university of groningen

Feature Selection

Team 14

Topics

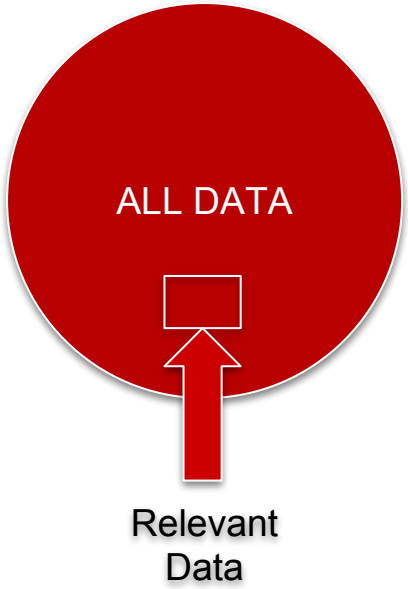
- Feature Selection DAAN
  - What is it?
  - How does it work?
- Curse of Dimensionality EMILE
  - Combinatorial explosion EMILE
  - Distance concentration ESER
- Approaches
  - Filter Methods CARLOS
  - Wrapper approaches PANAGIOTIS
  - Embedded approaches DAAN
- Ranking techniques DANNY

Feature Selection

- Data sets are large
- Not all data is necessary

- Feature Selection:
  - Selecting a relevant subset
  - Disregard unneeded features

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



Feature Selection

- How?
  - Algorithms
    - Greedy
    - Best-first
    - Exhaustive
  - Machine Learning
    - Neural Networks
    - Cross-validation
- 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

