MACHINE LEARNING FOR I-O 2.0

MQ Liu
Amazon

DISCLAIMER: Views and opinions do not represent Amazon's policy or positions.

SYMPOSIUM OVERVIEW

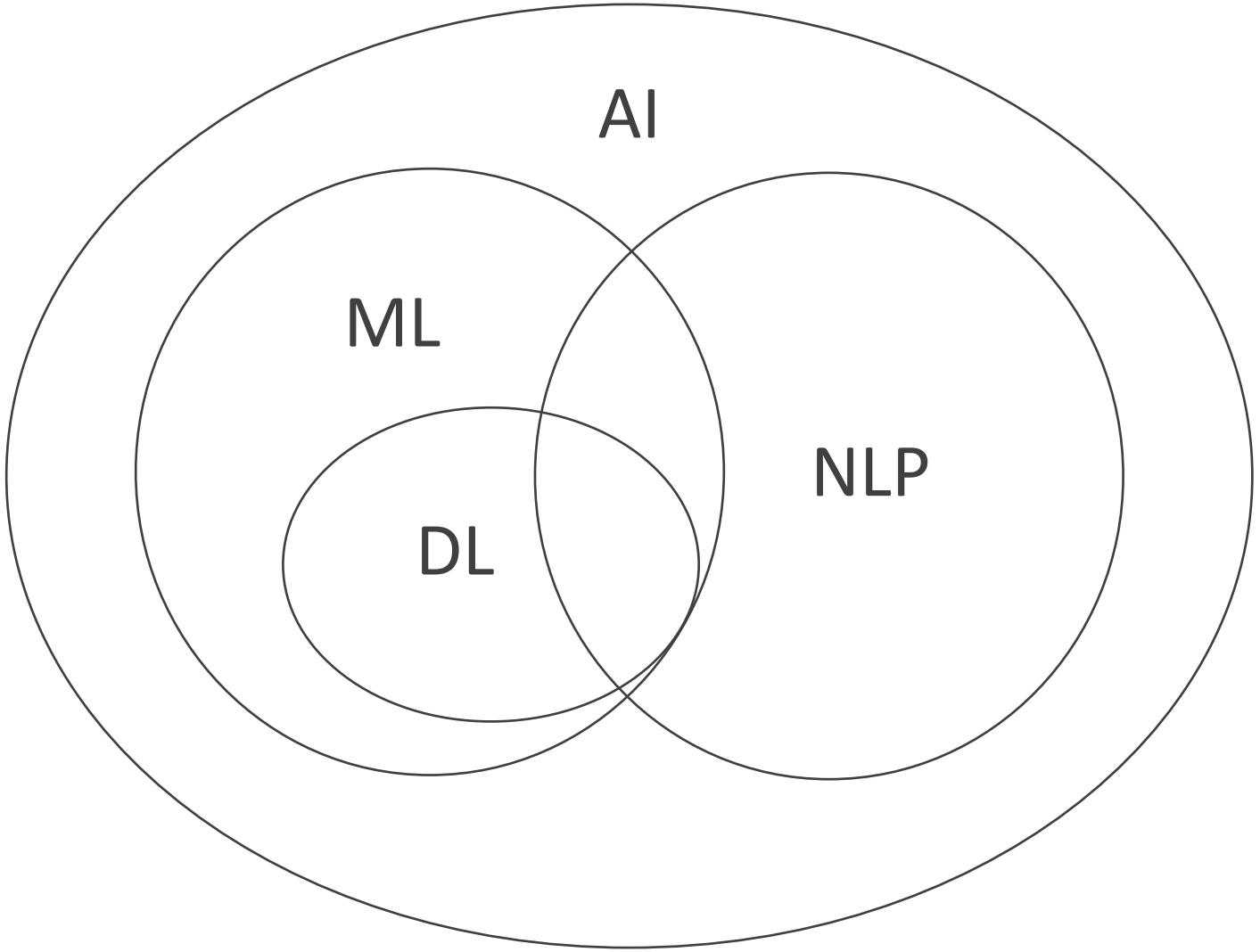
WHY?

- The increasing popularity of ML in I-O calls for a need to better understand its techniques and best practices.
- Literature regarding ML techniques and applications in I-O is still lacking.

Goal

Presents various ML techniques applicable to I-O and demonstrate how they can be used to address issues that span the employee life cycle.

DEFINITIONS

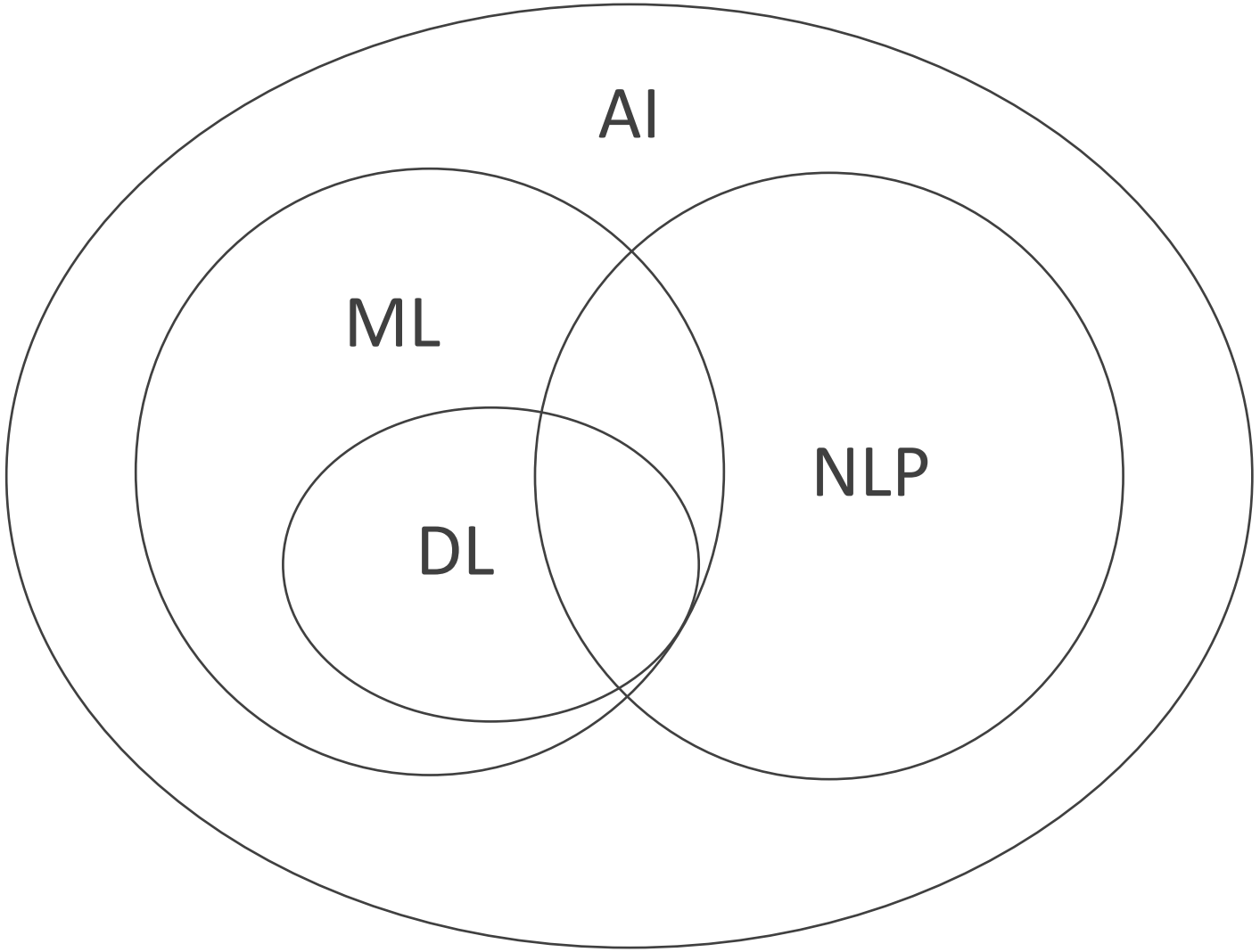


Artificial Intelligence (AI)

"…the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence…"

(John McCarthy, Father of AI)

DEFINITIONS

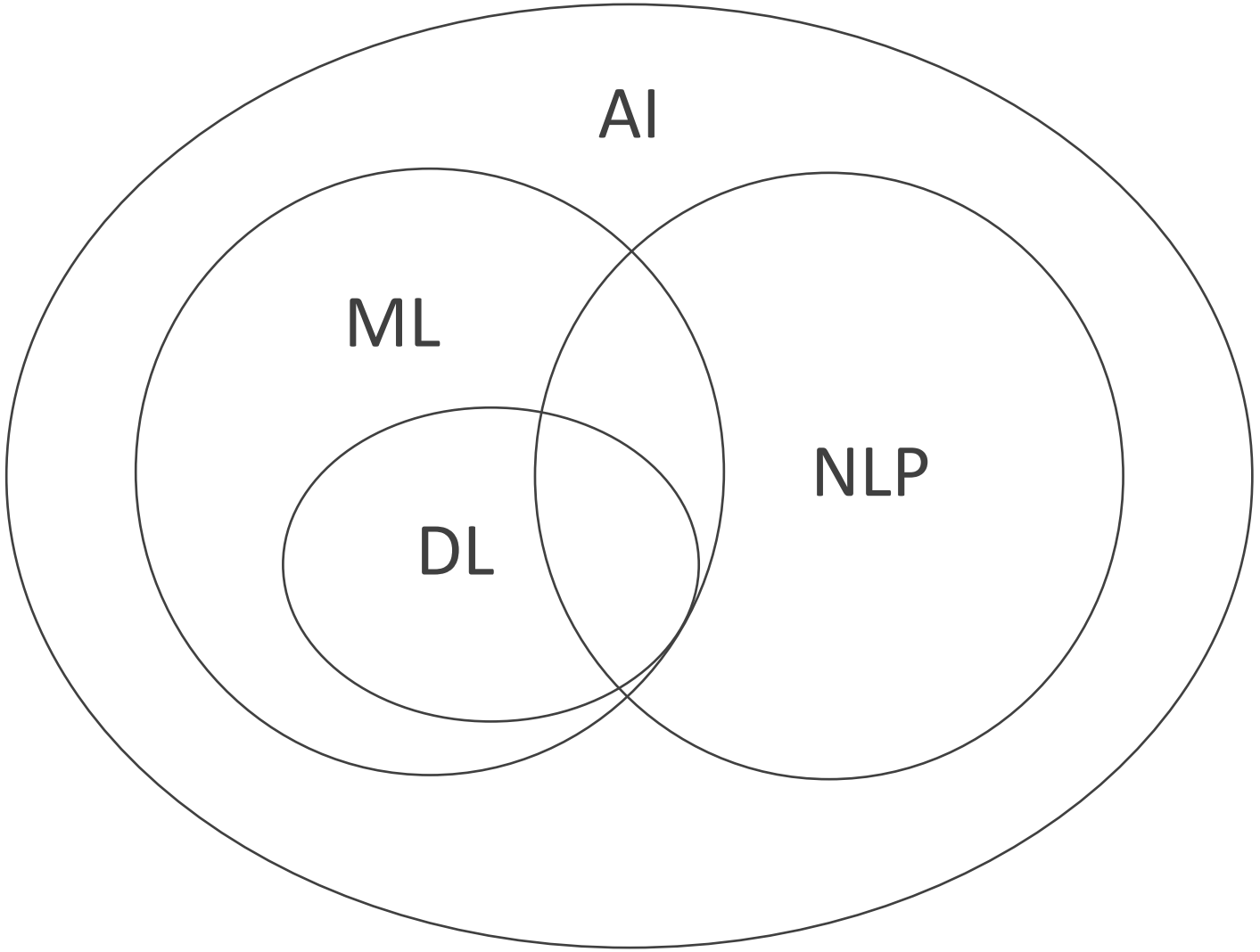


Machine Learning (ML)

A subfield of computer science that aims to construct computer programs that can learn and improve with experience automatically.

