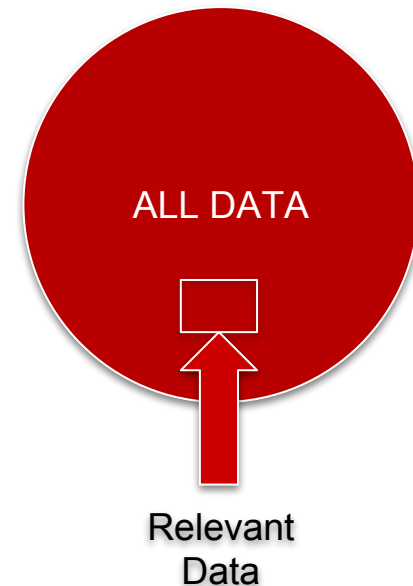# Feature Selection

university of groningen

Team 14

Topics

› Feature Selection
  - What is it?
  - How does it work?
› Ranking techniques
› Curse of Dimensionality
  - Combinatorial explosion
  - Distance concentration
› Approaches
  - Filter Methods
  - Wrapper approaches
  - Embedded approaches

Feature Selection

> Why?
> - Simplify models
> - Shorten training time
> - Reduce overfitting
> - Avoid Curse of Dimensionality

ALL DATA

Relevant Data

Feature Selection

› How?
  - Algorithms
  - Machine Learning

› Standard approaches:
  - Filter Methods
  - Wrapper approaches
  - Embedded approaches

Variable Ranking Techniques

Exploratory analysis
Scalable and efficient filter for further test

E.g: Eigenfaces pixels ranked by F statistic
per variable classification performance

Variable Ranking Techniques

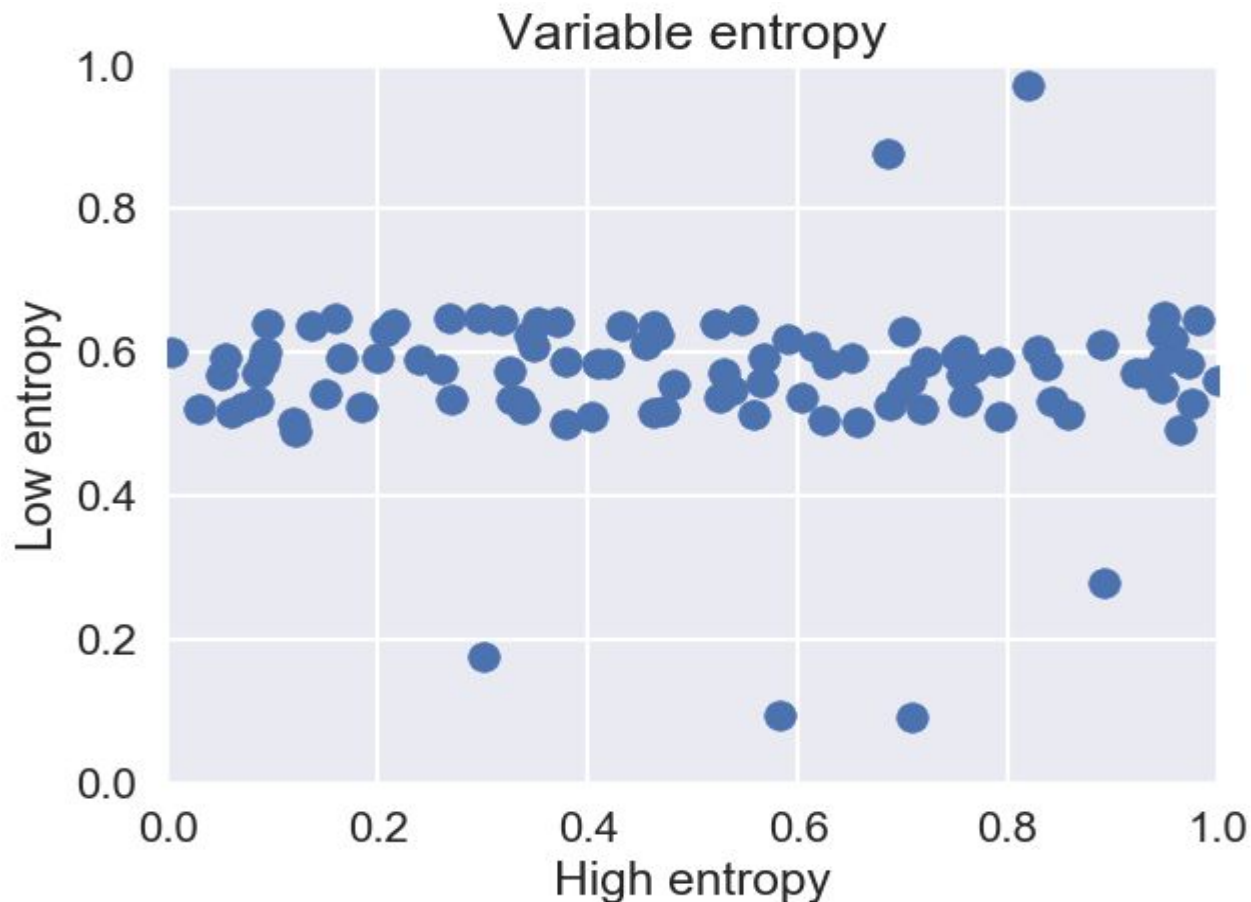Ranking with variable interactions
    E.g. Relief algorithm

