



DATA SCIENTIST : LES RISQUES DU MÉTIER

M2 Data Science :
Health, Insurance and Finance

Romain BOUCHER

28 janvier 2022

université
PARIS-SACLAY

Objectif de la séance :



Pour y arriver :

Qui est le data scientist ?

- Ère de l'information, ère des compétences
- Le bluff technologique
- Au service de quoi ?

Bilan des avancées

- Qualité de vie
- Environnement
- Démocratie et libertés

Quelques réflexions plus générales

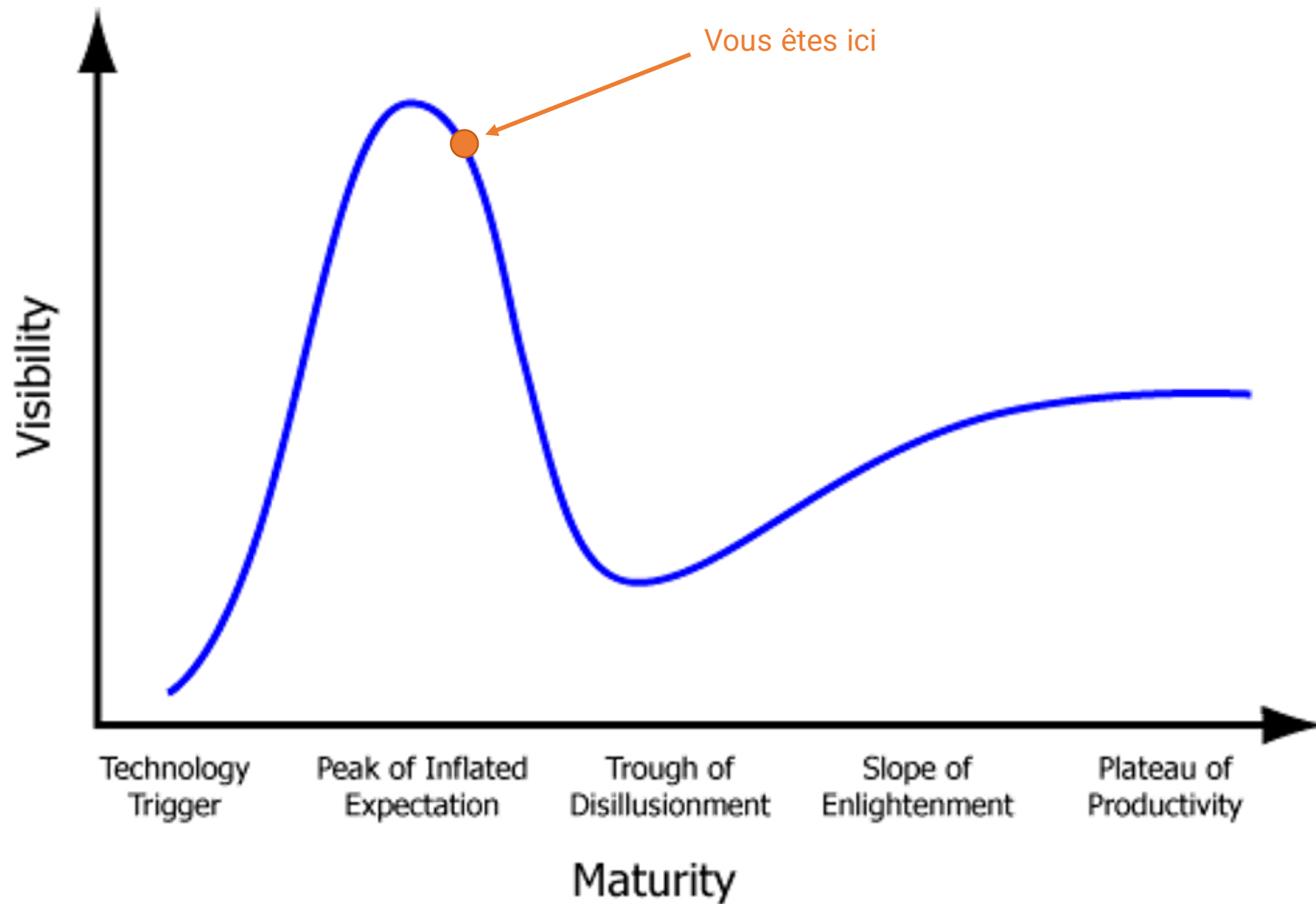
Qui vous parle ?

