Universidad de los Andes
Colombia

Facultad de Ingeniería

Departamento de Ingeniería
de Sistemas y Computación

Computación Visual
Imagine
Universidad de los Andes | Bogotá, Colombia

# Opening the black-box:

## Towards more Interactive and Interpretable Machine Learning

**Fabian C. Peña, John A. Guerra**

*fc.pena@uniandes.edu.co, ja.guerrag@uniandes.edu.co*

## The Problem

Most of the current real world Machine Learning (ML) systems use models as black-boxes. Given this, the big question that arises is: **What about if model performance is not the unique requirement to be fulfilled?** To address this, two new sub-fields of research have been proposed to help users to interact and understand ML models: **Interactive ML** and **Interpretable ML**.

## Interactive Machine Learning

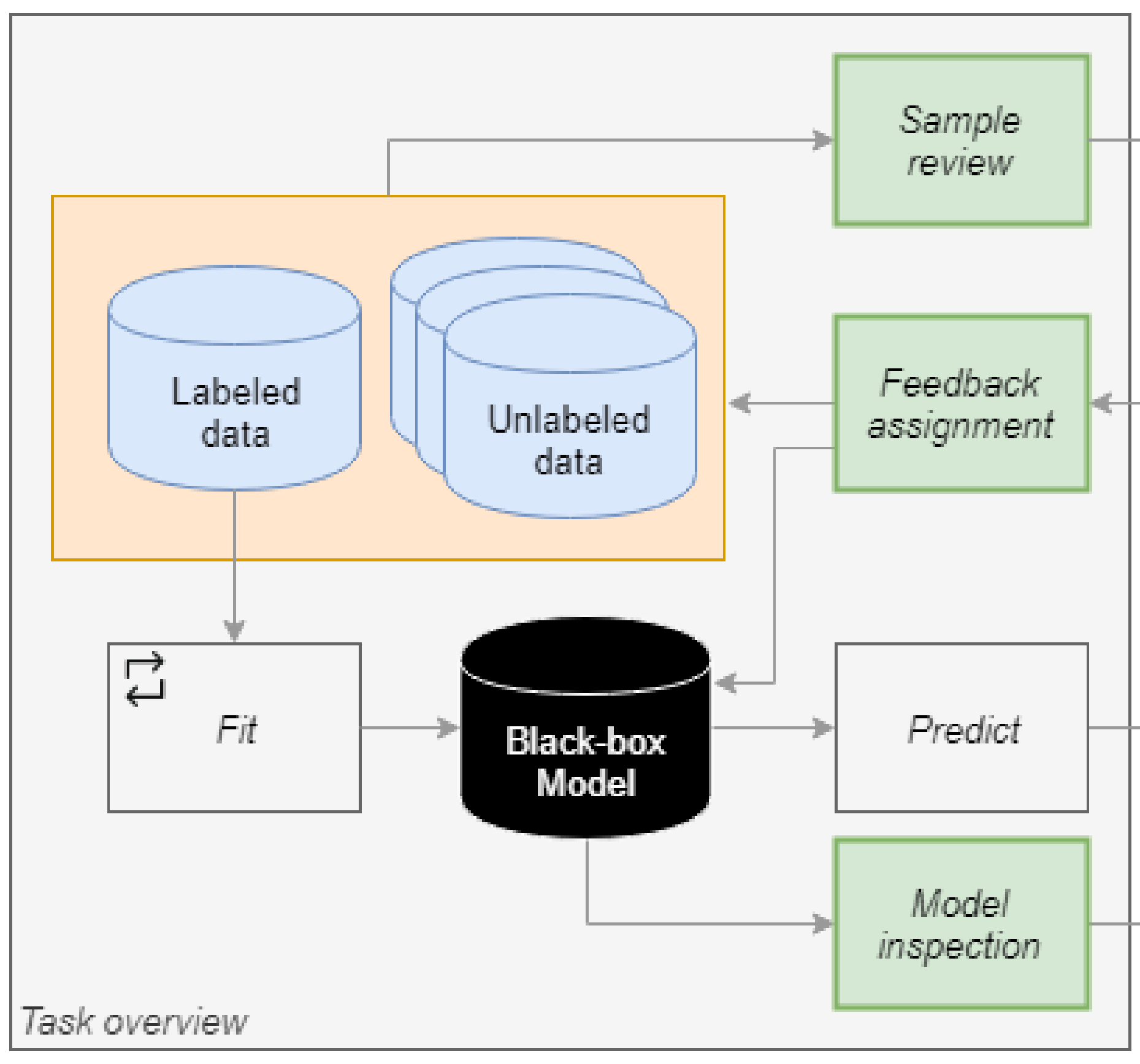


**Figure 1.** Elements to consider when designing Interactive ML systems. **Sample review**, **feedback assignment**, **model inspection** and **task overview** (Dudley, 2018).

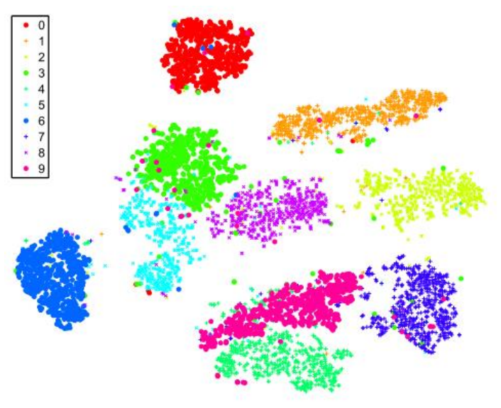


**Figure 2.** Dimensionality Reduction for **sample review**. E.g. t-SNE (Van der Maaten, 2008).

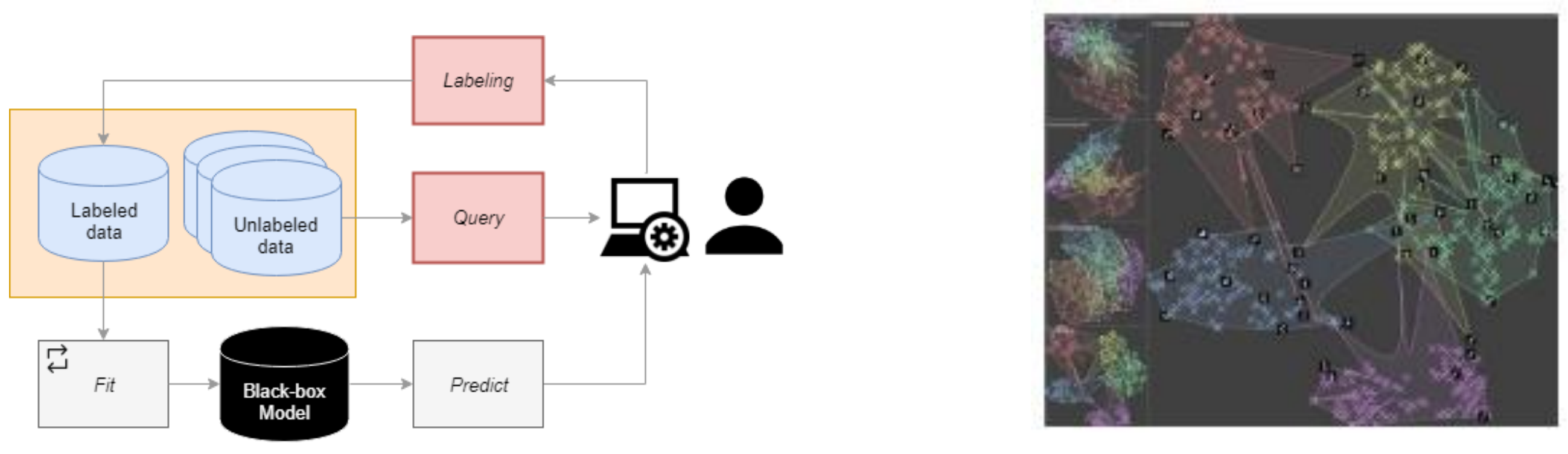


**Figure 3.** Active Learning (left) and Visual Interactive Labeling (right) for **feedback assignment** (Bernard, 2017).

