L'univers du Big Data

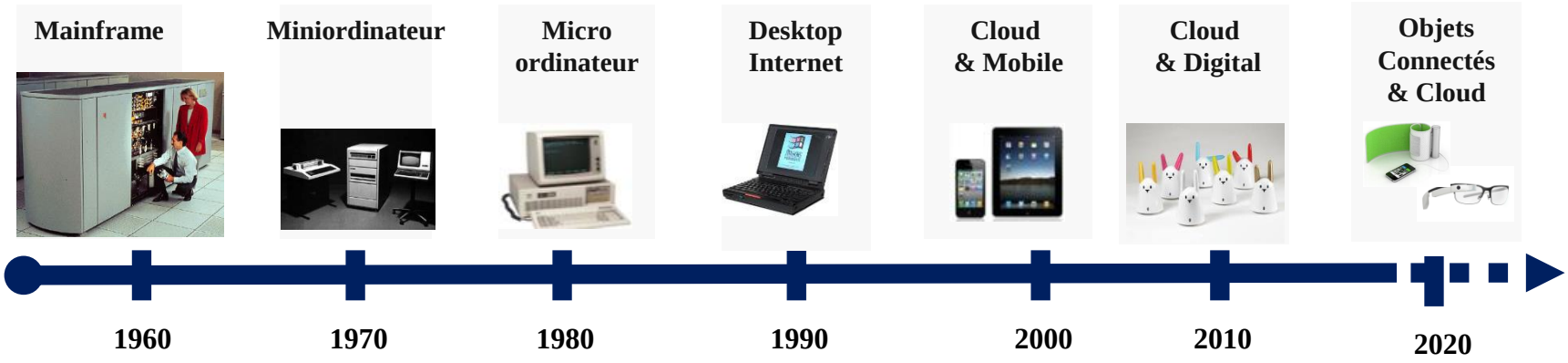
Par JF GOGLIN – Conseiller national SIS

Conférence de Territoire Yvelines

FEHAP
FÉDÉRATION DES ÉTABLISSEMENTS HOSPITALIERS & D'AIDE À LA PERSONNE
PRIVÉS NON LUCRATIFS


Les données produites autour de chacun d'entre nous ne cessent de proliférer

Du mainframe aux objets connectés…

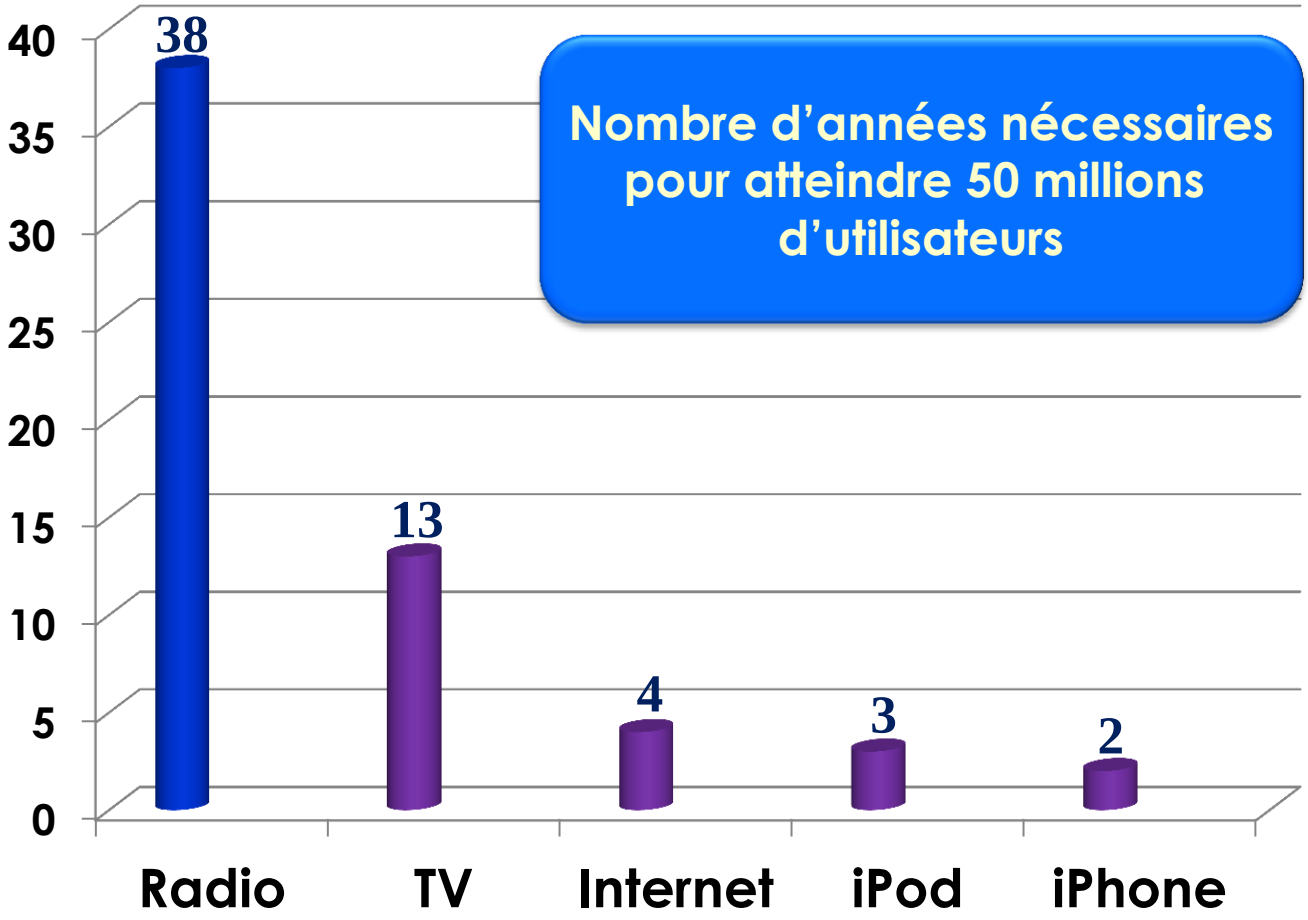


Informatique institutionnelle
1 terminal pour n utilisateurs

Informatique personnelle
1 poste de travail pour 1 utilisateur

Le numérique déferle
n objets connectés par utilisateur

- **Le numérique accélère le changement :**



The Internet in Real-Time

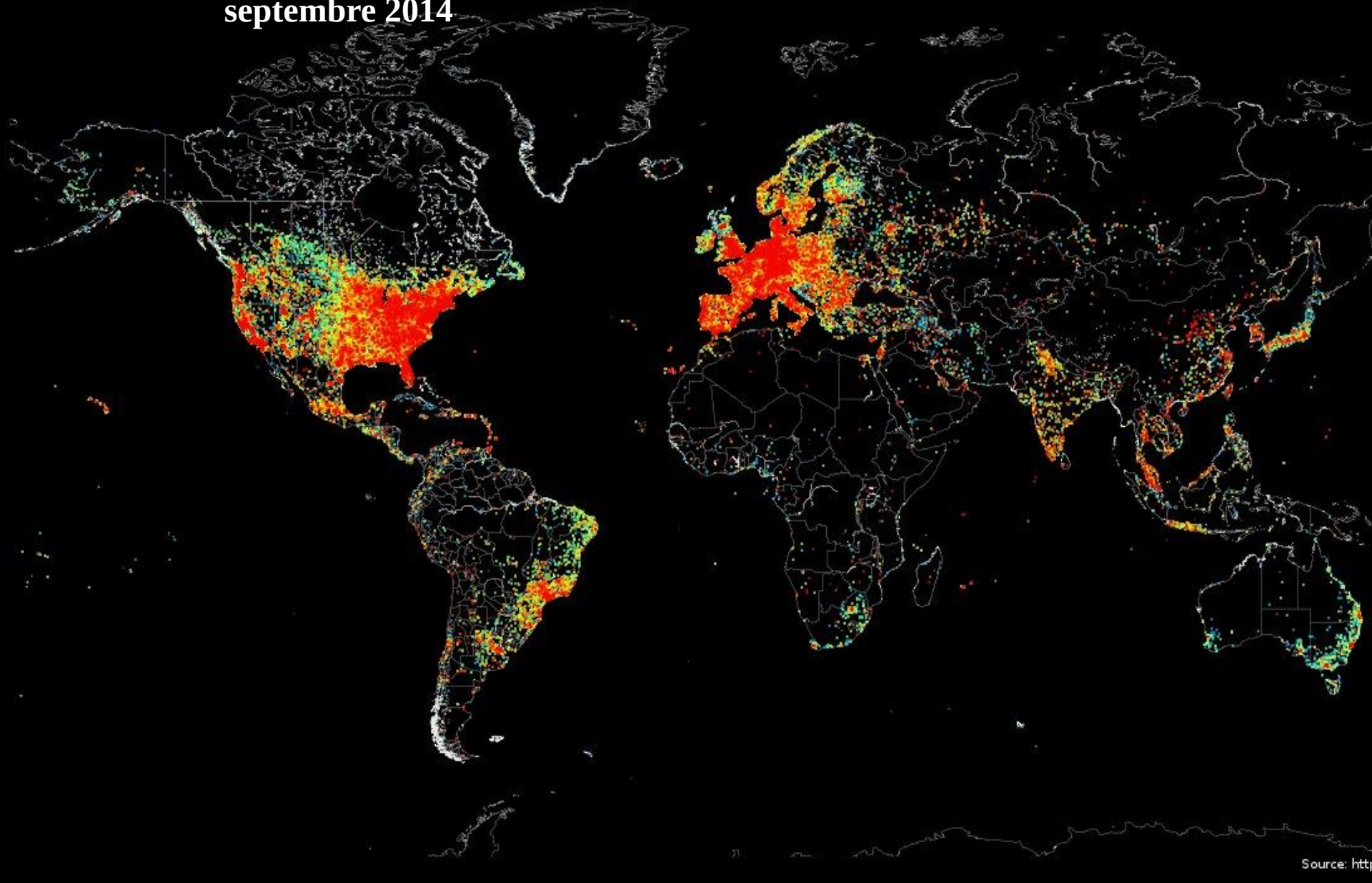
How Quickly Data is Generated

Click here to watch as these internet giants accumulate wealth in real-time.