Ingénieur des Mines de Saint-Étienne (2017)

Consultant data scientist chez Sia Partners (2017-2020)

Cofondateur du collectif Vous N'êtes Pas Seuls

Chercheur en Histoire des Sciences, des Techniques et des Savoirs à l'EHESS



Connaissances

Mathématiques

+

Statistiques

+

Informatique

+

Domaine
d'expertise

=

Data Sciences

→

Computer Vision
Data Science
Machine Learning
NLP
Cloud
Deep Learning
Bot
IA
Big Data
Dématérialisation

Compétences

Acquisition



Traitement



Analyse



Modélisation



Visualisation



Le job le plus sexy du XXIème siècle

Data Scientist:

The Sexiest Job of the 21st Century

**Meet the people who
can coax treasure out of
messy, unstructured data.**

by Thomas H. Davenport
and D.J. Patil

W

hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't

seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review October 2012



Le job le plus sexy du XXIème siècle

CE SONT DES HUMAINS ÉLEVÉS EN LIBERTÉ



**ILS PEUVENT SORTIR POUR SE NOURRIR
PENDANT PRESQUE UNE HEURE PAR JOUR**



La dématérialisation

La dématérialisation de plus en plus à la mode

Depuis déjà plusieurs années, internet a pris de plus en plus d'ampleur, au point que de nombreuses entreprises misent maintenant sur la dématérialisation.

IBM dépasse les prévisions de chiffre d'affaires grâce à l'informatique dématérialisée et aux services de conseil

The Sustainable Impact Of A Paperless Office

La dématérialisation des documents au service des TPE et PME

Comment l'essor du numérique peut servir la transition écologique

Malgré une empreinte carbone grandissante, le développement des outils numériques est favorable à la baisse des émissions de gaz à effet de serre de la France, selon une étude de l'Institut Rexecode publiée ce jeudi.

La Cour des comptes épingle les insuffisances du Plan de transformation numérique de la justice



La dématérialisation



1,2 millions de kilomètres de câbles sous les océans



Fibre optique : 25 milliards d'euros d'installation



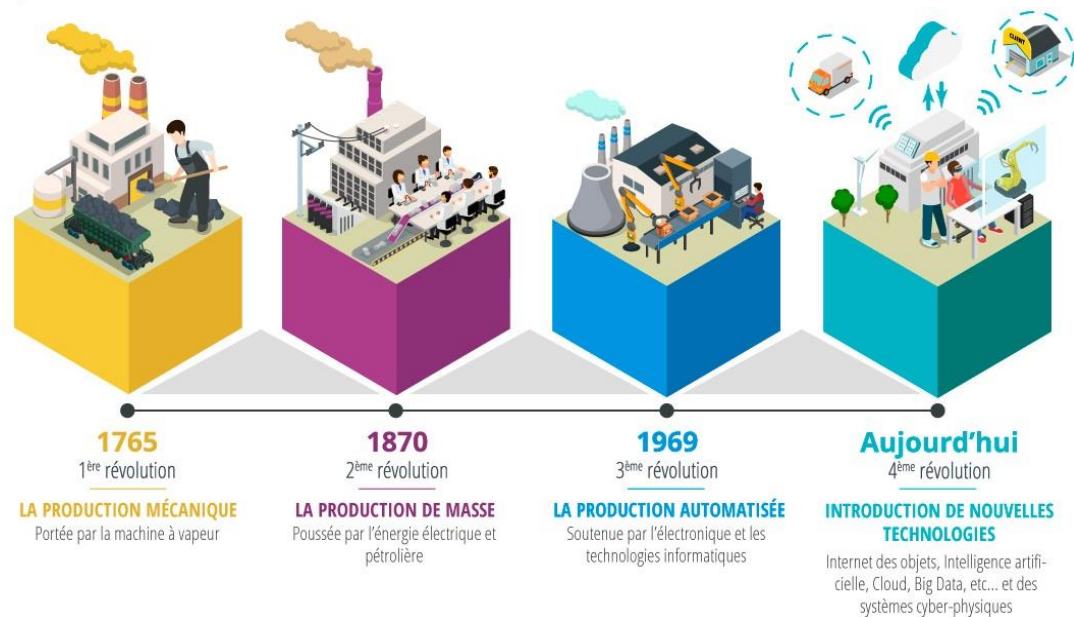
Des centaines de milliers de serveurs et datacenters



50 000 antennes 4G, 20 000 antennes 5G



L'industrie 4.0



L'industrie 4.0



Explosion de Lubrizol, Illinois, juin 2021



Explosion de Lubrizol, Rouen, sept. 2019



Marée noire au Pérou, janvier 2022



Marée noire à l'Ile Maurice, juillet 2020



Un écosystème complexe

Infrastructure



Python stack



Data Visualization



Big data processing



Data storage / formats / querying



Un écosystème complexe

Métier

Data scientist

Modèles, recherche fondamentale

Données

Système d'exploitation
et navigateur web

Plateformes
cloud

Outil de travail :
ordinateur

Infrastructures physiques de stockage,
de calcul et de transport d'information



Secteur	Montant levé (en M€)	Part du montant global (en %)	Nombre d'opé- rations
Marketing, ad- tech	728,4	16,0	84
Biotech	515,17	11,3	39
e-RH	466,81	10,3	48
Retail	447,82	9,8	64
Fintech	411,20	9,0	68
Mobilités	190,10	4,2	31
Immobilier	143,25	3,1	38
IoT	116,65	2,6	29
Cloud	56,48	1,2	6
Big Data	33,00	0,7	3
Autres	1342,19	31,7	252
Total	4451,07	100,0	662



- Ultra personnalisation des biens et services = "KYC"
- Automatisation ou accélération de processus pour "s'affranchir des tâches à peu de valeur ajoutée"
- Augmenter les marges de profits
- Se faire bien voir de sa hiérarchie



69% des offres d'emplois de data scientist sont consacrés de près ou de loin au marketing ou à la personnalisation de contenu



RATIONALISER



Plan de contrôle de toxicité des eaux

Objectif : 2000 agents en moins pour contrôler le taux de nitrate des cours d'eau



Un robot pour répondre aux appels

Objectif : ne plus avoir besoin d'embaucher des CDD lors des pics d'appels



Algorithmes de détection de fraude

Objectif : se focaliser sur la masse et se passer des agents humains

LÉGITIMER



Aide au diagnostic

Algorithmes de détection de forme pour les maladies de l'œil, cancer du sein..



