Cette cinquième semaine traite la question de la gestion de données respectueuse de la vie privée dans les villes intelligentes.

La première séquence traitera des architectures proposées pour gérer ces données face au respect de la vie privée. Ensuite, nous aborderons la gestion de la vie privée dans les réseaux sociaux, puis la gestion des données confinées, embarquées dans des puces et des capteurs en support au privacy by design. Ensuite, la gestion de la vie privée dans les applications mobiles participatives pour terminer par les traitements de données globaux respectueux de la vie privée.

Cette première séquence traite des architectures de gestion de données face au respect de la vie privée.

Une ville intelligente par définition vise à numériser tous les procédés analogiques, mécaniques, procédures papier et les procédés de communication.

Ça conduit inévitablement à générer des masses de données très importantes sur la ville, sur son activité et concernant ses citoyens.

Cette génération de données est bien sûr sous-jacente à tous les services urbains qu'on souhaite mettre en place.

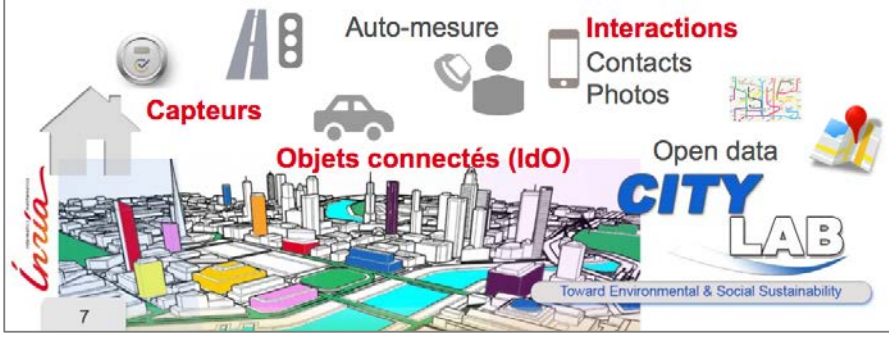
Du point de vue de la gestion de données, ces services qu'on souhaite offrir aux citoyens se basent sur des croisements de données :

- des croisements de données publiques mises en accès libre en open data par les villes,

- des données personnelles qui sont produites par chacun des citoyens urbains, que ces données soient collectées automatiquement ou délivrées de manière participative par les citoyens

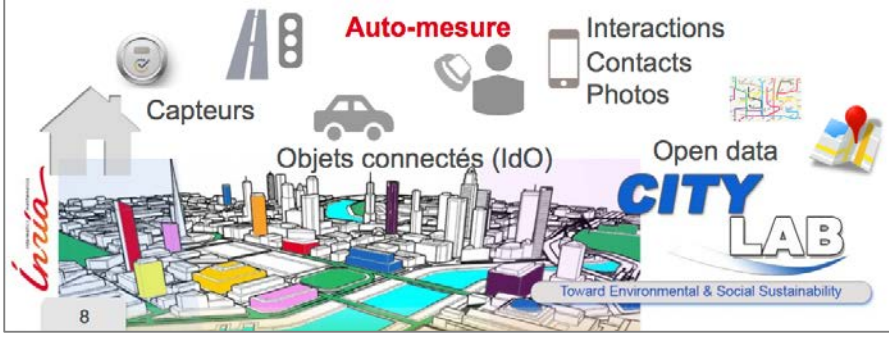
Les données dans les villes intelligentes

- Numérisation des procédés ➔ génération massive de données personnelles
- Des données sur la ville, son activité, ses citoyens ➔ services urbains
  - ✓ croisement de données publiques, personnelles, collectées automatiquement ou de manière participative

