# Agenda:

- model for detecting changes in a stream of data collected over time.

- outliers, and the need for data preparation.

**Sub-courses**

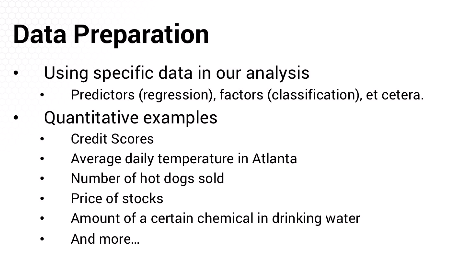
1. **Basic data preparation**
2. **Change detection**

**Links**

<http://charuaggarwal.net/outlierbook.pdf>.

<https://support.sas.com/documentation/onlinedoc/qc/132/cusum.pdf>

# 1.1.Basic preparation



* Factor value Xij is exact value of such predictors

1. Sometimes, we need to manipulate data

* Like, scaling



Different factors with different scaling can lead algorithm in wrong path

1. Sometimes, data has extraneous information that can complicate model or ability to correctly interpret results
2. Data prep step is covered in this topic(before the data is consumed for modelling)

# 1.2 Outlier detection (Finding Outliers)

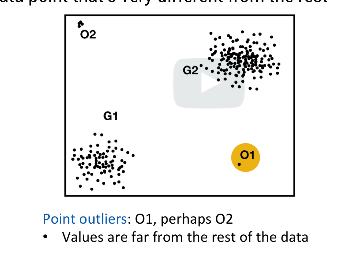
**Agenda:**

What is outlier?

How to detect Outliers

**Outliers :**

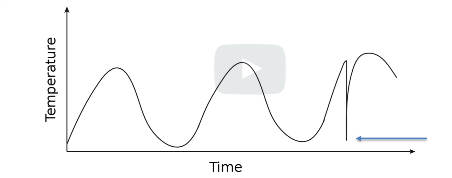
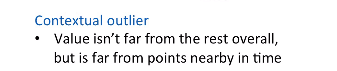
* philosophical and statistical issues
* data point that is different from others
* “O1” is “Point Outlier”



* “O2” is also outlier..really outlier? Or reason to be there?

**Types of Outliers**

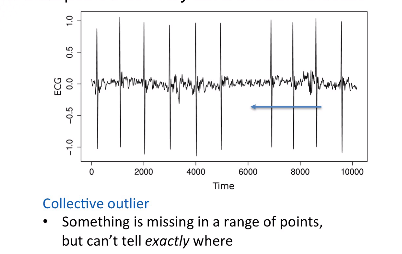
1. **Contextual Outlier**

**** 

-time series data

-temperature point itself is not a outlier; but the time it occurs is a outlier. Called “Contextual Outlier”

**2) Outlier by Omission/Collective Outlier**

****

* Expected large beep in ECG around 6000, but it is not there
* There is no wrong, but between 5000-7000, there is something missing
* The data points collectively seems to be an issue
* If we take time between beats, we expect a beat every 1000 beats, but in this case, the difference is 2000 millisecond.

**Finding Outlier in Automated way**

1. Box & Whiskers ; for 1 dimension

Top & bottom – 20% and 75 percentile of the bar;

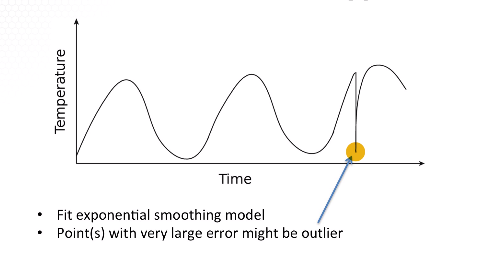
Middle – 50% median

Vertical lines are whiskers (to reason range of values) –sometimes 5 or 75% or 10 or 90%

Beyond that whiskers is possible outliers

\*Not good option to detect outlier for multi dimension

1. Build a model; fit the parameters; the point with most errors are possible outliers
2. E.g in graph below, we can fit an exponential smoothing model.

* It is a smooth model.
* At each point, it will fit the actual data point to smooth model.
* All points will be smooth, except the drop , when it has highest error. The model expects a peak and the actual is a dip.
* 

# 1.3 Dealing with Outliers

**Agenda:**

* what to do when we see outlier

**why outlier?**

-bad data ;

**Examples:**sensor malfunction;

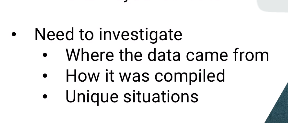
experiment is contaminated;

