conférences
éducation
évènements
la médiation scientifique
ateliers
animations
formation
exposés

Terra Data

4 avril 2017 – 8 janvier 2018

Enseignants de cycle 4 et de lycée

Département Éducation et Formation
Cité des sciences et de l'industrie
30 avenue Corentin-Cariou
75019 Paris
www.cite-sciences.fr

2017

Sommaire

I Liens avec les programmes scolaires

Cycle 4

Enseignements pratiques interdisciplinaires possibles, thématiques « information, communication, citoyenneté », « sciences, technologie et société » – en lien avec la physique-chimie, les sciences de la vie et de la Terre, l'éducation aux médias et à l'information, les mathématiques, l'histoire des arts.

Histoire et géographie
Compétences travaillées : « s'informer dans le monde du numérique ».

Technologie
Comprendre le fonctionnement d'un réseau informatique. Écrire, mettre au point et exécuter un programme.

Mathématiques
Algorithmique et programmation.

Éducation aux médias et à l'information
Utiliser les médias et les informations de manière autonome. Exploiter l'information de manière raisonnée. Utiliser les médias de manière responsable. Produire, communiquer, partager des informations.

Lycée – voie générale

En 2de

Enseignement moral et civique
La personne et l'État de droit.

Enseignement d'exploration
Informatique et création numérique. Création et innovation technologiques.

Mathématiques
Statistiques et probabilités : statistique descriptive, analyse des données. Échantillonnage.
Algorithmique : instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

En 1re

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Histoire et géographie
Mobilités, flux et réseaux de communication dans la mondialisation.

Mathématiques
Statistiques et probabilités : Statistique descriptive, analyse des données. Échantillonnage.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Sciences économiques et sociales
Sociologie générale et sociologie politique : groupes et réseaux sociaux. Contrôle social et déviance.

Enseignement spécifique 1re S : sciences de l'ingénieur
Analyse d'un système. Communiquer.

En terminale

Enseignement moral et civique
Biologie, éthique, société et environnement.

Histoire et géographie
Les dynamiques de la mondialisation.

Mathématiques
Probabilités et statistique.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Droit et grands enjeux du monde contemporain : enseignement de spécialité en série L
Des sujets du droit : Internet et le droit.

Physique-chimie
Transmettre et stocker de l'information.

Sciences de l'ingénieur : enseignement spécifique terminale S
Analyse d'un système. Communiquer.

Sciences de la vie et de la Terre : enseignement de spécialité en série S
Atmosphère, hydrosphère, climats : du passé à l'avenir.

Informatique et sciences du numérique : enseignement de spécialité en série S
Représentation de l'information, algorithmique, langages et programmation, architectures matérielles.

Lycée – voie technologique

En 1re

Enseignements obligatoires communs aux séries STI2D, STL et STD2A

Géographie
La mondialisation.

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Enseignements obligatoires communs aux séries STI2D et STL

Mathématiques
Statistiques et probabilités : Statistique descriptive, analyse des données. Échantillonnage.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Enseignements technologiques transversaux et spécifiques des spécialités - série STI2D

Tronc commun
Outils et méthodes d'analyse et de description des systèmes (outils de représentation, approche comportementale). Solutions technologiques (structures matérielles et/ou logicielles, constituants d'un système).

Spécialité architecture et construction
Projet technologique. Conception d'un ouvrage. Vie de la construction.

Spécialité énergie et environnement
Projet technologique. Conception d'un système. Transports et distribution d'énergie, études de dossiers technologiques. Réalisation et qualification d'un prototype.

Spécialité innovation technologique et éco-conception
Projet technologique. Conception mécanique des systèmes. Prototypage de pièces.

Spécialité systèmes d'information et numérique
Projet technologique. Maquettage des solutions constructives. Réalisation et qualification d'un prototype.

Enseignements obligatoires spécifiques - série STL

Mesure et instrumentation
Instrumentation : instruments de mesure, chaîne de mesure numérique.

Sciences physiques et chimiques en laboratoire
D'une image à l'autre. Images et information.

Enseignements obligatoires - série STD2A

Physique-chimie
Images photographiques.

Design et arts appliqués
Les outils infographiques

Enseignements obligatoires - série STMG

Sciences de gestion
Information et intelligence collective. Temps et risque.

Programme d'enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Mathématiques
Statistique et probabilités.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Enseignements obligatoires - série ST2S

Sciences et techniques sanitaires et sociales
Études contribuant à la connaissance de l'état de santé et de bien-être des populations.

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Mathématiques
Statistique et probabilités.

En terminale

Enseignements obligatoires communs aux séries STI2D, STL et STD2A

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Enseignements obligatoires spécifiques - série STI2D

Mathématiques
Probabilités et statistique.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Physique-chimie
Erreurs et notions associées. Incertitudes et notions associées. Expression et acceptabilité du résultat.

Enseignements technologiques transversaux et spécifiques des spécialités

Tronc commun
Outils et méthodes d'analyse et de description des systèmes (outils de représentation, approche comportementale). Solutions technologiques (structures matérielles et/ou logicielles, constituants d'un système).

Spécialité architecture et construction
Projet technologique. Conception d'un ouvrage. Vie de la construction.

Spécialité énergie et environnement
Projet technologique. Conception d'un système. Transports et distribution d'énergie, études de dossiers technologiques. Réalisation et qualification d'un prototype.

Spécialité innovation technologique et éco-conception
Projet technologique. Conception mécanique des systèmes. Prototypage de pièces.

Spécialité systèmes d'information et numérique
Projet technologique. Maquettage des solutions constructives. Réalisation et qualification d'un prototype.

Enseignements obligatoires spécifiques - série STL

Mathématiques spécialité « sciences physiques en laboratoire »
Probabilités et statistique.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Mathématiques spécialité « biotechnologies »
Statistique et probabilités.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Physique-chimie spécialité « sciences physiques en laboratoire » et « biotechnologies »
Erreurs et notions associées. Incertitudes et notions associées. Expression et acceptabilité du résultat.

Chimie-biochimie-sciences du vivant
Les systèmes vivants contiennent, échangent et utilisent de l'information génétique.

Biotechnologies
Initiation à la biologie moléculaire et au génie génétique.

Sciences physiques et chimiques en laboratoire
Erreurs et notions associées. Incertitudes et notions associées. Expression et acceptabilité du résultat.
Capteurs électrochimiques.
Traitement du signal.

Enseignements obligatoires - série STD2A

Physique-chimie
Voir des objets colorés, analyser et réaliser des images (images photographiques).

Design et arts appliqués
Les outils infographiques.

Enseignements obligatoires communs - série STMG

Économie - Droit
Quelles sont les grandes questions économiques et leurs enjeux actuels ? (les échanges économiques). Comment se crée et se répartit la richesse ? (la combinaison des facteurs de production et l'évolution des technologies). Une régulation des échanges internationaux est-elle nécessaire ? Quels sont les droits reconnus aux personnes ? Qu'est-ce qu'être responsable ?

Histoire-Géographie
Les territoires dans la mondialisation. La mondialisation : acteurs, flux et réseaux.

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Management des organisations
Le rôle du management dans la gestion des organisations.
Le management stratégique : Le choix des objectifs et le contrôle stratégique.
Le processus et le diagnostic stratégiques.

Mathématiques
Feuilles automatisées de calcul. Information chiffrée.
Statistique et probabilités.
Algorithmique : Instructions élémentaires. Boucle et itérateur, instruction conditionnelle.

Enseignements obligatoires spécifiques - série STMG

Gestion et finance, mercatique (marketing), ressources humaines et communication, systèmes d'information de gestion
Construire une image de l'entreprise (comment faciliter l'échange d'informations financières ? Qu'apporte l'environnement technologique au traitement de l'information financière ?)
Accompagner la prise de décision (qu'apporte l'analyse des coûts à la prise de décision ?)

Enseignements obligatoires - série ST2S

Biologie et physiopathologie humaines
Génétique moléculaire : expression de l'information génétique.

Histoire-géographie
Les territoires dans la mondialisation. La mondialisation : acteurs, flux et réseaux.

Enseignement moral et civique
Les enjeux moraux et civiques de la société de l'information.

Mathématiques
Statistique et probabilités.

Sciences et techniques sanitaires et sociales
Quels politiques et dispositifs de santé publique pour répondre aux besoins de santé ?
Quels politiques et dispositifs sociaux pour favoriser le bien-être social ?
Comment les organisations sanitaires et sociales mettent en place un plan d'action pour améliorer la santé ou le bien-être social des populations ?

Enseignements obligatoires - série TMD

Mathématiques
Probabilités. Statistiques.

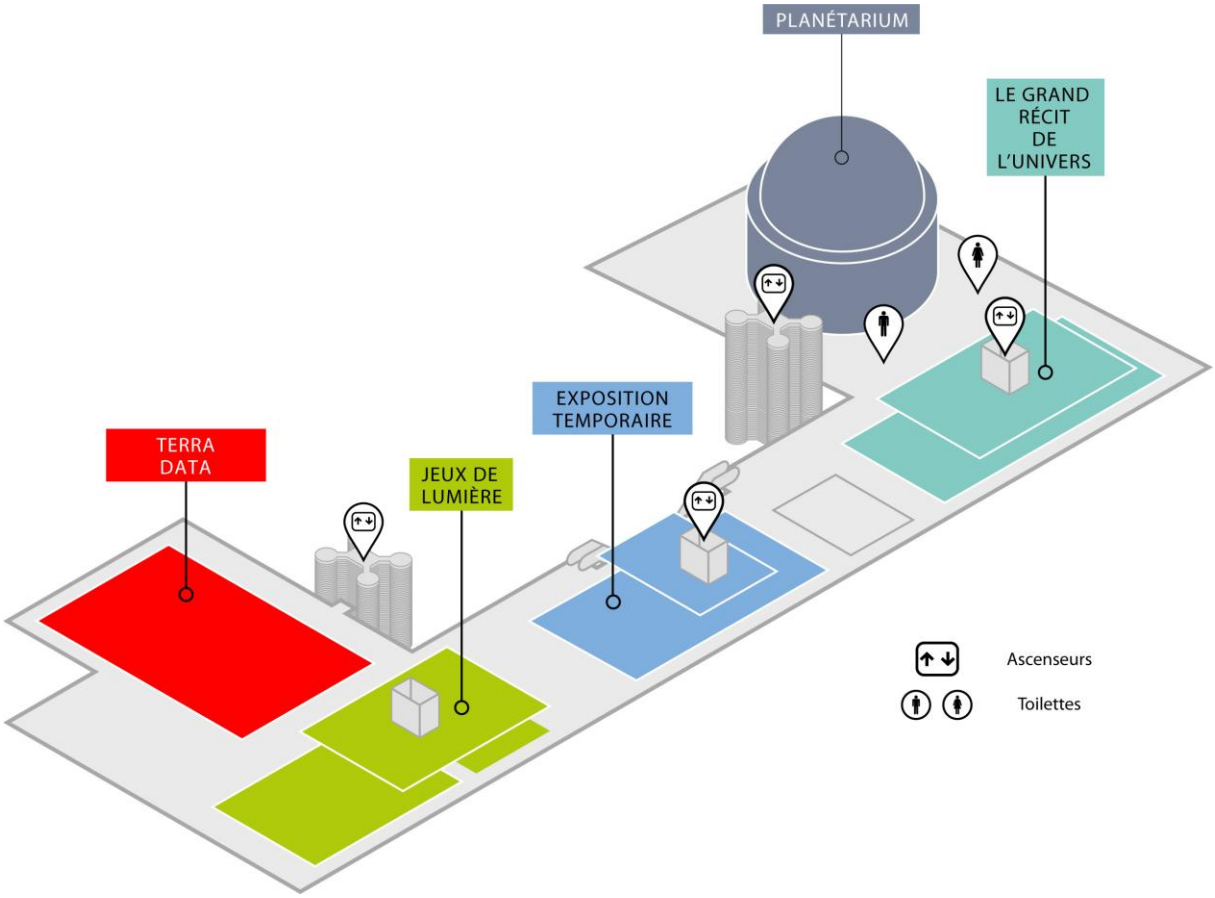
Enseignements facultatifs - série TMD

Arts
Pratique artistique. Histoire des arts.

II L'exposition *Terra Data*

II.1 Situation et plan de l'exposition

L'exposition « Terra Data. Nos vies à l'ère du numérique », qui occupe une surface totale de 500 m^2, prend place au niveau 2 (plateau L2) de la Cité des sciences et de l'industrie.

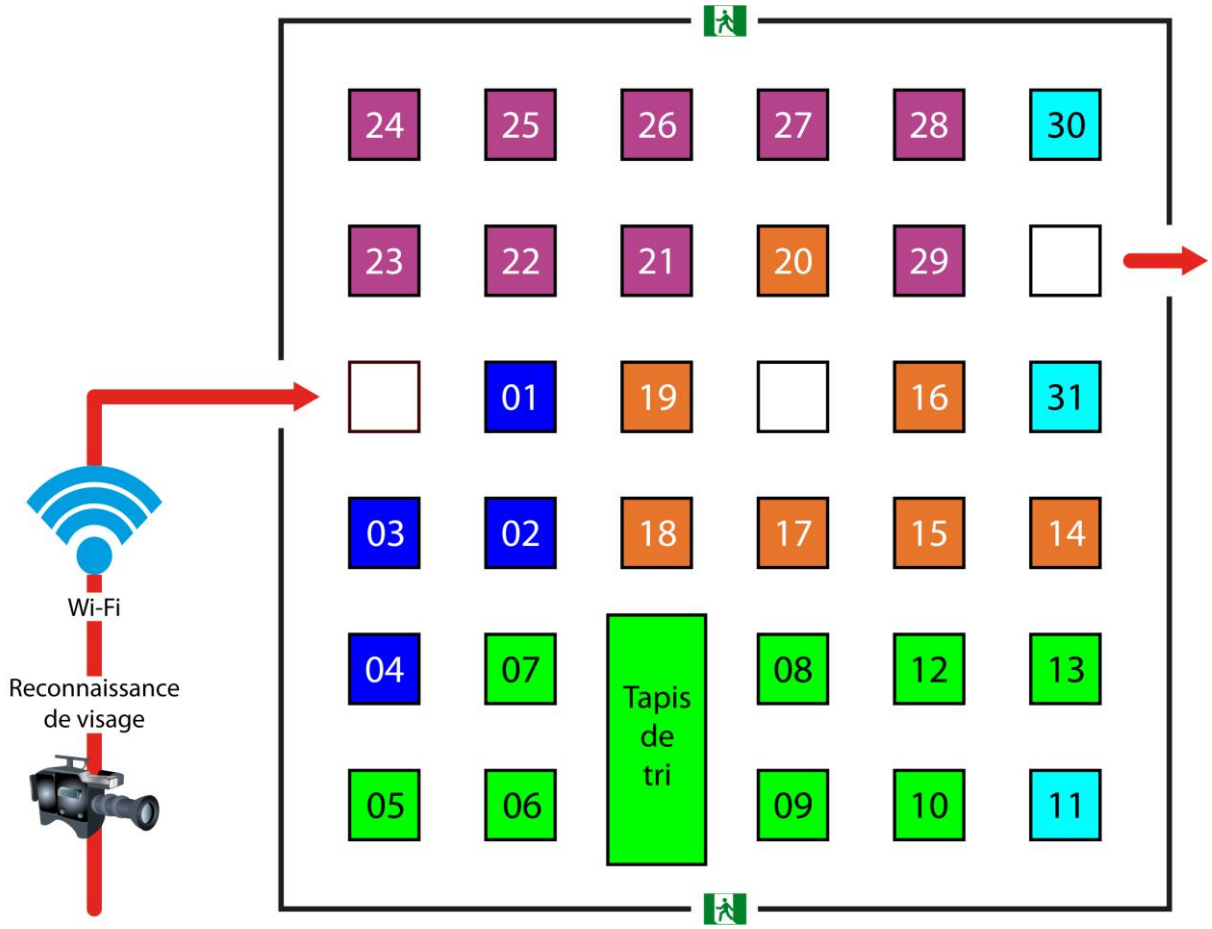


Terra Data, ce sont :

- 4 thèmes et 37 sujets ;
- 31 tables fonctionnelles et un jeu de 72 miroirs ;
- 1 tapis de jeu ;
- 12 audiovisuels et 14 multimédias ;
- 2 expériences numériques en temps réel et 1 concertation publique ;
- 1 médiation intégrée ;
- 1 livre d'exposition ;
- 10 partenaires (Inria, CNIL, ISC-PIF, Insee, Res Publica, Qwant, ESJ Pro, Keyrus, MAIF, Safran).

Avant même votre entrée dans l'exposition, vos élèves et vous vous verrez proposer, si vous le souhaitez, de passer devant un système de reconnaissance de visage et de vous connecter à un réseau Wi-Fi créé pour l'occasion. Ces deux dispositifs, validés par la CNIL, seront exploités un peu plus loin dans l'exposition (tables 10 et 24).

Le graphique ci-dessous révèle l'organisation schématique de *Terra Data* et suggère un sens de parcours parmi les 31 tables fonctionnelles. Dans la suite du document, nous présenterons un aperçu de ces éléments d'exposition dans l'ordre croissant.



L'exposition est en français. Le contenu des audiovisuels et des multimédias est également accessible en anglais et en italien, tout comme les contenus graphiques via un smartphone lisant les codes QR. Toutes les tables sont accessibles aux personnes à mobilité réduite. Pour les malentendants, neuf audiovisuels et trois multimédias sont traduits en langue des signes française. Enfin, six tables sont munies de prises jack et permettent aux malvoyants d'accéder à du contenu grâce à un casque, non fourni.

II.2 Propos et contenu de l'exposition

Le vivant, les objets, la société et les hommes deviennent, semble-t-il, mesurables en tout. Ils produisent en retour d'énormes masses de données. Quant à nos outils connectés, ils cachent des algorithmes qui traitent en continu et en accéléré ces données pour nous « prévoir » et nous « servir ».

Jusqu'à présent, nous semblons avoir massivement adopté les technologies numériques. Les multiples fonctionnalités des objets numériques, l'expérience sociale d'une interactivité en réseaux, l'enrichissement du champ perceptif individuel par des informations sur mesure, de nouveaux rapports connectés à la santé, au confort, aux loisirs, à la productivité personnelle et professionnelle, sont largement plébiscités.

Par ailleurs, le nombre grandissant de domaines concernés – science, information, industrie, santé, ville, transport, commerce, travail, finance, culture et, bien sûr, liens sociaux et vie privée – couplé à l'extraordinaire puissance de calcul de nos ordinateurs, ont un impact en temps réel sur nos vies et nos sociétés.

Les technologies des données connotent l'innovation technologique, la création de nouveaux services toujours plus personnalisés et la maîtrise du futur. D'autant plus que l'on attend d'elles une connaissance plus fine du monde et une amélioration des décisions – au point de chercher à devancer les événements.

Pour comprendre le monde que nous sommes en train de bâtir avec ces technologies, nous devons aller voir dans leurs « boîtes noires » et découvrir comment nous les faisons opérer et à quoi nous les utilisons. Car si elles ne sont ni bonnes ni mauvaises a priori, elles sont tout sauf neutres.

Dans ce contexte, complexe et problématique, cette exposition sur les données a pour ambition de proposer aux visiteurs de la Cité des sciences et de l'industrie – en particulier à vous, enseignants, et peut-être encore plus à vos élèves – une information scientifique et technique accessible. Cette culture partagée leur donnera des clefs pour se forger une opinion sur les enjeux des mutations en cours.

L'exposition se présente comme un parcours de découverte en quatre temps :

Temps 1 – Les données, qu'est-ce que c'est ?
Temps 2 – Les données, comment les traite-t-on ?
Temps 3 – Les données, qu'est-ce que ça change ?
Temps 4 – Les données, où ça nous mène ?

L'exposition est enrichie par la présence en son sein d'un médiateur scientifique qui proposera tous les jours des visites et des ateliers de courtes durées à tous les publics. Enfin, l'exposition veut donner à entendre et à voir la parole incarnée d'hommes et de femmes, acteurs et penseurs des données, via la présentation d'interviews sur des problématiques centrales. Car le monde des données n'est pas un univers froid et immatériel ; c'est avant tout un monde de passion, d'intérêt et de questionnement humains.

II.2.1 En guise d'introduction

Le projet *Venice Time Machine*, lancé par l'École polytechnique fédérale de Lausanne (EPFL) et l'université Ca' Foscari de Venise, a pour objectif la construction d'un modèle multidimensionnel de la ville de Venise et de son évolution sur plus de mille ans. Au moment même où vous lisez ces lignes, plus de 80 kilomètres de rayons de manuscrits anciens sont en cours de numérisation, de transcription et d'indexation. Des millions de photographies sont traitées à l'aide d'algorithmes et stockées dans un format adapté au calcul de haute performance. Des milliers de monographies indexées et consultables complètent ces sources primaires. Les renseignements extraits de toutes ces sources sont organisés dans un réseau sémantique de données et déployés dans l'espace et dans le temps sur une image à haute résolution de la ville.

La *Chronique de Nuremberg* (*Liber Chronicarum)* de Hartmann Schedel (1440 – 1514) est un incunable imprimé par Anton Koberger (v. 1440/1445 – 1513) en 1493. Il inclut de nombreuses illustrations rehaussées à la main comme cette vue de Venise.

II.2.2 Les données, qu'est-ce que c'est ?

Les données sont la matière première de la révolution numérique qui fait émerger de nouveaux rapports entre les citoyens, les États et les entreprises. La croissance exponentielle et vertigineuse de leur production justifie leur place centrale dans le discours et les représentations de l'exposition. La profusion des données a mené à la naissance d'un nouveau domaine technologique : le *Big Data* ou *mégadonnées*, terme d'ailleurs recommandé par la Délégation générale à la langue française et aux langues de France, un service rattaché au ministère de la Culture et de la Communication. Il est important de comprendre dès maintenant que l'objet du Big Data n'est pas l'information mais bien la donnée elle-même. Même s'il manque d'une définition sérieuse, le Big Data regroupe une famille d'outils qui permettent de stocker, traiter et analyser des déluges de données hétérogènes afin d'en faire ressortir de la valeur et, dans le cas des entreprises, de la création de richesse. Ces outils répondent à une problématique triple : variété, volume et vélocité. C'est la règle des 3V… qui à vrai dire, est plus un slogan susceptible d'évoluer en fonction de stratégies commerciales qu'une véritable règle. Les premiers outils du Big Data ont été créés par les entreprises chefs de file du web comme Google, Amazon ou Yahoo.

