Engineering Observations Sequences and Hidden States

of a

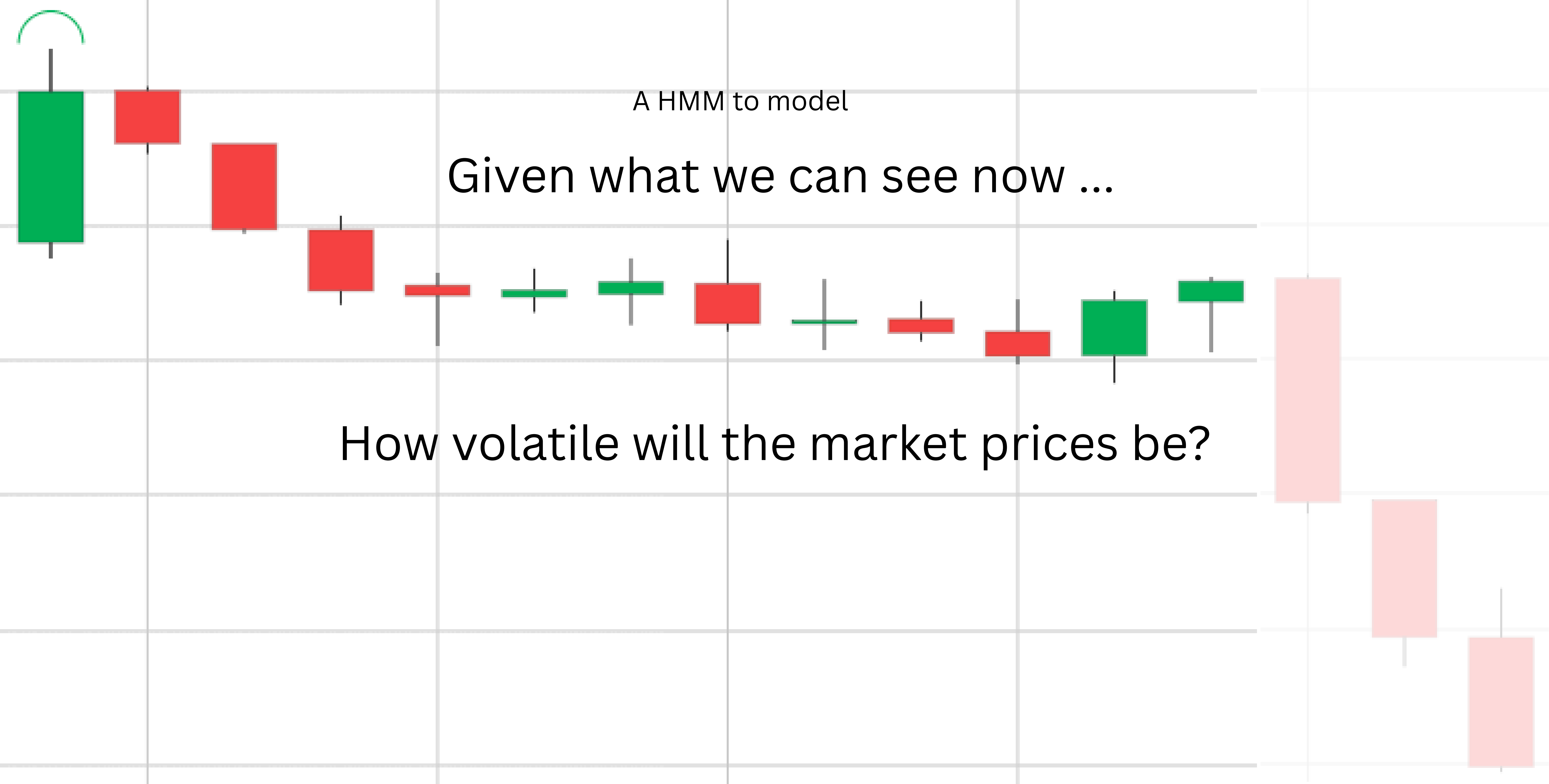
Hidden Markov Model (HMM)

Author: Ben Darby
Class: BINF6250 Spring 2025
Professor: Marcus Sherman

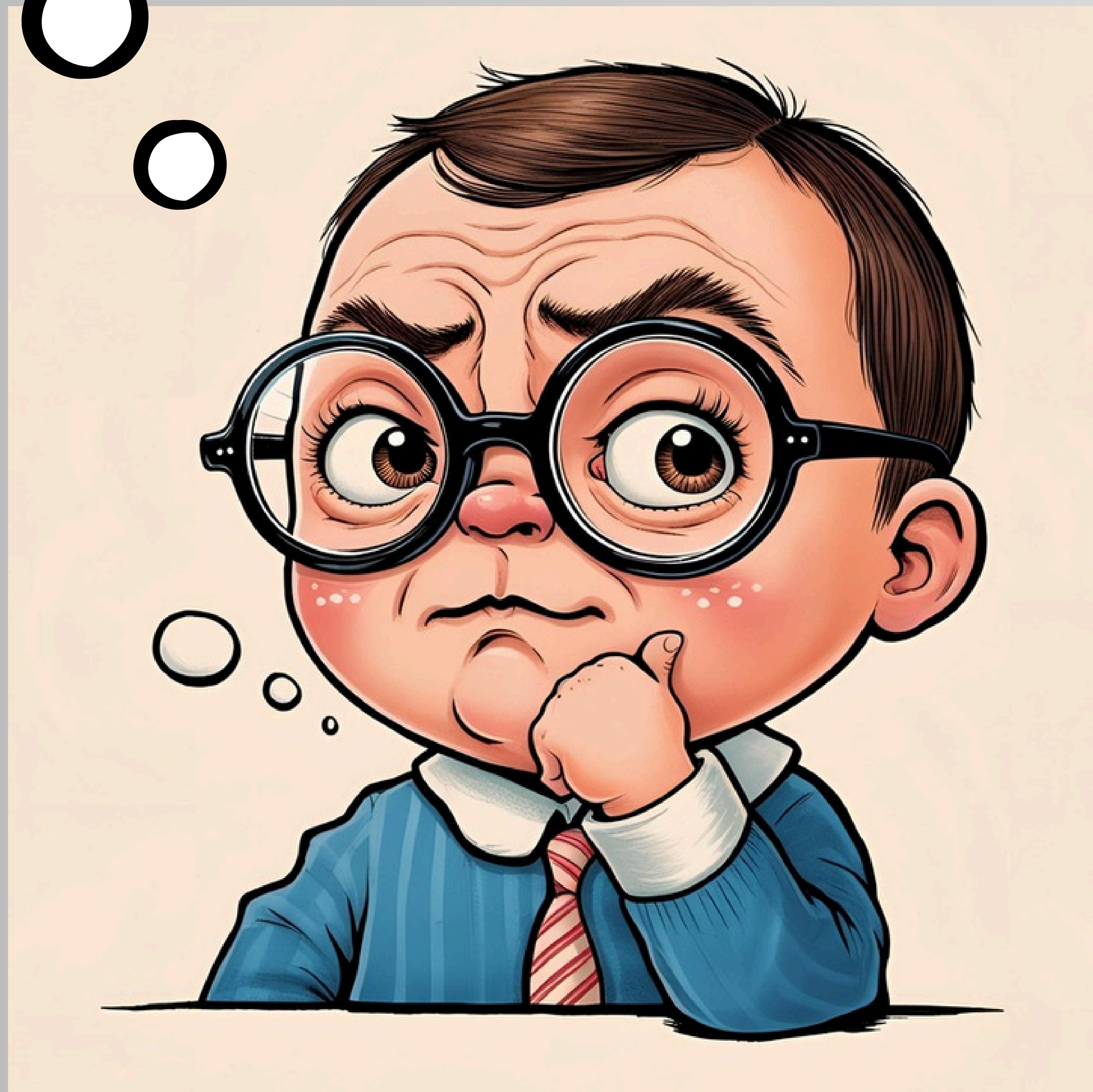
A HMM to model

Given what we can see now ...