Variable Ranking Techniques

Unsupervised:

E.g. Density, reliability, smoothness
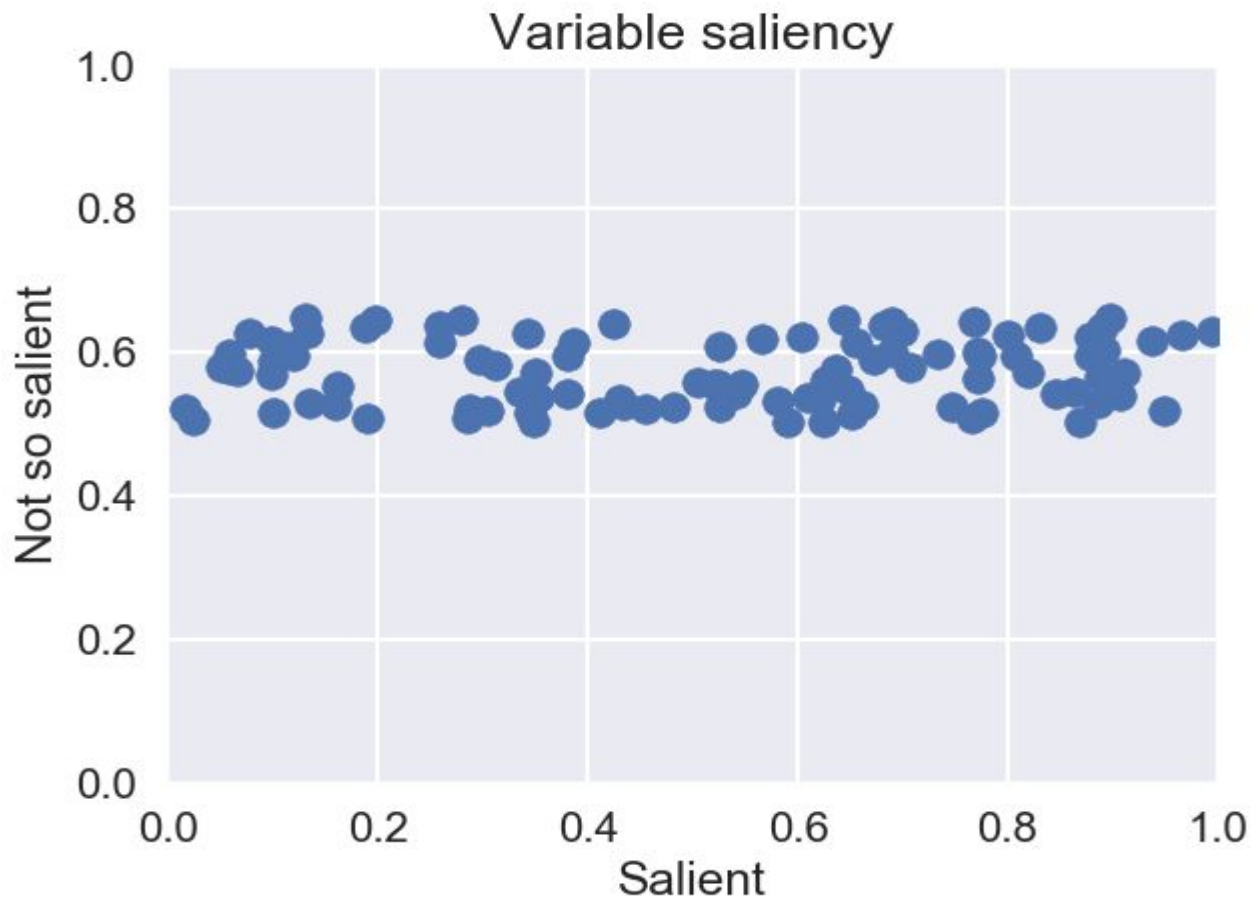
Variable Ranking Techniques

Unsupervised:

