BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

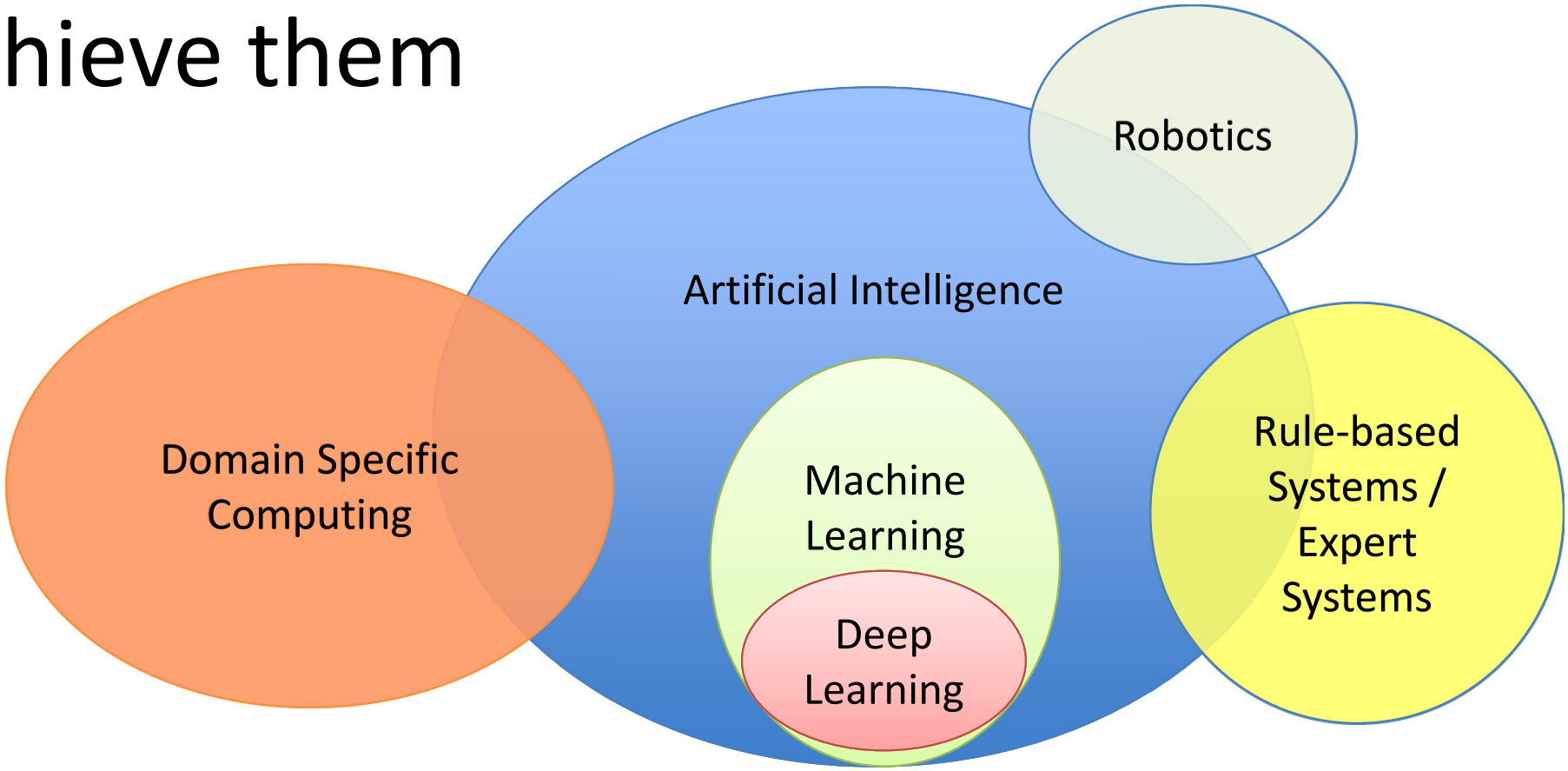
1. What is machine learning?
2. Types of machine learning
3. How to choose the best algorithm in each case?
4. Supervised learning
    1. K-Nearest Neighbours
    2. Decision trees
    3. Linear Regression
    4. Decision trees for regression

