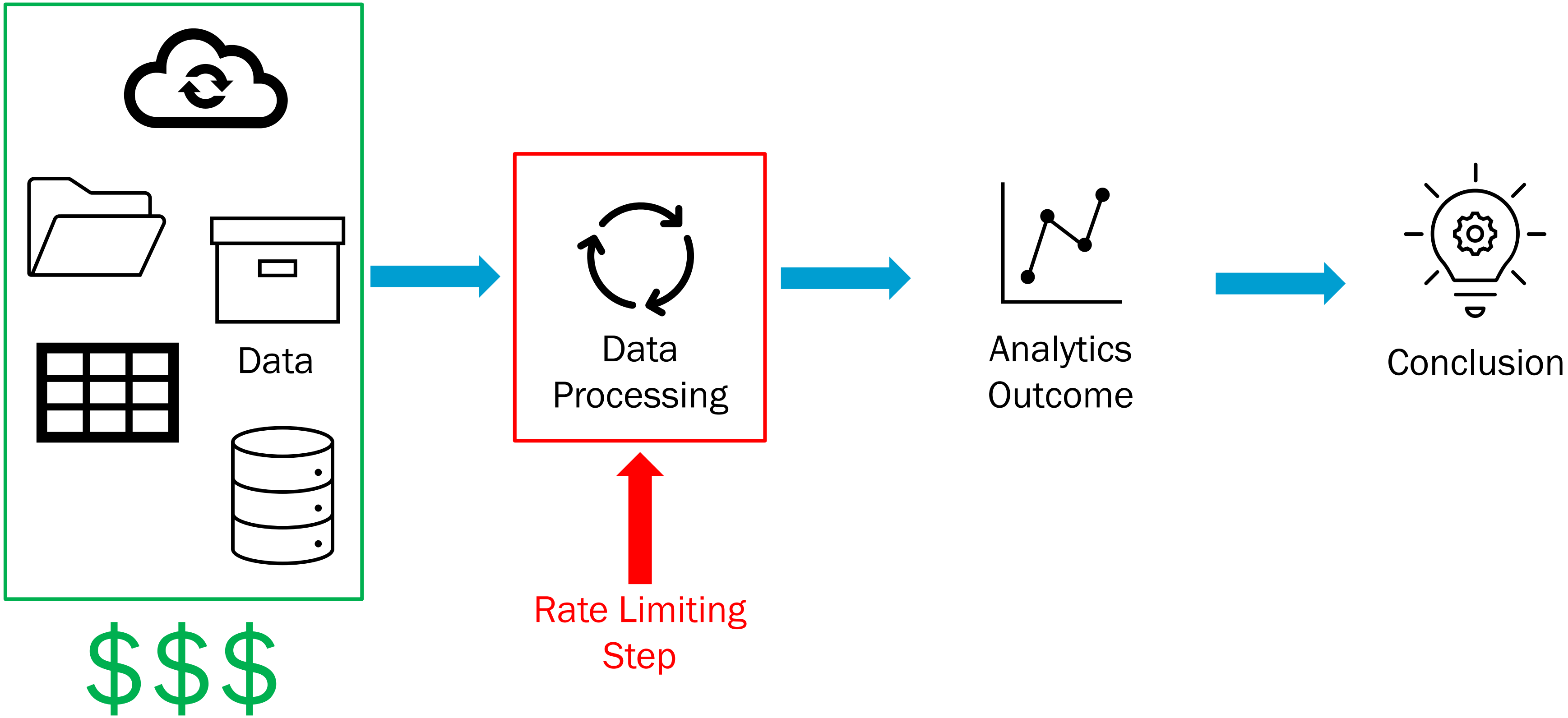
Brown AND Caldwell

Advancing Beyond Excel: Applying the R Software Environment for Water Quality Data Analysis

April 29, 2022

The Purpose of Data



Data

$$$

Data Processing

Rate Limiting Step

Analytics Outcome

Conclusion

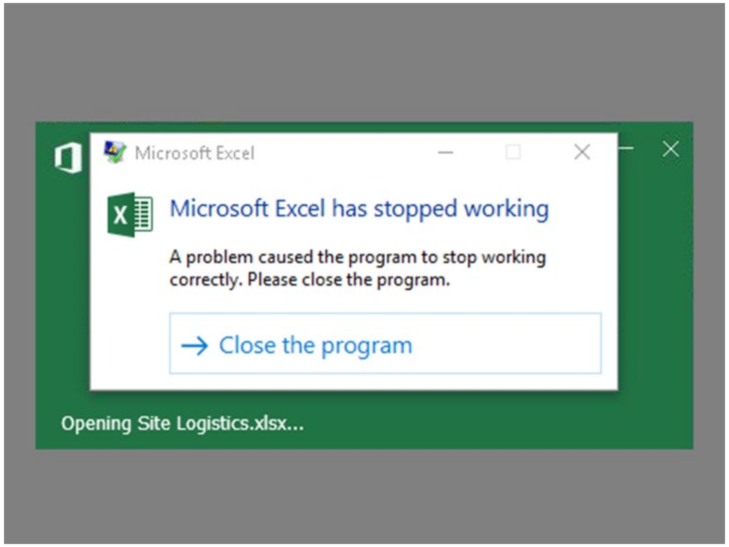
What are Excel’s Limitations?

What is Excel?

- Spreadsheet program introduced in 1987
- Organizes data into rows and columns
- Can perform calculations
  - Solver
  - Built in functions
  - Conditional formatting
- Can also provide graphic visualizations
- User friendly

Why can Excel be problematic?

