WATSON GETS PERSONAL: NOTES ON UBIQUITOUS PSYCHOMETRICS

MARC BÖHLEN

RealTechSupport,
Department of Media Study,
University at Buffalo, USA

marcbohlen@acm.org

Keywords
Cognitive Computing
Machine Learning
Qualculation
Cultural Techniques
Psychometrics
Public Computational Media

Novel computational methods and platforms have opened a new front in psychometrics, the search for measurable personal traits in artefacts created by humans. In particular, texts produced on social media channels have been queried for insights into behavior patterns and sentiment.

This text describes an experiment in querying computing platforms that offer psychometric text evaluation. The aim of the project is to reveal some of the dynamics and assumptions hidden in the code underlying computational text analysis. The project suggests “asymmetrical coding” as a practical intervention to build interfaces for opaque deep-learning systems that act outside the direct reach of individuals, yet produce conditions that effect multitudes.

2016.
xCoAx
.org

Computation
Communication
Aesthetics
& X
Bergamo, Italy

1 BACKGROUND

In 1991 Mark Weiser suggested that technologies weave themselves into the fabric of everyday life and disappear into the background (1991). Computers, like automobiles before them, would be perceived as everyday objects and not technical devices once they became ubiquitous. Callon and Law later coined the term *qualculation* to describe the mutual territory and dependencies between calculative and noncalculative actions (2005). Thrift expanded the concept of disappearing technology and qualculation as the product of ubiquitous computing events at the infrastructural level (2004). With massive deployment and aggressive distribution, these ubiquitous systems become constitutive; they pervade not only physical but also mental spaces and yet remain largely undetected when in operation.

Deep (machine) learning attempts to model high-level abstractions to detect or extract features hidden in large datasets. While feature extraction technologies have been in operation for many years, new efforts within big data research proposes to find features that were not anticipated, promising the discovery (not just the recovery) of knowledge and even the production of prediction (Mckenzie 2015). The activities of deep machine learning produce a new dimension of the qualculative condition. As opposed to the Thriftian framework in which computing and its infrastructure is hidden and operates in the background, the mental modes produced by machine learning enter into the foreground; their existence is acknowledged and celebrated by the engineering industry.

Cognitive computing is a catch phrase used to describe large-scale deep machine learning research applied to IBM business analytics. Cognitive computing is an offspring of several parents, expanding specifically on previous neural network research that, in turn, borrows from research into biological information processing systems and their ability to detect patterns within large quantities of unstructured information. Cognitive computing operates on an industrial scale; it is a platform level activity, dependent on and flourishing in cloud-scale centralized computing environments with access to vast amounts of data from distributed and changing sources. Cognitive computing no longer exists as a background flow but operates as a foreground fact that actively intervenes into everyday life, offering new approaches to problems of general interest and commercial value.

2 PUBLIC COMPUTATIONAL MEDIA

Computational systems have long been recognized as cultural territory (Agre 1997), and artists have responded in various ways to this condition. My own practice has produced contributions to the field in the past. Recently, I have become interested in specific responses to computational systems *at scale*. This inquiry is part of an ongoing research agenda of *public computational media* (PCM), the study of various aspects of computation systems and their ramifications for public life. The goal of PCM is to contribute to the debate through experiments that materialize ideas and test procedures in ways that text-centric methods do not. Machine learning and its corporate derivative, cognitive computing, are part of a new active computing infrastructure that acts on data of public interest, and as such are territory for PCM.

3 STATE OF THE ART COGNITIVE COMPUTING: WATSON

Cognitive computing is conceptually premised on cognitive science research that attempts to explain how thought occurs in the human mind. While the quest for synthetic consciousness remains evasive, and misgivings about the project's grandiose goals openly questioned (Marraffa 1999), the goal of synthetic problem solving has made rapid advances in recent years. In particular, data-centric deep learning systems have been created that are able to reliably detect patterns from a corpus of data.

