

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



PRÉSENTATION

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



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

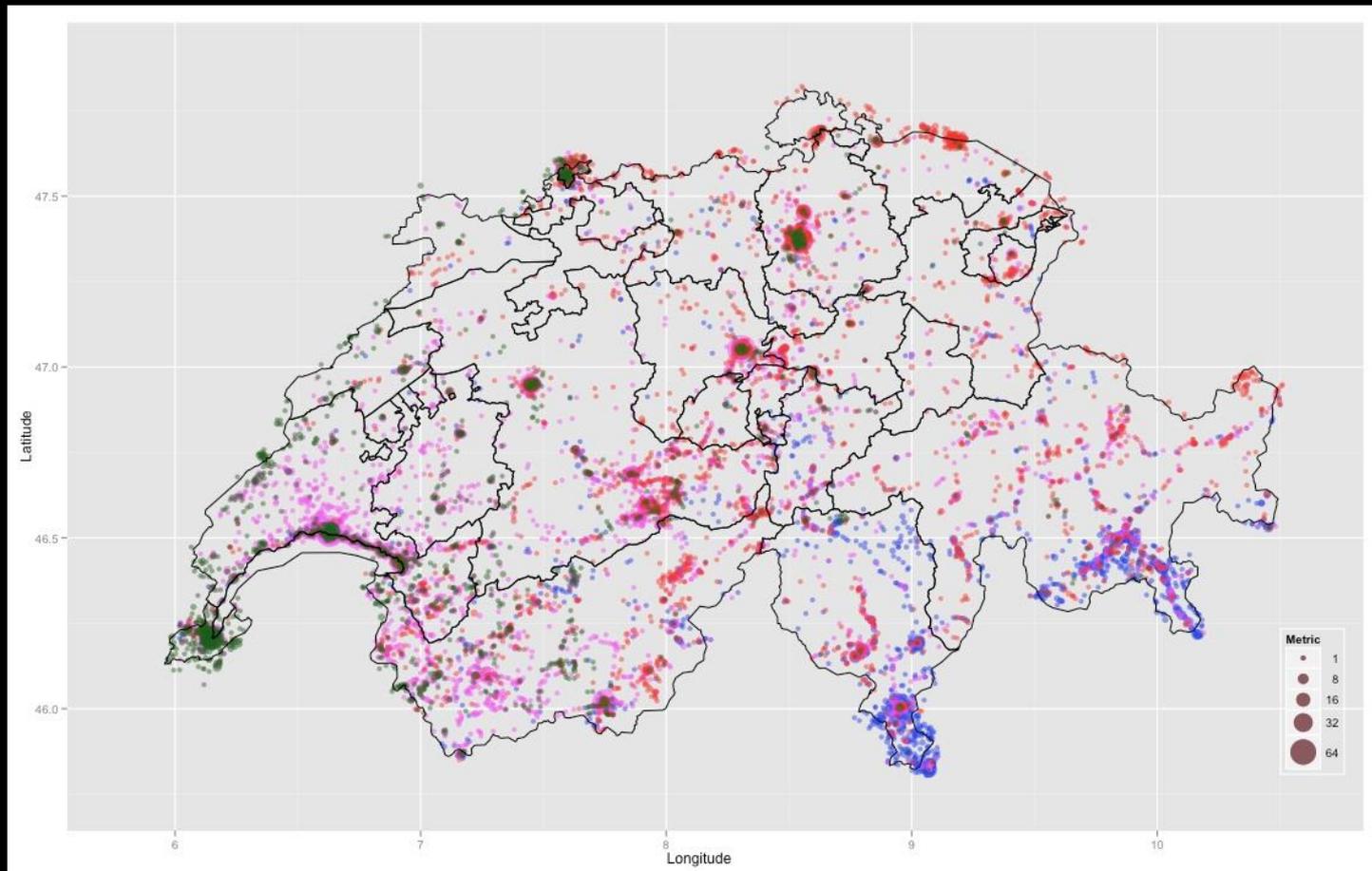
LES TRACES NUMÉRIQUES COMME SOURCES D'ENQUÊTES

- 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...

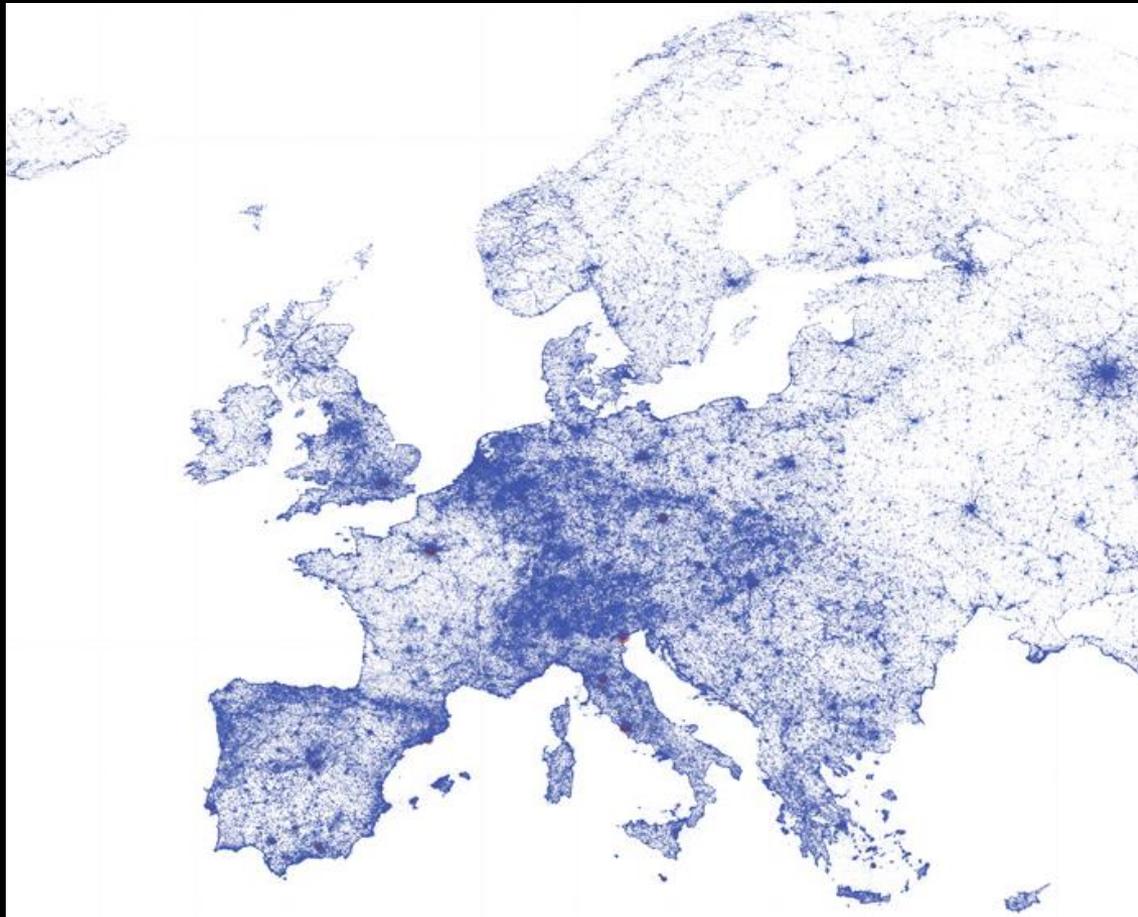
DES TRACES AUX CONNAISSANCES

- 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



L'EUROPE VUE PAR LES PHOTOGRAPHES DE PANORAMIO

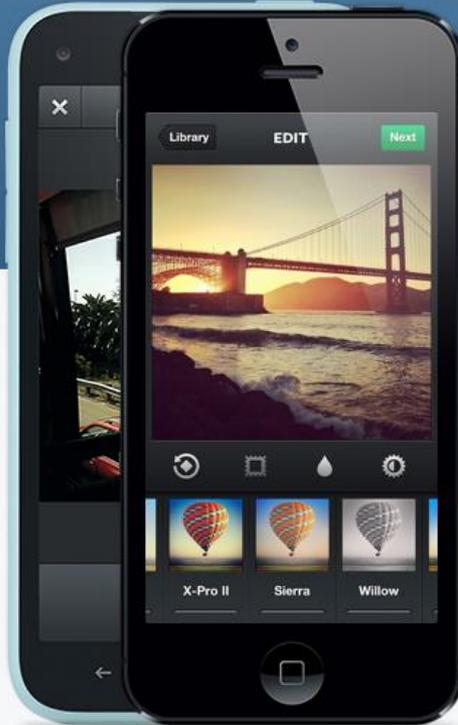


SOURCE DE DONNÉES

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

SOURCES DE DONNÉES IMAGES

Panoramic



Instagram [Connectez-vous](#)

Instagram, qu'est-ce que c'est ?

Instagram is a **fast, beautiful** and **fun** way to share your life with friends and family.

Take a picture or video, choose a filter to transform its look and feel, then post to Instagram — it's that easy. You can even share to Facebook, Twitter, Tumblr and more. It's a new way to see the world.

Et encore mieux, c'est gratuit !



