Statistique & Numérique en entreprise

À la découverte des métiers de la donnée

Ketsia Guichard

ketsia.guichard@univ-rennes.fr

Ingénieure en statistiques - Ensai 2015’

Data Scientist (Meetic, Vuitton, Leboncoin)

Enseignante en statistiques

Thèse en cours en statistiques/économie/mobilités (IRMAR - CREM)

Explosion des données disponibles

Source : https://www.domo.com/learn/infographic/infographic-data-never-sleeps

Le buzz autour des métiers de la donnée

Analytics And Data Science

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 DJ Patil

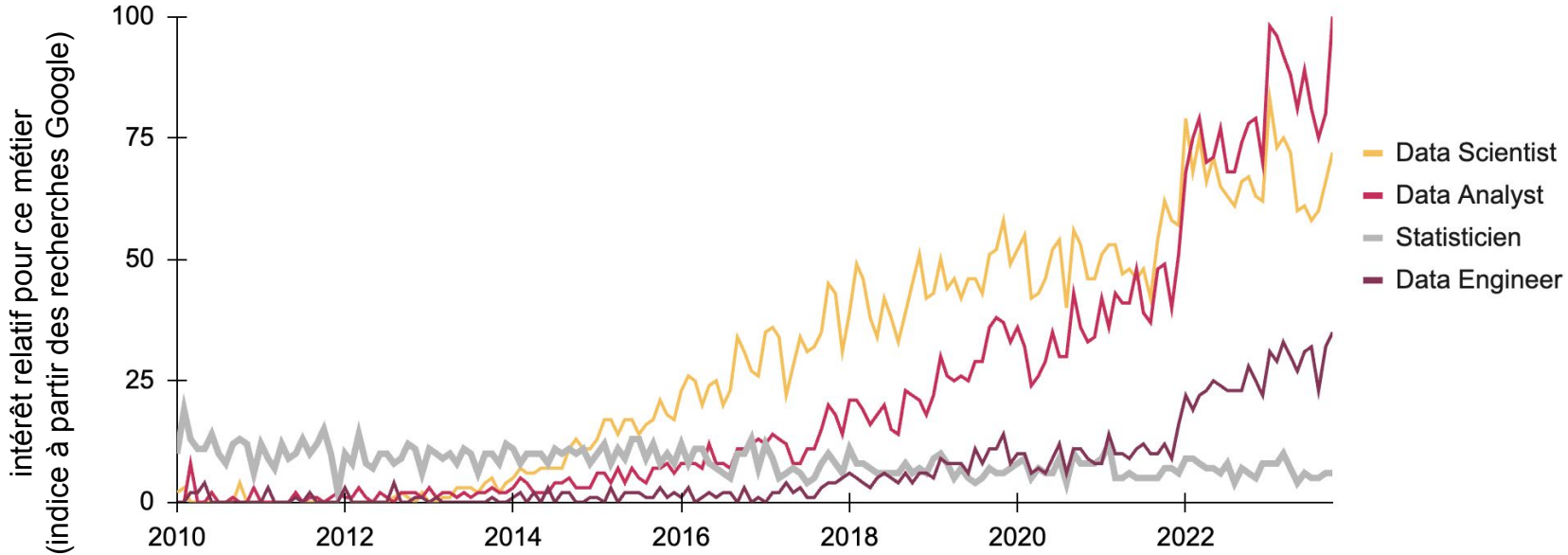
From the Magazine (October 2012)

Source : https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Le buzz autour des métiers de la donnée

Explosion des métiers "Data"

Evolution des recherches Google liées aux métiers de la donnée (2010-2021)



Source : Google Trends, France, https://t.ly/cb-U

Le buzz autour des métiers de la donnée

Data ninja

Data Wizard

Global Data Steward

Objectifs

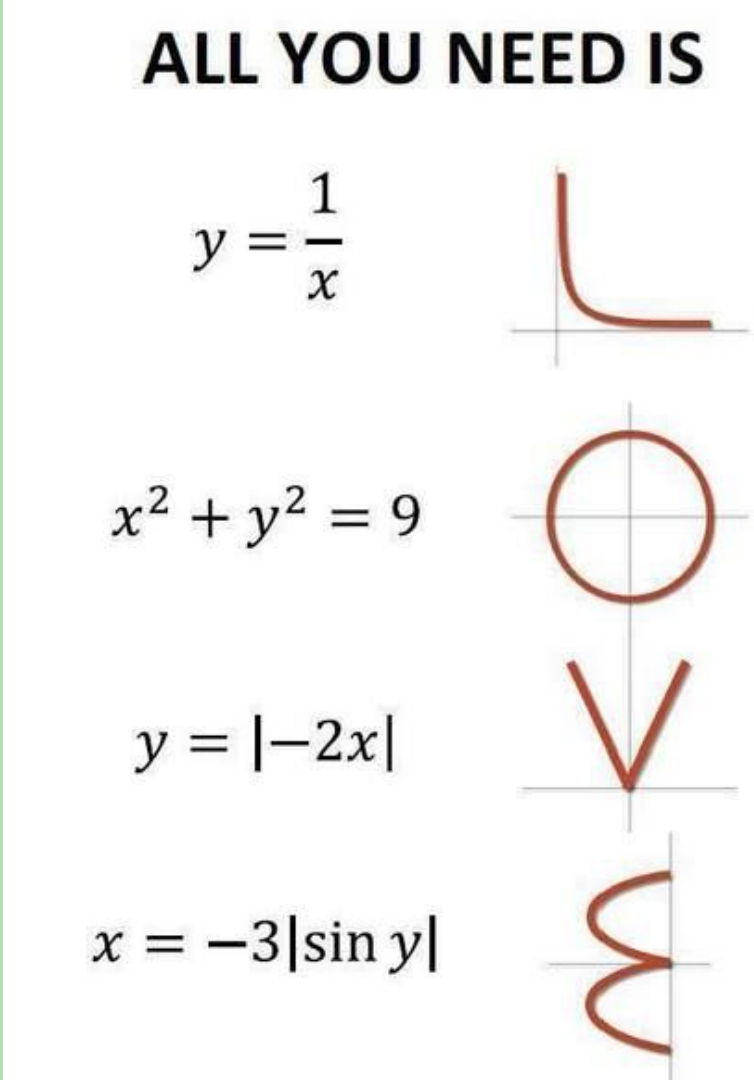
- **clarifier** l'utilisation des données et des statistiques ;
  - Comment nos données personnelles sont-elles récoltées ?
  - Quelles méthodes statistiques sont employées ?
  - Quels sont les différents profils qui collaborent sur ce sujet ?

- faire connaître et **rendre accessible** les métiers liés à la donnée.

Au programme

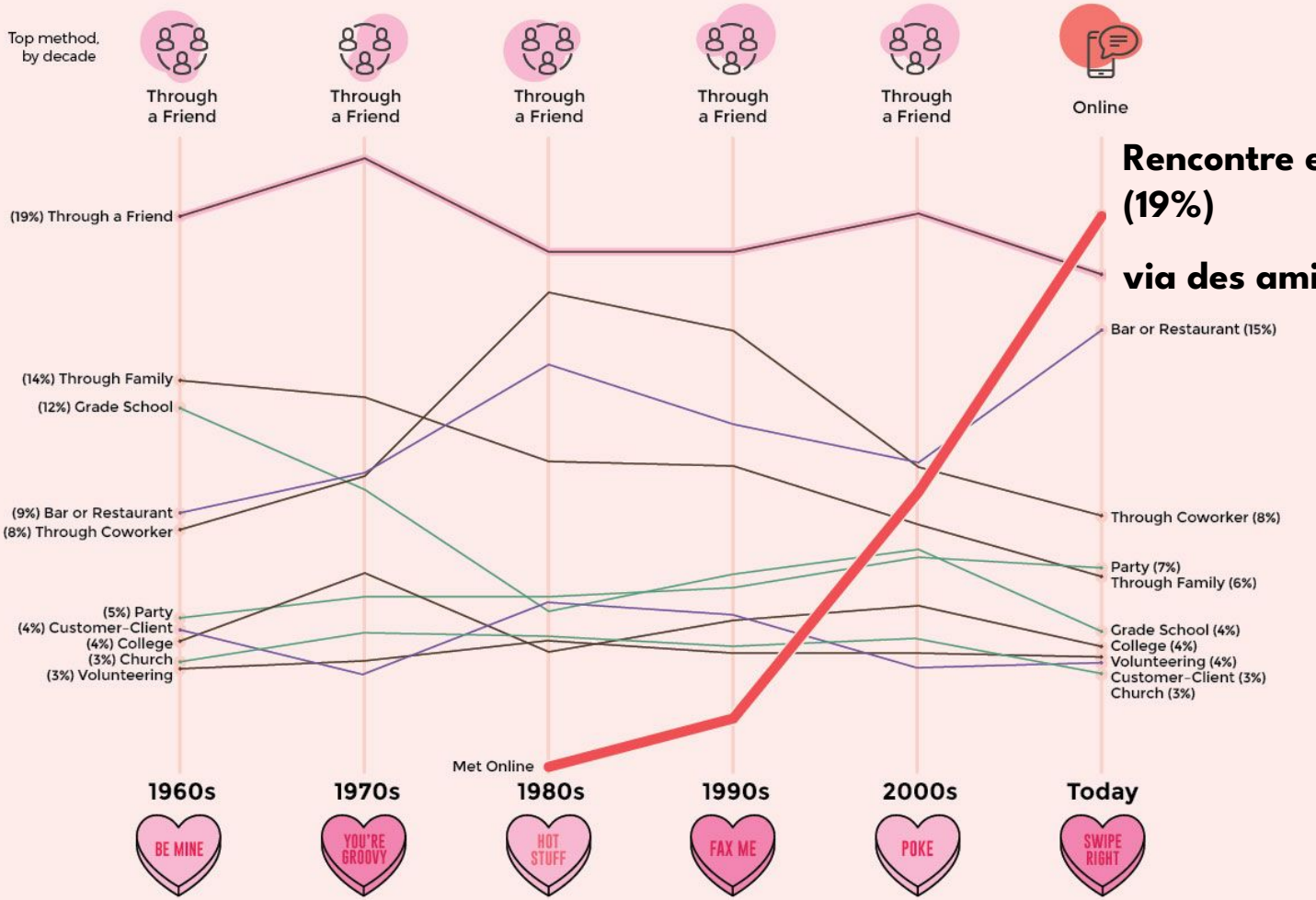
1. **Cas pratique** d'utilisation de nos données et des statistiques
2. Récapitulatif des **métiers de la donnée**
3. Focus sur la **place des femmes** dans ces métiers

Cas pratique :
les mathématiques de l'amour



Source : https://www.lenfisherscience.com/

Les sites de rencontres



Rencontre en ligne (19%)

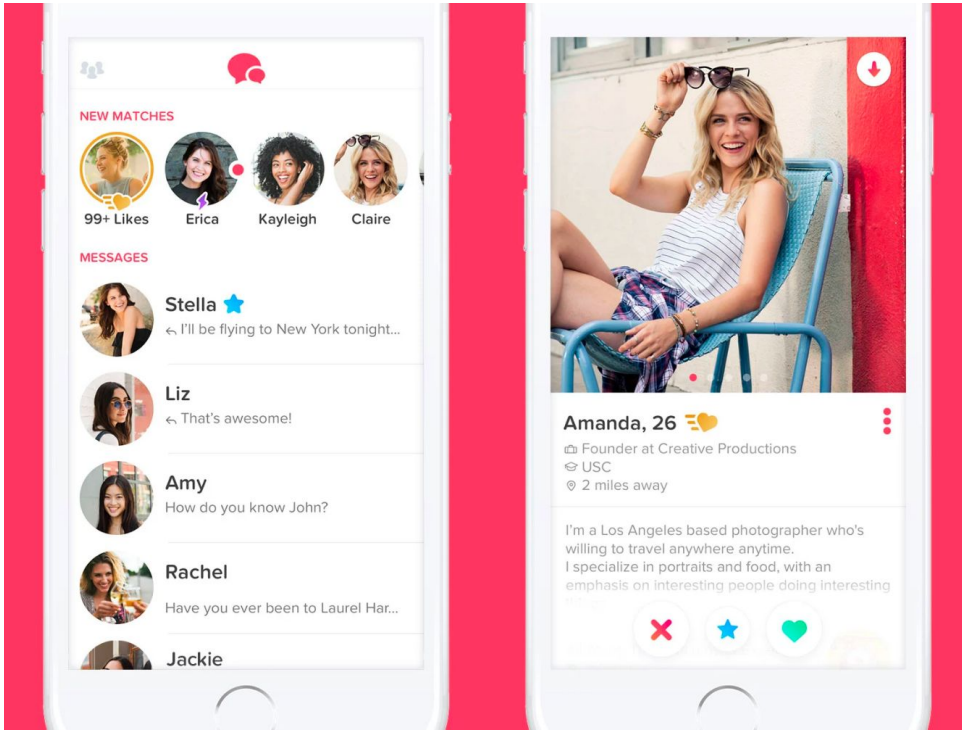
via des amis (17%)

Source : https://www.visualcapitalist.com/online-dating-big-business/

Rosenfeld, Michael J., Reuben J. Thomas, and Sonia Hausen (2019). How Couples Meet and Stay Together 2017

visualcapitalist.com

Aperçu d'un site de rencontre



Exemple de données

Profil des utilisateurs

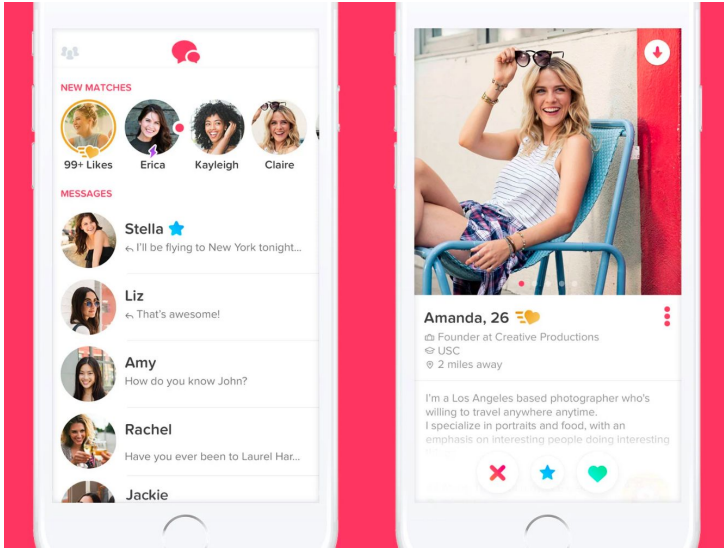
User	Genre	Âge	Cherche	Localisation	Photos	Distance max (km)	Age Min	Age Max
Ketsia	Femme	30	Homme Femme	Bruz Rennes	4	21	18	32

+

ancienneté, fréquence des connexions, type de téléphone, mode d'acquisition, temps passé sur l'application, sommes dépensées, abonnement…

Exemple de données

Actions et échanges entre utilisateurs



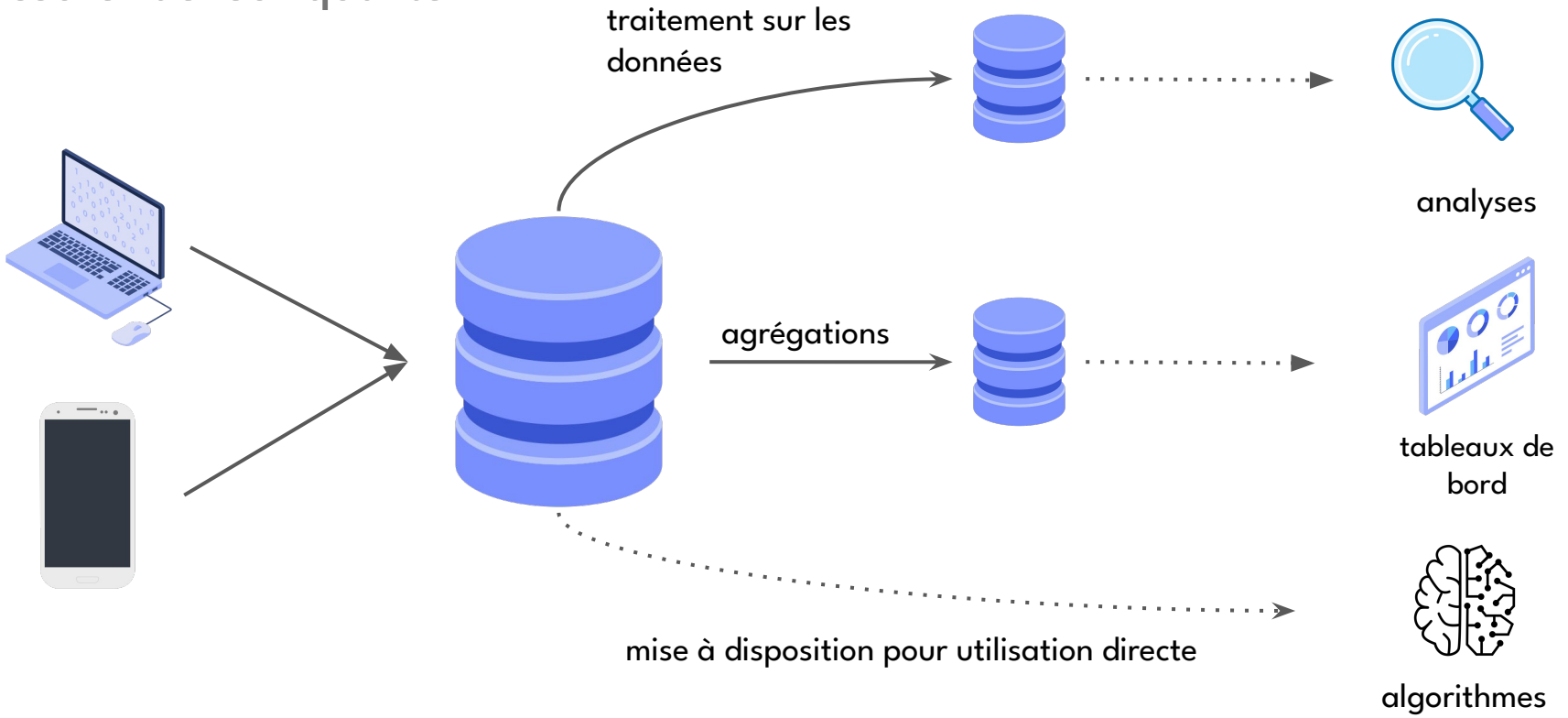
User	Profil	Choix	Match	Conversation	A initié le dialogue	Nombre de messages
Ketsia	David	Nope	Non	-	-	-
Ketsia	Marc	Nope	Non	-	-	-
Ketsia	Amanda	Like	Oui	Non	-	-
Ketsia	Stella	Like	Oui	Oui	Non	2
Ketsia	Jacky	Like	Oui	Oui	Oui	30

+

attractivité du profil (taux de like), taux de réponse, temps de réponse, longueur des messages envoyés…

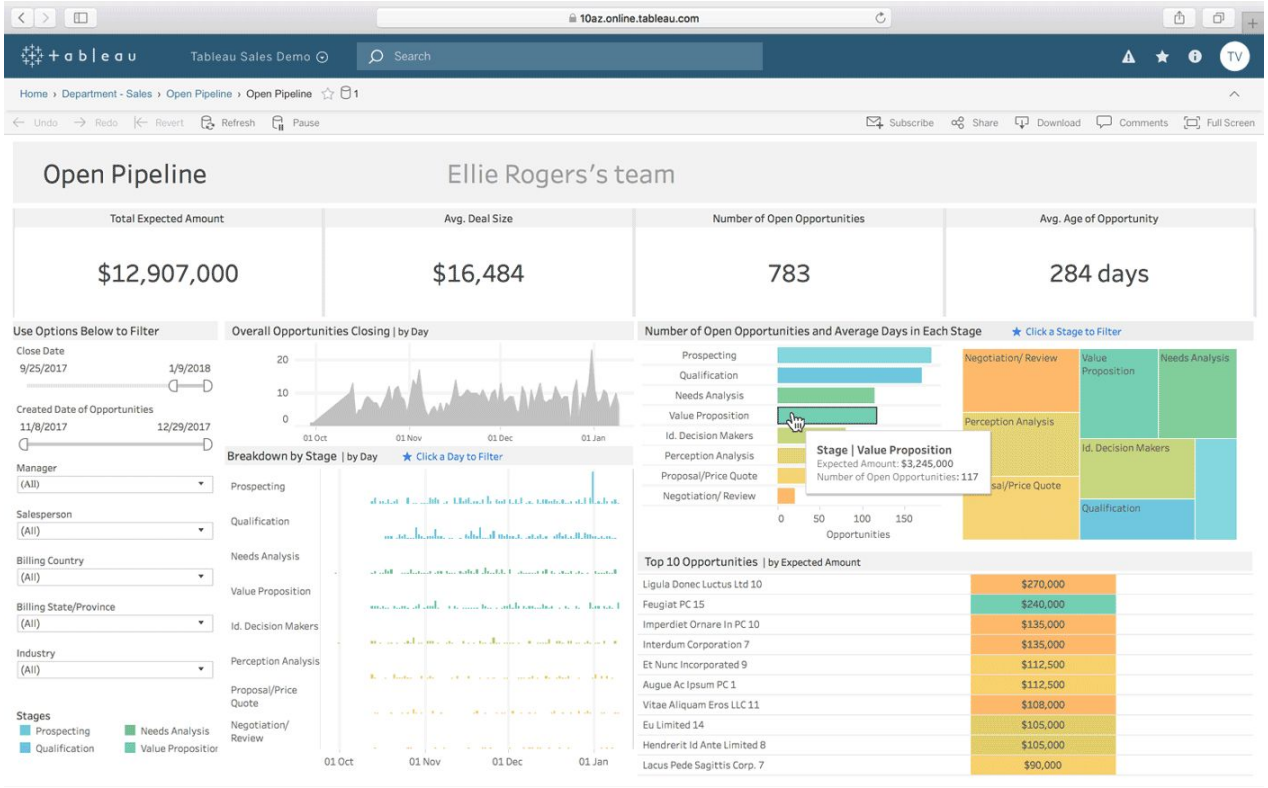
Stocker les données

Et s'assurer de leur qualité



Tirer de la valeur des données

Suivi business



Source : https://www.tableau.com/fr-fr/products/viewer

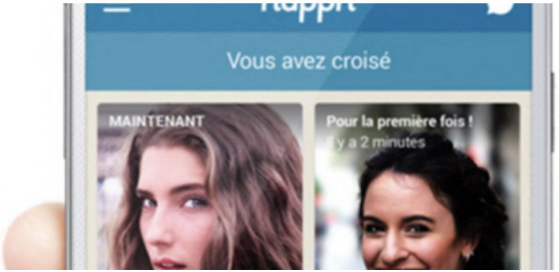
Tirer de la valeur des données

Analyses et collaborations presse

Actu › Occitanie › Hérault › Montpellier

Coups de foudre virtuels : quel est le classement de Montpellier pour "crusher" ?

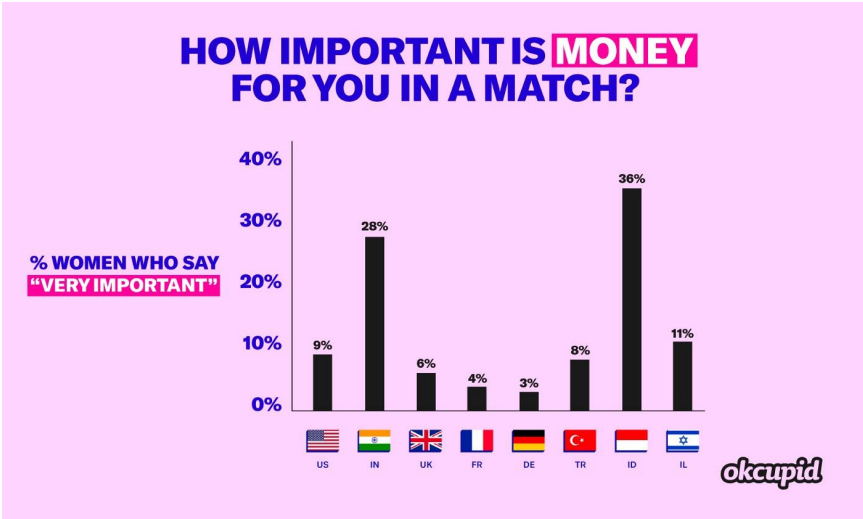
L'appli de rencontres Happn classe Montpellier dans le Top 10 des villes où les jeunes "crushent" le plus. Mais quelle place occupe t-elle ? Et au fait, c'est quoi un "crush" ?!



Actu › Occitanie › Haute-Garonne › Toulouse

Sur cette application, Toulouse est la deuxième ville la plus active de France, juste derrière Paris

Meetic, l'application de rencontres, a dévoilé le classement des villes françaises où les célibataires sont les plus actifs. Et Toulouse se classe en deuxième position !



Source : https://theblog.okcupid.com/tagged/data

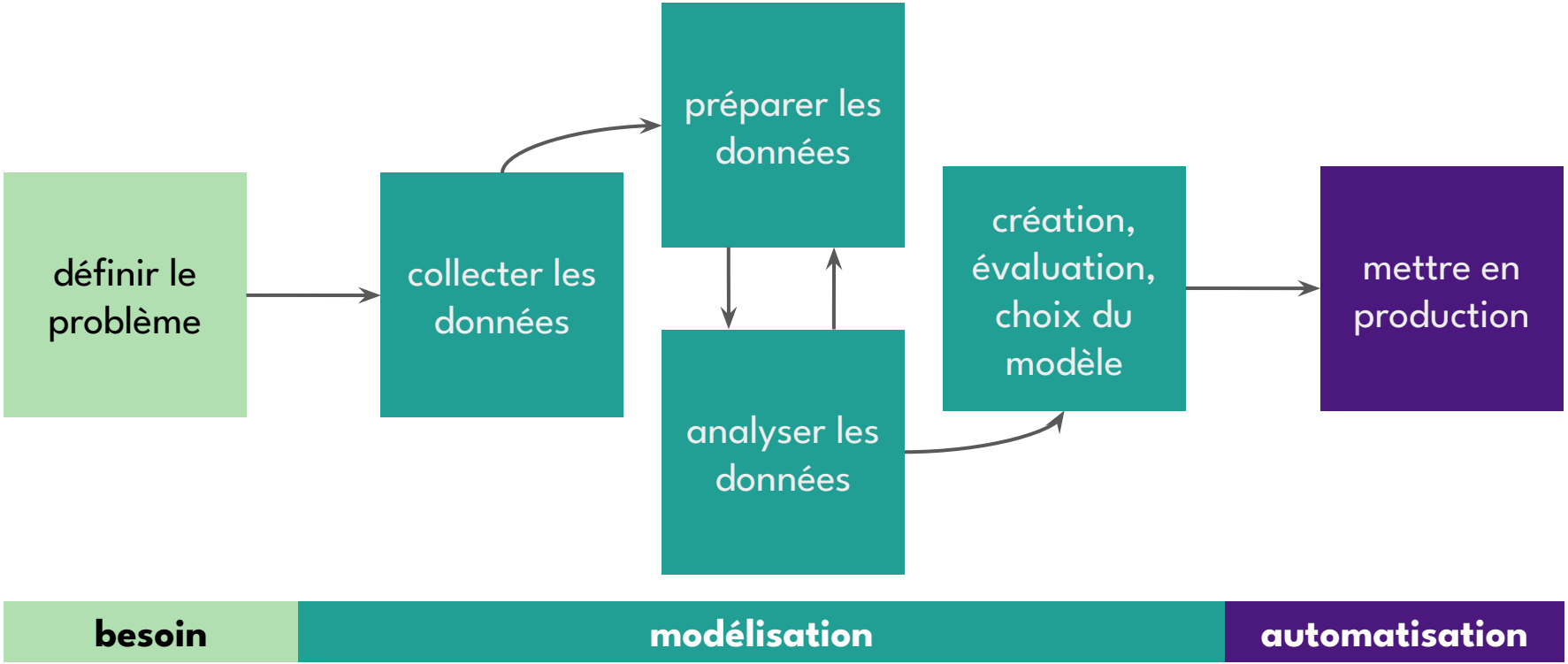
Tirer de la valeur des données

Les algorithmes, au support des différents métiers de l'entreprise

- **Marketing** : segmentation des utilisateurs → *statistiques descriptives*
- **Finance** : prédiction de chiffres d'affaires → *séries temporelles*
- **Relation client** : offres spéciales d'abonnement → *score d'appétence*

Tirer de la valeur des données

Processus de création d'un algorithme d'apprentissage automatique



Tirer de la valeur des données

Les algorithmes, au support des différents métiers de l'entreprise

- **Marketing** : segmentation des utilisateurs → *statistiques descriptives*
- **Finance** : prédiction de chiffres d'affaires → *séries temporelles*
- **Relation client** : offres spéciales d'abonnement → *score d'appétence*
- **Modération** : aide à la modération
- **Produit** : recommandation de profils → *algorithmes de recommandation*

Tirer de la valeur des données

Aider la modération

Les différents rôles de la modération :

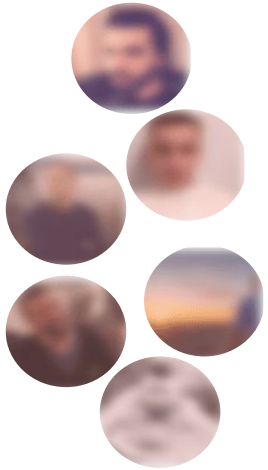
- repérer les faux profils → *score de risque*
- modérer les messages envoyés → *aide à la définition des mots clés*
- appliquer les règles du site, y compris **l'envoi de photos dénudées** (particulièrement quand elles sont non consenties)

Source : https://bumble.com/the-buzz/bumble-open-source-private-detector-ai-cyberflashing-dick-pics
Repo : https://github.com/bumble-tech/private-detector

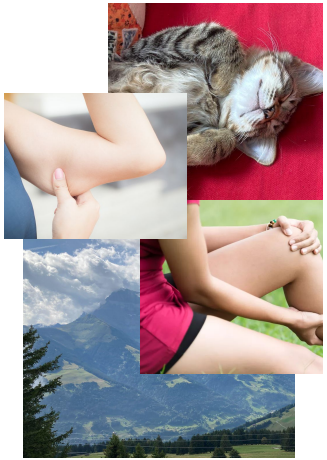
Repérer automatiquement les images à censurer

Deep Learning & Analyses d'images

Constitution des jeux de données (apprentissage/test)



Photos de profil



Photos (légitimes) échangées entre utilisateurs

Photos à censurer (dick pics) annotées par la modération

Source : https://bumble.com/the-buzz/bumble-open-source-private-detector-ai-cyberflashing-dick-pics
Repo : https://github.com/bumble-tech/private-detector

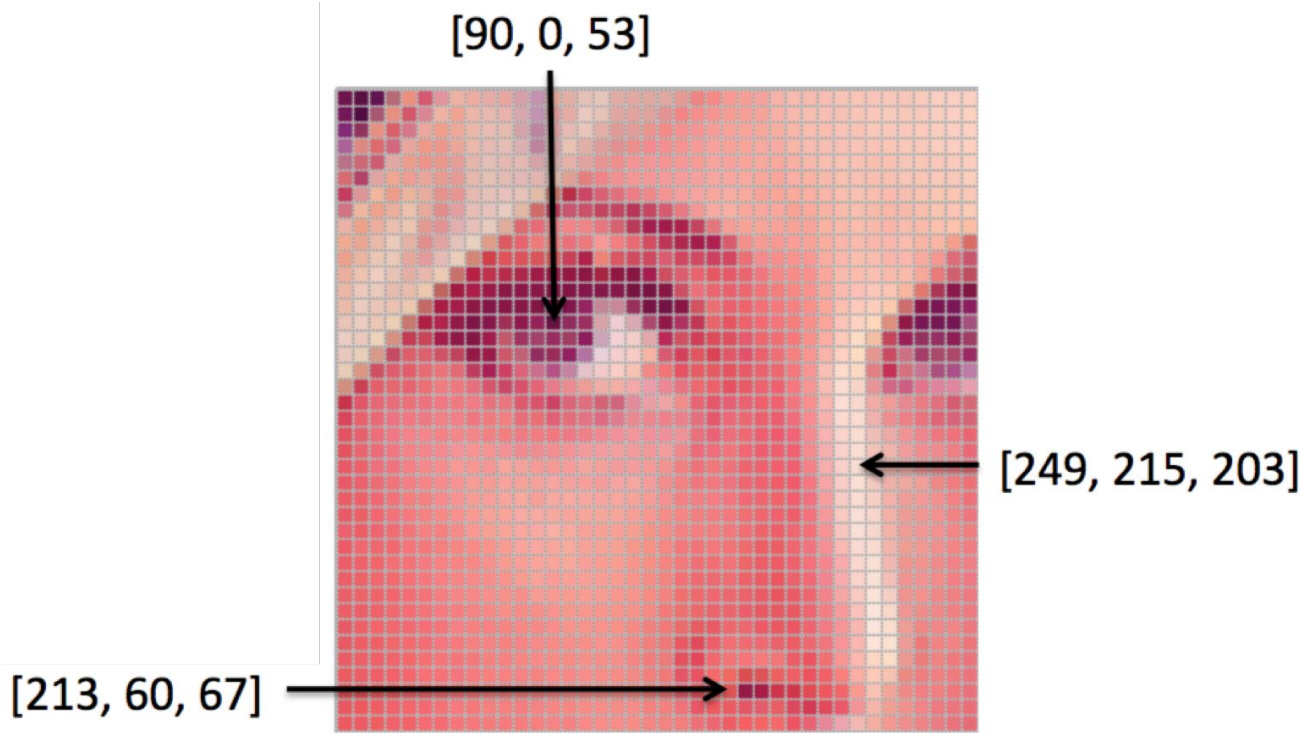
Repérer automatiquement les images à censurer

Mais… une image c'est des données ?

R = 255
V = 0
B = 0

R = 0
V = 255
B = 0

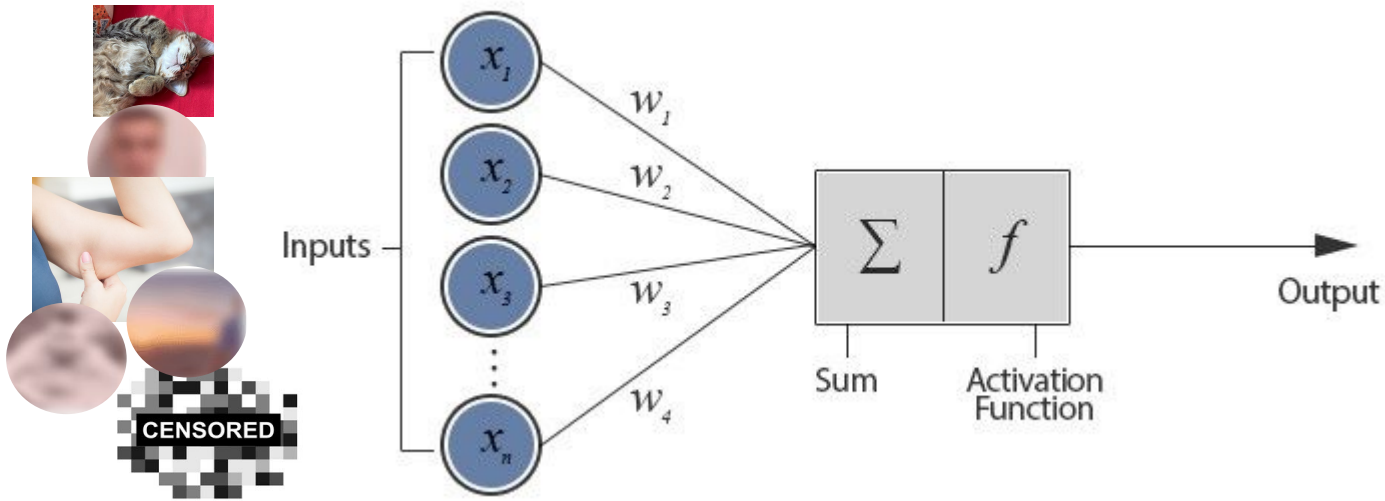
R = 0
V = 255
B = 0



Source : https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html

Repérer automatiquement les images à censurer

Deep Learning & Analyses d'images



Jeu d'apprentissage

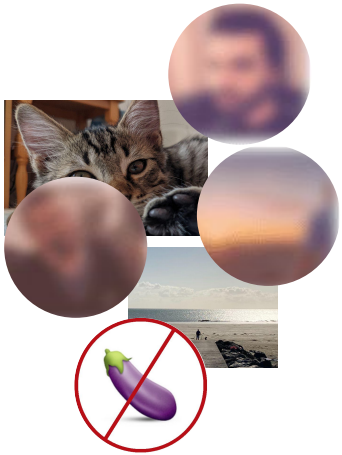
Image à modérer ou non

(score)

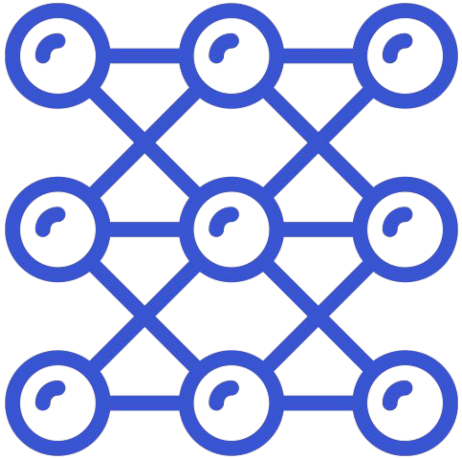
Source : https://towardsdatascience.com/conv2d-to-finally-understand...
Ce réseau est simpliste et à but d'illustration

Repérer automatiquement les images à censurer

Evaluation des résultats



Jeu de test
(nouvelles données)



Evaluation :

- taux de biens classés *(accuracy)*
- taux d'images effectivement à censurer parmi celles prédit comme telles *(précision)*
- taux d'images prédits comme à censurer parmi toutes celles effectivement à censurer *(rappel)*
- ...

Source : https://bumble.com/the-buzz/bumble-open-source-private-detector-ai-cyberflashing-dick-pics

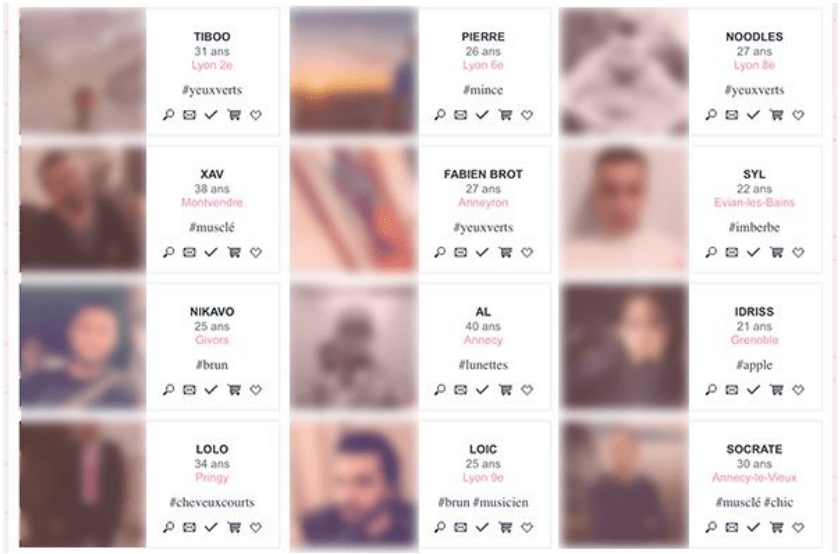
Tirer de la valeur des données

Les algorithmes, au support des différents métiers de l'entreprise

- **Marketing** : segmentation des utilisateurs → *statistiques descriptives*
- **Finance** : prédiction de chiffres d'affaires → *séries temporelles*
- **Relation client** : offres spéciales d'abonnement → *score d'appétence*
- **Modération** : aide à la modération
- **Produit** : recommandation de profils → *algorithmes de recommandation*

Algorithme de recommandation

Comment faire le tri parmi des millions de profils ?



En appliquant les filtres de recherche de l'utilisateur·rice, on obtient toujours des milliers de profils !

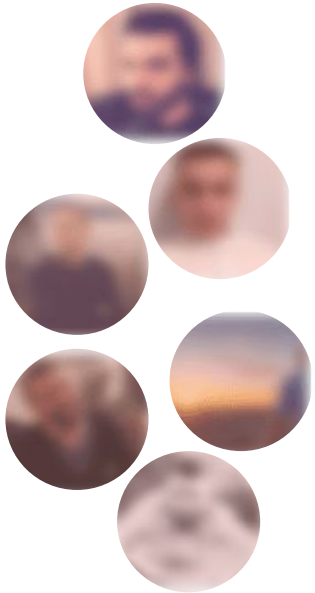
Extrait d'adopteunmec

Algorithme de recommandation

Comment faire le tri parmi des millions de profils ?

Profil
30 ans
Rennes

Recherche
Rennes +/- 21kms
18-32 ans

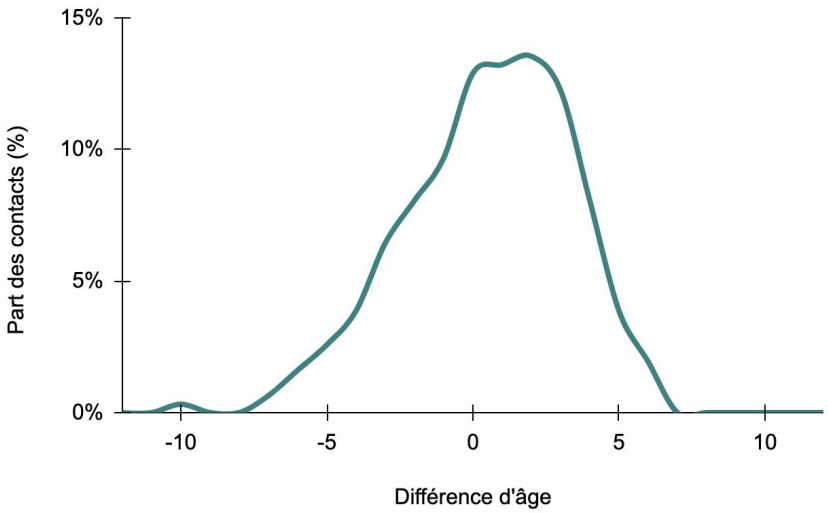


User	Profil	Choix	Ecart d'âge	Distance
Ketsia	David	Nope	- 8 ans	12 km
Ketsia	Marc	Nope	+ 12 ans	10 km
Ketsia	Amanda	Like	+ 1 an	5 km
Ketsia	Stella	Like	- 10 ans	2 km
Ketsia	Jacky	Like	**+ 5 ans**	**50 km**

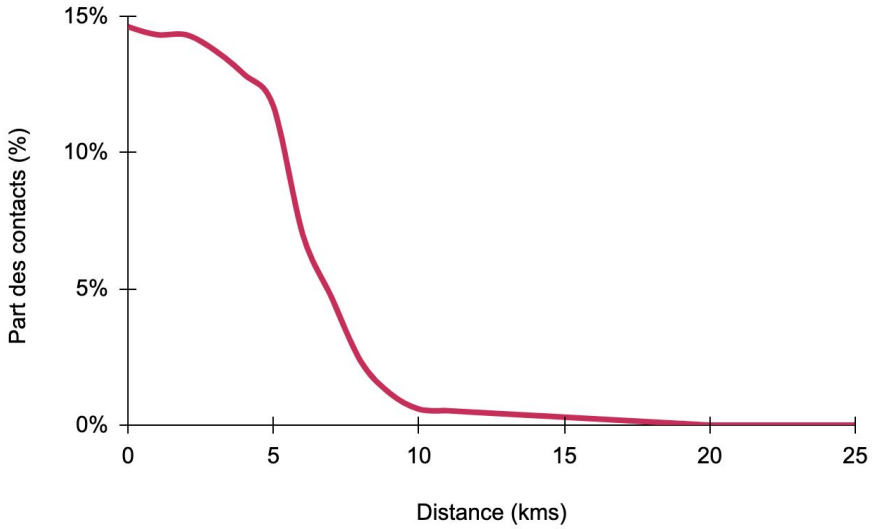
Algorithme de recommandation

Comment faire le tri parmi des millions de profils ?

Distribution des contacts en fonction de l'écart d'âge



Distribution des contacts en fonction de la distance



Données fictives

Algorithme de recommandation

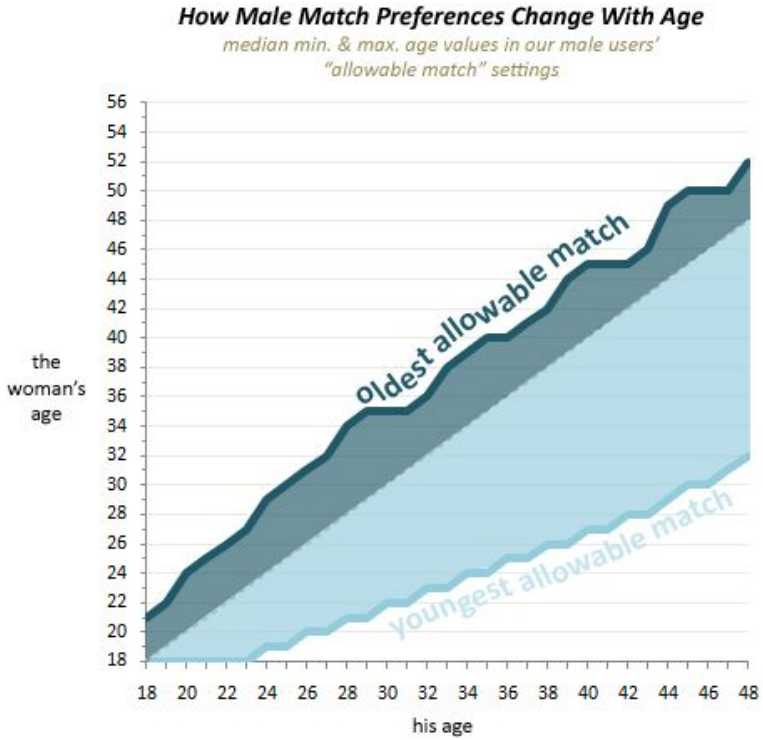
Score final

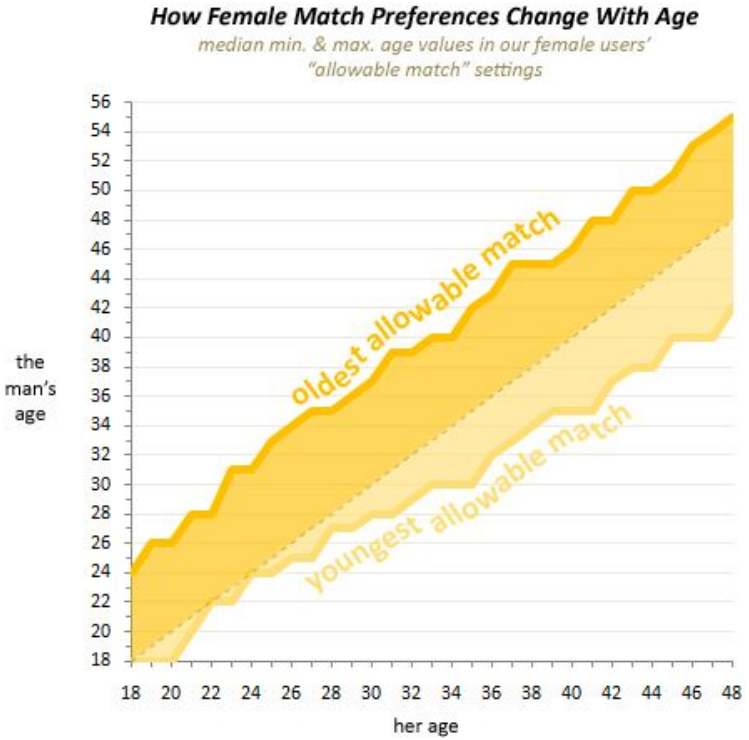
$$score_{woman} = \alpha_w \times f_w(distance\ géographique) + \beta_w \times g_w(écart\ d'âge) + \ldots$$

$$score_{man} = \alpha_m \times f_m(distance\ géographique) + \beta_m \times g_m(écart\ d'âge) + \ldots$$

Algorithme de recommandation

Pourquoi des scores différenciés par genre ?





Source : https://medium.com/@okcupid/the-case-for-an-older-woman-99d8cabacdf5

Références pour jouer avec les données de l'amour

Les mathématiques de l'amour, Hanna Fry, 2015

contient de nombreux exemples d'utilisations des mathématiques et des statistiques sur le thème de l'amour : calculer son nombre de partenaires potentiels, comment optimiser le placement des invités à un mariage, etc. Très agréable à lire, drôle et donnant beaucoup d'idées de petits contenus pédagogiques !

Jeux de données (Kaggle)

- ➔ Speed Dating Experiment, données issues d'une expérience de psychologie sociale
- ➔ OkCupid Profiles, 60 000 profils OKCupid anonymisés !

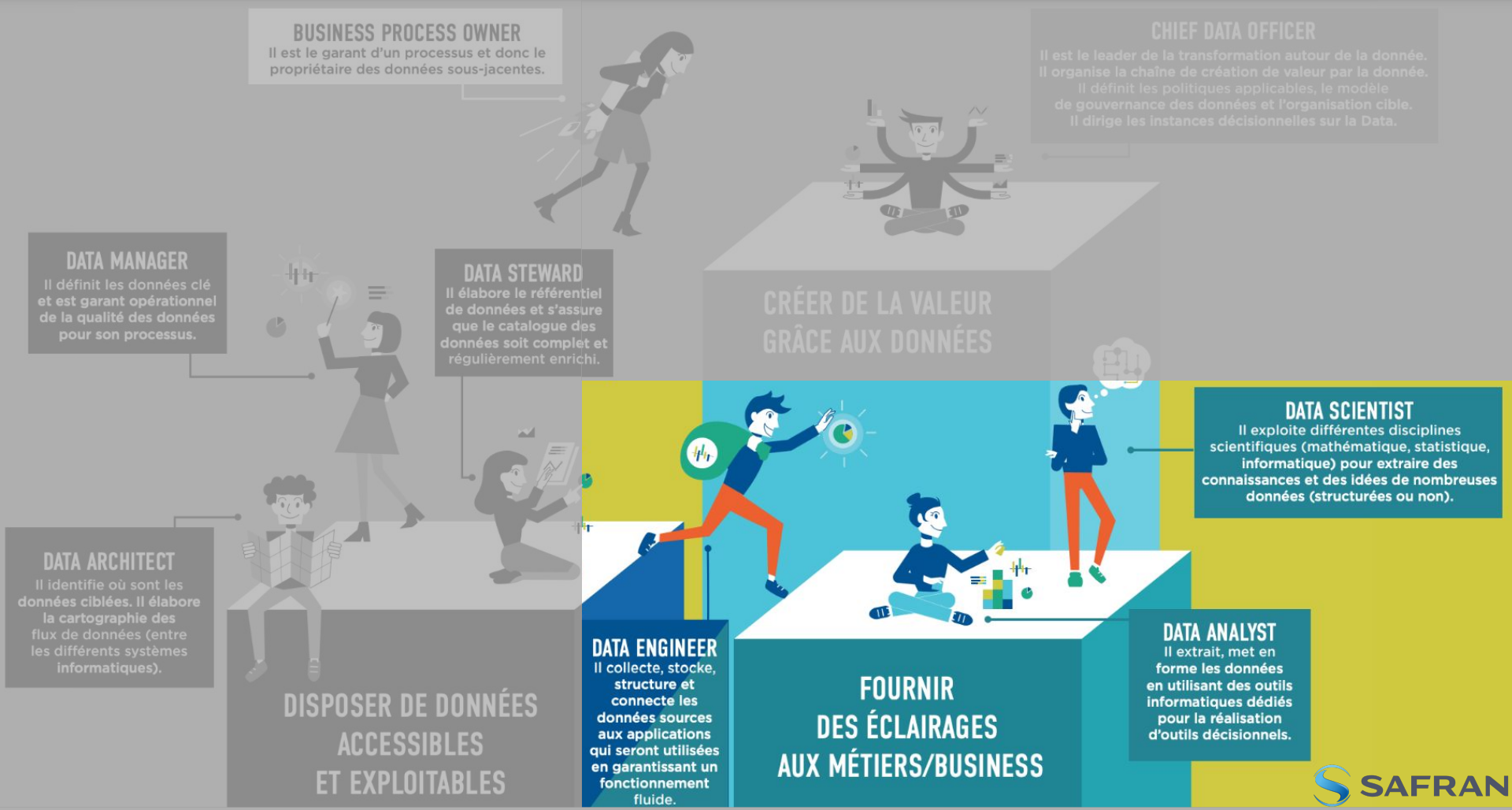
Jouer avec les modèles ou comprendre les algorithmes

- ➔ Bumble Private Detector - Bumble Buzzwords
- ➔ Finding Love on a First Data: Matching Algorithms in Online Dating, Harvard Data Science Review

Récapitulatif :
Les métiers de la donnée

SAFRAN

Source : https://www.safran-group.com/...



Source : https://www.safran-group.com/...

