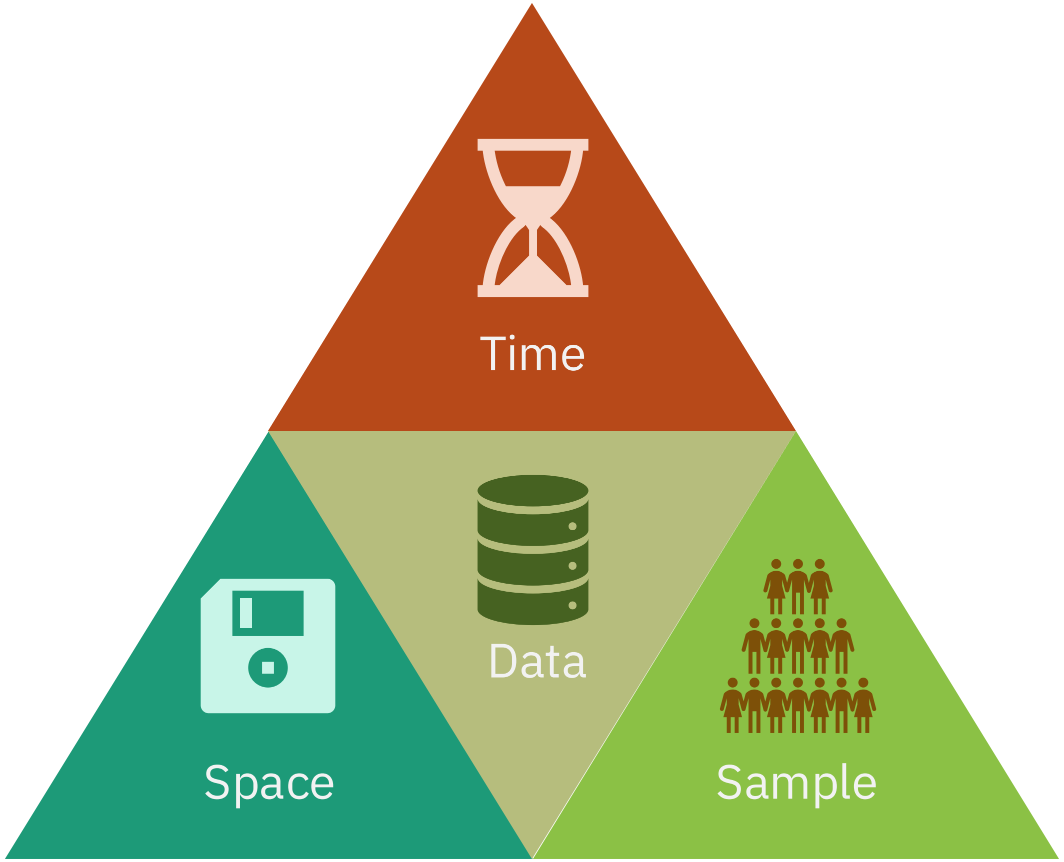
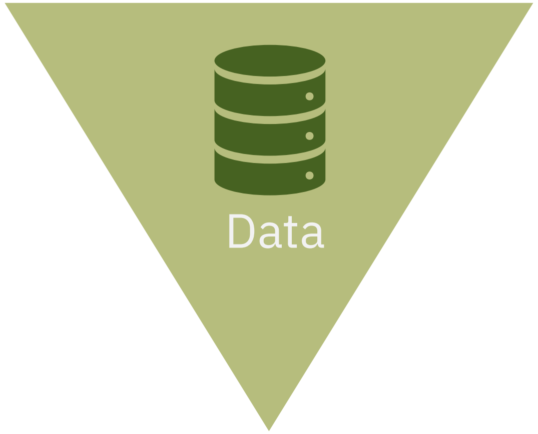
Characterizing Data Complexity in Machine Learning

