

Des faits aux données : le paradigme prédictif (Enass, 22 juin 2011)

Atelier 1 - la production des données : de l'Insee à Google

Dans l'histoire de la prédiction, la question récurrente est celle de la volonté face à l'aléa. Le modèle du jeu fonctionne sur le principe de données rares, coûteuses, difficilement manipulables. Une nouvelle ère s'ouvre avec une surabondance de données très malléables. S'agit-il d'un renouvellement des sciences humaines ?

Production et analyse des données

Le coût économique du stockage des données a beaucoup diminué ; ce qui coûte aujourd'hui 10 euros en coûtait 1 000 il y a 30 ans. Parallèlement, l'entrée des données, très coûteuse à l'époque de la mécanographie, se fait désormais de multiples façons. A la notion de donnée enregistrée une fois pour un usage précis s'oppose la réalité de milliards de données stockées, non structurées et facilement utilisables. « Big data » et réalité augmentée sont les tendances actuelles de la fourniture et du traitement de données.

L'accès à l'information et l'usage qui en est fait font partie de la problématique générale du stockage et/ou de la destruction des données.

Parallèlement à l'explosion de la numérisation se pose la question de l'énergie nécessaire, dont la consommation augmente, et de son prix. Tout est possible si le coût économique est acceptable par le plus grand nombre. La numérisation va s'étendre à l'ensemble de la connaissance. Le problème de la rémunération des droits reste posé ; une meilleure exploitation des données permettrait d'affiner la répartition de ces droits.

L'exemple de la génétique

La prédiction peut concerner l'individu ou le fonctionnement d'un système. La numérisation de l'univers s'adresse aussi à l'individu et à la matière et le traitement qui en est fait est lié à l'éthique. Par ailleurs, la production de données est indépendante de leur qualité : comment limiter le risque d'erreur et assurer la plus haute qualité possible ?

La médecine prédictive fait face à une inflation de données : ainsi, la production des données qui décrivent le génome humain équivaut à 3 milliards de caractères. Comment les conserver, y avoir accès, les analyser ?

Se pose la question de la propriété de ces données. Appartiennent-elles à celui qui paye pour les recueillir, c'est-à-dire le promoteur – par exemple l'Inserm –, à celui qui les gère, c'est-à-dire le scientifique ? En France, il n'existe pas de règles d'accessibilité. Aux Etats-Unis, au contraire, toute donnée publiée devient publique.

Le séquençage du génome permet de comprendre le processus d'une maladie, mais pas de prédire qui en sera atteint. La numérisation ressemble à une jungle et pose plusieurs problèmes non résolus (stockage, distribution, analyse, éthique).

Actuellement, la science entame le 4^{ème} paradigme de son évolution : elle se construit à partir de données planétaires traitées ensemble et mises à la disposition de tous les scientifiques, voire de tout un chacun. Les données s'enrichissent par interactions entre elles et plus uniquement par le jeu

d'algorithmes. La production de données correspond à la fourniture de services. Les interconnexions de bases de données entre elles permettent de créer de nouveaux services. La frontière entre monde expert et monde profane tend à s'estomper.

Le principe des « open data » se développe : une ville met par exemple à disposition des données de toutes sortes issues des bases de données qu'elle a constituées. Les données sont accessibles gratuitement, mais des applications en temps réel créées à partir de données publiques peuvent être payantes. Il s'agit de définir un modèle économique efficace. On voit apparaître un nouveau concept d'« open data » dans lequel des individus fournissent et partagent volontairement des données dans un but positif. L'utilisation des données devient alors citoyenne, voire politique.

Les principales tendances de l'évolution de la numérisation de données

Les producteurs de données deviennent concepteurs.

Les capacités à traiter les informations progressent moins vite que la masse de données disponibles. Une information d'intérêt a priori restreint rendue disponible par un individu peut toujours se révéler utile pour un autre individu.

Il y a croisement des différentes identités d'un internaute selon les communautés auxquelles il appartient.

La serendipité se révèle être une forme enrichissante de recherche intellectuelle.

