

La Boîte à outils

de la

Stratégie big data

Romain Rissoan |
Romain Jouin |

DUNOD

Maquette de couverture : Caroline Joubert
Photo de la boîte : © Mega Pixel
Pictos de couverture :
© bioraven-Shutterstock.com
© Eliricon from Noun Project
© I Putu Kharismayadi from Noun Project

Mise en page : Belle Page

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du

Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, 2018
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
978-2-10-077898-0

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

VOUS AUSSI, AYEZ LE RÉFLEXE

Boîte à outils

Des outils
classés par
dossiers
thématiques

5
DOSSIER


IMAGE DE ET NOTORIÉTÉ

“
Être le meilleur est bien,
car tu es le premier.
Être unique est encore mieux,
car tu es le seul.”

Wilson Kanadi

Une présentation
visuelle de chaque outil

Exercices



EXERCICE 1 : FINALEMENT SA CONCENTRATION

- Fermez les yeux, représentez-vous le chiffre 1.
- Lorsque vous le voyez clairement en pensée, effacez de votre esprit l'image du chiffre 1.
- Représentez-vous le chiffre 2. Continuez ainsi jusqu'à 10.

EXERCICE 2 : LA MÉTHODE DE « L'ÉCOUTE AVEC LE CŒUR »

> La technique se résume en cinq questions

1. Qui s'est-il passé ?

Quelle émotion avez-vous ressentie ?

Quelle a été la plus difficile pour vous ?

Outil 33 Le Personal Branding

“
Aujourd'hui,
à l'ère de l'individu,
vous devez
être votre propre
marque.”

En quelques mots

Le Personal Branding ou la gestion de sa marque personnelle est un outil de réflexion et de mise en œuvre d'actions définies visant à contribuer à la construction de son image personnelle. En marketing de soi, le Personal Branding est l'ensemble des moyens, techniques et canaux que l'on utilise afin de construire son identité, se rendre visible et se promouvoir de façon pertinente et efficace. À l'instar des entreprises qui créent des marques, les rendent visibles, développent leur notoriété et travaillent leur image. Il est possible et utile de construire et mettre en avant sa propre « marque ».

LES COMPOSANTES DE LA VALEUR DE L'EXPIÉRIENCE POUR LE CLIENT

| | |
|--|---|
| Composants de la valeur perçue dans l'expérience | Facteurs associés par l'entreprise à l'impact de cette valeur |
| Émotion Fait espérer ou gîger de l'argent | des offres spéciales, des ventes flash, des cadeaux à gîger, des charts ou des algorithmes géniaux... |
| Reconnaissance Fait gîger du temps ou respecte le temps souhaité par le client | une aventure 24h/24, une livraison instantanée... |

Des exemples,
cas ou exercices
pour approfondir



La Boîte à outils

DES OUTILS OPÉRATIONNELS TOUT DE SUITE

MEGA Boîte à Outils

Manager leader - 100 outils

Coordonnée par Pascale Bêlorgey, Nathalie Van Laethem

Digital - 100 outils

Coordonnée par Catherine Lejealle

MÉTIERS

Acheteur, 3^e éd.

Stéphane Canonne, Philippe Petit

Assistante, 2^e éd.

Christine Harache, Héliène Tellitocci

Auditeur financier, 2^e éd.

Sylvan Boccon-Gibod, Eric Vilmint

Chef de produit, 2^e éd.

Nathalie Van Laethem, Stéphanie Moran

Chef de projet, 2^e éd.

Jérôme Maes, François Debois

Coach en entreprise, 2^e éd.

Belkacem Ammiar, Omid Kohneh-Chahri

Commercial, 3^e éd.

Pascale Bêlorgey, Stéphane Mercier

Community Manager

Clément Pellerin

Comptabilité, 2^e éd.

Bruno Bachy

Consultant, 2^e éd.

Patrice Stern, Jean-Marc Schoettl

Contrôle de gestion

Caroline Selmer

Création d'entreprise, 2018

Catherine Léger-Jarniou, Georges Kalousis

E-commerce

Christian Delabre

Formateurs, 3^e éd.

Fabienne Bouchut, Isabelle Cauden, Frédérique Cuisiniez

Management, 2^e éd.

Patrice Stern, Jean-Marc Schoettl

Micro-entrepreneur

Jacques Hellart, Caroline Selmer

Pilote des systèmes d'information, 2^e éd.

Jean-Louis Foucard

Publicité

Servanne Barre, Anne-Marie Gayrard-Carrera

Responsable communication, 3^e éd.

