LOCALIZED ANALYSIS FRAMEWORK USING SMARTPHONES

Bommae Kim

Seoul, South Korea

M.A., University of Virginia, 2014

A Dissertation presented to the Graduate Faculty of
the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Psychology

University of Virginia
April, 2016

Committee Members:
Steven M. Boker (Chair)
Timo von Oertzen
Xin Cynthia Tong
Matthew Gerber (Systems and Information Engineering)

Abstract

Mobile phones have become an intrinsic part of our daily lives. Any data collected from them is especially reflective of our day-to-day reality. By using smartphone apps, researchers can now observe behavioral patterns in mobile data, all without soliciting any responses or relying on self-reports. Such behavioral data includes the frequency and length of phone calls, the contents of text messages, and users' physical locations. However, this immediately brings up privacy issues. Many people do not want others to harvest their mobile phone's private data. To utilize mobile data without invading privacy, I propose a localized analysis framework using smartphones. In this approach, researchers do not directly collect raw data. Instead, software on each individual's smartphone accesses the raw data locally stored in the device, and only the processed results are sent to the researchers. Researchers then analyze the aggregated individual results and find meanings at the whole sample level.

The current study aims to demonstrate utilities of the proposed localized analysis framework using smartphones with an emphasis on exploratory analysis using machine learning. While there are many possible methods to employ within this framework, in this study I adopt support vector machines (SVMs) for localized analysis at an individual level and visually inspect the aggregated patterns at the whole whole sample level. The simulation study results suggest that the proposed framework can perform as well as a conventional data collection and analysis paradigm. In this study, the aggregation of multiple individual models recovered the underlying true patterns as well as the single model built

with the whole data. The localized analysis framework opens a new way of treating data to protect user privacy.

Table of Contents

1. Background

Smartphones as data collection tools

The number of smartphone users is estimated to grow to 6.1 billion–70% of the global population–by 2020, which will outnumber fixed phone subscriptions (Erricson, 2015). Given that wide usage, smartphones are useful tools to collect data for social and behavioral science research: They offer great convenience to both researchers and study participants (Miller, 2012; Shah, Cappella, & Neuman, 2015).

Smartphone users can participate in studies using their own mobile phone whenever and wherever they want to. They don't have to commit a big chunk of time to participate in studies; rather, they can utilize their spare time. Researchers have already been adopting smartphones to replace traditional data collection methods, such as a phone interview or a paper-and-pencil (e.g., Killingsworth & Gilbert, 2010; Könen, Dirk, & Schmiedek, 2015). It is especially useful for longitudinal data collection, where participants provide data multiple times over the period of a study. Researchers no longer need to call their participants to fill out study materials every evening; smartphones prompt participants to do so at the preset time by researchers (e.g., Killingsworth & Gilbert, 2010).

Mobile data collection can go beyond traditional methods and provide unique information compared to traditional methods (Miller, 2012; Shah et al., 2015). In addition to numerical values and texts typically collected in traditional surveys or tests, a smartphone can easily include different forms of data such as images, audios, or videos (Lu, Pan, Lane, Choudhury, & Campbell, 2009). A mobile phone, by its nature, contains interpersonal data, including phone call logs, text messages, and contacts. This is valuable information for understanding interpersonal behavior (Calabrese et al., 2011). Even more, a smartphone can provide surrounding information collected by mobile sensors; mobile sensing data include

location or speed (Gonzalez, Hidalgo, & Barabasi, 2008; Song, Qu, Blumm, & Barabási, 2010). In fact, after consent, smartphone users can passively participate in studies, as their smartphones are constantly collecting surrounding data.

Moreover, using customized apps, a smartphone can prompt study materials based on contexts (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007; Gerber, 2015; Raento, Oulasvirta, Petit, & Toivonen, 2005). It can ask different questions based on a user's behavior. A question could pop up on the screen when the user enters their home or office. It could ask a series of questions at any point after the user takes photos or videos. It could prompt different questions when the user makes phone calls with family or with unsaved numbers.

For instance, if one would like to study how interpersonal contacts and social context influence a person's happiness, one could measure interpersonal contacts by frequency of phone calls and text messages, or length of phone calls in minutes and size of text messages in bytes, or by the number of bluetooth signals detected as a proxy of nearby people. One could also estimate social context (e.g., home or work) based on GPS location. One could measure happiness by self-report, by analyzing emotional words in text messages, or by analyzing smiles in photos.

New framework for data collection and analysis

A smartphone is a convenient and effective tool to gather a wide spectrum of data about an individual. It is even cost effective, as researchers do not need to purchase them for participants. It seems like researchers' dreams are coming true. However, this rich data immediately brings up serious concerns about privacy. Although aforementioned information is already being used for business purposes by commercial smartphone apps, research ethics inhibit researchers from obtaining extensive information from people's phone. To take advantage of such information, researchers should acquire participants' informed consent. But, even for research purposes, who wants someone else poking around their phones

and pulling out personal text messages and photos?

In fact, researchers often are not interested in each individual data value. Rather, they are interested in aggregated data and emerging patterns across many individuals. If we can achieve information about the whole sample without looking at original responses of individuals, then we do not need to obtain individual raw data. If we can assure that researchers cannot access raw data in their phones, smartphone users are more likely to readily agree to participate in studies and provide truthful answers.

I propose a new mobile data collection and analysis framework. A smartphone is computationally powerful enough to analyze locally stored data using its processors and memory. We can utilize its computing power for data analysis at an individual level. Then, each mobile phone sends the locally analyzed individual results to a study host server by a researcher. Researchers analyze the collected individual results at an aggregated level and find patterns and meanings. A comparison between the conventional framework and the proposed new framework is illustrated in Figure 1.1.

Consider a hypothetical study process using this framework. A researcher wants to study how interpersonal contacts influence happiness depending on social context in college students. Interpersonal contacts are measured by frequency of phone calls and text messages. Happiness is evaluated by emotional words in text messages. Social context (e.g., home or school) is recorded based on GPS location. The researcher defines features to include in analysis and customizes the research app for data collection and analyses. Then the researcher uploads the customized research app at a research hub website operated by a non-profit research organization.

Jamie is a community college student, who is always willing to contribute to science as a citizen-scientist. She is a member at the research hub website and occasionally receives study participation invitations. Jamie downloads the research app and installs it on her smartphone. The app shows a consent form that explains potential risks and benefits of study participation and then asks for her agreement. The app clearly explains which data

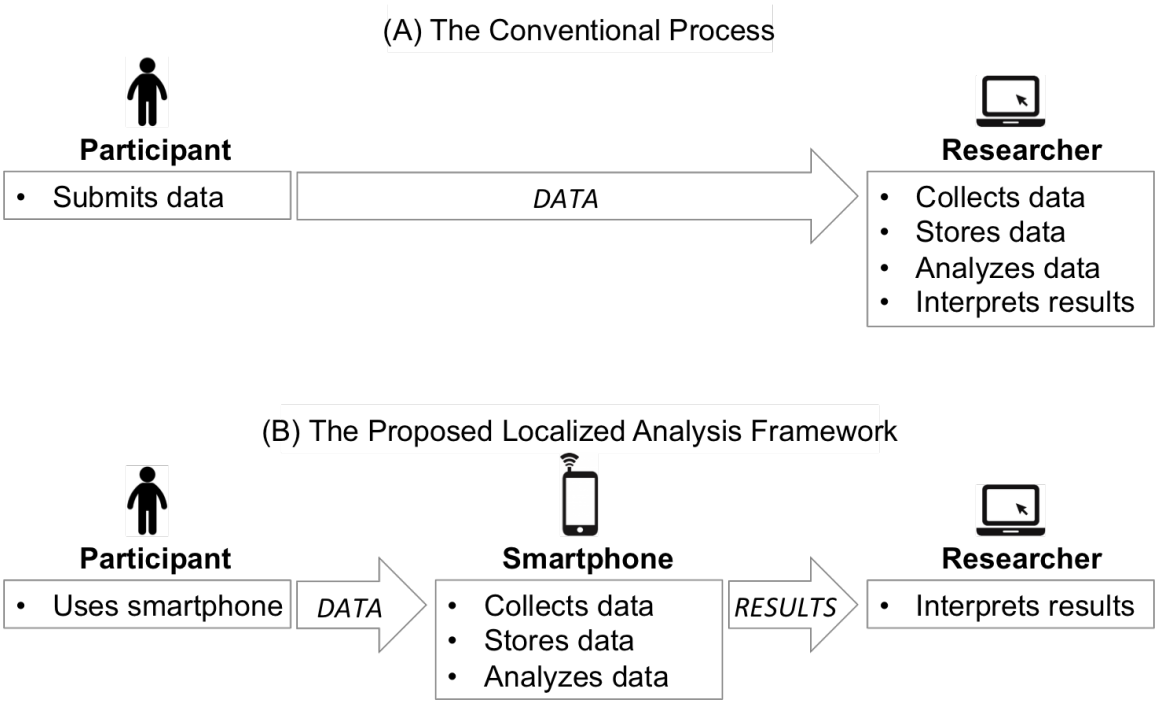

Figure 1.1: Data collection and analysis process in the conventional framework and in the proposed localized analysis framework. Note that original raw data never leave the user's device. Researchers recieve only processed results.

will be utilized for the study, although no raw data will be collected without consent. The research app allows her to opt in and out from the study whenever she wants to, just like conventional studies.

While Jamie uses her own smartphone as usual, the app accesses and stores mobile data needed for the study (i.e., frequency of phone calls and text messages, location, and emotion ratings in text messages). Once the data reaches the predefined period (e.g., four weeks) or the amount of data needed (e.g., 100 text messages), data analysis runs on her smartphone when the phone is not in use (e.g., when the phone is recharging at night). The research app sends the analysis results to the cloud, hosted by the research hub, which collects multiple analysis results from the whole participants. In addition to the individual results, Jamie agrees to send additional demographic information (e.g., gender, age, and occupation). The researcher downloads all of the participants' results from the research hub website and analyzes the aggregated results.

Benefits of the new framework

This framework has many benefits for both participants and researchers. First of all, it better protects participant's privacy while utilizing their private information. Researchers do not need to see individual raw data such as survey responses, text messages, or photos. Second, it provides better security. Because data are distributed in each mobile device and not centralized at a location (e.g., a researcher's computer), it has a lower risk from a data breach. Third, it saves computational resources. It is surprisingly easy to recruit study participants via online sources. Even without compensation, many people have voluntarily participated in studies at internet websites; many sites draw thousands of participants, and some draw millions (e.g., projectimplicit.com, yourmorals.org). However, if a researcher would like to design a highly powered study with a large number of paticipants, it could require significant centralized computational resources. These computational resources can be crucial if a researcher would like to take advantage of various forms of data–texts, images, and videos.

Finally, the framework helps to achieve higher response rates to sensitive research topics. Study participants are likely to feel more comfortable and less offended when answering sensitive questions in this framework. Sexual behavior and drug use are known sensitive topics that have very low response rates (Tourangeau & Yan, 2007). Participants are more likely to respond to sensitive questions in a more candid way if they are assured that no one will see their original response.

Using smartphone features beyond the traditional questionnaire maximizes the benefits of the proposed localized analysis framework. Phone call logs and location information generated by a mobile phone are considered very private and people are reluctant to share these data with someone else. Or, even if study participants allow researchers to access personal data, such as photos and videos, it can be computationally expensive to collect multimedia data at the central storage and analyze them all together. The proposed framework provides means to maximize the benefits of the unique types of data generated

by smartphones.

Previous research

A few methods have been developed and used to protect participant privacy in survey research in the social and behavioral sciences. Unmatched count (Raghavarao & Federer, 1979) is a technique to preserve privacy by counting positive responses in multiple items (e.g., Ahart & Sackett, 2004; Coutts & Jann, 2011; Dalton, Wimbush, & Daily, 1994). Participants are randomly split into control and target groups, answer a series of binary questions, and report only the total number of positive responses to researchers. In this process, the target group answers one additional question, which is highly sensitive and the real interest of the research. Using the control groups result as a baseline, researchers can deduce the number of positive responses to the question of interest. Because participants report only the total number of positive response to all questions, their answer to the sensitive question remains unknown to researchers.

Instead of randomly assigning participants into two groups, as in unmatched count, randomized response (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; Warner, 1965) uses known probabilities as a baseline rate (e.g., Aoki & Sezaki, 2014; Lee, Sedory, & Singh, 2013; Ostapczuk, Moshagen, Zhao, & Musch, 2009). For example, researchers would ask participants to flip a coin and to give a positive response if the coin comes up heads (regardless of the truth) and a true answer if the coin comes up tails. Because the baseline probability of coin flipping is known, researchers can deduce the probability of positive responses to the question.

While these methods were developed to protect participants privacy during the data collection process, more recent advancement was initiated in order to conceal individual privacy during the data retrieval process after data are collected completely. In a database management context, a differential privacy framework was proposed by Dwork and her colleagues (Dwork & Roth, 2014; Dwork, McSherry, Nissim, & Smith, 2006). In the differen-

tial privacy framework, privacy is preserved by introducing randomness in query outputs. They suggested mechanisms that can be used for differential privacy, such as the Laplace mechanism for real values and the exponential mechanism for discrete values. Differential privacy aims to obtain statistical information about the population without revealing anything about individuals.

Although differential privacy inspired further methodological advancement (e.g., Chaudhuri & Monteleoni, 2009; Lei, 2011), the framework was introduced as a database query context and little work was done in a distributed multi-agent context. Recently, Boker et al. (2015) have suggested a framework to promote participants' privacy using mobile devices. In their suggestion, a pre-specified model is sent to a participant's device where data are stored and only model fit information is returned to a researcher's host computer. Then a model fit for the whole sample is calculated via the optimization process. The suggested method can effectively replace centralized data collection in traditional theory-driven data collection and analysis, where researchers have a strong belief in their hypotheses and fit a precisely specified model to data as confirmatory analysis.

However, the advantages of mobile data cannot be fully achieved by using only this confirmatory approach. Mobile data may not be easily modeled in advance (e.g., movement measured by accelerometer and gyroscope), and in fact they may not even be in a quantified format to be analyzed (e.g., texts and images). To take full advantage of mobile data, it is necessary to employ a flexible approach for data analysis.

Paradigm shift in data collection

Data acquisition in the social science tends to involve human administration. Traditionally, researchers initiate data generation by taking measures or asking questions. Social and behavioral scientists have used archival data (e.g., books and pictures that were already produced by people), but they are labor intensive–expensive to work with. To utilize texts and images, an army of research assistants is hired to rate and quantify them so that

researchers can use quantified summary data for analysis.

Nowadays, data are being created all the time. Some are actively produced by users with intentions. Some are generated by machines without users being aware. In either way, data already exist. We no longer must constrain ourselves to the traditional sense of data collection.

This shift in data collection paradigm brings up new research opportunities for the social and behavioral science. For decades, there was no other way to find what people think or how they behave other than asking directly (or measuring by indirect proxy). In recent years, a part of our lives has become constantly recorded and stored as digitized data. Using such data, we can observe how people naturally behave without intrusive questioning. For example, we could analyze emotions in text messages using text mining methods. Furthermore, since data are being generated all the time, we can study human behavior retrospectively. For example, we could have studied how people reacted when the 9/11 attack happened using their phone call logs, text messages, and internet search logs. This is a completely new type of study.

However, the challenge is that it is not easy to deal with this type of machine generated data only by traditional analysis methods. The data are not quantified, the formats are heterogeneous, and the model specification is not known. To utilize such data, a first step could be exploratory analysis, using machine learning and data mining.

The proposed framework can provide flexible applications using various analysis methods, both at an individual level and at an aggregated level. For example, one could analyze text messages using text mining methods at an individual mobile phone and visually explore the collected results at an aggregated level. Machine learning and data mining methods will play an important role in utilizing non-traditional data such as text messages, photos, and videos.

2. Localized Data Collection and Analysis

In localized data collection and analysis framework, a smartphone analyzes stored data about the user of the device and sends the analysis results to researchers. Researchers aggregate multiple individual results and interpret meanings of emerging patterns at the whole sample level. The framework is readily feasible to implement. Mobile phone apps for data collection are already developed and being used by researchers (e.g., Dufau et al., 2011; Froehlich et al., 2007; Killingsworth & Gilbert, 2010; MacKerron & Mourato, 2013). In terms of mobile computing technology, running data analysis on a smartphone is feasible. However, because this is a new framework, little is known about which analysis methods to use.

