

# Data Science Accelerator Level 2: Power-up or reboot?

[richard.dean@esr.cri.nz](mailto:richard.dean@esr.cri.nz)

11-Feb-2021

**E/S/R**  
Science for Communities





# Data Science Accelerator Level 2: Power-up or reboot?

## Today's objectives:

- 1) What happened since eResearch2020
- 2) Review progress
- 3) Look forward to 2021



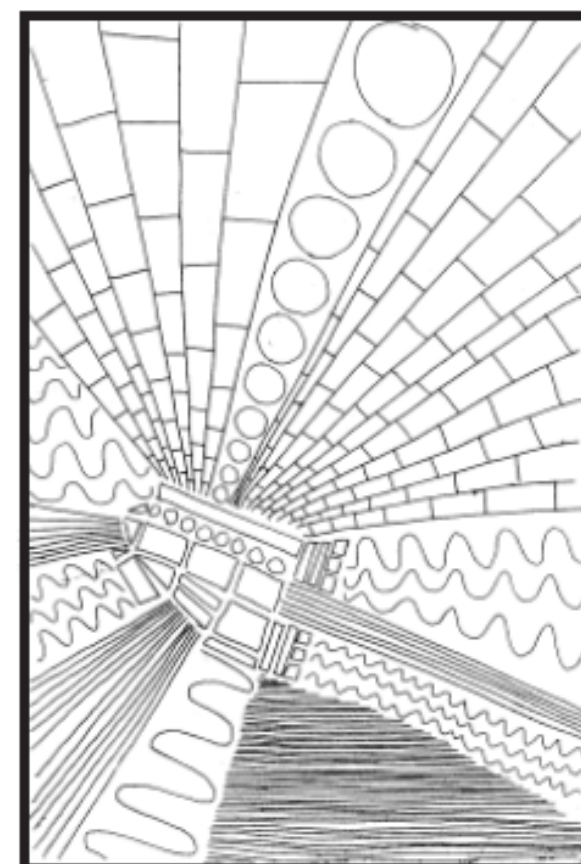
# SIM

# CITY 2000®

THE ULTIMATE CITY SIMULATOR



USER'S  
MANUAL



**E/S/R**  
Science for Communities



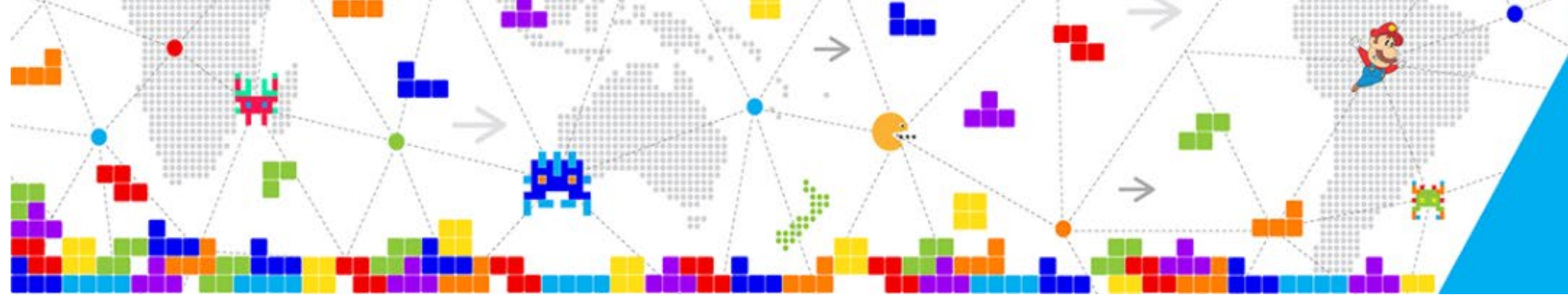
*This series of vignettes about cities and city planning was provided by Richard Bartlett, AIA, Architect. Spread throughout the manual, they give a historical and humanistic perspective to planning that you may wish to incorporate into your city designs.*

Cities are for people: a place for their hopes and dreams, their work and play, their homes and homes for their children. Cities are alive and have personalities, each different from all others and each in constant change. A living organism made up of its collective inhabitants, a city is many things, but it is above all a storehouse of human characteristics.



I





## Level Up

- Increasing speed, agility, scale and collaboration
- Future proofing
- Value and importance of skills training, professional development and mentoring





Dealing With  
Disaster





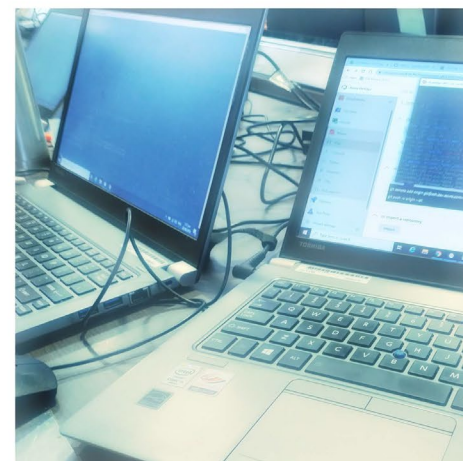
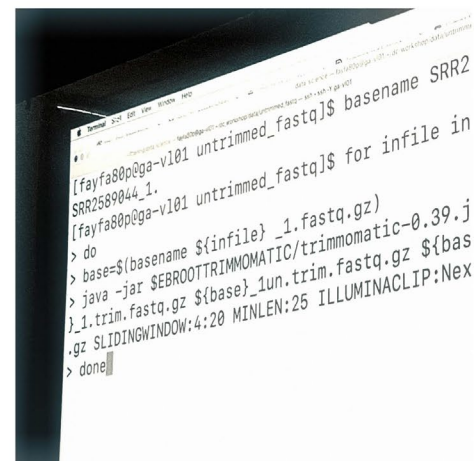
Dealing With  
Disaster



**E/S/R**  
Science for Communities



# Levelling up at ESR





# What's a **data science accelerator**?

A data science project that tackles a business problem

Access to required data

Participant able to commit 1 day per week [3 months | 15 weeks] to the project

Support from line manager and senior manager

Coding experience is useful but not essential.



# After eResearch 2020...

	Cohort 1	Cohort 2
Number of projects	4	6
Participants	6	12
Mentors	4 (of which 1 was also a participant)	7 (of which 2 acted as participant + mentor)

