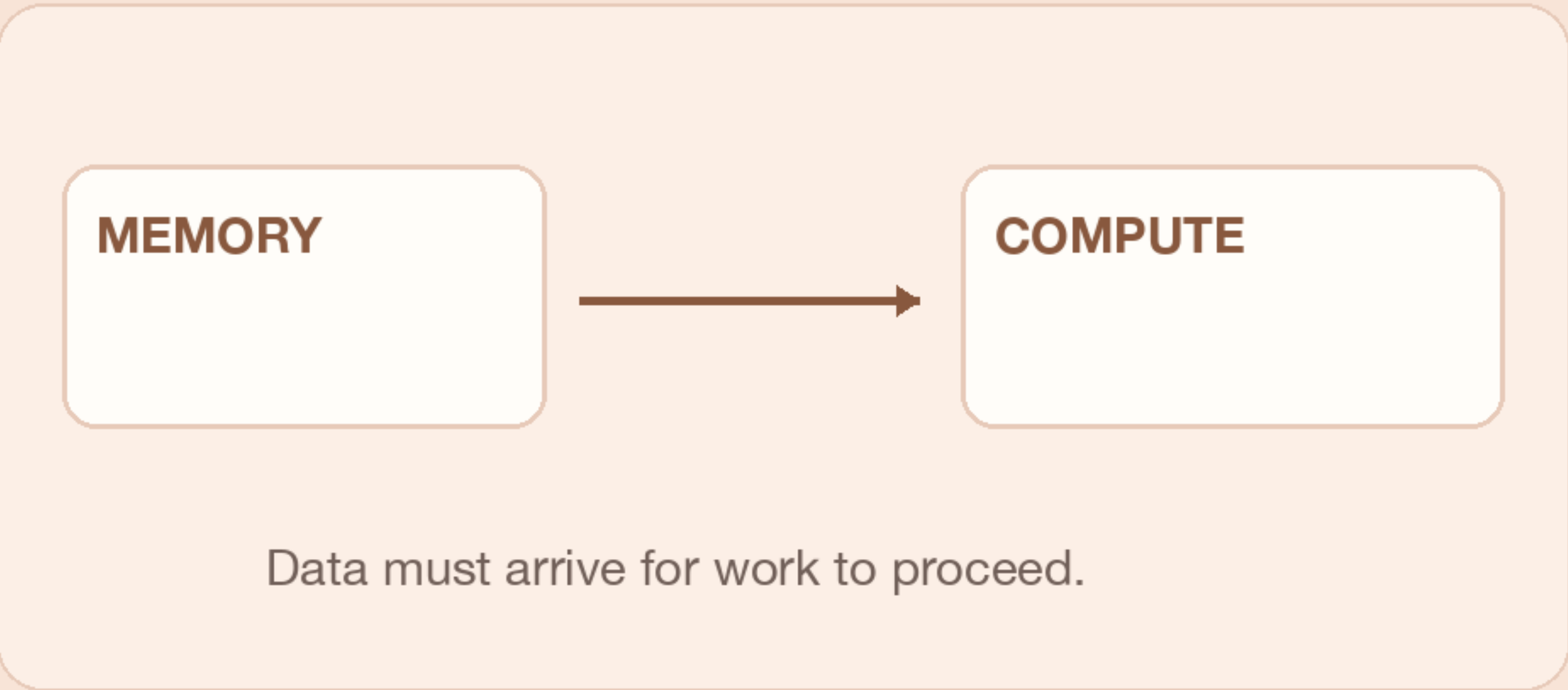


AI needs data.

Not just arithmetic.

Microsoft’s 12 October Maia announcement highlights data movement in AI infrastructure. Here is the idea, without repeating its performance forecasts.



Original conceptual diagrams. No measured product/site result.

Training learns. Inference uses.

Training adjusts a model. Inference uses a trained model to produce a result from an input. Today's chip story focuses on inference.

