75 PNR
Big Data
Programme national de recherche

Big Data

applications, technologies et aspects sociétaux

Résumé
du Programme national de recherche
«Big Data» (PNR 75)

Fonds national suisse

Avant-propos

La généralisation de la collecte des données et de leur traitement a entraîné de profonds changements dans notre mode de vie. Le big data s'avère aujourd'hui indispensable pour une économie compétitive et axée sur la technologie. Il recèle un énorme potentiel pour faire progresser le savoir ainsi que pour l'ensemble de la société. En parallèle, il génère de nombreux défis sociétaux. Il constitue une priorité majeure pour la recherche de plus haut niveau.

Le Programme national de recherche «Big Data» (PNR 75) a ciblé la recherche technologique fondamentale et appliquée ainsi qu'à orientation sociétale. Un jury international a sélectionné 34 projets de recherche parmi les propositions soumises à l'appel organisé par le Fonds national suisse. L'objectif du PNR 75 n'était pas de répondre à des questions technologiques ou sociétales prédéfinies, mais de faire progresser les capacités en Suisse dans ce domaine par une intensification de la recherche. A travers leurs travaux, des chercheuses et chercheurs en ingénierie et en sciences naturelles et sociales ont lancé des débats sur des questions touchant au droit, à la confidentialité et la souveraineté en matière de données, à la fracture numérique, à l'équité, à la responsabilité des algorithmes ainsi qu'au niveau approprié de réglementation.

Après cinq ans de travail, les recherches du PNR 75 démontrent les possibilités offertes par les applications du big data. Elles soulignent qu'une recherche fondamentale de haute qualité menée en Suisse peut directement contribuer aux technologies essentielles au big data. Ces travaux soulignent que développer des solutions nécessite une perspective globale qui inclut dès le départ les dimensions éthiques, juridiques et socio-économiques.

Le PNR 75 complète d'autres initiatives de recherche telles que Digital Lives (2018–2019) et le Programme national de recherche «Transformation numérique» (PNR 77, 2020–2025), qui étudient des sujets sociétaux spécifiques liés à la digitalisation. Ces actions renforcent les capacités

de recherche, d'innovation et de formation en Suisse, comme le prévoit la stratégie «Suisse numérique» du Conseil fédéral.

Les travaux réalisés dans le cadre du PNR 75 soulignent l'importance d'efforts interdisciplinaires pour le développement de ces technologies émergentes et pour leur utilisation responsable. Parmi les défis se trouvent notamment l'équilibre entre protection de la vie privée, sécurité des données et possibilités d'accéder et de partager les données, ainsi que la nécessité de développer les compétences et recruter les talents nécessaires pour concevoir, développer et intégrer les connaissances et l'expertise dans les processus courants.

Le PNR 75 a largement contribué à renforcer les compétences en matière de big data en Suisse, les cadres propices à l'innovation interdisciplinaire ainsi que les capacités à trouver des solutions sociales et juridiques appropriées. Ce résumé représente un extrait précieux des résultats et des conclusions du Programme national de recherche «Big Data».

Bert Müller
Délégué de la division Programme du Conseil national de la recherche du Fonds national suisse depuis janvier 2021

Friedrich Eisenbrand
Délégué jusqu'à décembre 2020

Executive summary

La numérisation de la société en cours se traduit par la collecte de jeux de données de très grande taille. Ce que l'on appelle le big data recèle un potentiel de création de valeur sociétale, industrielle et scientifique considérable, à condition d'être exploité de manière efficace. De 2017 à 2021, le Programme national de recherche «Big Data» (PNR 75) a mené 37 projets qui ont développé des applications concrètes, inventé de nouvelles technologies et étudié des aspects sociétaux en lien avec le big data, renforçant ainsi la capacité de recherche et d'innovation de la Suisse dans ce domaine.

Quinze projets de recherche ont produit des applications du big data concrètes à travers des collaborations réunissant des spécialistes en informatique et des secteurs concernés (chapitre 2). Les résultats montrent l'impact concret que ce type d'innovation peut avoir dans des domaines tels que la planification des énergies renouvelables, la surveillance des personnes hospitalisées, l'élaboration de politiques basée sur les faits ou encore la recherche scientifique. Les projets soulignent l'importance de constituer des équipes interdisciplinaires capables de s'adapter aux contextes éthiques, juridiques et opérationnels.

Onze projets en informatique ont développé ou amélioré des technologies nécessaires à l'exploitation du big data. Ils couvrent des aspects liés à l'infrastructure, notamment l'accès aux données et leur nettoyage, indexation et prétraitement, ainsi que des défis liés à l'analyse des données pour en extraire des connaissances, en particulier le traitement des requêtes, l'exploration des données et l'apprentissage automatique (chapitre 3). Ces avancées contribuent à améliorer la fonctionnalité et les performances des applications du big data, par exemple en renforçant la confidentialité ou en réduisant les ressources de calcul et de données utilisées pour entraîner des modèles d'apprentissage automatique.

Huit projets ont étudié des aspects sociétaux, éthiques et juridiques du big data (chapitre 4), y compris des cas concrets d'utilisation du big data dans les ressources humaines et les assurances. Leurs résultats soulignent l'importance d'adapter la législation aux évolutions technologiques, d'élaborer des lignes directrices, d'accroître la sensibilisation à l'éthique et à la transparence, et de suivre de près l'impact du big data sur la démocratie.

Trois projets transversaux ont analysé les obstacles au partage des données scientifiques, renforcé la participation des femmes dans la recherche sur le big data et

élaboré une vue d'ensemble des enjeux sociétaux majeurs liés au big data.

L'évolution technologique rapide dans ce domaine offre des opportunités dans de nombreux secteurs comme la production industrielle, les énergies renouvelables, la cybersécurité ou encore le commerce électronique (chapitre 5). Elle soulève également de nombreuses questions, notamment sur la consommation croissante d'énergie pour le traitement des données, l'équilibre entre le respect de la vie privée et la création de valeur, les risques de discrimination ainsi que les questions de responsabilité. Il est crucial de relever ces défis pour optimiser la création de valeur à partir des données par les entreprises et les institutions publiques.

Les conclusions du Comité de direction du PNR 75 suggèrent des pistes pour encourager la création de valeur responsable à partir du big data qui peuvent contribuer aux débats politiques et professionnels sur cette nouvelle ressource. Elles sont résumées ci-dessous et sont détaillées dans le rapport (chapitre 6).

Conclusions du comité de direction

Favoriser un environnement approprié pour le développement du big data
(1) Améliorer la formation au big data
(2) Soutenir le conseil juridique et éthique pour les projets de R&D en big data
(3) Permettre la certification des applications du big data

Intégrer le big data dans les organisations publiques et privées
(4) Accroître l'exploitation des technologies du big data dans le secteur de la santé
(5) Renforcer l'élaboration et l'évaluation des politiques grâce au big data
(6) Promouvoir la collecte partagée des données, l'open-source et les benchmarks

Actualiser et créer des réglementations adéquates
(7) Poursuivre une réglementation proactive du big data
(8) Assurer la confidentialité des données et la souveraineté numérique
(9) Renforcer l'harmonisation transnationale des réglementations

1.

Introduction: big data, grands changements

Des technologies matérielles et logicielles de plus en plus sophistiquées permettent de collecter et d'analyser des volumes de données sans précédent. Celles-ci offrent un potentiel de création de valeur de grande envergure dans les secteurs public et privé. La mise en œuvre responsable d'applications nécessite des recherches sur tous les aspects du big data, notamment les infrastructures informatiques, les méthodes d'analyse des données ainsi que les recommandations éthiques et les cadres juridiques. Le PNR 75 a apporté des contributions importantes tout au long de cette chaîne de valeur et a renforcé la capacité à développer les technologies nécessaires, à déployer des applications et à adapter la réglementation en Suisse.

1.1 Le rôle croissant des données

De la numérisation au big data

Alimentées par la recherche et l'innovation publiques et privées, les avancées technologiques se produisent à un rythme accéléré et affectent profondément notre mode de vie. Les progrès des technologies de l'information entraînent une numérisation généralisée de la société et une profusion de données avec un impact grandissant sur notre façon de vivre et de travailler.

Les capacités de traitement, de stockage et de transfert des données se développent à un rythme exponentiel depuis plusieurs décennies. La densité des transistors sur les micropuces a ainsi doublé tous les deux ans, entraînant une accélération similaire de la puissance des processeurs et des débits de transmission des données.

Ces avancées spectaculaires sont rendues possibles en partie par les très grandes économies d'échelle réalisées dans les secteurs des semi-conducteurs et des technologies de communication, ainsi que par la croissance très rapide du marché. Couplées aux progrès des technologies logicielles (notamment les systèmes d'exploitation et de gestion des données, les langages de programmation et compilateurs ainsi que les méthodes d'analyse telles que l'apprentissage automatique), elles ont permis des gains sans précédent en matière de traitement de l'information, ont créé d'innombrables nouveaux outils, et modifié profondément les secteurs économiques, les pratiques professionnelles et les habitudes de vie.

Les sphères publiques, économiques et privées toujours plus numérisées produisent des jeux de données toujours plus importants. Le terme «big data» fait généralement référence à des jeux de données dont les propriétés posent des défis aux technologies de l'information et de la communication disponibles, et exigent ainsi de nouvelles solutions. Au-delà de sa taille (ou volume), le terme «big data» peut également se référer à la vitesse de création et de traitement des données, à leur variété et à leur véracité (voir «Qu'est-ce que le big data?», p.11). Le champ d'application du big data évolue donc continuellement au gré des avancées technologiques.

Le big data documente une part croissante de notre vie sociale, professionnelle et individuelle, des transactions commerciales, des dispositifs industriels ou encore de la recherche scientifique. Certains grands jeux de données sont de nature personnelle, d'autres non. Les données sont collectées par les sites Web, les applis, les caméras et les capteurs déployés dans les smartphones, les véhicules, les chaînes de production industrielle, ou encore les dispositifs de surveillance de l'environnement.

Créer de la valeur

Les données sont considérées comme une ressource extrêmement précieuse qui n'a pas encore été pleinement exploitée, un nouvel «or noir» pouvant alimenter de nombreux processus et jouer un rôle central dans la société.

Les données ont peu de valeur en soi. Leur valeur est extraite sous la forme de résultats d'analyse qui sont exploitables. Une planification est nécessaire pour identifier les questions à poser et les données susceptibles de fournir des réponses. Il faut déterminer comment les données peuvent être collectées

ou rendues accessibles, développer des outils d'analyse efficaces et transformer les résultats de ces analyses en actions créatrices de valeur. Les éventuels effets secondaires doivent être évalués tout au long du processus.

Cette approche n'est pas nouvelle. Elle est suivie depuis des décennies dans les études de marché, les enquêtes auprès de la clientèle, la planification financière ou encore l'épidémiologie. Le big data apporte cependant une nouvelle dimension par la taille des jeux de données, les infrastructures de calcul et de communication requises, la complexité des algorithmes ainsi que l'étendue des applications et les défis qu'elles posent. De nombreux exemples montrent comment le big data peut permettre des modèles plus fins permettant d'améliorer des processus commerciaux, industriels et gouvernementaux.

L'utilisation des données a toutefois généré de nouveaux risques et défis pour la société, en termes de sécurité, d'équité et de cohésion sociale. Tout cela nécessitera des solutions adéquates et proportionnées.

Applications pour le numérique et l'analogique

Les exemples familiers d'applications du big data comprennent des services en ligne tels que les réseaux sociaux, les services de messagerie et de courrier électronique, les moteurs de recherches et la publicité, ainsi que les plateformes de commerce électronique ou de divertissement.

Les domaines de l'énergie et des transports utilisent également le big data. Par exemple, des éoliennes équipées de dizaines de capteurs peuvent générer des dizaines de mesures par seconde. Ces informations peuvent être analysées en temps réel pour affiner le réglage du pas des pales des turbines et maximiser leur efficacité. Pendant le vol d'un avion, les capteurs peuvent produire un téraoctet de données. Correctement transférées, stockées et analysées, elles permettent de surveiller les composants et de programmer leur entretien avant qu'une défaillance ne survienne, ce optimisant la gestion de la flotte et réduit les temps d'arrêt[1].

De très grands jeux de données sont couramment utilisés pour la surveillance et la gestion de l'environnement, permettant par exemple aux Nations unies d'évaluer la perte de biodiversité et de suivre les effets du changement climatique[2], ou aidant à prévoir la répartition des polluants dans l'air ou dans l'eau[3]. Un consortium de journalistes a mis au jour des évasions fiscales internationales systématiques en analysant des millions de documents ayant fait l'objet de fuites[4].

La recherche produit du big data depuis des décennies dans le cadre de collaborations internationales en physique ou en astronomie, comme au CERN. Le big data est utilisé désormais dans un nombre croissant de disciplines, de la biologie aux sciences humaines. Par exemple, une avancée majeure dans les sciences de la vie a eu lieu en juillet 2022 avec la publication d'une énorme base de données des

Qu'est-ce que le big data?

Le big data est un concept en constante évolution, car il décrit par définition des jeux de données dont les propriétés défient les technologies disponibles, qui s'améliorent continuellement.

Le **volume** (taille) des données dépasse généralement les gigaoctets (Go) pour atteindre les téraoctets (1 000 Go), voire les pétaoctets (1 000 To), ce qui nécessite une infrastructure de stockage et de traitement très puissante. La **vitesse** des données (le taux de production, de transfert ou d'analyse) peut dépasser 1 Go/seconde, ce qui exige un matériel rapide et des logiciels efficaces.

Les applications combinent souvent des types de données hétérogènes (texte, chiffres, coordonnées, images, son, vidéo, etc.) aux caractéristiques très différentes: une trace GPS est très précise alors que la sémantique textuelle reste souvent ambiguë. Cette **variété** nécessite des algorithmes capables de gérer de multiples formats et types de données.

Les données sont rarement sans erreurs, précises, représentatives et complètes – des propriétés englobées dans le concept de **véracité**. De nombreuses applications du big data reposent sur des modèles plus ou moins précis entraînés avec des données d'apprentissage de qualité variable, ce qui influence la **validité** des résultats.

D'autres «v» sont parfois utilisés pour décrire une application du big data, comme la variabilité des données, leur vulnérabilité, leur visualisation ou leur valeur.

---

1 The case for an industrial big data platform, General Electrics (2017)
2 World environment situation room, United Nations Environment Programme, https://data.unep.org/
3 A new area of utilizing industrial Internet of Things in environmental monitoring, HH Lou et al. (2022) Front. Chem. Eng. 4:842514.
4 The Panama papers: exposing the rogue offshore finance industry, , International Consortium of Investigative Journalists., https://www.icij.org/investigations/panama-papers

La taille du big data
Les applications traitent des jeux de données jusqu'à l'échelle du pétaoctet (un million de gigaoctets).

1 Gigaoctet (Go)

– 1 film

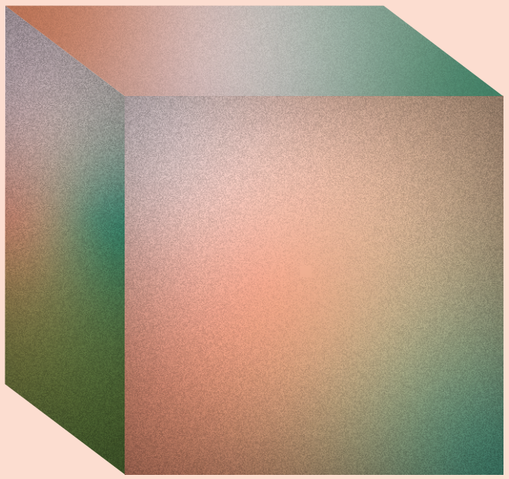
– 1000 livres

1 Téraoctet (To)

– 1 disque dur externe

– 1000 films

– données pour entraîner des modèles de langage artificiels



1 Pétaoctet (Po)

– images médicales produites dans un hôpital en un mois

– vidéos uploadées chaque jour sur YouTube

1 Exaoctet (Eo)

– trafic mobile mondial quotidien

formes possibles de la quasi-totalité des 200 millions de protéines connues, prédites par des algorithmes d'apprentissage automatique. Il s'agit d'informations cruciales pour comprendre les rôles des protéines dans les processus physiologiques, prédire leurs actions et concevoir de nouveaux traitements[5].

Assurer un usage responsable

La puissance des applications du big data génère des risques variés. Créer de la valeur de manière responsable nécessite des solutions appropriées et proportionnées aux défis liés à la vie privée, à l'éthique, aux affaires, au droit et à la gouvernance.

Par exemple, des pirates informatiques exposent régulièrement les informations privées de millions de personnes. Certaines plateformes Internet ont abusé des données (y compris médicales) des personnes les utilisant, tandis que le profilage de la clientèle peut compromettre la vie privée. Les algorithmes d'apprentissage automatique amplifient les biais présents dans les données utilisées pour les entraîner et peuvent produire des résultats discriminatoires[6]. Les entreprises de l'économie des plateformes (par exemple dans le domaine des transports ou de l'hébergement) ont déjà bouleversé le marché du travail et remis en cause ses réglementations. La résolution de ces problèmes constitue une tâche globale impliquant les institutions privées et publiques qui produisent, stockent, transfèrent, analysent et utilisent le big data ainsi que les administrations publiques régionales, nationales et internationales, les ONG et la population. Une adaptation rapide de la législation actuelle et l'adoption de nouvelles lois sont nécessaires pour garantir que les applications du big data respectent les principes fondamentaux de vie privée, d'équité, de transparence, de responsabilité et de non-discrimination.

1.2 L'importance d'une recherche solide

La recherche scientifique joue un rôle central pour les fondements des applications du big data. Elle contribue à développer les infrastructures nécessaires, telles que des capteurs à faible consommation d'énergie produisant des données pour l'Internet des objets, ou des technologies capables de stocker, transférer et traiter de grandes quantités de données. Elle fournit aussi de nouveaux outils pour développer les analyses et les prédictions sur lesquelles se basent les applications, tels que des méthodes statistiques avancées, des algorithmes d'exploration de données et l'apprentissage automatique.

Il est également est crucial de mener des recherches pour mieux comprendre les impacts sociétaux du big data. Des recommandations fondées sur les faits sont nécessaires pour relever les défis existants et émergents, notamment en matière de surveillance, de réglementation et de nouvelles pratiques au sein des entreprises et gouvernements. Une recherche universitaire solide constitue également la

---

[5] The entire protein universe, Ewen Callaway, Nature 608, 15-16 (2022). Quelque 200 millions de structures de protéines sont stockées par le consortium international Uniprot, né de l'initiative suisse Swissprot en 2003.

[6] Why algorithms can be racist and sexist, Rebecca Heilweil (Vox, 2020)

base d'un enseignement de haute qualité, un élément central de la Stratégie Suisse numérique[7].

La Suisse a besoin de compétences de pointe en matière de big data

Le maintien d'une recherche suisse hautement compétitive sur le big data revêt une importance stratégique.

— Il est crucial de rester en contact avec les développements internationaux qui façonnent le big data. Les projets de recherche et d'innovation n'ont accès aux spécialistes de pointe et aux connaissances les plus récentes que s'ils ont quelque chose à offrir, les meilleurs travaillant uniquement avec les meilleurs.
— Le haut niveau de la recherche et l'éducation en Suisse a attiré d'importantes entreprises de big data et de grands centres de recherche industrielle.
— La Suisse doit s'assurer qu'elle peut former, attirer et retenir des spécialistes du big data, de plus en plus recherchés au niveau international.
— De nombreux aspects du big data, notamment les aspects éthiques, juridiques et sociétaux, sont spécifiques à la Suisse et requièrent l'avis de scientifiques sur place.
— Les entreprises multinationales repoussent les limites des technologies du big data et déterminent la progression du domaine. Une communauté de recherche publique forte est nécessaire pour garder un certain contrôle sur la technologie et ses orientations.
— La recherche contribue à l'information de la population. Les scientifiques diffusent les nouveaux résultats de recherche et aident le public, le gouvernement et les entreprises privées à contribuer de manière informée aux processus et prises de décisions démocratiques.
— La recherche de pointe contribue à une éducation de haute qualité, élément essentiel pour une main-d'œuvre privée et publique qualifiée.

Ce sont les principales raisons pour lesquelles le renforcement de la recherche sur le big data revêt une importance stratégique pour la Suisse. Une contribution considérable a été apportée par le Programme national de recherche «Big Data».

1.3 Le Programme national de recherche «Big Data»

Nouvelles perspectives sur les technologies, les applications et les aspects sociétaux

Le Programme national de recherche[8] «Big Data» (PNR 75) a été conçu en 2014/2015. Il complète les programmes stratégiques nationaux de soutien à la numérisation tels que la Stratégie Suisse numérique du Conseil fédéral, l'initiative intersectorielle DigitalSwitzerland, la Swiss Digital Initiative ainsi que des programmes de recherche dédiés tels que Digital Lives.

---

[7] https://digital.swiss

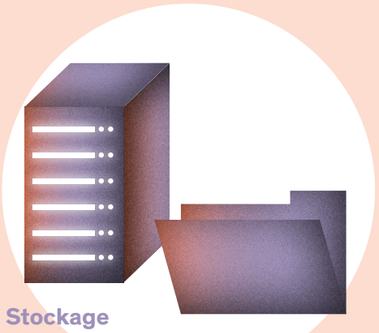
[8] Les programmes nationaux de recherche (PNR) permettent à des consortiums de recherche thématique d'aborder des sujets d'importance pour la Suisse. Ils sont proposés au secrétariat d'Etat à l'éducation, à la recherche et à l'innovation par des unités administratives, des instituts de recherche ou des particuliers. Ils sont approuvés par le Conseil fédéral et mis en œuvre par le Fonds national suisse (FNS).

La chaîne de valeur du big data
Créer des applications à partir de big data passe par différentes étapes, chacune ayant ses propres défis. Les **données** sont produites par des activités sur le Web ou des capteurs avant d'être acquises, nettoyées, vérifiées et anonymisées. Des **technologies** matérielles et logicielles sont nécessaires pour l'infrastructure du big data: stocker les données, en gérer l'accès et le sécuriser, effectuer le prétraitement et mener des analyses.

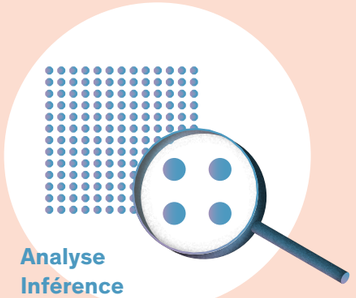
Les **applications** modélisent les données pour produire des analyses, prédictions et recommandations. Leur **utilisation** en pratique exige de les valider et de les intégrer dans les processus existants. Des **questions juridiques, éthiques et sociétales** se posent tout au long de la chaîne de valeur, de la confidentialité des données aux biais algorithmiques et à la réglementation. Le Résumé du PNR 75 couvre l'essentiel de ces étapes.

Le chapitre 2 décrit les applications développées par le PNR 75 en santé, durabilité, socio-économie et recherche.

Production
Acquisition
Accès
Nettoyage
Vérification
Anonymisation



Stockage
Partage
Contrôle d'accès
Sécurité
Compression des données
Traitement
Prétraitement



Analyse
Inférence
Prédiction
Modélisation
Augmentation

Fiabilité
Applicabilité
Intégration
Acceptation
Entretien

Données → Technologies → Applications → Utilisation

Propriété des données
Contrôle
Accès
Biais

Accès aux sources de données
Accès aux infrastructures
Accès à la technologie
Cybersécurité

Accès aux données personnelles
Confidentialité
Consentement éclairé
Équité
Black box

Vie privée
Asymétrie numérique
Régulation et certification
Autonomie
Confiance
Innovation

Questions juridiques, éthiques et sociétales

Le chapitre 3 résume les progrès des recherches sur les infrastructures du big data (matériels et logiciels) et sur les méthodes d'**analyse** nécessaires au prétraitement des données, à la gestion des serveurs ou à l'amélioration de l'apprentissage automatique.

Le chapitre 4 aborde les questions sociétales, juridiques et éthiques touchant l'ensemble de la chaîne de valeur – la propriété des données, le contrôle sur elles, leur accès, le consentement éclairé, l'asymétrie numérique et la vie privée – et discute la réglementation et les lignes directrices.

Le chapitre 5 propose une perspective sur le big data et discute l'impact, les opportunités et les défis liés à son utilisation croissante dans la société.

Le PNR 75 a reçu 25 millions de francs suisses qui ont permis de financer un portefeuille de projets de recherche répondant à des critères stricts d'excellence scientifique[9]. Ils ont été menés entre 2017 et 2021 et regroupent trois catégories:

— les innovations fondamentales dans les infrastructures informatiques nécessaires aux applications du big data,
— les projets de recherche orientée vers l'application développant des applications concrètes
— les études sur les interactions entre le big data et la société, y compris les aspects juridiques, éthiques et sociétaux.

Quatre objectifs ont été fixés dans l'appel à propositions du PNR 75:

— réaliser des avancées dans le domaine de l'informatique et des technologies de l'information;
— relever les défis sociétaux, économiques, réglementaires (tant locaux que globaux) et éducatifs;
— développer les champs d'application;
— renforcer les capacités de recherche.

Le programme a fait des avancées scientifiques qui contribuent à des infrastructures de big data plus efficaces, a développé des applications concrètes dans plusieurs domaines, a fourni des pistes pour aborder les aspects sociétaux et a renforcé le potentiel de recherche et d'innovation en matière de big data en Suisse.

Contributions majeures du PNR 75

Le PNR 75 a été conçu en 2014, alors que de nombreuses technologies et questions liées au big data qui sont bien connues aujourd'hui ne faisaient qu'émerger. Leur développement et déploiement dans la société très rapides ont constitué un défi pour les projets de recherche, qui ont dû montrer de la flexibilité pour adapter leurs objectifs. Le cadre du programme ainsi que les projets financés ont couvert des questions essentielles tout au long de la chaîne de valeur du big data (voir «La chaîne de valeur du big data», p. 16).

La décision prise il y a huit ans d'inclure les défis sociétaux dans le programme s'est avérée justifiée, comme le montrent les nombreuses discussions actuelles sur l'équité et les biais dans l'intelligence artificielle, sur la souveraineté des données, ou encore sur l'impact des nouvelles applications sur la population et le personnel.

Le financement de projets visant à développer des applications concrètes a créé de nouvelles collaborations entre les disciplines qui ont réuni des spécialistes des domaines concernés ainsi qu'en informatique et des partenaires des secteurs public et privé. Cela a permis de renforcer le savoir-faire interdisciplinaire nécessaire au développement d'applications du big data et de créer une expérience précieuse pour la décennie à venir, à la fois en matière de big data et de science des données. Ces collaborations devraient jouer un rôle de plus en plus important dans la résolution des problèmes de société globaux, définis par exemple dans les Objectifs de développement durable des Nations unies, et concernant le changement climatique, la crise environnementale ou encore le vieillissement des populations.
Les projets de recherche sur les défis technologiques pour développer les infrastructures du big data ont renforcé l'expertise disponible en Suisse pour façonner ces technologies.

---

[9] Voir l'Annexe «Le Programme national de recherche ‹Big Data› (PNR 75)», p. 92.

