Mondragon Unibertsitatea

Escuela Politécnica Superior

Introduction

Machine Learning

Mondragon Unibertsitatea
Escuela Politécnica Superior

Machine Learning (ML) is the field of study that gives computers the ability to learn without being explicitly programmed.

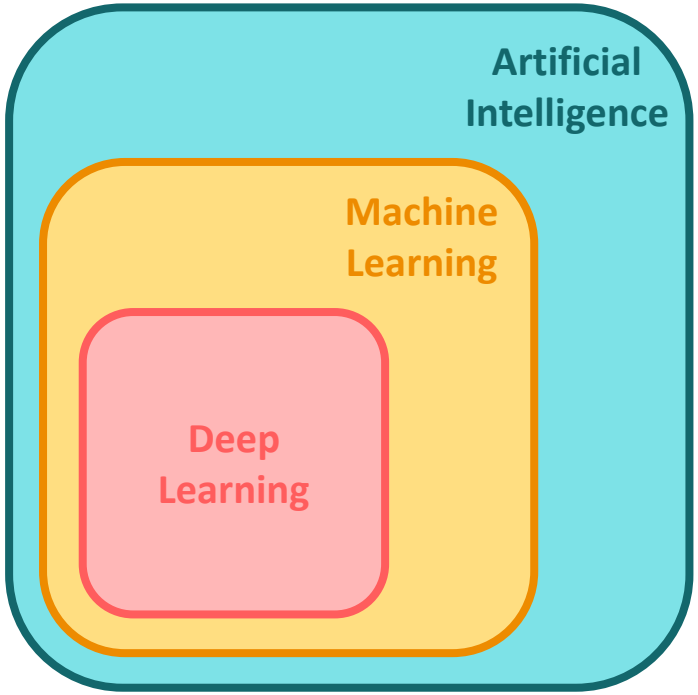
Arthur Samuel (1959)

Machine Learning

Mondragon Unibertsitatea
Escuela Politécnica Superior

Machine Learning (ML) is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel (1959)



Machine Learning – Problem types

Mondragon Unibertsitatea
Escuela Politécnica Superior

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning.[3][4] In its application across business problems, machine learning is also referred to as predictive analytics.

Problem taxonomy:

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Supervised problems

Mondragon Unibertsitatea
Escuela Politécnica Superior

- Classification

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	1
5	1	...	5	6	0
4	6	...	5	5	0
7	89	...	23	85	1
3	435	...	3	1	1
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	0
9	8	...	6	8	1
4	77	...	3	132	0
8	9	...	1	8	0
9	8	...	4	8	?

The category (class) to which a sample belongs must be predicted

We distinguish between **binary classification** (2 classes) y **multiple classification** (3 or more)

- Regression

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	3.1
5	1	...	5	6	1.4
4	6	...	5	5	6.1
7	89	...	23	85	8.0
3	435	...	3	1	9.3
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	4.5
9	8	...	6	8	3.6
4	77	...	3	132	4.5
8	9	...	1	8	2.7
9	8	...	4	8	?

A numerical value (generally a real number) must be predicted

Unsupervised problems

Mondragon Unibertsitatea
Escuela Politécnica Superior

- Clustering

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	-
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	-
3	435	...	3	1	-
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	-
8	9	...	1	8	-
9	8	...	4	8	-

A certain number (predefined or not) of groups (*clusters*) to which the samples belong must be defined, according to certain pre-selected **similarity measure**

- Dimensionality reduction

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	-
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	-
3	435	...	3	1	-
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	-
8	9	...	1	8	-
9	8	...	4	8	-

The number of dimensions (data columns) must be diminished by selecting a subset or defining new columns (less than n) using the n original ones

Unsupervised problems

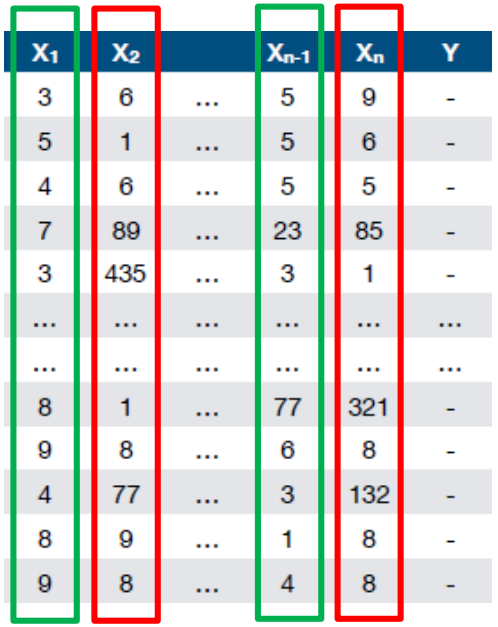
Mondragon Unibertsitatea
Escuela Politécnica Superior

- Clustering

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	-
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	-
3	435	...	3	1	-
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	-
8	9	...	1	8	-
9	8	...	4	8	-

A certain number (predefined or not) of groups (*clusters*) to which the samples belong must be defined, according to certain pre-selected **similarity measure**

- Dimensionality reduction



X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	-
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	-
3	435	...	3	1	-
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	-
8	9	...	1	8	-
9	8	...	4	8	-

The number of dimensions (data columns) must be diminished by selecting a subset or defining new columns (less than n) using the n original ones

Semi-supervised problems

Mondragon Unibertsitatea
Escuela Politécnica Superior

- Partial information

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	3.1
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	8.0
3	435	...	3	1	9.3
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	4.5
8	9	...	1	8	-
9	8	...	4	8	?

