En 1956, IBM a commercialisé le premier disque dur commercial. Celui-ci pouvait stocker un peu plus de 4 mégaoctets.

À l'heure actuelle, la plus grande clé USB disponible dans le commerce possède une capacité de 2 téraoctets, soit 2 billions (millions de millions) de mégaoctets.

Le mot «téra» est issu du grec et signifie «monstre». Cette énorme unité de mesure est déjà trop petite: Google traite près de 24 pétaoctets par jour.

On estime qu'un cerveau humain peut contenir 2,5 pétaoctets de données.

D'ici 2025, la quantité de données stockée dans le monde devrait atteindre 163 zettaoctets, soit environ dix fois plus qu'en 2016.

Méga, giga, téra, …

1 bit	(plus petite unité numérique)	
1 octet	= 8 bits	
1 kilooctet	= 1000 octets	
1 mégaoctet	= 1000^2 octets ou 1000 kilooctets	
1 gigaoctet	= 1000^3 octets ou 1000 mégaoctets	
1 téraoctet	= 1000^4 octets ou 1000 gigaoctets	
1 pétaoctet	= 1000^5 octets ou 1000 téraoctets	
1 exaoctet	= 1000^6 octets ou 1000 pétaoctets	
1 zettaoctet	= 1000^7 octets ou 1000 exaoctets	
1 yottaoctet	= 1000^8 octets ou 1000 zettaoctets	

Impressum
SATW Technoscope 01/20 | Février 2020
www.satw.ch/technoscope
Concept et rédaction: Beatrice Huber | Ester Elices
Collaboration rédactionnelle: Christine D'Anna-Huber | Alexandra Rosakis
Graphisme: Andy Braun
Photos: Adobe Stock
Photo de couverture: Adobe Stock
Traduction: Ars Linguae
Impression: Egger AG

Abonnement gratuit et commandes supplémentaires
SATW | St. Annagasse 18 | CH-8001 Zürich
technoscope@satw.ch | Tel +41 44 226 50 11

Technoscope 2/20 paraîtra en mai 2020 sur le thème «Food»

satw it's all about technology

Les recherches sur le Big Data

Après avoir collecté de grandes quantités de données, le travail n'est pas terminé pour autant. Les données doivent ensuite être enregistrées, gérées, traitées, analysées, représentées et reliées. Ces tâches sont facilement gérables pour les petits ensembles de données, mais constituent un véritable défi pour le Big Data. La recherche est sollicitée, par exemple au sein du Programme national de recherche 75 «Big Data» (PNR 75). Les projets de recherche menés dans le cadre du PNR contribuent à résoudre les grands problèmes actuels et à relever les défis auxquels est confrontée la Suisse. Le PNR aborde notamment les domaines de recherche suivants:

Le temps nécessaire aux algorithmes pour traiter les données est un élément crucial pour le Big Data. L'objectif est de concevoir des algorithmes plus rapides et plus performants, en particulier dans le domaine de l'apprentissage automatique.
http://bit.ly/nfp75_algorithmen

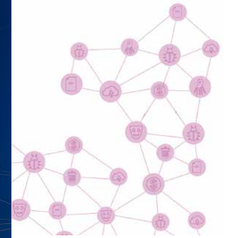
L'enregistrement correct et ergonomique d'énormes quantités de données constitue un autre défi. L'analyse des données en temps réel est une approche visant à réduire le besoin de stockage dans laquelle les données entrantes ne sont pas d'abord stockées, mais directement analysées. L'objectif est de faire en sorte que ces systèmes de traitement de flux de données soient également utilisables par les spécialistes de disciplines autres que l'informatique.
http://bit.ly/nfp75_Datenstromanalytik

Les entrées de base de données peuvent comprendre plusieurs centaines de colonnes, ce qui n'est pas pris en considération dans les langages de programmation actuels et complique donc l'utilisation des bases de données. C'est pourquoi un autre projet de recherche vise à améliorer l'interaction des différents langages de programmation et bases de données.
http://bit.ly/nfp75_Datenstromanalytik

Le Big Data représente à lui seul un vaste domaine de recherche dans lequel de nombreuses questions restent sans réponse. La participation des chercheurs de données, mais également des spécialistes d'autres disciplines, est nécessaire pour que le Big Data puisse développer son plein potentiel.

