#WaterData18

Hack Pad

Friday 18 May & Saturday 19 May

Add all of your notes, photos, discussion, concepts, etc here!

OS API Key: R4Pd6agEdzOqr0Tm0GJrF9Q6vAHUYg1j
Supporting documentation: https://apidocs.os.uk/docs/os-places-endpoints

Yorkshire Water are releasing open data for the first time, so there are still improvements to be made, For example, the DMA data can be linked to other things but the acoustic logger currently doesn't even have DMA data in it. The loggers are often fitted to property stop taps (which is personal data) so location can't be published.

#WaterData18 Teams

Customer's Index

Duncan Macintyre - Yorkshire Water - duncan.macintyre@yorkshirewater.co.uk
Timothy Ip - PA Consulting - timothy.ip@paconsulting.com
Kara Hazelgrave - water@leeds / University of Leeds - K.Hazelgrave@leeds.ac.uk
Ian Longley - ianlongley@live.co.uk
Liam Coatman - liam@asidatascience.com (DJVu recommendation)
Paul Hazelgrave - TugMeLug Productions - contact via KH

- Shazaam for acoustic loggers - aka training loggers to identify sounds in order to better "see" and profile leaks more accurately (i.e. know what is a leak and potentially quantify)
    - Run loggers for longer and different times
    - Train different sound profiles & sound at household / street / DMA
        - Sounds to profile (acoustic fingerprinting):
            - Domestic (taps, showers)
            - External "other" (lamp posts, pressure valves, etc.)
            - Leaks (size)
        - Repeat profile for differing pipe types; seasonal change
- Getting customers to care about leakage
    - Yorkshire could have the lowest levels of consumption per person per day!
- Profiling DMAs by characteristics, including customer behaviours
- Questions:
    - Is pitch sufficient for accurate identification?
    - Is it possible to identify an audio pitch that = stress likely to result in burst (e.g. indicate reduced pressure)
    - Would other data be helpful? E.g. prioritise more strongly if specific sound + weather in area &/or customer contact spikes - use this to improve the DMA RAG (prioritisation, Red Amber Green)
    - Don't appeal to customer's pocket, appeal to their conscious

- Combine algorithm with in-situ PIGs (in pipe robotic) - could PIG act as roaming acoustic logger
    - Could include epoxy resin repair for leaks

- Tool recommended: Open Source - DJVu http://djvu.sourceforge.net/
- Create a library of sounds of interest, add to library of sounds; compare sound spectrums

- ORIGINAL DIGITAL LISTENING STICKS - collect audio tracks
    - Sound library accessible by stick
    - Pitch / audio reading results in stick providing operator with suggestions

- Share tech with public for phones & provide discounts for customers who take readings? Tap the app
    - UBER technology for offering customers discounts for providing a reading

+ Team recommends that loggers are adapted to be self-powering, e.g. water power so that loggers could be run for longer periods

RE: SOFTWARE (consultation with PH)
Consider alternatives to DJVu because of compression constraints within DJVu;
https://www.steinberg.net/en/home.html Steinberg have excellent products, including Cubase & Nuendo

- Sadie - top end professional recording software
  - http://www.sadie.com/support/download_v6.php
  - Free 30 day trial
    - Consider: LRX2 flexible location recorder (has more features than needed, made for TV outdoor locations, "rugged" design) cost not listed: http://www.sadie.com/products/lrx2/main.php
- Cubase, by Steinberg
  - https://www.steinberg.net/en/products/cubase/comparison.html
- Nuendo, by Steinberg
  - https://www.steinberg.net/en/products/steinberg_trial_versions/nuendo.html
  - Free 60 day trial Nuendo 8
- Reaper
  - https://www.reaper.fm/download.php
  - Free 60 day trial; $225 purchase price for commercial use (not clear if this is an annual fee, a one time fee or if further fees might be associated with future upgrades)

But get what you pay for--to capture top end noise or specific

Compression through microphone on mobile phone will cause problems - want to be able to put you own compression ratios on - the high and low ends will be the discriminating information required to diagnose leakage & quantify.

From 0 to 50,000 htz -- we can't hear the extremes, we can only HEAR 20 htz to 20,000 khz

Vibration / movement sensors also required

Time Lords

- Scott Stevenson - ASI Data Science - scott@asidatascience.com
- Gary Willis - ASI Data Science - gary@asidatascience.com
- Liam Coatman - ASI Data Science - liam@asidatascience.com
- Jan Freyberg - ASI Data Science - jan@asidatascience.com
- Laurence Cowton - ASI Data Science - laurence.c@asidatascience.com
- Andy Sleigh - University of Leeds - p.a.sleigh@leeds.ac.uk
- Scott Clay - Oakland Data - scott.clay@theoaklandgroup.co.uk

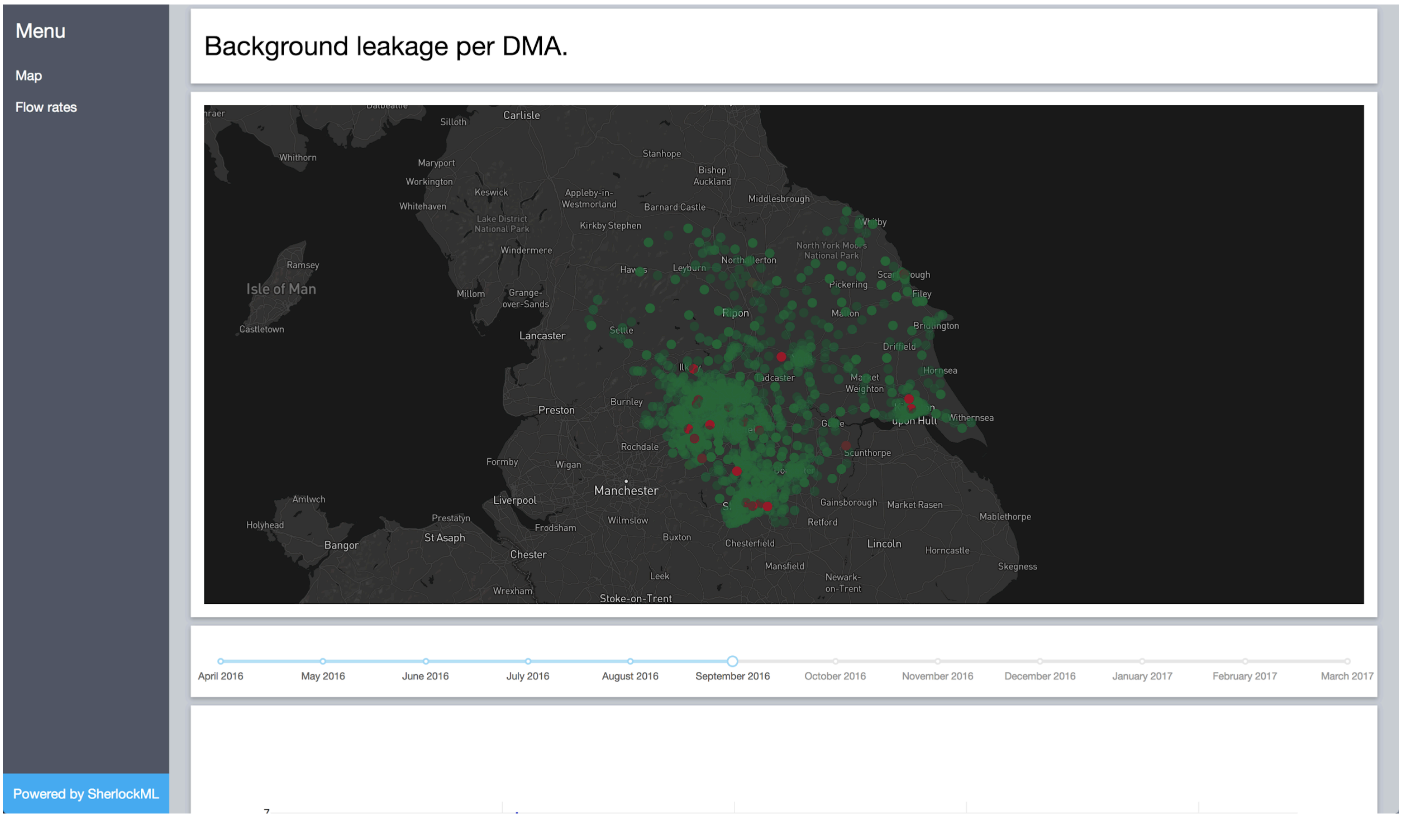
Project: Predicting Background Leakage

We used a novel combination of linear regression and random forest regression to predict the flow rate in each DMA using features such as the number of properties per DMA and data extracted from open data to calculate the average rooms per house. Using these predictions we identified outliers that were indicative of anomalously high flow. This data was then passed to the plotly dash app (below).

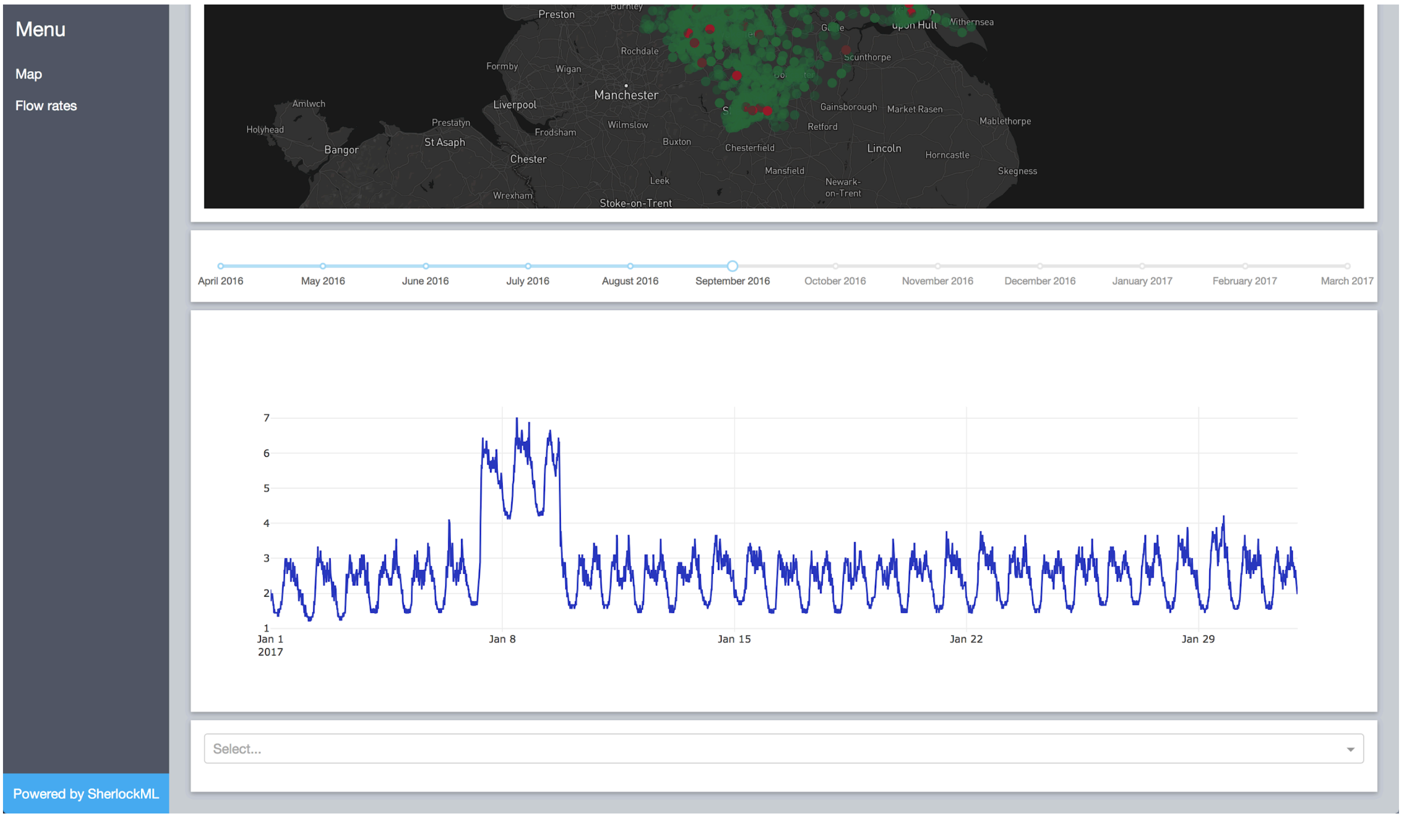
Here is a git archive of some example map views of the data provided.
https://bitbucket.org/andy_sleigh/ywdatamapping/src

In particular it plots the postal sector boundaries and can be linked to the DMA based leakage and/or usage to provide some visualisation a geographic context. Warning: It's a bit of a HACK, but should be easy to follow. It also includes the example that Clive from from OS provided as an example of the API they have given us.

