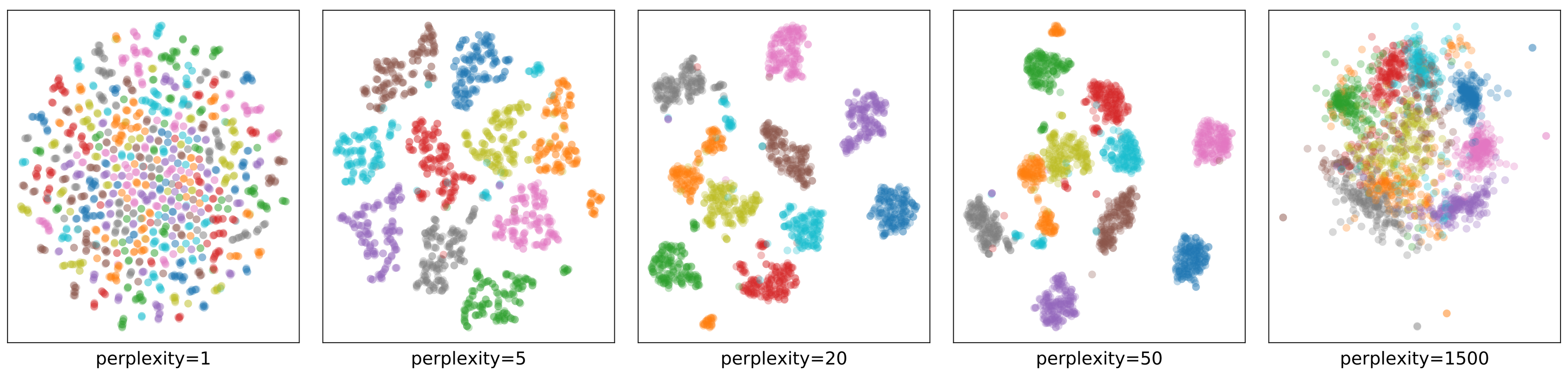
Tuning of Visualization Algorithms with User Constraints for t-SNE

Viet Minh Vu, Adrien Bibal, Benoît Frénay

Difficulty in choosing a good parameter for a visualization algorithm (t-SNE)

Problematic and Motivation

Goal: Visualize the **high dimensional** data



t-SNE is sensitive to the *perplexity* parameter, which is **important** but very **hard to understand and to tune**.

Proposed Solution

Use the users' **feedback** to steer the visualization.

- Let users define their requirements in form of **pairwise constraints**[A] between examples.
- The *perplexity* is automatically chosen based on the user's **constraint scores**[B].
- Evaluate the proposed visualization in quantitative comparison with the state-of-the-art **quality metrics**[C].

User pairwise constraints [A]

What are expressed by user (in high dim.)

Two similar examples → **Must link**.
Two dissimilar examples → **Cannot link**.

