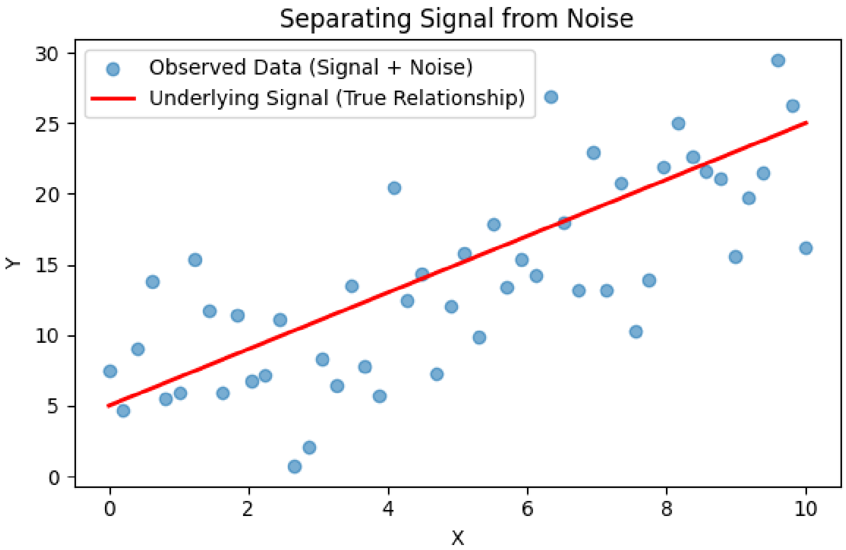
Probability and Inference

- The world is messy — every dataset is full of randomness and error.
- Our goal as data scientists:
  - Predict outcomes
  - Understand relationships between variables
- But we never see the "true" world directly — only noisy samples.
- Probability & inference give us tools to separate real signal from random noise.

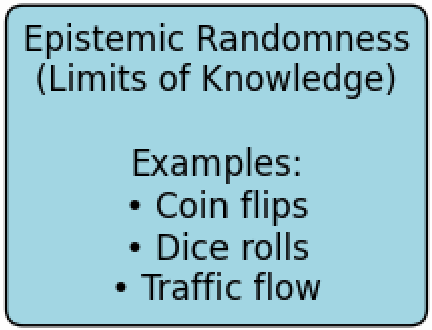
Signal vs. Noise

- Data are noisy: every measurement has randomness.
  - Measurement noise, sampling variability, natural variability, unobserved variables, random process, data processing
- Goal: separate real patterns (signal) from chance fluctuations (noise).
- Statistical inference = the formal framework for this separation.



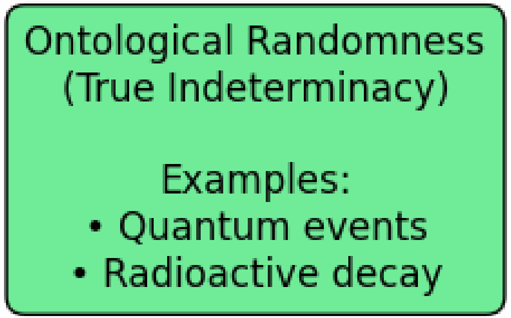
Random Variables

- Random variable = a numerical outcome of a random process.
  - What’s an example of a non-random process?
- Variability in data → modeled as distributions (Normal, Binomial, etc.).
- Probability quantifies uncertainty, not certainty.



Epistemic Randomness
(Limits of Knowledge)

Examples:
• Coin flips
• Dice rolls
• Traffic flow



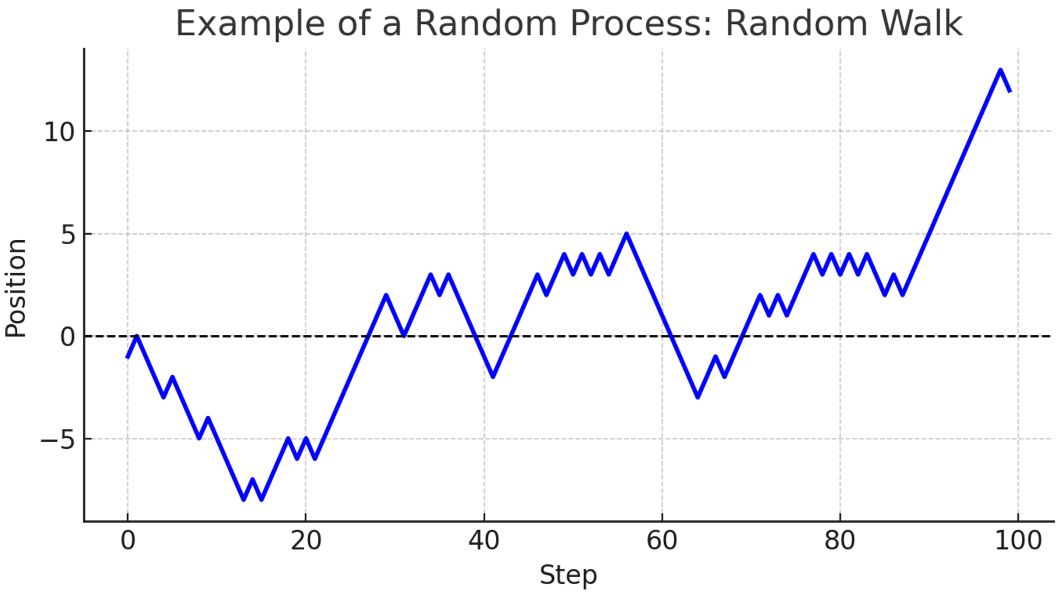
Ontological Randomness
(True Indeterminacy)