This dissertation investigates analysis methods for the localized analysis framework. I investigate potentially useful analysis methods at an individual level and at an aggregated level. The study focuses on exploratory analysis using machine learning in order to utilize various forms of mobile data, although the framework itself is not limited to these methods. The purpose of the study is to demonstrate the utilities of the framework.

I employ support vector machines (SVMs; Cortes & Vapnik, 1995; Vapnik, 1998) for a localized analysis on a mobile device and I examine aggregated results to find meanings at the whole sample level. Because of its simplicity to use and flexibility to different tasks (Bennett & Campbell, 2000), SVMs are one of the most widely used machine learning methods and have shown excellent performance in applications (Wang & Pardalos, 2015). Besides, SVMs can be easily understood by researchers in social science, where the framework will be used. In social and behavioral sciences, logistic regression and linear regression analysis are most widely used; an analogue can be made between logistic regression and linear regression and SVMs (for classification and for regression), with a

distinction in estimation mechanism. More details about SVMs will be presented later. Although SVMs were selected for this study, many other methods can be applicable in the proposed framework.

2.1 Assumptions and limitations

First, as a consideration for implementation, the characteristics of the processed results on individual devices play a crucial role in privacy protection. In the localized analysis framework, a smartphone analyzes locally stored data and sends the analysis results to researchers. The researchers receive only the processed results and never the original raw data. The nature of the processed results is the key to protecting user privacy.

Any analysis method creates multiple products (e.g., coefficients, standard errors, test statistics, or residuals, for linear regression); it is important to take into account that some of the products are more or less suitable to conceal private information than the others. The choice of processed results to be sent is an integral consideration in this framework. More details will be discussed with a specific case of this study in the Methodology section.

Second, a localized analysis at an individual device requires multiple observations per variable. Time-invariant variables such as demographic information (e.g., gender and age) cannot be incorporated at the localized analysis for an individual. Such variables need to be collected as traditional data collection process in order to be incorporated in analysis at an aggregated analysis for the whole sample.

Third, this study focuses on a data-driven approach (i.e., exploratory analysis) with the consideration of, and intention towards, an extension to heterogeneous mobile data. A hypothesis-driven approach based on model specification (i.e., confirmatory analysis) is not considered in this study. For confirmatory analysis that preserves privacy, consider the framework suggested by Boker et al. (2015).

Lastly, the analysis methods employed here are only a small set of candidate meth-

ods. The current study does not intend to provide an extensive range of analysis; the purpose of the study is to demonstrate the utilities of the localized data collection and analysis framework.

2.2 Significance of the study

Traditional data analysis assumes that analysts can use the data that are collected from the sample at any time for data analysis with no limitation. If access to the original data is limited, we need new analysis methods to find patterns and meanings without the full original data. This is the first attempt to analyze each individual's data locally, and examine the aggregated results to find patterns at the whole sample level.

Although the localized analysis framework has many benefits and can be readily implemented, there will be no practical use if we do not know what analysis methods to use and how to interpret the results. Investigating analysis methods to employ in the proposed framework is the foremost task in order to adopt the new framework to take advantage of mobile data.

It is especially beneficial to adopt machine learning and data mining methods to analyze mobile data because mobile data are different from traditionally well-defined measures. The traditional analysis methods widely used in social sciences are not suitable for non-quantified data (e.g., text messages, photos, and videos). Additionally, even when the data consist of numerical values, model specification will be challenging for non-traditional data such as accelerometer measurements.

The current study employs SVMs among many other machine learning methods because of their simplicity to use and flexibility to different tasks. Although this study uses quantified data for analysis, the SVM method can be applied to non-quantified data.

3. Literature Review of SVMs

SVMs comprise one category of machine learning algorithms for classification. In machine learning, a typical classification task is to predict class labels of new instances based on known instances. The basic idea of SVM is to construct a separating hyperplane with the maximum margin between two classes. SVMs apply a linear method to data in a high-dimensional feature space that is potentially non-linearly related to the input space.

SVMs have been successfully applied to a number of domains, such as computer vision (Han & Davis, 2012; Mohammed, Minhas, Jonathan Wu, & Sid-Ahmed, 2011; Rosten, Porter, & Drummond, 2010), bioinformatics (Che, Liu, Rasheed, & Tao, 2011; Inza et al., 2010; Saeys, Wehenkel, Geurts, et al., 2012; Upstill-Goddard, Eccles, Fliege, & Collins, 2013), geosceinces (Li, Kuo, Lin, & Huang, 2012; Mountrakis, Im, & Ogole, 2011; Pradhan, 2013), and finance and business (Huang, 2012; Yang, Deb, & Fong, 2011). Because of the strong performance and flexibility of the method, the SVM has been extended to tasks beyond classification, such as regression (Vapnik, 1998) and one-class classification for novelty detection (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001; Tax & Duin, 1999).

In this chapter, I will explain the basic mechanism of SVMs for classification, regression, and novelty detection (one-class classification). Further extensive explanation about SVMs can be found in Vapnik's books (Vapnik, 1998, 1999).

3.1 Classification

Linear separation

In classification, an SVM separates different classes of data by a hyperplane. In Figure 3.1, we want to find a hyperplane to separate negative instances from the positives. There are infinitely many hyperplanes to separate the two classes. The SVM finds the hyperplane that maximizes the gap between data points on the boundaries, that is, the distance of the two dashed lines as shown in Figure 3.2. There are always some points on the dashed line (because otherwise they could be pushed apart further); these points are called support vectors (marked in bold in Figure 3.2). Support vectors carry all information about the solution.

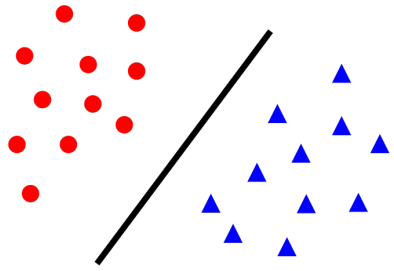

Figure 3.1: Classification task.

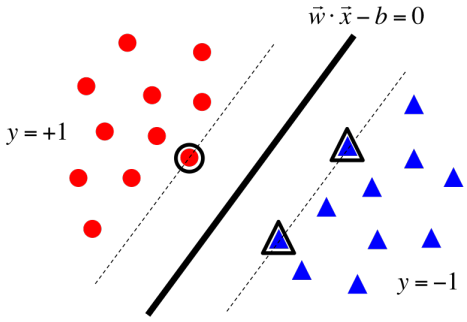

Figure 3.2: Illustration of basic idea of SVMs.

The gap between these supporting planes is $\frac{2}{||w||}$. Maximizing the margin is equiv-

alent to minimizing $\frac{1}{2}||w||^2$ in the following quadratic programming problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ s.t \quad & y_i(w \cdot x_i - b) \geq 1 \end{aligned} \tag{3.1}$$

The primal form of linear SVM shown above can be formulated as the dual problem by the Lagrange duality (Vapnik, 1998) as below:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i \cdot x_j - \sum_{i=1}^{m} \alpha_i \\ s.t \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & \alpha_i \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.2}$$

Although the primal optimization problem and the dual optimization problem yield the same solution ($w = \sum_{i=1}^{m} \alpha_i y_i x_i$), the dual problem is preferred due to its computational convenience using kernel methods, which will be illustrated later.

Finally, given a new instance, the class is decided by the sign of a decision value, $f(x) = sign(w \cdot x - b)$. The magnitude of the decision value indicates certainty of prediction. The decision values are close to 0 where the model is less certain about the class decision.

No perfect separation: soft-margin

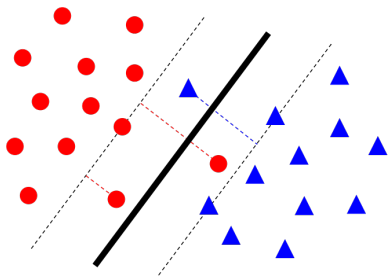

Figure 3.3: Illustration of soft-margin in SVMs.

Ideally, we want no points to be misclassified and no points to fall in the margin

between the two classes. Due to noise, data would not be perfectly separable in reality, as shown in Figure 3.3. Data points on the wrong side of its supporting plane are considered errors. In soft-margin' SVMs, a slack (or error) variable z weighted by C is added as a penalty. We want to maximize the margin as before while minimizing the error. Note that maximizing the margin regulates overfitting.

$$\begin{aligned} \min_{w,b,z} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} z_i \\ s.t \quad & y_i(w \cdot x_i - b) + z_i \geq 1 \\ & z_i \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.3}$$

The error z is weighted by C. When C is very large, the soft-margin SVM is equivalent to the hard-margin SVM. When C is smaller, we allow misclassifications in order for the margin to be larger. The cost parameter C must be chosen. The dual form of soft-margin SVM is (Cortes & Vapnik, 1995; Vapnik, 1998):

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i \cdot x_j - \sum_{i=1}^{m} \alpha_i \\ s.t \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & C \geq \alpha_i \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.4}$$

Non-linear separation: kernels

In Figure 3.4, the two classes are not linearly separable; we need a quadratic function, such as a circle, to separate them. A classic solution to convert a linear function into a nonlinear function is to simply add new features to the original data which are computed from the original features in a non-linear way. For example, the function of a circle needs x^2, y^2, and xy, in addition to the original data x and y. By applying the nonlinear function in the original input space, the linear function is applied in the feature space of expanded data.

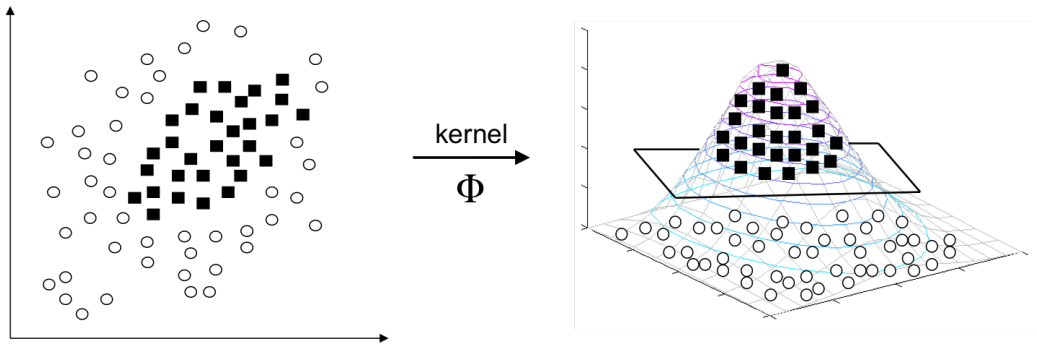

Figure 3.4: Illustration of the kernel trick (adopted from Statnikov et al., 2015). In the left, data are not linearly separable in the input space. In the right, data are linearly seperable in the feature space obtained by a kernel.

However, the dimensionality of the feature space may explode exponentially when adding all non-linear combinations of a certain degree. A way to get around this issue is using a special property of the evaluation function $\theta(x)$. Let us define $\theta(x) : R^n \rightarrow R^{n'} \quad n \gg n'$ and introduce this nonlinear mapping in the optimization problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \theta(x_i)\cdot\theta(x_j) - \sum_{i=1}^{m}\alpha_i \\ s.t \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & C \geq \alpha_i \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.5}$$

The mapping occurs only at the inner product. The inner product of the mapped points can be evaluated using the kernel function without explicitly defining the mapping (Cortes & Vapnik, 1995); this is often referred as the *kernel trick*. SVMs use an implicit mapping between the original input space and a high-dimensional feature space. Substitut-

ing the kernel into the dual form yields:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{m} \alpha_i \\ s.t \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & C \geq \alpha_i \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.6}$$

Using the kernel trick, we can apply the linear method to nonlinear problems. Data are linearly separable in the feature space that is implicitly mapped by a kernel. The performance of the kernel type depends on the type of application at hand. Some kernels frequently implemented in SVM softwares are given below. Parameters (γ, d, and c) must be chosen beforehand.

- Linear : $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + c)^d$
- Radial Basis Function (RBF): $K(x_i, x_j) = exp(-\gamma |x_i - x_j|^2)$
- Sigmoid: $K(x_i, x_j) = tanh(\gamma x_i^T x_j + c)$

Parameter tuning and scaling

The performance of SVMs depends on the parameters. It is necessary to find the optimal parameters to obtain the best results. In practice, it is widely suggested to do a grid search of a combination of a range of parameters (Chang & Lin, 2011; Karatzoglou, Meyer, & Hornik, 2005).

Another consideration for applying SVMs is that data should be scaled. Given that data are heterogeneous in domains where machine learning applies, the range of values of raw data can vary extremely across features. To prevent a feature with a wider range dominating variability in data, it is important to scale them into same ranges beforehand. A common practice is to standardize data with mean 0 and variance 1 for SVMs.

Summary of SVM classification

Figure 3.5 depicts data with three variables: social context (work or home), frequency of interpersonal contacts, and emotion ratings. Each observation (*vector*) is represented in the two dimensions of interpersonal contacts and emotion, while the social context is coded as circles and triangles. An SVM finds a linear decision surface (*hyperplane*) that can separate the two classes with the largest distance (*margin*) between borderline observation (*support vectors*). If such linear decision surface is not found in the original space (*input space*), the data is mapped into a higher dimensional space (*feature space*) where the separating decision surface exists. The feature space is constructed by mathematical projection (*kernel trick*).

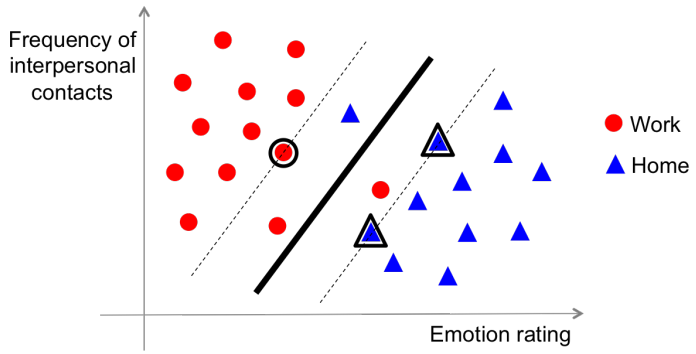

Figure 3.5: Illustration of SVMs.

3.2 Multiple classes

An SVM is a binary classifier that assigns either a positive label or negative one, which is decided by the sign of decision value of an instance. Many real-world problems, however, have more than two classes. When data has multiple classes, the class is decided by voting, using voting schemes such as one-against-one or one-against-all. In the one-against-all method, k binary classifiers are built to separate one class from the rest, where k is the number of classes. The class of an instance is decided by the classifier that produces the highest decision value for the instance. In the one-against-one classification method,

$\binom{k}{2}$ classifiers are built in $\binom{k}{2}$ subsets that contain only two classes of the data. The class of an instance is decided by the most frequently assigned class by the all pairwise classifiers.

3.3 Regression

The SVM approach can be applied to real values instead of just binary classes (Vapnik, 1998). The main idea of this so-called SV regression is to build a model to predict the outcome that allows a certain range (ϵ) of deviation. If the deviation between the actual value and the predicted value is less than ϵ, the deviation is not considered an error. As shown in Figure 3.6, it allows deviations in a band of size 2ϵ around the linear prediction; any points outside this band are considered errors.

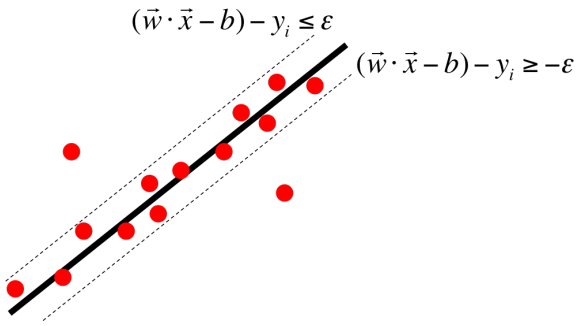

Figure 3.6: Illustration of ϵ-SV regression. Points outside the dashed band ($\pm\epsilon$) are errors.

