The Benefits of Personalized Data Mining Approaches to Human Activity Recognition with Smartphone Sensor Data

By

Jeffrey William Lockhart

B.S., Fordham University 2013

MASTERS THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE DEPARTMENT OF COMPUTER AND INFORMATION
SCIENCES AT FORDHAM UNVIERSITY
NEW YORK

MAY, 2014

Acknowledgments

I would like to express my sincerest gratitude to Dr. Gary Weiss, my thesis advisor. His guidance and support on this research and all of my scholarly endeavors over the last five years have been invaluable and formative. This work was also made possible by the community and resources he fosters in the Wireless Sensor Data Mining Lab.

I want to express my thanks also to Dr. Damian Lyons and Dr. Xiaoxu Han, for serving on my committee and for their valued feedback on this thesis and my work more broadly.

Table of Contents

1 Introduction

Activity recognition (AR) on mobile devices is a rapidly growing field. The ability of a device to recognize its user's activity is important because it enables context-aware applications and behavior. Activity recognition also makes it possible to develop mobile health applications that track a user's activities, such as our Actitracker app [1], which can help address the many health concerns that arise due to inactivity, including childhood obesity [25]. The work described in this paper relies on Android smartphones, but the tri-axial accelerometer present in these smartphones—illustrated in Figure 1—is very similar to those found in other smartphones and mobile devices.

In this paper we employ a straightforward approach for implementing AR. We collect accelerometer data from users as they walk, jog, climb stairs, sit, stand, and lie down, and then aggregate each 10 seconds of data into a single labeled example. We then induce an AR model by applying common classification algorithms to the generated training examples. The work in this paper includes data from our set of 59 test subjects [46], as well as data from 414 subjects in the HASC 2010 and 2011 data sets [22].

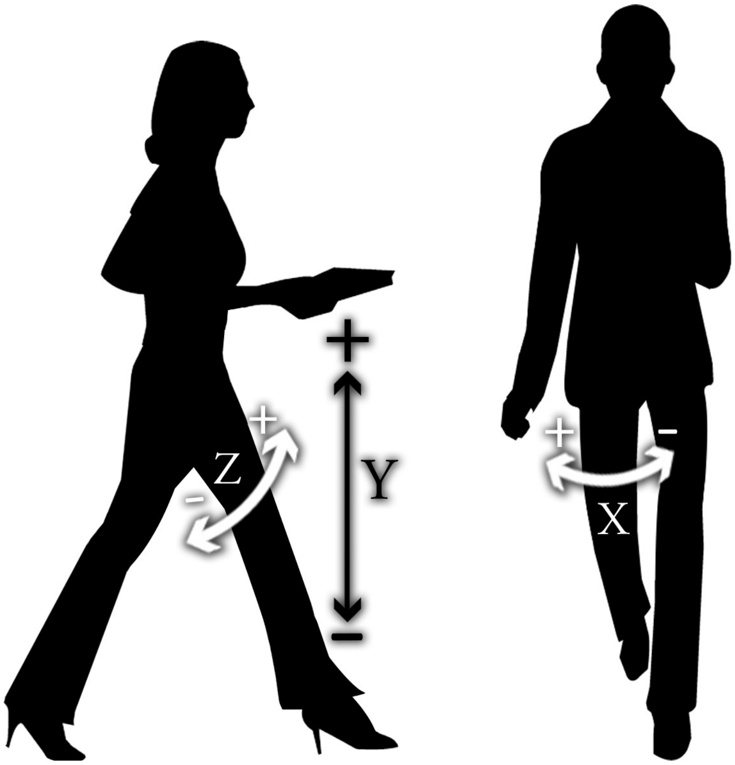


Figure 1: Axes and Placement of Smartphone Accelerometers

1.1 Background

There has been much prior work on AR [7, 26, 27, 45], and a smaller but growing body of work on smartphone-based AR [8, 23, 29, 51]. While some work has used personal models [36, 51], which are built exclusively using labeled training data from the intended user, most work has focused on universal models [8, 16, 30, 38], which are built using data from a panel of users who are not the intended users of the model. Although newer work [30] has aimed at tailoring universal models to individuals, little work has compared personal and universal models on a reasonably-sized population, and no work has carefully analyzed the relative performance of these types of models. This paper provides a thorough analysis of the relative performance of these models. The extensive review of others' findings in Chapter 1.2 demonstrates that existing publications follow the trend we identify as well. We conclude that personal AR models are extraordinarily effective and vastly superior to universal models, even when built from a very small amount of personal training data.

Personalized AR models are quite feasible for smartphone-based AR, since a smartphone typically is used by a single user and because little training data is required. Thus, we advocate for the development of personal AR models. We have in fact incorporated the ability to generate such models into our publicly available AR app, Actitracker [1]. This app allows a user to quickly perform "self-training" and then replaces the default universal model with an automatically generated personal one.

1.2 Related Work

There has been prior work related to personal, universal, and hybrid models of AR, although in virtually all cases the work has not had these topics as their primary focus. Several AR studies only analyze universal models, with limited results [8, 16, 38]. Other AR systems using smartphones have achieved relatively higher accuracy, but used only personal models [8, 36, 51]. One paper described personal models that could be incrementally trained to adapt to changes in a user's gait [2]. However, their results were similar to those of our universal models, indicating that feature selection and choice of algorithms is still an important part of AR.

There has been relatively little comparative analysis of the different types of AR models. Three studies did compare personal and universal models, but both employed five accelerometers---thus any conclusions would not necessarily apply to a smartphone--based system. The first of these studies concluded that universal models always outperform personal ones, due to the additional training data; it further showed that when the universal and personal training set sizes are equalized, personal models only slightly outperform universal ones [7]. Our results clearly and dramatically contradict this result (but for smartphone-based systems). In the second study the personal models outperform the universal models but there is virtually no analysis or discussion of this, as it is not the paper's focus [45]. In the third study, personal models again outperform universal ones, but the only activities evaluated are three different walking speeds, making it difficult to generalize their results to AR with more diverse activities [10].

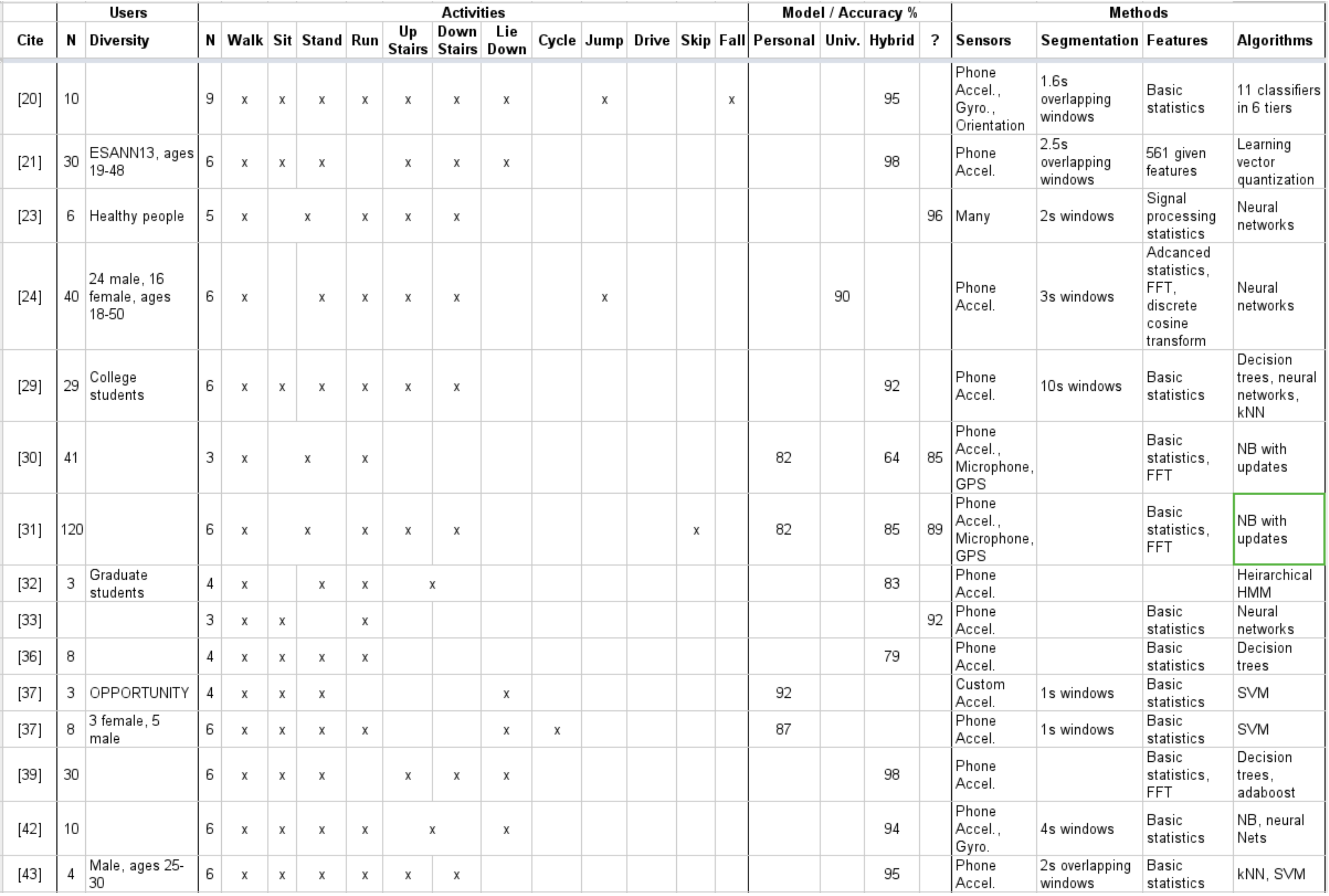
One paper already discussed [30] recognizes the poor performance of some universal models and develops methods for improving their accuracy by selectively training on similar users. However, our results show that beyond a certain point (i.e., the point they [30] reached and the point we start from with our set of similar users), user similarity is not able to compensate for the differences in gait. Further, our analysis of the learning curves for various model types shows that very small amounts of personal data dramatically outperform these best-case universal models, and that no amount of data for universal models will perform competitively with personal models. Thus our paper is the most comprehensive study on the impact of model-type on AR---especially as it relates to single accelerometer, smartphone-based systems. Furthermore, we additionally evaluate hybrid models and include many more users in our study than prior studies, as well as more activities which are more challenging to distinguish, leading to more reliable, and general, results.

