
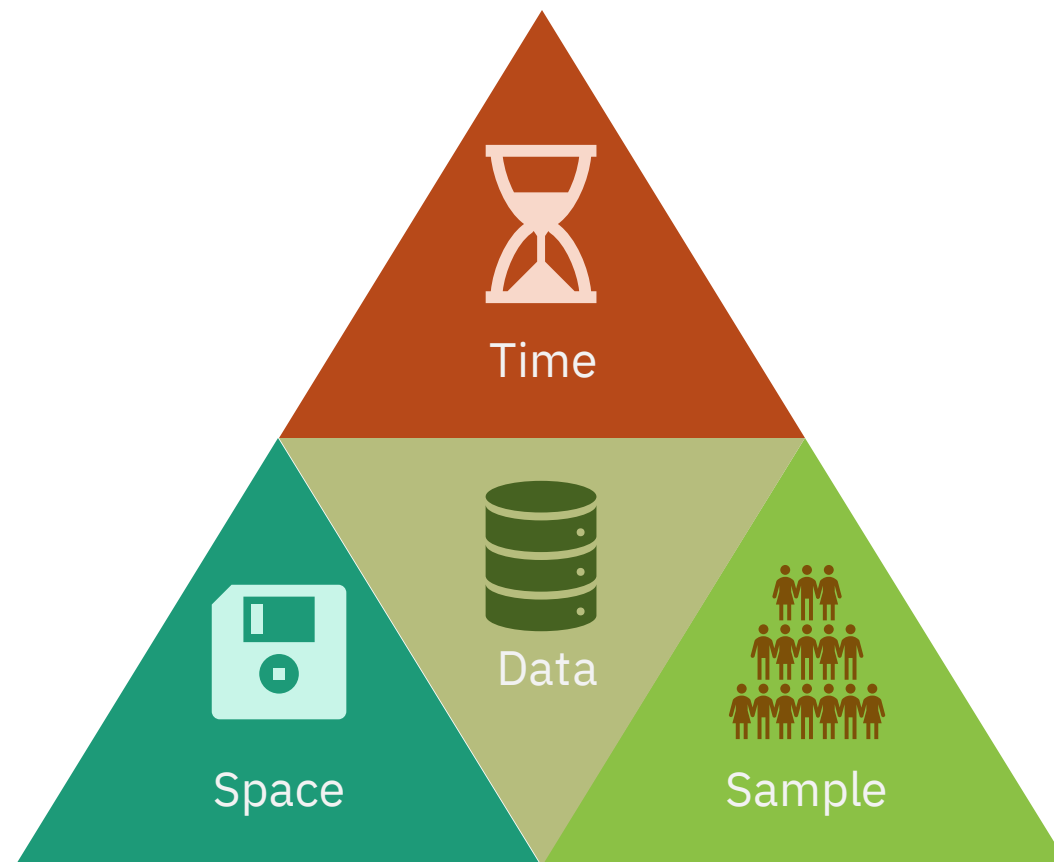


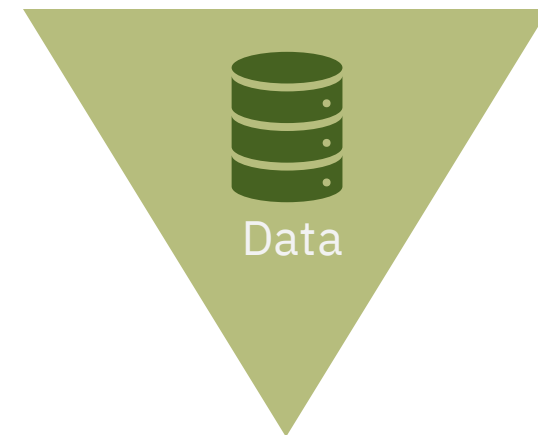
# Characterizing Data Complexity in Machine Learning

The background features a series of concentric circles in a lighter shade of blue. Overlaid on these circles is a network diagram consisting of several interconnected nodes (circles) and lines, suggesting a complex data structure or machine learning architecture.

# 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

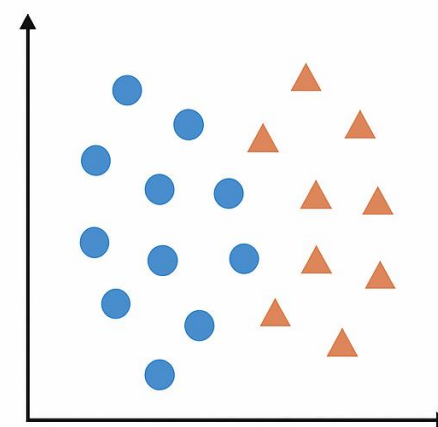
## Extrinsic

Complexity from external factors dependent on the algorithm or preprocessing.

