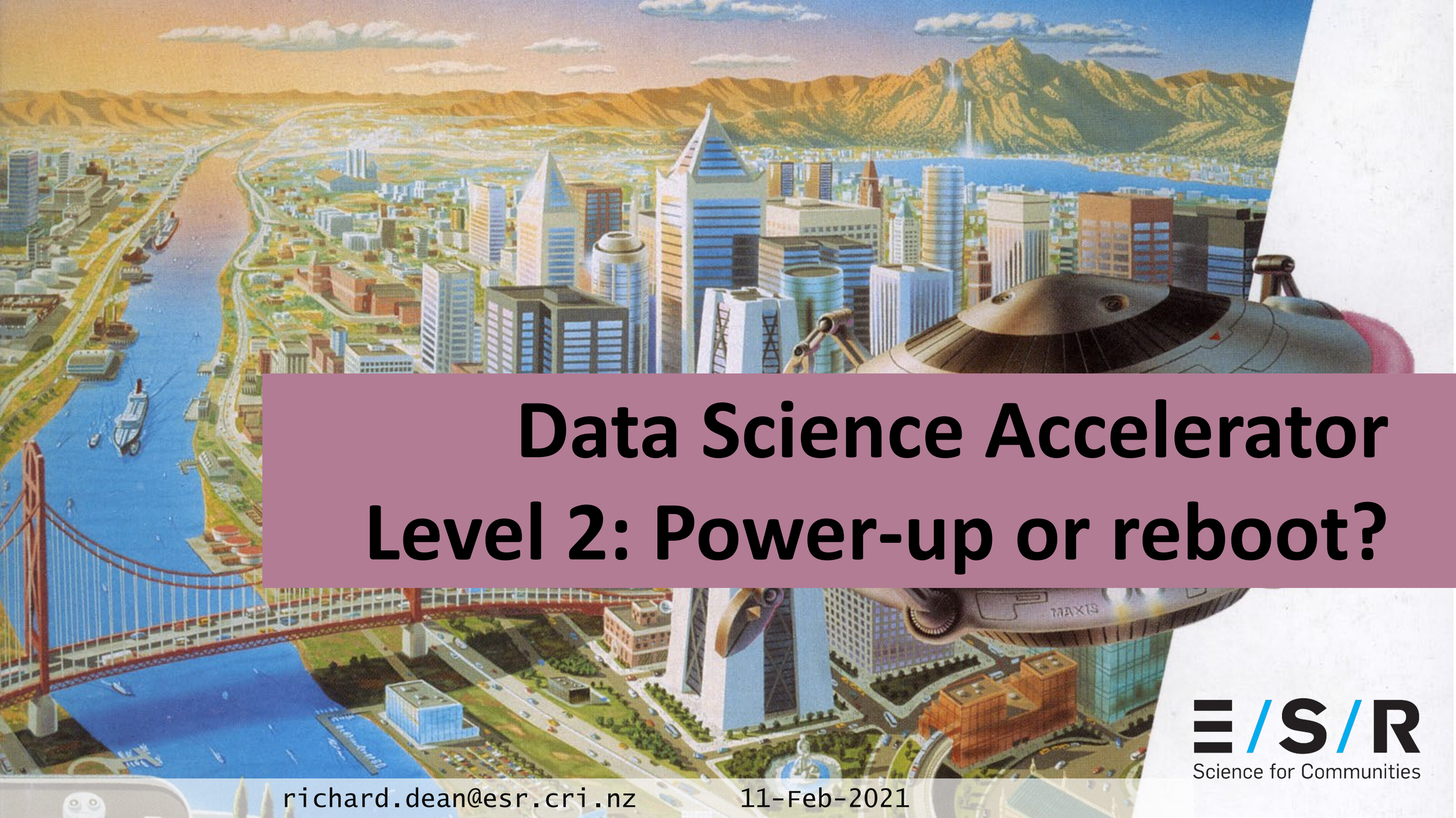


Data Science Accelerator
Level 2: Power-up or reboot?

richard.dean@esr.cri.nz

11-Feb-2021

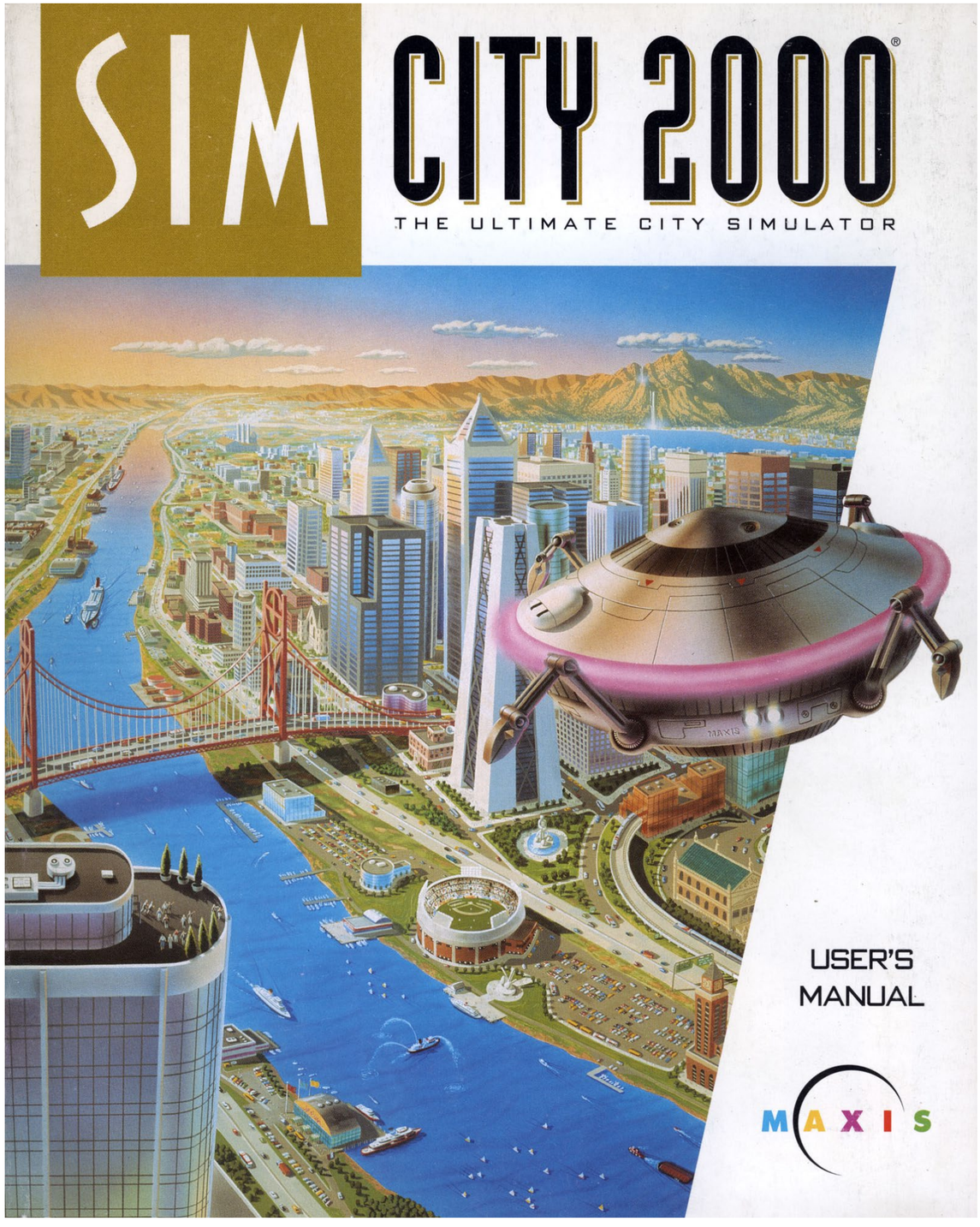
E/S/R
Science for Communities

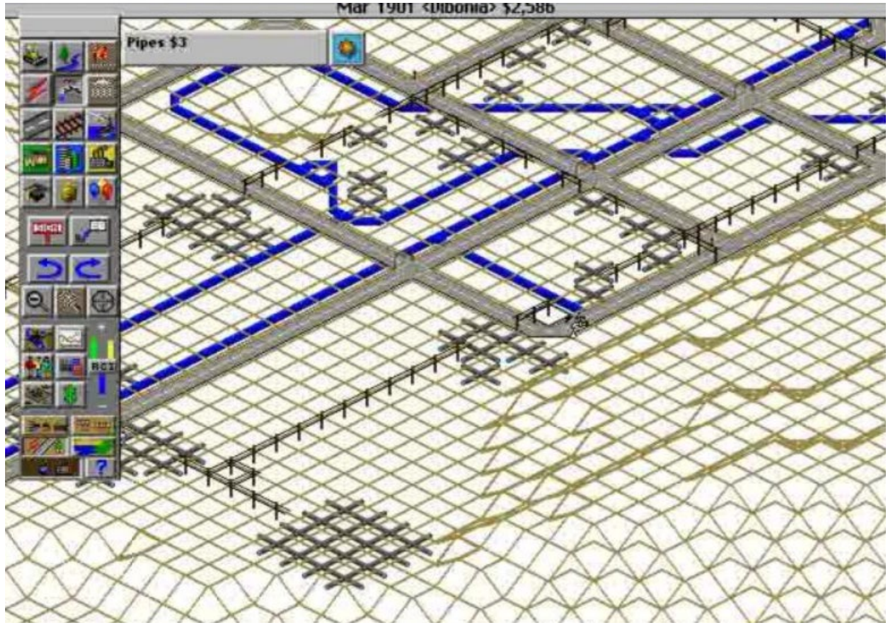
Data Science Accelerator
Level 2: Power-up or reboot?

