Fusion statistique de données d'enquêtes

Dernières avancées pour les mesures d'audience

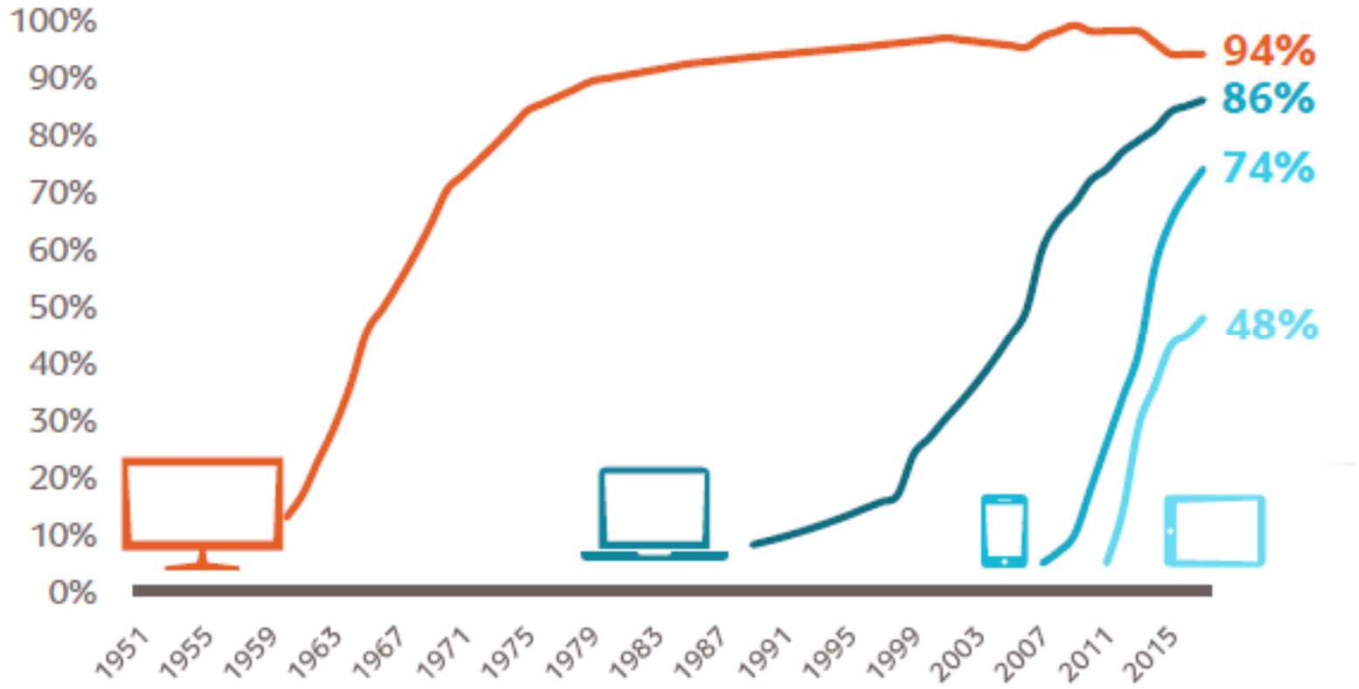
Lorie Dudoignon

Lyon, 25 novembre 2018

Mediametrie

- Contexte
- Etats des lieux
- Internet Global 1ère Génération
- Internet Global 2ème Génération

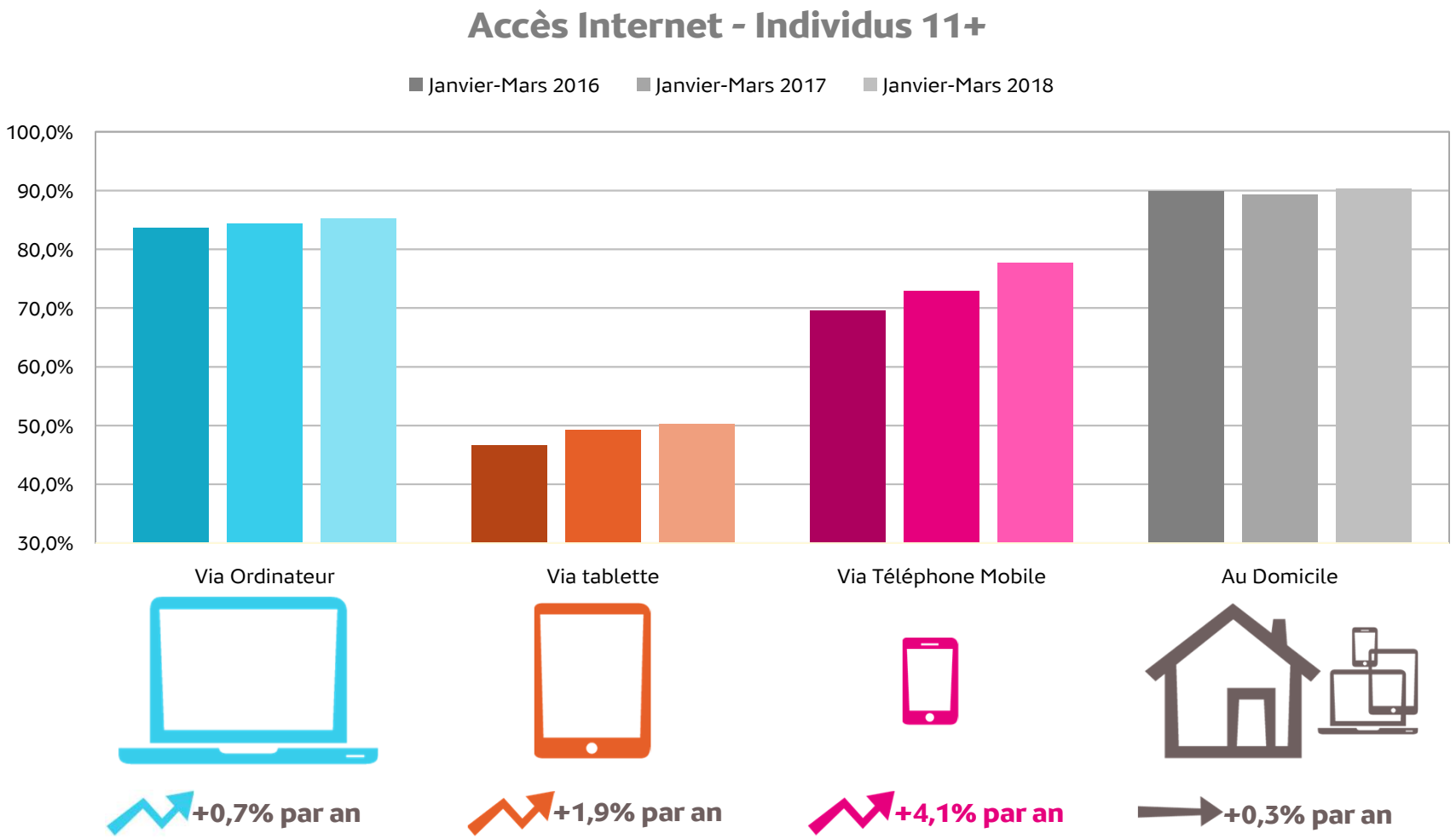
Accélération de l'adoption de nouvelles technologies



- En moins de 10 ans plus de 70% des foyers français possèdent au moins un smartphone
- Plus de 40 ans pour atteindre un tel niveau d'équipement pour la télévision

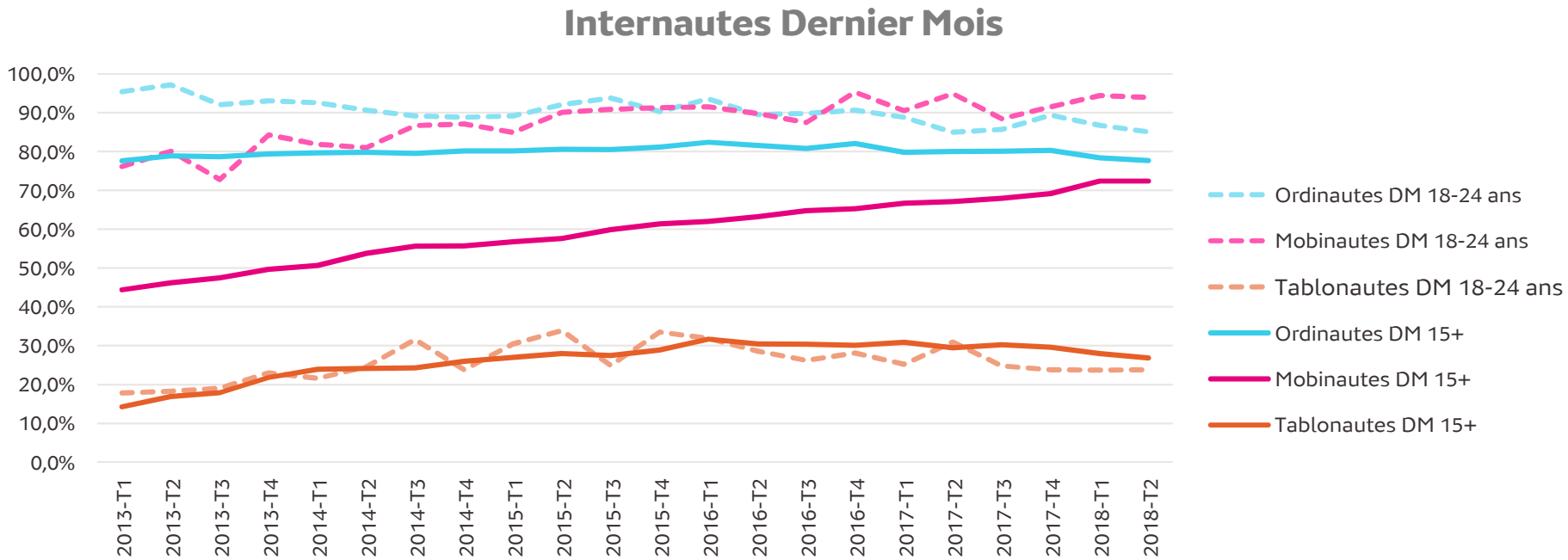
Source : Médiamétrie – La Référence des Equipements Multimédia -Home Devices;

