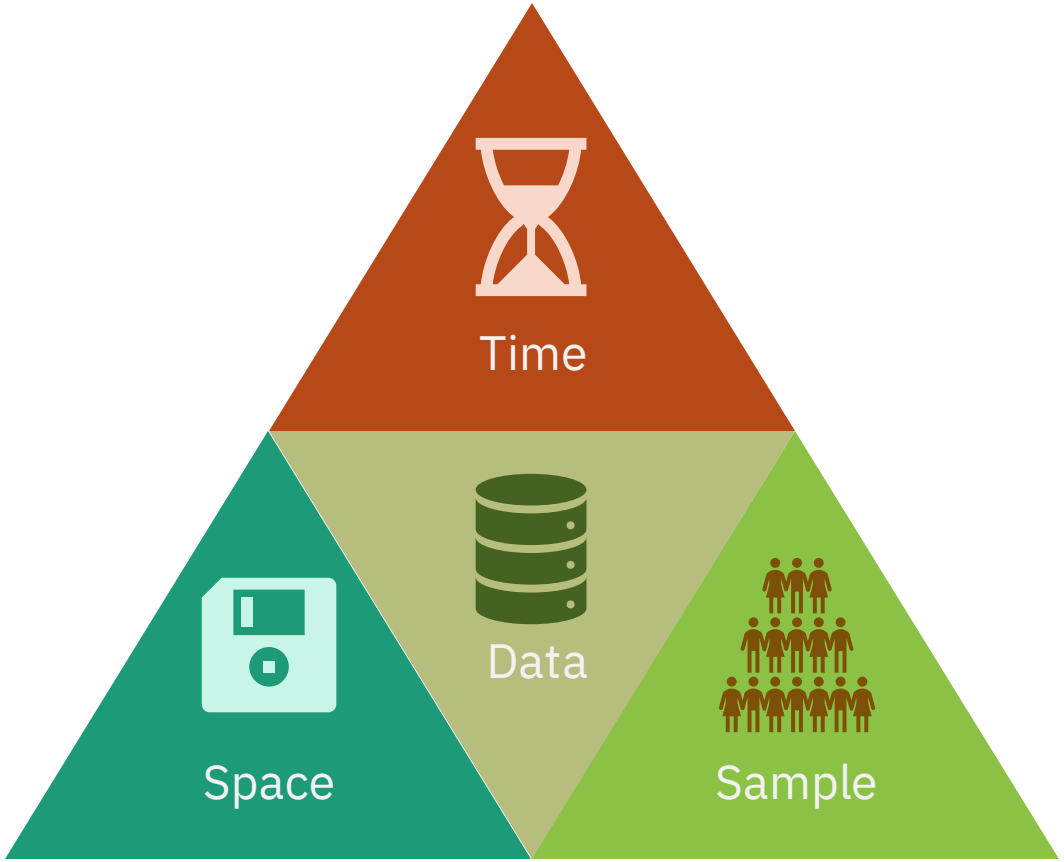
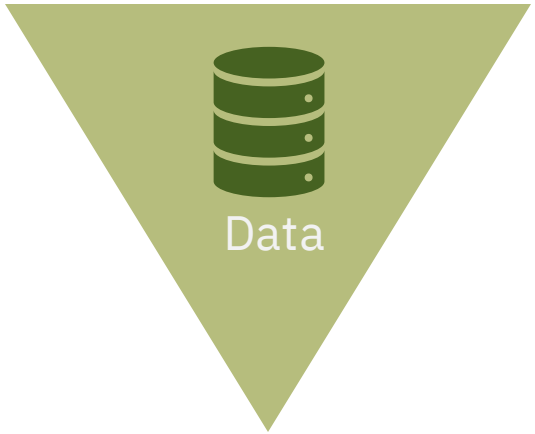
Characterizing Data Complexity in Machine Learning

Complexity in Machine Learning



Why data complexity is important?



- Impacts model selection
- Explains learning difficulty
- Generalize to unseen data
- Helps in meta-learning and benchmarking

Why does my favorite model perform better on dataset X but not on dataset Y ?

Data complexity types

Intrinsic

Inherent structure of the data which makes it difficult to learn independent of the algorithm.

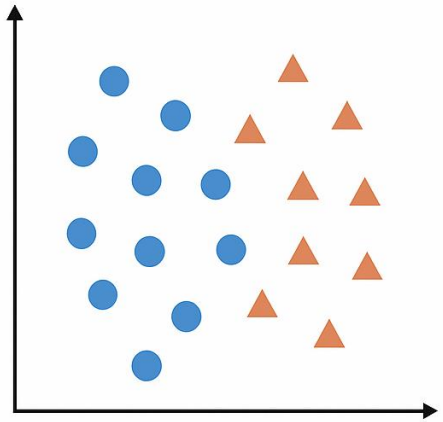
- Class distribution
- Non-linear decision boundaries
- Higher-order correlations
- Noise

Extrinsic

Complexity from externa factors dependent on the algorithm or preprocessing.

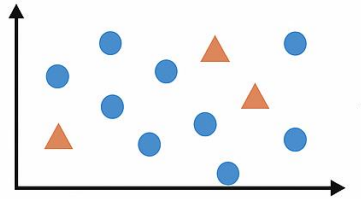
- Preprocessing issues
- Misalignment between model and data
- Learning limitations of models

Intrinsic Complexity



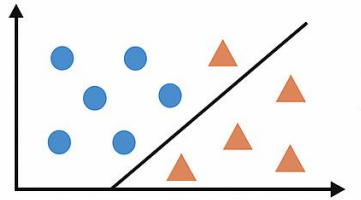
Overlapping class distributions in feature space

Extrinsic Complexity



Poor feature transformation

Non-linear data



Good feature transformation

Linearly separable data

Data complexity measures

Intrinsic

Dimensional

- Intrinsic Dimension (Rank)
- Manifold (Fractal Dimension)
- Volume
- Effective rank
- Eigenspectra

Distributional

- Kurtosis & Skewness
- Mutual Information
- Sparsity
- Entropy
- Condition Number

Geometric

- Manifolds
- Clusters
- Density
- Topological Data Analysis
- Graph-based measures

Sampling

- Class imbalance ratio
- Class overlap measures
- Margin of separation between classes
- Sampling density variation

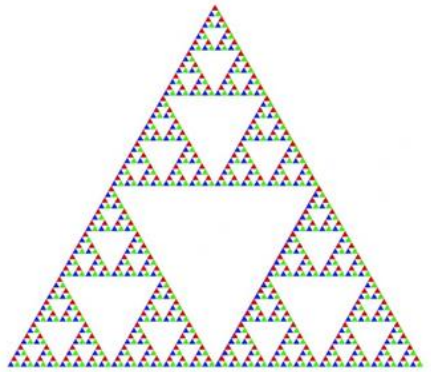
Data complexity

Dimensional

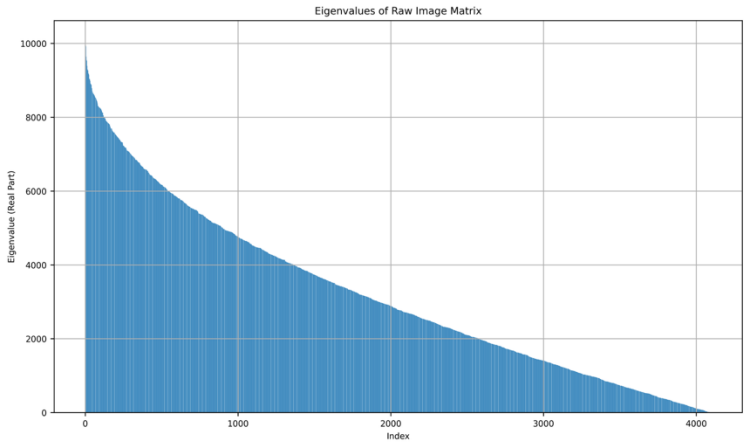
- Rank of data (k)

