Présentation de l’audit

Présentation de Data Checking

Objectif :

Connaître la qualité des informations socio démographiques présentes dans des bases de données client (éditeurs, opérateurs, annonceurs)

Méthodologie :

Croisement des informations :

- D'une base client
- Avec la base de référence de Médiamétrie (300 000 contacts qualifiés d'individus interrogés par téléphone, dans le cadre d'enquêtes ou de la constitution des panels)

Périmètre :

Jointure entre les deux bases grâce aux adresses emails.
A date, comparaison de 2 variables : l'âge et le sexe.
À terme, l'outil prévoit aussi de comparer les CSP.

Indicateurs :

Médiamétrie met à disposition deux types d'indicateurs :

- **Indice de concordance** permettant d'analyser le pourcentage d'informations communes entre la base de référence et la base client
- **Indice Kappa**, indicateur de qualité de concordance, intégrant la probabilité d'une réponse commune à une variable en fonction du nombre de modalités proposées

Etapes du déploiement de la solution

Sources

- Liste précise de sources
  - Adresses issues des panels de Médiamétrie : Weball, Mediamat / PaME
  - BABU : 126 000, OUI-TSM, 50 000 TV, Autres (Grille d'Été, TV Locales, Médialocales, ..)

- Contact exclusivement téléphonique et recueil des informations très majoritairement par téléphone

- Lors de la prise d'adresses mails, demander explicitement s'il s'agit d'une adresse personnelle ou si cette adresse est utilisée par d'autres personnes, pour comparer uniquement les adresses personnelles.
- Collecter l'exhaustivité des adresses personnelles

Etapes du déploiement de la solution

Actualisation et traitement

- Le chiffrage des données et la sécurisation du stockage
- Actualisation trimestrielle de la base de référence et processus de création de la base de référence clairement défini
- Récence des données de sexe et d'âge (5 ans maximum)
- Intention de comparer les données de CSP sur la base de données plus récentes que celles sur l'âge et le genre, en raison de la plus grande instabilité de cette variable
  NB : la base de référence sera donc significativement plus réduite pour les CSP

- Améliorer la récolte de la date de dernière actualisation

Analyse de la fiabilité de la base

- Volume conséquent d'adresses disponibles par âge à partir de 15 ans
- Plus de 90% des adresses proviennent d'études auditées par le

	Étude	*Poids dans la base*	
Base panel	Mediamat	Entre 5% et 6%	Étude auditée par le CESP
	PaMe		Non auditée
	WebAll	Entre 37% et 47%	Étude auditée par le CESP
BABU	126 000	Entre 44% et 51%	Étude auditée par le CESP
	OUI	3%	Étude auditée par le CESP
	50 000 TV	<1%	Non auditée
	Autres sources	Entre 0% et 3%	Non auditée

- Faible taux d'erreur :
  - Sexe : 1 erreur sur 1201 cas (0,1%)
  - Année de naissance : 3 erreurs sur 1201 cas (0,2%)
  - CSP : 44 erreurs sur 1201 cas (4%)
    Recueil complexe et méthode du client potentiellement différente de celle utilisée par Médiamétrie

- Stabilité des résultats :
  - Par source : WebAll, Mediamat / PaME, BABU
  - Dans le temps : T2 2018, T3 2018 et T4 2018

