EEP 596: Adv Intro ML || Lecture 12

Dr. Karthik Mohan

Univ. of Washington, Seattle

February 15, 2023

Last Time

- Anomaly Detection Use-Cases

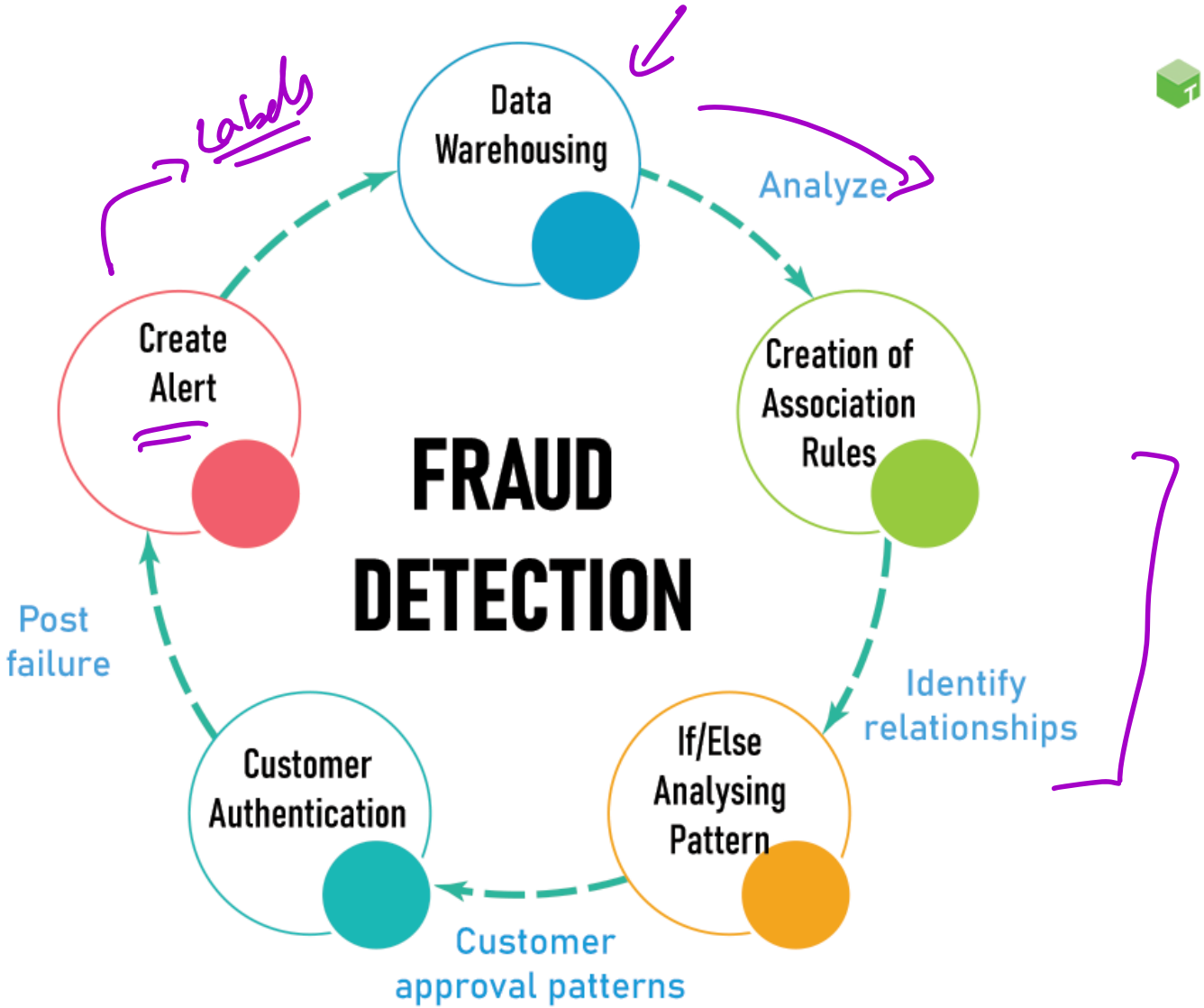
Last Time

a. Anomaly Detection Use-Cases
b. Anomaly Detection Baselines

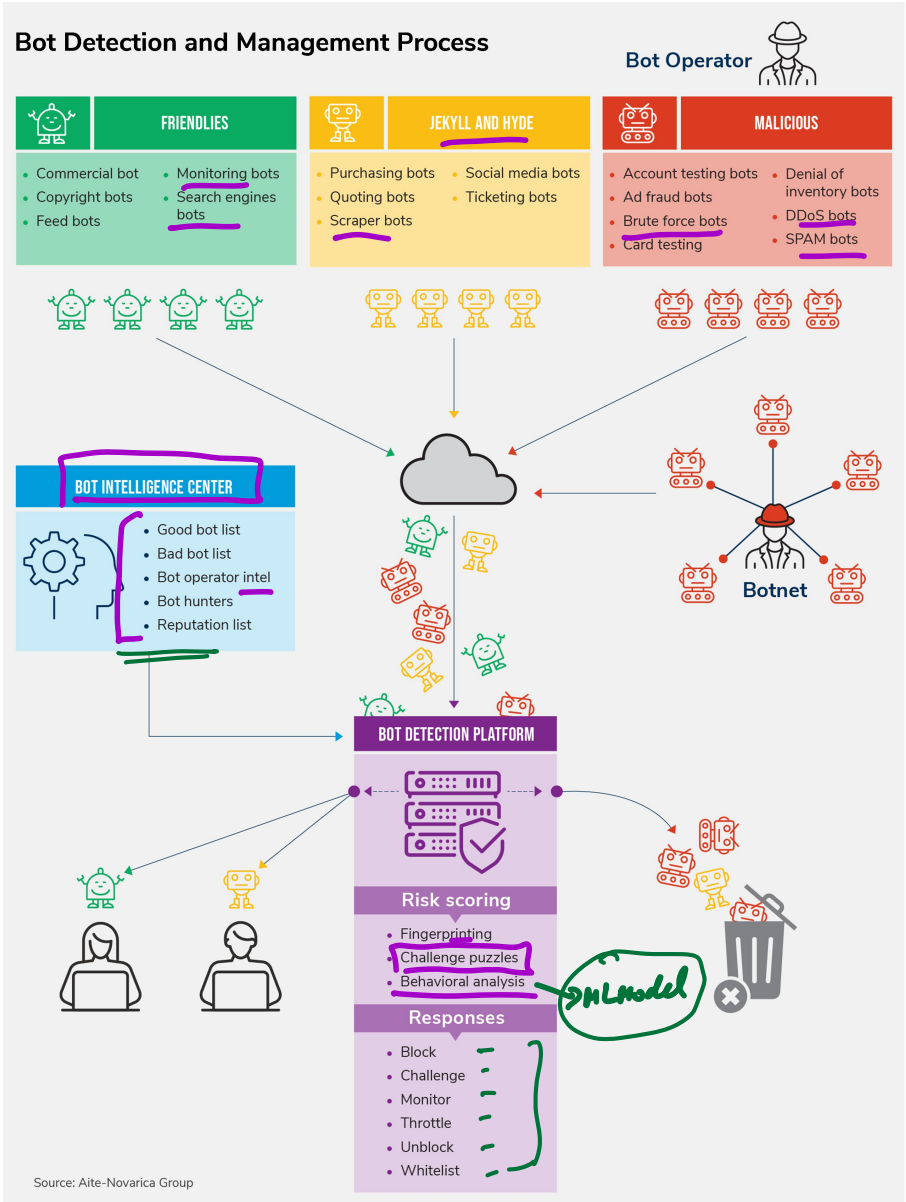
Today

- **More Baselines** for anomaly detection
- GMMs
- Anomaly Detection for Time-series
- **STL model** for anomaly detection

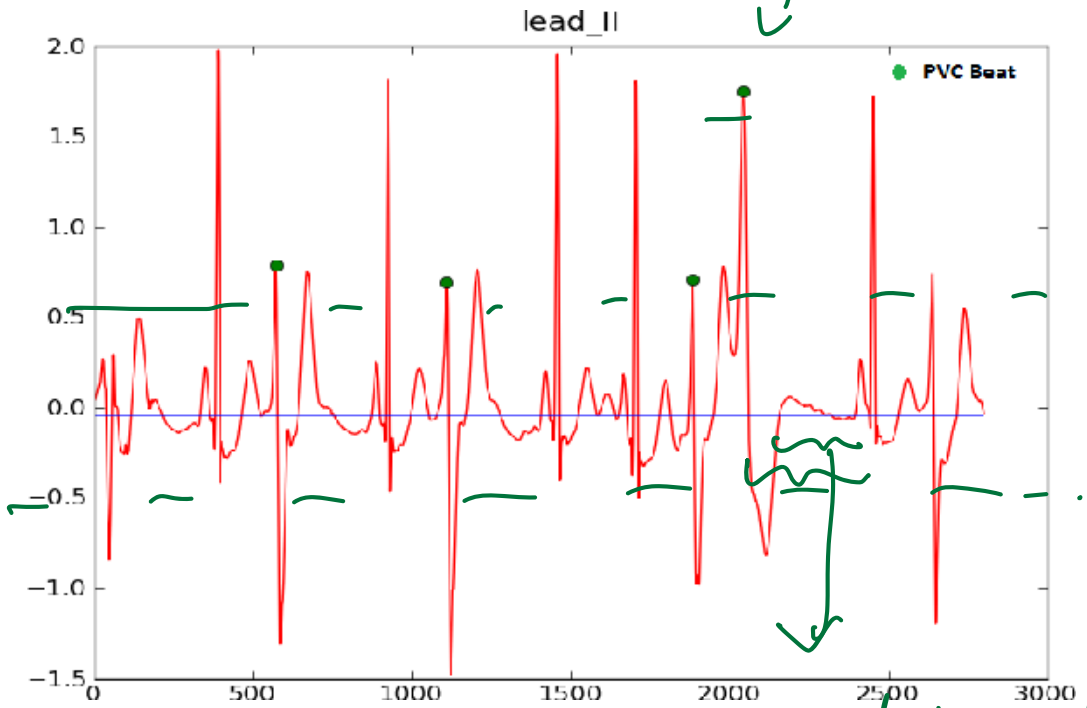
Fraud Detection



Got Bot?

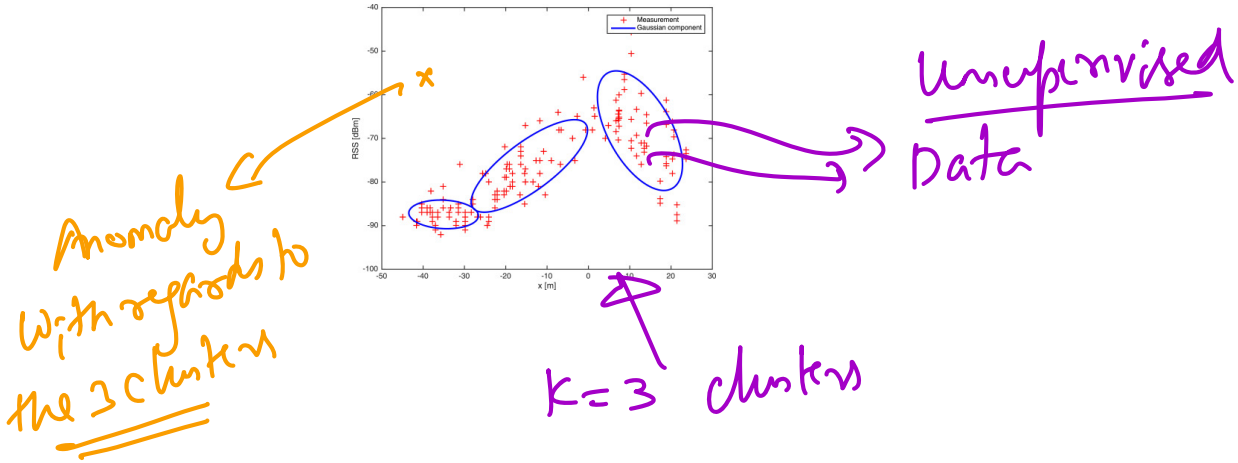

Bot Detection and Management Process
Bot Operator
FRIENDLIES
Commercial bot
Copyright bots
Feed bots
Monitoring bots
Search engines bots
JEKYLL AND HYDE
Purchasing bots
Quoting bots
Scraper bots
Social media bots
Ticketing bots
MALICIOUS
Account testing bots
Ad fraud bots
Brute force bots
Card testing
Denial of inventory bots
DDoS bots
SPAM bots
BOT INTELLIGENCE CENTER
Good bot list
Bad bot list
Bot operator intel
Bot hunters
Reputation list
Botnet
BOT DETECTION PLATFORM
Risk scoring
Fingerprinting
Challenge puzzles
Behavioral analysis
DNLmodel
Responses
Block
Challenge
Monitor
Throttle
Unblock
Whitelist
Source: Aite-Novarica Group

Arrhythmia Detection



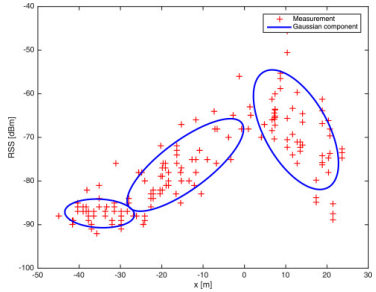
Types of Anomalies

1. **Point Anomaly:** Deviation from a set of data points.

