ML for Finance: Intro to ML

S. Yanki Kalfa

UCSD - Rady SOM

June 13, 2022

Outline

Who am I?

- Yanki Kalfa
- PhD candidate Finance
- MIEF class 2018
- IMF (2 years)
- 6+ years working on quatitative modeling and forecasting

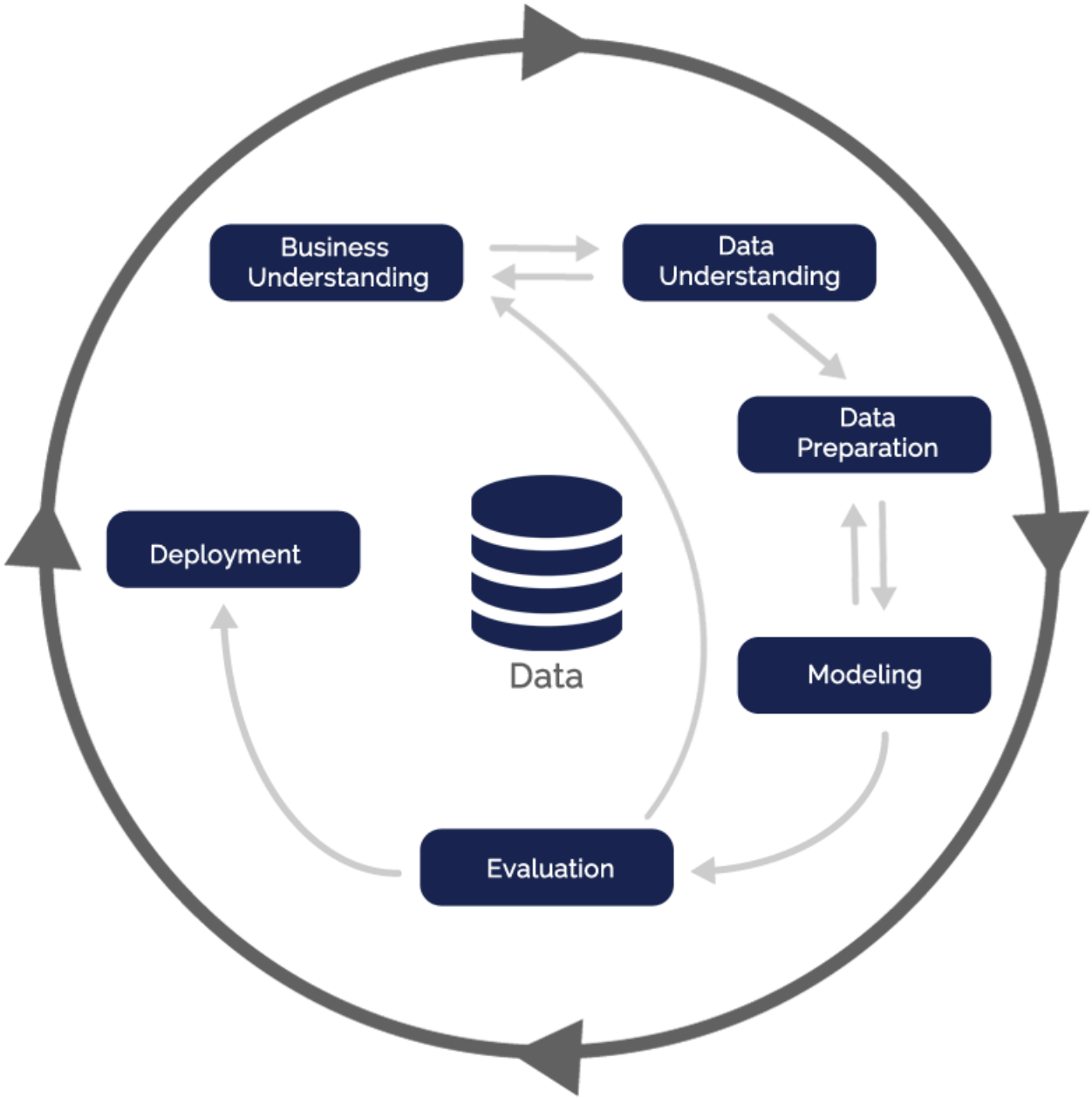
Applications of ML in Finance

- Risk Modeling: logistic regression, discriminant analysis, trees
- Portfolio Management: real-time asset allocation
- Fraud Detection: Reduce false positives
- Client default predictions

