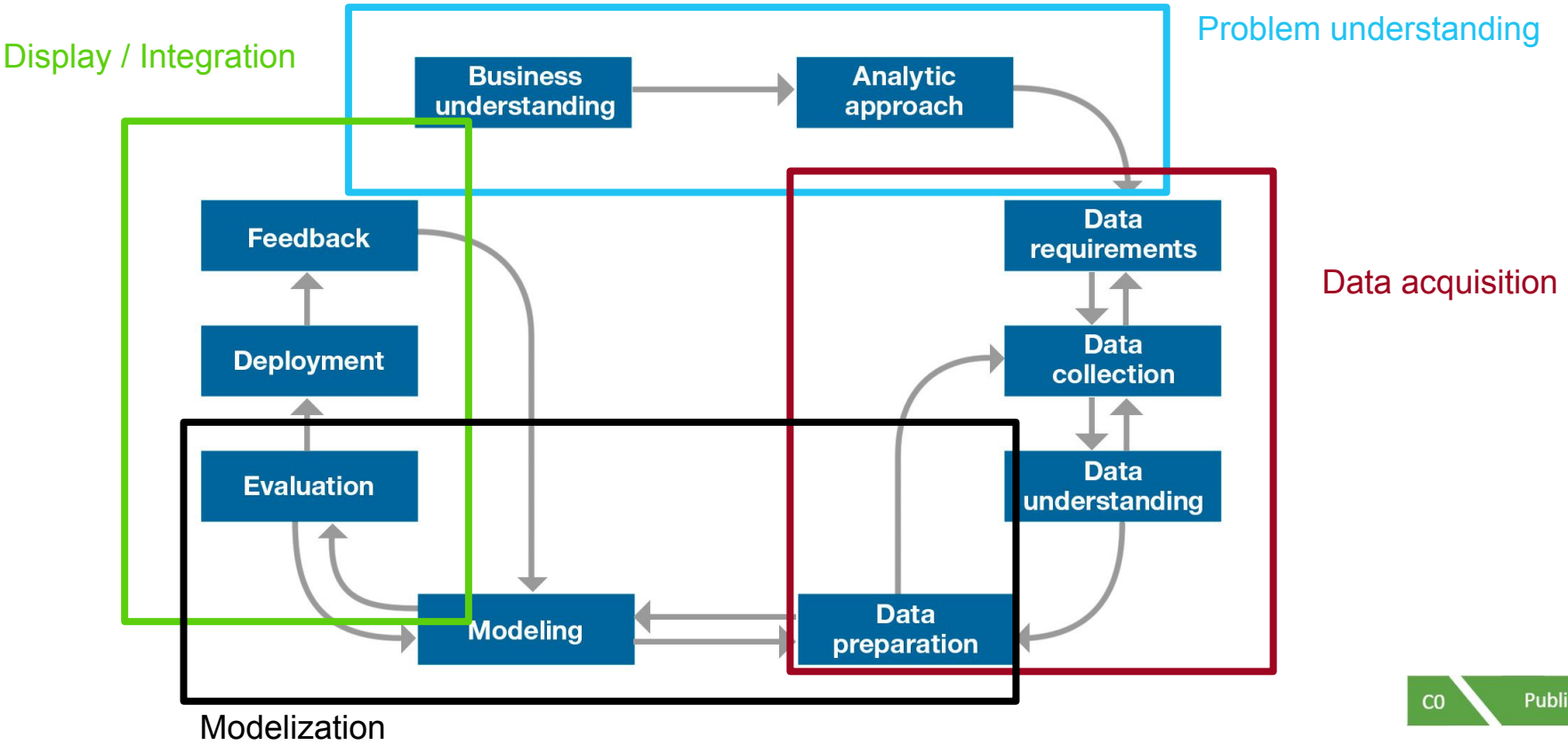
HUPI
VALORISER ANALYTIQUEMENT VOTRE DIFFÉRENCE

Data scientist presentation

2023ko Otsailaren 9a

C0 Public

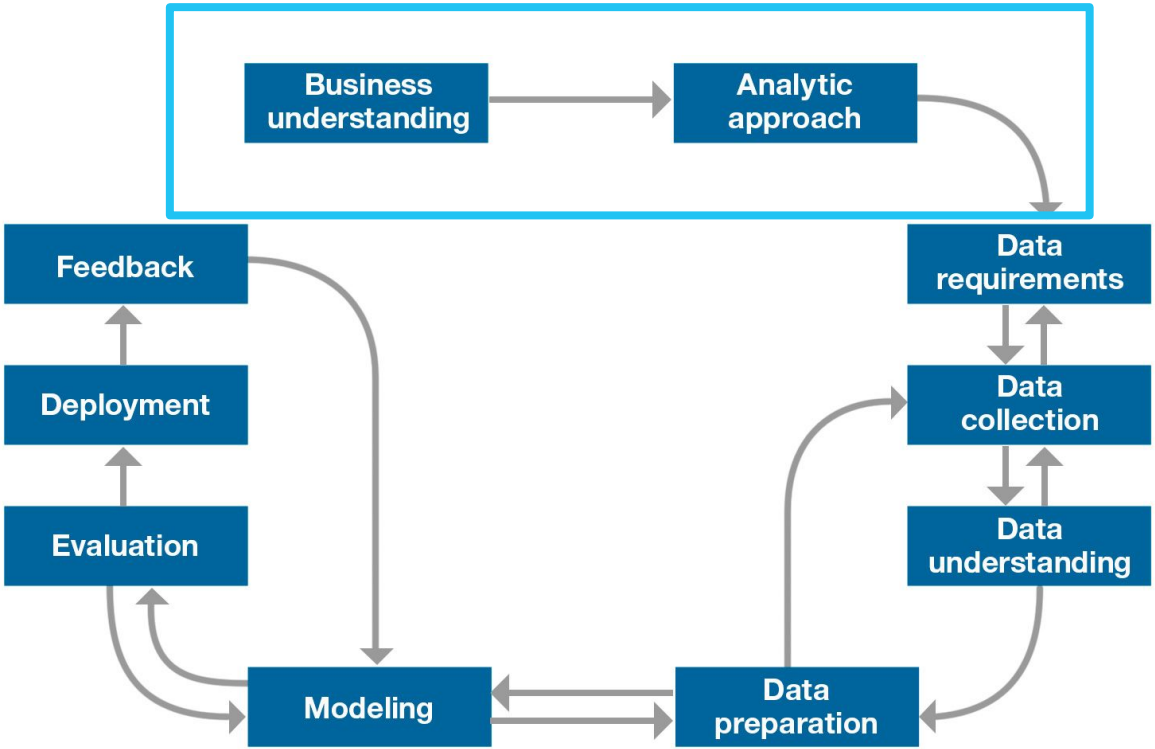
POSSIBLE ROADMAP



C0 Public

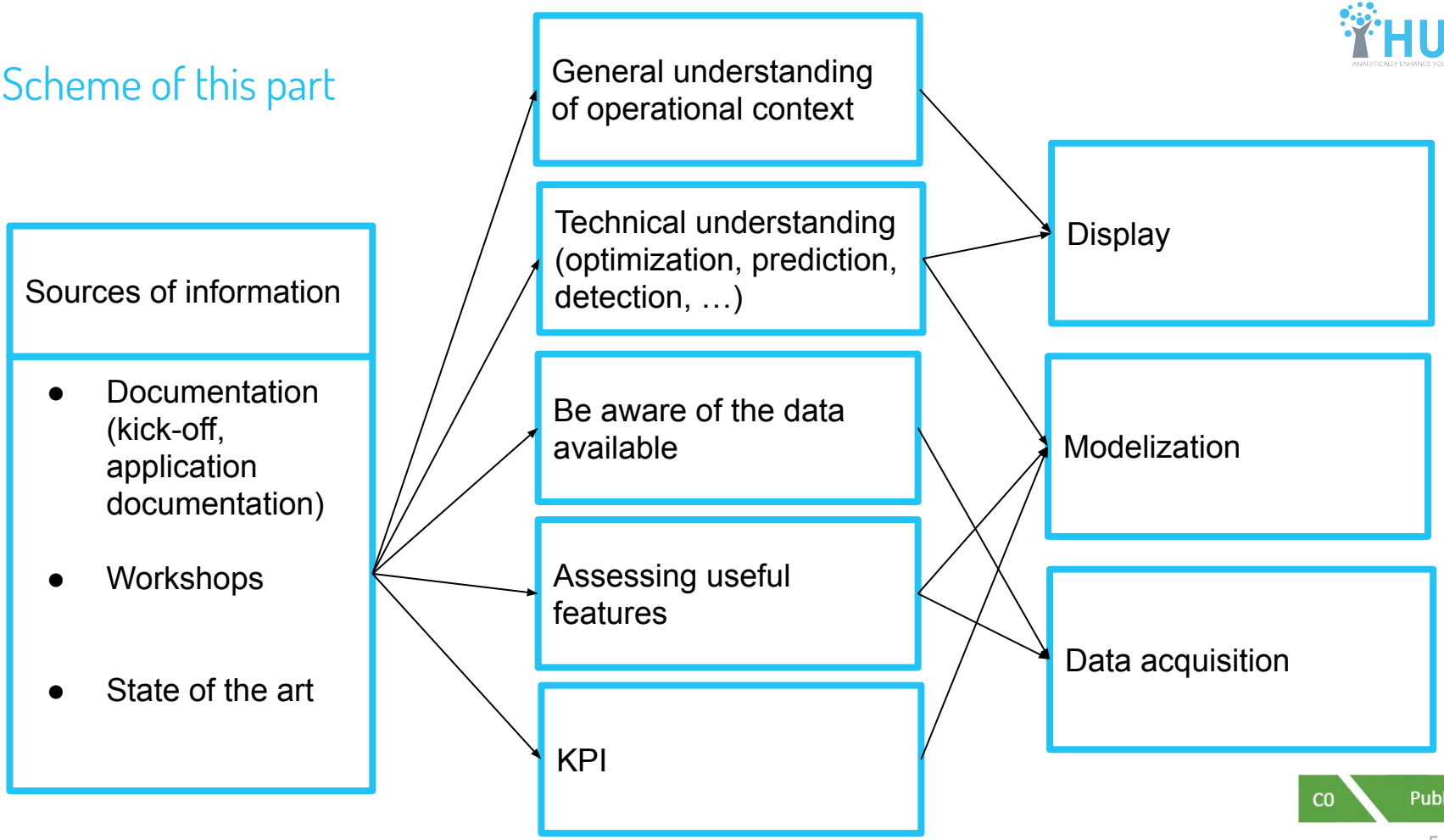
A-Problem understanding

POSSIBLE ROADMAP



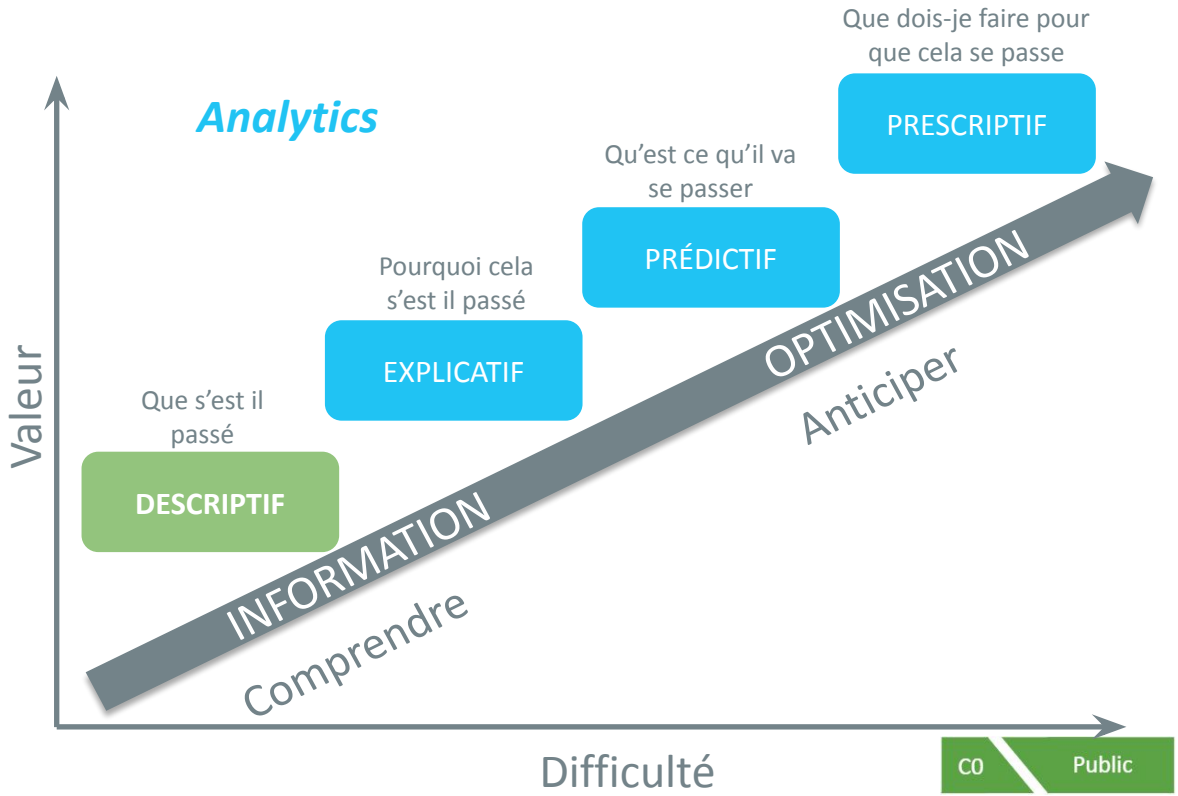
