R, VISUALISATION ET APPRENTISSAGE

ANALYSE DE COMPORTEMENTS TOURISTIQUES À PARTIR DE DONNÉES PHOTOGRAPHIQUES GÉOTAGGÉES

4 avril 2014, Chatou

B.Branchet – G.Chareyron – J.Da-Rugna

ESiLV LÉONARD DE VINCI

EIREST
Équipe Interdisciplinaire de REcherches Sur le Tourisme
– PANTHÉON - SORBONNE – UNIVERSITÉ PARIS 1

LMCS 2014

PRÉSENTATION

- Bérengère Branchet
- Gaël Chareyron
  - Enseignants-Chercheurs en informatique
  - Big Data, Open Data, Data Mining, Traitement d'images, Visualisation
  - Le monde ☺

ESiLV
LÉONARD DE VINCI

EIREST
Équipe Interdisciplinaire de REcherches Sur le Tourisme
UNIVERSITÉ PARIS 1 – PANTHÉON - SORBONNE –

PLAN

- Problématique
- Sources de données
- Visualisation des données
- Analyse des données
- Conclusion

PROBLÉMATIQUE

Utiliser les données issues d'internet pour :

- Déterminer des comportement touristiques
- Déterminer des parcours touristiques
- Caractériser des populations touristiques
- Mesurer la e-réputation d'un site touristique

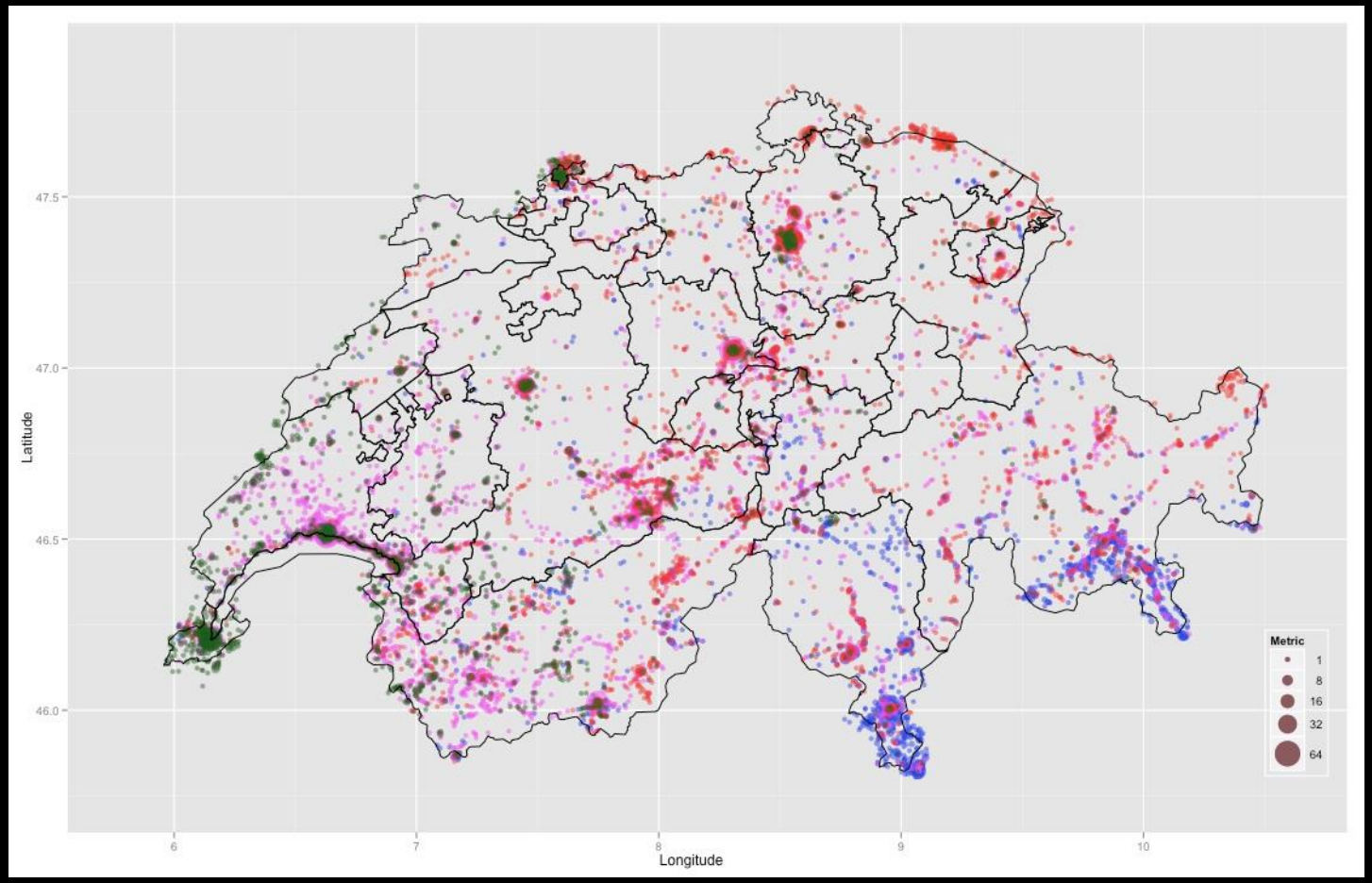
LES TRACES NUMÉRIQUES COMME SOURCES D'ENQUÊTES

- Des informations en quantité importante voire illimitée
  - Quel que soit le lieu
  - Quelle que soit l'échelle
- Des informations temps réel
  - Mise à jour des sources constantes
  - Toujours plus d'utilisateurs, toujours plus diversifiées
- Des informations complexes, variées, hétérogènes
  - Réseaux sociaux
  - Blogs, forums
  - Sites internet…

DES TRACES AUX CONNAISSANCES

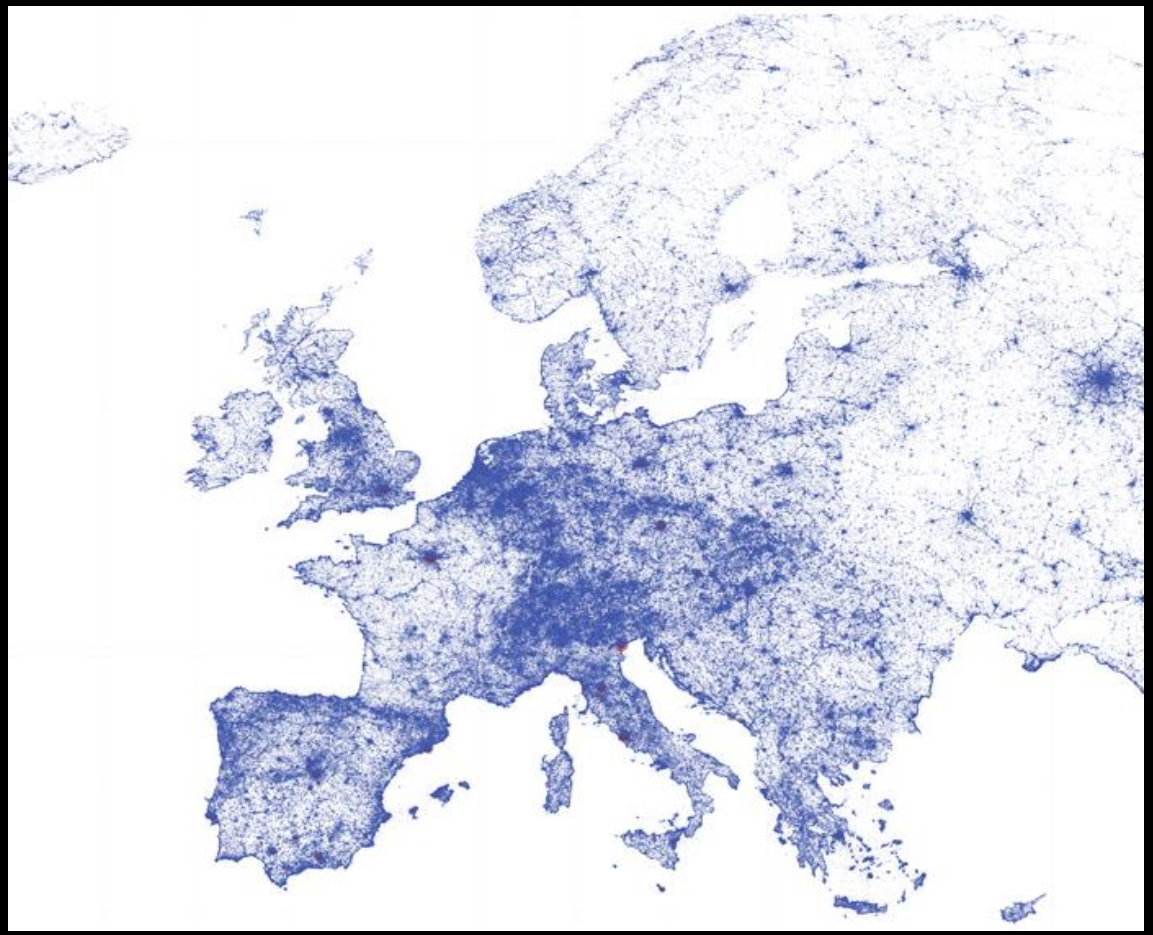
- L'enjeu final :
  - croiser l'analyse automatique des traces avec l'interprétation des contenus.
- Présenter quelques éléments sur la production des données, de leur sélection à leur interprétation.
- Modifier les approches et les méthodologies :
  - méthode inductive sans connaissance ni préjugé sur les pratiques des visiteurs, ou la valeur des sites.

