BIG data

opportunités & stratégies

Raphaël Jolivet - février 2018

Sommaire

- **Introduction** : généralités
- Nouveaux **paradigmes** & nouveaux **métiers**
- **Stats / IA** / machine **learning**
- Utilisations pratiques
- Question de droit : **sécurité** & **vie privée**
- Stratégies

Introduction

BIG comment ?

- **1995 : 30 Gb**
- **1997 : 2 Tb (x60)**

Introduction

2014

	Bande passante / jour	**Stockage**
Facebook	600 Tb	300 Pb
Google	100.000 Tb	15,000 Pb
Twitter	100 Tb	?
NSA	29.000 Tb	10.000 Pb

1 Tb (Terabytes) = 1000 Gb
1 Pb (Petabytes) = 1000 Tb

Introduction

Evolution des usages

- **Réseau sociaux** : Crowd sourcing
- **Smartphones** / **IoT** : senseurs automatiques
- Massification d'internet :
  3.4 milliards (40 % vs 1% en 1995)

Introduction

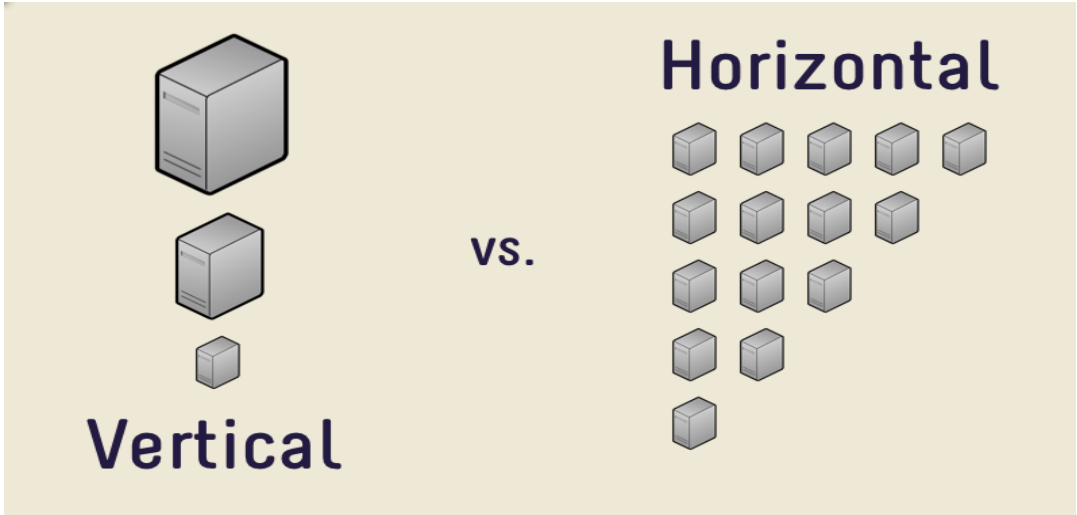
Evolution des technologies

- **Prix** du **stockage** de données
- **Centralisation** d'internet (GAFAM)
- **Virtualisation** des systèmes :
  Cloud // IaaS (infrastructure as a service)
- Technologies de **stockage** / **calcul distribué**
- **APIs** (Wikipedia, Google, Facebook)
- **Horizontal** scaling

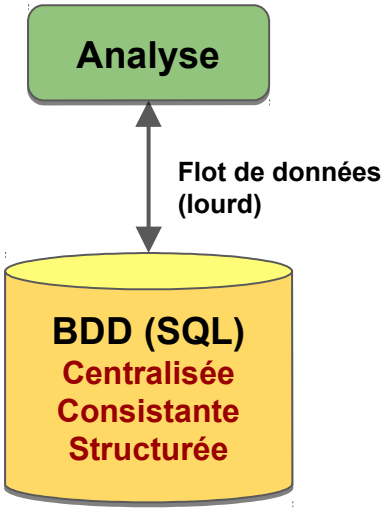
Nouveaux paradigmes : horizontal scaling