Sub-objectives of AI

- Perception
- Reasoning
- Navigate
- Natural language processing
- Represent Knowledge

4 MACHINE LEARNING ALGORITHMS

How to achieve them



4.1. What is machine learning

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

Machine learning (ML) is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed.[2]

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.[8]

4.1. What is machine learning

L. Sebastiá

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

Day-to-day ML applications

from https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

1. Virtual Personal Assistants
2. Predictions while Commuting
3. Video Surveillance
4. Social Media Services
5. Email Spam and Malware Filtering
6. Online Customer Support
7. Search Engine Result Refining
8. Product Recommendations
9. Online Fraud Detection

4.1. What is machine learning

L. Sebastiá

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

Tom Mitchell's definition of Machine Learning:

> *"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience"*

4 MACHINE LEARNING ALGORITHMS

7 steps of the machine learning process

Gathering data

- The quality and quantity of data gathered will directly determine how good the predictive model can be

Data preparation

- Load the data into a suitable place and prepare it for use in the machine learning training (normalization, error-correction, etc.)
- Do visualizations of the data, to see if there are any relevant relationships between different variables that can be helpful, as well as see if there are any data imbalances
- Split the data in two parts: training (the majority of the dataset) and test (will be used for evaluating the trained model's performance)

Choosing a model

- Select the most apropriate algorithm

7 steps of the machine learning process (continued)

Training
- Data is used to incrementally improve the model's ability to predict

Evaluation
- Test the model against data that has never been used for training

Parameter tuning
- There are a few parameters implicitly assumed when training, so now test those assumptions and try other values

Prediction
- Give some input values and obtain an answer

4 MACHINE LEARNING ALGORITHMS

Tom Mitchell's formalism of a Machine Learning problem:

> A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Example: *I need a program that will tell me which tweets will get retweets.*

- **Task T**: Classify a tweet that has not been published as going to get retweets or not
- **Experience E**: A corpus of tweets from an account where some have retweets and some do not
- **Performance P**: Classification accuracy, the number of tweets predicted correctly out of all the tweets considered as a percentage

From https://machinelearningmastery.com/how-to-define-your-machine-learning-problem/

Exercise

Formalize the following problem:

I need a program that tells me the price of a house

- **Task T**: Predict the price of a house that I don't know how much it costs
- **Experience E**: A corpus of houses with their price
- **Performance P**: Prediction accuracy, the error between the predicted price and the real price of a subset of houses used as testing set

Which characteristics about the house would be interesting to predict the price?

Features

A *feature* is an individual measurable property or characteristic of a phenomenon being observed

Usually: numerical or categorical data

Example: Features for House Price Prediction

https://www.kaggle.com/erick5/predicting-house-prices-with-machine-learning/data

Data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. It contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values.

Numerical data:

GarageArea: Size of garage in square feet

Categorical data:

OverallQual: Overall material and finish quality

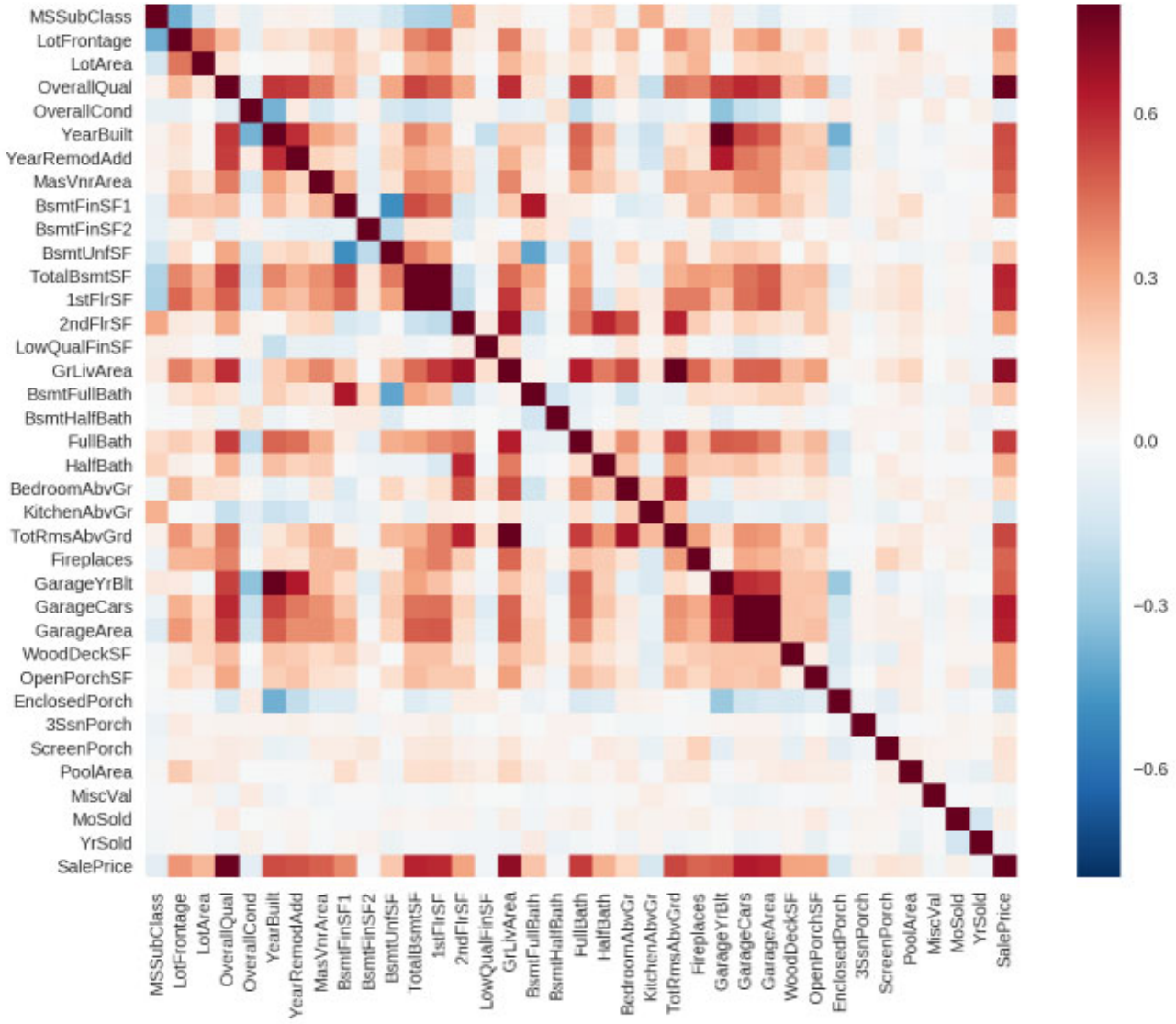
(1 = Very Poor, 10 = Very Excellent)

