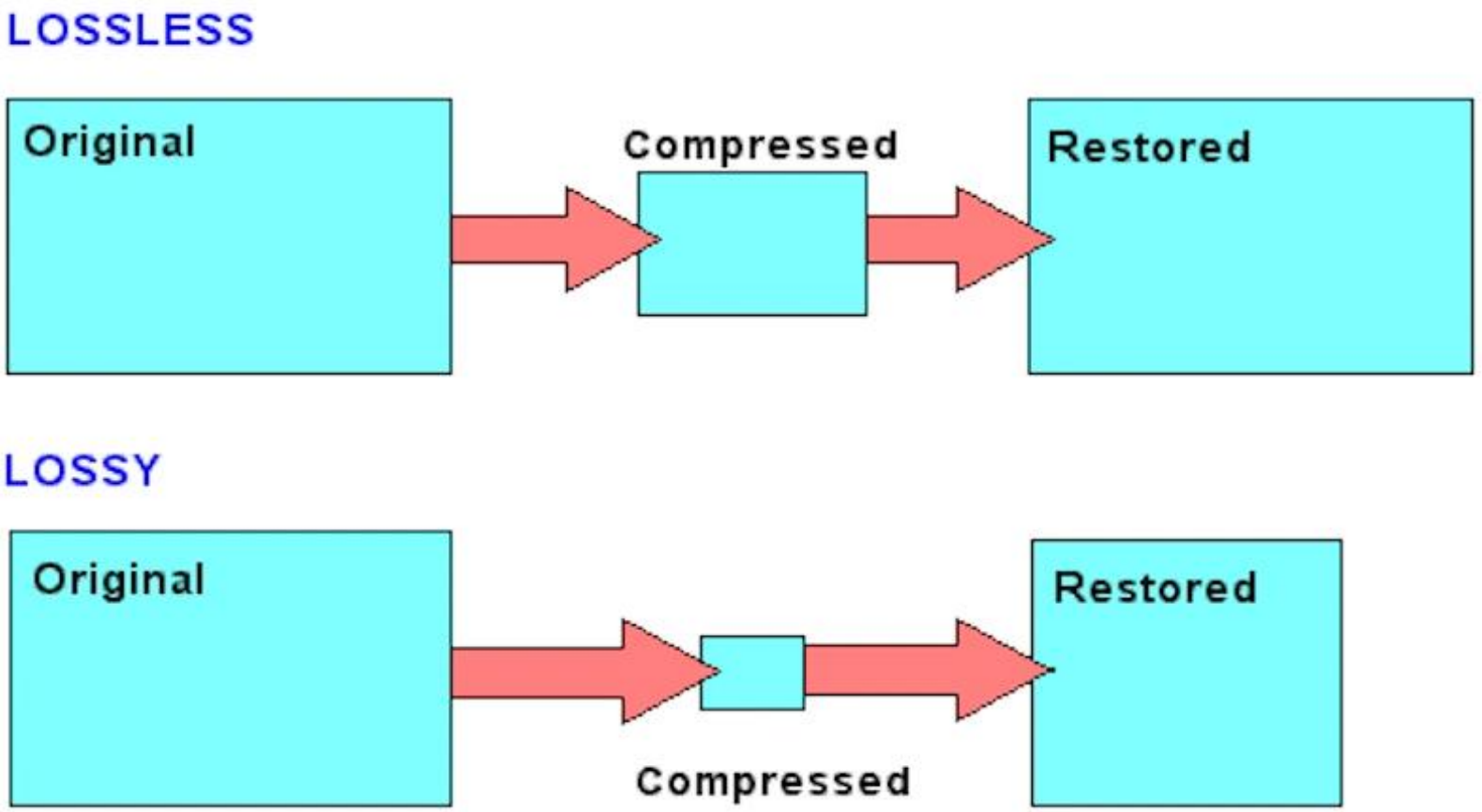


Sparse Autoencoders Find Highly Interpretable Features in Language Models. ICLR, 2024.