Hard to explore all the informative power of the data

- Two logical options

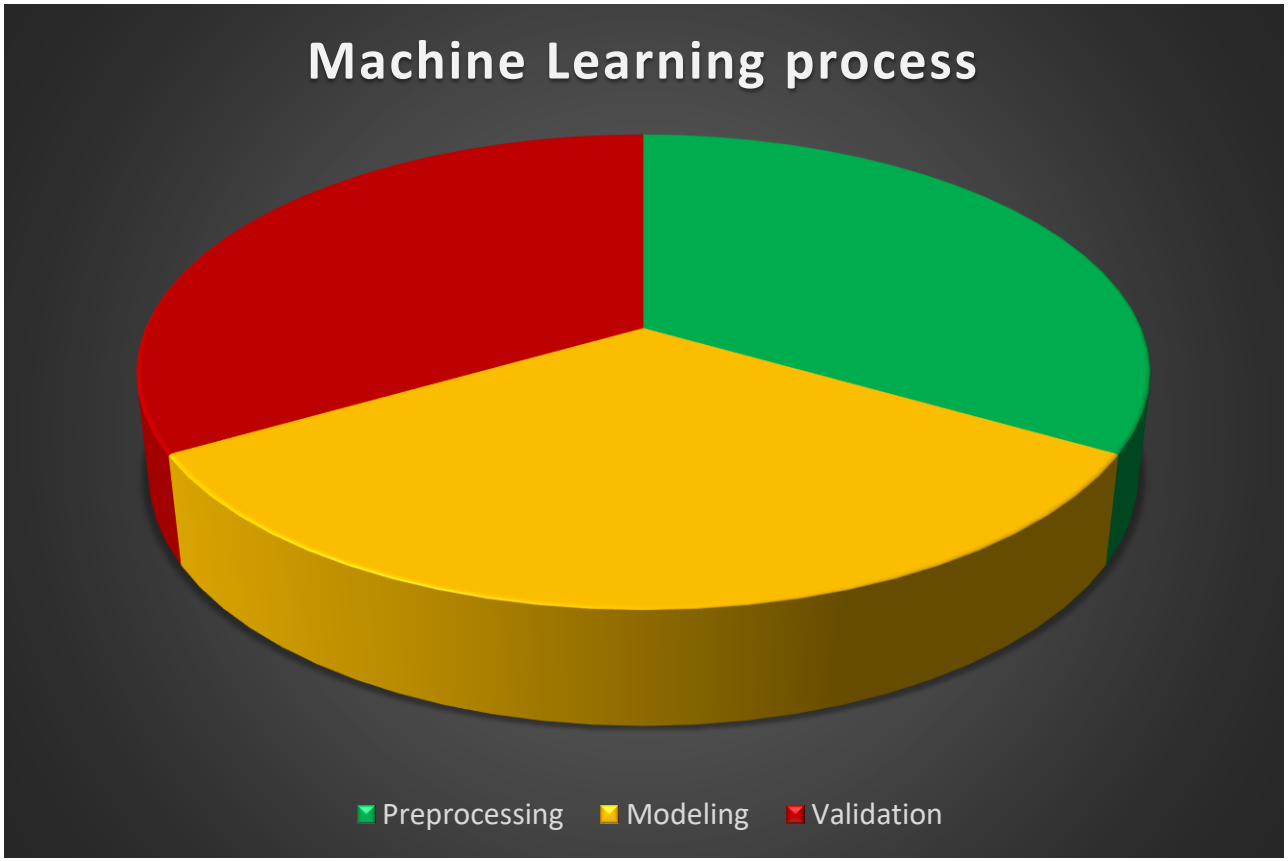
X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	-
5	1	...	5	6	-
4	6	...	5	5	-
7	89	...	23	85	-
3	435	...	3	1	-
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	-
9	8	...	6	8	-
4	77	...	3	132	-
8	9	...	1	8	-
9	8	...	4	8	?

1. Treat it as a clustering problem and predect the class or numerical value using the clusters

X_1	X_2		X_{n-1}	X_n	Y
3	6	...	5	9	3.1
5	1	...	5	6	1.5
4	6	...	5	5	5.9
7	89	...	23	85	8.0
3	435	...	3	1	9.3
...	...	...	...	...	...
...	...	...	...	...	...
8	1	...	77	321	4.5
9	8	...	6	8	3.7
4	77	...	3	132	4.5
8	9	...	1	8	2.5
9	8	...	4	8	?

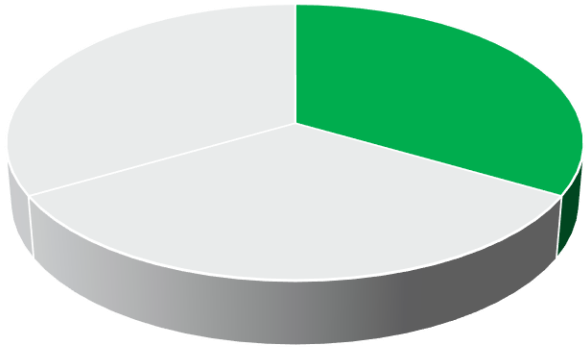
2. Treat it as a supervised problem, filling somehow the unknown values

Machine Learning



Machine Learning

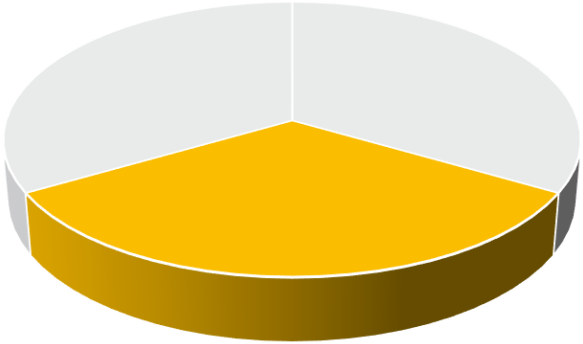
Mondragon Unibertsitatea
Escuela Politécnica Superior



PREPROCESSING

- Data checking:
  * Outlier detection
  * Missing values
- Data transformation:
  * Mean centering
  * Normalization
  * Standardization
- Data compression:
  * Variable discretization
  * Variable selection
  * Variable extraction
- Imbalanced data:
  * Cost-sensitive methods
  * Sampling methods

Machine Learning



MODELING

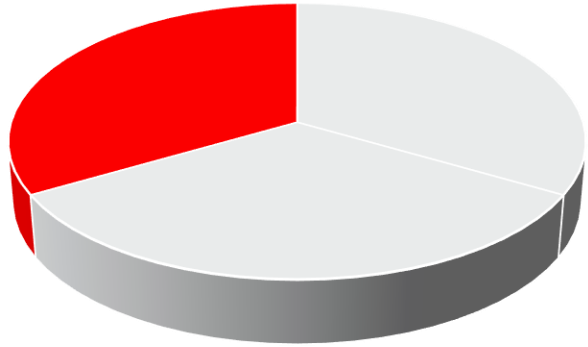
- Algorithm(s) selection:
  * Type(s) of approaches (problem dependent)
- Parameter tuning:
  * Grid search
  * Cross validation
  * Evolutionary algorithms

Machine Learning

Mondragon Unibertsitatea
Escuela Politécnica Superior

VALIDATION

- Testing on independent test sets
  * Train/Test (stratified) split
- Uncertainty estimation:
  * Confidence
  * Sensitivity
- Statistical significance:
  * Mann-Whitney-Wilcoxon test

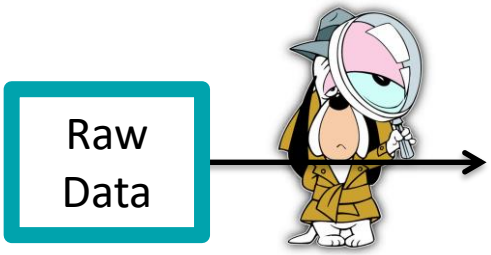


Machine Learning workflow

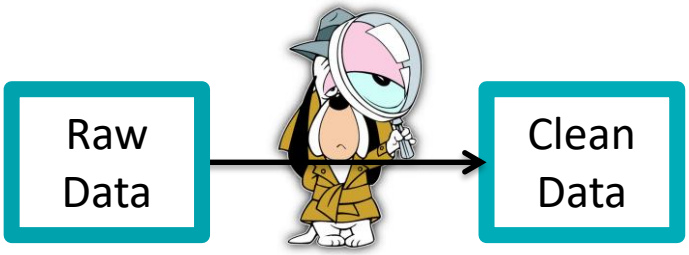
Mondragon Unibertsitatea
Escuela Politécnica Superior

