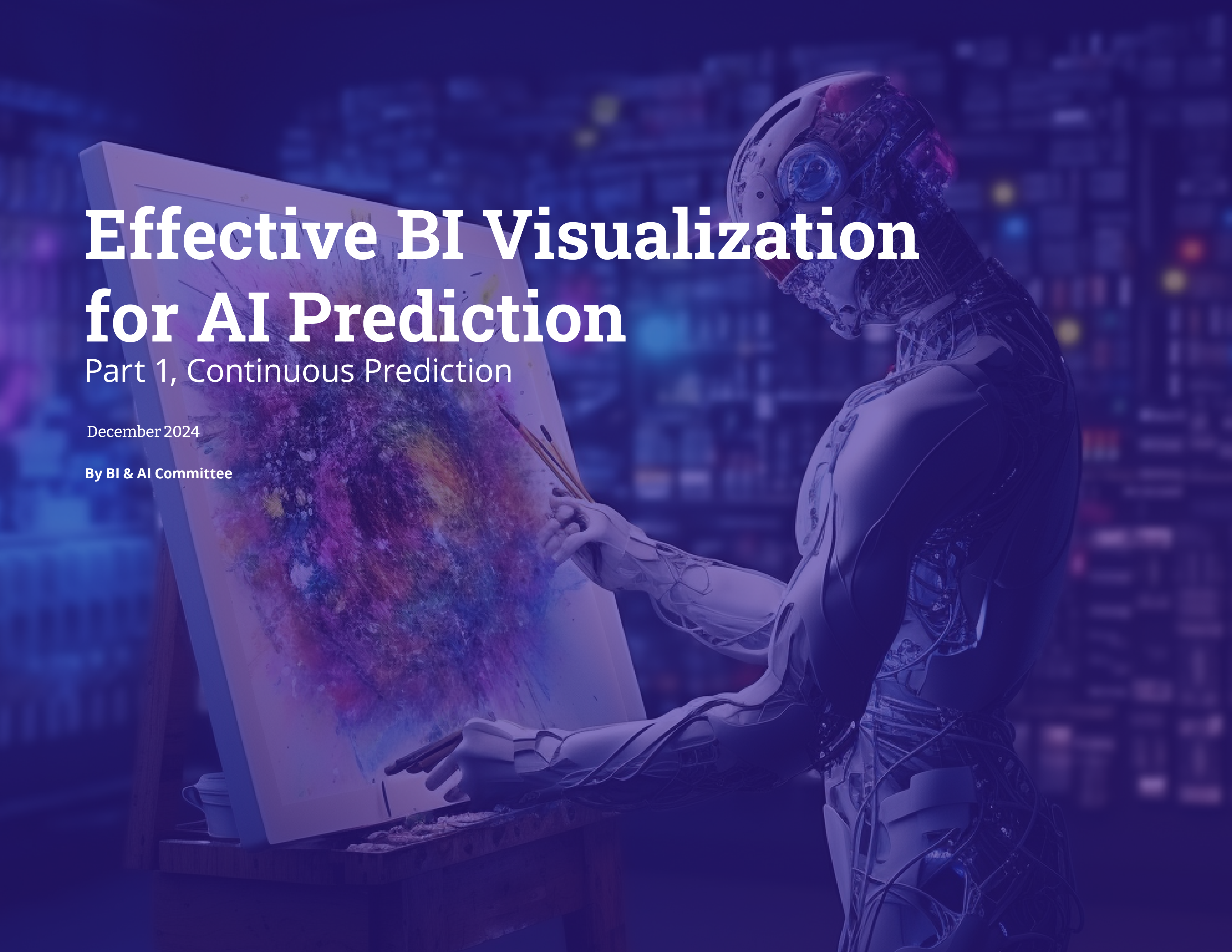


# Effective BI Visualization for AI Prediction

## Part 1, Continuous Prediction

December 2024

**By BI & AI Committee**

# Introduction

With so much material online to consume we know you aren't reading this by mistake. You are either here because you totally understand the issue and would like help, or you are curious to see why in a world so full of Large Language Model hype we are writing about visualizing AI predictions. After all, you simply take the output and display it on the screen. Right?

