GPS, capteurs, suivi en temps réel : analyser des données fonctionnelles dans le sport

Groupe Statistique et Sport de la SFdS

Marie Chion, Sébastien Déjean, Arthur Leroy, Christian Derquenne

SFdS

Plan

- Présentation du groupe Statistique & Sport de la SFdS et objectifs de l'atelier
- Introduction & contexte
- Se familiariser avec les aspects techniques des données
- Mise en pratique sur R
- Prétraitement des données
- Bonnes pratiques
- Bibliographie

Présentation du groupe Statistique & Sport de la SFdS

Le groupe Statistique et Sport

- Créé en 2018 par des professionnels du sport, des statisticiens, des chercheurs, des enseignants appartenant à l'INSEP, à l'INJEP, la FFR, l'ENSAI, L'Oréal et EDF

Objectif : disposer d'un **langage commun** afin de **construire un dialogue** entre statisticiens et professionnels du sport avec un **choix de thèmes communs**

Voies potentielles :

- fournir des outils pour la prise de décisions dans les clubs et les fédérations
- organiser des séminaires de recherche et applications
- développer des formations pour augmenter la compétence statistique dans le sport

Activités du groupe Statistique et Sport

- ✓ Un à deux séminaires par an à l'IHP en général, mais aussi à l'INSEP et à l'ENSAI
- ✓ Session spéciale + sessions libres aux Journées de Statistique (JdS)
- ✓ Au total, une centaine d'exposés
- ✓ Deux Cafés de la Statistique (Paris et Lyon)
- ✓ Une journée satellite aux JdS 2024 à Bordeaux
- ✓ Un poster SFdS Statistique et Sport
- ✓ Atelier Statistique et Sport

Atelier Statistique – Statistique et Sport

Ce que nous ferons dans cette formation :

- ✓ Lecture de fichiers pour récupérer des données (R, txt, csv, Excel, html, pdf)
- ✓ Prétraitements simple et complexe des données sur différentes problématiques
- ✓ Techniques de visualisation et d'analyse des données classiques
- ✓ Analyse de données fonctionnelles simple et complexe
- ✓ Fournir les garde-fous nécessaires pour l'utilisation des méthodes et l'interprétation de leurs résultats

Ce que nous ne ferons pas dans cette formation :

- ✓ Captation/collecte de données, analyse de vos propres jeux de données,
- ✓ Entrer dans les détails math des méthodes …
- ✓ …

Problématiques, données et prétraitements

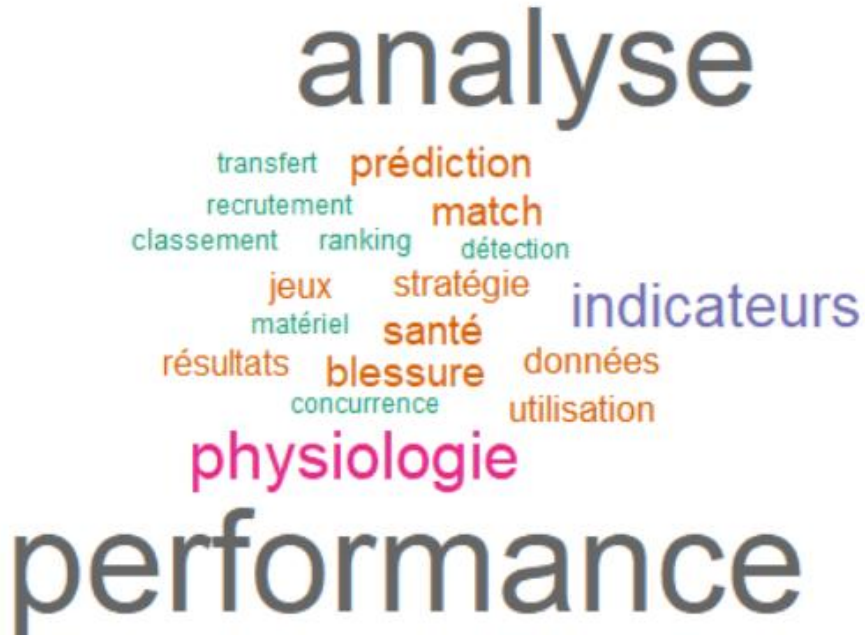
Groupe Statistique et Sport de la SFdS

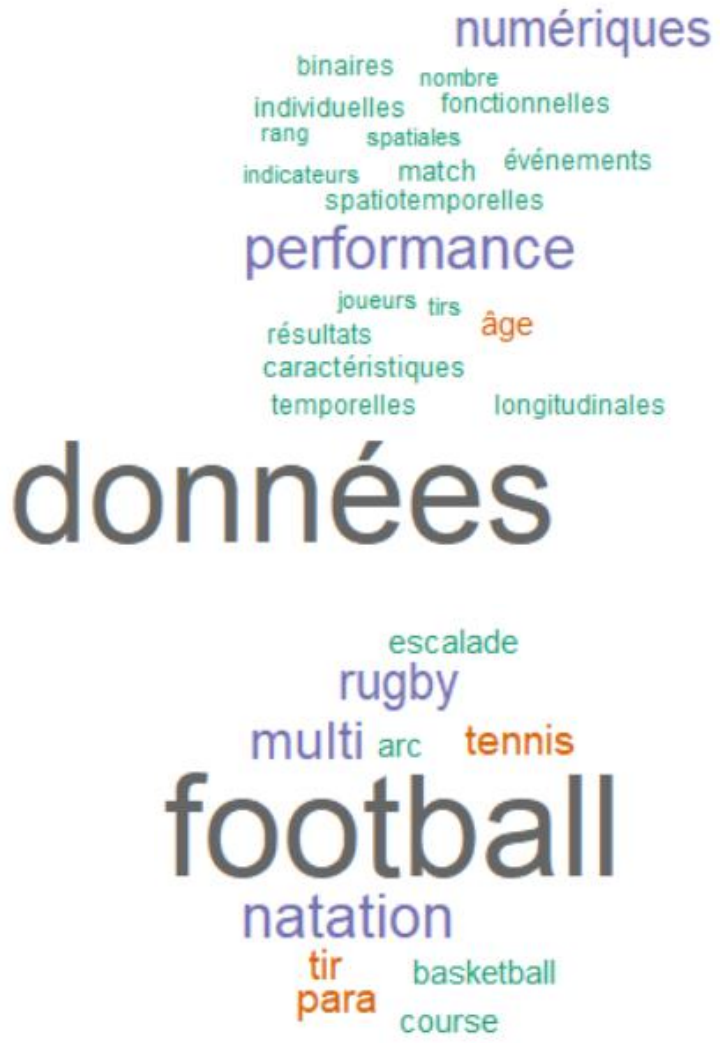
Marie Chion & Christian Derquenne

Introduction & Contexte

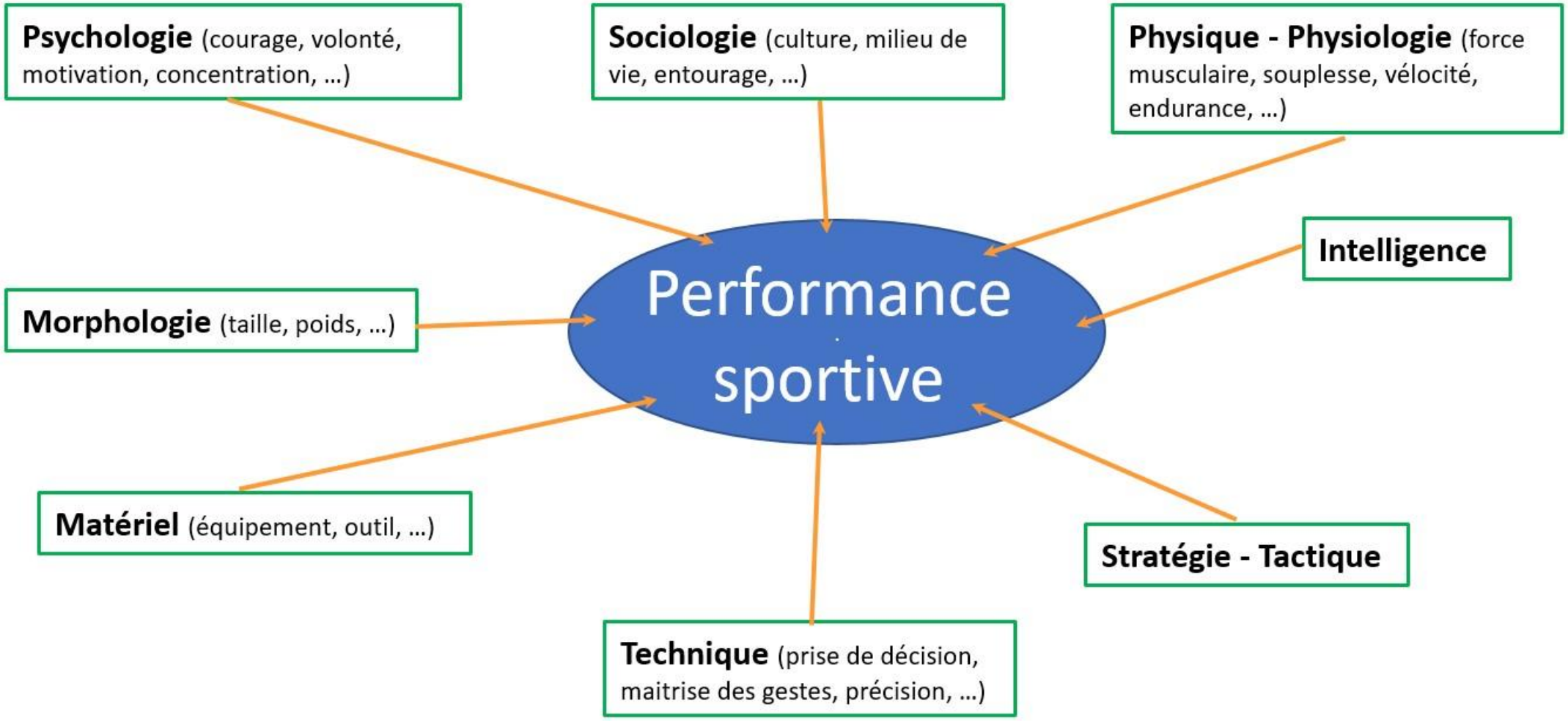
Les titres des exposés avec un nuage de mots