Raw Data

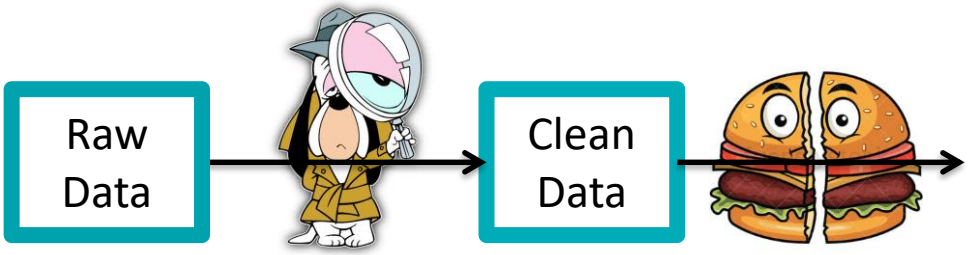
Machine Learning workflow



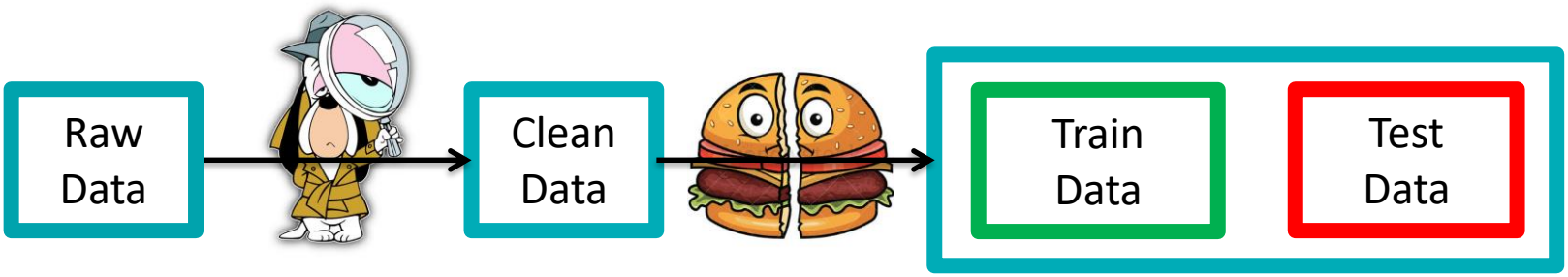
Machine Learning workflow



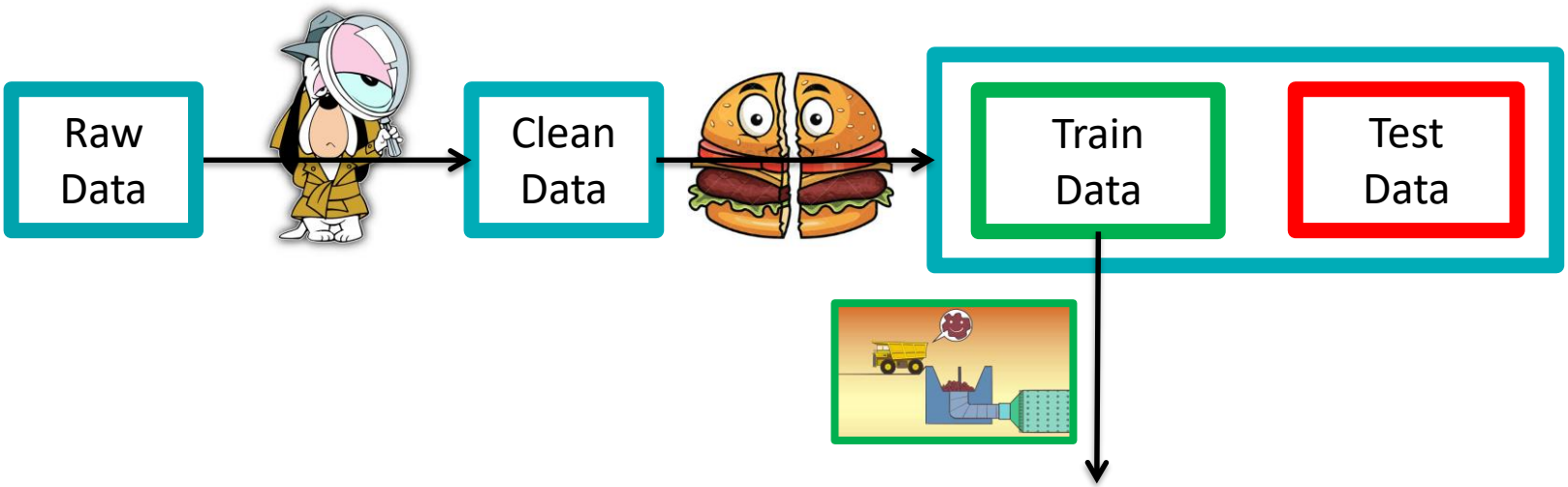
Machine Learning workflow



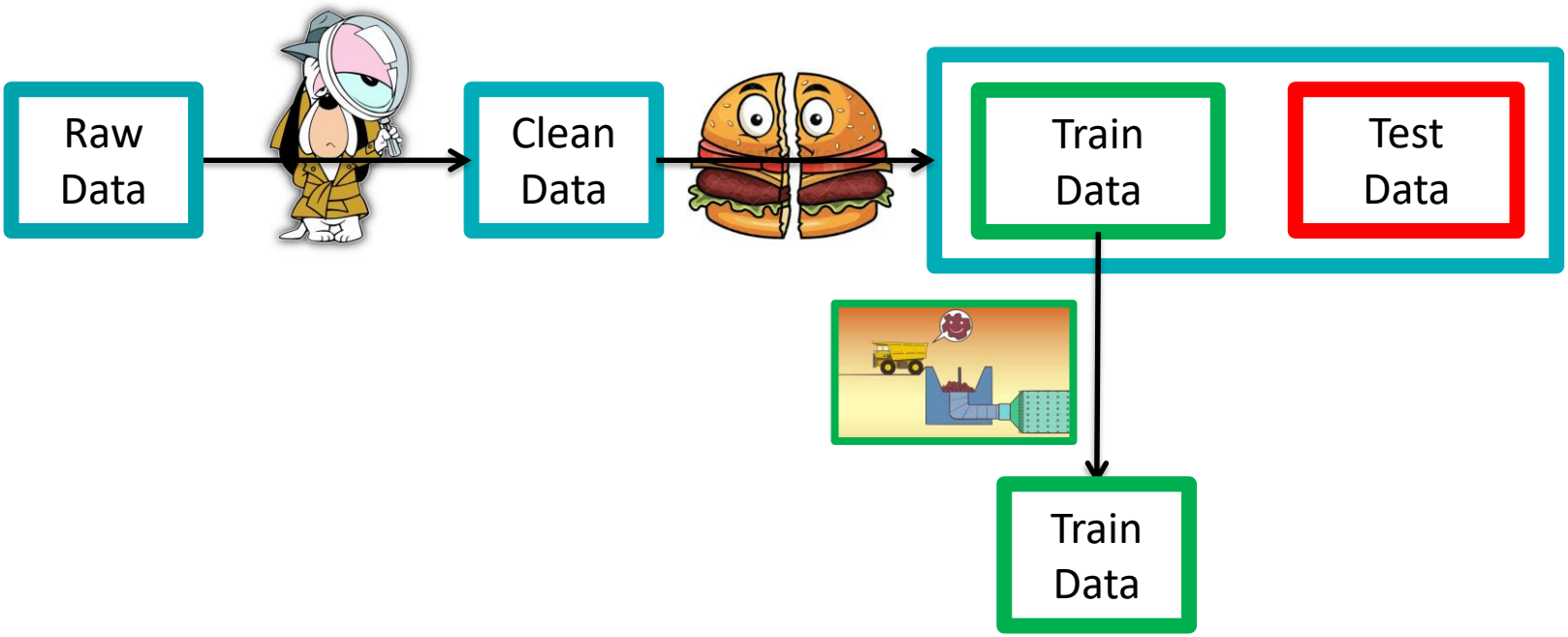
Machine Learning workflow



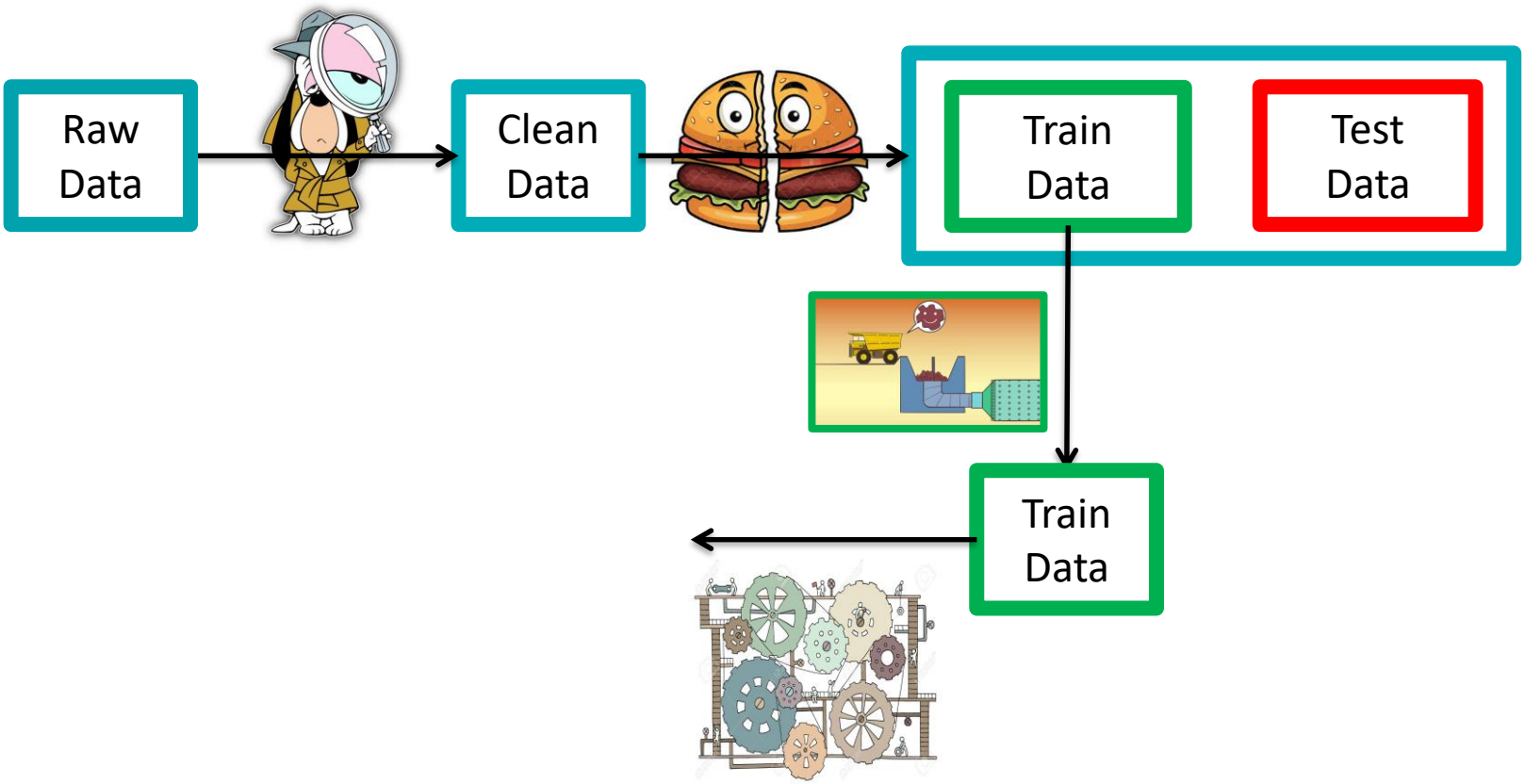