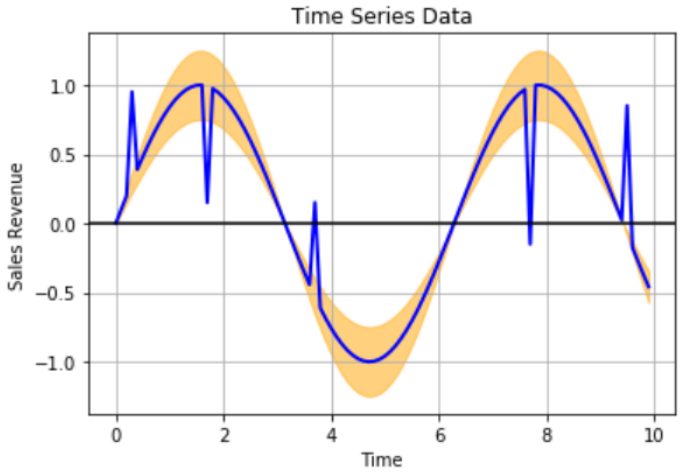


Types of Anomalies

1. **Point Anomaly:** Deviation from a set of data points.

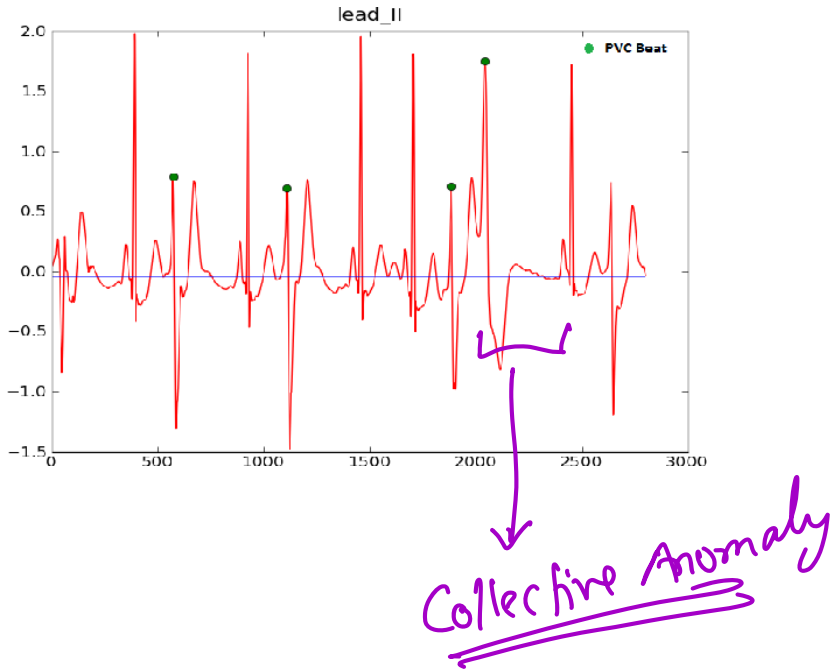


2. **Contextual Anomaly:** Depending on the context, a data point could be an anomaly or not. For instance 35 degrees is not an anomalous temperature for Seattle winter but it is for Seattle summer. Same is true for anomalies in a time-series data e.g. Sales Revenue data.



Types of Anomalies

3. **Collective Anomalies:** No one data point is anomalous but a collection of them become anomalous. E.g. the Arrhythmia time series.



ICE #1

Anomaly Type

You are tasked with detecting if a particular stock is trending downwards or upwards. So far its been trending downwards. You need to identify if at the end of next week, it is still trending downwards or upwards. This is an example of:

a. Point anomaly detection
b. Contextual anomaly detection
c. Collective anomaly detection
d. Not an anomaly detection problem

ICE #2

Anomaly Type

You are tasked with identify which clusters of servers are under-utilized based on cpu-utilization rates and re-purpose them as needed. This is an example of:

a. Point anomaly detection
b. Contextual anomaly detection
c. Collective anomaly detection
d. Not an anomaly detection problem

Broadly Speaking

1. **Un-Supervised Methods:** Clustering based, statistical methods (mean, standard deviation based alarming), etc
2. **Supervised Methods:** Requires enough labels for positives and negatives

Anomaly Detection Baselines

Baseline Algorithm

Classify a data point as an anomaly if your data point is α standard deviations (σ) away from the mean. Here α is typically greater than 3.

cpu-utilization @ 4PM $< \overline{\text{mean utilization}} - 3\sigma$
$\Rightarrow$ under-utilized

Anomaly Detection Baselines

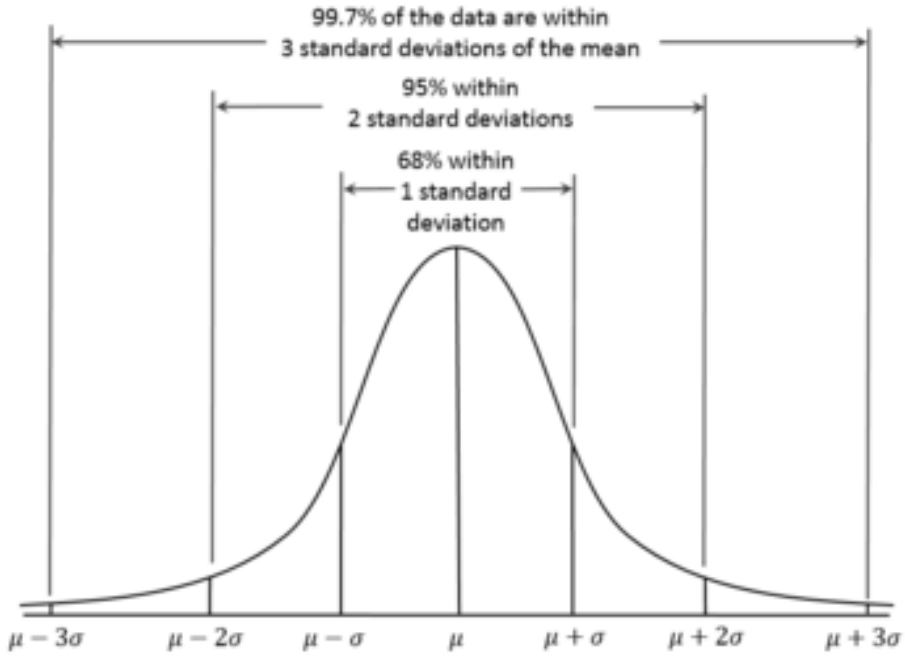
Temperature Example

The mean human body temperature is 98.4 F. Assume now that the thermometer is accurate but normal body temperature fluctuations are expected to be within 0.5 degrees F, then there is cause for concern if temperature deviates beyond 98.4 ± 1.5 for $\alpha = 3$.

Anomaly Detection Baselines

Temperature Example

The mean human body temperature is 98.4 F. Assume now that the thermometer is accurate but normal body temperature fluctuations are expected to be within 0.5 degrees F, then there is cause for concern if temperature deviates beyond 98.4 ± 1.5 for $\alpha = 3$.



Anomaly Detection Baselines

Temperature Example

	α	Outcome
1	2	Lots of false positives and un-necessary trips to urgent care
2	6	Almost no false positives. Might miss early signs of a flu
3	4	Fewer false positives. Get to urgent care at the right time!

Anomaly Detection Baselines

Faulty thermometer

Assume you only have a faulty thermometer to measure your body temperature. Sometimes its accurate and sometimes it is not. You get to know that its reading can fluctate up to 3 degrees over or below the true temperature. You measure your temperature and its 102 degrees F. Should you head to the urgent care?

Anomaly Detection Baselines

Faulty thermometer

Assume you only have a faulty thermometer to measure your body temperature. Sometimes its accurate and sometimes it is not. You get to know that its reading can fluctate up to 3 degrees over or below the true temperature. You measure your temperature and its 102 degrees F. Should you head to the urgent care?

False positives vs False Negatives

Anomaly Detection methods get caught between controlling false positives and not missing True positives (i.e. having false negatives). Would you rather flag a social media post as inappropriate and capture 90% of inappropriate posts but cause a stir with remaining 10% users or would you rather capture 80% of inappropriate posts and have zero false positives?

