Time Series Analysis

GA DAT3

Agenda

- About Time Series Analysis
- What is Time Series Data
- Common Analysis For Time Series Data

About Time Series Analysis

TIME SERIES ANALYSIS

- In this class, we’ll discuss analyzing data that is changing over time.
- In most of our previous examples, we didn’t care which data points were collected earlier or later than others.
- We made assumptions that the data was *not* changing over time.
- This class will focus on statistics around data that is changing over time and how to measure that change.

TIME SERIES ANALYSIS

- In this lesson, we will focus on Identifying problems related to time series.

- Additionally, we will discuss the unique aspects of Mining and Refining time series data.

What is Time Series Data?

WHAT IS TIME SERIES DATA?

- Time series data is any data where the individual data points change over time.
- This is fairly common in sales and other business cases where data would likely change according to seasons and trends.
- Time series data is also useful for studying social phenomena. For instance, there is statistically more crime in the summer, which is a seasonal trend.

WHAT IS TIME SERIES DATA?

- Most datasets are likely to have an important time component, but typically we assume that it's fairly minimal.
- For example, if we were analyzing salaries in an industry, it's clear that salaries shift over time and vary with the economic period.
- However, if we are examining the problem on a smaller scale (e.g. 3-5 years), the effect of time on salaries is much smaller than other factors, like industry or position.

WHAT IS TIME SERIES DATA?

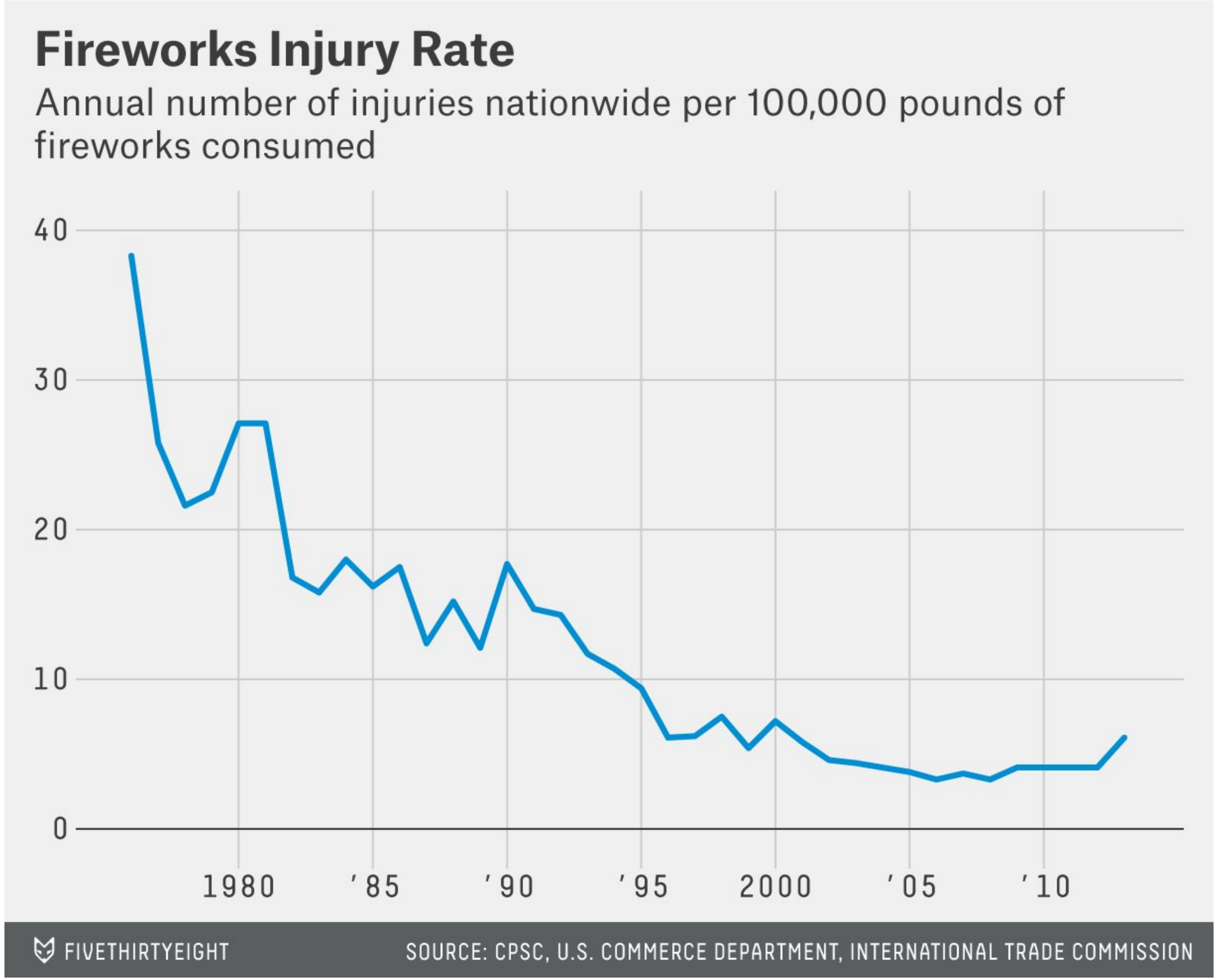
- When the time component *is* important, we need to focus on identifying the aspects of the data that are influenced by time and those that aren't.
- Typically, time series data will be a sequence of values. We will be interested in studying the changes to this series and how related individual values are.
- For example, how much does this week's sales affect next week's? How much does today's stock price affect tomorrow's?