(Mitchell, 1997)

DEFINITIONS

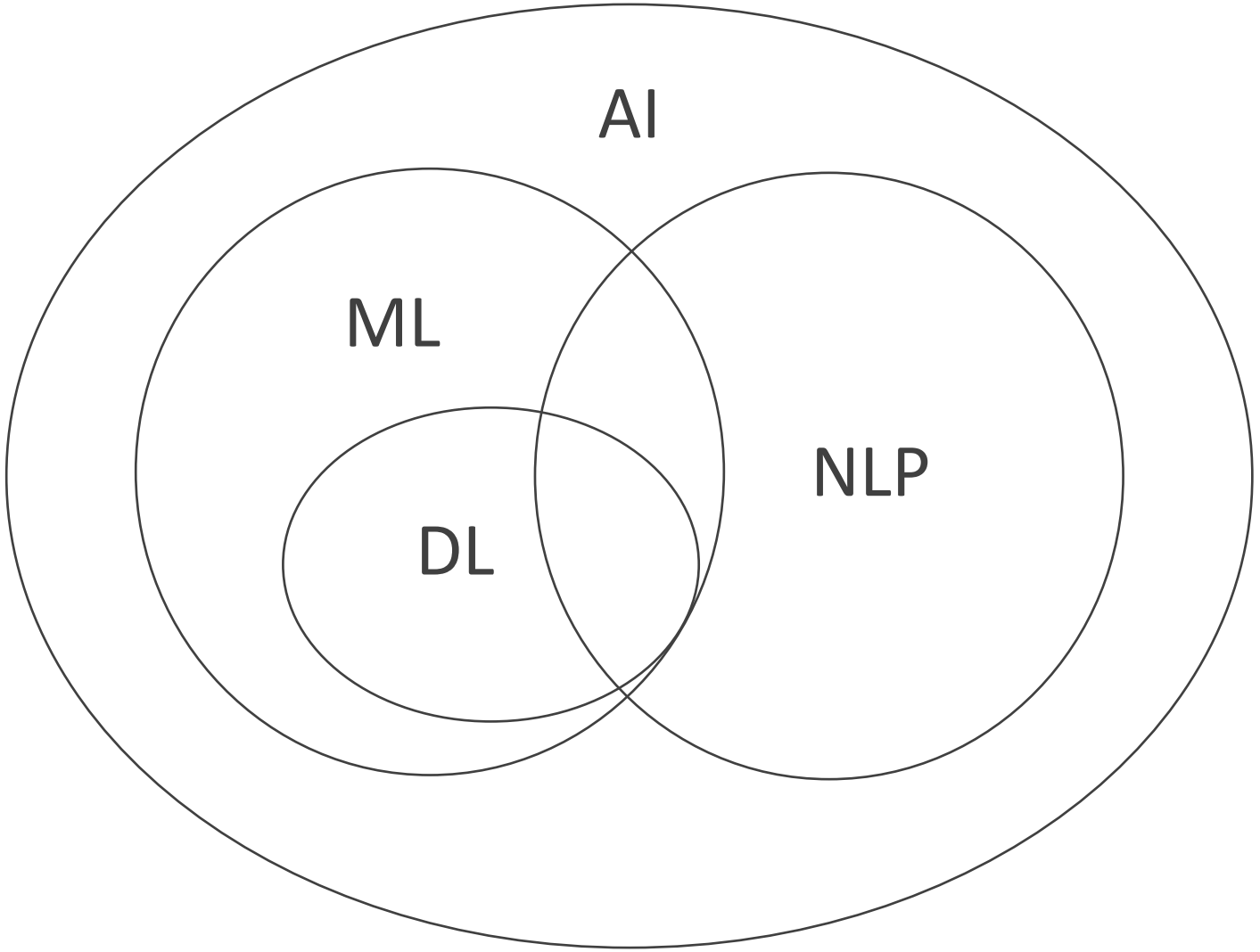


Deep Learning (DL)

A subfield of ML that focuses on “computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction”

(LeCun, Bengio, & Hinton, 2015)

DEFINITIONS



Natural Language Processing (NLP)

A discipline that aims to program computers that can automatically process and learn human natural language data

(Manning & Schütze, 1999)

PAPER OVERVIEW

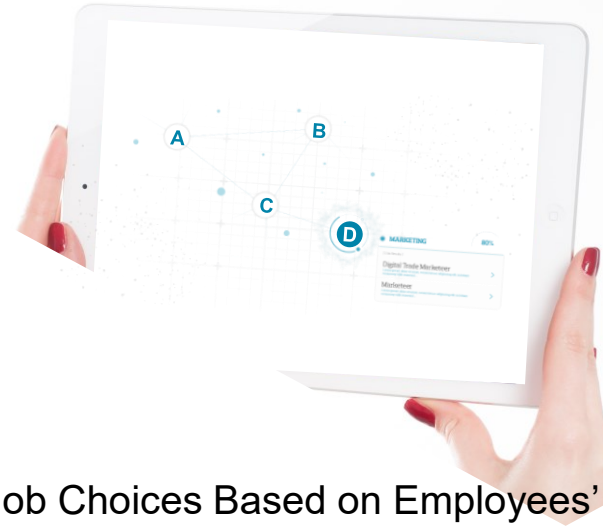
Authors	Paper
Guan, Gaertner, & Garner	An application of DL on job movement and recommendations
Hernandez, Sanders, Kim, & Towe	Using DL to infer personality traits and cognitive ability from résumé style
Hickman, Bosch, Tay, Ng, Saef, & Woo	Applying ML models in a multimodal system to predict personality from video interviews
McCune, Lewris, & Westerhoff	Using state-of-the-art NLP techniques to derive meaning from employee survey comments

Dr. Fred Oswald, Discussant

Professor, Herbert S. Autrey Chair in Social Sciences, Director of Graduate Studies @ Rice University

- ✓Extensive research and teaching on workforce readiness and quantitative methodology (including ML) for over 20 years.
- ✓Numerous peer-reviewed journal publications and book chapters, with 10k+ citation count on Google Scholar.
- ✓A Fellow and past president of SIOP.
- ✓An Associate Editor for 3 journals and on the editorial boards for 10 journals.

RICE



Identifying Alternative Job Choices Based on Employees' Job Profiles

Ada Guan PhD, Senior Data Scientist, Human Capital Solutions
Stefan Gaertner PhD, Partner, Human Capital Solutions
Amy Garner, Consultant, Human Capital Solutions

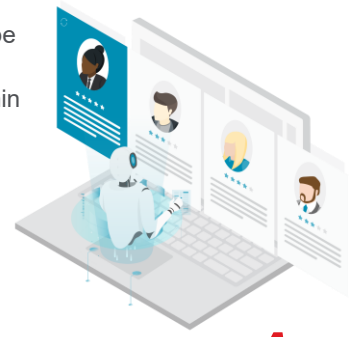
AON
Empower Results®

Future of Work is NOW

In response to skill shortage, automation, and a rapidly changing job market, organizations need to understand their employees' full potential and better leverage their capabilities.

Employers should be able to:

1. Leverage internal resources
2. Identify right developmental areas for their employees so that they can be trained appropriately
3. Improve employee work satisfaction by facilitating lateral movement within an organization



AON
Empower Results®

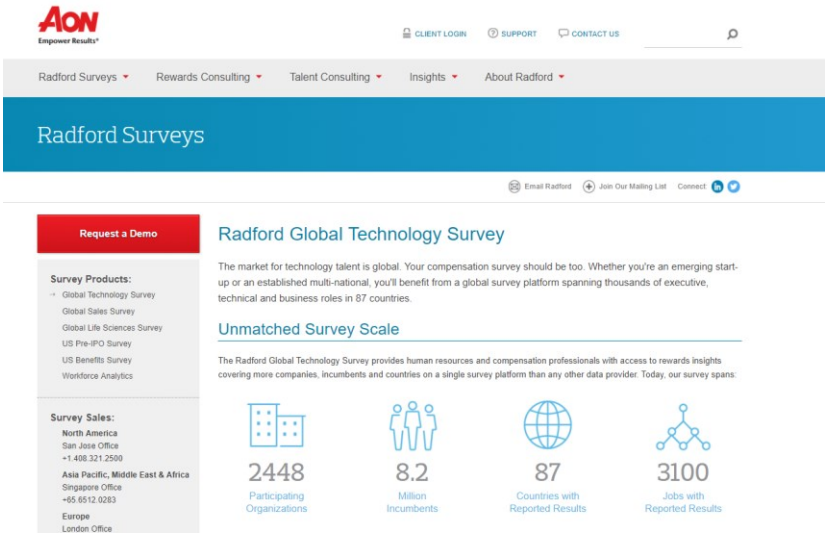
It is critical to identify potential job choices based on employees' profile..

But, how to find out your employees' future jobs based on what we know?

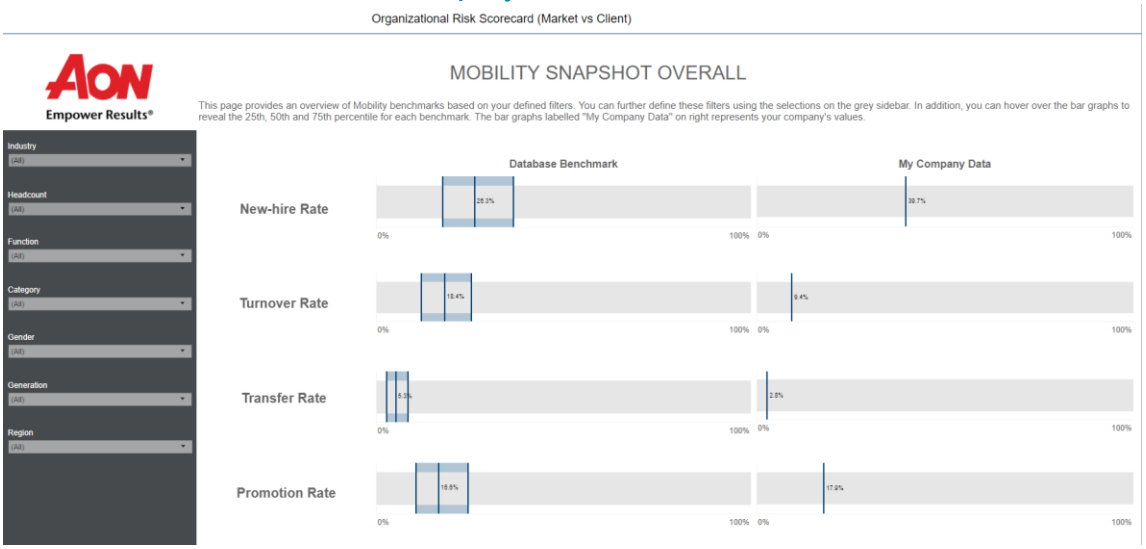
AON
Empower Results®

3

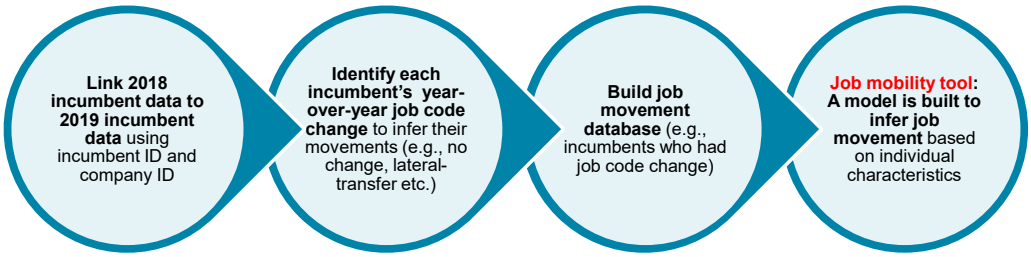
Database Used to Predict Employees' Future Jobs



Database Used to Predict Employees' Future Jobs



Modeling Process & Model Information



Job mobility database:

- +200k incumbents who had job code change over the past year

Machine learning algorithm (Random Forest):

- Movement across 300 job families from Support, Professional & Management levels
- Prediction accuracy: ~75%

AON
Empower Results®

Job Prediction Examples

Tool input

- Individual information: current job family, grade level, function, experience, tenure, & work location
- Company information: industry & company headcount