Watson, a computation framework created by IBM Research, is one of the most prominent examples of corporate machine learning today. The Watson palette consists of a variety of subsystems that can be applied to different types of data analysis problems. Watson was designed with the aim of processing structured and semi-structured information more efficiently than a human being. The project has a long history and has achieved prominent success. In 1997, a Watson predecessor, Deep Blue, won a six game match against the then World Chess Champion, Garry Kasparov. In 2011, a first version of Watson, designed specifically as a question-answering system with the full text of Wikipedia loaded into its four terabytes of disk memory (Ferrucci 2012), emerged as the winner of the quiz show Jeopardy!

Watson is a curious name for a system seeking superhuman intelligence. While the system's makers link the name to the founder of IBM, Thomas J. Watson (Ferrucci 2012), the name is also reminiscent of the fictional character John H. Watson, Sher-

lock Holmes's faithful and reliable assistant who can never quite match up to his master's superior deductive abilities.

But silicon Watson has loftier goals. The synthetic Watson has been created to find reliable answers in unstructured data more effectively than standard computational search solutions. According to IBM internal evaluations, Watson meets or exceeds performance metrics of other state-of-the-art search technologies (Saon 2015),

4 PSYCHOMETRICS AND COMPUTATION

Psychologists have historically sought measures that reveal the secret, hidden, or distorted real self. The theory of values posits that every person has a set of values or goals that motivate their actions (Schwartz 1994). This construct, referred to as the theory of Basic Human Values, maps desirable, trans-situational goals of people's lives independent of cultural boundaries onto 10 basic values: universalism, benevolence, conformity, tradition, security, power, achievement, hedonism, stimulation, and self-direction. Similarly, the Five-Factor Model (Big 5) categorizes human personality traits into five categories: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness (Norman 1963), (Goldberg 1981).

The field of psycholinguistics sees in language a window into hidden self-valuation systems. From that perspective, the use of language is the conduit through which these hidden values become exposed. This idea has a long history (Galton 1884, Allport 1936), and has led researchers to seek personal traits by comparing texts first with value orientation surveys, and then later by harnessing computational linguistics to perform the comparisons (Fast 2008, Fleishman 2009, Chen 2014). The increased use of social media and computing platforms has made it easier to (attempt to) detect human values in social media textual artefacts, and to (attempt to) automatically understand people through their use of social media production. More recent research has sought to expand the list of basic values, inferring even darker traits from social media text production, such as narcissism (Sumner 2012). By including more direct indicators of interests such as "liking", some researchers have claimed to detect explicit personal traits such as sexual and political orientation (Kosinski 2013).

One of the most popular text analysis packages is LIWC: Linguistic Inquiry and Word Count. LIWC has two central components, a processing node and a set of dictionaries with categories. When a LIWC based procedure is applied to a text, it calculates

the percentage of words for each LIWC category. Each of the 64 categories contains dozens to hundreds of words. LIWC has been employed in hundreds of studies to tally words in psychologically meaningful categories (Tausczik 2010, Matthews 2015).

Procedurally, the relationship between text and value is established by associating a specific trait with words that describe aspects of the sought trait. An input text is scanned and its words are parsed into the existing set of categories, as in a traditional linguistic parsing for pronouns and verbs, etc. The parsing becomes problematic when tallying qualitative features, such as emotion, for example. Which words should fall into the category of anger? While numerical tallying within a given category can be automated, the creation of categories themselves cannot. The LIWC system creates its own lists gleaned from dictionaries compiled by human helpers, and employs human word judges to categorize tricky entries (Tausczik 2010).

Qualitative features such as cruelty can be articulated in a text in many subtle ways, as any reader of Primo Levi will know. LIWC, however, is blind to sarcasm, irony and context in general. Systems such as LIWC are currently limited to detecting qualitative events on a per word level. Emotions are assessed by the detection of declared emotion words, and sadness might be found in the occurrence of words such as "hurt, sad, depressing, and disappointing". In addition to key words, sentence structure and language elements (such as pronoun and auxiliary word use) have been found to correlate with language emotionality, suggesting, according to psycholinguists, "a deeper importance of the expression of emotion and thinking styles" (Tausczik 2010). Other categories such as social coordination, honesty and deception are assessed in a similar process of combining detectable words and specific linguistic constructs. Some of the combinations are less convincing than others, however. One study makes the rather odd observation that an increased use of causal and insight words can be associated with—somehow—greater health improvements (Pennebaker 1997).