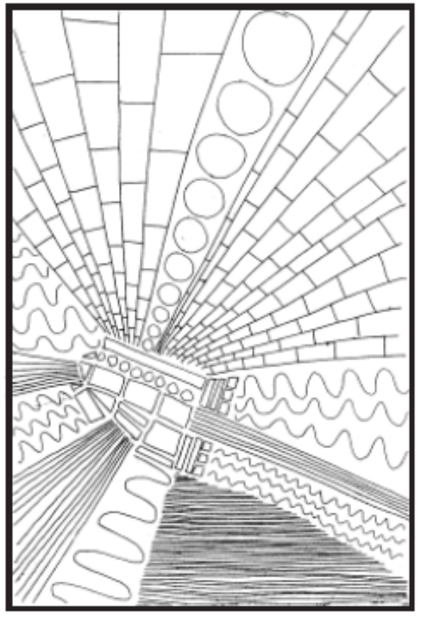
Today's objectives:

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

E/S/R
Science for Communities



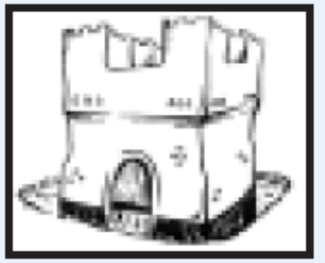




E/S/R
Science for Communities

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.

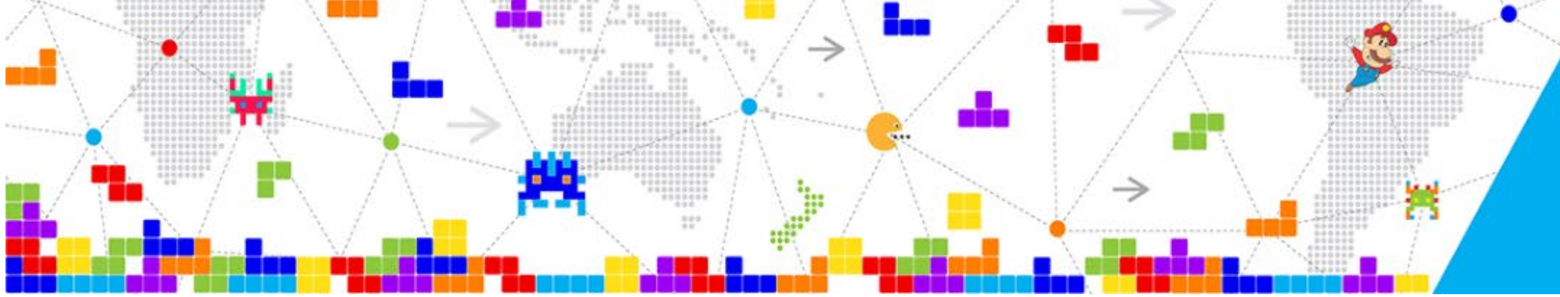
Cities are for people: a place for their hopes and dreams, their work and play, their homes and homes for their children. Cities are alive and



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

E/S/R
Science for Communities



Level Up

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

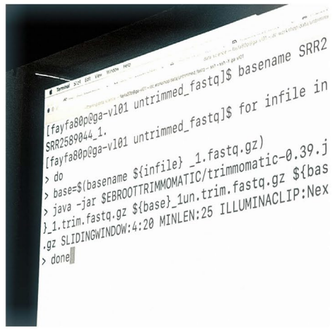
Dealing With Disaster

Science for Communities

Dealing With Disaster

E/S/R
Science for Communities

Levelling up at ESR



E/S/R
Science for Communities

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.

Science for Communities

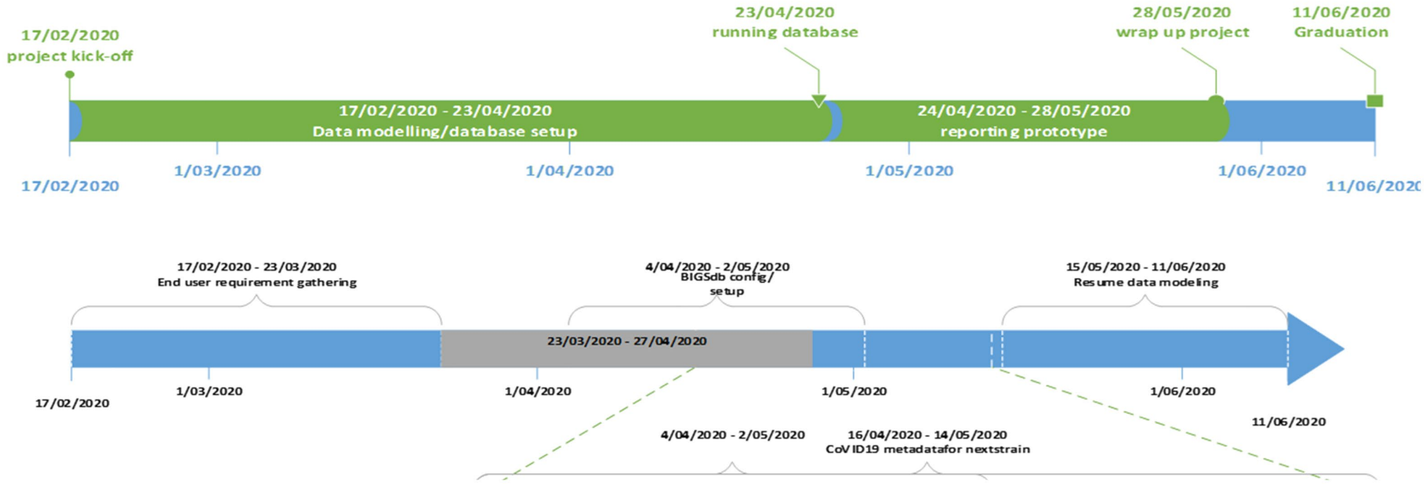
After eResearch 2020…

	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)