Wrong!

When we last saw Sally Sue Somebody in [The Convergence of AI, BI and Data](), her career had been evolving quickly. As expected, she was being asked to take advantage of the shiny new Automated Machine Learning tool that the company had purchased. It really was automated and wowzah, she was able to throw data in and get predictions out. No problem there. Woo-hoo, Sally Sue Somebody was able to add "Citizen Data Scientist" to the ever-growing list of roles she put on her LinkedIn profile.

Visualizing the predictions is where Sally Sue was really struggling. For binary predictions, she was converting 0's to the word No, and 1's to the word Yes just as she had been doing when she was given batched output files from the Data Scientists within the organization. But now that she was generating the predictions, she had started seeing the prediction scores as well as the 0's and 1's.

She realized that the 0's and 1's weren't binary. Mind-blowing, we know. But she reasoned that a 1 with a score of 0.5000001 was closer in nature to a 0 with a score of 0.4999999 than it was to a 1 with a score of 0.9999999. She reasoned, and perhaps you have as well, that end users needed to see both the prediction and the scores. In her mind, there were "I flipped a coin 1's" and "I'm pretty sure 1's" and "I pinky promise it's definitely a 1's."

*The purpose for Data Visualization in Business Intelligence isn't to show pretty versions of the words "Yes" or "No." It is to inform users so that they can make decisions on that knowledge and then act on it.*

We totally agree with her. The purpose for Data Visualization in Business Intelligence isn't to show pretty versions of the words "Yes" or "No." It is to inform users so that they can make decisions on that knowledge and then act on it. Should the actions of a "Yes" with a score of .5000001 be the same as the actions of a "Yes" with a score of .9999999 be the same? Of course not.

It's not just about the 0's and 1's of binary predictions either. Her mind was flooded with thoughts of how to best visualize the Classification, Continuous and Forecasting predictions as well. Certainly, the same concepts of combining the scoring and the prediction had to be applied. Hence Sally Sue's predicament and the reason we are writing this piece.

# Continuous Prediction

Let's first talk about arguably the most common prediction of all - Continuous prediction, also known as regression. There are many examples in our daily life. For instance, House Price Prediction uses machine learning models to predict house prices based on features such as location, size, amenities, proximity to schools or parks, and more. Moreover, Health Risk Scores are often calculated using continuous prediction models. To predict a person's likelihood of developing certain conditions like heart disease or diabetes based on their current health data (e.g., BMI, cholesterol levels, blood pressure).

Continuous prediction differs fundamentally from other tasks, such as classification or clustering. This is also why effective visualizations used in Continuous Prediction differ from those used in others. We must understand the key differentiators and corresponding visualizations to appreciate the uniqueness.

## Nature of Continuous Prediction

### Predictive Outcomes

In continuous prediction, the output is a continuous quantity. For example, predicting the temperature for the next week or forecasting the stock market prices. On the contrary, classification problems have discrete outcomes, like determining whether an email is spam, while clustering problems involve dividing the dataset into distinct groups.

Because of this, visualizations suitable for continuous prediction must effectively demonstrate constant change. Line graphs, scatter plots, heat maps, and contour plots are great tools for this purpose. However, visualizations that work well for discrete outcomes, like bar charts and pie charts, may not be appropriate for continuous prediction.

*Continuous prediction differs fundamentally from other tasks, such as classification or clustering. This is also why effective visualizations used in Continuous Prediction differ from those used in others.*