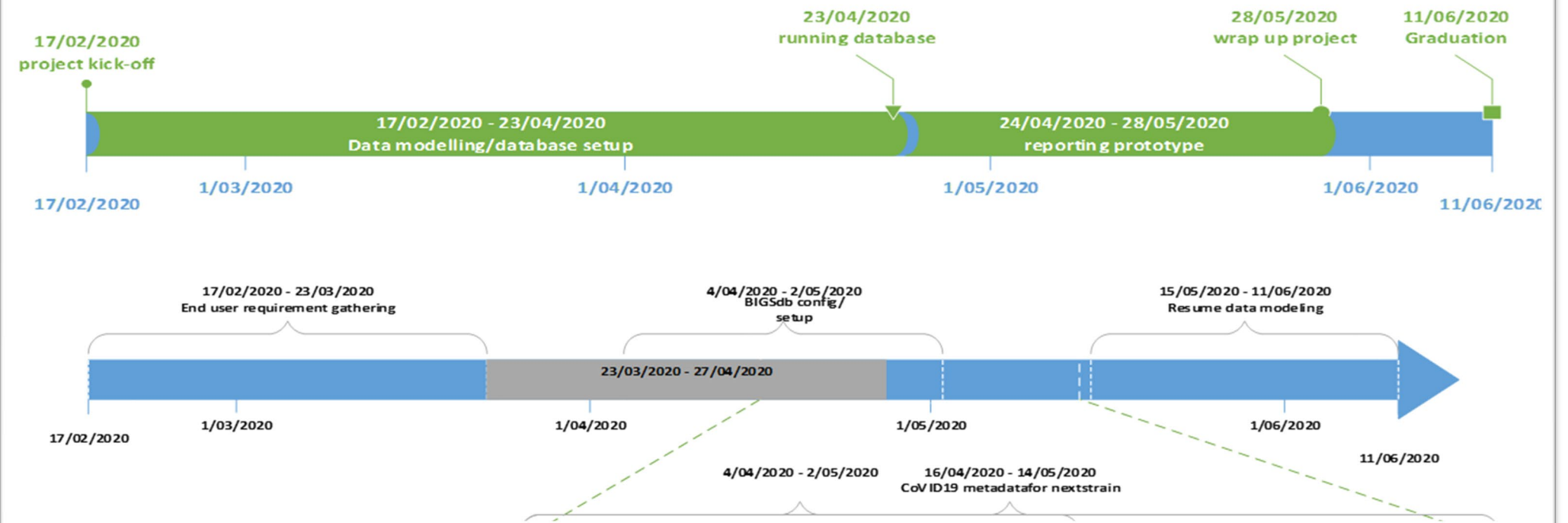
Level 2



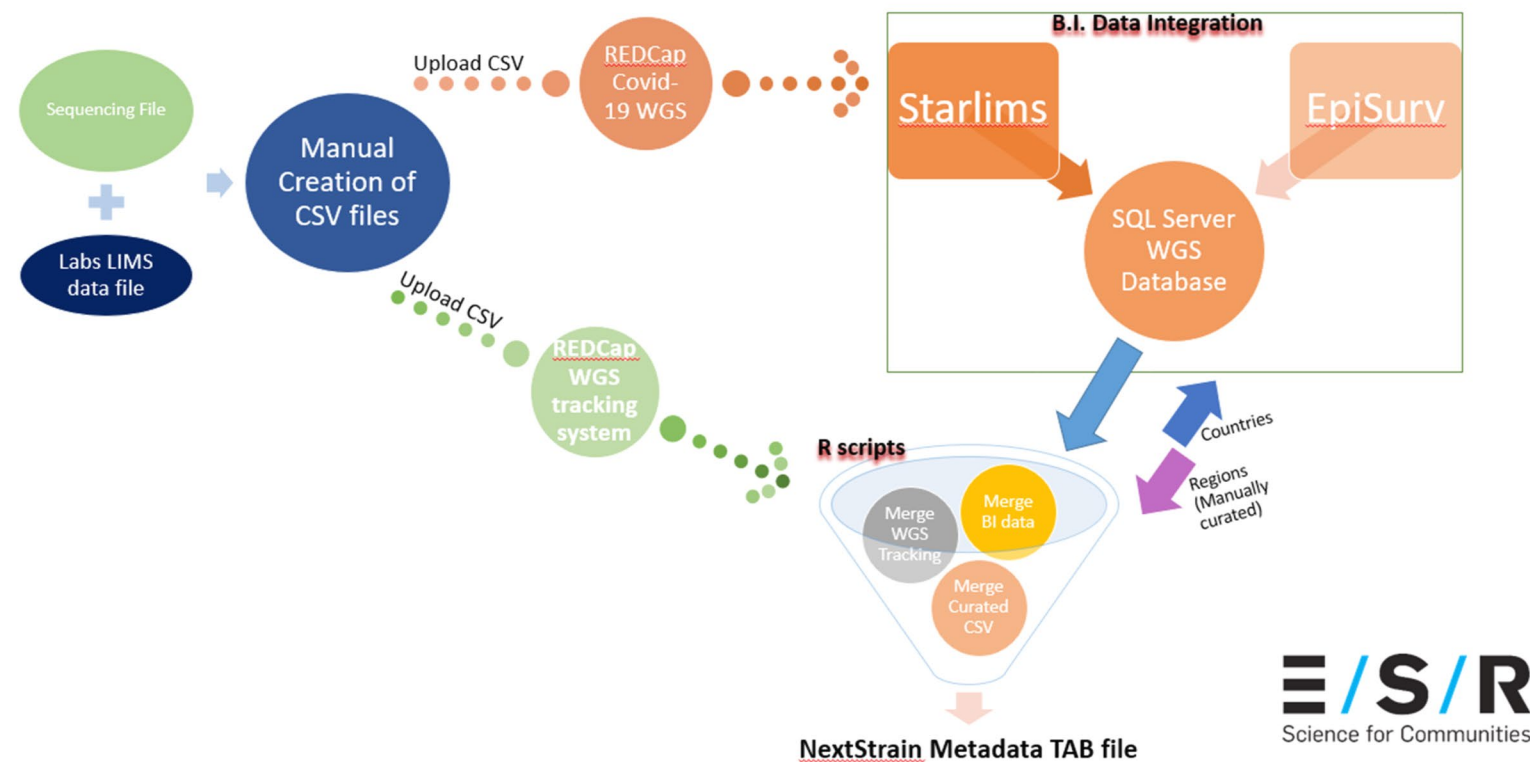




## Planning Vs Reality



## Nextstrain COVID-19 Data Integration and automation

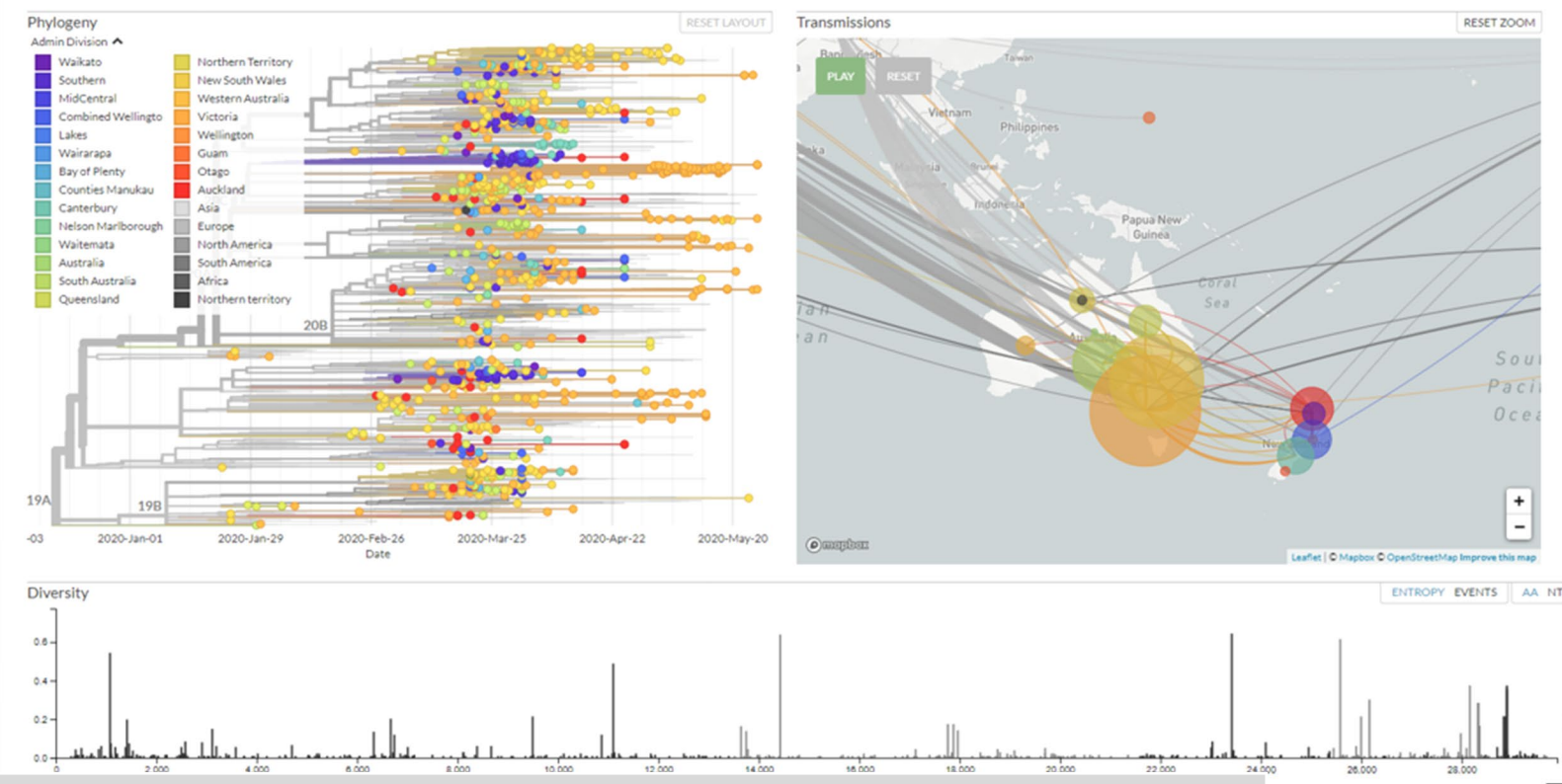


## Spin-off Project During Lockdown (Nextstrain COVID-19 Metadata)

Genomic epidemiology of novel coronavirus - Oceania-focused subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 901 of 2354 genomes sampled between Jan 2020 and May 2020. Filtered to Oceania [901].





# Powering on through lockdown

Science / ToxQC / Overview / Wiki

Wikis > ToxQC Wiki

Wiki

Tom Sheehan 28/11/2020 Revisions

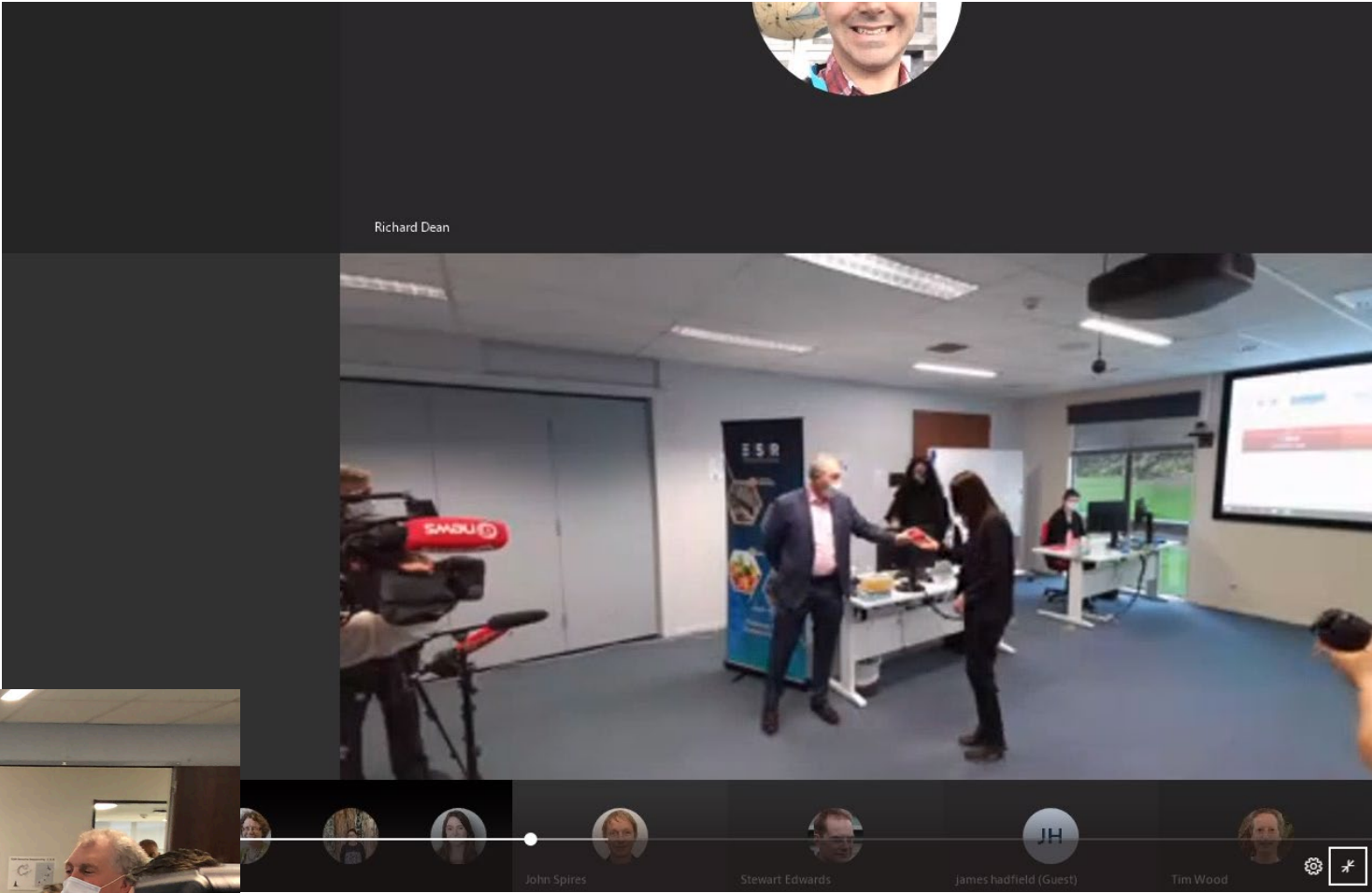
17	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Buprenorphine 3	45101.74070860796	2446.41238995854	3.993400241654996	0.004	0.00411215014094549	102.803753529637	N/A
18	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clobazam 1	50263.5152588191	28168.1147445305	4.7398145684394	0.01	0.00974735177874221	97.4735177874221	N/A
19	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clobazam 2	17158.7310182324	8954.69184431742	4.73981436052863	0.01	0.0104566789272647	104.566789272647	N/A
20	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clobazam 3	2966.84568236214	1888.95730471779	4.74031484747869	0.01	0.011419829378508	114.19829378508	N/A
21	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clonazepam 1	37790.7645964161	21365.2638129537	4.52278851425893	0.01	0.0105408629344515	105.408629344515	N/A
22	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clonazepam 2	14762.1522401186	9157.93380039248	4.51963074287158	0.01	0.010768942715567	107.68942715567	N/A
23	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Clonazepam 3	9977.20847747781	4317.64633435027	4.52045632116226	0.01	0.0102341996238519	102.341996238519	N/A
24	6	CRDOA 20200103	S2	Standard	CRDOAu1F.dam	Cocaine 1	123458.982708556	66837.0477213006	3.50792045574357	0.006	0.00602617367350163	100.436227891694	N/A

Showing 1 to 27 of 1,000 entries, 13 total columns

Richard pointed out an issue with having our server conection password as plain test in our code. Not very secure! We found a solution using "getpass" package.

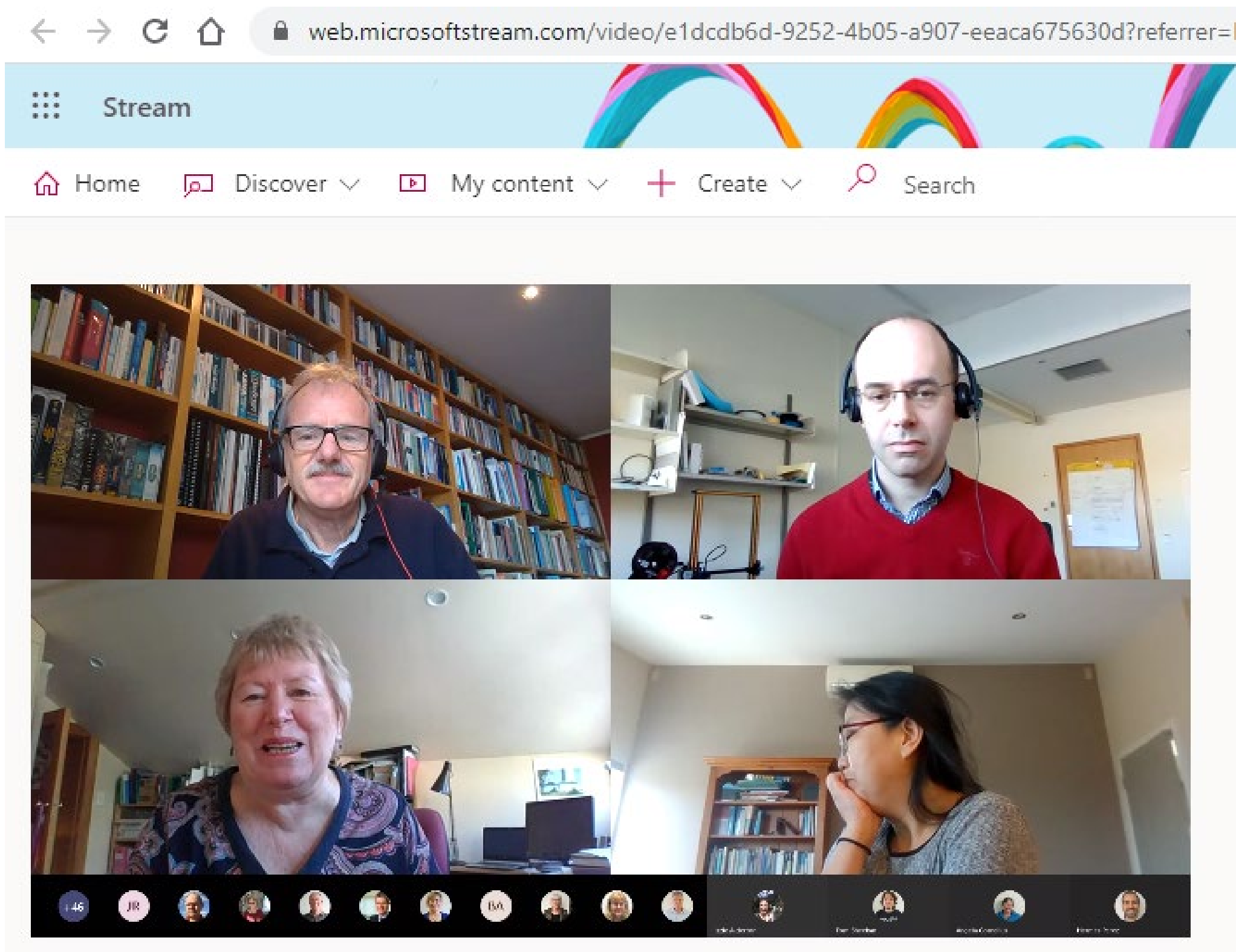
Day 4/5

Lockdown Sucks:





# Cohort 2 graduation





# Automating Foodborne Disease Data Consolidation

Data Accelerator Project by  
Beverley Horn, Isabelle Pattis, Bridget Armstrong

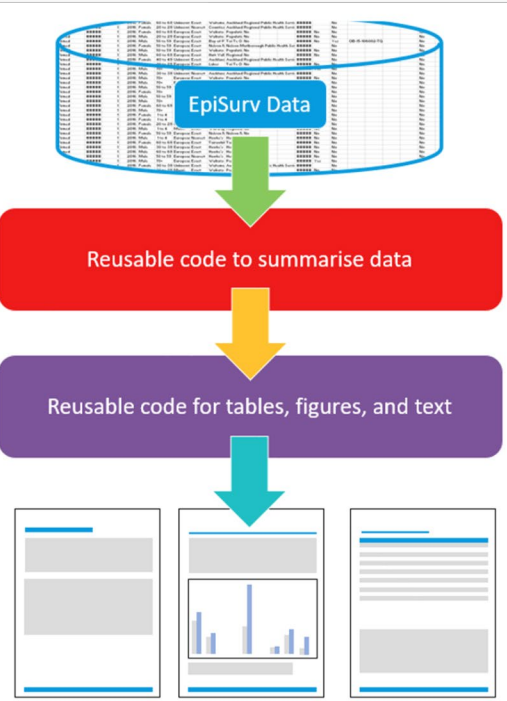
Creating a quality report for the Health of New Zealand

## The Plan

EpiSurv database  
Automate data summary  
Automate report elements

- Tables
- Figures
- Text

Create Foodborne Disease Report



## New Skills

- TFS
- Algorithms
- R, functions, package dev
- Formal function testing
- Data checks
- Teams collaboration
- Familiarity with common databases
- Flextables

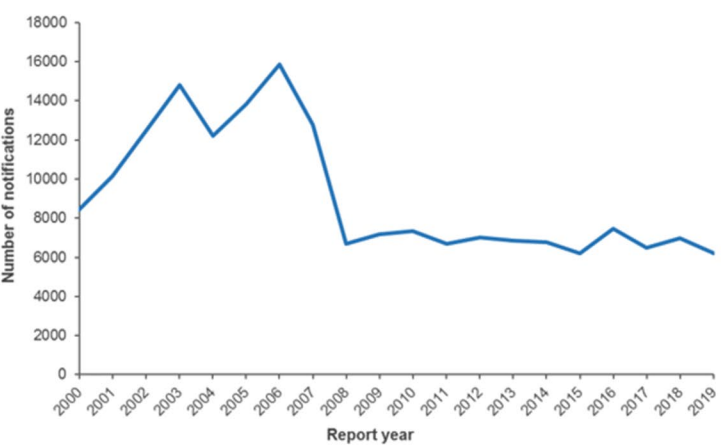
## The Products

Table 11. Details of *Campylobacter* spp. outbreaks, 2019

PHU	Month	Suspected Vehicle	Exposure Setting	Preparation Setting	No. ill
MidCentral	Mar	Raw milk	Camp	Farm	1C 3P
South	Apr	Raw milk	Home	Home	2C
Toi Te Ora	Jul	Re-heated rice	Long term care facility	Long term care facility	2C
Toi Te Ora	Aug	Raw Milk	Other food outlet	Farm	3C
Regional	Aug	Unknown	Restaurant/cafe/bakery	Restaurant/cafe/bakery	3C
Auckland	Sep	Unknown	-	-	1C 1P
Waikato	Dec	Hot and cold chicken meals, food also being hoarded in rooms	Prison	Prison	4C 58P
Auckland	Dec	Unknown	Restaurant/cafe/bakery	-	3C

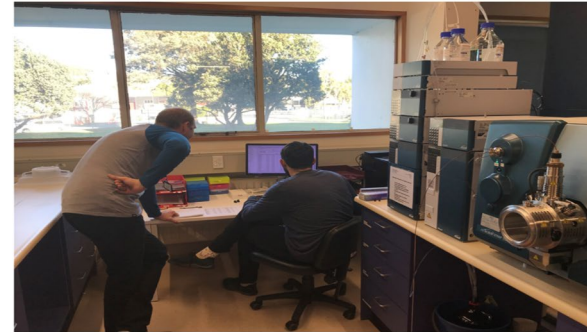
PHU = Public health unit, MidCentral: MidCentral Public Health Service, South: Public Health South, Toi Te Ora: Toi Te Ora - Public Health, Regional: Regional Public Health, Auckland: Auckland Regional Public Health Service, Waikato: Population Health Service  
Waikato, C: confirmed, P: probable

Figure 4. *Campylobacteriosis* notifications by year, 2000 - 2019



Outbreaks reported as caused by *Campylobacter* spp.  
In 2019, eight (40.0%) of the outbreaks caused by *Campylobacter* spp. and 81 (51.9%) of the associated cases were reported as foodborne (Table 10). There were three hospitalisations due to a foodborne *Campylobacter* spp. associated outbreak. An outbreak is classed as foodborne in this report if food was recorded as one of the likely modes of transmission applicable to the outbreak. It is important to note that a single outbreak may have multiple pathogens, modes of transmission, settings where exposure occurred, or settings where preparation of food was conducted. *Campylobacter* outbreaks accounted for 4.0% (20/499) of all enteric outbreaks and 2.0% (156/7824) of all associated cases reported in 2019.





