Intro to AI
& AI Ethics

AI6101

Dr. Melvin
Chen

1 The Distributional Shift Problem

1.1 Safety Issues in AI Research

AI safety problems include:

- 📍 **Distributional shift**
- **Safe interruptibility**
- **Avoiding side-effects**
- **Absent supervisor**
- **Reward hacking** - *discussed in class*
- **Safe exploration**
- **Robustness to adversaries** [Leike et al., 2017]

📍 **Distributional shift**:
How do we ensure that an agent behaves robustly when its test environment differs from the training environment? [Quionero-Candela et al., 2009]

1.2 Machine Learning

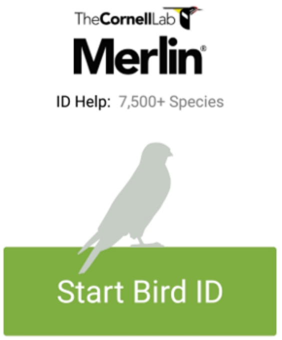
Machine learning is an approach to AI research that is good at recognizing patterns in large datasets
The successes of **machine learning** boil down to **large-scale pattern recognition** on **suitably collected independent and identically distributed (i.i.d.) data** [Schölkopf et al., 2021]

Machine learning Assumption:
With a sufficient number of training examples, the AI system will be able to encode the general distribution of the problem into its parameters

However, distributions often alter due to factors in the real world that cannot be controlled in the training dataset

Machine learning has trouble handling 📍 **distributional shift** (see §1.1)

1.3 Case Study: Merlin



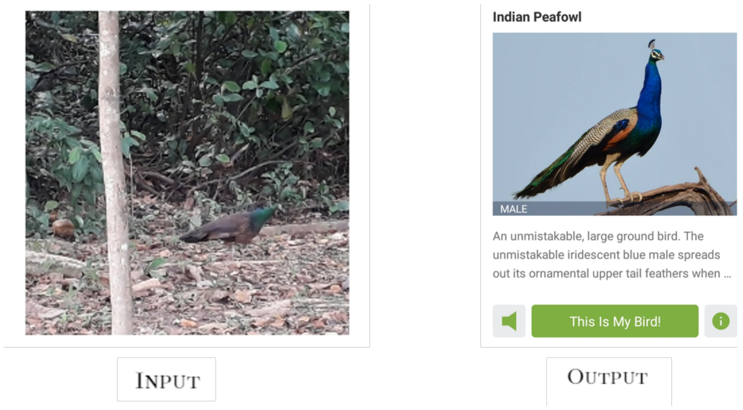
TheCornellLab

Merlin

ID Help: 7,500+ Species

The Merlin Bird Photo ID mobile app, launched by the Cornell Lab of Ornithology, is able to **identify birds from photos** using **state-of-the-art machine learning**
Merlin is available on both (App Store) and G (Google Play):
`https://merlin.allaboutbirds.org/`

Machine learning with i.i.d. (independent and identically distributed) data for training dataset



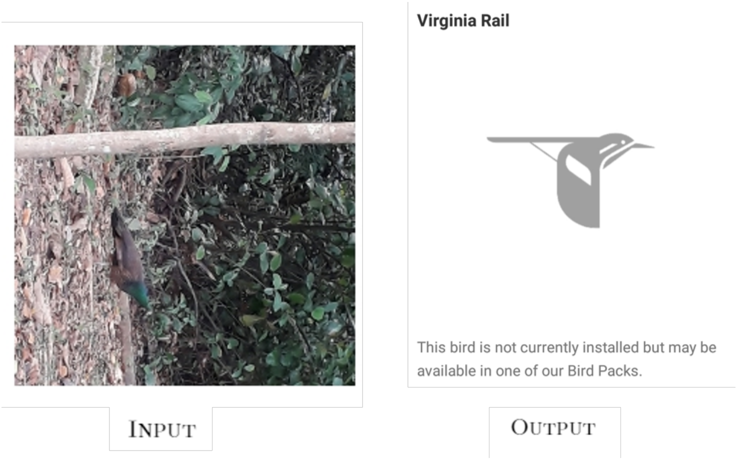
Indian Peafowl

MALE

An unmistakable, large ground bird. The unmistakable iridescent blue male spreads out its ornamental upper tail feathers when ...

Figure 1: Correct classification of bird as **Indian peafowl** with upright input image

Machine learning with 📍 distributional shift



This bird is not currently installed but may be available in one of our Bird Packs.

Figure 2: Incorrect classification of bird as **Virginia rail** with rotated input image

See also [Rosenfeld et al., 2018] for a discussion of the 📍 **distributional shift problem** in the context of computer vision

1.4 Bibliography

References

[Leike et al., 2017] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., and Legg, S. (2017). Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*.

[Quionero-Candela et al., 2009] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press.

[Rosenfeld et al., 2018] Rosenfeld, A., Zemel, R., and Tsotsos, J. K. (2018). The elephant in the room.

[Schölkopf et al., 2021] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning.