A 90-MINUTE WORKSHOP

The Math of AI

Using the elegant mathematics behind modern AI to motivate first-year calculus and linear algebra

For high school math teachers

and first/second-year university math instructors

Narrative arc inspired by Why Machines Learn (Ananthaswamy, 2024)

Goals & Agenda

What you will leave with

- A clear map from first-year math to modern AI
- Two by-hand worked examples (linear algebra + calculus)
- An honest, math-first model of how an LLM works
- A thought provoking surprise at the end

90-minute agenda

0:00	5 min	Welcome & goals
0:05	15 min	History: perceptron → LLMs
0:20	5 min	Why first-year math is enough
0:25	20 min	Linear algebra in AI + activity
0:45	20 min	Calculus in AI + activity
1:05	15 min	LLMs: pulling it together
1:20	8 min	Classroom pair-share
1:28	2 min	Wrap-up & reflection

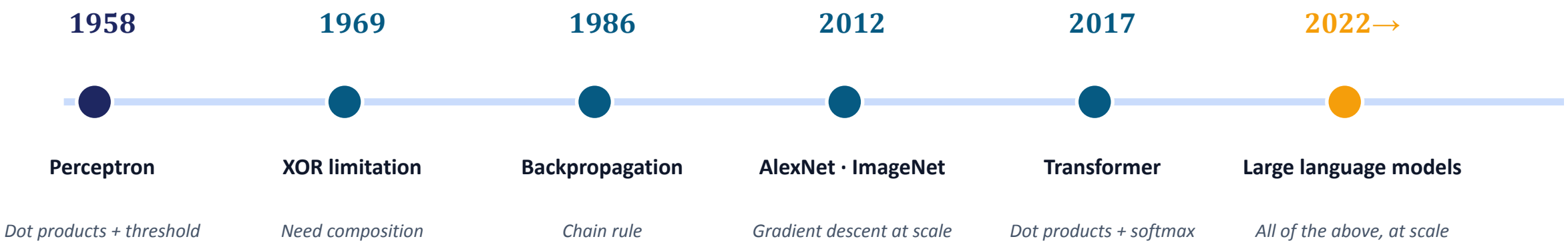
01

A Brief History of AI

From the perceptron (1958) to large language models

Every historical advance was an application of accessible mathematics

Six milestones, one mathematical thread



The thread: *nothing in this story is conceptually beyond a strong first-year student. The barrier was always engineering, data, and computing — not new mathematics.*