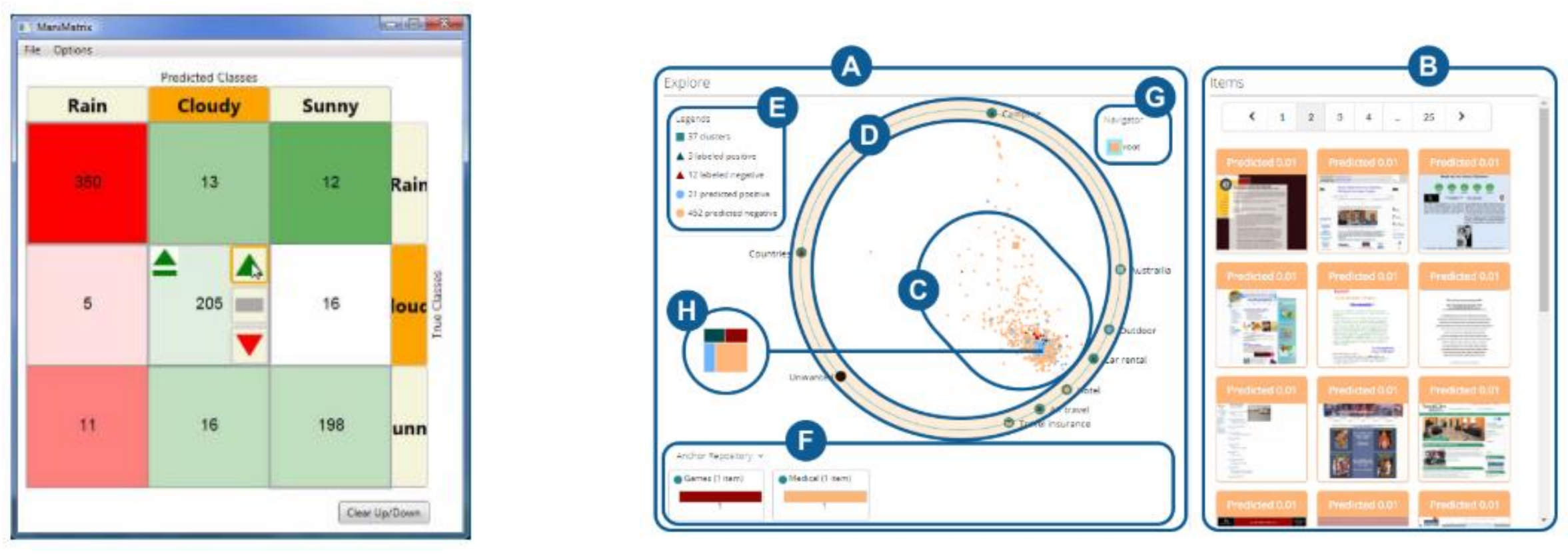
