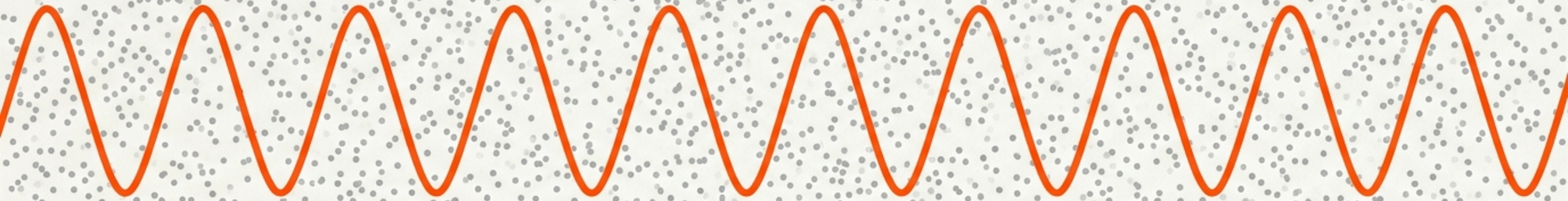
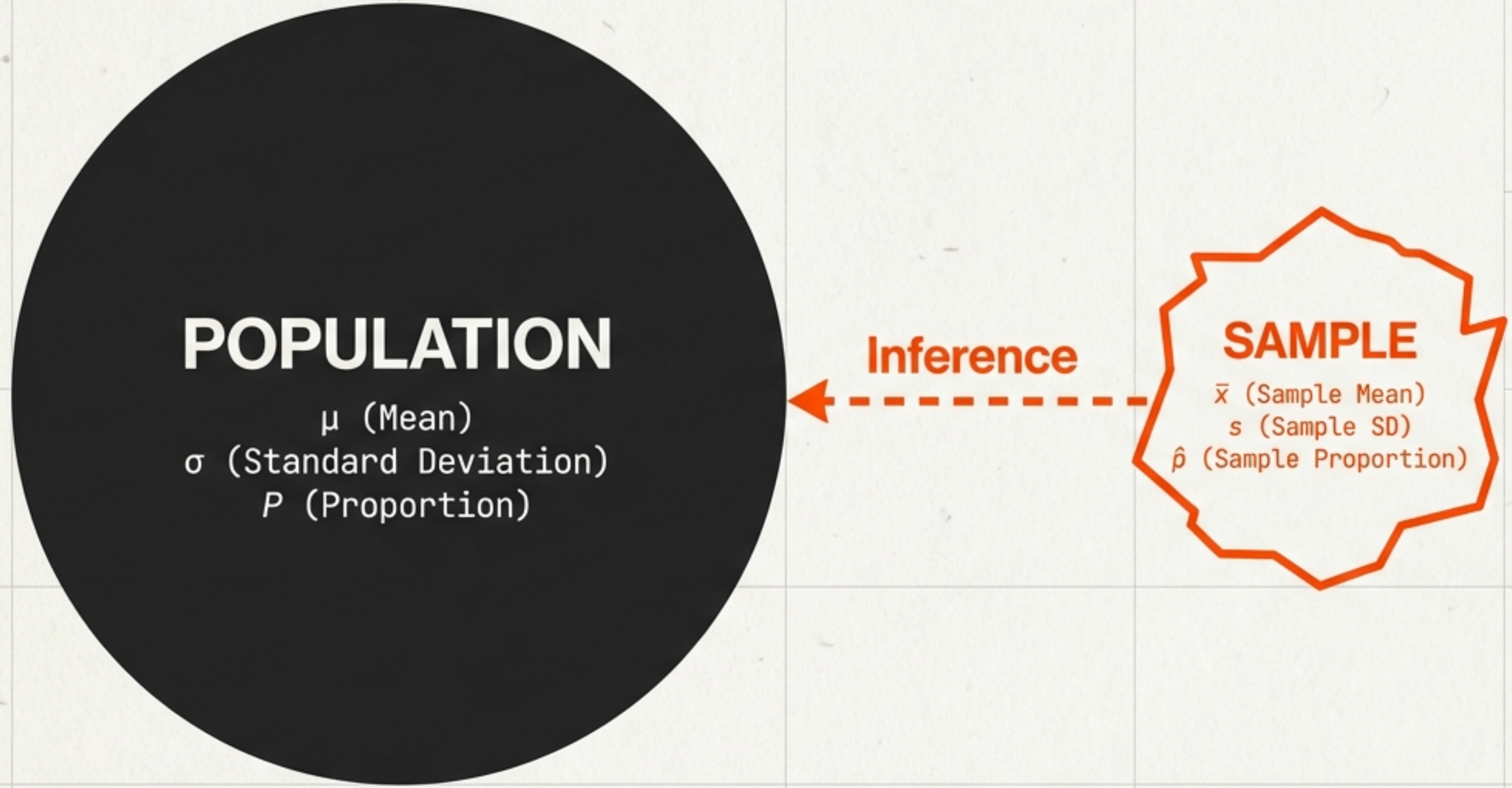
Extracting Truth from Chaos: The Signal in the Noise

