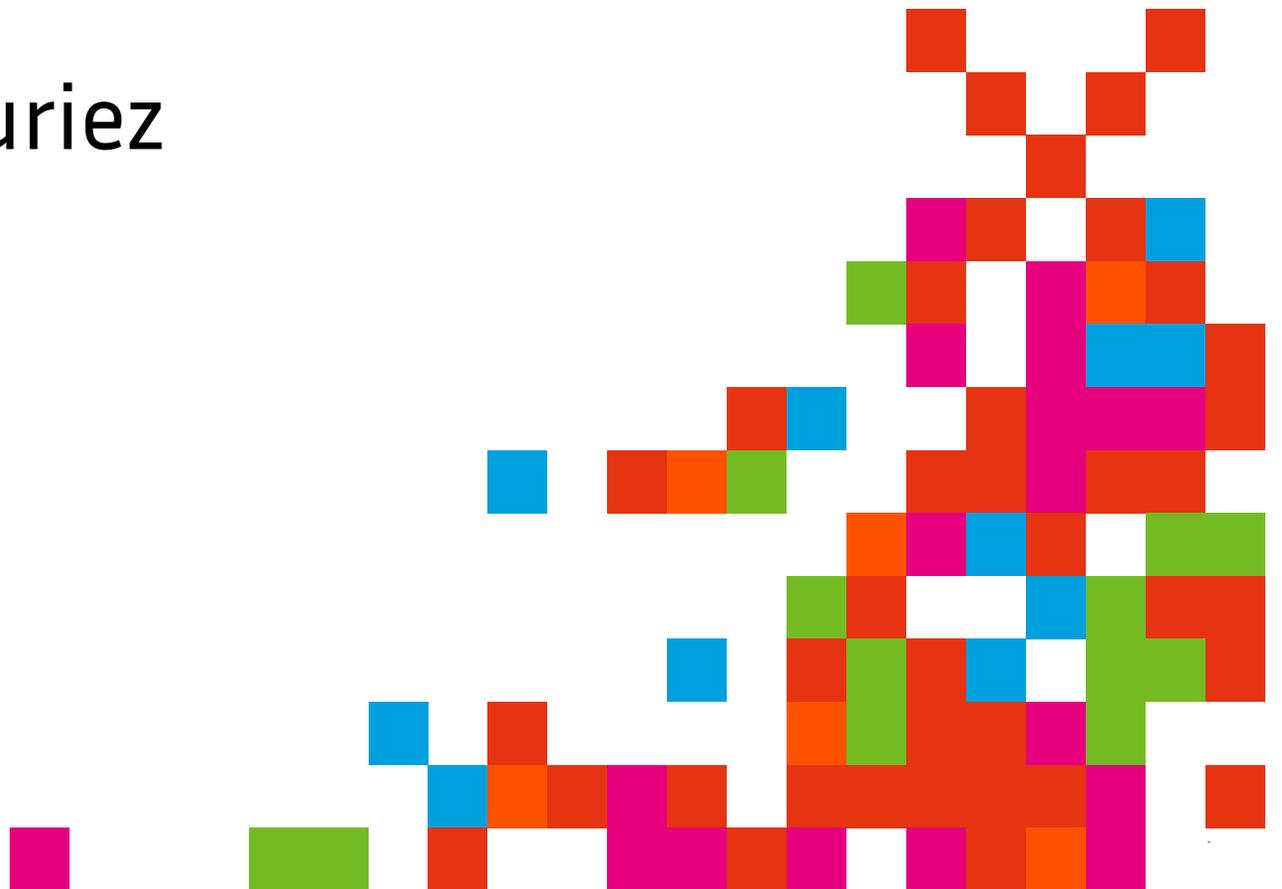


Les Données en Question

Patrick Valduriez



Beaucoup de jargon

- **Données**
 - Forfait data
 - Full data
 - Big data
 - Small data
 - Lean data
 - Smart data
 - Open data
 - ...
- **Gestion des données**
 - Data center
 - Data hub
 - Data base
 - Data bank
 - Data science
 - Data intelligence
 - ...



Donnée et Information

- **Donnée**

- Description d'un fait connu
 - Ex. température, note d'examen, message, photo, transaction, etc.
- Peut être complexe
 - Ex. une image satellite
- On peut en avoir beaucoup

- **Information**

- Obtenue par intégration, interprétation et analyse pour nous renseigner dans un *contexte*
- Permet de comprendre notre monde
 - Ex. évolution du climat, classement d'un élève, reconnaissance d'éléments dans une image satellite, etc.