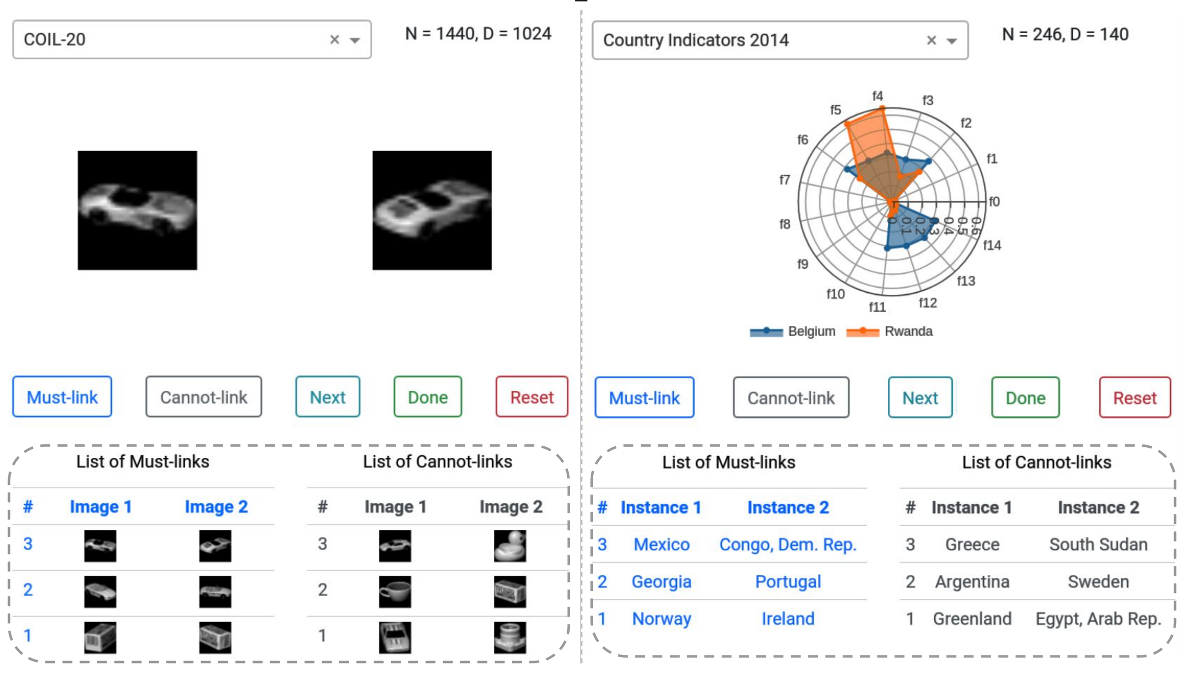


Figure: The interface for collecting users' feedback for image and tabular data.

What are translated to the algorithm (in low dim.)

Points connected by a **Must link** ($\mathcal{M}$) → must stay **close together**.
Points connected by a **Cannot link** ($\mathcal{C}$) → must stay **far apart**.

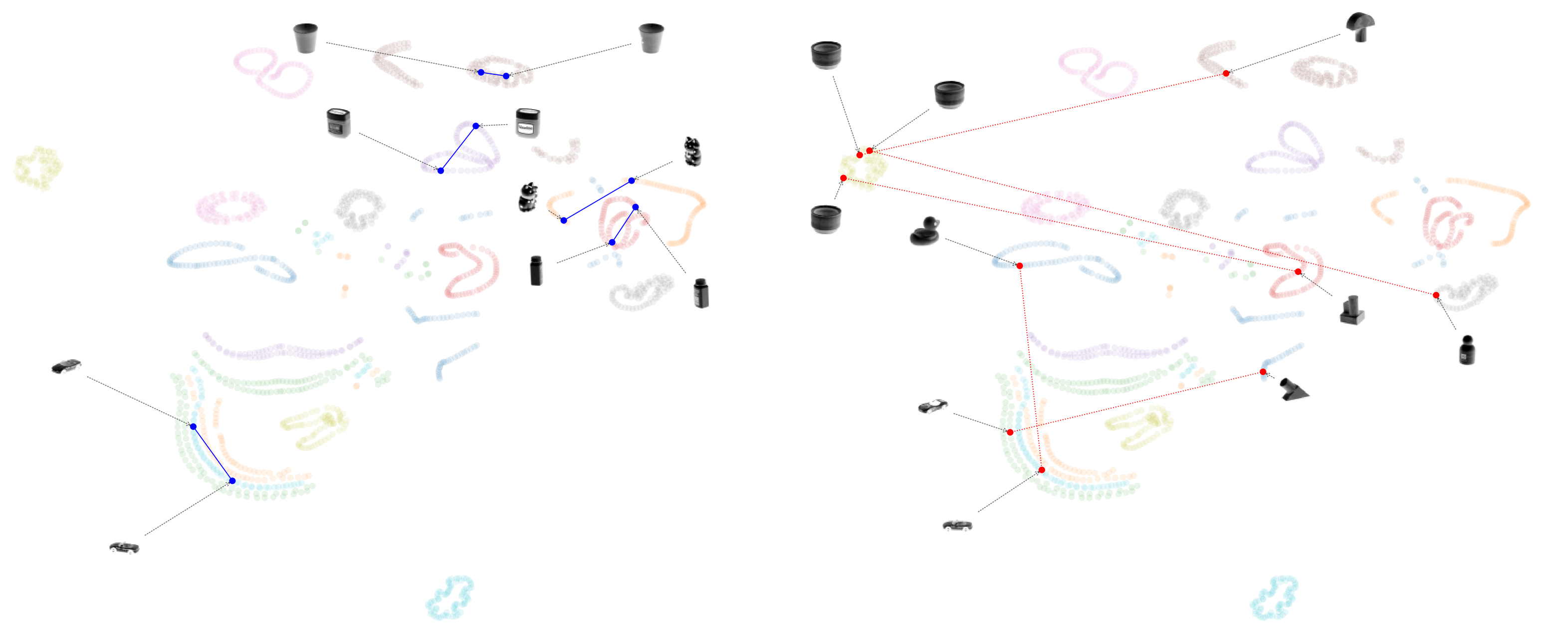


Figure: The user constraints in the visualization of the COIL20 dataset.

Constraint-Preserving Scores [B]

- Consider the points in the visualization (low dim.)
- q_{ij} = probability of i and j being neighbors.
- $S_{\mathcal{M}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \log q_{ij}$.
- $S_{\mathcal{C}} = -\frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} \log q_{ij}$.
- $S_{\mathcal{M}+\mathcal{C}} = S_{\mathcal{M}} + S_{\mathcal{C}}$.