Data Analyst

Activités principales

- structurer les données
  - définir les règles de nettoyage des bases
  - maîtriser la qualité des données
- analyse et exploration de données
  - études
  - réaliser des tableaux de bord
  - visualisation de données
- communiquer avec les autres équipes
  - traduire les besoins de manière analytique
  - présenter les résultats

Compétences techniques

- requêtage de bases de données
- connaissance des tests et méthodes statistiques
- langages d'analyse de donnée (Python, R)
- outils de data visualisation (Tableau, PowerBI…)

Diplômes

- Bac+3 en statistique, traitement de l'information
- Bac+5 en statistiques, Big Data, parfois Marketing ou école de commerce

Source : Les métiers de la Data, APEC, 2019

Data Scientist

Activités principales

- ➔ structurer les données
  - définir les règles de nettoyage des bases
  - maîtriser la qualité des données
- ➔ analyse et exploration de données
- ➔ élaborer les algorithmes
  - construire les données d'entraînement
  - créer et tester des algorithmes d'apprentissage automatique
  - amélioration continue des modèles
- ➔ industrialiser les algorithmes

Compétences techniques

- ➔ requêtage de bases de données
- ➔ langages d'analyse de donnée (Python, R)
- ➔ très bonnes connaissances en statistique
- ➔ expertise en méthodes d'apprentissage automatique (Machine Learning, Deep Learning) et environnements de développements associés (TensorFlow, PyTorch, Keras…)

Diplômes

- ➔ Bac+5 en informatique, Data science ou statistique
- ➔ Bac+8 : doctorat en statistique ou informatique

Source : Les métiers de la Data, APEC, 2019

Data Engineer

Ingénieur·e data/big data, Développeur·euse data

Activités principales

- ➔ acheminer la donnée
  - collecter la donnée
  - choisir les solutions de stockage
  - développer l'infrastructure de donnée
- ➔ mettre à disposition la donnée
  - automatiser le nettoyage des données
  - maintenir et documenter les bases
- ➔ industrialiser les algorithmes

Compétences techniques

- ➔ gestion de bases de données
- ➔ langages de programmation (Scala, Python, Java…)
- ➔ connaissances de base en statistiques et intelligence artificielle

Diplômes

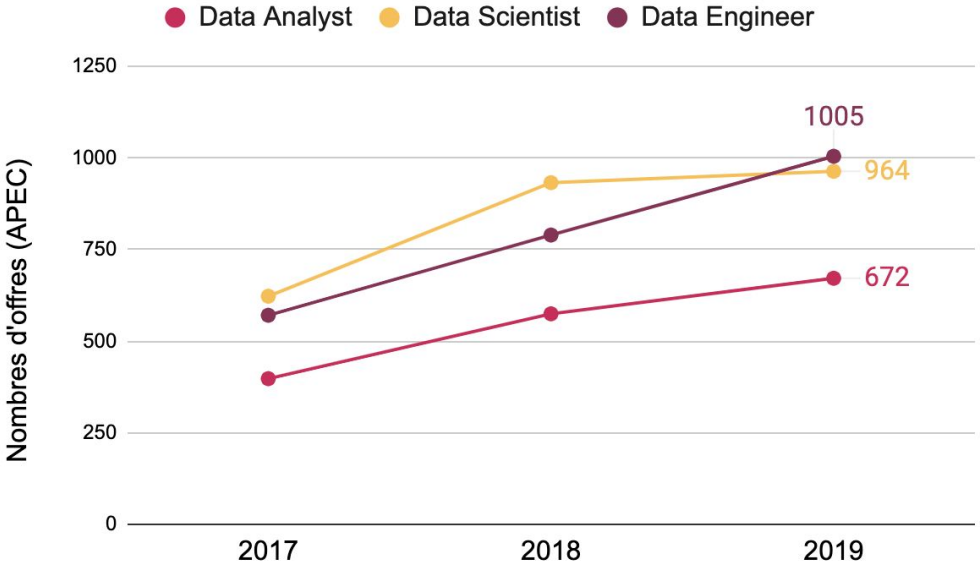
- ➔ Bac+5 en informatique, Data science ou statistique
- ➔ Bac+2 en statistique ou informatique avec une expérience dans le traitement des données

Source : Les métiers de la Data, APEC, 2019

Tendances du marché

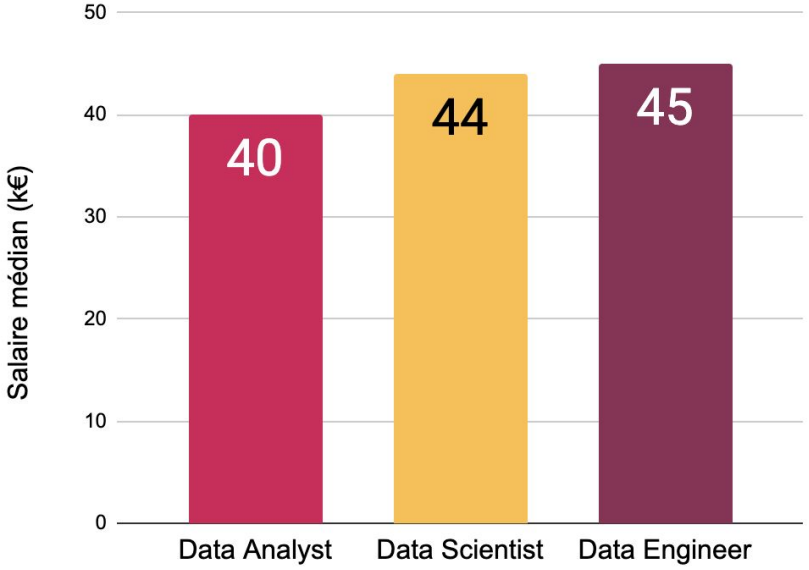
Evolution du nombre d'offres

Données APEC (2019)



Salaire médian (2019)

Données APEC



Source : Les métiers de la Data, APEC, 2019

Ressources

- ➔ Les métiers de la Data, APEC
- ➔ BD Data, la Data en 10 métiers, Syntec Conseil
- ➔ Le futur des métiers de la Data, Syntec Conseil
- ➔ Le marché du travail numérique en France, Indeed

Focus :

La place des femmes dans les métiers de la donnée

Etat des lieux : les femmes dans la tech

Selon le rapport DESI (Digital Economy and Society Index) de la Commission européenne,

- **19%** des spécialistes en TIC* sont des femmes en 2023 dans l'UE (**19% en France**) ;
  **TIC = Technologies de l'information et de la communication*

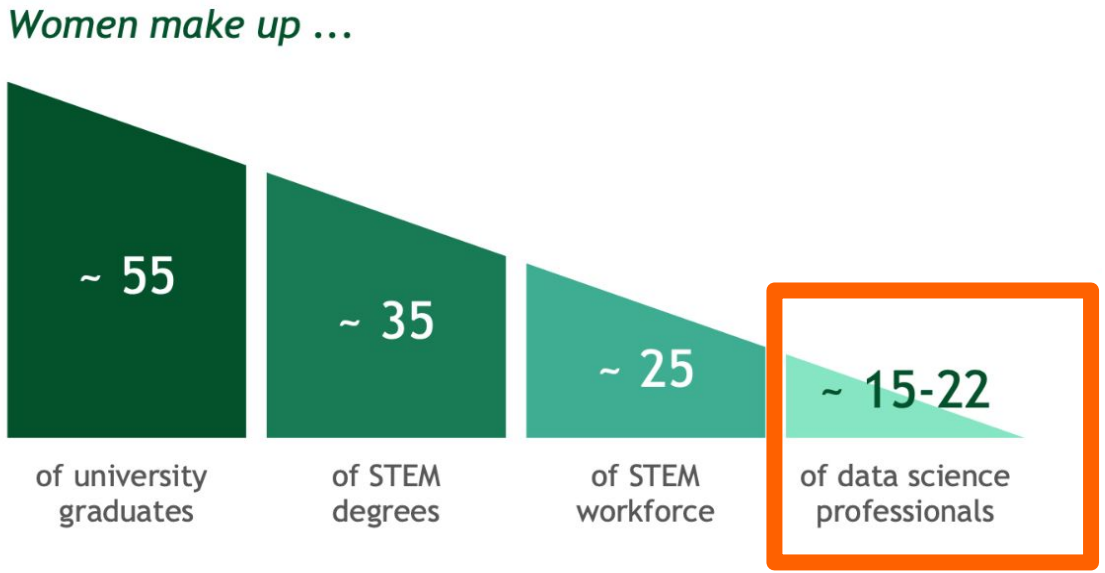
- Ce taux a baissé en France l'année dernière après une faible progression
  (-1pt en France, -0.2pt dans l'UE entre 2022 et 2023)
  (-0.4pts en France, +1.8pts dans l'UE entre 2018 et 2023)

- Le programme "Path to the digital decade" a donc ajouté en 2022 comme objectif **la convergence des genres** pour les spécialistes en TIC.

Source : https://digital-strategy.ec.europa.eu/fr/policies/desi
https://digital-strategy.ec.europa.eu/fr/policies/europes-digital-decade

Etat des lieux : les femmes dans la data

Share of women at each stage of the STEM talent funnel (in %)



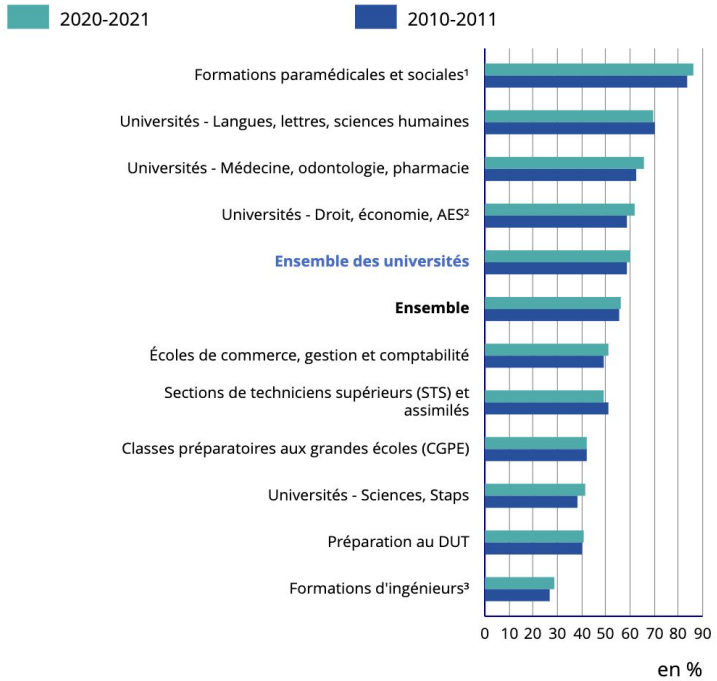
Quelques causes selon BCG :

- la data science a une image de secteur théorique, très "geek";
- réputation de compétitivité ;
- manque d'exemple de réels impacts pour les entreprises et la société.