Internet relie de façon neutre les données créées par des communautés très diverses. C'est un écosystème associant le global et le local, le réel et le virtuel.

A la différence des précédentes grandes révolutions industrielles et technologiques, la révolution numérique est au cœur de la vie quotidienne.

Atelier 2 : méthodes d'analyse et d'exploitation des données

Que sont devenues les probabilités ?

On observe aujourd'hui un déplacement. On est passé du calcul du risque, la probabilité, à une concentration sur le danger. La récente décision du gouvernement allemand de ne plus produire de l'énergie d'origine nucléaire suite au tremblement de terre et au tsunami au Japon est justifiée par le danger potentiel que représente cette énergie plutôt que par l'évaluation rationnelle et objective des risques. Cela signifie-t-il que les probabilités n'ont plus lieu d'être ? Au contraire. Les probabilités sont au cœur de tous les débats sur la maîtrise du risque et son évaluation.

La probabilité est une notion complexe :

- elle a un côté objectif : elle est un calcul et a une logique,
- mais elle a également un côté subjectif : chacun peut interpréter ces données différemment.

Données et risques

Grâce à l'entropie, il est possible de coder des variables de façon astucieuse et les transposer. Néanmoins, aujourd'hui nous sommes confrontés à une multitude de « données sales » selon Michel Bera, Professeur au Cnam. Ce sont des données entachées d'erreurs ou incomplètes qui sont très nombreuses en médecine, en marketing etc. On doit aller toujours vite pour mettre en place un plan marketing sauf que l'on s'appuie sur des données erronées. Dans le monde médical, avoir une base de données importante permet aux chercheurs d'aller beaucoup plus vite.

S'appuyer sur des « données sales » peut conduire à des situations de risques extrêmes comme ce fut le cas à Fukushima. Lors de la construction de la centrale nucléaire, on connaissait la présence d'une faille à proximité. Seulement, on s'était appuyé sur des données fausses, on pensait la faille nettement plus petite alors qu'en réalité elle est immense et a conduit à la catastrophe que l'on connaît.

La politique est également concernée par les risques et l'on observe des bouleversements politiques accélérés par Internet, comme ce fut le cas dans les pays du Maghreb. Par ailleurs, les Etats doivent prendre en compte les données pour éviter les problèmes, améliorer les situations sans entraver les libertés individuelles. C'est le prochain métier de la Cnil...

Données et interprétation

Avec l'arrivée de données en masse, on remarque une inversion. On passe du modèle qui découle d'une théorie à l'observation des données qui suggèrent des faits.

- Statistiques publiques : beaucoup d'observation mais peu de variables
- Recherche biomédicale : beaucoup de variables mais peu d'individus
- Et bientôt ? Tout sur tout le monde ?

On a un flux de données mais ce que l'on ne connaît pas, on l'estime. Le risque d'erreur est donc très important si l'on n'est pas prudent. L'outillage conceptuel du statisticien doit changer. Il doit trouver de nouvelles techniques pour concevoir des modèles. Mais faudrait-il concevoir des modèles pour comprendre ou pour prévoir ? Un « bon modèle » ne signifie pas nécessairement de bonnes prévisions au niveau individuel. On peut prévoir sans comprendre.

Quelles perspectives peut-on envisager ?

Nous sommes confrontés à une perte d'efficacité des critères classiques, de type sociodémographique, pour l'étude des comportements. Les données issues des réseaux sociaux permettent d'améliorer les prévisions. Le citoyen doit être situé dans son environnement et avec ses valeurs. Allons-nous vers de nouveaux critères de tri ?

Atelier 3 : domaines en jeu (génétique, climat, marketing, finance...)

La collecte et l'analyse des données peuvent avoir des utilisations diverses selon les domaines.

Jean-Claude Seys – Président fondateur de l'Institut Diderot

Il existe dans l'assurance un retard dans la collecte des données par rapport aux banquiers qui se servent de toutes les opérations de leurs clients pour les analyser.

