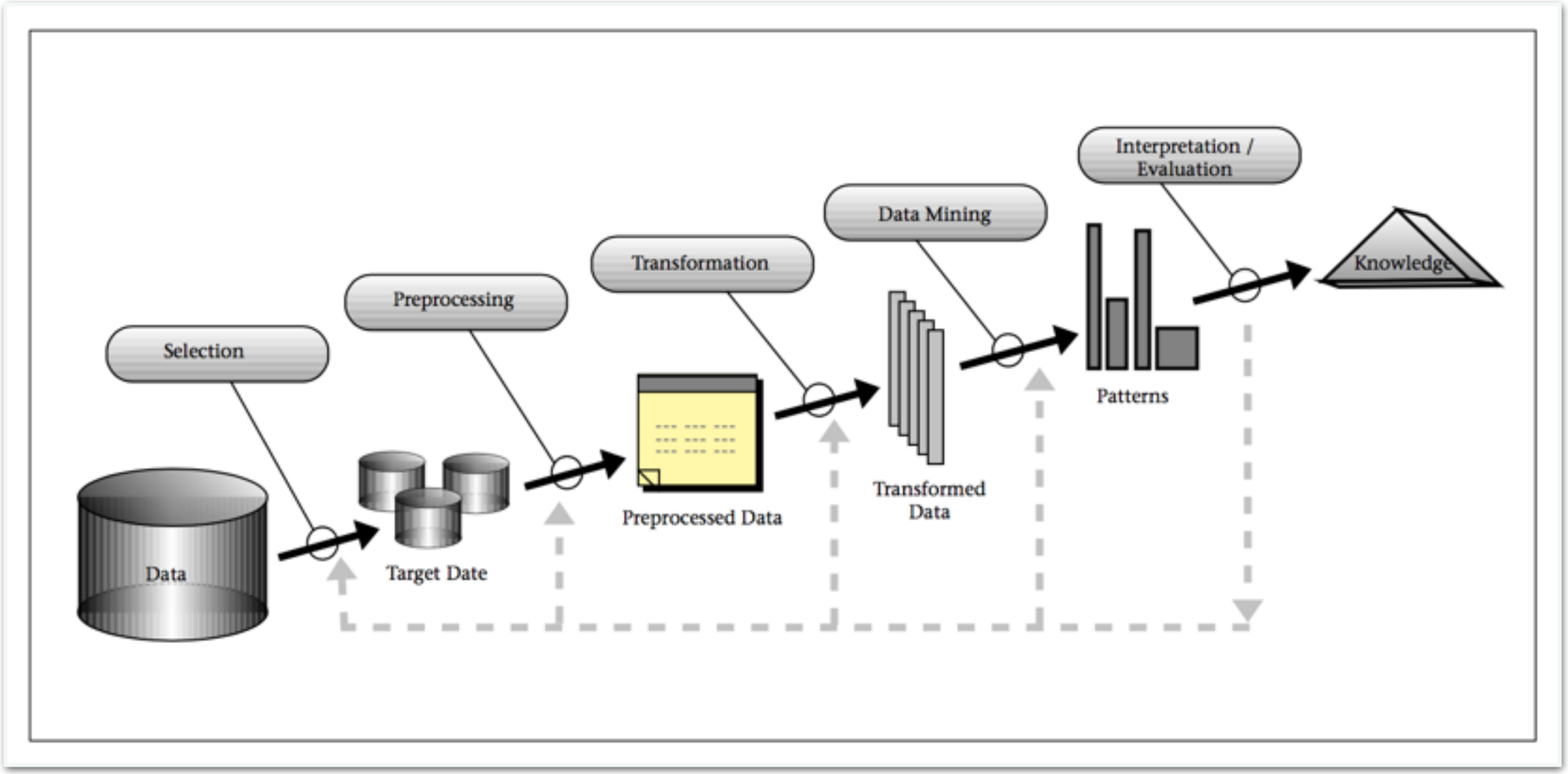
Data mining

How to extract informations from data?

