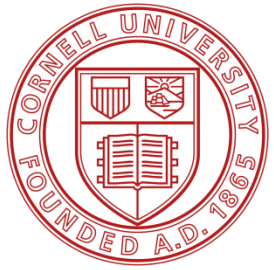


Data Analysis

Marcos López de Prado, Ph.D.

Advances in Financial Machine Learning

ORIE 5256

What are we going to learn today?

- Financial Data Structures: **The X Matrix**
  - Sample asynchronous data
  - Form datasets amenable for ML algorithms
- Labeling: **The Y Vector**
  - Predict something that is worth predicting
- Sample Weights: **The W Vector**
  - Not all observations are equally important
- Fractionally Differentiated Features
  - Obtain stationary features with minimum memory loss

Financial Data Structures

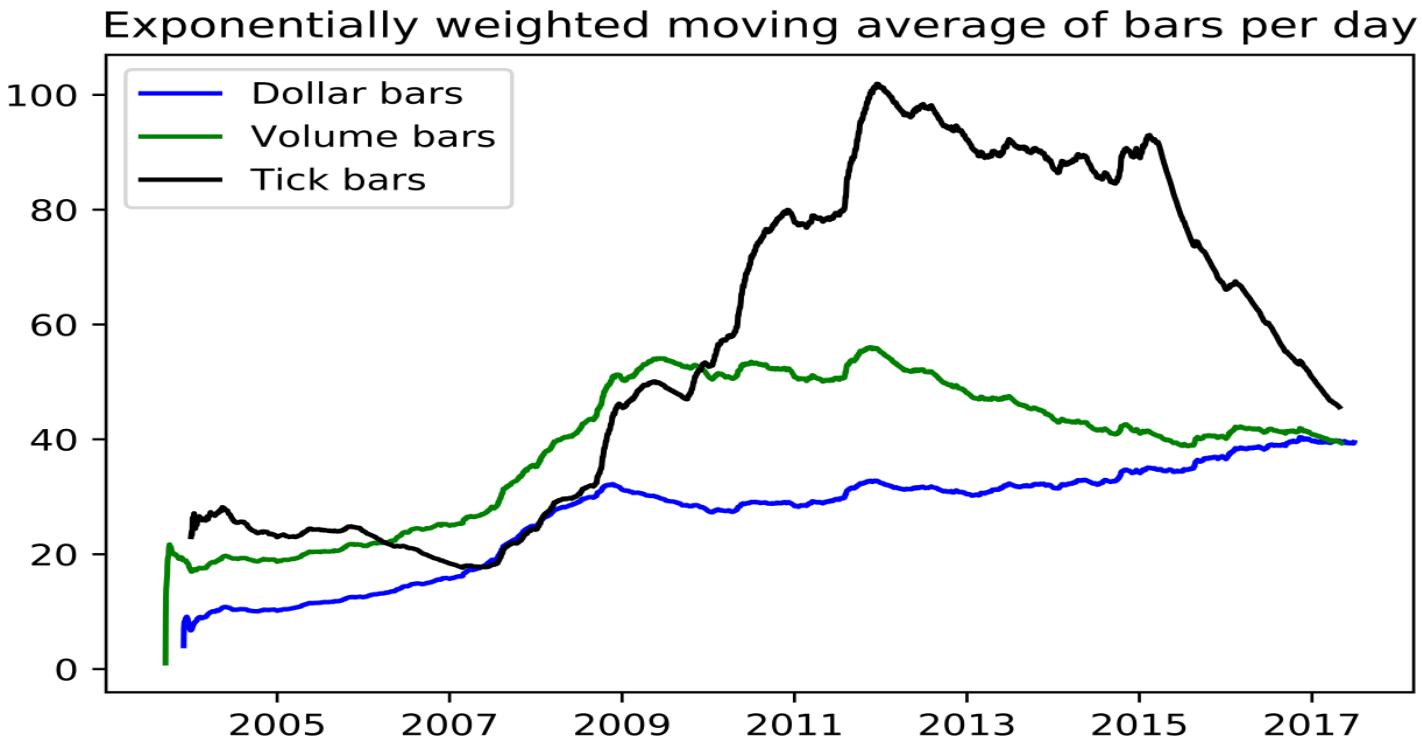
Essential Types of Financial Data

Fundamental Data	Market Data	Analytics	Alternative Data
• Assets • Liabilities • Sales • Costs/earnings • Macro variables • . . .	• Price/yield/implied volatility • Volume • Dividend/coupons • Open interest • Quotes/cancellations • Aggressor side • . . .	• Analyst recommendations • Credit ratings • Earnings expectations • News sentiment • . . .	• Satellite/CCTV images • Google searches • Twitter/chats • Metadata • . . .

Forming Bars

- Information does not arrive to the market at a constant entropy rate.
- Sampling data in chronological intervals means that the informational content of the individual observations is far from constant.
- A better approach is to sample observations as a subordinated process of the amount of information exchanged:
  - Trade bars.
  - Volume bars.
  - Dollar bars.
  - Volatility or runs bars.
  - Order imbalance bars.
  - Entropy bars.

Example 1: Sampling Frequencies



Three bar types computed on E-mini S&P 500 futures.

Tick bars tend to exhibit a wide range of sampling frequencies, for multiple microstructural reasons.

Sampling frequencies for **volume bars** are often inversely proportional to price levels.

In general, **dollar bars** tend to exhibit more stable sampling frequencies.

Example 2: Dollar Imbalance Bars (1/2)

- Let's define the imbalance at time T as $\theta_T = \sum_{t=1}^{T} b_t v_t$, where $b_t \in \{-1,1\}$ is the aggressor flag, and v_t may represent either the number of securities traded or the dollar amount exchanged.
- We compute the expected value of θ_T at the beginning of the bar

$$\begin{aligned}\mathrm{E}_0[\theta_T] &= \mathrm{E}_0\left[\sum_{t|b_t=1} v_t\right] - \mathrm{E}_0\left[\sum_{t|b_t=-1} v_t\right] \\ &= \mathrm{E}_0[T](\mathrm{P}[b_t = 1]\mathrm{E}_0[v_t|b_t = 1] - \mathrm{P}[b_t = -1]\mathrm{E}_0[v_t|b_t = -1])\end{aligned}$$

- Let's denote $v^+ = \mathrm{P}[b_t = 1]\mathrm{E}_0[v_t|b_t = 1]$, $v^- = \mathrm{P}[b_t = -1]\mathrm{E}_0[v_t|b_t = -1]$, so that $\mathrm{E}_0[T]^{-1}E_0[\sum_t v_t] = \mathrm{E}_0[v_t] = v^+ + v^-$. You can think of v^+ and v^- as decomposing the initial expectation of v_t into the component contributed by buys and the component contributed by sells.