Component Name	SE	MF	MD	MR	MA	MF	MD	MR	MA
	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005	Q1M 2005
ComponentGroup 1	95	105	95	104	102	104	92	102	94
Component 1	102	105	95	104	102	104	92	102	94
Component 2	93	105	95	107	103	100	96	100	94
Component 3	94	105	95	107	103	100	96	100	94
Component 4	93	113	113	109	109	109	96	106	111
Component 5	93	113	113	109	109	109	96	106	111
Component 6	93	113	113	109	109	109	96	106	111
Component 7	93	113	113	109	109	109	96	106	111
Component 8	93	113	113	109	109	109	96	106	111
Component 9	93	113	113	109	109	109	96	106	111
Component 10	93	113	113	109	109	109	96	106	111
Component 11	93	113	113	109	109	109	96	106	111
Component 12	93	113	113	109	109	109	96	106	111
Component 13	93	113	113	109	109	109	96	106	111
Component 14	93	113	113	109	109	109	96	106	111
Component 15	93	113	113	109	109	109	96	106	111
Component 16	93	113	113	109	109	109	96	106	111
Component 17	93	113	113	109	109	109	96	106	111
Component 18	93	113	113	109	109	109	96	106	111
Component 19	93	113	113	109	109	109	96	106	111
Component 20	93	113	113	109	109	109	96	106	111
Component 21	93	113	113	109	109	109	96	106	111
Component 22	93	113	113	109	109	109	96	106	111
Component 23	93	113	113	109	109	109	96	106	111
Component 24	93	113	113	109	109	109	96	106	111
Component 25	93	113	113	109	109	109	96	106	111
Component 26	93	113	113	109	109	109	96	106	111
Component 27	93	113	113	109	109	109	96	106	111
Component 28	93	113	113	109	109	109	96	106	111
Component 29	93	113	113	109	109	109	96	106	111
Component 30	93	113	113	109	109	109	96	106	111
Component 31	93	113	113	109	109	109	96	106	111
Component 32	93	113	113	109	109	109	96	106	111
Component 33	93	113	113	109	109	109	96	106	111
Component 34	93	113	113	109	109	109	96	106	111
Component 35	93	113	113	109	109	109	96	106	111
Component 36	93	113	113	109	109	109	96	106	111
Component 37	93	113	113	109	109	109	96	106	111
Component 38	93	113	113	109	109	109	96	106	111
Component 39	93	113	113	109	109	109	96	106	111
Component 40	93	113	113	109	109	109	96	106	111
Component 41	93	113	113	109	109	109	96	106	111
Component 42	93	113	113	109	109	109	96	106	111
Component 43	93	113	113	109	109	109	96	106	111
Component 44	93	113	113	109	109	109	96	106	111
Component 45	93	113	113	109	109	109	96	106	111
Component 46	93	113	113	109	109	109	96	106	111
Component 47	93	113	113	109	109	109	96	106	111
Component 48	93	113	113	109	109	109	96	106	111
Component 49	93	113	113	109	109	109	96	106	111
Component 50	93	113	113	109	109	109	96	106	111
Component 51	93	113	113	109	109	109	96	106	111
Component 52	93	113	113	109	109	109	96	106	111
Component 53	93	113	113	109	109	109	96	106	111
Component 54	93	113	113	109	109	109	96	106	111
Component 55	93	113	113	109	109	109	96	106	111
Component 56	93	113	113	109	109	109	96	106	111
Component 57	93	113	113	109	109	109	96	106	111
Component 58	93	113	113	109	109	109	96	106	111
Component 59	93	113	113	109	109	109	96	106	111
Component 60	93	113	113	109	109	109	96	106	111
Component 61	93	113	113	109	109	109	96	106	111
Component 62	93	113	113	109	109	109	96	106	111
Component 63	93	113	113	109	109	109	96	106	111
Component 64	93	113	113	109	109	109	96	106	111
Component 65	93	113	113	109	109	109	96	106	111
Component 66	93	113	113	109	109	109	96	106	111
Component 67	93	113	113	109	109	109	96	106	111
Component 68	93	113	113	109	109	109	96	106	111
Component 69	93	113	113	109	109	109	96	106	111
Component 70	93	113	113	109	109	109	96	106	111
Component 71	93	113	113	109	109	109	96	106	111
Component 72	93	113	113	109	109	109	96	106	111
Component 73	93	113	113	109	109	109	96	106	111
Component 74	93	113	113	109	109	109	96	106	111
Component 75	93	113	113	109	109	109	96	106	111
Component 76	93	113	113	109	109	109	96	106	111
Component 77	93	113	113	109	109	109	96	106	111
Component 78	93	113	113	109	109	109	96	106	111
Component 79	93	113	113	109	109	109	96	106	111
Component 80	93	113	113	109	109	109	96	106	111
Component 81	93	113	113	109	109	109	96	106	111
Component 82	93	113	113	109	109	109	96	106	111
Component 83	93	113	113	109	109	109	96	106	111
Component 84	93	113	113	109	109	109	96	106	111
Component 85	93	113	113	109	109	109	96	106	111
Component 86	93	113	113	109	109	109	96	106	111
Component 87	93	113	113	109	109	109	96	106	111
Component 88	93	113	113	109	109	109	96	106	111
Component 89	93	113	113	109	109	109	96	106	111
Component 90	93	113	113	109	109	109	96	106	111
Component 91	93	113	113	109	109	109	96	106	111
Component 92	93	113	113	109	109	109	96	106	111
Component 93	93	113	113	109	109	109	96	106	111
Component 94	93	113	113	109	109	109	96	106	111
Component 95	93	113	113	109	109	109	96	106	111
Component 96	93	113	113	109	109	109	96	106	111
Component 97	93	113	113	109	109	109	96	106	111
Component 98	93	113	113	109	109	109	96	106	111
Component 99	93	113	113	109	109	109	96	106	111
Component 100	93	113	113	109	109	109	96	106	111
Component 101	93	113	113	109	109	109	96	106	111
Component 102	93	113	113	109	109	109	96	106	111
Component 103	93	113	113	109	109	109	96	106	111
Component 104	93	113	113	109	109	109	96	106	111
Component 105	93	113	113	109	109	109	96	106	111
Component 106	93	113	113	109	109	109	96	106	111
Component 107	93	113	113	109	109	109	96	106	111
Component 108	93	113	113	109	109	109	96	106	111
Component 109	93	113	113	109	109	109	96	106	111
Component 110	93	113	113	109	109	109	96	106	111
Component 111	93	113	113	109	109	109	96	106	111
Component 112	93	113	113	109	109	109	96	106	111
Component 113	93	113	113	109	109	109	96	106	111
Component 114	93	113	113	109	109	109	96	106	111
Component 115	93	113	113	109	109	109	96	106	111
Component 116	93	113	113	109	109	109	96	106	111
Component 117	93	113	113	109	109	109	96	106	111
Component 118	93	113	113	109	109	109	96	106	111
Component 119	93	113	113	109	109	109	96	106	111
Component 120	93	113	113	109	109	109	96	106	111
Component 121	93	113	113	109	109	109	96	106	111
Component 122	93	113	113	109	109	109	96	106	111
Component 123	93	113	113	109	109	109	96	106	111
Component 124	93	113	113	109	109	109	96	106	111
Component 125	93	113	113	109	109	109	96	106	111
Component 126	93	113	113	109	109	109	96	106	111
Component 127	93	113	113	109	109	109	96	106	111
Component 128	93	113	113	109	109	109	96	106	111
Component 129	93	113	113	109	109	109	96	106	111
Component 130	93	113	113	109	109	109	96	106	111
Component 131	93	113	113	109	109	109	96	106	111
Component 132	93	113	113	109	109	109	96	106	111
Component 133	93	113	113	109	109	109	96	106	111
Component 134	93	113	113	109	109	109	96	106	111
Component 135	93	113	113	109	109	109	96	106	111
Component 136	93	113	113	109	109	109	96	106	111
Component 137	93	113	113	109	109	109	96	106	111
Component 138	93	113	113	109	109	109	96	106	111
Component 139	93	113	113	109	109	109	96	106	111
Component 140	93	113	113	109	109	109	96	106	111
Component 141	93	113	113	109	109	109	96	106	111
Component 142	93	113	113	109	109	109	96	106	111
Component 143</									

- R programming
- Collaboration
- Handling Data

## Project Goals

## Saving Everyone's Time

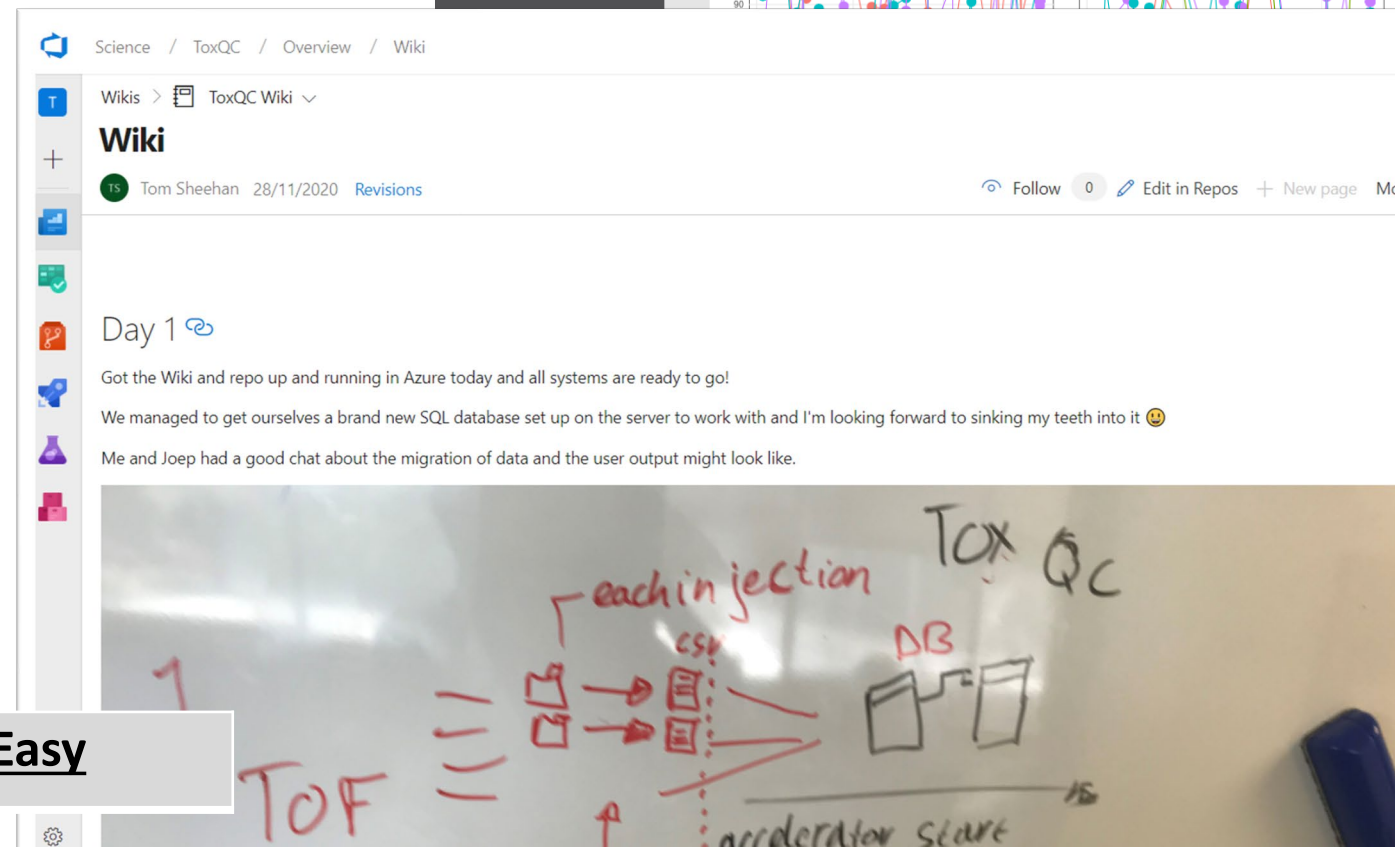
- Copying to spreadsheets alone cost around 30 min a day.
- Finding data in too many different spreadsheets takes time.
- Review multiple spreadsheets take time.