SOURCES DE DONNÉES

« COMMENTAIRES »

Hotels.com®
Cliquez, Réservez, Partez™

Connectez-vous / Créez un compte | Compte Réservations Hôtels sélectionnés

Accueil Promos Service Clients Commentaires sur le site

Sites internationaux : [FR] Les prix sont affichés en : EUR Réservez en ligne ou par téléphone : 01 57 32 47 84

Paris, France 1 chambre, 2 adultes, Modifiez la recherche

Affichez la carte

Affinez la recherche : 2 067 établissements

Nom de l'hôtel

Prix (total) 0 € à 500 €+

Nombre d'étoiles

Avis voyageurs 0 à 5

Quartier / Zone

Arrivée 23/11/2013 Départ 24/11/2013 Recherchez

Welcome Rewards

Tri par Meilleures ventes

Hôtel des Écrivains

8 Rue Coypel Paris, Paris, 75013 France, 01 57 32 43 16

★★★★☆

Très bien 4,0 / 5

169 avis voyageurs

81€ 74€ au total taxes et frais compris

Promo

Il reste 1 chambre sur notre site

Sélectionnez

Hôtel le Twelve Dernière réservation : il y a 1 heure

82 Avenue Du Docteur Arnold Netter Paris, Paris, 75012 France, 01 57 32 43 16

★★★☆☆

Correct 2,6 / 5

15 avis voyageurs

84€ 76€ au total taxes et frais compris

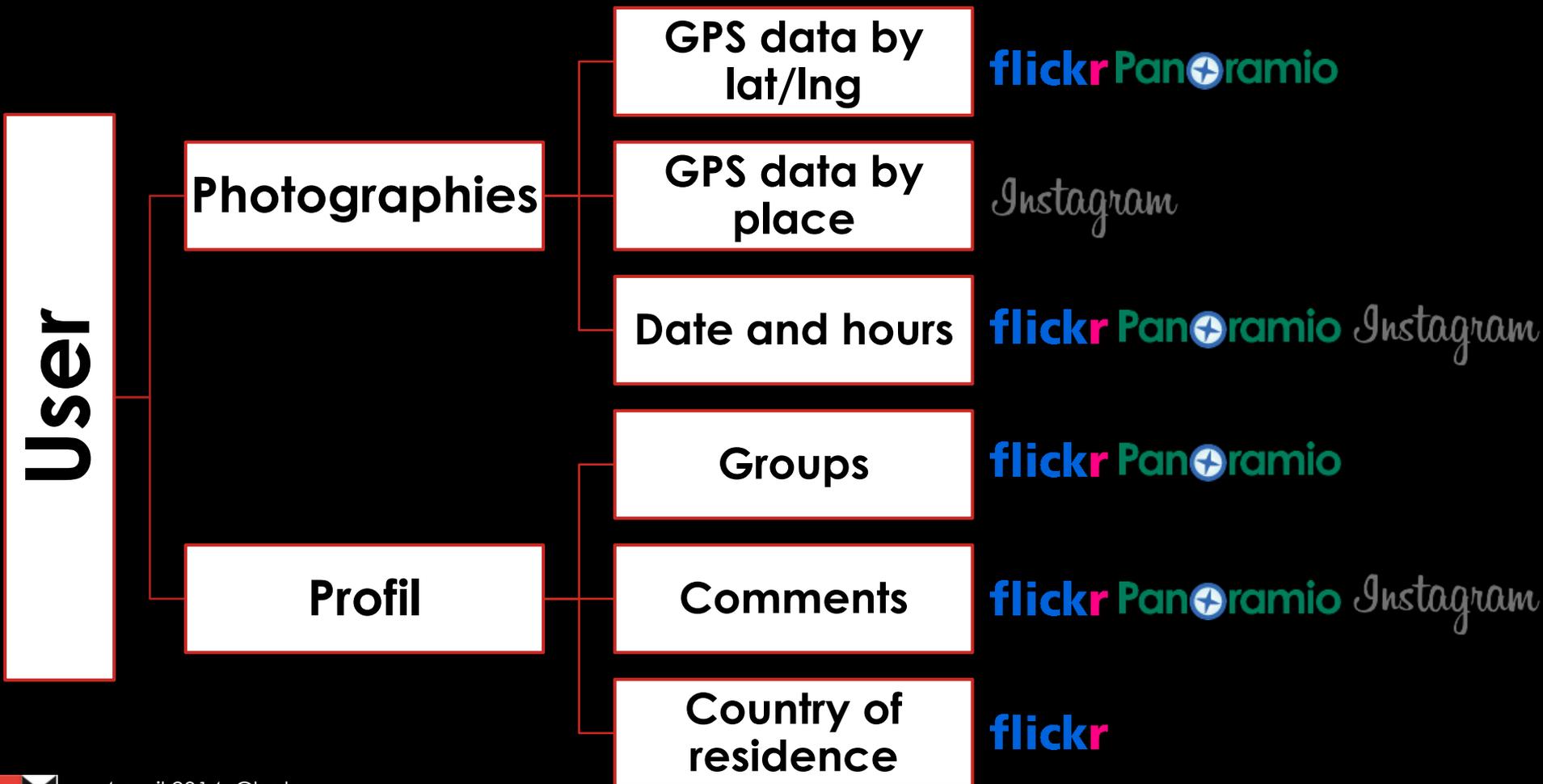
Promo

Sélectionnez

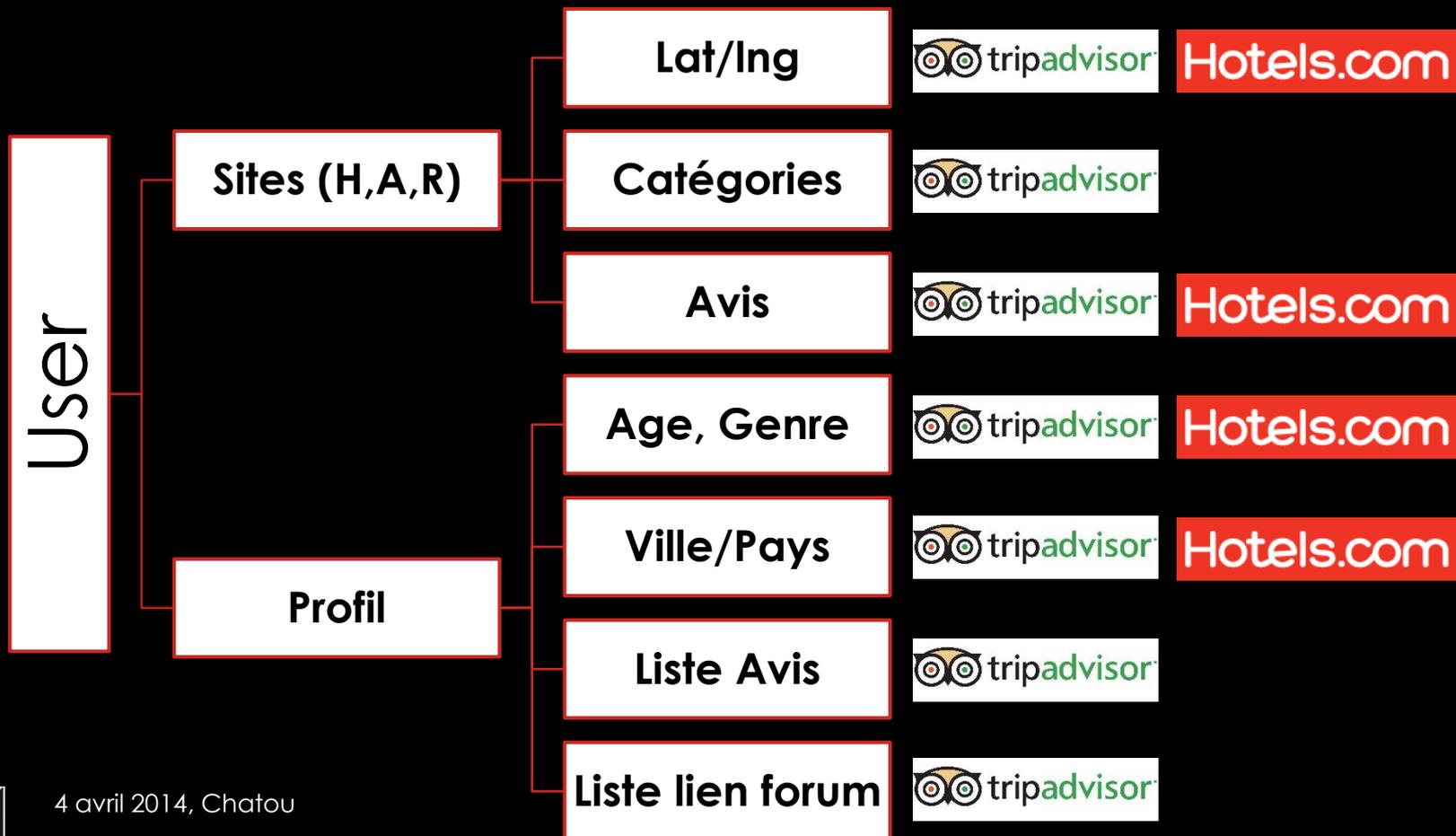
Hôtel Paradis Dernière réservation : il y a 2 heures

