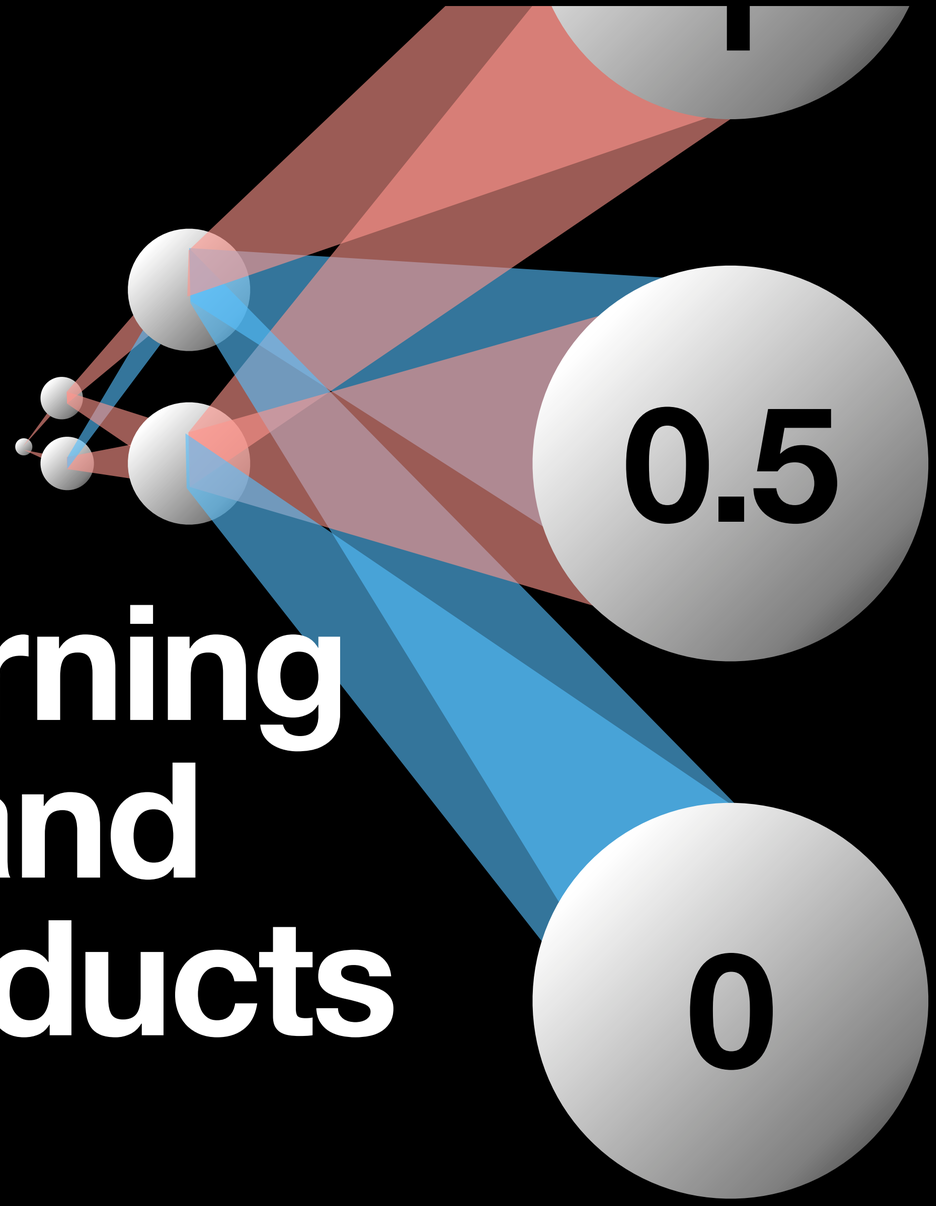
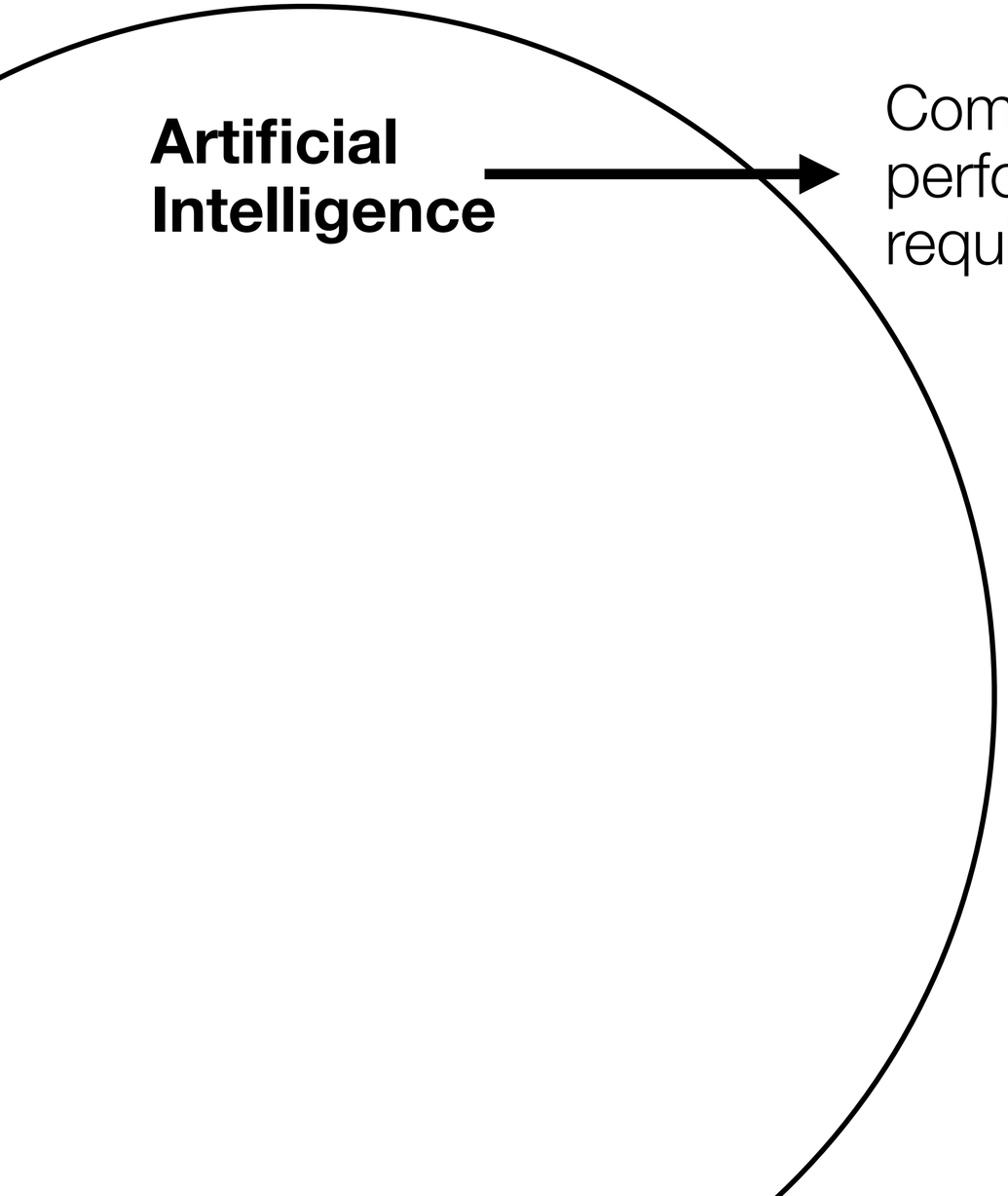