Example 2: Dollar Imbalance Bars (2/2)

- Then, $\mathrm{E}_0[\theta_T] = \mathrm{E}_0[T](v^+ - v^-) = \mathrm{E}_0[T](2v^+ - \mathrm{E}_0[v_t])$
- In practice, we can estimate $\mathrm{E}_0[T]$ as an exponentially weighted moving average of *T* values from prior bars, and $(2v^+ - \mathrm{E}_0[v_t])$ as an exponentially weighted moving average of $b_t v_t$ values from prior bars.
- We define a bar as a T^*-contiguous subset of ticks such that the following condition is met

$$T^* = \arg\min_T \{|\theta_T| \geq \mathrm{E}_0[T]|2v^+ - \mathrm{E}_0[v_t]|\}$$

where the size of the expected imbalance is implied by $|2v^+ - \mathrm{E}_0[v_t]|$.

- When θ_T is more imbalanced than expected, a low *T* will satisfy these conditions.

Multi-Product Series

```
def getRolledSeries(pathIn,key):
    series=pd.read_hdf(pathIn,key='bars/ES_10k')
    series['Time']=pd.to_datetime(series['Time'],format='%Y%m%d%H%M%S%f')
    series=series.set_index('Time')
    gaps=rollGaps(series)
    for fld in ['Close','VWAP']:series[fld]-=gaps
    return series
#———————————————————————————————————————————
def rollGaps(series,dictio={'Instrument':'FUT_CUR_GEN_TICKER','Open':'PX_OPEN', \
    'Close':'PX_LAST'},matchEnd=True):
    # Compute gaps at each roll, between previous close and next open
    rollDates=series[dictio['Instrument']].drop_duplicates(keep='first').index
    gaps=series[dictio['Close']]*0
    iloc=list(series.index)
    iloc=[iloc.index(i)-1 for i in rollDates] # index of days prior to roll
    gaps.loc[rollDates[1:]]=series[dictio['Open']].loc[rollDates[1:]]- \
        series[dictio['Close']].iloc[iloc[1:]].values
    gaps=gaps.cumsum()
    if matchEnd:gaps-=gaps.iloc[-1] # roll backward
    return gaps
```

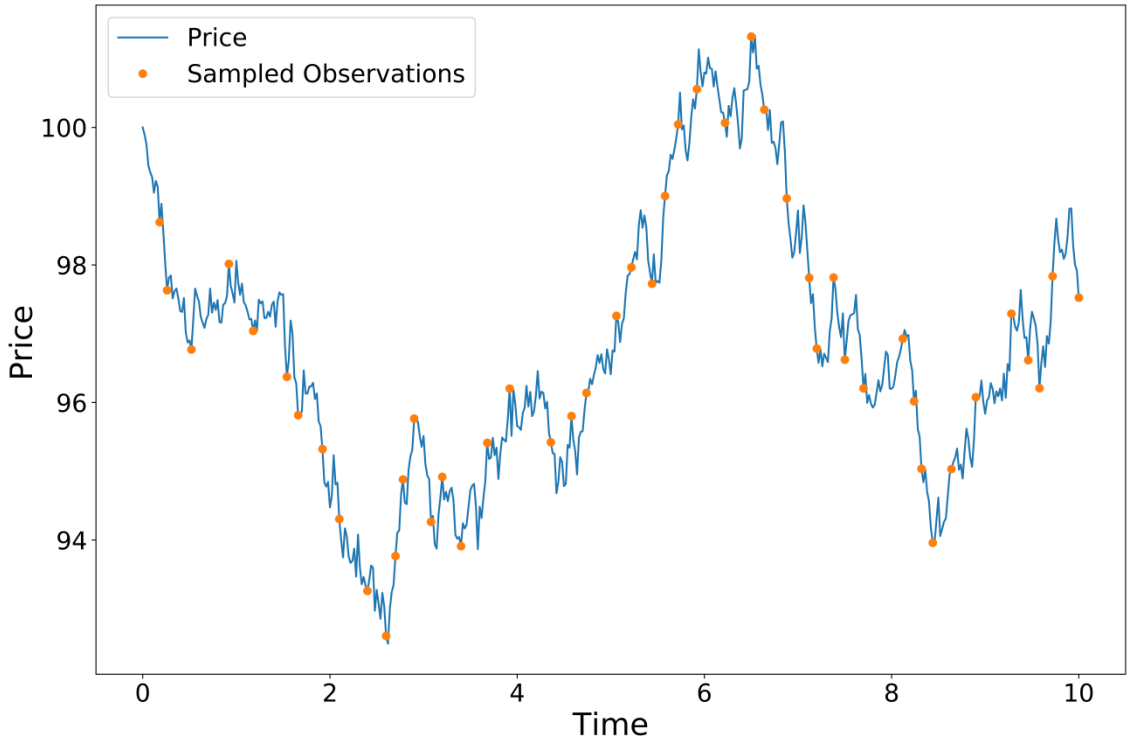
(*) For rolling baskets of futures & options, see also Section 2.4.1 (the “ETF Trick”)

Sampling Features

- We may want to sample the features at irregular frequencies, to train the algorithm to predict the outcomes of specific events.
- In that case, form a matrix X with as many rows as possible, which we can then downsample to the subset of rows (events) that are predictable by the selected features.

```
def getTEvents(gRaw,h):
    tEvents,sPos,sNeg=[],0,0
    diff=gRaw.diff()
    for i in diff.index[1:]:
        sPos,sNeg=max(0,sPos+diff.loc[i]),min(0,sNeg+diff.loc[i])
        if sNeg<-h:
            sNeg=0;tEvents.append(i)
        elif sPos>h:
            sPos=0;tEvents.append(i)
    return pd.DatetimeIndex(tEvents)
```

Example of CUSUM filter



Labeling

Labeling in Finance

- Virtually all ML papers in finance label observations using the fixed-time horizon method.
- Consider a set of features $\{X_i\}_{i=1,\dots,I}$, drawn from some bars with index $t = 1, \dots, T$, where $I \leq T$. An observation X_i is assigned a label $y_i \in \{-1,0,1\}$,

$$y_i = \begin{cases} -1 & \text{if } r_{t_{i,0},t_{i,0}+h} < -\tau \\ 0 & \text{if } \left|r_{t_{i,0},t_{i,0}+h}\right| \leq \tau \\ 1 & \text{if } r_{t_{i,0},t_{i,0}+h} > \tau \end{cases}$$

where τ is a pre-defined constant threshold, $t_{i,0}$ is the index of the bar immediately after X_i takes place, $t_{i,0} + h$ is the index of *h* bars after $t_{i,0}$, and $r_{t_{i,0},t_{i,0}+h}$ is the price return over a bar horizon *h.*

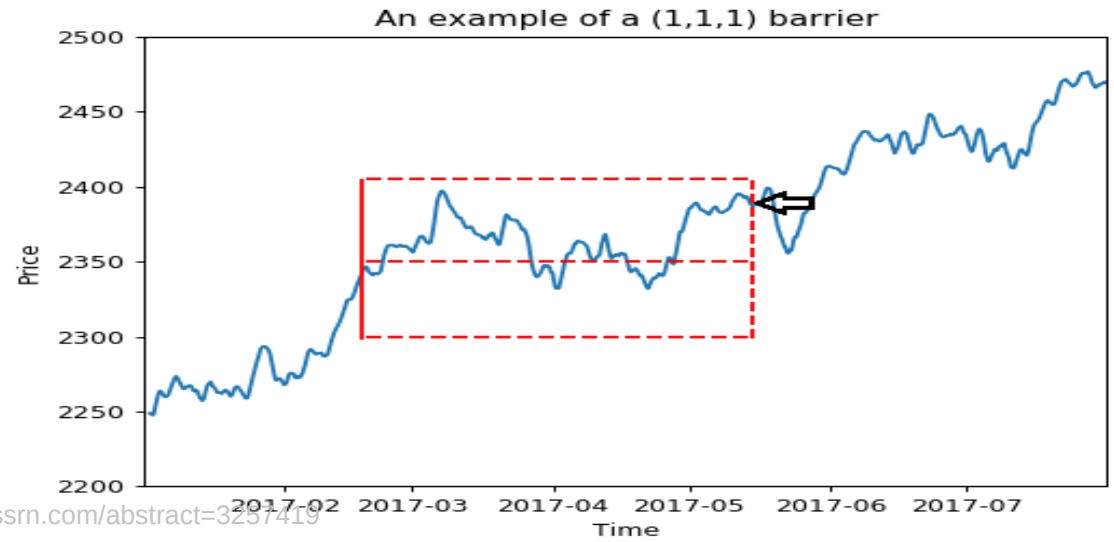
- Because the literature almost always works with time bars, *h* implies a fixed-time horizon.

Caveats of the Fixed Horizon Method

- There are several reasons to avoid such labeling approach:
  - Time bars do not exhibit good statistical properties.
  - The same threshold τ is applied regardless of the observed volatility.
    - Suppose that $\tau = 1E-2$, where sometimes we label an observation as $y_i = 1$ subject to a realized bar volatility of $\sigma_{t_{i,0}} = 1E-4$ (e.g., during the night session), and sometimes $\sigma_{t_{i,0}} = 1E-2$ (e.g., around the open). The large majority of labels will be 0, even if return $r_{t_{i,0},t_{i,0}+h}$ was predictable and statistically significant.
- A couple of better alternatives would be:
  - Label per a varying threshold $\sigma_{t_{i,0}}$, estimated using a rolling exponentially-weighted standard deviation of returns.
  - Use volume or dollar bars, as their volatilities are much closer to constant (homoscedasticity).
- But even these two improvements miss a key flaw of the fixed-time horizon method: The *path* followed by prices. We will address this with the Triple Barrier Method.

The Triple Barrier Method

- It is simply unrealistic to build a strategy that profits from positions that would have been stopped-out by the fund, exchange (margin call) or investor.
- The Triple Barrier Method labels an observation according to the first barrier touched out of three barriers.
  - Two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility (whether realized or implied).
  - A third, vertical barrier, is defined in terms of number of bars elapsed since the position was taken (an expiration limit).
- The barrier that is touched first by the *price path* determines the label:
  - Upper horizontal barrier: Label 1.
  - Lower horizontal barrier: Label -1.
  - Vertical barrier: Label 0.



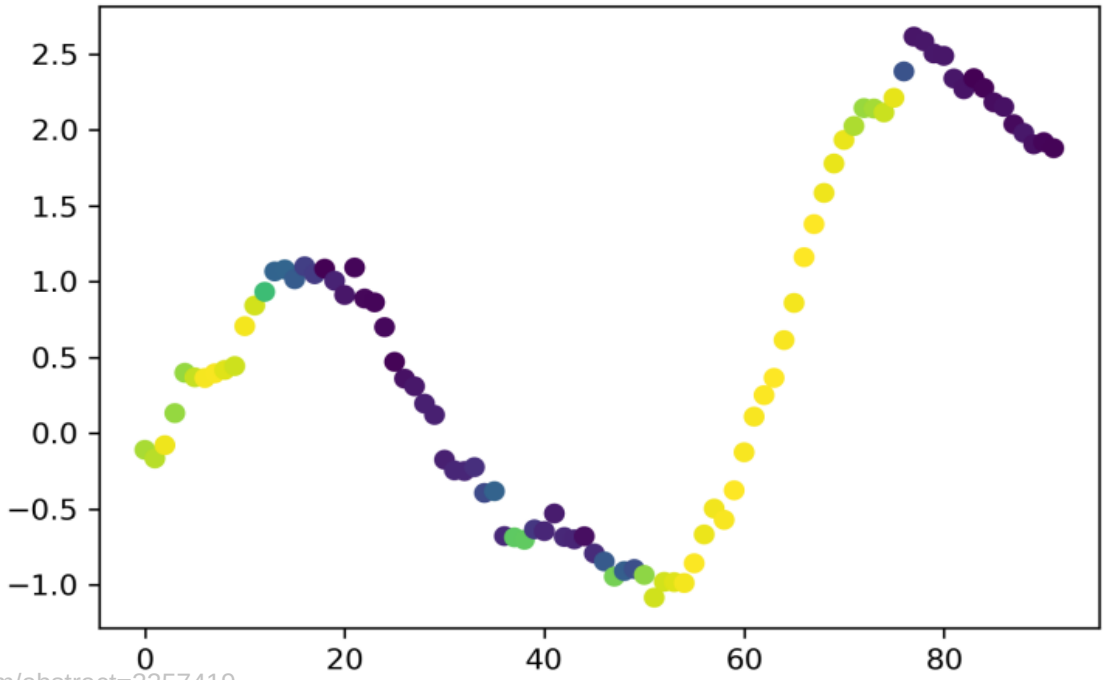
Trend Scanning Method

Consider a series of observations $\{x_t\}_{t=1,\dots,T}$, where x_t may represent the price of a security we aim to predict. We wish to assign a label $y_t \in \{-1,0,1\}$ to every observation in x_t, based on whether x_t is part of a downtrend, no-trend, or an uptrend. One possibility is to compute the t-value ($\hat{t}_{\beta_1}$) associated with the estimated regressor coefficient ($\hat{\beta}_1$) in a linear time-trend model,