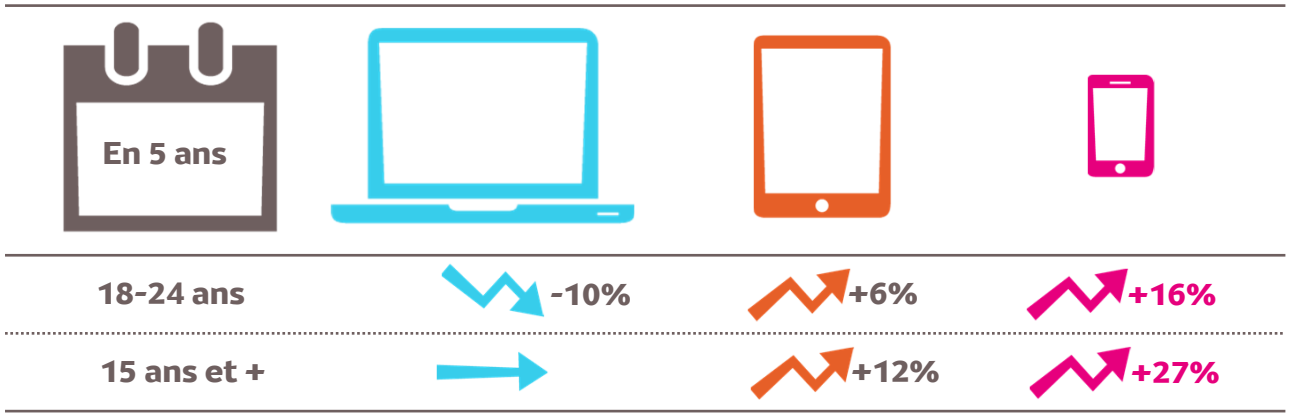
Multiplication des accès Internet



Source : Médiamétrie – Observatoire des Usages Internet;

Multiplication des accès Internet





Sources : Médiamétrie – Observatoire des Usages Internet & Web Observatoire

Les mesures d'audiences Internet

Mesure Ordinateur

Mesure Mobile

Mesure Tablette

Les mesures d'audiences Internet

Mesure Ordinateur

Panel
18.000 individus

- Panel Mediametrie//Netratings
- Mesure 2+ :
  - Grappage foyer
  - Meter Netsight de Nielsen + Déclaration

Les mesures d'audiences Internet

Mesure Mobile

Panel
10.000 individus

- Panel Mediametrie
  - Naissance en 2010 : Logs opérateurs
  - 2016 : Meter Reality Mine
- Mesure 11+
  - Individu

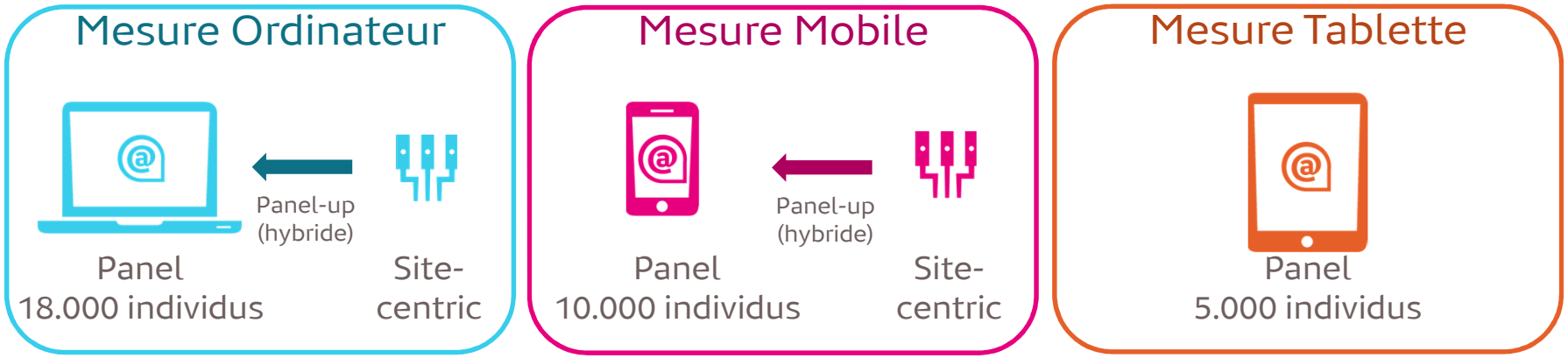
Les mesures d'audiences Internet

Mesure Tablette

Panel
5.000 individus

- Panel Mediametrie
  - Naissance en 2012
  - 2014
- Mesure 2+
  - Grappage foyer
  - Mesure Proxy + Appli de déclaration

Les mesures d'audiences Internet

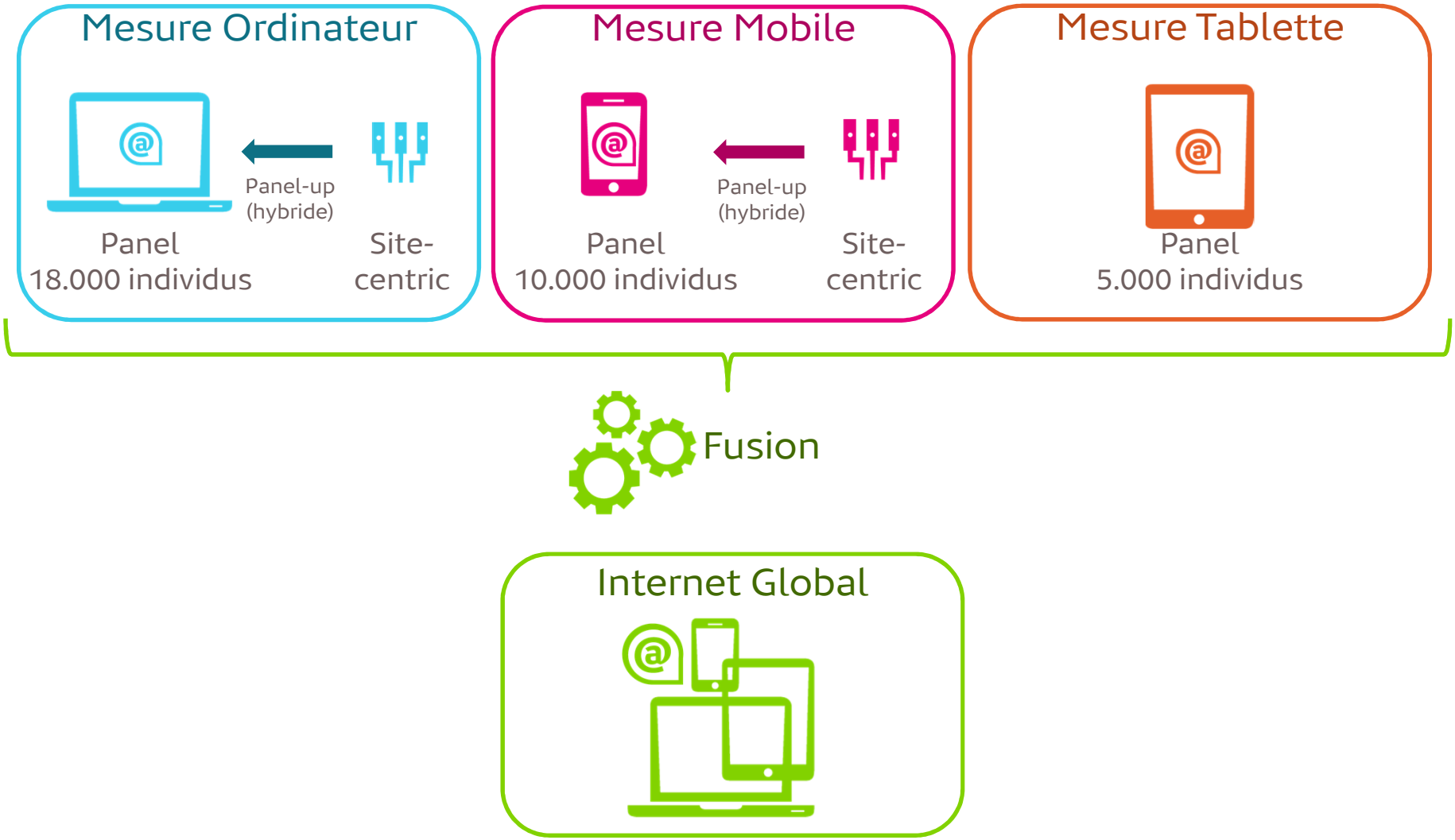


- Site-Centric
  - Marqueur Site / Appli
  - → Nombre de visites

- Hybridation Panel-up
  - Site-Centric = Information auxiliaire
  - Calage du nombre de visites

Références :
Dudoignon, L. & Zydorczak, L. Enquête et données exhaustives : un nouveau défi pour les mesures d'audience, Colloque Francophone sur les Sondages, Rennes, 2012.