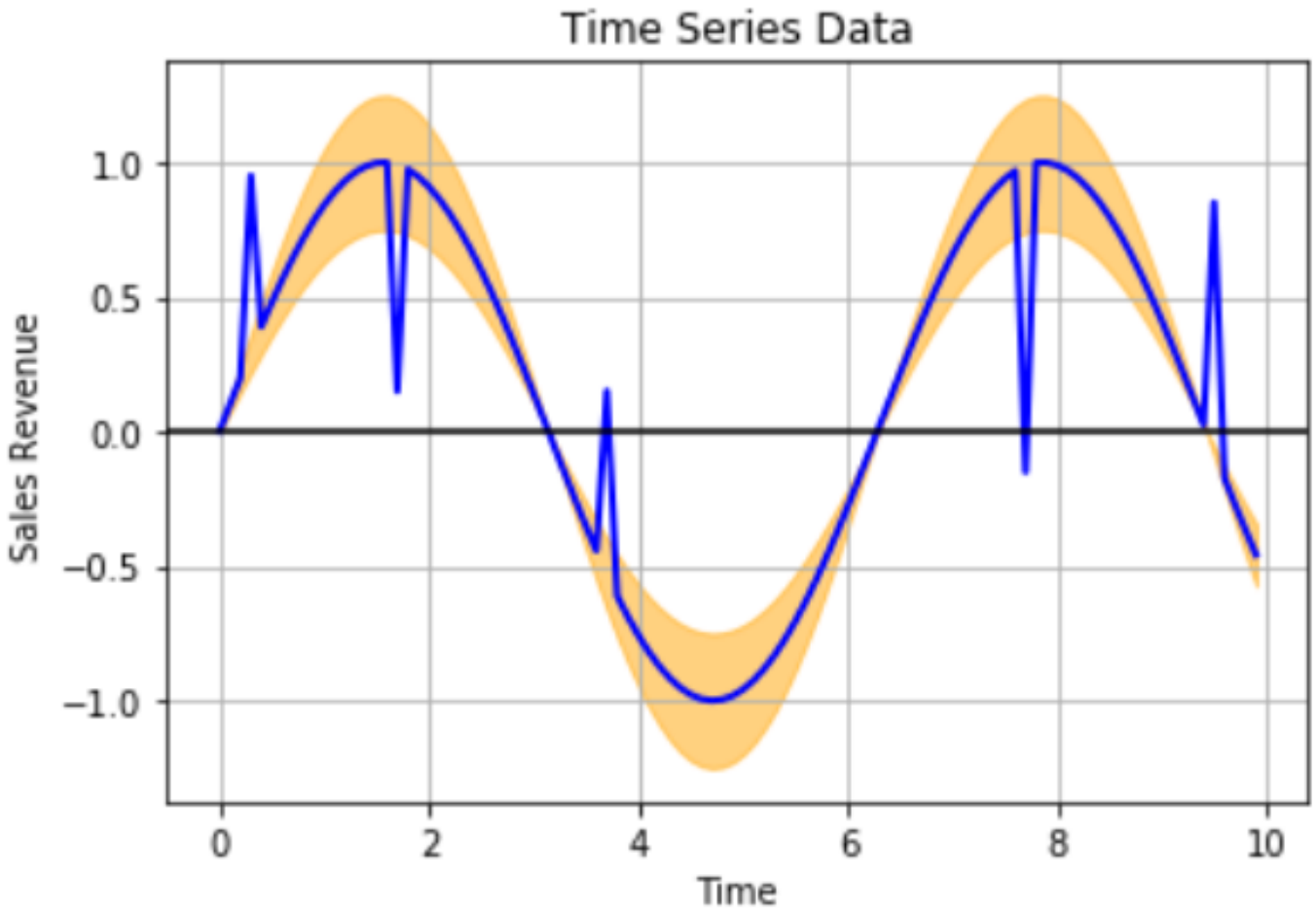
Detecting bots/in-appropriate posts

Biasing the metrics

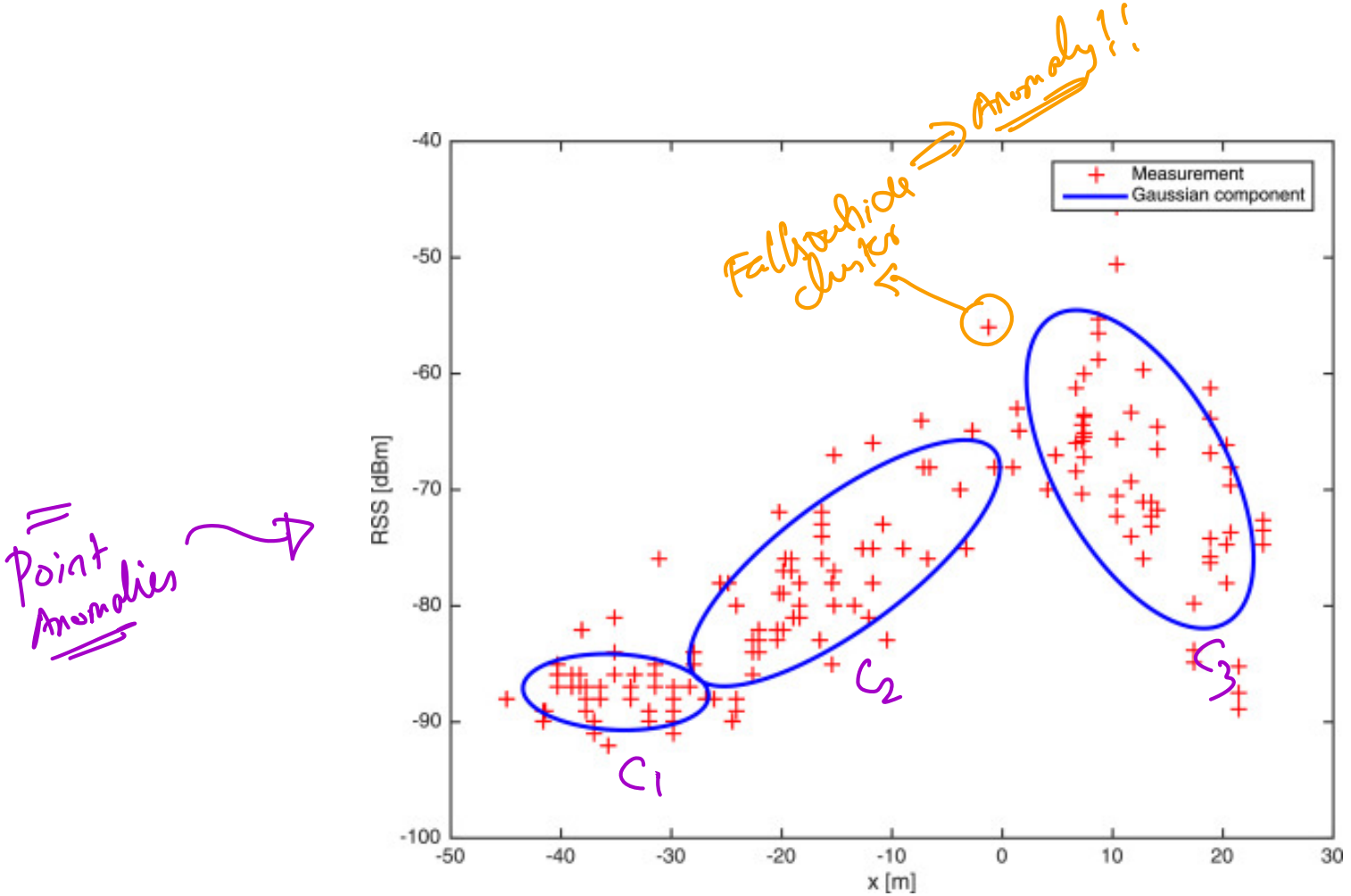
Do you bias more towards high precision or high recall? Is there a middle ground? Can we have higher recall (i.e. detect the bots/in-appopriate posts) without pissing people off with incorrect flags?