- Lesson Data Mining



The **knowledge discovery in databases (KDD) process** is commonly defined with the stages:
(1) Selection
(2) Pre-processing
(3) Transformation
(4) *Data mining*
(5) Interpretation/evaluation.[5]

- Current hardware and database technology allow efficient and inexpensive reliable data storage and access

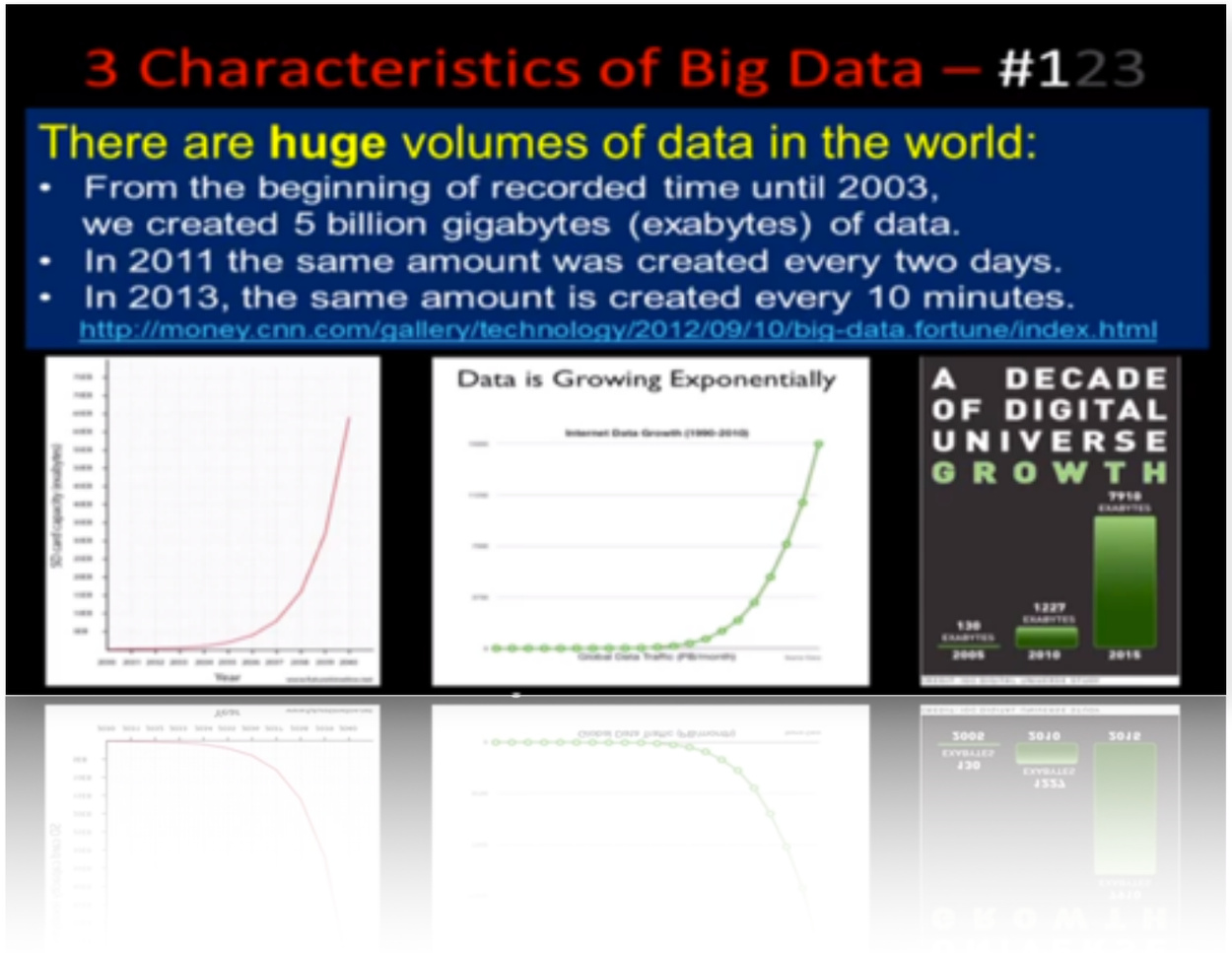
- Databases are increasing in size in two ways: the number N of records, or objects, in the data- base, and the number d of fields, or attributes, per object. Databases containing on the order of N = 10^9 objects are increasingly common in, for example, the astronomical sciences. The number d of fields can easily be on the order of 10^2 or even 10^3 in medical diagnostic applications.

