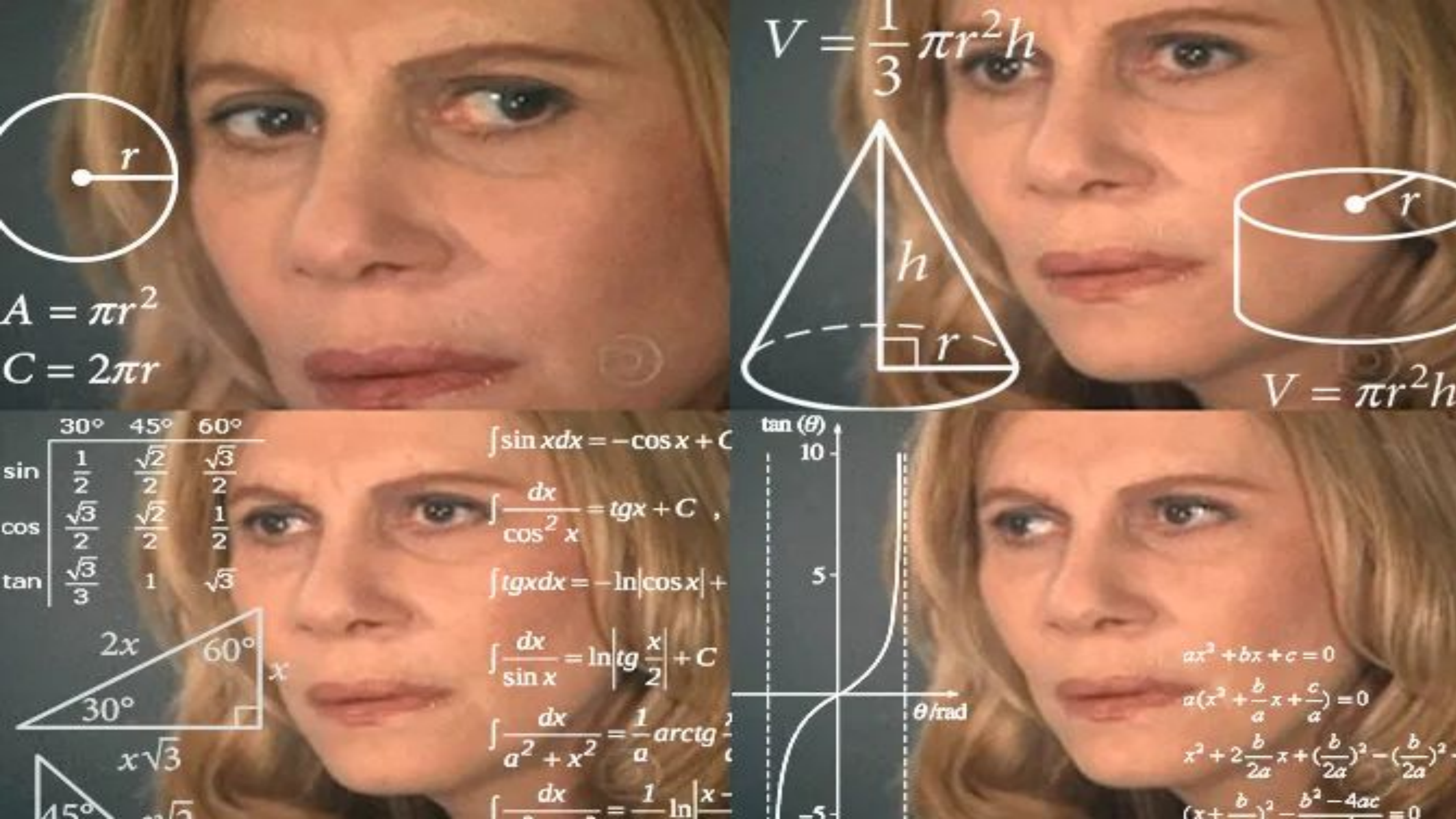
Mechanistic Interpretability 101

Episode 3

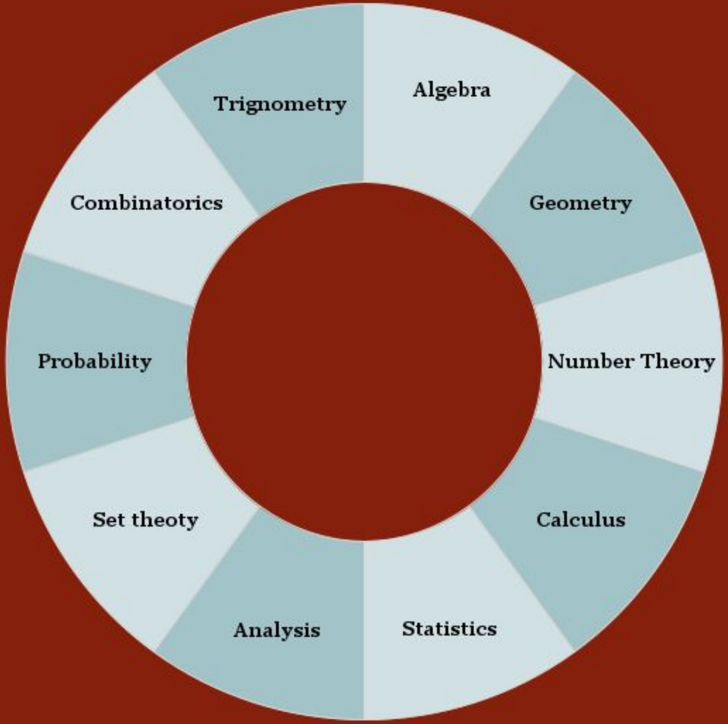
Calculus



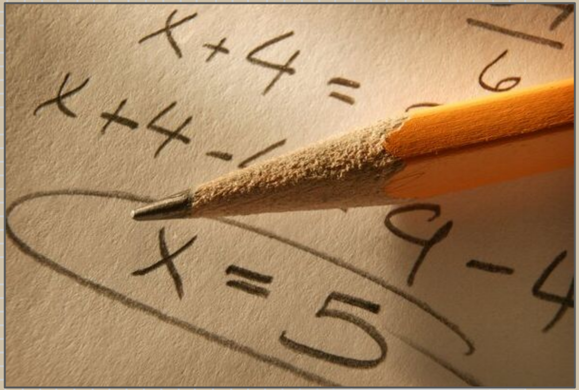
r
$A = \pi r^2$
$C = 2\pi r$
$V = \frac{1}{3}\pi r^2 h$
h
r
r
$V = \pi r^2 h$
30° 45° 60°
sin $\frac{1}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{\sqrt{3}}{2}$
cos $\frac{\sqrt{3}}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{1}{2}$
tan $\frac{\sqrt{3}}{3}$ 1 $\sqrt{3}$
$2x$
60°
x
30°
$x\sqrt{3}$
45°
$\int \sin x dx = -\cos x + C$
$\int \frac{dx}{\cos^2 x} = tgx + C$