1.2.1 Data Size and Diversity

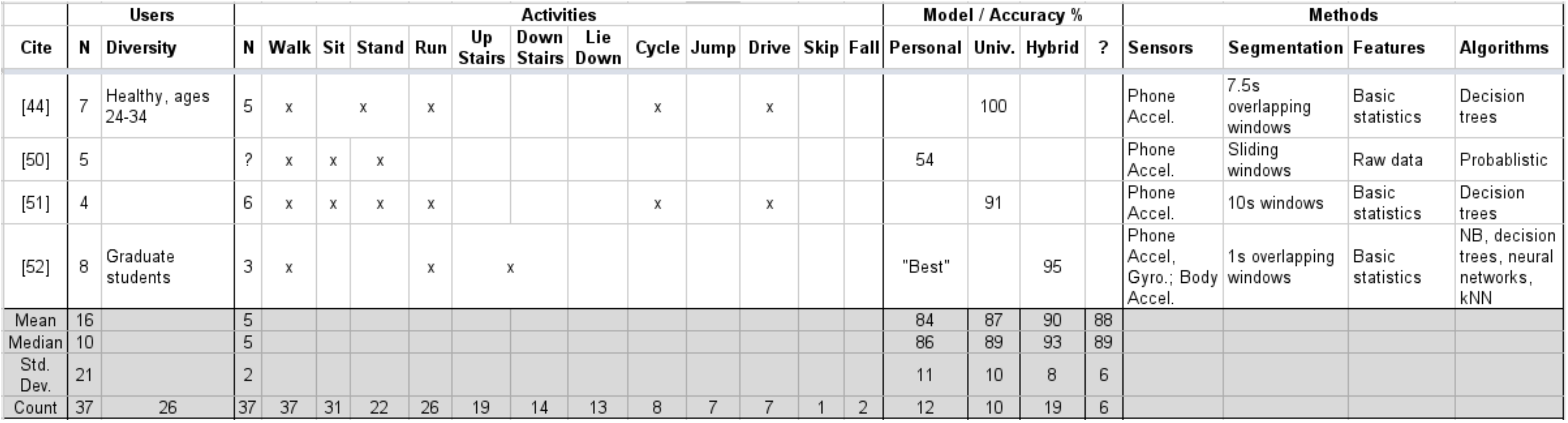
Many studies use very limited datasets, often with fewer than 5 users [2, 3, 12, 15, 16, 18, 32, 33, 37, 43, 51], or 10 users [11, 13, 14, 17, 20, 23, 36, 37, 42, 44, 52]. Compounding the issue, the most widely used AR datasets, COSAR and OPPORTUNITY, have data from only 4 and 12 users, respectively [40, 41]. Larger sets, such as HASC 2010 and 2011, contain simplified sets of activities and smaller amounts of data. This motivated us to release our AR dataset [46]. A detailed breakdown of data set size can been seen in the second column of Table 1.

Table 1 contains a summary of 38 related experiments, a few of which come from a single paper. Although most papers evaluated several methods on one data set, only the most successful method from each paper is included in Table 1. Where a paper had multiple distinct data sets, each data set was represented with its own row. Table 1 details the number and diversity of users in a data set, the activities represented, the model types and accuracies published, and information about the authors' methods (sensors used, data segmentation, feature transformation, and classification algorithms). Several papers were excluded because they lacked sufficient methodological detail, or because their experimental conditions varied dramatically from the setup described in this work (e.g. smart-homes, dozens of on-body sensors).

	Users		Activities													Model / Accuracy %				Methods			
Cite	N	Diversity	N	Walk	Sit	Stand	Run	Up Stairs	Down Stairs	Lie Down	Cycle	Jump	Drive	Skip	Fall	Personal	Univ.	Hybrid	?	Sensors	Segmentation	Features	Algorithms
[2]	3	OPPORTUNITY	3													86	82		80	Many	Windows	Raw data	Clustering
[3]	4	OPPORTUNITY	4	x	x	x				x							63			Many			kNN
[4]	30	Ages 19-48	6	x	x	x		x	x	x								89		Phone Accel., Belt	2.56s overlapping windows	Basic statistics, FFT	SVM
[5]	30	Ages 19-48	6	x	x	x		x	x	x								96		Phone Accel., Belt	Windows	561 advanced statistics	SVM
[6]	30	Ages 19-48	6	x	x	x		x	x	x							89			Phone Accel., Belt		Basic statistics, FFT	SVM
[8]	?		6	x	x	x		x	x						x	80				Many	Windows	Raw data	kNN
[10]	25	Students and staff	3	x												100	88			Phone Accel.	Data-dependent	Raw data, or basic statistics	kNN, SVM
[11]	10	Undergradate students	8	x	x	x	x	x		x	x		x					93		Phone Accel., Gyro.	Overlapping windows	Basic statistics	Decision trees, neural networks, K-star
[11]	10	Undergradate students	8	x	x	x	x	x		x	x		x					84		Phone Accel.	Overlapping windows	Basic statistics	Decision trees, neural networks, K-star
[12]	5	Ages 6-67	4	x	x		x						x					90		Phone Accel., GPS	1s overlapping windows	Basic statistics	Neural networks
[13]	10	7 male, 3 female, healthy, diverse age/weight/heigh	7	x			x	x	x		x	x						95		Phone Accel.	3s windows	Basic statistics	Genetic algorithms, EFM
[14]	10	7 male, 3 female, unsupervised data collection	7	x			x	x	x		x	x						95				Basic statistics	Clustering, fuzzy probability
[15]	4		4	x	x	x				x						86				Many			Incremental NB
[16]	3		4	x	x		x				x					83				Phone Accel.	Sliding window	Intensity	Neural netoworks
[17]	10	Graduate students	4	x	x		x						x					90		Phone Accel.	3s sliding windows	Basic statistics, FFT	GMM
[18]	3	Graduate students	5	x	x	x	x						x			90				Phone Accel., Gyro., GPS			Adaptive Bayes
[19]	11	9 male, 2 female	4	x	x		x					x					97			Custom Accel.	5s overlaping windows	Discrete cosine transfrm	SVM
[19]	11	9 male, 2 female	4	x	x		x					x					84			Custom Accel.	5s overlaping windows	Basic statistics	SVM



	Users		Activities													Model / Accuracy %				Methods			
Cite	N	Diversity	N	Walk	Sit	Stand	Run	Up Stairs	Down Stairs	Lie Down	Cycle	Jump	Drive	Skip	Fall	Personal	Univ.	Hybrid	?	Sensors	Segmentation	Features	Algorithms
[20]	10		9	x	x	x	x	x	x	x		x			x			95		Phone Accel., Gyro., Orientation	1.6s overlapping windows	Basic statistics	11 classifiers in 6 tiers
[21]	30	ESANN13, ages 19-48	6	x	x	x		x	x	x								98		Phone Accel.	2.5s overlapping windows	561 given features	Learning vector quantization
[23]	6	Healthy people	5	x	x		x	x	x										96	Many	2s windows	Signal processing statistics	Neural networks
[24]	40	24 male, 16 female, ages 18-50	6	x		x	x	x	x			x					90			Phone Accel.	3s windows	Adcanced statistics, FFT, discrete cosine transform	Neural networks
[29]	29	College students	6	x	x	x	x	x	x									92		Phone Accel.	10s windows	Basic statistics	Decision trees, neural networks, kNN
[30]	41		3	x	x		x									82		64	85	Phone Accel., Microphone, GPS		Basic statistics, FFT	NB with updates
[31]	120		6	x	x		x	x	x					x		82		85	89	Phone Accel., Microphone, GPS		Basic statistics, FFT	NB with updates
[32]	3	Graduate students	4	x		x	x	x										83		Phone Accel.			Heirarchical HMM
[33]			3	x	x		x												92	Phone Accel.		Basic statistics	Neural networks
[36]	8		4	x	x	x	x											79		Phone Accel.		Basic statistics	Decision trees
[37]	3	OPPORTUNITY	4	x	x	x				x						92				Custom Accel.	1s windows	Basic statistics	SVM
[37]	8	3 female, 5 male	6	x	x	x	x			x	x					87				Phone Accel.	1s windows	Basic statistics	SVM
[39]	30		6	x	x	x		x	x	x								98		Phone Accel.		Basic statistics, FFT	Decision trees, adaboost
[42]	10		6	x	x	x	x	x		x								94		Phone Accel., Gyro.	4s windows	Basic statistics	NB, neural Nets
[43]	4	Male, ages 25-30	6	x	x	x	x	x	x									95		Phone Accel.	2s overlapping windows	Basic statistics	kNN, SVM



	Users		Activities													Model / Accuracy %				Methods			
Cite	N	Diversity	N	Walk	Sit	Stand	Run	Up Stairs	Down Stairs	Lie Down	Cycle	Jump	Drive	Skip	Fall	Personal	Univ.	Hybrid	?	Sensors	Segmentation	Features	Algorithms
[44]	7	Healthy, ages 24-34	5	x		x	x				x		x				100			Phone Accel.	7.5s overlapping windows	Basic statistics	Decision trees
[50]	5		?	x	x	x										54				Phone Accel.	Sliding windows	Raw data	Probablistic
[51]	4		6	x	x	x	x				x		x				91			Phone Accel.	10s windows	Basic statistics	Decision trees
[52]	8	Graduate students	3	x			x	x								"Best"		95		Phone Accel, Gyro.; Body Accel.	1s overlapping windows	Basic statistics	NB, decision trees, neural networks, kNN
Mean	16		5													84	87	90	88				
Median	10		5													86	89	93	89				
Std. Dev.	21		2													11	10	8	6				
Count	37	26	37	37	31	22	26	19	14	13	8	7	7	1	2	12	10	19	6				

Table 1: Related Work

1.2.2 Test Set Membership

Generally speaking, when evaluating classification models it is best practice to test their accuracy on different data than was used initially to train the model (i.e. to have distinct, non-intersecting testing and training sets of data). This ensures that models are not “over fit” to one small set of data, and that their performance can generalize to new, unseen data. A key question in this research was whether implementing this method in activity recognition required ensuring not just that the training and testing sets contained distinct data elements, but whether they also must contain only data from distinct users. The hybrid experiment, where some data from a single user are present simultaneously in both sets, is most common in the literature, but its validity is largely untested.

The majority of smartphone-based systems only evaluate hybrid models [4, 5, 11, 12, 13, 14, 17, 20, 21, 29, 32, 36, 39, 42, 43, 49]. These authors often claim that their results are generalizable to new users, but as we will show, the results of evaluating hybrid models are better than we can expect for previously unseen users. Additionally, a number of other studies [9, 23] do not provide enough information in their methodology to determine the model type. We view this as problematic because of the dramatic difference in performance and applications between different model types. In one study of Parkinson's disease patients, the researchers conclude that models trained on healthy people perform more poorly on people with Parkinson's disease [49]. However, their methodology shows that they used hybrid models when evaluating their accuracy on healthy subjects and universal models when evaluating their accuracy on Parkinson's patients. Our work shows that the discrepancy in their results is likely due, at least in part, to their mixing of model types and not due to differences between Parkinson's patients and healthy people.

The average accuracy of hybrid models across all studies in Table 1 is 90%. Although their performance varies between papers, unusually high or low performance is often associated with external factors. For instance, by far the worst accuracy of a hybrid model was 64%, published in 2011 by Lane et al [30]. In this work, they evaluated just three activities (viz. still, walking, and running), which tend to be the easiest activities to predict and differentiate [29], and they have a comparatively large data set of 41 people so it is less likely that their results are heavily swayed by outliers. Yet even under these ideal circumstances, their methodology only achieves 82% accuracy under personal model conditions. This suggests that external factors, such as data quality in their proprietary set or their custom classification algorithm, are causing generally lower accuracy across all model types with this research team.

Conversely, one of the best hybrid models in Table 1 was produced by Reiss et al. [39]. However, the authors admit that their 98% accuracy is achieved only in the unnatural setting where the phone is secured firmly to the user's chest. This positioning, while dramatically reducing noise in the data, is wholly unrealistic for the kind of everyday users smartphone-based AR typically targets.

Personal models, which we will argue are the most accurate type of AR modeling, only average a surprisingly low 84% accuracy. We will explore some of the reasons personal models in related work may perform poorly, but ultimately this work is concerned more with the potential for each model type than with ineffective implementations of them.

The least accurate personal models were published by Yan et al. [50], and scored only a 54% accuracy. This work is difficult to evaluate, since it provides little information about its data set (even the list of activities is unspecified). It is clear, however, that data was gathered “in the wild,” a process likely to introduce noise and ambiguity in the labels used as ground truth during classification [50]. Additionally, the paper is not easily comparable to other AR research because it makes two novel changes to traditional methods. First, their method does not extract features or signals from the raw accelerometer data. Second, they develop a new probabilistic classification algorithm. Thus, this paper should not be used to judge the potential accuracy of other methods, which have demonstrated consistently superior performance.

Another poorly performing personal model was published by Gyorbiro et al. in 2008 [16], who were the only research team to publish an AR paper that used just a single feature (combined intensity of accelerations) as the input to their classifier. Subsequent research using features has always used multiple features, and nearly always outperformed this experiment. It is worth noting that the authors of this study [16] achieved this performance by including their training data in their test sets, a deviation from the standard of keeping entirely separate training and testing sets. As such, this study [16] is a strong indicator that intensity alone is insufficient for delineating activities, but it does not indicate a limited potential for methods considering a broader array of features.

1.2.3 Feature Selection

Nearly all research segments the accelerometer data into discrete windows for classification, since most classification algorithms do not understand raw time-series data [47]. Window lengths vary from 1 to 10 seconds. The large majority of smartphone AR research uses basic statistics (e.g. mean, standard deviation, binned distribution) derived from accelerometer data in each window as the input features for classification [4, 10-14, 19, 20, 29, 33, 36, 37, 42, 44, 51, 52]. Research using these features has been extremely successful with personal and hybrid model types, often achieving accuracies in the high 90's [29, 34]. universal models using these features have often performed generally in line with work using more advanced features [34]. The key advantage to using only basic statistics is that they are computationally lightweight to calculate [29]. This enables them to be generated on smartphones, which have limited CPU and battery resources. Features which are quick to calculate are also advantageous in real-time scenarios and in aggregate central-processing schemes, where computation time is a key concern.

A growing trend is to use more advanced signal processing techniques [4, 6, 17, 24, 30, 31, 39]. While the 'basic statistics' approach extracts features from the time domain, this approach uses Fourier transforms (almost always discrete, fast Fourier transforms, hereafter “FFT”) to transform the time-series data into frequency domain data. New features are then extracted from the FFT output by an additional processing step according to either domain expertise (e.g. knowledge about the rate of human gait), or to signal processing standards from other fields (e.g. audio analysis). Studies using FFT sometimes [39], but not always [30], perform well, and overall there is inconclusive evidence to suggest that these features improve activity recognition potential. Due to the substantially increased processing load of these techniques, it is best to avoid using them in AR applications until conclusive evidence of their usefulness is demonstrated.

Researchers have, of course, also varied from these norms. Several papers [2, 8, 50] have used the raw, time series accelerometer directly as input to classification algorithms (usually nearest neighbor). While this approach shows promise, it presents researchers with new challenges. First, the problem of segmenting and aligning the data becomes much more important, since misaligned segments will classify incorrectly. Second, when each 3-dimensional accelerometer reading (sampled at 20-100Hz) is used as a feature in nearest-neighbor algorithms, the computation time grows rapidly in comparison to statistical feature approaches, which typically only use a few dozen features. Other research teams have constructed features from more esoteric statistics [5, 21, 23, 24], or used data transformations other than FFT such as discrete cosine transforms [19, 24]. While each method shows promise, there is little consistency across these studies when compared with studies using more common methodologies. It is worth noting that these methods generally require more computational resources than basic statistics.

2 Experimental Setup

Algorithms from the open source WEKA suite of machine learning tools [47] were used to build and validate a variety of classifiers, in combination with custom scripts to automate the process and prepare our data. The experiments in this thesis were conducted offline on a multiprocessor lab server, although we have also constructed an online environment to gather this data from smartphones and apply our techniques in real-time [1].

2.1 Data Sets

We collected data by having 59 users carry an Android-based smartphone in their pocket while performing six everyday activities: *walking, jogging, sitting, standing, climbing up and down stairs*, and *lying down*. Our research team members directed the participants to perform the various activities and input the activity labels into our data collection app. The sensor data is stored on the phone and also transmitted to our server. For this study, we use a sampling rate of 20Hz. This sampling rate has been shown to reliably produce data even on old Android devices, and to be sufficiently fast to capture human activities [29]. We used fifteen different Android smartphone models in our study, and all of their accelerometers appeared to generate similar results.