Tool output

- Provide 1-5 job suggestions of each individual

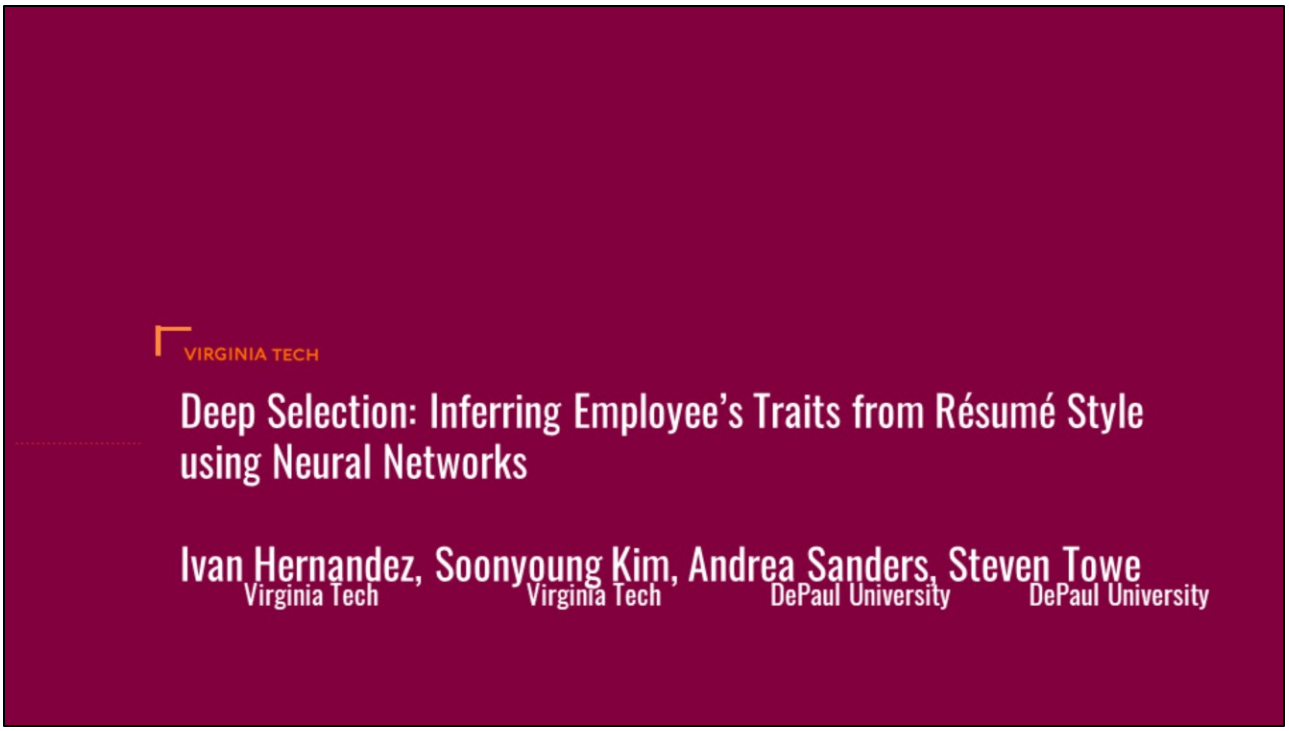
Incumbent ID	Current Job Family	Cat Level	Function	Experience	Tenure	Region	Industry	Company HC	Predicted Movement
1	Software Engineer(Apps)	P1	Professional Services Consulting/Outsourcing Services	Missing	Missing	Europe	Financial Technology	4000-5000	Game Producer, Configuration/Release Engineer, UI/HumanFactors Engineer, Software Engineer (Sys)
2	Administrative Assistant	S5	Finance and Administration	Experienced	1 -< 3 years	U.S.: West (CA Only)	Software Products/Services	10000-50000	Project (Design) Manager, Executive Assistant
3	Project (Design) Manager	P5	Professional Services Consulting/Outsourcing Services	Experienced	Missing	U.S.: West (CA Only)	Semiconductor Components	>=100000	Systems Design/Architecture Engineer, Development Engineering Mgmt, Development Engineer, Project/Program(Admin) Mgmt
4	Semiconductor Assembler	S3	Operations/Manufacturing	Experienced	Missing	Asia Pacific	Semiconductor Components	10000-50000	QC Inspector, Hardware Development Engineer
5	Sales Acct Manager-Direct-Existing Accts	P1	Sales	Some Experience	1 -< 3 years	Asia Pacific	Other Technology	50000-100000	Sales Acct Manager-Product Specialist/Overlay, InsideSales Representative -Own Quota, SalesAcctManager-OEM/VAR

AON
Empower Results®

For questions – please contact:

Ada Guan, PhD
Senior Data Scientist, Human Capital Solutions
li.guan@aon.com

Thank you!

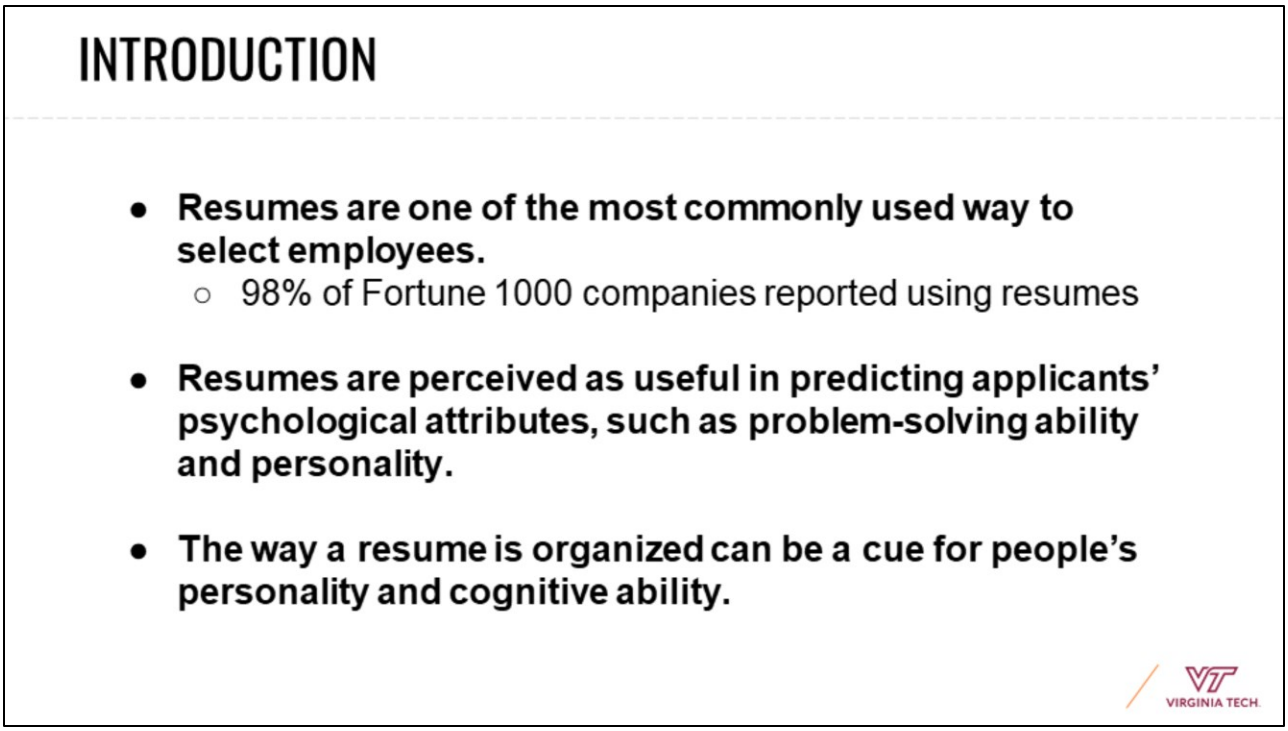


Assessing cognitive ability and personality traits facilitate predicting future job performance (Schmidt & Hunter, 1998)

Administering these an entire pool of applicants is expensive and time consuming, however.

Instead, most organizations use résumés as an initial screening item.

The current project describes how additional information about an applicant's cognitive ability and personality can be inferred from résumés. Specifically, we apply deep learning techniques to learn distinctive visual features from a résumé, and use those features as predictors in a machine learning model.



Résumés are an important part of the selection process.

In a survey of 150 Fortune 1000 companies surveyed, 98% reported using résumés as a selection technique (Piotrowski & Armstrong, 2006). No other selection method was used more widely.

Because résumés are request early in the selection process their evaluation has long-term consequences for applicants.

Recruiters treat résumés as containing informative biodata, and make inferences about the applicant's traits such as leadership, motivation, intelligence, and interpersonal skills. The more a résumé contains biodata reflecting attribute requirements of the jobs the more attractive recruiters evaluate the applicant (Brown & Campion, 1994).

The ubiquity of résumés means that drawing additional valid inferences about people's cognitive and personality traits have implications across a variety of organizations.

Although subjective interpretations of résumé biodata are prone to recruiter biases, résumé biodata may be able to help indicate an applicant's individual traits without the administration of lengthy tests. Some promising relationships between GMA, personality traits, and specific elements of résumé content have already been

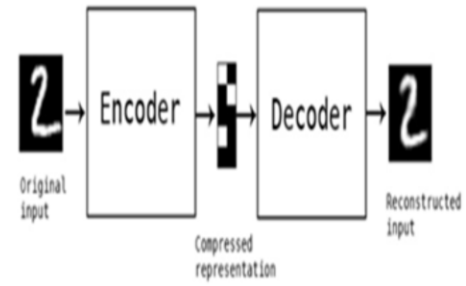
identified (Burns, Christiansen, Morris, Periard, & Coaster, 2014; Cole, Feild, & Giles, 2003).

An avenue we seek to explore is examining the visual arrangement of résumés as an additional cue to the person's traits. Like a blank canvas, résumés provide a space for applicants to portray a collection of information however they desire. Research examining people's living and work spaces finds that the arrangement of these open environments describes people's tendencies and preferences (Gosling, 2018; MacKinnon, 1977).

PROPOSED APPROACH

- We can apply **three methods** below as computational image analysis to quantify the visual features in resumes.

1. **Convolutional Neural Networks** can deconstruct an image into its basic features by examining the image in small "blocks" and encoding the shapes present.
2. **An Autoencoder** is a neural network that tries to reconstruct its input using information from a smaller number of values. This process condenses an image into its core features.
3. **"Machine Learning"** builds a predictive model that finds association between the visual features of a resume and applicants' personality & cognitive ability.



VIRGINIA TECH

Humans have limitations that would limit using human coding for visually inspecting résumés.

One limitation is that artistic judgements are difficult to verbalize (Wilson, Lisle, Schooler, Hodges, Klaaren, & LaFleur, 1993). Research suggests that the forcing verbalization of artistic judgements alters the value and quality of the judgments.

Second, verbalization of an artistic coding scheme relies on features that are explicit. That is, only features and aspects to an image that can be easily verbalized and perceived would be found within a coding scheme.

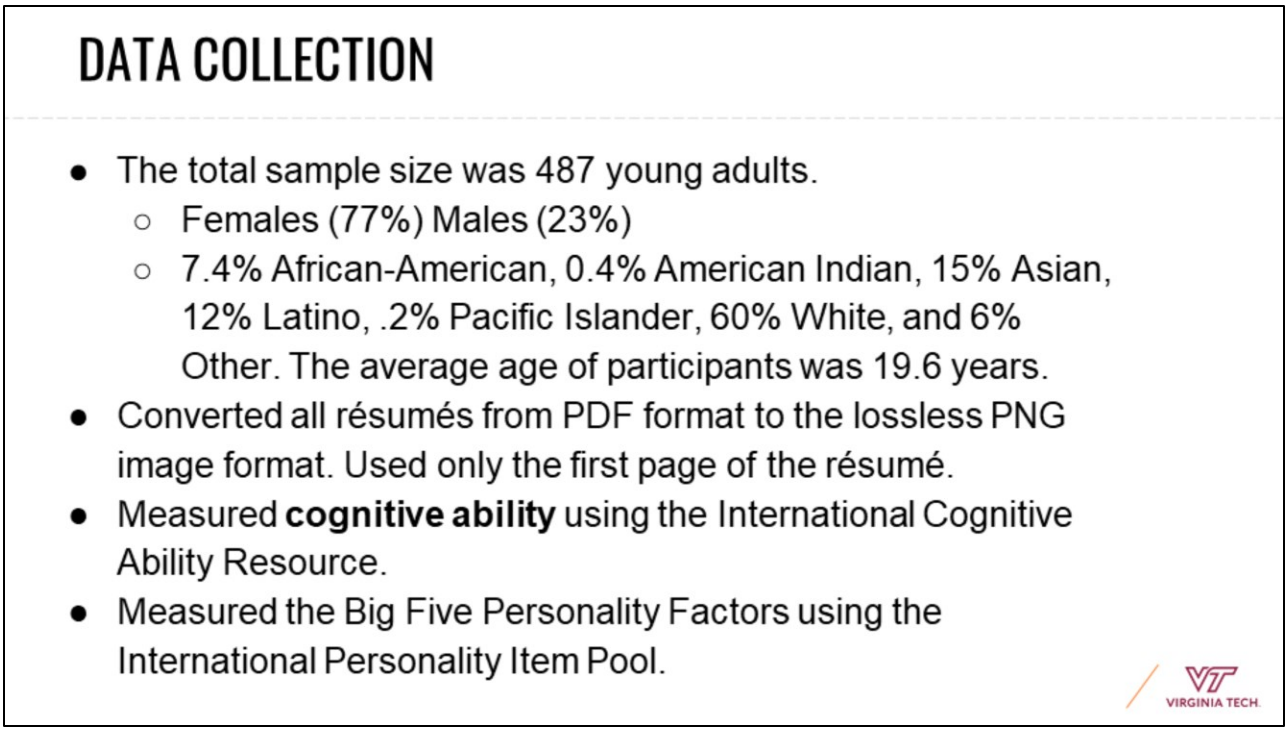
Thus, relying on human intuition inhibits discovering novel features. To address these limitations, we incorporate computational image analysis via convolutional autoencoders to extract common visual features, and use machine learning to relate those visual features to the desired outcomes.