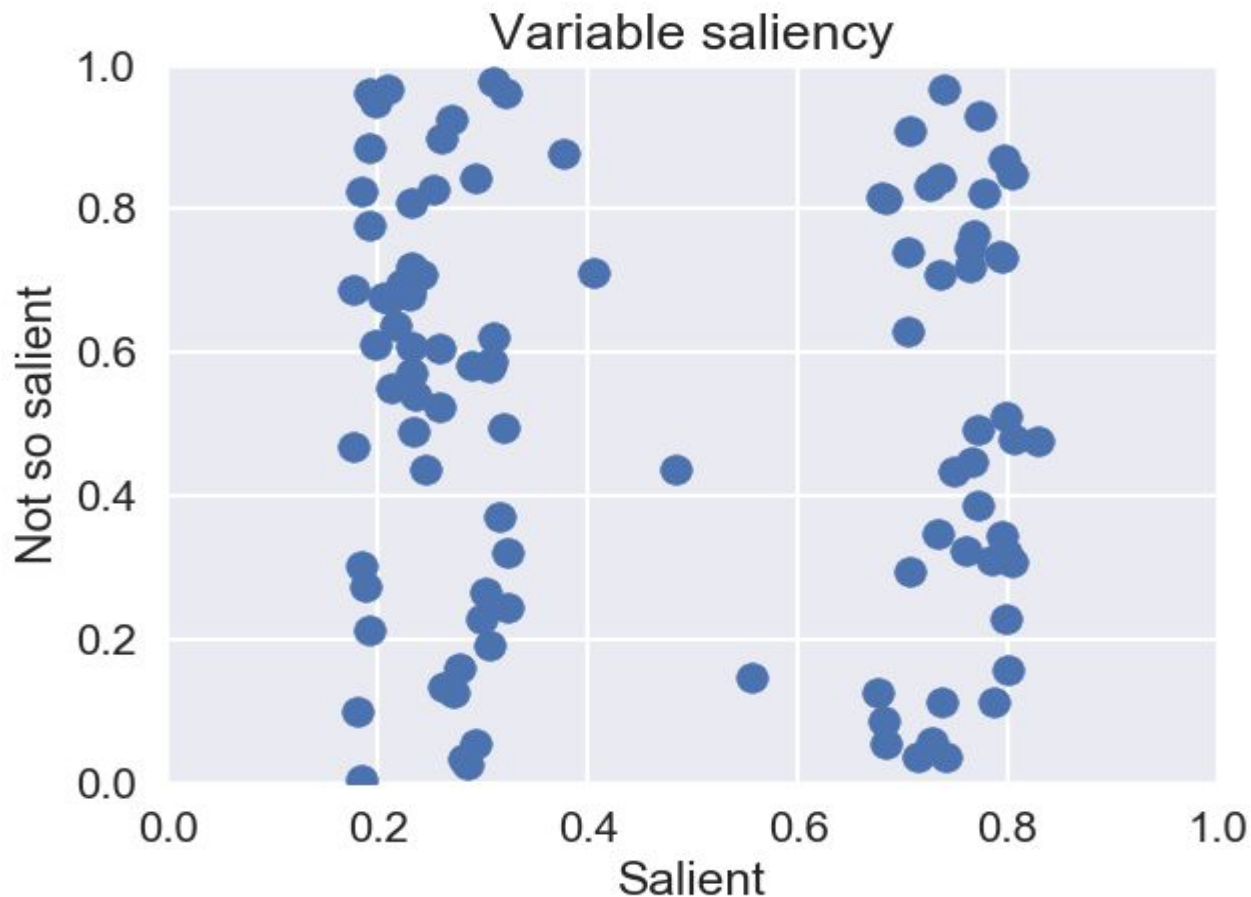E.g. entropy

Variable Ranking Techniques

Unsupervised:
 E.g. saliency

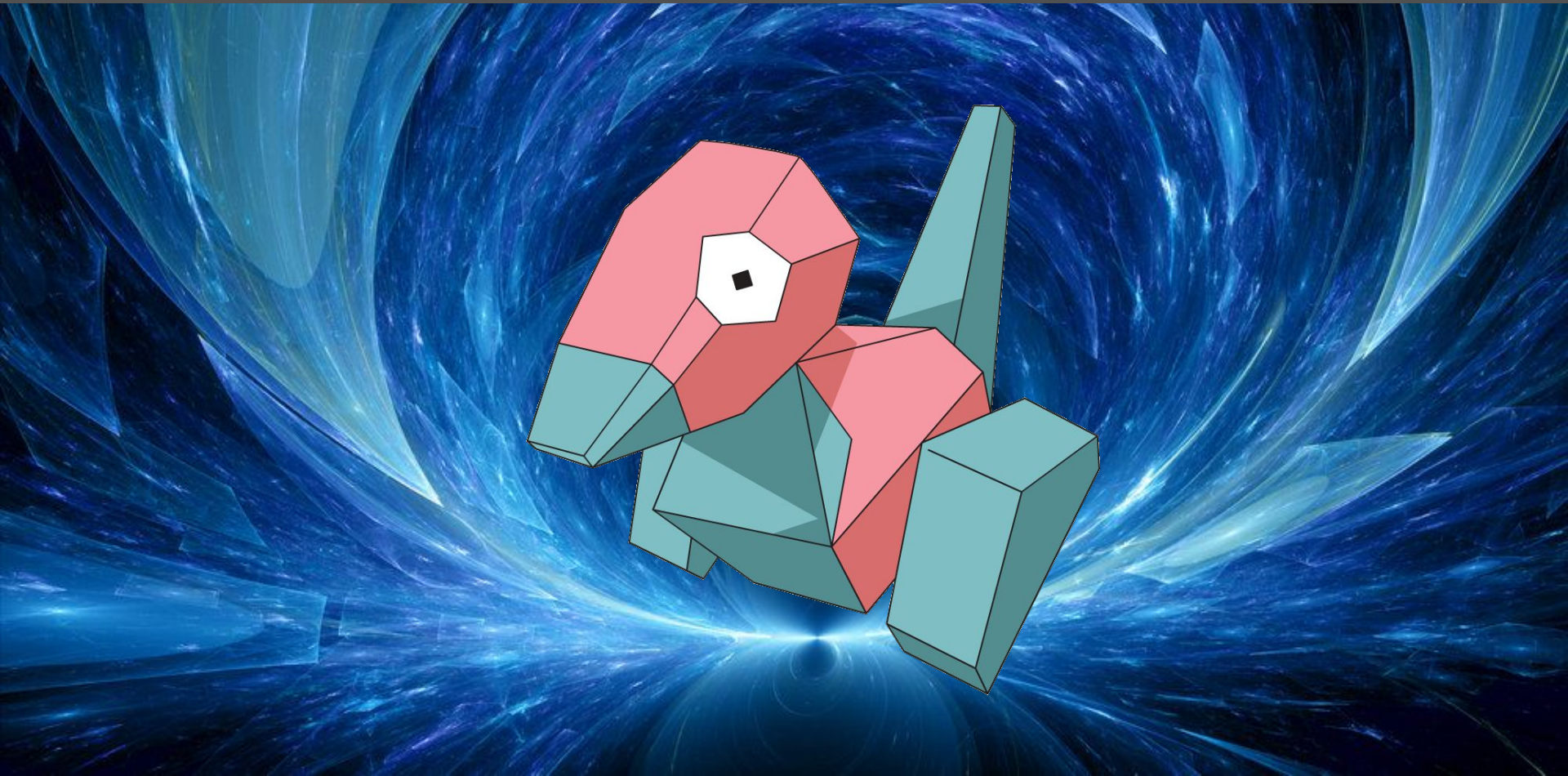Variable saliency

Variable Ranking Techniques
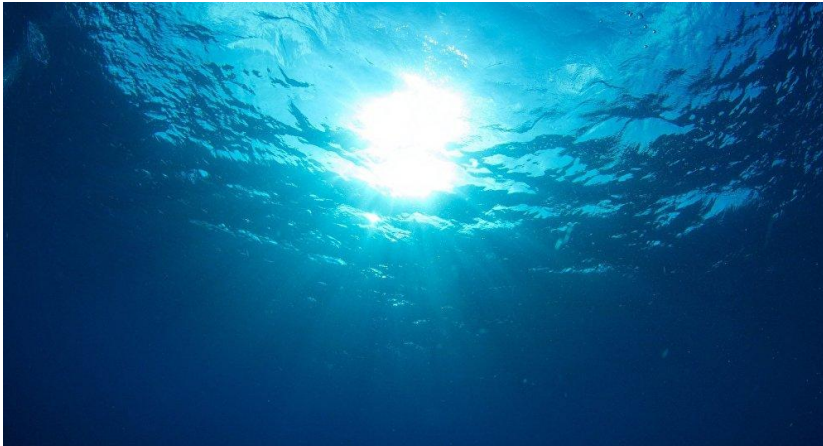
Unsupervised:
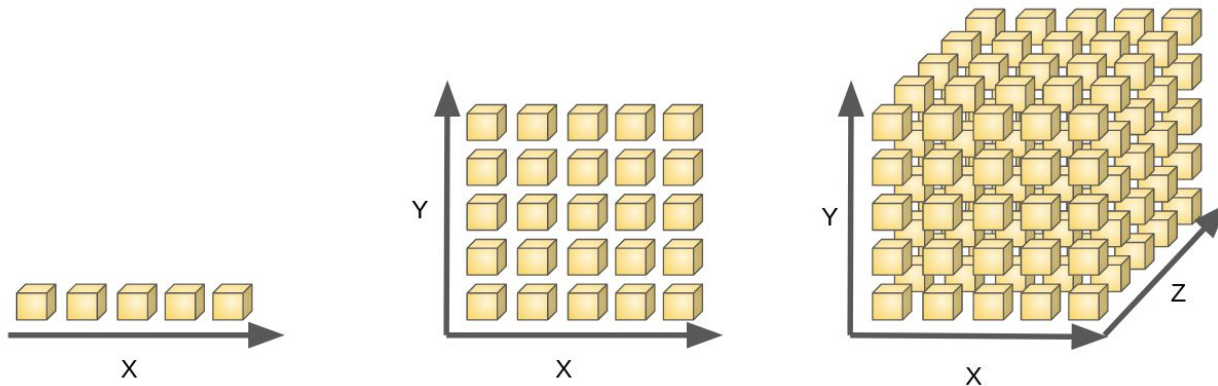E.g. saliency

Curse of Dimensionality

Curse of Dimensionality

- High-dimensional spaces:
  - Images
  - Videos
  - Genes

Curse of Dimensionality

- Analyzing/organizing data in **high-dimensional** spaces

- Data becomes **sparse**

- Data needed to support the result grows **exponentially**

- Organization strategies become **inefficient**

Curse of Dimensionality

Combinatorial explosion

- Puzzles (sudoku, etc...)
- Factorial in arithmetics
- Boolean system

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

N	$N!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5,040
8	40,320
9	362,880
10	3,628,800

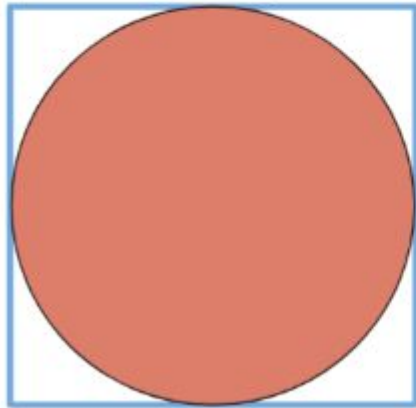| A	B	C	D	Result
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	1
1	0	1	1	1
1	1	0	0	0
1	1	0	1	0
1	1	1	0	1
1	1	1	1	1