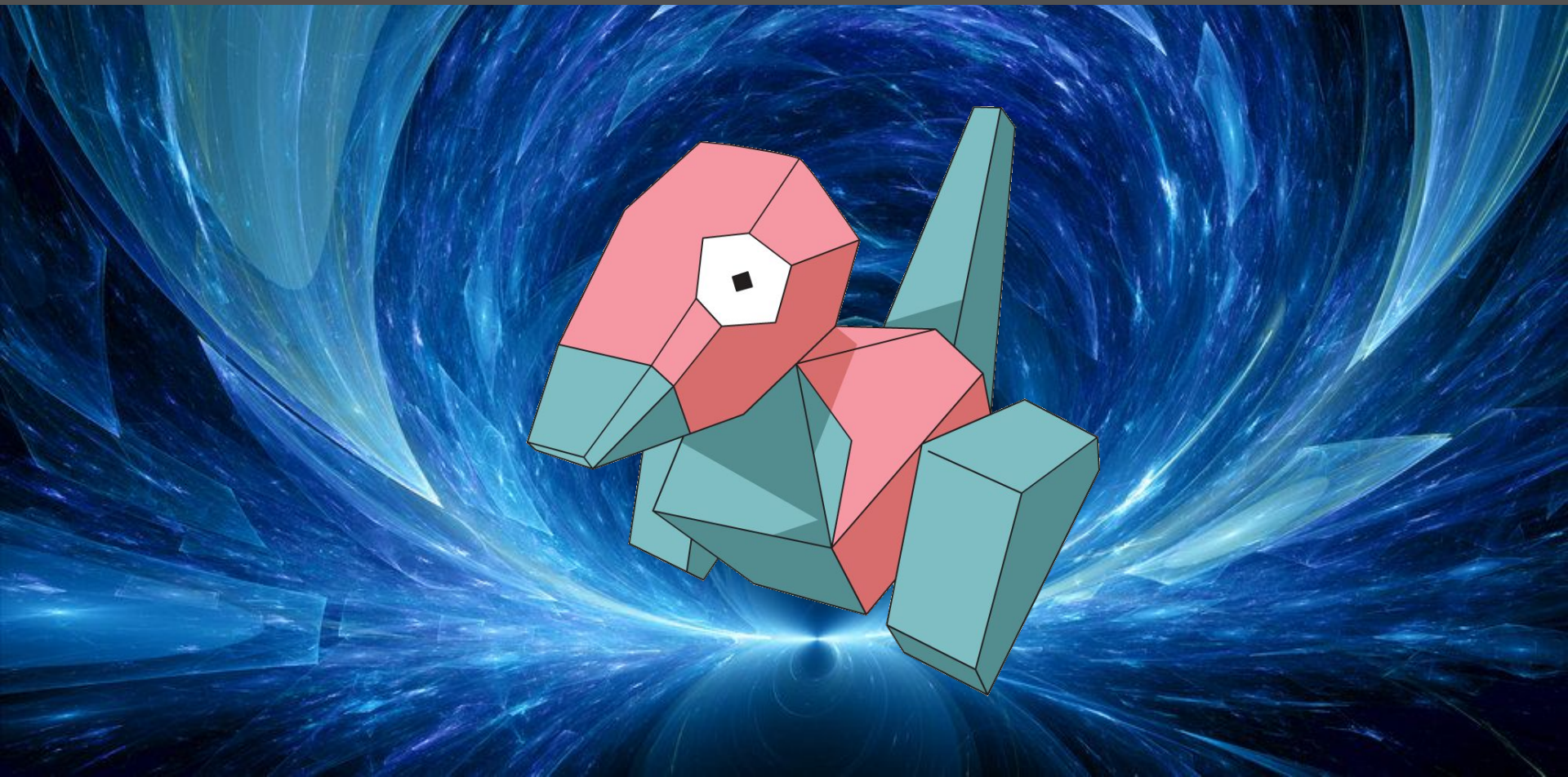
- Relief

Variable Ranking Techniques

Unsupervised:

- variable entropy

Curse of Dimensionality

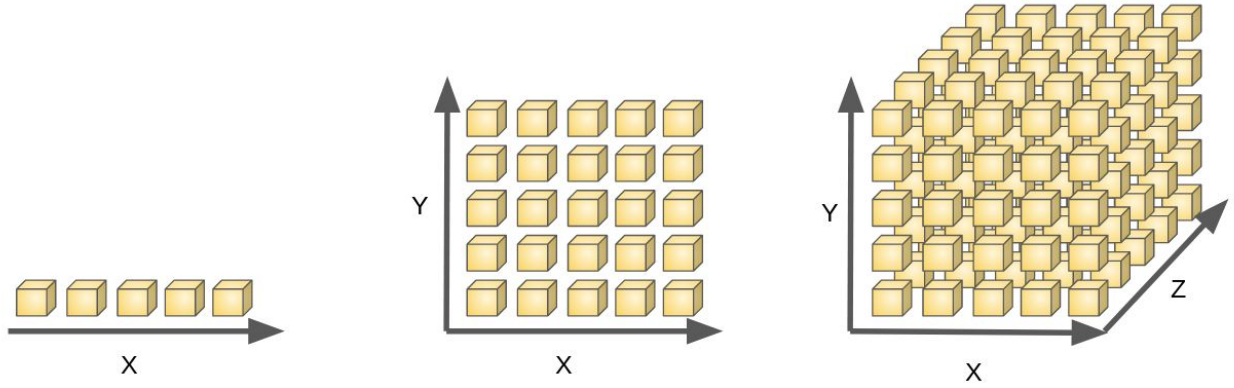


Curse of Dimensionality

- 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

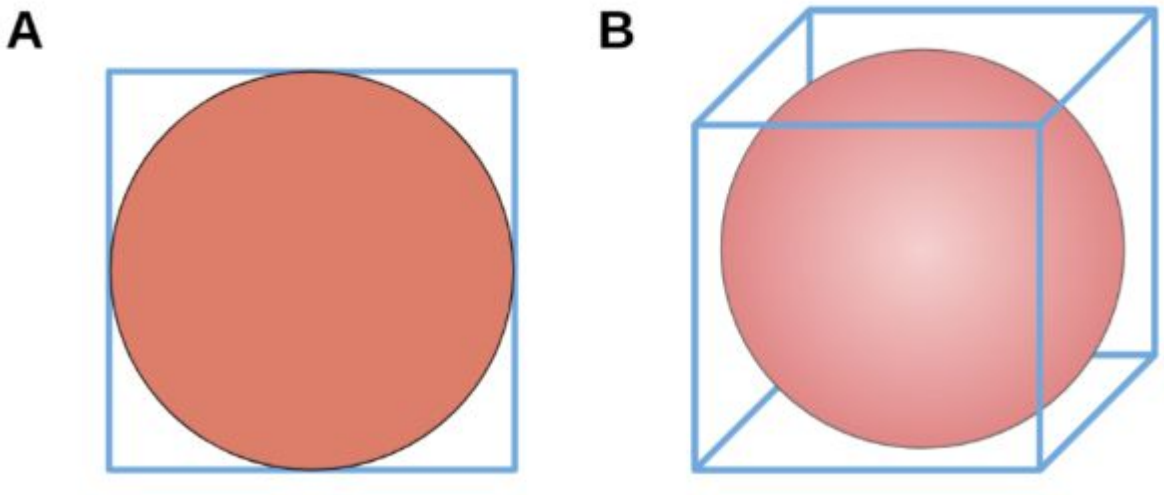
Combinatorial explosion

- Each combination of possible values must be considered

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

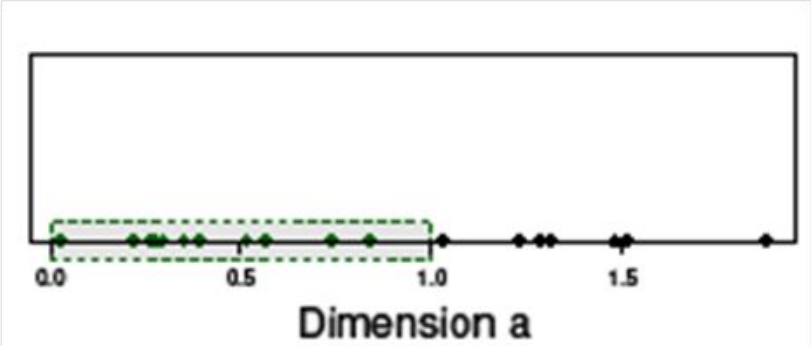
Curse of Dimensionality

Distance concentration

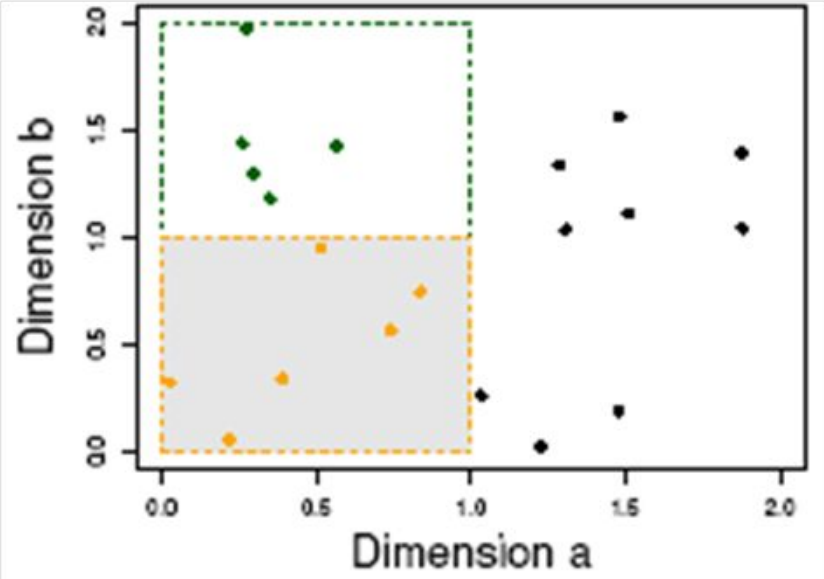


Curse of Dimensionality

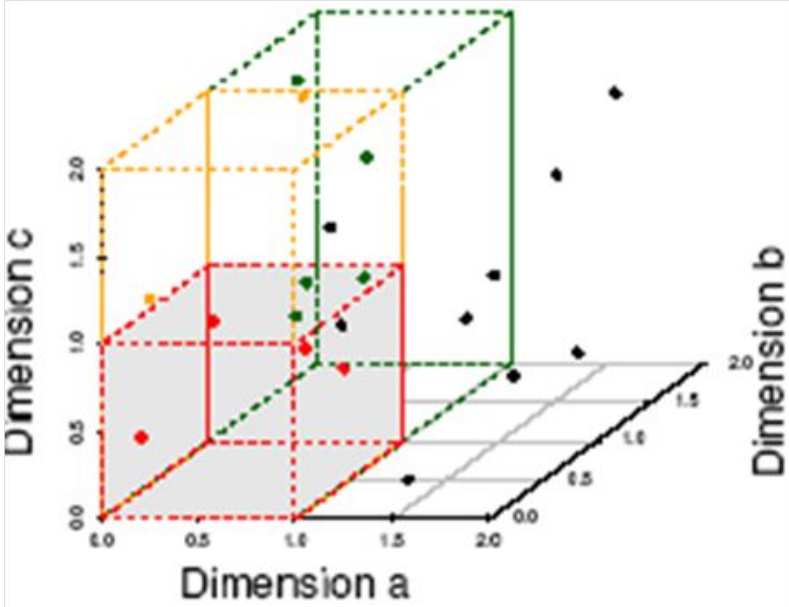
$$\frac{\frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)}}{2r^n} = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma(\frac{n}{2}+1)}$$



(a) 11 Objects in One Unit Bin



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

Curse of Dimensionality

Distance concentration

Draft

{ Distance concentration: with increasing dimensionality pairwise distances may converge to the same value (lack of contrast). Since many data analysis machine learning techniques base on distances this may be problematic.

Tip: The article An Introduction to Variable and Feature Selection by Isabelle Guyon and

Andre Elissee
, JMLR 3 (2003) pp. 1157-1182 constitutes a rich overview paper. There's a link

in the Assignments Wiki.

Distance concentration is also named as "concentration phenomenon". As dimensionality grows, differences between vectors using usual metric tends to be constant.

b) As dimensions increase, “contrast-loss” [3, 10, 17] occurs. Distances between points tend to a constant, with traditional clustering metrics becoming ill-defined [3, 5]. This is considered part of the curse of dimensionality [17, 1].

