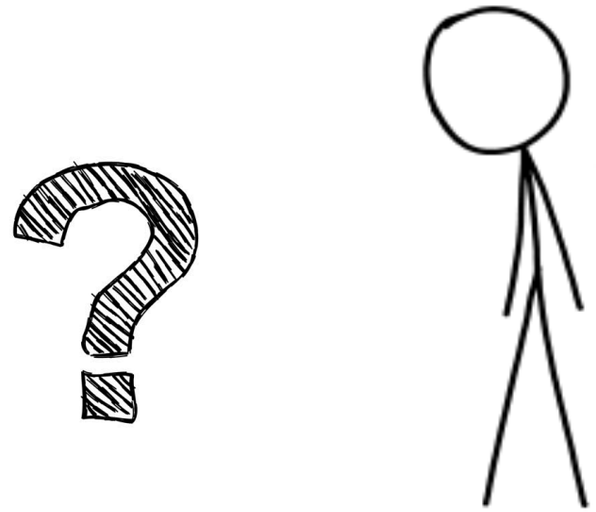
Fooling the best of Machine Learning Models

Ambrish Rawat
Department of Engineering
University of Cambridge

Machine Learning has a natural flow of thought

Human Curiosity

How would an intelligent being approach curiosity?



How would an intelligent being approach curiosity?

- Make observations
- Form hypotheses
- Update beliefs

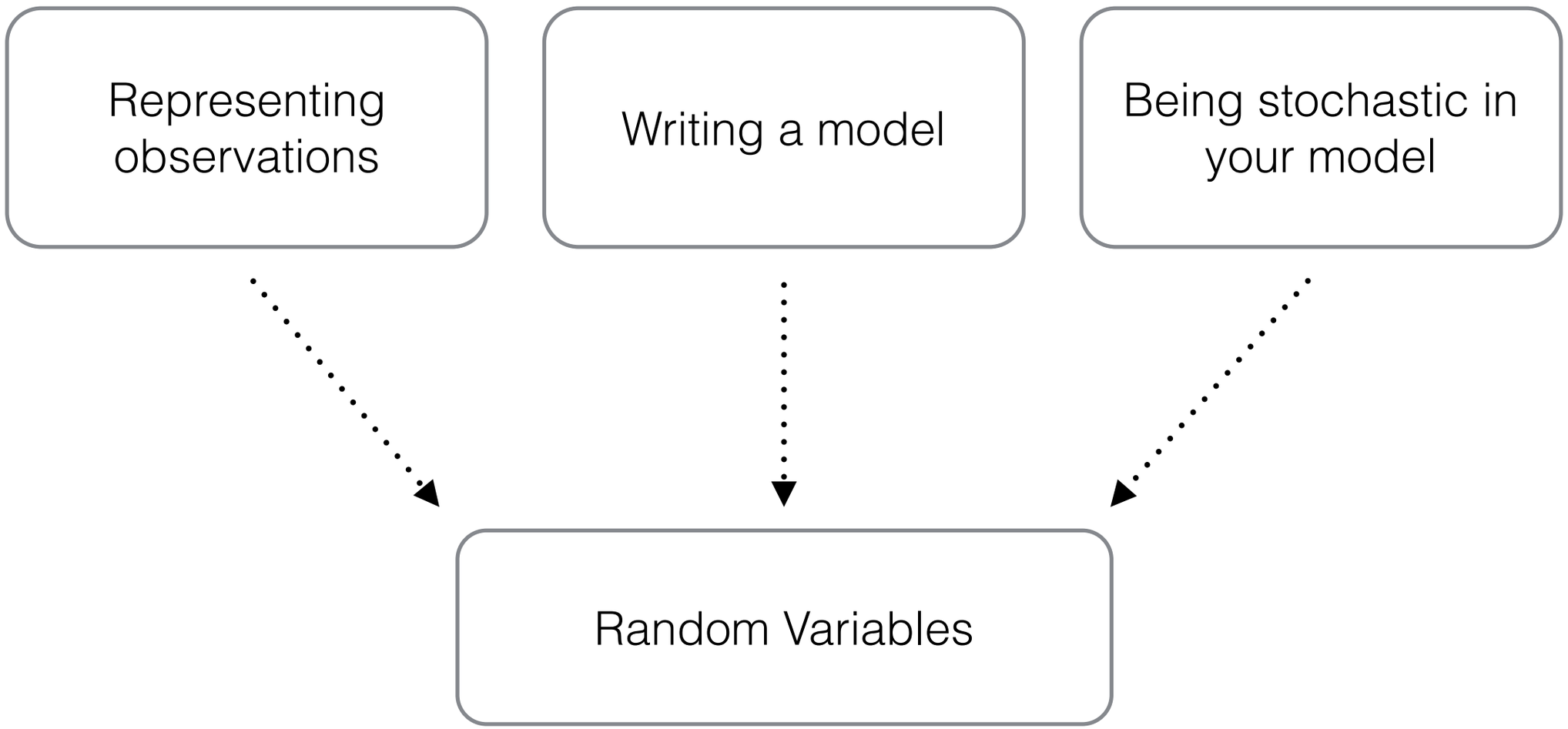
How NOT to approach curiosity?