1. Human Bias
2. Privacy Concerns

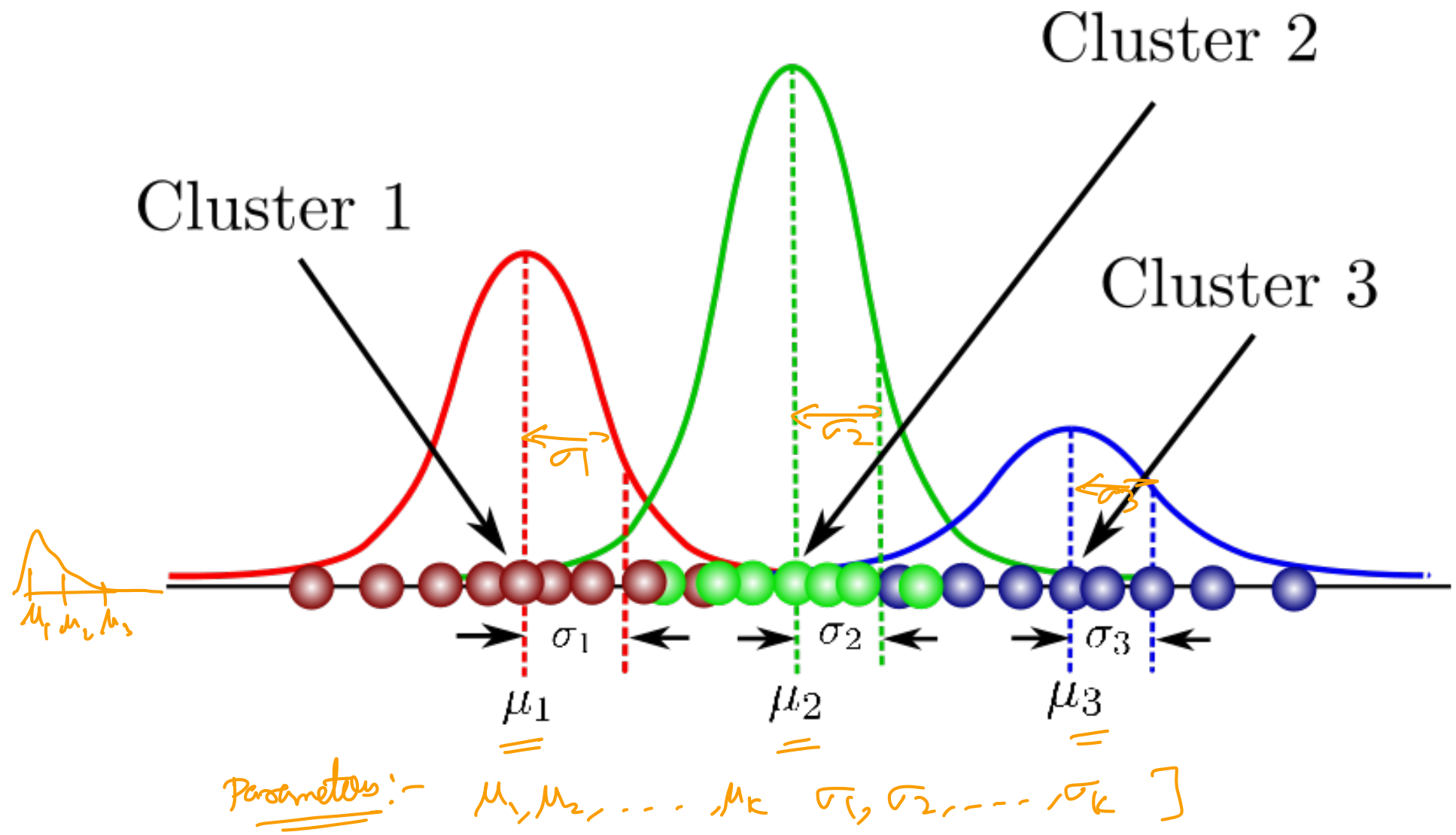
Time series anomalies



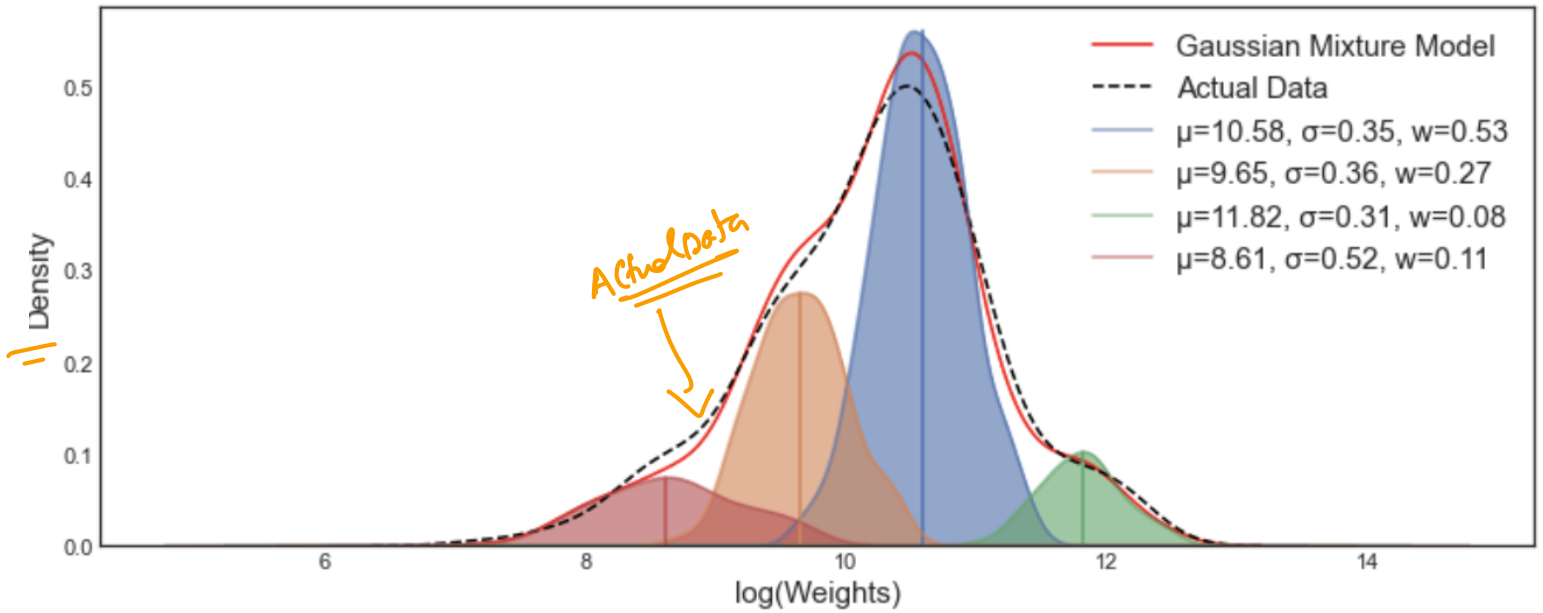
GMM - Better than Baseline



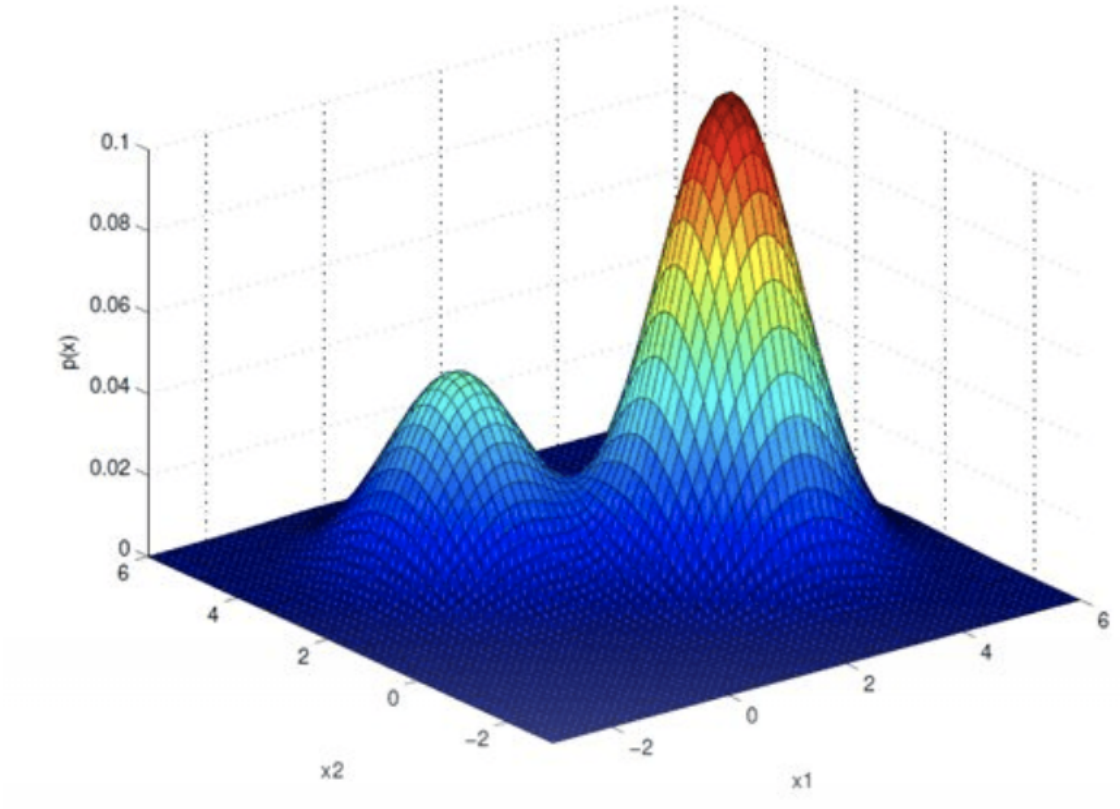
GMM - Better than Baseline



GMM - Better than Baseline



GMM - Better than Baseline



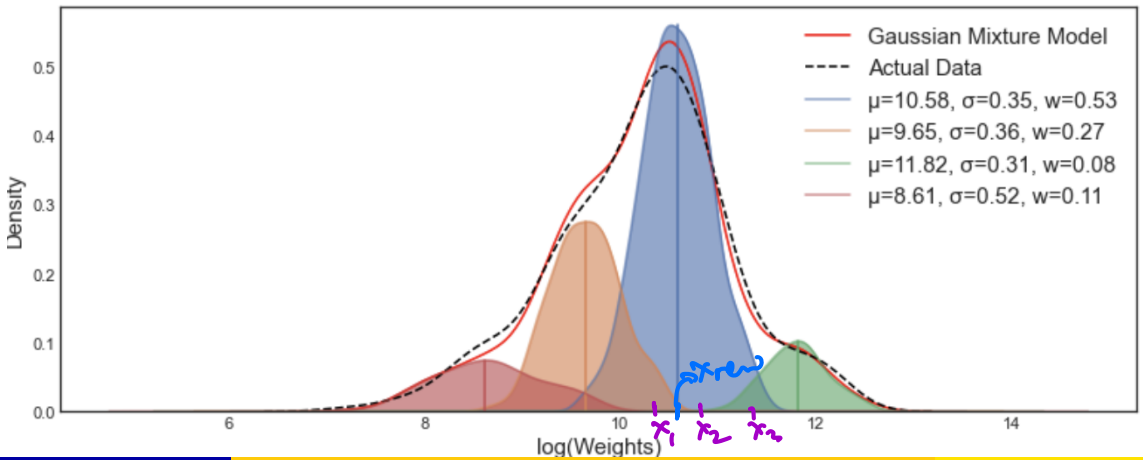
GMM - PDF

PDF

The **Probability Density Function** (PDF) for GMM at any given point x:

$$\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

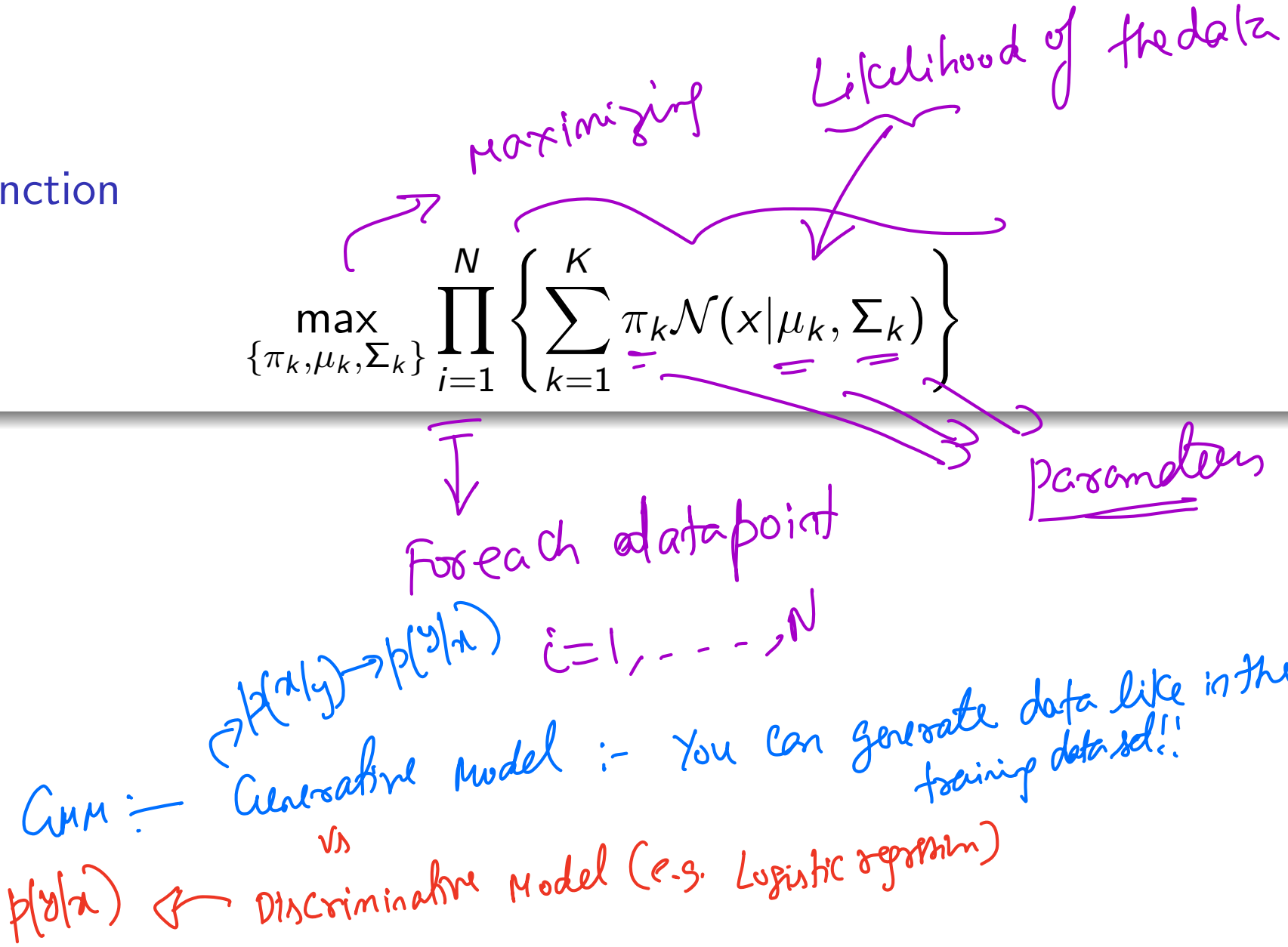
where $\sum_{k=1}^{K} \pi_k = 1$, μ_k are the means/centroids of the k clusters and Σ_k are the co-variances! π_k represents the contribution of cluster k to the overall density.



GMM - Loss function

Loss function

$$\max_{\{\pi_k,\mu_k,\Sigma_k\}} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \right\}$$



GMM - Loss function

~~Loss function~~

$$\max_{\{\pi_k,\mu_k,\Sigma_k\}} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \right\}$$

Negative Log-likehood loss function for learning GMM

$$\min_{\{\pi_k,\mu_k,\Sigma_k\}} -\sum_{i=1}^{N} \log \left(\left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \right\} \right)$$

Identifying anomalies from GMMs

Good for → Point Anomaly Detection

Algorithm for Anomaly Detection based on GMM

a. **Fit** a GMM model to the data with K clusters. K is a modeling choice! (k-means++ can be used!)

b. For a candidate data point x_j, identify the probability of x_j belonging to each of the clusters - call it p_k.

c. If $\max_k p_k < \alpha$ where α is an **anomaly probability threshold**, then flag x_j as an anomaly.

↓ Hyper-parameter

Identifying anomalies from GMMs

Algorithm for Anomaly Detection based on GMM

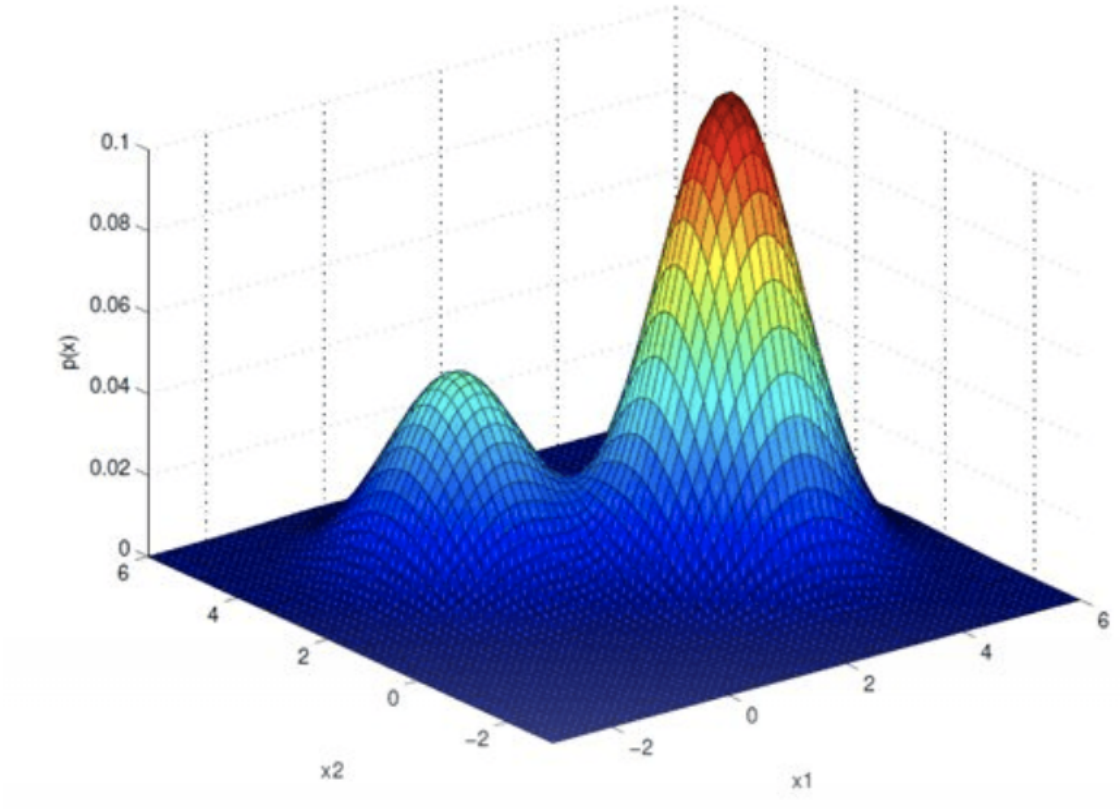
a. **Fit** a GMM model to the data with K clusters. K is a modeling choice!

b. For a candidate data point x_j, identify the probability of x_j belonging to each of the clusters - call it p_k.

c. If $\max_k p_k < \alpha$ where α is an **anomaly probability threshold**, then flag x_j as an anomaly.

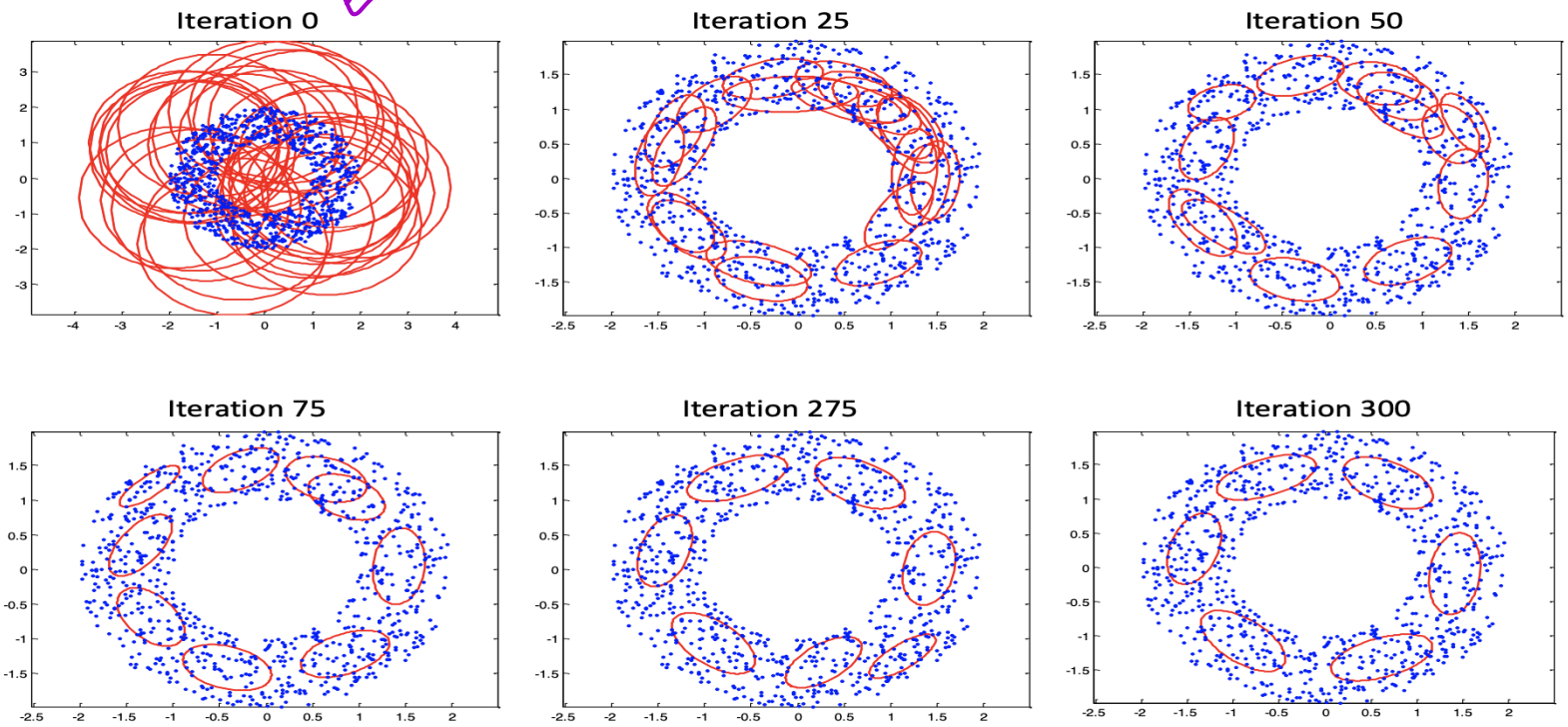
How to compute the probability p_k?