Despite the ongoing popularity of trait analysis, fundamental issues with the field persist. Boyle points out that there is no established theoretical basis for the Big-Five, that these features cannot be replicated consistently in different samples, and that even when detected, they provide only a static account of behavior regularities (Boyle 2008). Furthermore, there is surprisingly little attention devoted to differentiating the category of text within computational trait analysis. While a literature-centric approach to trait inquiry might consider poetry, legal proceedings and fictional accounts as vastly disparate territories, text

forms tend to be lumped together where text quality is secondary to text content. Recent research has started to investigate this deficit and inquire as to how different social media platforms influence the assessment of traits. Haber, for example, reports that pronouns are less frequent in wikis as opposed to blogs, and profanity is much more frequent on Twitter than in business-oriented media (Haber 2015). Consequently, media specific variations in word use are reflected in models created based upon those word use patterns.

5 WATSON DOES PSYCHOMETRICS

According to published IBM reports, Watson is premised on the same theoretical assumption as standard computational psycholinguistics, namely that human traits can be detected in language use, and that this process can be automated with software. Watson staff (IBMWatson) cite in their justification prior efforts in the field (Fast 2008, Chen 2014), and describe how the research team expanded this existing framework into social media data sets and new personality features. The Watson team developed its own models to infer scores for the Big-Five with several additional dimensions from other models, including the aforementioned Basic Human Values (and Needs) system. The model for the Big-Five personality characteristics was trained on data from blogs, while the model for Values was trained on forum posts, and the Needs model from Twitter data. With these augmented models in place, the Watson service infers characteristics from textual input by tokenizing the input and matching the tokens with the LIWC psycholinguistic dictionary in order to compute scores for each of its categories.

Depending on the particular set of characteristics, Watson uses a weighted combination from the LIWC category scores to form its own final score. For example, the Big-Five uses coefficients reported by one source (Yarkoni 2010), whereas the coefficients for the Values were gleaned from another source (Chen 2014). Interestingly, the domain specificity of the models has less of a negative effect than one might imagine. IBM organized a study (Gou 2014) in which models from different sources were applied to Twitter data. The researchers found that for a large majority (> 80%) of the Twitter users, scores for personality traits that were inferred for these models correlated significantly with survey-based scores.

6 ASYMMETRICAL CODING

According to Watson engineers (Ferrucci 2012), the Watson system has been exposed to books "from the Gutenberg Project". But precisely which texts Watson was exposed to is not known. Watson, as is the case with all machine learning systems, will be challenged to deal with ill-defined input, unusual samples and small data sets. As other researchers have lamented (Scott 2014), outsiders rarely have access to the innards of commercial algorithms. Recent work in black box auditing has shown that it is possible to investigate how a classification model takes advantage of features in datasets without knowing how the models themselves are constructed (Adler 2016). While this line of research is extremely promising for algorithm auditing in general, my goal here is not to detect the predictive qualities of a given algorithm but to observe artefacts of classification created through exposure to unusual materials.

With developer access to Watson's Personality Insights analytic engine API, one can observe how Watson performs the assessment of character traits on arbitrary text input. As opposed to working with textual materials from current social media platforms, I have confronted Watson with media texts from old platforms, which are not influenced by the dictates of social media text production. In this experiment, Watson is asked to respond to input text it might not have been exposed to previously. The selection of texts includes: Sun Tzu's *The Art of War*, 5th century BCE, Plato's *Republic*, around 380 BCE, Lucretius' *On the Nature of Things*, 1st century BCE, Ovid's *Metamorphoses* of 1CE, Orwell's *1984* of 1949, Carrol's *Alice in Wonderland* of 1865, de Sade's *120 Days of Sodom* of 1785, Shelley's *Frankenstein* of 1818, Austen's *Sense and Sensibility* of 1811, Marx' C*ommunist Manifesto* of 1848, Kaczynski's (aka Unabomber) *Industrial Society and Its Future* (Manifesto) of 1995, as well as IBM's *Annual Report* of 2014.