- Default data management program on most computers
- Data limited
- Limited options for graphics
  - Formatting is time intensive
  - Plots have a limit for data display
- VBA helps, but doesn't fix everything



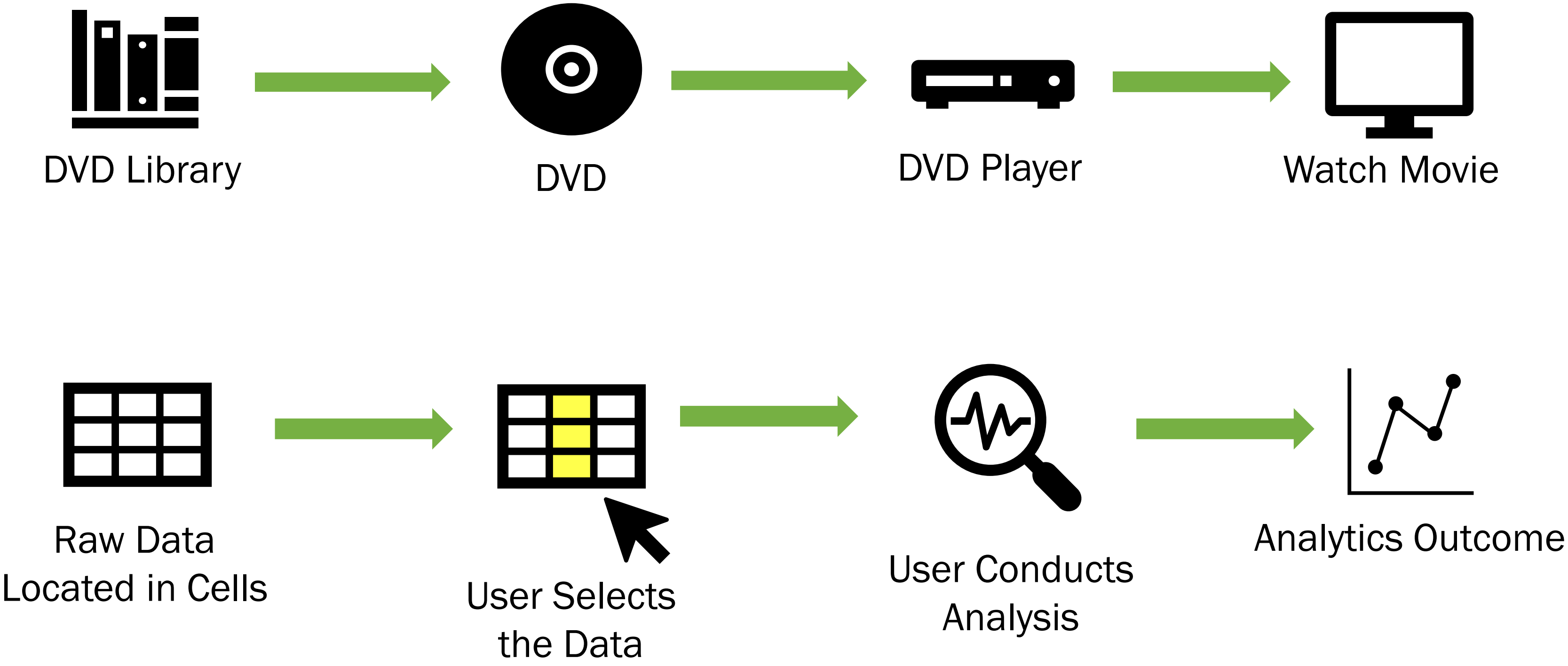
Microsoft Excel has stopped working

A problem caused the program to stop working correctly. Please close the program.

Opening Site Logistics.xlsx...

Excel’s Cell Structure

- Excel requires a location in the spreadsheet to manage data



DVD Library → DVD → DVD Player → Watch Movie

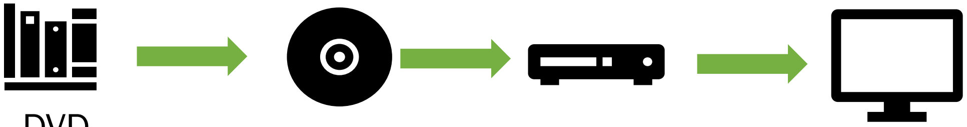
Raw Data Located in Cells → User Selects the Data → User Conducts Analysis → Analytics Outcome

Excel Cell Structure Continued

DVD put back into the wrong case → Wrong cell referenced

DVD gets scratched → Your raw data is deleted or changed

You must get off the couch to get the DVD → The user is the engine for conducting all the analysis



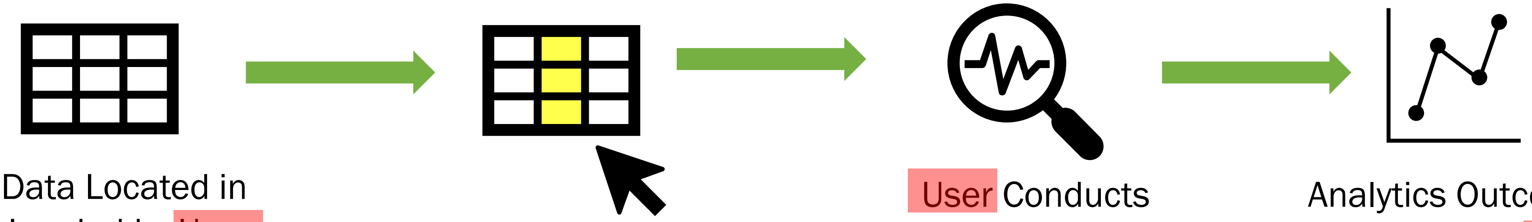
DVD Library → DVD → DVD Player → Watch Movie

Introduction to Data Analytics

What does an alternative solution look like?

- Scripted solutions
- Python, R, Matlab etc.
- It does not particularly matter what software you use
- Some key items to consider:
  - Open source?
  - Are their packages already programed?

What makes R different than Excel?