44 Rue Des Battois Enlève Paris, Paris, 75010 France, 01 57 32 43 16

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

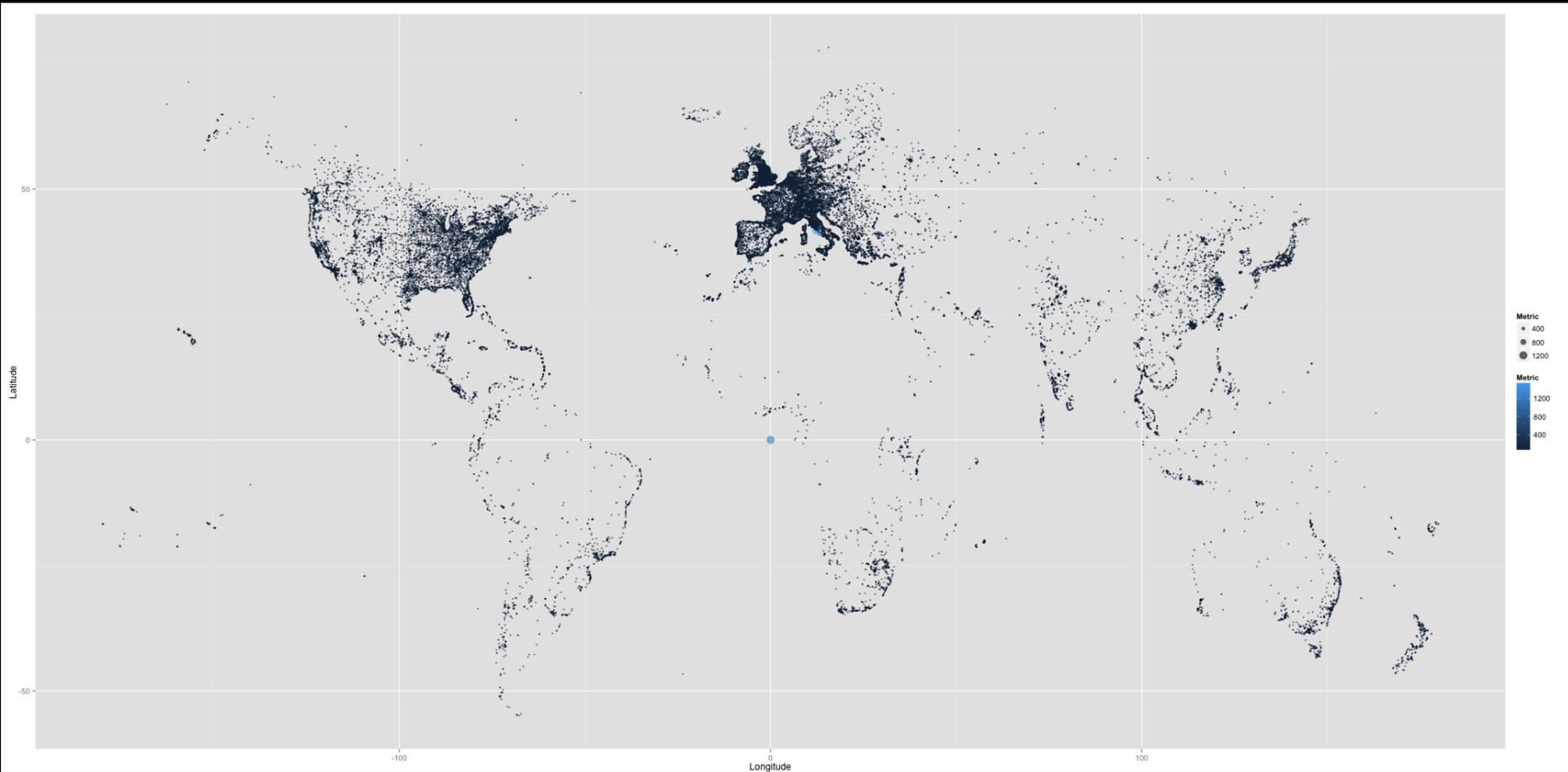
- Respect vie privée

Représentativité

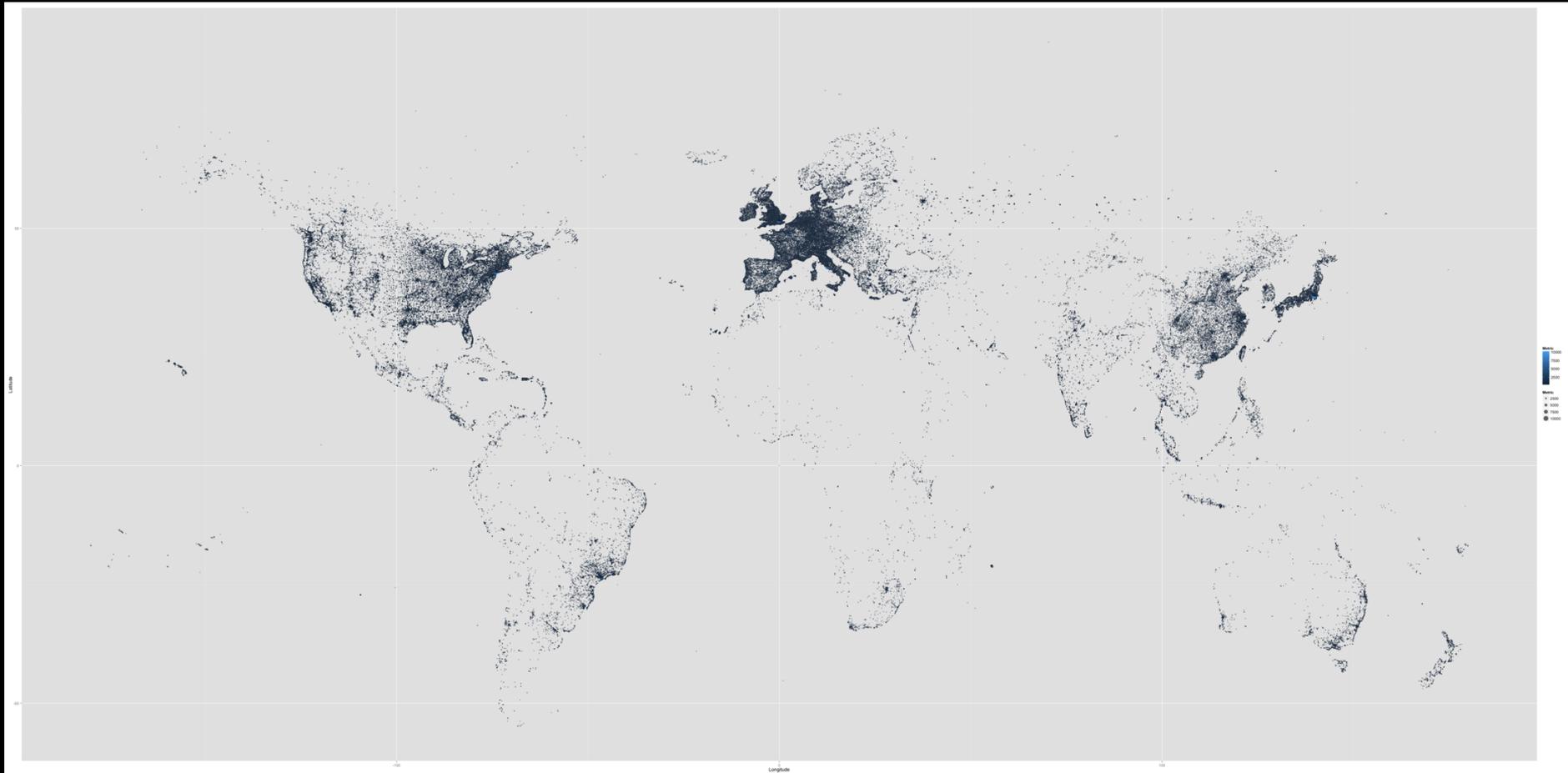
- Progression intrinsèque des réseaux sociaux
- Biais sociétaux
- Interprétation et intelligibilité des résultats

VISUALISATION DES DONNÉES

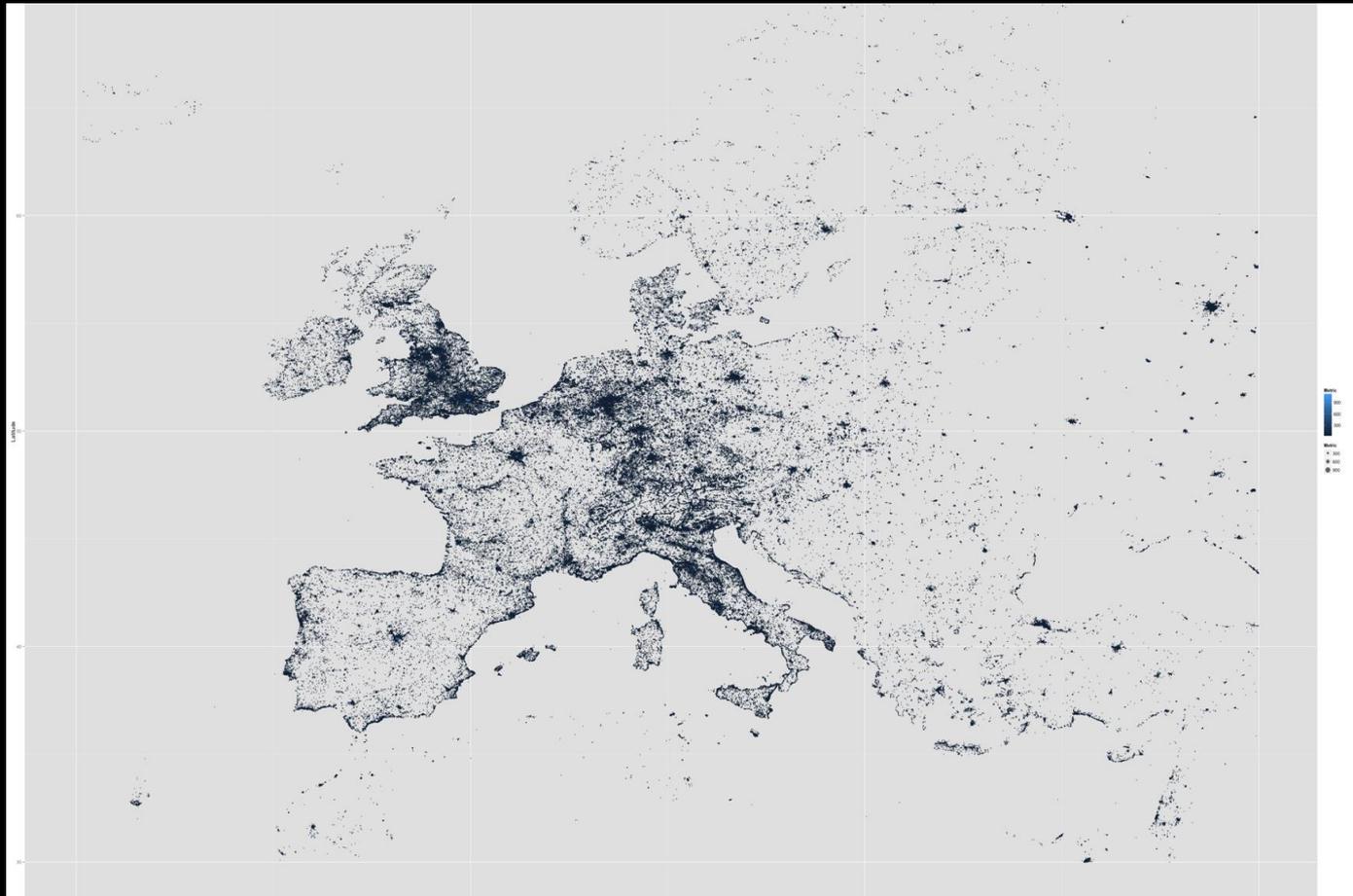
VISUALISATION (HOTEL.COM)



VISUALISATION (TRIPADVISOR)



VISUALISATION (TRIPADVISOR)



VISUALISATION DES DONNÉES GÉOLOCALISÉES

A heatmap visualization of geolocalized photo data overlaid on a city map. The map shows various districts and roads, with the heatmap using a color scale from green (low density) to red (high density) to indicate the concentration of photos. Three red rounded rectangular boxes are overlaid on the map, each containing a label in white text. The top box is labeled 'Tous les points Photo', the middle box is labeled 'Densité', and the bottom box is labeled 'Densité et diffusion'. The heatmap shows high density in the central urban areas, particularly around the Bois de Boulogne and the center of Paris.