Les données recueillies peuvent servir à ce que l'assuré devienne acteur de sa protection. L'assureur agit en « ange gardien » et prévient d'un risque comme par exemple informer son assuré de l'arrivée d'une tempête en lui envoyant un sms.

Dans le cas de la formule « pay as you drive », l'assuré tend à être moins assujéti à l'assureur qui en contrepartie recueille un certain nombre de données le concernant.

Il faut cependant nuancer cette utilisation des données dans le milieu de l'assurance car cela peut susciter des problèmes juridiques et déontologiques...

Antoine-Eric Sammartino – Chargé de mission – conduite du changement LaSer

En matière de marketing client pour les enseignes ou les marques, les données recueillies sont le fondement de la connaissance du client.

Il existe plusieurs types de données :

- identitaires : nom, âge, revenu, statut matrimonial...
- produits : code barre, prix, marge, stock...
- transactionnelles : informations du ticket de caisse (date, quantité, tarifs, remises...)
- déclaratives : enquêtes de satisfaction, aspiration, attitudes...
- externes : Insee, mégabases comportementales...

Un autre type de données apparaît également : les Big data. Il s'agit de ce que dit le client lors d'échanges sur le web (forums, blogs, réseaux sociaux) ou lors d'un appel à un call center.

On comptabilise également en données les actions du client sur un site e-marchand : comment il est arrivé sur le site, par quoi il commence ses achats, comment il remplit son chariot, où il se trouve...

Tout cela est possible grâce à des défis technologiques concernant le transfert, l'analyse et le stockage de données.

Un travail de nettoyage des données est nécessaire avant l'analyse, comme de savoir d'où viennent les données, leur cycle de vie..., en somme les contextualiser.

L'exploitation des données permet d'établir une segmentation (décrire le client) et de déterminer un scoring (noter le client), pour mettre en œuvre un marketing différencié qui implique le client dans son rapport à l'enseigne, qui le fidélise et l'influence, et avec lequel l'enseigne communique. Il s'agit d'un enjeu économique, d'image et relationnel pour l'enseigne.

A l'avenir on tend vers une analyse précise des personnes.

Concernant le scoring, il deviendra dynamique et on-line et pourra influencer le contenu d'un site marchand en fonction de la typologie du client.

Il se peut que le modèle relationnel évolue vers un partage des données avec le client afin qu'il puisse agir sur celles-ci, comme par exemple remonter un mauvais scoring en agissant différemment vis-à-vis de l'enseigne ou de la marque.

Amy Dahan-Dalmenico – Directrice de recherche, CNRS, Centre Alexandre Koyré (CNRS-EHESS-MNHN)

Concernant la relation de la climatologie avec les données, on peut dire que beaucoup de données sont récoltées et qu'elles impliquent beaucoup de visions. Il n'existe pas une seule définition du climat il s'agit plutôt d'un modèle qui le définit. Dans ce domaine on a donc plus à faire à des modèles qu'à des données.

Les controverses climatiques résultent de la définition de la limite des analyses, de ce qu'on inclut dans ces analyses... De plus on n'utilise jamais un seul modèle dans l'analyse car « la moyenne des modèles est meilleure que le meilleur modèle ».

En climatologie les données dépendent également de la qualité et de la précision des observations et des enregistrements automatiques. On utilise également des simulations pour les analyses afin de corréliser les données. Ces modèles climatiques permettent de prévoir les éléments extrêmes et de les communiquer aux personnes concernées.

Une telle complexité fait la spécificité de la climatologie dans son rapport aux données.

Bruno Racine – Président de la Bibliothèque nationale de France

Dans le monde des livres et des publications de toutes sortes (journaux, revues, documents audiovisuels, pages web...), la Bibliothèque nationale de France est un gigantesque stock de données.

La numérisation des collections physiques en mode plein texte et non plus en simple fac-similé métamorphose l'objet livre en un ensemble de données fragmentaires. On peut accéder au livre par un mot et non plus par son objet. Cela implique des questionnements sur le classement de ces données et sur l'évolution de la typologie des catalogues de bibliothèques.

