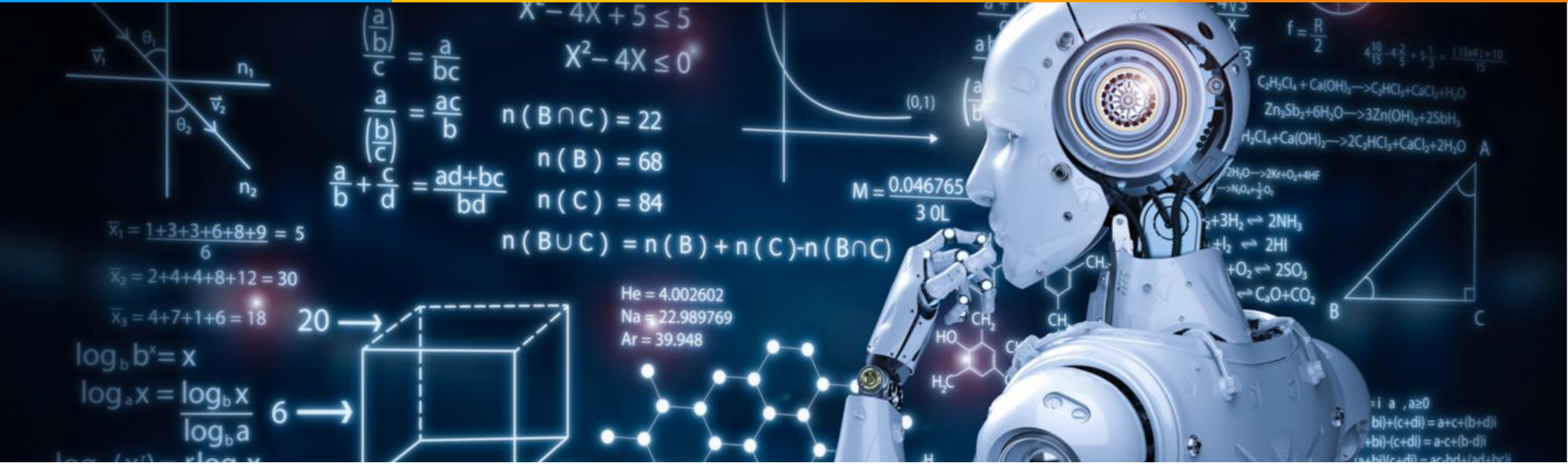
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML



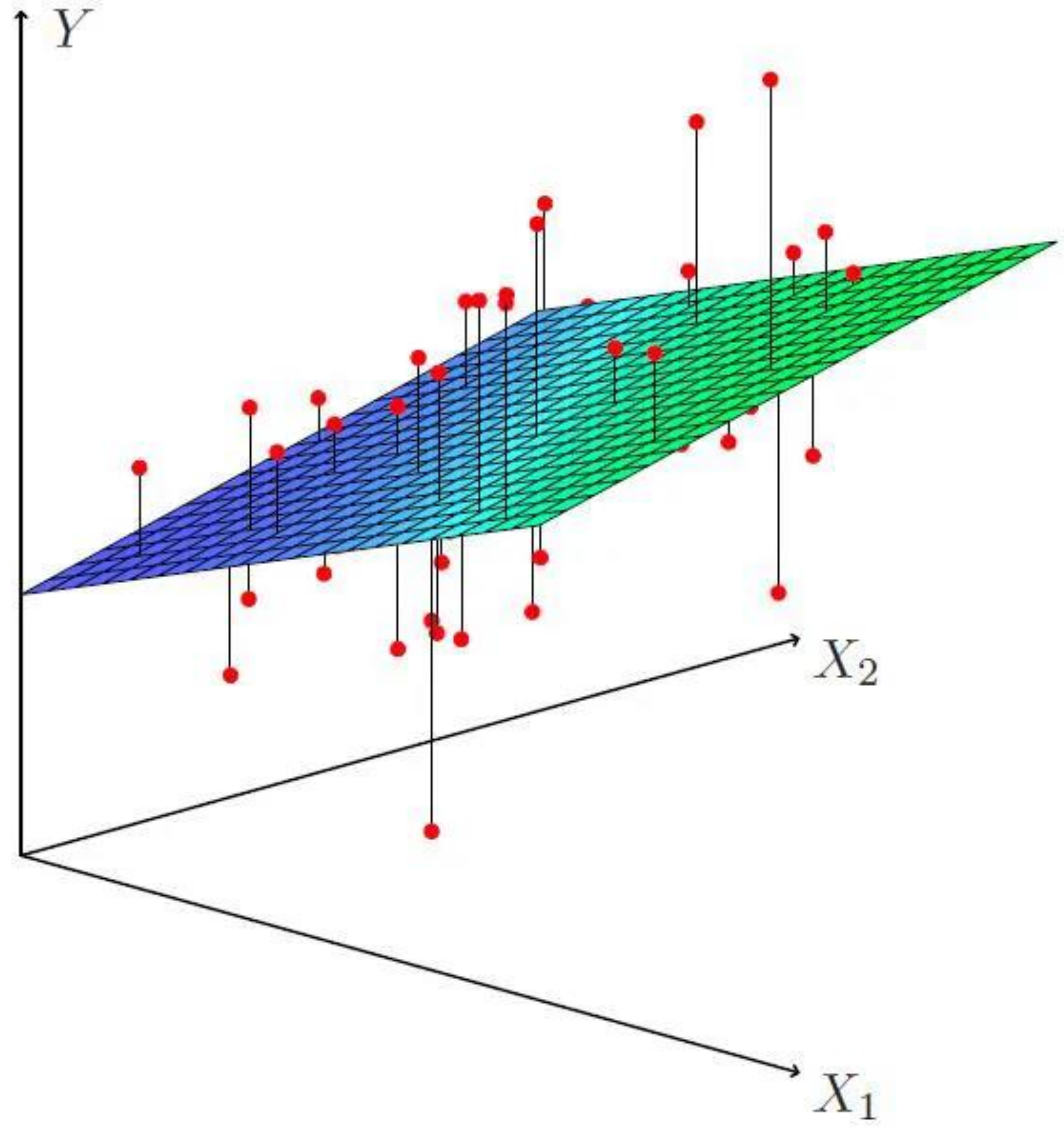
Lecture 22 & 23: Linear Regression Part 2

Recap

- Population and Sample Regression
- Simple Linear Regression Intuition
- Linear Regression Algorithm
- Gradient Descent
- Impact of Scaling in Gradient Descent
- Closed form analytical solution
- Types of Gradient Descent
- Coding Linear Regression

Simple Linear Regression

Multiple Linear Regression



CAUTION

This is going to be mathematically very intensive

Multivariate calculus refresher & Gradient intuition

Multivariate calculus refresher

$$y = f(x) : \mathcal{R} \rightarrow \mathcal{R} \qquad \frac{df}{dx}$$

$$y = f(x_1, x_2, \dots x_n) : \mathcal{R}^n \rightarrow \mathcal{R}$$

$$\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots \frac{\partial f}{\partial x_n}$$

$$y = f(\mathbf{x}) : \mathcal{R}^n \rightarrow \mathcal{R}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \quad \nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Input →	scalar x	vector $\mathbf{x}$
scalar f	$\frac{\partial f}{\partial x}$	$\frac{\partial f}{\partial \mathbf{x}}$
vector $\mathbf{f}$	$\frac{\partial \mathbf{f}}{\partial x}$	$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$