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

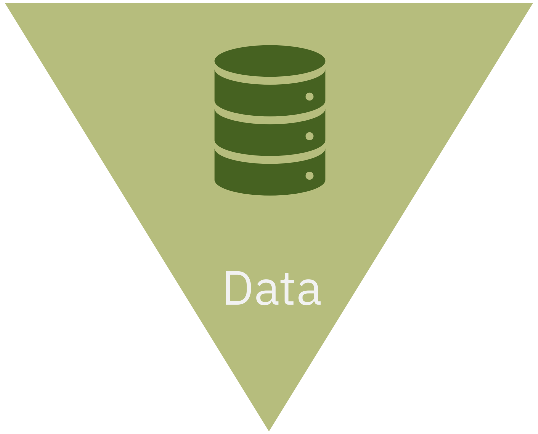
Inherent structure of the data which makes it difficult to learn independent of the algorithm.

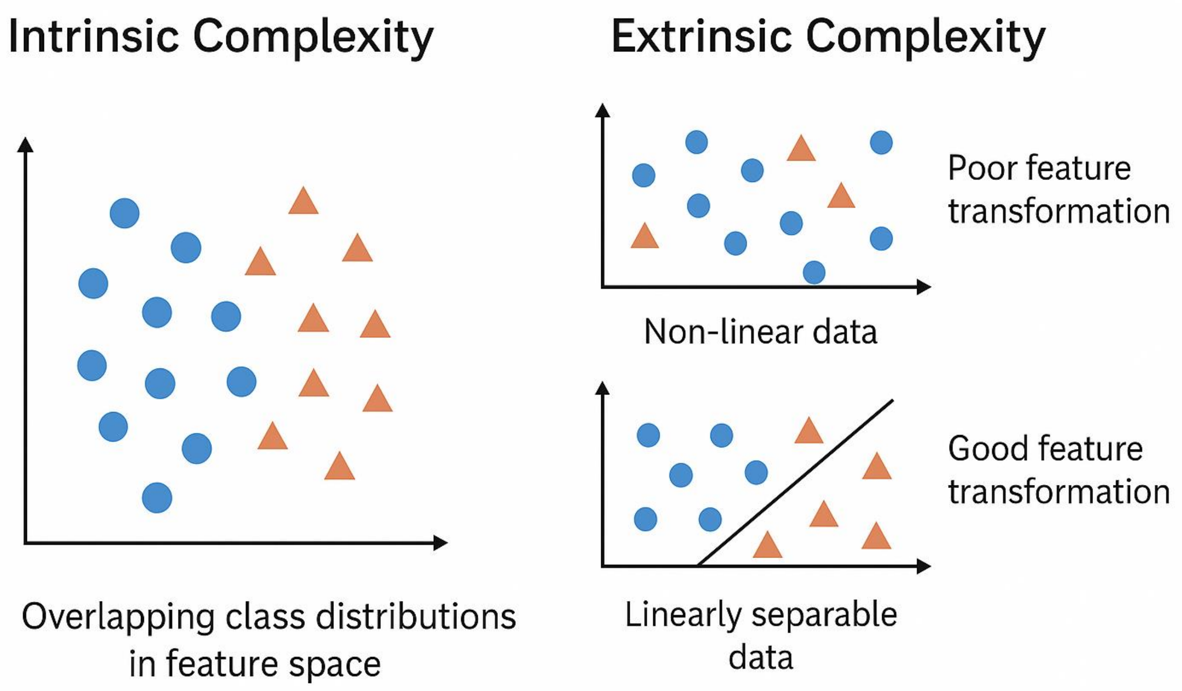
- Class distribution
- Non-linear decision boundaries
- Higher-order correlations
- Noise

Extrinsic

Complexity from externa factors dependent on the algorithm or preprocessing.

