“La Clinique des données”

Pr *Pierre-Antoine GOURRAUD*

« Un check-up de l’IA pour la santé numérique, c’est grave docteur(s) ? »

Séminaire Aristote - jeudi 27 février 2020,

École Polytechnique, Amphithéâtre Arago
91120 Palaiseau

ATIP-Avenir Team 5 «Translational Immunogenomics of Transplantation and Autoimmunity »,
ITUN - CRTI - UMR Inserm 1064 -CHU de Nantes

Pôle Hospitalo-Universitaire 11 : Santé Publique, Santé au Travail et Pharmacie,

Hôpital St-Jacques - CHU de Nantes

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale

« Un check-up de l’IA pour la santé numérique, c’est grave docteur(s) ? »

→ ***Passez donc à la « Clinique des données »***

1. Quelques notions clés
2. Un nouveau service – et ses motivations: “la clinique des données”
3. La technique au service de la gouvernance des données : “les avatars”

Quelques notions en filigrane

Partie 1

A two-fold re-evolution…

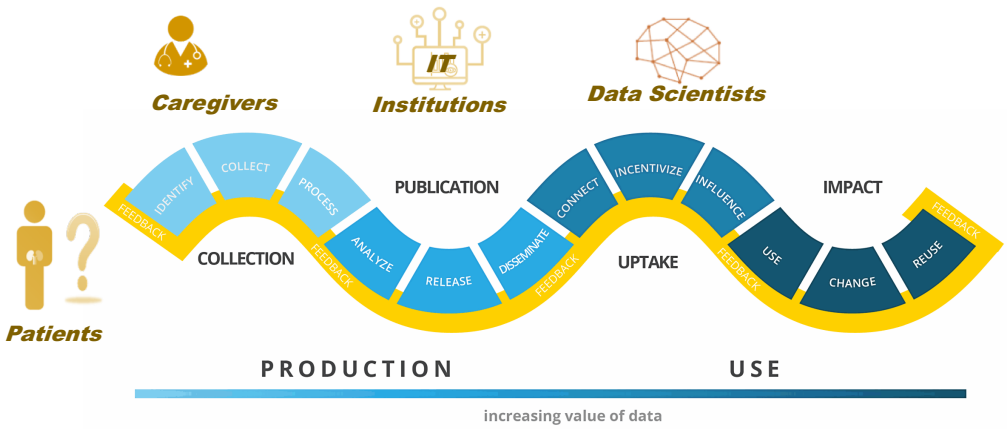
Ever lower cost of data access

- Information become technically available
- Central role of search (how to access what)

Ever lower cost of computation

- Computation
- On-demand
- Just computations…

Multiples stakeholders and contributions to the data chain value.

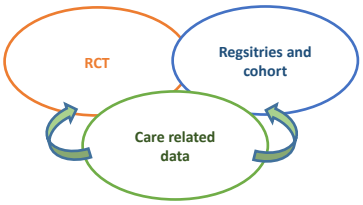


(Adapted From https://opendatawatch.com)

A shift in health Data for research

Clinical Trials

- ***Controlled via randomization***
- **Protocoled**
- **1 main question**
- **Selective inclusion**
- **Limited N (10^2 10^3)**
- **€€€**
- **Hypothesis-driven**

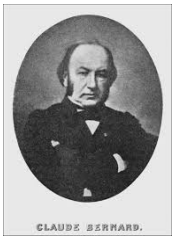


Registries/Cohorts

- **Systematic**
- **Comprehensive**
- **Selected disease and population**
- **Answer a limited set of questions**
- **Open N but limited (10^2 - 10^5)**
- **€€€**
- **Hypothesis-driven**

EMR/EHR

- ***Data are non- intentionally collected for research***
- **Real life evidence – local practice –local evidence**
- **Sparse**
- **Biased**
- **Large N (10^4 10^6)**
- **Agnostic**
- **Recycling €**



Claude Bernard, 1813-1878

All innovation in health now deals with IT …. A framework

Digital/ Numeric/ IT Technology
Application to Health of generic IT technology
Non-disruptive practice – no fundamental change.
« Altius Citus Foritus » + Healthism
Ex: On-line Drug catalog

Big Data Technology
Application to Health of generic IT technology
Infusive change of paradigm – "re-evolution" – slow disruption
Ex: Simulation vs Models - Machine Learning

Buzz Words
[3-4-6]P Medicine,
Precision Medicine
Beyond (NGS) Genetically informed

Phenotyping technology
Disruption of medical practice by emerging
To newly see & measure
Ex: Stethoscope – MRI - DNA sequencing - Biomarkers

Population-based approach of health (Public Health)
Vs.
Pseudo-singular approach of clinical medicine

Performative technology
Direct disruption of medical techniques / intervention
To differently do
Ex: Micro Surgery – drugs – targeted therapies

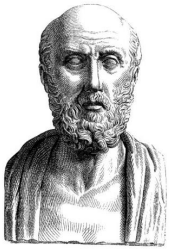
Cognitive technology
Disruption of explanatory model - understanding differently
Ex: Systems biology – network based drugs design – HMM – SEM – AI explicatbility – AI response

Data exceptionalism in health « not matter how big…"

- **Technically complex, Legally ruled, Socially sensitive**
  - **Health data at the crossroad of patients, health care systems, and care givers (vs.genomics)**
  - **V : Variety ; N : Numbers ; C : Categories ; P : Probabilistic**
  - **T : Time is needed, duration matter**
  - **Health data are sensitive – Privacy**

"Primum Non Nocere"

- ***Data come from patients…***
- ***Patients come first***
- ***It applies to … Data protection in health research***
  - ***First do no harm***
  - ***First respect patient privacy***



Hippocrates of Kos;
c. 460 – c. 370 BC

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

La « cliniques des données » et l'entrepôt de données biomédicales du CHU de Nantes

Partie 2

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Un réflexion sur l'enjeu de données

Constat partagé : Verrou des données

« Bonnes » données, un préalable à de bonnes analyses

Les données: c'est l'expertise du domaine

Compréhension et qualité des données sont un enjeu croissant à mesure que la disponibilité technique est résolue par les outils numériques.

Les données méritent une « clinique » au CHU de Nantes

Réfléchir, Traiter, Intégrer = Consultation, Open-Space, Formation

- **Dans son rôle d'accompagnateur méthodologique des projets de recherche et d'évaluation , la « Clinique des Données » devient médiateur de l'accès à l'entrepôt de données du CHU de Nantes (source de données)**
  - Fait la requête et fournit une extraction utile = décompte ou datamart pseudo-anonymisées(principe de parcimonie)
  - Garant de la protection du patient-usager-payeur-citoyen (Transparence et contrôle des usages et finalités)

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Trois dimensions à articuler

1. **Gouvernance** garantissant le respect éthique, réglementaire et scientifique de chaque projet de recherche (=structure de confiance vis-à-vis des patients et des soignants)
2. **Information des patients**, pour chaque exploitation des données avec opposition possible, par affichage site web CHU
   - Communication grand public (avril 2018→)
3. **Outil des gestion données générées par le soin au CHU de Nantes**
   - Ehop (Pr M Cuggia - Rennes)= Transformation/intégration de données + Entrepôt de données type « Big Data » + Requêteur « intelligent »
   - GIRCI GO en 2016-2018 (réseau des RiCDC)- HUGO


CHU de Nantes
info.CHU

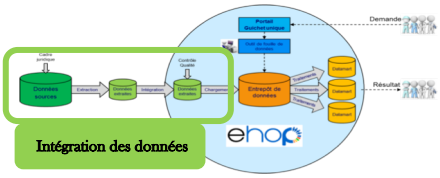
Le traitement de vos données personnelles au CHU de Nantes

GO!

CNIL

Juillet 2018

Extractions réalisées pour charger les antériorités dans MILLENNIUM	Antériorités
Reprise Patients	Depuis 1984
Reprise Séjours	Depuis 2000
Reprise RDV à venir	18 mois
Reprise Bureautique Clinicom Comptes-rendus Word	Depuis 1998
Reprise Radiologie Comptes-rendus Texte	Depuis 2004
Reprise Anapathologie Comptes-rendus Texte Comptes-rendus PDF	Depuis 2004 Depuis 2013
Reprise Biologie PDF Comptes-rendus PDF tous laboratoires	Depuis 2014
Reprise Biologie Texte	
Comptes-rendus Texte Bactériologie	Depuis 2002
Comptes-rendus Texte Mycobactérie	Depuis 2002
Comptes-rendus Texte Parasitologie	Depuis 2002
Comptes-rendus Texte Pharmacologie	Depuis 2002
Comptes-rendus Texte Virologie	Depuis 2002
Comptes-rendus Texte Biochimie Générale	Depuis 2010
Comptes-rendus Texte Biochimie Spécialisée	Depuis 2010
Comptes-rendus Texte RIA	Depuis 2010
Comptes-rendus Texte Hématologie	Depuis 2010
Comptes-rendus Texte Immunologie	Depuis 2010
Reprise Biologie structurée	
Biochimie Générale	Depuis 2008
Hématologie	Depuis 2008
Biochimie Spécialisée	Depuis 2008
Radio-Immunologie	Depuis 2008



Flux 1
Flux 2
Flux 3
Flux n

Entrepôt de données

Exploitation recherche :
- Requêtage
- Exports

2,7 Millions de patients

118 Millions de données structurées

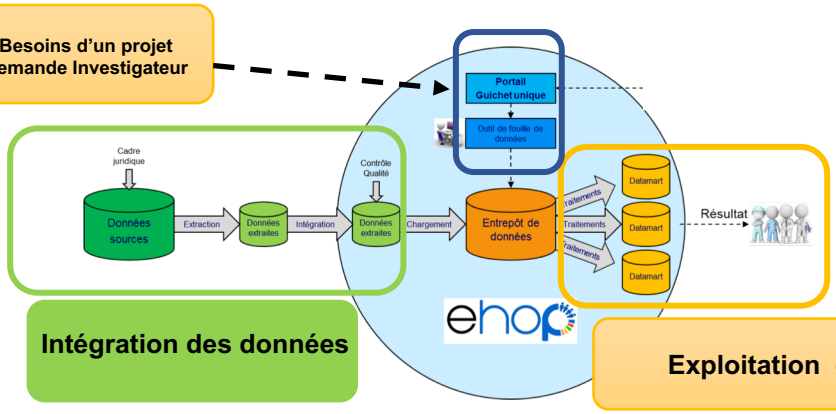
30 Millions de documents textuels

EDBN CHU NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

ehop

Besoins d'un projet
Demande Investigateur



Intégration des données

Exploitation des données

Exigences réglementaires (CNIL, RGPD) pour garantir la protection des données personnelles

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Exemple Screening RectoColite Hémorragique – Combinaison des requêtes

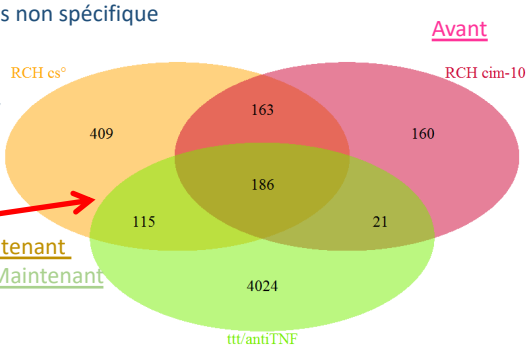
«Décompte de patients atteints de « RCH »

- Patients suivis en consultation pour RCH (codé (avant) ou mentionné dans le texte des CR de consultation)
- Hospitalisé ou non pour diagnostic principal de RCH codé en sur cette période (avant)
- et traités avec anti-TNF relatif à la pathologie mais non spécifique

On retrouvé des patients non identifiés « RCH » par le PMSI mais vus en hospitalisation :

- *Codage CIM-10 : Z512 « Autres formes de chimiothérapie »*

115 patients



"TAL" exclusion

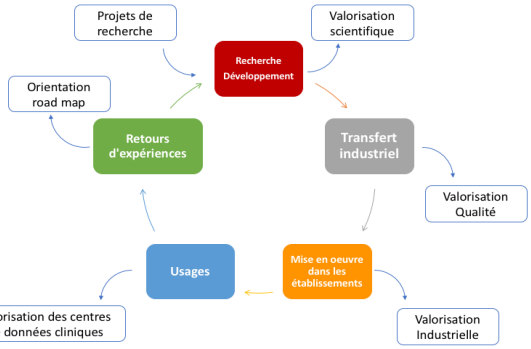
- **Les hypothèses rejetées : « pas d'argument pour une recto colite hémorragique »**
- **Les cas non avérés : suspicion, en prévention de…**
- **« son mari/père/épouse … est atteint(e) de recto colite hémorragique »**
- **Patients transférés au RCH (= rez-de-chaussée haut)**

14

Perspective : Les CHU "carrefours des données" de Santé

- **Articulation entre modes d'accès nationaux – accompagnement local**
  - Circuit d'information individuelle systématisée
  - Proximité de la production des données : « *En santé, un data scientist est d'abord un expert du contexte dans lequel naît la donnée avant d'être un expert des méthodes de traitement de ces données !"*
- **Médiation par un tiers-expert – Structure locale labellisée**
  - *« les données parleraient d'elles-mêmes »*
  - *"Jamais seul face aux données"*
  - HUB local - Clinique des données – Centres de Données Cliniques
- **Enjeu de transformation – nouvelle épidémiologie de données.**
  - On parle trés vite en millions…
  - Données en « vie réelle » - qui requièrent plus de méthodes

Regional Context - Millions of Medical Data from University Hospitals



girci grand ouest
LA RECHERCHE CLINIQUE

Accessible by project for public & private investigators

Biomedical data warehouse on the move

4,6 millions patients

112 millions clinical documents

A single IT for 8 institutions at the crossroad of healthcare system for 11 millions people

5+ millions Visits

1.3 billions structured data

The 3 ingredients hierarchy of Biomedical Data Warehouse

3- A common IT infrastructure
When data meets IT – interoperability is not guaranteed

2- A regulatory framework to inform patients
Nantes Biomedical warehouse was approved on April 2018

1- Governance
Mediated access to the wealth of dat a
Internal for research, screening etc

Stratégie Interrégionale

Projet GAVROCHE

Une question clinique

La variabilité glycémique permet-elle d'affiner le pronostic vital des patients admis pour insuffisance cardiaque aiguë ?