Examples:
• Quantum events
• Radioactive decay

Random Processes

- A sequence of random variables evolving over time or trials.
- Examples:
  - Rolling a die repeatedly.
  - Stock price changes.
  - Customer arrivals in a store.
- We model them as random to quantify uncertainty.

Random Walk



- np.random.seed(42)
- steps = np.random.choice([-1, 1], size=n_steps)
- walk = np.cumsum(steps)

- Is it really random?

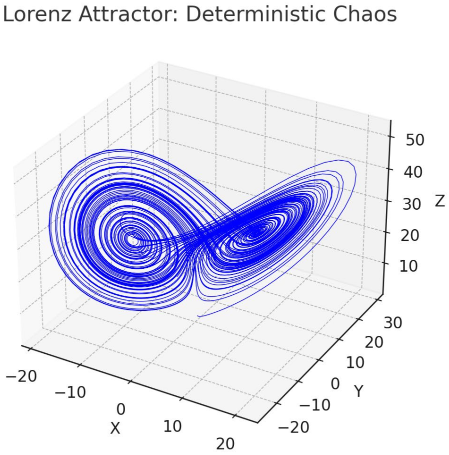
True Randomness

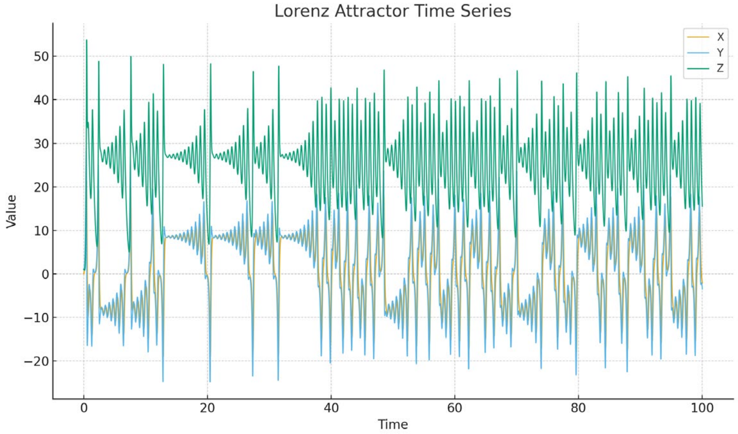
Chaos is not Random

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y$$

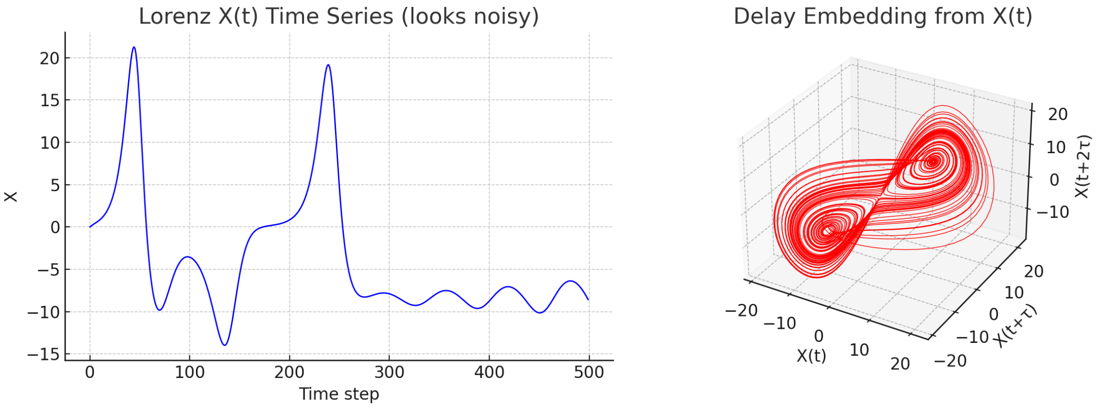
$$\frac{dz}{dt} = xy - \beta z$$





Embedding in Timeseries

- Left: $x(t)$x(t) time series — looks noisy and random.
- Right: Delay embedding $(x(t),x(t+\tau),x(t+2\tau))$(x(t),x(t+τ),x(t+2τ)) — the butterfly structure emerges.



From Noisy Time Series to Structured Signal via Embedding

Lorenz X(t) Time Series (looks noisy)

Delay Embedding from X(t)