## Reducing Rates of Error

- Copy and paste errors
- Incorrect naming
- Transcription error

## Making the Job Easier

- Easy data upload
- One-stop shop for many tasks
- User Friendly environment





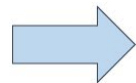
My data science journey

*Mentors:*

A decorative graphic consisting of several overlapping sine waves in various colors (blue, green, yellow, orange, pink, purple) against a white background. The waves are of different phases and amplitudes, creating a complex, layered pattern.

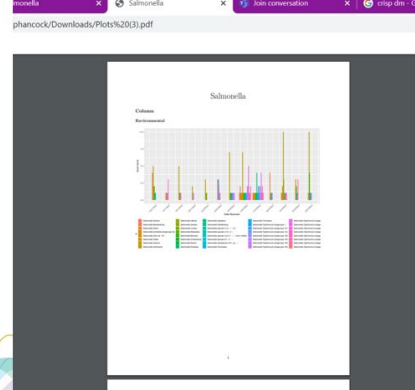
[illegible]

## The process

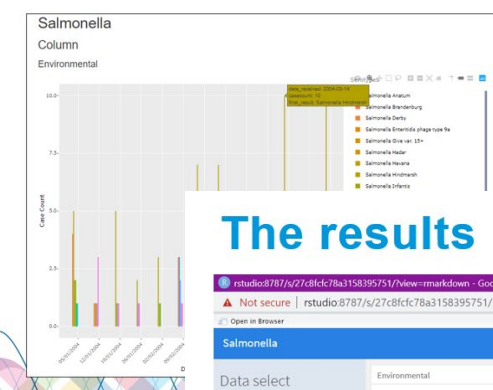


## What can be done

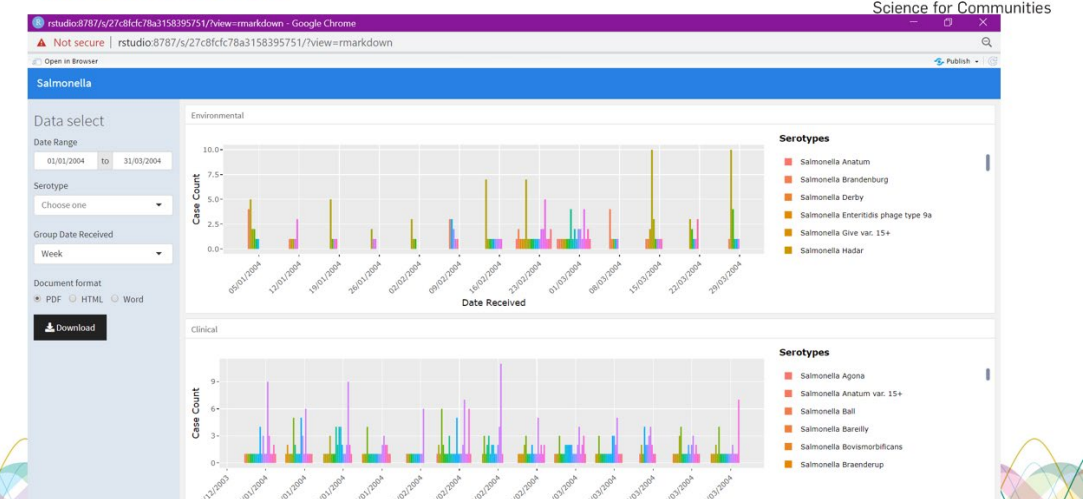
- PDF or Word extracts



- HTML extracts



## The results



## Penelope Hancock

## Retroactive study of the link between Human and Non-human Salmonella serotype for potential link discovery for future samples

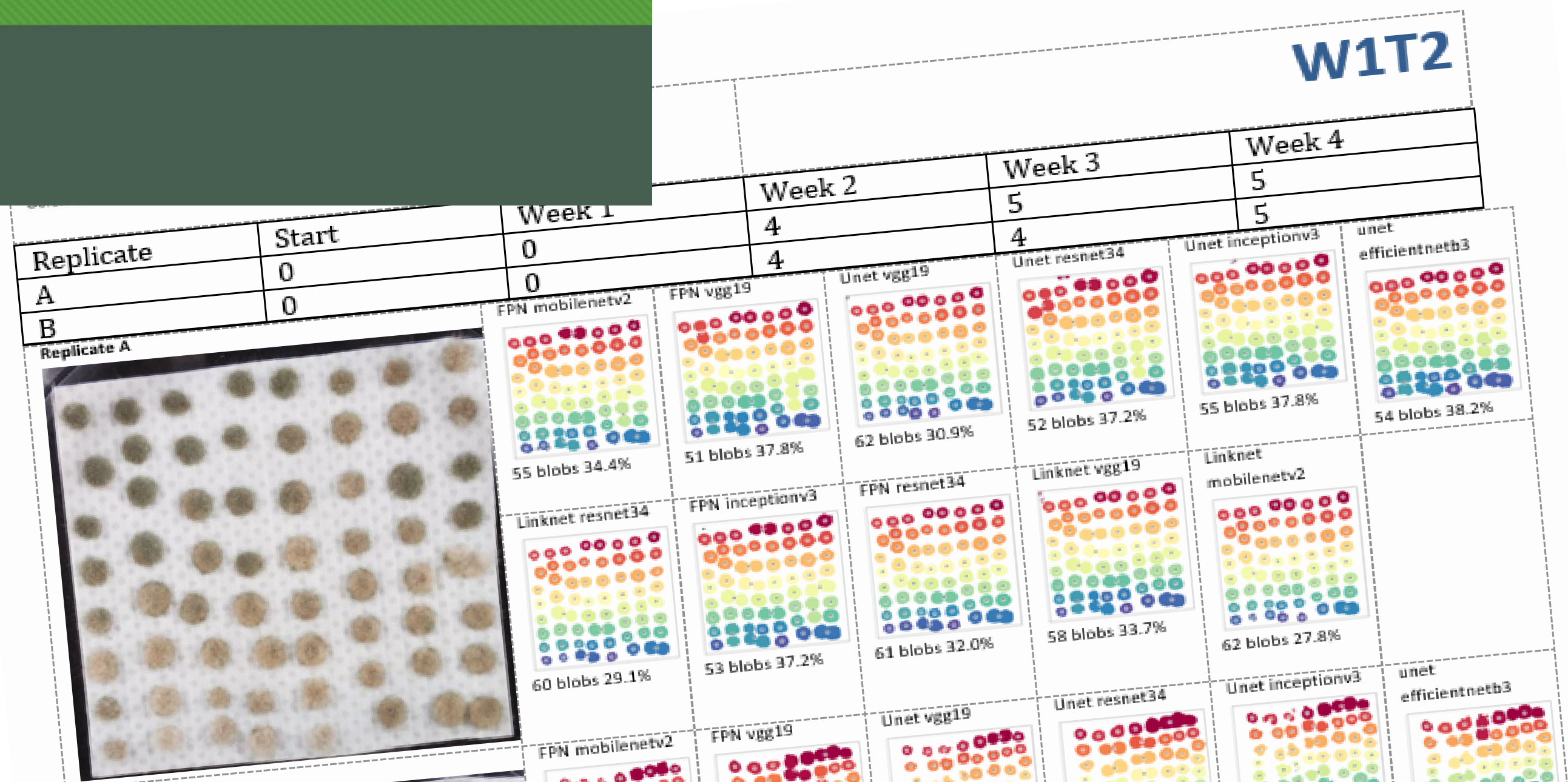
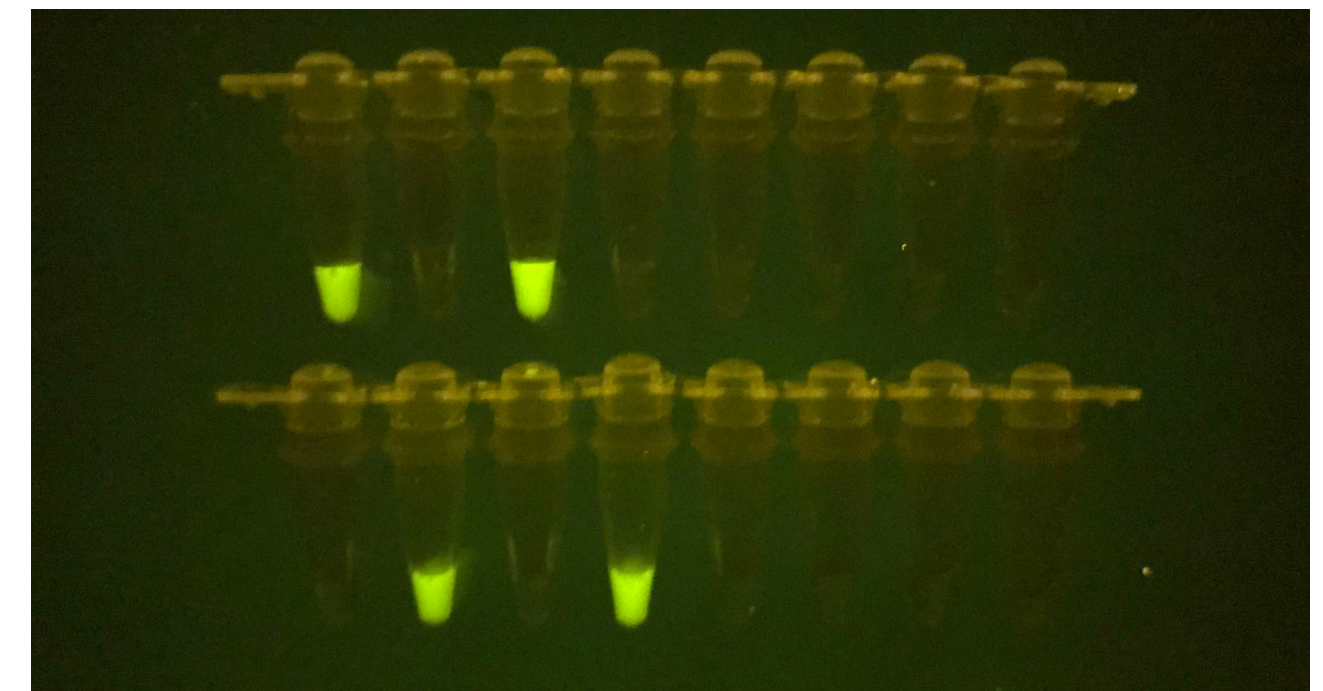
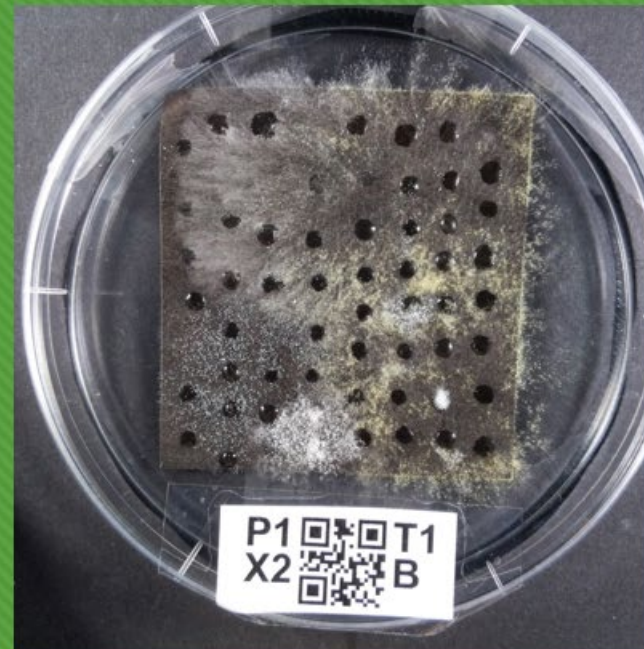