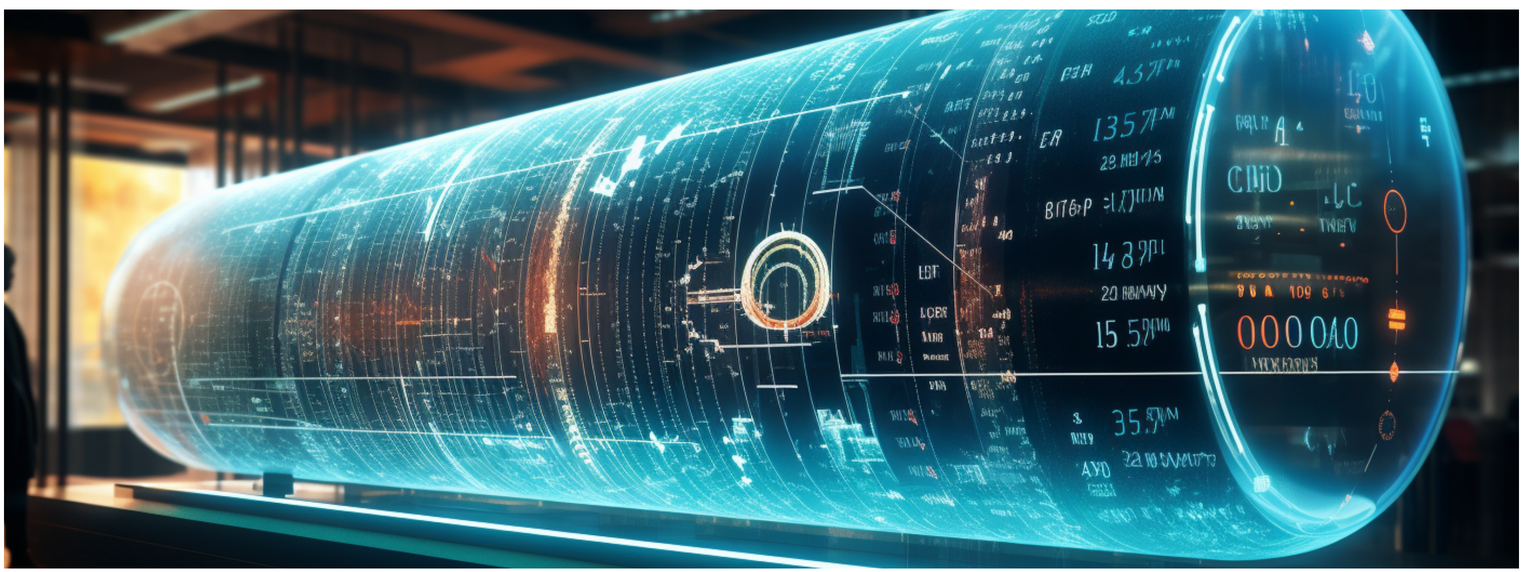
### The Complexity of the Relationship

Continuous prediction often deals with more complex relationships between variables, as it determines whether variables are related and how changes in one or more variables directly impact the outcome.

Hence, visualization techniques like heat maps or contour plots, which can show the correlation and density of variables, can be helpful. These techniques allow viewers to understand the intricate dependencies among multiple variables and their cumulative effect on the prediction.

### Non-Time-Dependent

Yes, we know what you are thinking. Forecasting is also continuous prediction. But since it involves a unique "Time" dimension, we branch it out and will talk about this in the later section and focus



on Continuous prediction to non-time dependent. Therefore, non-time forecasting continuous prediction doesn't primarily focus on trends, seasonality, or time-series data. It's more about predicting an outcome based on current conditions and relationships among variables

### Sensitivity to Fluctuations

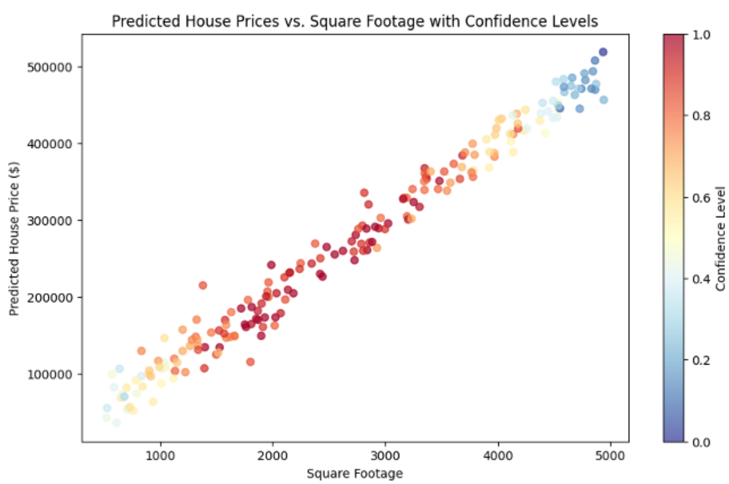
Continuous predictions are often sensitive to slight changes in the data. It's essential for visualization techniques to represent these fluctuations accurately.