C0 Public

Scheme of this part



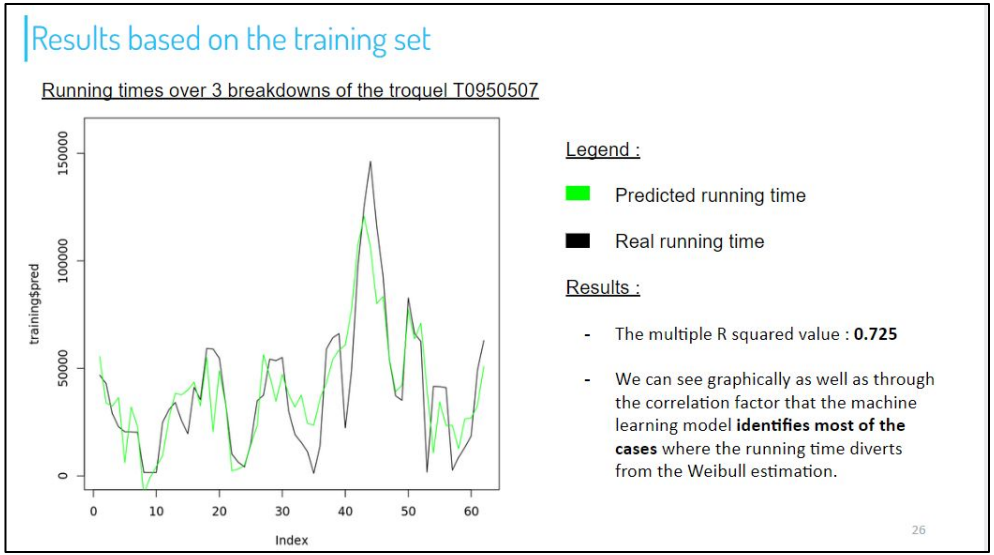
Spécifications de la solution

Dans le machine learning il y a 4 familles d'assistant virtuel, chacun dépendant de l'étape précédente. Pour faire un assistant poussant de la prescription, il faut du prédictif, de l'explicatif et du descriptif.