# Fluffy moulds

## Image feature detection for a Mould Growth Index

Izzie Alderton

Mentor – Richard Dean





# Why?

**ES/R** Faecal Source Identification Request

www.waterquality.org.nz email: faecalsource@esr.cri.nz

ESR, Christchurch Science Centre, 27 Croyke Road, Ilam, Christchurch : Ph: (03) 3516011

Contact Person: (for queries)

Name of Sampling Officer

Name and Address for Report if different from contact person

Purchase Order # 1132204

**ES/R** Faecal Source Identification Request

www.waterquality.org.nz email: faecalsource@esr.cri.nz

Contact Person: (for queries)

Name: \_\_\_\_\_

Phone: \_\_\_\_\_

Email: \_\_\_\_\_

Sample Receipt (lab use)

Date & time received: \_\_\_\_\_

Sample checked: Yes/No

General comments: \_\_\_\_\_

Fluorescent Whitering Agent

At least 100 ml of sample is required to be stored in dark at 4°C

Instructions to Lab: Store only

Molecular testing (see water quality.org.nz for details on specific assays or contact esr.cri.nz or Paula Scholten, Ph.D.)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

Instructions: (Please write sample details)

=VLOOKUP(\$A16,'G:\Environmental Health\~ Shared\Molecular Biology Results\MfE pilot study 2020\[MfE qP

Interpretations / Report		S	T	U	V	W	X	Y	Z	AA
ESR No	Date Received (CSC)	General GenBac / 100 mls	Human HF183 / 100 mls	HumanCrA ssphage / 100 mls	Human BiADO / 100 mls	Ruminant BacR / 100 mls	Proportion Ruminant	Ruminant Sheep / 100 mls	Ruminant Cow / 100 mls	Avian GF / 100 m
CMB200045	4/02/2020	83,939	1,549	17,845	54	0	ND	not tested	not tested	
CMB200046	4/02/2020	784,103	137,333	23,400	3,288	127	1% or less	not tested	not tested	
CMB200047	4/02/2020	944,556	0	770	22	150,323	50-100%	0	1,065	
CMB200048	3/02/2020	555,168	10,400	8,133	260	0	ND	not tested	not tested	
CMB200049	3/02/2020	76,891	385	0	4	124	1% or less	not tested	not tested	
CMB200051	4/02/2020	183,537	0	0	0	0	ND	not tested	not tested	
CMB200052	4/02/2020	216,631	0	0	9	0	ND	not tested	not tested	
CMB200053		0	0	0	0	0	ND	not tested	not tested	
CMB200065	11/02/2020	880,015	28,340	51,200	7,048	71	1% or less	not tested	not tested	
CMB200066	11/02/2020	309,196	2,376	2,494	507	0	ND	not tested	not tested	
CMB200067	11/02/2020	72,920	151	271	0	1,532	10-50%	0	0	
CMB200068	11/02/2020	125,724	2,572	655	32	1,439	1-10%	0	0	
CMB200069		1,807,862	148,500	337,000	30,224	1,847	1% or less	41	0	

**ES/R**  
Science for Communities

12 March 2020

To:

Catherine Yeatman  
New Plymouth District Council  
Private Bag 2025  
NEW PLYMOUTH 4342

Email: [Catherine.Yateman@npdc.govt.nz](mailto:Catherine.Yateman@npdc.govt.nz)

From:

ESR Christchurch Science Centre  
PO Box 29181  
CHRISTCHURCH 8540

Email: [faecalsource@esr.cri.nz](mailto:faecalsource@esr.cri.nz)

## REPORT ON FAECAL SOURCE TRACKING ANALYSIS

The following samples were received on 26<sup>th</sup> February 2020 and were analysed for faecal source PCR markers.

ESR Number	Client Reference	Date Sampled	Site Description	Enterococci No /100mL
CMB200206	SW2	25/2/20 14:15	Surface Water Open Drain, Point 2	1,360
CMB200207	SW3	25/2/20 14:09	Surface Water Open Drain, Point 3	1,505
CMB200208	SW4	25/2/20 14:05	Surface Water Open Drain, Point 4	4,100

### Notice of Confidential Information:

If you receive this report in error, please notify the sender immediately. The information contained in this report is legally privileged and confidential. Unauthorised use, dissemination, distribution or reproduction of this report is prohibited.

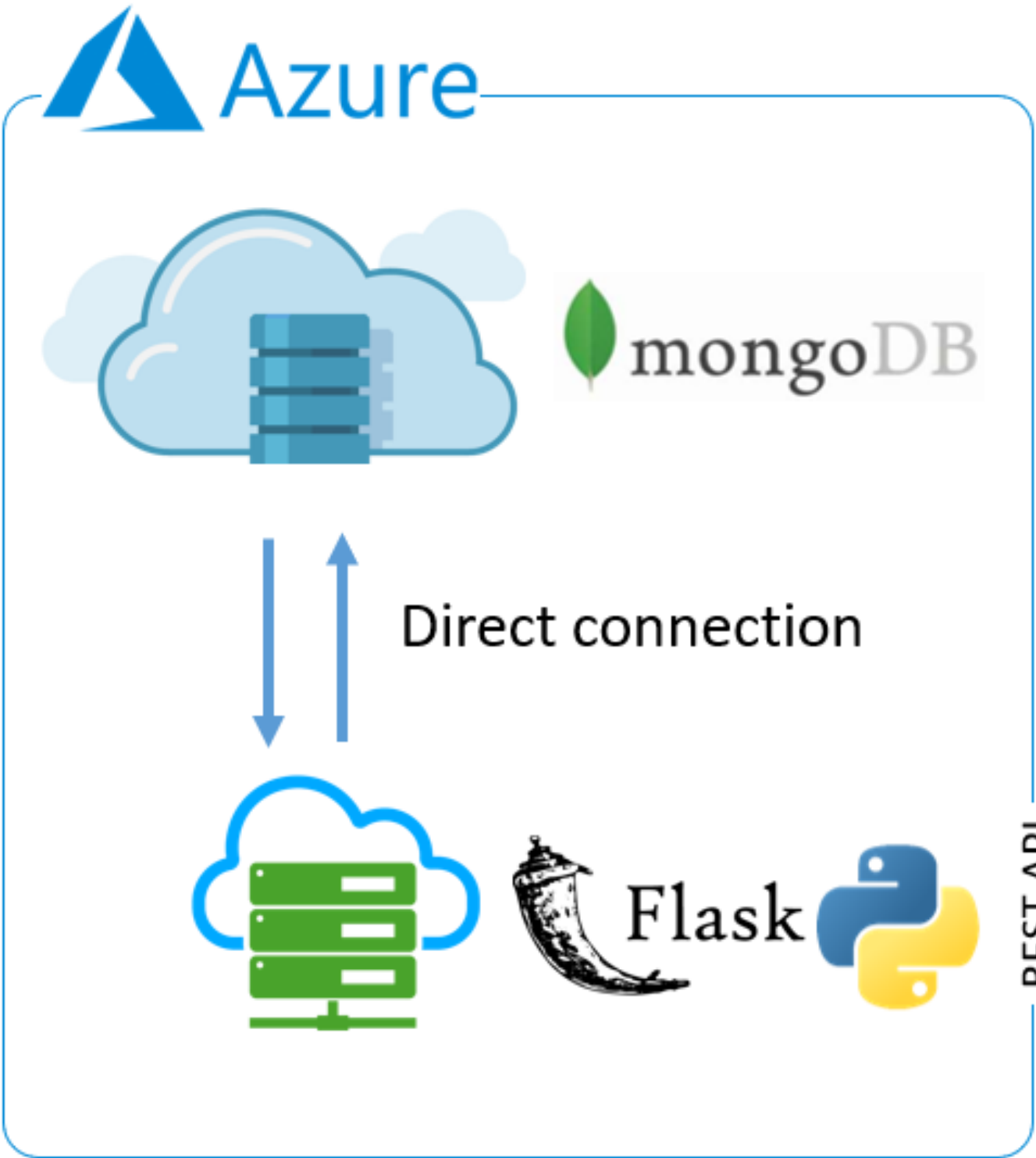
Angela Cornelius + Pierre Dupont + Phani Atmakur

