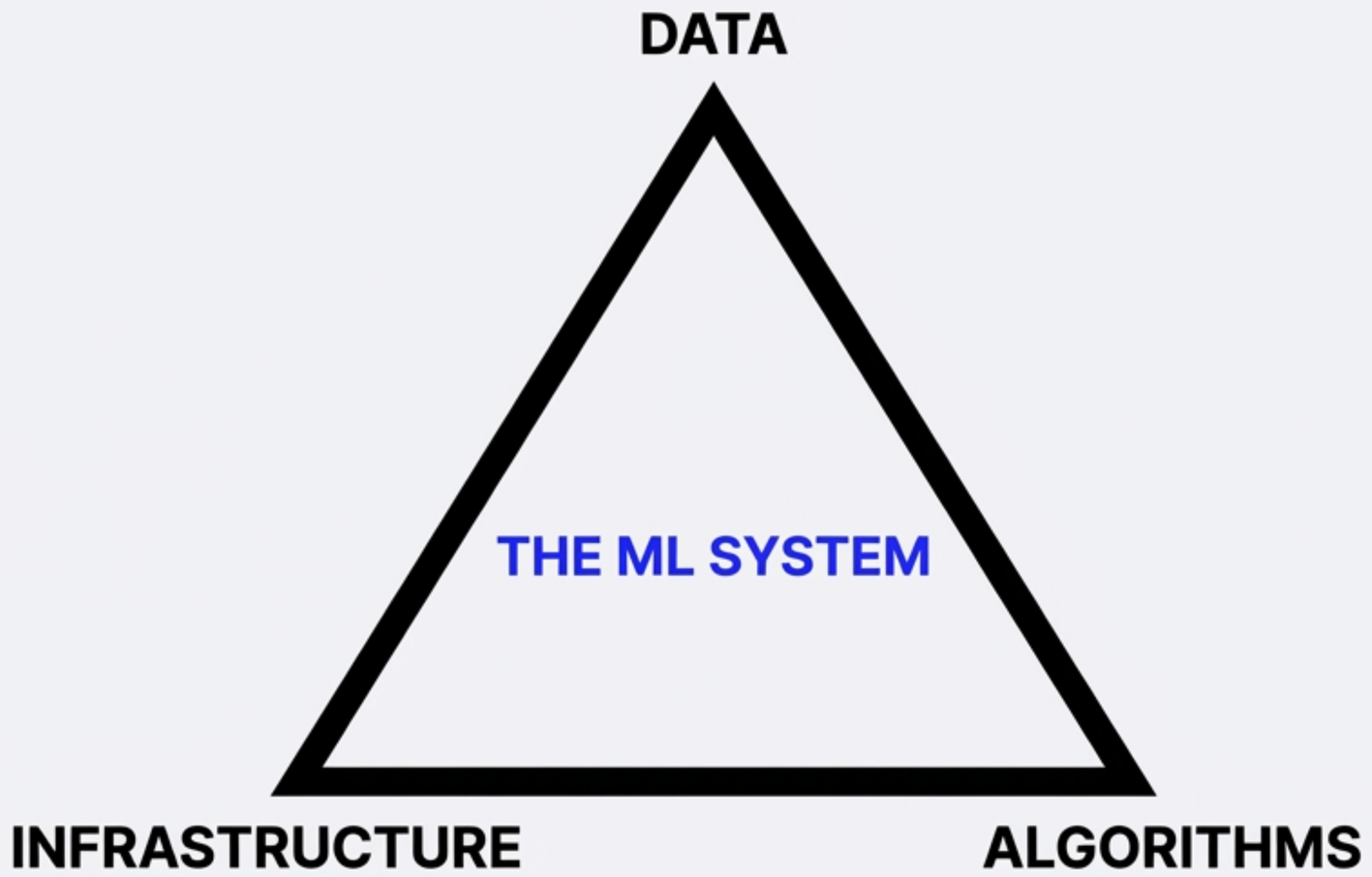
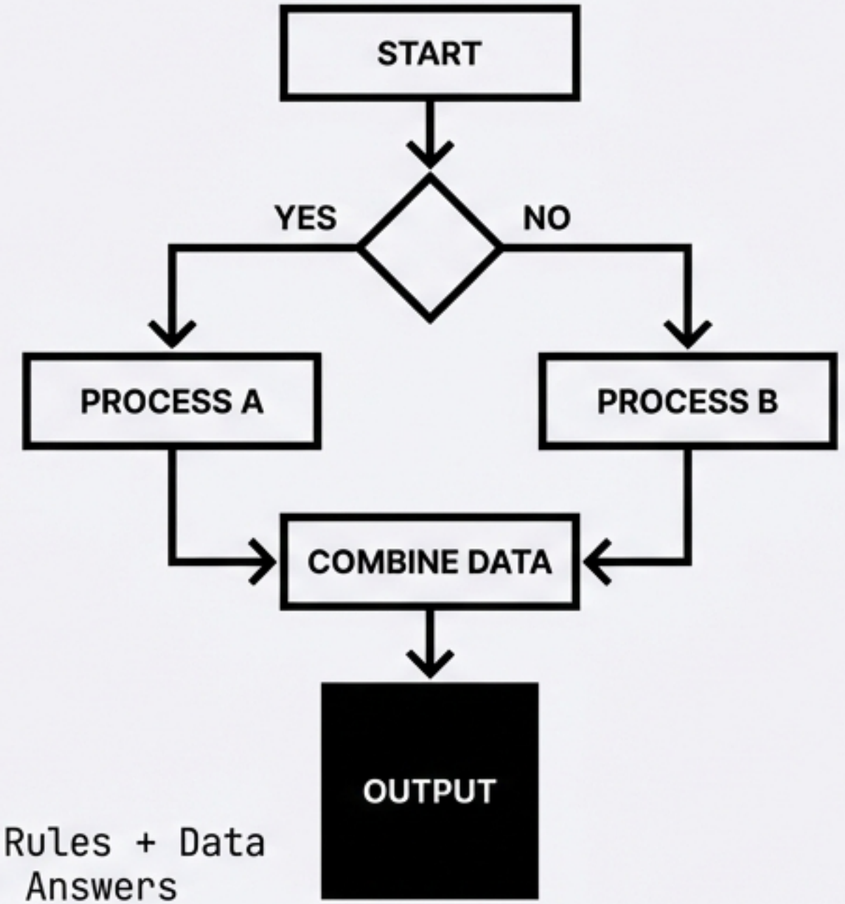
PART I: SYSTEMS FOUNDATIONS

