Lambton College

School of Computer Studies

Lecture 2

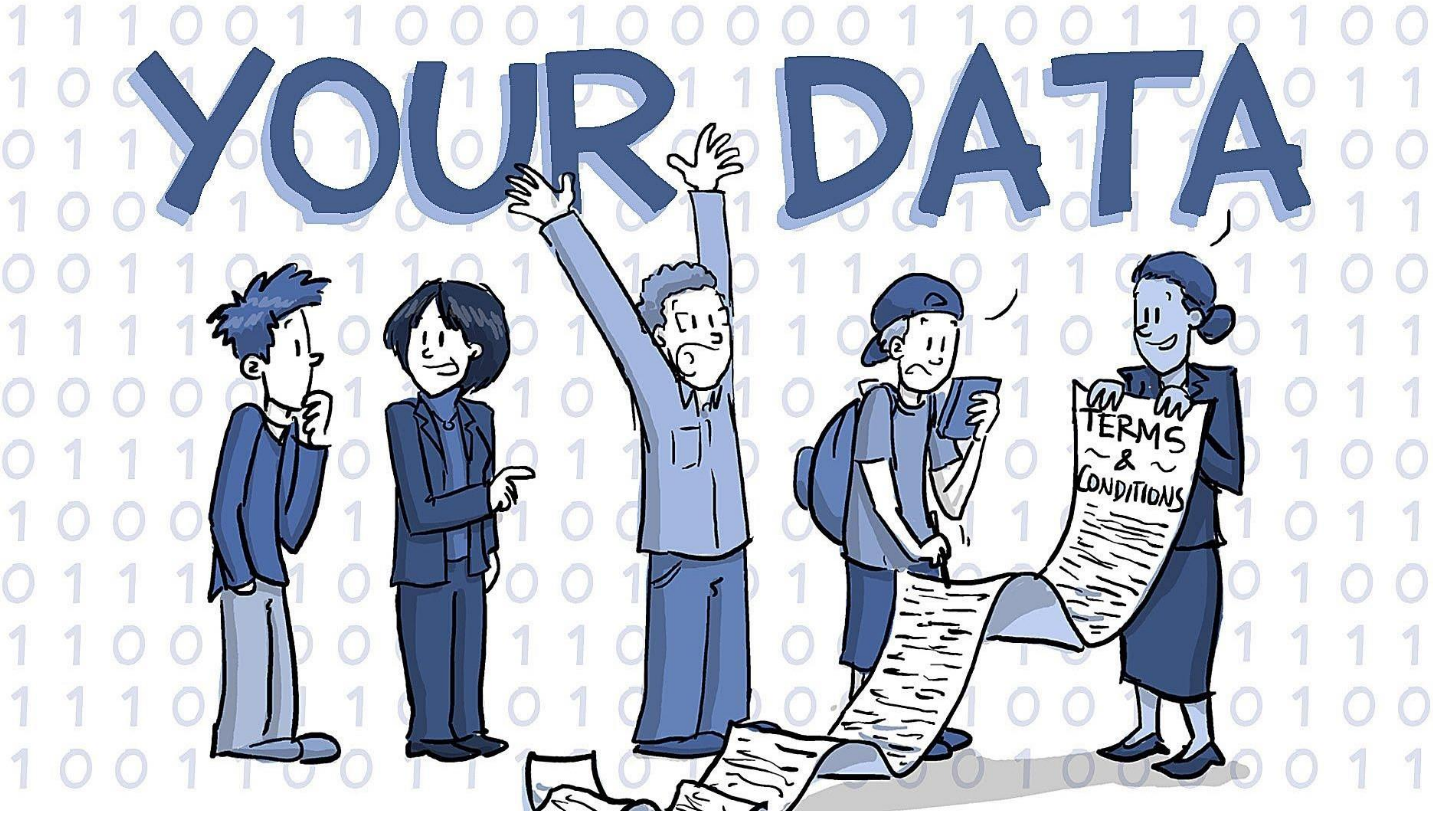
CBD-3335 Data Mining and Analysis

Learning Outcomes

- Data Taxonomies.
- Data sources, and types
- Textual Data Challenges.
- Feature Selection.
- Theory of Measurements

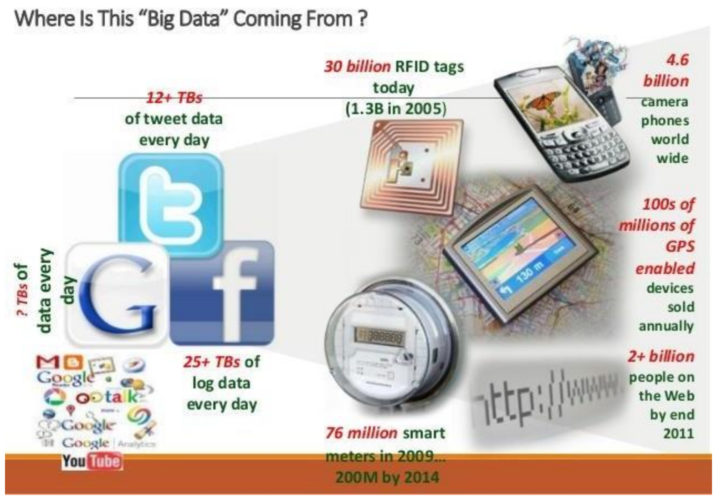
Data Taxonomies

- Categorizing data from different aspects
  - Data source
  - Data type
  - Structure
  - Time
  - Dimensionality
  - Quality



Data Sources

Where data comes from ?

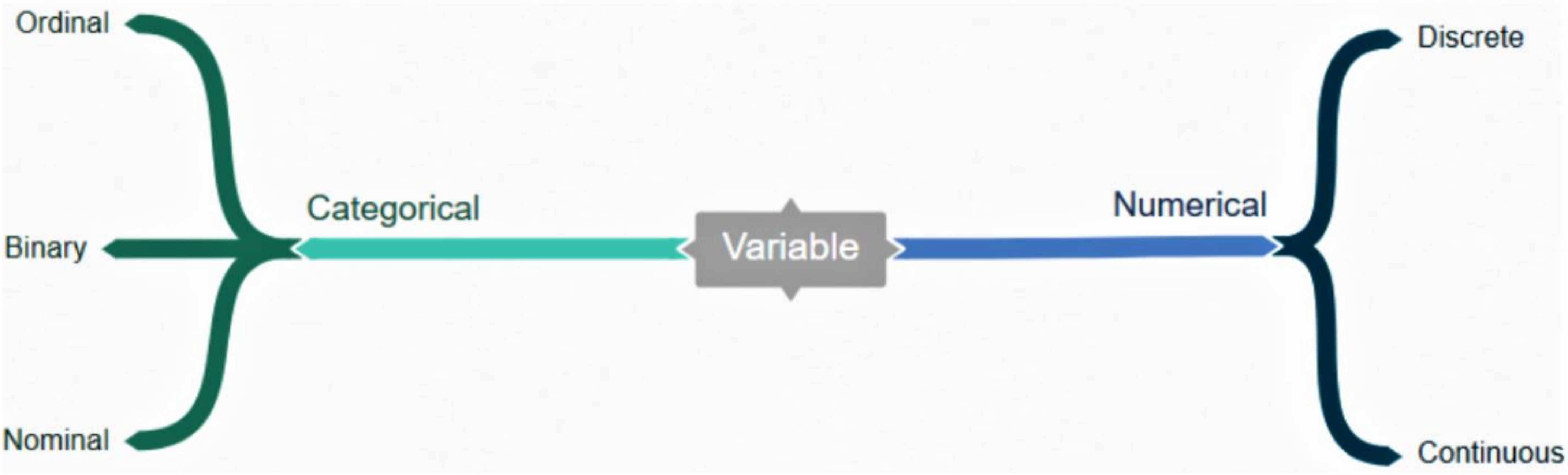

Where Is This "Big Data" Coming From ?
12+ TBs of tweet data every day
30 billion RFID tags today (1.3B in 2005)
4.6 billion camera phones world wide
? TBs of data every day
25+ TBs of log data every day
100s of millions of GPS enabled devices sold annually
76 million smart meters in 2009… 200M by 2014
2+ billion people on the Web by end 2011
http://www.
You Tube

- Database and data warehouses: Queries
- Sensory data (usually real-time): temperature data
- Data entry: questionaries' data, data surveys
- Online data: data from other computers
- Embedded data: data from computers inside other devices such as mobile data
- Web data: data collected from web resources
- User-generated data: Content generated by user

Data Types

Data

- Numerical
- Categorical



Data Types

Numerical data is information that is measurable and represented as numbers. It can be

a) Discreate (Interval), or

b) Continuous (Real, Ratio).

1. Discreate: Numerical data that have a logical end. Examples: Variables for days in the month, or number of bugs logged.

2. Continuous : Numerical numbers that don’t have a logical end.

Examples: Variables that represent money or height.

Data Types

Categorical data is any data that isn't number; which can mean a string of text or date. It can be mainly

a) Ordinal, or

b) Nominal.

1. Ordinal: Categorical data that have a set order to them.
   Examples: Having a priority on a bug such as "Critical" or "Low" or the ranking of a race as "First" or "Third".

2. Nominal: represent values with no set order to them.

Examples: Variables such as "Country" or "Marital Status".

Data Types

Binary data a special type of categorical data type having only two values – yes or no.

- This can be represented in different ways such as “True” and “False” or 1 and 0.
- Often used to represent one of two conceptually opposed values, e.g: the outcome of an experiment ("success" or "failure")

- Binary data occurs in many different technical and scientific fields, where it can be called by different names:
  - "bit" (binary digit) in computer science,
  - "truth value" in mathematical logic and related domains,
  - "binary variable" in statistics.

Structured & unstructured data

Structured data:

- Data that can be stored in a tabular form
- Every instance has the same structure
- Can be easily stored, organized, searched, recorded and merged with other structured data.
- Suitable for integration into an analytics records.

Example: The demographic data for a population where each row in the table describe one person (attributes: name, age, date of birth, gender, address, education, employment status etc.)

Structured & unstructured data

Unstructured data:

- Structure of data might not necessarily the same in every instance
- Each instance might have its own internal structure
- More common data type in real world; email tweets, text, posts, image, music, video, input from sensors etc. can be some examples.
- Difficult to analyze due to variation in structure.

Example: Dataset of webpages; each website might have data of an unique type).