Visualizations such as line graphs and scatter plots can capture even minor changes in the data with the correct scale in place, making it possible to identify and understand these subtle shifts. However, techniques such as bar graphs, which may generalize data too much, may not be ideal for visualizing continuous predictions.

## The fun part – Visualization

### Color-Graded Scatter Plots

A color gradient can represent the degree of certainty in the predictions.




*Visualization techniques like heat maps or contour plots, which can show the correlation and density of variables, can be helpful. These techniques allow viewers to understand the intricate dependencies among multiple variables and their cumulative effect on the prediction.*

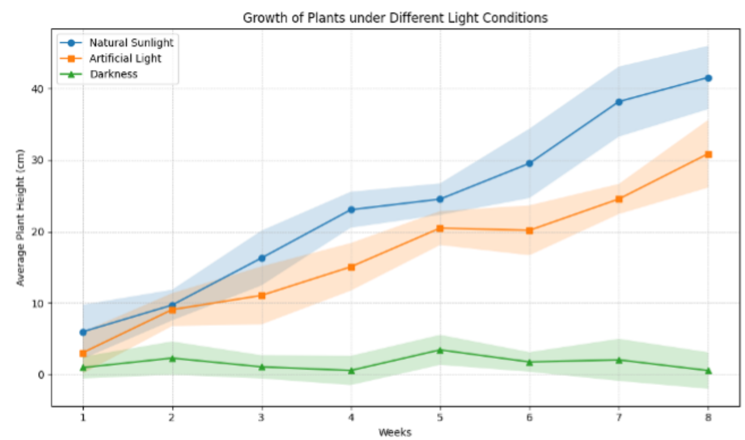
For example, highly confident predictions could be colored in a dark shade, while those with lower confidence are in a lighter shade.

**Pros:** Offers an immediate, intuitive depiction of confidence levels alongside the actual predictions.

**Cons:** Precision can be lost due to the subjective nature of color perception, and color blindness can pose accessibility issues.

### Error Bars or Ribbon Plots

Error bars or ribbons around the regression line in a scatter plot can depict the uncertainty in predictions. These could represent confidence intervals, for example.




**Pros:** Gives a sense of the range of possible values for each prediction and immediately shows where the model is more or less certain.

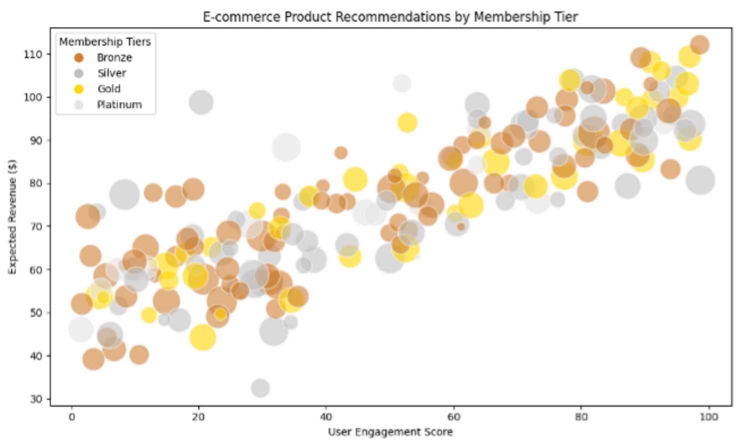
**Cons:** Can become visually cluttered, especially with larger datasets. It might be complex to interpret for non-technical stakeholders.