Additionally, we analyzed the data from the HASC 2010 and 2011 data sets [22]. In order to apply the same methods to this data, we selected only the accelerometer data from pocket or waist located sensors and down-sampled the readings from 100Hz to 20Hz. The resultant subset included 414 people performing 4 activities: *standing, walking, running*, and *skipping*. The results for the two data sets are presented separately, since they represent different activities and are recorded with different hardware/software.

Table 2 shows the number and distribution of the transformed examples, per activity, for our data. Walking is the most common activity. The time spent jogging and stair climbing was limited because these activities are strenuous, and we limited the time spent on the static activities because we found they were easy to learn. By comparison, after transformation the HASC set contains 10,718 examples from 414 users. This is a different and smaller set of activities than in our data, and the class distribution is much more skewed.

	Total	**Walk**	**Jog**	**Stair**	**Skip**	**Sit**	**Stand**	**Lie**
WISDM	n = 9,291	3,397	1,948	1,549		1,143	689	565
	100%	36.6%	21.0%	16.7%		12.3%	7.4%	6.1%
HASC	n = 10,718	6,420	1,393		1,393		1,618	
	100%	59.9%	13.0%		13.0%		15.1%	

Table 2: Number and Distribution of Transformed Examples Per Activity

2.2 Features

Standard classification algorithms cannot directly handle time-series data, so we first transform the raw accelerometer data into examples. To accomplish this each example summarizes 10 seconds of data (this time is sufficient to capture several repetitions of periodic motions and was empirically shown to perform well). Given 20 samples per second and 3 axes, this yields 600 accelerometer values per example. We then use the following 6 basic features to generate 43 features from these raw values (the number of features generated for each feature-type is noted in brackets):

- Average [3]: Mean acceleration (per axis).
- Standard Deviation [3]: Standard deviation (per axis).
- Average Absolute Difference [3]: Mean absolute difference between the value of each of the 200 values in an example and their mean (per axis). $$AAD=\frac{\sum_{i=0}^{n}\left|x_i-mean(x)\right|}{n}$$
- Average Resultant Acceleration [1]: Mean of the square roots of the sum of the values of each axis squared, over the example: $$ARA=\frac{\sum_{i=0}^{n}\sqrt{x_i^2+y_i^2+z_i^2}}{n}$$
- Binned Distribution [30]: The fraction of the 200 values that fall within each of 10 equally spaced bins, spanning the range of possible values (per axis).
- Frequency [3]: the frequency of the periodic wave associated with repetitive activities (per axis).

The frequency feature uses a heuristic method to identify all of the clearly distinct peaks in the wave and then calculates the average time between successive peaks. For samples where at least three peaks cannot be found, a special null value is used. It is presented as Algorithm 1.

```
identify_peaks(list L){
     //return elements where L[i]>L[i-1] && L[i]>L[i+1]
}
peaks = identify_peaks(L);

factor = 0.9;
while(factor > 0 && max_peaks.size() < 3){
     cutoff = peaks.max() * factor;
     max_peaks.clear();
     foreach(p in peaks){
          if(p >= cuttoff){
               max_peaks.add(p);
          }
     }
     factor -= 0.1;
}

if(max_peaks.size() >= 3){
     t = max_peaks.last_time() - max_peaks.first_time();
     frequency = t / max_peaks.size();
} else{
     frequency = null;
}
```

Algorithm 1: Frequency Heuristic Feature

2.3 Modeling

This section describes the data mining and classification model building techniques used in this work. Critically, three kinds of models are evaluated: Personal, Universal, and Hybrid. Each model type is constructed under different conditions and each of which comes with different assumptions and applications. Sections 2.3.1 through 2.3.3 define each model in greater detail. Section 2.3.4 describes the general algorithms and software that were used to implement the abstract model types.

2.3.1 Personal

Personal models use training data from only the user for whom the model is intended. These models require a training phase to collect labeled data from each user. The training and test data come from the same user, but contain distinct examples. Personal models have the advantage that they may match the idiosyncrasies of the intended user, but they require each user to provide training data, limiting the amount of data available and potentially also this model type's performance.

10-fold cross validation is applied to each user's data and thus with the WISDM data, 590 (59 × 10) personal models are evaluated, and the HASC set requires the construction of 4,140 personal models since it contains 414 users. Since each user has a very limited amount of data (on average 160 examples in the WISDM set, and only 26 in the HASC set), 10-fold cross validation is essential. The confusion matrices are created by summing the counts in each cell over all 59 or 4,140 runs.

To generate the learning curves for personal and hybrid models, we generate k folds for each user (where k is the total number of examples a user has, divided by the number of examples we want in the training set). This required the generation of tens of thousands of totally independent models for the WISDM set, and hundreds of thousands for the HASC set. The results for all folds of each user are combined into the user's score, and then the results for all of the users are averaged.

2.3.2 Universal

Universal models are built with training data from a panel of users that will not subsequently use the model (thus the training and test sets have no common users). These models are applied to a new user without requiring additional labeled training data or model regeneration for the new user. Universal models have the advantage that they can be built once and then apply to all people. They can also include data from many users for training purposes, potentially leading to more accurate models.

To construct universal models, data from 58 WISDM users (or 413 HASC users) is placed into the training set and data from 1 user is placed into the test set. This process is repeated 59 (or 414) times, which allows us to generate reliable performance metrics and characterize the performance on a per-user basis.

To generate the learning curves for universal models, it was impractical to generate every possible combination. Instead, we randomly selected the required number of users and then randomly selected the required number of examples from each selected user to build training sets. This process was repeated 50 times for each number of users and each amount of training data, and the results are averaged. Again, tens of thousands of independent models were evaluated to determine the aggregate results.

2.3.3 Hybrid

Hybrid models are a mixture of universal and personal models. The training set has data from both the test subject and other users, but the test set's examples are distinct. The hybrid model also requires training data and model generation for each user, but can potentially outperform the personal model because it utilizes additional training data from other users.

To build hybrid models, one could take the most direct and common approach and simply place all the data form all of the users in a single file, then have the classification software automatically randomly partition the data into 10 folds for cross validation. However, to match the rigor of the other experiments presented here, and to have more detailed results, we did not use this method. Instead, we took the training and testing files for the personal models' learning curves as a base and combined each training file with the corresponding universal model's training file. This resulted in training sets that contained a controlled amount of data from the target user, and all of the data from every other user. It also enabled us to easily repeat the cross-validation already built into the personal model experiments.

2.3.4 Classification Algorithms

Our AR models are induced from the labeled examples using the following WEKA [47] classification algorithms: decision trees (J48 and random forest, RF), instance-based learning (IBk), neural networks (Multilayer perceptron, NN), rule induction (J-Rip), naive Bayes (NB), and logistic regression (LR). Default settings from WEKA are used for all learning methods except NB, where kernel estimation is enabled, and IBk, where we set k=3 (IB3) so we use 3 nearest neighbors.

3 Results and Discussion

The predictive accuracy associated with the personal, hybrid, and universal models on our data set is displayed in Table 3. These results make it quite clear that for every classification algorithm the personal models perform best, the hybrid models perform second best, and the universal models perform worst. Furthermore, the personal models always achieve a very high level of accuracy and perform dramatically better than the universal models. While this result may seem easy to justify, since people move differently from one another, the result is far from obvious since the personal models are trained from dramatically less data.

	RF	NB	LR	IB3	NN	J48	J-Rip	Avg
Personal	98.4	97.6	97.7	98.3	98.7	96.5	95.1	97.4
Hybrid	95.0	82.8	84.6	96.5	92.1	91.8	91.1	88.7
Universal	75.9	74.5	72.7	68.4	67.8	69.1	70.2	70.9
Average	89.8	85.0	85.0	87.7	86.2	85.8	85.5	85.7

Table 3: Accuracies of All Classifiers and Model Types Using the WISDM Data Set

The hybrid models typically perform closer to the personal models than the universal models. Given how well the personal models do, this is a bit surprising. It implies that the hybrid models can make effective use of the personal data in the data set, even though only a small fraction of the data (on average 1/59) is personal data. This means that the classification algorithms can effectively identify the movement patterns of a particular user from among a host of users. In retrospect this is not so surprising, since our recent work has shown that biometric models induced from accelerometer data can identify a user from a set of users with near perfect accuracy [28]. Because the hybrid model performs more poorly than the personal model, but still requires the acquisition of labeled training data from the target user, there seems little reason to utilize the hybrid model. The only exception might be when the amount of personal data is extraordinarily small, thus increasing the importance of the relatively common universal data. But as we see later in this section, even when there is very little personal data the personal models outperform the universal models. This result surprised us, but the real surprise is just how effective personal data is, even in extremely small quantities.

The main focus of this thesis is on the comparative performance of the three types of AR models, but our results also suggest which classification methods may be best suited to AR, given our formulation of the problem (see the underlined values in Table 3). For personal models, NN does best, although RF and IB3 also perform competitively; for hybrid models IB3 does best but RF performs competitively; for universal models, RF does best. Averaged over the three types of models, RF does best. Due to space considerations, many of our detailed results focus only on RF, IB3, and NN, the three best performers.

