INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Machine Learning 2 – 5A
Data preprocessing

Team: E. Alata, D. Dragomirescu, P. Leleux, V. Migliore

Université Fédérale Toulouse Midi-Pyrénées

Section 1

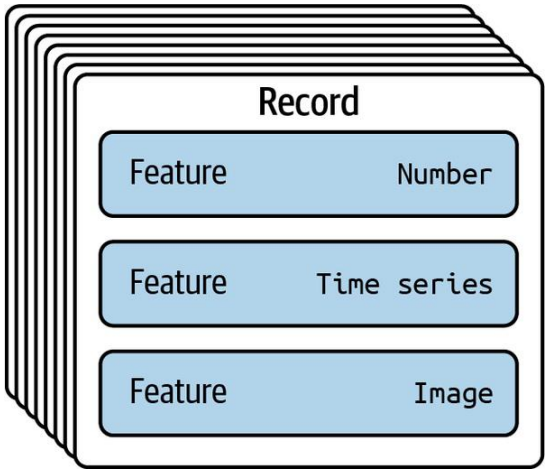
Data collection and representation

Data
Types

What types of data in this lecture ? Can be anything

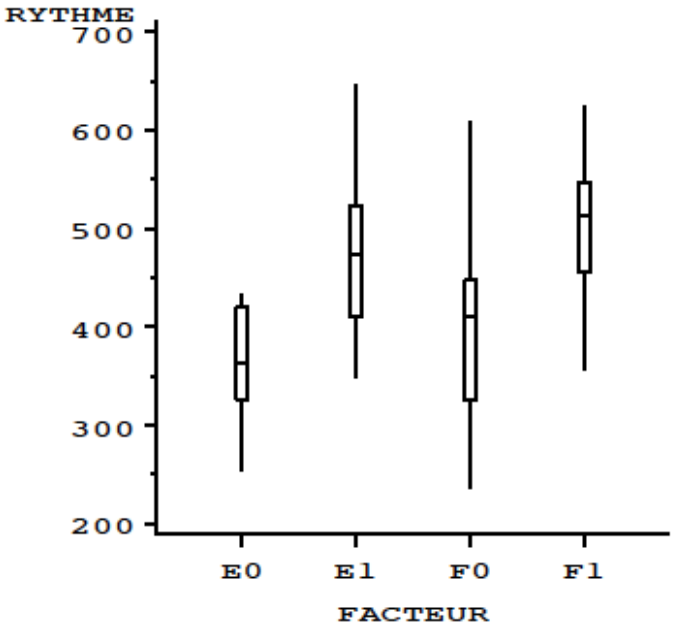
1. Quantitative and qualitative data : typically metadata of devices,
2. Time series **(lecture 3)** : data collected by sensors (e.g. accelerometers)
3. Audio (e.g. from smartphone mic)
4. Image **(lecture 4)** (e.g. medical imagery)
5. Others : Video, logs, network data, etc.