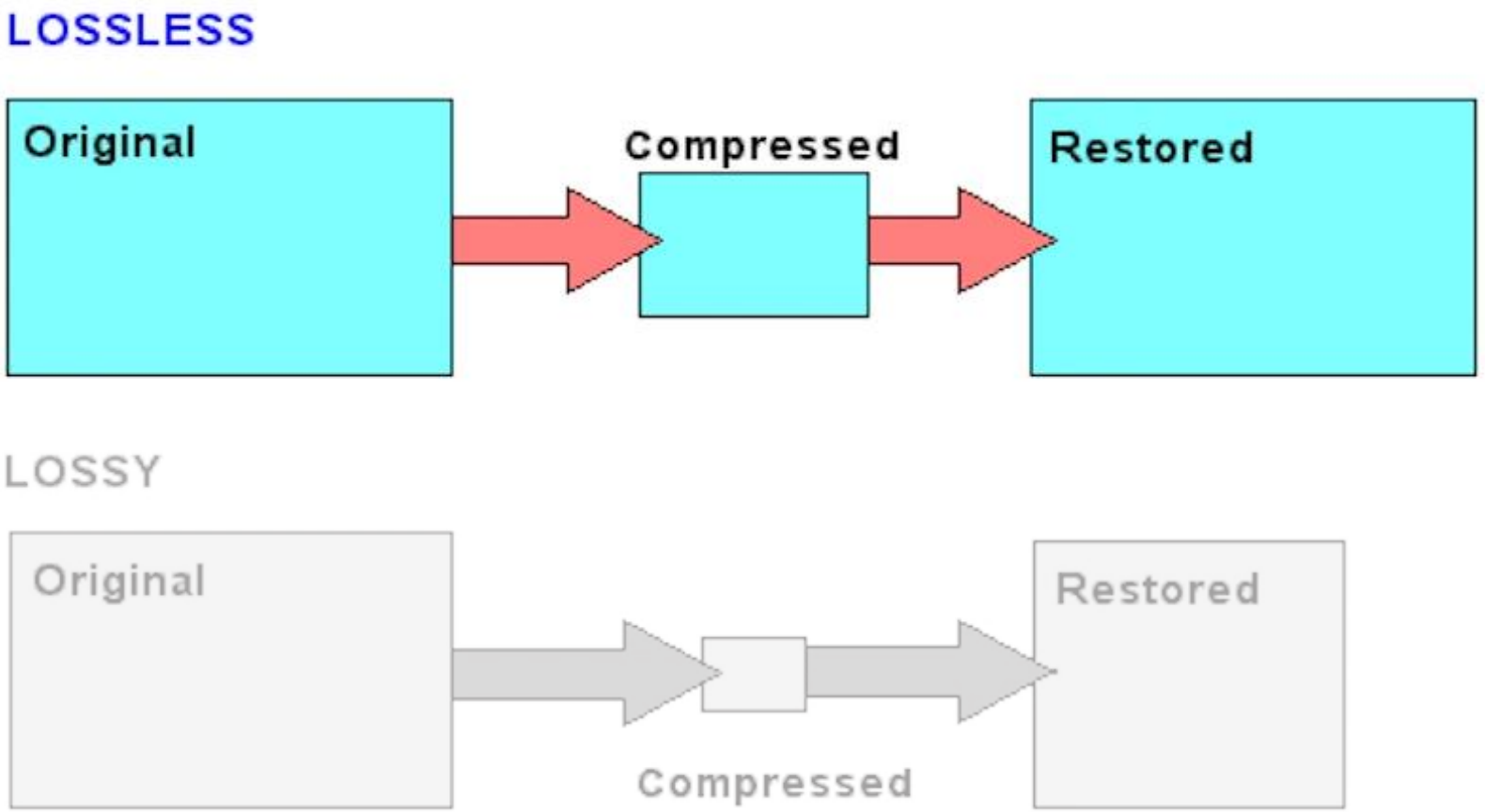
Data Compression



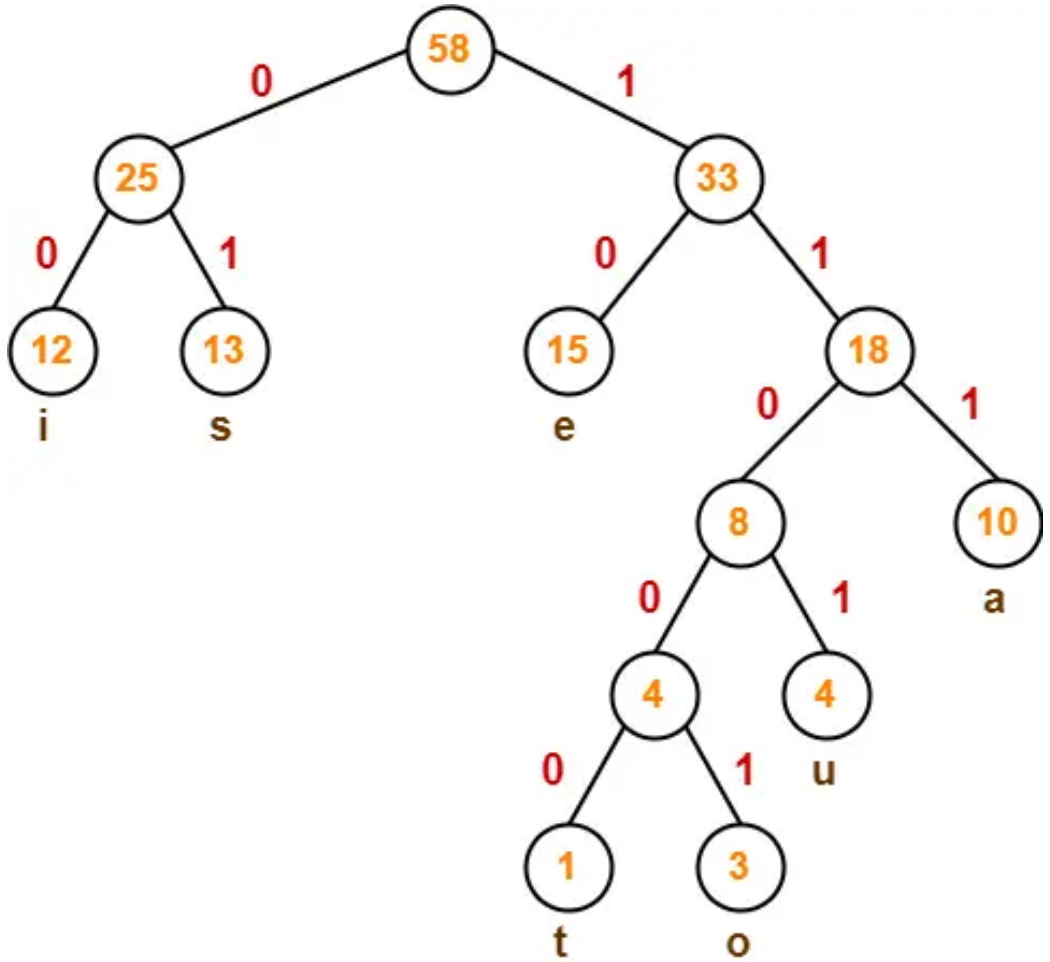
Lossless vs. Lossy



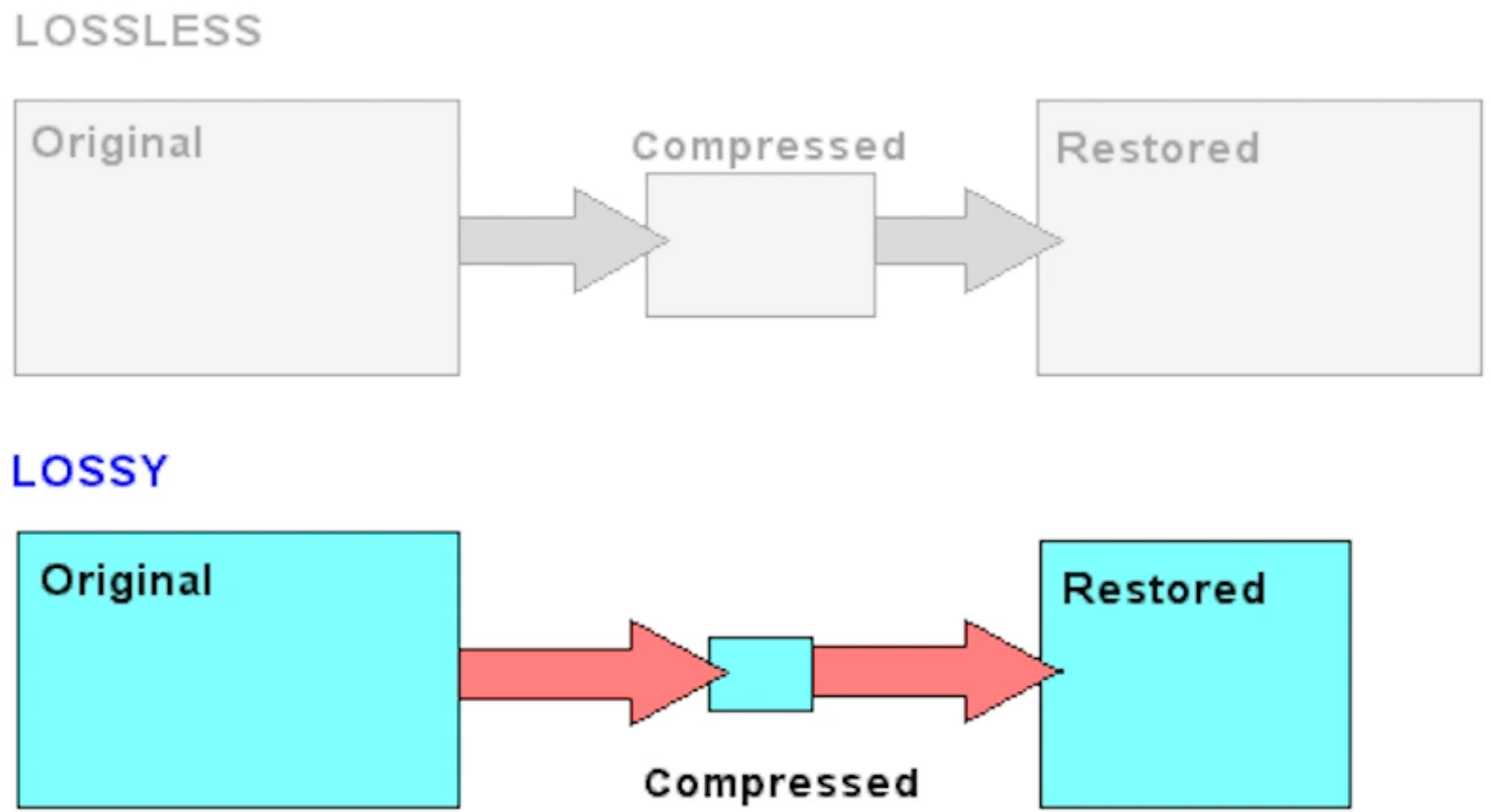
Lossless Compression



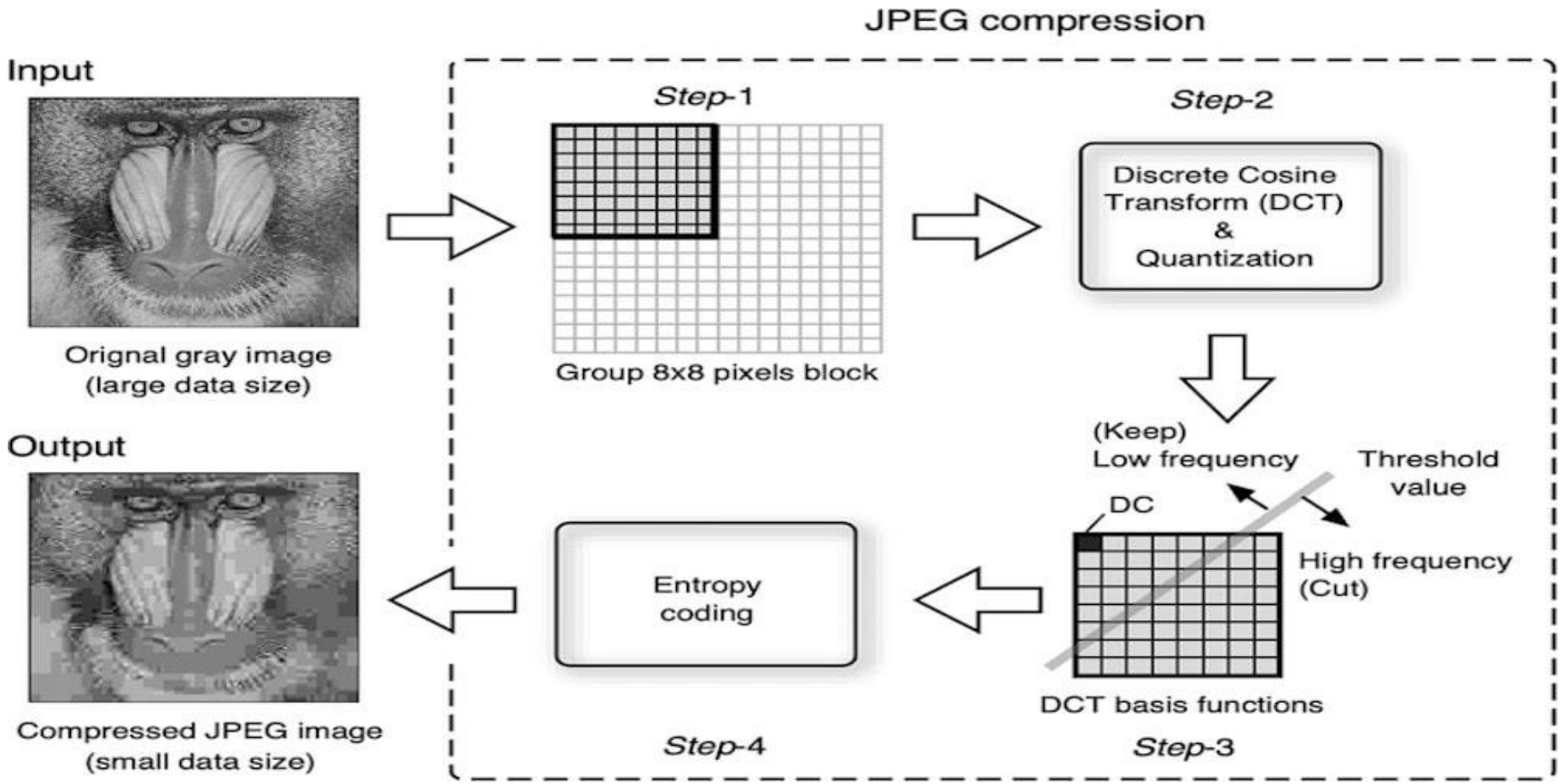