La BNF est en train de développer un outil qui fera en sorte que les moteurs de recherche trouvent le livre numérisé et le placent en tête des résultats lors d'une recherche. Pour l'instant il faut directement aller sur le site de Gallica pour accéder à ce document car le catalogue n'est pas identifiable par le moteur de recherche. Ce futur catalogue intègre en fait des métadonnées qui décrivent le contenu du livre et non plus seulement son objet (mots clés, résumé, biographie...).

La bibliothèque physique demeure dans un but patrimonial mais la numérisation permet un accès aux données à tout le monde.

Se posent alors les problèmes de la conservation pérenne de ces données numériques et de leur stockage qui ne prend plus de place mais qui est devenu techniquement complexe.

Désormais, on laisse le choix au lecteur entre l'accès physique, sur place, et l'accès numérique, à distance. De son côté, le livre numérique ne se résume pas seulement à un fac-similé du livre physique, il peut inclure des ajouts comme un résumé, des critiques, des sources...

La définition actuelle de notre droit d'auteurs français va à l'encontre de la logique communautaire du web et de l'accès numérique. Il s'agit là d'un des obstacles à la diffusion de ces données numériques, car tout est numérisable mais tout n'est pas communicable.

Philippe Ravaud – Professeur, Centre d'épidémiologie clinique, Hôtel Dieu, Université Paris Descartes et Directeur du centre Cochrane français

Concernant le milieu médical, l'utilisation des données sert à évaluer les traitements. La connaissance du meilleur état de la science permet la qualité des soins, même s'il existe un délai entre la connaissance et la pratique.

La médecine est désormais définie par les preuves car les traitements efficaces peuvent aussi être dangereux. Le plus bas niveau de preuve correspond aux experts scientifiques à cause des conflits d'intérêts avec les labos, du problème d'actualisation des connaissances... !

Il existe plusieurs méthodes pour évaluer les traitements :

- les essais randomisés (2 traitements effectués sur 2 groupes de patients). Environ 500 000 essais par an et 75 par jour
- les méta analyses (méthode statistique qui combine les résultats individuels de plusieurs études indépendantes en un résultat commun). Leur durée de vie est de 2 ans.
- les études observationnelles

Malgré tous ces essais et analyses, les Américains ont pris récemment conscience qu'on ne connaît pas vraiment les traitements qui marchent ! En effet il y existe un délai entre la transposition des résultats et la pratique.

Les médecins de leur côté ont des difficultés à actualiser leurs connaissances (peu de formation, peu de lectures scientifiques). 90% de l'information sur les médicaments vient des laboratoires qui les commercialisent.

Enfin les publications négatives sur les études sont limitées à cause notamment de la non-transparence des données sur les essais effectués : pas d'accès aux données individuelles seulement à des informations sur le protocole utilisé. L'accès aux bases de données des résultats et des publications pour les chercheurs est plus difficile. D'autre part l'accès aux données est également restreint par la Cnil et divers acteurs comme les industriels, les assureurs... On est donc loin de la médecine personnalisée. Le monde des données médicales ne peut efficacement fonctionner que s'il y a un décloisonnement des secteurs.

Atelier 4 : questions éthiques : prélèvement, conservation, usages des données

Isabelle Falque Pierrotin, Conseiller d'Etat, Vice-présidente de la CNIL

Les données numériques sont des données au carrefour d'enjeux complexes et pour lesquels il est primordial de définir la gouvernance. Nous sommes aujourd'hui confrontés à un flux inédit de données, données qui ne sont pas limitées par les capacités traitement, des données qui peuvent

donc se développer à l'infini et qui se sont de plus en plus « personnalisées ». Pour Isabelle Falque Pierrotin, CNIL, ces données sont le « pétrole du système internet » : un individu accepte de partager ses données en échange d'un service gratuit qui est financé par la publicité.

La mission de la CNIL dans un tel système est de trouver le meilleur équilibre pour que le gisement de données disponibles puisse être utile et puisse servir l'individu sans pour autant mettre en cause sa liberté individuelle.