Bernadette Jézéquel, Philippe Gérard

Responsable financier, 2^e éd.

Caroline Selmer

Responsable marketing omnicanal, 3^e éd.

Nathalie Van Laethem, Béatrice Durand-Mégret

Responsable qualité, 3^e éd.

Florence Gillet-Goinard, Bernard Seno

Ressources humaines, 2^e éd.

Annick Haegel

Santé - Sécurité - Environnement, 3^e éd.

Florence Gillet-Goinard, Christel Monar

TPE

Guillaume Ducret

COMPÉTENCES TRANSVERSALES

Conduite du changement

David Autissier, Jean-Michel Moutot

Créativité, 2^e éd.

François Debois, Arnaud Groff, Emmanuel Chenevier

Design management

Bérangère Szostak, François Lenfant

Développement durable et RSE

Vincent Maymo, Geoffroy Murat

Gestion des conflits

Jacques Salzer, Arnaud Stimec

Innovation, 2^e éd.

Géraldine Benoit-Cervantes

Intelligence collective

Béatrice Arnaud, Sylvie Caruso-Cahn

Intelligence économique

Christophe Deschamps, Nicolas Moinet

Lean

Radu Demetrescoux

Leadership, 2^e éd.

Jean-Pierre Testa, Jérôme Lafargue, Virginie Tilhet-Coartet

Management de la relation client, 2^e éd.

Laurence Chabry, Florence Gillet-Goinard, Raphaëlle Jourdan

Management transversal

Jean-Pierre Testa, Bertrand Déroutède

Marketing digital

Stéphane Truphème, Philippe Gastaud

Mind mapping

Xavier Delengaïne, Marie-Rose Delengaïne

Mon parcours professionnel

Florence Gillet-Goinard, Bernard Seno

Négociation, 2^e éd.

Patrice Stern, Jean Mouton

Organisation, 2^e éd.

Benoît Pommeret

Prise de décision

Jean-Marc Santi, Stéphane Mercier, Olivier Arnould

Réseaux sociaux, 4^e éd.

Cyril Bladier

Sécurité économique

Nicolas Moinet

Stratégie, 2^e éd.

Bertrand Giboin

Stratégie digitale omnicanale

Catherine Headley, Catherine Lejealle

Supply chain

Alain Perrot, Philippe Villemus

DÉVELOPPEMENT PERSONNEL

Bien-être au travail

Clothilde Huet, Gaëlle Rohou,
Laurence Thomas

Confiance en soi

Annie Leibovitz

**Développement personnel
en entreprise**

Laurent Lagarde

Efficacité professionnelle

Pascale Bélorgey

Gestion du stress

Gaëlle du Penhoat

Gestion du temps

Pascale Bélorgey

Intelligence émotionnelle

Martine-Eva Launet, Céline Peres-Court

Marketing de soi

Nathalie Van Laethem, Stéphanie Moran

Motivation

Sophie Micheau-Thomazeau,
Laurence Thomas

Pleine conscience au travail

Sylvie Labouesse, Nathalie Van Laethem

Avant-propos

“

Un bon data scientist est intéressé par la résolution de problèmes, pas par de nouveaux outils.

KDNuggets

L'ouvrage que nous vous proposons est dédié au domaine de la data. Il a pour ambition de dresser un panorama des notions, concepts et outils les plus éprouvés sur le marché actuellement. Sans être technique, il vous permettra de connaître les cas d'usage des différents outils et comment ils se distinguent les uns des autres.

Le monde de la data

En 2012, la *Harvard Business Review* (*HBR*) publiait un article : « Data-scientist, the sexiest job of the 21st century ». Depuis, tout le monde veut être data scientist : le gouvernement français a fait de la data science un des piliers de la « nouvelle France industrielle » en 2015, la BPI finance à tour de bras des start-up proposant du machine learning ou du big data et la récente étude « France IA » montre à quel point la France a une carte à jouer dans ce nouveau domaine de l'analyse des données, mais que la partie n'est pas gagnée.

Ce que la *HBR* avait oublié de nous dire, c'est que data scientist est un travail compliqué. Un data scientist doit avoir 3 compétences :

1. informatique
2. mathématique
3. business

Complétées par une excellente capacité à communiquer, c'est-à-dire transmettre ces trois compétences majeures à un auditoire qui ne les a pas.