Data, Algorithms, and Infrastructure



Machine Learning Systems are not just models; they are integrated computing systems where hardware constraints shape algorithmic possibilities.

TRADITIONAL SOFTWARE
Deterministic / Logic-Based



Input: Rules + Data
Output: Answers

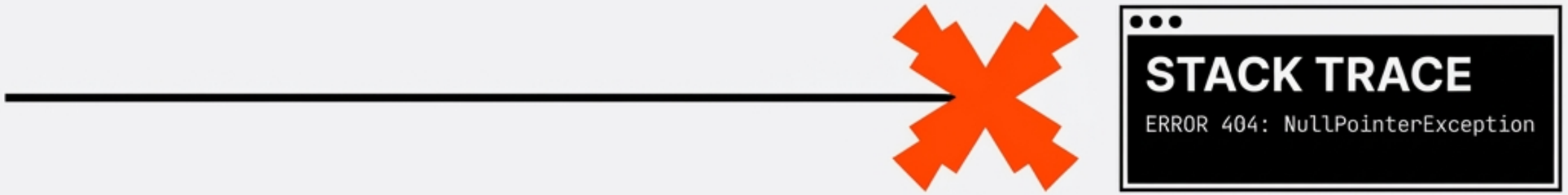
MACHINE LEARNING
Probabilistic / Data-Driven

Input: Answers + Data
Output: Rules

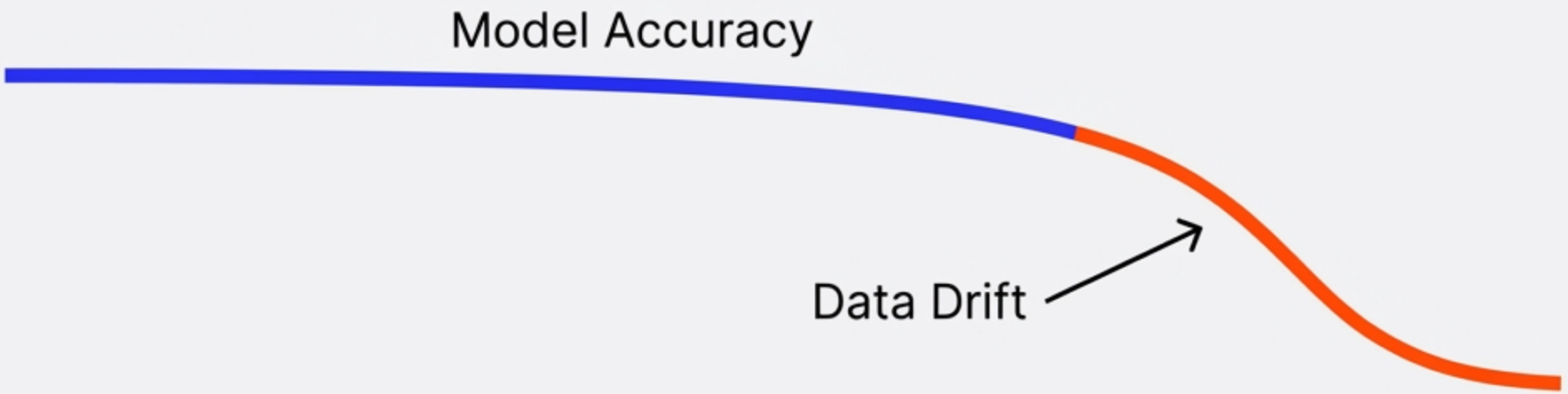
The Paradigm Shift: From writing rules to learning patterns.