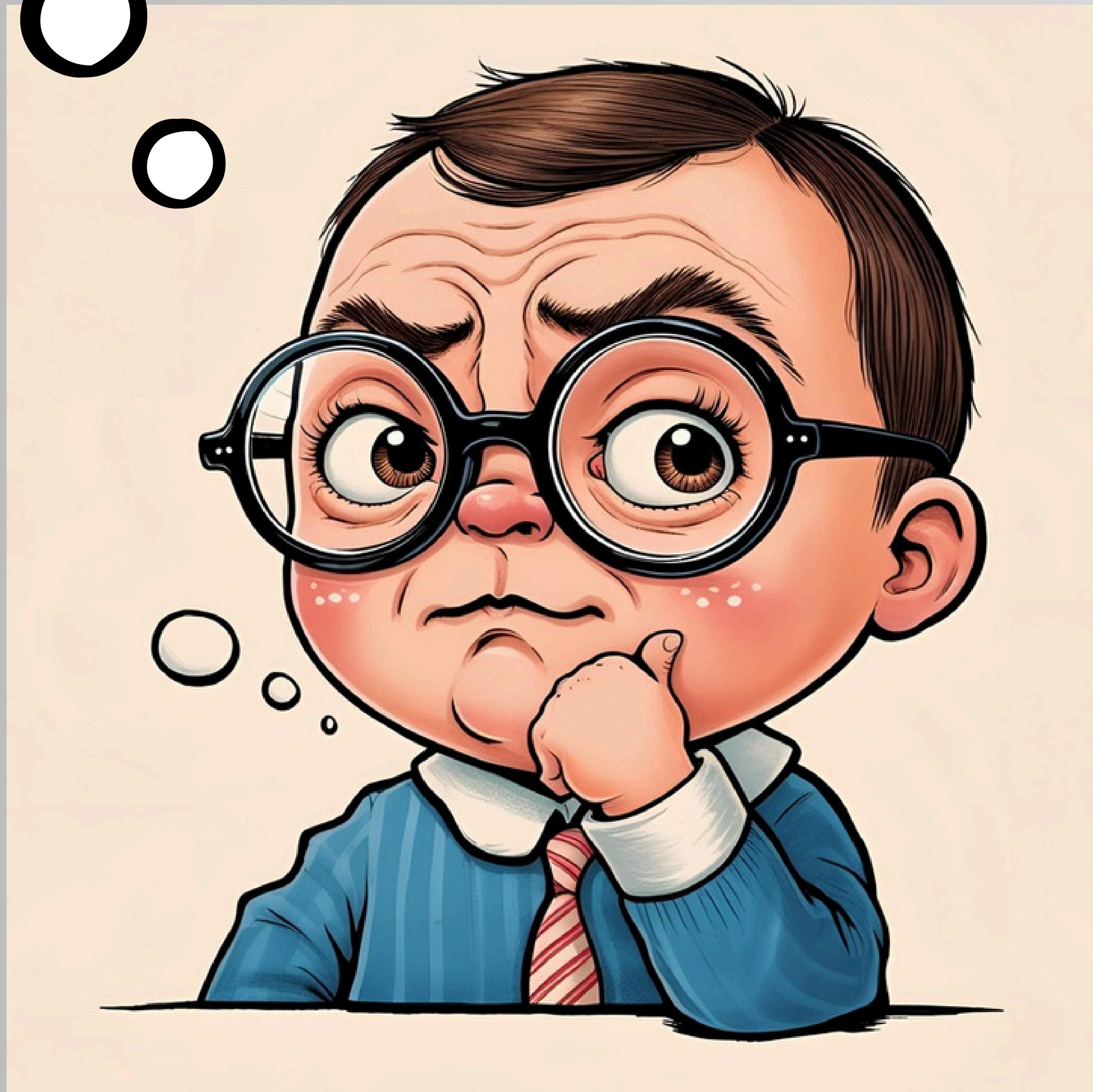
How volatile will the market prices be?



What does a HMM require?

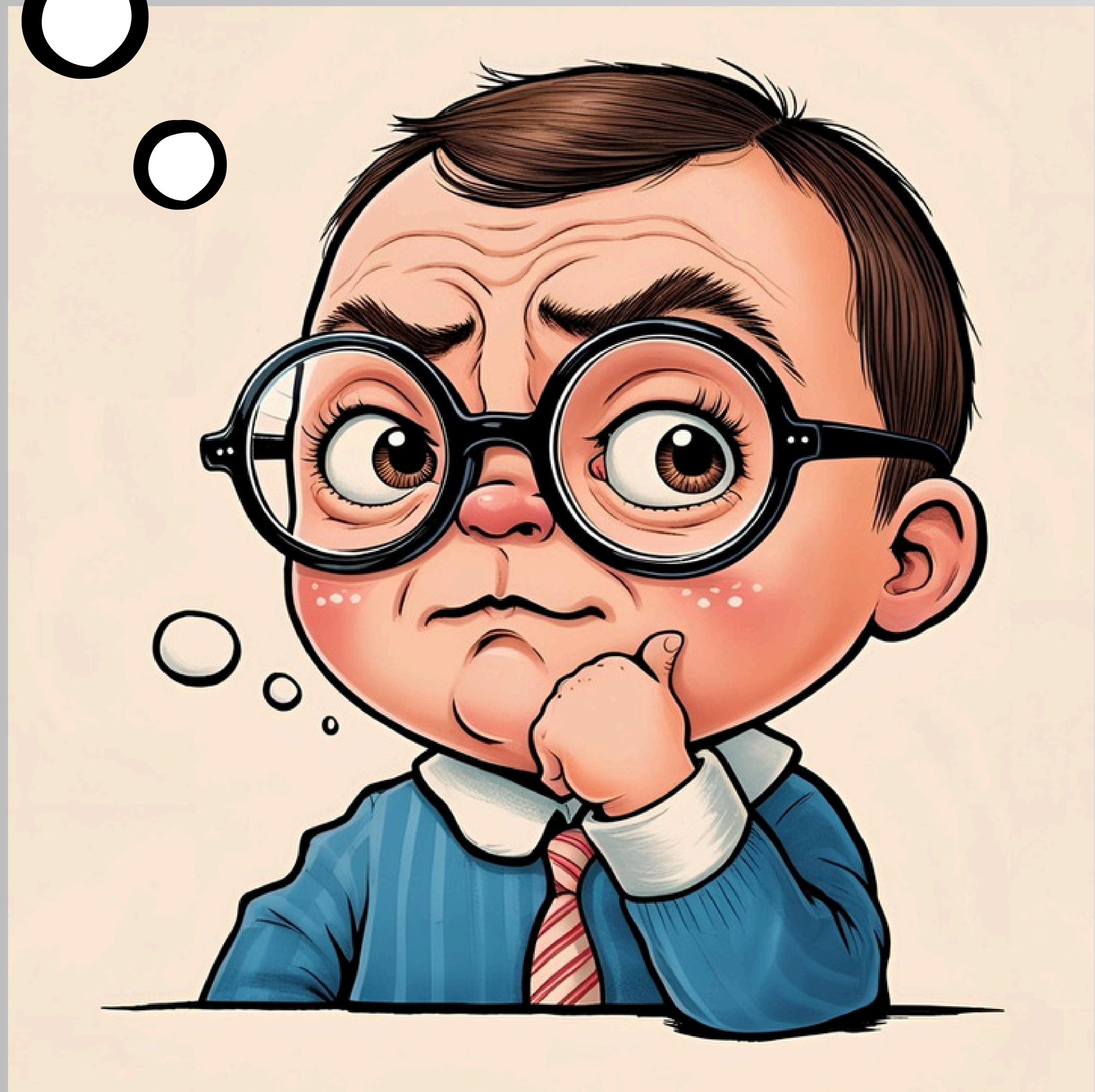


What does a HMM require?



1. **Observable Sequences**
2. **Hidden States: S_1, S_2, S_3**
3. **Model Parameters λ (π, A, B)**

What does a HMM require?



