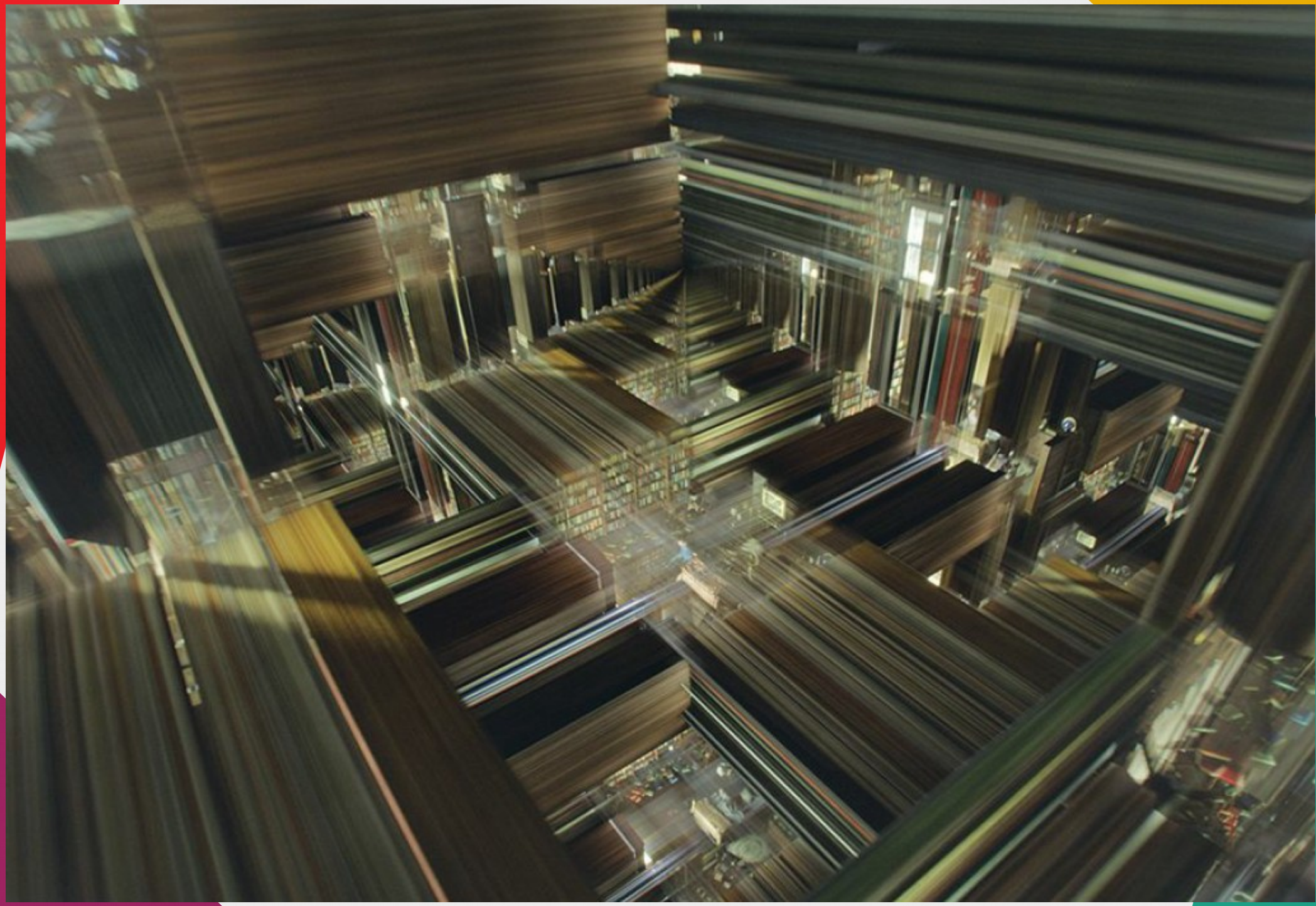
# Feature Selection

Arun Aniyan
SKA SA

# Curse of Dimensionality



# Feature Selection (FS)

- Feature selection is also called variable selection or attribute selection.
- Selection of variables for better class separation
- Feature selection is different from dimensionality reduction.

# Feature Selection vs Dimensionality Reduction

- Dimensionality reduction methods are mostly unsupervised while feature selection is most usually supervised.
- The objective function for dimensionality reduction is not class separability but better representation of data, where are FS methods optimize for class separability.
- PCA etc reduces dimensions by creating new combination of attributes. FS methods selects subset of features.

