Affini-TECH

BIG DATA et DONNéES SEO

Vincent Heuschling
vhe@affini-tech.com
@vhe74

Agenda

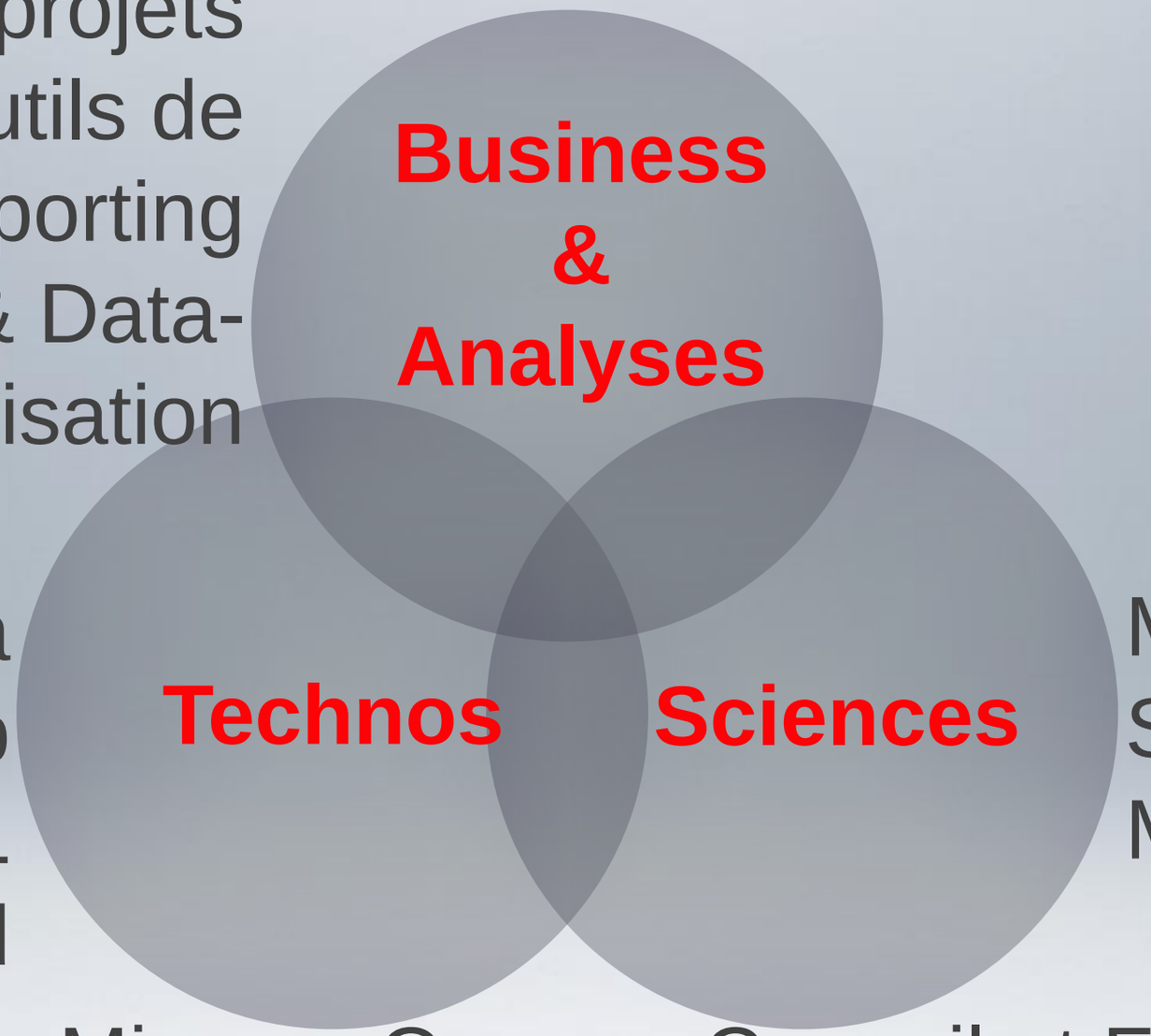
- Affini-Tech
- SEO ?
- Application
- Généralisation