$\int tgxdx = -\ln|\cos x| +$
$\int \frac{dx}{\sin x} = \ln\left|tg\frac{x}{2}\right| + C$
$\int \frac{dx}{a^2+x^2} = \frac{1}{a} arctg$
tan (θ)
10
5
-5
θ /rad
$ax^2 + bx + c = 0$
$a(x^2 + \frac{b}{a}x + \frac{c}{a}) = 0$


CALCULUS, I SEE
YOU!

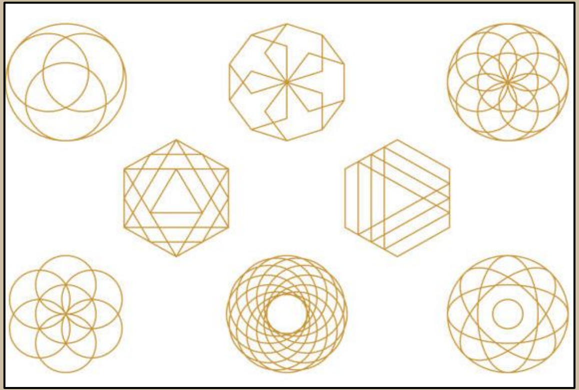
Fundamental branches of Math



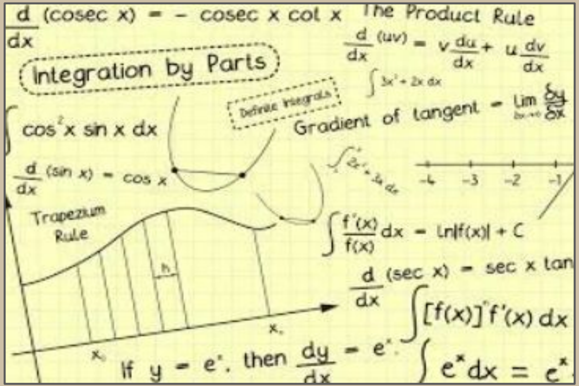
Fundamental branches of Math



Algebra



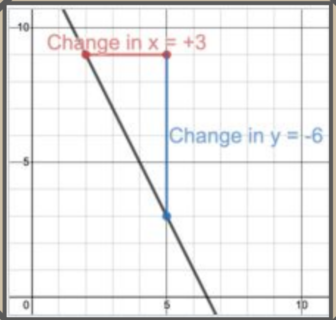
Geometry



Calculus

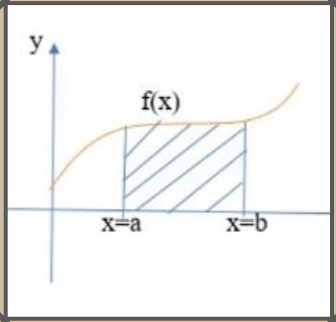
Let us ‘Change’ this outlook about Change today