Statistics is not merely a collection of formulas; it is a set of philosophical and practical tools designed to extract *signal* from the noise of raw data.



I. Foundations — II. The Bridge (CLT) — III. Inference — IV. The Minefield

The Gap Between What We Have and What We Seek.



The Whole Truth
/ Unknowable.

The Estimate
/ Data Collection.

Parameter: A fixed, usually unknown number describing the entire population (e.g., the true median income of all university students).

Statistic: A known number describing a sample (e.g., the median income of 850 surveyed students).

Quantifying the Chaos: Variance and Standard Deviation

Context Matters: ±26% Variation.

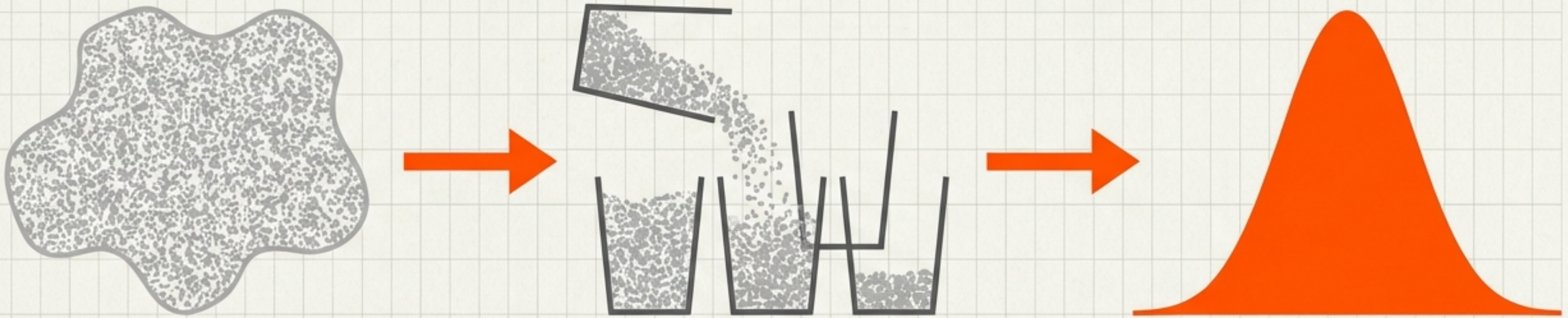
Factory Parts
A ±26% variation is disastrous.

Forest Trees
A ±26% variation is natural.

Variance (σ^2): The average squared deviation from the mean. Mathematically useful, but units are squared and hard to visualise.

Standard Deviation (σ): The square root of variance. Returns metric to original units (e.g., cm), making it the primary tool for visualisation.