EXEMPLE : LE TOURISME EN SUISSE



R + ggplot

L'EUROPE VUE PAR LES PHOTOGRAPHES DE PANORAMIO

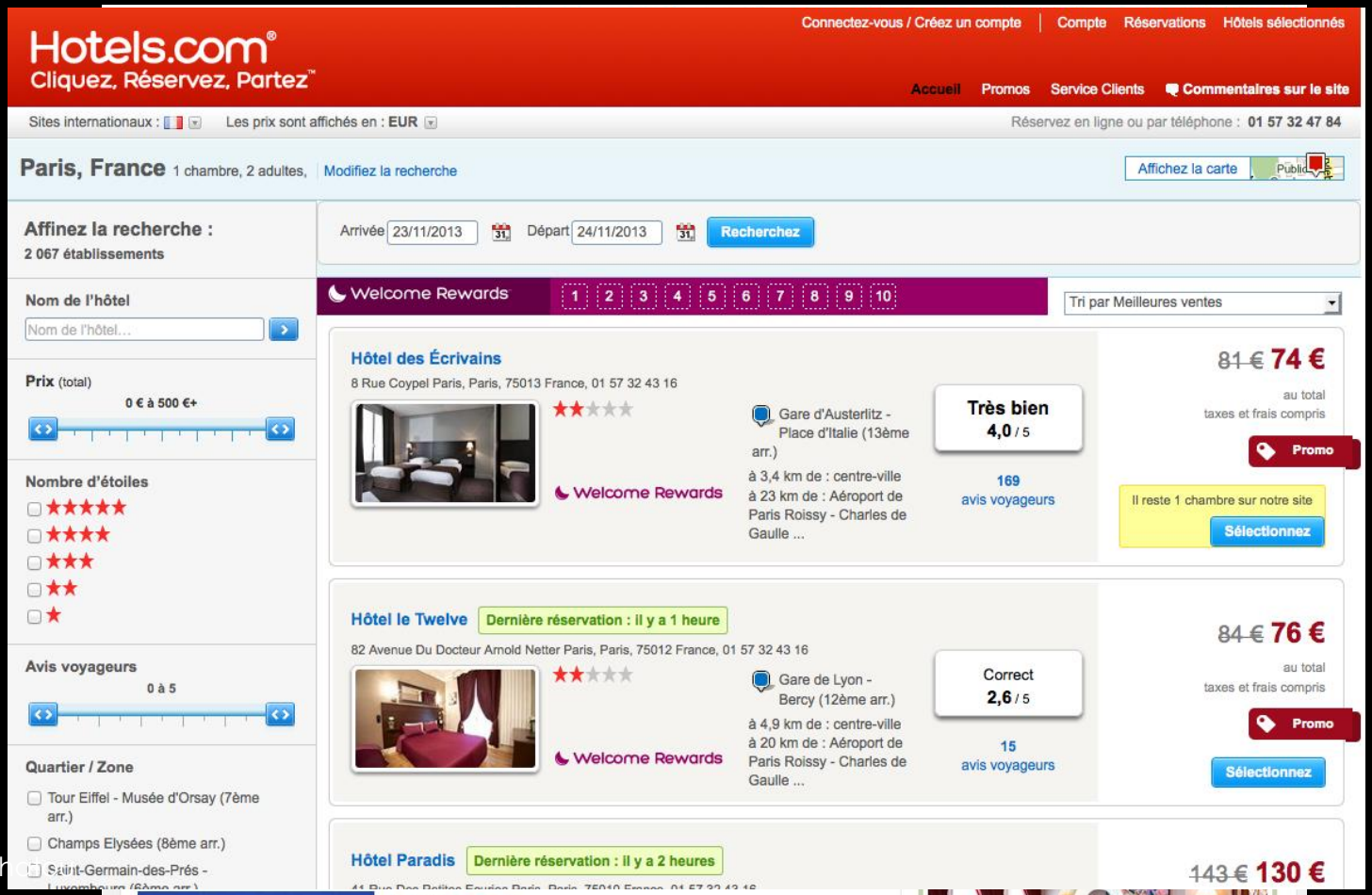


SOURCE DE DONNÉES

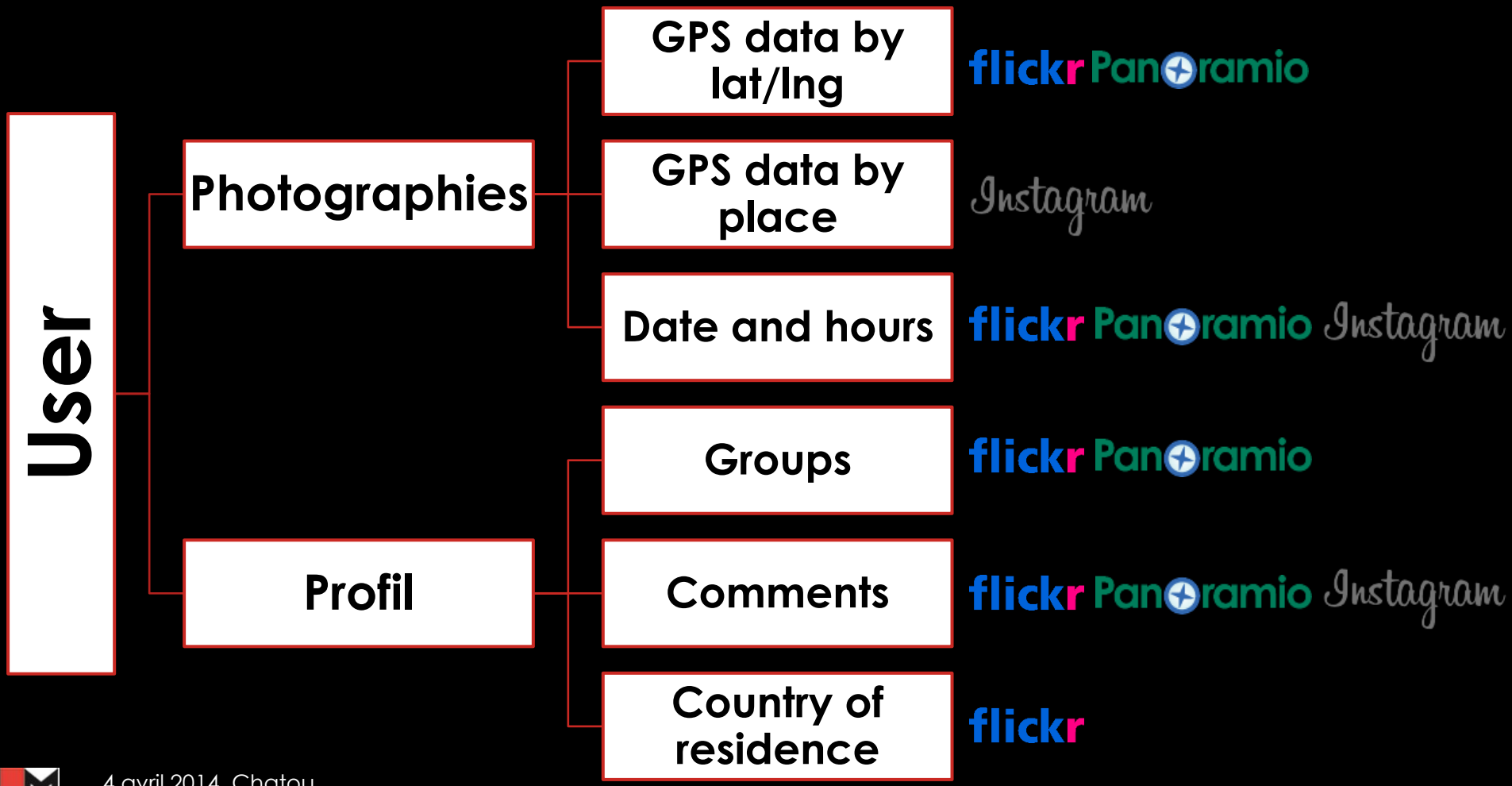
- Données Images
  - Flickr
  - Panoramio
  - Instagram
- Données commentaires
  - Hotel.com
  - TripAdvisor
- Blogs, forums …

SOURCES DE DONNÉES IMAGES

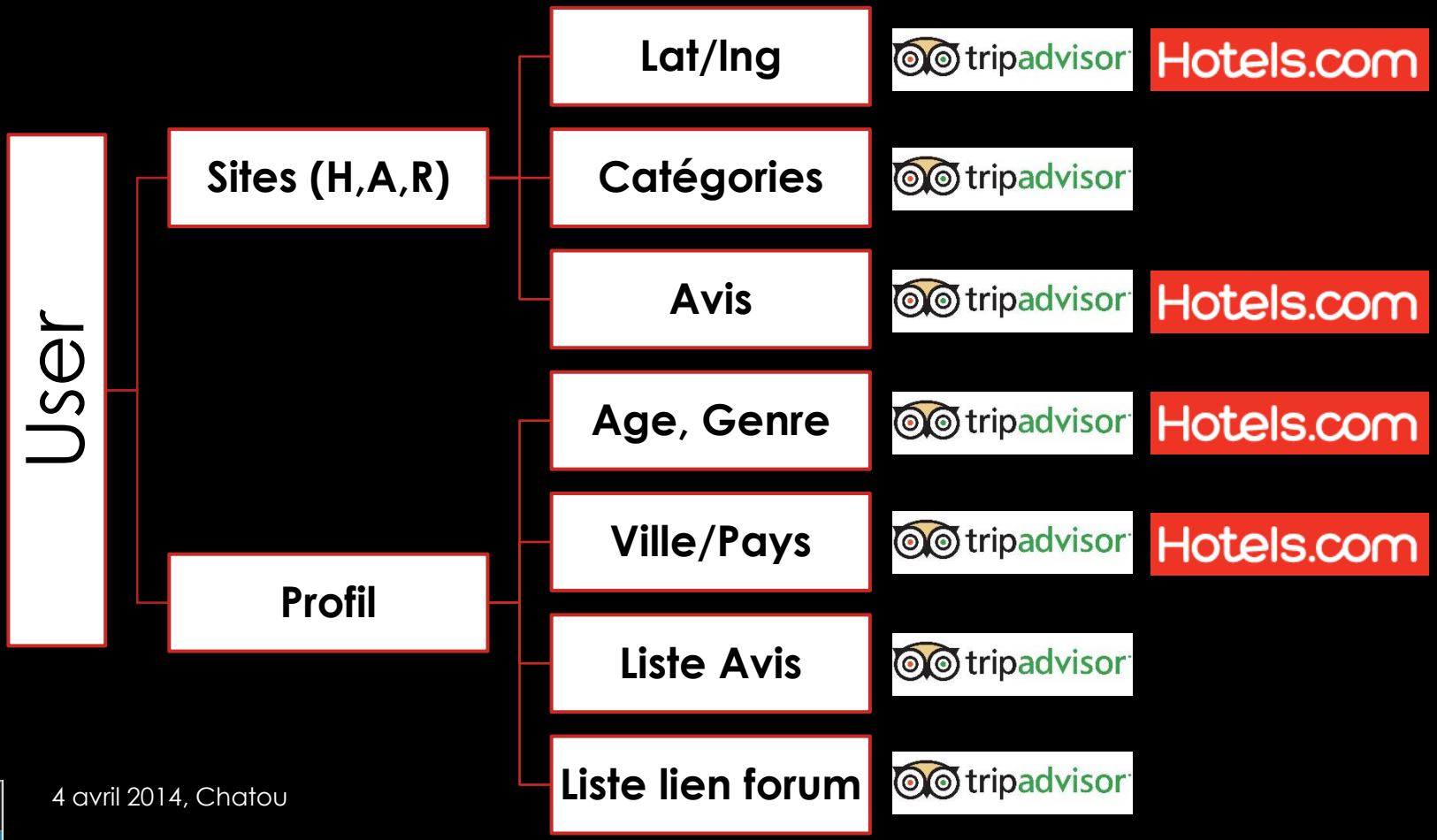
SOURCES DE DONNÉES « COMMENTAIRES »



LES SOURCES PHOTOS



LES SOURCES « AVIS ET COMMENTAIRES »



QUEL VOLUME POUR LE MONDE ?

