Research Statement

Weixi Gu
guweixi@berkeley.edu

The explosive growth of data over past years unleashes huge potentials to understand the world from novel perspectives. Many existing statistics models and deep learning frameworks have offered plug-and-play tools for the data analytics, but they are still hard to completely uncover the implicit structures such as the homogeneity and heterogeneity, thus missing its hidden values and essences.

My principal research interests are motivated by these points, which lie in the areas of designing theoretical machine learning and statistical methodology based on the proprieties of real data, and developing reproducible large-scale mobility solutions to tackle complex applications in the context of high-dimensional, multi-model, and dynamic data. In general, my research works balance between broad algorithmic theory and narrower domain/task-specific solutions. My current research work involves, 1) foundations of statistical learning, including information theory and algorithm on the automatic temporal and spatial feature engineering; 2) deep learning frameworks on high-dimensional mobility data modelling, to uncover inherent dynamic associations among a set of random variables; 3) large-scale applications of statistical learning on health sensing, transportation, and human dynamics. I will illustrate some of my works based on the above three topic in the following section.

Research Contributions

Feature Selection of High-dimensional Mobile Data. With the recent advancement in instrumentation and measurement technologies, researchers from a wide variety of disciplines now have access to rich and real-time measurement data. Often, those datasets present themselves in the form of high-dimensional co-evolving streams and are pre-processed to extract tentative patterns/features for data mining applications. Among the vast quantity of information generated by such process, some features are correlated with the target application while others may be less relevant or redundant. As such, identifying the most informative patterns or features from the observed data, *i.e.*, feature selection, is one of the underpinnings for the success of any data analytic methods, especially for the high-dimensional data analytics.

• *Adaptive Temporal Feature Selection and Online Learning by Causality:* Prior research focuses on the "static" feature selection where data is assumed to be generated independently from some distributions. In this work, we consider the problem of selecting patterns from streaming datasets. Inspired by a quantity from information theory that calibrates both instantaneous and temporal relations, we formulate the feature selection problem into cardinality constrained directed information maximization, which can be solved by a near optimal greedy algorithm for temporal feature selection. The performance of this method can be guaranteed by its monotonicity and approximately submodular. The features selected by the directed information, moreover, consider the causality, which can be regarded as the "real sources" of hidden variables. To show the effectiveness of DI on feature selection, we also provide theoretical bounds of probability error P_e given some certain directed informations, shown as Fig. 1.

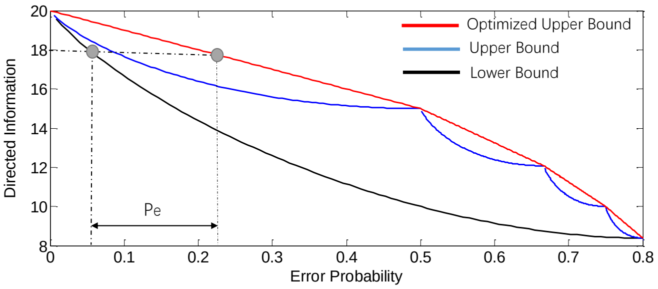


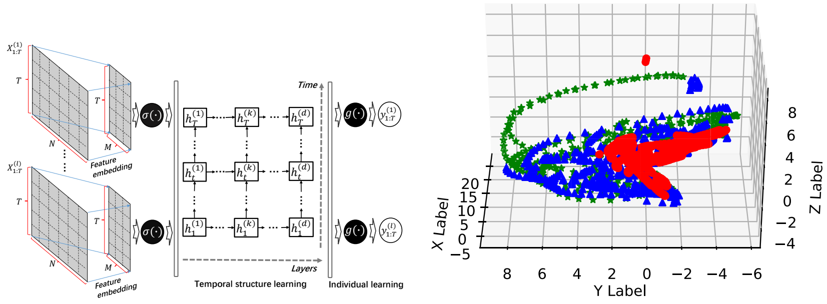
Figure 1: The Error Probability Bounds Based on Directed Information

Further, we apply the direction information on feature selection and infer the data dynamics via online learning method. The proposed method and its theoretical outcomes are also evaluated on both simulated data and real-world data involving blood glucose, flight price and stock price.

Structure Modeling of High-dimensional Mobile Data. High-dimensional data (*e.g.*, finance time series, biological sequences and neuroimaging) are challenging due to its inherent dependence and multi-order correlations.

How to establish a model uncovering the implicit data structure directly impacts the inference results. My research works are also stepped from this part, focusing on the spatial-temporal model construction of the high-dimensional data, *e.g.*, blood glucose concentration and WIFI signal.