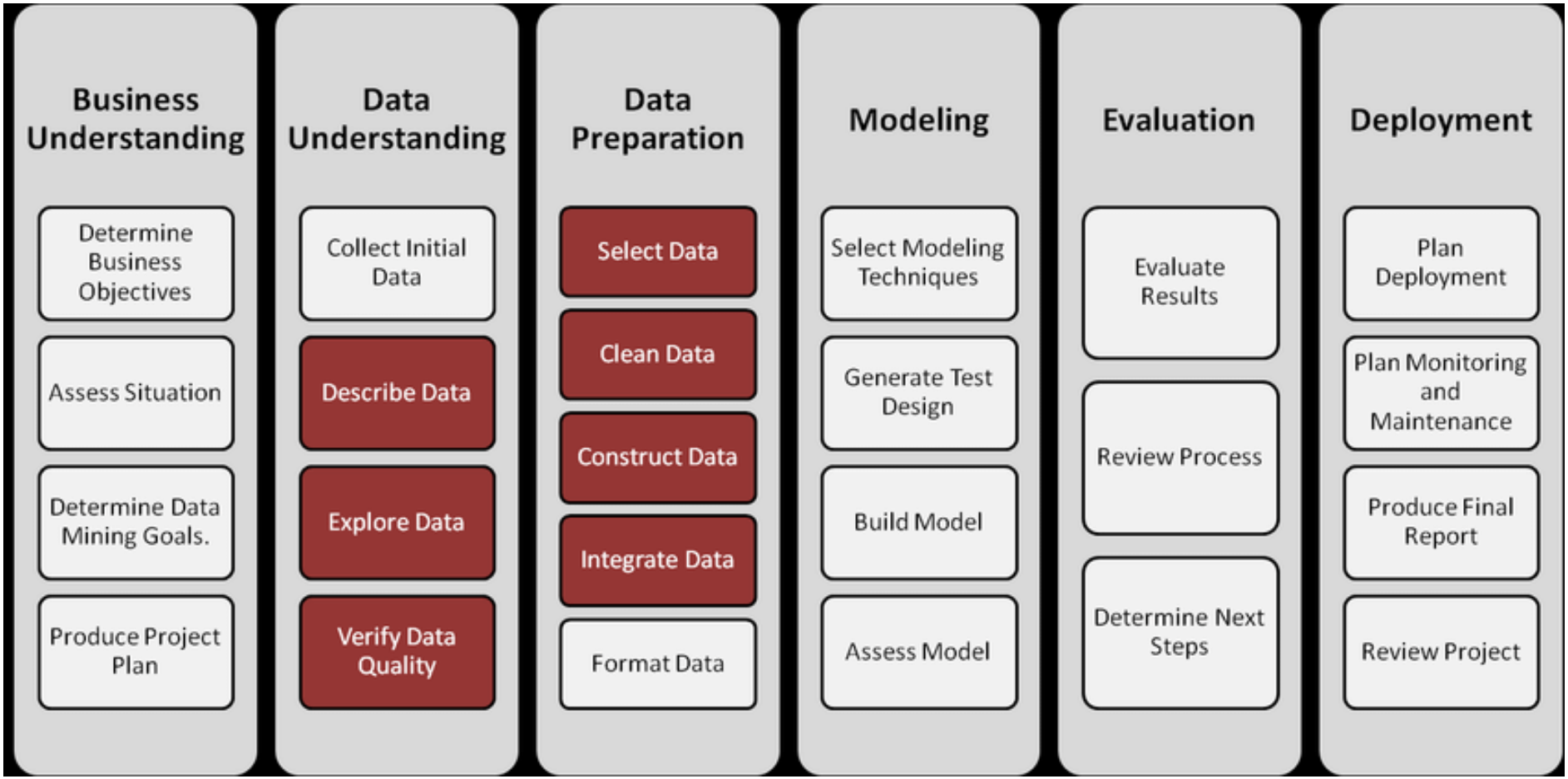
Applications of ML in Finance Cont'd

- Virtual Agents: compliance agents to answer queries from etnerprise personel
- Text Analytics: financial contracts, detect contractual risk
- Unique Identity: fraud detection
- Video Analytics: audit, report automation

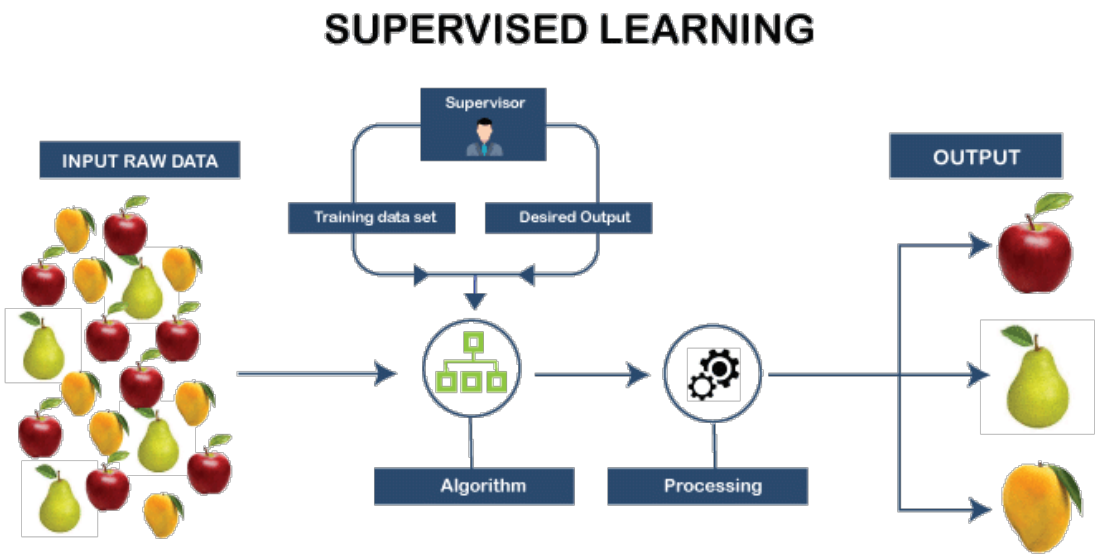
Cross-Indsutry Standard Process for Data Mining