Affini-TECH

Société

3 Piliers

Méthodes projets
Outils de reporting
& Data-visualisation



BigData
Hadoop
NoSQL
Cloud

Modélisation
Statistiques (R)
Machine Learning

Intégration, Mise en Oeuvre, Conseil et Formation

Une démarche intégrée de bout en bout

Affini-TECH

Collecter
Stocker
Traiter

Analyser
Valoriser

Présenter
Organiser

BigData

Data-Science

Data-Viz

Votre infrastructure

Notre Cloud

Affini-TECH

Partenaires sectoriels

Mktg & Ventes	**Finance**	**Production**
Stats	**Apps**	**Data-Viz**

Infrastructures

Partenaires technologiques

Affini-TECH

Métiers

Applications

Infrastructures

Opportunité Bigdata

Data-Mining

Applications

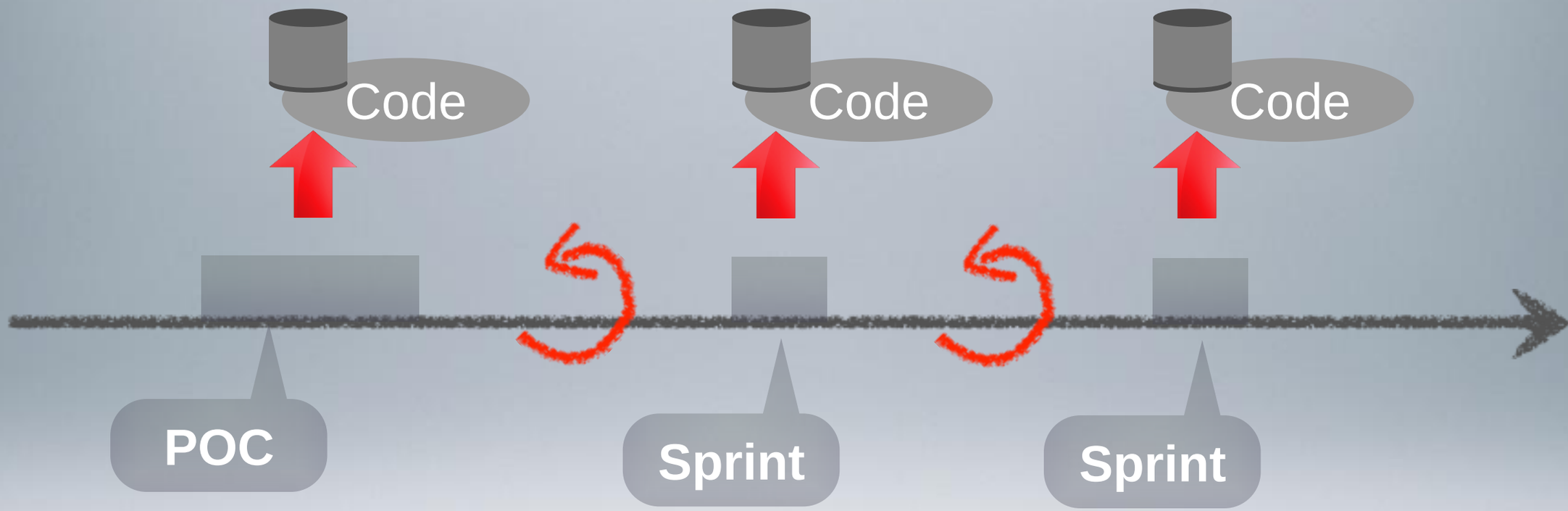
Data-visualisations

Hybridation

Infrastructures

Affini-TECH

Agile Data



Affini-TECH

COLLECTER | STOCKER | ANALYSER | PARTAGER

fluentd

hadoop

R

D3.js

APACHE HBASE

MAPR TECHNOLOGIES
EASY. DEPENDABLE. FAST.

Ruby
A Programmer's Best Friend

mongoDB
{name: "mongo", type:"DB"}

redis

mahout

Gephi

syncsort

elasticsearch.

python™

+ableau SOFTWARE

Google

Windows Azure

Affini-TECH

Affini-TECH

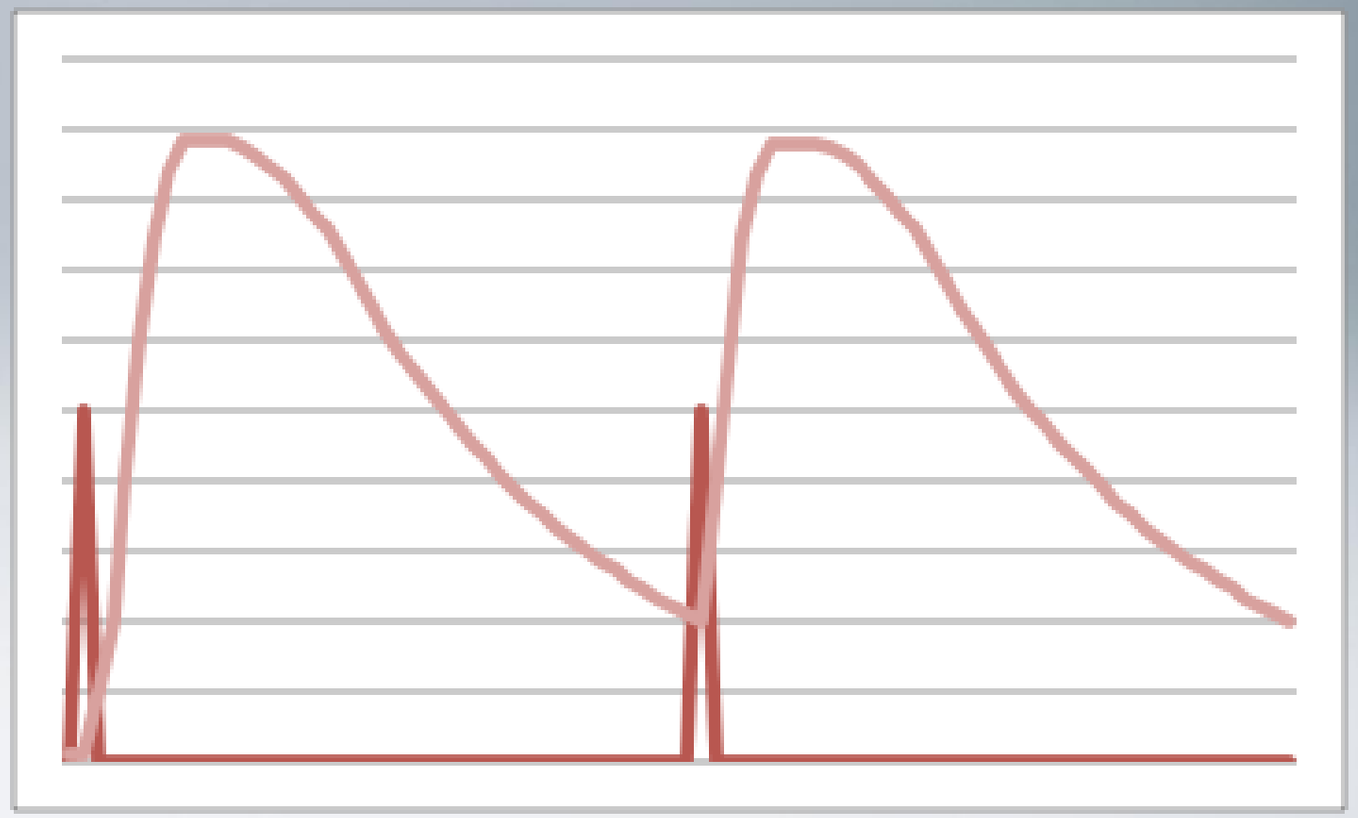
SEO ?

Obtenir les meilleures positions dans la page de réponse de Google.