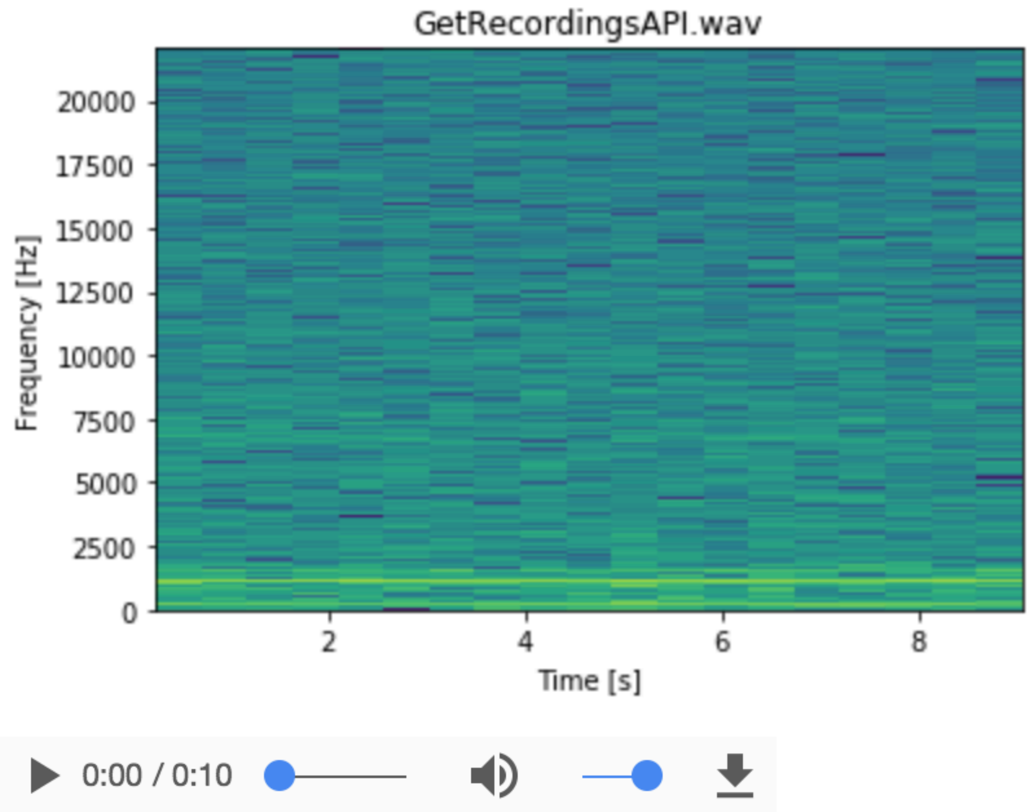
We built the output of this machine learning model into a map-based visualisation that flags up DMAs with unusually high background (night-time) leakage. This is currently displayed on a monthly basis, but with live data could be turned into a live-updating map:



In addition, to support the work of detecting "outbreaks" that the leakage team currently does, we linked this to visualisations of the flow rate on a daily basis. This should allow analysts to detect weather high night-time flow is due to background leaks or due to "outbreaks":



We also explored the acoustic recordings. Below we show a spectrogram (a plot showing the intensity as a function of frequency and time). The different recordings (with and without leaks) have very different spectrograms. With more data, we propose to build a machine learning algorithm to identify audio different sources. We will then be able to alert the operator when a ‘leak’ signal is detected.



The Detectives

Salma Afzal @SalmaAfzal_
Martyn Rayner

Objective: Understand how the demographic of population impacts on localised DMA usage / leakage calculations.

Using understanding from the demographic to look at trends and insights to understand usage.

Change of times for using Acoustic Loggers dependant upon, night workers, property users / owners.

Individual loggers used for certain commercial usage. Greater information collected for this type of usage.

Less information / understanding for residential (smaller commercial) usage / leakage. Could be dependant upon demographic details: no. of people, age of people, age of property, type of property, type / age of appliances, ethnicity, religious groups, affluence, students, shift workers / working patterns, rural…….non- 9 to 5 households, multiple jobs. Weekend trends

What are the future job type / working hours / patterns?

Use of water meters to record water usage (24/7) backed up with human response (user research/user needs / diary) ....paper, video, app, electronic / paper version. Different interactions with different people. Possible link to other energy / utility usage...join up water usage to power usage?

What data is available? What data do we need to find? How do we do this? To record the data- need a flow logger on the house, it's a blue box which records water flow every 15 mins.
https://ore.exeter.ac.uk/repository/bitstream/handle/10871/21312/Using%20Smart%20Meters%20for%20Household%20Water%20Consumption%20Feedback.pdf?sequence=1
Something like https://www.smappee.com/uk/gas-and-water-monitor (identifies appliances and their usage to use for discussion)

Diary record to capture :

- What they are doing and when (log times) ? E.g running a bath, shower, cooking, toilet flushes
- Interviews to find out when they go out ? what they do in the morning ?
- Patterns of behaviours - water use age can predict jobs in future.
- Google water usage questionnaire
- Paper diary or app option - visit the diary - collect the diary.
- Incentive £50 to take part, off their next bill

Day 2

Objective: Understand how the demographic of population impacts on localised District Metered Area usage / leakage calculations.

Using understanding from the demographic to look at trends and insights to understand usage.

Focus on people who work night shifts in a District Metered Area to gather additional insight, normal flow rate in the day versus unusual activity of water use age spikes to narrow down actual water leakage or a false alarm?

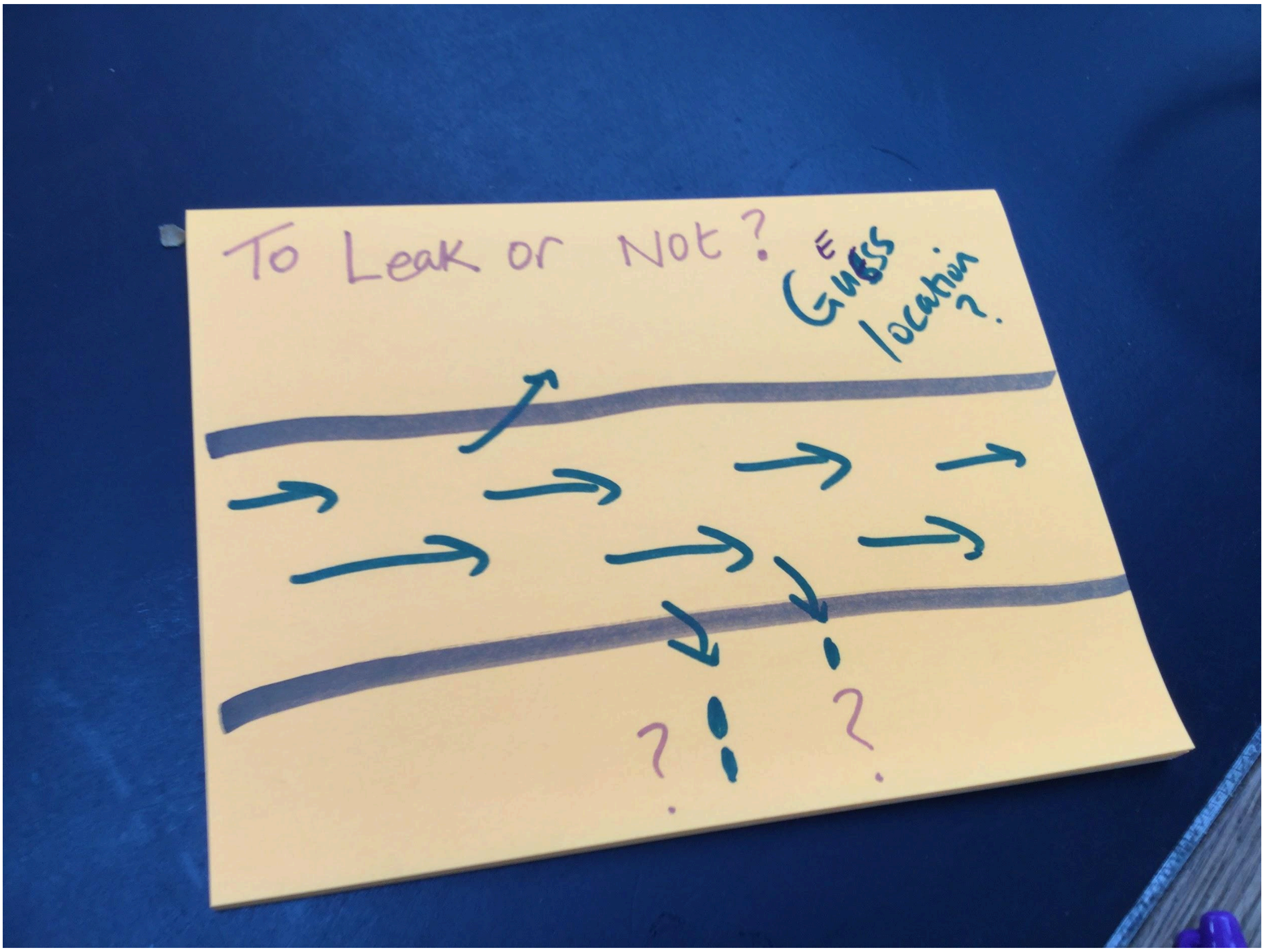
Advantages

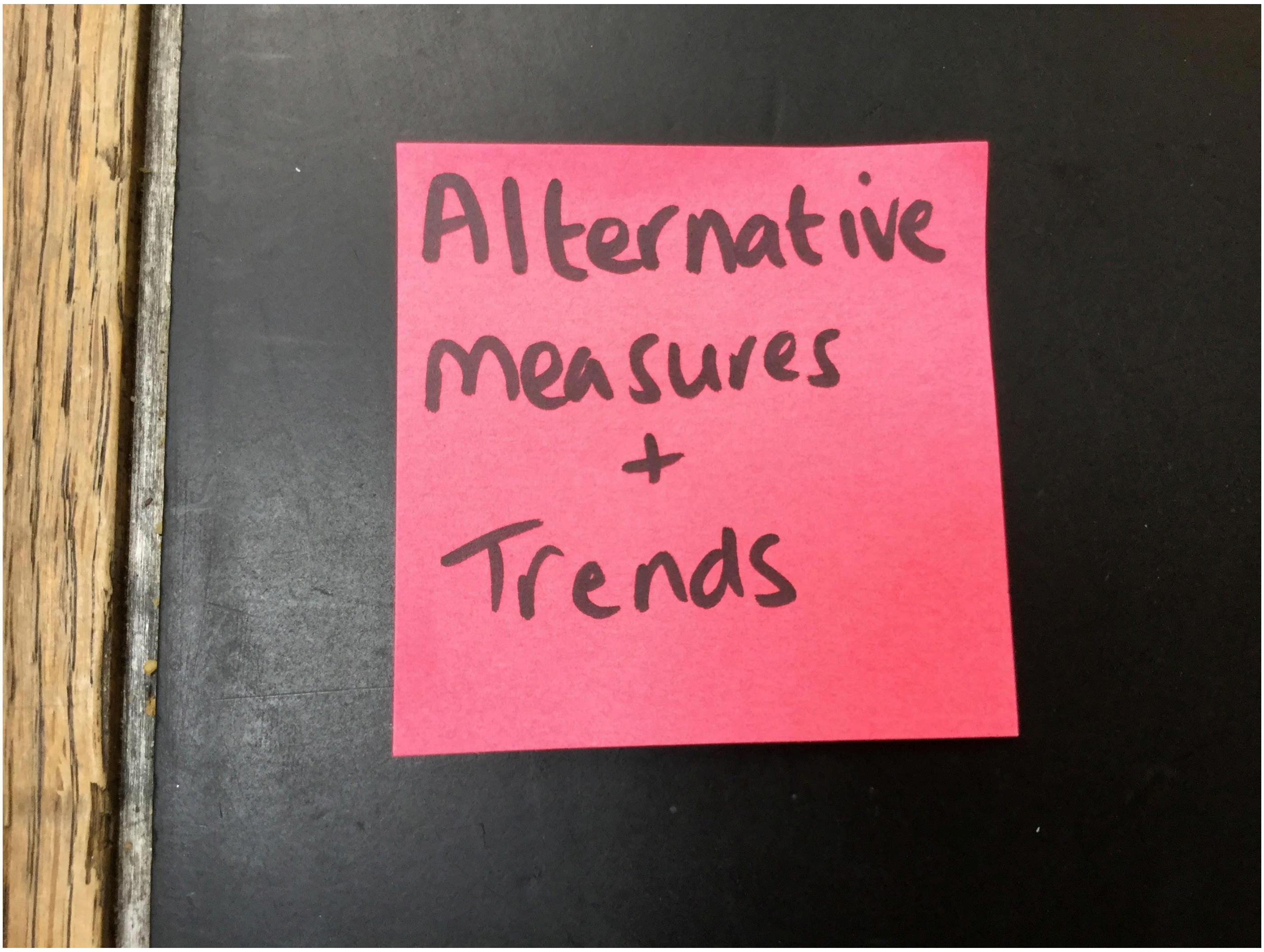
- From DMA easier to identify sample required
- Can obtain additional information about any data gaps
- If more information is obtained about the unknown then extra useful data can be gathered to quantify and narrow down area of leakage.
- If at night then can make comparisons between night and day to identify any trends or unusual activity