GMM Anomaly Detection Example

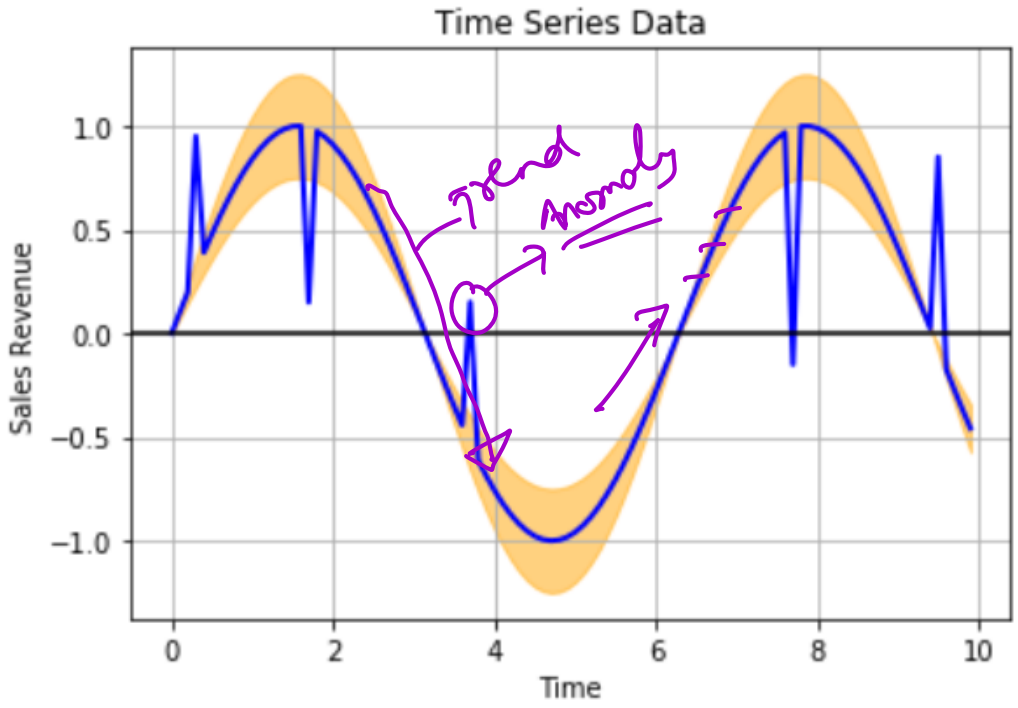


GMM Learning through EM algorithm

Expectation Maximization



Time-series Anomaly Detection



Local window anomalies

Fit a linear model that captures local trends and compute a probability for a new data point being an anomaly w.r.t local model.

Time-series Anomaly Detection

Un-supervised Learning

If we don't have enough labels for anomalies (or positive class), we have no choice but to resort to **un-supervised learning**.

Time-series Anomaly Detection

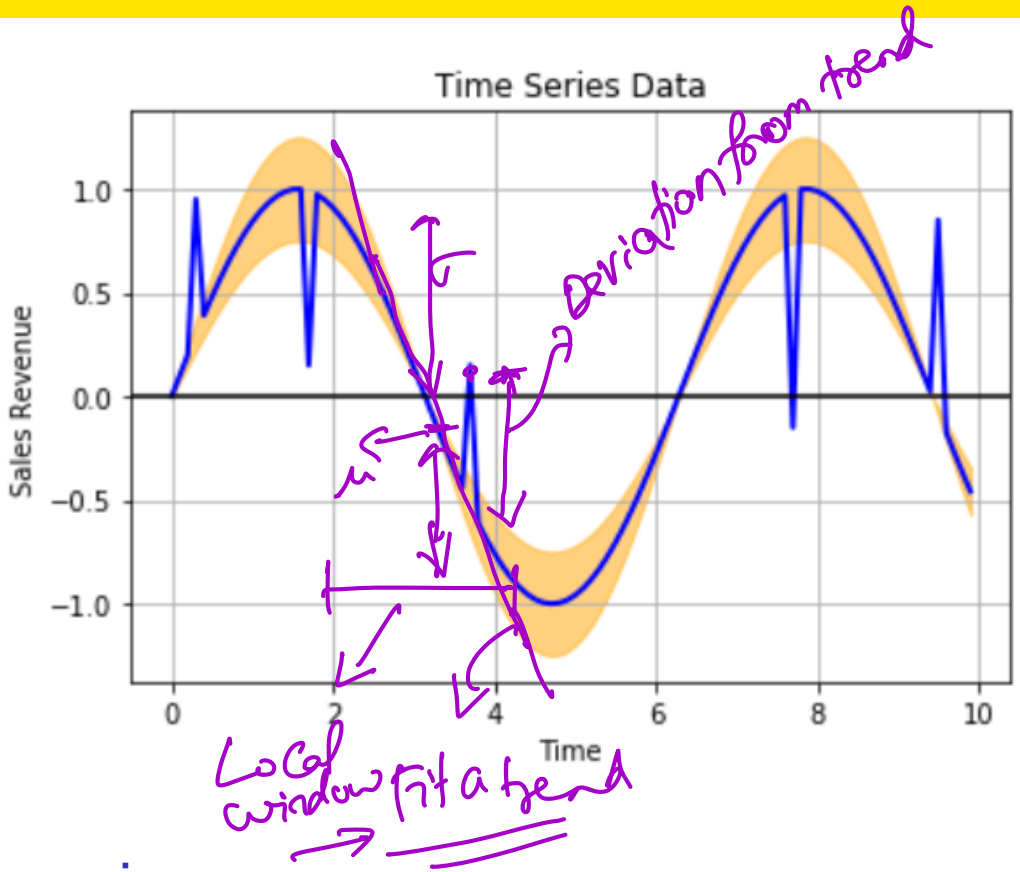
Un-supervised Learning

If we don't have enough labels for anomalies (or positive class), we have no choice but to resort to **un-supervised learning**.

Un-supervised Learning

However, un-supervised learning for anomaly detection is fraught with issues. What are they?

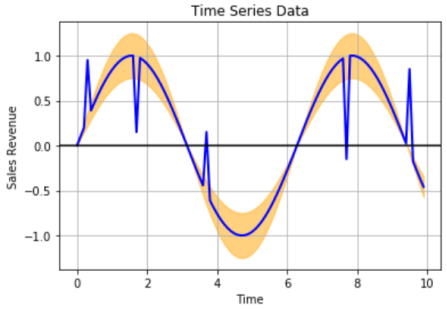
Time-series Anomaly Detection



Semi-supervised Learning

Learn good features from un-supervised learning and use a simple classifier - such as logistic regression model to fine tune the probability computations! **Example features:** Deviation from a local linear regression model fit on a local window. Deviation from median of a local window.

Time-series Anomaly Detection



ICE #3

What are the hyper-parameters for the semi-supervised logistic regression based anomaly detection approach we just described?

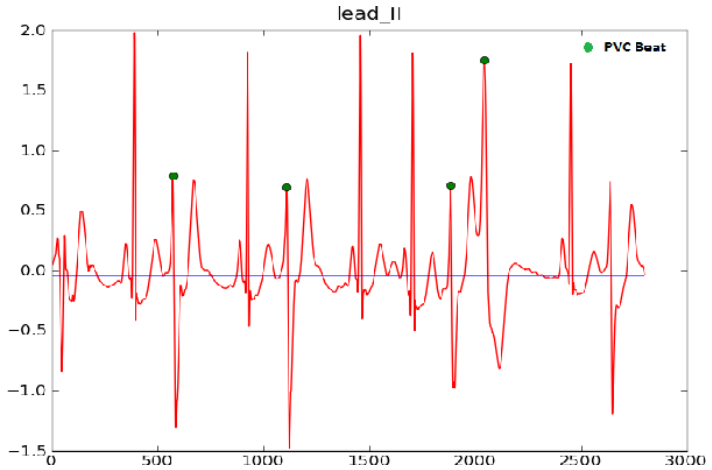
a. The weights for the different features learned from un-supervised learning that are then combined to get a probability prediction from logistic regression

b. The number of (unsupervised learning) features used in the logistic regression

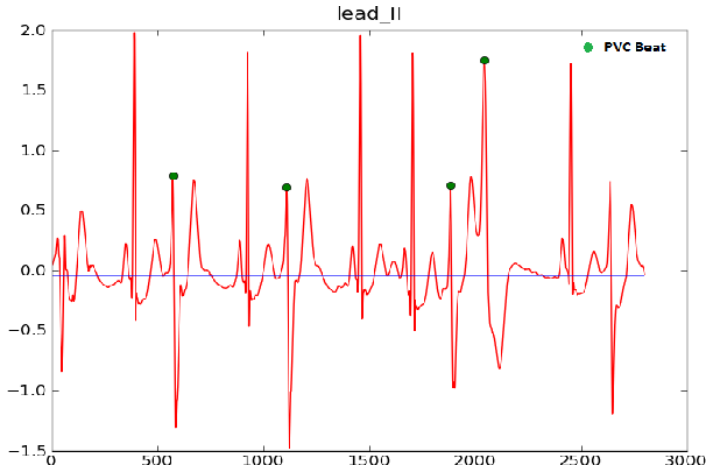
c. The size of the local window used to compute these features

d. The probability of a data point being an anomaly

Arrythmia detection



Arrythmia detection



Next Assignment

ICE #4 Automated Arrhythmia Detection

You want to build an automated algorithm for Arrhythmia detection from time-series data on heart beats. What would be a baseline un-supervised learning algorithm you can think of Arrhythmia detection? If you wanted to do supervised learning for arrhythmia detection, what features would you use? How would you cast it as a machine learning problem? How would you evaluate the performance of your automated algorithm? What would be the metrics you would use? Discuss in groups - We will implement this as part of the next programming assignment.