$$x_{t+l} = \beta_0 + \beta_1 l + \varepsilon_{t+l}$$

$$\hat{t}_{\beta_1} = \hat{\beta}_1 \Big/ \hat{\sigma}_{\beta_1}$$

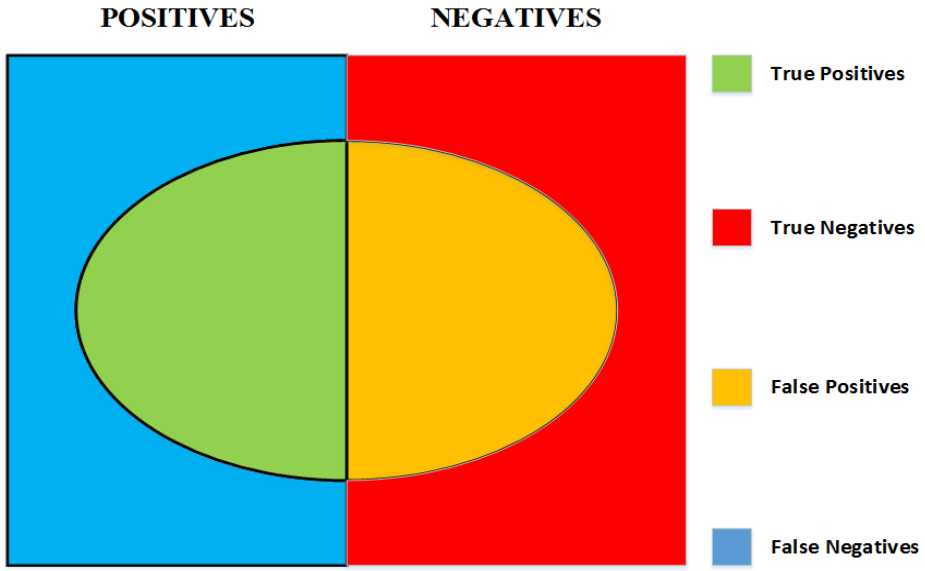
where $\hat{\sigma}_{\beta_1}$ is the standard deviation of $\hat{\beta}_1$, and $l = 0, \dots, L-1$, and L sets the look-forward period, with $L \leq t$. Different values of L lead to different t-values. To solve this indetermination, we can try a set of values for L, and pick the value that maximizes $\left|\hat{t}_{\beta_1}\right|$. In this way, we assign to x_t the most significant trend observed in the past, out of multiple possible look-forward periods.



Meta-Labeling

Turning a Weak Predictor into a Strong Predictor

- Suppose that you have a model for making a buy-or-sell decision:
  - You just need to learn the ***size*** of that bet, which includes the possibility of no bet at all (zero size).
  - This is a situation that practitioners face regularly. We often know whether we want to buy or sell a product, and the only remaining question is how much money we should risk in such bet.
  - Meta-labeling: Label the outcomes of the primary model as 1 (gain) or 0 (loss). See Sections 3.6-3.8 of AFML.
  - The goal is not to predict the market. Instead, the goal is to predict the success of the primary model.



- Meta-labeling builds a secondary ML model that learns how to use a primary exogenous model.
- The secondary model does not learn the *side*. It learns only the *size*.
- Meta-labeling is particularly useful when outcomes are asymmetric. In those cases, giving up some recall in exchange for improving the precision can yield a significant improvement in Sharpe ratio.

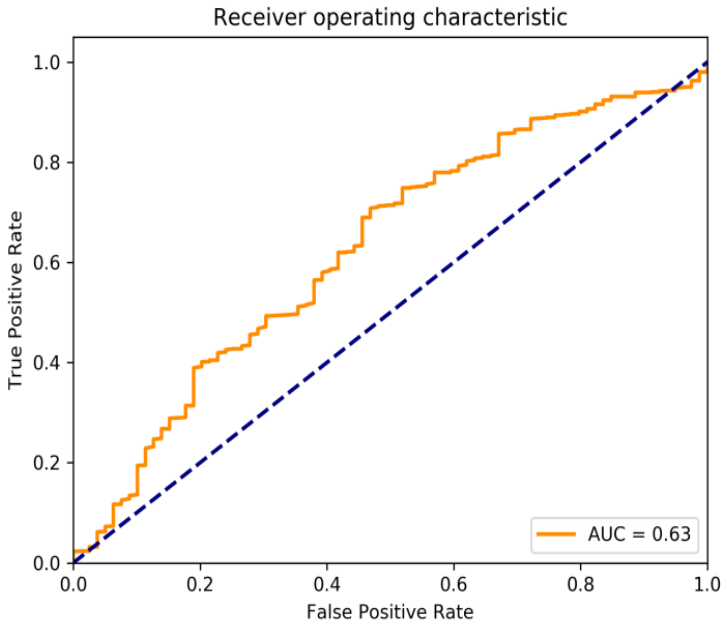
Why Meta-Labeling Works

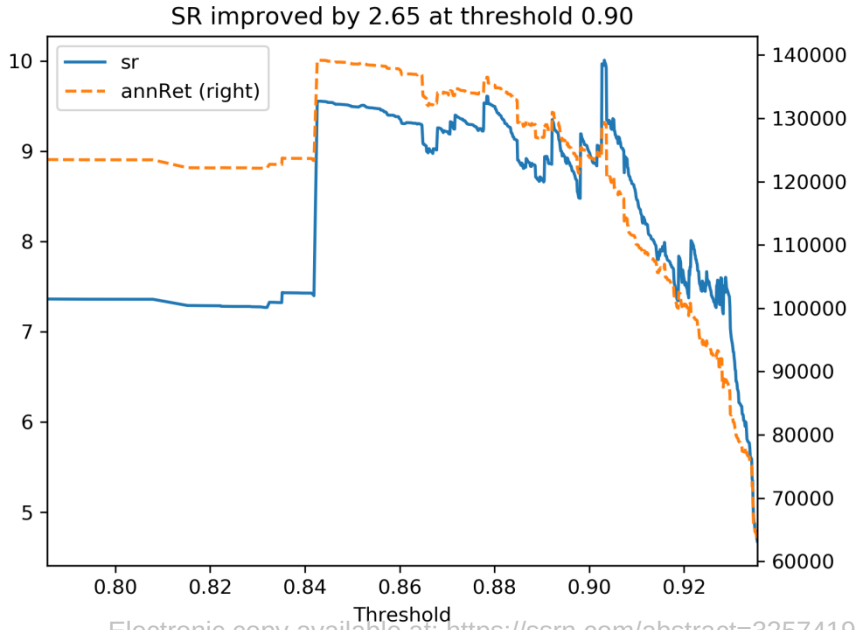
- The Sharpe ratio associated with a binary outcome can be derived as

$$\theta[p, n, \pi_-, \pi_+] = \frac{(\pi_+ - \pi_-)p + \pi_-}{(\pi_+ - \pi_-)\sqrt{p(1-p)}}\sqrt{n}$$

where $\{\pi_-, \pi_+\}$ determine the payoff from negative and positive outcomes, p is the probability of a positive outcome, and n is the number of outcomes per year (see Section 15.3 of AFML).

- When $\pi_+ \ll -\pi_-$, it may be possible to increase $\theta[.]$ by increasing p and the expense of n.





The primary model determines $\{\pi_-, \pi_+\}$, and the secondary model adjusts the acceptance threshold to regulate $\{p, n\}$.

In this example, a strategy's Sharpe ratio increased by 2.65 thanks to Meta-Labeling's ability to avoid the largest losses.

How to use Meta-labeling

- Meta-labeling is particularly helpful when you want to achieve higher F1-scores:
  - First, we build a model that achieves high recall, even if the precision is not particularly high.
  - Second, we correct for the low precision by applying meta-labeling to the positives identified by the primary model.
- Meta-labeling is a very powerful tool in your arsenal, for three additional reasons:
  - ML algorithms are often criticized as *black boxes*. Meta-labeling allows you to build a ML system on a white box.
  - The effects of *overfitting* are limited when you apply meta-labeling, because ML will not decide the side of your bet, only the size.
  - Achieving high accuracy on small bets and low accuracy in large bets will ruin you. As important as identifying good opportunities is to *size bets* properly, so it makes sense to develop a ML algorithm solely focused on getting that critical decision (sizing) right.
- Meta-labeling should become an essential ML technique for every discretionary hedge fund
  - It allows the seamless combination of discretionary inputs (primary model) with a quantitative overlay (secondary model).