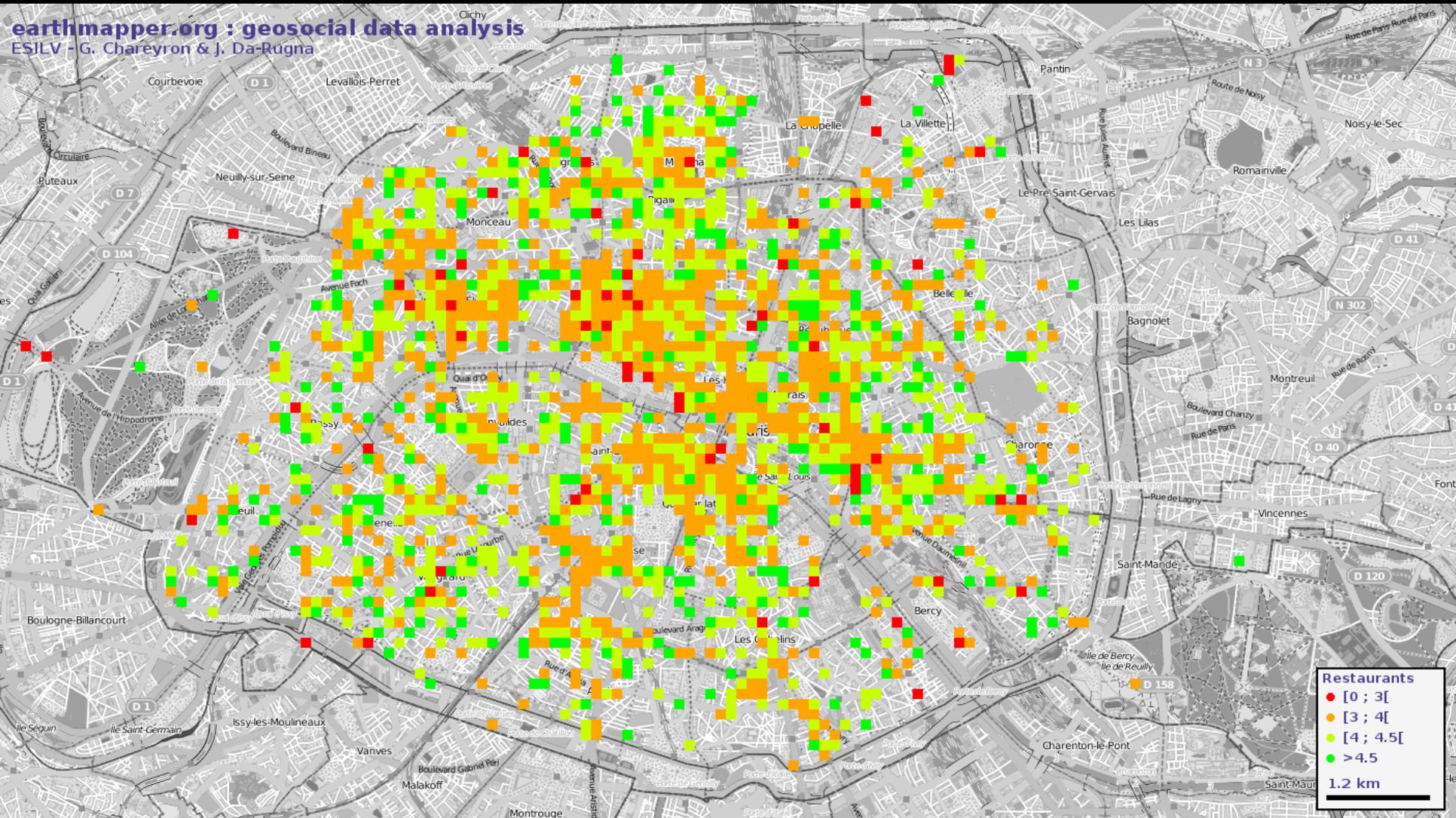
Tous les points Photo

Densité

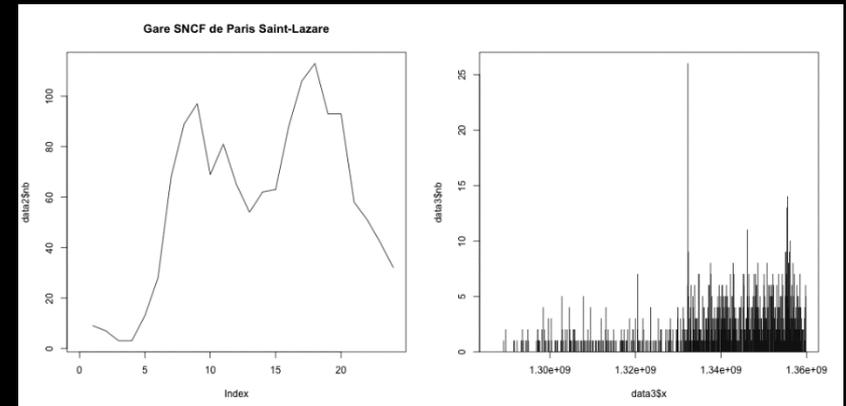
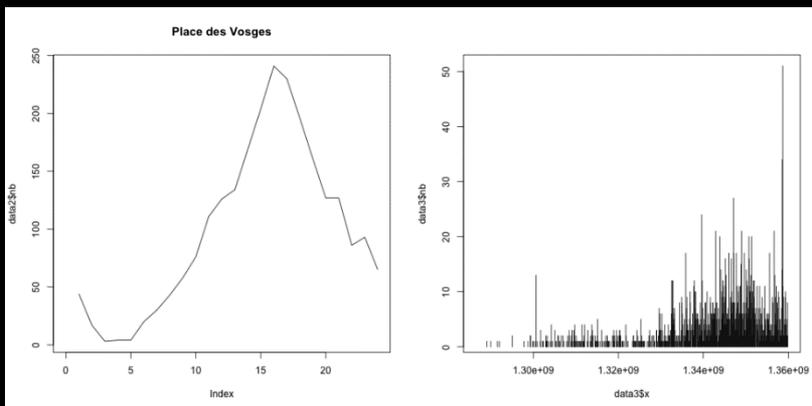
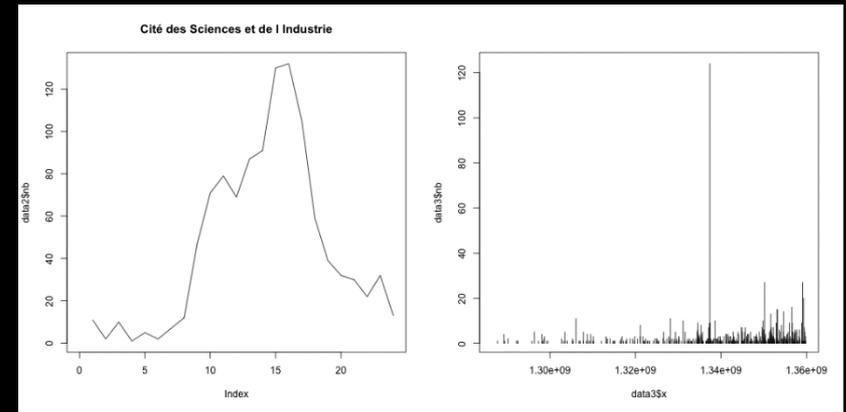
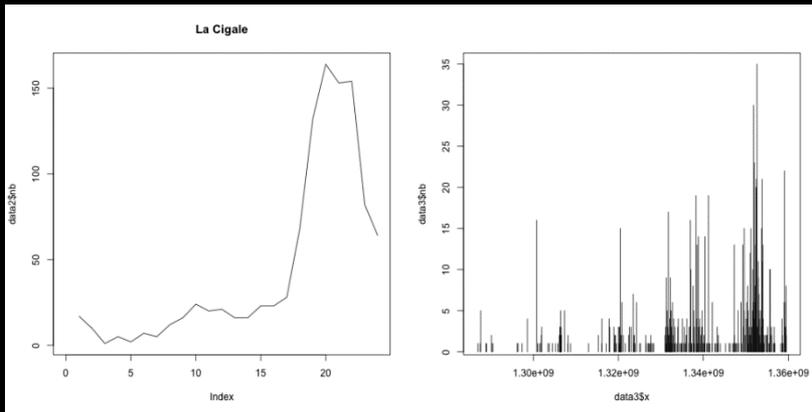
Densité et diffusion