- Définir quels éléments du site sont à forte valeur
- Les promouvoir vers les moteurs de recherches (linking, etc...)
- Mesurer et étudier le positionnement du site sur des recherches vis à vis de sa concurrence

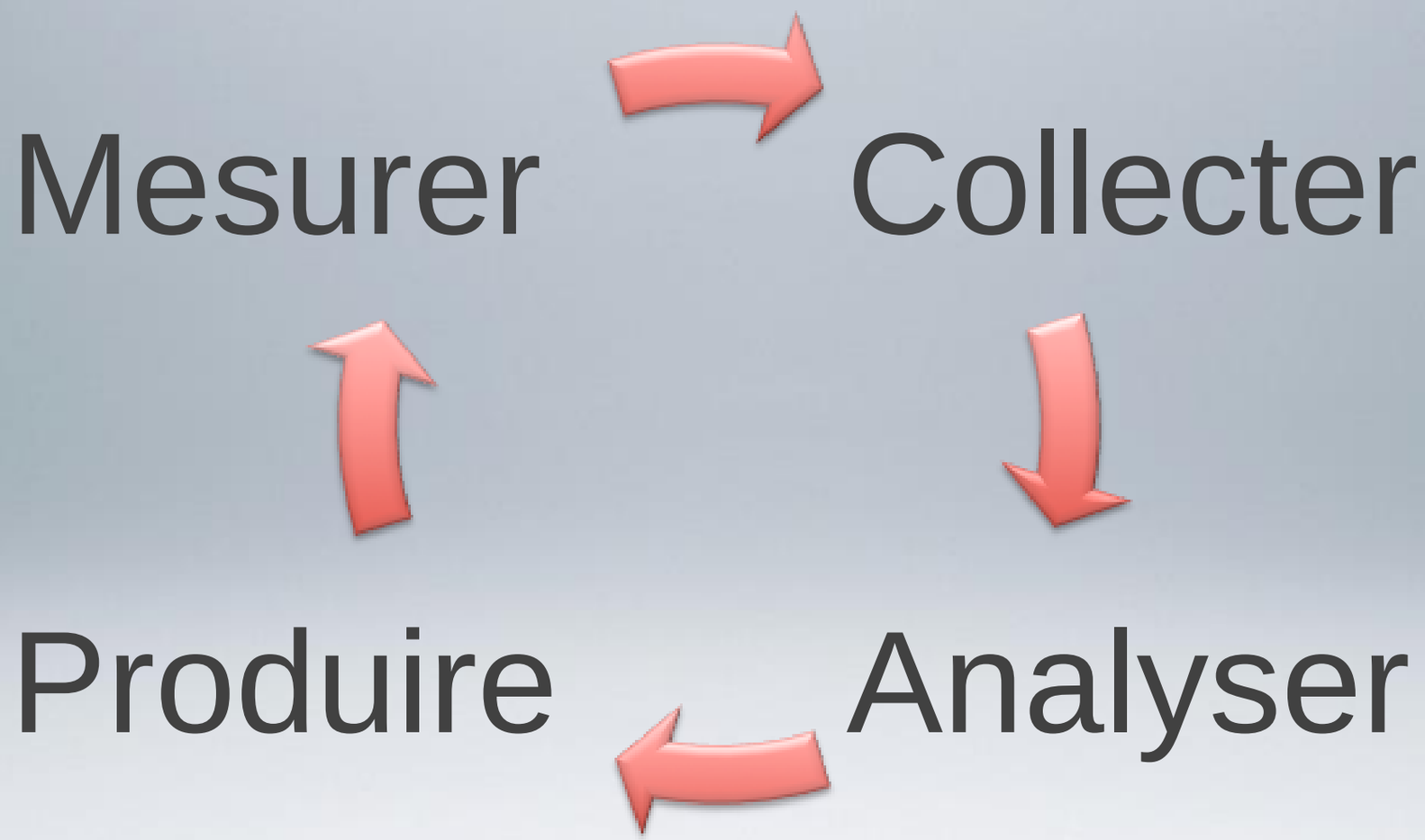
CRAWL et VISITES

- Organiser le contenu des pages (Pagerank)
- Faire Crawler les pages par Google
- Augmentation directe du trafic



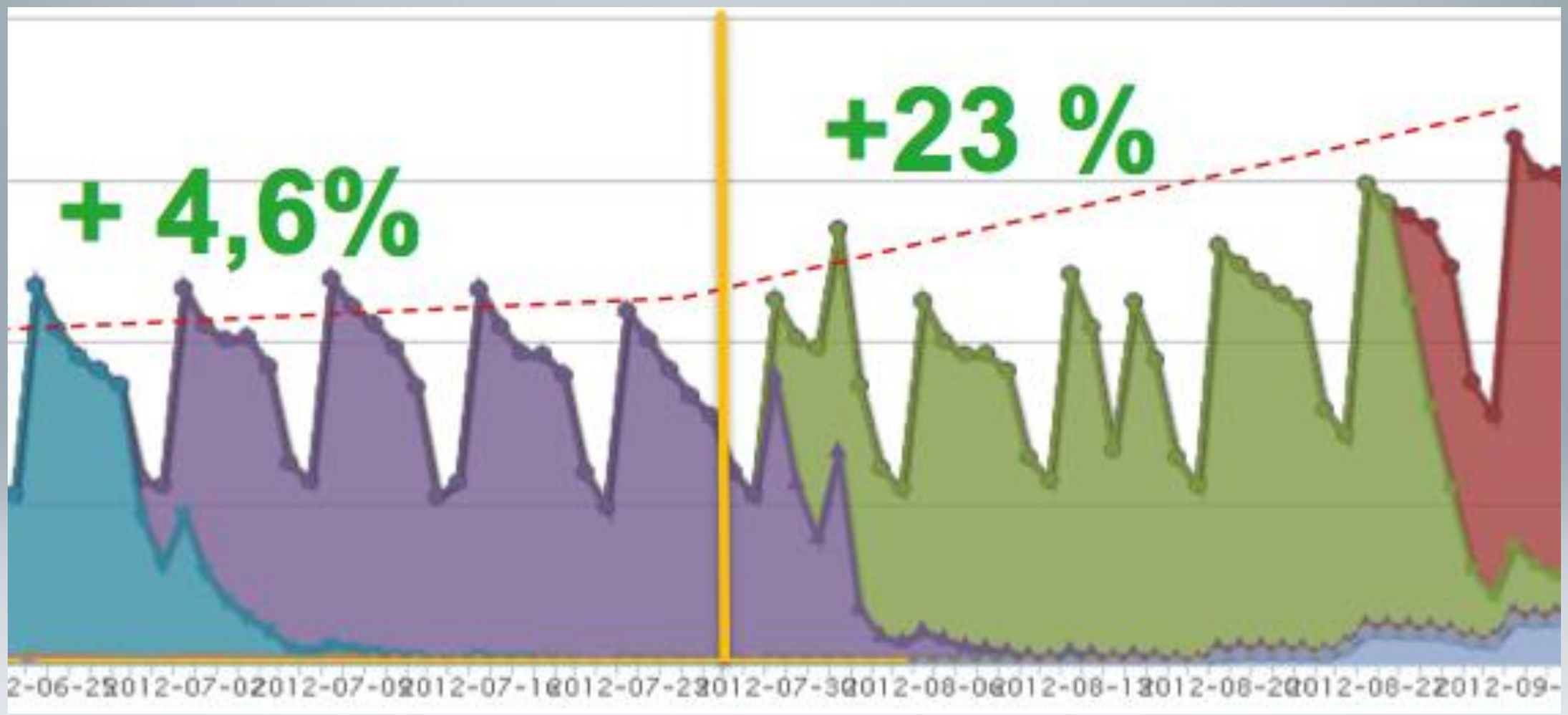
Affini-TECH

Cercle VERTUEUX de la DATA



Affini-TECH

RésUltats



Affini-TECH

Affini-TECH

Application

ANNUAIRE

- 2000 Professions
- 40000 Communes
- 100 M de requêtes par mois
- Small data : SEO = env 100 Go /an

Combien ?

