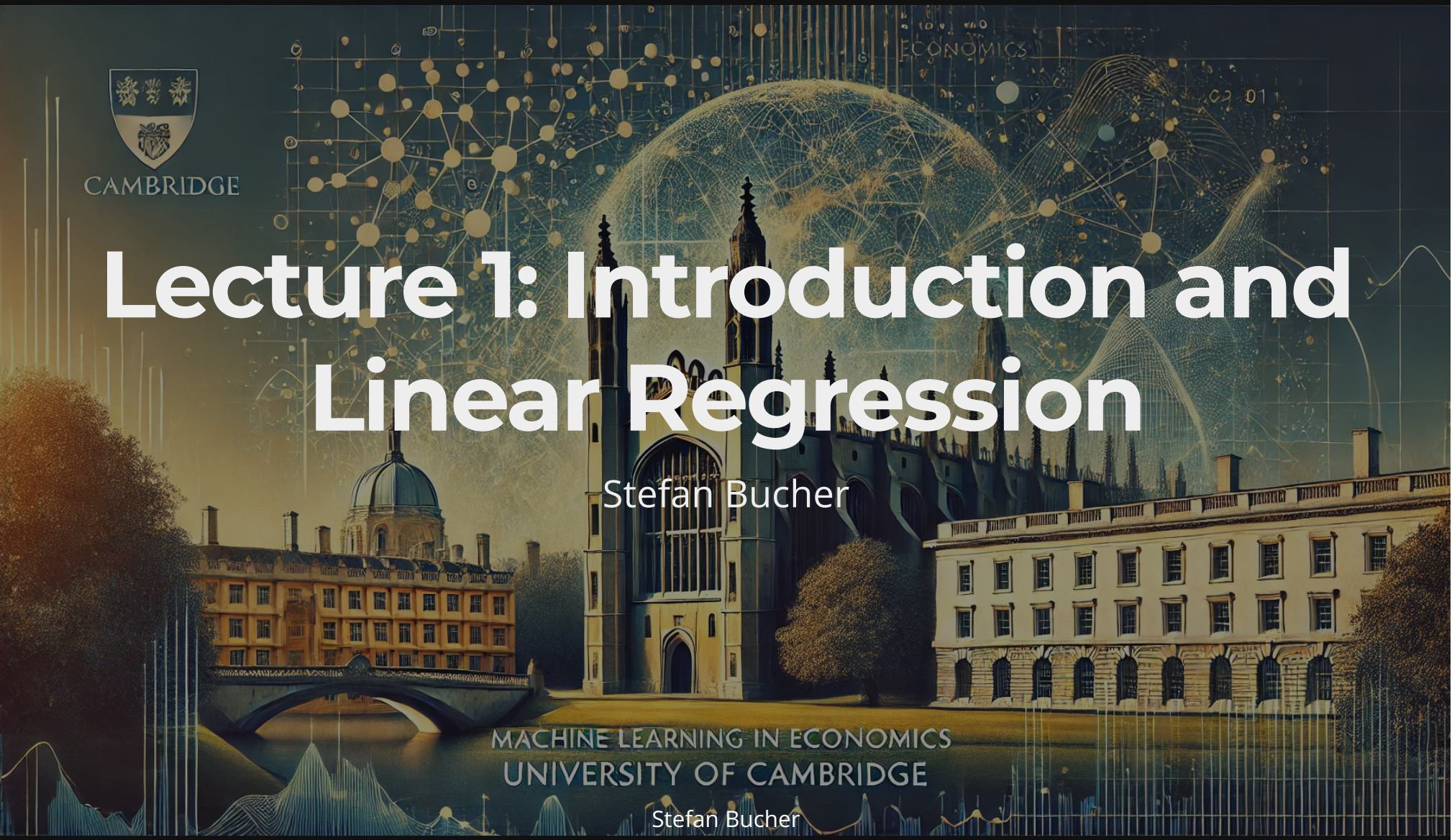


Lecture 1: Introduction and Linear Regression

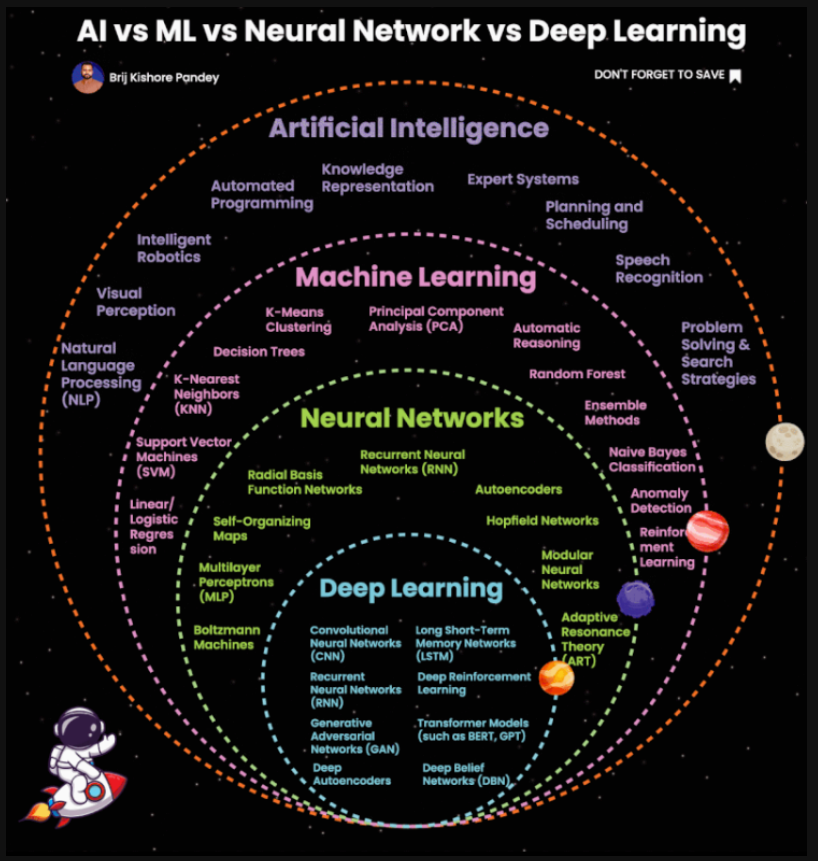
Stefan Bucher

Introduction to ML

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

