slow, sincenk for big matrx it isease rameesrayze image 10001 column ning**Facebook Project: User Profile in Social Media**

Yaqun Yu

University of Washington

yaqunyu@uw.edu

Xinhe Wang

University of Washington

xinhehe@uw.edu

Lan Lv

University of Washington

lyul@uw.edu

**ABSTRACT**

Social network analytics has been a trending topic in data mining and machine learning. There are many researches have been going on. In this project, we have 9500 Facebook users' information, which contains their posts, profile pictures and their id of the posts they liked. We have implemented many experiments by using different machine learning algorithms, different feature selection methods and different types of data for doing gender and age-group classification as well as big five scores of personality prediction. We finally got 81\% on gender classification, 56\% on age group classification and about reach the baseline of Rooted Mean Squared Error big five scores of personality prediction.

**Keywords**

Facebook data, User profile, Predict, Gender, Age, Five Big O Personalities, Machine Learning, Algorithms.

# INTRODUCTION

With the appearance of the social network, more and more people like to share their own information on the website. Meanwhile, researchers are becoming more interested in mining this data for using the result in personalized information access services, recommend systems, tailored advertisements and other applications that can benefit form personalization. This project is aim to building a system to predict the age group, gender and personality (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism) of Facebook users automatically when given their personal profile such as status updates, profile pictures and “likes” information. By using different Machine Learning techniques we are able to build up different model for prediction on different purpose. We separate image, text for different approach in order to increase overall accuracy. Along with different machine learning models we used many libraries focused on image and text processing, such as openCV and NLTK language toolkit. We will get into details for these techniques in the following sections.

# METHOLOGY

## Previous work:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

After brief introduce to this project, we started analysis data that current available, determine the model type and separate data for different predict approach. At this point, our data resources are:

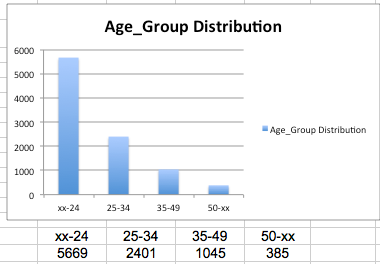
1. Training Data: Distribution for age group and gender is in Figure 1 and Figure 2

* One user profile csv file contains user ID, gender, age, and personality score
* 9500 user profile pictures at size 200 x 200 pixels
* 9500 user status text
* One like profile indicates relationship between user ID and the page ID they liked.

1. Test Data:

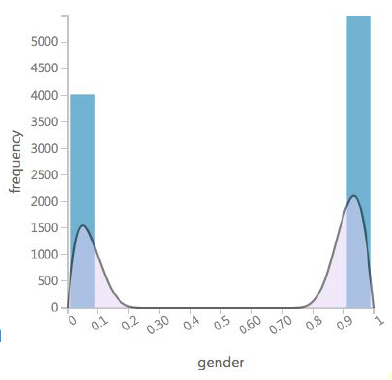
* Blank profile csv only contains header and user ID
* 334 user profile pictures at size 200 x 200 pixels
* 334 user status text
* One like profile csv indicate relationship between user ID and the page ID they liked

1. Papers: the paper can be roughly split into two categories: image process and text process. We went through these papers, inspired by their data processing skills and model selection.
2. Text book and machine learning knowledge come from class



**Figure 1. Age Group Distribution**

With study the resources, we determined we would build our model from two approaches: image and text. Each will be using supervised model because we have all information from training data and we can gain our accuracy through that. We will try different model based on their suitable level to our dataset. After compare accuracy for these two approaches, we will select the best for each part, and the combination will be our overall model.



**Figure 2. Gender Distribution.**

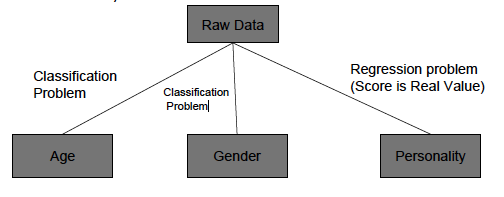


Figure 3. Model Separation figure.

