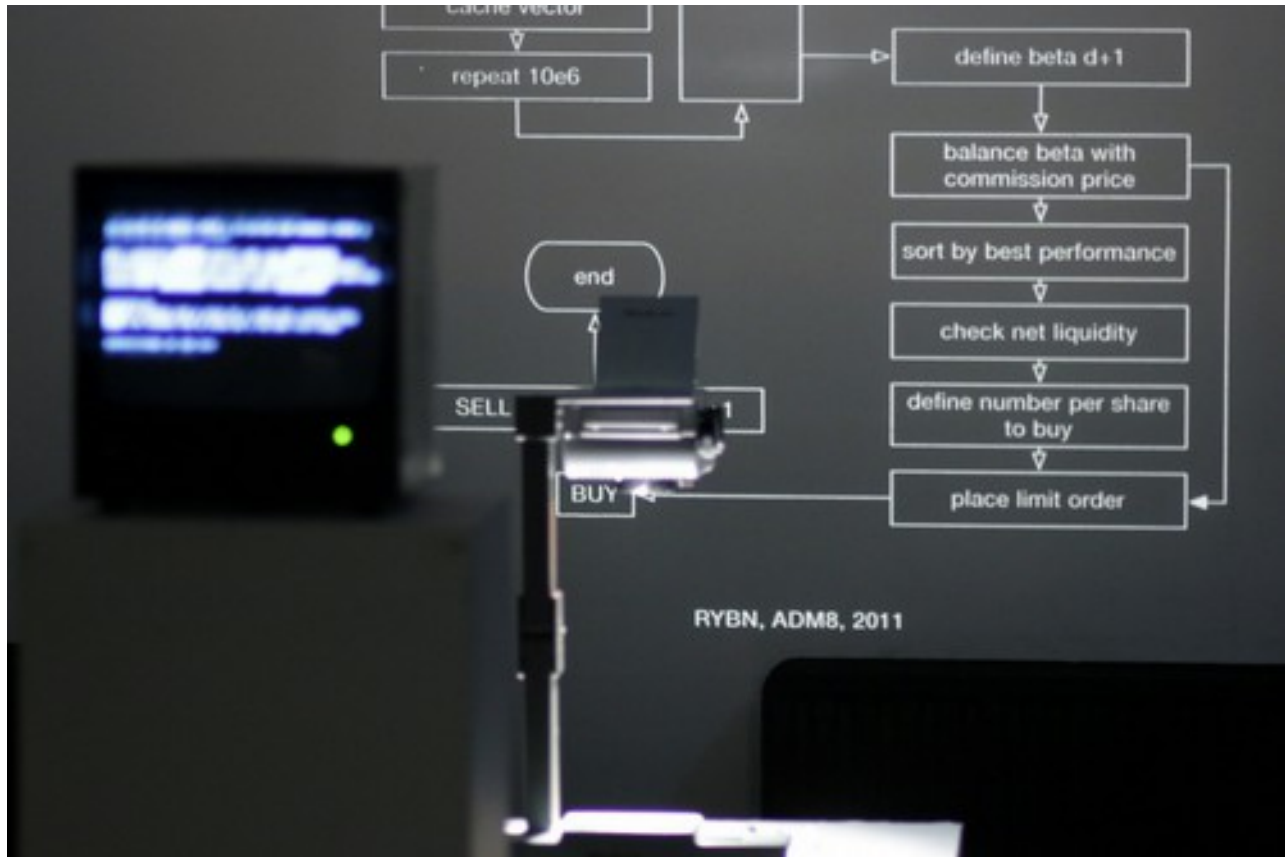


Dossier pédagogique

DATA VIZ ET DATA ART



RyBN, ADM8 (détail), 2011-2016

Donner forme aux données

Une exploration des données p. 2

Le traitement des datas p. 3

Rendre les datas perceptibles

Un processus d'interconnexions p 4

Le Data-Art p 6

Petit lexique p 7

À découvrir p 8

Donner forme aux données

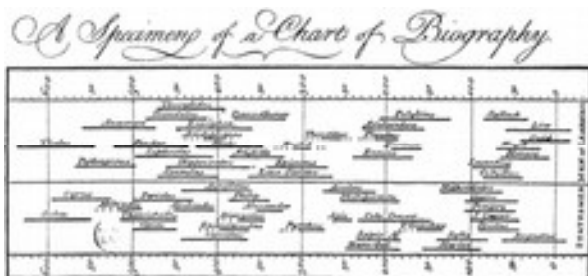
Une exploration des données

L'une des étapes les plus importantes dans l'analyse d'une série de données est sa représentation graphique. Peu à peu, divers moyens de présenter des données en fonction du type d'informations - concepts, stratégies, métaphores, informations combinées - voient le jour.

La génération de modèles à partir d'un grand nombre de données n'est pas un phénomène récent. En Chine, on prête à l'empereur Tang Yao la volonté de recenser les récoltes en 2 238 av. J.-C. En Égypte, le pharaon Amasis organise le recensement de sa population au Ve siècle av. J.-C. Mais ce n'est qu'au XVII^e siècle que l'on commence à vouloir analyser les données pour en rechercher des caractéristiques communes et au XVIII^e que commencent les recherches liées à leur représentation.

Histogramme et diagramme

En 1765, soit vingt ans avant les premières publications de William Playfair dont nous parlerons juste après, le chimiste Joseph Priestley a le premier recours à une frise chronologique dans laquelle des barres superposées de différentes longueurs permettent de comparer les époques auxquelles vivaient différentes personnes.



A specimen of a Chart of Biography, 1765