While the current version (March 18, 2016) of the Personality Insights module supports Arabic, English, Spanish and Japanese, the texts in this experiment are mostly originally written in English or have been translated into English. The choice of text materials is guided by the measure of enduring cultural impact, with a corporate annual report the exception to the rule. The goal is not to see how well Watson classifies the documents, but to collect and reflect on unexpected relationships between business language and literature that Watson might generate. Or: how do the classics fare under Watson? Can (author) character traits from sophisticated textual production be differentiated from ephemeral writing? Figures 1 and 2 show an attempt to ad-

dress the questions. The figures show code-generated views into Watson-generated similarities between (the authors of) texts that could hardly be more dissimilar: the Communist Manifesto and the IBM annual report. Both sources get high marks in "achievement striving", and a maximum score for "imagination", for example.

Watson's Personality Insights assumes that the input text is produced by a single person. Arguably this is not the case for annual reports where a cohort of anonymous writers put together documents spanning hundreds of pages. Yet the voice that emerges from the annual report is singular. It represents the corporation and, at least in the USA, corporations enjoy many of the same important rights and responsibilities of individuals. While other countries do not explicitly support the US model of corporate personhood (Blair 2013), features such as corporate social responsibility (May 2015) and the unified entity they suggest have become person-like actors across the global business landscape. While the applicability of personhood to the annual report production team may appear spurious at first, it fits the historical pattern of the making of the corporate persona (Blair 2013) though the mechanism of the "artificial person" that, like a cyborg, can be similar to a living person yet completely different "inside". This is the rational for applying personality analysis to a corporate annual report.

Annual reports are an odd combination of frank assessment and reflection created by corporations around the world. Companies report on their activities, successes, investments and future plans in their annual reports. Annual reports define a unique form of language and language use; a mixture of hype, promotion and bureaucratic write-easy that is inspirational, self-congratulatory and obfuscating at the same time. Like blogs and wikis, the annual report is at home on social media. Yet the annual report is not an established textual category in the way that blogs, wikis and posts are, and hence it has not yet been declared worthy of model creation for text analysis. Here it serves as an example of a new global text production category, one well-suited to be read by machines; a reference point by which to compare an algorithm's responses to earlier textual production categories.

The Watson team is continuously tuning and refining the Watson code base and training materials. The publically available changelogs track the sequence updates but do not describe the technical details and operational changes. While this visualization allows one to see the current state of the algorithm, it is not sufficient to show changes between updates. However, re-applying the visualization to generate a time-series sequence of the

outputs of the invisible algorithm might show when changes occur and how the evaluation of the stable texts changes over time. As these texts remain fixed, the differences in the positions of the text markers will be due to changes in the Watson code itself. Such changes give a low-resolution lumped view of the temporality of code modifications over time, giving these texts a new role as markers for change and progress in otherwise inaccessible algorithm design. Figures 4 and 5 show an unsuccessful attempt at finding such differences. From these graphs, we can see that the system V2 did not change between November 2015 and April 2016 as far as the evaluation of the category "Melancholy" is concerned.