Representation ? arrays / matrices of integers / floating point numbers

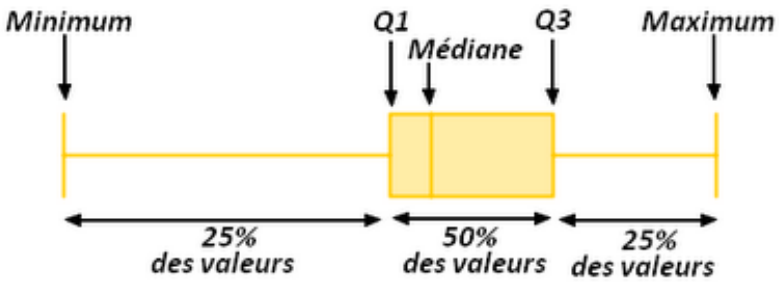


Data
Representation : quantitative features

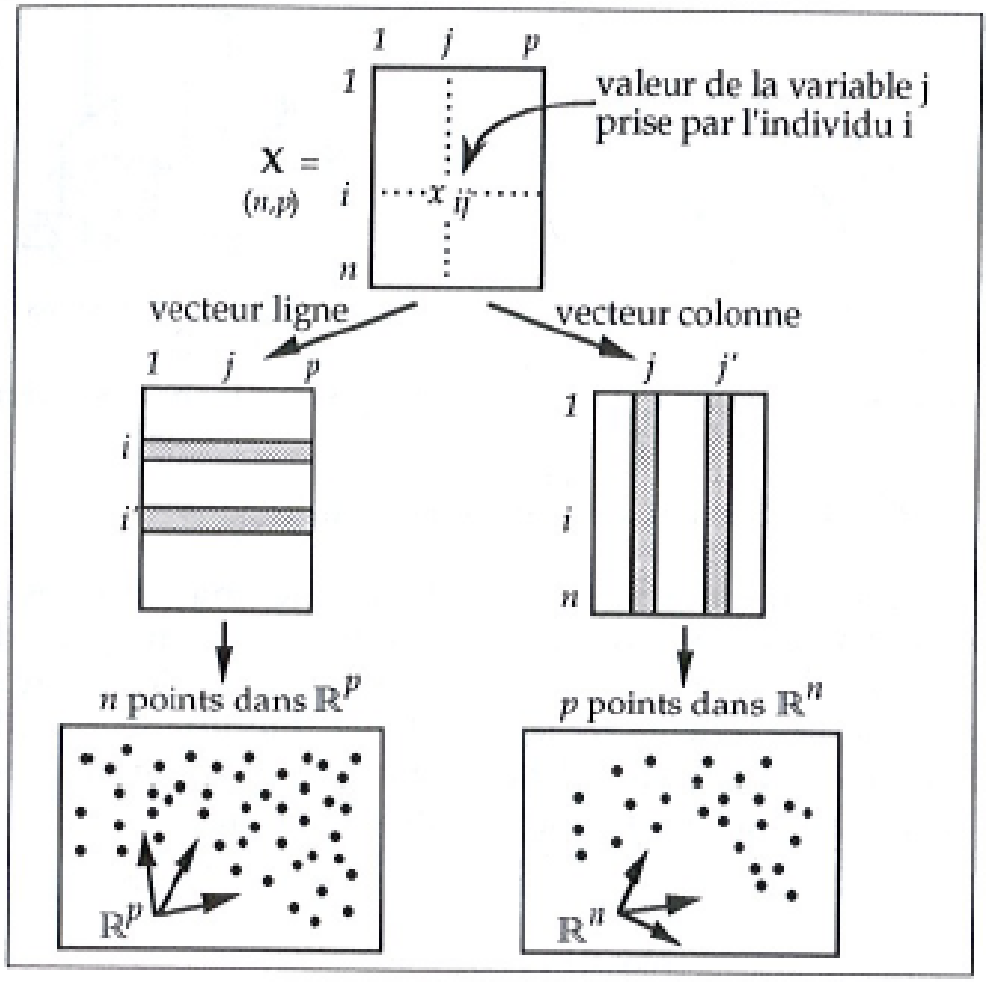
Parallel boxes



Box plot



Data
Representation : quantitative features

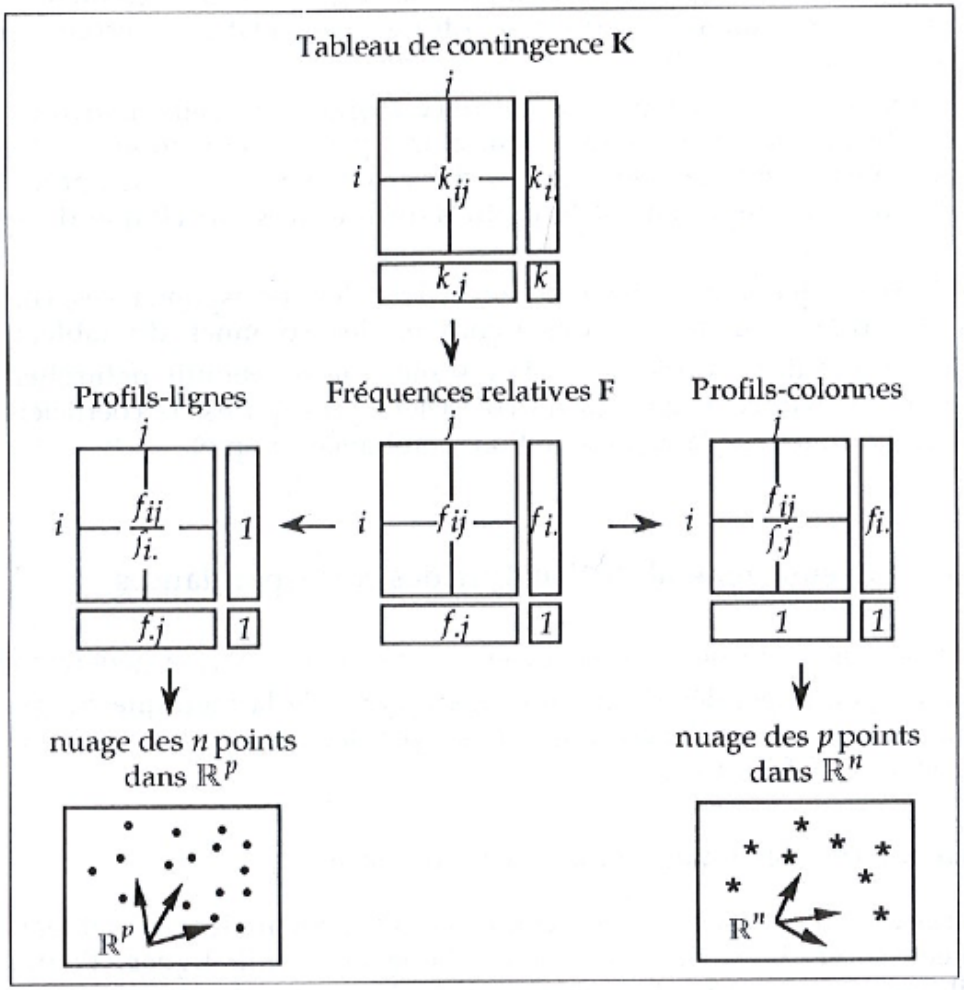


Data