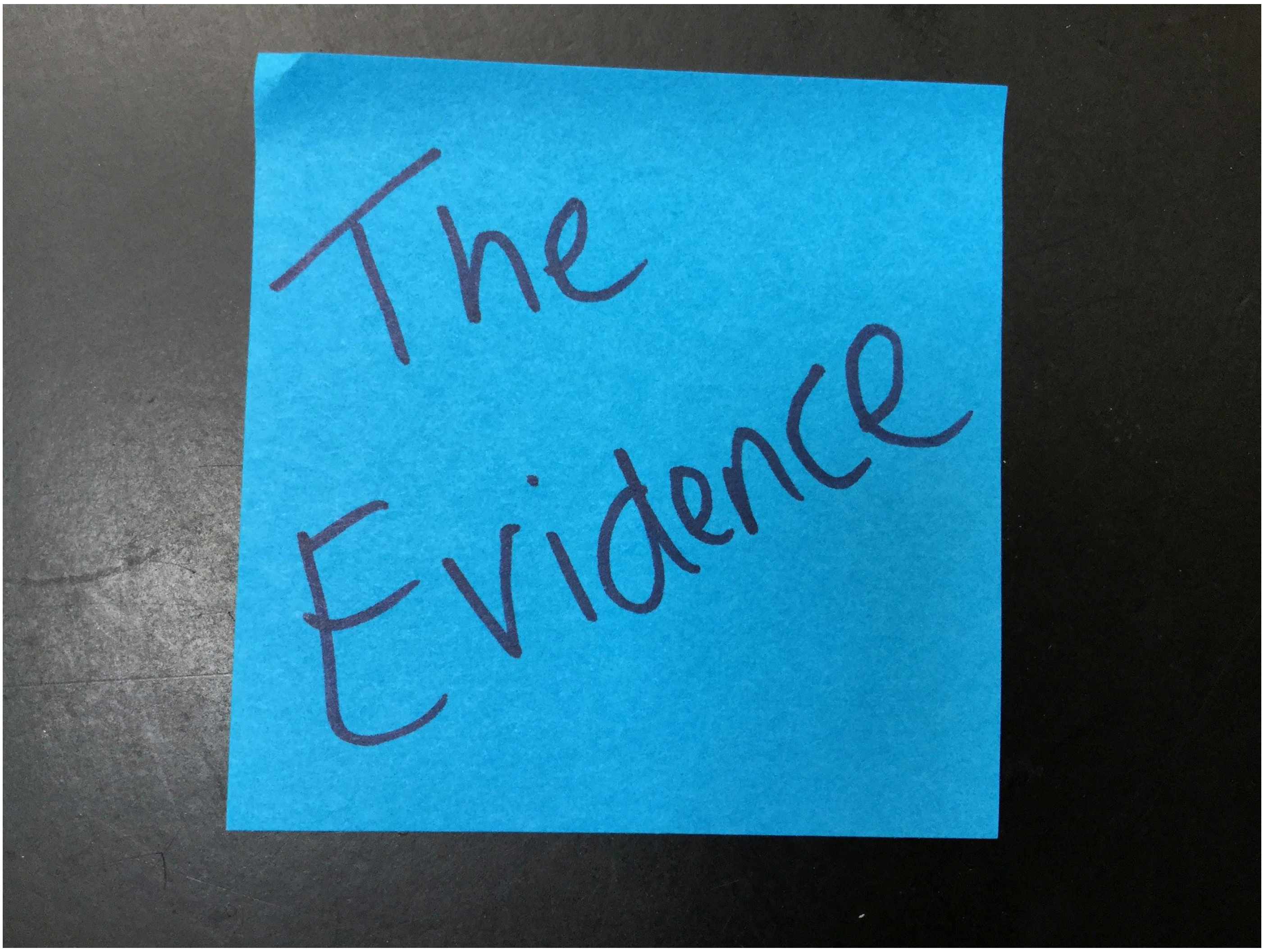
Water Detectives- Show and Tell

The Crime Scene (CSI)

Name that leak ? Or is it name that location ?

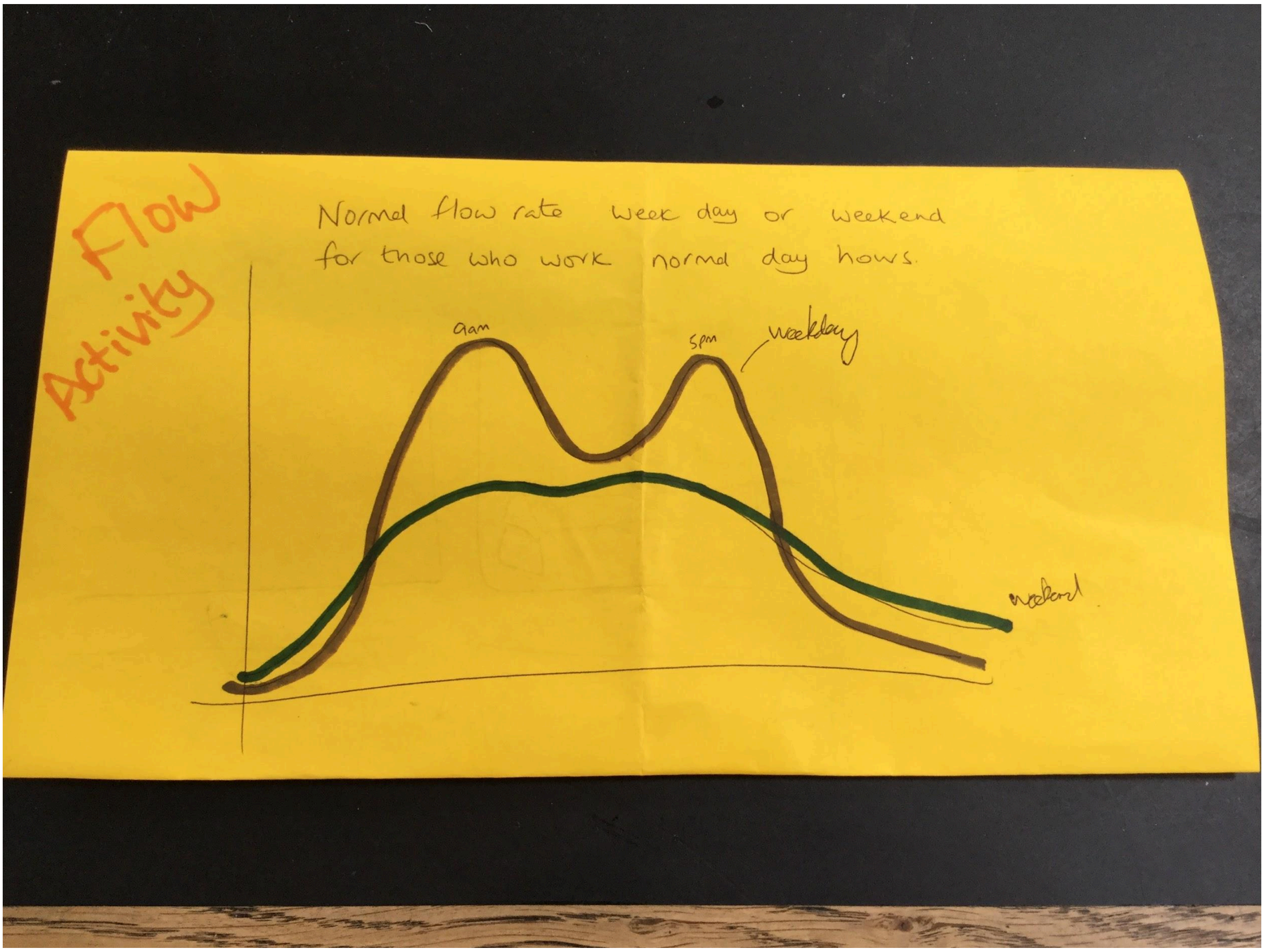



Alternative
measures
+
Trends

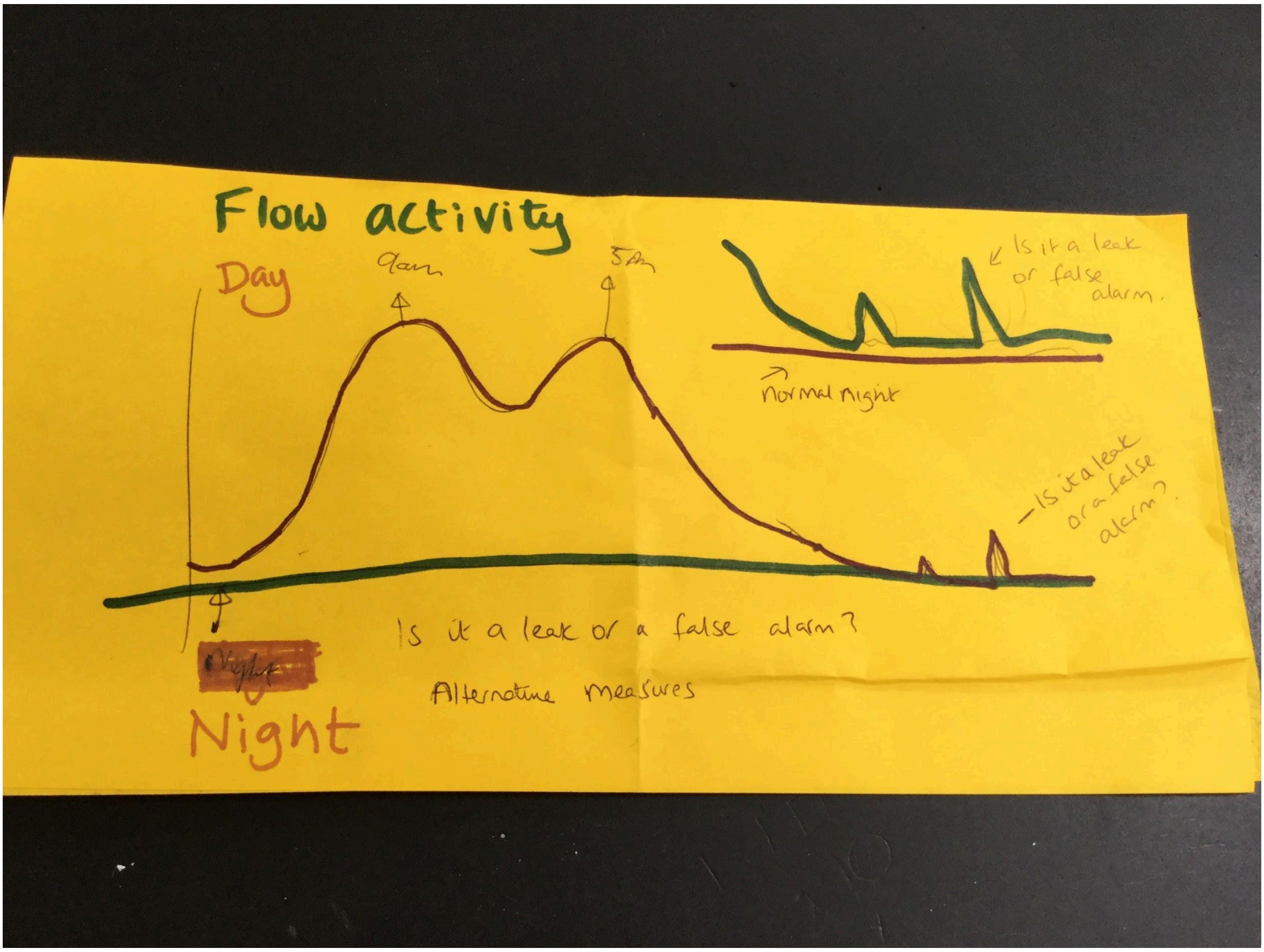


What do we already know ? Is it a leak or a false alarm?