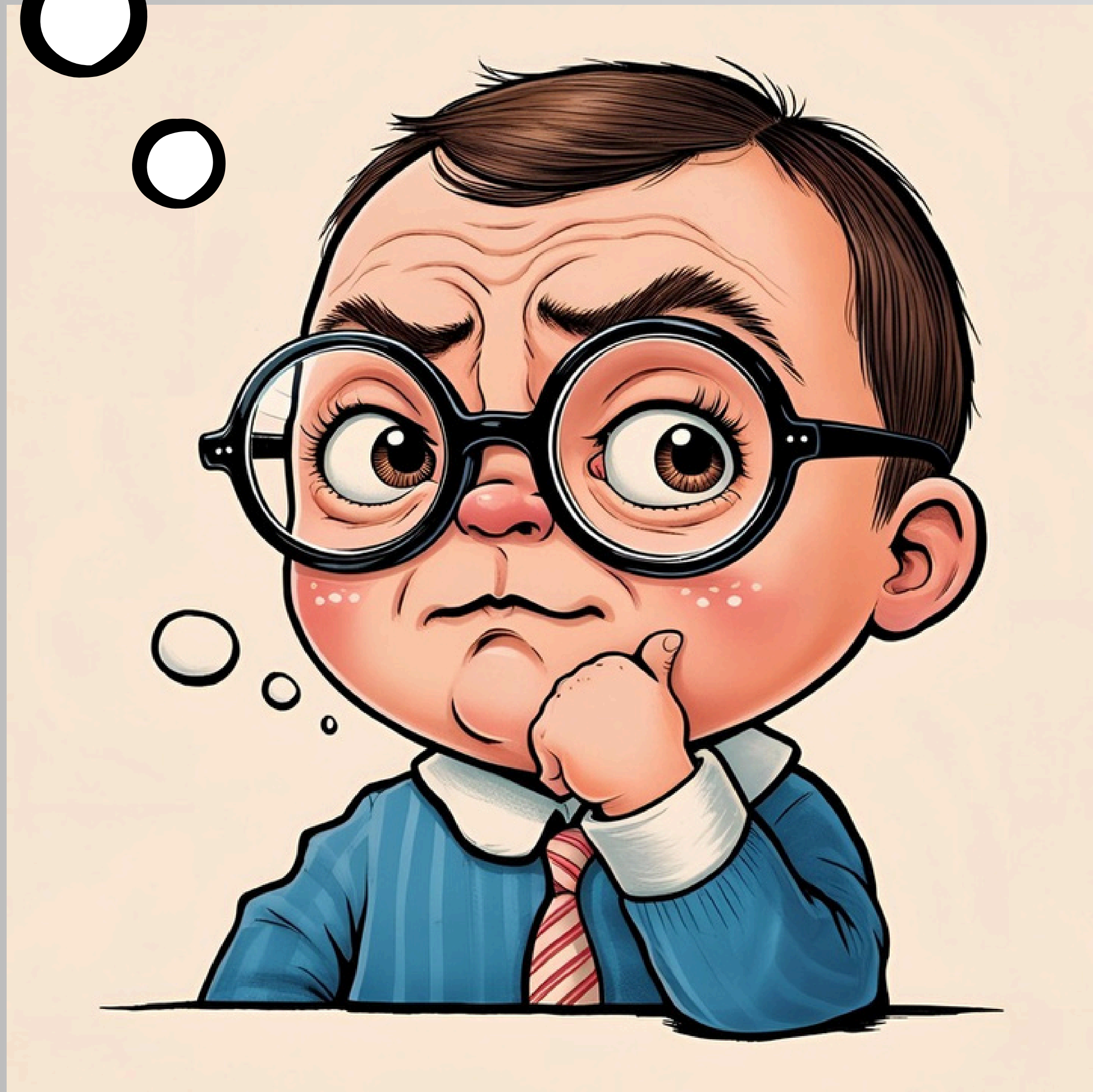
1. Observable Sequences

- Sequential data where order matters
- Data that is identifiable (visible / measurable)
- Must have predictive relationship to hidden states
- Conditional probability critical for state inference

What does a HMM require??

2. Hidden States: S_1, S_2, S_3

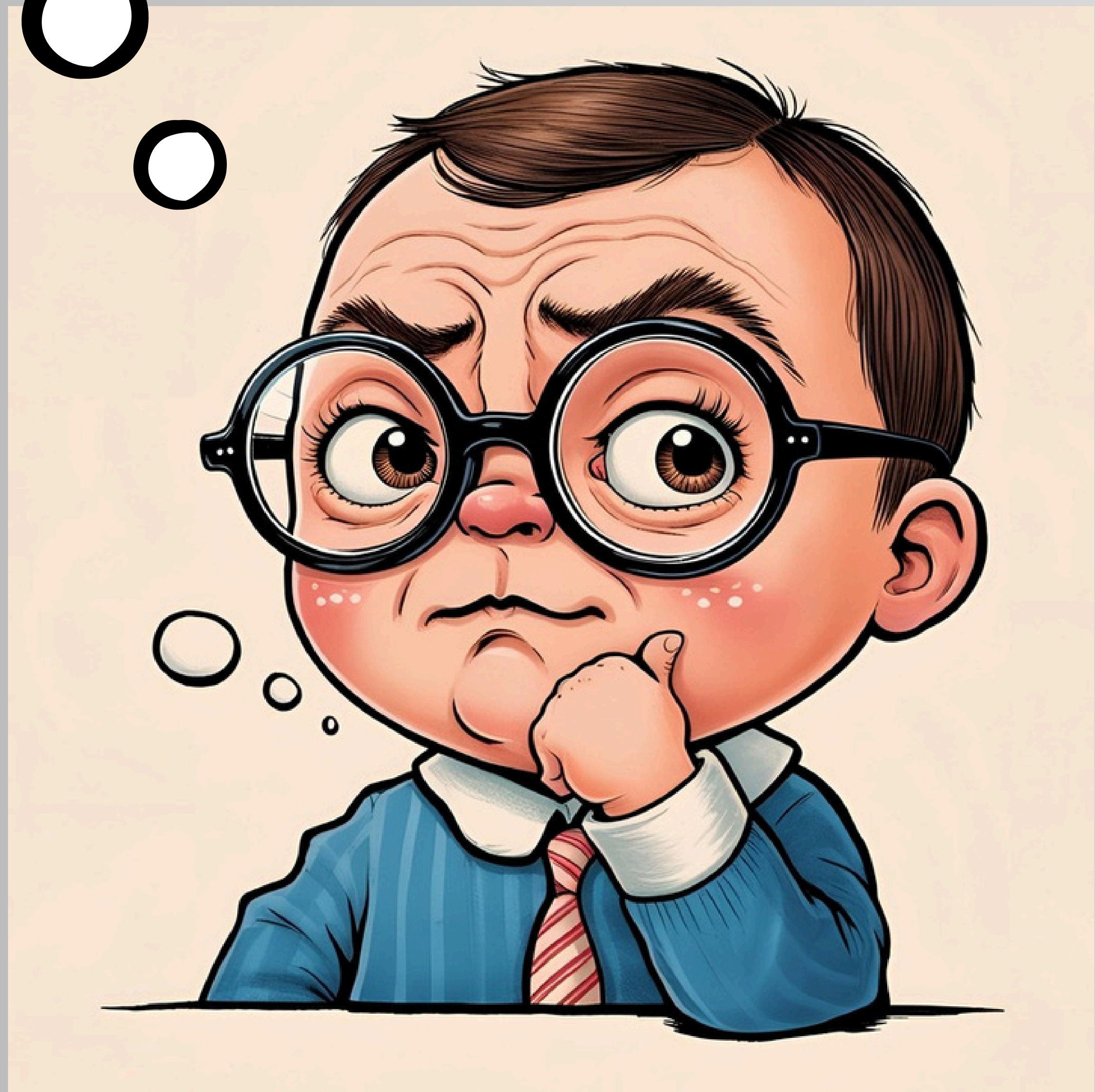
- Underlying, unobservable conditions of a system.
- Inferred from observable data.
- The system exists in exactly one state per time step.
- The system transitions between them via probabilities.



What does a HMM require??

2. Hidden States: S_1, S_2, S_3

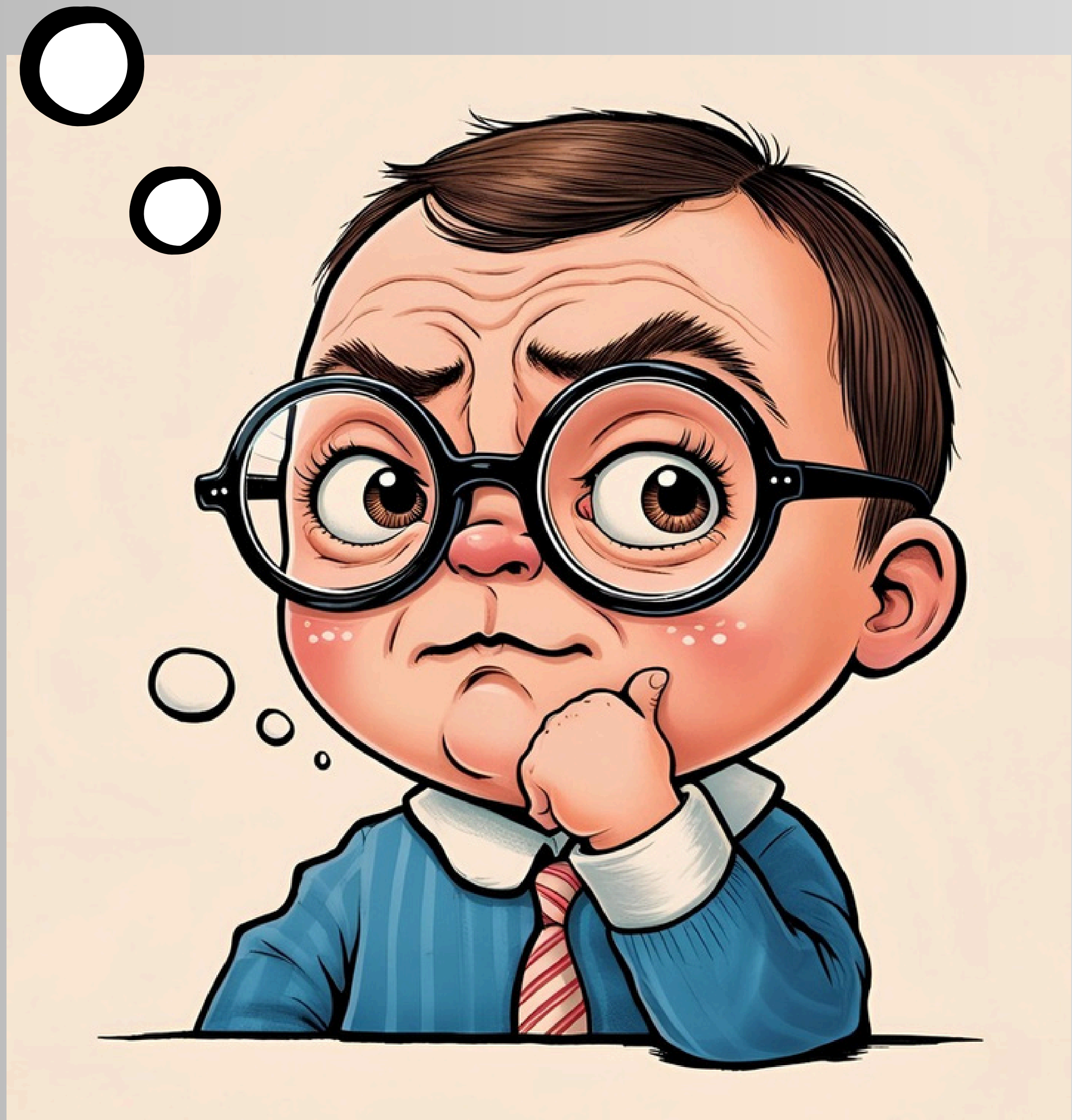
- S_1 - Price Volatility: Low
- S_2 - Price Volatility: Normal
- S_3 - Price Volatility: High



Price Volatility Range

"Compared to the average time period, is the range of prices as measured by the high price minus the low price for this time period...Low, Normal or High."