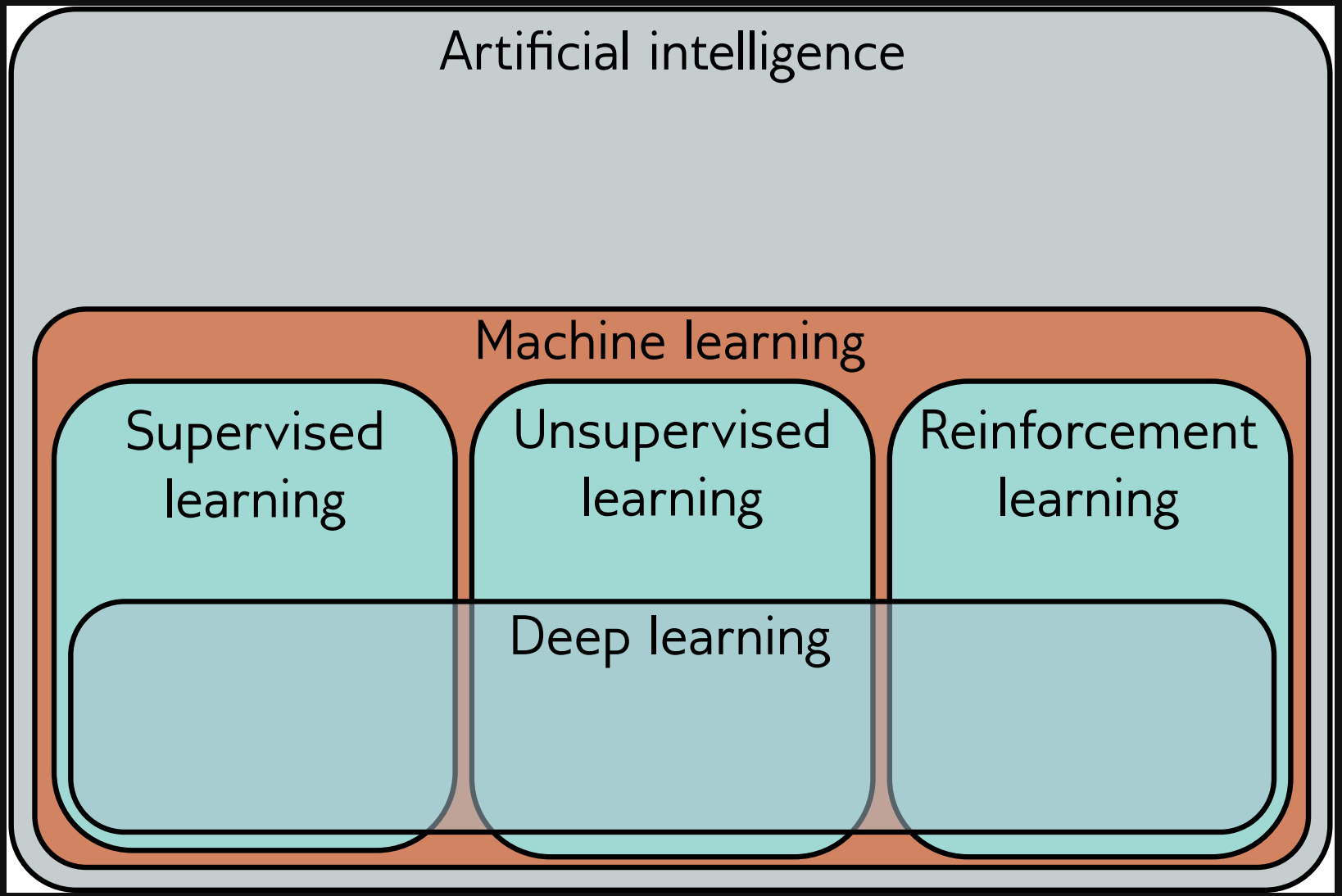
. Figures taken or adapted from Prince (2023). All rights belong to the original author and publisher. These materials are intended solely for educational purposes.