Bit, byte et octet : quelques définitions et rappels d'informatique

- Un **bit** (pour **b**inary dig**it**) est l'élément de base avec lequel travaille un ordinateur. Sa valeur est 0 ou 1, que l'on peut interpréter en oui/non ou vrai/faux. La capacité d'une puce de mémoire s'exprime couramment en bits et les taux de transfert de données, en bit/s.

- **Un octet est un ensemble de 8 bits**. Son symbole est *o*. Un ordinateur ne travaille jamais sur 1 bit à la fois mais sur un ou plusieurs octets, toujours donc sur des multiples de 8 bits. Les premiers ordinateurs personnels étaient 8 bits et ne comptaient que sur un octet à la fois. Les systèmes d'exploitation ont évolué pour passer de 8 bits à 16 bits (la série des Windows 1 à Windows 3.xx), à 32 bits (Windows 95 à Windows 10 32 bits) et enfin à 64 bits (Windows XP 64 bits à Windows 10 64 bits).

- Un **byte** est la plus petite unité adressable d'un ordinateur. Son symbole est *B*. Les bytes de 8 bits ayant très largement supplanté les autres en informatique, on exprime généralement les capacités de mémoire informatique en octet, surtout en français. ***Octet* et *byte* sont alors synonymes**. Sachez toutefois que, jusque dans les années 1970, il existait des processeurs avec des bytes de 6, 7 et 9 bits. Aujourd'hui encore, des processeurs utilisant des mémoires adressables par quantité de 4 bits sont utilisés pour programmer des automates ou des équipements industriels simples. De même, les bytes peuvent contenir plus de 8 bits dans le langage C. En bref, en anglais comme en français, on utilise le mot *octet* si l'on veut désigner explicitement une quantité de huit bits ; pour exprimer l'unité d'adressage indépendamment du nombre de bits, on utilise le terme *byte*.

Bit, byte et octet : leurs multiples

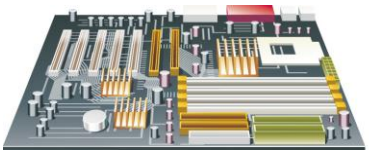
Comme nous venons de le voir, un bit ne peut avoir que deux valeurs, 0 ou 1. Un bit correspond donc à 2^1 = 2 possibilités, deux bits à 2^2 = 4 possibilités, trois bits à 2^3 = 8 possibilités et dix bits à 2^{10} = 1 024 possibilités. Ainsi, les puissances de deux interviennent-elles naturellement dans la mesure de la capacité des mémoires d'ordinateurs et des périphériques de stockage. Or, par une coïncidence arithmétique, il se trouve que 1 000, qui correspond au préfixe *kilo*, est une bonne approximation de 2^{10} = 1 024. De même, 10^6 (préfixe *méga*) est assez proche de 2^{20} (= 1 048 576) et 10^9 (préfixe *giga*), de 2^{30} (= 1 073 741 824). Les premiers informaticiens ont profité de ces coïncidences pour appliquer les préfixes du Système international d'unités *kilo*, *méga*, *giga*, etc. à des unités qui nécessiteraient des préfixes différents. Si l'erreur ainsi commise était faible pour les premières capacités mémoires qui s'exprimaient en kilooctets (2,4 %), elle est devenue difficilement tolérable avec les capacités actuelles qui se montent en gigaoctets (7,4 %) voire en téraoctets (10,0 %). À la fin des années 1990, la Commission électrotechnique internationale a donc publié une norme qui, d'une part, stipule que les préfixes du Système international d'unités ont toujours leurs valeurs de puissances de dix et ne doivent jamais être utilisés comme puissance de deux, et, d'autre part, introduit les préfixes binaires suivants pour représenter les puissances de deux : *kibi* (pour **ki**lo **bi**naire), *mébi* (pour **mé**ga **bi**naire), *gibi* (pour **gi**ga **bi**naire), *tébi* (pour **té**ra **bi**naire), etc.

Pour les multiples de l'octet, cela donne :

1 kibioctet	(kio)	= 2^{10} octets = 1 024 octets	
1 mébioctet	(Mio)	= 2^{20} octets = 1 024 kio	= 1 048 576 octets
1 gibioctet	(Gio)	= 2^{30} octets = 1 024 Mio	= 1 073 741 824 octets
1 tébioctet	(Tio)	= 2^{40} octets = 1 024 Gio	= 1 099 511 627 776 octets
1 pébioctet	(Pio)	= 2^{50} octets = 1 024 Tio	= 1 125 899 906 842 624 octets
1 exbioctet	(Eio)	= 2^{60} octets = 1 024 Pio	= 1 152 921 504 606 846 976 octets
1 zébioctet	(Zio)	= 2^{70} octets = 1 024 Eio	= 1 180 591 620 717 411 303 424 octets
1 yobioctet	(Yio)	= 2^{80} octets = 1 024 Zio	= 1 208 925 819 614 629 174 706 176 octets

Les préfixes du Système international d'unités correspondent aux mêmes multiplicateurs que dans les autres domaines. Ainsi :

1 kilooctet	(ko)	= 10^3 octets = 1 000 octets	
1 mégaoctet	(Mo)	= 10^6 octets = 1 000 ko	= 1 000 000 octets
1 gigaoctet	(Go)	= 10^9 octets = 1 000 Mo	= 1 000 000 000 octets
1 téraoctet	(To)	= 10^{12} octets = 1 000 Go	= 1 000 000 000 000 octets
1 pétaoctet	(Po)	= 10^{15} octets = 1 000 To	= 1 000 000 000 000 000 octets
1 exaoctet	(Eo)	= 10^{18} octets = 1 000 Po	= 1 000 000 000 000 000 000 octets
1 zettaoctet	(Zo)	= 10^{21} octets = 1 000 Eo	= 1 000 000 000 000 000 000 000 octets
1 yottaoctet	(Yo)	= 10^{24} octets = 1 000 Zo	= 1 000 000 000 000 000 000 000 000 octets



L'usage traditionnel et erroné reste largement en vigueur chez les professionnels comme le grand public (le document que vous lisez ne fait pas exception), même si c'est en contradiction avec les recommandations qui définissent clairement d'autres préfixes. L'usage des préfixes binaires reste très confidentiel et ne se répand presque pas dans le langage courant, alors que les valeurs représentées par ces unités en puissance de deux sont très utilisées dans les applications, notamment les systèmes d'exploitation. Cependant, leur utilisation commence à se répandre.

Cette distinction entre préfixes binaires et décimaux est nécessaire, car la confusion est utilisée depuis longtemps par les fabricants de disques durs. Le fait que l'usage (pour une même capacité) de préfixes en puissances de dix permette d'afficher commercialement des valeurs supérieures à celles données par les puissances de deux peut introduire une erreur d'appréciation de la part d'utilisateurs non avertis. Ainsi, un disque dur de 2 To (2.10^{12} octets) contient le même nombre d'octets qu'un disque de 1,819 Tio ($1{,}819 \times 2^{40}$ octets).

Table 01

Interview audiovisuelle de Serge Abiteboul, directeur de recherche à l'Institut national de recherche en informatique et en automatique (Inria).

Table 02

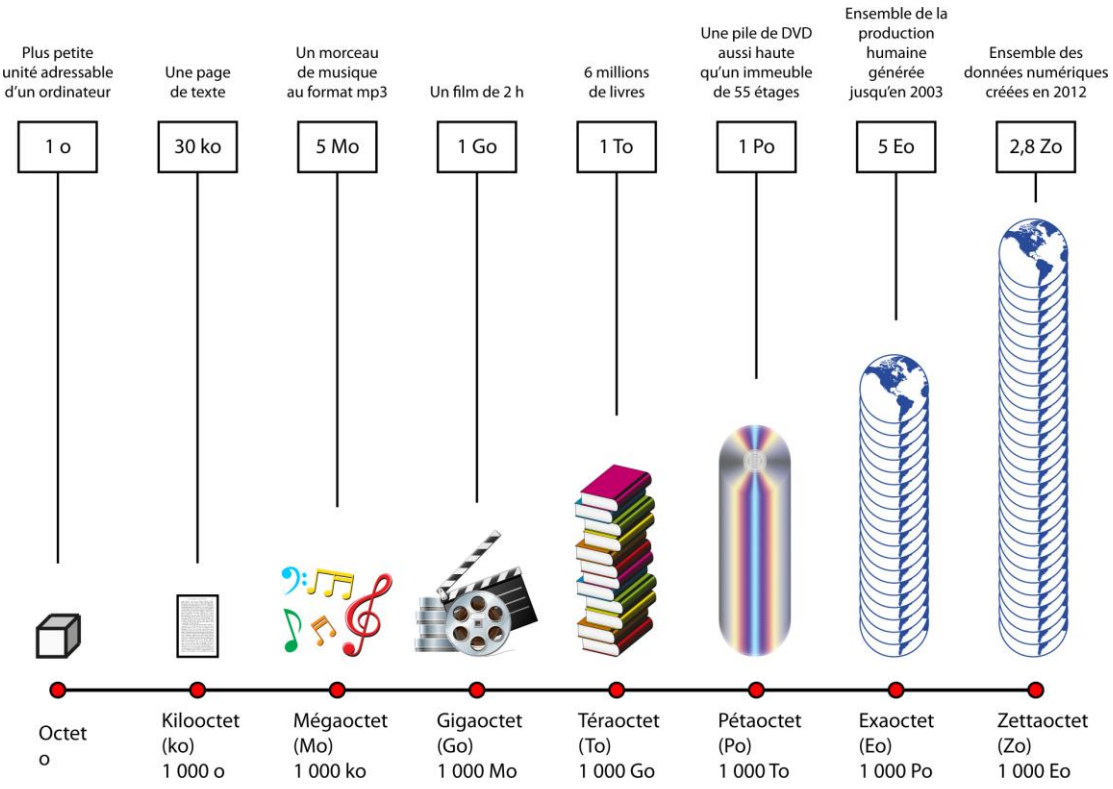
Revenons à la « règle » des 3V qui caractérisent le Big Data.

V comme… variété des données

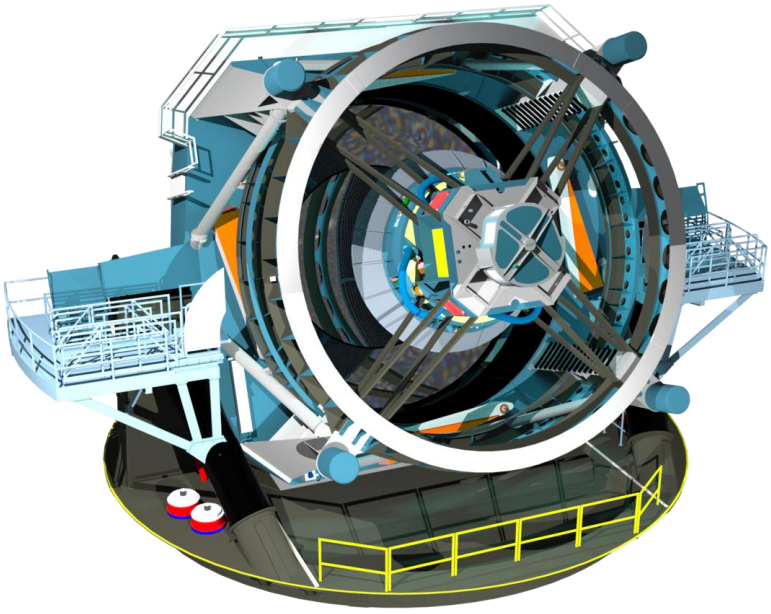
Une donnée est la représentation d'une information sous une forme qui permettra de la stocker, de la transmettre, de la transformer et de l'analyser. Dans le domaine informatique, on parle de données numériques, c'est-à-dire des données compréhensibles par un ordinateur. Qu'elles soient structurées dans des tableaux répertoriant, par exemple, des personnes avec noms, dates de naissance, numéros de téléphone etc., ou non-structurées comme les images, les vidéos ou les textes qui circulent par les réseaux, la variété de ces données s'enrichit à mesure que se développent de nouveaux outils de collecte. Web, smartphones, objets connectés, les capteurs se multiplient dans notre quotidien et entreprennent de numériser le monde réel en y prélevant des données de tous types. L'élément d'exposition présent sur cette table vous permettra d'évaluer la quantité de dioxyde de carbone que vous rejetez en expirant, de capter votre température et de mesurer votre fréquence cardiaque.

V comme… volume des données

La quantité de données générée dans le monde est en pleine expansion et suit une loi quasi exponentielle. Les chiffres parlent d'eux-mêmes. Nous générons aujourd'hui en une journée plus de données qu'il n'en a été produit entre les débuts de l'humanité et l'an 2000. Chaque seconde, plus d'une heure de vidéos est téléchargée vers le site web d'hébergement YouTube et plus de 1,5 million de courriers électroniques sont envoyés à travers le monde.



Les scientifiques ne sont pas en reste. En seulement huit ans (de 2000 à 2008), le *Sloan Digital Sky Survey* (SDSS), un programme de relevés des objets célestes utilisant un télescope dédié de 2,5 mètres de diamètre au Nouveau-Mexique, a enregistré 140 térabytes d'images, soit 140 000 gigaoctets. Son successeur, le *Large Synoptic Survey Telescope* (LSST) collectera la même quantité de données tous les cinq jours ! Ce télescope, dont le miroir primaire possédera un diamètre de 8,4 mètres, devrait voir sa première lumière dans les montagnes chiliennes en 2019 et sera muni d'une caméra CCD de 3,2 gigapixels. Le plus puissant accélérateur de particules construit à ce jour, le *Large Hadron Collider* (LHC), produit, après filtrage, 25 pétaoctets de données (25 millions de gigaoctets) chaque année soit l'équivalent de plus de trois millions de DVD. Enfin, le projet le plus ambitieux est le *Square Kilometre Array* (SKA), un réseau de plusieurs dizaines de milliers d'antennes réparties entre l'Australie et l'Afrique du Sud, qui sera équivalent à un seul radiotélescope doté d'une surface collectrice de 1 km^2. SKA devrait être construit en deux phases. À la fin de la première phase en 2023, alors que le radiotélescope ne sera exploité qu'à 10 % de ses capacités, ce ne seront pas moins de 157 téraoctets de données brutes – soit l'équivalent de plus de 30 000 DVD – que les superordinateurs devront traiter… chaque seconde !



Vue d'artiste du Large Synoptic Survey Telescope. Crédit : LSST.

Tous ces nombres sont impressionnants et donnent le vertige. Est-ce toujours le cas lorsqu'on les compare à la quantité d'informations générée par le vivant ? L'information contenue dans l'ADN des êtres vivants possède la plus forte concentration d'informations connue à ce jour. Les calculs montrent qu'en théorie, un seul gramme d'ADN pourrait stocker environ 400 exaoctets de données, soit 400 milliards de gigaoctets !

Table 03

V comme… vitesse des données

Accumuler beaucoup d'informations est une chose mais encore faut-il savoir les traiter à temps pour qu'elles restent pertinentes. La vitesse est donc l'un des facteurs d'émergence de ce nouveau rapport aux données numériques et se caractérise par l'utilisation de technologies de pointe pour obtenir des performances inédites. Produites de plus en plus rapidement, transmises à la vitesse de la lumière dans des fibres optiques et traitées par des ordinateurs de plus en plus puissants, les données sont ainsi transformées en informations exploitables à une vitesse proche du temps réel. Le risque pour l'Homme est de perdre une grande partie de la maîtrise du système quand, dans le cas des transactions boursières, les opérateurs sont des algorithmes informatiques capables de lancer des ordres d'achat ou de vente en quelques microsecondes, sans disposer de tous les critères pertinents d'analyse pour le moyen et long terme.

On emploie diverses unités de mesure pour quantifier la puissance de calcul d'une machine et déterminer ses performances de manière universelle. Parmi ces unités : les FLOPS, ou « opérations en virgule flottante par seconde » (en anglais : *FLoating-point Operations Per Second*). Un téléphone d'un gigaFLOPS peut réaliser un milliard d'opérations sur des valeurs non entières en une seconde.

La puissance de calcul des ordinateurs a tellement évolué que les machines personnelles rivalisent aisément avec les machines les plus imposantes d'il y a quelques années. Le système de guidage de la fusée Apollo 11, en 1969, n'était que deux fois plus puissant que la console grand public Nintendo Entertainment System (NES) commercialisée 16 ans plus tard.

Voici quelques comparaisons particulièrement révélatrices.

1 superordinateur Cray-2 de 1985 (1,9 gigaFLOPS) **≈ 1** Apple iPhone 4 de 2010 (1,6 gigaFLOPS)

1 Samsung Galaxy S6 de 2015 (34,8 gigaFLOPS) **≈ 5** PlayStation 2 de 2000 (6,2 gigaFLOPS chacune)

1 ordinateur de bureau typique de 2017 (90 gigaFLOPS) **≈ 8** superordinateurs Deeper Blue de 1997 (11,4 gigaFLOPS). Cet ordinateur, spécialisé dans le jeu d'échecs, avait battu le champion du monde de l'époque Garry Kasparov.

1 superordinateur Sunway TaihuLight de 2016 (93,0 pétaFLOPS) **≈ 22 100** PlayStation 4 Pro de 2016 (4,2 téraFLOPS chacune). Ce supercalculateur chinois est le plus puissant au monde en mars 2017.

Unité centrale du supercalculateur Cray-2 de 1985 (à gauche) et son système de refroidissement (à droite) exposés à l'École polytechnique fédérale de Lausanne (EPFL). Crédit : Rama / English Wikipedia.

V comme… véracité, valeur, vert

Que ce soit dans la collecte des données, dans la manière de les recouper, de les croiser, il devient de plus en plus crucial de veiller à la véracité des données, leur précision, leur pertinence par rapport au domaine dans lequel elles sont étudiées. Mais une donnée isolée n'a aucune valeur. C'est dans l'accumulation et le recoupement que réside la valeur des données. Aujourd'hui, c'est la création de valeur (économique, scientifique, culturelle) qui est le principal moteur du développement d'outils. De plus, la dimension écologique des données ne doit pas être négligée. En effet, les premiers ordinateurs, tel l'ENIAC, consommaient autant d'électricité qu'une petite ville. Aujourd'hui, un téléphone est plus puissant qu'eux, mais il ne consomme presque rien. Toutefois, les ordinateurs sont nombreux. Leur consommation globale est donc importante. Ils sont responsables de 2 % des émissions de dioxyde de carbone, ce qui est comparable aux émissions du transport aérien. S'ils participent au problème, ils aident aussi à chercher des solutions : les simulations numériques sont essentielles pour comprendre les dérèglements climatiques.
Quelques chiffres : en 2010, Google utilisait 900 000 serveurs répartis dans différents centres de données à travers le monde. Leur consommation globale s'élevait à 260 millions de watts, comme une ville française de 430 000 habitants.

Finalement, en quoi les techniques du Big Data diffèrent-elles des techniques traditionnelles d'analyse des bases de données ? La différence est de nature conceptuelle. La fouille de données classique s'appuie sur un modèle, ce qui revient à adopter un raisonnement déductif. Au contraire, le Big Data consiste à chercher, par induction, des modèles prédictifs dans de grands volumes de données à faible densité en information.

Le Centre de données de l'Utah (Utah Data Center) est un centre de stockage et de traitement de données géré par la NSA (National Security Agency) pour le compte de la Communauté du renseignement des États-Unis. La capacité de stockage de ce centre est sujette à débat, l'information étant, bien évidemment, confidentielle. Des estimations font état de quelques exaoctets…. à un millier de zettaoctets (ce que l'on appelle un yottaoctet). En l'état actuel de nos connaissances, cette dernière valeur est peu crédible des points de vue technologique, énergétique et financier.

Table 04

Un peu d'histoire

Les données sont la matière première d'une « révolution numérique » qui nous touche aujourd'hui très profondément. Mais elles ne sont pas une nouveauté. Elles sont inscrites dans une longue histoire humaine et multiculturelle. Il en est de même de l'algorithmique et de la statistique, ces sciences grâce auxquelles les données prennent sens pour nous aider à nous représenter le monde, à le comprendre et à agir.

Cet élément d'exposition propose un survol historique en trois périodes :

- de Sumer au XVI^{e} siècle avec l'apparition des données, de l'algorithmique et des premiers instruments de capture et de diffusion de l'information ;
- du XVII^{e} au XIX^{e} siècle avec la naissance et la montée en puissance de la statistique ;
- du XX^{e} siècle à nos jours avec l'apparition de l'ordinateur et la révolution informatique et numérique.

II.2.3 Les données, comment les traite-t-on ?

Pour rendre intéressantes les données captées et stockées en quantité énorme, l'informatique et l'algorithmique sont indispensables à l'extraction de l'information. L'une des grandes forces des technologies des données réside dans leur capacité scientifique et technique à croiser entre eux d'immenses volumes de données inaccessibles au traitement humain.

Table 05

Interview audiovisuelle de Françoise Soulié-Fogelman, professeur d'informatique à l'université de Tianjin (Chine).

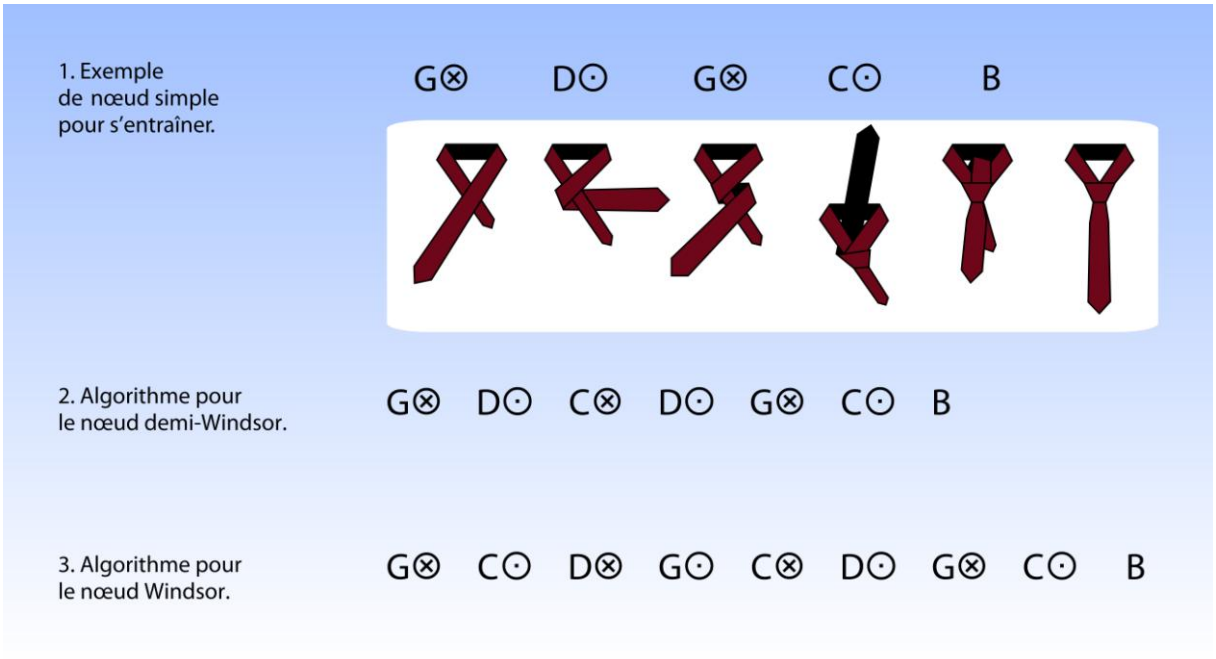
Table 06

Un algorithme est une méthode systématique permettant d'arriver à un résultat en un nombre fini d'opérations. Si les premiers algorithmes datent de plusieurs milliers d'années, le mot *algorithme* provient du nom de l'érudit perse al-Khwârizmî (v. 780 – v. 850), qui, dans son livre *Abrégé du calcul par la restauration et la comparaison*, pose les fondations de l'algèbre et expose un ensemble de méthodes de résolution des équations du second degré.