Problématiques, données, sports et méthodes

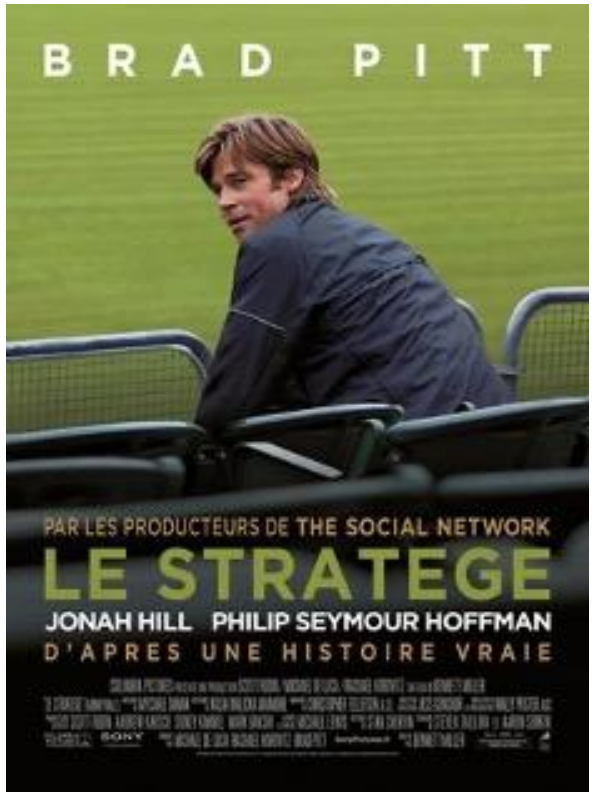




Un premier exemple : *la performance sportive*

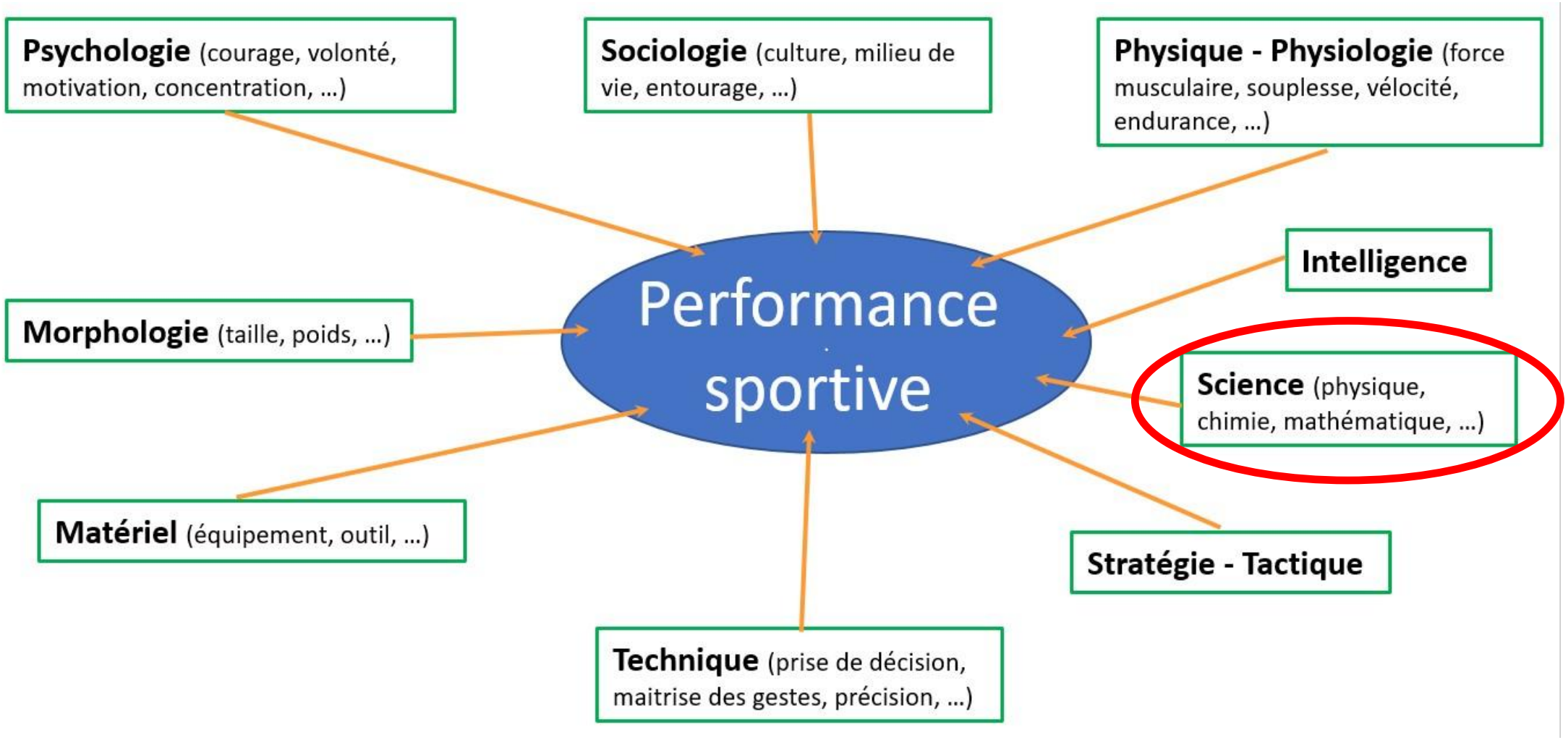


Quand les données changent la manière de gagner



- *Moneyball (Oakland Athletics, 2002)* : première équipe à recruter sur statistiques, pas sur instinct.
- Résultat : une révolution mais aussi un changement de culture : **mesurer ne suffit pas, il faut comprendre ce que l'on mesure**.
- Aujourd'hui, chaque fédération cherche ses propres "indicateurs cachés" de performance.

La performance sportive et la Science



Quelques problématiques sportives

- Analyse de la performance individuelle
- Optimisation de la performance individuelle
- Classement des sportifs, probabilité de victoire
- Sportifs durables – détection de talents – blessures
- Equipement sportif
- Analyse de match en sport collectif
- Détection de dopage
- Signature bio-mécanique d’un geste sportif

Comment se traduisent-elles en statistique ?

Analyse de la performance individuelle

- ✓ Construction d'indicateurs, classification des sportifs, identification des indicateurs pertinents pour la victoire ;
- ✓ Analyse exploratoire (classification d'individus et de variables, analyse factorielle), sélection de variables ;
- ✓ Approches de prévision de la victoire (rég. Logit, random forest, …)

Optimisation de la performance individuelle

- ✓ Adapter un entraînement pour atteindre une valeur d'indicateur lors d'un match ;
- ✓ Modèle de Banister (charge d'entrainement = quantité d'entrainements absorbée : https://www.sportifeo.com/blog/charge-entrainement/le-trimp-pour-suivre-la-charge-dentrainement/), modèle non-linéaire, inférence

Comment se traduisent-elles en statistique ?

Classement des sportifs, probabilité de victoire

- ✓ Modéliser la probabilité de victoire d'un point, puis remonter à la victoire du match, puis d'un tournoi ;
- ✓ Régression logistique, scoring, classement Elo, optimisation numérique ;
- ✓ Handicap de jeu en utilisant le classement Elo

Sportifs durables – détection de talents – blessures

- ✓ Prévention des blessures (fatigue, sport co), identification des facteurs de risque, aspects physiologiques et psychologiques (en particulier chez les très jeunes) ;
- ✓ Modélisation de durée de vie, régression logistique

Comment se traduisent-elles en statistique ?

Equipement sportif

- ✓ Usure, performance, ergonomie, interaction avec l'être humain ;
- ✓ Analyse multidimensionnelle, fiabilité, analyse sensorielle

Analyse de match en sport collectif

- ✓ Analyse vidéo et stratégies de jeu, schéma de jeu performant (expected goal) ;
- ✓ Données GPS, diagramme de Voronoï, réseaux de passes, chaînes de Markov, graphes

Comment se traduisent-elles en statistique ?

Détection de dopage

- ✓ Passeport biologique, contrôle inopiné ;
- ✓ Détection de trajectoires / valeurs atypiques

Signature bio-mécanique d'un geste sportif

- ✓ Modélisation du squelette ;
- ✓ Prévention de la blessure ;
- ✓ Swing au golf, service au volley, foulée à la course ;
- ✓ Analyse du mouvement, données temporelles et multivariées

Comprendre les données

Comprendre la nature des données de performance

	Mesure directe	**Mesure indirecte**
Spécificité des données	Plutôt sous forme classique : • tableau rectangulaire (ind × var) • nombre de points, • temps en secondes • longueur, hauteur, distance en mètres • poids en kg, …	Données complexes : • résultats de tests en continu, • capteurs de mouvements, • GPS, • plate-forme de force
Exemples de sports	tennis, badminton, athlétisme, cyclisme, haltérophilie, …	escalade, suivi en ligne d'un match de rugby, foot, …

Combiner mesure objective et critères subjectifs

- Performance notée = score avec une part de subjectivité

Distance parcourue + style (envol, réception, phase d'atterrissage)

Difficulté du mouvement + déduction des fautes

Attention aux choix de modélisation

Images provenant de FlatIcon (Freepik, amoghdesign, Andrean Prabowo)

Variété des sports d'opposition

→ Oppositions individuelles vs oppositions collectives

→ La structure de succession des points durant le match peut être différente

Exemples de sports :

- basket-ball (1, 2 ou 3 points), rugby (2, 3 ou 5 points) – en temps limité
- football, hand-ball (seulement 1 point par but) – en temps limité
- badminton, tennis, volley-ball (1 point par succès) – nombre de sets gagnés

Bien définir la question statistique

- Définir sa variable d'intérêt
- Définir les variables explicatives
- Définir les groupes éventuels
- Que sont les individus qui composent l'échantillon ?
- Les données permettent-elles de répondre à la question ?