Semi-structured data

- Most XML data

- Wordnet:

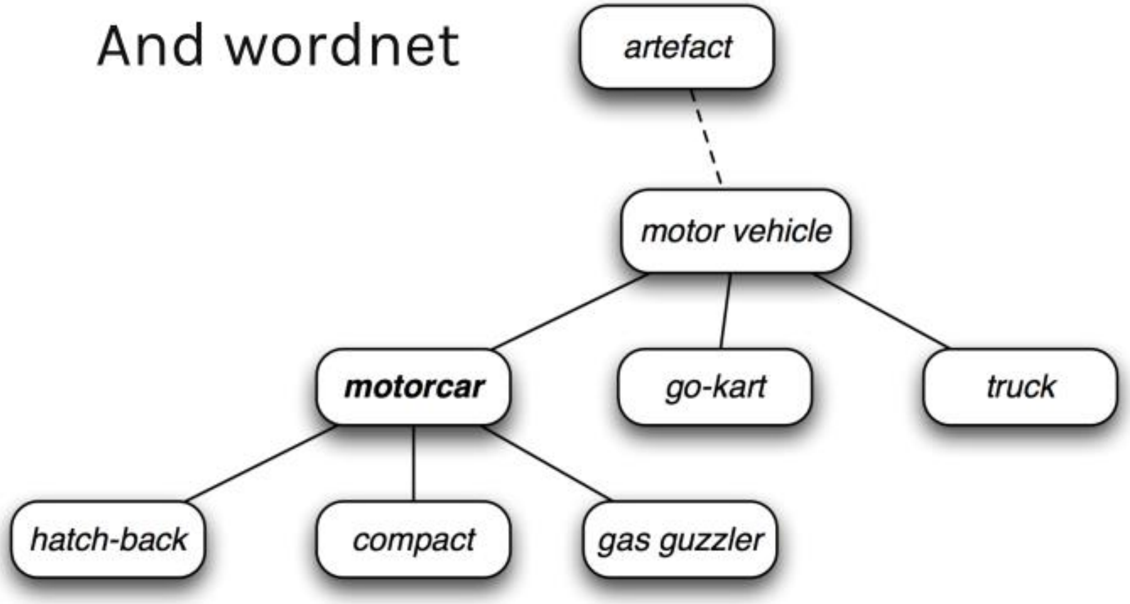
  - WordNet is a lexical database of semantic relations between words

  - WordNet links words into semantic relations including synonyms, hyponyms, and meronyms.

python™

Natural Language Analyses
with NLTK

And wordnet



Time

- Temporal data: financial data, twitter streaming data
  - Real-time data: time sensitive, if we miss reading any data, the consequence might be disastrous.
  - Non-time sensitive: We may miss some data without any dramatic consequences.
  - If data is numeric in type, data is a time series.
  - Static data: Fingerprint or biometric data

Dimension

- One dimensional:
    - body temperature, crime index, financial data
- Two-dimensional:
    - Image data (nxn matrix of pixels)
- N- dimensional:
    - Demographic data (age, height, weight, eye-color, race, DOB, POB, gender, occupation, …..)
- High dimensional:
    - Text data, Gene-expression data

Quality (1)

- Good quality data: Twitter of COVID-19 about recent Pandemic.
- Noise:
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Outlier: IQ>160 when collecting IQ of high school students in public schools
- Inconsistent: (salary = **-5000$**)
  - containing discrepancies in codes or names

Quality (2)

- Twitter data example
- Incomplete: A broken tweet, (John, 21, male, ??, 160 lb, American, ??)
  - lacking attribute values, lacking certain attributes of interest, or containing only aggregated data
- Missing: Some missing tweets because of rate of sampling
- Duplicate: Many copies of a single tweet
- Irrelevant: Lady GAGA concert in Chicago

Quality (3)

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Textual Data Challenges

- Information is in **unstructured** textual format
- **Large** textual database
- **Very high** number of possible "**dimension**s" (but sparse):
  - all possible words and phrase types in the language!!
- **Complex** and subtle **relationships** between concepts in text
  - "AOL merges with Time-Warner" "Time-Warner is bought by AOL"

Textual Data Challenges

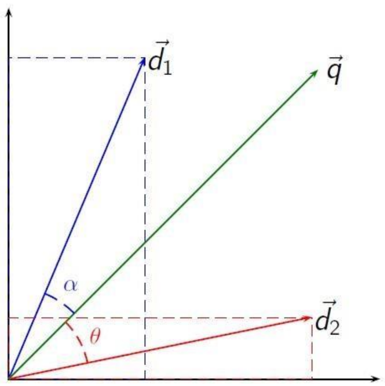
- **Word ambiguity** and context sensitivity
  - automobile = car = vehicle = Toyota
  - Apple (the company) or apple (the fruit)
- **Noisy data**: Spelling mistakes

Features (1)

- The piece of input data for which an output value is generated is formally called an instance.
  - Record, instance, object, observation, data
- The instance is formally described by a vector of features, which together constitute a description of all known characteristics of the instance.
  - Field, feature, variable, measurement
- The feature vectors can be seen as defining points in an appropriate multidimensional space.

Features (2)

- Vector methods can be correspondingly applied to them, such as computing the dot product or the angle between two vectors.
- Vector space model: Mostly in text data but can be used in other Pattern Recognition and data mining tasks.
  - Every data is a vector in a multi (high) dimensional space



Type of Features (1)

- Categorical aka nominal: consisting of one of a set of unordered items
  - Such as a gender of "male" or "female", or a blood type of "A", "B", "AB" or "O"
- Ordinal: consisting of one of a set of ordered items
  - Such as "large", "medium" or "small"
- Integer-valued
  - Such as a count of the number of occurrences of a particular word in an email

Type of Features (2)

- Real-valued
  - Such as a measurement of blood pressure
- Often, categorical and ordinal data are grouped together; likewise for integer-valued and real-valued data.
- Many algorithms work only with categorical data, such Naïve Bayes Classifier. How can we use numerical features? and require that real-valued or integer-valued data be discretized into groups (e.g., less than 5, between 5 and 10, or greater than 10).

Feature Selection (1)

- When do we employ feature selection?
  - For **very high dimensional data**, in which feature extraction might be expensive
  - **Features are not numeric**
  - **We are looking for meaningful features**

$$y_j = a_{j1}x_1 + a_{j2}x_2 + \square + a_{jm}x_m$$

Feature Selection (2)

- Feature Selection
  - Searching the feature space for a subset of features maximizing an objective function (quality index)
  - Wrappers
  - Filters
    - Feature ranking
  - Embedded
  - Markov Blanket

Feature Selection (3)

- Search strategy: search the power set of the feature set to find the optimum feature subset
  - Exhaustive search: the order of the search space is $O(2^m)$
  - Search strategy to reduce the size of the search space
    - Sequential Forward Selection (SFS)
    - Sequential Backward Selection (SBS)
    - Beam search
    - Simulated annealing

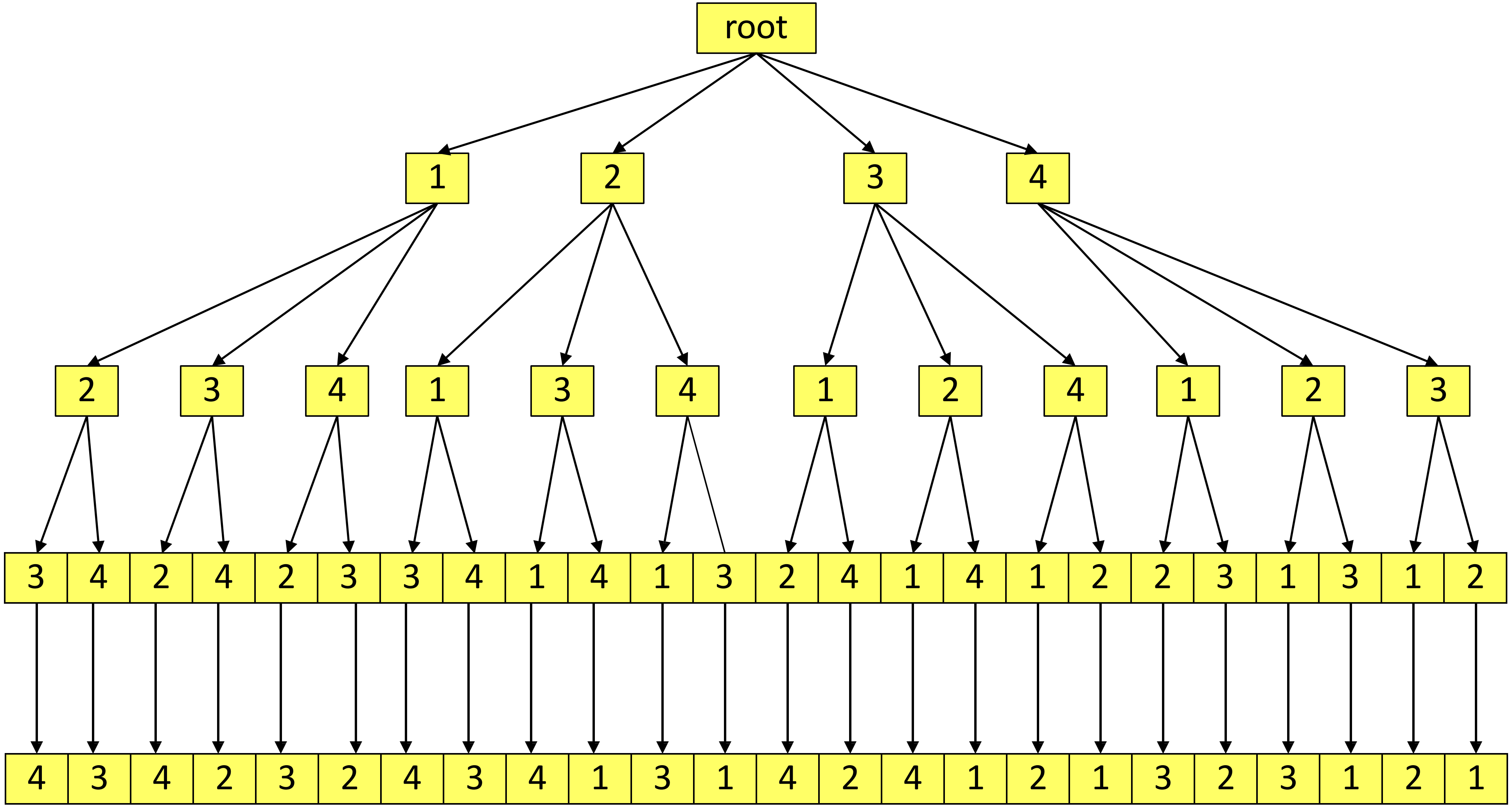
Search Strategies (2)

- Sequential Forward Selection (SFS)
  1. Start with empty set: Y← { }
  2. Select the next best feature that maximizes the objective function of the selected features

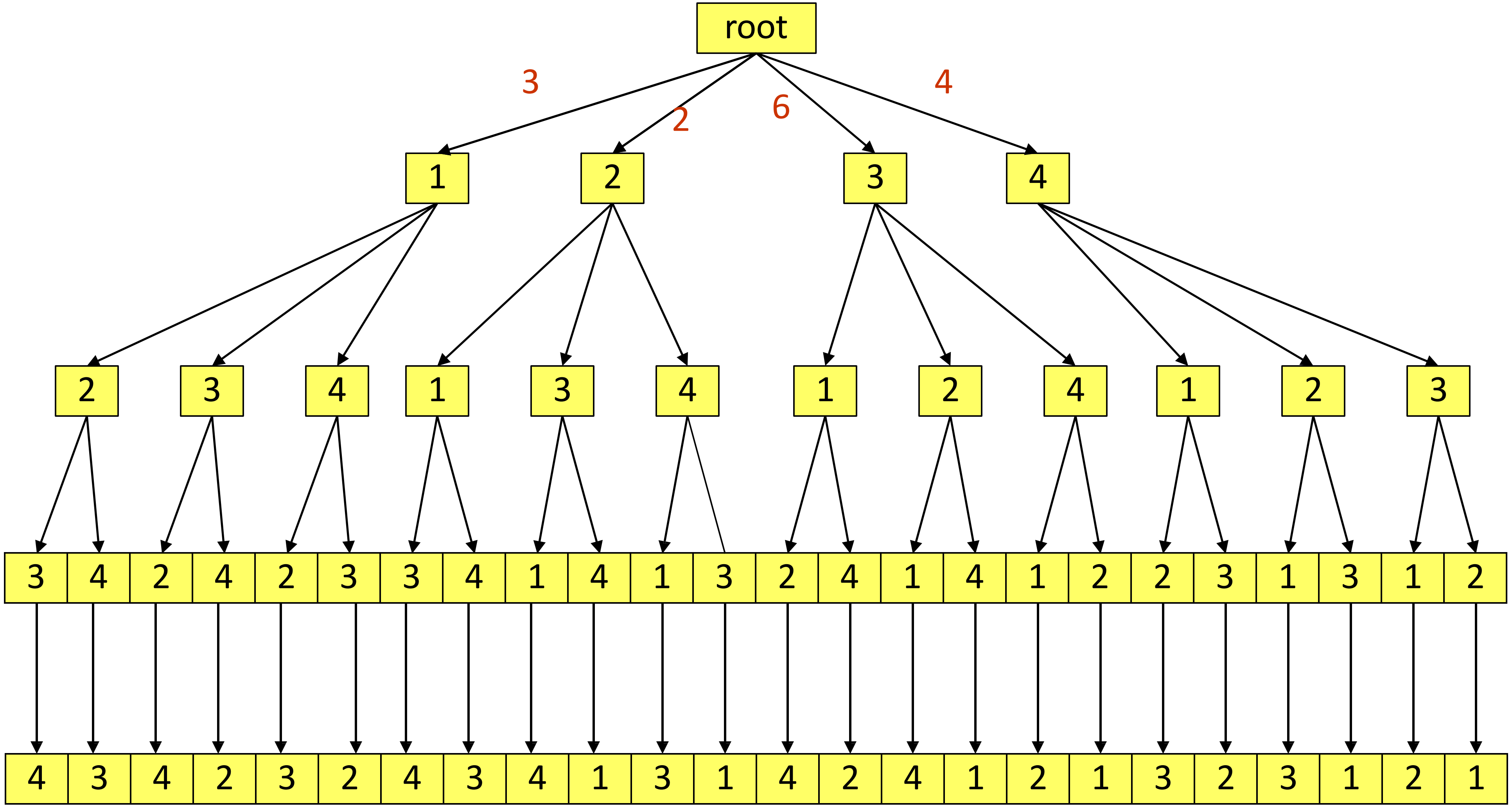
$$z \leftarrow \arg\max_{x \notin Y} [h(Y + \{x\})]$$