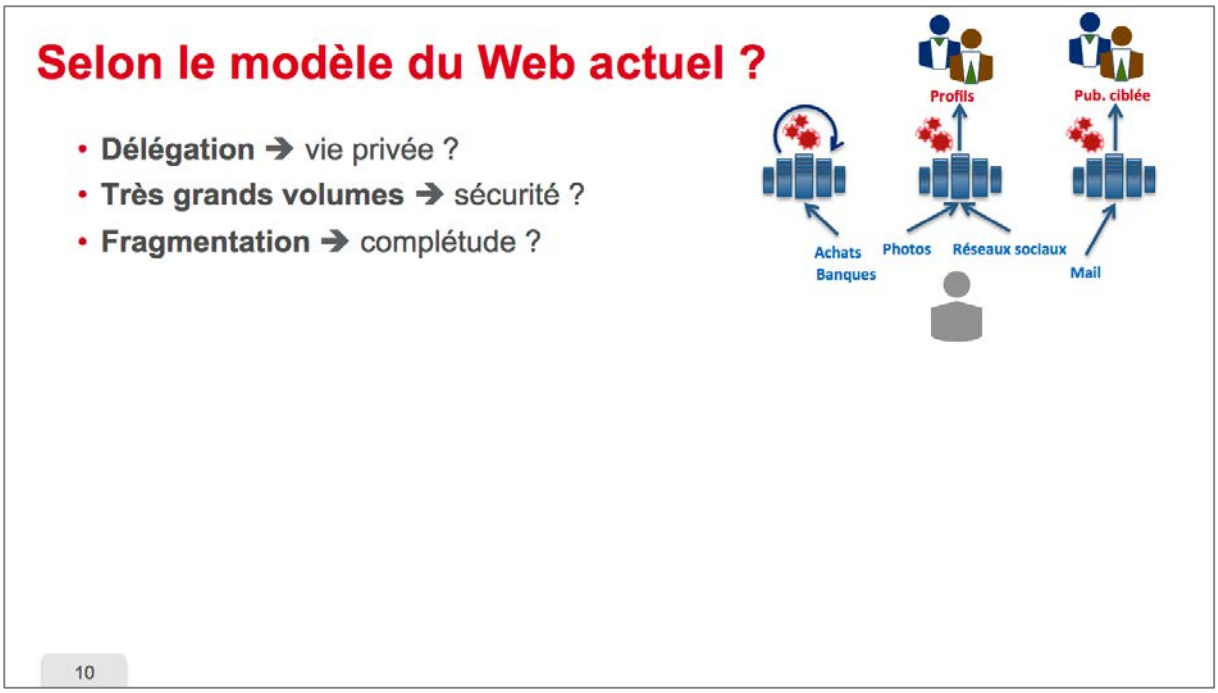


Les données dans les villes intelligentes

- Numérisation des procédés ➔ génération massive de données personnelles
- Des données sur la ville, son activité, ses citoyens ➔ services urbains
  - ✓ croisement de données publiques, personnelles, collectées automatiquement ou de manière participative



Naturellement, la question se pose de savoir **comment gérer ces données dans le respect de la vie privée des citoyens urbains**.



La première solution serait de gérer les données de la ville selon le modèle du web actuel. Comment ce modèle fonctionne-t-il ?

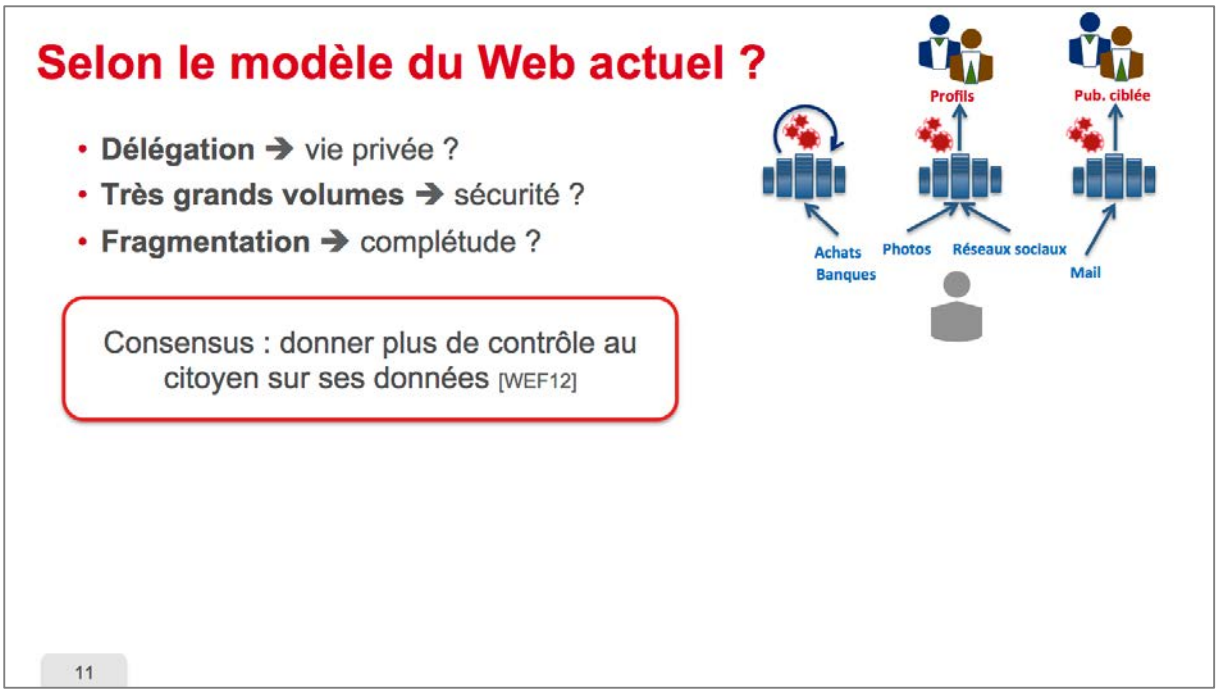
De façon synthétique, les données sont produites par les individus et sont extraites via les applications en ligne, pour être stockées finalement dans des data centers organisés sous forme de silos de données. L'avantage pour l'individu est que ces données sont durables et accessibles.
Mais ce modèle a **3 caractéristiques intrinsèques qui posent problème** aujourd'hui, notamment pour la vie privée.