Une source de données massive inter-entrepôts du Grand Ouest

+ de 20 000 patients admis, + de 100 000 mesures de glycémie, > 8 ans d'activité

Des données hétérogènes mais accessibles

- Structurées : biologie (DXLAB) et diagnostics (PMSI)
- Non structurées : clinique extraite du DPI par traitement automatique du langage

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

=> Un parmi 4 projets fédérateurs du GIRCI Grand Ouest

Contexte national

Projet HEALTH DATA HUB

- Accès facilité aux données de santé de sources multiples (hôpitaux, ville, pharmacie, biologie…)
- Environnements sécurisés
- Valorisation du « patrimoine » français de données de santé
- Hub national et hubs locaux

➡ **Entrepôt(s) HUGO**

La technique au service de la gouvernance des données : “les avatars”

Partie 3

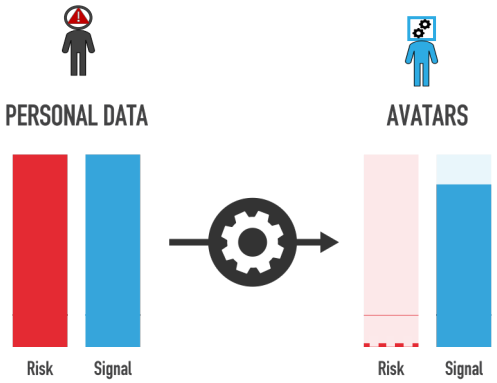
Risk of re-identification in (pseudo-)anonymous data

"Primum non nocere" = lower risk
Systematic introduction of the use of avatars for each project of the "Data Clinic"

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

WEDATA

Public- private PARTNERSHIP



WeData method **transforms** personal data into signal-retaining avatars with **privacy priority**

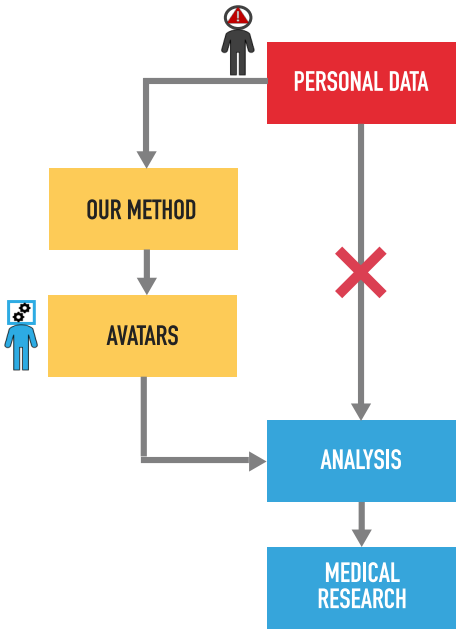
DEEPTECH bpifrance

LAURÉAT Réseau Entreprendre ATLANTIQUE

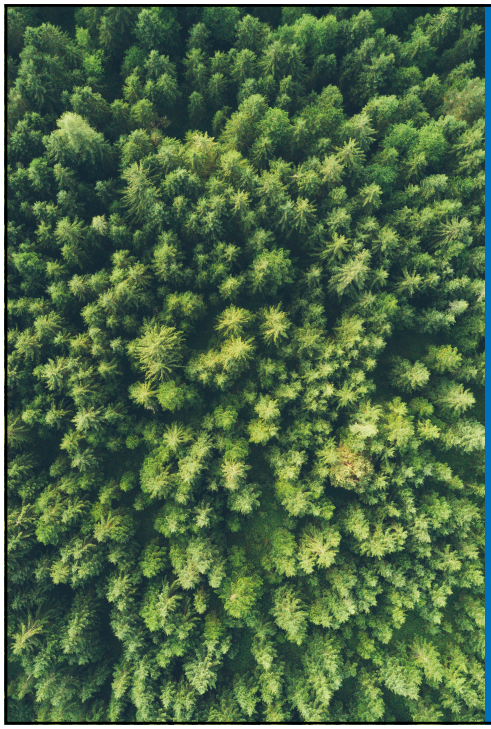
ATLANPOLE Land of innovation

21

Enable analyses to draw **comparable conclusions** without compromising **privacy**



22



ALGORITHM

"Blend into the crowd"

- Based on "**look-alike**" approaches
- Each individual's **surroundings** are used to create an avatar
- **Locality** sensitive method
- An Avatar is a **synthetic** individual

23

EXAMPLE

- Original sensitive data
- Avatar

• Application of the method

Example of avatars dataset

Individual	age	height	color
…	…	…	…
165	~~12~~ 9	~~147~~ 129	~~Blue~~ Blue
166	~~13~~ 11	~~158~~ 140	~~Green~~ Red
167	~~9~~ 12	~~132~~ 146	~~Blue~~ Red
168	~~11~~ 8	~~139~~ 124	~~Red~~ Blue
169	~~6~~ 12	~~110~~ 150	~~Blue~~ Blue
…	…	…	…

EXAMPLE

- Original sensitive data
- Avatar

- Application of the method

Example of one individual avatarized multiple times

Individual	age	height	color
…	…	…	…
165	~~12~~ 9	~~147~~ 129	~~Blue~~ Blue
166	~~13~~ 11	~~158~~ 140	~~Green~~ Red
167	~~9~~ 12	~~132~~ 146	~~Blue~~ Red
168	~~11~~ 8	~~139~~ 124	~~Red~~ Blue
169	~~6~~ 12	~~110~~ 150	~~Blue~~ Blue
…	…	…	…