What does a HMM require??



3. Model Parameters λ (π, A, B)

- **π: Initial Probabilities**
  - a vector of probabilities showing how likely the system is to start in each hidden state.
- **A: Transition Probabilities**
  - Shows probability of moving between states
- **B: Emission Probabilities**
  - Shows how likely each state is to generate each observation.

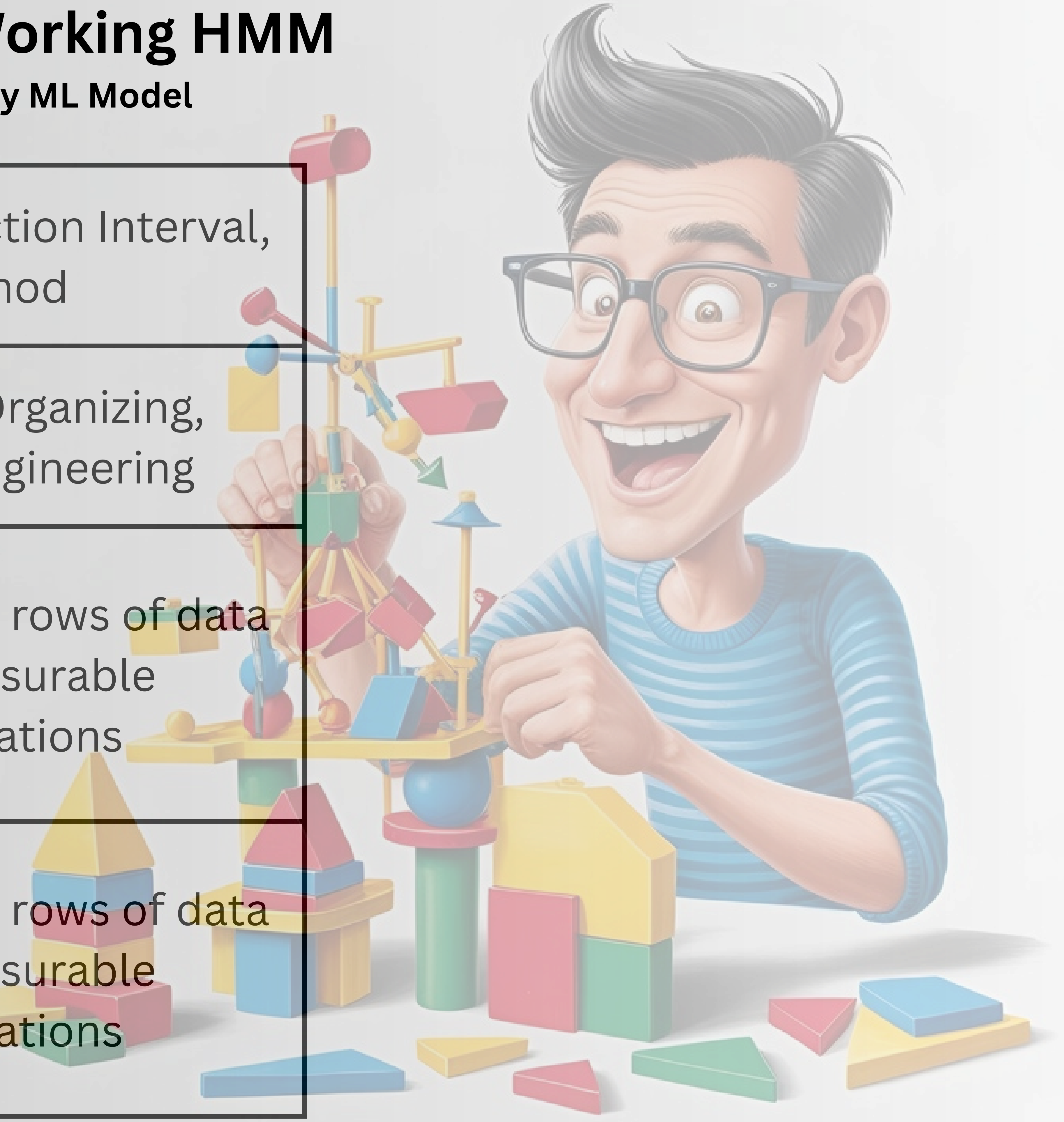
Applying HMM’s to Unrelated Domains

Model Attributes	Bioinformatics	Financial Volatility
Hidden States S	"H" (high GC) and "L" (low GC)	"High Volatility", "Medium Volatility", "Low Volatility"
Observations O	DNA sequences (A,C,G,T)	Volume Regime, (I, S, C) Price Range Class (E, S, D)
Initial Probabilities π	Starting in H or L state	Starting in each Volatility State
Transition Probabilities A	Moving between GC states	Moving between Price Volatility States
Emissions B	Nucleotide probabilities in each state	Volume & Range Regime Probabilities in each State

Building a Working HMM

or really any ML Model

Get the Data	Source, Collection Interval, Method
Prepare the Data	Cleaning, Organizing, Feature Engineering
Engineer Hidden States from Data	Transform the rows of data into measurable observations
Engineer Observation Sequences from Data	Transform the rows of data into measurable observations



Sequences of ***Observations*** are calculated from market data which looks like this

Each bar is a 1 minute Data

1 Min Price Data from 15:15-15:16

1-Min Volume Data from 15:29-15:30

Get the Data

OHLC Bars

Open - Edge of colored rectangle
High - Top wick
Low - Bottom wick
Close - Edge of colored rectangle

Can we engineer these bars into observation sequences?

Do they meet the criteria?

Market Data Sample

QQQ 1 Minute data

Get the Data

Prices

Open **High** **Low** **Close**

Time

Row #

1	timestamp	open	high	low	close	volume
112405	2023-04-11 15:20:00	317.09	317.1	317.02	317.07	57863
112406	2023-04-11 15:21:00	317.07	317.11	317.05	317.0998	25850
112407	2023-04-11 15:22:00	317.095	317.19	317.09	317.1299	47240
112408	2023-04-11 15:23:00	317.12	317.21	317.1	317.16	72582
112409	2023-04-11 15:24:00	317.16	317.17	317.03	317.0404	70439
112410	2023-04-11 15:25:00	317.05	317.07	316.91	316.914	70723
112411	2023-04-11 15:26:00	316.9199	316.9616	316.91	316.96	27630
112412	2023-04-11 15:27:00	316.96	316.965	316.87	316.886	58381
112413	2023-04-11 15:28:00	316.89	316.985	316.87	316.93	47827
112414	2023-04-11 15:29:00	316.94	317.0299	316.93	316.98	23636

Raw Data

This is the import format which all features, aka. observation sequences are engineered from.

Prepare Data
Engineering Hidden States

Price range, the ***hidden state,*** is an engineered feature created by discretizing the results of ten 1-minute price observation into classes.

10 Min Price Bar from 15:30:01 - 15:40:00

Each bar is actually made from ...

How do we transform these bars into hidden states?

Do they meet the criteria?

Calculating
Volatility Class (HS)
from QQQ 1-Minute data

Prepare Data
Engineering Hidden States

Prices

Open High Low Close

Time

Row #

Engineered Feature:
10 Min Price Ranges

Created by taking the high minus the low price across ten 1-minute price observations.

1	timestamp	open	high	low	close	volume
112415	2023-04-11 15:30:00	316.99	317.04	316.9	317.01	36309
112416	2023-04-11 15:31:00	317.004	317.03	316.96	316.99	23679
112417	2023-04-11 15:32:00	316.99	316.99	316.805	316.8501	34407
112418	2023-04-11 15:33:00	316.86	316.93	316.85	316.8749	51073
112419	2023-04-11 15:34:00	316.8885	316.8885	316.75	316.77	68194
112420	2023-04-11 15:35:00	316.77	316.79	316.68	316.69	54281
112421	2023-04-11 15:36:00	316.69	316.74	316.58	316.64	72531
112422	2023-04-11 15:37:00	316.63	316.67	316.55	316.61	65585
112423	2023-04-11 15:38:00	316.61	316.61	316.27	316.33	108482
112424	2023-04-11 15:39:00	316.335	316.335	316.12	316.24	163811

Prepare Data
Engineering Hidden States

Calculating
Volatility Class (HS)
from QQQ 1-Minute data

Extract HS from:
10 Min Price Ranges

Create the Hidden State by fitting the observed price range in a precalculated class range, representing the boundaries of Low, Normal or High Volatility.