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





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

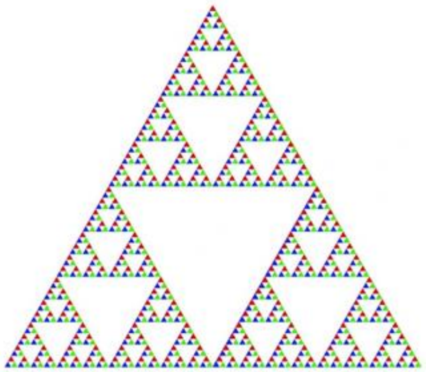
- Class imbalance ratio
- Class overlap measures
- Entropy
- Margin of separation between classes
- Sampling density variation

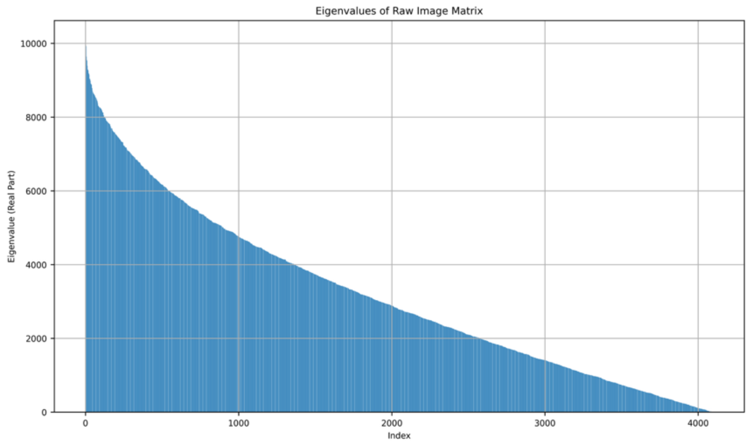
Data complexity

Dimensional

- Rank of data (k)
- Fractal Dimension
- Eigenspectrum

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





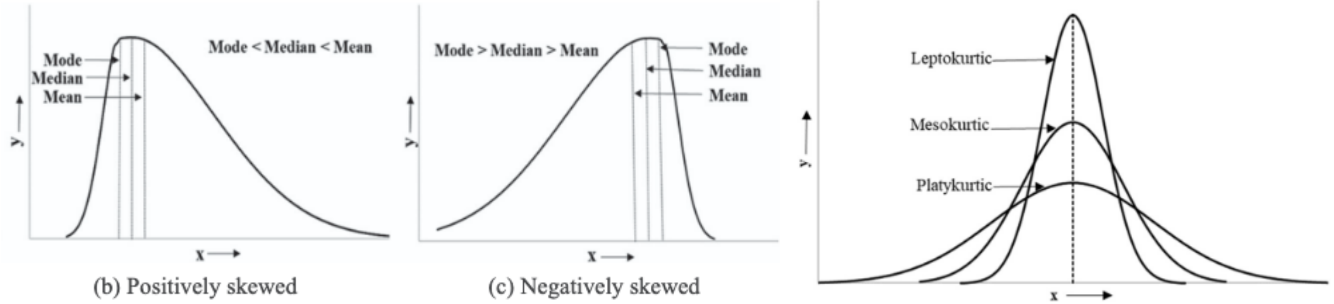
low high

Desired value

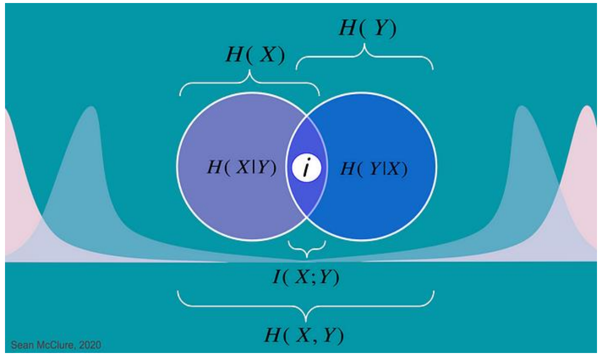
Data complexity

Distributional

- Skewness & Kurtosis



- Mutual Information



- Sparsity