- 68 M photos
- 2 M users

Panoramio 

- 196 M photos
- 1,9 M users

Flickr 

- 495 M photos
- 34 M users

Instagram 

- 175 000 hôtels
- 10 M reviews

Hotel.com 

- 2,5 M sites
- 45 M d'avis
- 10 M users

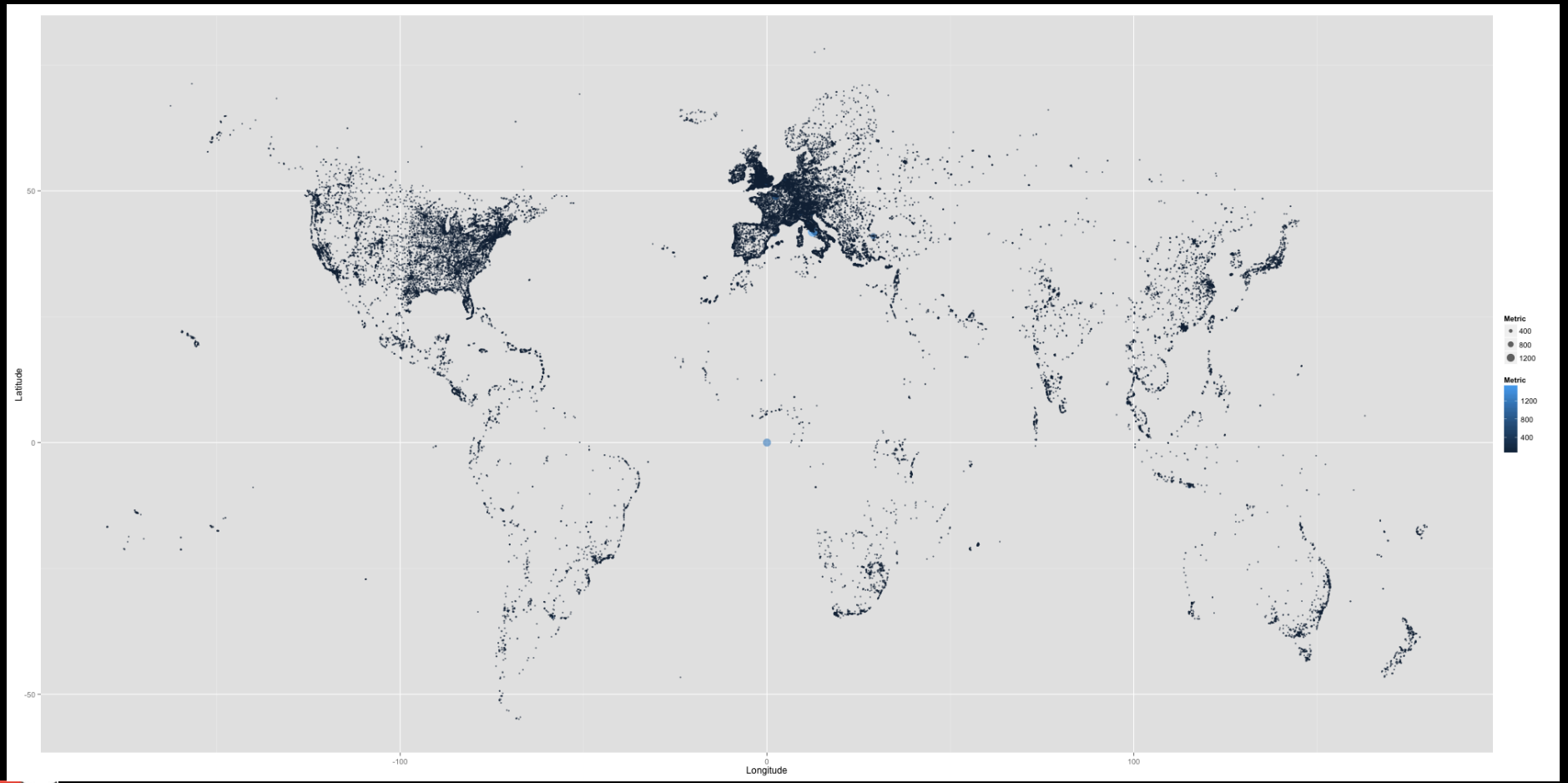
Tripadvisor 

PROBLÉMATIQUES SOCIAL DATA & BIG DATA ?

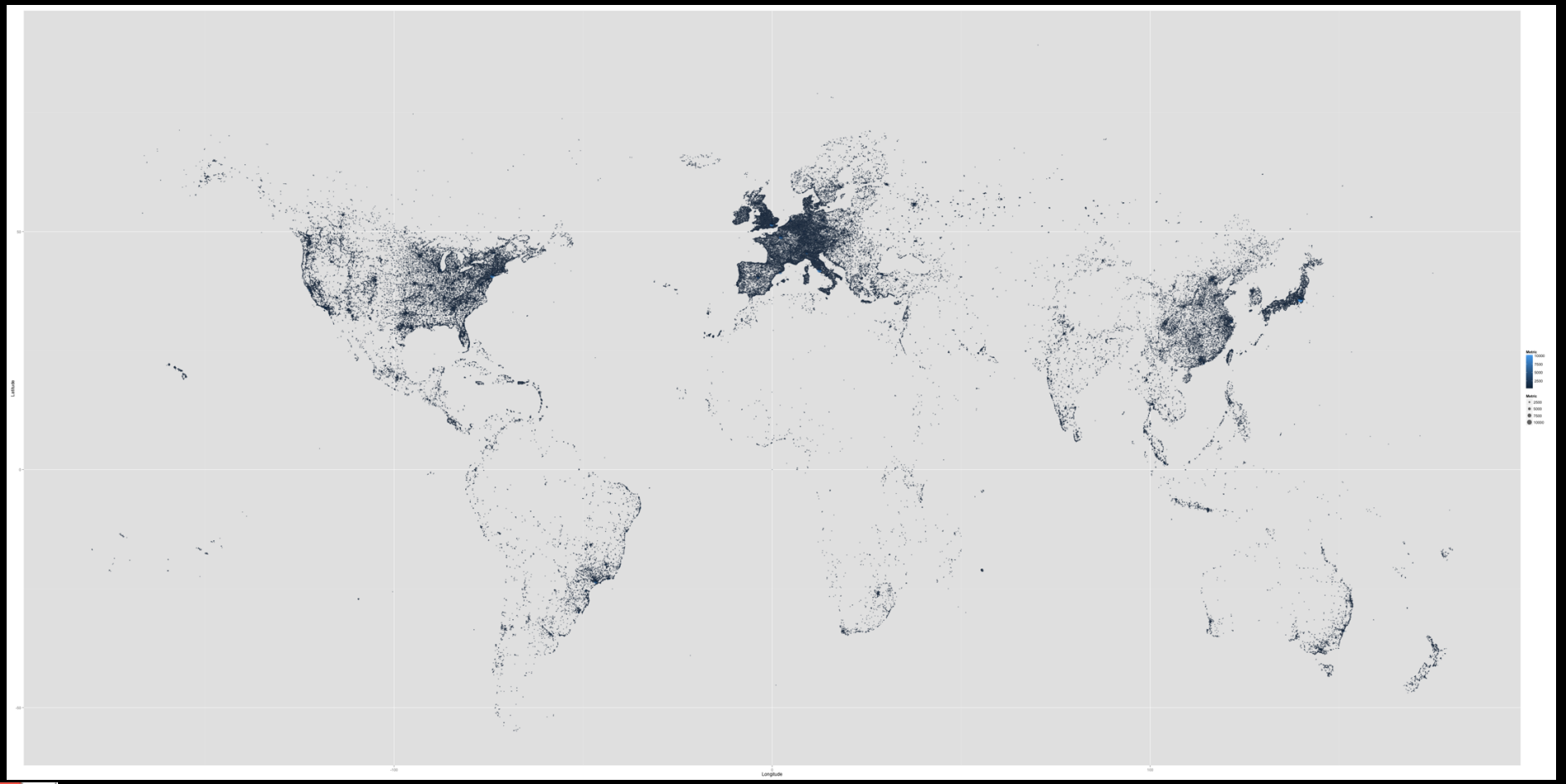
Technique	• Données aberrantes • Récolte et stockage des données • Traitement de données massives
Ethique	• Respect vie privée
Représentativité	• Progression intrinsèque des réseaux sociaux • Biais sociétaux • Interprétation et intelligibilité des résultats

VISUALISATION DES DONNÉES

VISUALISATION (HOTEL.COM)

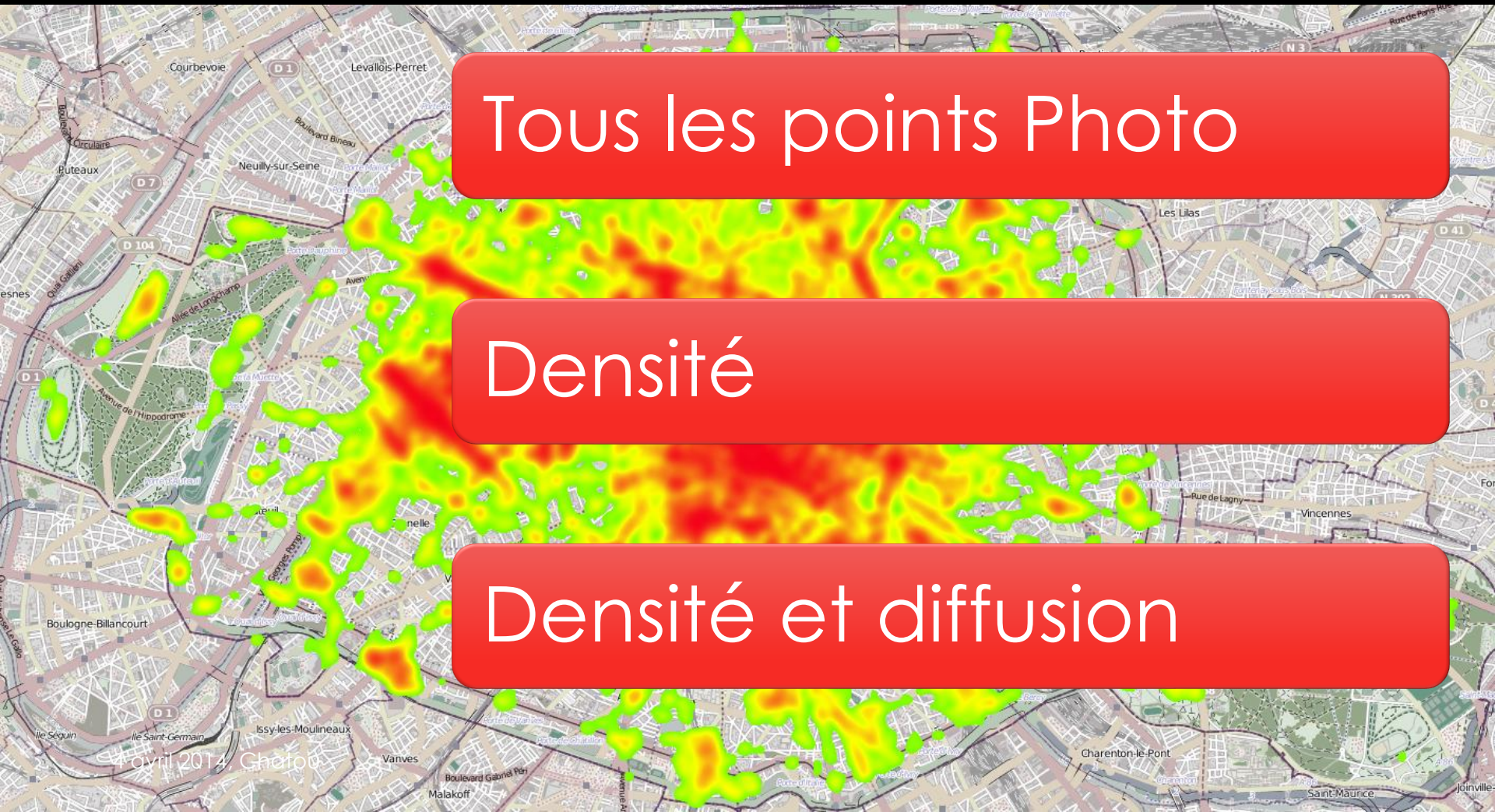


VISUALISATION (TRIPADVISOR)

