Workshop on "Big Data & Machine Learning Applications for Central Banks"

October 22nd 2019

Centro Carlo Azeglio Ciampi

Institutional sector classification

A Machine Learning Application

Paolo **Massaro** | Divisione Informazioni Anagrafiche
Dipartimento **ECS**

Oliver **Giudice** | Divisione Ricerca sulle Tecnologie Avanzate
Dipartimento **IT**

BANCA D'ITALIA
EUROSISTEMA

Problem statement

Given

a set of **features** of a company

Numeric and non-numeric: name, number of employees, balance sheet data, whether publicly held or not, etc.

Determine

the appropriate **SAE** code to assign to it

SAE="SETTORE DI ATTIVITA' ECONOMICA" is a code defined by Circ. 140/97 meant to cluster companies into one of 116 "institutional sectors" (e.g., public institution, productive company, financial holding, etc.)

Problem statement

Given

a set of **features**
of a company

Numeric and non-numeric: name, number of employees, balance sheet data, whether publicly held or not, etc.

Determine

the appropriate **SAE** code
to assign to it

SAE="SETTORE DI ATTIVITA' ECONOMICA" is a code defined by Circ. 140/97 meant to cluster companies into one of 116 "institutional sectors" (e.g., public institution, productive company, financial holding, etc.)

Machine Learning approach

SAE

We start from existing **data**; a "**machine learning model**" is trained from companies already labeled (by hand); on the basis of this "past experience" it learns to predict what SAE any new company belongs to

Provided

the machine is given several (**tens of thousands** of)
prior samples of correctly labeled companies

Why and when should ML help here?

Machine Learning approach

We start from existing **data**; a "**machine learning model**" is trained from companies already labeled (by hand); on the basis of this "past experience" it learns to predict what SAE any new company belongs to

AS-IS: Who classifies companies into SAEs?

Type of company	Classified by Institution	Data trustworthiness
Public Administrations	ISTAT	authoritative
Supervised Entities	Bank of Italy	authoritative
Other (Productive Companies, non-supervised Companies, etc…)	Financial Intermediaries	may be: ▶ incorrect → fix ▶ inconsistent → spot ▶ stale → update ▶ missing (~30%) → autofill

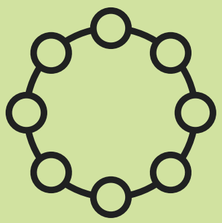
Why and when should ML help here?

Machine Learning approach

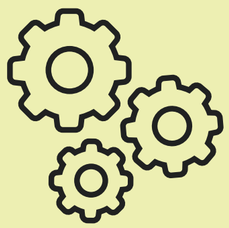
We start from existing **data**; a "**machine learning model**" is trained from companies already labeled (by hand); on the basis of this "past experience" it learns to predict what SAE any new company belongs to

AS-IS: Who classifies companies into SAEs?

Type of company	Classified by Institution	Data trustworthiness
Public Administrations	ISTAT	authoritative
Supervised Entities	Bank of Italy	authoritative
Other (Productive Companies, non-supervised Companies, etc…)	Financial Intermediaries	may be: incorrect → fix inconsistent → spot stale → update missing (~30%) → autofill

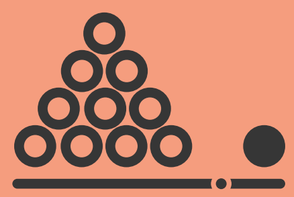


Data



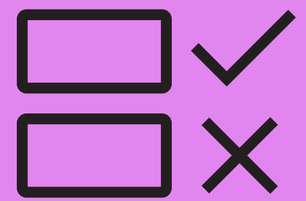
Preprocessing

Feature extraction



Imbalanced learning

Classification



Results

Original datasets

Dataset	#	Origin
Anagrafe Soggetti	**42**M	Bank of Italy
Listed Companies	**1**K	Bank of Italy
ATECO	**3.6**M	Ag. Entrate
Balance Sheet et al.	**2.2**M	CERVED
Info Imprese	**2.2**M	INFOCAMERE

Original datasets

Platform

Dataset	#	Origin
Anagrafe Soggetti	**42**M	Bank of Italy
Listed Companies	**1**K	Bank of Italy
ATECO	**3.6**M	Ag. Entrate
Balance Sheet et al.	**2.2**M	CERVED
Info Imprese	**4.8**M	INFOCAMERE

Data

Data ingestion (ETL)

Dataset	#	Origin
Anagrafe Soggetti	**42**M	Bank of Italy
Listed Companies	**1**K	Bank of Italy
ATECO	**3.6**M	Ag. Entrate
Balance Sheet et al.	**2.2**M	CERVED
Info Imprese	**4.8**M	INFOCAMERE