Calculus



DIFFERENTIAL

SLOPE



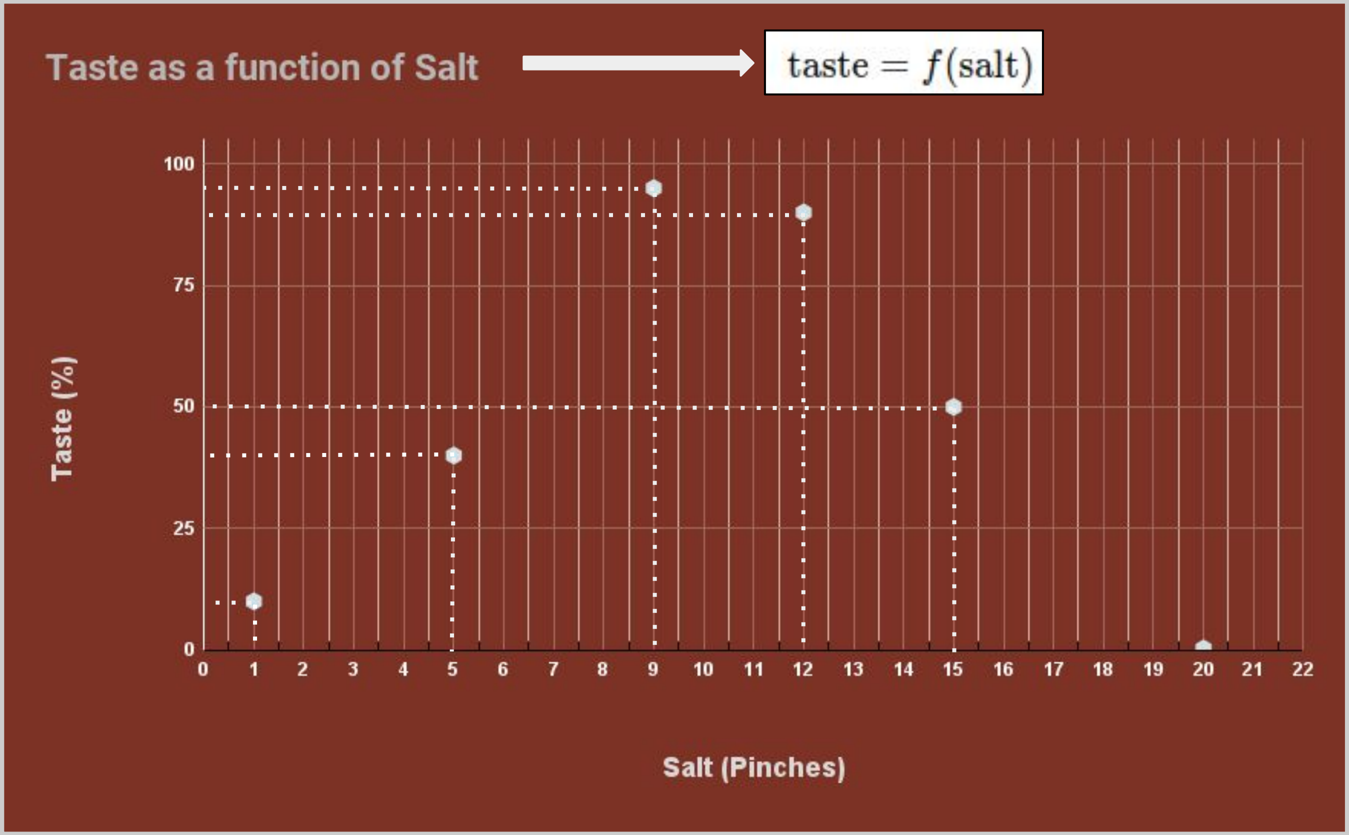
INTEGRAL

AREA UNDER THE CURVE

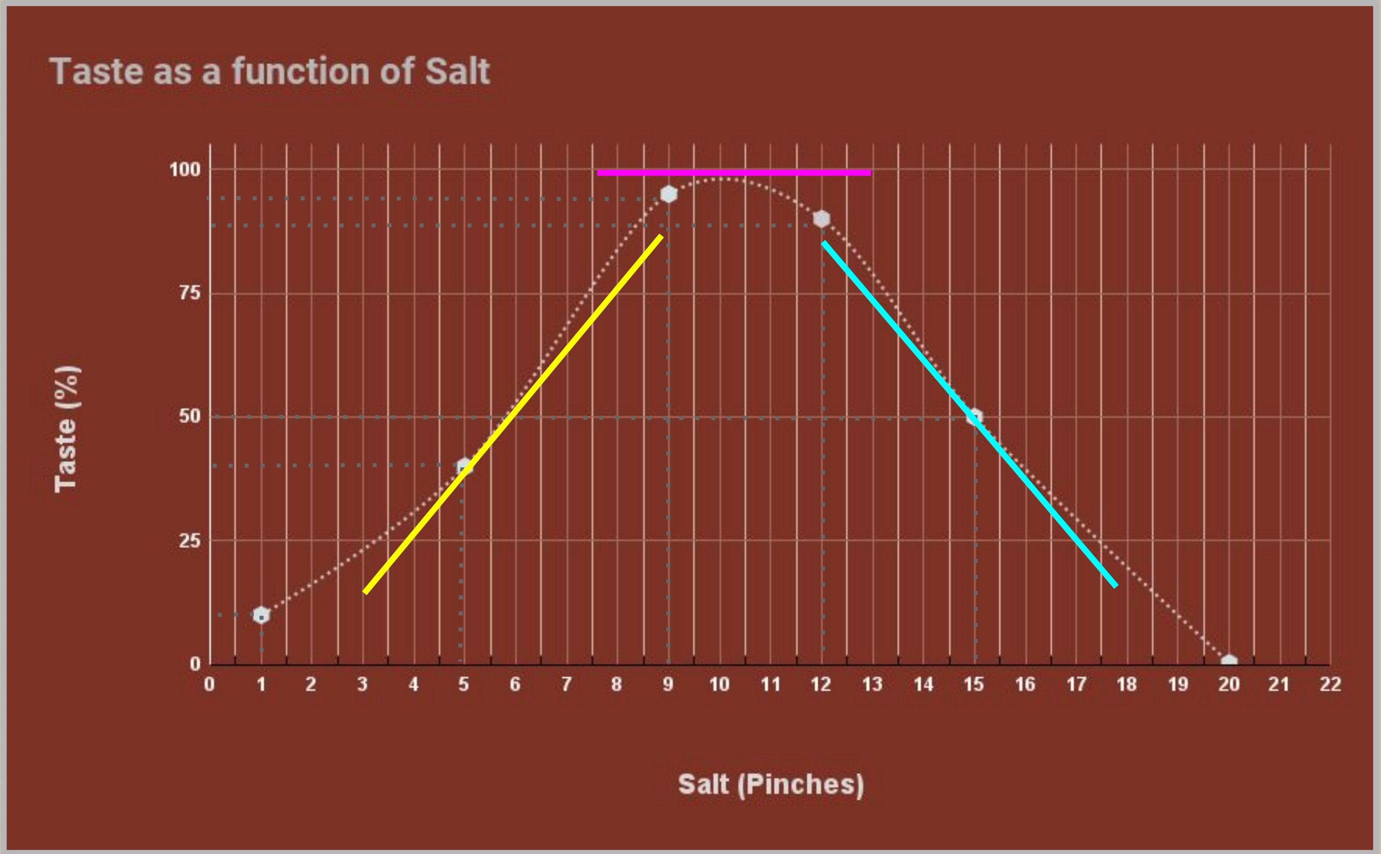