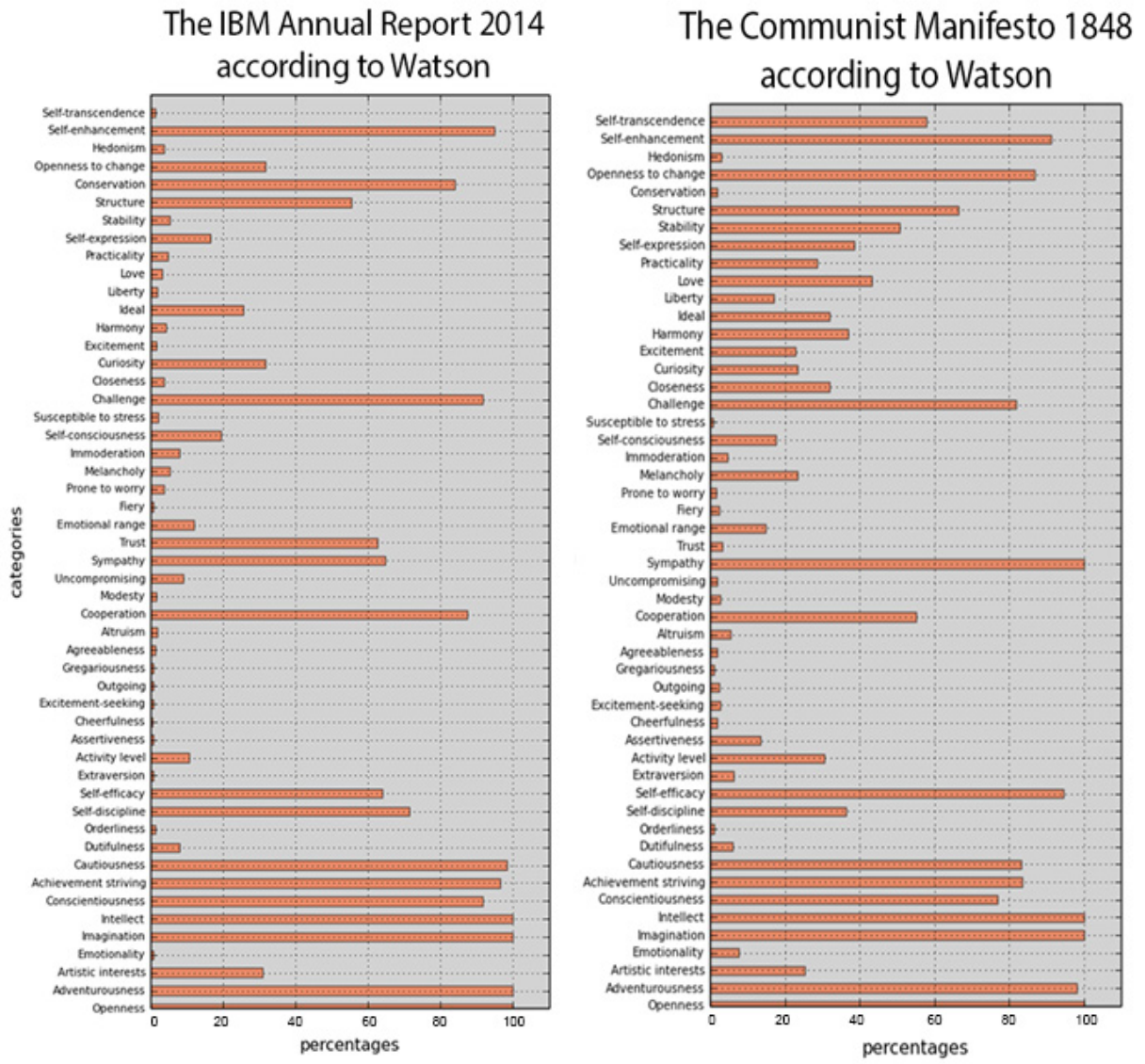

Fig. 1 and 2. Watson personality assessment engine applied to the IBM annual report of 2014 and the Communist Manifesto of 1848.

