EEP 596: Adv Intro ML || Lecture 11

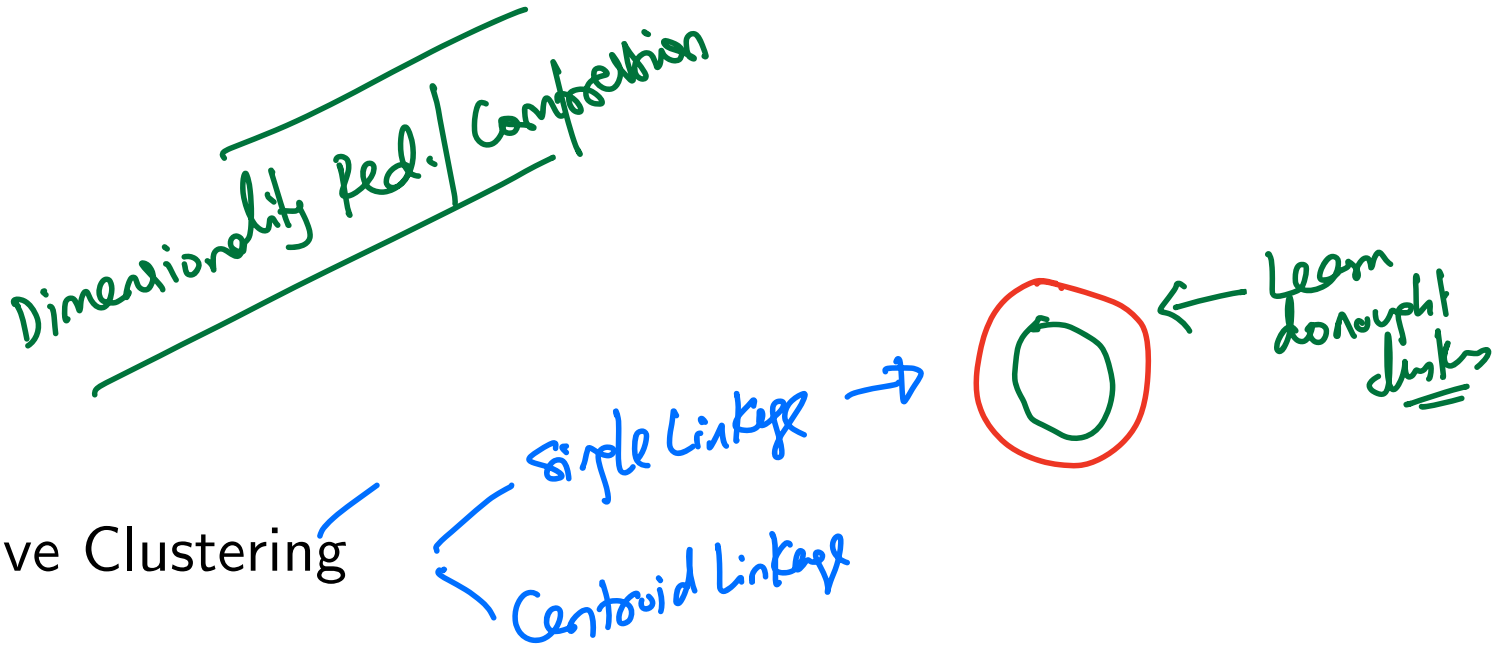
Dr. Karthik Mohan

Univ. of Washington, Seattle

February 10, 2023

Last Time

- SVD
- PCA
- Agglomerative Clustering



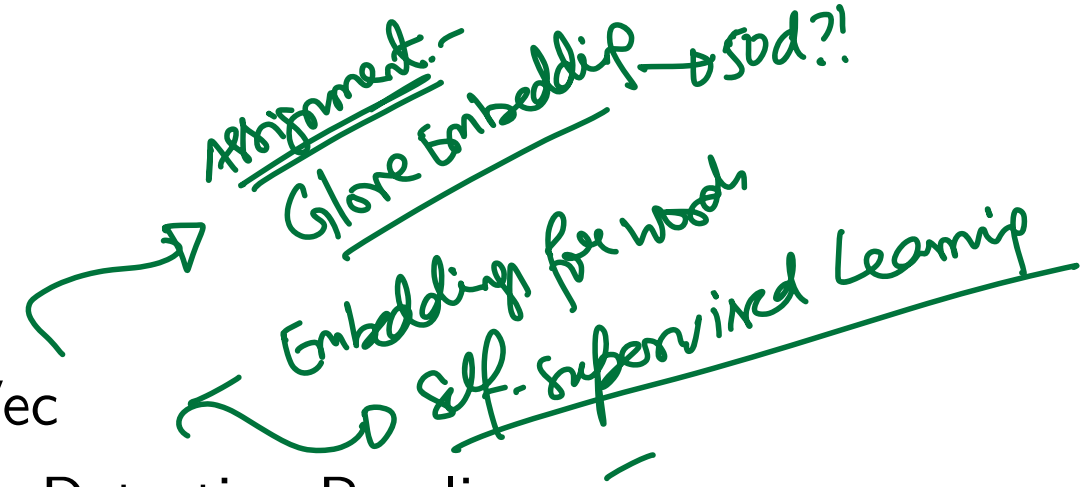
Today

- Word2Vec

Today

a. Word2Vec
b. Anomaly Detection Baselines

Today



a. Word2Vec
b. Anomaly Detection Baselines
c. GMMs

Today

a. Word2Vec
b. Anomaly Detection Baselines
c. GMMs
d. Anomaly Detection for Time-series

Dunn Index - Metric that measures goodness of clusters

Dunn Index

$$D = \frac{\min_{1 \leq i < j \leq K} d(i,j)}{\max_{1 \leq j \leq K} d'(j)}$$

Dunn Index ICE

ICE #1

Say you had a single-linkage and k-means clustering applied to a data set to produce K clusters each. Call them A and B. When would you say single-linkage produces better clustering than k-means?

a. $D(A) > D(B)$
b. $D(B) > D(A)$

PCA for Images

PCA for Images - Eigen Faces

PCA for Images - Re-construction

ICE #2

Images PCA (Work in groups of 2)

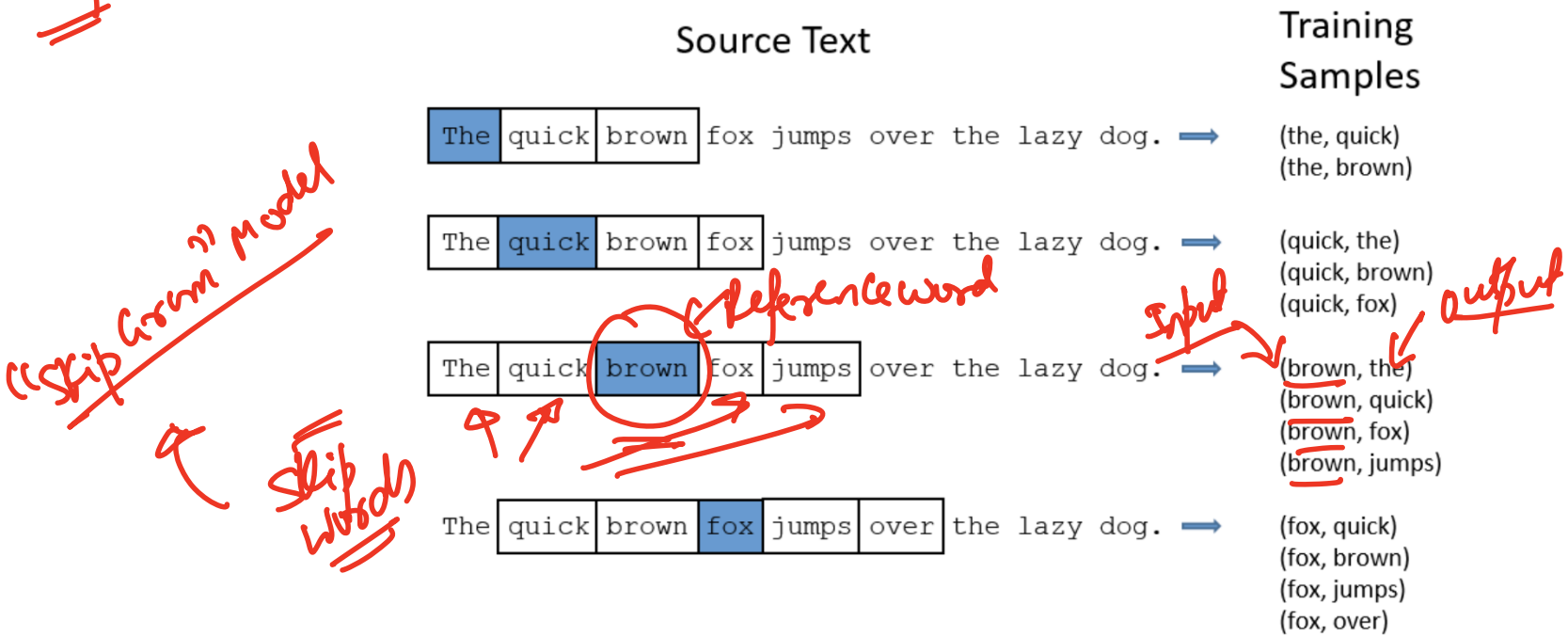
If you have 1000 face images and did a PCA on these images and found that 10 Eigen faces would be sufficient to reconstruct the images accurately. You stored compressed representations of the images on your laptop and to reconstruct the image, you send it to a server that then gives you back a re-constructed image. What would be the compression factor for the compressed representation you have on your laptop and obtained from PCA? Assume that each image is 1000×1000 pixels?