SALT

Understanding Functions

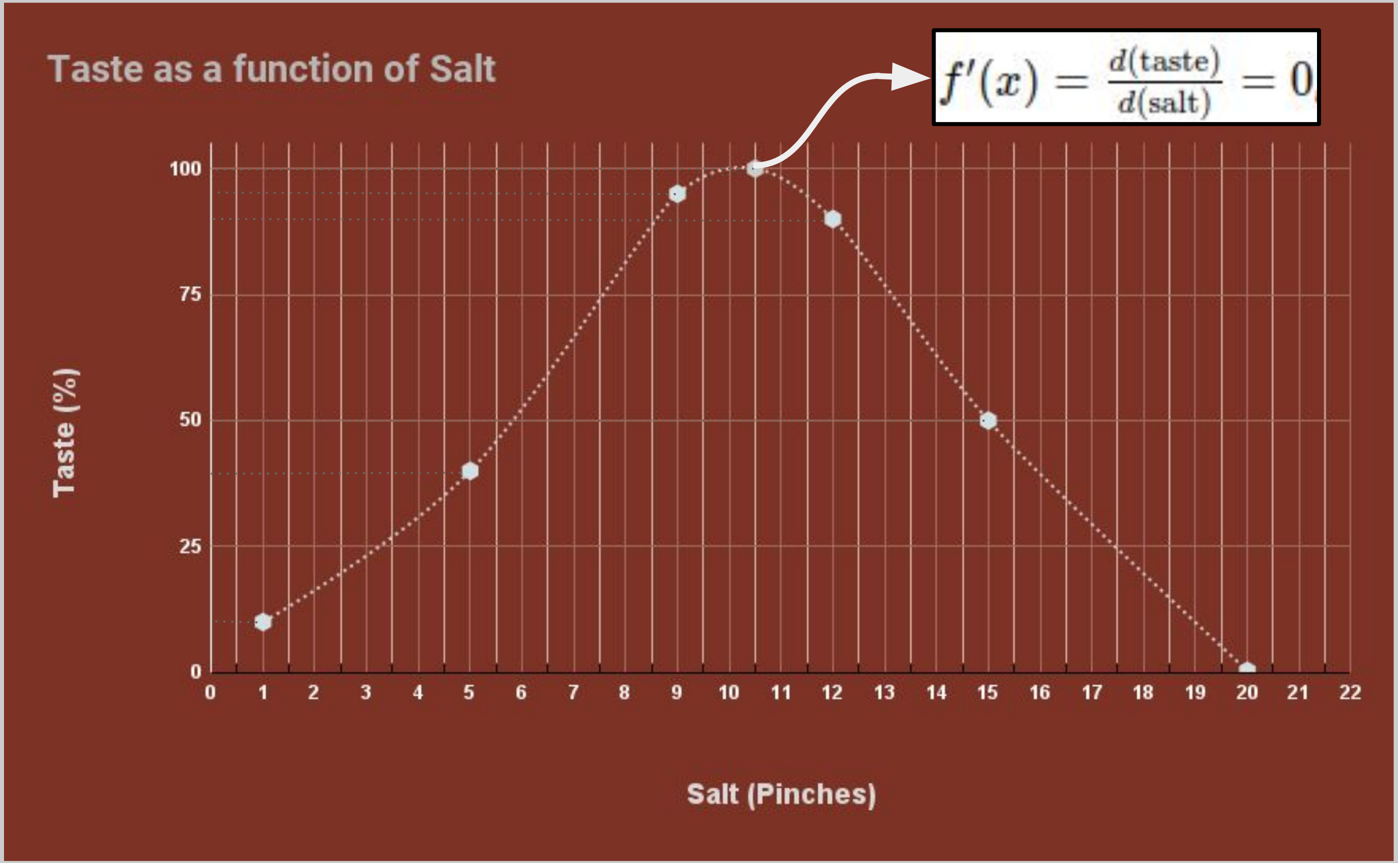
Salt (in Pinches)	Taste (in %)
1	10
5	40
9	95
12	90
15	50
20	0.3



