



CAMBRIDGE

# Lecture 0: ML Fundamentals

Stefan Bucher

MACHINE LEARNING IN ECONOMICS  
UNIVERSITY OF CAMBRIDGE

Stefan Bucher



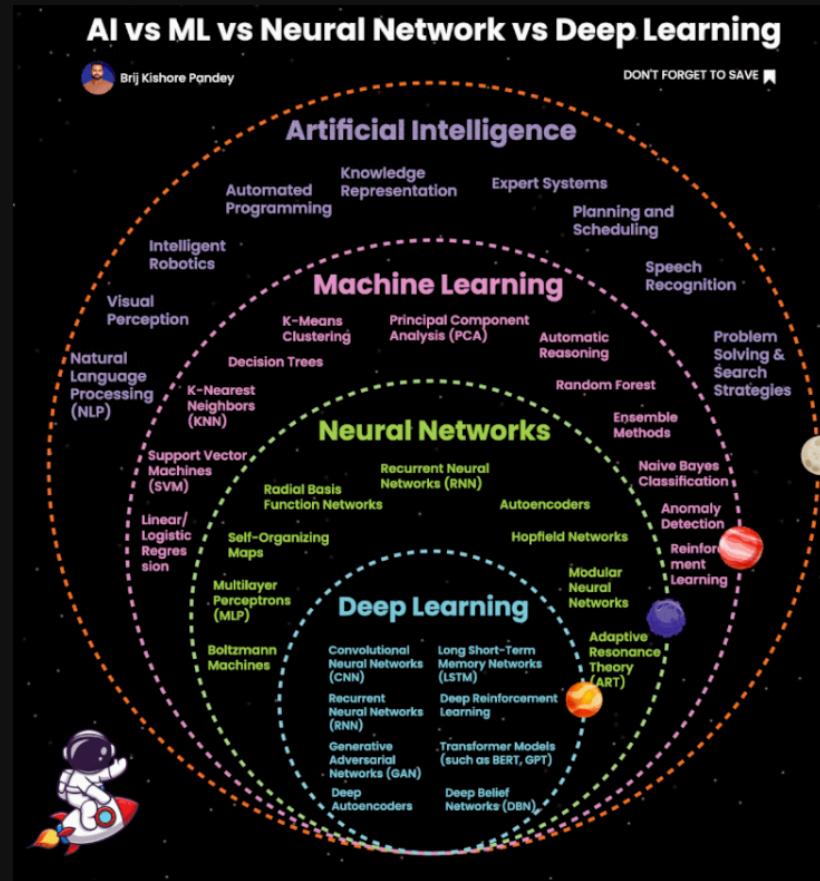
UNIVERSITY OF  
CAMBRIDGE

# Introduction to ML

Prince (2023, chap. 1 and Appendix C).<sup>1</sup>

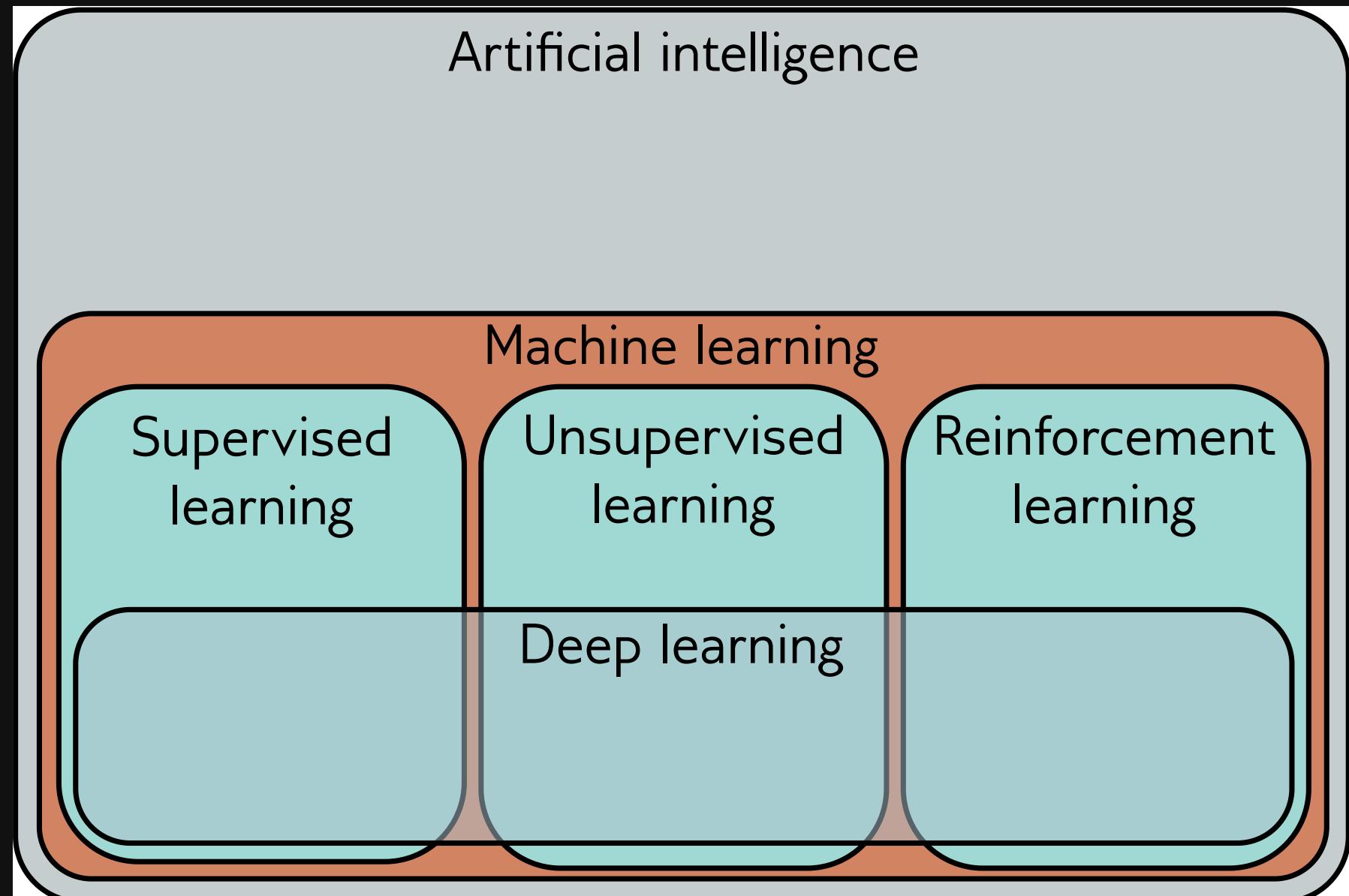