On peut écrire des algorithmes pour systématiser des actions comme nouer ses lacets, chercher un mot dans le dictionnaire, trier des objets, situer des villes sur une carte, multiplier deux nombres, extraire une racine carrée… ou encore faire un nœud de cravate ! En informatique, l'algorithme est traduit en langage de programmation et devient ainsi un programme qui sera exécutable par une machine. Pour une même tâche à accomplir (un tri, un calcul), plusieurs algorithmes sont possibles. Certains sont plus gourmands en calculs et en étapes que d'autres et il s'agit aussi de chercher la méthode la plus rapide pour arriver au résultat attendu.

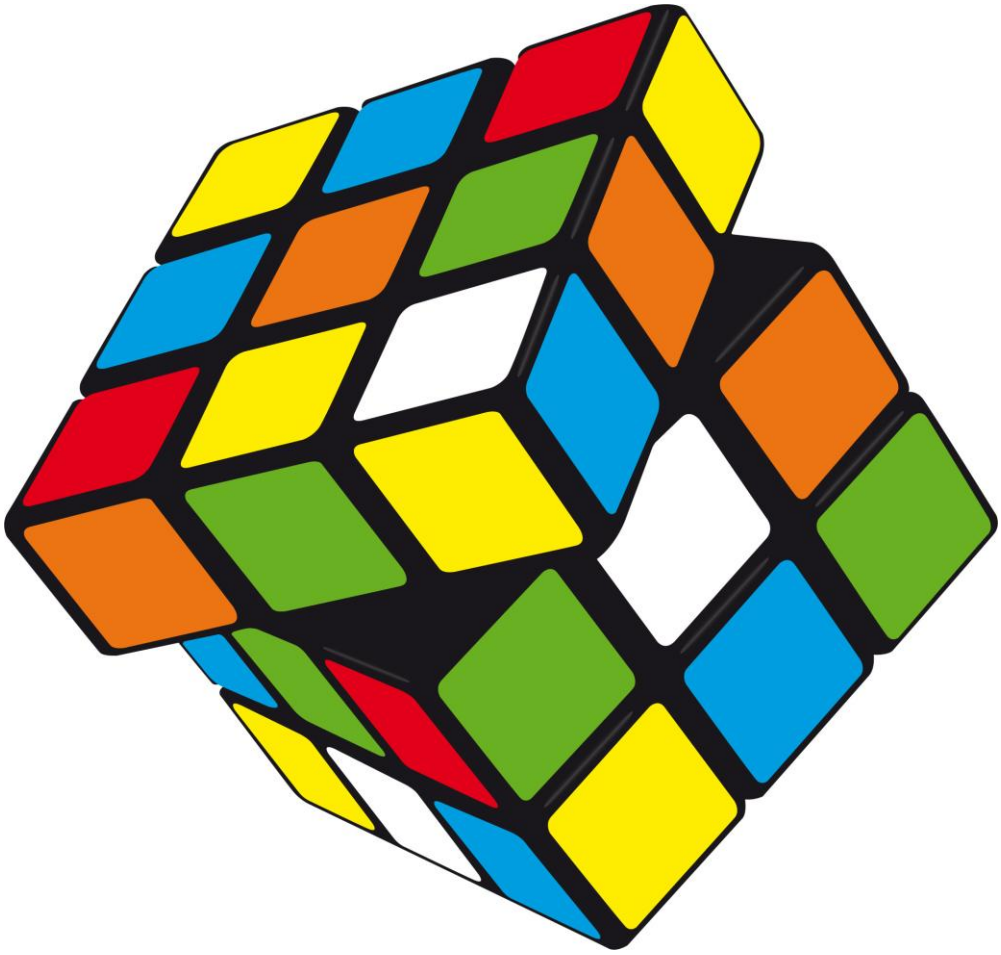
Ceux qui portent une cravate la nouent en exécutant quasi machinalement une série de gestes. Combien de nœuds différents peut-on réaliser ? En 1999, Thomas Fink et Yong Mao du laboratoire Cavendish de l'université de Cambridge ont publié le petit article *Designing tie knots by random walks* dans la prestigieuse revue *Nature*, où ils décomposent l'opération en gestes décrits et notés par des signes conventionnels. Cette notation leur permet de générer des algorithmes pour réaliser 85 nœuds différents. Ils retrouvent les quatre nœuds les plus populaires : le nœud simple, le Windsor, le demi-Windsor et le Pratt.

Ici, exercez-vous à faire un nœud de cravate en manipulant seulement son extrémité la plus large.



Faites votre nœud.

À titre d'exemple, les lignes qui suivent décrivent deux algorithmes non présentés dans l'exposition *Terra Data*. Il s'agit de l'**algorithme d'Euclide** et l'**algorithme de Nelder-Mead**. Le premier permet de déterminer le plus grand commun diviseur (PGCD) de deux entiers sans connaître leur factorisation ; le second, la résolution des problèmes d'optimisation de fonctions non linéaires. Son application à la détermination de la plus petite distance entre deux orbites planétaires dans un espace tridimensionnel est largement inspirée du chapitre 69 *The Simplex method and the least distance between two planetary orbits* de l'ouvrage *More Mathematical Astronomy Morsels* de Jean Meeus, paru en 2002 aux éditions Willmann-Bell.



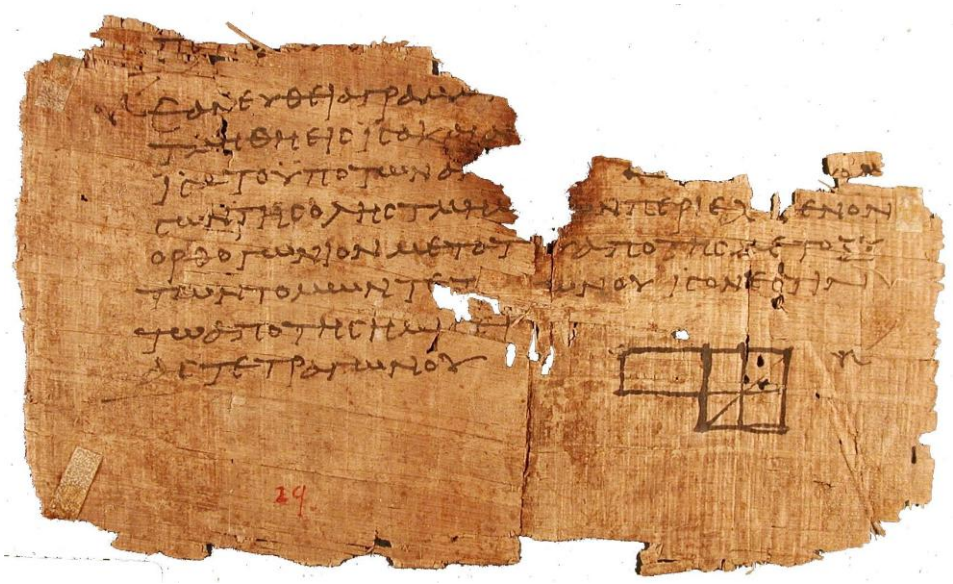
Avez-vous joué avec le Rubik's cube ?
Il existe plusieurs méthodes systématiques ou algorithmes pour en venir à bout.

L'algorithme d'Euclide

L'algorithme d'Euclide, formulé géométriquement par le grand mathématicien il y a 2 300 ans, utilise le fait que le PGCD de deux entiers ne change pas lorsqu'on remplace l'entier le plus grand par la différence entre lui et l'entier le plus petit. Ainsi, le PGCD de 252 et de 105 est également le PGCD de 105 et de 252 – 105 = 147. Puisque ce remplacement réduit le plus grand des deux nombres, la répétition du processus donne des paires de plus en plus petites, jusqu'à ce que les deux nombres soient égaux. Cette valeur commune est alors le PGCD, ici 21. L'algorithme d'Euclide est le suivant : Soit deux entiers *a* et *b* avec *a* > *b*. On commence par calculer le reste de la division de *a* par *b*, qu'on note *r*. On remplace ensuite *a* par *b* puis *b* par *r* et on réapplique le procédé depuis le début. On obtient ainsi une suite qui vaut 0 à un certain rang. Le PGCD cherché est le terme précédent de la suite. Exerçons-nous sur deux exemples et cherchons tout d'abord le PGCD de 782 et 221.

782 = 3 x 221 + 119
221 = 1 x 119 + 102
119 = 1 x 102 + **17**
102 = 6 x 17 + 0. Le PGCD de 782 et 221 est 17.

Cherchons ensuite le PGCD de 1 327 et 753.
1 327 = 1 x 753 + 574
753 = 1 x 574 + 179
574 = 3 x 179 + 37
179 = 4 x 37 + 31
37 = 1 x 31 + 6
31 = 5 x 6 + **1**
6 = 6 x 1 + 0. Le PGCD de 1 327 et 221 est 1. Ces deux entiers sont donc premiers entre eux.



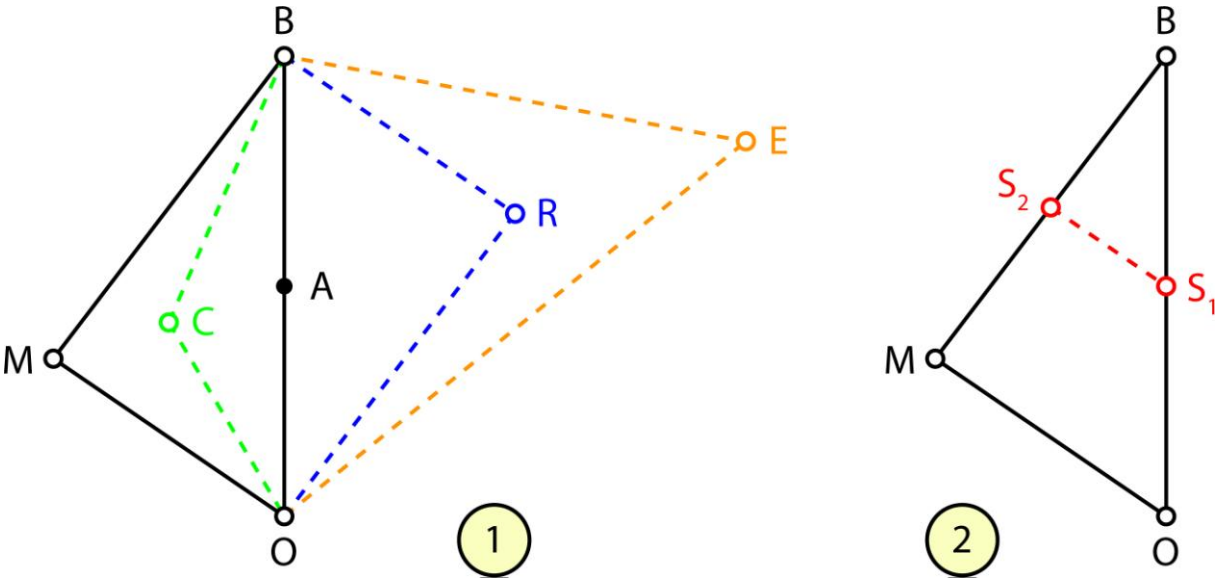
Un des plus anciennes versions connues des *Éléments*, découverte en 1896-97 sur le site d'Oxyrhynque en Égypte par Bernard Grenfell et Arthur Hunt. Ce fragment de papyrus daterait de la fin du I^er^ siècle de notre ère.

L'algorithme de Nelder-Mead

L'algorithme de Nelder-Mead a été décrit par John Nelder et Roger Mead dans l'article *A Simplex Method for Function Minimization* paru en 1965 dans la revue à comité de lecture *The Computer Journal*. Il fut popularisé par le numéro de mai 1984 du magazine informatique *Byte* grâce à l'article *Fitting Curves to Data. The Simplex algorithm is the answer*. En voici le principe.

Soit *f* une fonction de deux variables indépendantes *x* et *y*. **Pour quelles valeurs de *x* et *y* la fonction passe-t-elle par un minimum ?**

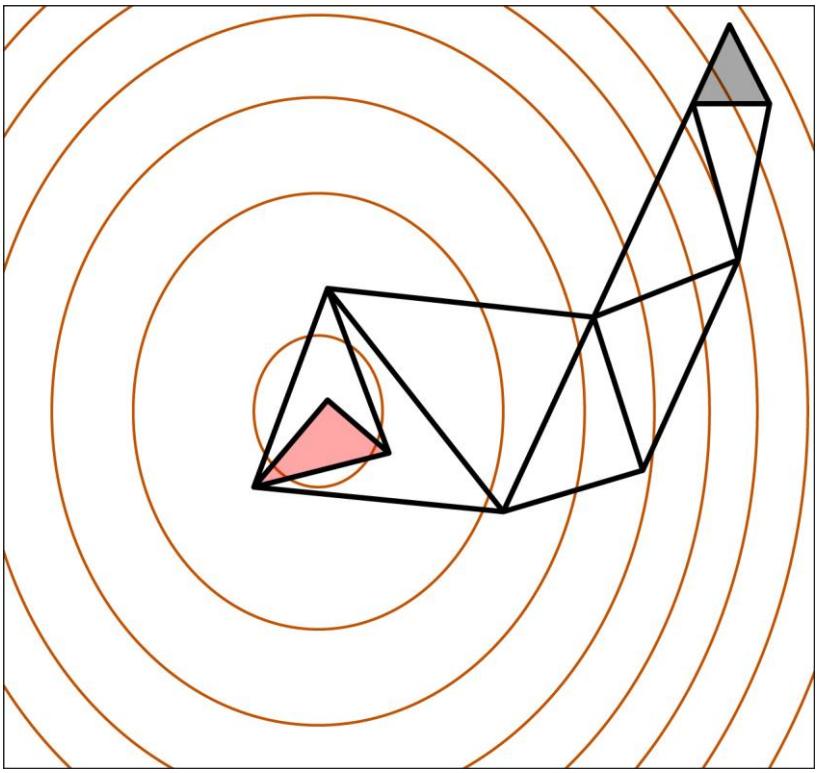
On commence par calculer la valeur de la fonction *f* pour trois points (x_1, y_1), (x_2, y_2) et (x_3, y_3) dont on pense qu'ils pourraient être assez proches du couple recherché. Une estimation grossière est suffisante puisque l'algorithme de Nelder-Mead convergera, de toute façon, vers un minimum. **Les trois points (x_1, y_1), (x_2, y_2) et (x_3, y_3) définissent un triangle dans le plan *x*-*y***. Pour atteindre le minimum de la fonction *f*, l'algorithme déplace le triangle dans le « sens de la descente », l'accélérant, le ralentissant ou le déformant si besoin est.
La première étape de l'algorithme consiste à repérer, parmi les trois couples (x_i, y_i), celui qui donne la valeur la plus faible à la fonction *f* (le point *B* de la figure ci-dessous) et celui qui lui donne la valeur la plus élevée (le point *M*). Comme on cherche le minimum de *f*, le couple donnant la valeur la plus élevée – donc le point *M* – doit être rejeté. On lui substitue un nouveau couple qui correspond à l'application de l'une des quatre transformations suivantes au sommet *M* : symétrie centrale, extension, contraction ou rétrécissement.



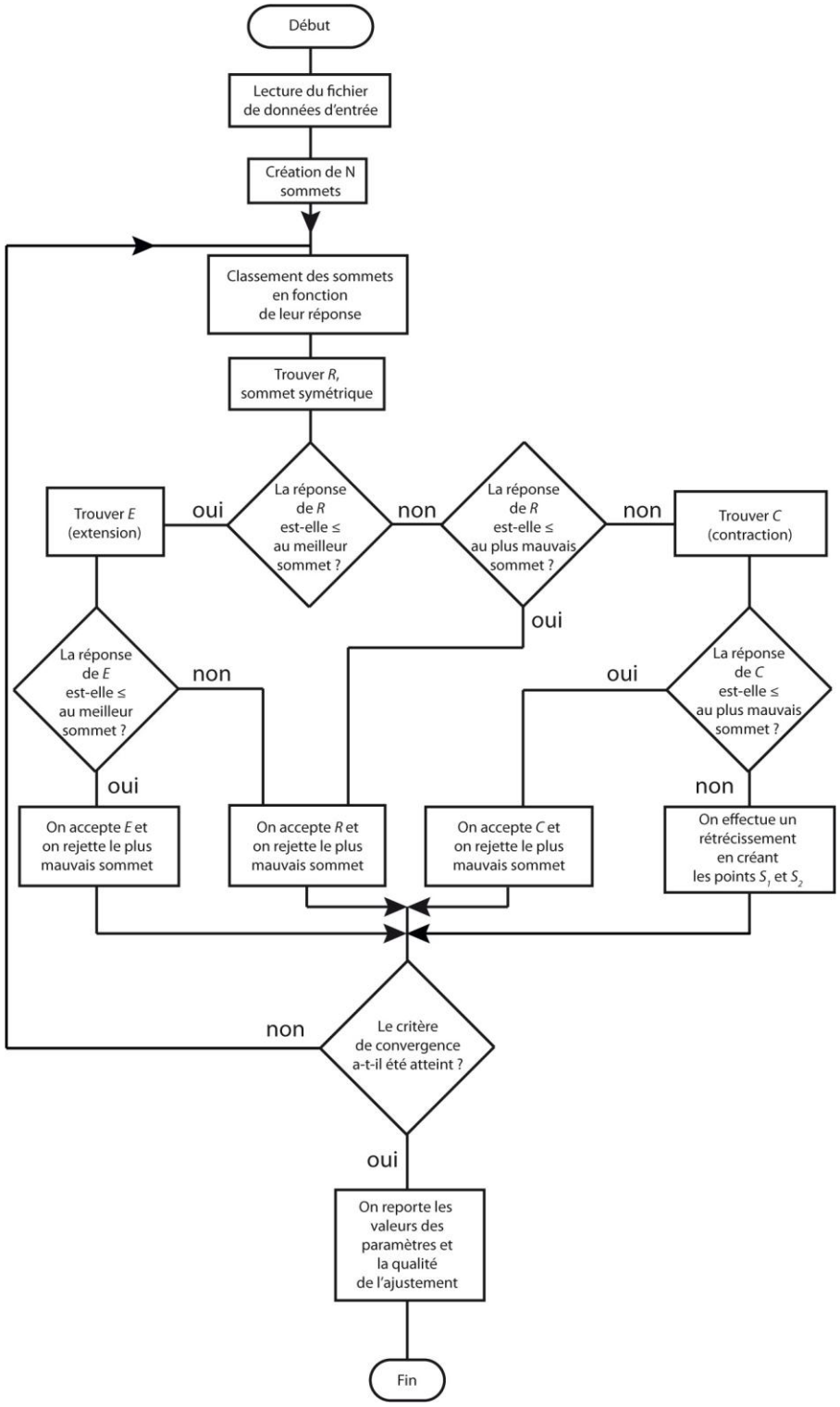
BMO est notre triangle de départ. *B* est le meilleur sommet. Parmi les trois couples (x_i, y_i), il représente celui qui donne la valeur la plus faible à la fonction *f*. *M* est le plus mauvais sommet et c'est lui qui est rejeté. Le schéma illustre les quatre transformations de l'algorithme de Nelder-Mead : ① *R* est le symétrique de *M* par rapport à *A* (milieu du segment [OB]), *E* le sommet étendu et *C* le sommet contracté ② S_1 et S_2, les sommets rétrécis.

On construit le point *R*, symétrique du point *M* par rapport à *A*, le milieu du segment [OB]. On teste alors le couple (x_R, y_R) et s'il donne une réponse qui n'est ni meilleure que le couple (x_B, y_B), ni plus mauvaise que le couple (x_M, y_M), il est conservé.
S'il donne une réponse meilleure (donc une valeur plus faible) que le couple (x_B, y_B), on considère que la direction trouvée est la bonne et l'on construit le point *E*, symétrique de *A* par rapport à *R*. On parle d'extension. Le point *E* est conservé s'il donne une réponse meilleure que le point rejeté *M* ; sinon, on garde *R*.
Si le point *R* donne une réponse plus mauvaise que le point *M*, on teste une contraction en construisant le point *C*, milieu du segment [MA]. Là aussi, *C* est conservé s'il donne une réponse meilleure que le point rejeté *M* ; sinon, on effectue un rétrécissement en substituant aux points *O* et *M* les milieux respectifs S_1 et S_2 des segments [OB] et [MB].

La figure suivante montre l'exemple d'un triangle se déplaçant sur les lignes de niveau d'une fonction. Le triangle initial est grisé. Après huit transformations, on le retrouve déformé sous la forme du triangle rose près du minimum.