Convolutional Neural Networks are type of machine learning method that can deconstruct an image into its basic features by examining the image in small "blocks" and encoding the shapes present. These basic features represent the presence of angles, borders, lines, etc. They are spatially and rotationally invariant.

These convolutional networks can be incorporated into an autoencoder, which is a type of neural network architecture that accepts and input, has the input go to a

bottle-neck layer, which contains a smaller number of neurons that the input has values, and then tries to reconstruct its input using only the information from the bottle-neck layer. This process condenses an image into its core features, and serves as a form of dimension reduction, with non-linear mappings of features.

After training an autoencoder, the values produced by the bottle-neck layer after being provided an input image serve as a condensed representation of the input. These higher dimensional values are then able to be provided to a machine learning model, which finds association between those core visual features of a resume and applicants' personality & cognitive ability



We collected a sample of 487 young adults from universities in the midwest and mid-atlantic.

Participants completed a survey online for course credit. This survey asked participants to provide their most recent résumé, and to complete personality and cognitive abilities measures.

We measured participants' cognitive ability using the International Cognitive Ability Resource measure (ICAR; Condon & Revelle, 2014). The ICAR is a publically available measure of cognitive ability with four item types: Three-Dimensional Rotations, Letter and Number Series, Matrix Reasoning, and Verbal Reasoning. To ensure that the length of the survey remained manageable for an online study, participants completed 16 items from the ICAR, with 4 items corresponding to each of the higher-level dimensions. Following the recommendations of the scale authors, participants were untimed and scores were summed to create a single cognitive ability composite.

Their personality scores for the Big Five were assessed using the 50-item IPIP (Goldberg et al., 2006). Specifically, participants completed 10 items per personality facet. For each facet, 5 of the items positively loaded on to the higher order factor, and 5 of the items negatively loaded on to the higher order factor. Participants scores on the higher order factor were computed by summing the items within each factor, reverse scoring the negative loading items.

HUMAN ASSESSMENT

- Two raters assessed applicants' personality & cognitive ability using adjective rating scales for each trait.
- Averaged raters' scores to assign each participant a single human evaluation score for each trait.
  (a) neuroticism = envious, discontented, relaxed, stable, and calm
  (b) extraversion = enthusiastic, sociable, energetic, extraverted, and active
  (c) openness to experience = versatile, wild interests, adventurous, creative, and insightful
  (d) agreeableness = kind, cooperative, warm, charming, and unselfish
  (e) conscientiousness = hardworking, organized, thorough, responsible, and systematic
  (f) cognitive ability = bright, intellectual, quick, intelligent, and smart.

VIRGINIA TECH

As a comparison to the predictions made by the machine learning model, two raters (advanced undergraduate students in I-O psychology) assessed applicants' personality using the five adjective trait rating scales developed in the pilot studies. The Cronbach's alpha within each Big Five factor, were above .70, and so the average rating made by the raters within a factor were averaged to compute 5 separate factor scores for each participant.

These factor scores represent the evaluation an applicant would receive from a human evaluator of their résumé. They are useful for comparing to the participant's actual personality and cognitive ability scores, and contrasting that correspondence with how well the machine learning predictions correspond to the actual trait ratings.

FEATURE EXTRACTION AND MODEL TRAINING

- Tried different autoencoder architectures for each personality trait. Trained for a period of ~24 hours.
- Extracted résumé features by passing a résumé through network and measuring the output from the smaller dimensional hidden layer.
- Passed each résumé to the autoencoder and extracted values representing the presence of different visual features.
- Applied a cross-validation procedure: Used 90% of the data to train the model, and 10% of the data to test the accuracy of the predictions - repeated 10 times.

VIRGINIA TECH

We used an autoencoder to receive a resume as an input image (a résumé rescaled to be 200 x 200 pixels). Only the first page of the résumé was used. We trained this model to minimize the difference between the output image reconstructed from a smaller number of nodes and the original input image. This process is iterative, and we trained the model for 24 hours.

After the autoencoder was trained, extracted résumé features by passing a résumé through the autoencoder network and measuring the output from the smaller dimensional hidden layer.

Passed each résumé to the autoencoder and extracted values representing the presence of different visual features.

To obtain the model's predictions for use in subsequent regression analyses, we applied a cross-validation procedure: Used 90% of the data to train and calibrate a random forest model, and then predictions using the trained on the remaining 10% of the data to test the accuracy of the predictions. We repeated this process 10 (10-fold cross-validation) making predictions about an independent set of data not previously used as the outcome.

We combined all of the cross-validated outcome predictions as a single “neural network assessment prediction”

This process was repeated for each personality trait and cognitive ability, providing six separate neural network based assessments.

RESULTS: PREDICTIVE VALIDITY

H1: Correlation between Neural Network Predictions and Trait Values	
Cognitive Ability	r = .20***
Extraversion	r = .11*
Agreeableness	r = .14**
Conscientiousness	r = .19**
Neuroticism	r = .15**
Openness	r = .11*

- Neural Networks can predict personality and cognitive ability better than what would be expected by guessing.

VIRGINIA TECH

We hypothesized that a résumés' visual features extracted from neural network would correlate with people's true trait ratings.

All correlations were statistically significant at the $p < .05$ level

The strongest effects were found for cognitive ability and conscientiousness. These traits are generally the strongest predictors of job performance (though the type of job is important to consider when evaluating predictive validity of traits and job performance).

The traits least predicted by the neural network model were extraversion and openness.

Therefore, there is predictive validity above guessing by using the visual features inferred by the neural networks.

RESULTS: INCREMENTAL VALIDITY

H2: Prediction Improvement with Neural Networks			
Predicted Trait	Human Prediction Only	Human & Neural Network Prediction	Increase in R^2
Cognitive Ability	$R^2 = .093$	$R^2 = .108$	.015*
Extraversion	$R^2 = .037$	$R^2 = .057$	.020***
Agreeableness	$R^2 = .033$	$R^2 = .069$	.036***
Conscientiousness	$R^2 = .040$	$R^2 = .063$	.023***
Neuroticism	$R^2 = .028$	$R^2 = .047$	.019**
Openness	$R^2 = .020$	$R^2 = .082$	.062***

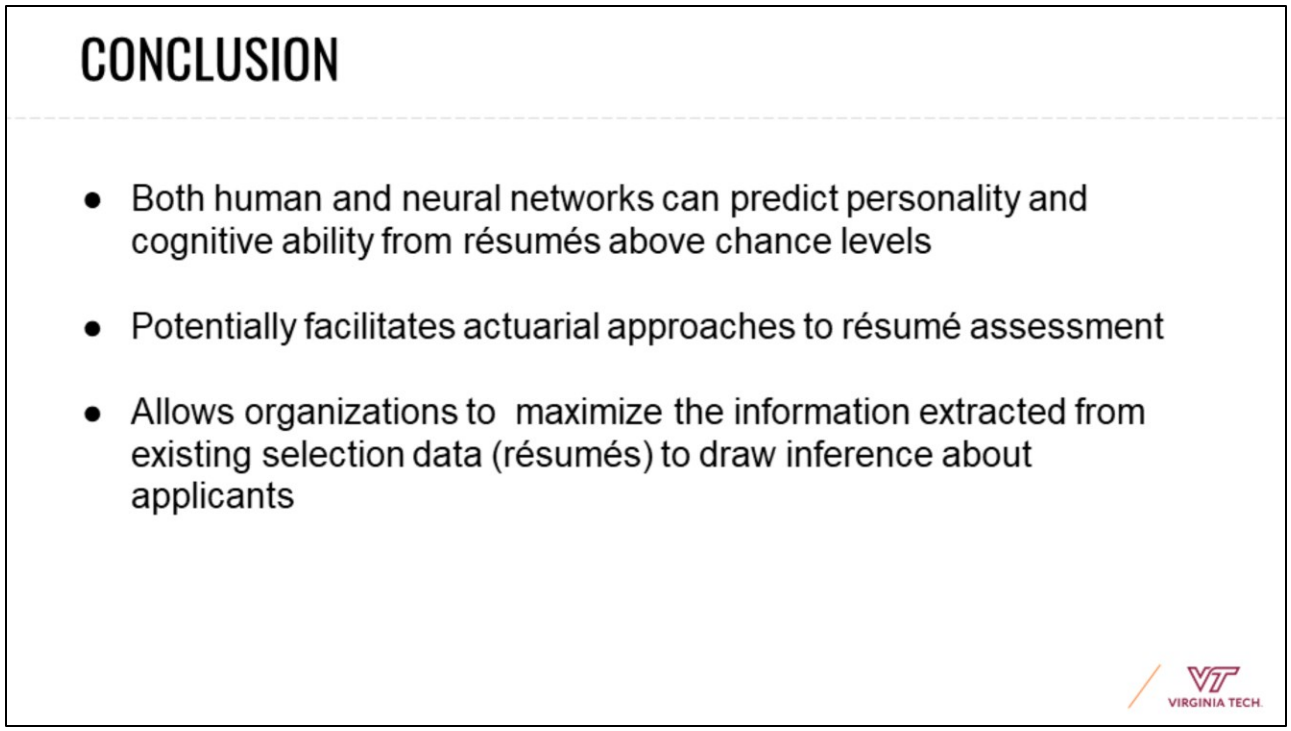
- Neural Networks significantly improved the explanatory ability of the model compared to only using human ratings.

VIRGINIA TECH

We also examined the incremental validity of the neural network features compared to Human assessments of the applicants' traits.