En 1786, l'ingénieur et économiste écossais William Playfair publie *The Commercial and Political Atlas*, premier à offrir des graphiques statistiques. Cet ouvrage comporte 43 séries chronologiques et un histogramme. Persuadé que les dessins valent mieux que des tableaux de chiffres, il développe par la suite une visualisation sous forme de diagramme circulaire. Ses graphiques de séries chronologiques passent encore aujourd'hui pour des modèles de clarté.



Statistical Breviary (excerpt), 1801

Influençant de nombreux domaines de recherche, Florence Nightingale s'appropriera très vite le principe des diagrammes circulaires développés par William Playfair. Après la Guerre de Crimée, elle développe une charte graphique et une version améliorée de ces diagrammes (équivalant aux histogrammes circulaires d'aujourd'hui), afin d'illustrer les causes saisonnières de mortalité des patients de l'hôpital militaire qu'elle gère.

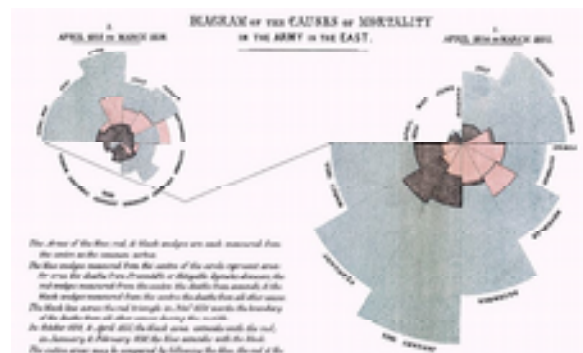


Diagram of the causes of mortality in the army in the East, 1858

Le mot histogramme quant à lui a été proposé par Karl Pearson en 1895 alors qu'il tentait d'appliquer le principe des méthodes statistiques à l'étude de la sélection naturelle de Darwin.

Collecte et statistiques

Dans l'objectif d'accélérer et d'améliorer le passage aux caisses tout en collectant les données de vente afin de favoriser la disponibilité des produits et l'efficacité des promotions, le code à barres est imaginé en 1949 par Norman Woodland et Bernard Silver, tous deux ingénieurs. Ce premier symbole, surnommé Bull's eye, était composé de cercles concentriques mais il ne sera jamais utilisé en pratique.