a. 1000x
b. 10000x
c. 100000x
d. 1MMx

Word2Vec

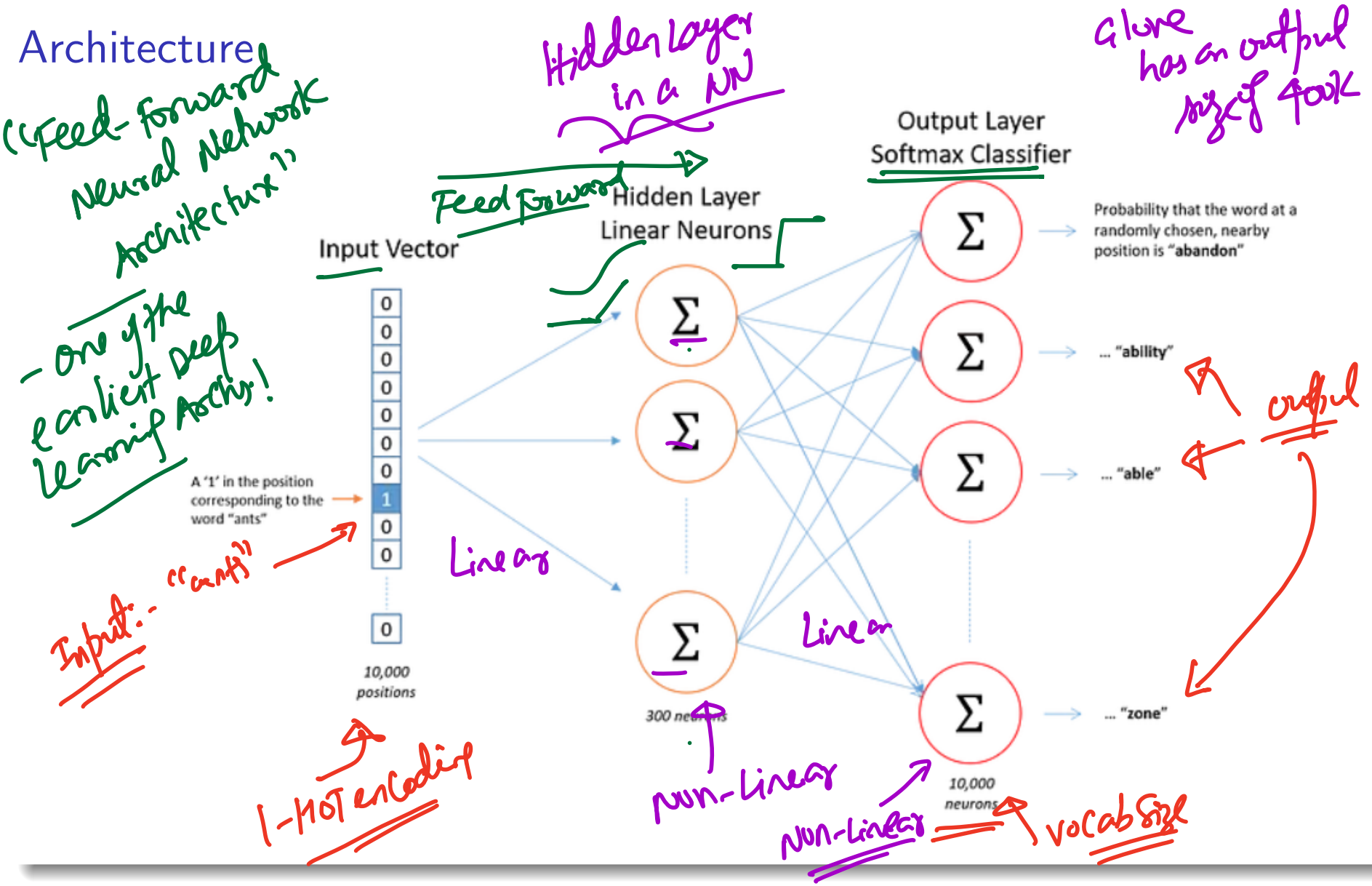
Skip Gram Model

Is based on the skip-gram model! How is training done? It's ~~semi~~-supervised!!



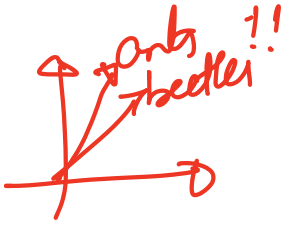
Word2Vec

Architecture



"ants" 1

"beetle" 2



Word2vec

1. Based on self-supervised learning
2. Uses a FF NN architecture
3. Loss function:- Categorical cross-Entropy
4. Popular implementations:- Glove
5. Embedding size!:- 50 to 1000...

Softmax

	Prob.
5	0.3
6	0.001
7	0.4
0.05	0.05
2	0.1
3	0.002

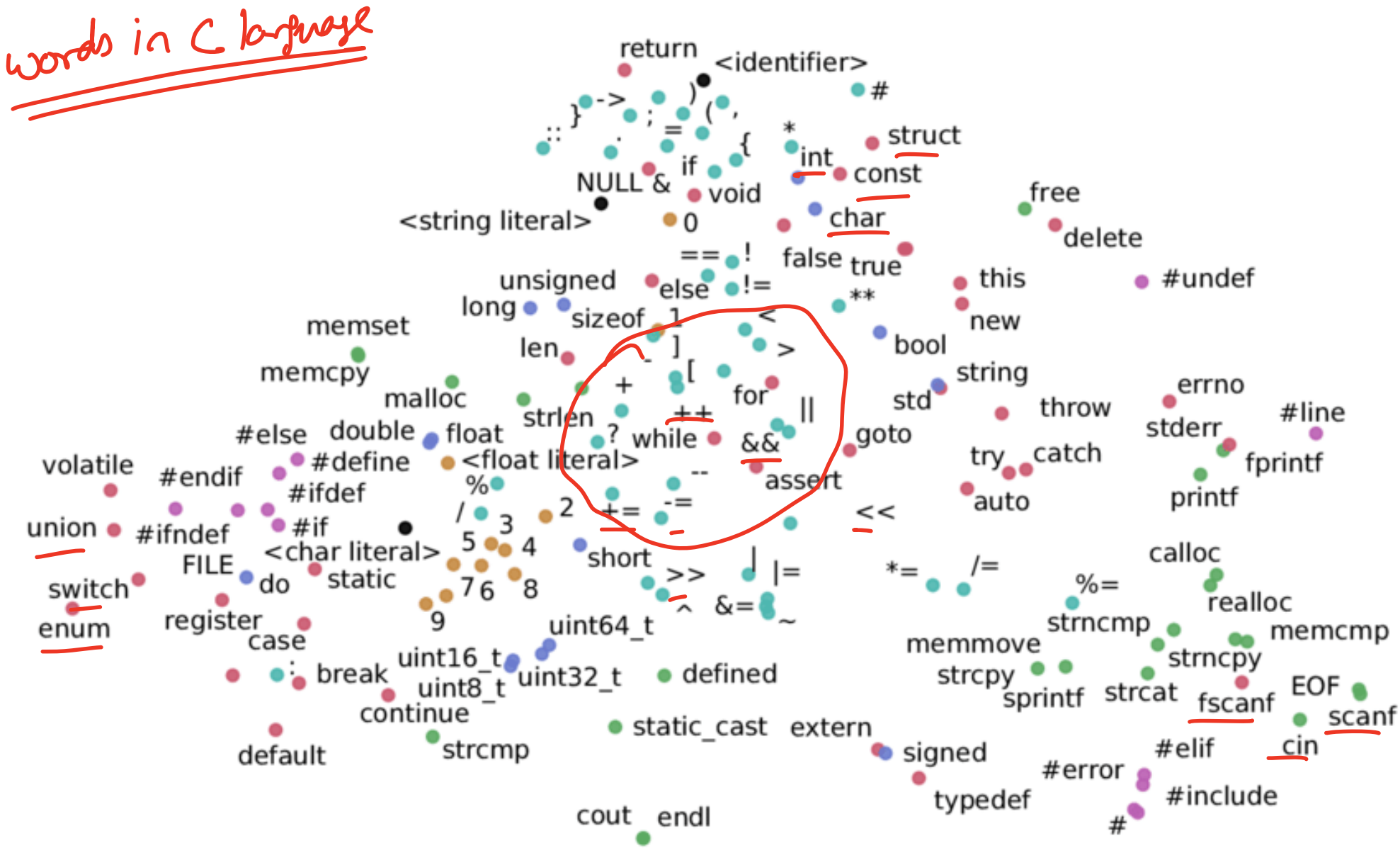
Prob.

"DUCK" → Top 5 similar words to "Duck"

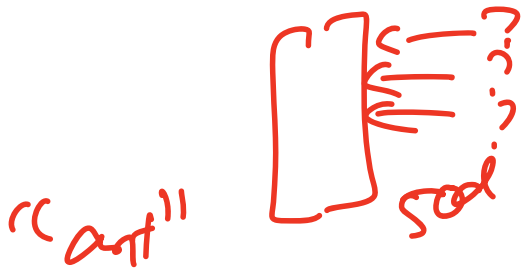
↳ Exact NN search

↳ Approx. NN search

Word2Vec representation



ICE #3



What do the embedding dimensions of word2vec represent?

1. Fixed words decided by word2vec
2. Topics that are common among the words
3. Parts of speech of the words (nouns, adjectives, etc)
4. Book titles that these words came from

Product2Vec

X2vec

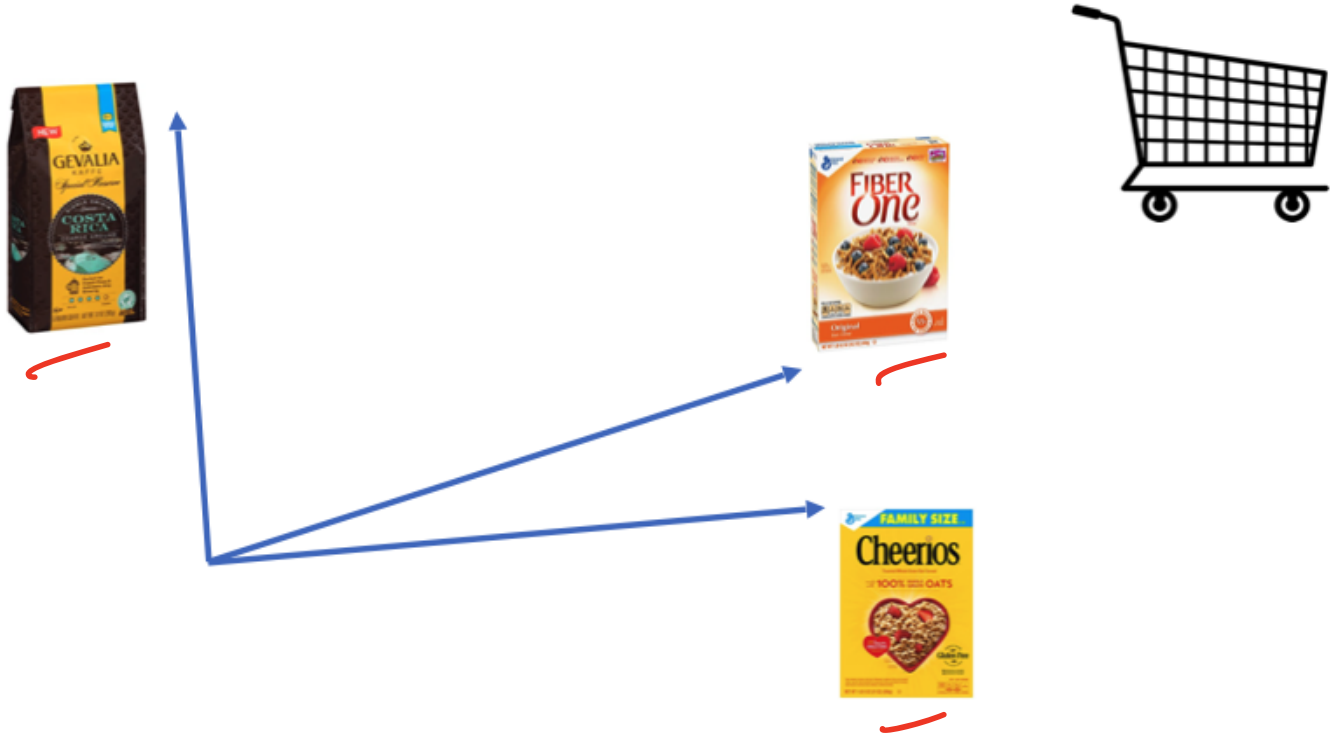
1. word2vec
2. sentence2vec
3. Doc2vec
4. ...

