Gauguin

Open-source Post-impressionist Data Visualization

Research Head
Professor Nicholas Brown nikbearbrown@gmail.com

Coordinator
Prabhu Subramanian
subramanian.pr@northeastern.edu

How to get started: -

1. Come to the weekly meetings which will start after **August 28th, 2020**, it will be weekly meetings
2. Learn about the basic idea of the project, get involved in the discussion, and get your questions answered
3. Start working on the project interested and present your progress weekly and get the feedback and constructive criticism to improve it
4. Provide weekly update as you work with the Research by presenting your weekly work in the meeting organized every Thursday 4 PM – 5 PM.

Gauguin is a project with 7 objectives:

Project 1: Extending existing library - Ex: Facets / AutoViz
Project 2: AutoFE visualization book library
Project 3: Bots and Database
Project 4: Visualizing bias book
Project 5: Guagan.ai
Project 6: GauguinDesigner (Hiatus)
Project 7: Propose your own idea under the same topic(Proposal template)

1. **Project 1: Extending existing library** – This is the idea to work and develop a library like the available open projects. Ex: Facets / AutoViz - to create a python based AutoEDA tool similar to the many AutoML frameworks where a user uploads some data selects a target of interest and the tool creates thousands of visualizations giving insight into the data ranked by interestingness
2. **Project 2: AutoFE visualization book library –** Work in visualizing the AutoFE(Book) and expressing the idea of Feature Engineering. This aims to **Post-impressionist -** Post-Impressionism has an emphasis on more symbolic content, formal order, and structure. The idea is to create a picture that reflects the essence of something.
3. **Project 3: Gauguin Bot and Data Journalism Database** - to create a python-based bot to gather social media data visualizations. The web bot will find data visualizations and their data sources and annotate them. A neural network will be used to classify images according to their type.

4. **Project 4: Visualization of Bias – Book** – this is also a part of **Post-Impressionism**. Here the intention is to research on visualizing bias in a data.
5. **Project 5: Gauguin.ai website (Hiatus)** - the website is intended and a portal and community for those interested in data visualization research
6. **Project 6: Gauguin Designer (Hiatus)**- the idea is to use a data set and a target or an existing graphic and have the tool to make design suggestions. This project will be kept on hold until Project 3 gathers enough data to start working for Project 6

Explanation of each Project

1. Gauguin (Library)

The idea automates EDA - to create a python based AutoEDA tool similar to the many AutoML frameworks where a user uploads some data, selects a target of interest and the tool creates thousands of visualizations giving insight into the data ranked by interestingness.

We will start by refactoring the AutoViz and Auto_ViML libraries https://github.com/AutoViML.
Also look at the open source libraries to get started and working to extend the same idea - Facets / AutoViz.
The AutoViz is missing many important EDA plots such as individual conditional expectation (ICE), leave-one-covariance (LOCO), local feature importance, partial dependency plots, tree-based feature importance, standardized coefficient importance, accumulated local effects (ALE) plots and Shapley values.
The AutoViz plots are ugly. An important feature would be to add "themes". That is, design parameters like font, leading, color sets, etc. There should be themes like "New York Times", "The Economist," etc. which make the plots adhere to a visual style.

The parameters of the theme should be configurable in a JSON file so a user and use custom themes. There seems to be no ranking of the plots to show interesting plots first in AutoViz.The library needs to be easily extendable. As new plots get developed in the research aspects of the project, they should be able to be added to the library in a standard and easy manner.

2. Gauguin.ai website is intended and a portal and community for those interested in data visualization research.

- It should look a little like the New York Times crossed with 538.
- It should allow users to publish Medium style article.
- It should also have a kind of article called a "wiki article". This means that the author will originally write an article but then allow others to edit and extend it.
- It should house the Data Journalism Database.
- It should have Yelp-like reviews of data visualization tools.
- It should have a database of data visualization freelancers and professionals.
- It should have a database of data sources.
- It should have a database of data visualization conferences and journals.
- The front page should be a newsfeed and recommended articles. This should be personalized if a user logs-in.
- All articles should be auto tagged (classified) according to a glossary of data visualization jargon.

3. Gauguin Bot and Data Journalism Database

The idea is to create a python-based bot to gather social media data visualizations. The web bot will find data visualizations and their data sources and annotate them. A neural network will be used to classify images according to their type

The types of graphs classified will start with:

Arc Diagram, Area Graph, Bar Chart, Box & Whisker Plot, Brainstorm, Bubble Chart, Bubble Map, Bullet Graph, Calendar, Candlestick Chart, Chord Diagram, Choropleth Map, Circle Packing, Connection Map, Density Plot, Donut Chart, Dot Map, Dot Matrix Chart, Error Bars, Flow Chart, Flow Map, Gantt Chart, Heatmap, Histogram, Illustration Diagram, Kagi Chart, Line Graph, Marimekko Chart, Multi-set Bar Chart, Network Diagram, Nightingale Rose Chart, Non- ribbon Chord Diagram, Open-high-low-close Chart, Parallel Coordinates Plot, Parallel Sets, Pictogram Chart, Pie Chart, Point & Figure Chart, Population Pyramid, Proportional Area Chart, Radar Chart, Radial Bar Chart, Radial Column Chart, Sankey Diagram, Scatterplot, Span Chart, Spiral Plot, Stacked Area Graph, Stacked Bar Graph, Stem & Leaf Plot, Stream Graph, Sunburst Diagram, Tally Chart, Timeline, Timetable, Tree Diagram, Tree map, Venn Diagram, Violin Plot, and Word Cloud.

See https://datavizcatalogue.com/ for descriptions of these graphs.

Visual Techniques such as small multiples, 3D, interactive, video, animation, dashboard, etc. will be annotated. Further, the colors used, negative space, data to ink ratio, position along a common scale, positions along with nonaligned identical scales, length, area, volume, shading, color saturation, direction, angle, curvature, font, leading, etc. will be annotated.

The database will be open-source and available as a data dump and well as searchable and browsable via the Gauguin.ai site.

4. Gauguin Designer

For Gauguin Designer the idea is to use a data set and a target or an existing graphic and have the tool make design suggestions.

One idea is to use a data set and a target and possibly other tune able latent variables and use GANs to generate and rank "design sketches" for visualization. That is, to use GANs to suggest interesting visualization designs that can be replicated more precisely using the Gauguin library once a design is decided. Another area of research is to allow one to upload an image and get critique on the type of graph chosen, font, color, etc.,

Another area of research is to allow one to upload an image and find similar images from the Data Journalism Database ranked by the quality of the related images.

5. Gauguin Visualizing the Abstract and Symbolic

Post-Impressionism has an emphasis on more symbolic content, formal order, and structure. The idea is to create a picture that reflects the essence of something. This is a research-oriented sub-project interested

in creating novel visualizations and extending existing graphs and algorithms to visualize abstract ideas like bias or confounders in data. These novel visualizations are part of a python library called Gauguin.