WHAT IS TIME SERIES DATA?

- Time series analysis is useful in many fields: sales analysis, stock market trends, studying economic phenomena, social science problems, etc.
- Typically, we are interested in separating the effects of time into two components:
  - Trends - significant increases or decreases over time
  - Seasonality - regularly repeating increases or decreases

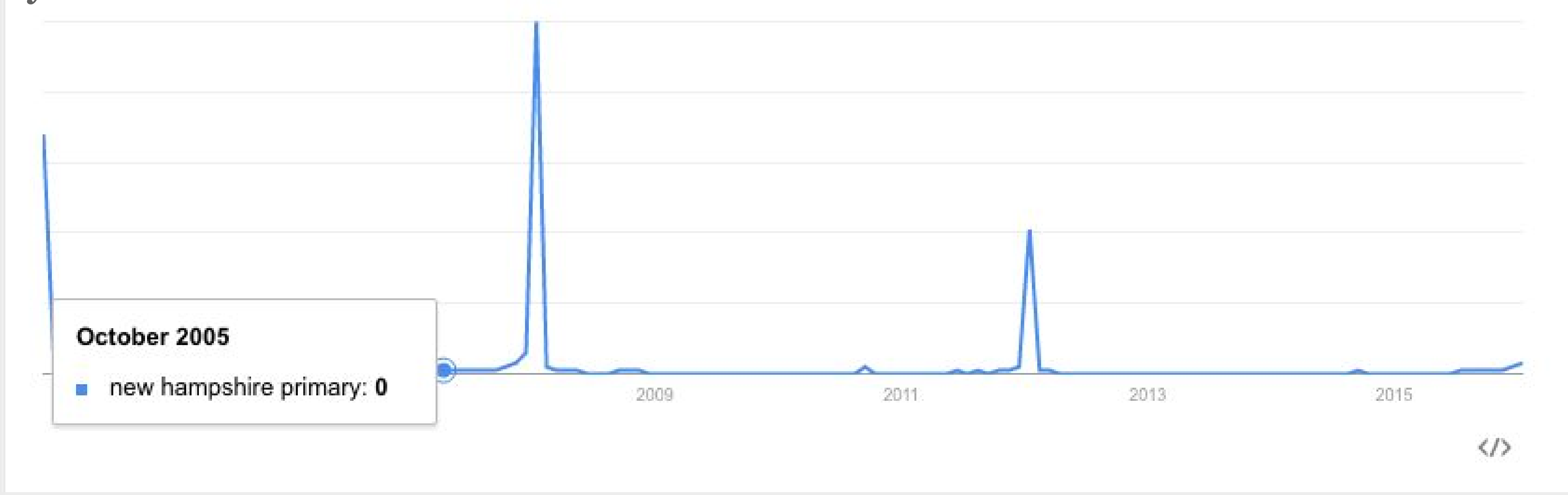
WHAT IS TIME SERIES DATA?