Bull's eye, 1949

Les années 50 voient l'apparition du calculateur et des techniques de calcul sur ordinateur. Simultanément des méthodes et des techniques d'analyse de données voient le jour, telles que la segmentation, les réseaux de neurones, les algorithmes génétiques, puis les arbres de décision et la méthode des centres mobiles.

Le traitement des datas

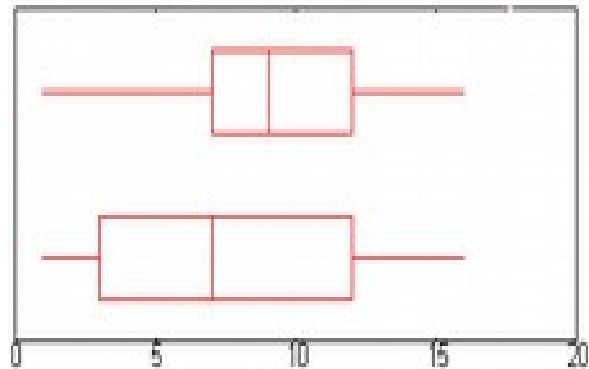
Les données ou « datas » sont aujourd'hui disséminées dans nos systèmes, notamment sur le web, sans structure, sous des formes variées. La capacité de comprendre ces données et d'en tirer une information utile est désormais de plus en plus valorisée et surtout rendu possible par la puissance des outils informatiques.

La Data visualisation

Les données que l'on possède peuvent ainsi être extraites de différentes sources pour ensuite être converties, testées, nettoyées, complétées et, enfin, croisées et représentées par les Data Visualisations, aussi appelées Data Viz. L'institut de gestion des médias et de la communication de l'université de Saint-Gallen en Suisse s'est ainsi spécialisé dans le domaine de la Data Visualisation. Ralph Lengler et Martin J. Eppler y ont développé en 2006 un tableau périodique des méthodes de visualisation. On peut y voir une liste non exhaustive de divers moyens de présenter des données en fonction du type d'informations et la centaine de grands classiques des différents types de visualisations.

Ces techniques permettent aux chercheurs d'exploiter et de découvrir des modèles d'utilisation de données de plus en plus précis.

En 1977, le statisticien américain John Tukey publie, avec John Wilder, un livre de référence dans le domaine de l'analyse des données statistiques : *Exploratory Data Analysis*. Il y présente le principe de la boîte à moustaches, ou diagramme de quartiles, un diagramme de représentation de données quantitatives particulièrement simple. En effet, la boîte à moustaches résume seulement quelques caractéristiques de position du caractère étudié (médiane, quartile, minimum, maximum ou déciles). Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes.



Comparaison de deux diagrammes en boîte à moustaches, 1977

Dès les années 80, l'arrivée du micro-ordinateur stimule la recherche et les analyses statistiques se font dès lors plus nombreuses et précises.

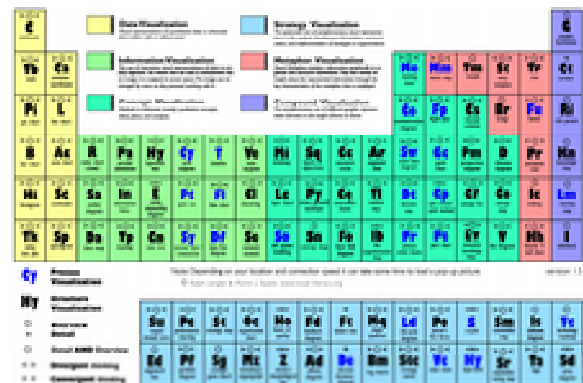
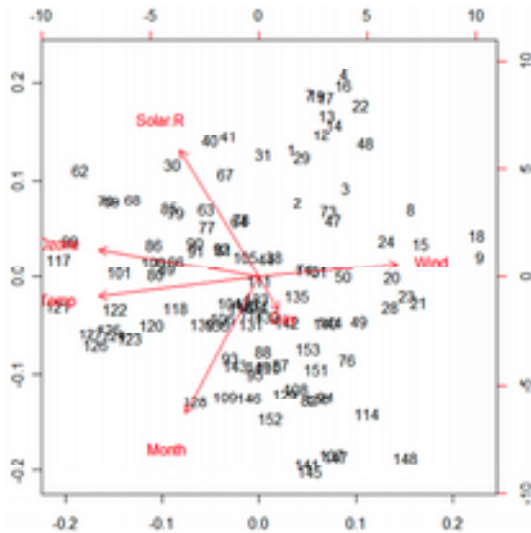