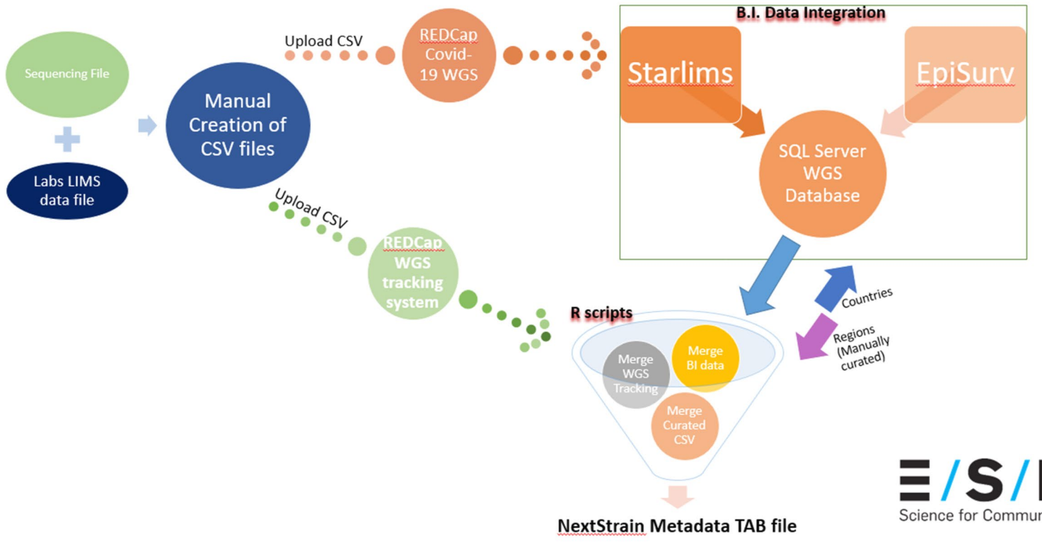
Level 2

Science for Communities

Planning Vs Reality

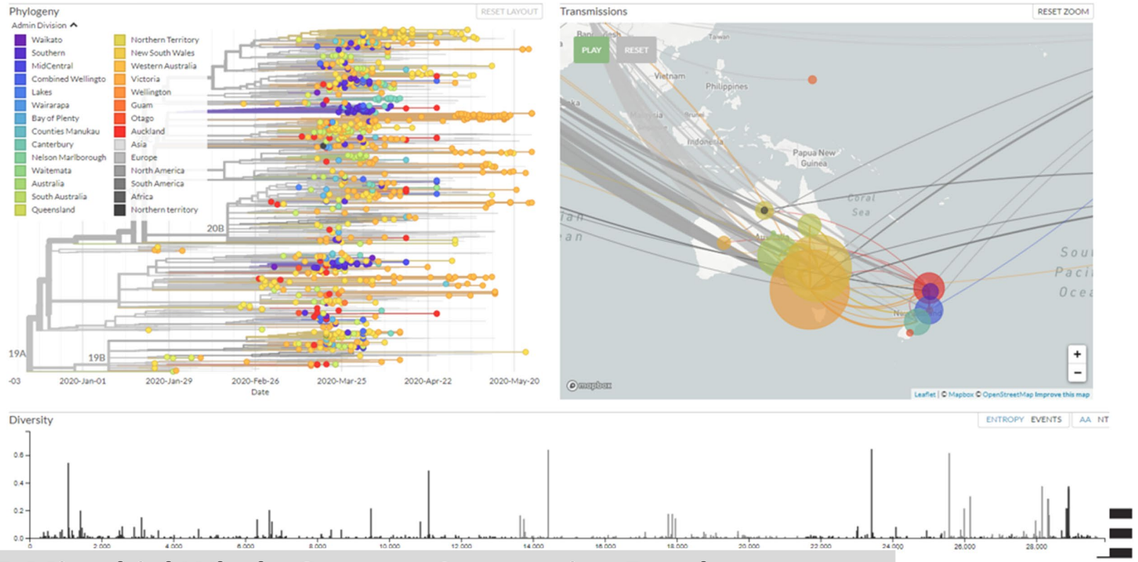


Nextstrain COVID-19 Data Integration and automation



E/S/R Science for Communities

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

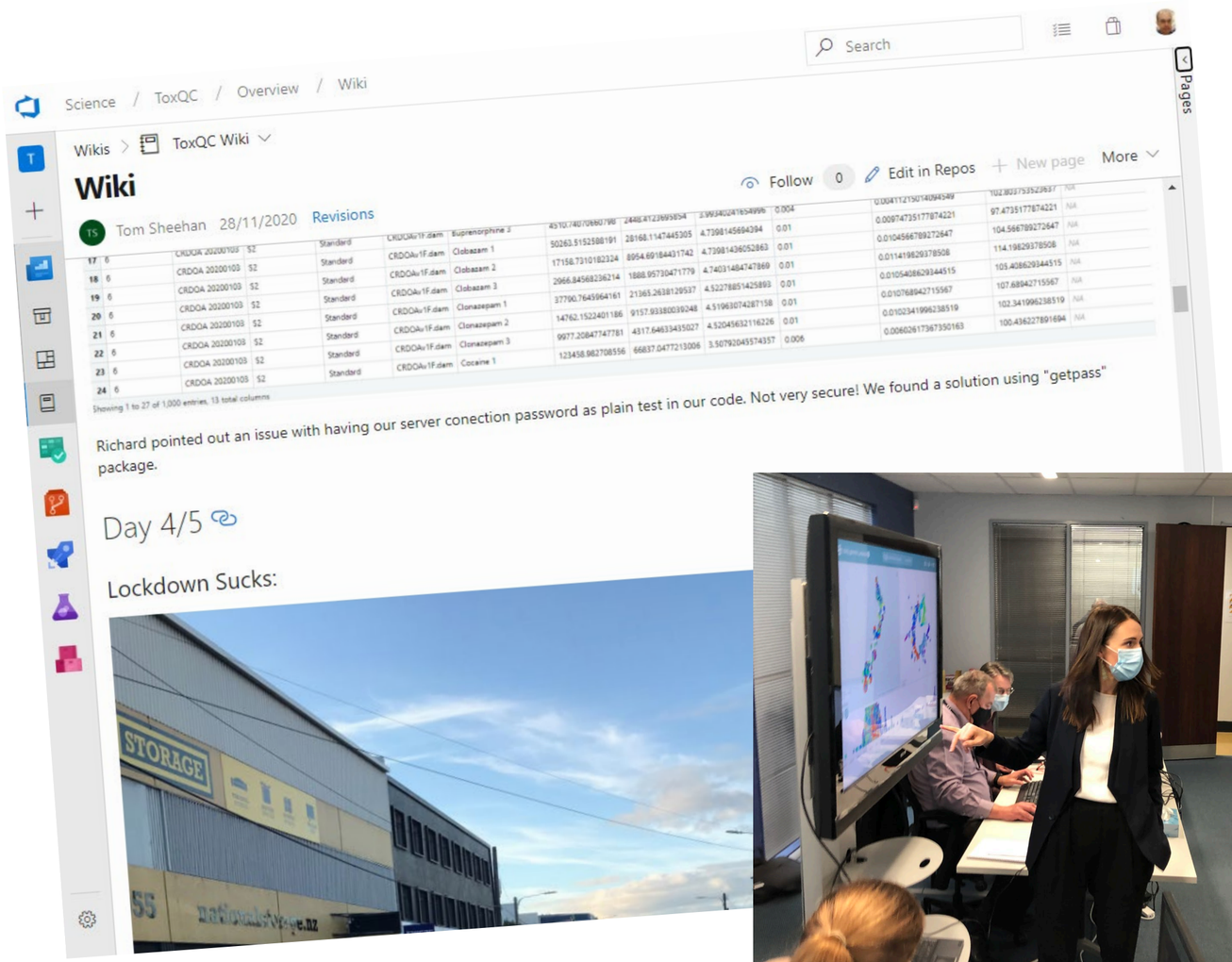


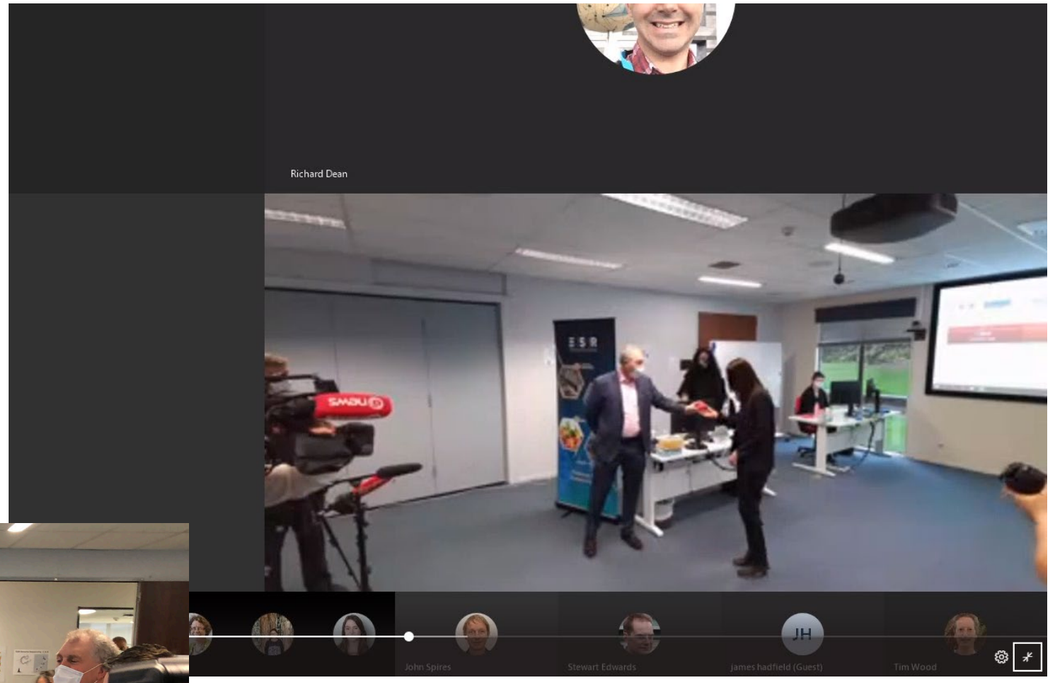
E/S/R Science for Communities

Jing Wang + Hermes Perez

A Unified Reporting Dashboard for ESR Microbial Whole Genome Sequencing Results

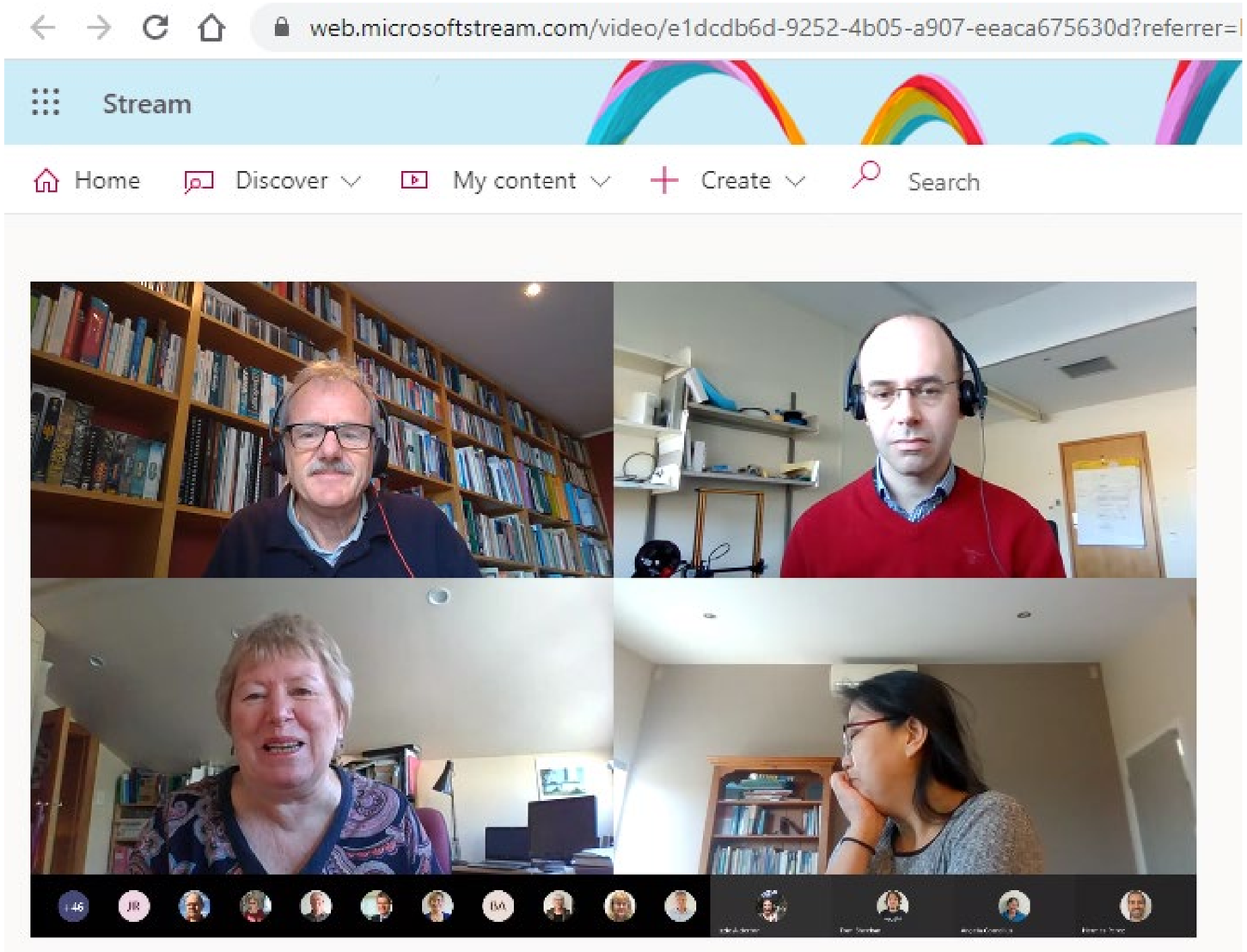
Powering on through lockdown





Science for Communities

Cohort 2 graduation



Science for Communities

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

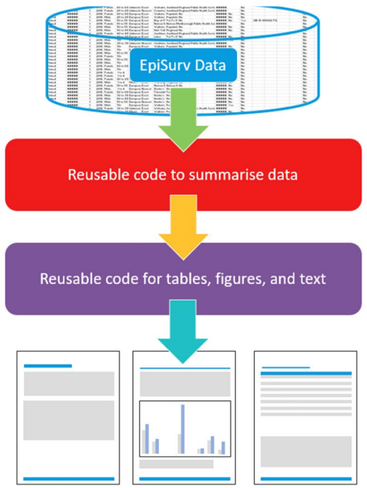
EpiSurv database

Automate data summary

Automate report elements

- Tables
- Figures
- Text

Create Foodborne Disease Report



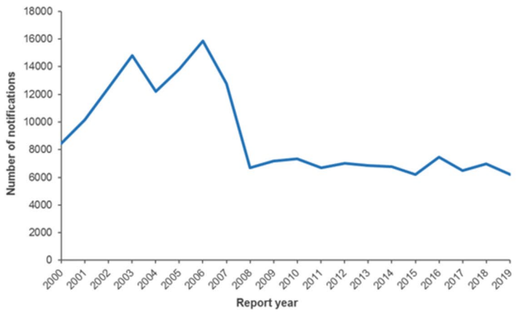
The Products

