Comment travailler sur des données sans y avoir accès?

Thomas Baudel, IBM France Lab

Séminaire CERNA ‘Anonymisation des données en recherche’ 3/7/19

Sommaire

Enjeux de la protection des données à IBM

- 100 ans d'expérience
- données sensibles avant d'être personnelles.
- 'les données sont au client'

Socle commun pour la protection des données

- Formation généralisée et répétée (inspirant le cours Ethics & STICs)
- Audit indépendant
- Lignes de défenses

Conclusion

Eventail de techniques pour travailler éthiquement et légalement sur des données sensibles (pas de cas d'usage de l'anonymisation)

Accepter les coûts induits par une gestion responsable des données.

7 histoires vécues d'utilisation ou accès à des données sensibles

1. Thèse économie industrielle sur la détection et prévention de la fraude aux mutuelles: données de santé, résultats confidentiels.
2. SmartDeliveries: projet de recherche sur des tournées de livraison, avec publications
3. Tests de performance chez un processeur de transactions bancaires sans accès aux données
4. Mise à jour de logiciel en production gérant des données sensibles
5. Prototypage d'un système de notification géolocalisée à des fins marketing
6. Visualisation de dossiers patients pour un service hospitalier.
7. Requête de suppression de données personnelles cross-entreprise

Enjeux de la protection des données pour une très grande entreprise de technologie informatique

- IBM, 300000 employés dans presque tous les pays du monde, une entreprise de plus de 100 ans, fondée pour traiter la donnée personnelle en masse.
- 1890: tabulatrices Hollerith pour traiter les données du US census.
- Tout le système de transfert interbancaire repose sur des mainframes IBM depuis 50 ans.

1. Sécurité des données et des process: une priorité générale, non-spécifiques aux données personnelles. Nos cadres contractuels usuels sont plus contraignants que le RGPD.
2. Pour nous distinguer de la concurrence, le slogan ‘vos données sont à vous’ est un point d’accroche important pour l’entreprise. Nous comptons dessus pour nous distinguer.
3. Nombreux métiers: conseil, infogérance, développement, recherche… avec une exposition au risque et des exigences variées.

Un socle commun pour la protection des données (et la conformité en général)

Formation

- Formation obligatoire annuelle (2 heures) pour tout le personnel, sous forme de MOOC.

-> inspiration directe de la formation 'Ethics & STICs' pour U. Paris-Saclay (avec le support de lq CERNA)

- + formations orientées 'conformité et éthique' spécialisées par métier: commercial, consultant, développeur, technicien… avec certifications.
- + centre(s) de ressources

Audit

- Organisation d'audit interne 'Business controls', rattachée à la direction mondiale
- Responsabilité au-delà de la seule protection des données: processus, bâtiments, contenu des systèmes…
- Un 'comité d'éthique' aux pouvoirs et budget conséquents.
- Pratique courante (pluri-centenaire) des industries fortement réglementées.

  https://www.ethics.org/

En préoccupation additionnelle, mesure de l'efficacité globale du dispositif:
Nombre et gravité des anomalies constatées ou projetées +
Pertes de productivité entrainées par la formation et les procédures de conformité.

Autour du dispositif

The Institute of Internal Auditors (IIA) Three Lines of Defense Model:

- The IIA issued "The Three Lines of Defense in Effective Risk Management and Control" Position Paper in January 2013. The Three Lines of Defense model provides a simple and effective way to enhance communications on risk management and control by clarifying essential roles and responsibilities.
- The **first line of defense** is the Operational Line owner who owns and manages risk on a day to day basis.
- The **second line of defense** includes Risk Management, Business Controls and Compliance functions that provide frameworks and oversight across the enterprise to monitor and assist the first line of defense in effective management of known and emerging risks.
- The **third line of defense** is Internal Audit that provides independent assurance.

(on le verra en application dans les exemples)

Sur les données elles-mêmes et l'anonymisation:

L'accent est mis sur les *usages permis ou non* et le contrôle d'accès plutôt que sur l'utilisation de transformations pour rendre des données ou process sensibles accessibles à d'autres usages.

On suppose qu'il n'est pas possible de détourner une donnée ou un process pour un usage autre que celui pour lequel il est conçu: pratique pour la conformité, ***plus ennuyeux pour la découverte 'par sérendipité'.***

Le RGPD est aussi conçu dans cet esprit: ce n'est pas tant l'existence des données et traitements qui est critique, mais bien l'usage qui en est fait.