Ce contexte rend la tâche du data scientist quasi impossible à atteindre. C'est pourquoi nous travaillons en équipe, avec chacun ses compétences. Avec ce livre, vous apprendrez le vocabulaire nécessaire au travail en équipe indispensable dans le monde de la data. Vous pourrez ainsi mettre vos compétences particulières à disposition de vos collègues.

Ce livre est avant tout un livre sur la « stratégie big data » : aussi, ses quatre premiers dossiers s'attachent à expliciter les enjeux qui se cachent derrière ce phénomène :

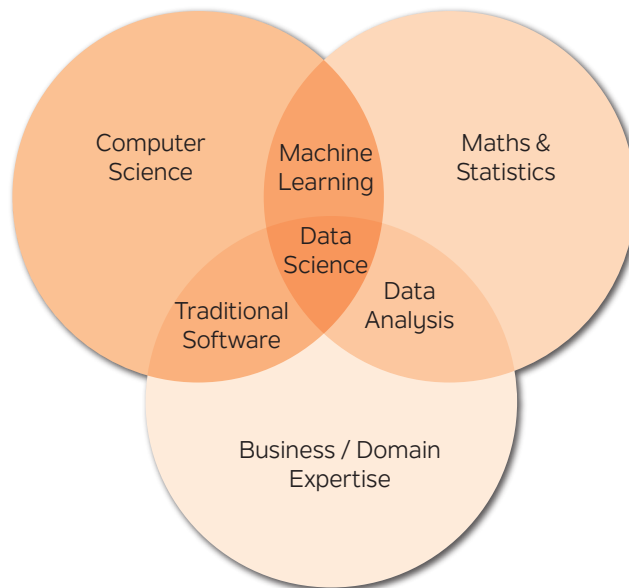
- Qu'est-ce qui a changé ces dernières années vis-à-vis de la data ?
- Quels sont les différents types de data ?
- Comment en tirer parti pour la stratégie de son entreprise ?
- Quels impacts techniques ?

Ces dossiers intéresseront plutôt le manager avide de comprendre comment profiter de cette opportunité qu'est le big data.

Viennent ensuite trois dossiers plus opérationnels, allant de la gestion de projets data à l'introduction des concepts de machine learning, et la différence avec les concepts de big data. Les personnes récemment nommés chef de projets dans un environnement big data y trouveront les clefs pour communiquer et comprendre les équipes avec lesquelles elles devront travailler.

Le dernier dossier, sur le passage en production, est une introduction aux différentes technologies usuellement mises en place dans les infrastructures big data. Il permet d'en comprendre les concepts, sans rentrer dans le détail technique opérationnel en tant que tel. Ainsi, les différents chapitres sont relativement indépendants, et s'adressent à un public de managers, ou chefs de projet, curieux du phénomène « big data ».

Classiquement, la data science nécessite six expertises distinctes, que l'on cherche chez le data scientist. Évidemment, il est bien difficile de disposer de l'ensemble, c'est pourquoi les projets sont généralement effectués en équipe. Il faut alors s'assurer de partager le même vocabulaire pour que le projet avance correctement.