This analysis first ran a regression with the human assessments of the trait as the predictor variable and the trait being assessed as the outcome. We controlled for the amount of text content on the résumé as well as the total darkness of the résumé (how much visual information is present in the resume. After adding the trait prediction from the neural network, all regressions showed in a statistically significant ($p < .05$) change in R^2

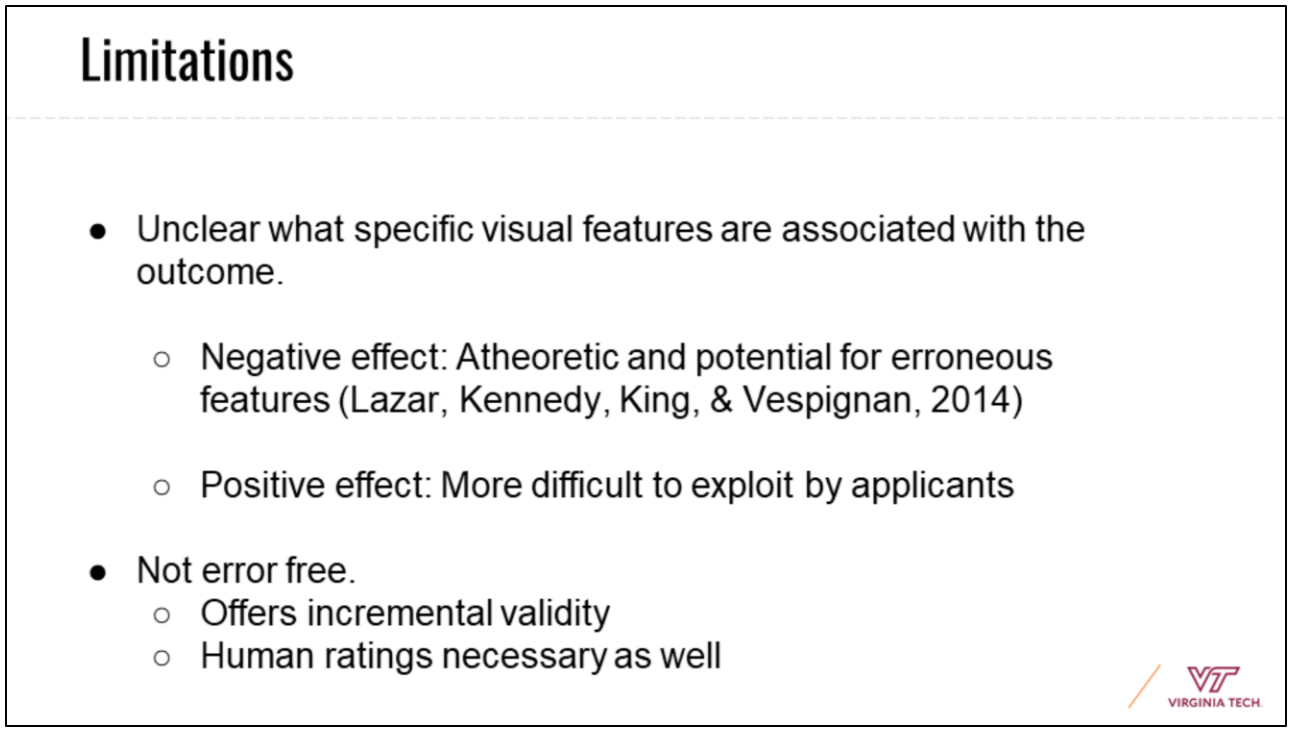
Therefore, although human assessment seem to perform better than the neural network method, including the neural network assessments improve the overall validity of the predictions.



When predicting a person's true personality traits, the baseline model found that human judges could accurately predict a person's true level, beyond chance.

By providing additional validity we hope this research raises the utility of, and consequently the motivation to, adopt more actuarial methods of assessing résumés.

This research would allow companies to maximize existing information from applicants, such as their résumés, to infer desired selection traits. This approach scales to thousands of resumes easily and works with commonly available selection data.



While using deep neural networks to analyze résumés offers many benefits, there are limitations to the method. One limitation inherent to deep neural networks is the black-box nature of the neural network. Because deep neural networks are a collection of numeric weights and non-linear mathematical transformations, they do not provide a clear theoretical understanding of the intermediate process between the input and output.

This limitation is also a strength because it minimizes exploitation by applicants. Most research examining the relationship between résumé content and employee traits uses simple linear models and describes the ideal content. Because the current method does not directly indicate what a desirable résumé resembles, neither the applicant nor the employer are aware of how to game the system.

Another limitation is that the method is not error free. The association between predictions and outcomes is not perfect and the method does not supersede human judgments. Using the method does not supplant human ratings and serves as a complement to further reduce the unexplained error variance of existing methods.

Questions

Ivanhernandez@vt.edu

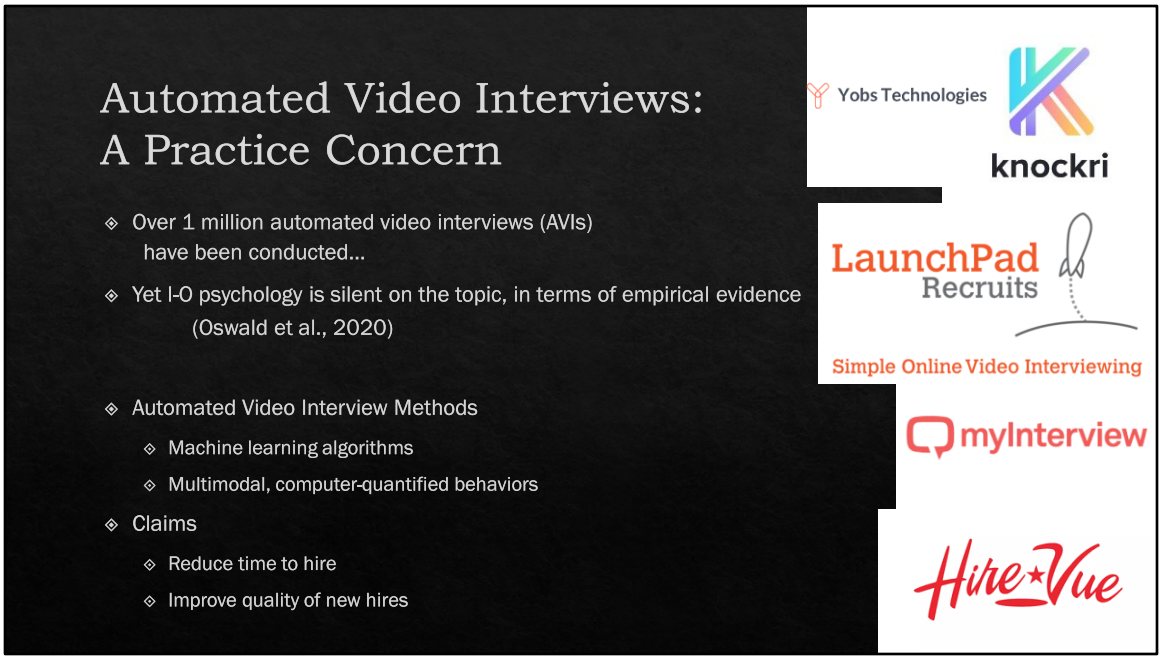
VIRGINIA TECH

Investigating Emotion Analytics for Predicting Personality in Video Interviews

Louis Hickman[1], Nigel Bosch[2], Louis Tay[1], Vincent Ng[3], Rachel Saef[4], and Sang Eun Woo[1]

[1]Purdue University, [2]University of Illinois, [3]University of Houston, [4]Northern Illinois University

Thank you for your interest in our research. Our team is comprised primarily of students, past students, and professors at Purdue University, but we are also grateful to have Nigel Bosch on the team. Nigel is a computer scientist who is an Assistant Professor in the School of Information Sciences at the University of Illinois.

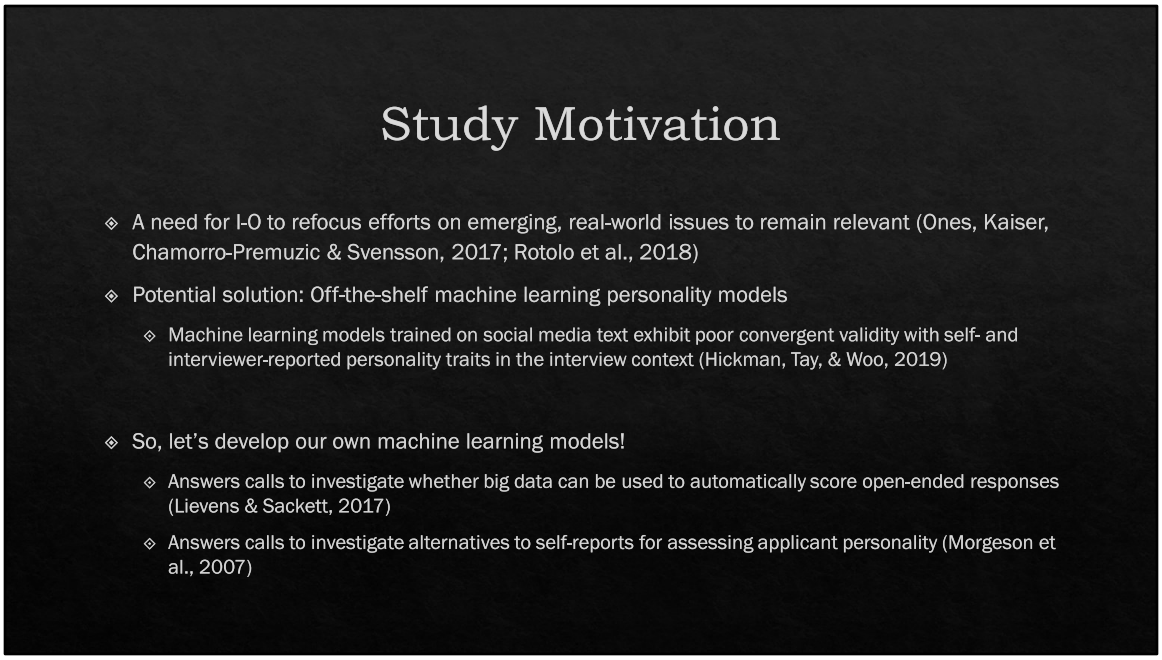


Automated video interviews are, inherently, a practice concern. Automated video interviews are being adopted by many organizations, and HireVue claimed that by mid-2019, they had already conducted over 1 million automated video interviews. At least 6 vendors are marketing automated video interview solutions to human resources professionals and departments across the world, suggesting many more have been conducted, and many more will be conducted.

Unfortunately, to date, I-O psychology has been largely silent on the topic of automated video interviews. Computer scientists have conducted the initial research in this area.

How do automated video interviews work? They use computers to quantify interviewee behavior (verbal, paraverbal, and nonverbal behavior), then use those behaviors as predictors in machine learning algorithms. By automating and standardizing the interview process, automated video interviews hold potential to reduce time to hire and improve the quality of new hires. Yet, since there is little publicly available validity evidence, more research is needed before organizations should adopt these tools. This presentation details some of our initial investigations into the psychometric properties of automated video interview personality trait

assessments.



Some researchers have raised concerns that I-O psychology is increasingly focusing on methodological minutiae and theoretical models that have little relevance to real world issues and applications. They suggest that I-O psychology should refocus its efforts on real world issues that affect today's organizations. We believe automated video interviews is one important area that deserves attention, since millions of workers may be affected by these technologies, yet we know little about their potential reliability or validity.

One solution we investigated in a prior paper was whether off-the-shelf machine learning personality models could be effective for scoring personality in interviews. We applied IBM Watson Personality Insights, which was trained on Twitter posts, to a set of mock video interviews. In that investigation, the model's trait scores showed little to no convergence with either self-reported or interviewer-judged personality traits.

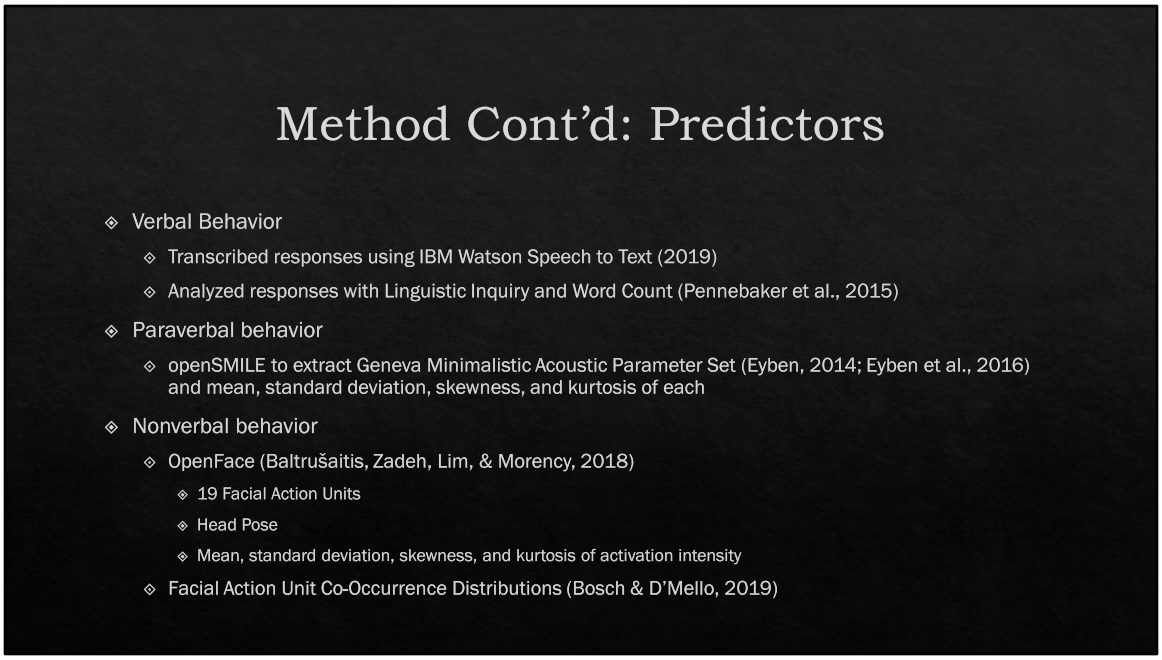
Therefore, we developed our own machine learning models for automatically scoring personality traits in interviews. These models are native to the interview context, as they are trained on interview data. In evaluating these models, we make two primary contributions to the literature. First, we investigate whether big data methods can be

used to automatically score open-ended responses. Interviews require open-ended answers from interviewees, and one-way/asynchronous interviews can be used to score personality traits. This leads us to our next contribution, which is to investigate alternatives to self-reported traits for assessing applicant personality, a goal raised by scholars who are skeptical of the value of self-reported traits in personnel selection.