Deep Learning for Anomaly Detection

Deep Learning

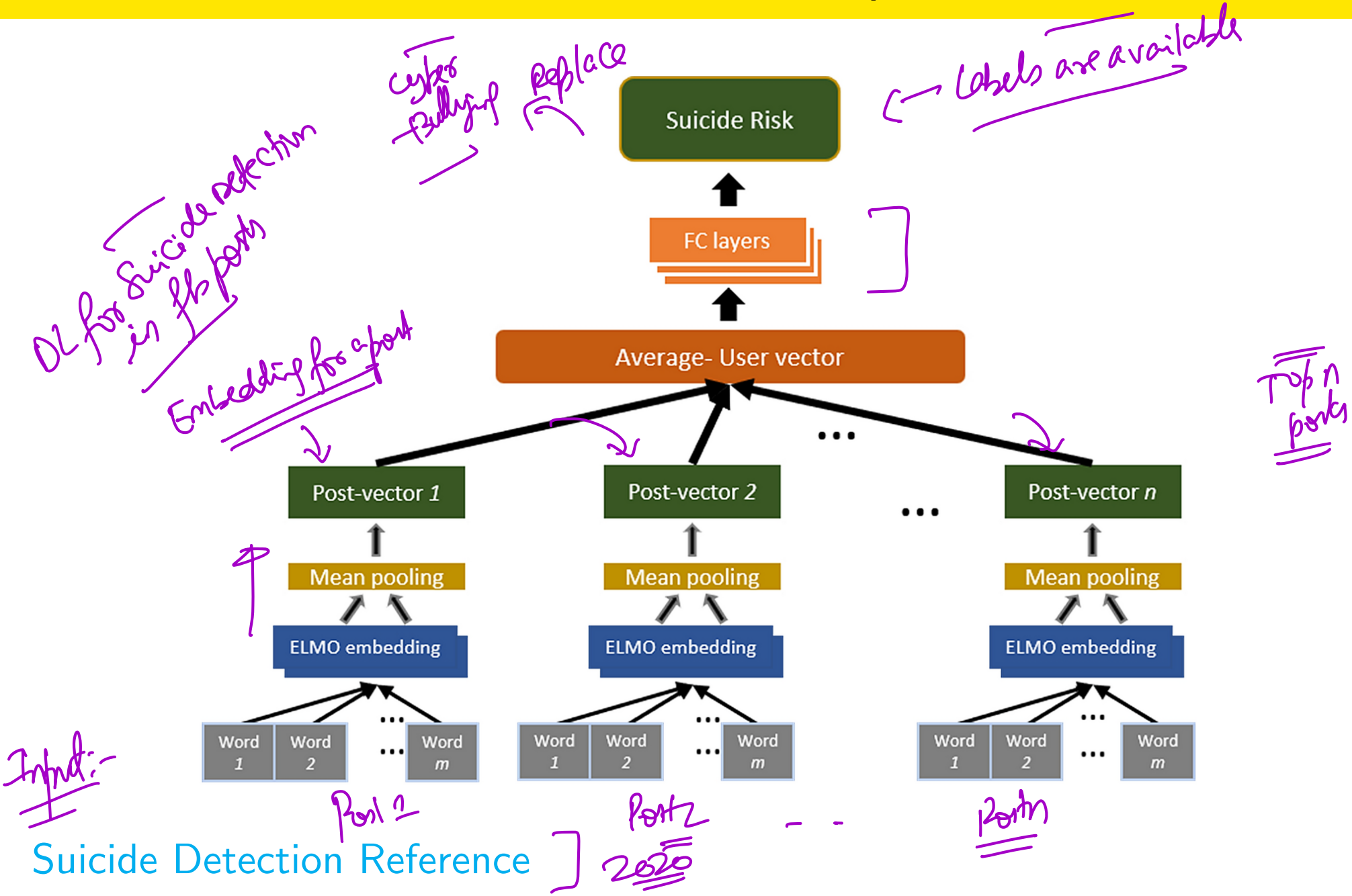
Deep Learning can provide powerful non-linear supervised models for anomaly detection provided there is enough data (both positive and negative examples) and we account for over-fitting. On the upcoming assignment, can also try out deep learning!

Detecting bots/in-appropriate posts

Biasing the metrics

Do you bias more towards high precision or high recall? Is there a middle ground? Can we have higher recall (i.e. detect the bots/in-appopriate posts) without pissing people off with incorrect flags?

Suicide detection from Social Media posts



Anomaly Detection Methods so far

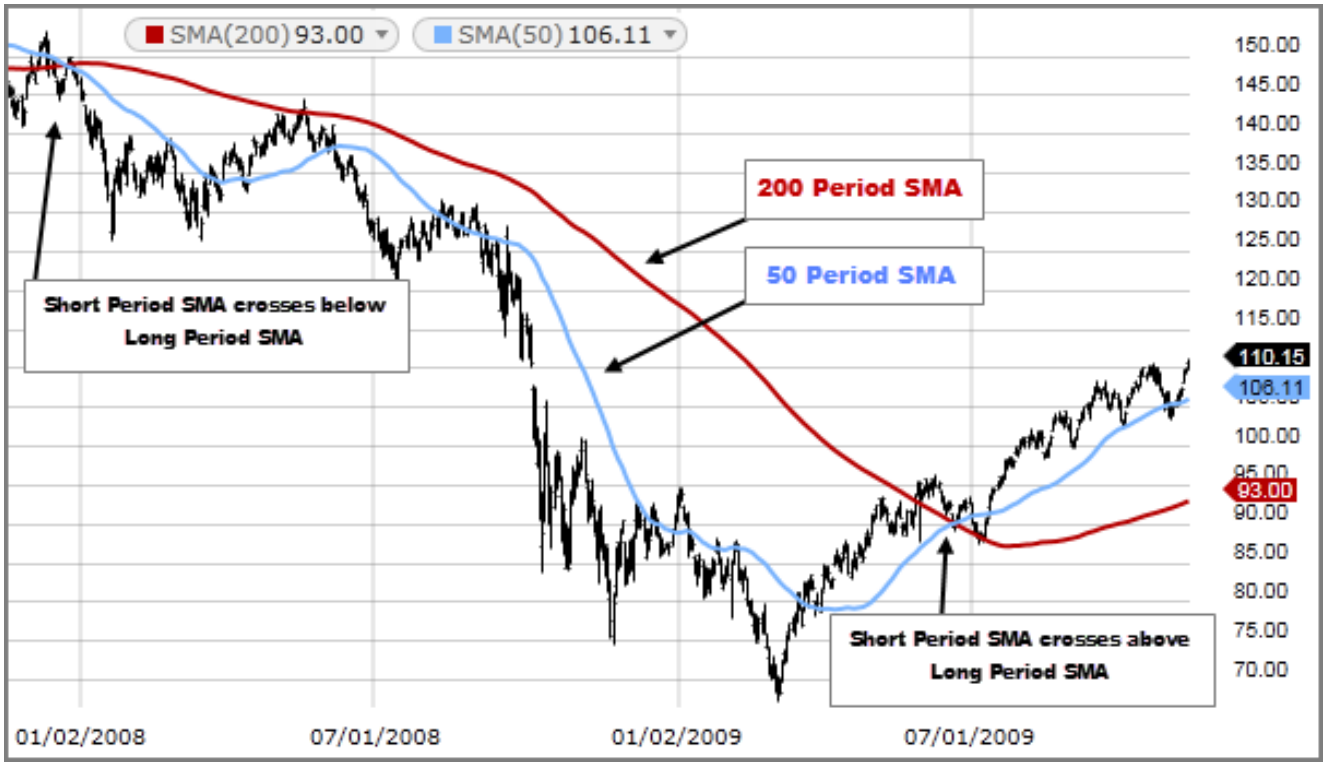
	Method	Pros	Cons
1	Mean/Std Deviation	Identifies some anomalies	False positives
2	Supervised Learning	Precise detection	As good as features

3. GMM — unsupervised model

Anomaly Detection Methods - Coming up

	Method	Pros
1	Mean/Std Deviation	Identifies some anomalies
2	Supervised Learning	Precise detection
3	**Simple Moving Average (SMA)**	Improves on mean/std deviat
4	**Exponential Moving Average (EMA)**	More sensitive then SMA
5	**STL**	Accounts for seasonality

Moving Averages - Simple Moving Average



Simple Moving Average and Anomalies

SMA

1. There is a window size that helps you track the **moving** average.
2. 50-SMA is a 50 day moving average
3. $\text{50-SMA}(i) = \frac{1}{50} \sum_{j=i-50}^{i-1} x_j$

Simple Moving Average and Anomalies

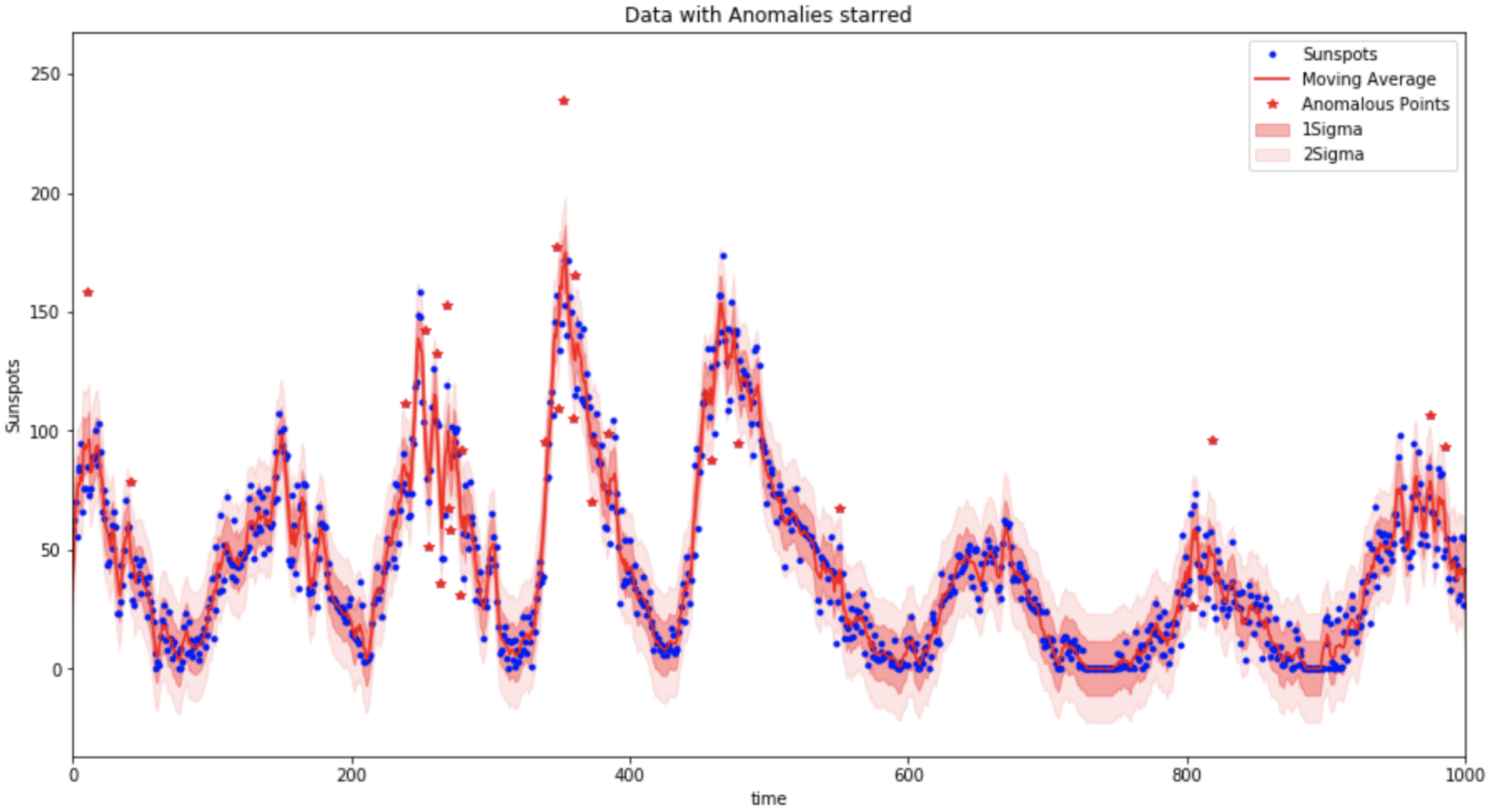
SMA

1. There is a window size that helps you track the **moving** average.
2. 50-SMA is a 50 day moving average
3. $\text{50-SMA}(i) = \frac{1}{50}\sum_{j=i-50}^{i-1} x_j$

Anomaly detection

1. x_i is an anomaly if $\|SMA(i) - x_i\|$ deviates above a $t \times SD(i)$ where $SD(i)$ is the standard deviation and N is the size of the window, t is the threshold.

SMA example



Github Library to try!

ICE #5

Moving mean computation

Let's say you wanted to implement SMA yourself. Let the window size be 100. You have $SMA(i-1)$. How do you compute it from $SMA(i-1)$?

a. $SMA(i) = SMA(i-1) + (x_i - x_{i-1})/N$
b. $SMA(i) = SMA(i-1) + x_i/N$
c. $SMA(i) = SMA(i-1)$
d. $SMA(i) = SMA(i-1) - x_{i-1}/N$

ICE #6

Moving mean computation

Based on the previous question, what's the computational complexity and memory/storage complexity of SMA at point i, i.e. $SMA(i)$?

a. $O(N), O(N)$
b. $O(1), O(N)$
c. $O(N), O(1)$
d. $O(1), O(1)$

Moving Variance computation for SMA

Moving Variance

Same principle as computing the moving mean for SMA.

Exponential Moving Average and Anomalies

EMA

1. Similar to SMA - Except the moving window is **soft**
2. Weight more of the recent terms than before and weight it exponentially.
3. $EMA(i) = (1 - \beta) * EMA(i - 1) + \beta * x_i$ where $0 \leq \beta \leq 1$
4. $EMA(i) = \beta x_i + \beta(1 - \beta)x_{i-1} + \beta(1 - \beta)^2 x_{i-2} + \ldots$
5. EMA has a hyper-parameter β instead of window size N as in SMA.