What is Machine Learning (ML)?



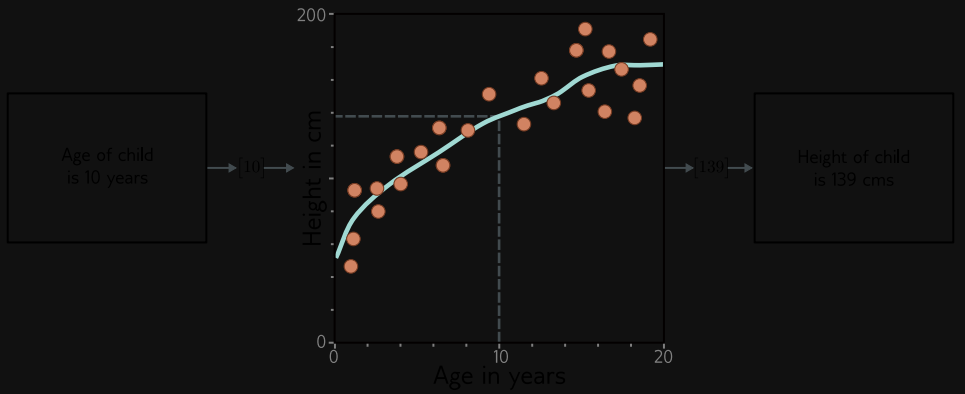
> "Machine learning is a field of study that gives computers the ability to learn [from data] without being explicitly programmed." — Arthur Samuel (1959)

