BizPro: Extracting and Categorizing Business Intelligence Factors from News

Wingyan Chung, Ph.D.
Institute for Simulation and Training
wchung@ucf.edu

Definitions and Research Highlights

- BI Factor: qualitative evidence that influences market reactions on a company
  - Represented as a textual sentence extracted from online news
- BizPro: An intelligent system that we developed to extract and categorize BI factors from textual news
- We examined the system's capability in profiling four IT companies' BI factors.
  - Statistical Classification
    - Naïve Bayes
    - Logistic Regression
    - SVM
    - OneR
- We discuss implications for BI analysis and data mining

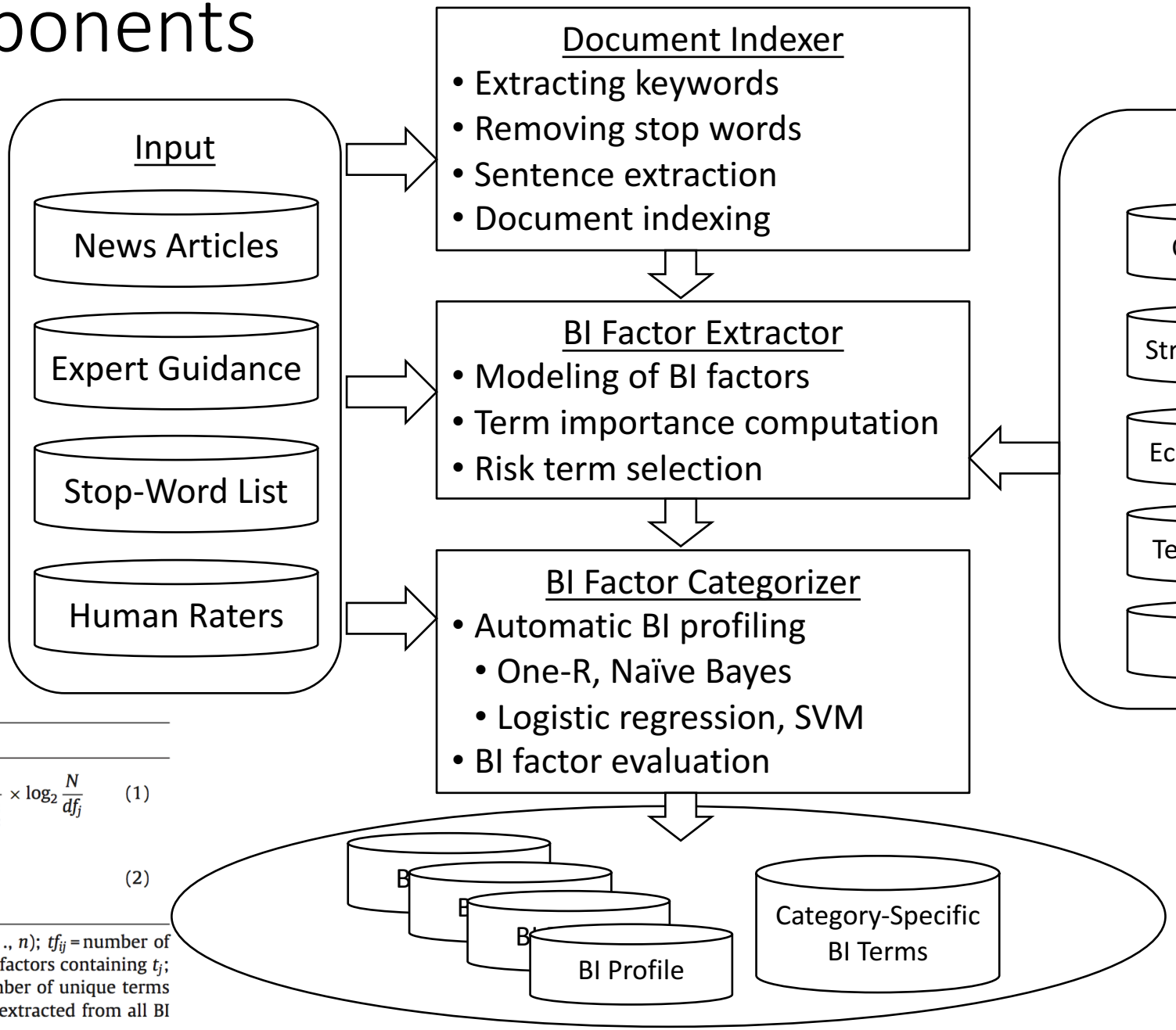
Related Work

- Qualitative information was found useful in improving understanding of accounting information
- Various textual clues have been studied for market prediction and fraud detection
  - Stock price prediction, market sentiment analysis, earning announcement analysis, news content analysis
  - Financial fraud detection, bankruptcy prediction, deception detection
- Text categorization and feature selection
  - Data mining (NB, LR, SVM, NN, etc.), computational linguistics (NMF, LDA, etc.)
  - Feature combination and heuristics for feature selection
- *Scarce work is found on analyzing company BI factors; the approaches may not consider different BI classes in textual documents*