+ High Price:	**317.04**
- Low: Price:	**316.12**
10 Min Price Range:	**0.92**

Hidden State - Classification

S_1 - Low:	**0.00 - 0.28**
S_2 - Normal:	**0.28 - 0.65**
S_3 - High:	**0.66 - INF**

The Hidden State for the period 15:30:01 - 15:40:00 is classified as: **High - H**

1st type of **observation**
Volume Expansion Regime

Prepare Data
Engineering Observations

Which Volume Pattern?

- **Increasing - I**
- **Steady - S**
- **Contracting - C**

Observing the pattern of volume distributed across a group between time intervals

Increasing Volume?

1st type of **observation**
Volume Expansion Regime

Data Sample

Engineering an ***observation***
from QQQ 1-Minute data

Prepare Data
Engineering Observations

Row #

Time

Volume
by minute

1	timestamp	open	high	low	close	range	volume
112405	2023-04-11 15:20:00	317.09	317.1	317.02	317.07	0.08	[illegible]
112406	2023-04-11 15:21:00	317.07	317.11	317.05	317.0998	0.06	25850
112407	2023-04-11 15:22:00	317.095	317.19	317.09	317.1299	0.10	47240
112408	2023-04-11 15:23:00	317.12	317.21	317.1	317.16	0.11	72582
112409	2023-04-11 15:24:00	317.16	317.17	317.03	317.0404	0.14	70439
112410	2023-04-11 15:25:00	317.05	317.07	316.91	316.914	0.16	70723
112411	2023-04-11 15:26:00	316.9199	316.9616	316.91	316.96	0.0516	27630
112412	2023-04-11 15:27:00	316.96	316.965	316.87	316.886	0.095	58381
112413	2023-04-11 15:28:00	316.89	316.985	316.87	316.93	0.115	47827
112414	2023-04-11 15:29:00	316.94	317.0299	316.93	316.98	0.0999	23636

Which Regime?

1st type of **observation**
Volume Expansion Regime

Prepare Data
Engineering Observations

Engineering a Volume Regime Observation
from QQQ 1-Minute *data*

Engineered Feature: Volume Expansion Regime

Created by identifying the minimum sum of squares from the ideal shape using high and low volume across the time periods.

Lowest Variance from Expected Value: 27, 104

Observation-Classification	Euclidean Distance to Expected Value
Increasing:	27, 104.67
Steady:	44,819.76
Contracting:	77,542.30

The 5 Min Volume Expansion Regime from 15:21:01 - 15:26:00 is classified as: **Increasing - I**

2nd type of ***observation***
Range Class

Data Sample

Engineering an ***observation***
from QQQ 1-Minute data

Prepare Data

Engineering Observations

Row #

Time

New Column

Range

by minute

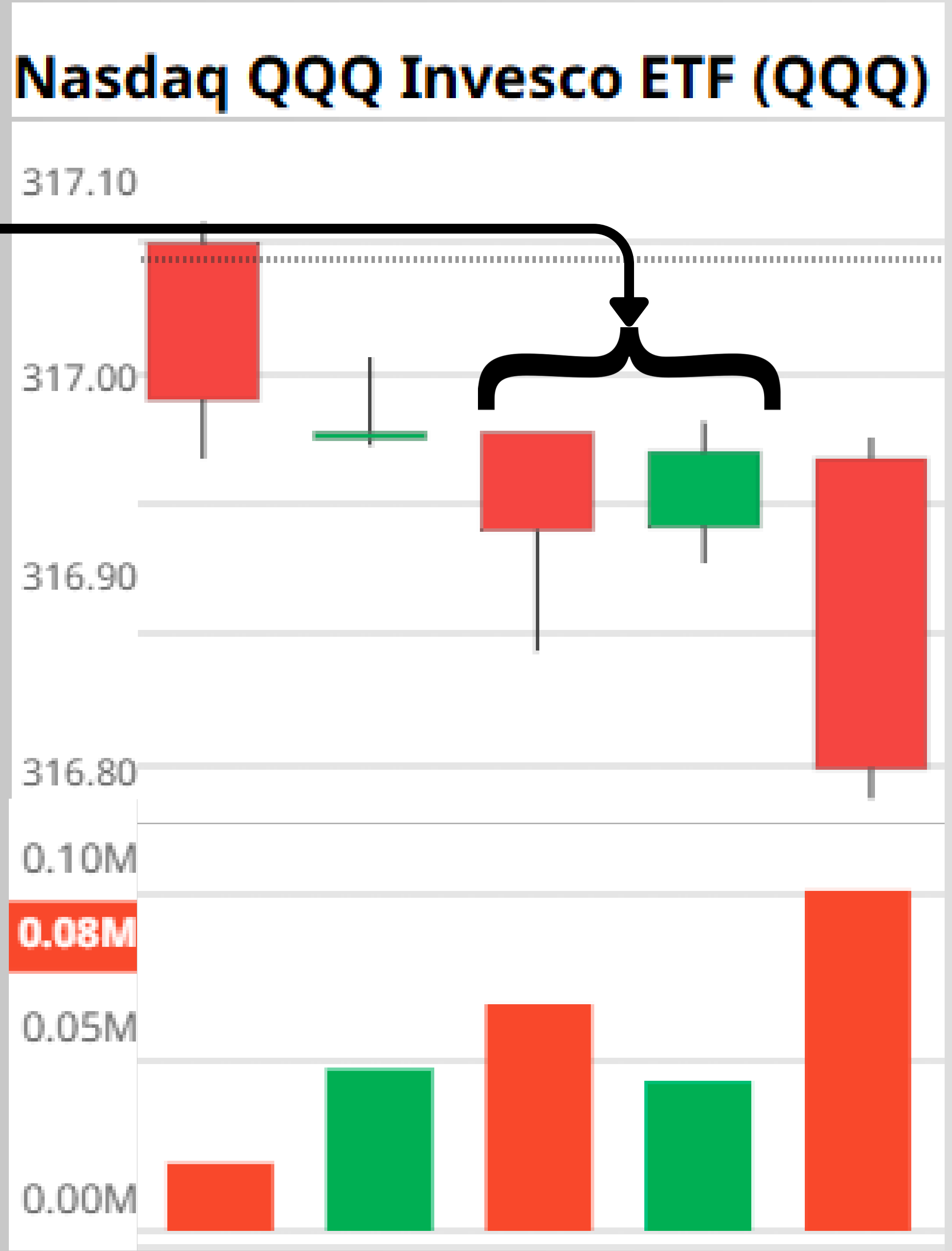
1	timestamp	open	high	low	close	range	volume
112405	2023-04-11 15:20:00	317.09	317.1	317.02	317.07	0.08	57863
112406	2023-04-11 15:21:00	317.07	317.11	317.05	317.0998	0.06	25850
112407	2023-04-11 15:22:00	317.095	317.19	317.09	317.1299	0.10	47240
112408	2023-04-11 15:23:00	317.12	317.21	317.1	317.16	0.11	72582
112409	2023-04-11 15:24:00	317.16	317.17	317.03	317.0404	0.14	70439
112410	2023-04-11 15:25:00	317.05	317.07	316.91	316.914	0.16	70723
112411	2023-04-11 15:26:00	316.9199	316.9616	316.91	316.96	0.0516	27630
112412	2023-04-11 15:27:00	316.96	316.965	316.87	316.886	0.095	58381
112413	2023-04-11 15:28:00	316.89	316.985	316.87	316.93	0.115	47827
112414	2023-04-11 15:29:00	316.94	317.0299	316.93	316.98	0.0999	23636

Which Range Class?

2nd type of ***observation***
Range Class

Prepare Data
Engineering Observations

Contracting Range?



Observing the change in price range between time periods

Which Volume Pattern?

- **Increasing - I**
- **Steady - S**
- **Contracting - C**

2nd type of ***observation***
Range Class

Prepare Data
Engineering Observations

Engineering a Volume Regime Observation
from QQQ 1-Minute *data*

Engineered Feature:
Range Class

Created by setting a threshold for change between classes across a single time period.