1. Figures taken or adapted from Prince (2023). All rights belong to the original author and publisher. These materials are intended solely for educational purposes.

# What is Machine Learning (ML)?



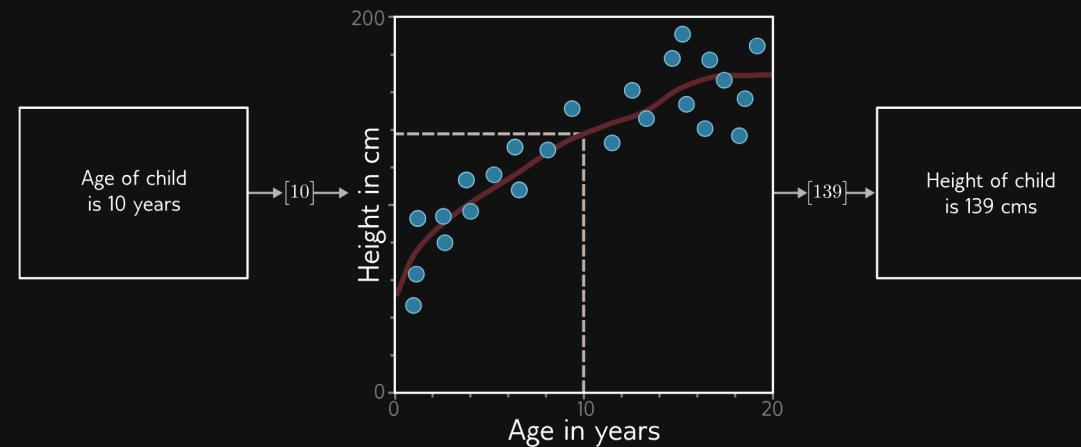
“Machine learning is a field of study that gives computers the ability to learn [from data] without being explicitly programmed.” — Arthur Samuel (1959)