Proposed System: *BizPro*

- *BizPro*: an intelligent system that automatically extracts and categorizes company BI factors from news articles
  - Input: BI category profile; training documents pre-categorized into BI categories; testing documents (document = key sentence of company news)
  - Output: Categorization of testing documents into BI categories
- *BizPro* incorporates domain knowledge of BI analysis, BI factor identification heuristics, and BI categories developed based on literature review and expert guidance.

System Components of *BizPro*



Name	Formula	
Importance of Term j in BI factor i	$d_{ij} = \frac{tf_{ij}^2}{\sum_{k=1}^{n_i} tf_{ik}^2} \times \log_2 \frac{N}{df_j}$	(1)
Importance of Term j in all the BI factors	$d_{t_j} = \sum_{i=1}^{N} d_{ij}$	(2)

t_j = a term extracted from the collection of BI factors ($j = 1, \ldots, n$); tf_{ij} = number of occurrences of t_j in BI factor i ($i = 1, \ldots, N$); df_j = number of BI factors containing t_j; N = total number of BI factors in the collection; n_i = total number of unique terms extracted from BI factor i; n = total number of unique terms extracted from all BI factor.

Research Questions

- How can an intelligent system be developed to model and extract different classes of BI factors from company news articles?
- What is the performance of our proposed system, *BizPro*, in comparison with a benchmark algorithm in BI factor categorization?
- When used in categorizing company-specific BI factors, what are the performances of the automatic categorization techniques used in *BizPro*?
- How does *BizPro*'s use of textual features extracted from company-specific news articles contribute to predicting risk categories?

BI Categories and Expert Verification

- Developed based on literature research, industry standards, and government sources
  1. Operations management
  2. Economic / environment factor
  3. Strategic management
  4. Technological change
  5. Legal issue
- Expert verification by a senior VP (with a PhD in business and 7+ yrs experience) in risk mgmt of a brokerage firm
  - Examined the BI categorization and provided comments that were used in developing systems requirements and functionality

Case Study: Profiling BI Factors of IT Companies from News

- Dataset
  - 6859 sentences extracted from 231 news articles about 4 technology companies

Table 2
Summary of data used in the case study.

Company	Apple	Google	Microsoft	Intel
Number of articles	101	75	18	37
Number of sentences extracted from all articles	2750	2602	503	1004
Number of sentences sampled and tagged by independent human raters	500	373	92	185
Total number of sentences having same BI category agreed by raters	312	273	76	142
Agreement level	62.4%	73.19%	82.61%	76.8%
Cohen's kappa	43.01%	58.01%	72.08%	68.76%
Z value	13.391	14.8879	8.8879	15.9489
p-Value	2.2e−16	2.2e−16	2.2e−16	2.2e−16

Sample News Headlines

Titles of selected news articles used in this study.

Apple Inc.	Google Inc.	Intel Corp.	Microsoft Corp.
Brace yourselves, new Macs are on the way	*Why try browsing with Google Chrome?*	*Intel Labs Aims to Reinvent How People Experience Computing*	*Bach departs amid Microsoft shake-up*
With iPad, Apple aims for sweet spot	*In Europe, Challenges for Google*	*Intel posts biggest quarterly profit in a decade*	*Microsoft succeeds in claim over Android*
Four things we learned from iPhone 4 'Antennagate'	*Google hopes to retain business unit in China*	*Intel faced hacker attack same time as Google*	*Microsoft Kin Discontinued in 48 Days*
Unhappy iPhone 4 users file class action in Maryland	*Hungry for new content, Google tries to grow its own in Africa*	*Intel updates low-power chip line*	*Microsoft and Apple set to roll out new phones*
After prevaricating, Apple releases Safari 5	*Stores see Google as Ally in E-Book market*	*Intel to invest $190 million in Mexican plant*	*After fumble, Microsoft redoes phone software*
Cybercrooks take shine to Apple lineup	*France calls Google a Monopoly*	*Intel touts 50 Gbps silicon optics*	*For Yahoo, Microsoft search is on*
	Google denies fault in lawsuit over Web ads		

Top Keywords

Examples of category terms (only top 20 shown in each category).