L'impact sociétal du PNR 75

Les activités de sensibilisation du PNR 75 ont permis de familiariser un public plus large au sujet de l'impact du big data sur la société. Elles ont abordé les aspects juridiques et éthiques, notamment les questions d'égalité, et fourni aux écoles du matériel pédagogique sur ce thème. Dans l'ensemble, le PNR 75 a contribué à ce que la Suisse soit en mesure de bénéficier du big data de manière responsable.

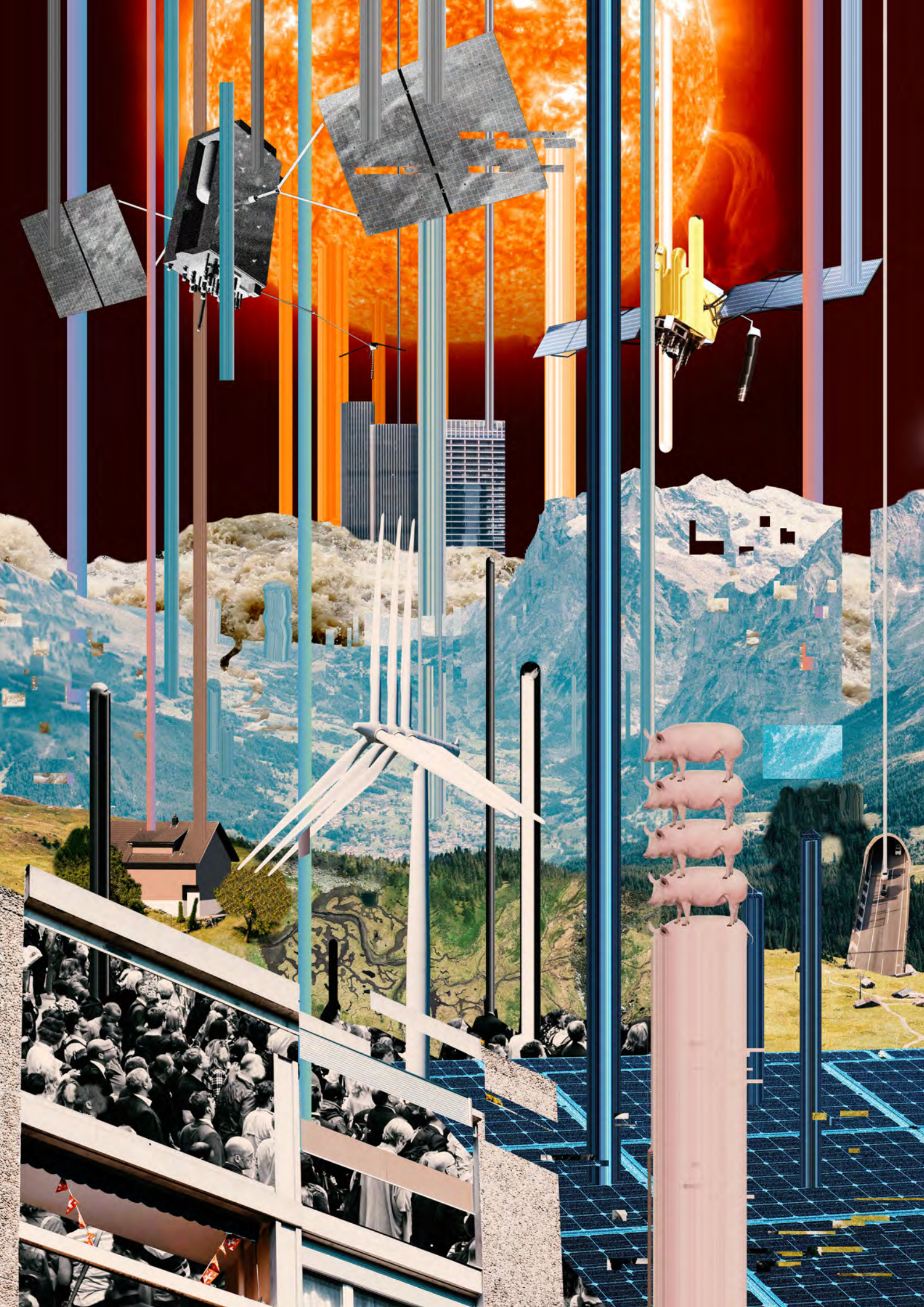
Certains projets ont apporté des informations utiles à l'élaboration de mesures politiques au-delà du thème du big data en analysant des données du monde réel dans les domaines des énergies renouvelables, de la gestion de l'environnement et de la socio-économie (voir «Résultats au-delà du big data», p. 33).

Les défis du big data ne peuvent pas être résolus une fois pour toutes. Les relever exige de maintenir des capacités de pointe en matière de recherche, d'éducation et d'innovation. Le PNR 75 a attiré l'attention des scientifiques ainsi que des acteurs privés et publics sur les enjeux du big data en finançant la recherche au moyen d'un appel ouvert et compétitif avec des normes de qualité élevées. Le programme a contribué à fournir une base solide à la recherche suisse sur le big data et à renforcer sa pertinence et son impact. Ce dernier point bénéficiera aussi du Programme national de recherche «Transformation numérique» (PNR 77) lancé trois ans après le PNR 75 et qui étudie les effets de la numérisation sur l'éducation, le marché du travail, la gouvernance et la confiance.

1.4 La structure du résumé

Le résumé du PNR 75 condense les principaux résultats des projets de recherche et présente une perspective sur les défis du big data.

Le chapitre 2 «Applications du big data» détaille les opportunités offertes par l'analyse de grands jeux de données dans des applications concrètes et explorées par les projets du PNR 75 dans les domaines de la santé, de durabilité et de la socio-économie. Le chapitre 3 «Technologies du big data» résume les progrès réalisés sur les défis techniques posés par le big data, notamment pour des infrastructures de calcul plus efficaces et pour de nouvelles approches de l'analyse des données. Le chapitre 4 «Aspects sociétaux, juridiques et éthiques du big data» offre de nouvelles perspectives sur les défis sociétaux du big data, en particulier liés à la propriété des données et à leur confidentialité ainsi qu'à l'équité, et présente des directives concrètes élaborées par les projets du PNR 75. Le chapitre 5 «Réflexions et perspectives» fournit un point de vue plus général sur les opportunités et risques liés au big data susceptibles de gagner en importance dans les années à venir. Les conclusions du Comité de direction du PNR 75 sont présentées au chapitre 6. L'Annexe fournit les informations principales sur le PNR 75. Ce résumé présente les principaux aspects du big data tout au long de la chaîne de valeur allant des données brutes aux applications concrètes, comme le montre l'infographie «La chaîne de valeur du big data».



2.
Applications du big data

Les applications du big data offrent des possibilités dans de nombreux domaines. Leur développement nécessite néanmoins un travail important: établir des partenariats avec les parties prenantes, assurer l'accès technique et légal aux données, développer des modèles analytiques et valider les applications auprès des usagers et usagères. Ce chapitre présente la douzaine d'applications développées dans le cadre du PNR 75, dans les domaines de la santé, durabilité, élaboration de politiques et recherche scientifique, et résume les enseignements tirés.

L'utilisation toujours plus grande des technologies numériques génère des volumes croissants de données. De nombreux acteurs souhaiteraient exploiter cette ressource pour créer de nouveaux outils ou rendre les instruments existants plus précis et plus efficaces. Cela entraîne un nombre croissant d'applications du big data dans de nombreux domaines.

Malgré l'engouement pour le big data, la création d'applications reste une tâche complexe, longue et parsemée de nombreux défis. Leur développement nécessite de réunir des partenaires, de résoudre les problèmes juridiques et éthiques liés à la vie privée et à l'équité, d'accéder aux données, de les analyser, de développer des modèles, de programmer les applications avant d'évaluer leur précision et leur utilité. Les dernières étapes consistent à valider les applications auprès des usagères et usagers, à les intégrer dans les flux de travail existants et à assurer une maintenance appropriée.

Le Programme national de recherche «Big Data» (PNR 75) a développé des applications basées sur le big data dans des domaines tels que la santé, la durabilité, la socio-économie et la recherche scientifique. Ils ne représentent qu'un petit sous-ensemble des secteurs qui explorent ou intègrent le big data. Ils vont des domaines avec une numérisation avancée, tels que la banque, le marketing et la santé, à des secteurs commençant à intégrer les données, comme l'agriculture, le journalisme ou le gouvernement.

Les projets financés par le programme ont amélioré les méthodologies existantes et en ont développé de nouvelles pour des applications dans certains domaines. Ils ont mis en évidence les avantages potentiels du big data pour la société et l'économie, comme pour la médecine personnalisée, la planification des transports, le déploiement intégré des énergies renouvelables, ou l'évaluation de l'effet des politiques socio-économiques.

Les diverses applications du PNR 75 ont atteint différents stades de développement, depuis de premiers modèles et prototypes jusqu'à des systèmes avancés. Cette diversité fait écho aux opportunités et aux défis liés à la création d'applications pratiques du big data.

Les applications développées dans le cadre du PNR 75 et les messages clés issus de ces projets sont regroupés dans quatre domaines:
— la santé dans la section 2.1;
— la durabilité – y compris le transport, l'énergie, l'approvisionnement alimentaire et la gestion environnementale – dans la section 2.2;
— les questions socio-économiques dans la section 2.3;
— et la recherche dans la section 2.4.

La section 2.5 présente un résumé des perspectives gagnées par le PNR 75 sur la création d'applications du big data. Les projets de recherche sont décrits dans la section 2.6.

2.1 Améliorer et personnaliser les soins de santé

De nombreuses approches veulent adapter les soins aux caractéristiques et aux besoins spécifiques des individus et des groupes de population, selon le paradigme «P4» de la médecine

prédictive, préventive, personnalisée et participative. Il repose sur un meilleur accès aux données, le développement d'outils d'analyse robustes et la mise en place de collaborations étroites avec le personnel de santé et la patientèle afin d'adapter les nouveaux outils numériques à la pratique. Les données de santé proviennent de bases de données traditionnelles telles que les dossiers électroniques des patient·es (DEP) et de nouvelles sources telles que les smartphones et les capteurs portables. Un dossier médical étant considéré comme sensible, les applications du big data doivent respecter des lois strictes en matière de protection des données. Elles varient d'une juridiction à l'autre, posant un défi aux collaborations interinstitutionnelles, comme discuté au chapitre 4. Les applications du big data ont un impact important sur la recherche, l'éducation et les soins de santé dans les institutions médicales ainsi qu'à domicile.

Les projets de recherche du PNR 75 sur les soins de santé (voir leur description dans la section 2.6) ne constituent qu'un petit sous-ensemble des applications possibles. Ils fournissent néanmoins des exemples concrets de la manière dont les soins de santé peuvent bénéficier du big data, comme:

— une gestion plus efficace des situations critiques dans les unités de soins intensifs, en surveillant les personnes soignées et en anticipant l'évolution de leur état (voir le projet de recherche du PNR 75 *Soins intensifs*),
— la personnalisation de la gestion des douleurs dorsales grâce à une appli pour smartphone (voir le projet *Douleurs dorsales*),
— de nouvelles techniques pour la recherche biomédicale (voir les projets *Genetic big data*, C*omparaison de génomes* et *Bases de données bioinformatiques*, aussi discutés dans la section 2.4).

Messages clés sur les applications dans le domaine de la santé

Le PNR 75 démontre le potentiel des applications du big data pour soutenir des soins de santé modernes et efficaces, mais révèle également de nombreux défis. En particulier, le développement d'applications utiles et pratiques nécessite une infrastructure solide pour le partage des données, y compris des solutions appropriées pour les questions techniques, organisationnelles, et juridiques.

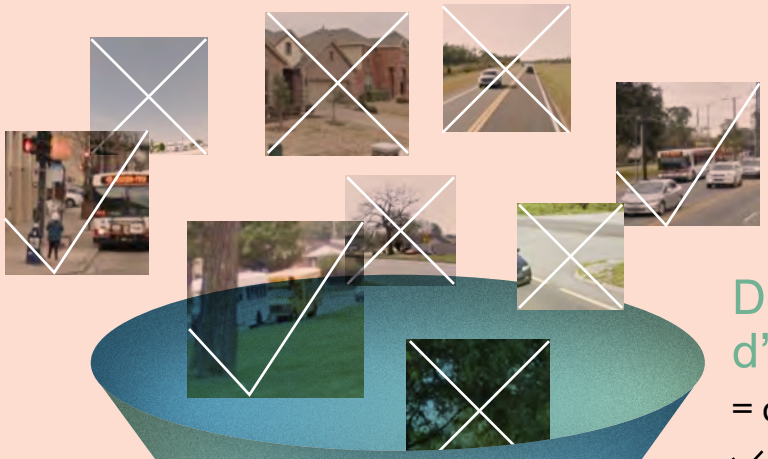
Impact sociétal

Automatiser l'analyse du monitoring des personnes hospitalisées peut améliorer la qualité des soins avec des alertes ayant un faible taux de fausses alarmes *(Soins intensifs).* Les applications mobiles peuvent soutenir les exercices de physiothérapie faits à domicile et recueillir des données pour évaluer l'impact de la thérapie *(Douleurs dorsales).* La recherche biomédicale peut bénéficier du big data (voir aussi section 2.4): plusieurs avancées méthodologiques et techniques facilitent l'utilisation de la génomique en recherche clinique, épidémiologie et biologie environnementale *(Genetic big data, Comparaison de génomes)* ainsi que l'exploitation des bases de données biomédicales par un plus large éventail de spécialistes *(Bases de données bioinformatiques).*

Nouvelles compétences en big data dans le secteur de la santé

Une application a pu intégrer des données multimodales (vidéo, données de capteurs, dossiers médicaux) dans un algorithme d'apprentissage automatique *(Soins intensifs).* Une autre a combiné des informations implicites et

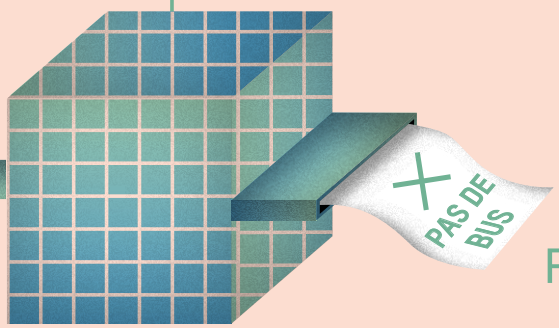
Comment fonctionne l'apprentissage automatique supervisé
Un algorithme apprend à résoudre une tâche (ici : reconnaître un bus) à partir de données labellisées. Plus la base de données d'apprentissage est grande, plus l'application sera précise.



Données d'apprentissage
= données labellisées
✓ un bus ✗ pas de bus

Algorithme d'apprentissage automatique
trouver des corrélations dans les données d'apprentissage

Données concrètes



Modèle

Résultats

explicites, montrant que les données recueillies par des capteurs peuvent être utilisées pour évaluer la véracité d'observations telles que l'auto-évaluation d'exercices physiques faits à domicile *(Douleurs dorsales).*

Défis

Il est nécessaire d'élaborer de nouvelles méthodes pour la collecte et la gestion des données de santé, généralement hébergées dans des silos de données. Mettre en place un système de santé national fondé sur des données transparentes et largement partagées faciliterait l'émergence de nouvelles solutions numériques en matière de soins *(Douleurs dorsales).* Le regroupement de diverses sources de données est entravé par la grande hétérogénéité des systèmes de santé et des réglementations en Suisse, et par le fait que les données recueillies par les dispositifs médicaux sont habituellement de nature propriétaire (en raison du manque d'incitations commerciales et réglementaires à l'intégration des données) *(Soins intensifs).* La collecte prospective de données de santé correctement labellisées et la gestion des processus réglementaires pour le traitement des données demandent beaucoup de temps et d'efforts.

Les applications de santé mobiles proposant des exercices physiques personnalisés pour la rééducation ou la gestion de la douleur peuvent aider les gens à respecter les protocoles de traitement, mais elles dépendent toujours de leur motivation *(Douleurs dorsales).* La numérisation des données de santé peut également susciter la méfiance du public à l'égard du gouvernement et des entreprises privées du secteur, comme en témoignent les critiques formulées à l'encontre de l'appli de traçage des contacts SwissCovid.

Pistes de solutions

Des mesures peuvent être prises pour améliorer le développement des applications de santé et favoriser l'utilisation du big data dans les soins.

— Généraliser la collecte de données pour certaines maladies dans tous les centres médicaux de Suisse, la configuration traditionnelle de centres d'étude individuels liés à des hôpitaux ou à des cantons spécifiques n'étant pas adaptée à l'utilisation du big data en santé. C'est un point important pour générer suffisamment de données et fournir des informations sur des maladies complexes. Il convient de prêter attention aux progrès réalisés par les entreprises technologiques multinationales dans la collecte, le stockage et l'analyse des données de santé privées et publiques.
— Rationaliser les processus nécessaires à la conformité juridique, comme l'utilisation de signatures numériques ou l'obtention du consentement des patientes et des patients pour l'utilisation de leurs données dans la recherche médicale. Veiller à ce que les organismes de réglementation tels que les commissions d'éthique connaissent bien les derniers développements en matière d'applications du big data *(Douleurs dorsales);* voir également le chapitre 4.
— Harmoniser les formats de données utilisés par les dispositifs médicaux et promouvoir une plus grande utilisation des formats standard dans les dossiers médicaux électroniques.
— Expliquer à toutes les parties prenantes qu'il est important de maintenir des métadonnées de haute qualité, un aspect crucial pour indexer et rechercher efficacement des grandes bases de données *(Soins intensifs).*

2.2 Soutenir la durabilité

Le développement d'une société durable demande d'optimiser les interactions entre les nombreuses composantes des systèmes d'énergie, de transport, d'approvisionnement ou d'alimentation. Par exemple, les modèles de mobilité dépendent de l'offre et de la demande de transport, avec une population ayant des besoins et des préférences variables, ainsi que d'interactions complexes déterminées par des facteurs tels que la météo, l'acceptation du travail à domicile et l'existence d'incitations financières.

Les applications du big data peuvent considérablement améliorer l'efficacité et la durabilité de ces systèmes. Il s'agit par exemple d'analyser des données détaillées en temps réel telles que le déplacement en transports publics ou la performance de panneaux solaires. Ces informations permettent d'optimiser en permanence l'offre pour répondre à la demande prévue, par exemple, en adaptant les horaires des transports ou en faisant varier la production d'électricité.

Le PNR 75 a démontré le potentiel d'une demi-douzaine d'applications innovantes pour soutenir la durabilité en:

— produisant des cartes à haute résolution à l'échelle régionale et nationale montrant comment l'énergie solaire, éolienne et géothermique pourrait profiter aux bâtiments individuels *(Potentiel des énergies renouvelables);*
— créant des modèles numériques de villes en 3D et en haute résolution *(Modèles numériques urbains)* et étudiant l'optimisation des transports collectifs et privés grâce à des traces GPS *(Gestion des transports);*
— produisant des outils pour quantifier automatiquement l'étendue des terrains érodés à l'aide d'images aériennes *(Érosion des sols)* et pour utiliser des vidéos mobiles ou de surveillance afin d'automatiser la détection des inondations *(Détection d'inondation);*
— développant une plateforme pilote pour collecter et analyser les données relatives à la filière porcine en Suisse *(Pig data).*

Messages clés sur les applications en durabilité

Le développement durable peut fortement bénéficier des applications du big data, notamment au vu de la longue expérience de l'utilisation des données en ingénierie ainsi que dans le monitoring de l'efficacité et de la sécurité. La numérisation, qui inclut le déploiement de capteurs pour l'Internet des objets, génère des jeux de données de plus en plus volumineux qui améliorent les perspectives d'application. Celles-ci passent par l'intégration de données provenant de sources variées et le développement de nouveaux outils analytiques puissants. Le besoin urgent d'une durabilité plus étendue dans notre société appelle à davantage exploiter le big data.

Impact sociétal

Les applications développées dans le cadre du PNR 75 montrent que l'utilisation du big data peut contribuer à soutenir la durabilité. Il peut notamment contribuer à évaluer et adapter les stratégies nationales et locales en matière d'énergies renouvelables, à créer des scénarios réalistes pour la gestion des transports, à produire des cartes pour la planification urbaine, à

surveiller la dégradation de l'environnement et à soutenir la gestion des risques naturels.

Nouvelles compétences en big data dans le secteur de la durabilité

Le PNR 75 a montré que de nouvelles techniques peuvent compléter des jeux de données et produire des modèles de production d'énergie à haute résolution avec des incertitudes quantifiées *(Potentiel des énergies renouvelables).* Des algorithmes peuvent produire des modèles numériques 3D à haute résolution de zones urbaines à l'aide de caméras à balayage omnidirectionnel montées sur des véhicules *(Modèles numériques urbains).* Il est possible d'obtenir un aperçu sur des traces de mobilité individuelles à partir de bases de données globales, agrégées et anonymes *(Gestion des transports).* L'apprentissage automatique supervisé peut identifier les principaux facteurs de l'érosion des sols et des glissements de terrain *(Érosion des sols).* On peut concevoir des interfaces efficaces pour visualiser des informations complexes et interagir avec elles à différentes échelles si l'on intègre les usagers et usagères de manière précoce *(Détection d'inondation).*

Défis

Les applications conçues pour la durabilité dépendent de la disponibilité de données complètes et de haute qualité, comme dans d'autres domaines. Les données environnementales nécessaires au déploiement de dispositifs de production d'énergie renouvelable sont souvent absentes ou incomplètes *(Potentiel des énergies renouvelables).* L'accès aux données de déplacement des gens nécessite le soutien de l'industrie, notamment des opérateurs de communication, et peut être entravé par les réglementations sur la protection de la vie privée. Des moyens permettent de limiter le risque de fuite des données, par exemple en ne donnant accès qu'à un back-end de stockage sécurisé *(Gestion des transports).*

Les algorithmes d'apprentissage automatique supervisé entraînés par des données labellisées ne fonctionnent pas toujours de manière fiable une fois utilisés sur le terrain. Des approches plus avancées peuvent être bénéfiques, comme l'apprentissage non supervisé, qui n'utilise pas de données d'apprentissage. Des métadonnées importantes, telles que la localisation GPS des vidéos, sont souvent absentes des données produites par crowdsourcing, ce qui rend les analyses difficiles et chronophages *(Détection d'inondation).* Les variations naturelles liées à la photographie aérienne (ombre, champs fauchés, etc.), compliquent l'analyse de l'évolution temporelle des terrains *(Érosion des sols).*

Il est difficile de développer des applications du big data dans les secteurs où l'utilisation des données reste fragmentée, et dans lesquels les analyses sont entravées par la faible qualité des données, leur portée limitée, ou l'absence d'ontologies cohérentes (la manière dont les informations sont conceptualisées, catégorisées et représentées). C'est également le cas lorsque la confidentialité est importante et que les données sont partagées avec réticence *(Pig data).* Enfin, les changements de personnel dans les organisations partenaires peuvent mettre un terme soudain à une collaboration prévue avec l'équipe de recherche.

Pistes de solutions

— Développer des approches de type «privacy-by-design» qui intègrent des solutions aux problèmes de protection de la vie privée dès le début

Faire connaître le big data dans les écoles

L'utilisation des technologies de l'information fait partie des plans d'études scolaires suisses («Éducation numérique» dans les cantons romands), et l'acquisition de compétences en matière de données a besoin de nouveaux supports pédagogiques. En collaboration avec le Musée de la communication à Berne, le PNR 75 a produit un matériel pédagogique sur le big data pour les niveaux secondaires inférieur et supérieur, comprenant huit modules de deux leçons chacun. Lancé au début 2020, le matériel a été téléchargé plus de 12 000 fois. L'initiative a été promue par le Musée de la communication et l'Académie suisse des sciences techniques (SATW).

Big data: outil pédagogique pour les cycles secondaires
(PNR 75 et Musée de la communication, Berne)

du projet et impliquent très tôt des spécialistes des aspects juridiques et sociétaux *(Optimisation des transports, Modèles numériques urbains);* voir aussi le chapitre 4.
- Élaborer des mesures afin de faciliter le partage de données sensibles entre l'industrie et le monde universitaire à des fins de recherche, tout en préservant la confidentialité et la vie privée *(Pig data).*
- Encourager une culture de la numérisation, sensibiliser aux problèmes et défis liés à l'utilisation des données et soutenir la formation continue dans ce domaine. Inclure dans les projets de recherche des gens ayant de l'expérience dans le domaine des sciences techniques et appliquées afin de combler le fossé entre le monde universitaire et l'industrie, en particulier dans les domaines dans lesquels la culture des données est encore émergente *(Pig data).* Mettre en place des plateformes gouvernementales pour le partage de données crowdsourcées présentant un intérêt pour le public en veillant à ce que les métadonnées nécessaires soient incluses *(Détection d'inondation).*
- Impliquer les usagers et usagères dès le début pour comprendre leurs besoins et développer des outils pratiques pour interpréter et visualiser les données *(Détection d'inondation).*
- Investir dans l'apprentissage non supervisé et dans d'autres approches réduisant la dépendance envers des données d'apprentissage qui sont potentiellement biaisées *(Détection d'inondation).* En contrepartie, développer l'apprentissage automatique supervisé lorsque des données d'apprentissage adéquates sont disponibles *(Érosion des sols).*

2.3 Analyser les interactions socio-économiques

Des approches d'analyse sophistiquées sont de plus en plus utilisées dans les domaines sociaux et politiques. Les données nécessaires ne sont néanmoins pas toujours disponibles et l'analyse s'avère être plus complexe que dans les sciences naturelles et techniques, les causes et les effets étant difficiles à démêler en raison de leurs multiples facettes et de leurs différents contextes socio-politiques. Le fait que davantage de données soient collectées offre néanmoins de nombreuses opportunités pour l'évaluation et l'amélioration de mesures politiques.