# The Three Pillars of ML

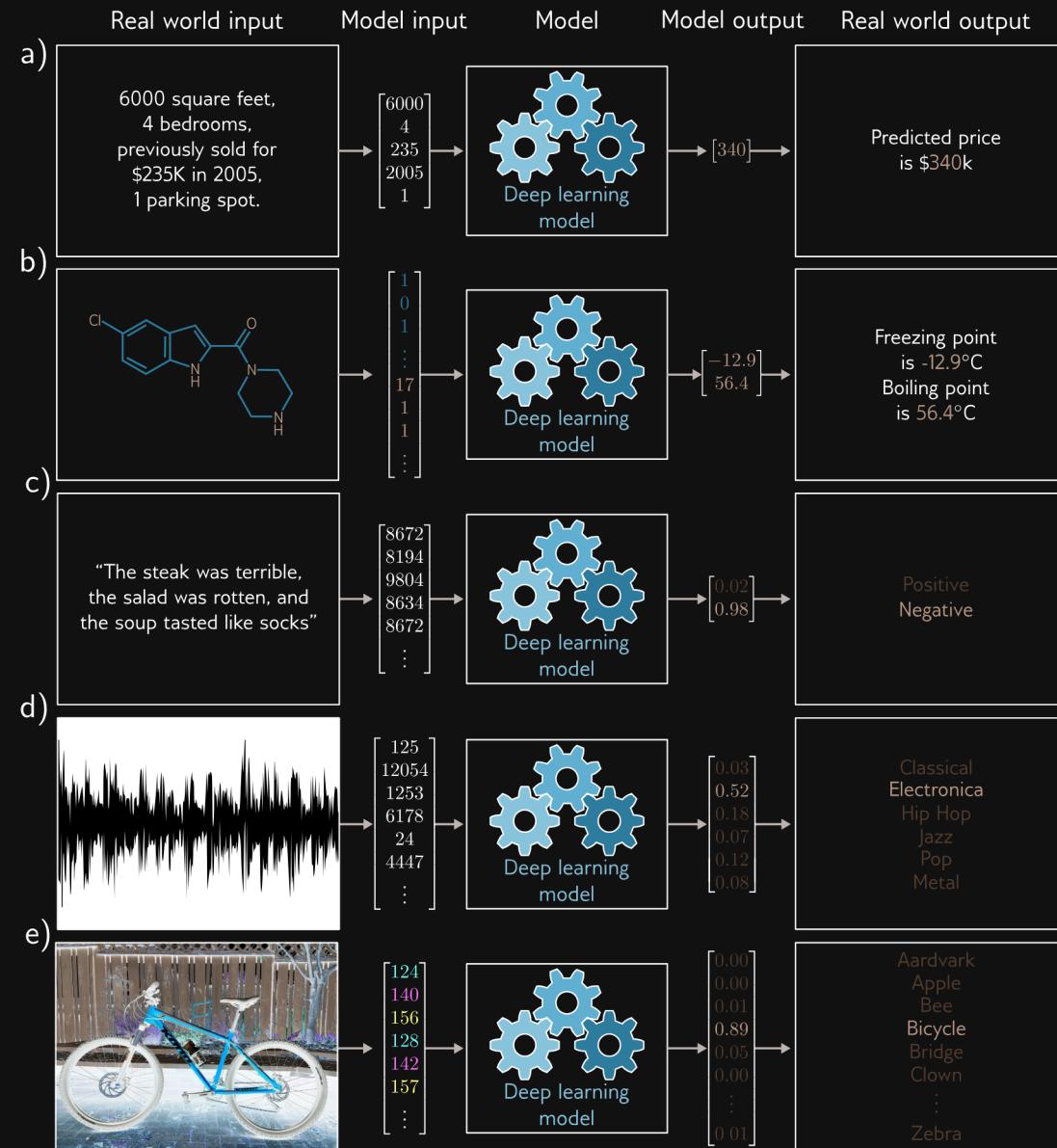


# Supervised Learning

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



# Regression & Classification

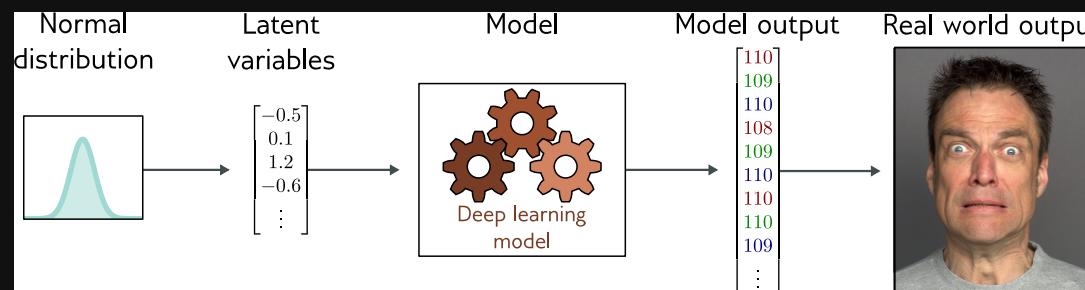


Stefan Bucher

# Unsupervised Learning

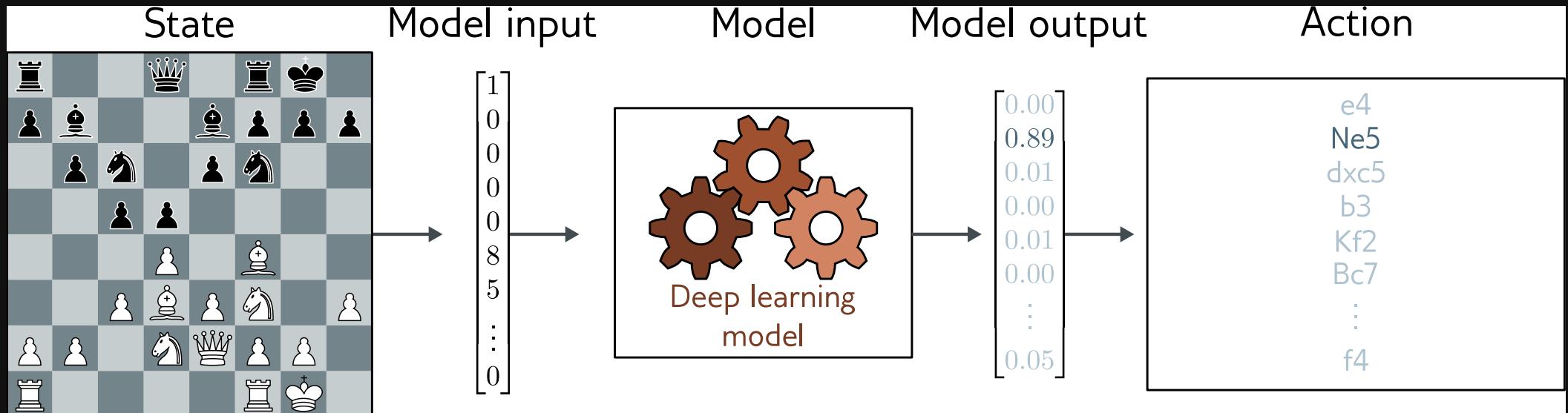


Key: Learning a distribution



Focus on generative unsupervised models