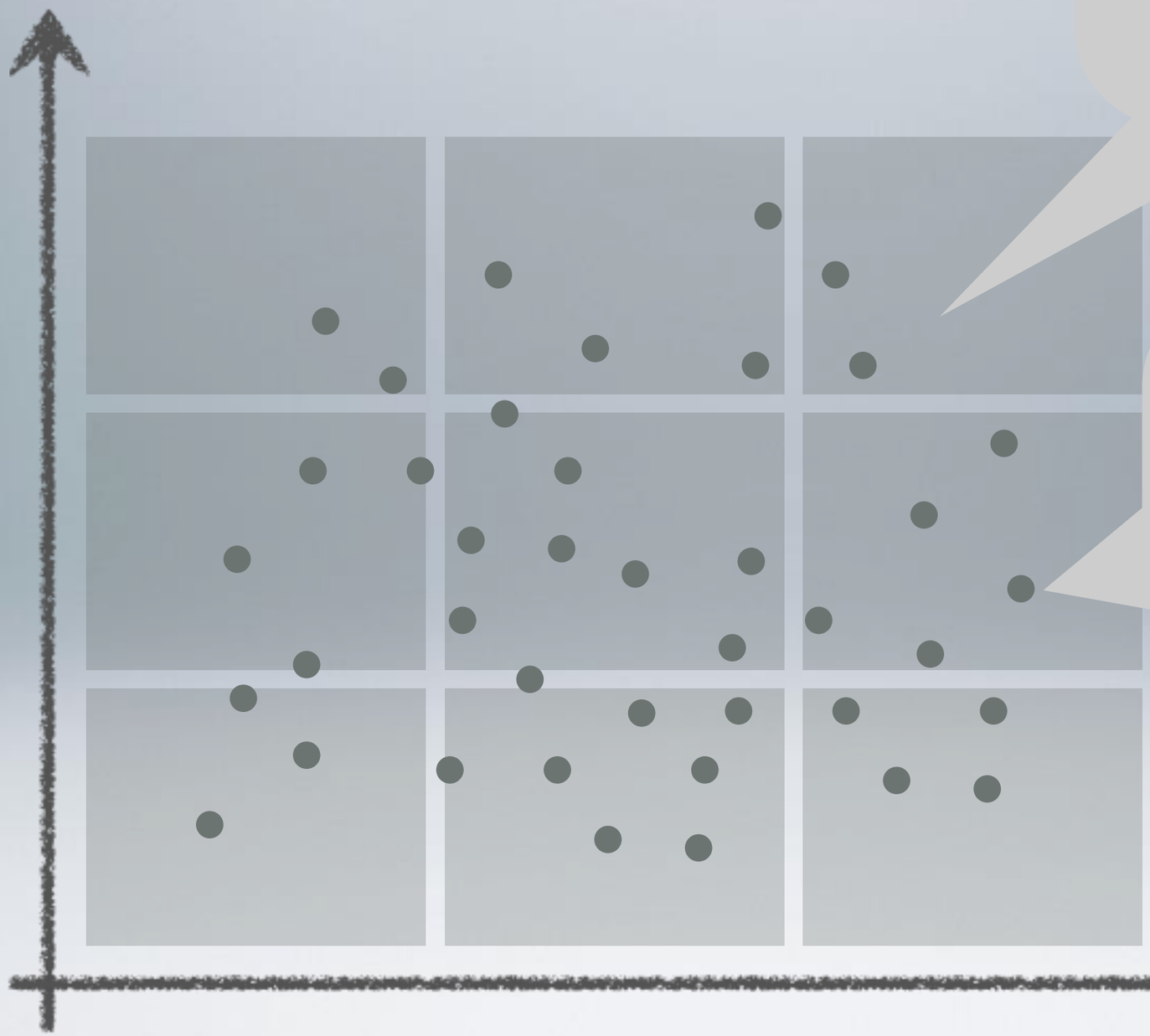
- 10 visites SEO (hors marque)
- 30 visites SEO (marque)
- 90 visites non SEO
- 20 crawl
- x7 à x10 au total (pages + ressources)
- Nécessité de filtrer à la source

TROUVER 400K NOUVELLES URLS À PROMOUVOIR PARMI 84M ?

- Similarités et Classifications
- Recommandation & intelligence collective
- OpenData