## Text

There are 9500 text data files in total. Each file for each specific user contains the posts the user has published in Facebook. In this project, we are using these profiles as our dataset by doing text analytics; we tried to predict the gender, age groups and big five personality scores. In data pre-processing, token function separates the words by delimiters such as commas and quotes, but not question marks and exclamatory mark since we consider the number of those emotional marks will be able to help with big five-personality prediction. We also kept the combination of punctuations that forms the emotions, such as ^^, :), :(. In addition, we generalize all the numbers to be "\_\_NUMBERS\_\_". And we generalized the long repetitive terms to be "12\_LONG\_TERMS" and "17\_LONG\_TERMS" if their length is longer than 12 or 17 respectively. We believe those information are somehow involved into classifying gender, age groups and related with users' personalities. The algorithm, features selection/dimension reduction method have been used and experiments design are different in gender, age group and personality prediction. Since the experiment for gender classification gave the best result from image processing this section will just focus on text analytics on age group classification and personality prediction. In addition to the text files provided for all the users, there is also a "LIWC" textual file contains the well-organized and representative 82 terms for each user, which gave the best result for personality prediction results.

### Age Group Classification

The users are splitted into groups based on their ages: "xx-24" for users' age less than 24, "25-34" for age between 25 and 34, "35-49" for age between 35 and 49 and "49-xx" for age older than 49. We regard this problem as multiple class classification problems. The distribution of the age group is very right skewed and data are not uniformly distributed for different age groups. We implemented experiments in mainly textual data including all the posts and LIWC file. Firstly, the terms appeared less than two documents and more than 85% of documents are ignored, from selected terms, 50000 most frequent terms are selected. Using the selected terms, we built the document-term matrix and for each term we counted the tf-idf score. Tf-Idf represents term-frequency, inverse document frequency, which represents the relative frequency (how important) of a word to a document in a corpus. TF-idf is often used as weighting factor in information retrieval and text mining. In addition, we also applied the truncated singular value decomposition (SVD) for dimension reduction, which is also has been referred as Latent Semantic Indexing (LSI), method for selecting informative subspaces of feature spaces with the goal of obtaining a compact representation of document. The singular value decomposition of an m\*n matrix is a factorization of the form of unitary matrix, m\*n rectangular diagonal matrix with non-negative real numbers on the diagonal and an n\*n unitary matrix. The diagonal entries are known as the singular values of matrix. In the experiments the combination of SVD with using support vector machine algorithm gave the highest accuracy for age group classification. The usage of the combination of SVD for dimension reduction and SVM for model fitting framework has been using in many fields, such as automatic classification scheme. We chose the C-support SVM, which refers the ridge parameter for the SVM, controls the complexity of the classifier and help to avoid over-fitting by separating the patterns of each class by large margins. The best kernel function is linear function for SVM, which gave the best accuracy in result. Since the non-linear methods are too powerful and hard to cause over-fitting. From many experiments, when the number of components is 20 which gave the highest accuracy. In SVM model setting, C parameter is 1 and gamma is 1/number of features selected in this case is 1/50000. The result from 10 fold cross validation is around 61% for age group classification.

### Personality Prediction

For personality big five score prediction, which is a regression problem using textual data. We have the similar set up for the feature selection method and algorithm as the experiments for classifying age groups. The evaluation of the regression problem is measuring the rooted square mean error (RMSE) of the predicted real values with the true values. The some of the results are slightly higher than the baseline of about 0.01 or 0.02 and others are the same as baseline RMSE. Therefore, we tried to use the LIWC file which contains very organized, representative, clean terms which are usually used very related with the personality of people. The model we chose is Linear Regression Model, which is a least squares estimator linear classifier. In previous experiments, we also used the textual data for predicting the big five score of personality. The similar experiment procedures gave us the lowest RMSE: firstly, choose top 50000 most frequent terms and reduce the dimension to 50 using truncated SVD and finally use Bayesian Ridge Regression model. Bayesian Ridge Regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. In our experiments, Bayesian Ridge Regression performs better than linear regression in predicting the big five scores for personality. The RMSE is about 0.01 or 0.02 lower than RMSE given by linear regression.

