The predictive reframing of machine learning applications: good predictions and bad measurements (Alexander Martin Mussgnug 2022)

I. Main claim: ML should be considered as ***automatically-calibrated model-based measurements*** rather than ***prediction***.
    A. Why? – An interpretation of supervised machine learning applications to measurement tasks as automatically-calibrated model-based measurements internalizes questions of <u>construct validity</u> and <u>ethical desirability</u> critical to the measurement problem these applications are intended to and presented as solving
II. Part 1. Predictive Reframing
    A. Claiming supervised ML is neither ethically nor epistemically neutral
        1. Why?
            a) With the term "prediction", we associate inferences about the future
            b) During development, one might understand the inferences of supervised ML models as predictions: the model predicts a value in a dataset that is later disclosed to it.
            c) However, in applying these models people should not claim prediction
            Ex: "poverty prediction" → they are predicting Global MPI, which is a measurement of poverty and not prediction of poverty
    B. This is not just an act of lexical convention
        1. Reframes epistemic aim
        2. Claim their supervised ML outcomes as second-order rather than first-order

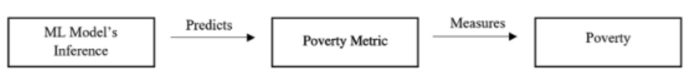


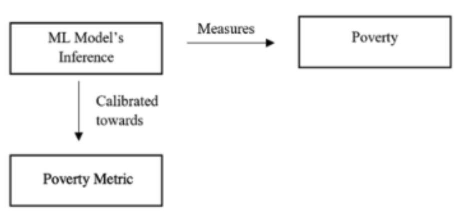

**Fig. 1** The Predicting Reframing of the Original Measurement Task

III. Part 2. Problems
    A. ML models do not question whether the given measurement is the right measurement for the context. Ex: a particular way of measuring poverty is taken for granted
    B. ML people should communicate limitations
        1. acknowledge the implications of the predictive reframing and, thereby, appropriately relativize the ML model's predictions
        2. However this is not enough
    C. Residual errors
        1. Evaluation 2-step
            a) Accuracy in reproducing measurement
            b) That reference measurement is appropriate

        2. Shortcoming: the relation between the ML model's prediction, the reference measurement, and the latent property of interest is not necessarily transitive
        3. Understanding supervised ML applications as predictive limits the advancement of the discipline
            a) Epistemic aim: predicting the results of a given reference procedure → improvements in accuracy: reducing errors with respect to the primary epistemic aim

IV. Part 3. Automatically Calibrated Model-based Measurement
    A. Calibration process
        1. Example: Watch
            a) Calibrate the watch to the atomic clock
                (1) Not predicting the reference (such as predicting the hypothetical measurement of the atomic clock or predicting the performance relative to the set) But measuring the property of interest
        2. Two step process
            a) Forward calibration - iteratively determines the relationship between quantity values provided by a reference procedure and the indications of the instrument being calibrated.
            b) Backward calibration - infer the measurement outcome from an instrument indication
        3. In ML terms: training – calibration of a measurement model & test – to see if it generalizes
    B. Supervised ML models are automatically calibrated model-based measurements
        1. Interpret their application as measuring rather than predicting a hypothetical measurement

**Fig. 2** Supervised ML as automatically calibrated model-based measurement




V. Conclusion: a closer association of supervised machine learning and metrology can help expand machine learning developers' toolkit when working toward epistemically and ethically more successful application

VI. Discussion Points
    A. Returning to the poverty example, what would be the difference when we see this as a measurement?
        1. I would say there is not that much difference?
    B. Isn't measurement a stronger claim with less uncertainty than prediction?
    C. Not all ML is about prediction, some are classifications, generation, etc. Do we see any similar issues with different claims?
    D. The author states "understanding supervised ML applications as predictive limits the advancement of the discipline" how does it do so? Do we see this?

**NOTES FLORIAN BOGE - TWO DIMENSIONS OF OPACITY**

GUIDING QUESTIONS *Is the sort of knowledge that can be produced by deep learning methods deployed in science different from classical methods? Will the widespread adoption of ANNs/DNNs disrupt scientific practice? Can we derive scientific understanding and explanation from the use of deep learning methods?*

Boge argues that there are 2 aspects in which scientific models can be instrumental, 2 aspects in which scientific models can be opaque

**C and R Instrumentality**

R-instrumentality is making idealisations that are counter to fact

C-instrumentality refers to contentlessness of various features of the model

**2 kinds of Opacity** - “how opacity” and “what opacity”.

The first dimension of opacity, model-relative opacity, is about the complexity of the internal workings of the model; we do not know how it arrives at its particular conclusions; the internal mechanisms \textit{of the model} are uninterpretable.

