Return To Learn (RTL) Automation Project

Yijian (Cris) Zong, Richard Duong, Nick Lin

We are the RTL data science team and today we will share a story about how we managed to automate most of processes for the on campus RTL team and ITS. And now, Richard would share an overview for the project and background knowledge.
Aaron feedback:
1. Broad context (intro to RTL)
2. current situation (good vs evil)
3. your approach (automation) -- broad
4. "results" (front end description)
5. "methods" -- tech description of serverless arch

Rob Feedback notes:
- Data is currently being underutilized. What if we could use the data to predict where viral infections will take place,
- Presentation should be a lot more enthusiatic
- Full Screen graphics > text
- Have fun with the title / story
    - “How to use robots to fight COVID-19 on campus”
    - Hook -> problem -> journey to solution
- Identify the key points
- Uniqueness of the journey
    - From no monitoring of the campus
    - Using poop to find the spread of covid
    - Giant amount of data being flushed down the toilet
    - Ramen??

    - Emphasize collaboration (“cross-disciplinary”)
    - Emphasize speed
        - Couple of days to couple of hours
        - Continually decreasing the time
    - Talk about the paper, new knowledge, keeping community safe
    - 75% of covid cases were detected through the program
        - UCSD infection rates much lower than surrounding san diego community
            - Think its like 1% for UCSD while 10% for SD???

Show enthusiasm: automate, poop resources to tell cases on campus and keep us safe
NYC:15, UCSD 100+
excited: predict the future

using materials out of paper -> automation
already making an impact

Don’t have the background. Poop data => genome => helps predict =>
Change the titles

1. Broad context (intro to RTL)
2. current situation (good vs evil)
3. your approach (automation) -- broad
4. "results" (front end description)
5. "methods" -- tech description of serverless arch

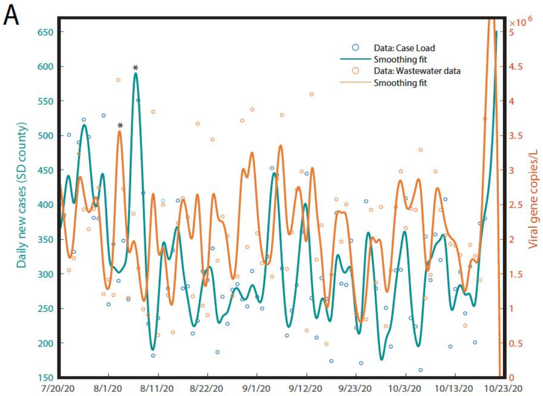
Overview

RTL Wastewater Sampling Project on UCSD

What started out as a small sampling process in which a handful of samples were collected from manholes at specific locations around campus became the leading indicator of forecasting COVID-19 cases. The scope of the monitoring covers over 7000 students in 239 different buildings on campus. Upon detection, students are notified of exposures by means of the wastewater notification program, where specific students in specific buildings were informed of exposure and ultimately tested and isolated quickly and effectively if needed. There were a ton of bottlenecks regarding the sampling process, eventually, the sampling process was assisted heavily by automation, and the turn-around time for the sampling was 4.5 hours from sampling collection at each manhole to automated data reporting and notification. For scale, all of NYC has 15 robot manhole sampling locations while UCSD has over 200 sampling locations.

Trends inferred from SARS-CoV-2 signal lead clinically confirmed cases

A



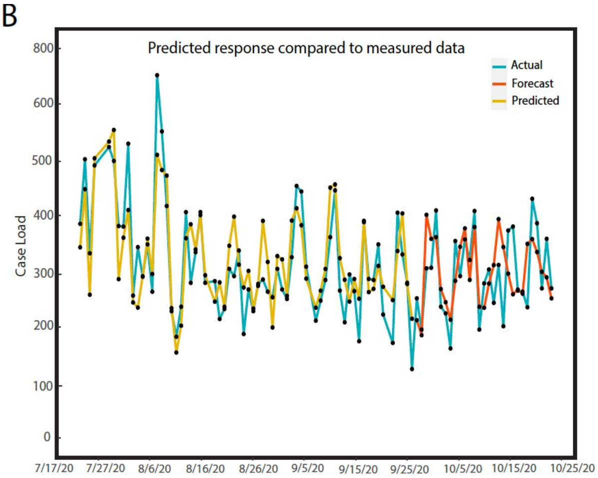
Daily caseload and wastewater viral concentration data shown for a period of 13 weeks, where a spline smoothing is applied to each time series to demonstrate general trends