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 ir 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.

New Skills

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

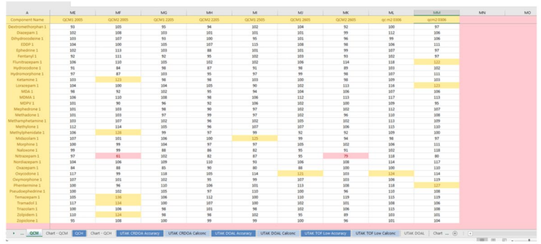
E/S/R
Science for Communities

MAPE

Making **A**ssurance **P**retty and **E**asy

Instruments Collecting Data

Spreadsheets We Use Currently



Our Journey

- R programming
- Collaboration
- Handling Data

Project Goals

Saving Everyone's Time

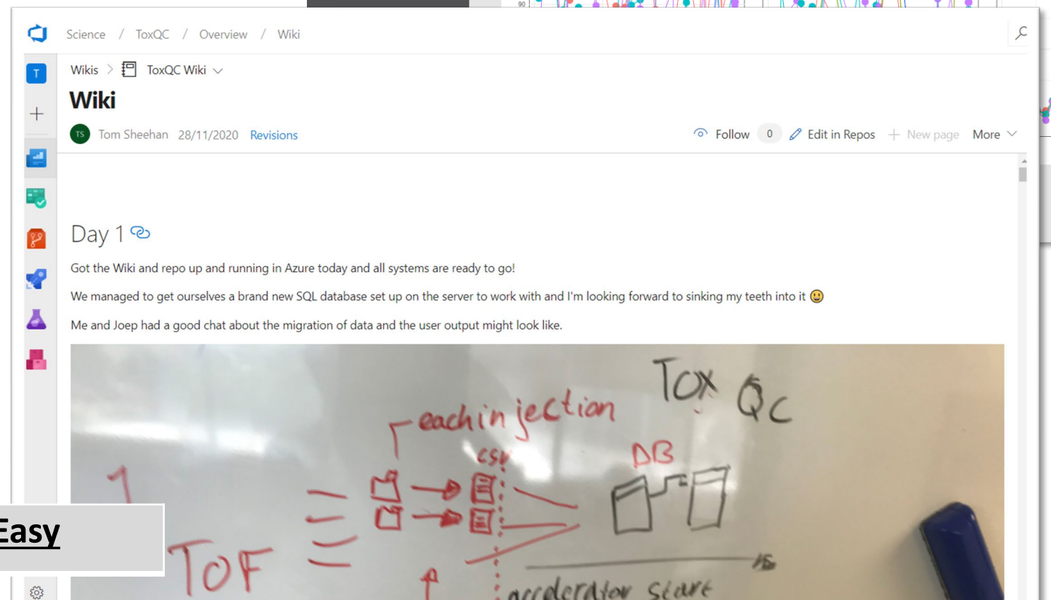
- Copying to spreadsheets alone cost around 30 min a day.
- Finding data in too many different spreadsheets takes time.
- Review multiple spreadsheets take time.

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



Tom Sheehan + Andrew Murray

Making QA Pretty and Easy

Science for Communities

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

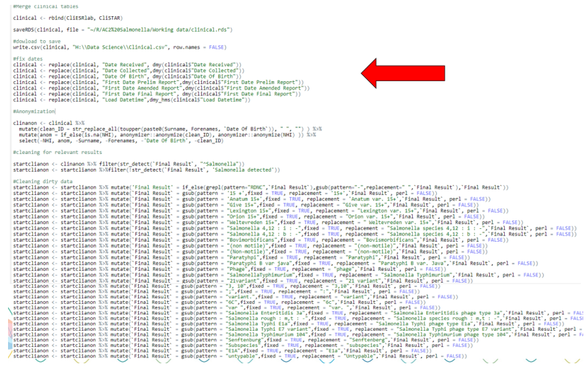
My data science journey

Penelope Hancock

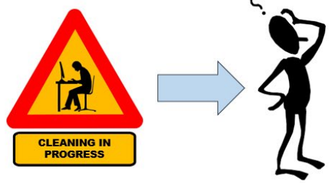
Mentors:

Tim Wood & Ben Waite

The process



The process

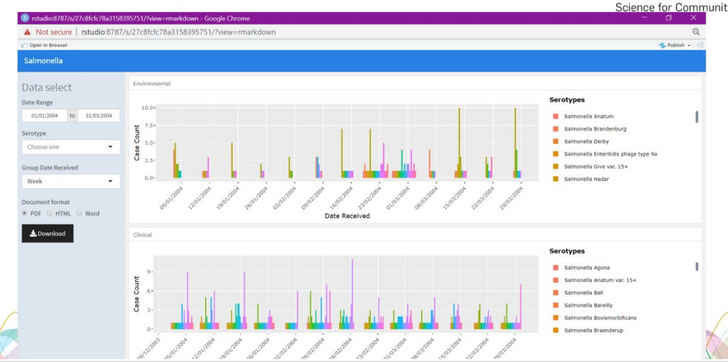


What can be done

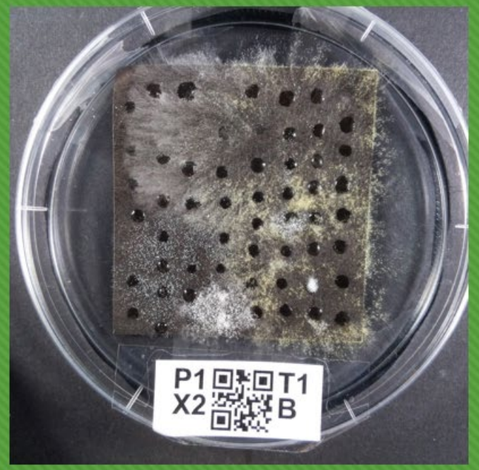
- PDF or Word extracts

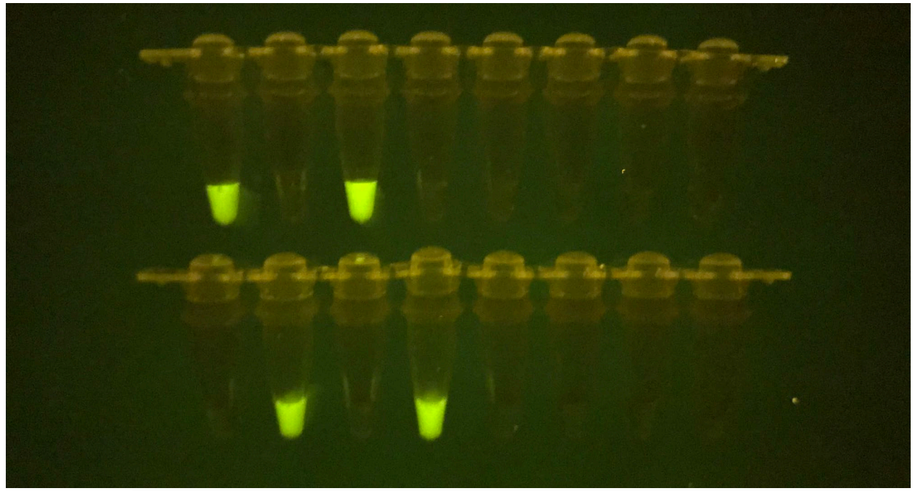
- HTML extracts

The results



Fluffy moulds
-
Image feature detection for a Mould Growth Index

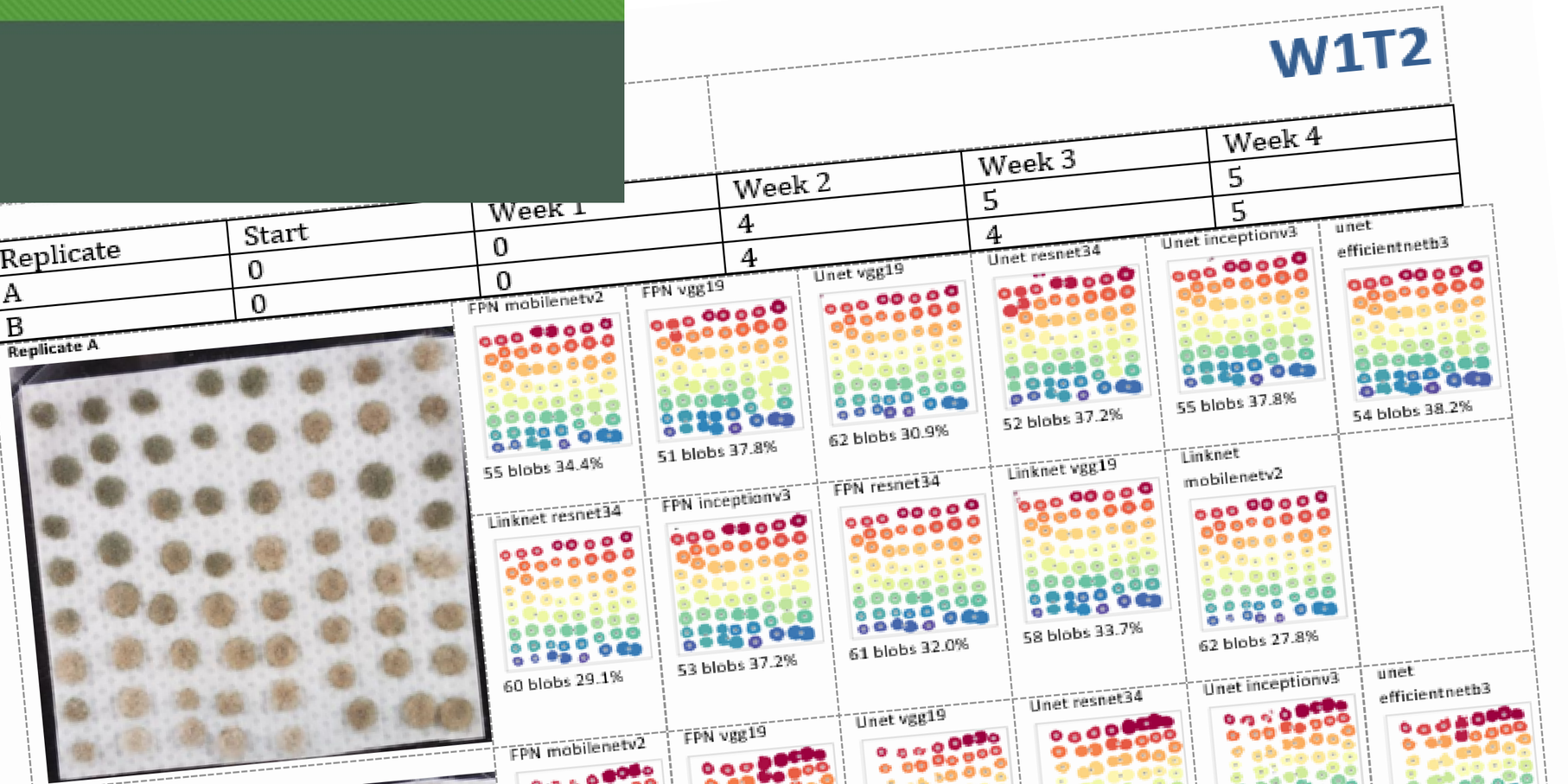




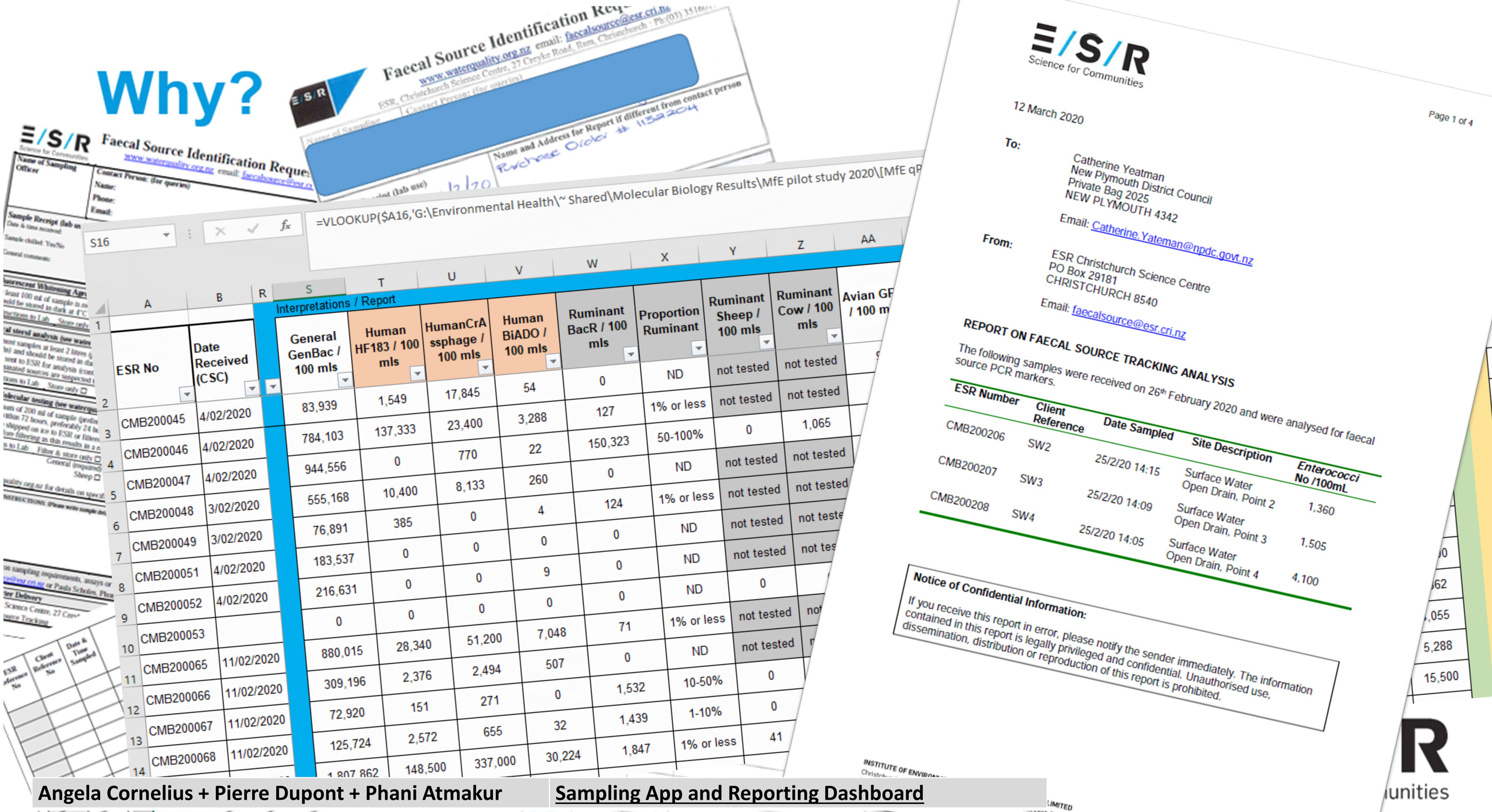
Izzie Alderton
Mentor – Richard Dean

W1T2

Replicate	Start	Week 1	Week 2	Week 3	Week 4
A	0	0	4	5	5
B	0	0	4	4	



Why?