- However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. COMMUNICATIONS OF THE ACM November 1996/Vol. 39, No. 11

Background

- The amount of data stored within informatic systems is growing exponentially



Problem

- Traditional techniques of data analysis could become ineffective

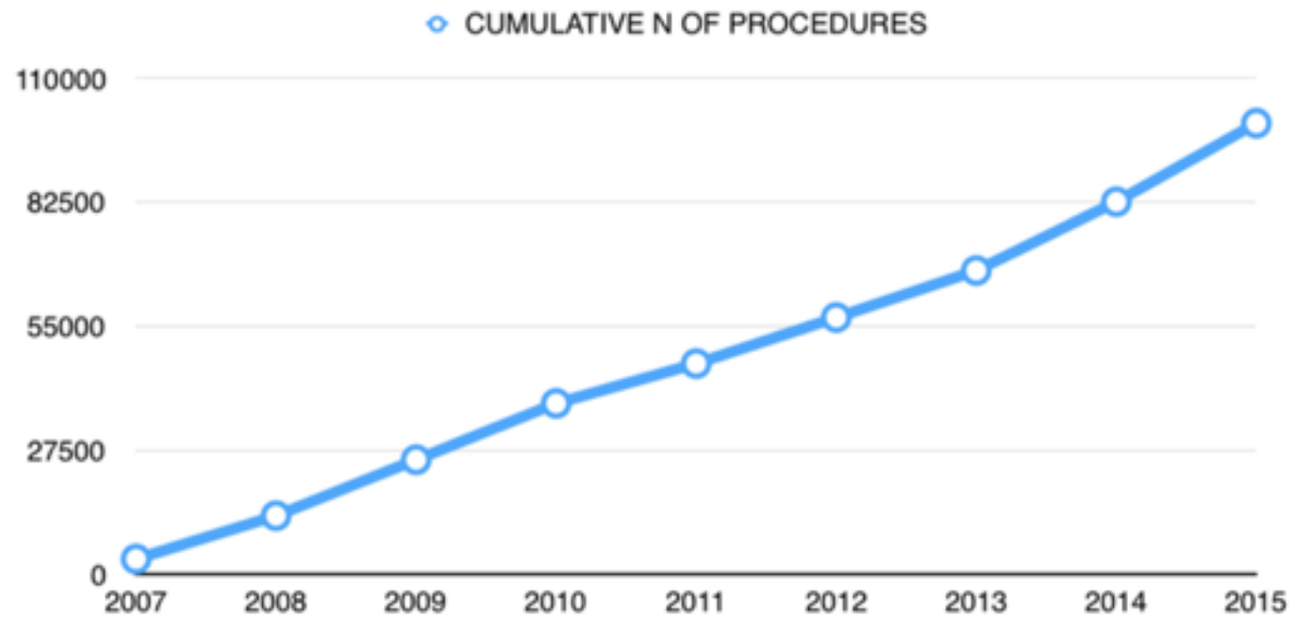
- DATA MINING could help researchers at classifying and clustering data as well as making hypothesis

Solution?

Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

YEARS	2007	2008	2009	2010	2011	2012	2013	2014	2015
N OF PROCEDURES	3433	13018	25441	37929	46759	56959	67373	82608	100043



THE DATA GAP

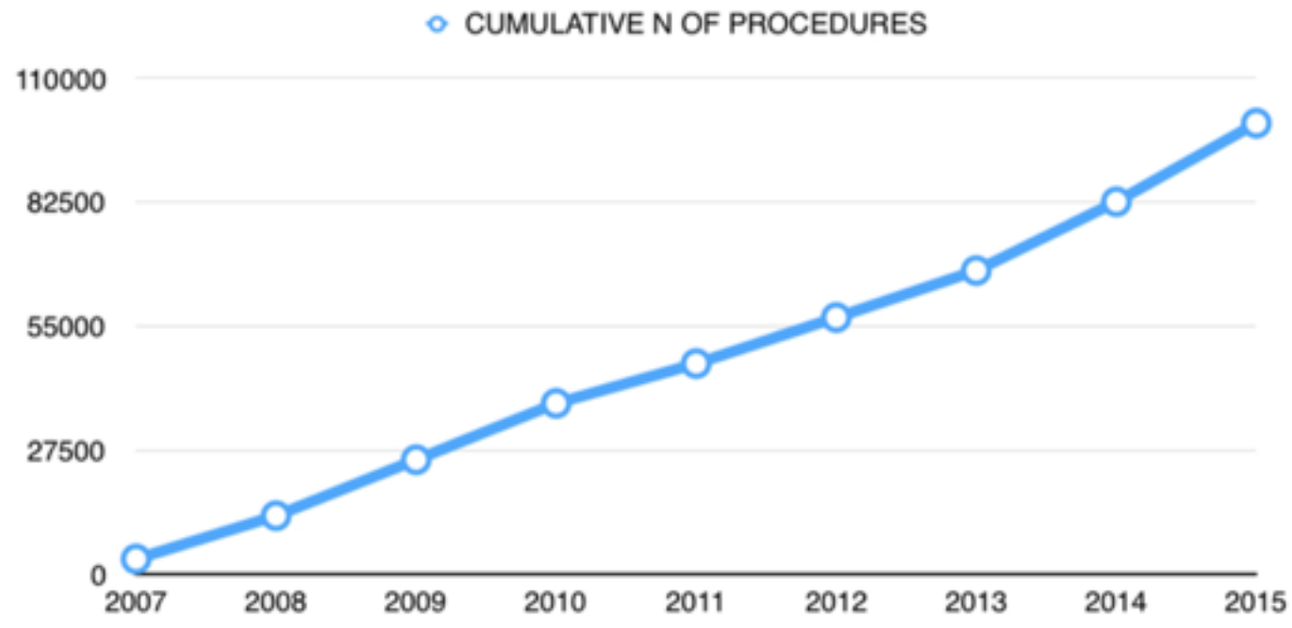
- ESTS Registry
- Growth not exponential
- In ten years 40 times larger amount of data
- Same analytic procedures