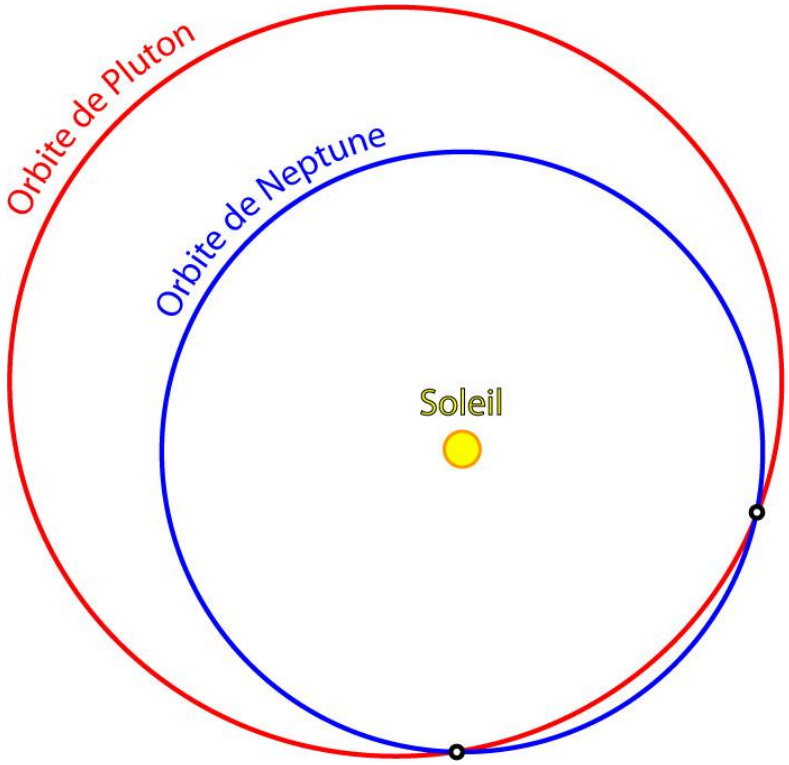
Un inconvénient de la méthode classique de Nelder-Mead est qu'elle peut aboutir à un minimum local. Pour des fonctions ayant plusieurs minima, le point vers lequel l'algorithme converge dépend des points initiaux et de la taille initiale du triangle.



Organigramme de programmation de l'algorithme de Nelder-Mead.

Il y a différentes façons d'exprimer le critère de convergence permettant de mettre un terme à l'algorithme. Dès le départ, on peut se donner un nombre maximal d'itérations, *i*. On peut également se donner une aire prédéfinie *j* ; l'algorithme s'arrête une fois que le triangle possède une surface qui lui est inférieure. On peut enfin se dire que la convergence est satisfaisante lorsque *k* évaluations successives de la fonction diffèrent de moins d'un seuil prédéfini *l*.

Dans son ouvrage *More Mathematical Astronomy Morsels*, Jean Meeus applique l'algorithme de Nelder-Mead au calcul de la plus petite distance entre l'orbite de la Terre et celle de l'astéroïde (3838) Épona, petit corps de quelques kilomètres à l'orbite très excentrique découvert en 1986. Nous l'appliquerons au calcul de la plus petite distance entre les orbites de la planète Neptune et celle de la planète naine Pluton. **Attention, nous parlons bien de la distance minimale entre les deux orbites et pas de la distance minimale entre les deux corps !**



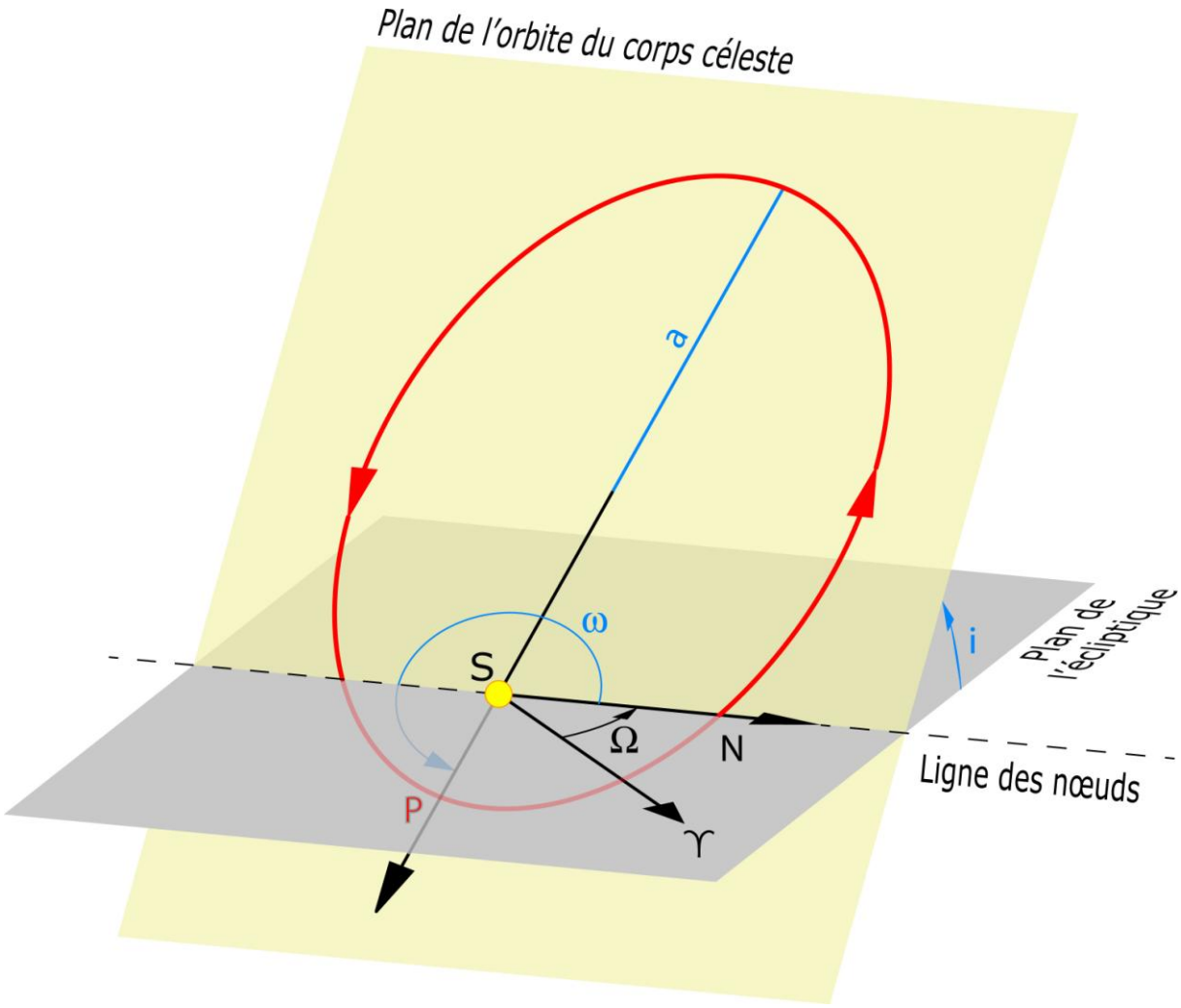
Projection des orbites de Neptune et de Pluton autour du Soleil sur le plan de l'écliptique.

Comme le montre le schéma ci-dessus, les orbites de Neptune et de Pluton autour du Soleil se croisent en deux points lorsqu'on les projette sur le plan de l'écliptique – le plan dans lequel la Terre tourne autour du Soleil. La collision entre les deux corps célestes est-elle seulement possible ?

Comment s'attaquer à ce problème ?

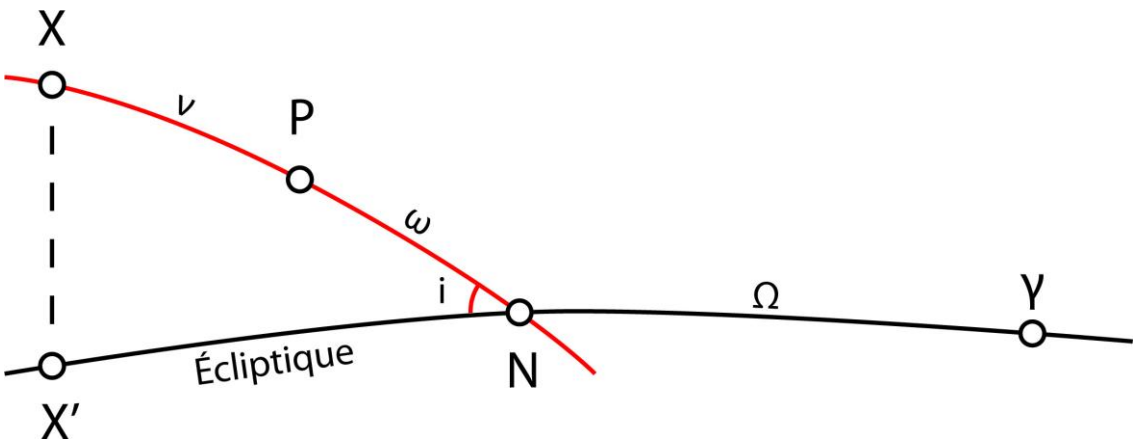
Il convient tout d'abord de caractériser correctement les orbites elliptiques des deux corps autour du Soleil, c'est-à-dire de préciser leur forme et leur orientation dans l'espace. On se donne donc un plan de référence, le plan de l'écliptique déjà présenté quelques lignes plus haut. On cherche ensuite dans la littérature spécialisée une série d'éléments appelés *éléments orbitaux*, qui permettent de préciser, à une date donnée, l'orbite elliptique des corps considérés par rapport à ce plan de référence. Ils sont au nombre de cinq :

- le **demi-grand axe** *a*, qui n'est autre que la moitié du plus long diamètre de l'ellipse. C'est une mesure de la taille absolue de l'ellipse et dans le système solaire, on l'exprime volontiers en *unité astronomique* (ua), une unité de distance s'élevant à 149 597 870,7 kilomètres ;
- l'**excentricité** *e*, qui est une mesure de l'aplatissement de l'ellipse ($0 \leq e < 1$) ;
- l'**inclinaison** *i* du plan considéré sur le plan de l'écliptique. Si *i* est compris entre 0° et 90°, le mouvement est dit *direct*. S'il est compris entre 90° et 180°, il est dit *rétrograde* ;
- la **longitude du nœud ascendant** *Ω*, qui est l'angle entre la direction du point vernal γ (position occupée par le Soleil à l'équinoxe de mars) et la ligne des nœuds (droite d'intersection du plan orbital avec le plan de l'écliptique) ;
- l'**argument du périhélie** *ω*, c'est-à-dire l'angle entre le nœud ascendant et le périhélie, mesuré dans le plan orbital et dans la direction du mouvement.

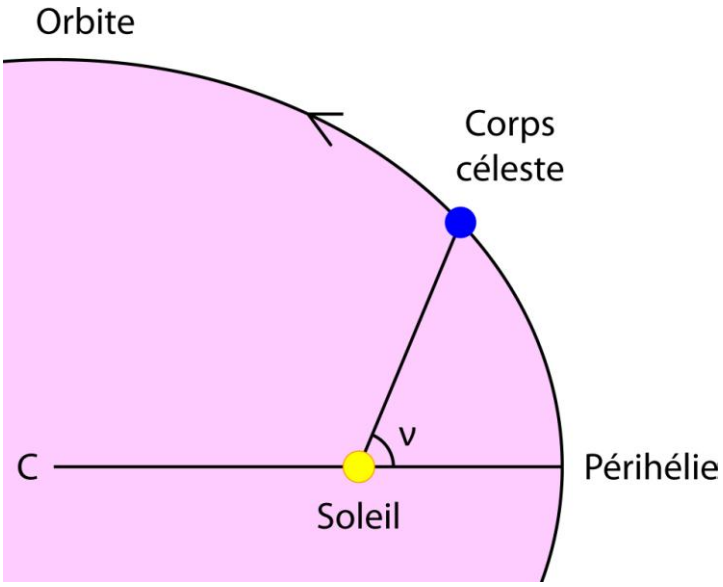


S est le Soleil. Le plan de l'écliptique, notre plan de référence, est en gris. Il porte la direction du point vernal *γ*. Le plan orbital du corps céleste est en jaune. Il est incliné d'un angle *i* par rapport au plan de l'écliptique. Les deux plans se coupent selon une droite, la ligne des nœuds. Dans le plan orbital est représentée, en rouge, l'ellipse décrite par le corps. On note *a* son demi-grand axe, *P* le périhélie et *ω* l'argument du périhélie.

Dans la figure suivante, l'arc *γNX'* représente une partie de l'écliptique vu depuis le Soleil. Il s'agit de l'intersection du plan de l'écliptique avec la sphère céleste. *NPX* est une partie de l'orbite du corps céleste, qui est l'intersection de son plan orbital avec la sphère céleste. *γ* est le point vernal (l'origine des longitudes), *N* le nœud ascendant de l'orbite, *P* le périhélie et *X* la position du corps céleste à un instant donné. L'arc *XX'* est une partie du grand cercle perpendiculaire à l'écliptique et passant par *X*.



L'arc *γNX'* est donc la *longitude écliptique λ* du corps céleste ; toutefois, ce n'est pas d'elle dont nous avons besoin mais de la *longitude vraie l*, qui est la somme des arcs *γN* = *Ω* et *NPX* = *ω* + *v* où *v* est l'anomalie vraie, à savoir l'angle entre la direction du périhélie et la position de l'objet sur son orbite, mesuré au foyer de l'ellipse (voir schéma ci-dessous). Notez que ces arcs sont mesurés dans deux plans différents.



Le Soleil occupe l'un des deux foyers de l'ellipse décrite par le corps céleste. L'anomalie vraie *v* est l'angle entre la direction du périhélie et celle du corps, mesuré depuis le Soleil.

Pour toutes les valeurs de la longitude vraie *l*, nous pouvons calculer les deux angles :

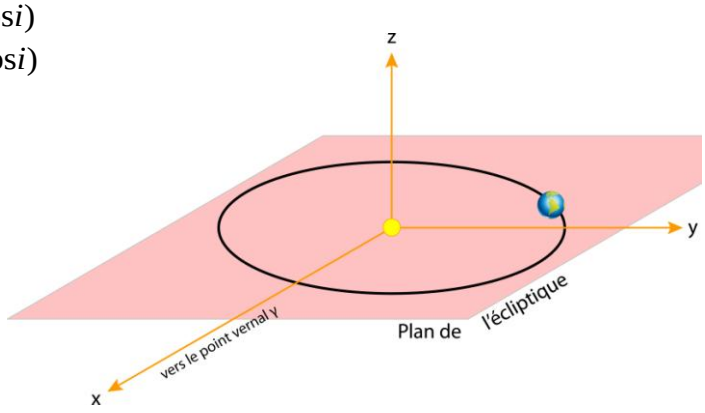
u = *l* - *Ω* = arc *NX* (argument de latitude)
v = *l* - *Ω* - *ω* = arc *PX* (anomalie vraie)

Nous avons désormais tous les outils en main pour mener à bien la mission que nous nous sommes assignée. En effet, on peut maintenant déterminer :

- le rayon vecteur *r* du corps céleste (sa distance au Soleil) grâce à la formule $r = \frac{a(1-e^2)}{1+e\cos v}$;

- les coordonnées rectangulaires de ce corps dans un référentiel héliocentrique

$$x = r(\cos\Omega\cos u - \sin\Omega\sin u\cos i)$$
$$y = r(\sin\Omega\cos u + \cos\Omega\sin u\cos i)$$
$$z = r\sin i\sin u.$$



De la même manière, si *a'*, *e'*, *i'*, *Ω'* et *ω'* sont les éléments orbitaux de l'orbite du second corps, on peut calculer *u'*, *v'*, *r'* et les coordonnées rectangulaires *x'*, *y'* et *z'* pour toute valeur de la longitude vraie *l'* sur cette orbite. Il convient, bien évidemment, d'exprimer les éléments des deux orbites au même instant.

Trouver la distance minimale entre les deux orbites revient à déterminer le couple de longitudes vraies *l* et *l'* pour lequel le carré de la distance entre les deux orbites $d^2 = (x-x')^2 + (y-y')^2 + (z-z')^2$ soit minimal. Voici donc la sympathique fonction sur laquelle nous allons travailler :

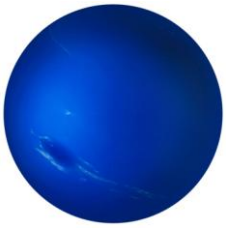
$$d^2(l,l') = \left(\frac{a(1-e^2)}{1+e\cos(l-\Omega-\omega)}(\cos\Omega\cos(l-\Omega)-\sin\Omega\sin(l-\Omega)\cos i) - \frac{a'(1-e'^2)}{1+e'\cos(l'-\Omega'-\omega')}(\cos\Omega'\cos(l'-\Omega)-\sin\Omega\sin(l'-\Omega')\cos i')\right)^2$$
$$+\left(\frac{a(1-e^2)}{1+e\cos(l-\Omega-\omega)}(\sin\Omega\cos(l-\Omega)+\cos\Omega\sin(l-\Omega)\cos i) - \frac{a'(1-e'^2)}{1+e'\cos(l'-\Omega'-\omega')}(\sin\Omega'\cos(l'-\Omega')+\cos\Omega'\sin(l'-\Omega')\cos i')\right)^2$$
$$+\left(\frac{a(1-e^2)}{1+e\cos(l-\Omega-\omega)}(\sin i\sin(l-\Omega)) - \frac{a'(1-e'^2)}{1+e'\cos(l'-\Omega'-\omega')}(\sin i'\sin(l'-\Omega'))\right)^2.$$

Puisqu'il faut bien commencer quelque part, nous considérerons d'abord des longitudes vraies égales ($l = l'$) sur les deux orbites.

Les éléments orbitaux de Neptune et de Pluton le 1er janvier 2000 sont donnés par E. M. Standish, X. X. Newhall, J. G. Williams et D. K. Yeomans dans le chapitre 5 *Orbital Ephemerides of the Sun, Moon and Planets* de l'ouvrage *Explanatory Supplements to the Astronomical Almanac*, édité par P. K. Seidelmann de l'U.S. Naval Observatory (University Science Books, 1992).

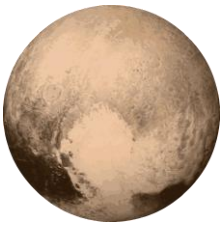
Neptune

Demi-grand axe a	30,068 963 48 ua
Excentricité e	0,008 585 87
Inclinaison i	1,769 17°
Longitude du nœud ascendant Ω	131,721 69°
Argument du périhélie ω	273,249 66°



Pluton

Demi-grand axe a'	39,481 686 77 ua
Excentricité e'	0,248 807 66
Inclinaison i'	17,141 75°
Longitude du nœud ascendant Ω'	110,303 47°
Argument du périhélie ω'	113,763 29°



Le tableau ci-dessous présente les valeurs calculées de d (en unité astronomique) correspondant à des valeurs de $l = l'$ croissant par sauts de 10°.

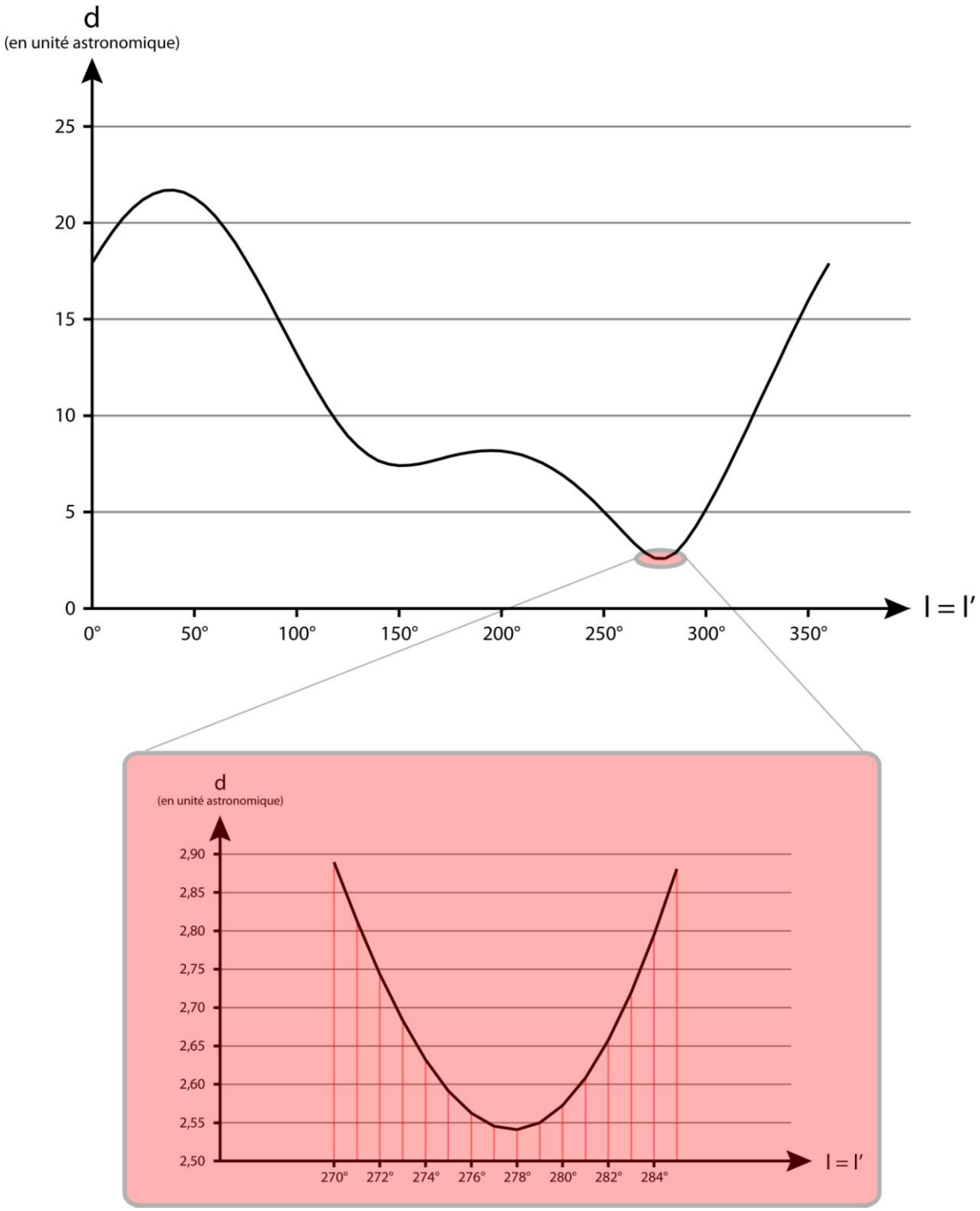
$l = l'$	0°	10°	20°	30°	40°	50°	60°	70°	80°
d (ua)	17,90	19,53	20,77	21,50	21,68	21,29	20,35	18,94	17,18

$l = l'$	90°	100°	110°	120°	130°	140°	150°	160°	170°
d (ua)	15,21	13,18	11,27	9,619	8,384	7,651	7,397	7,488	7,741

$l = l'$	180°	190°	200°	210°	220°	230°	240°	250°	260°
d (ua)	8,001	8,161	8,156	7,951	7,530	6,891	6,046	5,024	3,901

$l = l'$	270°	280°	290°	300°	310°	320°	330°	340°	350°
d (ua)	2,890	2,572	3,464	5,130	7,143	9,326	11,58	13,82	15,96

Enfin, les graphiques en page suivante montrent de façon plus détaillée l'évolution de d en fonction de $l = l'$.