https://arxiv.org/pdf/1804.02624.pdf

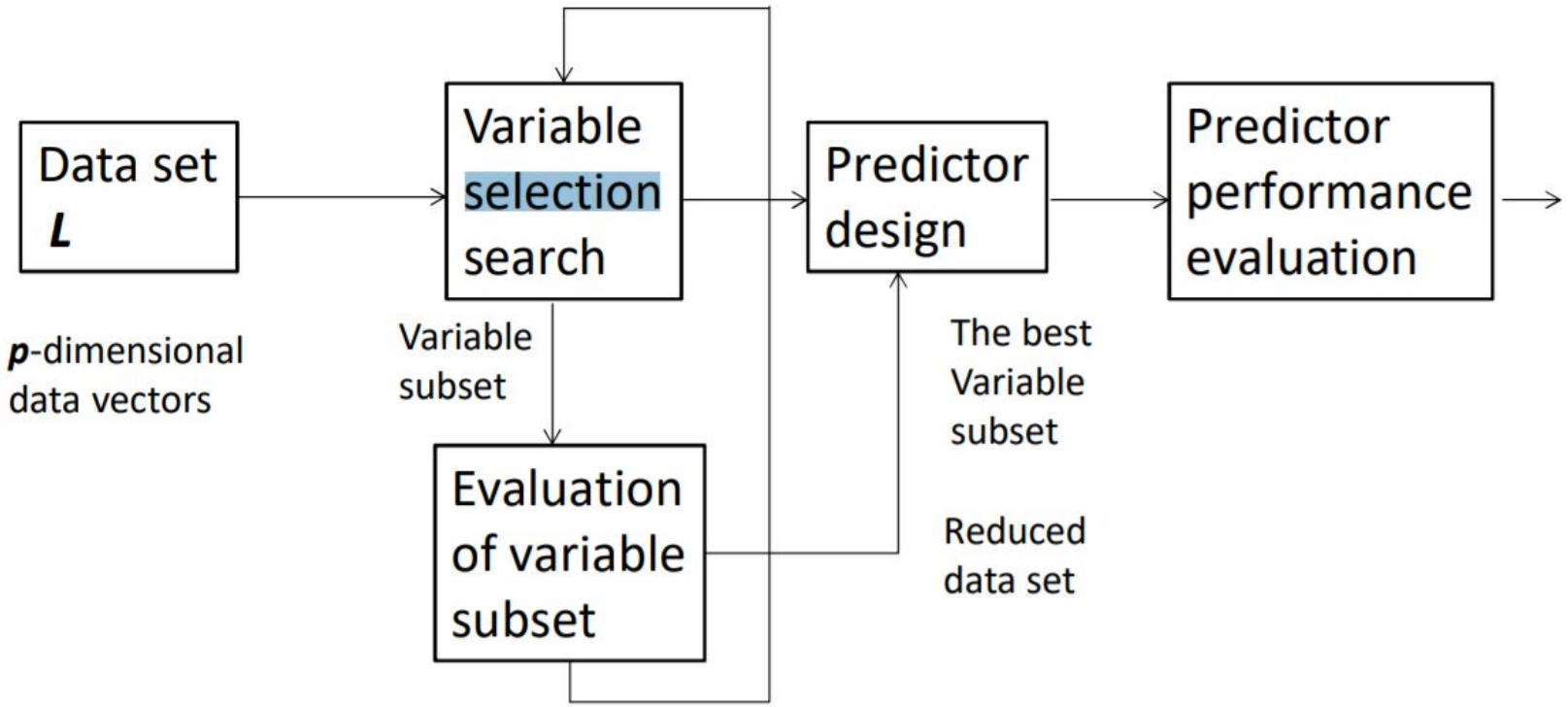
What was previously interpreted as “contrast-loss” is actually the law of large numbers causing instances of a distribution to concentrate on a thin “hypershell”. The hollow shells mean data points from apparently overlapping distributions do not actually mingle, making choatic data intrinsically separable

3-https://bib.dbvis.de/uploadedFiles/155.pdf

10-https://members.loria.fr/moberger/Enseignement/Master2/Exposes/beyer.pdf

We show that under certain broad conditions (in terms of data and query distributions, or workload), as dimensionality increases, the distance to the nearest neighbor approaches the distance to the farthest neighbor. In other words, the contrast in distances to different data points becomes nonexistent.