Represent products in product space with a large matrix of embedding coordinate vectors "*L*"

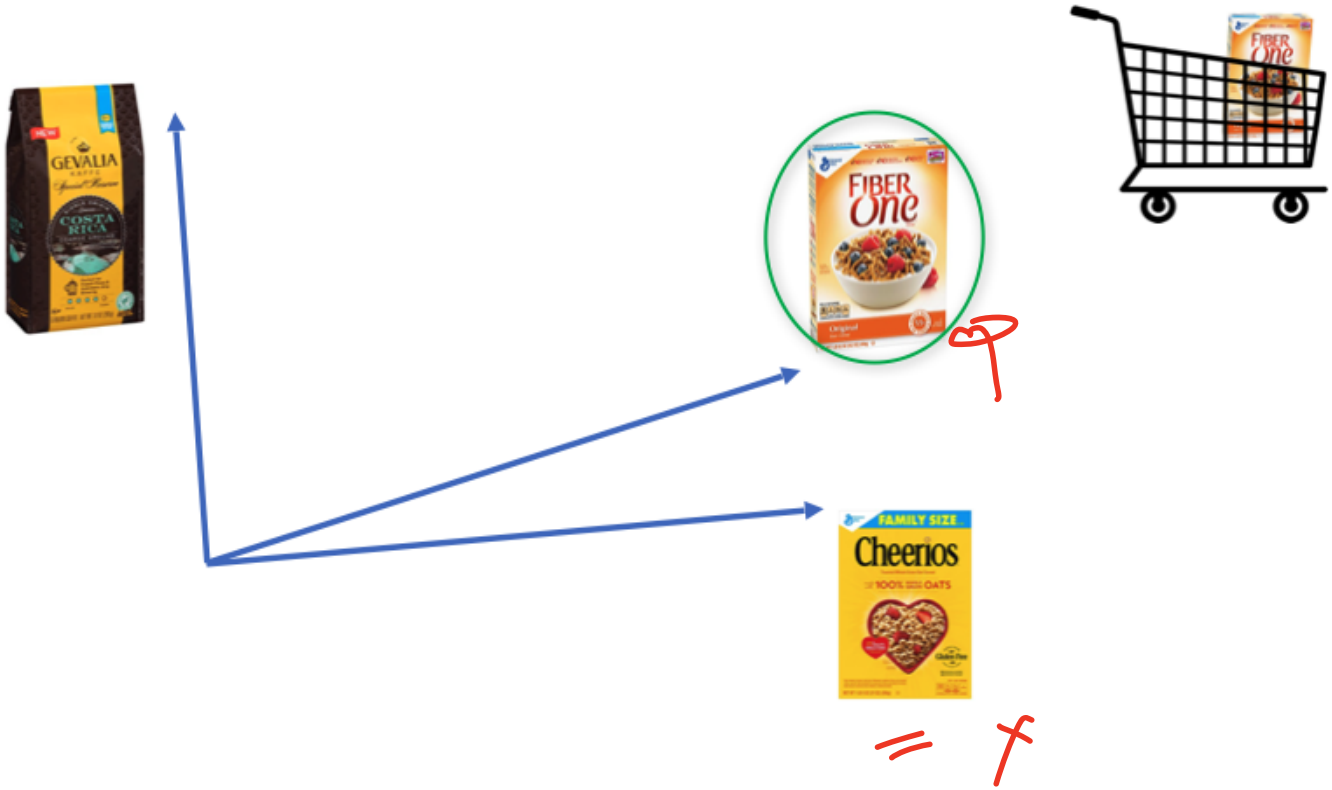
$$L = \begin{pmatrix} 1.5 & 1.9 & 1.8 & 1.4 & \cdots & 0.4 \\ 0.6 & 0.1 & 1.0 & 1.6 & \cdots & 1.9 \\ 0.6 & 1.6 & 1.6 & 1.6 & \cdots & 1.8 \\ 0.6 & 1.0 & 0.1 & 1.6 & \cdots & 0.6 \\ 0.8 & 1.4 & 1.9 & 0.8 & \cdots & 0.7 \end{pmatrix}$$

We obtain these embedding vectors from the Product2Vec service [London et al, 2017]

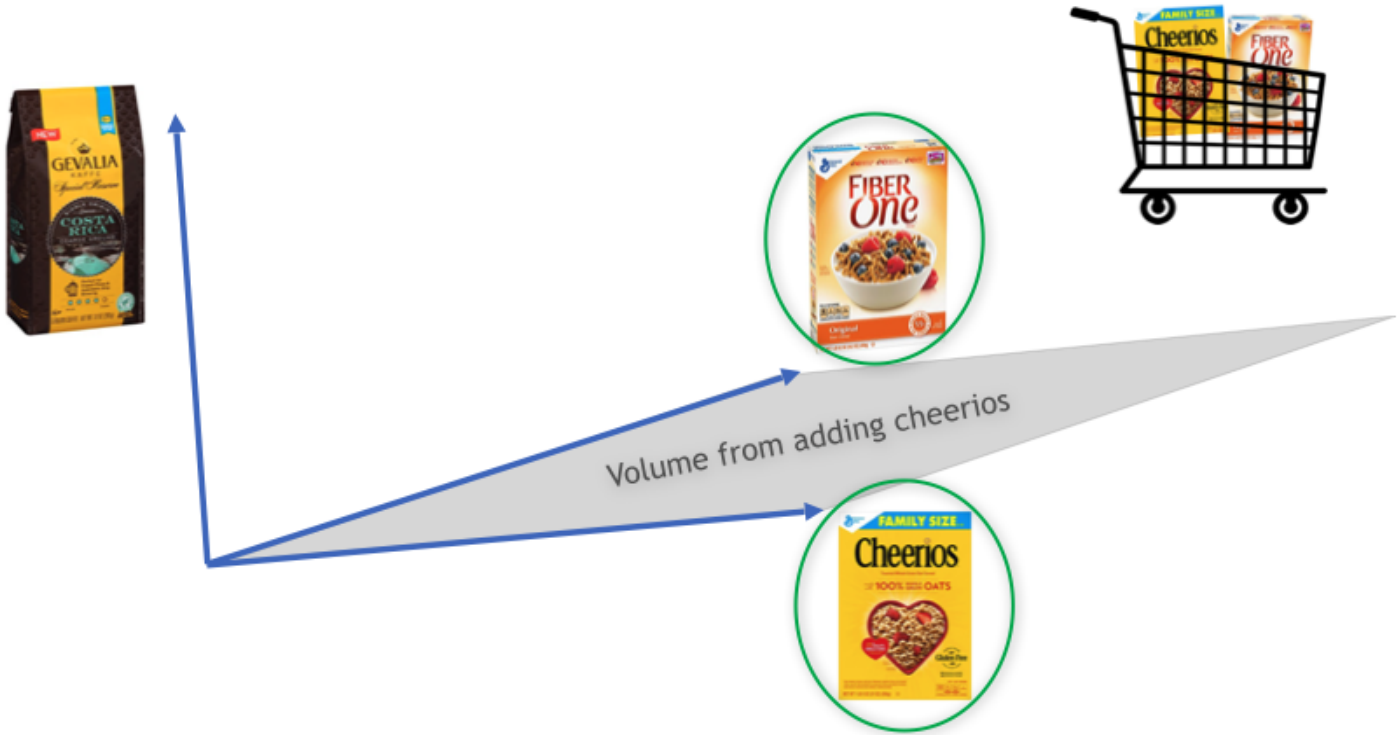
Product2Vec application



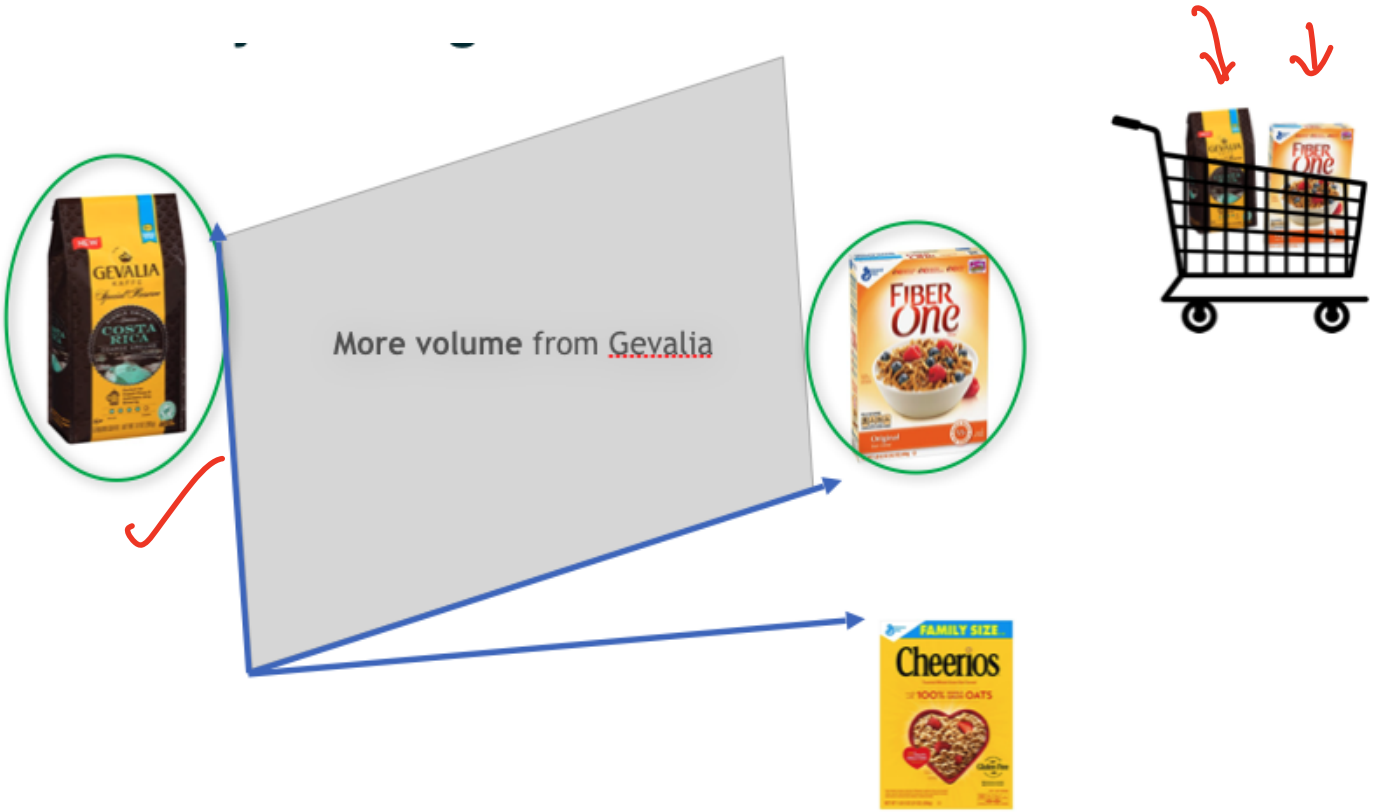
Product2Vec application



Product2Vec application



Product2Vec application



Breakout: Discuss your favorite X2Vec!