Platform

Big Data Analytics Platform

open source

Op

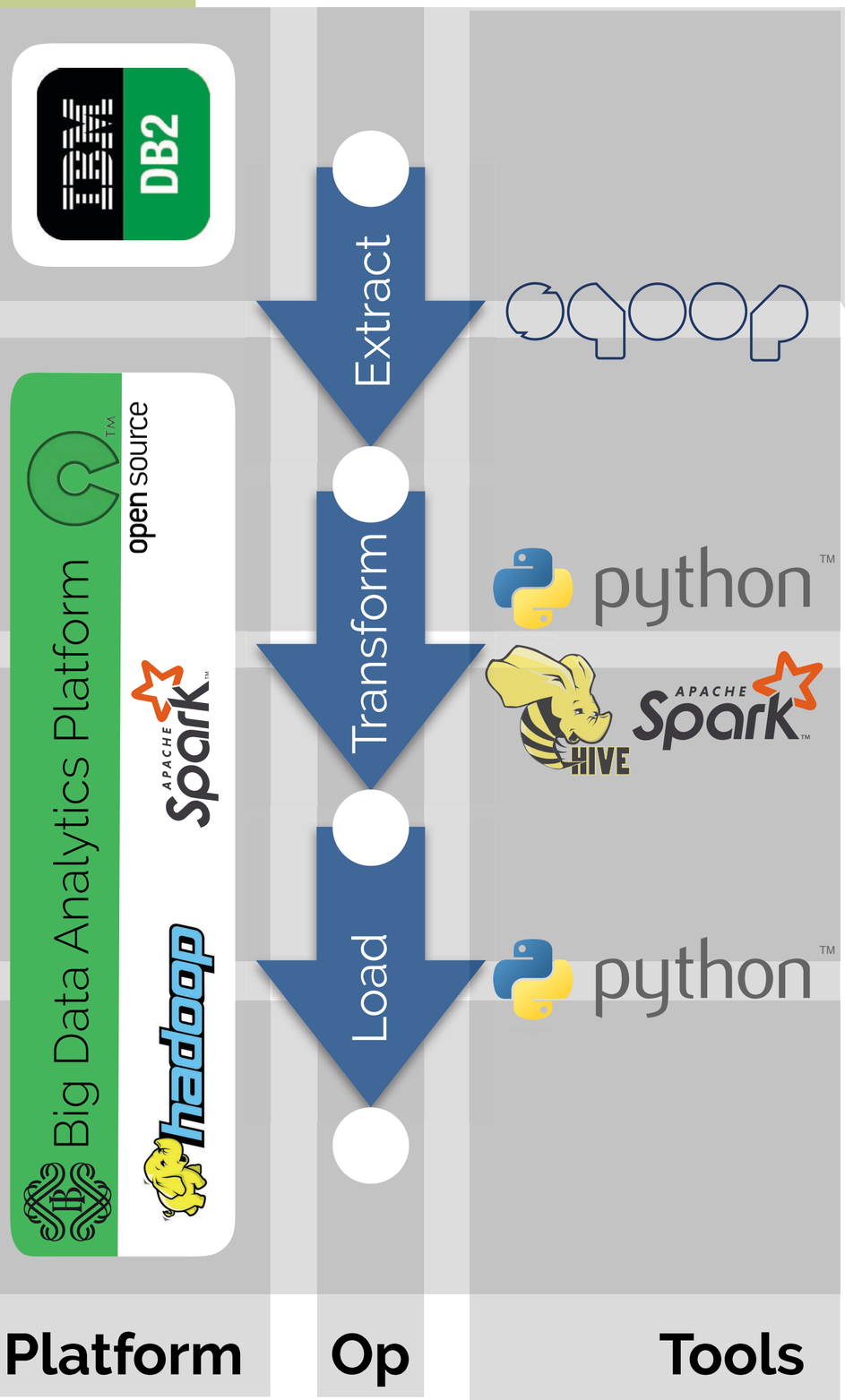
Extract

Transform

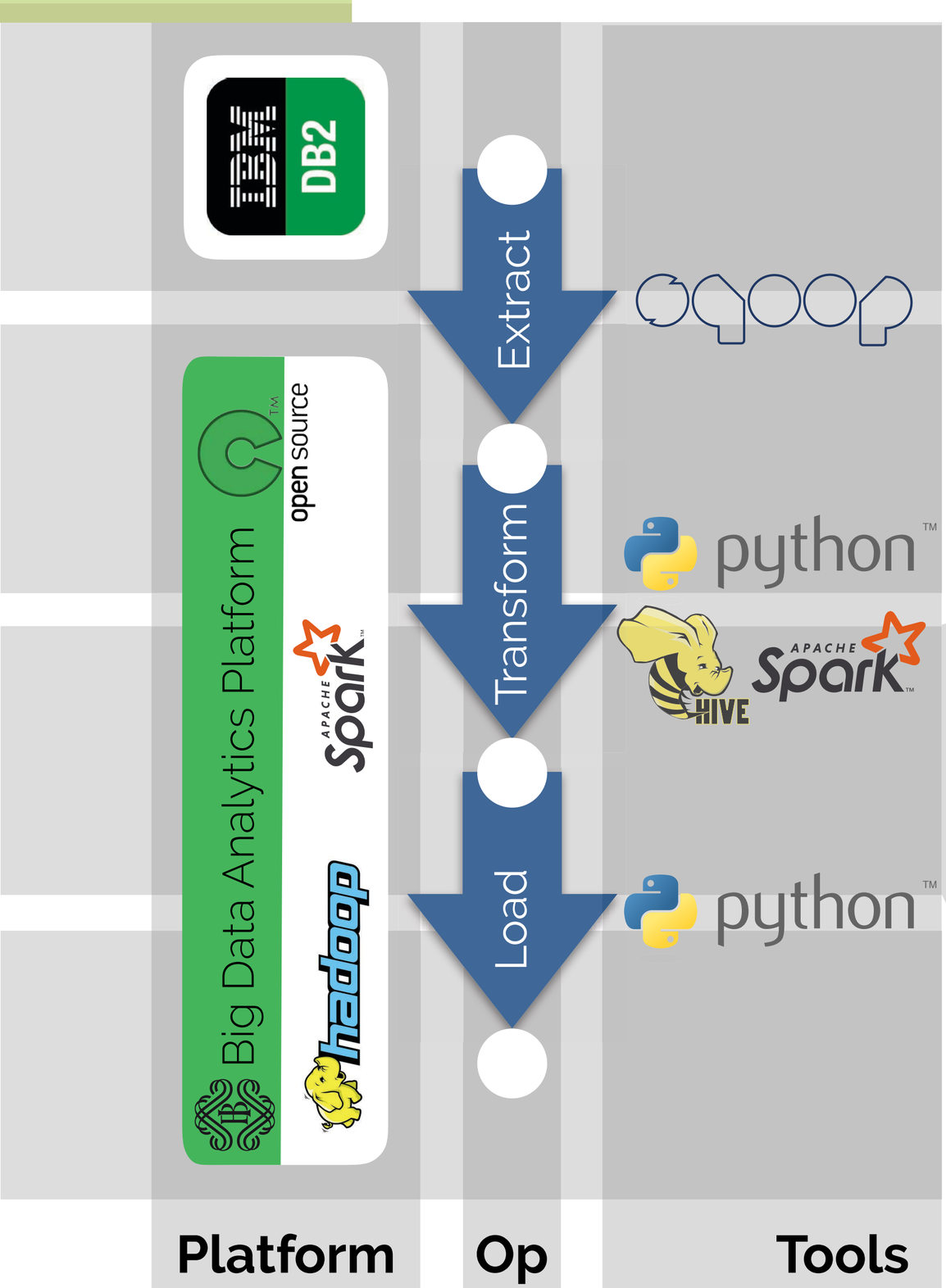
Load

Data ingestion (ETL)

Dataset	#	Origin
Anagrafe Soggetti	**42**M	Bank of Italy
Listed Companies	**1**K	Bank of Italy
ATECO	**3.6**M	Ag. Entrate
Balance Sheet et al.	**2.2**M	CERVED
Info Imprese	**4.8**M	INFOCAMERE



Data ingestion (ETL)



Input to ML machinery

Single text file

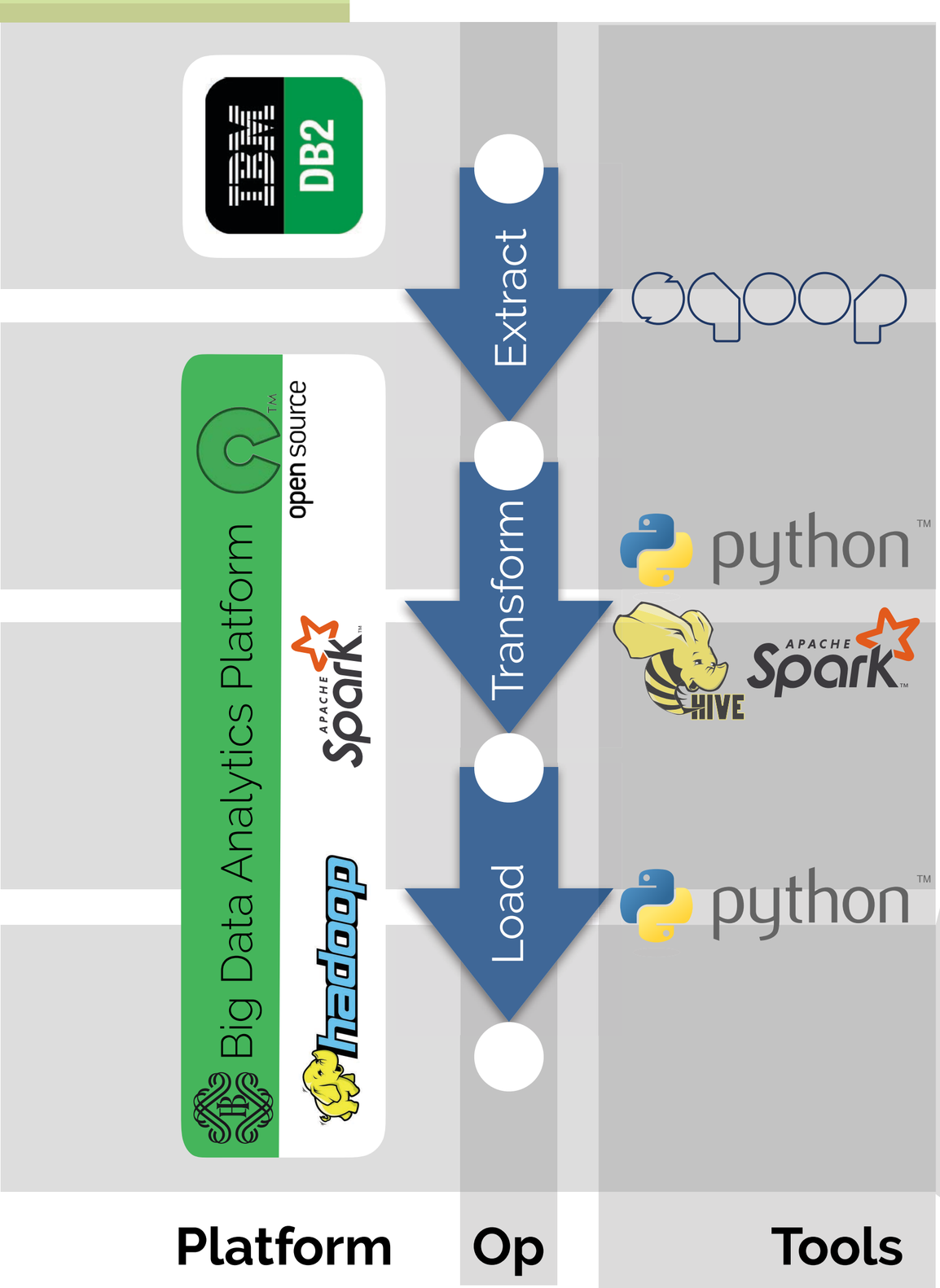
1.4M records, **400**MBytes

Each record contains info about:

a. Company structure
b. Balance sheet
c. Other info
d. **SAE**

Data

Data ingestion (ETL)



