

# **Title: “Looking Beyond the Lamppost: Bringing Light into the Dark Alleys of Complex Data”**

**Jeffrey S. Morris**

**The University of Texas MD Anderson Cancer Center, Houston, TX USA**

**Abstract:** Modern science is characterized by new measurement instruments producing an explosion of data, ever-growing in their quantity and complexity. These data raise numerous quantitative challenges, among them the challenge of efficiently and reliably extracting the valuable scientific information they contain while managing the practical challenges raised by their size and subtleties. The absence of sufficiently flexible methods and frameworks often forces scientists to first simplify their data using simple summaries to eliminate some of their vexing complexities. This can work well if these summaries retain all relevant information in the data, but many times that is not the case. This convenience-driven oversimplification of complex data is like “looking for the lost keys under the lamppost,” where we hope them to be, while ignoring the “dark alleys,” which may in fact be where the keys reside. One primary objective of modern statistics is to develop efficient, flexible methods and modeling frameworks that can model the complex data as they are, while avoiding oversimplifications, and thereby better modeling the systems that generate the data and potentially uncovering more of the treasure trove of information they contain. We could say that these methods are intended to “bring light to the dark alleys of complex data,” emboldening researchers to venture to places they fear to go and perhaps uncover new insights as a result. I will illustrate these principles through several real-world applications: a colon carcinogenesis molecular marker study, a children’s activity study involving accelerometers from the Planet Health initiative, a brain proteomics study of addiction, a DNA copy number cancer genomics study, and an fMRI study of smoking cessation. I will summarize and discuss a flexible and efficient framework for modeling complex object data such as functions and images developed in the past several years. Our framework is based on the functional mixed model framework, a generalization of linear mixed models that can model simultaneous effects of multiple factors on the objects and handle correlation between objects induced by the experimental design. The multi-domain modeling approach and software developed for this framework is appropriate for high-dimensional objects with complex features, can accommodate missing data and outliers, and yields Bayesian inference for all model components. This work serves as an example of methodological development motivated by the premise of developing flexible, efficient frameworks for modeling complex object data. There are similar ongoing efforts currently underway by many other researchers, and I expect these joint efforts will collectively produce a rich, flexible set of modeling tools that we as the statistical community can offer to the broader scientific community to help them uncover insights they could not find without our contributions.