7 histoires vécues d'utilisation ou accès à des données sensibles

1. Thèse économie industrielle sur la détection et prévention de la fraude aux mutuelles: données de santé, résultats confidentiels.
2. SmartDeliveries: projet de recherche sur des tournées de livraison, avec publications
3. Correctifs de performance chez un processeur de transactions bancaires, sans accès aux données
4. Mise à jour de logiciel en production gérant des données sensibles
5. Prototypage d'un système de notification géolocalisée à des fins marketing
6. Visualisation de données patients pour un hôpital.
7. Requête de suppression de données personnelles cross-entreprise

Thèse en économie industrielle: contexte

Contrat de services + infogérance sur un nouveau système de détection de fraude aux remboursements mutuelles.

- Analyse les demandes de remboursement, établi un profilage (par règles) et remonte des demandes suspectes à remonter pour audit (ou non).
- Le contrôleur des données est la mutuelle.

Plus-value du contrat: une thèse coencadrée avec un laboratoire d'économie industrielle pour analyser le retour sur investissement du système.

- En lutte contre la fraude, la prévention est un outil important. Prévention, Détection et Audit doivent aller ensemble. Comment? C'est le but de la thèse.

Nature des travaux

- A/B testing: informer (ou pas) les prestataires sur l'emploi de nouveaux outils de détection, mesurer le changement des comportements et les levées d'alertes.
- Installer le nouveau de système de détection, et mesurer les retours du système et son impact indépendamment des mesures prises précédemment.
- Modéliser pour estimer la non-détection et les faux positifs.

⇒ accès complet ou presque aux demandes de remboursement, qui sont des données médicales et financières. La pseudonymisation n'est pas réaliste.

Centre d’accès sécurisé aux données

La thèse démarrée fin 2015 sera soutenue fin 2019.

OPTIMOD'LYON

INVESTISSEMENTS D'AVENIR

ADEME
Agence de l'Environnement et de la Maîtrise de l'Energie

GRANDLYON
communauté urbaine

Develop high-value services, with self-sustaining business models

3 axes of development

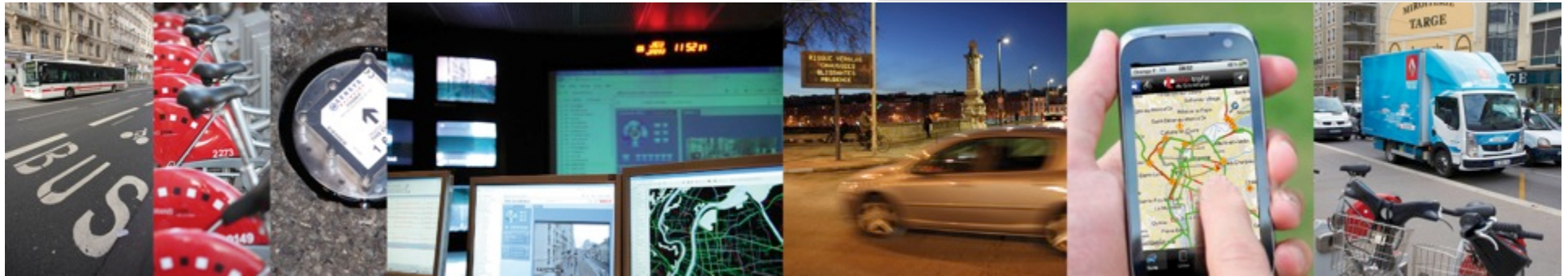
Novel real time data gathering systems (traffic and parking)

- **Mobility portal**
- **Centralized data warehouse**

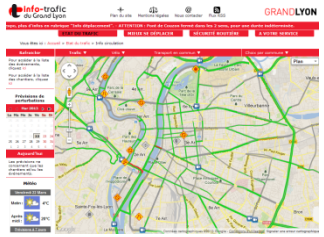
Optimizing traffic regulation Through 1h prediction

Real-time, multimodal, accessible travel planner

Optimizing freight and professional moves



Optimisation de tournées de livraison



City Monitoring center
(such as IBM IOC-IIT)