- This plot of fireworks injury rates has an overall *trend* of fewer injuries with no *seasonal* pattern.



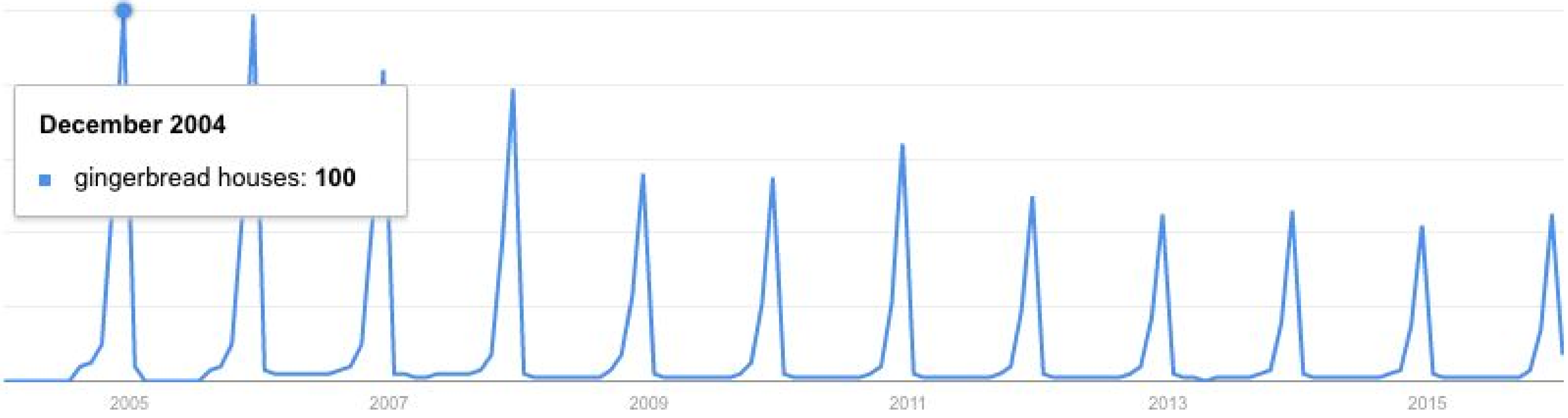
WHAT IS TIME SERIES DATA?

- Meanwhile, the number of searches for the New Hampshire Primary has a clear *seasonal* component - it peaks every four years and on election years.



WHAT IS TIME SERIES DATA?