26

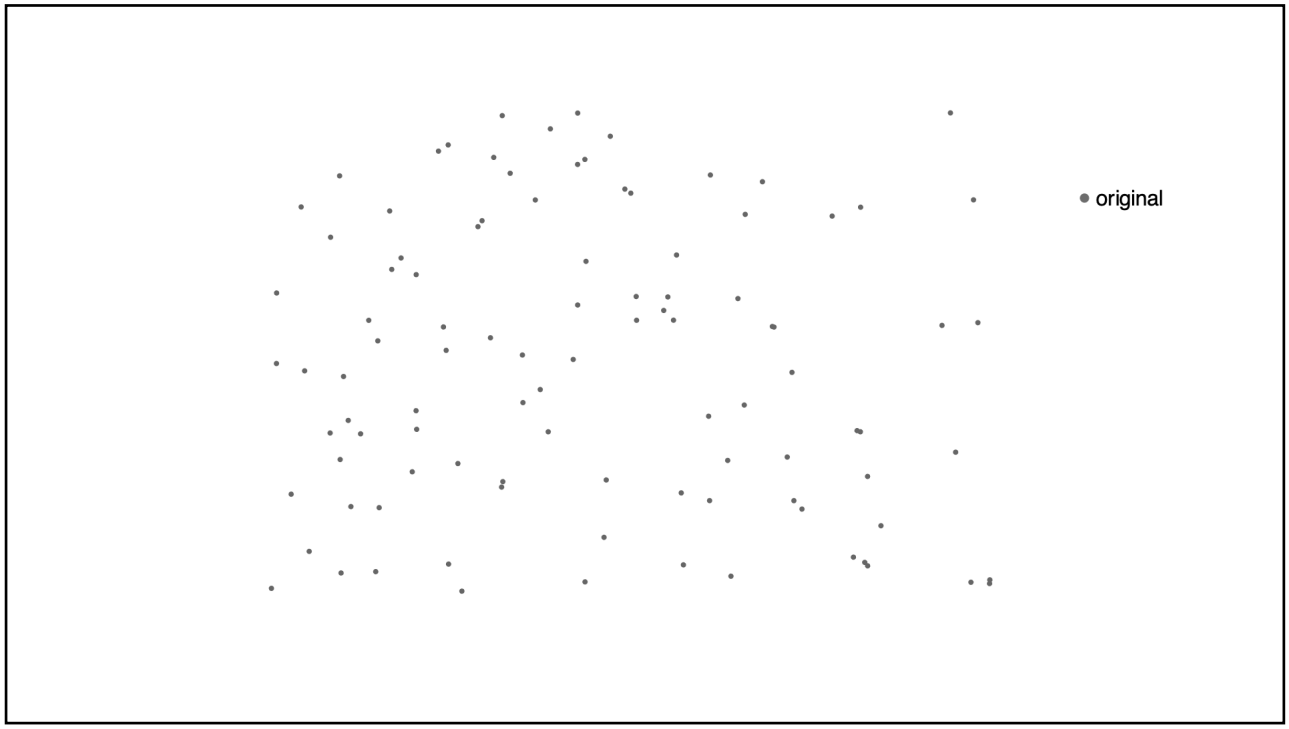
EXAMPLE

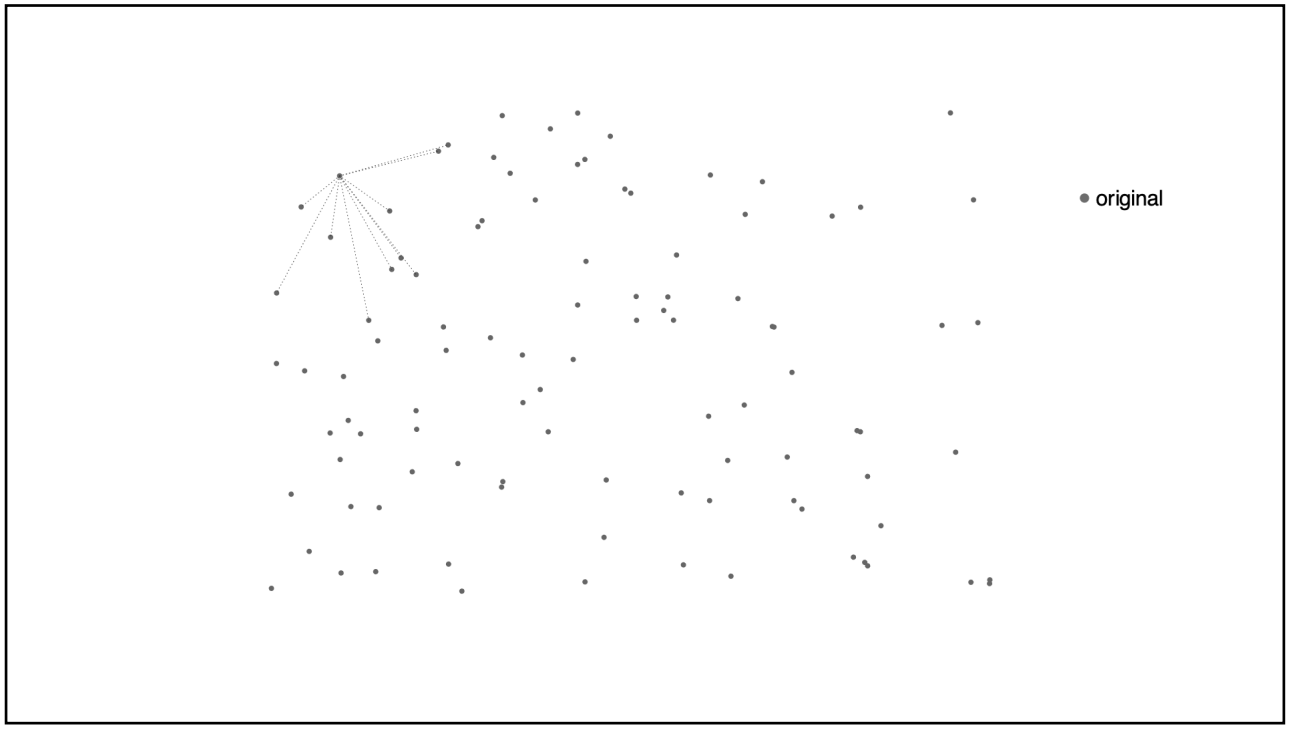
- Original sensitive data
- Avatar

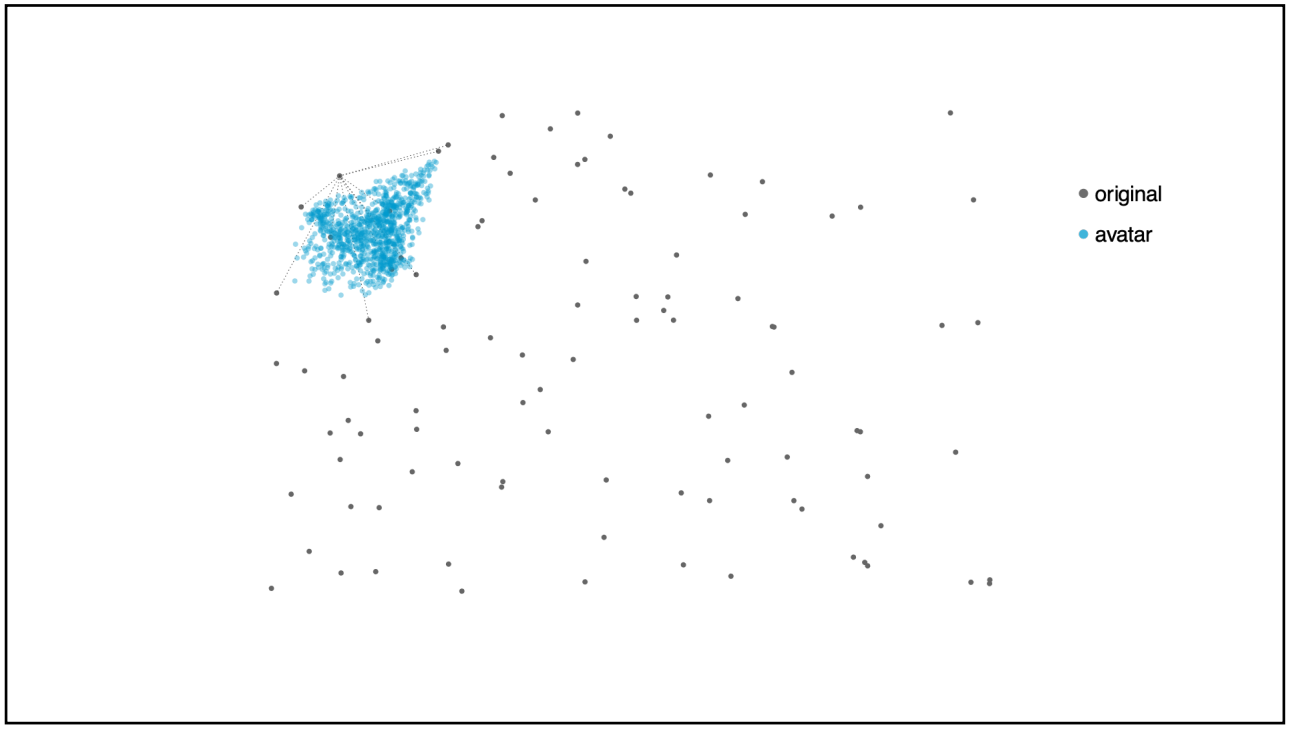
- Application of the method

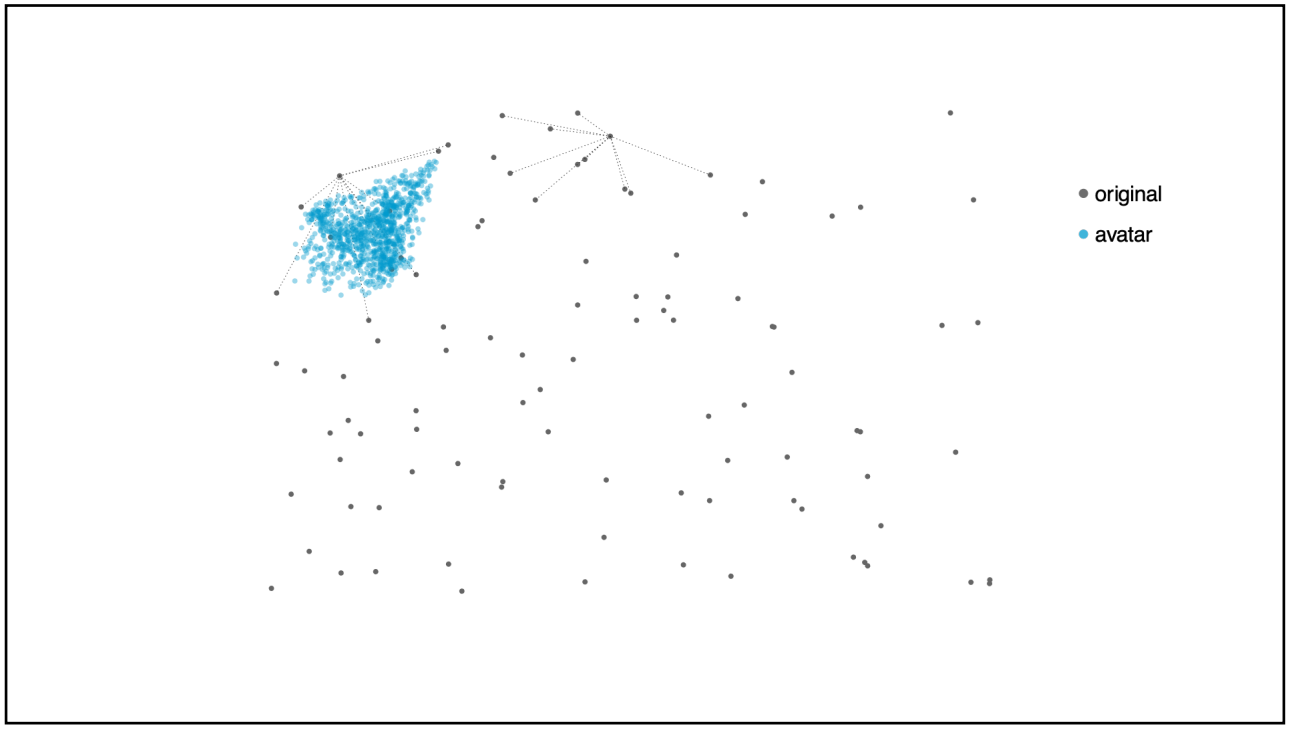
Example of one individual avatarized multiple times

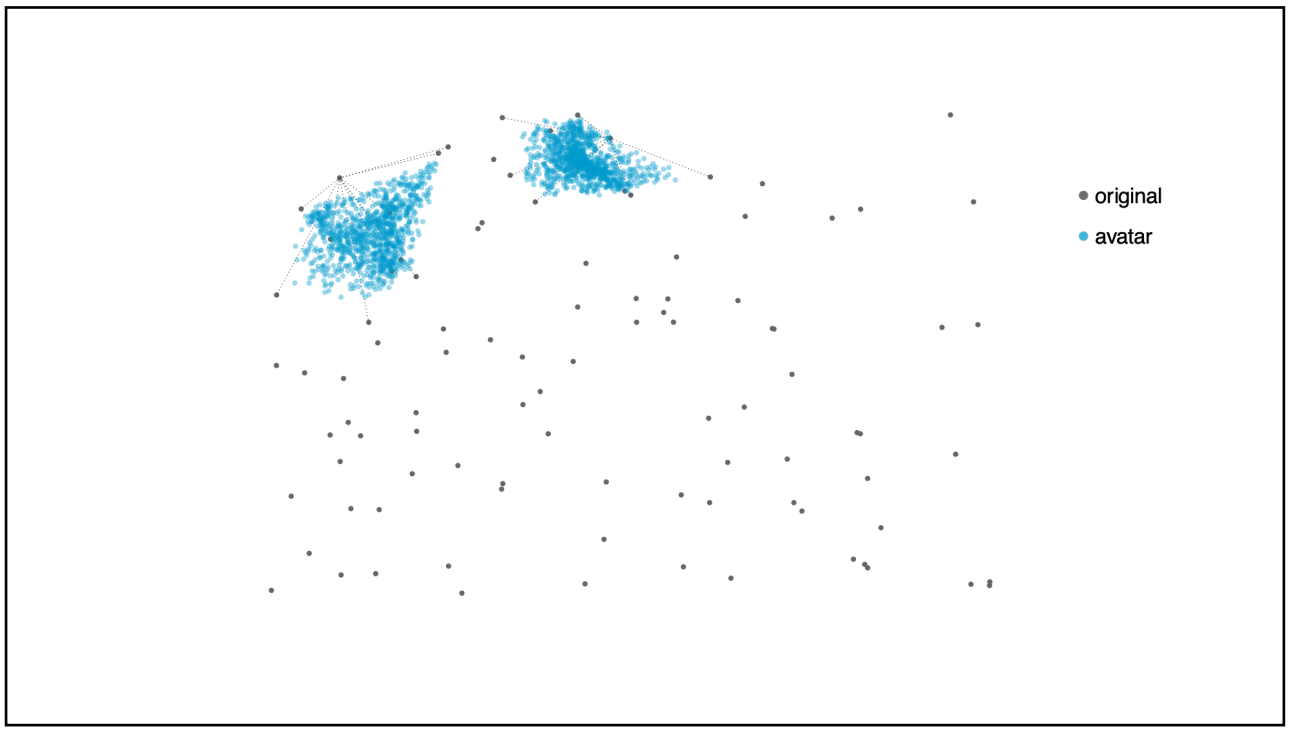
Individual	age	height	color
…	…	…	…
Original 167	9	132	Blue
Avatar 167 (1)	~~9~~ 11	~~132~~ 140	~~Blue~~ Blue
Avatar 167 (2)	~~9~~ 12	~~132~~ 146	~~Blue~~ Red
Avatar 167 (3)	~~9~~ 8	~~132~~ 124	~~Blue~~ Blue
Avatar 167 (4)	~~9~~ 10	~~132~~ 129	~~Blue~~ Blue
…	…	…	…

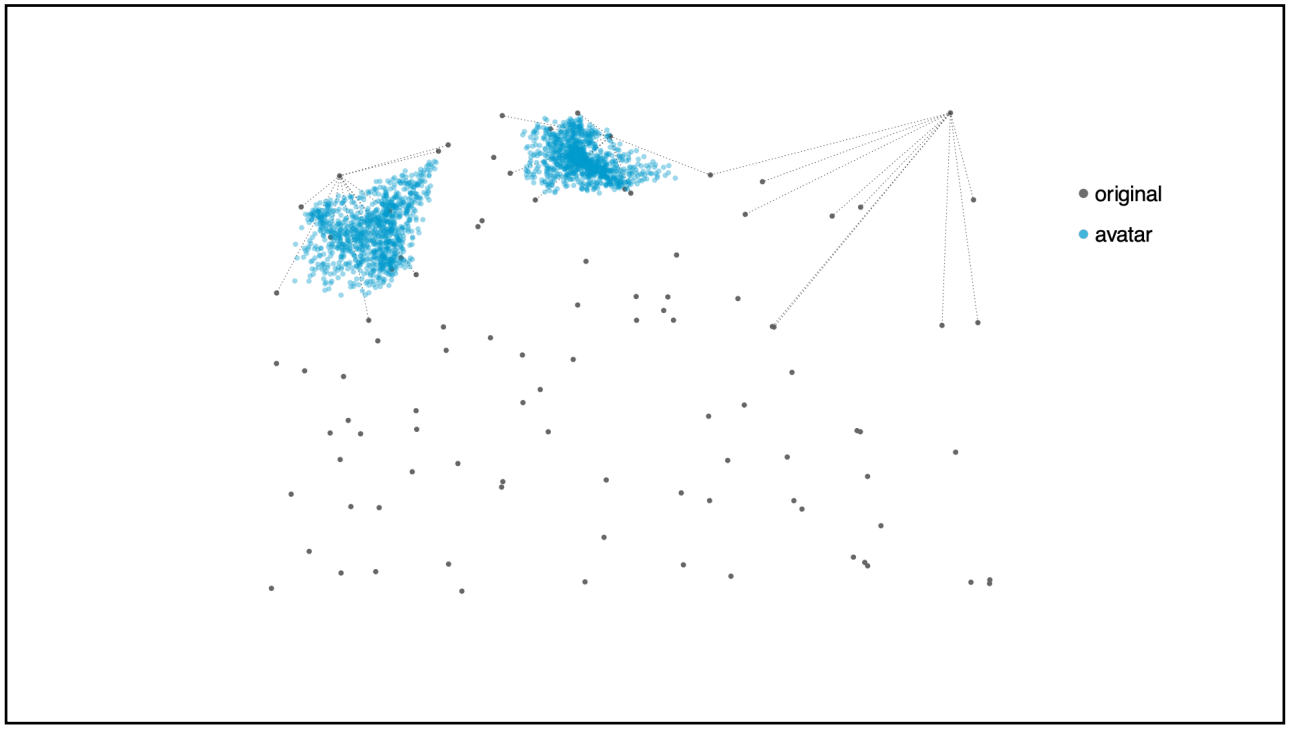

original

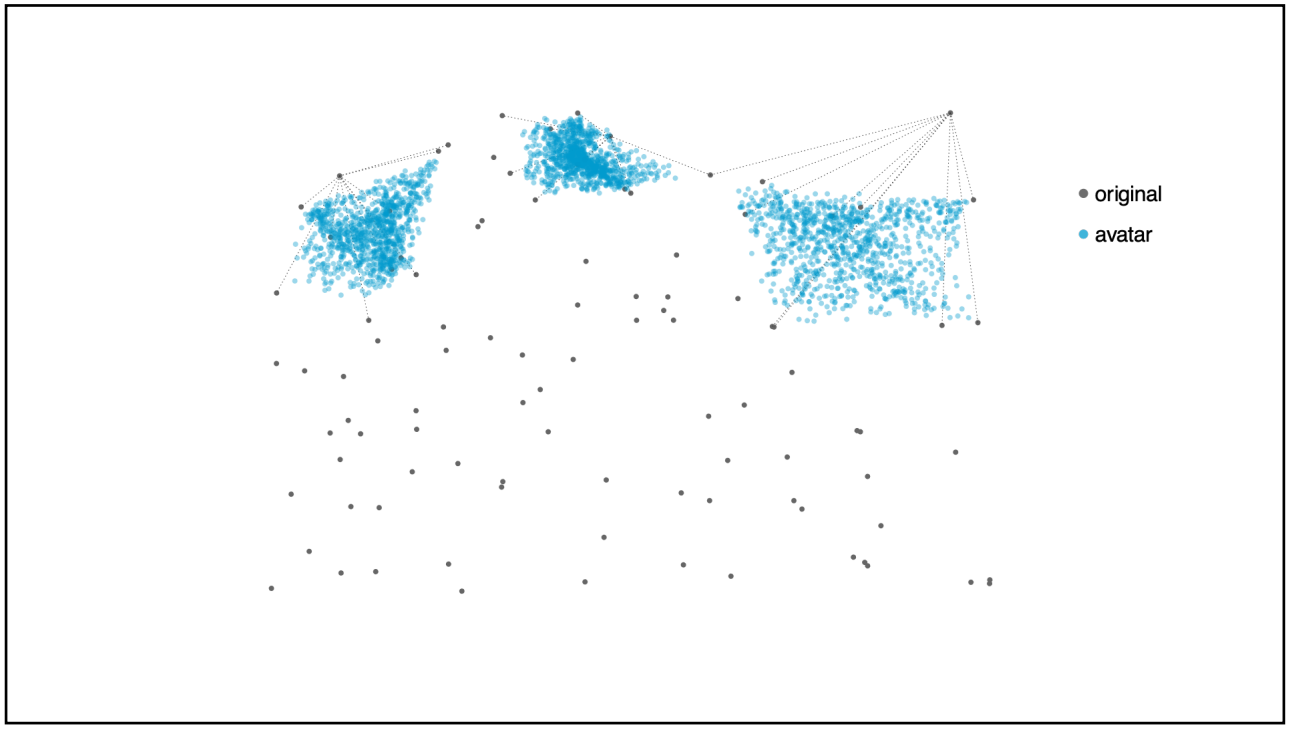

original

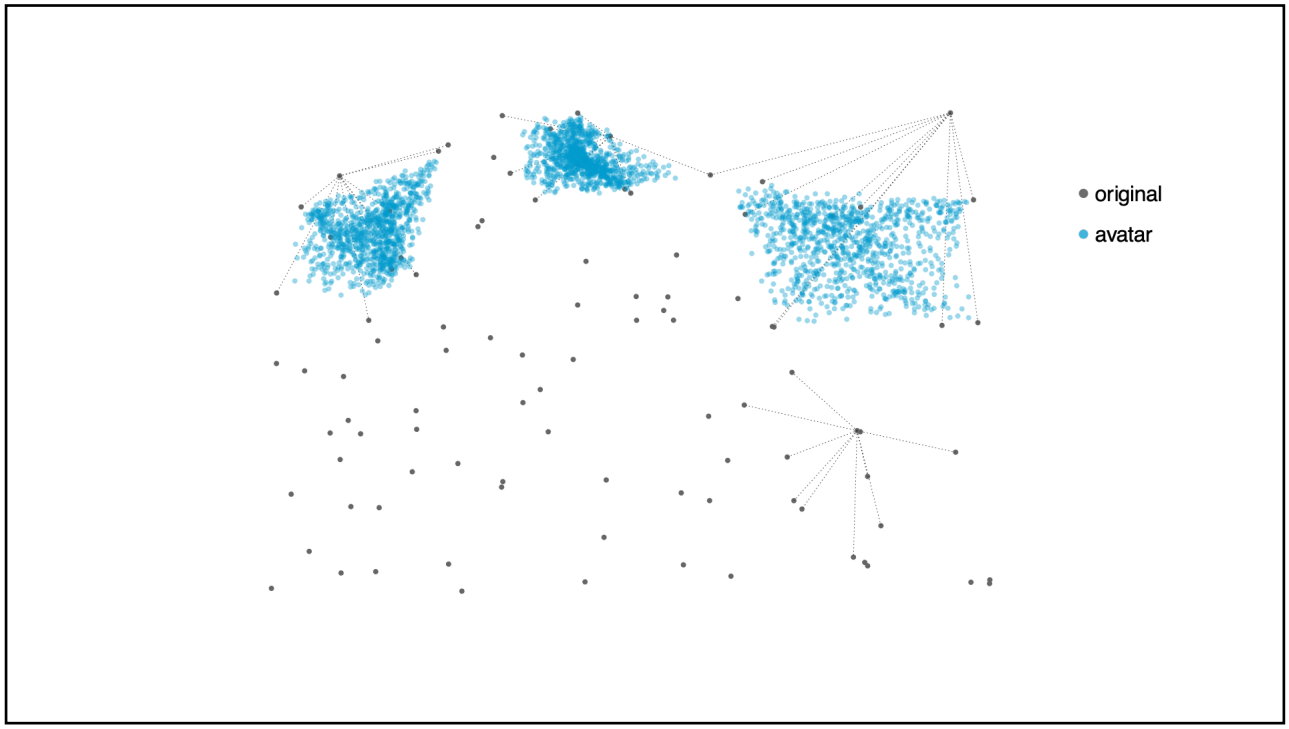

original
avatar

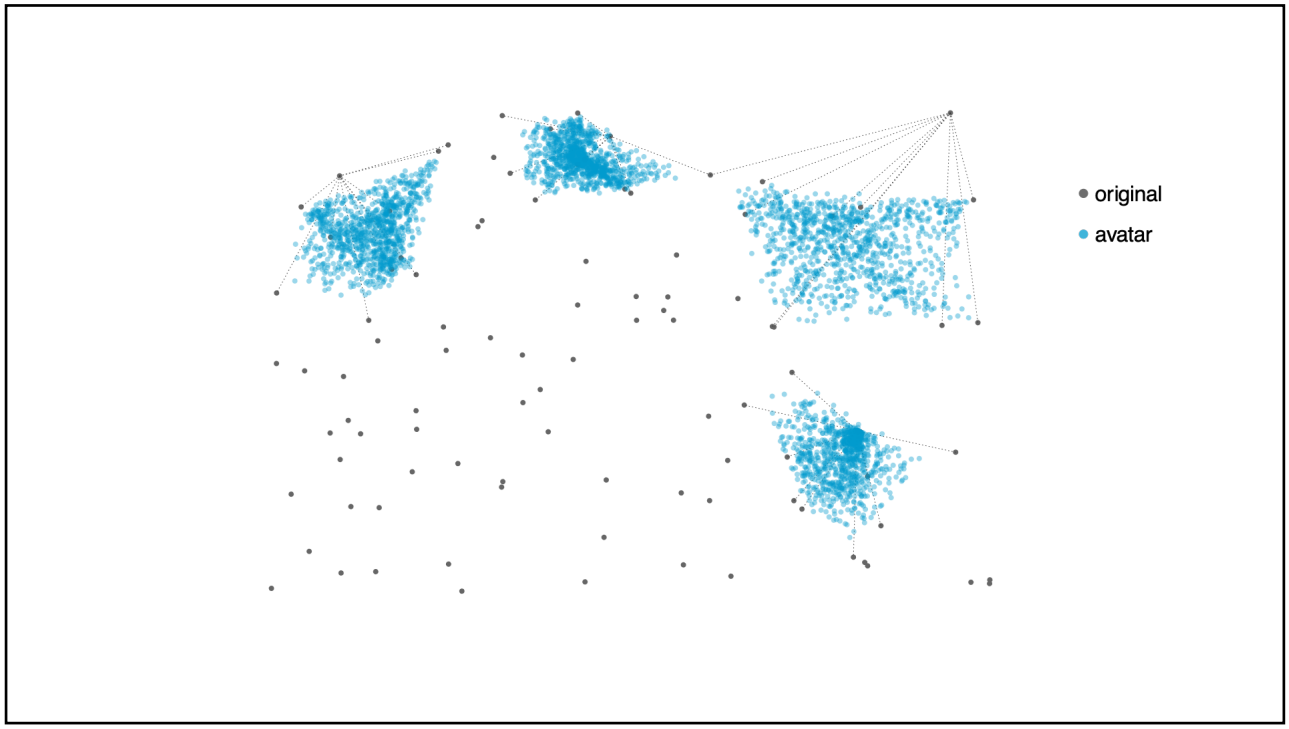

original
avatar

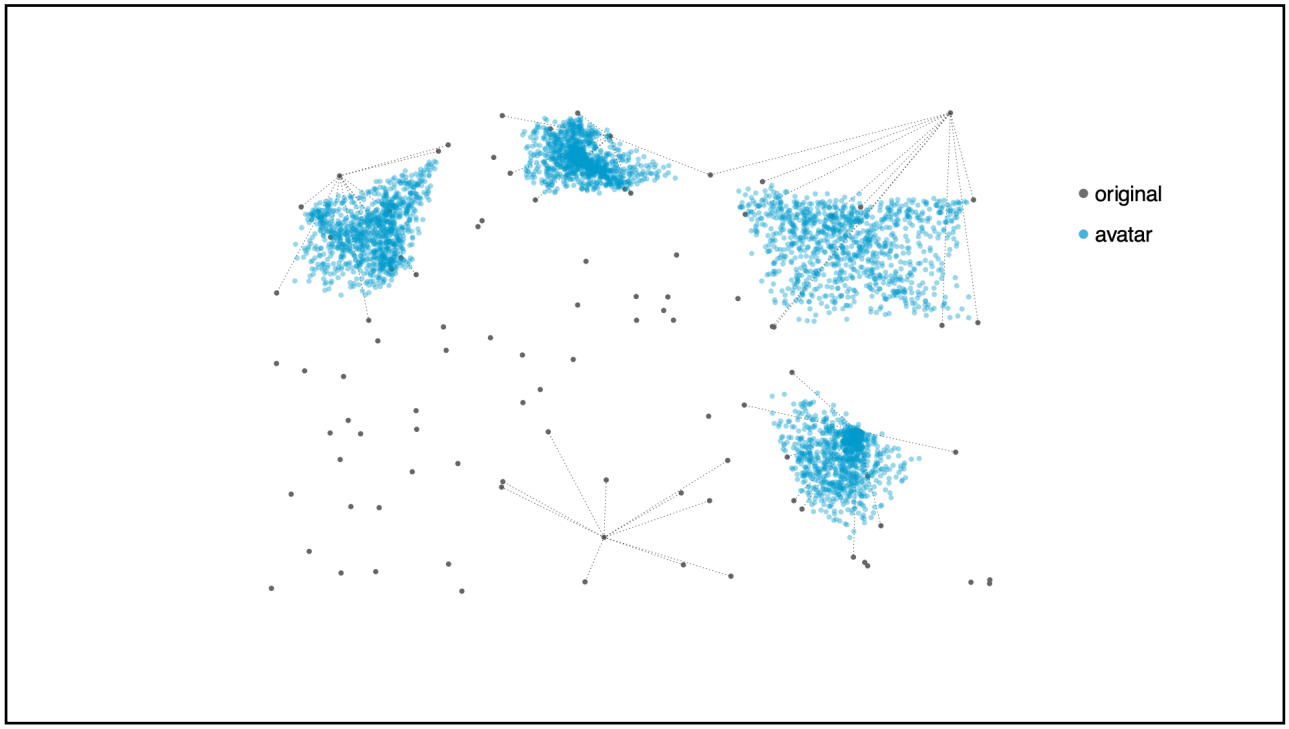

original
avatar

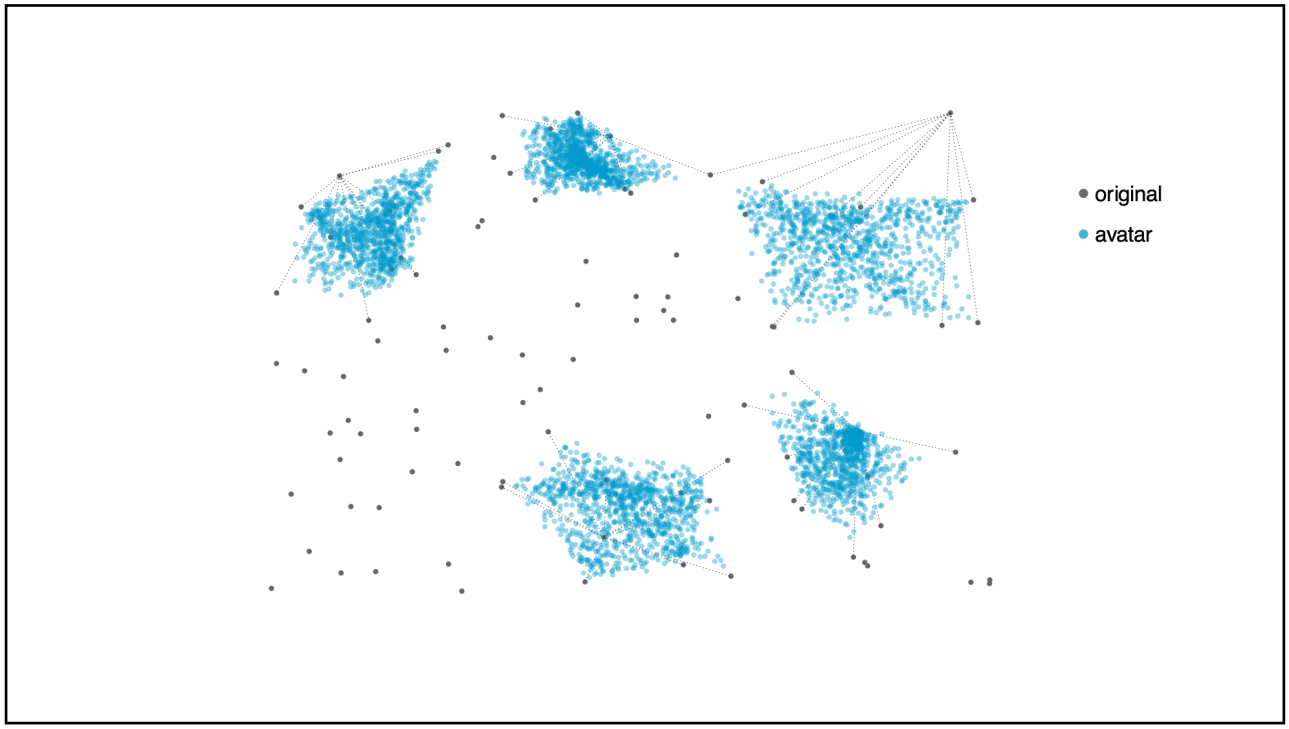

original
avatar

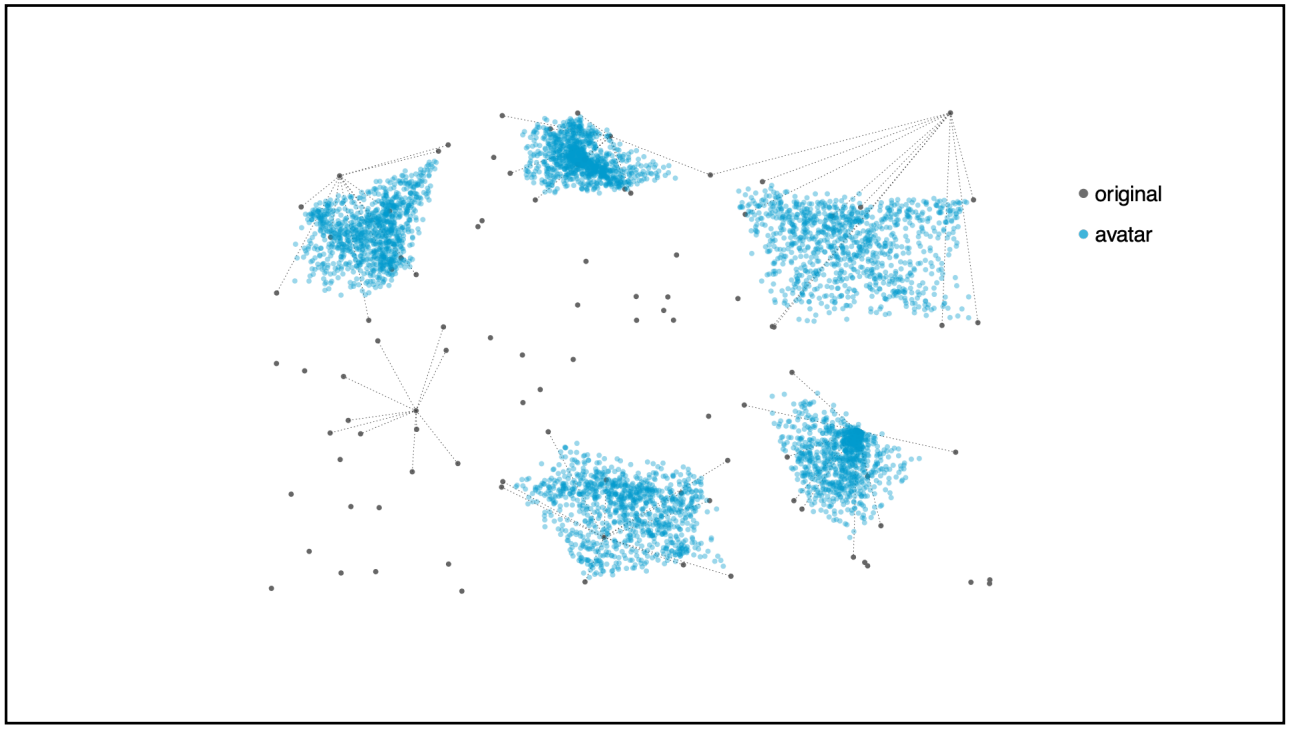