Patterns of behaviour

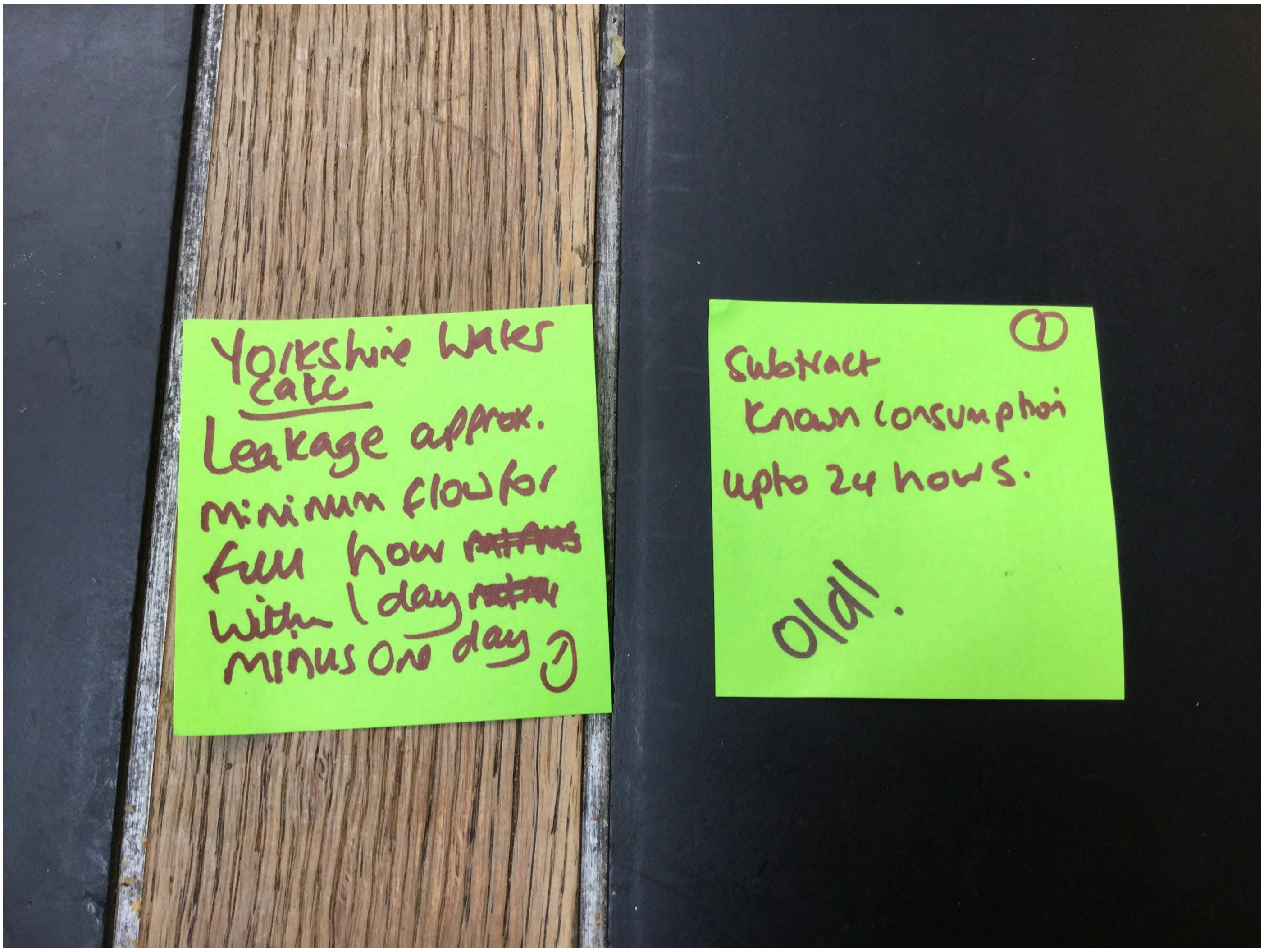


Flow Activity
Normal flow rate week day or weekend
for those who work normal day hours.
9am
5pm
weekday
weekend



Yorkshire Water calculation

Are these evidence methods outdated ?



How can we distinguish between leaks and false alarms ?

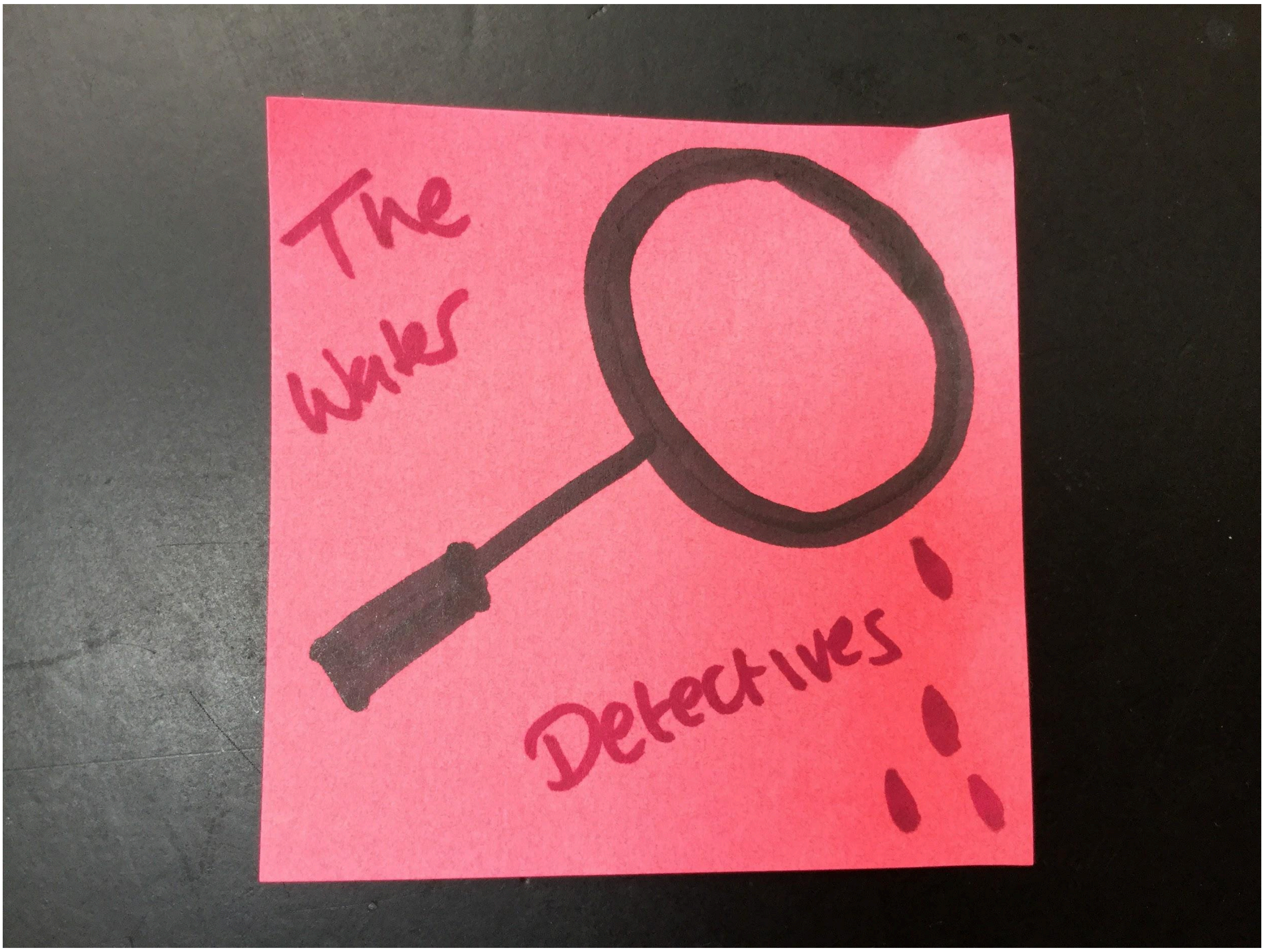
- Understand how the demographic of population impacts on localised District Metered Area (DMA)
- Using understanding from the demographic to look at trends and insights to understand usage.
- Focus on people who work night shifts in a District Metered Area to gather additional insight, normal flow rate in the day versus unusual activity of water useage spikes to narrow down actual water leakage due to a burst pipe or a false alarm?
- A flow meter in areas to measure water flow from night shift workers to understand habits and behaviours compared with day activity to identify and narrow down areas of investigation
- Focus on sample from District Metered Area, 1 device needed so cheaper
- More targeted and accurate allocation of resources i.e send out an inspector for investigation

Advantages

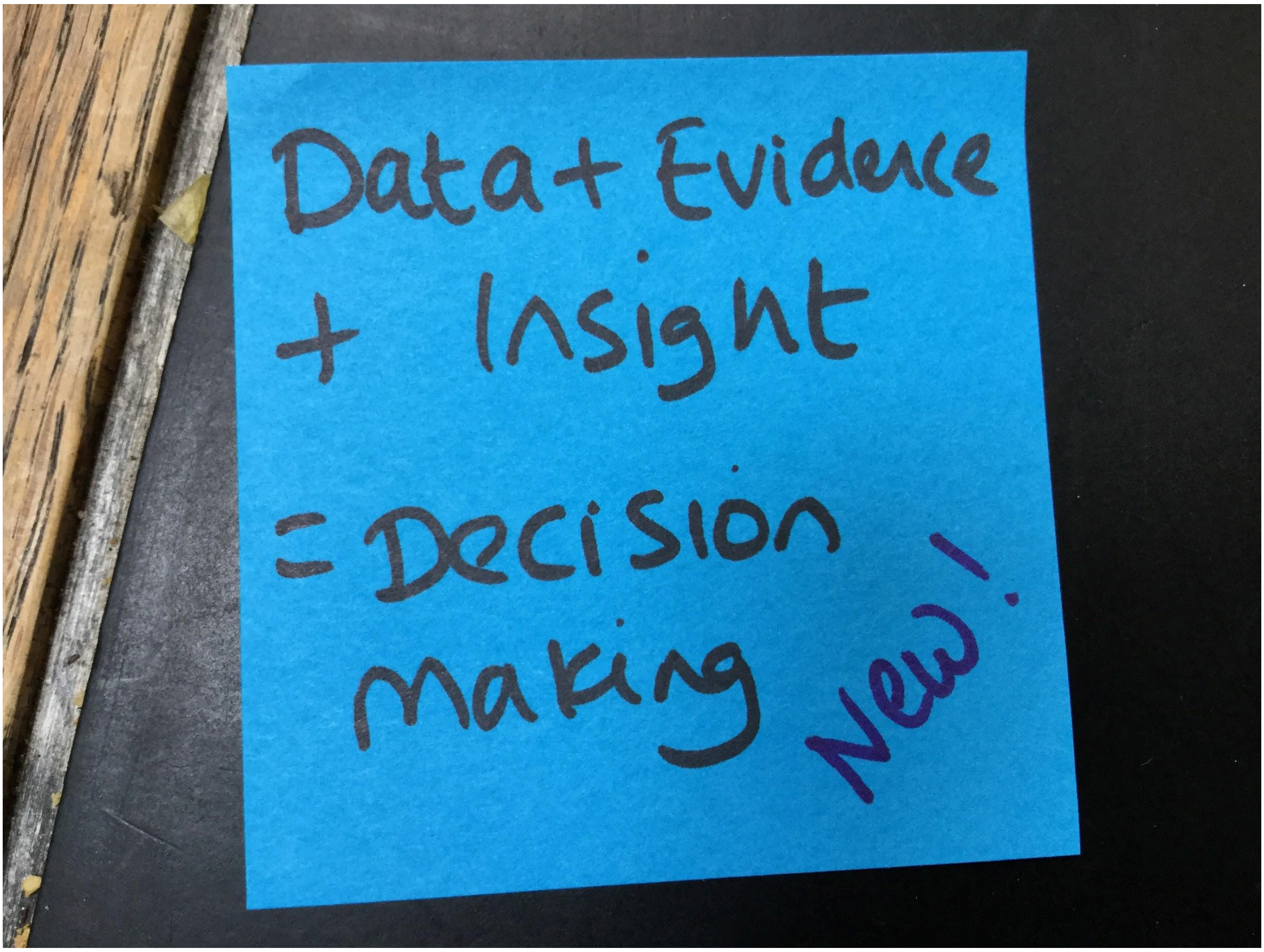
- From DMA easier to identify sample required
- Can obtain additional information about any data gaps
- If more information is obtained about the unknown then extra useful data can be gathered to quantify and narrow down area of leakage.
- If at night then can make comparisons between night and day to identify any trends or unusual activity

A case for the Water Detectives on the scene

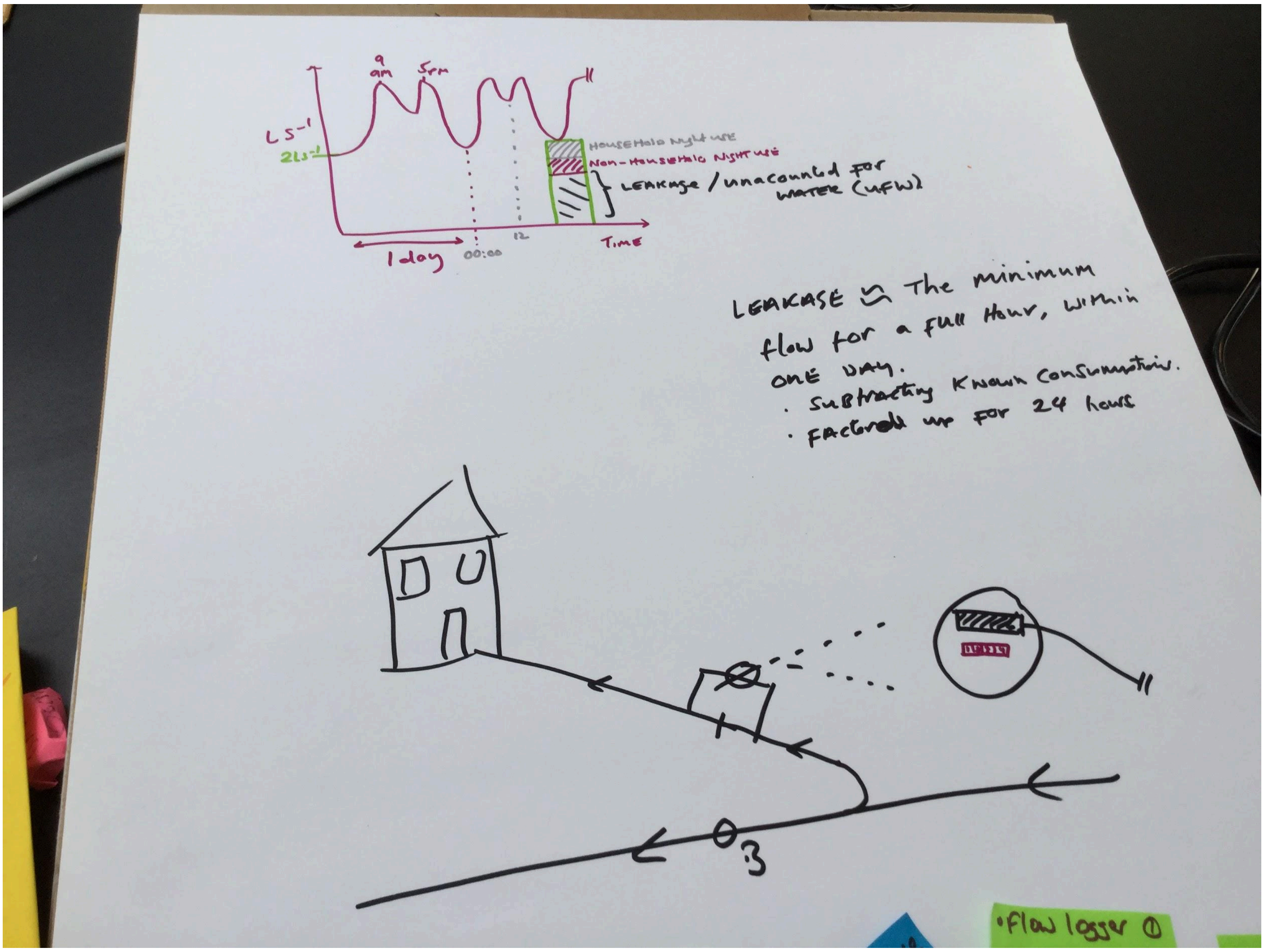
The Water Detectives calculation - New methods



