Reference: WaterBot
Industry: Energy & Utilities
https://www.ibm.com/case-studies/waterbot

Goal of WaterBot project: build a smart water management with an early warning detection system and can diagnose any leaks up-front to ensure efficient water management.

Analytics problem statement: how to detect and predict whether contaminant exist in drinking water. Is the water safe to drink? If there is a water leak happening, what is the cause?

I will propose two kinds of steps about how to solve this problem:
1. IoT infrastructure
2. Analytics

1. IoT infrastructure
Put sensor in each house that can measure each contaminant. We need to put multiple sensors, each of them will measure particular contaminant. In addition, sensor to measure volumetric flow rate, temperature, and pressure will also be installed to detect any anomaly in water flow.
Beside putting sensors inside the house, we also have to put sensor alongside the water distribution pipe to detect contaminants and finding out if leak is happening.

There are several substances that are considered as “drinking water contaminants”. The list below shows some of the major contaminants in drinking water.

- Inorganic Compounds: Arsenic, Nitrate, and Lead
- Organic Compounds: Atrazine, DEHP, TCE, and PCE
- Disinfection Byproducts: Trihalomethanes and Haloacetic Acids
- Radionuclides: Uranium and Radium
- Infectious agents: Bacteria, Viruses, and Parasites

Source: https://trackingcalifornia.org/water-quality/types-of-drinking-water-contaminants

The data will then be sent to cloud so analytics can be performed. Water management control command will have analytics and visualization tool to monitor the entire system.

2. Analytics
Use three analytics models:

1. Detect if a particular contaminant is high in each house using CUSUM.
Because there are contaminants, we have to make models for each of them. In this example I will just give example for bacteria detection.

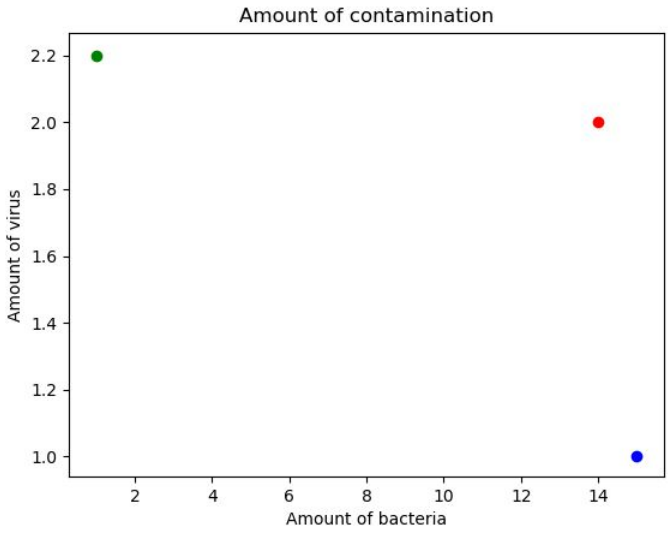
Given: bacteria sensor data
Use: CUSUM to detect increase in contaminant concentration
Result: alert that will be sent to both the house owner and water management control command when bacteria increases

The advantage of using CUSUM to detect change instead of simply using threshold on raw data is it is more robust to sudden change due to noise in sensor measurements. CUSUM model will trigger alert only when there is significant changes in sensor reading.

This individual contaminant detection is important because we have to satisfy government regulation regarding drinking water. Maximum contamination level for each substance is listed here: https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations

2. Classify whether the water in each house is safe to drink.
Previous model that uses CUSUM will only detect increase in each contaminant.
Consider the following example case (the numbers are just for illustration purpose):
Virus alert will be triggered when the amount of virus is 2.2 or more (point green will be considered dangerous level).
Bacteria alert will be triggered when amount of bacteria is 14.5 or more (point blue will be considered dangerous level).
But there might be a case where amount virus is less than threshold and amount of bacteria is less than threshold, but because both virus and bacteria is present in the water it becomes very dangerous (point red).

Point	Amount of bacteria	Amount of virus
green	1	2.2
blue	15	1
red	14	2



Example that I just gave is only for 2 variables, we have to consider all the variables of contaminants.

Before using data as input into classification model, I will smooth the data to increase signal to noise ratio of the data.
To smooth out the sensor reading, we can use exponential smoothing. The alpha value will depend on our confidence in each sensor reading. The better the sensor, the smaller the alpha will be because we can be confident about our current measurement.

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

where: F_t = new forecast
F_{t-1} = previous period forecast
A_{t-1} = previous period *actual* demand
α = smoothing (weighting) constant

Given: each contaminant time series data
Use: exponential smoothing
Result: noise reduced data

After getting contaminants data I will train the model using SVM with soft margin to classify whether the water is safe to drink or not. One important thing when choosing SVM is to make trade off between false positive (classifying safe water as unsafe) and false negative (classifying unsafe water as safe). I will argue that the effect of false negative is much worse than false positive because this is related to human health and public trust. Thus it is better to be on the safe side.

Given: noise reduced contaminants data
Use: support vector machine
Result: classification of safe or unsafe water

When there is a home that classified as unsafe, it will trigger an alarm to house owner and to water system command center.

3.Use analytics to find the root cause or find the main pipe that causes water leak.

Pipe that is leaking will have lower volumetric flow rate and lower fluid pressure. Using this data we can use analytics to automatically find where is the source of water leak. The leak can happen in the house or main water distribution pipe.
Approach to solve this problem is to use graph partitioning. This algorithm is not taught in the current course, but you can find the detailed algorithm here https://arxiv.org/pdf/1606.01754.pdf.

Given: volumetric flow rate, and pressure data
Use: spectral graph partitioning
Result: leakage localization

When leakage location is determined, the water company can send repair team to fix it. We can add features such as when leakage is happening, it directly send short messages or email to predetermined repair team and this will save the company money (leaked water is a waste of money).