City regulation center

Known and forecast trafic conditions

Planned vehicle tours

Transporters, Maintenance crew managers…

Web application

advance information on planned mobility demand

Smart Deliveries

mission tracking

```
Original demand
10:00 -> 120 trucks at Part-Dieu
11:00 -> 160 trucks at Presqu'Ile
```

Global demand taken into account to spread trafic optimally

```
Optimized plans
10:00 -> 60 trucks at Part-Dieu,
         80 trucks at Presqu'Ile
11:00 -> 80 trucks at Presqu'Ile
         60 trucks at Part-Dieu
```

Optimized tour plans and trafic alerts

Drivers

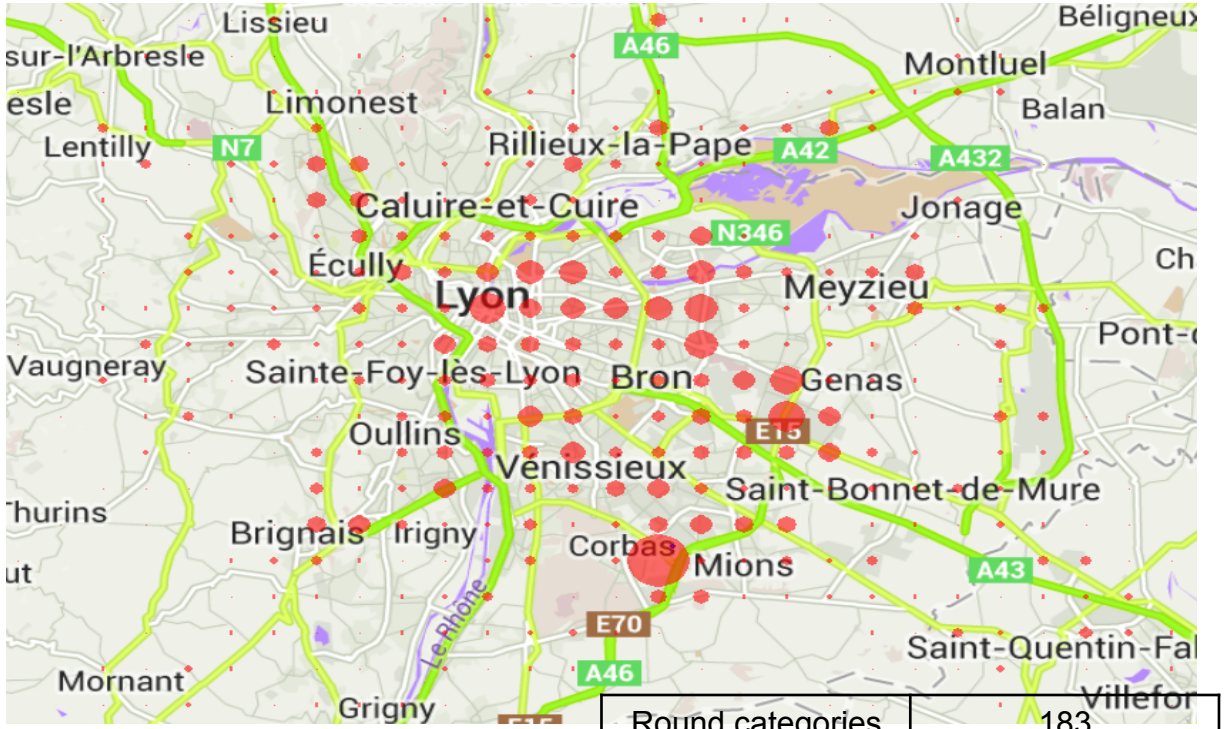
Mobile application

Optimisation de tournées de livraison

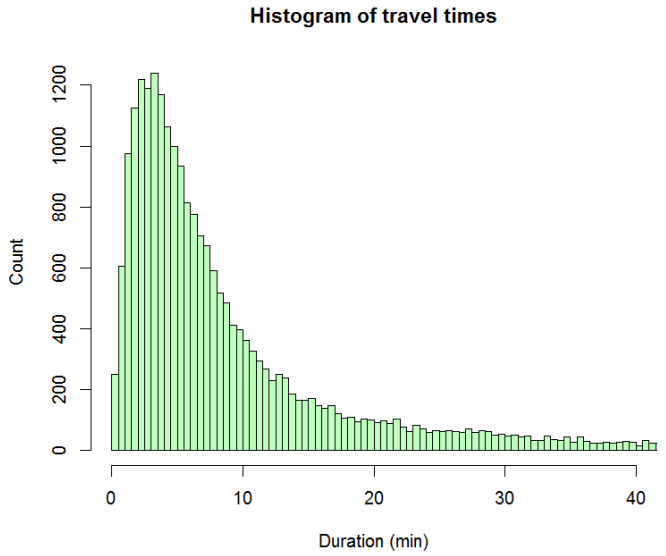
- 2012-2013: les transporteurs commencent à généraliser la géolocalisation de leurs camions
- 3 partenaires gros transporteurs acceptent de fournir des données de tournées réalisées pour optimisation, intéressés par le résultat potentiel.
- Les données sont cette fois-ci fournies par les transporteurs (contrôleurs) avec un contrat spécifique.
- Les destinations sont des commerces, les tournées sont numérotées: à priori, pas de données personnelles, mais données sensibles.

Pour la publication

- Les données de la ville sont publiques, avec une licence spécifique (droit de regard sur les usages): data.grandlyon.fr
- **Agrégation, floutage, et communication de certaines colonnes seulement** (temps de parcours, sans origine ni destination) à un chercheur demandant ces données.



	Actual	Optimized	Savings
distance:	63km	47km	25%
time:	12630s	10744s	20%
Arrives at	12h19	11h48	30min



Round categories	183
Rounds	1,715
Routes	~65,000
Routes after full cleansing	31,444
Routes per round	18
Average round travel time	2h24
Average trip time	10 min
Stddev trip time	15 min

Résoudre un bug sans accès aux données ni aux programmes.

- Un des plus grands centres de traitement de transactions par carte au monde: des millions de transactions par jour, SLA maximal.