Inside the ML machinery, for each company

Company structure

15 numeric features
- Num. of employees
- PA-owned shares
- ……

Balance sheet

14 numeric features
- Share capital
- Personnel costs
- ……

Other info

3 structured features
- Listed (y/n)
- ATECO
- Comune

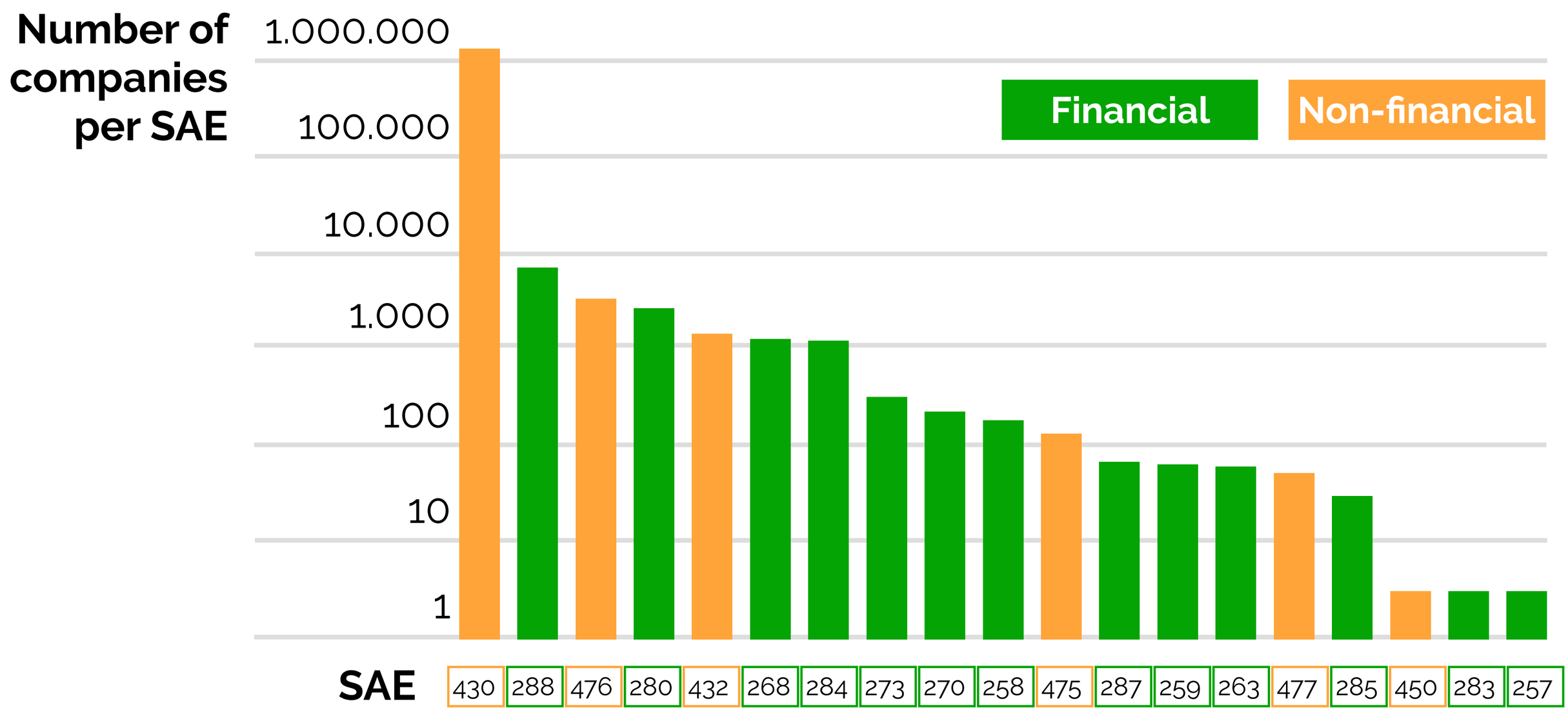
Name & notes

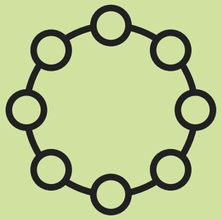
2 textual features
- company name
- **balance notes**

NEW

SAE

SAE: (Un)balanced data





Data

Preprocessing

Feature extraction

Imbalanced learning

Classification

Results

Specific w.r.t data type

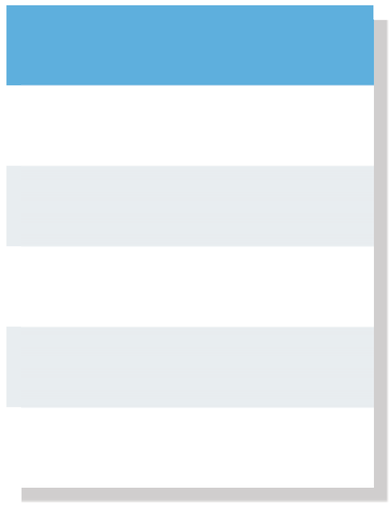
Dealing with Missing

Structured Data	Normalization	Try to fix or infer: • Use “zero” or average value • Regression on other variables • etc.
(un)structured Data	Lemmatization Stemming Dictionary-checking	Ignore

It is textual and can be divided:

- company denomination (always present)
- **balance notes** (missing in almost 50% of the dataset)

Types of features



Types of features

numeric quantity

categorial property

textual property

[direct]

[one-hot-encoding]

[tf-idf]