original
avatar

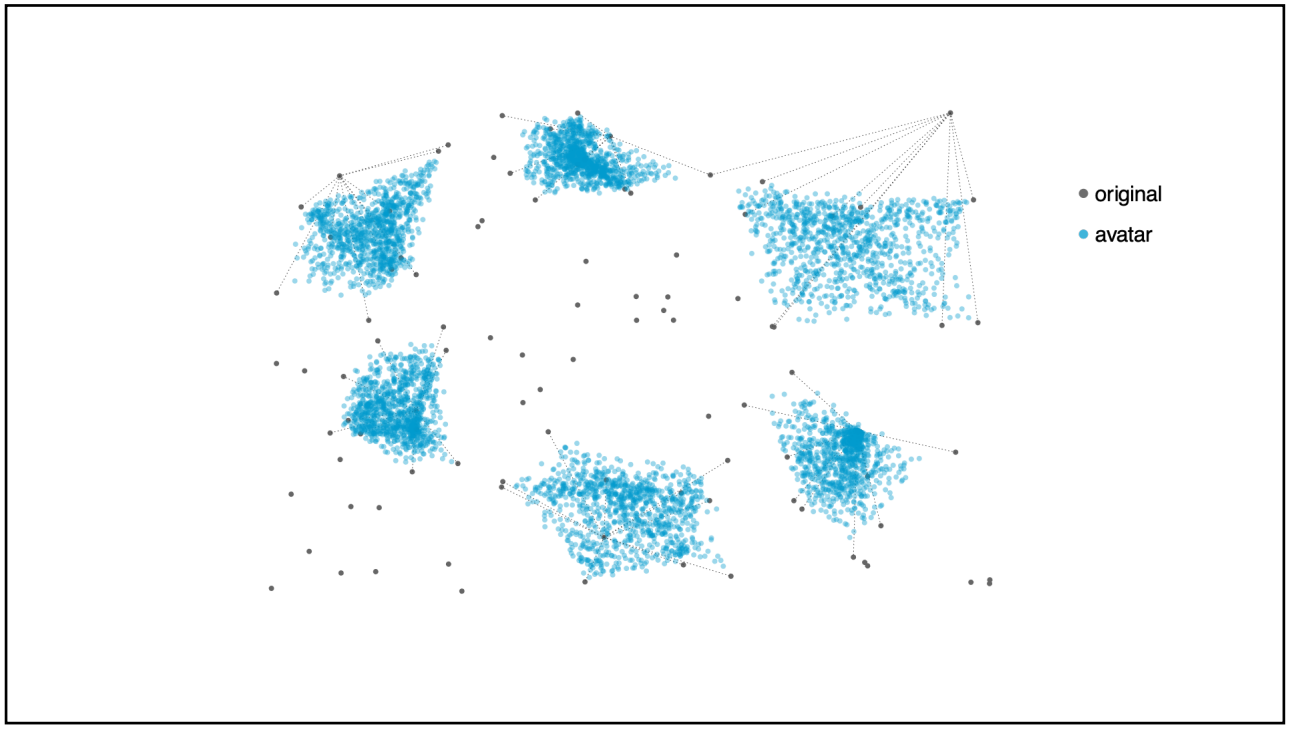

original
avatar

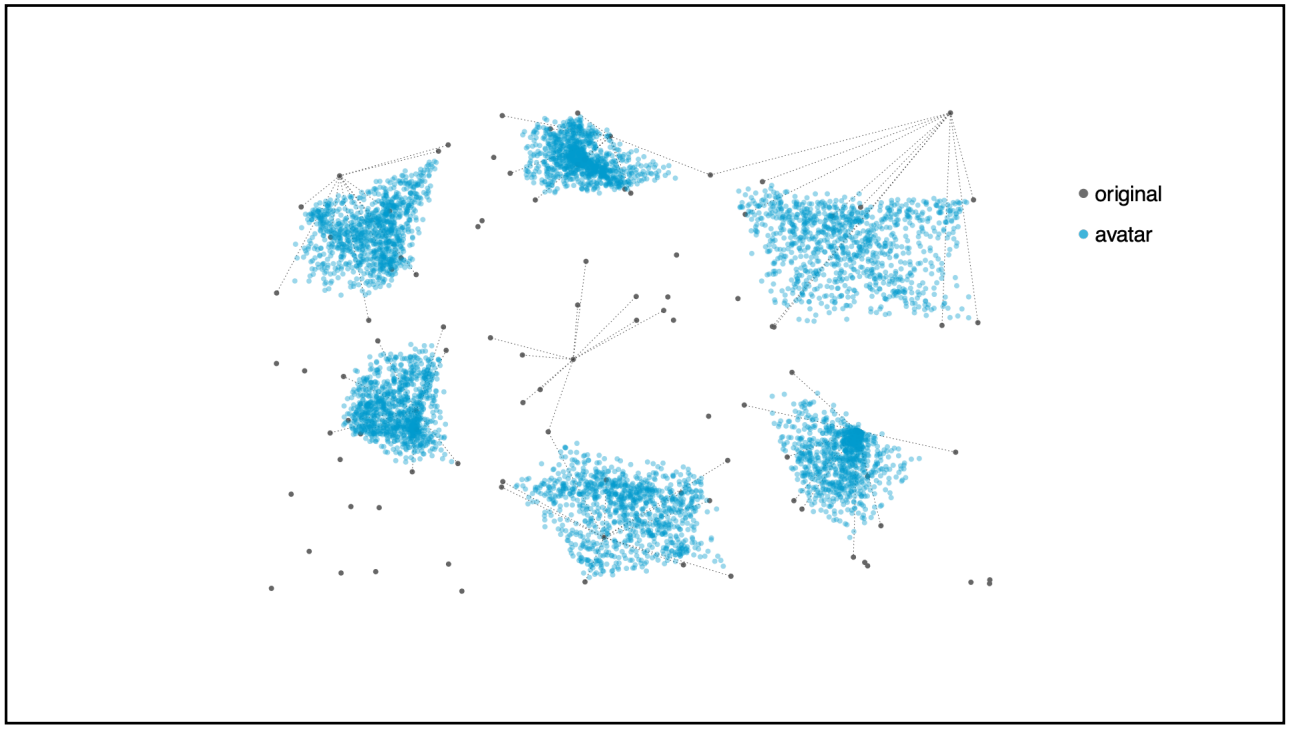

original
avatar

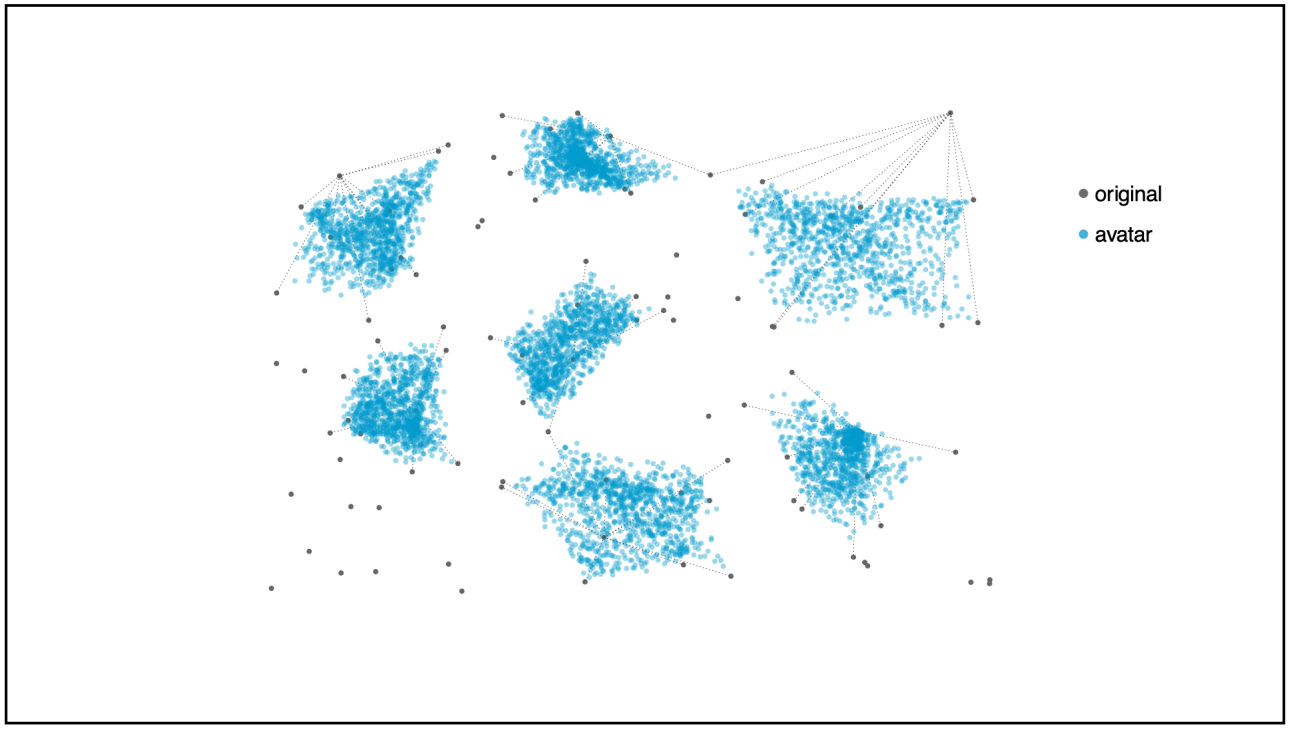

original
avatar

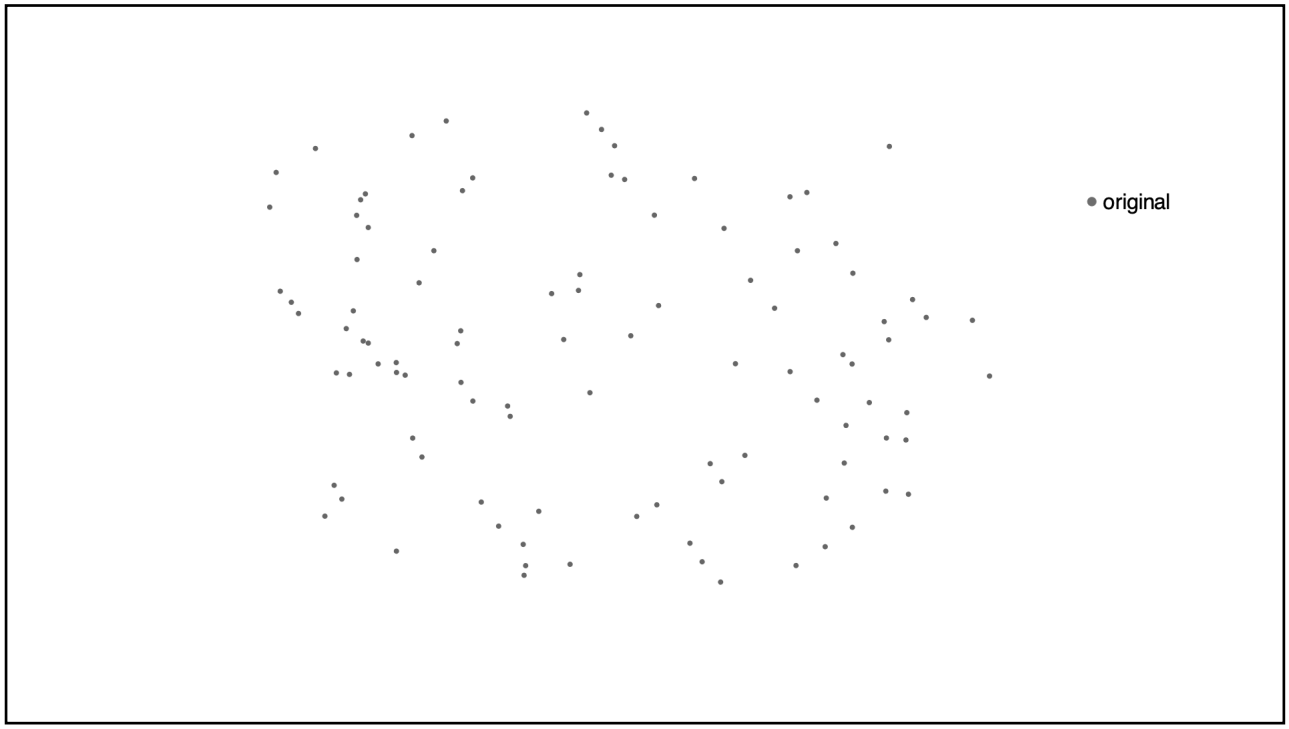

original
avatar

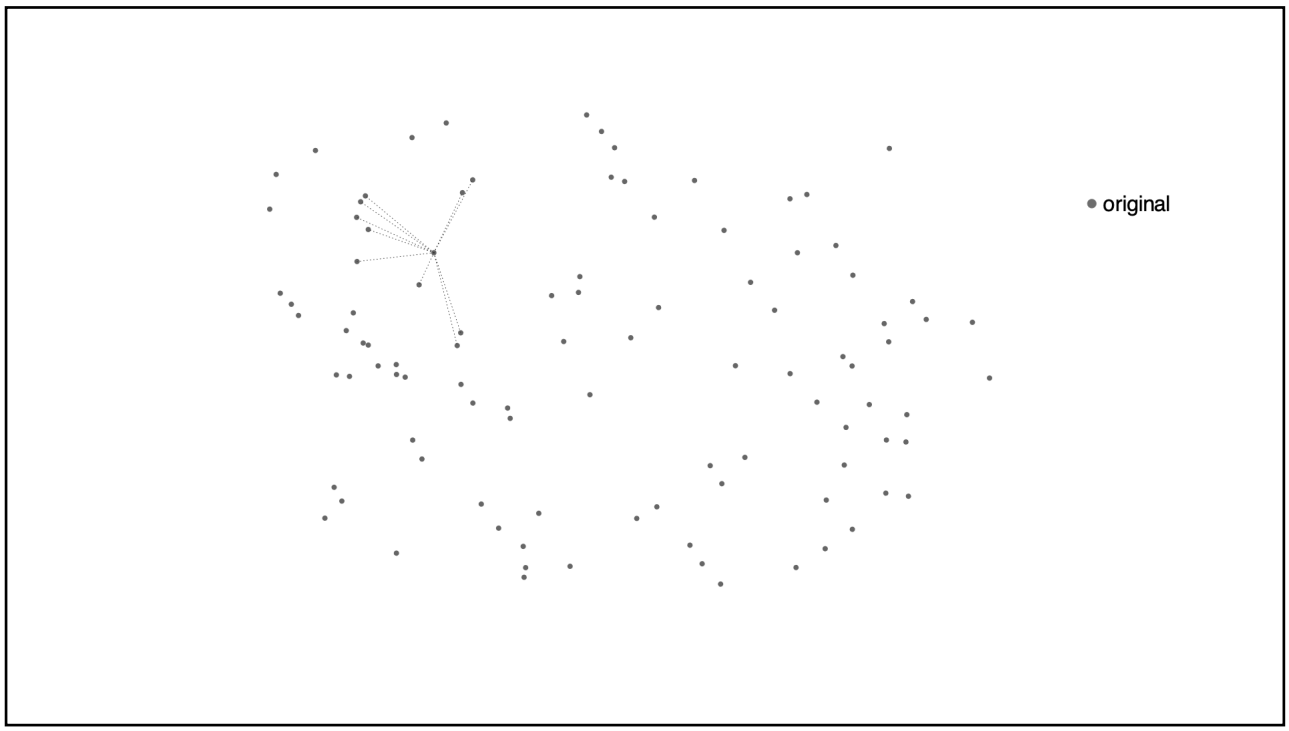

original
avatar

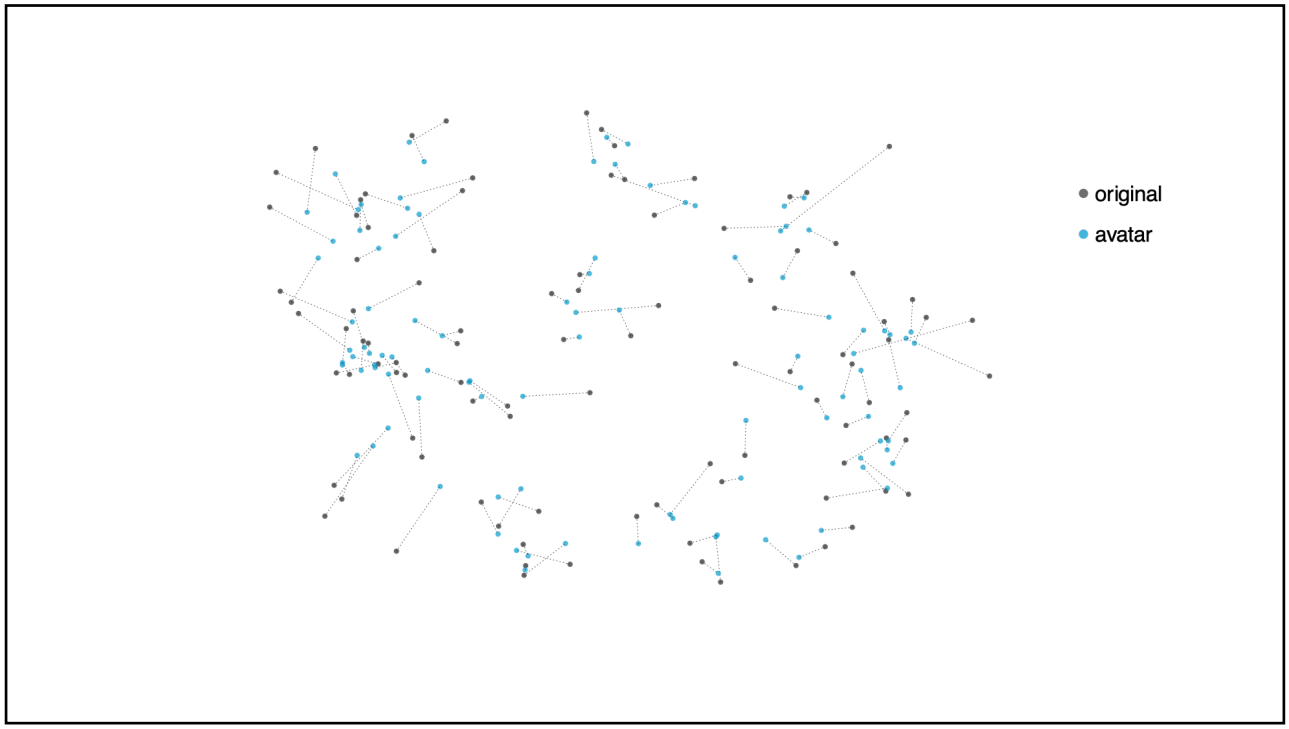

original
avatar