$$A \in R^{m \times n} \approx \tilde{A} \in R^{m \times k}$$

- Fractal Dimension



- Eigenspectrum



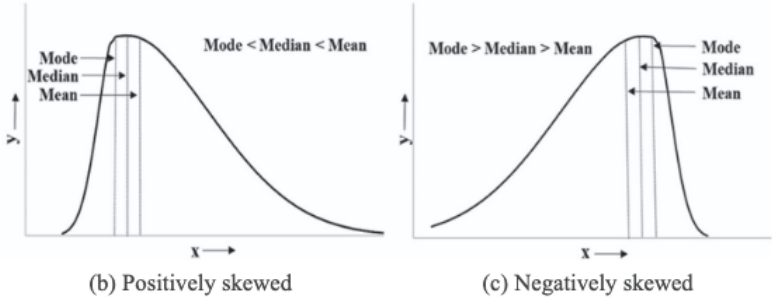
low high

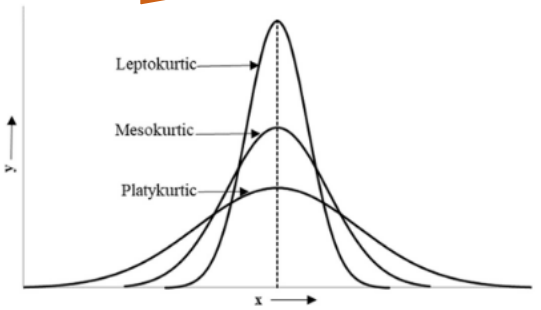
Desired value

Data complexity

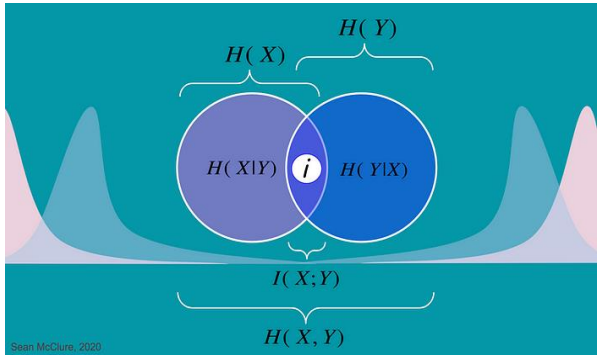
Distributional

- Skewness & Kurtosis





- Mutual Information



- Sparsity

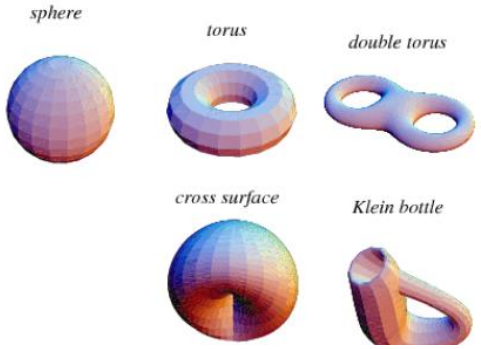
low high

Desired value

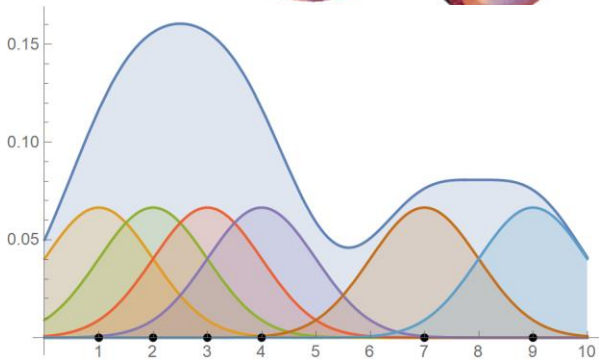
Data complexity