Les compétences des data scientists

Sommaire

| | | |
|------------------|---|----|
| | Avant-propos | 6 |
| DOSSIER 1 | L'ÈRE DE LA DATA | 10 |
| | • <i>Outil 1</i> La data (la donnée) | 12 |
| | • <i>Outil 2</i> Le sens de la donnée (donnée brute) | 14 |
| | • <i>Outil 3</i> Deep Learning, machine learning, intelligence artificielle | 18 |
| | • <i>Outil 4</i> Le V du sens de la donnée | 20 |
| | • <i>Outil 5</i> Le V de la diversité | 24 |
| | • <i>Outil 6</i> Le V de la puissance des données | 28 |
| | • <i>Outil 7</i> Vie privée et big data | 32 |
| DOSSIER 2 | ENJEUX ORGANISATIONNELS DU BIG DATA | 36 |
| | • <i>Outil 8</i> Smart data, fast data, big data | 38 |
| | • <i>Outil 9</i> Les métiers du big data | 40 |
| | • <i>Outil 10</i> Informatisation et modernisation des systèmes informatiques | 42 |
| | • <i>Outil 11</i> Big data vs Business Intelligence au service de la marque employeur | 44 |
| | • <i>Outil 12</i> Communication transparente et spontanée | 46 |
| | • <i>Outil 13</i> L'ouverture et l'intégration : Open Source, SAAS et Webservice | 48 |
| | • <i>Outil 14</i> EDM, MDM, DMP et ETL | 50 |
| | • <i>Outil 15</i> S'appuyer sur le plan de l'entreprise | 52 |
| | • <i>Outil 16</i> S'appuyer sur les objectifs annuels | 54 |
| | • <i>Outil 17</i> S'appuyer sur des ateliers d'idéation | 56 |
| DOSSIER 3 | ENJEUX STRATÉGIQUES DU BIG DATA | 58 |
| | • <i>Outil 18</i> Analyse de l'environnement : Deeplist | 60 |
| | • <i>Outil 19</i> Analyse des enjeux big data : le SWOT | 64 |
| | • <i>Outil 20</i> Influenceurs du big data : les parties prenantes | 68 |
| | • <i>Outil 21</i> Appétence et maturité : le Cycle de vie | 72 |
| | • <i>Outil 22</i> Expectatives : la matrice BCG | 76 |
| | • <i>Outil 23</i> Intensité concurrentielle : la matrice Porter | 78 |
| | • <i>Outil 24</i> Digitalisation des services : la Chaîne de valeur | 80 |
| | • <i>Outil 25</i> Les risques du big data | 82 |
| | • <i>Outil 26</i> Choisir de devenir Data Driven | 84 |
| DOSSIER 4 | ENJEUX TECHNIQUES DU BIG DATA | 86 |
| | • <i>Outil 27</i> Données structurées et non structurées | 88 |
| | • <i>Outil 28</i> Lac de données | 90 |
| | • <i>Outil 29</i> Stockage distribué | 92 |
| | • <i>Outil 30</i> Calcul distribué | 94 |
| | • <i>Outil 31</i> Théorème de CAP | 96 |

| | | |
|------------------|--|-----|
| DOSSIER 5 | GESTION DU PROJET | 98 |
| | • <i>Outil 32</i> Gérer une équipe..... | 100 |
| | • <i>Outil 33</i> Formuler une question..... | 102 |
| | • <i>Outil 34</i> Comprendre les données..... | 104 |
| | • <i>Outil 35</i> Visualiser les données : la data visualisation..... | 106 |
| | • <i>Outil 36</i> Data cleaning : que faire des valeurs aberrantes ?..... | 108 |
| | • <i>Outil 37</i> Tester des modélisations..... | 110 |
| | • <i>Outil 38</i> Performance d'une classification binaire..... | 112 |
| | • <i>Outil 39</i> Performance d'une régression..... | 116 |
| | • <i>Outil 40</i> Le storytelling..... | 118 |
| | • <i>Outil 41</i> Présenter des résultats actionnables..... | 120 |
| | • <i>Outil 42</i> Enrichir le data set..... | 122 |
| | • <i>Outil 43</i> Agilité et Scrum..... | 124 |
| DOSSIER 6 | USAGE ET MAÎTRISE DES ALGORITHMES | 126 |
| | • <i>Outil 44</i> Typologies des algorithmes de machine learning..... | 128 |
| | • <i>Outil 45</i> Principes des algorithmes d'apprentissage supervisé..... | 130 |
| | • <i>Outil 46</i> Principes des algorithmes d'apprentissage non supervisé..... | 132 |
| | • <i>Outil 47</i> Principe des algorithmes par renforcement..... | 134 |
| | • <i>Outil 48</i> Principes du Deep Learning..... | 136 |
| | • <i>Outil 49</i> Les arbres de décision..... | 138 |
| | • <i>Outil 50</i> Le couteau suisse du data scientist : le Random Forest..... | 140 |
| DOSSIER 7 | CHOIX DES TECHNOLOGIES BIG DATA | 142 |
| | • <i>Outil 51</i> Python..... | 144 |
| | • <i>Outil 52</i> R : statistiques, cran, ggplot..... | 148 |
| | • <i>Outil 53</i> Scala et la programmation fonctionnelle..... | 150 |
| | • <i>Outil 54</i> Plateforme sur le Web : Dataiku..... | 152 |
| | • <i>Outil 55</i> L'écosystème Hadoop et la distribution Hortonworks..... | 154 |
| | • <i>Outil 56</i> Spark vs Flink..... | 158 |
| | • <i>Outil 57</i> SMACK : Spark, Mesos, Akka, Cassandra, Kafka..... | 162 |
| | • <i>Outil 58</i> Lambda et Kappa architecture..... | 166 |
| | • <i>Outil 59</i> Aller dans le cloud ?..... | 168 |
| DOSSIER 8 | MISE EN PRODUCTION | 170 |
| | • <i>Outil 60</i> Penser l'infrastructure..... | 172 |
| | • <i>Outil 61</i> DevOps..... | 174 |
| | • <i>Outil 62</i> Docker..... | 176 |
| | • <i>Outil 63</i> Infrastructure as code..... | 178 |
| | • <i>Outil 64</i> Haute disponibilité et Redondance..... | 180 |
| | • <i>Outil 65</i> Mise à jour des modèles prédictifs..... | 182 |
| | Glossaire..... | 185 |
| | Crédits iconographiques..... | 191 |