Sample Weights

Weighting observations by uniqueness (1/2)

- Two labels y_i and y_j are concurrent at t when both are a function of at least one common return, $r_{t-1,t} = \frac{p_t}{p_{t-1}} - 1$.

1. For each observation $t = 1, \ldots, T$ we form a binary array, $\{1_{t,i}\}_{i=1,\ldots,I}$, with $1_{t,i} \in \{0,1\}$, which indicates whether its outcome spans over return $r_{t-1,t}$.
2. We compute the number of labels concurrent at t, $c_t = \sum_{i=1}^{I} 1_{t,i}$.
3. The uniqueness of a label *i* at time t is $u_{t,i} = 1_{t,i} c_t^{-1}$.
4. The average uniqueness of label *i* is the average $u_{t,i}$ over the label's lifespan, $\bar{u}_i = \left(\sum_{t=1}^{T} u_{t,i}\right)\left(\sum_{t=1}^{T} 1_{t,i}\right)^{-1}$.

Weighting observations by uniqueness (2/2)

5. Sample weights can be defined as the sum of the attributed absolute log returns, $\left|r_{t_{i-1},t_i}\right|$, over the event's lifespan, $\left[t_{i,0}, t_{i,1}\right]$,

$$\widetilde{w}_i = \left| \sum_{t=t_{i,0}}^{t_{i,1}} \frac{r_{t-1,t}}{c_t} \right|_1$$

$$w_i = \widetilde{w}_i I \left(\sum_{j=1}^{I} \widetilde{w}_j \right)$$

- The rationale for this method is that we weight an observation as a function of the absolute log returns that can be attributed *uniquely* to it.
- We can use these weights for sequential bootstrap (section 4.5).

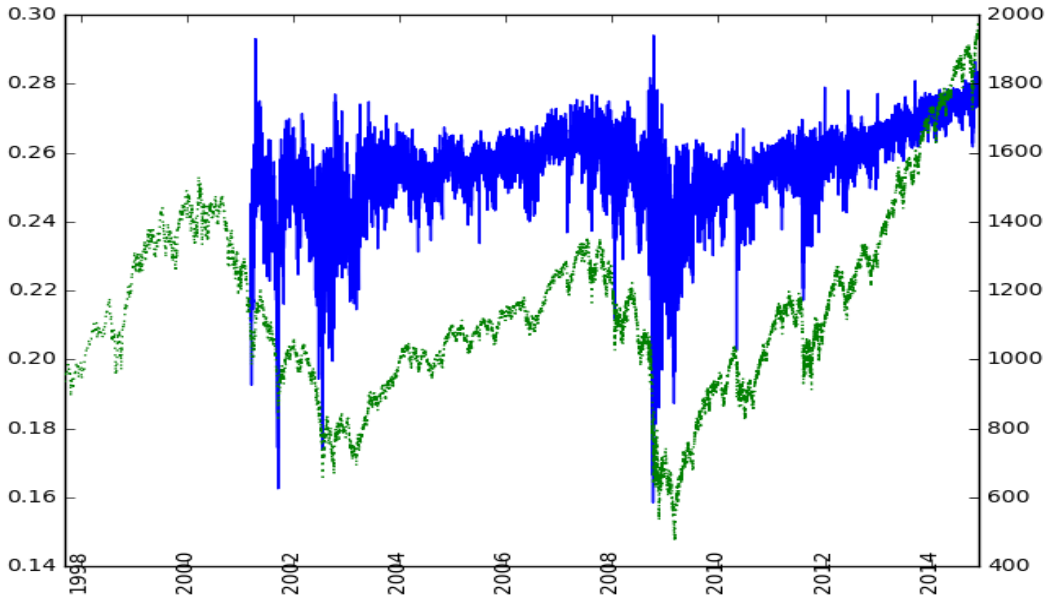
Fractionally Differentiated Features

The Stationarity vs. Memory Dilemma

- In order to perform inferential analyses, researchers need to work with invariant processes, such as
  - returns on prices (or changes in log-prices)
  - changes in yield
  - changes in volatility
- These operations make the series stationary, at the expense of removing all memory from the original series.
- Memory is the basis for the model's predictive power.
  - For example, equilibrium (stationary) models need some memory to assess how far the price process has drifted away from the long-term expected value in order to generate a forecast.
- The dilemma is
  - returns are stationary however memory-less; and
  - prices have memory however they are non-stationary.

The Optimal Stationary-Memory Trade Off

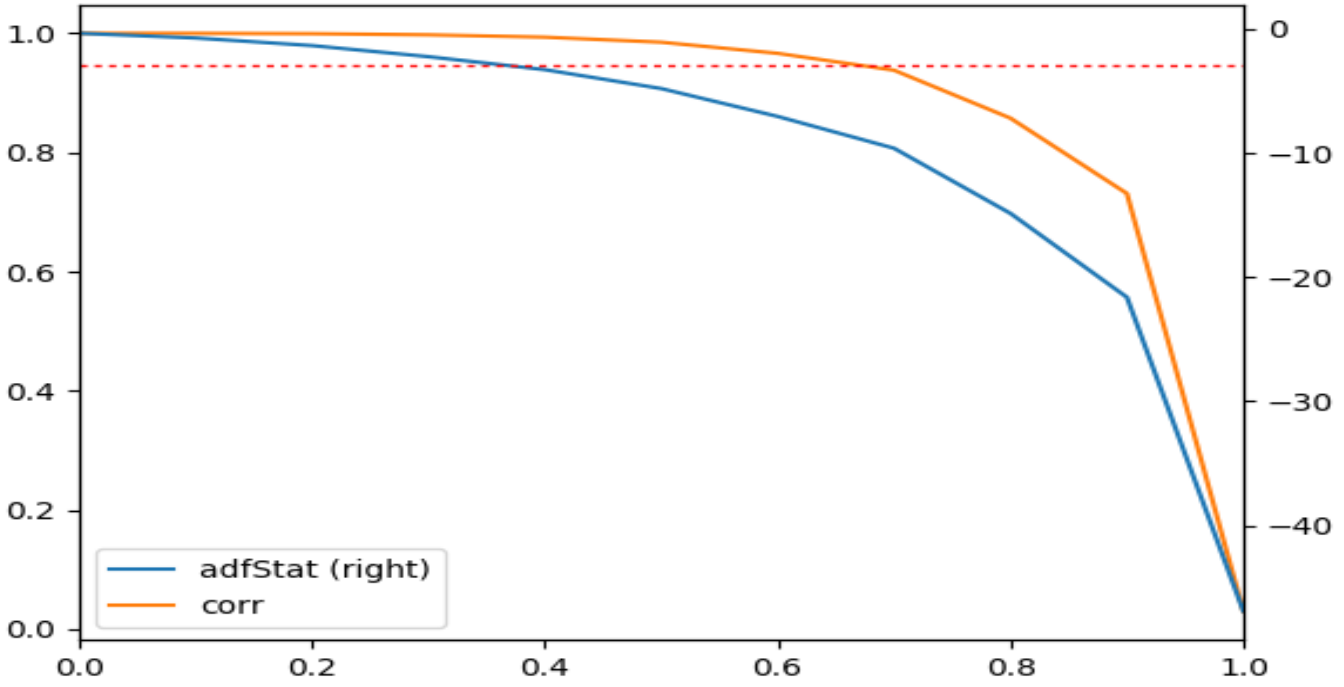
- Question: What is the minimum amount of differentiation that makes a price series stationary while preserving as much memory as possible?
- Answer: We would like to generalize the notion of returns to consider stationary series where not all memory is erased.
- Under this framework, returns are just one kind of (and in most cases suboptimal) price transformation among many other possible.