  3. Update Y: $Y \leftarrow Y + \{z\}$
  4. Go to 2

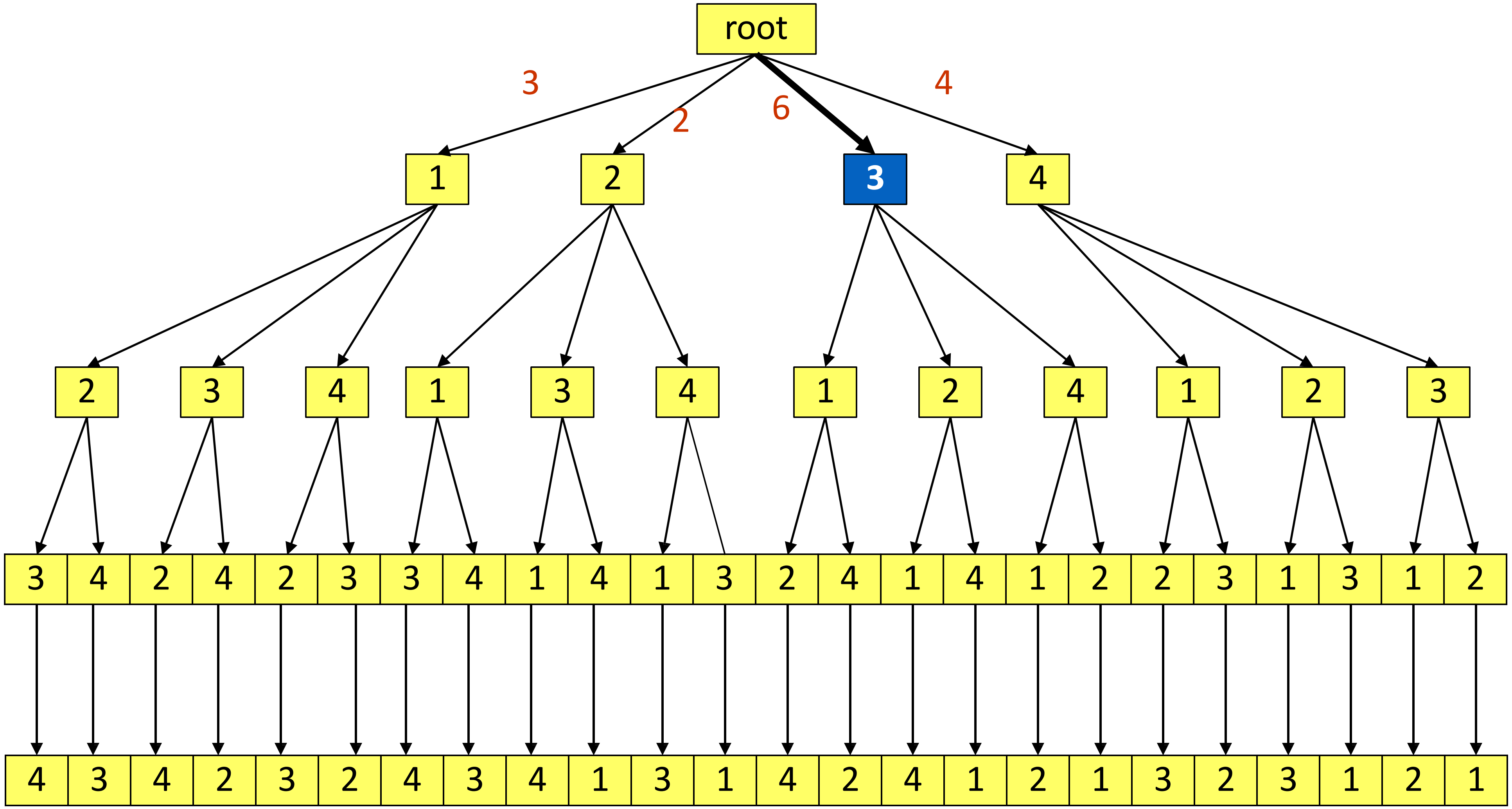
- Example:
  - Select the best feature subset among $\mathbf{X=\{x_1,x_2,x_3,x_4\}}$
  - Objective function
- $\mathbf{h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4}$



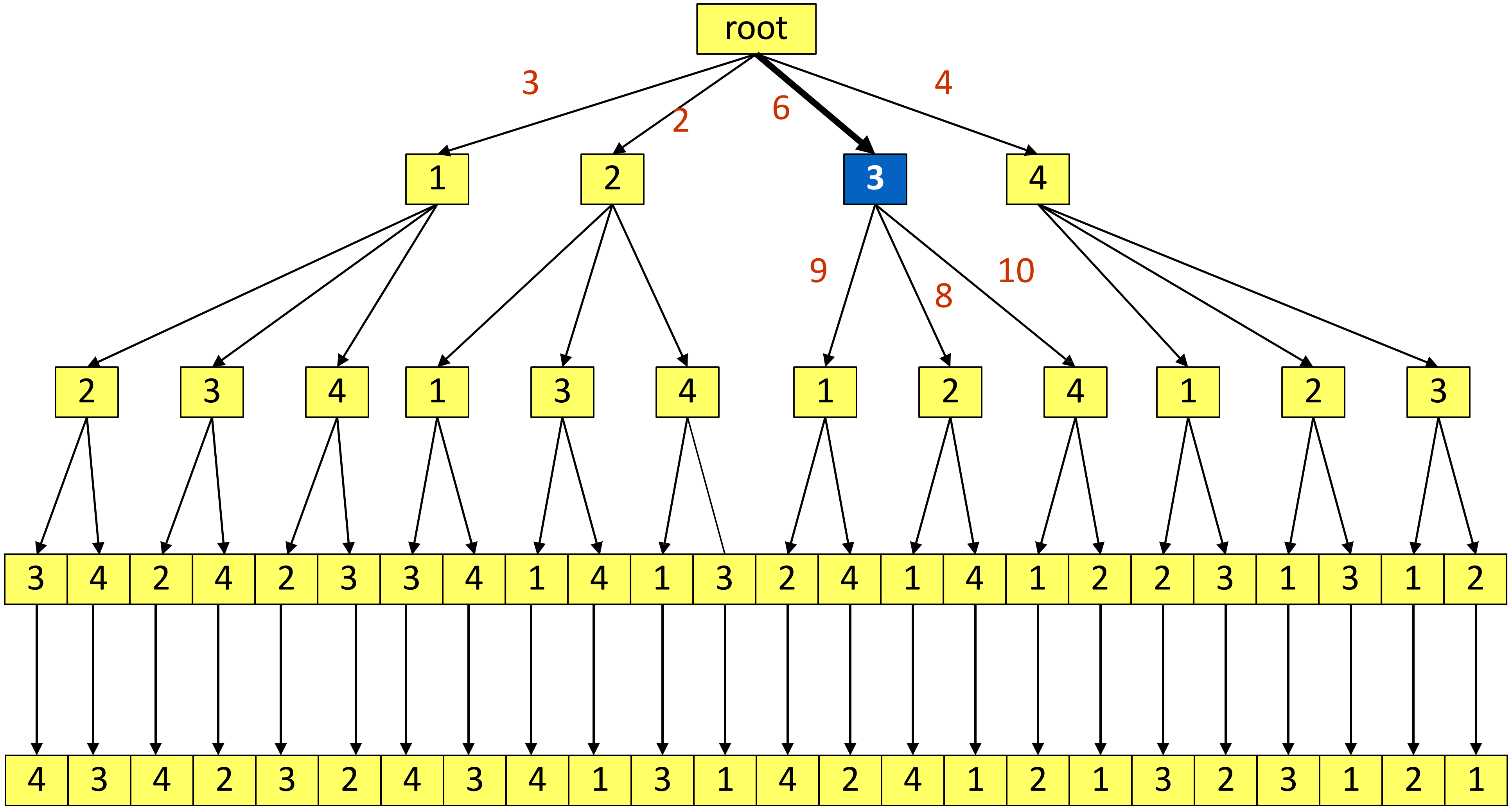
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



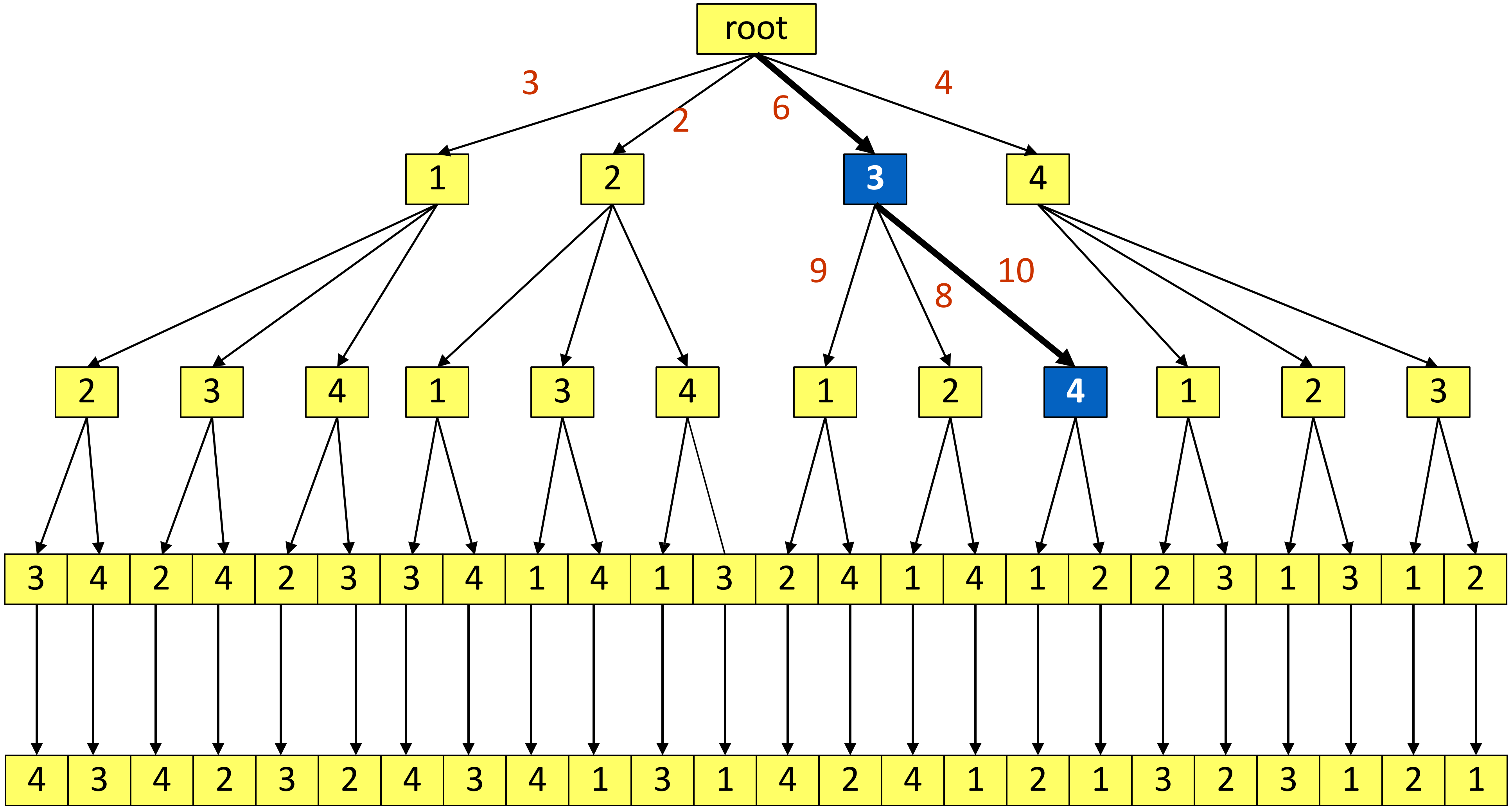
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