By the way, in the 60 seconds you've been on this page, approximately 1354440 GB of data was transferred over the internet.

SHODAN

Etat des connexions internet dans le monde, par by John Matherly, septembre 2014



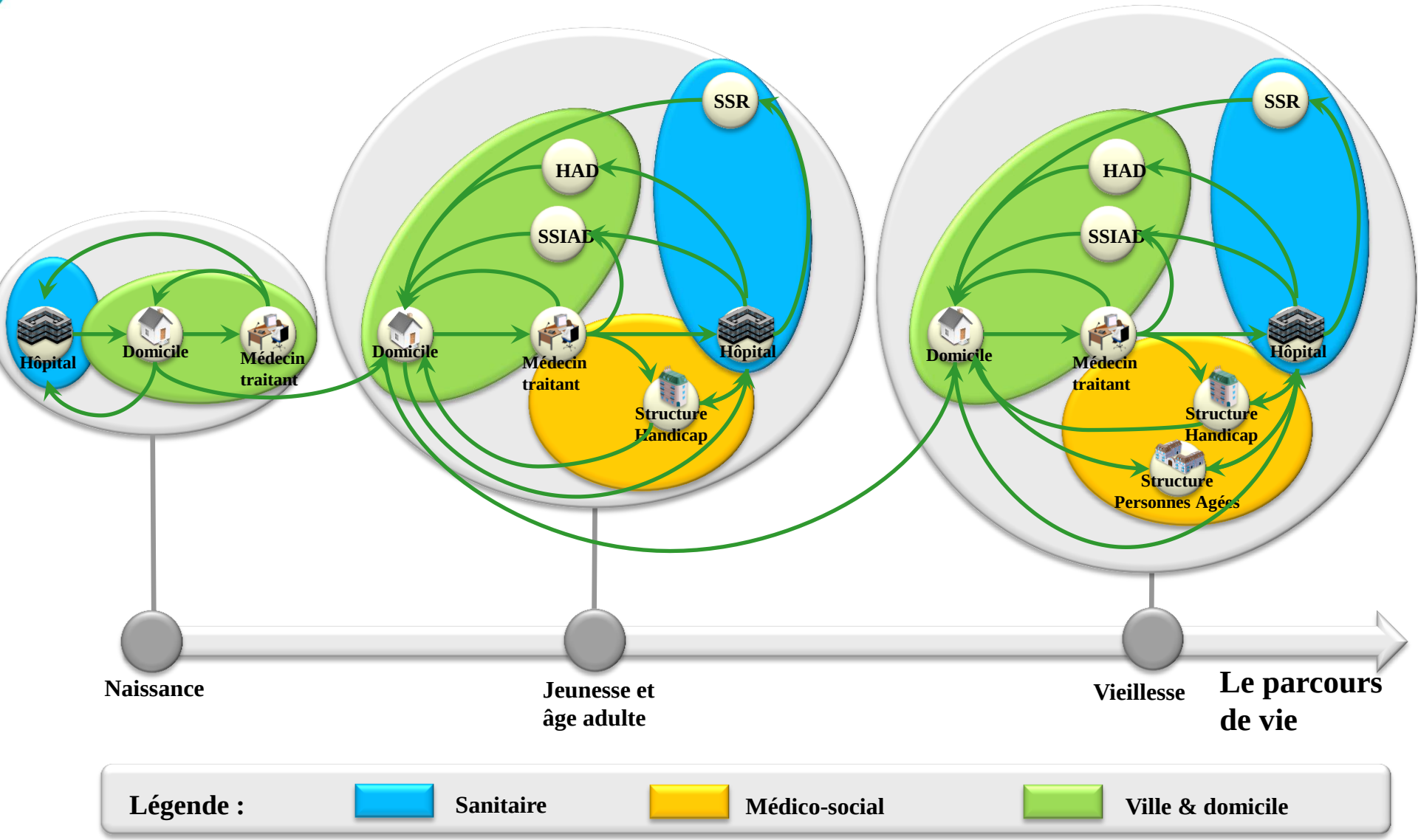

La lame de fond des objets connectés
commence à déferler sur nos SIS

Une lame de fond déferle sur nos SIS

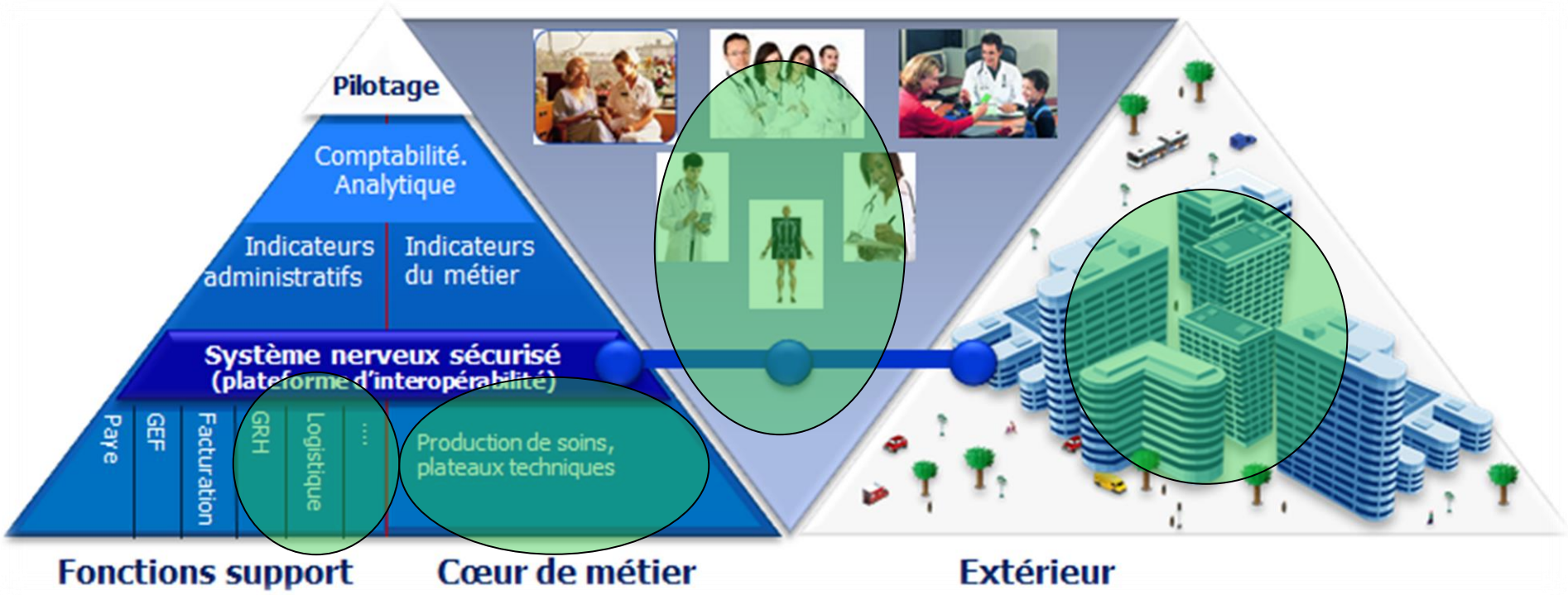
- **Les objets connectés sont déjà parmi nous….**
  - Environ 15 milliards d'objets connectés sont recensés aujourd'hui.
  - 76,8% des internautes ont déjà entendu parler des objets connectés.
  - 6% des français ont une balance intelligente connectée.
  - 14% des 18-24 ans disposent d'un appareil connecté comme une montre connectée, un coach, une balance intelligente.
  - 11 millions de français auront un objet connecté en 2017.
  - Selon les estimations, il y aura entre 80 et 100 milliards d'objets connectés dans le monde d'ici 2020 .
  - Toutes les tranches d'âge sont concernées.

- **L'observance, le suivi à distance, la télésurveillance, la prévention sont autant d'objectifs visés par l'essor et la généralisation des objets connectés liés au domaine de la santé.**

Une lame de fond déferle sur nos SIS



Une lame de fond déferle sur nos SIS



Le monde du sanitaire

Le monde du médico-social

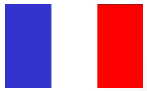
Le patient/usager chez lui, en action dans sa vie quotidienne.

Légende : **Objets connectés**

Qu'est ce qu'un objet connecté ?

- L'expression "Internet des objets" (Internet of Things ou «IoT») a été inventée voici près de 20 ans par des enseignants du MIT (Massachusetts Institute of Technology) pour décrire un monde dans lequel des objets intelligents munis de capteurs sont connectés afin de partager et d'utiliser des données collectées.

- Les objets connectés recueillent des données sur leur environnement, communiquent avec un ou plusieurs utilisateurs, voire avec d'autres objets connectés, et peuvent interagir avec un écosystème en fonction de variables données.