Human error with input and error

- outliers may be real ;then decide if this is required or not

Need investigation

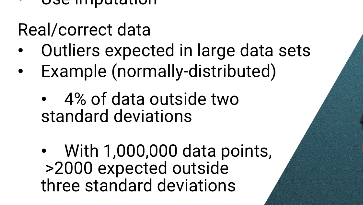
**Factors for investigation**

****

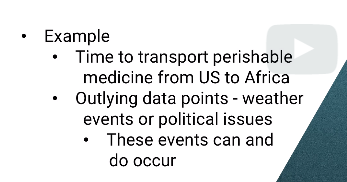
**How to handle?**

1. Bad data –
   1. OMIT
   2. Use data imputation
2. real correct outlier
   1. think about system
   2. there is enough randomness, that it has some outlying values
   3. **example**: if normally distributed randomness, then more than 4% of data will be more two standard deviations .

for larger data set, it will be worser

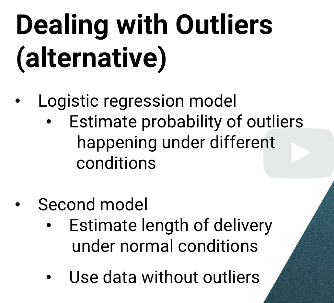


\*important to consider if magnitude of the model’ error is part of the measure of its vaue. In that case, it is important to keep the outliers in data set



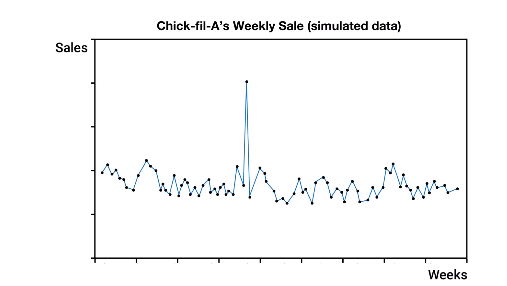
In above example, considering these outliers is important

1. Dealing with Outliers



\*Sometimes, unpredictable outliers though real data, needs to be removed from model.

Example: chik –fil-a –weekly sales data- expect one week –huge peak



* Assumed bad dab. But it was true data point;

(It turns out that a Chick-fil-A executiv had said something that turned out to be controversial

and some people called for a boycott of Chick-fil- as a protest, and in response, other people decided

they would eat at Chick-fil-A as a counter-protest, and overall there are a lot more extra people eating

at Chick-fil-A that week than there were boycotters **not eating there, so there was a huge spike in the data.**

**Resolution**: removed data point; always good to investigate data[[1]](#footnote-1)

# 2.1 Introduction to Change detection

-determinse something changed

-time series data

**Examples**

Type-1

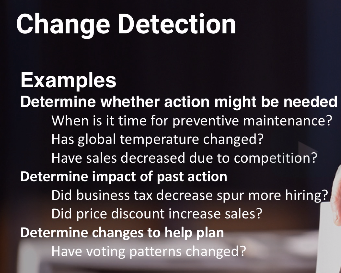
1. Manufactorue : machine showing variation is random or needs maintainence?
2. Global temperature has changed?

Type-2

1. If some change has a impact or not? Promotion impacted sales or not?
2. Business tax decrease more hiring?

Type-3

1. Difference over time plan or analyze
2. Changes in voting pattern

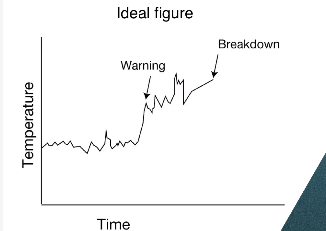
**or**

**Own example#1**

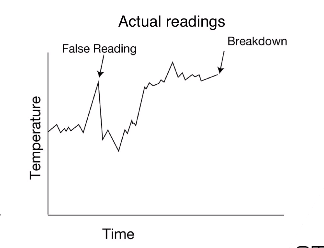
1. Semiconductor planning- cannot be done with human due to risk of dirt
2. So machine run overhead machine to machine (200 machine(
3. Chanllenges : traffic (central control to route the machine(
4. Required: will there be traffic congestion after 5 minutes even there is no traffic now

**Own example#2**

* Railroad axles fail and lead to derail
* Railroad sensors detect axle temperature to indicate possible maitanance
* when temperature increased and needs action to change



But actual data



* **Challgenges:**
* **1)** used hypotheses testing .but they are slow to detect

We cannot wait too long before we can change the axle.

**So , We have to accept some false positives**

# 2.2 CUSUM for Change detection

-Cumulative Sum

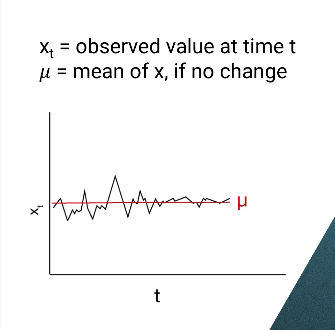
-has the mean of the observed distribution gone beyond a critical threshold

-can detect if process gets to a higher level & beyond ; lower level & before

**Example:**

For example

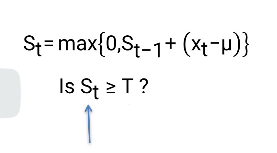
1. x sub t might be the price of a stock **on day t,**
2. **the measure temperature of a rail car axle**