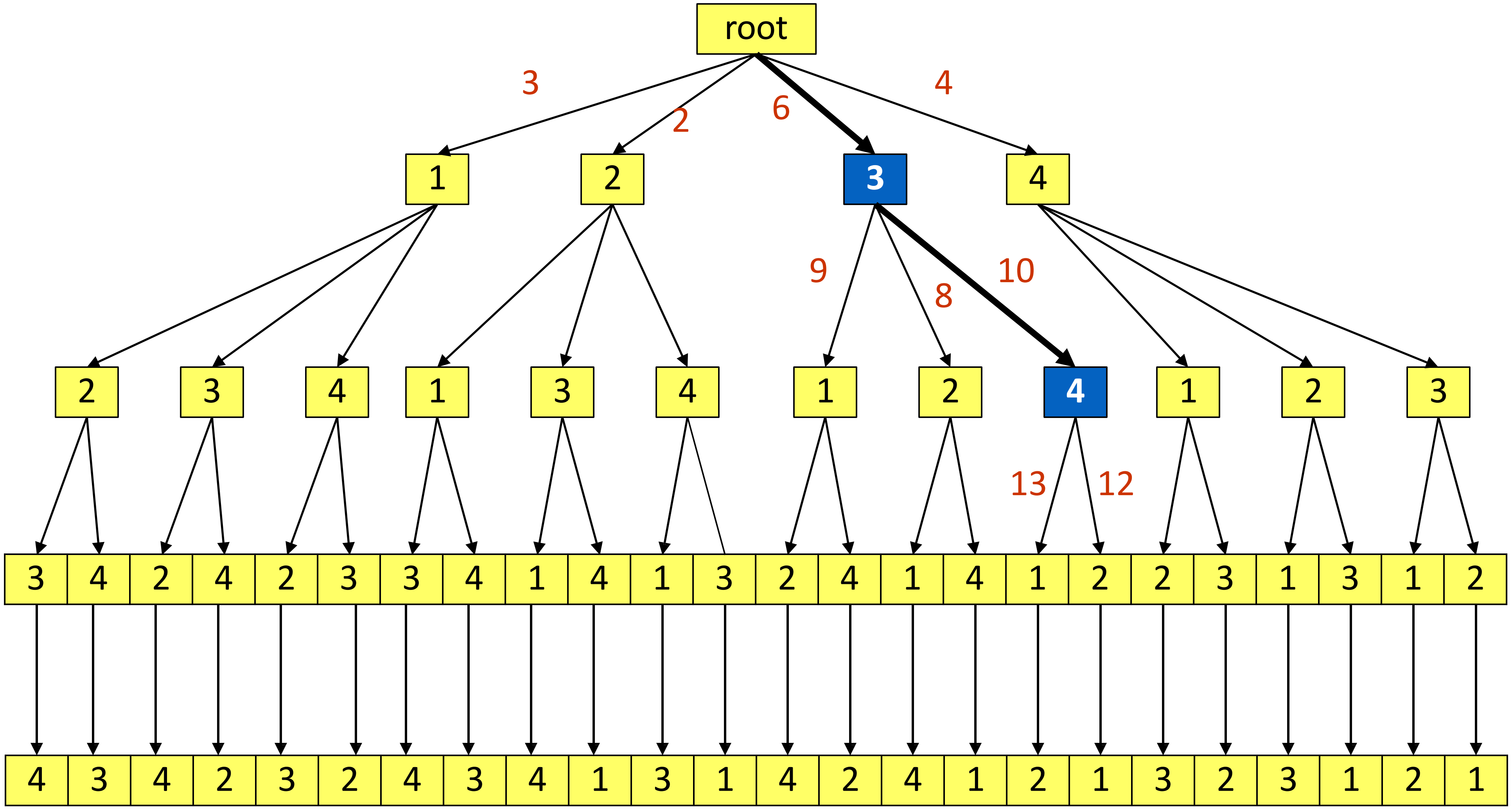
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



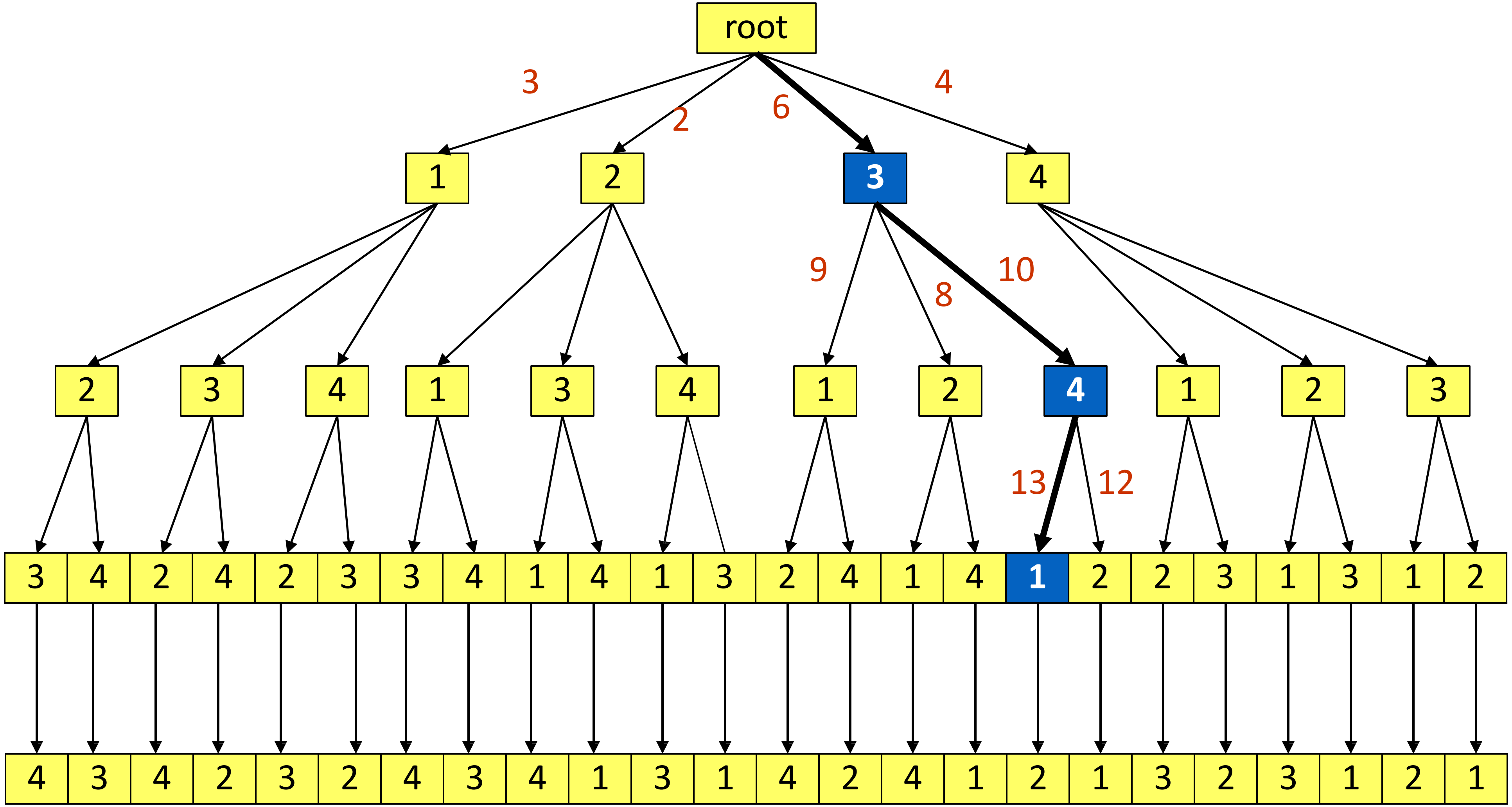
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



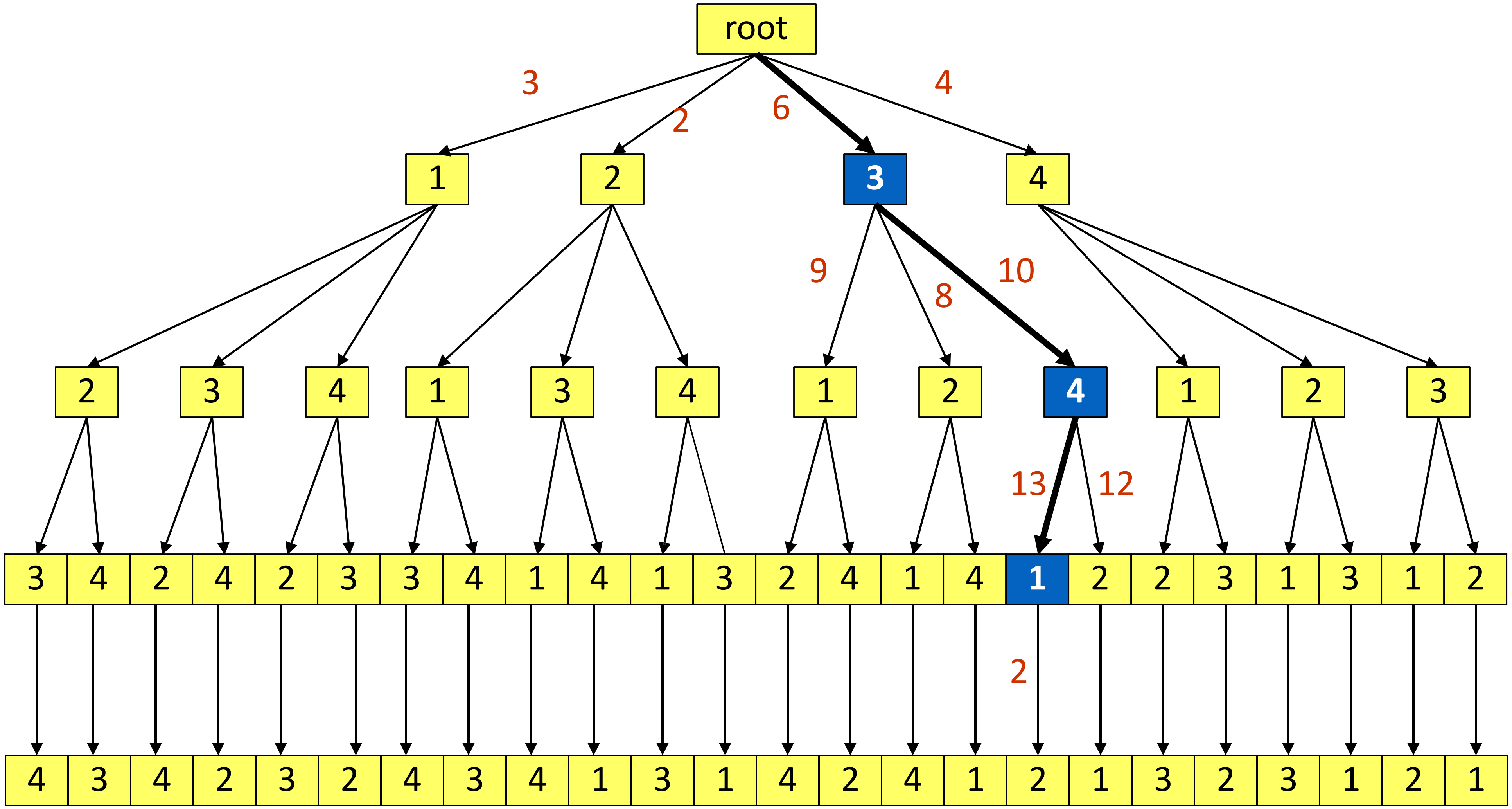
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$