La CNIL fait donc en permanence un bilan des bénéfices et dangers pour l'individu dans l'utilisation des données.

⇒ **Les bénéfices peuvent être d'ordres divers :**

- Offrir un meilleur service (en marketing, en médecine prédictive, en assurance...)
- Nous aider à prendre des décisions (choisir un client, lutter contre la fraude...)

⇒ **En revanche, derrière ces applications positives, il peut y avoir des risques :**

- Les décisions assistées par ordinateur déresponsabilisent les personnes qui prennent ces décisions. Ils se réfugient derrière la preuve apportée par la machine. Par exemple, en matière de criminologie, la preuve digitale fait aujourd'hui foi, au risque d'erreurs de jugement. Pour prendre un exemple plus trivial, les sites internet de rencontres opèrent une sélection des individus correspondant à des critères définis. Là encore, l'individu est totalement déresponsabilisé dans ses choix. Isabelle Falque Pierrotin estime donc que nous entrons dans une zone à risque dès lors que la loi stipule « qu'une décision ne peut être prise sur le seul fondement d'un traitement informatique ».
- Par ailleurs, le gisement des données n'est pas toujours fiable. Isabelle Falque Pierrotin prend l'exemple du fichier des antécédents judiciaires qui n'est pas à jour. Celui-ci est consulté dans le cadre de l'emploi de certaines professions (sécurité...) On imagine donc les risques encourus en se fiant uniquement à cet outil.
- Enfin, Isabelle Falque Pierrotin indique que la CNIL est particulièrement vigilante au risque de « profilage » des individus, qui sont catégorisés à travers leur mode de consommation, d'action... Les individus sont finalement jugés sur les traces qu'ils laissent et non sur ce qu'ils sont vraiment.

La CNIL agit donc à deux niveaux pour définir une sorte « d'ordre public de protection de l'individu » :

⇒ **Dans le flux de données, il importe de créer de la « discontinuité » :**

- La loi de 2002 sur le droit des malades indique que les assureurs ne peuvent utiliser les données génétiques
- Le droit à l'oubli est un engagement des acteurs internet (Google, Facebook...) à désindexer des informations qui n'ont pas été modifiées depuis un certain temps.

⇒ **Il faut imposer des process respectueux de l'individu :**

- Encourager la collégialité dans les prises de décisions
- Permettre une transparence dans la catégorisation des individus par les ordinateurs

Thomas Berns, Philosophe politique, Université libre de Bruxelles

Parmi les pratiques politiques, il existe deux façons de se référer au réel :

- Celles qui gouvernent – dominant – le réel
- Celles qui se construisent à partir du réel, c'est à dire qui partent des faits, des données pour agir

Nos sociétés démocratiques sont nettement dans le 2^e cas de figure. Le gouvernement garantit la liberté des sujets par une gestion « statisticienne » de la société. Et l'individu accepte cette relation « rationnelle ». Or, le phénomène actuel de l'automatisation de la récolte des données met à mal cette relation de confiance. Cette automatisation se retrouve à 3 niveaux :

1. La récolte et la conservation des données

La récolte des données est totalement décorrélée de l'usage qui en est fait. Des quantités infinies de traces sont laissées par les individus sans qu'ils en aient réellement conscience, sans aucune intentionnalité, sans aucun consentement « éclairé ». Nous abandonnons des données car nous ne les considérons que comme des traces inoffensives. Or ces données sont d'une telle « réalité », d'une telle « précision », qu'elles apparaissent comme des doubles de la réalité. Elles sont en quelque « irréfutables ».

2. Le traitement des données

On se retrouve face à une production de savoirs tout à fait hétérogènes. Les « réalités » déduites des données sont produites en dehors de toute hypothèse préalable. Et ceci renforce encore l'objectivité des « preuves laissées ».

3. L'usage des données

Là encore le sujet est totalement « désobjectivé ». Le gouvernement ne s'exerce plus sur des individus mais sur les multiples facettes de ces individus.

Dans cette société « inédite », le sujet doit donc se « réinventer » au risque de ne devenir qu'une « somme de données ».