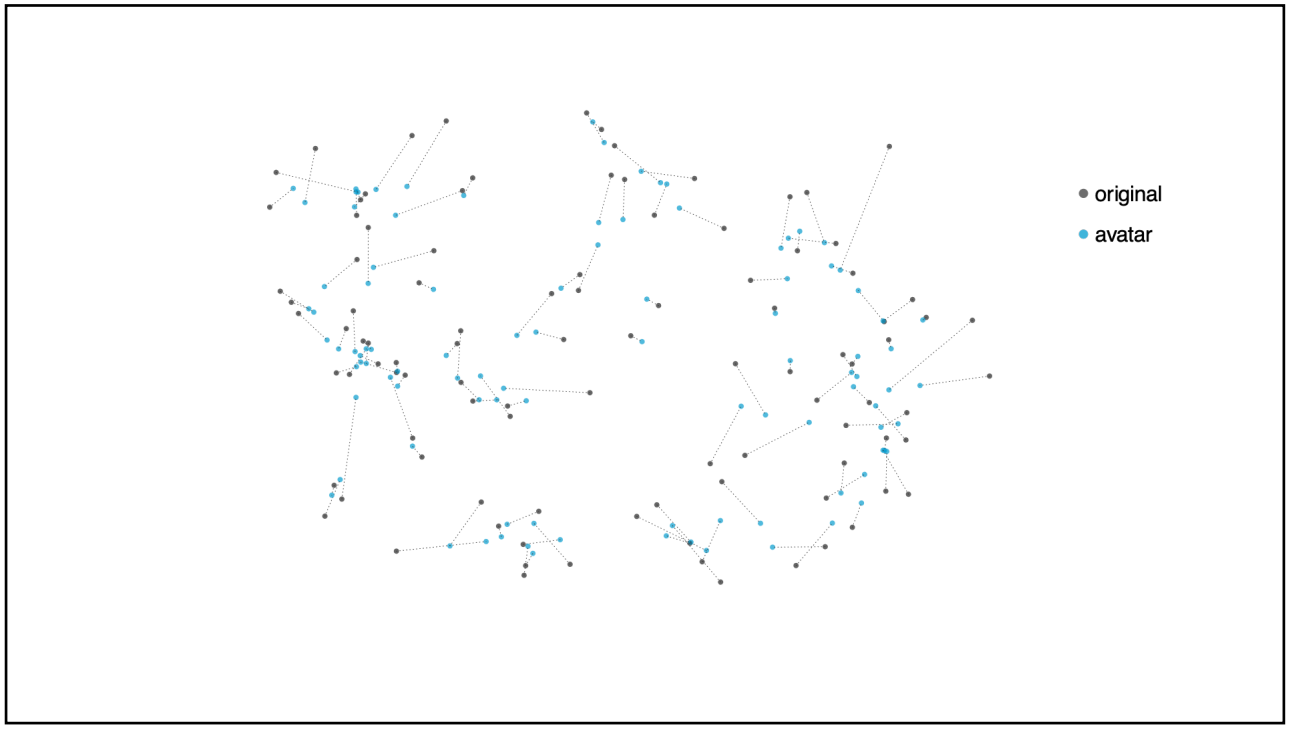

original
avatar

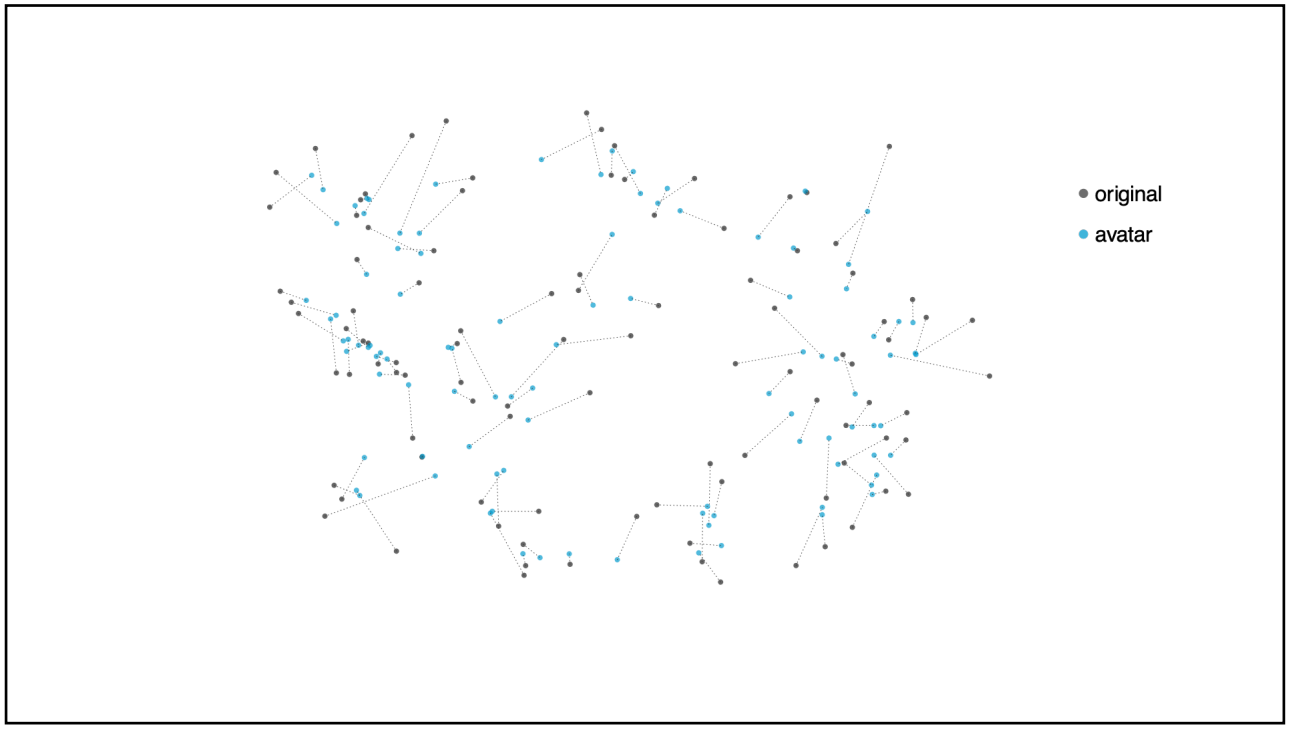

original
avatar

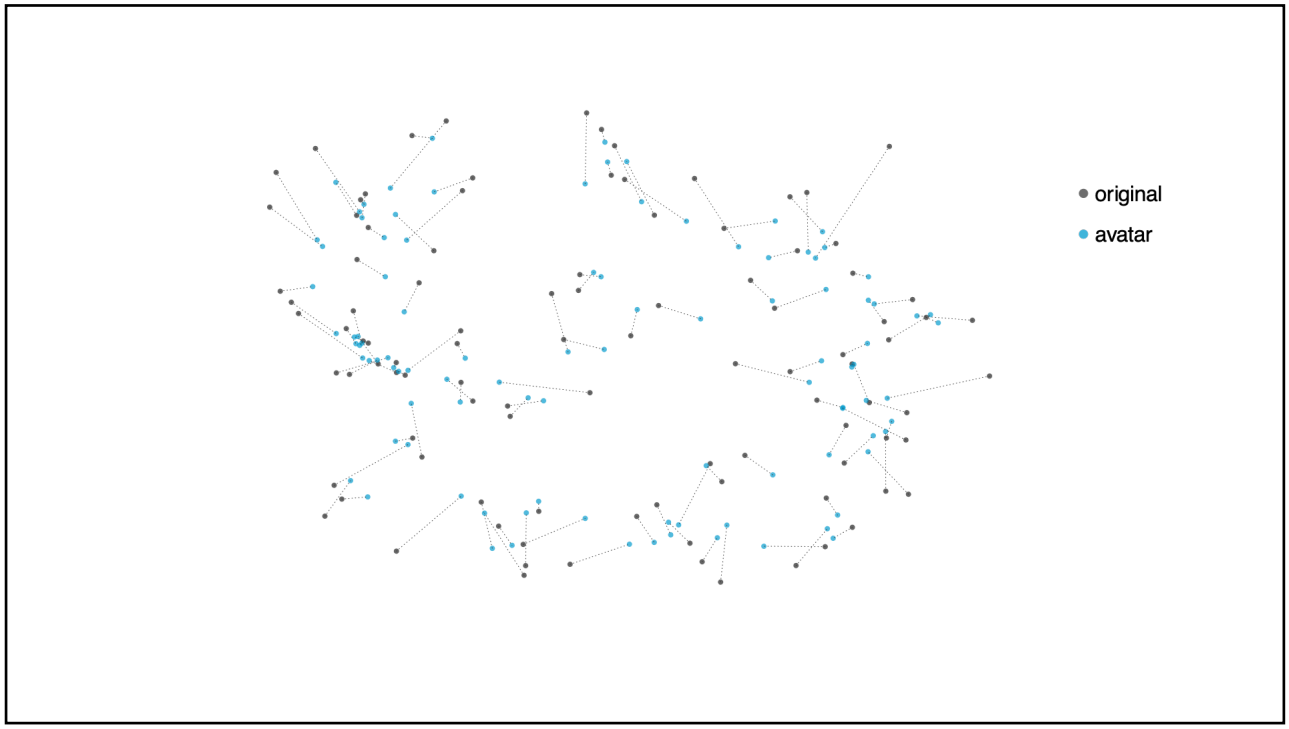

original

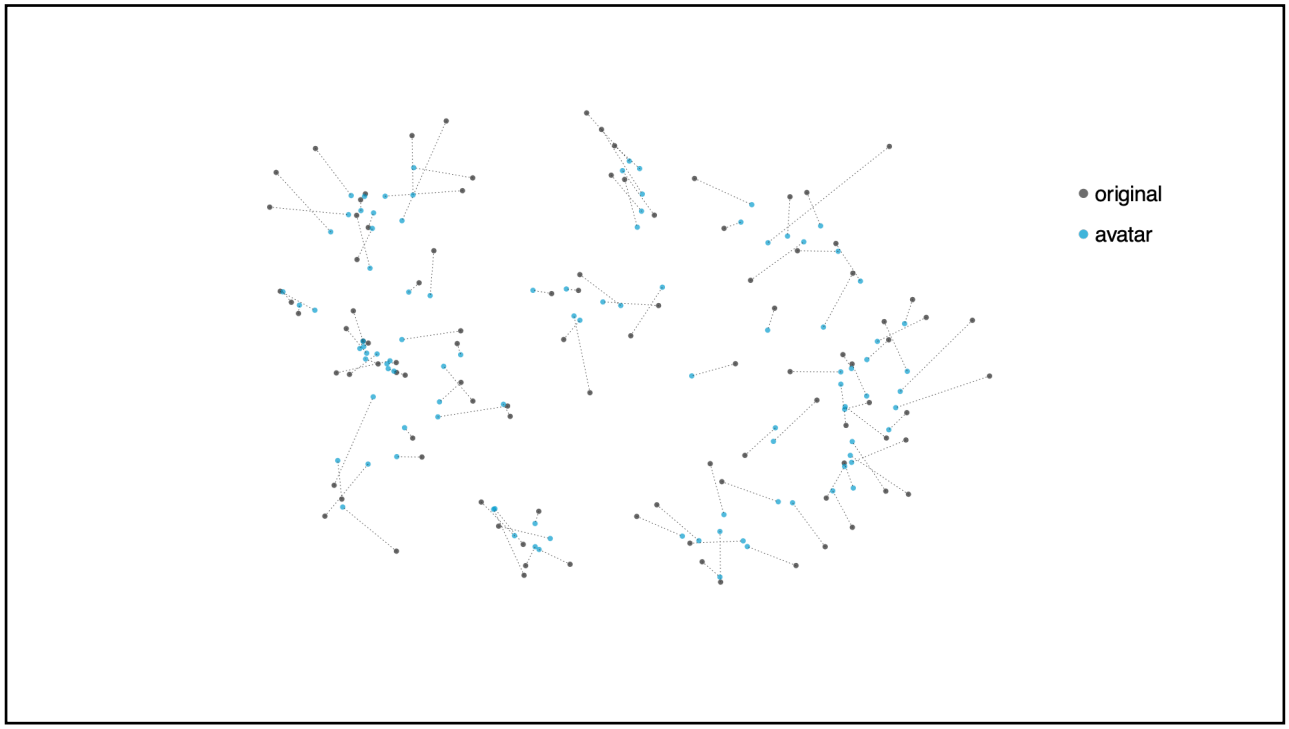

original

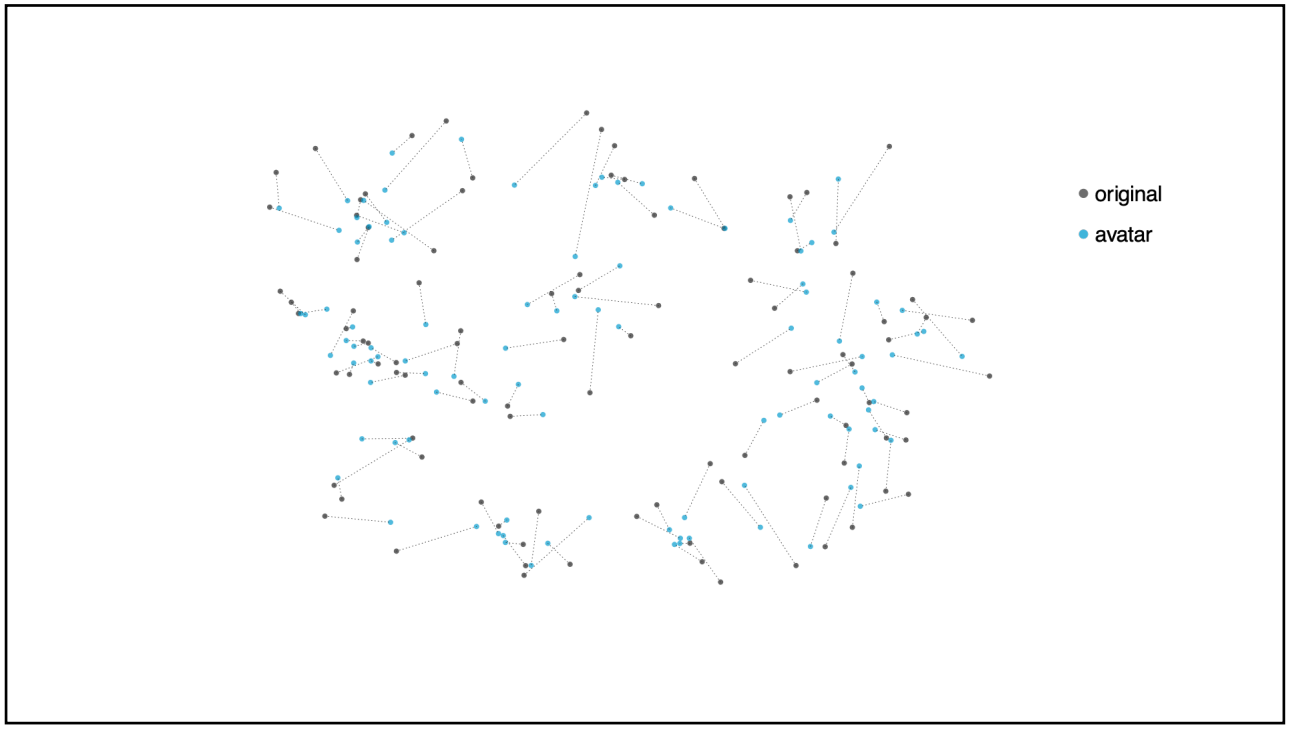

original
avatar

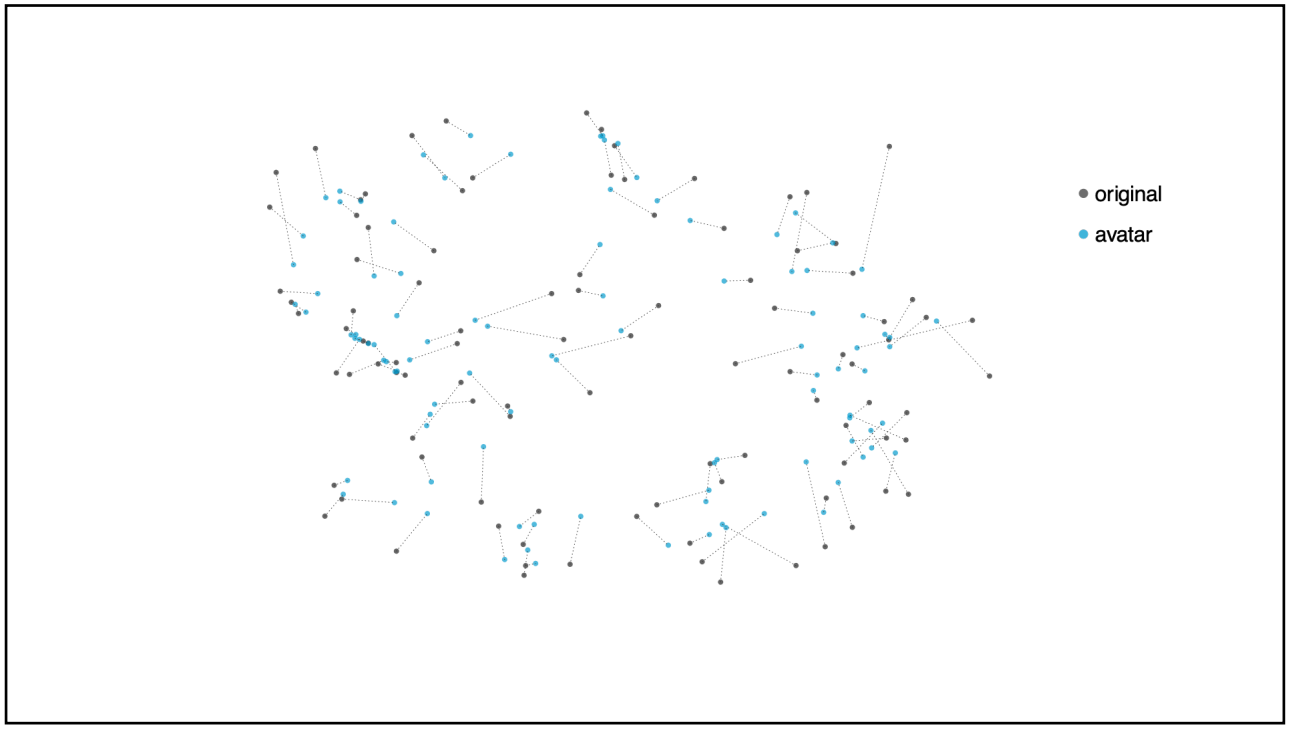

original
avatar

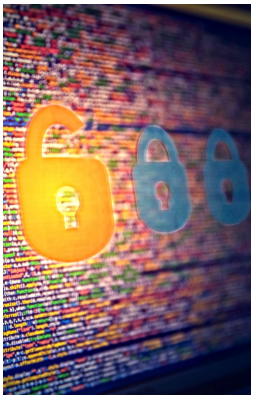