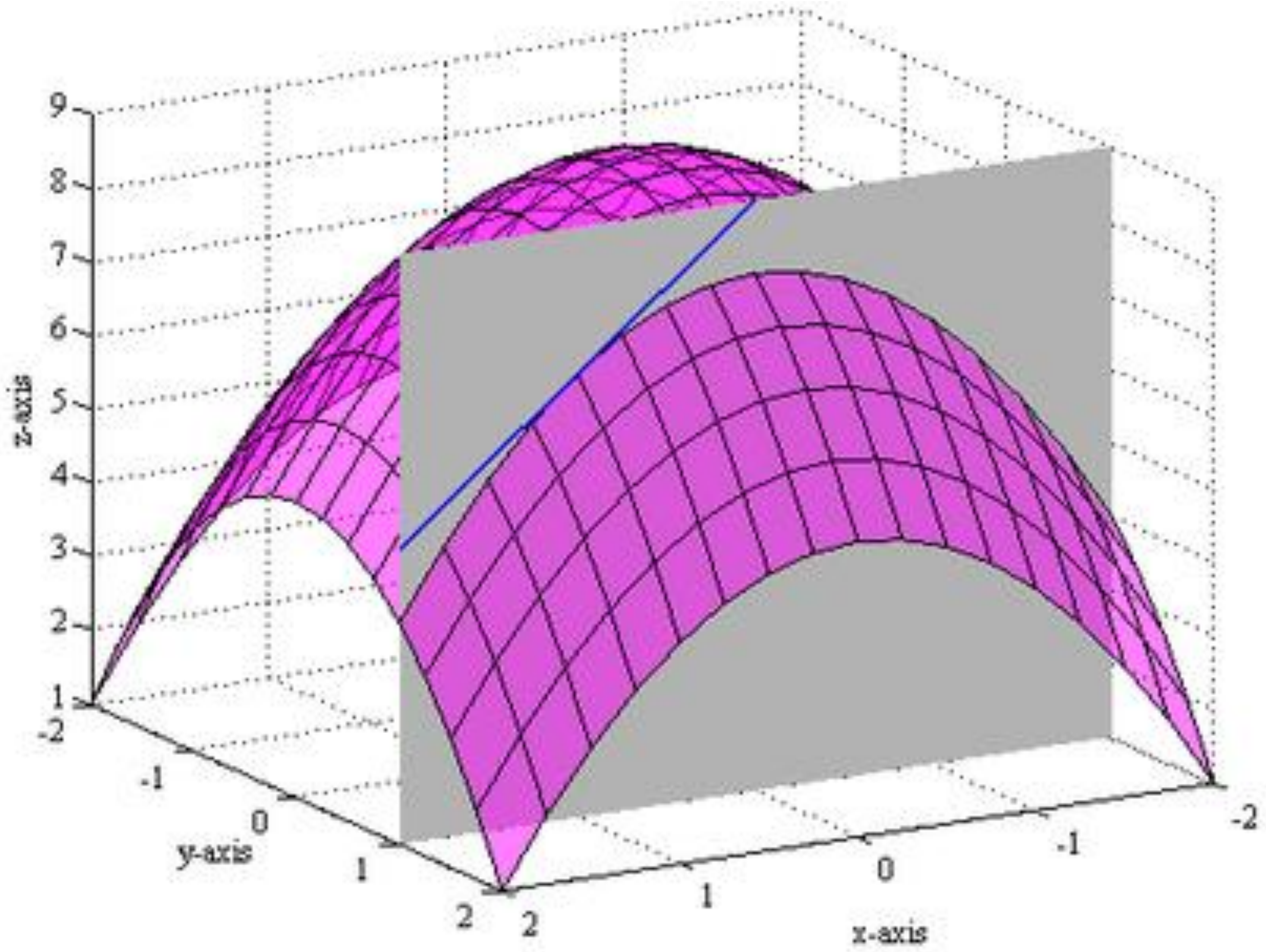
$$\nabla f \Big|_{(a,b)} = \begin{bmatrix} \frac{\partial f}{\partial x_1}\Big|_{(a,b)} \\ \frac{\partial f}{\partial x_2}\Big|_{(a,b)} \end{bmatrix}$$

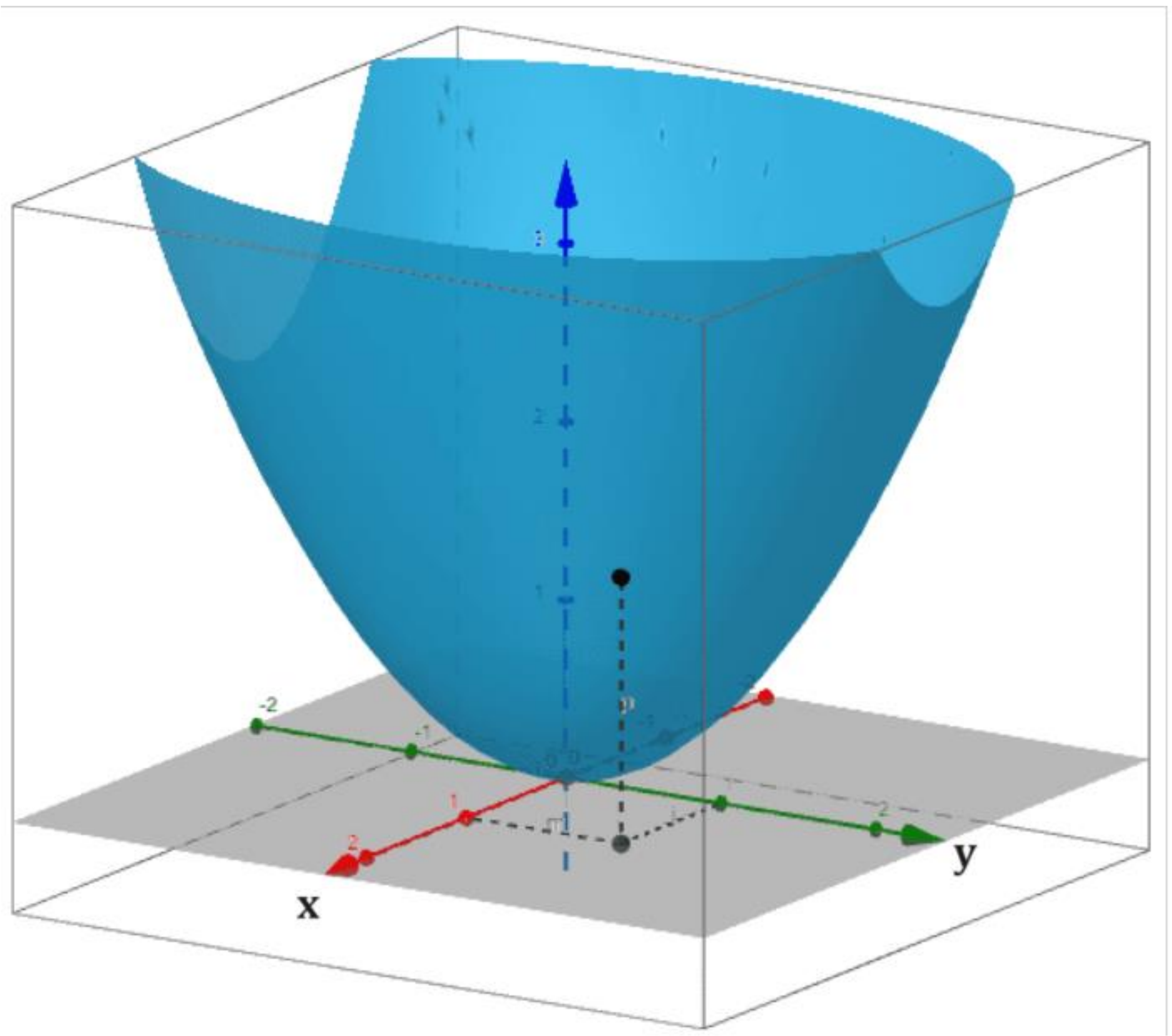
What is partial derivative exactly?

$$z = f(x, y) : \mathcal{R}^2 \to \mathcal{R}$$

$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$$

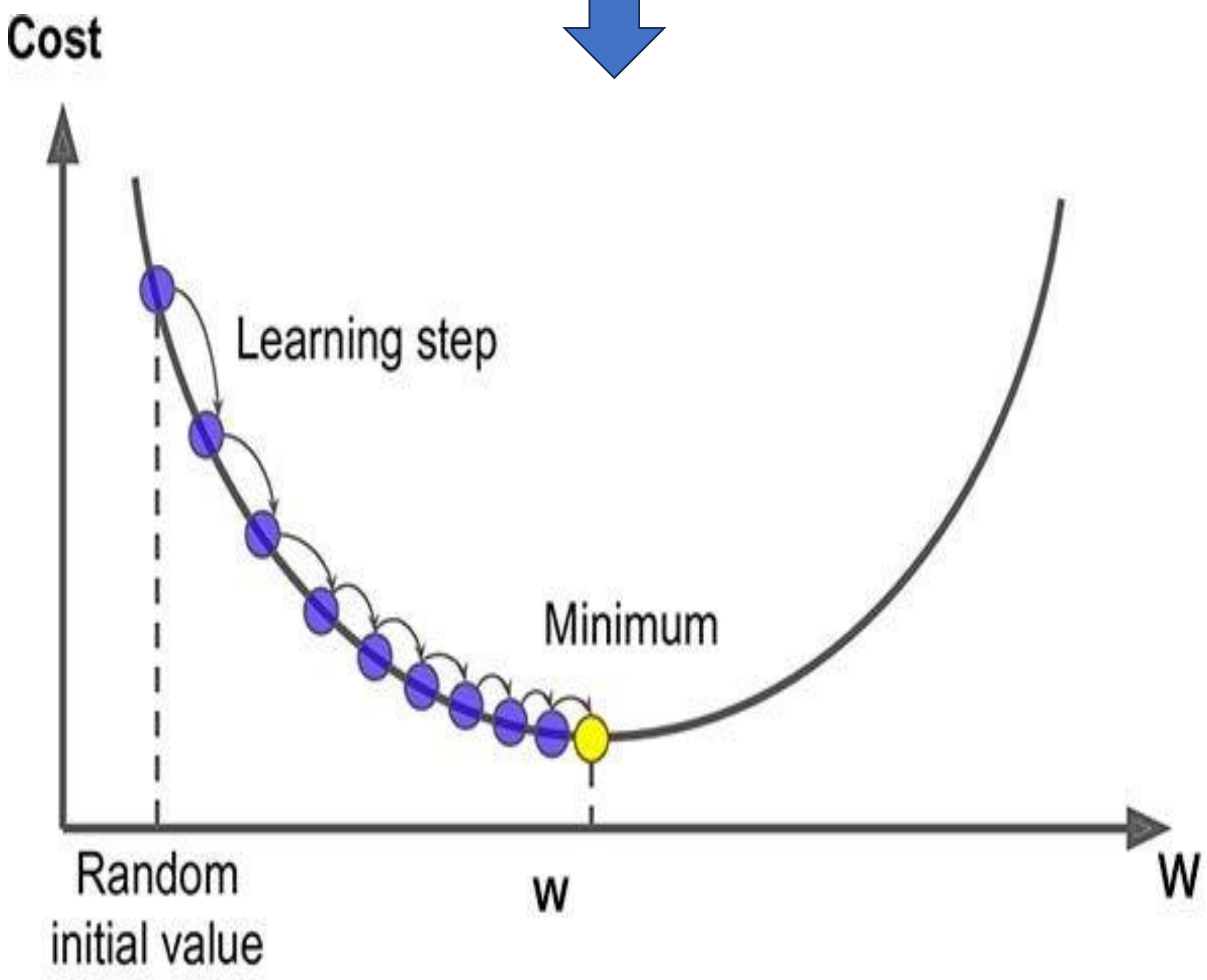
Derivative along one axis using another axis aligned plane