Approaches: Filter Methods



Approaches: Filter Methods

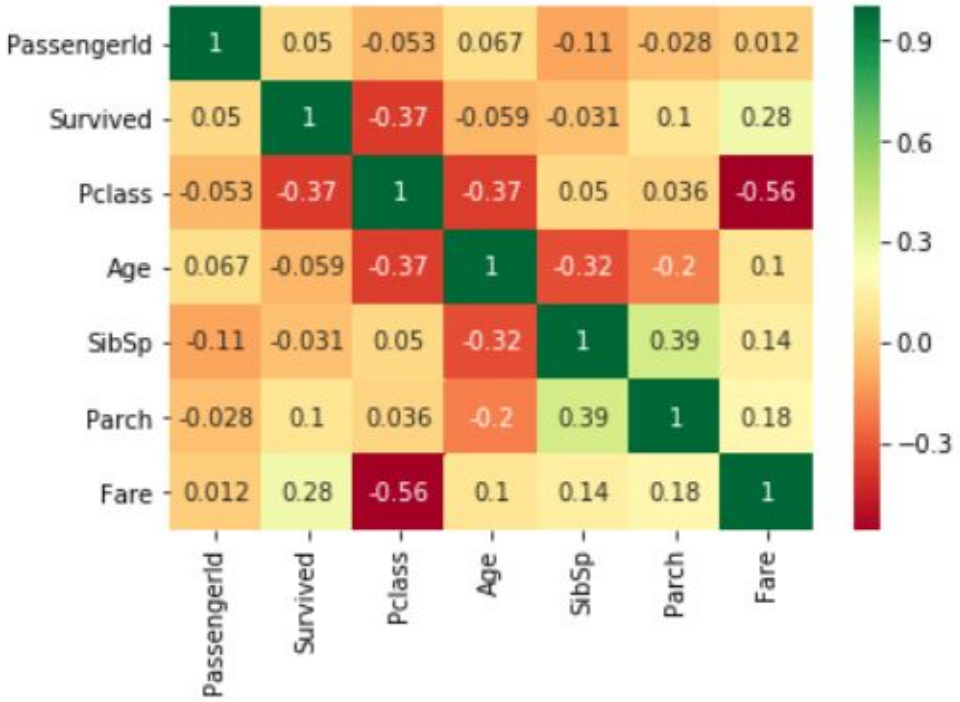
- Use
  - Intrinsic properties of the data
  - Statistic methods: chi-square, ANOVA, Correlation
- Calculate