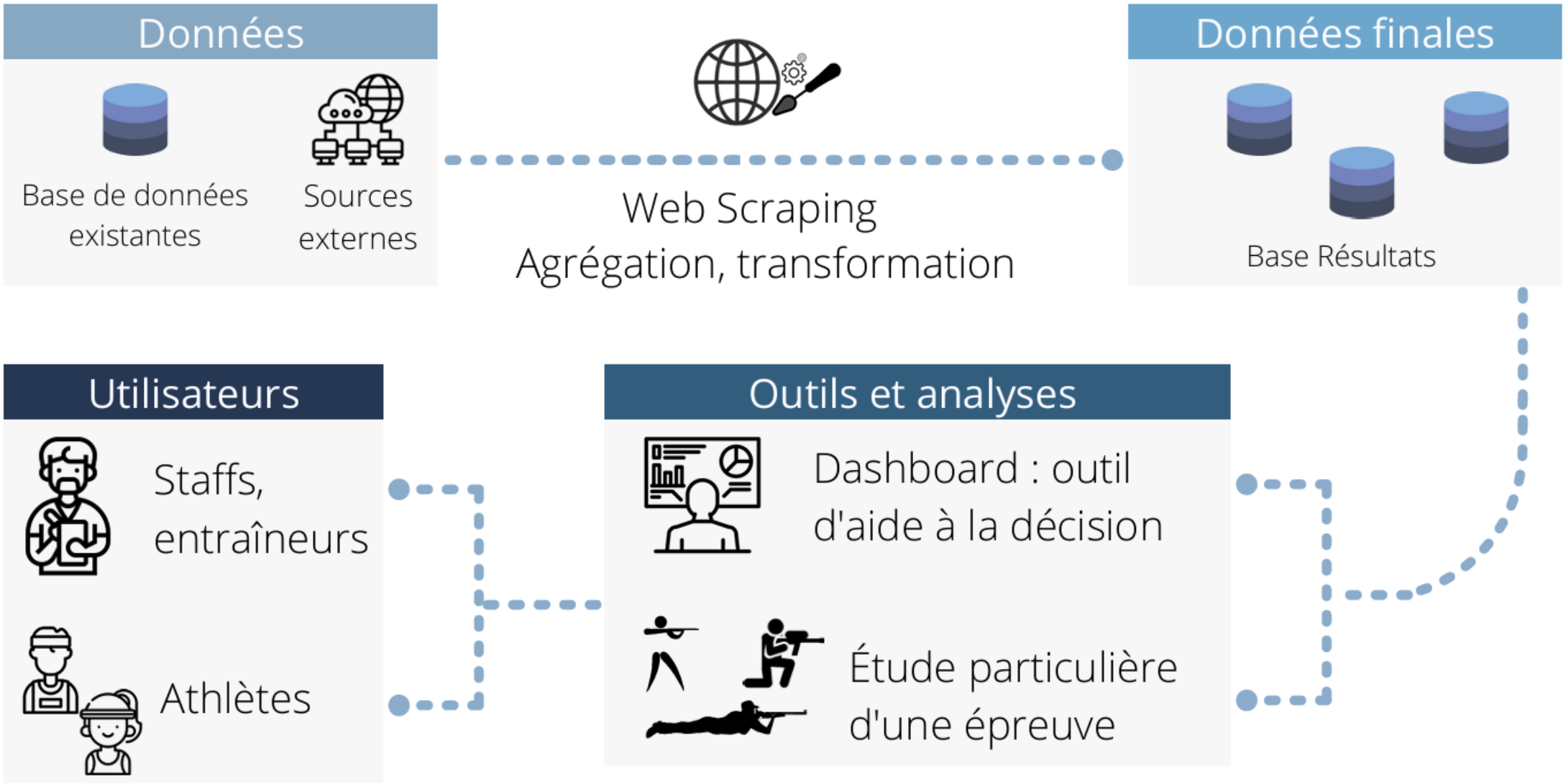
Prenons un exemple : *le projet Paraperf (INSEP)*

Objectif métier : mettre à disposition des outils pédagogiques répondant aux besoins des Fédérations et, à destination des coaches et des par-athlètes

Les objectifs scientifiques et techniques du projet (JP 2024) :

- Collecter et construire des bases de données pour chaque discipline
- Créer des indicateurs pour visualiser les trajectoires individuelles et identifier les déterminants de la progression
- Estimer la probabilité de gagner des médailles aux Jeux paralympiques
- Créer des modèles de détection individualisés en fonction des données disponibles selon la discipline

Le projet Paraperf *: architecture du projet*



Le projet Paraperf : *le para-tir sportif*

Distance	Epreuve	Discipline	Genre	Catégorie
10 mètres	R1	Carabine à air « couché »	Homme	SH1
	R2	Carabine à air « debout »	Femme	SH1
	R3	Carabine à air « couché »	Mixte	SH1
	R4	Carabine à air « debout »	Mixte	SH2
	R5	Carabine à air « couché	Mixte	SH2
	P1	Pistolet à air	Homme	SH1
	P2	Pistolet à air	Femme	SH1
25 mètres	P3	Pistolet	Mixte	SH1
50 mètres	P4	Pistolet	Mixte	SH1
	R6	Carabine « couché »	Mixte	SH1
	R7	Carabine 3 positions	Homme	SH1
	R8	Carabine 3 positions	Femme	SH1
	R9	Carabine « couché »	Mixte	SH2

Compétition en deux temps :

- **Match** : Nombre fixe de séries. Les 8 meilleurs sont qualifiés en finale
- **Finale** : Nombre fixe de séries puis coups à élimination.

Le para-tir sportif : résultats individuels

31		36	1137		FRA RICHARD Didier Sport Class: SH1	101.4	99.3	103.0	100.6	104.2	101.4	10.165	**609.9**

Series 1	Series 2	Series 3	Series 4	Series 5	Series 6
					
101.4	99.3	103.0	100.6	104.2	101.4

No	Score	No	Score
1	10.2	6	10.9
2	9.1	7	10.6
3	10.1	8	10.3
4	10.6	9	9.4
5	10.0	10	10.2

No	Score	No	Score
11	9.7	16	9.3
12	10.5	17	10.4
13	9.6	18	10.0
14	10.4	19	9.9
15	9.4	20	10.1

No	Score	No	Score
21	10.7	26	10.2
22	10.5	27	10.2
23	9.6	28	10.5
24	10.4	29	10.0
25	10.3	30	10.6

No	Score	No	Score
31	10.2	36	9.4
32	10.2	37	10.0
33	10.1	38	10.4
34	10.2	39	9.9
35	10.0	40	10.2

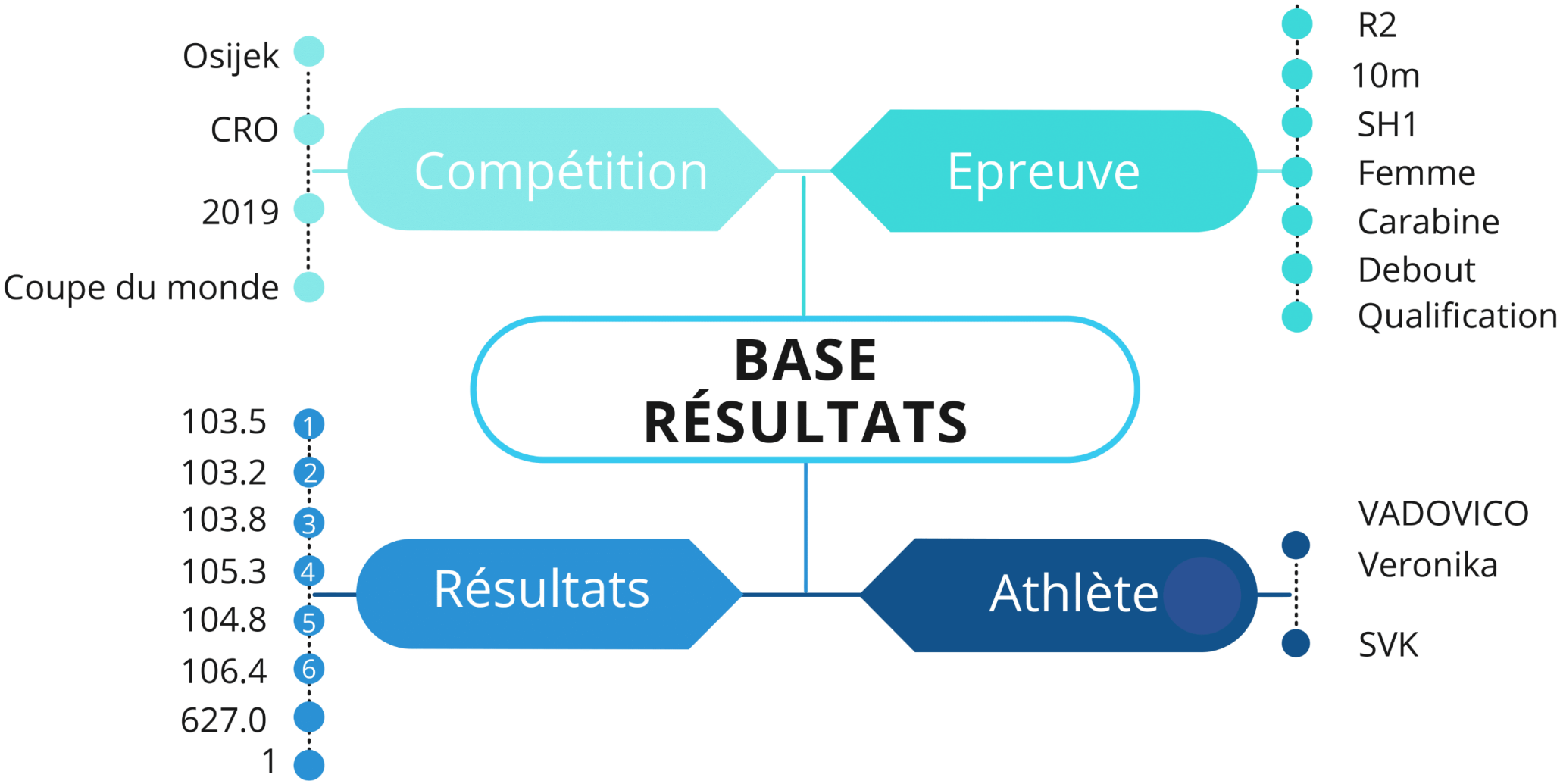
No	Score	No	Score
41	10.4	46	10.9
42	10.8	47	10.4
43	10.5	48	10.7
44	10.4	49	10.3
45	9.7	50	10.1

No	Score	No	Score
51	10.1	56	10.2
52	10.5	57	10.3
53	10.7	58	10.6
54	9.6	59	10.1
55	9.1	60	10.2

Aperçu des données

Lieu	Pays	Annee	Competition	Distance	Type	Classe	Sexe	Arme	Position	Phase	Rank	Name	Npc	Serie_1	Serie_2	Serie_3	Serie_4	Serie_5	Serie_6
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	106.4
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	104.3
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	6	HILTROP Natascha	GER	102.3	102	101.2	102.6	102.8	105
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	7	SEELIGER Elke	GER	101.5	101.7	104	102.2	101.6	103.9
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103	101.4
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101	97.3	101.2
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	12	HUANG Shu-Hua	TPE	98	99.5	98.9	96.5	96.3	102.3
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Qualification	DNS	LAMBERT Lorraine	GBR	NA	NA	NA	NA	NA	NA
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21	21	20.7
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21	21.3	20.1
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	4	FARMER Taylor	USA	51.3	50.3	20	21	20.2	20.8
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	6	SEELIGER Elke	GER	50	49.3	21.3	20.5	20	NA
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8	NA	NA
Osijek	CRO	2019	Coupe du Monde	10m	R2	SH1	Femme	Carabine	Debout	Final	8	HILTROP Natascha	GER	49.7	51.3	18.9	NA	NA	NA

Architecture des données



Une problématique : *analyse de la performance*

🎯 **Estimer la probabilité de gagner des médailles aux Jeux Paralympiques 2024**

Définissons la question statistique:

- ***Variable d'intérêt*** : podium/non podium en finale
- ***Variables explicatives*** : âge, nombre de points, indicateurs de perf individuelle
- ***Groupes*** : compétition, classe d'handicap, genre, compétition, année, distance
- ***Individus :*** un.e par-athlète renseigné.e par ses caractéristiques et ses résultats
- ***Les données permettent-elles de répondre à la question ?*** oui, mais pas assez détaillées

Quel type de variables avons-nous à disposition ?

