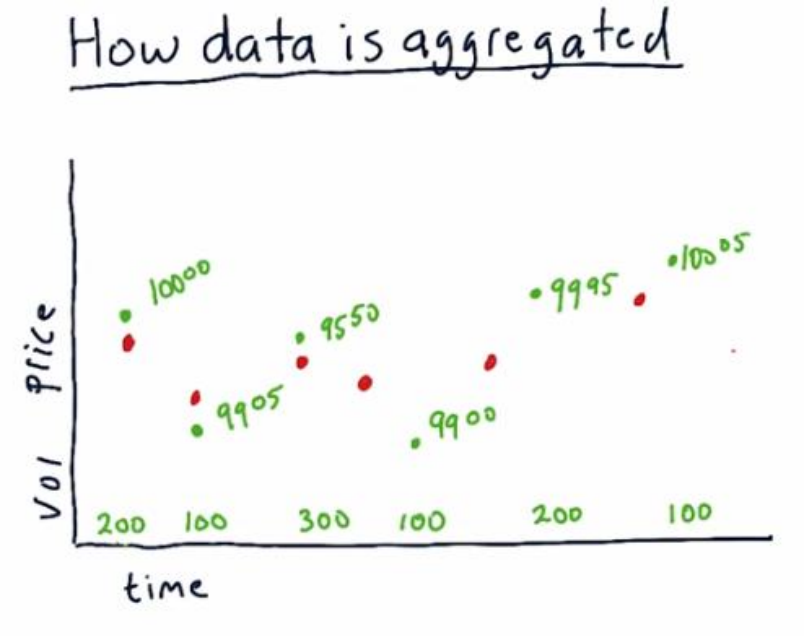
02-07 Dealing With Data

Compiled by Shipra De, Fall 2016

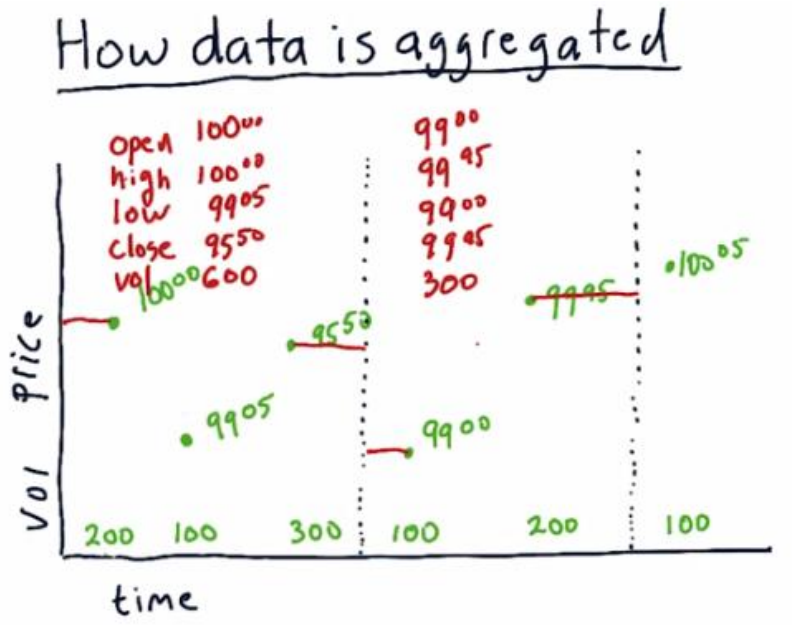
Lesson Overview

- Data is, of course, very important for computational investing.
- The core data and the primary data we will work with in this course is historical price and volume data. That's what we'll focus on in this lesson.