## Bubble Plots

The size of the bubbles in the plot could represent the confidence in the prediction, with more giant bubbles signifying higher confidence.

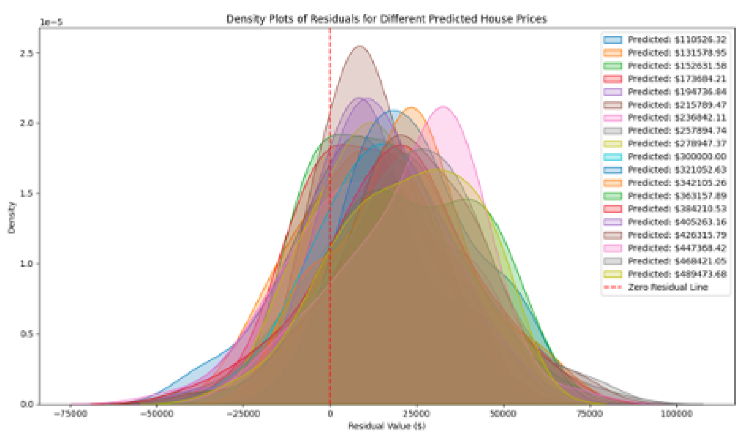
**Pros:** Visually intuitive, with the size of the bubbles offering a quick sense of the confidence levels.

**Cons:** Precision can be difficult to gauge from bubble size alone, and the plot can become cluttered with overlapping bubbles.




## Density Plots

Density plots can show the distribution of residuals or errors. For each predicted value, you can create a density plot (a smoothed histogram) showing the distribution of the residuals around that value. Wider distributions indicate less certainty.

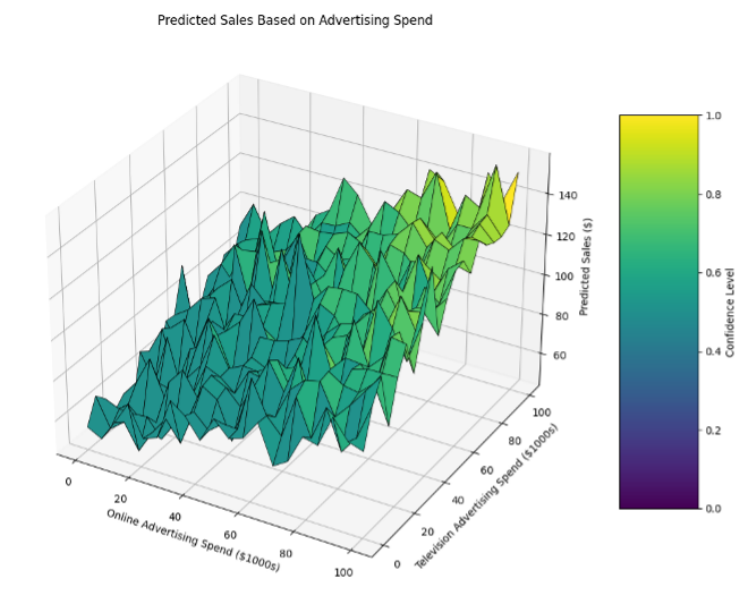



**Pros:** It can show much more detailed information about the confidence levels for each prediction.

**Cons:** It can be difficult to interpret, especially for non-technical users.

## 3D Surface Plots with Color Gradient

In the case of multiple predictors, you can use a 3D surface plot where the color gradient depicts the confidence level in the prediction. These plots can also represent the relationship between multiple predictors and the predicted response, with color indicating the degree of certainty.




**Pros:** Visually appealing and can convey complex multivariate relationships and confidence levels.

**Cons:** It can be hard to interpret, especially with many variables. Not suitable for printed format and color-blindness can pose accessibility issues.