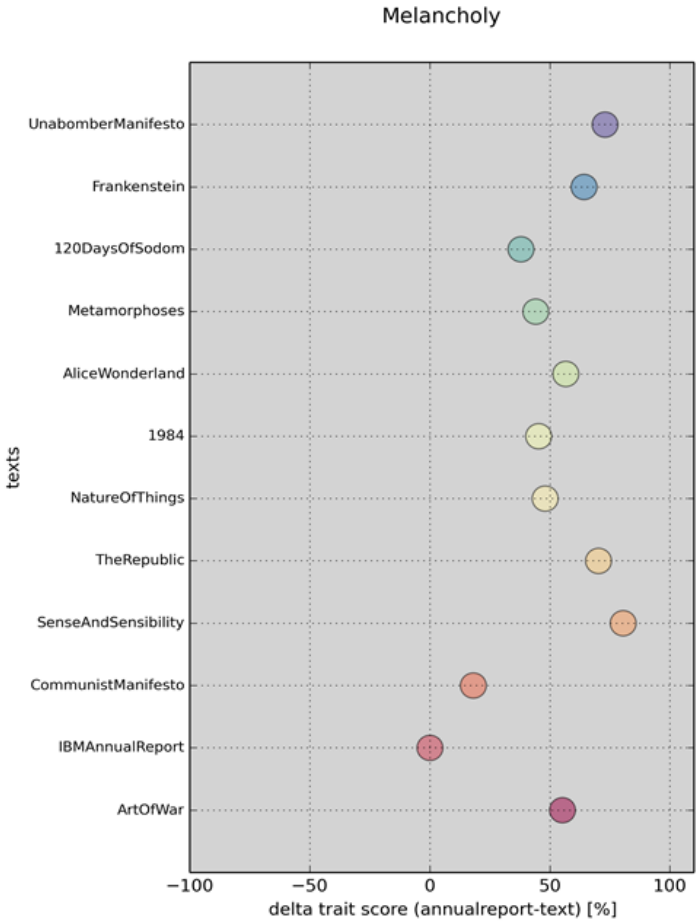

Fig. 3. Watson personality trait Melancholy (Personality Assessment algorithm version March 2016, filtered with a sampling error < 0.075) across a selection of text sources. Values are differential scores using the IBM Annual Report of 2014 (pale red at 0.0) as a reference. All results are normalized with regards to a sample population "based on a corpus of more than a quarter of a million Twitter users" (IBM Watson 2015).

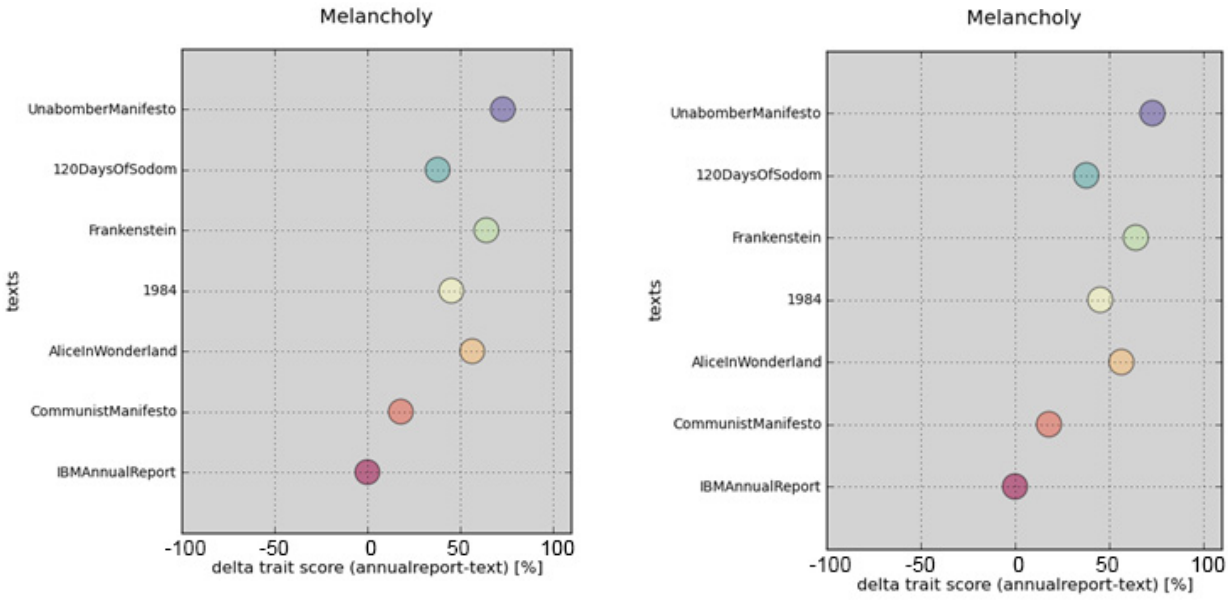

Fig. 4 and 5. No change detected in the evaluation of a subset of the sample texts between November 2015 (left) and April 2016 (right) within version 2 (v2) of the Personality Insights module.