1

DOSSIER

L'ÈRE DE LA DATA

par Romain Risoan

“

C'est une erreur capitale que de bâtir des théories tant qu'on n'a pas de données. Insensiblement, on se met à torturer les faits pour les faire cadrer avec les théories, au lieu d'adapter les théories aux faits.

Arthur Conan Doyle (créateur de Sherlock Holmes)

Le V du sens des données

- Valeur
- Véracité
- Viabilité
- Validité

Le V de la diversité des données

- Variété
- Variabilité

Le V de la puissance de données

- Volume
- Vitesse
- Vélocité
- Visualisation
- Vulnérabilité
- Volatilité

Les *a priori* sur le big data sont nombreux. D'abord, nous sommes tentés de considérer que ce sujet est celui du Directeur des Systèmes informatiques ou du moins des informaticiens. Puis, nous sommes tentés de confier ce sujet à un mathématicien. Celui-ci nous ramène alors à la question de ce que nous voulons et de nos objectifs. C'est alors qu'intervient le Stratège, le dirigeant, le manager, ou le chef de projet afin de poser les perspectives d'un projet. Or ce dernier, bien souvent, a besoin de compréhension technique alors que celle-ci est bien souvent structurée à l'envers : en big data on exploite souvent les données dont on dispose et rarement les données que l'on a souhaité obtenir, car la mise en place de collecte de données en masse prend beaucoup de temps, tant sur le plan technique que réglementaire.

Comprendre le sujet

Ce sujet très populaire qu'est le big data est soumis à de multiples discours de synthèse qui mènent à de nombreuses simplifications en tout genre. Dans le même temps, nous avons besoin de cette simplicité pour avancer et prendre des décisions.

Mais globalement, il y a de nombreux manques de compréhension de points fondamentaux. Par exemple sur la définition même de ce qu'est une donnée.

Aussi, il s'agit d'avoir une approche à la fois inductive et déductive sur le sujet pour comprendre celui-ci dans sa profondeur et sa complexité.

Comprendre l'avenir de la data

Pour beaucoup, le big data ne veut rien dire et ne prend pas de sens. Ceci pour deux raisons : d'abord nous n'avons pas suffisamment de données pour s'imaginer les conséquences de l'entrée dans l'univers big data. Ensuite, il s'avère que certains d'entre nous font du big data sans le savoir.

Dans ce contexte, il convient de comprendre le monde de demain, représenté par des acteurs majeurs de la data (Facebook, Amazon), de leur gestion de celle-ci et de leurs traitements des données. Ce sont de parfaits pionniers de cet univers, qui nous aident à concevoir le fonctionnement du monde de demain.

Les outils

| | | |
|---|--|----|
| 1 | La data (la donnée) | 12 |
| 2 | Le sens de la donnée (donnée brute) | 14 |
| 3 | Deep Learning, machine learning, intelligence artificielle | 18 |
| 4 | Le V du sens de la donnée | 20 |
| 5 | Le V de la diversité | 24 |
| 6 | Le V de la puissance des données | 28 |
| 7 | Vie privée et big data | 32 |



La data (la donnée)

par Romain Rissoan

“

La rétention de l'information est une forme de constipation du savoir.

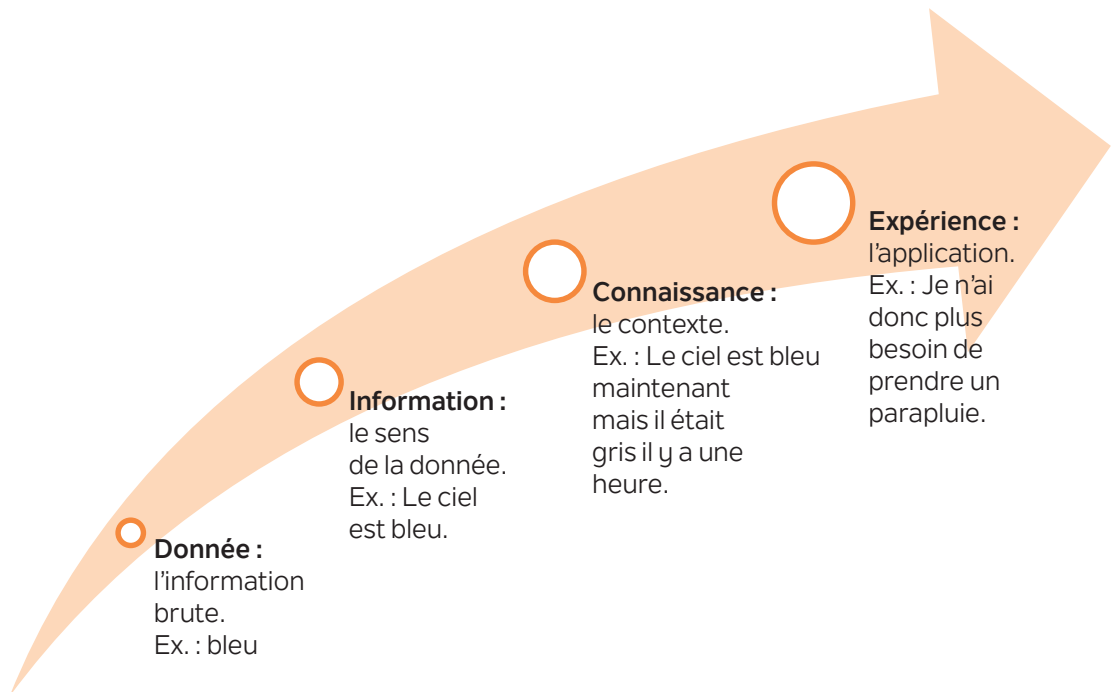
Théophraste Renaudot

En quelques mots

La donnée est l'**information élémentaire** dont nous avons besoin pour créer une cascade d'informations.

Plus nous avons de données, plus nous pourrons créer des **informations fiables** et de **qualité**. Et, plus nous travaillerons proche de la donnée, moins il y aura de **distorsion** et plus nous appliquerons des méta-analyses plutôt que des analyses.

LA DATA (LA DONNÉE)





POURQUOI L'UTILISER ?

Objectif

Apprendre à travailler avec des données et pas nécessairement avec des fichiers.

Contexte

La représentation ainsi que la valeur perçue d'une donnée ont bien évolué depuis la naissance de l'informatique.

Sur le plan informatique, des séquences de 0 et de 1 sont des **données** ou **data** (en anglais). Suite à une interprétation, ces données prennent un sens exploitable pour l'homme et pour la machine au travers de fichiers que l'on appelle alors **information**.

Sur le plan utilisateur, la **donnée** serait par exemple un nom, un prénom, une date de naissance, un sexe, alors que l'information serait un âge (calculé sur la date de naissance), le nombre de fois où cette personne s'est connectée à votre site Internet.

En résumé, une information est le résultat du traitement d'une donnée. Mais le traitement d'une donnée donne naissance à une nouvelle information.

Depuis que l'informatique s'est accélérée, ranger ses e-mails et ses fichiers Excel dans son ordinateur est moins efficace.

On doit travailler le plus en amont possible de ce flux. Ainsi, on parle plus de gérer ses données (ses flux d'informations) que de gérer ses fichiers (ses stocks d'information).



COMMENT L'UTILISER ?

Étapes

1. Prenez l'exemple d'un fichier Excel que vous partagez avec 10 collaborateurs sur un projet à fortes interactions. Rangez ce fichier dans un dossier dans votre ordinateur.
2. Déposez-le sur un espace de travail collaboratif comme Google Drive (ou Excel 365) pour pouvoir travailler de manière interactive sur ce fichier avec vos autres collaborateurs.

3. Créez un second fichier de type tableur dont les données évoluent en fonction du premier fichier.

4. Vous constaterez donc qu'à chaque fois que le fichier initial est modifié par un de vos collaborateurs, l'autre est modifié. Vos données vivent toutes seules. Vous êtes ainsi concentré sur des données plus que sur des fichiers.