## Image

### **Image For Gender Prediction**

Users’ profile pictures contains lot of information, from picture we can extract different features that would help us building the model for gender and even age. In pictures contain face, because there would be huge different between male and female, we can let machine learning these information and let them know which information belongs to female user and which belongs to male users. Based on this thought, our team started our first step on image processing: feature extraction. After we extract all features contains in an image, which model would be a better fit is our next challenge, with the knowledge of different popular machine learning algorithm, we are using them to find the best fit. During the fitting process, we faced the problem that the information matrix is 200x200=400000 for each pictures, if we using all information to fit our model it will take a long time before we can run our accuracy test. So next step is how to reduce features without losing important information. The structure for our remaining sections is arranged in following order: Feature extraction, feature reduces, Modeling, Improvement and Evaluation, hidden set test on VM and result analyze.

#### Feature Extraction

First of all, we plot the distribution of gender in our training data. This gives us a brief idea of the distribution for male and female, which we can take as a baseline to compare with our models for a better idea.

We did our research, and find the main popular library for image and video process is openCV, even though it is more powerful for live streaming data like video or camera, it build in functions still have great power for our image process. As mentioned in beginning of this section, to change image pixel information into our flatten text is our main task. From openCV tutorial we find that it has a build in functions for object detection using Haar feature-based cascade classifiers [1]. This is a machine learning based approach where a cascade function is training from a lot of positive and negative images. By using these it helps us to transfer image from visual to digital value, which contains pixels information.

#### Feature Reduces

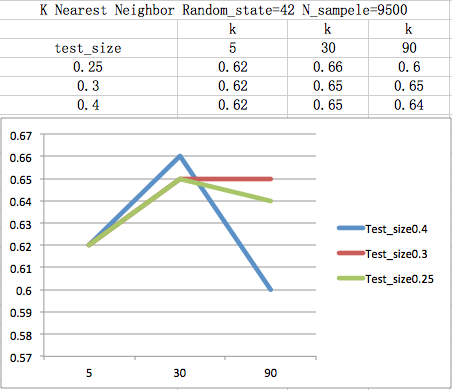
During the implement we find the color image is not help us a lot but increase complexity and running time. For many applications of image processing, color information doesn't help us identify important edges or other features and it is difficult for visualization. So we transfer image into gray scale. And after that we extract only face area as our input. This helps us reduce the feature contains in image. In our dataset, we have three kind of image: the one with one face, the one with multiple faces and the one with no face. For one face situation, it is easy to crop that face. For the one with multiple faces, we chose the biggest face as our user’s image. For the one with no face, we use the entire image and resize it to uniform size.

#### Modeling, Improvement and Evaluation

Since the problem is a classification problem, we choose the most useful models for classification. They are K Nearest Neighbor, PAC & SVM, Neural Network and Naive Bayes. For each algorithm, we change the parameter that may matters to accuracy and analyze why this matters

K-Nearest Neighbor (KNN): In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). In our test, with five fold cross validation, we have accuracy of 0.66%，k=30 gives best accuracy. The more neighbors it considers, the more accuracy majority would be under certain limit. Like in our case, 30 is the limitation, after 30, accuracy reached to the limit of this model.

PCA & SVM [2]: PCA is a method of summarizing some data with less characteristics, PCA [3] is not selecting some characteristics and discarding the others. Instead, it constructs some new characteristics. It finds the best possible characteristics, the ones that summarize the list of our target as well as only possible. SVM [4] are a set of supervised learning methods. It effective in high dimensional spaces, also memory efficient uses a subset of training points in the decision function. The combination of these two algorithms can give us good model for gender predict. With parameters: n\_components=200, 30x30, 0.25 random\_state=42, we have accuracy up to 71% in five fold cross validation. Normally PCA reduce the dimension of feature, it should increase the speed, on the contract, it rather slow than other models, so even it gives highest accuracy we did not select it for deep research.