Internet Global



Approche simple du type donneur - receveur

Base receveuse

Individus	Variables A	Variables B	Variables C
1			
i			?
N			

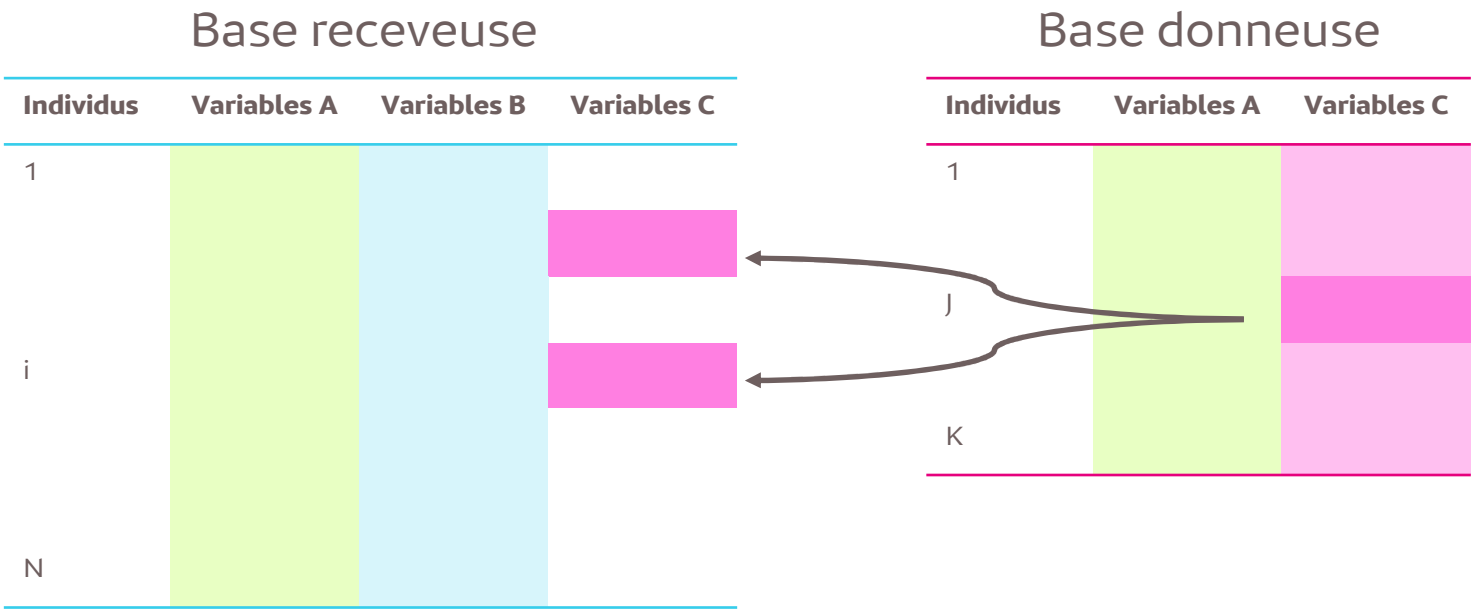
Base donneuse

Individus	Variables A	Variables C
1		
J		
K		

- Les variables A sont renseignées dans les 2 bases. → **Variables de pont**
- Les variables B sont spécifiques à la base receveuse.
- Les variables C sont spécifiques à la base donneuse. → **Variables à transférer**

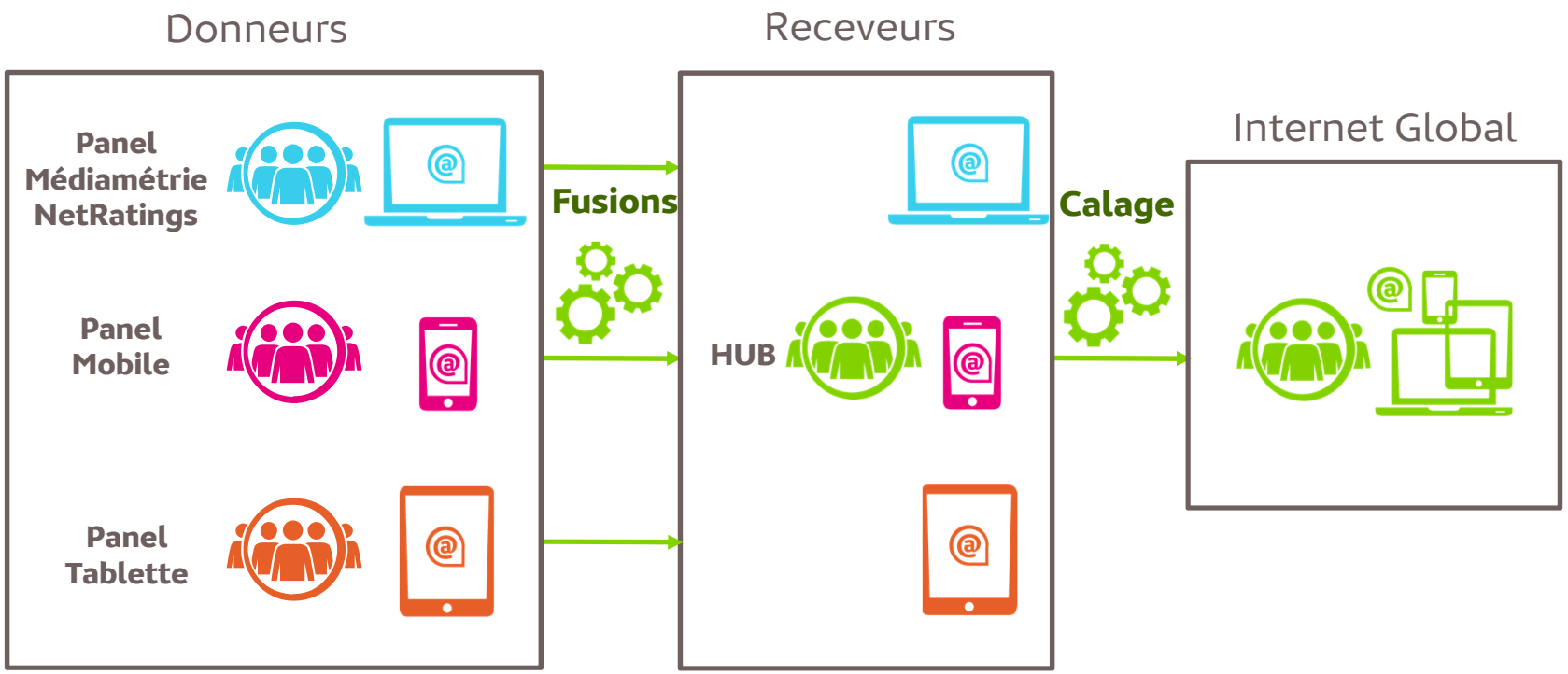
Références :
Fisher, N. (2004), Fusion statistique de fichiers de données, Thèse de doctorat, Montpellier

Approche simple du type donneur - receveur



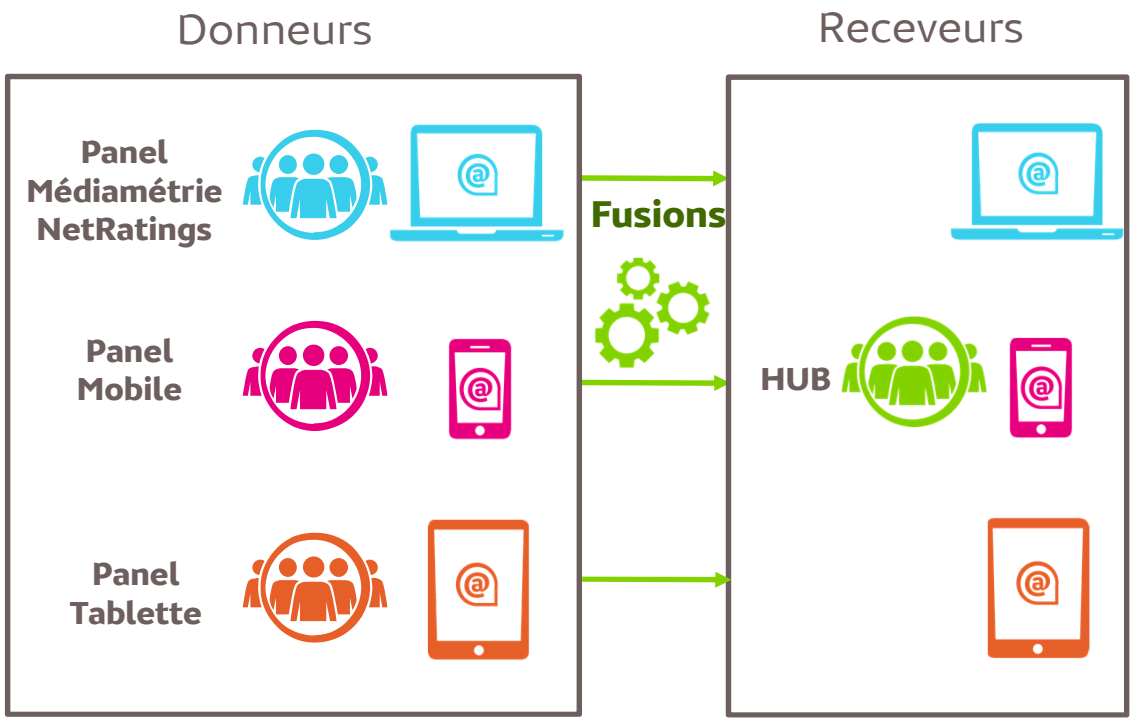
- Pour chaque individu i receveur :
  - Trouver un donneur j qui lui « ressemble »
  - Transférer les données de j à i (pour les variables C).
- Pour déterminer si deux individus se ressemblent, toutes les variables communes A peuvent être utilisées
- Un même donneur peut être utilisé pour plusieurs receveurs.