As before, we want to maximize the band width in order to avoid overfitting. To account for errors outside of the band, we introduce slack variables z and z^*, which compute errors of overestimation and underestimation. Non-linear prediction can be made by the kernel trick as before. The primal form and the dual form of ϵ-SV regression are:

$$\begin{aligned} \min_{w,b,z,z_i^*} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}(z_i + z_i^*) \\ s.t \quad & (w \cdot x_i - b) - y_i \geq \epsilon - z_i \\ & (w \cdot x_i - b) - y_i \leq -(\epsilon - z_i^*) \\ & z_i, z_i^* \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.7}$$

$$\begin{aligned} \min_{\alpha,\alpha^*,b,z,z^*} \quad & \sum_{i=1}^{m}\alpha_i + \sum_{i=1}^{m}\alpha_i^* + C\sum_{i=1}^{m}z_i + C\sum_{i=1}^{m}z_i^* \\ s.t \quad & y_i - (\epsilon - z_i^*) \leq \sum_{j=1}^{m}(\alpha_i^* - \alpha_j)K(x_i, x_j) - b \leq y_i + (\epsilon - z_i) \\ & \alpha, \alpha^*, z, z^* \geq 0 \qquad i = 1, ..., m \end{aligned} \tag{3.8}$$

3.4 Novelty detection (one-class classification)

The SVM approach has been extended to detect novelty or outliers that are distinctively different from the majority. The main idea is to find a hyperplane, where data lie predominantly on one side, and none or a few lie on the other side (see Figure 3.7).

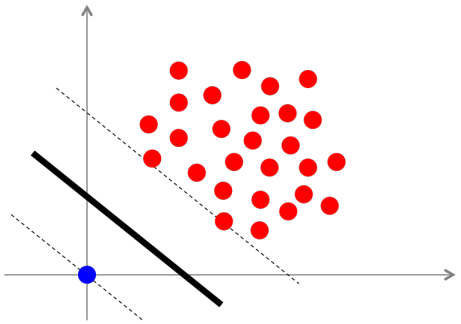

Figure 3.7: Illustration of novelty detection (one-class classification). The origin is the only member of the other class.

Again, we want to maximize the distance between two classes. We use a slack variable z just as we do in binary classification, but the decision function is constrained only to be positive except for small errors (z). We introduce ν ($0 \leq \nu \leq 1$), which controls the number of outliers. If the ν is smaller, the decision function allows less outlier and consequently finds a larger predominant class. The primal form and the dual form of one-

class classification are:

$$\begin{aligned} \min_{w,z,b} \quad & \frac{1}{2}||w||^2 + \frac{1}{m\nu}\sum_{i=1}^{m} z_i - b \\ s.t \quad & w \cdot x_i - b \geq -z_i \\ & z_i \geq 0 \\ & i = 1, ..., m \end{aligned} \tag{3.9}$$

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \\ s.t \quad & \sum_{i=1}^{m} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq \frac{1}{m\nu} \qquad i = 1, ..., m \end{aligned} \tag{3.10}$$

4. Methodology

4.1 Goal of the study

The purpose of this study is to demonstrate utilities of the proposed localized data collection and analysis framework. In this framework, a smartphone analyzes locally stored data about the user of the device and sends the analysis results to researchers. Researchers never collect the raw data; they only have access to the processed individual results. Researchers aggregate locally analyzed individual results and interpret meanings of emerging patterns at the whole sample level. The goal is to investigate exploratory analysis using machine learning and data mining in the proposed framework.

Exploratory analysis takes place when researchers have no preconceived behavioral patterns to look for. To look for patterns that may or may not exist, visualization is a useful approach to exploring data. The methods to locally analyze data at an individual level should be able to transfer data patterns to researchers without actual data so that researchers can visually inspect emergent patterns at the whole sample level.

Therefore it is crucial to employ an analysis method at the individual level that can transfer data patterns as accurately as possible to the whole sample level. I plan to employ SVMs for the localized individual analysis in order to capture individual data patterns. SVMs have shown high prediction accuracy across many application domains, for example, 87% in text categorization (Joachims, 2002), 91.4% in protein positions (Hua & Sun, 2001), 96.8% in face recognition (Moghaddam & Yang, 2002), and 96.8% in handwritten digit recognition (Decoste & Schölkopf, 2002). Although an SVM learns existing patterns and predicts unknown data successfully, it is not known whether multiple SVM models can reconstruct underlying patterns when aggregated.

The proposed method is multi-layered: localized analysis at an individual level and aggregated analysis at the whole sample level. Will the proposed analysis method show underlying patterns in the population? How accurately does aggregated individual analyses perform, in comparison to a single analysis using the whole data? Ultimately, can the localized analysis framework work as well as the conventional data collection and analysis paradigm? I will outline how to answer these research questions in the next section.

4.2 Methods

In the proposed framework, a smartphone analyzes locally stored data about the user of the device and sends the analysis results to researchers. The researchers receive only the processed results and never the original raw data. Therefore, the characteristics of the processed results play a crucial role in protecting user privacy. It is important to take into account that any analysis method creates multiple products and some of the products are more or less suitable to conceal private information than the others. The choice of processed results to be sent is an integral consideration in this framework.

In this study I suggest fitting an SVM and reconstructing individual patterns by the SV model on a smartphone, and sending reconstructed individual patterns to the research host and recovering the underlying overall patterns (see Figure 4.1). In this process, one might think that we could have the smartphone send an SV model instead of reconstructed individual patterns; it would be more efficient if smartphones send models and pattern reconstruction occurs at the research host. Indeed, that was the first inclination when designing this study.

Upon further thought, it became apparent that a model embeds raw information in itself in SVMs. As described in the Literature Review, SV models contain support vectors that are a part of raw data. Although support vectors exposed to researchers would be a relatively small portion given that mobile data may contain a myriad number of data

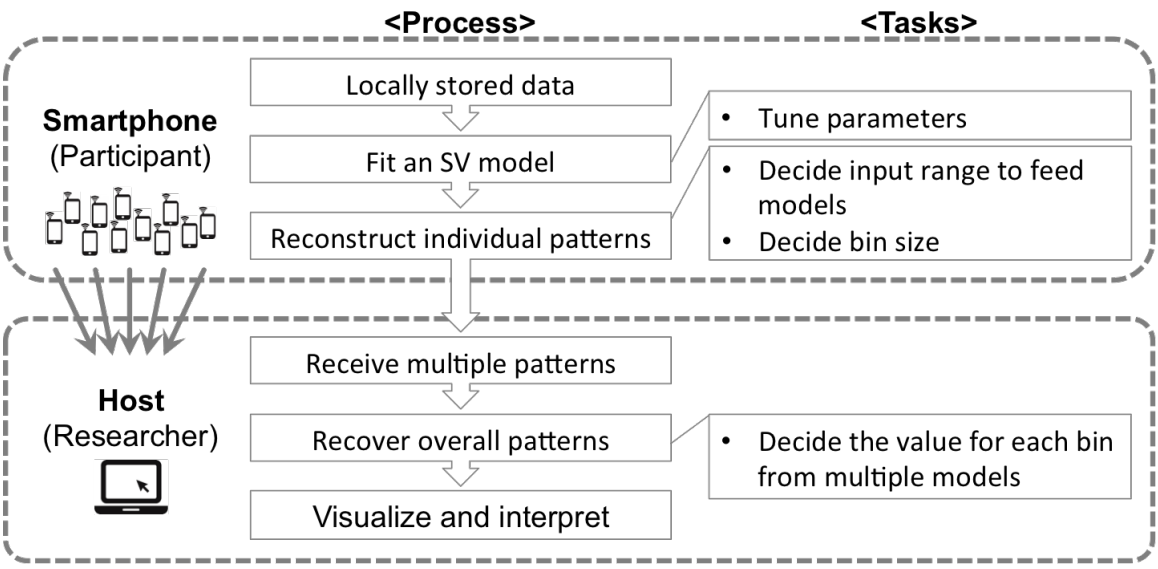

Figure 4.1: Flowchart of the data collection and analysis using SVMs. The Process shows steps of analysis with individual smartphones, as well as afterwards on the researcher's computer where it is aggregated. The Tasks shows related issues at each analysis step. Note that the server receives individual patterns from many participants and that the received multiple results will be aggregated into a single overall pattern. Also note that at no time the researcher has access to the original raw data of the participants, which allows the researcher to use even fairly private information.

points, this is not ideal for privacy protection. The choice of the processed results to be sent is crucial, as discussed earlier in the Proposal.

The proposed framework provides a simple solution: processing the analysis results even further on participants' devices. We can have participants' devices reconstruct individual patterns using SV models and send the reconstructed individual patterns to the research host, not the SV models that contain partial raw data. This alternative provides complete protection for privacy at the small expense of computational and network cost on smartphones.

This specific process using SVMs is designed to provide complete privacy protection. Of course, one can alter the process depending on analysis methods. In most analysis methods, user privacy will be preserved by only collecting models. The general process in the proposed framework is to: (1) fit an individual participant model on a smartphone (2) send individual models to the research host, and (3) recover the underlying patterns for the

whole sample. In the case that a model itself contains raw data like SVMs, at step (2) we could reconstruct individual patterns on a smartphone and send individually reconstructed data instead. The flexibility of the framework allows various approaches.

Overview of the study

The analysis process using SVMs in this study is that: (1) an SVM learns individual patterns at each user's smartphone, (2) the SV model reconstructs individual patterns at each user's smartphone, (3) researchers receive the reconstructed individual patterns from multiple participants, (4) the reconstructed individual patterns are combined into overall patterns at the whole sample level, and (5) researchers visually or analytically inspect and interpret the overall patterns (see Figure 4.1). Consequently, the effectiveness of visual inspection depends on the accuracy of reconstructed patterns.

There are several decision-making points that we should consider regarding pattern reconstruction and accuracy evaluation. Considerations should be given for parameter tuning, input data to reconstruct patterns, and accuracy evaluation, which I will discuss in detail in the following sections. Because this is a novel approach, there is no established convention to follow; we must investigate how different specifications could result differently in order to understand the results of pattern recovery, which is our main interest.

In Study 1, I will first investigate how analysis specifications influence pattern recovery. In Study 2, based on the results of Study 1, I will run simulation studies to investigate how well SVMs perform to reconstruct patterns at the whole sample level. I will evaluate the accuracy of the reconstructed patterns and show samples of visualization in Study 2.

4.2.1 Study 1

In Study 1, I will show how different analysis specifications could influence pattern reconstruction. In the following, I will first discuss considerations in analysis process and

evaluation of the method in detail. Second, I will describe study conditions to investigate the effects of analysis specifications.

Parameter tuning

To achieve best performance, SVM parameters must be tuned when an SVM runs on a participant's device. Idealy, it is best to tune parameters as precisely as possible to obtain better results, and it would take several minutes or longer on each device. In practice, this wouldn't cause problems for users, because we can opt it to run analysis when the smartphone is not in use (e.g., when being recharged). For this simulation study, however, it will cost time if we try to tune parameters very precisely, because a simulation consists of multiple SV models, which will be distributed across multiple smartphones in practice.

As parameter tuning costs significant computational time, I will initially show how parameter tuning can influence pattern recovery by comparing rough tuning results and fine tuning results. In rough tuning, each parameter will be tuned among two or three candidate values (e.g., 0.01, 0.1, or 1 for γ); in fine tuning, each parameter will be tuned by finer increments until no substantial improvement occurs in the results. This will let us understand how parameter tuning could influence pattern recovery. After that, the parameters will be only roughly tuned for each individual data in the rest of the study.

Input data

One of the challenges of this analysis method is that we do not have information about the original inputs. To reconstruct patterns on the researcher's side, not the smartphones, we need data to feed the SV models. The input data will be created as a grid space of the input features. The created input data may or may not exist in the input space of the original raw data and one should define a reasonable range of input. In this study, the range of $[-2, 2]$ in a standardized scale for each dimension will be used (approximately 95% of data falls in the defined range in the case of normal distribution).

Another consideration is that pattern recovery depends on the resolution of the grid that SVMs take as input. If the grid has fine resolution, SVMs show a precise picture of the pattern; if the grid has crude resolution, SVMs show a rough picture of the pattern (see Figure 4.2). However, a perfect recovery of original raw data does not mean successful recovery of behavioral patterns in social sciences because the original data contain noise in the first place. The goal is not a perfect recovery of original individual raw data, but rather a recovery of overall patterns at the whole sample level. The choice of resolution is up to the researcher. In this study, I will show a range of resolution and corresponding results so that one can make an informed decisions.

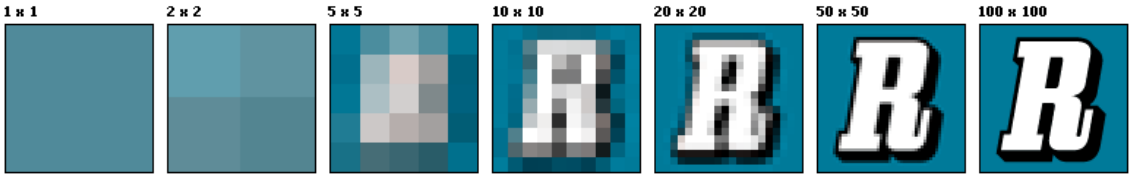

Figure 4.2: Illustration of grid size and resolution (Wikipedia, 2015). Pattern recovery depends on the resolution of the grid that SVMs take as input.

Classification results

On the researcher's side, we need to create an overall impression of all the classifiers received from the smartphones. One way is to create one overarching classification scheme that uses all classifiers. There are two different ways to decide the class of the interest with using multiple SVM models from individuals: voting or averaging decision values. Using the voting system, the class of an instance can be decided by the more frequently assigned class by the all classifiers. On the other hand, using decision values, of which sign decides the class, the class of an instance can be decided by the sign of average decision values across the all classifiers.

The results from the two mechanisms diverge if decision values are close to zero, that is, an SVM predicts classes with less certainty. For example, the two results wouldn't

agree if there are more positive classes with very small decision values and less negative classes with very large absolute decision values.

Once the patterns are reconstructed, I will evaluate the accuracy of the reconstructed patterns compared to the aggregated patterns of the original data. The original individual data exist separately and I need to combine them into one, in a comparable format to the reconstructed data, where the input space is a researcher-created grid. I will convert the input space of the original individual data into bins defined as the input space of the pattern reconstruction. This will result in multiple corresponding output values in the same bin. The final class of each cell is decided by the more frequent class across all outputs. Now the aggregated patterns are comparable to the reconstructed patterns and we can compute overall matching accuracy between the aggregated patterns and the reconstructed patterns.

Regression results

For SV regression, the predicted outcome by multiple SV models is averaged to reconstruct patterns. The original individual data is averaged for each bin as well, to construct the aggregated overall patterns at the whole sample level. However, unlike classification, the accuracy evaluation is more challenging in regression because the outcomes are real values (not classes as classification) and it is unlikely to find them exactly matching each other.

I suggest evaluating accuracy of regression results by permitting errors. I will first measure the difference between each pair of the reconstructed values and the original values. The values will be defined as matching if the difference falls in a certain range. The accuracy is the rate of the matching outcomes between the data. Note that the precision of accuracy depends on the size of error permitted.

In addition to the suggested approach, I will also report R^2 as an indicator of the portion of errors. Note that R^2 more severely penalizes for larger deviation whereas the suggested new approach does not reflect the magnitude of deviation. The two could yield

different results if original data include outlying values that are hardly captured in models.

The best judgement to compare patterns is human discretion, however. In this study, accuracy will be computed with an intention to provide a summary of massive simulation results. I will present sample visualized results at certain accuracies in order to help understand how numerical accuracy reflects actual similarity judged by human cognition.

Simulation conditions

The data will contain three variables (two predictors and one outcome) with arbitrary effects, 100 observations per person, and 100 individuals (which yield 100 SV models). There will be 200 replications per condition. For classification tasks, the following will generate data:

$$
\begin{aligned}
& z = -1 \quad if \quad x < y \\
& z = 1 \quad\;\; if \quad x > y \\
& where \quad x \sim N(0,1), \quad y \sim N(0,1), \quad e \sim N(0,\sigma^2)
\end{aligned}
$$

For regression tasks:

$$
\begin{aligned}
& y = x + z + e \\
& where \quad\;\; x \sim N(0,1), \quad z \in \{-1,1\}, \quad e \sim N(0,\sigma^2)
\end{aligned}
$$

The specific questions for classification are that: (a) How does parameter tuning influence pattern recovery? (b) How does bin size of input grid influence pattern recovery? (c) How does class decision method influence pattern recovery? The results will be presented in accuracy rates (Table 4.1).