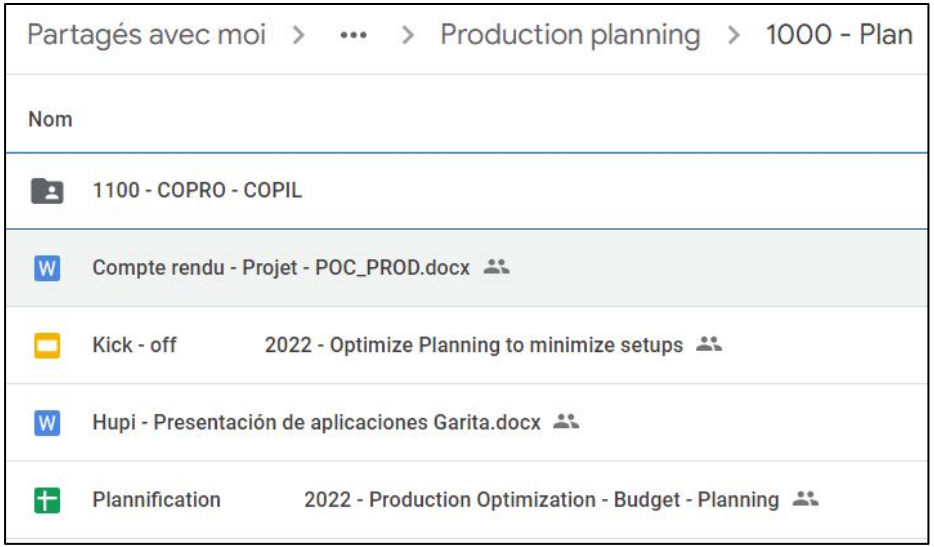
Garita : understanding the project

- 2nd phase of a project
  - Reading documentation about the first phase
    - Breakdown risk prediction (CIR, presentation)



Garita : understanding the project

- Getting to know the goal of the 2nd phase



-> Optimizing the planification to limit setups and breakdowns

Highest production and lowest delay (KPI)

Understanding the project and the operational context

PLANIFICACION SEMANAL S29/S30 2022

Hecho por:
Fecha:

	Turnos	M27	M25	M26	M20	M18	M17	M21	N1	M33	MOD
DOMINGO	22:00										5
	6:00										
LUNES	6:00	0950507	035/036015	009/0100527F10		005/006066		017089	0100504	007/0080565F10	6
	14:00		8000			8000		7250			
	14:00	001/0020600	035/036015	015057		005/006066	0010583		0100504	007/0080565F10	6
	22:00			8000							
	22:00	033/034015	035/036015	141067F10			0010585F10		0100504		7
	6:00	8000		25000			24000				
MARTES	6:00	033/034015	0600507F10	141067F10		mantenimiento	0010585F10	0030582OP20		37507/38507	6
	14:00		4000	380x1,8						6000	
	14:00	033/034015	0600507F10	001/0020565		mantenimiento	0010585F10		0100504	37507/38507	6
	22:00		440x2,2	10000							
	22:00	0030587	0840507	001/0020565	024089		0010585F10	9559	0100504		7
	6:00	4500	12000		6000			18000			
MIERCOLES	6:00	0040587	0840507	6554			0030523		0100504	37507/38507	6
	14:00	4500		4750			2500				
	14:00	0040587	0840507		041/042089	157			0100504	37507/38507	6
	22:00	515x3			4000	8000					
	22:00	097/0980507F10	857/957F10	0990554	118067	007/008066		9559	0100504		7
	6:00	10000	8000	8250	6000	10000					
JUEVES	6:00	097/0980507F10	857/957F10		001007F10	007/008066			0100504	027/028015F10	6
	14:00				20000					8000	
	14:00	097/0980507F10	0010523F10		001007F10	007/008066			0100504	027/028015F10	6
	22:00		11750								
	22:00	0960507F10	0010523F10		0060581F10	001/0020559F10	001022F10	035/036089	0100504		7
	6:00	10000			1000	4000	8000	7500			
	6:00	0960507F10	0010523F10	0440587F10			0010583	035/036089	0100504		6
	14:00			4000			40000				
	14:00	0960507F10	0010523F10		007/008525F10						

Rev.	1
Prior	Centro de Trabajo
1	M25
2	M27
3	CM2
4	M20
5	M26
6	M18
7	M17
8	M21
9	M29
10	M28