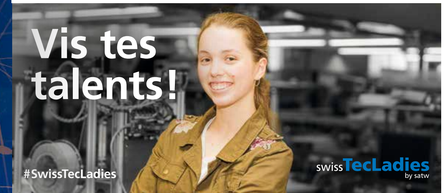
Le Big Data à l'école

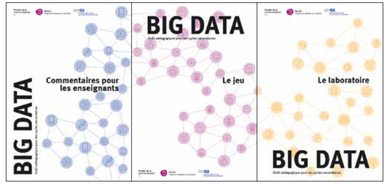
Le Big Data est un sujet complexe mais trop important pour ne pas être compris. C'est pourquoi le Programme national de recherche PNR 75 «Big Data» et le Musée de la communication ont uni leurs forces et développé un outil pédagogique attrayant sur le Big Data pour les élèves du niveau secondaire I et II.



Avec cet outil pédagogique, qui est divisé en plusieurs parties, l'informatique peut être comprises en se basant sur le quotidien des jeunes. Les élèves savent ce qu'est le Big Data et jettent un regard critique sur le sujet. De cette façon, ils prennent conscience qu'ils sont tous concernés. Car les jeunes ne sont pas de simples consommateurs. C'est à eux qu'il appartient de façonner l'avenir.

Outil pédagogique «Big Data»
www.mfk.ch/fr/bigdata





«Big Data – Le jeu»
Dans la première partie, une approche ludique est choisie en guise d'introduction. Il s'agit de résoudre une énigme. Une jeune américaine tombe inopinément enceinte. Elle ne le dit à personne et essaie de se comporter aussi discrètement que possible. Même son père ignore tout. La chaîne de supermarchés «Target» envoie à sa fille des bons de réduction pour des vêtements de grossesse et des articles pour bébé. Pourquoi donc? Une question clé est posée aux élèves au début de la période. Au cours du jeu, ils collectent des informations qui les aideront à répondre ensemble à la question et à résoudre l'énigme.

«Big Data – Le laboratoire»
La seconde partie de l'outil pédagogique approfondit sept aspects du «Big Data». Selon le niveau et les connaissances préalables de la classe, on choisira parmi ces sept aspects.

- «Pourquoi des secrets?» – Vie privée
- «Rien n'est gratuit.» – Données précieuses
- «Que nous donne-t-on à voir?» – Bulle de filtres
- «Car ils savent exactement ce que nous faisons…» – Surveillance totale
- «Comment se protéger sur Internet?» – Protection des données
- «Sauver le monde avec le Big Data?» – Opportunités offertes par le Big Data
- «Rendez-nous nos données!» – Open Data

«Big Data – Le trail»
La dernière partie de l'outil pédagogique est un jeu éducatif multimédia au Musée de la communication de Berne. En termes de contenu, «Le trail» complète «Le Laboratoire», mais il peut aussi être joué sans avoir traité le sujet au préalable en classe. «Le trail» se déroule dans le futur, en 2080. Les personnes ont complètement abandonné le contrôle de leurs données au groupe industriel Amathron. Mais les rebelles se défendent et veulent récupérer le contrôle de leurs données. Les joueurs se trouvent dans un simulateur d'entraînement. Ils participent à des unités d'entraînement pour pouvoir se confronter aux défis et à l'aventure dans le monde numérique. À la fin des unités d'entraînement, les joueurs, devenus activistes, démarrent leur première mission. Ils sont envoyés dans le passé – donc notre présent – pour y influencer les décisions importantes.

Choix d'études et de carrière

Corinne Giroud, Office cantonal d'orientation scolaire et professionnelle – Vaud

A l'école, nous avons pu jouer à un serious game sur le Big Data et notre prof d'informatique nous a expliqué comment nos données personnelles peuvent être utilisées. Comment devient-on spécialiste des données et dans quels domaines peut-on travailler? Marian, 14 ans

Salut Marian,
Tu sais maintenant que tes messages et tes photos publiées sur un réseau social, ton adresse e-mail inscrite sur un site d'achat en ligne et tes recherches sur Internet laissent des traces numériques pour toujours. Prudence, donc! Il en va de même pour toutes les actions que nous réalisons en ligne pour par exemple organiser une sortie, chercher le meilleur itinéraire ou toute autre information. Ces interactions numériques et leurs données représentent une véritable mine d'or - pour autant qu'on sache les extraire, les analyser et les valoriser. Le Big Data a fait émerger de nouveaux métiers où l'informatique, la statistique et les systèmes de communication se disputent la première place.