In addition to (a) and (b) as in classification, the specific question for regression is that (c) how does accuracy evaluation reflect pattern recovery? The results will be presented both in accuracy rates (Table 4.2) and in sample visualized patterns in correspondence with the resulted accuracy. Additionally, R^2 will be reported.

Table 4.1: Study conditions for classification. The resulted accuracy rate of pattern recovery will replace dashes in the cells.

Rough tuning		Class decision	
		Majority vote	Average value
Input bin	0.05	-	-
	0.10	-	-
	0.15	-	-
	0.20	-	-
	0.25	-	-

Fine tuning		Class decision	
		Majority vote	Average value
Input bin	0.05	-	-
	0.10	-	-
	0.15	-	-
	0.20	-	-
	0.25	-	-

Table 4.2: Study conditions for regression. The resulted accuracy rate of pattern recovery will replace dashes in the cells.

Rough tuning		Permitted error size				
		0.05	0.10	0.15	0.20	0.25
Input bin	0.05	-	-	-	-	-
	0.10	-	-	-	-	-
	0.15	-	-	-	-	-
	0.20	-	-	-	-	-
	0.25	-	-	-	-	-

Fine tuning		Permitted error size				
		0.05	0.10	0.15	0.20	0.25
Input bin	0.05	-				
	0.10		-			
	0.15			-		
	0.20				-	
	0.25					-

4.2.2 Study 2

In Study 2, I will examine how accurately the proposed method reconstructs different patterns and how visualization can capture such patterns. Based on the results of Study 1, I will decide the bin size of input, the class decision method, and the accuracy measure to use in Study 2. Parameters will be roughly tuned as discussed. I will introduce non-linear patterns and different effect sizes in Study 2. The output of Study 2 is accuracy of pattern recovery across different data patterns and visualization of the reconstructed patterns.

Effects of features

In simulation studies in social and behavioral science, an effect of a variable is typically made in a shape of data patterns. For example, one could define an effect as a distance between two group means (i.e., Cohen's d). However, I would like to approach from another aspect in this study. This study aims to reconstruct underlying patterns as accurately as possible; an effect can be defined as distinctiveness of emerging patterns. Data patterns are less distinctive and less likely to be recognized if random errors (noise) dominate data. Figure 4.3 shows that the underlying patterns are less distinctive with a

larger variance in random errors added to a model that generates data.

For features at an individual level, I will define an effect as distinctiveness of emerging patterns in this study. I will manipulate variance (σ^2) in random errors, $e \sim N(0, \sigma^2)$, to control the effect sizes. With a larger effect size (i.e., a smaller error variance), it is easier for an SVM to learn data patterns and make a better prediction.

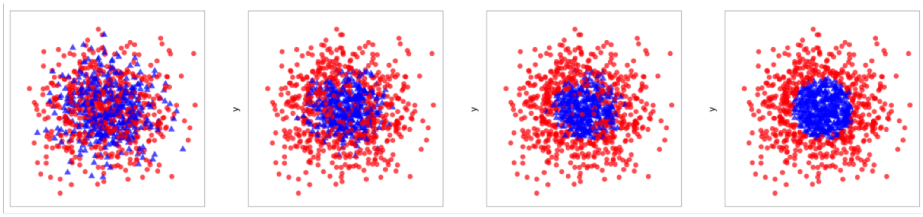

Figure 4.3: Illustration of data generation for classification. The distinctiveness of patterns depends on the random errors added to the model. $Error \sim N(0, \sigma^2)$.

Data patterns for classification

For classification tasks, the following will generate data.

$$
\begin{aligned}
z = -1 \quad & if \quad x^2 + y^2 < 1 + e \\
z = 1 \quad & if \quad x^2 + y^2 > 1 + e \\
where \quad & x \sim N(0,1), \quad y \sim N(0,1), \quad e \sim N(0, \sigma^2)
\end{aligned}
$$

This generates a circle inside another circle as shown in Figure 4.3. For example, a student's mood and frequency of text messages are moderate at home while they are extreme at school, as shown in Figure 4.3 (D). Error variance will manipulate effect sizes. With smaller random errors, the classes are more clearly separated (Figure 4.3).

Data patterns for regression

For regression, two non-linear patterns will be examined. For Pattern 1 (mean shift), the following will generate data:

$$y = \frac{1}{2}bz + bx^2 + e$$

$$where \qquad b = 0.3, \quad x \sim N(0,1), \quad z \in \{-1,1\}, \quad e \sim N(0,\sigma^2)$$

For Pattern 2 (opposite openings):

$$y = bx^2 z + e$$

$$where \qquad b = 0.3, \quad x \sim N(0,1), \quad z \in \{-1,1\}, \quad e \sim N(0,\sigma^2)$$

These generate parabolas as shown in Figure 4.4. For example, as for opposite openings, at home, positive emotion reflected in text messages increases along the frequency of text messages up to a certain point and decreases if one uses text messages excessively. On the contrary, at school, the patterns are in the opposite direction and positive emotion decreases up to a certain point and increases along the frequency of text messages. Error variance will manipulate effect sizes. With a larger effect size (i.e., a smaller error variance), the patterns emerge more clearly.

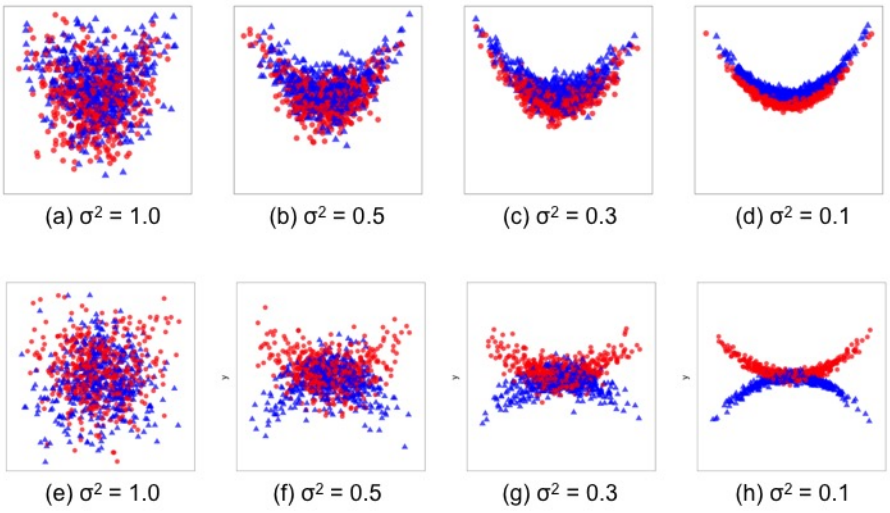

Figure 4.4: Illustration of data generation for regression. (a)-(d) Two U-shape parabolas with different mean levels (Pattern 1). (e)-(h) Two parabolas with opposite openings (Pattern 2). The distinctiveness of patterns depends on the random errors added to the model. $Error \sim N(0,\sigma^2)$.

Simulation conditions

The data will contain three variables (two predictors and one outcome), 100 observations per person, and 100 individuals (which yield 100 SV models). There will be 200 replications per condition. As described earlier, non-linear models will generate data. The effects will be made in terms of distinctiveness of patterns and it will be manipulated by random errors (noise). The results will be presented both in accuracy rates (Table 4.3) and in visualization of the simulation conditions. For regression, R^2 will be reported too.

Table 4.3: Study conditions for classification (left) and regression (right). The results of pattern recovery will replace dashes in the cells. Sample patterns will be plotted.

Effects	σ^2	Accuracy
Minimal	4.5	-
	3.5	-
	2.5	-
	1.5	-
Large	0.5	-

Effects	σ^2	Accuracy	R^2
Minimal	1.0	-	-
	0.8	-	-
	0.6	-	-
	0.4	-	-
Large	0.2	-	-

5. Results

5.1 Study 1

5.1.1 Classification

Parameter tuning: rough tuning vs. fine tuning

As shown in Table 5.1, parameter tuning did not show differences in pattern recovery accuracy in this study. The fine tuning condition did not show higher accuracy than the rough tuning condition. Therefore, in the rest of the Classification Results section, I will show only the rough tuning results. Future consideration and implication will be discussed later.

Class decision method: majority voting vs. averaging decision values

As shown in Table 5.1, the accuracy of averaging decision values was greatly lower than the accuracy of majority voting. SVM models did not reconstruct patterns accurately when the class was decided by averaging method. Therefore, in the rest of the Results, I will show only the majority voting results.

Bin size

As the bin size increased, the resulted accuracy increased. Note that the SVM models are identical across bin sizes; the accuracies vary because of the different input data (defined by bin sizes), not due to the model's prediction performance.

Table 5.1: Pattern recovery accuracy by class decision methods and by tuning methods.

Input bin size	Voting		Averaging	
	Rough tuning	Fine tuning	Rough tuning	Fine tuning
0.01	79.13%	79.12%	53.07%	52.89%
0.05	84.12%	84.12%	53.53%	53.39%
0.10	90.29%	90.28%	54.11%	53.99%
0.15	92.78%	92.83%	54.29%	54.13%
0.20	94.17%	94.18%	54.37%	54.22%
0.25	95.01%	95.01%	54.51%	54.44%

Accuracy of aggregated results from multiple SVMs

The current study suggests locally analyzing individual data and combining the multiple results from the individually trained models. To evaluate the effectiveness of the proposed method, I need to compare this approach to the conventional approach, that is, collecting and analyzing the whole raw data. However, direct comparison between the two approaches are not suitable: the input data are different (binned data and raw data) and therefore the predicted outcomes (reconstructed patterns) are not comparable. In addition, the accuracy of proposed method varies based on bin sizes and there is no standard for which bin size is comparable to raw data. Instead, I compare the two approaches using the same binned input data. This comparison shows, under the same condition (by using binned data), what can be achieved if the whole data were collected. The Table 5.2 shows the aggregated results from the individually-trained multiple SVM models (as in the proposed framework) and the accuracy results from an SVM model built on the whole raw data (as in the conventional paradigm). Note that the models in the two approaches are built on different data, but that they produced outcomes (accuracy of the reconstructed data) using the same binned data. As shown in Table 5.2, although an SVM model based on the whole data shows slightly higher accuracy than multiple individual SVMs, the differences are very small (less than 1% across different bin sizes). For comparison, I present results of using raw data (not binned data). When using raw data, the aggregated results of multiple individual SVM models showed 82.82% accuracy, whereas an SVM model trained on the

whole raw data showed 80.48% accuracy.

Table 5.2: Accuracy comparison of the aggregated results from multiple individual models and the single result from a model built with the whole data.

Data type		Model	
		Aggregated SVMs	Single SVM
Binned data (bin size)	0.01	79.13%	79.19%
	0.05	84.12%	84.19%
	0.10	90.29%	90.38%
	0.15	92.78%	92.95%
	0.20	94.17%	94.35%
	0.25	95.01%	95.25%
Raw data		82.82%	80.48%

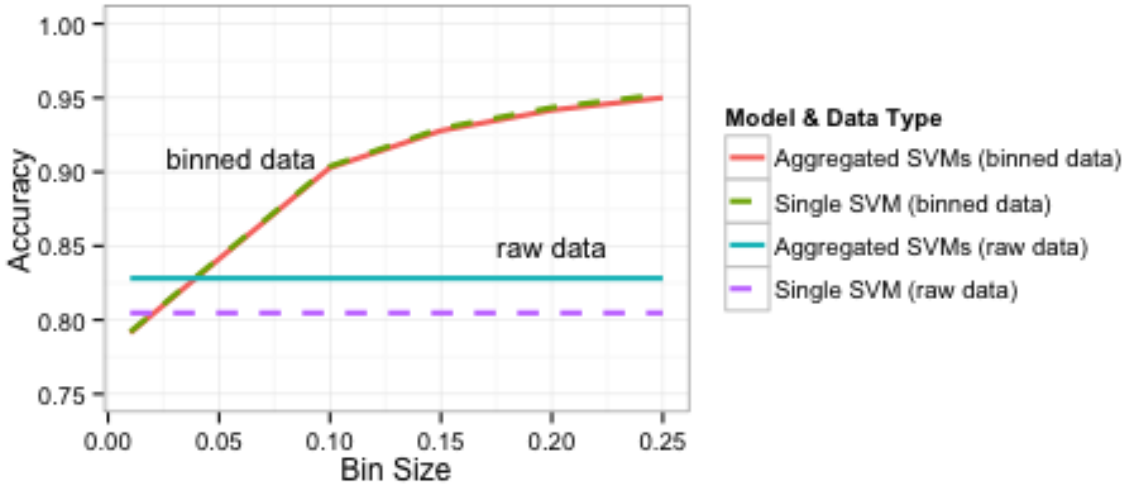

Figure 5.1: Pattern recovery accuracy by bin sizes.

Visualization

The outcomes of this study are visualized data patterns to enable exploratory analysis using this framework. Figure 5.1-(d), (e), and (f) show sample visualizations of the results of three bin sizes. Although, in all bin sizes, the decision boundaries were depicted correctly as in the visualized true model shown in Figure 5.2-(b) and (c), bin sizes larger than 0.10 showed crudeness in resolution. Note that the original raw data and the reconstructed data do not necessarily look similar due to the noise in the raw data. A good method

will allow investigation of underlying mechanisms of phenomena, thus resemblance of patterns between the reconstructed data and the true model are desired, not a representation of the original raw data. The original raw data shown in Figure 5.2-(a) looked more dense in the middle and less so in the outlying area because the raw data was simulated based on a normal distribution, which may or may not be the case in real mobile data. The reconstructed patterns by SVMs are based on binned input data, which are distributed uniformly between -2 to 2 with an interval of a certain bin size. The recovered patterns by SVMs can be cosmetically more similar to the input data if different input data (based on a normal distribution) were used, or if the original raw data were generated differently (based on a uniform distribution). For comparison, the true model based on a uniform distribution is presented as an example in Figure 5.2-(c).

5.1.2 Regression

Parameter tuning: rough tuning vs. fine tuning

As in the classification results, parameter tuning did not show differences in pattern recovery accuracy and R^2 in this study (Table 5.3). The fine tuning condition did not show higher performance over the rough tuning condition. Therefore, in the rest of the Regression Results section, I will show only the rough tuning results. Future consideration and implication will be discussed later.

Table 5.3: Accuracy and R^2 by tuning methods. The accuracy is averaged across a range of permitted errors from 0.1 to 1.0 in this result.