Synthèse de dossier patient

Algorithme de synthèse des dossiers patients pour les réunions de médecins



Analyse automatique des REX

Automatiser l'analyse des retours d'expérience pour lever automatiquement l'alerte sur l'exposition aux rayonnements

CONTRÔLER



Ciblage des contrôles d'hygiène

Algorithme de détection de risque sanitaire dans les établissements de restauration, à partir d'un historique d'enquête et des commentaires en ligne.



Détection de cabanisation

Utilisation d'images aériennes pour détecter des bâtis non déclarés ou non conformes au plan local d'urbanisme. Expérimentation de la DDTM de l'Hérault.

Un bilan du progrès



Qualité de vie

Santé/Médecine

Des progrès en :

- Génomique
- Diagnostic sur image (cancers du sein, maladies de l'œil)
- Prévention, suivi des maladies chroniques
- Pharmacologie (mise au point de médicaments, vaccins)

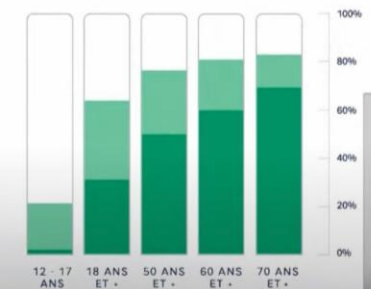
Gestion du covid, télémedecine, santé connectée ?



Le vaccin, notre bouclier collectif

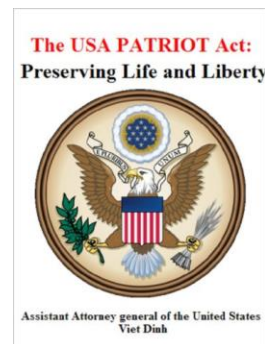
TAUX DE VACCINATION PAR TRANCHE D'ÂGE
AU 7 JUILLET 2021

● UNE SEULE DOSE ● COMPLÈTEMENT VACCINÉS



Qualité de vie

Santé/Médecine



Hébergement de données de santé du Health Data Hub chez Microsoft : saisine du Ministre de la Santé pour potentiel manquement aux règles de la commande publique **Mars 2020**

HEALTH DATA HUB: LA CNIL DEMANDE L'ARRÊT DE L'HÉBERGEMENT DES DONNÉES DE SANTÉ PAR MICROSOFT **Octobre 2020**

Health Data Hub: Véran s'engage à retirer l'hébergement à Microsoft d'ici «deux ans» **Novembre 2020**

VACCINATION: DOCTOLIB PASSE LE CAP DES 5 MILLIONS DE RENDEZ-VOUS **5 mars 2021**

Des syndicats de médecins contestent devant le Conseil d'Etat le partenariat entre l'Etat et Doctolib qui héberge les données chez Amazon **8 mars 2021**

Doctolib : le chiffrage des données incomplet ? **8 mars 2021**

Campagne de vaccination : le Conseil d'Etat appelé à se prononcer sur le partenariat avec Doctolib **8 mars 2021**

