Quelques repères (suite)

Système de Gestion de Base de Données (SGBD) - Quelques repères

Une base de données (relationnelle) c'est essentiellement une collection structurée d'informations

non nécessairement du même type (format)

Depuis l'avènement du big data (fouille des données, data mining), il existe des bases de données non structurées

(NoSQL) mais il faut passer par une structuration des données pour pouvoir les analyser et les exploiter…

⇨ **Il va donc falloir apprendre à stoker et organiser l'information en vue de son exploitation**

il existe des méthodes pour cela; cela ne s'improvise pas.

⇨ **Pas de place pour le bazar…**

⇨ **mais la logique et le bon sens ont tous leurs droits!**

Une base de données ne reflète pas la réalité, c'est une vision des choses.

De plus, il existe toujours une grande probabilité de présence d'erreurs de frappe

ou liées à une information biaisée, mal reportée, mal acquise ou d'erreur(s) intentionnelle(s) ou de négligence.

Système de Gestion de Base de Données (SGBD) - Quelques repères

- Une base de données c'est essentiellement une collection structurée d'informations

  non nécessairement du même type (format)

- Une base de données est usuellement localisée en un seul lieu et un seul support (dupliqué en fait)

  qui est généralement informatique (numérique)

- Pièce centrale des dispositifs informatiques qui servent à la collecte, le stockage et l'utilisation des informations

- SGBDR, acronyme de Système de Gestion de Base de Données Relationnelles :

  logiciel moteur qui pilote la base et en permet la manipulation et l'exploitation (interrogation).

- Le langage le plus souvent utiliser pour piloter, interroger et interfacer une base de données est le SQL

  System Querry Language

- Toutes les secondes le volume d'information ne cesse d'augmenter contribuant au Big Data. Mais sans analyse

  et sans base de données (contribuant à préparer les données pour leur analyse), le Big Data n'est rien.

Du jeu de données (data set) à la base de données

Premier contact avec la notion de base de données

Exemple : Gestion d'arbres disposés en parcelles

Fichier texte ASCII (arbres_sample1.txt)
support portable universel

Chaque **enregistrement** (lignes) contient un nombre défini de **champs** (colonnes)
Le caractère « **;** » est ici le **séparateur** de champs

Fichier Edition Format Affichage Aide

```
ID;arbre;parcelle;hauteur;circonf;etat;valeur
1;sapin;P133;10,9;1,23;bon;3
2;bouleau;P095;8,21;1,15;moyen;1
3;pin;P022;9,42;1,5;bon;2
4;bouleau;P095;9,51;1,03;bon;3
5;bouleau;P022;7,22;0,95;mauvais;0
6;chêne;P287;9,2;1,25;bon;3
7;pin;P287;9,51;2;moyen;2
8;pin;P204;10,83;NA;bon;4
9;bouleau;P133;11,74;NA;mauvais;0
10;bouleau;P095;11,39;3;moyen;3
11;peuplier;P022;10,44;NA;bon;3
12;chêne;P287;9;NA;NA;2
13;chêne;P204;9,52;NA;moyen;2
14;peuplier;P204;10,84;NA;mauvais;1
15;sapin;P133;10,9;NA;moyen;0
16;chêne;P095;10,01;NA;moyen;3
17;chêne;P022;11,35;NA;bon;
18;sapin;P204;11,99;NA;moyen;2
```

Cette (petite) base de données est constituée d'éléments ordonnés de façon bien précise ; l'ordre est donné par la première ligne (=entête). L'entête contient l'identificateur des champs [field] de chaque enregistrement [record].

;;=pas d'info pour ce champ > Ceci ne constitue pas une erreur, on ne dispose pas de l'information relative à la valeur de ce champ pour cet enregistrement