Input bin size	Accuracy		R^2	
	Fine tuning	Rough tuning	Fine tuning	Rough tuning
0.01	77.68%	77.66%	94.21%	94.20%
0.05	92.36%	92.34%	98.91%	98.90%
0.10	95.48%	95.40%	99.43%	99.41%
0.15	97.00%	96.92%	99.62%	99.60%
0.20	97.38%	97.28%	99.68%	99.67%
0.25	97.90%	97.82%	99.74%	99.72%

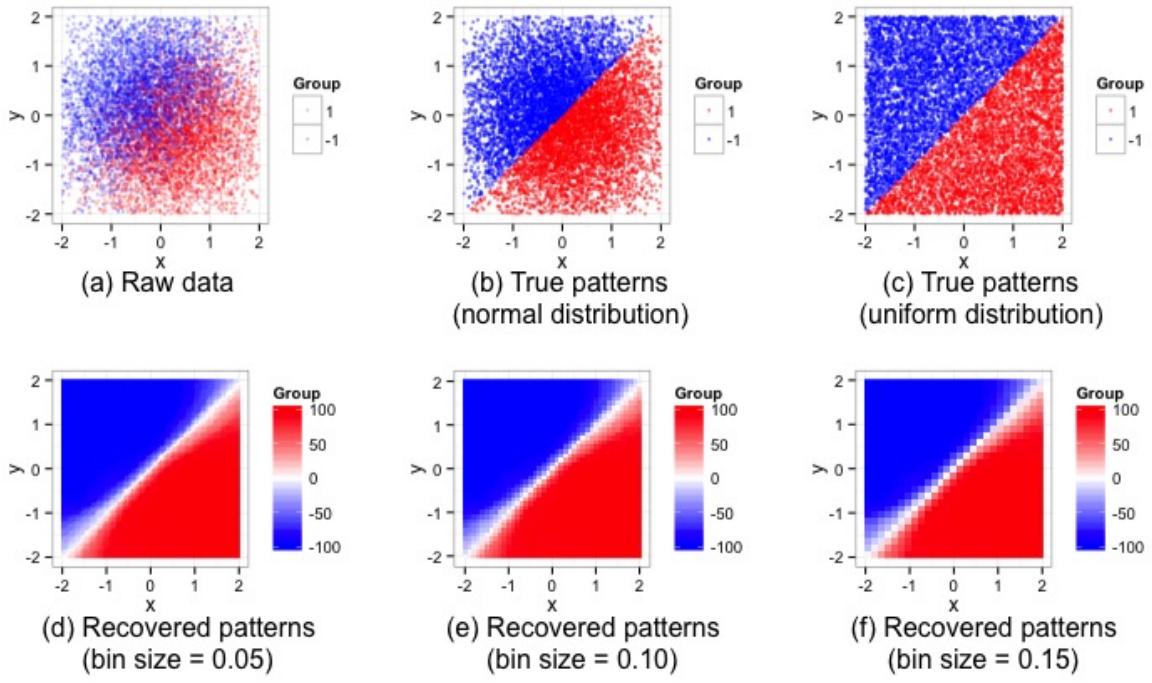

Figure 5.2: Sample visualization of the original raw data, the true model, and the reconstructed data by SVMs within the localized analysis framework (55th simulation in the study). Note that the original raw data and the reconstructed data do not necessarily look similar due to the noise in the raw data. We seek resemblance of patterns between the true model and the reconstructed data to evaluate the framework. (a) The original raw data. All individual data are combined. Each individual data was simulated based on a normal distribution. (b) Visualized patterns of the true model that generated the raw data based on a normal distribution. (c) Visualized patterns of the true model based on a uniform distribution. (d) Visualized patterns of the reconstructed data with bin size = 0.05. (e) Reconstructed data with bin size = 0.10. (f) Reconstructed data with bin size = 0.15.

Bin size

As in the classification results, when the bin size increased, the resulting accuracy increased (Table 5.4). Note that the SVM models are identical across bin sizes; the accuracies vary because of the different input data (defined by bin sizes), not due to the model's prediction performance. Implications will be discussed later.

Performance measure: regression accuracy using permitted errors

Along with R^2, the regression prediction accuracy was measured using a range of permitted errors (Figure 5.3). Prediction accuracy varies more widely than R^2. Note that

Table 5.4: Pattern recovery accuracy and R^2 by bin sizes and by permitted error sizes.

Input bin size	Error size for regression accuracy							R^2
	0.01	0.05	0.10	0.20	0.30	0.40	0.50	
0.01	2.62%	12.89%	25.57%	47.88%	65.38%	77.47%	85.51%	94.20%
0.05	5.71%	28.03%	52.40%	81.92%	93.42%	97.34%	98.92%	98.90%
0.10	7.91%	37.87%	66.20%	91.12%	97.50%	99.35%	99.86%	99.41%
0.15	9.97%	45.59%	75.12%	95.31%	99.02%	99.81%	99.95%	99.60%
0.20	11.01%	50.08%	78.04%	95.90%	99.05%	99.86%	100.00%	99.67%
0.25	11.75%	53.12%	82.04%	96.78%	99.47%	99.94%	100.00%	99.72%

the SVM models are identical across different error sizes; the different numerical results do not indicate different performance across error sizes in each bin size. The differences come from how to measure the prediction accuracy. In this study, error size 0.1 and 0.2 correspond respectively to 2.5% and 5% of the data range (-2 to 2). In the rest of the results section, I will only report the results of permitted error size 0.1 along with R^2. Implications will be discussed later.

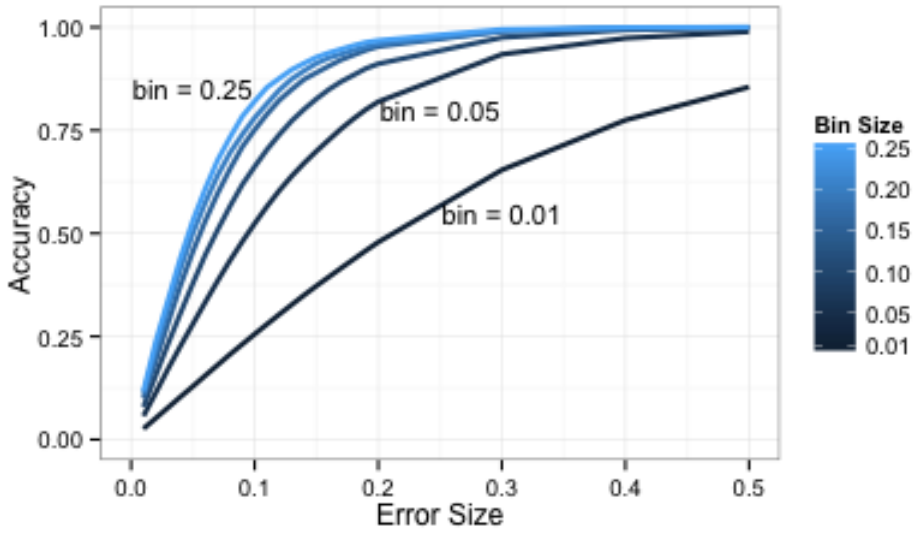

Figure 5.3: Pattern recovery accuracy by sizes of permitted error and bin.

Accuracy of aggregated results from multiple SVMs

The performance of the proposed framework and the conventional paradigm is next compared using the same binned input data. The pattern recovery accuracy of the localized analysis framework is evaluated based on the aggregated results from multiple SVM models individually built on each individual's data; the accuracy of the conventional paradigm is evaluated based on one SVM model built on the whole data. In addition to using binned input data, I compared multiple individual models with one model using raw input data too. This comparison shows that how locally analyzing small chunks of data behaves compared to analyzing the whole data. The Table 5.5 shows the aggregated results from the individually-trained multiple SVM models (as in the proposed framework) and the accuracy results from an SVM model built on the whole raw data (as in the conventional paradigm). Note that the models in the two approaches are built on different data (one is a subset of the other), but that they produced outcomes (accuracy of the reconstructed data) using the same binned data. The resulting comparisons between the aggregated results from multiple models and the single result from one model are very similar to the classification results (Figure 5.4 and 5.5). Although an SVM model based on the whole data shows slightly higher accuracy and R^2 than multiple individual SVMs, the differences are small. When using raw data, the aggregated results of multiple individual SVM models showed slightly better performance than an SVM model trained on the whole raw data. However, unlike the classification results as shown earlier, the differences in performance (either accuracy or R^2) were large between binned data and raw data. Implications will be discussed later.

Visualization

Figure 5.6-(c), (d), and (e) show sample visualizations of the results for three bin sizes. The regression predictions were depicted correctly as in the visualized model shown in Figure 5.6-(b). Unlike classification, data generation process (which probability distribution used) did not affect visualization results of regression, because the visualized patterns

Table 5.5: Comparison of the aggregated results from multiple individual models and the single result from a model built with the whole data. Accuracy error size = 0.10.

Input data type		Accuracy		R^2	
		Aggregated SVMs	Single SVM	Aggregated SVMs	Single SVM
Binned data (bin size)	0.01	25.57%	25.58%	94.20%	94.27%
	0.05	52.40%	53.52%	98.90%	98.97%
	0.10	66.20%	68.30%	99.41%	99.49%
	0.15	75.12%	78.22%	99.60%	99.67%
	0.20	78.04%	82.19%	99.67%	99.74%
	0.25	82.04%	86.21%	99.72%	99.80%
Raw data		9.79%	7.93%	67.23%	66.65%

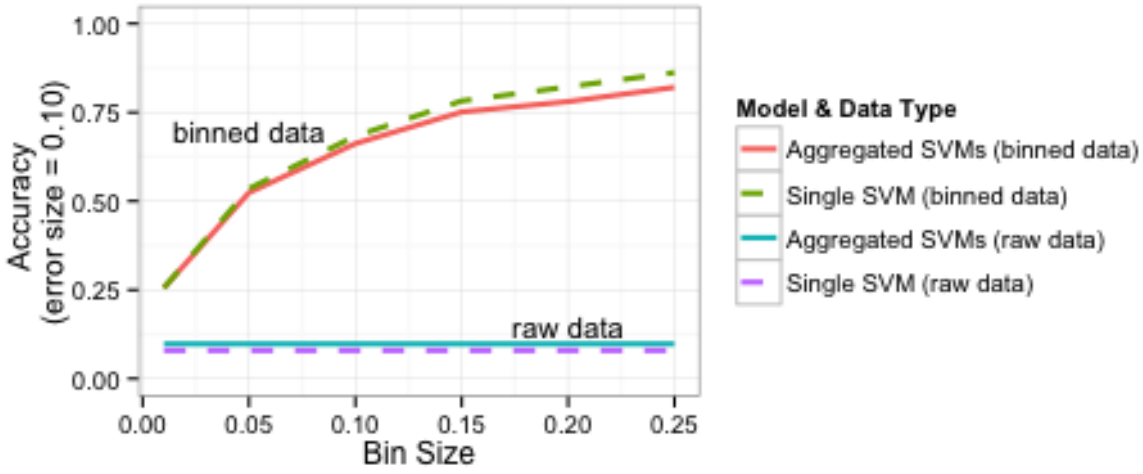

Figure 5.4: Pattern recovery accuracy by bin sizes. Error size = 0.10.

of outcomes (predicted values) are sparser in regression than in classification.

5.1.3 Discussion

Parameter tuning

In both classification and regression, contrary to the typical expectation, more precise parameter tuning did not improve the results. However, this may be the case due to the simplicity of the models in this study, where only three variables are included and the true models are simple. It is possible that more precisely tuned parameters could improve the accuracy of SVMs depending on data features, which researchers are interested in. In

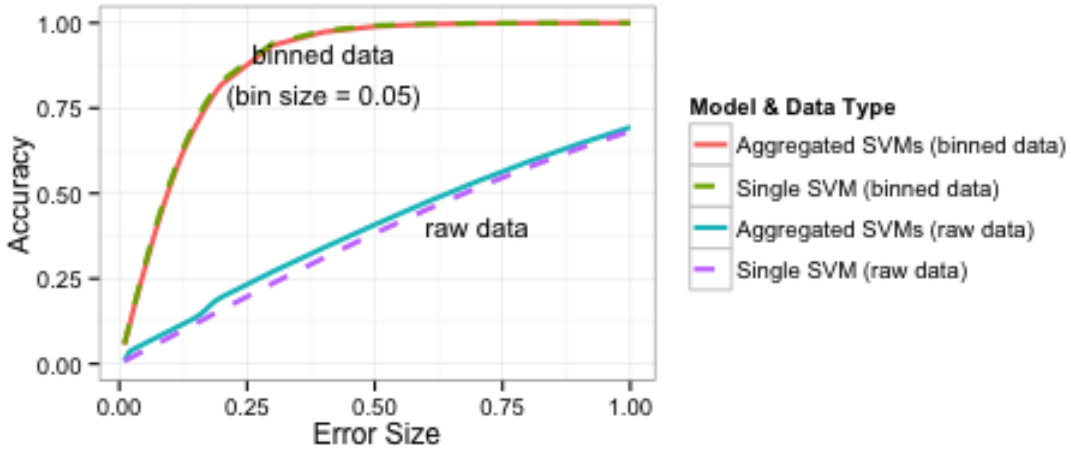

Figure 5.5: Pattern recovery accuracy by error sizes. Bin size = 0.05.

practice, before running studies at scale, I suggest running a pilot study to examine the costs and effects of parameter tuning using data features of interest.

Accuracy measure for regression

In this study, I introduced a new measure of regression prediction performance. By accepting a certain range of errors, we can evaluate the accuracy of real value predictions. Unlike R^2, this accuracy measure allows us to control precision of the accuracy measure. For example, in this study, the R^2s were already close to 1 (ranged .942 to .997) and they couldn't precisely differentiate performances across conditions and models. By choosing error size 0.1, the study was able to evaluate the results in a more interpretable way. Although the proposed regression accuracy measure and the conventionally used R^2 show different results, the actual performance of model predictions are identical. Note that the differences come from the metric of performance measure, not the actual performance. Also note that choice of error size can be post-hoc and has no significant computational cost for computation.

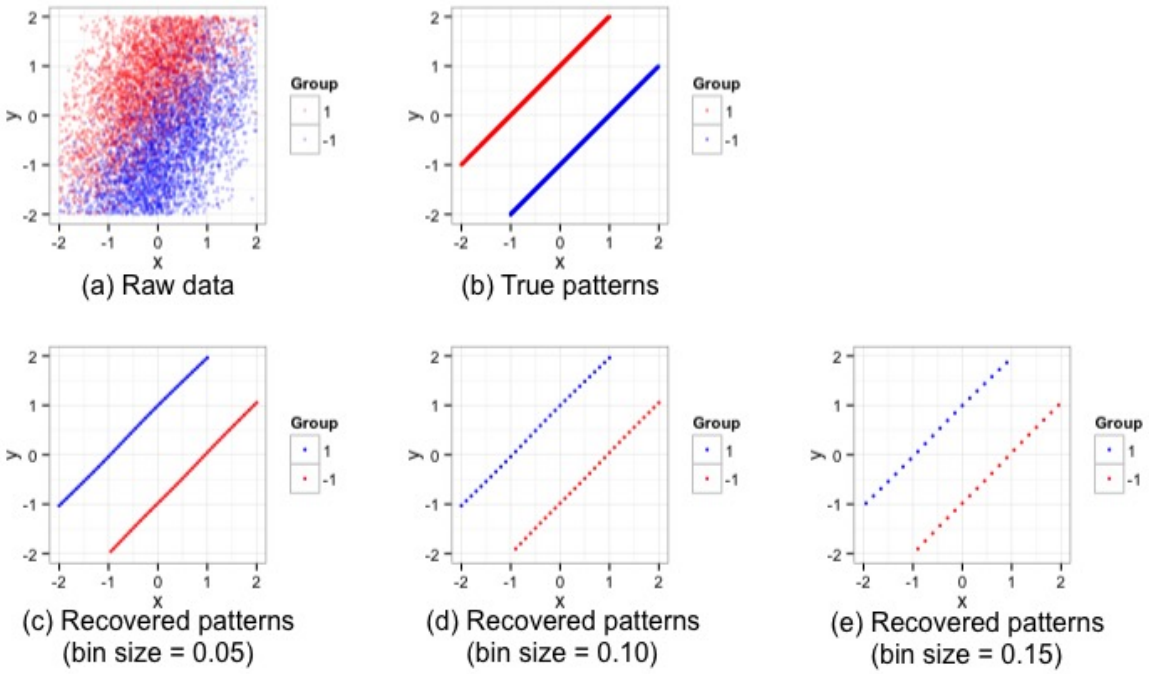

Figure 5.6: Sample visualization of the original raw data, the true model, and the reconstructed data by SVMs within the localized analysis framework (166th simulation in the study). (a) The original raw data. All individual data are combined. (b) Visualized patterns of the true model. (c) Visualized patterns of the reconstructed data with bin size = 0.05. (d) Reconstructed data with bin size = 0.10. (e) Reconstructed data with bin size = 0.15.

Effects of binning

The measured performance increased as bin size increased. However, this does not mean the SVM models performed differently: the models are identical across bin sizes. The differences come from how to create binned original data and reconstructed data using the models. The bin size defines the resolution of binned original data and reconstructed data while the accuracy of pattern recovery depends on the resolution of data defined by bin sizes. I reason that the accuracy of pattern recovery depends on the binned raw data, which the reconstructed data are compared with in order to evaluate the performance of SVM models. Binning reduces noise in data. As shown in Figure 5.7, the binned raw data has less noise than the original raw data and resembles the true model. This results in the high accuracy of the reconstructed data, which resembles the true model as shown in Figure 5.7.