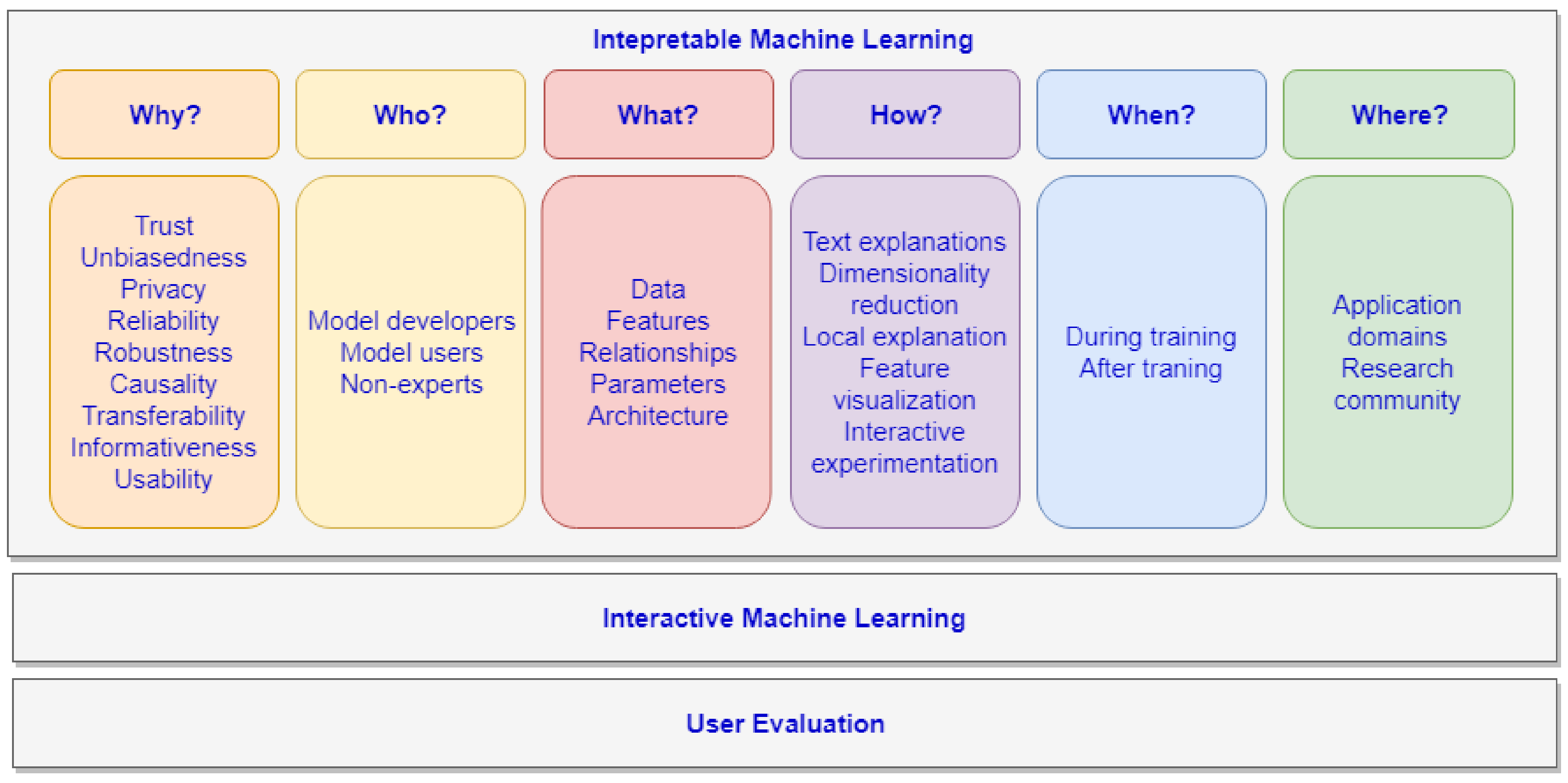