ID; arbre ; num_parcelle ; hauteur ; circonférence ; état

000001 ; sapin ; #133 ; 9,35 ; 1,23 ; bon

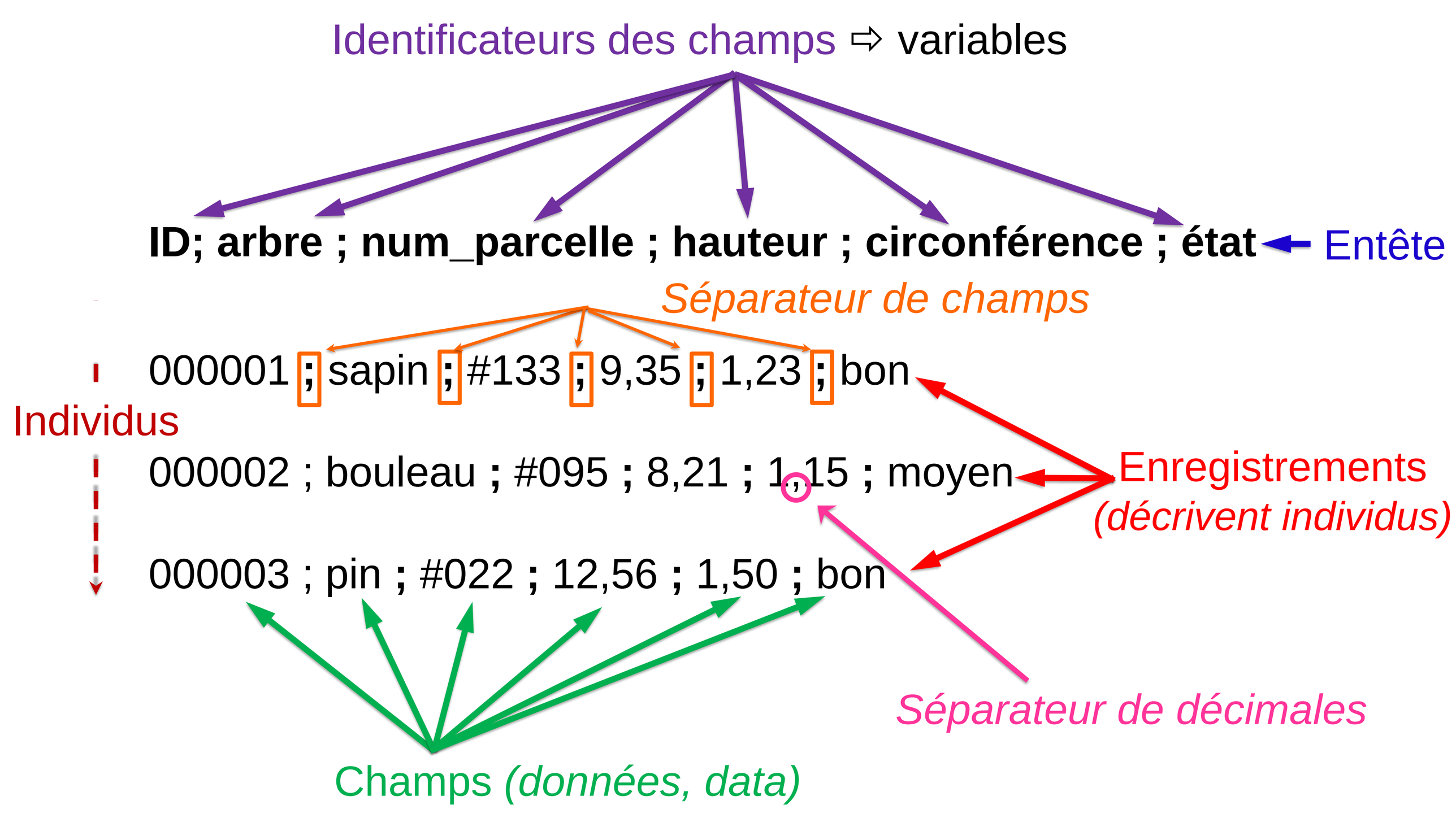
000002 ; bouleau ; #095 ; 8,21 ; 1,15 ; moyen