Role of Machine Learning in Dairy Ingredients and Fermented Dairy Products

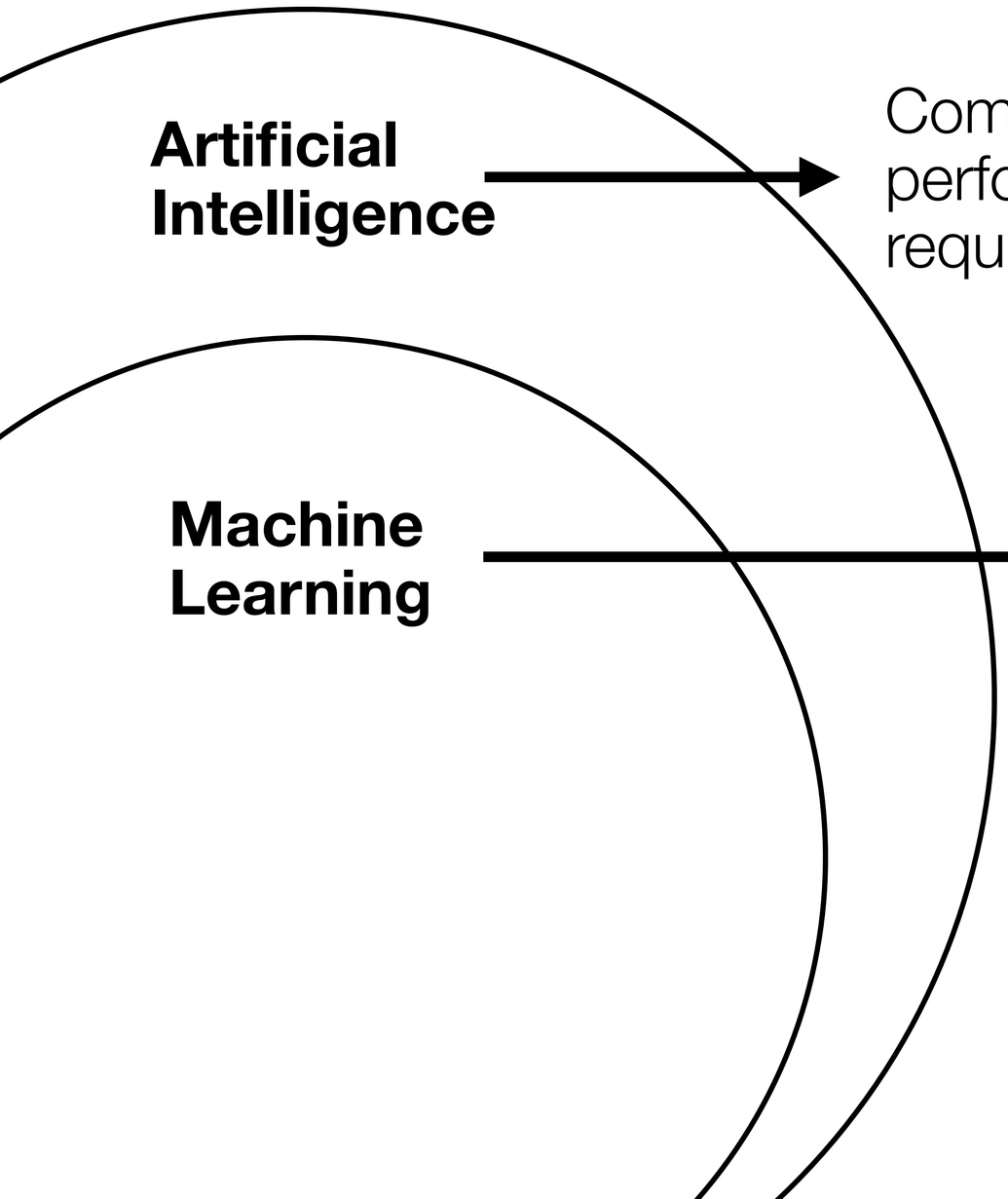
Austin Weber, PhD Candidate Food Science and Technology

Definitions



Computer systems
performing tasks that normally
require human intelligence

Definitions



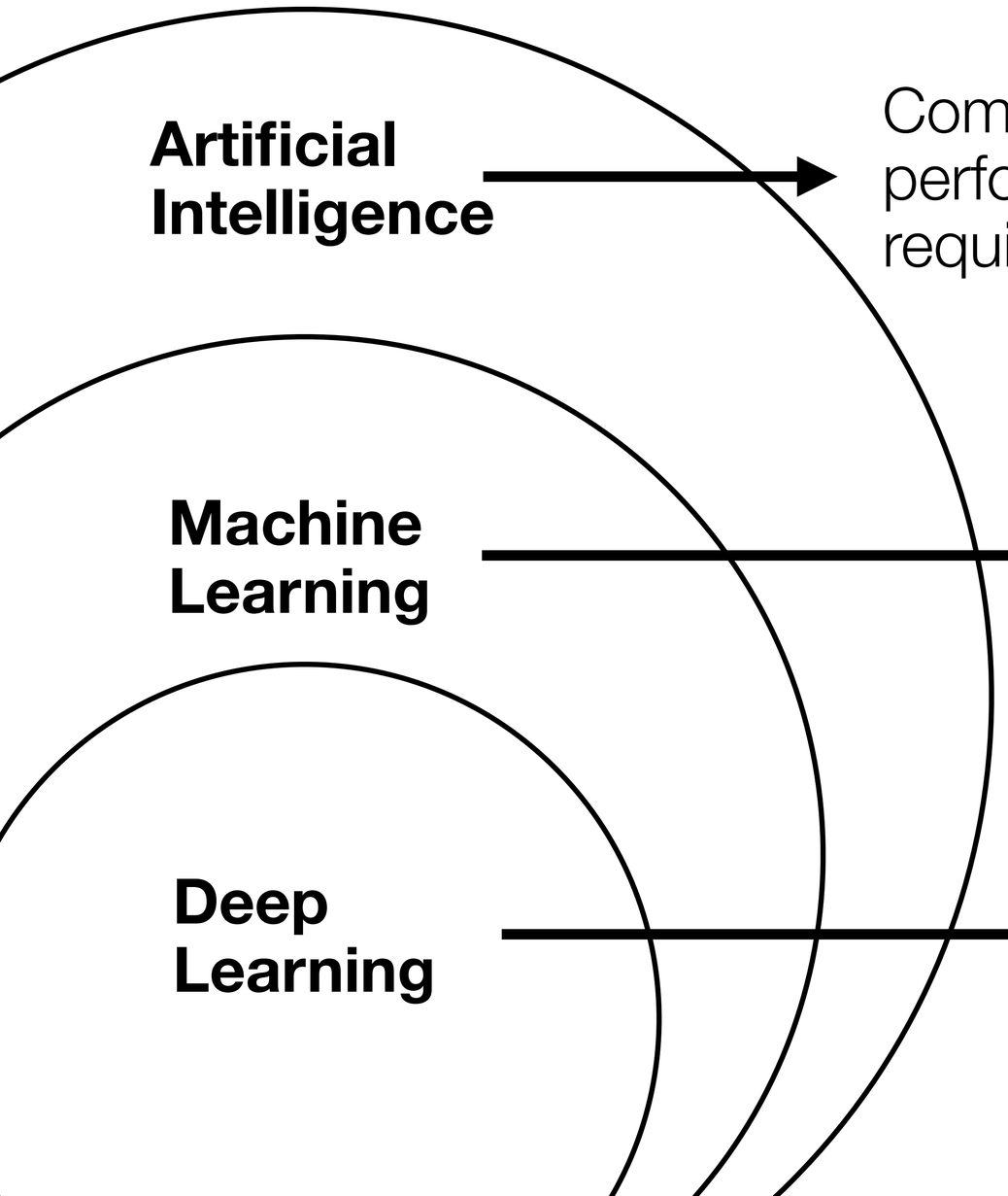
Artificial Intelligence

Computer systems performing tasks that normally require human intelligence

Machine Learning

Models learning patterns from data, rather than following predefined rules

Definitions



Artificial Intelligence

Computer systems performing tasks that normally require human intelligence

Machine Learning

Models learning patterns from data, rather than following predefined rules

Deep Learning

A type of machine learning **model** that can discover complex patterns from large amounts of raw data

Not included:

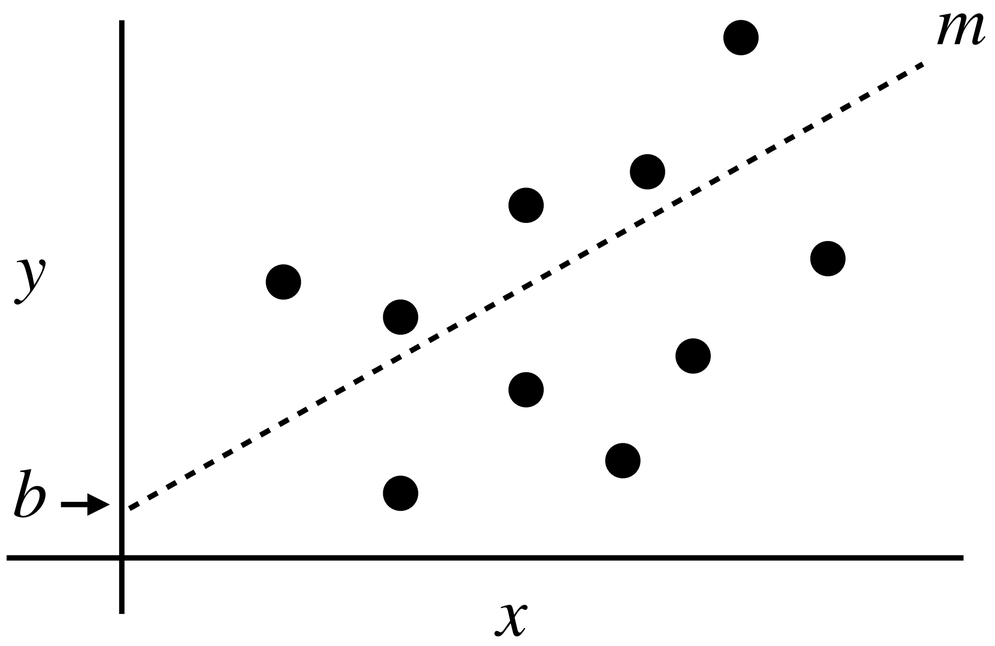
LLMs

Agentic AI

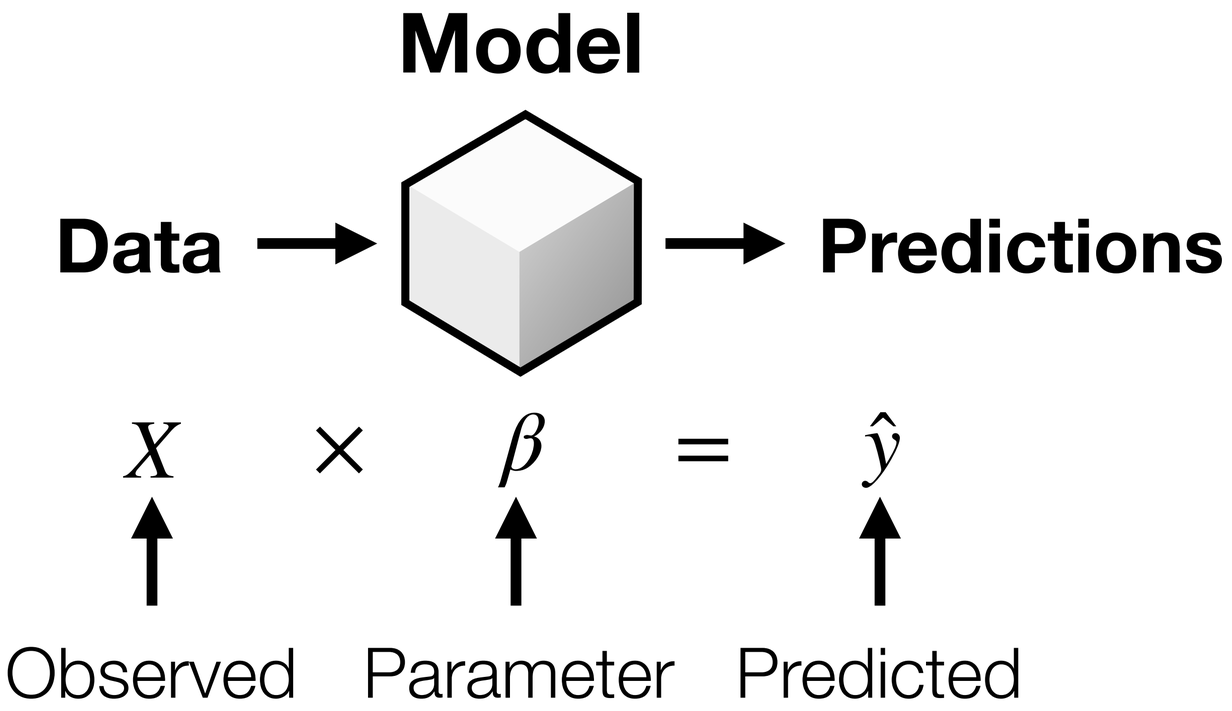
When we think of a **model**, think of calculating the slope of a line:

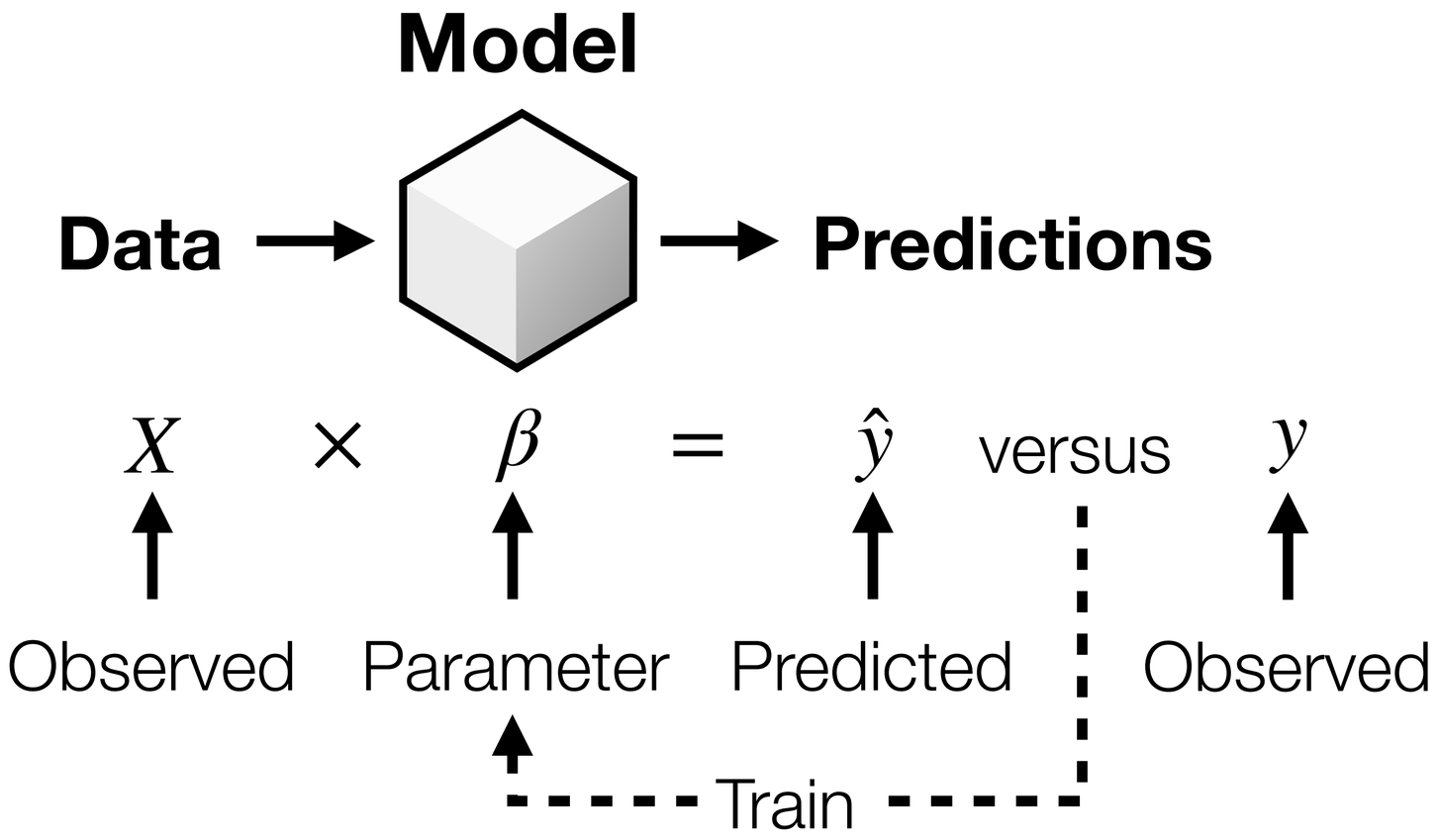
$$y = mx + b$$

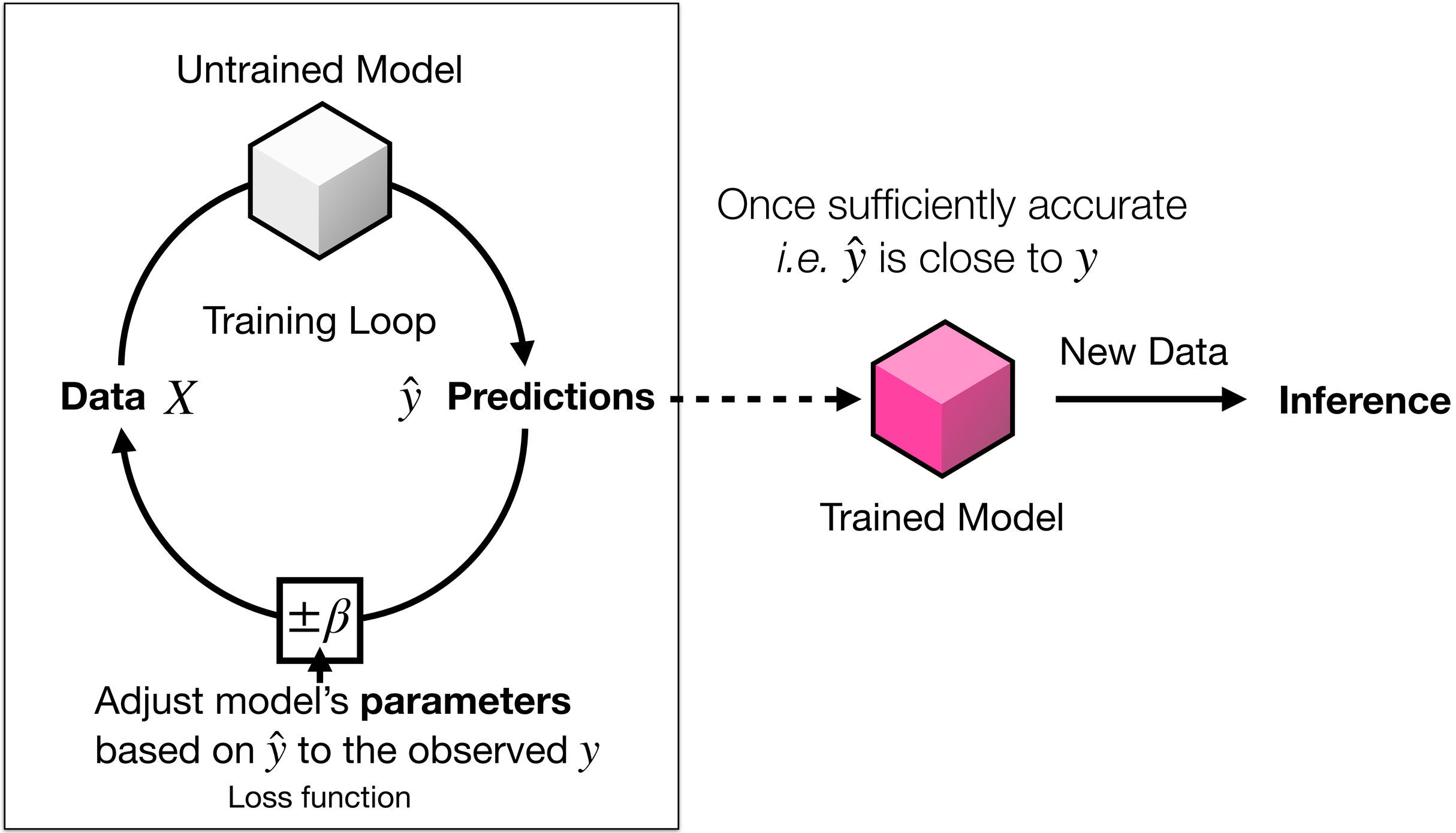
Given the observed values of x and y, we calculate m and b:



This is essentially what a model *does*, give observed data to *train* parameters


Model
Data
Predictions
$X \times \beta = \hat{y}$
Observed
Parameter
Predicted


Model
Data
Predictions
X
$\times$
β
$=$
$\hat{y}$
versus
y
Observed
Parameter
Predicted
Observed
Train


Untrained Model
Training Loop
Data X
$\hat{y}$ Predictions
$\pm\beta$
Adjust model's **parameters**
based on $\hat{y}$ to the observed y
Loss function
Once sufficiently accurate
i.e. $\hat{y}$ is close to y
Trained Model
New Data
Inference

Dairy AI applications

Integrative Modeling: Labeled data + machine learning models

Labeled = human identified (*e.g.* minutes since pasteurization started)

Feature Discovery: Unlabeled / vast data + deep learning models

Unlabeled = unidentified (*e.g.* a large database of product ingredient concentrations, without annotations describing product type, quality, texture, preference, etc)

Integrative Modeling

Integrative Modeling

	Type / Input	Prediction
1	**Sensor & Multimodal processes**: Temperature, pH, time, moisture, pressure	Processing state, composition, quality

Integrative Modeling

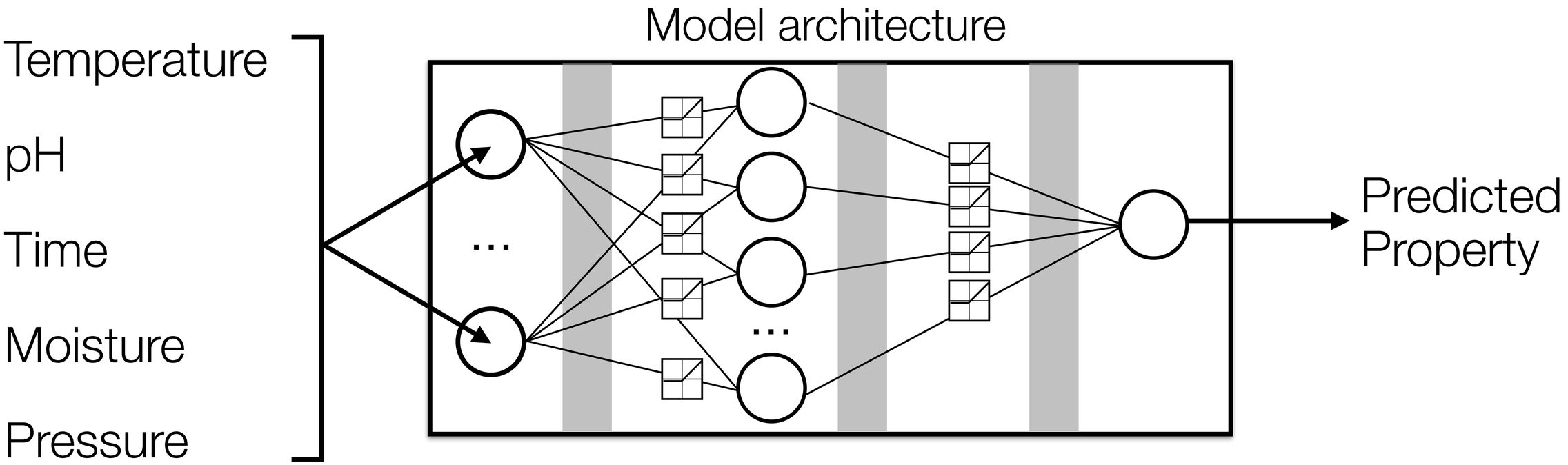
	Type / Input	Prediction
1	**Sensor & Multimodal processes**: Temperature, pH, time, moisture, pressure	Processing state, composition, quality
2	**Computer vision**: Imaging as a signal	Processing state, composition, quality, defects, spoilage

Integrative Modeling

	Type / Input	Prediction
1	**Sensor & Multimodal processes**: Temperature, pH, time, moisture, pressure	Processing state, composition, quality
2	**Computer vision**: Imaging as a signal	Processing state, composition, quality, defects, spoilage
3	**Formulation**: Predict new product designs	Final product structure, product stability and sensory properties

1 Multimodal Processes

Model is trained on labeled physical data, chemical data, or sensor data, to predict a functional property of food or processing