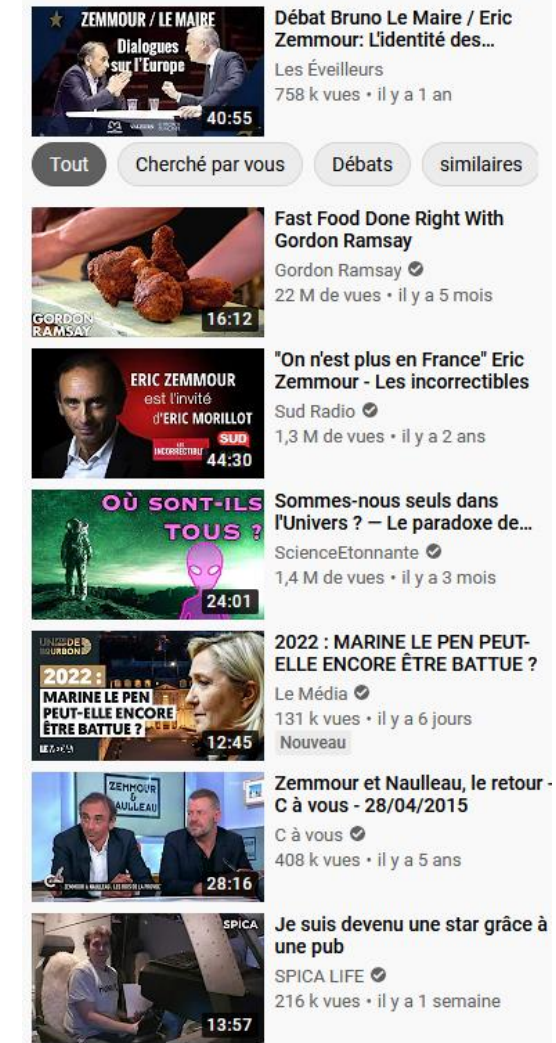
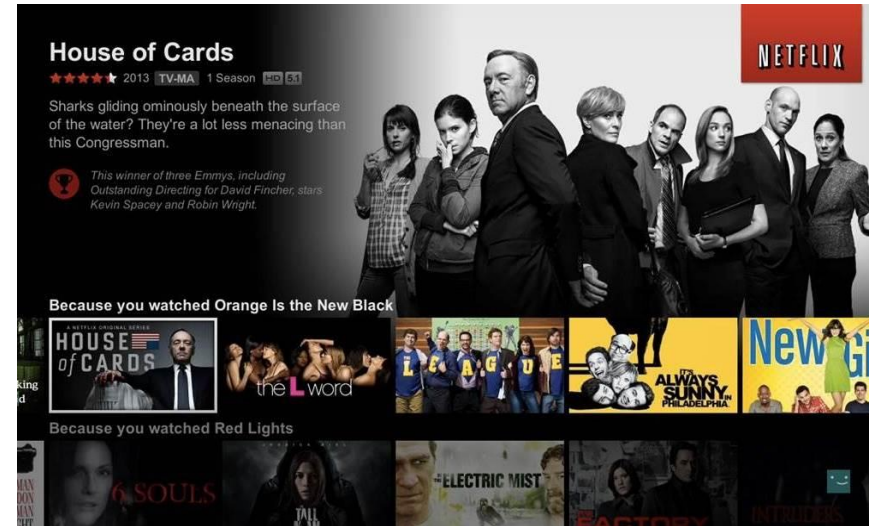
Qualité de vie

Algorithmes de recommandation

Normalisation de la culture, des mœurs, des idées et des imaginaires

Economie de l'attention :

- 70% du temps passé sur YouTube est du aux algorithmes de recommandations
- 80% du contenu regardé sur Netflix est déterminé par les algorithmes



Qualité de vie

Des modes de plus en plus incitatifs



Statistiques
Informatique décisionnelle
Aide à la décision



Apprentissage automatique
Suggestions
Recommandations



Apprentissage profond
Incitation



"Intelligence Artificielle"
Injonction
Coercition

PROGRÈS

La complexification technologique de la société s'accompagne de :

- Structures de pouvoir toujours plus asymétriques
- Perte d'autonomie
- Désengagement politique

Environnement

Un levier pour lutter contre le réchauffement climatique ?

- Améliorer les prévisions de consommation d'énergie, réduire les obstacles à l'adoption des véhicules électriques
- Découvrir de nouveaux matériaux
- Optimiser les chaînes d'approvisionnement
- Aider à rendre les bâtiments plus efficaces
- Rendre possible l'agriculture de précision à l'échelle
- Améliorer le suivi de la déforestation
- Pousser les consommateurs à changer notre façon de consommer