Yet the principle is more important than the example. By generating code of codes, algorithms of algorithms, it is possible to perform more discerning observations of algorithm effects in practice (Scott 2014). While the current version of this code of codes is only a start, the principle suggests a new kind of contribution to investigations into large and inaccessible software systems that evaluate text production.

7 TEXT ANALYSIS, ASYMMETRIC CODING AND CULTURAL TECHNIQUES

Recently, researchers have begun to question the results produced from automated trait analysis of social media artefacts. Models created in one domain and applied to another can introduce spurious artefacts. Specific features such as style and message length vary between domains, and can make comparison of traits observed across platforms difficult. For example, character-count limited Twitter showed in one study the least Big-Five variability for a given sample size while email showed higher variability (Haber 2015). Neuman describes how the Watson analysis framework completely misses the mark on a text produced by a recent mass murderer, and then proposes adding semantic similarity measures to existing text analysis methods (Neuman 2015) to counter the detected deficiency.

The accuracy of Watson's Personality Insights has also been tested in less formal settings. Contributors to QUORA, for example, have tested the system with a variety of personal and nonsense texts, finding a tendency for flattery in the generated output (QUORA).

However, in this investigation, the goal is not to assess how correctly or incorrectly Watson performs, nor to improve the Watson engine, but rather to find new ways of observing system status. How does it change, how does it see the materials at any given moment? What kind of relationships are created between ephemeral and canonical texts, and how might even minor glitches scale, given the industrial level deployment of ubiquitous text analysis? The codes that produce such observations and relationships are a code-based form of cultural technique (Siergert 2013); they are new kinds of difference-producing operators.

The impetus to query this territory leads us back to the start of the paper; it is given by the ubiquity of machine learning and cognitive computing. The Watson system – as are other corporate machine learning frameworks - is being deployed in industries ranging from travel planning to weather monitoring and, most

recently, health care. Gaining insights, even in small measures, into how these systems operate is important, both to appreciate what they can accomplish but also to understand their shortcomings and failings.

The focus of text analysis systems such as the Watson personality trait assessment is indicative of the obsession with social media production that is altering the definition of text. The fact that an annual report and the Communist Manifesto both achieve top scores in "imagination" tells us that the system is blind to the contexts in which the texts operate. Political vision and business innovation become similar only through the system that evaluates them; a new category of machine learning enabled semantic glitch.

Training text classifiers on ephemeral text materials such as tweets elevates tweets to the status of formal text, previously held by literature. Will we find ourselves in the position of deciding to write (and think) the way computing platforms expect us to in the future, simply in order to minimize classification errors? Once cognitive computing becomes ubiquitous and scans electronic correspondence for signs of mental instability, thus impacting health insurance, we just might be inclined to do so.

As the reach of machine learning expands into new territory, the problem of ill-defined input will develop two along two different trajectories. First, it will be a concern for algorithms faced with normalizing new data sources such that results are computationally sound. Second, it will be a concern to people who do not want to be reduced to a computational compatibility issue. One response to this situation might be to offer additional training materials to classifiers in order to educate them on variations in texts and people in meaningful ways. Let the machines take over, finally, but perhaps we could have them read a few good books first.

REFERENCES

Adler et al., P. Auditing Black-box Models by Obscuring Features. arXiv: 1602.07043v1 [stat.ML], 2016.

Agre, P. Computation and Human Experience (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge University Press, 1997.

Allport, W., and S. Odbert. Trait-names: A psycho-lexical study. Albany, NY: Psychological Review Company. 1936.

Blair, M. Corporate Personhood and the Corporate Persona. U. Ill. L. Rev. 785, No. 3, 2013.

Boyle, G. Critique of the five-factor model of personality. In G. J. Boyle, G. Matthews & D. H. Saklofske (Eds.), The Sage handbook of personality theory and assessment: Vol. 1 personality theories and models. pp. 295-312. Sage Publications, 2008.