- Chaque transaction engendre le déclenchement de règles de conformité, développées en interne et confidentielles, pour détecter des irrégularités potentielles.
- Le client se plaint de problèmes de performance, la R&D est impliquée.
- Aucun accès, ni aux données, ni aux programmes n'est autorisé.
- Seule une description du système installé et de la volumétrie des bases de règles sont fournies, ainsi que la possibilité de demander des statistiques sur les profils d'éxécution.

Solution:

- Reproduction de la solution matérielle complète dans un datacenter de test
- Création de bases de règles synthétiques
- Création d'un système d'alimentation en données synthétiques
- Tuning des données et règles synthétiques jusqu'à obtenir des profils de réponse similaire aux profils de réponse constatés chez le client
- Résolution des problèmes.
- **Beaucoup plus de travail que si données et programmes étaient accessibles.**

La synthèse de données artificielles à partir d'indicateurs ou de modèles (réels ou imaginés) devrait être un champ de recherche plus actif. Quelques articles, mais beaucoup de cas d'usage (tests, performance, démos…).

Maintenance logicielle ‘en nuage’

- C. est delivery manager pour une ligne de produits d’automatisation de la décision, fournie dans un service en nuage.
- Périodiquement, nécessité de mettre à jour le logiciel (continuous delivery) qui accède à toutes les données clients et fourni le service.
- Dans ce cas, les machines du centre de données sont accessibles par double authentification, avec traçage intégral de toutes les commandes réalisées par le mainteneur.

⇒ Pour chaque machine à mettre à jour, il faut une double-authentification pour lancer le script de mise à jour: ce qui pourrait se faire par un simple script allant sur toutes les machines réclame des manipulations fastidieuses.

⇒ acceptation des coûts supplémentaires au nom de la sécurité.