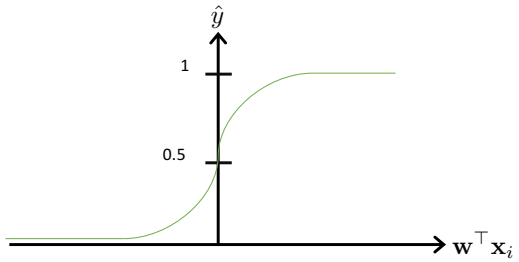
Operation Mgmt	Econ./Env. Ftr.	Strategic Mgmt.	Technological change	Legal issue
apple	cent	google	getting	intel
iphone	intel	apple	computing	apple
google	market	company	yesterday	data
people	apple	microsoft	desktop	oracle
company	company	china	laptop	ftc
phone	shares	deal	continue	google
mr	sales	intel	market	users
search	world	companies	computers	skyhook
users	friday	market	mr	antitrust
app	mobile	search	interview	settlement
billion	dropped	android	onstage	company
software	google	mobile	morph	software
years	ratings	yahoo	short	court
year	relative	internet	falls	include
intel	buy	technology	areas	decide
devices	strength	biggest	standards	regulators
problem	trading	software	web	case
tv	quarter	year	flash	settles
phones	industry	content	open	denied
microsoft	estimates	iphone	pc	sun

Text Classifiers

- Naïve Bayes
  - Finding the best class for a given input vector **x** by maximizing the product of all feature's probabilities of belonging to the class

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

- Logistic Regression
  - Predict y = 1 if $\hat{y} \geq 0.5$
  - Predict y = 0 if $\hat{y} < 0.5$



$$\hat{y} = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}} \Rightarrow \ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = \mathbf{w}^\top \mathbf{x}_i$$

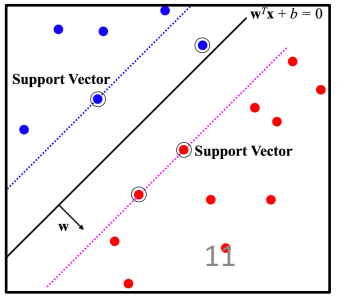
- Support Vector Machine
  - Finding best weights to form a margin to classify data points

$$\min_{\mathbf{w} \in \mathbb{R}^d, \xi_i \in \mathbb{R}^+} \left\|w^2\right\| + C \sum_{i}^{N} \xi_i$$

$$\text{subject to}$$

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \text{ for } i = 1 \ldots N$$

- One-R (benchmark)
  - Choose the variable with the smallest predictive error to form one rule for classification