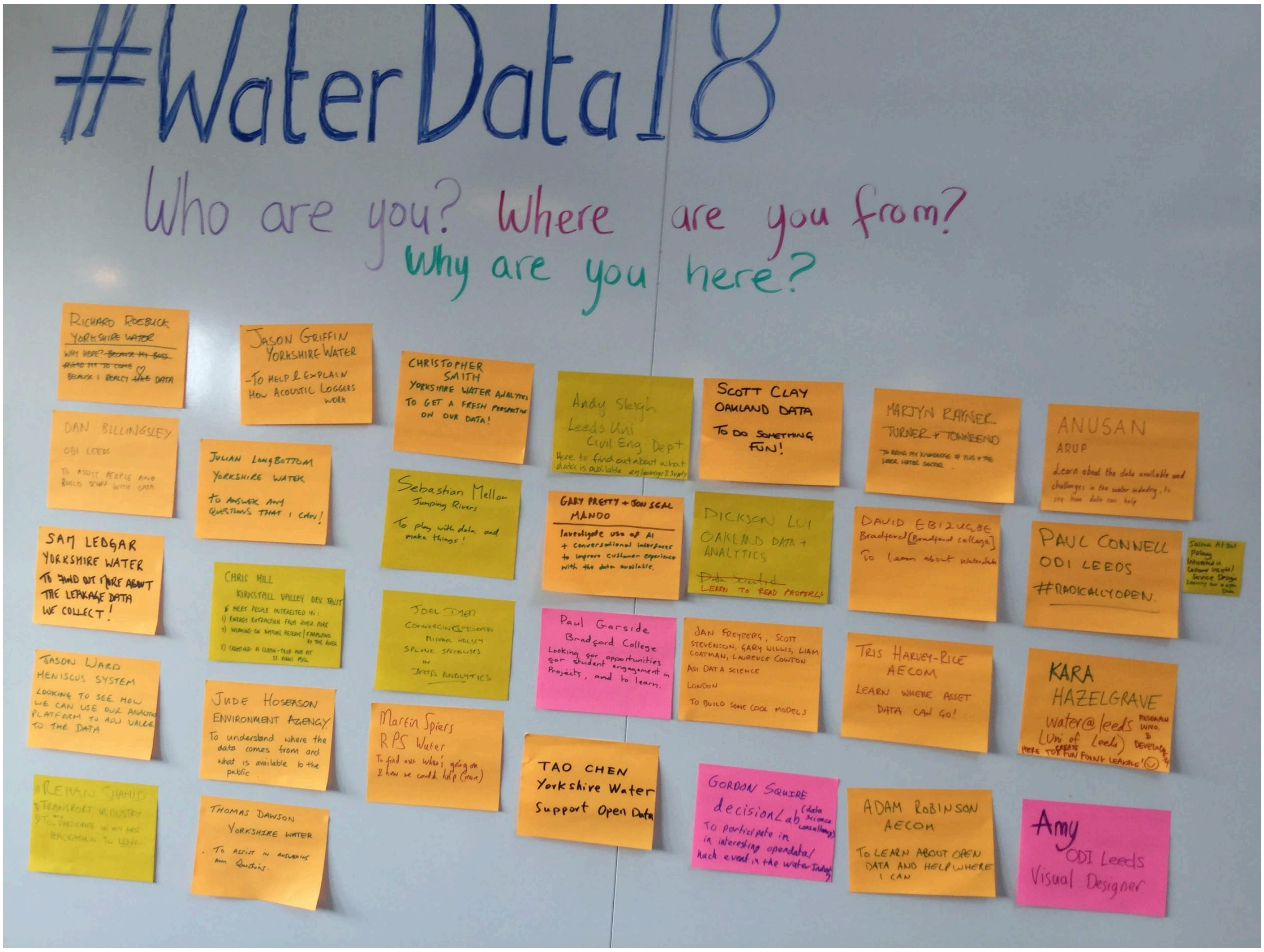
Futuristic methods to predict behaviour would add more value to existing data and identify trends and insights to understand customer behaviour better.


Data + Evidence
+ Insight
= Decision
making
New!

Interviews with a range of water experts to understand leakages - what's real and what's a fake ? Red herring? Confused?



And further additional expert advice which was either accepted or rejected



Further exploration of the problem and methods is needed to plug gaps

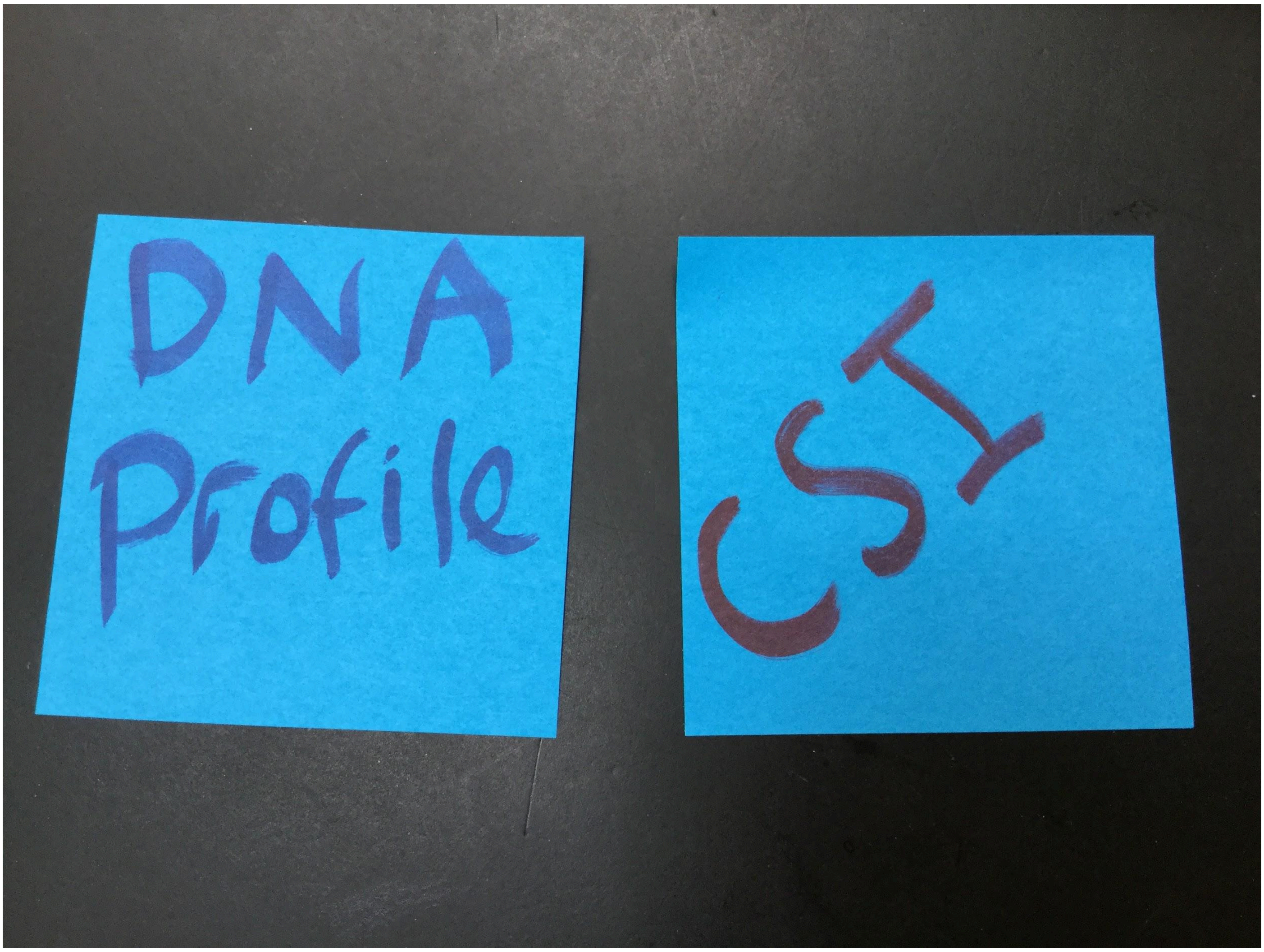
- Trial on night shift workers to better understand customer demographics as a future predictor of behaviour ?
- Can we predict future occupations in a District Metered Area from times of day or night water based on water consumption.
- Anomalies can help narrow leakage areas

Water Detective App prototype

An app/ diary for customers to log daily water usage

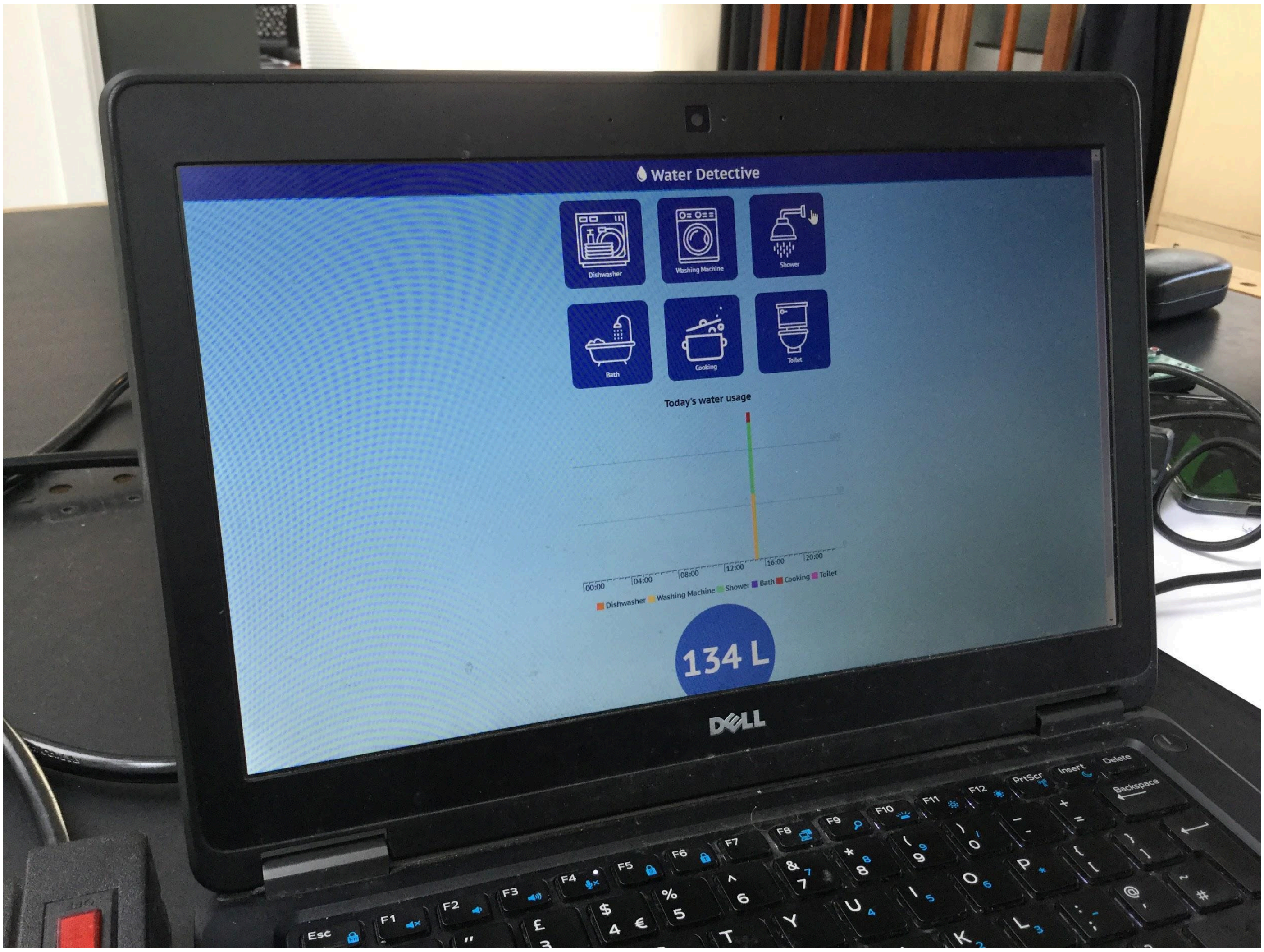
Pre-programmed with normal average consumption and 24 hr clock
Press a icon button linked to activity identifiable by visual pictures
Press select e.g shower this informs you on average how much you have used & is recorded on the water monitor stacked bar chart. The different colours tell you the breakdown of the different activities logged.

