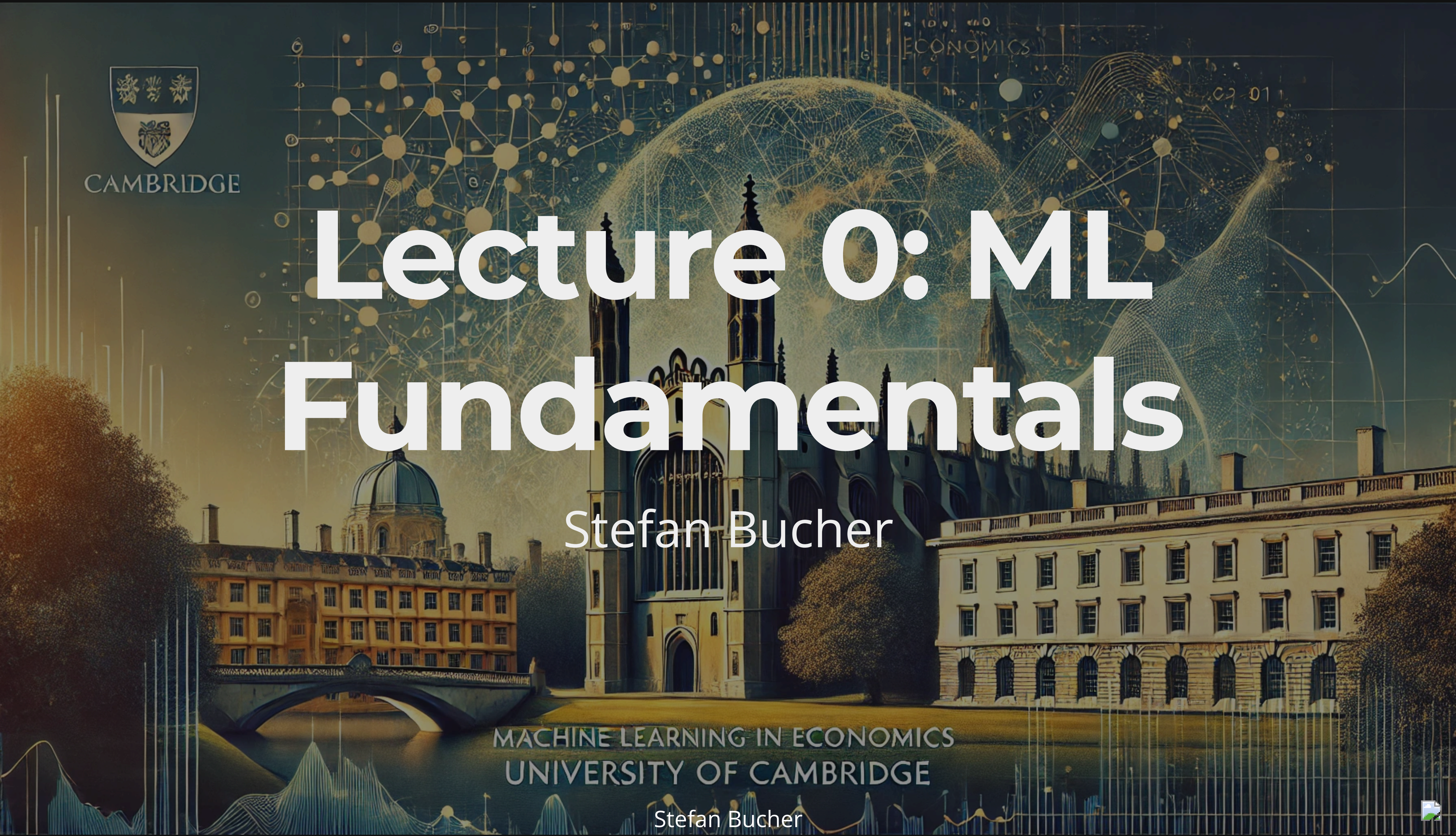


Lecture 0: ML Fundamentals

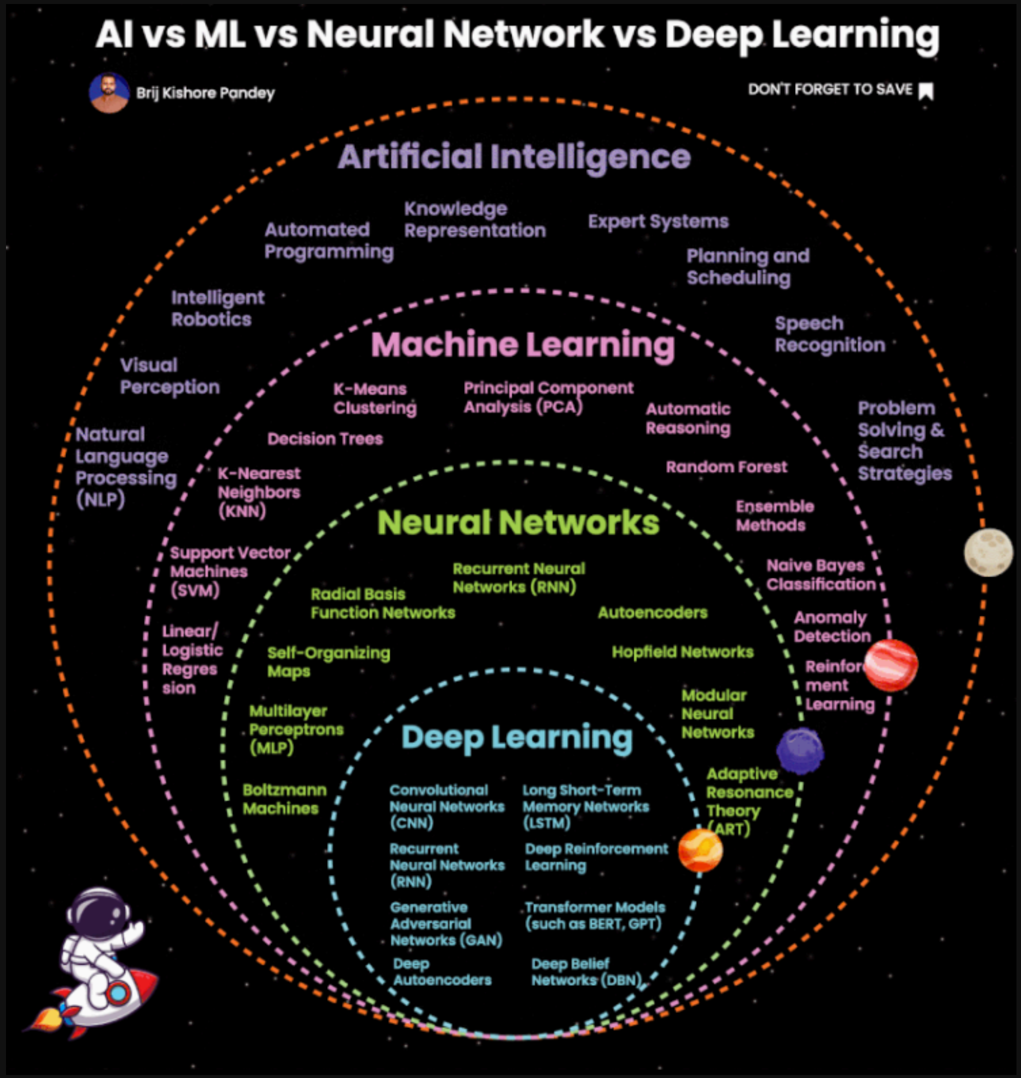
Stefan Bucher

Introduction to ML