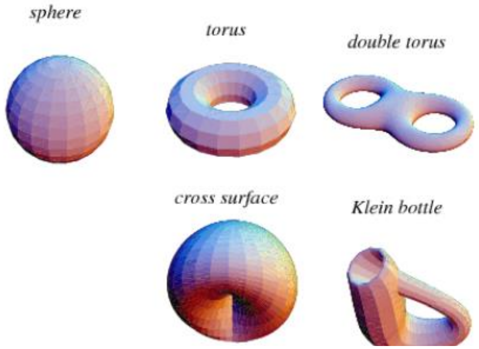
low
high

Desired value

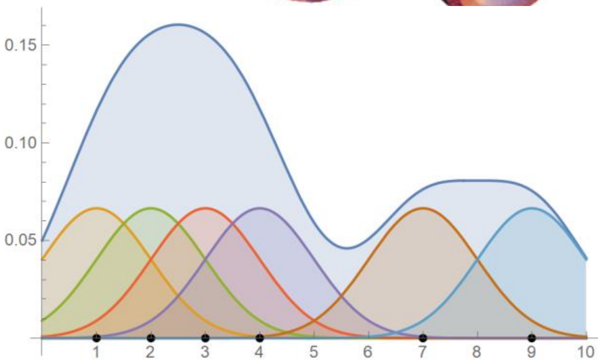
Geometric

Data complexity

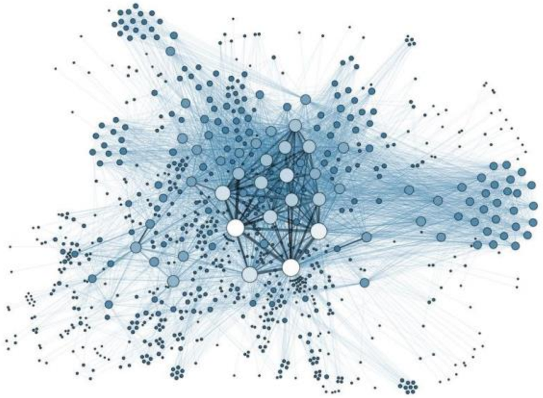
- Manifold



- Kernel Density



- Networks



low high

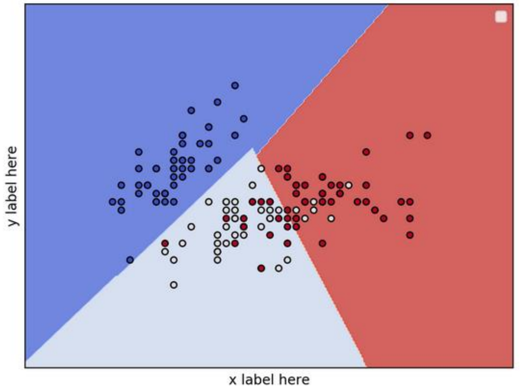
Desired value

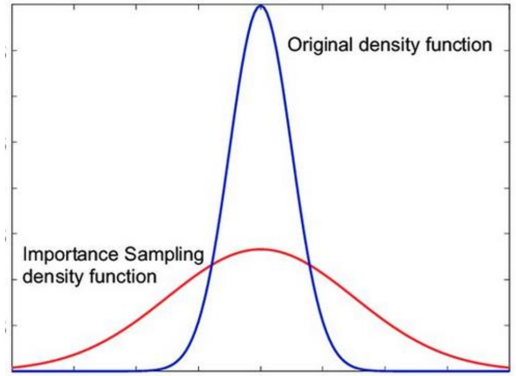
IBM Research

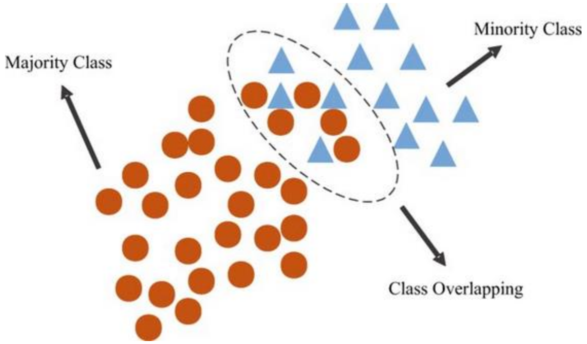
Data complexity

Sampling

- Class separation
- Sampling variation
- Class overlap







low high

Desired value

Graph Complexity Measures

We study graph complexity measures to understand which complexity measures are drivers of the quantum vs. classical graph representation learning for downstream tasks.