original
avatar


original
avatar


original
avatar


original
avatar


original
avatar



PRIVACY is the key concern

- Evaluated through quantitative metrics

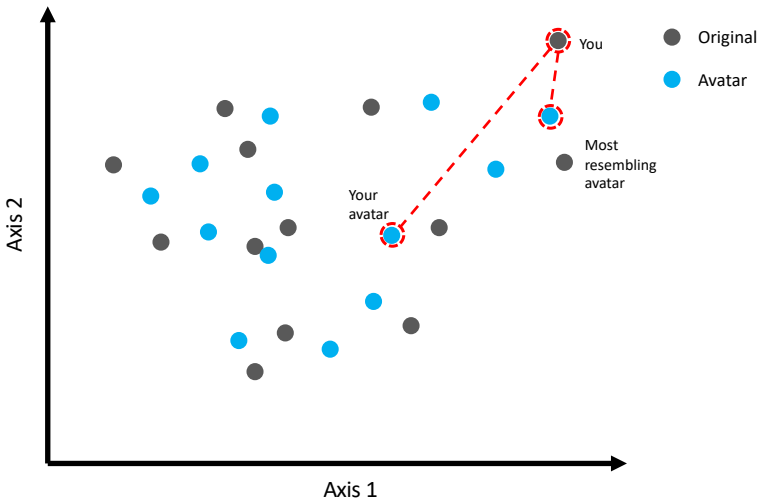
Before: Identify most at-risk records

During: define avatarization parameters

After: verify that the link between the record and the avatar is unmanageable to establish