Method: Sample and Procedures

- 490 psychology subject pool participants
  - Self-reported Big Five traits using 50-item IPIP (Goldberg, 1992; 1999)
  - Mock video interview consisting of three questions, 2-3 minute answers for each (*M* = 6 min 51 s)
    - Please tell us about yourself
    - Please tell us about at time when you worked effectively in a team
    - Please tell us about a time when you demonstrated leadership
  - 467 participants completed the study in full; self-reports were available for 396 of these participants after removing those who failed attention checks
- Ground truth (i.e., interviewer judgments of personality)
  - Undergraduate research assistants participated in frame of reference training
  - At least three from a pool of eight rated each interviewee
  - Used an observer version of the Ten Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003)

We used psychology subject pool participants for the study. Many prior studies of interviews have used mock interviews with students. In our study, they completed a Big Five personality self-report as well as a mock video interview that was modeled after prior research by computer scientists and made to be applicable to a wide range of jobs.



As mentioned, we will use computerized descriptions of interviewee behavior as predictors. According to the model of interviewee performance (Huffcutt, Van Iddekinge, & Roth, 2011), interviewee performance consists of verbal, paraverbal, and nonverbal behaviors. This is what you say (verbal), how you say it (paraverbal), and what you do while in the interview (nonverbal). We quantify all three types of behavior, then develop machine learning models using each separately, two of the types of behaviors, as well as all three together to see how they contribute to overall accuracy.

To quantify verbal behavior, we first transcribed responses using IBM Watson Speech to Text. Although computerized transcriptions can introduce errors into the analysis, we thought it important to use computerized transcription since the vendors of automated video interviews are also using computerized transcription. Because the dataset is relatively small, we used Linguistic Inquiry and Word Count to describe verbal behavior. We could use open vocabulary text mining, but due to overfitting, it would likely not cross-validate well in this relatively (for text mining) small dataset.

To quantify paraverbal behavior, we used openSMILE to extract the Geneva Minimalistic Acoustic Parameter Set. This includes things like pitch, indices of voice

quality including jitter and harmonics-to-noise ratio, frequency, loudness, speech rate, and more. We extracted these features in 30-second windows of time, sliding by 1 second within these windows, then aggregated the results using means, standard deviations, skewness, and kurtosis.

To quantify nonverbal behavior, we used OpenFace to extract 19 facial action units and head pose features. Facial action units are best known from Ekman's pioneering work on universal facial expressions/emotions. Note that we did *not* use this software to extract discrete facial emotions (e.g., happy, sad, angry) because facial expressions may be heavily influenced by context, so do not correspond one-to-one with basic emotions (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). The facial action units and head pose features were described by their mean intensity, as well as the standard deviation, kurtosis, and skewness of the intensity. The cooccurrence features are measured via Jensen-Shannon divergence and index the similarity between two action units, which is a way of describing facial expressions without making a priori assumptions about the emotions to which they correspond.

Method Cont'd: Analytic Strategy

- 10-fold cross-validation using all predictors
  - Tested multiple algorithms: Elastic Net Regression, XGBoost, and Random Forest
- Use the best performing algorithm to test the contribution of each type of behavior
- Used the caret R package

If you have been introduced to 10-fold or k-fold cross-validation before, feel welcome to go on to the next slide. 10-fold cross-validation involves splitting the data into 10 equally sized parts, training a machine learning model on nine of the ten parts, then testing the accuracy of its predictions on the remaining tenth, and conducting this process a total of 10 times, using each fold once and only once for testing. Accuracy is reported as the highest average cross-validated convergence with the ground truth the algorithm was designed to predict. In this case, we trained the algorithms to predict interviewer-reported traits.

Results

Table 1

Cross-validated accuracy using video, audio, and language data to predict observer ratings

	E		A		C		ES		O	
Algorithm	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r
Elastic Net	**.868**	**.67**	**.590**	**.49**	**.506**	**.45**	.562	**.32**	**.783**	.44
XGBoost	.907	.62	.621	.43	.585	.30	.595	.22	.859	.33
Random Forest	.875	.66	.632	.44	.518	.44	**.547**	**.32**	.792	**.45**

Note: E = extraversion. A = agreeableness. C = conscientiousness. ES = emotional stability. O = openness to new experiences. **Bold** indicates highest accuracy for that trait.

Here is the first set of results. We ran 10-fold cross-validation to identify optimal hyperparameters, and the accuracy reported here is for the optimal set of hyperparameters for each algorithm. Across the five traits, on average, Elastic Net Regression performed best. It had the lowest RMSE and highest r for three traits, while having RMSE and correlations that were comparable with Random Forest on the remaining two traits. This initial evidence suggests that although more complex mathematical algorithms can sometimes improve prediction, in this case, regularized regression (i.e., Elastic Net) performs just as well, if not better. Emotional stability was the least accurately inferred trait, and this may be because interviewers find it hard to judge emotional stability, so may have relatively inconsistent criteria for what leads them to decide whether an interviewee is emotionally stable or neurotic. The results for extraversion are particularly promising, as convergence > .6 is often as good as can be achieved (cf., Campion, Campion, Campion, & Reider, 2016).

Results Cont'd

Table 2
Cross-validated accuracy using features separately to predict observer ratings

	Extraversion			Agreeableness			Conscientiousness			Emotional Stability			Openness		
Data Mode	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2
Verbal	1.01	.49	.24	**.589**	**.49**	**.24**	.532	.40	.16	.561	.29	.09	.799	.40	.16
PV	.962	.58	.33	.637	.34	.12	.529	.36	.13	.561	.31	.09	.828	.34	.12
NV	1.04	.44	.20	.632	.36	.13	.562	.25	.06	.569	.24	.06	.856	.19	.04
PV + V	.926	.61	.37	.596	.46	.21	.519	.42	.17	**.556**	.29	.09	.794	.43	.18
PV + NV	.895	.63	.39	.627	.37	.14	.528	.39	.15	.562	.30	.09	.911	.40	.16
V + NV	.935	.60	.36	.595	.47	.22	.520	**.46**	**.21**	.566	.23	.05	.792	.42	.17
V + PV + NV	**.868**	**.67**	**.44**	.590	**.49**	**.24**	**.506**	.45	.20	.562	**.32**	**.10**	**.783**	**.44**	**.19**

Note: V = verbal behaviors. PV = paraverbal behaviors. NV = nonverbal behaviors. **Bold** indicates highest accuracy for that trait.

Because Elastic Net Regression performed best with the full set of features, we then investigated how accuracy changed when using each set of behaviors (i.e., verbal, paraverbal, and nonverbal) separately, in pairs, and all together. This informs us about which traits are relevant to each type of behavior, as well as which types of behavior provide the most useful information.

Importantly, of the three types of behaviors, verbal behaviors had the lowest RMSE and highest r/R-squared when used as the only predictors of traits. Further, of the three possible pairs of behaviors, the two pairs with verbal behavior (PV + V and V + NV) have lower average RMSE and higher r compared to the pairing of paraverbal and nonverbal behaviors. Indeed, we can see that for Agreeableness, Verbal behavior alone was as accurate as using all three types of data together. Further, verbal behavior alone was quite strong at predicting conscientiousness, emotional stability, and openness--in each case its results were similar in magnitude to using all three types of behavior as predictors. The only trait where we observed sizeable increases as we more types of behaviors were added was Extraversion, widely considered the most visible of the Big Five traits. Paraverbal behaviors were the type of behavior most strongly related to extraversion judgments.

We also investigated the convergence of interviewer-rated and automated video interview score personality traits with self-reported personality. Interviewer-ratings had an average monotrait correlation $r = .25$ with self-reports, while automated video interview personality scores had an average monotrait correlation $r = .18$ with self-reports.

Discussion

- Outside of I-O, researchers have shown that personality can be inferred in interviews (e.g., Naim et al., 2018) or from digital footprints like Facebook posts (e.g., Park et al., 2015)
- This research took initial steps toward bringing I-O into these conversations
  - Verbal behavior was the best predictor of most traits
  - Extraversion judgments were predicted quite accurately
- Past work has focused only on convergent-related evidence of validity…
  - Future work
    - Reliability
    - Discriminant-related evidence of validity
    - Criterion-related evidence of validity

References

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930
- Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.
- Bosch, N., & D'Mello, S. K. (2019). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2019.2908837
- Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975. https://doi.org/10.1037/apl0000108
- Eyben, F. (2014). *Real-time speech and music classification by large audio feature space extraction*. https://doi.org/10.1007/978-3-319-27299-3
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., … Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*(1), 26–42.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1
- Hickman, L., Tay, L., & Woo, S. E. (2019). Validity Investigation of Off-the-Shelf Language-Based Personality Assessment using Video Interviews: Convergent and Discriminant Relationships with Self and Observer Ratings. *Personnel Assessment and Decisions*, *early onli*, 1–23.
- Huffcutt, A. I., Van Iddekinge, C. H., & Roth, P. L. (2011). Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review*, *21*(4), 353–367. https://doi.org/10.1016/j.hrmr.2011.05.003
- IBM. (2019). IBM Watson Speech to Text. Available at https://www.ibm.com/watson/services/speech-to-text/
- Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, *102*(1), 43–66. https://doi.org/10.1037/apl0000160
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x
- Naim, I., Tanveer, I., Gildea, D., & Hoque, M. E. (2018). Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing*, *9*(2), 191–204. https://doi.org/10.1109/TAFFC.2016.2614299
- Ones, D. S., Kaiser, R. B., Chamorro-Premuzic, T., & Svensson, C. (2017). Has industrial-organizational psychology lost its way? *The Industrial-Organizational Psychologist*, *54*(4). Retrieved from http: //www.siop.org/tip/april17/lostio.aspx
- Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for Organizational Research and Practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1). https://doi.org/10.1146/annurev-orgpsych-032117-104553
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: Pennebaker Conglomerates.
- Rotolo, C. T., Church, A. H., Adler, S., Smither, J. W., Colquitt, A. L., Shull, A. C., … Foster, G. (2018). Putting an End to Bad Talent Management: A Call to Action for the Field of Industrial and Organizational Psychology. *Industrial and Organizational Psychology*, *11*(2), 176–219. https://doi.org/10.1017/iop.2018.6