Geofencing pour applications marketing

- Que peut-on offrir comme nouvelles applications de l'informatique mobile avec des fonctions de capture du contexte (position et notifications diverses)?
- Travail expérimental mené par une équipe de développeurs avec une grande enseigne. Le but de l'expérience est de permettre de créer des notifications du type: *S'il pleut et que l'utilisateur est à proximité du magasin XX, alors proposer le message 'nous vous offrons un café en attendant la fin de l'averse'*

Geofencing II

Etudes de faisabilité technique locale (les développeurs comme sujets de leur expérience)

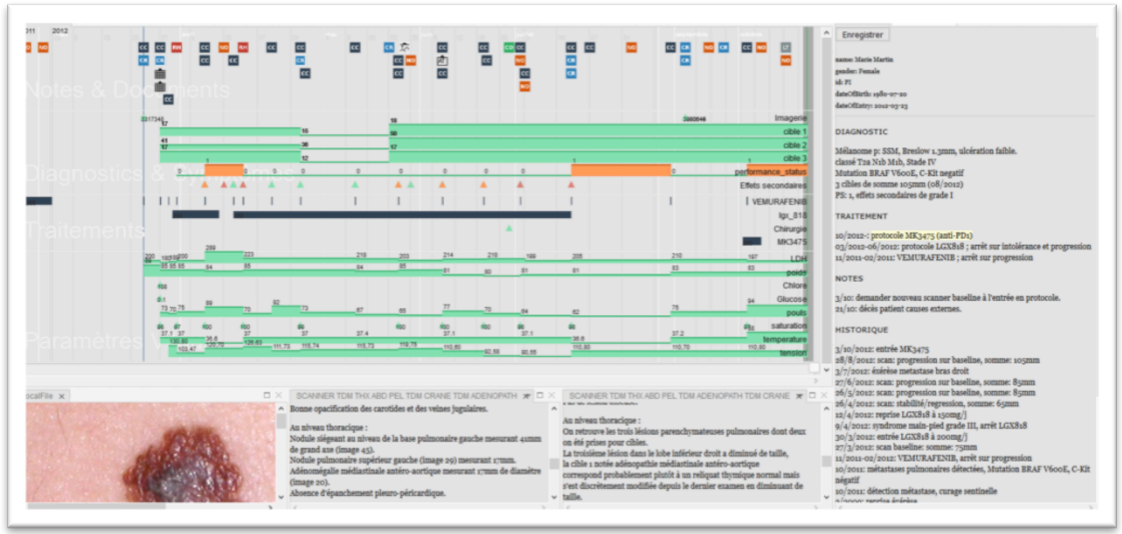
Réalisation d'une console permettant littéralement de superviser toutes les notifications reçues de tous les sujets.

Etude et discussions pour envisager une expérimentation in vivo.

"…Just as the war is too important to be left to the generals, human experimentation is too important to be left to the researchers and lawyers. If an experiment is good enough for your best customer, it's good enough for your best friend."

M. Schrage

Visualisation de dossiers patients



Les médecins ont besoin d'accéder à des visualisations de leurs cas patients. Données médicales.

Constitution de dossiers artificiels, « à la main » inspirés de cas réels, par une secrétaire médicale et un interne. Ce sont toutes les données de travail que nous avons.

Mise en place du logiciel et tests dans l'enceinte de l'hôpital, dans le service concerné (données non-anonymes, sinon pas de testabilité par les médecins).

Conclusion I : et l'anonymisation dans tous cela?

Les technologies utilisées pour effectuer des calculs sur données sensibles:

- Calcul sans accès aux données (CASD)
- Agrégation/Floutage pour rendu public
- Synthèse de données artificielles (de plusieurs types)
- Traçage intégral des actions réalisées
- Être son propre cobaye pour des applications à caractère sensible.
- Minimiser l'usage d'identifiants explicites/traçage intégral des flux de données lorsque des identifiants explicites sont utilisés.
- Travailler à l'aveugle ou presque (avec fortes limites)

+ formation généralisée avec rappels réguliers et 3 lignes de défense de protection des données.

Conclusion II

- Anonymisation – pseudonymisation: pas vraiment de cas d'usage flagrant dans les cas présentés. L'anonymisation fait craindre la perte d'information utiles, la pseudonymisation est un simple garde-fou, mais très insuffisante.
- Accepter les surcouts liés à la protection des données, en toutes circonstances.
- Selon une enquète de stackoverflow, en analyse de données: 60% du temps passé en collection de données et formatage, 20% features engineering et analyse, 20% reporting. Avec données sensibles, ce ne peut être que plus, soit des coûts de 3 à 5 fois le temps d'étude proprement dit.
- Une piste de recherche: synthèse de données artificielles
  - À base de statistiques externes (modèle graphique construit à la main)
  - À base de données sensibles (synthèse de modèle graphique et regénération)
  - Calcul homomorphique 'simplifié'