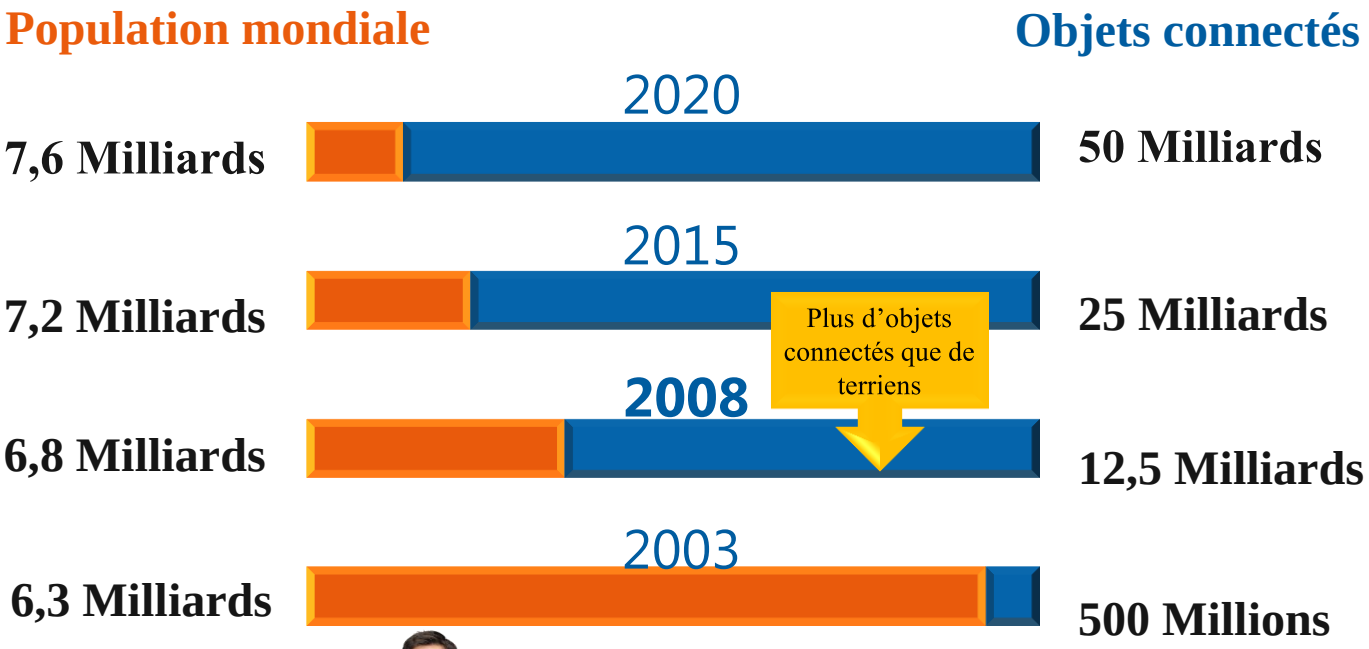
23% des français utilisent déjà un objet connecté.

Et vous ?

- **Il existe différents types d'objets connectés**

Santé			
Médical	**Bien-être**	**Nutrition-alimentation**	**Sport**
Glucomètre, tensiomètre, pilulier,…	Balance, montre, podomètre,…	Fourchette, cocotte minute connectée,…	Bracelets, chaussures, GPS…

De plus en plus d'objets connectés…



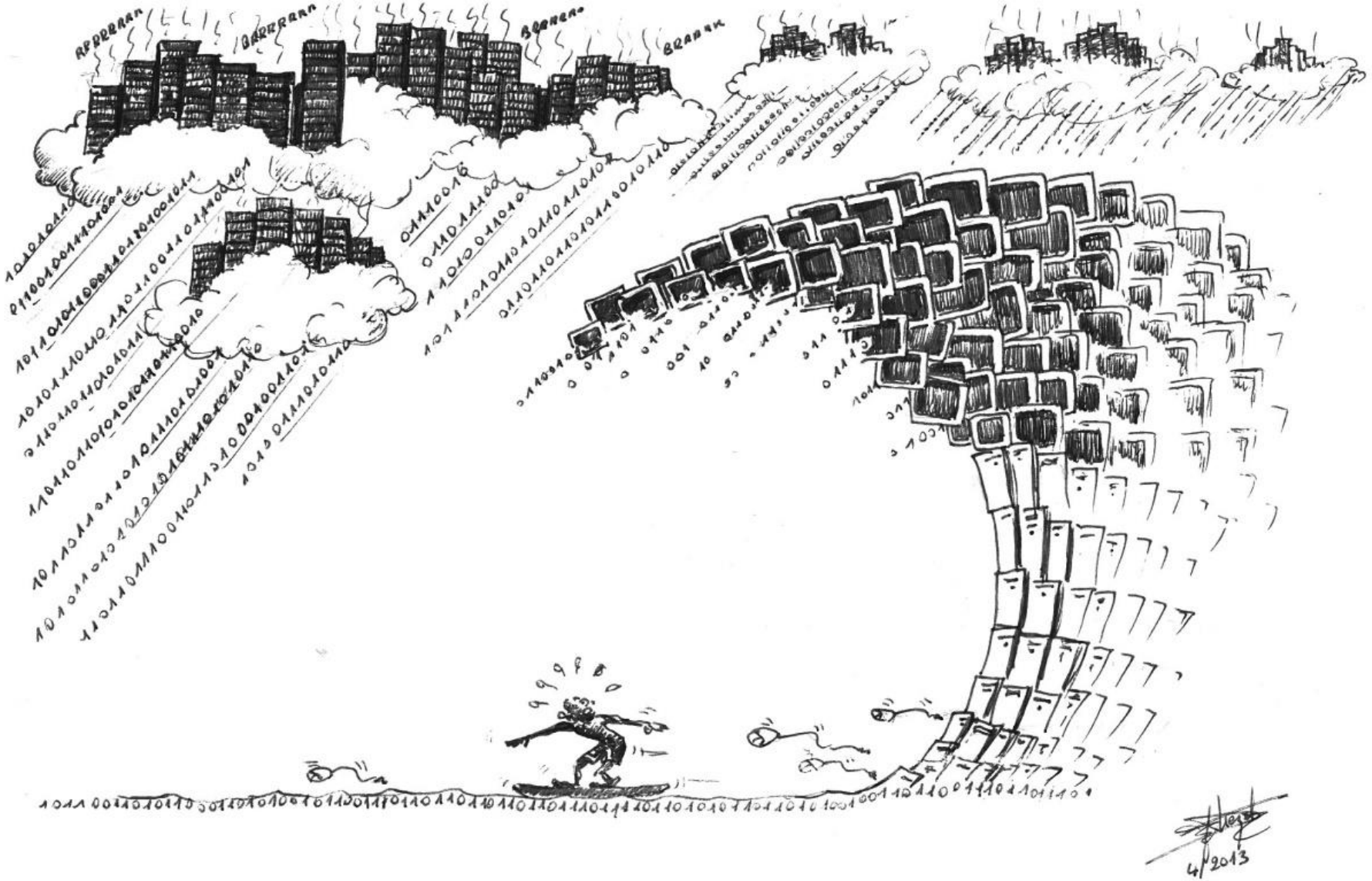
En 2015, il y aura 3 fois plus d'objets connectés que de personnes sur terre.

Notre société laisse derrière elle une empreinte digitale.

Les citoyens-usagers-patients vivent connectés, et s'expriment en postant des mails, des commentaires, des opinions, des "like" ou des "dislike" …

De fait, nous produisons un océan de données.

Un océan de données



Un océan de données….
Avec beaucoup de bruit de fond.

Les données produites… l'or du 21e siècle

- **Comment traiter ces océans de données ?**
  La production massive de données issue notamment de tous ces objets connectés va créer d'énormes gisements de création de valeur.
  La compréhension et la maitrise de ces données, qu'elles soient structurées ou non, pour en faire des connaissances et de la prédiction constitue la mine d'or du 21e siècle.

L'IDC estime que le volume des données numériques va être multiplié par 44 dans la prochaine décennie.

Plus de données seront produites dans les 18 prochains mois qu'il n'y en a eu depuis le début du numérique.

Les données produites… l'or du 21e siècle

- **Ce déluge de données a des caractéristiques particulières :**
  - Elles proviennent de **sources très disparates et très hétérogènes** (téléphones mobiles, tablettes, objets connectés, réseaux sociaux, capteurs, téléviseurs connectés, PC fixes, PC portables, …), de façon désordonnée et non prédictible.
  - Ce sont essentiellement des **données non structurées***.
  - Elles sont produites en **temps réel**.
  - Elles arrivent mondialement en **flots continus**.
  - Certaines données (comme une photo, une prescription, un acte,…) sont associées à **des données de localisation, d'horodatage**, ….

* Une donnée structurée est prédéfinie dans son format : exemple, le nom du patient, un code postal, un code d'acte.

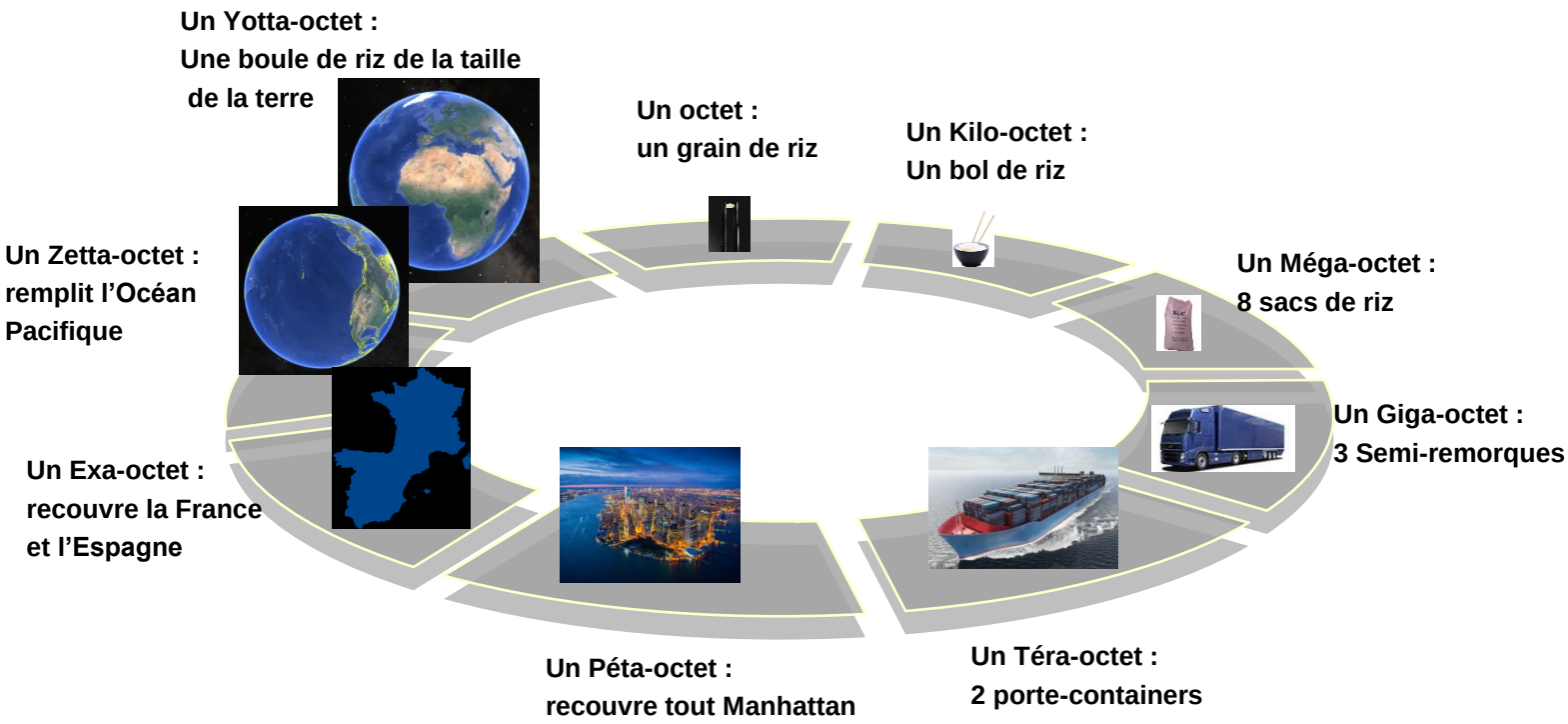
Une donnée non structurée n'a pas de format prédéfini : exemple, un compte-rendu, un mail, un article de journal, une observation, un film, une image, une bande son….