Décrire les données

Identifier les types de variables

Variables quantitatives (numériques)	Variables qualitatives (catégorielles)
• **Intervalle** : valeurs possibles sur la droite ex: score, âge, distance • **Ratio** : valeurs possibles entre 0 et 1 (0-100%) ex: possession, taux de victoires • **Comptage** : valeurs discrètes ex: nombre de perfects au tir • **Temporelle** : séquences de valeurs ordonnées en temps discret ou en temps continu (données fonctionnelles)	• **Binaire (booléenne, dichotomique)** : valeurs 0/1, ex: victoire/défaite • **Ordinale** : valeurs ordonnées ex: classement, niveau de difficulté • **Nominale** : valeurs non ordonnées ex: catégorie sportive, type de handicap

Illustration sur les données Paraperf

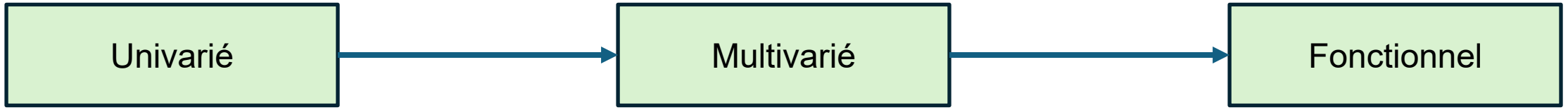
Binaire Ordinale Nominale Intervalle Ratio

Phase	Rank	Name	Npc	Serie 1	Serie 2	Serie 3	Serie 4	Serie 5	Serie 6	Serie 7	Serie 8	Serie 9	Total	I_perf
Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	106.4				627.0	0.9587
Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	104.3				621.2	0.9498
Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5				620.9	0.9494
Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4				618.2	0.9453
Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7				618.1	0.9451
Qualification	6	HILTROP Natascha	GER	102.3	102.0	101.2	102.6	102.8	105.0				615.9	0.9417
Qualification	7	SEELIGER Elke	GER	101.5	101.7	104.0	102.2	101.6	103.9				614.9	0.9402
Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9				606.8	0.9278
Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7				606.7	0.9277
Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103.0	101.4				606.6	0.9275
Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101.0	97.3	101.2				600.0	0.9174
Qualification	12	HUANG Shu-Hua	TPE	98.0	99.5	98.9	96.5	96.3	102.3				591.5	0.9044
Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98.0				589.3	0.9011
Qualification	DNS	LAMBERT Lorraine	GBR										0.0	
Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21.0	21.0	20.7	21.1	20.8	21.1	249.5	
Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7	21.0	21.1	21.3	248.2	
Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21.0	21.3	20.1	20.5	21.5		226.8	
Final	4	FARMER Taylor	USA	51.3	50.3	20.0	21.0	20.2	20.8	20.9			204.5	
Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5				183.4	
Final	6	SEELIGER Elke	GER	50.0	49.3	21.3	20.5	20.0					161.1	
Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8						140.8	
Final	8	HILTROP Natascha	GER	49.7	51.3	18.9							119.9	

Temporelle discrète

Comprendre la structure mathématique des données

Pour un individu donné :



Une observation unique (un scalaire)

ex:

- classement à la fin d'une course
- nombre de points marqués
- taille d'un athlète

Une observation de plusieurs variables (un vecteur)

ex:

- résultats des épreuves d'un décathlon
- statistiques d'un match (possession, tirs, …)
- résultats d'analyse anti-dopage

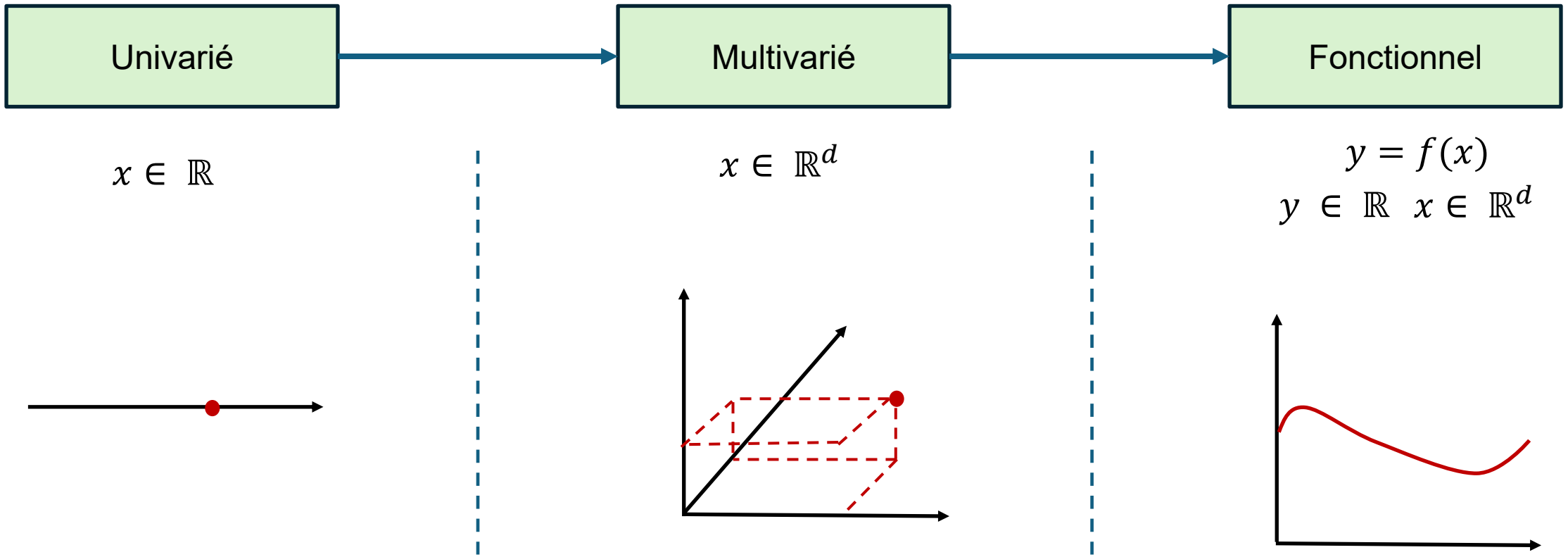
Une série de plusieurs (infinie) observations ordonnées et localement corrélées (une fonction)

ex:

- vitesse au cours du temps
- expected goals en fonction de la position sur le terrain
- Poids en fonction de la taille

Comprendre la structure mathématique des données

Pour un individu donné :



Le format tidy

R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

Les 3 commandements :

Chaque variable doit avoir sa propre colonne.

Chaque observation doit avoir sa propre ligne.

Chaque valeur doit avoir sa propre cellule.

En observant plusieurs individus

On désigne par X_1 un vecteur comportant n éléments (observations, individus), de même nature.
Alors X_1 est de **type univarié.**

$$X_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1i} \\ \vdots \\ x_{1n} \end{pmatrix}$$

On désigne par $X = (X_1, X_2, \dots, X_j, \dots, X_p)$ une matrice (ensemble de vecteurs juxtaposés) comportant p colonnes (variables) et n lignes (observations, individus). Les p variables peuvent être de nature différente.
Alors X est de **type multivarié.**

$$X = \begin{array}{c} \begin{matrix} X_1 & \cdots & X_j & \cdots & X_p \end{matrix} \\ \begin{pmatrix} x_{11} & \cdots & x_{j1} & \cdots & x_{p1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1i} & \cdots & x_{ji} & \cdots & x_{pi} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{jn} & \cdots & x_{pn} \end{pmatrix} \end{array} = \mathbf{x}_i$$

Illustration vectoriel vs fonctionnel

On s'intéresse aux performances de décathloniens. Vectoriel ou fonctionnel ?

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70	1	8217	Decastar
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50	2	8122	Decastar
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20	3	8099	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40	5	8036	Decastar
WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.10	6	8030	Decastar
ZSIVOCZKY	11.13	7.30	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268.00	7	8004	Decastar
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.10	8	7995	Decastar
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.60	4.92	52.33	262.10	9	7802	Decastar
HERNU	11.37	7.56	14.41	1.86	51.10	15.06	44.99	4.82	57.19	285.10	10	7733	Decastar
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.10	4.72	55.40	282.00	11	7708	Decastar
NOOL	11.33	7.27	12.68	1.98	49.20	15.29	37.92	4.62	57.44	266.60	12	7651	Decastar
BOURGUIGNON	11.36	6.80	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.70	13	7313	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	OlympicG
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	OlympicG

Illustration vectoriel vs fonctionnel

On s'intéresse aux performances de décathloniens. **Vectoriel !**

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70	1	8217	Decastar
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50	2	8122	Decastar
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20	3	8099	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40	5	8036	Decastar
WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.10	6	8030	Decastar
ZSIVOCZKY	11.13	7.30	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268.00	7	8004	Decastar
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.10	8	7995	Decastar
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.60	4.92	52.33	262.10	9	7802	Decastar
HERNU	11.37	7.56	14.41	1.86	51.10	15.06	44.99	4.82	57.19	285.10	10	7733	Decastar
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.10	4.72	55.40	282.00	11	7708	Decastar
NOOL	11.33	7.27	12.68	1.98	49.20	15.29	37.92	4.62	57.44	266.60	12	7651	Decastar
BOURGUIGNON	11.36	6.80	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.70	13	7313	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	OlympicG
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	OlympicG

Multivarié

Univarié

Aspect vectoriel *vs* fonctionnel

	X	frame_id	player	x	y
1	61530	4102	1	67.06116	14.048823
21	61550	4103	1	67.06116	14.048823
45	61574	4104	1	67.03518	14.039895
58	61587	4105	1	67.05643	14.033388
72	61601	4106	1	67.05643	14.033388
76	61605	4107	1	67.07178	14.045569
99	61628	4108	1	67.10484	14.077651
108	61637	4109	1	67.14853	14.106479
131	61660	4110	1	67.20757	14.147489
150	61679	4111	1	67.39765	14.199015
152	61681	4112	1	67.61251	14.217626
170	61699	4113	1	67.81205	14.300023
183	61712	4114	1	68.06592	14.483962
203	61732	4115	1	68.26075	14.626891
213	61742	4116	1	68.50281	14.772240
238	61767	4117	1	68.83931	14.922432
250	61779	4118	1	69.10378	15.027149

On s'intéresse à la position de joueurs de rugby au cours d'un match.
(frame_id correspond au 1/10e de seconde écoulé)

Vectoriel ou fonctionnel ?