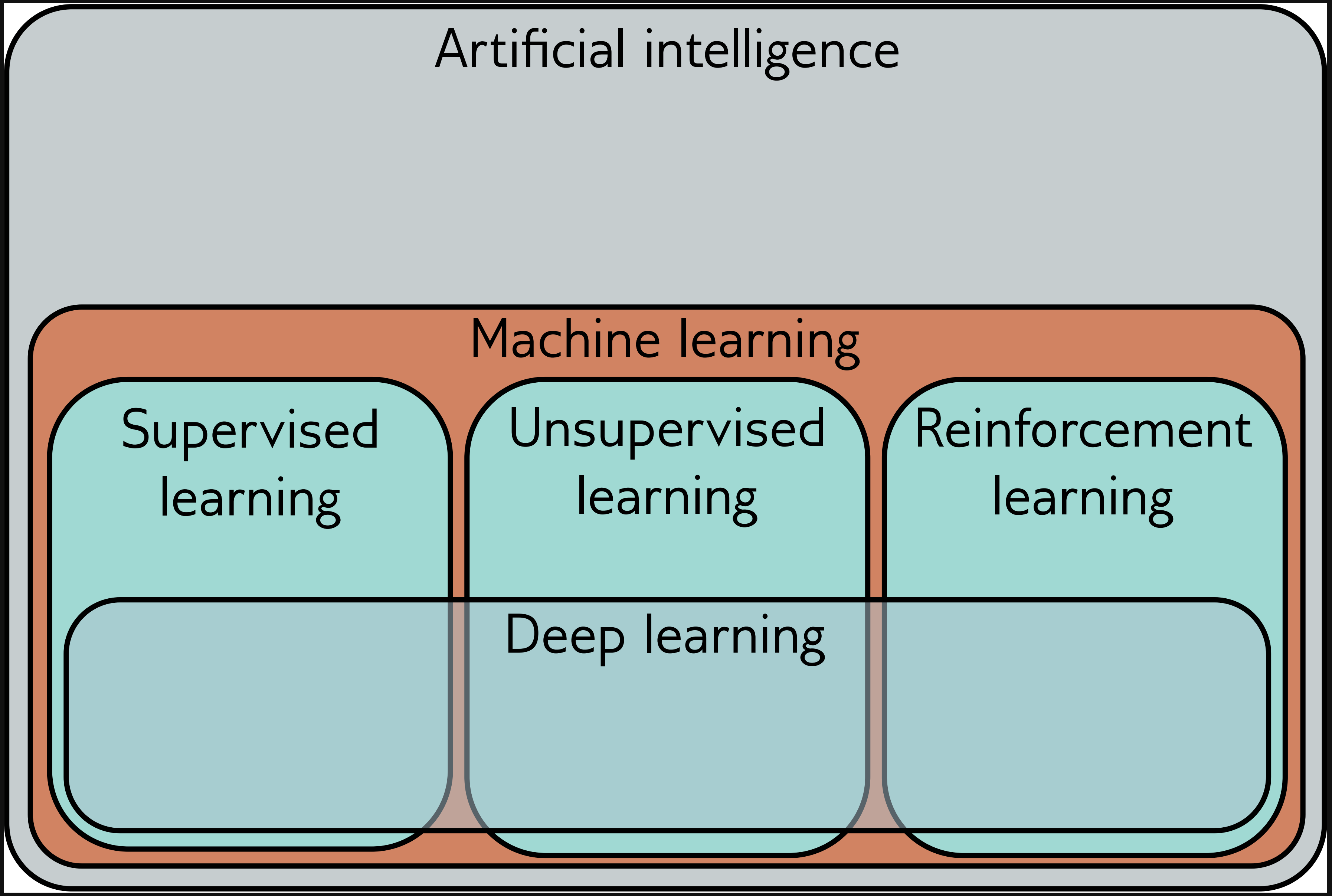
Prince (2023, chap. 1 and Appendix C).[1]

1. Figures taken or adapted from Prince (2023). All rights belong to the original author and publisher. These materials are intended solely for educational purposes.

What is Machine Learning (ML)?