VISUALISATION PAR ZONE

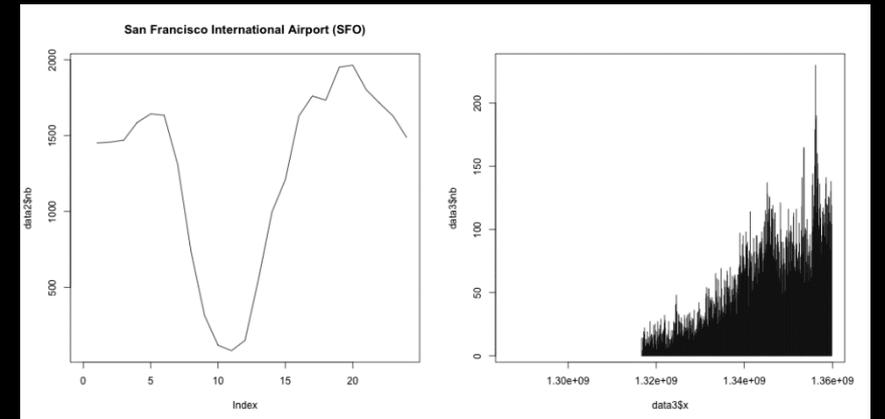
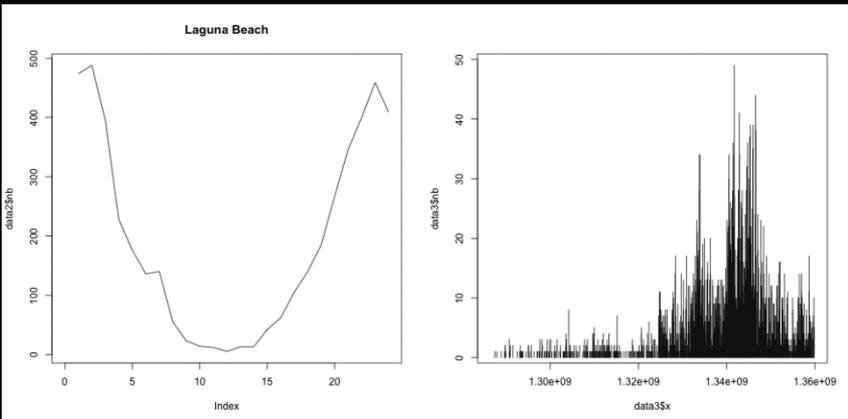
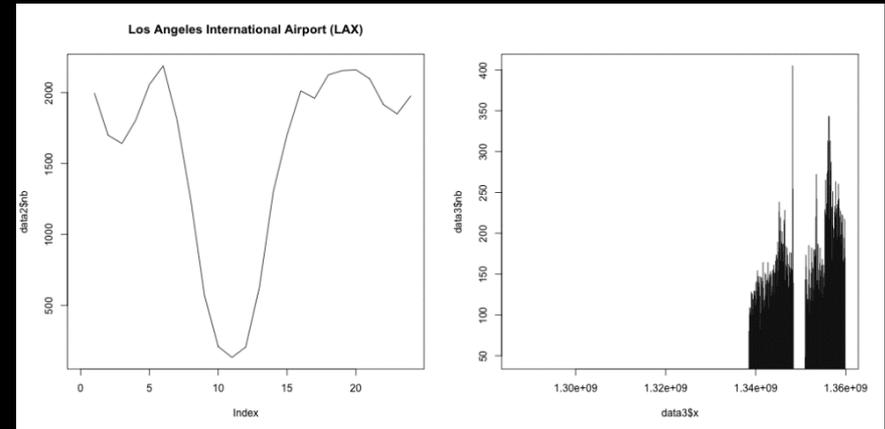
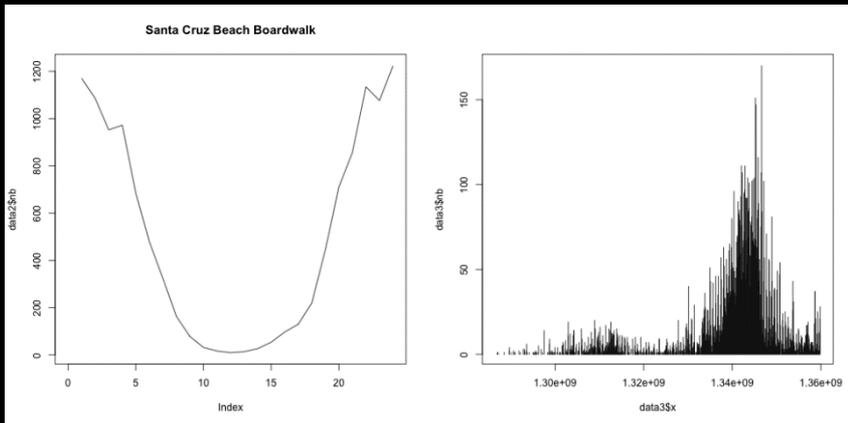
21



INSTAGRAM ET TEMPORALITÉ



INSTAGRAM ET TEMPORALITÉ



UN PEU D'INTELLIGENCE ?

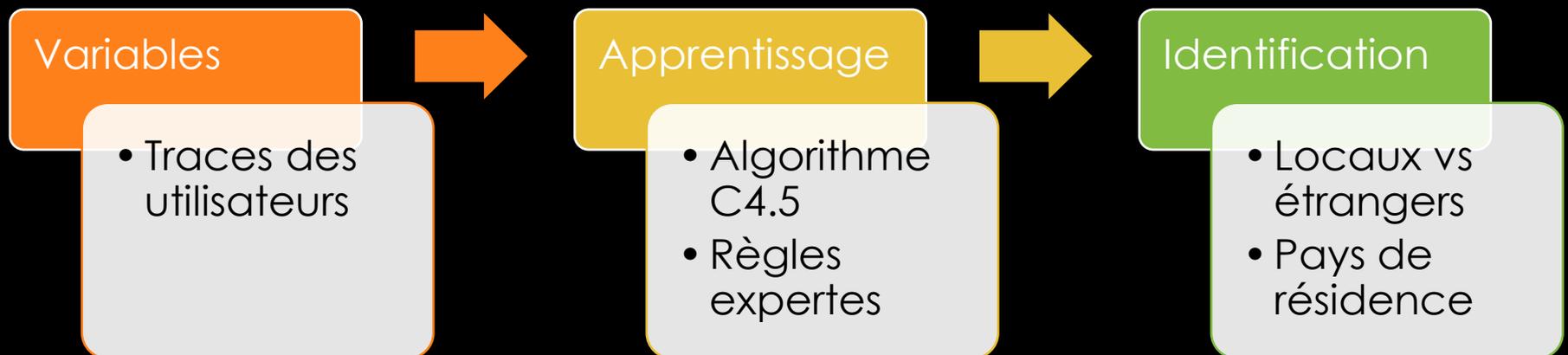
- Un potentiel immense, largement inexploité
- Détection automatique de comportements
 - typiques / atypiques
- Détection des évolutions, des transformations
- Mesurer autrement

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

ANALYSES DES DONNÉES

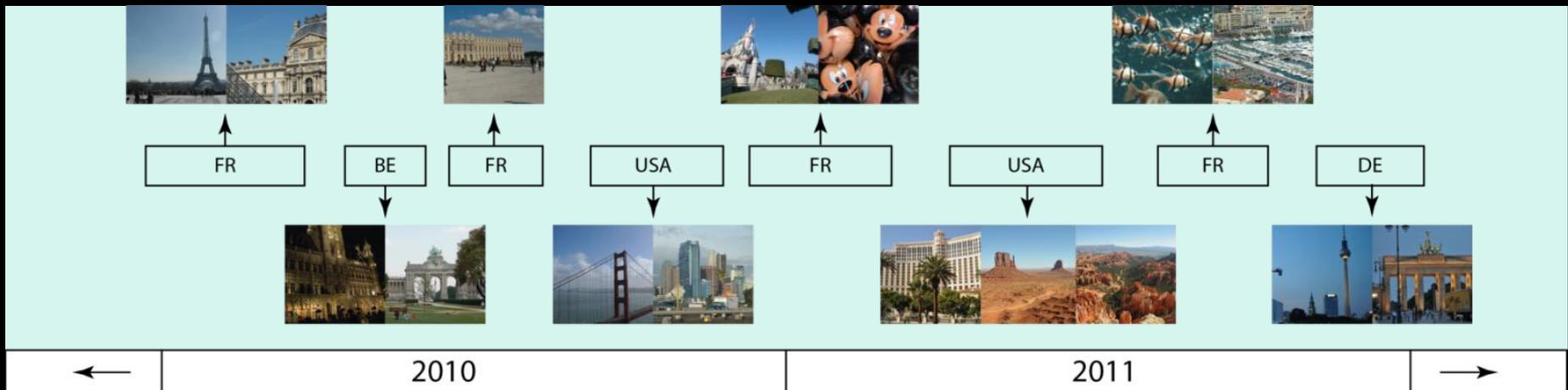
ENRICHISSEMENT DES PROFILS

- 30% des utilisateurs sont « localisés »
- Peut-on élargir cette localisation pour mieux connaître 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