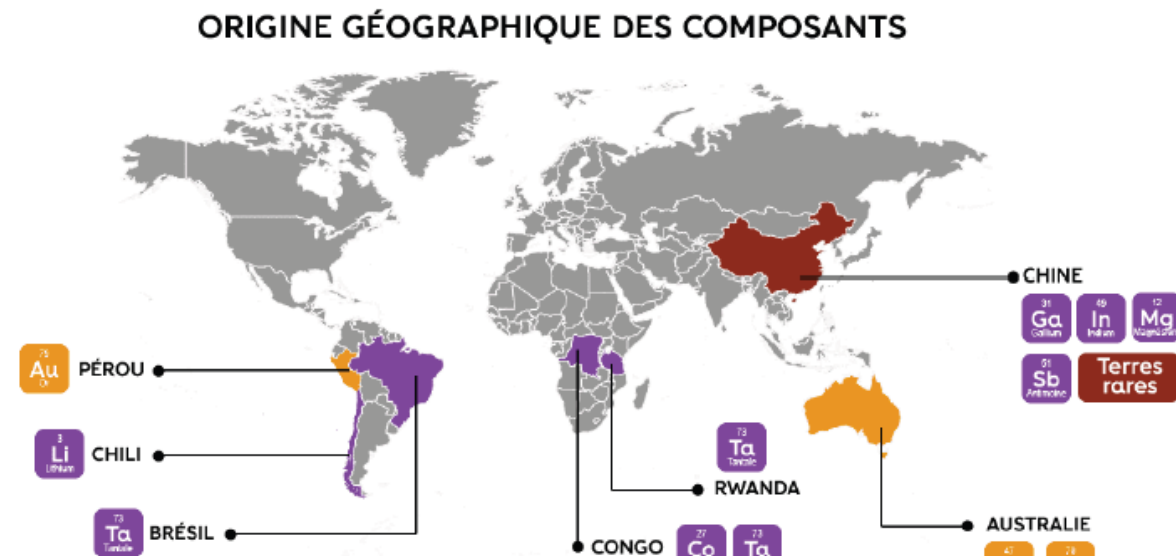
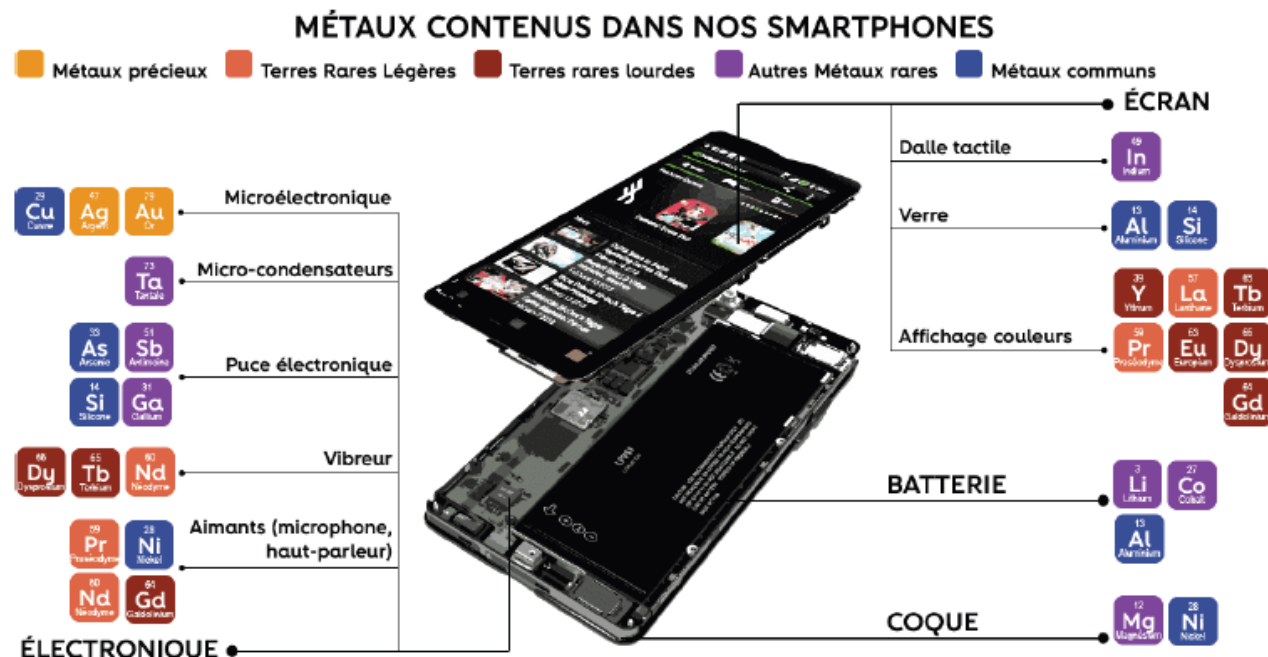
Environnement

Le mythe de la dématérialisation

- Sur les cinq dernières années, environ 1,4 milliards de smartphones ont été vendus dans le monde chaque année et un seul smartphone requiert environ 60 éléments différents (principalement des métaux rares ou critiques)
- La fabrication d'un ordinateur de 2kg mobilise environ 800kg de matières premières



- Après le bâtiment, l'industrie technologique est le secteur le plus extracteur de sable
- Conséquences de cette surexploitation : recul des terres, disparitions de certaines îles, instabilité des infrastructures, danger pour l'agriculture...



[Infographie BFM, 2018]

Environnement

Le mythe de la dématérialisation

- Nous avons multiplié les usages de ces métaux dans deux domaines qui sont les piliers essentiels de la transition énergétique:
- Les technologies supposément « vertes »;
- Le numérique.
- Car c'est de la convergence des green tech et de l'informatique que devrait naître un monde meilleur...
- Les premières (éoliennes, panneaux solaires, véhicules électriques), grâce aux métaux rares dont elle sont truffées, produisent une énergie décarbonée qui va transiter par des réseaux d'électricité dits « ultraperformants » qui permettent des économies d'énergies.
- Or ceux-ci sont pilotés par des technologies numériques, elles aussi farcies de tels métaux.

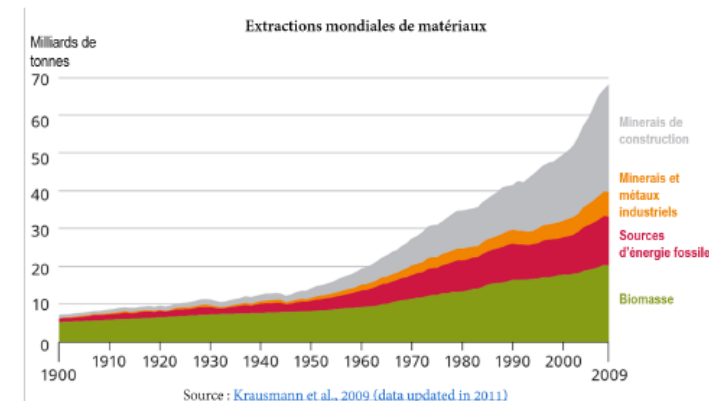
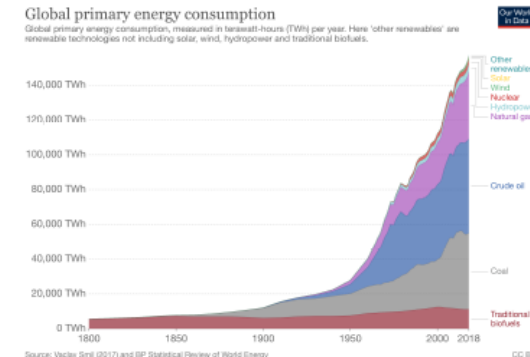
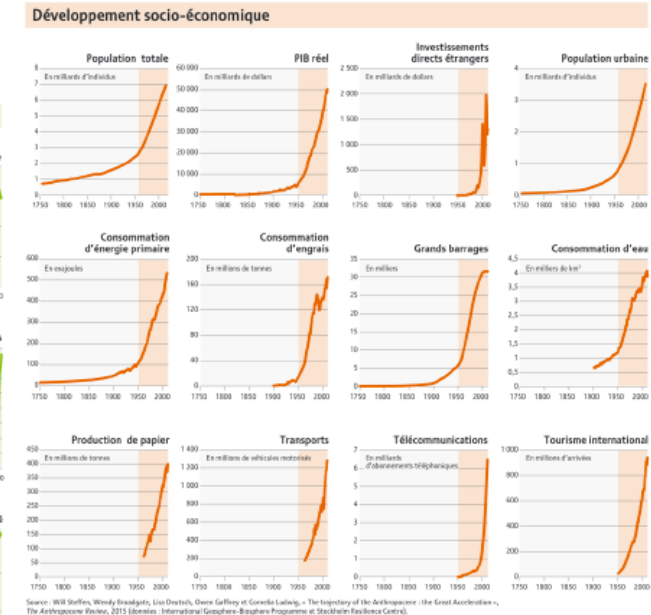
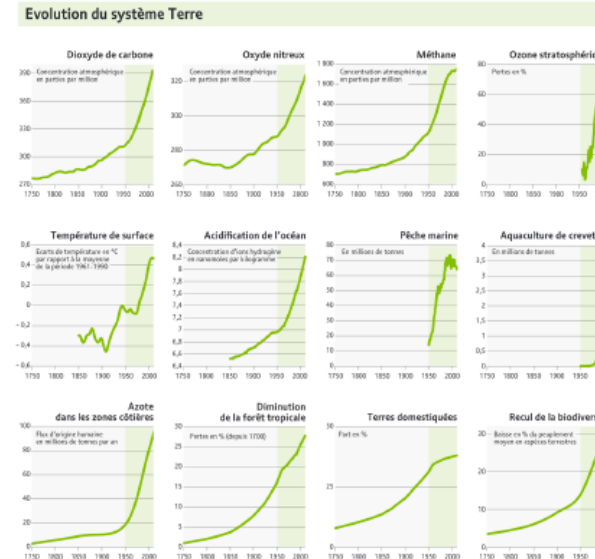
Ressource	Utilisations
Antimoine	Retardants du feu (additifs dans les plastiques), catalyse du polyéthylène
Baryte	Boues de forage pétroliers et gaziers, industrie du verre, radioprotection, santé, métallurgie, pyrotechnie
Béryllium	Télécoms et électronique, industrie aéronautique, nucléaire civil et militaire
Cobalt	Portables, ordinateurs, véhicules hybrides, aimants
Germanium	Photovoltaïque, fibres optiques, catalyse, optique infrarouge
Indium	Puces électroniques, écrans LCD
Graphite naturel	Véhicules électriques, aéronautique, industrie nucléaire
Silicium métal	Circuits intégrés, panneaux photovoltaïques, isolateurs électriques
Terres rares (voir annexe pour détails)	Aimants permanents, voiture électrique, éoliennes, TGV, scanners médicaux, lasers, transmission de données par fibre optique ...

Principales utilisations industrielles de quelques minerais rares
[Source du tableau: Guillaume Pitron]

Environnement

La grande accélération

- Représente un tableau de bord de l'Anthropocène avec l'évolution de 24 paramètres du système Terre depuis 1750.
- Pour les 9 les plus significatifs, des scientifiques se sont intéressés aux points de basculement concernant la biodiversité, la pollution de l'air et de l'atmosphère, la perturbation des cycles biogéochimiques ou l'anthropisation des terres.
- Une limite à ne pas franchir pour ces 9 grands paramètres a été évaluée. Mais pour 4 d'entre eux, le seuil de danger de basculement brutal du système Terre vers des états catastrophiques est d'ores et déjà approchée ou dépassée : cycles de l'azote, émissions de GES, extinction de la biodiversité, cycle du phosphore



Environnement

Consommation énergétique et empreinte carbone du numérique

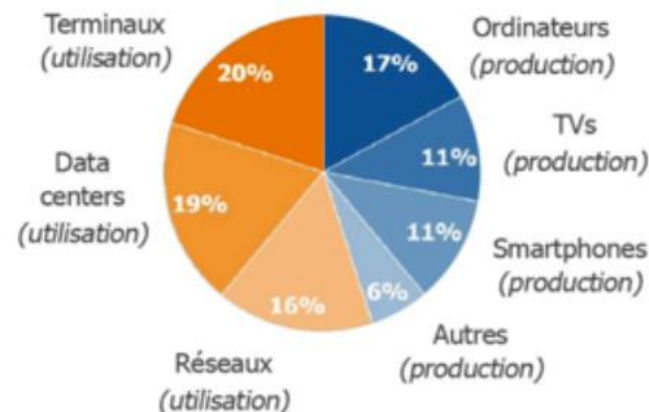
- La phase de production des équipements occupe une part très significative, environ 45% en 2020, dans l'empreinte énergétique totale du Numérique, ainsi que dans les émissions de GES qui en découlent.
- Un utilisateur de smartphone (s'il garde son appareil 2 ans) verra ainsi la consommation énergétique totale induite au cours du cycle de vie de cet équipement se réaliser à plus de 90% avant même son achat*. Ce poids de la phase de production dans l'impact énergétique est de l'ordre de 60% pour une télévision connectée et >80% pour un ordinateur portable.
- Une bonne partie des enjeux environnementaux du Numérique n'est donc pas liée à son usage, mais en grande partie au volume de matériel produit, à son processus de production, et à sa durée de vie.

Common carbon footprint benchmarks

in lbs of CO2 equivalent

Roundtrip flight b/w NY and SF (1 passenger)	1,984
Human life (avg. 1 year)	11,023
American life (avg. 1 year)	36,156
US car including fuel (avg. 1 lifetime)	126,000
Transformer (213M parameters) w/ neural architecture search	626,155

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper



Distribution de la consommation énergétique du Numérique par poste (2017)
[Source : [Lean ICT Materials]]

* Nous parlons bien ici de la consommation électrique propre du smartphone : si l'on prend en compte également le surcroît de consommation électrique du réseau occasionné par l'utilisation du smartphone, la proportion est plutôt de l'ordre de 50%. Voir (Ercan, 2013).

Démocratie et libertés

Surveillance



Les géants du numérique ont désormais la mainmise sur les services informatiques, les infrastructures et les données des utilisateurs.

Un pouvoir qui leur donne un poids considérable dans la politique des États occidentaux, dont les gouvernements n'hésitent pas à collaborer avec ces multinationales pour se maintenir.

<https://jenairienacacher.fr/>



Le crédit social chinois, paroxysme de l'hyper-rationalisation rendue possible par le Big Data, témoigne des effets de centralisation du pouvoir et de son incompatibilité avec la démocratie.

Voir sur ARTE : *Tous surveillés - 7 milliards de suspects*

Démocratie et libertés

Surveillance

Loi Sécurité Globale validée au Sénat le 3 mars 2021:

- Autorisation et banalisation de la surveillance par drone
- Reconnaissance faciale en direct (caméras drone et piéton)
- Mutualisation des moyens de surveillance publics/privés sur tout le territoire

Décret du 10 mars 2021

Autorisation au recours à la vidéo intelligente pour mesurer le taux de port de masque dans les transports

Décrets PASP du 8 novembre 2020

- Fichage massif de militants et d'opposants politiques (et de leur entourage)
- Surveillance automatisée sur les réseaux sociaux
- Recensement des opinions politiques et des données de santé
- Fichage des victimes
- Elargissement de la reconnaissance faciale

Valenciennes - 2019

Installation de 240 caméras offertes par Huawei et équipées d'un système de reconnaissance faciale

Paris - 2020

Détection du port du masque dans le RER C

Metz - 2016

Installation d'un dispositif de reconnaissance faciale dans un stade de foot

Start-up Nation

Saint-Etienne - 2019

Un projet de surveillance sonore est retoqué par la CNIL

Nice - 2019

- Système de reconnaissance faciale israélien testé sur la foule du carnaval
- Détection des émotions des passagers de tramways
- Portiques de reco faciale dans deux lycées

Cannes - 2020

Détection du port du masque

Hérault - 2019

Détection de bâtis non déclarés par images satellite

Marseille - 2019

Caméras détectant des comportements suspects sur la voie publique

Démocratie et libertés

Les données du Grand Débat

Une opportunité unique d'analyser longuement de grands volumes d'expression démocratiques, en mettant à contribution notamment des ingénieurs (algorithmes de traitement du langage naturel), des linguistes, des sociologues, des historiens, etc. Mais :