D'abord le modèle est basé sur le principe de la délégation.
L'usager n'est pas dépositaire des données, c'est celui qui les collecte, c'est le silo, qui en est propriétaire et l'usager n'a souvent pas accès au contenu de toute l'information le concernant.
De plus, il n'a pas la garantie que les usages qui seront faits de ses données seront clairement exposés. Elles pourraient être l'objet de passe-droit et puis seront accédées par des partenaires commerciaux pour réaliser des usages secondaires souvent cachés.
Tout ça à l'insu des usagers pour satisfaire des modèles d'affaires sous-jacents.
La question qui se pose est quels acteurs pourraient être habilités à jouer le rôle de tels gestionnaires de données ?

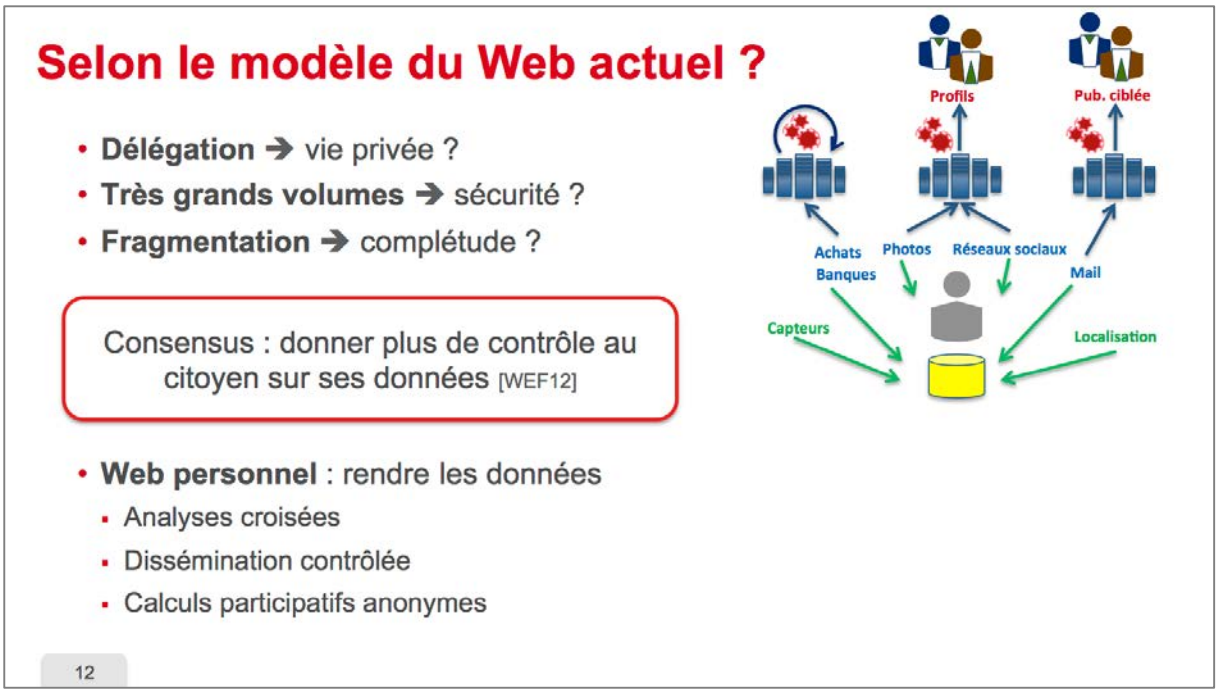
La deuxième caractéristique intrinsèque, c'est la concentration.
Une simple négligence ou une attaque conduit à affecter des millions d'enregistrements.
On assiste aujourd'hui à des attaques répétées, extrêmement sophistiquées, récurrentes, vu le bénéfice obtenu en cas de succès.
Le risque pour la vie privée peut être comparé à celui d'un risque nucléaire pour l'énergie.
La question est celle donc du coût à payer en mesure de sécurité pour pallier ce risque.