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

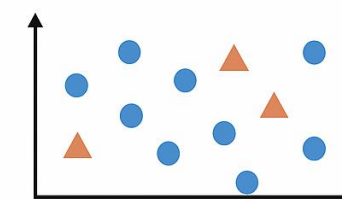


## Intrinsic Complexity



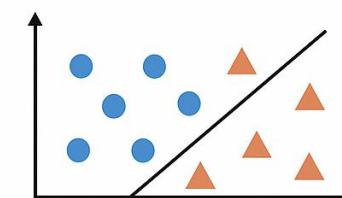
Overlapping class distributions in feature space

## 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
- Condition Number

## Geometric

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

## Sampling

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

# Data complexity

Dimensional



Rank of data ( $k$ )



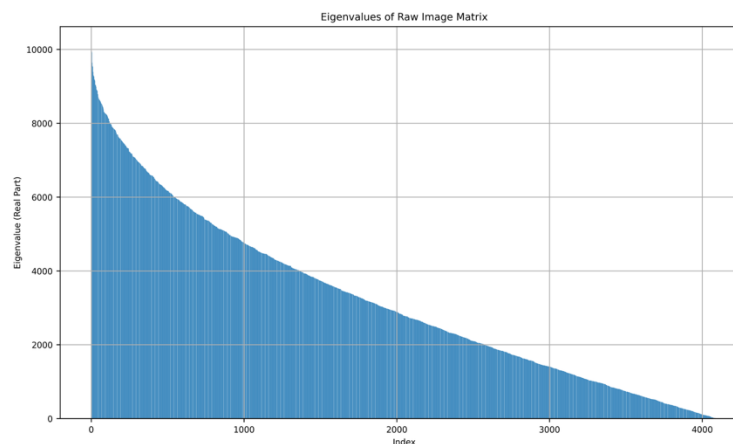
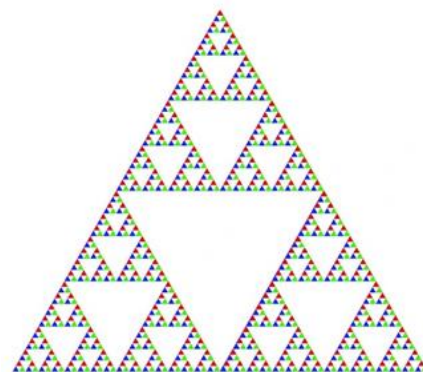
Fractal Dimension



Eigenspectrum



$$\left[ A \in R^{m \times n} \right] \approx \left[ \tilde{A} \in R^{m \times k} \right]$$



low high



Desired value

# Data complexity

Distributional



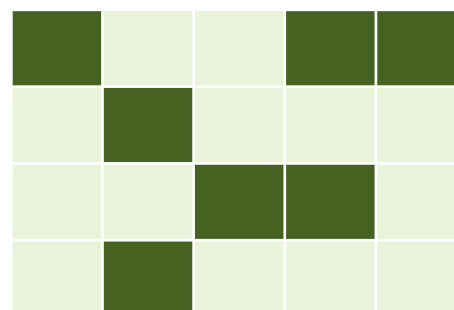
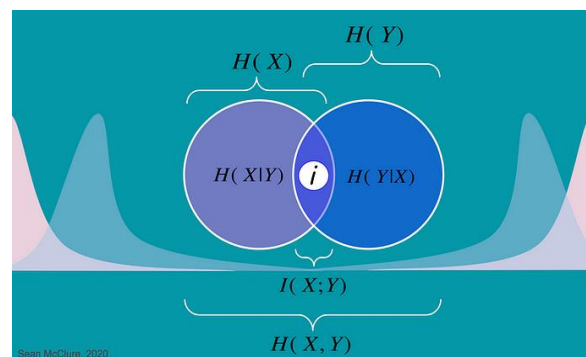
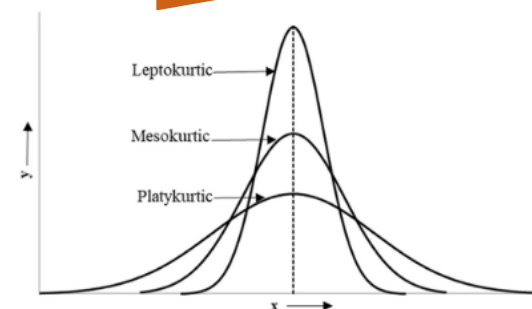
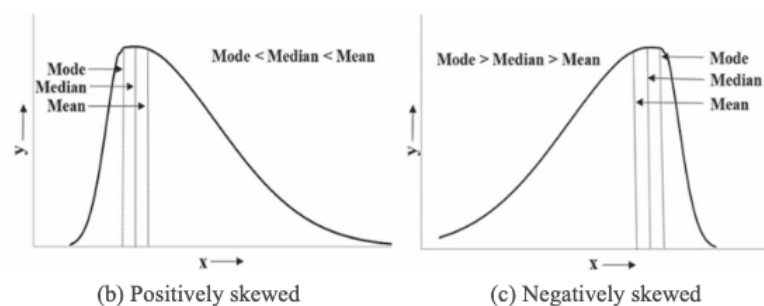
Skewness & Kurtosis