Les données produites… l'or du 21e siècle

Pour Bernard Ourghanlian, Directeur technique chez Microsoft, **il existe un océan de données de santé, alimenté par l'ensemble des objets connectés.**

Cet océan de données combine les données les unes avec les autres et voit émerger de ce semblant de chaos de nouveaux territoires à explorer.

Les méga-données, à partir de quel volume?



Pour passer d'un segment à l'autre, on multiplie le volume initial par 1024.

Un océan de données

Entre la naissance de notre civilisation et 2003, l'humanité a généré 5 exa-octets de données. Actuellement, nous produisons 5 exa-octets tous les deux jours, et cela ne fait que s'accélérer.

Eric Schmidt, ***Executive Chairman, Google***

Qui produit ces volumes de données?

Peut-on stocker de tels volumes?

Un océan de données

Le gouverneur Herbert, de l'Utah, a annoncé le 11 juin 2012 que le nouveau super datacenter de la NSA construit à l'extérieur de Salt Lake City serait le premier à héberger un yotta-octet de données.

Ce centre est opérationnel depuis septembre 2014.

Qu'est-ce que le Big Data?

Qu'est-ce que le Big Data?

Les Big Data, ou mégadonnées, parfois appelées données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information.

On entend parfois parler de «datamasse» par similitude avec la biomasse.

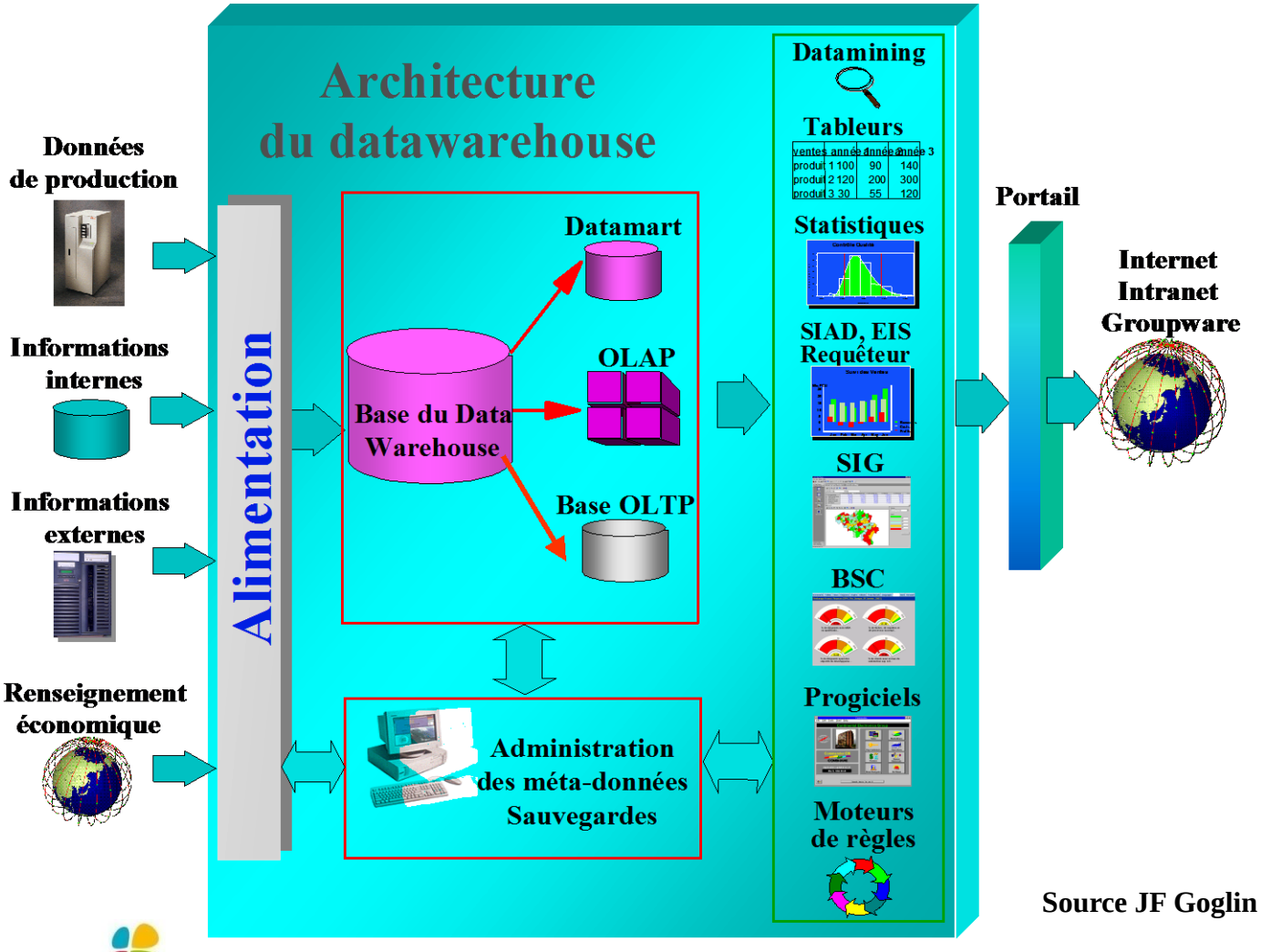
Qu'est-ce que le Big Data?

Le Big Data est caractérisé par la **maitrise des « 5 V » :**

- **La variété** des données qu'il faut rapprocher et rendre cohérentes (structurées, non structurées)
- Le **volume** des données produites qui peut être énorme (des centaines de Teraoctets à des Exaoctets).
- La **vitesse** de traitement, obligatoire pour obtenir des informations intéressantes dans un laps de temps suffisamment court permettant de les exploiter.
- La **véracité** des données (qualité, vraisemblance et précision)
- La **valeur** des données.


Comment ça marche?

Un système «Big Data » s'apparente à un système décisionnel de type datawarehouse



Source JF Goglin

Le Big Data permet de faire :

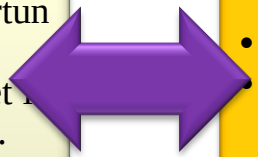
- *De l'analyse descriptive (Que s'est-il passé? Qui l'a fait? Pourquoi est-ce arrivé?)*
- *De l'analyse prédictive (Que va-t-il se passer?)*
- *De l'analyse prescriptive (Que va-t-il se passer? Quand? Pourquoi? Comment pouvons-nous bâtir le futur?)*

FEHAP

Mais un système «Big Data » est un peu différent…

Au sein d'un système décisionnel traditionnel (datawarehouse)

- Les sources de données sont essentiellement internes, connues et structurées.
- L'alimentation du système est rarement en temps réel, mais plutôt en temps opportun ou en batch.
- Les données utilisées sont structurées et le modèle des données est connu et stable.
- La qualité des données est relativement maitrisée.
- Les phénomènes à étudier sont connus et modélisés sous forme d'indicateurs.
- La gestion des droits et des autorisations de collecte et d'utilisation est relativement simple.
- Les rapports produits sont prédéfinis et récurrents.
- .

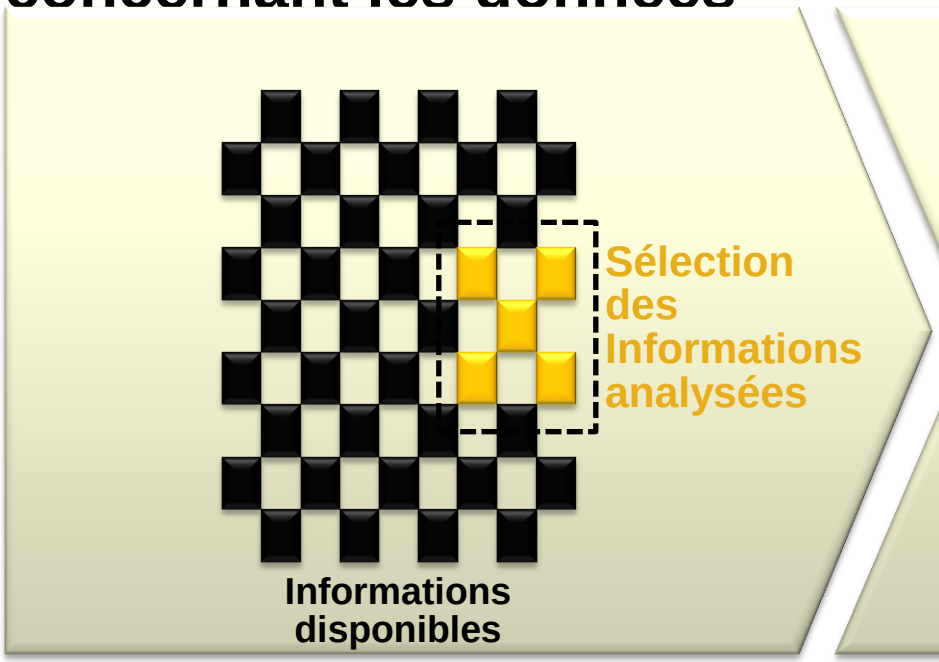


Big Data

- De nombreuses sources de données sont externes.
- Beaucoup de données sont non-structurées.
- Les données arrivent en flux continu.
- Les volumes à traiter sont très importants.
- La qualité des données est un problème important.
- Besoin d'itérations rapides pour expérimenter des hypothèses.
- L'analyse est faite sur des données qui restent dans leur état brut.
- La propriété des données doit être étudiée avec soin afin de respecter la Loi.