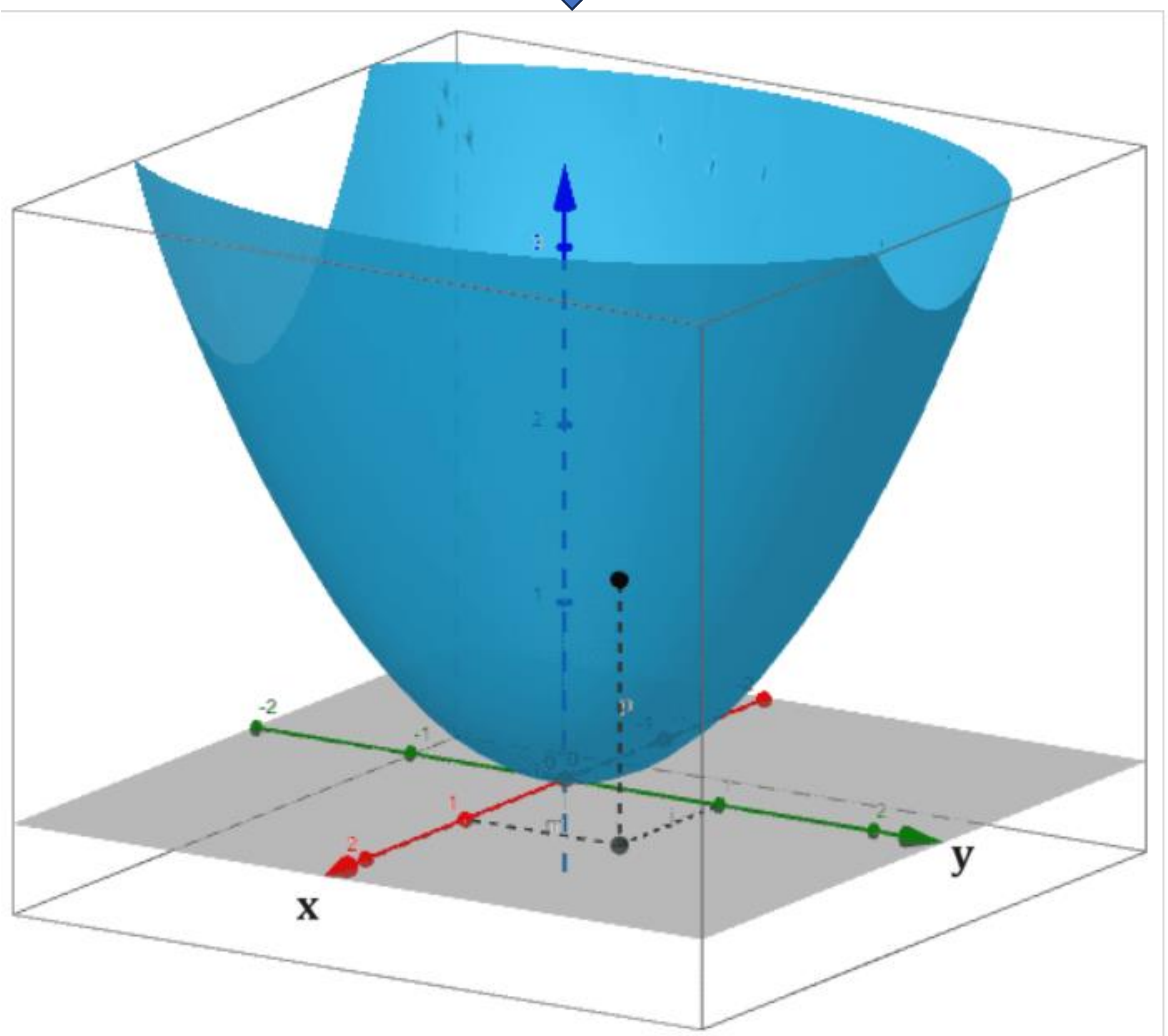


Which direction to do gradient descent?

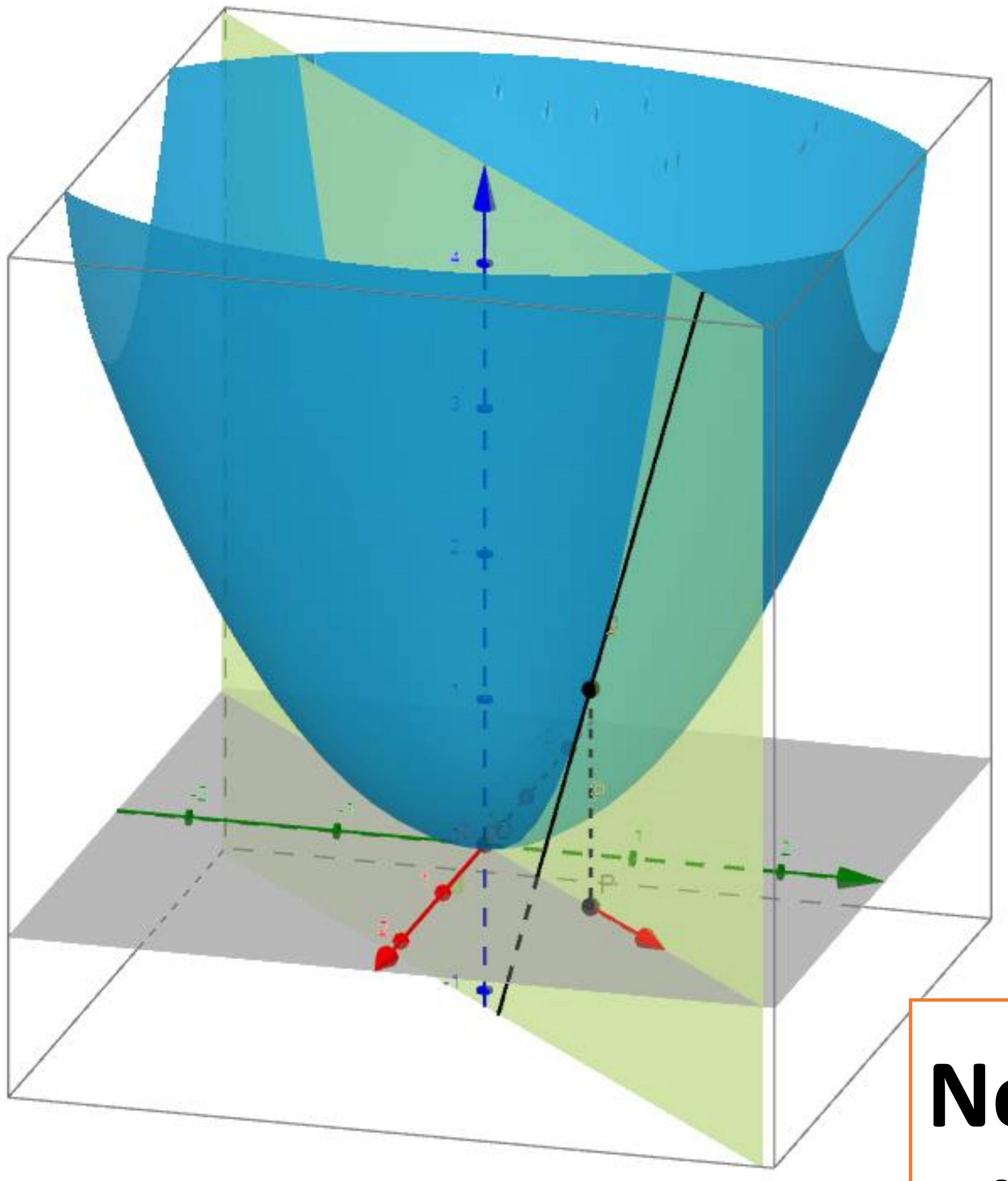
- In the direction of slope
- 2 directions – left & right