Curse of Dimensionality

Combinatorial explosion

- Each combination of possible values must be considered

- Each additional dimension increases **exponentially** the numbers of possibilities
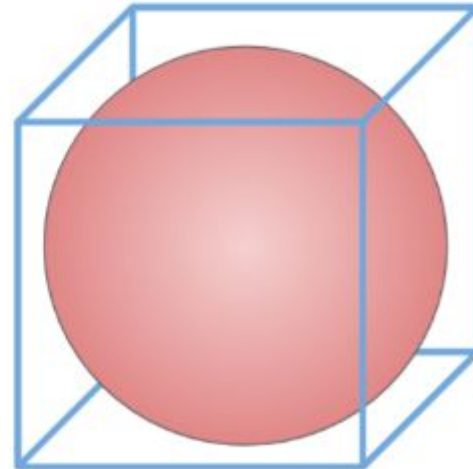
Curse of Dimensionality

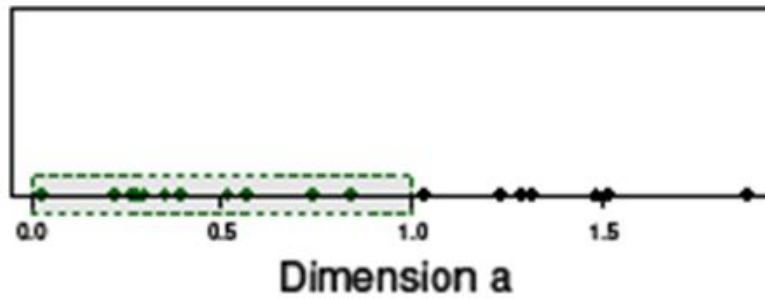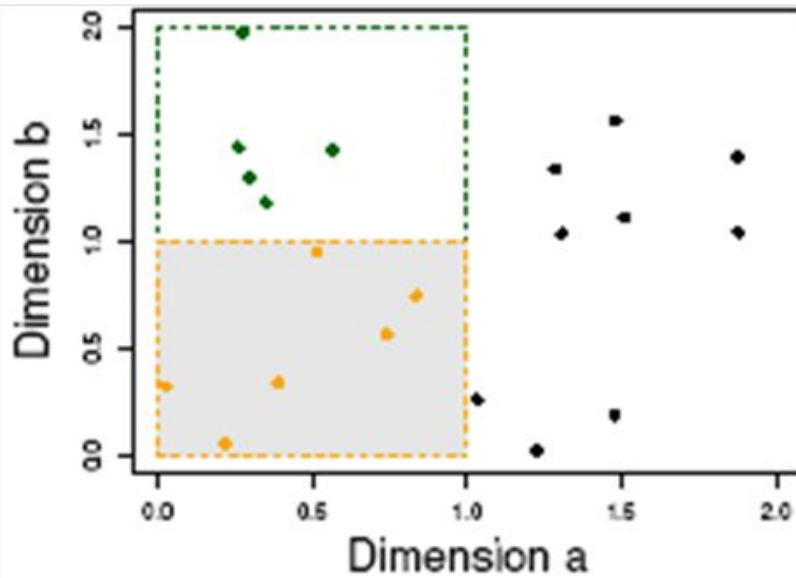**Distance concentration**