Chen et al., J Understanding individuals' personal values from social media word use. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). ACM, New York, NY, USA, pp. 405-414, 2014.

Callon, M., and J. Law. On Qualculation, Agency and Otherness. Environ Plan D, October, vol. 23 no. 5 pp. 717-733, 2005.

Fast, L., and D. Funder. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behaviour. Journal of Personality and Social Psychology, Vol 94(2), 2008.

Ferrucci, D. Introduction to This is Watson, in IBM Journal of Research and Development, vol.56, no.3.4, pp.1:1-1:15, May-June 2012. Also: http://www.aaai.org/Magazine/Watson/watson.php

Fleischmann et al., K. Automatic classification of human values: Applying computational thinking to information ethics. In ASIST, 46(1), pp. 1-4. 2009.

Galton, F. Measurement of Character. Fortnightly Review 36: pp. 179–185.1884.

Goldberg, R. Language and individual differences: The search for universals in personality lexicons. In Wheeler (ed.), Review of Personality and social psychology, vol. 1, pp. 141–165. Beverly Hills, CA: Sage. 1981.

Gou, L., M. Zhou, and H. Yang. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14). ACM, New York, NY, USA, 955-964.2014.

Haber. E. On the Stability of Online Language Features: How Much Text do you need to know a Person? ArXiv: 1504.06391, 24 April 2015. IBMWatson Developer Cloud. The science behind the Personality Insights service. http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/personality-insights/science.shtml https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/personality-insights/output.shtml#raw

Kosinski, M., M. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behaviour. PNAS 2013. LIWC. Linguistic Inquiry and Word Count: http://liwc.wpengine.com/ Mackenzie, A. The production of prediction: What does machine learning want? European Journal of Cultural Studies, Vol. 18(4-5), pp. 429-445. 2015.

Mckenzie, A. The production of prediction: What does machine learning want? European Journal of Cultural Studies 18(4-5):429-445, 2015.

Marraffa, M. Cognitive Computing and its Discontents. Review of 'Two Sciences of Mind' by Nuallain, McKevitt and Aogain, Psyche, 5 (27), 1999.

May, C. Global Corporations in Global Governance. Routledge, London, New York, 2015.

Norman, W. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. Journal of Abnormal and Social Psychology, Vol. 66(6) (June 1963): pp. 574-583.

Neuman, Y., and Y. Cohen. A Novel Methodology for Automatically Measuring Psychological Dimensions in Textual Data. The Computer Journal 2015. Public Computational Media. http://www.realtechsupport.org/RESEARCH!/research/public-realm.html

Pennebaker et al., J. Linguistic predictors of adaptive bereavement. Journal of Personality and Social Psychology, 72, 863-871. 1997. Quora. https://www.quora.com/How-accurate-is-IBMs-Watson-Personality-Insights-application

Saon et al., G. The IBM 2015 English Conversational Telephone Speech Recognition System. May 21, 2015. arXiv: 1505.05899, 2015.

Scott, S., and W. Orlikowski. Entanglements in Practice: Performing Anonymity Through Social Media. MIS Quarterly 38.3 (2014): 873-893.

Siegert, B. Cultural Techniques: Or the End of the Intellectual Post war Era in German Media Theory. Theory, Culture & Society, vol. 30 no. 6, pp. 48-65, 2013.

Schwartz, S. Are There Universal Aspects in the Structure and Contents of Human Values? Journal of Social Issues, 50(4), pp. 19-45, 1994.

Sumner et al., C. Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets. Conference proceedings at the IEEE 11th International Conference on Machine Learning and Applications ICMLA, 2012.

Tausczik, Y., and J. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, pp. 24-54, 2010.

Thrift, N. Movement-space: The changing domain of thinking resulting from the development of new kinds of spacial awareness, Economy and Society, Vol. 33, No. 4, pp. 582-604. 2004.

Yarkoni, T. Personality in 100'000 Words: A large-scale analysis of personality and word use among bloggers. J Res Pers. June 1; 44(3): pp. 363–373, 2010.

Weiser, M. The computer for the 21st century. Scientific American, 1991.