Evaluation de la méthodologie de concordance
Indicateurs utilisés

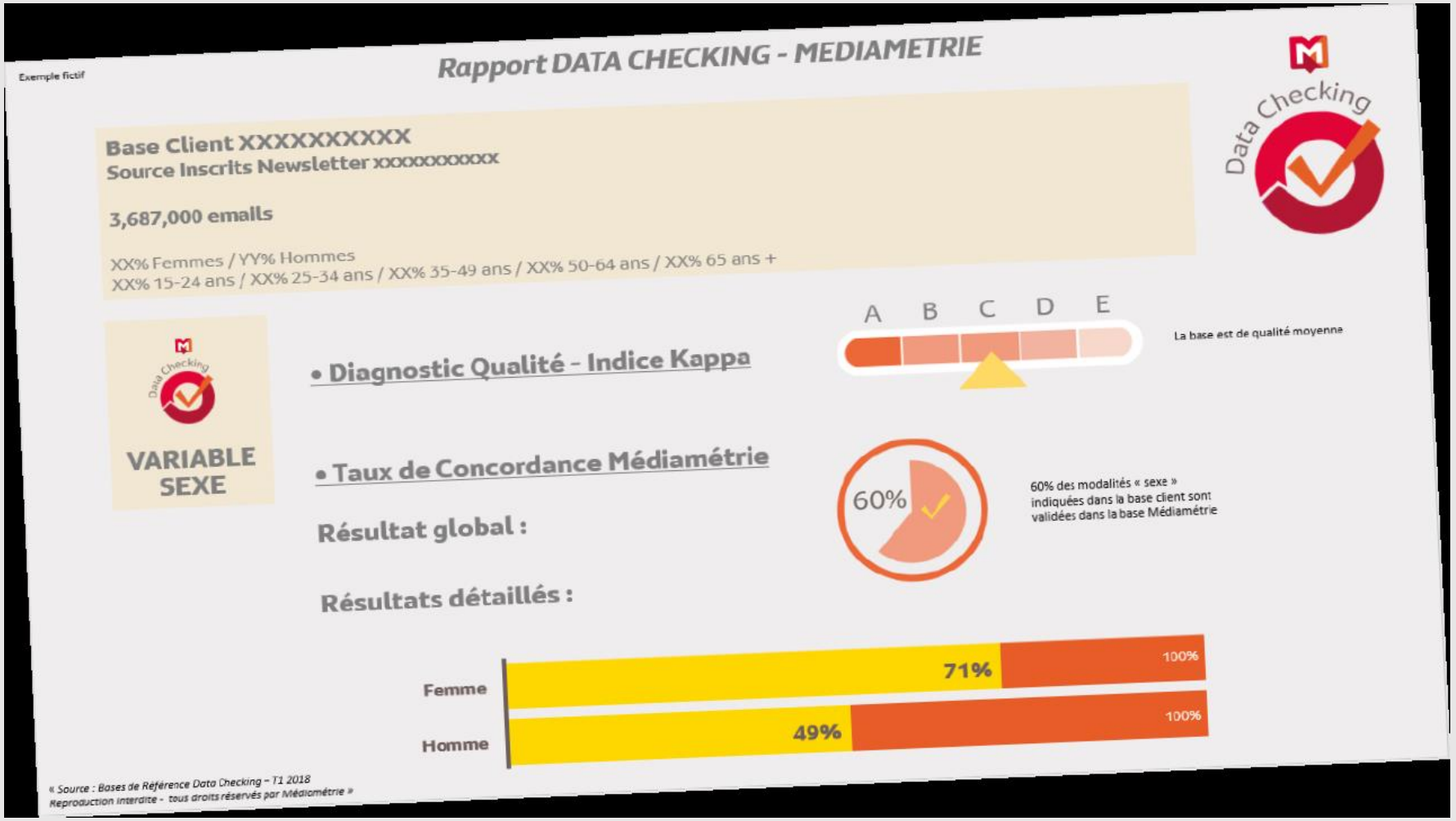
- Indicateurs de concordance simples, précis, adaptés:
  - Concordance stricte (Sexe, âge détaillé, tranches d'âge)
  - Concordance détaillée : part des individus bien classés par modalité de variables binaires (Homme versus femme, tranche d'âge versus les autres)
  - Concordance pondérée (âge détaillé)

Ecart d'âge	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pondération	1	0,93	0,87	0,80	0,73	0,67	0,60	0,53	0,47	0,40	0,33	0,27	0,20	0,13	0,07	0,00

- Limitation de l'amplitude à 10 ans maximum

- Indice Kappa:
- Permet d'étalonner la qualité de la base étudiée par rapport à une norme de 0 à 1
  - Kappa classique : pour des variables binaires (sexe)
  - Kappa normalisé : pour des variables non-binaires (tranches d'âge)

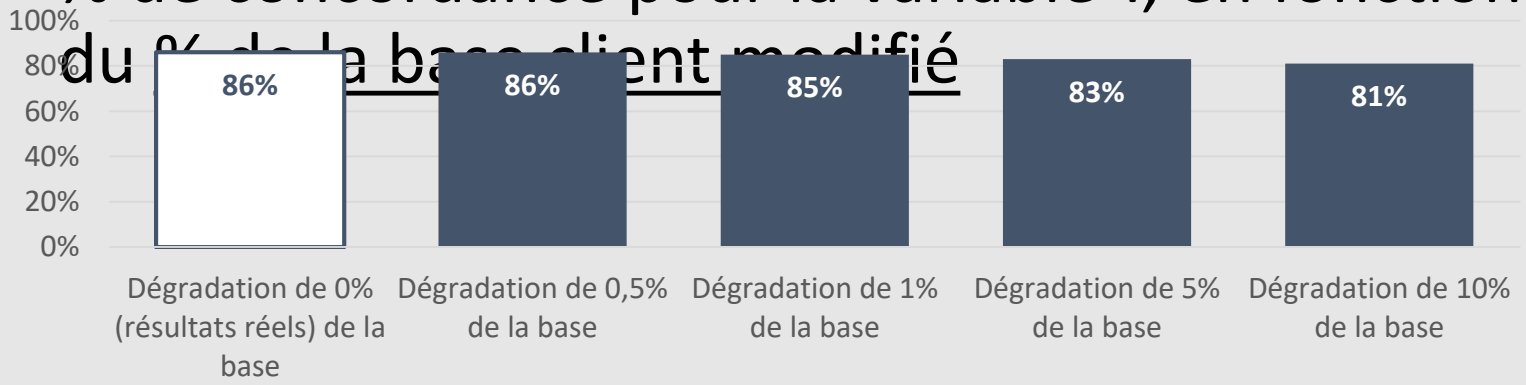
Evaluation de la méthodologie de concordance
Exemple fictif de livrable



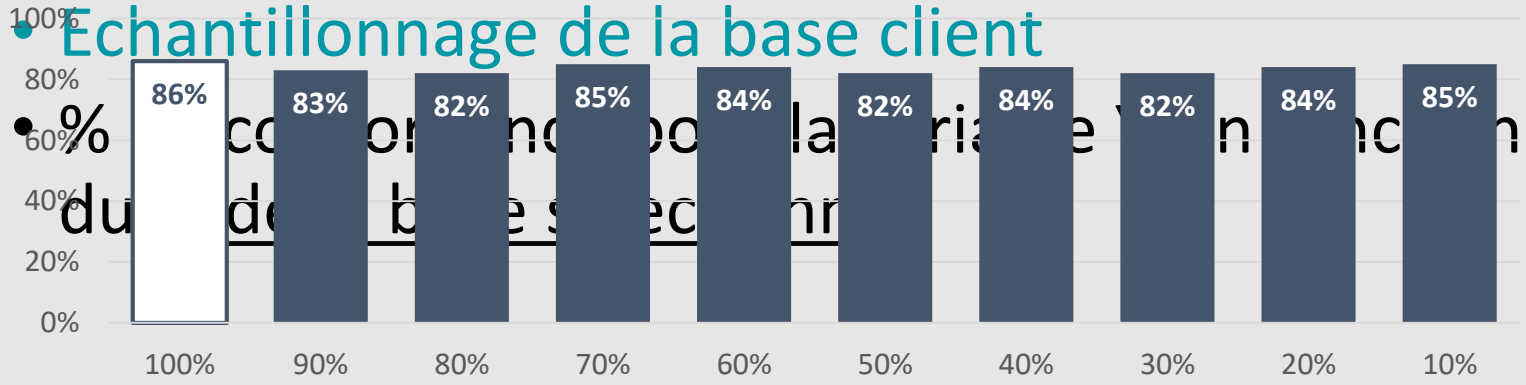
Evaluation de la méthodologie de concordance

Test de sensibilité de l'échantillon

- Dégradation volontaire de la base client
- % de concordance pour la variable Y, en fonction du % de la base client modifié



- Echantillonnage de la base client
- % de concordance pour la variable Y, en fonction du % de la base sélectionnée



Base client : 302 388 emails dont 2 454 communs avec la base de référence

Restitution client

0,9 – 1,0	Qualité presque parfaite	A+
0,8 – 0,9	Excellente qualité	A-
0,7 – 0,8	Très bonne qualité	B+
0,6 – 0,7	Bonne qualité	B-
0,5 – 0,6	Qualité moyenne	C+
0,4 – 0,5	Qualité très moyenne	C-
0,3 – 0,4	Qualité faible	D+
0,2 – 0,3	Qualité très faible	D-
0,1 – 0,2	Mauvaise qualité	E+
0,0 – 0,1	Très mauvaise qualité	E-

- L'échelle Ka[illegible]dre en compte les valeurs de κ négatives :
  - Lorsque κ= 1, la concordance est parfaite.
  - Lorsque κ= 0, la concordance n'est pas meilleure que le hasard.
  - Lorsque κ<0, la concordance est moins bonne que le hasard ou nulle
- La taille de l'échantillon doit être rappelée sur toutes les pages comportant des résultats et non uniquement sur la première page.
- Si l'échantillon induit un intervalle de confiance ne permettant pas d'attribuer une graduation unique de qualité de la base, le CESP recommande de communiquer la plage de notes d'évaluation Kappa comprises dans l'intervalle statistique.

Synthèse de l'audit (1/2)

Principales parties	Items	Conclusions du CESP		
		Satisfaisant (*)	Amélioration suggérée (**)	Changement nécessaire (***)
Etapes du déploiement de la solution	Sources	✔		
	Univers			⚠ Comparer uniquement les adresses strictement personnelles et collecter les différentes adresses personnelles de la personne interrogée pour la comparaison
	Actualisation	✔	Améliorer la récolte de la date de dernière actualisation	
	Traitement	✔		
Analyse de la fiabilité de la base	Recueil	✔		
	Structure	✔		
	Analyse par source	✔		
	Stabilité	✔		

(*) Satisfaisant : Correspond aux bonnes pratiques
(**) Amélioration suggérée : Recommandation pour optimiser l'outil
(***) Changement necessaire : Nécessité d'adapter l'outil pour correspondre aux bonnes pratiques

cesp audited

Synthèse de l'audit (2/2)

Principales parties	Items	Conclusions du CESP		
		Satisfaisant (*)	Amélioration suggérée (**)	Changement nécessaire (***)
Evaluation de la méthodologique de concordance	Taille de l'échantillon utilisé	✔		
	Concordance stricte, détaillée et pondérée	✔		
	Kappa classique et normalisé	✔	Limiter l'amplitude maximum du Kappa pondéré à 10 ans	
	Recalcul des résultats	✔		
Restitution client	Résultats restitués		Rappeler le nombre d'emails exploitables pour la comparaison des bases sur toutes les pages, ainsi que les intervalles de confiance	Intégrer l'intervalle statistique du Kappa dans la communication du niveau de qualité de la base client
	Echelle Kappa	✔	Prendre en compte les valeurs de kappa négatives	

(*) Satisfaisant : Correspond aux bonnes pratiques
(**) Amélioration suggérée : Recommandation pour optimiser l'outil
(***) Changement necessaire : Nécessité d'adapter l'outil pour correspondre aux bonnes pratiques

cesp audited

Data Checking

Collège Data CESP

20 Juin 2019

Mediametrie

Pourquoi un audit ?

Dans ce marché où tout le monde cherche à valoriser ses données, toutes ne se valent pas.

« Il n'y a jamais eu autant de défiance et de suspicion par rapport à la donnée, à la façon dont elle est collectée, partagée et utilisée »

Jean-Luc Chetrit, Président UDA.

« Sur les projets data, il y a encore beaucoup de déclaratif, peu de contrôle, pas d'organisme tiers »

Capucine Pierard, DGA Havas Media & Chief Data Officer.

Objectif Data Checking → **plus de transparence sur la qualité des données utilisées par le marché**.

Data Checking a été récompensé aux derniers Trophées Etudes & Innovations Catégorie Data Intelligence, Trophée argent

→ Mise en place de l'audit CESP

S'auto-appliquer le principe de transparence en auditant notre base de référence et les indicateurs de qualité du Data Checking

Suite aux recommandations du CESP

- **12 sur 14 des points étudiés reconnus satisfaisants**

- **Recommandations concernant la collecte des emails :**

  Identifier lors de la passation des différents questionnaires téléphoniques si l'adresse mail communiquée est strictement personnelle et recueillir l'ensemble des adresses mail personnelles de l'interviewé

  → Cette recommandation sera étudiée avec nos équipes panels dès le 2ème semestre 2019.

- **Recommandations concernant la communication des évaluations des bases aux clients :**

  Limiter l'amplitude de l'indicateur de concordance pondérée à 10 ans et prendre en compte les valeurs négatives du Kappa

  Communiquer les intervalles de confiance des indicateurs de qualité et ajuster la notation au sein de l'échelle de résultat

  Communiquer le nombre d'emails sur lequel est fondée la comparaison sur toutes les pages du rapport et non uniquement sur la première page

  → Cette recommandation sera mise en œuvre dès les prochaines restitutions clients.

Quelles sont les retours d'expérience de nos clients ?

Disposer

du «label» Médiamétrie, gage de qualité et de confiance pour une utilisation commerciale

Hiérarchiser

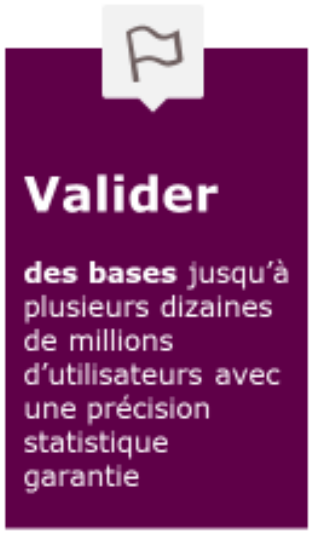
leurs bases de données internes en fonction de leur qualité et de leur source, y compris probabilistes.

Améliorer

leurs datas en supprimant leurs bases de données les moins pertinentes

Contrôler

la qualité de bases externes issues de partenaires, fournisseurs…



Valider

des bases jusqu'à plusieurs dizaines de millions d'utilisateurs avec une précision statistique garantie

Prochaines étapes :

La solution Data Checking sur la base des emails et des devices Id est opérationnelle

R&D en cours pour un Data Checking sur la base des cookies courant S2 2019