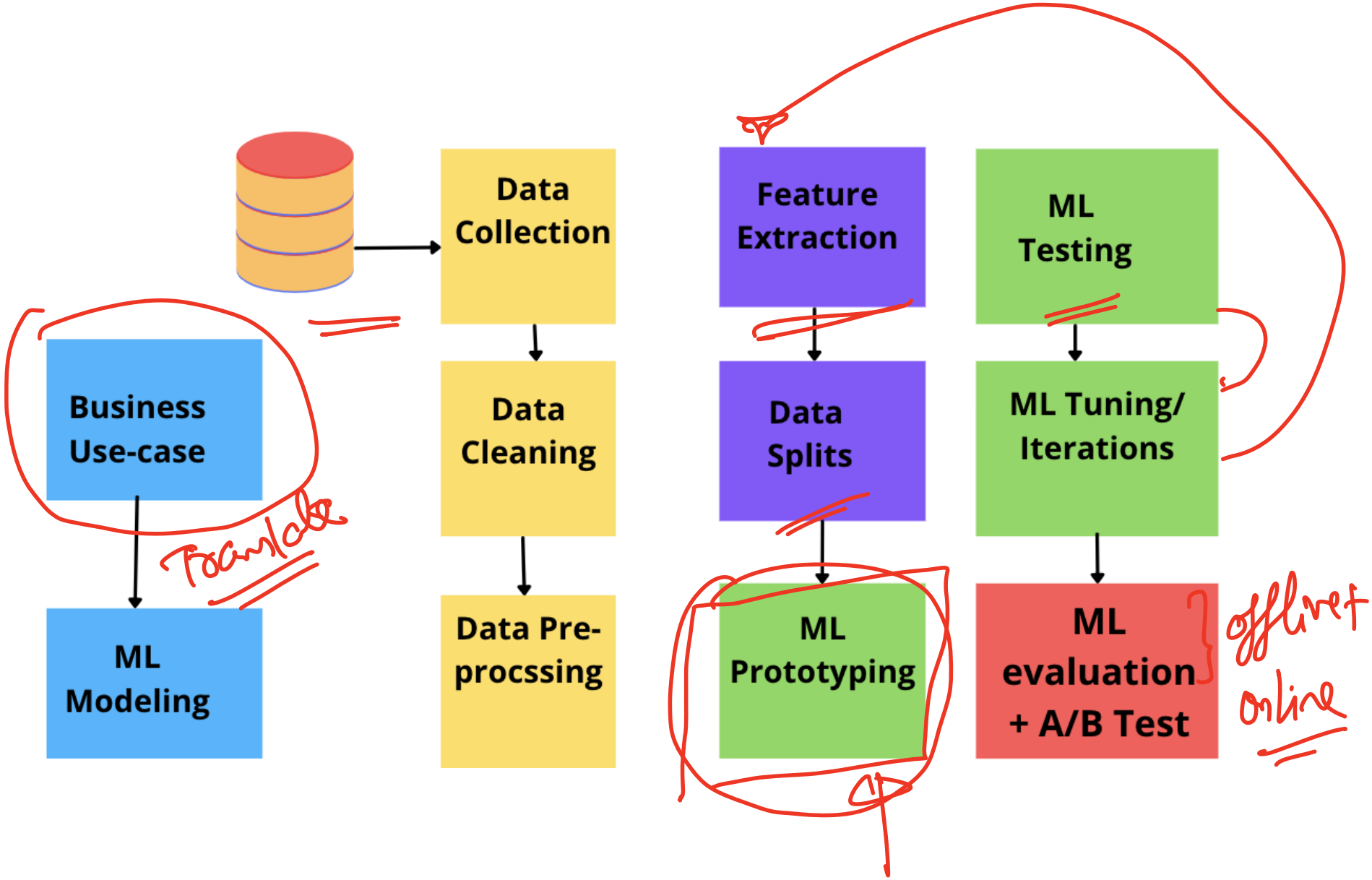
X2Vec Breakout

In your group - Discuss an application that requires machine learning. Be specific about it - Example, data, features, the type of problem (classificaiton, clustering, etc). Can you see how X2Vec would benefit your application. What would be your X in this case? How would you learn X2vec for your application? And how would you use it?

Comparison of Dimensionality Reduction Methods

Method	Utility	Pros	Cons
SVD	Low-dim embeddings	Easily	Scalability
		available	Accuracy
PCA	Same as SVD	EigenFaces	Outlier issues
Word2Vec	Semantic understanding	Non-linear	
Sentence2Vec	Comparing sentences		
Tweet2Vec	Understanding Tweets		
Product2Vec	Recommending products		
X2Vec			

ML Pipeline

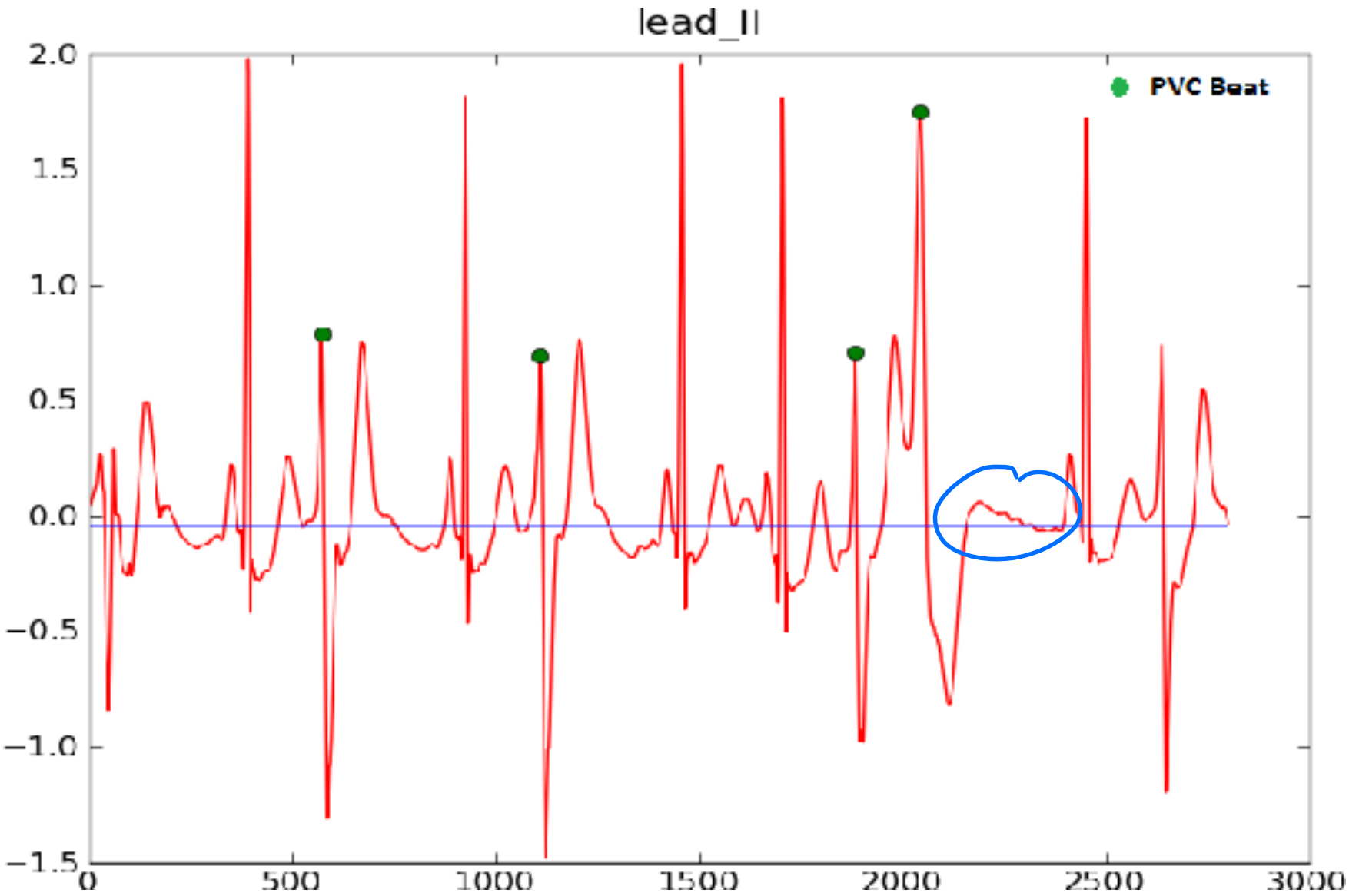

Data Collection
Feature Extraction
ML Testing
Business Use-case
Data Cleaning
Data Splits
ML Tuning/ Iterations
ML Modeling
Data Pre-procssing
ML Prototyping
ML evaluation + A/B Test
offline
online

ML Modeling - Simple Examples

	Business Use case	ML Modeling
1	Increase sales for products	Recommendation Systems?
2	Reduce server hardware failure cost	Anomaly Detection?
3	Automatically caption images	Image2Text Sequence models
4	Reduce storage cost in Data pipelines	Dimensionality Reduction
5	Automatically Diagnose heart issue	Classification?
6	...	...