Machine Learning workflow



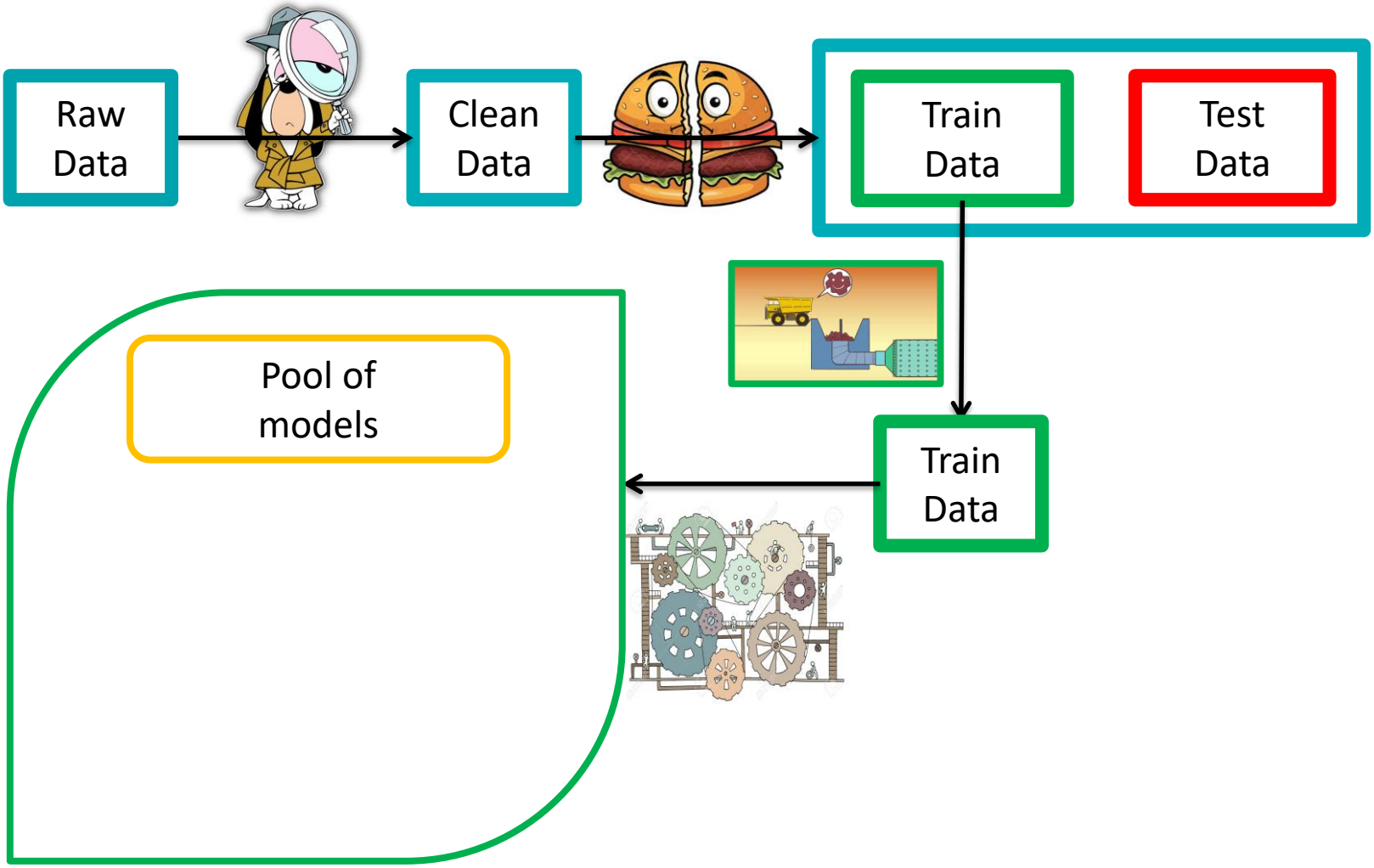
Machine Learning workflow



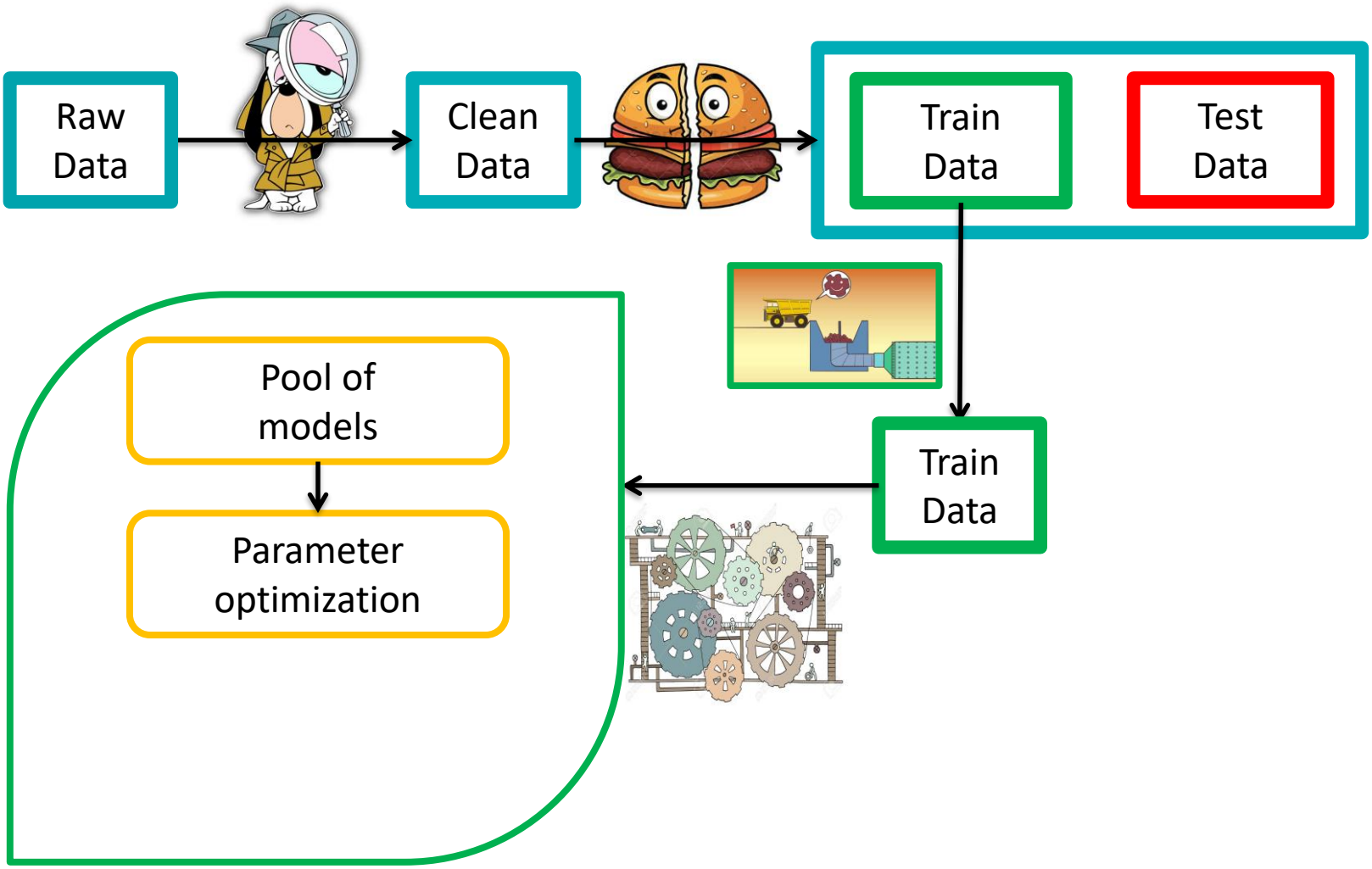
Machine Learning workflow



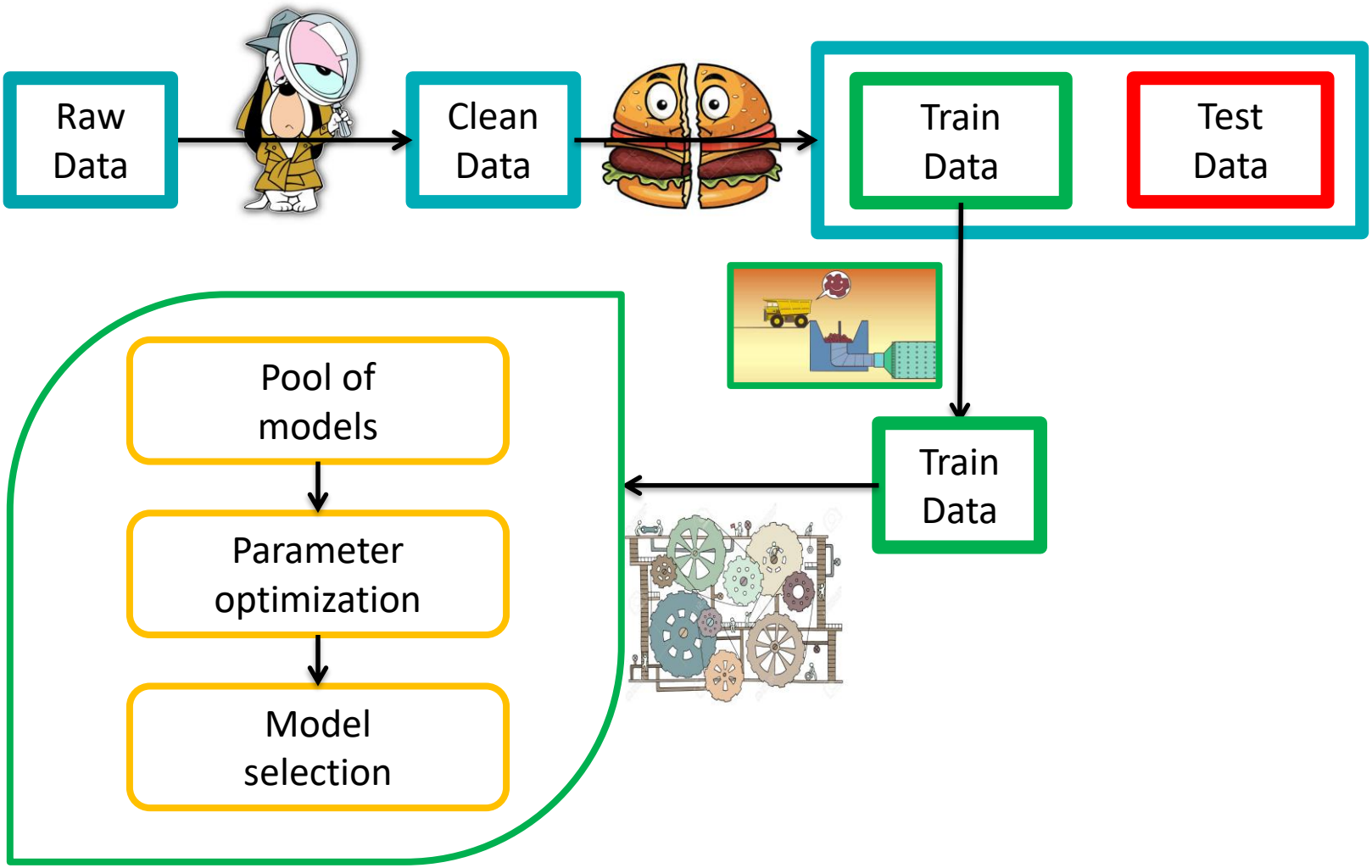
Machine Learning workflow



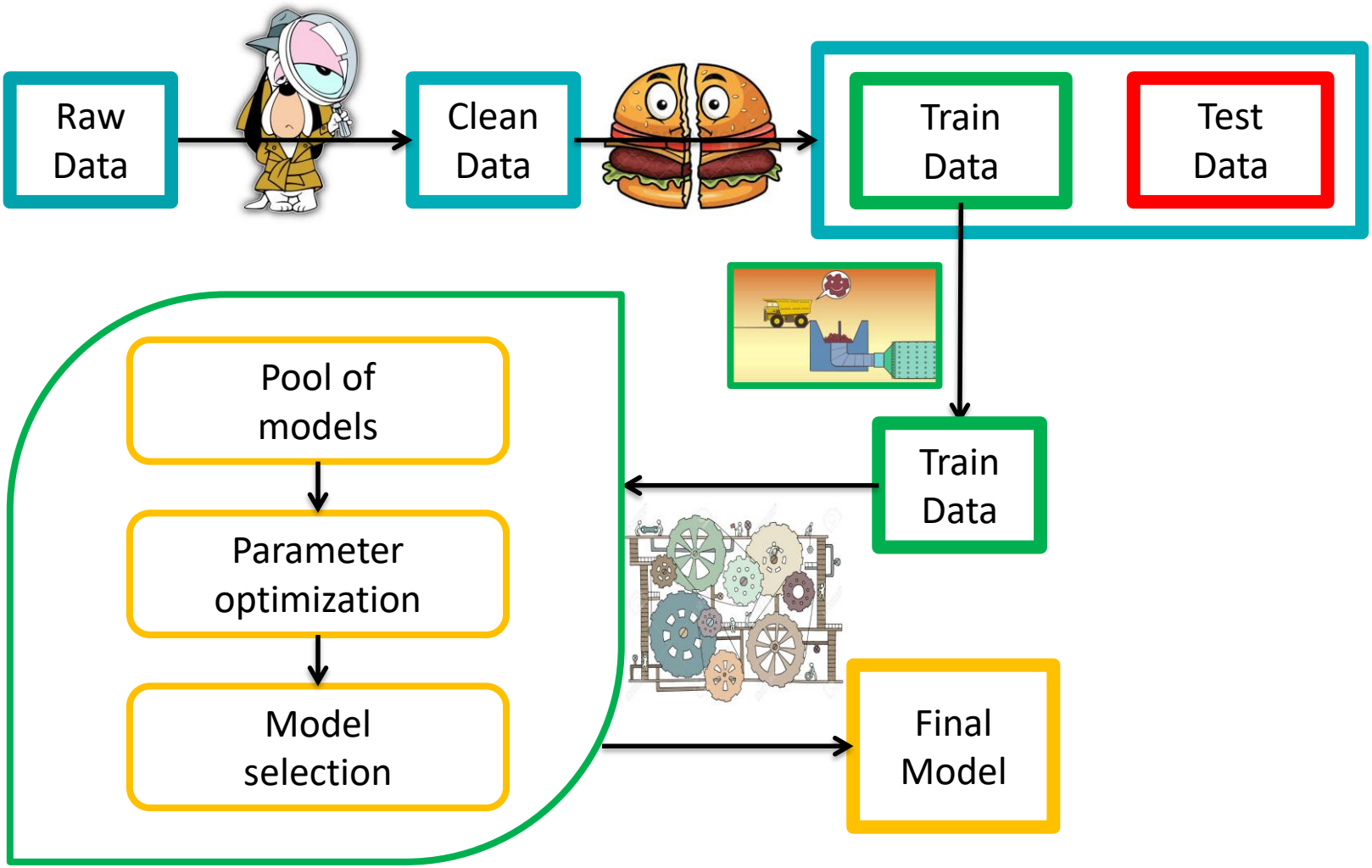
Machine Learning workflow



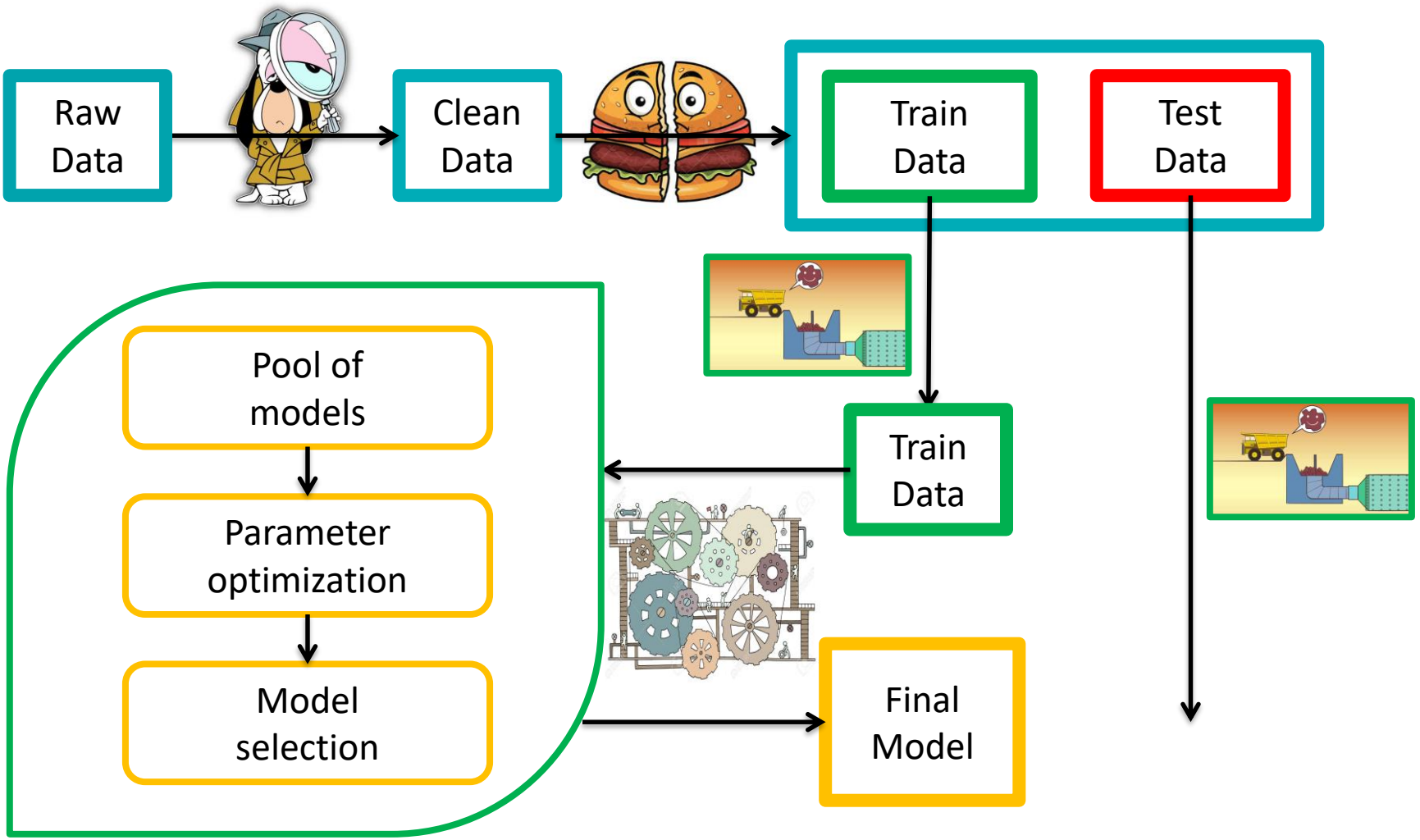