De nombreux domaines concernés
Le numérique a révolutionné la science des données. Obtenir en temps réel le meilleur itinéraire sur la base du trafic en cours et des moyens de transport disponibles, profiter des avancées de la science médicale ou améliorer ses performances sportives grâce à l'analyse de données recueillies dans les hôpitaux ou les salles de fitness: ces exploits sont le fruit du travail des spécialistes des données (data scientists) Du côté de l'économie et des entreprises, l'intérêt est énorme: les quantités de données disponibles peuvent servir à développer de nouvelles stratégies et atteindre de nouveaux clients. La statistique et l'informatique sont complétées ici par des compétences en marketing digital. Quant à la question de la sécurité des données, une préoccupation largement partagée, elle est le domaine d'expertise des ingénieurs en cybersécurité.

Polyvalence
La polyvalence caractérise les spécialistes des données: rechercher des données et les analyser à l'aide de logiciels spécifiques, puis les exploiter et les valoriser, tout cela demande des compétences en informatique, en mathématiques et en statistiques, ainsi qu'une maîtrise de l'anglais technique. De la rigueur et de l'organisation, de la curiosité, une capacité d'analyse et d'abstraction sont des qualités requises de professionnels qui passent une bonne partie de leur temps à se tenir informés des évolutions fulgurantes dans leur domaine.

Zoom sur quelques formations du big data

Data scientist
Ingénieur / Ingénieure en informatique et systèmes de communication
Statisticien / Statisticienne
Il existe de nombreuses possibilités de formation dans le domaine de la science des données (big data) en Suisse, de différents niveaux. Voir **orientation.ch** > Formations > Big Data.

TechnoScope
by satw 1/20

Big Data

Le traitement de tes données

www.satw.ch/technoscope

Big Data

Acquérir des connaissances à partir des données

Smartphones, cartes de crédit, GPS ou fitness trackers: où que nous allions, nous laissons une trace de nos données. Et cela ne se limite pas aux personnes, de plus en plus de capteurs d'appareils et de machines sont reliés à Internet et produisent également des données: jusqu'à 4 téraoctets de données par jour par exemple pour une voiture autonome, un téraoctet équivalant à un 1 suivi de 12 zéros. Un nombre difficilement imaginable.

Le Big Data implique que de plus en plus de données sont générées à partir de sources très diverses. Aujourd'hui, ces données peuvent être conservées de manière plus efficace et moins coûteuse sur des disques durs et dans des mémoires flash, et de plus en plus souvent dans le cloud. Les experts estiment qu'à l'avenir, la quantité de données stockée continuera de doubler tous les trois ans.

Des informations précieuses obtenues grâce aux données

Pourtant, la particularité du Big Data ne tient pas aux énormes quantités de données, mais plutôt à la capacité des ordinateurs toujours plus puissants de tirer des informations précieuses de ces montagnes de données. La technologie a littéralement dopé le traitement des données: des logiciels spécifiques et des algorithmes capables d'apprentissage analysent instantanément les montagnes de données enregistrées et recherchent des modèles, ainsi que des corrélations et des règles cachées, permettant ainsi de tirer des conclusions et de dégager des tendances dans d'innombrables domaines – de la recherche et la médecine à l'agriculture, en passant par les finances et le marketing. Ou faut-il craindre une mauvaise récolte? Quels patients répondent le mieux à un médicament? Quel groupe de clients le nouveau produit pourrait-il intéresser?

Chaque jour, 294 milliards d'e-mails parcourent le monde. À cela s'ajoutent 5 milliards de requêtes, 65 milliards de messages WhatsApp et 500 millions de tweets.

Le Big data expliqué en 3 minutes:
https://www.youtube.com/watch?v=uH813u7_b0s

D'où proviennent toutes les données du Big Data?

Une recherche de traces

Toute activité sur Internet est enregistrée: quel film YouTube tu as regardé, quels produits tu as achetés, quelle est ta musique préférée, combien de likes tu as reçu sur Instagram, qui sont tes amis, quelle est ta condition physique (grâce à la Smartwatch), ou tu es allé (grâce au vélo de location). Rien n'est perdu: les données que tu n'as peut-être jamais fournies, p. ex. ton âge, ton niveau d'études, tes hobbys. Ces données personnelles valent de l'or pour de nombreuses entreprises car elles leur permettent d'ajuster leurs services et leurs publicités.

L'échange de données

Les données sont largement vendues et achetées. De nombreux utilisateurs n'y voient aucun inconvénient car ils «n'ont rien à cacher». Mais les données peuvent aussi être utilisées de manière abusive ou bien une personne peut être pénalisée à cause de ses données – p. ex. par une prime d'assurance-maladie plus élevée en raison d'habitudes malsaines – ou même être la victime d'une usurpation d'identité. En Chine, la surveillance par l'État, ainsi que la récompense ou la sanction des citoyens sur la base de leurs données personnelles, sont déjà une réalité. En Suisse, la protection des données est réglementée par une loi fédérale. Cette loi est actuellement remaniée afin de mieux répondre aux nouvelles réalités d'Internet.