- Green line: E-mini S&P 500 futures trade bars of size 1E4
- Blue line: Fractionally differentiated ($d = .4$)
- Over a short time span, it resembles returns
- Over a longer time span, it resembles price levels

Example 1: E-mini S&P 500 Futures

- On the x-axis, the *d* value used to generate the series on which the ADF stat was computed.
- On the left y-axis, the correlation between the original series ($d = 0$) and the differentiated series at various *d* values.
- On the right y-axis, ADF stats computed on log prices.



The original series ($d = 0$) has an ADF stat of -0.3387, while the returns series ($d = 1$) has an ADF stat of -46.9114.

At a 95% confidence level, the test's critical value is -2.8623.

The ADF stat crosses that threshold in the vicinity of $d = 0.35$, where correlation is still very high (0.995).

Example 2: Optimal FradDiff Stationarity (1/2)

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AD1 Curncy	-1.7253	-1.8665	-2.2801	-2.9743	-3.9590	-5.4450	-7.7387	-10.3412	-15.7255	-22.5170	-43.8281
BO1 Comdty	-0.7039	-1.0021	-1.5848	-2.4038	-3.4284	-4.8916	-7.0604	-9.5089	-14.4065	-20.4393	-38.0683
BP1 Curncy	-1.0573	-1.4963	-2.3223	-3.4641	-4.8976	-6.9157	-9.8833	-13.1575	-19.4238	-26.6320	-43.3284
BTS1 Comdty	-1.7987	-2.1428	-2.7600	-3.7019	-4.8522	-6.2412	-7.8115	-9.4645	-11.0334	-12.4470	-13.6410
BZ1 Index	-1.6569	-1.8766	-2.3948	-3.2145	-4.2821	-5.9431	-8.3329	-10.9046	-15.7006	-20.7224	-29.9510
C 1 Comdty	-1.7870	-2.1273	-2.9539	-4.1642	-5.7307	-7.9577	-11.1798	-14.6946	-20.9925	-27.6602	-39.3576
CC1 Comdty	-2.3743	-2.9503	-4.1694	-5.8997	-8.0868	-10.9871	-14.8206	-18.6154	-24.1738	-29.0285	-34.8580
CD1 Curncy	-1.6304	-2.0557	-2.7284	-3.8380	-5.2341	-7.3172	-10.3738	-13.8263	-20.2897	-27.6242	-43.6794
CF1 Index	-1.5539	-1.9387	-2.7421	-3.9235	-5.5085	-7.7585	-11.0571	-14.6829	-21.4877	-28.9810	-44.5059
CL1 Comdty	-0.3795	-0.7164	-1.3359	-2.2018	-3.2603	-4.7499	-6.9504	-9.4531	-14.4936	-20.8392	-41.1169
CN1 Comdty	-0.8798	-0.8711	-1.1020	-1.4626	-1.9732	-2.7508	-3.9217	-5.2944	-8.4257	-12.7300	-42.1411
CO1 Comdty	-0.5124	-0.8468	-1.4247	-2.2402	-3.2566	-4.7022	-6.8601	-9.2836	-14.1511	-20.2313	-39.2207
CT1 Comdty	-1.7604	-2.0728	-2.7529	-3.7853	-5.1397	-7.1123	-10.0137	-13.1851	-19.0603	-25.4513	-37.5703
DM1 Index	-0.1929	-0.5718	-1.2414	-2.1127	-3.1765	-4.6695	-6.8852	-9.4219	-14.6726	-21.5411	-49.2663
DU1 Comdty	-0.3365	-0.4572	-0.7647	-1.1447	-1.6132	-2.2759	-3.3389	-4.5689	-7.2101	-10.9025	-42.9012
DX1 Curncy	-1.5768	-1.9458	-2.7358	-3.8423	-5.3101	-7.3507	-10.3569	-13.6451	-19.5832	-25.8907	-37.2623
EC1 Comdty	-0.2727	-0.6650	-1.3359	-2.2112	-3.3112	-4.8320	-7.0777	-9.6299	-14.8258	-21.4634	-44.6452
EC1 Curncy	-1.4733	-1.9344	-2.8507	-4.1588	-5.8240	-8.1834	-11.6278	-15.4095	-22.4317	-30.1482	-45.6373
ED1 Comdty	-0.4084	-0.5350	-0.7948	-1.1772	-1.6633	-2.3818	-3.4601	-4.7041	-7.4373	-11.3175	-46.4487
EE1 Curncy	-1.2100	-1.6378	-2.4216	-3.5470	-4.9821	-7.0166	-9.9962	-13.2920	-19.5047	-26.5158	-41.4672
EO1 Comdty	-0.7903	-0.8917	-1.0551	-1.3465	-1.7302	-2.3500	-3.3068	-4.5136	-7.0157	-10.6463	-45.2100
EO1 Index	-0.6561	-1.0567	-1.7409	-2.6774	-3.8543	-5.5096	-7.9133	-10.5674	-15.6442	-21.3066	-35.1397
ER1 Comdty	-0.1970	-0.3442	-0.6334	-1.0363	-1.5327	-2.2378	-3.2819	-4.4647	-7.1031	-10.7389	-40.0407
ES1 Index	-0.3387	-0.7206	-1.3324	-2.2252	-3.2733	-4.7976	-7.0436	-9.6095	-14.8624	-21.6177	-46.9114
FA1 Index	-0.5292	-0.8526	-1.4250	-2.2359	-3.2500	-4.6902	-6.8272	-9.2410	-14.1664	-20.3733	-41.9705
FC1 Comdty	-1.8846	-2.1853	-2.8808	-3.8546	-5.1483	-7.0226	-9.6889	-12.5679	-17.8160	-23.0530	-31.6503
FV1 Comdty	-0.7257	-0.8515	-1.0596	-1.4304	-1.8312	-2.5302	-3.6296	-4.9499	-7.8292	-12.0467	-49.1508
G 1 Comdty	0.2326	0.0026	-0.4686	-1.0590	-1.7453	-2.6761	-4.0336	-5.5624	-8.8575	-13.3277	-42.9177
GC1 Comdty	-2.2221	-2.3544	-2.7467	-3.4140	-4.4861	-6.0632	-8.4803	-11.2152	-16.7111	-23.1750	-39.0715
GX1 Index	-1.5418	-1.7749	-2.4666	-3.4417	-4.7321	-6.6155	-9.3667	-12.5240	-18.6291	-25.8116	-43.3610
HG1 Comdty	-1.7372	-2.1495	-2.8323	-3.9090	-5.3257	-7.3805	-10.4121	-13.7669	-19.8902	-26.5819	-39.3267
HI1 Index	-1.8289	-2.0432	-2.6203	-3.5233	-4.7514	-6.5743	-9.2733	-12.3722	-18.5308	-25.9762	-45.3396
HO1 Comdty	-1.6024	-1.9941	-2.6619	-3.7131	-5.1772	-7.2468	-10.3326	-13.6745	-19.9728	-26.9772	-40.9824
IB1 Index	-2.3912	-2.8254	-3.5813	-4.8774	-6.5884	-9.0665	-12.7381	-16.6706	-23.6752	-30.7986	-43.0687
IK1 Comdty	-1.7373	-2.3000	-2.7764	-3.7101	-4.8686	-6.3504	-8.2195	-9.8636	-11.7882	-13.3983	-14.8391
IR1 Comdty	-2.0622	-2.4188	-3.1736	-4.3178	-5.8119	-7.9816	-11.2102	-14.7956	-21.6158	-29.4555	-46.2683
JA1 Comdty	-2.4701	-2.7292	-3.3925	-4.4658	-5.9236	-8.0270	-11.2082	-14.7198	-21.2681	-28.4380	-42.1937
JB1 Comdty	-0.2081	-0.4319	-0.8490	-1.4289	-2.1160	-3.0932	-4.5740	-6.3061	-9.9454	-15.0151	-47.6037
JE1 Curncy	-0.9268	-1.2078	-1.7565	-2.5398	-3.5545	-5.0270	-7.2096	-9.6808	-14.6271	-20.7168	-37.6954
JG1 Comdty	-1.7468	-1.8071	-2.0654	-2.5447	-3.2237	-4.3418	-6.0690	-8.0537	-12.3908	-18.1881	-44.2884
JO1 Comdty	-3.0052	-3.3099	-4.2639	-5.7291	-7.5686	-10.1683	-13.7068	-17.3054	-22.7853	-27.7011	-33.4658
JY1 Curncy	-1.2616	-1.5891	-2.2042	-3.1407	-4.3715	-6.1600	-8.8261	-11.8449	-17.8275	-25.0700	-44.8394
KC1 Comdty	-0.7786	-1.1172	-1.7723	-2.7185	-3.8875	-5.5651	-8.0217	-10.7422	-15.9423	-21.8651	-35.3354
L 1 Comdty	-0.0805	-0.2228	-0.6144	-1.0751	-1.6335	-2.4186	-3.5676	-4.8749	-7.7528	-11.7669	-44.0349