Aspect vectoriel *vs* fonctionnel

	X	frame_id	player	x	y
1	61530	4102	1	67.06116	14.048823
21	61550	4103	1	67.06116	14.048823
45	61574	4104	1	67.03518	14.039895
58	61587	4105	1	67.05643	14.033388
72	61601	4106	1	67.05643	14.033388
76	61605	4107	1	67.07178	14.045569
99	61628	4108	1	67.10484	14.077651
108	61637	4109	1	67.14853	14.106479
131	61660	4110	1	67.20757	14.147489
150	61679	4111	1	67.39765	14.199015
152	61681	4112	1	67.61251	14.217626
170	61699	4113	1	67.81205	14.300023
183	61712	4114	1	68.06592	14.483962
203	61732	4115	1	68.26075	14.626891
213	61742	4116	1	68.50281	14.772240
238	61767	4117	1	68.83931	14.922432
250	61779	4118	1	69.10378	15.027149

On s'intéresse à la position de joueurs de rugby au cours d'un match.
(frame_id correspond au 1/10e de seconde écoulé)

Fonctionnel !

On peut s'intéresser à :

- x en fonction de frame_id
- y en fonction de frame_id
- une autre variable en fonction de x et y

Données fonctionnelles complexes

- **Multivariées** (avec $x \in \mathbb{R}^d$)

  ex: Données spatio-temporelles type GPS

- **Haute fréquence** (beaucoup d'observations)

  ex: Données de fréquence cardiaque

- **Irrégulièrement observées** (tous les individus n'ont pas le même nombre de points)

  ex: Données de suivi de âge-performance

Prétraitement des données

Pourquoi le prétraitement est indispensable ?

- Biais, données brutes non adaptées à la méthode statistique
- Valeurs absentes (manquantes), non applicable ;
- Valeurs atypiques (aberrantes), anomalies ;
- Fortes corrélations entre les variables, entre les observations

Identifier les données manquantes

Exemple

Phase	Rank	Name	Npc	Serie 1	Serie 2	Serie 3	Serie 4	Serie 5	Serie 6	Serie 7	Serie 8	Serie 9	Total
Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	106.4				627.0
Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	104.3				621.2
Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5				620.9
Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4				618.2
Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7				618.1
Qualification	6	HILTROP Natascha	GER	102.3	102.0	101.2		102.8	105.0				
Qualification	7	SEELIGER Elke	GER	101.5	101.7	104.0	102.2	101.6	103.9				614.9
Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9				606.8
Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7				606.7
Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103.0	101.4				606.6
Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101.0	97.3	101.2				600.0
Qualification	12	HUANG Shu-Hua	TPE	98.0	99.5	98.9	96.5	96.3	102.3				591.5
Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98.0				589.3
Qualification	DNS	LAMBERT Lorraine	GBR										0.0
Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21.0	21.0	20.7	21.1	20.8	21.1	249.5
Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7	21.0	21.1	21.3	248.2
Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21.0	21.3	20.1	20.5	21.5		226.8
Final	4	FARMER Taylor	USA	51.3	50.3	20.0	21.0	20.2	20.8	20.9			204.5
Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5				183.4
Final	6	SEELIGER Elke	GER	50.0	49.3	21.3	20.5	20.0					161.1
Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8						140.8
Final	8	HILTROP Natascha	GER	49.7	51.3	18.9							119.9

Identifier les données manquantes

Exemple

Phase	Rank	Name	Npc	Serie 1	Serie 2	Serie 3	Serie 4	Serie 5	Serie 6	Serie 7	Serie 8	Serie 9	Total
Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	106.4				627.0
Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	104.3				621.2
Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5				620.9
Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4				618.2
Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7				618.1
Qualification	6	HILTROP Natascha	GER	102.3	102.0	101.2	na	102.8	105.0				na
Qualification	7	SEELIGER Elke	GER	101.5	101.7	104.0	102.2	101.6	103.9				614.9
Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9				606.8
Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7				606.7
Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103.0	101.4				606.6
Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101.0	97.3	101.2				600.0
Qualification	12	HUANG Shu-Hua	TPE	98.0	99.5	98.9	96.5	96.3	102.3				591.5
Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98.0				589.3
Qualification	DNS	LAMBERT Lorraine	GBR	na	na	na	na	na	na				0.0
Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21.0	21.0	20.7	21.1	20.8	21.1	249.5
Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7	21.0	21.1	21.3	248.2
Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21.0	21.3	20.1	20.5	21.5	na	226.8
Final	4	FARMER Taylor	USA	51.3	50.3	20.0	21.0	20.2	20.8	20.9	na	na	204.5
Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5	na	na	na	183.4
Final	6	SEELIGER Elke	GER	50.0	49.3	21.3	20.5	20.0	na	na	na	na	161.1
Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8	na	na	na	na	na	140.8
Final	8	HILTROP Natascha	GER	49.7	51.3	18.9	na	na	na	na	na	na	119.9

Pas pour la même raison

Types de données manquantes

- **Complètement aléatoire (MCAR)** : absence de la valeur pour une raison inconnue

  ex : Absence des points de Natascha Hiltrop au tir à la carabine

- **Aléatoire (MAR)** : absence liée aux valeurs d'une ou plusieurs autres variables dans le jeu de données

  ex : Non réponse à la question sur le nombre de flexions dans une enquête sur la pratique du sport, lié à l'âge

- **Non aléatoire (MNAR)** : absence liée aux caractéristiques de la variable elle-même

  ex : Non réponse des hauts salaires par des footballeurs dans une enquête

- **Non applicable (NA)** : ne peut pas exister d'après la structure du tableau de données

  ex : le nombre de points au tir à la cabine d'une par-athlète dans la phase finale, alors qu'elle vient d'être éliminée au tour précédent

Gérer les données manquantes

Selon le type de données manquantes (MCAR, MAR, MNAR)

X	Y
a	e
b	NA
c	g
d	h

Cas complet

X	Y
a	e
c	g
d	h

Imputation simple

X	Y
a	e
b	f
c	g
d	h

Moyenne, médiane, + proche voisin, …

Imputation multiple

Imp	X	Y
1	a	e
1	b	f_1
1	c	g
1	d	h
2	a	e
2	b	f_2
2	c	g
2	d	h

MICE, EM + bootstrap, lissage fonctionnel, PCA fonctionnelle, …

Attention aux biais et à l'incertitude

Identifier les données atypiques

Exemple

Phase	Rank	Name	Npc	Serie_1	Serie_2	Serie_3	Serie_4	Serie_5	Serie_6	Serie_7	Serie_8	Serie_9	Total
Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	120				640.6
Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	109				626
Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5				620.9
Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4				618.2
Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7				618.1
Qualification	6	HILTROP Natascha	GER	102.3	102	101.2	102.6	102.8	105				615.9
Qualification	7	SEELIGER Elke	GER	101.5	101.7	104	102.2	101.6	103.9				614.9
Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9				606.8
Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7				606.7
Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103	101.4				606.6
Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101	97.3	101.2				600
Qualification	12	HUANG Shu-Hua	TPE	98	99.5	98.9	96.5	96.3	102.3				591.5
Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98				589.3
Qualification	DNS	LAMBERT Lorraine	GBR	NA	NA	NA	NA	NA	NA				0
Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21	21	20.7	21.1	20.8	21.1	249.5
Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7	21	21.1	21.3	248.2
Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21	21.3	20.1	20.5	21.5		226.8
Final	4	FARMER Taylor	USA	51.3	50.3	20	21	20.2	20.8	20.9			204.5
Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5				183.4
Final	6	SEELIGER Elke	GER	50	49.3	21.3	20.5	20					161.1
Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8						140.8
Final	8	HILTROP Natascha	GER	49.7	51.3	18.9							119.9

Identifier les données atypiques

Exemple

Phase	Rank	Name	Npc	Serie_1	Serie_2	Serie_3	Serie_4	Serie_5	Serie_6	Serie_7	Serie_8	Serie_9	Total
Qualification	1	VADOVICOVA Veronika	SVK	103.5	103.2	103.8	105.3	104.8	120				640.6
Qualification	2	LEKHARA Avani	IND	104.7	102.1	102.7	103.2	104.2	109				626
Qualification	3	SHCHETNIK Iryna	UKR	101.9	103.2	104.4	102.4	104.5	104.5				620.9
Qualification	4	FARMER Taylor	USA	101.6	104.2	102.9	105.2	102.9	101.4				618.2
Qualification	5	NORMANN Anna	SWE	102.5	102.8	103.9	102.4	102.8	103.7				618.1
Qualification	6	HILTROP Natascha	GER	102.3	102	101.2	102.6	102.8	105				615.9
Qualification	7	SEELIGER Elke	GER	101.5	101.7	104	102.2	101.6	103.9				614.9
Qualification	8	AL-WAELI Farah	IRQ	100.7	101.7	100.5	100.4	101.6	101.9				606.8
Qualification	9	LEUNGVILAI Wannipa	THA	104.3	100.9	101.1	101.9	99.8	98.7				606.7
Qualification	10	SAENLAR Chutima	THA	100.2	98.6	101.1	102.3	103	101.4				606.6
Qualification	11	PANTOVIC Jelena	SRB	100.5	100.3	99.7	101	97.3	101.2				600
Qualification	12	HUANG Shu-Hua	TPE	98	99.5	98.9	96.5	96.3	102.3				591.5
Qualification	13	BABSKA Emilia	POL	96.7	98.3	100.6	99.9	95.8	98				589.3
Qualification	DNS	LAMBERT Lorraine	GBR	NA	NA	NA	NA	NA	NA				0
Final	1	VADOVICOVA Veronika	SVK	50.9	52.1	20.8	21	21	20.7	21.1	20.8	21.1	249.5
Final	2	LEKHARA Avani	IND	50.4	50.3	21.4	20.8	21.2	20.7	21	21.1	21.3	248.2
Final	3	SHCHETNIK Iryna	UKR	50.9	50.7	20.8	21	21.3	20.1	20.5	21.5		226.8
Final	4	FARMER Taylor	USA	51.3	50.3	20	21	20.2	20.8	20.9			204.5
Final	5	NORMANN Anna	SWE	51.2	50.4	19.9	20.6	20.8	20.5				183.4
Final	6	SEELIGER Elke	GER	50	49.3	21.3	20.5	20					161.1
Final	7	AL-WAELI Farah	IRQ	48.4	50.8	20.8	20.8						140.8
Final	8	HILTROP Natascha	GER	49.7	51.3	18.9							119.9

Pas le même type de données atypiques

Données atypiques

Pourquoi la détection des données « atypiques » est primordiale ?
Meilleure connaissance des données initiales (influence sur les résultats issus des méthodes même si le choix est bon)

Exemple : Un point atypique peut influencer fortement la pente de la régression par la méthode des moindres carrés

La robustesse de la plupart des méthodes statistiques est souvent liée à l'homogénéité des données

Attention ! Détecter ne veut pas dire éliminer, mais étudier la nature et les causes de l'état du point « anormal » afin de décider de son statut ultérieurement

Type de données atypiques

- **Point contaminant** : valeur qui perturbe fortement mais qui reste tout à fait possible

  ex : 9,71s aux 100 mètres de Tyson Gay deuxième au Championnat du Monde d'athlétisme le 16 août 2009 derrière Usain Bolt avec 9,58s – record encore non détrôné