Sampling App and Reporting Dashboard

**R**  
unities



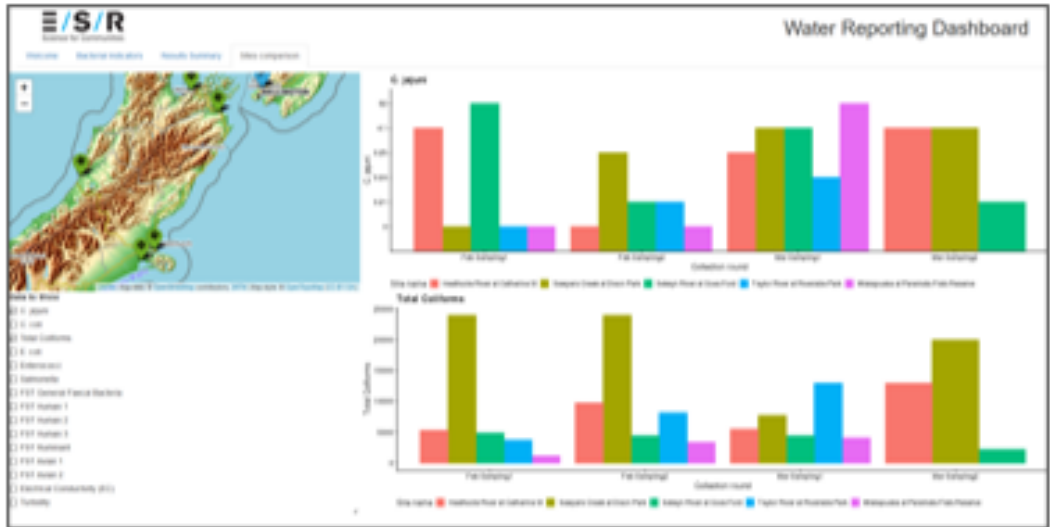
# Structure



On site data entry



Reporting dashboard

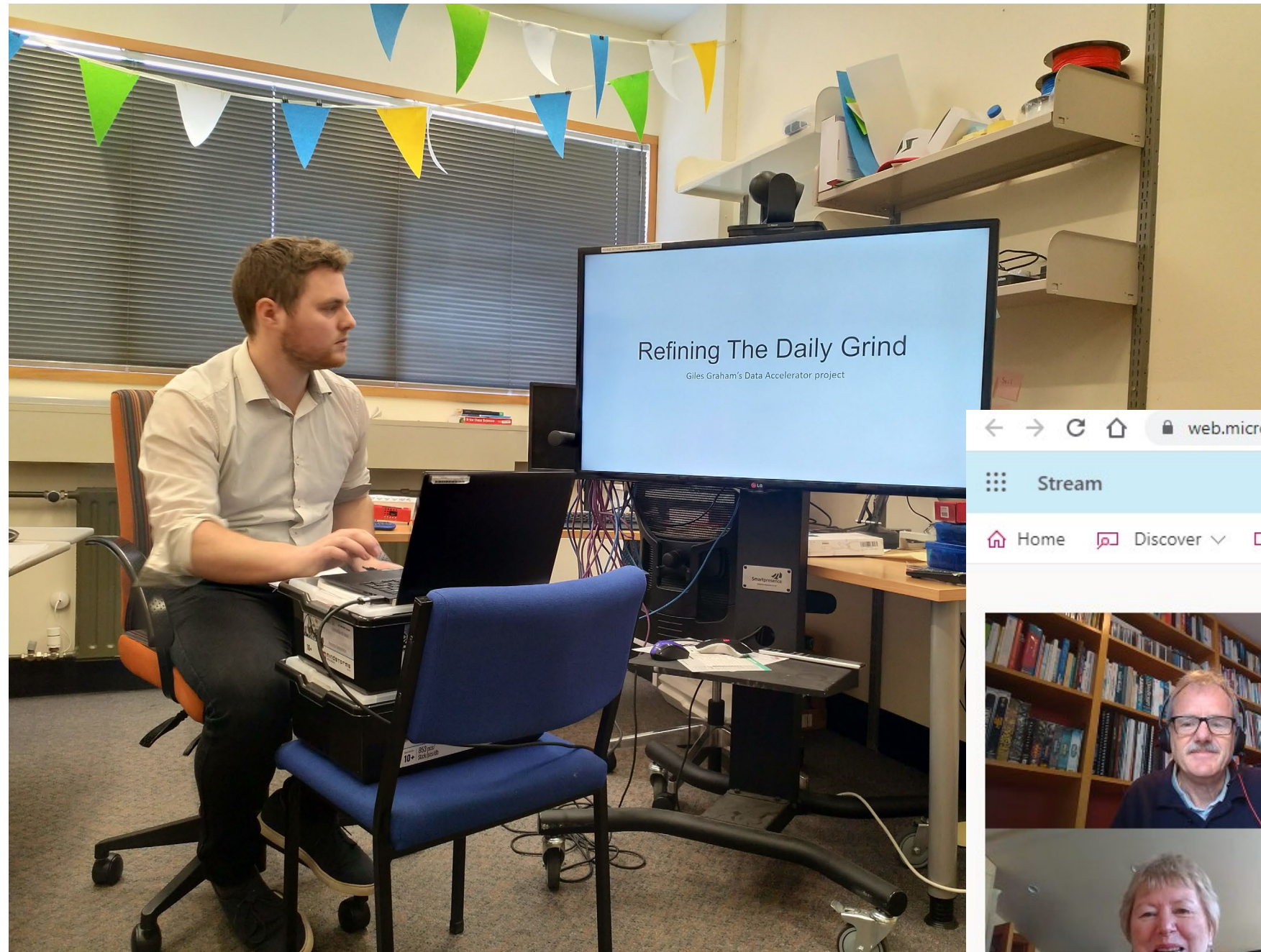


Data from internal and external labs





# Presentation skills



## Housekeeping

- Everyone on mute
- Video off except presenters
- Each presentation – max 7 minutes – warning at 6 mins
- Questions afterwards via email

web.microsoftstream.com/video/e1dcdb6d-9252-4b05-a907-eeaca675630d?referrer=

Stream

Home Discover My content Create Search




E/S/R  
Science for Communities

E/S/R  
Science for Communities



# Power up

 Home Code of Conduct License Improve this page

## Plotting and Programming in Python

### ESR

Online

Sept 22-23, 2020

9:00am - 1:00pm

**Instructors:** Richard  
Shane Sturrock, C  
Jing Wang, Janet  
**Helpers:** Bridget  
Leah Kemp

Stream

Home Discover My content Create Search

Details Interactivity

DSWeek - MedTech, wearables and innovation

Published on 24/09/2020 by Richard Dean Company 4 0

Stream

Home Discover My content Create Search

Microsoft Teams

ESR Data Science week - MedTech, wearables and innovation

2020-09-23 22:03 UTC

Recorded by Richard Dean

Organized by Richard Dean

Play


0:00:01 / 1:55:49

Details Interactivity

DSWeek - MedTech, wearables and innovation

Published on 24/09/2020 by Richard Dean Company 4 0

More from my videos

 DSWeek - Machine Learning ... 1 view 47:45

## General Information

Software Carpentry aims to help researchers get their work done in less time and with less pain by teaching them basic research computing skills. This hands-on workshop will cover basic concepts for programming and plotting in python. Participants will be encouraged to help one another and to apply what they have learned to their own research problems.

For more information on what we teach and why, please see our paper "Best Practices for Scientific Computing".

This course is aimed at ESR staff, who would like to get an introduction to working with python. You don't need to have any previous knowledge of python to attend this course.

Where: This training will take place online. The instructors will provide you with the information you will need to connect to this meeting.