Les données ouvertes: une opportunité

Mais le Big Data ne se limite pas aux données personnelles. Les «Open Data» sont librement accessibles et peuvent être utilisées et traitées par tout le monde. Il s'agit par exemple de portails de connaissances, tels que Wikipedia, ou de statistiques qui sont publiées par l'Office fédéral de la statistique, d'informations routières, de données sur les incidents environnementaux actuels, mais également de logiciels et d'œuvres artistiques telles que des photos ou des vidéos. La plateforme Open-Data des transports publics suisses met notamment à disposition des données sur les horaires et les arrêts à partir desquels il est possible de concevoir des applications ou d'autres produits ou d'établir des statistiques.

Où suis-je?


©Verkehrsbetriebe Zürich

La mobilité du futur
Les données contribuent à une meilleure fluidité du trafic

Imagine que tu doives choisir la liaison de transport public la plus rapide pour arriver à l'heure à ton rendez-vous mais que tu disposes seulement de l'indicateur des chemins de fer officiel suisse au format papier – un véritable défi. Heureusement, beaucoup de choses ont changé depuis la publication du premier indicateur en 1905. Désormais, différentes applications t'évitent d'avoir à chercher la meilleure liaison, te vendent directement le billet correspondant et relient les informations concernant les changements de train à des données actuelles comme les retards ou les déviations.

Une meilleure fluidité du trafic grâce aux connexions live

La ville de Zurich souhaite aller plus loin encore et interconnecter en live tous les usagers – transports publics, car sharing, taxi, location de vélos sur **www.stadt-zuerich.ch/vbz/de/index/mobilitaet-der-zukunft**. Même les véhicules privés communiqueront entre eux et avec l'infrastructure des transports afin de contribuer à une meilleure fluidité du trafic. Google Maps permet déjà de s'informer en temps réel sur les embouteillages. Pour cela, Google détecte la position déterminée par GPS des smartphones et, d'après leur nombre et leur localisation, reproduit l'état de la circulation sur la carte géographique.

Des données qui servent à nous orienter

En parlant de carte géographique, ce sont les Big Data qui nous permettent aujourd'hui de trouver si facilement un chemin menant de A à B. Google travaille avec d'innombrables partenaires qui fournissent des coordonnées détaillées permettant de représenter les réalités géographiques de manière précise et actuelle. Street View fournit également des données permettant d'améliorer la carte en lisant les panneaux de signalisation, puis en les comparant à la carte. Les données satellite, quant à elles, sont utilisées pour détecter les modifications géologiques ou structurelles éventuelles. Autre avantage: chacun de nous peut contribuer à rendre la carte plus utile en fournissant des photos ou des évaluations. Des algorithmes, ainsi que des collaborateurs en chair et en os, gèrent, comparent et relient ces ensembles de données multicouches afin d'améliorer les cartes en continu.

Big Data by you

en jouant à un jeu. L'application **allyscience.ch** permet par exemple aux personnes atteintes du rhume des foins de documenter leurs maux actuels et de contribuer ainsi au développement des futurs systèmes de prévention et aux thérapies pour les personnes allergiques au pollen.

De nombreuses questions scientifiques nécessitent la collecte d'un grand nombre d'ensembles de données. Il ne s'agit pas seulement de paramètres complexes mesurés en laboratoire, mais également de données que nous pouvons tous collecter dans notre entourage immédiat. Citizen Science – une recherche menée par et pour les citoyens – regroupe des amateurs de sciences dans le but de collecter rapidement de telles données.

Tout le monde aide tout le monde

En général, ce sont les chercheurs qui lancent et dirigent les projets, analysent les données et les mettent à la disposition du public. Selon le projet, ils sont impliqués dans la saisie des données, mais également dans la planification des projets et l'analyse des résultats. L'université de Zurich et l'ETH Zurich ont notamment créé le centre de compétences Citizen Science **https://citizenscience.ch/**. L'éventail des questions de recherche qui peuvent être traitées avec la participation des bénévoles est large. Différents projets sont présentés sur le site Web **www.schweiz-forscht.ch**, de la prolifération des salamandres noires au code quantique, en passant par les différents dialectes en Suisse. Toutes les personnes intéressées peuvent s'inscrire et contribuer à la recherche en observant, en collectant, en photographiant ou même

Une qualité des données irréprochable

Lors de cette saisie variable de données, le contrôle de la qualité est toutefois difficile. En particulier dans une saisie de données très complexe, il peut facilement arriver que les données soient faussées involontairement. Outre le contrôle direct de la communauté, les données sont donc vérifiées au moyen d'algorithmes afin d'éliminer les entrées erronées.