One might wonder if we could simply bin data as a means of processing original

raw data. Although it could work to a certain degree, it can reveal raw data if we choose a very small bin size so that only a datum fits in a cell. As suggested in the proposed framework, when we use a predictive model to reconstruct data patterns, it is not possible to triangulate original raw data by binning. The original raw data are protected because we do not use the original raw data as input data and thus the reconstructed patterns are generated by models using researcher-defined input data. The original raw data does not take part in this framework after individual models are trained with the raw data.

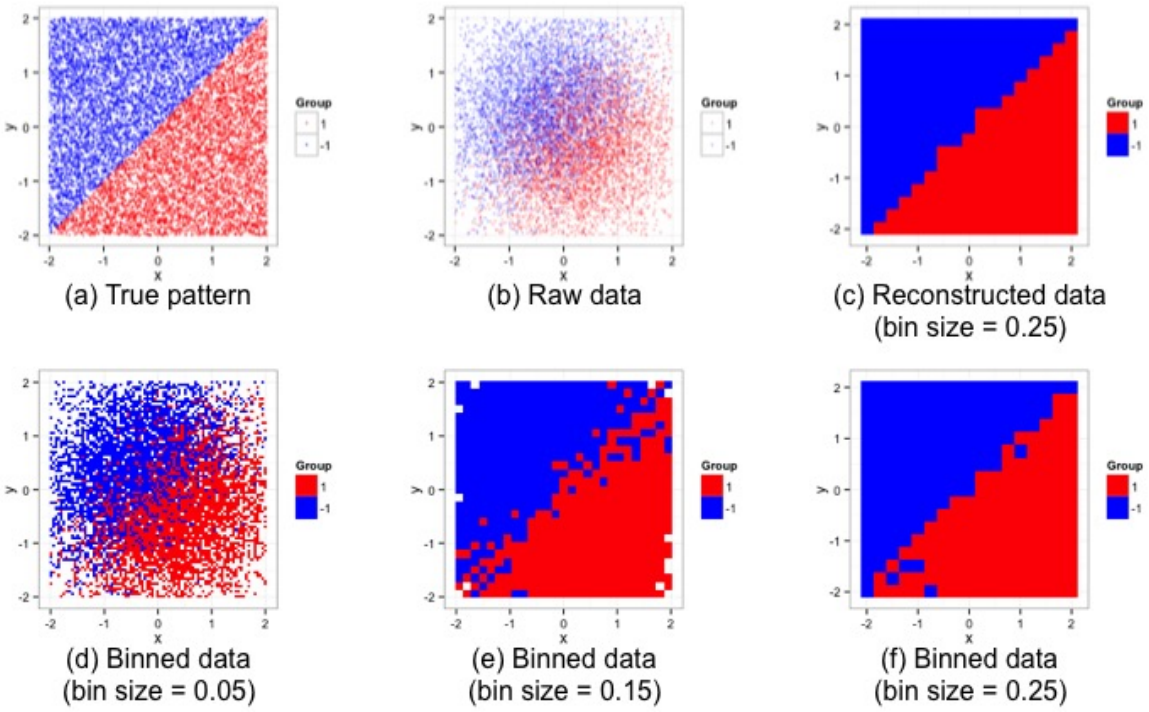

Figure 5.7: Comparison of the true patterns, raw data, binned data, and reconstructed data by SVMs. The data patterns of binned data (d, e, and f) resemble the true patterns (shown in a) by simply binning. With a larger bin size (shown in f), the binned data look more like the true underlying patterns. The reconstructed data (shown in c) results in high accuracy because the binned data it is to be compared with looks more similar than the raw data does. (a) The true underlying patterns (a uniform distribution). (b) The raw data simulated based on the true model (a normal distribution). (c) The reconstructed data by SVMs (bin size = 0.25). (d) Binned data (bin size = 0.05). (e) Binned data (bin size = 0.15). (f) Binned data (bin size = 0.25).

Classification and regression

The binning effects can explain the differences in the classification results and the regression results. Whereas the aggregated individual results and single model results in classification were seemingly comparable between binned data and raw data, there was a huge gap in the results between the two data types in regression. As I explained earlier, however, raw data results and binned data results are not comparable since the input data that was used to generate patterns were different. Even so, the regression results showed unexpected differences between the aggregated individual results and the single model results, compared to the classification results. In classification, because an outcome is binary (two bins), I suspect that the classification task (assigning into one of the two bins) itself functions as a noise filter. In regression, because an outcome is a real value, nothing filters the noise.

Aggregated individual models vs. single model with the whole data

The most important point is that there was no practical difference between the results of aggregated individual models (as in the proposed framework) and the results of a single model using the whole data (as in the conventional paradigm). This is the case both using binned data and using raw data. These results suggest that the localized analysis framework has similar analytical power compared to the conventional paradigm. In fact, when raw data is used, aggregated individual models show higher performance than a single model. This was unexpected. I expected that a model with the whole data would perform better than multiple models with subsets of the data. This opens new possibilities in research methodology especially with big data, which is computationally difficult to process with a single machine. For future research, it is worth investigating whether this result holds with smaller data (fewer participants or fewer observations per participant).

Input data type

Although binned input data were used in this study, any data can be used in application. Binning was employed for raw comparison data, and input data to reconstruct data from given models. By exactly matching features after binning, it was possible to compare the raw data and the model generated data, and to compute the pattern matching accuracy. In practice, however, we no longer need binning and researchers are free to choose any type of data as input in order to recover underlying patterns. Consider that, as shown in the results, the aggregated individual results show higher accuracy than the single model results when using real values as input. Input data to feed locally trained models can be binned just as in this study, or randomly generated based on a normal or other distributions. The choice of data distribution will depend on data of interest from the mobile device. One possibility is to estimate underlying distributions of data on smartphones. The effective choice of input data type requires further research.

Performance compared to the true model

Although this study measured the pattern recovery accuracy in comparison to raw data (generated based on a true model), we could measure the accuracy compared to the true model to evaluate local analysis methods. The analysis goal is ultimately to find an underlying true model that generates raw data, rather than raw data itself. Even if reconstructed data patterns do not resemble raw data, the methods are good solutions if they effectively represent the true models. In future research, it is worth examining the accuracy compared to true models. We can measure the accuracy compared to true models without binning and we can easily compare different types of input data.

5.2 Study 2

5.2.1 Classification

As shown in Table 5.6 and Figure 5.8, the classification accuracy increased as the distinctiveness of patterns increased (less noise in data). The aggregated individual SVMs and the single SVM built with the whole data showed very minimal difference in the performance. With raw data, the aggregated SVMs showed higher accuracy when the patterns are less distinctive than the single SVM. The difference between the aggregated SVMs and the single SVM decreased as the variance decreased (more distinctive patterns). Sample patterns are visualized in Figure 5.9.

Table 5.6: Pattern recovery accuracy comparison of the aggregated results from multiple individual models and the single result from a model built with the whole data. Bin size = 0.05.

Data type	Condition		Model	
	Effects (patterns)	σ^2	Aggregated SVMs	Single SVM
Binned data	Minimal (less distinctive)	4.5	61.56%	61.75%
		3.5	64.53%	64.71%
		2.5	69.13%	69.38%
		1.5	77.07%	77.26%
	Large (more distinctive)	0.5	91.14%	91.22%
Raw data	Minimal (less distinctive)	4.5	69.80%	61.29%
		3.5	70.83%	63.84%
		2.5	73.16%	67.88%
		1.5	78.48%	74.86%
	Large (more distinctive)	0.5	92.34%	90.22%

5.2.2 Regression

Pattern 1: Mean shift

As shown in Table 5.7, Figure 5.10, and Figure 5.11, the accuracy and R^2 increased as the distinctiveness of patterns increased (less noise in data). Remember that the accuracy

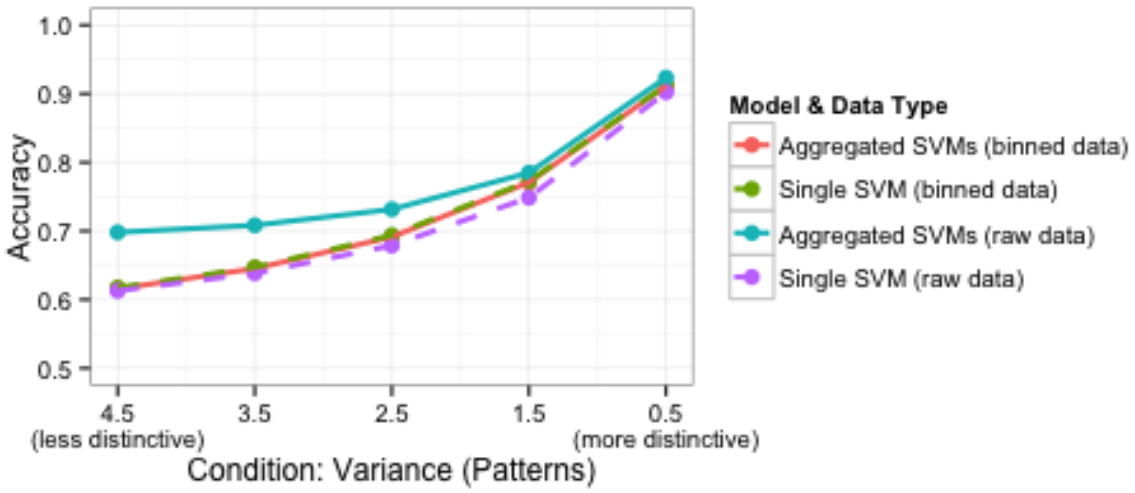

Figure 5.8: Pattern recovery accuracy by conditions. Bin size = 0.05.

or R^2 of two methods should be compared with the same type of data. The results of binned data and raw data are not comparable. Either by accuracy or R^2, the performance of the aggregated individual SVMs was on par with a match with the single SVM built with the whole data. Sample patterns are visualized in Figure 5.12. Regardless of the fuzziness of the raw data, the recovered patterns represent the true model very well.

Table 5.7: Pattern recovery accuracy and R^2 for the pattern of mean shift. The aggregated results from multiple individual models and the single result from a model built with the whole data are compared. Bin size = 0.05 and error size = 0.10.

Data type	Condition		Accuracy		R^2	
	Effects	σ^2	SVMs	Single SVM	SVMs	Single SVM
Binned data	Minimal	1.0	52.57%	53.28%	85.71%	86.64%
		0.8	62.69%	62.97%	90.20%	90.69%
		0.6	74.95%	75.91%	94.36%	94.67%
		0.4	89.42%	89.81%	97.43%	97.54%
	Large	0.2	99.09%	99.11%	99.35%	99.37%
Raw data	Minimal	1.0	9.93%	7.97%	19.66%	16.92%
		0.8	13.22%	9.98%	26.32%	24.07%
		0.6	16.97%	13.28%	37.30%	36.05%
		0.4	23.92%	19.84%	55.67%	55.81%
	Large	0.2	42.70%	38.36%	82.90%	83.48%

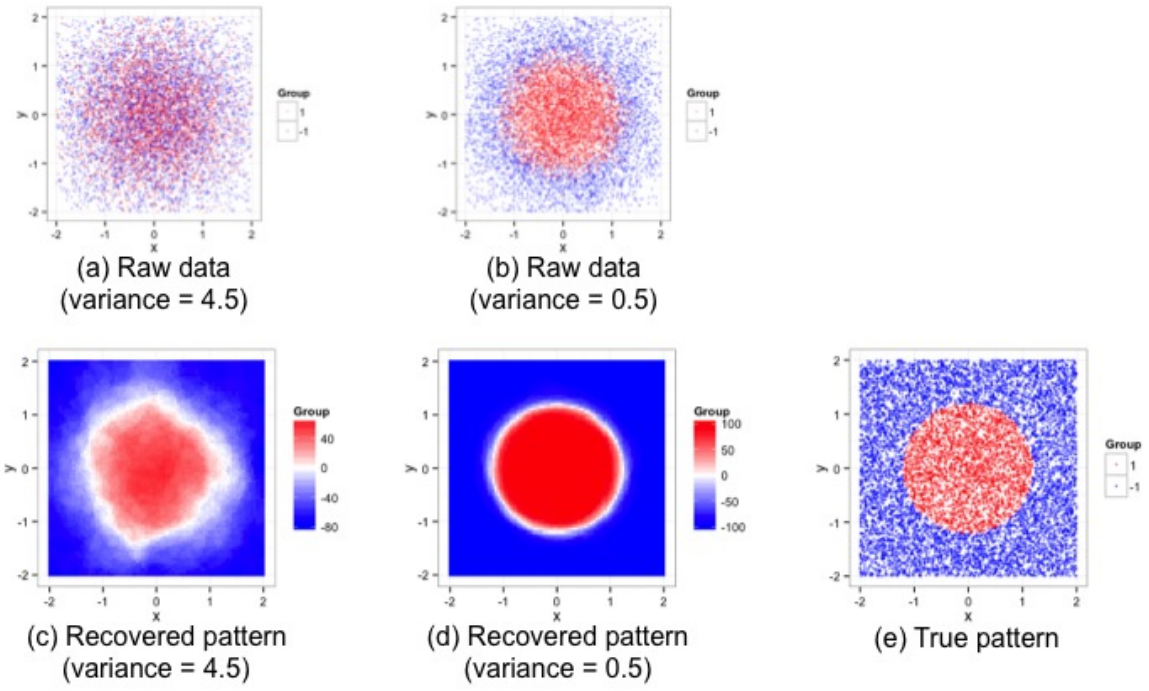

Figure 5.9: Sample visualization of the original raw data, the reconstructed data by SVMs within the localized analysis framework, and the true underlying patterns. Note that the original raw data and the reconstructed data do not necessarily look similar due to the noise in the raw data. We seek resemblance of patterns between the true model and the reconstructed data to evaluate the framework. (a) The original raw data of Condition 1 (least distinctive patterns, σ^2 = 4.5). All individual data are combined. Each individual data was simulated based on a normal distribution. (b) The original raw data of Condition 5 (most distinctive patterns, σ^2 = 0.5). (c) Visualized patterns of the reconstructed data of Condition 1 (least distinctive patterns, σ^2 = 4.5). Bin size = 0.05. (d) Visualized patterns of the reconstructed data of Condition 5 (most distinctive patterns, σ^2 = 0.5). (e) Visualized patterns of the true model based on a uniform distribution.

Pattern 2: Opposite opening

The overall results showed similar trends for both patterns of mean shift and opposite openings. As shown in Table 5.8, Figure 5.13 and Figure 5.14, the accuracy and R^2 increased as the distinctiveness of patterns increased (less noise in data). Remember that the accuracy or R^2 of two methods should be compared with the same type of data. The results of binned data and raw data are not comparable. Either by accuracy or R^2, the performance of the aggregated individual SVMs was on par with the single SVM built with the whole data. Sample patterns are visualized in Figure 5.15. Regardless of fuzziness of

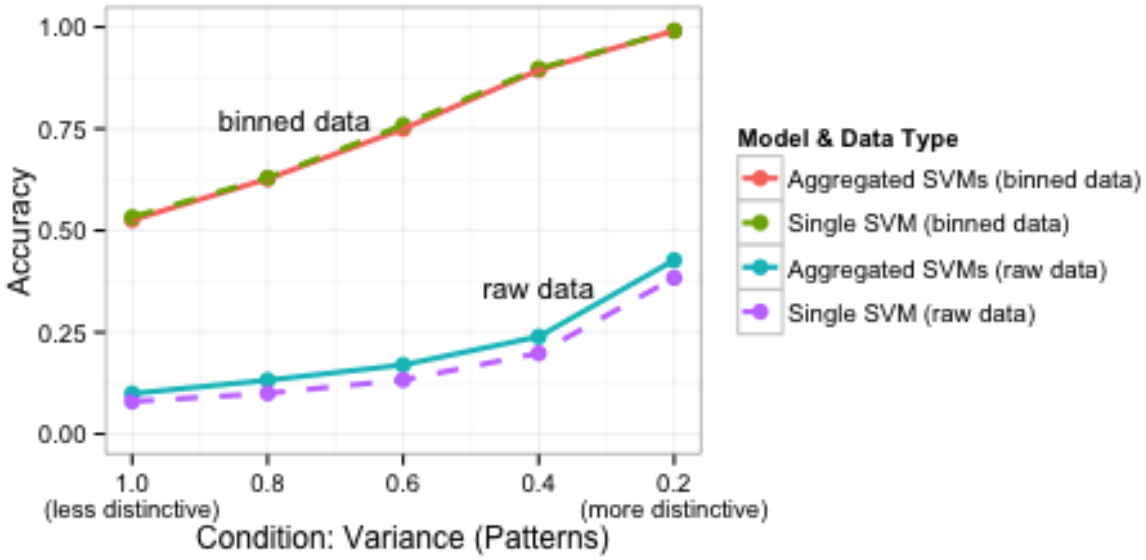

Figure 5.10: Pattern recovery accuracy by conditions for the patterns of mean shift. Bin size = 0.05, error size = 0.10.