3 fusions + 1 hub



- HUB
  - 10.000 interviews / an
  - Questionnaire d'Habitude d'Ecoute auto-administré
    - Habitudes par écran au global
    - Habitudes par écran x sites

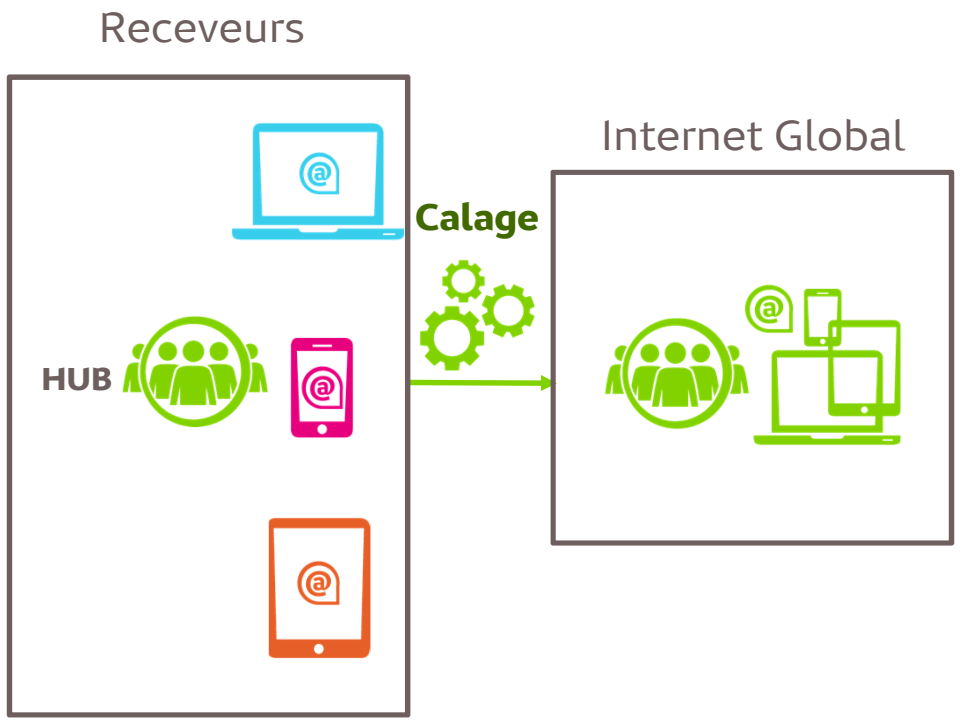
3 fusions + 1 hub



Fusions

- Approche Donneur → Receveur
  - Contrainte sur le nombre de réplication d'un même donneur
- « Distance scorée » sur les habitudes d'écoute
  - Score(Visiteur, Visiteur) > Score(Non visiteur, Non visiteur)

3 fusions + 1 hub



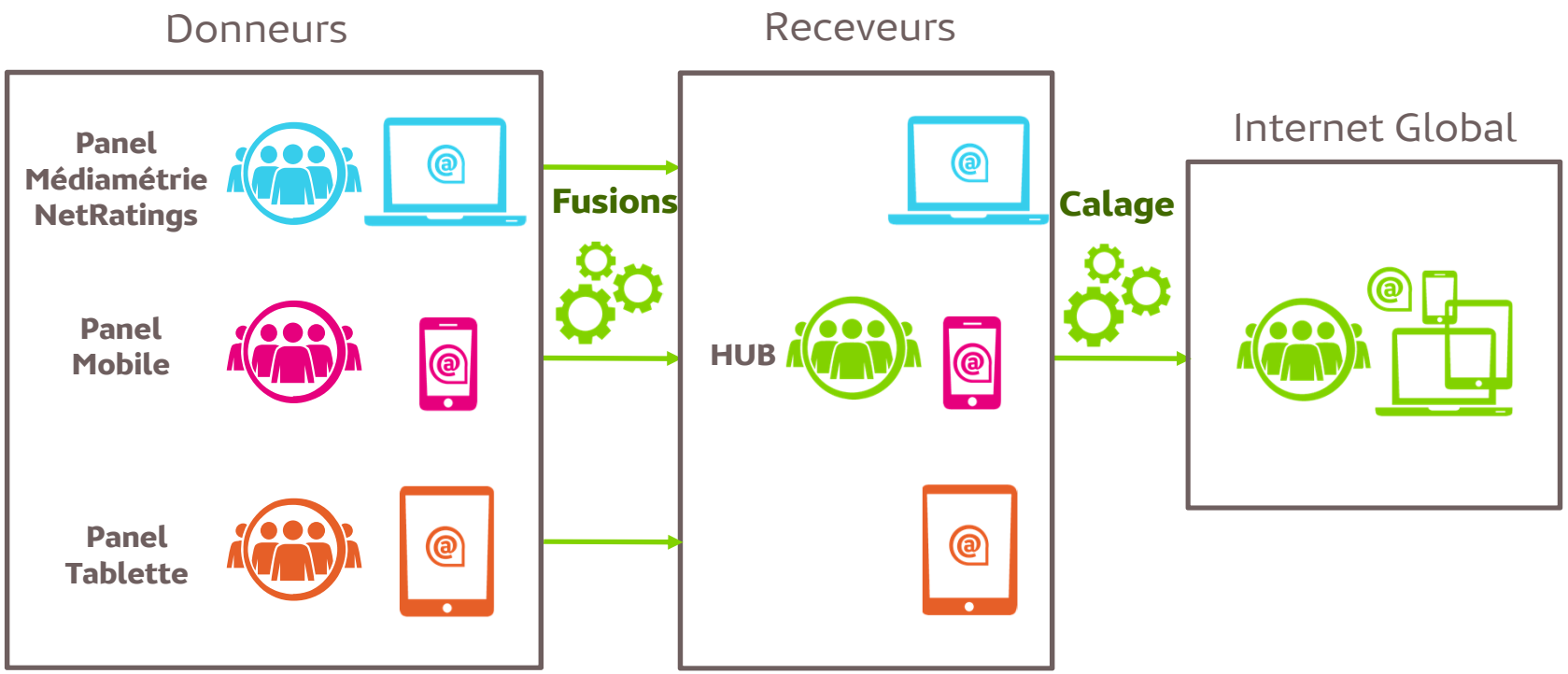
Calage

- Redressement Ridge
  - Tolérance sur les marges issues des enquêtes donneuses

Références :

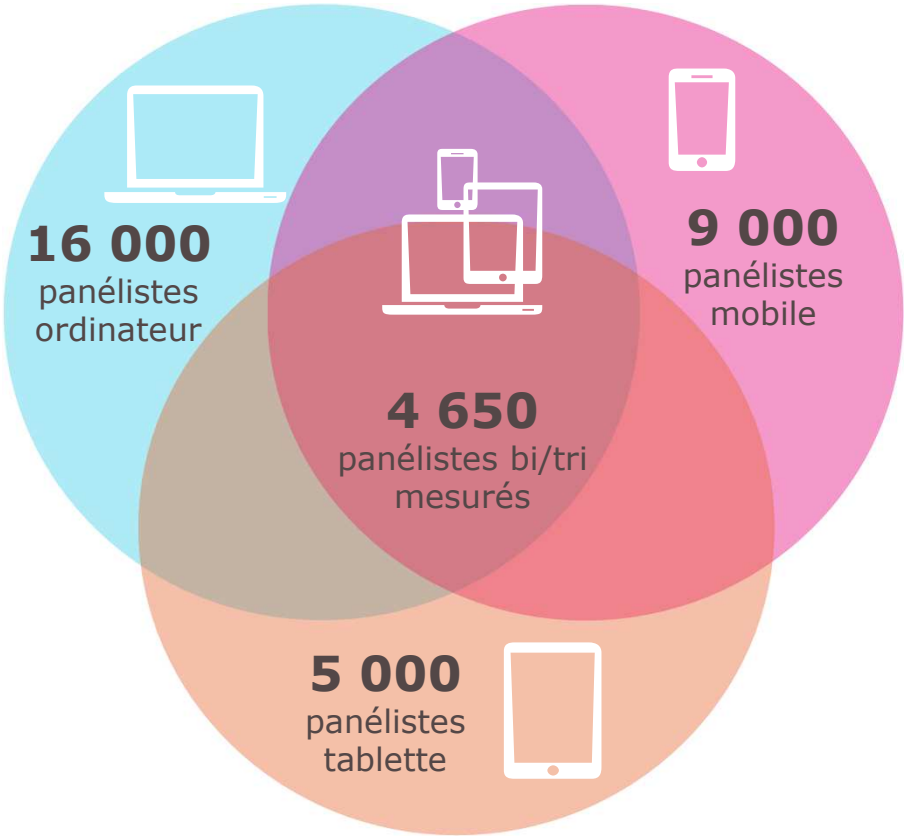
Beaumont, J.-F. and Bocci, C. (2008), Another look at ridge calibration, Metron-International Journal of Statistics, LXVI, 5–20.

3 fusions + 1 hub



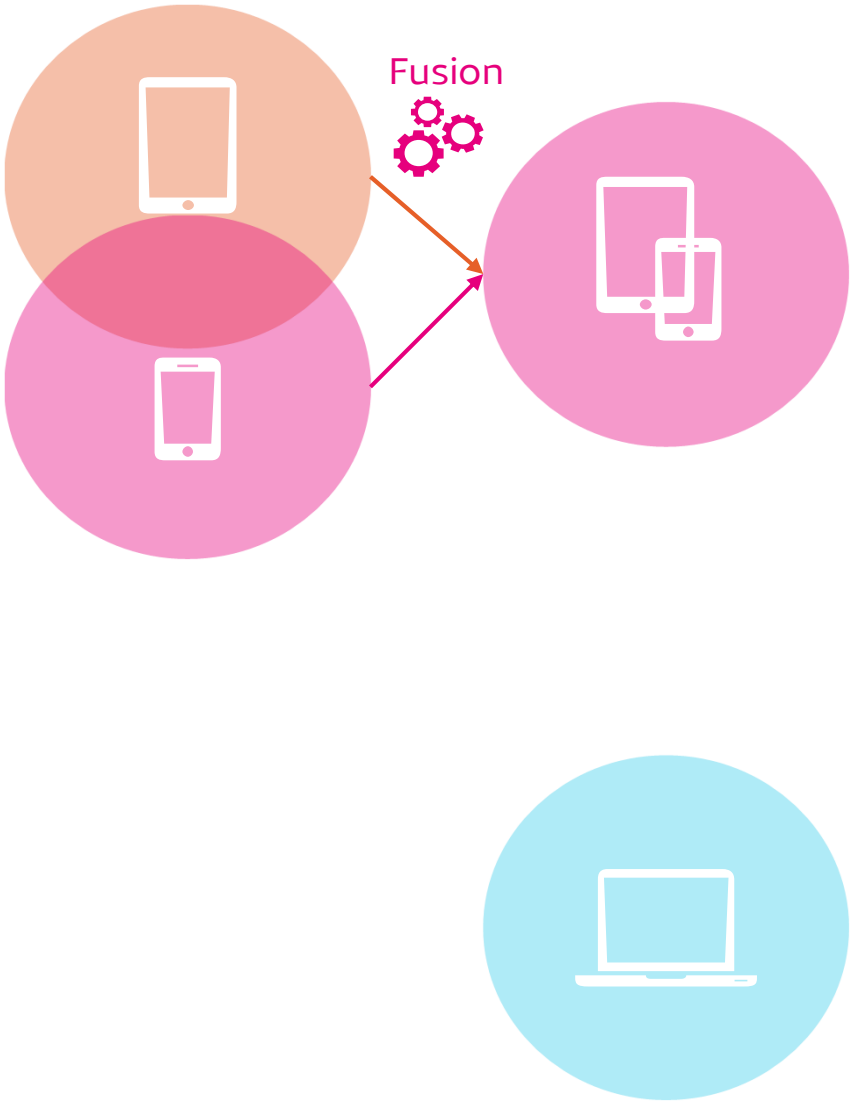
- **Résultats par écran ≠ Etude de référence**
- **HUB mis à jour 2 fois par an**