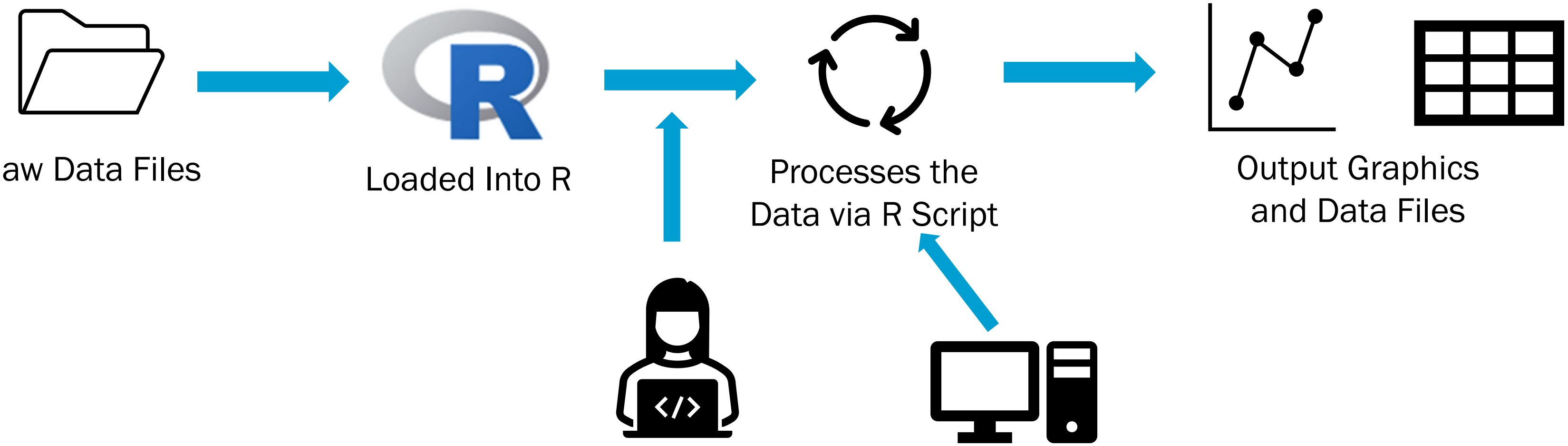


Raw Data Located in Cells Loaded by User

User Selects the Data

User Conducts Analysis

Analytics Outcome Formatting by User



Raw Data Files

Loaded Into R

Processes the Data via R Script

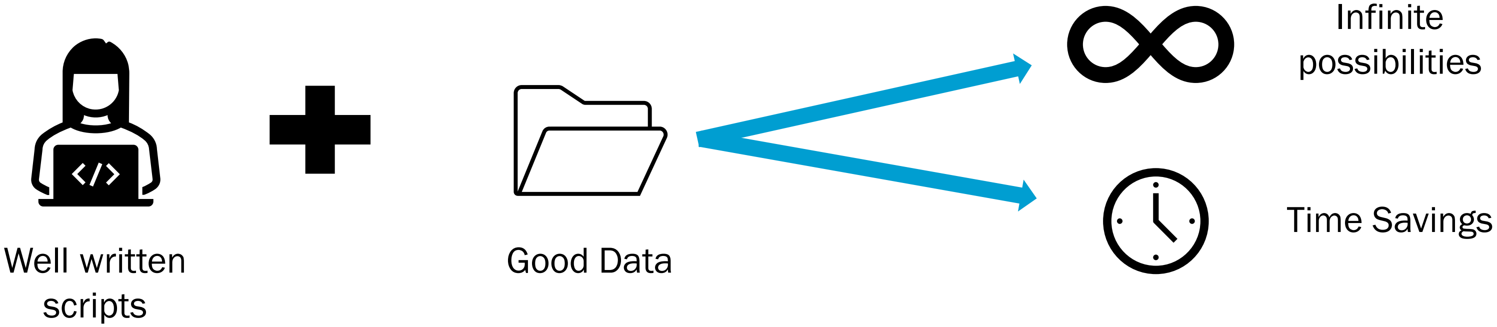
Output Graphics and Data Files

Script written by User

Computer Does the Work

Some Scripting Advantages

- Process is repeatable
- Easy to organize inputs and outputs
- Multiple data sources
- Data frames are dynamic
- Data can be “created” then “destroyed”
- Data storage
- Well written scripts provide a step-by-step guide