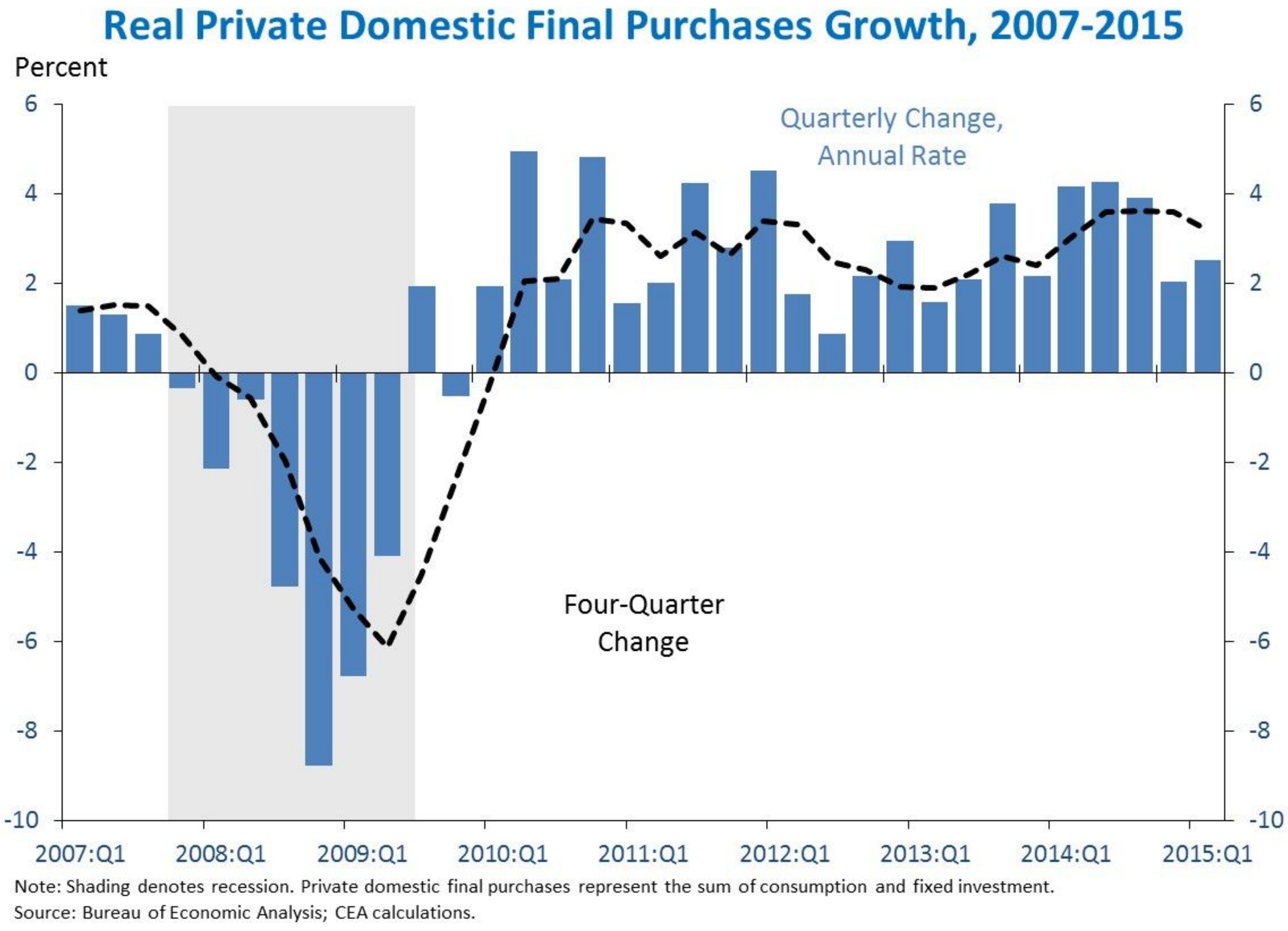
- Similarly, searches for ‘gingerbread houses’ spike every year around the holiday season



- These spikes recur on a fixed time-scale, making them *seasonal* patterns.

WHAT IS TIME SERIES DATA?

- Many other types of regularly occurring up or down swings may occur without a fixed timescale or *period* (e.g. growth vs. recession for economic trends).