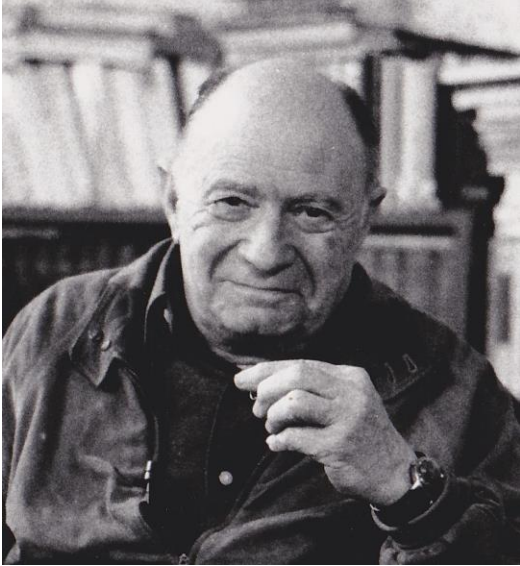
- L'analyse des données du Grand Débat est confiée, **sans appel d'offre**, à un consortium de 3 entreprises dirigé par le cabinet Roland Berger.
- Une synthèse est rédigée et publiée, sans que les revendications au cœur du mouvement des Gilets Jaunes soient entendues
- Les cahiers de doléances recueillis par l'Etat ne **sont jamais publiés**, sous prétexte d'un trop grand volume de données



Synthèses des contributions en ligne

Questionnaires rapides et propositions partagées sur la plateforme





Jacques Ellul (1912-1994)

"La technique n'est ni bonne, ni mauvaise ; elle est ambivalente. On ne peut dissocier ses effets positifs des effets négatifs, ses effets prévus de ses effets imprévus. Mais surtout, la technique n'est pas neutre : ce n'est pas l'usage qui va être fait de telle ou telle technique qui en définit la caractéristique pour la communauté humaine. [...]"

Une technique ne peut être utilisée que selon ses propres règles. En retour, elle exige une indispensable conformité du comportement humain à son mode d'emploi. [...]"

Une innovation technique provoque en général des effets indésirables et malencontreux. La seule voie pour résoudre le problème sera, dans le registre de la technique, la recherche d'une autre innovation technique."



Dominique Pestre (1950-)

"La science moderne n'est jamais « naturellement » attentive aux conséquences qui surgissent de la boîte de Pandore qu'elle ré-ouvre constamment. En d'autres termes, qu'il faut penser les sciences et techniques comme favorisant certaines manières d'être au monde et de se développer, comme portant en elles-mêmes une politique."



"Il est difficile de faire comprendre quelque chose à quelqu'un lorsque son salaire dépend du fait qu'il ne la comprenne pas."

Upton Sinclair, écrivain socialiste
états-unien (1878-1968)

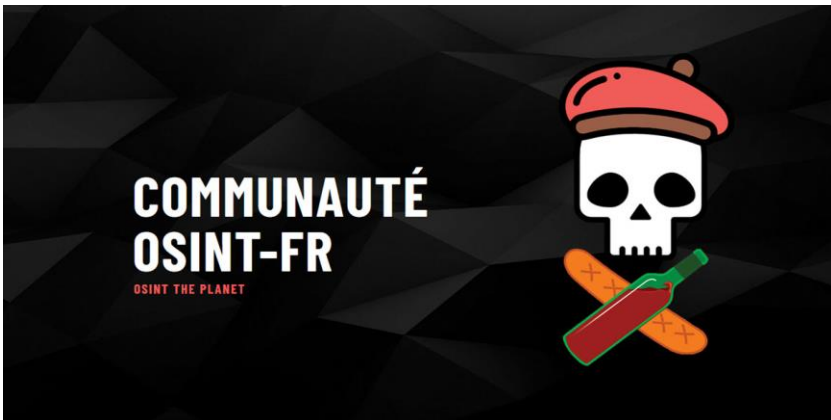
Mes options



Ne pas prendre en compte ces élucubrations et devenir trader en crypto.



Trouver une activité qui optimise ma courbe travail/épanouissement.



Utiliser mes connaissances et la technologie uniquement si cela sert le bien commun.

Documentaires Arte sur le contrôle social

- George Orwell, Aldous Huxley : "1984" ou "Le meilleur des mondes" ?
- Tous surveillés, 7 milliards de suspects

Se tenir informé

- <https://www.laquadrature.net/>
- <https://jenairienacacher.fr/>

Quelques bons livres

- *La Zone du Dehors*, Alain Damasio
- *Le Procès*, Franz Kafka
- *1984*, George Orwell
- *Fondation*, Isaac Asimov

Penseurs de la technique

- Ivan Illich, *La convivialité*
- Jacques Ellul, *Le bluff technologique*
- Hannah Arendt, *Condition de l'homme moderne*
- Günther Anders, *L'obsolescence de l'homme*
- Gilles Deleuze, *Post-scriptum sur les sociétés de contrôle*
- Theodore Kaczynski, *Révolution Anti-tech*

Réflexions contemporaines

- Les ouvrages du collectif Pièces et Mains d'Œuvre
- Les dernières publications d'Éric Sadin

Sur la science et la recherche

- Guillaume Carnino, *L'invention de la science*
- Dominique Pestre, Conférences sur Youtube
- Conférence d'Alexandre Grothendieck au CERN en 1972 (<https://www.youtube.com/watch?v=ZW9JpZXwGXc>)