Representation : qualitative features

yeux / cheveux	marron	noisette	vert	bleu	Profil moyen
brun	11	3	1	3	18
châtain	20	9	5	14	48
roux	4	2	2	3	12
blond	1	2	3	16	21
profil moyen	37	16	11	36	100

Table – Transformation of the contingency table

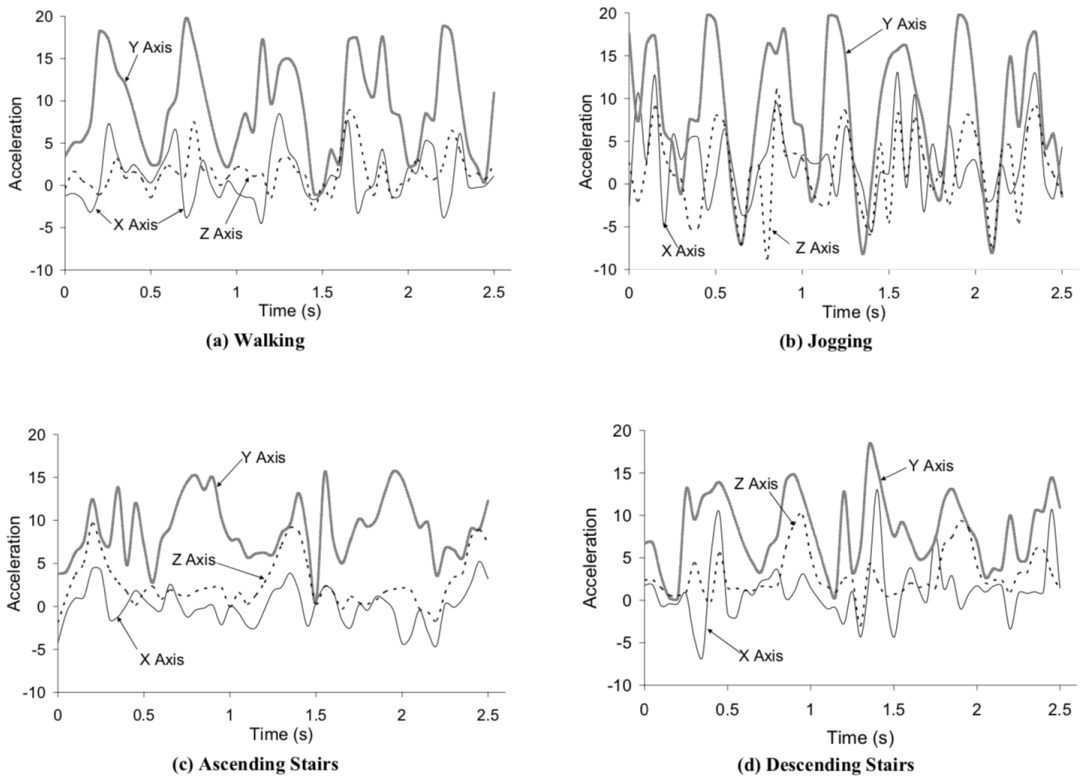
Data
Representation : qualitative features