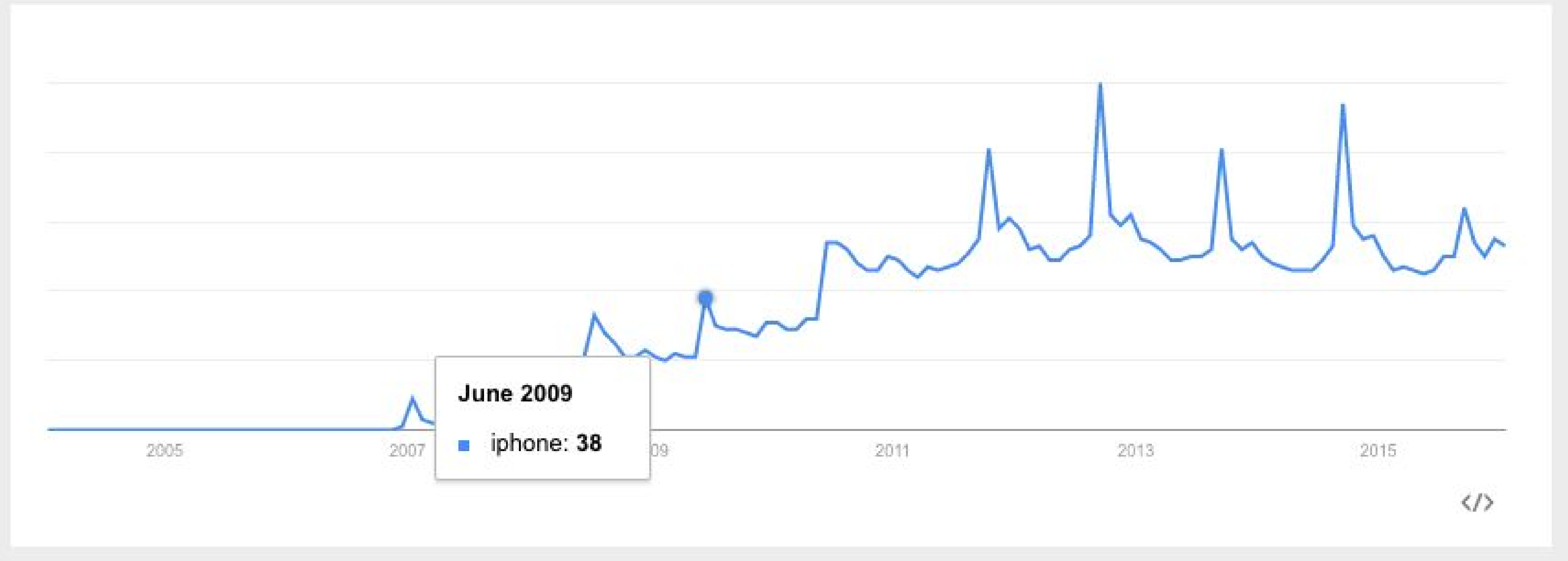
WHAT IS TIME SERIES DATA?

- These aperiodic patterns are called *cycles*.

- While identifying aperiodic cycles is important, they are often treated differently than seasonal effects. Seasonal effects are useful for their consistency, since prior data is useful as a predictor.