Scaling horizontal

- économique
- sans coupure
- sans limite de taille

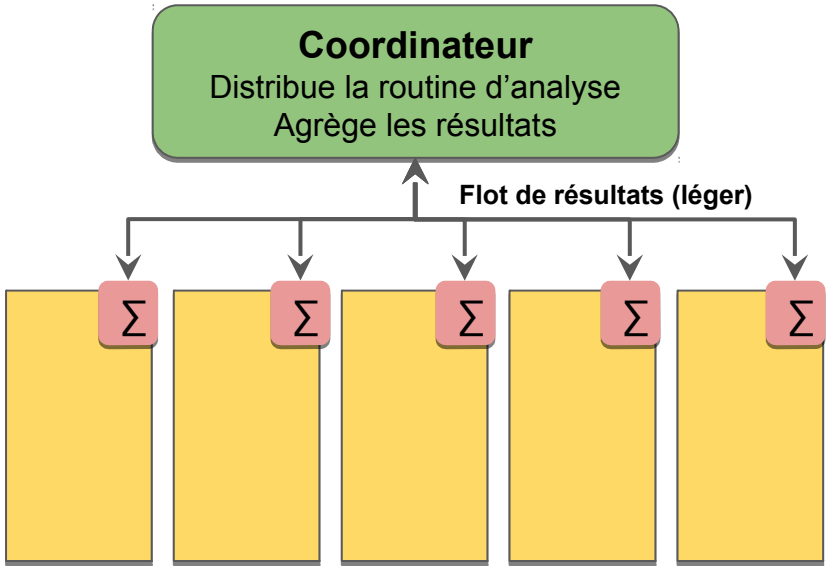


Nouveaux paradigmes : BDD



Centralisé

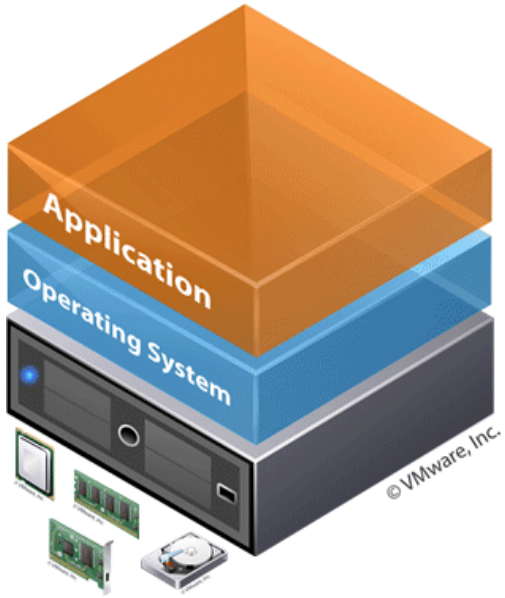
Distribué



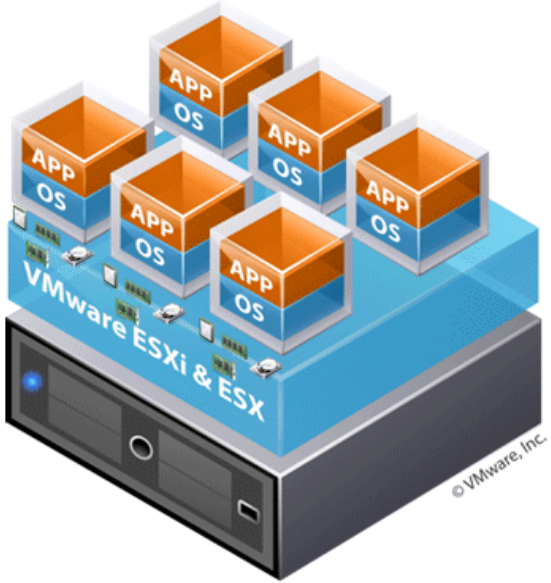
Données distribuées
Hétérogènes
Partielles (sparse)
Peu structurées (NoSQL)