Well written scripts

Good Data

Infinite possibilities

Time Savings

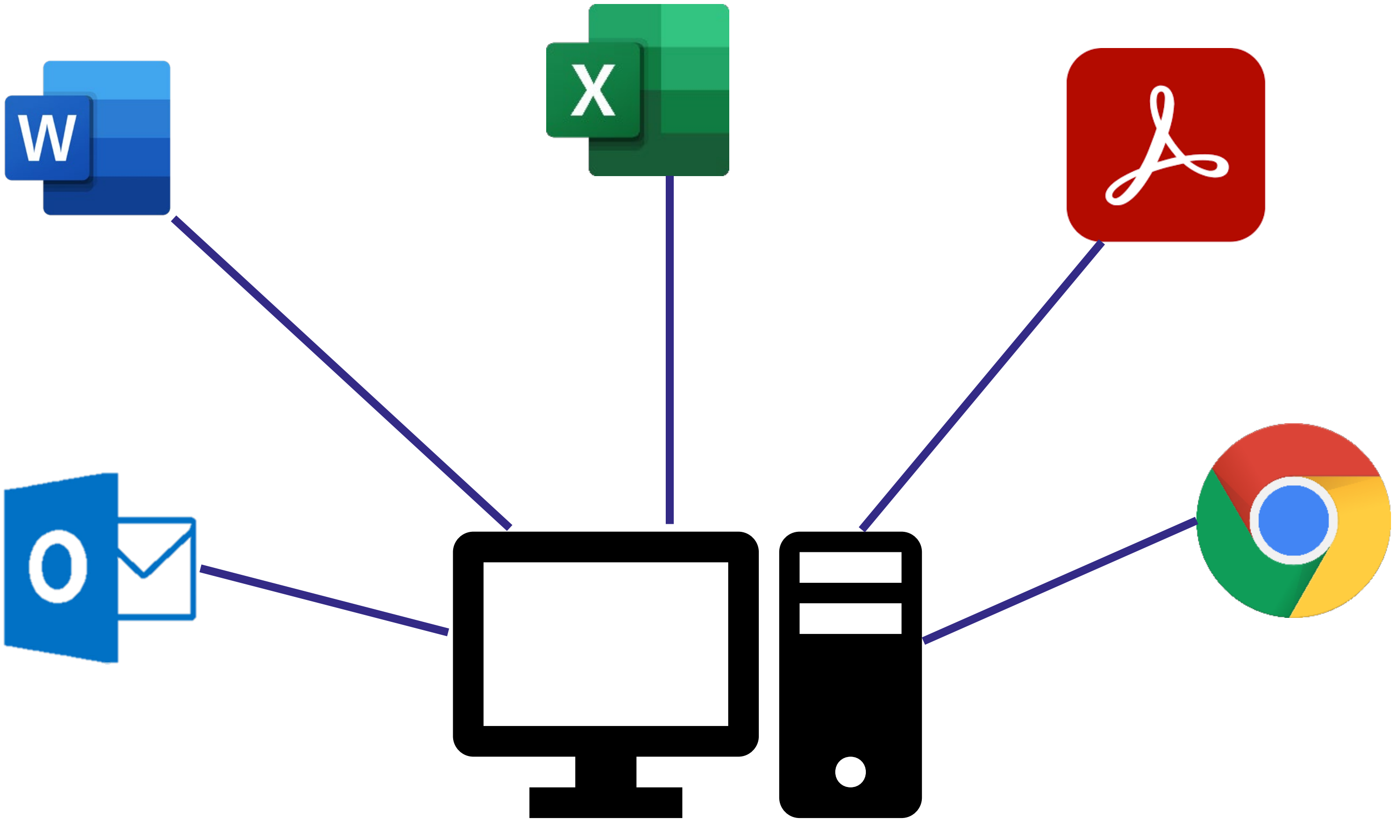
My Scripting Background

- Background in civil and environmental engineering
- I do not have any formal data analytics or computer software training
- My programming experience was basic
- I am proof of where determination, lots of internet searches, and asking “what if?” can get you

Be-leaf in yourself

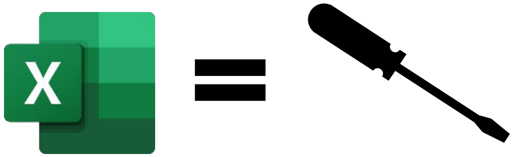
When should I use R?

Your Engineering Toolbox



The Computer is the Engineer’s Toolbox

The Right Tool for the Job



- Is this the only time I will need to do this?
- Is my data set limited?
- Can the analysis be easily done manually without a lot of effort?

- Is this going to be a repeated task?
- Is there a lot of data?
- Is the analysis going to require a lot of manual effort?

Is this a screwdriver task or a power drill task?

Data Analysis Case Studies

Water Quality Analysis Survey

- Timeseries data for 10+ years of data for 20+ water quality parameters
- All the data had been presented in timeseries plots, but the client then requested boxplot representations

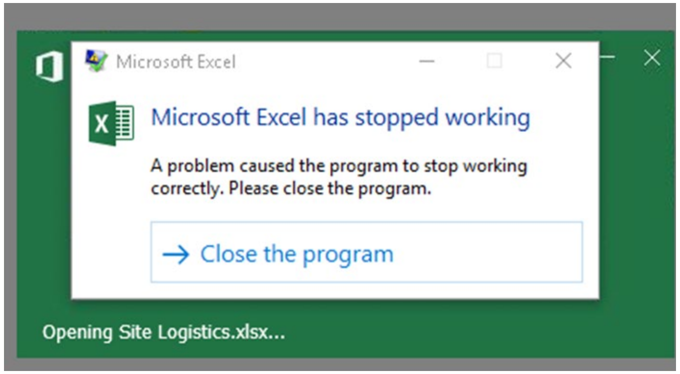
Seems like a power drill task!

1. Coded a boxplot template in R
2. Created a loop to cycle through each water quality parameter
3. Output 40 plots in less than 1 minute of processing
4. Bonus: The formatting was automatically consistent across each plot!

What About Pivot Tables?

Analysis of plant operating data for solids production calculation:

- 11 years of hourly plant SCADA data were provided
  - Raw data file size: ~38 MB
- Pivot Table made Excel unusable, processed data file size: ~390 MB



Microsoft Excel

Microsoft Excel has stopped working

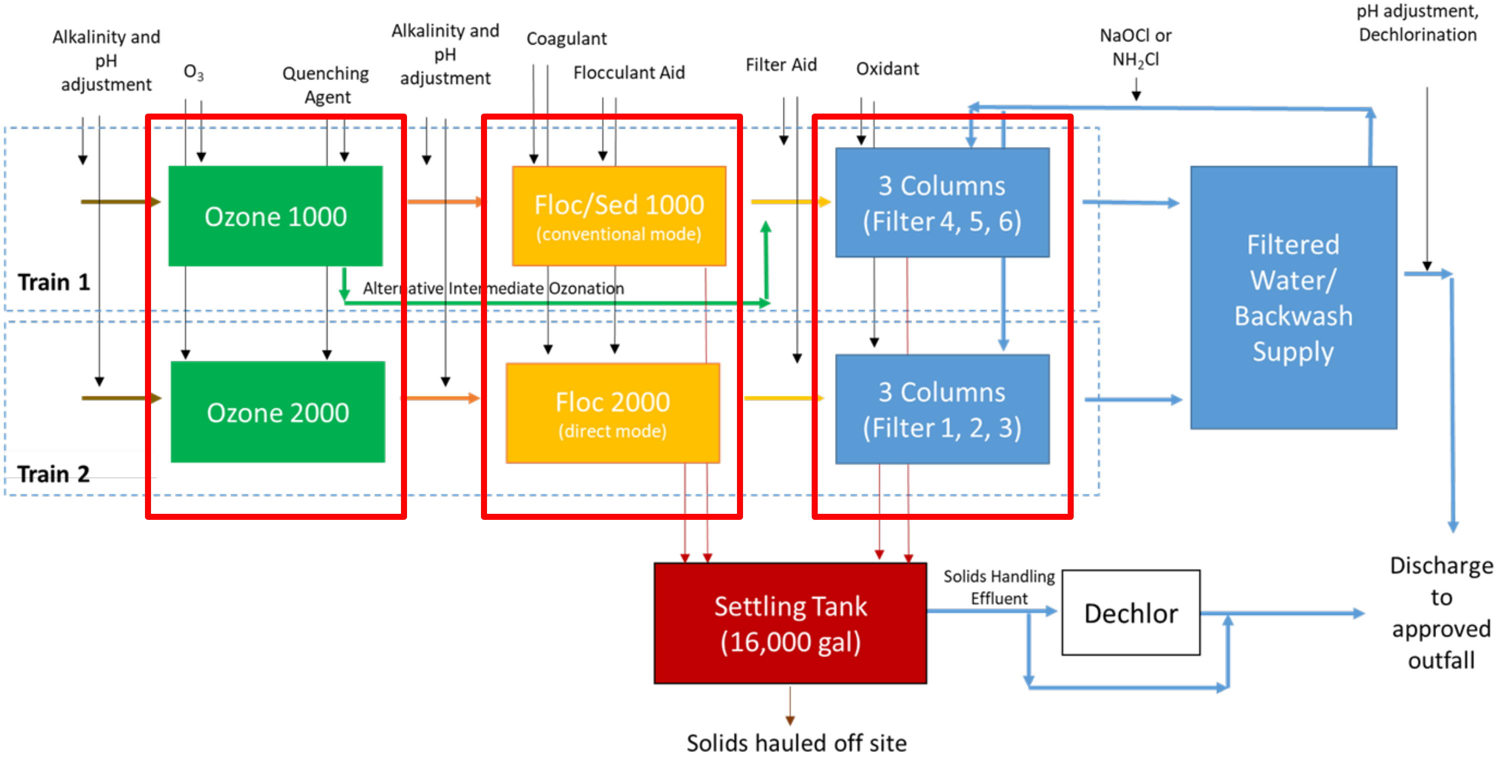
A problem caused the program to stop working correctly. Please close the program.