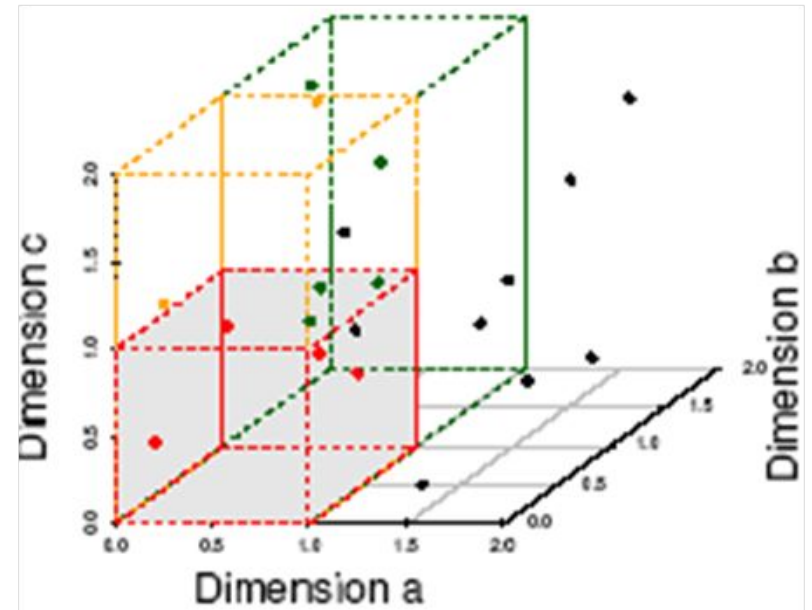A — 21.5% Empty

B — 47.6% Empty

Curse of Dimensionality

$$\frac{\dfrac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)}}{2r^n} = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma(\frac{n}{2}+1)}$$
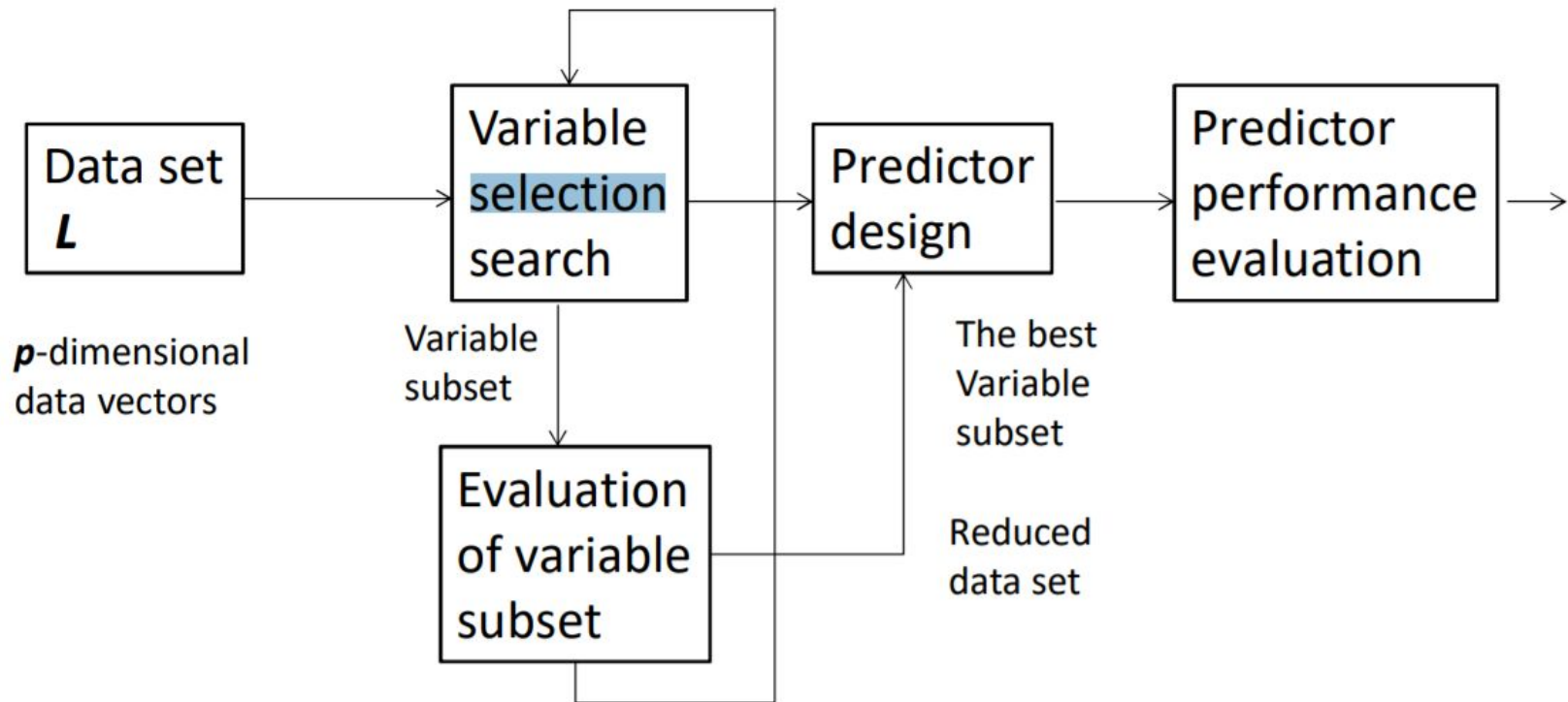
(a) 11 Objects in One Unit Bin

(b) 6 Objects in One Unit Bin

(c) 4 Objects in One Unit Bin

Approaches: Filter Methods
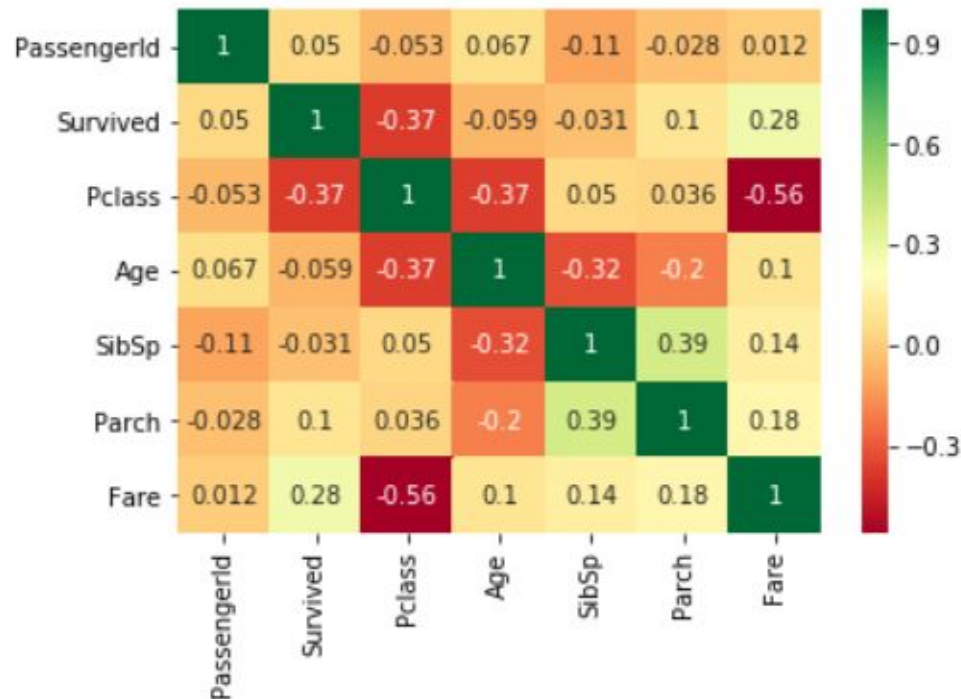
Approaches: Filter Methods