Angela Cornelius + Pierre Dupont + Phani Atmakur

Sampling App and Reporting Dashboard

Structure

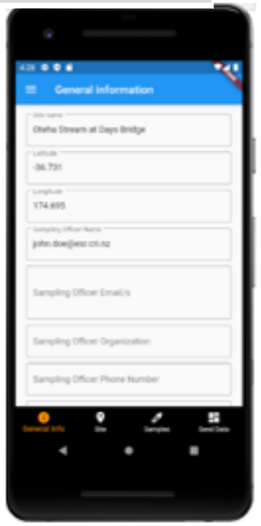
Azure

mongoDB

Direct connection

Flask

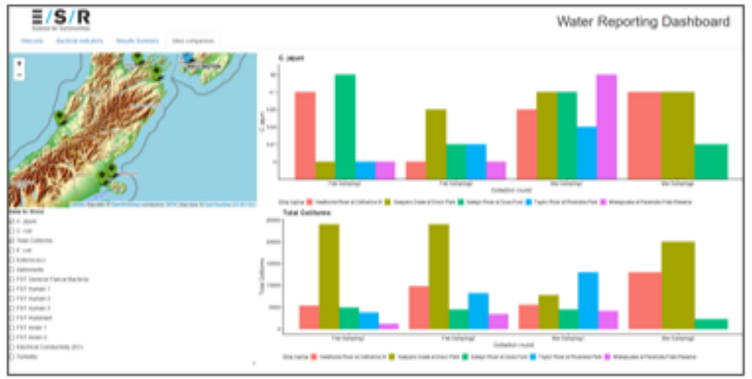
REST API



On site data entry

android

Reporting dashboard

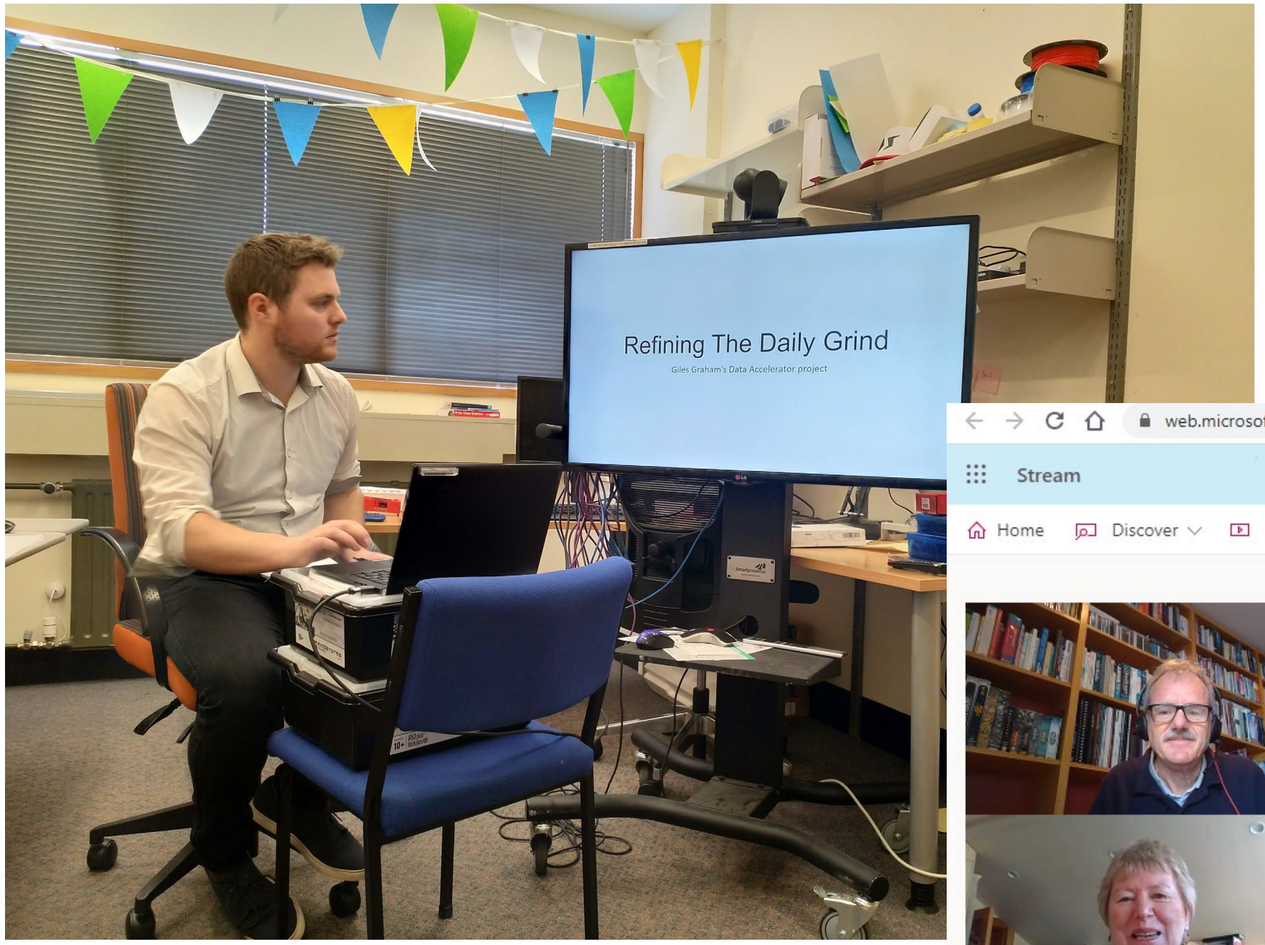


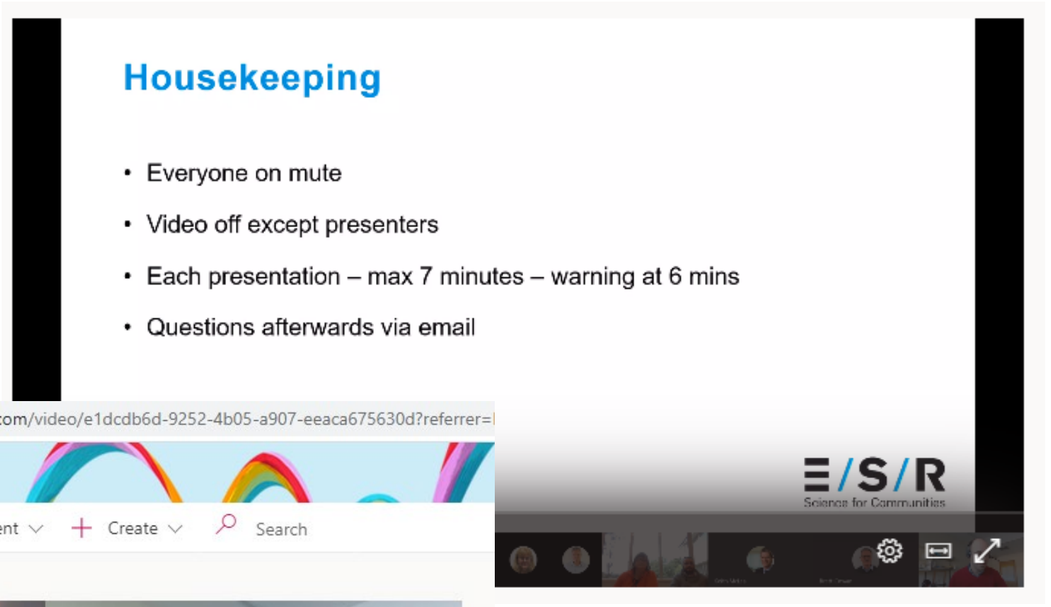
Shiny

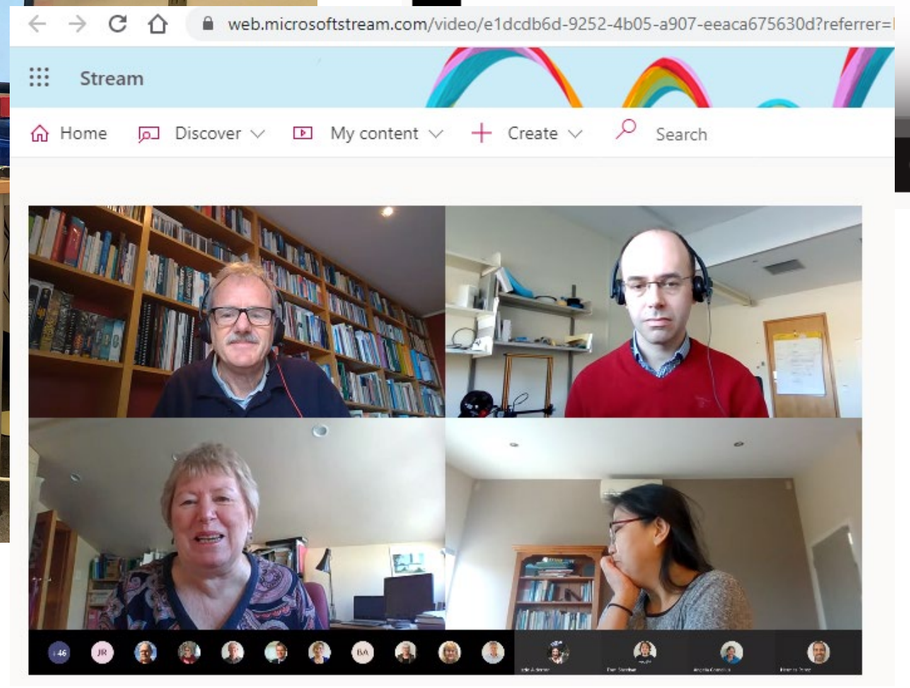
Data from internal and external labs

E/S/R
Science for Communities

Presentation skills







E/S/R
Science for Communities

Power up

Home Code of Conduct License Improve this page

Plotting and Programming in Python

ESR