The personal results in Table 3 are mirrored by our results on the HASC data:.using RF, the personal accuracy is 97.2%. However, the universal accuracy is higher than for our data set, at 85%. We examine the causes of this discrepancy more thoroughly throughout this section.

3.1 Sources of Confusion

Table 4 shows the AR performance for the personal and universal models for each activity, using the three best-performing classification algorithms and a baseline strategy. The baseline strategy always predicts the specified activity, or, when assessing overall performance, the most common activity. The baseline allows us to consider class imbalance. Personal models outperform universal models for every activity, usually by a substantial amount, although universal models still outperform the baseline.

	% of Examples Correctly Classified						
	Personal			Universal			Base-line
	IB3	RF	NN	IB3	RF	NN	
Walking	99.1	98.9	99.0	65.2	73.0	56.8	36.6
Jogging	99.5	99.6	99.8	89.0	95.2	92.1	21.0
Stairs	96.4	96.8	98.0	65.1	61.5	68.0	16.7
Sitting	98.2	98.7	98.1	67.6	81.5	66.7	12.3
Standing	96.4	97.8	97.5	75.2	91.9	88.0	7.4
Lying Down	95.9	95.0	97.5	34.0	45.1	45.5	6.1
Overall	98.3	98.4	98.7	68.4	75.9	67.8	36.6

Table 4: Accuracy Per Activity for the Top Three Classifiers Using the WISDM Data Set

Table 5 provides the confusion matrices associated with the Random Forest learner for the universal and personal models. These results show that most of the errors, for both the universal and personal models, are the result of confusing walking with stairs and lying down with sitting.

(a) Universal		Predicted Class					
		Walk	Jog	Stairs	Sit	Stand	Lie
Actual Class	Walking	**2480**	66	819	22	8	2
	Jogging	51	**1854**	41	1	0	1
	Stairs	518	69	**953**	2	4	3
	Sitting	7	5	3	**931**	19	178
	Standing	3	0	12	19	**633**	22
	Lying down	7	0	5	284	14	**255**

(b) Personal		Predicted Class					
		Walk	Jog	Stairs	Sit	Stand	Lie
Actual Class	Walking	**3359**	3	30	1	3	1
	Jogging	5	**1940**	3	0	0	0
	Stairs	40	5	**1500**	2	2	0
	Sitting	3	0	1	**1128**	2	9
	Standing	5	0	8	2	**674**	0
	Lying down	3	2	4	18	1	**537**

Table 5: Confusion Matrices for RF Models on the WISDM Data Set

The confusion between walking and stairs may be due to the similar time between footsteps and exacerbated by the differences that individual people exhibit when performing these activities (jogging probably does not exhibit this problem due to the shorter time between footsteps and the more extreme acceleration values). It is easy to see why lying down and sitting are confused, since the orientation of one's pocket will be similar for both of these stationary activities. While the results for personal models in Table 5b show that these activities are still confused the most, the frequency of such errors is reduced by more than a factor of 10. *This indicates that it is possible to learn the* *user-specific differences in these two sets of activities and that these differences are not the same for all people*. This is a key argument for the use of personal models and perhaps the most important conclusion of this paper.

The confusion matrices in Table 5 are presented in aggregate over many users and cross-validation folds. To better measure this phenomenon, we also looked at the rates of errors produced by the individual classifiers. We use majority voting (MV) to identify instances where an individual classifier misclassifies one of a user's activities more than 50% of the time (e.g. the majority of a user's *walking* is confused with *stairs*). Each user had up to six majority votes, one for each activity. The results are presented in Table 6. They indicate that while this very serious error (being wrong about at least one activity the majority of the time) does not happen at all in personal models, it is extremely likely (95%) in universal models. Table 6 also shows that when these errors happen in universal models, they usually happen for multiple (1.8) activities simultaneously. This is less promising than their aggregate performance in Table 5, which shows universal models correctly predicting each activity the majority of the time. To our knowledge, no metrics like those in Table 6 have been published in prior work using universal models. These results represent a major challenge to claims about the performance of universal models.

	Personal	Hybrid	Universal
% of Incorrect MV	0	6.9	30
% of Users with One or More Incorrect MV	0	9.8	95
Mean Incorrect MV's per User with Incorrect MV	0	1.2	1.8

Table 6: Majority Voting Results for Individual Activities Using RF and the WISDM Data Set

The confusion between sitting and lying down is not a factor with the HASC data set, where standing is the only stationary activity. This strongly contributes to the increased performance of universal models on the HASC set; the most difficult to identify classes are missing. For this reason, we focus more on our own, more challenging, data set than on the HASC set.

3.2 Consistency Across Users

The results presented thus far are averages over all users. However, it is informative to see how AR performance varies between users. Figure 2 provides this information for the personal models on the WISDM data set and shows that these models perform consistently well for almost all users. The minor outliers that do show poor performance are primarily due to the high levels of class imbalance in those users' data. For example, the user with the second highest error rate has 59 examples of walking data but only between 5 and 8 examples for each of the other activities. The user with the worst accuracy had a similar class distribution, and also had a leg injury. Thus, the few problems that do occur for the personal models appear to be due to high levels of class imbalance or due to an injury.

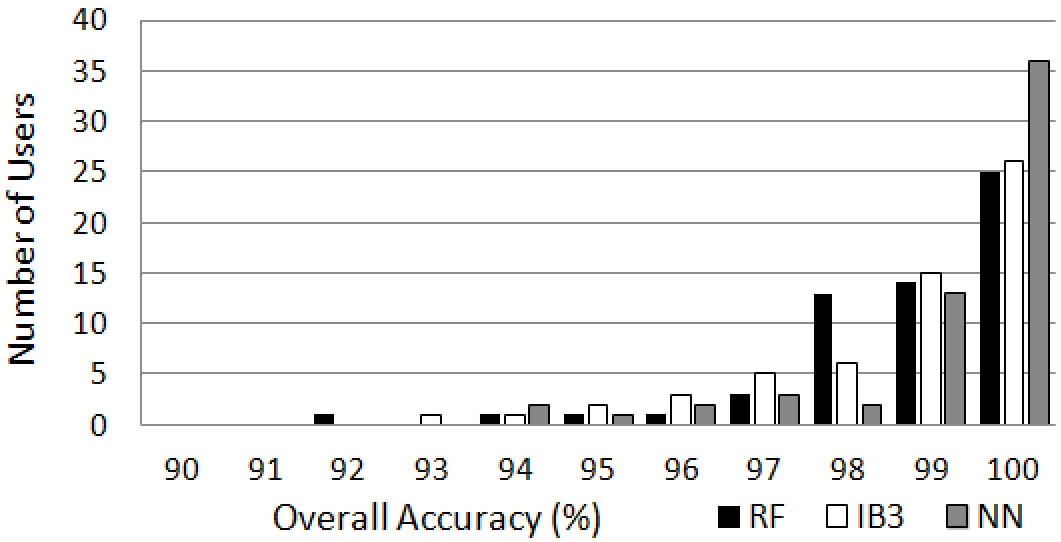


Figure 2: Histogram of Personal Model Accuracies for Individual Users

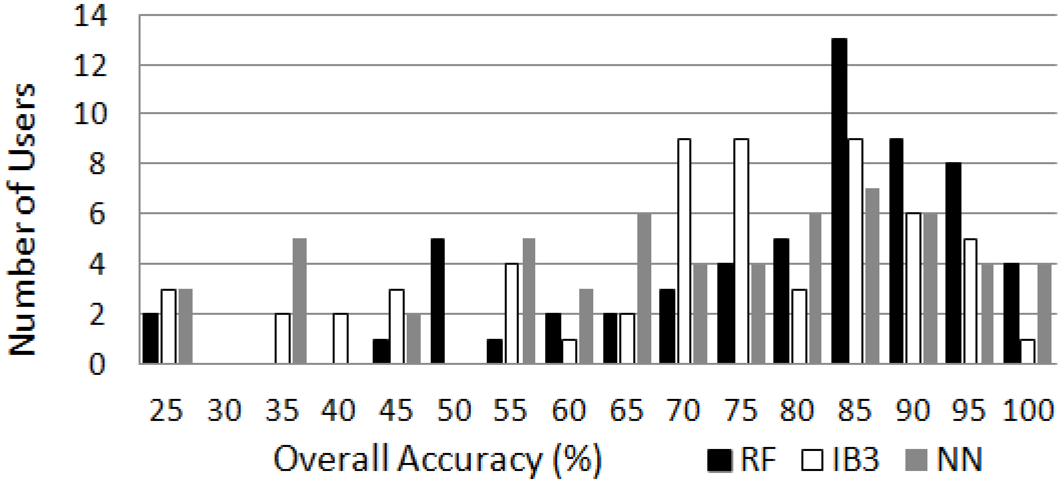


Figure 3: Histogram of Universal Model Accuracies for Individual Users

Figure 3 shows a much broader distribution of performance for the universal models. There are still some users with classification accuracies in the 95-100% range, but the accuracies vary widely and there are some users with extremely low accuracies. Detailed analysis showed that most of these very poor-performing users performed quite well when using personal models. These results further support the view that there are many users who move differently from other users, which “confuses” the universal models, while the personal models can learn these user-specific differences to achieve consistently good results.