Opening Site Logistics.xlsx...

- Leveraged R’s ability to handle dates and manage dynamic data frames
- Logic of a Pivot Table can be controlled by the user instead of relying on a pre-programed function
  - R script size: ~4 KB
  - Output file size: ~1 MB

PWB Filtration Pilot Project

- Filtration Pilot
  - Pilot study period: 7/1/2019 to 6/30/2020
  - 2 Treatment Trains


Alkalinity and pH adjustment
O_3
Quenching Agent
Alkalinity and pH adjustment
Coagulant
Flocculant Aid
Filter Aid
Oxidant
NaOCl or NH_2Cl
pH adjustment, Dechlorination
Train 1
Ozone 1000
Floc/Sed 1000 (conventional mode)
3 Columns (Filter 4, 5, 6)
Filtered Water/ Backwash Supply
Alternative Intermediate Ozonation
Train 2
Ozone 2000
Floc 2000 (direct mode)
3 Columns (Filter 1, 2, 3)
Settling Tank (16,000 gal)
Solids Handling Effluent
Dechlor
Discharge to approved outfall
Solids hauled off site

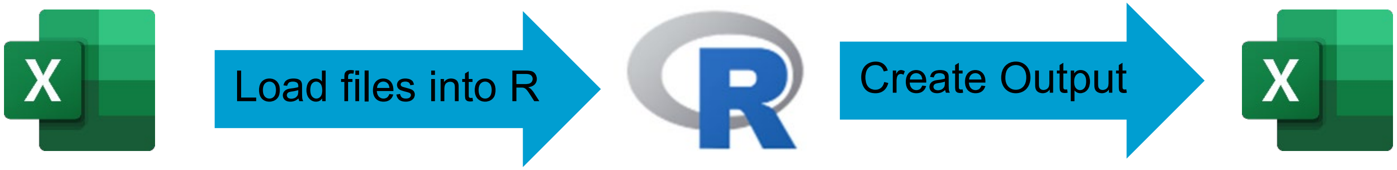
Piloting – There’s a LOT of Data!

- Data collected every 5 minutes
  - Approximately 5.18 million data generated for *one* filter
- Filter columns programmed to backwash at a setpoint beyond the piloting goal to compare media designs
- Operational challenges (i.e., chem feed pump delivery loss, etc.) occasionally resulted in poor data quality

<u>Key Points</u>

- Needed consistent way to clean data
- Needed data to be objective
- Large quantity of data