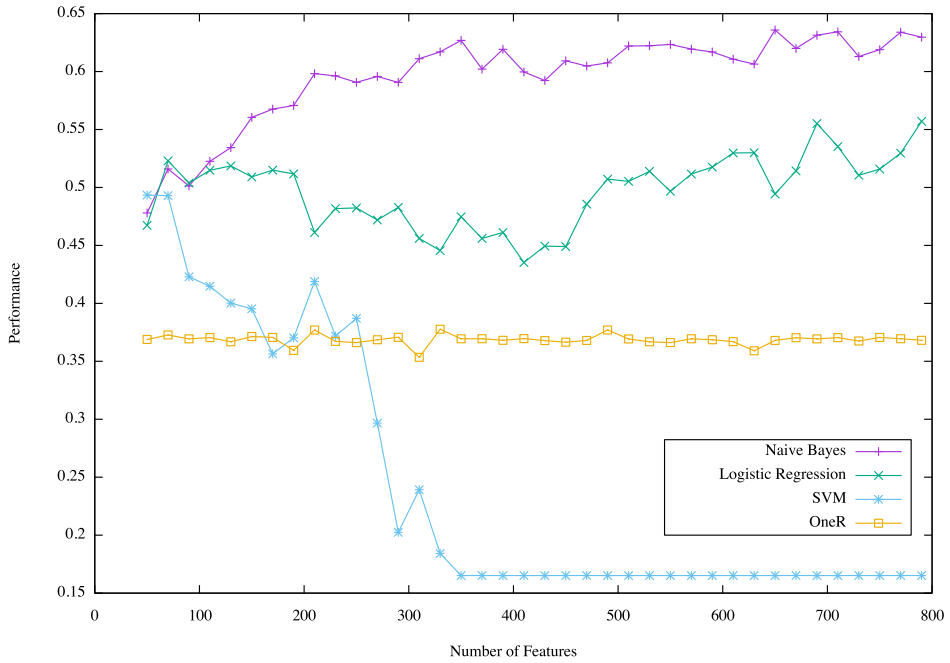


Performance Metrics

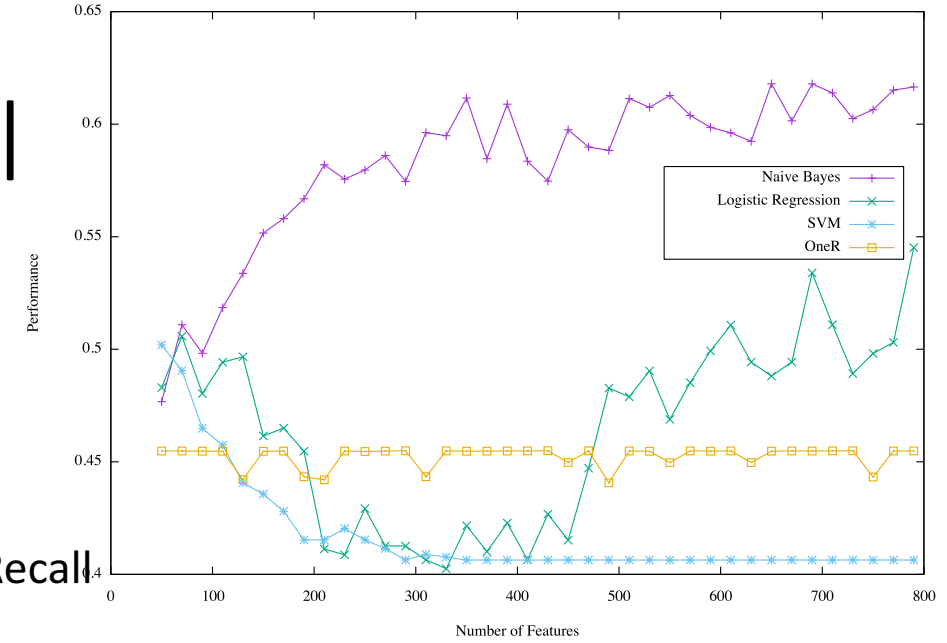
- We used ten-fold cross validation to evaluate the techniques' performances
- We tested performance of classifying the first 5 statements in an article by the first *m* top-ranked terms, where *m* increases from 50 to 790 with 20 as each increment

Metric	
$\text{Weighted precision} = \sum_{i=1}^{m} \left(w_i \times \frac{\text{number of correctly classified sentences in Category } i}{\text{number of sentences classified as Category } i \text{ by the algorithm}} \right)$	(4)
$\text{Weighted recall} = \sum_{i=1}^{m} \left(w_i \times \frac{\text{number of correctly classified sentences in Category } i}{\text{number of all sentences classified as Category } i \text{ by the gold standard}} \right)$	(5)
$\text{Weighted } F\text{-measure} = \sum_{i=1}^{m} \left(w_i \times \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \right)$	(6)
Area under ROC curve: measures the impact of changes in the probability threshold, which is the decision point used by the categorization model ($> 50\%$ for accurate prediction of a class).	(7)