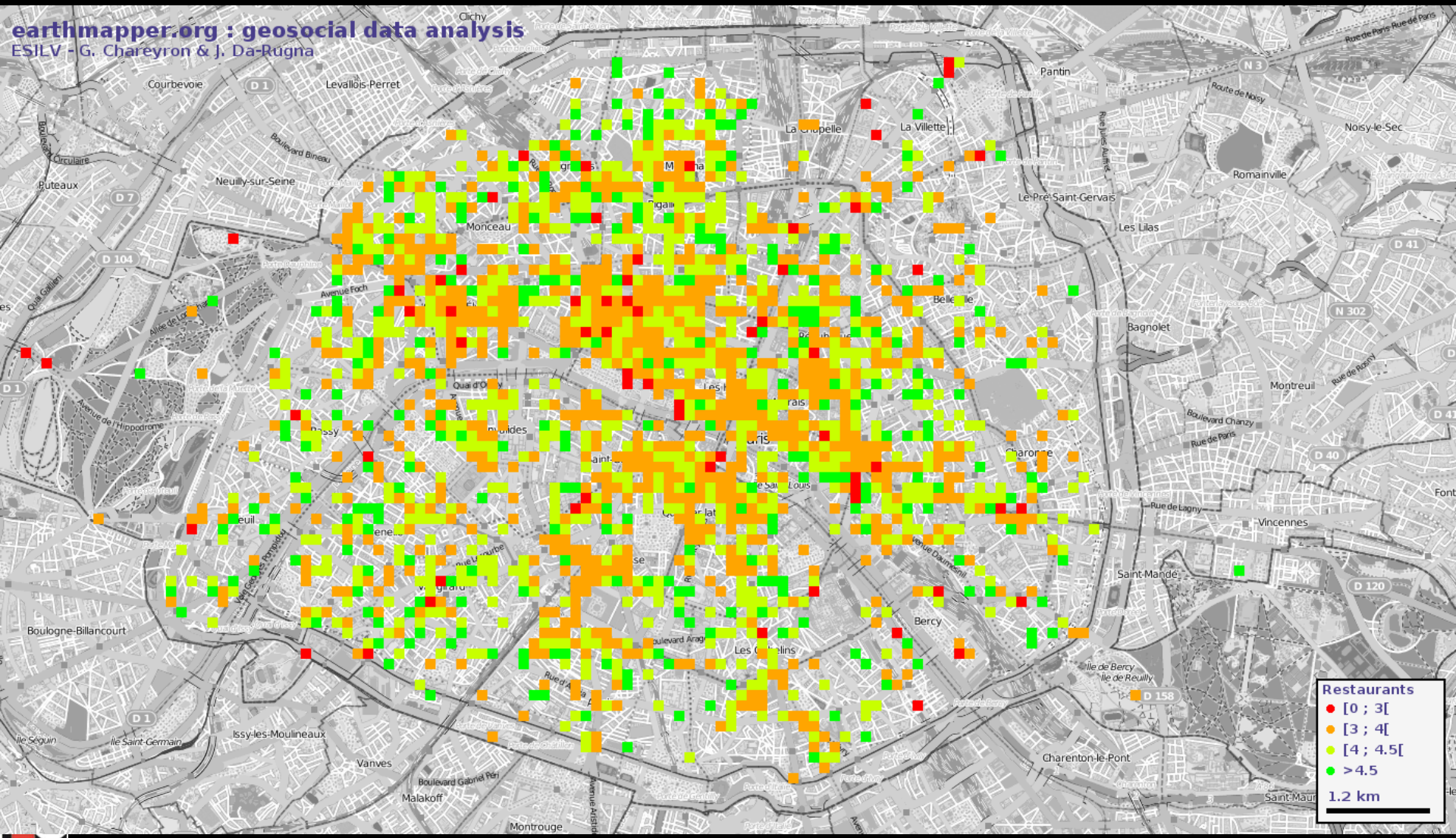


VISUALISATION (TRIPADVISOR)