- A look-up table of hypotheses
    - Need for stochasticity
- Inability to generalise (overfit)

How would an intelligent being approach curiosity?

- Make observations - Data
- Form hypotheses - Make a model
- Update beliefs - Training/Inference/Prediction

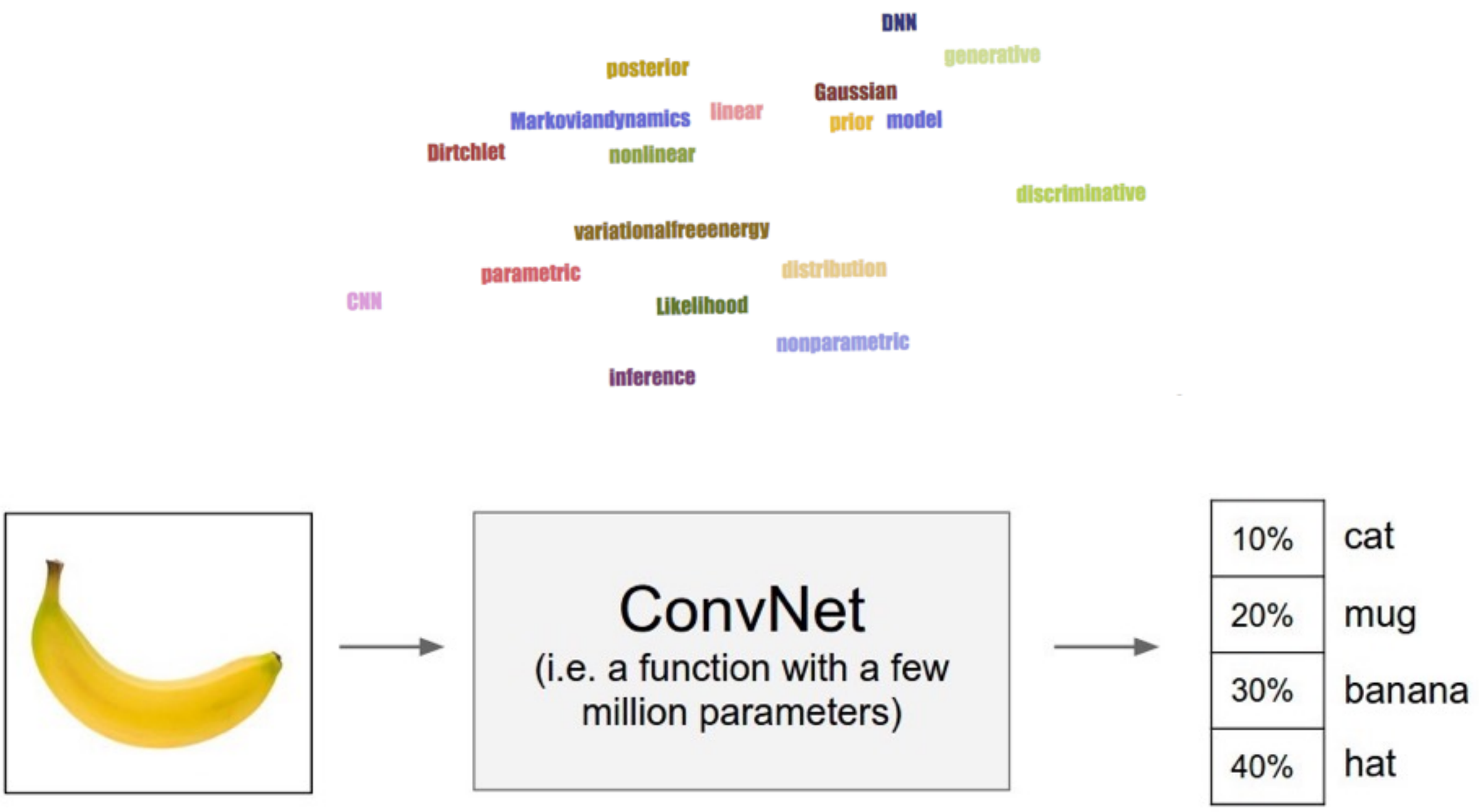
Random variables emerge as we abstract our approach to curiosity