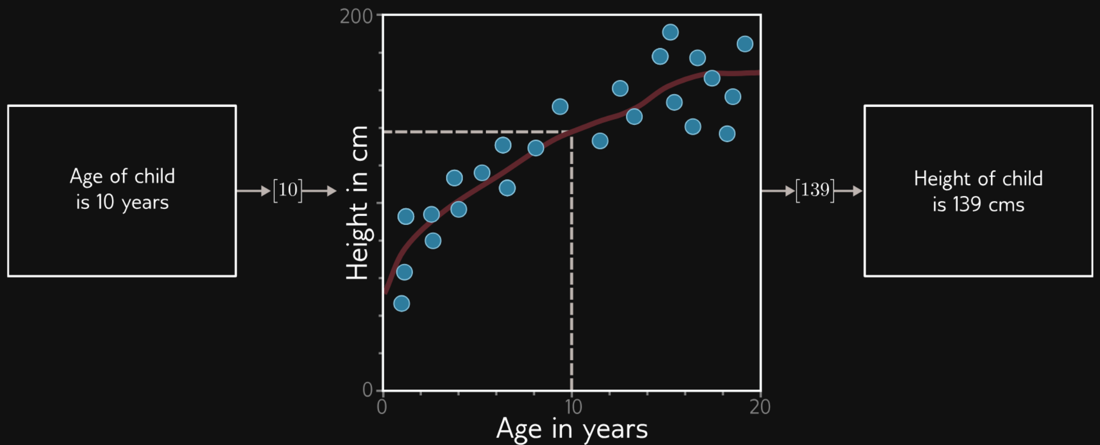


The Three Pillars of ML

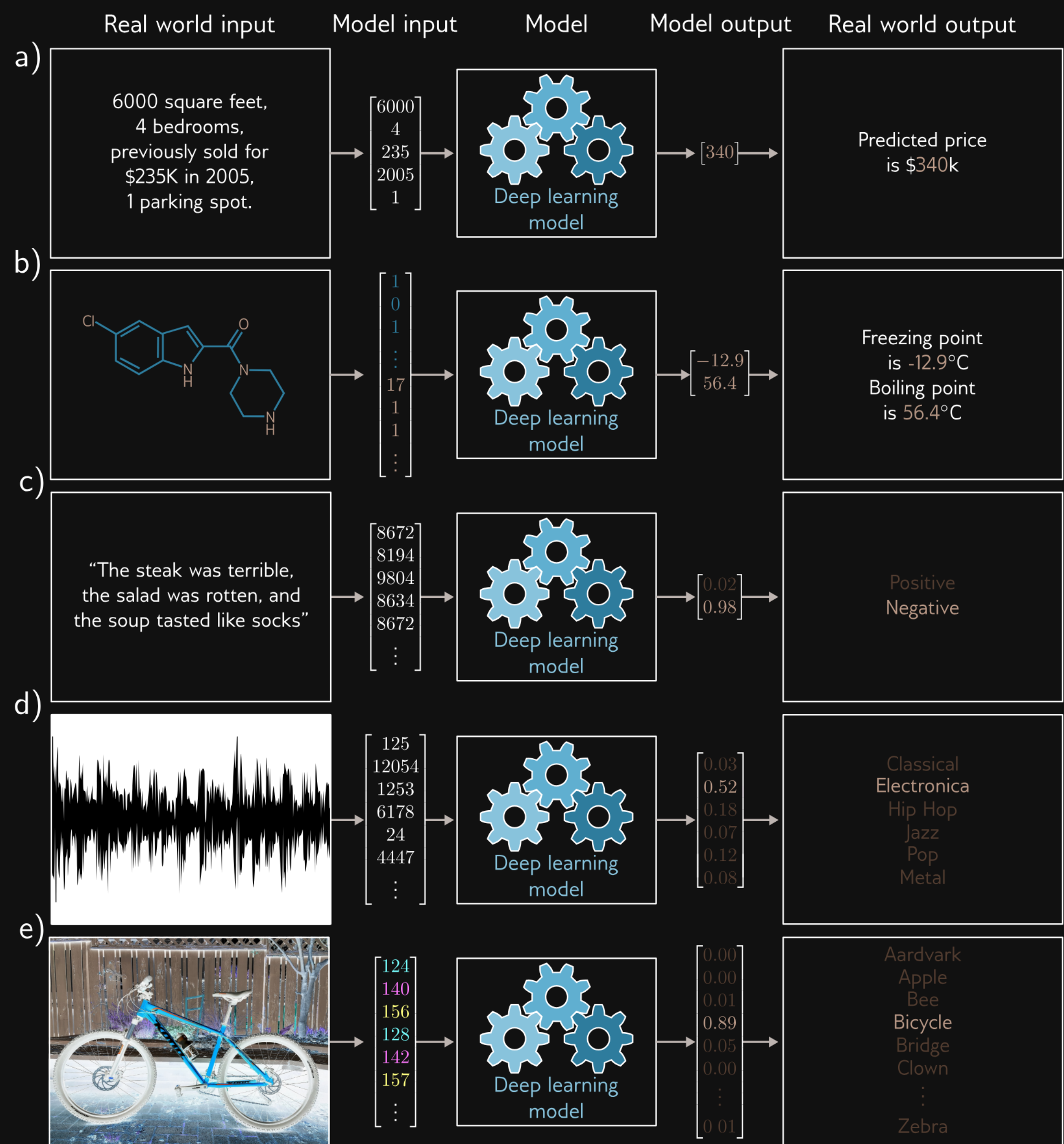


Supervised Learning

- Economists use regressions to predict outcomes
- ML does the same
- Key: Learning input-output mapping