**Figure 4.** Parameter tuning (left) and Error discovery (right) for **model inspection** (Kapoor, 2010 and Chen, 2018).

## Interpretable Machine Learning



Intepretable Machine Learning

| Why? | Who? | What? | How? | When? | Where? |
|---|---|---|---|---|---|
| Trust<br>Unbiasedness<br>Privacy<br>Reliability<br>Robustness<br>Causality<br>Transferability<br>Informativeness<br>Usability | Model developers<br>Model users<br>Non-experts | Data<br>Features<br>Relationships<br>Parameters<br>Architecture | Text explanations<br>Dimensionality reduction<br>Local explanation<br>Feature visualization<br>Interactive experimentation | During training<br>After traning | Application domains<br>Research community |

Interactive Machine Learning

User Evaluation

**Figure 5.** Aspects to consider when designing Interpretable ML systems (Hohman, 2018, Doshi-Velez, 2017 and Lipton, 2016).

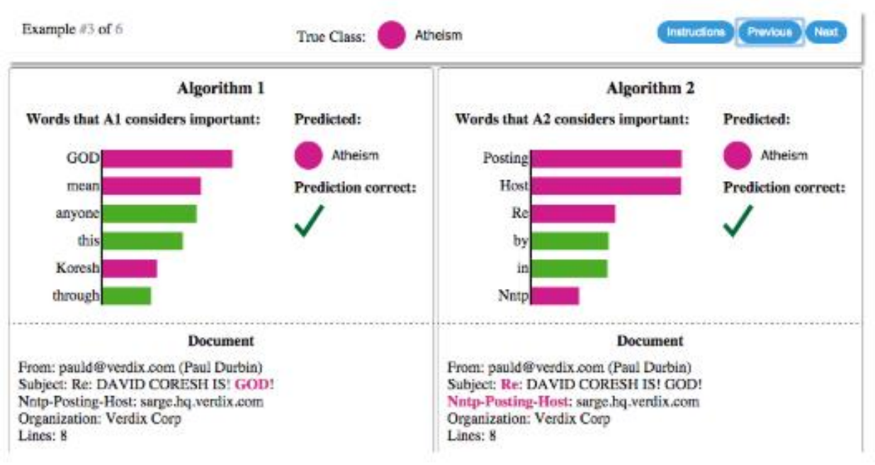
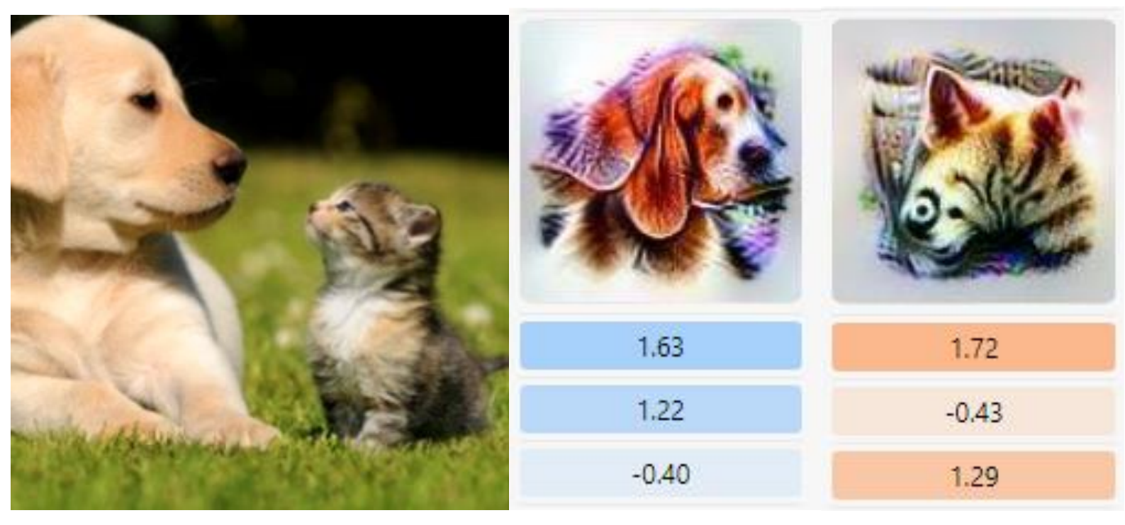


**Figure 6.** Local interpretability with LIME (Ribeiro, 2016).

**Figure 7.** Feature visualization in combination with attribution (Olah, 2018).



## Our Current Work

LAC-URBAN HEALTH
*Urban Health Network for Latin America and the Caribbean*

**SALURBAL Case Sudy**

Working in contexts related to town planning and public health and data from Latin American cities, we want to produce interpretability for Clustering and Dimensionality Reduction algorithms in an interactive way.