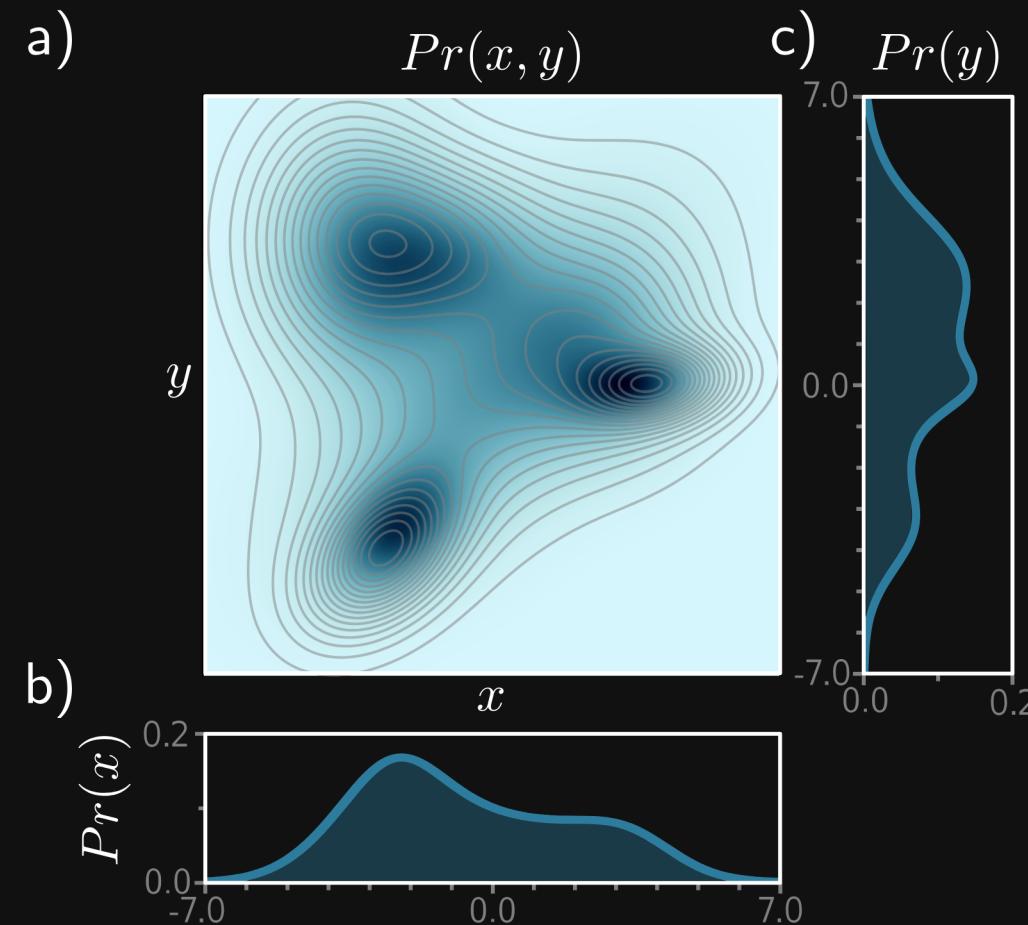
# Reinforcement Learning



Key: Learning an action policy

# Probability and Information Fundamentals

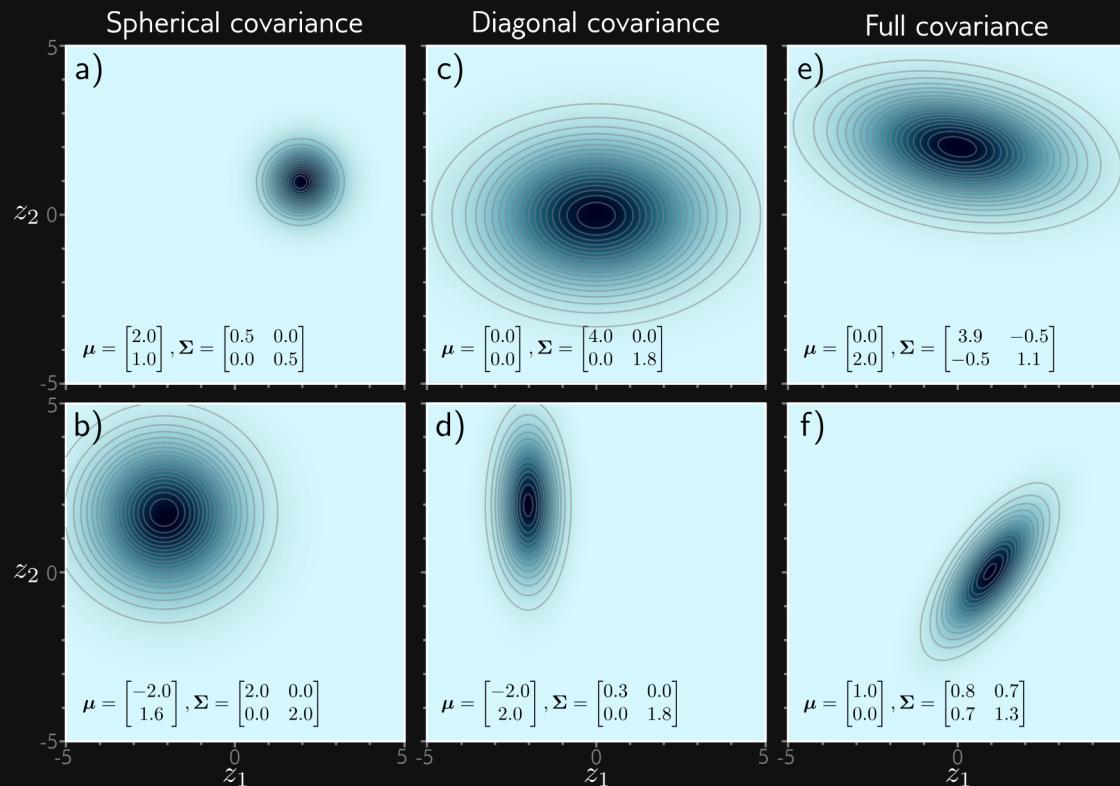
# Joint and Marginal Distributions



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

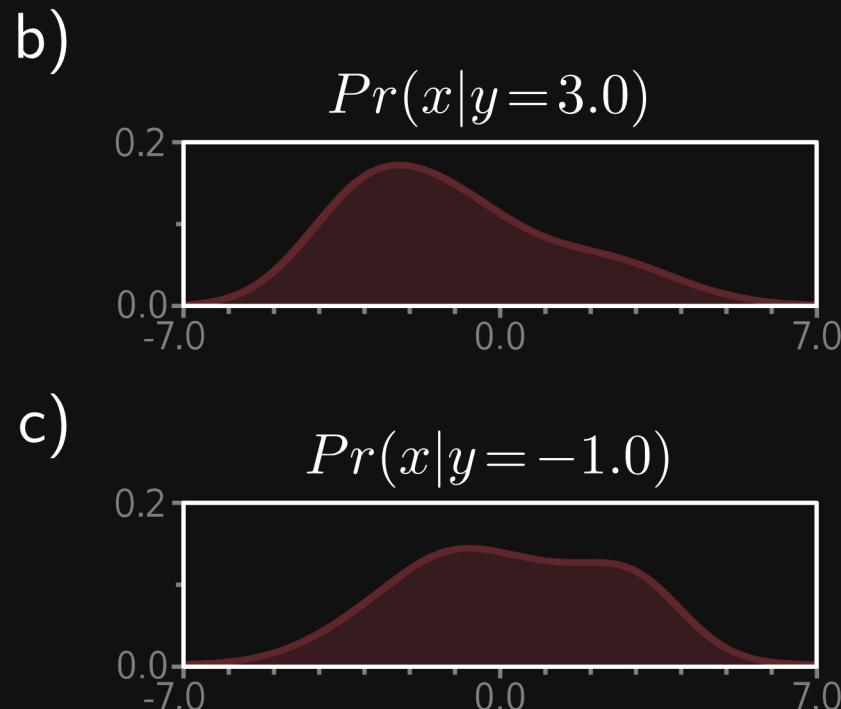
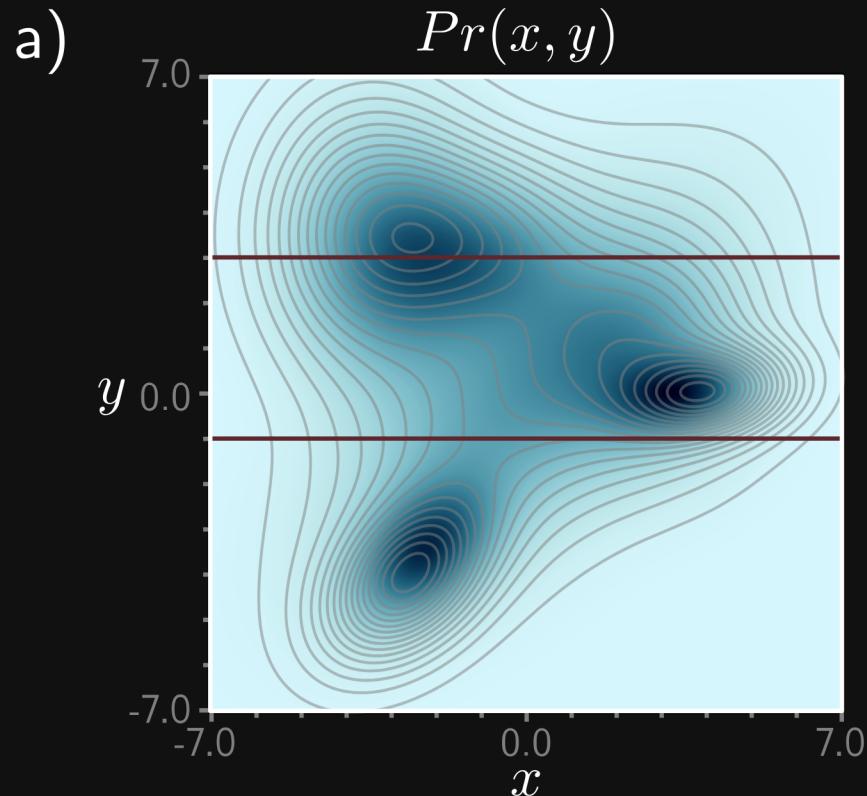
Stefan Bucher

# Bivariate Gaussian



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

# 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)  $P$

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

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

$$H(X) = - \sum_x p(x) \log p(x) = \log(N) - D_{KL} (p(x) || p_l)$$

In continuous case differential entropy (Shannon)

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

or better limiting density of discrete points (Jaynes)

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

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

# Mutual Information

KL divergence of joint from product of marginals

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

is the expected reduction in entropy (information gain).

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

# Evidence Lower Bound (ELBO)

$$\begin{aligned}
 L(\phi, \theta; x) &= \int \log \frac{p_\theta(x, z)}{q_\phi(z|x)} q_\phi(z|x) dz \\
 &= \log p_\theta(x) - D_{KL}(q_\phi(z|x) || p_\theta(z|x)) \leq \log p_\theta(x)
 \end{aligned}$$

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

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

# References

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