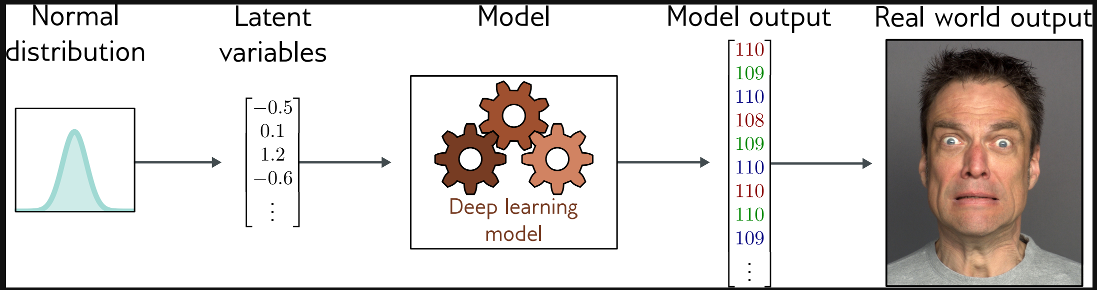
Regression & Classification



Stefan Bucher

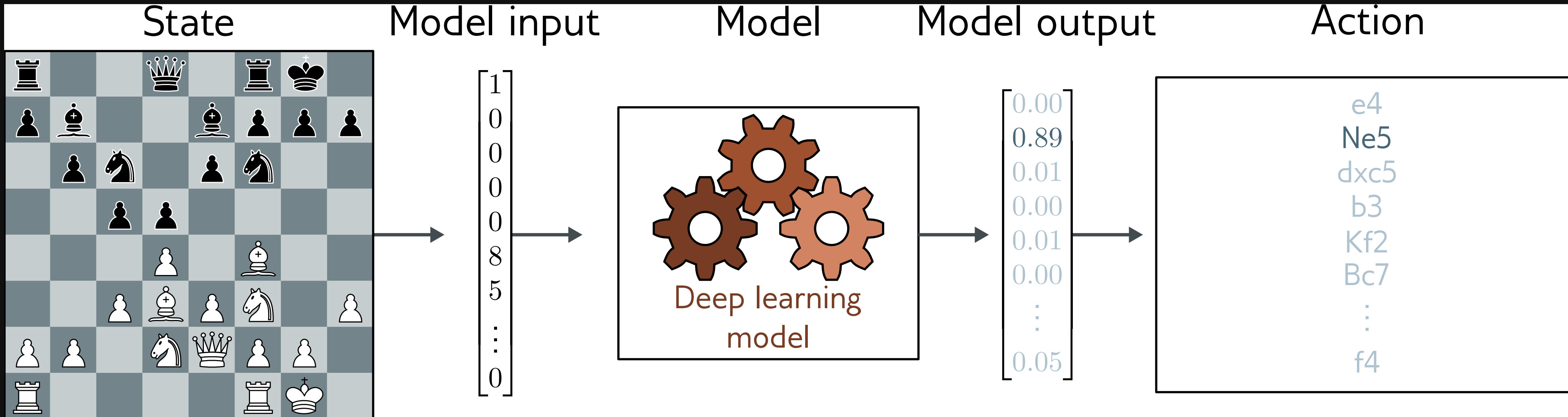
Unsupervised Learning

Key: Learning a distribution