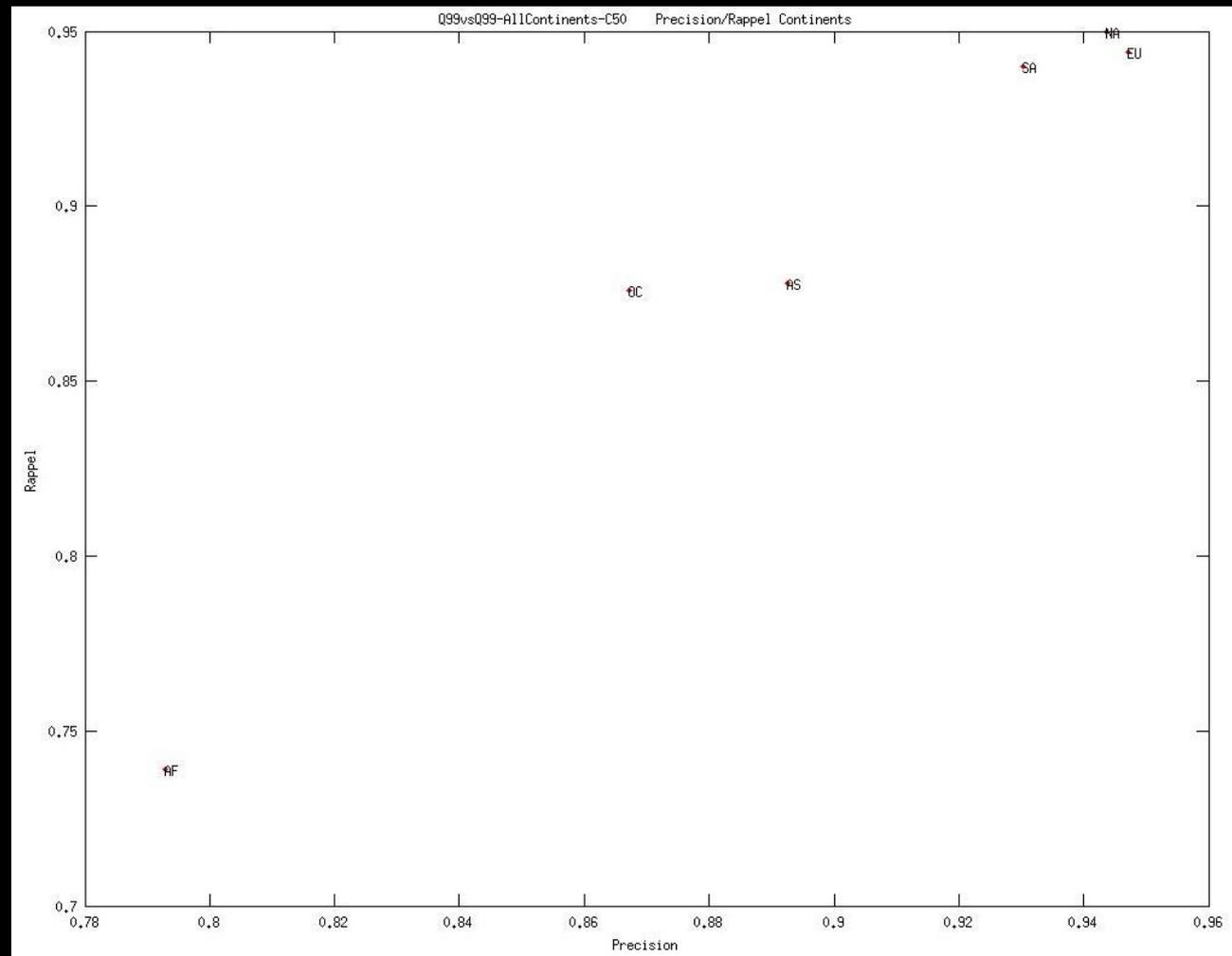
CLASSIFICATION

- 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

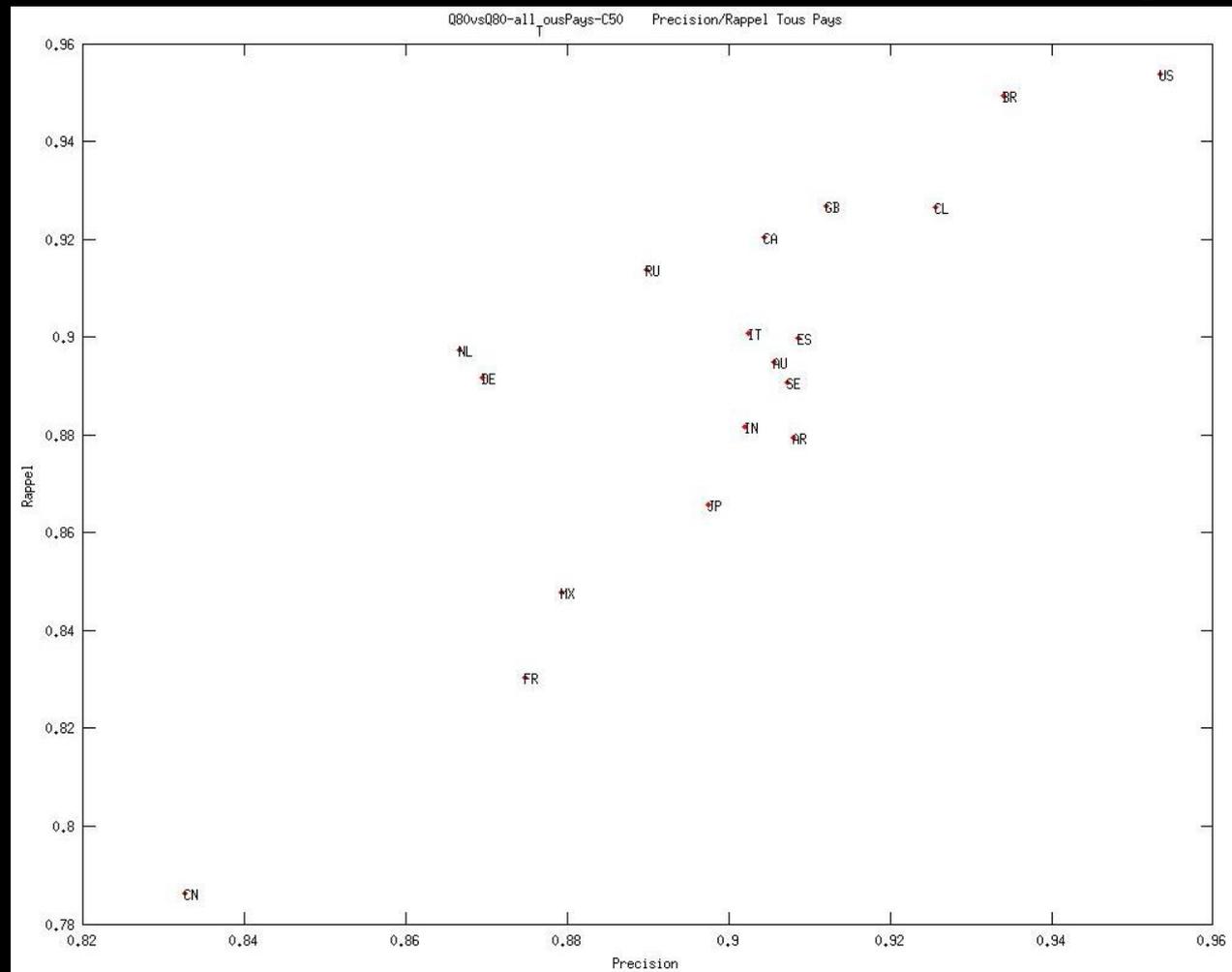
CLASSIFICATION PAR CONTINENTS



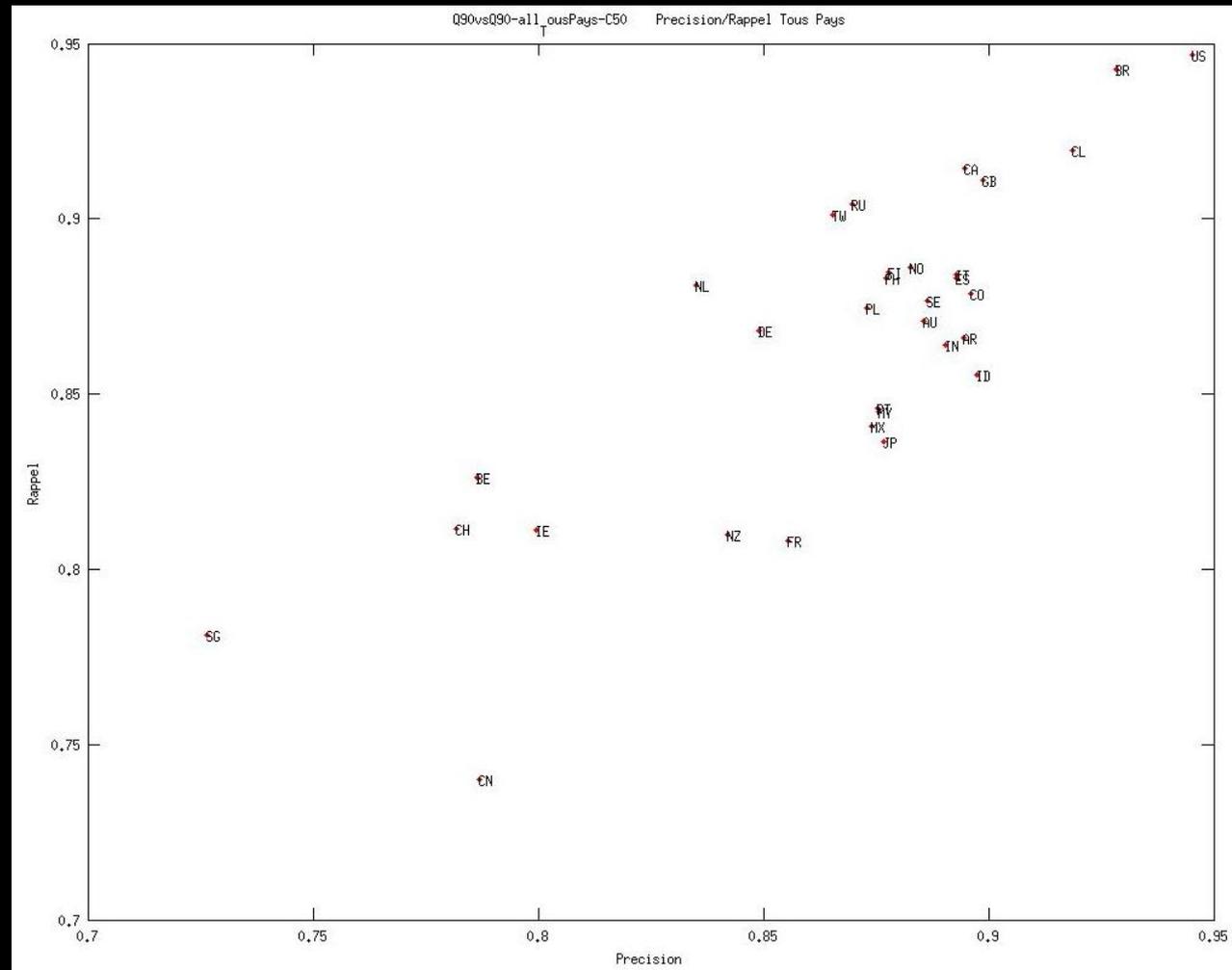
CLASSIFICATION PAR PAYS

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