D'après le graphique du haut, le minimum de distance est atteint pour *l* = *l'* compris entre 250° et 300°. Une étude plus fine (graphique du bas) place ce minimum vers 278°.

L'application de l'algorithme de Nelder-Mead nécessite le calcul de trois valeurs de la fonction d = *f*(*I, I'*) pour définir le triangle dans le plan *I* – *I'*. Soit donc les trois points A, B et C, choisis arbitrairement tels que leurs coordonnées *I* et *I'* soient proches de 278°. Les trois valeurs de *I* ne doivent pas être les mêmes, tout comme les trois valeurs de *I'*. Pour chacun des points, nous calculons la distance *d* correspondante entre les deux orbites.

A	*I* = 278°	*I'* = 278°	→	d_A = 2,540 958 791
B	*I* = 277°	*I'* = 277°	→	d_B = 2,545 228 443
C	*I* = 278°	*I'* = 277°	→	d_C = 2,549 778 353

Le triangle *ABC* est caractérisé par une amplitude en *I* de 1°, en *I'* de 1° et en *d* de 0,008 819 562 unité astronomique, soit un peu plus de 1,3 million de kilomètres.

Dans le triangle *ABC*, *A* est le meilleur sommet et *C* le plus mauvais. On construit d'abord *R*, symétrique de C par rapport à *O*, le milieu du segment [AB]. Ses coordonnées sont I_R = 277° et I'_R = 278°.

R	*I* = 277°	*I'* = 278°	→	$d_R = 2{,}646\,968\,521 > d_C$

Visiblement, *R* donne une réponse plus mauvaise que le point *C*. On teste donc une contraction en construisant le point *D*, milieu du segment [CO]. Ses coordonnées sont I_D = 277,75° et I'_D = 277,25°.

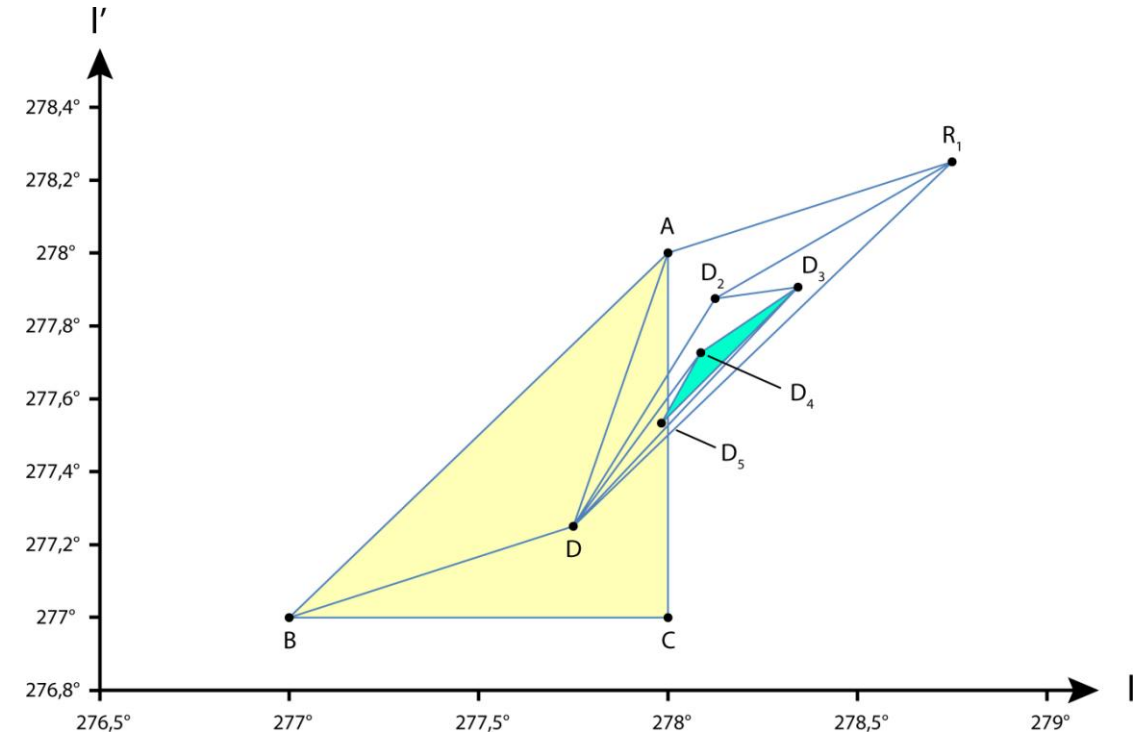
D	*I* = 277,75°	*I'* = 277,25°	→	$d_D = 2{,}531\,138\,967 < d_C$

Le point contracté *D* est donc accepté.

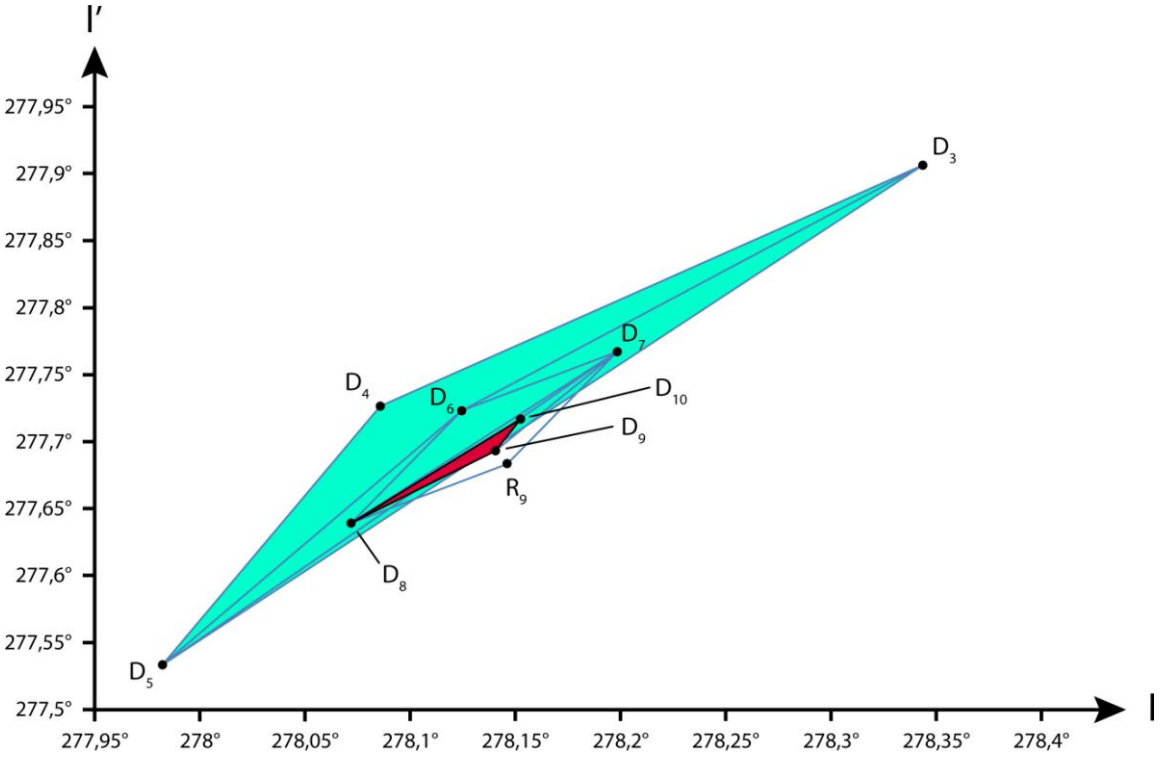
Nous nous retrouvons avec un nouveau triangle, *ABD* :

A	*I* = 278°	*I'* = 278°	→	d_A = 2,540 958 791
B	*I* = 277°	*I'* = 277°	→	d_B = 2,545 228 443
D	*I* = 277,75°	*I'* = 277,25°	→	d_D = 2,531 138 967

Les calculs, dont nous vous faisons grâce, ont ensuite été menés jusqu'à la 12e étape. Les deux graphiques suivants présentent les triangles associés.



Le triangle de départ *ABC* est en jaune. Six transformations plus tard, on obtient le triangle bleu $D_3D_4D_5$.



Un zoom a été effectué sur le triangle bleu $D_3D_4D_5$. Après six transformations, on obtient le triangle rouge $D_8D_9D_{10}$.

Le triangle final de cette étude est $D_8D_9D_{10}$.

D_8	l = 278,071 990 967°	l' = 277,639 190 674°	d_{D8} = 2,529 758 165 ua
D_9	l = 278,140 693 665°	l' = 277,693 229 675°	d_{D9} = 2,529 746 657 ua
D_{10}	l = 278,152 475 357°	l' = 277,716 711 044°	d_{D10} = 2,529 743 501 ua

Il est caractérisé par les amplitudes :
Δl = 0,080 484 390° ≈ 4’ 50’’ $\Delta l'$ = 0,077 520 370° ≈ 4’ 40’’
Δd = 0,000 014 664 ua ≈ 2194 km.

Conclusion : la distance minimale entre les orbites de Neptune et de Pluton est, selon toute vraisemblance, un peu inférieure à 2,53 unités astronomiques soit environ 380 millions de kilomètres. Sur l’orbite de Neptune, elle correspond à une longitude vraie proche de 278,1° et sur celle de Pluton, proche de 277,7°. Pas de collision possible !

En fait, pour des raisons de résonance entre les deux corps, Neptune et Pluton ne s’approchent jamais l’un de l’autre à moins de 17 unités astronomiques.

Nous venons de passer en revue les algorithmes d’Euclide et de Nelder-Mead. Il en existe, bien sûr, d’autres en très grand nombre. L’exposition *Terra Data* vous propose d’en expérimenter quelques-uns.

Table 07

Classer : les algorithmes de tri
Trier signifie « classer, répartir les différents éléments d’un ensemble en plusieurs classes, selon certains critères ». De manière plus restrictive, le terme de *tri* en algorithmique est très souvent attaché au processus de classement d'un ensemble d'éléments dans un ordre donné. Trier des nombres du plus petit ou plus grand, ranger une liste dans l’ordre alphabétique… On entend ainsi par *algorithme de tri* un algorithme procédant par comparaisons successives entre plusieurs éléments ou données. Cet élément met à votre disposition un tapis de tri vous permettant, à vos élèves et vous, de vous mettre à la place d’un ordinateur.

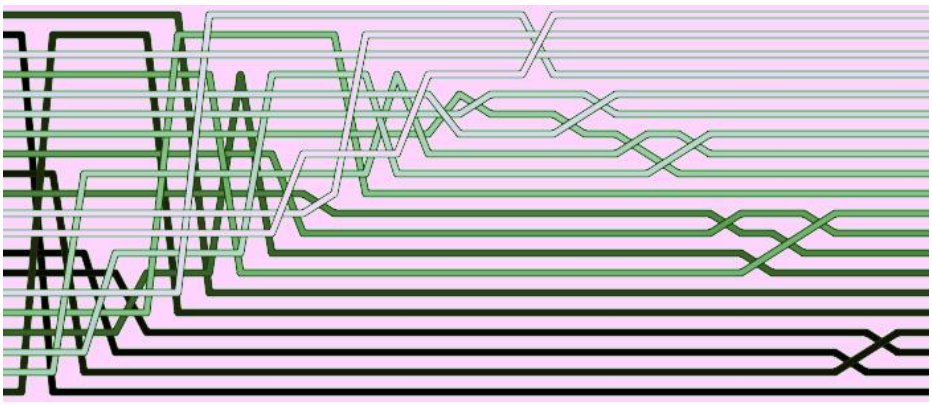


Table 08

Retrouver des données : les algorithmes d'indexation

L'indexation de données est la réponse à une question à laquelle nous avons toutes et tous été un jour confrontés : comment organiser au mieux une collection de documents afin de pouvoir plus tard retrouver facilement celui qui nous intéresse ? Face aux problèmes du déluge de données et de l'hétérogénéité croissante des documents que doivent traiter les moteurs de recherche, l'indexation automatique est une nécessité. Elle se base directement sur le contenu, dans le but d'obtenir des résultats univoques et cohérents.

Comprendre le fonctionnement d'un moteur de recherche

Nous utilisons tous les jours des moteurs de recherche qui nous orientent vers des sites internet, des articles, des réseaux sociaux. Mais comment font-ils leurs choix ? Qu'est-ce qui se cache derrière ces résultats et leur hiérarchisation ? L'utilisation quotidienne des algorithmes dans le référencement d'un moteur de recherche mérite d'être décortiquée pour en découvrir les différentes étapes et principes sous-jacents.

Table 09

Les algorithmes d'apprentissage automatique

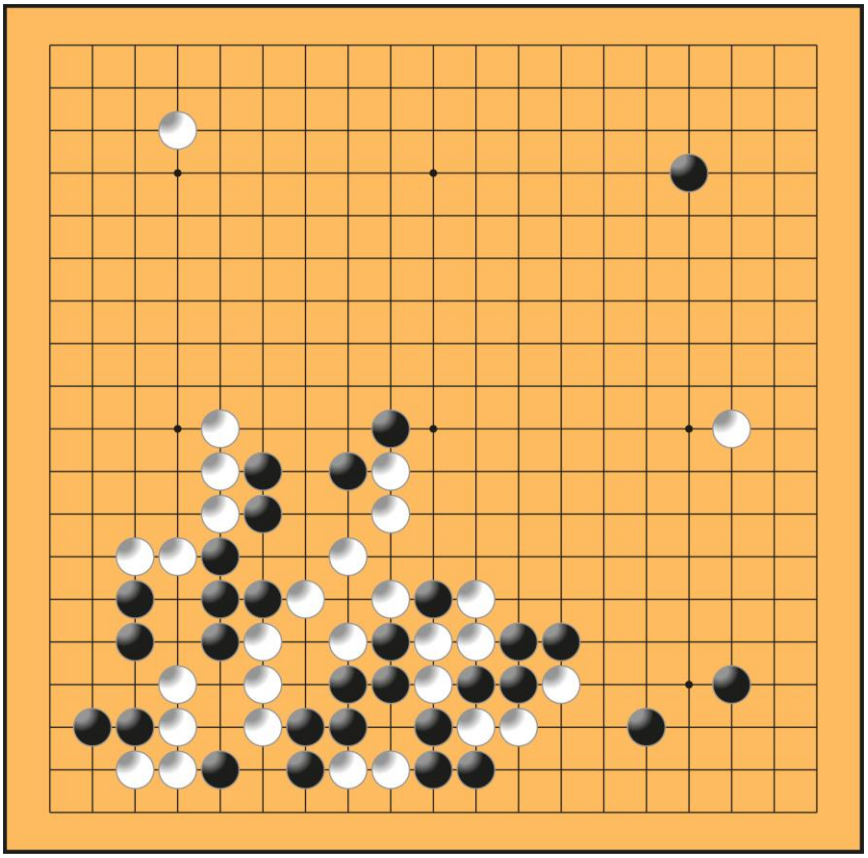
S'il existe de nombreux algorithmes pour résoudre divers problèmes (comme effectuer un tri rapide) ceux-ci fournissent toujours un résultat identique pour de mêmes entrées. Quand le problème est trop compliqué, on a recourt à des méthodes qui permettent aux machines d'apprendre à partir des données qui leur sont soumises.

Un humain n'a aucun mal à reconnaître un chien. Mais un ordinateur ? Formaliser toutes les variantes possibles du « chien » est bien trop complexe pour écrire un algorithme classique capable, par exemple, de repérer cet animal sur une image. Un algorithme d'apprentissage automatique va s'appuyer sur de nombreuses images où des personnes ont marqué la présence d'un chien. Il va apprendre à reconnaître l'animal sans passer par une définition exhaustive de « chien », un peu comme un enfant apprend à parler par imitation, sans connaître la définition exacte des mots et la grammaire. Des exemples familiers ? Votre filtre anti-spams, qui s'appuie sur vos choix pour écarter des mails indésirables, et les articles « que vous devriez aimer » sur les sites de vente en ligne.

C’est en 1997 que le superordinateur Deep Blue, et plus précisément sa version améliorée Deeper Blue, a remporté un tournoi d’échecs contre le champion du monde Garry Kasparov. Mais comment ? L’ordinateur s’est appuyé sur des parties perdues par Kasparov contre des humains. À chaque mouvement de Kasparov, Deep Blue puisait dans ce répertoire. Puis, lorsque suffisamment de pièces furent éliminées, le nombre de coups possibles a diminué. Il est devenu plus facile à Deep Blue d’envisager tous les cas possibles et de dominer la partie.

En mai 1997, le champion du monde d’échecs Garry Kasparov s’incline face à Deeper Blue, capable de calculer de 100 à 300 millions de positions par seconde. Deeper Blue défait Kasparov 3,5 à 2,5 dans un match à 6 parties.

Le jeu de go est un jeu de plateau très populaire en Asie de l’Est. Pour un ordinateur, il est plus complexe à maîtriser que le jeu d’échecs. Pourquoi ? Ceci est principalement dû au fait que le plateau de go est bien plus étendu, que la plupart des coups sont légaux et souvent plausibles, et également au fait que la capture des pions rend possible de rejouer dans les espaces ainsi libérés. Pourtant, l’algorithme développé par la société DeepMind, AlphaGo, a réussi à battre en mars 2016 le champion coréen Lee Sedol. Pour gagner, AlphaGo s’est appuyé sur l’apprentissage automatique. D’abord, il a analysé un très grand nombre de situations pour apprendre comment des experts avaient joué avec succès. Ensuite, il a joué contre lui-même des millions de parties pour explorer des situations inédites. C’est en combinant des puissances de calcul considérables et des algorithmes d’apprentissage nouveaux que le programme a vaincu ce joueur professionnel, l’un des meilleurs au monde.



Les 59 premiers coups d'une partie de go.

Les corrélations

Qu'est-ce qu'une corrélation ? En s'appuyant sur des données statistiques de l'Insee, l'animation présentée ici permet d'apprendre à ne pas confondre corrélation et causalité. Par exemple, certaines corrélations absurdes entre des phénomènes n'ayant aucun lien de causalité entre eux (cf. graphique ci-dessous) sont fallacieuses ou illusoires. Le visiteur apprend également à interpréter le sens d'une causalité et à exercer son esprit critique vis-à-vis de l'analyse de phénomènes corrélés, notamment dans le décryptage de l'actualité.

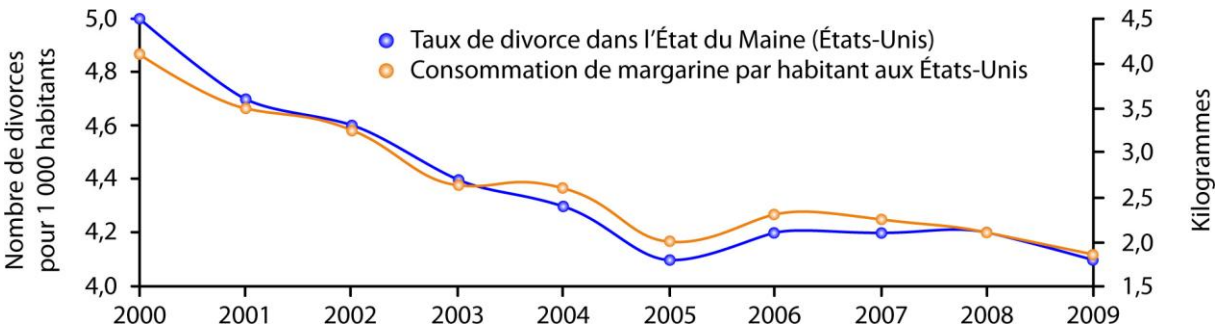


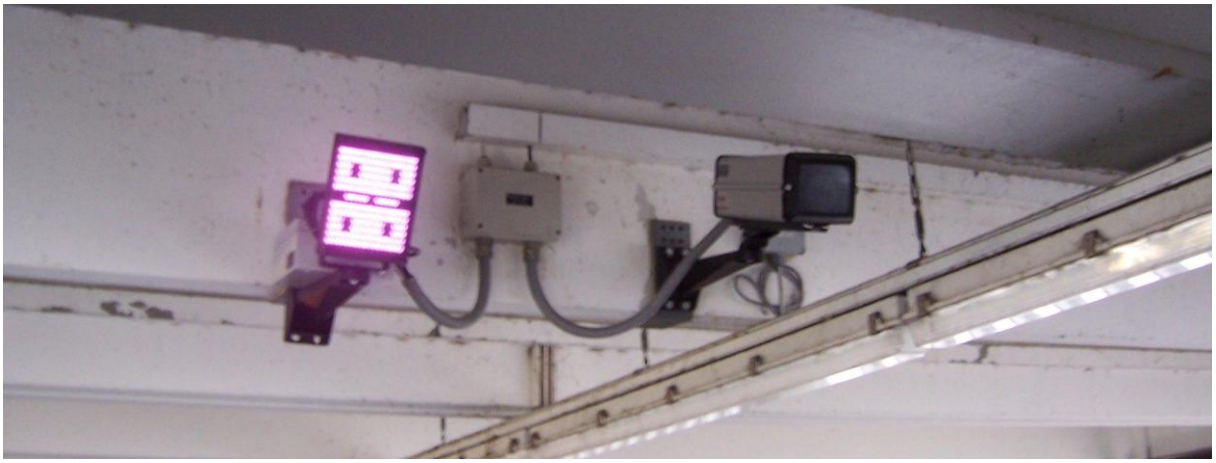
Table 10

Replica

Un humain n'a aucun mal à discriminer une forme dans une image et à la retrouver plus ou moins ressemblante dans une autre image. Mais un ordinateur ? Les technologies informatiques du *deep learning* (ou *apprentissage profond*) permettent aujourd'hui de révolutionner la lecture des images en utilisant les capacités de calculs et d'apprentissages de réseaux dits *neuronaux*. Elles permettent d'effectuer des recherches en détectant des similarités et des récurrences de motifs au sein de très grandes bases de données d'images. Elles pourraient ainsi modifier en profondeur nos connaissances en histoire de l'art en permettant de suivre la généalogie de tableaux et la dissémination historique des motifs. L'objectif de cet élément d'exposition est donc de vous montrer comment on enseigne à un système à retrouver des formes dans une base de données iconographiques – ici des tableaux vénitiens du XV^{e} siècle – et à en extraire tous les tableaux où elles apparaissent.