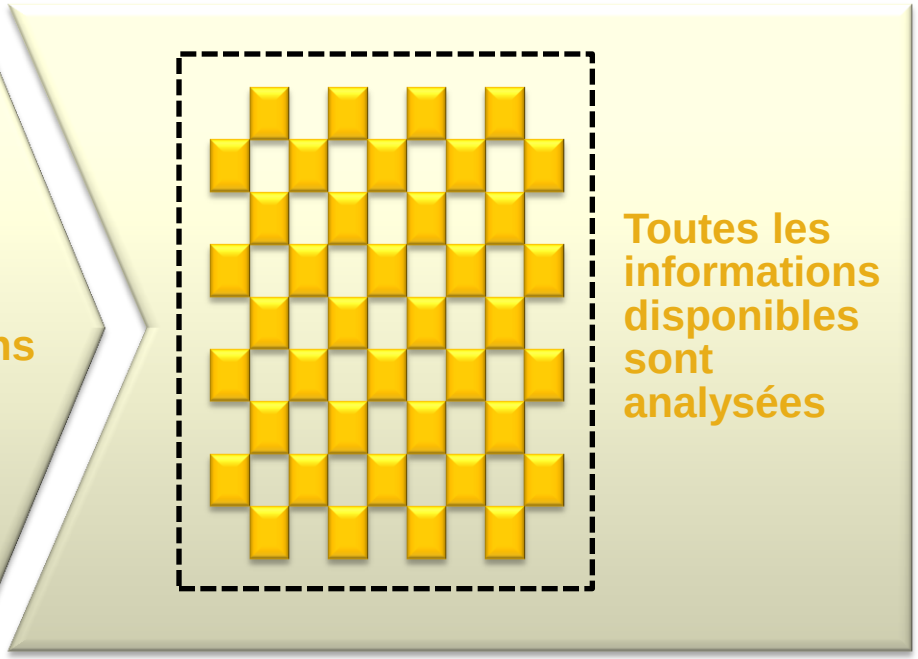
Mais un système «Big Data » est un peu différent…

L'approche traditionnelle concernant les données



On analyse de petits ensembles de données, de bonne qualité, plutôt de type structuré

L'approche Big Data



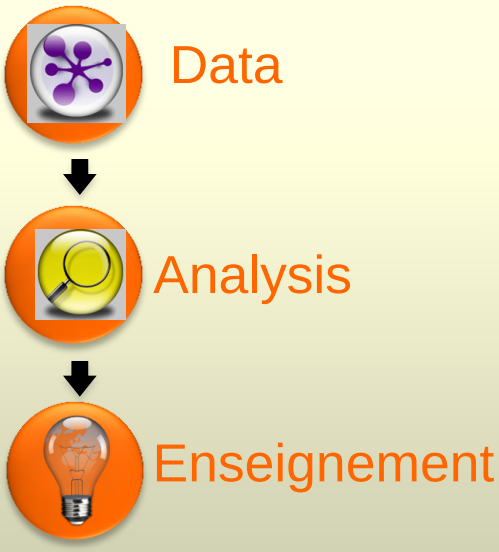
Toutes les données sont analysées qu'elles soient structurées ou non.

Mais un système «Big Data » est un peu différent…

L'approche traditionnelle

La donnée est analysée dès lors qu'elle est stockée dans le datawarehouse

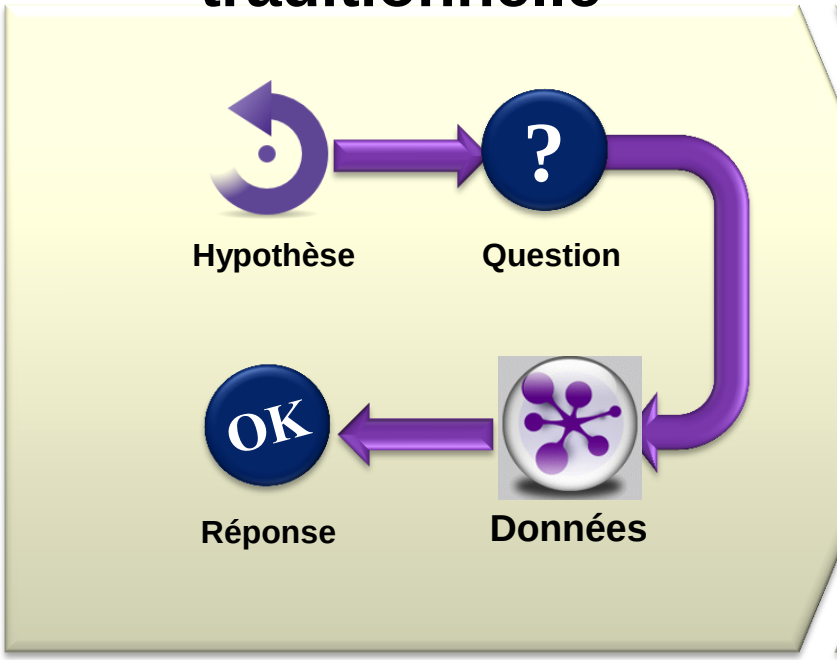
L'approche Big Data



La donnée est analysée en temps réel dès qu'elle est disponible.

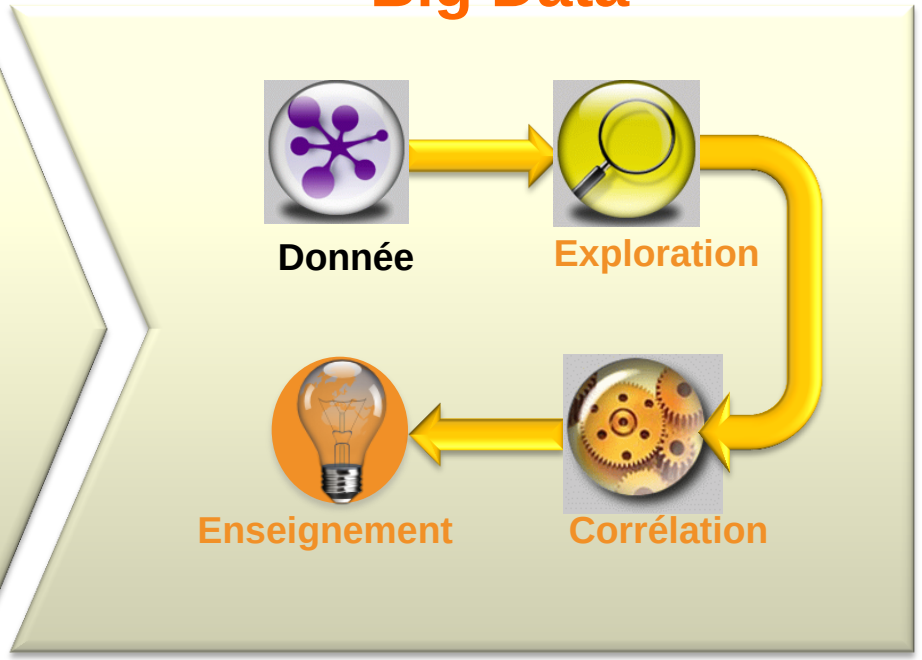
Mais un système «Big Data » est un peu différent…

L'approche prédictive traditionnelle



Une hypothèse est testée sur un jeu de données

L'approche prédictive Big Data



Toutes les données sont explorées pour identifier des corrélations

Ethique et Big Data

Mes données personnelles

- Mon identité et mes alias
- Mes coordonnées
- Mes données génomiques
- Ma géolocalisation et mes parcours
- Mes centres d'intérêt
- Mes données dans les réseaux sociaux
- Mes amis
- Mes connaissances
- Mes courriels
- Mes appels téléphoniques
- Mes achats,
- Mes données médicales,
- Mes données fiscales,
- Mes assurances…
- Ma vie privée…

Les données personnelles appartiennent à la personne

Les entreprises/états n'en sont que les dépositaires temporaires

Qui maitrise les données ?

- La production de la donnée peut relever d'un processus relativement complexe mêlant :
  - une activité humaine
  - des capteurs
  - un ou plusieurs traitements
  - divers enrichissements par différents acteurs.

- Les différents contributeurs peuvent revendiquer une participation à la valeur créée par l'exploitation des données.

L'enjeu de la protection des données personnelles

Les technologies actuelles du Big Data permettent de presque tout savoir sur un individu dans l'espace et dans le temps (temps passé et bientôt dans le temps futur) par la collecte et le traitement de données sur cette personne.

L'usage que l'on pourrait faire de ces technologies peuvent être attentatoires à nos libertés fondamentales.

Il y a donc un enjeu crucial, celui de la protection des droits de l'individu en environnement numérique.

Un enjeu éthique au-delà de la protection des données

Dès lors qu'une donnée a été associée à l'identité d'une personne, toute association à une identité virtuelle brise l'anonymat des autres données.

Source : Texas researchers Arvind Narayanan et Vitaly Shmatikov

Une personne +Big Data

=

Zéro anonymat

Le Big Data amplifie la transparence issue de la digitalisation de l'activité humaine pour les individus et les organisations

Un enjeu éthique au-delà de la protection des données

La violation des règles de protection des données personnelles est passible de sanctions :

- Administratives
  - Avertissement (public ou pas)
  - Injonction à cesser le traitement.
  - Retrait de l'autorisation.
  - Amendes allant de 150 K€ à 300 K€ si récidive, avec possibilité de les rendre publiques.
- Sanctions pénales
  - 5 ans d'emprisonnement
  - 300 K€ d'amende

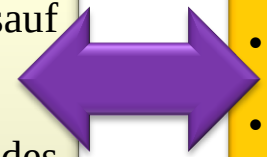
Google

Google a déjà écopé d'une amende maximale de la CNIL.

Des principes et des droits très différents, parfois contradictoires

Principes et droits applicables aux traitements de données personnelles

- Les données sont collectées et traitées de manière loyale et licite (pas de données sensibles sauf exception)
- Principe de finalité spécifique
- Proportionnalité : traitement des seules données nécessaires
- Qualité des données : exactes et mises à jour
- Durée de conservation limitée
- Transparence et information des personnes
- Droits des personnes : accès, rectification, opposition.