Troisième caractéristique intrinsèque, le cloisonnement.

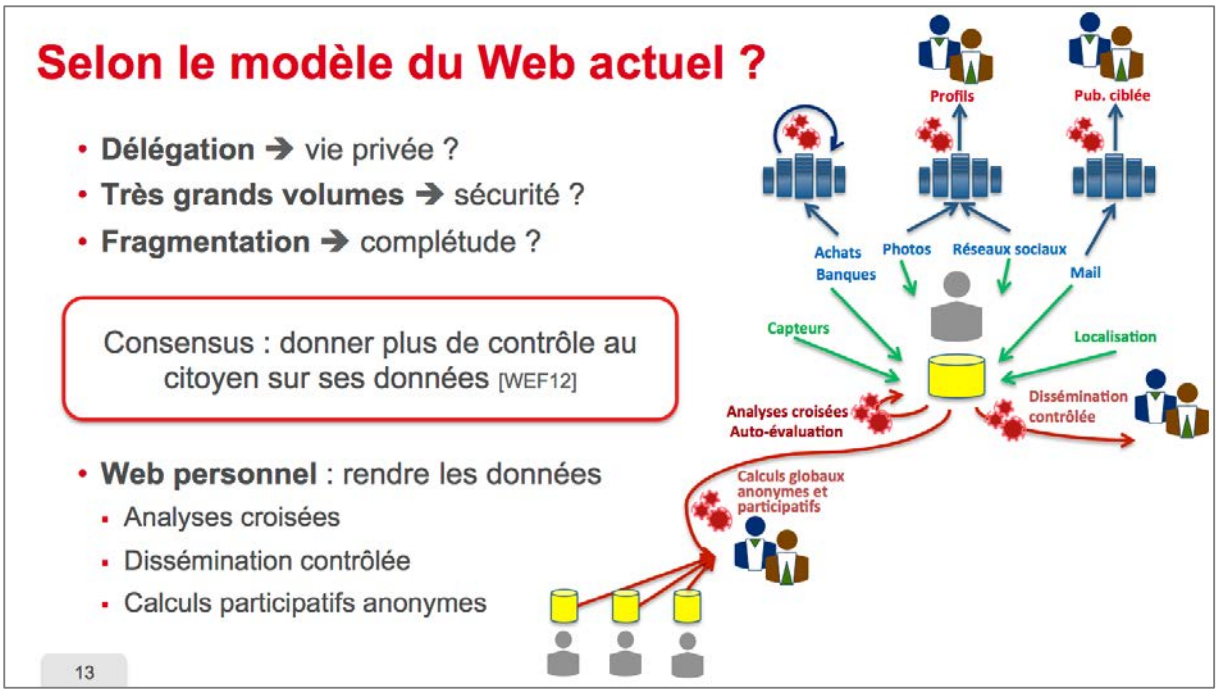
Chaque application nourrit un seul et unique silo de données et l'enjeu de chacun des acteurs du web va être celui de la complétude, c'est-à-dire que chaque silo va tenter d'accumuler le dossier le plus transverse et le plus complet possible sur les citoyens. Ça conduit à adopter des stratégies monopolistiques au sein des silos et à mettre en œuvre des techniques d'appareillement de données qui sont assez contestables et parfois à la limite de techniques de désanonymisation de données.