Les données en question

PAR
Stéphane Grumbach
Patrick Valduriez

NIVEAU DE LECTURE
Facile ● ● ●

PUBLIÉ LE
31/03/2016



Au cœur de la connaissance et de l'information, les données ont peu à peu pris une importance qui nous dépasse. Mais qu'entend-on exactement par données ? Quels sont les enjeux autour de leur gestion ou de leur analyse ? Quels impacts sur la société ?



© Fotolia - ptinphotof

Une donnée est la description élémentaire d'une réalité ou d'un fait, comme par exemple un

<https://interstices.info>

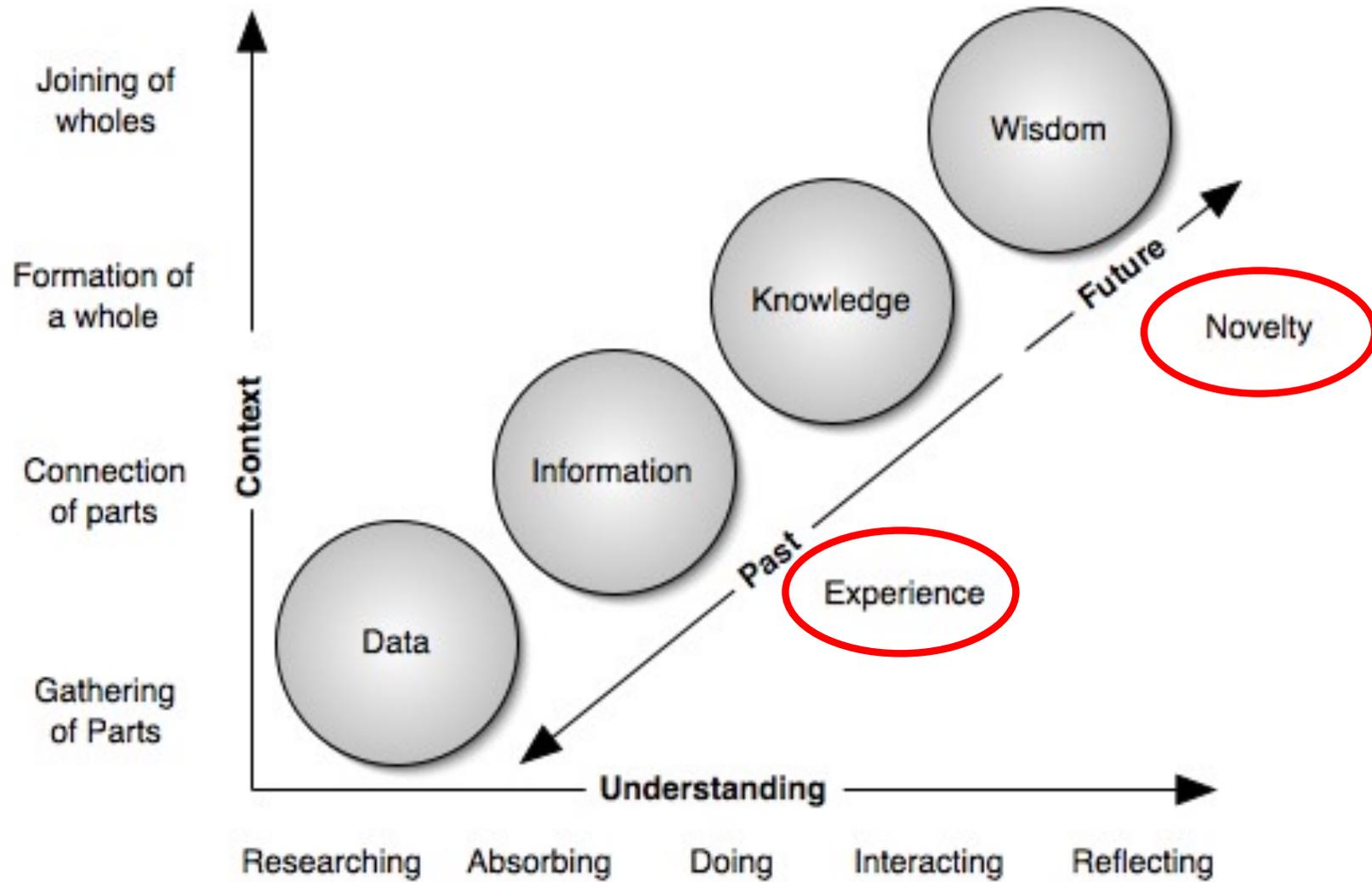
Donnée et Algorithme

"Le contenu sans méthode conduit à la fantaisie,
la méthode sans contenu au discours creux."