Big Data

- La légalité de la collecte des données n'est pas toujours maîtrisée ni acquise.
- Les données sont souvent réutilisées pour d'autres finalités.
- Les données sont parfois revendues sans le consentement des personnes.
- La conservation des données est parfois illimitée avec peu de possibilités d'effacement réel.
- Les personnes en sont pas ou trop peu informées.
- La multiplicité des acteurs rend difficile la gestion des droits et des personnes.

Pour vous aider dans la mise en œuvre de vos projets Big Data, une charte éthique existe sur internet.

http://www.cil.cnrs.fr/CIL/IMG/pdf/CharteEthiqueBigDatav5.pdf

Que peut-on faire avec le Big data?

Que peut-on faire avec le Big Data?

Le Big Data est historiquement tracté par plusieurs acteurs :

- Les services spéciaux qui nous écoutent.
- Le GAFA (Google, Apple, Facebook, Amazon) qui scrutent nos actes et nos commandes pour nous proposer de la publicité et de nouveaux services/produits.

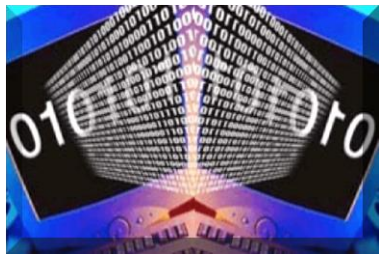
Que peut-on faire avec le Big Data?

La santé

La valorisation de la relation client

La finance

La sécurité informatique



La sécurité des biens

La sécurité routière

Les télécoms

La qualité

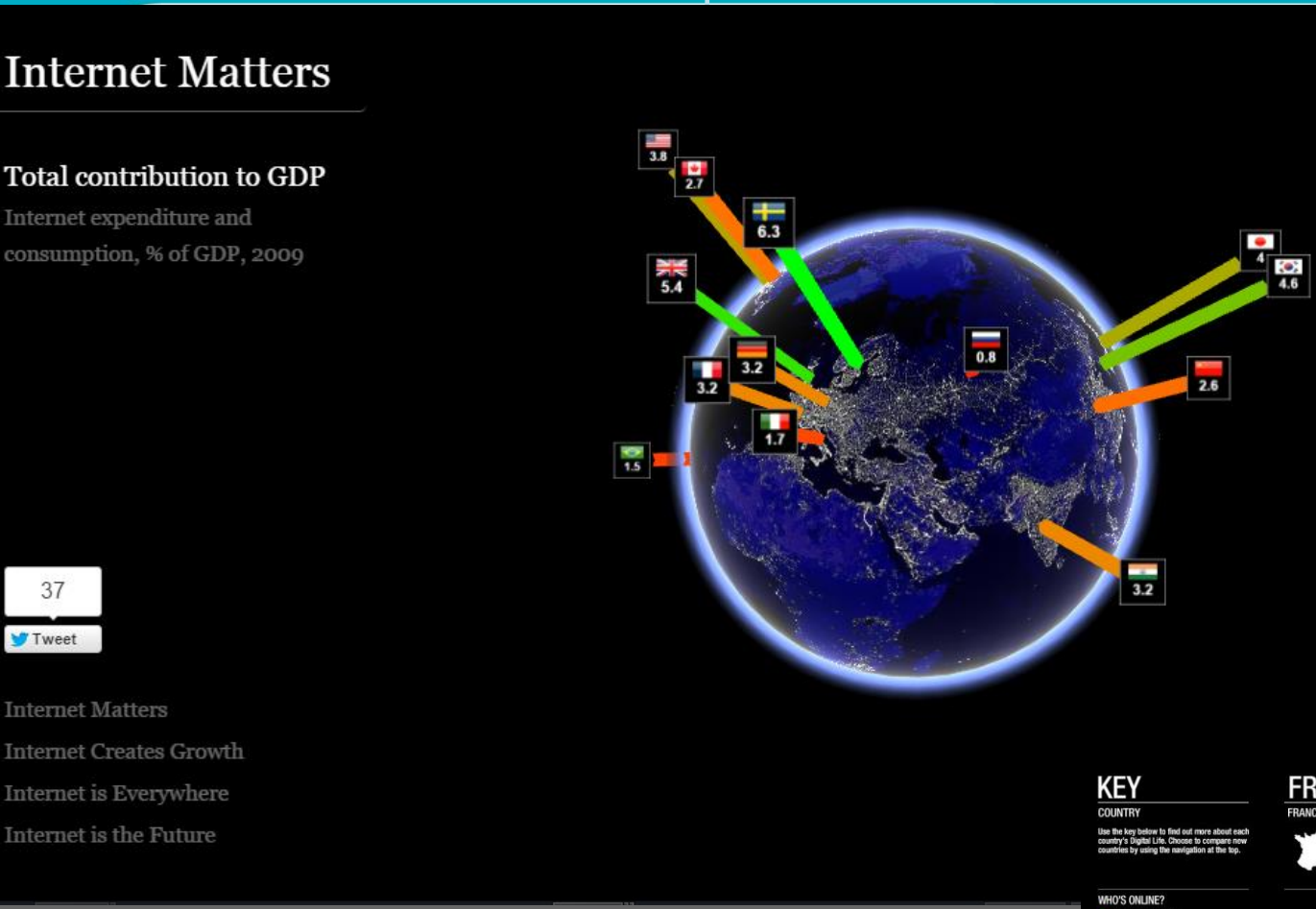
La fabrication de biens

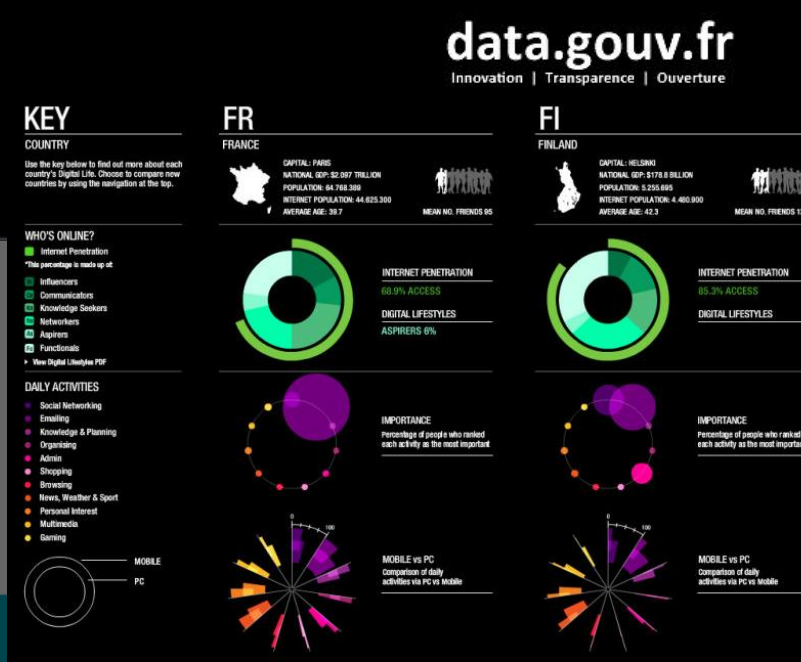
L'analyse des ventes

La détection de fraudes

L'atritition

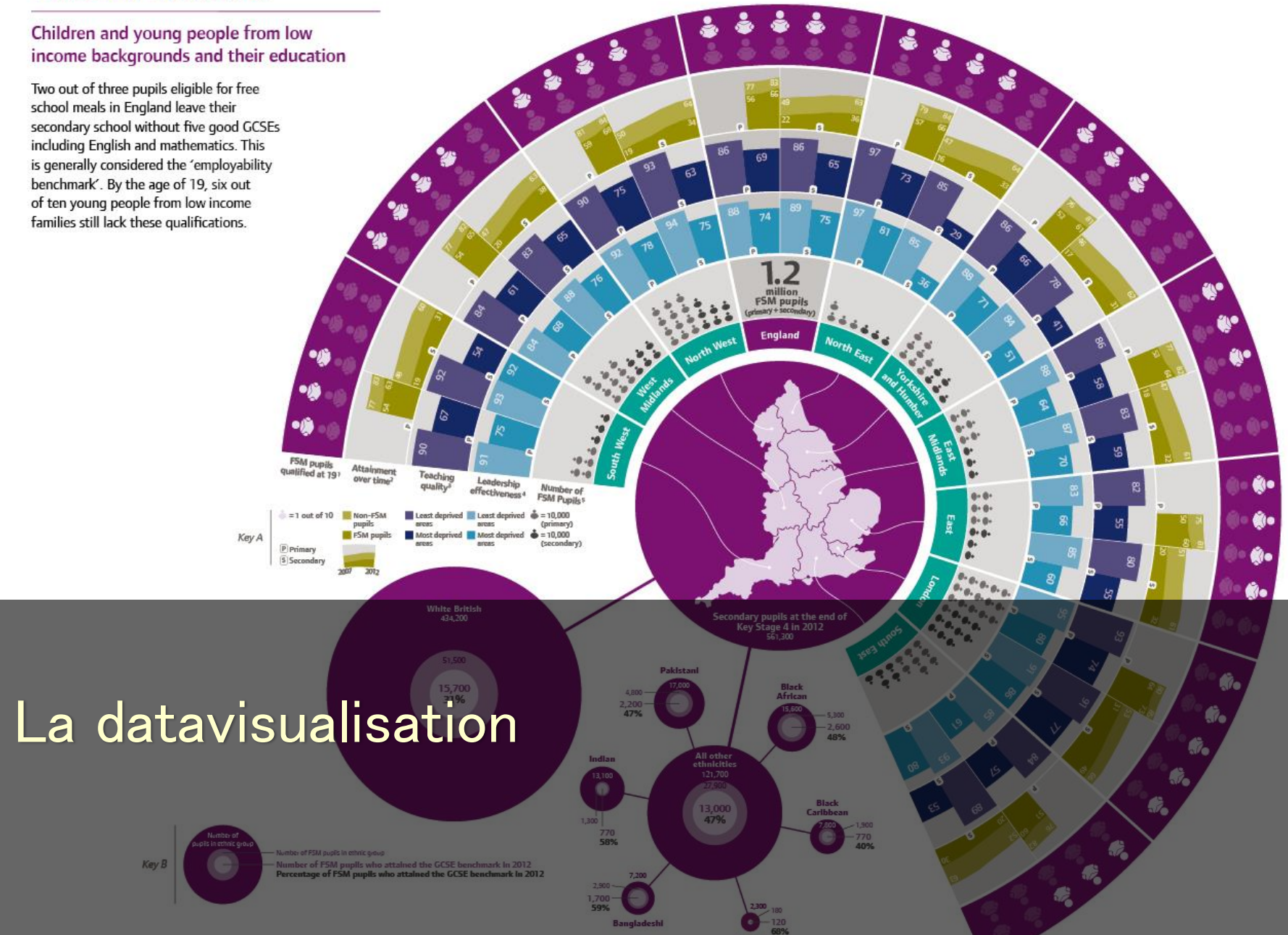
Que peut-on faire avec le Big Data?