Face à ce constat, un consensus émerge au niveau du législateur en Europe, du monde économique - le Forum économique mondial et ses rapports récents en attestent - et des citoyens. **Donner aux usagers plus de contrôle sur leurs données devient nécessaire.**



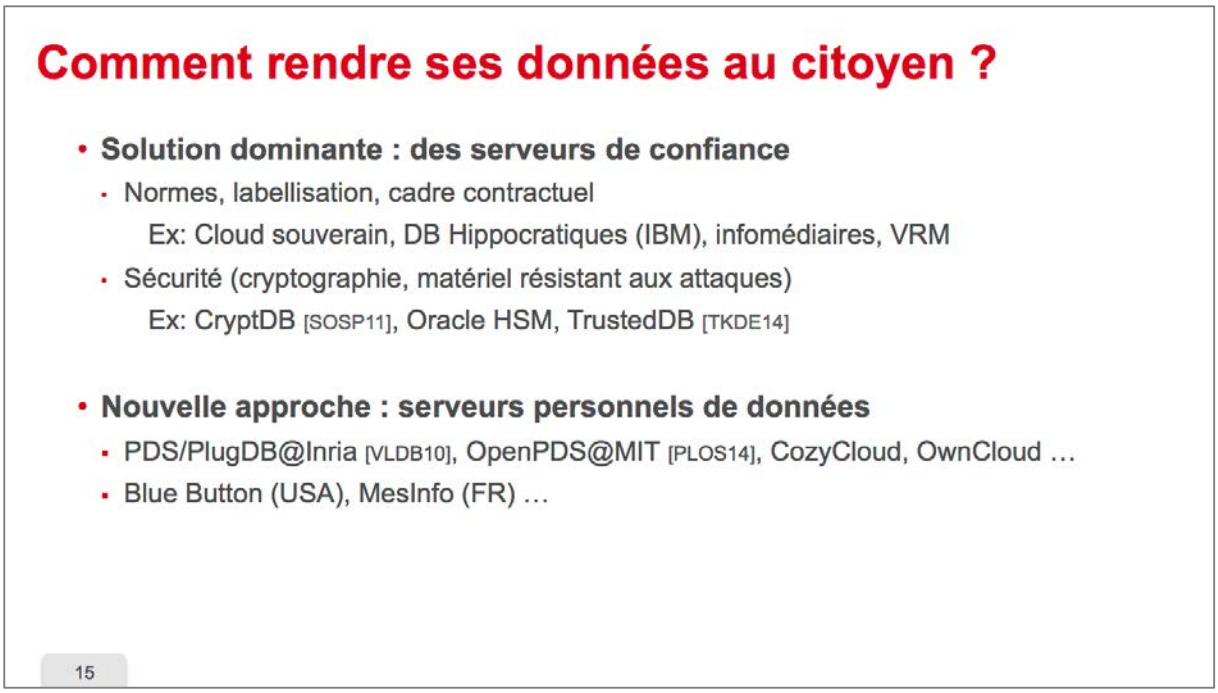
Le **modèle alternatif** qui se dessine actuellement est celui du **web personnel**.
Ce modèle peut cohabiter avec le web actuel ou avec des services en open data.
L'idée est de rendre leurs informations ou certaines informations aux usagers.
Des applications nouvelles pourront émerger et nourrir directement le web personnel.



Ce modèle conduit de facto à une information bien plus complète et ouvre à de nouveaux usages qui permettent à l'individu par exemple de conduire des analyses croisées sur ses propres données, dans la mouvance du quantify self, de l'auto-évaluation.

La vie privée est bien plus sous contrôle puisque l'individu peut réguler lui-même la dissémination de ses propres données. Il peut n'exposer que certains résultats plutôt que de donner un accès libre aux données de base qui permettent de les calculer.
Il peut aussi participer à des calculs globaux d'intérêt collectif sur la base du volontariat.

Rendre ses données à l'usager est une belle idée qui répond aux problèmes intrinsèques du modèle actuel. Mais comment faire ?
Si rien n'est fait, l'individu risque de devoir stocker ses données sur son **ordinateur personnel**. Il ne sera pas à même de les administrer correctement ni d'en réguler la dissémination.
S'il se tourne vers un **service cloud** avec le modèle du web actuel pour réaliser ça, ça revient à exacerber les problèmes mentionnés précédemment sous couvert de vie privée, ce qui serait paradoxal.
Il faudrait donc fournir aux usagers une autre solution. Les **principes fondateurs des architectures envisagées actuellement** sont **centrés sur l'individu**, sur la **propriété des données** - donc c'est l'individu qui doit en être propriétaire - et sur la **fourniture de services respectueux de la vie privée**.