› Use
  - Intrinsic properties of the data
  - Statistic methods: chi-square, ANOVA, Correlation
› Calculate
  - Subset of the variables based on thoose methods
› Rank
  - The variables according to a certain result
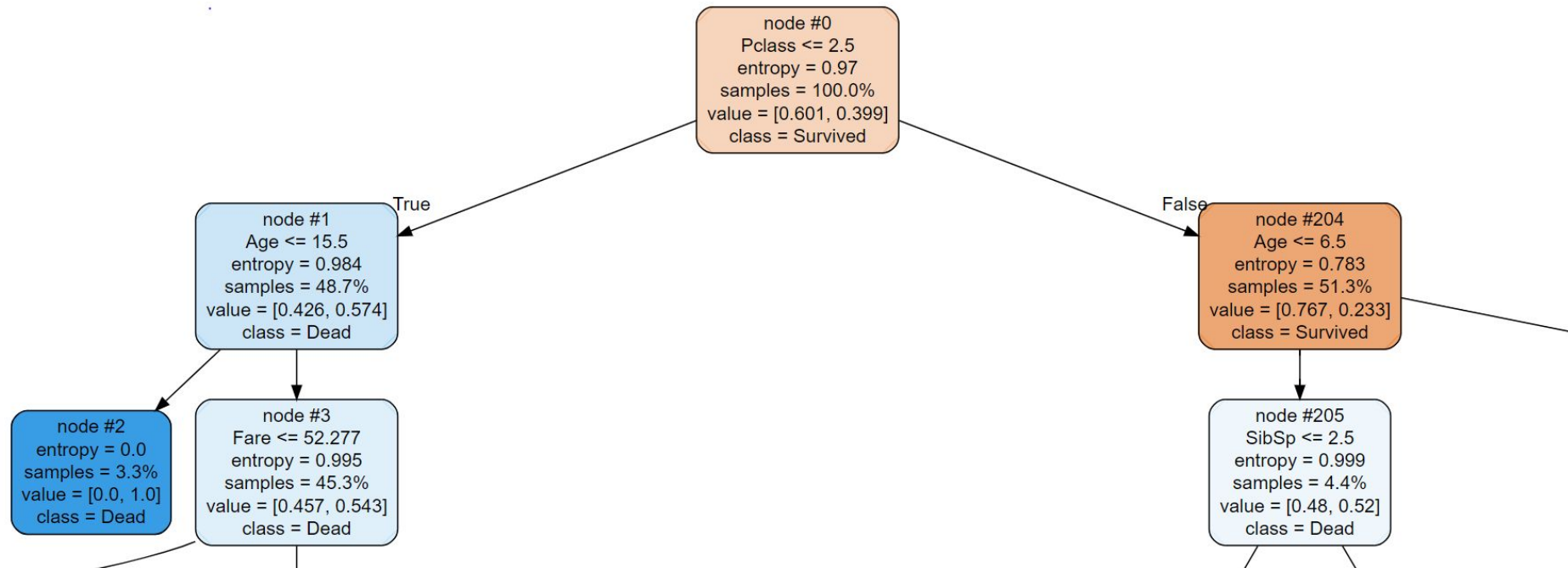
Approaches: Filter Methods