Reconnaissance de visage

Un système de reconnaissance faciale est une application logicielle qui reconnaît automatiquement quelqu'un grâce à son visage. Outre ses applications pour des besoins de sécurité, cette technologie est de plus en plus utilisée dans la vie courante pour valider un achat avec son smartphone grâce à un autoportrait photographique (ou *selfie*), assister le conducteur d'une voiture pour une conduite sécurisée grâce à la détection des cas de somnolence, passer la frontière avec un passeport biométrique, etc. Mais comment cela fonctionne-t-il ? Vous pouvez ici expérimenter une caméra qui permet la reconnaissance faciale en procédant par comparaisons successives et découvrir les principes de son fonctionnement algorithmique.



Système de surveillance utilisé en Suisse et en Allemagne. Il comprend un système de reconnaissance faciale, de reconnaissance de véhicule (marque, modèle, couleur) et un lecteur de plaque d'immatriculation.
Crédit : Maraparrac / English Wikipedia.

Table 11

Cette table, sans contenu, est utilisée pour la médiation humaine.

Table 12

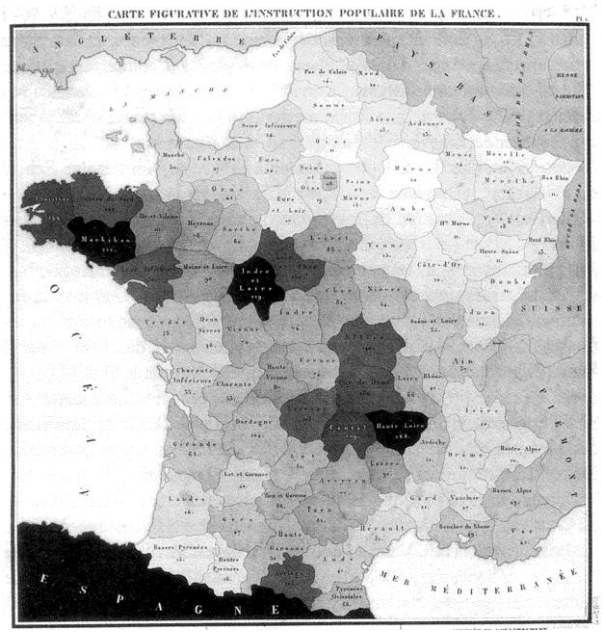
Qu'est-ce que la datavisualisation ?

Chaque mode de représentation, que ce soit une courbe, un graphe, un diagramme, une carte proportionnelle (*treemap* en anglais) a une fonction particulière : comparer, hiérarchiser, regrouper, etc. Selon le discours visé, il faut donc déterminer la meilleure transcription. Cette discipline, appelée *datavisualisation* ou *représentation graphique de données statistiques*, est née à la fin du XVIII^{e} siècle et a été théorisée et mise en pratique dès le milieu du siècle suivant. Elle s'est enrichie au fil de la numérisation croissante du monde. Ainsi, en 1869, l'ingénieur Charles-Joseph Minard a élaboré un graphe exemplaire retraçant la campagne de Russie de 1812. Vous retrouverez ce graphe, parmi d'autres, dans une composition présentée dans l'exposition.

Table 13

Les règles de la datavisualisation

Un ensemble de données complexe peut être rapidement lu et correctement compris si certaines règles de présentation sont respectées. Une mauvaise utilisation de ces règles peut conduire à faire dire aux représentations autre chose que le sens des données qu'elles étaient censées traduire. Face à l'obligation de décrypter de plus en plus de visualisations d'informations, la lecture des représentations doit être enseignée. À l'aide de quelques notions de psychologie cognitive, vos élèves comprendront comment ces règles peuvent être utilisées à bon escient ou, au contraire, détournées. Ils exerceront leur esprit critique face à l'instrumentalisation possible des modes de représentation.



La *Carte figurative de l'instruction populaire de la France* est une carte qui représente le taux d'élèves masculins scolarisés pour chaque département. Elle a été conçue en 1826 par l'ingénieur, mathématicien et homme politique français Charles Dupin (1784 – 1873) et réalisée à Bruxelles par le lithographe belge Jean-Baptiste Collon.

II.2.4 Les données, qu'est-ce que ça change ?

Le nombre de domaines concernés par le traitement des données est grand et va croissant : science, information, industrie, santé, ville, transport, commerce, travail, finance, culture et, bien sûr, liens sociaux et vie privée. Tout ceci a un impact sur nos vies et plus largement sur nos sociétés.
Les technologies des données connotent l'innovation technologique, la création de nouveaux services toujours plus personnalisés et la maîtrise du futur. D'autant plus qu'on attend de lui une connaissance plus fine du monde et une amélioration des décisions – au point de chercher à percer, voire à devancer les événements.

Table 14

Interview audiovisuelle de Dominique Cardon, sociologue.

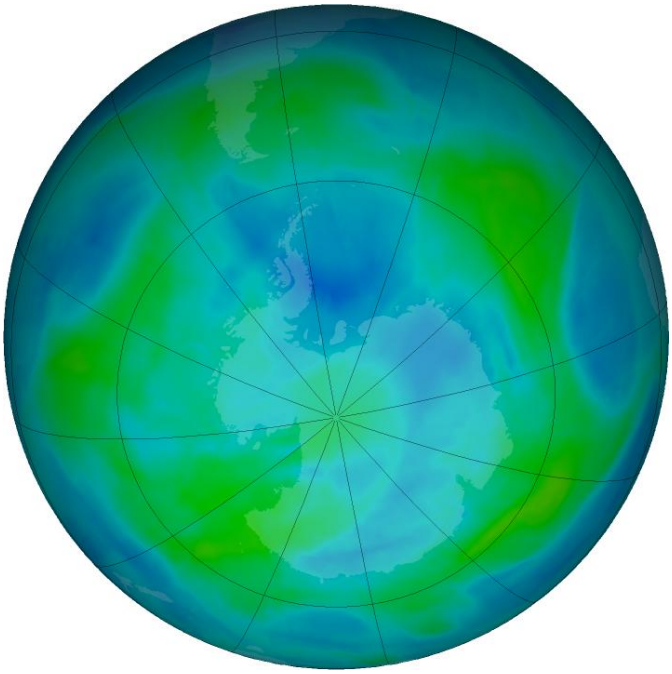
Table 15

Données et connaissances scientifiques
La récupération et le traitement des données permettent d'accroître et d'accélérer la connaissance dans des domaines comme la chimie, la biologie, la mécanique des fluides, la génomique, l'astronomie, la climatologie ou l'épidémiologie. Ces connaissances sont générées à la fois par l'analyse de données captées mais aussi par les simulations qui génèrent de nouvelles données exploitables.

Trois exemples issus de différents domaines de recherches scientifiques sont présentés ici sous forme d'audiovisuels :

- Dans le cadre du ***Blue Brain Project***, le superordinateur de l'École polytechnique fédérale de Lausanne a simulé l'activité d'un réseau de neurones du cerveau d'un rat. Une grande quantité de données expérimentales et de nombreuses caractéristiques (forme, taille, comportement électrique…) ont été prises en compte pour représenter chaque neurone et ses connexions. Les méthodes de simulation du *Blue Brain Project* se poursuivent au sein du ***Human Brain Project***. Celui-ci a pour objectif de rassembler les connaissances sur le cerveau humain, d'améliorer la compréhension des maladies neurologiques et de développer la conception de systèmes informatiques en imitant certaines fonctions du cerveau. Plus d'une centaine d'institutions scientifiques dans le monde collaborent au projet ;

- L'arrivée des satellites a apporté des données d'observation sur les différents constituants de l'atmosphère. Dans la stratosphère, une couche d'ozone (O_3) nous protège du rayonnement ultraviolet nocif du Soleil. On observe un « trou » dans la couche d'ozone au-dessus de l'Antarctique. Son étendue dépend de la saison (températures hivernales et lumière insuffisante) mais aussi de la présence de gaz à composés chlorés et bromés qui détruisent la molécule d'ozone. On a pu mesurer et suivre son évolution : peu étendu au début des années 1980, il s'est agrandi dans les années 1990 puis stabilisé lors des années 2000. C'est seulement depuis quelques années que l'on observe une reconstruction progressive de la couche d'ozone et ce, grâce à la réduction des émissions humaines de gaz destructeurs ;



Vue en fausses couleurs de la quantité totale d'ozone au-dessus de l'Antarctique le 10 mars 2017. Entre les zones vertes, plus fournies, et les zones bleues, plus pauvres, il existe un facteur deux. Crédit : NASA / Goddard Space Flight Center.

- L'observation des rayonnements qui nous arrivent de l'espace permet aux astrophysiciens de construire des modèles physiques pour expliquer leur origine et leur nature. La simulation présentée ici est celle d'un trou noir supermassif identique à celui qui serait situé au centre de notre galaxie, autour duquel des étoiles tournent tout en expulsant de la matière sous forme de vents stellaires. Cette accumulation de matière forme un disque d'accrétion. Par comparaison avec les données recueillies par le télescope spatial Chandra qui observe dans le rayonnement X, les scientifiques ont compris que ces étoiles, bien que situées sur des orbites proches du trou noir, ne pouvaient, à elles-seules, être impliquées dans l'accrétion.

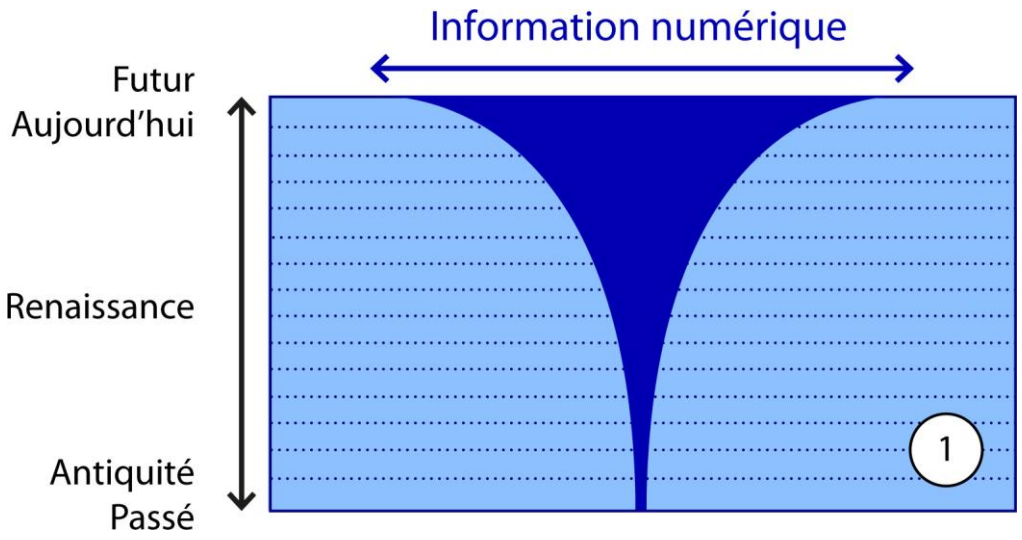
Table 16

Venice Time Machine

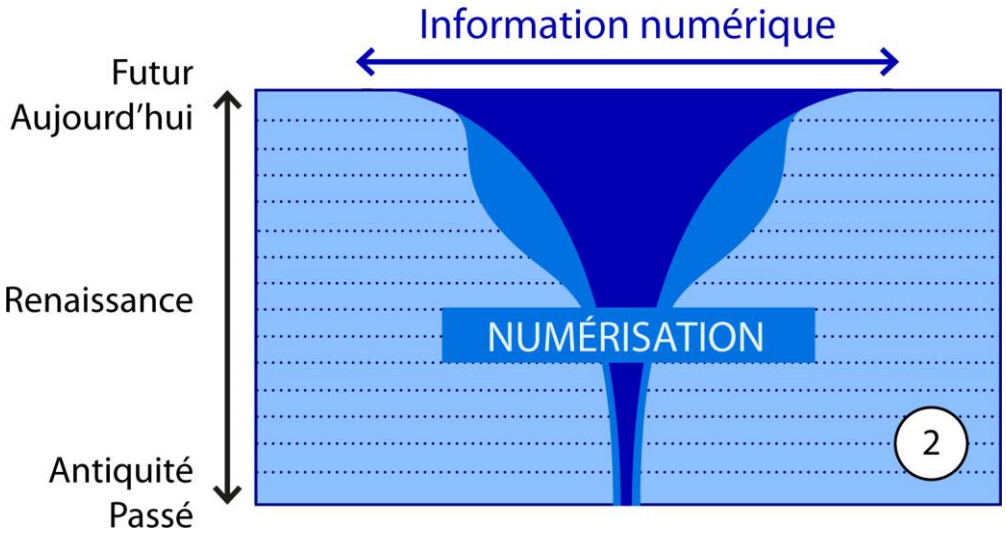
Nous avons déjà exposé le projet *Venice Time Machine* en introduction à ce document. Voyons comment les nouvelles technologies numériques, qui nous projettent déjà dans le futur, nous aident aussi à mieux comprendre le passé. Ainsi, rien que dans le domaine du traitement des données :

- la numérisation transforme un document physique (un texte manuscrit, par exemple) en une série d'images numériques ;
- l'analyse algorithmique des pages numérisées permet de repérer toutes les occurrences d'un même mot sur un document, puis sur l'ensemble des documents numérisés. Ceci permet d'établir des connexions entre toutes ces formes graphiques ;
- la modélisation permet de reconstruire les réseaux unissant les entités présentes dans les documents (personnes, lieux, etc.) ;
- la simulation, qui extrapole les données manquantes à partir des modélisations, permet de reconstruire plusieurs versions possibles du passé.

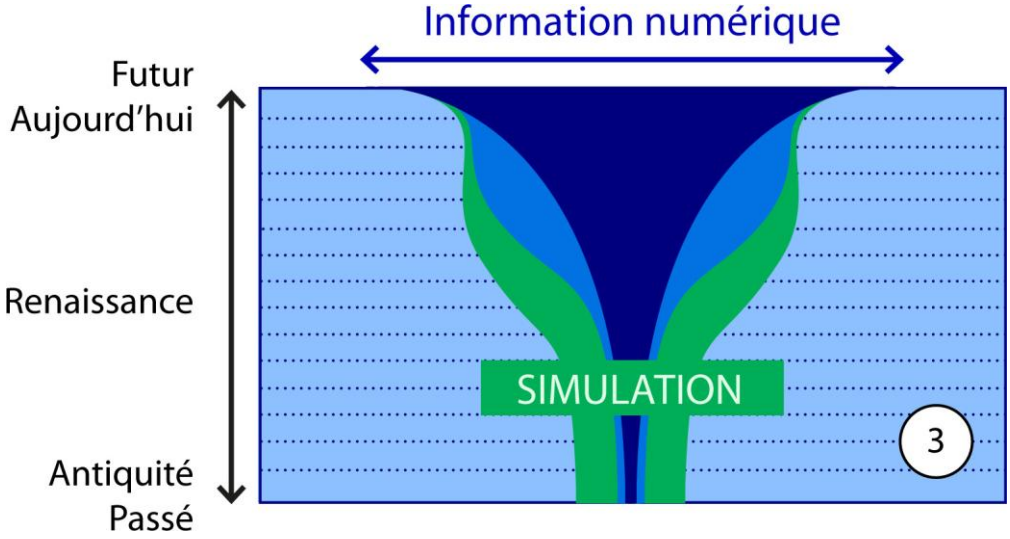
Afin d'utiliser les technologies du présent pour explorer le passé, nous avons besoin d'une densité d'informations comparable à celle du présent.



① Plus on va vers le passé, moins il y a de documents et d'informations disponibles pour raconter l'histoire de Venise.



② La première étape consiste à numériser les documents disponibles dans les archives. Entre aujourd'hui et la Renaissance, ils sont très nombreux.



③ Dans le passé plus lointain, il y a moins de documents d'archives et donc de données utilisables. Il faut dès lors extrapoler et simuler les données manquantes à partir des données existantes. Autrement dit, il nous faut élargir le pied du « champignon » informationnel.

Un audiovisuel propose une modélisation historique et géographique du quartier de Rialto entre 950 et 1850. La simulation a été réalisée par le Laboratoire d'humanités digitales de l'École polytechnique fédérale de Lausanne (EPFL).

Table 17

Des secteurs d'activité en mutation

Modèles, algorithmes et données sont les piliers d'une nouvelle donne. Avec ces nouvelles technologies, ce sont des secteurs entiers d'activité qui se transforment comme ceux des assurances, du renseignement ou encore celui de la santé publique, où l'on passe du concept de réparation à celui de la prévention.

Tourisme

Ce domaine a été complétement reconfiguré par les outils numériques : gestion, réservations, avis d'utilisateurs, etc. Les professionnels du secteur, dont la pratique reposait sur le contact direct avec les clients, doivent innover pour affronter les géants du net.

Médias

L'information est désormais accessible n'importe où, n'importe quand et à coût quasi nul. Les services numériques sont en mesure d'offrir des informations ciblées selon les goûts et les intérêts des utilisateurs. Le risque : une information peu diversifiée et manquant de neutralité. Point positif : l'analyse de données s'ajoute aux moyens d'investigation des datajournalistes.

Renseignement

En matière de renseignement, la collecte de données devient colossale : communications, localisation des téléphones, vidéo-surveillance, etc. Ces données doivent être analysées pour prendre sens. Par exemple, établir des liens entre des individus peut faire émerger des réseaux (terrorisme, pédophilie, délinquance financière…). Toutefois, la surveillance numérique massive des citoyens doit être encadrée pour préserver les libertés essentielles.

Santé

L'explosion des connaissances en sciences biomédicales et la mise à disposition de technologies comme le séquençage des gènes transforment les pratiques. L'analyse de données est un nouvel outil pour les chercheurs, par exemple pour évaluer les effets de combinaisons de médicaments, les risques d'exposition à des environnements nocifs ou les bénéfices de nouveaux traitements. Elle ouvre aussi la voie vers une médecine plus personnalisée et plus orientée sur la prévention.

Agriculture

Les pratiques agricoles vivent une mutation technologique : sur les tracteurs et les parcelles, des capteurs collectent des données qui permettent d'adapter les apports en eau, pesticides et engrais ; le bétail est équipé de puces électroniques. L'analyse de données est aussi utilisée par la recherche agronomique pour obtenir par sélection des lignées animales plus performantes et de nouvelles variétés de plantes cultivées.

Énergie

Le numérique apparaît comme une nécessité pour relever le défi écologique. La multiplication des capteurs fournit une masse de données qui permet de moduler la fourniture d'énergie. Il devient possible d'envisager une transformation en profondeur du réseau énergétique : production dans de petites unités à partir d'énergies renouvelables, consommation locale évitant les gaspillages du transport sur de grandes distances.

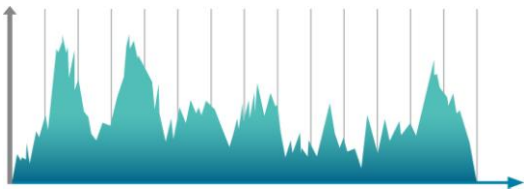
Un multimédia réalisé avec la société d'assurance mutuelle MAIF vous montrera combien la révolution des données impose des mutations profondes aux assurances. En effet, la donnée a toujours été la matière première de l'assurance, qui repose sur la prévision et l'estimation des risques. Aujourd'hui, elle est prise en compte à tous les niveaux. En voici quelques exemples : l'analyse décèle qu'un client se prépare à résilier son contrat, le marketing peut réagir ; les tarifs peuvent s'adapter au profil de risque des conducteurs grâce à l'étude de leur comportement ; le traitement des données révèle les éventuelles tentatives de fraudes dans des dossiers de sinistres. Du côté assuré, le *Big Data* promet une tarification plus juste et le règlement plus rapide des sinistres.

Toutefois, l'utilisation massive des données personnelles pose un problème éthique car les individus sont rarement tenus informés de l'usage qui est fait de leurs données et ne disposent pas des moyens de les exploiter pour eux-mêmes. De par ses valeurs mutualistes, la MAIF entend corriger cette asymétrie et s'intéresse particulièrement au courant *Self data*, qui vise à donner aux individus les moyens de se réapproprier leurs données personnelles, en organisant eux-mêmes leur production, leur exploitation et leur diffusion.

Table 18

Les métiers des données

Face à l'explosion quantitative des données numériques, les métiers à haute valeur ajoutée liés à leur traitement et leur analyse deviennent stratégiques. Des modes de stockage, d'analyse, de visualisation doivent être réinventés et de façon concomitante, une kyrielle de nouveaux professionnels du numérique émergent dans tous les secteurs. Ces profils sont convoités dans les startups, les grands groupes ou chez les professionnels du web. Selon la Fédération Syntec (un regroupement de syndicats professionnels), le numérique génère en France en moyenne 35 000 embauches annuelles, dont la moitié sont des créations de postes. Pour répondre au besoin constaté des entreprises, les écoles d'ingénieurs ont récemment créé de nombreuses formations spécialisées de niveau bac + 5 et plus.



Le développement de ces métiers est une tendance de fond. Déjà présents dans le monde de la finance, les métiers autour de l'analyse de données s'implantent dans la grande distribution, les télécoms, le secteur public, etc.

Les nouveaux métiers de la donnée exigent des compétences multiples :

- informatique mais également statistique ou mathématiques appliquées ;
- formation métier dans le domaine visé : finance, commerce, ressources humaines, etc. ;
- expertise dans l'analyse de données massives pour se poser les bonnes questions, suivre les bonnes pistes, bâtir les tests et expérimentations adaptés ;
- capacité de collaborer avec des équipiers d'univers différents car une même personne ne peut maîtriser tous les degrés d'expertise requis.

Les lignes qui suivent fournissent un inventaire et une brève description de certains métiers de la donnée. Elles pourraient servir de base à une réflexion sur le parcours professionnel futur de vos élèves. Les anglicismes sont omniprésents !

Data scientist

Selon la taille de l'entreprise et le secteur d'activité, le *data scientist* peut encadrer une équipe ou concentrer plusieurs fonctions dont celle de *data miner*. Il doit avoir des compétences en algorithmique et développement de modèles mais également en infrastructure et datavisualisation. Son rôle : fournir des informations fiables à son employeur pour la prise de décision.

Data miner

Parmi des milliers d'informations, le *data miner* extrait celle qui sera décisive pour l'avenir d'une entreprise. Informaticien et mathématicien, il utilise des techniques de fouille de données et met parfois à profit l'apprentissage automatique. Dans le secteur de la grande distribution, par exemple, il établit des profils de clients à partir de leurs achats afin de prédire leurs comportements futurs de consommateurs.

Web analyst

Spécialiste des sites internet, il joue un rôle crucial pour l'économie du web. Présent surtout dans les secteurs du commerce en ligne et des sites internet culturels, il recueille, vérifie et analyse les données d'audience, les profils clients, etc., pour fournir des réponses fiables aux questions de son employeur. Il identifie alors les améliorations à apporter en fonction de la stratégie marketing de son entreprise.