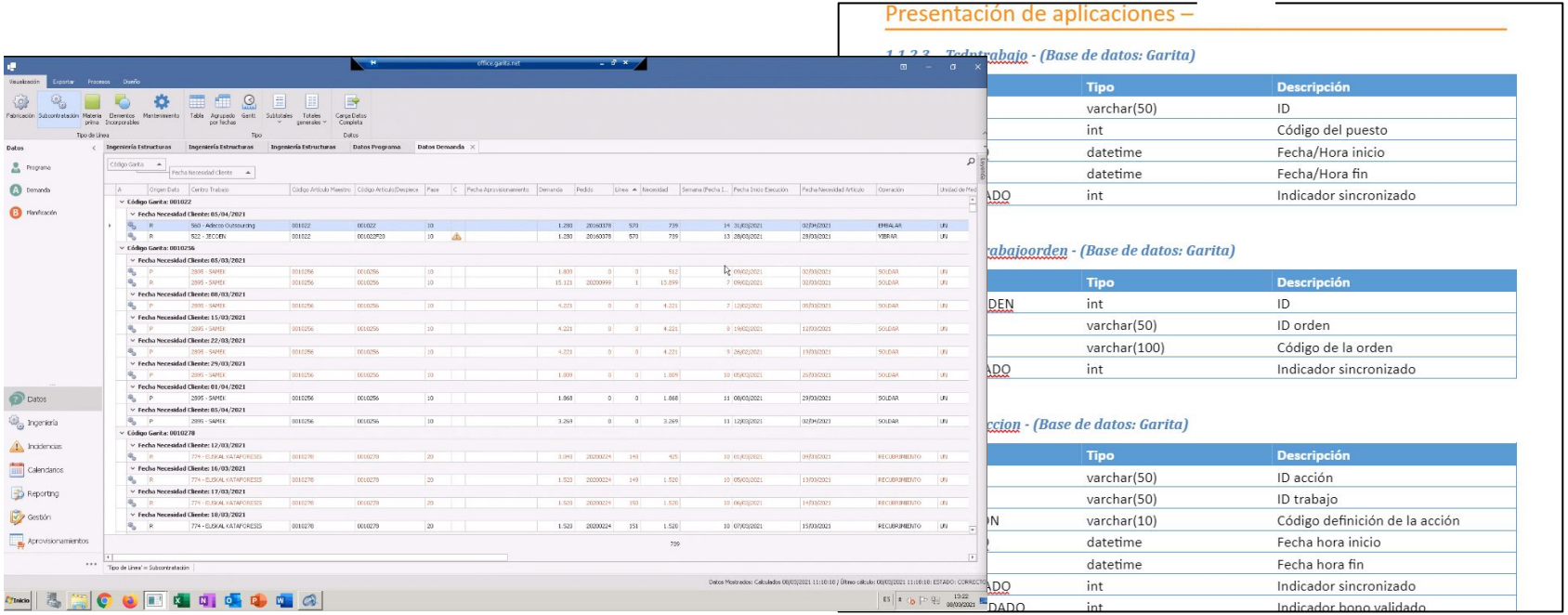
NOTAS

M27

M25

C0 Public

Assess the data available



Presentación de aplicaciones –

1.1.2.3 Tedxtrabajo - (Base de datos: Garita)

	Tipo	Descripción
	varchar(50)	ID
	int	Código del puesto
	datetime	Fecha/Hora inicio
	datetime	Fecha/Hora fin
ADO	int	Indicador sincronizado

rabajoorden - (Base de datos: Garita)

	Tipo	Descripción
DEN	int	ID
	varchar(50)	ID orden
	varchar(100)	Código de la orden
ADO	int	Indicador sincronizado

ccion - (Base de datos: Garita)

	Tipo	Descripción
	varchar(50)	ID acción
	varchar(50)	ID trabajo
N	varchar(10)	Código definición de la acción
	datetime	Fecha hora inicio
	datetime	Fecha hora fin
ADO	int	Indicador sincronizado
	DADO	int	Indicador bono validado

C0 Public

Further technical understanding with Galderak

State of the art

- Sci-hub
- Google scholar
- Researchgate
- Kaggle

-> what has been done in similar projects ?

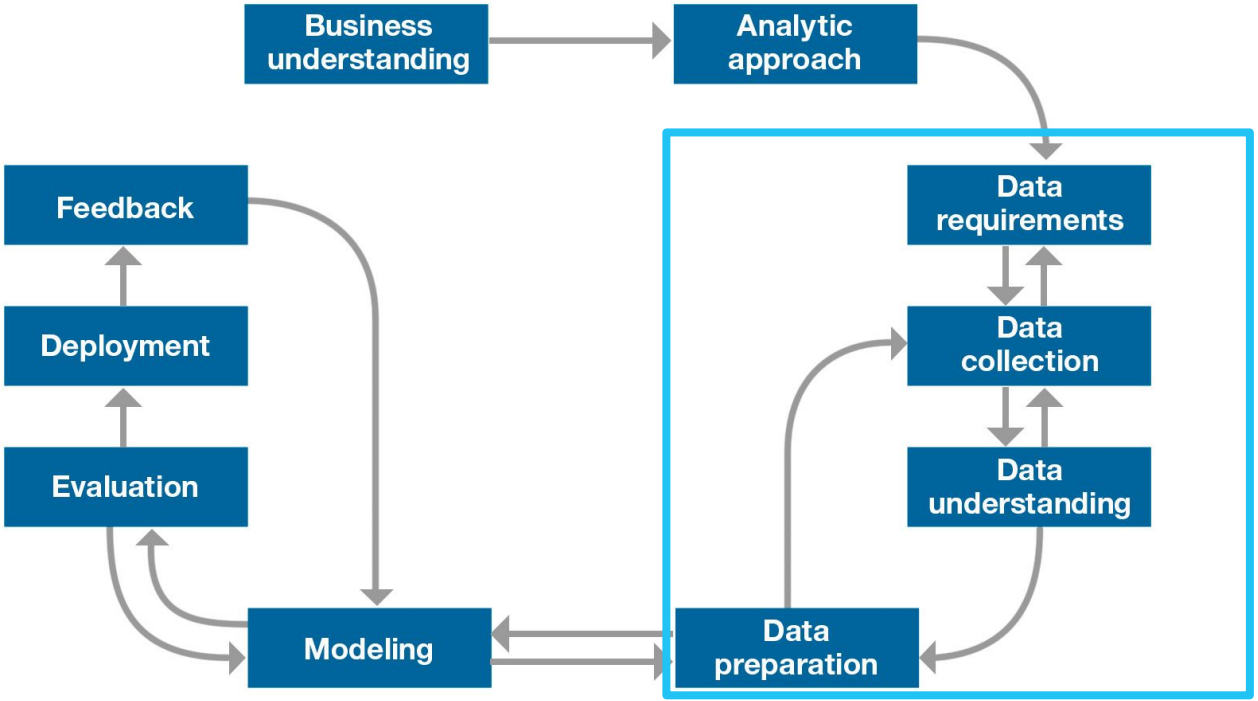
-> which technologies are used ?

"Optimizing scheduling with uncertainty"

"Multiple criteria optimization"

B-Data acquisition

ROADMAP



C0 Public

DATA TREATMENT PROCESS PRESENTATION

We will describe the following 4 main steps,

1. Data sources
   Objective: access to data
2. Data collection
   Objective: save data in Hupi
3. Data preparation
   Objective: accurate & clean data
4. Data analysis
   Objective: data validation

detailing at each step:

1. **The work done by data scientists at HUPI**
2. **The issues we may have & solutions**

WORK THROUGHOUT THE PROCESS (1 + 2)

1

Data sources
Objective: access to data

- **API** (for weather, geographical areas, …)
- **Customers** databases (SQL server, PostgreSQL, Oracle, MySQL)
- **Customers** Excel files, documents, …
- **Opendata** files

2

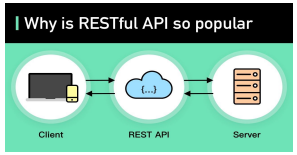
Data collection
Objective: save data in Hupi

How to proceed?

Analyze data > Select data > Save data

Where we store data?

- ***Mongo DB*** → most of the time
- ***HDFS*** → with big data



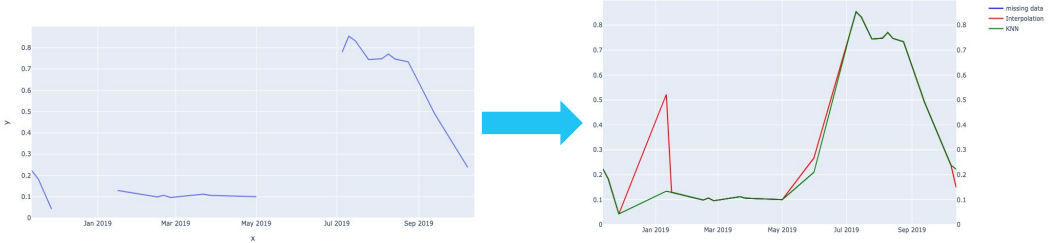
WORK THROUGHOUT THE PROCESS (3 + 4)

3

Data preparation
Objective: accurate & clean data

Transformation of raw data into usable data:

1. Identify problems: missing/inconsistent data
2. Cleaning: replace missing data, delete inconsistent data
3. Formatting: dummy encoding of variables, data typing

