Please quote as: Dellermann, D. & Calma, A. (2018): Making AI Ready for the Wild- the Hybrid Intelligence Unicorn Hunter. In: Social Science Research Network (SSRN).

Making AI Ready for the Wild- The Hybrid Intelligence Unicorn Hunter

Dominik Dellermann
vencortex
dominik.dellermann@vencortex.com

Adrian Calma
vencortex
adrian.calma@vencortex.com

Abstract
This demonstration paper proposes a hybrid intelligence system that combines the complementary strengths of human and machines for complex decision-making problems that require human "gut feeling" (i.e. success prediction for startups). The architecture extends principles of previous interactive machine learning systems by using continuous input from an expert crowd and explicitly leveraging the advantages of collective intelligence. This approach allows to augment machine learning techniques for generating features, intuitive and analytic labeling as well as troubleshooting.

Author Keywords
Interactive Machine Learning; Hybrid Intelligence; Crowdsourcing

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

Introduction
Progresses in machine learning techniques led to increasing use of artificial intelligence (AI) in real-world business contexts. In the wild, however, AI applications frequently struggle with complex decision-making tasks that are highly dynamic and time variant, where

User Interface

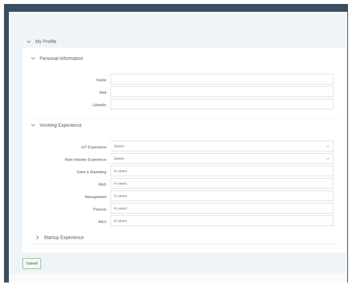

Figure 1: User profile for capturing domain expertise that is required for matching users with tasks.

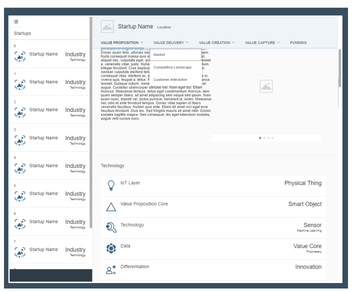

Figure 2: List of similar startups (left) that allows the user to compare startups and startup data along its business model dimensions.

training data might not be representative for real applications, only few data is available, or processing soft information and human gut feeling is necessary to predict extremely uncertain outcomes [3].

Previous research has shown that humans and machines have complementary benefits, which allows to leverage human intelligence to augment AI applications during training [e.g. 2, 7]. While human users are good at generating diverse ideas, or can annotate arbitrary data, machines are particularly superior in weighting multiple criteria and provide deep insights in the complex decision rules of human intuition. However, these interactive machine learning architectures fail short for highly complex and uncertain decision-making tasks that require continuous input of domain experts. Those, were not fully captured in previous applications of interactive architectures that leverage crowdsourcing (e.g. [1, 8]).

The purpose of our research is to propose a scalable hybrid intelligence system that combines collective human and machine intelligence to achieve better results than each of them could do in separation. For solving such highly uncertain tasks, the system architecture empowers a collective group of domain experts to teach an artificial intelligence by generating features, labelling, and actively debugging the learner. Vice versa the system provides feedback to the user to improve task performance over time. The long-term aim is therefore to develop a hybrid intelligence that has a better “gut feeling” then the best individual human expert. This prototype allows to draw generalizable conclusions that can be leveraged for designing hybrid intelligence architectures for several other complex decision-making problems requiring continuous learning (e.g. medicine, recruiting etc.).

On representative context for such problems is startup investing (i.e. finding unicorns). Those investment decisions are inherently complex as they require to assess large quantities of data including multiple data types that is often subject to interpretation to infer future success of an investment. This results time delayed feedback on the quality of a decision [5]. Such complex and highly uncertain tasks require so called gut feeling, a combination of both intuitive and analytical decision making [6] that can only be achieved by experts with extensive domain expertise. These demands make it highly difficult for machine learning techniques to predict. On the other hand, some human super forecasters proved to be astonishing successful in predicting outcomes that seem to be unpredictable [5].

System Description

Using humans for labelling tasks typically assumes that non-experts are crowdsourced for creating noisy labels with ground truth (e.g. labelling a cat as a cat), which does not require a high amount of domain knowledge. Our assumption, however, is that complex decision-making tasks require the subjective judgements of expert crowdsourcing that is based on human gut feeling. Such users are benevolent, error prone experts (some are better than others) but have individual biases and limits [6]. Moreover, our hybrid intelligence architecture capitalizes from the rational of collective intelligence [10]. The system actively uses the diverse knowledge and experience of users and aggregates their knowledge while it reduces the biases of individual decision makers. As there is no ground truth for

Crowdsourcing Task

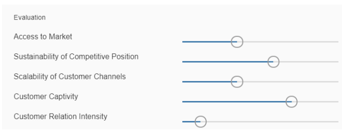

Figure 3: The analytical judgement of the startup is provided by 10 point Likert scales. The same mechanism is applied for troubleshooting.

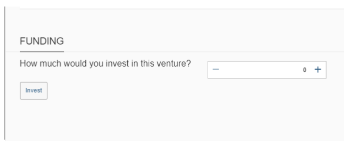

Figure 4: The users provide an intuitive judgement through a funding mechanism that allows them to invest a budget from 0 to 100.

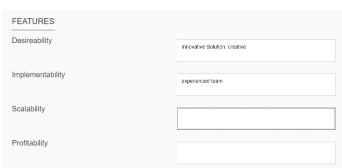

Figure 5: Features for each judgement dimension are provided in textual form to allow the machine learning engineer adapting the ontology and features that are applied.

judging startups, aggregation reduces error and is informative, which optimizes the labelling quality. Finally, interactive machine learning mostly uses human input for training. Our system however focuses on continuous learning from human input, which is required for such AI applications in the wild.