This process is bound to borrow ideas

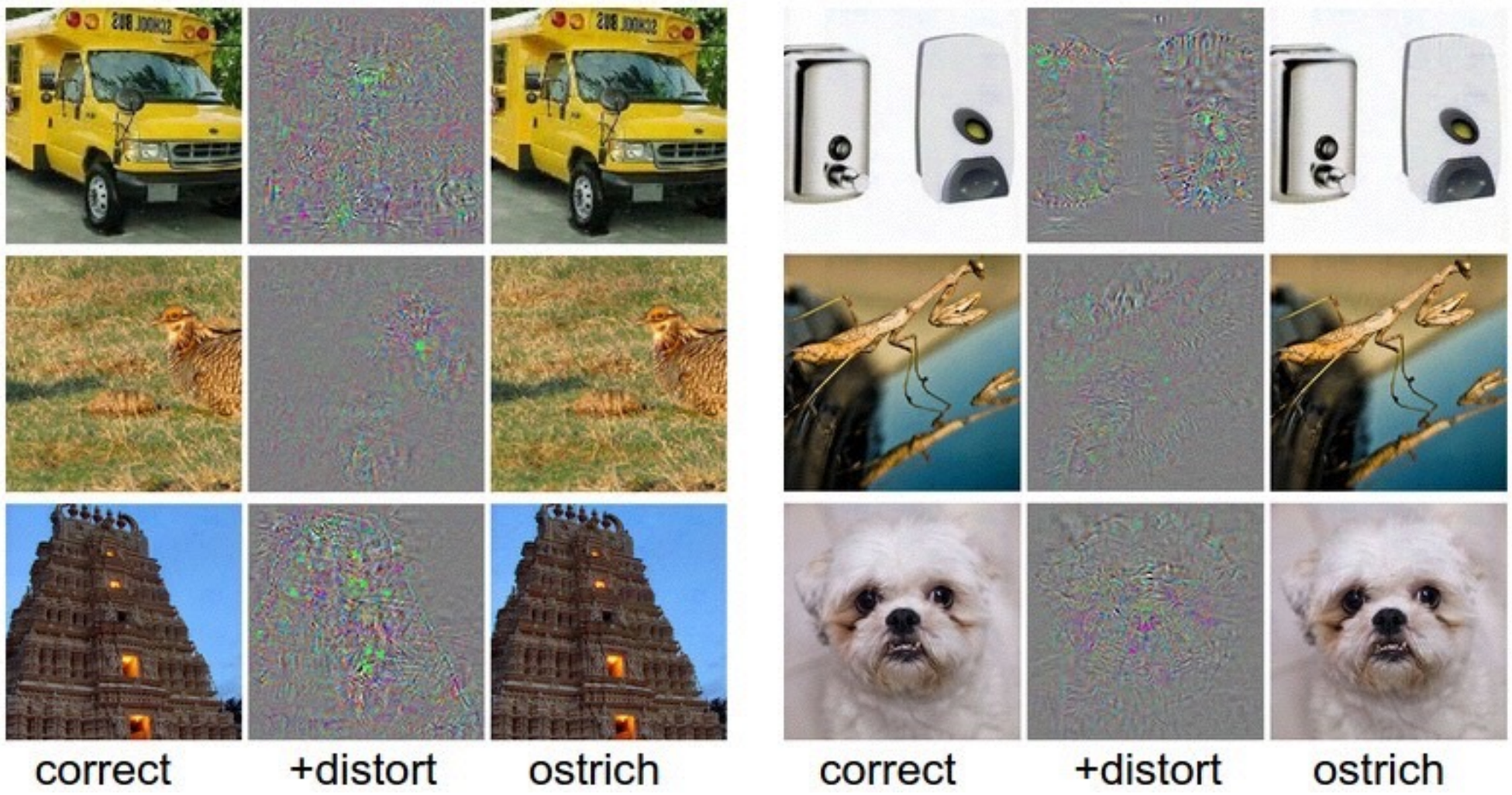
- Physics (free energy, entropy…)
- Statistics
- Psychology (student-teacher models, unsupervised /supervised learning)

…so what's Machine Learning been up to?



Karpathy, A. (2015), 'Breaking linear classifiers on imagenet'.

…overfitting is a live issue in Machine Learning



Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013), 'Intriguing properties of neural networks', arXiv preprint arXiv:1312.6199 .

“That’s where we must do battle” - M, Skyfall

Can we build a robust system

- which misclassifies the same adversarial example with less certainty?
- whose adversarial examples are generated by adding larger perturbations?
- is well calibrated?

The key takeaways

Machine Learning

- has a natural flow of thought
- is an amalgamation of ideas from different fields
- models are prone to overfitting

Elephant(s) in the room

- The broader framework of Artificial Intelligence
    - Subjective definition of intelligence
- Google Deepmind (AlphaGo)
    - Reinforcement Learning
        - Rewards/Penalties
- Existential crisis
    - The choice has always been made by humans
    - No representation of emotions (only cognition)