J.W. von Goethe (Maximes and Réflexions, 1892)

- Plus il y a de données, plus un algorithme est efficace
- Repères
 - 1997: IBM Deep Blue bat le champion du monde d'échecs Garry Kasparov
 - Algorithme de planification Negascout (1983)
 - Dataset de 700 millions de parties d'échecs
 - 2016 : Google Alphago bat le maître de Go Lee Sedol (4-1)
 - Algorithme basé sur la méthode de Monte Carlo (des années 1940) et un réseau de neurones
 - Dataset de 30 millions de mouvements de go

Le continuum de la compréhension



Gestion de données

- **Définition**

- La collecte, le nettoyage, l'organisation, le stockage et la manipulation des données pour produire des informations

- **1975 : ACM lance SIGMOD**

- Special Interest Group on the Management of Data

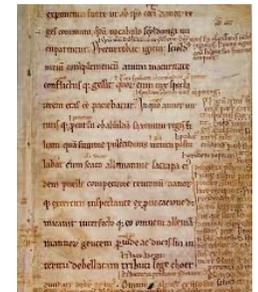
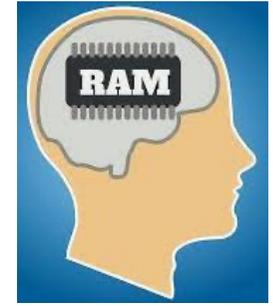
- **Mais tout a commencé il y a bien longtemps**

- 2238 avant JC : recensement de la production agricole par les Chinois
- 1700 avant JC : cadastre (pour la perception des impôts fonciers) par les Egyptiens
- Les origines de la statistique, de la comptabilité et d'autres disciplines scientifiques



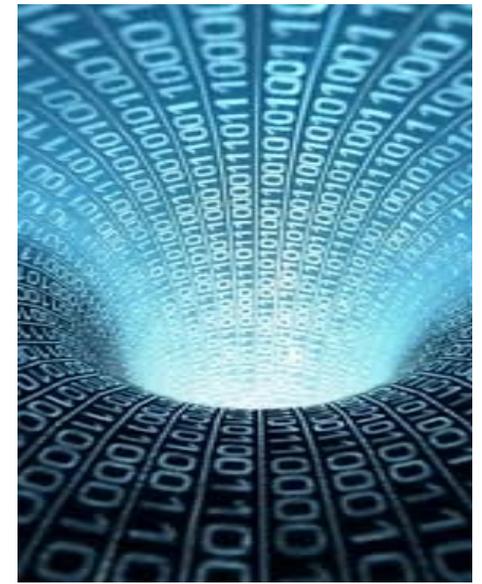
L'importance du support

- Pour stocker, organiser, partager et manipuler les données
- Histoire du support physique
 1. Mémoire humaine : rapide, peu fiable
 2. Argile, pierre : limitée, fiable
 3. Papyrus : léger, organisation séquentielle
 4. Parchemin : organisation en pages
 5. Papier : comme le parchemin, en mieux
- Aujourd'hui: le numérique
 - Indépendance des données par rapport au support physique
 - Les données peuvent être organisées, copiées, manipulées et communiquées facilement



Big data

- Très grandes collections de données
 - Teraoctets (10^{12} octets), Petaoctets (10^{15}), Exaoctets (10^{18})
- Provenant du Web
 - Pages web, moteurs de recherche, réseaux sociaux, etc.
- Mais aussi de toutes sortes d'appareils
 - Petits: objets connectés (IoT), capteurs
 - Grands: satellites, radiotélescopes, etc.
- Très utile dans un monde numérique
 - Information et connaissance à forte valeur ajoutée
- Mais difficiles à gérer
 - Volume, variété, vélocité, véracité, valeur, etc.