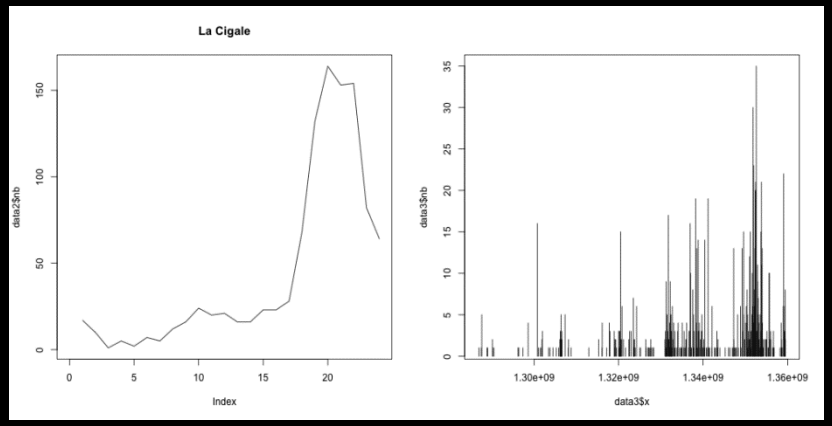
VISUALISATION DES DONNÉES GÉOLOCALISÉES

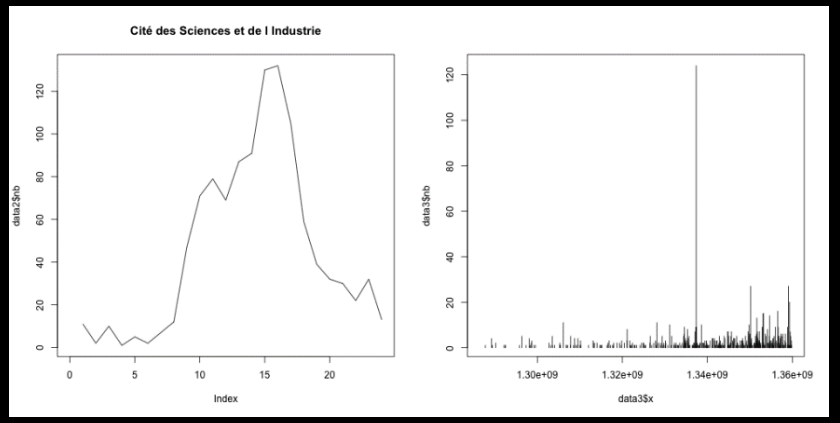


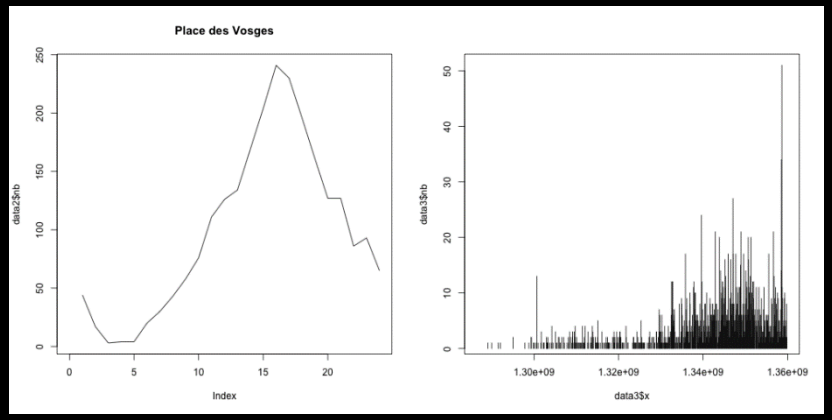
VISUALISATION PAR ZONE

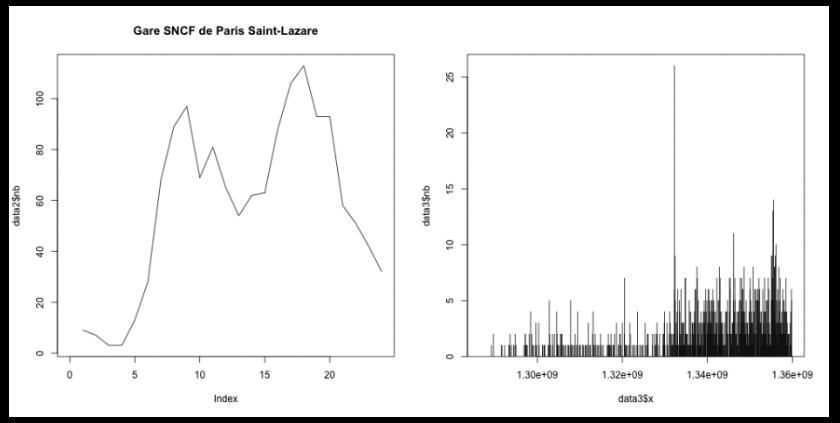


INSTAGRAM ET TEMPORALITÉ

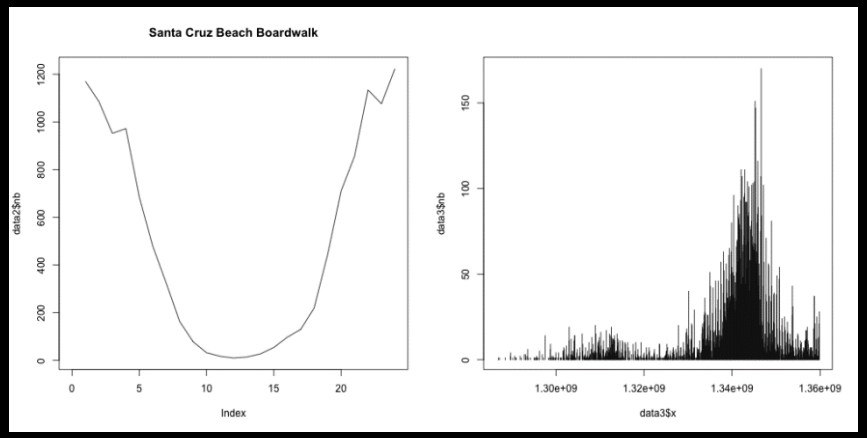


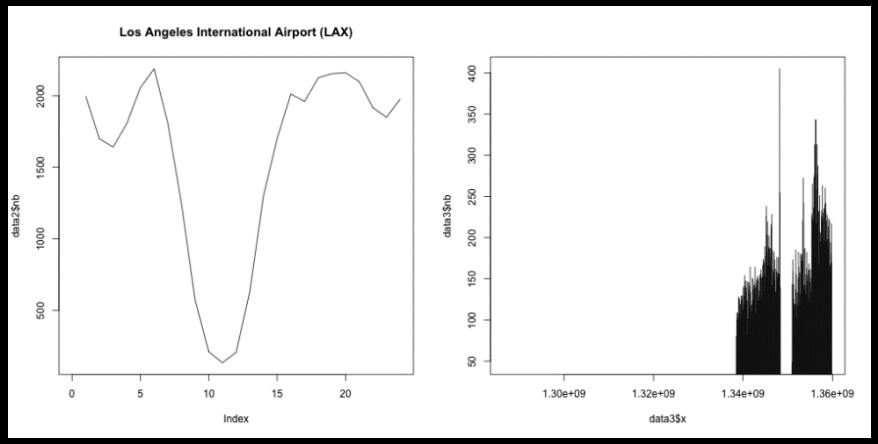


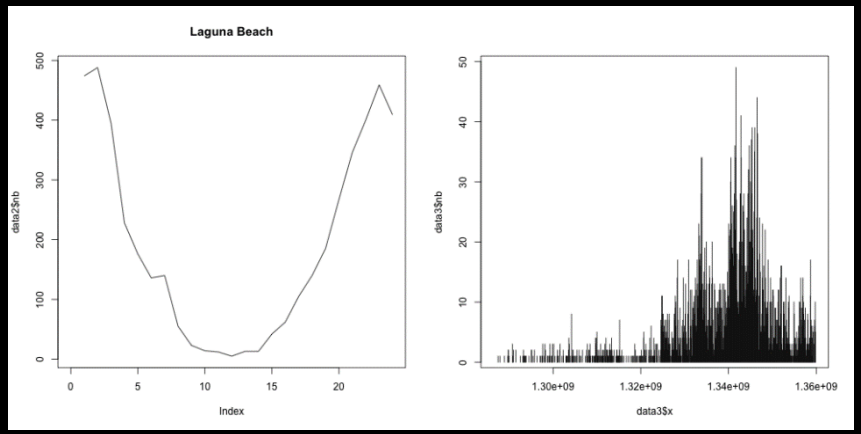


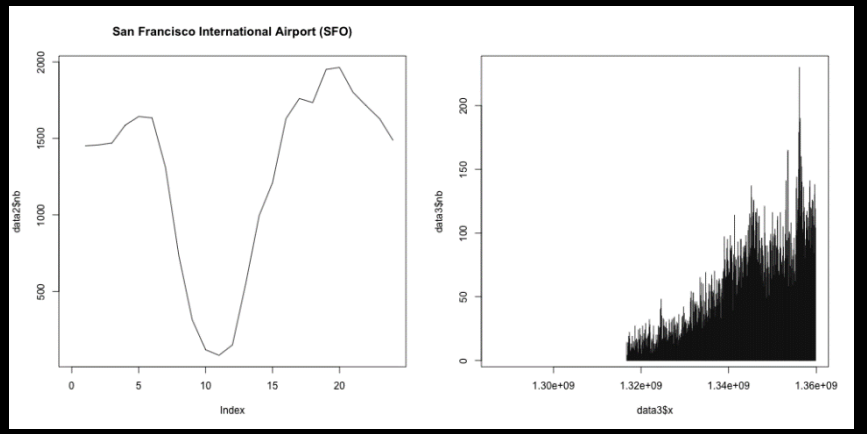


INSTAGRAM ET TEMPORALITÉ









UN PEU D'INTELLIGENCE ?

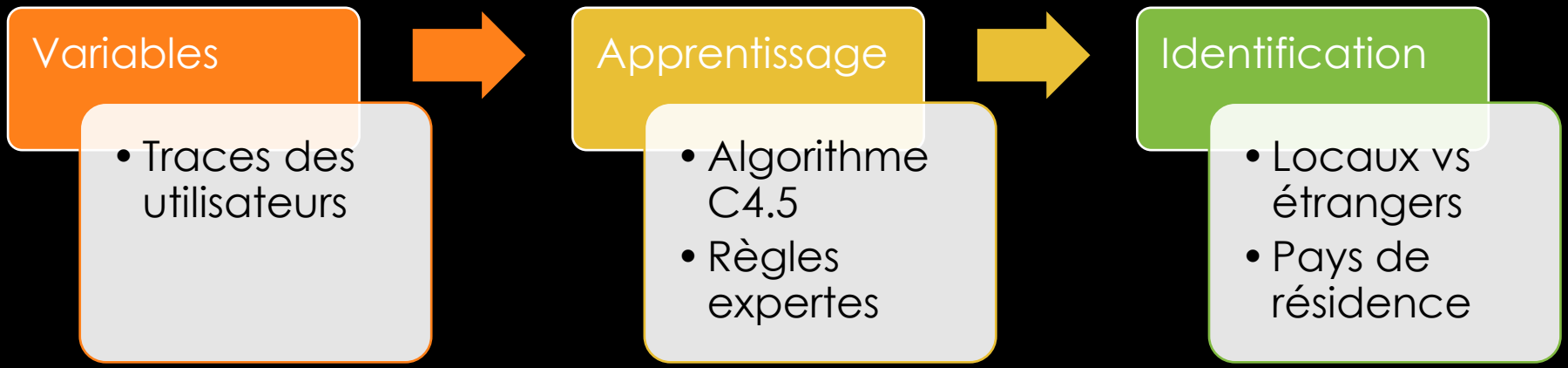
- Un potentiel immense, largement inexploité
- Détection automatique de comportements
  - typiques / atypiques
- Détection des évolutions, des transformations
- Mesurer autrement

Besoin de modélisation, d'outils de fouilles de données, de méthodologies et de pluridisciplinarité

ANALYSES DES DONNÉES

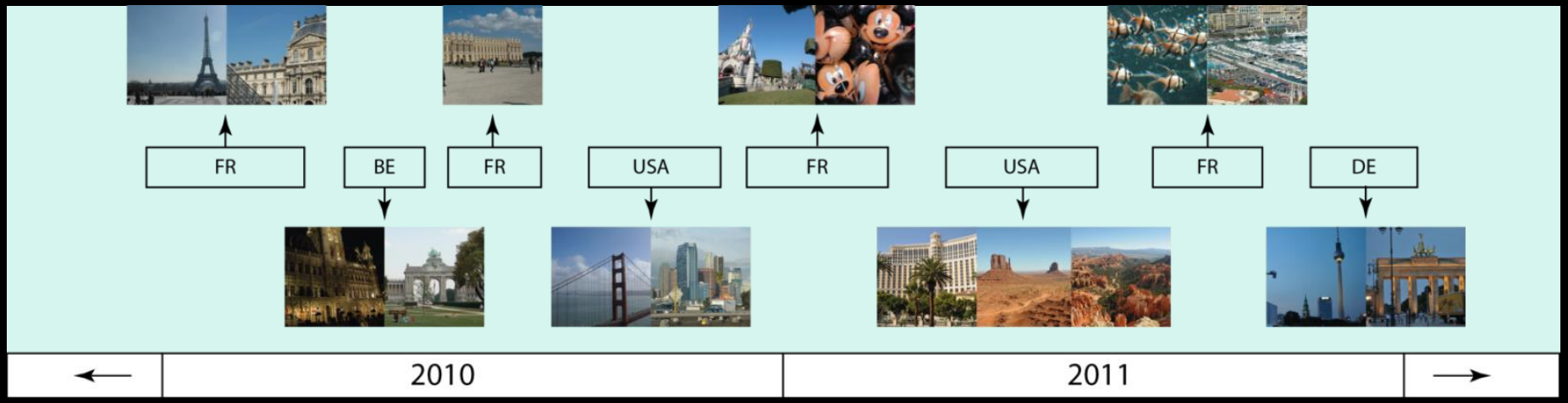
ENRICHISSEMENT DES PROFILS

- 30% des utilisateurs sont « localisés »
- Peut-on élargir cette localisation pour mieux connaitre notre échantillon ?



DESCRIPTION DES PHOTOGRAPHES

- Construction d'une « timeline » pour chaque utilisateur



DESCRIPTION DES PHOTOGRAPHES

- Nombre total de photographies
- Nombre total de pays visités
- Pour chaque pays visité :
  - Nombre de photographies
  - Nombre de sites visités
  - Nombre de visites dans le pays
  - Nombre de jours dans le pays
  - Nombre de jours entre 2 visites

CLASSIFICATION

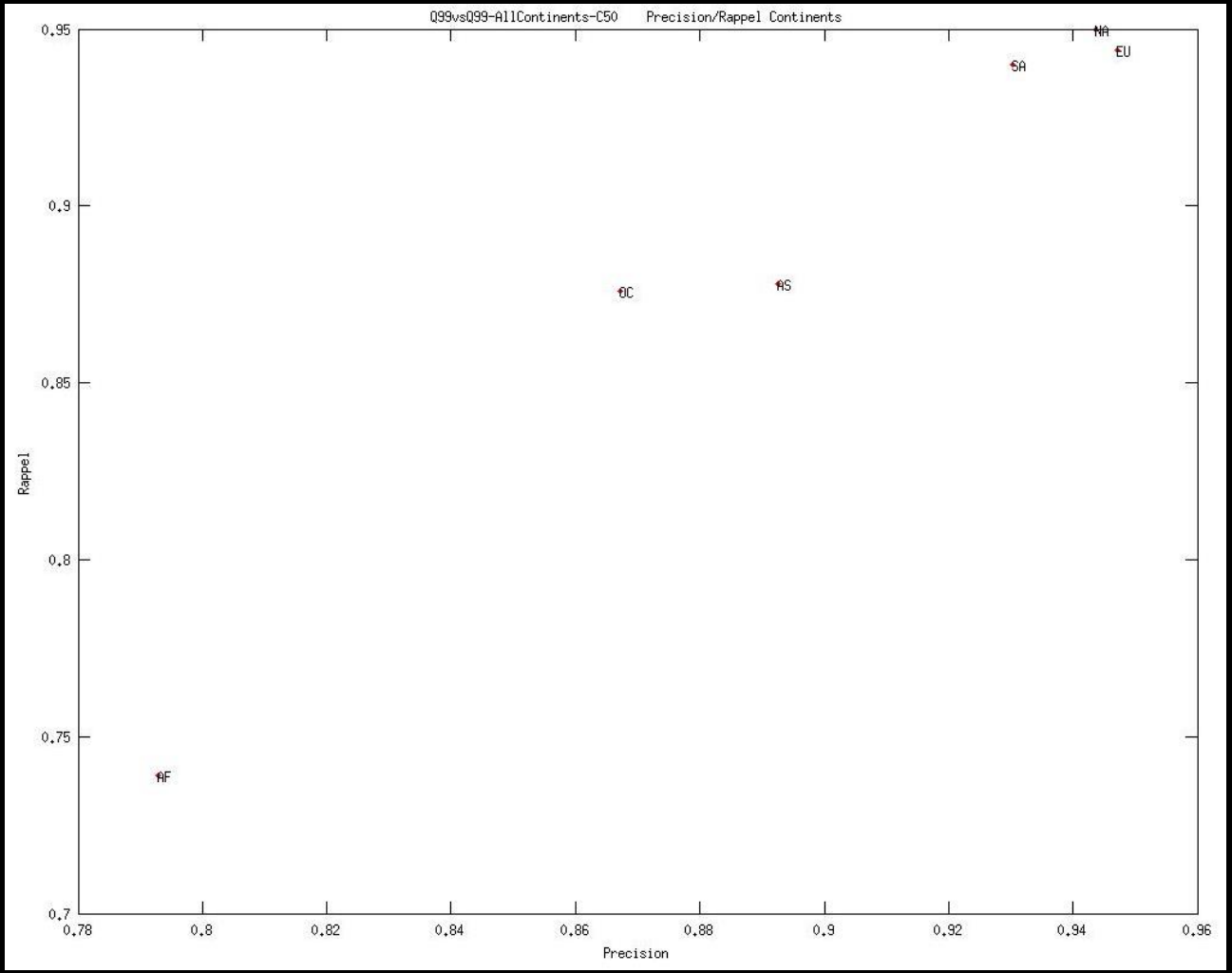
- Algorithme d'apprentissage C4.5 avec cross validation

- Logiciels :
  - Tanagra
  http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html
  - Weka
  http://www.cs.waikato.ac.nz/ml/weka/
  - R
  http://www.r-project.org/

CLASSIFICATION PAR CONTINENTS

	Rappel	Précision
Afrique	0,73902	0,79281
Asie	0,87808	0,8926
Europe	0,94399	0,94705
Amérique du Nord	0,94999	0,94357
Océanie	0,87607	0,86719
Amérique du Sud	0,93994	0,93031

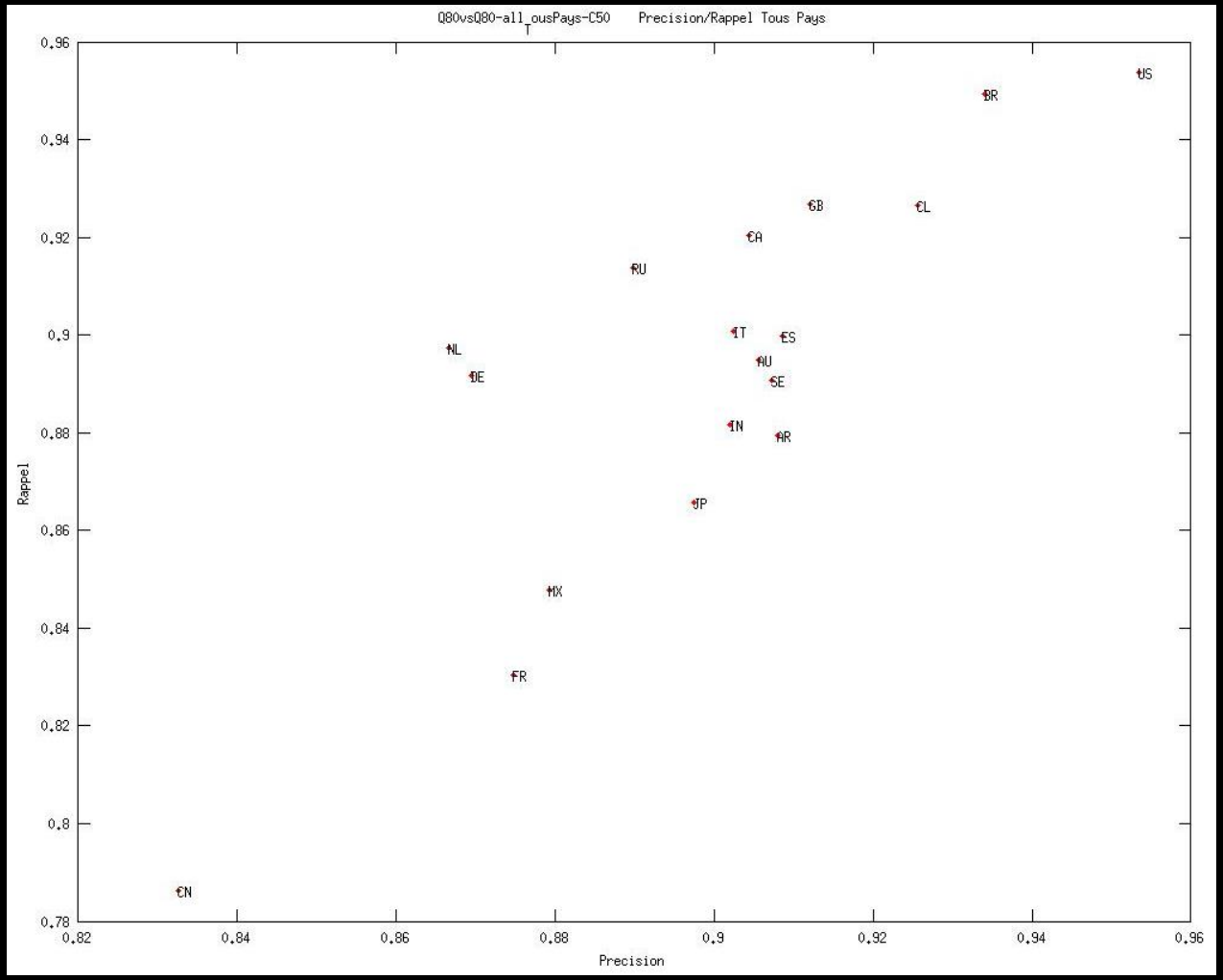
CLASSIFICATION PAR CONTINENTS



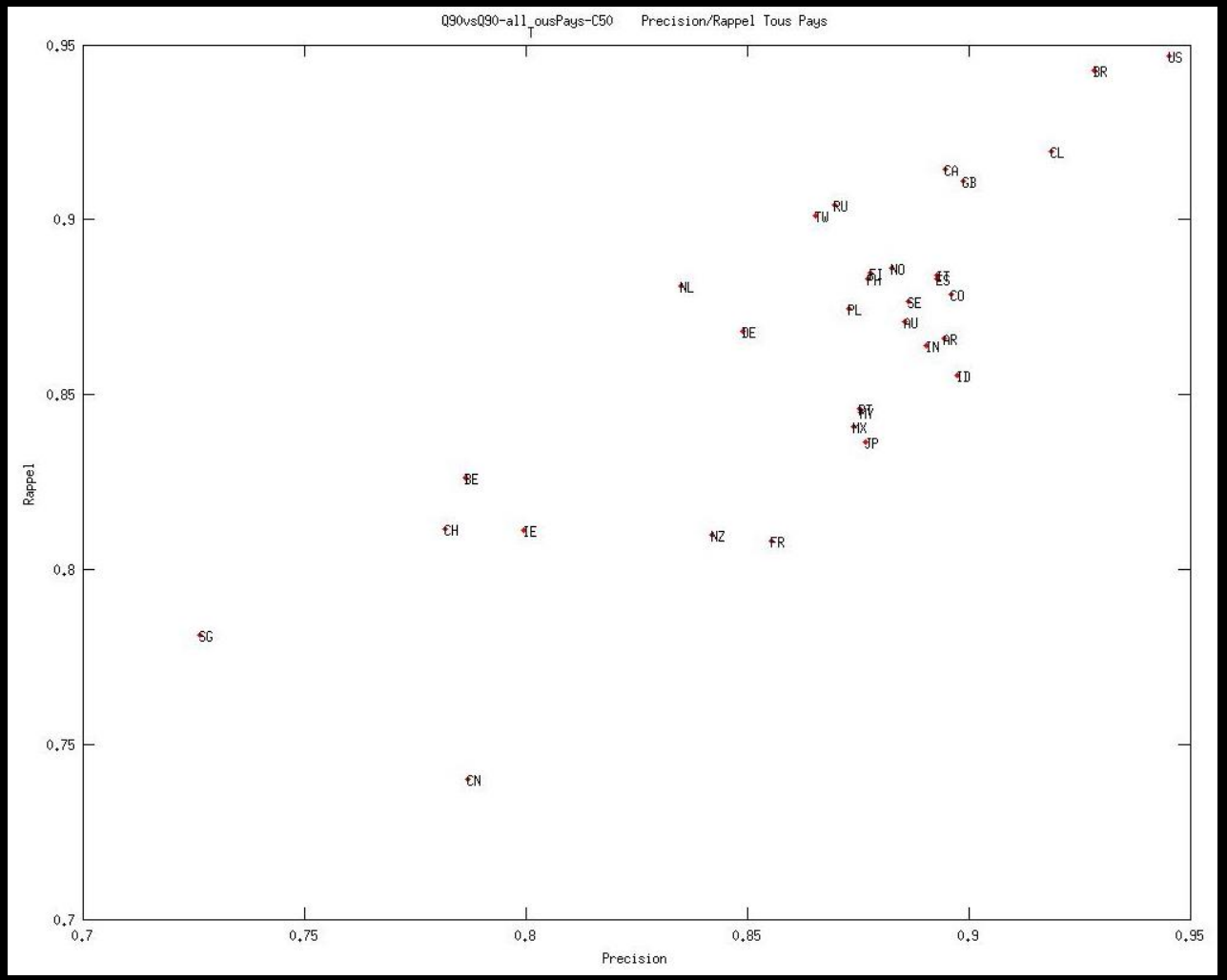
CLASSIFICATION PAR PAYS

- Quantile d'ordre 80% : Q80
  - 18 pays
- Quantile d'ordre 90% : Q90
  - 32 pays
- Quantile d'ordre 99% : Q99
  - 79 pays