The Three Pillars of ML

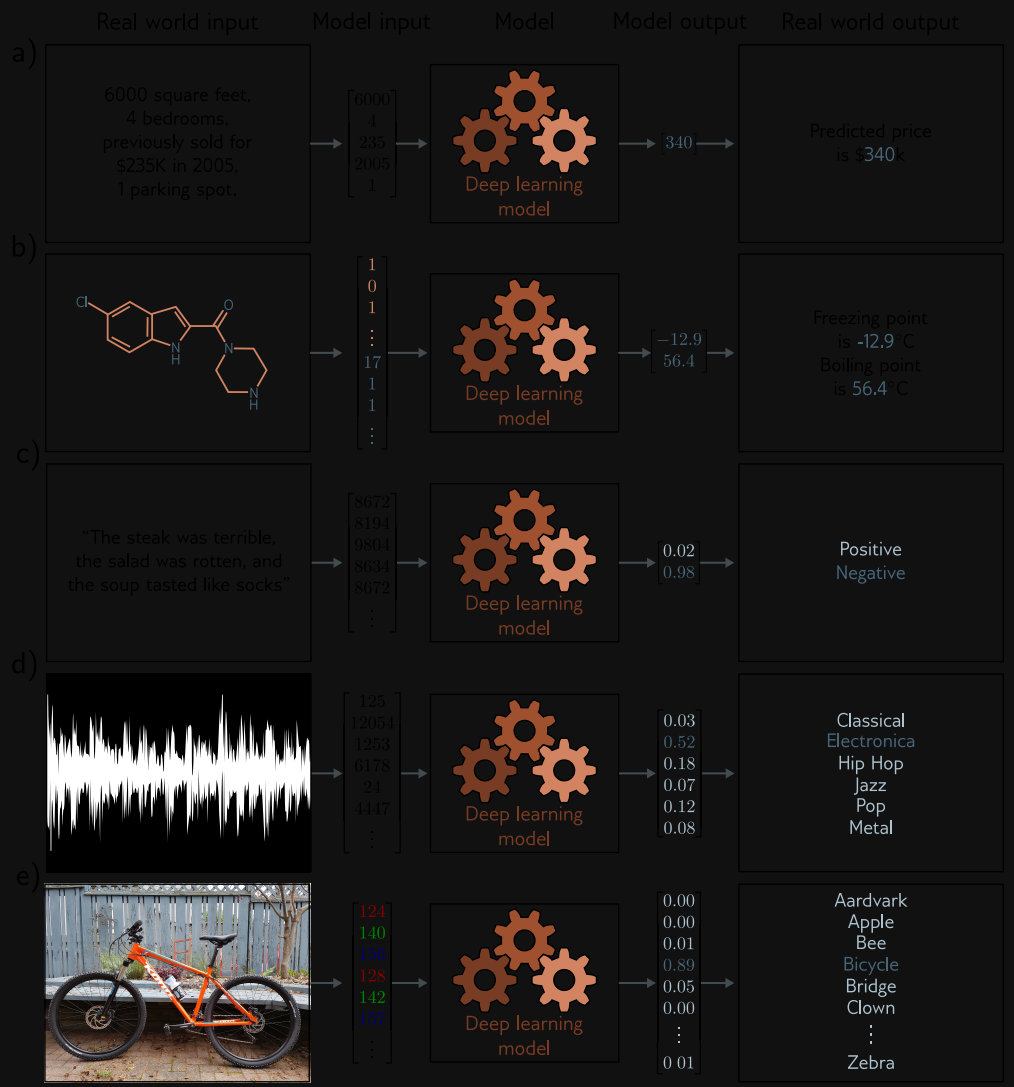


Supervised Learning

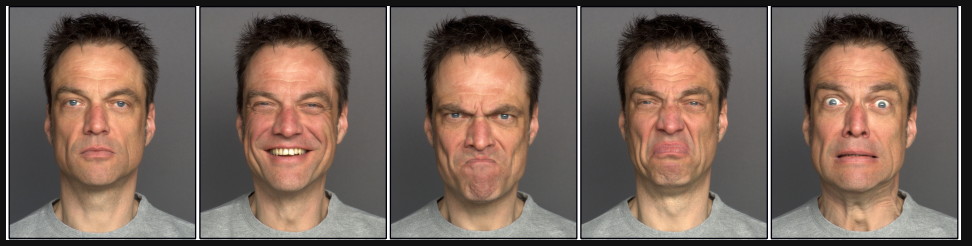
- Economists use regressions to predict outcomes
- ML does the same
- Key: Learning input-output mapping