Nouveaux paradigmes : Virtualisation

Sans virtualisation



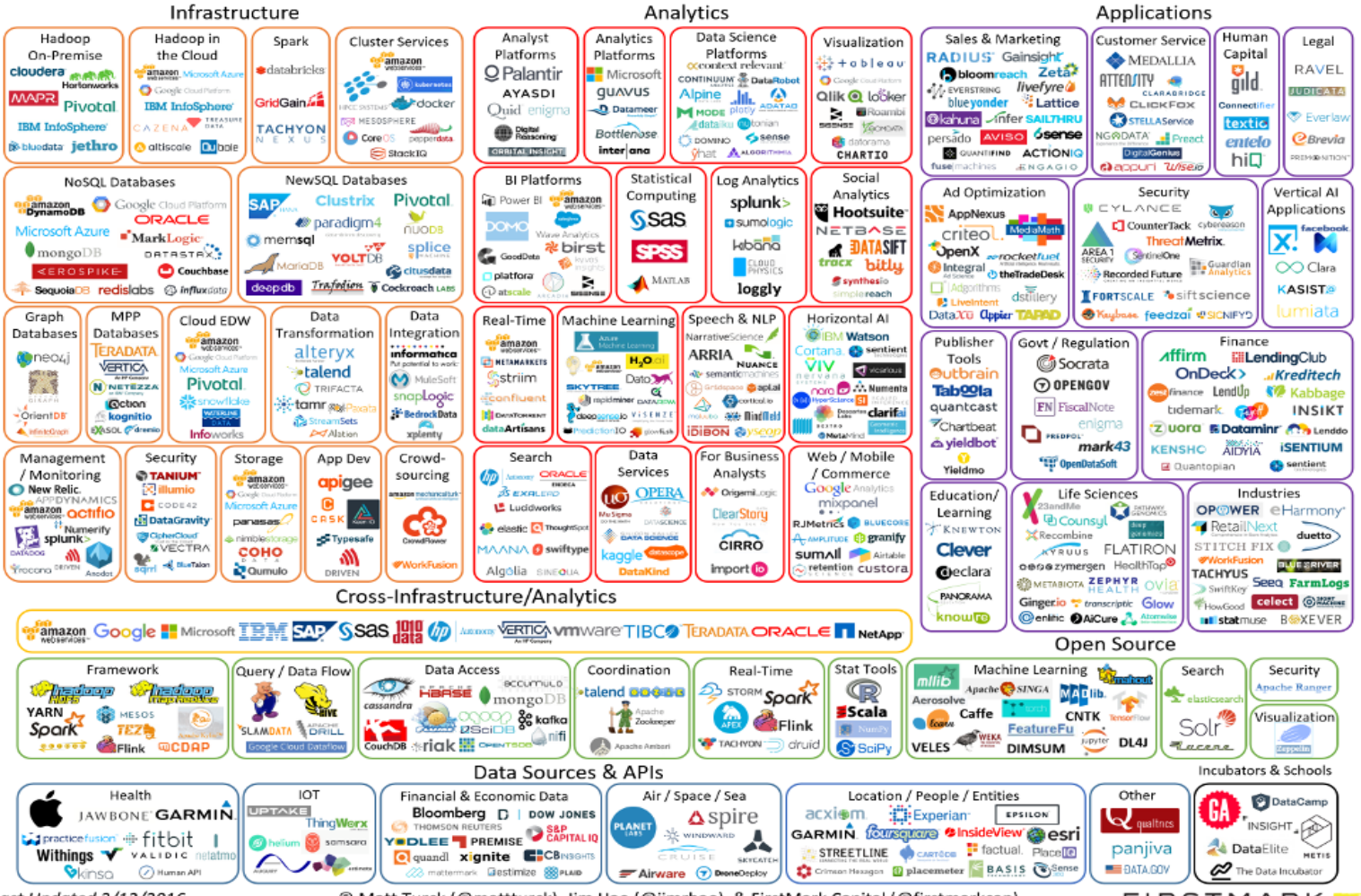
Avec virtualisation



Nouveaux paradigmes : Virtualisation

- **Isolation** des VMs ⇒ **Sécurité**
- **Décorrélation** physique / logique
- Adaptation à la **charge** (paiement **ressource** * **temps**)
- Provision rapide (qq **jours** ⇒ qq **minutes**)
- **Sauvegarde**, **clonage**, **migration** facile
- **Automatisation** de l'administration : **IaaS**

Big Data Landscape 2016 (Version 2.0)