- Slope in WHICH direction?
- Infinite directions at each point



What is gradient?



$$y = f(x_1, x_2, ...x_n) : \mathcal{R}^n \to \mathcal{R} \quad \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots \frac{\partial f}{\partial x_n}$$

What is directional derivative (aka slope)?

$$D_{\tilde{\mathbf{u}}} f(a,b) = (\nabla f)^T \vec{u}$$

Max value when theta = 0

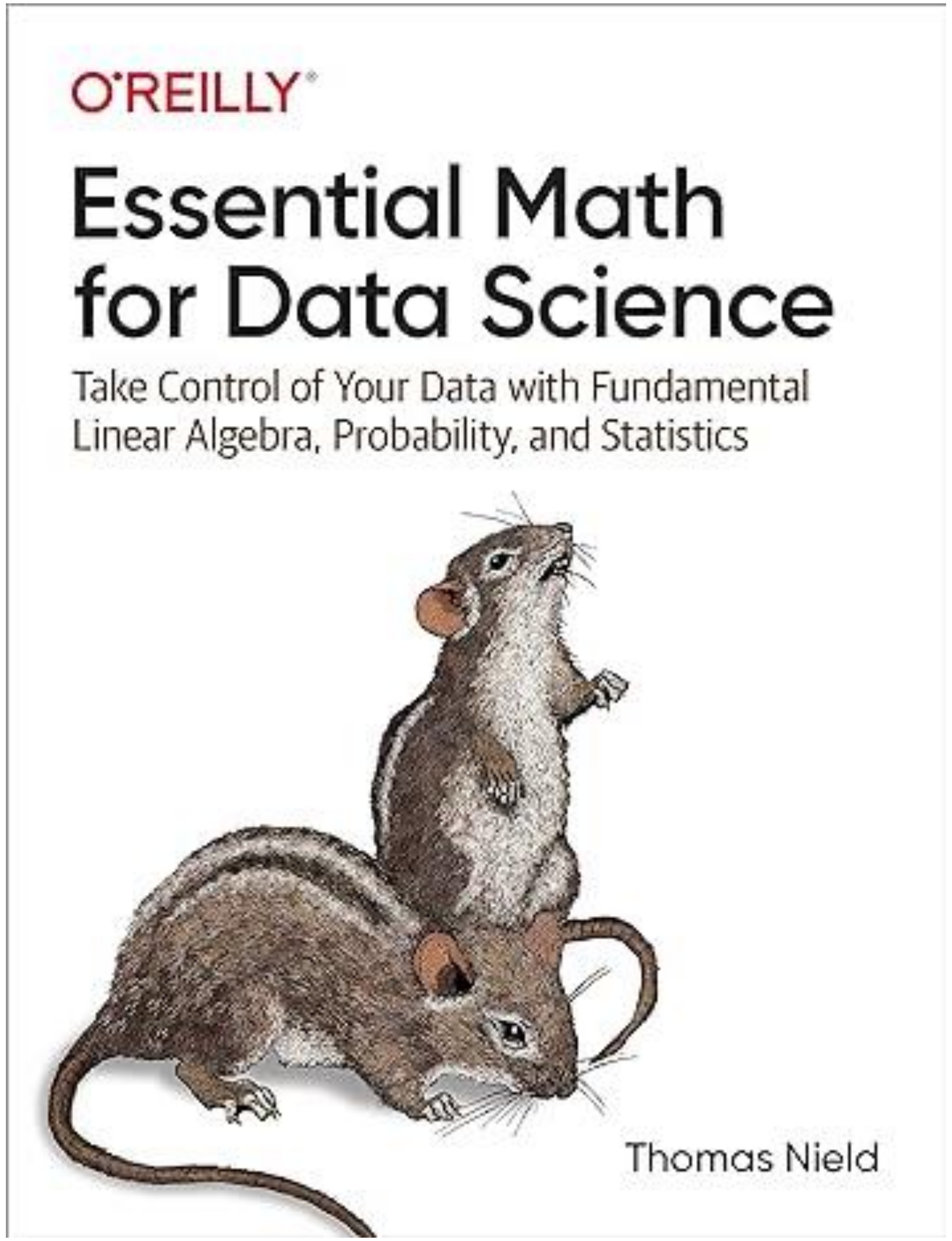
Gradient: Direction of steepest ascent

Negative gradient: Direction of steepest descent

$$\nabla f\big|_{(a,b)} = \begin{bmatrix} \frac{\partial f}{\partial x_1}\big|_{(a,b)} \\ \frac{\partial f}{\partial x_2}\big|_{(a,b)} \end{bmatrix}$$

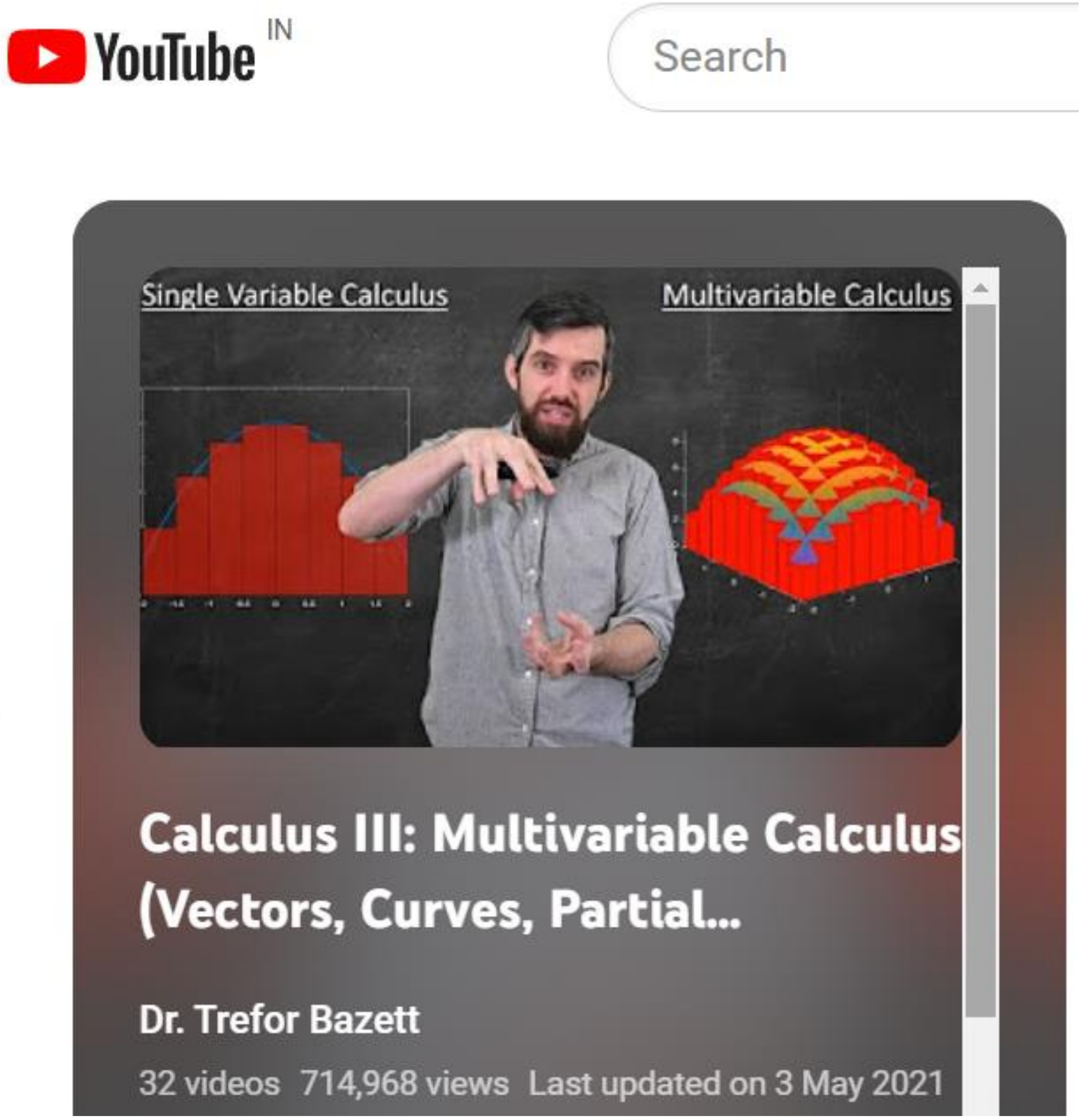
Supplementary Reading (Optional)

- Easy read
- Just very high level overview



Supplementary Videos (Optional)