Understanding Derivatives



Understanding Derivatives



Limits

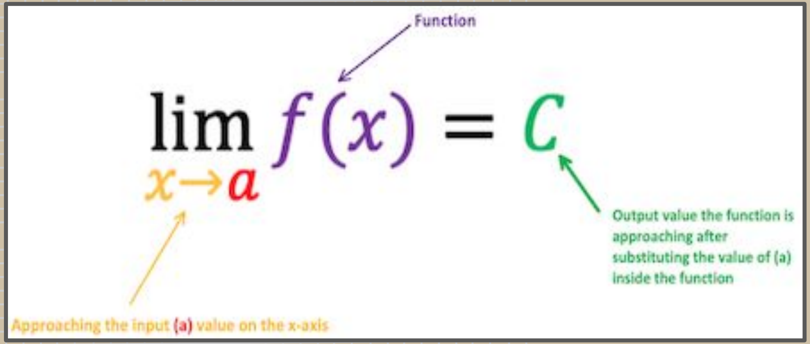
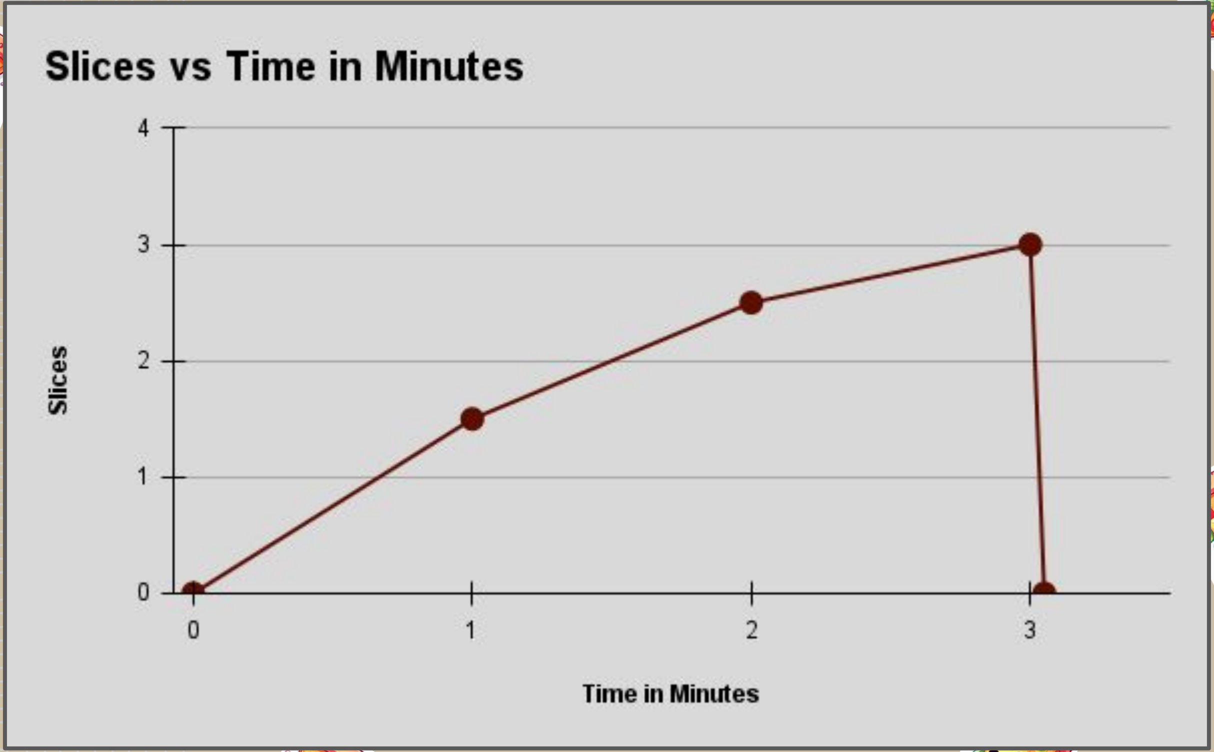
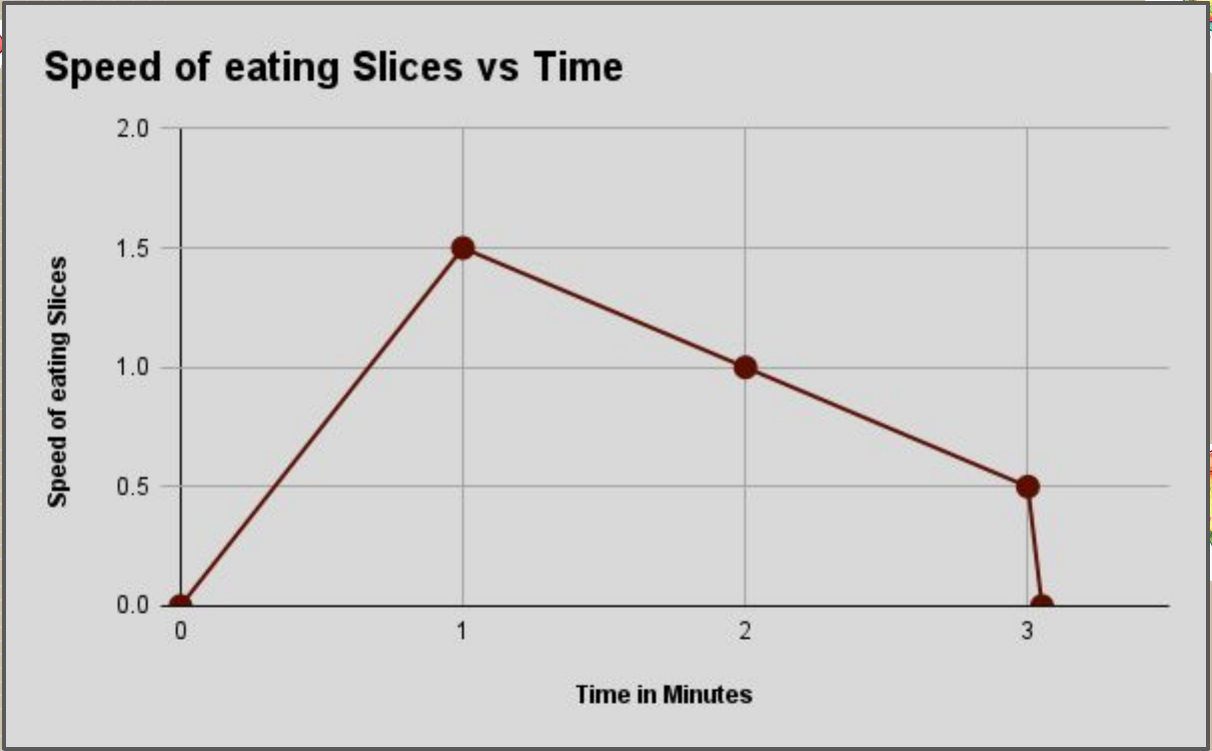