000003 ; pin ; #022 ; 12,56 ; 1,50 ; bon

000004 ; bouleau ; #095 ; 9,51 ; 1,03 ; bon

000005 ; peuplier ; #022 ; 7,22 ; 0,95 ; mauvais

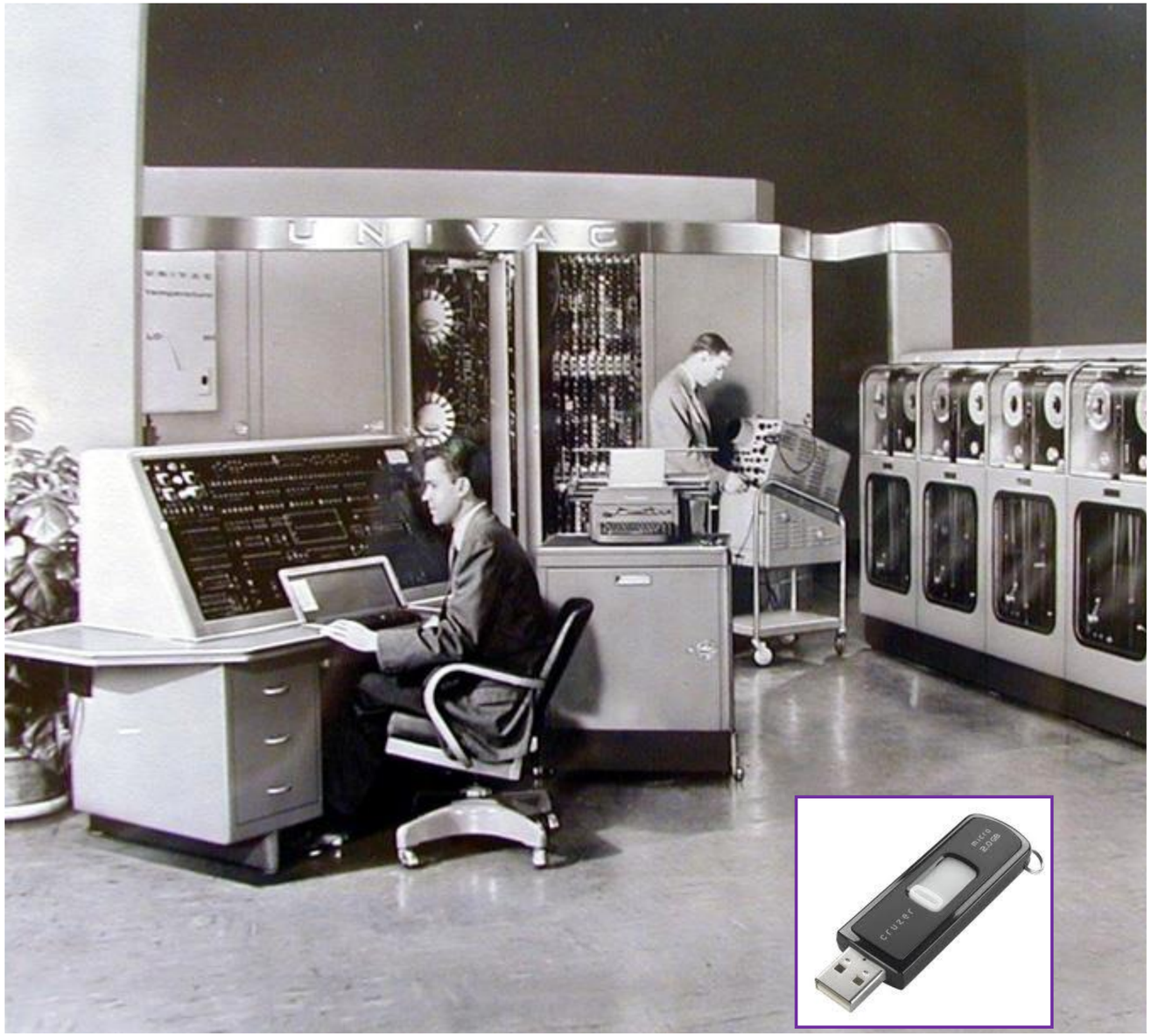
000006 ; chêne ; #133 ; 10,08 ; 1,25 ; bon