The Bridge: The Central Limit Theorem (CLT)



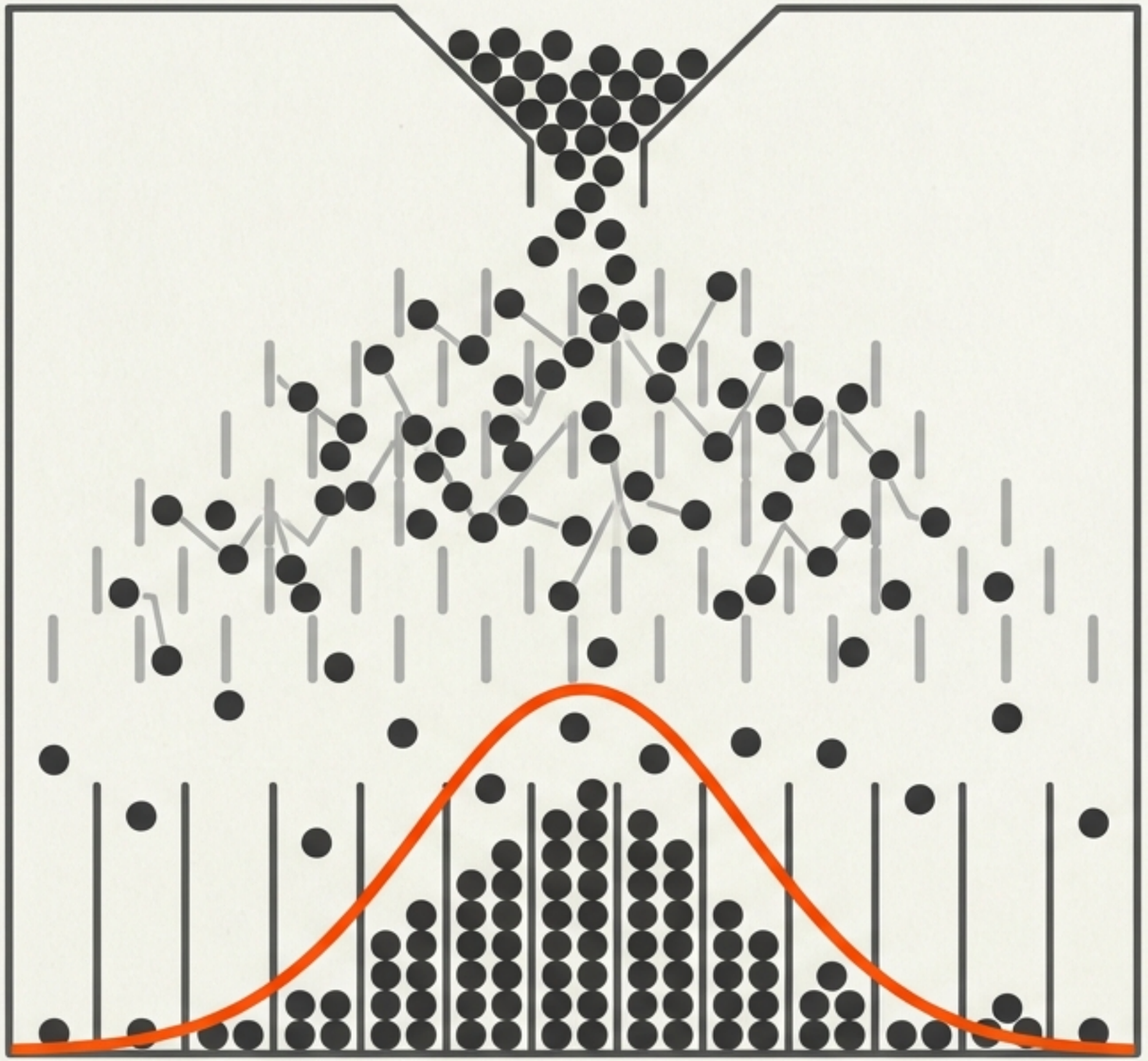
Population Distribution (Any Shape)

Random Sampling (n > 30)

Sampling Distribution of Means (Normal)

The CLT states that if you take sufficiently large independent random samples from ANY population and compute the means, the distribution of those means will approximate a Normal Distribution.

Order Emerging from Anarchy

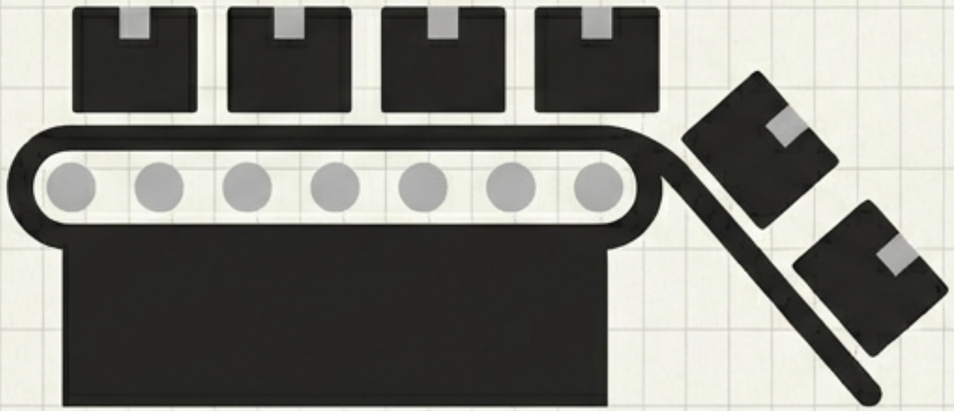


“I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ‘Law of Frequency of Error’. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason.”

— Sir Francis Galton (1889)

The Schism: Two Definitions of Probability

THE FREQUENTIST

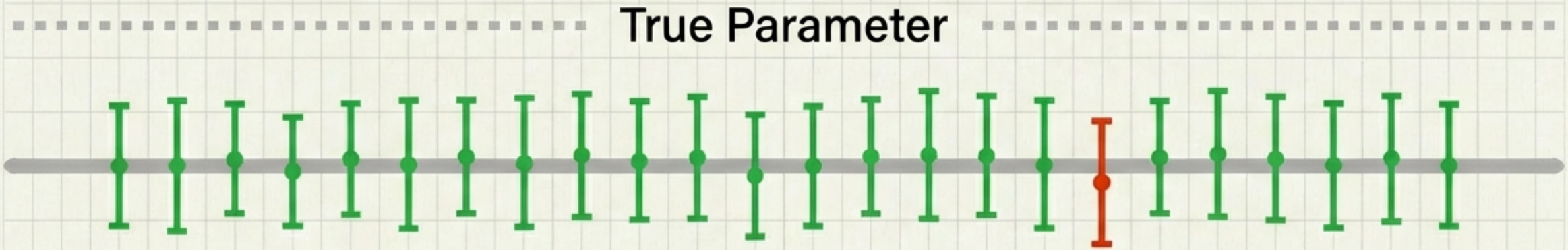


Probability is the LIMIT OF FREQUENCY in repeated measurements. Parameters are fixed values. Data is random.

THE BAYESIAN

Probability is a DEGREE OF BELIEF. Parameters are random variables described by probability distributions. Data is fixed.

The Frequentist: The Limit of Repetition

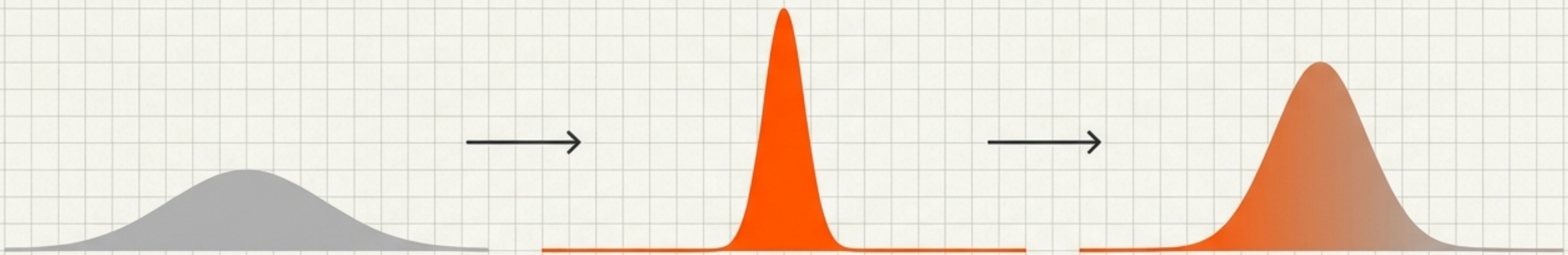


Tool: Maximum Likelihood.
“Which parameter value makes our observed data most probable?”