Une panélisation « à la carte »

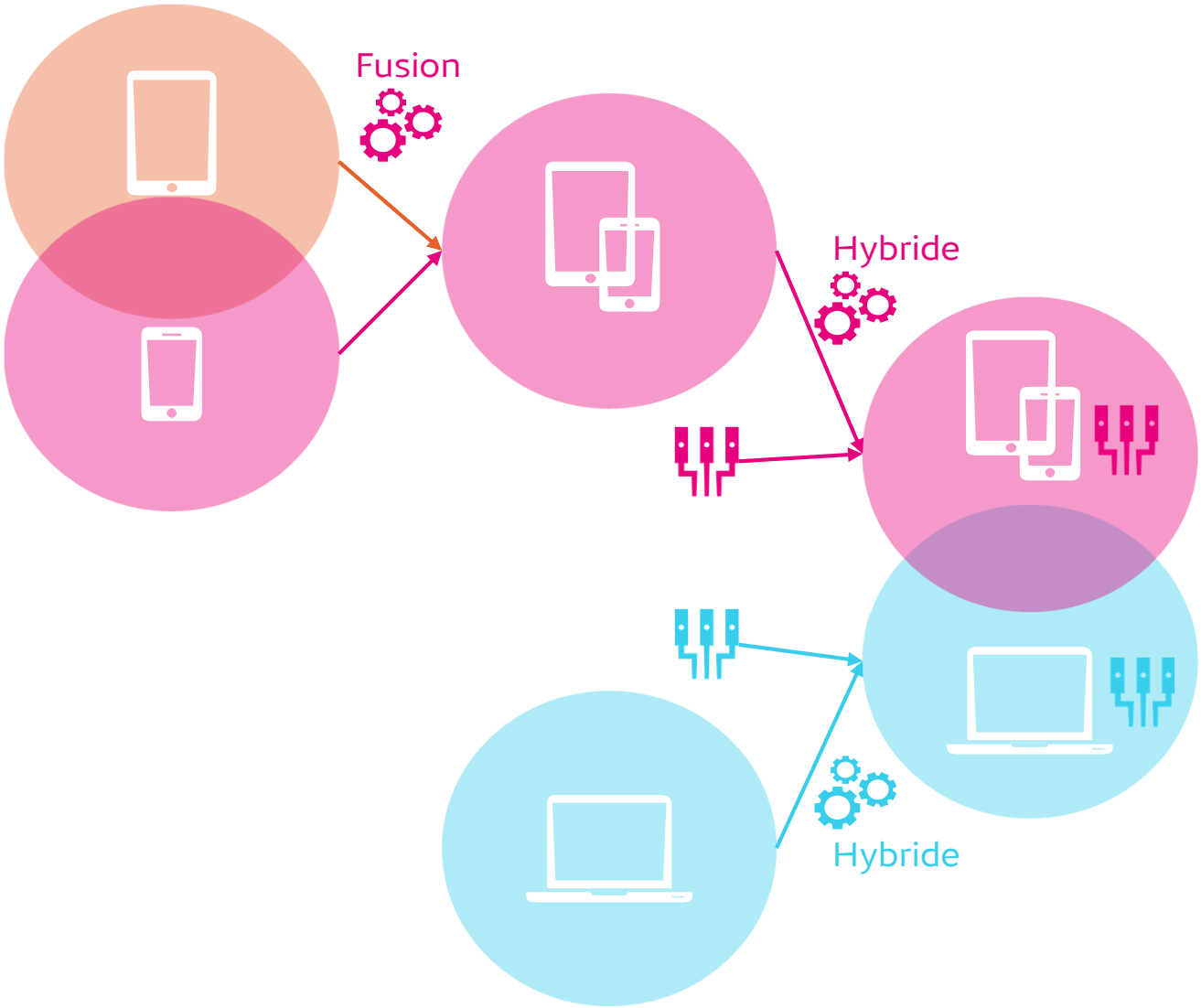


Objectifs fin 2018

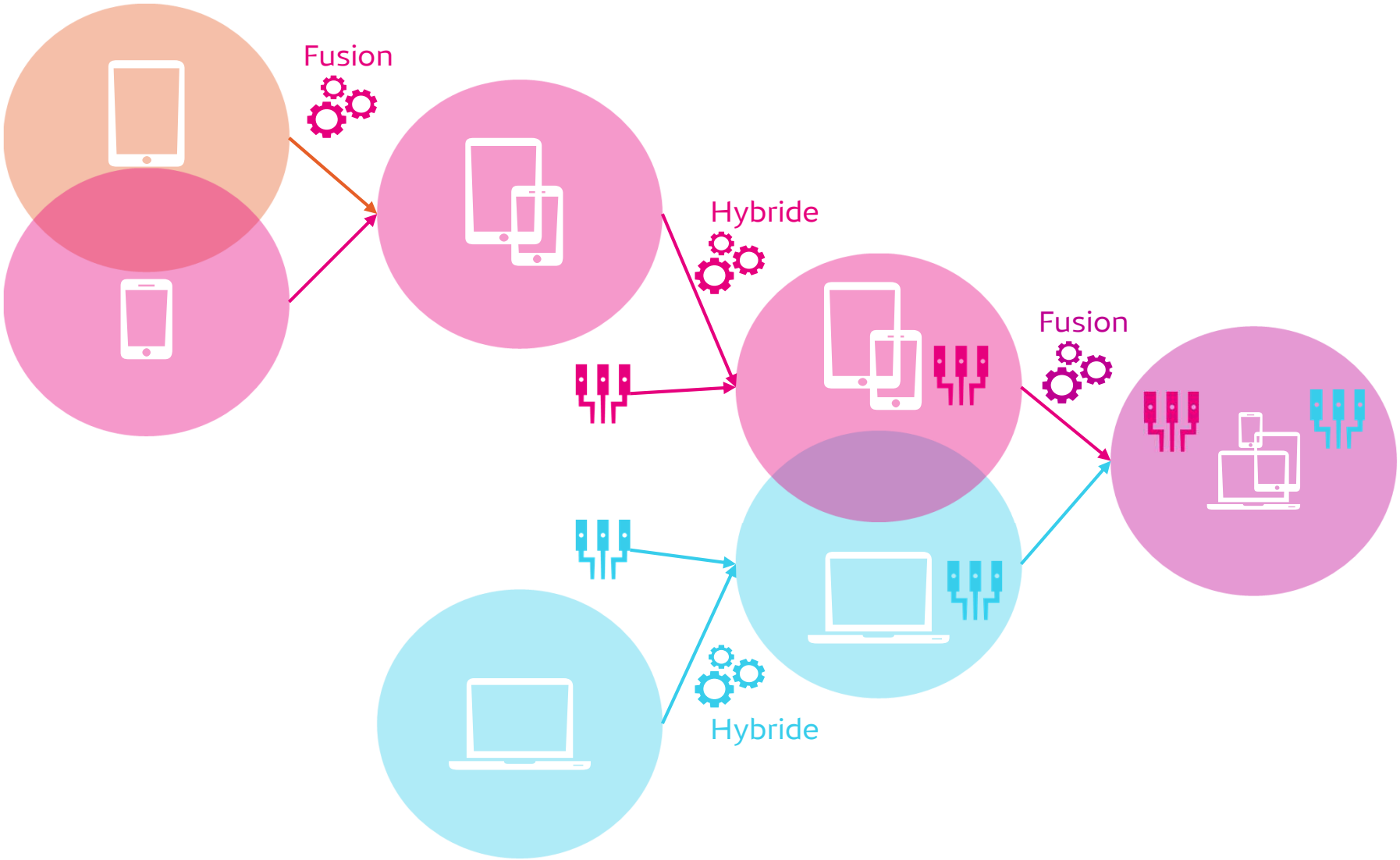
Mise en œuvre



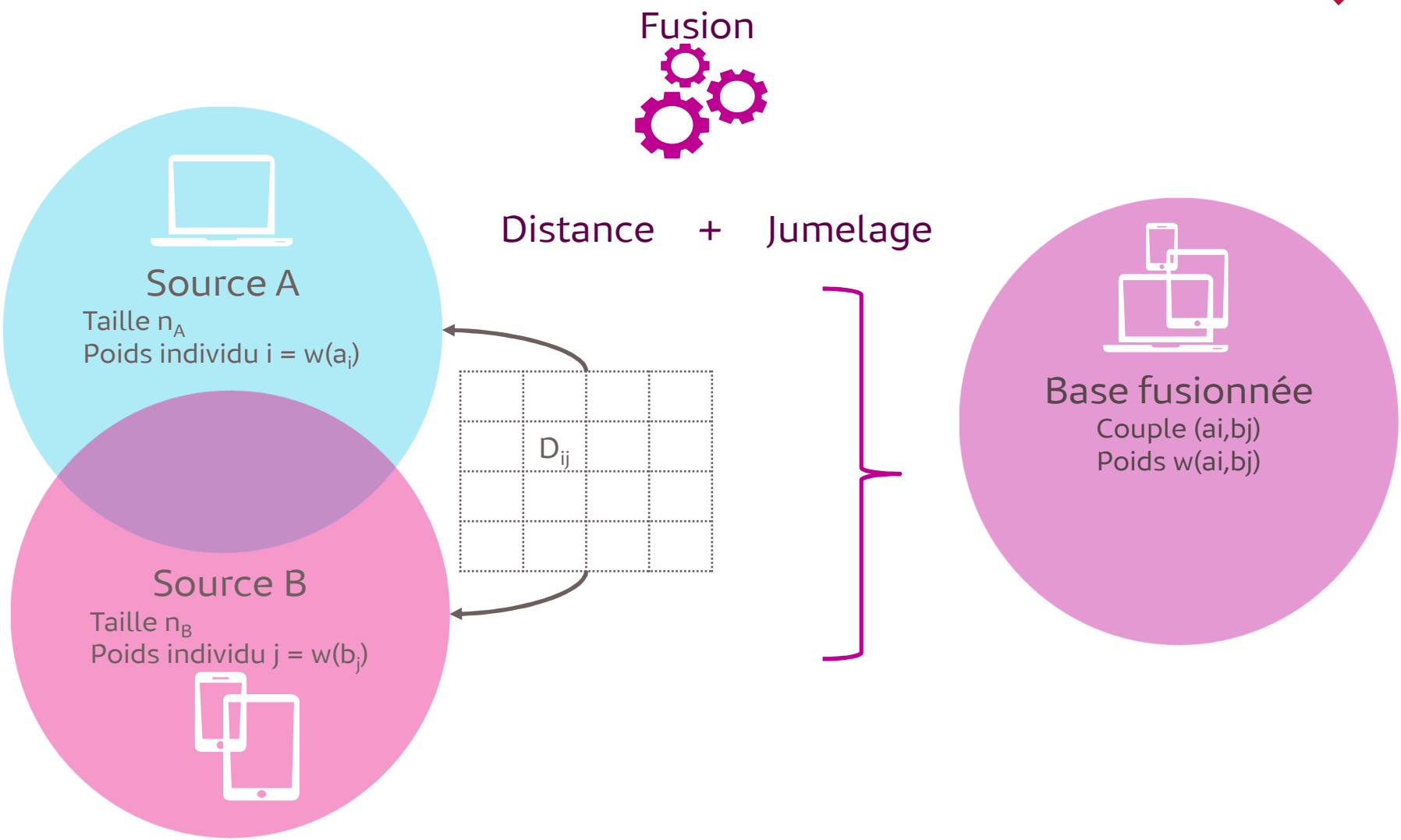
Mise en œuvre



Mise en œuvre



Fusion = Distance + Jumelage



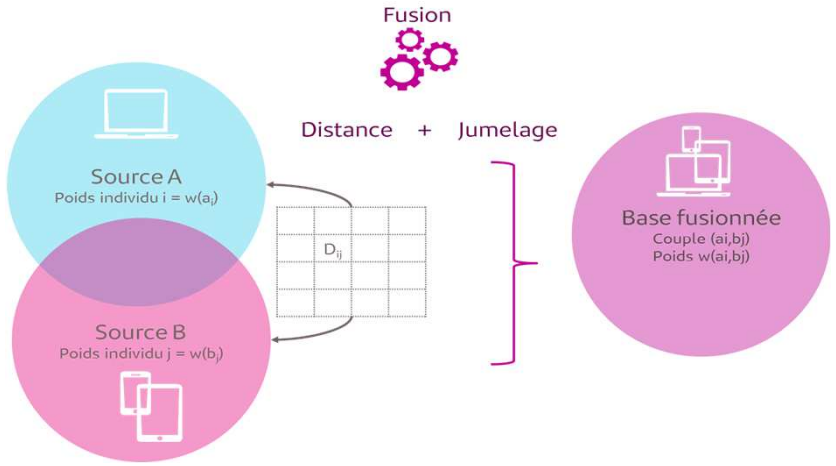
Jumelage sous contraintes

Objectifs :

- Conserver les résultats des 2 sources
- Conserver le « bénéfice » de l'hybridation

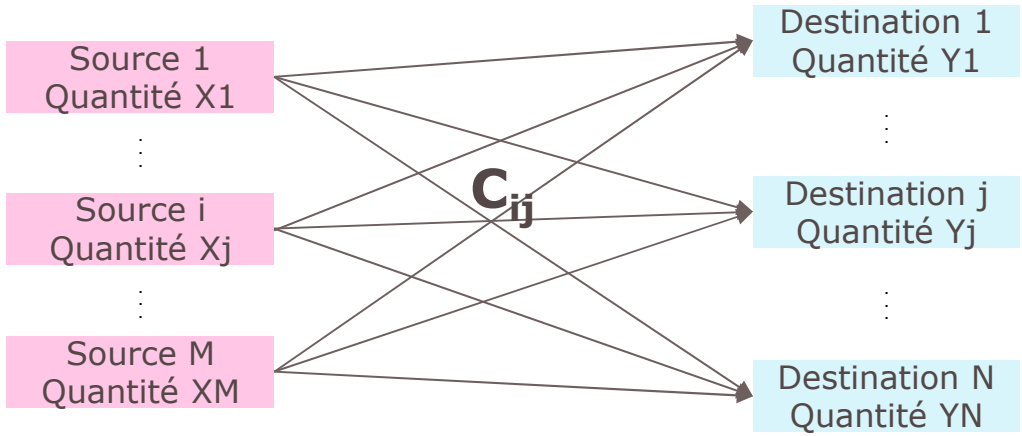
→Conserver les poids

- Minimiser le coût du jumelage
  - $Z = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} D_{ij} \times w(a_i, b_j)$
- Sous les contraintes de conservation des poids
  - $\sum_{i=1}^{n_A} w(a_i, b_j) = w(b_j)$
  - $\sum_{j=1}^{n_B} w(a_i, b_j) = w(a_i)$



Jumelage sous contraintes

Un problème de transport !



Hypothèse :
Offre = Demande

- Minimiser le coût du transport
  - $Z = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} C_{ij} \times w(a_i, b_j)$
- Sous les contraintes Offres et Demandes
  - $\sum_{i=1}^{n_A} w(a_i, b_j) = w(b_j)$
  - $\sum_{j=1}^{n_B} w(a_i, b_j) = w(a_j)$

→ Nombre de couple ≤ N + M -1

Références :
Mansi, S.G. (2011), A Study on Transportation Problem, Transshipment Problem, Assignment Problem and Supply Chain, Thèse de doctorat, Rajkot

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	50
Destination 2	3	8	7	40
Destination 3	4	5	2	60
Offre	50	65	35	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1				50
Destination 2				40
Destination 3				60
Reste	50	65	35	

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	50
Destination 2	3	8	7	40
Destination 3	4	5	2	60