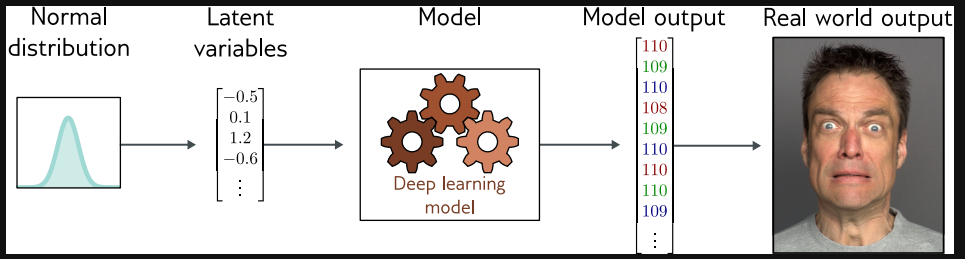
Regression & Classification



Unsupervised Learning

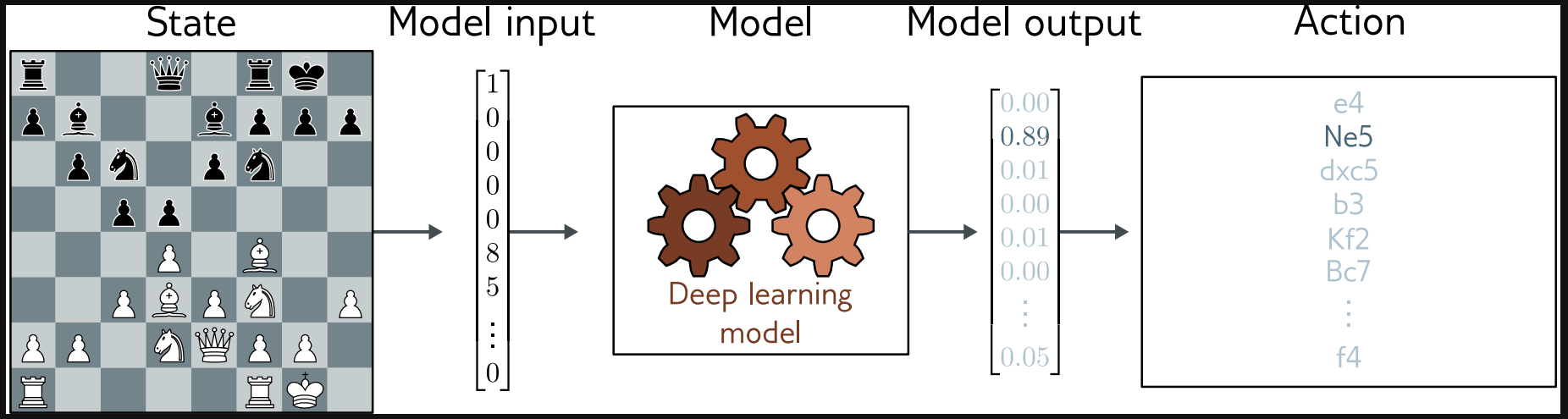


Key: Learning a distribution



Focus on generative unsupervised models

Reinforcement Learning



Key: Learning an action policy

Prediction: Linear Regression

Prince (2023, chap. 2)

Linear Model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Minimizing the mean-squared error (MSE)

$$\min_{\beta} \frac{1}{n}(y - X\beta)^T (y - X\beta)$$

Minimizing MSE Loss (2.1-2.3)

Analytical Solution

$$\min_{\beta} \frac{1}{n}(y - X\beta)^T(y - X\beta)$$

Setting the gradient with respect to β to zero

$$0 = \nabla_{\beta}[y^T y - 2\beta^T X^T y + \beta^T X^T X\beta] = -2X^T y + 2X^T X\beta$$

yields the BLUE estimator (Gauss-Markov)