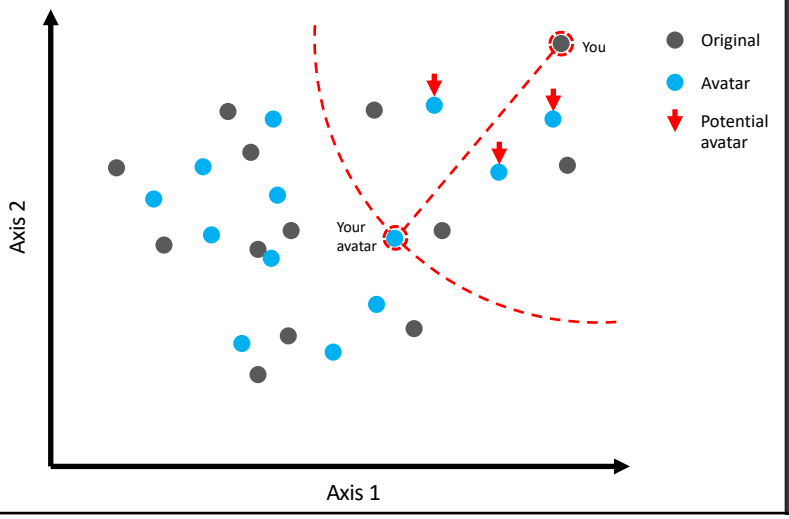
53

RISK ASSESSMENT



- Avatars **to quantify the risk** thanks to metrics
- These metrics **can't be used with other method**
- **Right First Hit** = protection percentage
- **Proof** for data owner that resembling avatar not been generated from his data

RISK ASSESSMENT



- Radius KNN = local cloaking
- The more avatars inside the radius, the stronger is the protection

AIDS ~ EXAMPLE 1 ~

BACKGROUND

Aids dataset

Trial comparing nucleoside monotherapy with three other combination therapy in HIV-infected adults

Dataset:

- 2139 individuals in the cohort
- 27 variables

57

COHORT

Aids dataset

wtkg	hemo	homo	drugs	cd40	...	cd420	days	arms
89.8128	FALSE	FALSE	FALSE	422	...	477	948	2
49.4424	FALSE	FALSE	FALSE	162	...	218	1002	3
88.452	FALSE	TRUE	TRUE	326	...	274	961	3
85.2768	FALSE	TRUE	FALSE	287	...	394	1166	3
66.6792	FALSE	TRUE	FALSE	504	...	353	1090	0

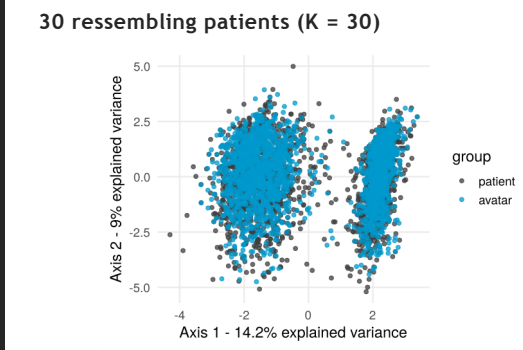
https://clinicaltrials.gov/ct2/show/NCT00000625

58

RESULTS

Aids dataset

Individuals with **marginal values** and by essence easy to re-identify are **protected** after avatarization



59

PRIVACY

Aids dataset

Almost 95% of records do not produce the closest avatar

On average, **a record is surrounded by 14 avatars:**

It is therefore **unmanageable** for an attacker to **establish** the link between the record and its avatar

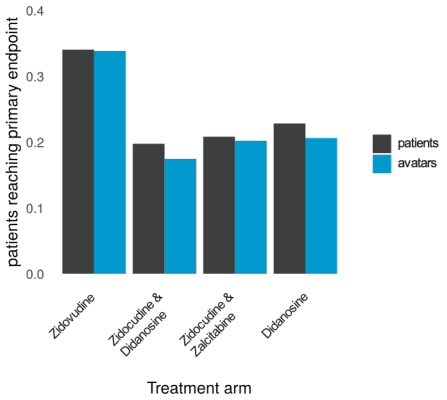
60

RESEARCH

Aids dataset

- The same conclusions can be drawn from the avatar dataset

Patients reaching primary endpoint depending on treatment



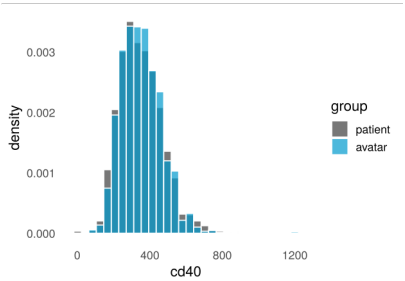
61

DISTRIBUTIONS

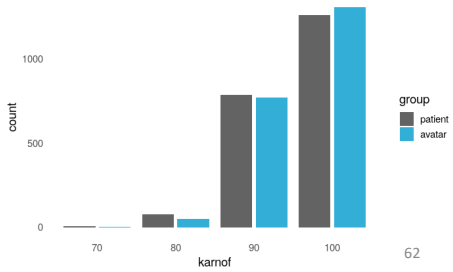
Aids dataset

- **Distributions** are also mainly conserved
- Marginal values are recentered

Continuous variable



Categorical variable



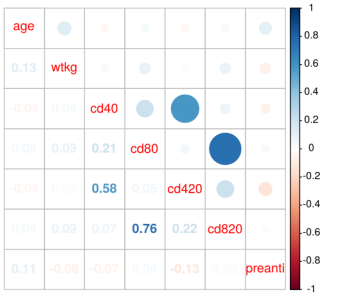
62

RELATIONS

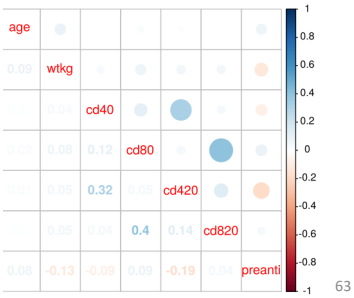
Aids dataset

- **Correlations** between variables remain comparable
- The avatarization seems to have **preserved signal** of the original dataset

Original dataset



Avatars dataset



63

BODY FAT
~EXAMPLE 2~

BACKGROUND

The objective of this study was to **evaluate** the calculation of percentage body fat from other simple body measurement

Dataset:

- 252 individuals in the cohort
- 15 variables

65

COHORT

Body fat dataset

siri	age	weight	height	adipos	...	chest	abdom	hip
22.8	31	148	70.75	21.6	...	88.5	83.5	94.5
14.9	68	179	68.25	27.2	...	100	91.6	96.4
27.2	34	205.75	71.25	27.6	...	105.2	98.8	108.3
25.3	22	176.75	70.5	25.4	...	101.8	87.2	98.5
28	52	219.15	74.25	26.1	...	107.5	104.2	107.5

Fitting Percentage of Body Fat to Simple Body Measurements, RW Johnson, Journal of Statistics Education, 1996

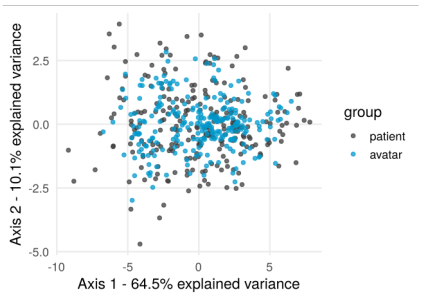
66

RESULTS

Body fat dataset

Individuals with **marginal values** and by essence easy to re-identify are **protected** after avatarization

25 ressembling patients (K = 25)



67

PRIVACY

Body fat dataset

Almost 95% of records do not produce the closest avatar

On average, **a record is surrounded by 7 avatars:**

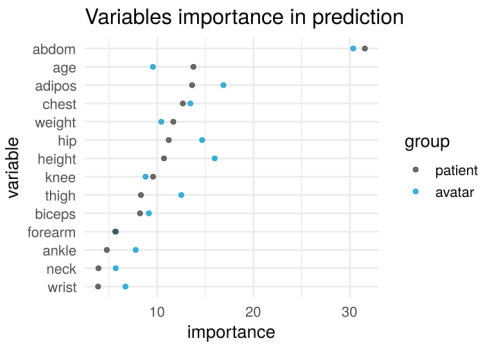
It is therefore **unmanageable** for an attacker to **establish** the link between the record and its avatar

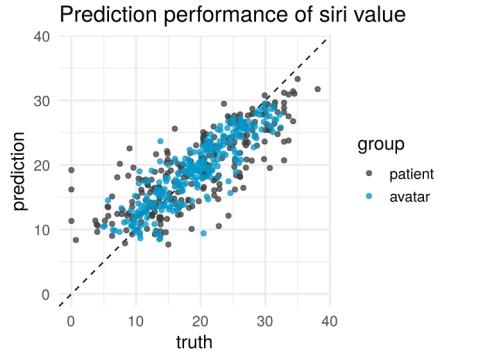
68

RESEARCH

Body fat dataset

- Variables importance in prediction are comparable
- Prediction performance remains



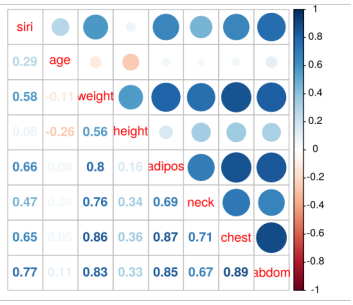


RELATIONS

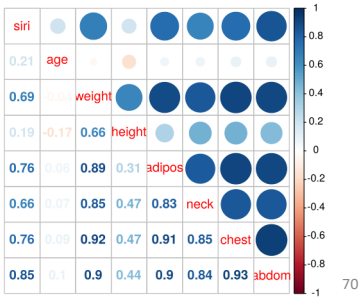
Body fat dataset

- **Correlations** between variables remain comparable
- The avatarization seems to have **preserved signal** of the original dataset

Original dataset



Avatars dataset



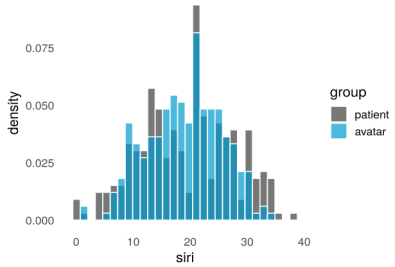
70

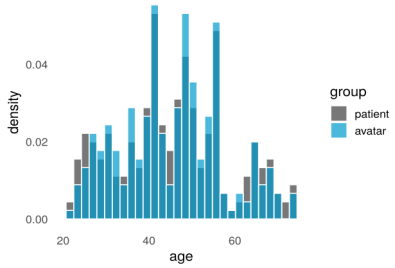
DISTRIBUTIONS

Body fat dataset

- **Distributions** are also mainly conserved
- Marginal values are recentered

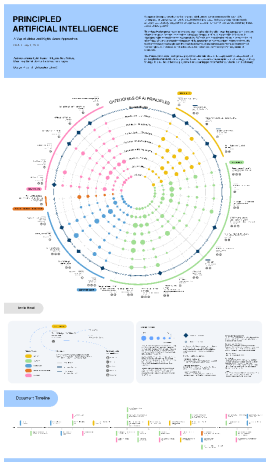
Continuous variable only





71

Conclusion : Bottom-up “ethics by-construction”



https://ai-hr.cyber.harvard.edu/primp-viz.html

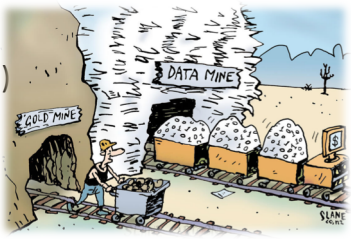
- *The systematic use of synthetic data (avatars) in our biomedical data warehouse is an ethical question of governance.*
- *Over 70 top- down approaches on AI ethics*
  - *Finish line ?*
- Collective building and revision of a set of ethical principles
  - Starting point for diverse public-private community
  - Contradictory principles promote ethics by construction

…

NAONEDIA

Conclusion

- **Les Technologies créent des opportunités techniques**
  - Tension existentielle dans le saisissement de ces outils.
- **Un rôle pour les données de santé ?**
  - Exceptionnalisme? Exemplarité ?
  - Intentionalité, consentement, transparence ?
  - Santéisme ?
- **Les entrepôts de données biomédicales**
  - "Mine" de données (multiples jeux de mots possibles)
  - Usage secondaires, non intentionnels
  - Opportunité de Soli-data-rité



"La Clinique des données"

Pr *Pierre-Antoine GOURRAUD*

« Un check-up de l'IA pour la santé numérique, c'est grave docteur(s) ? »

Séminaire Aristote 27 février 2020,

École Polytechnique, Amphithéâtre Arago
91120 Palaiseau

ATIP-Avenir Team 5 «Translational Immunogenomics of Transplantation and Autoimmunity »,
ITUN - CRTI - UMR Inserm 1064 -CHU de Nantes

Pôle Hospitalo-Universitaire 11 : Santé Publique, Santé au Travail et Pharmacie,

Hôpital St-Jacques - CHU de Nantes

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale