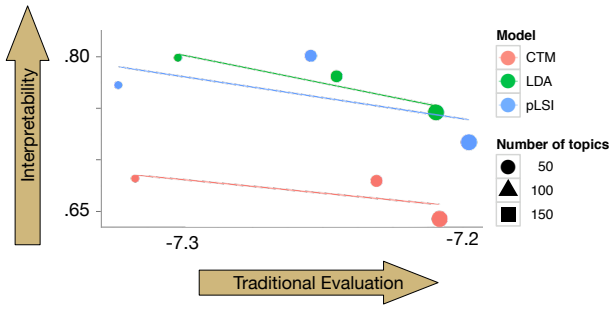
MACHINE LEARNING SHOULDN'T BE A BLACK BOX

JORDAN BOYD-GRABER, CU BOULDER

Machine learning is ubiquitous: detecting spam e-mails, flagging fraudulent purchases, and providing the next movie in a Netflix binge. But few users at the mercy of machine learning *outputs* know what's happening behind the curtain. My research goal is to demystify the black box for non-experts by creating *algorithms that can inform, collaborate with, compete with, and understand users* in real-world settings.

This is at odds with mainstream machine learning—take topic models. Topic models are sold as a tool for understanding large data collections: lawyers scouring Enron e-mails for a smoking gun, journalists making sense of Wikileaks, or humanists characterizing the oeuvre of Lope de Vega. But topic models' proponents never asked what those lawyers, journalists, or humanists needed. Instead, they optimized *held-out likelihood*. When my colleagues and I developed the *interpretability* measure to assess whether topic models' users understood their outputs, we found that interpretability and held-out likelihood were negatively correlated [3]! The topic modeling community (including me) had fetishized complexity at the expense of usability.



Since this humbling discovery, I've built topic models that are a collaboration between humans and computers. The computer starts by proposing an organization of the data. The user responds by separating confusing clusters, joining similar clusters together, or comparing notes with another user [6]. The model updates and then directs the user to problematic areas that it knows are wrong. This is a huge improvement over the "take it or leave it" philosophy of most machine learning algorithms.

This is not only a technical improvement but also an improvement to the social process of machine learning adoption. A program manager who used topic models to characterize NIH investments uncovered interesting synergies and trends, but the results were unpresentable because of a fatal flaw: one of the 700 clusters lumped urology together with the nervous system, anathema to NIH insiders [15]. Our tools allow non-experts to fix such obvious (to a human) problems, allowing machine learning algorithms to overcome the *social* barriers that often hamper adoption.

Topic Words (before)	Topic Words (after)
bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery	sci, spinal_cord, spinal_cord_injury, spinal, injury, recovery, motor, reflex, urothelial, injured, functional_recovery, plasticity, locomotor, cervical, locomotion

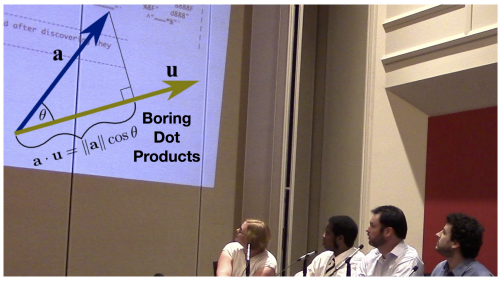
The machine learning tools that we developed to enable these interactions came from our attempt to model probabilistic lexicons [1]. We realized that much of the confusion came from words used in ambiguous contexts, and incorporating semantic knowledge (e.g., from WordNet) could help create clearer topics. However, WordNet is not always available (or the right answer), so our current interactive system allows users to create their own personal lexicon interactively to refine the results of unsupervised machine learning algorithms.

However, lexicons are not the only way humans can teach machines. *Simultaneous machine interpretation* [4] is another language-based task that requires significant human intuition, insight, and—for those who want to become interpreters—training. Because verbs end phrases in many languages, such as German and Japanese, existing algorithms must wait until the end of a sentence to begin translating (since English sentences have verbs near the start). We learned tricks from professional human interpreters—passivizing sentences and guessing the verb—to translate sentences sooner [5], letting speakers and algorithms cooperate together and enabling more natural cross-cultural communication.

The reverse of cooperation is competition; it also has much to teach computers. I've increasingly looked at language-based games whose clear goals and intrinsic fun speed research progress. For example, in *Diplomacy*, users chat with each other while marshaling armies for world conquest. Alliances are fluid: friends are betrayed and enemies embraced as the game develops. However, users' conversations let us predict when friendships break: betrayers writing ostensibly friendly messages before a betrayal become more polite, stop talking about the future, and change how much they write [14]. Diplomacy may be a nerdy game, but it is a fruitful testbed to teach computers to understand messy, emotional human interactions.

A game with higher stakes is politics. However, just like Diplomacy, the words that people use reveal their underlying goals; computational methods can help expose the "moves" political players can use. With collaborators in political science, we've built models that: show when politicians in debates strategically change the topic to influence others [10, 12]; frame topics to reflect political leanings [11]; use subtle linguistic phrasing to express their political leaning [8]; or create political subgroups with larger political movements [13].

Conversely, games also teach humans *how computers think*. Our trivia-playing robot [2, 7, 9] faced off against four former Jeopardy champions in front of 600 high school students.[1] The computer claimed an early lead, but we foolishly projected the computer's thought process for all to see. The humans learned to read the algorithm's ranked dot products and schemed to answer just before the computer. In five years of teaching machine learning, I've never had students catch on so quickly to how linear classifiers work. The probing questions from high school students in the audience showed they caught on too. (Later, when we played again against Ken Jennings,[2] he sat in front of the dot products and our system did much better.)



Advancing machine learning requires closer, more natural interactions. However, we still require much of the user—reading distributions or dot products—rather than natural interactions. Document exploration tools should describe in words what a cluster is, not just provide inscrutable word

[1] https://www.youtube.com/watch?v=LqsUaprYMOw
[2] https://www.youtube.com/watch?v=kTXJCEvCDYk

clouds. Deception detection systems should say *why* a betrayal is imminent. Question answers should explain *how* it knows Aaron Burr shot Alexander Hamilton: thus helping human players of trivia games either as a study partner or as a teammate at a competition.

Creating metrics to measure interpretability and systems that implement it is the subject of my recently awared NSF CAREER grant. By creating teams of humans and computers working together to solve language problems incrementally, we can create metrics that measure how much machine learning systems (and their visualizations) help or hurt the performance of their human teammates. We can then use these metrics to guide better visualizations using reinforcement learning. This complements machine learning's ubiquity with transparent, empathetic, and useful interactions with users.

I certify that this statement is a current and accurate statement of my professional record to the best of my knowledge

(October 7, 2017)

Full list of three book chapters, six journal publications, and fifty-five conference publications at
http://boydgraber.org/dyn-pubs/year.html

References

[1] Boyd-Graber, J., Blei, D.M., Zhu, X.: A topic model for word sense disambiguation. In: Proceedings of Empirical Methods in Natural Language Processing (2007), http://www.cs.colorado.edu/~jbg/docs/jbg-EMNLP07.pdf

[2] Boyd-Graber, J., Satinoff, B., He, H., Daumé III, H.: Besting the quiz master: Crowdsourcing incremental classification games. In: Empirical Methods in Natural Language Processing (2012), http://www.cs.colorado.edu/~jbg/docs/qb_emnlp_2012.pdf

[3] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proceedings of Advances in Neural Information Processing Systems (2009), http://www.cs.colorado.edu/~jbg/docs/nips2009-rtl.pdf

[4] Grissom II, A., He, H., Boyd-Graber, J., Morgan, J., Daumé III, H.: Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In: Proceedings of Empirical Methods in Natural Language Processing (2014), http://www.cs.colorado.edu/~jbg/docs/2014_emnlp_simtrans.pdf

[5] He, H., Grissom II, A., Boyd-Graber, J., Daumé III, H.: Syntax-based rewriting for simultaneous machine translation. In: Empirical Methods in Natural Language Processing (2015), http://www.cs.colorado.edu/~jbg/docs/2015_emnlp_rewrite.pdf

[6] Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Mach. Learn. 95(3), 423–469 (Jun 2014), http://dx.doi.org/10.1007/s10994-013-5413-0

[7] Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., Daumé III, H.: A neural network for factoid question answering over paragraphs. In: Proceedings of Empirical Methods in Natural Language Processing (2014), http://www.cs.colorado.edu/~jbg/docs/2014_emnlp_qb_rnn.pdf

[8] Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political ideology detection using recursive neural networks. In: Proceedings of the Association for Computational Linguistics (2014), http://www.cs.colorado.edu/~jbg/docs/2014_acl_rnn_ideology.pdf

[9] Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Association for Computational Linguistics (2015), http://www.cs.colorado.edu/~jbg/docs/2015_acl_dan.pdf

[10] Nguyen, V.A., Boyd-Graber, J., Resnik, P.: SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In: Proceedings of the Association for Computational Linguistics (2012), http://www.cs.colorado.edu/~jbg/docs/acl_2012_sits.pdf

[11] Nguyen, V.A., Boyd-Graber, J., Resnik, P.: Lexical and hierarchical topic regression. In: Proceedings of Advances in Neural Information Processing Systems (2013), http://www.cs.colorado.edu/~jbg/docs/2013_shlda.pdf

[12] Nguyen, V.A., Boyd-Graber, J., Resnik, P., Cai, D., Midberry, J., Wang, Y.: Modeling topic control to detect influence in conversations using nonparametric topic models. Machine Learning 95, 381–421 (2014), http://www.cs.colorado.edu/~jbg/docs/mlj_2013_influencer.pdf

[13] Nguyen, V.A., Boyd-Graber, J., Resnik, P., Miler, K.: Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In: Association for Computational Linguistics (2015), http://www.cs.colorado.edu/~jbg/docs/2015_acl_teaparty.pdf

[14] Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.: Linguistic harbingers of betrayal: A case study on an online strategy game. In: Association for Computational Linguistics (2015), http://www.cs.colorado.edu/~jbg/docs/2015_acl_diplomacy.pdf

[15] Talley, E.M., Newman, D., Mimno, D., Herr, B.W., Wallach, H.M., Burns, G.A.P.C., Leenders, A.G.M., McCallum, A.: Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods 8(6), 443–444 (May 2011)

E-mail address: jbg@boydgraber.org