

Calvin Thomas

New OOP Standards of Model Functionality Reusability - Genome-Free Data_Modeling_Project – First Step, Data Mining, Unstructured ML and Data Modeling

Every workday, from [your local time] 8AM to 5PM, developers at firms likely type many-multiple-millions of lines of code, and firm-level worldwide-used models as an amalgamated group creating billions of actionable signal insights via real-time model outputs for executive decision-making. Social Media on the weekend and on breaks creating a sprawling unmanageable dataset that cannot be possibly reviewed by individual developers, or truly reviewed row-by-row to truly navigate how to wrangle and classify the unstructured dataset. With this new territory comes increasing model complexity risk that internal review cannot fix, and generating novel insights from large unstructured datasets in their exact same category (undiversified datasets create latent “hidden variable” tail-risks) with only internal modeling, code, and data. This is a huge waste that can be immediately fixed for eliminating the largest corporate tail risk – outside developers non-biased review and assessment for cracks.

For private firms, etc, with non-sensitive data and information (GPT auto-checks needed for firm confirmation), we first categorize as structured or unstructured datasets and anonymize the dataset to highest possible standard. We can encourage open-sourcing as many structured and unstructured datasets to the public and high-level, high-quality internally-reviewed modeling and code released for public review and benefit. The exact same models should be standardized world-wide and created for highest-quality OOP standards as a community. This ensures the highest level of scrutinization for review and most importantly, highest-quality industry standard internal OOP-practices disseminated to the broader developer community and public to standardize high-quality coding as an ensemble group. The largest existing, combined dataset may seem disorganized at first, but it will immediately give the public and private access to the largest intellectual, unstructured ML challenge – incentivizing generating novel insights from already existing datasets and models for free and as a community to test and develop OOP standards uniformly together. Instead of the million-dollar industry acquisitions of new ones in the incremental pursuit of “1-step ahead” non-commodity, same source, alt data. Outsourcing the firm’s tedious task of finding insights with internal-bias structures; views for using same techniques in labeling, culling, and data mining that may prevent finding unique insights or “hidden variable” risks in datasets, modeling, and code. The combined dataset, same shared models, will generate novel insights and true programming standardization, not possible as individual firms creating redundant, disjoint internal datasets, models, and code. For OOP concepts, we should consider new code creation via existing code and modeling with accuracy rated for same functionality as a low-hanging fruit for code reusability.

If code programs/functions/models have been created before, the more versions/different variable names/etc of the exact same program/function/model we create, we clutter the code space with garbage non-uniform code and functionality translating into poor international universality in the interpretation of functions, code, and non-standardized model-level risks and universal code quality control. Written with our own individual programmer preferences and cultural expressions embedded in how use cases should be algorithmically encoded or modeled at individual firms; prevents the broader

public's access to this quality of standardization to learn for themselves with high-quality standard examples of datasets, code, modeling to become the best new developer talent right out the gate.