Confidence Interval (CI): If we repeated this experiment infinite times, 95% of the constructed intervals would contain the true parameter. We cannot say there is a 95% chance the parameter is in THIS specific interval.

The Bayesian: Updating Belief

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$



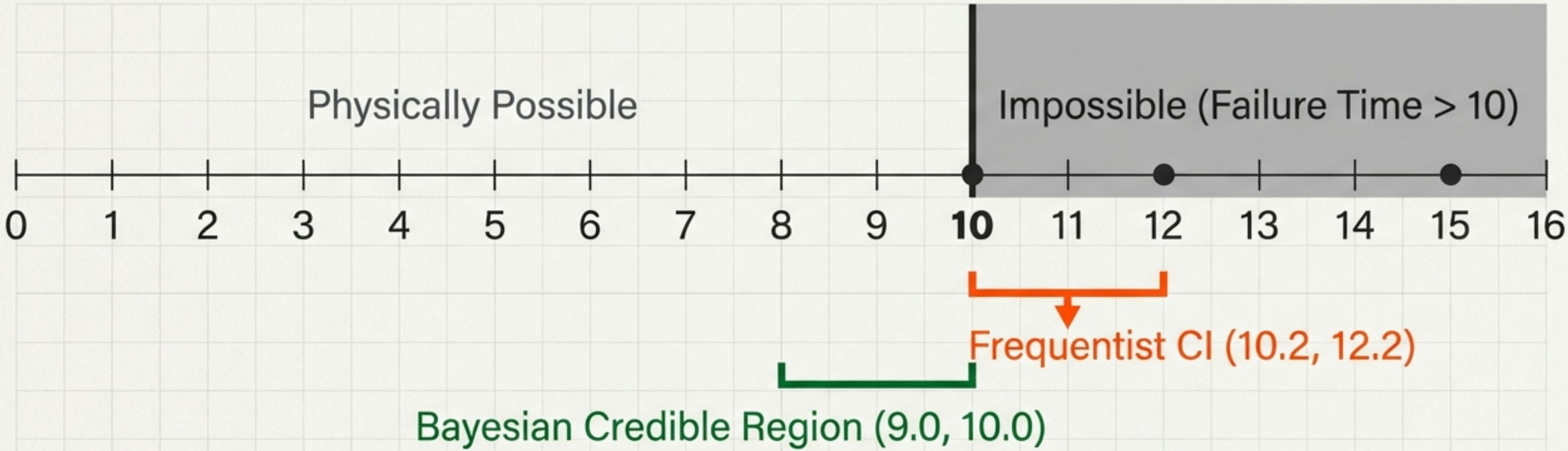
The Prior:
"What we thought before"

The Likelihood:
"What the data says"

The Posterior:
"What we think now"

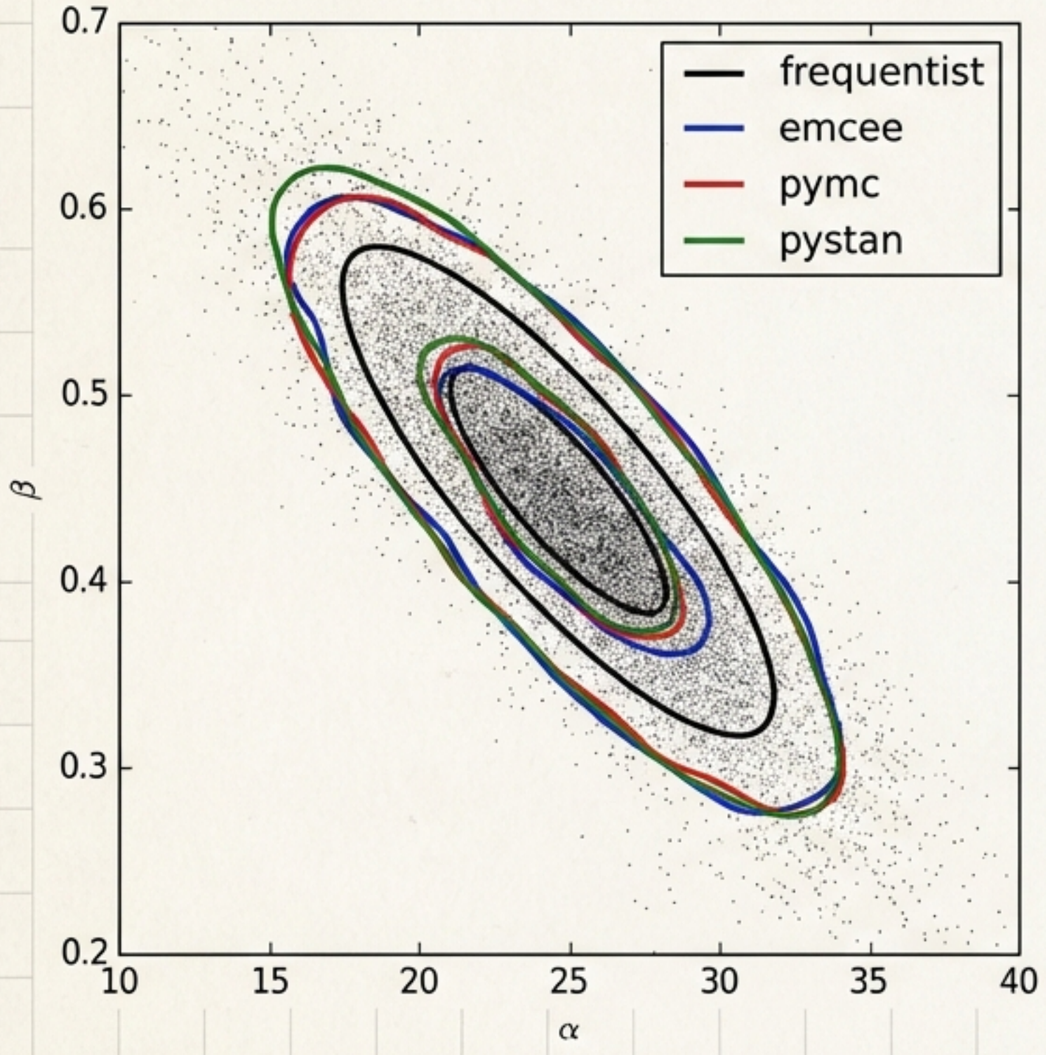
Critique: Frequentists argue Priors are subjective. Bayesians argue that Frequentist assumptions are just implicit priors.

The Battleground: Confidence vs. Credibility



Scenario: A component fails after time θ. Failures observed at t={10, 12, 15}. Common sense says θ must be ≤ 10. The Frequentist CI mathematically excludes the possible values. The Bayesian method incorporates the constraint.

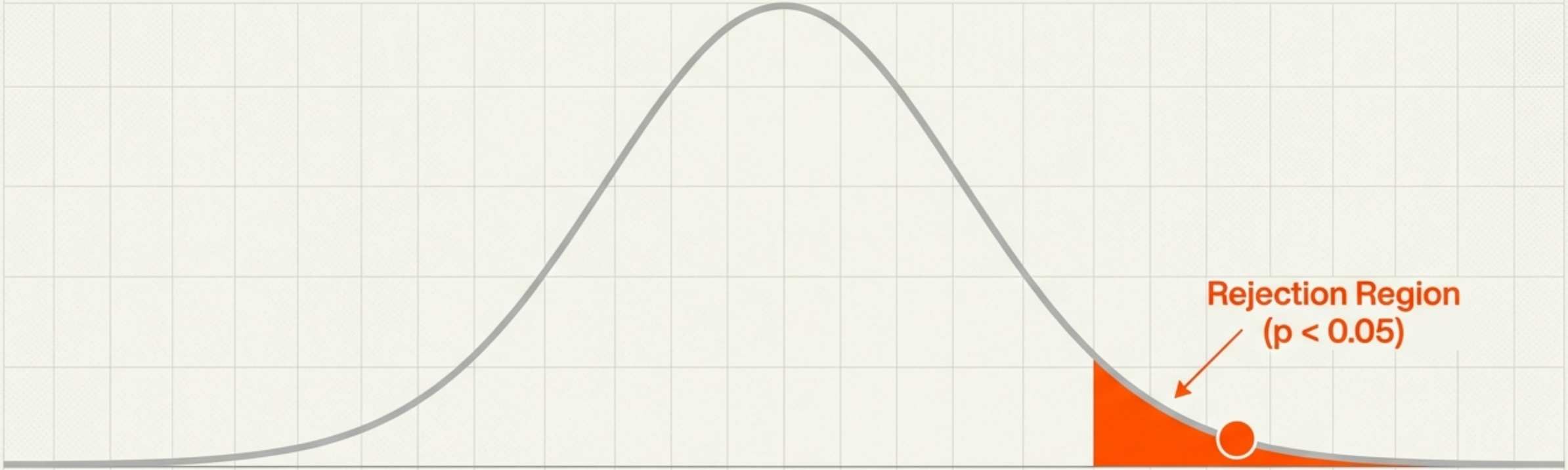
Visualising Inference in Practice



Comparison of Frequentist vs. Bayesian MCMC methods.

- The Black contour (Frequentist) assumes a single point-estimate for scatter, leading to overconfident bounds.
- The Coloured contours (Bayesian MCMC) explore the full shape of the posterior, capturing high-dimensional uncertainty.

Judgment Day: Hypothesis Testing



Null Hypothesis (H_0): The assumption of “no effect” or “nothing to see here”.

P-Value: The probability of observing a result this extreme, ASSUMING the Null is true.

Interpretation: A measure of surprise. A low p-value means the data is highly surprising under the Null, so we reject the Null.

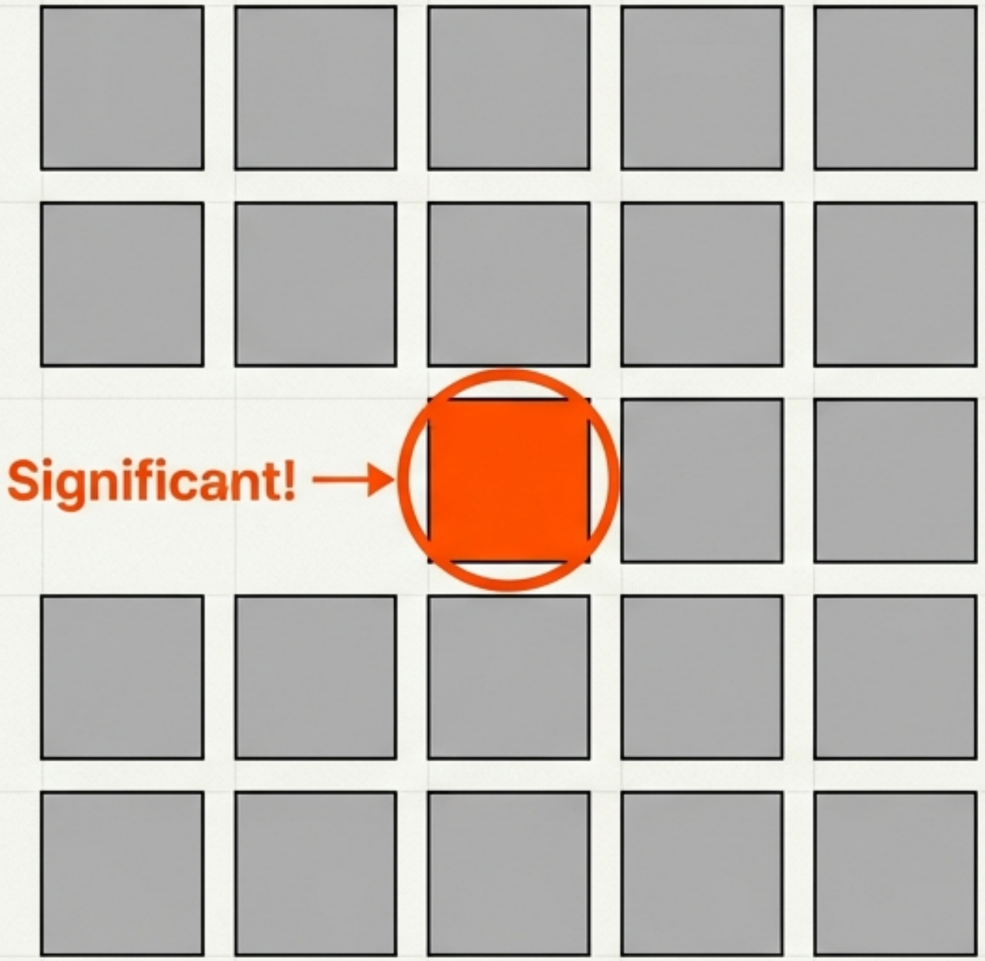
The Dark Side of Significance: P-Hacking

The Problem: Data Dredging.