Structural Topology

- $\log(\# \, of \, nodes)$
-
- Density
- Average Degree
- Modularity
- Largest CC fraction
- Cycle Density
- …

Spectral Structure

- Spectral Gap
-
- IPR
- Spectral Entropy
- Log Odd Girth
- Algebraic Connectivity
- …

Diffusion

- Average Path Length
- Heat Kernel Trace
- Approx. Conductance
- Non-backtracking radius
- …

Transport Geometry

- Oliver-Ricci Conjecture
- Pagerank Gini
- Betweenness Gini
- Closeness Gini
- …

IBM Research

Graph Complexity Measures

Structural topology

Degree heterogeneity (Gini)

Homogeneous degrees → weak hub interference

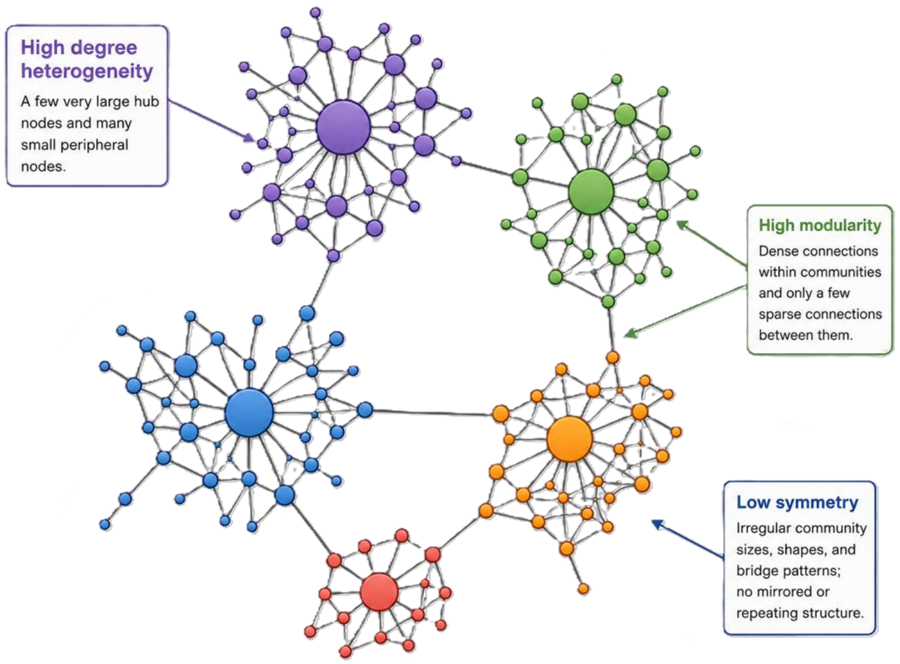
Modularity / community

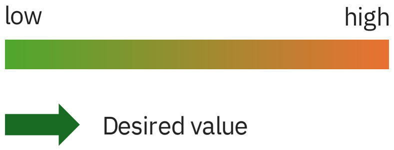
Few bottlenecks for QW to exploit

Symmetry (WL compression)

More automorphism → classical-tractable

*Quantum-favoring regime**





Classical-favoring regime: Homogeneous, weakly-clustered, high-symmetry graphs.

* Probable regime for quantum walks outperforming classical walks

Graph Complexity Measures

Spectral Structure

Eigenvector Localization (IPR)