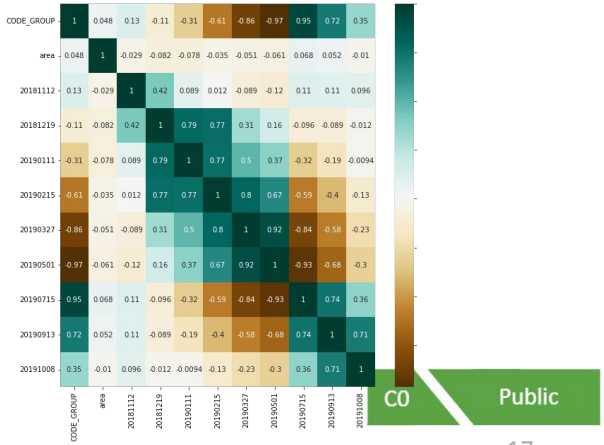


4

Data analysis
Objective: data validation

Data qualification:

Data exploration: normal behaviour detection / anomaly detection



C0 Public

ONE EXAMPLE : Prediction of bus delays on lines

1 Data sources
Objective: access to data

From a SQL Server :

DBeaver 22.2.3 - kb_145_data_billetique

Jour Jour: timestamp	Lit	Ligne	Nombre
2022-05-16 00:00:00.000	Retour	T1	3
2022-05-16 00:00:00.000	Retour	T1	1
2022-05-16 00:00:00.000	Retour	T1	2
2022-05-16 00:00:00.000	Aller	T1	2
2022-06-27 00:00:00.000	Retour	T2	2
2022-06-27 00:00:00.000	Aller	4	2
2022-05-16 00:00:00.000	Retour	T1	2
2022-05-16 00:00:00.000	Retour	T1	1
2022-05-16 00:00:00.000	Retour	T1	1
2022-05-16 00:00:00.000	Retour	T1	1
2022-05-16 00:00:00.000	Aller	T1	1
2022-05-16 00:00:00.000	Aller	T1	1

2 Data collection
Objective: save data in Hupi

Select data: select priority bus lines and not & filter the routes of buses returning to the depot

Save data:

```
db.getCollection('kb_28_tbCourse').find({})
```

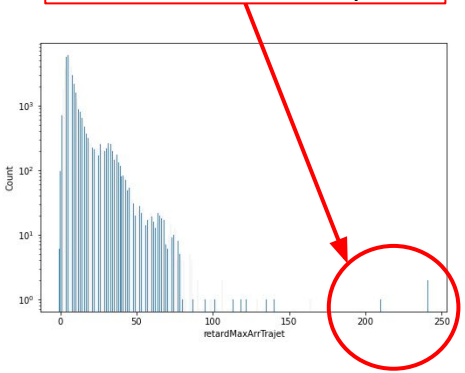
kb_28_tbCourse 0.084 sec.

Key	Value
(1) ObjectId("6256e59df2e1ff58cfd9051...	{ 12 fields }
_id	ObjectId("6256e59df2e1ff58cfd90515")
IIDCourse	4673284
stJournee	2021-04-13 00:00:00.000Z
szNomCourse	11316
szCodeHastus	727489
szNomLigne	4
stHeureDepTheo	2021-04-13 08:12:00.000Z
stHeureDepReel	2021-04-13 08:11:00.000Z
retard	-60
retardYesNo	avance
unixts	1618272000
dimTbCourse	257644

3 Data preparation
Objective: accurate & clean data

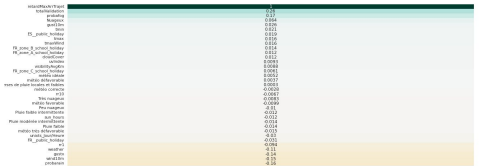
Identify & clean problems: find & remove outliers

abnormal behaviors: more than 4 hours delay



4 Data analysis
Objective: data validation

Data qualification: Correlation analysis between data collected and the target



C0 Public

ISSUES & SOLUTIONS THROUGHOUT THE PROCESS

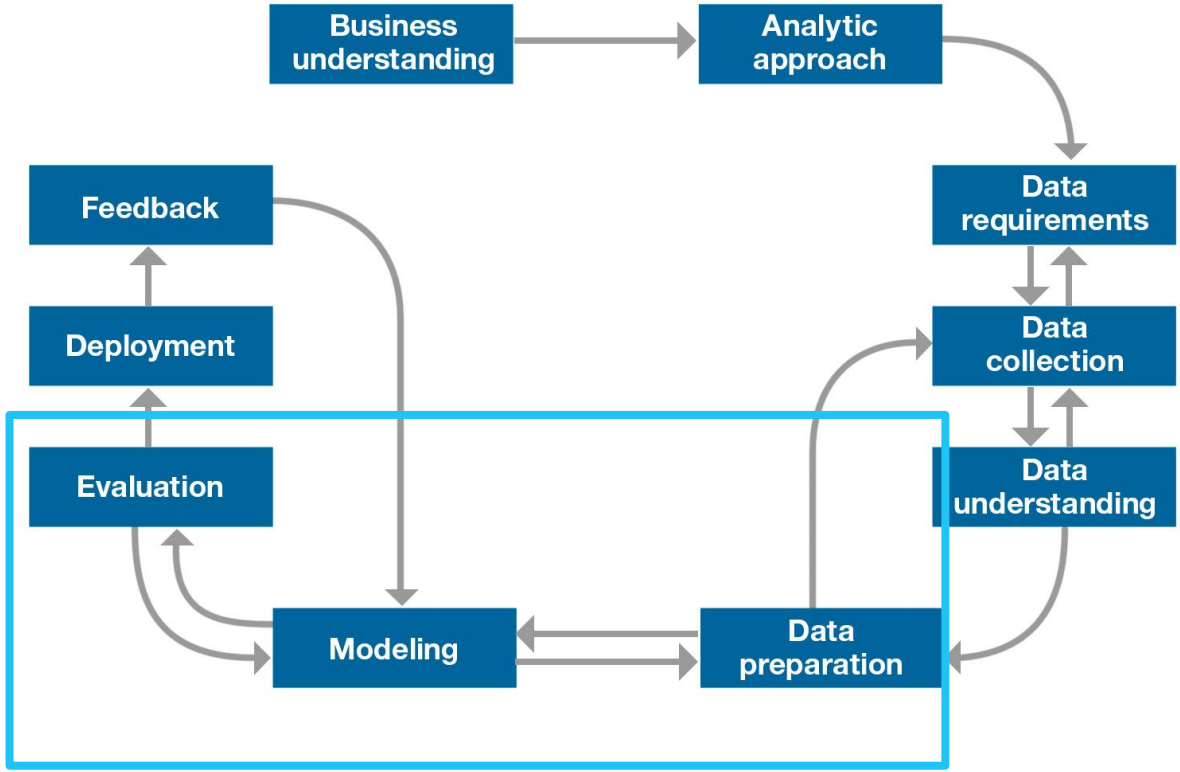
1 Data sources *Objective: access to data*	2 Data collection *Objective: save data in Hupi*	3 Data preparation *Objective: accurate & clean data*	4 Data analysis *Objective: data validation*
Problems with **database connections**	Problems of **massive storage**	**Heterogeneous data** to be processed	**Understanding** of data

Some solutions provided:

• Exchange with clients to adapt a solution • Use (& learn) different programming languages to connect	• Use (& learn) the adapted **Spark Scala language** • Use the power of HUPI's servers	• Clean up data on a case-by-case basis	• Exchange with clients with *galderak-erantzunak* platform • Search in scientific articles (state of the art)

C-Modeling

POSSIBLE ROADMAP



C0 Public

CREATION OF THE INPUT MATRIX

What is feature selection ?