- Calculus III – Multivariable Calculus – Dr. Trefor Bazett
- https://www.youtube.com/playlist?list=PLHXZ9OQGMqxc_CvEy7xBKRQr6I214QJcd

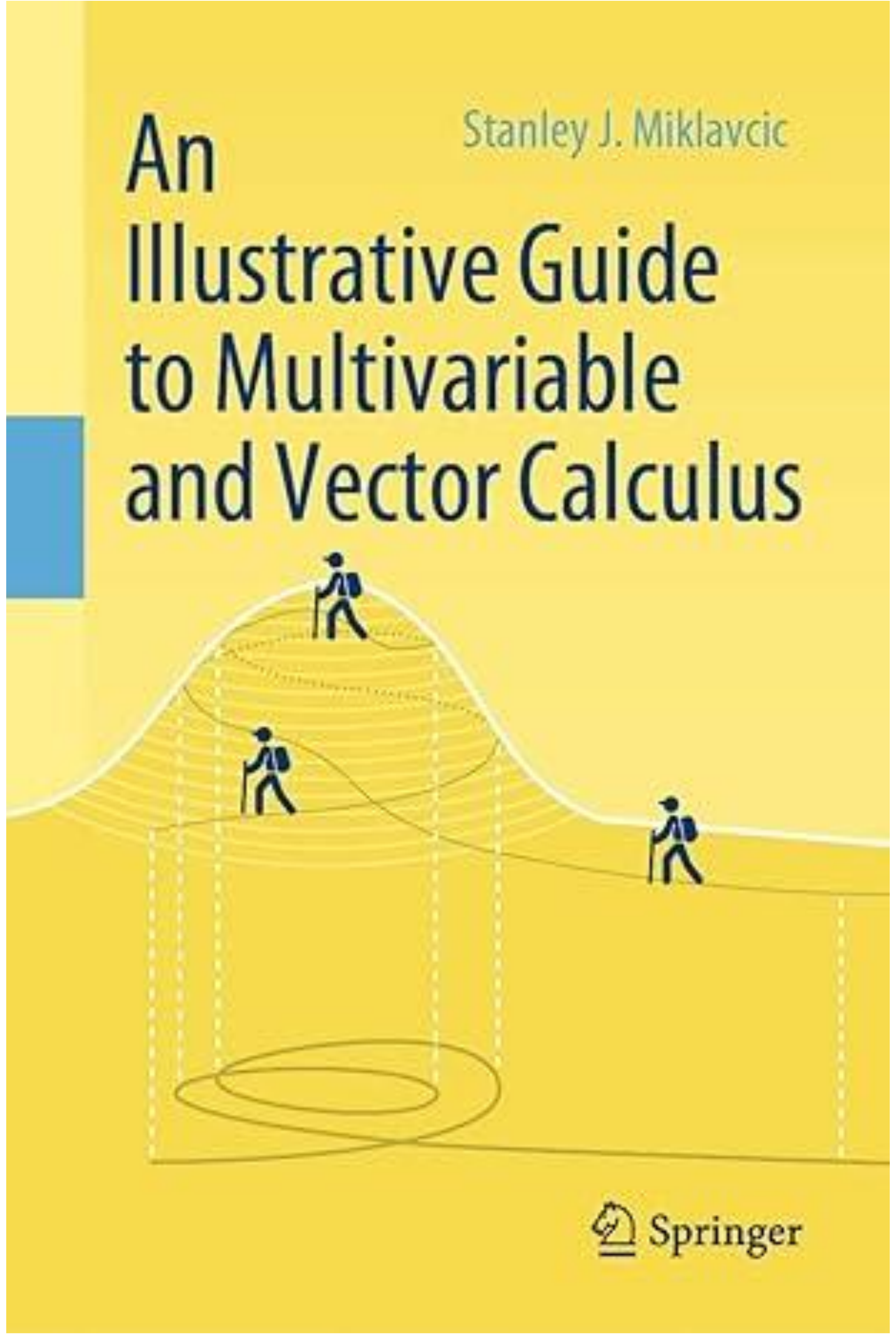


Supplementary Reading (Optional)

- The Matrix Calculus you need for Deep Learning
  - https://arxiv.org/abs/1802.01528
- Explanation
  - Part1: https://www.youtube.com/watch?v=pQ5HT8LylZs
  - Part 2: https://www.youtube.com/watch?v=rWRb8K-hcTo

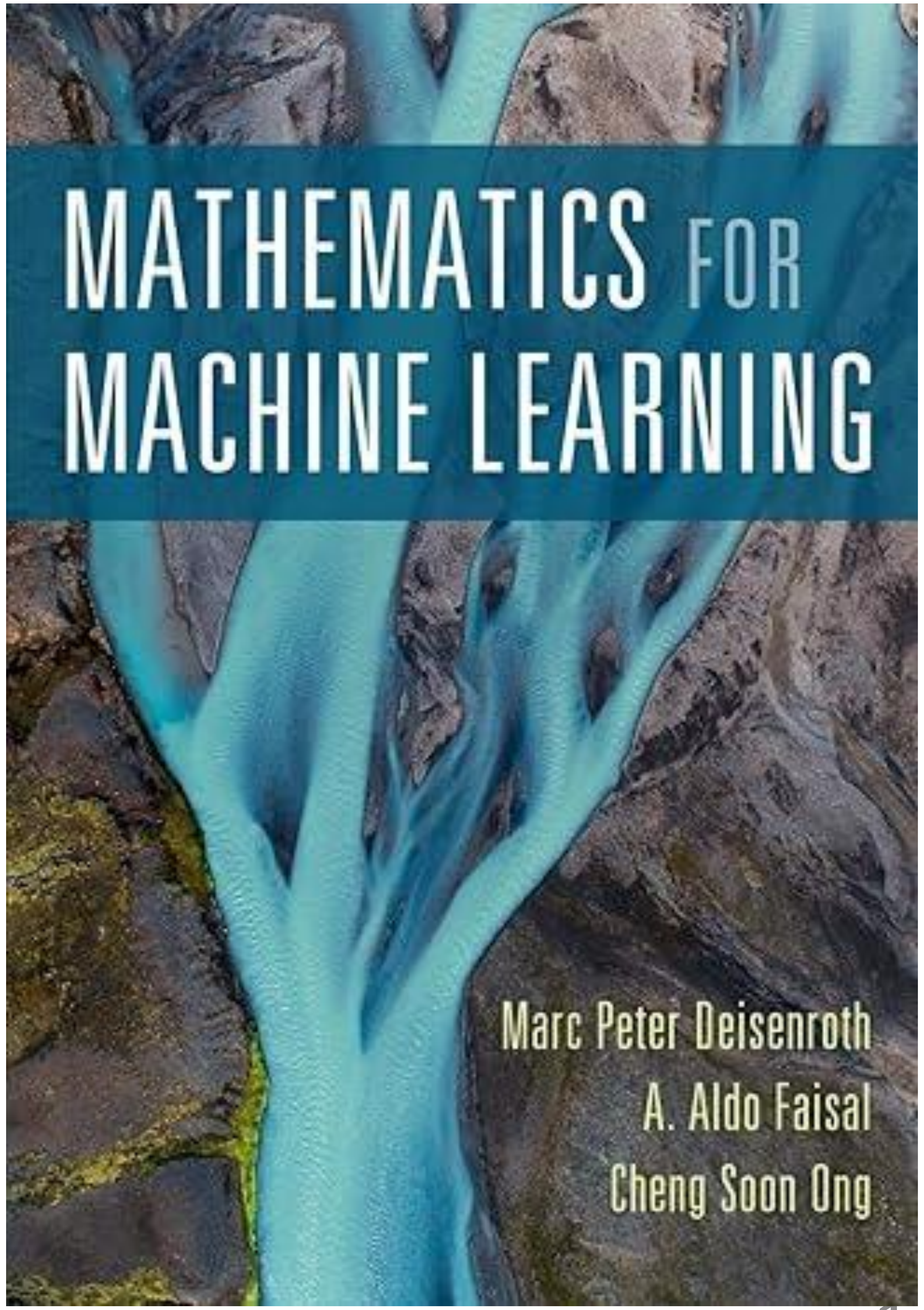
Supplementary Reading (Optional)

- An illustrative guide to Multivariable and Vector Calculus
  - Detailed
  - MATLAB code



Supplementary Reading (Optional)

- Mathematics for Machine Learning
  - https://mml-book.github.io/
- Chapter 5 Vector Calculus
- More of a reference/review book



Vectorized Linear Regression

Objective function - vectorized solution

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Non vectorized

$$\hat{y} = w_1 x^{(i)} + w_0$$

$$\mathcal{J}(w_1, w_0) = \frac{1}{m} \sum_{i=1}^{m} \left(w_0 + w_1 x^{(i)} - y^{(i)}\right)^2$$