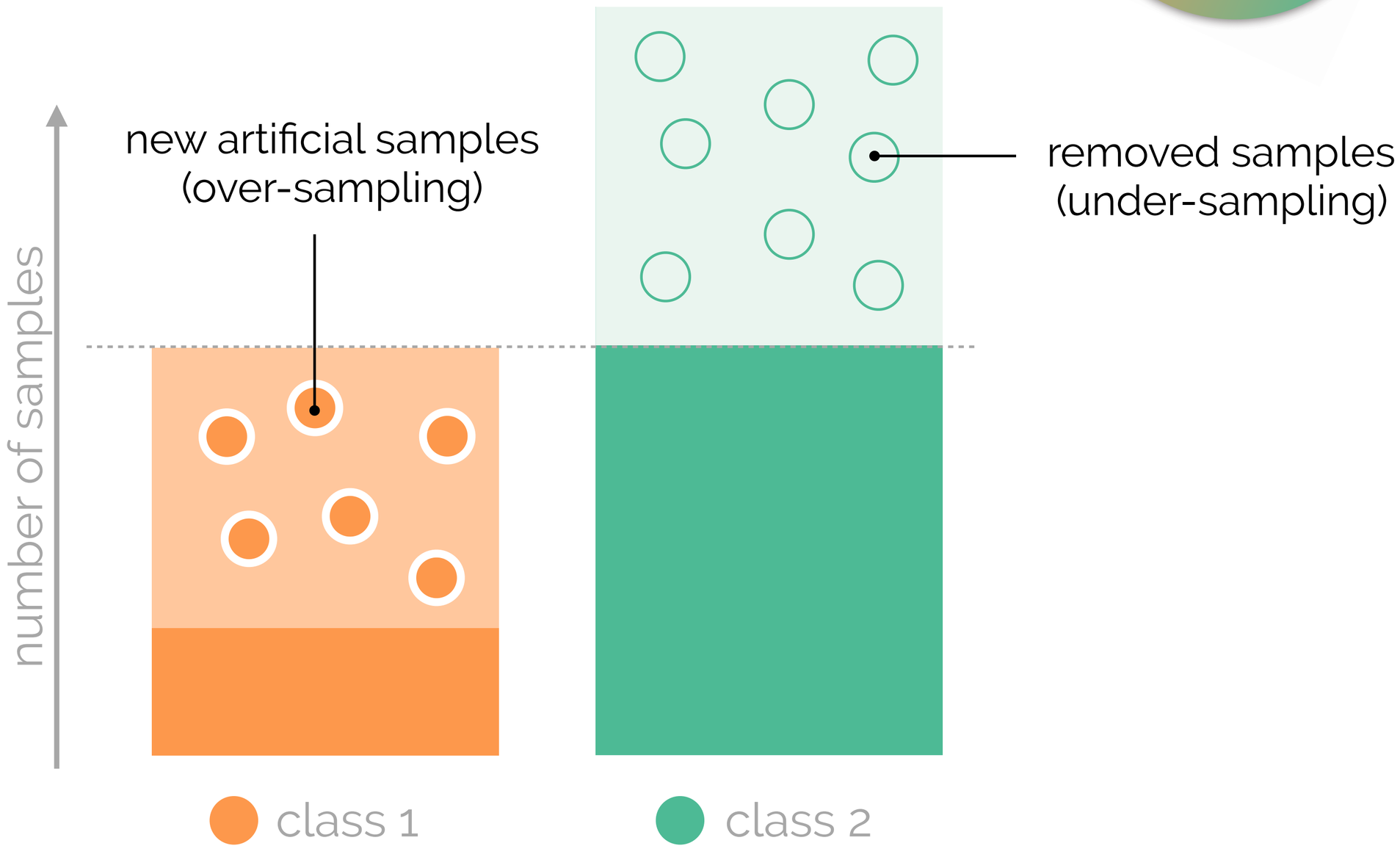
PCA

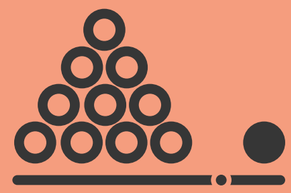
List of company **features**

A couple of unbalanced classes

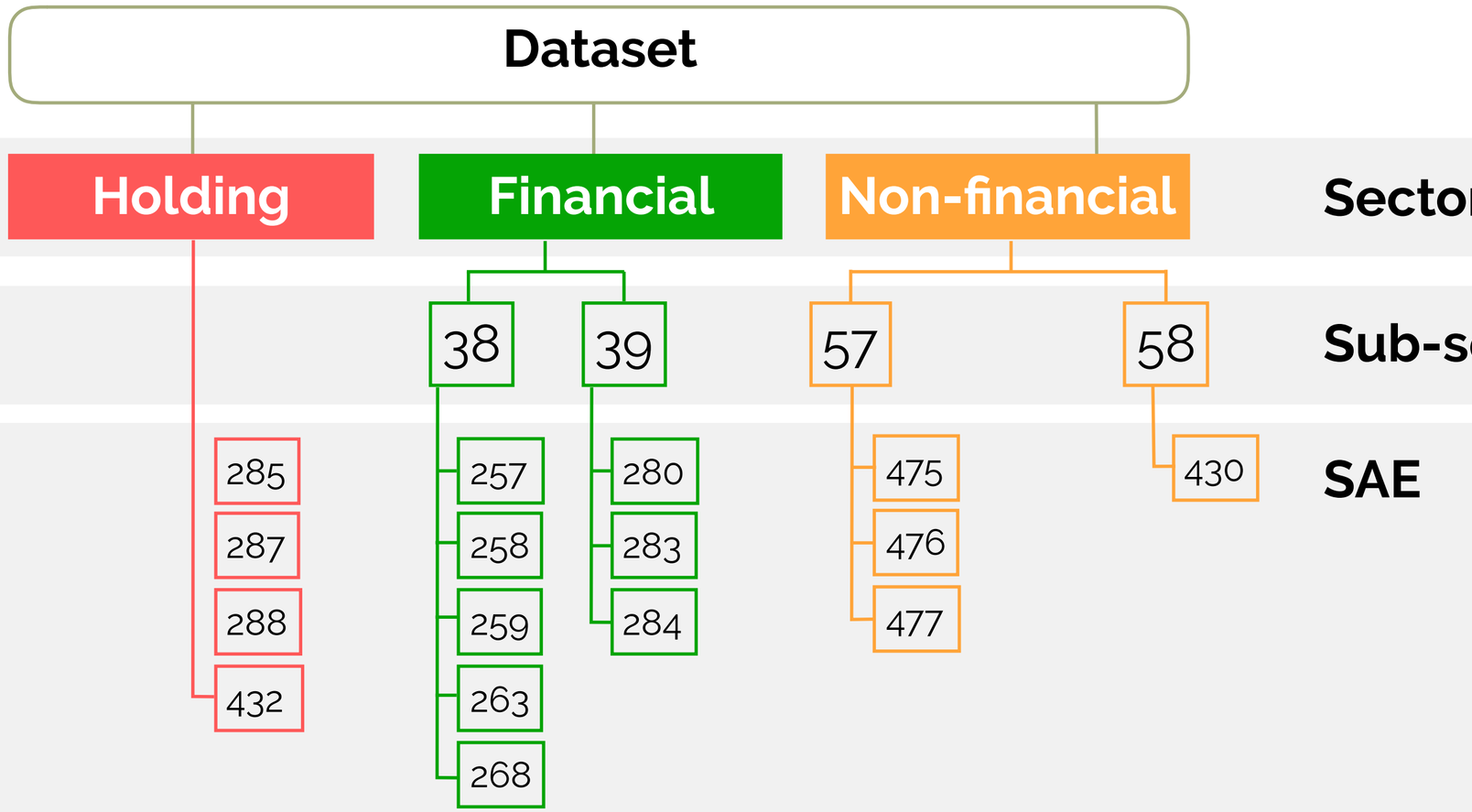
Under-sampling & over-sampling

SMOTE

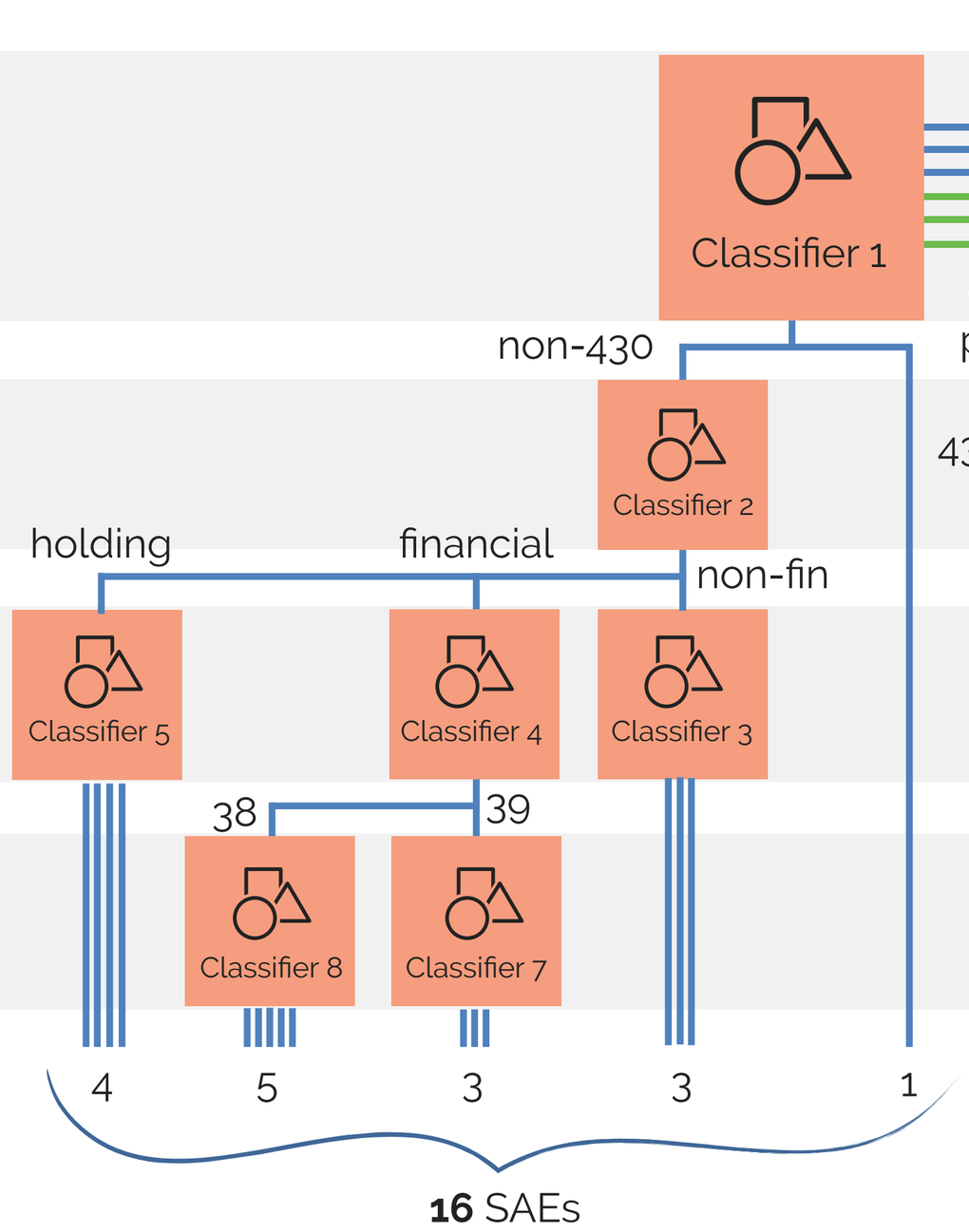




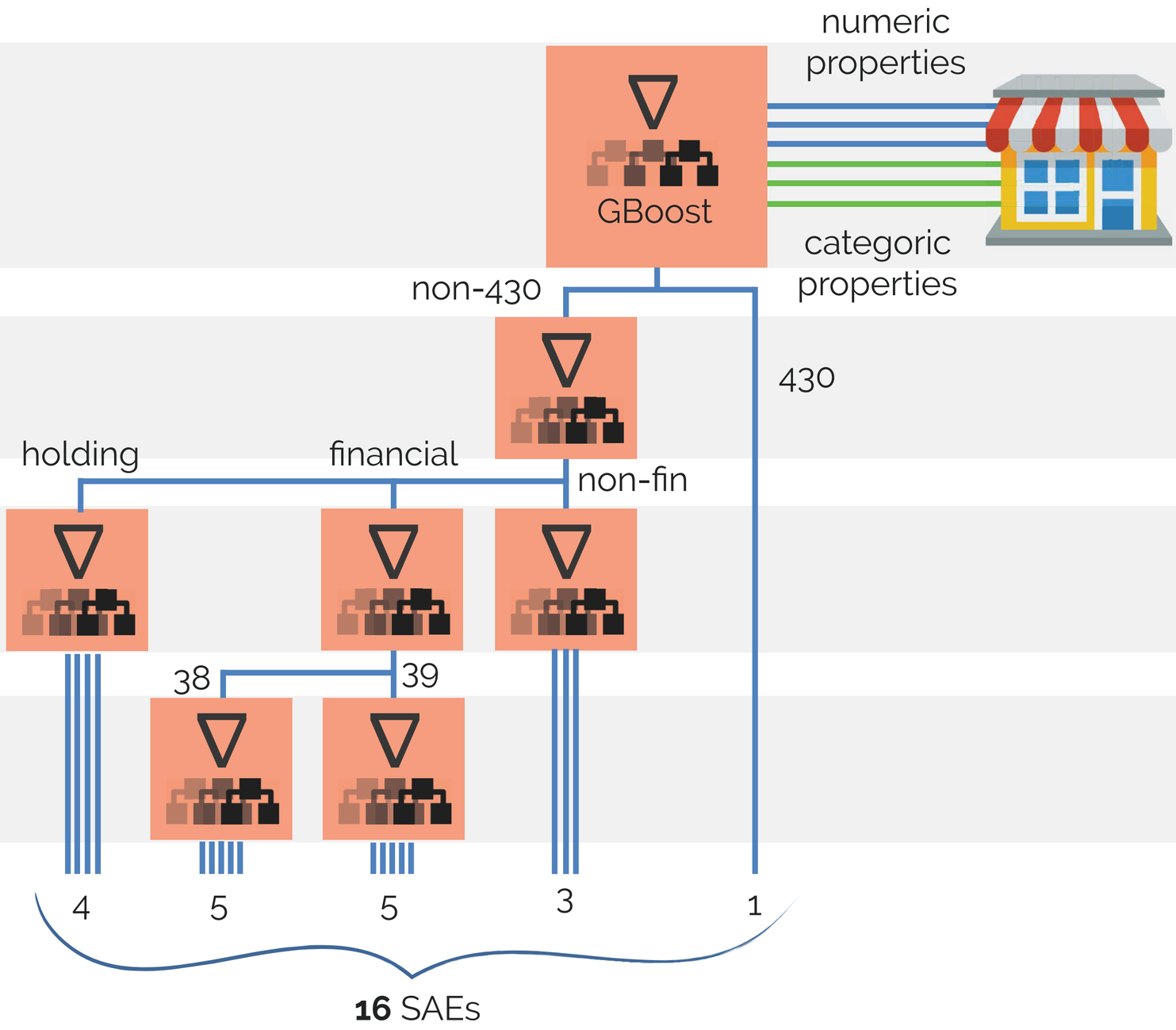
SAE hierarchy



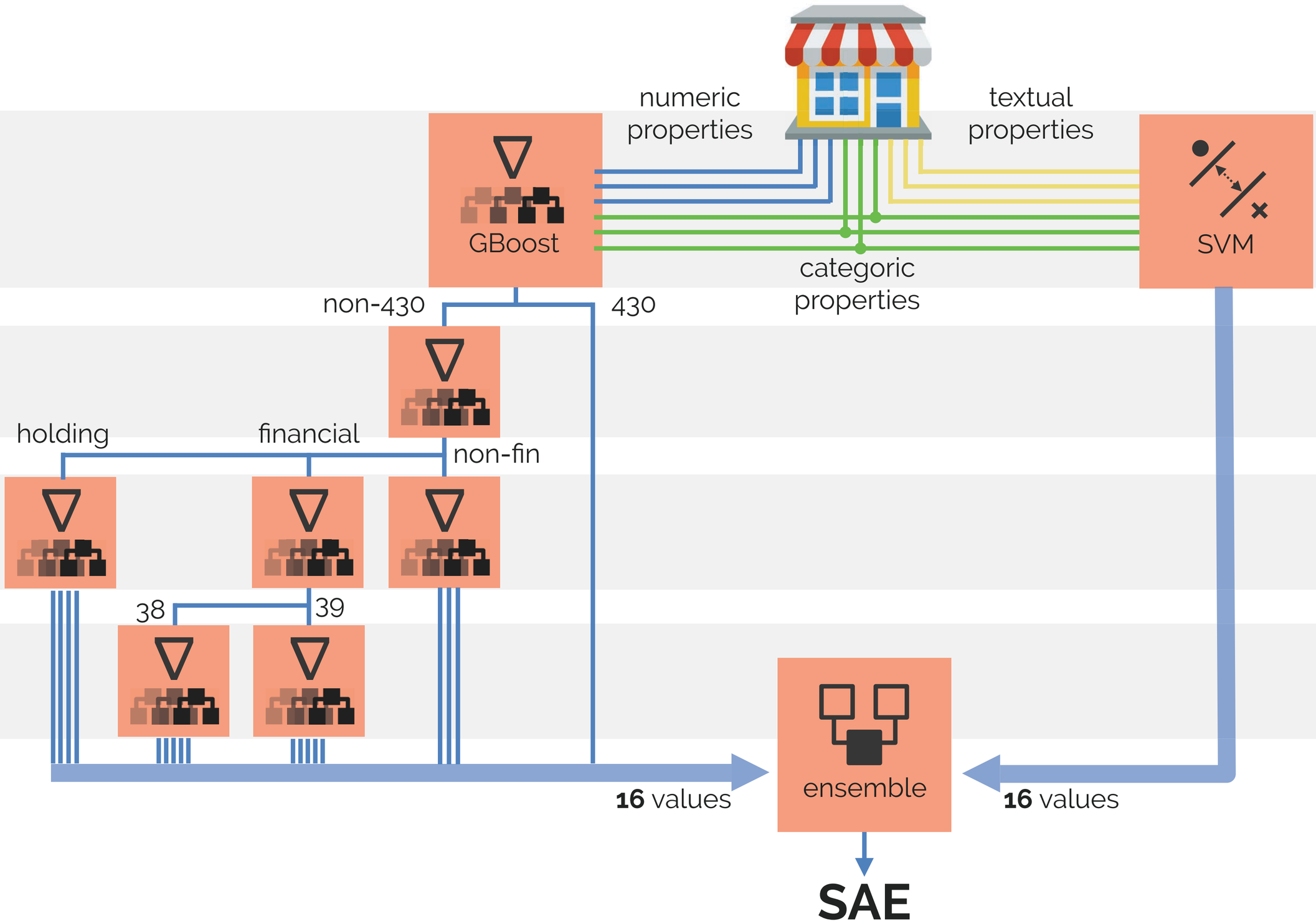
Classifier hierarchy



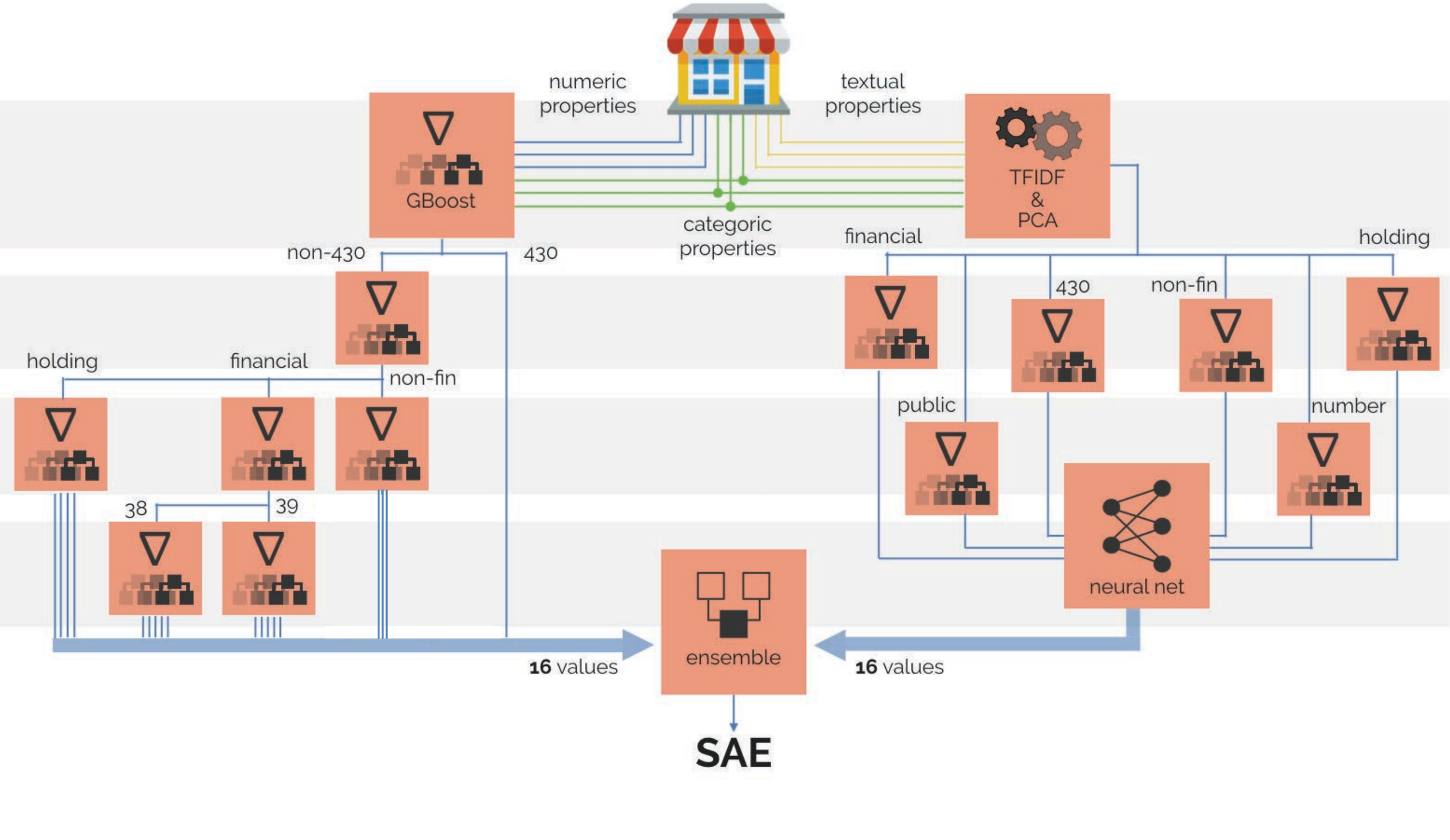
Classifier hierarchy



Ensemble classifier



Ensemble Neural classifier

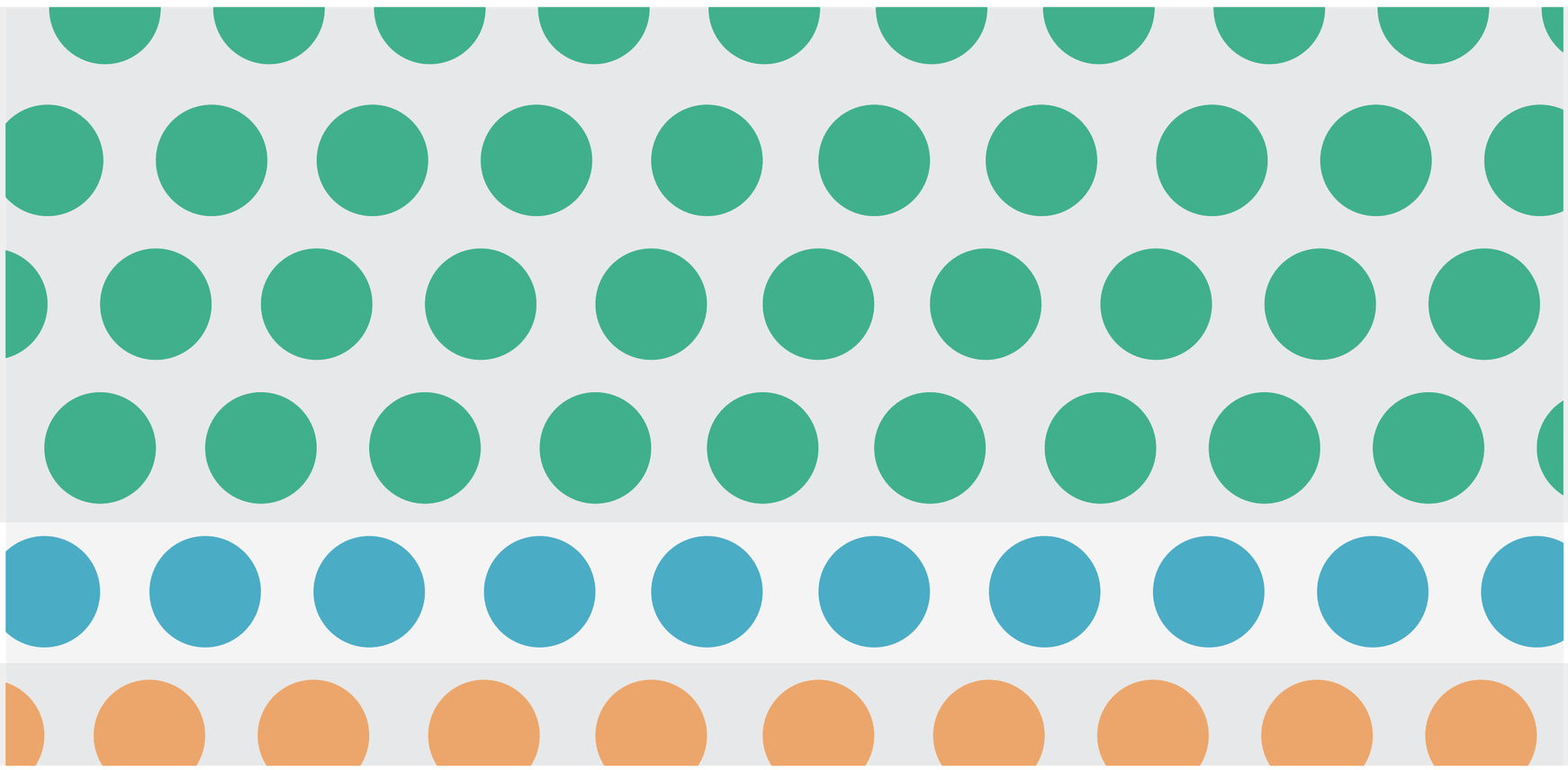


Datasets and performance

1.4 million records