Offre	50	65	35	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1			35	15
Destination 2				40
Destination 3				60
Reste	50	65	0	

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	15
Destination 2	3	8	7	40
Destination 3	4	5	2	60
Offre	50	65	0	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1			35	15
Destination 2	40			0
Destination 3				60
Reste	10	65	0	

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	15
Destination 2	3	8	7	0
Destination 3	4	5	2	60
Offre	10	65	0	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1		15	35	0
Destination 2	40			0
Destination 3				60
Reste	10	50	0	

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	0
Destination 2	3	8	7	0
Destination 3	4	5	2	60
Offre	10	50	0	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1		15	35	0
Destination 2	40			0
Destination 3	10			50
Reste	0	50	0	

Jumelage sous contraintes

Un problème de transport !

- Exemple d'algorithme « simple » : Least Cost Method

Coûts	Source 1	Source 2	Source 3	Demande
Destination 1	6	4	1	0
Destination 2	3	8	7	0
Destination 3	4	5	2	50
Offre	0	50	0	

Solution	Source 1	Source 2	Source 3	Reste
Destination 1		15	35	0
Destination 2	40			0
Destination 3	10	50		0
Reste	0	0	0	

Jumelage sous contraintes

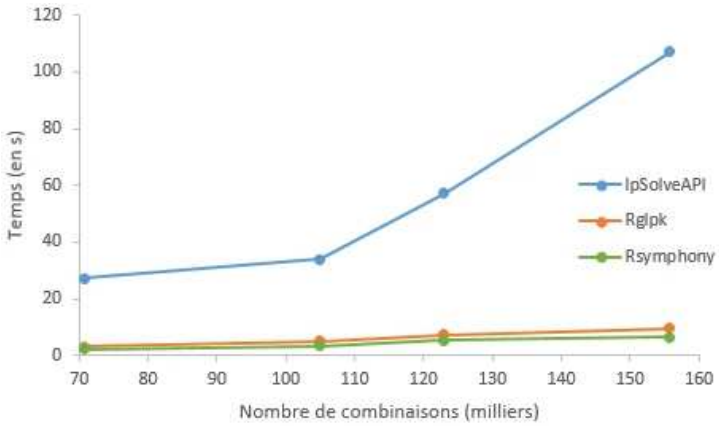
Un problème de transport !

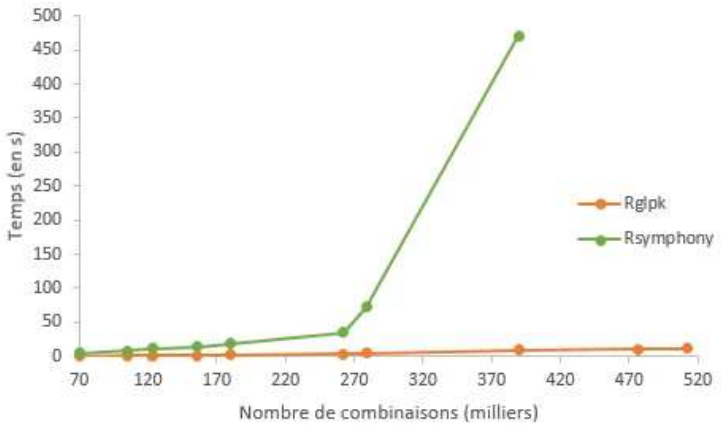
- MODI (MOdified DIstribution method)
  - Une solution initiale non optimale qui respecte les contraintes
  - Pour chaque couple non utilisé
    - Gain potentiel sur le coût total
    - Gain le plus important → Poids max
    - Baisse du poids des autres couples pour équilibrer

Jumelage sous contraintes

Un problème de transport !