Although informative, this time-lagged correlation alone is not enough for robust predictions. Instead, this served as the main motivation to build a predictive model for forecasting the number of new cases per day in San Diego County.

red peak in front of blue peak, few days before clinical cases, tell us where it gonna get worse

Wastewater SARS-CoV-2 detection enables forecasting of community infection dynamics in San Diego county

B



Data collected from 07/07/2020 to 09/28/2020 were used as the training dataset to predict the caseload for the following weeks (up to 10/25/2020).

broadcast dynamics, autoregressive moving average
few days lead time -> clinic
huge scale, single building?-> campus

Instead, Data-driven approach to train a prediction model that utilizes wastewater data and temporal correlations (embedded in the day of the week) in order to forecast the number of new positive cases in San Diego County.

The (predicted) number of new cases consist of lagged past values from all three series (number of new cases, wastewater data, day of the week) and each term can be thought of as the influence of that lagged time series on the number of new cases.

Correlation

Smruthi used auto-correlation and found there is ~0.75 correlation between wastewater data and official cases.

The issue

Question: If an infected individual has COVID-19 there is a period of time when they are asymptomatic, but still shed the virus. Is there a way to find the delay between the start of the viral shedding and when they report their illness to the county?

Solution: Find when the viral loads in the sewage and reports are most in sync!

wastewater signal correlation maximized

max five words summary of the issue

How might we?....

lunch + computers + emails => user friendly 30% of attention

get most of the info from the title

Pearson vs Spearman Correlation

Pearson:

Demonstrates the **linear** relationship between two continuous variables.

Spearman:

Demonstrates the **monotonic** relationship between two continuous variables.

Question: Which correlation should be most suitable to our scenario?

parametric or non parametric
spearsman: rank orders

illustrations => use graph instead of words

three images, two on the top

it is all about balance

comparisons

Solution

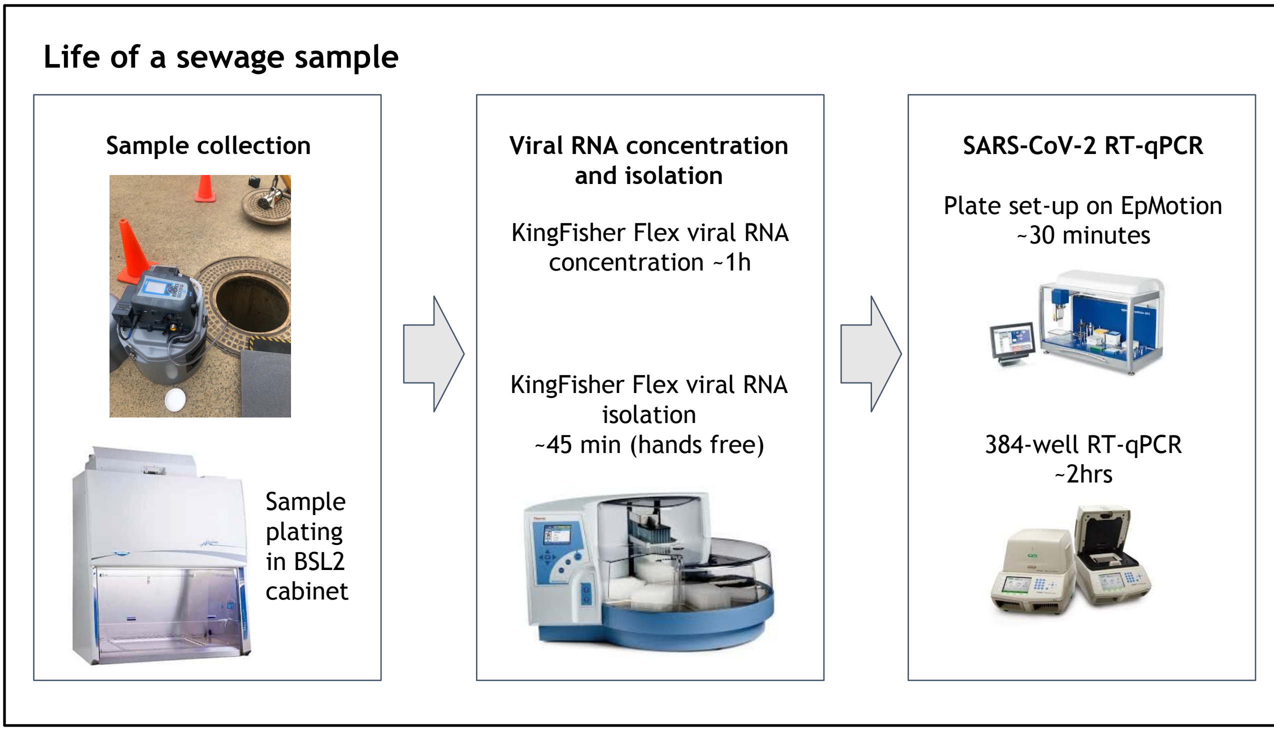
One of the biggest constraints with correlations is that they are not optimizable. The closest method is through brute force methodology. There is consensus that the a person is infective for about two weeks, therefore we decided on a brute force comparison of correlations for a two week period.