Datasets and performance

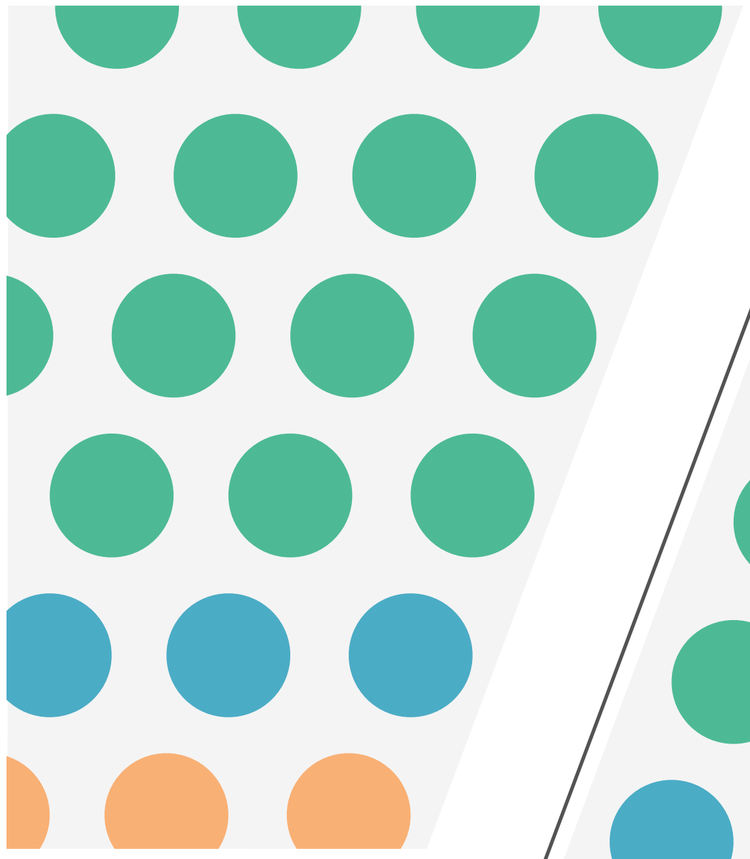
1.4 million records
of SAE-labeled data



Datasets and performance

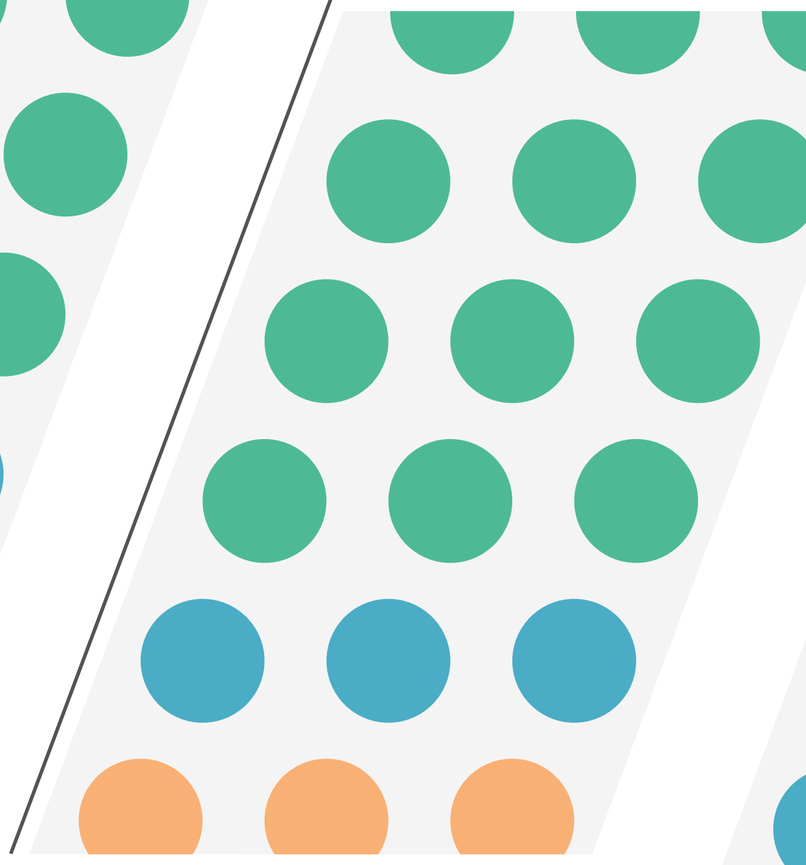
training set

used to automatically learn classifier **parameters**



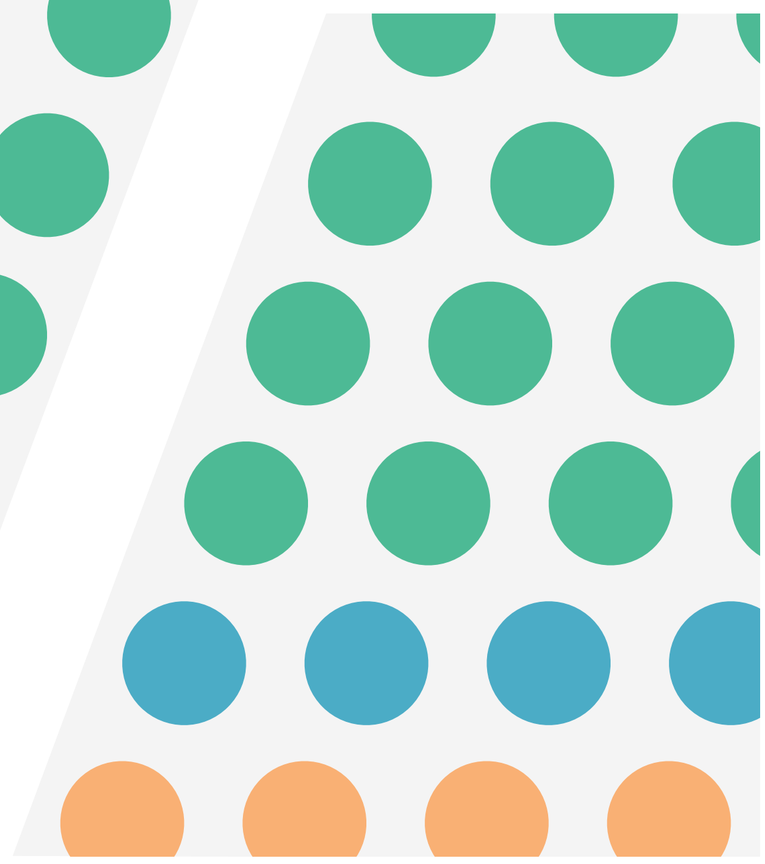
validation set

used to optimise classifier **hyperparameters**



test set

used to evaluate classifier **performance**



Datasets and performance

training set

used to automatically learn classifier **parameters**

validation set

used to optimise classifier **hyperparameters**

test set

used to evaluate classifier **performance**