1958 — The perceptron

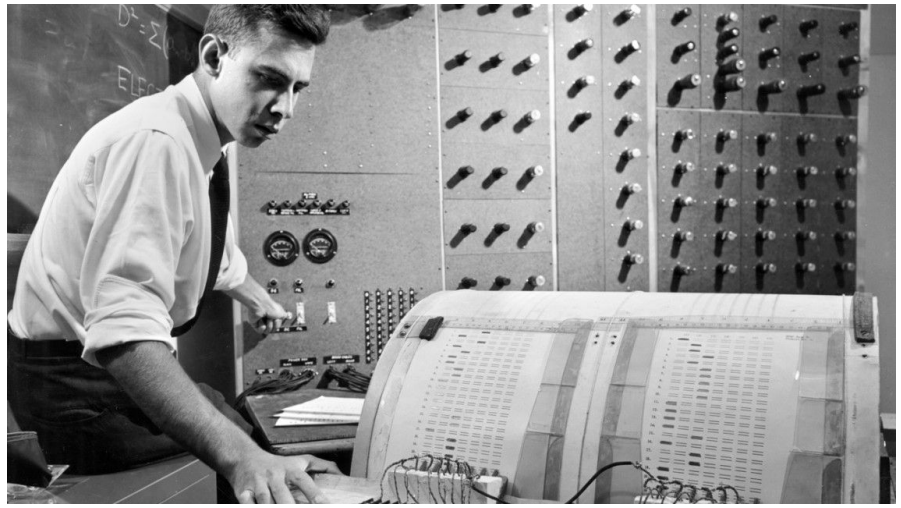
Rosenblatt's single neuron

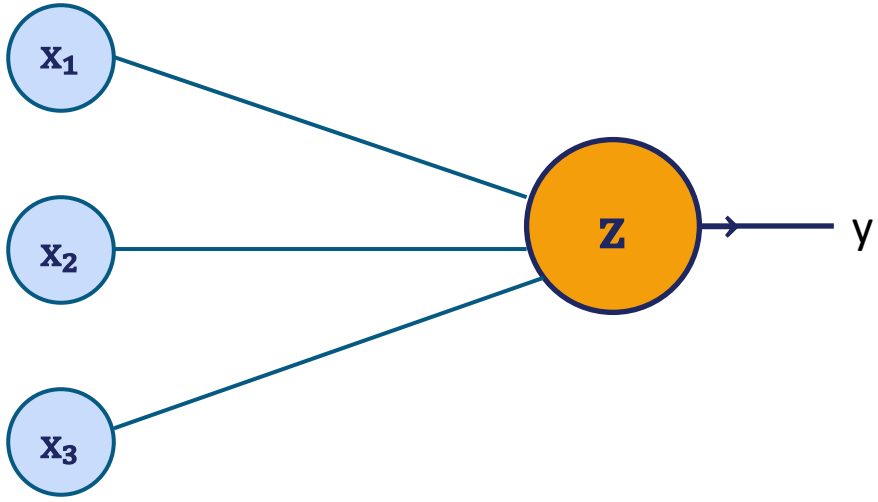
- Take an input vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$
- Compute a weighted sum $z = \mathbf{w} \cdot \mathbf{x} = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$
- Fire if z > θ; stay silent otherwise.
- Adjust weights when the prediction is wrong.

THE MATH

$$\mathbf{z = w \cdot x}$$

A dot product. That's it. That is the first learning algorithm in AI history.





A single neuron — the original AI architecture

SECTION TRANSITION

Why first-year and second -year math is enough

The vocabulary of modern AI fits on one slide.