Affini-TECH

Professions



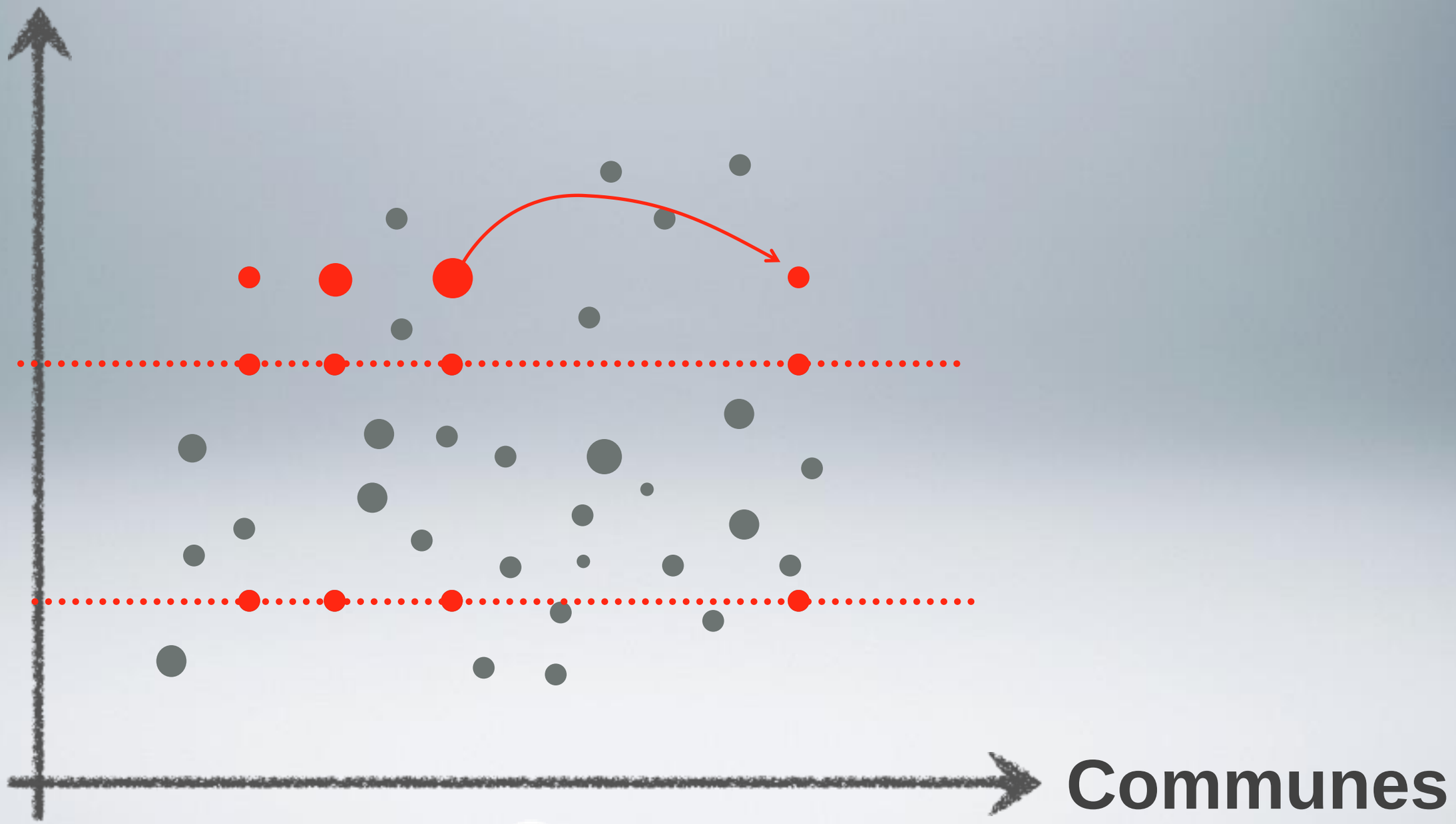
Grandes catégories, segments et moyennes

Analyse à la granularité la plus fine

Communes

RECOMMANDATIONS

Professions



AUGMENTER LA DONNÉE

Data

a	b	c
~~	~~~	~~~
~~	~~~	~~~
~~	~~~	~~~
~~	~~~	~~~
~~	~~~	~~~
~~	~~~	~~~

+ Insee

a	b	c	m	n	n
~~	~~~	~~~	~	~	~
~~	~~~	~~~	~	~	~
~~	~~~	~~~	~	~	~
~~	~~~	~~~	~	~	~
~~	~~~	~~~	~	~	~
~~	~~~	~~~	~	~	~

+ Opendata

a	b	c	m	n	n	x	y	z
~~	~~~	~~~	~	~	~	~	~	~
~~	~~~	~~~	~	~	~	~	~	~
~~	~~~	~~~	~	~	~	~	~	~
~~	~~~	~~~	~	~	~	~	~	~
~~	~~~	~~~	~	~	~	~	~	~
~~	~~~	~~~	~	~	~	~	~	~