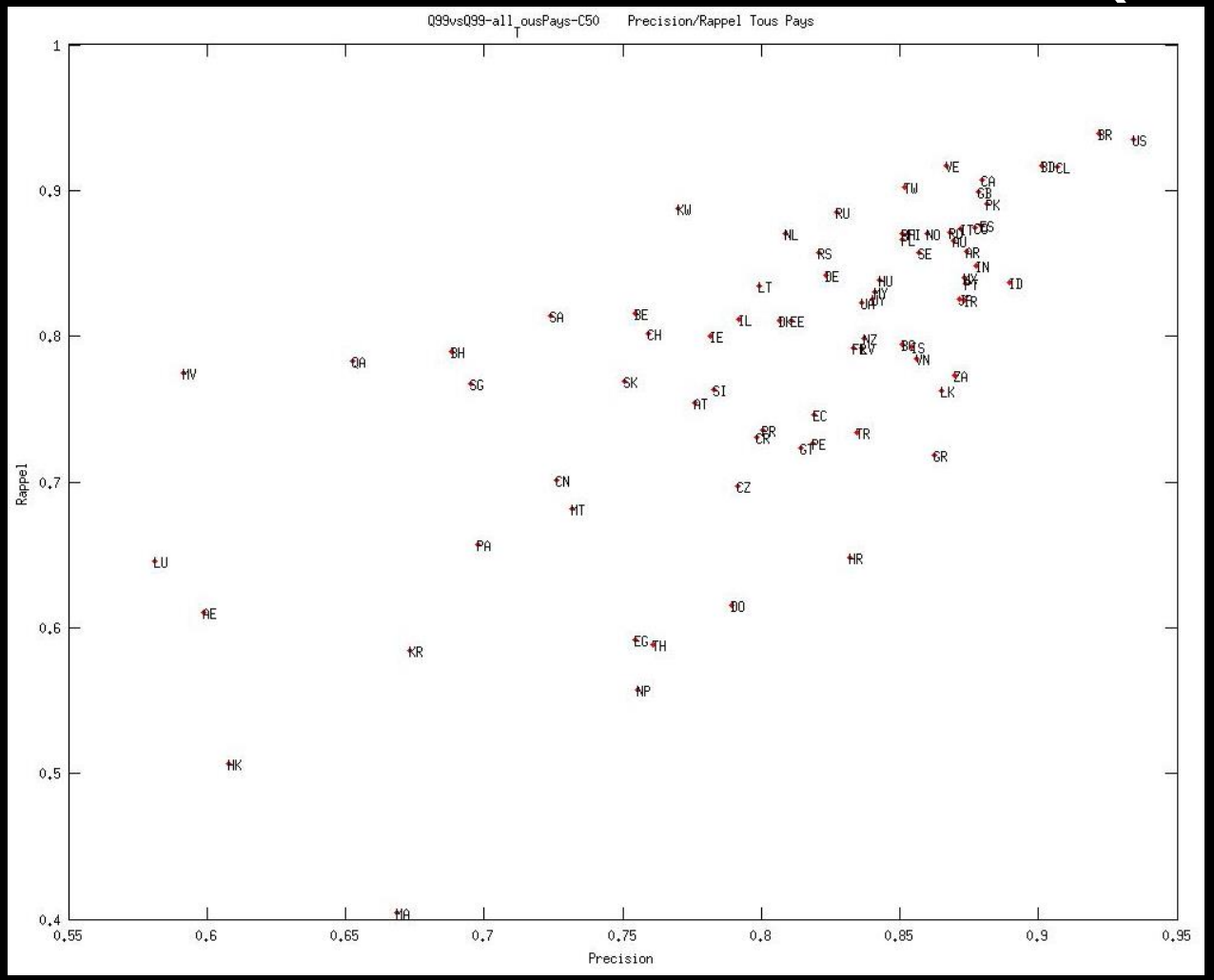
CLASSIFICATION PAR PAYS Q80



CLASSIFICATION PAR PAYS Q90



CLASSIFICATION PAR PAYS Q99

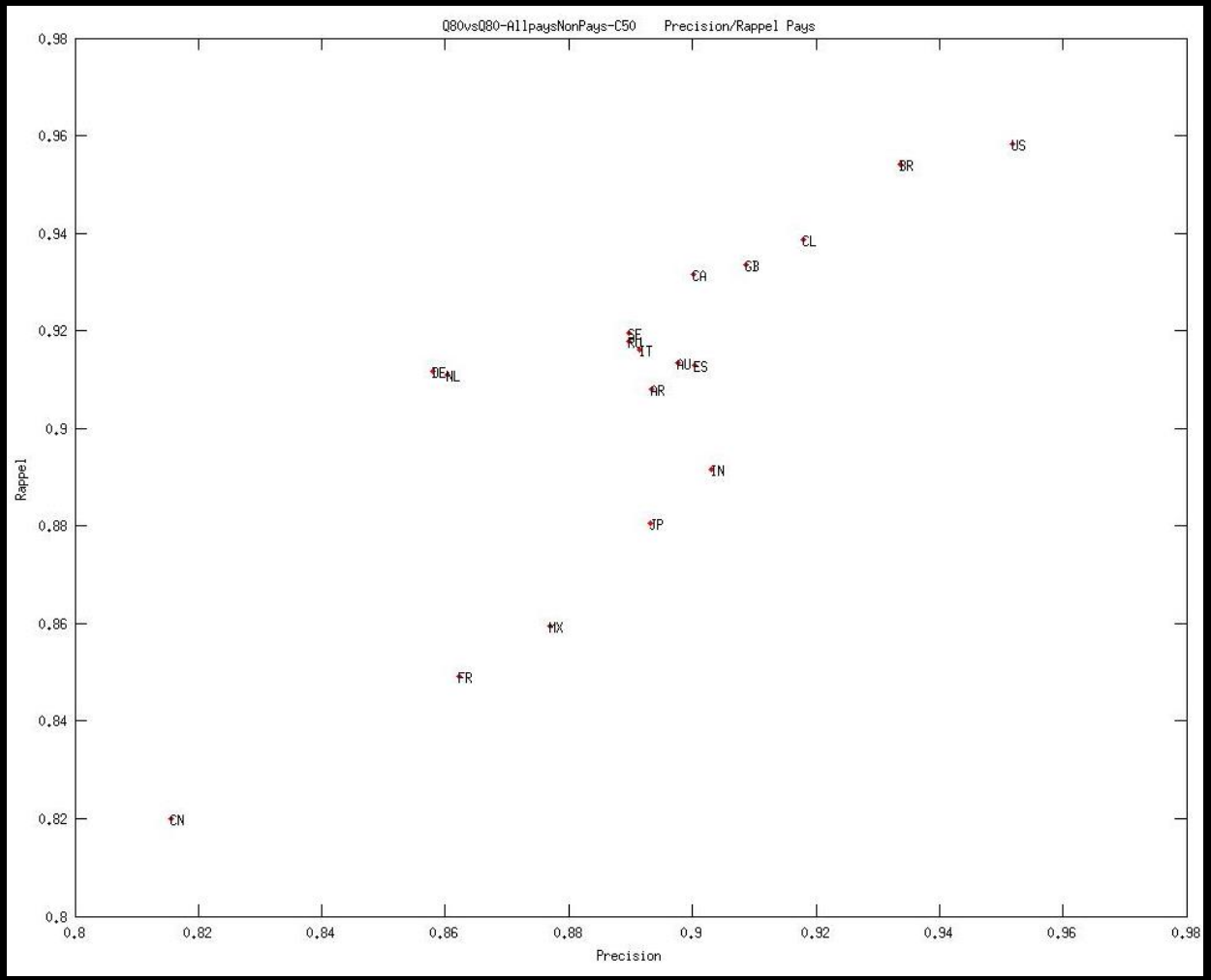


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

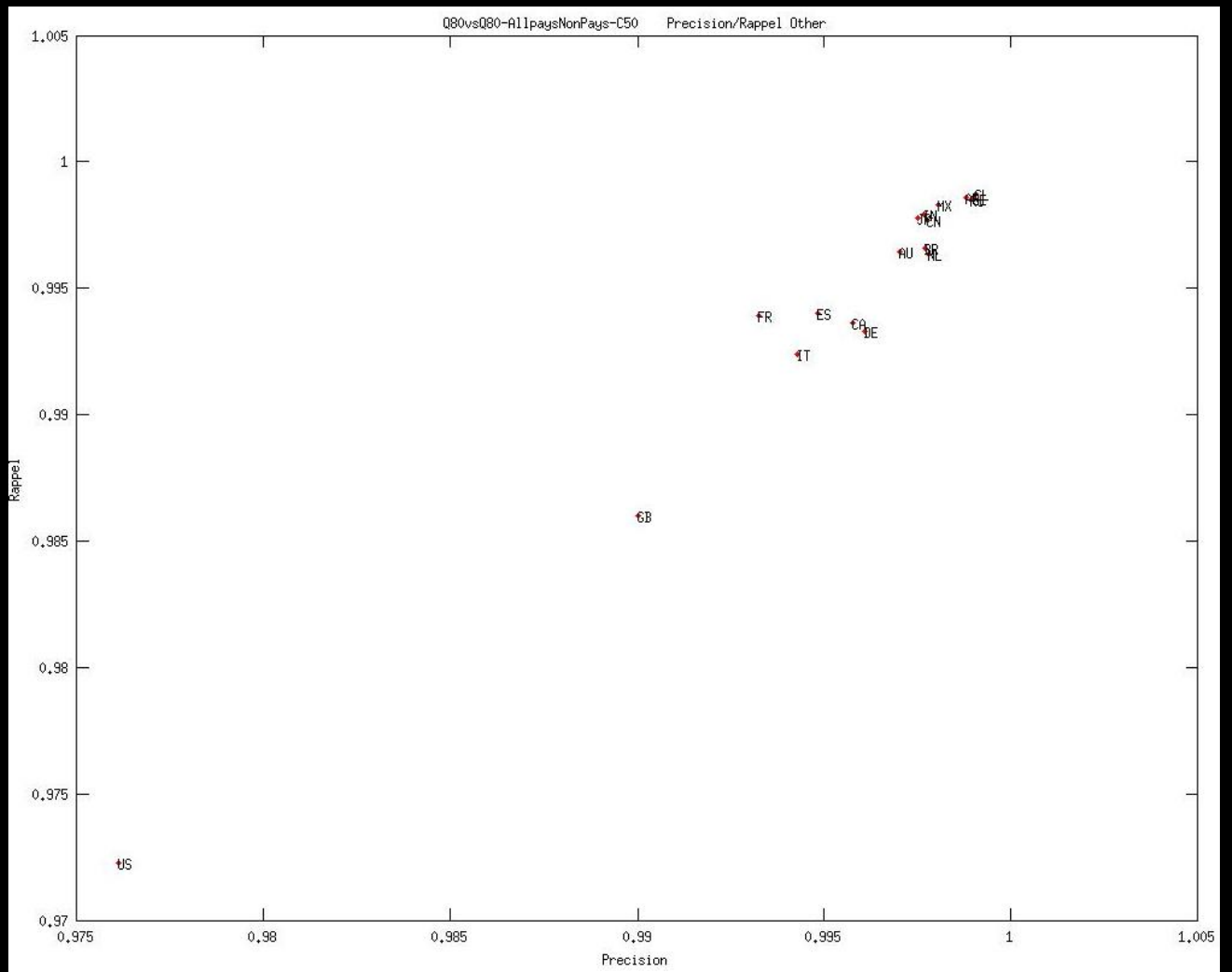
Précision-Rappel
Pour les « Locaux »



CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

Précision-Rappel
Pour les « Non-Locaux »

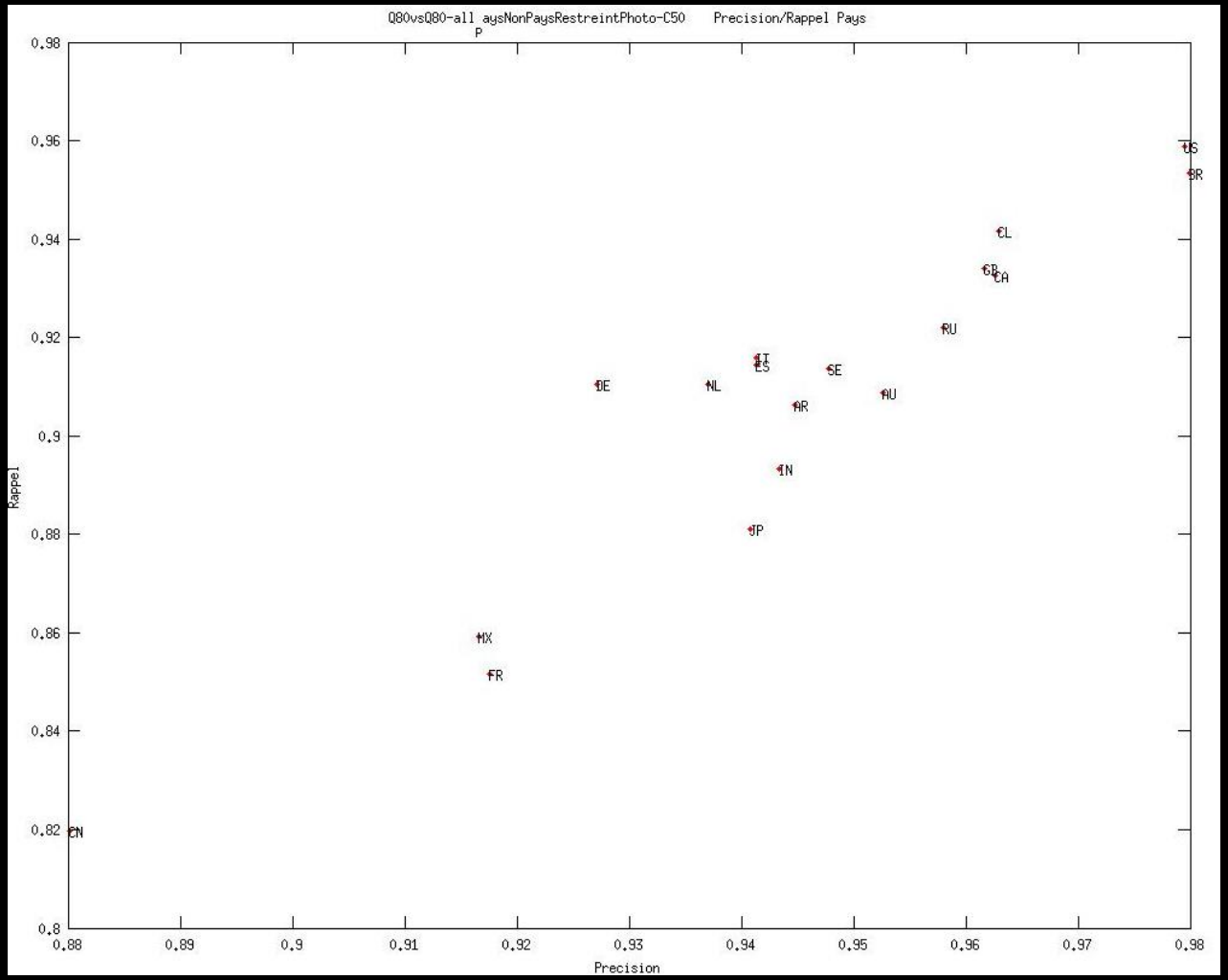


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

Précision-Rappel
Pour les « Locaux »