High IPR → QW collapses to classical diffusion

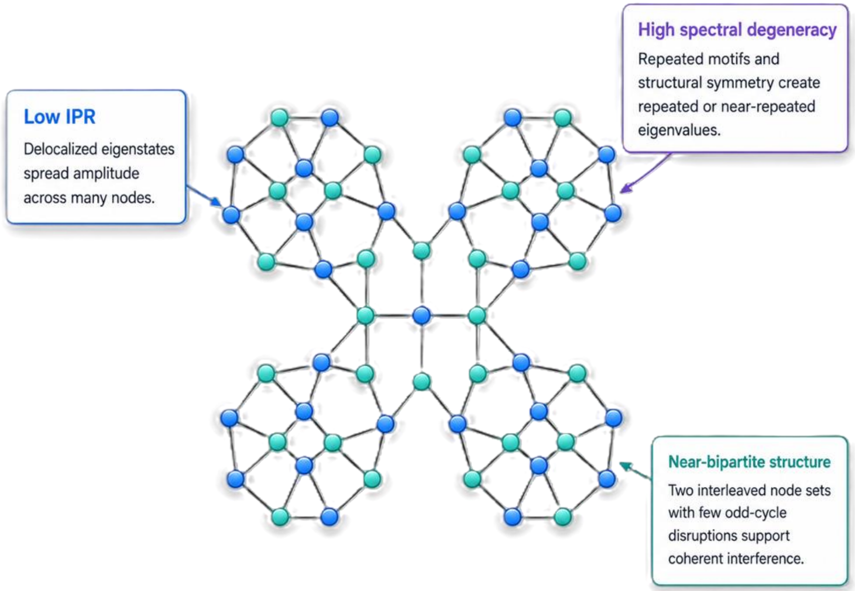
Spectral Degeneracy

Few degeneracies → weaker coherent interference

Bipartite Proximity

Far from bipartiteness → no DTQW channel

*Quantum-favoring regime**



low high

Desired value

Classical-favoring regime: localized eigenstates, non-degenerate spectrum, far from bipartite.

Note bipartite_proximity = $\max(0, 2 - \lambda_{max})$*: 0 means bipartite, so 'far from bipartite' is a HIGH metric value — arrow points high to match the code's sign convention.*

* Probable regime for quantum walks outperforming classical walks

Graph Complexity Measures

Diffusion & Mixing

Spectral Gap

Fast classical mixing → little room for QW gain

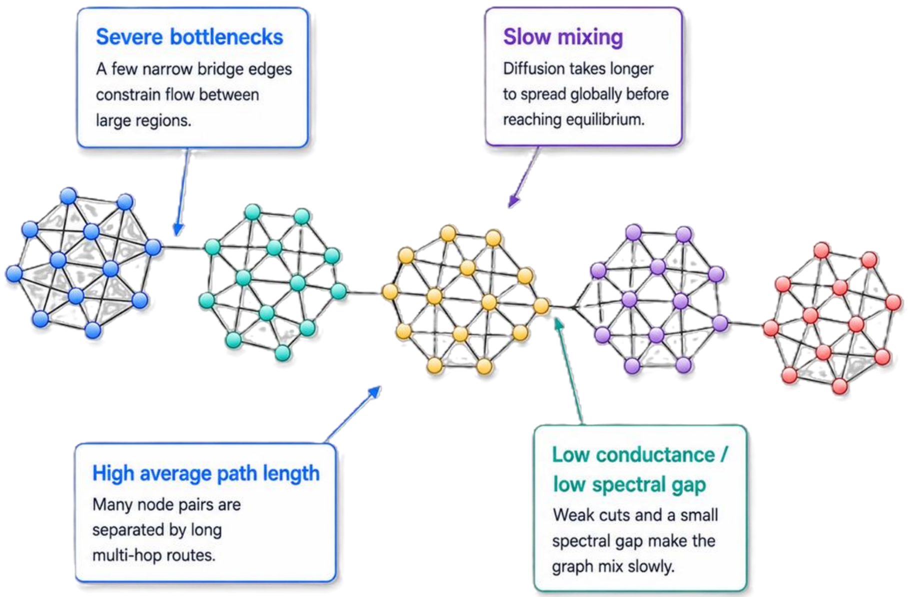
Conductance

No tight bottleneck → classical transport adequate

Average Path Length

Short paths → classical reach already efficient

*Quantum-favoring regime**



Classical-favoring regime: well-connected, fast-mixing graphs with no severe bottleneck.