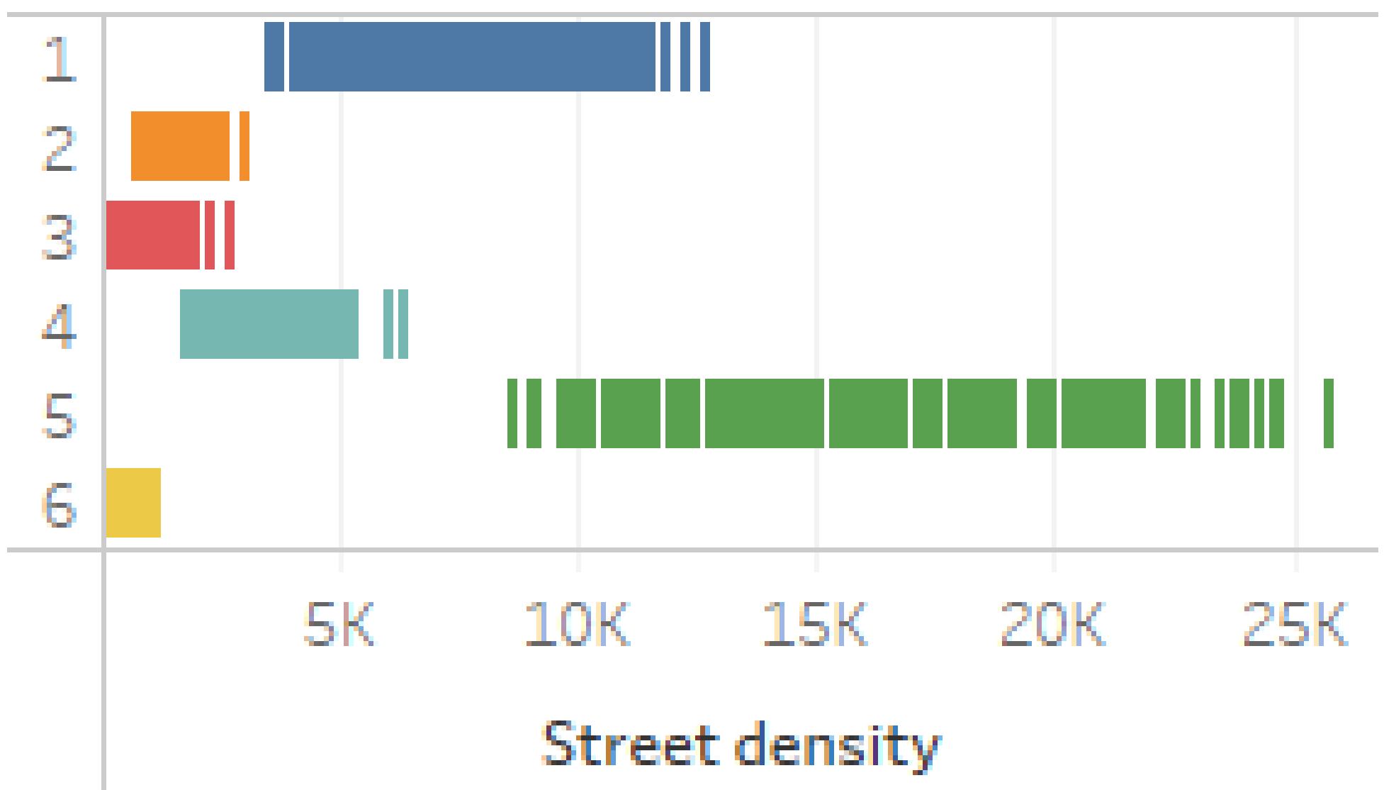


**Figure 8.** Validating feature distribution for each profile identified by Clustering algorithms.

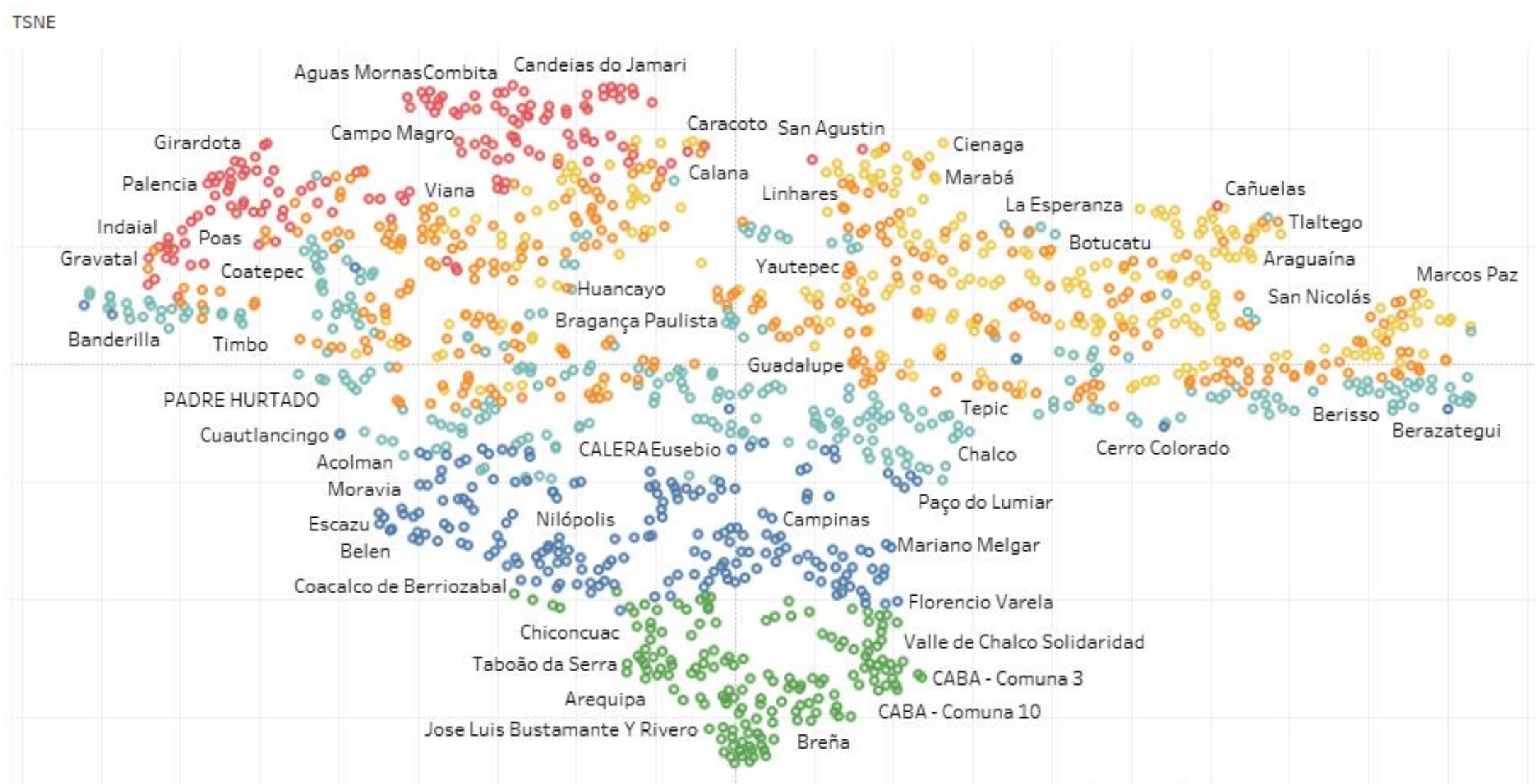
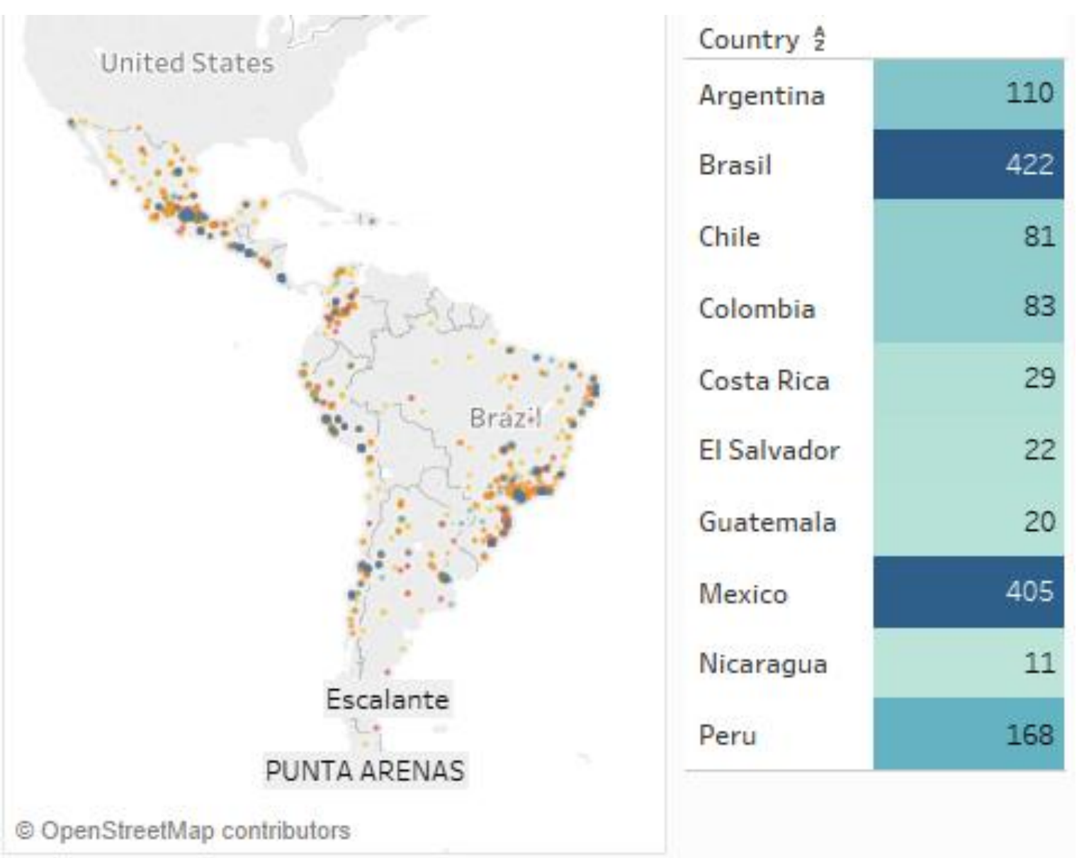


**Figure 9.** Applying Dimensionality Reduction to show the overall distribution of cities.



| Country |  |
|---|---|
| Argentina | 110 |
| Brasil | 422 |
| Chile | 81 |
| Colombia | 83 |
| Costa Rica | 29 |
| El Salvador | 22 |
| Guatemala | 20 |
| Mexico | 405 |
| Nicaragua | 11 |
| Peru | 168 |

**Figure 10.** Visualizing complementary features to enable more enriched insights.

http://sistemas.uniandes.edu.co /DISCUniandes /DISCUniandes