Vectorized

$$\hat{y} = w^T x^{(i)} \qquad \mathbf{x}^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \end{bmatrix}$$

$$\mathcal{J}(w_1, w_0) = \frac{1}{m} \sum_{i=1}^{m} \left(w^T x^{(i)} - y^{(i)}\right)^2$$

$$\mathcal{J}(w) = \frac{1}{m} \left(\begin{bmatrix} {x^{(1)}}^T w \\ {x^{(2)}}^T w \\ \dots \\ {x^{(m)}}^T w \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{bmatrix} \right)^T \left(\begin{bmatrix} {x^{(1)}}^T w \\ {x^{(2)}}^T w \\ \dots \\ {x^{(m)}}^T w \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(m)} \end{bmatrix} \right) \qquad X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & .. \\ 1 & x_1^{(m)} \end{bmatrix}$$

Minimize the objective function

$$\arg\min_{w} \quad \mathcal{J}(w) = \frac{1}{m}(Xw - y)^T (Xw - y)$$

Gradient - vectorized solution

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & .. \\ 1 & x_1^{(m)} \end{bmatrix}$$

Non vectorized

$$\mathcal{J}(w_1, w_0) = \frac{1}{m}\sum_{i=1}^{m}\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial \mathrm{J}}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m} 2\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)$$

$$\frac{\partial \mathrm{J}}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m} 2x^{(i)}\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)$$

Vectorized

$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathbf{x}^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \end{bmatrix}$$

$$\nabla_w J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \frac{\partial J}{\partial w_1} \end{bmatrix} = \frac{2}{m}\begin{bmatrix} \sum_{i=1}^{m} 1 \quad \left({x^{(i)}}^T w - y^{(i)}\right) \\ \sum_{i=1}^{m} x^{(i)}\left({x^{(i)}}^T w - y^{(i)}\right) \end{bmatrix}$$

$$\nabla_w J = \frac{2}{m}X^T(Xw - y)$$

$$
\begin{aligned}
\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\partial\left((X\mathbf{w}-\mathbf{y})^T(X\mathbf{w}-\mathbf{y})\right)}{\partial \mathbf{w}} && \text{(definition of } J\text{)} \\
&= \frac{\partial\left((X\mathbf{w})^T X\mathbf{w} - (X\mathbf{w})^T\mathbf{y} - \mathbf{y}^T(X\mathbf{w}) + \mathbf{y}^T\mathbf{y}\right)}{\partial \mathbf{w}} && \text{(expanding brackets)} \\
&= \frac{\partial\left((\mathbf{w}^T X^T X\mathbf{w} - \mathbf{y}^T(X\mathbf{w}) - \mathbf{y}^T(X\mathbf{w}) + \mathbf{y}^T\mathbf{y}\right)}{\partial \mathbf{w}} && ((AB)^T = B^T A^T,\ \mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}) \\
&= \frac{\partial\left((\mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{y}^T(X\mathbf{w})\right)}{\partial \mathbf{w}} && (\mathbf{y}^T\mathbf{y} \text{ is not a function of } \mathbf{w}) \\
&= \frac{\partial\left((\mathbf{w}^T (X^T X)\mathbf{w} - 2(X^T\mathbf{y})^T\mathbf{w})\right)}{\partial \mathbf{w}} && \text{(associativity of matrix multiplication)} \\
&= \frac{\partial\left(\mathbf{w}^T (X^T X)\mathbf{w}\right)}{\partial \mathbf{w}} - 2\frac{\partial\left((X^T\mathbf{y})^T\mathbf{w}\right)}{\partial \mathbf{w}} && \text{(derivatives of sum of functions)} \\
&= 2X^T X\mathbf{w} - 2\frac{\partial\left((X^T\mathbf{y})^T\mathbf{w}\right)}{\partial \mathbf{w}} && \left(\text{for a symmetric } A,\ \frac{\partial \mathbf{x}^T A\mathbf{x}}{\partial \mathbf{x}} = 2A\mathbf{x}\right) \\
&= 2X^T X\mathbf{w} - 2X^T\mathbf{y} && \left(\text{for any vector } \mathbf{u},\ \frac{\partial \mathbf{u}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{u}\right)
\end{aligned}
$$

Gradient descent - vectorized solution

Non vectorized

$$\mathcal{J}(w_1, w_0) = \frac{1}{m}\sum_{i=1}^{m}\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial \mathrm{J}}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m} 2\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)$$

$$\frac{\partial \mathrm{J}}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m} 2x^{(i)}\left(w_0 + w_1 x^{(i)} - y^{(i)}\right)$$

$$w_0 = w_0 - \eta\frac{\partial \mathrm{J}}{\partial w_0}$$

$$w_1 = w_1 - \eta\frac{\partial \mathrm{J}}{\partial w_1}$$

Vectorized

$$X = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & .. \\ 1 & x_1^{(m)} \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad \mathbf{x}^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \end{bmatrix}$$

$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

$$\nabla_w J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \\ \frac{\partial J}{\partial w_1} \end{bmatrix} = \frac{2}{m}X^T(Xw - y)$$

$$\mathbf{w} = \mathbf{w} - \eta\nabla_w J \qquad \mathbf{w} = \mathbf{w} + \eta(-\nabla_w J)$$

Linear Regression Summary

Non vectorized

$$\mathcal{J}(w_1, w_0) = \frac{1}{m}\sum_{i=1}^{m}(w_0 + w_1 x^{(i)} - y^{(i)})^2$$

$$\frac{\partial \mathrm{J}}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m} 2(w_0 + w_1 x^{(i)} - y^{(i)})$$

$$\frac{\partial \mathrm{J}}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m} 2x^{(i)}(w_0 + w_1 x^{(i)} - y^{(i)})$$

$$w_0 = w_0 - \eta\frac{\partial \mathrm{J}}{\partial w_0}$$

$$w_1 = w_1 - \eta\frac{\partial \mathrm{J}}{\partial w_1}$$

Vectorized

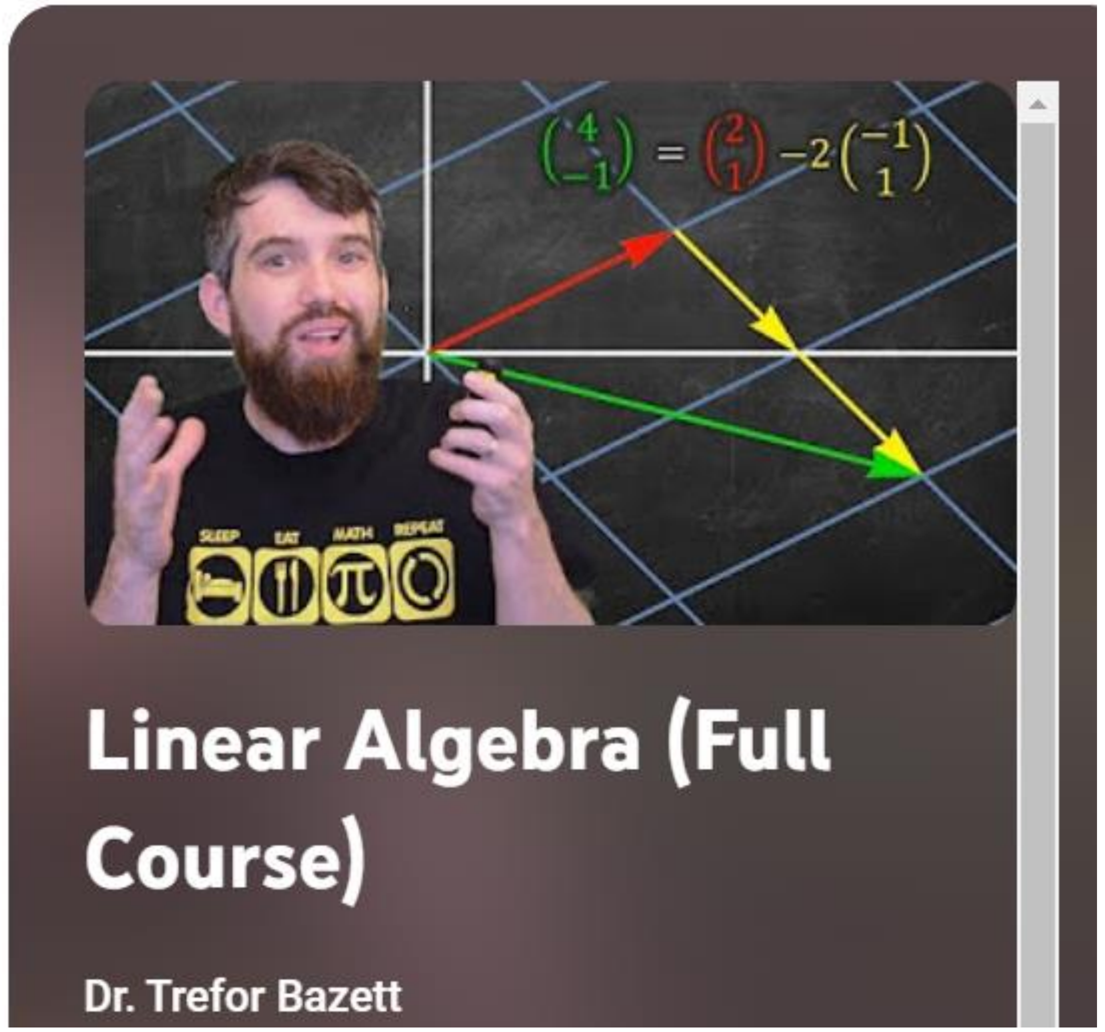
$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