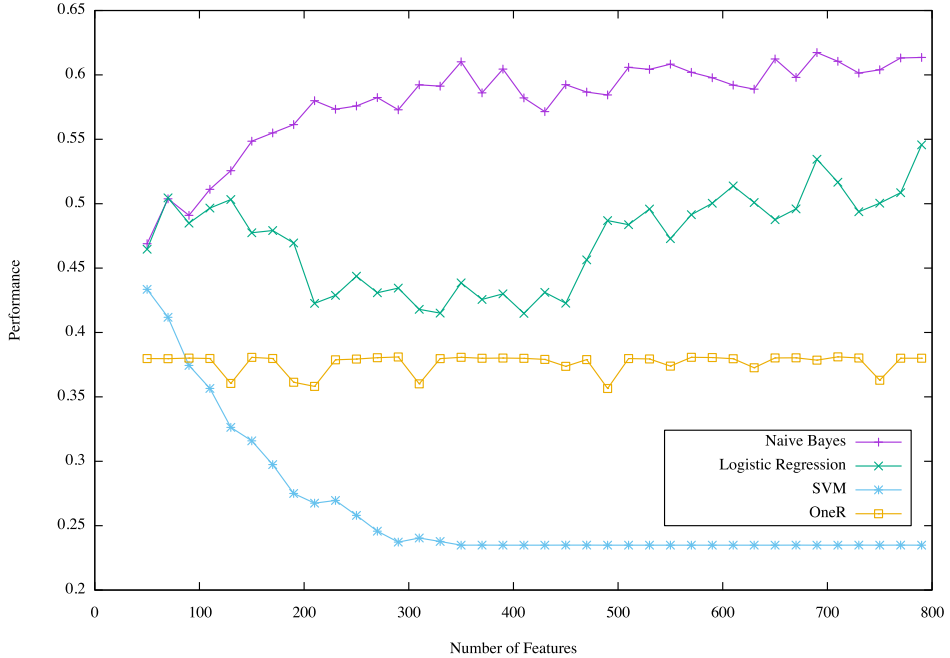
Experimental Results



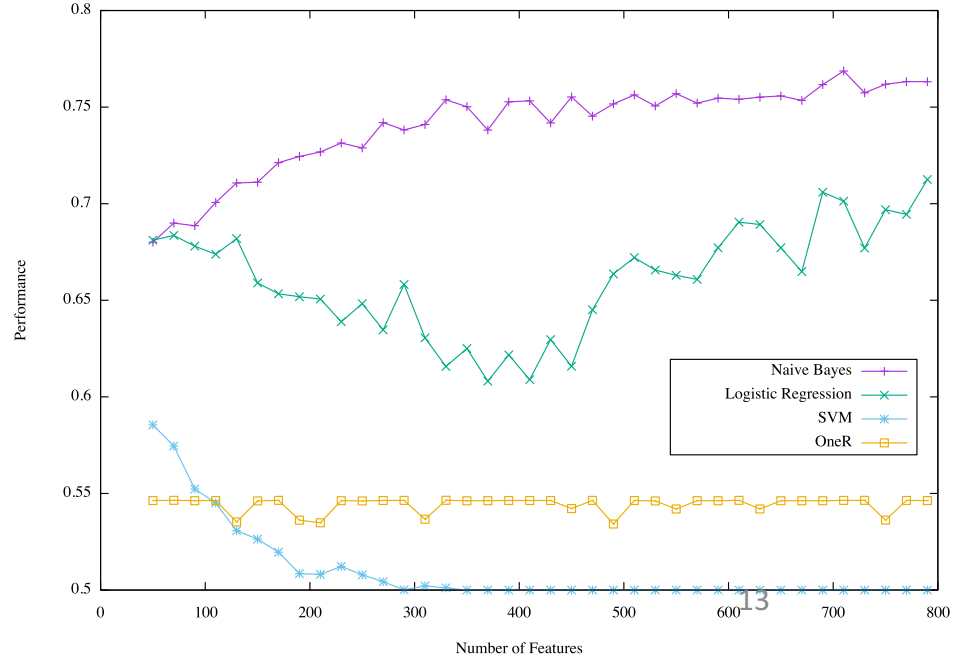
Precision



Recall



F-measure



AUC

Comparison

- Naïve Bayes demonstrates the highest classification performance in terms of precision, recall, and F-measure.
- NB outperformed SVM and LR in P/R/F, while LR outperformed SVM in P/R/F.
- LR's performance fluctuates highly with the number of features.
  - LR may produce higher variance than Naïve Bayes when training size is limited, making it less stable in performance.
- SVM's performance is adversely affected by the increasing number of features.
  - Increasing SVM cost parameter values dramatically increases its performance

Discussion and Implication

- The encouraging results indicate a high potential to use the proposed framework to supplement traditional risk assessment methods.
- Managers can use the techniques to identify and categorize BI factors from news articles.
- Provides strong implication for intelligence agencies and organizations that need to handle risk factors from online textual data.

Conclusion

- This research addressed the need for BI profiling and categorization from news articles
- An intelligent system call *BizPro* was developed to profile and categorize BI factors from online news articles
- A study of four technology companies show that *BizPro* outperformed benchmark technique in P, R, F, and AUC.
  - Feature selection, BI factor classification, and document indexing
- New application of NB, LR algorithms and their performance
- New insights on SVM algorithm tuning

Research Opportunities

- Cyber security and privacy
  - Online risk assessment
  - Prediction of threat
  - Modeling human behavior in online transaction
- Cyber Intelligence Lab at IST
  - http://cyber.ist.ucf.edu/
- Text classification and clustering
  - Feature selection
  - Simulation of data and activities
  - Development of classification techniques
- Funding available: NSF, DoD, DHS, DARPA, …, etc.
- PhD in Modeling and Simulation

Thank you!

Wingyan Chung, Ph.D.
Institute for Simulation and Training
Email: wchung@ucf.edu; Tel.: (407) 882-1435