Des applications concrètes en socio-économie ont été développées dans le cadre du PNR 75 pour:

— améliorer des méthodes pour dévoiler des relations causales dans des jeux de données socio-économiques et soutenir ainsi l'élaboration des politiques *(Politique fondée sur les faits);*
— développer des méthodes analytiques pour étudier la production et les citations de brevets afin de cartographier l'innovation et la production de connaissances internationales *(Cartographie de l'innovation).*

Messages clés sur les applications en socio-économie

Impact sociétal

La disponibilité croissante de données socio-économiques offre un grand potentiel pour une prise de décisions et une élaboration de politiques davantage fondées sur les faits, mais les analyses sont souvent plus difficiles que prévu. Le projet «Politique basée sur les faits» a développé des techniques basées sur l'apprentissage automatique pour démêler les effets causaux de simples corrélations. Le big data permet également de cartographier la diffusion des idées et des influences à l'échelle mondiale, ce qui devrait aider à mieux comprendre les facteurs soutenant l'innovation *(Cartographie de l'innovation).*

Impact sur les capacités de big data en socio-économie

L'utilisation d'algorithmes pour analyser des données sur des thèmes politiques peut contribuer à réduire les biais, les scientifiques pouvant préférer inconsciemment les résultats correspondant à leurs opinions. Elle permet de déterminer si une mesure politique affecte différemment certaines sous-populations que la moyenne, et ainsi de générer des informations plus fines sur son efficacité. Ces techniques sont désormais prêtes à être utilisées concrètement *(Politique fondée sur les faits).* De nouvelles méthodes peuvent analyser de grands réseaux en incluant des effets temporels et identifier les nœuds ayant des impacts rapides ou durables *(Cartographie de l'innovation).*

Défis

Même une analyse de données minutieuse peut être biaisée par les méthodes utilisées. Des relations peuvent être cachées par des données trop granulaires *(Politique fondée sur les faits).* Les jeux de données ont des limites intrinsèques, comme le fait que les citations de brevets ne peuvent pas être suivies à travers différents registres nationaux *(Cartographie de l'innovation).*

Pistes de solutions

— Promouvoir dans l'évaluation de mesures politiques l'utilisation d'outils analytiques avancés utilisant des données socio-économiques *(Politique fondée sur les faits).*
— Encourager l'agrégation des données à différentes granularités, par exemple au niveau de la population et des sous-populations, et promouvoir l'analyse causale pour déduire des relations solides à partir des données *(Politique fondée sur les faits).*
— Soutenir la compatibilité des jeux de données de différents pays *(Cartographie de l'innovation).*

2.4 Accélérer la recherche

Le big data soutient également l'innovation par son rôle croissant dans la recherche fondamentale. Ce fait est visible notamment dans les très grandes collaborations internationales telles que le CERN à Genève, des projets d'astronomie, d'observation par satellite ou de génomique. Un nombre croissant de disciplines scientifiques ont établi des normes pour la production et le partage des données. Cependant, la taille des jeux de données (jusqu'à plusieurs pétaoctets) complique l'accès, le stockage, la manipulation et l'analyse des données, et exige de disposer de méthodes analytiques et d'algorithmes d'apprentissage automatique efficaces.

Le PNR 75 a produit de nouvelles méthodes pour le traitement et l'analyse de très grandes bases de données en chimie, physique du soleil et génétique. Elles permettent de:

— accélérer la découverte de nouvelles molécules grâce à une simulation efficace de leurs propriétés *(Chimie computationnelle);*
— améliorer la compréhension et la prévision des éruptions solaires, à la source de tempêtes géomagnétiques sur Terre *(Éruptions solaires);*
— faciliter l'analyse de données génomiques et biologiques dans la recherche biomédicale, et favoriser le développement de nouvelles approches thérapeutiques et diagnostiques *(Genetic big data, Comparaison de génomes, Bases de données bioinformatiques).*

Messages clés sur les applications en recherche

Impact sociétal

Des outils d'analyse efficaces, notamment basés sur l'apprentissage automatique, peuvent exploiter des jeux de données scientifiques de plus en plus volumineux et permettre de faire des découvertes plus rapidement et des prévisions plus précises en science de la vie, chimie et astrophysique.

Impact sur les capacités de big data dans la recherche

Atteindre une grande précision de prédiction avec de grands jeux de données est possible même après les avoir compressés pour réduire le temps et le coût du transfert, stockage et traitement *(Éruptions solaires, Genetic big data).* L'apprentissage automatique peut considérablement réduire le temps nécessaire pour résoudre avec précision des problèmes mathématiques complexes, tels que la détermination des propriétés des molécules *(Chimie computationnelle).* Plusieurs projets du PNR 75 ont développé des outils pour gérer et explorer des jeux de données biologiques massifs à l'échelle du pétaoctet (million de gigaoctets) *(Genetic big data, Bases de données bioinformatiques).*

Défis

Les projets de recherche ne disposent souvent pas de données labellisées, limitant le potentiel des outils standard d'apprentissage automatique supervisé *(Éruptions solaires).* Les métadonnées (les informations sur les données) sont souvent de mauvaise qualité, ce qui entrave leur indexation et analyse *(Genetic big data).*

Pistes de solutions

— Encourager les milieux universitaires et industriels à partager et à conserver les grands jeux de données de

recherche dans le cadre du mouvement des données de recherche ouvertes ou «Open Research Data» (voir «Les défis des données de recherche ouvertes», p. 47).
— Evaluer les méthodes développées dès le début avec des données réalistes pour être en mesure d'adapter, si nécessaire, les objectifs du projet de recherche.
— Investir dans des projets de recherche fondamentale et appliquée de très haut niveau utilisant de grands jeux de données afin de favoriser les compétences en Suisse en matière de big data.

2.5 Messages clés sur les applications du big data

Attention à la hype

Les médias, l'industrie et les think-tanks décrivent parfois le big data comme une baguette magique: il suffit de trouver les données, d'ajouter un peu d'apprentissage automatique, d'entraîner les algorithmes et de construire une application pour être en mesurer de bouleverser les pratiques professionnelles et des secteurs économiques entiers. Cette vision trop simpliste cache de nombreux obstacles conceptuels, techniques, juridiques, collaboratifs et organisationnels.

La mise en place d'une application big data exige des efforts considérables: créer des partenariats avec les parties prenantes, identifier les sources de données, évaluer leur qualité, les stocker de manière sécurisée, les préparer pour l'analyse, trouver les algorithmes appropriés et les adapter ou en créer de nouveaux, évaluer leur performance et fiabilité, créer des interfaces pour utiliser les résultats et, finalement, intégrer la nouvelle application dans les procédures existantes.

La célébration excessive du big data obscurcit des questions très profondes, même si elles semblent évidentes. Les données existent-elles vraiment? Sont-elles accessibles? Sont-elles correctement décrites par des métadonnées? La vie privée peut-elle être préservée et les réglementations respectées? Les utilisateurs finaux ont-ils réellement besoin de l'application prévue? Ces questions doivent être envisagées dès le départ afin d'évaluer de manière réaliste le travail à accomplir.

Votre domaine est-il prêt pour le big data?

Les applications du big data peuvent être développées de manière relativement linéaire lorsqu'elles s'appuient sur des prototypes existants (Potentiel des énergies renouvelables, Gestion des transports) ou lorsque des données de haute qualité et standardisées sont disponibles comme en météorologie ou biologie *(Éruptions solaires, Genetic big data).* Les efforts peuvent alors se concentrer principalement sur les questions techniques, telles qu'établir un pipeline pour accéder aux données en temps réel, concevoir les algorithmes ou créer des interfaces interactives conviviales. Les projets du PNR 75 ont fait des avancées importantes dans la conception, la mise en œuvre et l'évaluation d'approches pratiques de l'ingénierie des données,

notamment pour la gestion, l'analyse, la visualisation, l'évaluation, l'audit, l'intégration et l'exploration des données.

À l'inverse, il est bien plus difficile de créer des applications du big data dans des domaines moins numérisés, manquant une culture des données et peu enclins à les partager *(Pig data, Cartographie de l'innovation).* Dans ce cas, il faut consacrer beaucoup d'efforts à des questions non techniques, comme la mise en place de partenariats entre des parties prenantes potentiellement réticentes. La disponibilité et la qualité des données doivent être évaluées dès le début, ce qui peut conduire à réviser les objectifs de l'application.

L'interdisciplinarité est nécessaire

La gestion de jeux de données à l'échelle du pétaoctet exige beaucoup de temps, de main-d'œuvre et d'efforts collaboratifs afin de résoudre de nombreux problèmes techniques ou juridiques. Créer une application qui a un impact nécessite le plus souvent une approche interdisciplinaire pour aborder avec les parties prenantes tous les problèmes potentiels dès le début, y compris la manière dont la solution imaginée sera utilisée concrètement *(Pig data, Détection d'inondation).* Les scientifiques, les spécialistes du domaine et les partenaires privés ou publics doivent interagir fréquemment pour garantir l'acquisition et le partage de données de haute qualité et s'assurer que les applications répondent aux besoins concrets. Les scientifiques du PNR 75 ont exploré de nouvelles façons d'interagir efficacement avec les différentes parties prenantes. Faire participer les usagères et usagers dès le début améliore la conception des applications *(Détection d'inondation).* Ces expériences contribuent à consolider une recherche universitaire à même de développer rapidement des applications concrètes lorsque nécessaire.

Le monde réel est plus complexe que les données d'entraînement

Les algorithmes d'apprentissage automatique peuvent échouer sur le terrain, avec des conséquences potentiellement graves lorsqu'il s'agit de santé ou de véhicules autonomes. C'est un problème qui touche l'apprentissage supervisé lorsque les données d'apprentissage sont incomplètes, imprécises ou trop homogènes. Le recours à un apprentissage non supervisé plus frugal pourrait déboucher sur des systèmes plus robustes *(Détection d'inondation).*

Penser au «privacy-by-design» dès le départ

Les questions liées au respect de la vie privée et à la réglementation doivent être traitées avec soin, notamment avec l'aide de spécialistes juridiques (voir chapitre 4). Les approches de type «privacy-by-design» doivent être envisagées et mises en œuvre le plus tôt possible. Il convient alors d'examiner attentivement les principes généraux tels que la limitation des objectifs, la transparence et la proportionnalité, ainsi que la minimisation, l'exactitude et la sécurité des données *(Soins intensifs, Modèles numériques urbains).* Le partage d'expériences, au sein d'un domaine ou entre plusieurs secteurs, favorise l'établissement de bonnes pratiques.

Résultats au-delà du big data

Plusieurs projets de recherche du PNR 75 ont produit des résultats ayant une pertinence sociétale directe, notamment dans les domaines de la durabilité et de la socio-économie. Ils illustrent le potentiel du big data à soutenir les analyses et décisions politiques.

Le projet *Potentiel des énergies renouvelables* a généré des connaissances concrètes pouvant soutenir **la stratégie énergétique 2050 du Conseil Fédéral.**

→ Il est possible d'atteindre la moitié de la production nationale d'électricité photovoltaïque potentielle avec seulement un dixième des toits existants en se concentrant sur ceux ayant le plus grand potentiel. Cela correspond à 12 TWh d'électricité par an, soit environ 20% de la consommation nationale.

→ Un millier d'éoliennes pourraient produire 4 TWh, l'objectif fixé par l'Office fédéral de l'énergie pour l'énergie éolienne en 2050.

→ Les cartes en haute résolution du potentiel géothermique à faible profondeur dans les cantons de Vaud et de Genève indiquent une production potentielle de 4 TWh de chaleur, soit 40% de la demande des deux cantons. L'utilisation de réseaux de chauffage urbain pour la distribuer entre les localités permettrait de doubler ce potentiel.

→ La réinjection de la chaleur dans le sol en été est importante pour une utilisation durable de l'énergie géothermique. En combinaison avec les réseaux de chauffage urbain, la géothermie de faible profondeur pourrait couvrir plus de 70% de la demande de chauffage et de refroidissement des bâtiments suisses d'ici à 2050.

Le projet *Érosion des sols* a permis de mieux comprendre **l'érosion dans les régions alpines.**

→ De 2007 à 2016, une augmentation de 80% de la surface touchée par l'érosion a été observée dans une région de 2 000 km² (11% des Alpes suisses) autour de Martigny (VS).

Le projet *Politique basée sur les faits* a établi des effets de causalité dans des **thèmes socio-économiques.**

→ Une formation à la recherche d'emploi ne permet pas, en moyenne, aux demandeurs d'emploi de trouver plus rapidement un travail. Elle peut néanmoins augmenter de 60% les chances de le faire pour certains sous-groupes, comme la population issue de l'immigration.

→ La pratique de la musique a un effet positif sur les développements cognitif et non cognitif des enfants.

→ Les arbitres de football sont plus susceptibles de pénaliser les équipes provenant de certaines zones linguistiques en Suisse.

→ Le projet *Cartographie de l'innovation* a produit des **éclairages sur l'innovation** en analysant des millions de brevets.

Les écosystèmes de brevets sont moins internationaux que prévu. Les citations de brevets sont souvent regroupées sur les plans géographiques et par disciplines, les plus interdisciplinaires ne semblant pas avoir davantage de succès. Les brevets de certains pays font l'objet d'un nombre particulièrement élevé de citations, comme la Suisse, connue comme un centre d'innovation prolifique.

→ Les entreprises et les organisations jouent des rôles différents dans la manière dont les brevets se citent les uns les autres, certaines d'entre elles agissant comme plaque tournante ou distributrice de connaissances.

Voir section 2.6 pour plus de détails

2.6 Les projets de recherche sur les applications du big data

Le PNR 75 a développé des applications basées sur le big data dans divers domaines: deux en santé, six en durabilité, deux dans le domaine socio-économique et cinq dans celui des capacités de recherche.

Projets de recherche en santé

Soins intensifs: un système d'alerte automatisé

Ce projet a créé une plateforme pour aider le personnel d'une unité de soins intensifs de neurochirurgie à réagir rapidement aux situations critiques, en particulier lors d'accidents vasculaires

cérébraux ischémiques et de crises d'épilepsie. Elle peut aider à organiser les interventions et à accroître la sécurité des patientes et des patients. Le projet a mis en place une collaboration entre la Neurocritical Care Unit de l'Hôpital universitaire de Zurich, l'ETH Zurich et IBM Research Zurich.
Le système prédit les situations critiques en intégrant et analysant de nombreux types de données: électroencéphalographie, flux vidéo et antécédents médicaux des patient·es, y compris l'imagerie cérébrale et les analyses de laboratoire. L'équipe a créé des technologies pour capturer des données biomédicales en temps réel à une haute résolution, jusqu'à une fréquence de 200 hertz. Les algorithmes développés peuvent détecter automatiquement les crises d'épilepsie à partir de vidéos et d'électroencéphalographies, et prédire les lésions cérébrales secondaires imminentes. L'application s'est basée sur les données de plus de 100 personnes souffrant d'hémorragie sous-arachnoïdienne, un type d'accident vasculaire cérébral. Deux algorithmes ont été développés pour réduire le taux de fausses alarmes, l'un utilisant l'apprentissage automatique et l'autre la surveillance vidéo du mouvement des patients. Le système a été intégré et testé dans un environnement clinique.

—

ICU-Cockpit: IT platform for multimodal patient monitoring and therapy support in intensive care and emergency medicine
Emanuela Keller (Hôpital universitaire de Zurich)

Douleurs dorsales: une solution personnalisée sur smartphone

Ce projet a développé un système basé sur smartphone pour la gestion des douleurs dorsales. L'équipe a créé l'application «Swiss Health Challenge» pour collecter, transmettre et stocker les données anonymes générées par des capteurs. Elle a évalué trois stratégies préventives, notamment afin de réduire les coûts des traitements (qui impliquent souvent des analgésiques, des séances de physiothérapie et des interventions chirurgicales). Les scientifiques de l'ETH Zurich ont mis au point des méthodes d'apprentissage automatique pour analyser les données dans le cadre d'une collaboration impliquant l'Hôpital universitaire Balgrist et l'entreprise suisse de dispositifs médicaux Hocoma.
Une approche préventive a régulièrement demandé aux personnes souffrant de lombalgie d'effectuer leur séance d'exercices à domicile. L'évaluation faite par des physiothérapeutes n'a pas révélé d'amélioration dans les cas non sévères. Cependant, l'étude a permis de mieux comprendre ce qui influence l'adhésion des patientes et patients aux séances d'exercices physiques et la façon dont la peur de faire un faux-mouvement peut affecter le balancement postural, à savoir les mouvements inconscients qui maintiennent l'équilibre.
L'application a été complétée par des capteurs de mouvement mesurant la charge physique pendant un entraînement de ski, susceptible d'entraîner des blessures au dos. Cette approche permet d'obtenir des informations bien plus détaillées sur le comportement en matière d'exercice que les auto-évaluations. Le projet a également mis en évidence l'importance de la motivation lors d'un programme d'exercices physiques.

—

Personalized management of low back pain with mHealth: big data opportunities, challenges and solutions
Robert Riener (ETH Zurich), Walter Karlen (Université d'Ulm)

Projets de recherche en durabilité

Potentiel des énergies renouvelables: estimations pour la Suisse

Ce projet a créé une plateforme numérique pour estimer le potentiel des énergies géothermique, éolienne et photovoltaïque pour le chauffage et la climatisation des bâtiments. Ces estimations nationales à hautes résolutions spatiale et temporelle peuvent aider à planifier les systèmes énergétiques aux niveaux local et régional, à optimiser les incitations et à adapter la stratégie énergétique nationale.

Le système intègre des données météo (vent et rayonnement solaire), environnementales (topographie, géologie et température du sol) et sur le milieu bâti (orientation des toits, espace disponible pour les forages). Il crée des cartes régionales et nationales du potentiel des énergies renouvelables, avec une résolution spatiale à l'échelle des bâtiments et une résolution temporelle d'environ une heure. De nouvelles techniques peuvent interpoler les points de mesure disponibles pour combler les lacunes. Par exemple, des cartes des vents à une échelle de 250 mètres ont été générées pour l'ensemble du pays à partir des données produites par 208 stations de surveillance de MétéoSuisse. Le projet a également quantifié l'incertitude des cartes générées et produit de nombreux résultats pertinents pour la politique énergétique (voir «Résultats au-delà du big data», p. 33).

—

Hybrid renewable energy potential for the built environment using big data: forecasting and uncertainty estimation
Jean-Louis Scartezzini (EPFL)

Gestion des transports: traces de mobilité individuelle anonymes

Ce projet a exploré les moyens de recueillir et d'analyser les données GPS de smartphones pour étudier la mobilité de la population. Plus de 4 000 personnes ont participé et installé une application, qui a enregistré plus d'un million de trajets. L'équipe de recherche a développé des méthodes pour rendre les données anonymes, identifier le type de transport utilisé (marche, vélo, bus, voiture, etc.) ainsi que l'objectif du déplacement (sport, travail, éducation, etc.). Les résultats du projet ont été utilisés dans d'autres grandes études sur la mobilité en Suisse qui examinent la véracité des auto-évaluations («Microrecensement mobilité et transports») ou la tarification de la mobilité («Comportement en matière de mobilité en Suisse»).

Ils ont également permis de générer et évaluer différents scénarios – modifier les horaires des transports publics, mettre en place une nouvelle gestion du trafic – sur une plateforme de simulation de la mobilité (Matsim). Ils montrent par exemple une très forte diminution de l'utilisation des transports publics à Zurich durant la pandémie de Covid-19 avec une augmentation significative des trajets à vélo.

Réalisé en collaboration avec l'opérateur Swisscom, le projet a souligné le potentiel des traces GSM pour la modélisation et gestion des transports. Des profils de mobilité individuelle anonymes peuvent être extraits de traces GSM agrégées, ce qui évite de devoir collecter activement des données auprès des individus ou de compromettre la vie privée.

—

Big data transport models: the example of road pricing
Kay W. Axhausen (ETH Zurich)

Modèles numériques urbains: scans 3D réalisés par un véhicule

Ce projet a conçu un algorithme et une caméra omnidirectionnelle pour produire un modèle numérique 3D complet d'une ville à partir de scans continus réalisés par la caméra montée sur

un véhicule. Un tel «jumeau numérique» d'une zone urbaine peut soutenir la planification urbaine et des transports.
Le projet a étudié les questions juridiques et éthiques, et développé une approche «privacy-by-design» basée sur les principes de limitation des objectifs, de transparence, de proportionnalité, de minimisation, d'exactitude et de sécurité des données. Il a par exemple automatiquement supprimé des détails sensibles tels que les plaques d'immatriculation des voitures ou les visages lors de tests menés à Sion (VS). L'équipe a discuté du transfert de technologie avec une start-up suisse offrant des services de numérisation urbaine.

—

ScanVan – a distributed 3d digitalization platform for cities
Frédéric Kaplan (EPFL)

Érosion des sols: quantification par photographie aérienne

Ce projet a créé des algorithmes d'apprentissage automatique pour identifier et cartographier de manière automatique les sols érodés à partir de photographies aériennes officielles. Il a étudié dix sites, principalement dans des régions montagneuses. Il a montré que les surfaces érodées dans la vallée d'Urseren entre Realp et Hospental (UR) ont presque triplé en 16 ans pour atteindre 0,4 km^2. Il a identifié les principaux facteurs influençant l'érosion et les glissements de terrain, à savoir la pente, la rugosité et l'orientation du terrain. L'analyse automatisée a été utilisée dans une étude détaillée d'une région de 2 000 km^2 (un dixième des Alpes suisses) autour de Martigny (VS). Elle a estimé que la surface dégradée avait augmenté de 80% entre 2007 et 2016 (voir Résultats au-delà du big data, p. 35).
Ces résultats peuvent rendre l'érosion visible auprès des instances politiques et soutenir les mesures de protection des sols en agriculture, tourisme et aménagement du territoire. Il s'agit d'un point important, le sol étant une ressource non renouvelable essentielle pour la production alimentaire, la biodiversité et la gestion des risques naturels. Ces approches seront poursuivies dans un projet financé par l'Office fédéral de l'environnement pour développer un outil capable de cartographier l'érosion à grande échelle.

—

WeObserve: integrating citizen observers and high throughput sensing devices for big data collection, integration, and analysis
Volker Roth (Université de Bâle)

Détection d'inondation: géolocalisation automatique de vidéos crowdsourcées

Ce projet a élaboré les premiers éléments d'une plateforme d'aide à la gestion des risques d'inondation. Il comprend l'analyse automatisée de vidéos pour reconnaître les situations de crise ainsi que leur localisation et la visualisation des résultats pour les personnes en charge de la gestion des catastrophes naturelles.
Le projet a souligné l'importance des algorithmes d'apprentissage automatique non supervisé. En effet, la technologie disponible de reconnaissance d'images, basée sur l'apprentissage supervisé entraîné avec des jeux de données annotées, peut échouer lors de l'analyse de vidéos prises sur le terrain.
L'équipe a testé dans une étude menée dans le canton de Bâle-Campagne des techniques pour localiser automatiquement des vidéos prises par exemple avec un smartphone sans contenir d'informations GPS. Elle a comparé les vidéos avec des images de rues disponibles. Cette stratégie repose sur des sources de qualité suffisantes, une météo adaptée

et la présence d'infrastructures ou de points de repère reconnaissables. En collaboration avec des spécialistes de gestion des risques naturels, les scientifiques ont testé des prototypes de cartes interactives visualisant les scènes d'inondation. Le projet a franchi plusieurs étapes importantes vers l'utilisation de l'intelligence artificielle pour la détection automatique des catastrophes naturelles.

—

EVAC – Employing video analytics for crisis management

Susanne Bleisch (Haute école du Nord-Ouest de la Suisse, FHNW)

Pig data: analyse de la filière porcine en Suisse

Ce projet a réuni les acteurs de la filière porcine suisse pour intégrer des informations sur des aspects tels que le transport des porcs, leur santé, les traitements, la qualité de la viande, les processus d'engraissement et la facturation. Il a développé des méthodes de prédiction en temps réel. Les analyses ont confirmé que la plupart des exploitations fournissent des carcasses répondant aux normes de qualité prescrites et ont identifié les aspects dans lesquels les spécifications de l'industrie ne sont pas respectées.

Le projet a apporté des réponses à 6 des 18 questions posées par les parties prenantes, telles que les facteurs influençant la qualité des carcasses, la structure du réseau de l'industrie porcine, ou encore la différence de qualité et de revenus entre les troupeaux d'une même race et les lots mixtes. Il a souligné l'importance de la numérisation de l'élevage pour les principales parties prenantes telles que l'Office fédéral de la sécurité alimentaire et des affaires vétérinaires, contribuant ainsi à la création d'un centre suisse de compétences et d'information sur la santé des porcs.

Le projet a révélé la difficulté à mener des analyses dans un domaine comptant de petites exploitations peu habituées à manier les données en comparaison avec d'autres secteurs économiques, ou à l'industrie porcine dans d'autres pays. Les défis spécifiques comprennent la qualité et la compatibilité des données ainsi qu'une réticence à partager les informations.

—

Pig data: health analytics for the Swiss swine industry

John Berezowski (Université de Berne)

Projets de recherche en socio-économie

Politique basée sur les faits: démontrer les causalités dans les données

Ce projet a créé de nouvelles techniques d'apprentissage automatique capables de démontrer des relations de cause à effet – au-delà de simples corrélations – dans de grands jeux de données socio-économiques. Il a mis en évidence des causalités dans des jeux de données réelles concernant les allocations de chômage, l'éducation et le sport (voir Résultats au-delà du big data, p. 33).

Ces nouvelles méthodologies permettent de découvrir des sous-groupes susceptibles de bénéficier d'une intervention même lorsque la population n'est, en moyenne, pas affectée. De tels résultats normalement cachés pourraient orienter la politique vers des mesures plus personnalisées.

Ce projet a fait des avancées importantes en matière de méthodologie, notamment en comparant et évaluant des approches statistiques existantes. Les jeux de données socio-économiques sont de plus en plus complexes, ce qui peut se traduire par une meilleure compréhension des phénomènes qu'elles capturent, mais aussi par des analyses plus difficiles. Le projet a

développé des approches d'apprentissage automatique pour répondre automatiquement à une question clé des analyses: quels facteurs doivent être analysés explicitement et lesquels doivent au contraire être considérés comme des facteurs confondants? Ces questions sont essentielles à l'élaboration de politiques fondées sur les faits et seront explorées dans la continuation de ce projet, financée par le Programme national de recherche «Transformation numérique» (PNR 77).

—

Causal analysis with big data
Michael Lechner (Université de Saint-Gall)

Cartographie de l'innovation: analyse des brevets

Ce projet a analysé des millions de brevets pour étudier comment l'innovation se répand dans le monde. Il a fusionné plusieurs bases de données contenant des informations sur les brevets et les entreprises, et créé un modèle pour saisir les changements et les effets temporels dans des réseaux complexes. Il a permis d'identifier les brevets influents et de mieux comprendre l'innovation mondiale (voir Résultats au-delà du big data, p. 33). L'équipe a adapté un modèle statistique aux réseaux de big data comptant plus d'un million de nœuds et l'a mis à disposition en open-source.

—

The global structure of knowledge networks: data, models and empirical results
Alessandro Lomi (Université de la Suisse italienne)

Projets de recherche sur les capacités de recherche

Chimie computationnelle: découverte de nouvelles molécules

Ce projet a amélioré les méthodes d'intelligence artificielle pour simuler les molécules, ce qui peut accélérer la découverte de composés utiles en santé ou pour l'industrie. Il a développé de nouvelles méthodes d'apprentissage automatique pour calculer les principales propriétés des molécules à partir de bases de données contenant des informations sur des centaines de milliards de molécules. S'il est très difficile d'analyser efficacement des sources aussi vastes, le projet a produit des jeux de données d'apprentissage et de test, et a pu caractériser les molécules bien plus rapidement et précisément qu'avec les méthodes existantes, même pour les composés de grande taille.

—

Big data for computational chemistry: unified machine learning and sparse grid combination technique for quantum based molecular design
Helmut Harbrecht (Université de Bâle)

Éruptions solaires: prédiction de tempêtes géomagnétiques

Ce projet a montré que les éruptions solaires peuvent être prédites à l'avance, aidant à anticiper d'éventuelles tempêtes géomagnétiques sur Terre. Celles-ci peuvent perturber des infrastructures critiques telles que les télécommunications, les réseaux électriques, les satellites ou encore le trafic aérien. L'équipe a utilisé 30 téraoctets d'observations solaires, développé des méthodes pour analyser des données compressées et montré que les signaux dans les longueurs d'onde ultraviolettes, presque inutilisés à l'heure actuelle, peuvent aider à prédire la formation d'une éruption solaire. L'algorithme a ainsi prévu une éruption solaire une demi-heure avant son observation. L'équipe a développé à la fois des algorithmes supervisés (avec des données d'apprentissage) et des techniques non supervisées afin de surmonter le problème la disponibilité limitée de données labellisées.