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

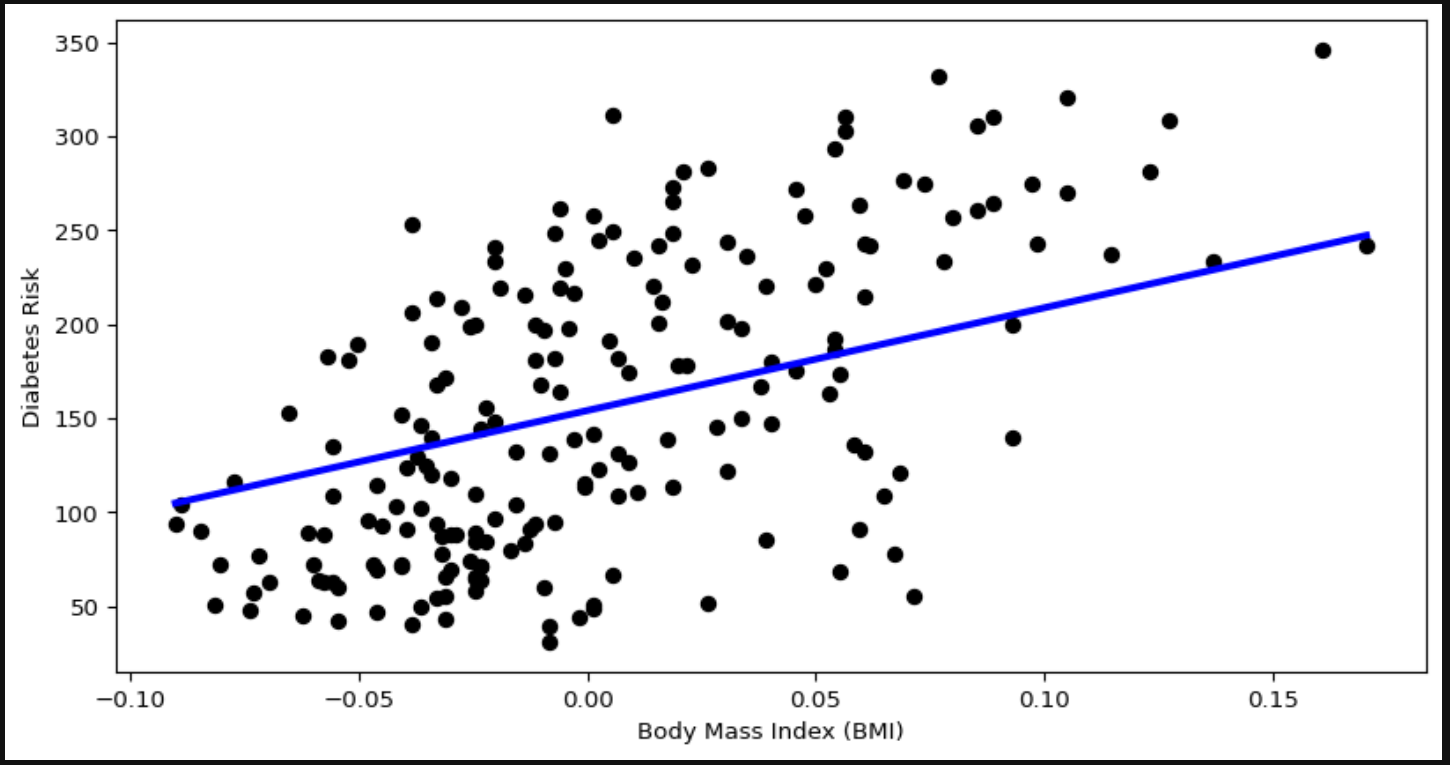
scikit-learn

scikit-learn

```
1 from sklearn.linear_model import LinearRegression
2 model = LinearRegression()
3 model.fit(X_train, y_train)
4 print(model.coef_[2], model.intercept_)
```

```
546.1898661888242 152.05949984245817
```

```
Text(0, 0.5, 'Diabetes Risk')
```