Exponential Moving Average and Anomalies

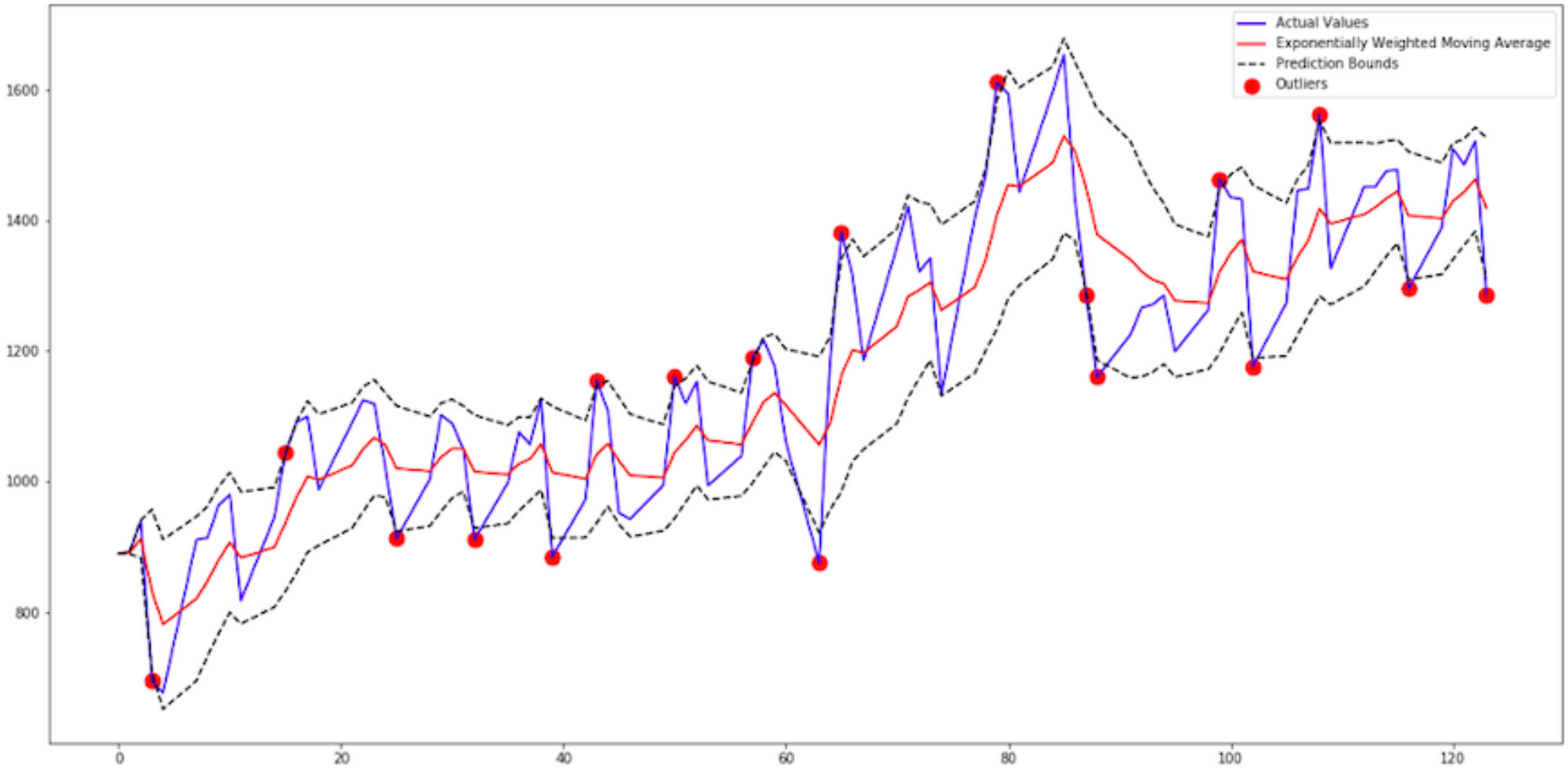
EMA

1. Similar to SMA - Except the moving window is **soft**
2. Weight more of the recent terms than before and weight it exponentially.
3. $EMA(i) = (1 - \beta) * EMA(i - 1) + \beta * x_i$ where $0 \leq \beta \leq 1$
4. $EMA(i) = \beta x_i + \beta(1 - \beta)x_{i-1} + \beta(1 - \beta)^2 x_{i-2} + \ldots$
5. EMA has a hyper-parameter β instead of window size N as in SMA.

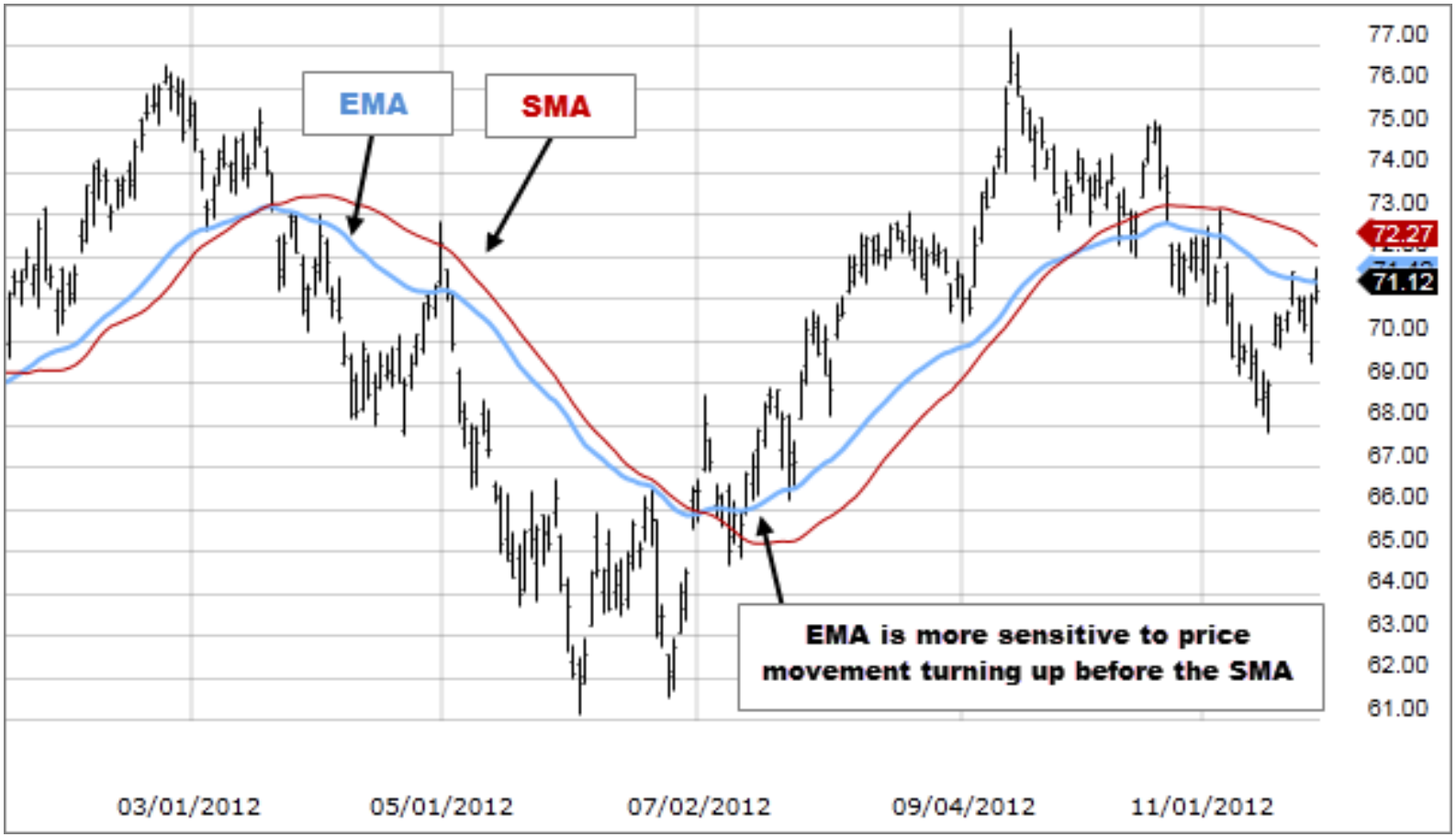
Anomaly detection

1. x_i is an anomaly if $\|EMA(i) - x_i\|$ deviates above a $t \times SD(i)$ where $SD(i)$ is the standard deviation of the deviation.