  - Subset of the variables based on thoose methods
- Rank
  - The variables according to a certain result

Approaches: Filter Methods

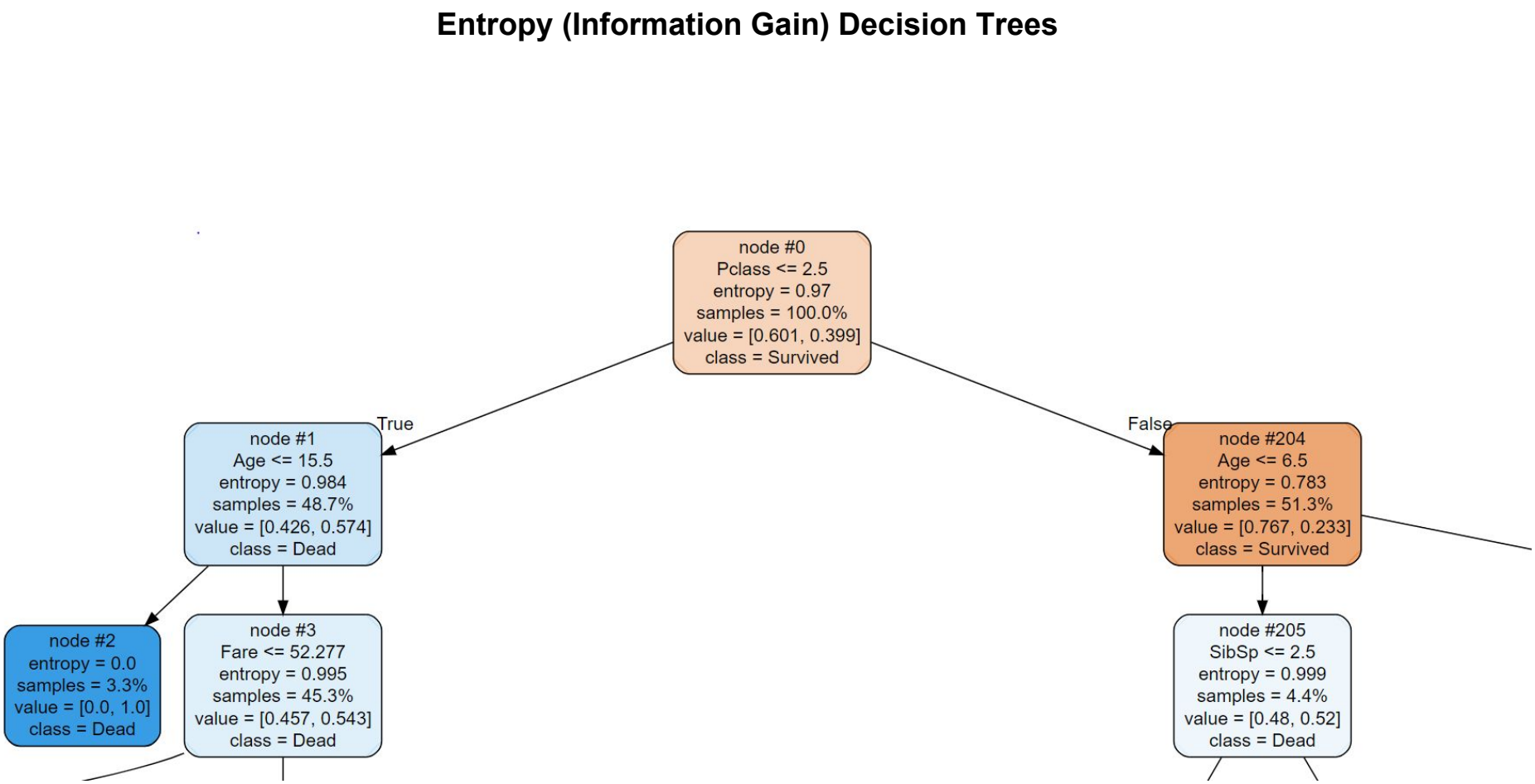
ANOVA

```
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
```