Traits we want in our loss function:

- "Balanced" between the two correlations
- Worse values have higher value
- Don't want to deal with negatives

Loss: 1 / (pearsonr(x1, x2)[0] * spearmanr(x1, x2)[0])**2.

we cannot optimize automatically. All possible offsets -> brute force
non negative



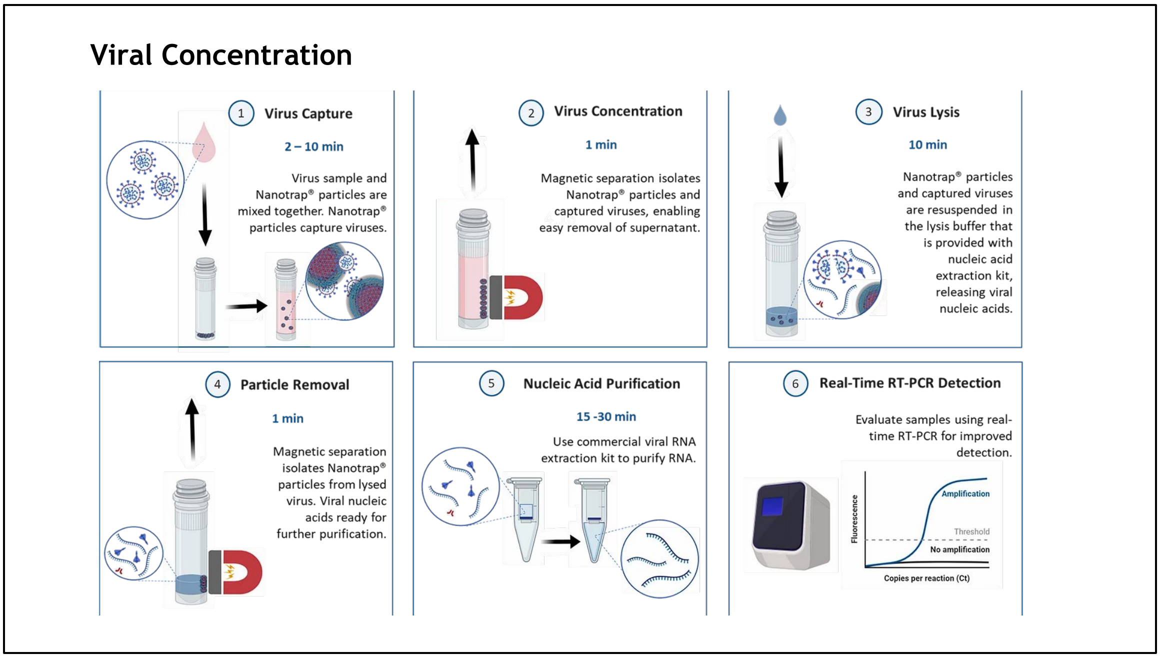
*still working out kinks / finalizing protocols
*of course there could be more sensitive alternatives…. But this is all using established protocls and best we have for now that can be quickly scaled!