Data
Representation : qualitative features (generalisation

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

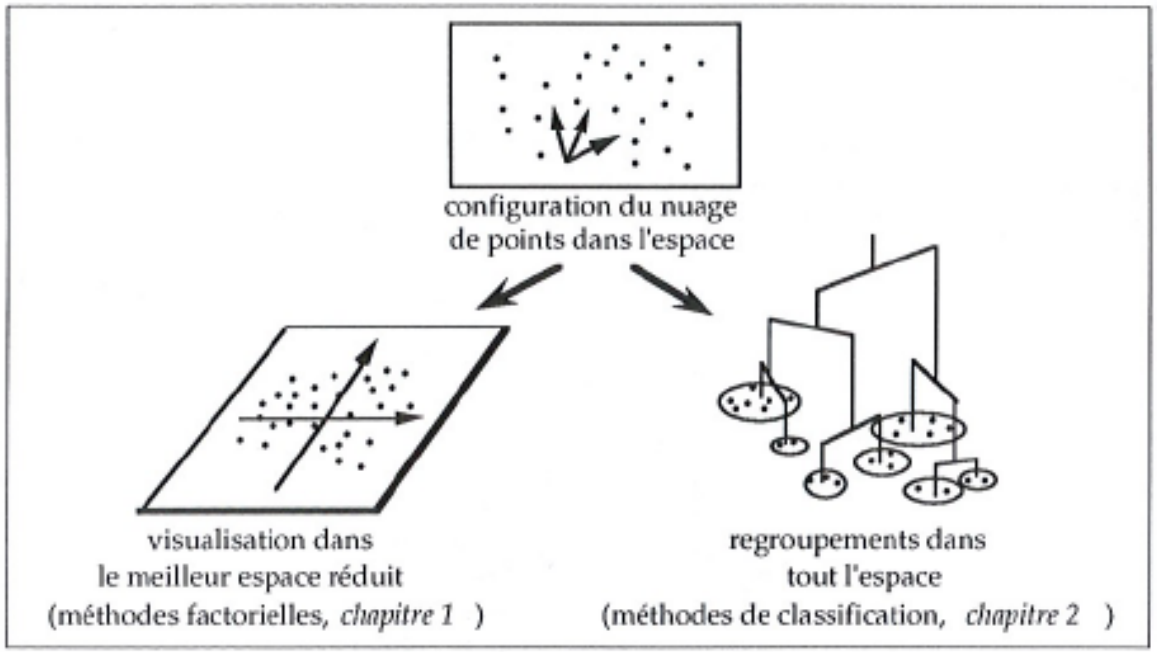
Data
Representation : temporal series



Principle of the graphical representation

Data
Towards factorial and classification methods

What happens when the feature space has a high dimension ?



Data Issues = Feature engineering

In your opinion, what are possible issues with data in embedded AI for IoT ?

- Data privacy ?
- Data transmission ?
- Data labeling ?
- Real-time data processing ?
- Heterogeneous data integration ?
- Scalability Issues ?
- Bias and ethical concerns ?
- Data quality and variety ?
- Limited storage and computation ?
- Real-time data capture ?
- Energy E□ciency Concerns ?

Data
Issues = Feature engineering

In your opinion, what are possible issues with data in embedded AI for IoT ?

- Data privacy ? access/use can be problematic
- Data transmission ?
- Data labeling ? can be inaccurate
- Real-time data processing ? expensive
- Heterogeneous data integration ? Techniques to ensure capability
- Scalability Issues ? Growing number of IoT devices
- Bias and ethical concerns ? data can be biased and lead to unfair/unethical model
- Data quality and variety ? real-world = noisy, incomplete, inconsistent BUT potentially better model
- Limited storage and computation ? "curse of dimensionality" + model complexity $\#param^p \times (\#samples \times \#features)^q$
- Real-time data capture ? data current/actual + direct (removes human errors and biases)
- Energy E□ciency Concerns ? Data processing can be energy-intensive

Issues :	low quality,	large amounts	+	real-time issues	+	ethics
Solutions :	data cleaning,	dimension reduction	+	out-of-scope	+	out-of-scope

Section 2

Data cleaning

Data cleaning
Transformation of data to quantitative features

Most machine learning techniques are designed to handle quantitative data ! All types of data can be transformed into it :

1. Qualitative :
   - To integers : $Eyes = \{blue, brown, green\} = \quad Eyes = \{0, 1, 2\}$
   - To binary : $Eyes = \{blue, brown, green\} = \quad Blue\_eyes = \{0, 1\}, Brown\_eyes = \{0, 1\}, Green\_eyes = \{0, 1\}$
2. Temporal series : array of floats
3. Image : matrix of integers/floats
4. Video : tensor of integers/floats
5. Text ? Tokenizing

Else, it is possible to build a machine learning model for each type of data then assemble the results : Ensemble Learning (Lecture 6)