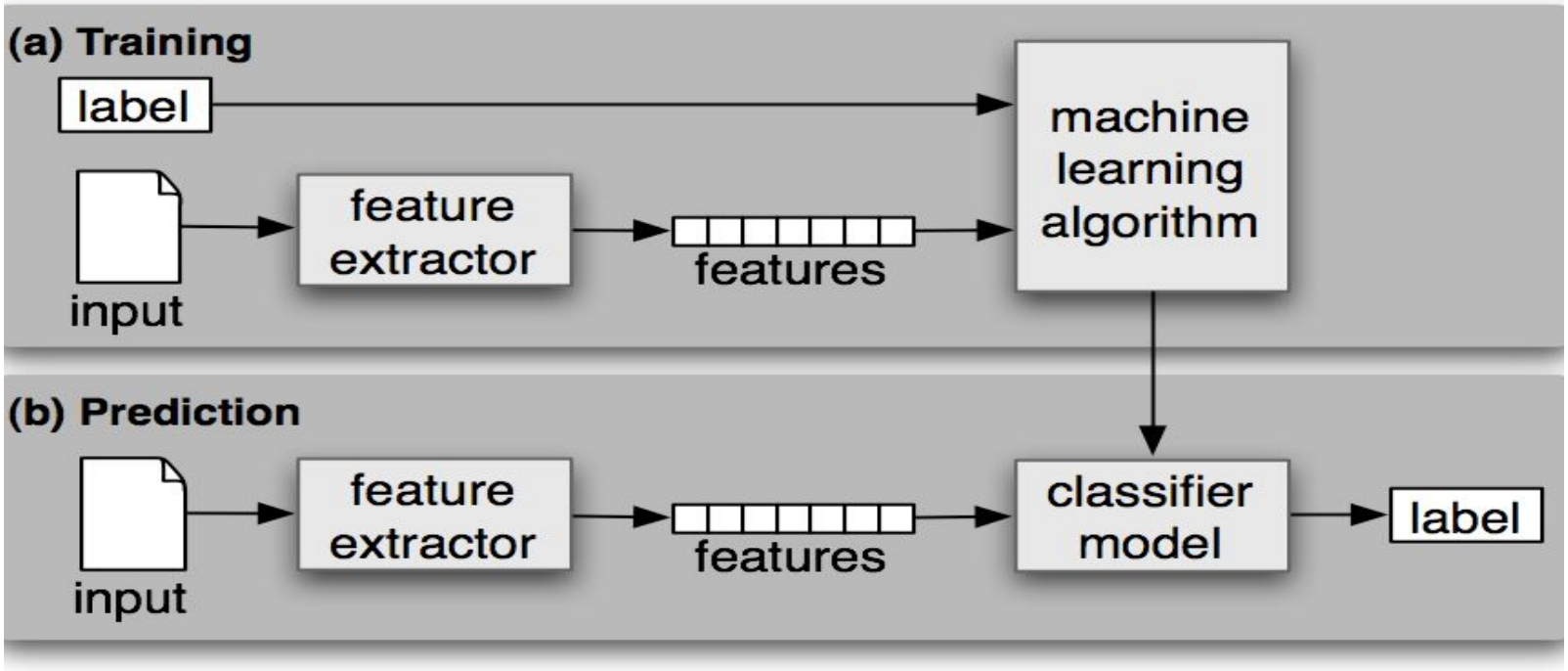
Feature selection is the process of identifying and selecting the features that contribute the most to the prediction variable or output that you are interested in, either automatically or manually.

Why feature selection ?

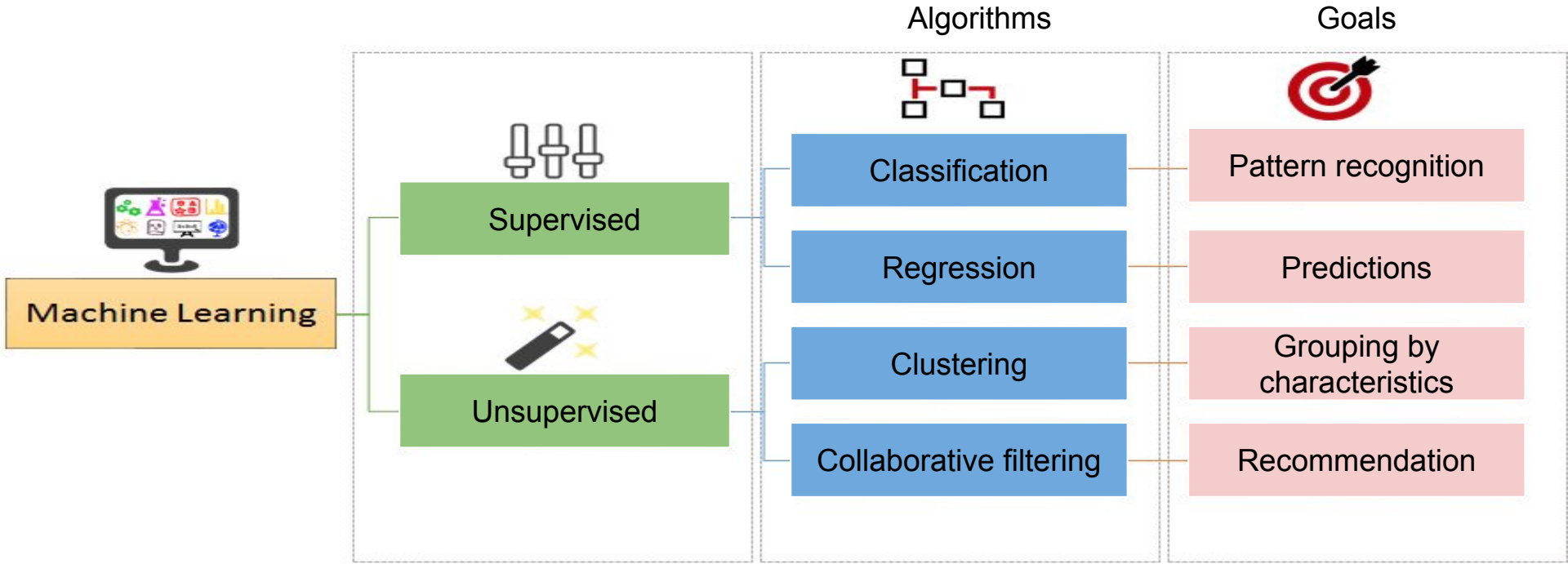
To train a model, we collect enormous quantities of data to help the machine learn better. Usually, a good portion of the data collected is noise, while some of the columns of our dataset might not contribute significantly to the performance of our model.
Further, having a lot of data can slow down the training process and cause the model to be slower. The model may also learn from this irrelevant data and be inaccurate.

	nightly_rate	property_id	nb_persons	nb_rooms
bien_id	0.089367	1.000000	0.043811	0.114561
mois	-0.005908	0.113943	-0.033987	-0.002945
price	0.981738	0.086359	0.361230	0.361920
pallier	0.981480	0.086491	0.360879	0.361859
nightly_rate	1.000000	0.089367	0.366480	0.371866
property_id	0.089367	1.000000	0.043811	0.114561
nb_persons	0.366480	0.043811	1.000000	0.793961
nb_rooms	0.371866	0.114561	0.793961	1.000000
nb_bedrooms	0.357416	0.118441	0.791099	0.982245
nb_bathroom	0.464984	0.099750	0.640176	0.651123
surface	0.397884	0.112158	0.561888	0.632538
sauna	0.222732	0.024847	0.132082	0.125360
balcony	0.056192	-0.007349	-0.007999	-0.196050
tennis	0.000290	-0.004638	0.034187	0.053254
pets	-0.072950	-0.098369	0.012027	-0.021015
fireplace	0.156403	0.016723	0.297880	0.376675