Search Strategies (3)

- SFS performs best when the optimal subset has a small number of features
- When the search is near the empty set, a large number of states can be potentially evaluated
- Towards the full set, the region examined by SFS is narrower since most of the features have already been selected

Search Strategies (4)

- Sequential Backward Selection (SBS)

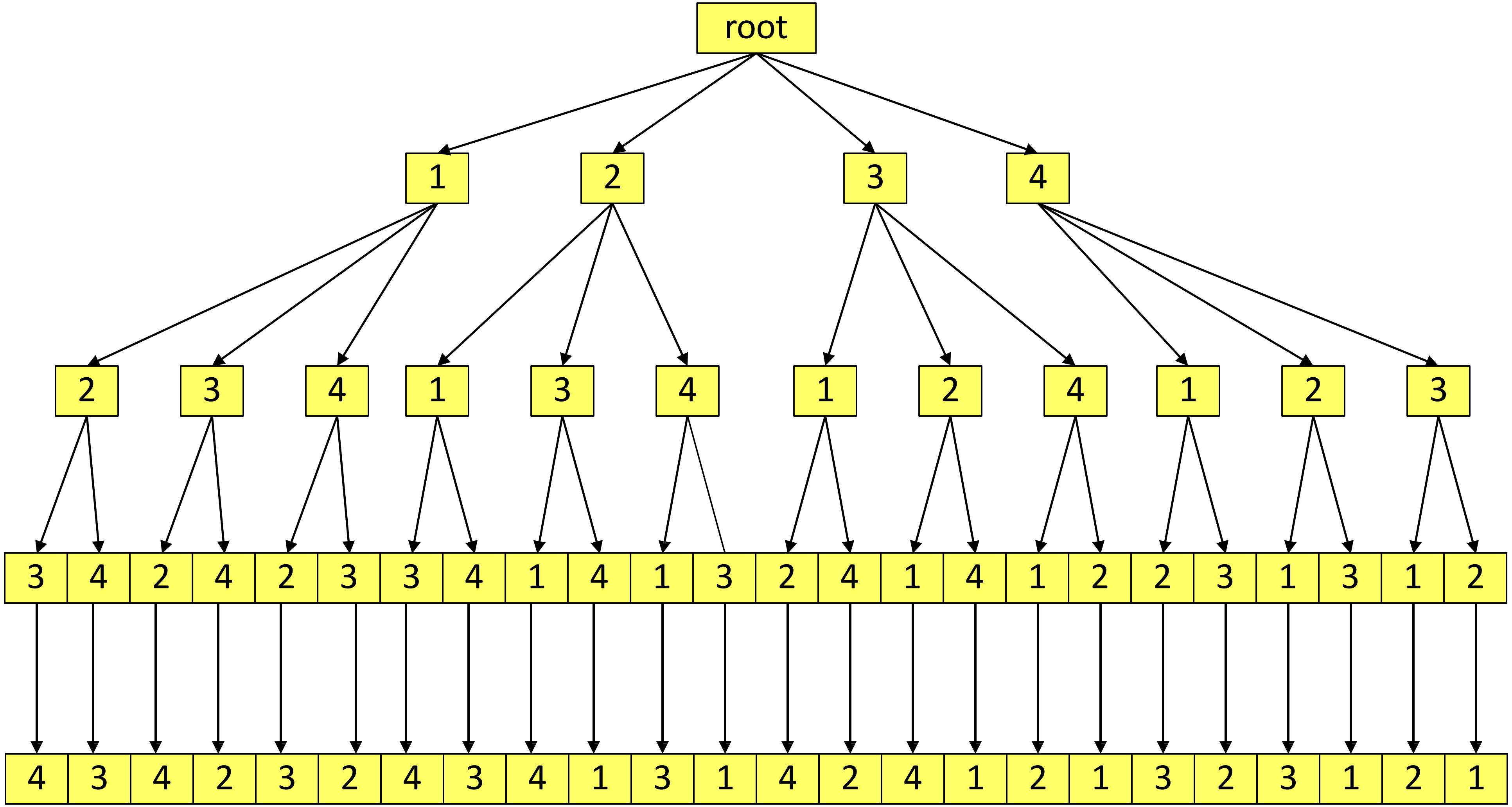
1. Start with full set: Y← X
2. Select the next worst feature that maximizes the objective function of the selected features

$$z \leftarrow \arg\max_{x \notin Y} [h(Y - \{x\})]$$

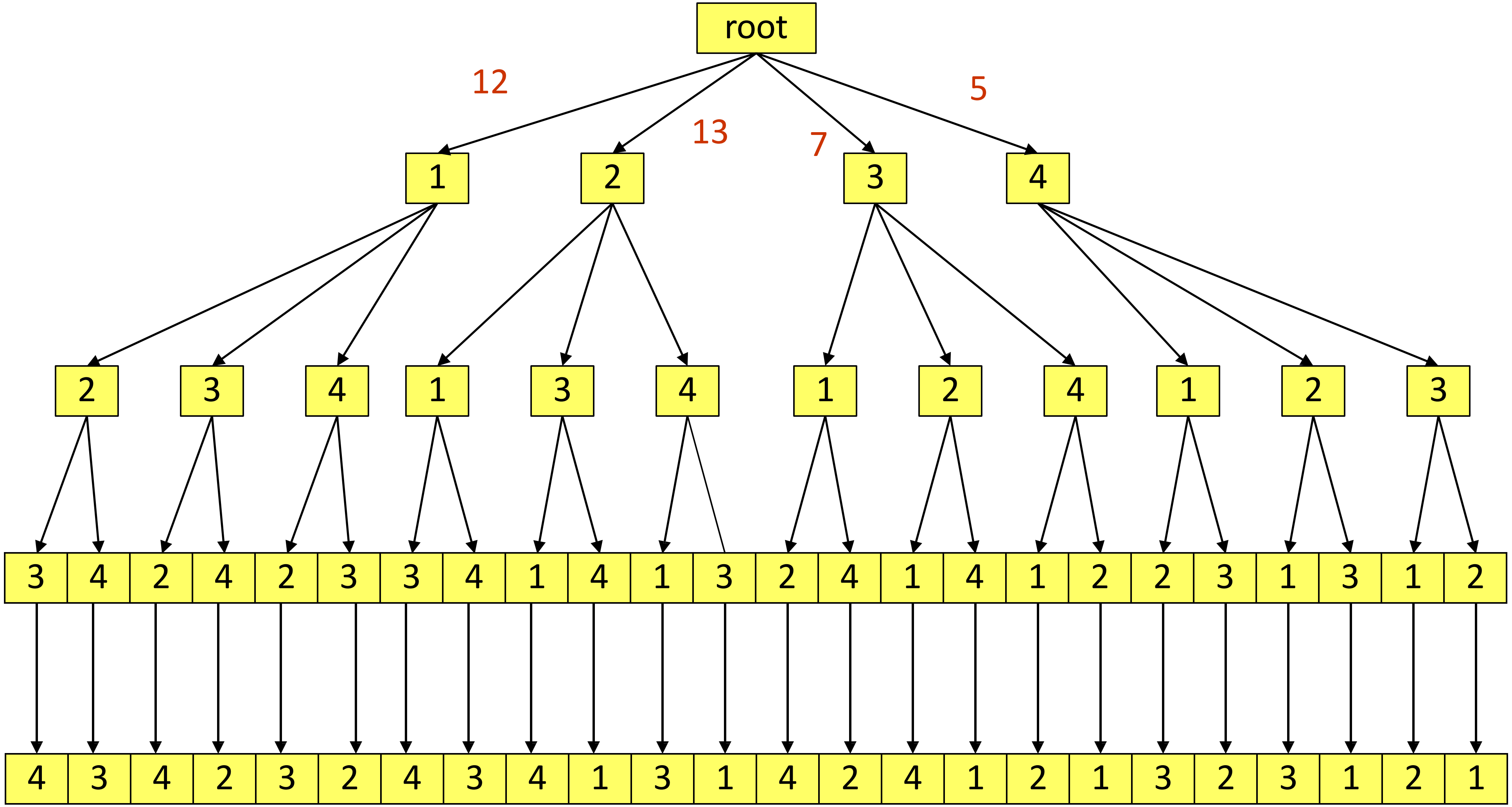
$$Y \leftarrow Y - \{z\}$$

3. Update Y:
4. Go to 2

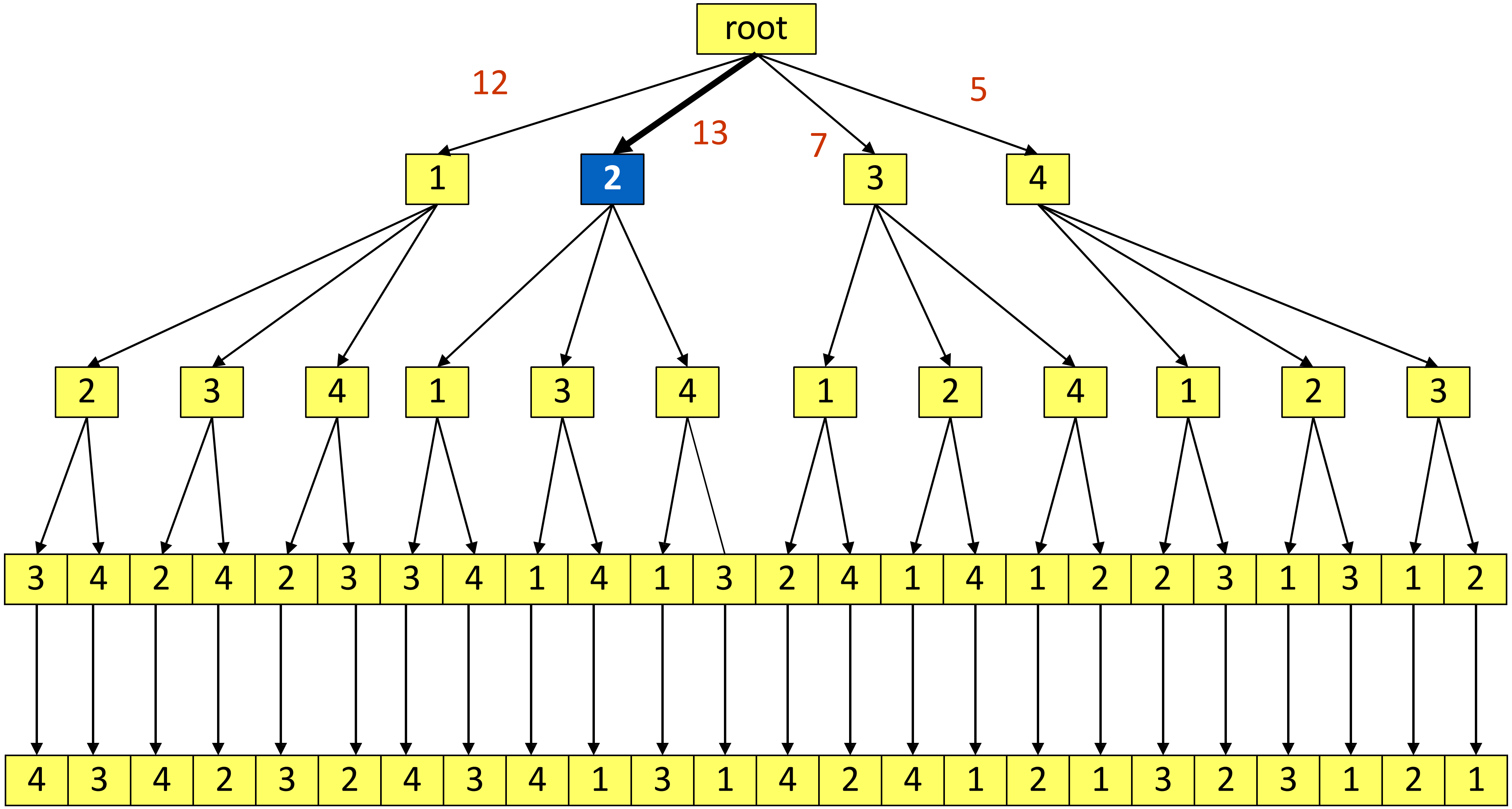
$$\mathbf{h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4}$$