Last Updated 2/12/2016

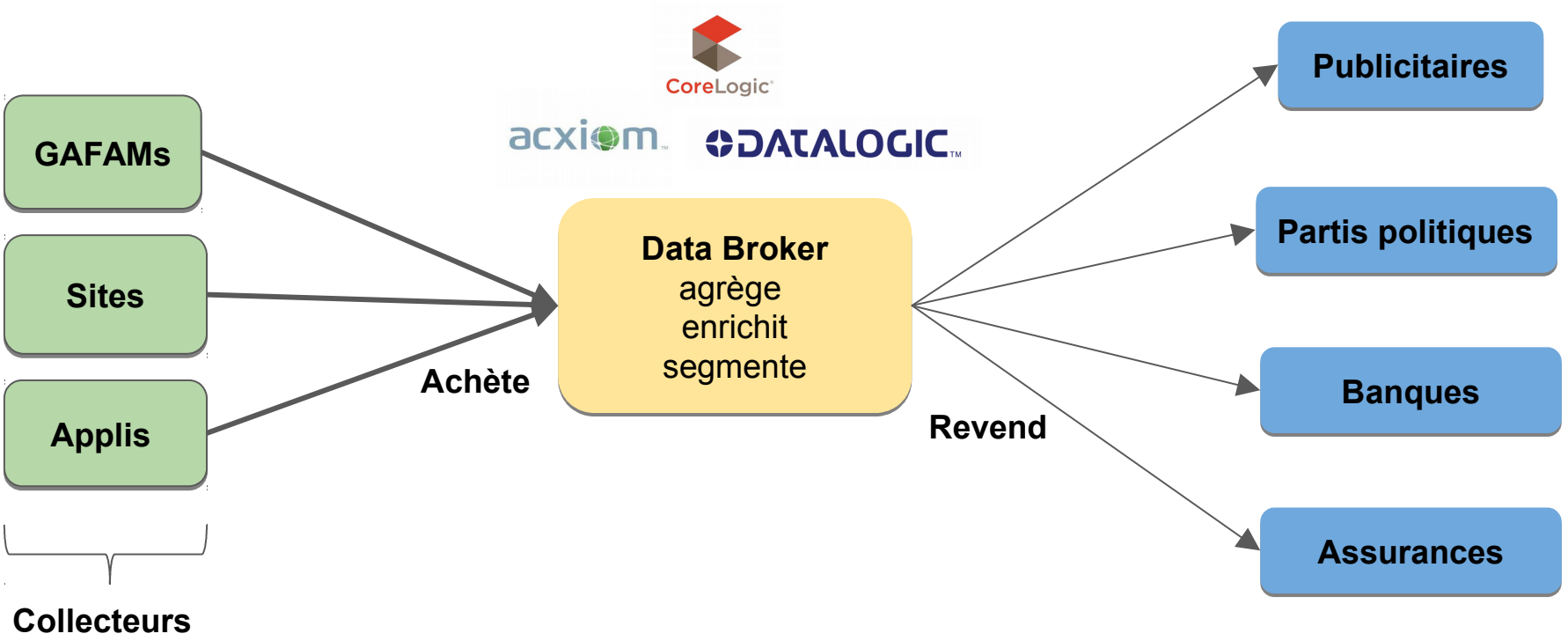
FIRSTMARK

Nouveaux acteurs

- **Infrastructure** / **Cloud** : *Google, Amazon, Microsoft, OVH*
- **BDD** distribuées, **NoSQL** : *CouchDB, MongoDB, Hive*
- **Framework** : *Hadoop, Apache Spark*
- **Ecosystème fourni** d'outils et **d'applications**
- **Intégration** complexe
- Domaine **évoluant** très **vite**