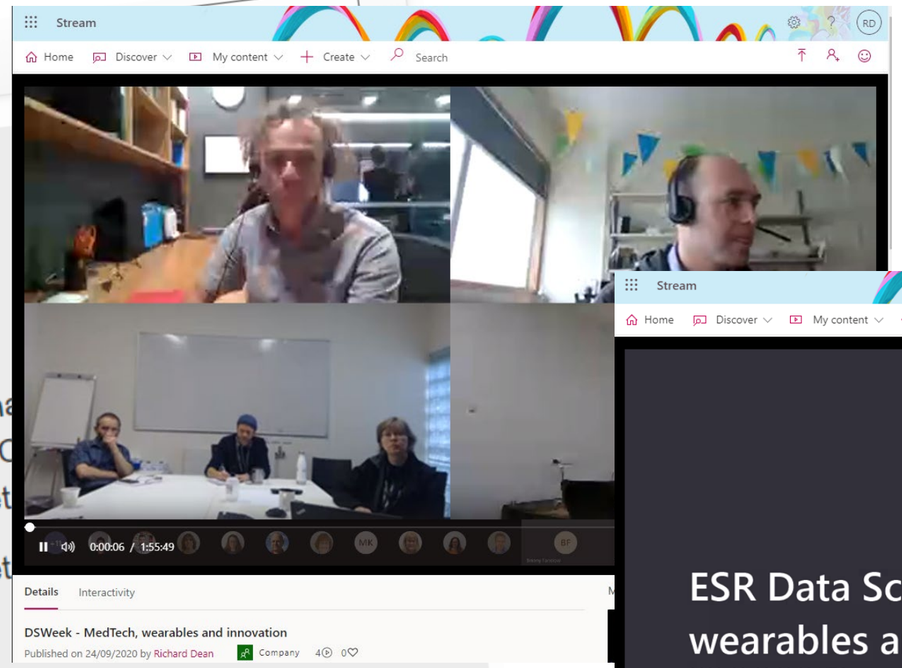
Online

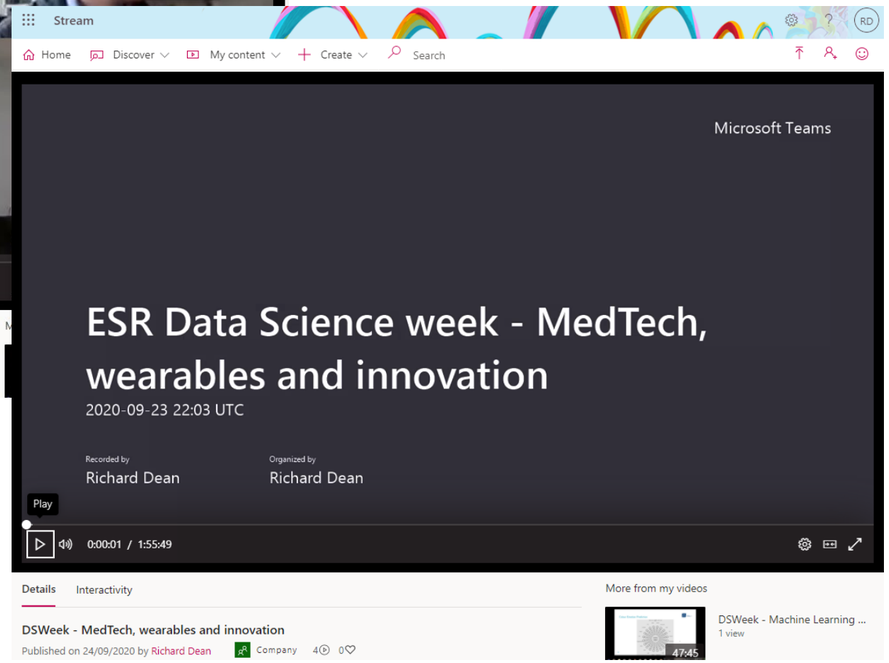
Sept 22-23, 2020

9:00am - 1:00pm

Instructors: Richa
Shane Sturrock, C
Jing Wang, Janet

Helpers: Bridget
Leah Kemp





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.

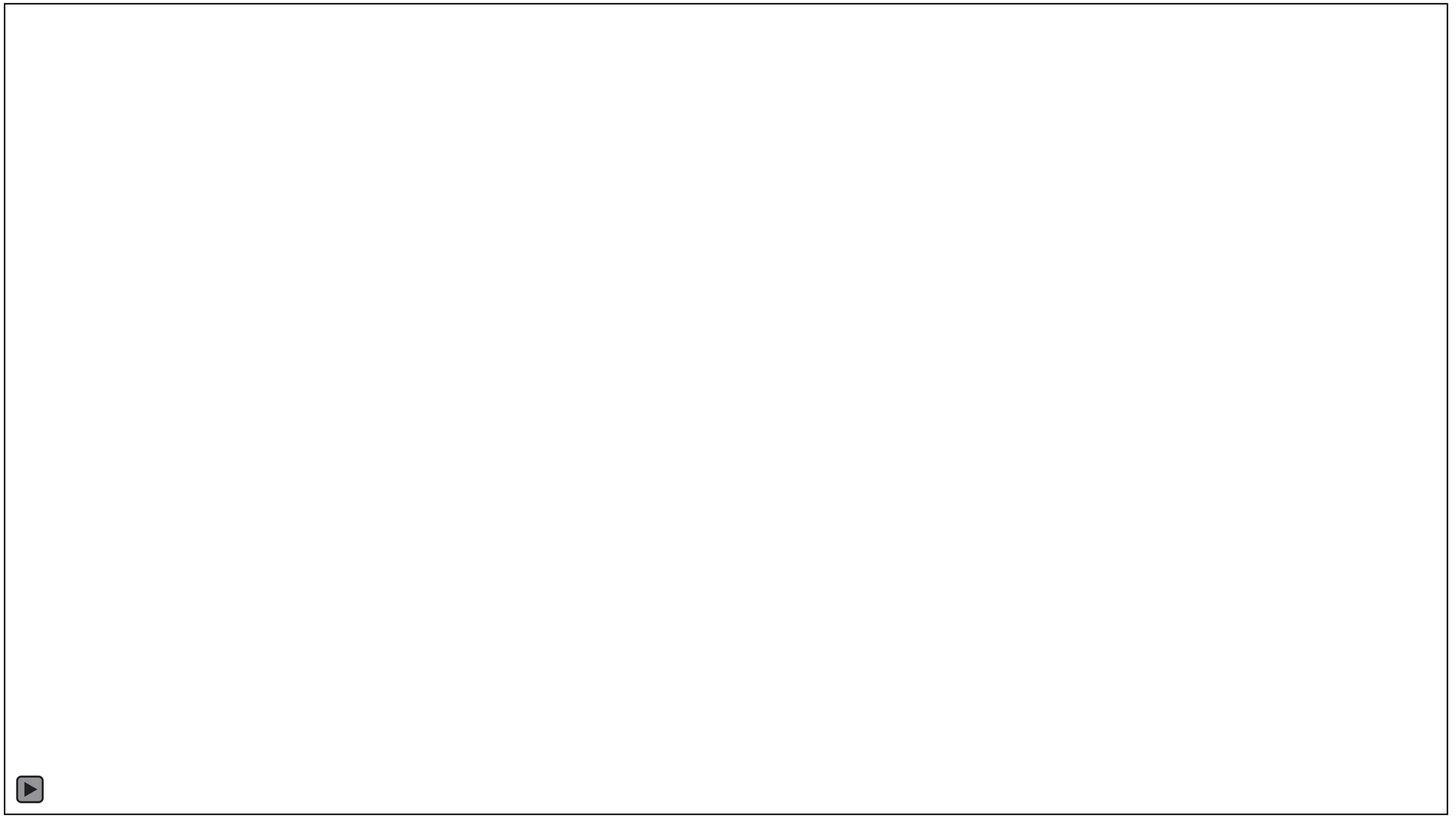
For more information on what we teach and why, please see our paper "Best Practices for Scientific Computing".

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.

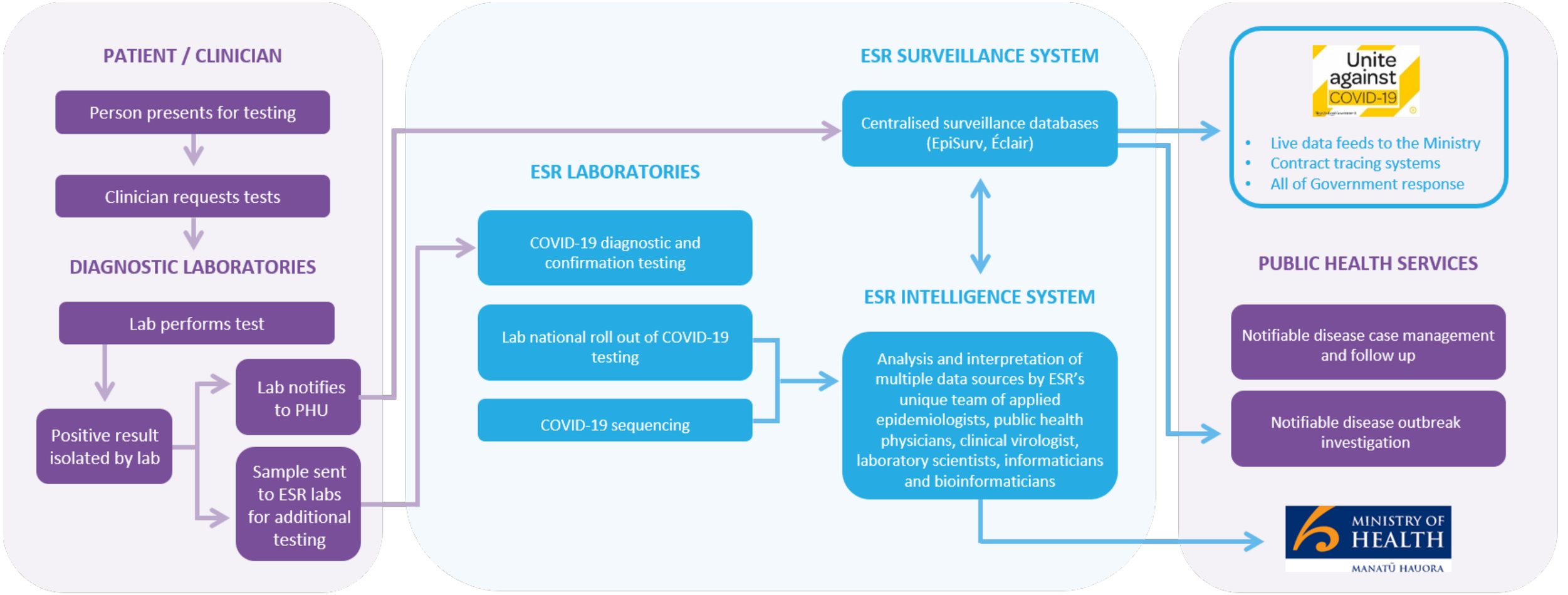
Where: This training will take place online. The instructors will provide you with the information you will need to connect to this meeting.