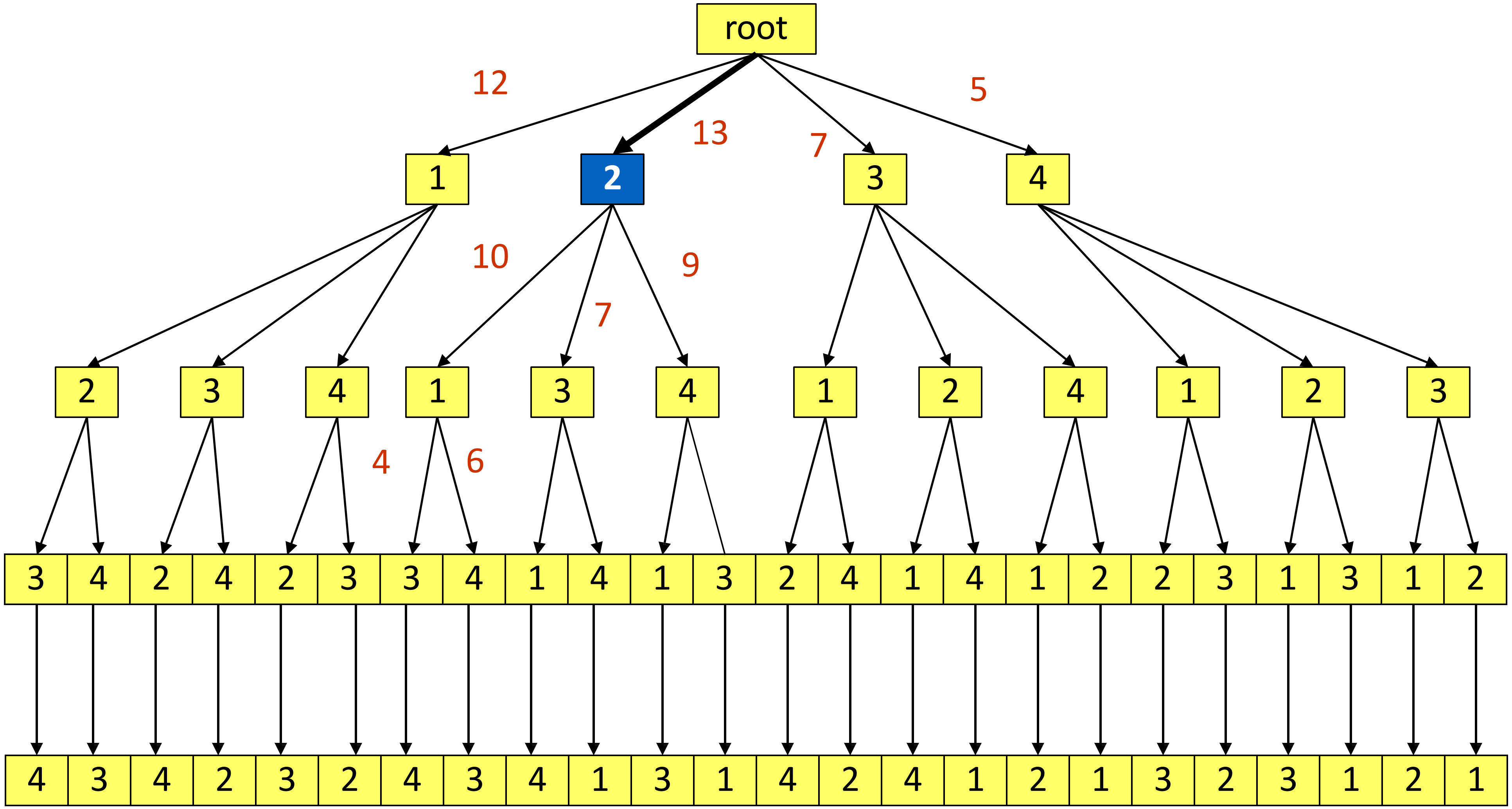
$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$



$$h=3x_1+2x_2+6x_3+4x_4-2x_1x_2-4x_1x_2x_3-7x_1x_2x_3x_4$$

Objective Function (1)

- Objective function to evaluate the feature subset during the search
  - The objective function is a classifier evaluating feature subsets by their predictive capacity (classifier performance) → Wrapper approach
  - The objective function evaluates feature subsets by their information content, relevance, interclass distance, statistical dependence or information-theoretic measures → Filter approach

Objective Functions (2)

- Distance-based measures: a good feature subset is increasing intra-class similarity and decreasing inter-class similarities
- Correlation-based measures: a good feature subset is correlated with the relevant class. All features should be uncorrelated with each other
- Information-theoretic measures: A good feature shares maximum information with the relevant class

Missing value and error in data

Missing Value

- A variable in an observation does not have any value recorded.
- Common in most real-world datasets (examples: incomplete or partial data for an observation, missing sequence, incomplete feature, reporting error etc.)

Importance

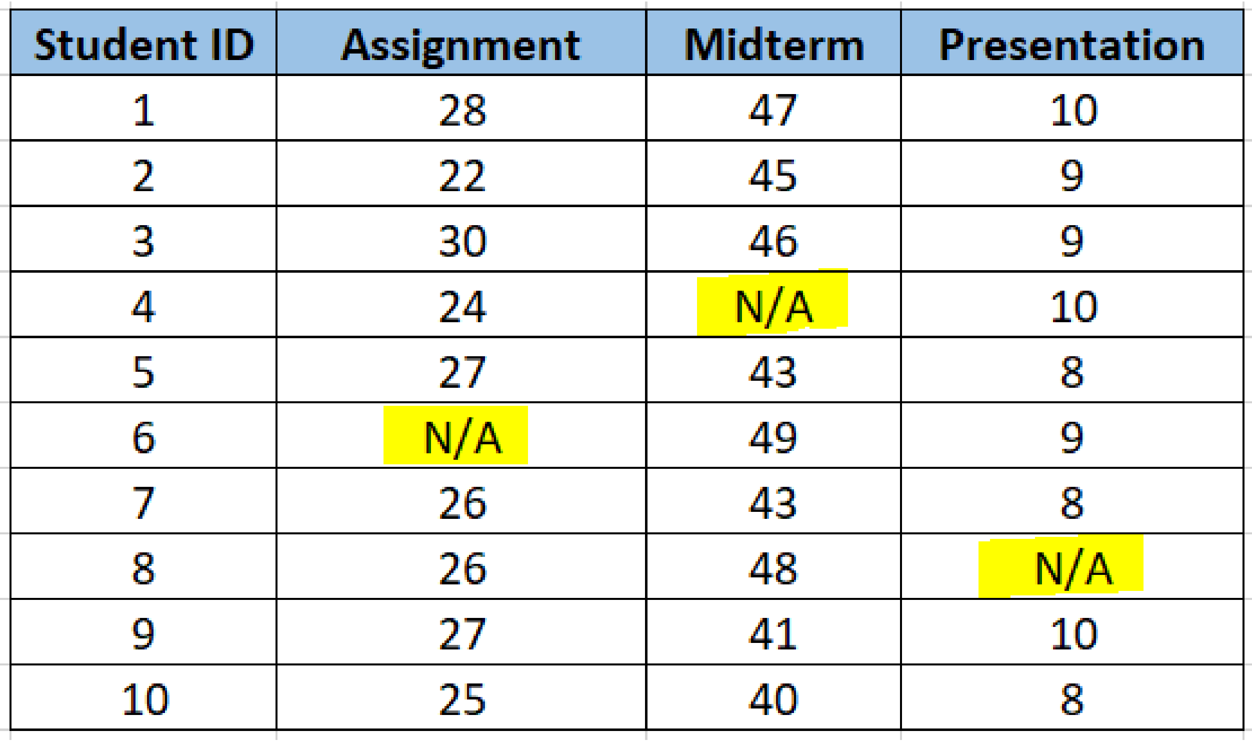
- Can have serious performance effect if not taken care of
- Missing data fields needs to be transformed to fit into ML modeling and further analysis

Missing value and error in data

Missing value example

- Student test scores of a class in certain point in time

- Student-6 missed the assignment
- Student-4 test score not recorded by mistake
- Student-8 presentation score on hold for re-submission



Student ID	Assignment	Midterm	Presentation
1	28	47	10
2	22	45	9
3	30	46	9
4	24	N/A	10
5	27	43	8
6	N/A	49	9
7	26	43	8
8	26	48	N/A
9	27	41	10
10	25	40	8

Missing value and error in data

Missing Value Handling

- Missing data reduces the representativeness of the sample.
- Makes it difficult to process the data for many analysis models / algorithms.
- Three main approaches to deal with missing values-
  1. Imputation
  2. Omission
  3. Analysis

Missing value and error in data

Missing Value Handling

- **Imputation**
  - Values are filled in the place of missing data
  - Works well for situation where analysis tools are not robust to missing values
  - Dataset sizes are not reduced but noise gets imposed with the imputation

Estimation methods: regression, maximum likelihood estimation and approximate Bayesian bootstrap

Missing value and error in data

Missing Value Handling

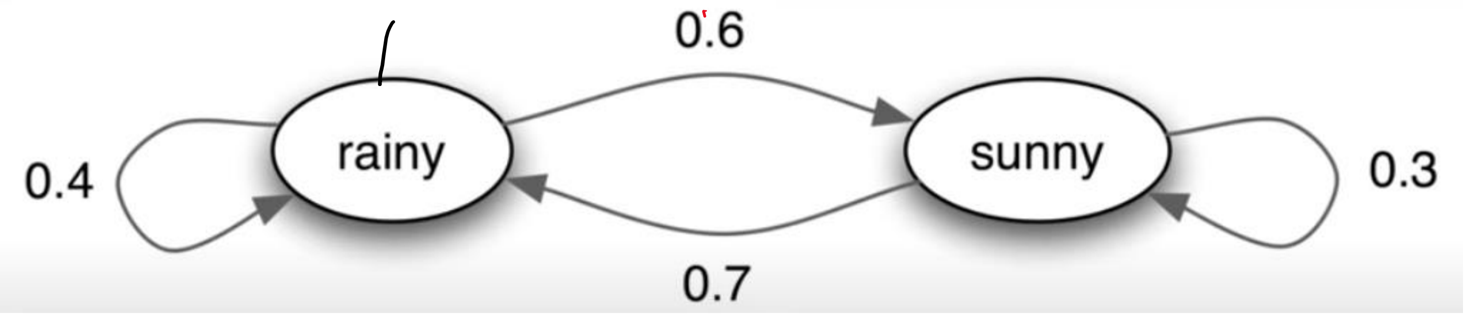
- **Omission**
  - Samples with invalid data are discarded from further analysis
  - Creates a subset of dataset with no missing values
  - Works well for models that are not robust against data missingness