Robots army….

QPCR

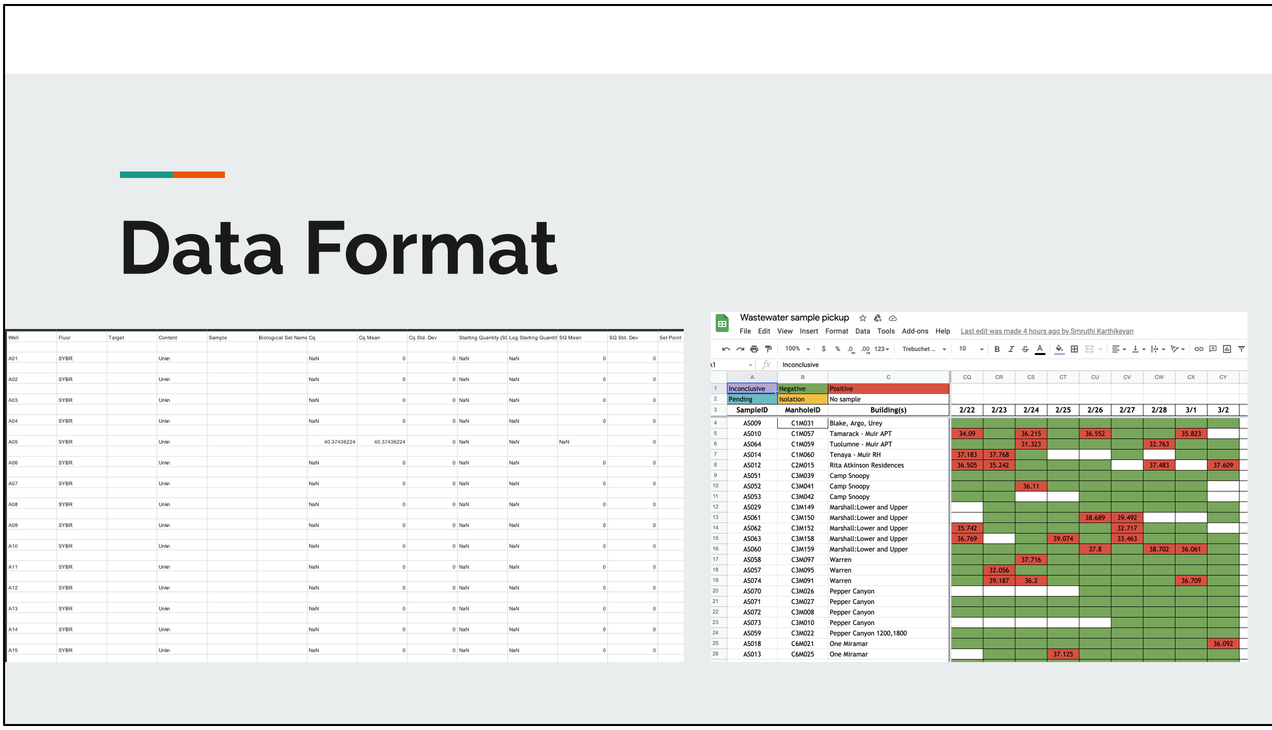
https://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction

This is how the data is retrieved



Steps of doing so. Talk about the sewer walks
Does the QPCR occur on site or at a lab?

quotes Smruthi



painpoints: possible manual input errors, hours of manual input time, cross reference with google sheet

Free the researchers from these laborious work!

arrows for the current process(diagram)

happy scientist => sad scientist

paper vs map --->

Problem Statement

How do we help the RTL team get their jobs done faster?

since the workflows of the RTL team are mult faceted, the solution should be portable, flexible, and scalable service that automate each part of the workflow independently in order to automate the whole process. So what could be a suitable solution?