E/S/R
Science for Communities

Power ups don’t last forever



Science for Communities

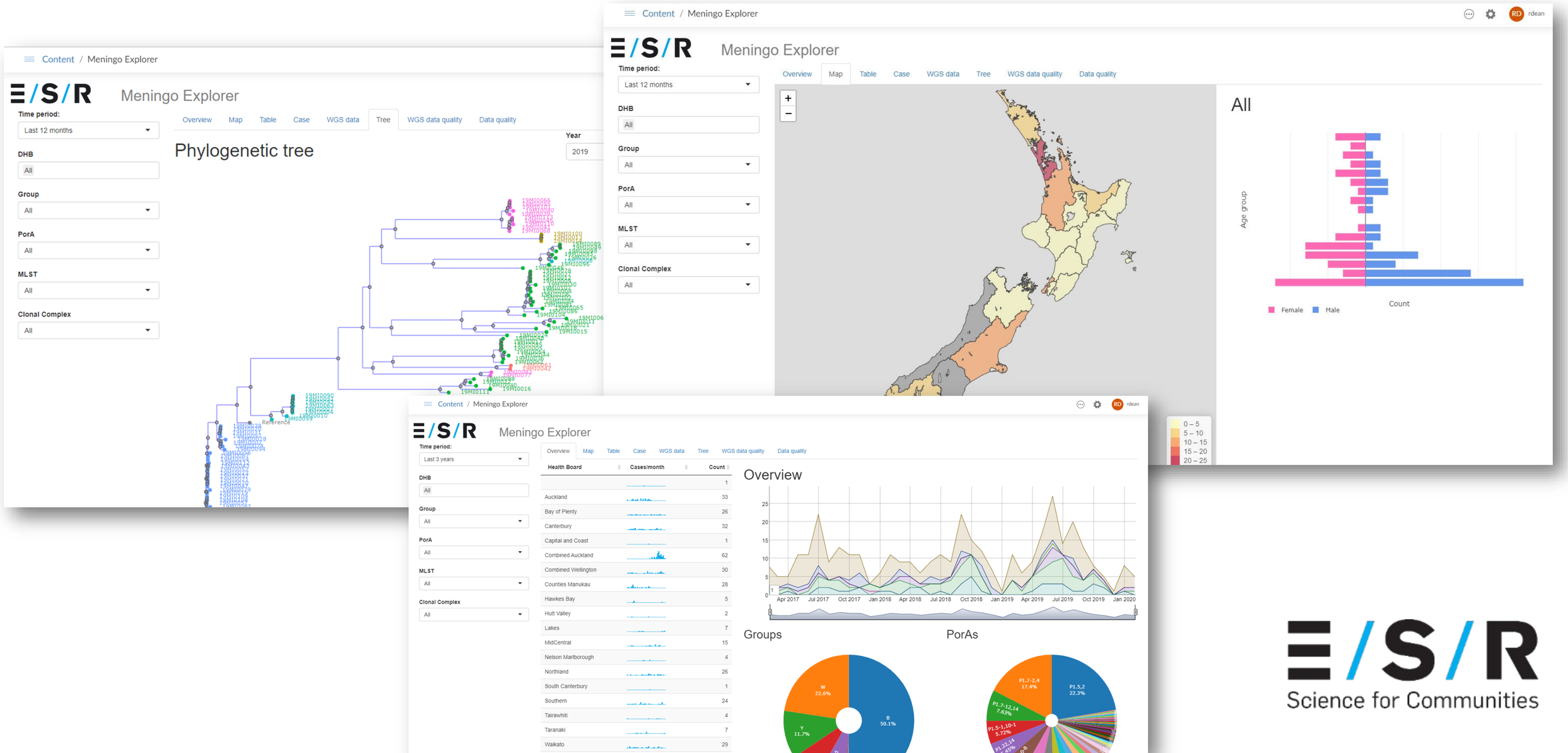


ESR

PUBLIC HEALTH

Science for Communities

Genomics to clinical wisdom



E/S/R
Science for Communities

Southern Hemisphere Influenza and Vaccine Effectiveness Research and Surveillance

ILI cases from sentinel GPs

Laboratory results

Emergency visits

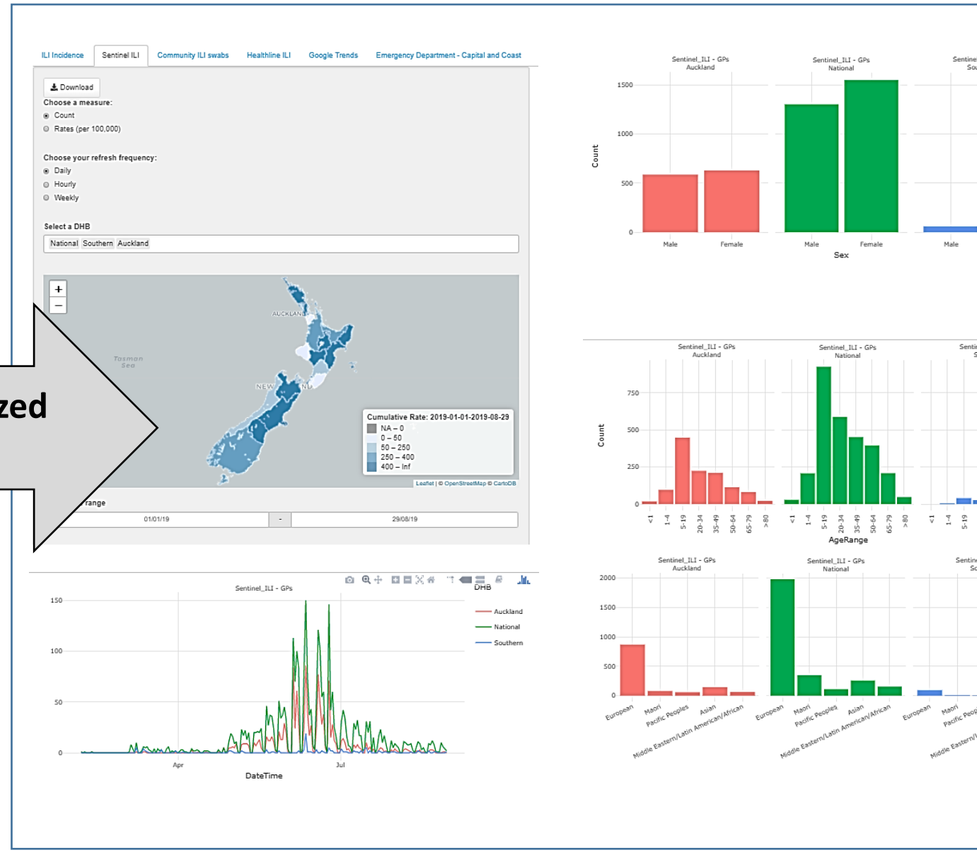
HealthLine

Google Trends

Automated scripts to retrieve data from feeds, appends in the ILI database

ILI Observatory Database

summarized data



Stats New Zealand data
- **Population**
- **Shape files**

Data Acquisition → Visualization

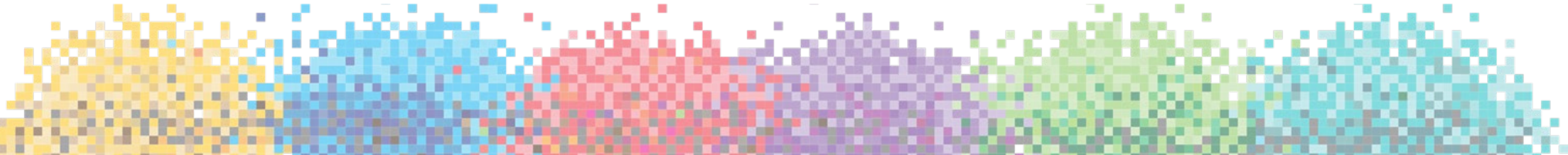
Science for Communities

E/S/R
Science for Communities

Data Science Pivot

September 2020

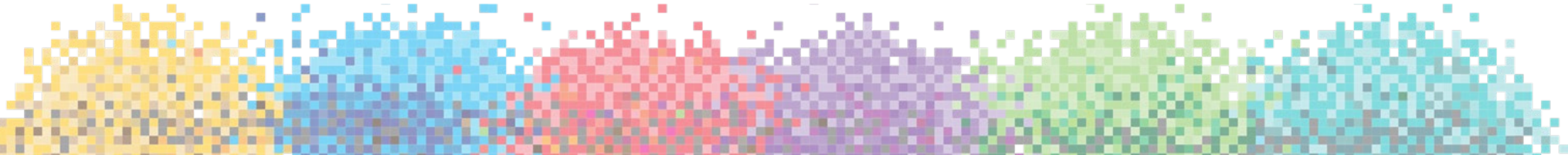
Richard Dean - Data Scientist



New plan – Primary goals

Science for Communities

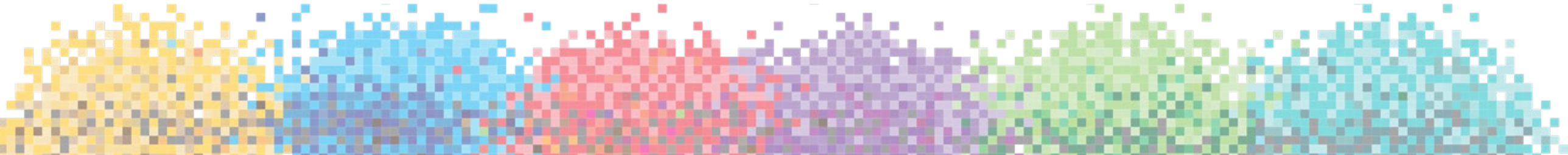
1. Run **Data Science Week** without a session on the IT roadmap
2. Deliver **Python software carpentry** during Data Science Week and repeat x2 in October-November
3. Run a cut down **accelerator cohort 3** from October-March – reduce pressure by extending timelines and only select essential projects 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

Science for Communities

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

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

Science for Communities