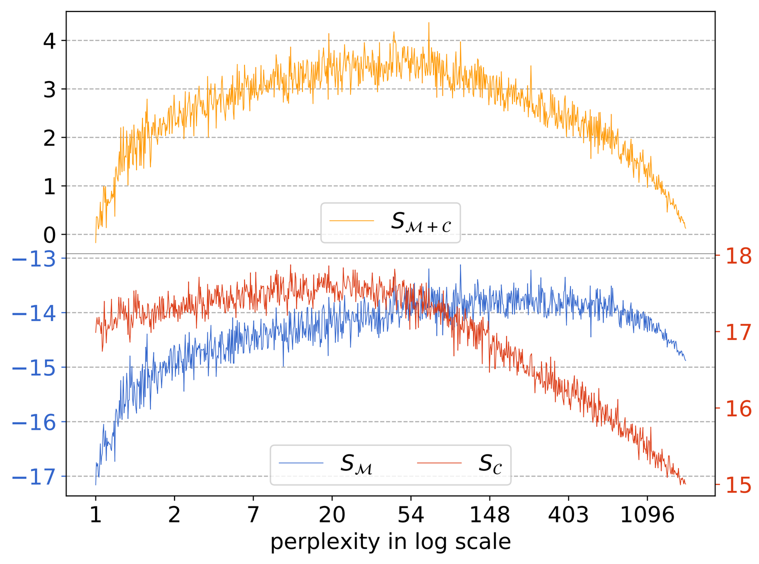


Figure :Constraint-preserving scores with 50 constraints for MNIST dataset.

☞ Can easily find the perplexity that maximizes $S_{\mathcal{M}}$, $S_{\mathcal{C}}$ or $S_{\mathcal{M}+\mathcal{C}}$.

Quality Metrics [C]

- **CC**: Pearson corr. coeff.
- **NMS**: Stress of pairwise distance orders comparison
- **CCA**: Stress with accent put on low dim.
- **NLM**: Stress with accent put on high dim.
- **AUC$_{log}$RNX**: How neighbors in high dim. are preserved in low dim.