Machine Learning



- Un courant de l'Intelligence Artificielle
 - Dit "connexionniste"
 - Permet d'identifier des connexions sous formes de règles ou de patterns dans les données
 - Versus l'IA symbolique
 - Manipule des représentations symboliques (lisibles par l'homme) pour automatiser un processus de déduction à partir de règles logiques
- Principe
 - S'appuyer sur du big data pour prédire des résultats statistiquement significatifs de l'évolution future
- L'application "killer"
 - Reconnaissance d'images (ex. Google Images)

Science des données



- Définition

- Utilisation des statistiques, de la gestion de données, du machine learning et de la visualisation pour collecter, nettoyer, intégrer, analyser et visualiser du *big data*

- Objectif

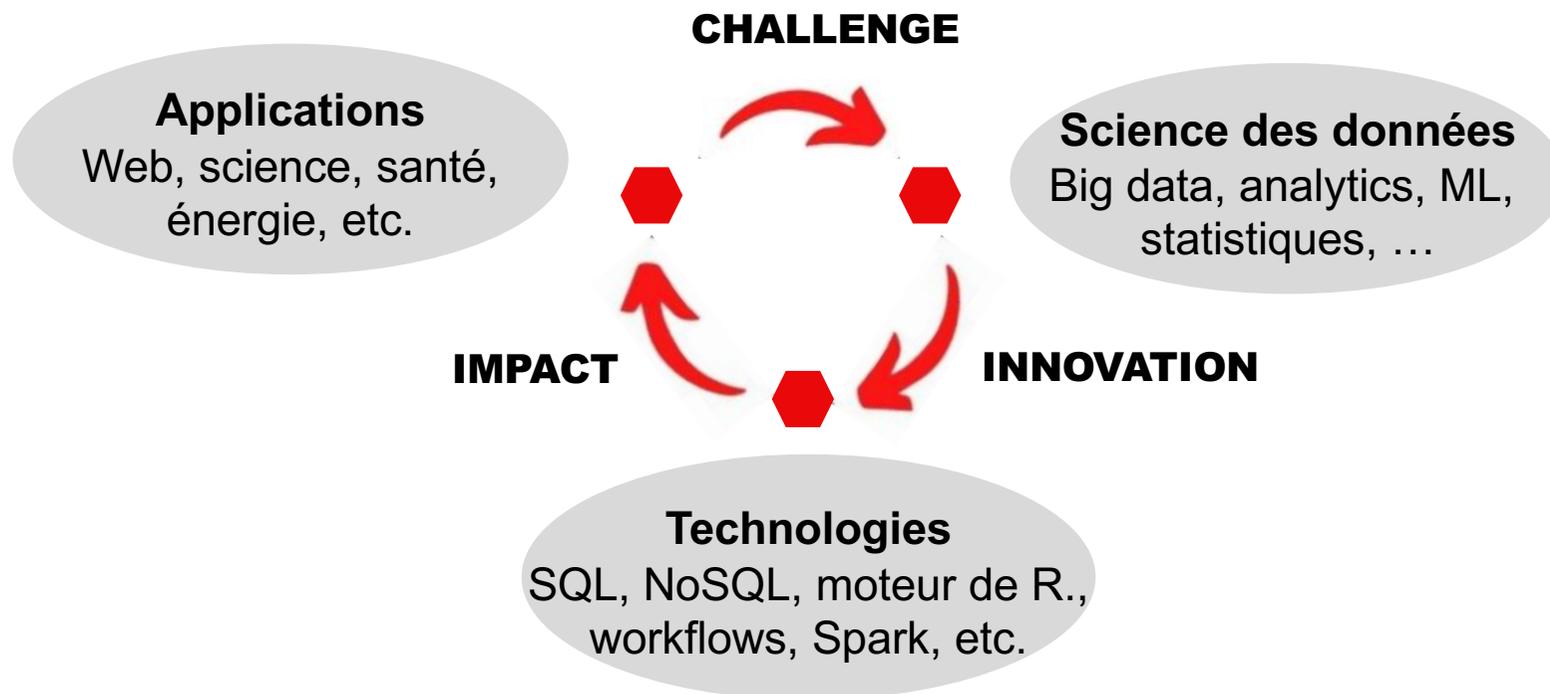
- Créer des produits et services de données
 - "Data is the new oil of the digital economy" (Wired, 2014)

- Data scientist

- Fortes compétences en statistiques, gestion de données et machine learning
- ET forte connaissance du domaine métier