Slide 1

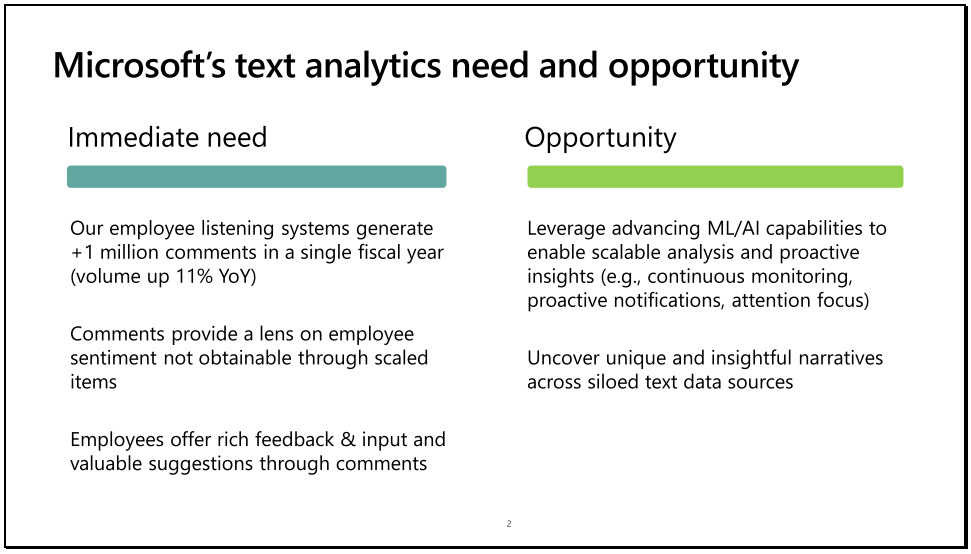
Scalable Analysis of Employee Comments Leveraging NLP and an Analytics Platform

Elizabeth A. McCune, Jason Lewris, Victoria Westerhoff

Prepared for the 2020 Virtual SIOP Conference

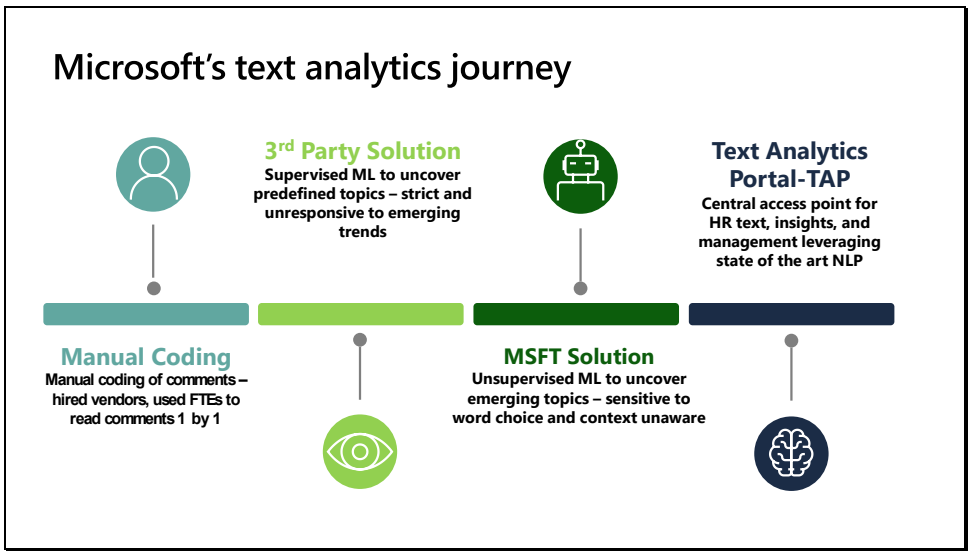
Elizabeth, Jason and Victoria are part of Microsoft's HR Business Insights team, which is the people analytics function within Microsoft. Elizabeth is the Director of Employee Listening Systems and Culture Measurement, Jason is a Data Scientist who specializes in natural language processing (NLP), and Victoria is a Business Analytics Specialist supporting the corporate functions within Microsoft.

Slide 2



We have both an immediate need as well as meaningful future opportunities in the space of scalable text analytics.
When it comes to our immediate need, our employee listening systems generate over one million comments in a single fiscal year, and the volume is growing. In the last year the number of comments we received increased by 11%. We also know that comments provide a lens on employee sentiment and experience that goes beyond what we can obtain through our quantitative items. What we've found in our work within the people analytics function at Microsoft, and based on the feedback we receive from leaders, is that employees offer rich feedback & input and valuable suggestions through comments that can be leveraged to improve the employee experience.
We have an opportunity to leverage advancing machine learning and artificial intelligence techniques to enable scalable analysis of this text and offer proactive insights. Moreover, these techniques can be leveraged to uncover insightful narratives across data sources.

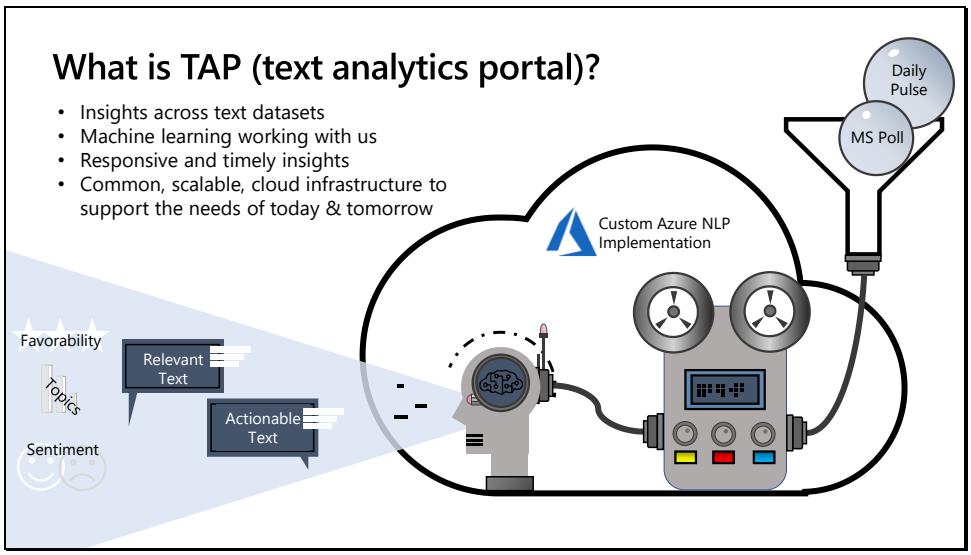
Slide 3



We, like many companies, started our text analysis journey with manual coding. We would pay vendors and leverage FTE time to read and code a random sample of comments received across all employees. Of course, managers and leaders were given access to the comments for their organizations, but in terms of generating insights at a company-wide level, the volume of comments received couldn't be adequately scaled to manual coding.

Our first foray into NLP for employee survey comments was a 3rd party solution leveraging a supervised ML model. Comments were assigned to a predefined list of topics set by the provider. While we found this tool helpful for gaining a high level understanding of the nature of comments, the rigidity of the topics was limiting and the models were really only suitable to narrow range of data sources. We then began using an NLP solution developed by tone of our internal data science teams. This tool uses an unsupervised ML approach which provides much more flexibility with regard to identifying topics that reflect the nature of the specific data source. This is a great solution that we continue to leverage today under certain scenarios. However, the amount of manual work required for analysts to upload data into the tool, and the learning curve required to build sufficient capability presented an opportunity for us to create a solution that would be custom built to the most common text analysis scenarios on our team, help analysts skip the labor-intensive data upload and cleaning steps, and ultimately yield faster insights.

Slide 4



The Text Analytics Portal (TAP) is an analysis tool that democratizes the ability to enrich narratives/stories/presentations with qualitative analysis by lowering the barriers to entry to get started with text analytics research using state of the art NLP methods on HR text data. Currently, the users are analysts in our people analytics team at Microsoft, however in the next year we anticipate releasing a streamlined version of TAP to several hundred HR professionals at Microsoft.

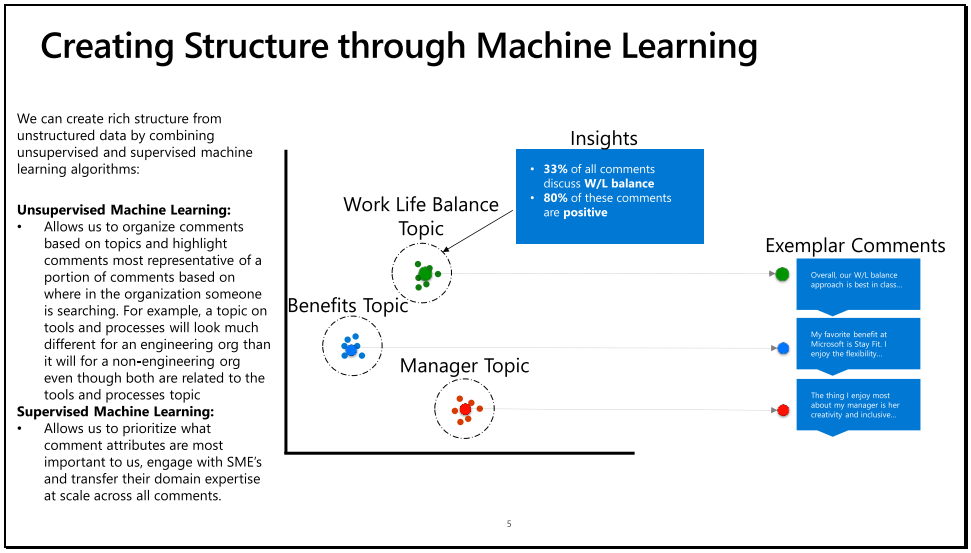
Our focus in developing TAP has been to centralize our survey text datasets and apply a library of common machine learning capabilities on top of this normalized data structure in the cloud. So far, we have ingested comments from two of our largest survey programs at Microsoft, including our annual MS Poll survey and our continuous Daily Pulse survey. Leveraging a custom Azure NLP implementation* we perform the following key functions:

1) Normalization of text data sources. There are several advantages to normalizing your different text data sources into a common format, including:
- Ease the load on analysts – no longer must relearn field names for each dataset of interest
- Ease the load on data scientists – downstream analysis/algorithm development has a known format in which folks can expect
- Normalize to a common base language – our employees respond in over a dozen different languages – we translate comments into a common base language using ML while retaining original comments
2) Assign a common set of NLP attributes to comments.

- Topic models group comments together based on underlying themes – it scores each comment with the likelihood of a topic belonging to any given comment
- Sentiment models provide positive/negative/or neutral classification codes to each comment
Organizing relevant text up front based on different parts of the organization users are exploring – embeddings allow you to measure relatedness of comments from each other and to groups of other comments and accounts for the context contained within a comment. Historically, comment similarity was defined based on the prevalence of keywords, however keywords can mean very different things based on the context in which they appear.

*What's meant by a custom Azure NLP implementation is an Azure subscription with carefully selected and customized services that support the ability to:
- Ingest new data sources and store them in a common, accessible format
- Update data sources as new data is provided by employees
- Pilot new machine learning ideas and put them into 'production' easily
- Serve data to down stream reporting solutions

Slide 5



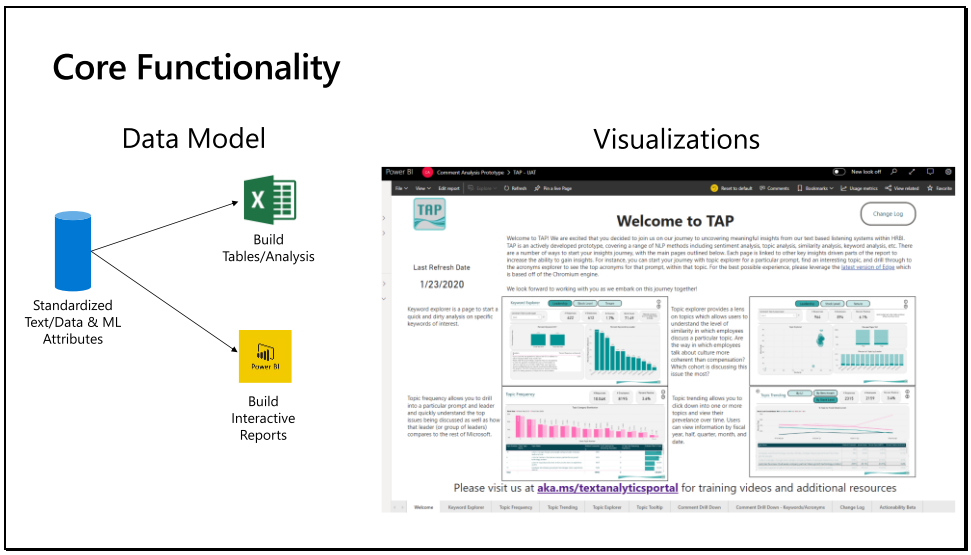
By combining unsupervised and supervised machine learning algorithms, we can create a rich structure from unstructured text. While the responsibilities of these types of machine learning algorithms can be quite different, they can be complementary. The example visual above projects unstructured text into a measurable 2d space, where these projections are learned from unsupervised learning algorithms that read through your text and create numerical representations of text in the form of embeddings. Embeddings can be projected onto 2d and 3d spaces for visualization purposes, but typically have hundreds of dimensions which in their raw form can be directly utilized to measure the distance between and among comments or groups of comments. This allows you to surface comments most typical of a topic, within a given set of user applied filters. You can further enrich this type of analysis by layering on supervised learning insights such as sentiment. For instance, referring back to the visual above, we can define the sentiment by topic grouping, or we can capture the sentiment for any given projected comment above.