<u>**Observation-Classification**</u>	<u>**Change Threshold: 0.04**</u>
Increasing:	> 0.04
Steady:	Between -0.04 & 0.04
Contracting:	< -0.04

Observed Change in Price Range: -0.084

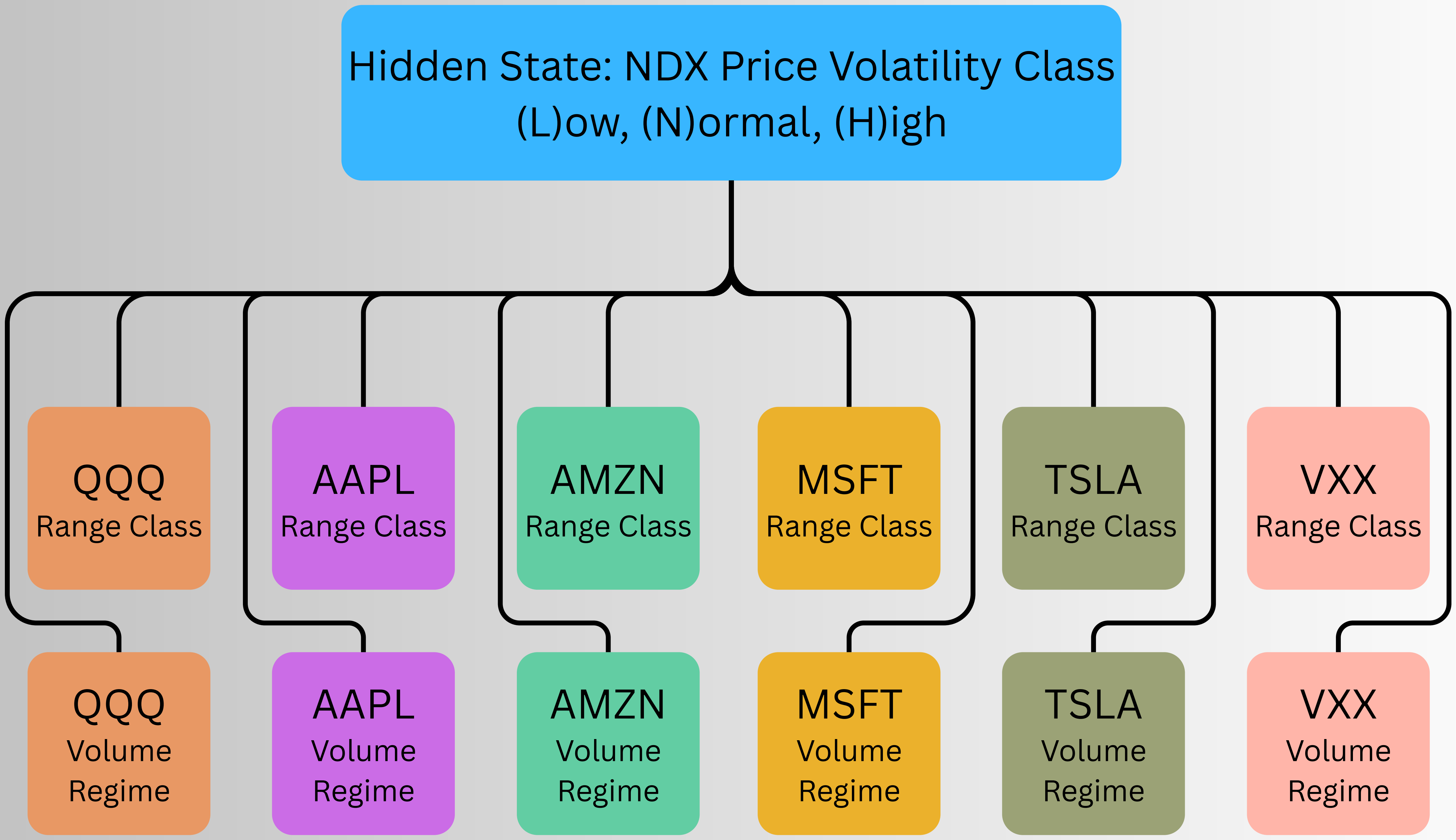
The 1 Min Volume Expansion Regime Between 15:25 and- 15:26 is classified as: **Contracting - C**

Using Multiple Securities to Inform the Hidden State

Model Structure Overview

Observations Types:
Range Expansion Class: 1-min obs
(E)xpanding, (C)ontracting, (S)teady)

Volume Regime: 5-min obs
(I)ncreasing, (D)ecreasing, (S)teady



The model uses 2 observations types from 6 securities to infer hidden volatility state of NDX

Get data ✔

Go to the Github Repo

Run the Code

Examine the Output

darbyatNE/BIN6250-HMM_Feature_Engineering

Contribute to darbyatNE/BIN6250-HMM_Feature_Engineering development by creatin...

GitHub

Prepare Data

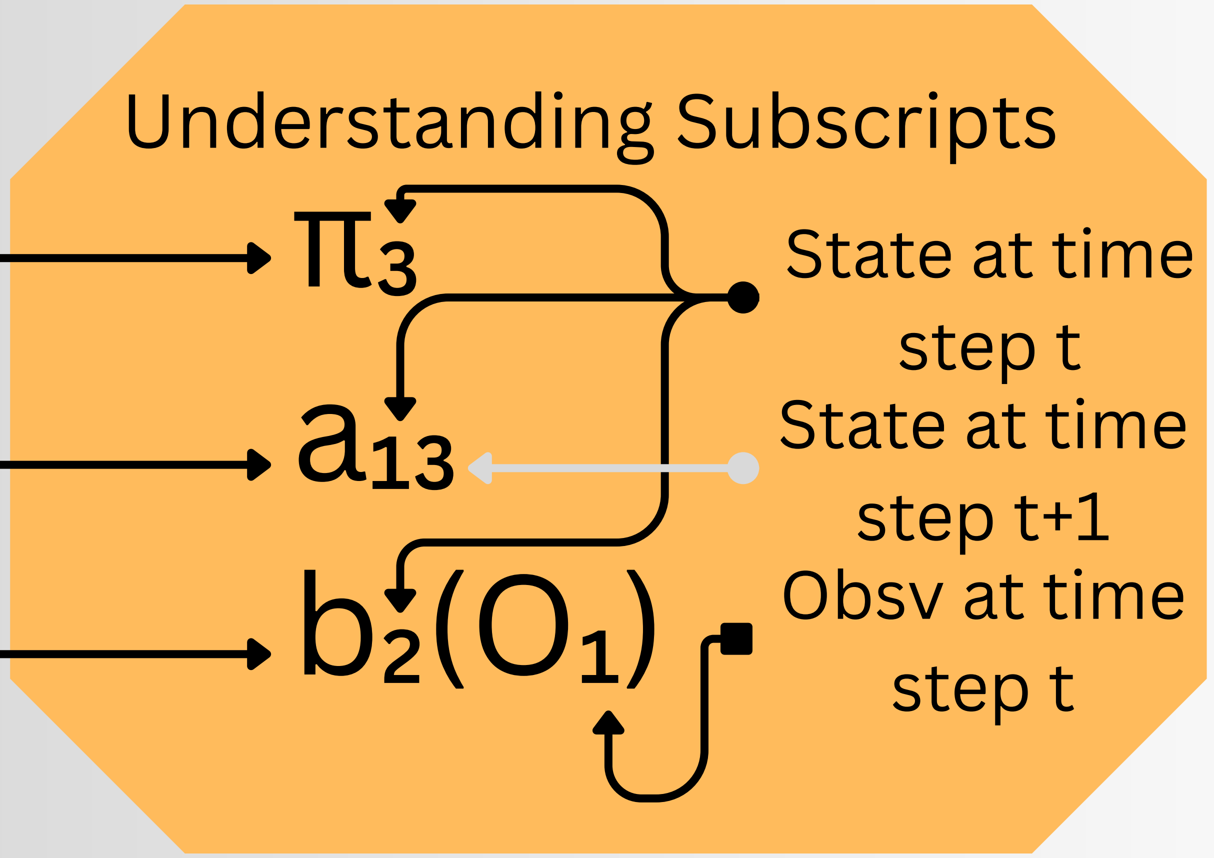
Model Parameters λ

(π, A, B)

- To get the optimal model parameters we will start with a "best guess"
- Based on domain expertise and intuition