420,000 records

the classifier has never seen before

Performance metrics

Raw, direct, relatable measure

1% of samples in the test set

4000 *max*

absolute number of errors

min 0

The standard performance measure

any classifiers that gets the large 430 “right” surpasses 99%

99% *min*

accuracy

max 100%

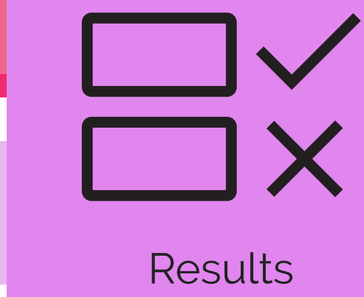
0% *min*

average F1 score

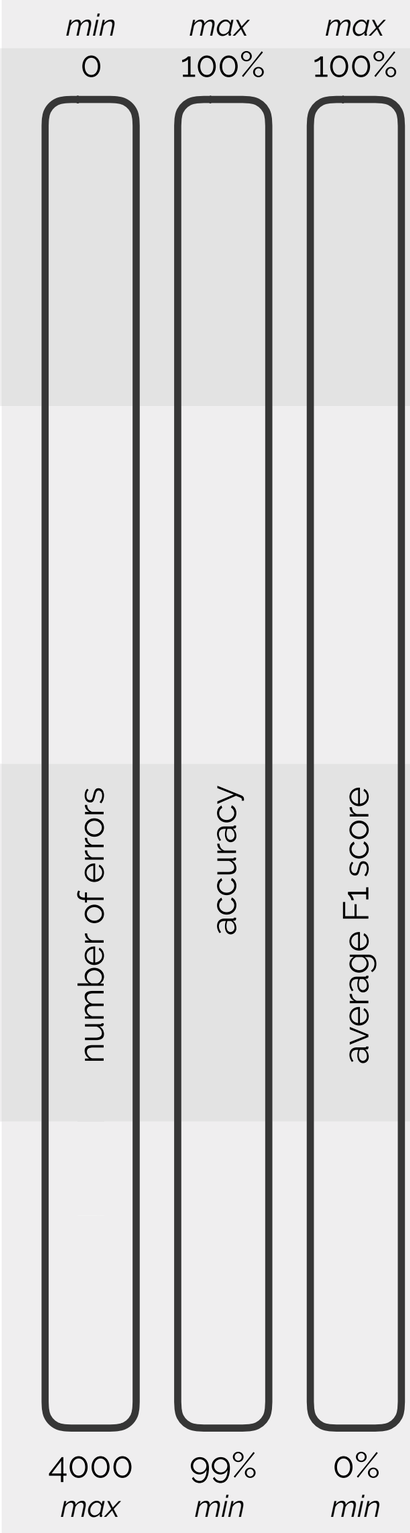
max 100%

Insensitive to class size, to score high here a classifier has to get most things right

Classification results



Results



Data | Preprocessing | Feature extraction | Imbalanced learning | Classification | Results