⇒ **Distributions** : *Cloudera*, *Mapr*

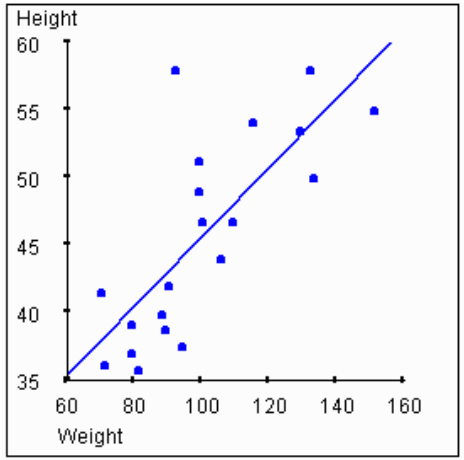
Nouveaux acteurs : Les Data Brokers



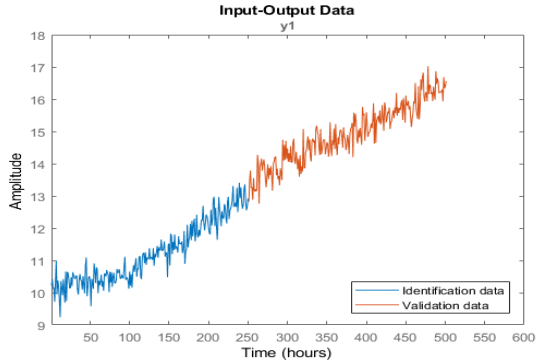
Nouveaux métiers

- Nouveaux **langages : Scala** (calcul parallèle, fonctionnel)
- **Spécialisations** en informatique
  - Data **analyst :** maths, stats
  - **Architecte** de **données**
  - IA / **Machine learning :**
    “**psychologues**” de réseaux de neurones.