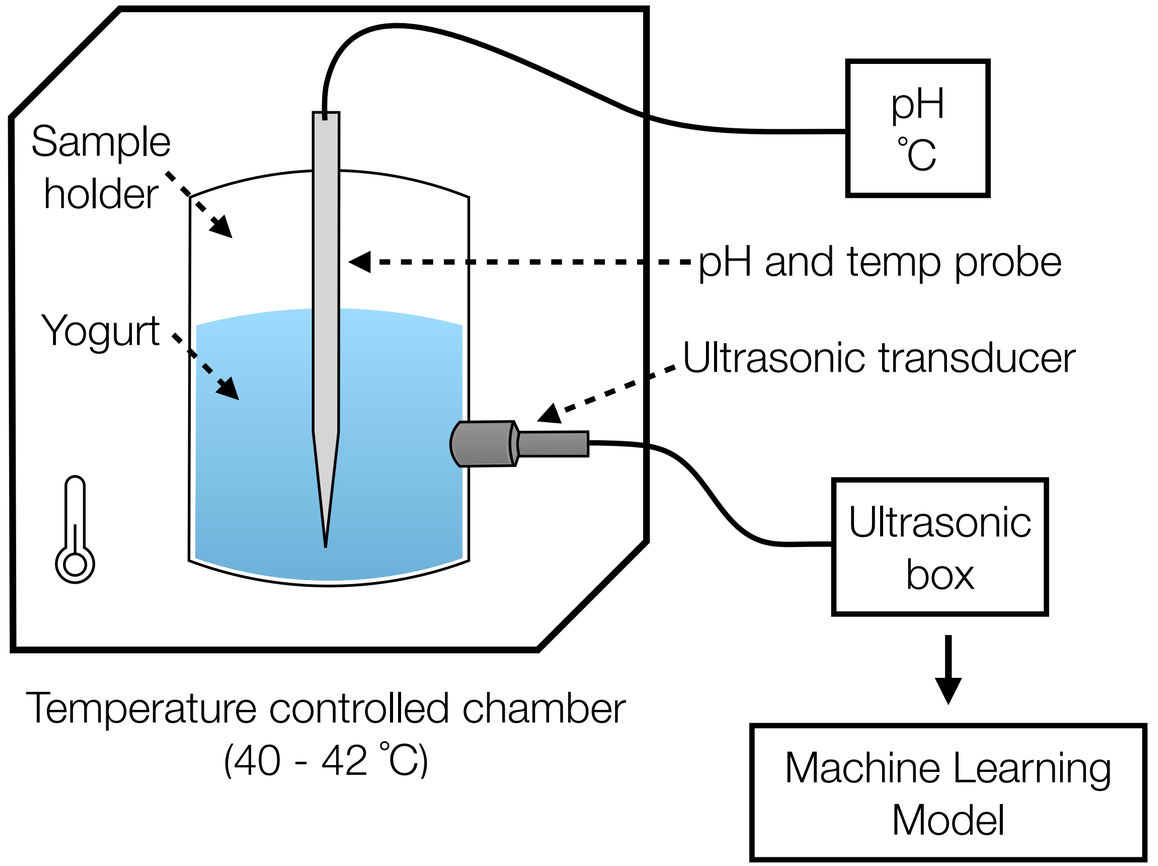


1 Multimodal Processes

Method:
Non-invasive yogurt fermentation monitoring. Integrating sensor and measurement data to predict realtime fermentation state

Goal: Monitoring enables fast corrective action, identify spoiled batches, and reduce energy and QA time loss

Result: Highly accurate predictions



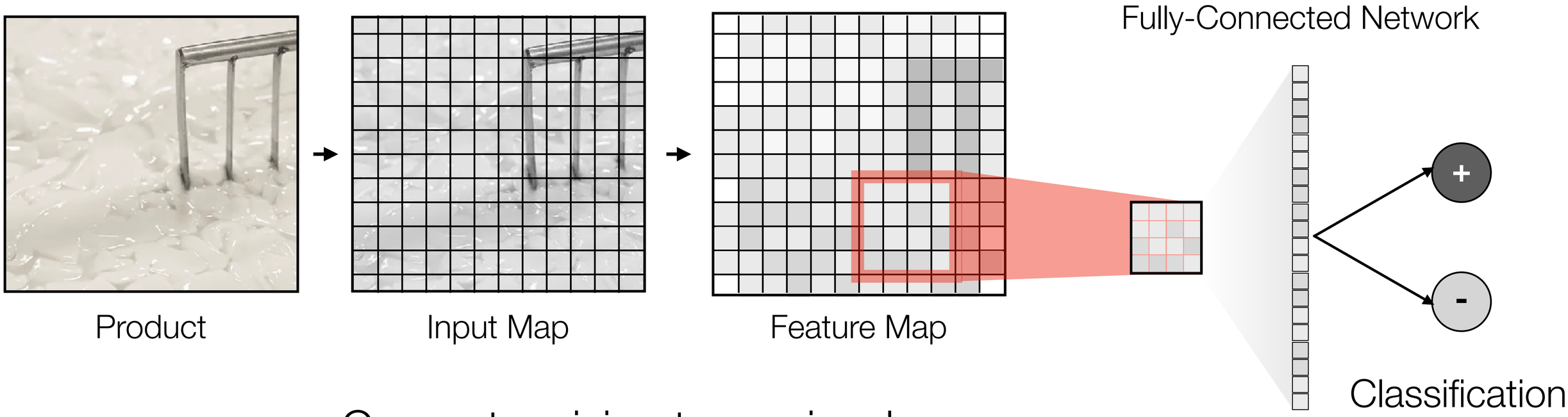
Bowler et al., 2023

❷ Computer Vision

Bins image into ‘feature maps’

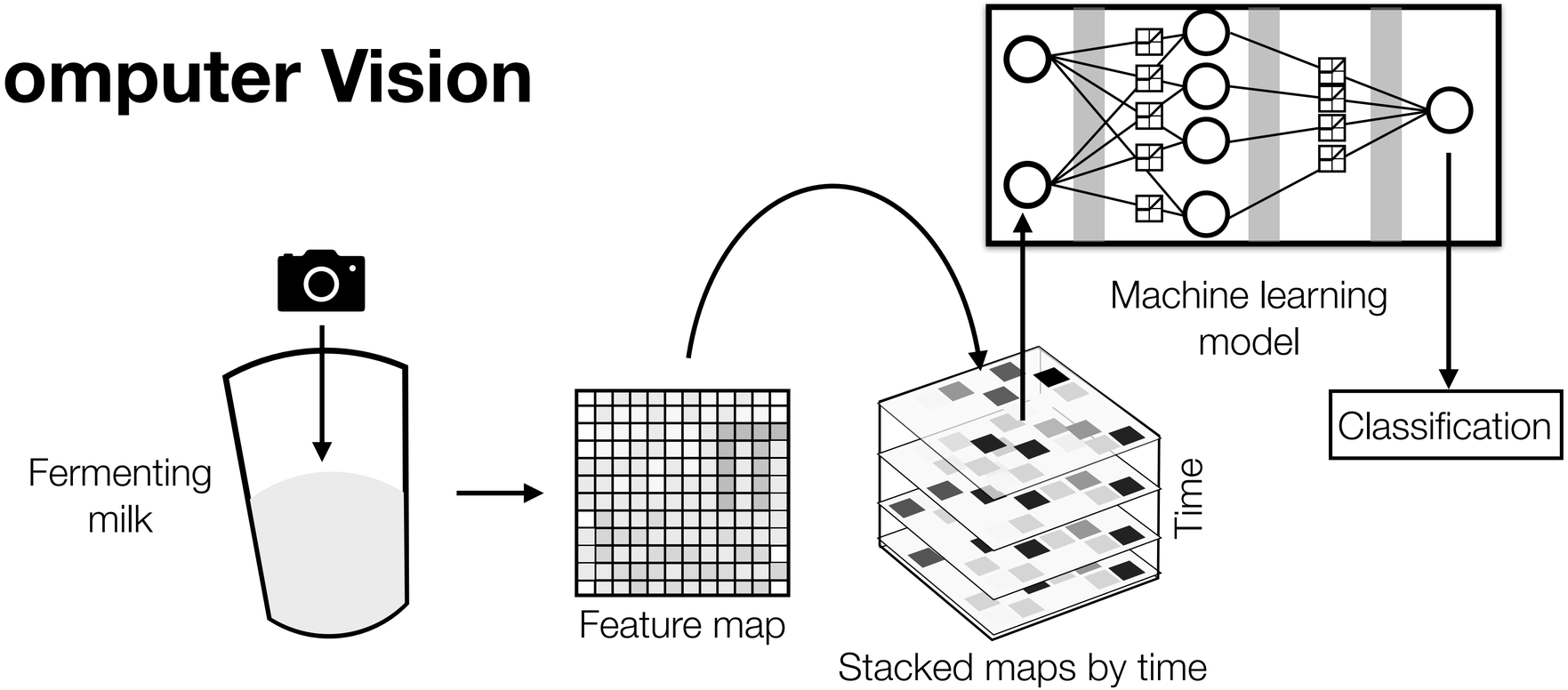
Feature hue/shade encoded as a scalar value (machine readable)

Model is then trained to predict a product classification



Computer vision turns visual process or product states into measurable variables

2 Computer Vision



Rationale: Curd cutting timing influences quality of resulting cheese, but coagulation timing is variable.