Geometric

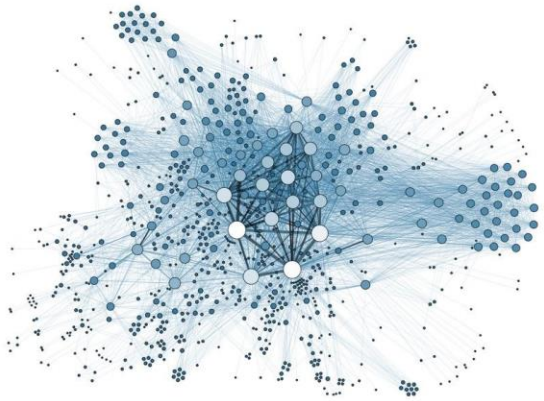
Manifold



Kernel Density



Networks



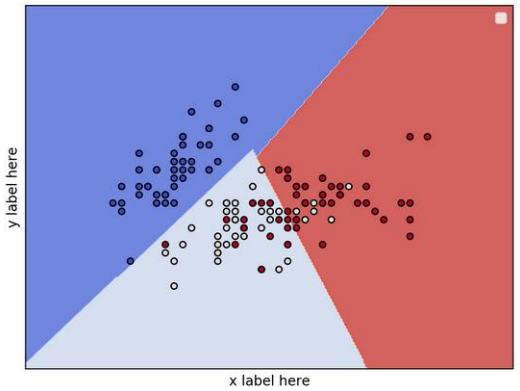
low high

Desired value

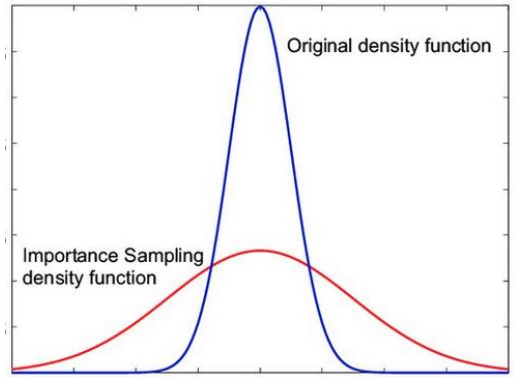
Data complexity

Sampling

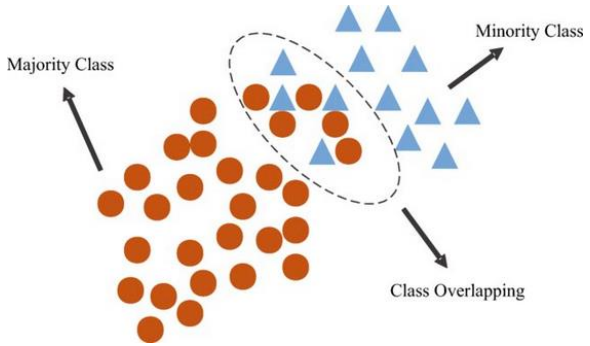
Class separation



Sampling variation



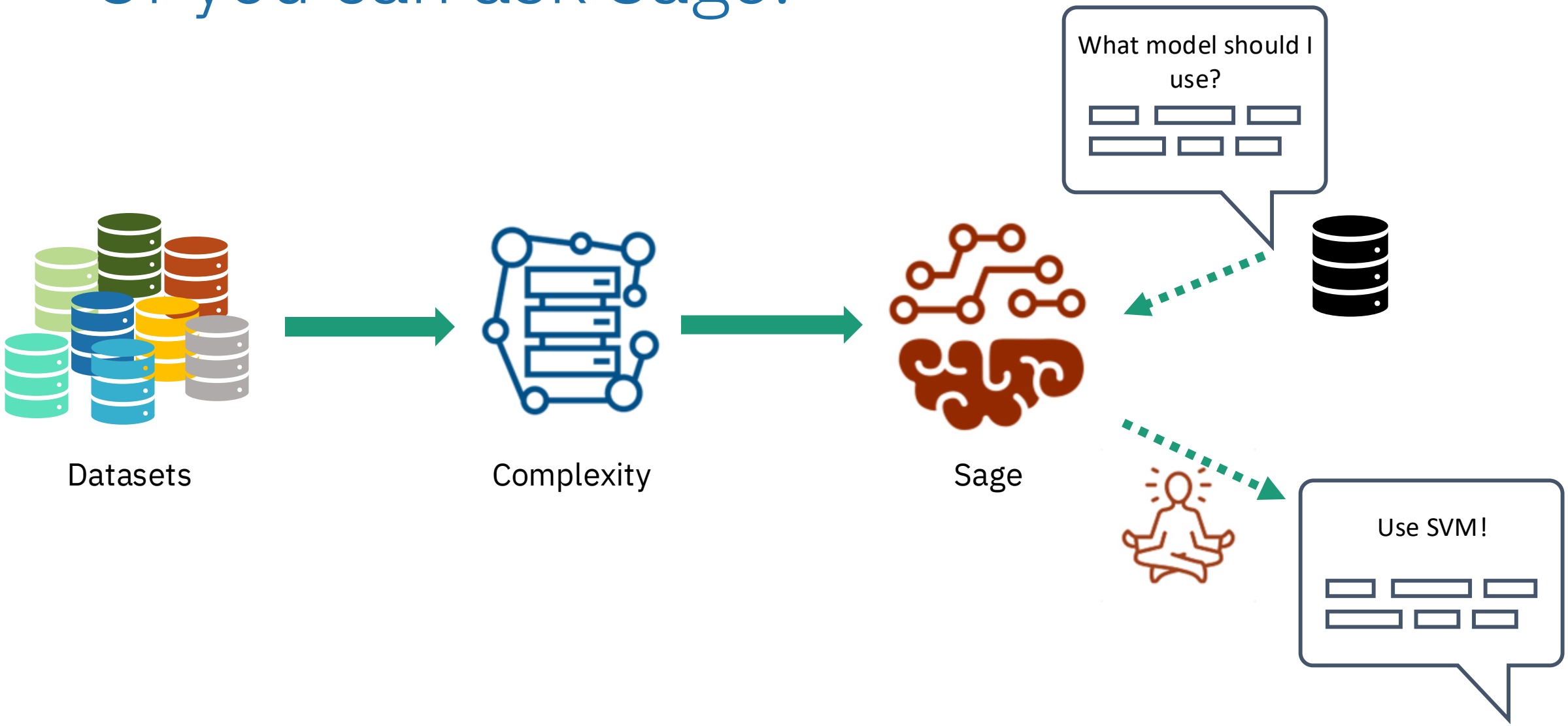
Class overlap



low high

Desired value

How should we use it?

👉 Helps in meta-learning and benchmarking

Or you can ask Sage!