Méthodologie et conseils

Ce mode de fonctionnement peut être nouveau pour certaines personnes. Aussi, il est important de manager le changement et non pas d'admettre que ce mode de fonctionnement est « La » vérité. Certains arguments sont justifiés : une utilisation épisodique de la donnée (une personne va voir le fichier une fois par an et a besoin de l'avoir bien rangé dans ses dossiers pour être certain de ne pas le perdre), l'absence potentielle de connexion à Internet (si je n'ai plus Internet, je n'ai plus accès à mes données) et le besoin de maîtrise de la donnée (si tout le monde peut modifier à n'importe quel moment mon fichier, je peux ressentir une perte de sécurité).

Dans ce contexte, il convient d'opérer ponctuellement et durablement des ateliers de transformation digitale pour inciter les utilisateurs à travailler autrement et ainsi permettre la migration vers la data.

Avant de vous lancer...

- ✓ **Toute donnée peut donner une information. Mais toute information peut donner naissance à une autre information.**
- ✓ **Le plein potentiel d'une donnée est préservé lorsqu'elle reste à l'état brut.**

“

La donnée est une chose précieuse qui restera plus longtemps que les systèmes.

Tim Berners-Lee

Le sens de la donnée (donnée brute)

par Romain Risoan

En quelques mots

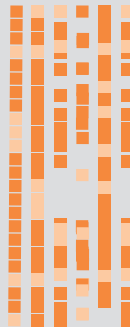
Le point de départ d'un projet big data est bien évidemment la data. Or, le sens même de ce que peut représenter la data n'est pas évident pour une organisation ou plutôt pour la culture de cette organisation. En effet, encore aujourd'hui, il n'est pas rare de constater que la majorité des structures considère que l'expression de besoin client peut se résumer à des informations écrites sur une feuille de papier blanc ajouté à la mémoire que le commercial en a retenu, ce qui finira au mieux en synthèse sur un CRM (un logiciel de gestion de relation client), au pire en un devis perdant ainsi toute trace de l'expression du besoin client.

LE SENS DE LA DONNÉE

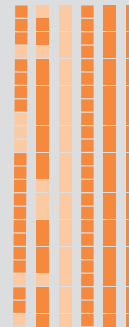
BIG DATA



ANALYTICS



DECISIONS





POURQUOI L'UTILISER ?

Objectif

Comprendre le véritable sens et intérêt d'une donnée à l'état brut.

Contexte

Au quotidien, nous travaillons principalement sur la base d'informations plus que de données. Fichiers Word, Excel, e-mails, tout ce que nous utilisons est en soi le résultat d'un traitement, d'une interprétation de nos données, les rendant ainsi orientées selon une analyse prédéfinie et une décision également prédéfinie.

L'idée du big data est justement de préserver la donnée à son état le plus pur afin de permettre des analyses différentes et des décisions innovantes.



COMMENT L'UTILISER ?

Étapes

1. Prenez votre appareil photo habituel.

Activez la prise de vue au format RAW (en plus du format .jpg habituel).

2. Prenez une photo au hasard.

Vous obtiendrez alors une photo à l'état brut (non compressée) au format .CR2 (fichiers appelés comme étant de type RAW).

Cette photo est la version non compressée et non déformée de votre prise de vue.

Elle est plus lourde que votre fichier .jpg habituel. Vous avez des exemples ici : <https://www.imaging-resource.com/PRODS/T21/T21THMB.HTM>

3. Lancez le logiciel Photoshop ou l'application <https://raw.pics.io/app> et ouvrez le fichier .CR2 avec l'application choisie.

4. Vous pourrez alors constater toutes les informations techniques lors de la prise de vue, comme le temps d'ouverture de l'objectif, les coordonnées GPS du lieu de la prise de vue, la focale, etc.

De plus, et surtout, vous pourrez alors modifier la photographie depuis son format d'origine et non depuis son expression au format .jpg, qui est un format compressé et déformé de la photo originale.

Méthodologie et conseils

Ayez toujours à l'esprit qu'une donnée telle qu'elle est exploitée aujourd'hui pourrait être exploitée différemment demain. Ainsi, plus vous stockerez de données à l'état brut, plus vous préservez l'avenir de celles-ci.

Ainsi :

- si vous traitez des fichiers audio, stockez des fichiers .FLAC,
- pour les photos stockez des .CR2,
- pour des fichiers texte stockez du .DOC, .CSV, .TXT ou .DOCX (mais pas du PDF).

Grâce à cela, vous préservez les fichiers originaux et leurs métadonnées associées.

Ces fichiers occuperont évidemment plus de place sur vos supports de stockages et surtout vos serveurs. N'hésitez pas à utiliser des offres low cost sur Internet.

Suite outil 2 →

Avant de vous lancer...