Tableau périodique des méthodes de visualisation (extrait), Ralph Lengler et Martin J. Eppler

Le Data Mining

L'exploration de données ou « Data Mining » a pour objet l'extraction d'un savoir d'une connaissance à partir de grandes quantités de données par des méthodes automatiques ou semi-automatiques. Utilisant un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, cette exploration amène à construire des modèles, à trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, afin d'en extraire un maximum de connaissances utiles, souvent dans le domaine de l'informatique décisionnelle.



Exemple d'analyse basée sur les données de la qualité de l'air

L'Open Data

Les grands volumes de données ou « Big Data » à traiter obligent aussi bien les professionnels du numérique et de l'information que les utilisateurs, à changer non seulement d'outils de conception, de partage et d'utilisation, mais également de mode de penser les contenus. Ainsi, alors que les technologies de l'information et de la communication connaissent une relative stabilisation des mutations technologiques, s'engage la bataille de la donnée. L'idée d'un web ouvert, offrant un accès à toutes les données, prend forme et s'oppose à la rétention d'information.



Nuage de l'Open Data, Richard Cyganiak, 2007

Rendre les datas perceptibles

Un processus d'interconnexions

« Tous ces territoires, en relation les uns avec les autres, forment eux-mêmes un seul et même système fondé sur des codes de fuite capables d'organiser des formes à l'intérieur de la communauté humaine. » Olivier Auber, artiste et chercheur

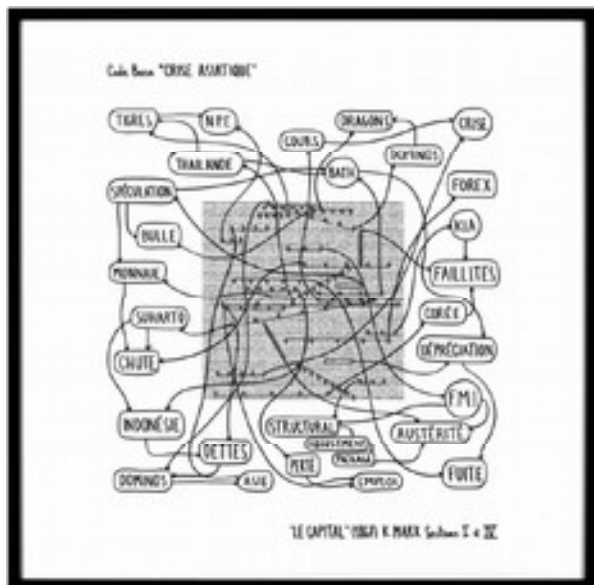
Des cartes heuristiques

Pour Electronic Shadow, chaque entité de mémoire, chaque information, se définit tant par sa valeur propre que par le réseau de connexion, volontaire ou non, qu'elle constitue et/ou auquel elle appartient. Ces multiples connexions mettent en place une toile de liens potentiels infinis entre les différentes entités. Dans *Palimpseste 3D Navigator*, c'est le cœur d'Internet qui fonctionne par une mise en interconnexions de réseaux d'informations. La matière première, ce sont les mots-clés qui représentent aussi bien l'information à laquelle ils renvoient, à savoir des pages web, que des accès à d'autres nœuds d'interconnexions.