Huffman Coding



Lossy Compression



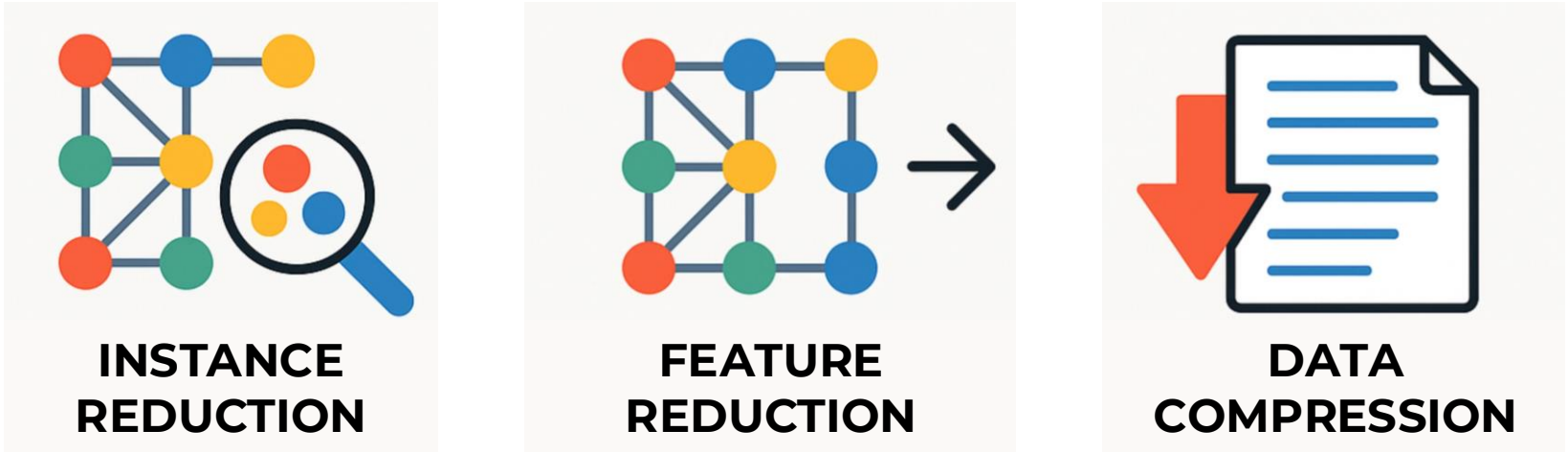
Lossy Compression



Lossless vs. Lossy: When to Use?

Category	Lossless Compression	Lossy Compression
Restoration	100% original data can be recovered	Approximate recovery (some info discarded)
Typical Data	**Text, code, executables, logs**	**Images, audio, video**
Examples	ZIP, PNG, GZIP, Parquet (Snappy, Zstd)	JPEG, MP3, MP4 (H.264), WebP
Advantages	Guarantees data integrity	Much higher compression ratio, efficient storage/transfer
Limitations	Limited compression ratio	Not identical to original (information loss)

Summary: Data Reduction



Summary: Data Pre-processing

