Plotly.plus an Improved Dataset for Visualization Recommendation

Luca Podo, *Sapienza University of Rome*
Paola Velardi, *Sapienza University of Rome*
Bardh Prenkaj, *Sapienza University of Rome*

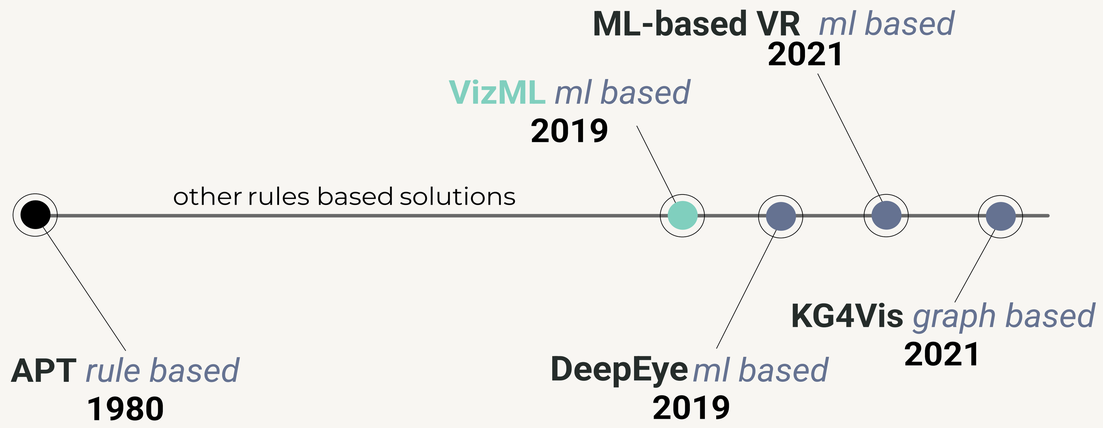
SAPIENZA UNIVERSITÀ DI ROMA

DATASET

PRESENTATION

Visualization Recommendation Systems (VRSs)

VRSs fill the gap between non-expert users and data analysis, providing automatic tools for insights discovery