Machine Learning workflow



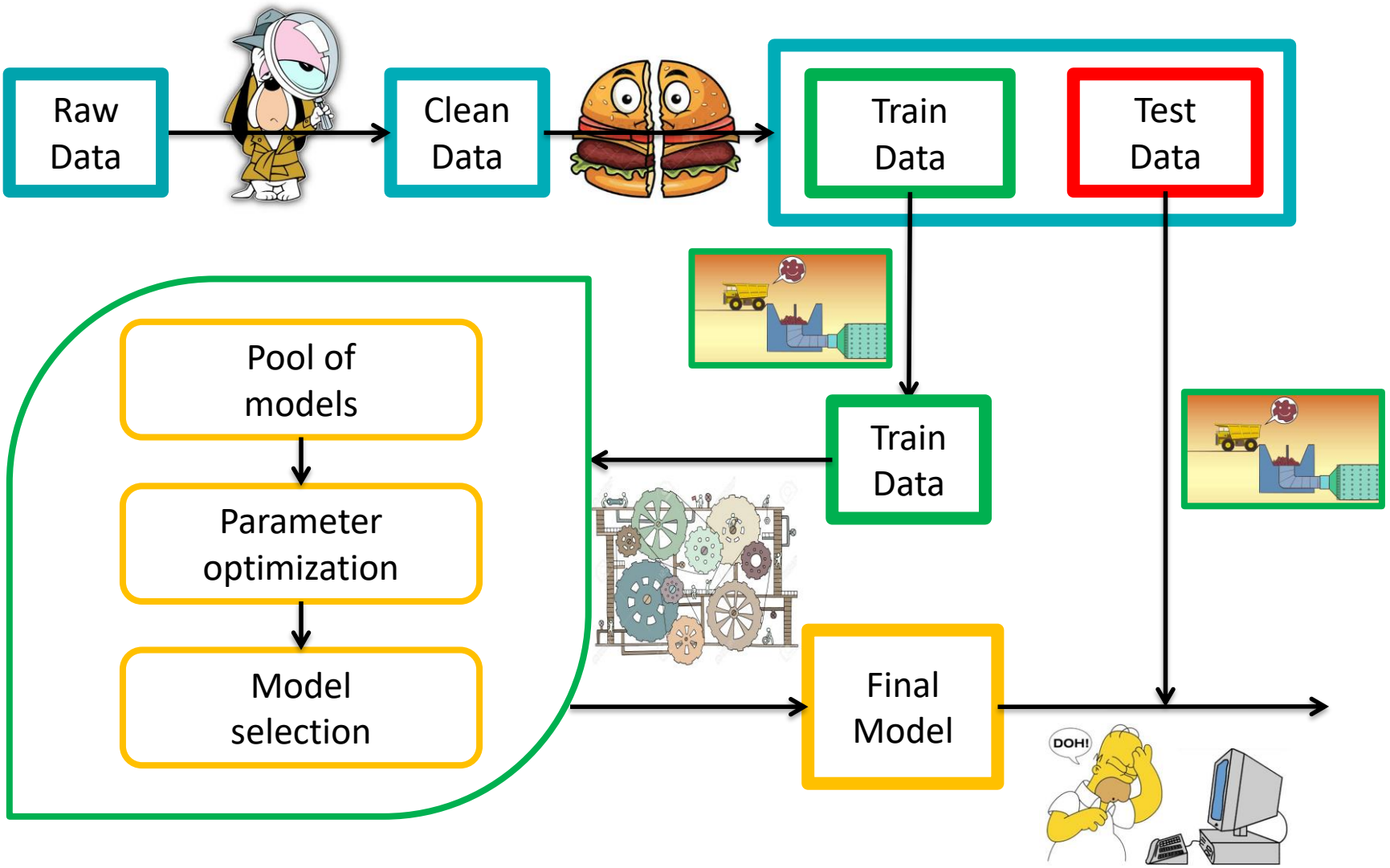
Machine Learning workflow



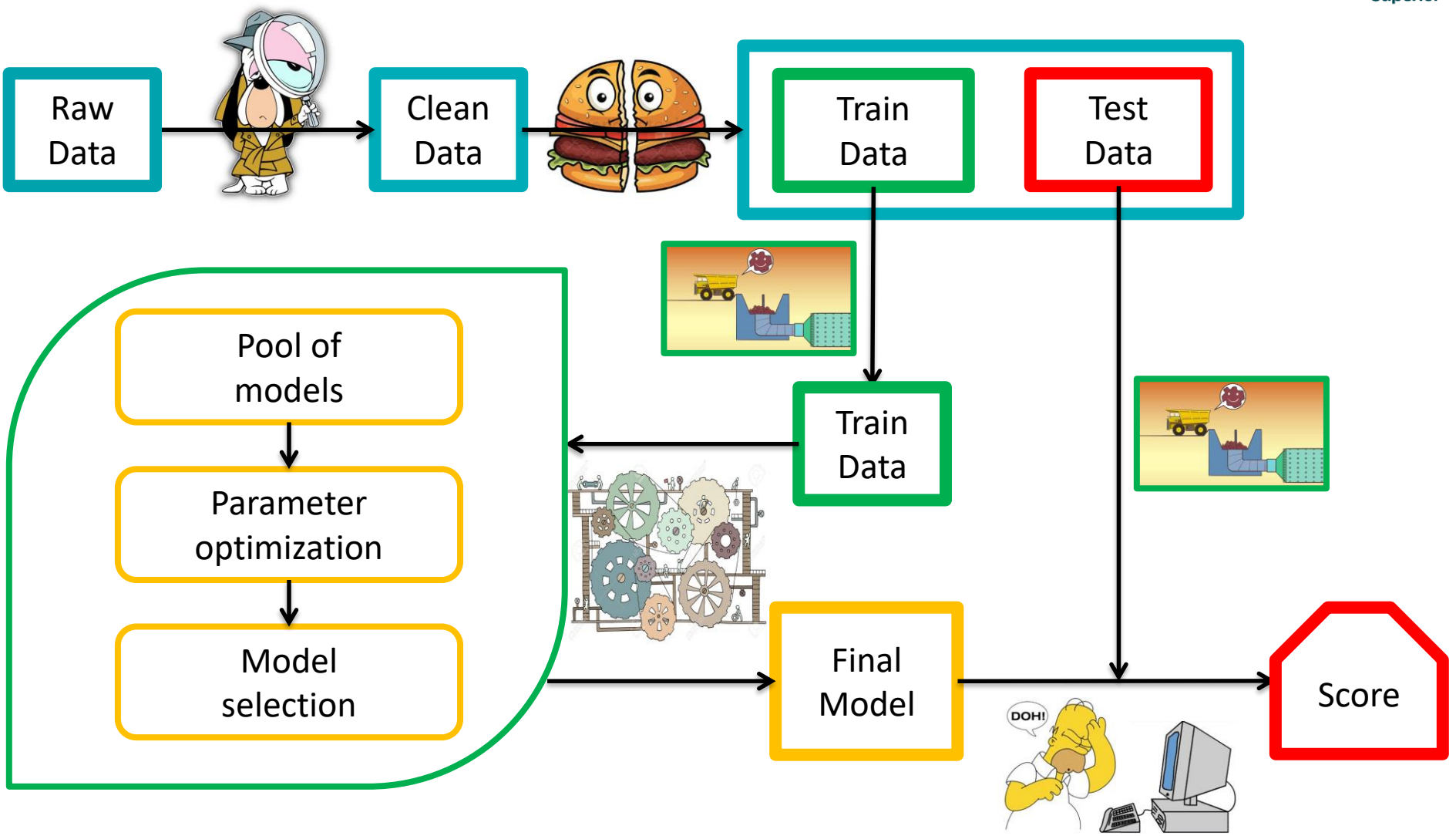
Machine Learning workflow



Machine Learning workflow



Machine Learning workflow



Calendar & schedule

Mondragon Unibertsitatea
Escuela Politécnica Superior

	Urria						Azaroa			