These tables show ADF stats for the most liquid futures contracts worldwide.

One row per instrument, and one column per differentiation value.

Highlighted in green are ADF values that do not reject the null hypothesis of unit root.

Highlighted in red are ADF values that reject the null hypothesis of unit root.

Example 2: Optimal FradDiff Stationarity (2/2)

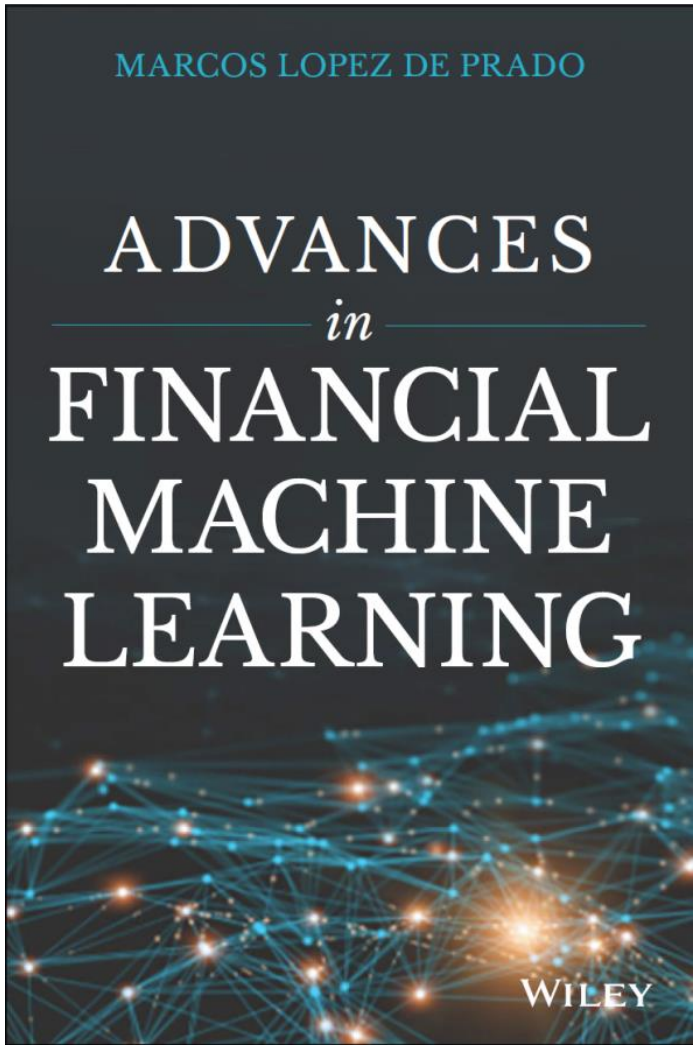
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
LB1 Comdty	-2.0133	-2.2043	-2.7692	-3.7363	-4.9980	-6.8712	-9.5572	-12.5024	-17.7300	-23.1173	-31.9508
LC1 Comdty	-3.0977	-3.2487	-4.0104	-5.1441	-6.8472	-9.1425	-12.4560	-16.0186	-21.8070	-27.1929	-34.2574
LH1 Comdty	-2.4059	-2.5980	-2.6847	-3.0616	-3.7269	-4.8461	-6.6899	-8.8143	-13.3179	-18.6747	-34.4944
MFS1 Index	-1.8618	-2.4061	-3.0316	-4.2111	-5.6544	-8.2728	-11.3954	-14.2083	-19.2276	-23.7318	-29.9174
NG1 Comdty	-1.2022	-1.2278	-1.2971	-1.5259	-1.9188	-2.5619	-3.5688	-4.7757	-7.4392	-11.2455	-41.3164
NI1 Index	-1.0865	-1.4354	-2.1171	-3.0946	-4.3528	-6.1476	-8.8056	-11.7667	-17.6428	-24.6738	-43.8325
NK1 Index	-0.8467	-1.1964	-1.8390	-2.7349	-3.8871	-5.5119	-7.9025	-10.5570	-15.8085	-22.0688	-38.7505
NQ1 Index	0.0153	-0.2883	-0.7985	-1.5227	-2.3900	-3.5965	-5.3719	-7.4372	-11.7580	-17.5718	-47.7300
NX1 Index	-1.2749	-1.6410	-2.3648	-3.4331	-4.8169	-6.8106	-9.7514	-13.0195	-19.3190	-26.5442	-43.2635
O 1 Comdty	-1.9643	-2.3536	-3.1711	-4.4057	-6.0102	-8.3139	-11.6484	-15.2893	-21.7540	-28.5592	-39.9112
OAT1 Comdty	-2.1234	-1.9151	-2.2928	-2.9948	-3.9627	-5.3126	-7.0749	-8.8556	-11.2388	-13.2080	-15.0069
OE1 Comdty	0.1688	-0.0863	-0.4587	-0.8500	-1.3174	-2.0411	-2.9760	-4.0461	-6.4504	-9.8420	-44.0898
PA1 Comdty	-1.4237	-1.6949	-2.2550	-3.1287	-4.2748	-5.9456	-8.4346	-11.2251	-16.6076	-22.8823	-37.8283
PE1 Curncy	-1.7713	-2.1928	-3.0869	-4.3894	-6.0523	-8.4218	-11.9137	-15.7241	-22.6601	-30.1037	-43.8788
PT1 Index	-1.9088	-2.2753	-3.0047	-4.1548	-5.6979	-7.9456	-11.2588	-14.8504	-21.5933	-28.9158	-43.4395
QS1 Comdty	-0.2084	-0.4919	-0.9675	-1.6192	-2.4490	-3.6160	-5.3075	-7.2161	-11.0838	-15.9596	-32.1660
RR1 Comdty	-0.0657	-0.4432	-0.9827	-1.6856	-2.5403	-3.7445	-5.4592	-7.4618	-11.4360	-16.4247	-33.0067
RTA1 Index	-0.4991	-0.8450	-1.4518	-2.2701	-3.3347	-4.8131	-7.0163	-9.4859	-14.4313	-20.5139	-38.4632
RX1 Comdty	0.3374	0.0368	-0.3370	-0.8033	-1.3293	-2.0307	-3.1201	-4.2717	-6.8379	-10.4035	-43.1525
S 1 Comdty	-2.3905	-2.5632	-3.0364	-3.8647	-5.0057	-6.7561	-9.4036	-12.4148	-18.2529	-24.9520	-39.1747
SB1 Comdty	-1.3895	-1.7489	-2.4806	-3.5180	-4.9204	-6.9044	-9.7911	-12.8777	-18.5958	-24.6554	-35.9220
SF1 Curncy	-2.4335	-2.8967	-3.8496	-5.3187	-7.2411	-9.9945	-13.9627	-18.2641	-25.8117	-33.5334	-46.1841
SI1 Comdty	-1.6435	-1.9468	-2.6104	-3.6207	-4.9544	-6.8834	-9.7471	-12.9306	-18.9448	-25.6872	-39.6744
SM1 Comdty	-2.1197	-2.0686	-2.2593	-2.7314	-3.5152	-4.6986	-6.5691	-8.7911	-13.3516	-19.1866	-37.8627
SM1 Index	-1.4716	-1.7336	-2.3942	-3.3732	-4.6921	-6.5834	-9.3968	-12.5018	-18.5601	-25.5175	-42.7253
SP1 Index	-0.5900	-0.9726	-1.6887	-2.6118	-3.7857	-5.4356	-7.8842	-10.6784	-16.4223	-23.8436	-50.2515
ST1 Index	-1.5957	-1.8926	-2.5130	-3.4803	-4.7593	-6.6294	-9.4127	-12.5153	-18.4786	-25.2546	-40.7900
TP1 Index	-1.2901	-1.6144	-2.2911	-3.3049	-4.5946	-6.4768	-9.2514	-12.3480	-18.5256	-25.9865	-46.2311
TU1 Comdty	-0.6340	-0.6768	-0.8529	-1.1306	-1.5256	-2.1951	-3.2065	-4.2674	-6.8060	-10.4758	-48.7361
TW1 Index	-1.1854	-1.5331	-2.2852	-3.3336	-4.6677	-6.5776	-9.3678	-12.4932	-18.5628	-25.6502	-42.5179
TY1 Comdty	-0.8208	-0.9876	-1.2585	-1.6069	-2.1026	-2.8142	-4.0467	-5.4328	-8.6137	-13.1678	-48.6412
UB1 Comdty	-0.3052	-0.5418	-0.9441	-1.4744	-2.1400	-3.0797	-4.4703	-6.0749	-9.4466	-13.9063	-36.3328
US1 Comdty	-0.8071	-1.1082	-1.5195	-2.0586	-2.8385	-4.0023	-5.7401	-7.7040	-12.0160	-18.0689	-47.9605
VG1 Index	-1.9920	-2.4127	-3.3269	-4.7189	-6.5700	-9.1847	-13.0116	-17.1131	-24.4105	-31.9086	-44.9058
VH1 Index	-1.5805	-1.9248	-2.7044	-3.8438	-5.3480	-7.5449	-10.7841	-14.3586	-21.2567	-29.0585	-46.5168
W 1 Comdty	-0.6236	-0.9148	-1.3959	-2.1267	-3.0507	-4.3849	-6.3497	-8.6538	-13.3216	-19.3053	-41.4181
XB1 Comdty	-2.2352	-2.4744	-2.9506	-3.7092	-4.9733	-6.7217	-9.4858	-12.5086	-18.3777	-25.0316	-39.5784
XG1 Comdty	-2.0082	-2.0972	-2.3756	-3.0026	-3.9027	-5.3023	-7.5000	-10.0158	-15.1353	-21.6376	-41.2603
XM1 Comdty	-0.9140	-1.1841	-1.8967	-2.8240	-4.0056	-5.6936	-8.2092	-11.0940	-17.0495	-24.7002	-51.5154
XP1 Index	-1.5053	-1.7699	-2.4437	-3.4436	-4.7258	-6.6019	-9.3891	-12.5294	-18.8368	-26.5249	-48.0102
YM1 Comdty	-1.1028	-1.1658	-1.6422	-2.3731	-3.3197	-4.6849	-6.7878	-9.1765	-14.2354	-20.9065	-49.2648
YS1 Comdty	-1.9101	-2.1735	-2.8727	-3.8500	-5.2679	-7.2488	-10.2821	-13.6430	-19.9992	-27.0788	-41.5913
Z 1 Index	-1.3096	-1.7242	-2.6045	-3.7736	-5.3196	-7.5241	-10.7341	-14.2851	-21.0992	-28.7746	-45.6802

Most financial series can be made stationary with a fractional differentiation of order $d < 0.5$.

However, most financial studies are based on returns, where $d = 1$.

The implication is that for decades most financial research has been based on **over-differentiated (memory-less) series**, leading to spurious forecasts and overfitting.

For Additional Details



The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

THANKS FOR YOUR ATTENTION!

Disclaimer

- The views expressed in this document are the authors' and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-