- Optimisation linéaire R
  - Contrainte du problème de transport = solutions entières
  - Suppression de cette contrainte → Gain important dans les temps de calcul
  - 3 packages testés : IpSolveAPI, **Rglpk**, Rsymphony



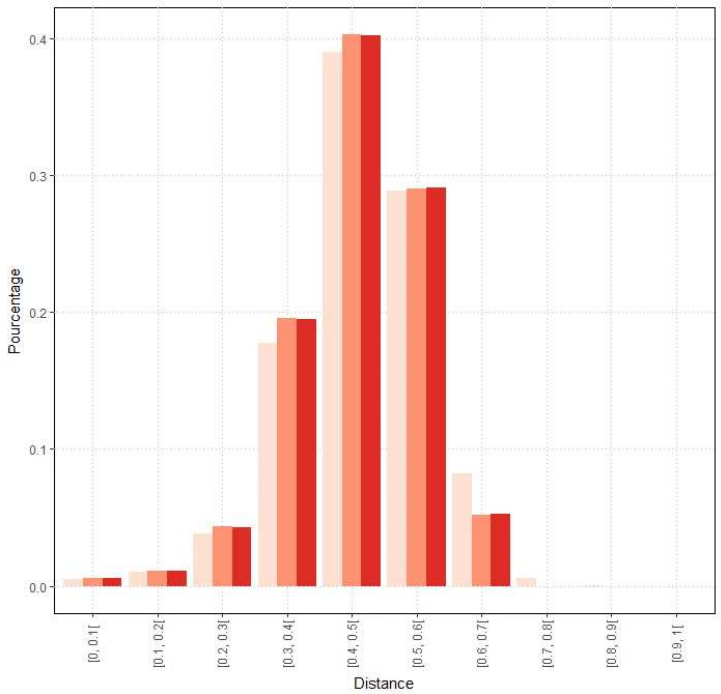


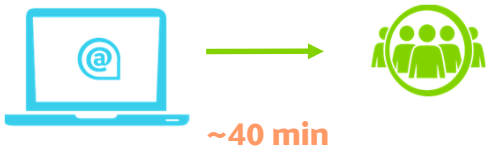
Jumelage sous contraintes

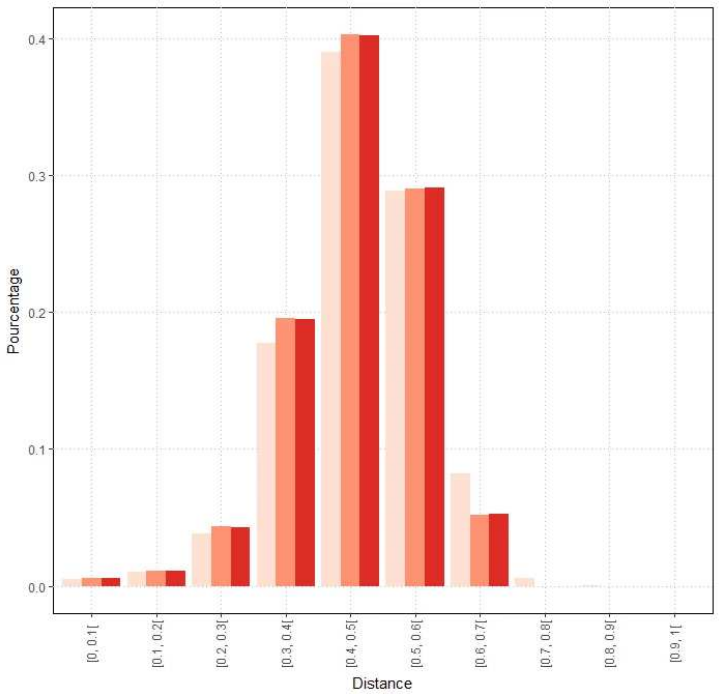
Comparaisons des solutions

- Distribution des distances

Internet Global 1ère génération
Optimisation linéaire R - Rglpk
MODI







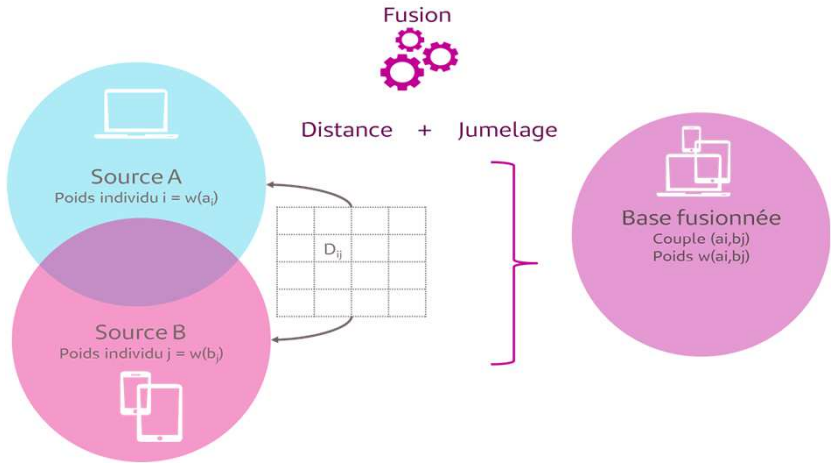
Distance

Objectifs :

- Utiliser la partie du panel pour laquelle la mesure est complète

- Transformation procustèenne
  - Rapprocher le + possible 2 nuages
  - En conservant les relations entre les points

- Sur la base des communs
  - Transformations optimales
  - Translation, homothétie, rotation, symétrie

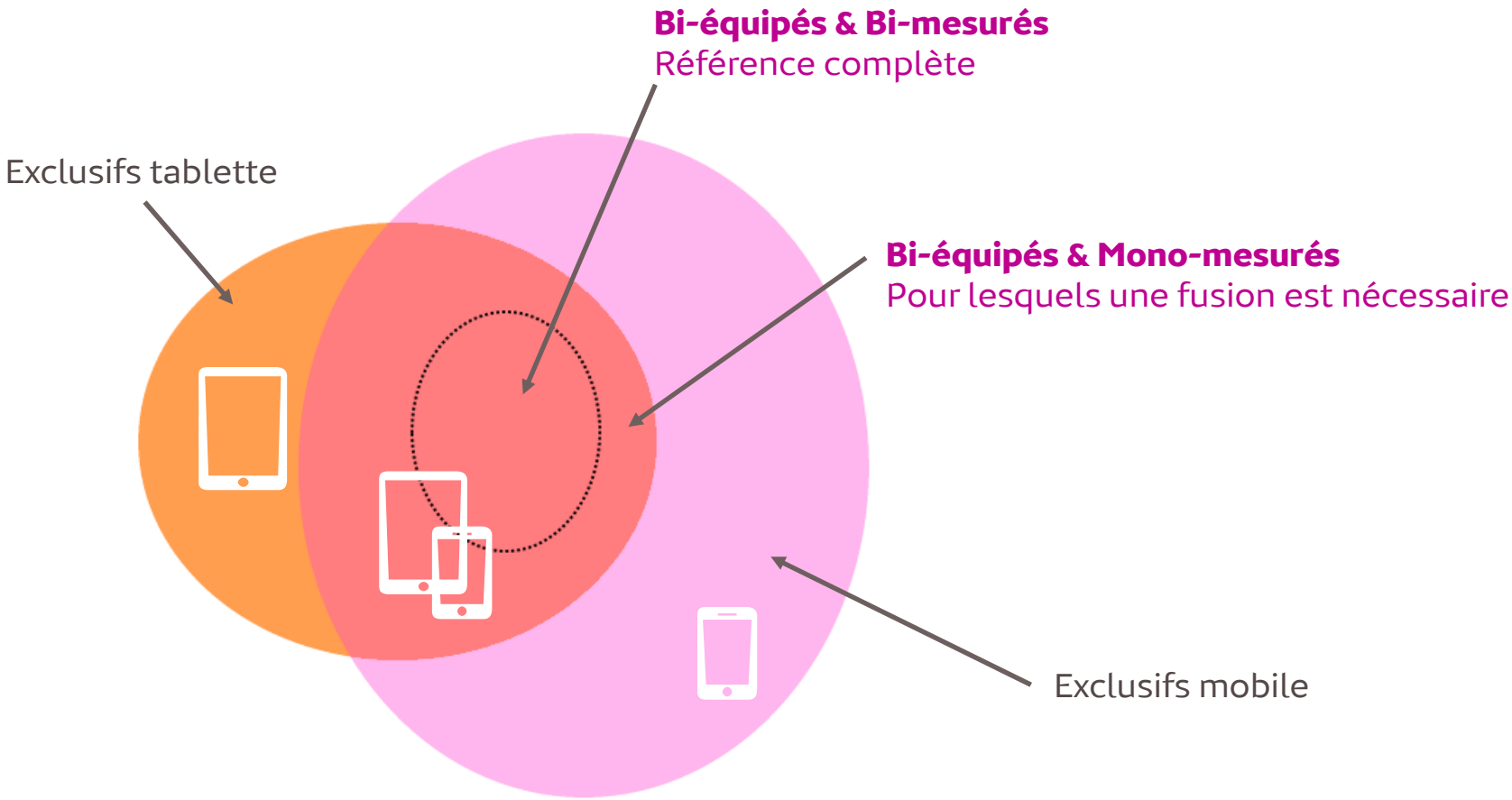


Références :

Borg I., Groenen P., Modern Multidimensional Scaling, Theory and Applications, 2005.