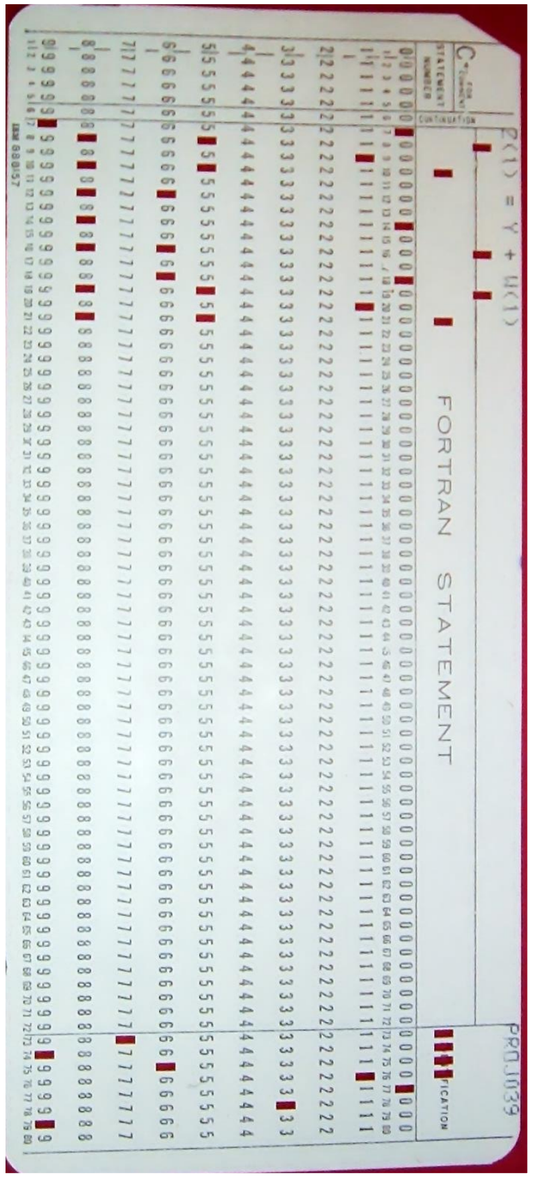

Identificateurs des champs ⇨ variables
ID; arbre ; num_parcelle ; hauteur ; circonférence ; état
Entête
Séparateur de champs
000001 ; sapin ; #133 ; 9,35 ; 1,23 ; bon
Individus
000002 ; bouleau ; #095 ; 8,21 ; 1,15 ; moyen
Enregistrements
(décrivent individus)
000003 ; pin ; #022 ; 12,56 ; 1,50 ; bon
Séparateur de décimales
Champs (données, data)

Origine du mot Enregistrement (Record)

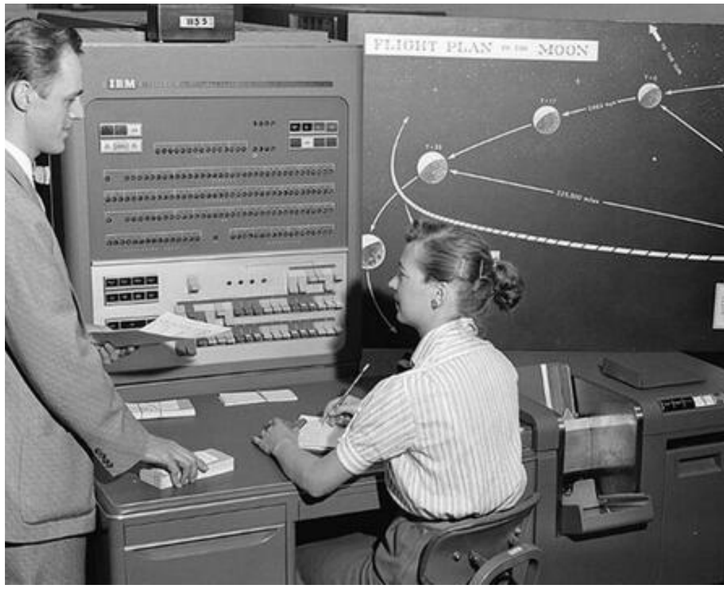
Les premiers gros ordinateurs et leur utilisation civile

L’Univac (1951)





Carte perforée (codant ici pour une instruction en Fortran)



Fichiers – Enregistrements – Variables - Individus

Enregistrement : compilation d'informations relatives à l'individu considéré

- **un enregistrement par individu** (unité statistique) ;
- découpé en **champs** de différentes natures/types ⇨ **Variables**

Pas de limite quant au nombre de **variables** engagées, ni à la diversité de leurs types