1. Many informations contained within a registry are not clearly evident

2. The analytic process for extracting knowledge could require a long time to be carried out

3. A large amount of data could remain not explored nor analyzed

YEARS	2007	2008	2009	2010	2011	2012	2013	2014	2015
N OF PROCEDURES	3433	13018	25441	37929	46759	56959	67373	82608	100043



NOTHING IN COMPARISON WITH BIG DATA

THAT ARE CHARACTERIZED BY

- NUMBER
- VELOCITY
- VARIETY

DATA MINING

1. EXTRACTING IMPLICIT UNKNOWN INFORMATIONS WITH POTENTIAL USEFULNESS (CLINICAL RELAPS) FROM DATA

2. DATA EXPLORATION AND ANALYSIS ARE PERFORMED BY AUTOMATIC OR SEMI-AUTOMATIC SYSTEMS ON LARGE AMOUNT OF DATA IN ORDER TO DISCOVER SIGNIFICATIVE PATTERNS

Data mining involves six common classes of tasks:[5]

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

ELSEVIER

European Journal of Cardio-thoracic Surgery 31 (2007) 518–521

EUROPEAN JOURNAL OF CARDIO-THORACIC SURGERY

www.elsevier.com/locate/ejcts

Measured FEV1 in the first postoperative day, and not ppoFEV1, is the best predictor of cardio-respiratory morbidity after lung resection☆

Gonzalo Varela [a,*], Alessandro Brunelli [b], Gaetano Rocco [c], Nuria Novoa [a], Majed Refai [b], Marcelo F. Jiménez [a], Michele Salati [b], Tindaro Gatani [c]

[a] Service of Thoracic Surgery, Salamanca University Hospital, 37007 Salamanca, Spain
[b] Unit of Thoracic Surgery, "Umberto I°" Regional Hospital, Ancona, Italy
[c] Division of Thoracic Surgery, National Cancer Institute, Naples, Italy

Received 26 September 2006; received in revised form 24 November 2006; accepted 27 November 2006; Available online 22 December 2006

Abstract

Introduction and objective: There is a low correlation between predicted postoperative FEV1 (ppoFEV1) and FEV1 measured the days after pulmonary resection, when most complications are developed. The hypothesis of this investigation is that ppoFEV1 does not predict postoperative morbidity in patients undergoing lung resection when immediate postoperative FEV1 is considered in the predictive model. **Methods:** One hundred ninety-eight consecutive patients undergoing lobectomy or pneumonectomy were included in a prospective, multiinstitutional study. Independent variables: age, body mass index, ppoFEV1, surgical approach (VATS or muscle-sparing thoracotomy), type of analgesia (epidural or intraveous), postoperative visual analogue pain score and FEV1 measured the day after the operation. Target variable: occurrence of postoperative cardio-respiratory complications. Method of analysis: classification tree (CART) dividing the population at random in two subsets and developing a bootstrap set of 100 trees resampling training data. The relative importance of each variable and the accuracy of both initial and committee trees to predict the outcome were presented. **Results:** One hundred seventy-seven lobectomies and 21 pneumonectomies were included. Overall cardio-respiratory morbidity was 22%. According to CART results, first day FEV1 was the most important variable to classify cases as primary splitter and as a surrogate of each primary splitter (100% importance). Patient age followed (51%) and ppoFEV1 was third (43%) with a score similar to postoperative pain score (42%) and type of analgesia (36%). Sensitivity and specificity of the initial tree were, respectively, 0.5 and 0.7; values for committee tree were 0.5 sensitivity and 0.7 specificity. **Conclusion:** Postoperative cardio-respiratory complications are more related to FEV1 measured in the first postoperative day than to ppoFEV1 value.