Cross-Indsutry Standard Process for Data Mining



Supervised Learning



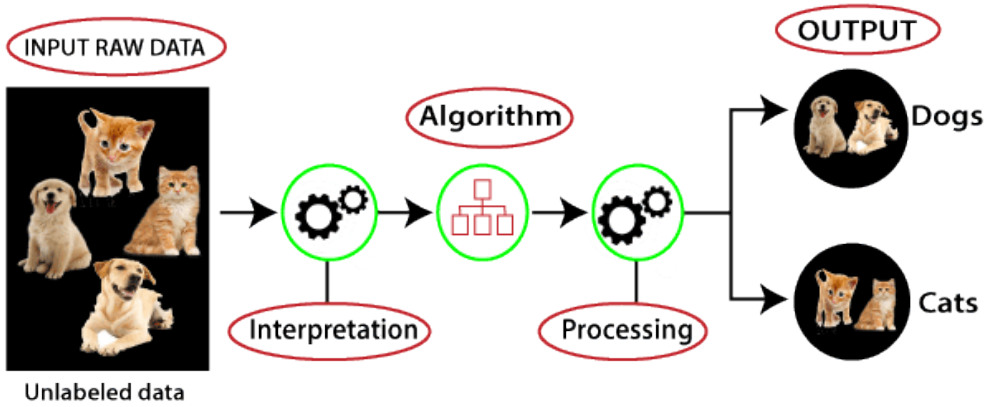
Supervised Learning

- Labeled Data
- Clear Outcome
- Which models
- Train and validate
- Results

Examples:

- Stock prices
- Macro variables

Unsupervised Learning



Unsupervised Learning

- Unlabeled Data
- Goal is to classify data of interest
- Use clustering
- Check if classification is correct
- Repeat

Example:

- Industry classification

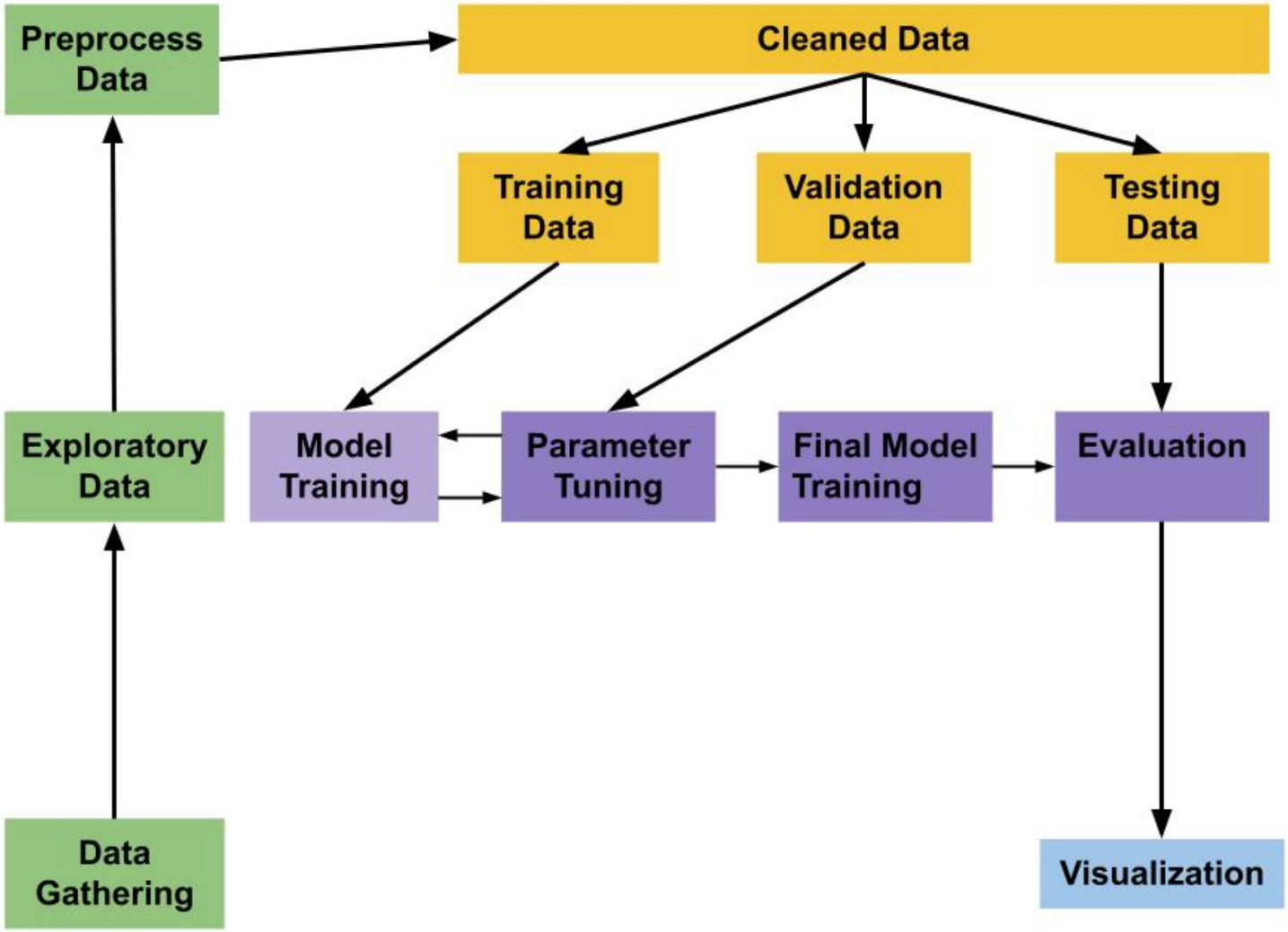
Parametric Models

- Regression
- VAR/SVAR/GARCH
- Assume some type of distribution
- Model conditional mean
- Generalizable
- Usualy efficient with bias

Non-Parametric Models

- Classification and Regression Trees
- Random Forests
- Boosted Trees
- No assumption of distribution
- Less generalizable
- Small bias

General Workflow



Gathering Data

Data can be gathered from many sources

- Databases
- Internet
- Local files

Some Financial and Macro data sources include:

- FRED
- IMF
- OECD
- Yahoo Finance
- On-chain data

Exploratory Analysis

- Autocorrelation
- Missing Data
- Pairwise Correlation

Preprocess Data

- Unit Roots?
- Level vs. Growth
- Dropping Variables

Get Cleaned Data

Training, Validation, and Testing Sets

- Train model
- Tune hyperparameter in validation set
- Iterative Process
- Tuning can be manual
- Cross Validation (K-fold vs. Time Series)
- Find best model

Test Set

- Run the best model in test set
- Evaluate model

Evaluation

- RMSE
- Diebold Mariano (competing models)
- Out of Sample R^2

Visualization

- Easy to communicate
- Easy to see where it performs better
- Cumulative SSE

How can we forecast?

- Heuristics
  - historical average (Prevailing Mean)
- Expert Forecasts
  - IMF Forecasts rely on data and subjective forecasts
  - Survey of Professional Forecasters
  - Bloomberg Forecasts
- Quantitative Forecasts
  - This is our focus

Reality Check

- When generating forecasts we simplify things
  - All models are bad, some are useful
  - We are almost surely misspecifying the relationship
  - Which models to choose
- Learning from the past
  - We rely on past data
  - Some data can be incomplete
  - There are outliers
  - Should we model the levels or growth ?

What can be forecasted

- Probability of an event
  - Recessions
- Time Series
  - Use past data
  - S&P 500
  - Inflation
  - GDP growth

What is the output of the model

- Point Forecasts
  - The output is just a single number that characterizes the conditional mean. We do not know how precise this is.
- Interval Forecasts
  - Give some confidence bands around the foreasted outcome. You can think of it as confidence intervals.
- Density
  - Give a full distribution of the probability of the forecasted object. Normal distribution, Gamma distribution.

Forecast Horizon

- We refer to horizon as h (short for horizon)
- We can forecast longer horizons
  - 3 year GDP growth
- Or medium term outcomes
  - 6 month inflation, quarterly GDP growth
- Or the short term
  - Minutes, Days, Weeks

The definition of horizon depends on what you are forecasting. 3 years for GDP could be medium term, but 3 years for stocks is a long period.Usually longer horizons are harder to predict because of shocks. This does not mean the short term is easy to forecast.

Information Set

In all cases, when we decide to forecast a variable we need to establish an information set. This set will determine the outcome of the model. We label this information set as I_t

- Should we just include the past values of the variable we are forecasting?

$$I_T = \{y_{t-1}, \ldots, y_{t-k}\}$$

- Should we include other variables? We need to know if the additional variables we are helping in forecasting the target.