- **Point aberrant** : valeur qui n'a jamais été observée depuis le recueil des données

  ex : 8,90m au saut en longueur de Bob Beamon aux JO 1968 – le précédent record était détenu par Ralph Boston avec 8,35m

- **Point extrême** : en queue de distribution, très faible probabilité de l'observer

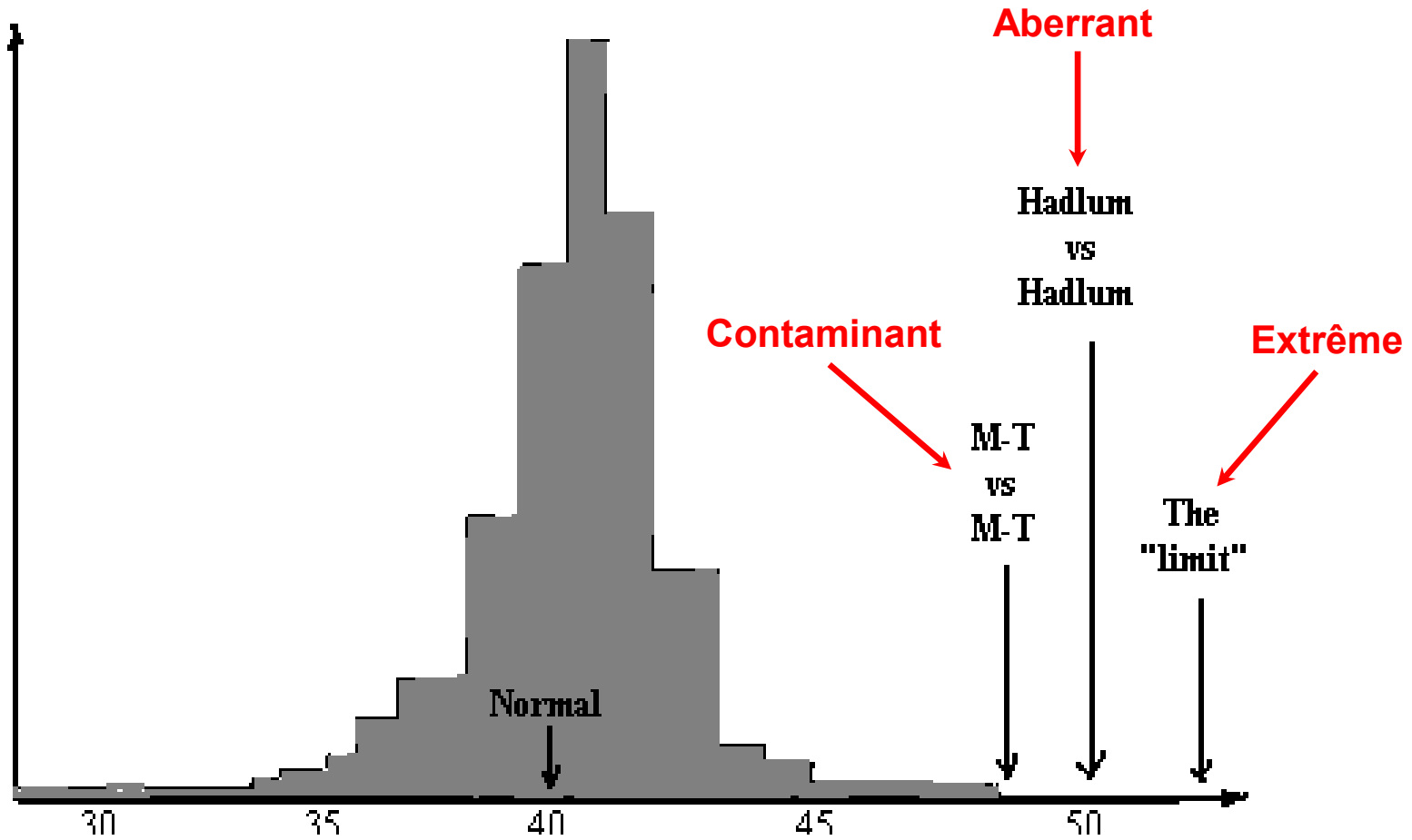
  ex : 109 points au tir à la carabine – c'est le maximum

- **Point impossible** : valeur ne rentrant pas dans l'univers de la variable

  ex : 120 points au tir à la carabine – impossible car ne peut excéder 109 points.

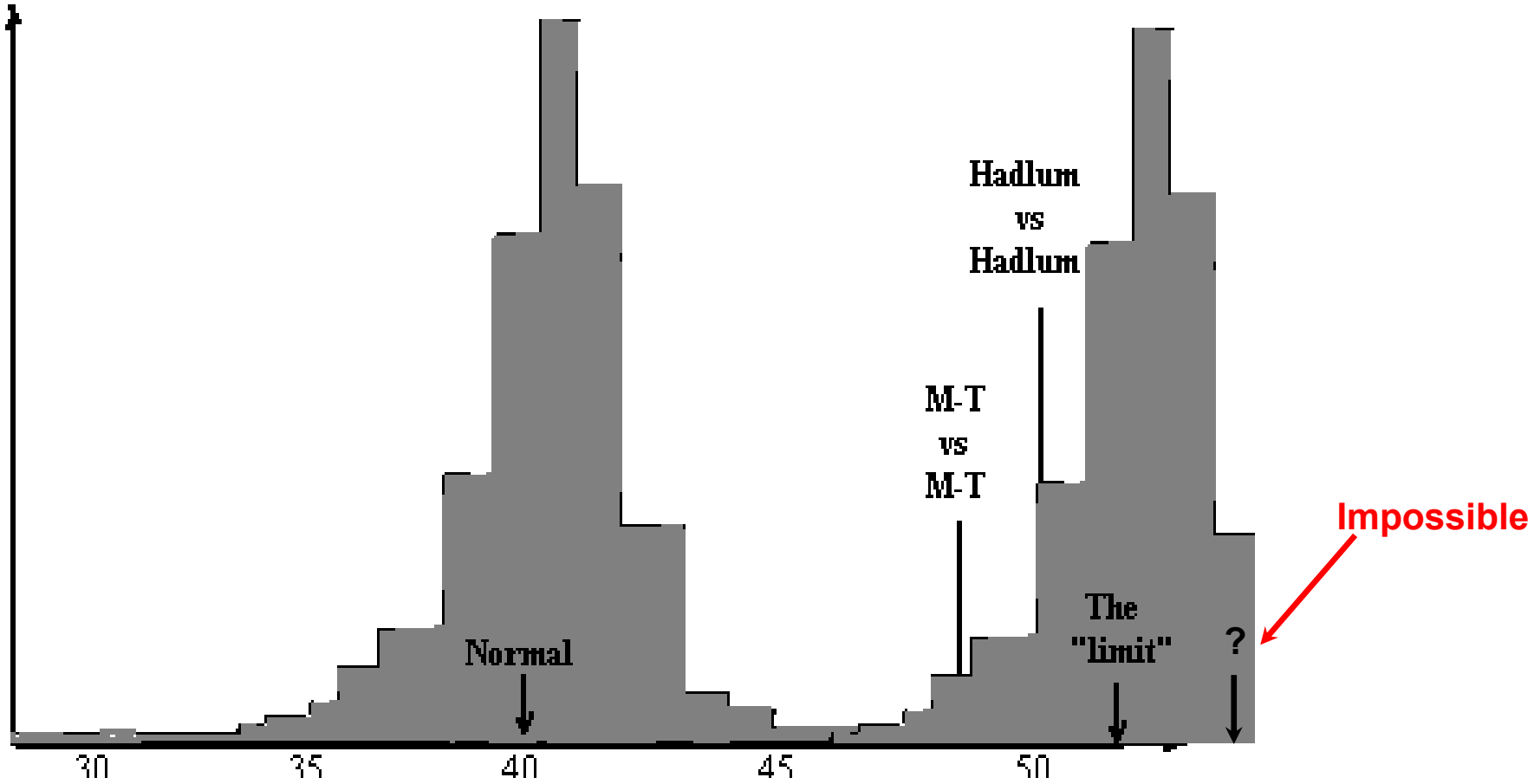
Données atypiques

Un exemple emprunté à Barnett & Lewis : *La durée de grossesse*



Données atypiques

… et son prolongement « extra-terrestre »

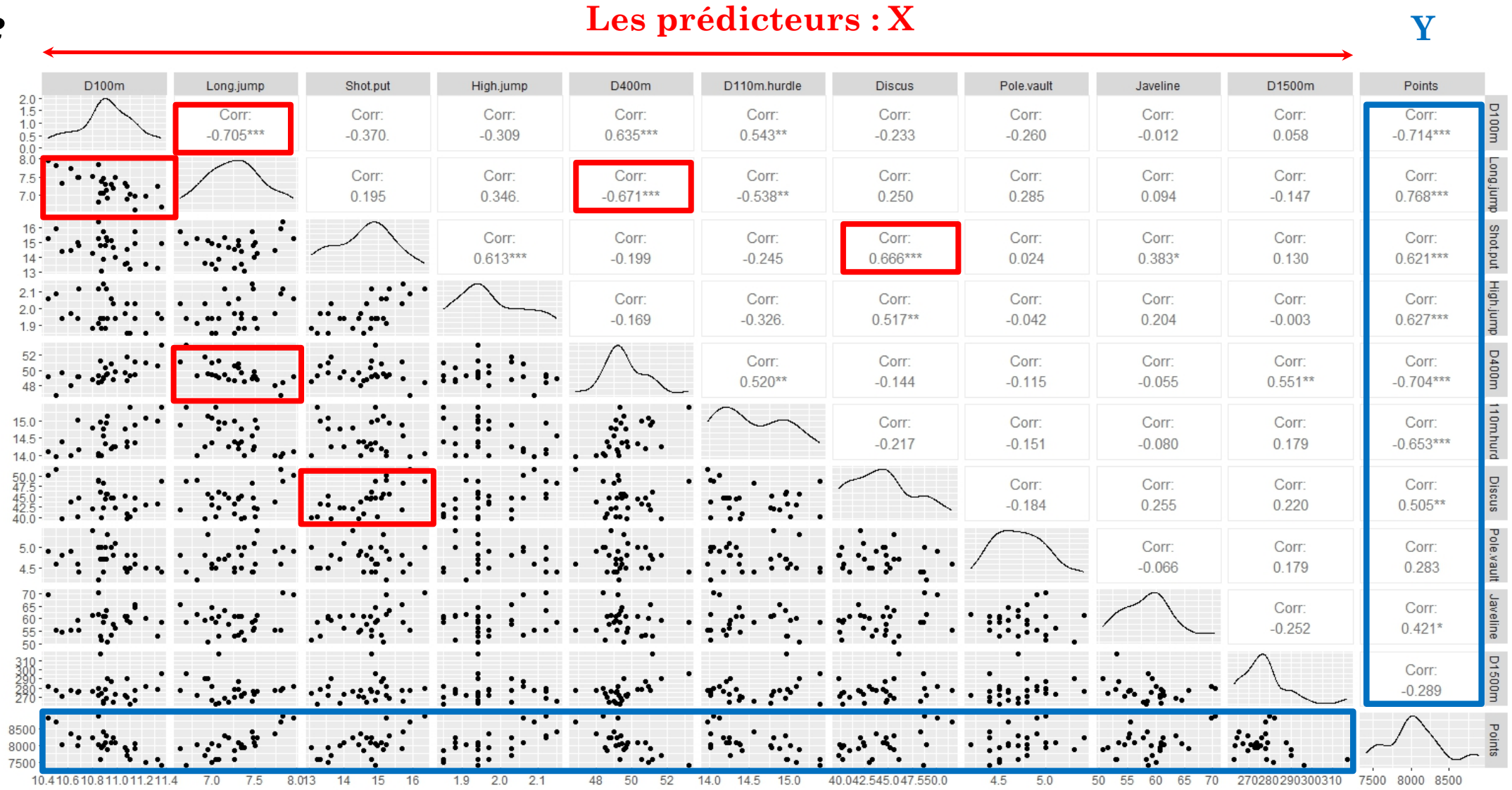


Gérer les données atypiques

- Rejet
- Correction
- Remplacement par une valeur manquante
- Révision du modèle
- Choix d’une méthode robuste