Method: Use machine learning to classify images of fermenting milk by coagulation state

Goal: Non-invasively identify gelation in realtime, achieve optimal cutting times

Result: Variable predictive accuracy, dependent on feature mapping

Loddo et al., 2025

3 Formulation

Alisia Heath
NotCo VP of R&D

"Giuseppe, [our] proprietary platform…enables optimization of ingredient choices and significantly reduces trial-and-error…the AI identifies which combinations of plants can replicate specific animal-based flavors and textures"

NotCo product line

source: https://dacodes.com/blog/how-notco-is-using-ai-to-lead-the-food-revolution

Integrative Modeling Limitations

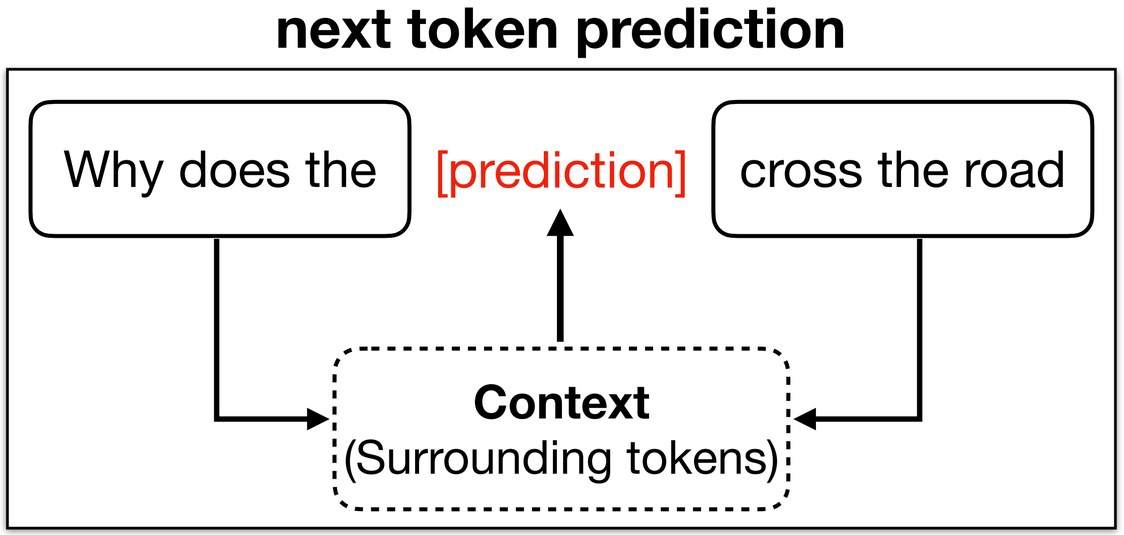
1	Data requirements	Huge structured, labeled, datasets
2	Model training requirements	Difficulty for model to generalize to new conditions / data
3	Implementation costs	Sensors, data infrastructure, calibration, integration with systems, computing resources, specialized personnel,…
4	Modeling sensory outcomes	Sensory perceptions are subjective, variable, difficult to measure

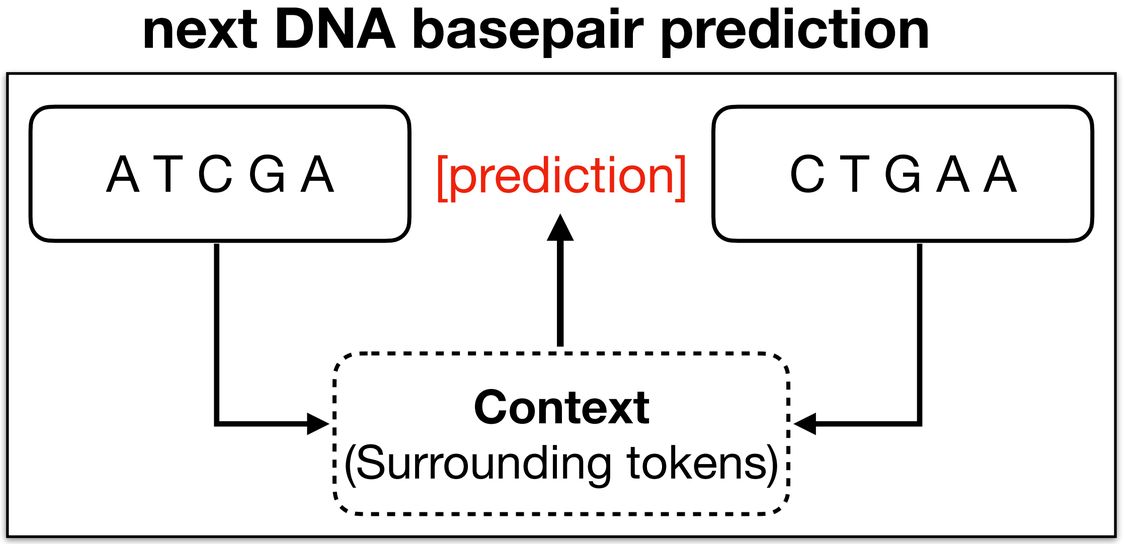
Feature Discovery

Genetic Feature Discovery

Use of LLM-like ‘next token’ prediction architecture on DNA sequences

DNA sequences are information-rich, like our language





Kaplan *et al.*, 2020

Genomic language models

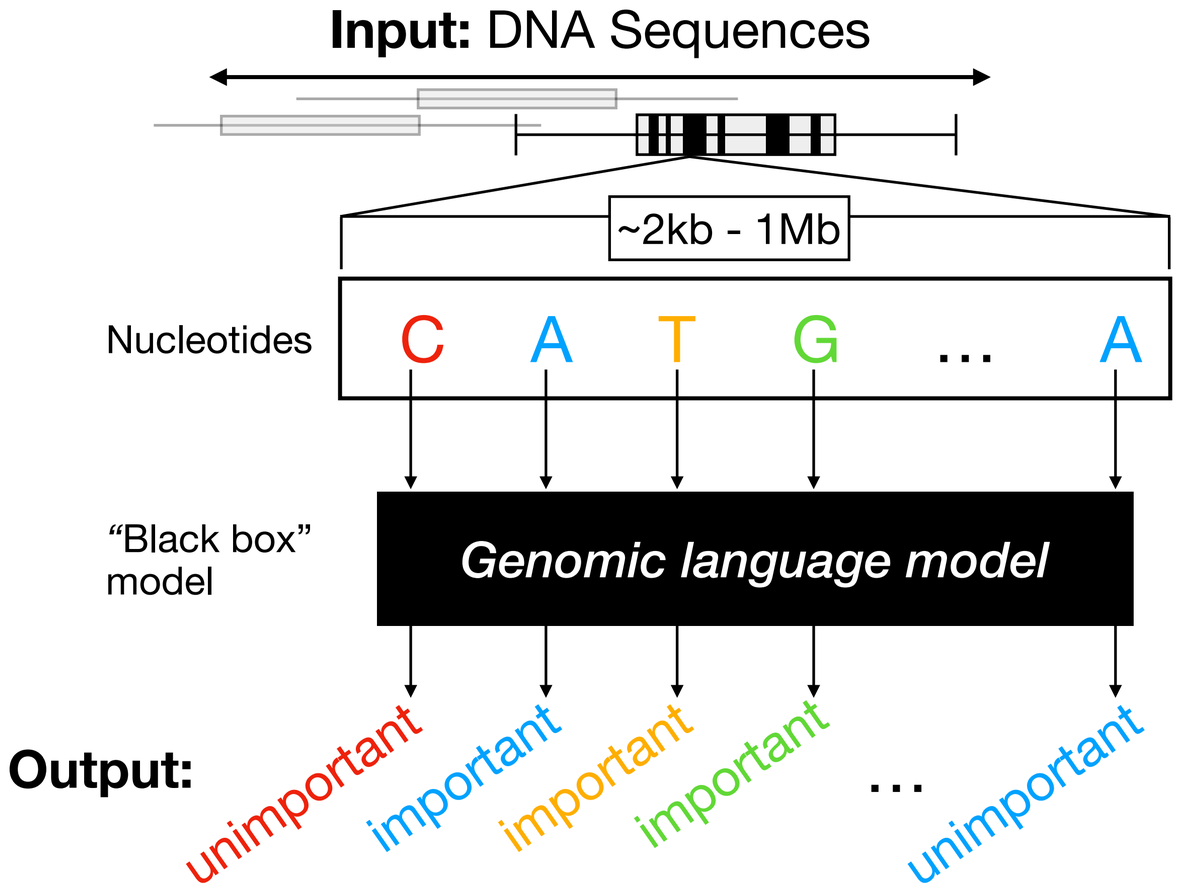
How important is each nucleotide, based on its neighbors?

ATG → Start codon, very important

CACACACA → Repeat, unimportant

Train model to learn the ‘language’ and ‘rules’ of DNA (dependencies)

Model predicts the ‘meaning’ of DNA sequences



Benegas *et al.*, 2025a

Evo 2 Arc Institute x NVIDIA

Just published

nature

Article | Open access | Published: 04 March 2026

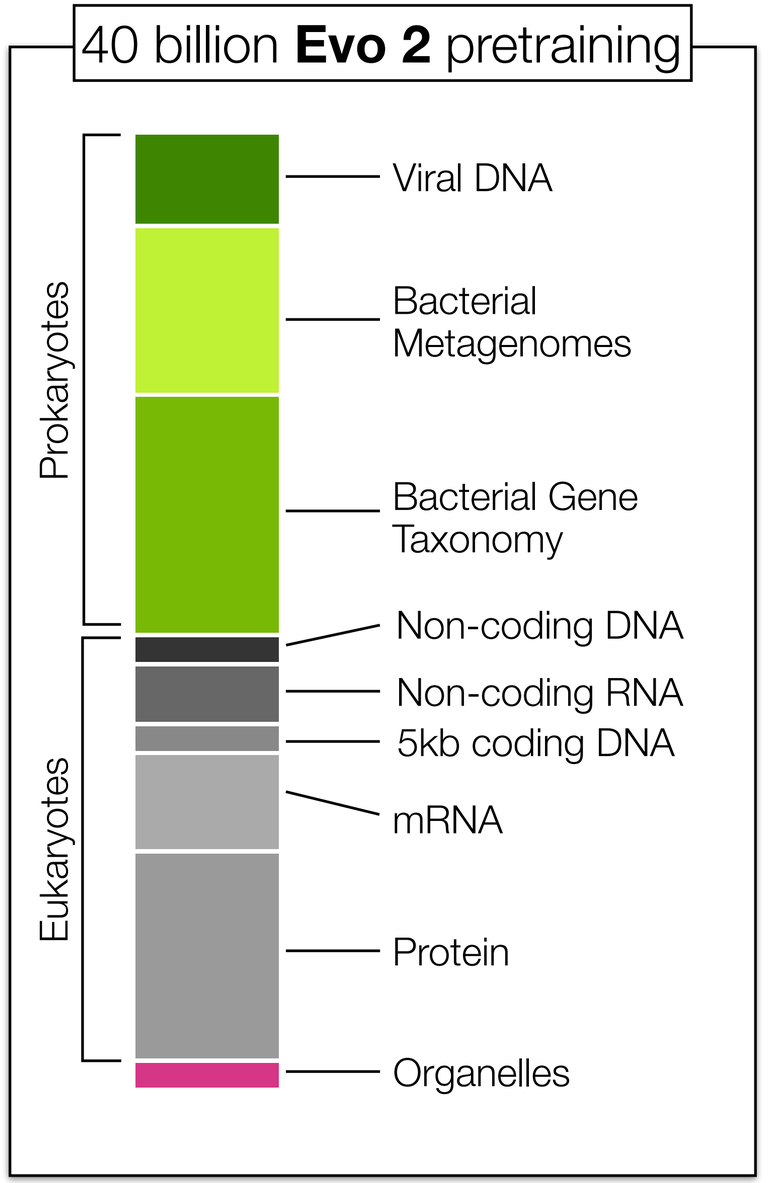
Genome modelling and design across all domains of life with Evo 2

40 billion parameter frontier deep learning model

Goal: Train model on a diversity of genetic evolution to predict nucleotide functional importance

Assumption: Natural selection drives genetic advantages

Training: DNA, RNA, protein sequences from bacteria, archaea, eukarya, and bacteriophages



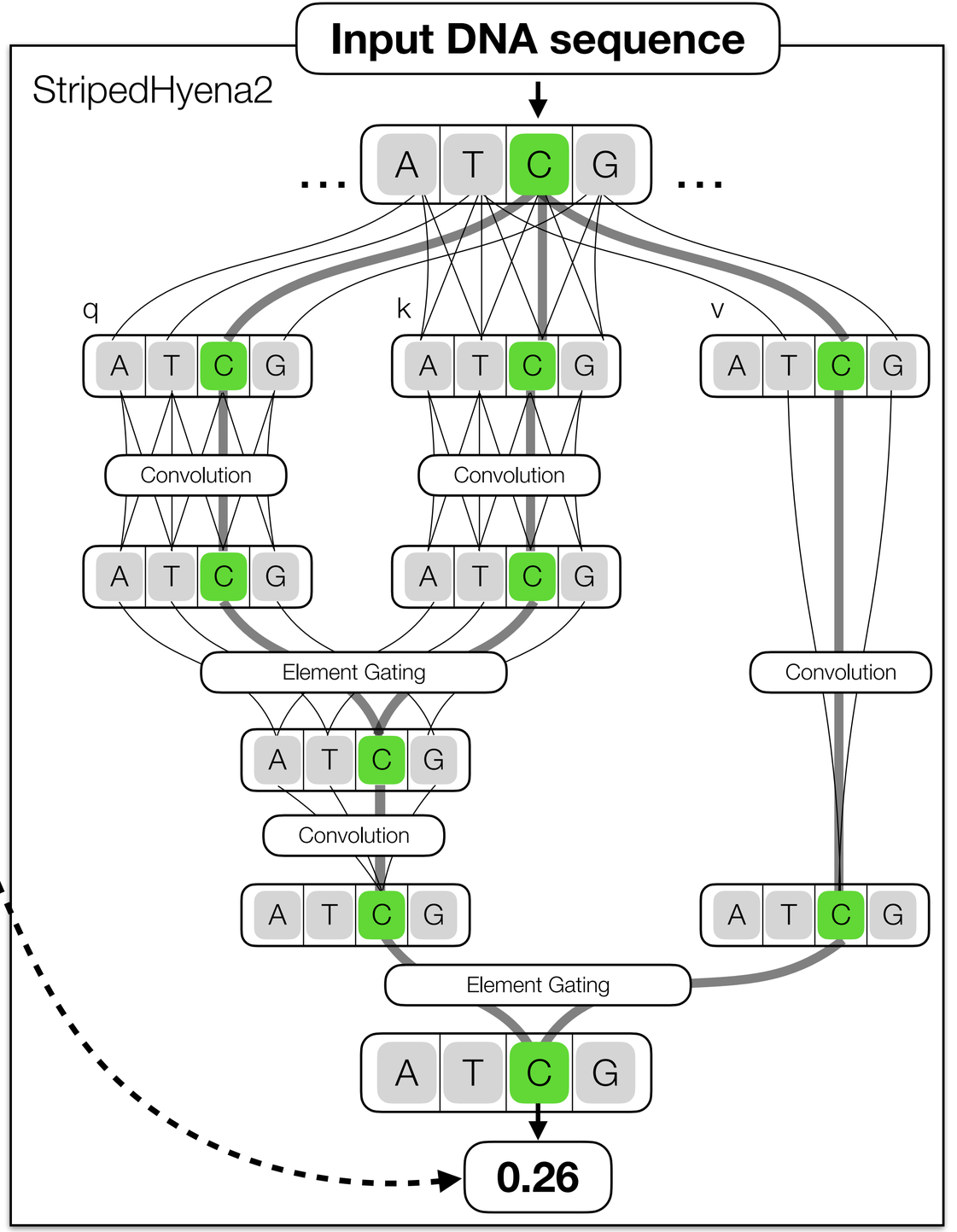
Brixi *et al.,* 2026; Wilber *et al.,* 2025

Evo 2

Output: predict ‘functional relevance’ of each basepair

Predicted score is a proxy for molecular function or organismal fitness

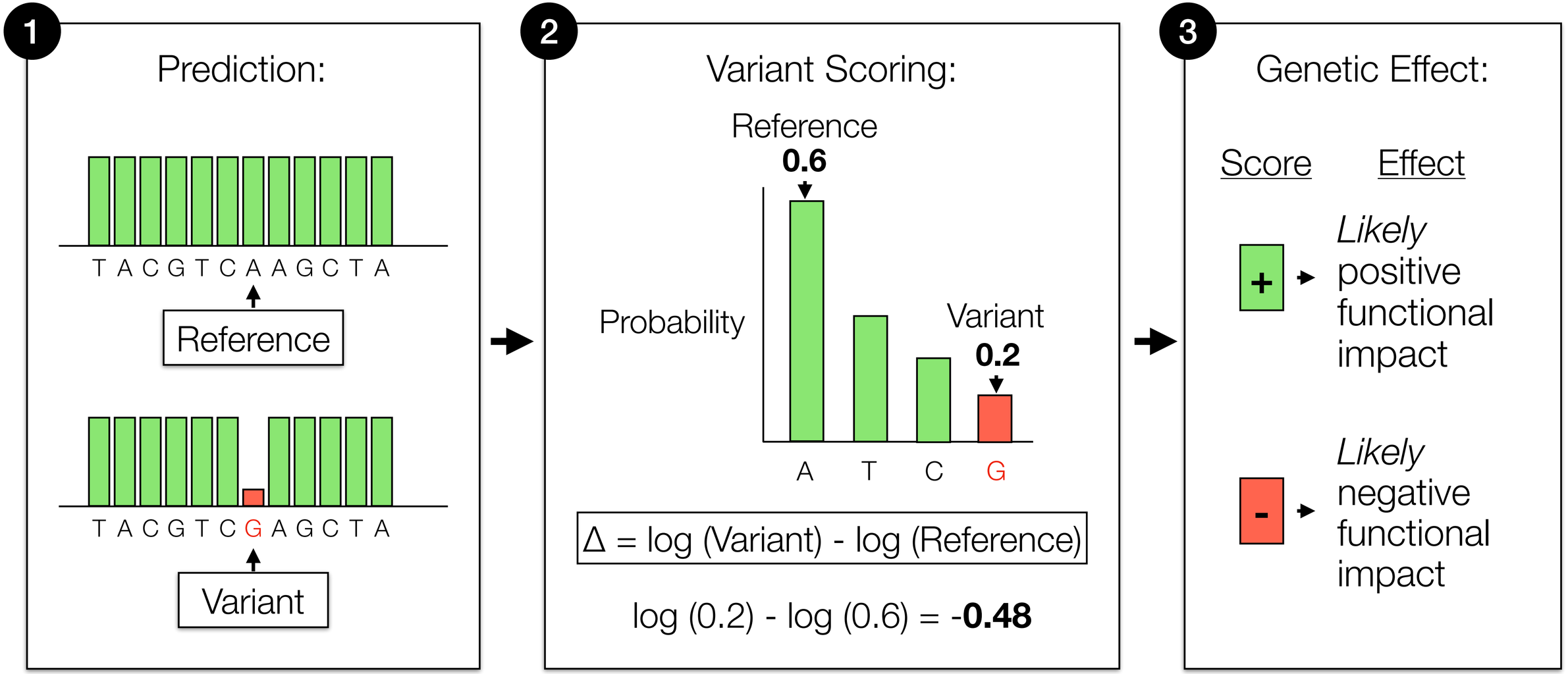
This design enables the calculation of **variant effects**



Brixi *et al.,* 2026; Wilber *et al.,* 2025

Evo 2 Variant Effect Prediction

Comparing 2 DNA sequences



"Functional impact" ~ will the gene/organism retain its function/fitness* ?

Brixi *et al.,* 2026; Wilber *et al.,* 2025

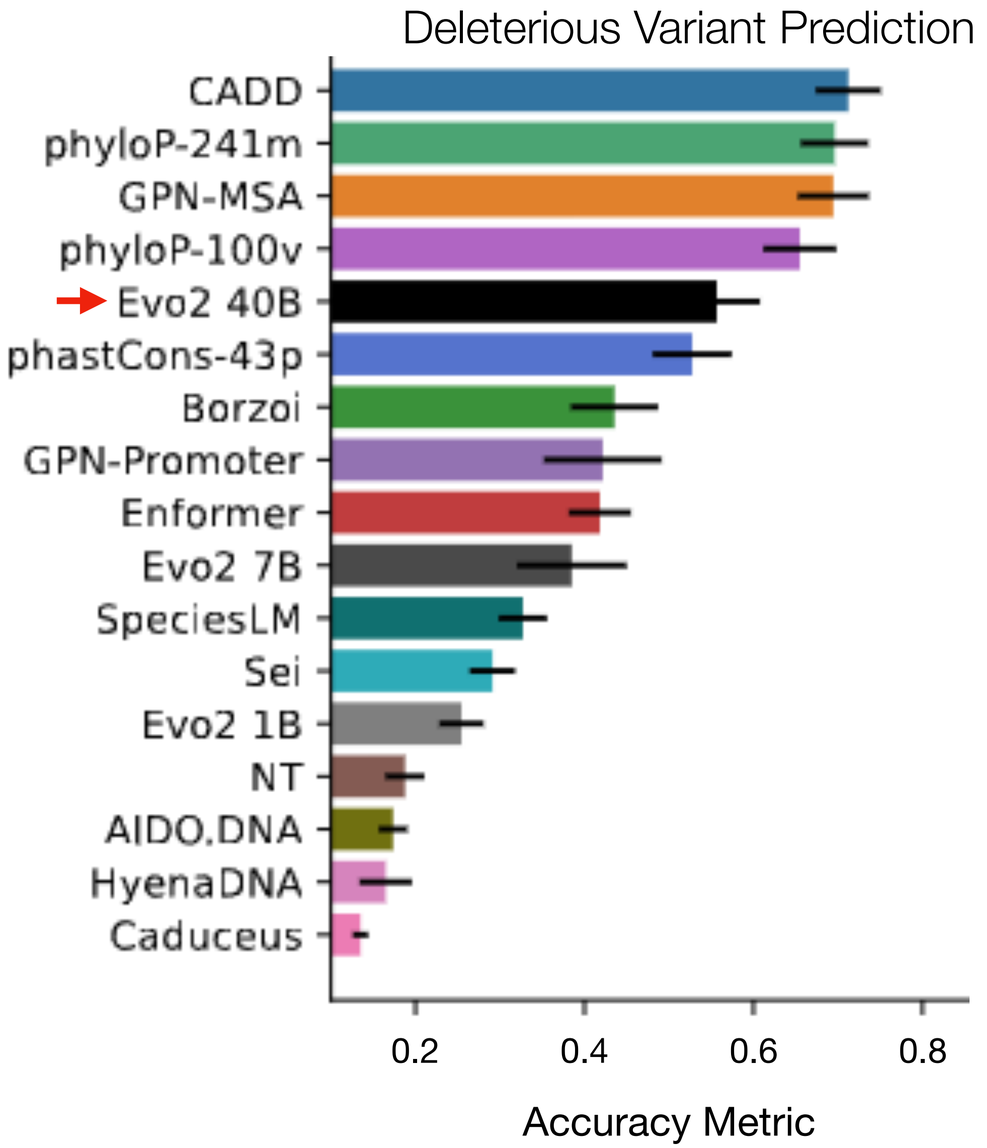
* Broadly speaking

Variant prediction accuracy

In a test of classifying disease-associated variants in human genomes, the 40B parameter Evo 2 model made some of the most accurate predictions

Evo 2 has been used to find novel variants associated with Alzheimer's disease

Ability to make accurate predictions opens up new possibilities to leverage models to discover novel genetic function



Benegas *et al.,* 2025b; Zhu *et al.,* 2025

Deep Learning Computational Demand

Computational requirements for pretrained deep learning model inferencing are lower than you might think

Runtimes on 10th gen i7, 64GB ddr4 RAM, 12GB GPU VRAM (~2 generations old):

AlphaFold2 model: ~18 minutes per protein

Evo 2 1B model: ~30 minutes for all sequences

Total cost to reproduce this set up: ~**1,400$***

This same level of compute on Amazon Web Services would cost ~**1121.62$/month**

12GB VRAM GPU - RTX 3060

**at the current market rate, excluding electricity*

Source: *aws-pricing.com/gpu.html*

Integrative Modeling

Use labeled datasets to predict specific outcomes or classifications

Pros

+ High accuracy
+ Automation
+ Solves specific challenges

Cons

- Labeled data requirements
- Training / implementation costs

Feature Discovery

Find patterns in vast unlabeled datasets to drive innovation

Pros

+ Pretrained models widely available
+ Low startup costs
+ Scalable discovery

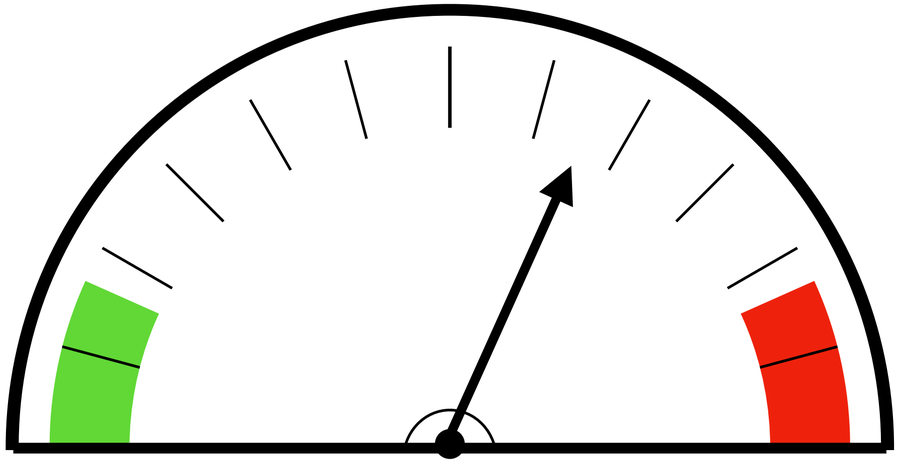
Cons

- Experimental validation burden
- Unknown accuracy

Conclusion

Current economics of implementing machine learning

Integrative modeling



Feature discovery

Without more concrete examples of integrative modeling that make *clear* financial sense, the current economic scales lean in favor of feature discovery implementations*

**In my opinion*