How Data is Aggregated

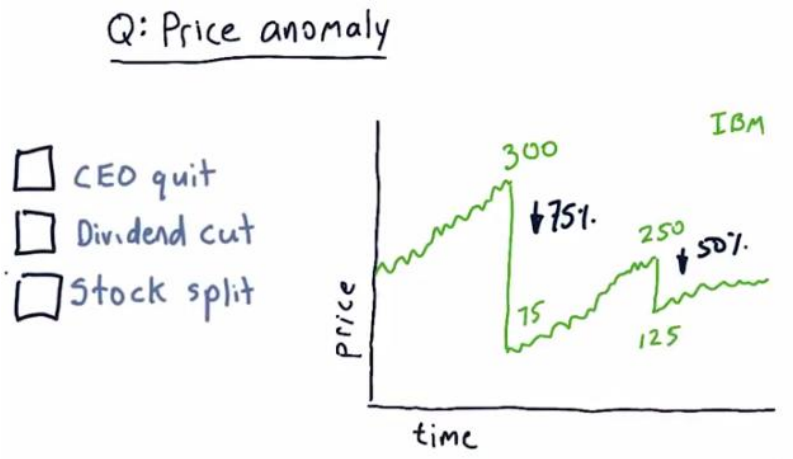


- Our first step in considering data is to think about how it's aggregated. In other words, imagine that many, many trades are happening on different exchanges. How is all that data combined and how is reported back to us so that we can analyze it and use it?
- The finest resolution of data is called a tick. A tick represents a successful by cell match or a successful transaction.
- So let's suppose that this time there's a successful transaction $100 was the price and 200 was the volume. So we record that here with a single point. A little bit later there was another transaction 9905, a 100 shares.
- Now, something to keep in mind is each of these transactions happens at no specific time. It happens only when the buy and sell are matched. So there's no guarantee that there's going to be a trade in any particular minute or any particular hour. It just happens when it happens.
- Each exchange provides its own data feed regarding these transactions. So you can subscribe to these feeds and see at each tick when a transaction happens on any particular exchange.
- So I've added some red dots here that represent transactions that might be happening on another exchange. Note that the prices are a little bit different, the volumes of course might be different. All of this happening simultaneously and the prices of different exchanges aren't guaranteed to be exactly the same.
- Now for highly liquid stocks, there may be hundreds or thousands or hundreds of thousands of transactions like these every second. Collecting and using all these ticks for all the exchanges over a long period of time would result in a lot of data, and it becomes very complex.

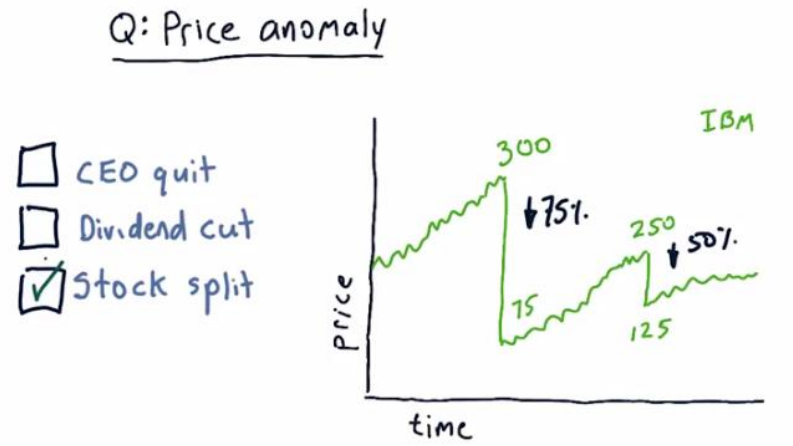


- Tick data is usually consolidated into minute by minute or hour by hour chunks. So I've drawn some boundaries here. Let's suppose those are minute boundaries, and we collect all the data together within each minute, and represent it as open, high, low, close and volume.
- So the open is the first transaction within the time period. So we came in here, we had a transaction at 100, so our open value is 100 within this time period. High, well you look at the entire time period and you see what was the highest price. The highest price here was also 100. Low was 99.05, close is the last transaction and that was 95.50. And volume is just the total volume during that time period. So it's 200 plus 100 plus 300, so 600 shares. So for this minute of time we've consolidated all that information into open, high, low, close, and volume.
- Similarly we go to the next minute and consolidate it in the same way. So in this next minute our open was 99, our close 99.95, low is 99, high was 99.95 and our volume was 300.
- And we will continue throughout the day to consolidate the data at each minute like this. And depending on your data feed, this might be the result say, for one exchange or it might be combined across multiple exchanges.
- In this course, these time boundaries are in days. So the data we'll be working with is daily data. And so we'll be looking at what is the data at the end of the day for each day.
- Now, all of the concepts that we teach here, you can just as easily apply at smaller time periods. It just requires more computing, faster computing, more and larger databases, and so on. So that's why we look at daily data in this course, is there's a little bit less data to work with. It's easier for you to download and work with and so on.

Price Anomaly

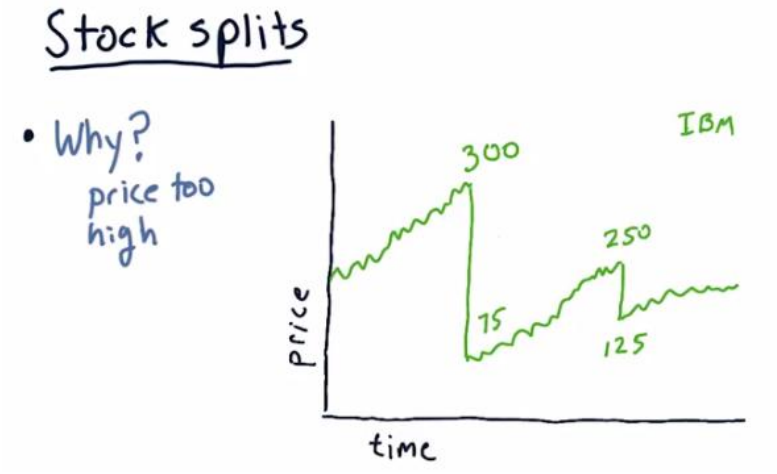


- If you look at this data over many years, you'll see that there are a couple sudden price drops. For instance, here it goes from $300 to $75 in one day. Here, it goes from 250 to 125 in one day. So those represent a 75% drop and a 50% drop respectively.
- Now surely the value of IBM did not drop that much in just one day. I want you to think about why this might've happened. In a moment, I'll give you some options that you can choose. But why might the price of IBM dropped by this significant amount over these days?
- Here are a few options for you to consider. Check the box that you think makes the most sense.