Avis identiques dans la bulle

Comment distinguer les fausses nouvelles des vraies nouvelles? Autrefois, des journaux réputés parcouraient la jungle d'informations, expliquaient les nouvelles et les classaient. À l'ère d'Internet, cette jungle d'informations est devenue plus dense, les sources d'informations sont plus nombreuses, et leur origine et leur véracité sont discutables. Parallèlement, les médias classiques ont perdu leur rôle de «gardien». En particulier les jeunes leur ont tourné le dos. Ils préfèrent s'informer via des agrégateurs de nouvelles tels que Reddit, des blogs et des forums en ligne. En Suisse, selon l'étude James de 2018, environ 51% de tous les jeunes utilisent chaque jour les réseaux sociaux comme source d'informations.

Des nouvelles personnalisées

Les agrégateurs de nouvelles se basent sur des algorithmes. Ces derniers font ce que faisait une rédaction autrefois: ils sélectionnent les informations à présenter aux lecteurs. Mais à l'ère du Big Data, cette sélection est précisément adaptée aux différents clients. En utilisant toutes sortes de données

telles que l'origine, l'âge et le sexe, combinées aux requêtes, aux consultations de pages et aux habitudes de navigation d'un utilisateur, les algorithmes établissent le profil de personnalité de celui-ci et en déduisent des informations susceptibles de l'intéresser ou de le divertir. Aucune autre nouvelle ne lui est présentée. À long terme, cela lui donne l'impression que le monde est constitué de gens ayant tous la même opinion que lui. Toute personne qui n'est jamais stimulée par l'expression d'avis différents ni poussée à réfléchir ou à apprendre se retrouve alors prisonnière de la bulle de filtres.

Sortir de la bulle

Le meilleur antidote est la diversité: autrement dit, trouver de quelle manière un thème est présenté ailleurs – peut-être dans les médias traditionnels pour changer – regarder qui se cache derrière une nouvelle, quelle est la fiabilité de la source et à qui s'adresse le message, puis se faire sa propre opinion.

Informations complémentaires:
https://www.jeunesetmedias.ch/fr/themes/fake-news-manipulation.html

Chocolat, prix Nobel et conclusions erronées

Les algorithmes intelligents permettent d'obtenir des renseignements précieux à partir de différents ensembles de données. Par exemple, de savoir pourquoi il y autant de lauréats du prix Nobel en Suisse. Étonnamment, cela s'expliquerait par le fait qu'en Suisse, la consommation de chocolat par habitant est plus élevée qu'ailleurs. C'est en tout cas que ce semble prouver l'illustration ci-dessous. Ou serait-ce totalement faux?

1 Suisse
2 Suède
3 Danemark
4 Autriche
5 Norvège
6 Royaume-Uni
7 Allemagne
8 Irlande
9 Pays-Bas
10 USA
11 France
12 Belgique
13 Finlande
14 Canada
15 Australie
16 Italie
17 Portugal
18 Grèce
19 Espagne
20 Japon
21 Brésil
22 Chine

Prenons un autre exemple: en Suisse, depuis la Seconde Guerre mondiale, aussi bien le nombre de cigognes que le nombre de bébés par famille ont diminué. Cela prouve-t-il que les cigognes amènent les bébés, comme on le croyait autrefois? Bien sûr que non. Le fait que les cigognes désertent nos toits aujourd'hui est certes malheureux, mais n'explique en rien la diminution du nombre de naissances.

L'étudiant de Harvard, Tyler Vigen, a relevé toute une série de corrélations trompeuses. Bon nombre d'entre elles sont très drôles. Mais elles posent un problème bien connu des statisticiens: «La corrélation n'implique pas de causalité». La relation entre deux éléments ne signifie pas forcément que l'un influence l'autre.

Il existe une jolie expression en informatique, «Garbage in, garbage out», qui signifie grosso modo «À données inexactes, résultats erronés». En termes de Big Data, cela signifie que la pertinence d'un résultat fourni par un algorithme ne dépend pas seulement des données qui l'alimentent, mais également de la manière dont il est programmé pour calculer les bons paramètres.

Des corrélations trompeuses:
http://www.tylervigen.com/spurious-correlations