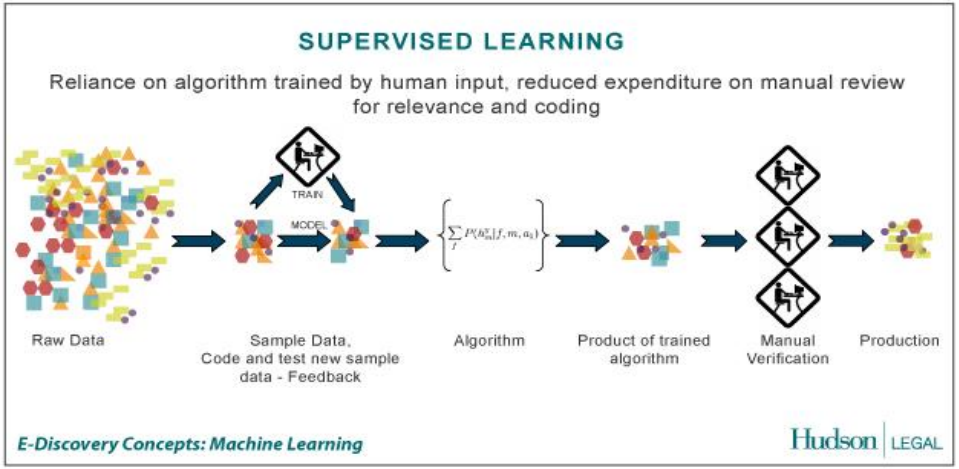
CHOICE OF MODEL

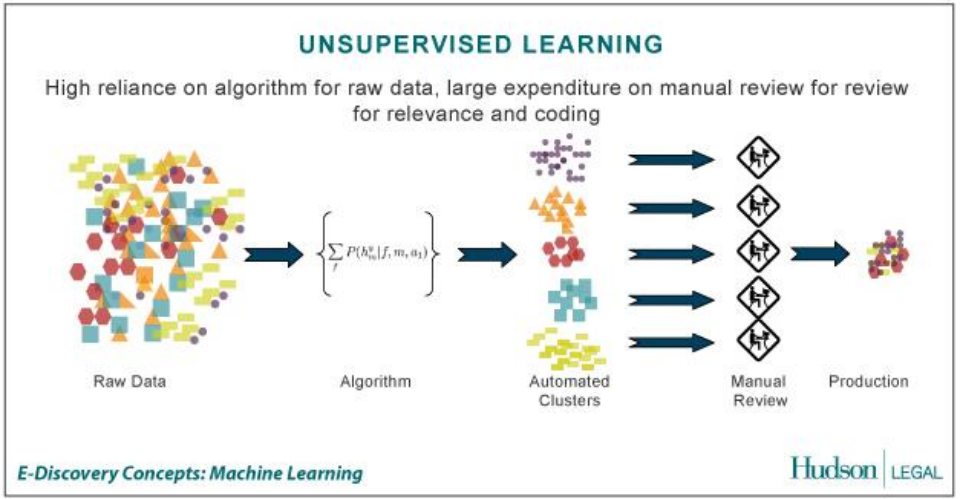


CHOICE OF MODEL

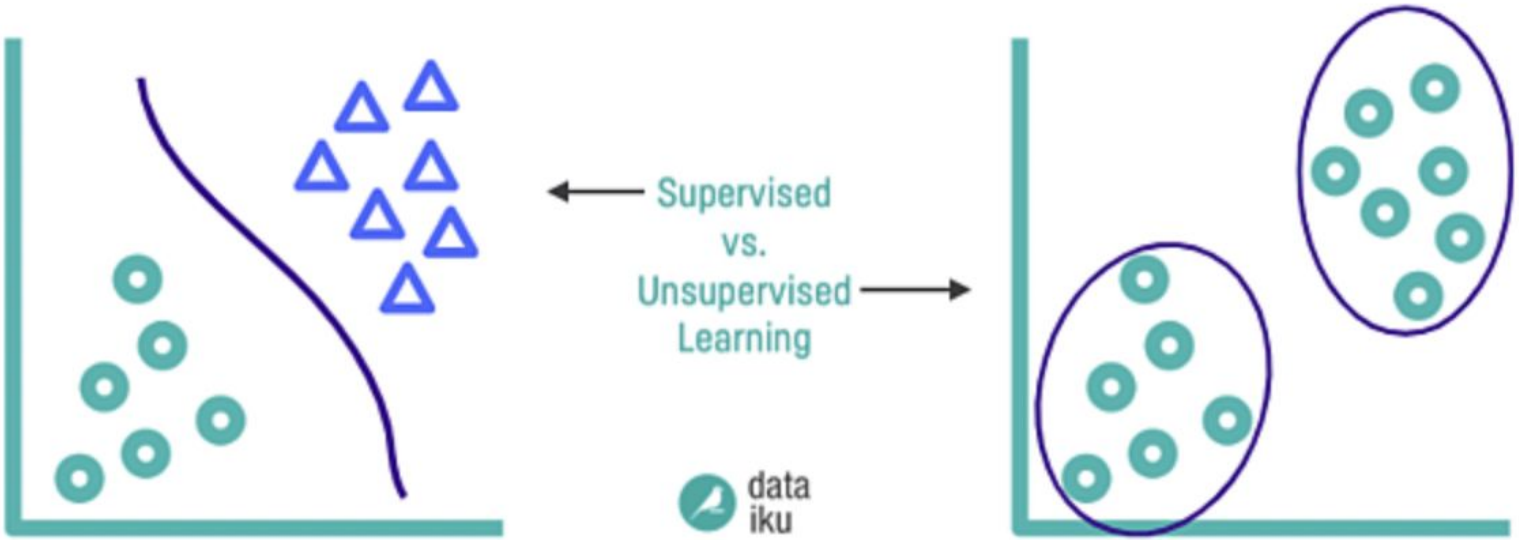


SUPERVISED VS UNSUPERVISED LEARNING





SUPERVISED VS UNSUPERVISED LEARNING



Supervised vs. Unsupervised Learning

CREATION OF STATISTICAL AND BUSINESS INDICATORS

Business indicators

Example of KPI : the number of stays sold for a dynamic pricing project

Statistical indicators

Classification :