Image Source: Table Salt under a Microscope by Stephanie Huesler

Let’s eat Pizza



Second Derivatives



Derivatives

Common Derivatives

$$\frac{d}{dx}(x)=1$$

$$\frac{d}{dx}(\sin x)=\cos x$$

$$\frac{d}{dx}(\cos x)=-\sin x$$

$$\frac{d}{dx}(\tan x)=\sec^2 x$$

$$\frac{d}{dx}(\sec x)=\sec x\tan x$$

$$\frac{d}{dx}(\csc x)=-\csc x\cot x$$

$$\frac{d}{dx}(\cot x)=-\csc^2 x$$

$$\frac{d}{dx}(\sin^{-1} x)=\frac{1}{\sqrt{1-x^2}}$$

$$\frac{d}{dx}(\cos^{-1} x)=-\frac{1}{\sqrt{1-x^2}}$$

$$\frac{d}{dx}(\tan^{-1} x)=\frac{1}{1+x^2}$$

$$\frac{d}{dx}(a^x)=a^x\ln(a)$$

$$\frac{d}{dx}(\mathbf{e}^x)=\mathbf{e}^x$$

$$\frac{d}{dx}(\ln(x))=\frac{1}{x},\ \ x>0$$

$$\frac{d}{dx}(\ln|x|)=\frac{1}{x},\ \ x\neq 0$$

$$\frac{d}{dx}(\log_a(x))=\frac{1}{x\ln a},\ \ x>0$$

Differentiation Rules

Constant Rule	$\frac{d}{dx}[C] = 0$
Power Rule	$\frac{d}{dx} x^n = nx^{n-1}$
Product Rule	$\frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$
Quotient Rule	$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$
Chain Rule	$\frac{d}{dx}[f(g(x))] = f'(g(x))\, g'(x)$

Partial derivatives

Multivariable Function: $f(x, y) = x^2 y$

$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial}{\partial x} x^2 y}_{\substack{\text{Treat } y \text{ as constant;} \\ \text{take derivative.}}} = 2xy$$

$$\frac{\partial f}{\partial y} = \underbrace{\frac{\partial}{\partial y} x^2 y}_{\substack{\text{Treat } x \text{ as constant;} \\ \text{take derivative.}}} = x^2 \cdot 1$$

Chain Rule

If $h(x) = f(g(x))$, then:

$$h'(x) = f'(g(x)) \cdot g'(x)$$

Given: $h(x) = (2x+3)^4$

Let $f(u) = u^4$ and $g(x) = 2x+3$.

Find $f'(u)$ and $g'(x)$:

$$f'(u) = 4u^3$$

$$g'(x) = 2$$

Now, using the chain rule $h'(x) = f'(g(x)) \cdot g'(x)$:

$$h'(x) = 4(2x+3)^3 \cdot 2$$

$$h'(x) = 8(2x+3)^3$$

Gradient Descent

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\[2ex] \dfrac{\partial f}{\partial y} \\[2ex] \vdots \end{bmatrix}$$

$$\nabla f(x_0, y_0) = \begin{bmatrix} \dfrac{\partial f}{\partial x}(x_0, y_0) \\[3ex] \dfrac{\partial f}{\partial y}(x_0, y_0) \end{bmatrix}$$

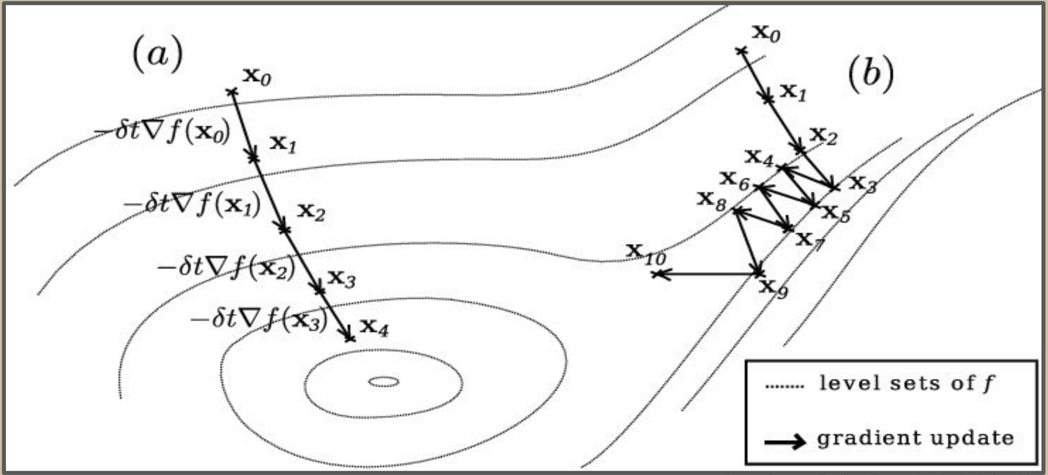
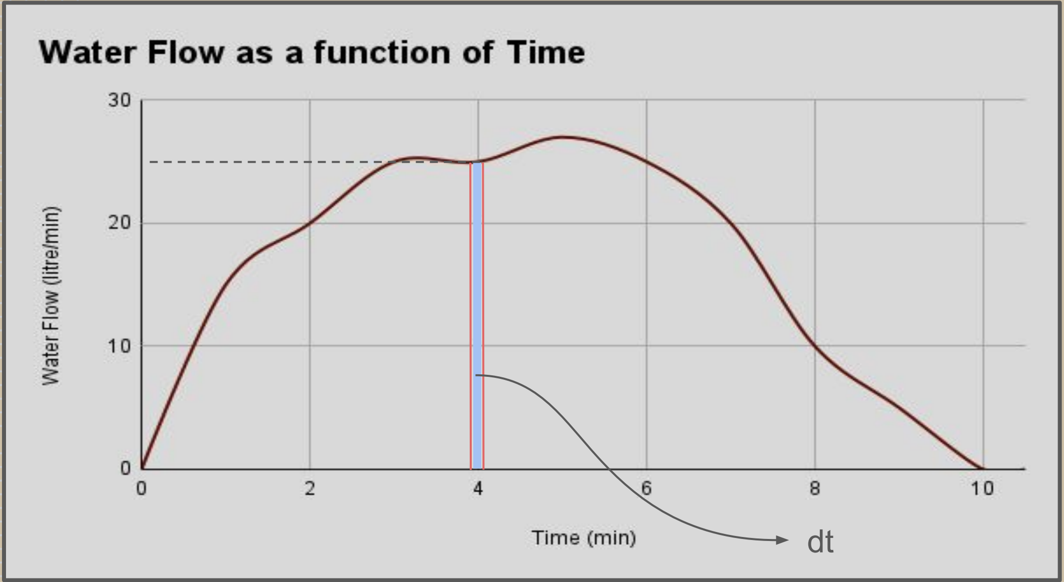


Image Source: Gradient Descent by Ludovic Arnold

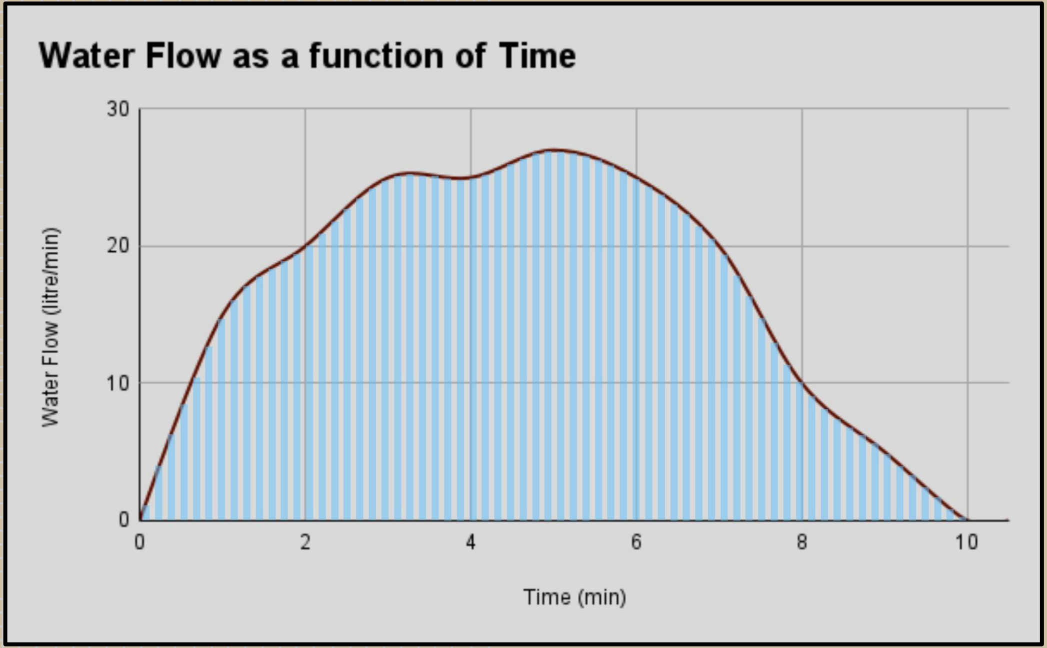
Integration



How much total water flowed in the time interval dt?

25dt

Integration



Total water flowed from $t = 0$ to $t = 10$:

$$\text{Total Water} = \int_0^{10} f(t)\,dt$$

Sum over all intervals from $t = 0$ to $t = 10$:

$$\int_0^{10} f(t)\,dt \approx \sum (\text{flow rate in each interval}) \times dt$$

Integration

Given: Water flow rate as a function of time $f(t)$.

Total water flowed from $t = 0$ to $t = 10$:

$$\text{Total Water} = \int_0^{10} f(t)\,dt$$

In the diagram, if $f(t) \approx 25$ liters/min at a specific interval dt:

$$\text{Water in interval} = f(t) \cdot dt \approx 25 \cdot dt$$

Sum over all intervals from $t = 0$ to $t = 10$:

$$\int_0^{10} f(t)\,dt \approx \sum (\text{flow rate in each interval}) \times dt$$

Recap

- Fundamental concept of Derivatives
- Understanding Types of Slopes and the insights they give
- Second Derivatives
- Partial Derivatives
- Rules of Differentiation
- Chain Rule
- Gradient Descent
- Integration