Mutual Information



Sparsity



low high



Desired value

# Data complexity

Geometric

Manifold



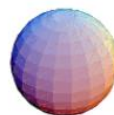
Kernel Density



Networks



sphere



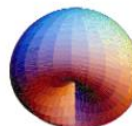
torus



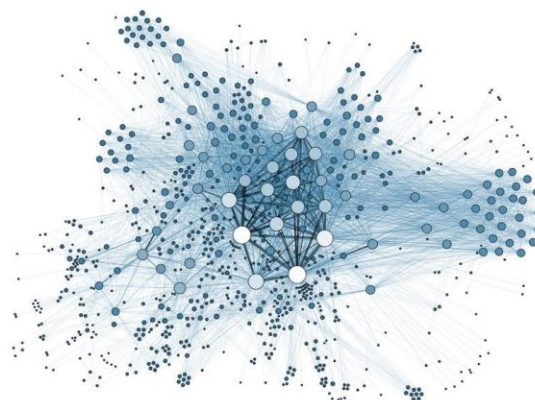
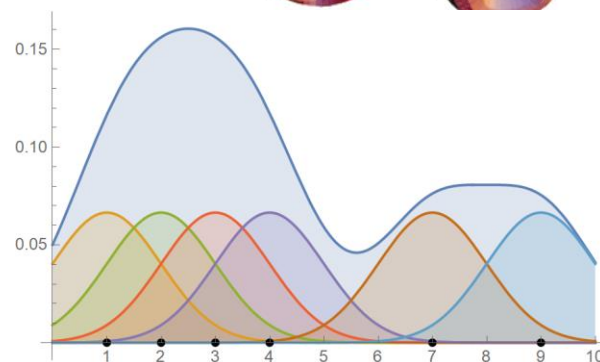
double torus



cross surface



Klein bottle



low high



Desired value



# Data complexity



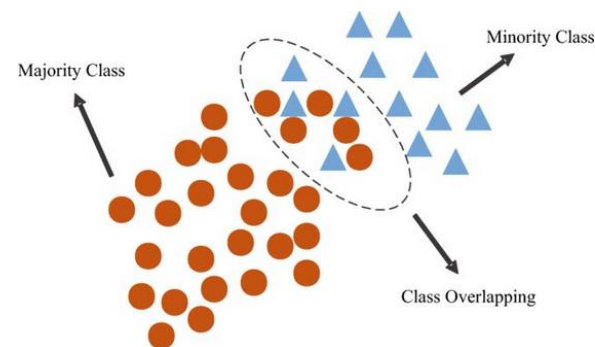
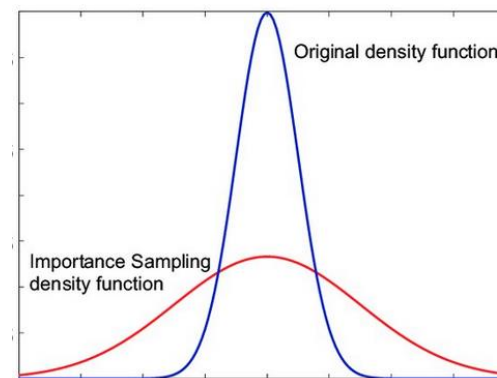
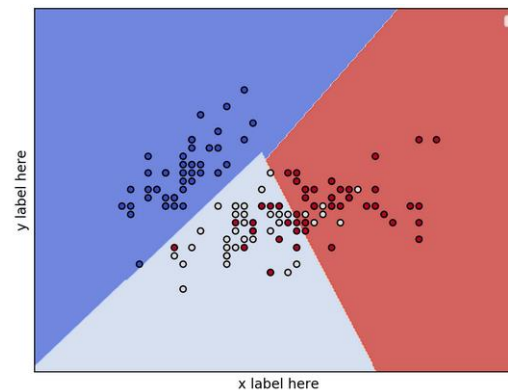
Class separation



Sampling variation



Class overlap



Sampling

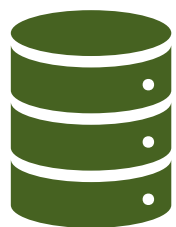
low high



Desired value

# How should we use it?

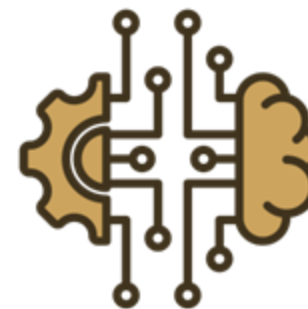
👉 Helps in meta-learning and benchmarking



Data



Complexity



Model

# Or you can ask QSage!