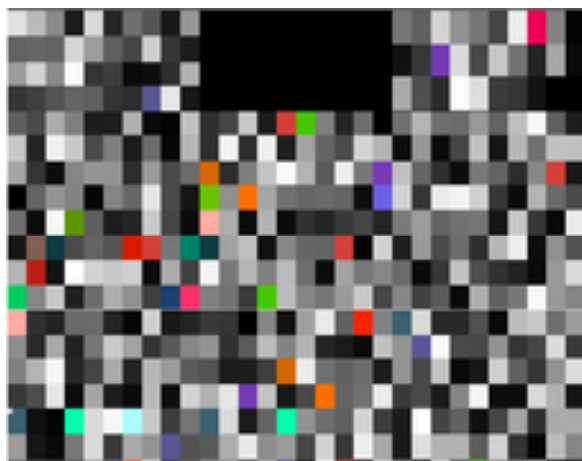
Palimpseste 3D Navigator, Electronic Shadow, 2001

Julien Prévieux tente plutôt de décrypter les significations cachées. Sa méthode de décryptage est connue sous le nom de « code de la Bible », un code appliqué par les moines au Moyen Âge pour extorquer des significations cachées aux textes sacrés. Dans les marges du *Capital* ou de *La Richesse des nations* se déploie ainsi une nébuleuse de mots-clés (dates, faits, personnalités) reliés par des flèches qui matérialisent des rapports d'appartenance ou de cause à effet. Au terme d'un fastidieux travail de déchiffrement, il dresse le portrait d'une réalité obscure et chaotique, égrenant scandales financiers et crises économiques en lieu et place du miracle attendu.



In Search of Economic miracle, Julien Prévieux, 2006-2007

Questionnant les réseaux sociaux et les « liens d'amitié », *The Big Picture* est une expérience artistique menée par Albertine Meunier et Olivier Auber permettant de suivre les mises à jours de statuts dans Facebook, de manière collective ou individuelle. Ce projet de détournement low tech des relations personnelles à l'œuvre sur les réseaux sociaux évolue en permanence et propose de créer une image vivante de statuts d'amis tout en gardant trace de cette mouvance.



The Big Picture, Olivier Auber et Albertine Meunier, 2008

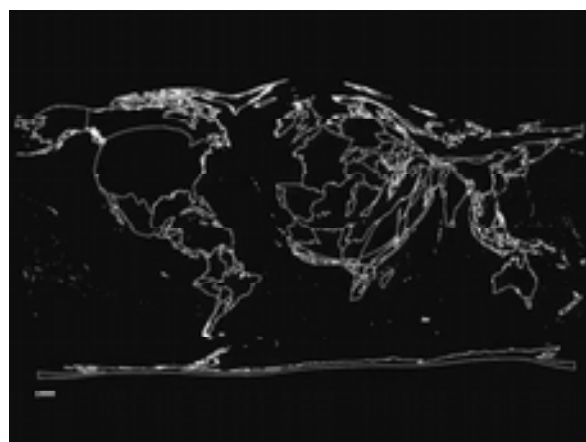
Des environnements interconnectés

Depuis 2005, Jonathan Harris et Sep Kamvar récoltent des sentiments humains à partir du web. Toutes les deux minutes, *We feel fine* recherche les entrées récemment publiées sur le réseau et contenant les mots « *I feel* » et « *I am feeling* ». Toutes les informations liées à la publication trouvée sont alors enregistrées - phrase, date et heure, auteur, sexe, localisation géographique, conditions météorologiques locales... - dans l'objectif d'identifier le « sentiment » exprimé dans cette publication - triste, heureux, déprimé... Le résultat est une base de données de plusieurs millions de sentiments humains, augmentant de 15 000 à 20 000 nouvelles sensations par jour.