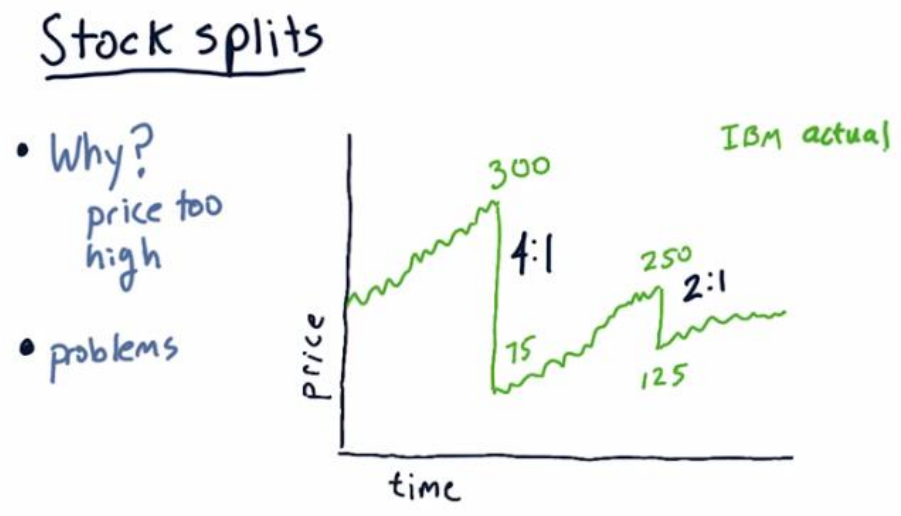


- So the correct answer is stock split. You might cry foul because I haven't told you yet what a stock split is. But don't worry, we're getting there. But just as advance warning, here's what it is.
- What happened on this day is that the stock was selling at 300, and if you had one share of the stock on this day, on the very next day, you had four shares. So your total value was preserved, you still had $300 worth of IBM, just now you had instead of one share, you had four shares. That's called a stock split.

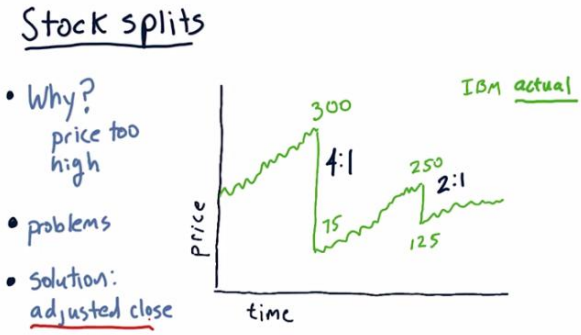
Stock Splits



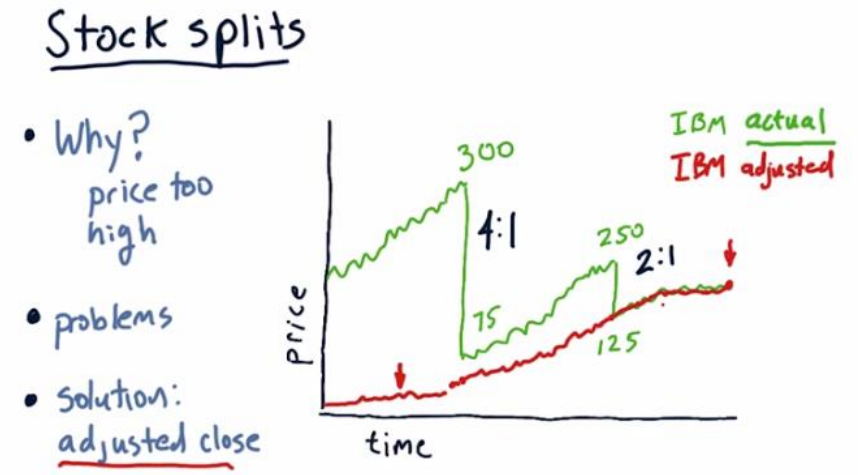
- So indeed these big price drops were caused by stock splits. Why do stocks split? The most common reason and as far as I know, the only reason, is that the price is too high.
- Why is a high price a problem? Well, consider for a moment that stock may be above say $500 per share. People like to buy stocks in groups of 100. So that means, for instance, it would be $50,000 to buy or sell a block of 100 shares of a $500 stock. Now, of course, you can buy it in smaller blocks. That's fine. But another aspect is that options when they're traded on stocks are usually traded with regard to 100 shares. So options covering 100 shares of $500 stocks, becomes suddenly very expensive and less liquid.
- So from the point of view of options, and also individual stock shares, very high prices are a problem. Even in the case where you want to buy just one share of the stock. Let's suppose you're setting your kid up with an account, and you want to buy one share of Apple. Earlier this year that would have cost you $600. So even buying one share of a very expensive stock can be a problem.
- Finally, one other issue here is if you're building a portfolio and you want to have a finely tuned proportion of each stock in your portfolio. If some of the stock prices are very high, it becomes difficult to get that fine resolution that you want. So, when the prices get very, very high, what the companies do is they say, look, let's take that 1 share that's price at $300 and break it into 4 shares at 75. So that's called a 4 of a 1 split.