Restriction aux users ayant pris des photos dans le pays considéré

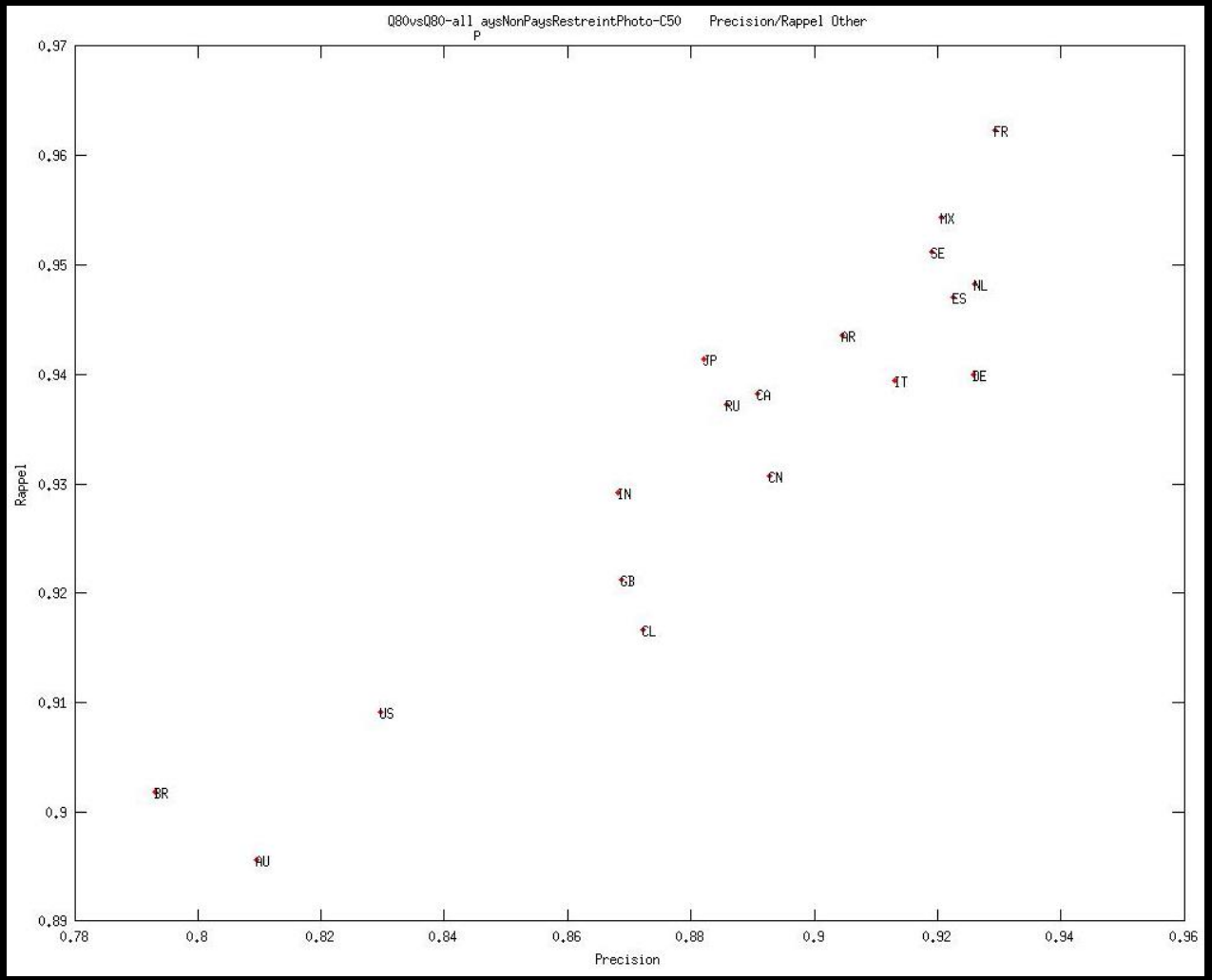


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

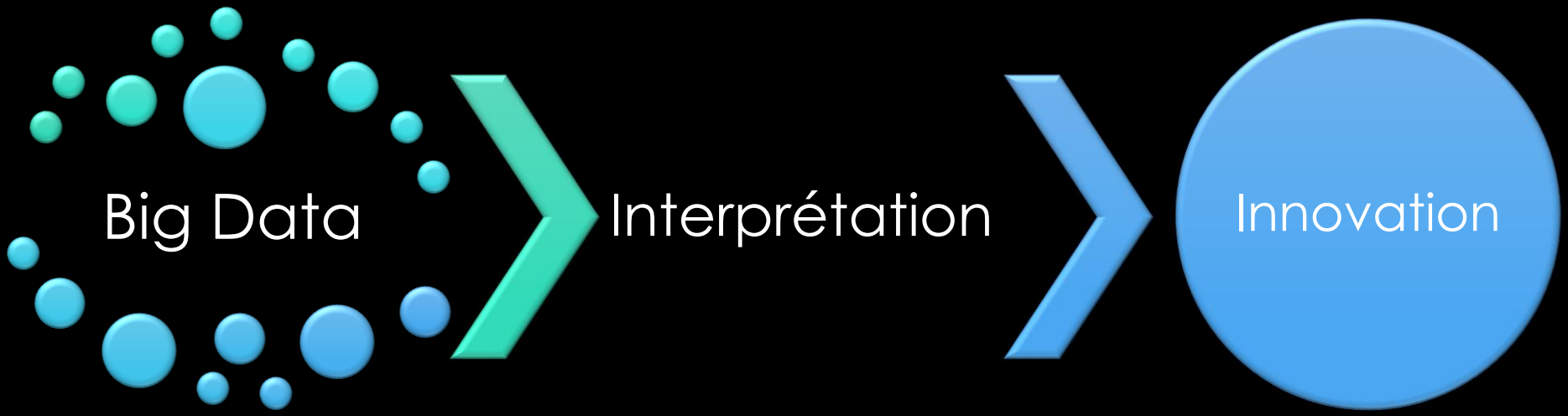
Q80

Précision-Rappel
Pour les «Non- Locaux »

Restriction aux users ayant pris des photos dans le pays considéré



CONCLUSION

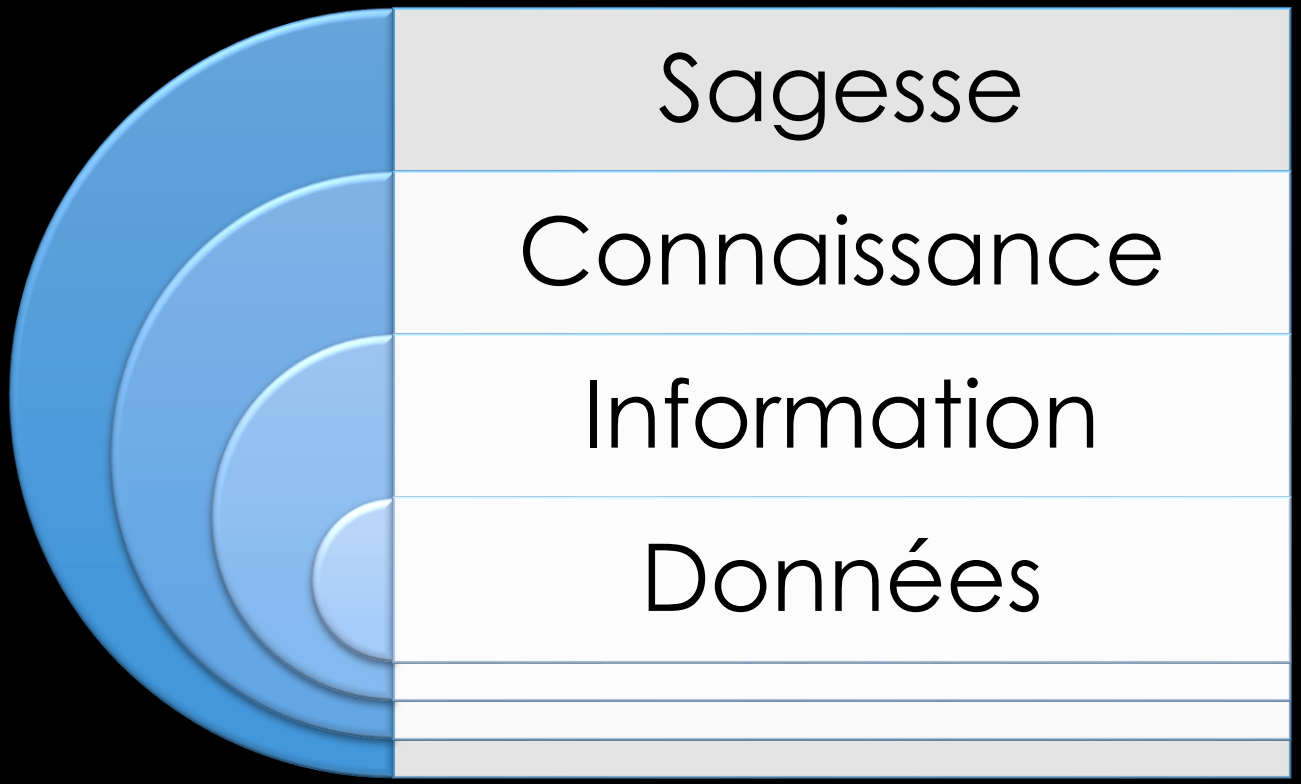


- Récolte
- Performance
- Modélisation
- Data mining

- Représentativité
- Confrontation
- Biais sociétaux

- Nouveaux services
- Nouvelles analyses

CONCLUSION



MERCI DE VOTRE ATTENTION

http://earthmapper.org
http://world.earthmapper.org