Next Topic: Anomaly Detection

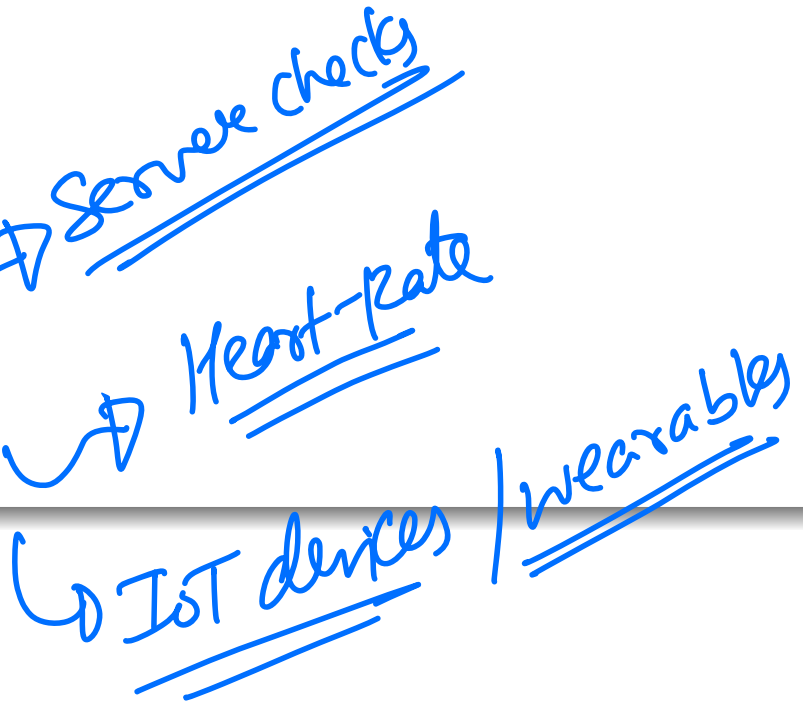
Anomaly Detection: Arrythmia



Broad list of methods

Categorization

- Offline anomaly detection
- Real-time anomaly detection



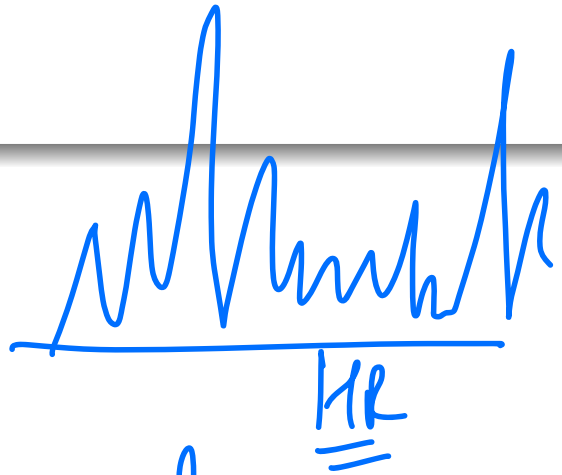
Broad list of methods

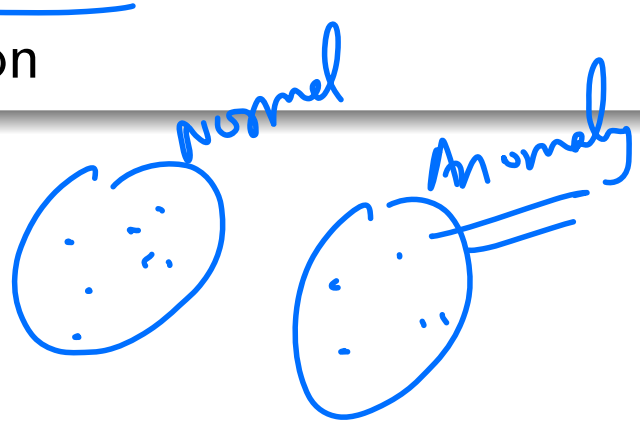
Categorization

- Offline anomaly detection
- Real-time anomaly detection

Categorization

- Time-series data anomaly detection
- Regular anomaly detection





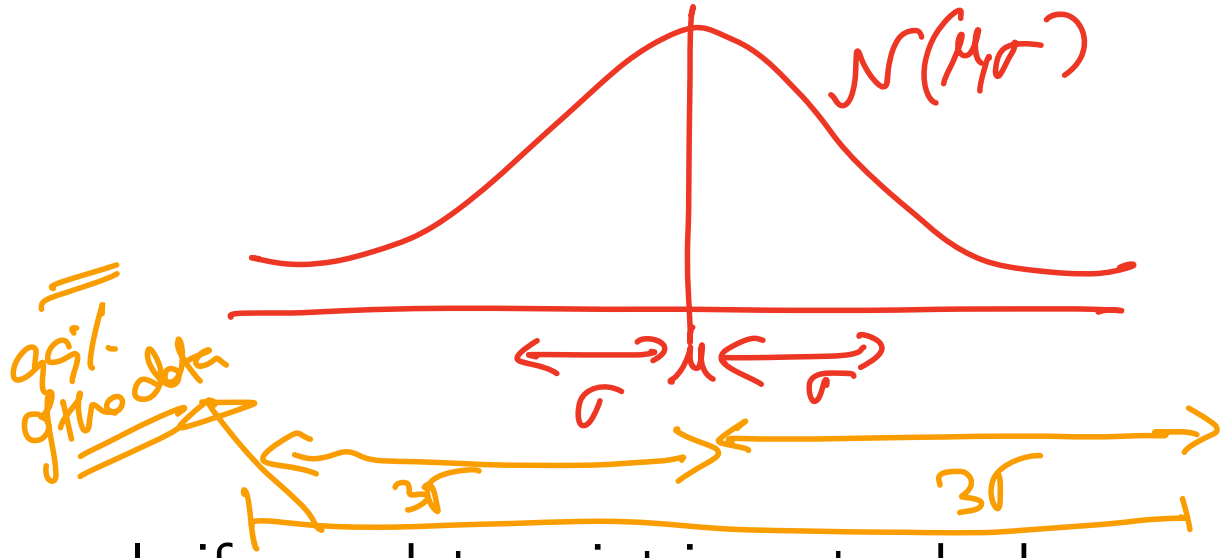
Anomaly Detection - Baselines

ICE #4 Temperature Control!

Let's say you have a thermometer that is pretty accurate but on an average is off by 0.5 degree Fahrenheit. You suspect a flu and measure your body temperature and it turns out to be 99.5. Would you say there is a cause for alarm and maybe a covid test?

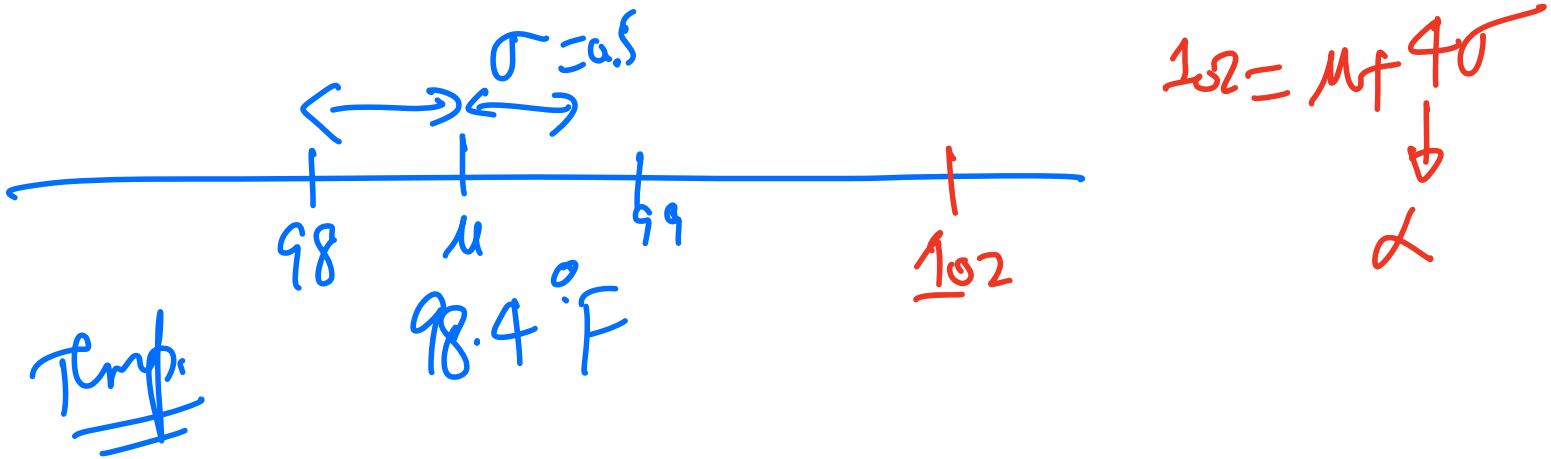
1. Yes
2. Maybe
3. No
4. Don't know

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.