get rid of problem statement and make it bold

Solution

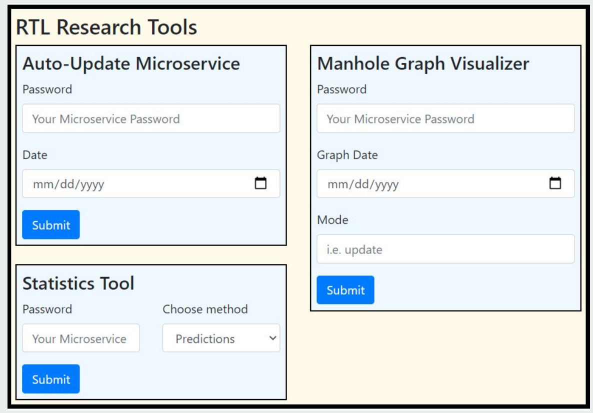
MAI

Microservice-based Auto Infrastructure (A serverless system)

Serverlerss system, Microservevices, rather than monolithic system that does all the business. Each component is broken down into individual microservices, consuming the product of each dependent microservice. This nicely matains the atomicity of the service and makes it easy to adjust to bebug and faster to roll out.

Connect the dots:

User Interface

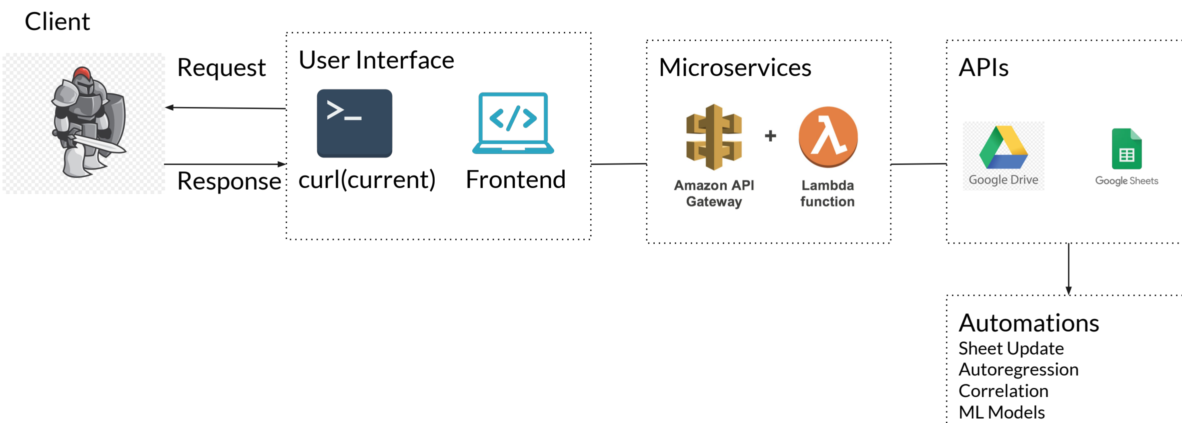


So to easily implement our serverless system and allow research team members to easily use the system without having to know the jargon, we implemented, tell the results the first. methods in details in the end