User Interface

The users can provide profile information capturing their domain knowledge and expertise. Relevant startup data is visualized and divided in different parts for different expertise requirements along an ontology that translates each dimension in proxies in machine readable format. This creates a shared understanding between humans and the learner. Each user is then provided with a list of startups with varying similarity. This is inspired by analogical encoding that enables a human to recognize similar attributes for identifying differences more deliberate judgements [4].

Crowdsourcing Task

The user workflow of the crowdsourcing tasks consists of four main components:

- matching each dimension of a startup with a crowd of users (approx. 5) that have the required domain knowledge. To leverage the advantages of collective intelligence, the system balances exploiting experts from the same industry and exploration (users from a different, related industry).
- generating learning features that experts use to make a judgement to adapt the ontology when decision criteria change. This approach allows to deal with time variant features in machine learning.
- the labelling task that requires the user to start with an intuitive, holistic assessment of the startup by providing a funding mechanism (from 0 to 100), which represents the overall degree of belief that a startup will succeed. The intuitive assessment is then proceeded with an analytical judgment sequence that requires the user to judge each startup along several criteria dimensions on a 10-point Likert scale.
- the troubleshooting task allows the learner to actively ask members of the crowd to debug the predicted judgement, when it is uncertain. The user can then adapt the predicted values.

Machine Learning Architecture

The machine learner includes several algorithms that learn from the human user input. First, it learns what features are perceived as relevant from human decision makers and how they judge certain values (e.g. how is a team of 3 novices perceived on a scale from 1 to 10). This allows the learner to study both the analytical and intuitive judgement rules of human experts as well as new features that are relevant for such decisions. Moreover, an uncertainty sampling active learning algorithm allows the learner to actively ask for troubleshooting. Finally, the learner captures how the input of each user should be weighted according to their expertise and skills to optimize label quality and thus prediction accuracy.

Feedback

Finally, the user receives feedback (cognitive and outcome feedback) [9] on the judgement task visualizing the individual analytical and intuitive assessment of each startup compared to the aggregated score of the crowds´ judgement (i.e. the

User Feedback

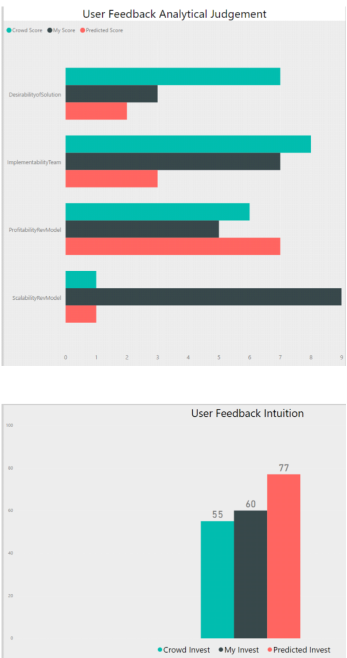

Figure 6: The users receive feedback on their decision rules (cognitive feedback) and their confidence (outcome feedback) compared to the learner´s prediction and the crowds´ collective judgement.

collective assessment) as well as the learner's prediction. This allows to teach the user and increase decision-making performance over time.

General Implications for CHI

The general goal is to design a hybrid intelligence system that can achieve superior performance by combining the complementary benefits of humans and machines through augmenting an artificial intelligence through crowdsourcing. We thus apply several rationales that were not fully discovered so far. First, we propose a hybrid architecture that allows to teach an artificial intelligence "gut feeling", which is defined as a combination of intuitive and analytical decision-making and adapt to complex problems under extreme uncertainty. Furthermore, this allows us to get deep insights into human intuition and decision rules as well as how they can be taught to a machine. Such architectures might be relevant for several applications in dynamic real-world settings. Second, we propose a task design for labelling situations without ground truth that starts the relevant cognitive procedures to allow better quality screening. Third, we explicitly consider the differences of individuals in the crowd and leverage the benefits of diverse skills and expertise for machine learning labelling tasks. Fourth, our system allows to examine the effect of feedback from different sources (i.e. other members of the crowd or AI predicted) on learning and the long-term performance of individuals in highly complex decision-making tasks.

References

[1] J. C. Chang, S. Amershi, and E. Kamar, Eds. 2017. *Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets*. ACM.

[2] Justin Cheng and Michael S. Bernstein, 2015. *Flock: Hybrid crowd-machine learning classifiers*. ACM.

[3] Dominik Dellermann, Nikolaus Lipusch, Philipp A. Ebel, Karl M. Popp, and Jan M. Leimeister. 2017. Finding the Unicorn: Predicting Early Stage Startup Success through a Hybrid Intelligence Method. *Proceedings of The International Conference on Information Systems, Seoul*, So

[4] Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. *Encyclopedia of human behavior* 130, 130.

[5] Laura Huang. 2017. The Role of Investor Gut Feel in Managing Complexity and Extreme Risk. *Academy of Management Journal,* 2017.1009.

[6] Daniel Kahneman and Amos Tversky. 2000. *Choices, values, and frames*. Cambridge University Press.

[7] Ece Kamar. 2016. Hybrid workplaces of the future. *XRDS: Crossroads, The ACM Magazine for Students* 23, 2, 22–25.

[8] B. Nushi, E. Kamar, E. Horvitz, and D. Kossmann, Eds. 2017. *On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems*.

[9] William Remus, Marcus O'Conner, and Kenneth Griggs. 1996. Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes* 66, 1, 22–30.

[10] Anita W. Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004, 686–688.