Anomaly Detection Baselines

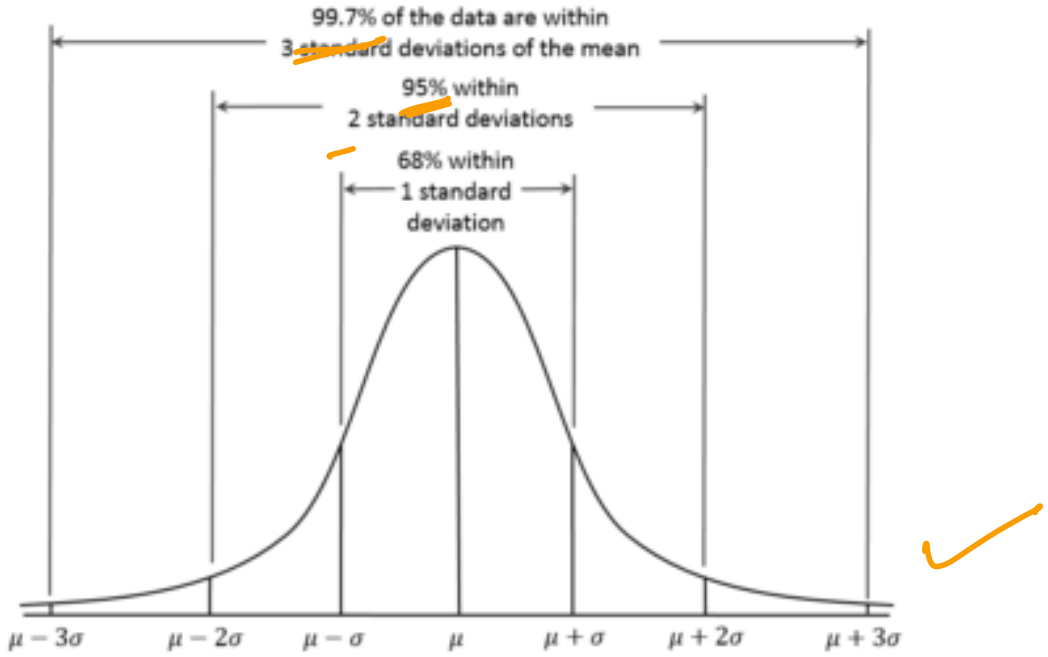
Temperature Example

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 95% of

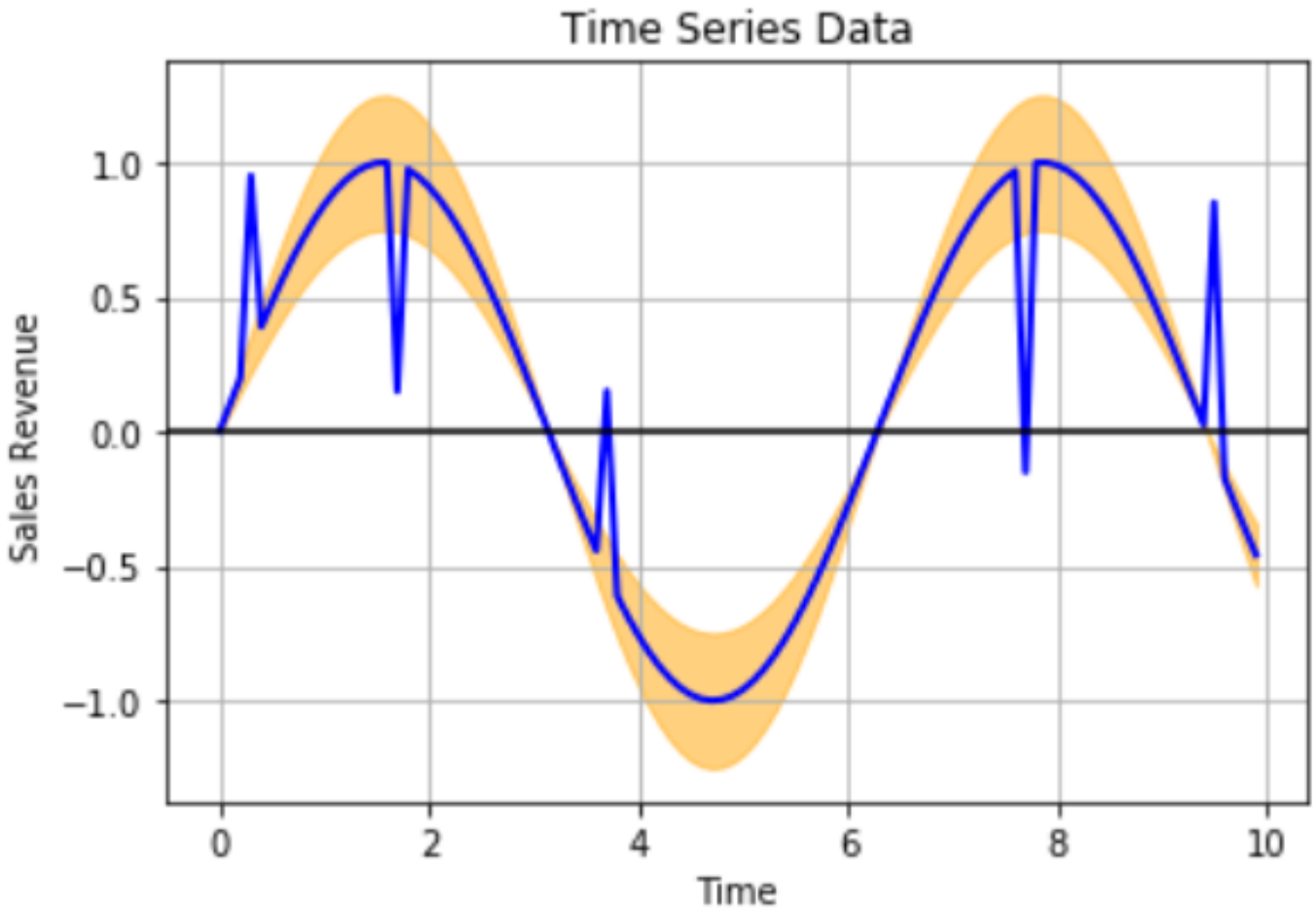
Anomaly Detection - Baselines

ICE #5

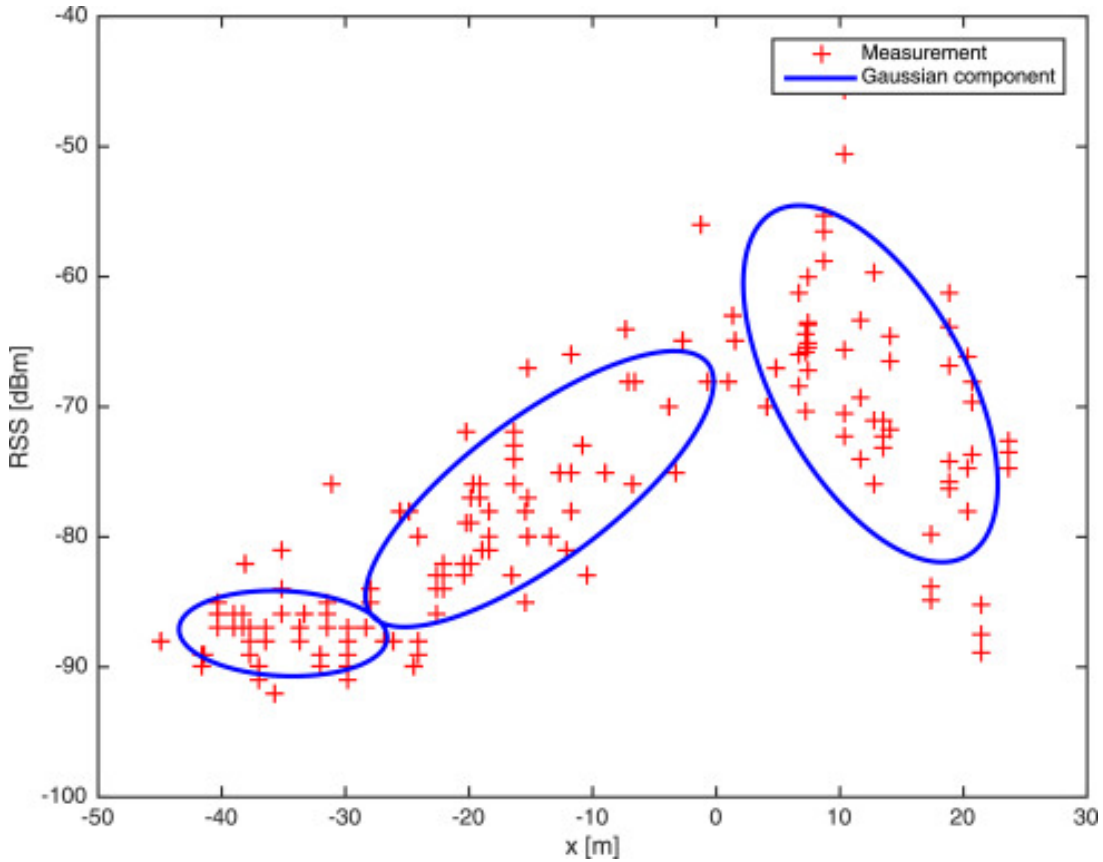
Suppose you have a data for product sales of all of groceries for a particular retail company by the hour for each day. Let the maximum product sales for any given hour is $30k$ and minimum is $5k$. Suppose you notice today that for the hour starting at 6 pm, the sales was $29k$ and for the hour starting 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