Good Research: Theorise effect → Collect Data → Test.

Bad Research: Collect massive data → Test every correlation → Publish only the one where $P < 0.05$.

Consequence: If you test 20 random variables, one will likely appear "significant" by pure chance.



Correlation ≠ Causation.

Correlation: Statistical association.
Causation: Change in A brings about change in B.

The Traps:

1. Third Variable Problem: Heat causes both ice cream sales and crime.
2. Directionality: Does low Vitamin D cause depression, or does depression cause low Vitamin D?
3. Spurious Correlations: Statistical artefacts (e.g., Storks vs. Birth Rates).

The Gallery of Deception

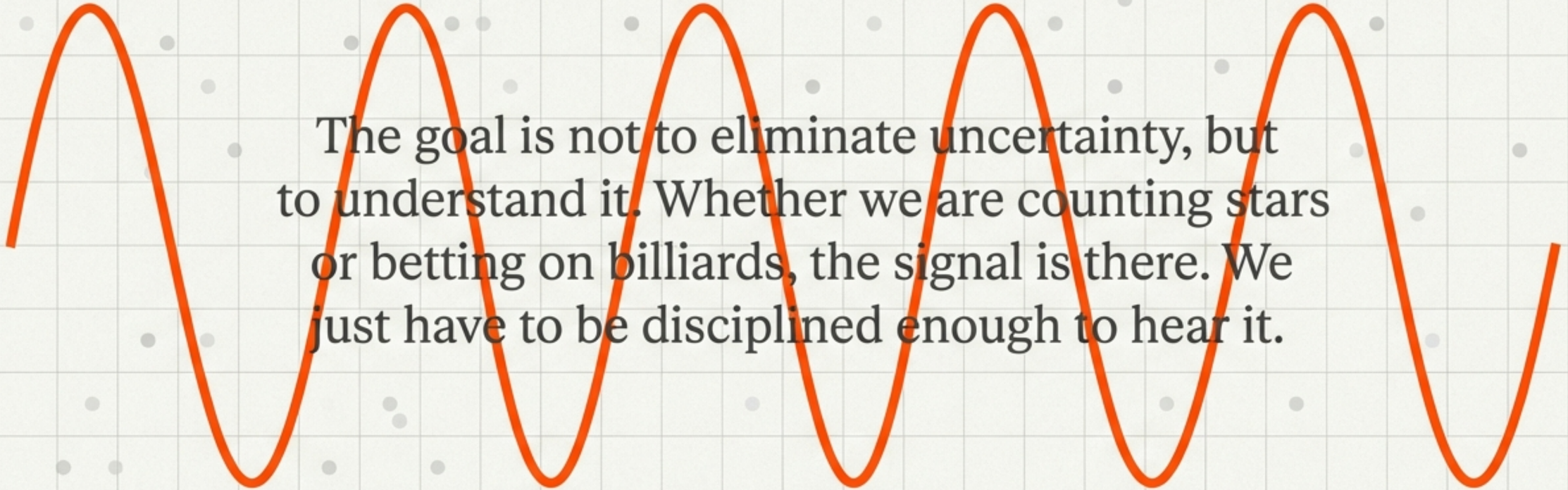
Survivorship Bias
(The Missing Bullet Holes)

The Cobra Effect
(Perverse Incentives)

Cherry Picking

The McNamara Fallacy
(Ignoring the Unmeasurable)

Navigating the Noise.



The goal is not to eliminate uncertainty, but to understand it. Whether we are counting stars or betting on billiards, the signal is there. We just have to be disciplined enough to hear it.

END OF PRESENTATION.