Like the results in Table 6, we were able to find little prior work that broke down the performance by user. Those few that did had fewer than 5 users, making it impossible to determine a trend [3, 18]. The inconstant performance across different users is a major barrier to the generalizability of universal models, and should be specifically evaluated by future work in the area. This is an important argument for larger data sets such as ours [46].

As part of our data collection protocol, we collect information about the physical characteristics of each user (height, weight, sex, shoe size, etc.). We analyzed this information to determine if people with particular or extreme characteristics are especially hard to predict for universal models, but partly due to the limited number of users, we could find only suggestive patterns. For example, of the 10 users that were hardest to predict using the universal RF models, 3 were among the oldest users in the study. In the future we plan to collect data from substantially more users so that we can better assess the impact of such factors.

3.2.1 Augmented Universal Models

In an effort to better understand the relationship between physical traits of users and the performance of our models, we repeated the experiments for the universal models several times, each time augmenting the training and testing sets with an additional attribute of personal information (viz. *height, weight, shoe size,* and *sex*). Such information requires less effort to obtain than labeled training data, and thus presents the possibility of building more targeted universal models, which may have some of the benefits of personal models, but without the cost. Additionally, we matched users with each other based on the similarity of their transformed accelerometer data and constructed another set of models using data from the most similar users. These are alternate versions of the protocols used in prior research [30].

Where the prior work showed improvement by restricting universal models to similar users, this improvement was limited and only brought the models' accuracies per activity to rates similar to those of our universal models in Table 3 [30]. Because our users are already very similar (primarily college students), while other studies' subjects are less homogenous, it is not very surprising that on our data set these techniques to select similar users do not improve accuracy. This suggests the poor performance of models that are not trained from labeled personal data is the result of differences that cannot be explained entirely by differences in user demographics, or compensated for by the similarity of unlabeled data. We conclude that simple demographic information, or even unlabeled data, is no substitute for labeled accelerometer data from that user, which may encode idiosyncrasies with the subject's movements.

3.2 Effects of Training Set Size

In the context of this paper, learning curves can be particularly insightful because the varying amount of training data may impact model types differently and acquiring labeled personal data can be quite costly. We begin by analyzing the learning curves for the personal and hybrid models, presented in Figures 4 and 5. These figures show that personal and hybrid models improve their performance rapidly. Figure 4 shows that with only 20 seconds (2 examples) of personal data, all three classifiers clearly outperform universal models.

Furthermore, our results show that after 2 minutes (only 12 examples) of labeled data for each activity from a user, RF models reach 98.7% accuracy. With 3 minutes of data, this increases to 99.2% and at 5 minutes we reach 99.6%. But the key takeaway here is that we need only a miniscule amount of personal data---one example per activity---in order to outperform an universal model built using far more data.

To generate the same curves for hybrid models in Figure 5, we used the universal model training sets as a base and added in the data from the personal model training sets used to generate Figure 4. The resulting training sets included all available universal data and the amount of personal data specified on the x-axis.

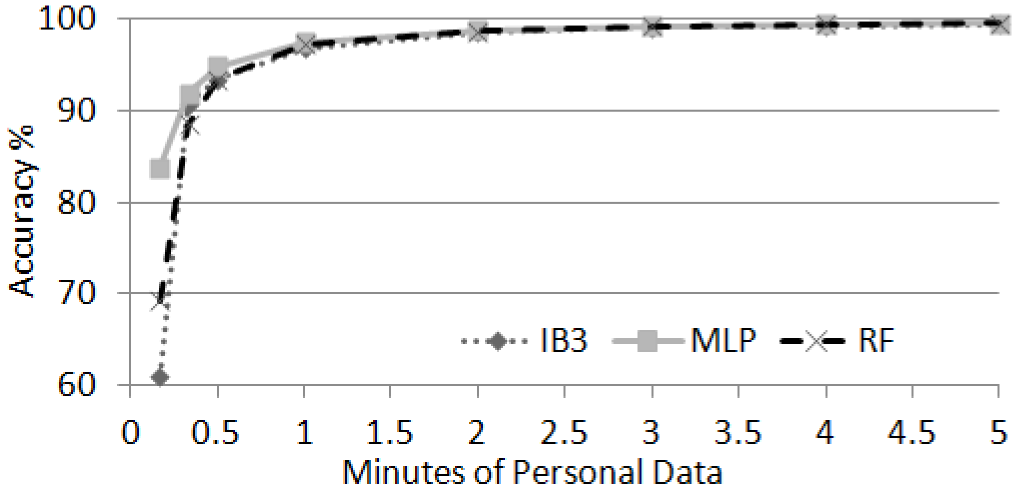


Figure 4: Learning Curves for Personal Models on the WISDM Data Set

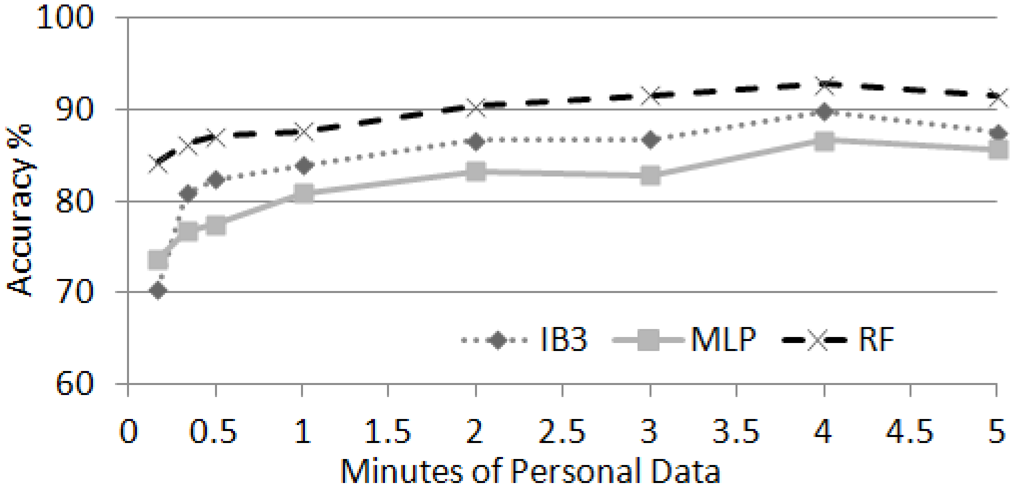


Figure 5: Learning Curves for Hybrid Models on the WISDM Data Set

Figure 4 shows a similar, though less dramatic pattern for hybrid models as we saw with personal models. With 10 seconds of labeled personal data for each activity, two out of our three hybrid models outperform personal models, and all outperform universal ones.

From 20-30 seconds and beyond, personal models dramatically outperform hybrid ones. universal models have an additional factor to consider, however: we can vary both the amount of data from each user and the number of users in the training set. Figure 6 shows these two factors plotted with classifier accuracy in 3-dimensions. The surface is shaded to make its accuracy dimension clearer.

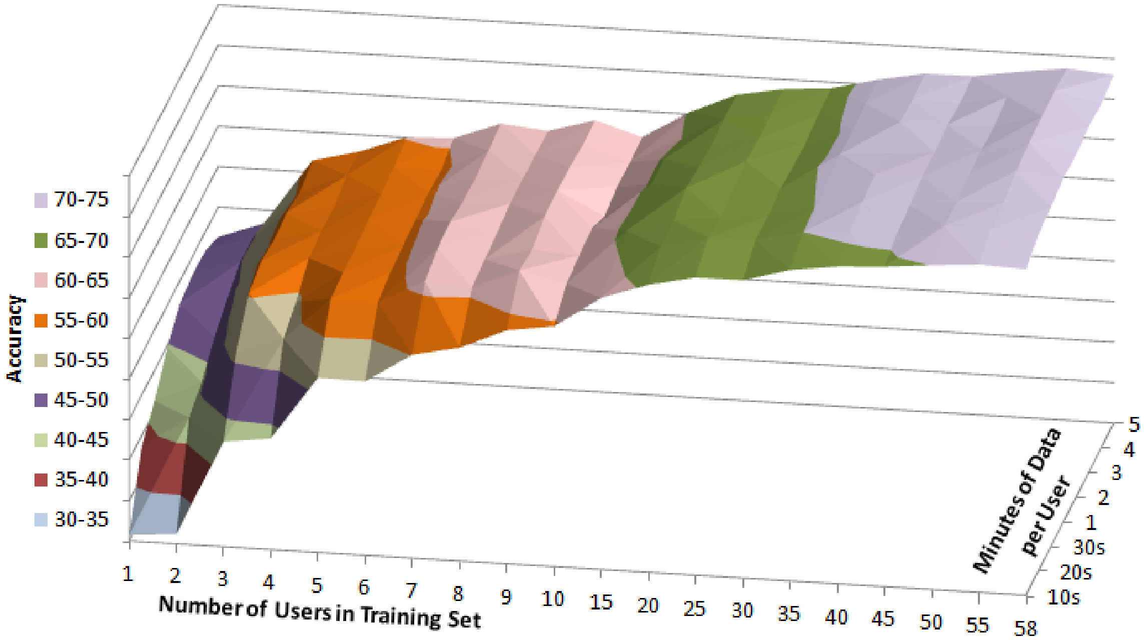


Figure 6: Impersonal Model Learning Curve for RF using WISDM Data Set

Figure 6 allows us to make several important observations, although some of these may be difficult to see at a quick glance, given that there are three dimensions involved. Like with personal and hybrid models, there is little to no improvement in accuracy from including more than two minutes of training data from each user. However, when there are few users in the training set, including more data per user (up to 2 minutes) dramatically improves performance. As the number of users in the training set increases, the value of additional data from each user decreases. With 58 users in the training set, there is little difference between using 10 seconds or 5 minutes of data per user per activity. Thus, the best way to improve the accuracy of universal models is to increase the number of users in the training set, rather than increasing the amount of data per user—this is another key lesson. This tradeoff is dramatic: a model generated from 5 users with 10 seconds of data per activity outperforms one generated from 2 users with 60 seconds of data per activity---even though the first data set has less than half the total data. Figure 6 also shows us that accuracy has not reached a plateau, and hence having data from more than 58 users will yield improvements in accuracy.