- Une information plus riche
- Un ciblage plus pertinent

Affini-TECH

OUTILS

Collecter

Traiter

Analyser

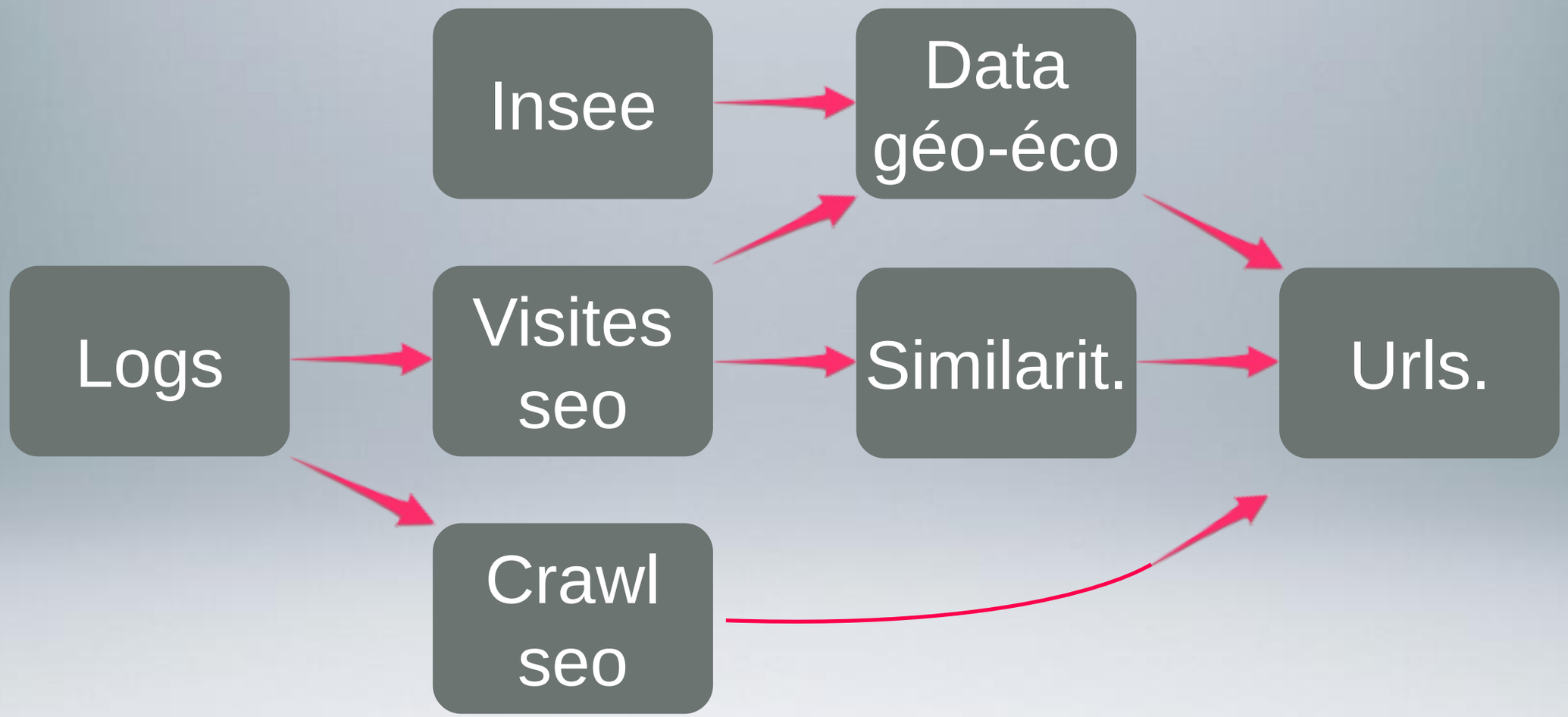
Pig

Affini-TECH

TYPES de REQUETES

- Analyse par zones de géographique
- Données socio-économiques
- Recherches de similarités
- Analyse au niveau Url (granularité fine)
- Impact du Crawl sur les visites

Pipeline



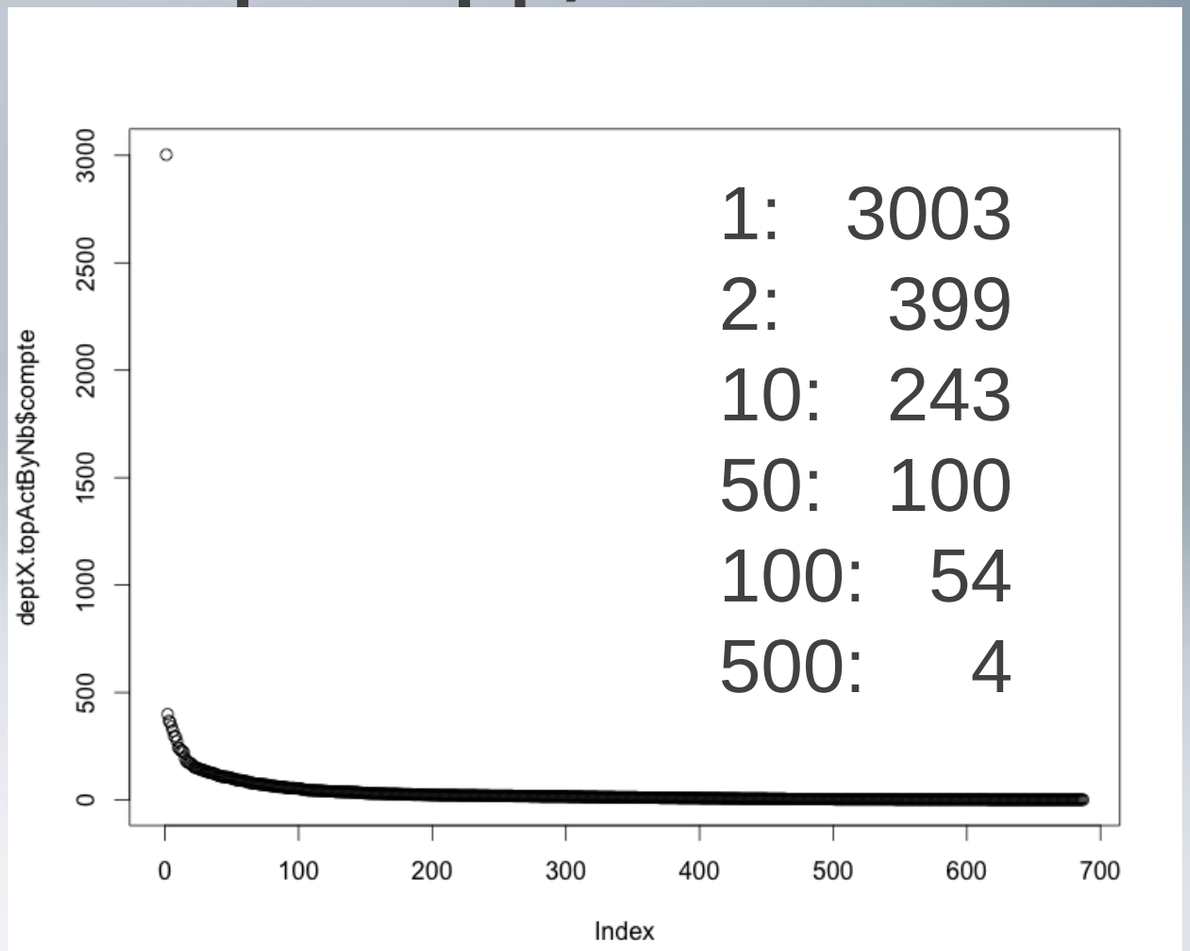
Affini-TECH

Long TAIL

- Très grosses concentrations d'activités sur certai

  - Map/Reduce !!

  - PIG Skewed joins



Affini-TECH

HEATMAPS

Départements (96)

			2			11	1	
			9	32		9	21	
			1	3		1		
			48	9			42	
150	27	61	29	77	15	67	73	58
1352	782	1199	757	2733	244	2255	2016	2187
45	35	115	274	100	9	36	105	112
		50	8			117	22	17
13	20	21		14		158	46	9
			8	5		2	7	
			2	2		6		
							1	1