Data analyst

Les organisations accumulent des données, mais celles-ci sont rarement structurées. De plus, les formats sont hétéroclites : fichiers créés par des traitements de texte et des tableurs, fichiers PDF, fichiers audio ou vidéo, images, courriers électroniques, etc. Le *data analyst* rend ces données utilisables. Il les agrège puis les modélise dans un format simple. Il doit savoir concevoir des programmes informatiques pour, par exemple, générer des données et les organiser sous forme de tableaux interactifs.

Data broker

Le *data broker* (courtier en données) explore des sources publiques et non publiques, pour en extraire des données personnelles : âge, revenu, numéro de sécurité sociale… voire historique médical. Segmentées par profils détaillés, ces données sont vendues aux banques et aux professionnels du marketing qui peuvent ainsi cibler leurs propositions. Très répandue aux USA, cette pratique est strictement réglementée en Europe : la collecte de masse est incompatible avec le respect du droit fondamental à la protection des données personnelles.

Architecte Big Data

L'architecte Big Data conçoit et met en place l'infrastructure informatique d'analyse des données. Il est compétent en systèmes informatiques et souvent en développement. Il collabore avec les *data scientists* et *miners* afin de bien dimensionner l'infrastructure (puissance de calcul, mémoire, espace disque, bande passante réseau, etc.) et recherche un compromis entre la performance et le coût.

Business analyst

Il fait le lien entre les techniciens informatique et l'entreprise qui fait appel à eux. Il ajoute donc à ses compétences techniques une parfaite connaissance du domaine d'activité dans lequel l'entreprise évolue. Il intervient dès le début d'un projet, analyse les besoins du client, estime les coûts, étudie la façon dont le projet s'inscrit dans la stratégie de l'entreprise, propose des améliorations et rédige le cahier des charges. Il fait ainsi en sorte que les techniciens comprennent parfaitement les attentes du client et il pilote la réalisation.

Data protection officer

Garant d'un traitement responsable des données, le *Data Protection Officer* (DPO) est chargé de la sécurité informatique et juridique. Chef d'orchestre de la conformité, il protège les données et applications informatiques de l'entreprise. Il veille aussi à la protection des données personnelles. À ce titre, sa présence sera obligatoire dans les administrations publiques dès mai 2018, en vertu d'un règlement européen. De même pour les entreprises effectuant un « *suivi régulier et systématique des personnes* » (profilage par exemple) et tout organisme « *traitant des données sensibles à grande échelle* » (Sécurité sociale).

Table 19

Le datajournalisme

Le data journalisme est un bon exemple de la transformation du cœur d'activité d'un métier. À travers l'exemple du scandale des *Panama papers*, nous décryptons les méthodes d'investigation journalistiques s'appuyant sur les Big Data et permettant pour la première fois de faire éclater certains scandales indécelables auparavant.

Face à l'avalanche de données désormais accessibles, que devient l'investigation journalistique ? Comment produire de l'information de qualité à partir de données numériques qu'il faut comprendre, puis collecter, recouper, analyser, traiter et éventuellement visualiser ? Un nouveau journalisme est en train d'émerger.

En quoi datajournalisme et journalisme classique sont-ils différents ?
L'enquête est toujours au cœur du métier, avec sa part de « flair », mais elle s'est enrichie de nouvelles sources et méthodes. Les journalistes acquièrent de nouvelles compétences ou travaillent avec des professionnels du traitement numérique des données, pour faire le tri dans la grande masse d'informations disponibles.

Les bases de données ont toujours existé mais depuis l'automatisation de leur traitement et l'existence de tableurs, on peut les exploiter plus rapidement et en profondeur. À partir de la collecte, le datajournaliste édifie des bases de données exploratoires avant d'approfondir son enquête. Ces bases de données peuvent être collaboratives. Le datajournaliste peut utiliser des données fournies par une source anonyme ou des données *hackées*, dérobées et utilisées sans autorisation du propriétaire. Il exploite aussi des données ouvertes (*open data*), dont l'accès et l'usage sont libres.

Une nouvelle écriture journalistique : entre vigilance et exigence
Les données numériques permettent d'étayer un article autour d'éléments les plus objectifs possibles comme des documents PDF, des chiffres et des photos, plutôt que des citations. Leur traitement fait apparaître des histoires indécelables autrement, en allant au-delà de la communication toute faite des entreprises et des institutions, par exemple.

Mais les données ne sont pas un reflet parfait du monde. Il faut d'abord s'interroger sur les sources dont elles proviennent et le mode de collecte, puis les travailler : sélection, filtrage, corrections, analyses, contrôles, vérification. Il faut ensuite les interpréter, les mettre en perspective et les contextualiser.

De la même façon que nous avions décrit les métiers de la donnée en table 18, les lignes suivantes donnent un aperçu des nouveaux métiers au cœur des rédactions.

<u>Datajournaliste</u>
Le datajournaliste mène ses investigations en exploitant données ouvertes, fournies ou piratées. Il analyse chiffres, statistiques, courriers et autres documents pour en extraire des informations à haute valeur ajoutée qu'il faut ensuite contextualiser. Enfin, il utilise la datavisualisation pour optimiser la présentation de ses articles et en faciliter la compréhension.

<u>Responsable d'édition (*front page editor*)</u>
Il suit l'actualité immédiate et le contenu des sites concurrents pour faire évoluer plusieurs fois par jour la page d'accueil de son média. Sa connaissance des algorithmes des moteurs de recherche et des réseaux sociaux lui permet d'y apparaître en bonne place. Il attire ainsi des lecteurs sur le site pour lequel il travaille.

Développeur

Les développeurs sont de plus en plus sollicités par les rédactions. Il peut s'agir de gros projets (création ou refonte d'un site internet) ou de projets ponctuels (création d'une application pour les Jeux olympiques, collaboration avec un datajournaliste sur une enquête…). Certains journalistes apprennent à coder pour gagner en autonomie et permettre aux développeurs de se concentrer sur les projets qui nécessitent un savoir-faire plus pointu.

Un audiovisuel d'une durée de quatre minutes, dédié aux *Panama papers*, clôt l'élément d'exposition. De quoi s'agit-il ? En 2016, une gigantesque enquête menée par 370 journalistes dans près de 80 pays a mis au jour un système mondialisé d'évasion fiscale. Au cœur de l'affaire : un cabinet d'avocats panaméen et 11,5 millions de documents riches en données. Le voile est levé : sur la planète entière, la fraude fiscale est devenue une institution. D'Argentine en Chine, d'Algérie en Islande, de savants montages permettent à des grandes fortunes d'échapper à l'impôt ou blanchir de l'argent sale. Des stratagèmes mis en place par des banques, des avocats, des professionnels de la finance pour des clients de plus de 200 nationalités différentes, par l'intermédiaire du cabinet Mossack Fonseca fondé dans les années 1970. L'audiovisuel dévoile comment ont travaillé les journalistes pour mener cette enquête exemplaire.

Table 20

Intermédiation algorithmique

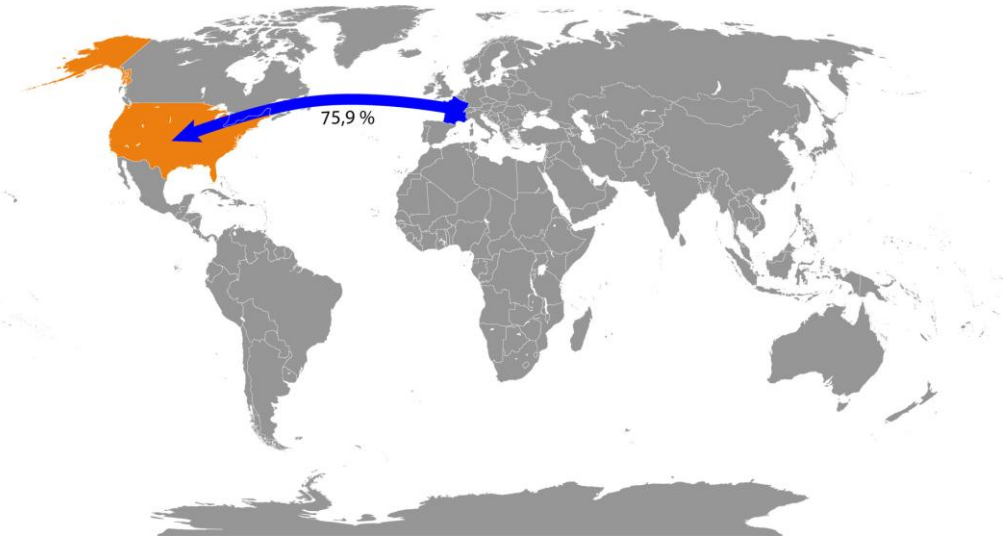
Les plateformes d'intermédiation mettent en relation directe les utilisateurs, supprimant les autres intermédiaires classiques entre les clients et les fournisseurs de services. Ce sont des plateformes marchés comme *Uber* ou *Airbnb* ou des plateformes sociales comme *Twitter*. Elles collectent d'abord des données, les traitent, les stockent puis recueillent les traces d'utilisation des services pour en offrir de nouveaux. Elles font payer de façon tout à fait différente les utilisateurs et les fournisseurs. Parfois, les utilisateurs ne paient pas mais c'est la revente de leurs données qui paie le service. L'objectif de cet élément audiovisuel est de présenter le rôle et l'influence des plateformes d'intermédiation ainsi que la façon dont elles changent notre façon de voir le monde.

Flux de données

Services, réseaux sociaux ou moteurs de recherche : l'efficacité des plateformes numériques tient en grande partie au fait qu'elles multiplient les modes de valorisation des flux de données collectées auprès des internautes.

Ces flux sont identifiés grâce à des indicateurs de trafic : nombre de visites sur une page, pages vues, nombre de visiteurs, etc. Sur le plan mondial, on peut représenter ces flux sur une carte et constater que ces données dépendent de plusieurs territoires : le pays où elles sont produites, le pays où est localisée la plateforme numérique qui les stocke et celui où est le siège social.

Il est, par exemple, intéressant et instructif d'apprendre qu'en Russie, plus du tiers des visites sur les pages web les plus populaires se font sur des sites ou des plates-formes dont le siège social se trouve aux États-Unis. En Iran, la proportion monte à plus de 50 %, en France à plus de 75 %, au Royaume-Unis à 90 % et aux États-Unis, elle vaut… 100 % (source : https://intermed.jumplyn.com/)



II.2.5 Les données, où ça nous mène ?

L'interaction entre la technologie et la société est telle aujourd'hui que les développements techniques ont des conséquences environnementales, sociales et humaines qui dépassent de loin les objectifs des appareils techniques et des pratiques elles-mêmes. Ainsi la numérisation croissante de l'univers physique et des existences humaines n'est pas sans affecter la gouvernance du monde, frapper l'imaginaire, interroger les pratiques et les consciences, inquiéter les libertés et faire émerger par rebond une culture et une éducation au phénomène numérique qui, à leur tour, vont agir sur le monde.

La collecte des données personnelles serait devenue le « mal nécessaire » de l'accès aux services numériques. Peut-on et doit-on se résoudre à cette perspective ? N'est-il pas essentiel de donner aux citoyens les moyens légaux de rééquilibrer l'asymétrie d'information et de pouvoir entre ceux qui émettent les données, volontairement ou à leur insu, et ceux qui les utilisent avec plus ou moins de transparence et de loyauté ?

Table 21

Interview audiovisuelle d'Isabelle Falque-Pierrotin, présidente de la Commission nationale de l'informatique et des libertés (CNIL).

Table 22

Les données personnelles

Avant l'arrivée d'internet, vos données personnelles étaient conservées sur des fichiers papier ou informatique par la Sécurité sociale, les impôts, l'Insee, votre mairie, votre employeur, votre médecin, etc.

Aujourd'hui, chaque fois que vous communiquez par mail ou sur les réseaux sociaux, réglez par carte bancaire ou vous déplacez avec votre smartphone, vos données personnelles sont traitées et stockées sur des serveurs distants un peu partout dans le monde, regroupés communément sous l'appellation de *cloud*.

Qu'est-ce qu'une « donnée personnelle » ?

C'est une information qui permet d'identifier une personne physique. Vous pouvez être identifié par votre état-civil (vos nom et prénoms) ou, indirectement, par votre numéro de Sécurité sociale, votre numéro de téléphone, l'immatriculation de votre véhicule ou encore par le recoupement de données (comme : date et lieu de naissance / lieu de résidence…).

Les données biométriques (empreintes digitales, iris, voix, visage…) sont rattachées à vos caractéristiques physiques et sont donc considérées comme des données personnelles. Des données techniques comme l'adresse mac de votre smartphone, l'adresse IP de votre ordinateur, les traces de vos navigations sur internet, votre géolocalisation, forment votre identité numérique. Elles permettent aussi de vous identifier.

La donnée personnelle, nouvelle monnaie du XXI^{e} siècle ?

Vos données ont une valeur économique : lorsque vous les communiquez à un opérateur internet, celui-ci en tire un bénéfice financier. Un réseau social, par exemple, vous offre un service gratuit mais pour y accéder, vous « monnayez » à votre insu vos informations personnelles. Né avec l'internet, ce nouveau modèle d'affaire porte un nom : l'économie de la gratification immédiate.

Les données personnelles, un bien commun ?

Vos données sont utiles aux chercheurs et aux statisticiens : meilleure connaissance des populations, suivi des épidémies, évaluation des politiques publiques… Elles sont anonymisées avant d'être rendues publiques, de plus en plus souvent, via *l'open data* ou *donnée ouverte*.

Traces numériques, sécurité nationale, cyber sécurité

Les États, au nom des impératifs de sécurité publique et de défense nationale, s'intéressent aussi aux traces numériques laissées dans le cyberespace, qu'il s'agisse de se protéger contre les attaques informatiques, de faciliter la recherche criminelle ou de lutter contre le terrorisme. Dans nos sociétés démocratiques, la loi définit les conditions dans lesquelles les services de police et la justice peuvent avoir accès à ces données.

Le saviez-vous ?

En analysant la base de données d'un opérateur téléphonique, des chercheurs ont pu identifier facilement 95 % des clients répertoriés sur ce fichier pourtant « anonymisé » ! Pour cela, il leur suffisait d'analyser les données techniques (date et lieu d'émission) de quatre appels téléphoniques seulement. Source : Yves-Alexandre de Montjoye, César Hidalgo, Michel Verleysen et Vincent Blondel, *Unique in the Crowd: The privacy bounds of human mobility*, journal en ligne *Scientific Reports* 3, article n°1376, 2013.

Le rôle de la Commission nationale de l'informatique et des libertés (CNIL)

Dans l'univers numérique, la CNIL est le régulateur des données personnelles. Elle informe les personnes sur leurs droits, les aide à les exercer et à maîtriser leurs données personnelles et instruit leurs plaintes (8 000 plaintes en 2015). Dans ce cadre, elle mène des contrôles (500 contrôles en 2015) et sanctionne les manquements à la loi informatique et libertés et en particulier, peut infliger des amendes pouvant aller jusqu'à 3 millions d'euros.

Parallèlement, elle conseille les professionnels pour leur permettre d'être en conformité avec la loi et contrôle les projets de traitement de données comportant des risques pour les personnes. Elle mène une réflexion sur les problèmes éthiques et les questions de société soulevés par le numérique. Enfin, elle travaille en étroite collaboration avec ses homologues européens et internationaux pour élaborer une régulation harmonisée.

Nous transmettons très souvent des données personnelles à des sites, administrations ou entreprises lorsque nous participons à un concours, demandons une carte de fidélité, remplissons un formulaire sur internet, ou simplement postons sur les réseaux sociaux. Grâce à la loi informatique et libertés, nous avons des droits sur les données qui nous concernent, notamment le droit de connaître l'usage qui en est fait, le droit de les faire rectifier, voire totalement supprimer. Tout ceci vous est expliqué dans un audiovisuel de trois minutes.

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

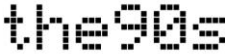
Table 23

Numérisation du monde et libertés individuelles

En 1978, en réponse à la mise en place de fichiers informatiques interconnectés (SAFARI), la commission nationale de l'informatique et des libertés (CNIL) est créée. Quarante ans plus tard, les libertés individuelles restent la préoccupation centrale des instances de régulations du numérique et plusieurs chantiers sont en cours. Par exemple, le projet *Mesinfos* (http://mesinfos.fing.org/) piloté par la Fondation internet nouvelle génération (Fing), vise à la production, l'exploitation et le partage de données personnelles par les individus, sous leur contrôle et à leurs propres fins. C'est le *self data*. Autre démarche : le centre d'accès sécurisé aux données (CASD) permet aux chercheurs de travailler sur des données individuelles soumises à la confidentialité dans des conditions de sécurité élevées. L'objectif de cet élément est de présenter le contexte juridique et éthique à travers quelques démarches institutionnelles et associatives.

Cet élément d'exposition propose un survol historique en quatre périodes :

- **les années 1970, la prise de conscience**. L'informatisation des administrations puis des entreprises soulève des débats sur la protection de la vie privée. Entre 1970 et 1980, la France et d'autres pays d'Europe se dotent de lois et d'autorités de contrôle pour protéger les personnes face au développement de l'informatique et encadrer la détention de fichiers sur les individus et l'exploitation abusive qui pourrait en être faite ;
- **les années 1980, les fichiers de l'État sous contrôle**. Les grands fichiers administratifs sont déclarés à la CNIL, que ce soit dans le domaine régalien comme dans ceux des impôts, de l'action sociale ou de la santé. La médecine préventive et l'action sociale font appel à l'informatique de gestion, des registres épidémiologiques sont créés. Avec l'essor du Minitel et des ordinateurs personnels, le monde du travail est peu à peu pénétré par l'informatique : logiciels de gestion des personnels, recrutement, contrôles d'activité et élections professionnelles. Les fichiers de prospection commerciale se développent ;
- **les années 1990, les données, carburant du marketing**. Internet est lancé. Les réseaux s'étendent, les données circulent. L'e-santé fait ses premiers pas : sites de santé, réseaux de soins, télémédecine, carte de santé… La vidéosurveillance entre en scène. Sur les lieux de travail, la cybersurveillance se met en place, avec le suivi et le contrôle de l'activité des salariés. Les mégabases comportementales et les méthodes de profilage prennent leur essor : la prospection commerciale cible les goûts et les habitudes de consommation, banques et assurances utilisent le calcul de probabilités pour identifier les risques socio-économiques des clients. Un nouveau cadre juridique apparaît pour favoriser la libre circulation des données personnelles en Europe tout en harmonisant le niveau de protection de ces données ;



- **depuis le début des années 2000, la protection des données à l'heure des géants du web**. Le web révolutionne les pratiques sociales. Empreintes digitales, ADN, iris, contour de la main et reconnaissance faciale : il faut livrer ses données biométriques pour passer une frontière, accéder à son bureau, utiliser son smartphone… Cartes bancaires, pass transport, smartphones, voitures connectées : les objets connectés se multiplient et tracent déplacements et comportements. Les données personnelles sont au cœur de l'économie numérique et voyagent à travers le monde. Les exigences sécuritaires conduisent à la création de nouveaux fichiers de police. La police et la justice accèdent plus largement aux communications échangées entre particuliers ou entre États.

Table 24

Les cookies

Les clics, « likes », liens et traces de passage sont autant de données qui en disent sur nous. Ces empreintes numériques s'appellent les *cookies*. Ce sont des petits fichiers texte déposés dans votre ordinateur à la demande du site que vous visitez. Ce site et d'autres serveurs partenaires comme des régies publicitaires, des éditeurs de service ou des réseaux sociaux, ont accès aux informations contenues dans les cookies et peuvent en déduire vos habitudes de consultation ou de consommation. Lors d'une prochaine connexion, ces serveurs reconnaîtront votre profil et vous proposeront des services de plus en plus personnalisés. À l'origine, les cookies ont été inventés pour mémoriser vos préférences de navigation, notamment la langue que vous utilisez et votre panier d'achat. Aujourd'hui, les acteurs du web s'en servent pour scruter la fréquentation des sites, fluidifier leur trafic, proposer de la publicité ciblée ou encore interagir avec les réseaux sociaux. L'application CookieViz (un outil de localisation des cookies développé par la CNIL) permet de mesurer les interactions entre son ordinateur, son navigateur et des sites ou des serveurs distants. L'outil permet de visualiser en temps réel les traces que nous laissons derrière nous en surfant sur le web. Puis, CookieViz affiche et identifie point par point les agrégateurs de données qui captent et revendent nos données personnelles à notre insu. On peut ainsi mesurer combien la captation de nos données à notre insu et leur utilisation en secondes mains est répandue.

Le Wi-Fi

Les smartphones, tablettes et montres connectées sont de formidables outils de surveillance à distance. Les signaux émis par ces appareils utilisant le Wi-Fi (une technologie de communication sans fil) permettent de tracer les personnes qui les portent, à des fins diverses : observation des axes de transport routier, statistiques et profilage pour les enseignes commerciales, ciblage des publicités urbaines, etc. Cette collecte de données de présence et de mobilité pose évidemment des problèmes de vie privée. Elle se fait généralement à l'insu de l'utilisateur ; ce qui n'est pas le cas dans l'exposition. Cet élément multimédia vous permet de vérifier si vous avez été suivi ou non.

Table 25

My Google search history

Google capte nos données personnelles et il est presque impossible pour nous, utilisateurs, de contrôler ce que celles-ci deviennent. En 2006, l'artiste Albertine Meunier s'empare du service *Search History* de Google, qui enregistre tous nos historiques de navigation. Elle réussit à décrypter les fichiers issus de son propre historique de navigation. Albertine Meunier décide alors de les compiler de manière exhaustive et de les publier dans un livre. Ces 160 pages de requêtes mises bout à bout racontent l'histoire d'Albertine mais aussi celle du réseau. Le tome 2 est paru fin 2016. Les deux tomes vous sont exposés sous vitrine. Dans une audiovisuel de trois minutes, l'auteure nous présente *My Google Search History*, une œuvre à lire et à écouter.

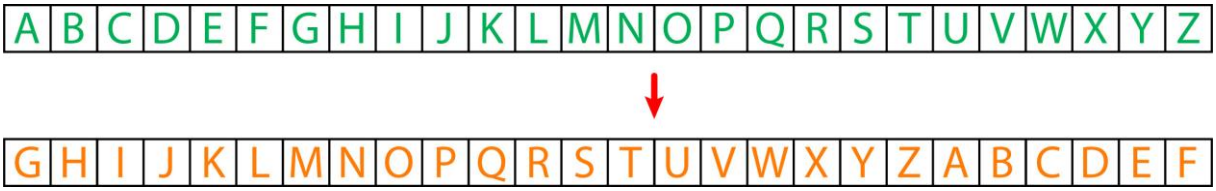
Crédit : Benjamin Boccas.