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

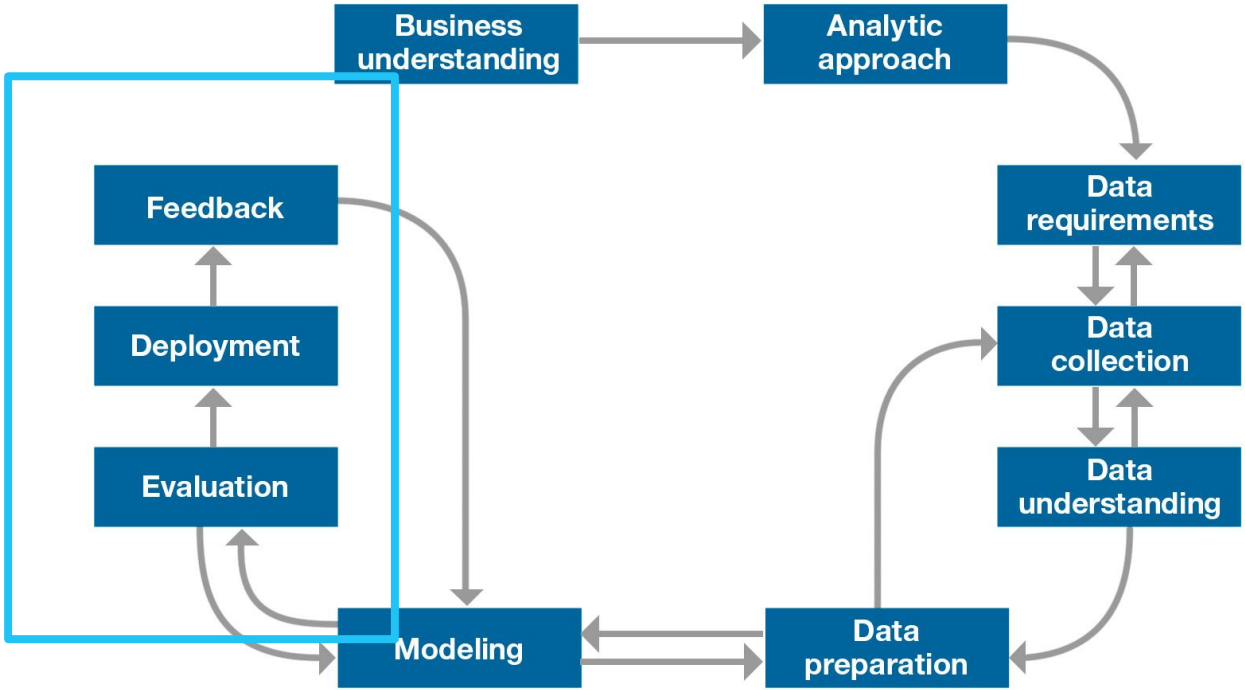
Regression metrics :

- R^2
- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)

Model performance increase doesn't always mean business growth. Monitoring and correlation AI model metrics with the business KPIs help in bridging the gap between performance analysis and business growth, integrating the whole enterprise to function more efficiently towards a set objective.

D-Display/Restitution

POSSIBLE ROADMAP

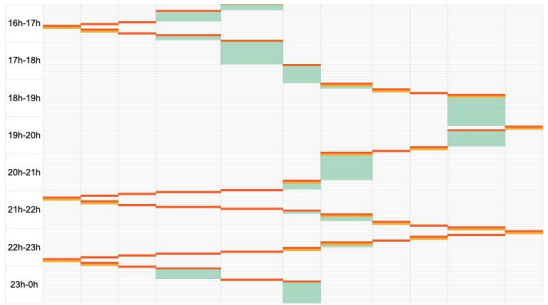


C0 Public

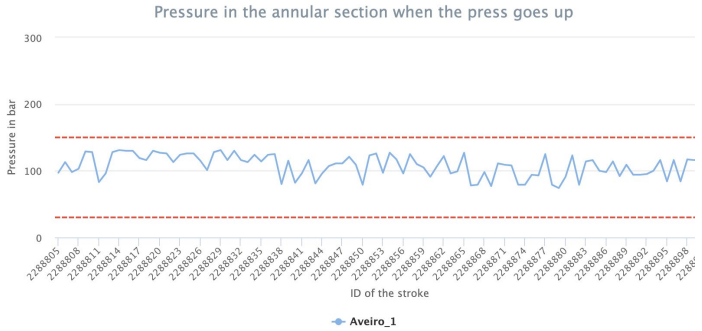
RESTITUTION OF THE RESULTS

Data sources
Objective: share data

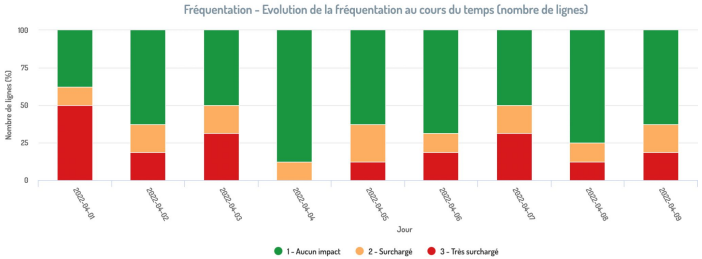
- **API** with HUPI Front
- **Widgets** on HUPI Front
- **Files** Excel files, documents, Slides …
- **Notebooks** Zeppelin



Excel file - EGIS



Graphic line - Loire Gestamp



Bar plot - Keolis-Bayonne

C0 Public

DEPLOYMENT

1 **Containerization**

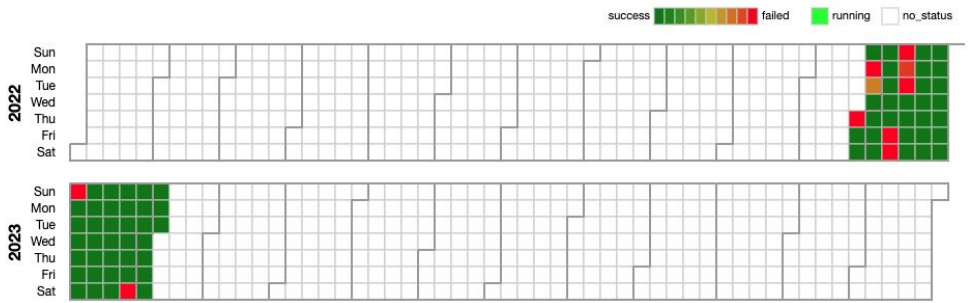
Objective: quick and efficient deployment of applications and models

2 **Automatization**

Objective: Run the jobs on automatically at a defined frequency

Apache Airflow

docker



In HUPI's servers or client's servers !

ex: Training 1 x month
Predictions 1 x day

C0 Public

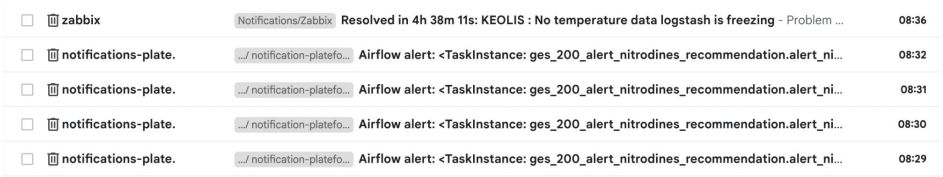
MONITORING

Checks the performance evolution of the models
Objective: Detect Big changes

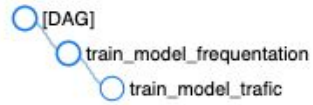
Supervise the models
Objective: Detect bugs

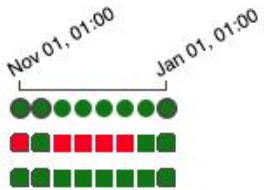
Mois	Valeur prédite	Valeur réelle	Taux d'erreur
Février	234929.0	66420.0	253.7 %
Janvier	542290.0	471927.0	14.91 %
Décembre	553806.0	551063.0	0.5 %
Novembre	544990.0	543119.0	0.34 %

Error percentage of passengers number prediction - Keolis Bayonne



Apache Airflow





FEEDBACK

Weekly Meetings
Objective: Vulgarize results with the clients

- **Frequency** ≈ 30 min each week
- **Support** Notebooks, slides, documents …
- **Decisions**
  - Add additional features
  - Bug reports
  - Suggestions for improvements

New version of the model !