We feel fine, Jonathan Harris et Sep Kamvar, depuis 2005

Pour Tim Schwartz, il s'agit avant tout de filtrer l'information, cette quantité énorme de données à travers une logique intuitive où l'Homme est sujet à étude. Dans *Geohistoriography*, il applique les principes du Data Mining en développant sa propre application et extrait du New York Times les chiffres correspondant au nombre d'articles écrits au sujet d'un pays chaque année à partir de 1851. Cette installation retrace alors la manière dont l'Amérique voit le monde à travers le prisme de ce média et met en évidence le changement de perspective dans le traitement des informations au cours des 150 dernières années.



Geohistoriography, Tim Schwartz, 2009

Stanza quant à lui interroge les données issues des capteurs qui parsèment aujourd'hui notre environnement – qualité de l'air, vitesse du vent, baromètre... *Data data data* interprète ainsi en temps réel les données transmises par ces capteurs avant de les projeter sur le tissu urbain. À travers cette projection brute des données, ce projet tente de transmettre des informations sur la structure même de nos villes, structure dans laquelle nos actions se répercutent sur l'environnement et où le monde physique de liquides et de gaz, d'atomes et de molécules se transpose en un monde virtuel de chiffres.



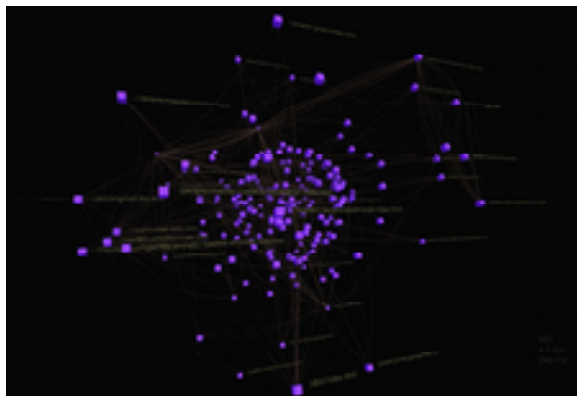
data data data, Stanza, 2012

Le Data-Art

« Ce n'est pas comme visualiser une particule au microscope. Il s'agit de prendre quelque chose que l'on ne peut pas voir et de l'amener sur le champ de la perception. » Aaron Koblin, directeur du Data Arts Team.

Une visualisation des flux

Spécialisé dans les méthodes de visualisation de grands ensembles de données provenant de sources ou de systèmes dynamiques, Ben Fry propose en 1997 un cadre conceptuel nouveau, lié au design de flux de données. L'approche développée s'inspire de propriétés organiques pour collecter et gérer des informations changeantes et complexes. En 2001, il présente *Valence*, un ensemble de logiciels qui explorent les structures et les relations à l'intérieur de très grands ensembles de données, un ensemble capable de trier qualitativement des données quantitatives, émises par des environnements interactifs.



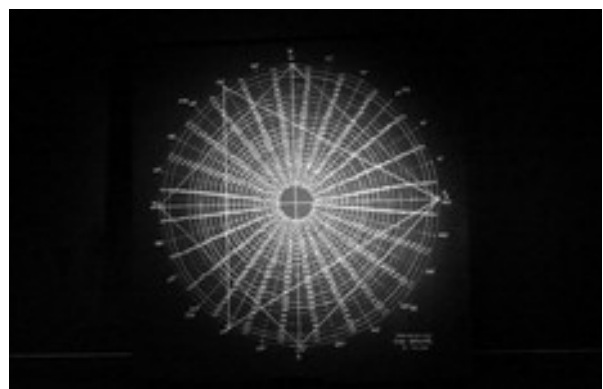
Valence, Ben Fry, 2001

Développé à l'origine comme une série d'expériences pour le projet *Celestial Mechanics* de Scott Hessels et Gabriel Dunne, *Flight Patterns* permet de visualiser les données de flux IP (adresses e-mail, peer-to-peer...) ainsi que des appels téléphoniques longue distance de suivi des avions transmises par la Federal Aviation Administration durant 24 heures. Aaron Koblin fait ainsi apparaître les relations spatiales entre infrastructure et communication, entre éveil, sommeil et déplacements humains dans chaque État.