Data Cleaning Workflow:
Defining Unit Filter Run Volumes



Load files into R

Create Output

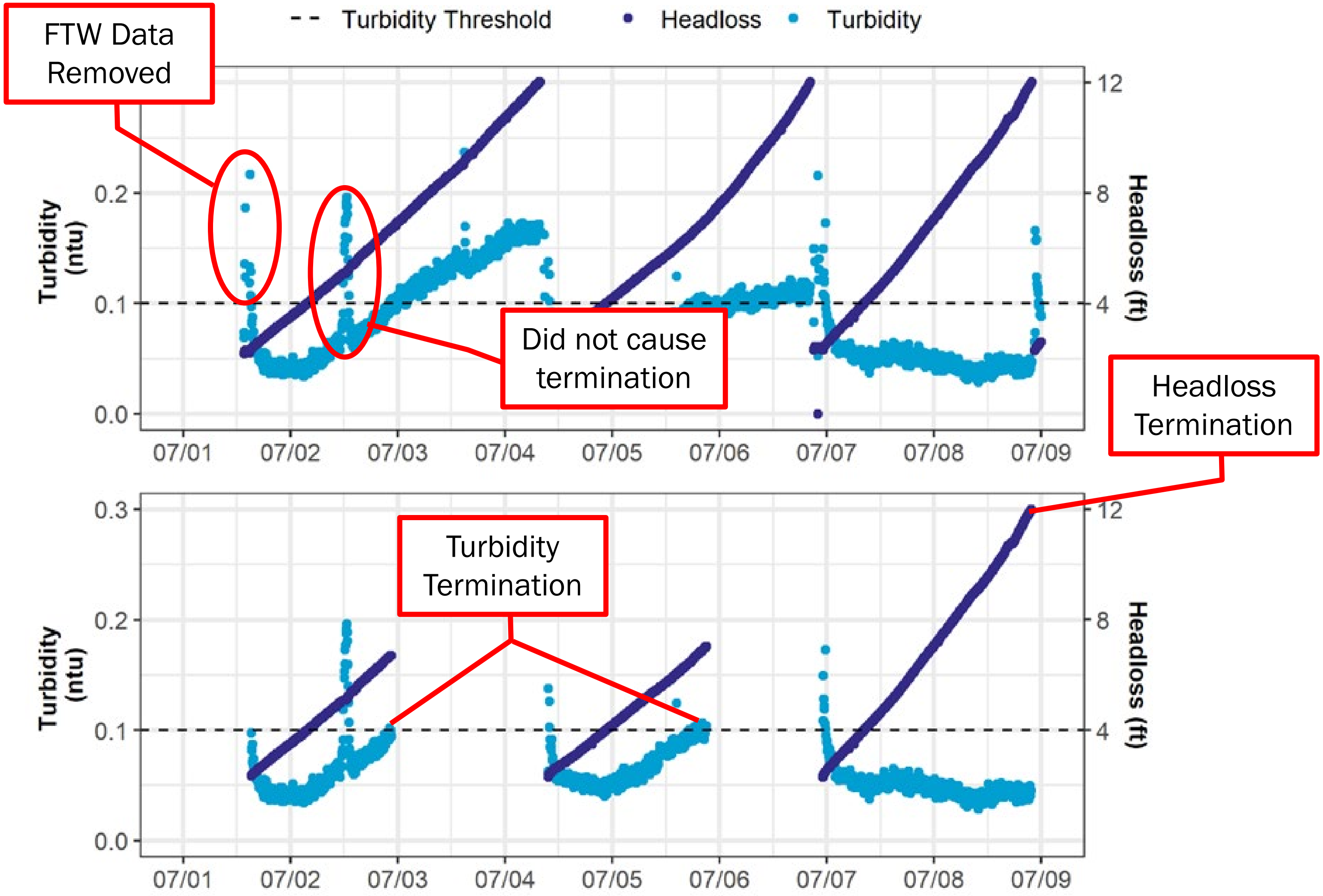
Filter Operation
CSV Files

UFRV Creation
File

UFRV Operations
CSV File

Raw Data

Cleaned Data



FTW Data Removed

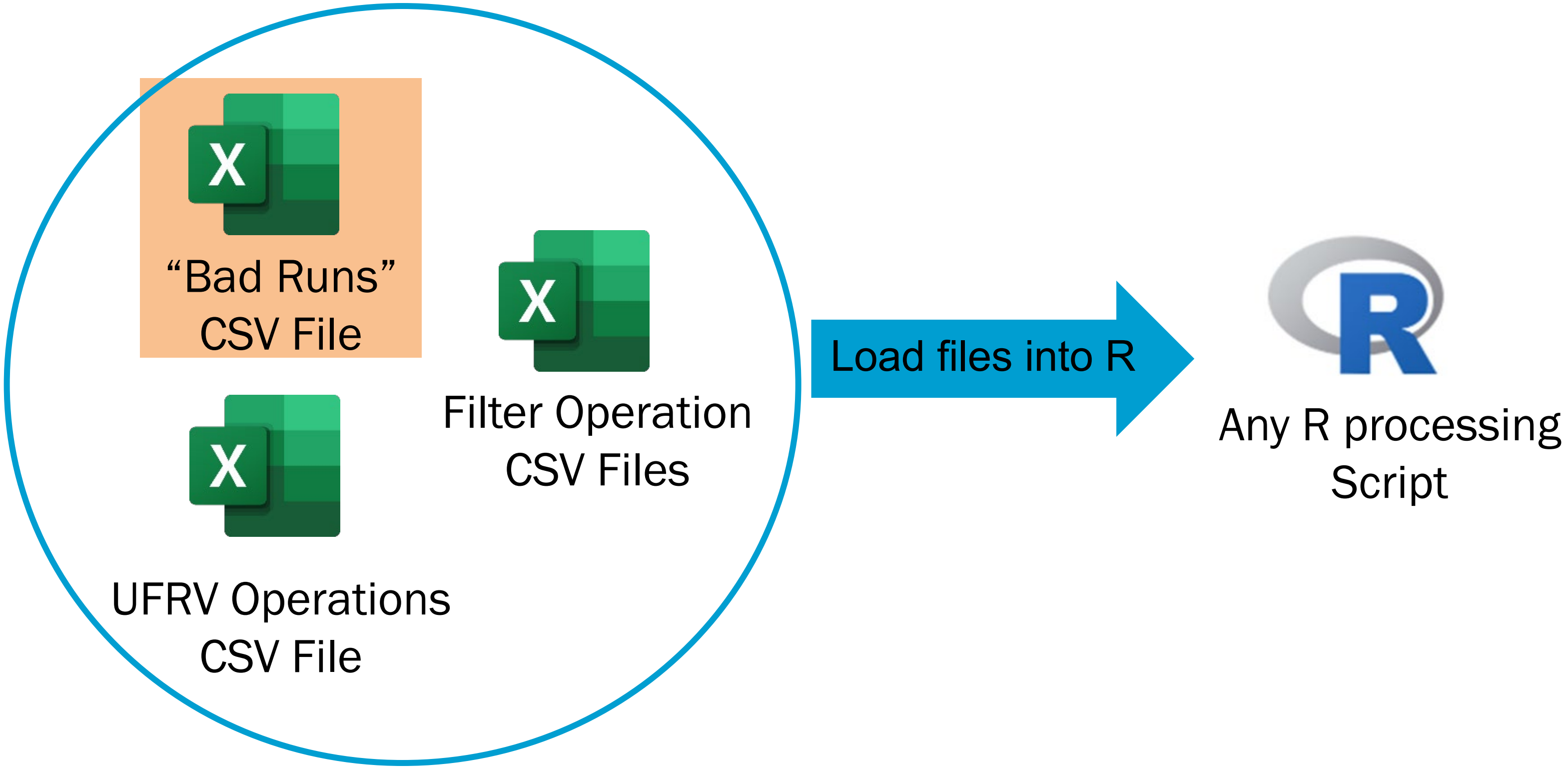
Did not cause termination

Headloss Termination

Turbidity Termination

Data Cleaning Workflow:
Removing Data due to Operational Issues