—

Machine learning based analytics for big data in astronomy
Svyatoslav Voloshynovskiy (Université de Genève)

Genetic big data: une indexation puissante

Ce projet a créé une méthode pour indexer de très grandes bases de données de séquences génétiques après les avoir compressées par un facteur mille. Cet outil public appelé Metagraph aide à explorer efficacement de grandes bases de données génétiques et à effectuer des analyses complexes. Cette approche a de nombreuses applications en médecine personnalisée, ou pour suivre les mutations de pathogènes. Elle contribue à tirer profit de bases de données génétiques de taille croissante, telles que l'Atlas du génome du cancer, qui contiennent souvent des pétaoctets (millions de gigaoctets). Le projet a indexé plus de 1,4 million de séquences génomiques complètes. Une plateforme interactive donne accès à des informations génétiques de 500 000 plantes, 450 000 bactéries et 120 000 champignons, ainsi que 240 000 métagénomes intestinaux humains.

—

Scalable genome graph data structures for metagenomics and genome annotation
Gunnar Rätsch (ETH Zurich)

Comparaison de génomes: des analyses plus rapides

Ce projet a créé des méthodes d'apprentissage automatique pour comparer les génomes de différents organismes malgré des données de qualité variable. Ces comparaisons permettent de mieux comprendre l'évolution de groupes de gènes impliqués dans des processus métaboliques spécifiques, en révélant leur association à des fonctions d'entretien ou à l'évolution. Le projet a développé des méthodologies pour trouver des gènes ou des protéines ayant la même fonction chez l'humain et dans des organismes modèles tels que la mouche ou la souris. Ce type de recherche est crucial pour les efforts à grande échelle, tels que l'Atlas européen des génomes de référence ou le projet Earth BioGenome, qui visent à séquencer les génomes de tous les animaux, plantes et champignons connus. Les résultats du projet contribuent à accélérer considérablement les comparaisons de génomes.

—

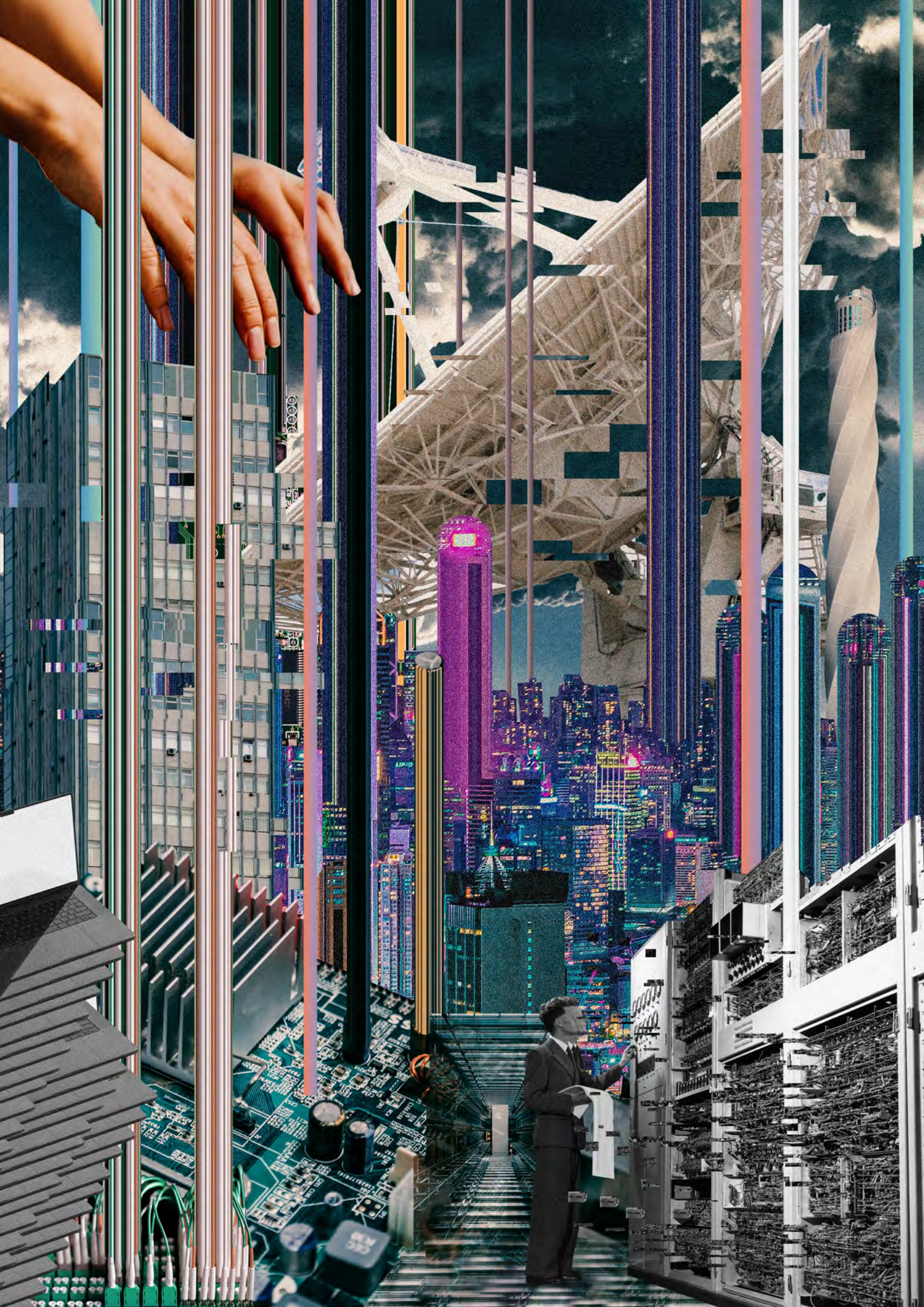
Efficient and accurate comparative genomics to make sense of high-volume low-quality data in biology
Nicolas Salamin (Université de Lausanne)

Bases de données bioinformatiques: recherches en langage naturel

Ce projet a conçu une interface intuitive pour interroger des bases de données bioinformatiques en utilisant le langage naturel. La plateforme fournit des indications visuelles pour faciliter la recherche de mots clés pendant la phase exploratoire et donne un feedback pour affiner les requêtes et améliorer les résultats. Le système a intégré des bases de données de protéines (UniProt), d'expression de gènes (Bgee) et d'expression de gènes inter-espèces (OMA). Cette approche pourrait être utilisée en dehors des sciences de la vie: elle est poursuivie dans le projet européen INODE qui englobe la recherche sur le cancer, l'astrophysique et l'élaboration de politiques d'innovation.

—

BIO-SODA: enabling complex, semantic queries to bioinformatics databases through intuitive searching over data
Kurt Stockinger (Université des sciences appliquées de Zurich, ZHAW)



3.
Technologies du big data

Les applications du big data ont besoin des technologies du big data: du hardware et du software capables de traiter des jeux de données massifs et de les analyser de manière fiable et efficace. Les résultats des recherches du PNR 75 montrent comment les scientifiques travaillant en Suisse contribuent au développement de nouvelles technologies pour les applications du big data et peuvent aider à déployer avec succès des solutions d'infrastructure et des méthodes d'analyse de nouvelle génération.

Utiliser le big data en pratique est confronté à de grands défis technologiques. L'un d'entre eux est le volume des données: le big data dépasse – selon la définition courante – les capacités des technologies existantes de stockage, gestion et analyse des données. De nombreuses infrastructures informatiques actuelles seront bientôt dépassées et devront être remplacées. Le big data appelle de nouveaux moyens de traitement et d'analyse de l'information.

C'est pourquoi la recherche fondamentale sur les infrastructures du big data et sr les technologies d'analyse est cruciale. Elle jette les bases sur lesquelles les applications futures seront construites et soutient l'approvisionnement en spécialistes de haut niveau. La Suisse produit une recherche en informatique de classe mondiale et doit continuer à la soutenir. Elle doit aussi s'engager pour de nouvelles infrastructures et pour la production et conservation de jeux de données fiables. Ces efforts la préparent à relever les défis technologiques du big data.

Le PNR 75 a renforcé la recherche fondamentale suisse dans ce domaine. Il a produit une douzaine de nouvelles approches pour les technologies qui sous-tendent les applications du big data. Elles peuvent être regroupées en deux domaines:

— l'infrastructure informatique – principalement logicielle – nécessaire pour accéder aux données et les nettoyer, stocker, indexer, contrôler, prétraiter et surveiller (voir section 3.1);
— l'analytique des données, à savoir les algorithmes permettant d'en extraire des connaissances (voir section 3.2).

La section 3.3 décrit les défis de la recherche sur les technologies du big data. La section 3.4 résume les messages clés, tandis que la section 3.5 présente les projets individuels.

3.1 Des infrastructures du big data plus efficaces

Le big data a besoin d'une infrastructure performante, de processus de bas niveau qui servent d'épine dorsale aux analyses de données de plus haut niveau. Cette infrastructure est matérielle et logicielle.

— Le hardware comprend les processeurs tels que les CPU, les GPU et les TPU (processeurs, cartes graphiques et unités de traitement de tenseur), la mémoire transitoire et le stockage permanent, les composants de communication, les unités d'alimentation et de refroidissement, etc.
— Le software permet l'accès aux données et leur prétraitement, la programmation, la surveillance des flux de données et du matériel, etc.

L'amélioration de l'infrastructure du big data nécessite des avancées tant au niveau matériel que logiciel. Alors que l'industrie est à l'origine des principales avancées en matière de matériel, la recherche universitaire apporte une contribution importante aux logiciels permettant un traitement plus rapide et plus efficace de grands jeux de données.

Le PNR 75 a exploré, proposé et testé une demi-douzaine d'approches – principalement logicielles – pour améliorer les infrastructures en:

Concepts technologiques clés du big data

Métadonnées Informations sur un point de données, telles que le lieu et la date de son acquisition, son type ou sa catégorie.
Anonymisation Suppression des informations susceptibles de révéler l'identité d'une personne afin de rendre les données anonymes (ou plutôt «pseudonymes»).
Réidentification Combinaison de plusieurs jeux de données anonymes dans le but d'identifier des individus.
Intelligence artificielle Algorithmes et machines démontrant un comportement «intelligent», ainsi que les méthodes sous-jacentes et applications.
Apprentissage automatique Techniques informatiques permettant aux algorithmes d'apprendre de manière autonome, par exemple grâce à des données d'apprentissage.
Apprentissage supervisé Approche d'apprentissage automatique dans laquelle les algorithmes apprennent à partir de données d'apprentissage labellisées.
Apprentissage non supervisé Approche d'apprentissage automatique dans laquelle les algorithmes découvrent des caractéristiques dans les jeux de données sans utiliser de données d'apprentissage labellisées.

— contribuant au domaine émergent de l'in-network computing, à savoir le traitement des données pendant leur transfert effectué dans les nœuds de communication *(In-network computing);*
— améliorant l'analyse des réseaux dynamiques de données en continu *(Exploration de graphes);*
— analysant automatiquement les descriptions de données hétérogènes afin de les combiner et de les préparer pour un traitement ultérieur *(Données peu structurées);*
— surveillant en temps réel la conformité des données avec des règles *(Flux de données)*
— créant des interfaces pour traiter de grands jeux de données dans le langage de programmation Scala *(Scala pour le big data).*

Le programme n'a couvert qu'une partie restreinte des questions soulevées par les technologies du big data et des approches explorées actuellement dans le monde entier.

Impact sur les infrastructures du big data

Le traitement des données réalisé dans les nœuds de communication pendant leur transfert permet d'augmenter considérablement la vitesse des processus et de réduire la latence des opérations *(In-network computing).* De nouveaux algorithmes parallélisés peuvent surveiller en temps réel si un flux rapide de données volumineuses est conforme à des règles déterminées. Les méthodes basées sur la logique formelle sont souvent considérées comme bien plus lentes que les approches d'apprentissage automatique, mais peuvent être utilisée à plus grande l'échelle *(Flux de données).*

Une nouvelle architecture informatique peut améliorer l'analyse de réseaux dynamiques massifs *(Exploration de graphes).* Un outil pour classer différents types de données aide à développer des applications combinant chiffres, textes, images et vidéos *(Données peu structurées).*

La sortie d'une nouvelle version du langage de programmation Scala en 2021 facilite l'interface avec de grands jeux de données et la création d'applications du big data dans de nombreux secteurs. Inventé en Suisse, Scala est utilisé par de grandes entreprises technologiques, bancaires et des médias *(Scala pour le big data).*

3.2 Nouvelles approches pour l'analyse du big data

L'analytique représente la composante la plus visible des applications du big data, créant de la valeur à partir des données. Les algorithmes modernes d'exploration des données et de requêtes analytiques multidimensionnelles combinent des méthodes statistiques avancées et des méthodes d'apprentissage automatique tels que l'apprentissage profond.

Le défi est d'analyser de très grands jeux de données de manière fiable, précise et dans un temps raisonnable.

Malgré la puissance des systèmes informatique distribués et des processeurs dédiés, l'entraînement d'un

algorithme avec des données d'apprentissage peut encore prendre des jours. Accélérer cette étape exige des algorithmes plus efficaces, ou des moyens d'élaguer les jeux de données d'entraînement, ou de les compresser sans réduire la qualité des résultats.

Le PNR 75 a amélioré des outils d'analyse du big data de six manières, notamment en développant des approximations efficaces pour accélérer le traitement et en réduisant la quantité de données d'apprentissage, sans affecter indûment la précision. Les recherches ont:

— amélioré les techniques d'analyse en continu de très grands jeux de données tout en préservant la vie privée *(Analyse de flux);*
— réduit la taille de jeux de données utilisés pour entraîner des modèles d'apprentissage automatique *(Coresets);*
— diminué le temps nécessaire à l'entraînement d'algorithmes de modélisation de données et de prédiction *(Algorithmes de prédiction rapide);*
— permis un monitoring en temps réel efficace de centres de calcul *(Data centres);*
— élargi la compréhension des limites des réseaux de neurones profonds *(Apprentissage automatique);*
— amélioré les modèles de langage pour des systèmes de dialogue automatisés *(Modèles de langage).*

Impact sur l'analytique du big data

La recherche du PNR 75 a généré des progrès dans plusieurs types d'analyse du big data. De nouvelles techniques permettent d'entraîner plus rapidement des algorithmes pour analyser les processus «gaussiens» – utilisés pour modéliser les données et faire des prédictions – tout en quantifiant plus précisément leur incertitude *(Algorithmes de prédiction rapide).* Certains types de traitements de données, comme l'analyse d'images, peuvent désormais être effectués en temps réel sur un flux de données entrant, sans avoir à attendre que toutes les données soient stockées. Cette approche peut également être utilisée pour protéger la vie privée lors de la génération et de la transmission des données grâce à l'ajout de perturbations aléatoires aux données afin de dissimiler des informations personnelles *(Analyse de flux).*

Une meilleure compréhension des limites des réseaux de neurones profonds, comme la possibilité de les généraliser et de les utiliser dans des contextes différents, leur permet d'être plus robustes et fiables *(Apprentissage automatique).* Les avancées théoriques dans le domaine de systèmes de dialogue renforcent leur sophistication et fiabilité *(Modèles de langage).*

La quantité de données peut être réduite sans compromettre la fiabilité, notamment celles utilisées pour entraîner les modèles d'apprentissage automatique, ce qui réduit aussi les ressources informatiques nécessaires *(Coresets).* Il est possible de monitorer et optimiser les processus dans les centres de calcul même à l'aide de données restreintes sur les ressources informatiques employées *(Data centres).* Une nouvelle méthode de synthèse de données tabulaires minimise le risque de fuite d'informations propriétaires ou confidentielles *(Data centres).*

3.3 Défis de la recherche sur les technologies du big data

Défis

Des défis similaires se retrouvent pour les projets de recherche sur les infrastructures et sur les techniques d'analyse du big data. Il peut être difficile d'accéder aux données nécessaires au développement d'un nouveau logiciel d'infrastructure, les entreprises n'étant pas toujours disposées à, ou en mesure de fournir des études de cas réalistes. Cela peut être dû au secret d'affaires et aux réglementations sur la vie privée telles que le RGPD *(Data centres),* à des changements de personnel ou de politique interne, ou encore à des attentes peu réalistes sur les applications *(Flux de données).* Certains projets ont dû recourir à des données synthétiques ou de benchmark, pas toujours bien adaptées à une utilisation pratique. Dans l'ensemble, les procédures régissant l'accès, le traitement et le partage des données, en particulier celles qui concernent la vie privée et le consentement, sont perçues par les scientifiques comme rigides, lourdes et chronophages *(Exploration de graphes).*

La recherche sur le big data est très compétitive, notamment en raison de l'implication des multinationales de la technologie. Ces organisations investissent massivement dans la R&D, disposent d'une infrastructure informatique inégalée et d'un accès aux plus grands jeux de données du monde, et recrutent des spécialistes de classe mondiale. Le monde universitaire peut collaborer avec l'industrie privée à un haut niveau, mais il doit aussi trouver des créneaux dans lesquels rivaliser avec elle *(Modèles de langage).* Avoir des financements à long terme et des plans de recherche flexibles est important afin de pouvoir saisir les opportunités survenant dans ce domaine en évolution rapide.

Mener des recherches de pointe exige de recruter les meilleurs scientifiques, mais la concurrence des multinationales de la technologie rend la tâche difficile. En outre, le monde universitaire accorde peu d'importance à l'innovation, notamment au développement de prototypes avec des partenaires publics et privés, ou à la création de logiciels open-source. Les incitations à poursuivre de tels projets manquent encore *(Scala pour le big data).*

Pistes de solutions

Les processus administratifs actuels concernant l'accès, le partage et le traitement des données en Suisse pourraient être rationalisés lorsqu'ils concernent la recherche publique. L'utilisation des données dans les applications peut avoir une incidence négative sur la vie privée, mais restreindre l'utilisation des données d'une manière générale présente des inconvénients, comme créer un frein à l'innovation. Protéger la vie privée a un certain coût, et celui-ci doit également être pris en compte.

Une piste serait de faire de la confidentialité un aspect inhérent, voire obligatoire, du traitement du big data. Les gens qui développent les technologies doivent être informés des solutions pour préserver la vie privée, de leurs avantages et de leurs inconvénients. Dans l'idéal, ils ont accès à des outils les aidant à optimiser les algorithmes suivant l'équilibre souhaité entre respect de la vie privée, efficacité

et qualité des services. La numérisation de l'information exige une analyse précise pour garantir une utilisation judicieuse des formats et des métadonnées.

L'évaluation du succès académique, en particulier pour la promotion des carrières et l'attribution de fonds, doit aller au-delà du nombre de publications et de citations et inclure l'impact des travaux en dehors du milieu universitaire (notamment lors de l'utilisation de protocoles open-source). Les scientifiques doivent disposer de liberté et de flexibilité pour être en mesure d'adapter leurs plans et de tirer le meilleur parti de domaines évoluant rapidement tels que le big data.

Le facteur humain peut s'avérer aussi important que l'accès à la technologie, cette dernière étant souvent disponible en open-source. Soutenir la recherche universitaire permet non seulement de faire progresser la technologie du big data au niveau local et international, mais aussi de former les spécialistes dont la société a besoin. Ces derniers ne se contentent pas de développer des technologies, mais apportent également leurs compétences sur des questions telles que la disponibilité des technologies, la confidentialité, la cybersécurité et l'inclusion des parties prenantes. À ce titre, ils contribuent aux décisions stratégiques des institutions publiques et privées en matière de numérisation.

3.4 Messages clés sur les technologies du big data

Le PNR 75 a produit de nombreux résultats de recherche de premier plan ouvrant de nouvelles voies pour améliorer l'infrastructure et l'analytique nécessaires à l'exploitation du big data. Ce type de recherche fondamentale est difficile, mais peut suivre un cheminement relativement bien connu et souvent plus simple que le développement d'applications. Par exemple, l'accès limité à des données réelles peut être parfois contourné avec des données produites artificiellement permettant de tester les nouveaux systèmes et de les optimiser. La recherche se concentre principalement sur la question de la vitesse à laquelle les systèmes peuvent traiter et analyser les données, et produire le résultat escompté avec une marge d'erreur connue. En d'autres termes, les problèmes de recherche sont bien définis. Cependant, leur environnement évolue rapidement et comprend des acteurs aux objectifs distincts, voire divergents.

Concurrence et coopération entre recherches privée et publique

L'intense concurrence internationale dans le domaine du big data met en danger l'autonomie numérique des États, mais offre aussi des opportunités de collaboration. L'économie privée, en particulier aux États-Unis et en Chine, se trouve à l'origine d'une grande partie des avancées sur les infrastructures et l'analyse du big data, qui constituent un

Les défis des données de recherche ouvertes

Le partage des résultats et des données rend la recherche plus efficace. Les données peuvent être réutilisées et combinées, et des nouvelles études peuvent s'inspirer des résultats existants. Le partage augmente la productivité en réduisant le temps nécessaire à la collecte des données et stimule la créativité en facilitant des expériences novatrices à faible coût. Il élargit également l'accès aux résultats de la recherche, y compris aux personnes extérieures au milieu universitaire. Cependant, les données de recherche restent dans les faits souvent inaccessibles et peu réutilisées. Les scientifiques qui souhaitent partager leurs données se heurtent encore à de nombreux obstacles pratiques, institutionnels et financiers.

Le PNR 75 a abordé ces questions avec l'activité transversale *Big data: open data and legal strings*. Celle-ci a examiné les défis concrets auxquels sont confrontés les scientifiques lorsqu'ils partagent, publient et réutilisent des données. Elle a mené des entretiens, proposé des conseils concrets sous la forme d'un guide pratique et formulé des recommandations claires à l'intention des institutions de recherche.

Les principaux défis

→ Des obstacles pratiques et juridiques entravent l'accès aux données de recherche existantes: faible qualité des données, formats obsolètes, identification des sources, diversité des lieux de stockage, dépôts inaccessibles, sites web non mis à jour, statut juridique des données et restrictions en matière de réutilisation, lois sur la protection des données, violation des droits d'autrui, etc.

→ Les scientifiques se heurtent également à des obstacles organisationnels, financiers et juridiques dans la publication des données de recherche: manque d'incitations, coûts, savoir-faire technique et juridique, engagement à long terme, risque de se voir devancés par d'autres scientifiques ou d'utilisation abusive des données par des partenaires extérieurs.

→ La faible standardisation des données de recherche limite le potentiel de réutilisation.

Pistes de solutions

→ Conseils juridiques sur des questions de propriété des données, propriété intellectuelle et droit d'auteur, accords contractuels avec des tiers et droit de la protection des données.

→ Standardisation: formats de données, stockage, processus d'anonymisation, procédures réglementaires.

→ Identification et publication des meilleures pratiques et de guides pratiques.

→ Incitations financières et de réputation pour le partage.

Big data: open data and legal strings
Sabine Gless (Université de Bâle)

bon tiers des travaux présentés lors de conférences scientifiques de haut niveau. Elles ont apporté des modèles de langage perfectionnés et le développement de matériel tel que les unités de traitement de tenseur, optimisées pour implémenter des réseaux neuronaux. La R&D industrielle se trouve au moins au même niveau que la meilleure recherche universitaire.

Les technologies du big data développées par les entreprises peuvent paraître universelles au vu de l'intérêt d'une large adoption – un processeur ou un algorithme peut être agnostique quant à la manière dont il sera utilisé. Mais les technologies se spécialisent de plus en plus pour répondre de la manière la plus efficace aux problèmes concrets et aux types de données (dynamiques ou statiques, homogènes ou hétérogènes, etc.). L'industrie influence grandement la gamme des applications possibles du big data, et il est crucial que la recherche publique soit en mesure de suivre le rythme rapide de l'industrie si l'on veut que la société ait son mot à dire sur l'avenir de la numérisation.

La recherche universitaire reste essentielle pour développer les technologies du big data, en particulier lorsqu'il s'agit d'atteindre des objectifs qui sont plus importants pour la société que pour les grandes entreprises, comme la réduction de la consommation d'énergie, ou la garantie du «privacy-by-design». Alors que l'industrie suit souvent des approches uniformes, la recherche publique peut être plus audacieuse en s'engageant dans des voies prometteuses, mais risquées. Elle a ainsi développé avec succès une gamme plus large de matériels et de logiciels pour les technologies du big data. Il s'agit notamment des commutateurs de réseau programmables, des analyses réalisées dans le réseau lui-même, de nouvelles approches de programmation pour des dispositifs spécifiques à un domaine, et d'algorithmes basés sur la logique formelle plutôt que sur l'apprentissage automatique.

Si la recherche universitaire est en général très en avance sur l'industrie privée (cette dernière s'appuyant sur l'innovation des spin-offs universitaires), cet écart est beaucoup plus faible dans certains domaines du big data. Cela encourage les collaborations entre le monde universitaire et l'industrie, d'autant plus que le premier a besoin de la puissance de calcul, des capacités de stockage et de l'accès aux données du second. Ces collaborations sont en principe bénéfiques aux deux parties, le monde universitaire bénéficiant des ressources de l'industrie, de problèmes issus du monde réel et de défis de taille, tandis que l'industrie profite d'une recherche de pointe et d'idées plus innovantes.

Un problème latent dans le monde universitaire est le manque de reconnaissance accordé aux scientifiques qui développent des applications, stimulent la collaboration et adoptent des logiciels open-source. Cela peut dissuader les talents de classe mondiale de s'attaquer à des problèmes concrets et de collaborer avec l'industrie. Il est donc nécessaire de diversifier les parcours professionnels dans la recherche publique et d'utiliser des indicateurs qui vont au-delà des publications scientifiques et des fonds de recherche.

La question du personnel

L'un des principaux défis du déploiement du big data est la pénurie de personnel qualifié tout au long de la chaîne de valeur, des technologies d'infrastructure aux applications, en passant par l'intégration commerciale et la réglementation. La concurrence pour les talents est intense, et les spécialistes les plus brillants sont nombreux à rejoindre de grandes multinationales, de petites et moyennes entreprises et des start-ups. La recherche universitaire en pâtit, car elle peine à garder les meilleurs scientifiques, même au niveau du doctorat. Les universités risquent de perdre leurs talents lorsqu'elles collaborent avec de grandes entreprises technologiques. Les changements rapides et fréquents de carrière, bien qu'ils apportent de nouvelles perspectives et de nouvelles connexions, constituent un problème pour les projets de recherche.

D'un autre côté, le haut niveau de la recherche suisse sur le big data permet de former les nombreux spécialistes dont les organisations publiques et privées ont besoin, et d'entretenir de bons contacts entre le monde universitaire et l'industrie. Cela rend la Suisse innovante et attrayante pour les entreprises multinationales et les organisations internationales.

Obtenir les données

Le second grand défi concerne la disponibilité de grands jeux de données de haute qualité, essentiels pour une évaluation réaliste des analyses et des applications du big data (voir chapitre 2).

Ce problème s'atténuera à mesure que les organisations publiques et privées développeront une culture des données, mais le résoudre nécessite une stratégie solide garantissant que les données sont de haute qualité, correctement décrites par des métadonnées et protégées par des pratiques «privacy-by-design». Les développeurs et développeuses des technologies du big data doivent connaître les différentes techniques de préservation de la vie privée pour être en mesure de trouver le bon équilibre entre vie privée, efficacité et qualité des services.