Example techniques: list-wise deletion, pair-wise deletion

Missing value and error in data

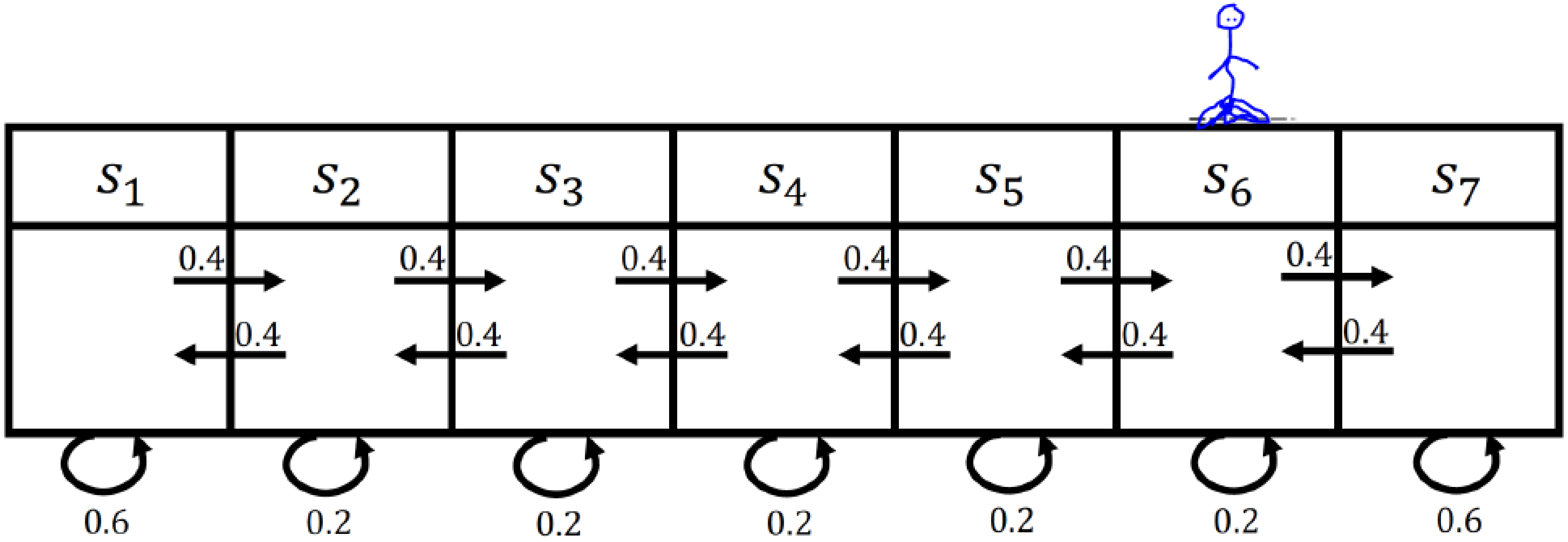
Missing Value Handling

- **Analysis**
  - Samples with invalid data are discarded from further analysis
  - Model-based techniques used to determine missing values
  - Various non-stationary Markov chain models can be used for time series data



Missing value and error in data

Analysis: Markov chain models can be used for time series data



Missing value and error in data

Error in Data

- The difference between the recorded data and true value
- Higher error rates in data makes it less representative

Types

- The major types of data error includes-

- Sampling error
- Non-sampling error

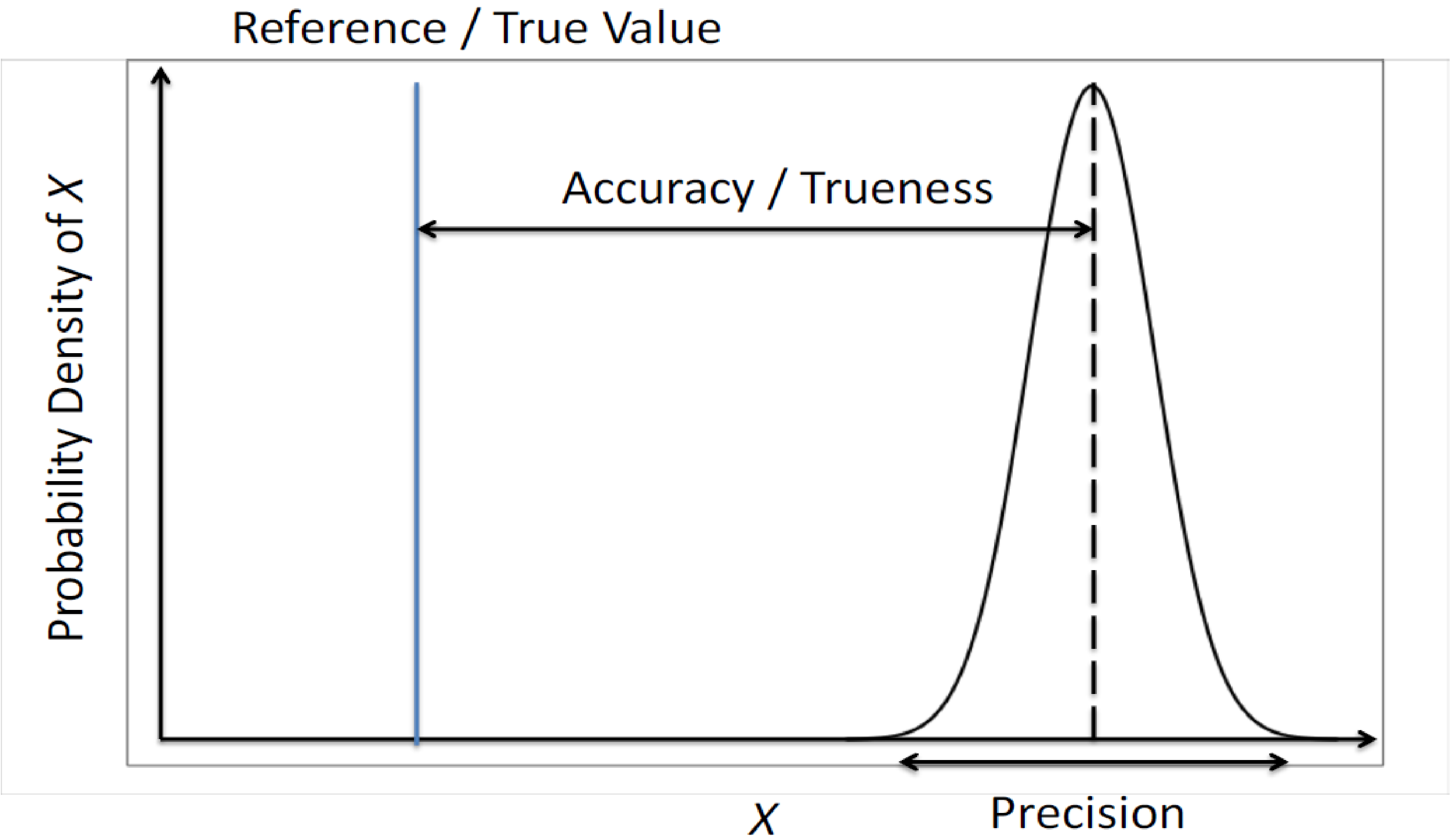
Missing value and error in data

Sampling error

- Error for using data from sample of the population, in place of entire population
- It’s the difference between the estimate from the sample and true vale for the population
- Could occur for very small sample size
- If the sampling is not random and have some bias