Flight Patterns, Aaron Koblin, 2005-2009

Antidatamining est une série d'œuvres réalisées depuis 2006 par le collectif RYbN. Portant sur le domaine opaque de la finance et s'inspirant des mythologies et rituels constitutifs de l'économie contemporaine, cette série prend la forme d'herbiers algorithmiques, de performances sur les marchés, de reconstitutions et d'interprétations ou de représentations audiovisuelles dynamiques des données financières qui transitent sur internet. Cette approche extra-disciplinaire repose sur des enquêtes et des mises en situation.



ADMX, RYbN, 2016

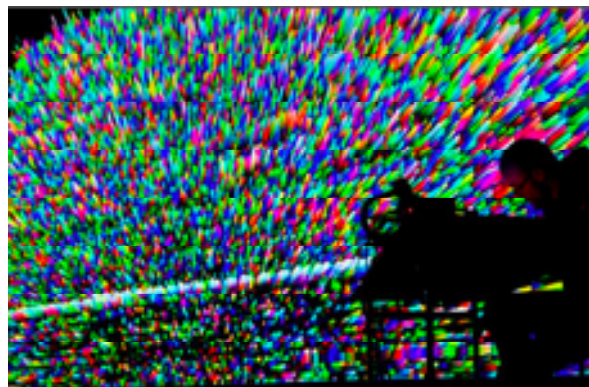
Les datas comme matière

Depuis 2004, Ryoji Ikeda a entamé un vaste projet intitulé *Datamatics* qui prend la forme de concerts ou d'installations audiovisuelles. Son ambition est de rendre perceptible le flot abstrait et invisible de données mathématiques qui informent le monde. *data.tron* est une installation audiovisuelle où chaque pixel de l'image est calculé très précisément selon des principes mathématiques, véritable association entre les mathématiques pures et le vaste univers de données présentes dans notre monde.



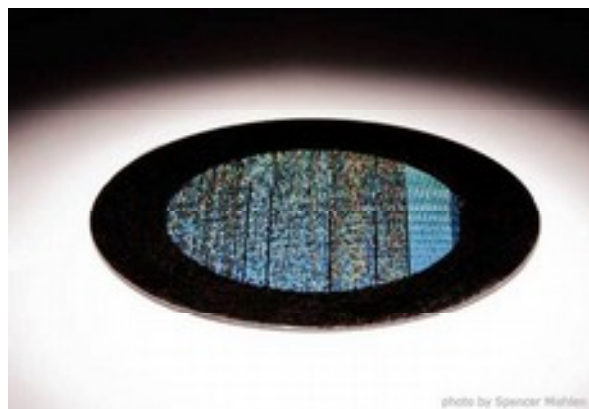
data.tron, Ryoji Ikeda, 2008

Nicolas Maigret expérimente la capacité des technologies contemporaines à auto-générer des langages sonores ou visuels spécifiques. Dans ses réalisations, les caractéristiques internes des médias sont rendues perceptibles à travers leurs erreurs, leurs dysfonctionnements, leurs états limites ou seuils de rupture. *System Introspection* peut être envisagé comme une observation de la machine par elle-même, proposant une expérience physique de la donnée numérique, des différents langages, formats et contenus. Ce travail repose sur une exploration concrète du code binaire contenu sur un disque dur et de ses qualités intrinsèques (structure, logique, rythme, redondance, compression) instantanément rendus par l'ordinateur sous forme de flux visuels et sonores.