TRAIN: adjust

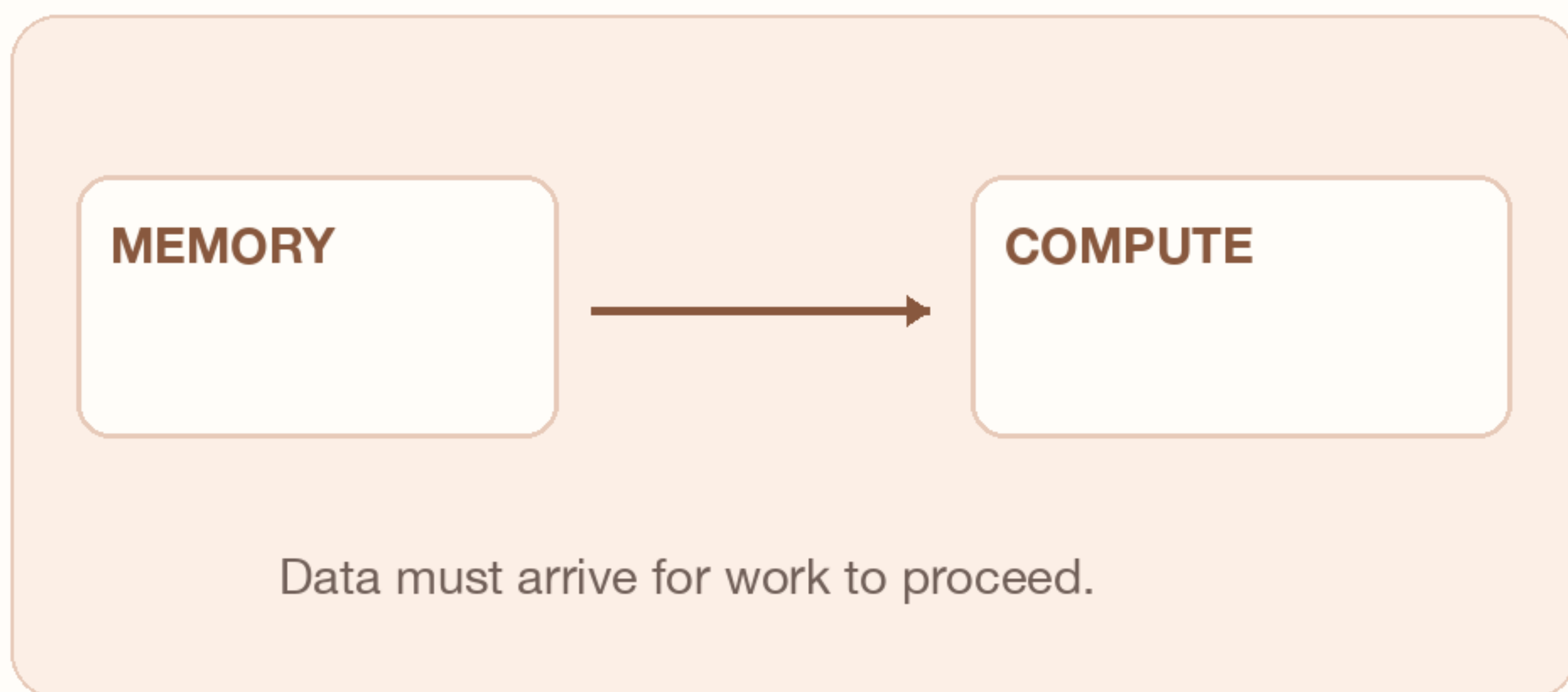
USE: inference

Two different stages of the workflow.

Original conceptual diagrams. No measured product/site result.

The chip needs the next pieces.

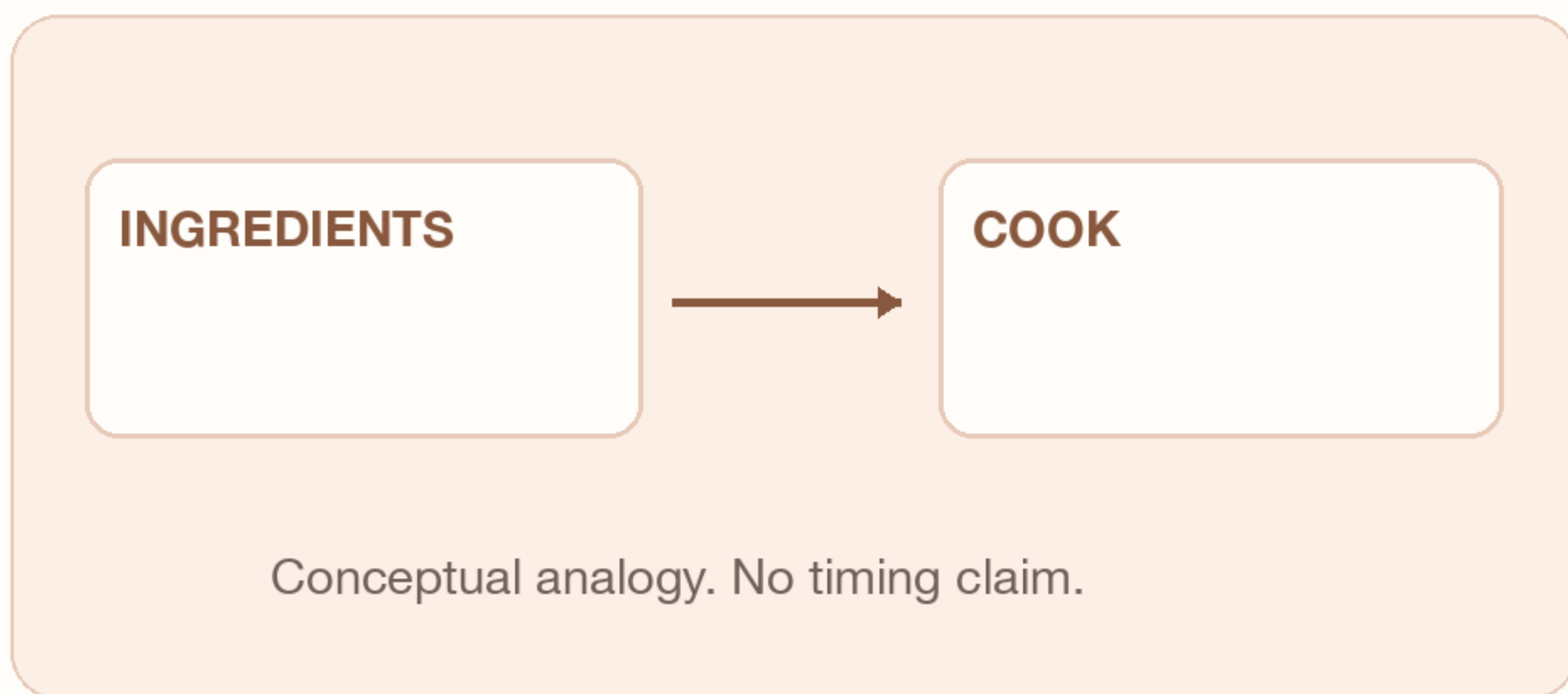
A computing unit performs operations. Model data and working information must reach it from memory. Waiting for those pieces can limit useful work.



Original conceptual diagrams. No measured product/site result.

A nearby shelf is a useful analogy.

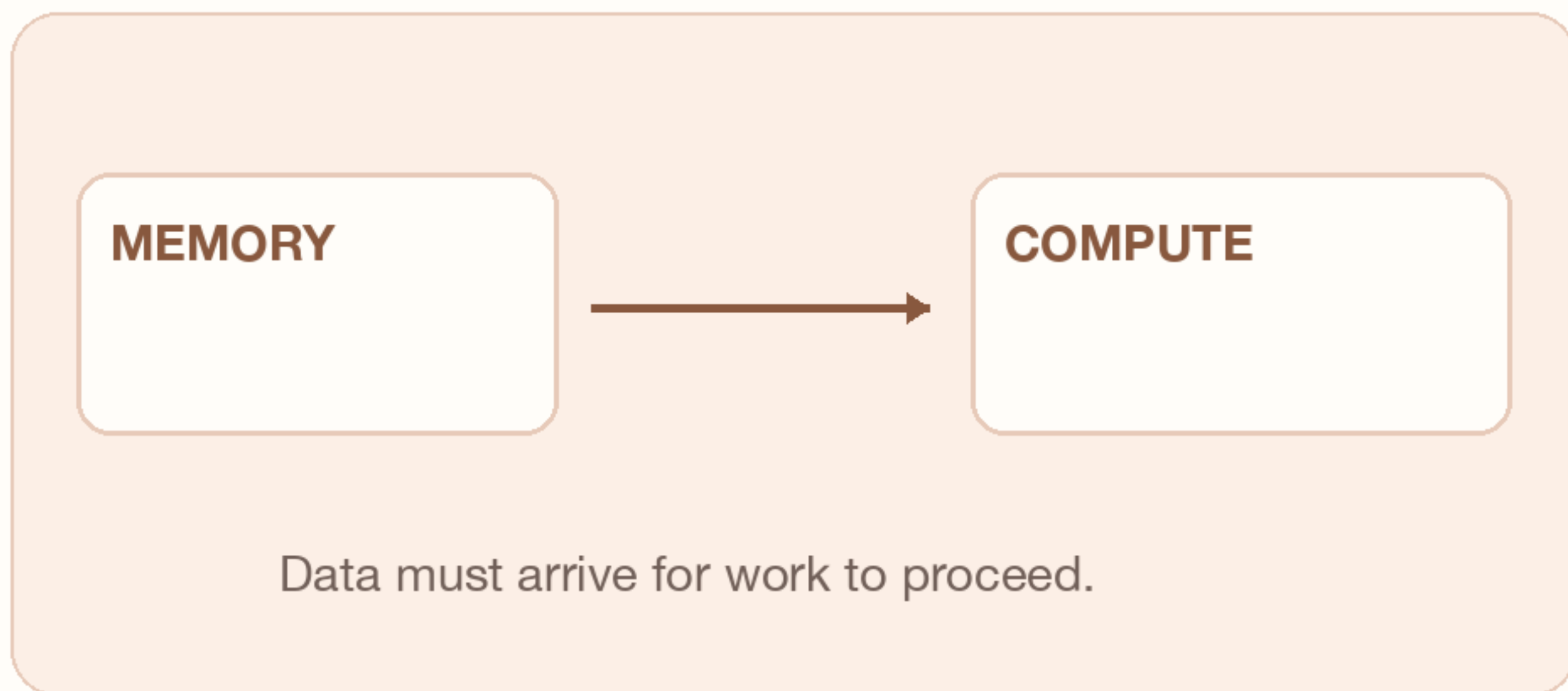
Imagine a cook and ingredients. A fast cook can still wait on deliveries. Nearby storage can reduce some trips. This is an analogy, not a chip diagram to scale.



Original conceptual diagrams. No measured product/site result.

More requests change the problem.

Serving many users also involves memory, communication and scheduling. The whole system matters; one headline chip number does not describe every workload.



Original conceptual diagrams. No measured product/site result.

Today and 2027 are separate.

Microsoft says Maia 200 serves production workloads. It plans Maia 300 production deployments in 2027. A plan is not evidence of deployment today.



Original conceptual diagrams. No measured product/site result.

Ask three workload questions.

What model and input size? How long until a useful reply? How much complete-system energy and cost? Keep the output quality requirement the same.

01 Define the workload

02 Locate the bottleneck

03 Test the full result

Original conceptual diagrams. No measured product/site result.

A faster number needs a context.

Our original framework: identify the workload, locate the bottleneck, then test the complete result. No chip test, measured saving or vendor ranking is claimed.

01 Define the workload

02 Locate the bottleneck

03 Test the full result

Original conceptual diagrams. No measured product/site result.