****

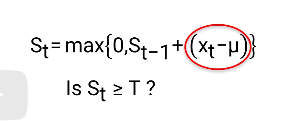
For each observervation,

Xt-Mu =>how much above expected the observation is at time “t”

**CUSUM:** obJective is to find “S(t)” and if it is greater than a threshold “T”

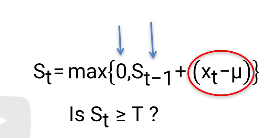
****

For each time, we get the observed value at time t and how far it goes above the mean(mu)

****

S(t-1) is the metrics for the previous time period to give running total

If running total is greater than 0, we keep it or replace it with “0”

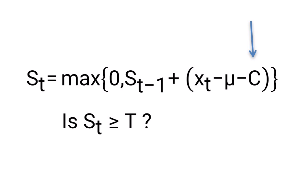


This helps CUSUM detect changes quickly

Challenge:

1. On the other hand, we don't want to be too sensitive and say there's a change when there really isn't.
2. We expect there to be some randomness, sometimes in fact maybe about half the time, X sub t will be higher than the expectation just at random.

So , we use “C”- it can be confidence interval to pull down the running total value



Meaning, if “C” is bigger, it is harder for S(t) to get bigger and it is less sensitive

Smaller the “c”, more sensitive the system is

**How to choose value for “C”?**

“C” and “T-Threshold” are model parameter that we have to use data to find the correct “C”

* factors to consider : how costly it is if the model takes a long time to notice a change and how costly it is if model found a change that is NOT there

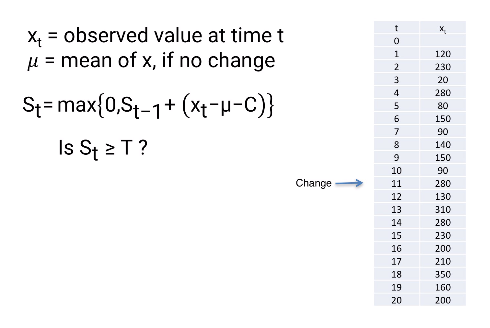
**example:**

1. in computer chip manafucturing, if a machine starts putting a component a little too far to the left, then each chip produced between when the chain starts and when it's detected will have to be discarded

which means if “model” detects change too soon , it is costly.

1. In rail road, if model predicts too late, then it is costly.

We have to check the results for different “C” value

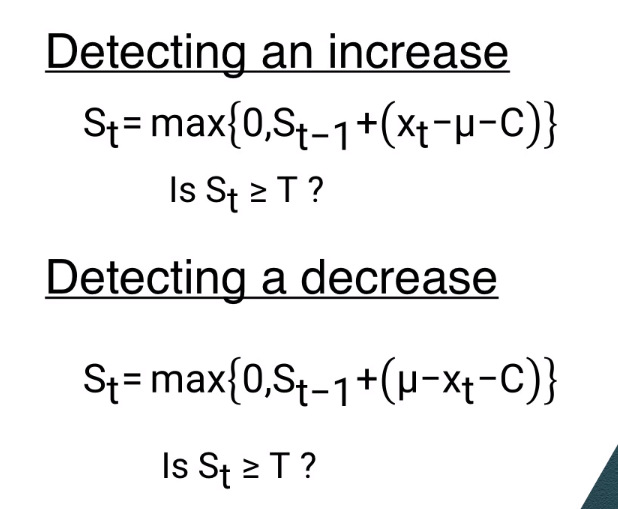


question : whether someone's reaction time slows down

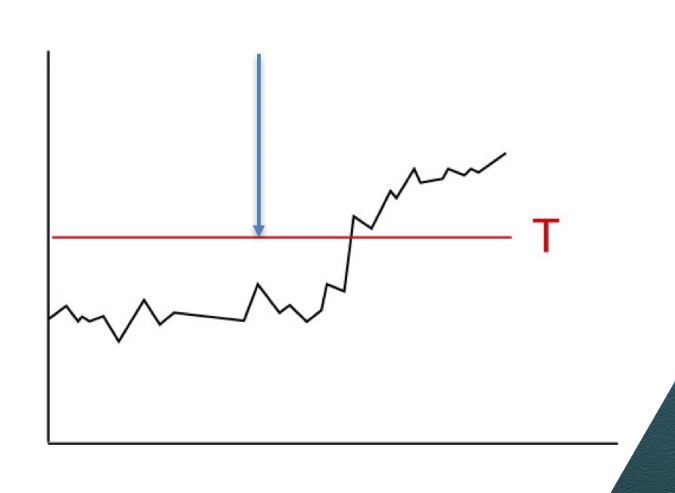
**what was done?** to test the CUSUM method, we artificially added 100 to every observation from number t equals 11 onward and checked to see how long it takes CUSUM to notice that there is a change.

Threshold “T” :450

C=0



Best way to visualize the change detection is through **control chart**



Tradeoff : early detection and false detection

1. [↑](#footnote-ref-1)