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

Example: Features for House Price Prediction

Variable to predict: SalePrice



4.1. What is machine learning

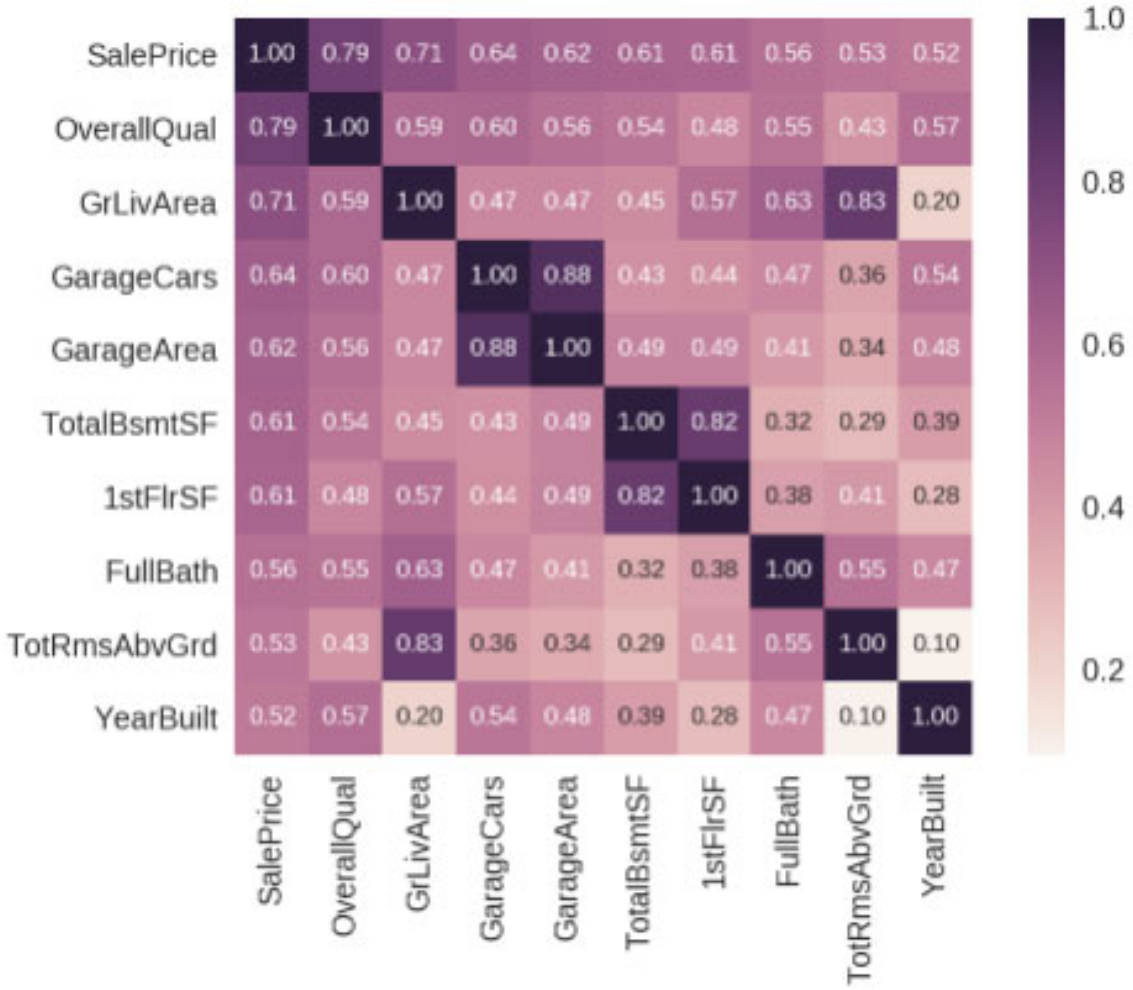
L. Sebastiá

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS

Example: Features for House Price Prediction

Variable to predict: SalePrice



4.1. What is machine learning

L. Sebastiá

Example: Features for House Price Prediction

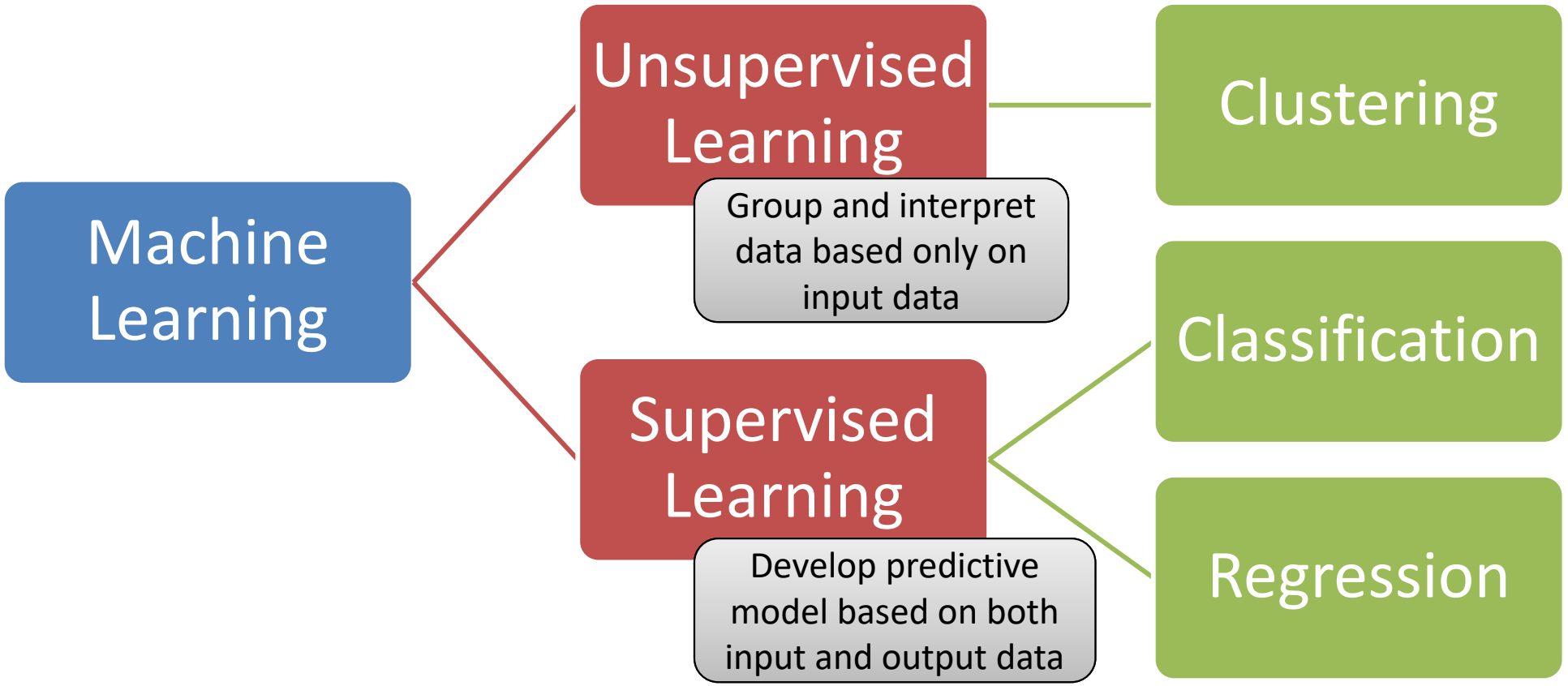
Variable to predict: SalePrice

Variables correlated with SalePrice:

- OverallQual: Rates the overall material and finish of the house
- GrLivArea: Above grade (ground) living area square feet
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- TotalBsmtSF: Total square feet of basement area
- 1stFlrSF: First Floor square feet
- FullBath: Full bathrooms above grade
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- YearBuilt: Original construction date

- Supervised machine learning:
  - The program is trained on a pre-defined set of *training examples*, which then facilitate its ability to reach an accurate conclusion when given new data
- Unsupervised machine learning:
  - The program is given a bunch of data and must find patterns and relationships therein