3. Model Parameters λ
(π, A, B)

1. Observation Sequence: $[o_2, o_3, o_1, \ldots o_n]$
2. Hidden States: S_1, S_2, S_3
3. Model Parameters λ: (π, A, B)



λ

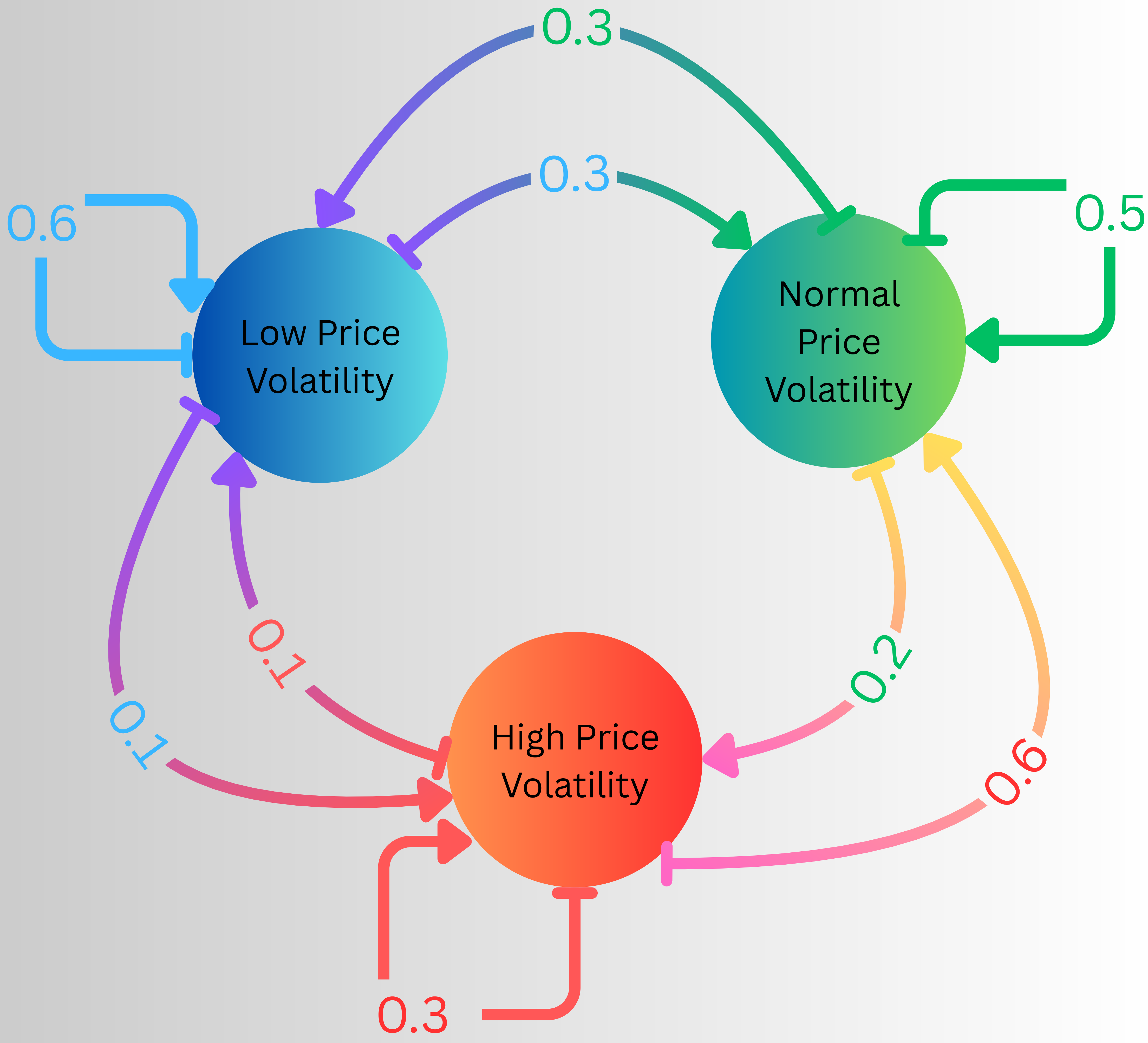
Initial (π):	$\pi_1 = 0.6$ (% start in S_1)	$\pi_2 = 0.3$ (% start in S_2)	$\pi_3 = 0.1$ (% start in S_3)
Transition (A):	$a_{11} = 0.6$ ($S_1 \to S_1$) $a_{12} = 0.3$ ($S_1 \to S_2$) $a_{13} = 0.1$ ($S_1 \to S_3$)	$a_{21} = 0.3$ ($S_2 \to S_1$) $a_{22} = 0.5$ ($S_2 \to S_2$) $a_{23} = 0.2$ ($S_2 \to S_3$)	$a_{31} = 0.1$ ($S_3 \to S_1$) $a_{32} = 0.6$ ($S_3 \to S_2$) $a_{33} = 0.3$ ($S_3 \to S_3$)
Emission (B):	$b_1(o_1) = 0.1$ $b_1(o_2) = 0.4$ $b_1(o_3) = 0.5$	$b_2(o_1) = 0.7$ $b_2(o_2) = 0.2$ $b_2(o_3) = 0.1$	$b_3(o_1) = 0.3$ $b_3(o_2) = 0.3$ $b_3(o_3) = 0.4$

3. Model Parameters: A

Hidden State Transition Probabilities

What is the probability of transition?

“What is the independent chance for each hidden state that between time step, t, and the next step, t+1, the state will either transition to itself or to a new state?”



3. Model Parameters: A

Hidden State Transition Probabilities

What is the total probability of transitioning?

“What are the chances that between time interval, t, and the next ,t+1, the hidden state will either transition to itself or new state?”

Next Hidden State at time step - t+1

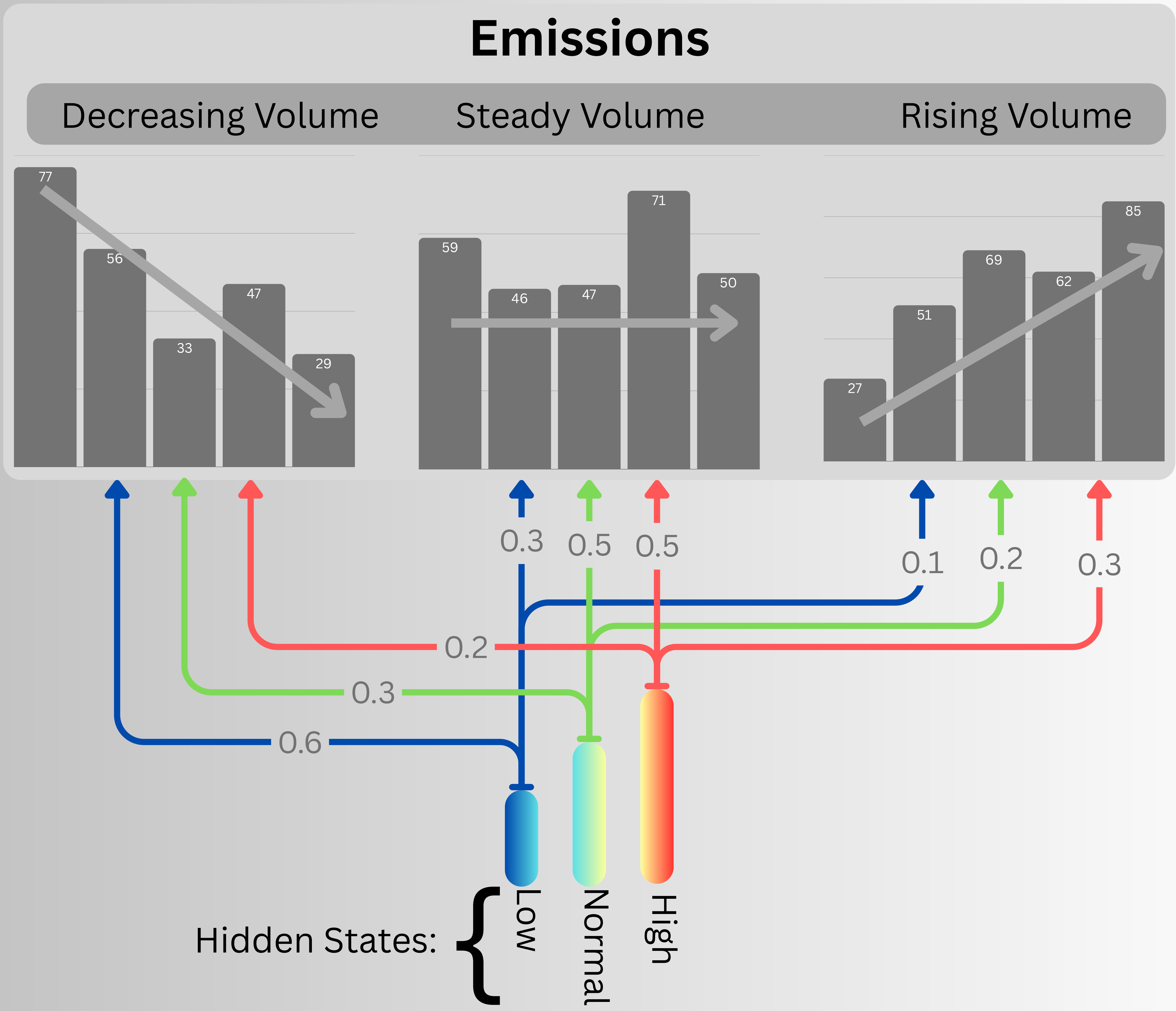
Original Hidden State at Time Step - t

	Low Price Volatility	Normal Price Volatility	High Price Volatility	**Total**
Low Price Volatility	60%	30%	10%	**100%**
Normal Price Volatility	50%	30%	20%	**100%**
High Price Volatility	30%	60%	10%	**100%**