$$\nabla_w J = \frac{2}{m}X^T(Xw - y)$$

$$\mathbf{w} = \mathbf{w} - \eta\nabla_w J$$

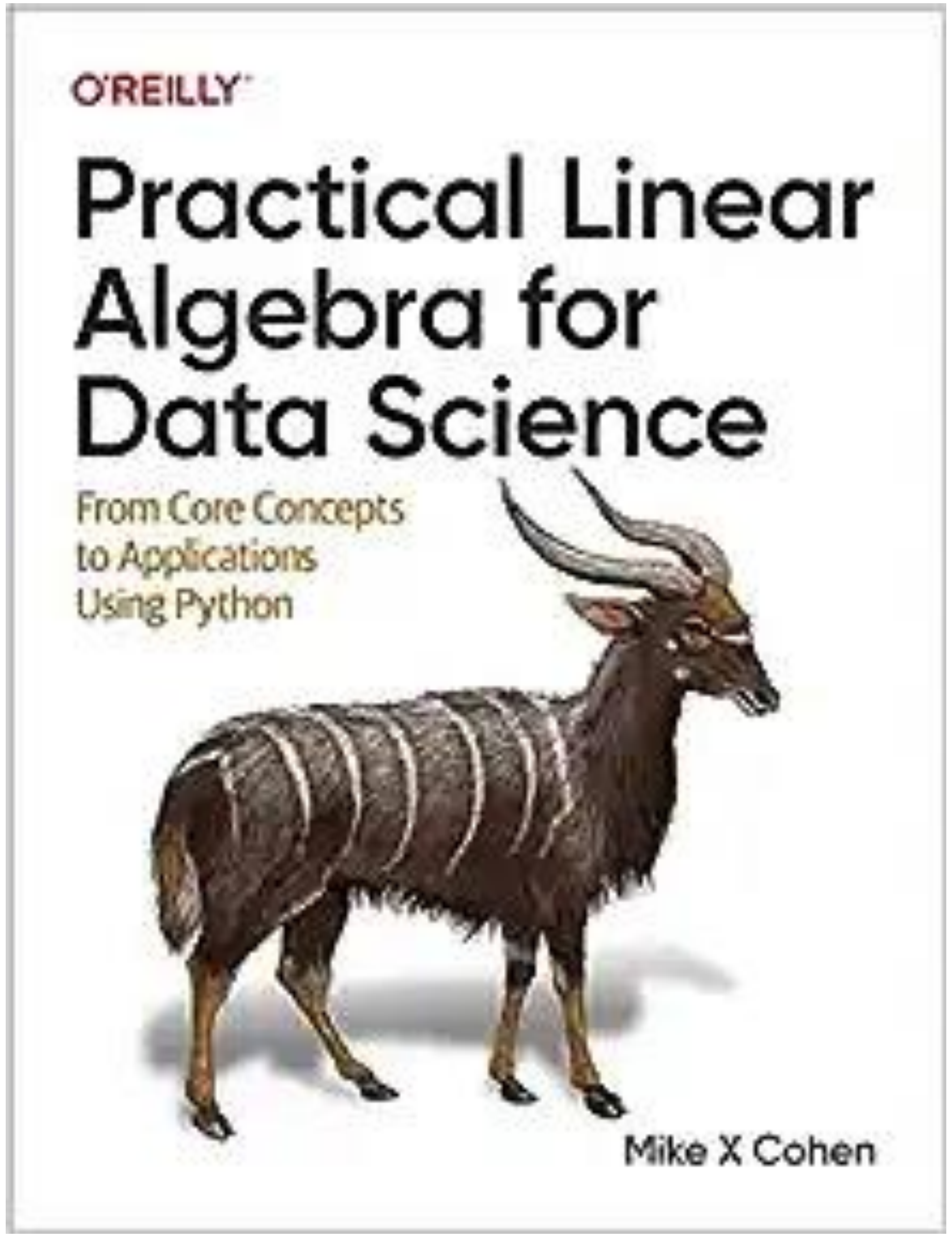
Supplementary Videos (Optional)

- Linear Algebra Full Course – Dr. Trefor Bazett
- https://www.youtube.com/playlist?list=PLHXZ9OQGMqxfUl0tcqPNTJsb7R6BqSLo6
- Fairly easy to understand 👍
- But theoretical, no ML application 👎
- Precursor to Gilbert Strang course – MIT



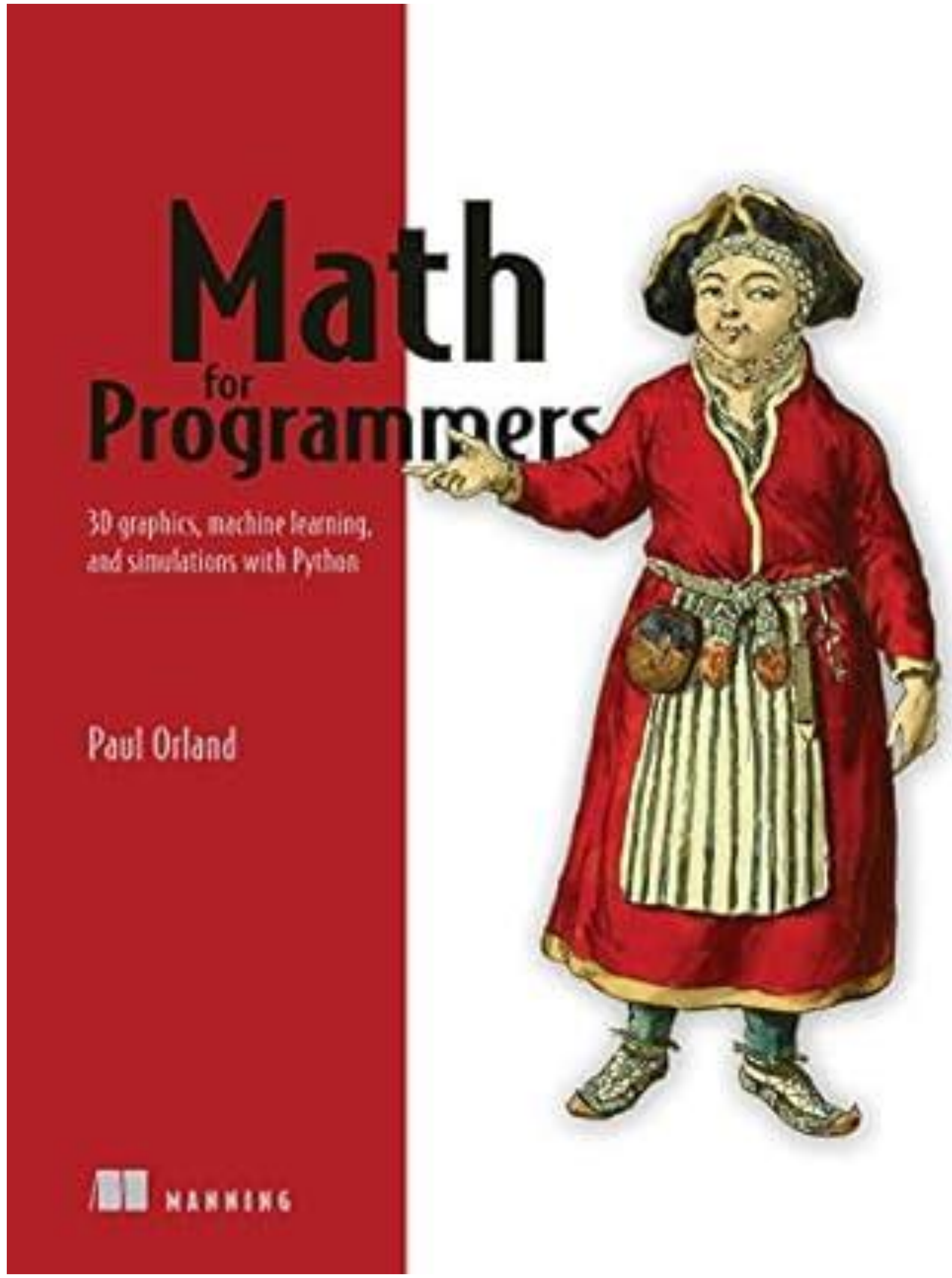
Other readable books

- Easy to read & understand 👍
- Not comprehensive 👎



Other readable books

- Easy to read & understand
- Can code immediately
- Applicable to Machine Learning