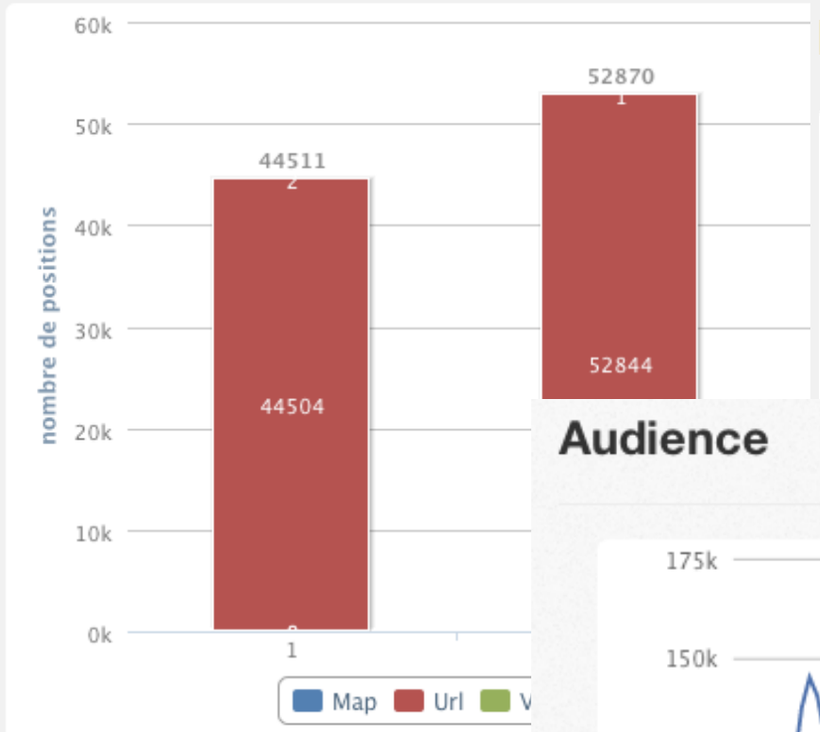
Activités (10500)

Volume de visites

MEsurer : Rankings

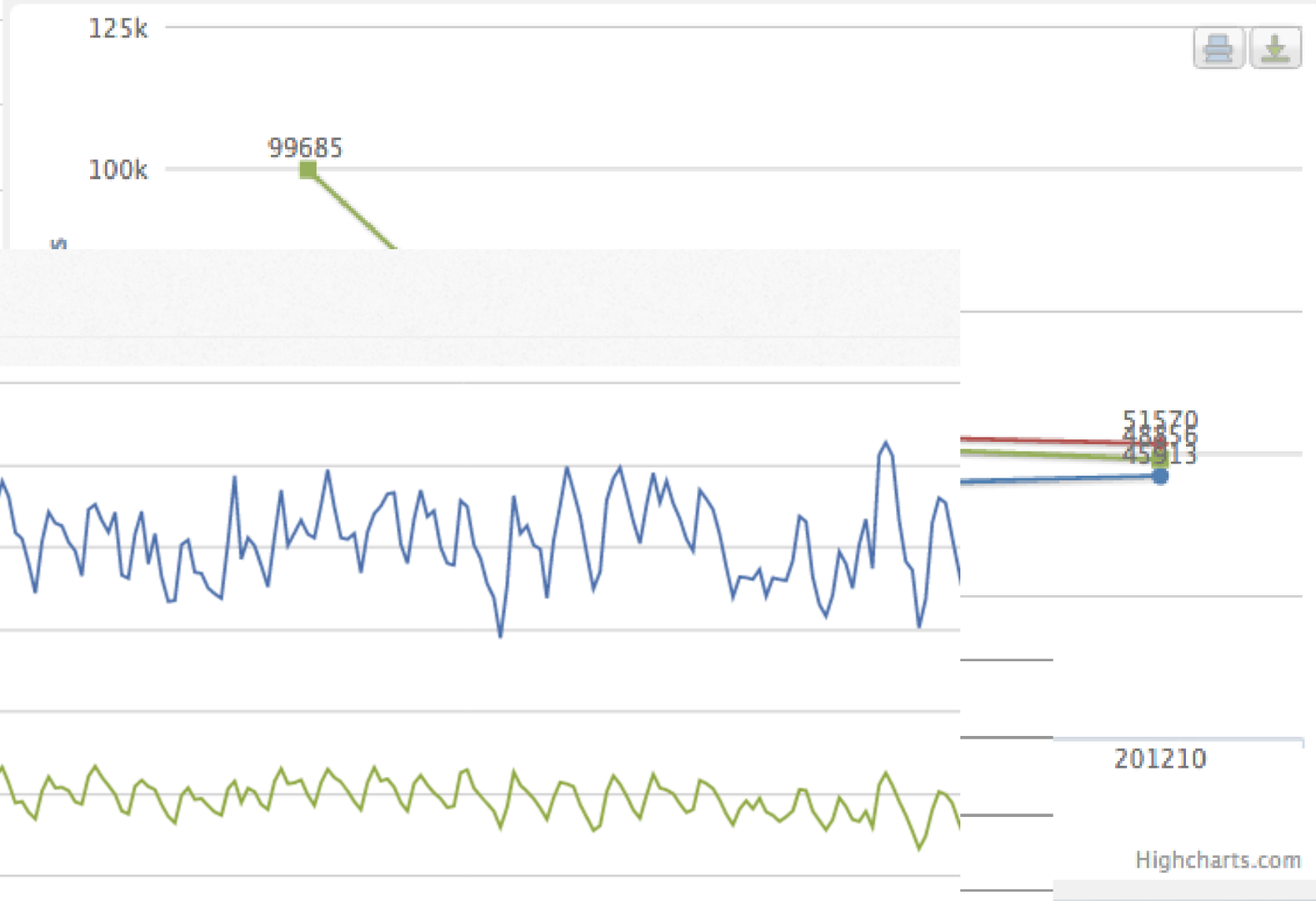
- Collecte des réponses Google (30x par recherche = dizaines de millions par mois)
- Forte croissance de la volumétrie
- Classifier et Segmenter par produit, par thématiques.

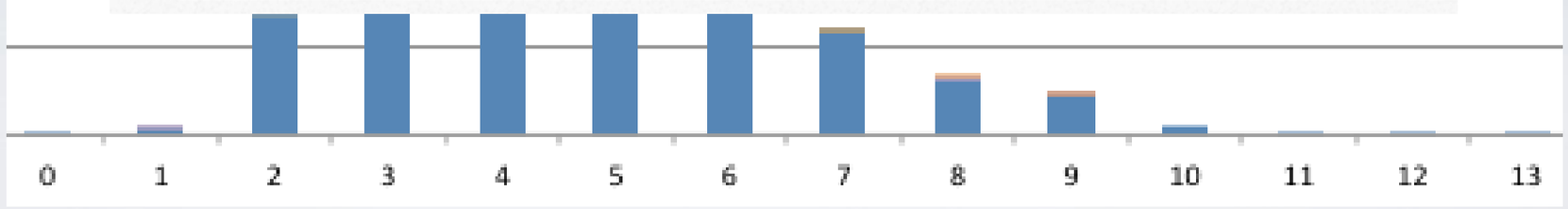
distribution des positions en pages 1



	1
Map	2
Url	4450
Video	5
Image	0
Adresse	0

Evolution du nombre de positions par classe dans le temps :