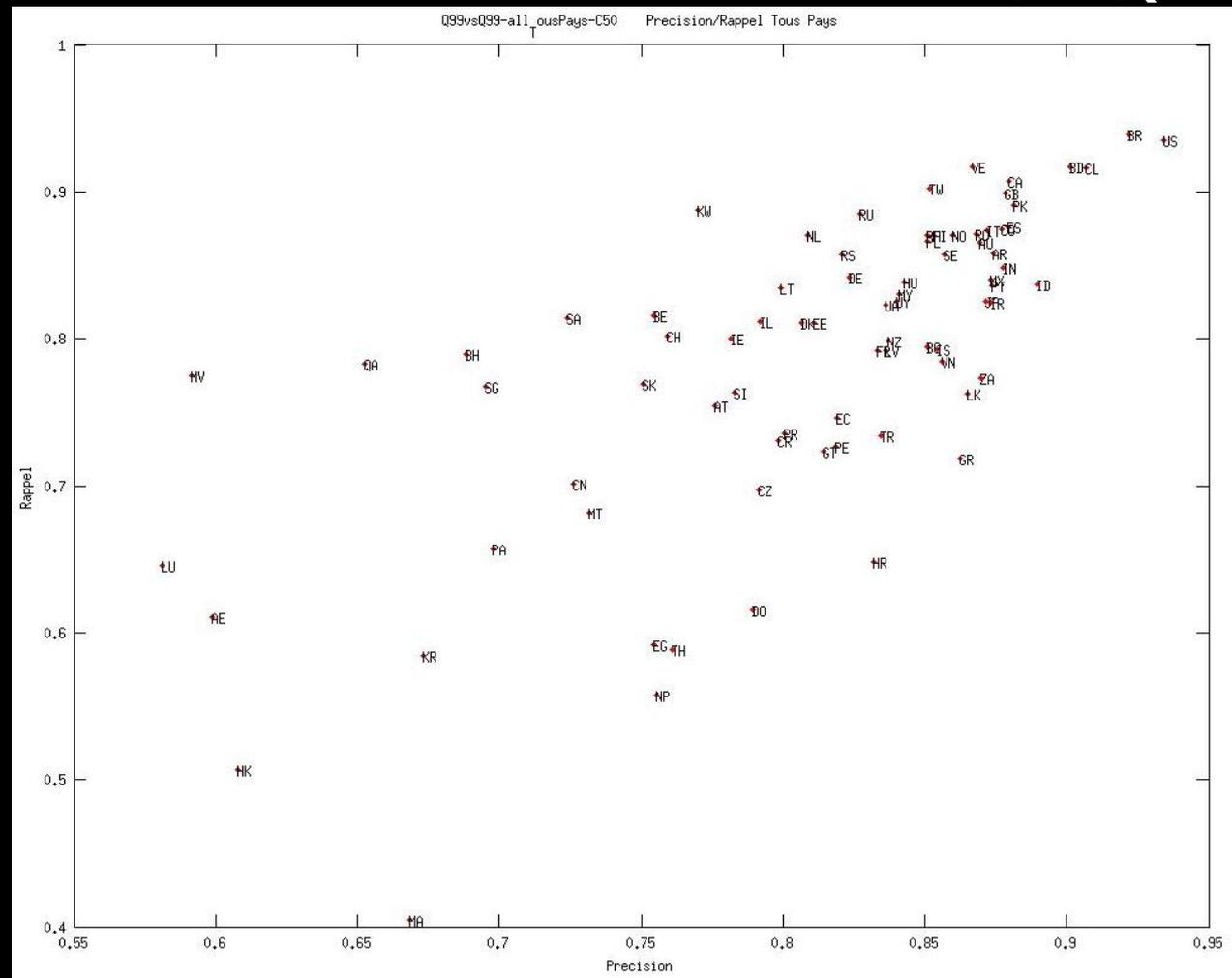
CLASSIFICATION PAR PAYS Q80



CLASSIFICATION PAR PAYS Q90



CLASSIFICATION PAR PAYS Q99

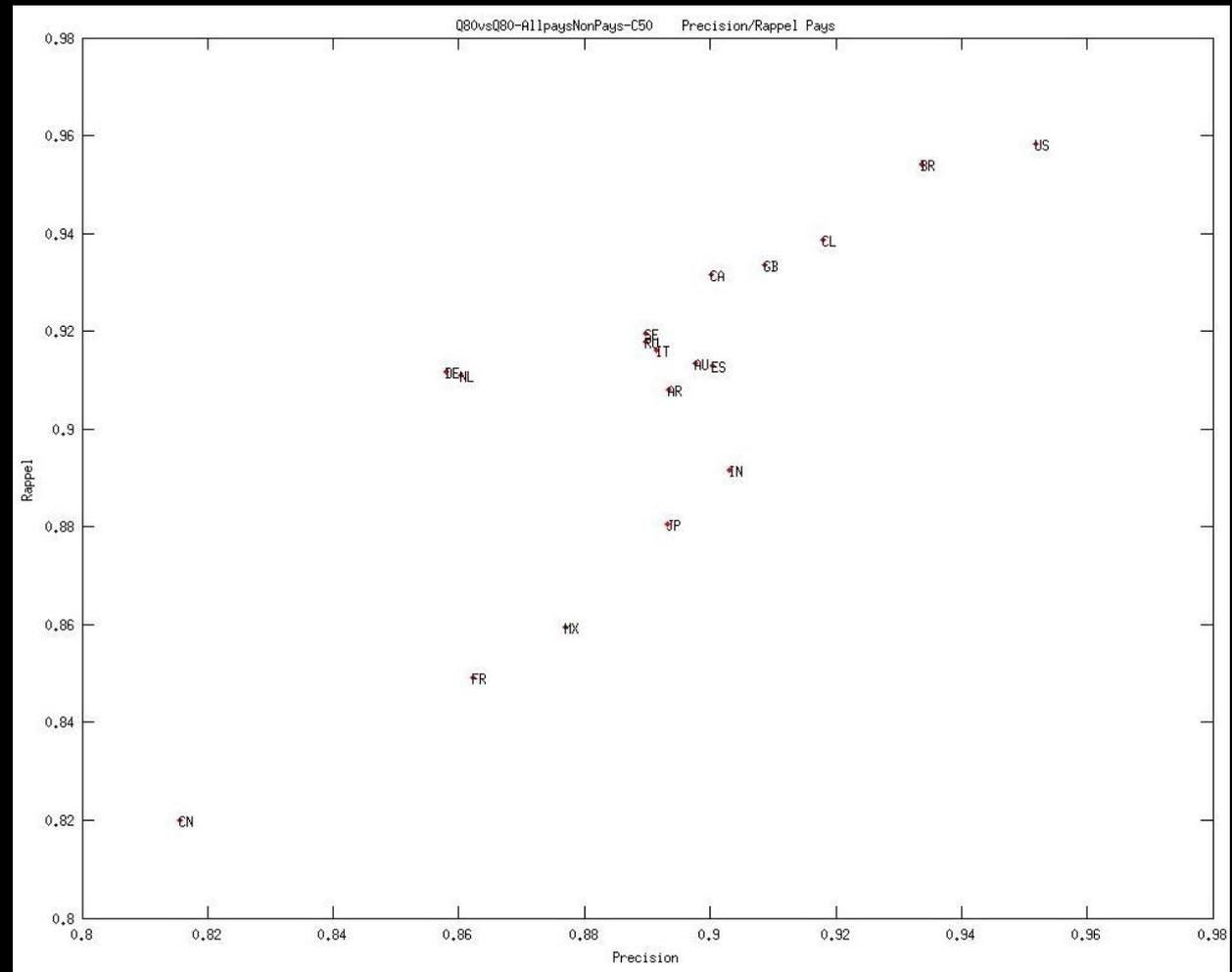


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

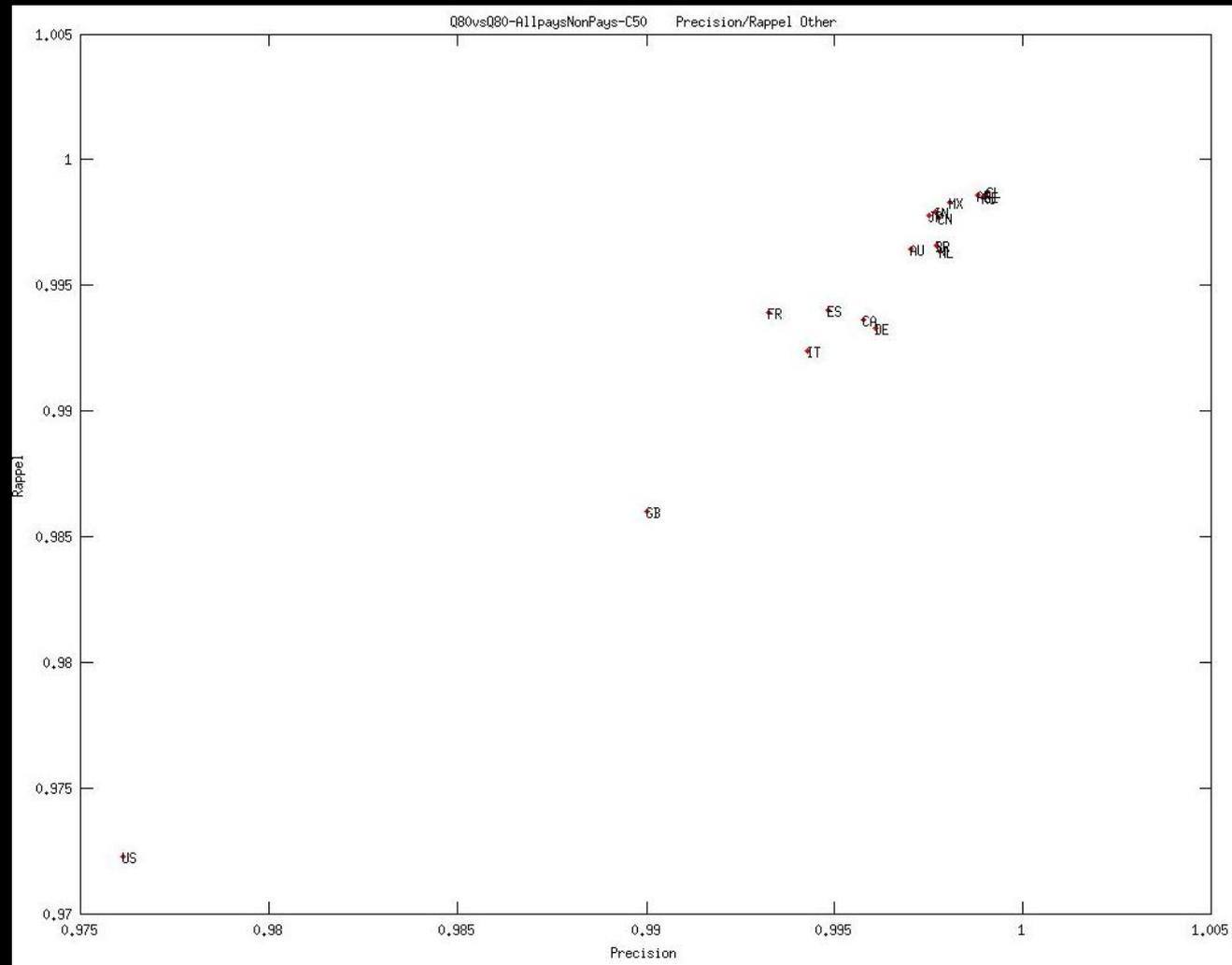
Précision-Rappel
Pour les « Locaux »



CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

Précision-Rappel
Pour les « Non-Locaux »

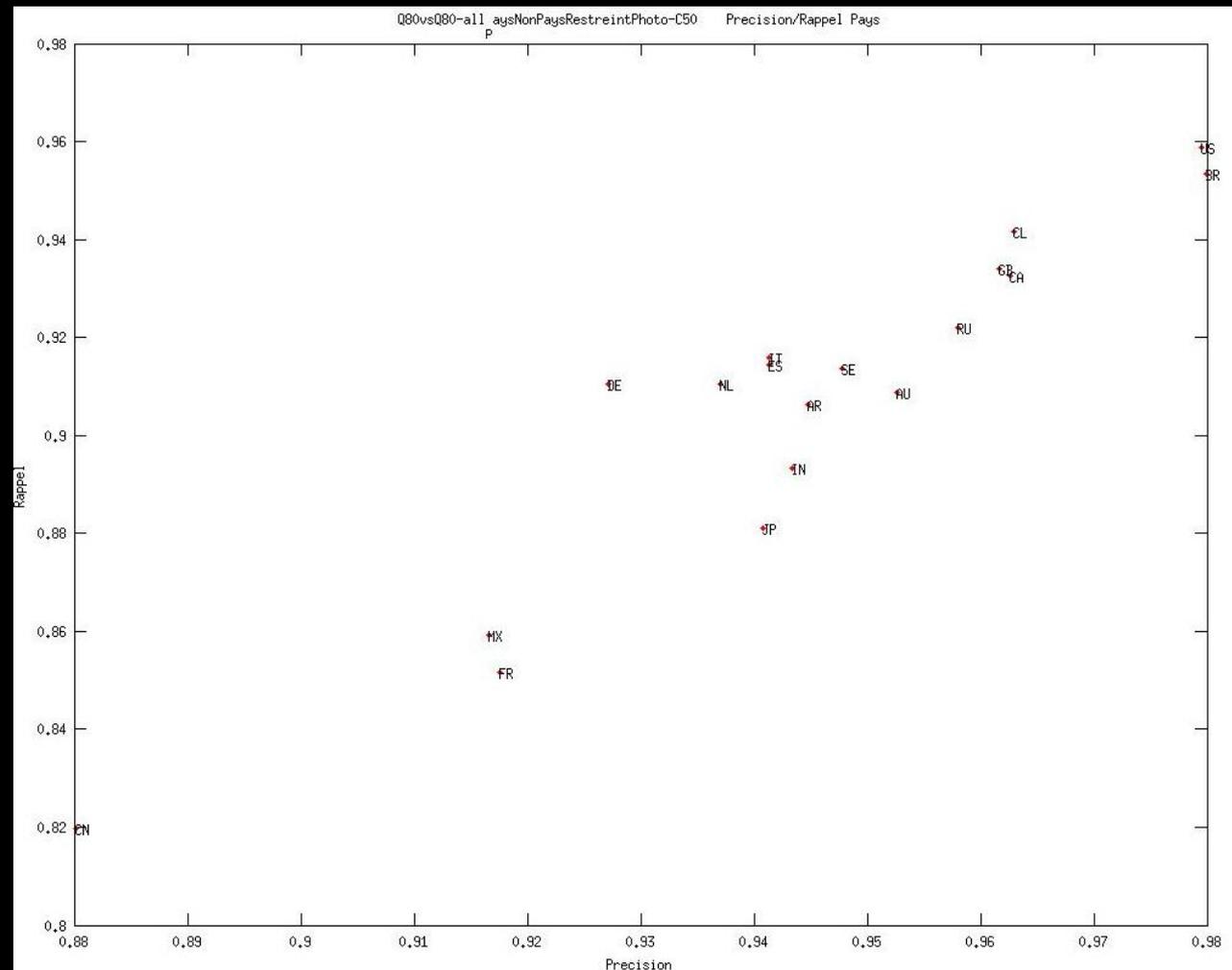


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

Précision-Rappel
Pour les « Locaux »

Restriction aux users
ayant pris des
photos dans le pays
considéré

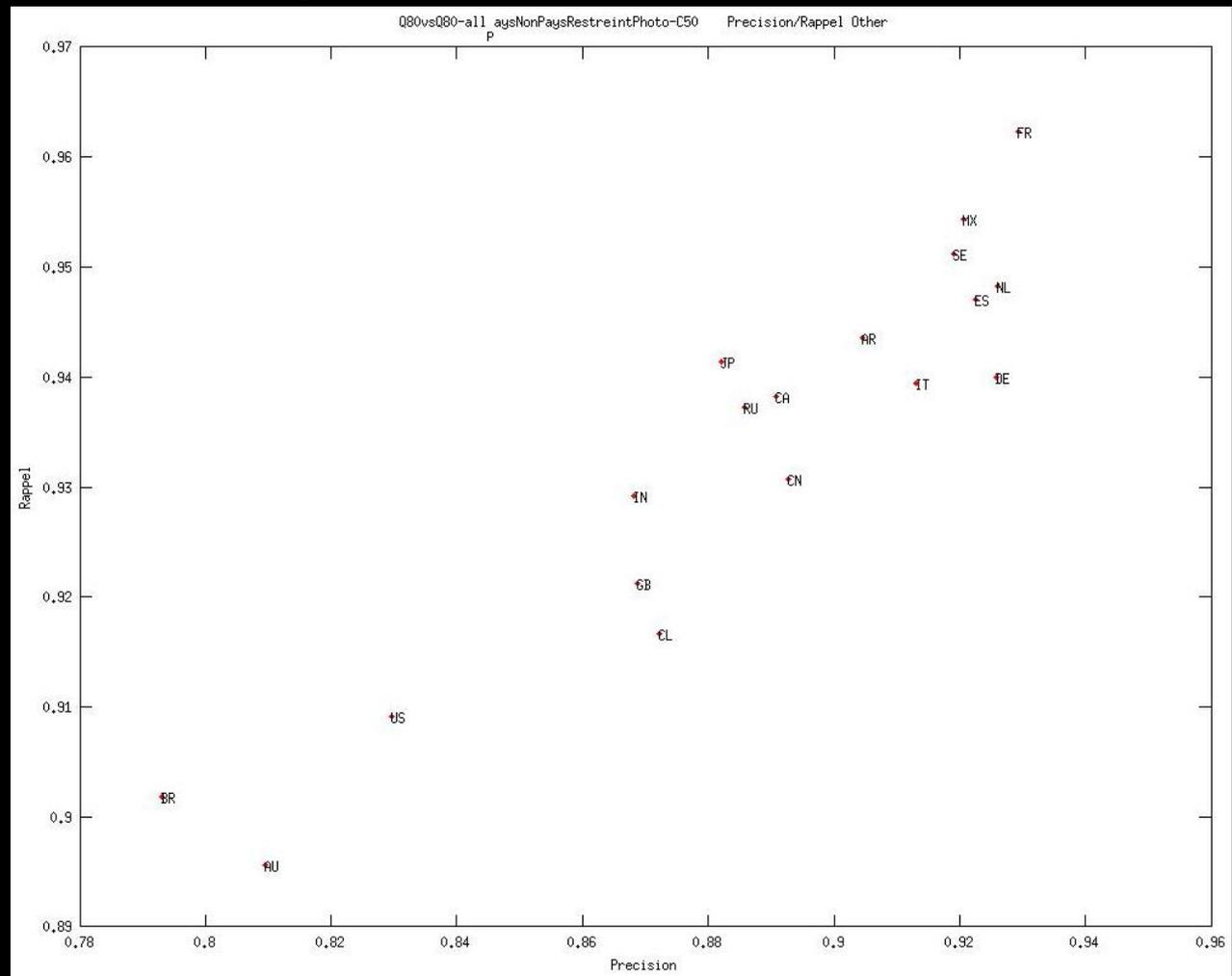


CLASSIFICATION « LOCAUX » / « NON-LOCAUX »

Q80

Précision-Rappel
Pour les «Non- Locaux »

Restriction aux users
ayant pris des photos
dans le pays considéré



CONCLUSION



Big Data

- Récolte
- Performance
- Modélisation
- Data mining



Interprétation

- Représentativité
- Confrontation
- Biais sociétaux



Innovation

- Nouveaux services
- Nouvelles analyses

CONCLUSION



MERCI DE VOTRE ATTENTION

<http://earthmapper.org>

<http://world.earthmapper.org>