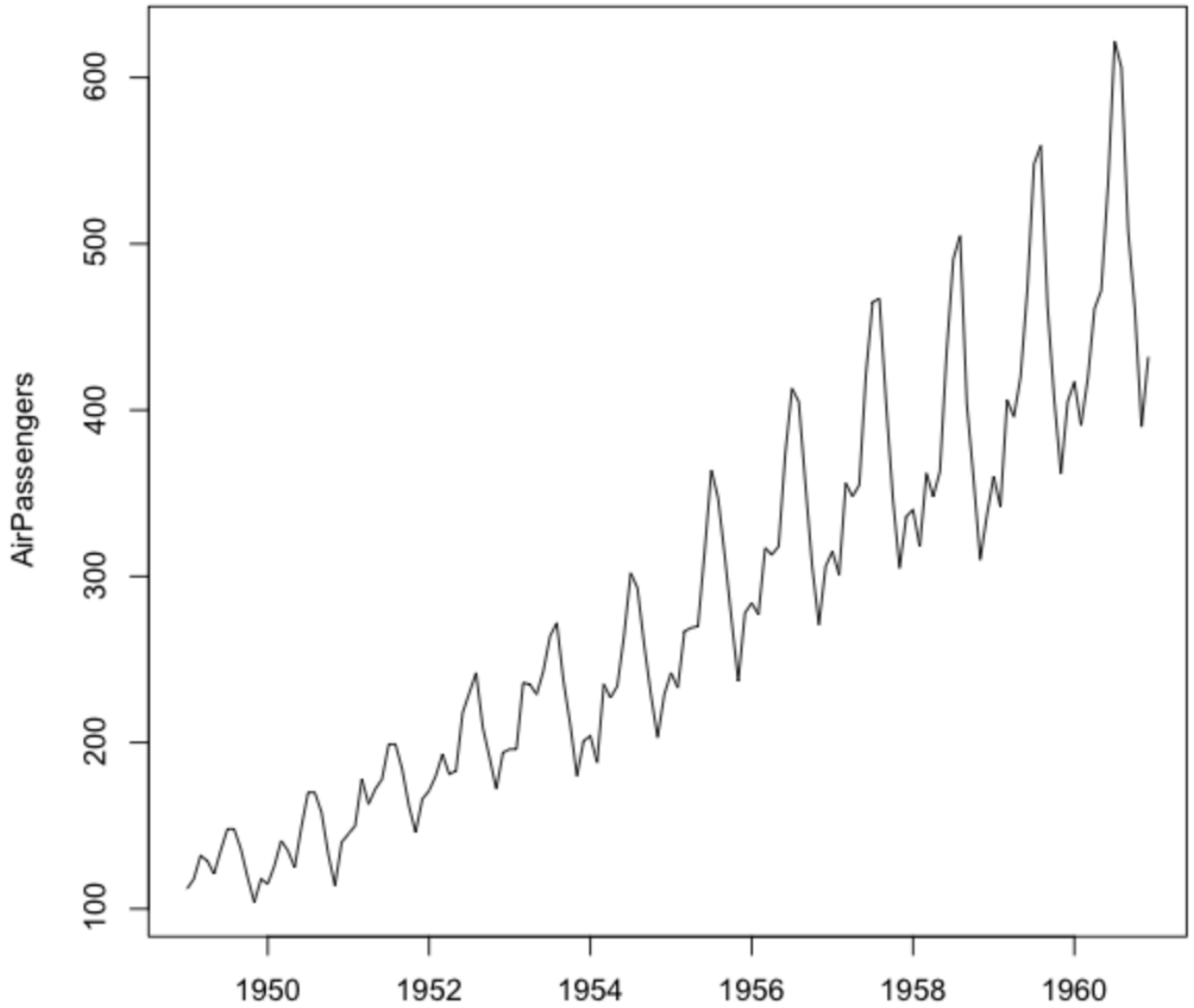
Exponential Moving Average



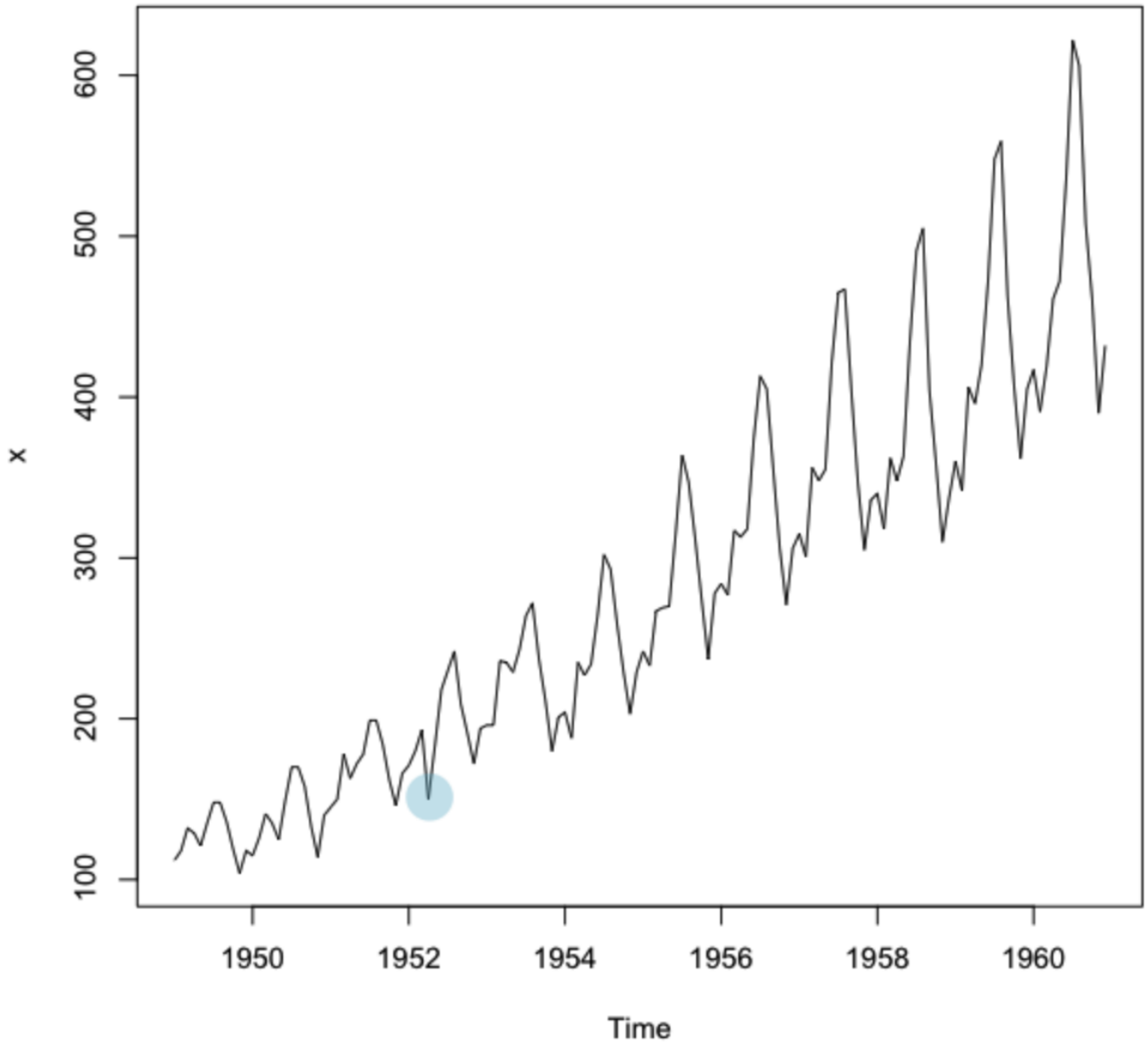
SMA vs EMA


EMA
SMA
EMA is more sensitive to price movement turning up before the SMA
77.00
76.00
75.00
74.00
73.00
72.27
71.12
70.00
69.00
68.00
67.00
66.00
65.00
64.00
63.00
62.00
61.00
03/01/2012
05/01/2012
07/02/2012
09/04/2012
11/01/2012

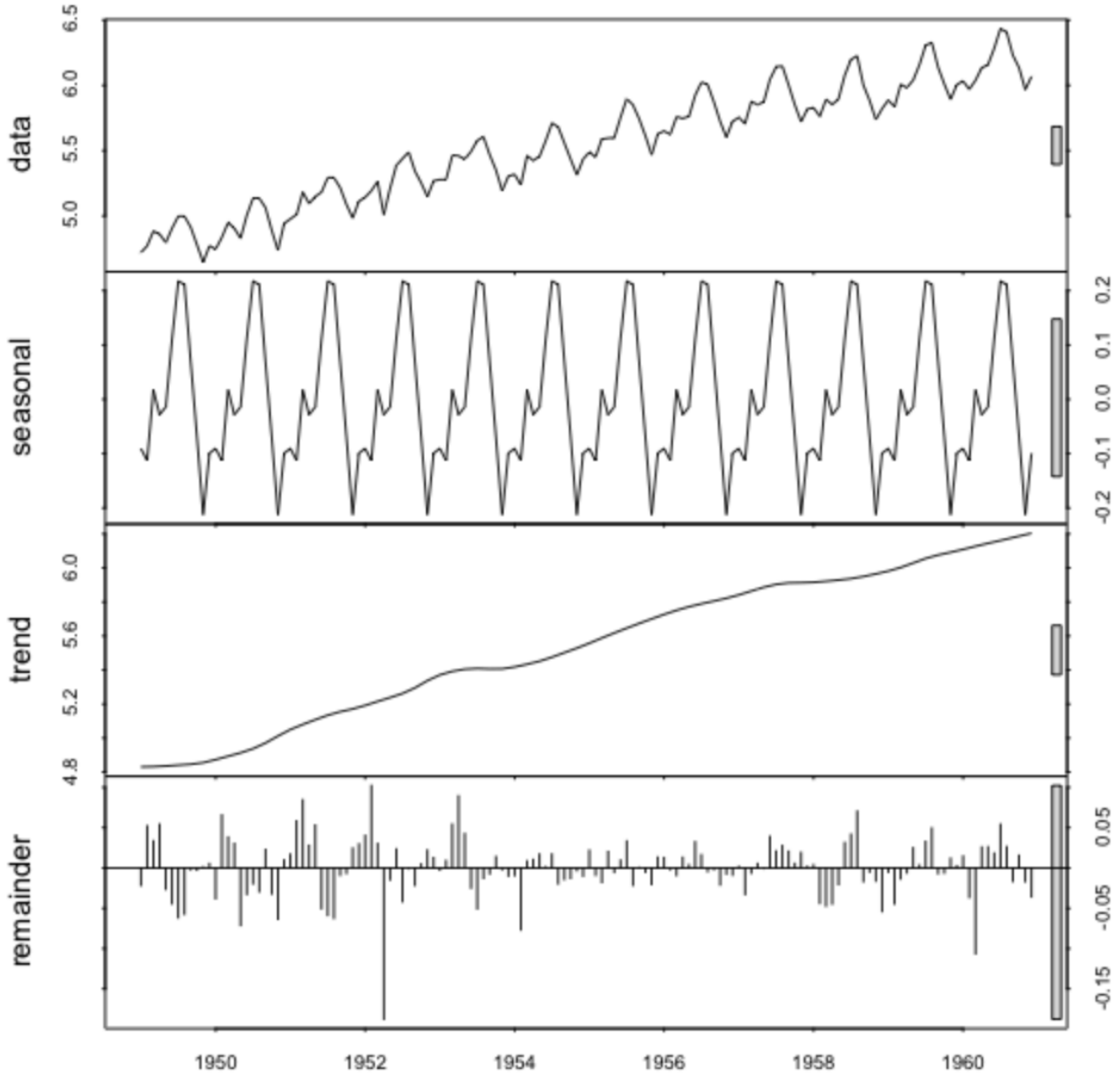
Accounting for Seasonality



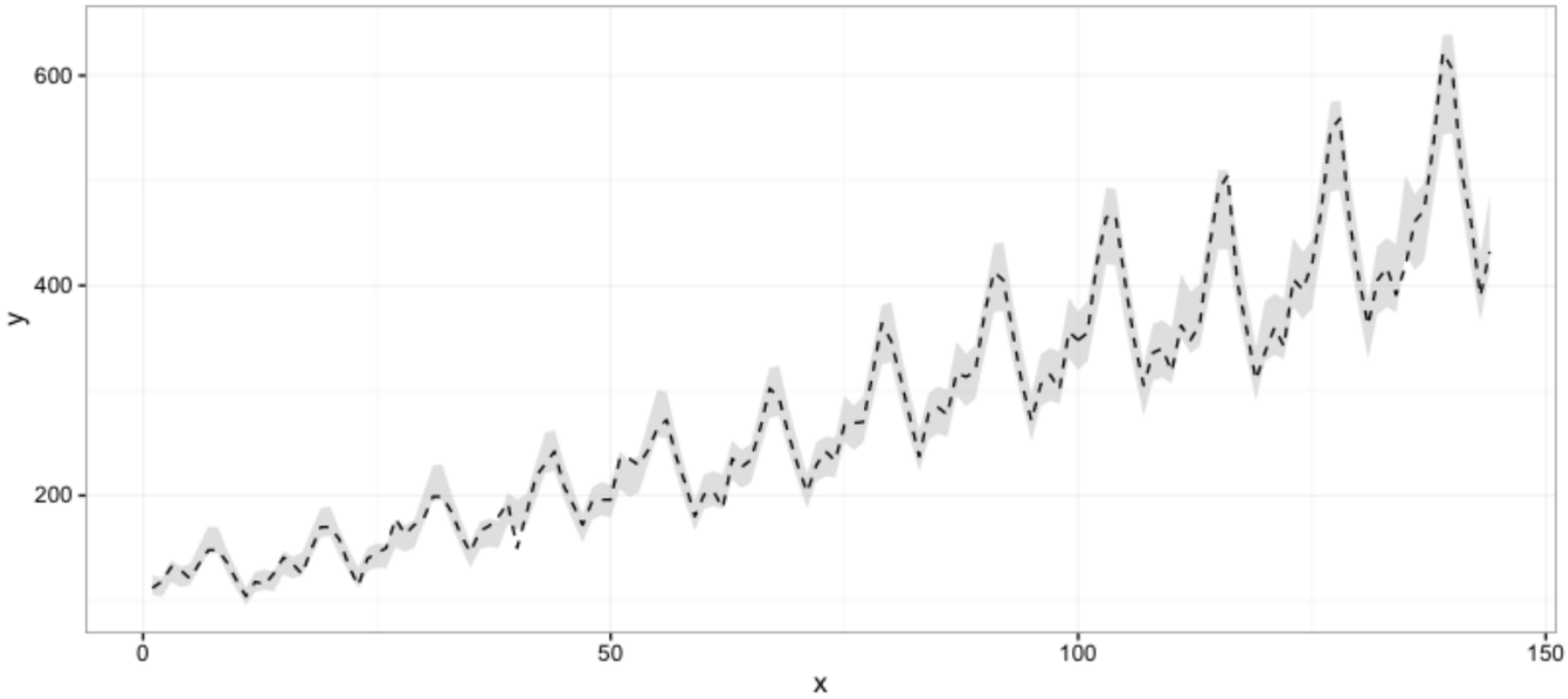
Accounting for Seasonality



STL: Accounting for Seasonality



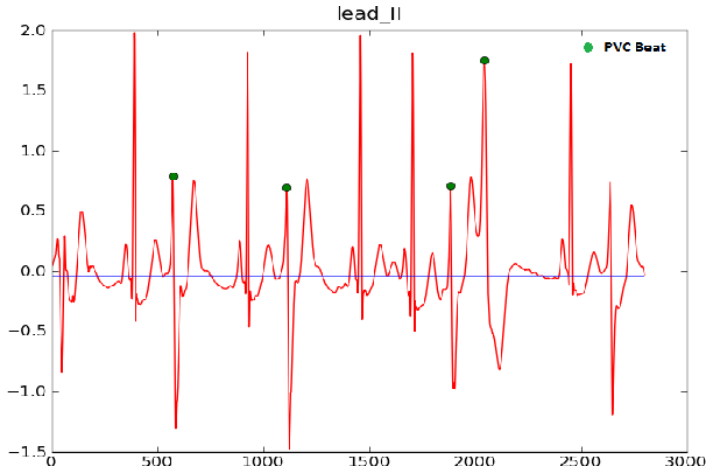
STL: Accounting for Seasonality



STL Library

Prophet Anomaly Detection

For the upcoming assignment on Arrhythmia Detection



1. Try SMA/EMA (unsupervised baseline)
2. Try a supervised method baseline like Logistic Regression
3. Try a deep learning model

Questions?

Extra Slides

Anomaly Detection - Baselines

ICE #1

Suppose you have a data for product sales of all of groceries for a particular retail company by the hour for each day. Let the maximum product sales for any given hour is $30k$ and minimum is $5k$. Suppose you notice today that for the hour starting at 6 pm, the sales was $29k$ and for the hour starting at 10 am, the sales was $6k$. Which would be more suspect to be an anomaly sales data point?

a. Sales at 10 am
b. Sales at 6 pm
c. Could be both
d. Neither

Stock Price Prediction

Can this be modeled as an anomaly detection problem?

Can you build a ML model that can predict when to buy a stock and when to sell a stock to maximize your profits. How exactly would you do it? And how could you cast it as an anomaly detection model?

Stock Price Prediction

Local Window

What would be the local window size you would choose for stock price prediction?