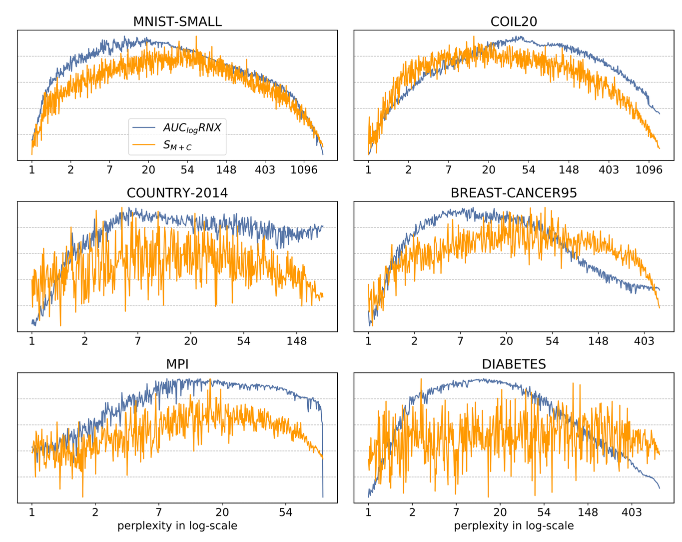
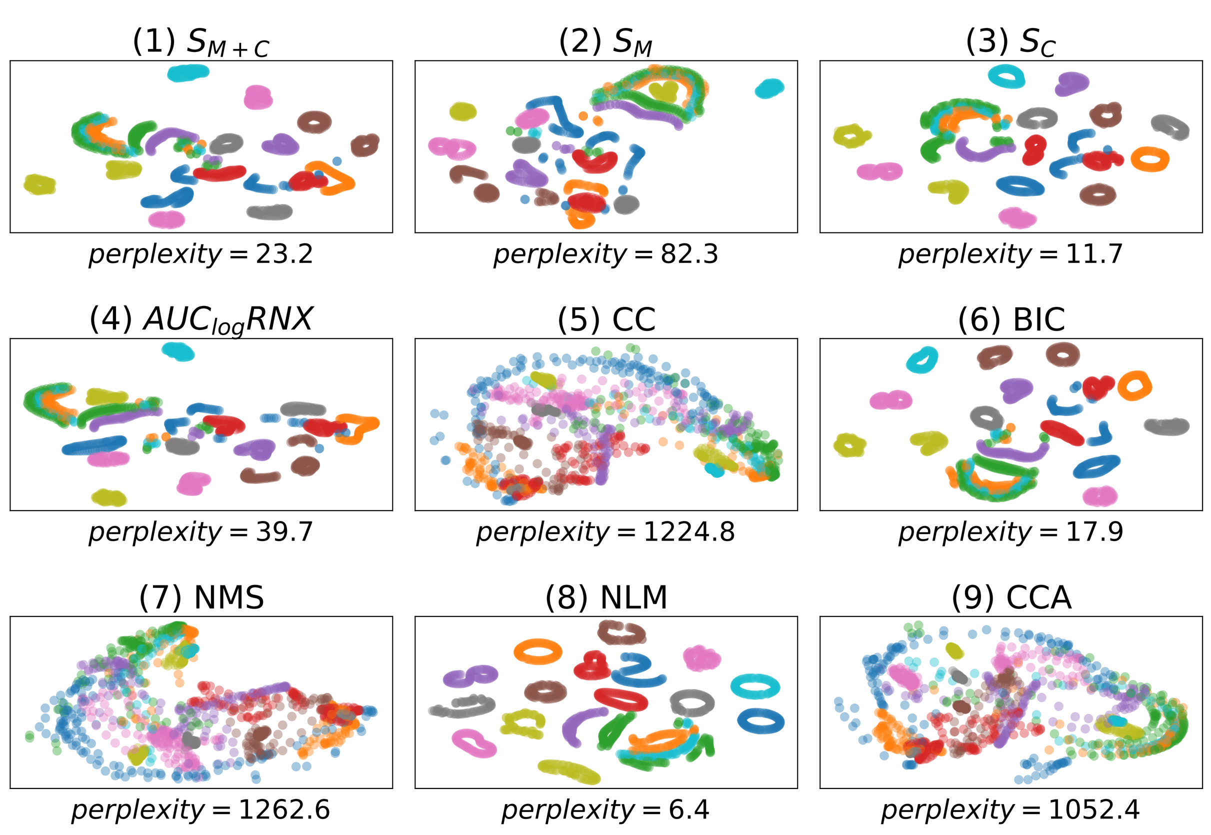
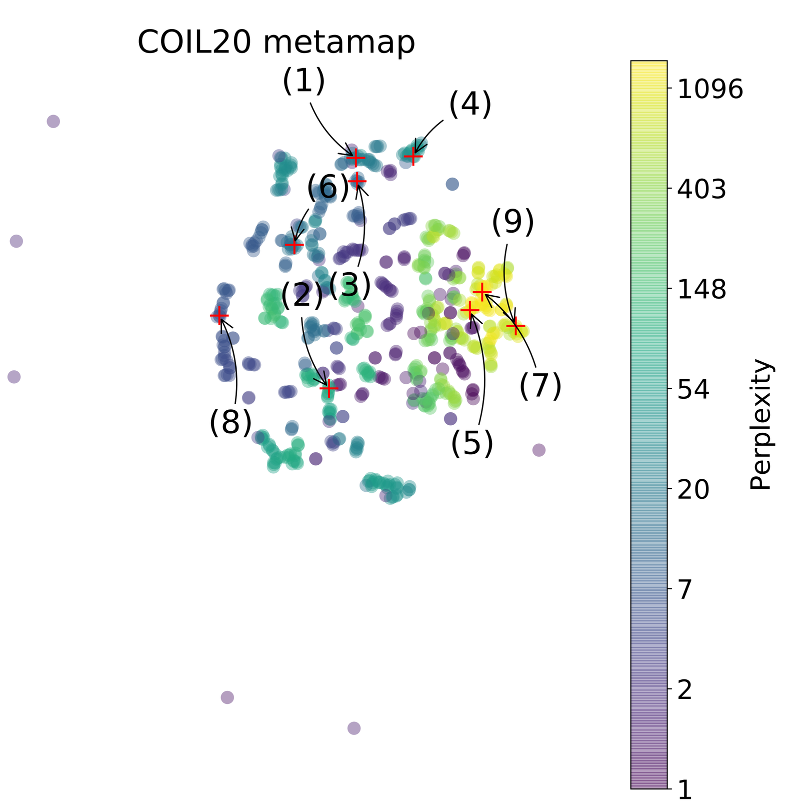


Figure: Compare $S_{\mathcal{M}+\mathcal{C}}$ and **AUC$_{log}$RNX** for six datasets.

☞ Our constraint scores agree with the quality metrics.

All visualizations in one place: Meta-plot





Conclusion

✓ Consider *user knowledge* under the form of constraints to find the most suitable visualization.

✓ Make complex visualization technique (t-SNE) *accessible* to users by freeing them from the tedious task of selecting the hyper-parameter.

✗ *Heavy computation* due to the pre-calculation of many possible embeddings.

Viet Minh Vu, Adrien Bibal and Benoît Frénay
✉ : *{vuvietminh, adrien.bibal, benoit.frenay}@unamur.be*