- Remove entire filter runs that were not representative of piloting performance



“Bad Runs” CSV File

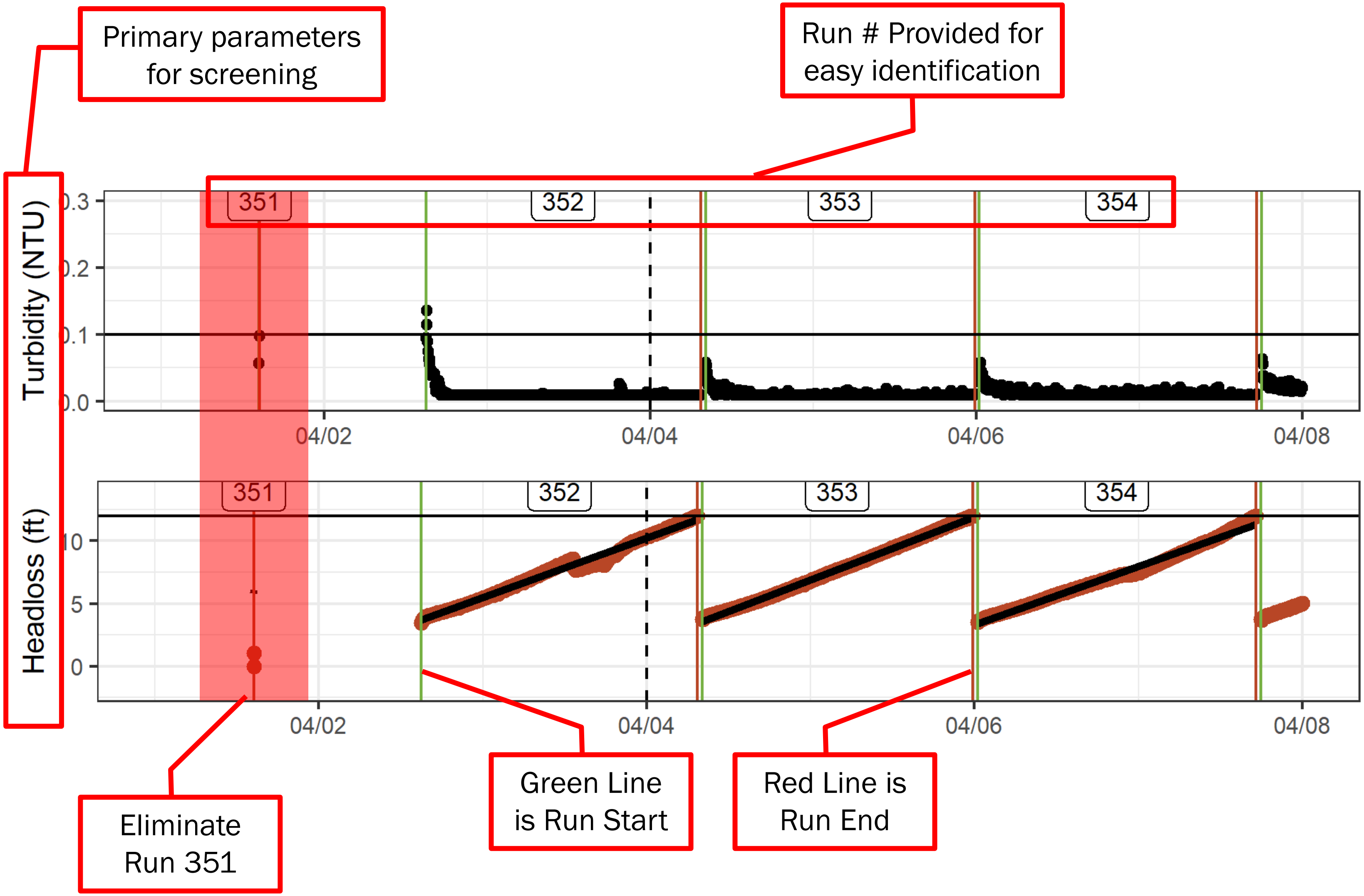
Filter Operation CSV Files

UFRV Operations CSV File

Load files into R

Any R processing Script

Weekly Data Processing



Primary parameters for screening

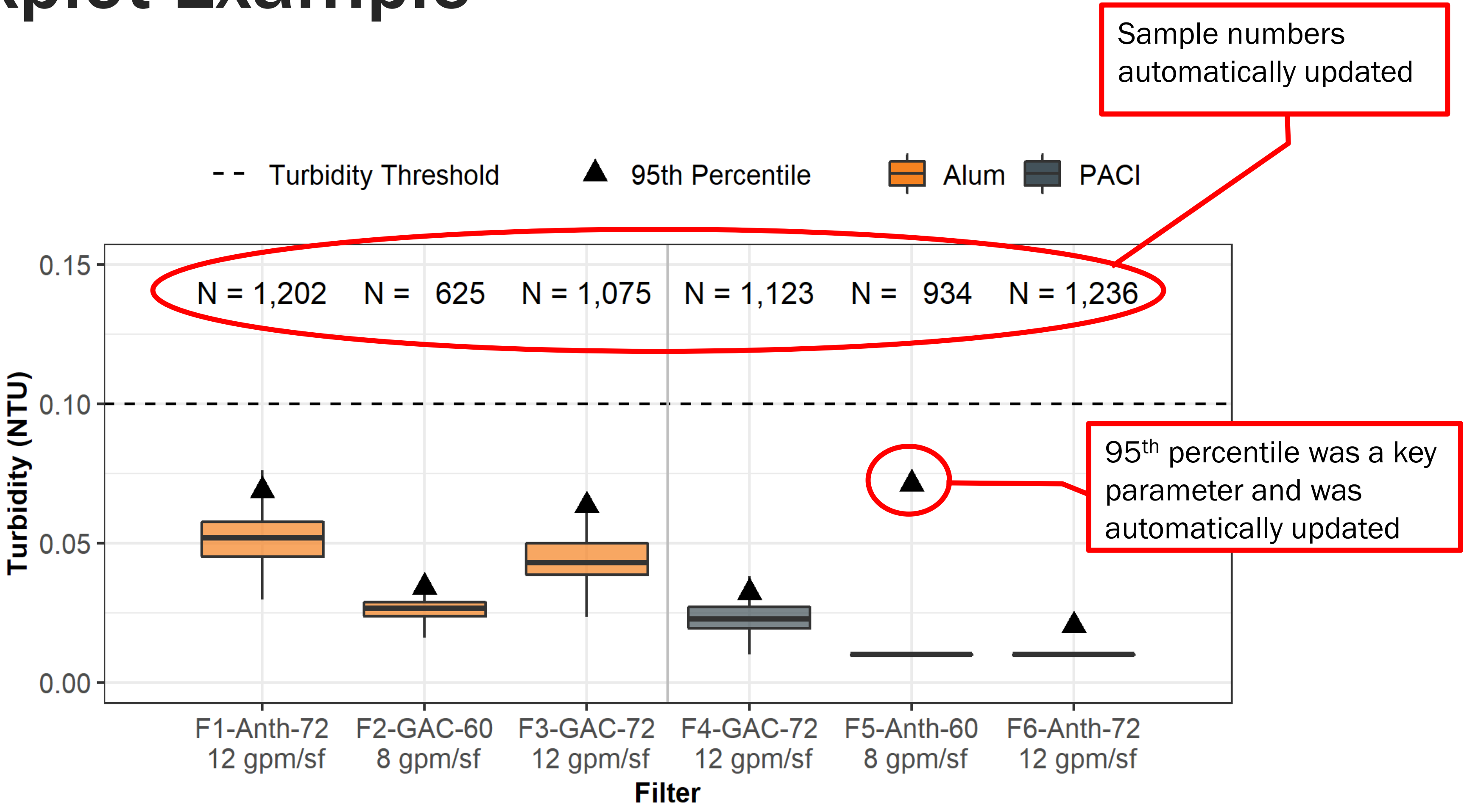
Run # Provided for easy identification

Eliminate Run 351

Green Line is Run Start