App data & logger data Generates a DNA profile



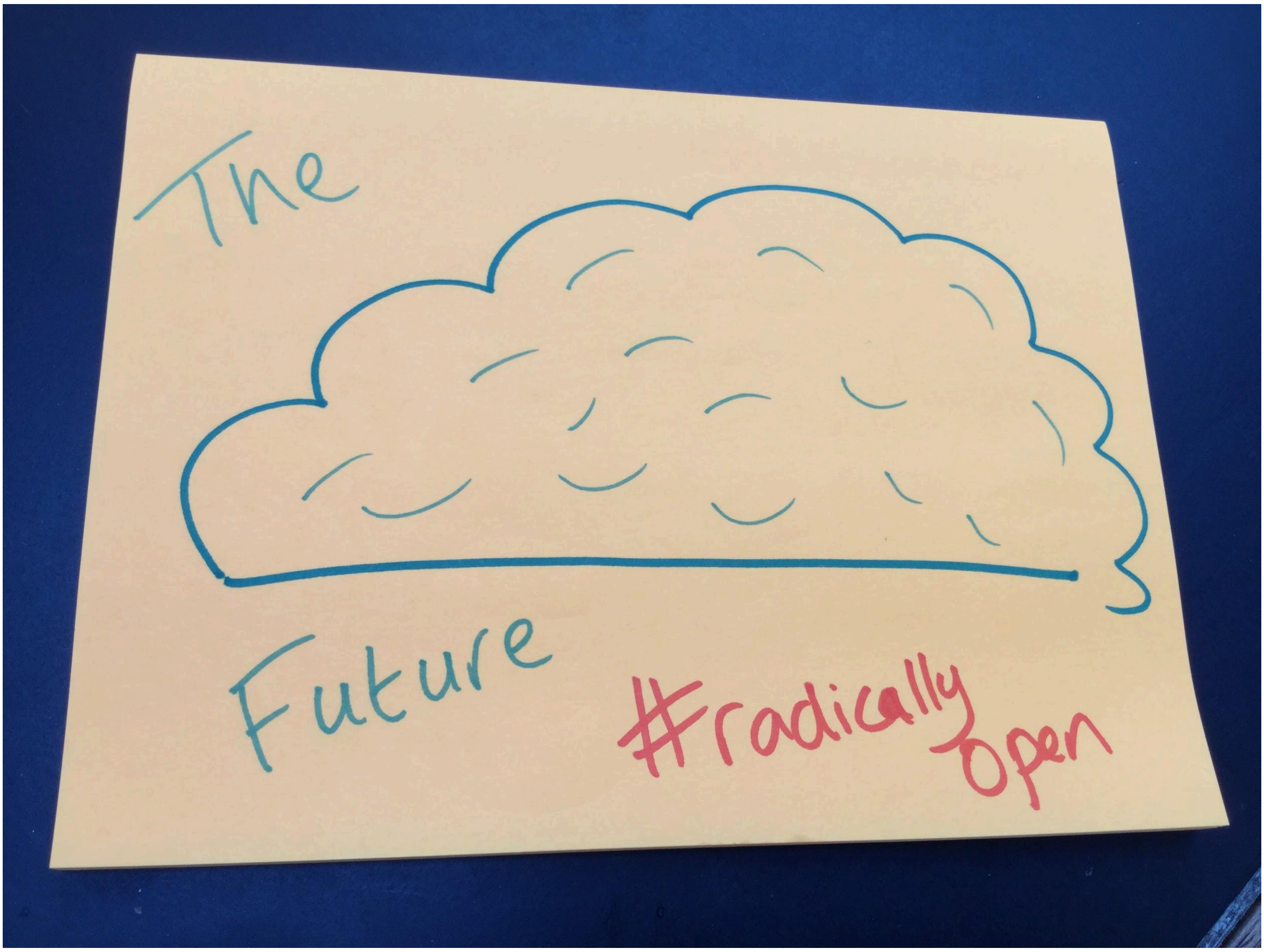
- Water consumption at specific times
- Latest logger status can help to prioritise resources and send an inspector out to investigate as this sends data every 15 minutes into a portal. Data then uploaded at 30 min intervals
- Grey leakage
- Purple Worth investigating
- Brown - normal

Prototype of the screen



The future

- Patterns of behaviours - water use age can predict future job types in District Metered Areas
- Customers may change habits based on feedback
- Night time investigation causes less disruptions if leaks need fixing
- Targeted resources and less waste as leaks are # radically open



Explorers

Which DMAs should be focussed on?

Based on

- Average profile for domestic customers - from analysis of recorded values
- Number of customers - from customer billing data
- Socioeconomic breakdown of customers - ACORN dat already available to YW
- Heavy users etc -
- Etc...?

\+ Observed profile for each DMA

End product - GIS tool to show high / medium / low risk DMAs - prioritise acoustic logger installation

Continual monitoring to see if baseflow in a DMA is increasing

Areas where growth curve in base flow is steeper, accounting for population growth
Or - where growth curve is abnormally low - what’s different about these?

Outline process for hackathon

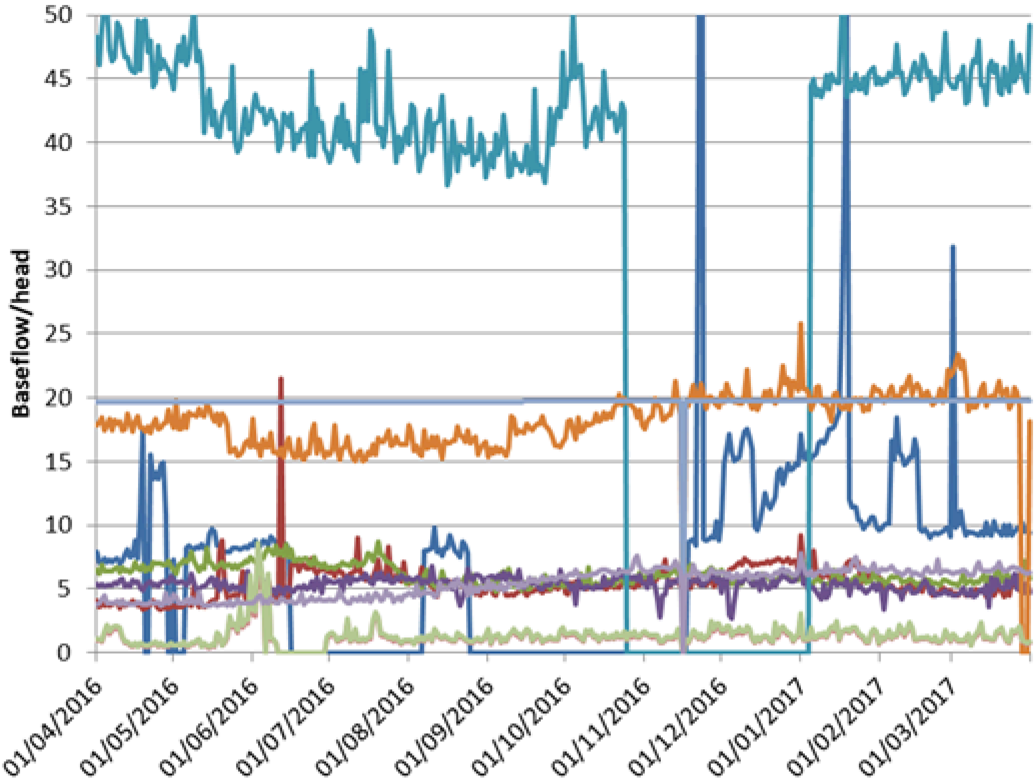
Step 1 - select 50-ish DMAs which are primarily residential (non-resi count <7)
Step 2 - estimate population for each DMA based on postcodes
Step 3 - normalise minimae by population
Step 4 - hence baseline consumption for each DMA
Step 5- extract minimae for selected DMAs
Step 6 - hence estimate infiltration = minimae - consumption & rank accordingly

We can’t actually visualise this in the way we imagined (as a GIS layer) as we don’t have geographic boundaries for the selected DMAs, which are anonymised. We’ll use some dummy polygons instead

Views

- Top 10 worst
    - With baseline and actual reading for specific time

Exploratory analysis: there are some big differences in per capita usage amongst the DMAs , this could be accounted for by variations in non-commercial flow and/or leakage and/or domestic consumption



===revised

1) Using a number of dmas that have few household or non-household, work out a baseline average usage per property
2) Using this baseline per property number, work out for each DMA property type, get the expected consumption and adding this together and taking it from the total flow given in the DMA flow data, work out the leakage
3) Rank this leakage by the Leakage flow value and show this on the voronoi diagram UI

Detail additional Influences that could affect this ranking and determinism of the leakage flow.

- SIC code for commercial
- Points of interest

- Socioeconomic - ACORN classification of residents
- Topography
- Asset data
- Metered / unmetered
- Pressure
- Ground conditions
- Weather

API call
https://map-rainuk4.meniscus.co.uk/Map/ReadData/CalcJson?exp=ToJson(%225af0be6b2381642d787f5965%22,%22average%20consumption%22,ToTime(2017,3,31,0,0,0),ToTime(2017,3,31,23,59,0),null)&apiKey=yyy7dR43!u

https://map-rainuk4.meniscus.co.uk/MAP/ReadData/Item/Childitems?idexp=GetItemId(%22Yorkshire%20Water\dma%20entity%20model\entities%22)&index=0&recursive=false&apikey=yyy7dR43!u

https://map-rainuk4.meniscus.co.uk/Map/ReadData/CalcJson?exp=ToJson(%225af0be6b2381642d787f5965%22,%22Est%20Dom%20Baseload%22,ToTime(2016,3,31,0,0,0),ToTime(2017,5,31,23,59,0),null)&apiKey=yyy7dR43!u

```
var _dataType = 'json';
 var _contentType = 'application/json; charset=utf-8';
```

Skynet

This may help other teams, logger samples could be submitted to them, Yorkshire could form a partnership with Stattus4:
https://www.economist.com/science-and-technology/2018/04/26/using-ai-to-trace-leaking-pipes

Using drones and thermal IR cameras to detect leaks, more useful in rural areas.

Neural Network:

Apply deepnet neural network to the existing data to obtain a predictive algorithm that can inspect live incoming data and alert when measures fall outside the established thresholds (i.e. a leak event is detected or future leak event predicted - thresholds established by the initial training task that produces the algorithm).

Explorers II

The Lone Ranger

p.garside@bradfordcollege.ac.uk

Where to look for the leak and where to dig your exploratory hole. Or…. digging the right size hole, in the right place, more often.

Ideas:

#1 After initial trigger, use multiple acoustic sensors (possibly mobile) to use distance of leak to help with triangulation.

- Where would the extra sensors be located for maximum effect? Use Predictive modeling to pre-define best locations, perhaps.
- How would these sensors be created? Arduino? Raspberry Pi? Off the shelf? Other?
- How are they attached to the infrastructure? Must be quick to aid fast response.
- How do they connect? Short range, Mobile data, LPWAN?
- Do they use existing acoustic sensor network and system?
- Use analytics to improve detection over time.

Prototype:
RPi or Arduino with acoustic sensor, physical connectivity to infrastructure, and network connectivity. Or Heath Robinson prototype.

#2 Use a nano drone submarine to navigate the mains.

- Must be sterile! Latex covered?
- Must be small.
- Battery powered or tethered?... What is maximum distance?
- Powered or goes with the flow… how do we capture at the end? Caught in a filter screen
- Remote controlled or automated? Auto-return to base.
- Camera or radar/sonar
- Use gps, can that work in a water pipe?
- or.. Track distance travelled by speed of travel or way points on inside of pipes.

Prototype:
RPi Nano or similar.

#3 Additive to water that can be detected via a sensor (hand-held, perhaps)

- Must be colourless, odourless & sterile.
- Like radioactive, but not radioactive.

- Additive can then be detected in the ground.
- Or additive strength detected at two points, then detect the loss of additive.

Prototype:
Could be dye in water in a leaking tube blotting out on to paper

#4 **Using doppler radar to see the internal flow of water through a main.**

Doppler radar works by sending out radio waves from an antenna. These radio waves are reflected back to the antenna by objects in the air. Through this process Doppler radar can detect precipitation in the air. It even detects frequency differences based on whether an object is moving away from the antenna or towards it. The frequency will be lower if the object is moving away from the antenna and higher if the object is moving toward the antenna. After the antenna detects an object it sends the information back to a computer that brings up the different frequencies as different colors. The colors used represent speed and direction to the user.

- Deploy two small doppler radar down manholes at either end of suspected leakage area. Like a hydrophone.
- View the flow of water to see anomalies in flow.
- look for water flowing to a leak or eddies caused as they pass by holes.