The second dimension of opacity is meant to be indicative of a lack of insight into what sorts of regularities are indeed being picked up on by the model or a lack of insight into what features in the data drive its outputs

*“When a DNN learns to approximate a desired function, it is hence not only opaque how, precisely, it achieves this goal: It is also opaque what it is about the data that drives this process”*

Boge acknowledges that some “traditional” models in science can be instrumental in both r-instrumental and c-instrumental senses and that some “traditional” computer simulations can exhibit h-opacity; however, Boge claims w-opacity to be unique to DL:

*“w-opacity is, ultimately, the distinctive factor which sets DL apart from all traditional models and, eventually, impairs our ability to acquire scientific understanding in a special way.”*

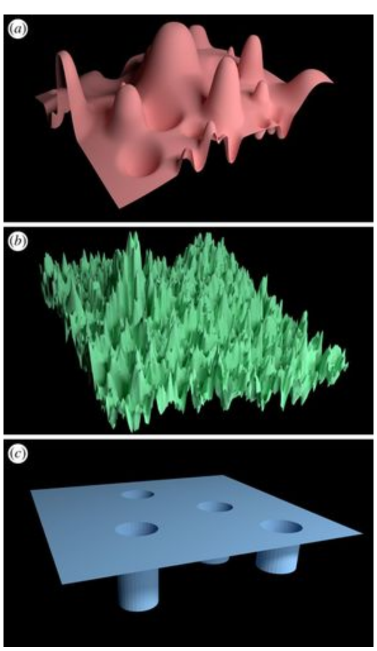
The unique combination of c-instrumentality and w-opacity in DL is meant to be prohibitive to arriving at explanation/understanding, especially when using *unsupervised* learning in *exploratory* contexts

Exploratory modelling contexts are those which Boge defines to be unguided by strong theoretical commitments; ones in which the conceptual schema which serve to make sense of the phenomena under study have yet to be defined. In such contexts, the usage of models which are both instrumental in the sense of some of their (formal) features being contentless and opaque in the sense that one cannot read off from the results of the modelling procedure what features of the data were decisive in lending that result Boge takes to be an impediment to understanding.

- Critiques a culture of collecting large amounts of data in order to answer scientific hypotheses that was developing (and continues today), mostly focused on biological fields - they call this **radical empiricism**
    - Argues that theoretical approaches are also needed, and that scientists should not just focus on predictive accuracy [contrast with earlier two cultures of statistical modelling paper]
    - Unlike the statistical modelling paper, the theory that the authors suggest we need is not about any statistical guarantees or features of the model, but more focused on attempting to understand underlying phenomena in biology

**Features of modern science**

- "Our ability to model Nature far outstrips our ability to conceive it; indeed, only the most mundane aspects of Nature are intelligible, those corresponding to our everyday experience".
- They list four properties of modern science as they conceive it:
    1. It takes the form of a model formulated in a mathematical system
    2. Precise relationships are specified between terms in the theory and measurements of corresponding events
    3. There are validating experimental data that confirm the theory in the future
    4. A rigorous statistical analysis supports acceptance of the theory based on concordance between the predictions and the measurements
- They note that biological and medical sciences are not quite "modern" in this sense, because they typically rationalize observations after the event rather than build predictive theories
- Although data gathering may be good for the sake of modelling in general, the data gathered has to be reliable, and not haphazardly collected
    - Replication crisis in medicine - 'most published research findings are false'
    - Gathering data with a lack of theoretical insight may not give any results - e.g. human genome project has gathered a great deal of data but not much to show for it
- There may also be pathological loss landscapes for which ML doesn't work well such as b) and c) below




**Radical empiricism and biology**

- Since ML models fit to data, they can only extrapolate outside their training distribution to a limited extent depending on how similar the test distribution is
- Local data e.g. about different biological systems may be hard to interpret without a global theory of how these data should be viewed together (parable about blind men and an elephant)
- Vastness of feature space in biology poses an especially large challenge - still don't know how to use an individual person's genome to create a bespoke treatment for them

**Dynamics and chaos**

- Although they have been critiquing biology, they note that there is some theory pertaining to population dynamics, biophysics, epidemiology, ecology, evolutionary theory etc
- However, biological theories don't reach the level of generality and power seen in physics due to a vast state space and chaotic behaviour
    - (there are actually multiple defns of chaos out there, but here's the most common one in a nutshell) a chaotic system is one that is dependent on initial conditions and whose trajectories seem 'unpredictable'
        - Many physical systems turn out to be chaotic in this sense, which makes it hard to study them (this area of math is rather new and not well developed yet compared to others)

**Conclusion**

- Somewhat pessimistic, it seems like at least in 2016 the collection of big datasets and ML have not really helped researchers in biology and medicine much
    - Not sure how accurate this is, don't personally know the impact in these fields myself…
- However, even if it doesn't help provide a scientific understanding, ML can still be useful, e.g. models that synthesize new compounds or materials
    - However, these models are uninterpretable, so it's hard for researchers to learn anything from them
- Although they focus on biology, the situation may be similar for many sciences that employ machine learning, as well as perhaps attempts to understand machine learning itself