OCTOPIZE
MIMETHIK DATA

#MACHINE LEARNING
#CONFORMITÉ RGPD
#USAGES SECONDAIRES

CAS CLIENT

Anonymiser des données RH pour les conserver sans limite et entraîner un modèle de Machine Learning

TOP TURNOVER / AYMAX

CHALLENGE

- Constituer une base de données multi-secteurs / multi-métiers
- Conserver les données sans limite de temps
- Alimenter un modèles de prédiction RH, tout en respectant la vie privée des collaborateurs

70%

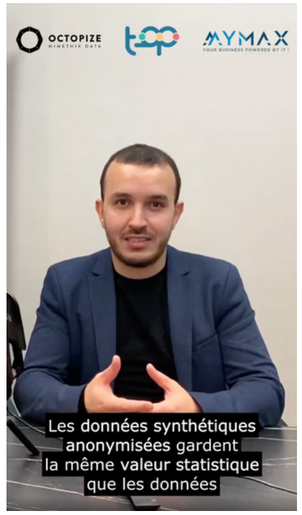
Précision du modèle prédictif (vs 68% sans anonymisation)

TÉMOIGNAGE

"La solution avatar d'Octopize nous a permis de maintenir les mêmes performances du modèle prédictif IA, et même mieux..."

"Les données anonymes avatar conservent les mêmes performances statistiques que les données originales et maintiennent les mêmes corrélations."

- Amine Menacer, Dr IA, CTO, CEO @Top Turnover



RETOUR SUR INVESTISSEMENT

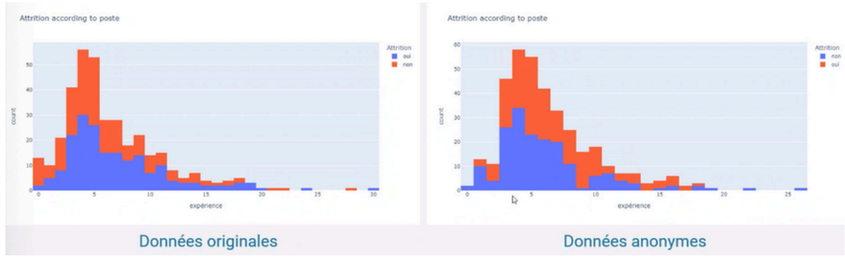
Objectif : prédire et anticiper les risques de démission avec les données RH

- **68%** de bonne prédiction avec le modèle entraîné avec les données d'origine
- **70%** de bonne prédiction avec le modèle entraîné avec les données anonymes
  - les données anonymes conservent la capacité prédictive des données d'origine
  - **2%** d'amélioration sur la précision globale du modèle prédictif (échantillon de 200 collaborateurs)
- Meilleure répartition des données
- Conservation sans limite de temps

Top turnover peut utiliser les données anonymes en production.

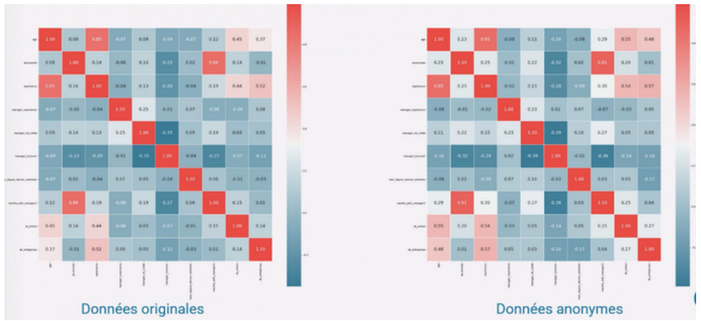
MAINTIEN DE LA QUALITÉ & DE L'UTILITÉ

- Comparaison de distribution univariée (expérience du collaborateur)

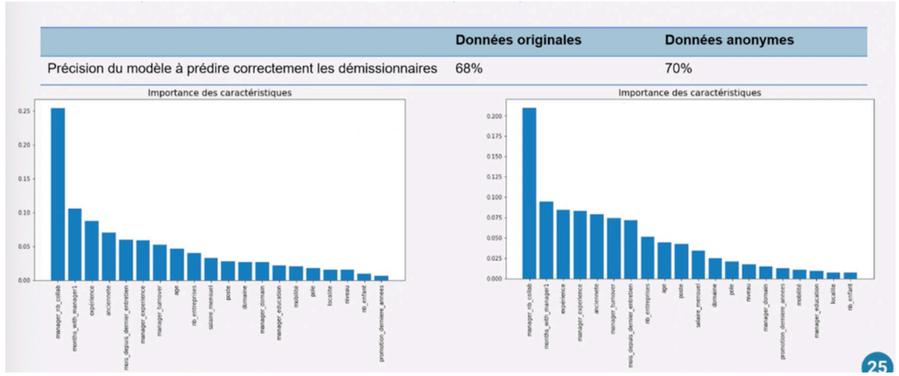


Les données anonymes conservent les distributions originales.

- Comparaison de distribution bivariée (corrélation)



Les données anonymes conservent les corrélations entre les variables.



L'interprétabilité du modèle est conservée.

OCTOPIZE
MIMETHIK DATA

#PROFILAGE
#USAGES SECONDAIRES

CAS CLIENT

Anonymiser des données de déplacements pour optimiser les services

ACTEUR DE LA MOBILITÉ

CHALLENGE

- Collecter massivement des données de géolocalisation.
- Utiliser de façon secondaire et éthique la donnée.
- Valoriser et ré-exploiter des données pour améliorer les services proposés et élaborer différents profils clients.
- Améliorer des services sans nécessiter de nouveau consentement.