WHAT IS TIME SERIES DATA?

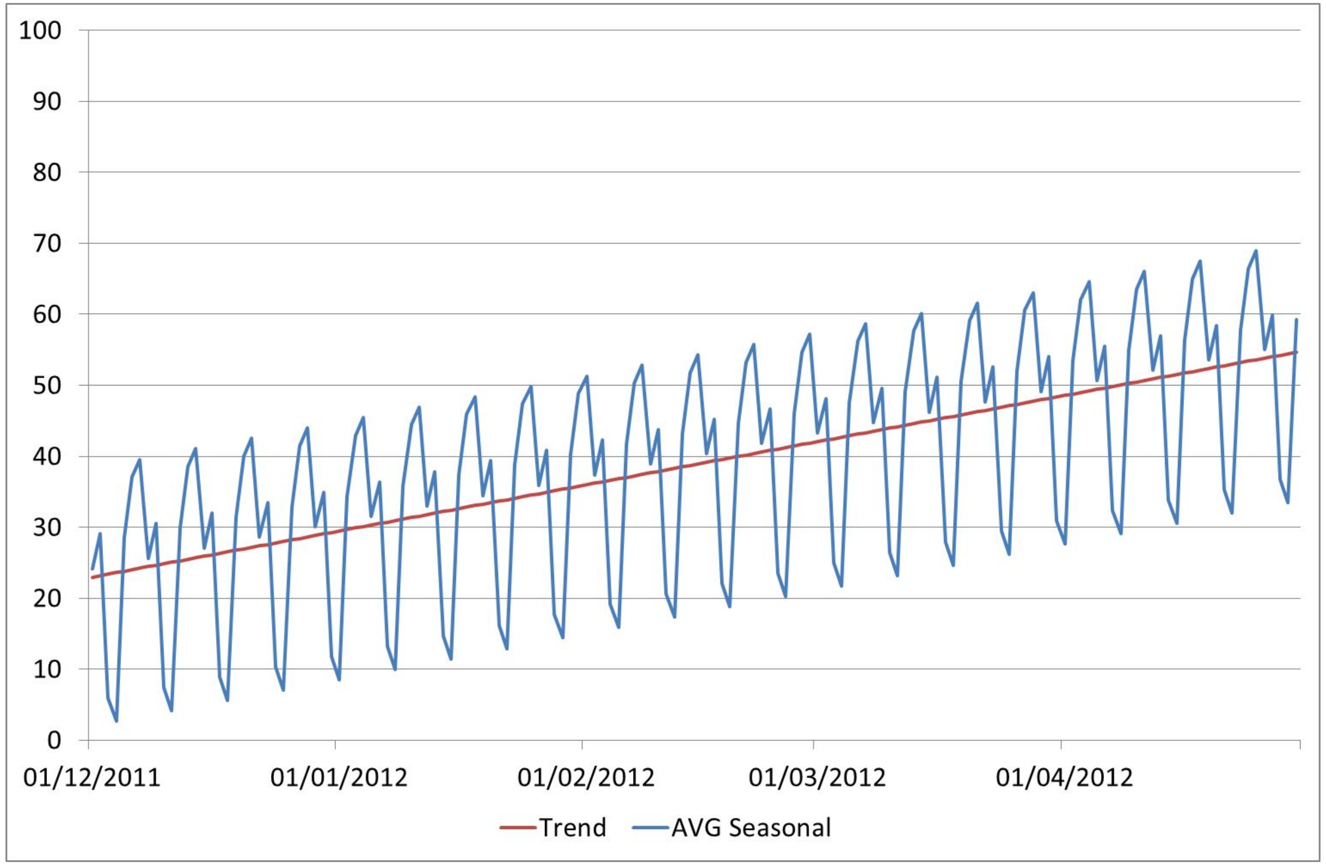
- Searches for “iphone” have both a general trend upwards (indicating more popularity for the phone) as well as a seasonal spike in September (which is when Apple typically announces new versions).