Big data & Statistiques

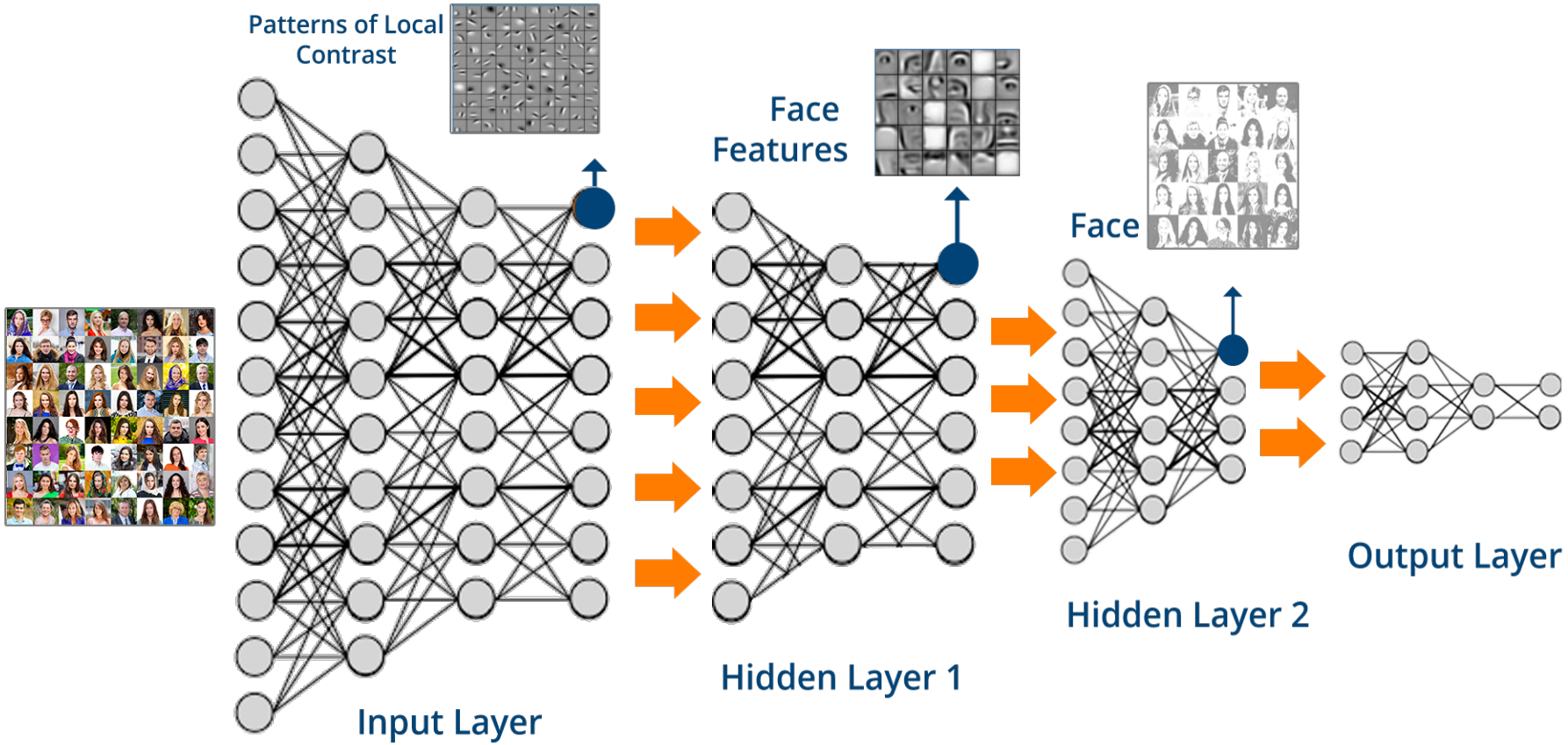


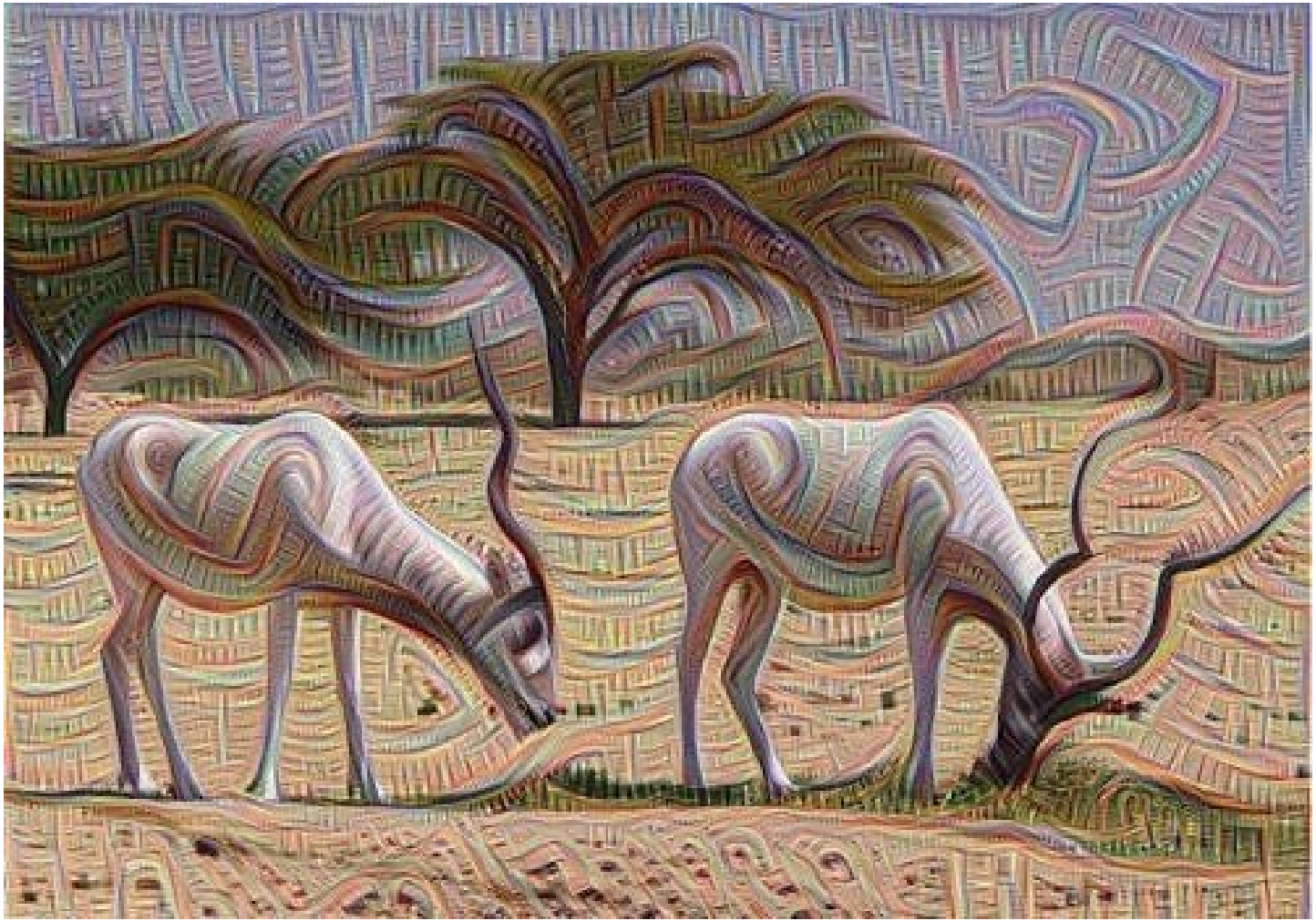
- **Analyse**, métriques
- **Prédictions**
- **Corrélations**, régression ⇒ Fonction
- **Segmentation**
- **Périodicité**, prévision **temporelle**