Classification results

data

preprocessed data

all features

balanced samples

neural ensemble

data

preprocessed data

num + categorical + name features

balanced samples

ensemble

data

preprocessed data

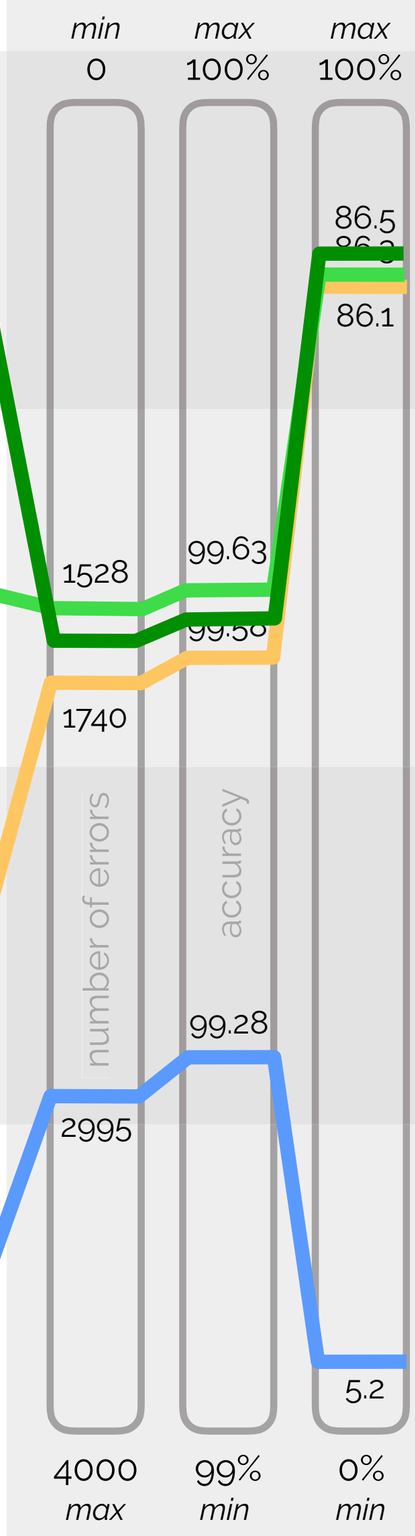
numerical and categorical

balanced samples

Hierarchical Classifier

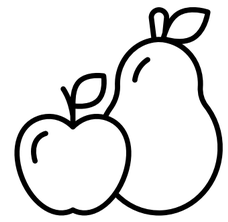
data

always 430



Conclusions

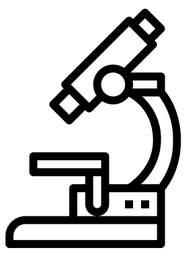
Dealing with **hybrid data** is complex and **different pipelines** (with ensemble techniques) are needed



Aa

Hierarchical structures give comfortable a-priori knowledge but are not well suited for “ambiguous” data

A scientific paper with details on all the techniques presented is currently under review and will be published soon.



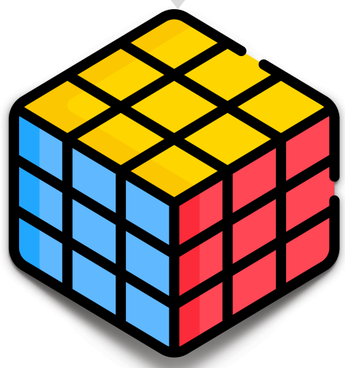
From  problem

A **business necessity** to improve DQM activity efficiency

A **Machine Learning** solution could solve the problem

A **research activity** was carried out in order to find the best solution

A **final solution** is being developed as an integration in the enterprise SW

To  solution

Thank you for your attention

Any questions?

BANCA D’ITALIA
EUROSISTEMA