- ✓ Pour faire du big data à long terme, stockez les données à l'état brut.
- ✓ Grâce à cette approche, vous disposerez d'une capacité d'analyse plus large et donc d'une capacité de décision plus importante.



Photo ABC SARL

De l'argentique au numérique

Pendant des années, nous avons pris des photos avec des appareils photos argentiques. Les négatifs étaient alors transformés en photographie papier, puis rangés avec plus ou moins de soin dans des pochettes en vue d'un stockage long. Puis, nous avons découvert un nouveau type de donnée, le format numérique. Les appareils photos prenaient alors des photos dans un format nommé BITMAP, rapidement abandonné pour laisser place à un format de donnée appelé JPEG, qui est un format compressé de la donnée photo pour pouvoir permettre le stockage de celle-ci. Par la suite, et après de nombreuses années, est arrivé le format RAW qui est un stockage à l'état brut de la donnée, permettant de nombreuses retouches mais prenant une place conséquente sur nos disques durs. Ce dernier format permet à n'importe quel logiciel de photo de recréer toutes les couleurs sur une photo existante. Ainsi il est par exemple possible de redessiner la couleur verte d'une prairie sur ces nouveaux formats.

Data et métadonnées

La société Photo ABC SARL a construit son modèle économique sur la vente de photos dans de nombreuses tailles : photos de portrait, posters, panneaux publicitaires. Pendant des années, ses photos ont été prises en argentique et elle a donc stocké des milliers et des milliers de négatifs dans des entrepôts. Elle possède toujours ces négatifs et a bien tenté de les numériser, mais les rendus sont dégradés et ne permettent pas des opérations professionnelles. À l'époque, on pouvait trouver une image grâce au moteur de recherche et l'arborescence des dossiers.

Par la suite, après avoir rejeté les formats BITMAP, elle a pendant de nombreuses années travaillé sur des photos numériques au format JPEG. Ces formats ont été stockés dans des dizaines de disques durs de plusieurs giga-octets mais le format compressé de ces fichiers JPEG a causé quelques pertes de fichiers rendant inexploitable environ 10 % de ceux-ci.

De plus, ces photos ne contiennent pas toutes les informations que l'on retrouve de nos jours dans les fichiers de type photo, et que l'on nomme les métadonnées : auteur de la photographie, marque de l'appareil qui a pris la photographie, date de prise de vue, etc.

En utilisant les métadonnées, il est désormais possible de rechercher des photographies également en recherchant par auteur, marque de l'appareil.

Depuis cinq ans, la société travaille désormais exclusivement sur la base de fichiers RAW. Ces fichiers pèsent environ 50 Mo, soit dix fois plus que les fichiers JPEG. Ils contiennent désormais les coordonnées GPS du lieu de la prise de vue et toutes les données numériques pour reconstruire la colorimétrie de la photographie de la manière la plus adaptée possible, selon les besoins. Ainsi, une même photo prise en plein jour peut être retravaillée pour donner un effet crépuscule. Ces fichiers RAW sont ensuite convertis en fichiers JPEG. Les fichiers JPEG étant toujours de taille importante, sont ensuite réduits et dépourvus de certaines données stratégiques comme les coordonnées GPS afin d'être diffusées aux clients en guise d'échantillon ou de vente directe.

Le stockage de la data

La société Photo ABC SARL doit donc évidemment stocker ses fichiers au format le plus riche possible, le format RAW, afin de permettre de puissants traitements ultérieurs. Grâce à ce stockage, certes coûteux, elle peut désormais solliciter par exemple un moteur de recherche sur la base d'une carte géographique ou bien même d'une reconnaissance de photo.

Ainsi en tapant par exemple « PACA » ou « mer » dans son moteur de recherche, elle pourra trouver toutes les photos prises en région PACA ou avec la mer présente sur la photo ou à proximité de la prise de vue.

Compte tenu de ses contraintes de stockage et du niveau de risque que cela représente pour la société de tout stocker en local (dans ses locaux), elle décide alors de stocker ses photos sur un serveur distant (on parle de cloud privé).

Puis, se rendant compte de la complexité de le gérer elle-même, elle le stockera sur un serveur distant appartenant à une société bien connue du grand public, Microsoft, sous la marque Azure (on parle alors de cloud public).

Pour cause, elle produit deux fois plus de photographies que cinq ans auparavant, chaque photo faisant une taille cinq fois plus importante et elle ne souhaite pas rajouter des disques durs sur ses serveurs et faire des copier/coller de téra-octets de données.