Prototype:

Further Investigation on idea #4 and development of prototype

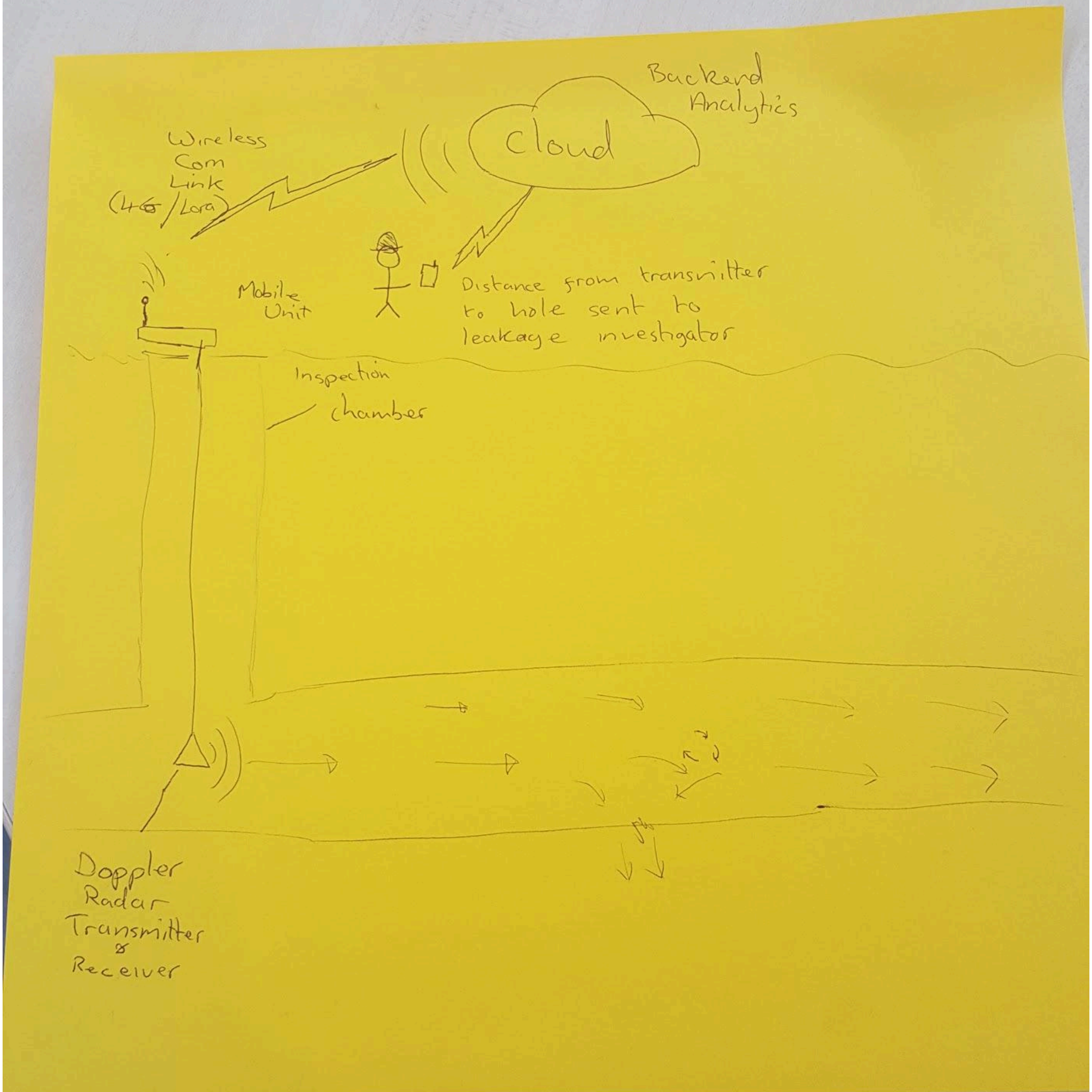
After initial alert of possible leak event from acoustic sensor(s) using possible triangulation from multiple event reports, Mobile Doppler Transmitter & Receiver (MDTR) deployed to the closest inspection chamber.

Only one MDTR would be required (50% saving already!)

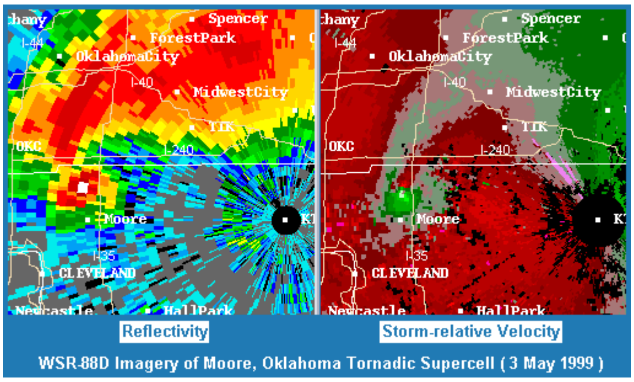
As water flows passed a hole in the pipe the direction and velocity of the flow may change. An eddy may form which could lead to water flowing in an opposite direction to normal flow.

MDTR can detect the direction and velocity of the flow of water. This can be shown on a map using colours.

The data collected would be transmitted via 4G or LPWAN to cloud back end analytics for mapping.


Backend Analytics
Cloud
Wireless Com Link (4G/Lora)
Mobile Unit
Distance from transmitter to hole sent to leakage investigator
Inspection chamber
Doppler Radar Transmitter & Receiver

If an eddy is formed this could be seen on the doppler mapping as a hook signature as often seen in doppler mapping of mesocyclones by meteorologists while attempting to track and predict tornadoes.



WSR-88D Imagery of Moore, Oklahoma Tornadic Supercell (3 May 1999)

The direction of flow can be shown in two colours, with the strength of the colour denoting the increase velocity of the flow.

Backend Analytics

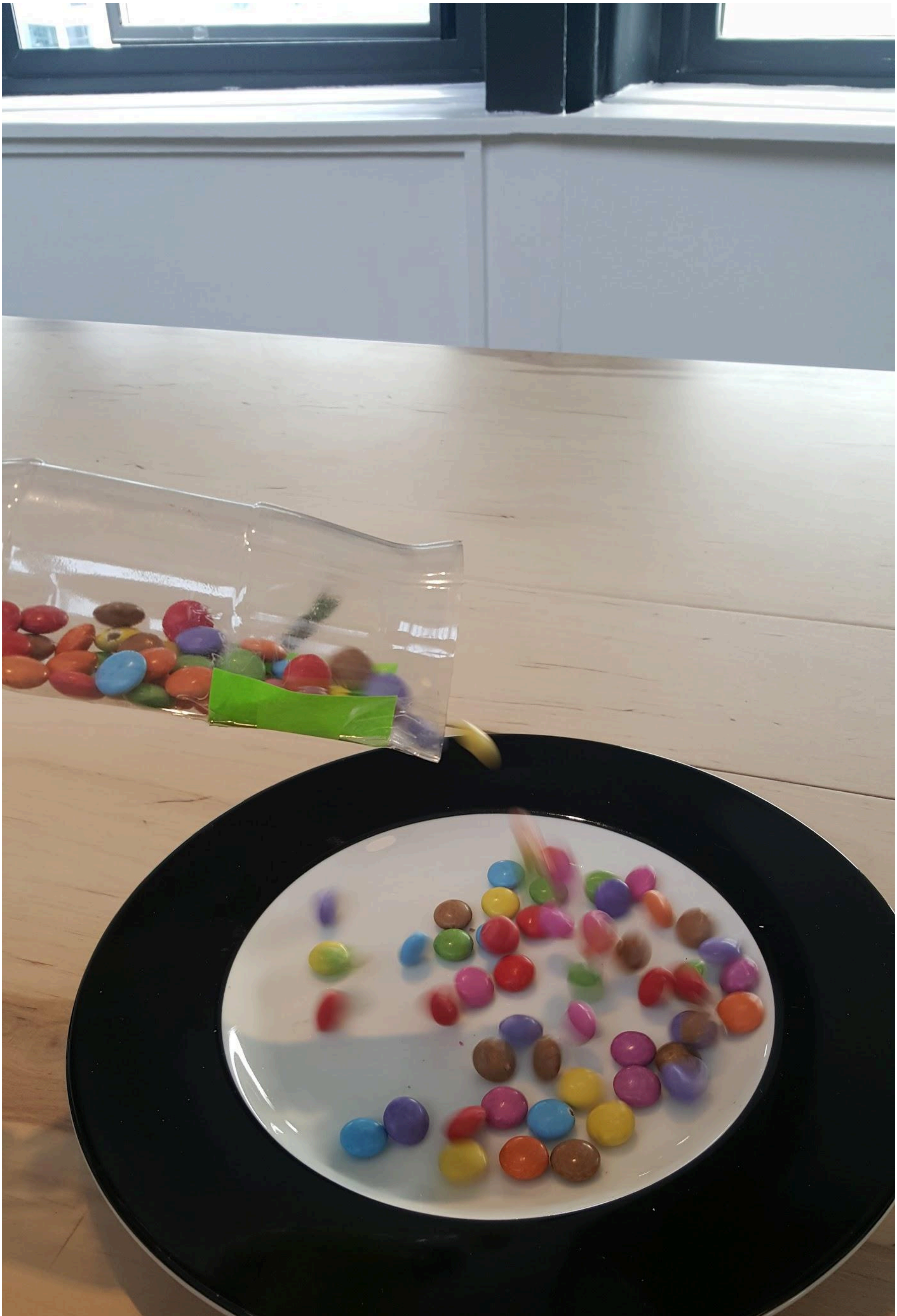
Doppler Velocity Measurement

shows the direction & velocity of water through colours plotted

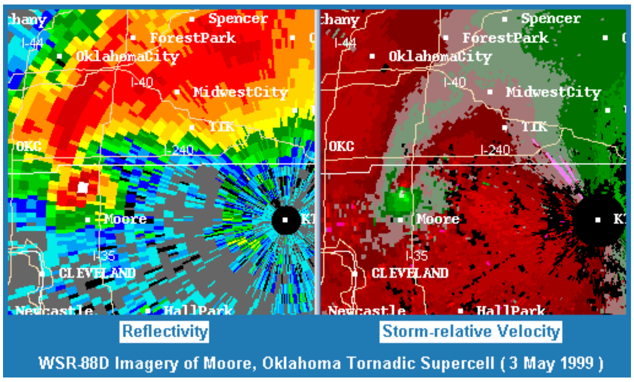
As water passes a hole in the pipe the direction of the water may change or ~~sf~~ speed up or slow down

Darker the Red or Blue, the faster the velocity

Eddy may be formed as the flow changes direction. These could be detected as a hook on the doppler radar image



The mapping could be interpreted by analysts either at head office or remotely accessing the cloud analytics. It could be accessed by on site technicians for analysis.



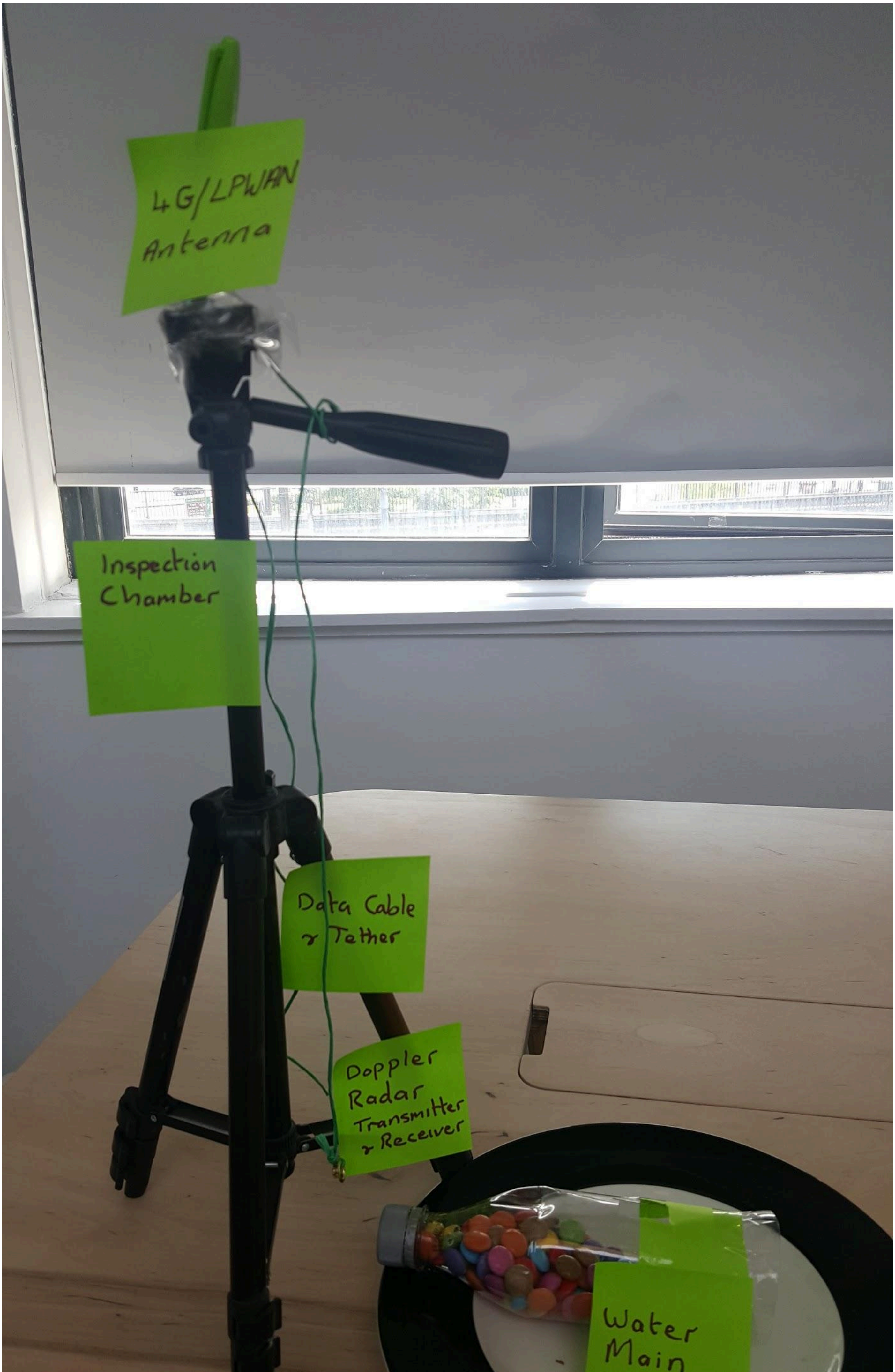
WSR-88D Imagery of Moore, Oklahoma Tornadic Supercell (3 May 1999)

Possibilities for automatic analysis to flag up potential areas of interest. The mapping would show the distance from the MDTR that the anomaly is occurring. This could be cross referenced with geo mapping information.