WHAT IS TIME SERIES DATA?

- Most often, we're interested in studying the *trend* and not the *seasonal* fluctuations.

- Therefore it is important to identify whether we think a change is due to an ongoing trend or seasonal change.



Common Analysis For Time Series Data

MOVING AVERAGES

- A *moving average* replaces each data point with an average of *k* consecutive data points in time.
- Typically, this is *k/2* data points prior to and following a given time point, but it could also be the *k* preceding points.
- These are often referred to as the “rolling” average.
- The measure of average could be mean or median.
- The formula for the rolling *mean* is

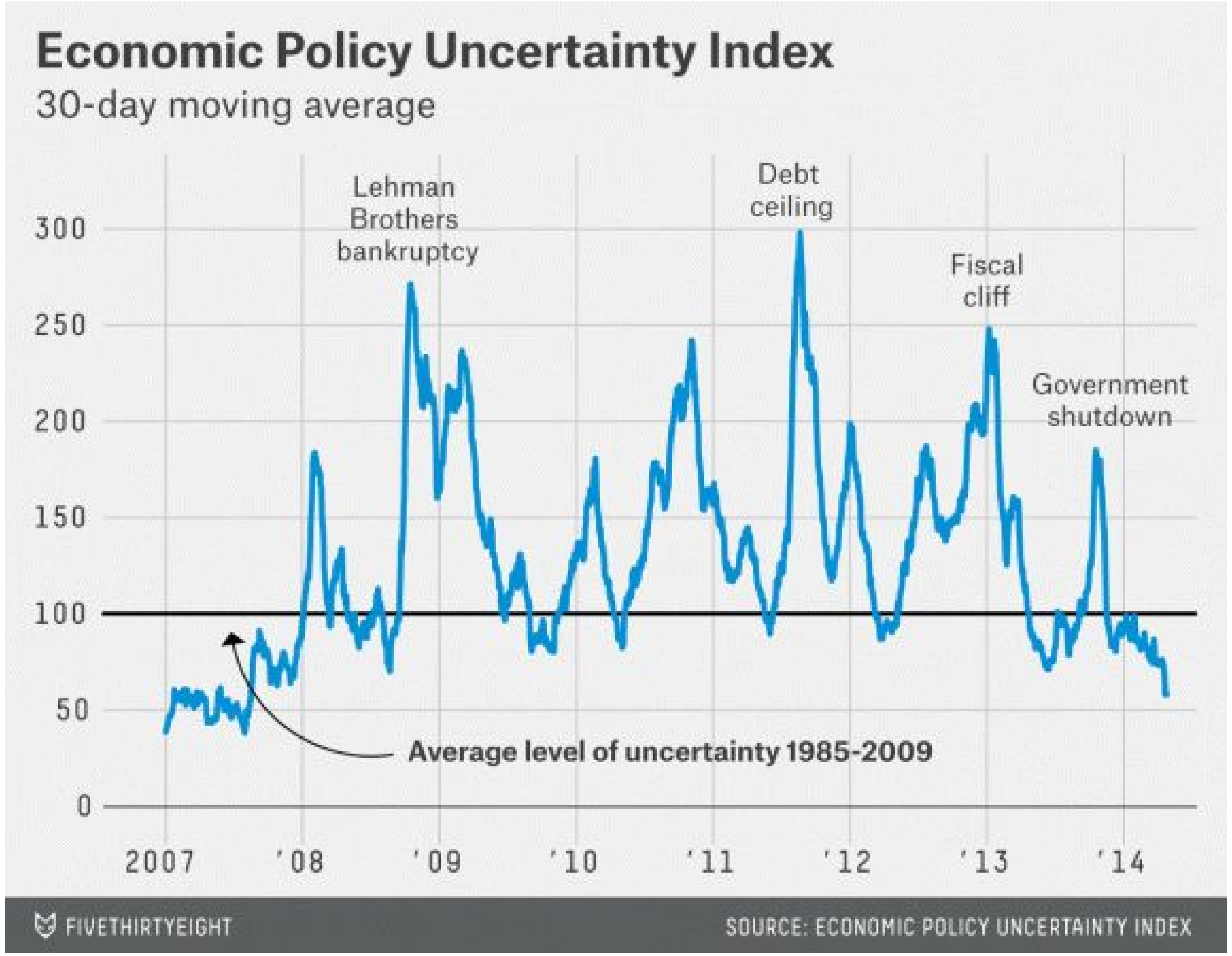
$$F_t = \frac{1}{p} \sum_{k=t}^{t-p+1} Y_k$$

MOVING AVERAGES

- A rolling mean would average all values in the window, but can be skewed by outliers (extremely small or large values).
- This may be useful if we are looking to identify atypical periods or we want to evaluate these odd periods.
- For example, this would be useful if we are trying to identify particularly successful or unsuccessful sales days.
- The rolling median would provide the 50 percentile value for the period and would possibly be more representative of a “typical” day.

WHAT IS TIME SERIES DATA?

- This plot shows the 30-day moving average of the Economic Uncertainty Index.
- Plotting the moving average allows us to more easily visualize trends by smoothing out random fluctuations and removing outliers.



MOVING AVERAGES

- While this statistic weights all data evenly, it may make sense to weight data closer to our date of interest higher.
- We do this by taking a *weighted moving average*, where we assign particular weights to certain time points.
- Various formulas or schemes can be used to weight the data points.