* Probable regime for quantum walks outperforming classical walks

Graph Complexity Measures

Transport Geometry

Negative ORC fraction

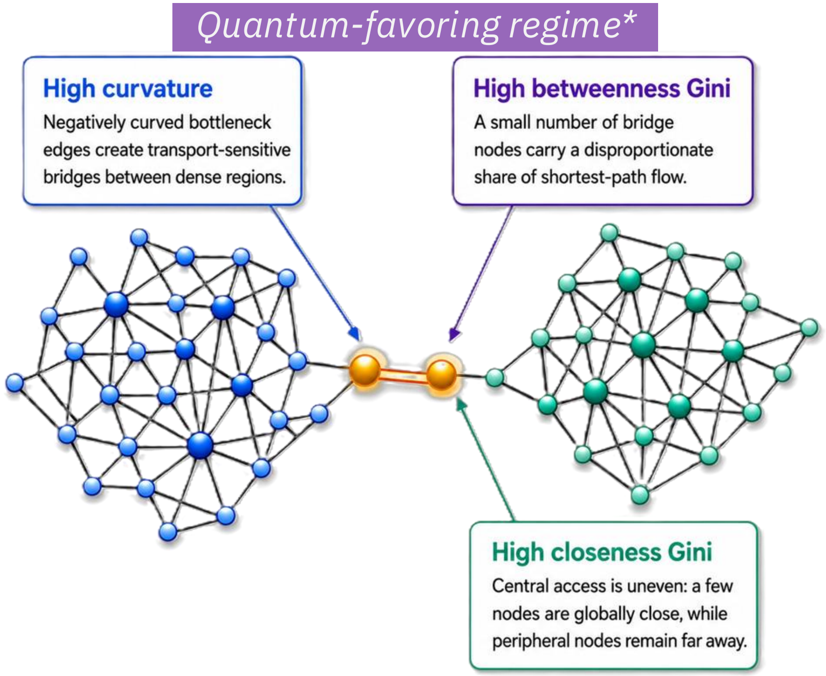
Fewer negatively-curved bottleneck edges

Betweenness Gini

Even load → no congestion for QW to bypass

Closeness Gini

Uniform reach → classical walk suffices



low high

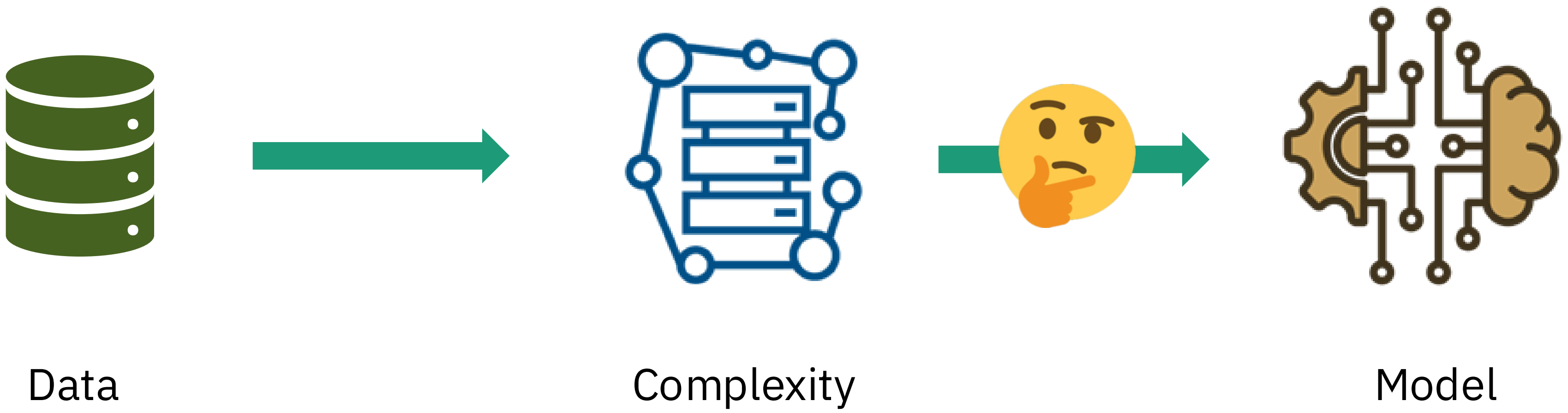
Desired value

Classical-favoring regime: near-flat curvatures and even load distribution.

* Probable regime for quantum walks outperforming classical walks

How should we use it?

- Helps in meta-learning and benchmarking



Or you can ask QSage!