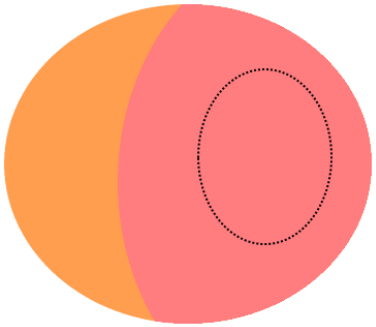
Distance

Illustration sur le panel Mobile - Tablette

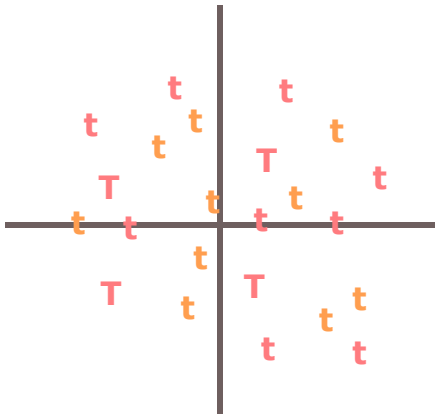


Distance

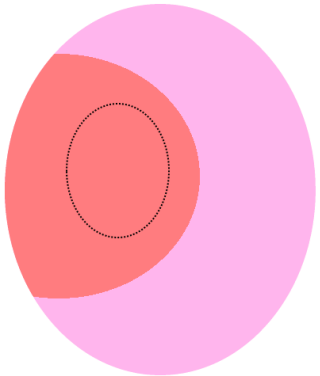
Illustration sur le panel Mobile - Tablette



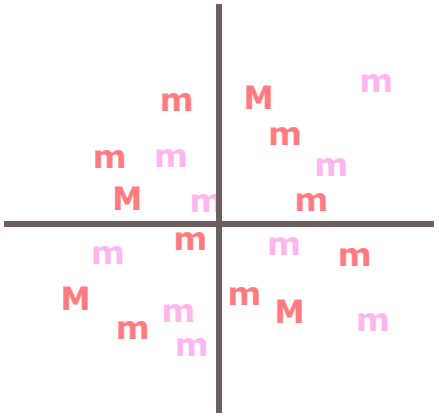
ACP



- **t** panéliste mesuré uniquement sur tablette
- **T** panéliste mesuré sur tablette **ET** mobile



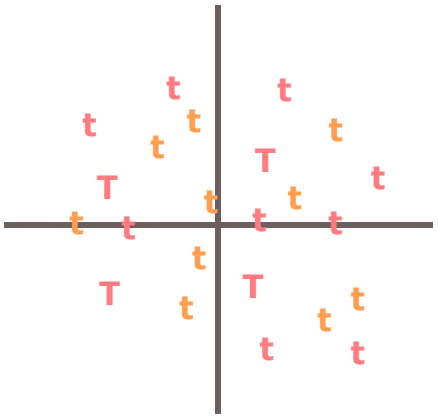
ACP

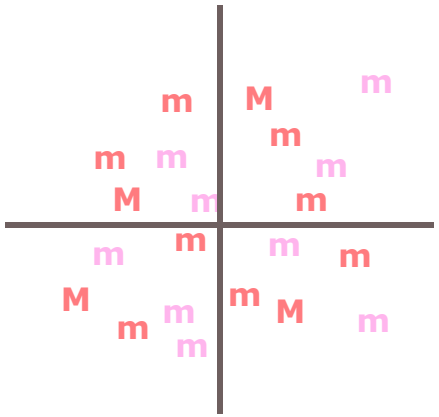


- **m** panéliste mesuré uniquement sur mobile
- **M** panéliste mesuré sur tablette **ET** mobile

Distance

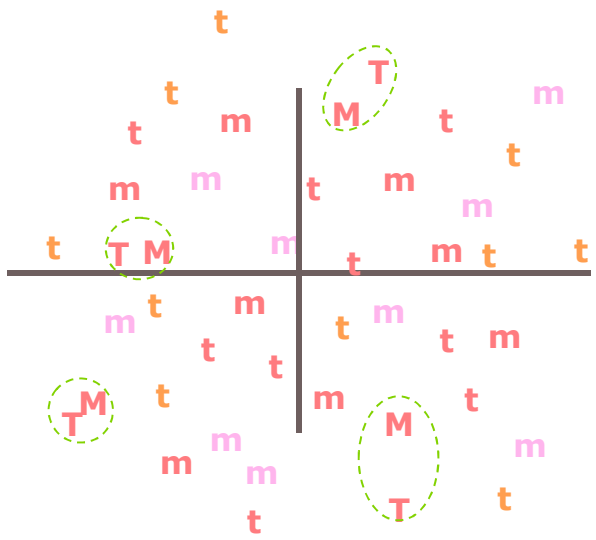
Illustration sur le panel Mobile - Tablette





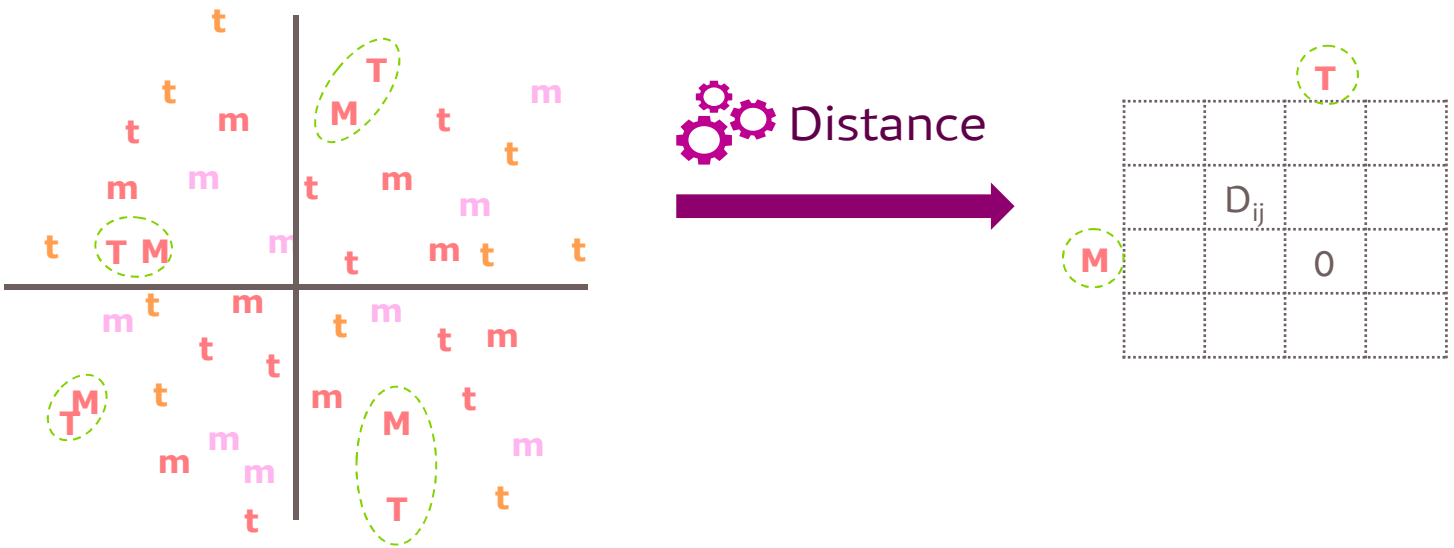
Transformation procustéenne

- Package R **MCMCpack**

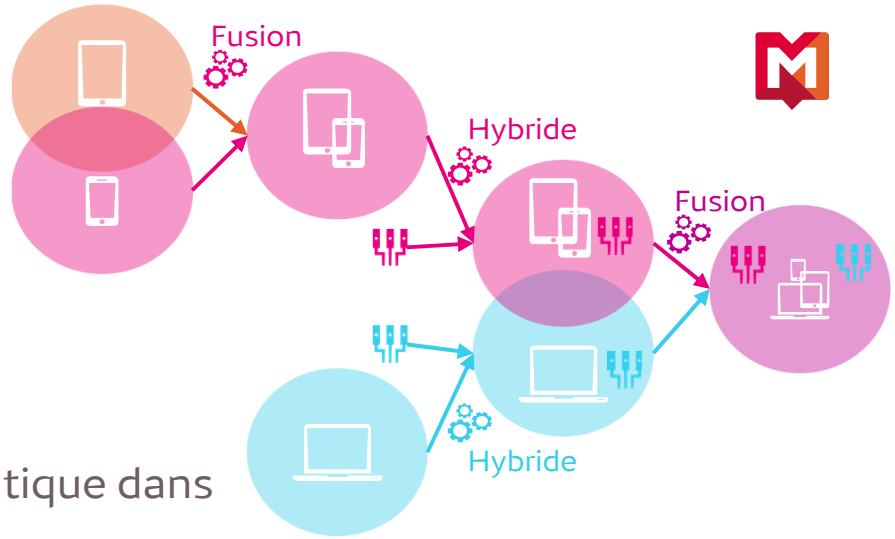


Distance

Illustration sur le panel Mobile - Tablette



Quelques détails pratiques

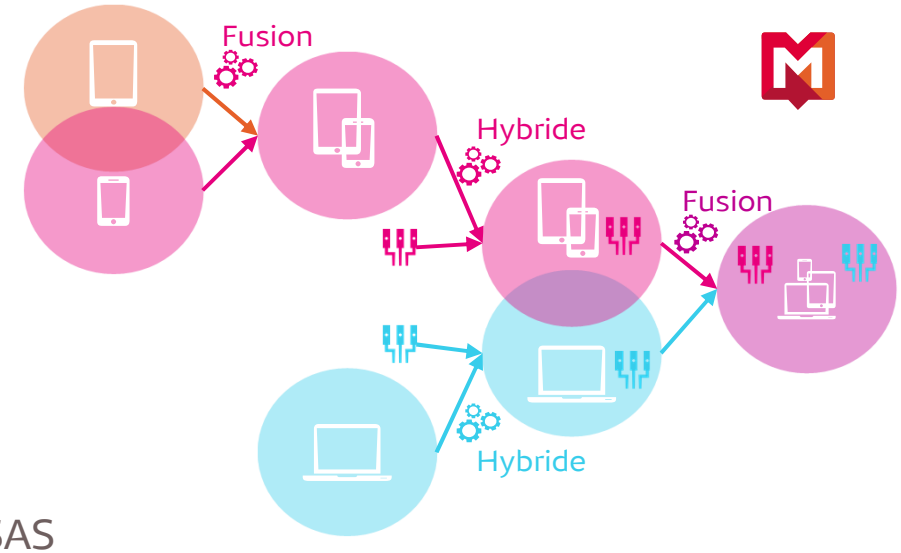


Stratification

- Le poids d'une même strate doit être identique dans les 2 sources avant fusion
- En particulier sur l'équipement :
  - Le poids des bi-équipés ordinateur-mobile doit être le même au sein du panel ordinateur et du panel mobilité

- Pour les étapes de fusion, le traitement des strates peut être parallélisé

Quelques détails pratiques



R sur le serveur SAS

- Utilisation de package R pour les fusions
- Chaînes de traitement historiques sous SAS

- Exécution du code R via la PROC IML de SAS

Merci pour votre attention