Big data : Deep learning

- **Machine** learning :
  apprentissage sans **programmation explicite**
- Le programme **s'améliore** avec **l'expérience**
- Deep learning = **imitation** du **cerveau**
- Réseau de **neurones artificiels**
- **Essai / erreur ⇒ renforcement**


Patterns of Local Contrast
Face Features
Face
Input Layer
Hidden Layer 1
Hidden Layer 2
Output Layer



Big data : Deep learning

- Apprentissage : nécessite **énormément** de **données**
- **BIG DATA** ⇒ progrès **fulgurant** des **IA :**
  - **Classification** d'images
  - Reconnaissance **vocale**
  - Reconnaissance **faciale**
  - **Conduite** automatique
  - **Traduction** automatique

Utilisations

- **Trafic** temps réel / bouchon (**Waze**)
  Smart cities ⇒ optimisation des flux :
  Circulation: lignes de bus.
- Smart **GRIDS** : réseau électrique:
  - Prédiction / pilotage de la consommation (**Linky**)
  - Prédiction de la production (**ENRs**)

⇒ **Pilotage** de stockage & **vente** sur le marché D+1

Utilisations

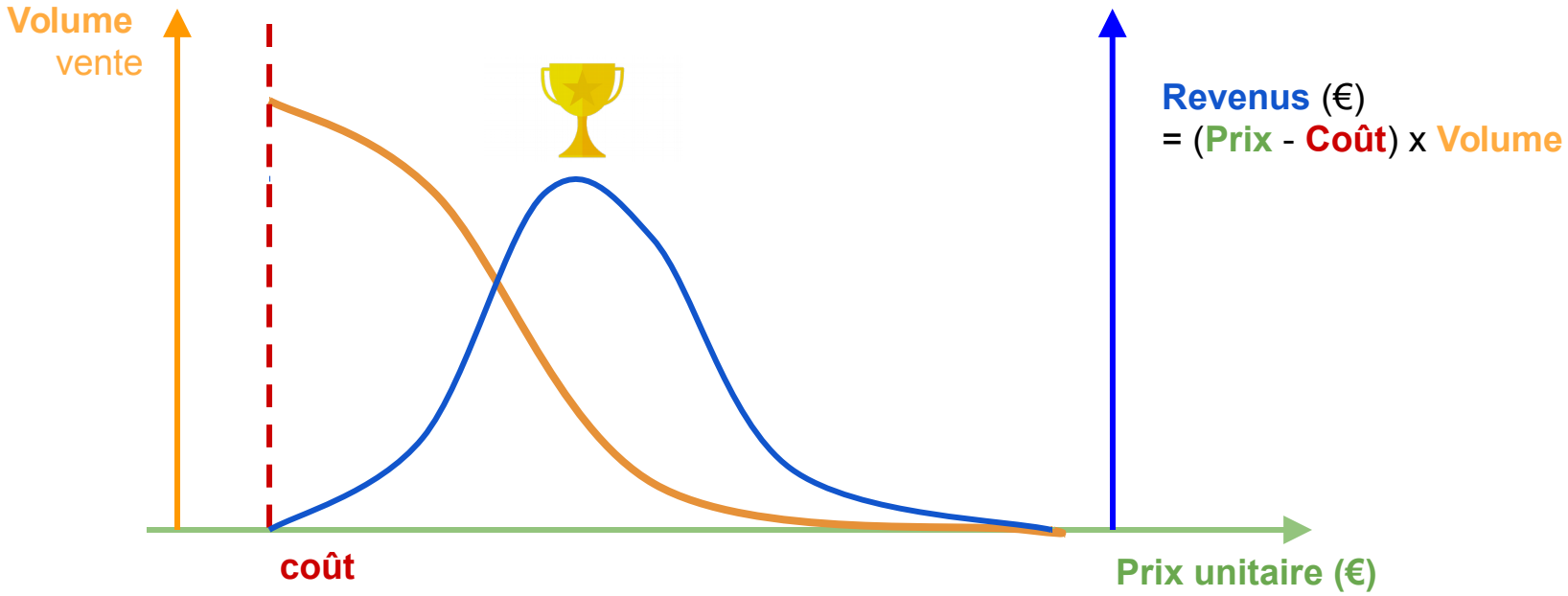
- Détection **épidémies** (google trends) / **catastrophes** (Twitter)
- Gestion des **communs** : Réseau **hydraulique** (fuites, suivi perso, quotas, pénurie ...)
- **Médecine** : aide au diagnostique
- **Droit** : aide aux **juristes**, recherche de jurisprudence