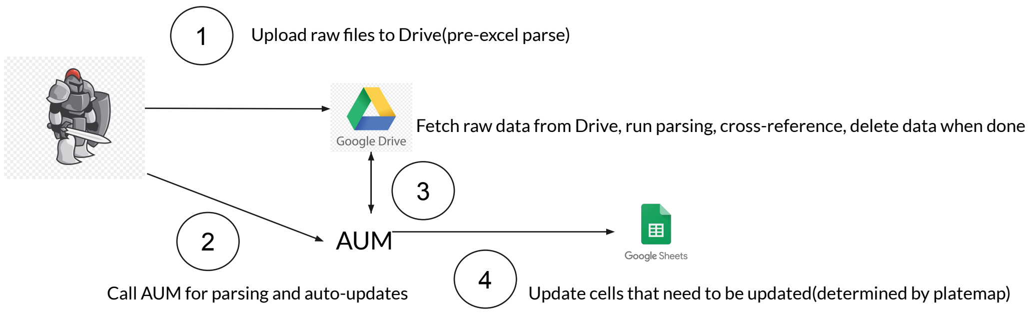
DEMO

Work in Progress

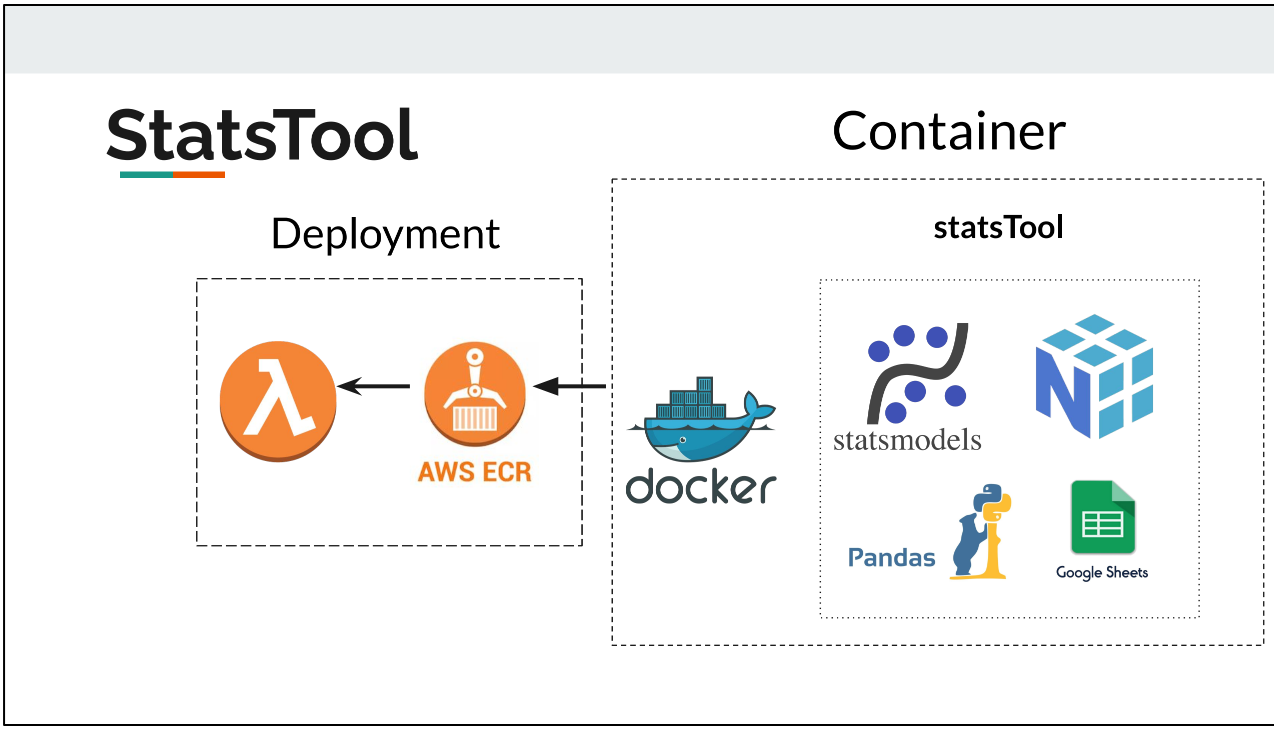
Basic Structure(REST-based)



AUM: Auto Update Microservice



This is the AUM -> before, people, were running script file against a raw files with cq values, then cross reference, then entering values in the script
already been used by lots of people in the sample collection process, saving them a considerable amount of time.



short and sweet
services built
general overview

if can automate, we can also , get tools for ds to use

Looking Ahead

- unit tests
- automation of remaining data integration process
- cases prediction on dashboard
- integration of the virus phylogenetic tree

Acknowledgements

Major thanks to the people who assisted us

- Rob Knight for his mentorship
- Smruthi Karthikeyan for her guidance
- Daniel McDonald for his technical assistance
- Andrew Nguyen for his data assistance
- Natasha Martin for her expertise on the subject matter
- Michiko Souza for organizing the meetings

References