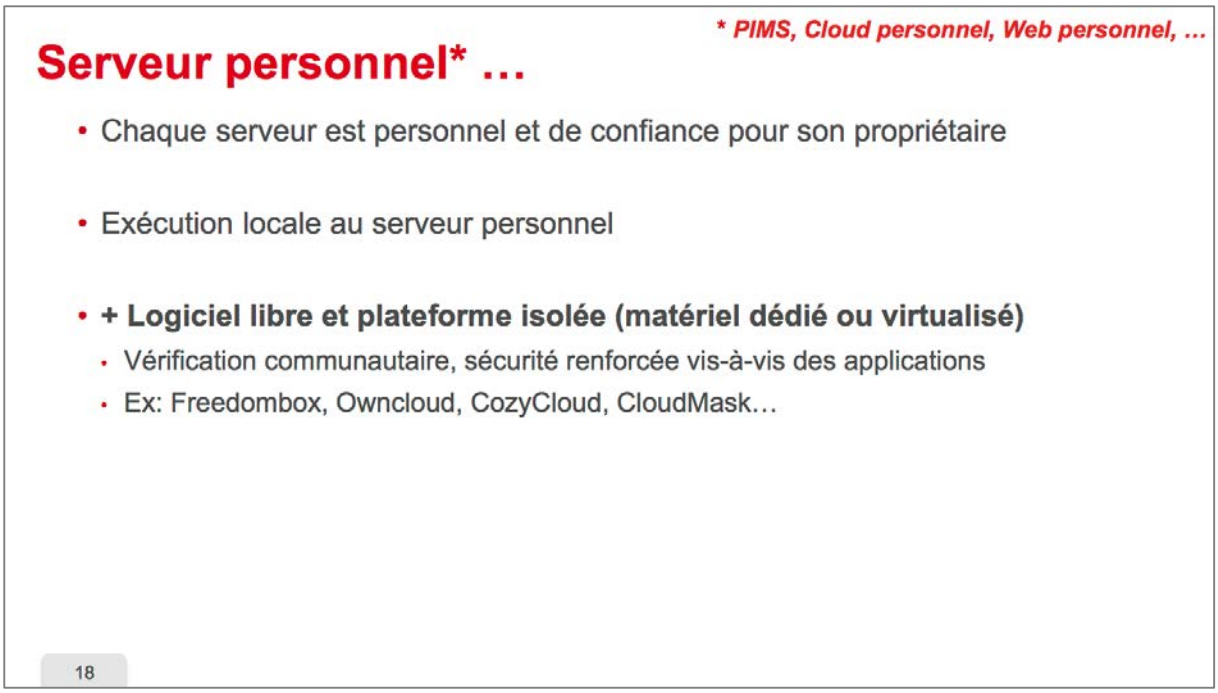


L'objectif est de se mettre dans un cadre qui est plus propice à une application de la législation et des recommandations des CNIL en matière de vie privée.
Pour les grands éditeurs du web, ça passera par la mise en place de **serveurs de confiance**. D'où viendrait cette confiance ?
De normes, de processus de labellisation, de critères communs dans un cadre contractuel et de modèles d'affaires plus propices, voire de consortiums publics-privés de confiance qui sont envisagés par exemple dans le cas du cloud souverain.
Les industriels comme IBM proposent aussi des serveurs de données implantant par exemple pour la santé des propriétés de respect de la vie privée spécifiques et conformes à la législation.
Des outils cryptographiques ou des composants offrant une sécurité matérielle peuvent aussi être adjoints au serveur. C'est proposé par Oracle, CryptDB au MIT est une solution qui se base aussi sur le chiffrement ou TrustedDB à l'université Stony Brook qui propose de gérer les données dans un serveur ultra sécurisé adjoint au serveur central.
Toutes ces mesures peuvent améliorer la situation mais ne jouent en rien sur les problèmes intrinsèques mentionnés précédemment.
La nouvelle approche en revanche, basée sur **l'introduction de serveurs personnels** dans le cadre du web personnel, répond mieux à cette situation.
Elle est soutenue par les start-ups et par certains académiques et aussi par certains grands groupes dont le modèle d'affaires repose sur la confiance que les usagers leur font comme La Poste ou EDF pour la France.
Certains projets nationaux comme le Blue Button aux États-Unis se basent aussi sur ce nouveau modèle.