$$I_T = \{y_{t-1}, \ldots, y_{t-k}, X_{t-1}, \ldots, X_{t-k}\}$$

Methods to generate forecast

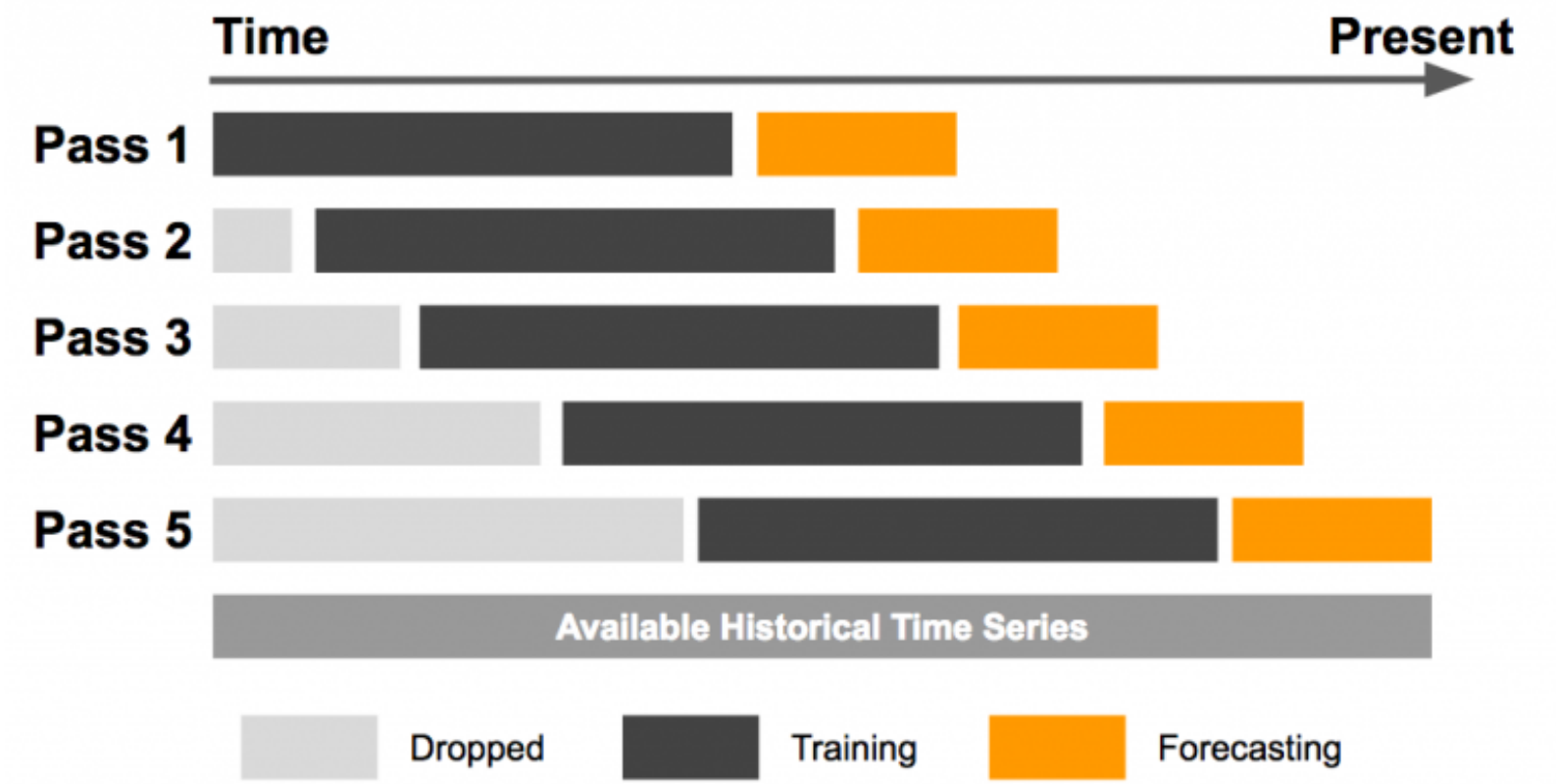
- Expanding Window
- Rolling Window
- FIxed Window

Expanding Window

Expanding Window

- Put equal weight on all observations to estimate the model.
- We add one new observation each time we forecast
- The information set grows as time goes on
- What if there is a break in the parameters?
- It is the most optimal way of forecasting, reduces estimation error

Rolling Window



Rolling Window

- Equal weight the most recent n observations
- We need to pick the length that makes most sense (commonly used in finance)
- It is good for accounting for parameter changes

Fixed Window

- Only uses the first n observations to forecast
- Assumes that parameters do not change
- It only works for stable relatinships
- Easy to work with in practice and analytically

Declining Weights

- If we suspect that the model is not stable. It might be a good idea to put less weight on past observations and more weight on more recent observations.
- We can use discounted least squares for that

$$\omega(s,t) = \begin{cases} \phi^{t-s} & 1 \leq s \leq t \\ 0 & \text{otherwise} \end{cases}$$

$$0 < \phi < 1$$

Recap

- Expanding window: add new observations to estimation sample each time we forecast; information set grows
- Rolling window: set a window length and estimate paramters; information set changes
- fixed window: set inital amount of observations required; information set remains the same.

Intro to ARMA

- Past information of a given variable is often valuable
- Time series are persistent
- Seasonality is an important part of building a model
  - Monthly inflation
  - Quarterly GDP
  - Electricity demand
  - Sales of Goods (December)
- ARMA models are the work horse of the forecasting industry
- Extensively used in many institutions

Intro to ARMA Cont'd

- We only need past information
- No need to fully specify the model
- It has great empirical success, it is usually very hard to beat an ARMA model.
- Often serves as a benchmark

Covariance Stationary Processes

We say that a time series $\{y\}_{t=-\infty}^{\infty}$ is "covariance" stationary if

- The unconditional mean of the process is the same at all times

$$\mathrm{E}[y_t] = \mu_t = \mu \ \forall t$$

- The autocovariance of the series does not depend on t but on the lag order

$$\mathrm{E}[y_t y_{t-j}] = \gamma_j$$

Covariance Stationary Processes

We need to assume some to type of stability when we are building models that depend on past values of the series