- *Blood Glucose Concentration Prediction with Multivariate Time Series Deep Learning* [3]: Predicting blood glucose concentration facilitates timely preventive measures against health risks induced by abnormal glucose events. Advances in continuous blood glucose monitoring devices and smartphones have made it convenient for measurements of blood glucose and its external impacting factors. However, accurate and personalized blood glucose concentration prediction is still challenging. Since many factors impact the dynamics of blood glucose and different individuals hold distinct blood glucose fluctuation laws, previous models yield low prediction accuracy either due to ineffective feature representation or the limited, imbalanced learning of the high-dimensional blood glucose data. In this work, we propose MT-LSTM, a multi-time-series deep LSTM model for accurate and efficient blood glucose concentration prediction. As shown in Fig. 2a, MT-LSTM learns feature representations in the first part, which compresses a high-dimensional feature matrix into a low-dimensional feature matrix by a non-linear transformation. Three principal embedded features learned by MT-LSTM have been displayed in Fig. 2b, which can be well distinguished after feature representation learning. MT-LSTM uncovers temporal dependencies of blood glucose dynamics, enforces data sharing among multiple users for efficient training in the second part, and utilizes an individual learning layer for personalized blood glucose prediction in the third part. The temporal similarities of blood glucose dynamics and distinct personalized blood glucose characteristics are well presented by MT-LSTM.



(a)

(b)

Figure 2: a) MT-LSTM architecture b) The distribution of feature representation

- *Multi-divisional Modelling of Blood Glucose Levels via Md^{3}RNN* [6]: Inferring abnormal glucose events such as hyperglycemia and hypoglycemia is crucial for the health of both diabetic patients and non-diabetic people. However, regular blood glucose monitoring is usually invasive and inconvenient in everyday life, resulting to the imbalanced and often limited measurements of blood glucose. To take full advantages of the *sparse*, *imbalanced* measurements to build *personalized* blood glucose level models, we propose Md^{3}RNN (multi-division deep dynamic recurrent neural network), an efficient learning paradigm that extracts general blood glucose level relevant features and preserves user-specific characteristics (in Fig. 3). Md^{3}RNN advances previous personalized recurrent neural network (RNN) structures via a group-shared input layer to extract distinctive feature representations within the same group (*i.e.*, non-diabetic, type I and type II diabetic). Given the high-dimensional features parsed by the physical model, Md^{3}RNN first depicts complex glucose dynamics via a deep RNN model, and extracts generic feature representations with a grouped multi-division framework in the second, and finally preserves individual differences using personalized outputs. It tackles the sparsity and imbalance problem, which is the main hurdle of highly-accurate personalized blood glucose level tracking. Generally, Md^{3}RNN can be regarded as a deep hierarchical RNN architecture. It is a combination of single user division and grouped user division learning, which well presents the inherent high-dimensional blood glucose features from branch to general perspectives. Evaluations show that Md^{3}RNN outperforms the state-of-the-art methods on the blood glucose level inference. Since many dataset yields the three divisional characteristics (*e.g*, flight price, earth quack density and shock price), Md^{3}RNN can be easily applied to many other applications.

- *Non-parametric Outliers Detection in Multiple Time Series* [8]: Outlier detection is an essential step for the abnormal points filtering, especially on the high dimensional time series data processing. In this work, we consider the problem of outlier detection with multiple co-evolving time series data. To capture both the temporal dependence and the inter-series relatedness, a multi-task non-parametric model is proposed, which can be extended to data with a broader exponential family distribution by adopting the notion of Bregman divergence. The learning problem, Albeit convex, might be hard as the time series accumulate. In this regards, an efficient randomized block coordinate descent

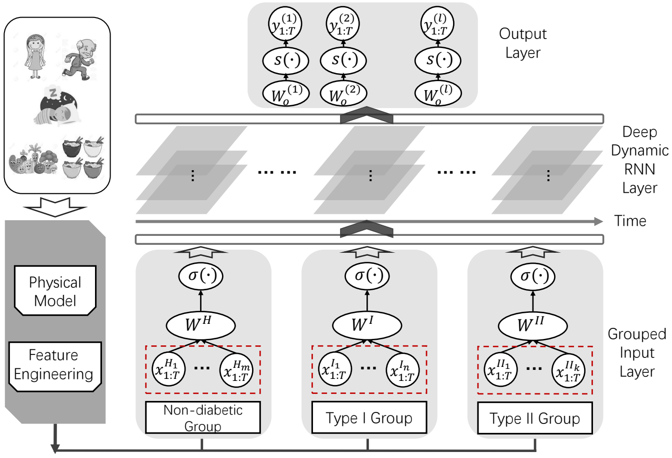


Figure 3: The architecture of Md^{3}RNN

(RBCD) algorithm is proposed. The model and the algorithm is tested with a real-world application, involving outlier detection and event analysis in power distribution networks with high resolution multi-stream measurements. It is shown that the incorporation of inter-series relatedness enables the detection of system level events which would otherwise be unobservable with traditional methods.