1 million

de déplacements anonymisés.

TÉMOIGNAGE

"Grâce à la solution avatar d'Octopize, nous préservons les mêmes propriétés statistiques tout en gérant efficacement la complexité de nos analyses de données. Leur technologie est particulièrement innovante et inclut un rapport d'anonymisation que nous pouvons présenter à la CNIL si nécessaire. C'est un outil complet qui nous est utile dans plusieurs de nos services."

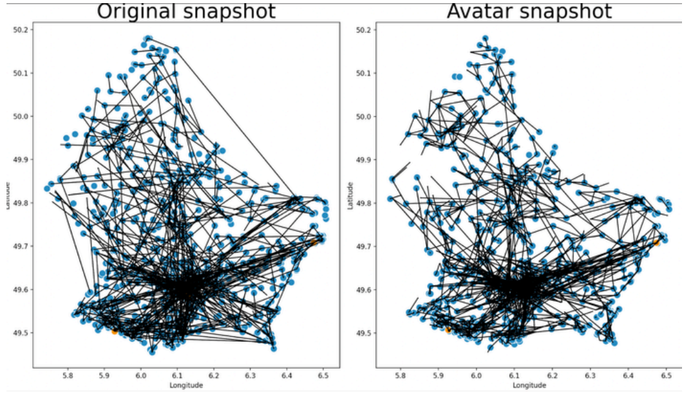
RETOUR SUR INVESTISSEMENT

- Reproductibilité (même **qualité d'analyse**) permise par les données avatar.
- Utilisation des informations issues des données avatar **sans nécessiter de nouveau consentement**.
- **Profilage** des clients avec des données anonymes exploité pour **optimiser les services** proposés.

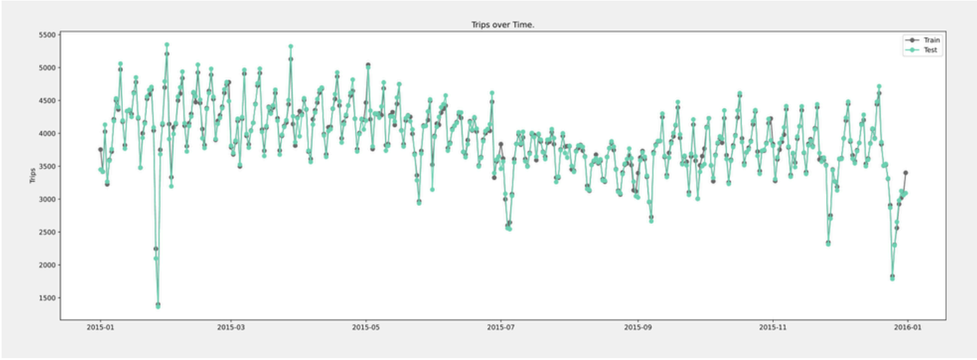
MAINTIEN DE LA QUALITÉ & DE L'UTILITÉ

- La structure originale des données est préservée.

- L'information géographique est préservée.



- L'anonymisation avec la méthode avatar garantit l'utilité des données.



OCTOPIZE
MIMETHIK DATA

#MACHINE LEARNING
#PROFILAGE
#USAGES SECONDAIRES

CAS CLIENT

Anonymiser des données cliniques pour optimiser les parcours de soin Covid

APHP / ECHOPEN

CHALLENGE

- **Évaluer la sévérité**, personnaliser la prise en charge, et soulager les établissements de santé en pleine crise sanitaire à partir des données cliniques de patients soupçonnés de Covid.

- Développer un **algorithme de Machine Learning** pour prédire le niveau de sévérité des patients, améliorant ainsi la qualité des soins, tout en garantissant la **confidentialité** des patients.

TÉMOIGNAGE

“Convaincu par les avatars, à l’issue d’expérimentations dans le cadre d’Epidemium et d’Echopen, j’ai proposé à Octopize de venir présenter la technologie des avatars au groupe de travail mensuel IA et Santé de l’Académie de Médecine."

- Olivier de Fresnoye, Directeur Général et Co-fondateur @echOpen

RETOUR SUR INVESTISSEMENT

- **Conservation des performances de prédiction du modèle** avec de très bons résultats

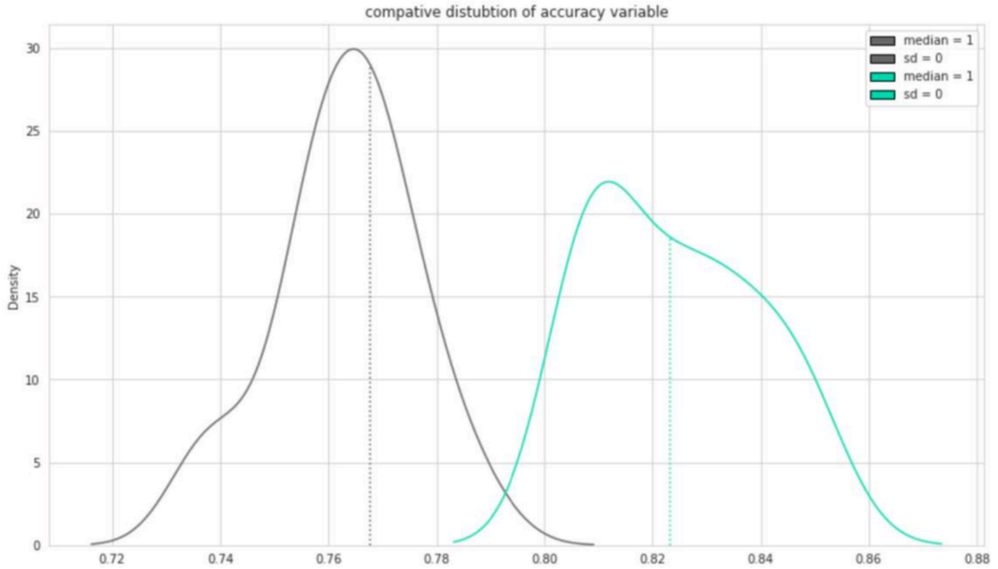
- Optimisation de la **gestion du flux** des urgences dans un cas de crise de santé publique (pandémie Covid 19)

- Amélioration et accélération de la **recherche**

MISE EN PLACE

Service d'anonymisation : Octopize s'est déplacé sur les lieux de l'APHP pour effectuer l'anonymisation de données personnelles directement sur ses serveurs.
Durée : 2 jours.