System Introspection, Nicolas Maigret, extrait 2012

Le projet Rosetta est un projet mené par The Long Now Foundation sur l'archivage de données, questionnant le problème de l'obsolescence numérique. Il tente également par ce biais d'attirer l'attention sur la perte brutale et accélérée des langues du monde, jusqu'à 90 % de la diversité linguistique pourraient être perdus au cours du siècle prochain. Matérialisant les datas reçues sur ce sujet, le *Rosetta Disk* contient plus de 13 000 pages d'informations sur plus de 1 500 langues humaines. Tenant dans la paume de la main, chaque information s'y inscrit en image décodée car, plutôt que d'archiver ces données en code binaire, chacune d'entre elles peut être lue par l'œil humain au moyen d'un très fort grossissement optique.



The Rosetta Disk, collectif, 2011

Petit lexique

Partitionnement de données

Le partitionnement de données ou « segmentation » est une des méthodes statistiques d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité.

Réseaux de neurones

Un réseau de neurones artificiels est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement des neurones biologiques. Le neurone formel est conçu comme un automate doté d'une fonction de transfert qui transforme ses entrées en sortie selon des règles précises. Une fonction des réseaux de neurones formels, à l'instar du modèle vivant, est d'opérer rapidement des classifications et d'apprendre à les améliorer.

Algorithmes génétiques

Les algorithmes génétiques utilisent la notion de sélection naturelle et l'appliquent à une population de solutions potentielles au problème donné. La solution est approchée par « bonds » successifs, comme dans une procédure de séparation et évaluation, à ceci près que ce sont des formules qui sont recherchées et non plus directement des valeurs. Leur but est d'obtenir une solution approchée à un problème d'optimisation, lorsqu'il n'existe pas de méthode exacte (ou que la

solution est inconnue) pour le résoudre en un temps raisonnable.

Arbre de décision

Un arbre de décision est un outil d'aide à la décision qui représente la situation plus ou moins complexe à laquelle on doit faire face sous la forme graphique d'un arbre de façon à faire apparaître à l'extrémité de chaque branche les différents résultats possibles en fonction des décisions prises à chaque étape.

À découvrir

Autour du Dataviz

Tableau périodique des méthodes de visualisation

http://www.visual-literacy.org/periodic_table/periodic_table.html

Data Viz, Open Data, Big Data

<http://123opendata.com/blog/data-viz-open-data-big-data/>

Quelques artistes citées

Electronic Shadows

<http://www.electronicshadow.com/>

Julien Prévieux

<http://www.previeux.net/>

Tim Schwartz

<http://www.timschwartz.org/>

Stanza

<http://www.stanza.co.uk/>

Ben Fry

<http://benfry.com/>

Aaron Koblin

<http://www.aaronkoblin.com/>

Ryoji Ikeda

<http://www.ryojiikeda.com/>

Nicolas Maigret

<http://peripheriques.free.fr/blog/>

Quelques œuvres en ligne

We feel fine

<http://www.wefeelfine.org/>

Antidatamining

<http://www.antidatamining.net/>

Quelques outils de Dataviz

Timetoast

<http://www.timetoast.com/>

Tweetography

<http://www.tweetography.com/>

DrasticTreemap

<http://www.drasticdata.nl/DrasticTreemapGApi/>

Weave

<http://oicweave.org/>

Lovely Charts

<http://www.lovelycharts.com/>

Orange

<http://orange.biolab.si/>

Text 2 Mindmap

<http://www.text2mindmap.com/>

Visalix

<http://visalix.xrce.xerox.com/>

Stat Silk

<http://www.statsilk.com/>

Open Spending

<http://openspending.org/>

Just Map It

<http://feeds.just-map-it.com/>

Tagxedo

<http://www.tagxedo.com/>

Infomous

<http://www.infomous.com>

Tiki-Toki

<http://www.tiki-toki.com/>

Sigmajs

<http://sigmaj.org/>

Sources : Data Flow / We make money not art / Wikipédia