Identifier les fortes corrélations entre les variables

Exemple



Les méfaits des fortes corrélations entre les variables

- Signes incohérents des coefficients de régression multiple par rapport à ceux des coefficients de corrélation simples
- Oubli de prédicteurs dans le modèle de régression
- Haut R^2

Gérer les fortes corrélations entre les variables

- Régression sur composantes principales
- Régression PLS (Partial Least Squares)
- Régressions Ridge, Lasso,

Henri Poincaré – La science et l'hypothèse (1902)

Je veux déterminer une loi expérimentale ; cette loi, quand je la connaîtrai, pourra être représentée par une courbe ; je fais un certain nombre d'observations isolées ; chacune d'elles sera représentée par un point. Quand j'ai obtenu ces différents points, je fais passer une courbe entre ces points en m'efforçant de m'en écarter le moins possible et, cependant, de conserver à ma courbe une forme régulière, sans points anguleux, sans inflexions trop accentuées, sans variation brusque du rayon de courbure. Cette courbe me représentera la loi probable, et j'admets, non seulement qu'elle me fait connaître les valeurs de la fonction intermédiaires entre celles qui ont été observés, mais encore qu'elle me fait connaître les valeurs observées elles-mêmes plus exactement que l'observation directe (c'est pour cela que je la fais passer près de mes points et non pas par ces points eux-mêmes).

[...] Les effets ce sont les mesures que j'ai enregistrées ; ils dépendent de la combinaison de deux causes : la loi véritable du phénomène et les erreurs d'observation.

[...] **Nous faisons passer un trait continu, aussi régulier que possible, entre les points donnés par l'observation.** Pourquoi évitons-nous les points anguleux, les inflexions trop brusques ? Pourquoi ne faisons-nous pas décrire à notre courbe les zigzags les plus capricieux ? C'est parce que nous savons d'avance, ou que nous croyons savoir que la loi à exprimer ne peut pas être si compliquée que cela

Transformations simples *vs* lissage

Exemples de transformations simples entre 0 et 1

I_perf
0.9587
0.9498
0.9494
0.9453
0.9451
0.9417
0.9402
0.9278
0.9277
0.9275
0.9174
0.9044
0.9011

$$I_{perf}(i) = \frac{nb\ pts(i)}{\max possible}$$

Par rapport à une référence générale (le max de pts possible)

I_perf_b
1.0000
0.8462
0.8382
0.7666
0.7639
0.7056
0.6790
0.4642
0.4615
0.4589
0.2838
0.0584
0.0000

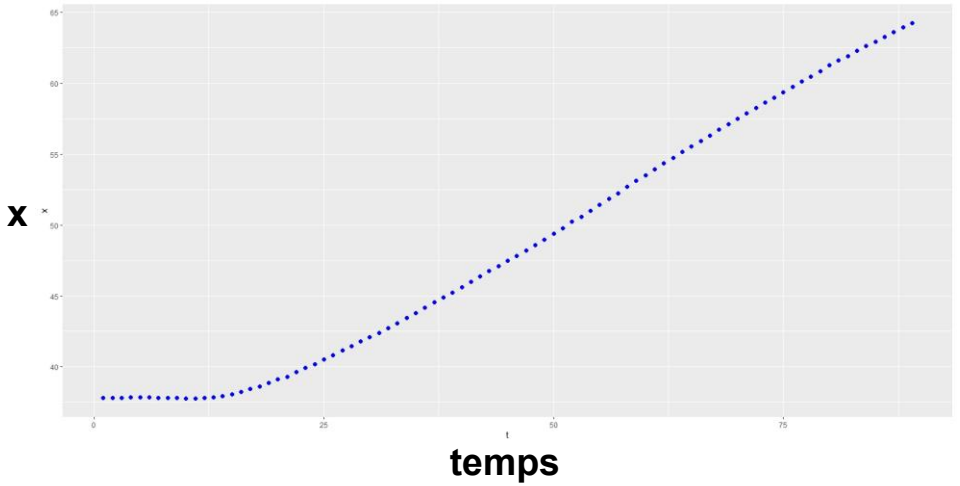
$$I_{perf_b}(i) = \frac{\left(nb\ pts(i) - \min_k nb\ pts(k)\right)}{\left(\max_k nb\ pts(k) - \min_k nb\ pts(k)\right)}$$

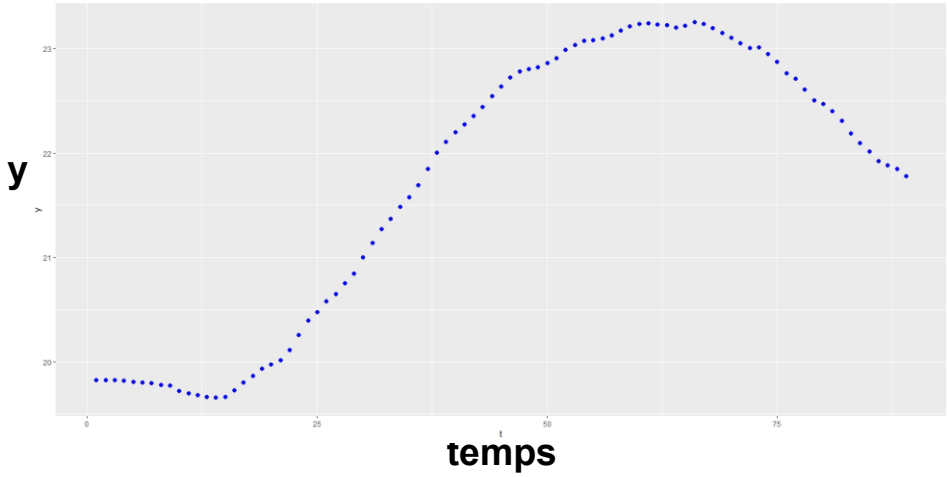
Par rapport aux données (le min et le max observes dans le tableau de données)

Transformations simples *vs* lissage

Exemples de transformations par lissage continu

	frame_id	player	x	y
1	4102	15	37.81618	19.82324
2	4103	15	37.81618	19.82324
3	4104	15	37.81618	19.82324
4	4105	15	37.83743	19.81673
5	4106	15	37.83507	19.80902
6	4107	15	37.85631	19.80251
7	4108	15	37.83034	19.79358
8	4109	15	37.82561	19.77815
9	4110	15	37.82325	19.77043
10	4111	15	37.78545	19.72291
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
80	4181	15	61.25134	22.47186
81	4182	15	61.61373	22.40308
82	4183	15	61.92181	22.30872
83	4184	15	62.26765	22.18592
84	4185	15	62.61233	22.09723
85	4186	15	62.91215	22.01385
86	4187	15	63.25683	21.92517
87	4188	15	63.59207	21.88158
88	4189	15	63.92969	21.84571
89	4190	15	64.23423	21.77776

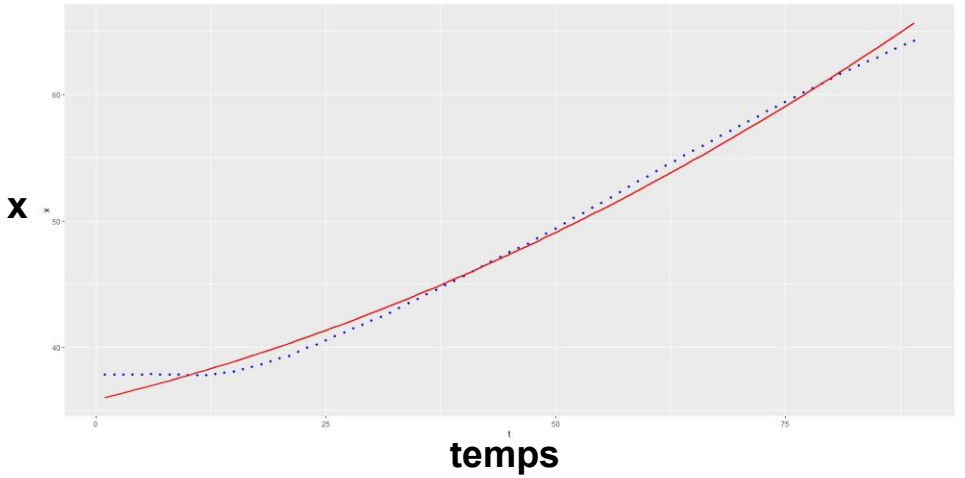


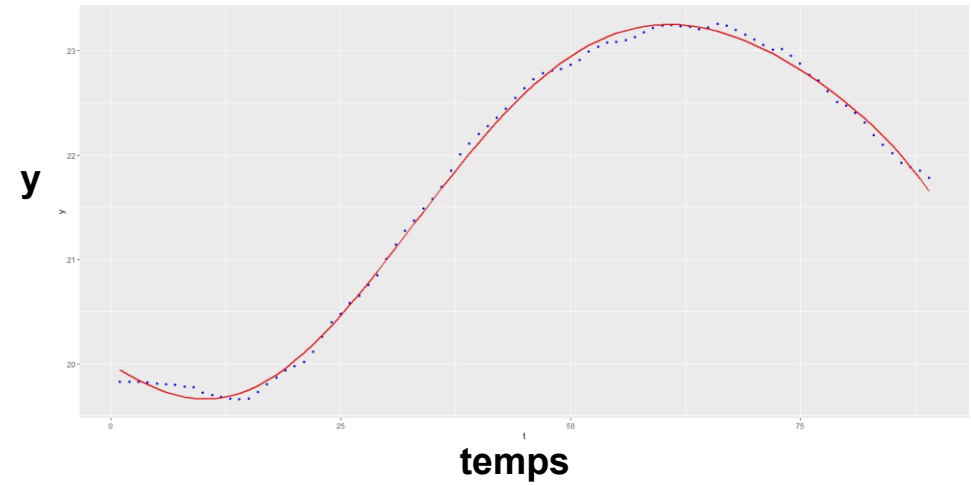


Transformations simples *vs* lissage

Transformation par lissage polynomial

	frame_id	player	x	y
1	4102	15	37.81618	19.82324
2	4103	15	37.81618	19.82324
3	4104	15	37.81618	19.82324
4	4105	15	37.83743	19.81673
5	4106	15	37.83507	19.80902
6	4107	15	37.85631	19.80251
7	4108	15	37.83034	19.79358
8	4109	15	37.82561	19.77815
9	4110	15	37.82325	19.77043
10	4111	15	37.78545	19.72291
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
80	4181	15	61.25134	22.47186
81	4182	15	61.61373	22.40308
82	4183	15	61.92181	22.30872
83	4184	15	62.26765	22.18592
84	4185	15	62.61233	22.09723
85	4186	15	62.91215	22.01385
86	4187	15	63.25683	21.92517
87	4188	15	63.59207	21.88158
88	4189	15	63.92969	21.84571
89	4190	15	64.23423	21.77776