OUTILS

Collecter

Traiter

Stocker

Visualiser

Pig

D3.js

Analyser

Affini-TECH

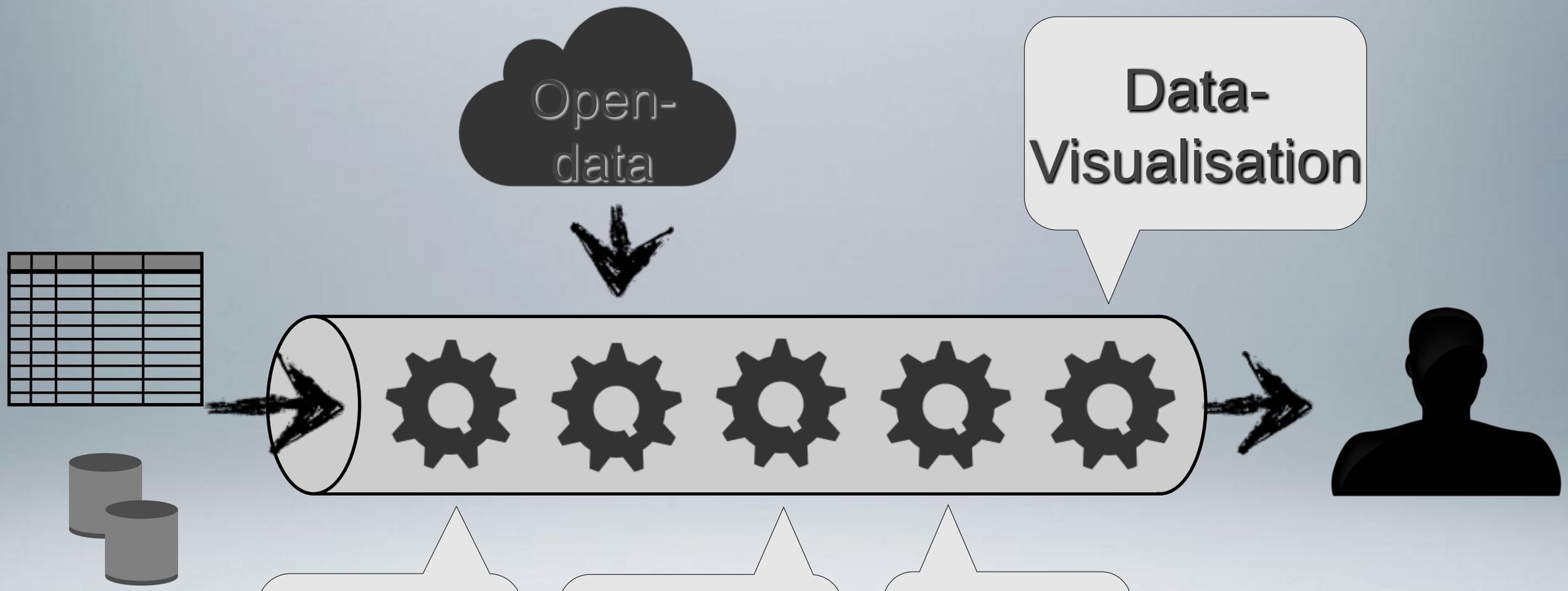
Affini-TECH

GENERALISATION

- **Applicable à toute transaction**
- **Des données brutes**
- **Augmenter la donnée**
- **Similarités et Classifications**
- **Recommandations**

Affini-TECH

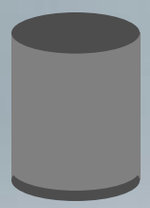
DAta-PIPELINE



Affini-TECH

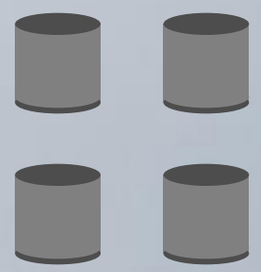
hadoop : ETL & DW

Non-Structuré

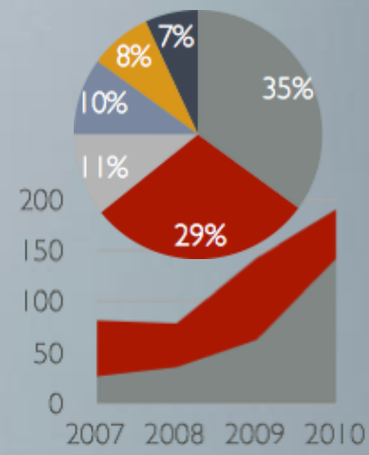


Transactionnel

ETL & DW



DataMarts

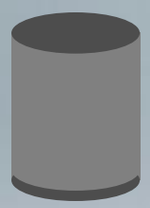


BI Applications

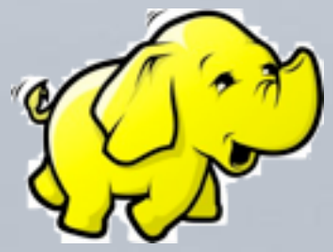
Affini-TECH

hadoop : EDW

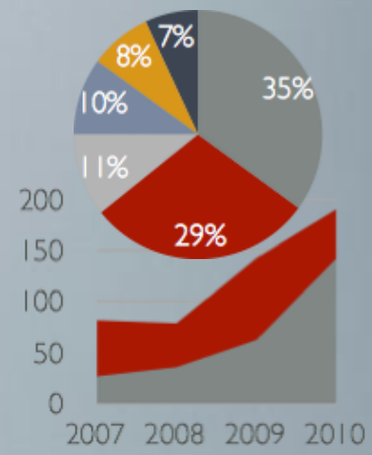
Non-Structuré



Transactionnel



ETL & DW & DataMarts

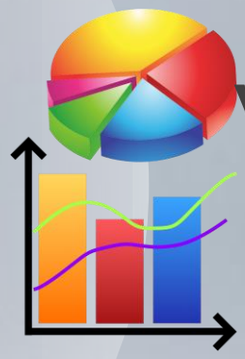


BI Applications

Affini-TECH

Applications
& Machine Learning

Opendata



Visualisations
Tableau & JS

Plateformes

- **"With data collection, 'the sooner the better' is always the best answer"**
- **Marissa Mayer, Yahoo CEO**

Merci !

- Vincent Heuschling
- Gsm : 06 61 88 76 71
- Email : vhe@affini-tech.com
- Web : http://www.affini-tech.com
- Twitter : @affinitech & @vhe74