Vectors & matrices

→ Embeddings, transformations

In AI
Embeddings, transformations

Dot products

→ Similarity, attention

In AI
Similarity, attention

Derivatives

→ Gradients, sensitivity

In AI
Gradients, sensitivity

Optimization

→ Loss minimization

In AI
Loss minimization

Chain rule

→ Backpropagation

In AI
Backpropagation

Probability & log

→ Softmax, cross-entropy

In AI
Softmax, cross-entropy

02

Linear Algebra in AI

Vectors, dot products, and the geometry of meaning

20 minutes — content + activity

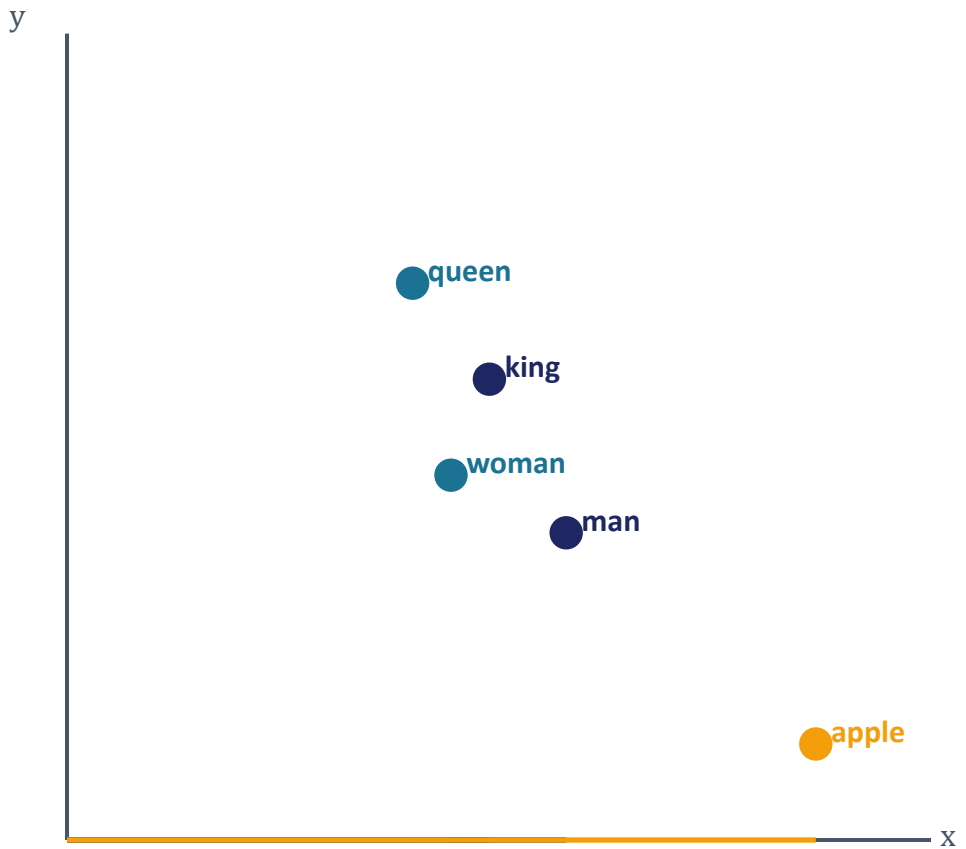
A word can be a vector

Each word in an AI system is represented as a vector of numbers.

- Similar meanings → similar vectors
  - Contrast this with similarity between species based on similar attributes
- Direction encodes meaning more than magnitude
- Cosine of the angle = similarity score
- Inside an LLM, vectors are ~ 10 000-dimensional

king − man + woman ≈ queen

Embedding arithmetic captures real semantic relations (meaning!).



Words plotted as 2-D vectors (illustrative)

Cosine similarity = first-year linear algebra

FORMULA

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||||\mathbf{b}||}$$

$$\text{where } \mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 \text{ and } \|a\| = \sqrt{a_1^2 + a_2^2}$$

≈ 1

Very similar

cos(θ)

≈ 0

Unrelated (orthogonal dot product)

cos(θ)

≈ −1

Opposite in meaning

cos(θ)

Activity 1 — Cosine similarity by hand

8 MINUTES · IN PAIRS

Compute the cosine similarity for each pair and decide: which pair is more similar?

Pair A

$$\mathbf{a} = (4, 3)$$

$$\mathbf{b} = (3, 4)$$

Compute: $a \cdot b$, $\|a\|$, $\|b\|$, $\cos(\theta)$

Suggested word interpretation: doctor / nurse

Pair B

$$\mathbf{a} = (4, 3)$$

$$\mathbf{b} = (1, 6)$$

Compute: $a \cdot b$, $\|a\|$, $\|b\|$, $\cos(\theta)$

Suggested word interpretation: doctor / orange

If you have more time, try a three-component example. Extend this example in a way that makes mathematical and meaning sense.

03

Calculus in AI

Loss landscapes, gradient descent, and the chain rule

20 minutes — content + activity

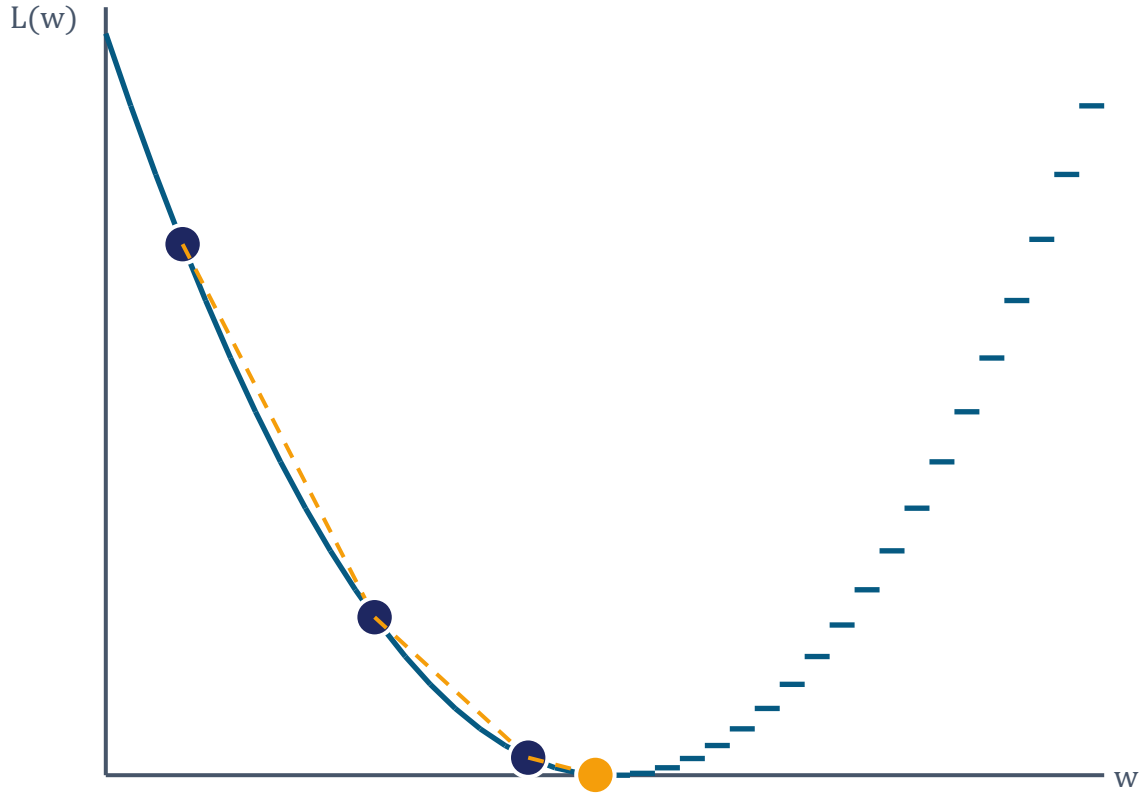
Training a model = minimizing a function

Choose parameters to minimize the loss.

- A loss function L(w) measures how wrong predictions are.
- Training = find w that minimizes L.
- First-year approach: set L'(w) = 0 and solve.
- AI's approach: take small steps downhill instead.
  - Think numeric differentiation
  - Think Euler’s method

Why not just solve L'(w) = 0?

Modern AI has billions of parameters. No closed form. The iterative approach is the only practical way.



Gradient descent rolls downhill

The update rule that trains every modern AI

GRADIENT DESCENT

$$w^{(n+1)} = w^{(n)} - \alpha \cdot f'(w^{(n)})$$

α is the learning rate. Small α: stable but slow. Large α: fast but may overshoot. Euler's method with a step size of alpha.

Derivative

Tells us the slope at the current point.

Step downhill

We move opposite the gradient direction.

Chain rule

Same idea, applied through many layers = backpropagation.

Activity 2 — Gradient descent by hand

8 MINUTES · IN PAIRS

Three iterations on the same parabola your students already know.

Setup

$$f(w) = w^2 + 2w$$

$$w_0 = 4, \quad \alpha = 0.1$$

Step 1: compute f ′(w) symbolically.

Step 2: do 3 iterations of the update rule.

Step 3: compare with the minimum of the function via calculus one techniques.

Stretch

What if $\alpha = 0.6$?

What if $\alpha = 1.1$?

Predict before you compute. Then verify.

Discuss: how do learning-rate failures show up when the loss landscape is more complicated than a parabola?

If you have time, google Kepler's Des solved using Euler's method error and see what happens and how error can grow

04

Large Language Models

All of the above, at unprecedented scale

15 minutes — putting the math together

Four stages, all powered by first-year math

1

Embedding

Each token becomes a vector. Similar meanings → similar vectors.

LINEAR ALGEBRA

2

Attention

Dot products between tokens, normalized by softmax. Output = weighted sum of vectors.

LINEAR ALGEBRA + PROBABILITY

3

Prediction

Output vector → probability distribution over next token. Pick or sample.

MATRIX MULTIPLY + SOFTMAX

4

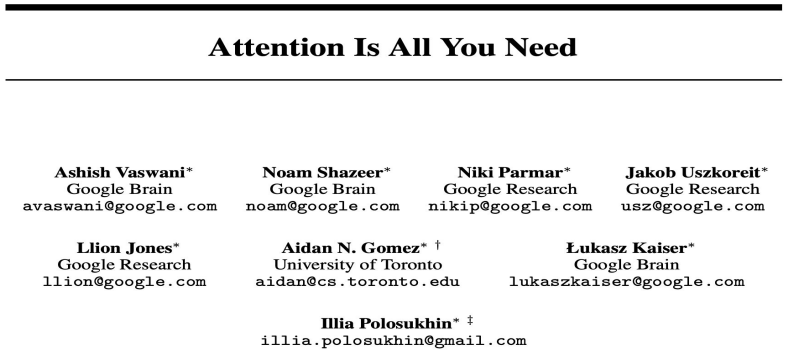
Training

Adjust billions of parameters by gradient descent on prediction error.

CHAIN RULE (CALCULUS)

Attention is dot products + softmax

The single equation behind every modern language model.



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

QK^T

Dot products between every pair of tokens. Linear algebra.

division by $\sqrt{d_k}$

Scale to keep gradients well-behaved.

softmax

Convert scores into a probability distribution that sums to 1.

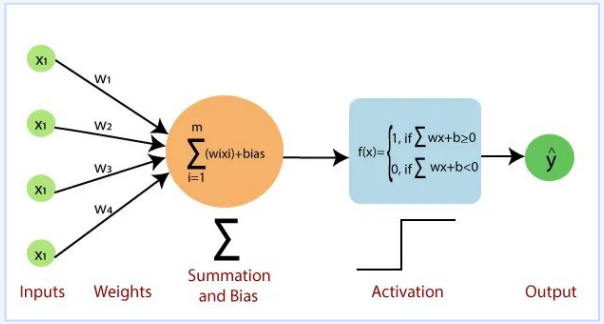
V

Weighted sum of value vectors. A linear combination.

The same math, at three scales

1958

Perceptron



PARAMETERS

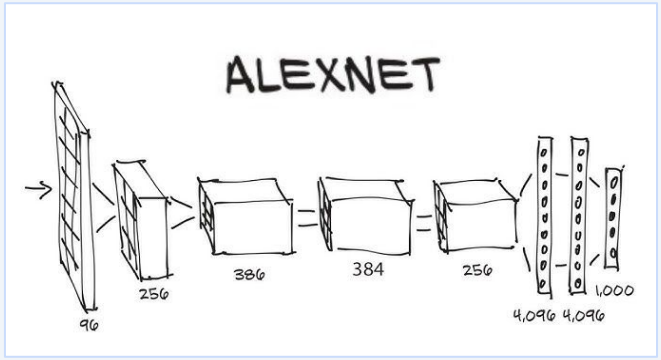
≈ 10

MATHEMATICAL CORE

Dot product

2012

AlexNet



PARAMETERS

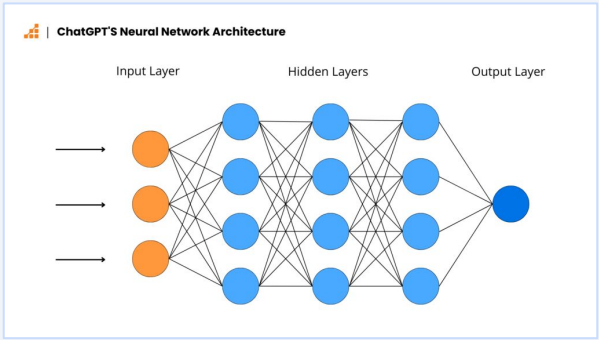
≈ 60 M

MATHEMATICAL CORE

Dot product + chain rule

TODAY

GPT/Claude/Deepseek



PARAMETERS

≈ 100 B – 1 T

MATHEMATICAL CORE

Same — at scale

Scale changed. The math did not.

Take it back to your classroom

PAIR-SHARE · 4 MINUTES

Pick one topic from your syllabus this term. Where could you embed an AI-motivated example?

1. **Dot product**

   Cosine similarity for music or document recommendation

2. **Matrix multiplication**

   Image pixels passing through a Convolutional Neural Network layer (more advanced)

3. **Least squares**

   Predicting student grades or housing prices

4. **Derivatives**

   Gradient descent to minimize a simple loss

5. **Chain rule**

   Backpropagation through a tiny two-layer net

6. **Probability & log**

   Cross-entropy as a measure of surprise

Thank you and a thought about using AI

Go teach the math behind the machine but should you or will you use the machine

EXIT REFLECTION

3 things you learned today

2 questions you are still curious about

1 How you are going to use AI to teach or not?

RECOMMENDED READING

Why Machines Learn

The Elegant Math Behind Modern AI

Anil Ananthaswamy · 2024

Plus: 3Blue1Brown (videos), Karpathy's Neural Networks Zero-to-Hero, TensorFlow Playground.

I need help to prepare a 90 minute workshop

The topic is using the math of AI to motivate learning mathematics for high school math teachers and university teachers of first and second year math.

First year calculus and linear algebra

large language models plus in the intro I would like a history of AI starting with the percetron

Much like the book Why Machines Learn the elegant math behind modern AI by Anil Ananthaswamy

Can you produce me an abstract of this workshop for the conference program?