Theory of Measurements

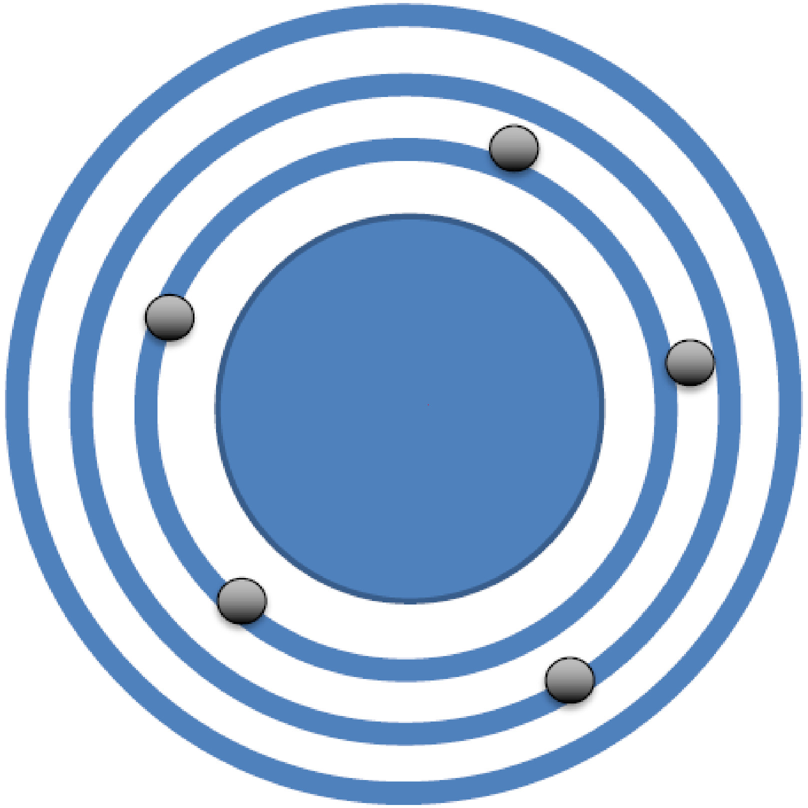
Accuracy & Precision



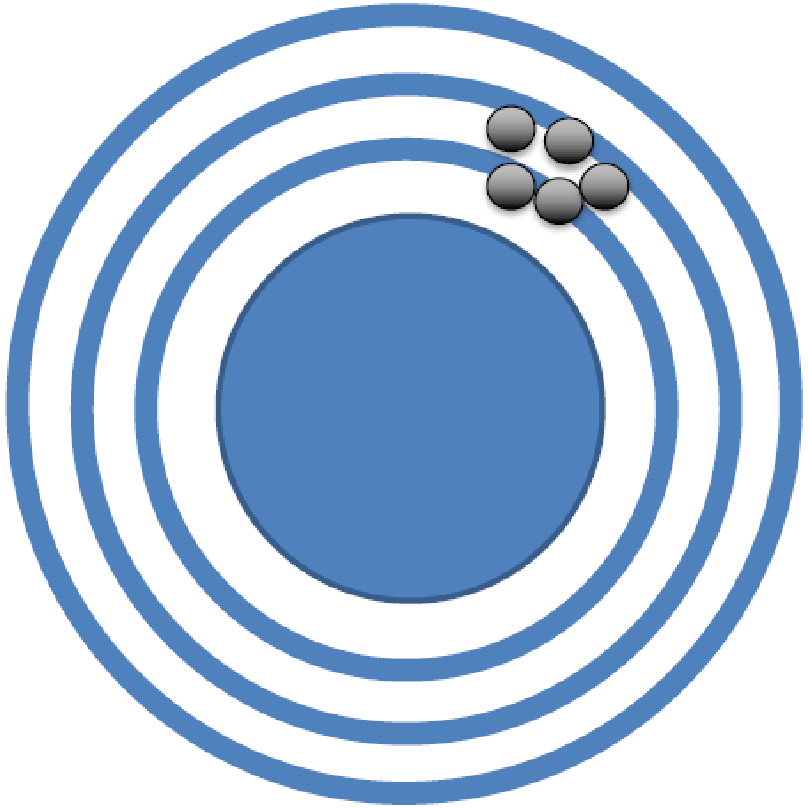
Theory of Measurements

Accuracy & Precision

Low Accuracy, Low Precision



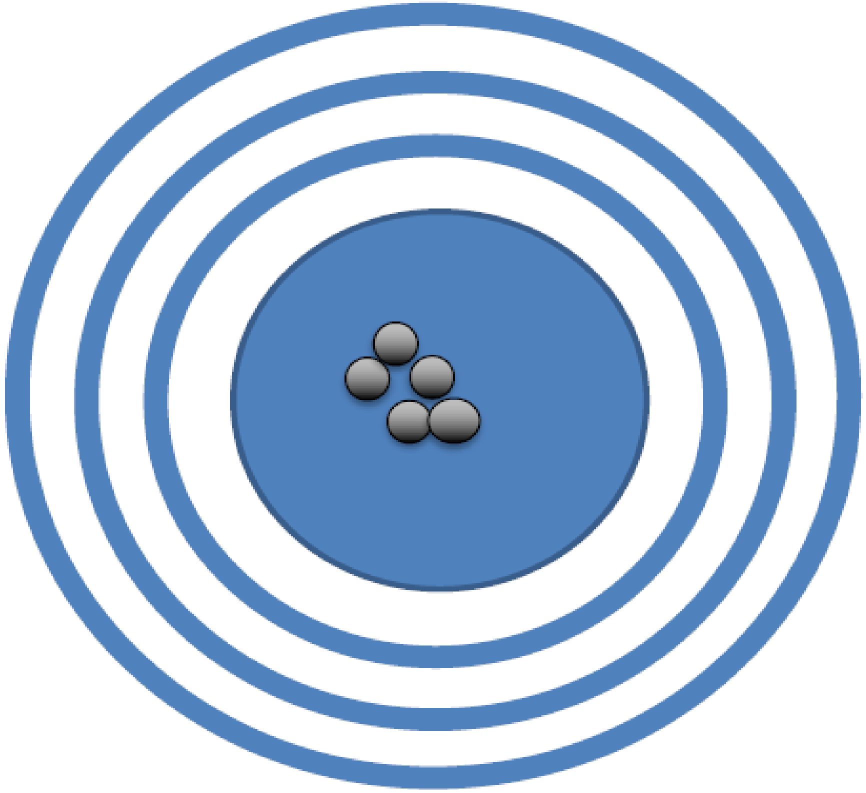
Low Accuracy, High Precision



Theory of Measurements

How would "High accuracy, High precision" look like?

Accuracy & Precision



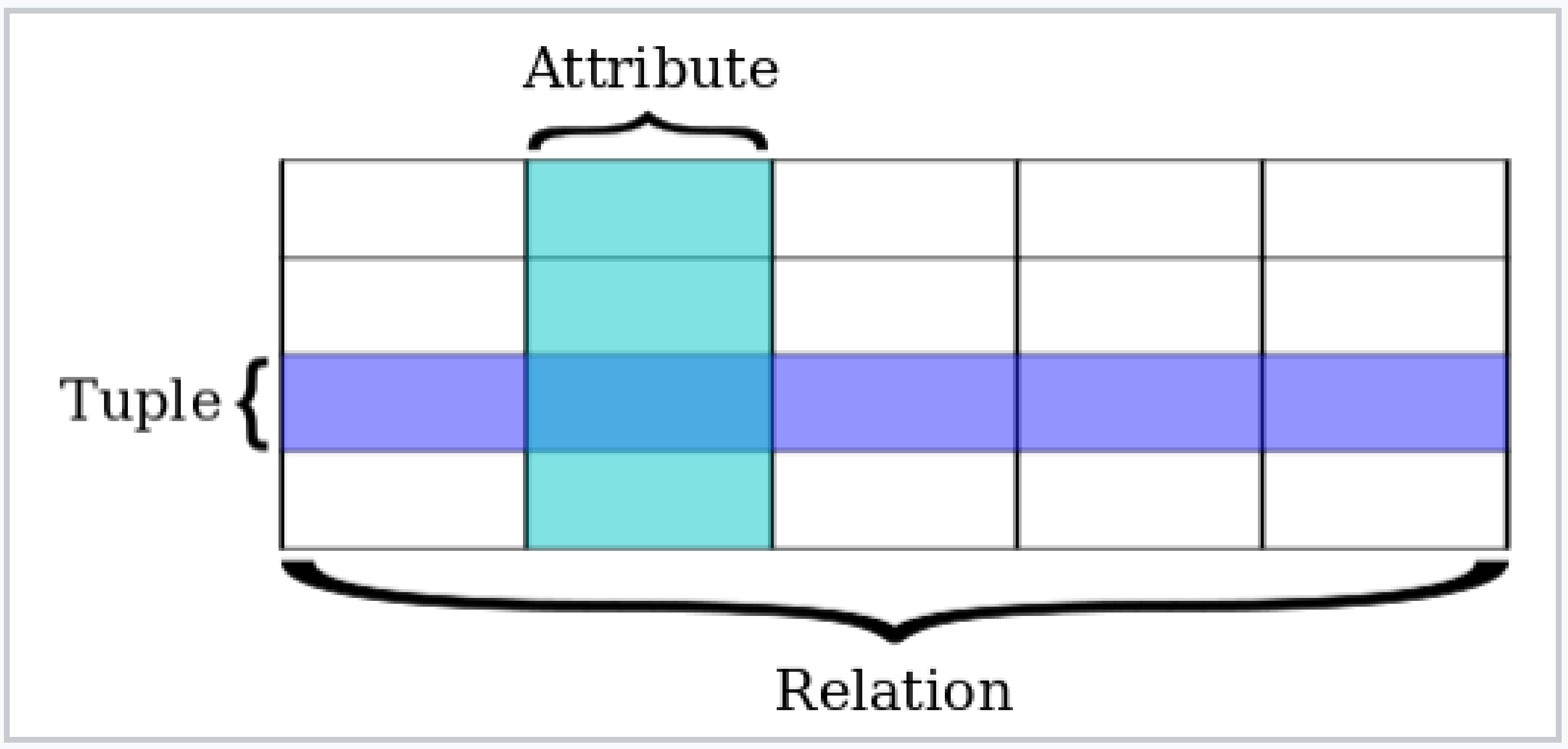
Relational Database

Relational Model

- Relational model uses table (called relation) to represent a collection of related data values
- Rows are called records or tuples
- Columns are called attributes
- The number of attributes (i.e., number of columns) is called the degree
- Example database: MySQL, PostgreSQL and SQLite3

Relational Database

Relational database terminology



Non-relational Database

Non-relational database

- Database that does not use the tabular schema of rows and columns; i.e. it don’t use relational model.
- Often refers to NOSQL (not only SQL); Data may be stored as
  - simple key/value pairs
  - JSON documents or
  - a graph consisting of edges and vertices.
- Most NOSQL systems are distributed databases or distributed storage systems
- Example DB: MongoDB, Oracle NoSQL, Apache CouchDB and Redis.

Non-relational Database

NOSQL Systems

- Database NOSQL systems focus on storage of “big data”
- Typical applications that use NOSQL
  - Social media
  - Web links
  - Marketing and sales
  - Posts and tweets
  - Road maps and spatial data
  - Email

Non-relational Database

NOSQL Systems

- BigTable

  - Google’s proprietary NOSQL system

  - Column-based or wide column store

- DynamoDB (Amazon)

  - Key-value data store

- Cassandra (Facebook)

  - Uses concepts from both key-value store and column-based systems

Non-relational Database

NOSQL Systems

- MongoDB and CouchDB
    - Document stores
- Neo4J and GraphBase
    - Graph-based NOSQL system
    - OrientDB
    - Combines several concepts

Non-relational Database

NOSQL characteristics

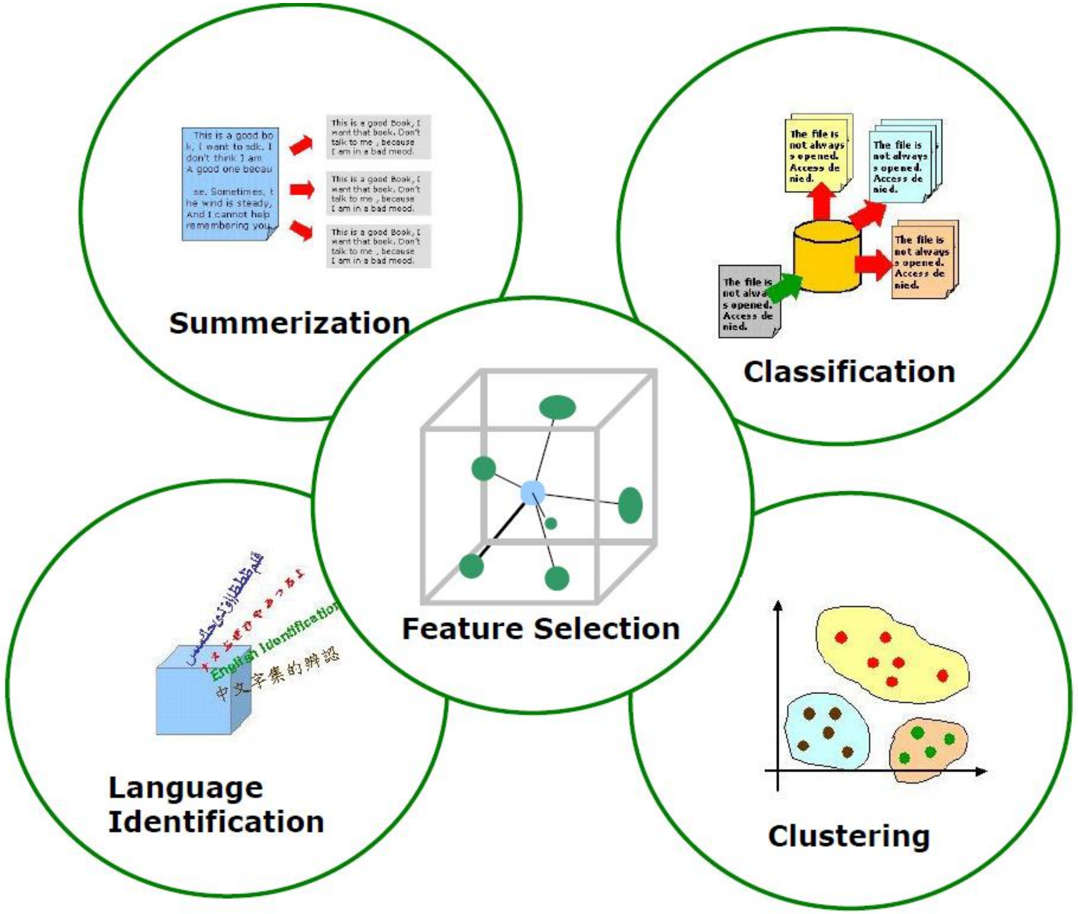
- With respect to Data models and query languages
  - Schema not required
  - Less powerful query languages
  - Versioning

Non-relational Database

NOSQL characteristics

- With respect to distributed databases and distributed systems
  - Scalability
  - Availability, replication, and eventual consistency
  - Replication models (Master-slave & Master-master)
  - High performance data access

Next session: Text Analysis


Summerization
Classification
Feature Selection
Language Identification
Clustering
English Identification