## Other Color Considerations

Color may be suitable for some, but it can be bad for colorblindness. Considering colorblindness when designing visualizations is crucial as it ensures that your graphics are accessible to a broader audience, including those with various types of color vision deficiency. Here are some tips on how to make your visualizations more colorblind-friendly:

- **Avoid Problematic Colors:** Red-green and blue-yellow are the most common types of color blindness. Therefore, avoid using these color combinations in your visualizations.
- **Use High Contrast:** Use colors that have high contrast with each other. This helps everyone, not just colorblind people, to distinguish different data points or series.
- **Use Patterns or Textures:** Instead of relying on color alone, use different patterns or textures to differentiate between data points or series. For example, use dotted or dashed lines in addition to solid ones.
- **Use Symbols:** In scatter plots or line graphs, use different symbols (like circles, squares, triangles, etc.) for each data series.
- **Provide Multiple Indicators:** If color indicates a specific value or category, provide another indication, like a label, size, or shape. This redundancy can help ensure that the message gets through.
- **Test Your Visualizations:** Use online tools like color blindness simulators to test your graphics. These tools can help you see your visualizations as a colorblind person would.

## Summary

Great, we have explored how to present Continuous Prediction effectively. As we mentioned, we intentionally omitted an important dimension, Time, in the discussion and saved that for the next Episode. Stay tuned!

*Note: Source code used in this section can be found at* [https://github.com/cupidchan/BIAI](https://github.com/cupidchan/BIAI)

# Authors

## Cupid Chan

Cupid Chan, an industry veteran, holds the Senior Fellow position at CDHAI in Johns Hopkins and the University of Maryland. His career began in developing a top-tier BI platform and contributing to Linux Foundation ODPi. As Chair of LF AI & Data's BI & AI Committee, he drives the fusion of AI and BI, forging Cognitive Intelligence (CI), melding AI's speed with human-driven insights.

## Deepak Karuppiah

Deepak Karuppiah, a Senior Architect in MicroStrategy's Augmented Analytics Group, merges AI and BI to craft technology stacks for AI-powered dashboards. Formerly, he specialized in secure data connectors for MicroStrategy. Proficient in machine vision and learning, he explores novel AI-BI intersections, evident through his role in the BI & AI committee. Beyond work, he volunteers technical skills to a local non-profit, showcasing dedication beyond his professional sphere.

## Dalton Ruer

Dalton is a Senior Solution Architect at Qlik. He is a Data Scientist Storyteller, Analytics Evangelist and is an impassioned student of Generative AI. He is a seasoned author, speaker, blogger and YouTube video creator who is best known for dynamically sharing inconvenient truths and observations in a humorous manner. The passion which Dalton shares through all mediums, moves and motivates others to action.

## Sachin Sinha

Sachin, Microsoft's Director of Technology Strategy, has honed his data engineering expertise post-graduation from the University of Maryland. Specializing in information management, he crafted systems aiding data-driven decision-making. His support extended to healthcare startups, fostering data-centric business models, and empowered public sector entities toward their missions via data-enabled decisions. Residing in Fairfax, VA, with his family, he passionately backs the Terps and Ravens while excelling in his professional pursuits.

## Stu Sztukowski

Stu, a senior product manager at SAS, focuses on user-friendly AI through visual analytics. With a BS in Statistics from North Carolina State University and an MS in Advanced Analytics, he transitioned from data science, excelling in forecasting and BI. He champions accessible high-performance AI, aiming for universal comprehension. A mentor and versatile leader, he simplifies complex analytics for data scientists and business analysts, driven by a passion for public speaking and democratizing analytics.

twitter.com/LFAIDataFdn

linkedin.com/company/lfai

youtube.com/@lfaidatafoundation9555

github.com/lfai

LF AI & DATA

## About the LF AI & Data Foundation

LF AI & Data is an umbrella foundation of the Linux Foundation that supports open source innovation in artificial intelligence (AI) and data. LF AI & Data was created to support open source AI and data, and to create a sustainable open source AI and data ecosystem that makes it easy to create AI and data products and services using open source technologies. We foster collaboration under a neutral environment with an open governance in support of the harmonization and acceleration of open source technical projects.Explore our current portfolio of projects and contact us to discuss hosting your open source AI or data project under our Foundation.

December 2023



To reference this work, please cite as follows:
https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=35160417#BI&AI-PastPublication