Science des données et innovation

- Cercle vertueux qui nourrit l'innovation



Masses de données scientifiques



Données
d'observation

Données
d'expérimentation



Données du Web

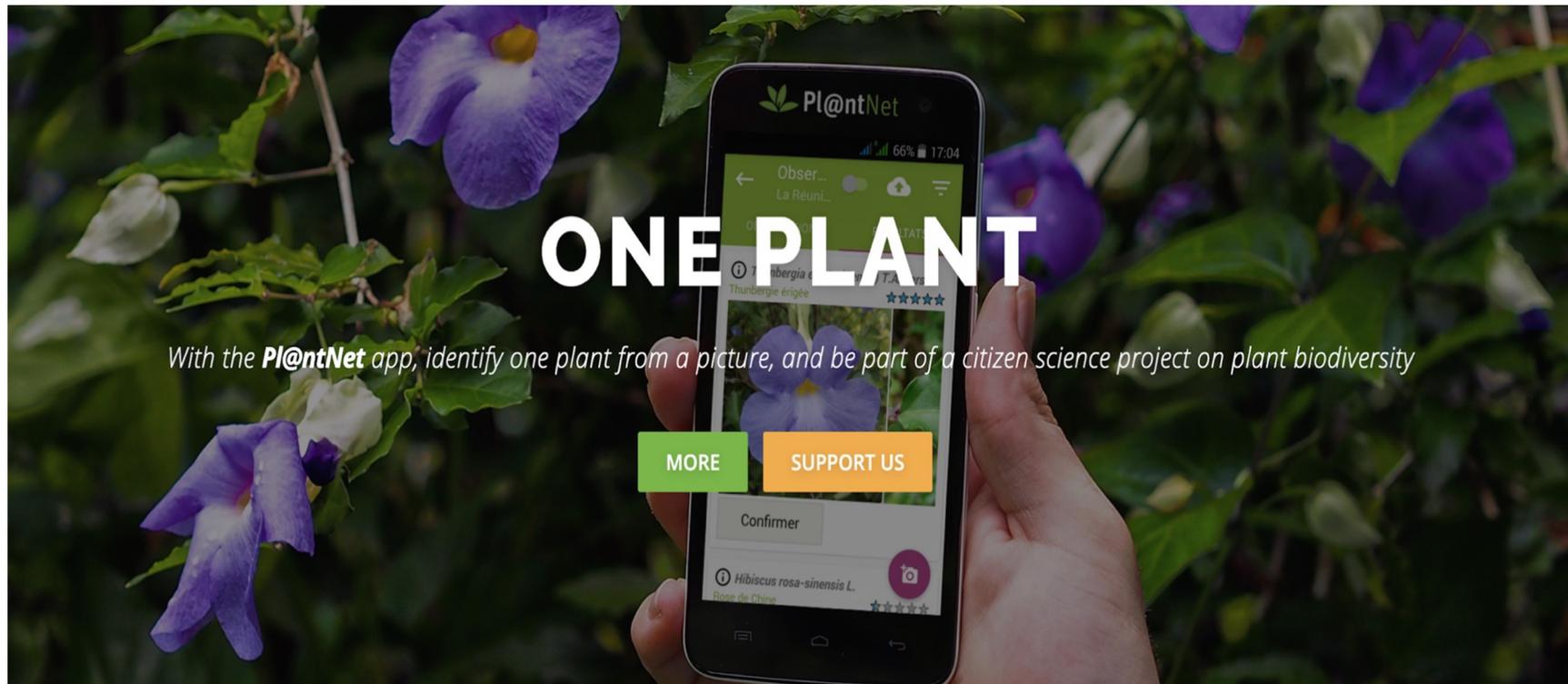


**Validation
Intégration
Analyse
Collaboration**

Information
Connaissance



Une plateforme de science citoyenne basée sur la science des données pour étudier la biodiversité végétale



25M utilisateurs
200+ pays
2M identifications par jour



Usage personnel



Nature



Jardinage



Phytothérapie

Usage professionnel



Agro-écologie



Gestion des ressources naturelles



Education, divertissement



Commerce

Apprentissage



Opportunités

- **Accès universel à l'information**
 - Données et algorithmes en accès libre
 - Open data, open source
 - Rappel: la nature humaine est d'échanger de l'information
- **Science des données et IA**
 - Progrès rapide des sciences expérimentales
 - Automatisation et assistance dans toutes les activités (humain-machine et machine-machine)
- **Economie de la donnée**
 - GAFAM
 - Réseaux bifaces: Uber, Airbnb, LeBonCoin, ...
 - Monnaies numériques: bitcoin, ethereum, ripple, ...

Questions



- Enjeux techniques, économiques, politiques, sociétaux
 - Cybersécurité
 - Comment faire? Combien ça coûte?
 - Consommation énergétique des data centers
 - L'objectif du 100% d'énergie renouvelable des GAFAM en 2030 est-il réaliste?
 - Souveraineté numérique des états
 - "L'information, c'est le pouvoir" ?
 - "Trop d'information tue l'information" ?
 - Impact sociétal de la science des données et de l'IA
 - Faut-il contrôler? Si oui, comment?
 - La vie dans le monde numérique
 - Comment éduquer?