**Sampling interval.**

Certain data observations pertaining to financial assets will arrive at irregular times.

There is also no upper limit on how many observations can be made at any given point in time for some data.

Computational costs grow with sample size.

To train a ML-model on every single observation would obviously not scale well.

Hence, observations should be sampled from a curated dataset.

The rows in this dataset are often referred to as bars in finance.

A balance must be struck between sampling too frequently and too rarely.

A straightforward approach is to sample information at fixed time intervals. That might not be the best approach, however.

As exchange activity varies throughout the day, sampling at fixed intervals may lead to oversampling during periods of low activity and undersampling during periods of high activity.

In addition, time-sampled series often exhibit poor statistical properties, such as non-normality of returns, autocorrelation, and heteroskedasticity \cite{de2018advances}.

Using some simplifying assumptions for the market, theoretical models simulate financial asset prices as Brownian motion.

The normality of returns assumption underpins several models in mathematical finance e.g., Modern Portfolio Theory \cite{markowitz1968portfolio} and the Sharpe-ratio \cite{sharpe1998sharpe}.

There is, however, too much peaking and fatter tails in the real distribution for it to be relative to samples from Gaussian populations \cite{mandelbrot1997variation}.

Mandelbrot showed in 1963 \cite{mandelbrot1997variation} that returns over fixed time periods can be approximated by a Lévy alpha-stable distribution with infinite variance.

This is known as a stable Paretian distribution.

However, in 1967 Mandelbrot and Taylor \cite{mandelbrot1967distribution} argued that returns over a fixed number of transactions may be close to independent and identically distributed Gaussian.

Several studies have since confirmed this \cite{clark1973subordinated}\cite{ane2000order}.

Later, Clark \cite{clark1973subordinated} discovered that sampling by volume instead of transactions exhibits better statistical properties. I.e., closer to IID Gaussian distribution.

Sampling by volume instead of ticks has an intuitive appeal.

While tick bars count one transaction of $n$ contractions as one, $n$ transactions of one contract count as $n$ transactions.

However, for volatile securities, sampling according to the volume of transactions might lead to large variations in sampling frequencies.

When the price is high the volume will be lower, and therefore the number of observations, will be lower, and vice versa, even though the same value might be transacted.

Therefore, sampling by the monetary value transacted may exhibit even better statistical properties \cite{de2018advances}.

Furthermore, sampling by monetary value exchanged make an algorithm more robust against corporate actions like stock splits and reverse splits, as well as stock offerings and buybacks.

To maintain a suitable sampling frequency, the sampling threshold may need to be adjusted if the total market size changes significantly.

**Information-driven bars.**

This idea is from book Advances in Financial Machine Learning by Marcos Lopez de Prado. After having both tried this myself and read about other people’s attempts to implement this: it doesn’t work but was just meant as a starting point/an example.

The EMWA of the expected tick-size converges to some fraction of the initial expectation.

The EMWA of the expected imbalance either explodes if the initial expectation is over some obscure threshold, or it becomes very small such that almost every tick breaks the threshold.

However, a constant threshold work quite well and clearly exhibits improved statistical properties.

There probably is a way of creating a more dynamic threshold that is better than the one suggested.

There are also some small problems like how to handle end of the day auctions and futures roll over.

Ideally one would sample data at times when new, in a market microstructural sense, information arrives to the market.

New information is detected through the presence of informed traders \cite{de2018advances}.

The presence of informed traders is detected through order imbalance.

This would present the optimal opportunity for the ML model to exploit new information before the market price reaches a new equilibrium.

Therefore, one would like to synchronize sampling with the presence of informed traders.

Using the knowledge from the last section we want to sample information on intervals related to the actual value exchanged. We can fine tune this approach with the new approach of sampling when new information arrives to the market.

Combining this gives Dollar Imbalance Bars (DIB), which is sampling information when the dollar imbalance diverges from our expectation \cite{de2018advances}.

In order to generate DIB we need to find an expectation for the imbalance, so that we can sample when this balance is exceeded.

First we define a tick as the sequence $\{p\_t, v\_t\}\_{t=1,...,T}$, where $p\_t$ and $v\_t$ is respectively the price and dollar amount exchanged for the tick at time $t$.

The tick rule is defined as a sequence $\{b\_t\}\_{t=1,...,T}\in \{-1,1\}$ where

\begin{equation}

b\_t=

{\begin{cases}

b\_{t-1} & if \Delta p\_t=0 \\

\dfrac{|\Delta p\_t|}{\Delta p\_t} & if \Delta p\_t\neq0 \\

\end{cases} }

\end{equation}

with the boundary condition that $b\_0$ is set to the same value as $b\_T$.

Intuitively the tick rule tells us if the price has increased or decreased at time $t$. If the price has not changed it is set to the value at time $t-1$.

The imbalance at time $T$ is defined as

\begin{equation}

\theta\_T =\sum\_{t=1}^T {b\_t v\_t}

\end{equation}

which represents the direction and size of the market imbalance.

The expected value of $\theta\_T$ at the start of the bar is

\begin{align\*}

E\_0[\theta\_T]={} & E\_0\left[ \sum^T\_{t|b\_t=1} v\_t \right] - E\_0\left[ \sum^T\_{t|b\_t=-1} v\_t \right] \\

={} & E\_0[T](P[b\_t=1]E\_0[v\_t | b\_t=1]-P[b\_t=-1]E\_0[v\_t | b\_t=-1] \\

={} & E\_0[T](v^+-v^-)

\end{align\*}

We define $E\_0[T]$ as the expected size of the tick-bar at the start of the bar. It can be estimated using an exponentially weighted moving average (EWMA) of $T$ values from previous bars.

We define $v^+=P[b\_t=1]E\_0[v\_t | b\_t=1]$ and $v^-=P[b\_t=-1]E\_0[v\_t | b\_t=-1]$. Since $P[b\_t=1]E\_0[v\_t | b\_t=1] + P[b\_t=-1]E\_0[v\_t | b\_t=-1] = E\_0[v\_t]$ we can rewrite $v^+-v^-=2v^+-E\_0[v\_t]$.

$2v^+-E\_0[v\_t]$ implies the expected imbalance, and can be estimated using an EWMA of $b\_t v\_t$ values from previous bars.

Now, we can define DIB as $T^\*$-contiguous subset of ticks where the following condition is satisfied

\begin{equation}

T^\*=\argmin\_T \{ |\theta\_T| \geq E\_0[T] | 2v^+-E\_0[v\_t] | \}

\end{equation}