Supervised learning allows us to scale out expertise from our SME's and apply them to an entire universe of employee comments. For instance, if a group of subject matter experts has been working hard on identifying comments that contain suggestions to improve a particular tool within one of their client organizations, supervised learning allows us to leverage the work they have already done in years past and help narrow the focus area for future iterations of the work. Additionally, supervised learning allows teams to prioritize what are the most important attributes to know about all comments and invest in making these attributes accessible, regardless of how you cut the data. Modern NLP practices have significantly reduced the volume of labelled data required to maximize the impact for supervised learning algorithms.

Teams no longer need tens of thousands of labels to make reasonable supervised machine learning algorithms. We have examples of creating successful models with as little as a few hundred labels and have put models like this into production, creating more efficient processes for our client teams.

Our primary form of supervised learning is in the form of deep learning transformer models like BERT. We have built on the work of the open source community and identified efficient ways of adapting open source models into the Microsoft domain to create high quality models with limited amounts of data.

Slide 6



By standardizing HR text data across several disparate data sources into a common format, analysts no longer need deep expertise of the underlying datasets in order to make use of it and build from it. Users simply need to know what dates they want, which program, and can query and build new tables, reports, and analysis at will. By doing it once, they can repeat that analysis across different HR text data sources. They have access to the same underlying machine learning text attributes that are made available in the reporting side of TAP and can use those attributes to sort and filter text to more efficiently identify what they are looking for. Additionally, we make available a number of key analysis visualizations to explore text in a more structured way. Visualizations are made available to explore topics, sentiment, emerging questions, keywords, etc. through a series of PowerBI reports.

Slide 7

Use Case: Leadership and Culture Sentiment

Question: How do different sub-groups within organizations resonate with, absorb, and enact corporate culture initiatives? How can executives change accountability to create sustainable culture programs?

Adaptive Topic Models: Reviewing the generated topics, analysts could observe that culture was prevalently spoken about in two ways, and associated with two distinct threads of sentiment

Context-Specific Models: Topics are generated based on prompts, but modeled across other employee variables, allowing dynamic cuts to explore how culture is discussed under different leaders and within varied populations

Key Words Across Models: Key word search allows analysts to narrow in on specific, tokened company culture terms, tracking prevalence of term use across leaders, sentiment with use, and differences in topic alignment

Ultimately, TAP enabled the discovery of two different discussions on culture, one **asking for modelling at a leadership level**, another on **managers' daily activation** of values. The context variables allowed for **detection of successes and differences between micro-populations** across the company, and key word propensity and sentiment approximated **programs' success in defining and spreading culture concepts and terms.**

Slide 8

Use Case: Career Sentiment Differences and Themes

Question: Are one-off cases of sentiment on career substantiated across the board? How to they differ across groups and what major themes inform the narratives in various sectors of the organization?

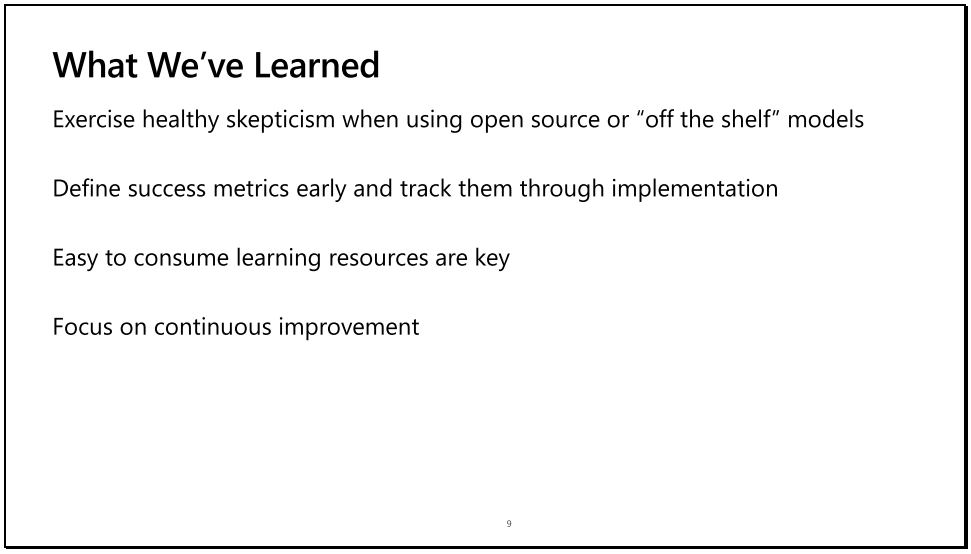
Working from specific comments to overarching insights, **keyword analysis on "career" expanded one instance of feedback into a guiding theme for overall text analysis**. While topic models are created based on the response population to a single survey question, key word investigation and analysis showed the **varied uses of a word across the entire response population**, revealing actionable sentiment and content differences between disciplines and under different leaders, informing custom solutions.

Adaptive Topic Models: The key word analysis interface reports the propensity of a word in response to different survey questions, informing analysts where career concerns were mentioned most and which topic models to drill into

Context-Specific Models: Tracking a single word, but investigating the context in which the word is used across questions allows analysts to disambiguate different conversations around career and isolate the relevant topics

Key Words Across Models: Analysts cut and view key word usage by organization alignment, hierarchy, tenure, and leader, giving unprecedented acuity into actionable insights on specific populations for leaders to implement.

Slide 9



We've learned so much in this process, and to close we've summarized a few of those key learnings.

It's important to exercise healthy skepticism when using open source or "off the shelf" models. We found that calibrating models on our company's domain yielded substantial improvements to both your supervised and unsupervised learning algorithms. Calibrating a model on your domain does not require labels. If you have large pools of unlabelled employee comments/text, you can calibrate models on your domain by essentially tasking the model with reading this text. Through the inherent structure of your text, the model can learn about your organization. This calibrated model can serve as your foundation for many supervised tasks, reducing the volume of labelled data required to make quality predictions.

The role of success metrics plays out in two areas for us: 1) the metrics we used to determine the accuracy and utility of the models, and 2) the metrics we used to track usage of TAP as a scalable solution. Success metrics for the model are critical of course, as you proceed through many iterations and need to assess performance through those iterations. But if you're also considering scaling your work to other users similar to what we have done, you'll also want to be sure to track the uptake and usage of the solution. Tracking this information continues to be essential for us, as it helps us identify who our "power users" are and the extent to which our user base is evenly spread across the teams we want to be using TAP.

We've taken a wide range of approaches to training the analysts on our people analytics team (our targeted users) in TAP. What we've found is that easy to consume learning resources that

target specific scenarios analysts encounter are key. Each training resource that we have developed is oriented around actual examples that analysts have encountered with their clients related to text analysis. The format in which we deliver these resources has varied from relatively lengthy and comprehensive recorded trainings, to 3-5 min videos for a specific scenario, to hands-on workshops. We have found that providing some of the foundational NLP knowledge that is important for interpretation of these models is best acquired through learning about these actual scenarios.

And finally, we continue to be reminded that change takes time, and the important thing to focus on is continuous improvement by continually seeking feedback from users and scanning for opportunities.

Machine Learning for I-O 2.0

Fred Oswald
Rice University
Discussant

Presentations

Mengqiao (MQ) Liu, Amazon, ***Chair***

Li Guan, Aon, Stefan Gaertner, Aon, Amy Garner, Aon Inc., ***Identifying Alternative Job Choices Based on Employees' Job Profiles***

Ivan Hernandez, Virginia Tech, Andrea Sanders, DePaul University, Soonyoung Kim, Virginia Tech, Steven Towe, DePaul University, ***Deep Selection: Inferring Employee Traits from Résumé Style Using Neural Networks***

Louis Hickman, Purdue University, Nigel Bosch, University of Illinois at Urbana-Champaign, Louis Tay, Purdue University, Vincent Ng, Purdue University, Rachel M. Saef, Northern Illinois University, Sang Eun Woo, Purdue University, ***Investigating Emotion Analytics for Predicting Personality in Video Interviews***

Elizabeth A. McCune, Microsoft, Jason Lewris, Microsoft, Victoria Westerhoff, Microsoft, ***Scalable Analysis of Employee Comments Leveraging NLP and an Analytics Platform***

Fred Oswald, Rice University, ***Discussant***

AI and Machine Learning: Important Distinctions

- **Technologies**

  intensive: e.g., mouse clicks, audio/video, text

  extensive: e.g., enterprise data merged across HR, departments, clients

- **Big Data**

  incidental: e.g., Facebook "likes", times that employees used their door card

  intentional: e.g., traditional job applicant measures (personality)

- **ML algorithms**

  predictive: e.g., convolutional neural nets, random forests, SVMs

  interpretable: e.g., lasso and elastic net regression, rotated PCA

- **Settings**

  local (and hyper-local): e.g., teams, jobs, employees over time

  broad: e.g., cross-industry, cross-cultural

AI and Machine Learning: Potential Benefits

- Faster throughput
- Wider scale
- ‘Natural’ data (text, video, game behavior)
- Applicant engagement
- Improved prediction

- However, opening the black box will raise new questions about I-O psychology and reliability, validity, fairness as much as answer them….

AI and Machine Learning: Potential Concerns

Concerns span science, practice, ethical, legal

- All selection practices should make earnest attempts to adhere to the SIOP *Principles*
- Unstandardized measures are concerning, subject to fairness issues (when data are 'natural' and people can respond however they like, similar to the problem with unstructured interviews)
- Need to systematically examine cost/benefit vs. reasonable alternatives (Is the AI solution better than other options? What is sufficient evidence to convince you, one way or the other?)
- Lack of interpretability (You have prediction, but do you know why? Was a job analysis involved to ensure the predictors and criteria are job-relevant?)
- Privacy issues (invasiveness, surveillance)
- Gaming the system (How will prediction change when job applicants know they are being watched?)

Hickman et al.

- Machine learning can extract personality trait information from interviews, given that they are shown to converge with the ratings of undergraduates (especially for extraversion)

- Appreciated seeing results for the verbal/paraverbal/nonverbal components of prediction (this is more specific information than similar studies)

- Because there is no gold standard, maybe some variance unique to machine learning and unique to ratings could also personality relevant. Future studies will need to obtain more convergent-discriminant validity information to better identify and understand personality relevant variance.

Guan et al.

- Job change happens for many external reasons (salary, career advancement, geographic change) and internal reasons (skill development, greater autonomy, more variety/interest)
- Both job-seekers and employers benefit from thinking about KSAOs at more refined levels than college degrees or other credentials (see ONET)
- Moving forward, a transition matrix (major-to-job, job-to-job) can be useful for modeling pipelines tied to personal development, employer opportunity, and policy efforts (what factors, such as those above, contribute to these transitions)
- Machine learning can model how things were; but may not predict how things will be (e.g., given macro-level economic changes such as COVID-19)

McCune et al.

- Text analytics and employee sentiment → specific context + “healthy skepticism” important for interpretation and value (a great appraoch)
- Both context and structure reflects and informs source, team, performance, satisfaction, climate, etc.
- Topics in isolation can become more important when they are re-evaluated in context
- Speed of text processing was consistent and necessary w/ daily-pulse-based data collection
- Appreciated how discussion of the data iterated with intervention

Hernandez et al.

- Neural networks can extract meaning from visual features of resumes…but what are these features?
- Consider whether structured resumes would be better than standard resumes that are non-structured
- Extracted information correlated with *g* and B5 – 3-10% variance and consistently higher than humans (why were humans much worse at openness…or neural nets much better)?
- What are the implications of these relationships (e.g., without knowing the resume characteristics…maybe longer resumes = higher C, but future gaming of the system would undermine this relationship)
- In future work, organizational criterion data and related validity information will be useful.

Thank You!

Fred Oswald
foswald@rice.edu
https://workforce.rice.edu