Focus on generative unsupervised models

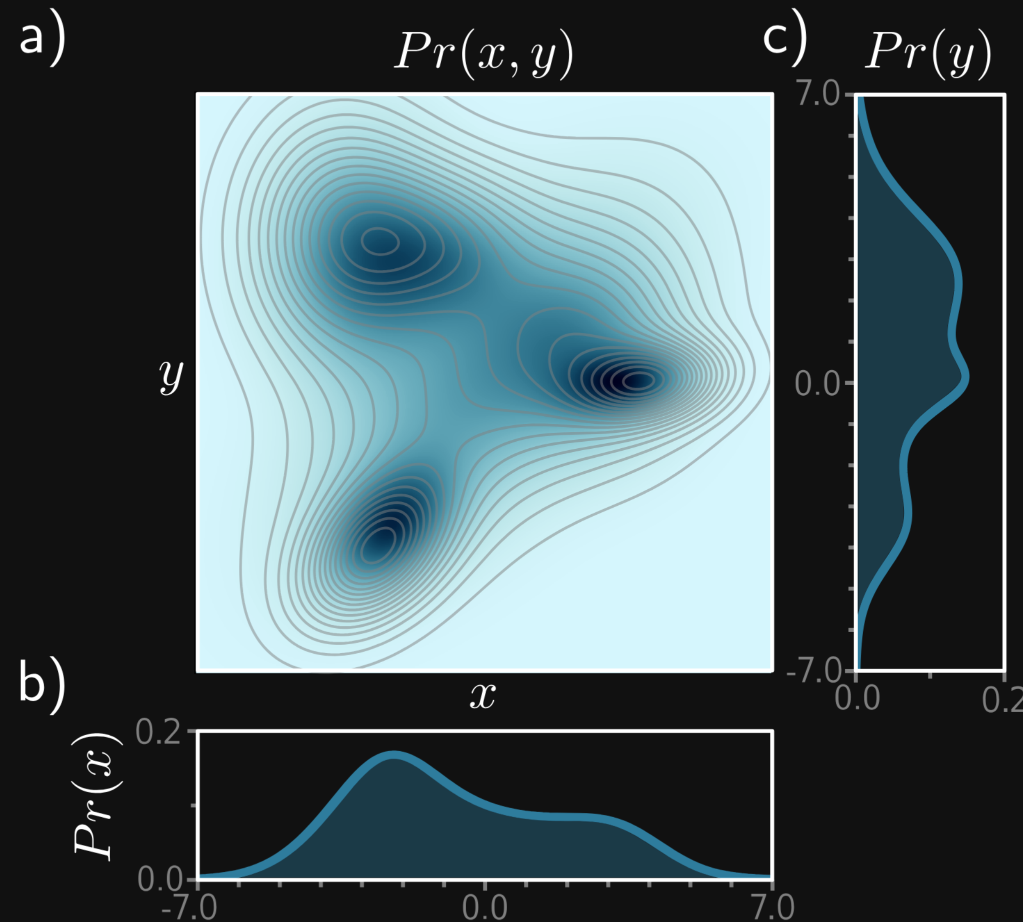
Reinforcement Learning



Key: Learning an action policy

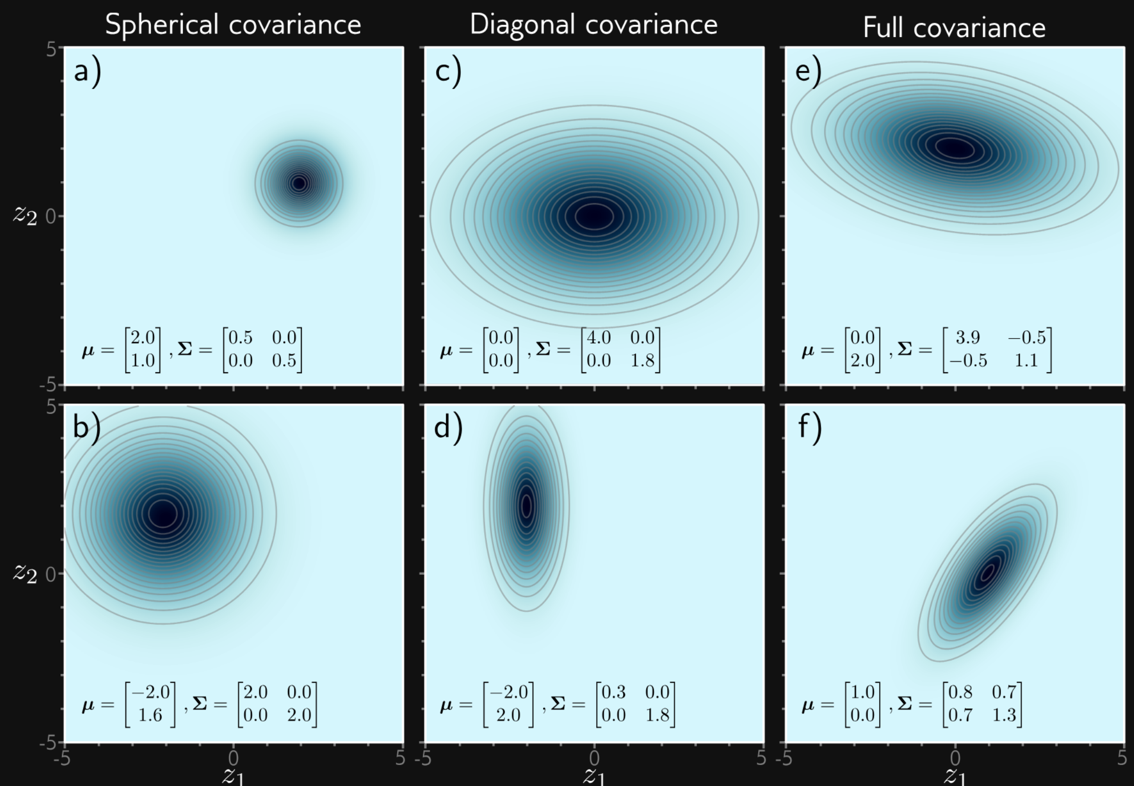
Probability and Information Fundamentals