In order to understand this trend better, we generated the same curves for the HASC data set, which has seven times as many users. The three-dimensional chart for this set has the same shape as Figure 6, except shifted slightly up due to the simpler activity set and greater class imballance. For ease of comparison, we show the HASC curve side by side with our own in Figure 7. Beyond 210, the increase in accuracy of universal models from additional users is negligible.

We also find that the performance gap between the same algorithms on the two data sets decreases as we increase the number of users in the set. With only one user, the gap is 10 points (46% to 56%), but at 55 users the gap is only 5 points (73% to 78%). Further, we fitted a logarithmic function, $f(x)=6.53*\ln(x)+47.75$, $R^2=0.97$, to the accuracy of universal models on our data set so we could compare it to the HASC set for greater numbers of users. The fitted curves for each set are shown as the thinner, smooth lines in Figure 7. Based on this function, increasing the number of WISDM users from 58 to 200 would bring the average accuracy to 82%, while the accuracy on the HASC set at 200 users is only 83%. Of course, these projections may not hold.

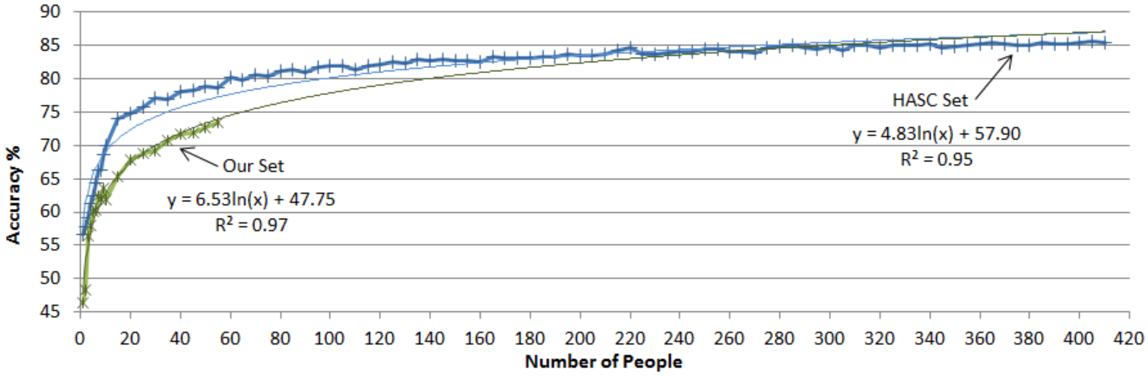


Figure 7: Impersonal Model Learning Curves for HASC and WISDM Data Sets

In any case, it is clear that personal and even hybrid models will substantially outperform universal models, while using dramatically less data. We anticipate that no number of users would allow an universal model to outperform our personal models, which are based on about 24 minutes of data divided over six activities. One of our key findings is that even universal models using data from a homogenous group are limited by the diversity of gait patterns, an obstacle easily overcome by building personal models.

4 Conclusions

In this paper we describe and evaluate a data mining approach for implementing activity recognition, using only smartphones. We demonstrate that nearly perfect results can be achieved if a personalized model is constructed, even using only a very small amount of user-specific training data. We further show that universal models perform much worse than personal models, even under the best conditions where they are trained on similar users. Analysis of the data shows that universal models cannot effectively distinguish between certain activities, whereas personal models can effectively learn the user-specific differences that confound universal models. We also show that while the poor performance of universal models is affected by some "idiosyncratic" users, whose activities cannot be accurately predicted, the problem is widespread and not restricted to a few problem users. Although there appears to be room for marginal improvement in the accuracy of universal models by increasing the number of people in the training set up to about 210, and related work [30] shows that selecting similar subsets of users also increases the accuracy of universal models to a point, it is clear that the personal models are able to substantially outperform universal ones while using only a very small amount of personal data. We also showed that if one does want to improve the performance of universal models, it is far better to obtain more users than to obtain more data per user.

In this paper we also evaluated the performance of hybrid models and showed that their performance was generally inferior to personal models but consistently better than that of even best-case universal models constructed with similar users. Given that hybrid models require user-specific training data, and personal models with the same amount of data generally perform better, one is better off using personal models. The one exception is the extreme case where we have only 10 seconds of data from a user, but that case is unlikely, since anyone already gathering personal data would easily be able to gather more than 10 seconds of it from healthy people. Thus we conclude that, hybrid models will never realistically outperform personal models.

This study provides the most thorough analysis of the relative performance of personal, universal, and hybrid AR models, and we view this as a key contribution of the paper (along with the various lessons learned that we listed above). This is also the first study to examine in great detail the effect of training set size on AR performance, as it varies by number of unique users and quantity of data per user. Additionally, it is the first study to consider in depth the reliability of universal models' performance for individual users rather than only in aggregate. We show that reporting only aggregate results is likely to be misleading about the performance of universal models for individual users. This work should greatly influence the design of future AR research, as well as the design of AR systems and the context representation systems which rely on them.

One of our key achievements is making this AR research available to smartphone users and researchers alike, via our downloadable app, Actitracker [1]. Our AR system tracks a user's activities and provides reports via a secure account and web interface. This mobile health application helps people ensure that they and their children are sufficiently active to maintain good health and avoid the many health conditions associated with inactivity.

Because of the results of this research, we have included a self-training mode in our system, so one can quickly generate personal labeled activity data to achieve good predictive performance. This data is uploaded to our server, which automatically generates a personalized AR model that then replaces the universal model for that user.

Bibliography

(1) actitracker.com

(2) Z.S. Abdallah, M.M. Gaber, B. Srinivasan, and S. Krishnaswamy, CBARS: Cludter Based Classification for Activity Recognition Systems, In Proc. First International Conference on Advances Machine Learning Technologies, 2012.

(3) Z.S. Abdallah, M.M. Gaber, B. Srinivasan, and S. Krishnaswamy, StreamAR: incremental and active learning with evolving sensory data for activity recognition, In Proc. 24th IEEE International Conference on Tools with Artificial Intelligence, 2012.

(4) D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine, In Proc. International Workshop of Ambient Assisted Living, 2012.

(5) D. Anguita, et al., A Public Domain Dataset for Human Activity Recognition Using Smartphones, In Proc. European Symposium on Artificial Neural Networks, Comtational Intellegence, and Machine Learning, 2013

(6) D. Anguita, et al., Energy Efficent Smartphone-Based Activity Recognition using Fixed-Point Arithmatic, Journal of Universal COmputer Science vol. 19, no. 9, 2013: 1295-1314

(7) L. Bao, and S. Intille, Activity recognition from user-annotated acceleration data, Lecture Notes Computer Science 3001, pp.1-17, 2004.

(8) T. Brezmes, J.L. Gorricho, and J. Cotrina, Activity recognition from accelerometer data on mobile phones, In Proc. 10th International Work-Conference on Artificial Neural Networks, pp. 796-799, 2009.

(9) Y. Cho, Y. Nam, Y-J. Choi, and W-D Cho, Smart-Buckle: human activity recognition using a 3-axis accelerometer and a wearable camera, HealthNet, 2008.

(10) M. Derawi and P. Bours, Gait and Activity Recognition Using Commercial Phones, Computers and Security vol. 39, 2013: 137-144.

(11) S. Dernbach, B. Das, N.C. Krishnan, B.L. Thomas, D.J. Cook, Simple and complex activity recognition through smart phones, In Proc. 8th International Conference on Intelligent Environments, pp. 214-221, 2012.

(12) T.M. Do, et al., HealthyLife: An Activity Recognition System with Smartphone using Logic-Based Stream Reasoning, Mobile and Ubiquitous Systems: Computing, Networking, and Services, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Volume 120, 2013, pp 188-199.

(13) M. Fahim, Evolutionary Learning Moels for Infoor and Outdoor Human Activity Recognition, PhD Thesis, Kyung Hee University, 2014

(14) M. Fahim, I. Fatima, and S. Lee, EFM: Evolutionary Fuzzy Model for Dynamic Activities Recognition Using a Smartphone Accelerometer, Applied Intellegence, 39.3, 2013: 475-488.

(15) J.B. Gomes, et al., Mobile Activity Recognition using Ubiquitous Data Stream Mining, In Proc. 14th International COnference on Data Warehousing and Knowledge Discovery, 2012

(16) N. Gyorbiro, A. Fabian, and G. Homanyi, An activity recognition system for mobile phones, Mobile Networks and Applications, vol. 14, no. 1, pp. 82-91, 2008.

(17) M. Han, An intergativ Human Activity Recognition Framework based on Smartphone Multimodal Sensors, Doctoral Thesis, Kyung Hee University, 2013

(18) M. Han, et al., HARF: A Hierarchical Activity Recognition Framework Using Smartphone Sensors, UCAmI Lecture Notes in Computer Science 8276, 2013: 159-166.

(19) Z. He, and L. Jin, Activity Recognition from Acceleration Data Based on Discrete Cosine Transform and SVM, In. Proc. IEEE International Conference on Systems, Man, and Cybernetics, 2009.

(20) Y. He and Y. Li, Physical Activity Recognition Utilizing the Built-in Kinematic Sensors of a Smartphone, International Journal of Distributed Sensor Networks, 2013.

(21) M. Kastner, M. Strickert, and T. Villmann, A Sparse Kernelized Matrix Learning Vector Quantization Model for Human Activity Recognition, In Proc. European Symposium on Artificial Neural Networks, Computational Intelligene, and Machine Learning, 2013.

(22) N. Kawaguchi, et al., HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings, In proc. 2nd Augmented Human International Conference, pp. 27, March 2011.

(23) A.M. Khan, Y.K. Lee, S.Y. Lee, and T.S. Kim, Human Activity Recognition via an Accelerometer-Enabled-Smartphone Using Kernel Discriminant Analysis, In Proc. 5th International Conference on Future Information Technology, pp. 1--6, 2010.

(24) A.M. Khan, et al., Exploratory Data Analysis of Acceleration Signals to Select Light-Weight and Accurate Features for Real-Time Activity Recognition on Smartphones, Sensors, vol. 13, 2013: 13099-13122.

(25) J.P. Koplan, C.T. Liverman, and V.I. Kraak, Preventing childhood obesity: health in balance, National Academies Press, Washington DC, 2005.

(26) N.C. Krishnan, S. Panchanathan, Analysis of low resolution accelerometer data for human activity recognition, In Proc. International Conference on Acoustic Speech and Signal Processing, ICASSP, 2008.

(27) N.C. Krishnan, D. Colbry, C. Juillard, S. Panchanathan, Real time human activity recognition using tri-axial accelerometers, In Proc. Sensors Signals and Information Processing Workshop, 2008.

(28) J.R. Kwapisz, G.M. Weiss, and S.A. Moore, Cell phone-based biometric identification, In Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems, 2010.

(29) J.R. Kwapisz, G.M. Weiss, and S.A. Moore, Activity recognition using cell phone accelerometers, SIGKDD Explorations, vol. 12, issue 2, pp. 74-82, 2010.

(30) N.D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A.T. Campbell, and F. Zhao, Enabling large-scale human activity inference on smartphones using community similarity networks (CSN), In Proc. 13th International Conference on Ubiquitous Computing (UbiComp 2011), pp. 355--364, September 2011.

(31) N.D. Lane, et al., Community Similarity Networks, Personal Ubiquitous Computing, 2013.

(32) Y. Lee, and S. Cho, Activity recognition using hierarchical hidden markov models on a smartphone with 3D accelerometer, In Proc. 6th international conference on Hybrid artificial intelligent systems, pp. 460--467, 2011.

(33) Y. Lee and S. Cho, Activity Recognition with Android Phone Using Mixture-of-Experts Co-Trained with Labeled and Unlabeled Data, Neurocomputing, 2013.

(34) J.W. Lockhart, and G.M. Weiss, The Benefits of Personalized Smartphone-Based Activity Recognition Models, In Proc. SIAM International Conference on Data Mining, 2014 (In Press).

(35) U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, Activity recognition and monitoring using multiple sensors on different body positions, In Proc. IEEE International Workshop on Wearable and Implantable Sensor Networks, vol. 3 no. 5, 2006.

(36) E. Miluzzo, N.D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S.B. Eisenman, X. Zheng, A.T. Campbell, Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application, In Proc. 6th ACM Conference on Embedded Networked Sensor Systems, pp. 337-350, 2008.

(37) X. Qi, et al., AdaSense: Adapting Sampling Rates for Activity Recognition in Body Sensor Networks, In Proc. IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS), 2013: 163-172.

(38) N. Ravi, N. Dandekar, P. Mysore, M.L. Littman, Activity recognition from accelerometer data, In Proc. 17th Conference on Innovative Applications of Artificial Intelligence, pp. 1541-1546, 2005.

(39) A. Reiss, G. Hendeby and D. Stricker, A Competitive Approach for Human Activity Recognition on Smartphones, In Proc. European Symposium on Artificial Neural Networks, Computational Intellegence, and Machine Learning, 2013: 455-460.

(40) D. Riboni, and C. Bettini, COSAR: hybrid reasoning for context-aware activity recognition, Personal and Ubiquitous Computing, vol. 15, no. 3, pp. 271-289, 2011.

(41) D. Roggen, et al., Opportunity: towards opportunistic activity and context recognition systems, In Proc. 3rd IEEE WoWMoM Workshop on Autononomic and Opportunistic Communications, 2009.

(42) N. Roy, A Misra, and D. Cook, Infrastructure-Assisted Smartphone-based ADL Recognition in Multi-Inhabitant Smart Environments, In Proc. IEEE International Conference on Pervasive Computing and Communications (PerCom), 2013: 38-46.

(43) M. Shoaib, et al., Towards Physical Activity Recognition Using Smartphone Sensors, In Proc. 10th International Conference on Ubiquitous Intelligence & Computing, 2013: 80-87.

(44) P. Siirtola and J. Roning, Ready-to-Use Activity Recognition for Smartphones, In Proc. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2013: 59-64

(45) E.M. Tapia, S.S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman, Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor, In Proc. 11th IEEE International Symposium on Wearable Computers, pp. 37-40, 2007.

(46) www.cis.fordham.edu/wisdm/dataset.php

(47) I.H. Witten, and E. Frank, Data mining: practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann, 2005.

(48) World Health Organization, Physical inactivity a leading cause of disease and disability, warns WHO, 2002, www.who.int/mediacentre/news/releases/release23

(49) W. Wu, S. Dasgupta, E.E. Ramirez, C. Peterson, G.J. Norman, Classification accuracies of physical activities using smartphone motion sensors, Jurnal of Medical Internet Research, vol. 14 no. 5, 2012 .

(50) Z. Yan, et al., An Exploration with Online COmplex Activity Recognition using Cellphone Accelerometer, UbiComp Adjunct, 2013.

(51) J. Yang, Tooward physical activity diary: motion recognition using simple acceleration features with mobile phones, in Proc. 1st International Workshop on Interactive Multimedia for Consumer Electronics at ACM Multimedia, 2009.

(52) B. Yuan, et al. Smartphone-based Activity Recognition Using Hybrid Classifier Utilizing Cloud Infrastructure for Data Aalysis, In proceeding of the 4th International Conference on Pervasive and Embedded Computing and Communication Systems. 2014.

Jeffrey William Lockhart

BS, Fordham University

The Benefits of Personalized Data Mining Approaches to Human Activity Recognition with Smartphone Sensor Data

Thesis directed by Gary Weiss, Ph.D.

Activity recognition allows ubiquitous mobile devices like smartphones to be context-aware and also enables new applications, such as mobile health applications that track a user's activities over time. However, it is difficult for smartphone-based activity recognition models to perform well, since only a single body location is instrumented. Most research focuses on universal activity recognition models, where the model is trained using data from a panel of representative users. In this paper we compare the performance of these universal models with those of personal models, which are trained using labeled data from the intended user, and hybrid models, which combine aspects of both types of models. Our analysis indicates that personal training data is required for high accuracy—but that only a very small amount of training data is necessary. This conclusion led us to implement a self-training capability into our Actitracker smartphone-based activity recognition system [1], and we believe personal models can also benefit other activity recognition systems as well.

VITA

Jeffrey William Lockhart, son of Kathleen Hennen and Tom Lockhart, was born on January 26, 1991, in Ames, Iowa. He entered Fordham University in 2009, supported by the Presidential Scholarship. Jeff received his bachelors of science in computer science and women's studies from Fordham University in 2013. During his time at Fordham, he was awarded numerous internal research support grants, and received both the Herbert W. Bomzer Award in Computer Science and the Women's Studies Essay Award.

While an undergraduate, Jeff enrolled in Fordham's accelerated masters program in computer science. There he continued to work as a research assistant in the Wireless Sensor Data Mining Lab, funded by the National Science Foundation. Upon graduation, Jeff will take up his place as a Gates-Cambridge Scholar at the University of Cambridge, England.