- If there are shifts that permanently change the structure of the series, we cannot rely on past data
- A stationary series is free of mean shifts, variance instabilities
- Stationarity allows us to use past values and forecast future values of a variable

White Noise Process

The Gaussian White Noise (WN) is the building block for the ARMA. We can represent the WN as follows:

$$\begin{aligned} \mathrm{E}[\varepsilon_t] &= 0 \\ \mathrm{E}[\varepsilon_t^2] &= \sigma^2 \\ \mathrm{E}[\varepsilon_t \varepsilon_s] &= 0 \text{ for } t \neq s \\ \varepsilon_t &\sim N(0, \sigma^2) \end{aligned}$$

MA(1)

Let $\{\varepsilon_t\}$ be a WN process, and suppose

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

We now calculate the moments of the above process.

$$\begin{aligned}
\mathrm{E}[y_t] &= \mathrm{E}[\mu + \varepsilon_t + \theta\varepsilon_{t-1}] \\
&= \mu + \mathrm{E}[\varepsilon_t] + \theta\mathrm{E}[\varepsilon_{t-1}] = \mu \\
\mathrm{E}[y_t - \mu]^2 &= \mathrm{E}[\varepsilon_t + \theta\varepsilon_{t-1}]^2 \\
&= \mathrm{E}[\varepsilon_t^2 + 2\theta\varepsilon_t\varepsilon_{t-1} + \theta^2\varepsilon_{t-1}^2] \\
&= \sigma^2 + 0 + \theta^2\sigma^2 \\
&= (1+\theta)\sigma^2 \equiv \gamma_0 \\
\mathrm{E}(y_t - \mu)(y_{t-1} - \mu) &= \mathrm{E}\left(\varepsilon_t + \theta\varepsilon_{t-1}\right)\left(\varepsilon_{t-1} + \theta\varepsilon_{t-2}\right) \\
&= \mathrm{E}\left(\varepsilon_t\varepsilon_{t-1} + \theta\varepsilon_t\varepsilon_{t-2} + \theta\varepsilon_{t-1}^2 + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}\right) \\
&= \theta\sigma^2 \equiv \gamma_1
\end{aligned}$$

MA(1)

It is clear that the moments of the MA(1) all exists and do not depend on time. Hence, it is a stationary process. Let's look at the autocorrelation function.

$$
\begin{aligned}
Corr(y_t, y_{t-j}) &= \frac{\mathrm{E}(y_t y_{t-j})}{\sqrt{\mathrm{E}(y_t-\mu)}\sqrt{\mathrm{E}(y_t-\mu)}} \\
&= \frac{\gamma_j}{\gamma_0} \equiv \rho_j \\
\rho_1 &= \frac{\theta\sigma^2}{(1+\theta)\sigma^2} = \frac{\theta}{1+\theta^2}
\end{aligned}
$$

MA(q)

An MA process can have more than 1 lags.

$$
\begin{aligned}
y_t &= \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \\
\mathrm{E}(y_t) &= \mu + \mathrm{E}(\varepsilon_t) + \theta_1 \mathrm{E}(\varepsilon_{t-1}) + \ldots + \theta_q \mathrm{E}(\varepsilon_{t-q}) = \mu \\
\mathrm{E}(y_t - \mu)^2 &= \mathrm{E}(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q})^2 \\
&\because \mathrm{E}(\varepsilon_t \varepsilon_{t-j}) = 0, \ \forall j > 0 \\
\therefore \gamma_0 &= (1 + \theta_1^2 + \ldots + \theta_2^2)\sigma^2 \\
\text{For } j &= 1, \ldots, q \\
\gamma_j &= \mathrm{E}[\theta_j \varepsilon_{t-j}^2 + \theta_{j+1}\theta_1 \varepsilon_{t-j-1}^2 + \ldots + \theta_q \theta_{q-j} \varepsilon_{t-q}^2] \\
\gamma_j &= \begin{cases} [\theta_j + \theta_{j+1}\theta_j + \ldots + \theta_q \theta_{q-j}] & \text{for } j = 1, \ldots, q \\ 0 & \text{for } j > q \end{cases}
\end{aligned}
$$

MA(2) Example

$$
\begin{aligned}
y_t &= \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \\
\mathrm{E}(y_t) &= \mu \\
\gamma_0 &= (1 + \theta_1^2 + \theta_2^2)\sigma^2 \\
\gamma_1 &= (\theta_1 + \theta_1 \theta_2)\sigma^2 \\
\gamma_2 &= \theta_2 \sigma^2 \\
\gamma_j &= 0 \ \forall j > 2
\end{aligned}
$$

AR(1)

The Autoregressive model uses the past values of variable to forecast the future values. Let's start with an AR(1) model.

$$y_t = c + \phi y_{t-1} + \varepsilon_t$$

By recursive substitution we have:

$$y_t = (c + \varepsilon_t) + \phi(c + \varepsilon_{t-1}) + \phi^2(c + \varepsilon_{t-2}) + \dots$$

$$y_t = \frac{c}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

The above holds if $|\phi| < 1$. For now, we will assume that is the case.