Distinguer la (les) variable(s) réponse(s) (*variable dépendante, VD, target, cible, à expliquer,...)*

des variables indépendantes *(explicative, régresseur,…)*

Fichiers de données (jeux de données, dataset)

► **Lignes** : individus - **Colonnes** : variables

de type « feuille de calcul », destinée à devenir un « data frame » dans le traitement des données

Le fruit de notre première réflexion

ID	arbre	parcelle	hauteur	circonf	etat	valeur
1	sapin	P133	10,9	1,23	bon	3
2	bouleau	P095	8,21	1,15	moyen	1
3	pin	P022	9,42	1,5	bon	2
4	bouleau	P095	9,51	1,03	bon	3
5	bouleau	P022	7,22	0,95	mauvais	0
6	chêne	P287	9,2	1,25	bon	3
7	pin	P287	9,51	2	moyen	2
8	pin	P204	10,83	NA	bon	4
9	bouleau	P133	11,74	NA	mauvais	0
10	bouleau	P095	11,39	3	moyen	3
11	peuplier	P022	10,44	NA	bon	3
12	chêne	P287	9	NA	NA	2
13	chêne	P204	9,52	NA	moyen	2
14	peuplier	P204	10,84	NA	mauvais	1
15	sapin	P133	10,9	NA	moyen	0
16	chêne	P095	10,01	NA	moyen	3
17	chêne	P022	11,35	NA	bon	
18	sapin	P204	11,99	NA	moyen	2
19	hêtre	P133	11,12	2	bon	3
20	hêtre	P287	10,16	NA	bon	4
21	hêtre	P022	11,13	NA	mauvais	1
22	peuplier	P095	12	NA	bon	3
23	pin	P204	9,9	NA	moyen	2
24	peuplier	P133	9,61	NA	mauvais	0
25	peuplier	P287	10,2	NA	bon	2
26	hêtre	NA	10,15	NA	moyen	2
27	pin	P022	9,9	1	mauvais	1
28	hêtre	P287	9,99	2,2	mauvais	1
29	bouleau	P287	10,05	3	NA	2
30	peuplier	P204	10,8	NA	bon	4
31	chêne	P133	10,9	NA	bon	3
32	bouleau	P095	9,85	1,5	moyen	2
			10,6	NA	mauvais	0
			9,04	NA	mauvais	1

Avec EXCEL

Les deux aspects d'un même fichier (CSV, séparateur = « ; »)

Ce petit fichier pourrait faire couler beaucoup d'encre !

```
ID;arbre;parcelle;hauteur;circonf;etat;valeur
1;sapin;P133;10,9;1,23;bon;3
2;bouleau;P095;8,21;1,15;moyen;1
3;pin;P022;9,42;1,5;bon;2
4;bouleau;P095;9,51;1,03;bon;3
5;bouleau;P022;7,22;0,95;mauvais;0
6;chêne;P287;9,2;1,25;bon;3
7;pin;P287;9,51;2;moyen;2
8;pin;P204;10,83;NA;bon;4
9;bouleau;P133;11,74;NA;mauvais;0
10;bouleau;P095;11,39;3;moyen;3
11;peuplier;P022;10,44;NA;bon;3
12;chêne;P287;9;NA;NA;2
13;chêne;P204;9,52;NA;moyen;2
14;peuplier;P204;10,84;NA;mauvais;1
15;sapin;P133;10,9;NA;moyen;0
16;chêne;P095;10,01;NA;moyen;3
17;chêne;P022;11,35;NA;bon;
18;sapin;P204;11,99;NA;moyen;2
19;hêtre;P133;11,12;2;bon;3
20;hêtre;P287;10,16;NA;bon;4
21;hêtre;P022;11,13;NA;mauvais;1
                                095;12;NA;bon;3
```

Avec Notpad

- Fichier texte (ASCII)
- Ligne/Colonnes
- Individus/Variables
- Variables à expliquer
- Variables explicatives
- Précision des données
- Type des informations
- Chaîne de caractères
- Numérique réel
- Echelle de Likert
- Variable ordinale
- Relation d'ordre
- Source des informations
- Données manquantes
- Autres champs à inclure
- Et bien d'autres choses...