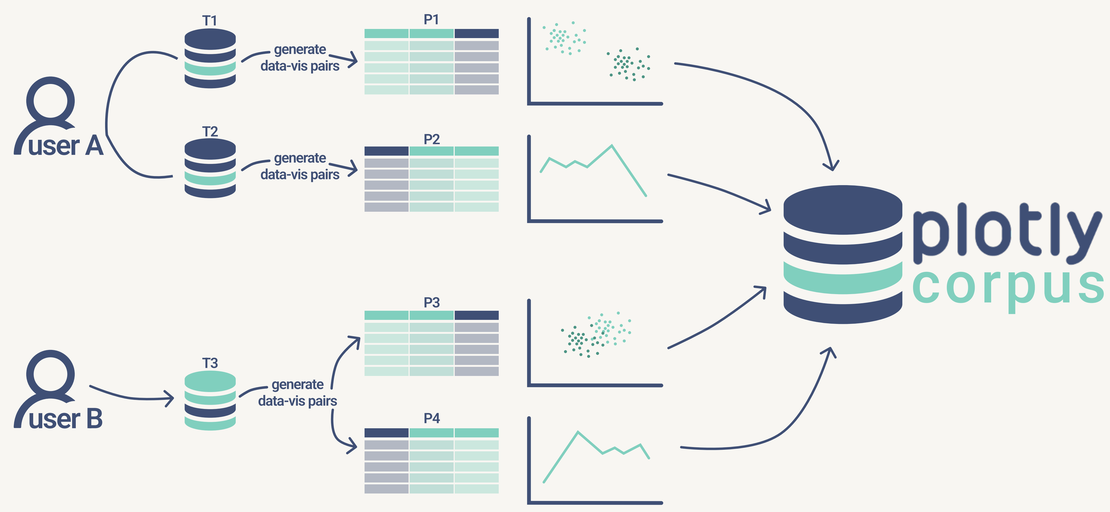


Plotly corpus dataset

Plotly corpus is the main reference dataset for Ml4VIs Application

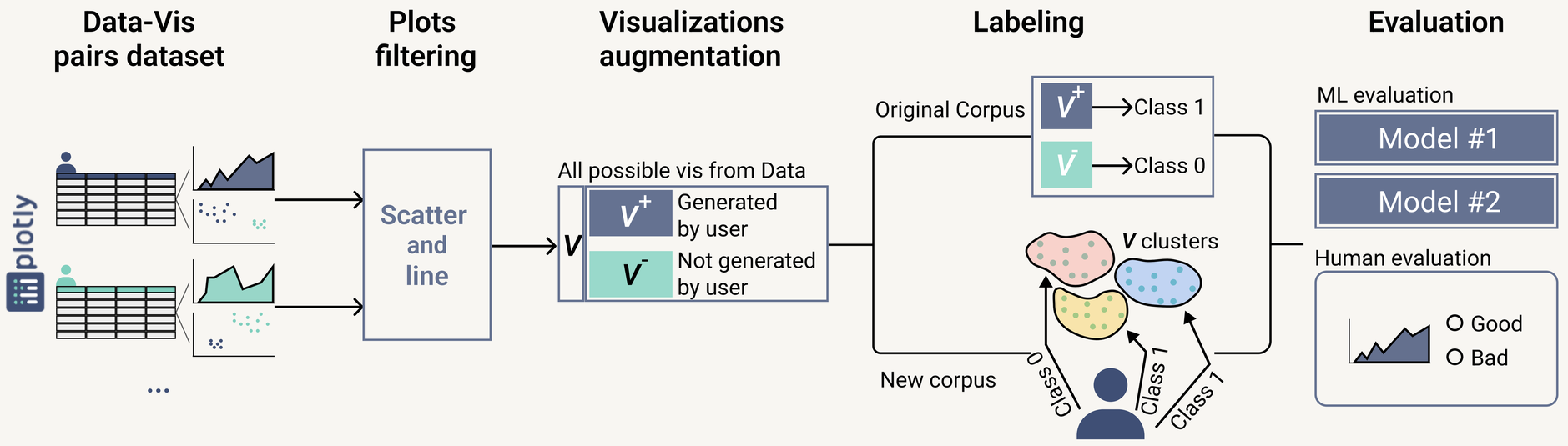
📊 The **Plotly corpus**, generated from Plot.ly by the authors of VizML and consists of data-visualization pairs from the public feed of Plot.ly

❌ The **main limitation** is that it contains data-vis pairs developed by both expert and non expert



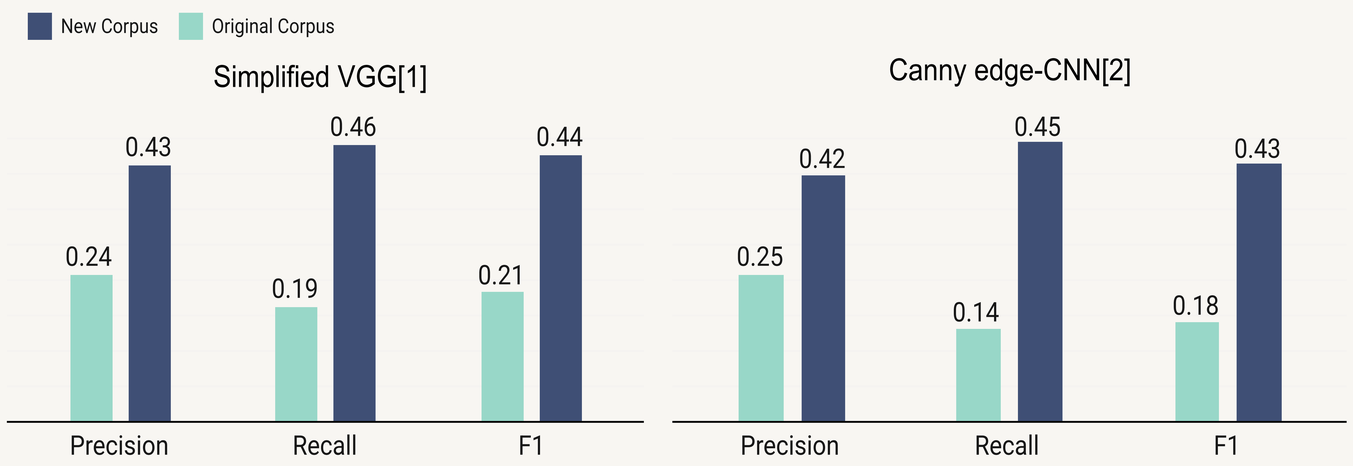
- Cluster on the entire set of charts, V, with both **V+** and **V-**
- ClustImage embeds the charts into a latent space
- Use hierarchical agglomerative clustering to group similar charts into the same clusters → **38 groups**
- Mean image = centroid of the clusters
- Human-in-the-loop to tag with either **1 or 0** the mean image of each cluster

Ploly.plus workflow



Evaluation

Crowdsourcing to confirm ~ 70% of the time the labels of the human annotator



[1] Bajić et al. 2019. *Chart classification using simplified VGG model*. IWSSIP.

[2] Kosemen et al. 2020. *Multi-label classification of line chart images using convolutional neural networks*

Next steps

🤖 **Workflow automatization**
No human factor involved in clustering ops

📦 **Larger subset relabelled**
More than 1 million images

🚀 **Self-supervised learning**
Triplet-loss to push V- instances away from V+.