AR(1)

We showed that an AR(1) has an equivalent MA(∞) representation. Let's calculate some moments.

$$\mathrm{E}(y_t) = \mathrm{E}\left(\frac{c}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\right) = \frac{c}{1-\phi} = \mu$$

$$\begin{aligned}
\gamma_0 = \mathrm{E}(y_t - \mu)^2 &= \mathrm{E}\left(\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\right)^2 \\
&= (1 + \phi^2 + \phi^4 + \ldots)\sigma^2 \\
&= \frac{\sigma^2}{1-\phi^2} \\
\gamma_j = \mathrm{E}(y_t - \mu)(y_{t-j} - \mu) &= (\phi^j + \phi^{j+2} + \phi^{j+4} + \ldots)\sigma^2 \\
&= \phi^j(1 + \phi^2 + \phi^4 + \ldots)\sigma^2 \\
&= \frac{\phi^j}{1-\phi^2}\sigma^2
\end{aligned}$$

ARMA(p,q)

The AR and MA processes are special cases of the ARMA(p,q) process. The general ARMA process can be written as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

An AR(1) process is an ARMA(1,0), and an MA(1) process is an ARMA(0,1). Let's introduce some life saving notation:Lag operators

$$L y_t = y_{t-1}$$
$$L^p y_t = y_{t-p}$$
$$\phi(L) = \sum_{i=0}^{p} \phi_i L^i$$

ARMA(p,q)

So we can re-write an AR(1) as:

$$y_t = \phi y_{t-1} + \varepsilon_t$$
$$(1 - \phi(L))y_t = \varepsilon_t$$

So we can write the ARMA process as

$$y_t = \phi(L)^{-1}\theta(L)\varepsilon_t$$

Non-Stationarity

Up until now we assumed that our models were stationary Specifically in the case of the AR(1) process we assumed that $|\phi| < 1$. What if that is not the case?

- Suppose that one of the root of the polynomial $\phi(L)$ is one. Then we call this process a unit root process or an integrated process.

Example AR(1);

$$(1 - z\phi) = 0$$
$$\therefore \phi = 1$$

We solve this issue by differencing

$$\omega(L)(1 - L) = \phi(L)$$

Non-Stationarity ARMA to Stationary

Define $\Delta y_t = (1 - L)y_t$, and rewrite ARMA as:

$$\omega(L)(1 - L)y_t = \theta(L)\varepsilon_t$$
$$\omega(L)\Delta y_t = \theta(L)\varepsilon_t$$

The roots are all outside of the unit circle, Δy_t is stationary. This is called an I(1) process. It requires us to difference the series once.

Stationarity is an important topic, if a variable is not stationary it is really hard to predict. This is because we rely on the autocorrelation of the stationary process to forecast future values.

Forecating with AR

$$y_t = \alpha + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

- Assume $\varepsilon \sim WN(0, 1)$

$$y_{t|t-1} = \alpha + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}$$

$y_{t|t-1}$ means forecast of y_t given information at $t-1$.

The Chain Rule

Forecasting more than one step ahead can be easily done. We replace the unknown values of y_{T+h} with the forecast.

$$y_{T+1|T} = \phi_1 y_T + \ldots + \phi_p y_{T-p+1}$$
$$y_{T+2|T} = \phi_1 y_{T+1|T} + \phi_2 y_T + \ldots + \phi_p y_{t-p+2}$$
$$y_{T+p+1|T} = \phi_1 y_{T+p|T} + \phi_2 y_{T+p-1|T} + \ldots + \phi_p y_{T+1|T}$$

AR(1) Example

$$
\begin{aligned}
y_{T+1} &= \alpha + \phi y_T + \varepsilon_{T+1} \\
y_{T+1|T} &= \alpha + \phi y_T \\
y_{T+2|T} &= \alpha + \phi y_{T+1|T} \\
&= \alpha + \phi(\alpha + \phi y_T) \\
&= \alpha(1 + \phi) + \phi^2 y_T \\
y_{T+3|T} &= \alpha + \phi y_{T+2|T} \\
&= \alpha + \phi(\alpha + \phi(\alpha + \phi y_T)) \\
&= \alpha \sum_{i=0}^{2} \phi^i + \phi^3 y_T
\end{aligned}
$$

AR(1) Example

Can we forecast up to infinity with an AR(1)

$$\lim_{h\to\infty} y_{T+h} = \lim_{h\to\infty} \alpha \sum_{i=0}^{h} \phi^i + \phi^h y_T + \sum_{i=0}^{h} \phi^i \varepsilon_{T+h-i}$$

$$\lim_{h\to\infty} y_{T+h|T} = \lim_{h\to\infty} \alpha \sum_{i=0}^{h} \phi^i + \phi^h y_T$$

$$= \frac{\alpha}{1-\phi}$$

This only works for stationary AR process. Notice that the AR(1) forecasts converge to the mean of the process.