Transformations simples *vs* lissage

Transformation par lissage polynomial

$$x = \beta_0 + \sum_{d \in D_{sel}} \beta_d t^d + \varepsilon$$

où *x* est la trajectoire, β_0 est une constante, $\beta_{d \in Dsel}$ est le coefficient de régression associé au degré *d* du temps *t* et ε est le résidu aléatoire

```
Call:
lm(formula = y ~ t + t_3 + t_4 + t_5 + t_6, data = joueur_15)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11421 -0.05265  0.00278  0.04794  0.12406

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.999e+01  3.133e-02  638.01   <2e-16 ***
t           -5.056e-02  3.583e-03  -14.11   <2e-16 ***
t_3          2.406e-04  9.583e-06   25.11   <2e-16 ***
t_4         -6.700e-06  3.324e-07  -20.16   <2e-16 ***
t_5          6.724e-08  4.219e-09   15.94   <2e-16 ***
t_6         -2.385e-10  1.852e-11  -12.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05964 on 83 degrees of freedom
Multiple R-squared:  0.9981,    Adjusted R-squared:  0.998
F-statistic:  8712 on 5 and 83 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = x ~ t + t_2, data = joueur_15)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41549 -0.60675  0.05567  0.54142  1.78984

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.585e+01  2.276e-01  157.53   <2e-16 ***
t           1.745e-01  1.167e-02   14.95   <2e-16 ***
t_2         1.801e-03  1.257e-04   14.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6996 on 86 degrees of freedom
Multiple R-squared:  0.9938,    Adjusted R-squared:  0.9937
F-statistic:  6902 on 2 and 86 DF,  p-value: < 2.2e-16
```

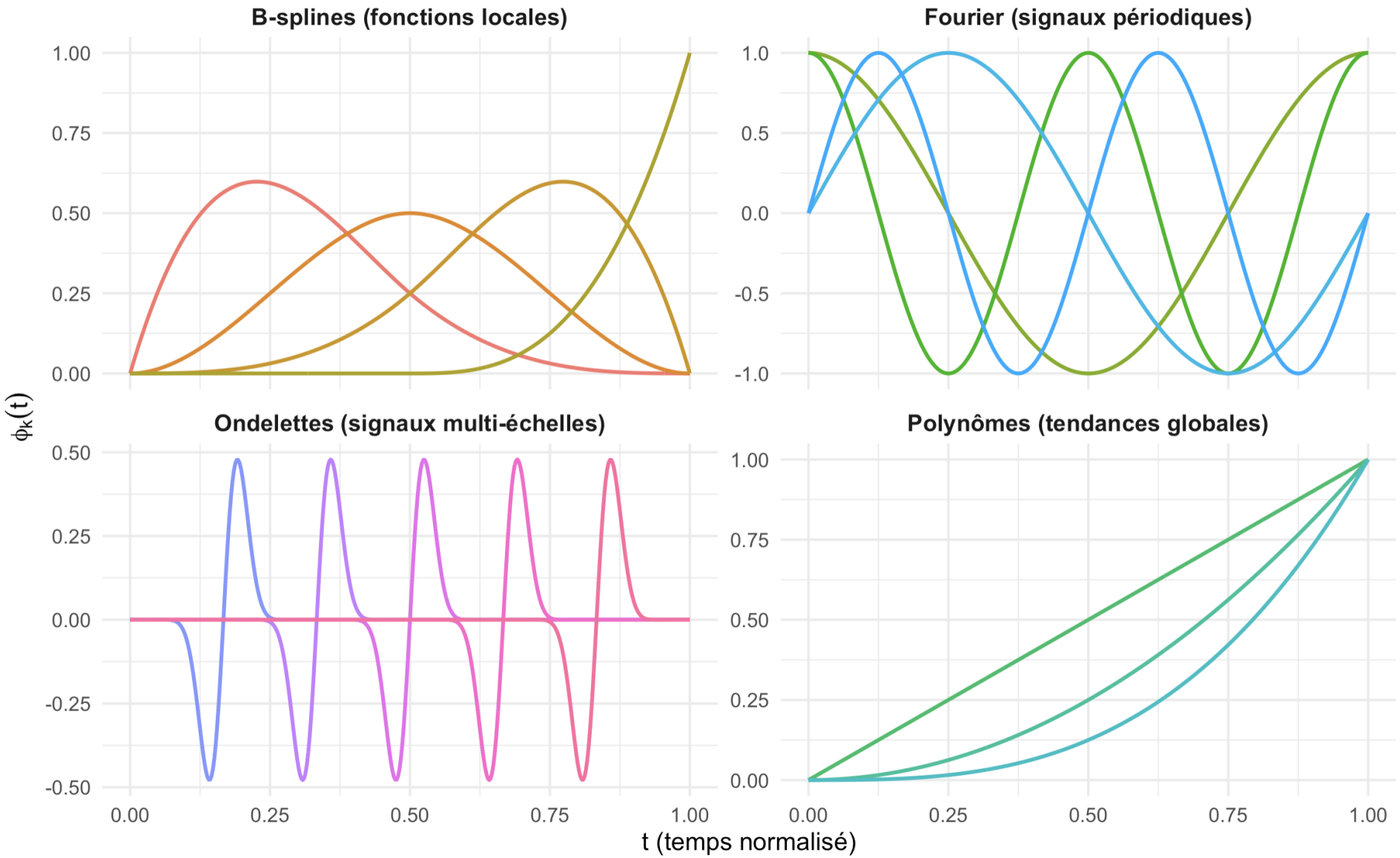
Décomposition en fonctions de base

Une fonction $f(t)$ peut être approchée comme une **combinaison pondérée** de fonctions de base :

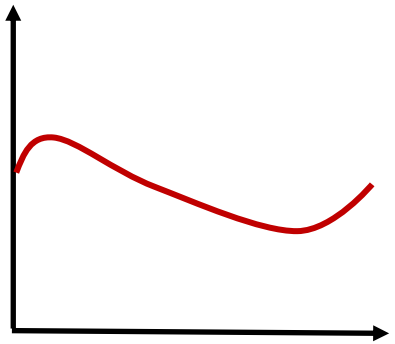
$$f(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$$

- Les $\phi_k(t)$ sont des fonctions de base (orthogonales), chacune définie en un nœud de l'espace des entrées.
- Les c_k sont les coefficients, caractéristiques de la fonction approximée.

Exemples de fonctions de base



Travailler sur les coefficients



$$f_i(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

$$\begin{pmatrix} c_{11} & \cdots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NK} \end{pmatrix}$$

Décomposition en fonction de bases

Estimation des coefficients (ex. moindres carrés)

Les coefficients peuvent alors être étudiés avec les méthodes d'analyse multivariée usuelles.

- Analyse en composantes principales
- Régressions
- Clustering

Bonnes pratiques

Bonnes pratiques

Liste non exhaustive

- **Le problème doit être bien posé** : répondre à des enjeux identifiés sur un périmètre donné, compréhensible et explicable
- **Les résultats doivent être interprétables, utiles et utilisables** : par les demandeurs afin qu'ils puissent se les approprier (coaches, sportifs professionnels, fédérations, …)
- **Les données doivent être fiables** : provenance, historique et recueil des données
- **Prétraitement des données** : elles doivent être utilisables et lisibles par la(les) approches statistiques afin d'éviter les biais d'interprétation
- **Adéquation des méthodes statistiques** : elles doivent pouvoir répondre à la question de départ grâce à l'aide d'outils d'interprétation fiables et adaptés

Bibliographie

Bibliographie : données atypiques

Barnett V. & Lewis T, (1987), *Outliers in Statistical Data*, Wiley & Sons, New-York

Hampel F.R, Ronchetti E.M., Rousseeuw P.J. & Stahel W.A., (1986), *The Robust Statistics, The Approach Based on the Influence Functions*, Wiley & Sons, New-York

Rousseeuw P.J. & Leroy A.M., (1987), *Robust Regression & Outlier Detection*, Wiley & Sons, New-York

Bibliographie : données manquantes

Chavent M., Kuentz V. & Liquet B., (2006), Données manquantes en ACM : l'algorithme NIPALS, ***SFC'09***, Grenoble, France

Rubin D.B., (1976), Inference and Missing Data, Biometrika, **63**, 591-597

Rubin D.B., (1987), *Multiple Imputation for Nonresponse in Surveys*, New-York, Wiley & Sons

Shafer J.L. & al, (2002), Missing Data: Our View of the Art, *Psychology Methods*, **7-2**, 147-177

Wold H., (1966), Estimation of the Principal Components and Related Models by Iterative Least Squares, Krishnaiah P.R. Editor, *Multivariate Analysis,* 391-420, Academic Press, New-York

Wold H., (1973), Nonlinear Iterative Partial Least Squares (NIPALS) Modelling some Current Developments, in Krishnaiah P.R. Editor, *Multivariate Analysis*, **III**, 391-420, Academic Press, New-York

Bibliographie : fortes corrélations entre les variables

Hoerl, A.E. & Kennard R.W., (1970) "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* **42** (1): 80–86.

Tibshirani, R., (1996), Regression Shrinkage and Selection via the Lasso, J. R. Statist. Soc. B, 58, No. 1, 267-288.

Wold S., Ruhe A., Wold H. & Dunn III W.J., (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput*., **5**, n°3, 735-743

Zou H. & Hastie T., (2005), Regularization and Variable Selection via the Elastic Net, J. R, Statist. Soc. B, 67, No. 2, 301-320.

Bibliographie générale

Jay Gould S., (1981) La Mal-mesure de l'homme, Edition Odile Jacob.

Servan Schreiber F. & Mauriac E. (2025), Futurs champions, le prix de la gloire https://www.arte.tv/fr/videos/115069-000-A/futurs-champions-le-prix-de-la-gloire/