**Figure 4. KNN for Gender.**

PCA & SVM [2]: PCA is a method of summarizing some data with less characteristics, PCA [3] is not selecting some characteristics and discarding the others. Instead, it constructs some new characteristics. It finds the best possible characteristics, the ones that summarize the list of our target as well as only possible. SVM [4] are a set of supervised learning methods. It effective in high dimensional spaces, also memory efficient uses a subset of training points in the decision function. The combination of these two algorithms can give us good model for gender predict. With parameters: n\_components=200, 30x30, 0.25 random\_state=42, we have accuracy up to 71% in five fold cross validation. Normally PCA reduce the dimension of feature, it should increase the speed, on the contract, it rather slow than other models, so even it gives highest accuracy we did not select it for deep research. SVM alone gives us accuracy on gender for 59%.

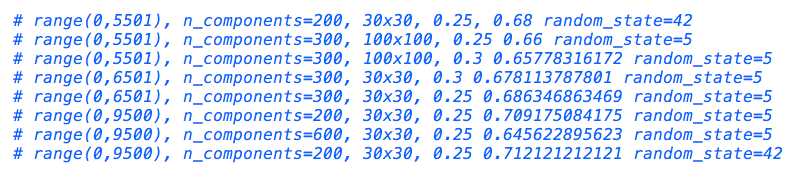
The other three models we tried are: SVD, gives us accuracy 59%, neural network, give 59% accuracy and Gaussian Naive Bayes, give us 61% accuracy.Macintosh HD:Users:yaqunyu:Desktop:GaussianNB.png

#### Hidden Set Test on VM

The hidden set test on VM is black box test. Aim to check if the model can output same accuracy as local. If the accuracy drops down dramatically, the original model might be over fitting.

#### Result Analyze

Features selection for image is essential for a good model, the more core feature selected, the better a model will be. Another influencing factor is the parameter in an algorithm. We tried many different sets of parameters in order for a better accuracy.



**Figure 5. Parameter Change**

### **Image For Age Group Prediction**

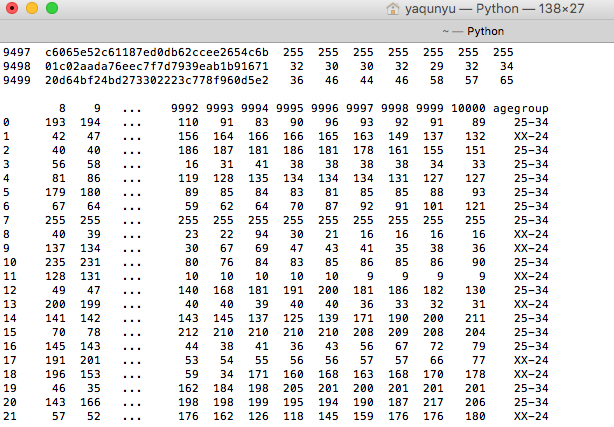
Age group prediction by image we follow the same routing, first process image, extract information that we are interested in, with the data frame we can fit in models for prediction. One thing need to pay attention is the information we have is not a request age range, so we need to transfer this information into 4 age groups first. Below is the distribution of our training data age group.

#### Feature Selection

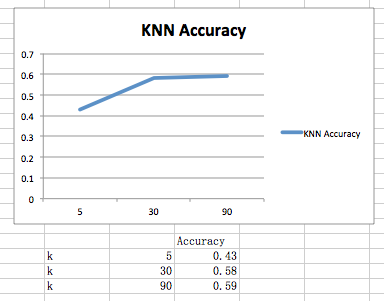
After gray scale image and resize it to 100x100, by using pandas data frame we change visual image into digital frame with 10000 image information columns with 1 age group columns. As show in Figure 6.

#### Modeling, Improvement and Evaluation

Similar to gender prediction, we select several Models as model candidates: K-Nearest Neighbor (KNN), PCA with SVM, Naive Bayes (Bernoulli) (Naive Bayes have three algorithms: Gaussian Naive Bayes, [Multinomial Naive Bayes](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB) , Bernoulli Naive Bayes)[4] Bernoulli Naive Bayes implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions.