AR(1) Forecast Errrors

$$
\begin{aligned}
y_{T+1} &= \alpha + \phi y_T + \varepsilon_{T+1} \\
y_{T+1|T} &= \alpha + \phi y_T \\
\mathrm{E}[(y_{T+1} - y_{T+1|T})^2] &= \mathrm{E}[(\varepsilon_{T+1})^2] = 1
\end{aligned}
$$

How about 2 step-ahead forecasts

$$
\begin{aligned}
y_{T+2} &= \alpha + \phi y_{T+1} + \varepsilon_{T+2} \\
&= \alpha(1+\phi) + \phi^2 y_T + \varepsilon_{T+2} + \phi\varepsilon_{T+1} \\
y_{T+2|T} &= \alpha + \phi y_{T+1|T} \\
&= \alpha(1+\phi) + \phi^2 y_T \\
\mathrm{E}[(y_{T+2} - y_{T+2|T})^2] &= \mathrm{E}[(\phi(y_{T+1} - y_{T+1|y_T}) + \varepsilon_{T+2})^2] \\
&= 1 + \phi^2
\end{aligned}
$$

AR(1) Forecast Errrors

How about the forecast errors of $h \to \infty$?

$$\mathrm{E}[(\lim_{h\to\infty} y_{T+h} - \lim_{h\to\infty} y_{T+h|T})^2] = 1 + \phi^2 + \phi^4 + \dots$$
$$= \frac{1}{1-\phi^2}$$

So the AR forecasts are well defined. The infinity forecast converges to the mean of the AR process and the forecast error converges to the variance of the process. This is not the case for the Random Walk ($\phi = 1$).

Forecasting with MA

Forecasting with an MA process is less cumbersome, but also less informative. Consider the MA(1) case:

$$y_{T+1} = \varepsilon_{T+1} + \theta\varepsilon_T$$

- ε_{T+1} is not in the information set I_T
- We cannot observe the shocks directly but can be estimated

MA(1) example

$$
\begin{aligned}
y_{T+1} &= \varepsilon_{T+1} + \theta\varepsilon_T \\
y_{T+1|T} &= \theta\varepsilon_T \\
y_{T+2} &= \varepsilon_{T+2} + \theta\varepsilon_{T+1} \\
y_{T+2|T} &= 0
\end{aligned}
$$

How about forecast errors

$$
\begin{aligned}
\mathrm{E}[(y_{T+1} - y_{T+1|T})^2] &= \mathrm{E}[(\varepsilon_{T+1})^2] = 1 \\
\mathrm{E}[(y_{T+2} - y_{T+2|T})^2] &= \mathrm{E}[(\varepsilon_{T+2} + \theta\varepsilon_{T+1})^2] = 1 + \theta^2
\end{aligned}
$$

Forecasting with ARMA

ARMA models combine the AR and MA processes.

$$y_{T+1} = \alpha + \phi_1 y_T + \phi_2 y_{T-1} + \phi_p y_{T-p+1} + \ldots + \varepsilon_{T+1} + \theta_1 \varepsilon_T + \ldots + \theta_q \varepsilon_{T-q+1}$$

Simple strategy, separate the AR and MA components and recursively forecast ahead. Replace future values unkown values with forecasted values.

ARMA(1,1) Example

$$
\begin{aligned}
y_{T+1} &= \alpha + \phi y_T + \varepsilon_{T+1} + \theta \varepsilon_T \\
y_{T+1|T} &= \alpha + \phi y_T + \theta \varepsilon_T \\
\mathrm{E}[(y_{T+1} - y_{T+1|T})^2] &= \mathrm{E}[\varepsilon_{T+1}^2] = 1 \\
y_{T+2} &= \alpha + \phi y_{T+1} + \varepsilon_{T+2} + \theta \varepsilon_{T+1} \\
y_{T+2|T} &= \alpha + \phi y_{T+1|T} \\
&= \alpha + \phi(\alpha + \phi y_T + \theta \varepsilon_T) \\
\mathrm{E}[(y_{T+2} - y_{T+1|T})^2] &= \mathrm{E}[(\phi(y_{T+1} - y_{T+1|T}) + \varepsilon_{T+2} + \theta \varepsilon_{T+1})^2] \\
&= \mathrm{E}[((\phi + \theta)\varepsilon_{T+1} + \varepsilon_{T+2})^2] \\
&= 1 + (\phi + \theta)^2
\end{aligned}
$$

Direct vs. Indirect Forecasting

- Indirect Forecasting:
  - Use chain rule to forecasting multiple steps ahead
  - If correctly specified this is more efficient
- Direct Forecasting
  - Match the lag on the RHS to horizon forecasted
  - Misspecification robust

Lag Length

- How to pick the lag order of the ARMA(p,q) model?
- Rely on judgement: ACF, PACF
- Information Criterion: AIC, BIC

$$IC_k = ln \underbrace{\hat{\sigma}_k^2}_{\text{SSE of model k}} + \underbrace{n_k}_{\text{Number of parameters } n_k=p_k+q_k+1} \times \underbrace{p(T)}_{\text{Penalty term function of sample size}}$$

For AIC

$$p(T) = 2T^{-1}$$

For BIC

$$p(T) = \frac{lnT}{T}$$