La datavisualisation

La datavisualisation



Suivi et prévision de la propagation de la grippe et du virus H1N1

Le recrutement

Certains spécialistes annoncent d'ores et déjà la mort du CV traditionnel.

En effet, avec le Big Data, toutes les données du candidat pourront être capturées sur l'ensemble du Net, pour obtenir ses vrais diplômes, ses habitudes sur les réseaux sociaux, …

WATSON d'IBM

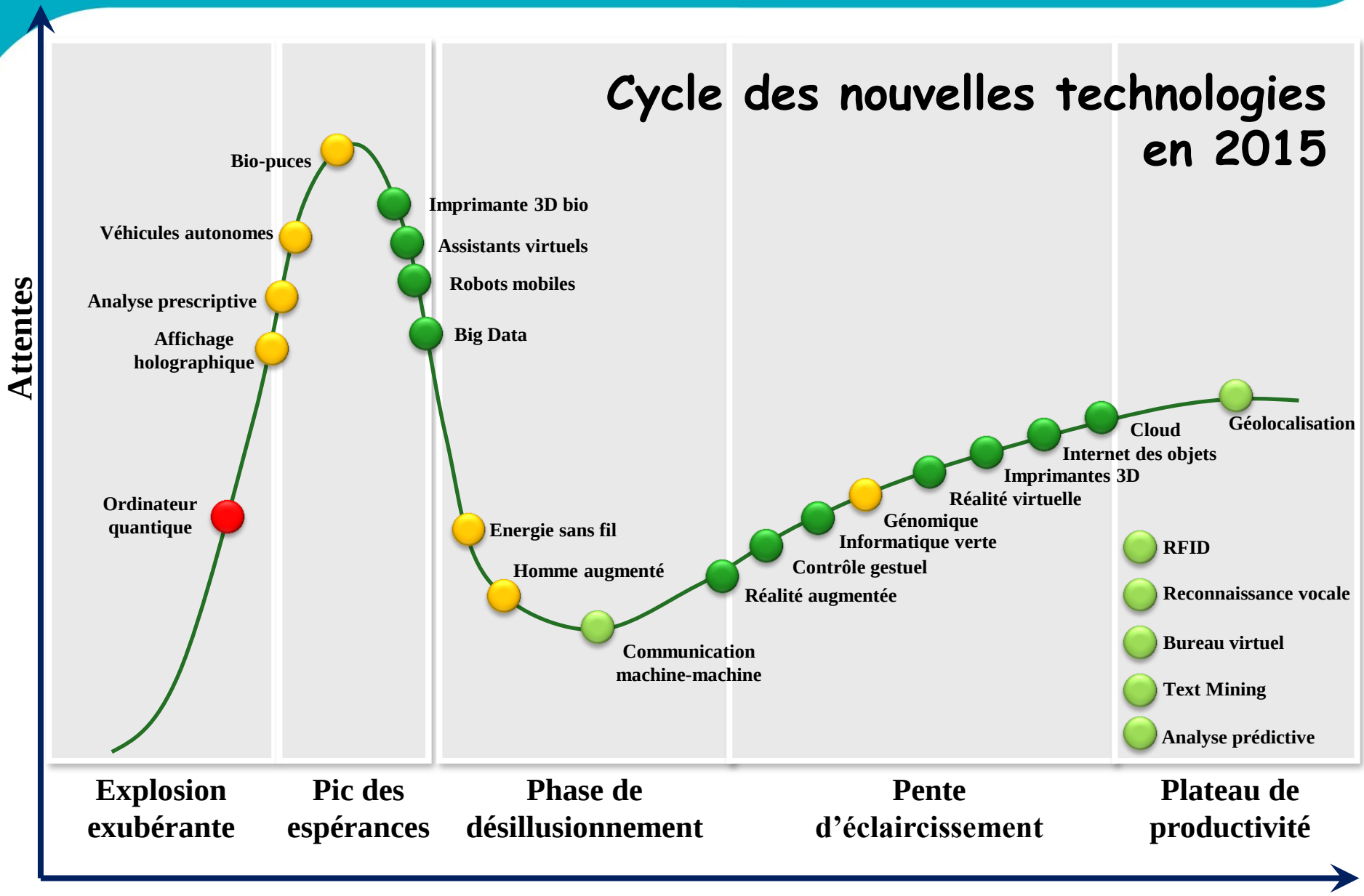
Selon Sloan-Kettering, **seulement 20% environ du savoir qu'un médecin utilise quand il réalise son diagnostic et prescrit un traitement est basé sur des études scientifiques vérifiées**.

De plus **cela prendrait au moins 160 heures de lecture hebdomadaire à chaque médecin pour se tenir à jour des nouvelles publications médicales** (pour rappel, une semaine = 168 heures).

Le système WATSON d'IBM, réussit cette performance, obtenant **un taux de réussite quant à la détection du cancer du poumon de 90% pour 50 % à un médecin humain.**

Le potentiel est énorme…

Le cycle de Hype FEHAP des nouvelles technologies pour 2015



Le pourquoi de la désillusion?

De grands espoirs sont nés, sur la base du fait que Le Big Data résoudrait beaucoup de problèmes de l'humanité….

- On a plus en plus de données bientôt toutes les donnée nécessaires…
- Nous allons vaincre le cancer, la pauvreté...
- Nous pouvons prévenir l'avenir individuel grâce à la génomique…

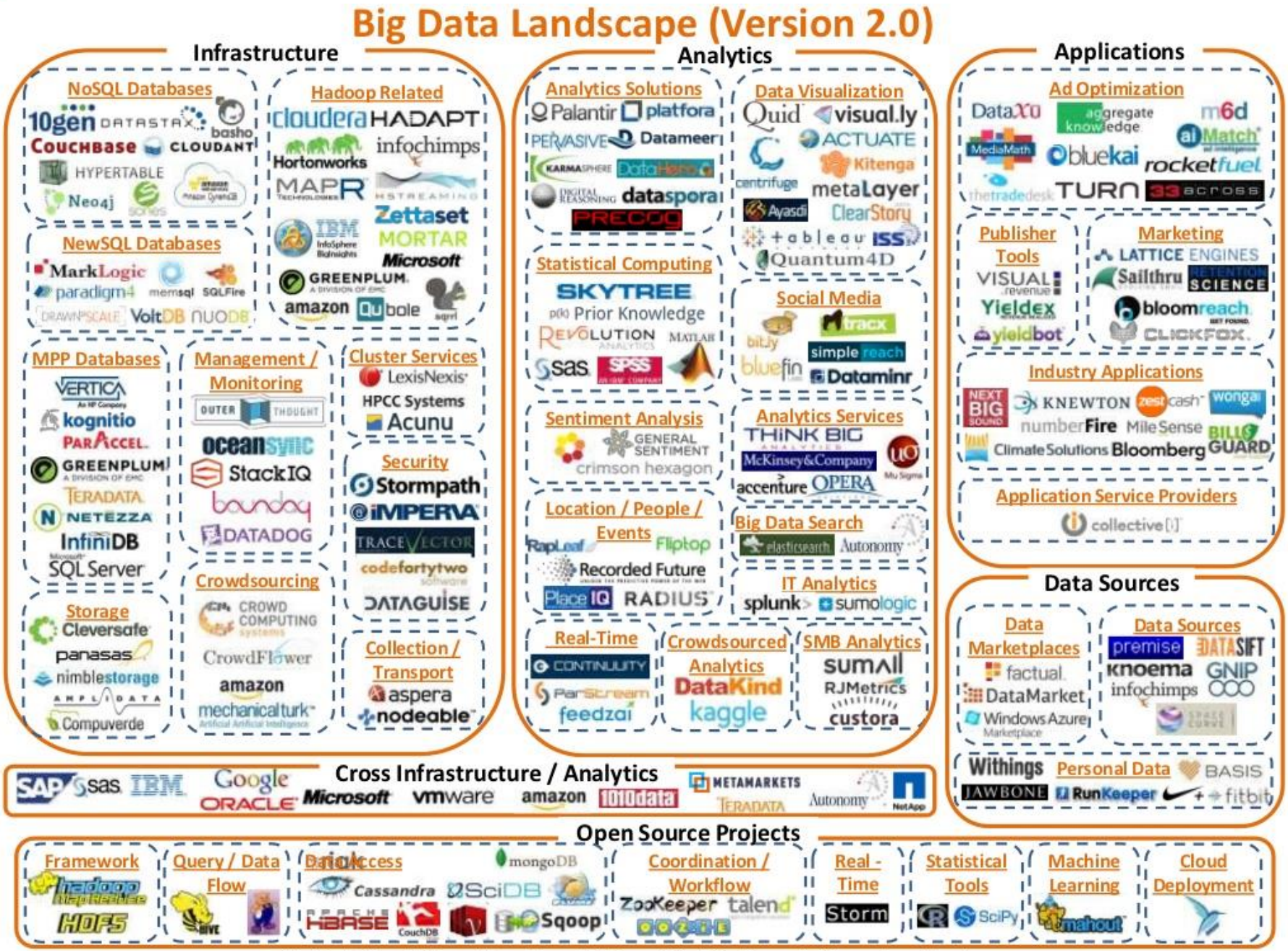
Mais en réalité ….

Même si beaucoup d'outils existent ou sont annoncés, les outils, accessibles qui permettront de réaliser du Big Data, à un prix accessible à toutes les entreprises n'existent pas encore.