Source : https://media-publications.bcg.com/Whats-keeping-women-out-of-data-science-bcg-gamma.PDF
Scope : Australia, Canada, China, France, Germany, India, Japan, Spain, the United Kingdom, and the United States

Etat des lieux : les femmes dans les études scientifiques

Figure 1 - Part des femmes dans les formations d'enseignement supérieur



L'INSEE (2022) constate une spécialisation des genres selon le type de cursus et les matières étudiées

"Au sein des disciplines scientifiques à l'université, seulement 23 % des étudiants en DUT dans les spécialités de production et d'informatique et 31 % des étudiants en sciences fondamentales et applications sont des femmes.

[...]

Ces écarts se sont ***peu réduits*** *durant la dernière décennie"*

Source : https://www.insee.fr/fr/statistiques/6047727?sommaire=6047805

L'orientation des lycéen·ne·s

Selon l'Observatoire sur la féminisation des métiers du numérique (Ipsos) les lycéennes ne se dirigent pas vers le numérique car :

- sous-estimation des compétences dans ces matières ;
- manque de représentation féminines ;
- crainte d'être désavantagées professionnellement dans ces voies ;
- parents moins encourageants vers ces voies.

94%

des lycéens pensent qu'il est important voire indispensable d'avoir un très bon niveau dans les matières scientifiques pour être admis et réussir dans une école d'informatique.

Or même lorsqu'elles ont plus de **14/20** de moyenne dans les matières scientifiques, les filles sont bien moins nombreuses que les garçons à penser avoir le niveau pour suivre une formation en école d'informatique (**43%** contre **78%**), et leurs parents sont du même avis.

37%

des lycéennes envisagent de s'orienter vers une école d'informatique ou une école d'ingénieur, contre **66%** des garçons. Pourtant **56%** des lycéennes sont intéressées par l'informatique / le numérique.

Source : Ipsos x Epitech

L'orientation des lycéen·ne·s

Le rapport *Les freins à l'accès des filles aux filières informatiques et numériques du lycée (Marianne Monfort et Manon Réguer-Petit - 2022)* prend l'exemple de l'enseignement de spécialité **Numérique et sciences informatiques (NSI).**

Elles soulèvent plusieurs pistes de l'exclusion des lycéennes :

- mise à l'écart des mécanismes d'entraide ;
- postulat selon lequel l'enseignement de l'informatique repose sur des acquis antérieurs qui seraient déterminants ;
- des stéréotypes de genre ;

"Leur impression grandissante de ne pas être « à leur place » est renforcée par l'interprétation, par elles-mêmes et par les professionnel·les qui les accompagnent, des hésitations qu'elles expriment comme autant de preuves d'une « erreur » d'orientation initiale"

Source : Les freins à l'accès des filles aux filières informatiques et numériques du lycée, Centre Hubertine Auclert

Pédagogie

*"**Une attitude soutenante** d'un point de vue social de la part de l'enseignant **n'est pas un élément particulièrement favorable** contrairement à un soutien davantage axé sur la relation académique (via les feed-back).*

*Se montrer attentif aux progrès et aux difficultés des élèves, **leur indiquer comment se corriger et s'améliorer** sont des comportements qui, s'ils sont perçus par les filles, **renforcent** sensiblement la vision qu'elles ont de l'utilité des mathématiques et **leur intérêt pour cette discipline**."*

Doriane Jaegers, Dominique Lafontaine
Aspirer à une carrière mathématique :
quel rôle jouent le soutien et les attentes de l'enseignant chez les filles et les garçons ?

Source : Doriane Jaegers et Dominique Lafontaine, « Aspirer à une carrière mathématique : quel rôle jouent le soutien et les attentes de l'enseignant chez les filles et les garçons ? », Revue française de pédagogie, 208 | 2020, 31-47.
https://m.centre-hubertine-auclert.fr/article/publication-de-l-etude-les-freins-a-l-acces-des-filles-aux-filieres-informatiques

Ressources

Place des femmes dans la tech/la data

- Rapports DESI (Europe/France)
- Women in Tech Statistics Show the Industry Has a Long Way to Go, Builtin
- RESETTING TECH CULTURE : 5 strategies to keep women in tech, GirlsWhoCode x Accenture
- What's Keeping Women out of Data Science?, BCG Gamma

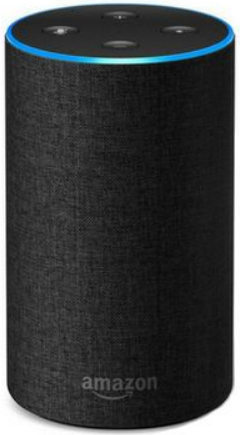
Part des femmes dans les études scientifiques

- Femmes et hommes, l'égalité en question (édition 2022), INSEE
- La disparition des filles dans les études d'informatique : les conséquences d'un changement de représentation, Isabelle Collet, Carrefours de l'éducation 2004
- Étudiantes en filières scientifiques : où en sommes-nous ?, L'Etudiant

Orientation dans le secondaire

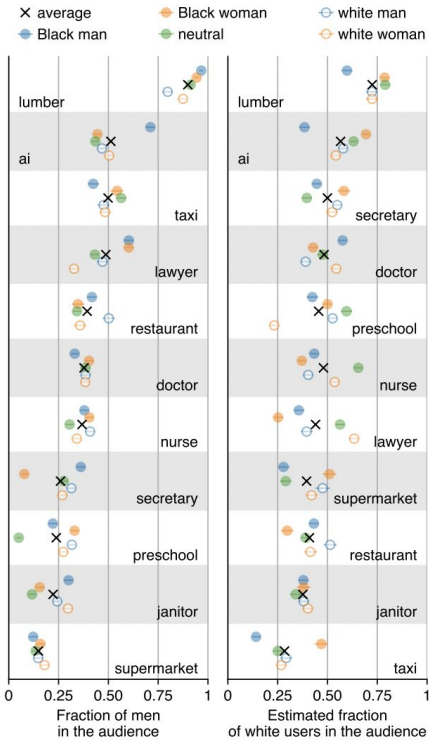
- Observatoire sur la féminisation des métiers du numérique, Ipsos x Epitech
- Réforme du lycée : quand les filles désertent les maths, les prépas scientifiques en pâtissent, l'Etudiant
- Do boys and girls differ in their attitudes towards school and learning?, PISA 2018
- Les freins à l'accès des filles aux filières informatiques et numériques du lycée, Centre Hubertine Auclert
- Égalité filles-garçons en mathématiques, Rapport de l'IGESR

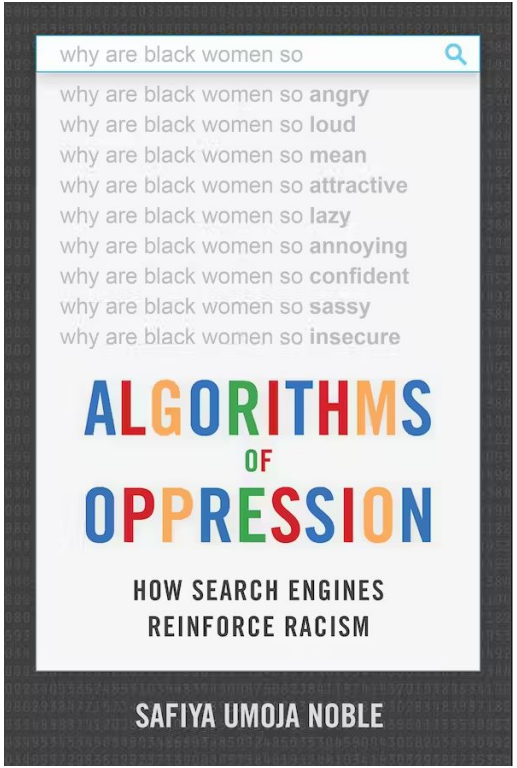
Les enjeux de la diversité



Source : Voice Recognition Still Has Significant Race and Gender Biases, Harvard Business Review
http://www.ethicsinnlp.org/workshop/pdf/EthNLP06.pdf

Les enjeux de la diversité





Sources: West, S.M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute.
Ali, Sapiezynski, Bogen, Korolova, Mislove, Rieke (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes
Safiya Umoja Noble (2018), Algorithms of Oppression : How Search Engines Reinforce Racism, NYU Press

Les enjeux de la diversité

"develop the advanced technical skills of women and girls so they can steer the creation of frontier technologies alongside men"

UNESCO, rapport EQUALS

Source : https://en.unesco.org/EQUALS/voice-assistants

Merci !

Des questions ?