Keywords: Thoracic surgical procedures; Lung volume measurements; Postoperative care; Postoperative pain; Classification and regression trees

CLASSIFICATION AND REGRESSION TREE: an example

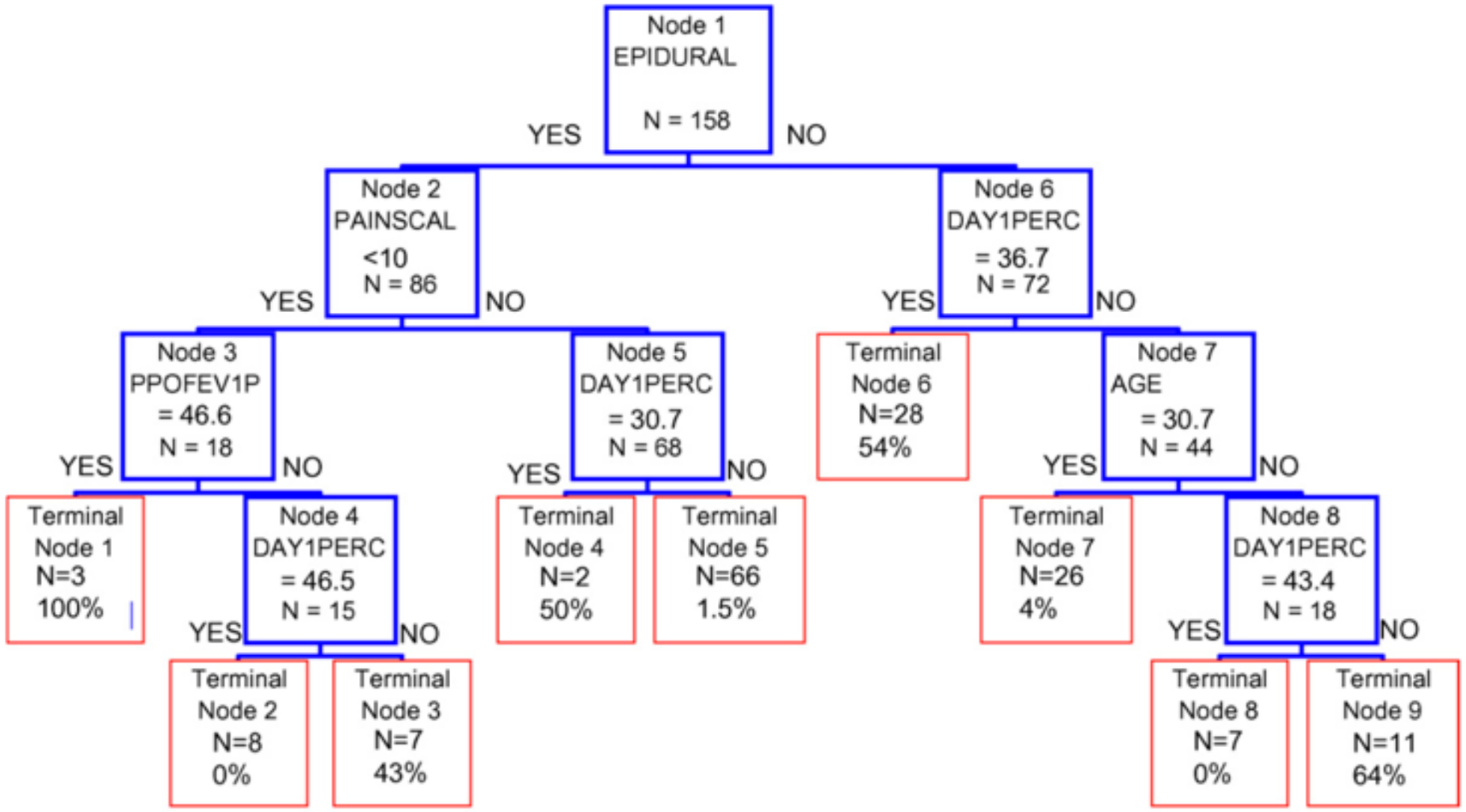


Fig. 2. Initial classification tree. Each node includes the independent variable, its cut value (only for continuous ones) and the number of cases. Figures in percentages represent the rate of cases with complications in the subset. Abbreviations: EPIDURAL: postoperative epidural analgesia; PAINSCAL: pain scale; DAY1PERC: FEV1% on first postoperative day; PPOFEV1P: estimated postoperative FEV1%.