- In this case, we had a 2 for 1 split. So that's why splits happen. Now let's look at this data. Suppose you read this data into your computer and you were going to analyze it and look for trading opportunities.
- Well, your algorithm might discover, hey, look, here's a great shorting opportunity. Whenever this condition occurs, whatever that was, you want to short the stock, and then you'll see a 75% drop, and you'll reap a significant reward. And also here, you might identify this time as a good opportunity to short as well.
- Well, of course, that's wrong because what's happening underneath, is that the value of the company isn't really decreasing. You just have more shares. So if you want to trade using this actual closed data, you have to account for all these splits. And that becomes cumbersome.

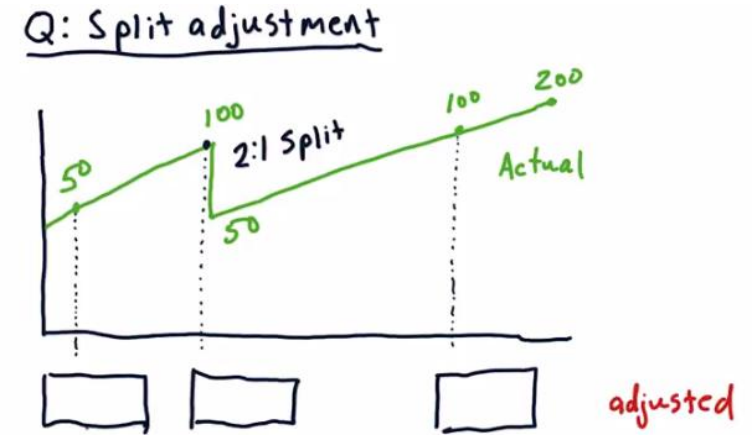


- So someone came up with a great solution to this problem. Mainly, adjusted close. Or adjusted prices. And the idea is to create a timeline of prices that are adjusted to account for these changes such that you can look back over adjusted close, simulate buying at a particular time, and seeing how the value increases over time accurately without having to account for these splits.
- Here's how it works.

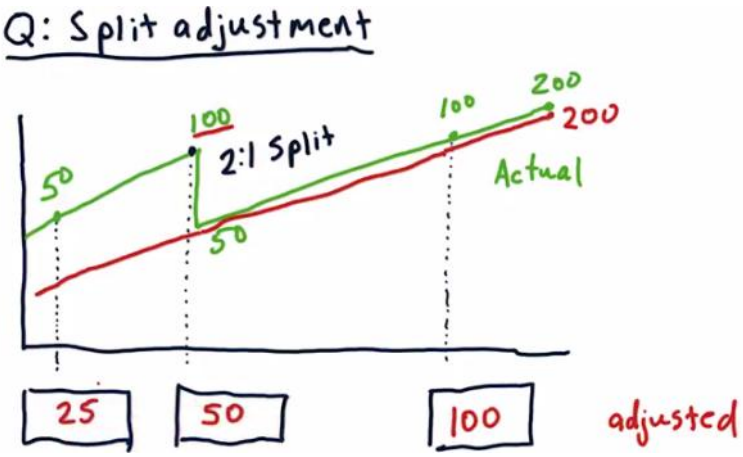


- The first thing to point out is that at the very last day, in other words, today, adjusted close and actual close are always the same. Now we track back in time and adjusted close and actual close are exactly the same.
- But then, on this day, when we see this 2 to 1 split. What we do is we go back over all the historical data, and we divide it all by 2. So the first day before the split, we get about a $125 price instead of that 250. Then when we get to this 4 to 1 split we divide by 2, and then by 4, so that on this 1 day before that split our price is in the neighborhood of about $70, and that continues back in time.
- As we go back we adjust for each of the splits in history. So now we have a nice smooth price timeline without these big jumps. And for instance, if you consider that you bought the stock on this day and held it until this day, the accumulation in value that you see there is accurate. That reflects all those splits and you would have a lot more shares here, but this would also correctly represent the increase in value.

Split Adjustment

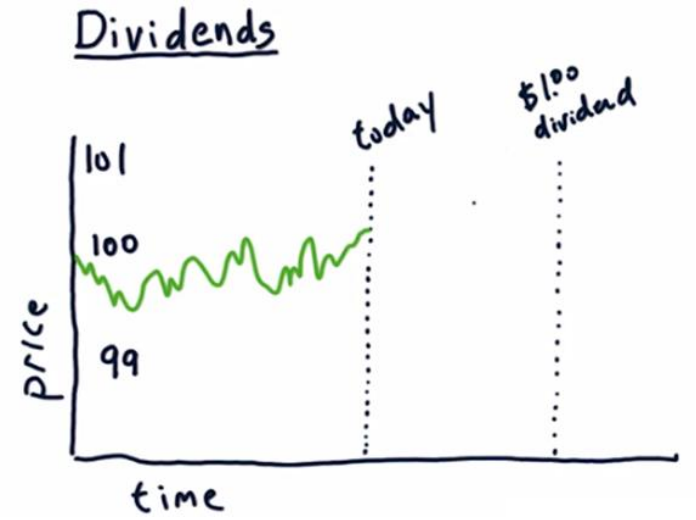


- Consider this situation. This green line represents the actual price for a stock over time. On this day, it had a two for one split and accordingly, it went from $100 to $50 on that day.
- So in light of this split on this particular day, I'd like you to consider these actual closing prices, and then calculate what the corresponding adjusted closing prices would be on these days.