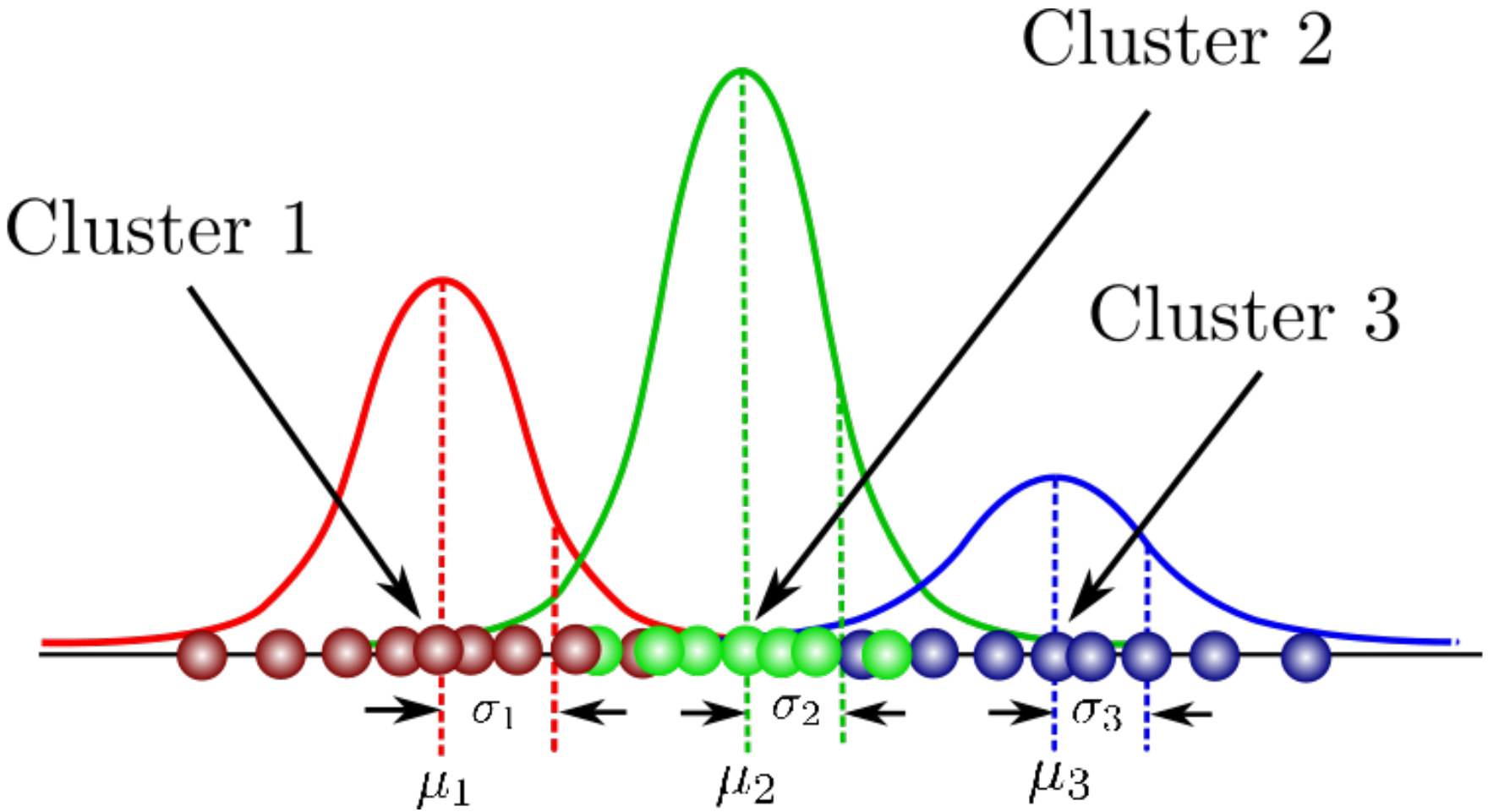
Time series anomalies



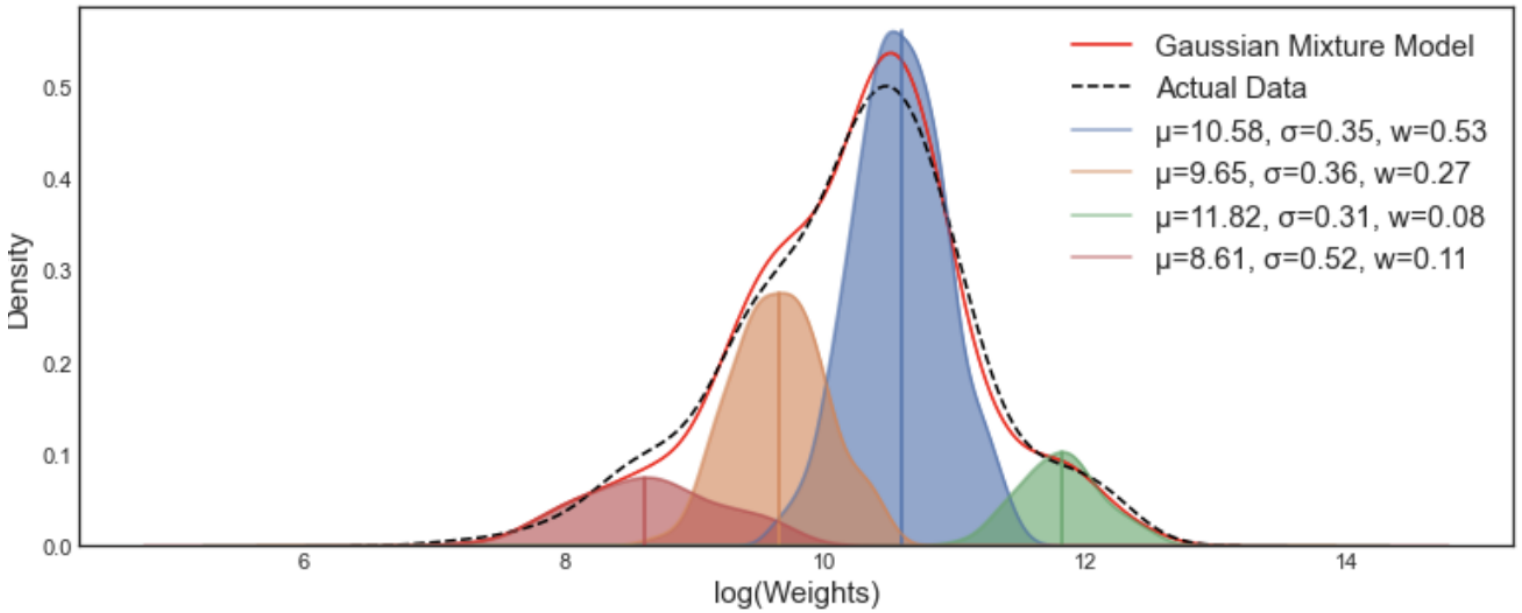
GMM - Better than Baseline



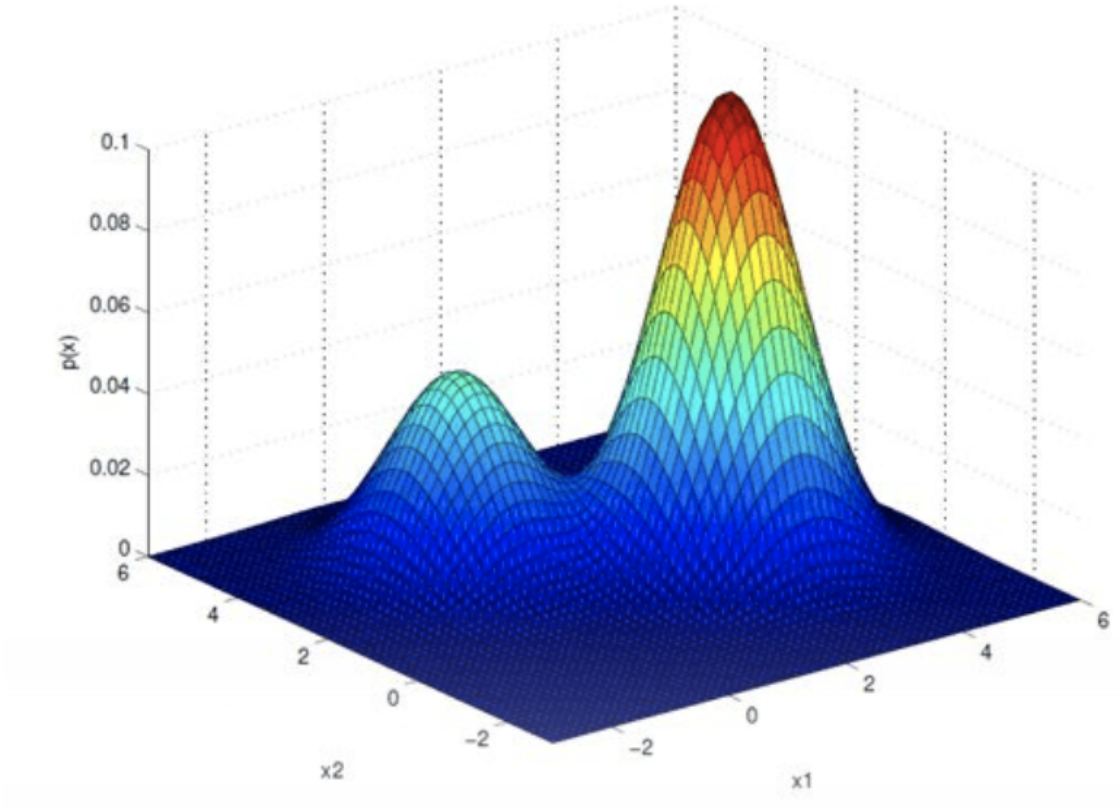
GMM - Better than Baseline



GMM - Better than Baseline



GMM - Better than Baseline



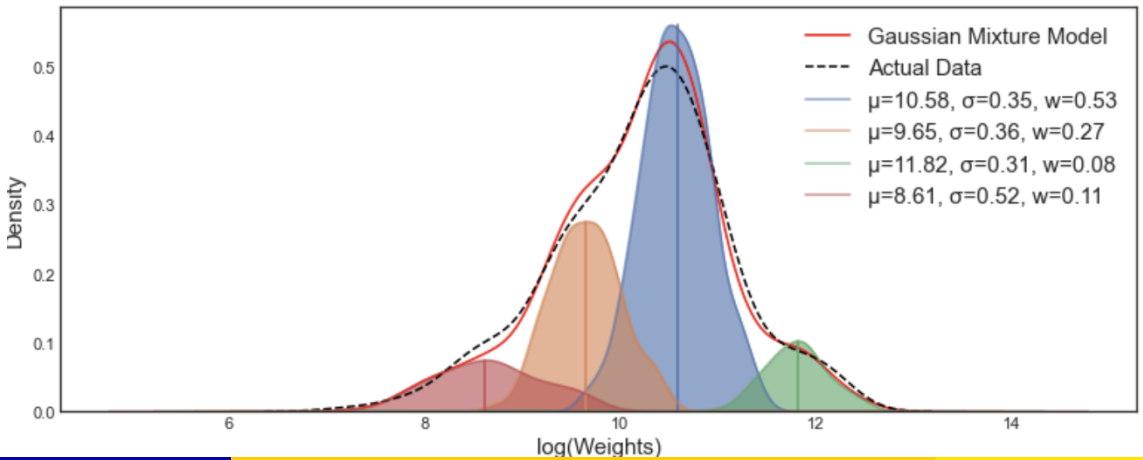
GMM - PDF

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

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.

Identifying anomalies from GMMs

Algorithm for Anomaly Detection based on GMM

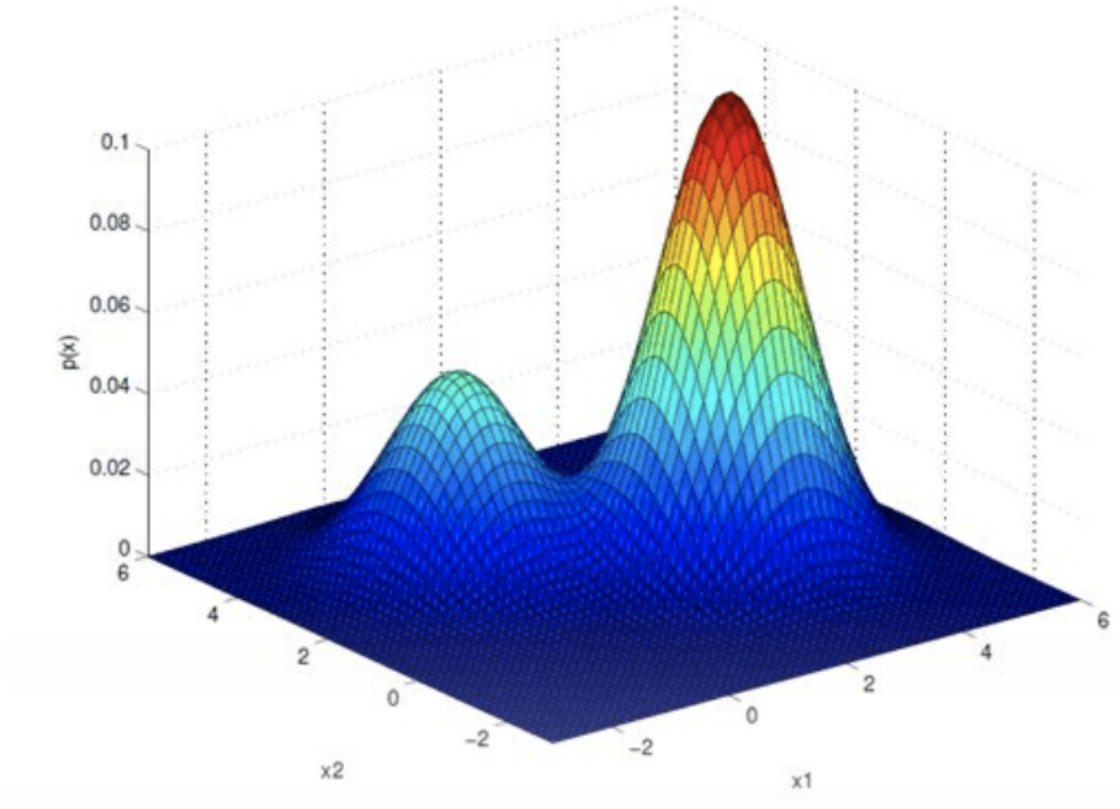
a. **Fit** a GMM model to the data with K clusters. K is a modeling choice!