La vie privée et la protection des données soulèvent de nombreuses questions, comme celle de savoir si la règlementation suisse ou européenne sur la protection des données fixe des limites appropriées en matière de gestion des données ou comment protéger la vie privée tout en encourageant l'innovation. Les processus éthiques, d'approbation et administratifs actuels encadrant l'utilisation des données, médicales et scientifiques en Suisse sont perçus comme lents et complexes, et pourraient être rationalisés et simplifiés. Mais il s'agit d'une discussion aux multiples facettes qui nécessite des approches multidisciplinaires, y compris l'implication des sciences sociales. Elle est traitée dans le chapitre 4.

3.5 Les projets de recherche sur les technologies du big data

Le PNR 75 a exploré une douzaine d'approches pour améliorer les technologies nécessaires au big data, la moitié dans le domaine des infrastructures, l'autre dans celui de l'analytique.

Infrastructures du big data

In-network computing: solutions pour l'analyse des graphes

Ce projet a fait plusieurs avancées dans l'analyse de grands graphes (réseaux) grâce au in-network computing, à savoir le traitement des données lors de leur transit avant stockage. Les messages sur les réseaux sociaux et les appels téléphoniques forment, par exemple des graphes très complexes d'une taille croissante, et dont les traces sont de plus en plus disponibles. Les nouvelles méthodes reposent notamment sur des composants tels que les circuits intégrés ASIC ou FPGA. Elles augmentent de plusieurs ordres de grandeur les performances de systèmes logiciels couramment utilisés, et peuvent traiter plus de quatre milliards d'événements par seconde.

—

Exploratory visual analytics for interaction graphs

Robert Soulé (Université de la Suisse italienne)

Analyse et exploration de graphes

Ce projet a extrait des inférences d'un réseau et étudié l'analyse de graphes sur différentes plateformes, y compris

des combinaisons de traitement in-core et out-of-core. Les résultats se sont révélés utiles pour les systèmes de stockage employés dans le traitement général du big data.

—

Building flexible large-graph analytics and mining systems on commodity hardware

Willy Zwaenepoel (University of Sydney)

Données peu structurées: nouveaux outils d'intégration

Ce projet a créé des outils pour combiner différents types de données (texte, PDF, images, etc.) et les préparer pour leur traitement et analyse. Cette variété représente un défi, en particulier à cause du manque de cohérence des métadonnées. Ces nouveaux outils accélèrent la transformation des données brutes en modèles et visualisations grâce à une intégration de jeux de données automatique ou semi-automatique, bien que toute contribution humaine n'ait pas disparu.

Ils ont été testés sur de nombreux jeux de données tels que des messages Twitter, des actualités, des documents PDF et des images médicales. L'équipe de recherche a développé un outil pour extraire des données peu structurées dans les documents d'archives, en collaboration avec les Archives fédérales suisses, et a pu identifier des documents sans données accessibles librement. Des collaborations avec les hôpitaux cantonaux de Fribourg et de Berne ont abouti au prototype d'une nouvelle architecture pour intégrer des informations sur le cancer de la prostate, en particulier des images médicales.

—

Tighten-it-all: big data Integration for loosely-structured data

Philippe Cudré-Mauroux (Université de Fribourg)

Flux de données: monitoring en temps réel

Ce projet a mis au point des algorithmes parallélisés efficaces pour surveiller en temps réel si un flux rapide de données volumineuses respecte des règles précises. Il est difficile de contrôler des règles complexe de manière efficace pour de très grands volumes de données. Les algorithmes de surveillance doivent être évolutifs pour pouvoir être exécutés de manière parallélisée dans des clusters informatiques.

Le projet a développé des algorithmes de contrôle pour des langages dits expressifs et les a testés dans deux applications concrètes. Un outil a été créé pour auditer les remboursements de frais professionnels. Une collaboration avec la société de télécommunications Huawei a montré la faisabilité d'outils pour vérifier qu'un traitement est neutre sur le plan des données, c'est-à-dire qu'il ne peut pas être contrôlé par le fournisseur. Le projet a montré que ces nouveaux algorithmes pouvaient rivaliser avec les systèmes de surveillance de niveau industriel.

—

Big data monitoring

David Basin, Dmytro Traytel (ETH Zurich)

Scala pour le big data

Ce projet a introduit plusieurs nouveaux concepts pour Scala, un langage de programmation développé à l'EPFL qui est devenu un leader pour les plateformes et outils de science des données. L'équipe a intégré plusieurs nouvelles technologies dans un ensemble cohérent d'abstractions pour l'interfaçage avec un grand jeu de données, et les a validées dans des projets open-source. La nouvelle version de Scala, publiée en 2021, devrait être adoptée par des centaines de milliers de développeurs de logiciels.

—

Programming language abstractions for big data

Martin Odersky (EPFL)

Analytique du big data

Analyse de flux: traitement rapide et préservation de la vie privée

Ce projet a créé plusieurs outils pour le traitement en continu et en temps réel de grands flux de données. Ils incluent une protection renforcée de la vie privée grâce à des perturbations ajoutées aux jeux de données. Le premier est un algorithme de traitement des images. Il a été testé avec succès sur les observations du Square Kilometre Array Pathfinder, un radiotélescope australien produisant jusqu'à 2 gigaoctets de données brutes par seconde. L'algorithme a pu reconstruire les images astronomiques en temps réel, sans avoir à attendre la réception de toutes les données d'image, ouvrant la voie à l'analyse en temps réel à l'échelle du pétaoctet.

Un deuxième outil aide les non-spécialistes à implémenter et utiliser la confidentialité différentielle, qui modifie de manière aléatoire des données afin d'éviter le risque de réidentification des personnes. L'équipe a créé des méthodes pour implémenter une confidentialité différentielle en temps réel dans un flux de données continu et proposé une nouvelle façon de représenter les paramètres contrôlant la quantité de perturbations aléatoires ajoutées. Les usagers et usagères peuvent décider du compromis entre le respect de la vie privée (avec des perturbations importantes) et la précision (perturbations plus faibles). L'équipe a également publié un langage de programmation modifié pour simplifier l'intégration et le contrôle des techniques de confidentialité différentielle pour les non-spécialistes. Elle a testé ces concepts en analysant les habitudes télévisuelles d'environ trois millions de personnes.

—

Privacy preserving, peta-scale stream analytics for domain-experts

Michael Böhlen (Université de Zurich)

Coresets: du big data avec moins de données

Ce projet a montré que l'on peut réduire la quantité de données nécessaires à l'entraînement de modèles d'apprentissage automatique sans diminuer de manière importante la précision des analyses statistiques clés et des processus d'apprentissage. Cette méthode fonctionne également avec des flux de données dynamiques. L'équipe a utilisé des «coresets», qui peuvent résumer un grand jeu de données à partir d'un petit échantillon tout en pouvant être traités de manière robuste et précise. Cette approche peut également améliorer la protection de la vie privée, les données individuelles étant modifiées de manière significative dans les échantillons de coresets.

—

Scaling up by scaling down: big ML via small coresets

Andreas Krause (ETH Zurich)

Data centres: monitoring efficace des performances

Ce projet a conçu de nouvelles méthodes d'analyse des performances dans les centres de données du cloud, une tâche importante pour gérer efficacement les ressources informatiques et minimiser la consommation d'énergie. Des analyses approximatives utilisant des sous-ensembles de données de performance (un journal des ressources informatiques virtuelles et physiques) ont pu prédire l'utilisation des ressources. Cette approche pourrait améliorer les systèmes de monitoring actuels, habituellement peu sophistiqués et lents.

Les résultats montrent qu'il est plus efficace d'entraîner un modèle avec un petit jeu de données pertinentes et propres qu'avec une grande quantité de données de faible qualité. Pour éviter une perte de précision des inférences, le choix du sous-ensemble approprié est essentiel, par exemple avec

un résumé systématique du jeu de données plutôt que des sous-échantillons uniformes, ou pris au hasard. L'équipe a mis au point une méthode pour synthétiser des données tabulaires de telle sorte que les données propriétaires fournies par des sociétés commerciales puissent être partagées sans être divulguées.

—

Dapprox: dependency-ware approximate analytics and processing platforms
Lydia Yiyu Chen (Université de technologie de Delft)

Apprentissage automatique: robustesse et généralisabilité

Ce projet a fait plusieurs avancées théoriques pour évaluer si les modèles d'apprentissage automatique sont robustes (fiabilité lorsque les données d'entrée sont perturbées) et généralisables (capacité à traiter des données du monde réel différentes de celles utilisées pour entraîner les modèles). Les résultats ont mis au jour un compromis entre l'efficacité et la généralisabilité d'un modèle. Ces avancées sont importantes pour améliorer l'interprétabilité et la reproductibilité des modèles, des ingrédients cruciaux pour garantir l'impartialité et la fiabilité des algorithmes.

—

Theory and methods for accurate and scalable learning machines
Volkan Cevher (EPFL)

Algorithmes de prédiction rapide

Ce projet a créé des algorithmes capables d'apprendre à partir de grands jeux de données plus rapidement que les méthodes actuelles à la pointe. Ils font appel à une méthode statistique puissante (les processus gaussiens) utilisée pour modéliser les données, tirer des inférences et faire des prédictions. Les prévisions sont exprimées non pas sous forme de valeurs uniques, mais de distributions de probabilité. Il s'agit d'un point particulièrement important lorsque les incertitudes se doivent d'être quantifiées, comme pour les prévisions météorologiques. L'équipe a utilisé une formulation d'un filtre de Kalman sur un ensemble réduit de données d'apprentissage (les points d'induction), ainsi que des approximations locales utilisant la méthode des produits corrélés d'experts. Le temps de calcul et la complexité n'augmentent que proportionnellement à la taille des échantillons de données, comparé à une croissance cubique avec les méthodes précédentes. Ces travaux offrent de bonnes perspectives pour des applications avec de très grands jeux de données, ainsi que pour mieux estimer l'incertitude des prédictions et inférences.

—

State space Gaussian processes for big data analytics
Marco Zaffalon (Istituto Dalle Molle di studi sull'Intelligenza Artificiale USI-SUPSI)

Modèles de langage: nouvelles méthodes pour agents conversationnels

Ce projet a réalisé plusieurs avancées théoriques dans le domaine des modèles de langage (des algorithmes qui génèrent du texte), notamment pour les agents conversationnels et les systèmes répondant à des requêtes. Cette tâche exige des réponses adaptées aux questions, et une certaine compréhension des inputs exprimés dans une langue naturelle, telle que l'anglais ou le mandarin. Ces systèmes sont utilisés dans le service clientèle, les moteurs de recherche, les réseaux sociaux ou encore le commerce électronique, et sont susceptibles d'avoir d'importantes conséquences économiques et sociales.

L'équipe a développé des technologies intégrables dans des agents conversationnels pour comprendre les éléments de texte se référant à un agent (détection et liaison d'entités), générer

des textes à l'aide de réseaux de neurones profonds, évaluer et améliorer les performances des algorithmes linguistiques (apprentissage par renforcement) et de l'apprentissage automatique géométrique (qui permet d'ajouter de la complexité et de la structure). Une partie du travail a été réalisée avec Google Zurich. Un spin-off a été lancé pour commercialiser des solutions de recherche sémantique de dossiers juridiques, et de rédaction de documents juridiques.

—

Conversational agent for Interactive access to information
Thomas Hofmann (ETH Zurich)

4.

Aspects sociétaux, juridiques et éthiques du big data

La collecte, l'analyse, le stockage et le partage de grands jeux de données soulèvent des questions sociétales, juridiques et éthiques nombreuses et profondes. Trouver des réponses nécessite des efforts interdisciplinaires rassemblant de multiples parties prenantes. Ce chapitre examine les nouvelles perspectives apportées par le PNR 75 sur les questions de propriété, contrôle, accès et transfert des données, de vie privée et de souveraineté numérique, de discrimination et d'équité, ainsi que de gestion des savoirs.

L'offre et l'utilisation toujours plus grandes de données dans la société ont attiré l'attention des gouvernements, des entreprises, des organisations et des particuliers, et soulevé de profondes questions éthiques, juridiques et sociales. Ces dernières font l'objet de nombreux débats, mais n'ont pas encore été entièrement cartographiées. Des lignes directrices adéquates sur la collecte, l'analyse, l'utilisation et le partage des données doivent encore être conceptualisées, développées, testées et prises en compte dans les réglementations gouvernementales et institutionnelles, et mises en pratique.

L'influence du big data s'observe dans tous les aspects de la société, par exemple sur les réseaux sociaux où la combinaison de nombreux flux d'informations serait impensable sans des techniques d'analyse avancées. Celles-ci rendent également possible des comportements nocifs et des manipulations délibérées, générant ainsi des risques nouveaux ou exacerbant les dangers existants à une échelle et une vitesse sans précédent. Les processus démocratiques sont affectés, positivement aussi bien que négativement.

La sensibilité politique autour des impacts négatifs potentiels du big data a considérablement augmenté ces dernières années, comme lors de la révision de la nouvelle loi sur la protection des données (nLPD), qui entrera en vigueur le 1er septembre 2023.

Les connaissances produites par le PNR 75

Le Programme national de recherche «Big Data» (PNR 75) a analysé de nombreuses questions éthiques, réglementaires et juridiques soulevées par la croissance rapide des applications et des pratiques du big data, tant à un niveau conceptuel que dans des contextes spécifiques.
Les scientifiques du PNR 75 ont étudié des domaines d'application concrets, tels que la santé, exploré la propagation des réglementations à l'international, rédigé des lignes directrices pour le secteur des assurances, analysé le potentiel de discrimination dans les ressources humaines, proposé des cadres pour l'utilisation éthique des données dans les soins de santé et examiné comment la profession de spécialiste des données a vu le jour. Ils se sont également penchés sur des questions génériques liées à la souveraineté et au contrôle des données, sur la réglementation de l'utilisation du big data dans la recherche, et sur la nécessité de faire face aux nouvelles incertitudes inhérentes aux prévisions produites par des modèles.
Les résultats des recherches du PNR 75 sur les aspects éthiques et juridiques du big data sont organisés selon quatre thèmes:

— propriété, accès et transfert des données dans la section 4.1,
— vie privée et souveraineté numérique dans la section 4.2,
— non-discrimination, équité et inclusion dans la section 4.3,
— production et gestion des connaissances dans la section 4.4.

Une perspective sur les questions sociétales est présentée dans la section 4.5. La dernière section 4.6 résume les huit projets du PNR 75 sur ce thème.

4.1 Données: droit de propriété, contrôle, accès et transfert

Les données sont désormais produites, collectées, analysées et partagées à une échelle sans précédent par des organisations et des personnes aux rôles variés. Les gouvernements publient une partie de leurs données (généralement agrégées et rarement personnelles) conformément au paradigme des données ouvertes, alors que des entités commerciales accumulent les données, ne les partageant qu'avec réticence.

Produites à très grande échelle, les données personnelles ont une valeur particulièrement élevée pour les gens concernés. Les individus partagent et donnent accès à une grande partie de leurs données en échange de services gratuits, efficaces, pratiques et dont il est souvent difficile de se passer, comme le courrier électronique, les messageries instantanées, le partage de photos et de vidéos, les cartes, les recommandations ciblées et les réseaux sociaux. Le public est conscient, dans une certaine mesure, des risques pour la vie privée ainsi que des possibilités de surveillance et d'abus, mais le partage très libéral des données personnelles se poursuit sans relâche. Des réglementations nationales et internationales – telles que la nouvelle loi sur la protection des données (nLPD) ou le règlement général sur la protection des données (RGPD) de l'Union européenne – sont passées ou actualisées, mais à un rythme relativement lent. Elles entrent en vigueur des années ou des décennies après que la collecte des données a débuté, réduisant d'autant leur efficacité.

Les projets du PNR 75 ont étudié la propriété des données, le contrôle sur elles, leur circulation transfrontalière (voir section 4.6 pour plus de détails). Ils ont analysé

— les défis juridiques soulevés par le big data, en particulier la propriété, la protection contre la surveillance indue et contre l'auto-incrimination par ses données, ainsi que l'application des droits légaux en cas d'infractions liées à l'utilisation des données *(Les défis juridiques du big data),*
— le risque de nuire aux personnes dont les données sont utilisées dans un cadre de recherche sur le big data, y compris des questions telles que la discrimination, l'interférence avec la vie privée et l'utilisation potentiellement abusive des données *(Réglementer la recherche recourant au big data),*
— les règles de l'Organisation mondiale du commerce (OMC) et les accords commerciaux internationaux conclus depuis 2000 ainsi que leur impact sur les réglementations nationales en matière de big data *(Accords commerciaux).*

Propriété des données

Plusieurs concepts relatifs à l'attribution de droits aux propriétaires de données ont été discutés dans le passé. La catégorie normative du droit de propriété se heurte à des problèmes, les dispositions légales ne concernant généralement que les objets physiques, et non les données. Ces dernières sont immatérielles et peuvent être copiées; elles sont également non rivales (elles peuvent être utilisées par de nombreuses personnes en même temps

sans réduire leur accessibilité ou utilité) et généralement non exclusives (disponibles à un très grand nombre); ces deux dernières propriétés constituant d'ailleurs une caractéristique centrale des biens publics. L'idée de propriété des données manque donc de pertinence *(Les défis juridiques du big data, Réglementer la recherche recourant au big data),* et les instances de régulation ne s'appuient habituellement plus sur ce concept.

Il est également difficile d'attribuer aux données un droit de propriété intellectuelle. Ce dernier doit fournir une position juridique exclusive pouvant être exercée contre n'importe qui, mais un bien incorporel doit répondre aux exigences spécifiques des lois applicables. Ces conditions ne sont pas remplies pour les données, comme démontré par la plupart des cas pratiques. Alors qu'un programme informatique est soumis à la protection du droit d'auteur, les données n'atteignent pas le niveau de création spécifique d'un esprit individuel requis pour l'attribution de la propriété intellectuelle. D'autres catégories juridiques, telles que les droits voisins *(un type particulier de droit d'auteur),* les règles de la responsabilité civile (compensations) ou les droits de la personnalité (contrôle de l'utilisation commerciale de son identité) peuvent s'appliquer dans des situations spécifiques. Ces dispositions normatives ne constituent toutefois pas une base adéquate pour un cadre général de propriété des données.

C'est l'accès aux données et le contrôle qu'on peut exercer sur elles (et non leur propriété potentielle) qui déterminent la manière dont elles seront utilisées. Les personnes ou institutions qui détiennent et traitent les données sont en fait en position de propriétaire et ont le pouvoir de décider de leur utilisation, stockage, suppression et transfert. Les régulations se basent donc généralement sur le concept d'accès aux données.

Accès aux données personnelles

Les règles d'accès aux données sont déterminantes pour les personnes directement ou indirectement impliquées dans leur traitement. Certaines entités gestionnaires de données accordent volontairement aux individus l'accès à leurs données personnelles, mais il existe d'autres types de droits d'accès.

— Certains instruments juridiques généraux traitent spécifiquement de l'accès à ses propres données. Par exemple, le droit à la portabilité des données (intégré à la nouvelle loi fédérale sur la protection des données, dont l'entrée en vigueur est prévue en septembre 2023) assura aux personnes le droit de transférer les données les concernant d'un fournisseur de services à un autre.
— Les réglementations actuelles mettant en œuvre les droits d'accès aux données en Suisse se trouvent principalement dans la loi sur les cartels, mais aussi en partie dans la loi sur la concurrence déloyale. Leur application se heurte à de nombreux défis, tels que la délimitation du marché, la définition du pouvoir de marché, ou encore le caractère correct et approprié des données. Les procédures en matière de cartels sont généralement coûteuses et longues, tandis que l'autorité de la concurrence rend souvent sa décision seulement après que la situation concrète a déjà évolué.
— L'accès aux données peut être restreint par des réglementations sectorielles, par exemple dans le domaine de la santé.

Questions sociétales et éthiques liées au big data

Vie privée Les personnes doivent être protégées contre l'accès indu à leurs données privées, leur partage et leur analyse par des tiers.
Accès Les gens doivent pouvoir accéder à leurs données personnelles stockées par les fournisseurs de services et pouvoir les supprimer.
Autonomie Les usagers et usagères doivent pouvoir contrôler quelles données personnelles sont collectées, comment et à quelles fins – au-delà de simplement autoriser ou non l'emploi de certains cookies.
Autonomie sociétale Le développement du big data est dirigé principalement par les entreprises, avec peu de contrôle par la population et les autorités.
Asymétrie du pouvoir La population, les entreprises et les autorités ne sont souvent pas en mesure de changer de fournisseur.
Réglementation Même des algorithmes portant une grande responsabilité sont largement déréglementés, contrairement aux produits médicaux physiques ou aux véhicules. La transparence et les projets de recherche transnationaux sont entravés par les différences entre les réglementations nationales.
Biais Les données ne sont pas neutres: elles reflètent les biais existants dans la société, comme la représentation limitée des minorités ou des corrélations de nature discriminatoire.
Équité Les algorithmes entraînés avec des données biaisées peuvent produire des résultats inéquitables..
Boîte noire Le résultat généré par un algorithme d'apprentissage automatique ne peut souvent pas être expliqué, ce qui nuit à la fiabilité et à la confiance.
Confiance Il doit exister une certaine confiance dans le big data et ainsi dans l'ensemble du processus, des données elles-mêmes (vie privée, accès et partialité) aux algorithmes (fiabilité et équité) et à l'utilisation des applications.
Innovation L'innovation a besoin d'une réglementation claire, stable et équilibrée.
Pratiques commerciales Les applications du big data nécessitent un partage des données, soulevant des questions de confidentialité.

Aperçu des enjeux éthiques, juridiques et sociaux du big data

Le PNR 75 a mis en place la *Task Force ELSI* afin d'étudier les questions éthiques, juridiques et sociales spécifiques au big data. Les messages clés des résultats sont résumés ci-dessous.

Big data et souveraineté numérique

Le concept de souveraineté numérique peut avoir deux significations: l'autonomie d'un État pour réglementer et protéger les données de sa population, ou l'autodétermination des usager·ères quant à la manière dont leurs données personnelles seront utilisées. Ces deux notions entrent en conflit, car le big data constitue un bien sur lequel aussi bien les individus que les États cherchent à exercer un contrôle. Le big data étant immatériel, un État ne peut pas le réglementer en se basant sur une souveraineté s'exerçant sur un espace physique fini. Les gouvernements doivent coopérer afin de réglementer la collecte, le stockage, le partage et le transfert du big data. La protection par l'État de son infrastructure numérique doit s'équilibrer avec l'autonomie de sa population afin d'éviter toute intrusion injustifiée dans la vie privée.

Les défis du consentement éclairé

Il serait peut-être plus transparent d'abandonner, au moins partiellement, l'obligation de demander le consentement éclairé des personnes lorsque l'on utilise leurs données. Ce consentement pourrait être remplacé par des mécanismes assurant la protection des données par l'anonymisation et garantissant une compensation appropriée en cas d'échec de ces mécanismes. Les personnes recevraient des informations claires sur la manière dont leurs données sont utilisées, tandis que les abus seraient soit improbables, soit détectables suffisamment tôt pour éviter tout préjudice important, indépendamment de la manière dont les gens souhaitent que leurs données soient utilisées.

L'importance des lignes directrices

Les réglementations en matière de protection des données sont importantes pour protéger les droits des individus, mais l'hétérogénéité du paysage juridique actuel rend ces réglementations difficiles à comprendre. Il est donc important de créer un pont entre la loi et les questions pratiques. Compte tenu de l'évolution rapide du big data, il peut être préférable de fournir des lignes directrices plutôt que de formuler des lois, en particulier dans le cas du consentement éclairé.

L'éthique du big data dans la santé et la recherche biomédicale

Le big data peut améliorer les services de santé. Il soulève néanmoins des questions sur les mécanismes de surveillance existants (tels que le recours à des commissions d'éthique) et sur la pertinence de l'évaluation de données publiques ou anonymes. La clarification du rôle de ces comités, qui doivent posséder une expertise dans des domaines tels que la science des données, peut contribuer à responsabiliser davantage les scientifiques. On peut envisager la création d'organismes complémentaires, tels que des commissions d'éthique des données.

L'asymétrie entre les entreprises et leur clientèle

Il existe une asymétrie entre les entités qui collectent, stockent et analysent de grands jeux de données, comme les entreprises, et celles dont les données sont collectées, comme la clientèle. Étant donné le caractère quasi inévitable de ce fossé dans la société actuelle, il serait peut-être préférable de ne pas tenter de l'éliminer pour privilégier la protection des personnes et entités menacées. Des valeurs telles que la non-malfaisance et l'équité seraient alors considérées comme plus importantes que l'autonomie et la vie privée.

Risques de discrimination

Les applications d'apprentissage automatique peuvent introduire des biais, discriminations et injustices. Combattre ces problèmes comprend la prise de conscience que les biais ont une composante normative aussi bien que technique, et que la politique doit exiger l'équité dans la conception d'algorithmes touchant à des enjeux importants. La discrimination ne peut pas être évitée en interdisant le traitement des informations relatives aux groupes, puisque ces informations sont nécessaires pour évaluer l'équité des algorithmes. Différentes normes statistiques peuvent être utilisées pour estimer si un algorithme est biaisé, et les mesures de lutte contre la discrimination doivent permettre d'utiliser différentes normes suivant les cas concrets.

ELSI-Task-Force for the National Research Programme Big Data,
Markus Christen (Université de Zurich)
Ethical, Legal and Social Issues of Big Data – a Comprehensive Overview,
Eleonora Viganò (Ed.) (PNR 75, 2022)

Internationalité

Les données personnelles constituent des biens économiques de plus en plus précieux. Leur transfert entre différentes juridictions nécessite des accords transfrontaliers et des réglementations harmonisées. Cette évaluation se heurte à la multitude de lois nationales encadrant la collecte et l'utilisation des données. Celles-ci sont également influencées par les accords commerciaux internationaux bi- ou multilatéraux *(Accords commerciaux).*

4.2
Vie privée et souveraineté numérique

La confidentialité et la sécurité des données font partie des questions politiques les plus débattues au sujet du big data. Les lois sur la protection de la vie privée régissent la collecte et le traitement des données à caractère personnel, mais il est souvent difficile de les distinguer des données non personnelles. En particulier, des éléments de données impersonnels peuvent révéler des informations personnelles une fois combinés par le recoupement de plusieurs bases de données (processus de désanonymisation). Des analyses peuvent générer des informations nouvelles grâce aux corrélations complexes existant dans les très grands jeux de données. D'une manière générale, les avancées technologiques suscitent des doutes quant à l'adéquation au big data des principes traditionnels de protection des données.