Macintosh HD:Users:yaqunyu:Desktop:Bernoulli NB.png 

**Figure 6. Age Group Data Frame**

K-Nearest Neighbor (KNN): In our test, with five fold cross validation, we have accuracy of 0.59%，k=90 gives best accuracy. Like the gender prediction, our age model has its limit. In our case, limitation is up to k =90, which is the close square root of number of sample. After 90 accuracy is cannot increase anymore, otherwise will be over fit. 

**Figure 7. KNN Age Group**

For SVM and PCA, SVM is short for Support vector machines. They are a set of supervised learning methods. For age group is a little bit different from gender classification because this is a multi-class. PCA still running slow than others. SVM along gives us accuracy around same level with knn, but it takes times of knn’s running time to complete. So we go straight to deep search of knn instead of other models. Others we Bernoulli Naïve Bayes, which gives us 0.48 (+/- 0.23), we can see from result that this accuracy is pretty unstable, it can go up to 71% and also down to 25%. This is the reason we do not continue with it. Decision rule:

https://lh5.googleusercontent.com/HZIAa25K-aetzsu-IMqPkjD5yOdaRIdnqtmz-duDHgGCMl_rgFESQqHYo4JbpUQu9y-7jLGwognm73KCLhP_AFL9hoWS7K2vNg8QX-rYRzlSNkoqHyQ1JhBqG2NT0C1N_pSkW5l2els

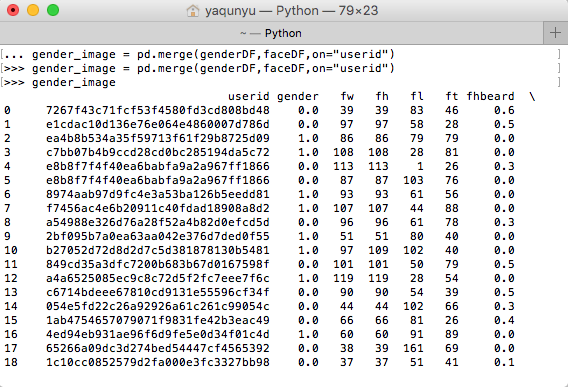
#### Hidden Set Test on VM

We also run multiple times for both pure image, pure text, oxford image combine with pure image on VM for the hidden set to get a better idea of what is more suitable for age group prediction. The result that pure text and oxford combine would be the best, both up to 56% and 55%, so we chose text as our final model. But we did not yet have time to try on oxford combine with pure image. We hope to take this offline as our future work.

### **Oxford for Gender and Age**

#### Oxford for Gender

Along with the progress of this project, we can access more and more information. The Oxford is an API that is powerful for extraction the feature from face image. After we get the profile file, we transfer it to a data frame that can be used in our models.



**Figure 8. Oxford Gender data frame**

In this file we only have users who contain face in their profile pictures that gives us 7915 records with 67 attributes. The other 1585 are the users who have no face in their image, so we need to separate our model for these two parts.

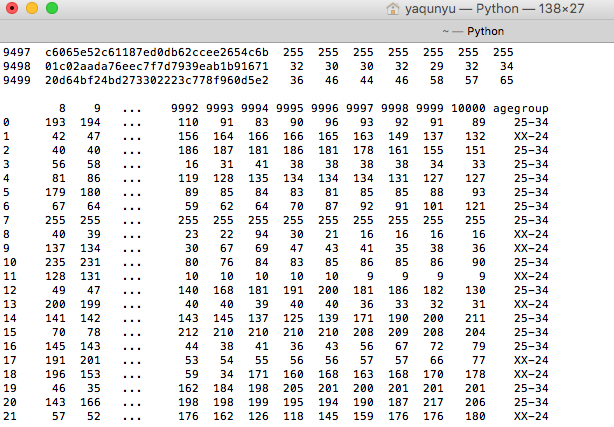
We first dealing with the one in oxford, there are some attributes related closely to gender, such as facial hair, the distribution among male will higher than female. So we both use manually selection and algorithm called Select K best for our feature selection. Manually select features: underLipBottom，facialHair\_mustache, facialHair\_beard, facialHair\_sideburns, faceRectangle\_top. We get accuracy from neural network up to 81%. From Select K best, k=3 with KNN (k=13), we can get accuracy up to 87%. So we chose select k best as our final model for user who has face in our oxford file.