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

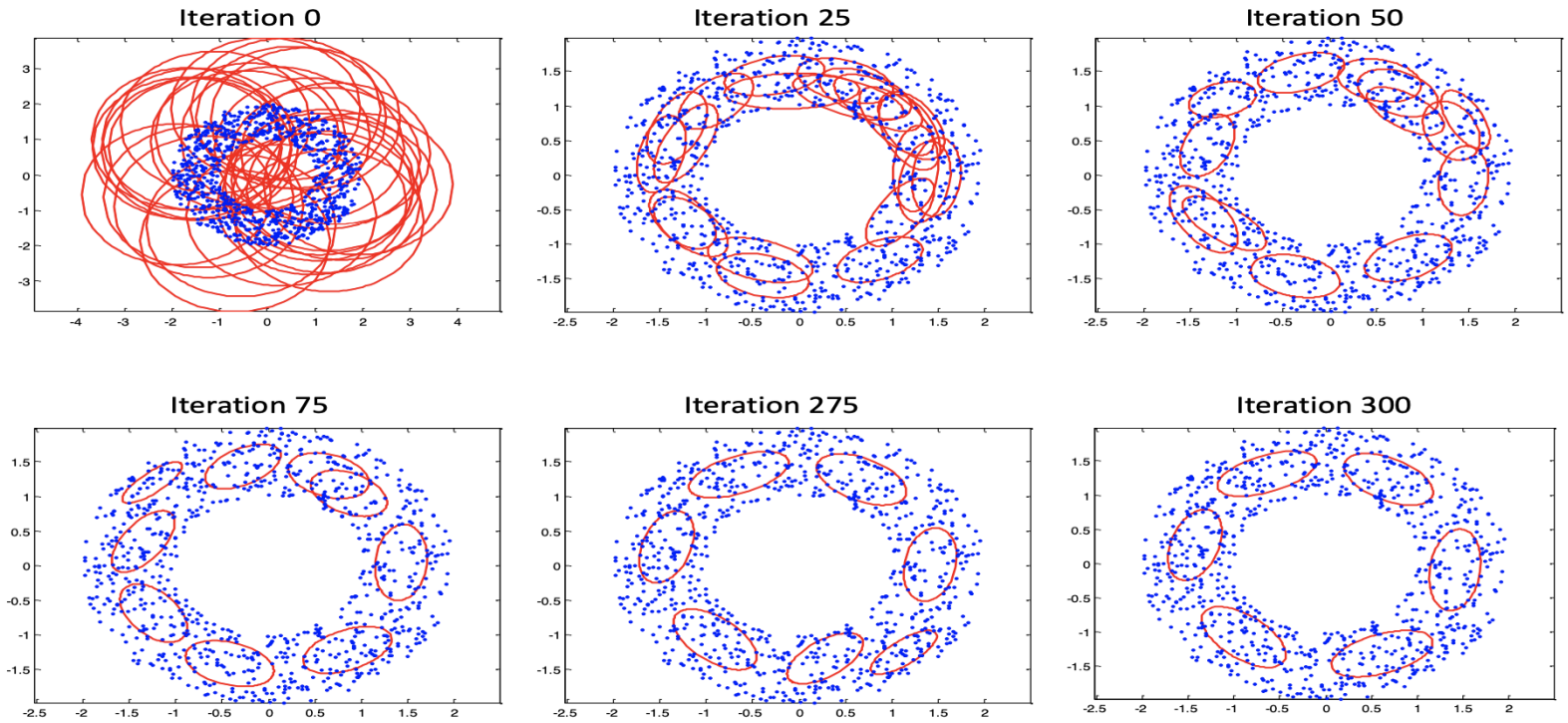
c. If $\max_k p_k < \alpha$ where α is an **anomaly probability threshold**, then flag x_j as an anomaly.

How to compute the probability p_k?

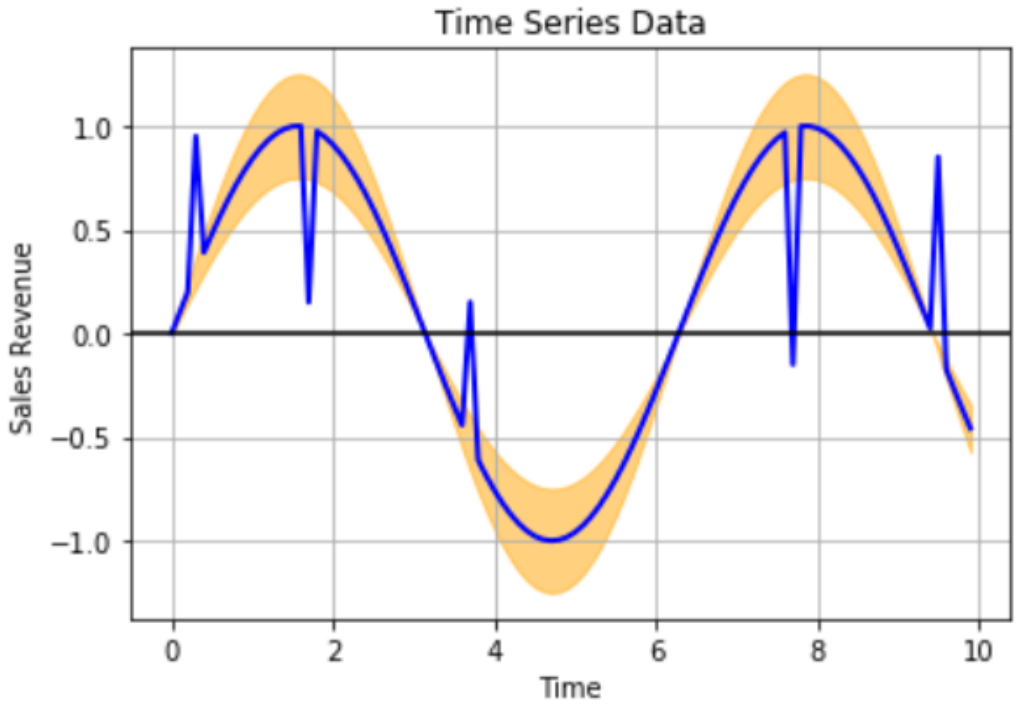
GMM Anomaly Detection Example



GMM Learning through EM algorithm



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

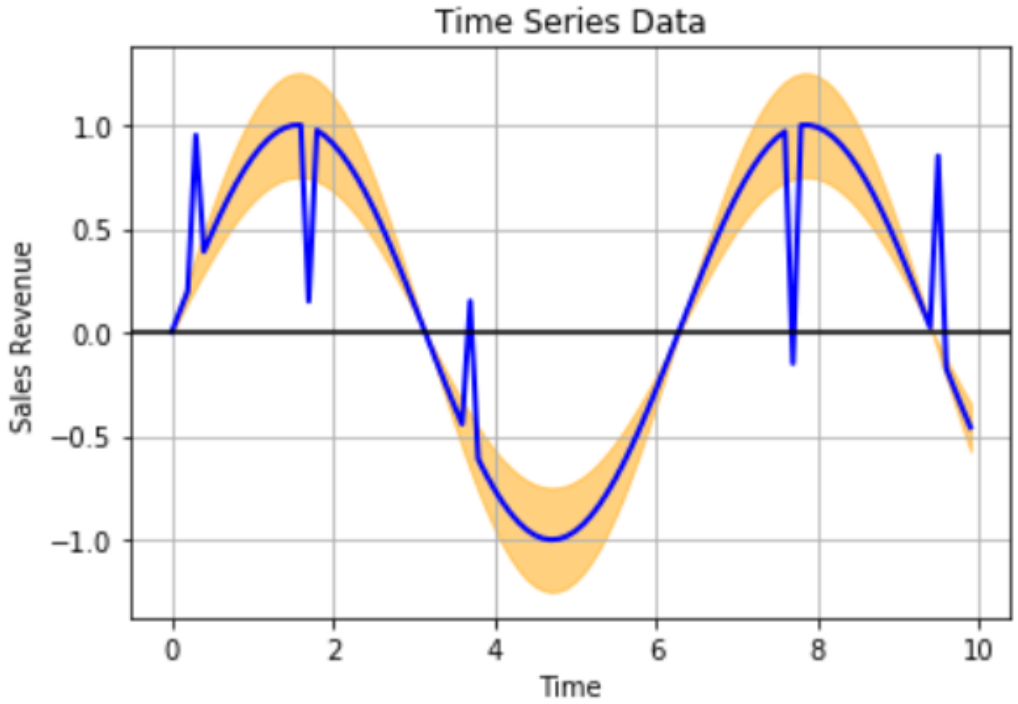
Un-supervised Learning

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?

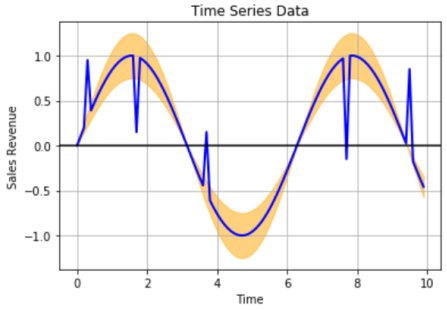
Time-series Anomaly Detection



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 #6

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

Stock Price Prediction

Can this be modeled as an anomaly detection problem?

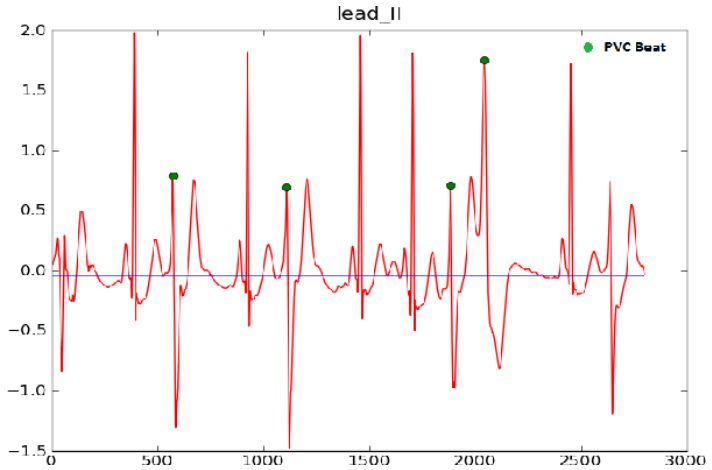
Can you build a ML model that can predict when to buy a stock and when to sell a stock to maximize your profits. How exactly would you do it? And how could you cast it as an anomaly detection model?

Stock Price Prediction

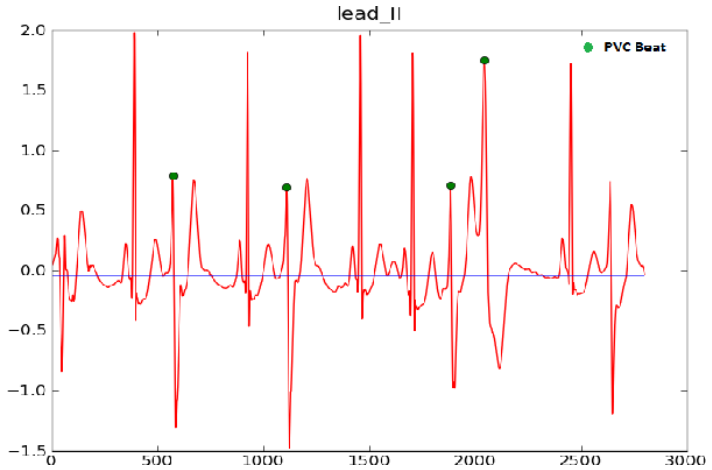
Local Window

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

Arrythmia detection



Arrythmia detection



ICE #7 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.