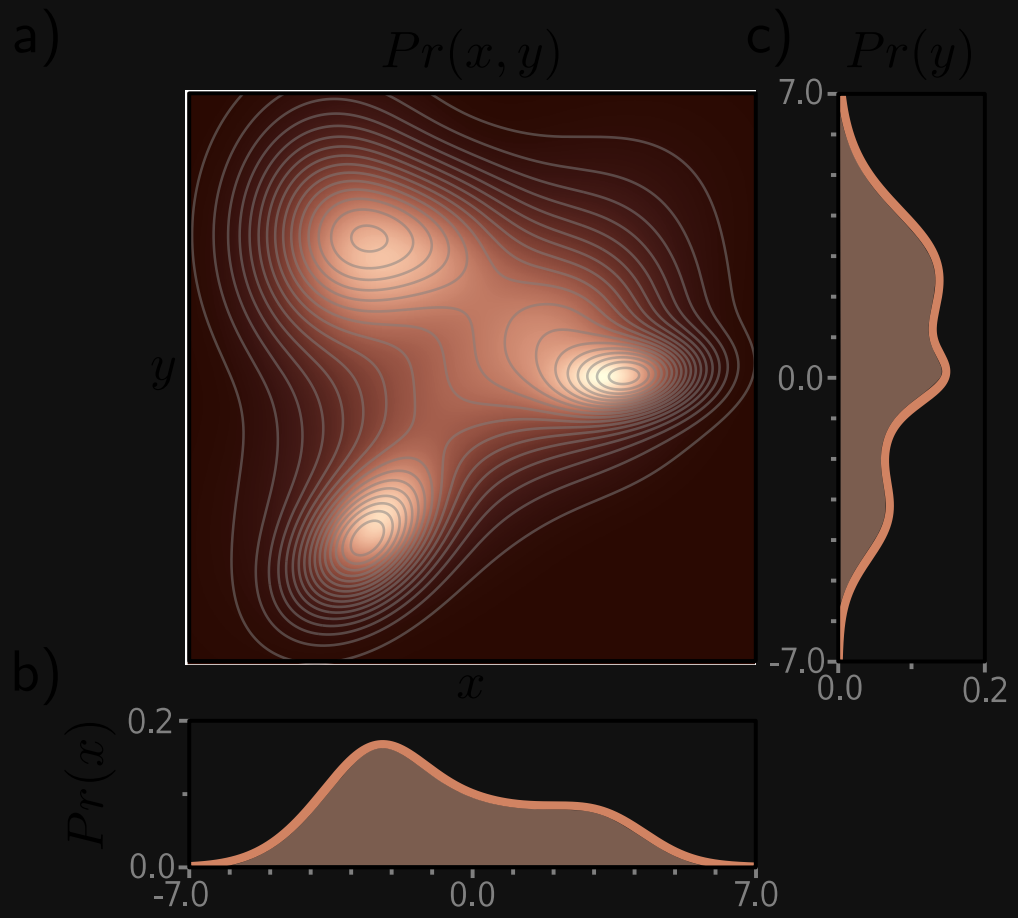
Review: Linear Regression

What did we do?

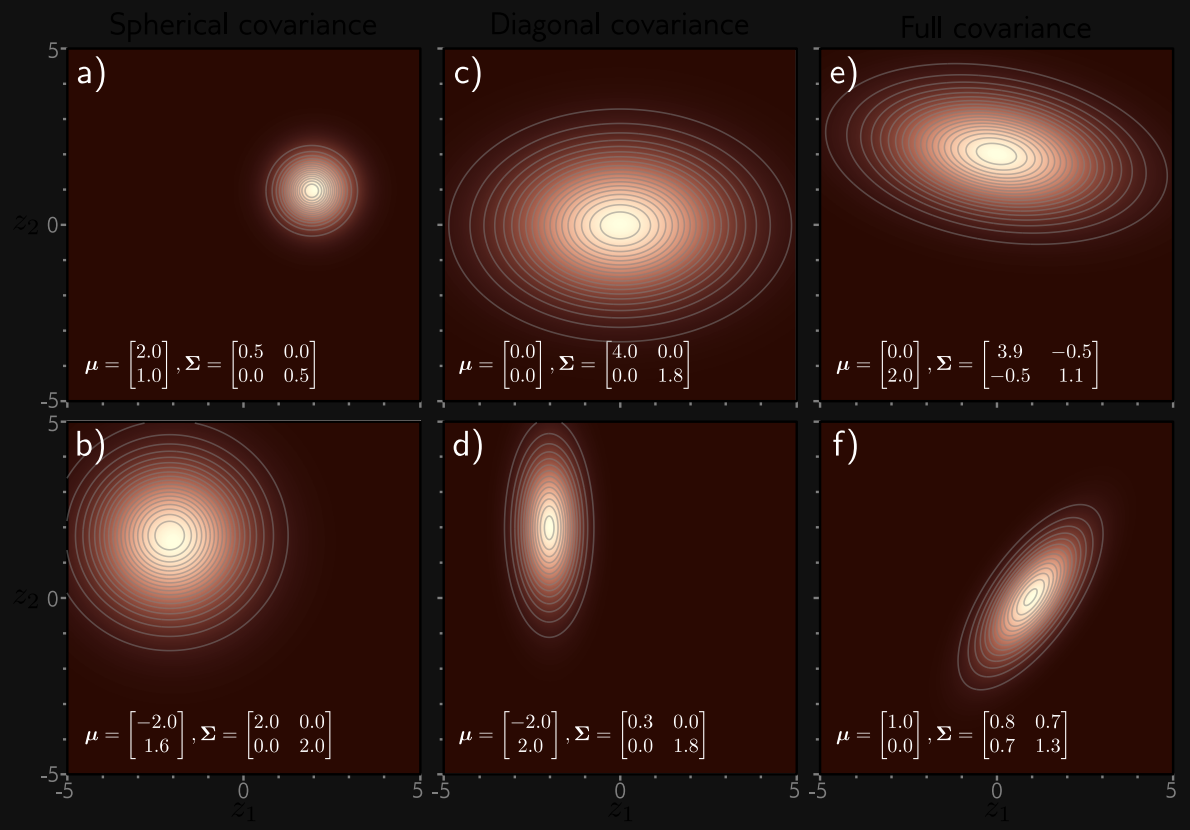
- We chose a model (distribution) with parameters
- We chose an objective function (loss function)
- We (analytically) found the optimal parameter values that minimize the loss function