Du chiffrement au déchiffrement

Pour cacher ses données, on peut chercher à les rendre illisibles pour ceux à qui elles ne sont pas destinées, notamment à l'aide du chiffrement. Ce procédé consiste à brouiller l'information à l'aide d'un code, d'une « clé de chiffrement », pour rendre impossible sa compréhension à celui qui ne possède pas cette clé. Le déchiffrement, processus inverse, sert à transformer les informations de façon à les rendre à nouveau compréhensibles. La sécurité d'un système de chiffrement repose donc sur le secret de la clé de chiffrement et non sur le transfert. Lorsqu'une personne cherche à obtenir l'information sans en avoir la clé, elle fait appel à des techniques de décryptage.

Le chiffrement des données est devenu un enjeu majeur. Auparavant relativement simples, les méthodes de chiffrement font appel aujourd'hui à des algorithmes complexes de calcul.

Le plus connu des algorithmes est le « chiffre de César », qui opère par décalage ou substitution. Jules César l'utilisait pour ses correspondances secrètes. On obtient le message chiffré en remplaçant chaque lettre du message original par une lettre située plus loin dans l'alphabet. Le décalage est fixe, c'est la clé de chiffrement. Par exemple, avec un décalage de 6 lettres (clé 6), le A devient G, le B devient H et ainsi de suite.



Longtemps, toutes les méthodes de chiffrement furent des variantes de plus en plus complexes du chiffrement par substitution. Ainsi, durant la Seconde Guerre mondiale, les Allemands utilisèrent la machine à chiffrer électromécanique Enigma.

L'histoire du chiffrement s'apparente à une bataille acharnée : les chiffreurs ne cessent d'élaborer de nouveaux algorithmes, les décrypteurs s'emploient à les « casser ».

1/ Ce message a été chiffré avec le chiffre de César, clé de chiffrement 6.
Hoktbktak g rg Iozk jky Yioktiky kz jk r'Otjayzxok
Pouvez-vous le déchiffrer ?

2/ Ce message a été chiffré avec le chiffre de César mais vous ignorez la clé.
Kxkt a'tmedhxixdc Itggp Spip !
À vous de décrypter !

Réponses en page suivante…

Réponses
1/ Bienvenue à la Cité des sciences et de l'industrie
2/ Vive l'exposition Terra Data ! (Décalage de 15)

Table 27

Disparition
Peut-on se soustraire au regard électronique des entreprises et institutions ? C'est le thème développé dans le multimédia proposé ici. Comment disparaître ? Faut-il déconnecter ordinateur et téléphone ? Renoncer au net et aux messageries ? Il existe aussi des systèmes d'anonymisation, comme le routeur Tor qui permet d'évoluer masqué sur le web. Toutefois, nul ne peut faire disparaître complétement ses données s'il possède carte bancaire, carte vitale, passeport biométrique, pass de transport, etc. Une grande partie de la vie sociale est aujourd'hui numérisée.

Concrètement, il existe des solutions pour interagir le moins possible sur les réseaux et éviter le traçage par les cookies : par exemple, installer des systèmes d'anonymisation ou couper les ondes de son portable. Si l'on ne souhaite pas se contraindre à des mesures si radicales, l'autre alternative possible est l'offuscation. Cette stratégie consiste à générer des informations superflues, inutiles, ambiguës ou inexactes, afin de rendre le ciblage peu précis et inefficace. Dans cet élément, vous apprendrez à déjouer les dispositifs de traçage en brouillant intelligemment vos données.

La disparition ou l'offuscation ne sont pas des solutions satisfaisantes à terme. Une protection véritable des données personnelles exige la participation des acteurs du numérique. L'entreprise à laquelle le citoyen confie ses données devrait garantir qu'elle ne les utilisera pas à d'autres fins que celles qui ont obtenu le consentement éclairé du citoyen. Les citoyens n'auront confiance dans ces technologies que si la pratique des firmes et institutions est transparente. En attendant, chacun doit s'efforcer de garder le contrôle de ses traces numériques en comprenant le fonctionnement des outils numériques qu'il utilise et ce qu'ils laissent filtrer de sa vie privée.

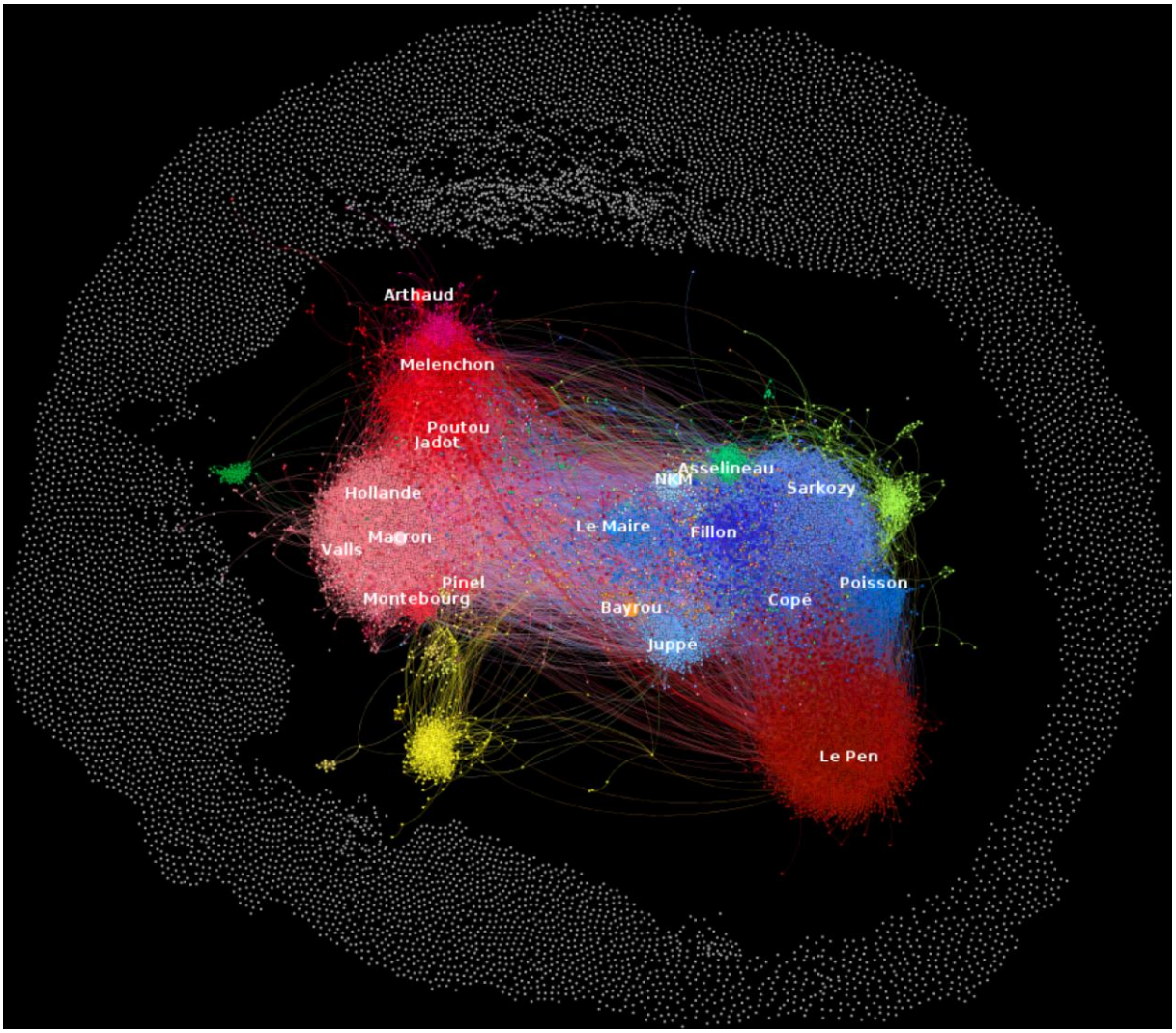
Table 28

Des réseaux sociaux dans nos urnes
La grande quantité de données désormais disponibles sur le web ouvre de nouveaux champs d'étude, avec un traitement des informations à la fois graphique et analytique. Par exemple, les données issues des interactions entre abonnés à Twitter permettent de cartographier les relations entre acteurs de la sphère politique.

L’Institut des Systèmes Complexes Paris Île-de-France a conduit une recherche à partir de l'observation des échanges de tweets liés à la campagne présidentielle de 2017, en vue de représenter et étudier ces interactions (y compris les occurrences où un abonné re-tweete un message à l'identique).

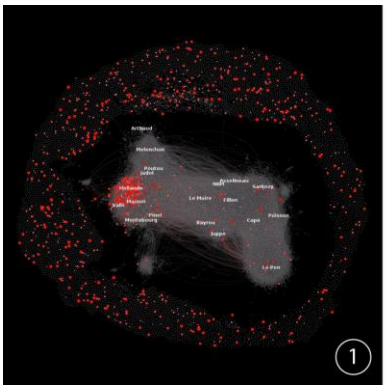
Les graphes générés par des algorithmes montrent la dynamique de diffusion de l’information liée à la politique dans ce réseau social. Chaque nœud correspond à un compte Twitter ayant produit du contenu politique. Chaque ligne liant deux nœuds indique une interaction forte entre ces comptes : ils se sont re-tweetés au moins 5 fois entre août et octobre 2016.

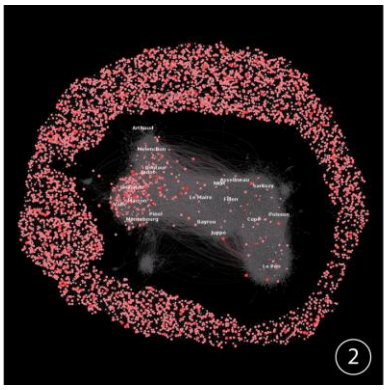
Le graphe total comporte environ 300 000 comptes utilisateurs dont 21 000 sont représentés ici.

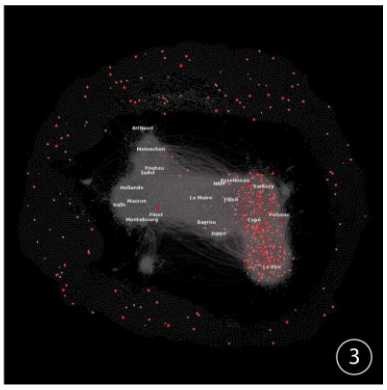


21 000 comptes Twitter sont représentés sur ce graphe :

- 11 500 comptes qui interagissent fortement avec au moins 5 autres comptes ainsi que leurs liens (femmes et hommes politiques, journalistes…). C'est le cœur de la « tweetosphère » politique ;
- 10 000 comptes qui forment une « couronne » éloignée du cœur de la tweetosphère politique : de 1 à 4 interactions fortes avec lui. Leurs liens ne sont pas représentés et ils sont disposés aléatoirement.







① Exemple de tweet dont la diffusion dans le cœur de la tweetosphère politique est circonscrite au voisinage proche d'une personnalité politique et modérément reprise dans la couronne.

② Exemple de tweet dont la diffusion est circonscrite à une orientation politique (la gauche au sens large) et largement reprise dans la couronne.

③ Exemple de tweet dont la diffusion est plutôt circonscrite aux personnalités politiques de droite et d'extrême-droite et relativement ignorée dans la couronne.

Le multimédia présent sur cette table montre comment les réseaux sociaux vont jusqu'à influencer directement le fonctionnement de nos démocraties. La campagne présidentielle en France (2016 – 2017) en est un bon exemple. Si on veut comprendre l'évolution de nos sociétés, on ne peut plus ignorer leur rôle dans la diffusion de l'information et la formation des opinions. Ces nouveaux médias sont des systèmes complexes où de multiples phénomènes collectifs s'auto-organisent, avec des règles qui diffèrent de celles des médias traditionnels : transmission d'informations plus décentralisée et horizontale, caractère viral de certains messages, multiplicité des acteurs…

Table 29

Les communs

Des groupes de citoyens recueillent des informations par leurs propres moyens pour ne plus dépendre des informations distillées à travers le filtre d'organismes privés ou publics, ayant éventuellement tel ou tel intérêt. Les données volontairement produites en biens communs ont une valeur de contre-expertise ou de contre-information. La donnée produite devient un outil mobilisable de manière collective qui produit des connaissances avec les habitants et peut devenir une aide précieuse à la décision des pouvoirs publics et servir au développement de leviers d'émancipation pour la société civile. Un exemple d'initiative réussie de récolte collaborative de données « à la source » est la mesure de la nuisance sonore à Paris dans le quartier de la Chapelle (Paris 18^{e}). Des habitants mobilisés par la problématique de nuisance sonore se sont équipés d'une application téléchargeable gratuitement appelée « SoundCity » (aujourd'hui « Ambiciti »). Elle a été développée par l'Institut national de recherche en informatique et en automatique (Inria) avec le soutien de la Ville de Paris dans le cadre de la Mission « Ville intelligente et durable ». D'autres exemples de projets collaboratifs de dimension internationale sont présentés : l'encyclopédie participative Wikipédia, la base de données géographiques OpenStreetMap et la base de données alimentaire Open Food Facts.

Ce sonomètre est muni d'un microphone. Il mesure le niveau de pression acoustique, une grandeur physique liée au volume sonore.

II.2.6 En guise de conclusion

Table 30

Expression publique
La concertation publique, organisée le 26 mars 2016 à la Cité des sciences et de l'industrie par Res publica (un cabinet de conseil en stratégie et ingénierie de la concertation), a permis de mettre en lumière des questionnements et des avis variés concernant les données personnelles et les technologies numériques du Big Data. Cette journée a été filmée. Nous vous proposons ici quelques minutes d'extraits.

Le numérique et vous
Les technologies numériques envahissent tous les secteurs de notre vie et brassent une profusion de données souvent personnelles. Elles nous poussent ainsi à nous interroger sur elles et les changements culturels qu'elles induisent, mais également sur nous-mêmes et sur nos pratiques. Un dernier multimédia vous propose de répondre à dix questions et vous donne l'occasion de vous situer par rapport aux autres visiteurs de l'exposition.

Table 31

Table de lecture
Vous trouverez ici le livre de l'exposition *Terra Data. Qu'allons-nous faire des données numériques ?* en vente à la boutique de la Cité des sciences et de l'industrie et en librairies. Édité chez Le Pommier, il a pour auteurs Serge Abiteboul et Valérie Peugeot. Bien évidemment, sa version numérique sera également présentée. Elle est disponible à l'achat sur les sites de librairies numériques. Comble du luxe, vous trouverez sur la table un espace vous permettant de recharger votre smartphone.

III Ressources

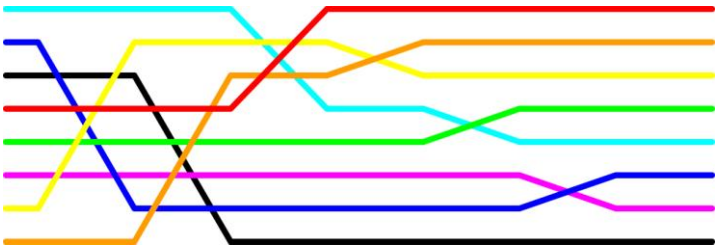
III.1 Au sein de l'exposition

À partir du 21 octobre, des facilitateurs proposeront tous les jours deux médiations à destination du grand public au sein même de l'exposition, de 11 h à 18 h.

1. Les algorithmes sur leur 31 : le tapis de tri

Les algorithmes sont partout, même dans la vie quotidienne ! Un algorithme, c'est une liste d'instructions simples qui mène à la réalisation d'une tâche plus complexe, et beaucoup d'actions peuvent être fractionnées ainsi : nouer une cravate, faire un gâteau, lacer ses chaussures… Pour traiter des données, les informaticiens ont inventé des algorithmes de tri. Ils permettent de ranger les données par ordre alphabétique, ou en fonction de leur date de création, de leur taille…

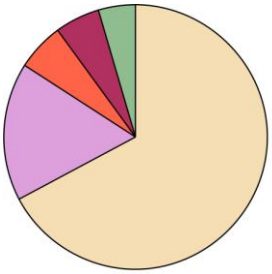
➔ En parcourant un réseau de tri dessiné sur un tapis, les participants se transforment en véritable machine humaine à trier.

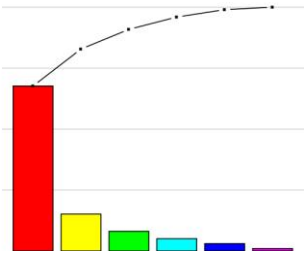


2. Jeu sur les représentations de données

Comment représenter le nombre de prix Nobel obtenu par un pays en fonction de la quantité de chocolat englouti dans ce même pays ? Faut-il utiliser un camembert ? Un histogramme ? Un graphique ?

➔ Les participants votent en groupe et s'ensuit une discussion pour déterminer la représentation la plus adéquate selon le jeu de données à manipuler. On verra aussi comment décrypter les « data visualisations » et leurs biais éventuels.





Des ateliers sont proposés aux groupes scolaires par les médiateurs de la Cité des sciences et de l'industrie.

Jusqu'à la fin de l'année scolaire 2016 – 2017

1. Le code, tout un programme !

Du CM1 à la 6e

Pourquoi programmer, et pour quoi faire ? Les élèves s'initient aux algorithmes de tri et au langage binaire. C'est sous forme de jeux que sont illustrés le traitement, le partage et la protection des données.

Objectifs

- Créer un algorithme simple pour coder des instructions.
- Comprendre le principe du langage binaire et la nécessité d'un langage de programmation.
- Sensibiliser les élèves à la gestion des données (stockage, cryptage et partage).

Déroulement

L'atelier commence par un jeu permettant de définir la notion de langage de programmation (langage binaire). Des mises en situation illustreront :

- les algorithmes (algorithmes de tri, de sélection …) ;
- le partage et protection des données.

2. Neurone artificiel, NA !

De la 5e à la terminale

Comment un ordinateur peut-il devenir champion du monde de Go ? Il a besoin de beaucoup de données, et de s'entrainer à jouer contre lui-même. Venez découvrir, au travers d'un exposé ludique, comment l'ordinateur apprend...

Objectifs

- Comprendre comment l'ordinateur lit les données.
- S'initier aux algorithmes, comprendre comment l'ordinateur utilise les données.
- Comprendre comment la révolution des « Big Data » a bouleversé les performances de certains algorithmes et ce, jusqu'aux limites de l'éthique.

Déroulement

Au travers d'anecdotes historiques, les participants découvrent comment l'ordinateur peut lire les données en utilisant le code binaire. Ils seront ensuite transformés en unité centrale, et exécuteront un algorithme simple permettant de faire des requêtes sur les bases de données.

Enfin, au travers d'un jeu, ils simuleront un algorithme appelé réseau de neurones. Cet algorithme de nouvelle génération existe grâce aux « Big Data » et leurs applications (reconnaissance faciale, devenir champion du monde de go, apprendre à reconnaitre un langage…) seront présentées.

La première partie sera développée plus longuement avec les classes de cycle 4, et la seconde avec les classes de lycées.

À partir de la rentrée scolaire de septembre 2017 et jusqu'au 22 décembre 2017

1. Les coulisses des « Big Data »

De la 5e à la 3e

Le « cloud », ça vous dit quelque chose ? Que se cache-t-il derrière ce nuage de données que l'on dit virtuel ? Venez découvrir, avec un algorithme simple, comment un ordinateur manie les données, et ce qu'il peut faire avec des données massives. Apprentissage et intelligence artificielle au programme…

2. Un nouveau monde, les « Big Data »

De la 2de à la terminale

Aujourd'hui, la quantité de données générée chaque seconde à travers le monde se compte en téraoctets. Quelles technologies ont rendu possible cette explosion ? Quels enjeux pour nos sociétés ? Venez découvrir ce nouveau monde qui bouleverse nos modes de vie, parfois aux frontières de l'éthique.

III.2 Suggestion bibliographique

. Serge Abiteboul et Gilles Dowek, *Le temps des algorithmes*, éd. Le Pommier, 2017.
. Sous la direction de Mokrane Bouzeghoub et Rémy Mosseri, *Les Big Data à découvert*, CNRS éditions, 2017.
. Emmanuel Lazard et Pierre Mounier-Kuhn, *Histoire illustrée de l'informatique*, éd. EDP Sciences, 2016.
. Dominique Cardon, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, éd. du Seuil, coll. La République des Idées, 2015.
. Pierre Delort, *Le big data*, éd. Presses Universitaires de France, coll. Que sais-je ?, 2013.
. Fabrice Demarthon, Denis Belbecq, Grégory Fléchet, *The big data revolution* dans *CNRS international magazine* n°28, January 2013.

→ Certains ouvrages de cette liste se trouvent à la **bibliothèque de la Cité des Sciences et de l'Industrie**, 30 avenue Corentin-Cariou, 75019 Paris.

Métro : Porte de la Villette (Métro ligne 7 ou Tramway ligne 3b).

Horaires : du mercredi au dimanche, 12 h – 18 h 45, le mardi 12 h – 19 h 45.

Description La bibliothèque met à votre disposition 120 000 documents (livres, revues, films, cédéroms, DVD) dans tous les domaines scientifiques et techniques. Possibilité de consultation sur place et d'emprunt de documents.

IV Informations pratiques

Adresse
Cité des sciences et de l'industrie
30 avenue Corentin-Cariou
75019 Paris
www.cite-sciences.fr

Accès
Métro : Porte de la Villette (L7)
Bus : 139, 150, 152
Tramway : Porte de la Villette (Ligne 3b)

Horaires d'ouverture
Du mardi au samedi de 10 h à 18 h, le dimanche de 10 h à 19 h.
Fermeture le lundi ainsi que les jours fériés suivants : 1er janvier, 1er mai et 25 décembre.

Élémentaire : 1 gratuité pour 12 entrées payantes
Secondaire : 1 gratuité pour 15 entrées payantes

Tarifs scolaires (valables du 1er septembre 2016 au 31 août 2017)
4,50 € (2,50 € pour les établissements en ZEP)

Tout billet acheté donne droit à une entrée au *Cinéma Louis Lumière* et au sous-marin *Argonaute* (dans la limite des places disponibles) + un accès aux ateliers et au Planétarium sur réservation.

Réservation groupes

Sur internet (devis en ligne)
http://www.cite-sciences.fr/fr/vous-etes/enseignants/votre-sortie-scolaire/infos-pratiques-et-reservation/devis-en-ligne/

resagroupescite@universcience.fr

01 40 05 12 12

01 40 05 81 90

Cité des sciences et de l'industrie
Service groupes
30 avenue Corentin-Cariou
75930 Paris Cedex 19