	5. astea		*6. astea*		*7. astea*		*8. astea*		*9. astea*	
	14	**17**	**21**	**24**	**28**	**31**	**4**	**7**	**11**	**14**
7:45-8:45	FAA	PI	AA	AAS	AAS	AAS	AA	AAS	AA	AAS
8:45-9:45	FAA	PI	AA	AAS	AAS	AAS	AA	AAS	AA	AAS
9:45-10:45	AA	AAS	SIR	SIR	SIR	SIR	SIR	SIR	VD	SIR
11:00-12:00	AA	SIR	SIR	PI	SIR	PI	SIR	PI	VD	SIR
12:00-13:00	PI	SIR	PI	PI	PI	PI	PI	PI	AAS	PI
14:15-15:15	PI	AA	PI	AA	PI	AA	PI	AA	AAS	PI
15:15-16:15	AAS	AA	AAS	AA	AA	AA	AAS	AA	PI	VD
16:30-17:30	AAS	AA	AAS	AA	AA	AA	AAS	AA	PI	VD

Preprocessing **Modeling & Validation** **Control points**

- Preprocessing control point [60'] (31/10/2019)
- Final control point [120'] (11/11/2019)

- Friday2Friday exercises [(1+2+1+2) x 180']
- Mark raising/compensating exercise [1 x ?']

Marks weights:

- Practice & exercises => 15%
- PBL => 25%
- CPs & indiv. presentations => 60%

Mondragon Unibertsitatea
Escuela Politécnica Superior

Eskerrik asko
Muchas gracias
Thank you

Carlos Cernuda
ccernuda@mondragon.edu

MGEP
Goiru, 2
20500 Arrasate – Mondragon
Tlf. 662420414