ANOVA

The p-value of Age is: 0.16169995412816476
The p-value of Class is: 5.487184140399378e-20
The p-value of # of Siblings is: 0.46609165802064034
The p-value of # of Parents and children is: 0.017105880263189474
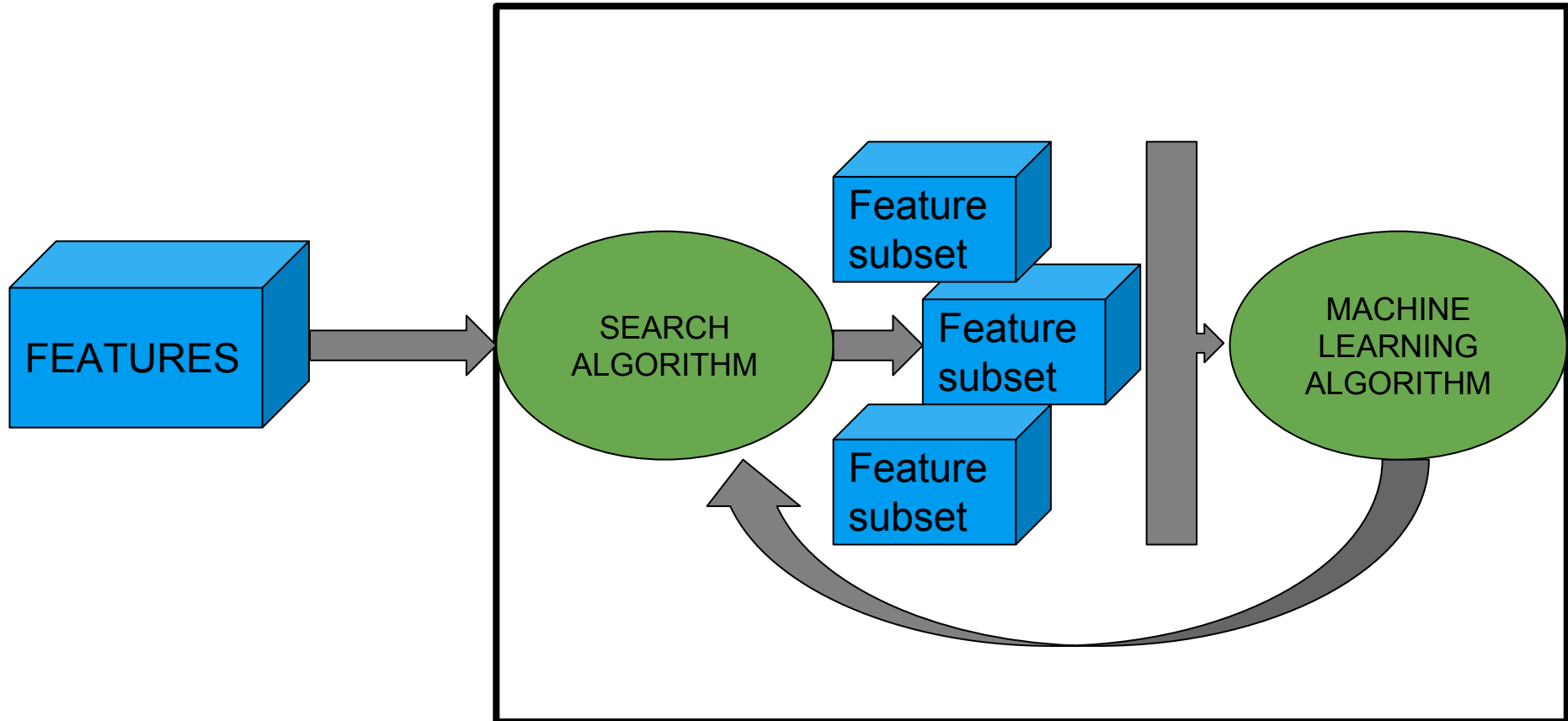The p-value of Fare is: 1.0265102576807696e-11

CORRELATION

Approaches: Filter Methods

**Entropy (Information Gain) Decision Trees**

Approaches: Wrapper

Approaches: Wrapper

› Require
  - state space: feature subsets
  - initial state
  - termination condition
  - search engine
› Search algorithm:
  - Exponential complexity
  - Forward selection/Backward elimination
› Machine learning algorithm:
  - Search criterion/-a of search
  - Feedback to search algorithm

Approaches: Embedded

"A learning algorithm that takes advantage of its own variable selection process and performs both feature selection as well as classification simultaneously."

Selecting the best subset

Set of all features → Generate a subset → Learning Algorithm + Performance
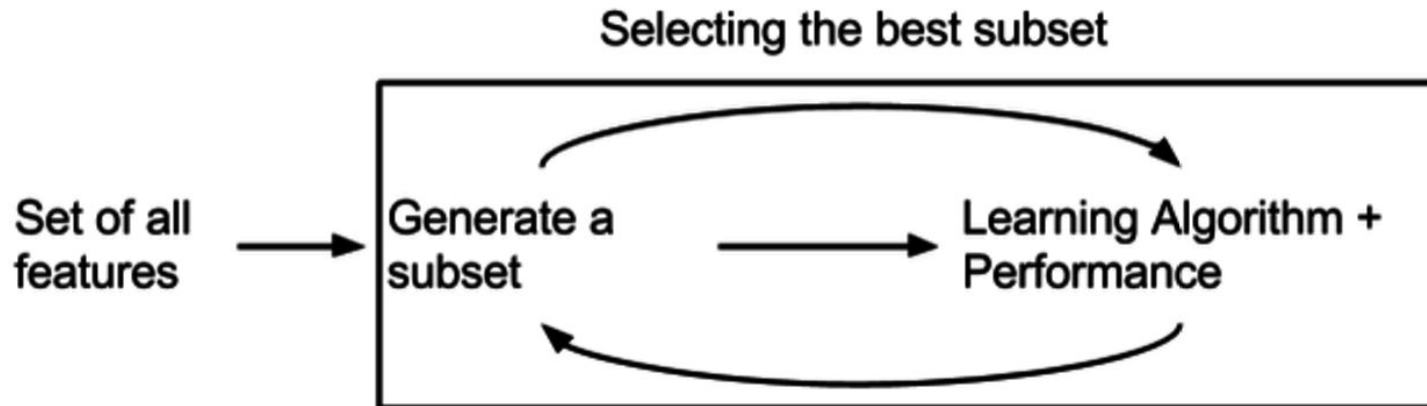
Example: Iterated Local Search

Algorithm: Genetic

Classifier: Support Vector Machine

Evaluation: Classification accuracy (tenfold)

Approaches: Embedded

Advantages

› All-in-one method

› Less computationally expensive

› Less prone to overfitting

Disadvantages

› Needs multiple algorithms

› Less flexible with backtracking

› Specific to a learning machine

Questions