This could be then sent to on site leakage investigators for them to decide where to dig their investigatory hole and how big a hole may be required.


N
W
E
S

All final outcomes of investigations, and data used (acoustic sensors, triangulation data, prediction data etc.) leading to direct interventions by leakage investigators (holes dug) could then be analysed to improve future outcomes.


4G/LPWAN Antenna
Inspection Chamber
Data Cable & Tether
Doppler Radar Transmitter & Receiver
Water Main

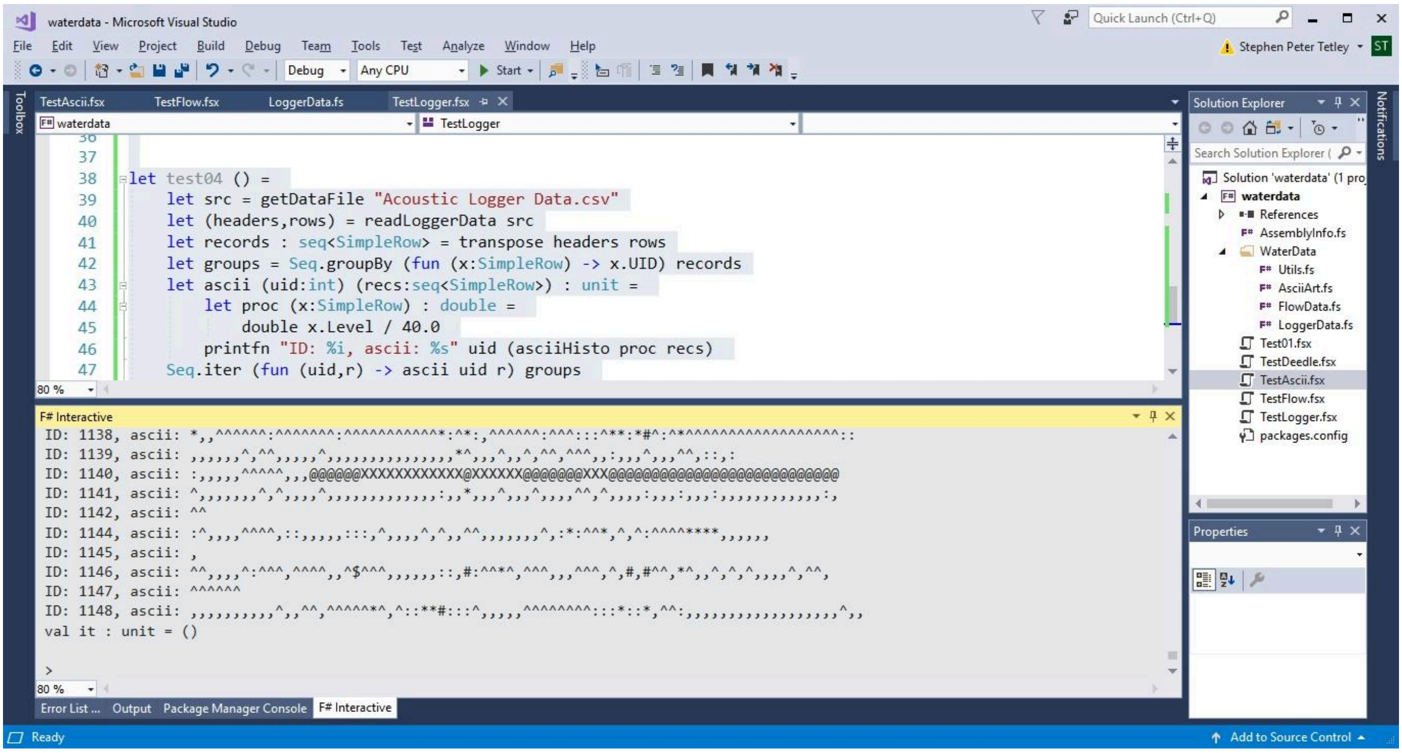
This idea would be complementary to many of the other ideas on show today and could be used in conjunction with many of them as part of a multifaceted approach to working efficiently at reducing leakage rates.

Timelord II

Stephen Tetley - stephen.tetley@gmail.com

Code repository here:
https://github.com/stephentetley/waterdata18

Looking at
Ascii art for simple histograms
Transposing the logger data so it grows vertically rather than horizontally…



Ascii art of Logger data - dark chars are high Levels.

Transposed Logger Data:

UID,Date,Level,Spread
1,02-May,13,4
1,01-May,13,5
1,30-Apr,13,5
1,29-Apr,14,4
1,28-Apr,14,4
1,27-Apr,14,4
1,26-Apr,14,7
1,25-Apr,13,21

2,02-May,11,16
2,01-May,10,5
2,30-Apr,11,6
2,29-Apr,11,9
2,28-Apr,11,8

The transposed data itself:
https://github.com/stephentetley/waterdata18/blob/master/data/SimpleLoggerOut.csv

WORK IN PROGRESS

Problem:
Analogue sensors have recently been introduced into the system to detect leaks.
The presence of leaks is indicated by the noise levels and spread of decibels detected by the loggers.
Current practice assumes that a leak is more than likely to be present if (noise - spread) >= 15.

Question posed: Can we use available data to improve the current threshold for leak detection?

Data available:
Date
Logger ID
Noise
Spread
Outcome of visit - Leak = NO, Leak = YES, Leak = N - PRV.

It is known that the presence of a PRV will automatically increase the noise detected.

Investigation conducted: files were explored to confirm the level of accuracy of alarm indicators.

	Y	N	N - PRV	Total
Y	44	20	5	69
N	8	8	1	17
Total	52	28	6	86

Based on the given data, 44 alarms believed to relate to a leak were a identified as a leak. However this is less than 50% of all alarms, ie: no better than chance. Furthermore there are 20 false-positives (23%) which would indicate a potential ineffective use of resources. We could conclude that he current threshold indicator can be improved.

What could be done to improve the threshold?

Option 1:
Decision tree modelling having three outcomes (No/Yes/NoPRV) and two inputs (Noise and Spread).

Outcome from this would be a probability of leak based on a series of ‘IF’ statements relating to conditions concerning noise and spread scores..

Option 2:
In the future when unique keys are available to link the flow and logger data then they could be engineered together as features to build a better prediction model.

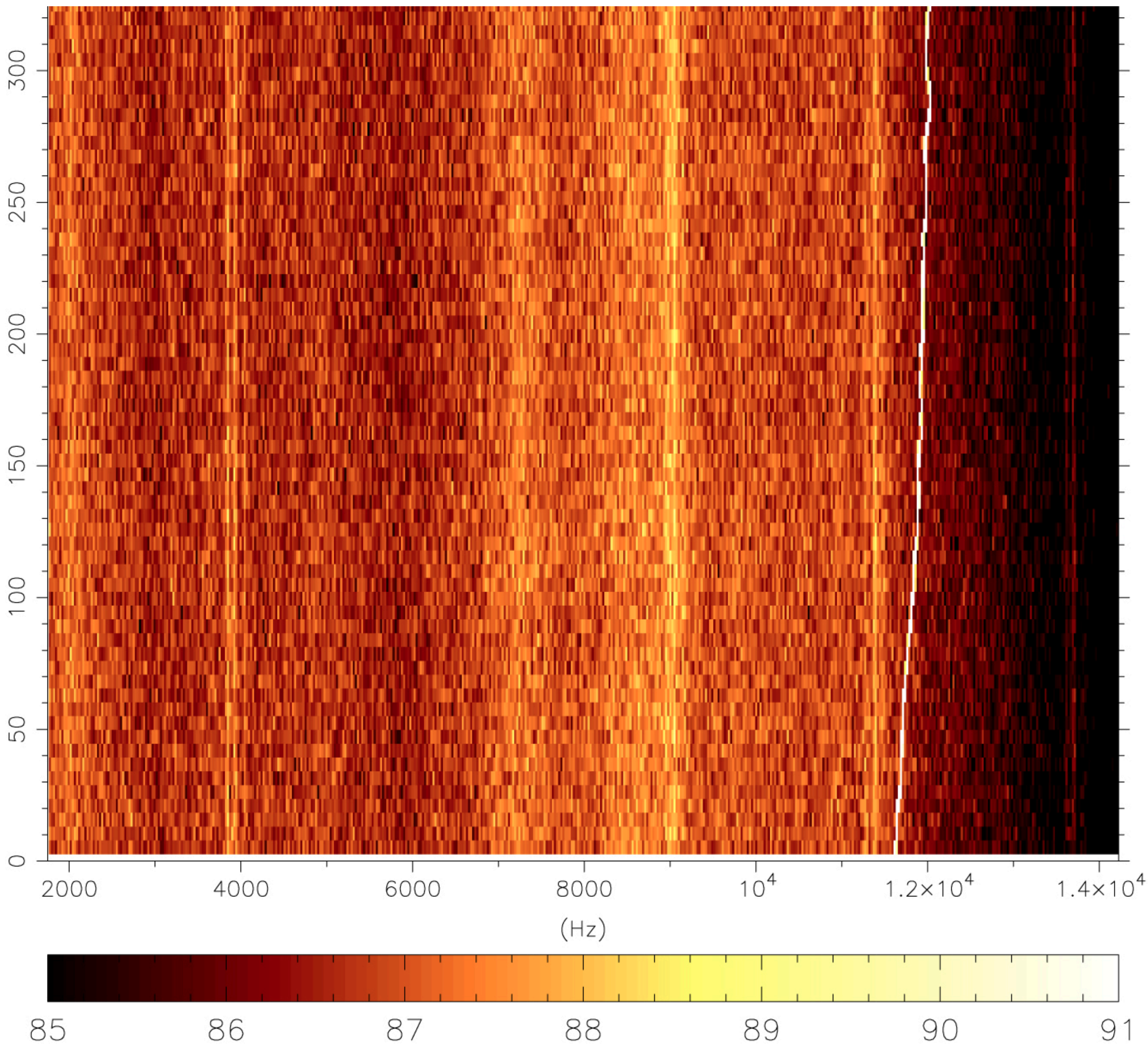
Points for consideration:

Dataset as it currently stands is biased - loggers are currently located where there is a relatively high natural rate of rise. Further bias introduced by the fact that little or no investigation is done if the threshold of 15 is not met. Currently we have no idea of the number of leaks missed (if any) due to the current threshold that is used.

Randomly investigate thresholds < 15 to aid in the reduction of bias within the dataset.

Additional data factors to consider for future:

Historic ‘leakiness’ of an area (ie: if it’s leaked once does this improve the
DMA/pipe related attributes
Marker as to whether or not a PRV is present
Estimation of baseline ‘noise’ and ‘spread’ for each logger - consider different thresholds for different loggers?
Creation of a frequency spectrum of the data from available sound files - to build up visualisations of how the spectrum of noises change and therefore able to equate patterns to potential leaks. Patterns, in time, to be analysed using neural networks



Example “waterfall” plot of frequency data from Jodrell Bank Centre for Astrophysics
http://www.jb.man.ac.uk/public/BeagleII.html

Visualisation

We’ve been visualising the level/spread/site-visit data from https://datamillnorth.org/dataset/yorkshire-water-daily-acoustic-logger-data using Javascript/HTML.

Every sensor is shown as a horizontal bar and stacked vertically (ID=1 at the top). There is a date slider to see how they vary over time. There are three readings of 255dB - equivalent to a nuclear blast - which seem implausible. So, the horizontal scale is determined by the data but ignoring levels over 200dB.

Sensors are coloured yellow if LEVEL-SPREAD >= 15. Otherwise they are grey. A sensor is given a coloured border if a site visit occurred on that date. A green border indicates a success (there was an alarm + a leak was detected or there wasn’t an alarm and there wasn’t a leak). A red border indicates that the alarm state doesn’t match the reality. A purple border indicates “NO-PRV”.

See the interactive at https://slowe.github.io/WaterAcousticLoggerExplorer/

z{