# When do you need ?

- Create accurate prediction model
- Find meaningful relationships
- Simpler model
- Faster and cost effective prediction
- Better understanding of underlying process

# If “YES” do FS

- Do you have domain knowledge ?
- Suspicion of feature interdependence
- Asses features individually
- Data is sparse
- Need a stable solution

# Simplest Feature Selection Methods

- Forward selection
- Backward Selection
- Floating search method

# Feature Selection

- Filter Methods
- Wrapper Methods
- Embedded Methods

# Filter Based Feature Selection

- Uses intrinsic properties of data
- Derive statistical  measures to evaluate quality of features.
- Ranking methods are applied (univariate and multivariate)
- Effective in computation time and robust to overfitting.

# Filter Based Feature Selection

- Fischer Score
- Relief Feature selection

# Wrapper Based Feature Selection

- Based on a learning algorithm
- Evaluated on different combinations of features
- Involves heuristics like hill climb, best first search etc
- Involves significant computational time
- Risk of overfitting

# Wrapper Based Feature Selection

- Tree based FS
- Evolutionary methods – GP
- Correlation based methods – CFS , FCBF
- Simulated annealing  (Boltzmann learning)

# Embedded Methods

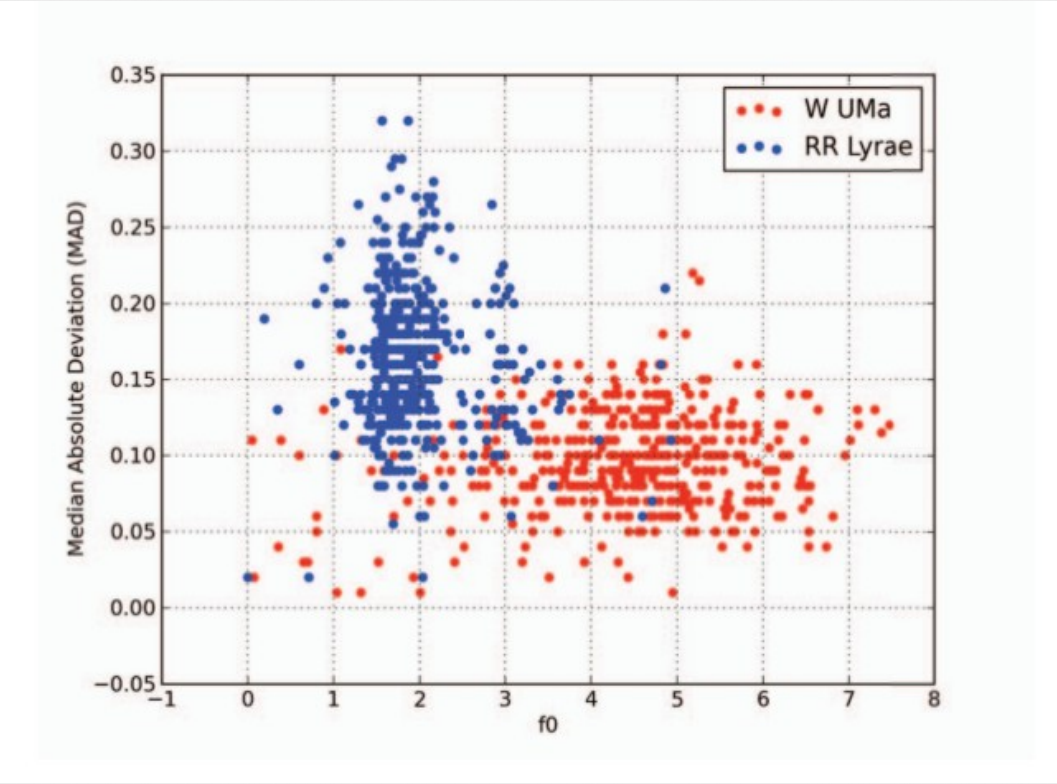
- Recently added type of methods
- Combination of regularization and manifold methods
- Significant computational time.
- MCFS (Multi Cluster Feature Selection)

# Applications in General

- Better visualization
- Better understanding of data and variables
- Better precise model
- Reduces training time
- Reduces storage requirement

# Applications in General

- Find interesting relationships between variables




Donalek, Aniyan et.al 2014

# Hobby / Toy Project

- Silhouette Score Based Feature Selection – Work in (very) slow progress