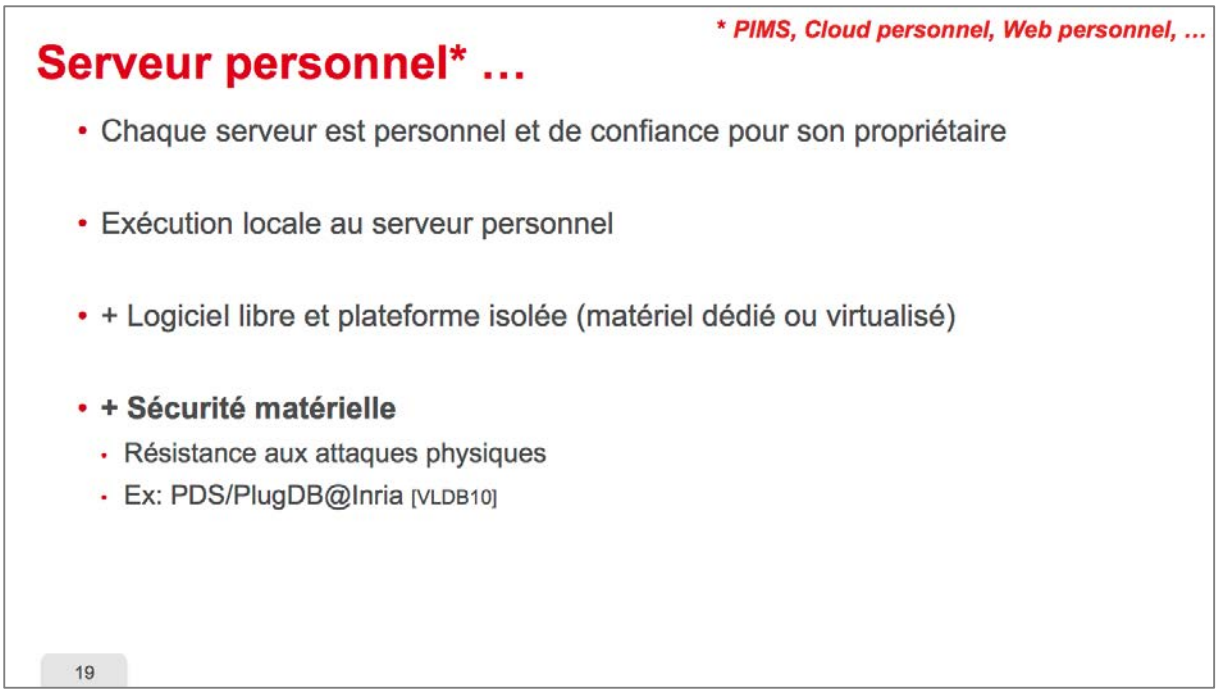
Donc ce nouveau modèle repose sur l'introduction d'un **serveur personnel.**
Ce serveur personnel doit être **de confiance pour son propriétaire** :

- Il est le serveur personnel d'un seul individu.
- Il ne va pas stocker les données de millions de personnes.
- Il est auto administré par l'individu
- Il n'y a pas de délégation.

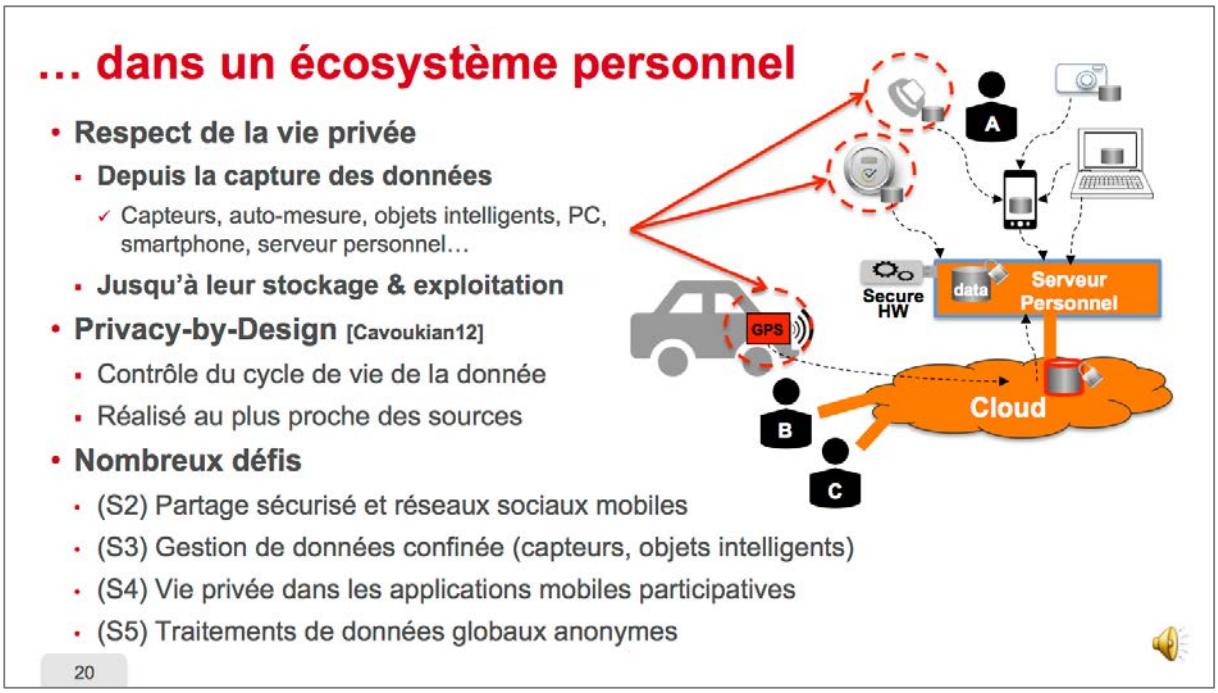
Les traitements de données sont locaux au serveur personnel. Seuls des résultats peuvent être exposés, mais absolument pas les données de base qui permettent de les calculer.



On peut ajouter à ça, le fait que le logiciel est souvent libre et ouvert de manière à permettre une vérification communautaire du code.
Le serveur est isolé des applications notamment, ce qui limite le risque d'attaque.



Enfin le serveur peut être sécurisé matériellement contre les attaques physiques en lui ajoutant un composant matériel sécurisé type carte à puce ou n'importe quel autre token assurant une sécurité physique.



C'est tout l'écosystème personnel qu'il va falloir protéger. La **vie privée** doit être assurée depuis la capture des données personnelles donc depuis les dispositifs qui entourent l'individu jusqu'à leur exploitation.
L'approche dite du **privacy by design** a justement pour objectif de contrôler intégralement le cycle de vie des données. Cette protection de bout en bout doit être assurée au plus proche des sources de données et ensuite au-delà.
Ça ouvre à de nombreux défis qui seront bien sûr étudiés dans les séquences qui suivent.

Références

- [Cavoukian12] Cavoukian, A. (2012). Privacy by design and the emerging personal data ecosystem. *Privacy By Design*.
- [CFP00] Catlett. J. Panel on infomediaries and negotiated privacy techniques. In Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions, CFP '00, pages 155–156, New York, NY, USA, 2000
- [JDDDM08] Mitchell, A., Henderson, I. and D. Searls. Reinventing direct marketing — with VRM inside. Journal of Direct Data and Digital Marketing Practice, 10(1):3–15, 2008
- [PLOS14] de Montjoye, Y. A., Shmueli, E., Wang, S. S., & Pentland, A. S. (2014). OpenPDS: Protecting the privacy of metadata through safeanswers. *PloS one*, *9*(7).
- [PlugDB] https://project.inria.fr/plugdb/
- [VLDB10] Allard, T., Anciaux, N., Bouganim, L., Guo, Y., Le Folgoc, L., Nguyen, B., ... & Yin, S. (2010). Secure personal data servers: a vision paper. *Proceedings of the VLDB Endowment*, *3*(1-2), 25-35.
- [SOSP11] Popa, R. A., Redfield, C., Zeldovich, N., & Balakrishnan, H. (2011, October). CryptDB: protecting confidentiality with encrypted query processing. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles* (pp. 85-100). ACM.
- [TKDE14] Bajaj, S., & Sion, R. (2014). TrustedDB: A trusted hardware-based database with privacy and data confidentiality. *Knowledge and Data Engineering, IEEE Transactions on*, *26*(3), 752-765.
- [WEF12] The World Economic Forum. Rethinking Personal Data: Strengthening Trust. May 2012
- VRM project, http://blogs.law.harvard.edu/vrm/projects/
- CozyCloud https://cozy.io/fr/ ; OwnCloud https://owncloud.org/ ; FreedomBox http://freedomboxfoundation.org/
- Wikipedia: Freedombox, Vendor Relationship Management