MAINTIEN DE LA QUALITÉ & DE L’UTILITÉ



Le modèle de prédiction entraîné sur des données synthétiques avatars est en moyenne **7% plus performan**t que le modèle de prédiction entraîné sur les données d’origine.

OCTOPIZE
MIMETHIK DATA

#USAGES SECONDAIRES
#RECHERCHE
#PARTAGE

CAS CLIENT

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Epidemium

Accélérer la recherche contre le cancer grâce au partage de données anonymisées

APHP / EPIDEMIUM

CHALLENGE

- Anonymiser le jeu de données (KORL) incluant des **données cliniques** et des images histologiques issues de biopsies de patients rigoureusement sélectionnés
- Faire progresser la **recherche** pour lutter contre le cancer en organisant des "data challenges"
- Faciliter la familiarisation du personnel à l'**analyse de données de vie réelle** en utilisant des données avatar
- Associer une donnée avatar avec une **image**

TÉMOIGNAGE

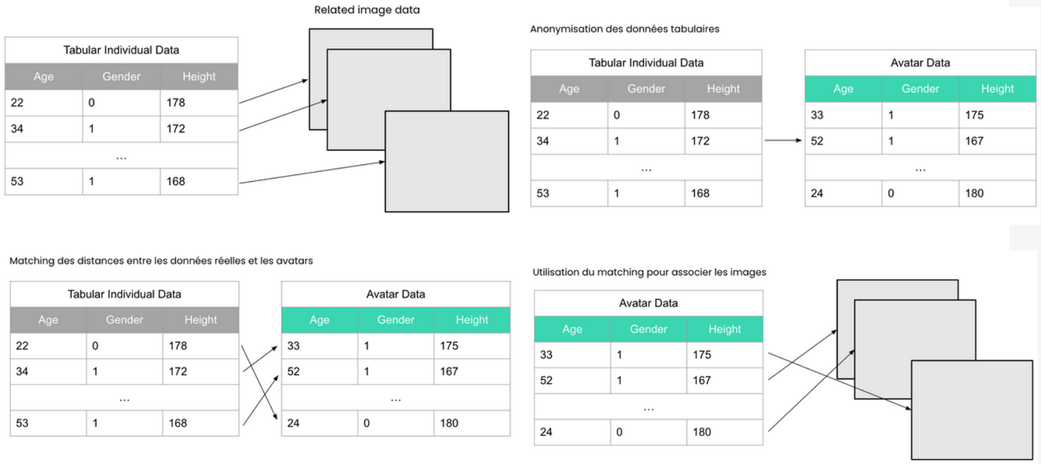
"Convaincu par les avatars, à l'issue d'expérimentations dans le cadre d'Epidemium et d'Echopen, j'ai proposé à Octopize de venir présenter la technologie des avatars au groupe de travail mensuel IA et Santé de l'Académie de Médecine."

- Olivier de Fresnoye, Directeur Général et Co-fondateur @echOpen

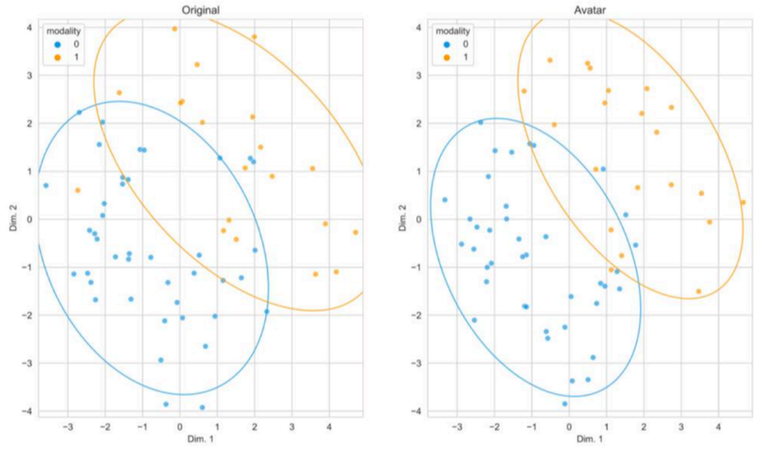
RETOUR SUR INVESTISSEMENT

- Accélération de la **recherche** sur le cancer
- Partage des données en **Open Data pour innover**
- Garantie de la **confidentialité** des patients et génération de **confiance** dans les données

PROCESSUS D'ASSOCIATION ENTRE UNE DONNÉE AVATAR ET UNE IMAGE



MAINTIEN DE LA QUALITÉ & DE L'UTILITÉ



- Les données anonymisées ont pu être utilisées en association avec des images (coupes d'anatomo-pathologie).
- Les propriétés structurelles du jeu original ont été conservées.
- Les données ont été mises à disposition des eunes praticiens pour les familiariser à l'utilisation de données de vie réelle.

OCTOPIZE
MIMETHIK DATA

#PARTENARIAT
#ACCÈS DONNÉES
DE VIE RÉELLES

CAS CLIENT

Évaluer la qualité des données avant l'établissement d'un partenariat

ROCHE / CHU DE BREST

CHALLENGE

- Améliorer le partage des données de santé pour faciliter les projets de R&D
- Évaluer la qualité des données avant une acquisition
- Accélérer les partenariats industriels-établissements de soin

1 an

gagné dans l'acquisition des données.

TÉMOIGNAGES

"Octopize propose une excellente solution dans le domaine des données synthétiques que nous sommes heureux de soutenir au sein de Roche !"

- Tania Aydenian, Innovation Senior Lead @Roche

"J'y vois un grand intérêt sur 2 points : un ROI au niveau RH mais aussi au niveau financier.
Ça prend moins de temps, donc on va pouvoir traiter plus de projets. Je vois vraiment ça comme une accélération, un game changer fort."

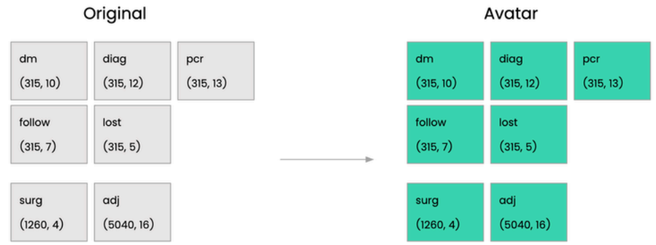
- Adrien Bussard, Chef de projet innovation @CHU de Brest

RETOUR SUR INVESTISSEMENT

- Accélération du **délai d'acquisition de 6 mois à 1 an**
- **Partage** de la valeur informative des données
- Résolution de l'asymétrie de l'information
- **Garantie du ROI** d'un partenariat
- **Valorisation** des données sans nouveau consentement

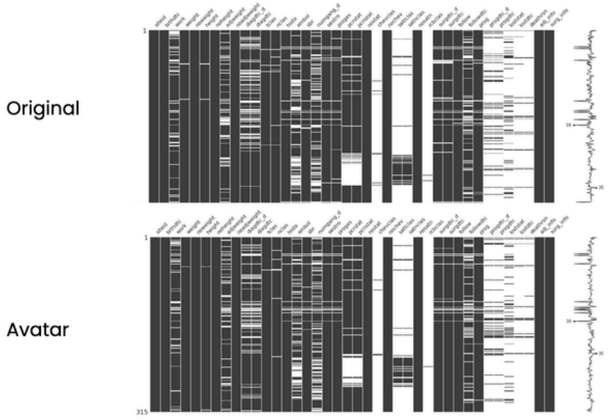
REPLAY WEBINAR

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ



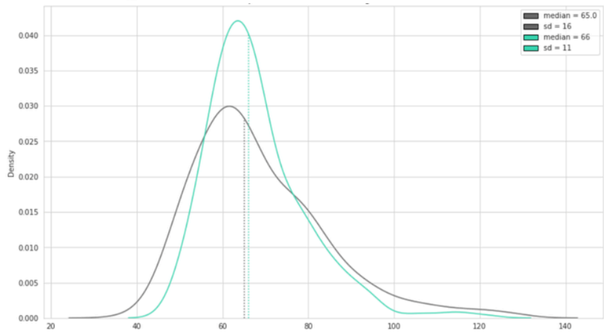
- Les données générées conservent la même structure de tables relationnelles que les données d'origines
- Chaque table synthétique présente le même nombre de lignes et le même nombre de colonnes

-> Les données synthétiques générées peuvent être utilisées sans modification dans les mêmes pipelines que les données d'origines



Dans ce schéma les zones blanches représentent les données manquantes et chaque bar verticale représente une colonne du jeu de données.

Ce graph comparatif illustre que les données synthétiques générées conservent le pattern des données manquantes.



Ce graph comparatif illustre que les données synthétiques générées conservent les distributions des données d'origine.

OCTOPIZE
MIMETHIK DATA

#MACHINE LEARNING
#CONFORMITÉ LÉGALE
#OPEN SOURCE

CAS CLIENT

pôle emploi

Proposer des services personnalisés via un modèle de Machine Learning éthique

FRANCE TRAVAIL (PÔLE EMPLOI)

CHALLENGE

- Ré-exploiter les données pour proposer des services personnalisés
- Alimenter un modèle de Machine Learning
- Être conformes aux règles européennes
- Assurer la conformité réglementaire sans nécessiter de nouveau consentement

1 an

gagné dans l'acquisition des données.

TÉMOIGNAGE

"La solution avatar d'Octopize est une solution **innovante**, **élégante** et qui, grâce à sa **rapidité** d'exécution, est facilement **exploitable** dans un système d'information."

- Laurent Guinard, Responsable du département Agence Data Service & Usine IA @Pole Emploi

RETOUR SUR INVESTISSEMENT

- **Conservation des performances de prédiction du modèle d'IA** avec des résultats très bons, aucune perte de performance voire même une hausse
- Publication des données en **Open Source**
- **Partage** de la valeur informative des données sans compromis
- **Rôle de réassurance des usagers** en cas d'audit

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ

Classement par max F1-score

k	cw	ncp	threshold	f1score	AUC	recall	preciseness	hidden_rate	local_cloaking
original	wo	NONE	0.30	0.517000	0.695600	0.685000	0.415000	0.000000	0.0
k20	cw1	ncp10	0.30	0.516976	0.687458	0.686522	0.414588	98.067481	52.0
k10	cw1	ncp30	0.25	0.512889	0.686425	0.747402	0.390394	97.159060	52.0
k20	cw1	ncp20	0.25	0.512482	0.687693	0.755112	0.387857	98.510993	53.0
k10	cw1	ncp20	0.25	0.512369	0.688300	0.747082	0.389880	97.857137	51.0
k20	cw1	ncp30	0.25	0.512098	0.685892	0.752554	0.388094	98.221803	54.0
k30	cw1	ncp20	0.25	0.511824	0.688280	0.749587	0.388572	98.819766	53.0
k30	cw1	ncp30	0.25	0.511629	0.685036	0.754437	0.387058	98.688528	55.0
k30	cw3	ncp30	0.25	0.511563	0.684983	0.733457	0.392746	98.649623	55.0
k10	cw1	ncp10	0.25	0.511392	0.688280	0.778757	0.380691	97.233497	51.0
k30	cw1	ncp10	0.25	0.510999	0.687412	0.781581	0.379587	98.479610	52.0
k100	cw1	ncp10	0.25	0.510610	0.686939	0.777158	0.380207	99.354314	53.0
k30	cw27	ncp30	0.30	0.509677	0.684352	0.688654	0.404540	98.772691	56.0
k30	cw9	ncp30	0.25	0.509041	0.683530	0.784673	0.376713	98.750386	56.0

F1 score original = 0.52
F1 score avatar = 0.52

AUC original = 0.70
AUC avatar = 0.69

OCTOPIZE
MIMETHIK DATA

#RECHERCHE
#PARTAGE

CAS CLIENT

CHU ANGERS
CENTRE HOSPITALIER UNIVERSITAIRE

Inserm

Publier des données de santé au travail et permettre la recherche

CHU d'Angers / INSERM

CHALLENGE

- Traiter les données sensibles liées à la **santé au travail**, incluant des informations confidentielles sur la santé et l'emploi.
- Démontrer que les **données anonymes** peuvent être utilisées pour la **science ouverte** et dans le secteur de la santé au travail.
- Permettre la transparence dans le cadre du travail de **publication scientifique**.

TÉMOIGNAGE

« Malgré un **jeu de données et un nombre de variables important**, le logiciel Octopize a permis de générer des données synthétiques exploitables pour mener des recherches à l'échelle d'une cohorte d'individus.

Les résultats sont meilleurs que ce que j'attendais ! »

- Pr. Descatha, Professeur des Universités et Praticien Hospitalier en médecine et santé au travail, @CHU d'Angers

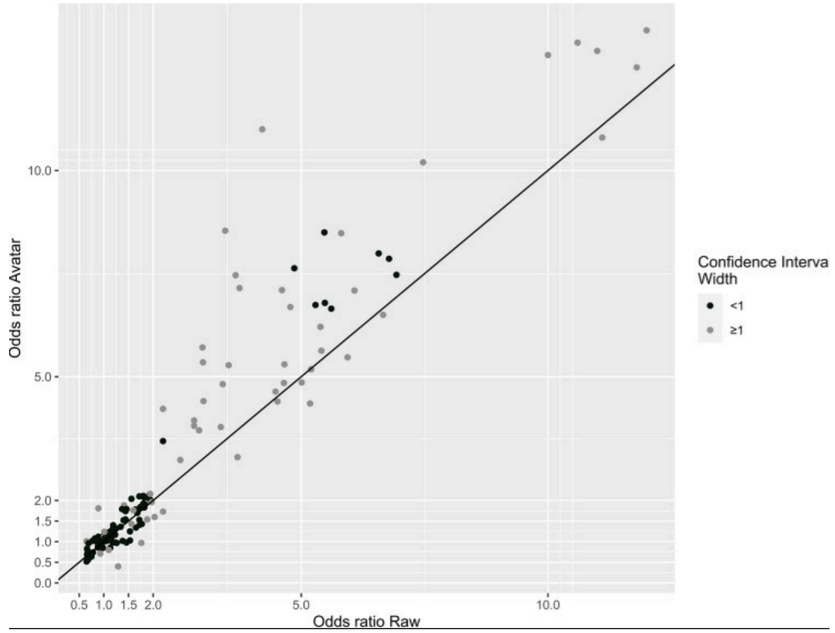
RETOUR SUR INVESTISSEMENT

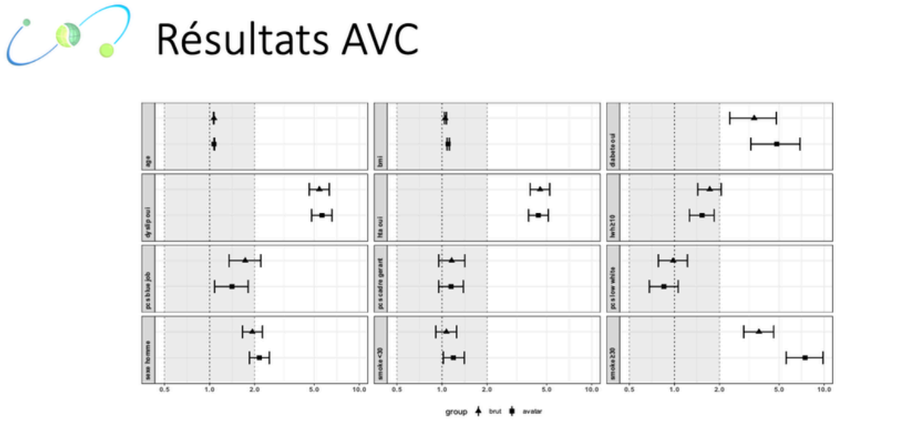
- **Partage des données** grâce au logiciel avatar (respectant qualité de l'information & confidentialité des individus)
- **Accélération de la recherche** en santé publique au travail (publication d'un article scientifique : une première rendue possible)

1ère

Publication avec des données anonymes de santé au travail rendue possible grâce à la méthode.

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ





UA FACULTÉ DE SANTÉ UNIVERSITÉ D'ANGERS
université angers
CHU ANGERS
irset
Inserm
Feinstein Institutes for Medical Research
Northwell Health
Descatha
DONALD AND BARBARA ZUCKER SCHOOL of MEDICINE AT HOFSTRA/NORTHWELL

OCTOPIZE
MIMETHIK DATA

#PARTAGE
#CONFORMITÉ LÉGALE
#PRIVACY SHIELD

CAS CLIENT

Développer un protocole de soin avec une filiale aux USA, en respectant le cadre réglementaire

BIOMÉRIEUX

CHALLENGE

- **Partager** des données issues d'une cohorte rétrospective (datant de 10 ans) **à un prestataire hors U**E (invalidation du Privacy Shield).
- **Accéder rapidement** à la valeur informative des données pour développer un nouveau protocole de soin.

1 an

de gagné dans le développement d'un nouveau protocole de soin.

TÉMOIGNAGE

"Octopize a répondu à nos besoins complexes : réaliser des **analyses pointues** tout en respectant les contraintes du **partage aux USA** de données sensibles de 15 000 patients sur une période de 10 ans. L'anonymisation s'est révélée être **notre meilleure option**, ouvrant des perspectives pour la recherche en assurant confidentialité et reproductibilité des analyses.

En plus de débloquer une impasse, Octopize a considérablement accéléré notre projet, nous faisant **gagner un temps précieux** et ouvrant de **nouvelles voies thérapeutiques.**

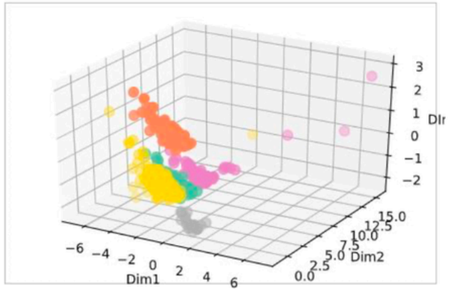
Une collaboration qui a dépassé nos attentes. "

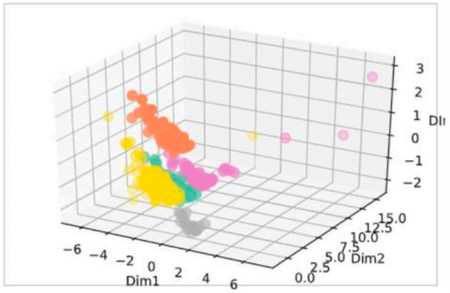
- Julien Textoris, Vice President & EME Medical Affairs @bioMérieux

RETOUR SUR INVESTISSEMENT

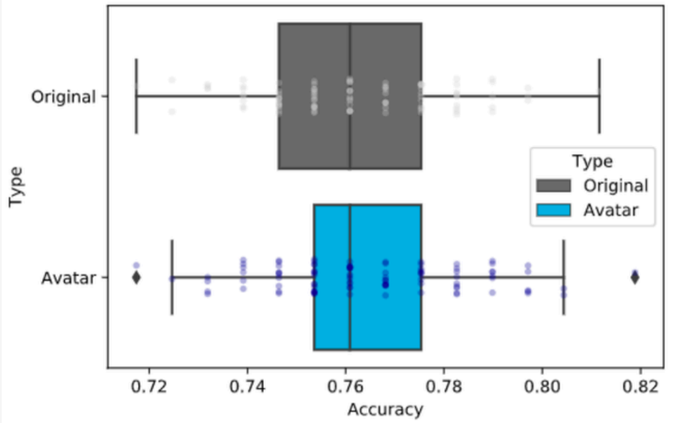
- **Partage de données vers une filiale aux USA** en respectant le contexte réglementaire
- **Accélération de la recherche** et développement d'un nouveau protocole de soin
- **Conservation de la valeur informative des données** sans compromettre la confidentialité des patients

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ





- Original à gauche, avatar à droite. Les patients sont représentés selon leur motif d'inclusion (5 sous groupes)
- Les caractéristiques des patients par cluster sont respectées -> données utilisable à l'échelle multifactorielle pour des analyses poussées



Le modèle de ML entraîné sur les données avatar présente les mêmes performances de prédiction que le modèle entraîné sur les données d'origine.

Comment lire ce graph ?

100 itérations ont été réalisées pour chaque modèle.
Chaque point représente le score de prédiction d'un de ces modèles (0.76 = 76% de bonne prédiction).
50% des valeurs des modèles sont comprises dans la boîte (cf rectangle bleu et gris) et la médiane de performance de prédiction des modèles est représentée par la barre verticale au sein de la boîte.
Il ressort donc que médiane original = 0.76 et que la médiane avatar = 0.76 également.

OCTOPIZE
MIMETHIK DATA

#PARTAGE
#CONFORMITÉ LÉGALE
#PRIVACY SHIELD

CAS CLIENT

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Accélérer la recherche de façon éthique et dans le respect des contraintes réglementaires

CHU DE NANTES

CHALLENGE

- Faciliter les **projets de recherche en santé** : utiliser les données pour réaliser des **analyses**, des **essais cliniques** ou encore entraîner des **logiciels d'IA**…
- Respecter les réglementations européennes (**RGPD**).
- Garantir la **confidentialité** des patients.

6

publications scientifiques réalisées avec des données avatar

TÉMOIGNAGE

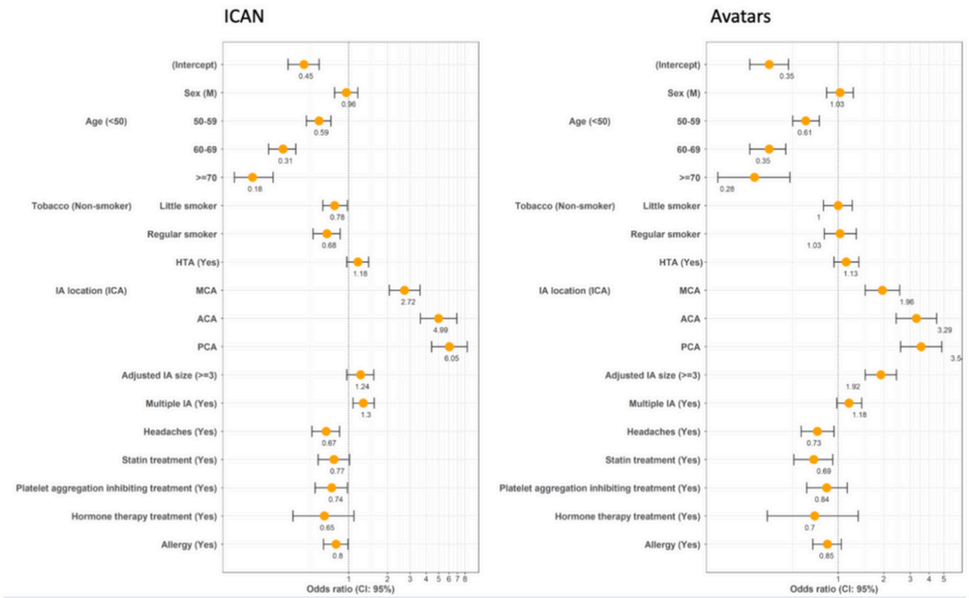
« Il n'y a plus de raison de manipuler des données de santé identifiantes pour faire des analyses statistiques.»

- Pr. Pierre-Antoine Gourraud, Pr. Faculté de Médecine @Nantes Université, Praticien hospitalier, @CHU de Nantes

RETOUR SUR INVESTISSEMENT

- **Levier politique** de gestion des données
- Contribution à la **reproductibilité de la Science**
- Accélération de la **recherche collaborative**
- Outil de **compliance** en recherche
- Outil de stratégie de **collaboration industrielle** et sur les logiciels

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ

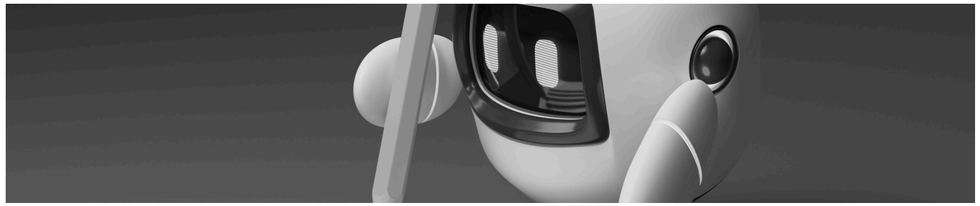


- Les données avatar ont été utilisées pour la publication d'articles scientifiques car elles reproduisent les résultats des données d'origine
- Source : **publication ICAN**

OCTOPIZE
MIMETHIK DATA

#ANALYSES SECONDAIRES
#CONFORMITÉ LÉGALE
#PARTAGE

CAS CLIENT



Wefight

Analyser et ré-exploiter des données de santé collectées par un chatbot

WEFIGHT

CHALLENGE

- Analyser et ré-exploiter les **données sensibles de santé** (texte libre) collectées par Vik, une intelligence artificielle qui accompagne les patients atteints de maladies chroniques
- Respecter la **confidentialité** des patients à l'origine des données et les **réglementations** en vigueur.

6

mois à 1 an de gain de temps dans le partage de données

TÉMOIGNAGE

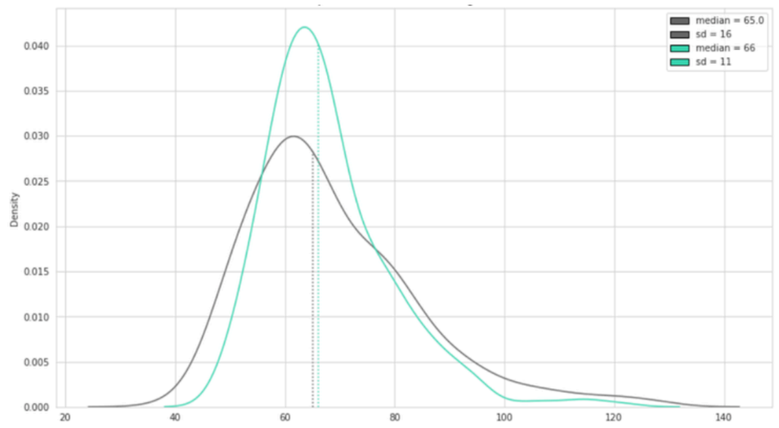
« La méthode avatar remplit ses promesses :
c'est une technologie **transparente, facile à utiliser**, bien **documentée** qui a permis à Wefight de passer un vrai gap dans l'anonymisation des données. »

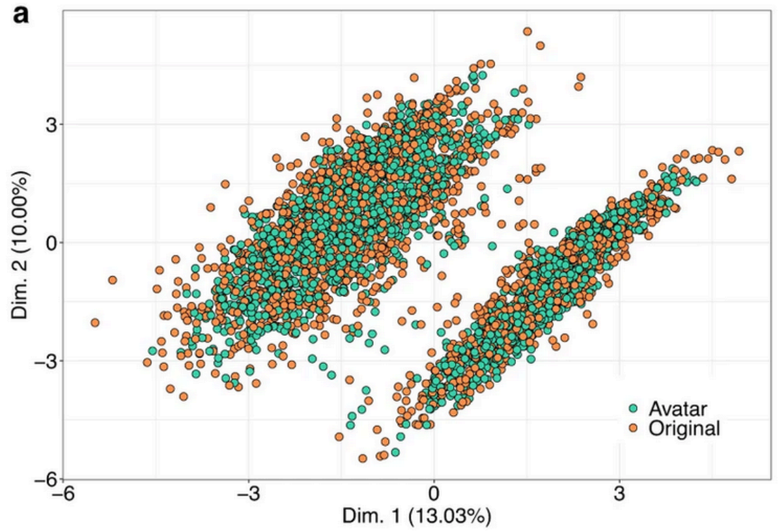
- Arthur Guillemassé, Head of Data @wefight

RETOUR SUR INVESTISSEMENT

- Accélération de la **recherche en santé.**
- Partage des données en conformité avec le **RGPD.**
- Ré-exploitation des données pour **l'industrie pharmaceutique.**
- Génération de **confiance** auprès des clients en garantissant l'anonymat sur les données.

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ



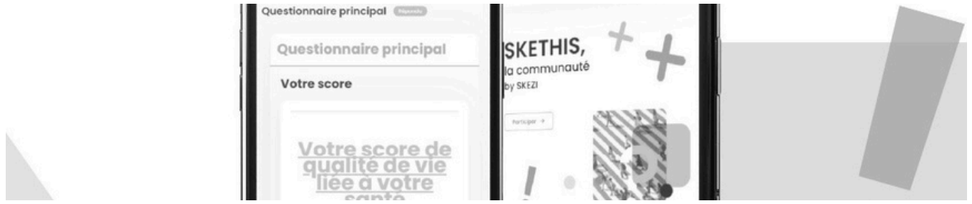


- Les données anonymes conservent les caractéristiques des répondants au questionnaire d'origine.
- La structure globale du jeu de données est également conservée permettant de grouper les répondants en cluster type.

OCTOPIZE
MIMETHIK DATA

#OPEN DATA
#VALORISATION
#ANALYSES

CAS CLIENT



SKEZI

Exploiter et valoriser les données publiques de la e-cohorte SKETHIS en toute conformité

SKEZI

CHALLENGE

- Exploiter et valoriser de façon **éthique**, les données de la e-cohorte SKETHIS (sur la qualité de vie liée à la santé).

- Assurer la **confidentialité** des données sensibles tout en conservant **l'information** utile.

TÉMOIGNAGE

"Ce partenariat est clé pour SKEZI, car la protection des individus est au coeur de nos préoccupations.

En garantissant aux volontaires de la e-cohorte SKETHIS l'anonymisation de leurs données, ils seront plus à même de partager leurs ressentis et participer à l'amélioration de la qualité de vie en France."

- Jean-Philippe Bertocchio, Médecin, Chercheur, CEO @SKEZI

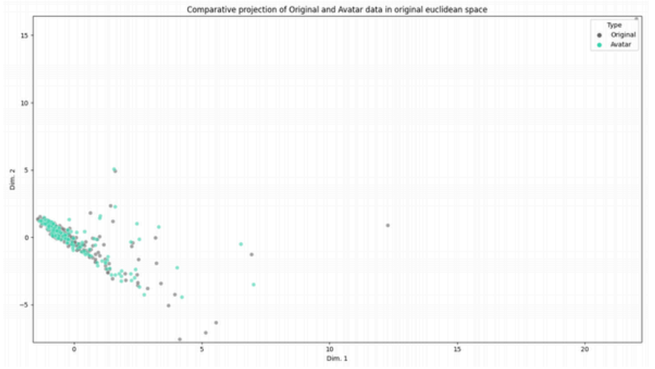
RETOUR SUR INVESTISSEMENT

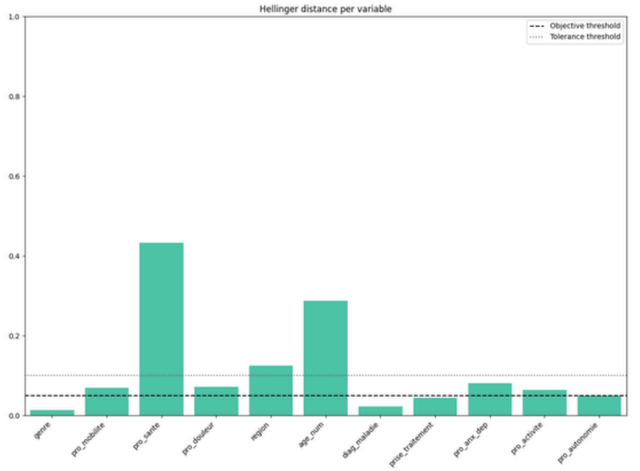
- Valorisation et analyses des données sans compromettre la confidentialité
- Conservation des données pour une durée illimitée
- Partage des données à des fins de recherche
- Génération de la confiance des participants

100 000

nombre de personnes visées dans la cohorte d'ici 2025

MAINTIEN DE LA QUALITÉ STATISTIQUE & DE L'UTILITÉ





Privacy metric	Value	Target
Distance to closest	0.48	> 0.2
Closest distance ratio	0.81	> 0.3
Closest rate	94.09 %	> 90 %
Correlation protection rate	100.0 %	> 95 %
Categorical inference	46.85 %	%
Categorical hidden rate	97.41 %	> 90 %
Column direct match protection	98.62 %	> 50 %
Row direct match protection	99.21	> 90