Plusieurs projets du PNR 75 ont étudié les questions de vie privée et de souveraineté numérique dans des contextes concrets. Ils ont

— analysé les questions éthiques, juridiques et sociétales soulevées par l'utilisation du big data par des compagnies d'assurance, et formulé des recommandations *(Big data dans l'assurance),*
— fait un état des lieux des questions éthiques liées à l'utilisation du big data dans le domaine de la santé, évalué les mécanismes de monitoring existants et élaboré un cadre éthique ainsi que des recommandations politiques *(Big data dans la santé),*
— analysé la manière dont les départements des ressources humaines exploitent le big data et évalué l'impact sur la confiance du personnel *(Big data dans les ressources humaines).*

Protection des données

La question de savoir quelles personnes ou institutions sont en mesure de décider de l'utilisation des données est fondamentale, un contexte dans lequel le concept de souveraineté numérique est souvent utilisé. Ce dernier est néanmoins ambigu: la souveraineté numérique peut signifier l'autodétermination informationnelle ou le contrôle personnel sur ses données, mais également le principe que les données collectées dans un pays ne devraient être soumises qu'à ses propres lois et non à celles d'autres États. Une analyse approfondie des concepts de souveraineté est importante pour les réglementations futures (voir le rapport du PNR *75 Ethical, legal and social Issues of big data – a comprehensive overview).*

La plupart des lois sur la protection des données impliquent le principe de

minimisation des données, à savoir la limitation de leur collecte, traitement et utilisation au strict nécessaire pour atteindre un objectif spécifique. Il vise à minimiser l'accumulation inconsidérée d'informations pour une utilisation additionnelle prévue ou imaginée plus tard. Cependant, les grands jeux de données sont généralement conçus pour trouver des corrélations et générer de nouvelles informations. Le fait que de tels développements ne peuvent pas être prévus peut être utilisé comme argument pour justifier la collecte et le traitement des données, puisque le consentement, par sa nature même, ne peut être donné que pour une procédure dont les résultats sont connus. Une motion parlementaire appelant à élaborer une loi réglant l'utilisation secondaire des données a été déposée le 22 août 2022 et adoptée au Conseil des États le 14 décembre 2022[10].

Le principe du consentement éclairé (selon lequel une personne doit donner son accord explicite et informé à la collecte, au traitement, à l'utilisation et au transfert de ses données) n'est pas facilement applicable en pratique. La plupart des gens ne sont pas en mesure d'évaluer le contexte contractuel et les risques liés à l'analyse de leurs données. Les principes traditionnels de protection des données étant mal adaptés au big data, il est urgent de les repenser.

Modèles de gouvernance et technologies pour renforcer la vie privée

Les concepts prometteurs de protection de la vie privée devraient inclure des éléments de gouvernance tels que:
— une analyse adéquate des risques pertinents et potentiels pour la protection des données,
— l'établissement d'une stratégie pour se conformer aux principes de la protection des données,
— l'application des politiques de protection des données,
— la mise en place de procédures pour remédier aux défaillances de la protection des données.

La vie privée peut être protégée non seulement par la réglementation, mais aussi par des technologies telles que le chiffrement de bout en bout, ou des techniques garantissant l'anonymat. Des infrastructures dédiées (comme l'écosystème Polybox de l'ETH Zurich) peuvent garantir que les données sont partagées de manière sécurisée tout en restant faciles à trouver, accessibles, interopérables et réutilisables *(Big data dans la santé).*

4.3 Équité, non-discrimination et inclusion

Les questions d'équité, de non-discrimination et d'inclusion dans le domaine technologique suscitent une grande attention. Comme pour la vie privée, les risques sont particulièrement élevés pour les segments sensibles de la société. Les implications éthiques du big data doivent être comprises à la fois de manière générale et dans le contexte d'applications spécifiques, telles que les assurances, le travail ou la santé.

[10] Élaboration d'une loi-cadre sur la réutilisation des données, Motion 22.3890 (2022)

Équité des données dans des applications concrètes

Assurances

Des données détaillées sur les personnes assurées permettent une tarification plus personnalisée. Cela peut menacer la mutualisation des risques et ainsi le principe de solidarité à la base du concept d'assurance *(Big data dans l'assurance).* Ces développements soulèvent une question provocante: le big data rend-il les gens inassurables? On ne peut exclure que certaines personnes se voient exclues du marché de l'assurance parce que considérées comme trop risquées. Les principes éthiques jouent ici un rôle important, de même que les réglementations légales. Des codes de conduite spécifiques ont pu être élaborés pour les compagnies d'assurance utilisant le big data (voir «Eclairages sur le secteur des assurances», p. 65).

Relations de travail

Les employeurs doivent adhérer aux principes de non-discrimination et d'équité, et atténuer les effets des algorithmes discriminatoires. L'intégrité de l'être humain est particulièrement importante dans les relations entre l'employeur et le personnel, ainsi que pour établir et maintenir la confiance sur le lieu de travail *(Big data dans les ressources humaines).*

Le contrôle des activités du personnel ne doit pas compromettre l'intégrité ou la réputation morale. Il est nécessaire de communiquer avec lui de manière transparente et de l'impliquer dans l'organisation de l'environnement de travail. Les ressources humaines constituent un domaine particulièrement sensible, leurs activités pouvant violer les droits des employées et employés ou nuire à leur réputation.

Santé

Les concepts de consentement éclairé, de risque minimal et de respect de la vie privée sont particulièrement importants en matière de santé *(Big data dans la santé, Réglementer la recherche recourant au big data).* Les lignes directrices et les outils doivent non seulement prescrire la conformité des données, mais également garantir leur utilisation éthique.

La recherche sur le big data en biomédecine implique la participation d'une variété d'acteurs – scientifiques et spécialistes en informatique, secteur privé, organisations professionnelles, responsables, population – ayant des intérêts divergents ou concurrents ainsi que des attentes différentes en matière de santé. Les mécanismes de contrôle existants doivent être renforcés ou réformés à différents niveaux et doivent s'adapter aux nouvelles technologies.

4.4 Production et gestion des connaissances

Les concepteurs et conceptrices d'applications du big data jouent un rôle central dans l'élaboration et le maintien de lignes directrices et de pratiques éthiques, tant générales que spécifiques à un domaine. En plus du projet *Réglementer la recherche recourant au big data*, deux autres équipes du PNR 75 ont étudié l'impact du big data sur les métiers du savoir.

— Une recherche a évalué avec une

approche épistémologique la manière dont le big data est utilisé dans les simulations informatiques pour la recherche scientifique, notamment en science du climat *(Incertitudes dans le big data).*

— Un projet a utilisé des outils de la sociologie et de l'ethnologie pour étudier le rôle du big data dans la sociologie ainsi que dans la science et le journalisme des données *(Le big data en pratique).*

Transdisciplinarité et nouvelles compétences

Les questions sociétales, juridiques et éthiques du big data s'étendent également aux sciences naturelles utilisant des jeux de données de grande taille, exigeant des approches interdisciplinaires.

L'inclusion des personnes dans les affaires qui les concernent doit être améliorée. L'inclusivité fait partie d'un changement fondamental dans la production des connaissances scientifiques et dans la compréhension des structures et mécanismes impliqués dans le maintien et l'évolution des domaines de connaissances. Les compétences en matière de données, d'images et de pensée computationnelle sont essentielles pour une participation étendue à la production de connaissances *(Le big data en pratique).*

L'importance du contexte

La variété de la recherche sur le big data implique que les questions éthiques ne peuvent pas être traitées à travers une réglementation globale et unique. Il faut mettre l'accent sur le contexte et la délibération plutôt que sur une standardisation inflexible *(Réglementer la recherche recourant au big data).* L'interdépendance des questions juridiques, sociales et éthiques ainsi que l'intégration de cultures différentes appellent une recherche interdisciplinaire et internationale.

Comprendre les connaissances sur le big data

La science emploie le big data de diverses manières, en synthétisant des données, en produisant des prédictions et en découvrant des relations. Les résultats de ces analyses s'accompagnent de nombreuses incertitudes; celles-ci doivent être évaluées, quantifiées et communiquées correctement afin que les résultats soient fiables et utilisables *(Incertitudes dans le big data).*

Eclairages sur le secteur des assurances

Le projet du PNR 75 *Big data dans l'assurance* a analysé les questions éthiques et juridiques soulevées par l'utilisation du big data dans les assurances personnelles. Il a formulé plusieurs suggestions concrètes à l'intention de l'industrie et des instances de réglementation.

→ Déterminer si et comment les compagnies d'assurances doivent être autorisées à personnaliser leurs contrats sur la base du big data ne devrait pas être fait de manière indirecte par l'application de lois générales de la protection des données et de lois antidiscriminatoires.

→ Le régulateur doit surveiller et anticiper en permanence l'utilisation du big data pour la personnalisation des contrats d'assurance. La législation sur les assurances doit être adaptée afin d'interdire les formes de personnalisation non souhaitées, ou de définir le type de personnalisation autorisée.

→ Les compagnies d'assurance doivent éviter d'utiliser des données sans rapport avec le risque assuré afin de préserver la confiance de la clientèle.

→ Elles doivent être conscientes que l'utilisation discriminatoire de l'apprentissage automatique peut affecter la prédiction, la tarification et la détection des fraudes.

→ Elles doivent montrer comment elles protègent la vie privée, l'équité et la solidarité lorsqu'elles utilisent le big data.

→ Elles doivent adapter leurs principes éthiques afin d'assumer leur responsabilité lorsqu'elles gèrent les questions soulevées par la numérisation du secteur.

Big data ethics recommendations for the insurance industry, PNR 75 (2019)

4.5 Défis et messages clés

Les défis des recherches sur les questions sociétales, juridiques et éthiques du big data

La recherche sur le big data dans la société fait face à l'évolution rapide des technologies et des questions réglementaires, comme la notion – en perte de vitesse – de propriété des données *(Les défis juridiques du big data).* Cette situation exige de la flexibilité dans les programmes de recherche et les financements.

Ces travaux – comme ceux sur les applications et les infrastructures – se confrontent à la difficulté de pouvoir accéder à des données, même universitaires, de qualité *(Le big data en pratique, Big data dans l'assurance, Incertitudes dans le big data).* Cela a été possible dans le domaine juridique *(Accords commerciaux).*

L'évaluation éthique et juridique des systèmes utilisant le big data est essentiellement interdisciplinaire, avec des spécialistes des données, de la modélisation et des aspects juridiques. Cela rend difficile l'adoption de terminologies, méthodologies et concepts communs *(Les défis juridiques du big data, Incertitudes dans le big data).*

Messages clés

Le PNR 75 n'a abordé que partiellement un sous-ensemble de toutes les questions sociétales, juridiques et éthiques liées au big data, mais ses recherches mettent en lumière certains aspects génériques.

— Les sphères publiques et privées doivent être plus transparentes quant à l'utilisation du big data.
— Des recherches universitaires doivent examiner l'impact potentiel du big data sur la démocratie. Il s'agit notamment du rôle de plus en plus important de l'analytique dans les réseaux sociaux, qui accélère la diffusion de fausses informations et les manipulations. De tels développements ne devraient pas être laissés à la discrétion de sociétés commerciales.
— Les aspects sociétaux – non pas seulement technologiques – du big data doivent être inclus dans les délibérations.

Aborder les questions sociétales

Les nouvelles technologies bouleversent la vie personnelle et professionnelle, l'agrégation et l'analyse des données ayant des effets majeurs sur des secteurs tels que la santé ou le travail. Il est important de mener des analyses contextualisées, d'anticiper des conséquences sociales, et d'élaborer des lignes directrices pratiques adaptées aux différents environnements *(Big data dans la santé, Big data dans les ressources humaines, Big data dans l'assurance, Réglementer la recherche recourant au big data).*

Élaborer des règlements appropriés

L'ordre juridique doit mettre en place un cadre normatif proportionné. Le concept de propriété étant mal adapté aux données, les instances législatives pourraient formuler le concept alternatif de détenteur de droits sur les données, centré sur le contrôle et l'accès aux données. De nouveaux instruments normatifs seront nécessaires dans des domaines tels que la blockchain. Il est important de trouver un niveau de réglementation équilibré.

Développer des directives éthiques

Des lignes directrices éthiques sont nécessaires. L'inclusion et l'équité ne sont pas bien prescrites par la loi et de nombreux principes de non-discrimination ne sont pas couverts par les constitutions.

Ces lignes directrices doivent être formulées à travers des processus concrets, adaptés à la situation et impliquant les multiples parties prenantes afin d'augmenter les chances qu'elles soient suivies *(Big data dans la santé, Big data dans l'assurance, Réglementer la recherche recourant au big data).*

4.6 Les projets de recherche sur les aspects sociétaux, juridiques et éthiques du big data

Questions sociétales

Le big data en pratique: sociologie, sciences des données et journalisme

Ce projet a analysé de manière sociologique et ethnographique la façon dont le big data est compris, enseigné ou utilisé en sociologie, en sciences des données et en journalisme de données. Il a analysé plus de 750 programmes d'études d'universités allemandes et identifié quatre cultures dans l'enseignement des méthodes sociologiques. Pour comprendre comment les sciences des données sont perçues et orientées en Suisse, l'équipe a analysé 4 300 offres d'emploi en ligne, 34 documents de politique et de stratégie ainsi que 40 nouveaux programmes d'études dans des établissements d'enseignement supérieur. Les résultats montrent que les sciences des données partagent un ensemble distinct de méthodes, d'outils et de pratiques, qui transcendent les frontières entre disciplines tout en se situant sur la ligne de front des tensions entre l'informatique et les statistiques.
Les études ethnographiques montrent que le journalisme de données nécessite des cultures épistémologiques et professionnelles spécifiques à coordonner avec les pratiques journalistiques existantes. Dans l'ensemble, les résultats soulignent que le big data requiert des compétences qui dépassent les frontières disciplinaires.

—

Facing big data: methods and skills needed for a 21st century sociology
Sophie Mützel (Université de Lucerne)

Incertitudes dans le big data: le cas des simulations climatiques

Ce projet a étudié de manière épistémologique les simulations informatiques développées pour la recherche climatique et basées sur le big data. Il montre que la science du climat utilise des données de plus en plus variées, comme les réseaux sociaux ou les recherches sur le Web, ce qui rend les calculs plus efficaces et permet de découvrir de nouvelles relations au sein des modèles. Toutefois, des incertitudes découlent des variations de la qualité des données et d'une compréhension incomplète de leur rôle. Les résultats de ces recherches étant régulièrement comparés aux nouvelles observations, le projet souligne l'importance de combiner les méthodes du big data avec les approches scientifiques

traditionnelles, telles que la compréhension des processus. Il appelle à des collaborations transdisciplinaires entre les spécialistes et les scientifiques des données, notamment pour évaluer les incertitudes.
L'équipe a analysé deux études de cas sur le climat: la modélisation et la prévision à haute résolution des îlots de chaleur urbains, et la dépendance de la température urbaine vis-à-vis de la végétation et d'autres facteurs. Dans le premier cas, la recherche différencie les incertitudes basées sur les limitations de l'algorithme de prédiction lui-même et celles dérivées des données d'apprentissage finies. Ces dernières affectent davantage des objets spécifiques que des catégories génériques.
—
Combining theory with big data? The case of uncertainty in prediction of trends in extreme weather and impacts
Reto Knutti (ETH Zurich)

Questions juridiques

Les défis juridiques du big data

Ce projet a exploré plusieurs questions juridiques soulevées par le big data. Il a constaté que la propriété des données ne peut pas être actuellement définie et que les marchés de données personnelles, tels que la publicité ciblée, opèrent essentiellement en dehors de réglementations légales. Les scientifiques ont évalué les alternatives possibles aux concepts normatifs existants en matière de droits des données.
Une analyse juridique indique que le fait d'accepter de partager des données personnelles avec un prestataire de services ne constitue qu'une renonciation partielle au droit à la confidentialité des données. Une autre étude souligne le conflit entre le mandat des autorités de protéger la société et le droit fondamental de la population ne pas être indûment surveillée. Par exemple, les données enregistrées par les véhicules permettent la reconstitution médico-légale d'accidents, mais peuvent violer des principes juridiques fondamentaux tels que le privilège contre l'auto-incrimination. Ces travaux abordent de nombreuses questions non résolues sur les droits des données, telles que la clarification de l'identité des victimes de crimes liés aux données, ou le fait que les personnes vivants en Suisse et en Allemagne peinent actuellement à faire valoir des intérêts légitimes dans le cadre de procédures pénales.
—
Legal challenges in big data. Allocating benefits. Averting risks
Sabine Gless (Université de Bâle)

Accords commerciaux: impacts sur le droit national

Ce projet a examiné des centaines d'accords commerciaux conclus au cours des deux dernières décennies afin de déterminer leur pertinence pour l'économie des données. Il a analysé les normes existantes et les dispositions toujours plus nombreuses sur le big data, telles que celles sur le commerce électronique, la protection des données et les données gouvernementales ouvertes.
L'équipe a analysé l'interaction entre les engagements internationaux et les politiques nationales. Elle a créé une base de données accessible librement, et utilisée entre autres par l'OCDE, le ministère britannique du commerce ou le WEF. Les travaux montrent que la réglementation des données exige une plus grande coopération internationale, les politiques dans des domaines tels que la protection des données et la sécurité nationale variant considérablement d'un pays à l'autre. Les résultats soulignent l'importance croissante du droit commercial et suggèrent des moyens de mieux utiliser ces lois dans une économie axée sur les données.

Le projet fait valoir que la Suisse pourrait jouer un rôle important en tant que pays innovant et connecté internationalement.

—

The governance of big data in trade agreements: design, diffusion and implications
Mira Burri (Université de Lucerne)

Réglementer la recherche recourant au big data

Ce projet a analysé les nombreuses questions éthiques soulevées par la recherche menée avec du big data, en particulier les problèmes de discrimination, de violation de la vie privée et d'utilisation abusive des données. Si les processus éthiques et réglementaires sont bien connus dans la recherche sur les êtres humains telle que la médecine ou la psychologie, l'utilisation de données anonymes dans le domaine scientifique soulève de nombreuses questions encore mal connues.

Les résultats constatent que la nature variée des travaux scientifiques entrave la mise en place d'un cadre global, harmonisé et normalisé pour la recherche sur le big data. Ils suggèrent que la réglementation devrait plutôt impliquer des décisions fondées sur le contexte, la délibération éthique et l'analyse des compromis, un processus potentiellement continu. Les commissions d'éthique de la recherche devraient être composées de professionnel·les, notamment de spécialistes du big data, et évaluer l'éthique des travaux de recherche tout au long de leur cycle de vie. L'évaluation éthique doit également porter sur les recherches menées par les entreprises privées, qui collaborent de plus en plus avec le monde universitaire. Les lignes directrices, les procédures et les codes de conduite devront être révisés régulièrement pour suivre l'évolution des technologies et des réglementations, comme le développement d'algorithmes questionnant l'efficience de l'anonymisation moins efficace ou encore la mise en œuvre de la réglementation européenne RGPD.

—

Ethical and legal regulation of big data research – Towards a sensible and efficient use of electronic health records and social media data
Bernice Simone Elger (Université de Bâle)

Questions éthiques

Big data dans la santé: un cadre éthique

Ce projet s'est penché sur l'éthique du big data dans le domaine de la santé et a évalué si ces questions peuvent être traitées efficacement par les mécanismes de surveillance existants tels que les commissions d'éthique de la recherche. Il a élaboré un cadre éthique et des recommandations politiques pour soutenir ces mécanismes, et proposé une boîte à outils utilisable par les scientifiques et les comités d'éthique.

Les résultats montrent que les scientifiques et les personnes développant des applications ont tendance à considérer que l'éthique du big data se réduit principalement à la conformité avec les réglementations existantes en matière de protection des données, ignorant souvent des questions telles que la responsabilité de la recherche, l'équité, l'autonomie individuelle ou le préjudice causé aux groupes de population. En Suisse, les commissions d'éthique admettent un manque d'expertise en matière de big data et expriment le besoin de formations adéquates. Le projet conclut qu'elles doivent modifier leurs règlements et leurs procédures afin de pouvoir superviser adéquatement la recherche biomédicale. L'équipe a élaboré une boîte à outils pour d'aider les commissions d'éthique à évaluer leur état de préparation à la recherche sur

le big data et une check-list pour faciliter l'examen des projets.

—

BEHALF – Big-data-ethics-health framework
Effy Vayena (ETH Zurich)

Big data dans l'assurance

Une équipe interdisciplinaire a étudié les questions éthiques, juridiques et sociétales du big data dans l'assurance privée, formulant des recommandations à l'intention des compagnies d'assurance disposant d'importants jeux de données sur leur clientèle (voir «Eclairages sur le secteur des assurances», p. XX).

Le projet a révélé que le respect de la vie privée est un sujet moins problématique que l'analyse prédictive, qui quantifie les risques et la propension des personnes assurées à payer leurs primes ou à commettre des fraudes. La granularité croissante de l'évaluation des risques rend plus probable la discrimination à l'encontre d'un comportement spécifique (comme le fait de ne pas pratiquer de sport), tout en discriminant indirectement les groupes de population les plus enclins à ce comportement. L'analyse de la législation montre que la Suisse est plutôt libérale et respecte le principe de la liberté contractuelle bien plus que, par exemple, la Californie.

—

Between solidarity and personalization – Dealing with ethical and legal big data challenges in the insurance industry
Markus Christen (Université de Zurich)

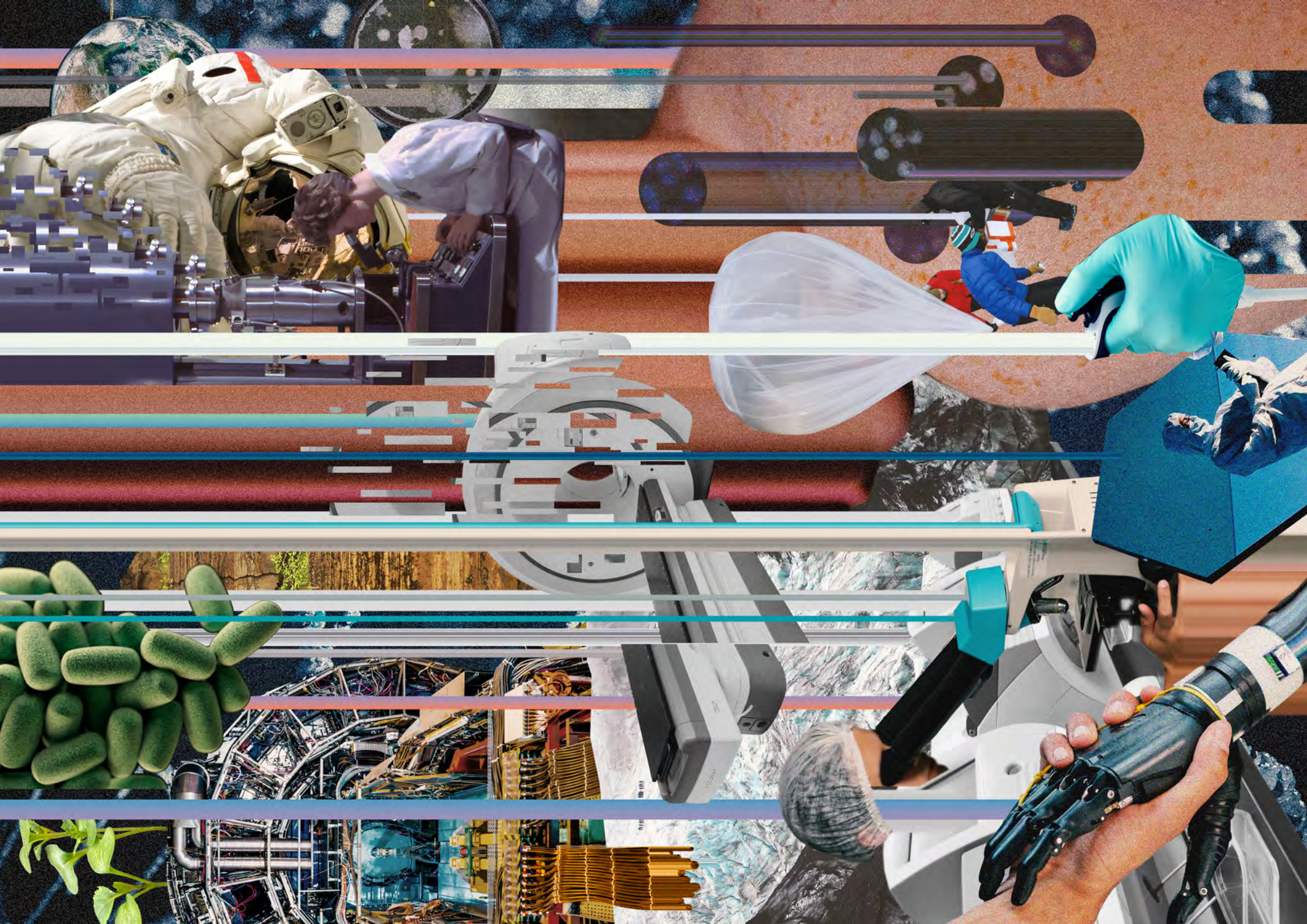
Big data dans les ressources humaines

Ce projet a analysé comment les départements de ressources humaines, notamment en Suisse, ont intégré le big data, et comment cela affecte la confiance entre personnel et employeur. Il a mis en évidence de grandes variations dans la transparence des entreprises en matière de collecte de données et de responsabilisation des employé·es, avec des niveaux différents de confiance. Pour renforcer cette dernière, il convient de protéger l'autonomie et la capacité d'action du personnel tout en préservant la vie privée, la transparence et le contrôle. Le personnel doit être associé aux décisions stratégiques et ses doléances doivent être écoutées.

Les résultats montrent que les lois suisses, en particulier, sont mal équipées pour faire face à des systèmes algorithmiques discriminatoires sur le lieu de travail, ce qui souligne la nécessité de garanties juridiques appropriées. L'équipe a développé une boîte à outils pour familiariser les responsables des ressources humaines avec la technologie du big data et permettre d'évaluer les questions juridiques, éthiques et professionnelles.

—

Big data or big brother? Big data HR control practices and employee trust
Antoinette Weibel (Université de Saint-Gall)



5.
Réflexions et perspectives

La société doit anticiper les bouleversements que les applications du big data et de l'apprentissage automatique peuvent entraîner. Une vue d'ensemble des opportunités et défis les plus importants est présentée ci-dessous.

Un apport central du Programme national de recherche «Big Data» (PNR 75) est d'avoir renforcé les compétences disponibles en Suisse nécessaires pour aborder les questions de technologie, d'applications et d'aspects sociétaux du big data. Le PNR 75 a fait progresser les technologies qui sous-tendent l'infrastructure du big data et a réuni des spécialistes en science des données et des experts de domaines concernés pour créer et mettre en œuvre des applications spécifiques. Il a également sensibilisé aux défis sociétaux qui accompagnent la production et l'analyse de données à grande échelle, et a contribué à développer la culture du big data nécessaire pour en profiter de manière responsable.

Les 37 projets financés par le programme n'ont couvert qu'une partie du big data, un domaine en pleine expansion. Ce chapitre va au-delà et donne une vision plus globale des opportunités et des risques liés au big data, en particulier de ceux qui pourraient prendre de l'importance dans les années à venir. L'analyse qui suit se fonde sur les connaissances acquises dans le cadre de la recherche du PNR 75 et sur les idées collectives des membres du Comité de direction du programme. Elle aborde les perspectives d'une plus grande utilisation du big data dans l'industrie et le secteur public, et discute les questions de durabilité, vie privée et responsabilité.
.