- *WiFi-based Human Identification via Convex Tensor Shapelet Learning* [9]: Human identification is a critical underpinning for the secure authentication and the tailored services to each individual. We provide a human identification system based on the CSI measurements from existing WiFi-enabled IoT devices. In order to improve the identification capacity, we propose convex tensor shapelet learning, which makes the identity estimation from the multi-stream temporal data. It formulates shapelet learning from tensors as a convex optimization problem and establishes an efficient generalized gradient based algorithm. Moreover, the incorporation of three concurrant regularization terms enables the automatic learning of the inter-dependence and the clustering effect of time series CSI tensor data. We believe this algorithm can also be applied on many other scenarios such as ”Parkinson Shaking” or ”Diabetes Gait”.

Understanding Individual Behaviors via Multi-dimensional Mobile Data. Mobile data collected by pervasive devices *e.g,* smartphones, wristbands bring much more opportunities to understand the individual behaviors compared with those of the past. By fusing the sensing data from embedded sensors of the wearable devices, human activities can be tracked. However, numerous inherent patterns cannot be directly mined by analyzing the data from single agent. How to collaborate the multi-dimensional sensing data from different sources and model their relationships with the hidden variable status are essential on the individual behavior analysis. My research works are stepped from this point, denoting to building and applying statistical methodologies to establish the inherent correlations between the sensing data and human patterns, and further infer the implicit behaviors.

- *Mining Inherent Sleep Stage via Mobile Data* [2, 4]: Sleep quality plays a significant role in personal health. A great deal of effort has been paid to design sleep quality monitoring systems, providing services ranging from bedtime monitoring to sleep activity detection. However, as sleep quality is closely related to the distribution of sleep duration over different sleep stages, neither the bedtime nor the intensity of sleep activities is able to reflect sleep quality precisely. To this end, we present Sleep Hunter, a mobile service that provides a fine-grained detection of sleep stage. It encodes the correlations of observable features and the hidden sleep stage transition for sleep quality monitoring and intelligent wake-up call. The rationale is that each sleep stage is accompanied by specific yet distinguishable body movements and acoustic signals. Leveraging the built-in sensors on smartphones, Sleep Hunter integrates these physical activities with sleep environment, inherent temporal relation and personal factors by a conditional random field model for a fine-grained sleep stage inference. By carefully designing the characteristic functions, the conditional random field model not only encodes the dependency of the observable features and the hidden sleep stages, but also represents the temporal transitions between the hidden sleep stages. Based on the inference results, Sleep Hunter further provides sleep quality report and smart call service for users. Experimental results from over 30 sets of nocturnal sleep data show that our system is superior to existing actigraphy-based sleep quality monitoring systems, and achieves satisfying detection accuracy compared with dedicated polysomnography-based devices

- *Learning Passengers Behavior Pattern on the Subway via Smartphones* [5, 1]: High variability interchange times often significantly affect the reliability of metro travels. Fine-grained measurements of interchange times during metro transfers can provide valuable insights on the crowdedness of stations, usage of station facilities and efficiency of metro lines. Measuring interchange times in metro systems is challenging since agent-operated systems like automatic fare collection systems only provide coarse-grained trip information and popular localization services like GPS are often inaccessible underground. In this work, we propose a smartphone-based interchange time measuring method from the passengers' perspective. It leverages low-power sensors embedded in modern smartphones to record ambient contextual features, and utilizes a two-tier cost sensitive classifier to model the spatial and temporal correlations of the mobile data, which is used to infer interchange states during a metro trip, and further distinguish 10 fine-grained cases during interchanges. Experimental results within 6 months across over 14 subway lines in 3 major cities demonstrate that our approach yields an overall interchange state inference F1-measurement of 91.0% and an average time error of less than 2 minutes at an inference interval of 20 seconds, and an average accuracy of 89.3% to distinguish the 10 fine-grained interchange cases. We also conducted a series of case studies using measurements collected from crowdsourced users during 3 months, which reveals findings previously unattainable without fine-grained interchange time measurements, such as interchange time variance of a day (in Fig. 4), the comparison of *waiting time* and *transfer time* during interchange through a day (in Fig. 5) and interchange directions, usage of facilities (stairs/escalators/lifts), and the root causes of long interchange times.

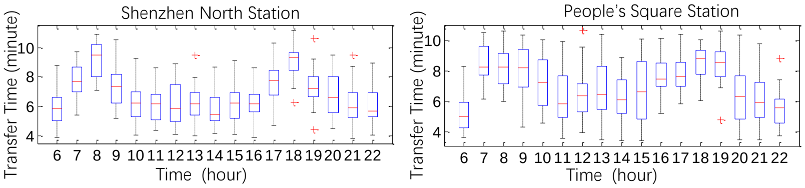


Figure 4: Interchange time variance of day

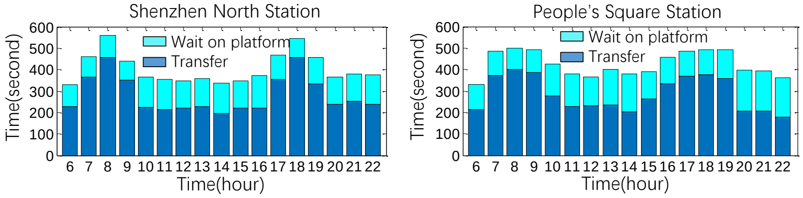


Figure 5: Comparison of Waiting Time and Transfer Time During Interchange at Different Hours of a Day

- *Modeling Multiple Cyclists Riding Patterns with Mobile Data* [7]: Detecting dangerous riding behaviors is of great importance to improve bicycling safety. Existing bike safety precautionary measures rely on dedicated infrastructures that incur high installation costs. In this work, we propose a ubiquitous bicycling behavior monitoring system based on smartphones, which invokes the embedded sensors to infer dangerous riding behaviors including lane weaving, standing pedalling and wrong-way riding. Considering the riding patterns of different cyclists are different, it is hard to leverage an uniform model to describe the riding traces of data and infer the dangerous events directly. To handle with this issue, we adopt transfer learning to reduce the overhead of training models for different users, and apply crowdsourcing to infer legal riding directions without prior knowledge. Experiments with 12 participants show that this platform achieves an overall accuracy of 86.8% for lane weaving and standing pedalling detection, and yields a detection accuracy of 90% for the wrong-way riding using crowdsourced GPS traces.

References

[1] Weixi Gu, Ming Jin, Zimu Zhou, Costas J Spanos, and Lin Zhang. Metroeye: Smart tracking your metro trips underground. (best paper runner-up award). In *MobiQuitous*, pages 84–93, 2016.

[2] Weixi Gu, Longfei Shangguan, Zheng Yang, and Yunhao Liu. Sleep hunter: Towards fine grained sleep stage tracking with smartphones. *IEEE Transactions on Mobile Computing*, 15(6):1514–1527, 2016.

[3] Weixi Gu, Costas Spanos, and Lin Zhang. Blood glucose concentration prediction with multivariate time series deep learning (submitted). In *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*.

[4] Weixi Gu, Zheng Yang, Longfei Shangguan, Wei Sun, Kun Jin, and Yunhao Liu. Intelligent sleep stage mining service with smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 649–660. ACM, 2014.

[5] Weixi Gu, Kai Zhang, Zimu Zhou, Ming Jin, Yuxun Zhou, Xi Liu, Costas J Spanos, Zuo-Jun Max Shen, Wei-Hua Lin, and Lin Zhang. Measuring fine-grained metro interchange time via smartphones. *Transportation Research Part C: Emerging Technologies*, 81:153–171, 2017.

[6] Weixi Gu, Yuxun Zhou, Zimu Zhou, Xi Liu, Han Zou, Pei Zhang, Costas J Spanos, and Lin Zhang. Sugarmate: Non-intrusive blood glucose monitoring with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):54, 2017.

[7] Weixi Gu, Zimu Zhou, Yuxun Zhou, Han Zou, Liu Yunxun, Costas J Spanos, and Lin Zhang. Bikemate:bike riding behavior monitoring with smartphones. In *MobiQuitous*, 2017.

[8] Yuxun Zhou, Arghandeh Reza, Han Zou, and Weixi Gu. Non-parametric outliers detection in multiple time series a case study: Power grid data analysis. In *AAAI*, 2018.

[9] Han Zou, Yuxun Zhou, Jianfei Yang, Weixi Gu, Lihua Xie, and Spanos Costas. Wifi-based human identification via convex tensor shapelet learning. In *AAAI*, 2018.