Probability and Information Fundamentals

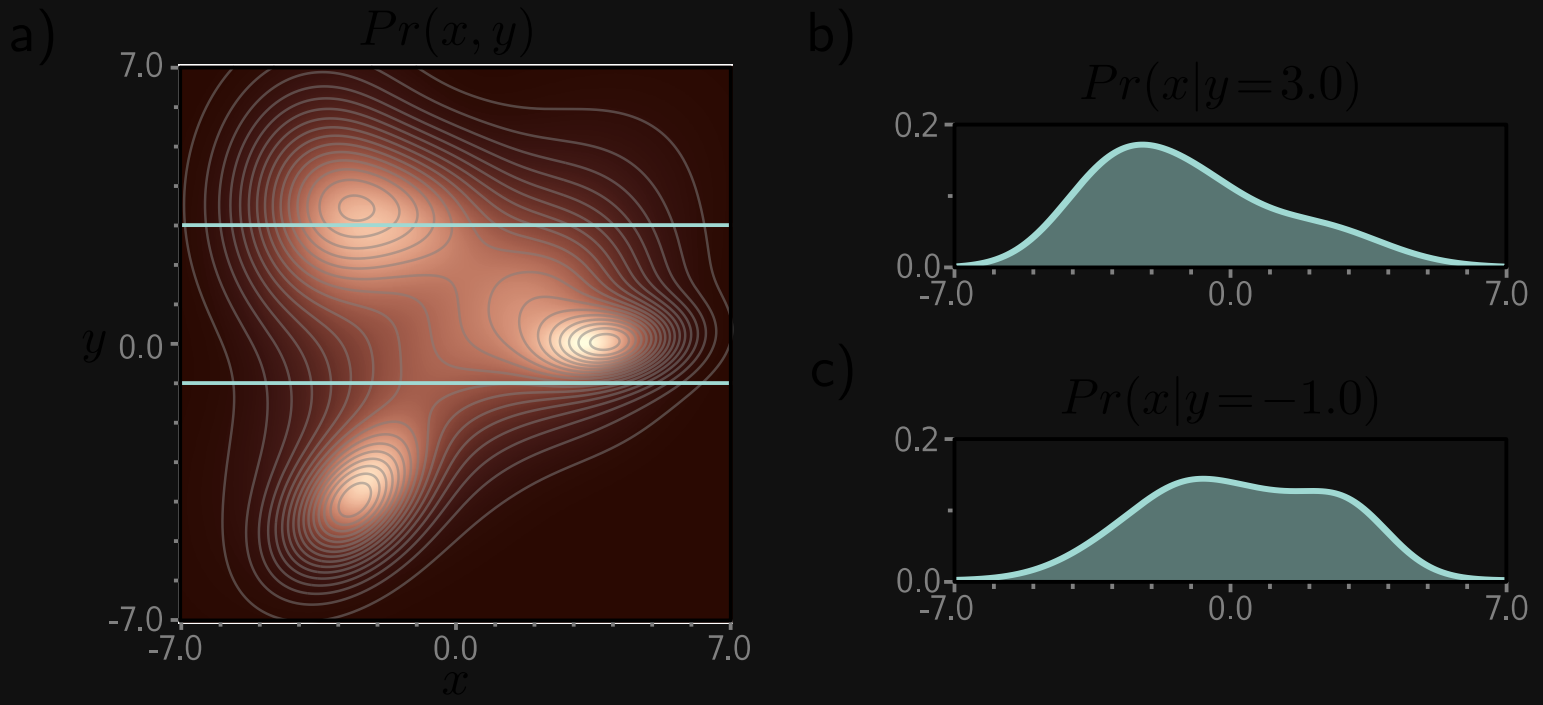
Joint and Marginal Distributions



Bivariate Gaussian



Conditional Dist. & Bayes' Rule



Kullback-Leibler Divergence

KL divergence of (true) from (model) 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 to posterior .

Shannon Entropy

Entropy is a measure of uncertainty

In continous case differential entropy (Shannon) not invariant, so better with respect to limiting density of discrete points (Jaynes)

Mutual Information

KL divergence of joint from product of marginals

is the expected reduction in entropy (information gain).

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

Evidence Lower Bound (ELBO)

is a lower bound on the *evidence* for data .

In variational Bayesian inference, loss minimization simultaneously maximizes evidence so that the easy generative model is good and minimizes KL divergence so that discriminative model approximates posterior well, yielding

References

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