Leur performance n'est pas toujours au rendez-vous du fait qu'il s'agit de traiter en temps réel d'énormes quantités de données.

Par ailleurs, même si ces outils fonctionnent , la qualité des données récupérées laisse souvent à désirer… ce qui n'est pas un problème nouveau.

Que peut-on faire avec le Big Data?



De l'analyse prescriptive (Que va-t-il se passer? Quand? Pourquoi? Comment pouvons-nous bâtir le futur?)

IBM
AYATA

Pour l'analyse prédictive (Que va-t-il se passer?)

IBM
Microsoft
Oracle
SAS
Teradata
SAP
Alteryx
Arcplan
KXEN
Opera
Intelligence
iTrend
Palantir
QlikTech
Tableau
River Logic
RJ Metrics
Targit

Pour l'analyse descriptive (Que s'est-il passé? Pourquoi est-ce arrivé?)

IBM
Microsoft
Oracle
SAS
ISS
Jaspersoft
Teradata
Advizor Solutions
Bluefin Labs
Centrifuge
Clearstory
Domo
FICO
GoodData
Metal Layer
Microstrategy
Panopticon
Panorama Software
Pentaho
Quidd
Bime
SAP
QlikTech
Tableau

FEHAP

Mais en réalité ….

De plus, on observe que des sociétés privées lucratives s'approprient et revendent des données personnelles par essence privées.

Loin de résoudre les problèmes de l'humanité, on constate trop souvent que le traitement des big datas permet de générer en temps réel de la publicité de mieux en mieux ciblée qui surcharge nos écrans et nos boites mails.

La surveillance des citoyens est l'un des objectifs du Big Data pour de plus en plus d'états, fragilisant ainsi les libertés individuelles.

A notre échelle, plus humaine, que peut-on faire?

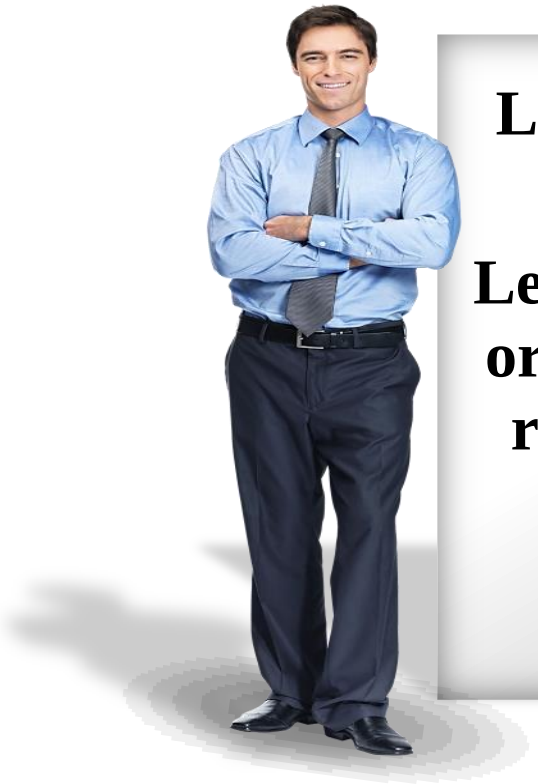
Que peut-on faire à notre échelle?

87% des entreprises sont convaincues que l'analyse des données redéfinira le paysage concurrentiel de leur secteur dans les trois prochaines années.

Reste à se mettre en capacité de faire de leur exploitation un outil d'aide à la décision.

Source Microsoft

Que peut-on faire à notre échelle?



Les mégadonnées constituent le nouvel or noir du 21e siècle.

Les entreprises, les gouvernements et les organismes qui pourront exploiter cette ressource auront un énorme avantage sur ceux qui ne le pourront pas.

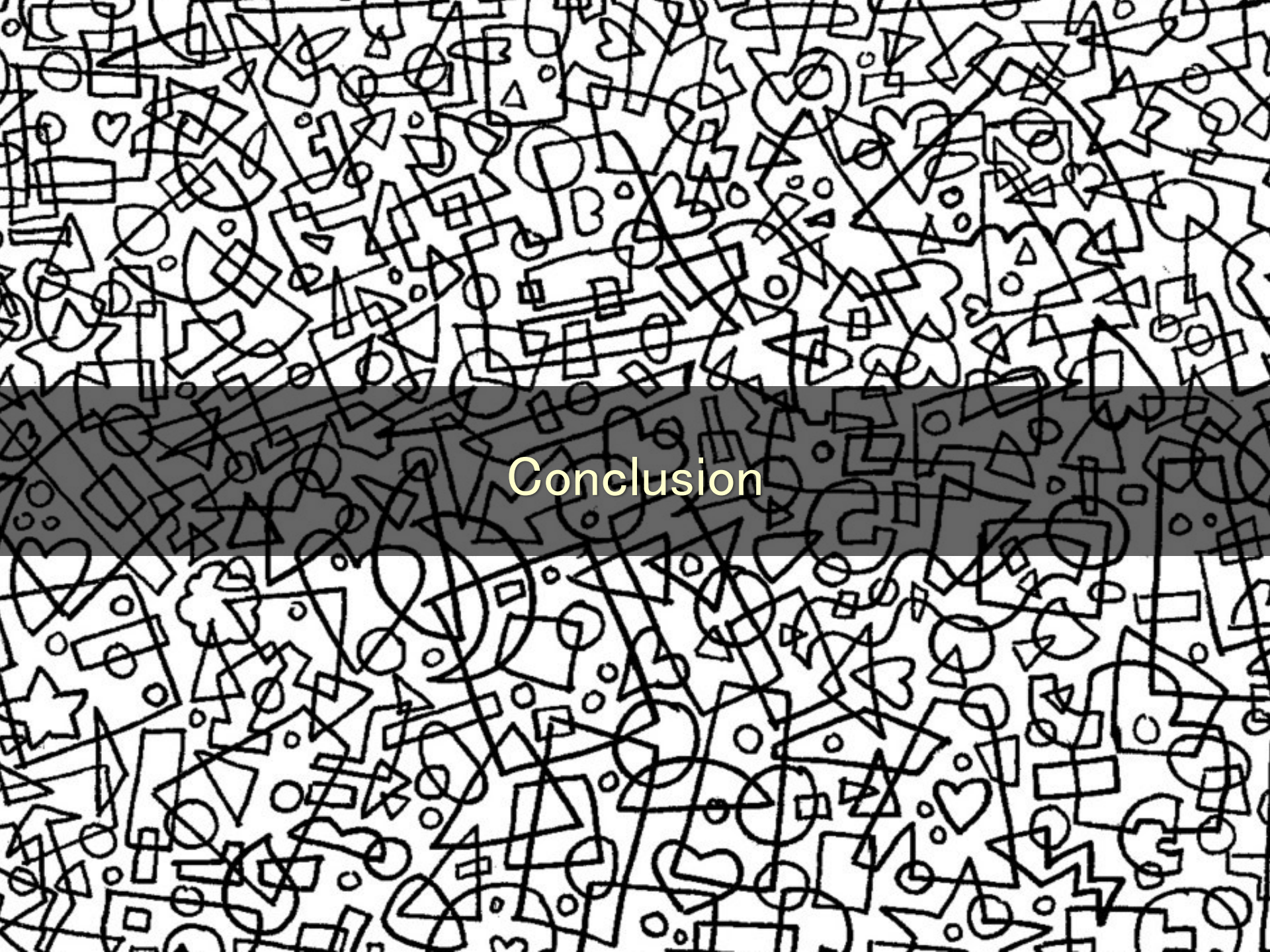
The Future of Big Data, Pew Internet, 20 juillet 2012 [traduction libre]

Que faire au niveau de l'établissement?

1. Mettre en place un outil de pilotage médico-économique doté d'un véritable outil d'extraction alimentation.
2. Doter l'outil d'extraction d'un connecteur internet permettant de récupérer des informations issues de bases externes.
3. Décloisonner les données en intégrant les données du parcours de vie du citoyen-usager-patient.
4. Valoriser les données aussi tôt que possible et se rapprocher de tableaux de bord en temps réel.
5. Mettre en place un outil de datamining permettant de rechercher des corrélations entre les données pour faire des prédictions.
6. Mutualiser l'infrastructure et les outils sur une plateforme de traitement commune tout en isolant les flux de données si celles-ci ne sont pas anonymisées.
7. Mettre en place un groupe de réflexion sur l'éthique.
8. S'intéresser à la datavisualisation.
9. Investir en formation et/ou en recrutement sur l'analyse des données.

Que faire au niveau national/régional?

1. Mettre en place des plateformes collectant les données anonymisées issues des parcours de vie et associer les données de l'open data.
2. La transformation de la santé sera accélérée grâce à l'intégration de la chaine de soins et notamment de ses systèmes d'information.
3. Réfléchir aux nouveaux rôles et au partage des rôles ainsi qu'à une meilleure coordination tout en impliquant le citoyen-usager-patient.
4. Mettre en place une charte nationale d'éthique concernant le Big Data en santé.
5. Veiller à la stricte interdiction de la revente des données personnelles collectées sans qu'il y ait eu consentement préalable du citoyen-usager-patient producteur.


Conclusion

- **Le développement des objets connectés accélère l'émergence de gisements de données personnelles susceptibles d'apporter de nouvelles connaissances.**
- **L'open-data en santé ne fera qu'accélérer ce phénomène.**

- **C'est la maîtrise du big data associée à la prédiction qui permettra de faciliter la prévention et de passer d'une médecine curative à une médecine préventive et personnalisée.**

- « Réaliser des prédictions est très difficile, surtout quand il s'agit du futur ».
  Niels Bohr

- « Les nouvelles technologies nous ont condamné à devenir intelligents. C'est à la fois une nouvelle catastrophique pour les grognons, mais c'est une nouvelle enthousiasmante pour les nouvelles générations. »
  Michel Serres 2007

Merci pour votre attention.

FÉDÉRATION DES ÉTABLISSEMENTS HOSPITALIERS & D'AIDE À LA PERSONNE
PRIVÉS NON LUCRATIFS

Jean-François Goglin
Conseiller national SIS
jeanfrancois.goglin@fehap.fr
tel : 06 62 79 27 81