3. Model Parameters: B

Probability of Emission

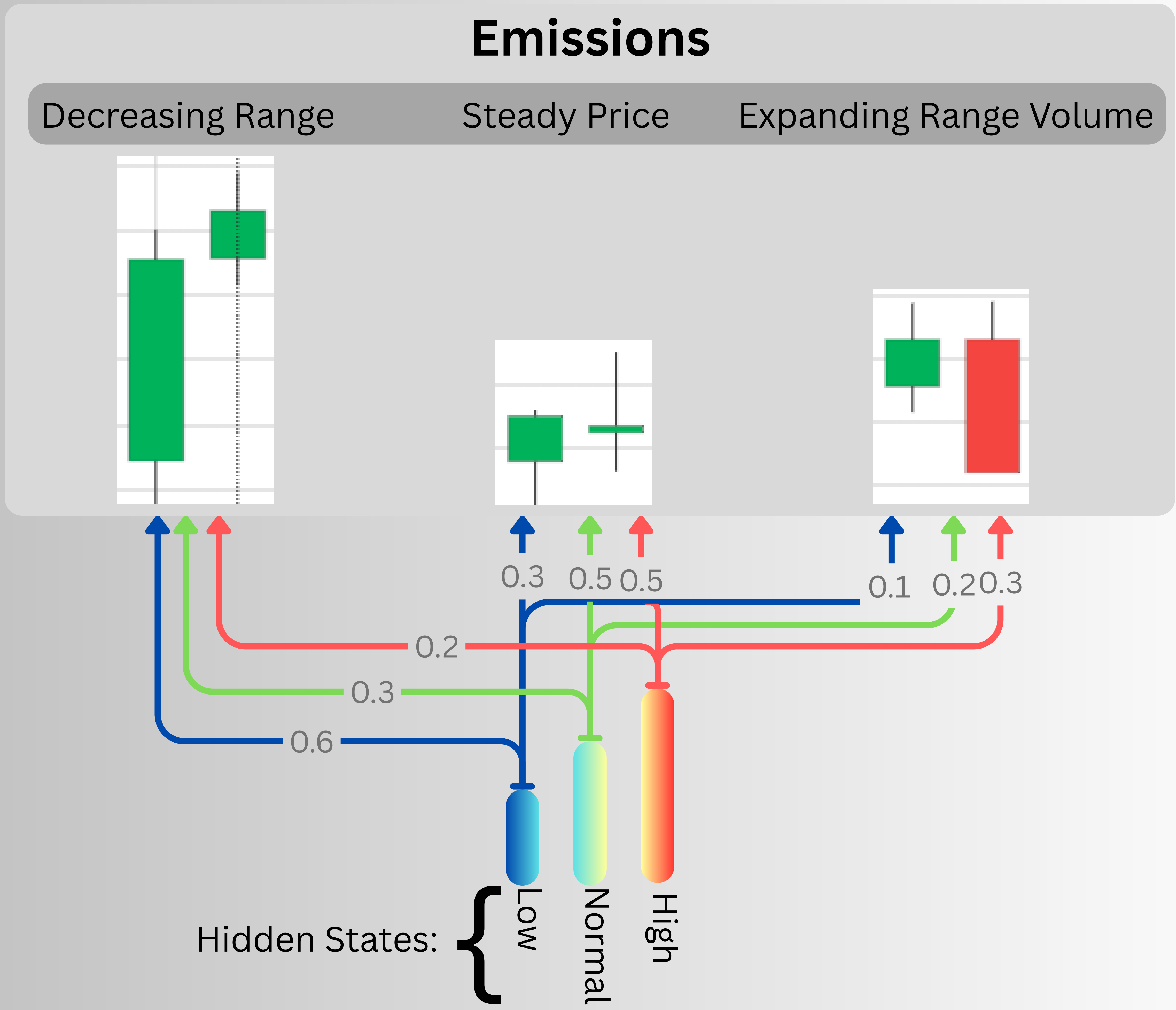
How likely are we to see each 5-Min volume regime?



3. Model Parameters: B

Probability of Emission

How likely are we to see each 1-Min price range class?



Next Steps

Understanding Temporal Alignment

Training vs. Inference

- Complete Historical Data
- Baum Welch for Parameter Estimation
- Training Optimizes Likelihood Using Chosen “Like Days”

Deployment Phase Using Adjusted Probabilities

- Apply Learned Parameters to observable data
- Use Forward Algorithm
- Make Prediction in Real Time
- No Look Ahead Bias

Training Phase

Baum Welch Algorithm

The learning algo for HMM’s

- **Initialization:**
  - Start with a "best guess" of HMM parameters
  - Each of: Transition probabilities, Emission probabilities, and initial state distribution
- **Purpose**:
  - Estimating the model parameters $\lambda = (A, B, \pi)$
  - Maximize the probability of observing the given sequences, locally.
- **Process:**
  - E-step: Using Forward & Backward algorithm to compute expected state occupancies
  - M-step: Re-estimating model parameters to maximize likelihood
    - Like finding your way up a hill in fog

Training Phase

Baum Welch Algorithm

Defining the Steps

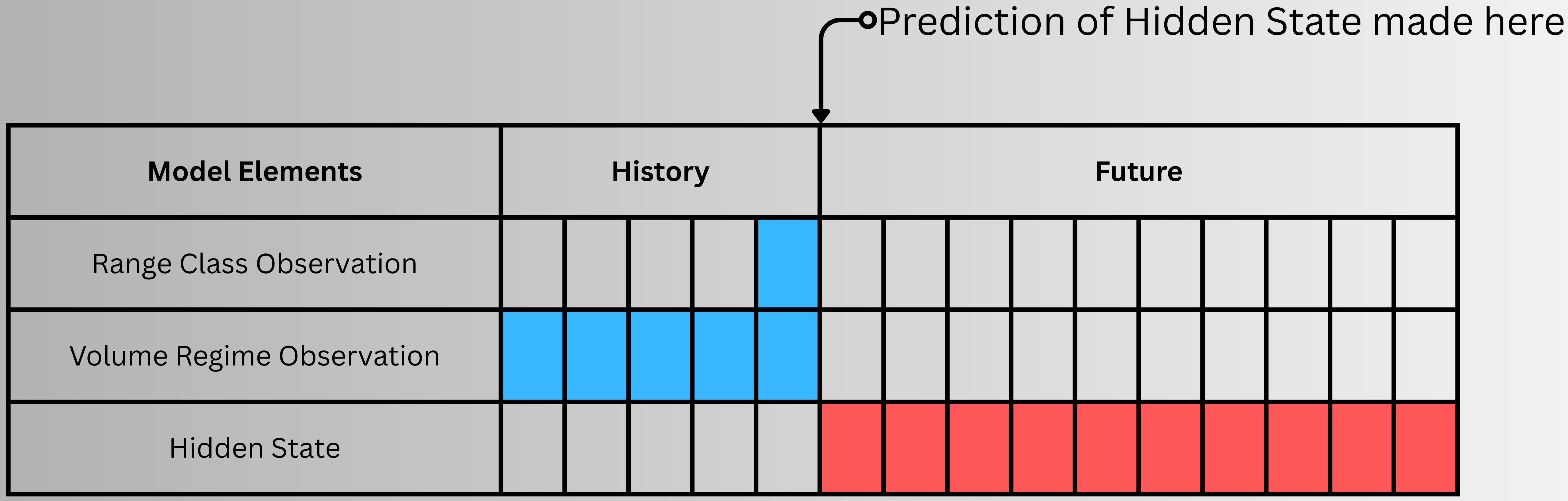
- **Forward Pass:**
  - Calculate the probability of reaching each state at each time step from the beginning of the sequence
- **Backward Pass:**
  - Calculate the probability of completing the sequence from each state at each time step
- **Calculate Expected Counts:**
  - Combine forward and backward probabilities to determine state occupancy and transition probabilities
- **Update Parameters:**
  - Re-estimate transition and emission probabilities based on the expected counts
- **Evaluate Improvement:**
  - Check if the model likelihood has increased, in not the model has converged

Deployment Phase

Inference

Defining the Steps

- Apply Learned Parameters to Observable Data
  - Use Forward Algorithm to make Hidden State Predictions
- Proper Observation Alignment
  - Rolling windows for multi minute observations
  - All observations are aligned to end at the current time point, creating a coherent view of the market state
- No Look Ahead Bias
  - The model should use only data available up to the current minute
  - Observations are strictly from completed time windows



Project Takeaways

Innovative Approach: Encourage creativity in model design and training.

Feature Identification / Creation: Imagine features and hyperparameters to enhance model flexibility.

Hyperparameter Tuner: Showcase flexibility in observations rather than doing things “the old way”.

Goal: Understanding and be able to explain why your model is useful. Not one “right way” to do things

Appendix I

Name of Resource	Purpose	Link
Claude	Answer debugging questions in HMM.ipynb	https://claude.ai/
NVIDIA RAPIDS cuDF Pandas	Researching processing 10min blocks efficiency	https://youtu.be/Ocql-0B4o5c?si=ZvtEM1W6RP1hpsgk
Baum Welch Reestimation	Processing steps of BW algo	https://youtu.be/9Igh_OKECxA?si=M-pp1iAIQ1LCM-Q1
Hidden Markov Models 12: the Baum-Welch algorithm	Explaining expectation step of SW algo	https://www.youtube.com/watch?v=JRsdt05pMoI&t=572s
MPLFinance Library	Creating and coloring a candlestick plot	**https://github.com/matplotlib/mplfinance**
Hidden Markov Model Clearly Explained	Explaining the flow of HMM transitions and emissions	https://www.youtube.com/watch?v=RWkHJnFj5rY&t=102s

Appendix II

Name of Resource	Purpose	Link
BarChart	Generating sample images of a NDX candlestick charts	**https://www.barchart.com/etfs-funds/quotes/QQQ/interactive-chart**
Yahoo Finance	Retrieving historical Quotes of market securities from an API	**https://developer.yahoo.com/api/**
Canva	Image Generation	**canva.com/design**