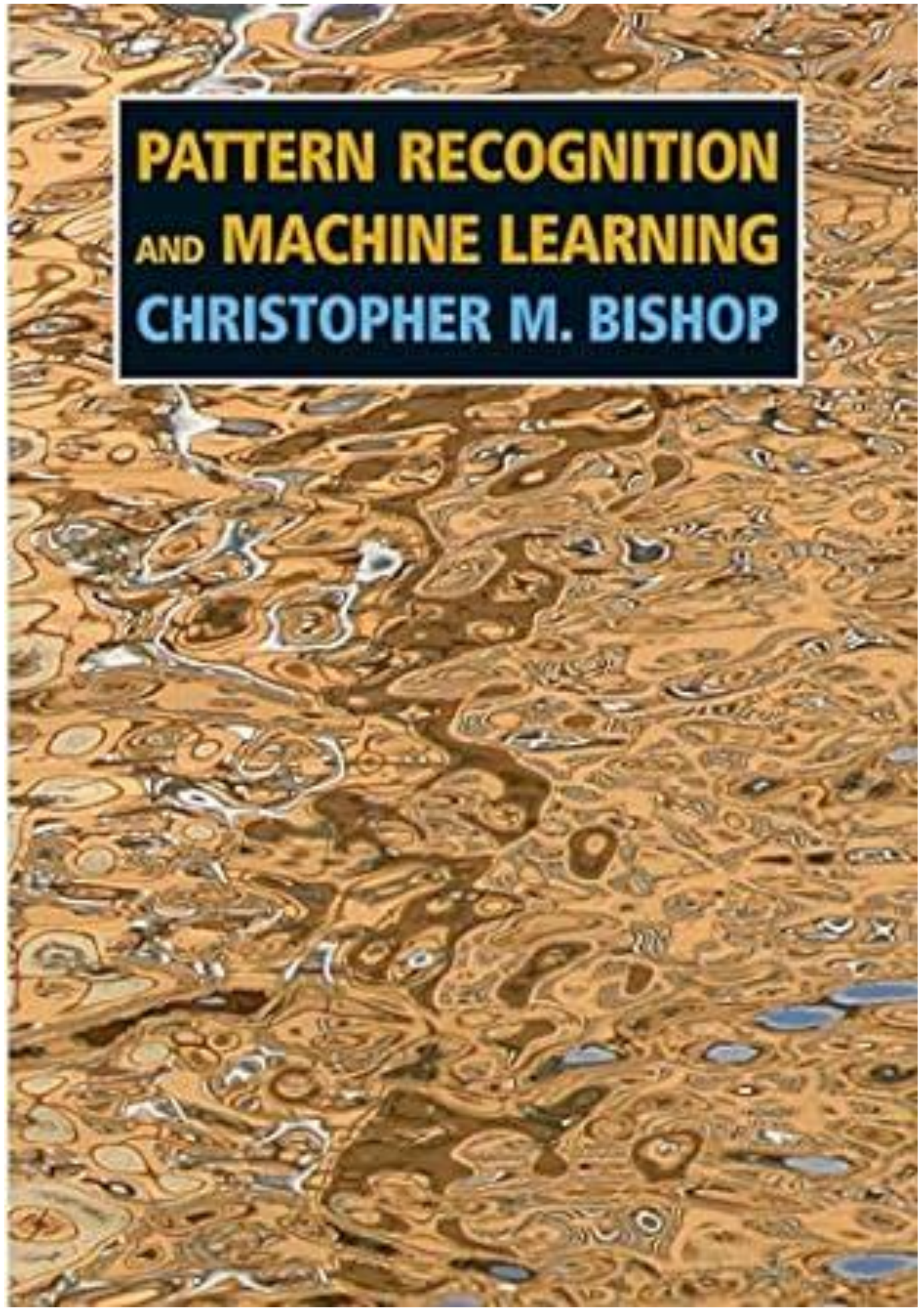


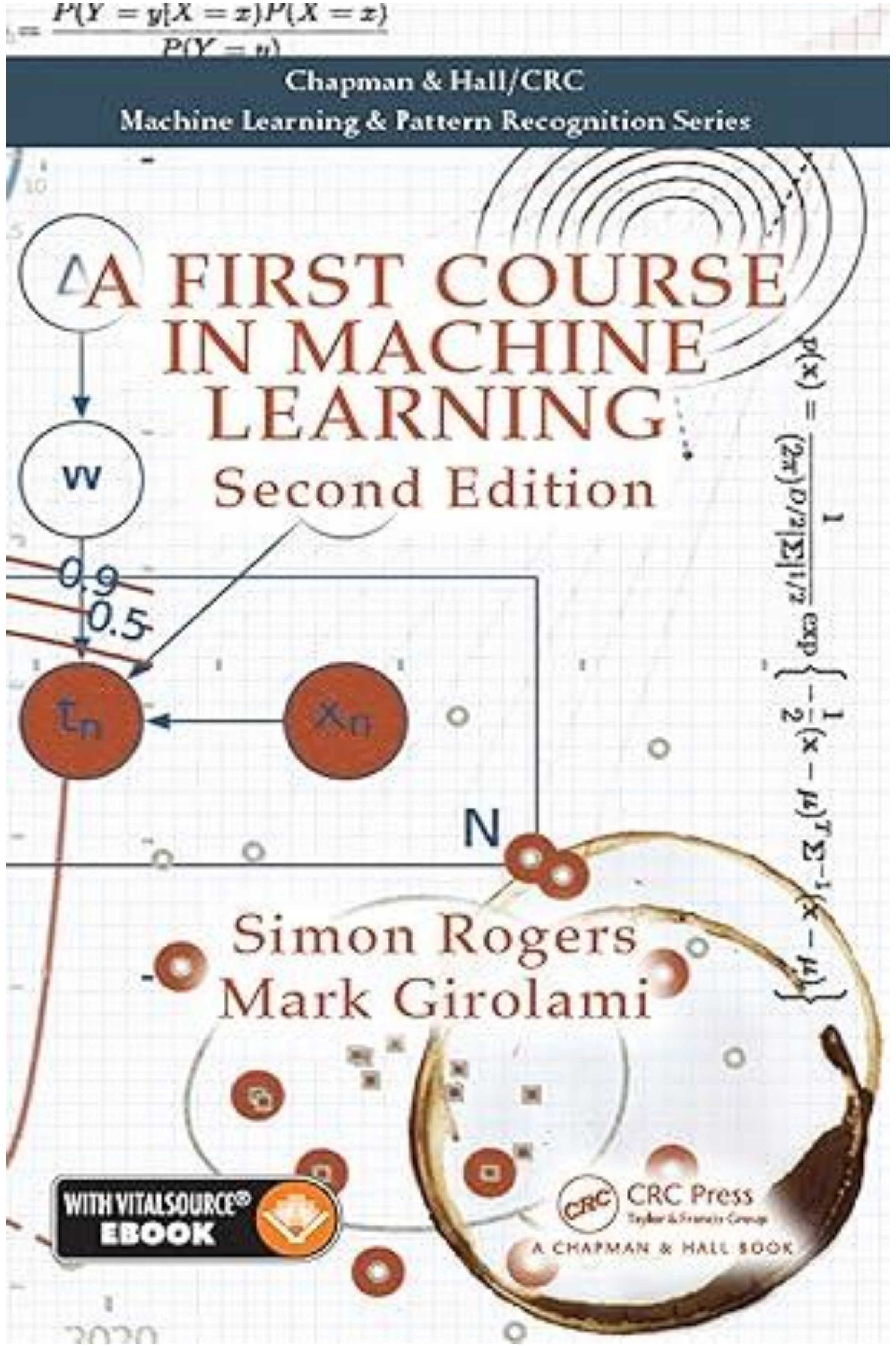
Supplementary Reading (Optional)

- ML books that use maths
- Math for Machine Learning – Richard Han
  - https://www.onlinemathtraining.com/wp-content/uploads/2016/04/Math-for-Machine-Learning-Book-Preview.pdf
- Machine Learning from scratch
  - https://dafriedman97.github.io/mlbook/content/introduction.html
- Mathematical Foundations of Machine Learning
  - https://skim.math.msstate.edu/LectureNotes/Machine_Learning_Lecture.pdf

Supplementary Reading (Optional)

- ML books that use maths
- A comprehensive guide to Machine Learning: UC Berkeley
- Data Driven Science and Engineering (Videos and Book)
  - https://bcourses.berkeley.edu/courses/1487769/files/75906444/download?verifier=8zmDRQWSpuX36ZTiEmSrRt8Eidu5w5bXPliIBaud&wrap=1
- Steve Brunton University of Washington
  - Data Science & ML: Mathematical & Statistical Methods
  - https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series
A FIRST COURSE IN MACHINE LEARNING
Second Edition
Simon Rogers
Mark Girolami
WITH VITALSOURCE® EBOOK
CRC Press
A CHAPMAN & HALL BOOK


QUESTIONS