For those who do not have face in their profile, we use whole image as test data. Training model we use the best one from previous work: KNN, and use all 9500 user image as our training data.

Combination of both model gives us 81% overall accuracy on VM black box test. Due to the research time limit, we have no time for other combination, so we will test it offline as our future work.

#### Oxford for Age

Similar to oxford for gender, we separate to two part, user in oxford and user who have no face, AKA not in oxford. In oxford we build up a matrix, see Figure 9, as our input data to train the best model we have so far for age: KNN, which gives us up to 59% on local 5 fold cross validation test. And for non-face users we again use whole image as input, and again, train on KNN model with k=90, and combine two model’s prediction as our output. The combination on VM black box test gives us accuracy for 55% overall. The pure text model gives us 56%. So we use our text model as our final model. We did not have time try on oxford with text combination. We also put this as our future work.



**Figure 9. Oxford Age Group data frame**

# RESULT AND DISCUSSIONS

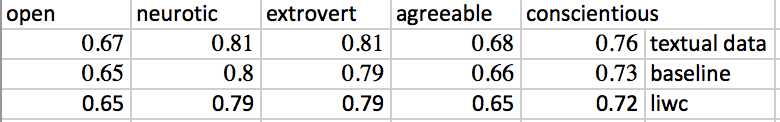
## Textual Data Analytics

### Age Group Classification

The accuracy from ran in hidden test data is 56% which is lower than the baseline 59%. In local the 5 folds cross validation the accuracy is 61%. The reason why we have lower accuracy about age group classification is following: Firstly, we chose 50000 words for training the model, which might not be the optimal chose since there must be many noises caused by our imperfect tokenize function. In addition, the classes for age groups are very unbalanced, and the classifier tried to fit the major classes very hard, which causes bias. Furthermore, we need to implement more experiments on using linear SVM for different c parameter and gamma.

### Personality Big Five Prediction

The results are in Table 1. The rooted mean squared errors for neurotic, agreeable and conscientious are all lower than baseline, others are the same.



**Figure 10. RMSE of Personality Big Five Prediction**

## Title and Authors

.[[1]](#footnote-1)

## First Page Copyright Notice

.

## Subsequent Pages

.

Table 1. Table captions should be placed above the table

|  |  |  |  |
| --- | --- | --- | --- |
| **Graphics** | **Top** | **In-between** | **Bottom** |
| Tables | End | Last | First |
| Figures | Good | Similar | Very well |

## References and Citations

## Page Numbering, Headers and Footers

# CONCLUSION

.

# FUTURE WORK

As we mentioned in early section, due to the work flow and time limitation. We do have some thoughts about improve our accuracy on each part. We will take them as our future work.

## Gender

Will try combination for model oxford with pure image, which only trained by the no face data from 9500.

## Age

Will try combination for oxford image model with text model only for no face users.

## Personality

Will try different combination for each model’s parameter and combine two types of models to see if this improves behavior.

# ACKNOWLEDGMENTS

Our thanks to Professor Ankur and Professor Ying for helping us with all suggestions and guidance.

# REFERENCES

1. Paul Viola and Michael Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features” Accepted Conference on Computer vision and pattern recognition, 2001.
2. <http://napitupulu-jon.appspot.com/posts/pca-ud120.html>,
3. <http://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues>
4. Naive Bayes Classifier algorithms: http://scikit-learn.org/stable/modules/naive\_bayes.html
5. Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.
6. Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.
7. Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.
8. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= <http://doi.acm.org/10.1145/964696.964697>.
9. Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= <http://dx.doi.org/10.1016/j.jss.2005.05.030>.
10. Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= <http://doi.acm.org/10.1145/90417.90738>.

1. [↑](#footnote-ref-1)