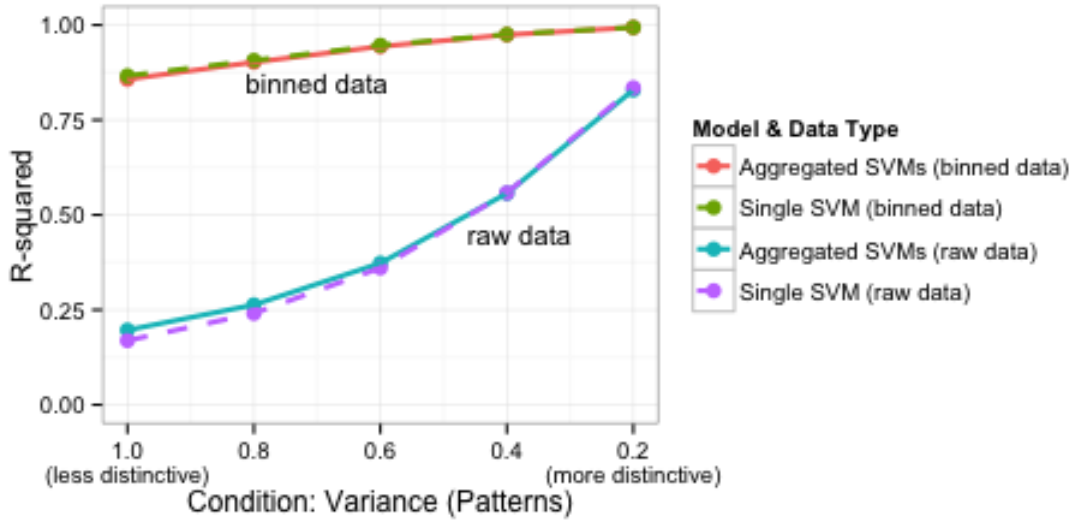

Figure 5.11: R^2 by conditions for the patterns of mean shift. Bin size = 0.05.

the raw data, the patterns represent the true model very well.

5.2.3 Discussion

First of all and most importantly, across all tasks, patterns, conditions, and data types, the aggregated individual models show practically equal performance as the single

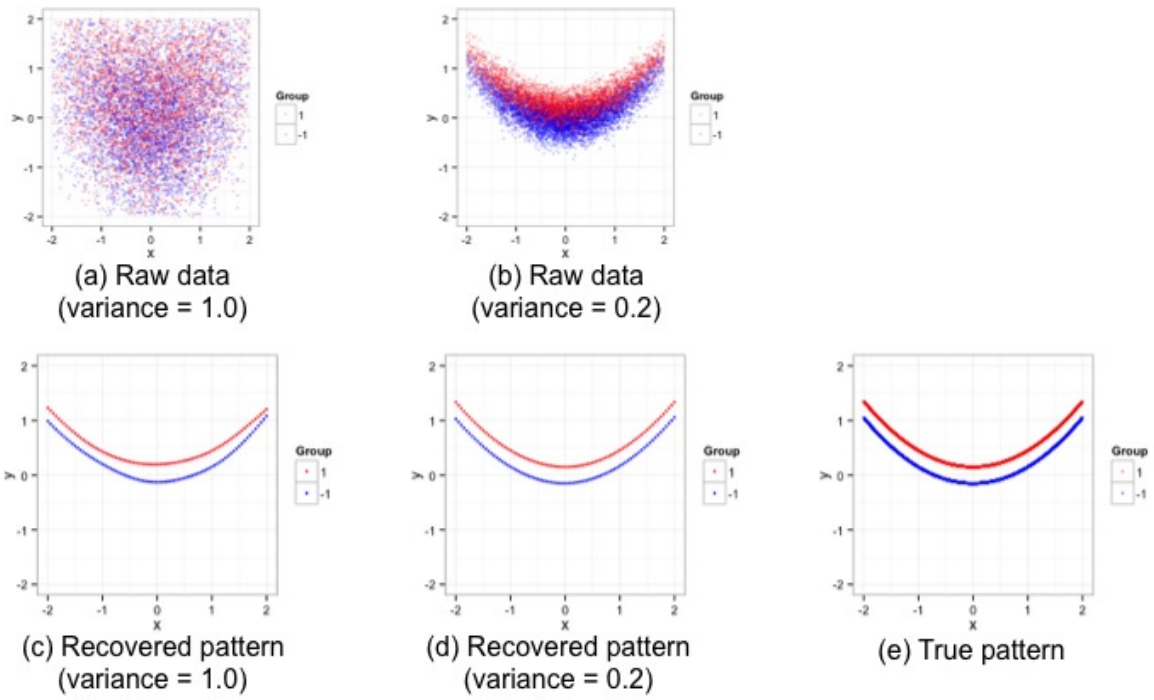

Figure 5.12: Sample visualization of the original raw data, the reconstructed data by SVMs within the localized analysis framework, and the true underlying patterns. Note that the original raw data and the reconstructed data do not necessarily look similar due to the noise in the raw data. We seek resemblance of patterns between the true model and the reconstructed data to evaluate the framework. (a) The original raw data of Condition 1 (least distinctive patterns, σ^2 = 1.0). All individual data are combined. Each individual data was simulated based on a normal distribution. (b) The original raw data of Condition 5 (most distinctive patterns, σ^2 = 0.2). (c) Visualized patterns of the reconstructed data of Condition 1 (least distinctive patterns, σ^2 = 1.0). Bin size = 0.05. (d) Visualized patterns of the reconstructed data of Condition 5 (most distinctive patterns, σ^2 = 0.2). (e) Visualized patterns of the true model.

model built with the whole data. With more distinctive patterns (less noise in data), as expected, models recovered patterns in data more accurately. Although R^2 results for binned data show ceiling effects, the proposed accuracy measure is able to show a better picture of the results.

Note that the better results with binned data than with raw data does not indicate the models perform better. The models are identical across binned data and raw data. The valid comparison is between the aggregated individual models and the single model within the same type of data. The higher accuracy when binned data are used is likely due to the

Table 5.8: Pattern recovery accuracy and R^2 for the pattern of opposite openings. The aggregated results from multiple individual models and the single result from a model built with the whole data are compared. Bin size = 0.05 and error size = 0.10.

Data type	Condition		Accuracy		R^2	
	Effects	σ^2	SVMs	Single SVM	SVMs	Single SVM
Binned data	Minimal	1.0	49.80%	53.89%	88.94%	92.32%
		0.8	60.34%	63.10%	93.29%	95.00%
		0.6	73.27%	75.65%	96.24%	97.11%
		0.4	87.21%	90.06%	98.33%	98.72%
	Large	0.2	98.06%	99.13%	99.58%	99.67%
Raw data	Minimal	1.0	10.24%	7.98%	23.52%	21.20%
		0.8	14.07%	10.02%	31.62%	29.67%
		0.6	18.68%	13.31%	43.85%	42.82%
		0.4	26.15%	19.83%	62.63%	62.85%
	Large	0.2	45.17%	38.33%	86.78%	87.13%

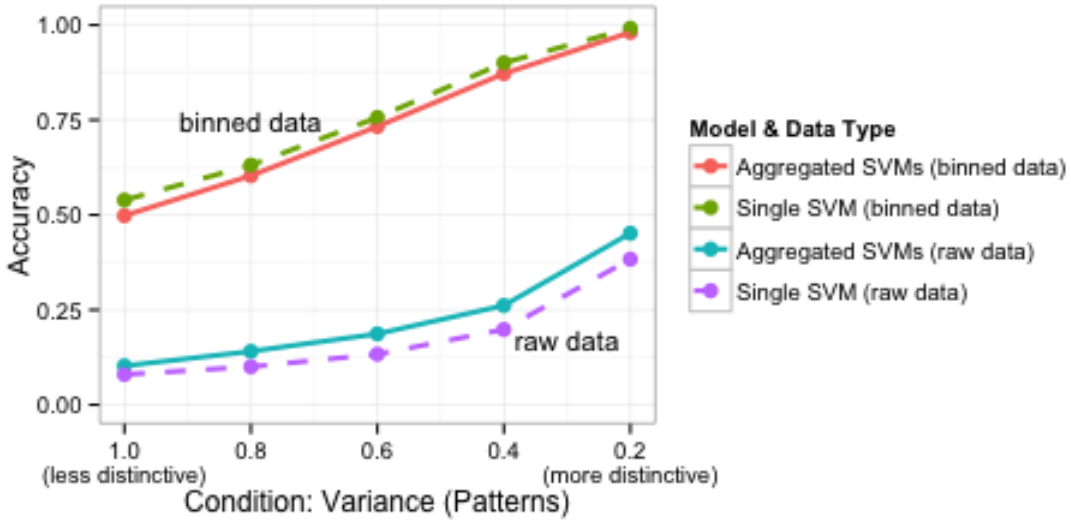

Figure 5.13: Pattern recovery accuracy by conditions for the patterns of opposite openings. Bin size = 0.05, error size = 0.10.

fact that the data compared with has less noise and is more close to the true underlying patterns, which the model-generated patterns resemble.

In fact, the underlying patterns in raw data may not be detectable by human eye (Figure 5.9-a, 5.12-a, and 5.15-a). SVMs successfully detect the true patterns and reconstruct them (Figure 5.9-c, 5.12-c, and 5.15-c). The study results suggest that we could use

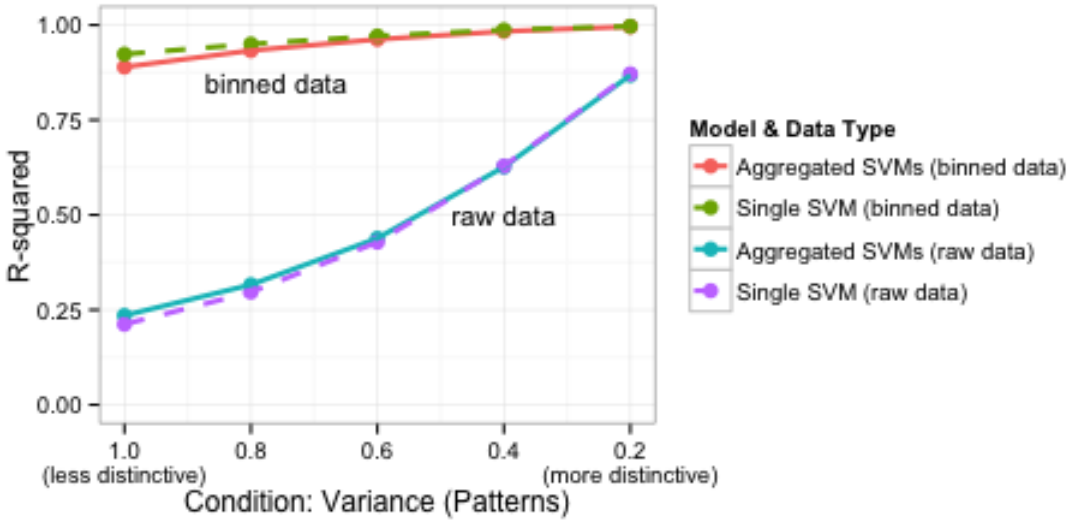

Figure 5.14: R^2 by conditions for the patterns of opposite openings. Bin size = 0.05.

predictive models to visualize seemingly invisible underlying patterns in data.

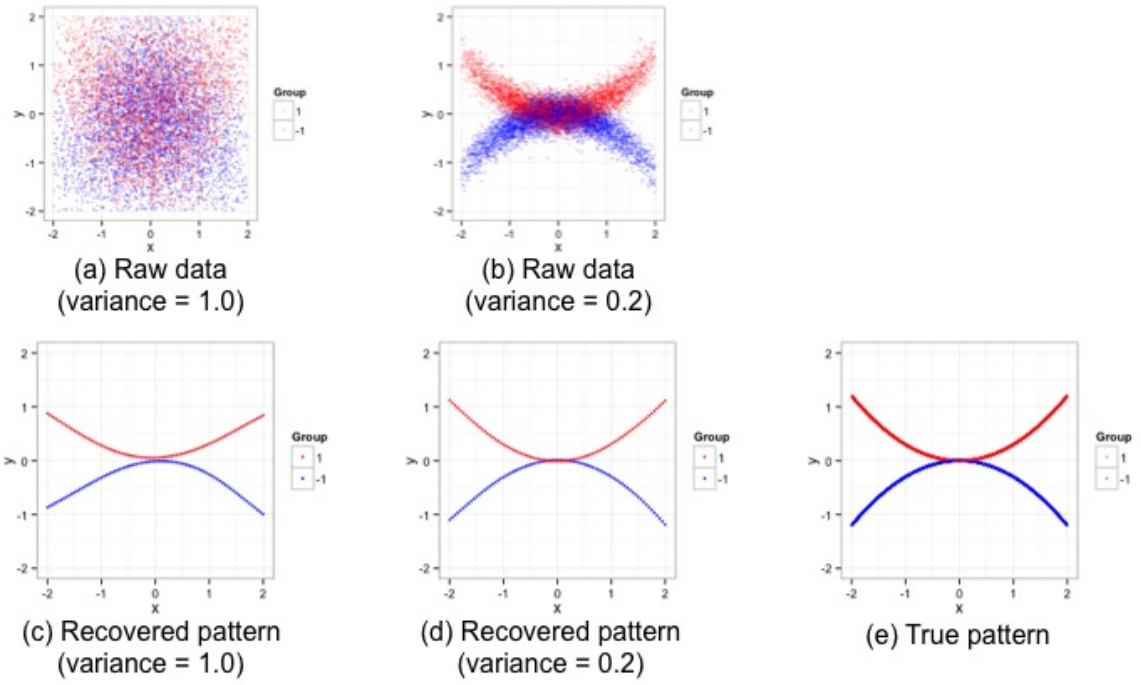

Figure 5.15: Sample visualization of the original raw data, the reconstructed data by SVMs within the localized analysis framework, and the true underlying patterns. (a) The original raw data of Condition 1 (least distinctive patterns, σ^2 = 1.0). All individual data are combined. Each individual data was simulated based on a normal distribution. (b) The original raw data of Condition 5 (most distinctive patterns, σ^2 = 0.2). (c) Visualized patterns of the reconstructed data of Condition 1 (least distinctive patterns, σ^2 = 1.0). Bin size = 0.05. (d) Visualized patterns of the reconstructed data of Condition 5 (most distinctive patterns, σ^2 = 0.2). (e) Visualized patterns of the true model.

6. General Discussion

In the localized analysis framework using smartphones, researchers do not directly collect raw data. Instead, software on each individual's smartphone collects the raw data, and only the processed results are sent to the researchers. Researchers then analyze the aggregated individual results and interpret them. The simulation study results suggest that the proposed localized analysis framework can perform as well as the conventional data collection and analysis paradigm. The pattern recovery results show that the aggregation of multiple individual models recover the underlying true patterns as well as the single model built with the whole data. In addition, this study found that binning can filter noise out in original data. As a result, the binned data resemble the underlying true model more, simply by binning. The noise filtering effects of binning is greater with a larger bin size.

6.0.1 Limitations of the study

Most importantly, the degree of privacy preservation was not measured in this study. To serve the purpose of privacy protection, it is beneficial to quantify how well the proposed method protects raw data and how much information the method can transfer. Differential privacy, for example, suggest adjusting the degree of privacy by using ϵ (ϵ-differentially private; Dwork et al., 2006). Although the current study showed how accurately the underlying data patterns can be transferred in comparison with the conventional data collection and analysis, how well privacy can be preserved is still in question. Because privacy and utility in data are likely to be a trade-off, it is especially important to provide measures for privacy preservation and information transfer. It will also allow us to evaluate different privacy protection approaches. The degree of privacy preservation and information loss should be considered in the future research.