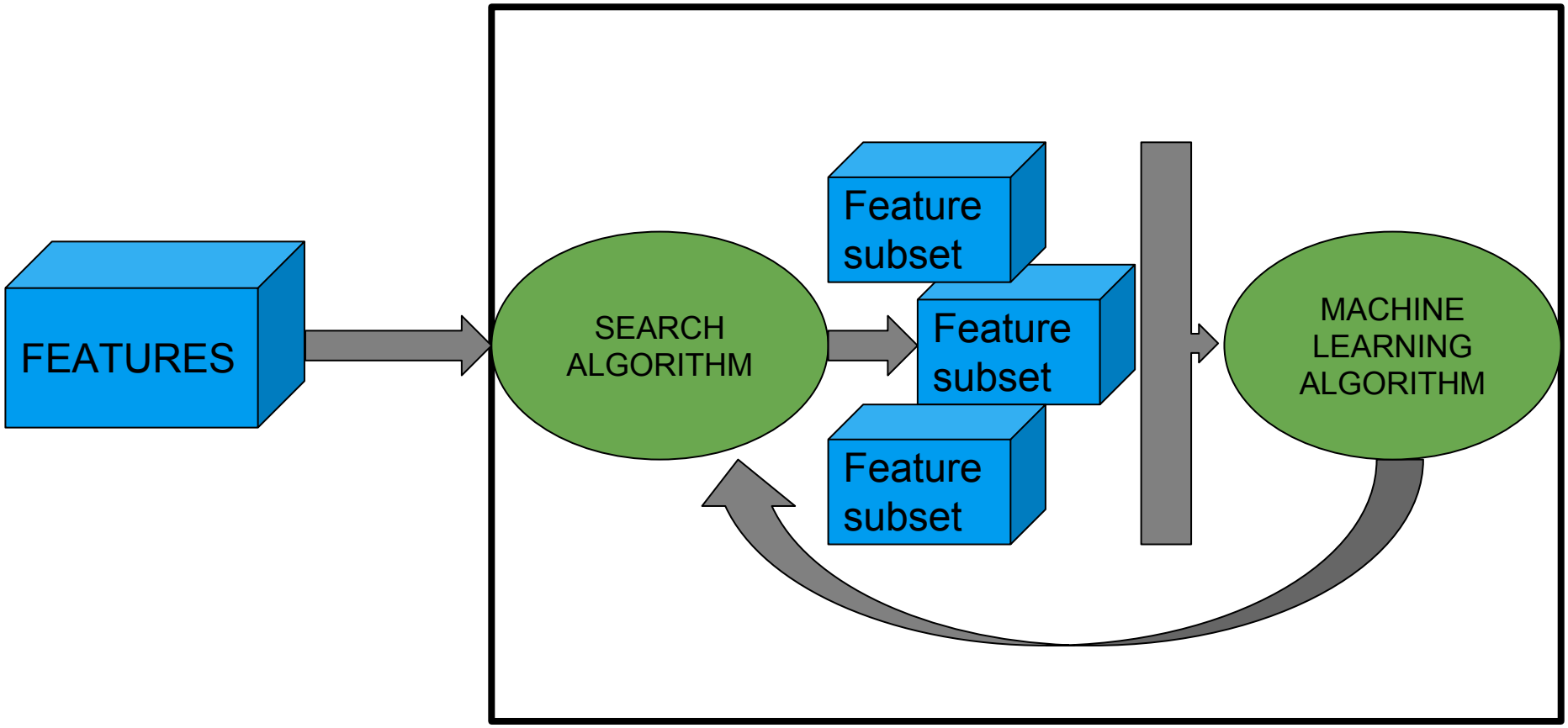
CORRELATION



Approaches: Filter Methods



Approaches: Wrapper

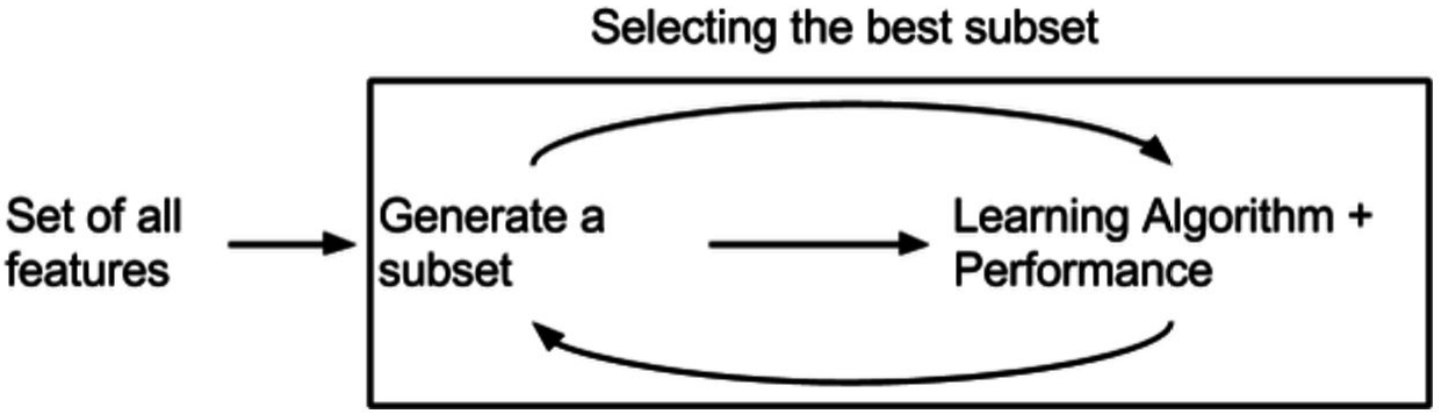


Approaches: Wrapper

- 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."



Example: Iterated Local Search

Algorithm: Genetic

Classifier: Support Vector Machine

Evaluation: Classification accuracy (tenfold)

Approaches: Embedded

Advantages

Disadvantages

university of groningen

Questions