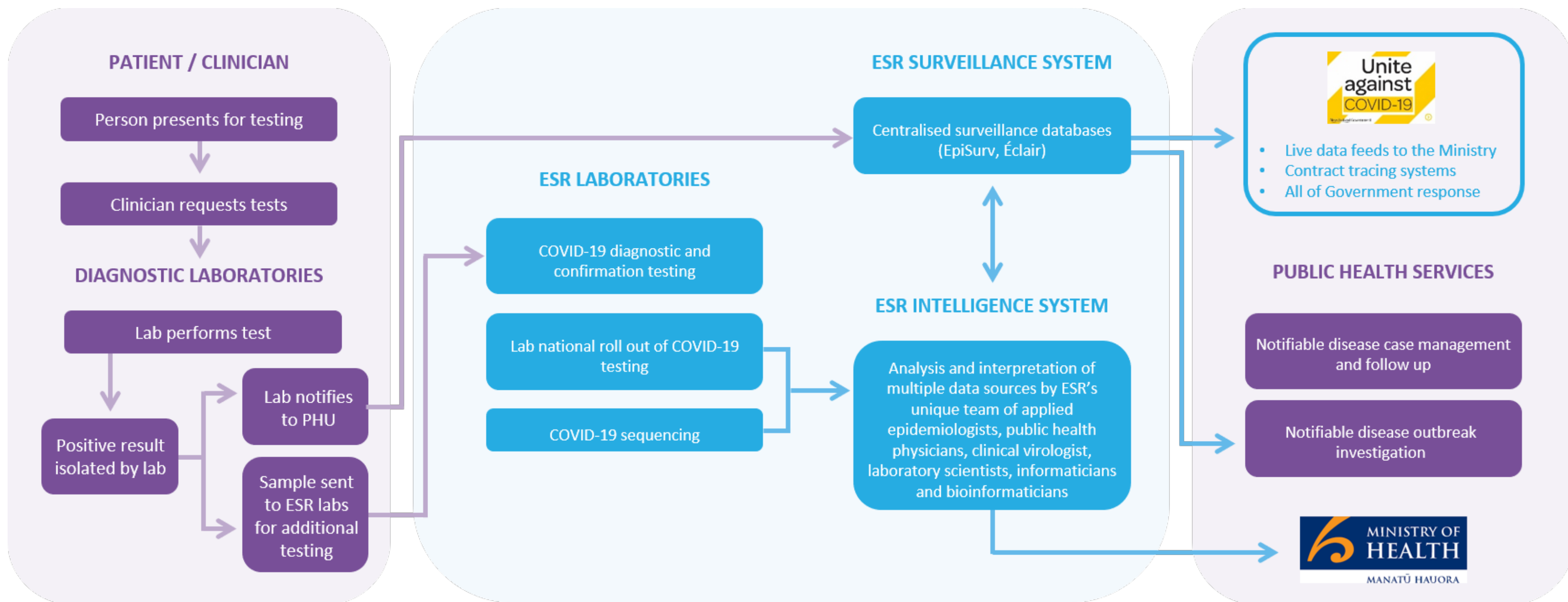




# Power ups don't last forever

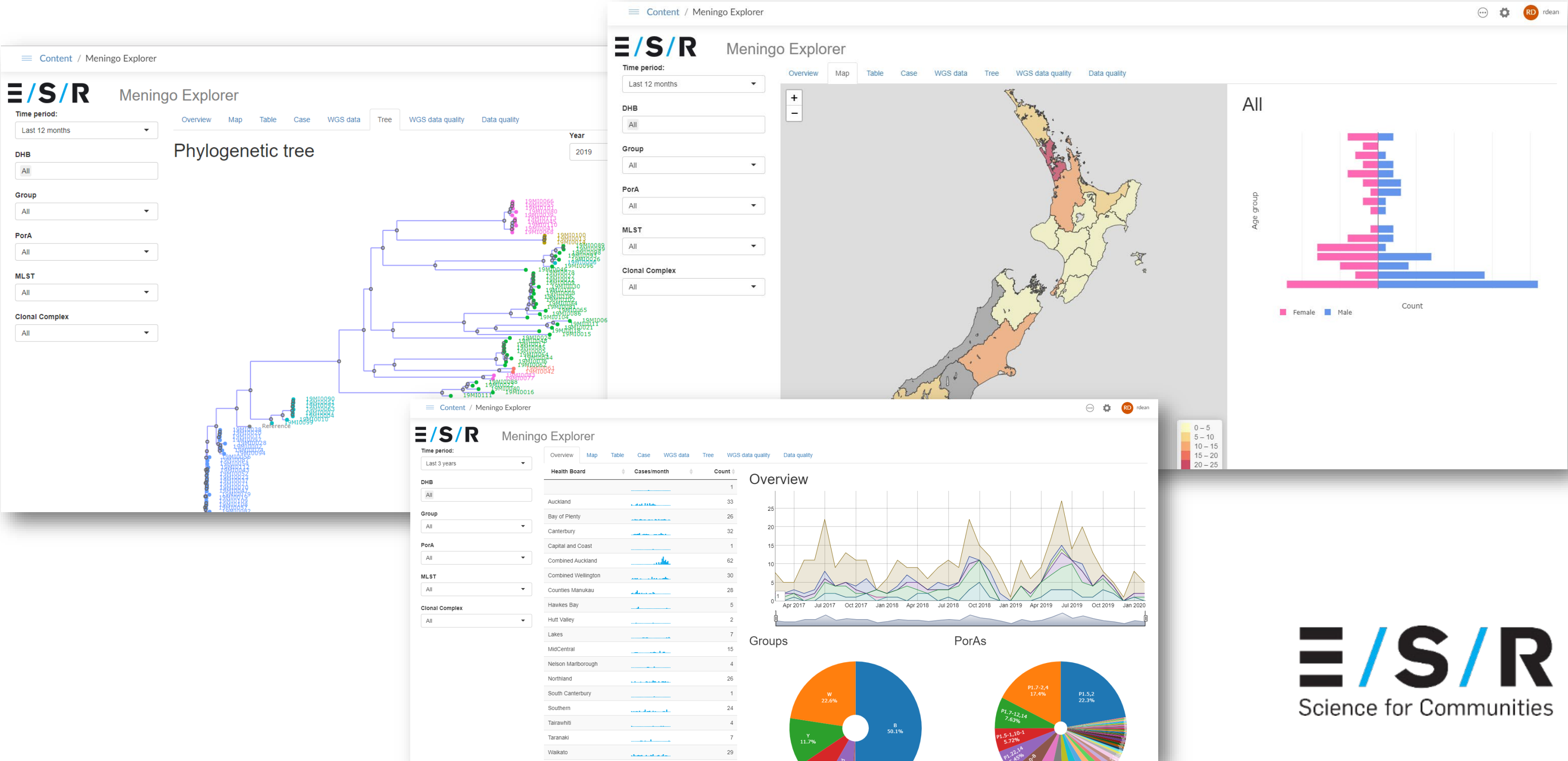






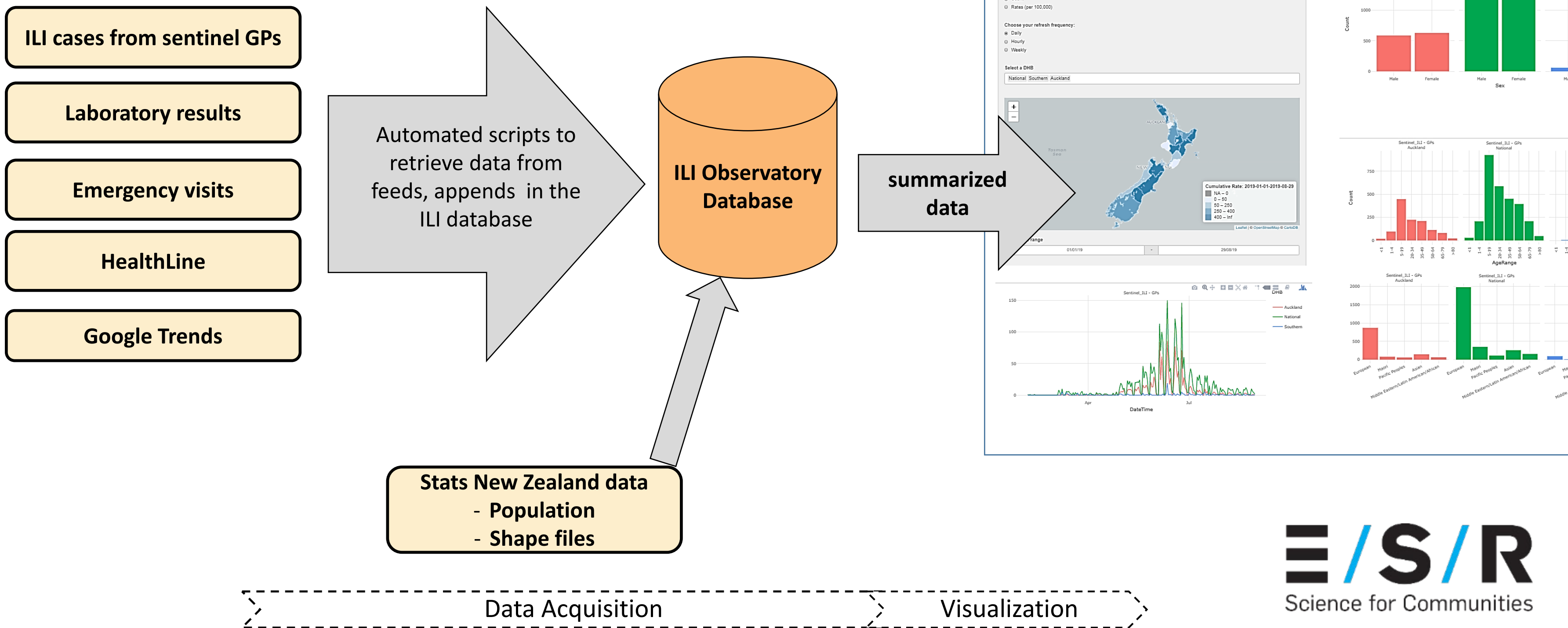


# Genomics to clinical wisdom





# Southern Hemisphere Influenza and Vaccine Effectiveness Research and Surveillance

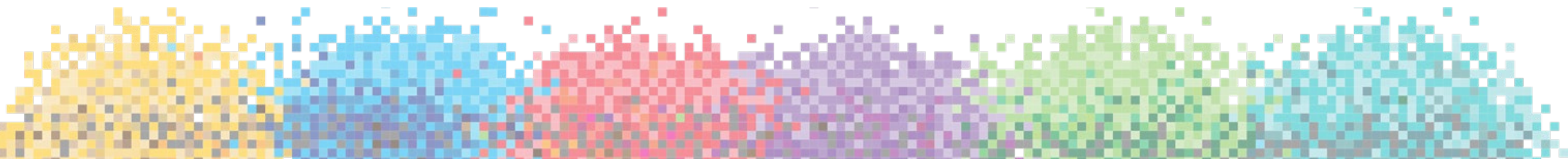




# Data Science Pivot

September 2020

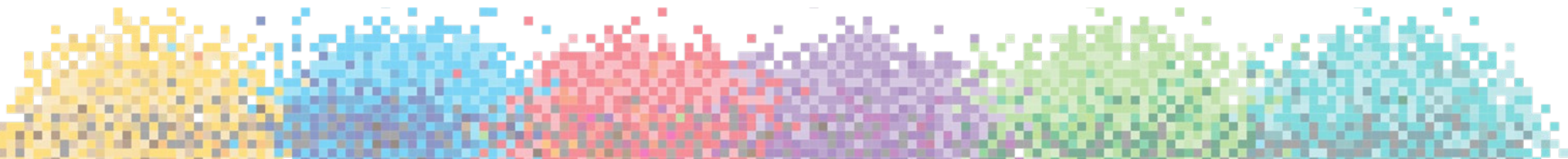
Richard Dean - Data Scientist





# New plan – Primary goals

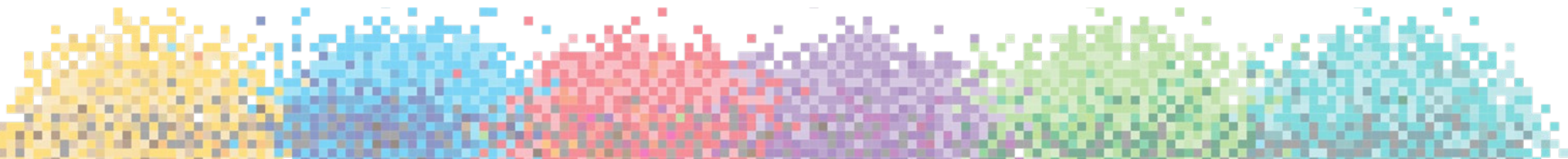
1. Run **Data Science Week** without a session on the IT roadmap
2. Deliver **Python software carpentry** during Data Science Week and repeat x2 in October-November
3. Run a cut down **accelerator cohort 3** from October-March – reduce pressure by extending timelines and only select essential projects which do not place additional pressure on IT
4. Increase availability for **Covid 19 projects** – Ensure that there is sufficient data science capability to assist during the pandemic





# New plan – Secondary goals

5. Establish **data science incubator** to provide ongoing access to data science mentors – e.g. forensic toxicology project from cohort 2
6. Developing stronger links with **NZ universities** – visits and presentations to build groundwork for future data science collaboration – access to expertise through adjunct and shared appointments
7. Connect with **data science teams in other CRIs** – follow recommendation 5 from the CRI review to work collaboratively to build data science capability
8. Begin groundwork for **data science week 2021**
9. Remain **agile** and **help others**





# Cohort 3

- 3 projects
- 3 sites
- November – March 2021
- 4.5 mentors
- 4 mentees

Site	Participant(s)	Title
CSC	Louise Weaver Judith Webber	Groundwater omics for machine learning predictions of groundwater health
KSC	Sarah Underwood	Automation of SARS-CoV-2 EQA programme reporting
MASC	Jason Min	Reagent supply management



# Data Science Accelerator Level 2: Power-up or reboot?

## Today's objectives:

- 1) What happened since eResearch2020
- 2) Review progress
- 3) Look forward to 2021