5.1 Un impact croissant

De nombreuses applications du big data continueront à être développées et déployées dans les années à venir. De nouveaux secteurs privés – au-delà du commerce électronique – et des administrations publiques s'efforceront de devenir «data-ready» afin de gagner en compétitivité par le développement de nouvelles capacités et par la réduction des coûts. Comme montré par les projets de recherche du PNR 75, le développement d'applications exige la bonne combinaison de compétences dans plusieurs domaines. Cela nécessite une stratégie solide en matière de données, notamment des approches «privacy-by-design», un savoir-faire analytique chez les spécialistes du secteur et des connaissances pointues de la main-d'œuvre. Un ingrédient crucial est de pouvoir trouver des spécialistes des données qui comprennent le domaine d'application concerné ainsi que des spécialistes du domaine familiarisés avec la science des données. Cela souligne l'importance de doter les nouvelles générations – et les plus anciennes – des connaissances et outils nécessaires pour s'attaquer aux applications du big data (voir Conclusion 1 dans le chapitre 6).

Une sélection de domaines susceptibles d'être fortement affectés par les applications du big data est présentée ci-dessous.

Production: amélioration du rendement et optimisation de la maintenance

De nombreux produits manufacturés intègrent des capteurs connectés à l'Internet des objets. Ils peuvent envoyer des informations en temps réel sur leurs performances, permettant aux fabricants d'identifier les composants à remplacer ou à améliorer, ou de renforcer la satisfaction des clients et la sécurité.

Dans l'agriculture, des systèmes robotisés autonomes utilisent la reconnaissance d'images pour éliminer les mauvaises herbes, détecter les maladies et les parasites, récolter les fruits, appliquer des engrais localement et surveiller les champs à l'aide de drones. Ces robots pourraient contribuer à réduire les pénuries de main-d'œuvre, à diminuer l'emploi d'engrais et à éviter celui de pesticides[11].

Gouvernement: améliorer les infrastructures et soutenir la transition énergétique

Les gouvernements peuvent utiliser le big data pour mettre en œuvre des politiques basées sur les faits pour l'allocation des ressources, la planification stratégique ou encore la surveillance des infrastructures publiques (Conclusion 5). L'analyse du big data peut améliorer la planification des transports *(Gestion des transports),* soutenir la planification, construction et exploitation des services publics (eau, électricité, éclairage, etc.) et assurer la surveillance de l'environnement *(Érosion des sols, Détection d'inondation).* Des analyses sophistiquées contribueront à réduire notre empreinte carbone en garantissant la flexibilité de l'approvisionnement, du stockage et de la distribution de l'énergie, et permettront en particulier aux réseaux électriques de gérer les sources d'énergie renouvelable décentralisées et intermittentes telles que les panneaux solaires ou les éoliennes *(Potentiel des énergies renouvelables).*

Services: automatisation dans la finance et la cybersécurité

Les institutions financières peuvent utiliser l'analyse des transactions en temps réel et les prévisions du marché pour un trading automatisé rapide, qui nécessite néanmoins des infrastructures efficaces *(Algorithmes de prédiction rapide, Exploration de graphes).* La quantification des risques individuels permet aux compagnies d'assurance d'optimiser leurs polices, mais menace potentiellement le principe de solidarité qui sous-tend les assurances *(Big data dans l'assurance).* Les dispositifs de suivi embarqués dans les véhicules pourraient récompenser les comportements atténuant les risques, mettant ainsi l'accent sur la prévention plutôt que sur la protection.

L'analytique peut contribuer à prévenir les cyberattaques en identifiant des anomalies dans les transferts de données en temps réel et en bloquant automatiquement les menaces *(Flux de données).* La reconnaissance d'images peut être utilisée pour détecter automatiquement les atteintes à la sécurité physique et d'autres irrégularités.

Santé: assister le personnel médical et personnaliser la médecine

L'apprentissage automatique est déjà utilisé, par exemple, pour identifier des anomalies dans les images cliniques[12] et pourrait considérablement contribuer à améliorer les soins de santé (voir Conclusion 4). Les nouvelles technologies devraient permettre des progrès majeurs en matière de prévention, de diagnostic et de thérapies ciblées en rassemblant d'énormes jeux de

[11] Voir par exemple https://www.agricultural-robotics.com pour un aperçu.
[12] Voir par exemple https://grand-challenge.org/aiforradiology pour un aperçu.

données provenant de tests de laboratoire, de dossiers médicaux et de la génétique. En particulier, le traitement avancé du langage naturel *(Modèles de langage)* permet l'extraction et l'interprétation automatiques d'informations à partir de textes non structurés dans les dossiers médicaux. Intégrer les flux de données provenant de divers dispositifs cliniques en temps réel peut faciliter la surveillance de l'état de santé des patient·es et la détection de cas d'urgences *(Soins intensifs).*

L'utilisation du big data pour les applications médicales nécessite des infrastructures dédiées. Des méthodes innovantes sont nécessaires pour générer des résultats fiables à partir de petits sous-ensembles de données médicales, une seule personne pouvant en générer plusieurs téraoctets. Pour les données génomiques, cela peut se faire par un prétraitement adéquat *(Genetic big data).*

Commerce électronique et divertissement: participation de la clientèle et créations synthétiques

La collecte, l'analyse et l'exploitation d'informations sur les consommateurs et consommatrices joueront probablement un rôle croissant dans le commerce électronique. Les entreprises en ligne utilisent déjà des recommandations personnalisées et des prévisions de tendances, et de nouvelles applications fondées sur les données pourraient intégrer les attentes de la clientèle dans le processus même de la conception des produits.

Les modèles de langage s'améliorent très rapidement. Ils permettent de mieux comprendre le sens, l'intention et le contexte de textes, d'en extraire des informations pertinentes et de générer des rapports synthétiques ou des conversations par des «chatbots». Des algorithmes peuvent produire de la musique en s'inspirant du style de compositeurs et compositrices. Des ordinateurs génèrent des images et des vidéos synthétiques convaincantes à partir de descriptions textuelles; on s'attend à ce que les logiciels soient bientôt capables de générer des films avec des personnes et des décors à l'aspect naturel que l'on ne peut distinguer de séquences filmées. Ces systèmes peuvent compléter ou remplacer les médias et les produits de divertissement actuels. Cependant, ils posent des défis majeurs pour la propriété intellectuelle[13] ainsi que pour la démocratie avec la production accélérée de canulars réalistes incluant images, sons et vidéos.

Recherche ouverte: accélérer les découvertes

Les scientifiques sont de plus en plus nombreux à mettre gratuitement à disposition les données de leurs recherches afin d'accélérer les découvertes et d'améliorer la reproductibilité (Conclusion 6). Mais comme tout autre dépôt de données, il est nécessaire de se conformer à certaines normes, telles que les «principes FAIR» (facilité de recherche, accessibilité, interopérabilité et réutilisation). Il s'agit de métadonnées standardisées, lisibles par ordinateur, contenant les explications et les descriptions nécessaires. Il s'agit d'un nouveau paradigme auquel le monde universitaire doit s'adapter *(Big data: open data and legal strings).*

---

[13] The lawsuit that could rewrite the rules of AI copyright, The Verge (2022)

Assurer la représentativité de la recherche sur le big data

Le PNR 75 a mené un programme visant à renforcer la communauté des femmes scientifiques actives dans la recherche sur le big data en Suisse. Seulement 22% des diplômés dans les matières techniques sont des femmes, soit l'un des taux les plus bas parmi les pays de l'OCDE[15]. Cette situation intensifie le manque de spécialistes, génère des sujets de recherche peu représentatifs de la société, et cadre les questions de manière biaisée. Pour les expertes, cela se traduit par un manque d'inspiration, d'encouragement et de soutien, ainsi que par un risque accru de harcèlement et de discrimination.

L'activité transversale du PNR 75 Women in big data s'est penché sur les sciences techniques aussi bien que sociales. Elle a lancé diverses actions pour promouvoir les carrières, comme l'aide à la création de réseaux, l'élimination des obstacles à l'excellence technique et l'encouragement des échanges interdisciplinaires sur les problèmes de genre dans le domaine du big data.

Women in big data
Lydia Yiyu Chen (Delft University of Technology)

5.2 Diminuer l'empreinte des infrastructures de données

Si le big data va certainement jouer un rôle important dans la lutte contre le changement climatique et dans la réduction de notre empreinte carbone, il contribue également au problème. Le stockage et le traitement de grands jeux de données nécessitent beaucoup d'énergie: 3,6% de la consommation électrique totale de la Suisse en 2019 était due aux centres de données, soit une hausse de 30% en 6 ans[14].

La gestion du big data ne se limite pas à la collecte et au stockage; les données doivent aussi être protégées contre les accès non autorisés, la corruption et la perte. Cela nécessite un contrôle d'accès, des protocoles de sauvegarde et des procédures pour corriger les données endommagées, incomplètes ou inexactes. Les bases de données doivent être préservées en étant continuellement adaptées aux nouvelles normes de stockage, de compression et d'analyse. Cela nécessite le travail de spécialistes des données et des domaines, et augmente les coûts des applications du big data. L'intelligence artificielle dite frugale vise à réduire la consommation d'énergie, par exemple en étant capable de travailler avec des jeux de données plus petits et grâce à des données d'apprentissage synthétiques qui économisent les ressources. Ce domaine nouveau et en pleine expansion appelle des efforts de recherche supplémentaires *(Coresets).*

5.3 Vie privée: trouver le bon équilibre

De nombreuses applications du big data, telles que celles utilisées dans

[14] Cela correspond à 2,1 TWh, soit un quart de la production de la centrale nucléaire de Gösgen. Voir «La consommation d'électricité des centres de calcul en Suisse continue d'augmenter», Office fédéral de l'énergie (2021).
[15] Suisse: 22% des diplômés dans les disciplines MINT (mathématiques, informatique, sciences naturelles et techniques) sont des femmes, contre 26% en Allemagne, 32% en France et 40% en Italie. Les sciences informatiques ont un taux encore plus bas, de seulement 16%. Geschlechterunterschiede in MINT-Studiengängen: Eine deskriptive Analyse, KOF, ETH Zürich (2020)

les domaines de la finance, de l'ingénierie ou de la surveillance de l'environnement, ne soulèvent pas de nouvelles questions sur la vie privée, car elles n'utilisent pas d'informations personnelles. Mais de nombreuses autres applications le font, et la quantité toujours croissante de données qu'elles collectent sur les individus soulève des problèmes éthiques et juridiques importants. Les gens ont généralement une connaissance limitée des données collectées et de qui peut y avoir accès et à quelles fins. Le fait que les fournisseurs de services en ligne contrôlent ces éléments a donné naissance aux termes «fracture numérique» et «asymétrie numérique».

Bien que les fournisseurs soient actuellement tenus d'informer leur clientèle et de lui demander son consentement lorsqu'ils collectent des données, ces mesures ne suffisent pas à protéger la vie privée, car la plupart des gens donnent leur accord de manière automatique et sans en connaître les conséquences. Le principal problème est que c'est aux usagers de comprendre les implications de leur consentement, même s'ils ne tirent aucun avantage immédiat de la collecte de données. Les autorités devront décider dans quelle mesure il convient de réglementer cette pratique (Conclusion 8).

L'anonymat complet est une illusion

Jusqu'à récemment, il était considéré comme sûr de partager des données contenant des informations sur des individus une fois qu'elles avaient été rendues anonymes, en supprimant les informations susceptibles d'identifier directement les gens, telles que leur nom, date de naissance et adresse. Il est devenu de plus en plus clair que la mise en relation de données provenant de différentes sources, même anonymisées, peut permettre de réidentifier des individus. Certains types de données, comme les génomes entiers ou les traces GPS d'un smartphone, contiennent un tel niveau d'informations personnelles sensibles qu'une anonymisation absolue n'est pas réaliste. La diffusion de données dont les informations personnelles identifiables ont été supprimées doit donc être considérée comme un continuum. La perte de vie privée doit ainsi être mise en balance au cas par cas avec la création de valeur.

Plusieurs approches peuvent entraver la réidentification. La confidentialité différentielle, par exemple, brouille les données en ajoutant des perturbations aléatoires, mais au détriment de la précision *(Analyse de flux).* Une autre option est de supprimer certains points de données ou de les combiner dans des catégories plus larges, comme avec la technique appelée «k-anonymity».

Analyser les données sans y accéder

Les données sensibles peuvent être stockées dans des enclaves avec un contrôle d'accès sophistiqué. Cela garantit que les analyses ne peuvent être effectuées que localement, seuls des résultats agrégés, qui protègent la vie privée, étant envoyés en dehors des enclaves. Une autre option en cours de développement est l'analyse fédérée, où les données sont conservées dans plusieurs systèmes locaux sans être partagées. Les calculs, y compris l'entraînement des algorithmes d'apprentissage automatique, sont effectués localement et en collaboration. Ici aussi, les seuls éléments partagés sont les résultats partiels et agrégés ou les mises à jour intermédiaires des modèles, tandis que les données originales ne sont jamais distribuées. Cela

permet de résoudre les problèmes difficiles de transfert transfrontalier de données, qui nécessitent des solutions juridiques au niveau international (Conclusion 9). D'une manière générale, les équipes de recherche qui développent des applications pour le big data doivent envisager très tôt le cadre éthique et juridique du traitement des données (Conclusion 2).

5.4 La responsabilité des algorithmes

Les applications du big data utilisent souvent des algorithmes d'apprentissage automatique capables de faire des prédictions sur la base de modèles entraînés avec certains jeux de données. Si ces algorithmes peuvent être très performants en matière de prédictions, on ne sait souvent pas exactement comment ils les ont générées. Ce problème peut soulever des questions éthiques et juridiques, comme abordé dans le chapitre 4 et dans le rapport *Ethical, legal and social issues of big data – a comprehensive overview*.

Le risque de discrimination

Les logiciels usuels suivent une série stricte d'instructions qui ont été (en grande partie) conçues par des humains chargés de la programmation. Des tests peuvent en principe garantir que les programmes fonctionnent comme prévu. La situation est différente avec de nombreux algorithmes d'apprentissage automatique: leurs résultats sont basés sur des modèles comportant un très grand nombre de paramètres, dont les valeurs sont générées automatiquement à partir de données d'entraînement. Leur comportement ne suit pas de règles codées par des humains.

Il est donc difficile de déterminer si ces résultats sont conformes aux normes éthiques établies, ou s'ils peuvent au contraire s'avérer discriminatoires à l'égard de certains groupes de population. Cela peut se produire si les données d'apprentissage ne sont elles-mêmes pas représentatives, ou si elles sont biaisées, périmées ou erronées, ce qui peut être le cas lorsqu'elles proviennent du Web. Les modèles d'apprentissage automatique dépendent des données d'entraînement, de sorte que leurs résultats peuvent reproduire les biais qui s'y trouvent. Et la suppression du paramètre «sexe» des données d'apprentissage peut ne pas empêcher l'obtention de résultats discriminatoires, car un modèle entraîné va peut-être automatiquement recréer la catégorie «sexe» à travers d'autres informations corrélées. Ce type de comportement peut échapper à la détection lors des premiers tests et n'apparaître que plus tard.

Comprendre l'apprentissage automatique

Les résultats produits par les réseaux neuronaux profonds et d'autres techniques d'apprentissage automatique peuvent être très difficiles à comprendre pour les humains, car les milliards de paramètres entraînables qui composent leurs modèles obscurcissent les mécanismes conduisant à des résultats particuliers. Il n'existe actuellement aucune solution reconnue pour surmonter complètement ce problème de «boîte noire» de l'intelligence artificielle.

Les scientifiques tentent de mieux comprendre ces systèmes automatisés afin

d'améliorer l'explicabilité et la traçabilité de leurs décisions. Ces objectifs sont cruciaux pour démontrer que les algorithmes sont non discriminatoires, responsables et dignes de confiance.

Une personne ou une entreprise affectée par un algorithme potentiellement biaisé n'a ni les connaissances ni la capacité d'argumenter de manière convaincante que le système a commis une erreur ou a été discriminatoire à son égard. Une possibilité serait d'inverser la charge de la preuve, en exigeant que les responsables d'un algorithme doivent démontrer qu'il a un comportement correct. Cela pourrait impliquer un processus de certification développé par une organisation publique ou privée (Conclusion 3). Il pourrait passer par la modification délibérée de jeux de données de test pour vérifier si les résultats sont conformes aux règles éthiques.

Qui est responsable des algorithmes?

Les progrès rapides de l'apprentissage automatique soulèvent la question de la responsabilité, comme cela a été largement débattu dans le cas des véhicules à conduite autonome. Qui doit être tenu pour responsable d'un accident: le ou la propriétaire du véhicule, le constructeur, ou bien personne? Il s'agit d'un domaine évolutif du droit et de la politique, et il n'existe actuellement aucun accord sur la réponse à donner à ce type de questions. Alors que les constructeurs doivent concevoir leurs voitures de manière à minimiser les risques dans des situations de conduite typiques, ils ne peuvent pas prévoir toutes les circonstances possibles. Il est essentiel de définir précisément les responsabilités afin que l'incertitude juridique n'entrave pas l'innovation.

Vers une nouvelle réglementation

Le cadre législatif est souvent en retard sur les progrès rapides de l'apprentissage automatique et sur la collecte de données en constante expansion. Jusqu'à présent, la législation s'est concentrée sur les droits individuels et la prévention des effets négatifs sur les individus, plutôt que sur les impacts sur la société dans son ensemble.

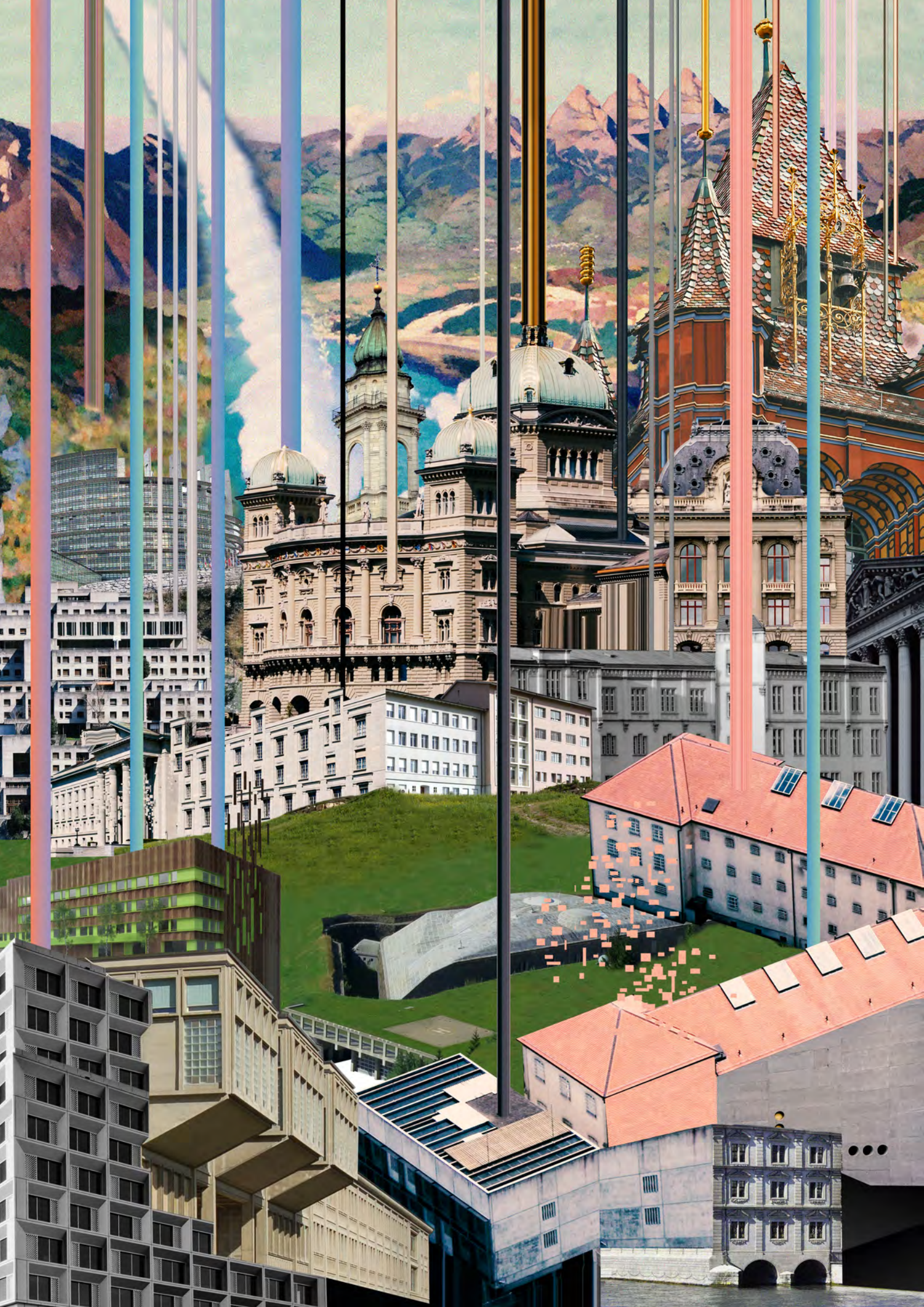
L'UE élabore actuellement une loi[16] visant à réglementer les applications de l'IA. Elle interdirait les applications considérées comme présentant un risque inacceptable, telles que les algorithmes manipulateurs ou les systèmes de notation sociale, tout en restreignant celles considérées comme présentant un risque élevé, telles que celles gérant des infrastructures critiques ou de sécurité. La Chine a également formulé une politique éthique en matière d'IA, privilégiant la sécurité sociale aux droits individuels[17]. Elle exclut le secteur public, libre de procéder à la reconnaissance faciale et au profilage social.

L'évolution rapide de la technologie, entraînée en grande partie par des entreprises internationales, pose un problème juridique complexe. La Suisse doit élaborer sa législation de manière proactive (Conclusion 7) afin de veiller à ce que les règles soient appliquées et respectées.

---

[16] Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union, Commission européenne (2021) ↗; voir également https://artificialintelligenceact.eu/

[17] New Generation Artificial Intelligence Code of Ethics, Ministère chinois des sciences et de la technologie (2021)



6.
Conclusions du comité de direction

Le comité de direction du PNR 75 a énoncé neuf conclusions sur le sujet du programme de recherche. Leur objectif est de présenter des observations aux institutions publiques, aux organisations privées et à la société en général pour qu'ils puissent se former une opinion sur les développements nécessaires pour exploiter le big data et pour assurer son utilisation responsable.

Le big data, l'intelligence artificielle et l'apprentissage automatique auront un impact considérable sur la société et des avantages potentiels dans tous les secteur. Le Programme national de recherche «Big Data» (PNR 75) a exploré des pistes variées pour accélérer le développement de nouvelles technologies et applications du big data, et aborder les défis sociétaux associés. Garder la maîtrise de cette évolution représente un défi majeur pour nos institutions publiques et privées, qui exige des efforts ciblés dans de nombreuses directions, de l'éducation à la réglementation, en passant par des initiatives sectorielles et des débats publics. Les conclusions du Comité de direction du PNR 75 ont été formulées sur la base des connaissances acquises au cours des cinq années de recherche ainsi que de réflexions collectives.

Ces conclusions ont pour but de promouvoir des développements appropriés et de soutenir les mesures déjà en cours. Il revient aux parties prenantes de décider comment – ou si – les transformer en actions concrètes. Elles explicitent les besoins et perspectives identifiées du point de vue de la recherche sur le big data. Ce condensé a été élaboré par les membres du Comité de direction sur la base des résultats des projets de recherche et de leurs propres connaissances et leurs expérience.

La recherche scientifique peut apporter des éléments de réponse aux questions soulevées par le big data. Ils peuvent néanmoins faire apparaître un manque de compatibilité entre les différentes approches. Il n'appartient pas aux scientifiques d'évaluer les priorités et les équilibres sociétaux, cette tâche revenant à la politique et aux processus démocratiques.

Le résumé du programme et ses conclusions représentent une contribution de la communauté scientifique à la formation de l'opinion, au débat politique et professionnel ainsi qu'à la planification des stratégies et mesures visant à développer des applications et des réglementations du big data. Il s'adresse notamment aux institutions qui définissent et façonnent l'espace suisse des données.

Favoriser un environnement approprié pour le développement du big data

1
Améliorer la formation au big data

Utiliser efficacement les technologies du big data exige de nouvelles connaissances et compétences. Les spécialistes de l'informatique, même formés il y a une dizaine d'années seulement, peuvent éprouver des difficultés face à des thèmes tels que l'utilisation responsable de grandes quantités de données, leur intégration et analyse, l'apprentissage automatique ou la visualisation. Il y a une pénurie de personnel qualifié tout au long de la chaîne de valeur du big data, ainsi qu'une concurrence intense dans le monde universitaire, les grandes entreprises informatiques et les start-ups pour recruter les meilleurs talents. Pour tirer parti des avantages du big data dans les entreprises, la société et la recherche, il est recommandé d'étendre l'enseignement du big data dans les formations scolaire, universitaire et continue.

La disponibilité limitée de personnel informatique qualifié constitue un goulot d'étranglement pour l'exploitation des technologies et des applications du big data. Cette situation appelle à une formation améliorée et plus large. Ce n'est que par l'éducation, la recherche et la sensibilisation que la Suisse pourra attirer les talents nationaux et internationaux du big data et renforcer sa position déjà forte dans ce domaine.

Les universités et hautes écoles spécialisées doivent enrôler davantage d'étudiantes et étudiants en informatique et proposer des programmes d'études axés sur le big data aux niveaux bachelor, master et doctorat. Des cours sur ses aspects pertinents doivent être proposés dans les domaines faisant fréquemment appel au big data. Cela pourrait accroître la diversité dans le domaine IT puisque le big data recouvre un domaine plus large que l'informatique et nécessite des compétences interdisciplinaires qui englobent également les aspects sociétaux, commerciaux et juridiques.

Il est important d'élargir les offres de formation continue sur les mégadonnées en couvrant tout le pipeline du big data, comme la collecte des données, leur préparation, intégration, nettoyage et mise en relation. Se former aux nouvelles bibliothèques open-source, qui évoluent rapidement, est nécessaire afin d'utiliser de manière qualifiée les infrastructures de stockage, de gestion et de valorisation des données. Il est important d'être en mesure d'utiliser des outils de visualisation de données. Enfin, le personnel doit être en mesure d'appréhender les questions juridiques, réglementaires et éthiques liées au big data, notamment en se formant dans des programmes interdisciplinaires.

Comme les développements rapides du big data toucheront tout le monde, l'éducation au big data doit être intensifiée et continuellement modernisée dans les écoles, les lycées et la formation professionnelle. Elle doit inclure la science des données et les questions sociétales.

2
Soutenir le conseil juridique et éthique pour les projets de R&D en big data

De nombreux projets de recherche et de développement ont besoin de conseils juridiques et éthiques sur des points tels que le droit d'utiliser ou de partager des données. Les responsables de projet peuvent être mis au défi de communiquer efficacement et de convaincre que la base légale et éthique de leurs travaux est solide. Ceci peut être facilité par des conseils compétents et des audits disponibles à un coût abordable.

Un éclairage juridique et éthique est de plus en plus souvent nécessaire dans les projets de recherche et développement pour déterminer quelles données peuvent être utilisées et partagées, et de quelle manière. Si l'expertise sur ces questions fait défaut, la décision prudente sera de ne pas utiliser ni partager les données disponibles, même dans le cas où cela serait parfaitement légal et éthique. Cela peut creuser un fossé entre les problèmes concrets et les cadres expérimentaux, avec des conséquences potentiellement négatives dans des domaines tels que la médecine personnalisée, la santé publique ou la durabilité.

Cette problématique touche notamment les hautes écoles et les institutions de recherche publiques. Des recherches prometteuses ne sont pas menées et les formations perdent de leur attrait. La situation est exacerbée par le risque d'une couverture médiatique négative et une protection du personnel limitée. Les responsables de projet pourraient tirer profit d'un service interne ou de l'administration publique leur fournissant des conseils compétents, dignes de confiance et fiables sur les questions juridiques, éthiques et de communication.

Comment s'assurer que des activités sont conformes à la loi et à l'éthique? Quelles sont les raisons d'éventuelles limitations? Comment communiquer sur des activités de manière transparente, en particulier au sujet de thèmes controversés? Un service pourrait fournir des conseils sur ces questions, des directives pour accompagner la mise en œuvre pratique, ainsi que des explications faciles à comprendre sur les considérations juridiques et éthiques. Il pourrait proposer des audits de l'utilisation et du partage des données, menés par des spécialistes qualifiés et certifiés.

3
Permettre la certification des applications du big data

Les applications du big data ont le potentiel d'améliorer un large éventail de processus, notamment dans l'administration publique et le secteur privé. Dans certains cas, les applications soulèvent des questions d'équité, de biais, de discrimination, de normes éthiques, ou encore de confidentialité. Il est recommandé de mettre à disposition des moyens de certifier les propriétés pertinentes des applications du big data afin de rendre leur adoption possible. Il s'agit à la fois de spécifier les propriétés pertinentes et de proposer des procédures pour certifier qu'une application les satisfait.

Le fonctionnement et les résultats des applications du big data manquent souvent de transparence, ce qui peut susciter la méfiance. Ceci peut être contré en définissant les propriétés pertinentes des applications du big data, telles que l'équité, l'explicabilité, la transparence, la responsabilité ainsi que l'absence de biais ou de discrimination. Il s'agit d'établir des moyens permettant aux fournisseurs d'applications de montrer que leurs applications possèdent ces propriétés. Il est recommandé de mettre en place des moyens de certifier les propriétés pertinentes des applications du big data afin d'instaurer une confiance raisonnable en celles-ci.

De nombreuses applications du big data sont utilisées dans des domaines non sensibles, offrant une amélioration des performances, de l'efficacité et des coûts. Des avantages considérables sont également attendus dans des contextes sensibles, comme les services sociaux, de santé ou de police. Des processus pour la conformité réglementaire et la certification de la conformité existent déjà dans des domaines importants tels que l'énergie, la construction ou le commerce. Il est recommandé de les étendre afin qu'ils incluent également les propriétés spécifiques aux applications du big data.

Intégrer le big data dans les organisations publiques et privées

4
Accroître l'exploitation des technologies du big data dans le secteur de la santé

La santé est un excellent exemple de secteur dans lequel le potentiel du big data est largement reconnu par les parties prenantes tout en restant loin d'être complètement exploité. Mettre l'accent sur une gestion et une prise de décision fondées sur les données pourrait transformer les pratiques actuelles et potentiellement améliorer la transparence, qualité, sécurité, efficacité et coordination des soins, tout en renforçant les compétences des patientes et des patients en matière de santé. Ce potentiel ne doit pas rester inexploité. Les aspects juridiques et éthiques doivent être abordés afin que le big data puisse être utilisée plus largement dans le secteur de la santé.

La collecte et l'utilisation de données de santé de haute qualité constituent des éléments clés de la médecine personnalisée et fondée sur les preuves. Elles peuvent améliorer les diagnostics, permettre la détection précoce des personnes à risque, favoriser la découverte de nouvelles interactions entre maladies, médicaments, et facteurs de risque, et améliorer l'adhésion des patient·es à leur traitement. L'utilisation du big data dans le domaine de la santé exige de surmonter des obstacles juridiques, éthiques, administratifs, financiers et informatiques. Le système de santé suisse est décentralisé et organisé localement, ce qui crée des défis de législation et d'administration pour permettre la collecte et l'utilisation de données à l'échelle nationale. Du point de vue de la science des données, cette organisation peut aboutir à des solutions informatiques locales fragmentées, et d'une sécurité et de qualité réduites. Il est important de d'améliorer considérablement l'accès aux données au bon endroit et au bon moment ainsi que les interfaces entre les différents acteurs.

La Suisse a mis en place le dossier électronique du patient (DEP) ainsi que des mesures de qualité dans le secteur de la santé, ce qui démontre la possibilité d'initiatives fédérales dans ce domaine. Dans sa forme actuelle, le DEP se résume à une collection non structurée de documents scannés. Il lui manque des résumés et des index structurés, des formats de données harmonisés, une interopérabilité sémantique et une terminologie standardisée. Il manque de compatibilité avec une analyse automatisée des données et des directives de standardisation. Le DEP a néanmoins un énorme potentiel si une couche d'anonymisation est introduite afin de permettre des analyses collectives des données médicales. Des technologies de protection de la vie privée peuvent être utiles, comme les enclaves, dans lesquelles les données sont uniquement traitées sans être accessibles directement de l'extérieur, ou encore l'analyse fédérée, qui peut traiter des données distribuées sans compromettre leur sécurité. Ces technologies doivent être davantage développées et ont besoin de comités compétents sur le plan éthique, juridique, scientifique et statistique pour la supervision de l'utilisation des données.

5
Renforcer l'élaboration et l'évaluation des politiques grâce au big data

Les possibilités accrues de collecte et d'analyse des données établissent une base solide pour renforcer une élaboration de politiques qui se base sur les faits. Il est possible de quantifier davantage les problèmes sociaux et économiques et d'évaluer l'efficacité des politiques et des réglementations. Ce potentiel doit être exploité d'une manière à la fois responsable et bénéfique.

Les technologies du big data ont un très grand potentiel dans de nombreux secteurs privés, mais également dans l'administration publique. Leur déploiement, couplé à la collecte de données, peut être utilisé dans l'élaboration des politiques publiques et dans l'évaluation de leur efficacité dans des secteurs tels que la santé, l'énergie, la finance, les transports, l'aménagement du territoire ou encore les sports. La collecte et l'évaluation des données permettent de comparer l'efficacité des politiques avec celles d'autres pays ainsi que d'identifier des politiques efficaces et des meilleures pratiques. Cela exige de considérer attentivement les aspects de sécurité et de respect de la vie privée.

En contrepartie, les données utilisées pour la conception et l'évaluation des politiques devraient être accessibles au public, ce qui nécessite à nouveau de considérer les questions de confidentialité et d'anonymisation. Dans l'ensemble, les entités administratives concernées doivent être soutenues afin de pouvoir gérer la charge croissante de travail et la communication. Les procédures et les mécanismes nécessaires à la mise en œuvre de telles coopérations doivent être développés et établis.

6
Promouvoir la collecte partagée des données, l'open-source et les benchmarks

Des infrastructures librement accessibles vont accélérer la création de valeur à partir des données. Il est recommandé d'affiner les politiques en matière de publication des données afin d'augmenter le volume de l'open data. Un meilleur soutien à la création de benchmarks et de cas d'usage pour les applications dans différents domaines scientifiques se justifie. Les logiciels open-source représentent souvent une alternative intéressante aux produits commerciaux onéreux. Il est recommandé d'allouer des fonds supplémentaires au développement de logiciels open-source afin de permettre des fonctionnalités et capacités open-source nouvelles et incluant par exemple des infrastructures informatiques de nouvelle génération et des outils d'apprentissage automatique.

Le développement d'applications du big data est accéléré par la disponibilité accrue des données, dont la collecte est généralement onéreuse. Pour garantir une collecte continue dans un cadre scientifique, il est important d'assurer un rapport coût-bénéfice convenable. La collecte de données doit être encouragée lorsqu'elle entraîne des coûts substantiels. On peut autoriser une publication différée afin de permettre aux collecteurs des données une première création de valeur avant la publication des données (à l'instar du système des brevets qui encourage les investissements dans l'innovation), veiller à ce que les organismes de financement couvrent les coûts générés par le respect des principes de gestion des données de recherche FAIR[18] – qui peuvent s'étendre au-delà de la fin d'un projet –, et d'exiger la publication des données de recherche comme condition préalable à celles des résultats. Il faut une approche nuancée du partage des données qui reconnaît que toutes les données n'ont pas la même valeur, afin que les ressources soient utilisées à bon escient.

Le développement des applications peut être accéléré par des benchmarks englobant des ensembles de données anonymes et des cas d'usage représentant des scénarios courants dans les domaines ciblés. Ces références sont utiles pour le développement et le test des applications ainsi que pour améliorer la précision et les prédictions des algorithmes. Elles peuvent faciliter le déploiement et la validation à grande échelle des applications du big data.

Les logiciels open-source peuvent accélérer la création de valeur à partir de données. Il est recommandé de mettre en place des incitations au partage des outils sous forme de logiciels open-source. Par exemple, l'évaluation des carrières académiques devrait considérer l'impact sur la société d'une adoption à grande échelle de logiciels libres développés par des scientifiques, au même titre que leur nombre de citations. De telles initiatives constitueraient un service public important, contribueraient à attirer des talents internationaux en Suisse, et constitueraient un élément-clé de la numérisation de la Suisse.

---

[18] Les principes FAIR visent à garantir que les données et autres objets numériques de la recherche soient trouvables, accessibles, interopérables et réutilisables. Voir https://www.go-fair.org/fair-principles ainsi que la Stratégie Nationale Suisse Open Research Data, Swissuniversities (2021)

Actualiser et créer des réglementations adéquates

7
Poursuivre une réglementation proactive du big data

Alors que les technologies du big data sont déployées à un rythme rapide, leur réglementation n'en est qu'à ses débuts et accuse un retard considérable sur les développements technologiques. Le manque de réglementation peut avoir des effets négatifs, notamment sur la démocratie, la santé mentale des jeunes, la concurrence ainsi que l'innovation, par exemple, en raison de biais ou d'une concurrence limitée. Il est recommandé de déployer des efforts généralisés pour accélérer les processus réglementaires, ceux-ci pouvant jouer un rôle important pour éviter les effets négatifs et d'encourager la création de valeur basée sur le big data.

Les applications du big data ont un impact profond et étendu sur la société. La réglementation peut accélérer la création responsable de valeur tout en limitant les effets négatifs. Une approche plus proactive de la réglementation peut donc promouvoir la création responsable de valeur, faciliter la concurrence et l'innovation, et servir la démocratie.

La fracture numérique s'étend au big data et décrit une relation asymétrique entre les institutions qui collectent, stockent et analysent les mégadonnées et les personnes qui y sont sujettes. Elle est une conséquence inévitable d'une société valorisant la liberté et la diversité. Au lieu de tenter d'éliminer cette fracture, il est recommandé que la législation identifie les préjudices qui peuvent en résulter et élabore des garanties juridiques pour les personnes désavantagées.

Le succès des applications du big data repose sur la confiance. Lors de la mise en place des cadres dans lesquels les données peuvent être collectées, analysées et utilisées, il convient non seulement d'insister sur la création de normes (en autorégulation) équilibrant les intérêts des entreprises et de leur clientèle, mais aussi de donner à celle-ci les moyens de prendre des décisions en connaissance de cause.

Dans l'ensemble, il est important de développer des garanties juridiques pour compenser les préjudices causés par la fracture du big data, à l'aide de normes pour la collecte, le partage et l'analyse des données, et pour protéger les groupes vulnérabilisés par le déploiement du big data.

8
Assurer la confidentialité des données et la souveraineté numérique

Le déploiement d'applications basées sur le big data comporte des risques pour la vie privée et les droits connexes des personnes. Même si des cadres juridiques de base existent en Suisse (nouvelle loi sur la protection des données à caractère personnel) et dans l'UE (règlement général sur la protection des données, RGPD), la conformité aux règles peut représenter un défi. Il est recommandé de sensibiliser aux questions de confidentialité et aux règles de protection des données les scientifiques et les ingénieur·es travaillant avec le big data, les propriétaires de données et les responsables de la protection des données. Il est également recommandé d'élaborer des normes complètes sur le respect de la vie privée et d'accorder une attention accrue à la sécurité des infrastructures numériques.

La réglementation fondée sur la souveraineté d'un État sur un espace physique déterminé est insuffisante lorsqu'il s'agit de big data. Une coordination et une coopération internationales sont nécessaires pour préserver la sécurité des infrastructures numériques ainsi que la vie privée et les autres droits numériques de la population. Néanmoins, des efforts peuvent et doivent être déployés aussi au niveau national afin de garantir la confidentialité des données. Les décideurs politiques, les administrations nationales et cantonales, les universités et les scientifiques doivent renforcer et compléter le cadre juridique national.

Un programme national de protection des droits liés aux données implique des acteurs et thèmes variés. Il est important d'encourager la création de liens solides entre les nombreuses parties prenantes dans les disciplines concernées. Il est recommandé d'élaborer des méthodologies pour établir les meilleures pratiques en matière de collecte et d'anonymisation des données, de stockage sécurisé et d'analyse préservant la vie privée. Par exemple, le développement d'enclaves de données offre une option intéressante pour garantir la confidentialité. Le concept de consentement éclairé devrait être complété par des mécanismes de protection spécifiques.

La recherche développe des techniques de préservation de la confidentialité, mais leur déploiement prend du temps. Certaines d'entre elles peuvent contribuer à une stratégie nationale de protection des droits liés aux données, par exemple la nomination d'administrateurs protégeant les données personnelles des individus, ou l'implémentation de critères d'équité dans l'analyse des mégadonnées afin d'éviter la discrimination. Le traitement quotidien des questions liées à la vie privée pourrait être facilité par l'élaboration de lignes directrices en matière de protection des données, ou par la création d'un centre de compétences spécialisé (par exemple au sein du Centre national pour la cybersécurité) en tant que service public chargé des questions juridiques liées à la vie privée lors du déploiement du big data.

9
Renforcer l’harmonisation transnationale des réglementations

Les données traversent souvent les frontières. L’accès aux données de l’étranger ainsi que le déploiement international de services basés sur le big data sont courants. Une perspective purement nationale sur l’application et la réglementation du big data est insuffisante. Il est nécessaire d’observer et de s’engager au niveau international. En raison des nombreuses organisations internationales ayant leur siège en Suisse, le pays se trouve dans une position unique pour soutenir les activités d’harmonisation par les institutions d’orientation transnationale. La Suisse a la possibilité de démontrer son engagement et son expertise dans les organisations internationales ainsi que dans la législation nationale.

La mondialisation des flux de données et le déploiement accru d’applications basées sur le big data rendent nécessaire la mise en place de cadres réglementaires transfrontaliers harmonisés et couvrant le commerce international. Tandis que les négociations au sein de l’Organisation mondiale du commerce (OMC) sont toujours en cours, les accords commerciaux bilatéraux et régionaux (préférentiels) réglementent de plus en plus le commerce des produits et services numériques, ainsi que les flux de données.. Les nouvelles règles englobent souvent des aspects de la protection des données, de la cybersécurité et de la confidentialité commerciale.

Alors que la Suisse s’engage activement dans les négociations, un soutien supplémentaire est recommandé. Il est fortement recommandé que la Suisse apporte sa contribution à l’élaboration en cours du Guide OCDE sur le devoir de diligence pour une conduite responsable des entreprises (CRE). La Suisse a joué un rôle important en tant que promoteur du Forum de l’ONU sur la gouvernance de l’Internet à Genève. Compte tenu des tensions accrues dans le monde numérique, il est recommandé que la Suisse fournisse des efforts pour contribuer à éviter la fragmentation dans la réglementation de l'économie basée sur les données.

Annexe:
Le programme national de recherche «big data» (PNR 75)

www.pnr75.ch

Informations clés

Chronologie

2014
Proposition d'un Programme national de recherche (PNR) sur le big data

2015
Mandat du Conseil fédéral au Fonds national suisse de la recherche scientifique pour la réalisation du PNR 75

2015
Appel à projets de recherche et sélection

2017–2021
Travaux de recherche

2022
Travail de synthèse et diffusion des résultats

2023
Publication du résumé du PNR 75

Chiffres

Budget
25 millions de francs suisses

Projects
34 projets de recherche et 3 activités transversales

Organisation

Comité de direction du PNR 75

Professeur Christian S. Jensen
Département d'informatique, Université d'Aalborg (Président)

Professeure Sihem Amer-Yahia
CNRS, Laboratoire d'Informatique de Grenoble LIG, Université Grenoble Alpes UGA (depuis le 12.07.2016)

Professeure Sabrina de Capitani di Vimercati
Département d'informatique, Université de Milan

Professeur Friedrich Eisenbrand
Institut de mathématiques, EPFL (depuis le 01.01.2021)

Professeur Joerg Huelsken
Institut suisse de recherche expérimentale sur le cancer ISREC, EPFL

Professeur émérite Erkki Oja
Département d'informatique, Université d'Aalto

Professeur Reinhard Riedl
Institut de gestion des technologies numériques, Haute école spécialisée bernoise

Professeure Caroline Sporleder
Institute for Digital Humanities, Université Georg-August de Göttingen (jusqu'au 31.12.2019)

Professeur Rolf H. Weber
Faculté de droit, Université de Zurich

Délégué de la division Programme du Conseil national de la recherche pour le PNR 75

Professeur Bert Müller
Biomaterials Science Center, Université de Bâle (depuis le 01.01.2021)

Professeur Friedrich Eisenbrand,
Institut de mathématiques, EPFL (jusqu'au 31.12.2020)

Responsable du programme PNR 75

Boris Buzek
Fonds national suisse, Berne (depuis le 01.11.2022)

Dr Stefan Husi
Fonds national suisse, Berne (du 01.11.2020 au 31.10.2022)

Dr Christian Mottas
Fonds national suisse, Berne (jusqu'au 31.10.2020)

Représentant de la Confédération au PNR 75

Dr Uwe Heck
Chancellerie fédérale, Secteur Transformation numérique et gouvernance de l'informatique (depuis le 01.01.2019)

Willy Müller
Unité de pilotage informatique de la Confédération (jusqu'au 31.12.2018)

Chargée du transfert de connaissances

Beatrice Huber
Académie suisse des sciences techniques (SATW), Zurich (depuis le 01.12.2018)

Dr Béatrice Miller
Académie suisse des sciences techniques (SATW), Zurich (jusqu'au 30.11.2018)

Les 34 projets de recherche

Module 1: Infrastructures du big data

Flux de données: monitoring en temps réel
David Basin, Dmytro Traytel, ETH Zurich
Big data monitoring

Analyse de flux: traitement rapide et préservation de la vie privée
Michael Böhlen, Université de Zurich
Privacy preserving, peta-scale stream analytics for domain-experts

Apprentissage automatique: robustesse et généralisabilité
Volkan Cevher, EPFL
Theory and methods for accurate and scalable learning machines

Data centres: monitoring efficace des performances
Lydia Yiyu Chen, Delft University of Technology, Pays-Bas (anciennement IBM Research Zurich)
Dapprox: dependency-ware approximate analytics and processing platforms

Données peu structurées: nouveaux outils d'intégration
Philippe Cudré-Mauroux, Université de Fribourg
Tighten-it-all: big data integration for loosely-structured data

Modèles de langage: nouvelles méthodes pour agents conversationnels
Thomas Hofmann, ETH Zurich
Conversational agent for interactive access to information

Modèles numériques urbains: scans 3D réalisés par un véhicule
Frédéric Kaplan, EPFL
ScanVan – a distributed 3d digitalization platform for cities

Coresets: du big data avec moins de données
Andreas Krause, ETH Zurich
Scaling up by scaling down: big ML via small coresets

Scala pour le big data
Martin Odersky, EPFL
Programming language abstractions for big data

In-network computing: solutions pour l'analyse des graphes
Robert Soulé, Université de la Suisse italienne
Exploratory visual analytics for interaction graphs

Algorithmes de prédiction rapide
Marco Zaffalon, Istituto Dalle Molle di studi sull'Intelligenza Artificiale USI-SUPSI
State space Gaussian processes for big data analytics

Exploration de graphes
Willy Zwaenepol, University of Sydney (d'abord EPFL)
Building flexible large-graph processing systems on commodity hardware

Module 2: Défis sociétaux et réglementaires

Accords commerciaux: impacts sur le droit national
Mira Burri, Université de Lucerne
The governance of big data in trade agreements: design, diffusion and implications

Big Data dans l'assurance
Markus Christen, Université de Zurich
Between solidarity and personalization – Dealing with ethical and legal big data challenges in the insurance industry

Réglementer la recherche recourant au big data
Bernice Simone Elger, Université de Bâle
Ethical and legal regulation of big data research – Towards a sensible and efficient use of electronic health records and social media data

Les défis juridiques du Big Data
Sabine Gless, Université de Bâle
Legal challenges in big data. Allocating benefits. Averting risks

Incertitudes dans le big data: le cas des simulations climatiques
Reto Knutti, ETH Zurich
Combining theory with big data? The case of uncertainty in prediction of trends in extreme weather and impacts

Le big data en pratique: sociologie, sciences des données et journalisme
Sophie Mützel, Université de Lucerne
Facing big data: methods and skills needed for a 21st century sociology

Big data dans la santé: un cadre éthique
Effy Vayena, ETH Zurich
BEHALF – Bigdata-ethics-health framework

Big data dans les ressources humaines
Antoinette Weibel, Université de Saint-Gall
Big data or big brother? Big data HR control practices and employee trust

Module 3: Applications du big data

Gestion des transports: traces de mobilité individuelle anonymes
Kay W. Axhausen, ETH Zurich
Big data transport models: the example of road pricing

Pig data: analyses de la filière porcine en Suisse
John Berezowski, Université de Berne
Pig data: health analytics for the Swiss swine industry

Détection d'inondation: géolocalisation automatique de vidéos crowdsourcées
Susanne Bleisch, Haute école du Nord-Ouest de la Suisse (FHNW)
EVAC – Employing video analytics for crisis managament

Chimie computationnelle: découverte de nouvelles molécules
Helmut Harbrecht, Université de Bâle
Big data for computational chemistry: unified machine learning and sparse grid combination technique for quantum based molecular design

Soins intensifs: un système d'alerte automatisé
Emanuela Keller, Hôpital universitaire de Zurich
ICU-cockpit: IT platform for multimodal patient monitoring and therapy support in intensive care and emergency medicine

Politique fondée sur les faits: démontrer les causalités dans les données
Michael Lechner, Université de Saint-Gall
Causal analysis with big data

Cartographie de l'innovation: analyse des brevets
Alessandro Lomi (Université de la Suisse italienne)
The global structure of knowledge networks: data, models and empirical results

Genetic big data: une indexation puissante
Gunnar Rätsch, ETH Zurich
Scalable genome graph data structures for metagenomics and genome annotation

Douleurs dorsales: une solution personnalisée sur smartphone
Robert Riener, ETH Zurich et Walter Karlen, Université d'Ulm (d'abord ETH Zurich)
Personalized management of low back pain with mHealth: big data opportunities, challenges and solutions

Érosion des sols: quantification par photographie aérienne
Volker Roth, Université de Bâle
WeObserve: integrating citizen observers and high throughput sensing devices for big data collection, integration, and analysis

Comparaison de génomes: des analyses plus rapides
Nicolas Salamin, Université de Lausanne
Efficient and accurate comparative genomics to make sense of high-volume low-quality data in biology

Potentiel des énergies renouvelables: estimations pour la Suisse
Jean-Louis Scartezzini, EPFL
Hybrid renewable energy potential for the built environment using big data: forecasting and uncertainty estimation

Bases de données bioinformatiques: recherches en langage naturel
Kurt Stockinger, Université des sciences appliquées de Zurich (ZHAW)
BIO-SODA: enabling complex, semantic queries to bioinformatics databases through intuitive searching over data

Éruptions solaires: prédiction de tempêtes géomagnétiques
Svyatoslav Voloshynovskiy, Université de Genève
Machine learning based analytics for big data in astronomy

Les 3 activités transversales

Big data: open data and legal strings
Sabine Gless, Université de Bâle

ELSI-Task-Force for the National Research Programme Big Data
Markus Christen, Université de Zurich

Women in big data
Lydia Yiyu Chen, Delft University of Technology, Pays-Bas (anciennement IBM Research Zurich)

Publications et matériel didactique

Ethical, legal and social issues of big data – a comprehensive overview, Eleonora Viganò (Ed.), PNR 75 (2022)

Big data: outil pédagogique pour les cycles secondaires, PNR 75 et Musée de la communication, Berne (2020)

Big data ethics recommendations for the insurance industry, PNR 75 (2019)

Impressum

Ce résumé du Programme national de recherche «Big Data» (PNR 75) intègre les résultats des 37 projets du PNR 75 dans une vue d'ensemble présentant les opportunités et les défis liés au big data. Les auteurs et autrices ont synthétisé ces résultats et apporté leur expertise et expérience scientifiques. Le texte a été consolidé et édité par un journaliste scientifique. Les conclusions présentées à la fin du document sont le résultat d'un processus collectif en plusieurs étapes: elles ont été rédigées, discutées et consolidées par les membres du Comité de direction, et représentent son consensus.

Ce résumé doit être considéré comme une contribution scientifique au processus de formation de l'opinion, au débat politique et spécialisé ainsi qu'à la planification de stratégies et de mesures pour les transformations politiques et sociales liées au big data. Le texte est la responsabilité collective du Comité de direction. Ses analyses et conclusions ne reflètent pas nécessairement celles des équipes de recherche ou du Fonds national suisse. Vous trouverez sur le site www.pnr75.ch de plus amples informations sur tous les projets de recherche du PNR 75 cités dans le résumé.

Auteurs et autrices

Professeure Sihem Amer-Yahia, Université Grenoble Alpes UGA

Professeur Joerg Huelsken, EPFL

Professeur Christian S. Jensen, Université d'Aalborg

Professeur émérite Erkki Oja, Université d'Aalto

Professeur Reinhard Riedl, Haute école spécialisée bernoise

Professeur Rolf H. Weber, Université de Zurich

Rédaction

Dr Daniel Saraga, Saraga Communications, Bâle

Coordination

Boris Buzek, Fonds national suisse, Berne

Dr Stefan Husi, Fonds national suisse, Berne

Beatrice Huber, Chargée du transfert de connaissances, PNR 75

Production

Traduction: STP Language Services, Stäfa

Mise en page et illustrations: Gabriel Alber, Alber Visuelle Kommunikation, Zurich

Impression: Druckerei Herzog AG, Langendorf (SO)

Éditeur

Comité de direction du Programme national de recherche «Big Data» (PNR 75)

Citation suggérée:

Comité de direction du PNR 75 (2023): «Big data: applications, technologies, et aspects sociétaux; Résumé du Programme national de recherche ‹Big Data› (PNR 75)», Fonds national suisse, Berne

Les Programmes nationaux de recherche (PNR) sont un instrument d'encouragement du Fonds national suisse (FNS). Leur objectif est de permettre des recherches contribuant à la solution de problèmes contemporains d'importance nationale. Le Programme national de recherche «Big Data» (PNR 75) s'est tenu de 2015 à 2022.

ISBN 978-3-907087-65-7