THE SILENT FAILURE MODE

TRADITIONAL CRASH

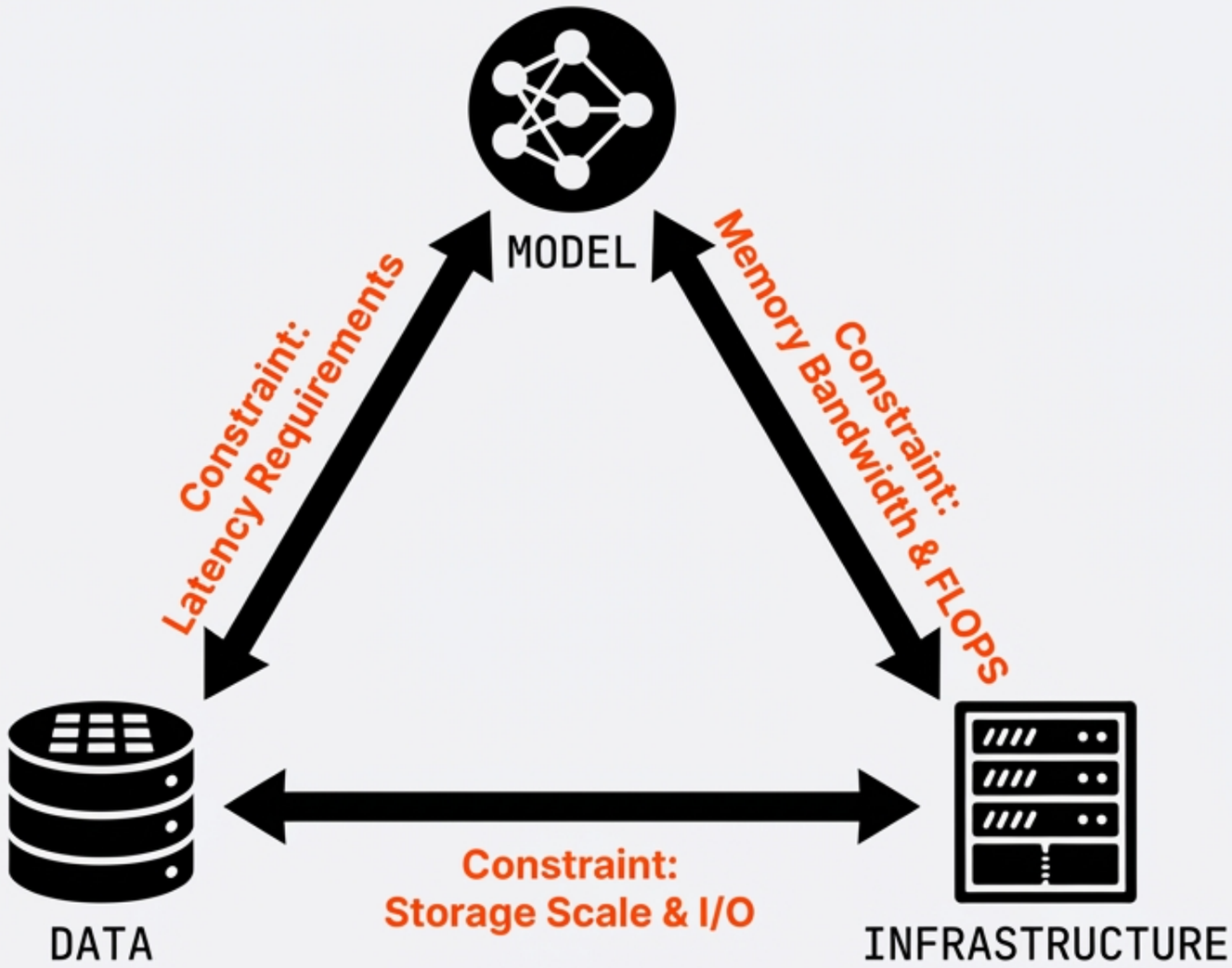


ML MODEL DRIFT



Traditional software crashes loudly. ML systems degrade silently.

SYSTEM INTERDEPENDENCE

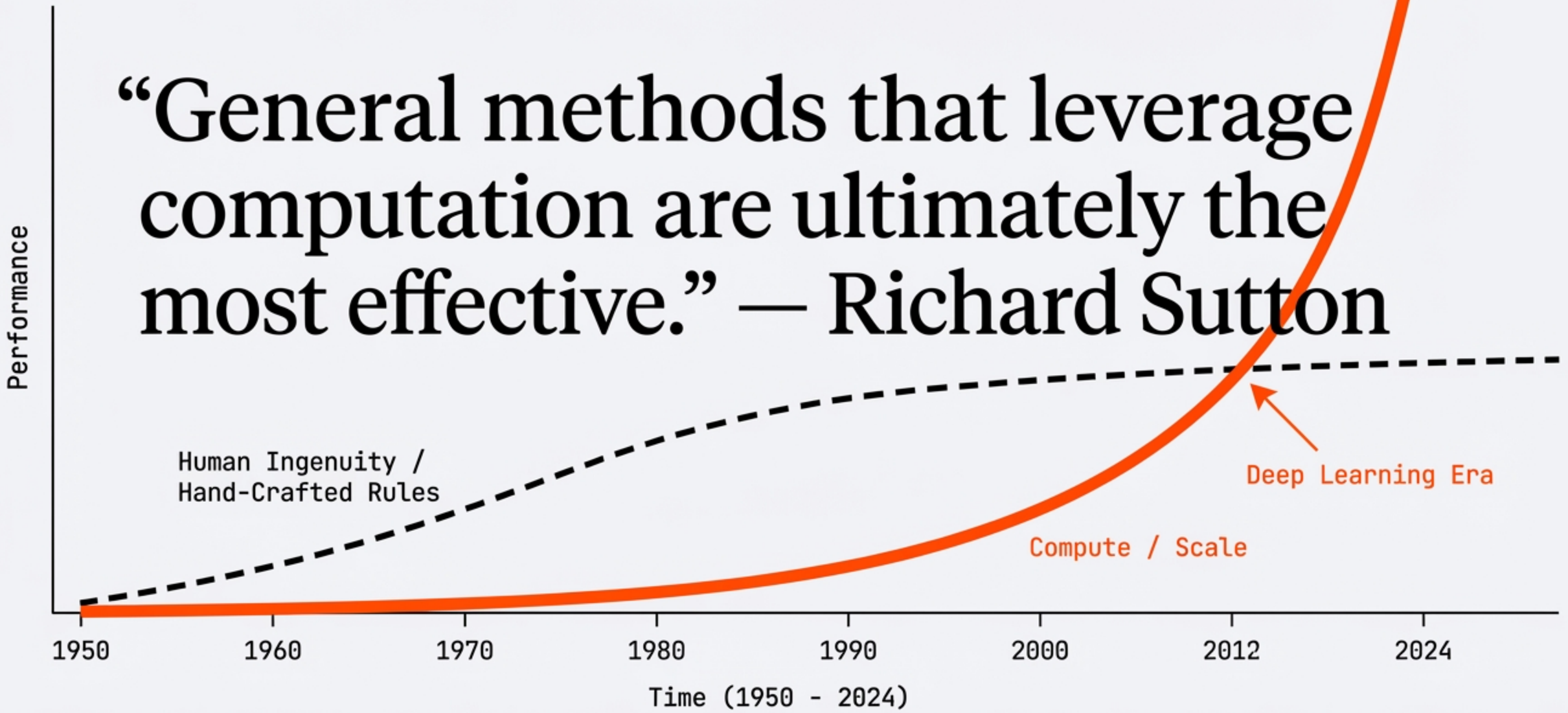


Optimizing one component in isolation leads to system-level failure.

Engineering Editorial

#F5F5F7

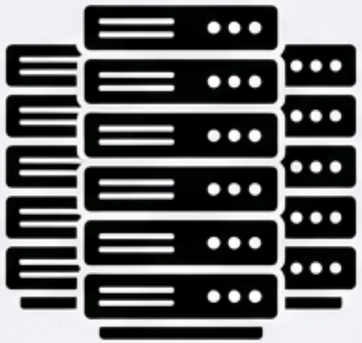
THE BITTER LESSON



Engineering Editorial

THE DISTRIBUTED INTELLIGENCE SPECTRUM

CLOUD



Infinite Resource /
High Latency

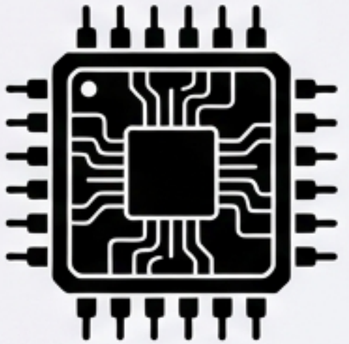
EDGE

Regional /
Moderate Latency

MOBILE

Battery Constrained

TINYML



Extreme Constraint (mW)

#2E2EFE
International Klein Blue

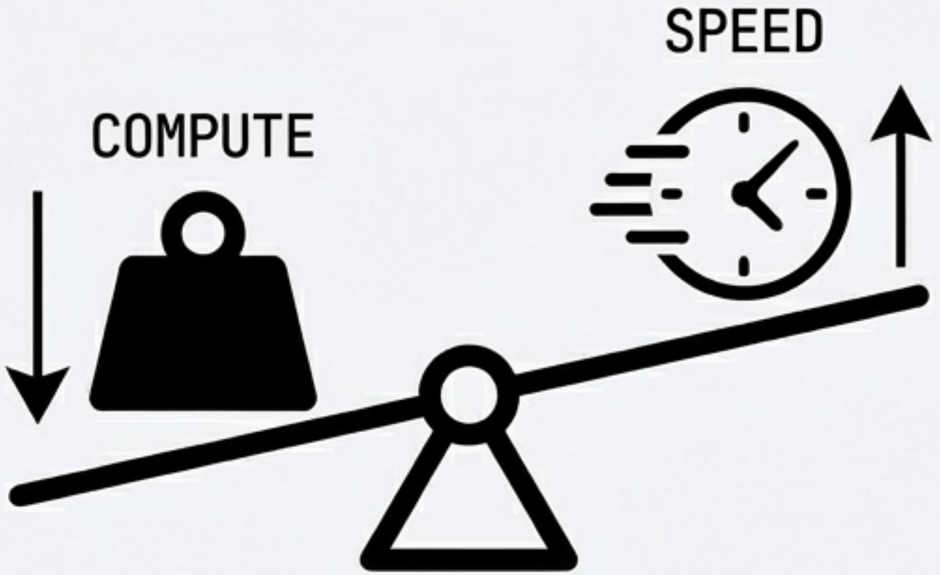
#FF4500
Safety Orange

Engineering Editorial

CLOUD vs. EDGE TRADE-OFF

CLOUD ML

- **Compute:** Unlimited (Petabytes)
- **Scaling:** Elastic
- **Latency:** 100-500ms (Network)
- **Privacy:** Data leaves premise



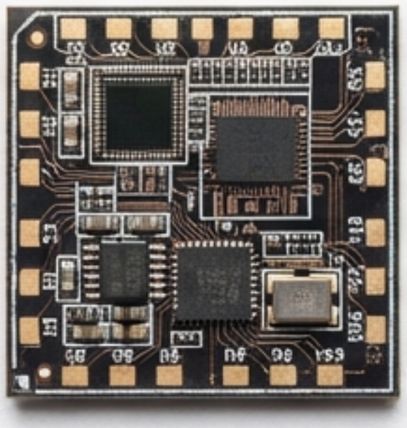
EDGE ML

- **Compute:** Strictly Limited (mW)
- **Scaling:** Hardware dependent
- **Latency:** <10ms (Real-time)
- **Privacy:** Data stays local

Physics dictates the trade-off: Massive compute or instant response.

THE FRONTIER: TINYML

Engineering Editorial



POWER:
< 1 mW

MEMORY:
256 KB RAM

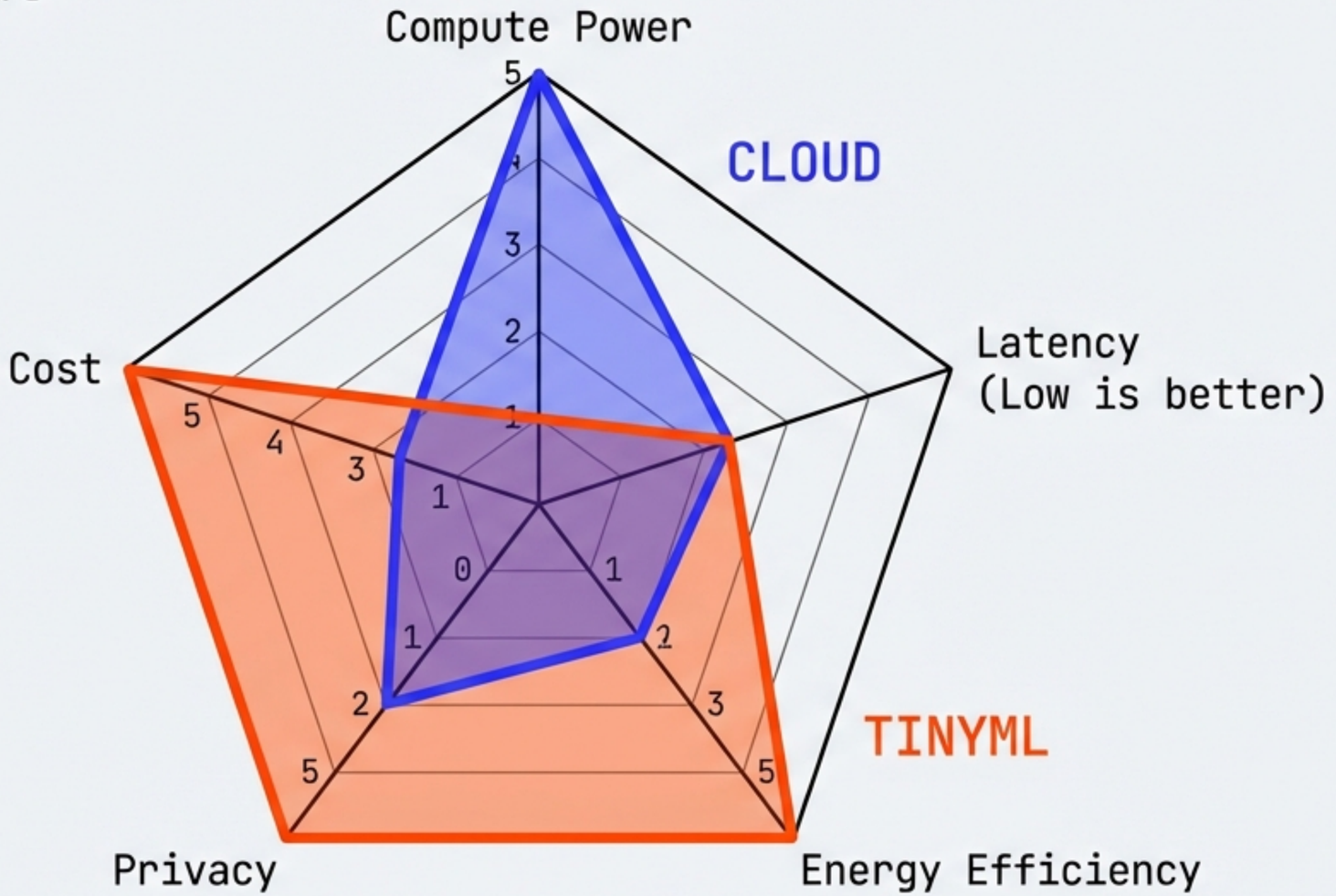
50,000x less than Cloud

COST:
< $5.00

Implementing intelligence under the strictest constraints of physics.

THE RADAR CHART OF TRADE-OFFS

Engineering Editorial



No 'best' system. Only the right trade-offs.

EVOLUTION OF PARADIGMS

Engineering Editorial

SYMBOLIC AI (1980s)

The Logic Trap

```
IF (condition)
THEN (action)
```

STATISTICAL LEARNING (2000s)

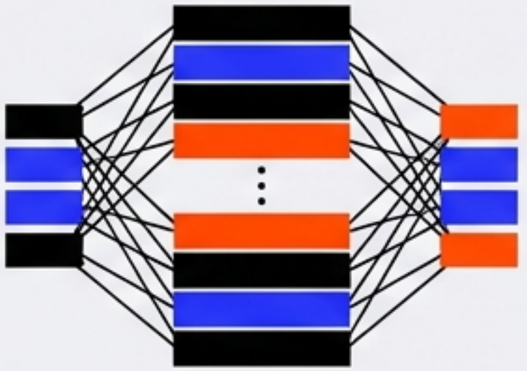
Feature Engineering

	Hand-Crafted Features
	Feature A: 0.5
	Feature B: 1.2
...	Feature C: -0.8

Automated Feature Extraction

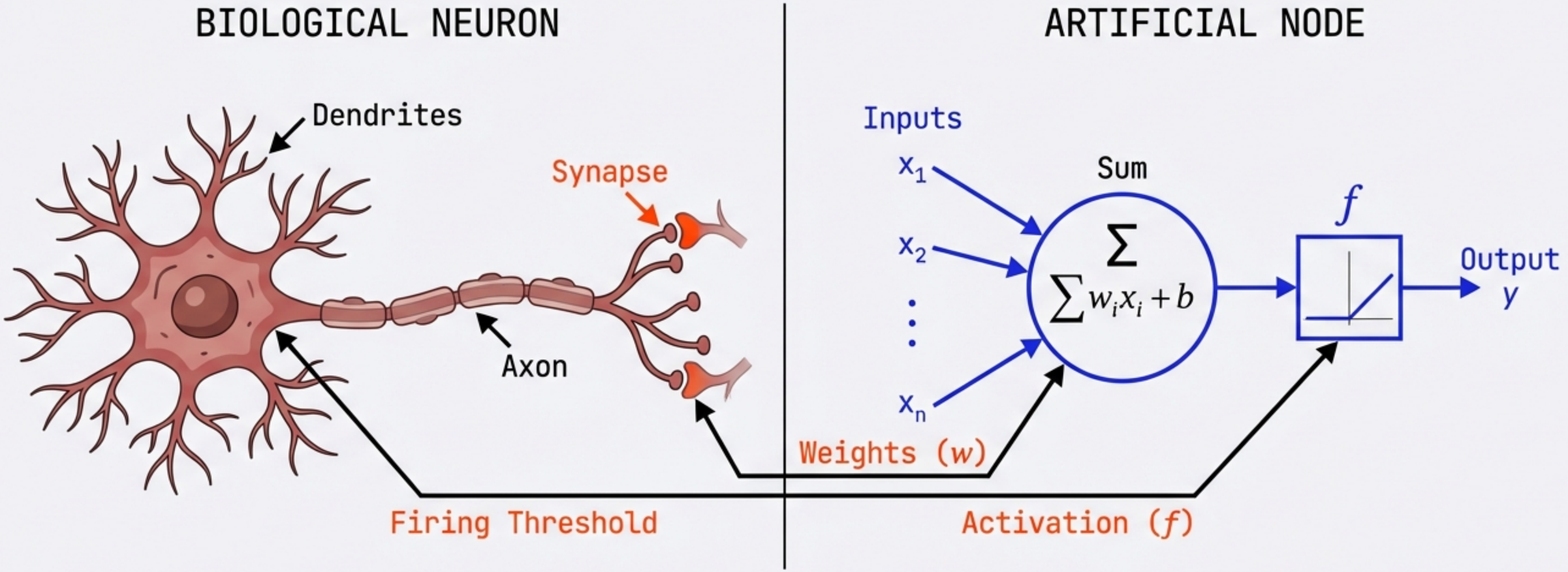
DEEP LEARNING (2010s)

Representation Learning



Performance now scales with Data Volume.

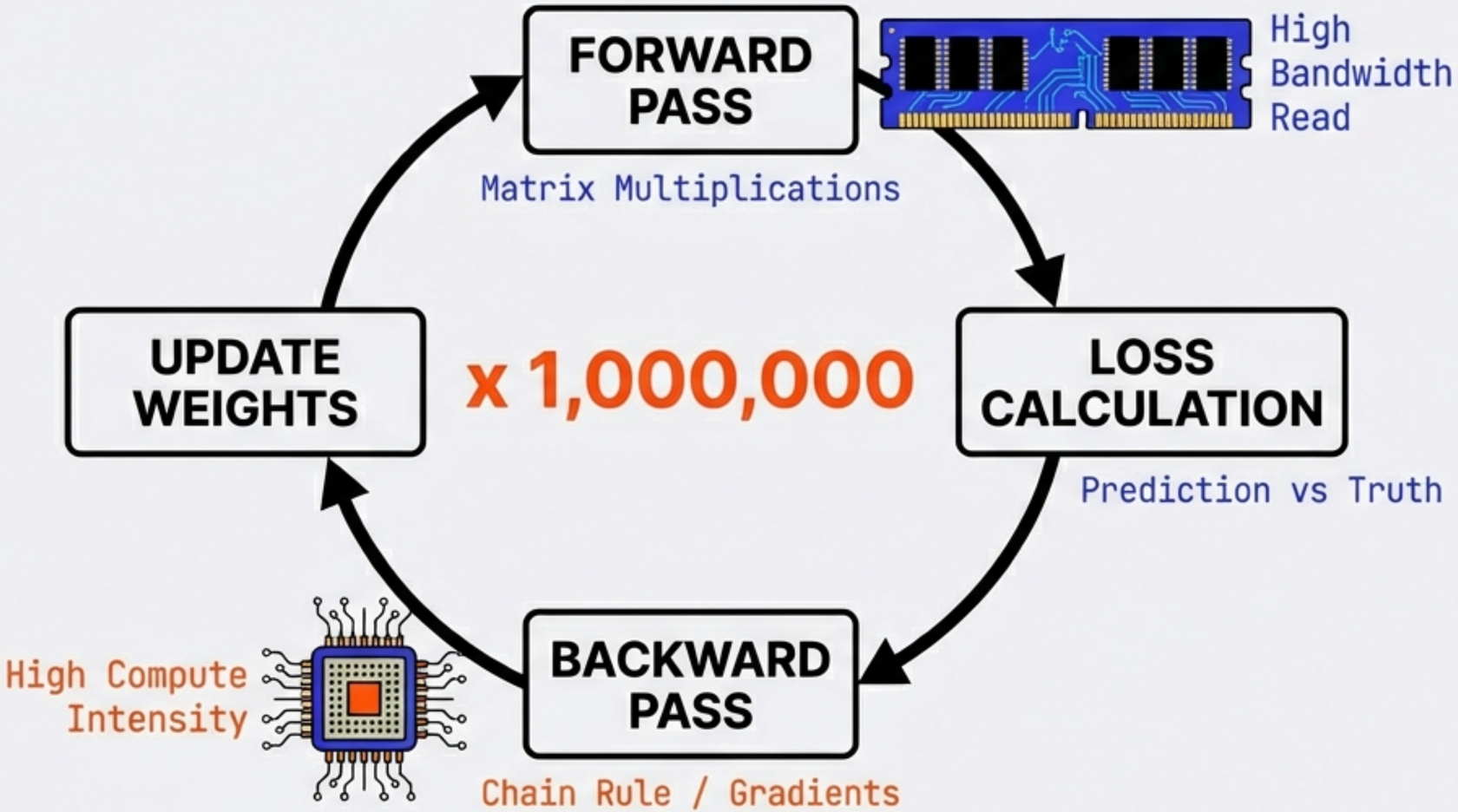
FROM BIOLOGY TO SILICON



We abstract biology into Matrix Multiplication to enable massive parallelism.

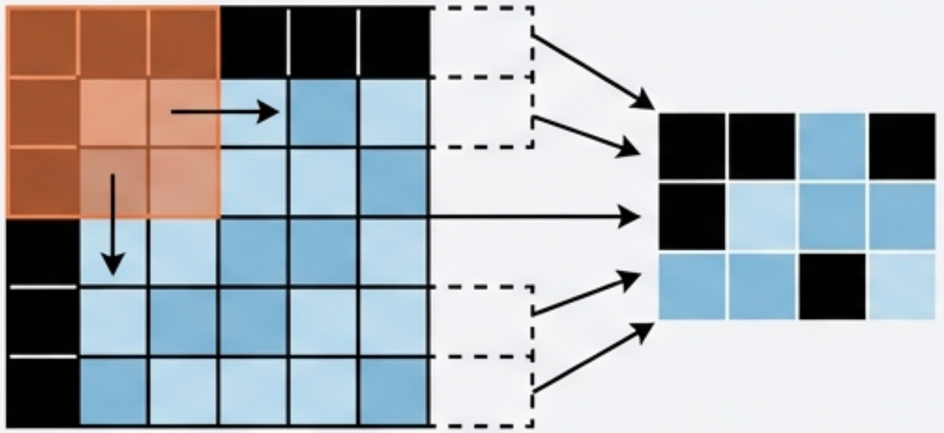
THE MATHEMATICAL WORKLOAD

Engineering Editorial



The Systems challenge is keeping the compute units fed with data.

PROCESSING SPACE: CNNs

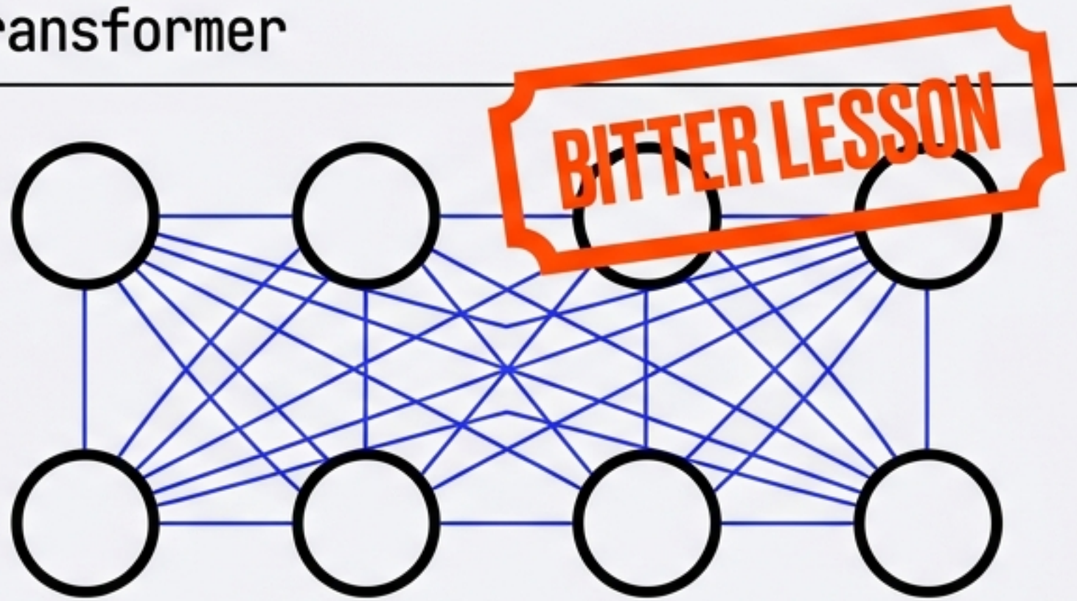


```
import torch.nn as nn

class SimpleNet(nn.Module):
    def __init__(self):
        super().__init__()
        self.conv = nn.Conv2d(3, 64, kernel_size=3)
        self.fc = nn.Linear(64, 10)

    def forward(self, x):
        x = self.conv(x)
        x = torch.relu(x)
        x = self.fc(x)
        return x
```

- Inductive Bias: Spatial Locality
- Hardware Benefit: Parameter Sharing
- Systems Result: High Memory Efficiency

Engineering Editorial

PROCESSING SEQUENCE: TRANSFORMERS

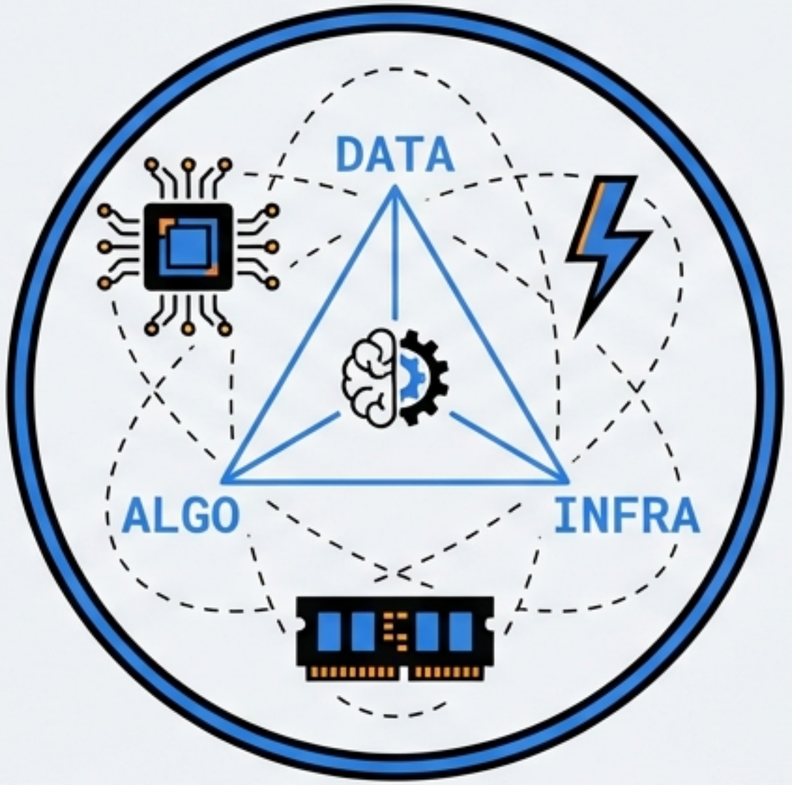
RNN	Transformer
	
Sequential Bottleneck	Parallel Attention

Constraint: Quadratic Complexity (N^2).

Hardware Requirement: Massive High Bandwidth Memory (HBM).

Engineering Editorial

THE AI ENGINEER



AI Engineering is the systems-level integration of algorithms, data, and infrastructure.

"If you want to go fast, go alone. If you want to go far, go together."