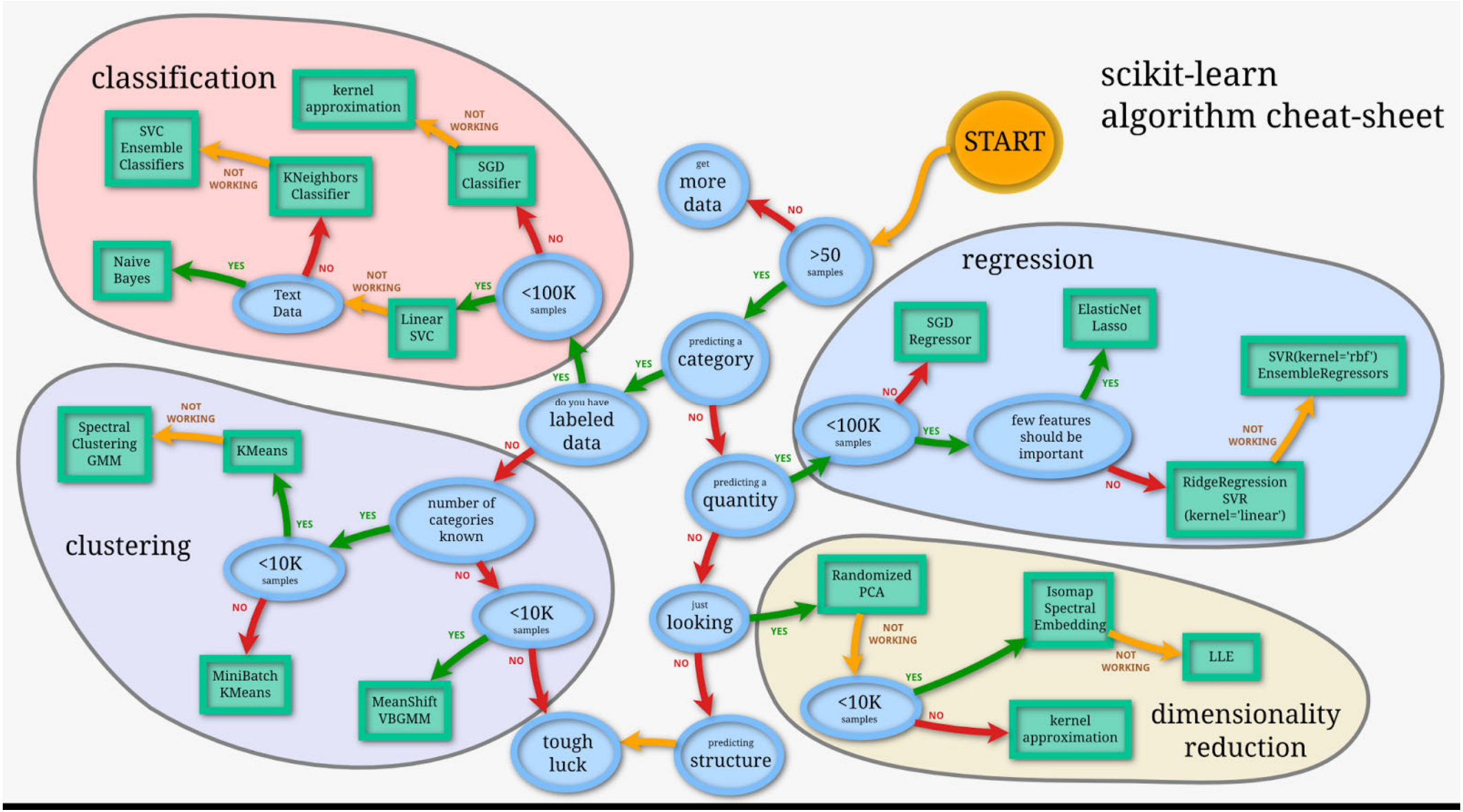
4 MACHINE LEARNING ALGORITHMS



From: https://es.mathworks.com/discovery/machine-learning.html

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEARNING ALGORITHMS



4.3. How to choose the best algorithm in each case? L. Sebastiá