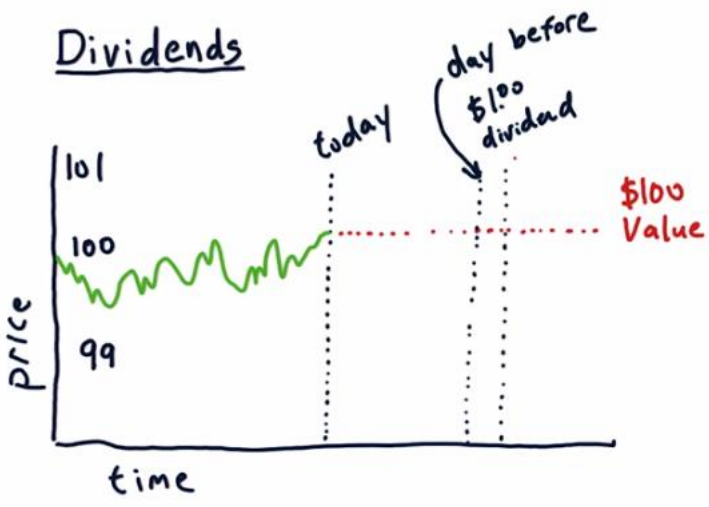


- So, let's start over here, going back from today, our adjusted price is the same as the price today. So, as we go back and we get to this day where there is a $100 in the actual. There's no change, it's still a $100 also for the adjusted okay, we go back further on this day of the split, again, no change that would have been $50, but we're looking to calculate the price just before $100.
- In other words, when the actual price was 100 just before that split, what would the adjusted close be? Well, we divide everything by 2 remember, so we would have 50 here. So it would cruise along like this. And again, here we divide by two because that's what our recent split was and we would have 25 here. So this correctly represents that if you bought the stock on this day it would double in value by the time you got to this day. And if you bought it on this day and held it all the way to here, you would actually get four times your value.

Dividends

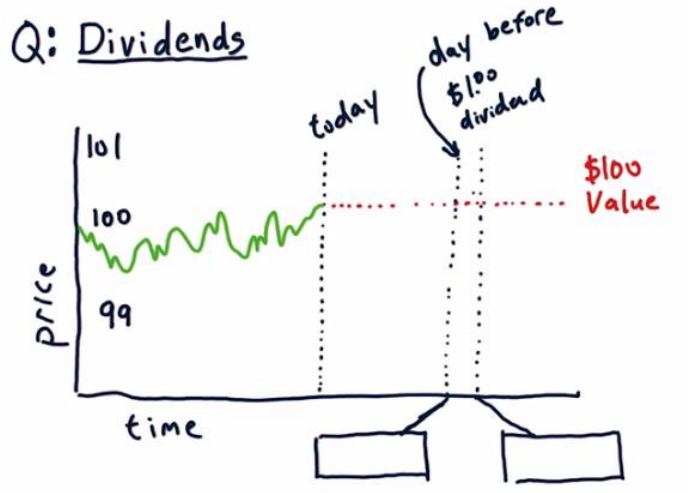


- Now remember earlier when we were talking about stock value, or I should say company value. And we looked at a way to compute company value based on dividends. So companies pay regularly, many companies, not all, dividends to their owners.
- And this can be worth a lot of money. So a stock that's priced at say around $100, can very often pay up to $1 or $2 per year in dividends, or 1 or 2%. Some stocks pay even more than that.
- Now dividends can have a significant effect on what happens to the actual price of the stock, and here's why. Suppose this stock is trading at about $100. And they announce, on this date in the future, we're going to pay a $1 dividend. So for every share of stock that you own, you'll get $1.
- What do you think is going to happen to the price of this stock between now and the date of that dividend?