Joint and Marginal Distributions



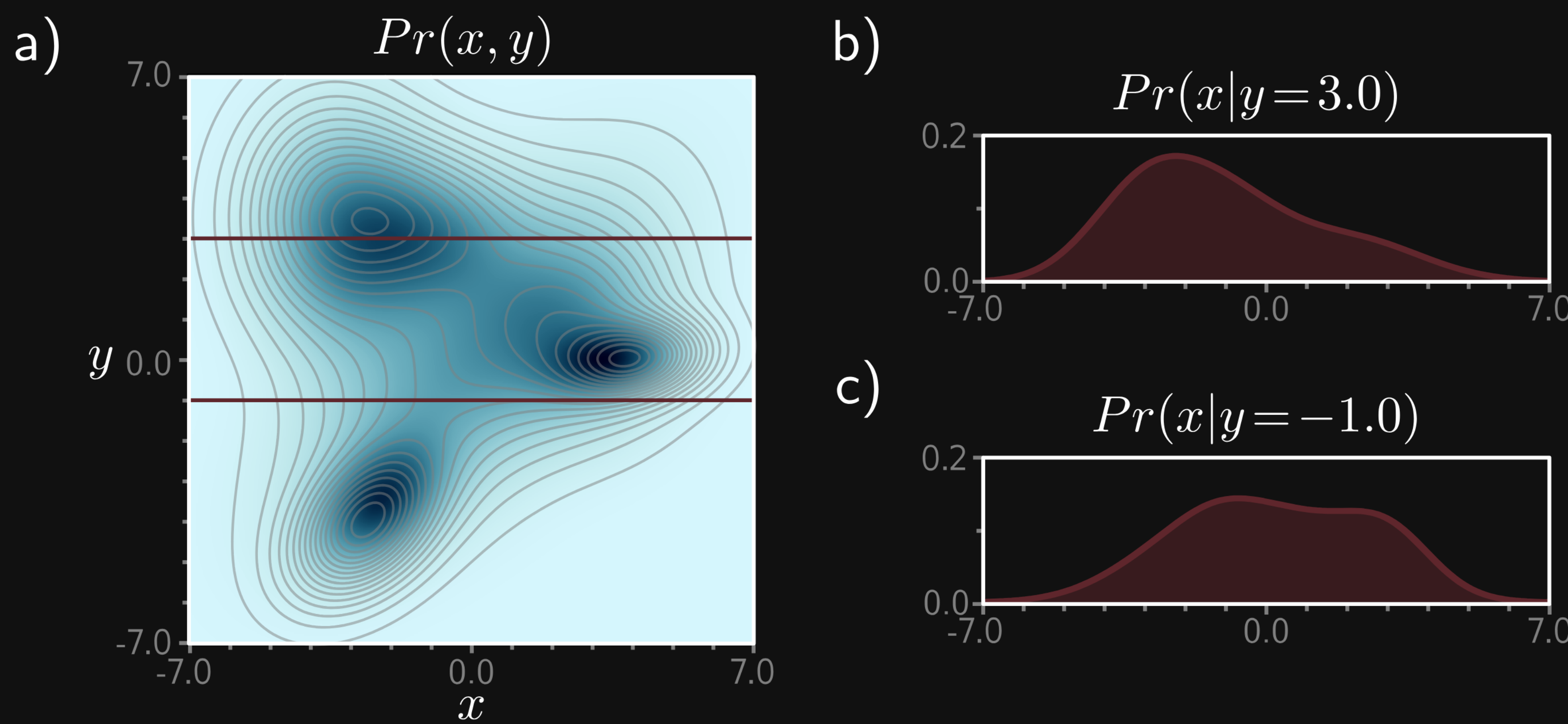
$$P(x) = \int P(x,y)dy$$

Bivariate Gaussian



$$P(\mathbf{x}; \mu, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mu)\right]$$

Conditional Dist. & Bayes' Rule



$$\frac{P(x,y)}{P(y)} \equiv P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Kullback-Leibler Divergence

KL divergence from (model) Q to (true) P

$$D_{KL}\left[P||Q\right] = \int_{-\infty}^{\infty} \log\left[\frac{p(x)}{q(x)}\right] p(x) dx$$

or relative entropy of P with respect to Q captures "distance" (*not* a metric!).

In Bayesian inference for instance, it can measure information gain from prior Q to posterior P.

Shannon Entropy

Entropy is a measure of uncertainty

$$H(X) = -\sum_x p(x) \log p(x) = \log(N) - D_{KL}\left(p(x)||p_u\right.$$

In continous case differential entropy (Shannon)

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$$

or better limiting density of discrete points (Jaynes)

$$H(X) = \log(N) - D_{KL}\left(p(x)||m(x)\right)$$

with $m(x)$ the limiting density of discrete points.

Mutual Information

KL divergence of joint from product of marginals

$$
\begin{aligned}
I(X;Y) &= \iint P_{(X,Y)}(x,y) \log\left[\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right] dxdy \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

is the expected reduction in entropy (information gain).

- Captures statistical dependence (zero iff independent)
  - also nonlinear depend. (unlike linear correlation)

Evidence Lower Bound (ELBO)

$$L(\phi, \theta; x) = \int \log \frac{p_\theta(x, z)}{q_\phi(z|x)} q_\phi(z|x) dz$$

$$= \log p_\theta(x) - D_{KL}\left(q_\phi(z|x) || p_\theta(z|x)\right) \leq \mathrm{lo}$$

is a lower bound on the *evidence* $\log p_\theta(x)$ for data x.

In variational Bayesian inference, loss minimization

$$\min_{\theta, \phi} -L(\phi, \theta; x)$$

simultaneously maximizes evidence so that the easy generative model $p_\theta(x|z)p(z)$ is good and minimizes KL divergence so that discriminative model $q_\phi(z|x)$ approximates posterior $p_\theta(z|x)$ well, yielding

$$p(x) \approxeq \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$$

References

Prince, Simon J. D. 2023. *Understanding Deep Learning*. Cambridge, Massachusetts: The MIT Press.