The study was limited to visually recovering underlying data patterns using SVMs. The employed method using SVMs is only one of many methods possible in the proposed localized analysis framework. SVMs might not be the best method to recover patterns. Or, data visualization might not be the best way to use the proposed framework. As an extension of this study, for example, we could cluster reconstructed individual patterns and look for distinctive demographic characteristics between the clusters. Or we could use completely different analysis algorithms at an individual level. Further studies exploring different methods and approaches are needed to find effective ways to employ the localized analysis framework.

In this study, the study conditions had 100 observations per person, 100 participants, and only three features. This would not be typical when we use mobile devices for research. With mobile devices, it is much easier to collect thousands of participants. Mobile sensing data could add up to thousands or more per day because they can log activities every second or less. Although the main idea and principle will hold at a larger scale, the small scale of this study is one of the limitations.

In the proposed data collection and analysis framework, time-invariant data such as demographic information (e.g., gender and age) cannot be analyzed at an individual level as it has no variability over time. The host server should collect time-invariant data along with individual models if one wishes to include them in analysis at the aggregated level. In this study I did not examine time-invariant variables, which can be called higher-level variables because they are included at aggregated analysis after individual analysis. Higher-level variables, such as demographic information, are important factors in the social and behavioral sciences and in the future research we should investigate how to incorporate them in this framework.

Temporal and spatial aspects of mobile data were not taken into account in this study. The study data were generated based on an assumption that each observation is independent of each other. In many cases, however, mobile data are likely to be time-

dependent as well as space-dependent. Although studies have shown that SVM performs consistently with non-i.i.d. data (Steinwart, Hush, & Scovel, 2009), further investigation will be beneficial.

6.0.2 Implications of the study

The significance of this study is twofold. Although the main purpose of this study is to demonstrate the proposed framework, the study methods and results contribute other aspects too. The method developed in this study and the results can be applied to analytical methods other than SVMs. I suggested a new method in which predictive models reconstruct data patterns and studied analysis specification systematically for this approach. The study results show how the analysis results can be influenced (or not) by analysis specification, that is, bin size and class decision from multiple classifiers. The majority voting method outperformed the average decision values in class decision for classification tasks. Binning functioned as a noise filter and the binned data itself resembled the true patterns more than the raw data did. Furthermore, I suggested a new approach for accuracy measurement for regression that permits a certain range of errors. By controlling the precision of accuracy, the results could avoid ceiling effects shown in R^2. These are all applicable to different analysis methods and contexts other than SVMs, regardless of the proposed framework.

In this study, I demonstrated how to find underlying patterns of the whole sample from individual results. This is a form of meta-analysis. We could try applying the approach in this study to another context such as meta-analysis of multiple results. For example, in biomedical research, it is common to have a very small sample size per specific condition (e.g., Zou et al., 2005). Instead, the results have hundreds of results (models) in total. We could apply this approach to such cases in order to look for an overall pattern from multiple sub-results.

Finally, I suggested a unique approach to use machine learning in social science

contexts. A black-box approach of machine learning could assist theory building in the social sciences. Despite the high performance in prediction, the social sciences, where the theory-driven approach is highly endorsed, show very slow movement toward adopting machine learning, whose approach is highly data-driven. In other domains such as biomedicine, where traditional statistical analyses have also been widely used, machine learning has been adopted and gained popularity in recent years. In this study, I attempted to show a unique approach to using machine learning to build hypotheses by visualizing data patterns.

Data-driven approaches can assist theory-building in many ways. In the humanities, it is a new trend for researchers to use text mining to support their claims (Sculley & Pasanek, 2008). The same can be done in the social and behavioral sciences. The evidence in existing massive data, such as online blogs, can be utilized by machine learning methods (e.g., Pang & Lee, 2008). Machine learning can be very useful to utilize pre-existing data in natural settings without soliciting people's response.

6.0.3 Future consideration

Temporal and spatial correlations in mobile data are not considered in this study. For example, emotions reflected in text messages would not be independent from the previously sent messages. Someone who is upset might send multiple messages and express such emotion to someone else. During texting, the emotion could evolve based on the interaction with the other person. As another example, places where a person visits are not independent either. A person visits places from one to another because geographically they're connected. Furthermore, temporal characteristics and spatial characteristics can be intertwined. These characteristics require additional attention in the future.

Missingness by non-activity is an interesting topic to consider. Data from different mobile features do not align well with each other because different activities do not necessarily occur together. For example, moving between location and texting behavior may

or may not occur at the same time. Also, some mobile sensors create data every moment while some record only when an activity occurs. If we try to analyze such data altogether by aligning them by timestamps, the data will have a large portion of missingness. However, they are not missing data in the conventional sense (Little & Rubin, 2002; Rubin, 1976). In the conventional way of data collection, missingness matters because respondents may intentionally choose not to answer certain questions. In mobile data, the nature of missingness is completely different; it rather indicates non-activity. How to analyze or collapse such misaligned multiple data would be an important consideration when using mobile data.

Future research is needed into how to incorporate higher-level variables (such as demographic information) in the localized analysis framework. For example, I could have plotted recovered patterns by demographic categories in this study. Or, I could have clustered individual patterns and investigated demographic characteristics of each clusters. The method for using higher level variables will be something to consider when planning studies using the localized analysis framework.

Although this is not limited to the proposed framework, we need to investigate how to adjust the representativeness of mobile data. Even though smartphones have proliferated widely throughout our daily lives, their usage varies across demographics, such as age, gender, race, and socioeconomic status (Askitas, Zimmermann, Zagheni, & Weber, 2015). We need to be cautious about drawing conclusions based on what we observe using mobile devices. By accounting for usage patterns across demographics, we will be able to better understand the results from advanced technologies.

Finally, as mentioned in the study limitations, a measure of privacy preservation and information transfer should be considered in the future. To what degree is the information altered (lost) by introducing privacy protection mechanisms in data collection and analysis process? To what degree can it preserve privacy by altering (losing) the information? Such metric will allow us to develop and evaluate methodologies in privacy preserving

data collection and analysis. A promising approach could be integrating differential privacy (Dwork et al., 2006) in the localized analysis framework. For example, we could introduce randomness (defined by ϵ) in a localized analysis on each device. This remains an important topic to investigate in further research.

6.0.4 Significance of the localized analysis framework

Traditional data analysis assumes that analysts can use data that are collected from the sample at anytime for data analysis with no limitation. If the access to the original data is limited, we need new analysis methods to find patterns and meanings without the full original data. This is the first attempt to analyze each individual's data by itself, and examine the aggregated results to find patterns at the whole sample level.

The proposed research has significant benefits in research methodology. First of all, the localized analysis framework allows researchers access to sensitive information while protecting participants' privacy. Because app users are ensured that no original data will be transferred to someone else, we will be able to recruit wider participation from the general public.

This is especially beneficial in health-related research because health information is highly sensitive and personal. For example, drug use or sexual behavior are sensitive topics that have shown low response rates and high attrition. Participants will be more likely to respond to sensitive questions in a candid way if they are assured that no one will see their original responses, and that any data collected from those responses will remain anonymous.

Secondly, the localized analysis framework opens a new way of treating data and conducting research. Although data collection is much easier than before, public concerns about data privacy and security have been greatly heightened through major data breaches in recent years. Conventionally, researchers have to collect data and place it at a central place in order to extract information from the data using computing resources. We no

longer need to confine ourselves to the traditional paradigm. Data can be collected and analyzed using distributed computing resources: mobile phones.

Researchers are often not interested in each individual's data, ultimately. Rather, we are interested in aggregated data and emerging patterns across many individuals. If we can use information about the whole sample level without looking at the original responses of individuals, then we do not need to obtain individual raw data. In that way, not only do we protect people's privacy, but we have better security from data breaches.

Finally, the localized analysis framework can be applied to a broad range of practical applications for broader populations. For example, researchers (and users) can monitor physical activity levels using mobile sensors for general health monitoring purposes. Or we could reach a more specific population, such as those with depression, and use both their interpersonal data and physical activities for intervention. People are more likely to readily agree to participate in studies or to adopt healthcare apps for their daily lives if we assure that no one accesses personal raw data on their phones.

The proposed framework provides practical benefits. It can utilize private information while protecting people's privacy. Does it mean it completely protects user privacy from any researchers? Of course not. A researcher with a malicious intention could use a pencil and paper and ask for social security numbers from participants. Likewise, one may try to obtain highly private data from a personal device (which is possible with any currently available smartphone app). It is not possible to perfectly prevent it. Absolute security does not exist. However, this proposed framework will help researchers with good intentions avoid accidental invasion of participants' privacy, if the method is employed properly.

As discussed earlier, the choice of the processed results to send to the research host is important. Some results are more or less revealing of raw data. One should decide how much to process and what results to obtain when studies are designed. For example, running text analysis and obtaining the frequency of each word used would not suffice as

privacy protection. However, if one intends to evaluate suicide tendency, we could analyze text messages and evaluate the frequency of warning words defined by the researcher. Then the researcher could obtain suicide tendency scores based on text messages without looking at participants' actual text messages.

6.0.5 Conclusion

Mobile phones have become an intrinsic part of our daily lives. Any data collected from them is especially reflective of our day-to-day reality. By using smartphone apps, researchers can now observe behavioral patterns in mobile data, all without soliciting any responses or relying on self-reports. Such behavioral data includes the frequency and length of phone calls, the contents of text messages, and users' physical locations. However, this immediately brings up privacy issues. No one wants other people to harvest their mobile phone's private data. To utilize mobile data without invading privacy, I have proposed a localized analysis framework using smartphones. In this approach, researchers do not directly collect raw data. Instead, software on each individual's smartphone accesses the raw data locally stored in the device, and only the processed results are sent to the researchers. Researchers then analyze the aggregated individual results and find meanings at the whole sample level.

The current study aimed to demonstrate utilities of the proposed localized analysis framework using smartphones with an emphasis on exploratory analysis using machine learning. While there are many possible methods to employ within this framework, in this study I adopted support vector machines (SVMs) for localized analysis at an individual level and visually inspected the aggregated patterns at the whole sample level. The localized analysis framework opens a new way of treating data to protect user privacy.

References

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods*, *7*(1), 101–114.

Aoki, S., & Sezaki, K. (2014). Privacy-preserving community sensing for medical research with duplicated perturbation. In *Communications (icc), 2014 ieee international conference on* (pp. 4252–4257).

Askitas, N., Zimmermann, K. F., Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, *36*(1), 13–25.

Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*, 1–13.

Boker, S. M., Brick, T. R., Pritikin, J. N., Wang, Y., Oertzen, T. v., Brown, D., ... others (2015). Maintained individual data distributed likelihood estimation (middle). *Multivariate behavioral research*, *50*, 706–720.

Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., ... Ratti, C. (2011). The connected states of america: Quantifying social radii of influence. In *Privacy, security, risk and trust (passat) and 2011 ieee third inernational conference on social computing (socialcom), 2011 ieee third international conference on* (pp. 223–230).

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

Chaudhuri, K., & Monteleoni, C. (2009). Privacy-preserving logistic regression. In *Advances in neural information processing systems* (pp. 289–296).

Che, D., Liu, Q., Rasheed, K., & Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. In H. R. Arabnia & Q.-N. Tran (Eds.), *Software tools and algorithms for biological systems* (pp. 191–199). New York, NY: Springer.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, *40*(1), 169–193.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (uct) to estimate base r. *Personnel Psychology*, *47*(4), 817.

Decoste, D., & Schölkopf, B. (2002). Training invariant support vector machines. *Machine learning*, *46*(1-3), 161–190.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., ... others (2011). Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. *PloS one*, *6*(9), e24974.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography* (pp. 265–284). Springer.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407.

Erricson. (2015, June). *Erricson mobility report* (Tech. Rep.).

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on mobile systems, applications and services* (pp. 57–70).

Gerber, M. (2015). *A cross-platform system for mobile crowd-sensing.* Retrieved from `https://github.com/predictive-technology-laboratory/sensus/wiki` (Online; accessed 7-November-2015)

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520–539.

Han, B., & Davis, L. S. (2012). Density-based multifeature background subtraction with support vector machine. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(5), 1017–1023.

Hua, S., & Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, *17*(8), 721–728.

Huang, C.-F. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, *12*(2), 807–818.

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. In R. Matthiesen (Ed.), *Bioinformatics methods in clinical research* (Vol. 593, pp. 25–48). New York, NY: Humana Press.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Karatzoglou, A., Meyer, D., & Hornik, K. (2005). Support vector machines in r. *Journal of Statistical Software*, *15*, 1-28.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932–932.

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry*, *56*, 171–182.

Lee, C.-S., Sedory, S. A., & Singh, S. (2013). Estimating at least seven measures of qualitative variables from a single sample using randomized response technique. *Statistics & Probability Letters*, *83*(1), 399–409.

Lei, J. (2011). Differentially private m-estimators. In *Advances in neural information processing systems* (pp. 361–369).

Li, C.-H., Kuo, B.-C., Lin, C.-T., & Huang, C.-S. (2012). A spatial–contextual support vector machine for remotely sensed image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, *50*(3), 784–799.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Lu, H., Pan, W., Lane, N. D., Choudhury, T., & Campbell, A. T. (2009). Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on mobile systems, applications, and services* (pp. 165–178).

MacKerron, G., & Mourato, S. (2013). Happiness is greater in natural environments. *Global Environmental Change*, *23*(5), 992–1000.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*, 221–237.

Moghaddam, B., & Yang, M.-H. (2002). Learning gender with support faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(5), 707–711.

Mohammed, A. A., Minhas, R., Jonathan Wu, Q., & Sid-Ahmed, M. A. (2011). Human face recognition based on multidimensional pca and extreme learning machine. *Pattern Recognition*, *44*, 2588–2597.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259.

Ostapczuk, M., Moshagen, M., Zhao, Z., & Musch, J. (2009). Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. *Journal of Educational and Behavioral Statistics*, *34*(2), 267–287.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. *Computers & Geosciences*, *51*, 350–365.

Raento, M., Oulasvirta, A., Petit, R., & Toivonen, H. (2005). Contextphone: A prototyping platform for context-aware mobile applications. *Pervasive Computing, IEEE*, *4*(2), 51–59.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40–45.

Rosten, E., Porter, R., & Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*, 105–119.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Saeys, Y., Wehenkel, L., Geurts, P., et al. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*, *28*, 1766–1774.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, *13*(7), 1443–1471.

Sculley, D., & Pasanek, B. M. (2008). Meaning and mining: the impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, *23*(4), 409–424.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*, 6–13.

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

Statnikov, A., Hardin, D., Guyon, I., & Aliferis, C. F. (2015). *A gentle introduction to support vector machines in biomedicine.* Retrieved from `http://www.med.nyu`

.edu/chibi/sites/default/files/chibi/Final.pdf (Online; accessed 7-November-2015)

Steinwart, I., Hush, D., & Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, *100*(1), 175–194.

Tax, D. M., & Duin, R. P. (1999). Support vector domain description. *Pattern recognition letters*, *20*(11), 1191–1199.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*, 859.

Upstill-Goddard, R., Eccles, D., Fliege, J., & Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in Bioinformatics*, *14*, 251–260.

Vapnik, V. N. (1998). *Statistical learning theory* (Vol. 1). Wiley New York.

Vapnik, V. N. (1999). *The nature of statistical learning theory*. Springer Science & Business Media.

Wang, X., & Pardalos, P. M. (2015). A survey of support vector machines with uncertainties. *Annals of Data Science*, *1*(3-4), 293–309.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Wikipedia. (2015). *Resolution illustration.* Retrieved from https://commons.wikimedia.org/wiki/File:Resolution_illustration.png (Online; accessed 7-November-2015)

Yang, X.-S., Deb, S., & Fong, S. (2011). Accelerated particle swarm optimization and support vector machine for business optimization and applications. In *Networked digital technologies* (pp. 53–66). Springer.

Zou, K. H., Greve, D. N., Wang, M., Pieper, S. D., Warfield, S. K., White, N. S., ... others (2005). Reproducibility of functional mr imaging: Preliminary results of prospective multi-institutional study performed by biomedical informatics research network 1. *Radiology*, *237*(3), 781–789.