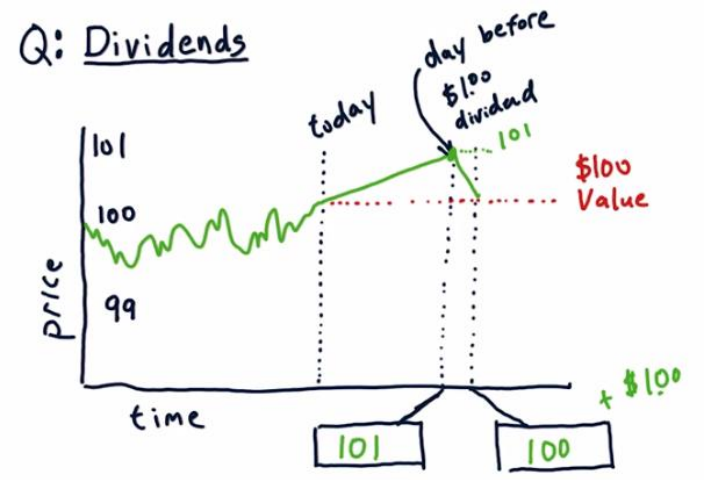


- Let me give you a little bit more information that might make things simpler. Let's suppose that the consensus on the value of the company, by whatever means you've determined that consensus, is that the company is worth $100 per share.
- So on this date, in order to we get that $1 dividend, we're going to have one share of the company that people think is worth $100 per share and we will have $1. So on this date, we'll have one share of a company supposedly worth $100 and $1. So think about it. What prices will we see this stock converge to on the day before the dividend is paid and the day the dividend is paid?

Dividends

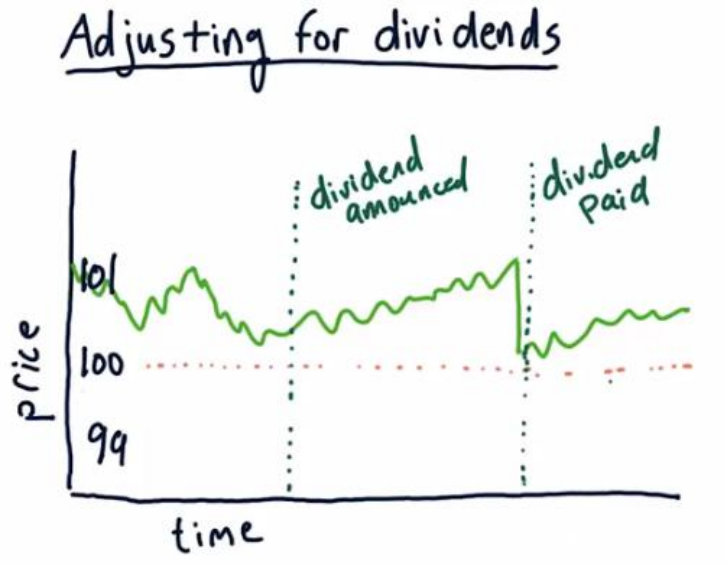


- So answer this as a quiz. Here fill in what would the price of the stock be the day before the dividend, and what the price of the stock be the day the dividend is paid when people get $1.00 and they also keep that share of stock?

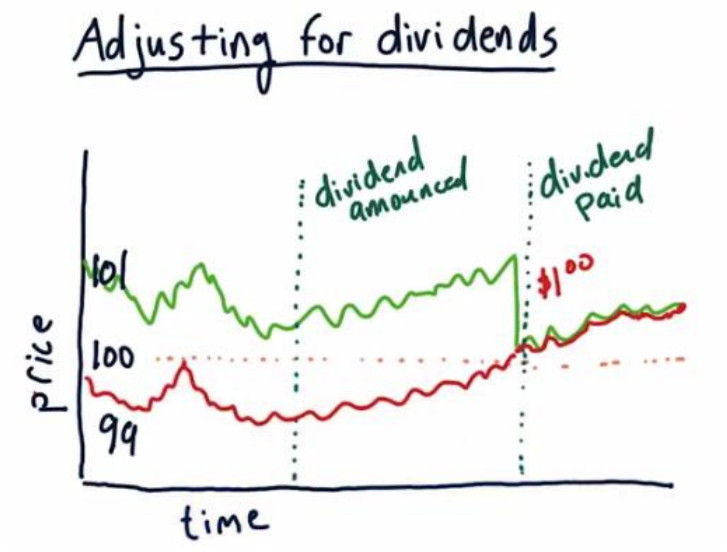


- So, what's going to happen is that the stock price is going to rise to 101 because everybody knows that on the next day, you're going to have one share of a stock worth $100 and $1. So, the total value of that is $101.
- On the very next day, the price is going to drop by $1. And everybody who held a share that day will have one share of stock and $1. So their value continues at this $101 level.
- So the answers are, the day before the price, we should expect to see, is $101. And the day the dividend is paid, we should expect to see $100. And don't forget, you've got your $1 dividend also.