MOVING AVERAGES

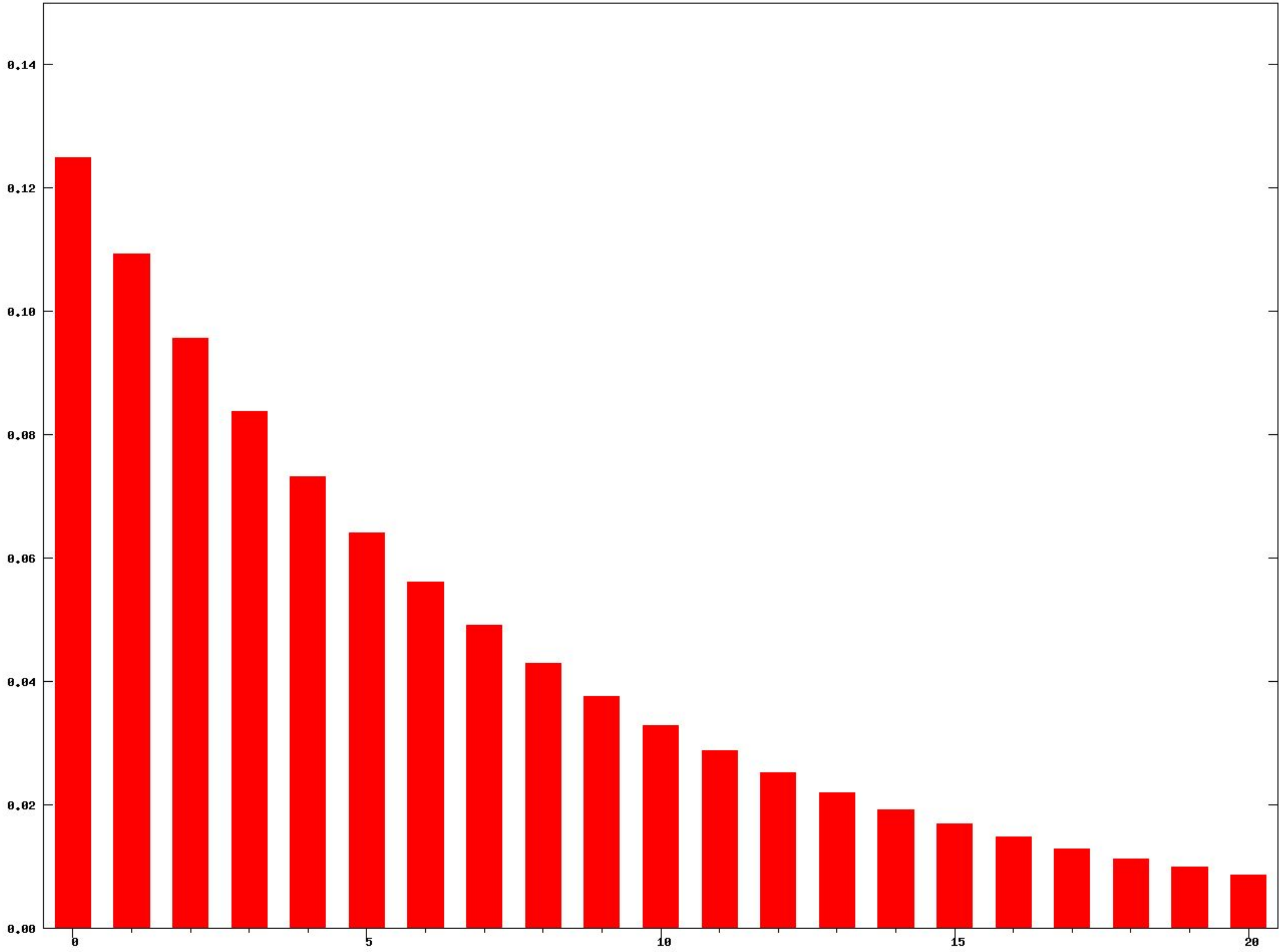
- A common weighting scheme is an *exponential weighted moving average (EWMA)* where we add a *decay* term to give less and less weight to older data points.
- The EWMA can be calculated recursively for a series Y.

For $t = 1$, $EWMA_1 = Y_1$

For $t > 1$, $EWMA_t = \alpha \bullet Y_t + (1 - \alpha) \bullet EWMA_{t-1}$

MOVING AVERAGES

- The weights for an exponential weighted moving average with k = 15.



AUTOCORRELATION

- In previous classes, we have been concerned with how two variables are correlated (e.g. height and weight, education and salary).

- *Autocorrelation* is how correlated a variable is with itself. Specifically, how related are variables earlier in time with variables later in time.

AUTOCORRELATION

- To compute autocorrelation, we fix a “lag” k. This is how many time points earlier we should use to compute the correlation.

- A lag of 1 computes how correlated a value is with the prior one. A lag of 10 computes how correlated a value is with one 10 time points earlier.

AUTOCORRELATION

- The following formula can be used to calculate autocorrelation.

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

Conclusion

CONCLUSION

- We use time series analysis to identify changes in values over time.
- We want to identify whether changes are true trends or seasonal changes.
- Rolling means give us a local statistic of an average in time, smoothing out random fluctuations and removing outliers.
- Autocorrelations are a measure of how much a data point is dependent on previous data points.

Q??