Utilisations : marketing

- Vers une **personnalisation** des offres
- **Fidéliser** le client : promos **ciblées,** rappels
- **Provoquer** l'achat : "Vous **aimerez aussi**"
- **Prédiction** de la **demande** :
  Gestion des **stocks**, demande **périodique** (WE, vacances), dépendant de la **météo** (Cinema, Restaurant)
- Etudes de **tendances** du **marché** :
  Ouvertures de **lignes aériennes**, création de **produits**

Utilisations : A/B testing

- La **courbe** de **prix** : Le **Graal** du commercial



Utilisations : A/B testing

- Variation aléatoire du prix (ou % promo, pub, design, ..) sur un échantillon de visiteurs (qq %)
- Impact **négligeable** sur les ventes **pendant la campagne**
- Puis **statistiques** ⇒ **courbe** de **prix**
- **Bonus** : segmentarisée !

⇒ **Action** : changement du prix, voire **variable** par **segment**

⇒ **Maximisation** des **revenus**

Question de droit : sécurité & vie privée

- Les données critiques :
  - Données **client** ⇒ Réputation en cas de vol
  - **Secret industriel** :
    Innovation ? ⚠ **NSA / GAFAM**
  - Secret **défense** / **légal** / **admin FR** ⇒ **Sol Francais**
- Droit, espionnage ⇒ préférer **UE** / **France** (OVH)

Question de droit : données personnelles

- Données personnelles : **définition** (FR / UE)
  - Identifiants uniques (nom, tel, sécurité sociales, ..)
  - **ou** identification par **recoupement**
- **Obligations** :
  - **FR** : Déclaration **CNIL +** accès, modif, suppression
  - + UE [**RGPD**] Mai **2018 :**
    - Collecte limitée, explicite, légitime ⇔ finalités

      Cf **droits applis** sous **android**
    - Expiration

Stratégies : Open Source

- **Open Source** : Logiciel avec **accès** au **code source** + droit de **modification** et **redistribution**
- Big DATA : Beaucoup d'outils **open source** :
  - Suite Apache Spark
  - Apache **Impala**
  - Elasticsearch
- **Gratuits**, **puissants**, **maintenus** + **communauté**
- Open Source != bénévole :
  - **support**, **formation**, **consulting**

Stratégies : résumé

	Outils ⇒ Données	Données ⇒ Cloud
Vie privée	• Déclaration **CNIL**	• Déclaration **CNIL** • + **Anonymiser** si possible • + Préférer **Cloud privé** (+ cher)
Sécurité	⚠ **Politique** stricte de sécu interne **compétences** interne	Souvent bonne politique sécurité ⚠ **Espionnage / NSA**
Outils d'analyse	**Open source** **Installation** spécifique Adaptation au business	• Open source • ou **logiciel "maison"** ⇒ Pas de communauté
Coûts / travail	• Consulting • **ou** compétences internes	• **Licence** / **abonnement** Cloud • **Préparation** des données • **Transfert** • **Config**

À emporter

- Secteur porteur : **marchés** & **métiers**
- Technologies **jeunes** / **mouvantes**
- Données = mines d’or
- **Analyse** = **extraction** de **valeur**
- Gardez-en le contrôle !

Questions ?

PUB

ATELIER DES COMMUNS

Des artisans développeurs, au service du bien commun.

http://atelier-des-communs.fr/
raphael@atelier-des-communs.fr