Adjusting for Dividends

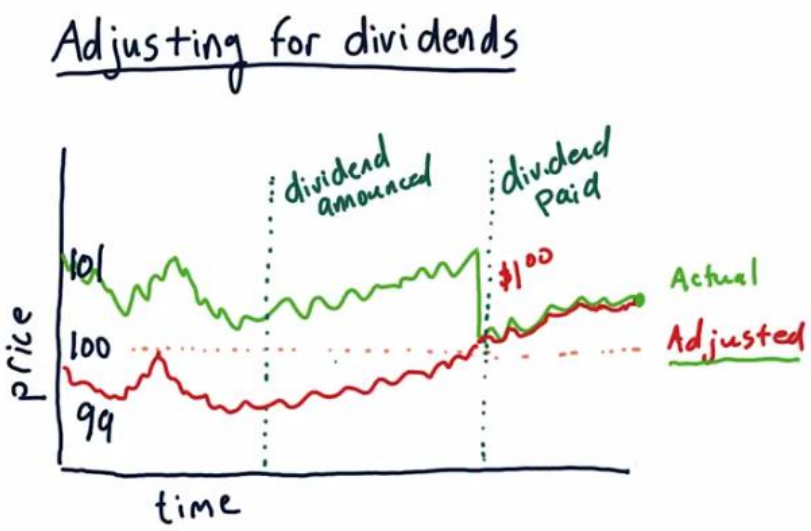


- So here I've redrawn that same scenario. We've got the price of a stock, trundling along. On this date, the dividend is announced that it will be paid on this date. And so we see, in general, the price rising up until that date and then a strong drop.
- And again, the consensus value for the company is about $100. Let's consider now how we might adjust historical prices for this situation.

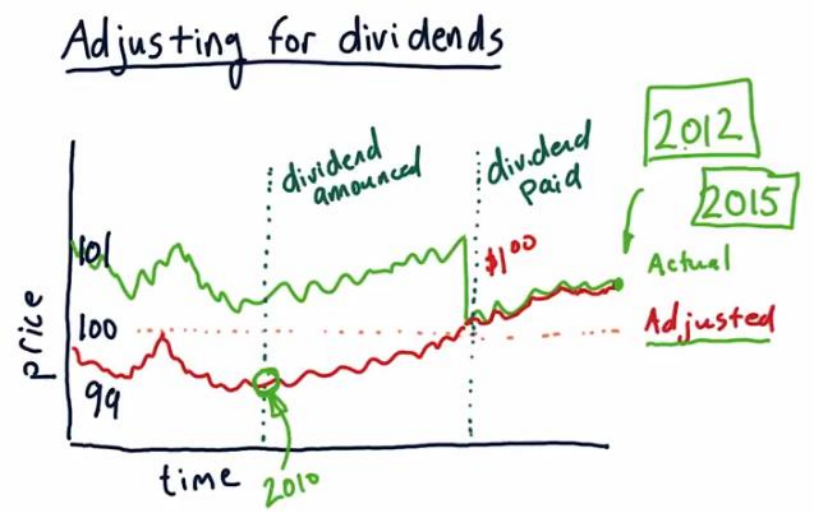


- Again, adjusted price as of today or the latest day in our data, is always the same as the actual price. And as we go back in time, it remains about the same until we hit one of these events like a dividend or a split.
- So we treat, historically, the price in the same way that we do with this split. In other words, just before the dividend is paid, we adjust all of these prices down by the proportion of the

dividend. So in this case we adjust everything down by about 1%, all the way back in history. So our adjusted price continues like this.

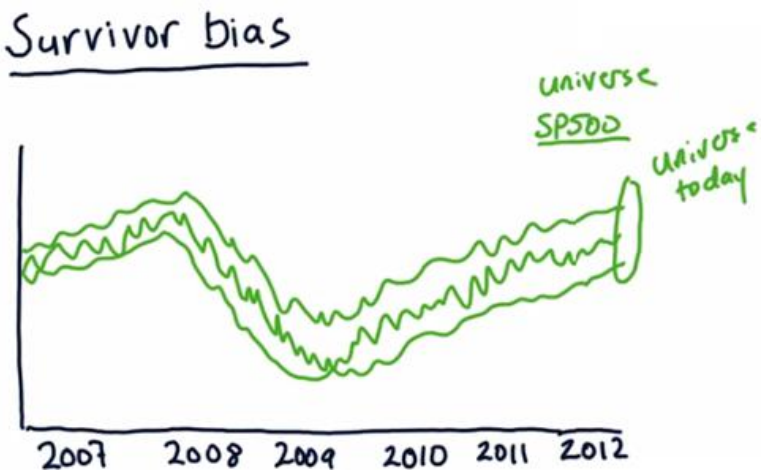


- Now to double check this for rationality, consider that you bought the stock around this date and you held it all the way though the dividend payment you would get an increase in value of your holding of about one dollar.
- So the adjustment is achieving its purpose, which is to allow us to not worry about dividends and splits. Just go back in time, buy, and hold the share, and see how much our value accumulates.
- So for the data we'll be working with in this class, which are daily close values, we have in our data both actual and adjusted. And in almost every instance I will assume that we're using adjusted closed prices in our calculations.
- Another couple things to reiterate about adjustments is the last day in our file, usually that's as of today, the adjusted and actual close value are the same. So any day that you look, say you've gotta go to Yahoo Finance or Google Finance and you get adjusted close and actual close, you'll see that today they're exactly the same price, but as you go back in time they will begin to diverge and the prices we use are adjusted for both dividends and splits, and it's very important that you always use adjusted close.

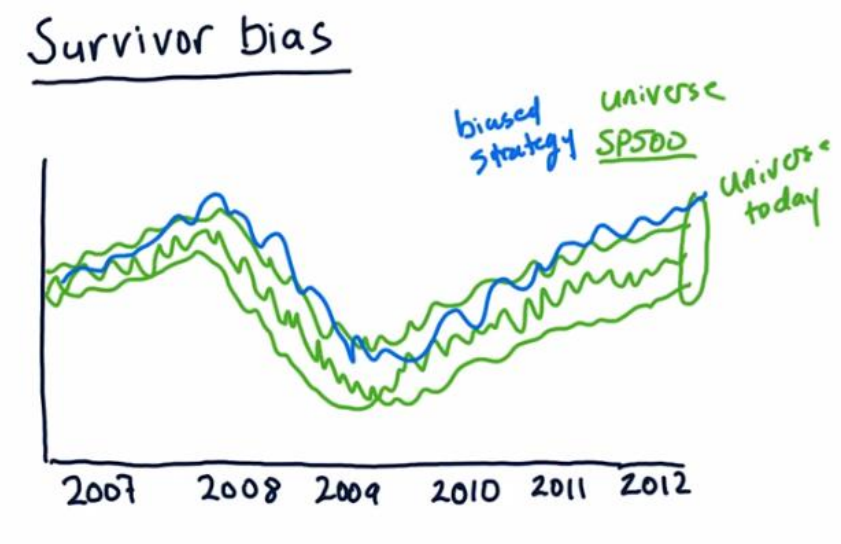


- One other thing to mention is suppose we had this data as of 2012. And so this last date here in our file is some date in 2012, and you go and you look at the adjusted close, let's suppose this is some date in 2010. That value, that adjusted close value, will be different if you gather data from, say, 2015.
- So if you go to Yahoo and get today's adjusted close, there will certainly, for most stocks, there will be dividends and splits that have occurred since 2012. And so this adjusted close price today in Yahoo's data, will be different, most likely lower than the adjusted close for that date, if you had gathered it in 2012.
- I know it's a little bit tricky, the key point is for projects in this class you need to use the data that's provided for the class. It was gathered as of 2012. If you go to yahoo and get new data, you're going to get different answers for the projects, so just keep that in mind.

Survivor Bias

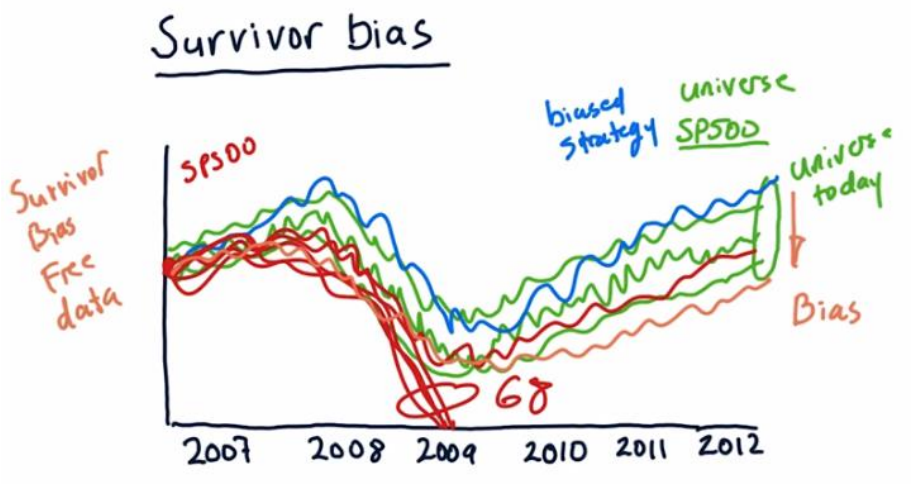


- One of the things we do in this class is to simulate strategies that we might develop. We roll back time, and pretend that we traded on certain dates according to certain signals, and we see what the result of our strategy might have been.
- Now to do that, you have to start with some universal stocks. One of the most common universes is the S&P 500. When you simulate your trading, you roll back time and you look at that universe of stocks. You apply your algorithm to choose which stocks you might buy.
- A very common mistake that people make is that they look at the membership of that universe, as of today. Then they go back in time and they use that list of stocks for their strategy.



- So let's suppose we go back in time, we take the current list of the S&P 500, and we run our strategy and our strategy is just doing great. I'm going to call that the biased strategy.

- So why is it biased? Well, we're selecting from stocks way back here that we knew were existing over here, so there's a built in bias that these stocks are going to do well because they weathered the storm here.



- Now consider, what if we use the S&P 500 as it was back in 2007? So a lot of the stocks from back then did just fine, they survived, but a lot didn't. 68 stocks from the S&P 500 died. They completely went away, from 2007 to 2009.
- So if you applied the same strategy that appears to be so awesome, but you use the real members of the S&P 500 from back then, you're probably going to have a significantly lower performance. And the difference between these two is the bias.
- The lesson learned is to use survivor bias free data and that's available from a number of providers. It's not usually free, but it's not necessarily that expensive either. But if you do that, you'll avoid this sort of false optimism for a strategy that you develop.
- Okay, that's it for dealing with data. I hope you found the lesson useful and we'll see you again soon. Bye-bye.