Red Line is Run End

Boxplot Example



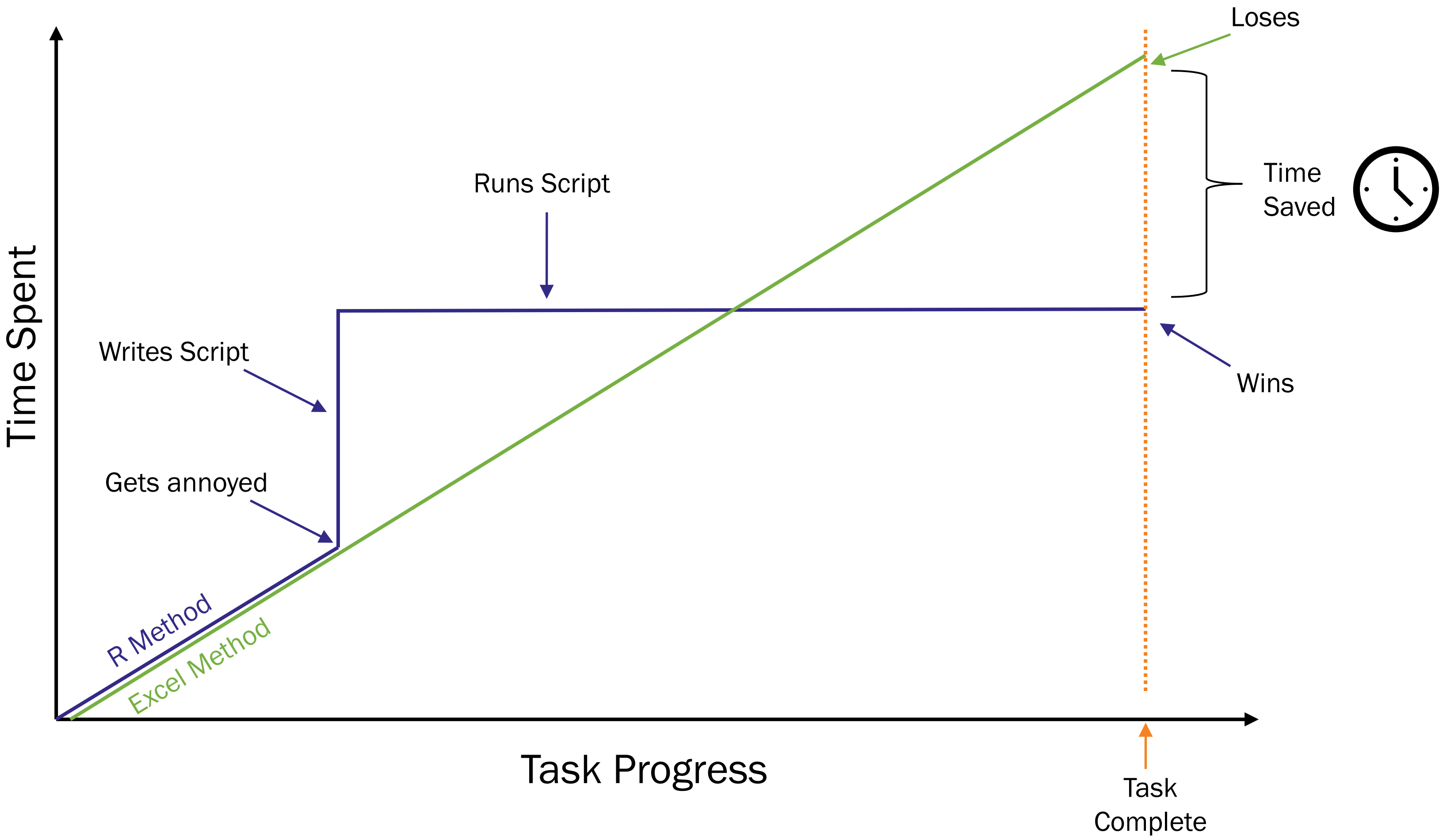
Sample numbers automatically updated

95th percentile was a key parameter and was automatically updated

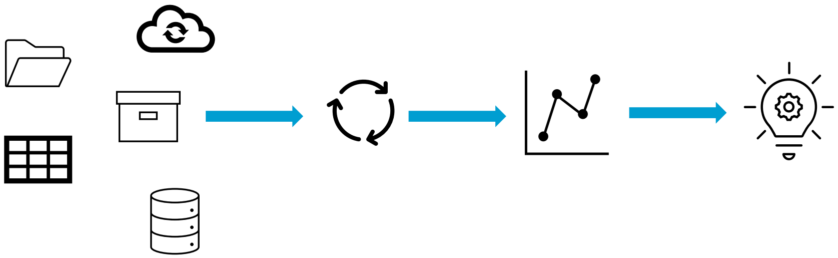
Filter effluent turbidities recorded during accepted filter runs during the side-by-side testing of alum and PACl with filter aid, August 5-12, 2019

Conclusions

The Reality of Scripting



Conclusions



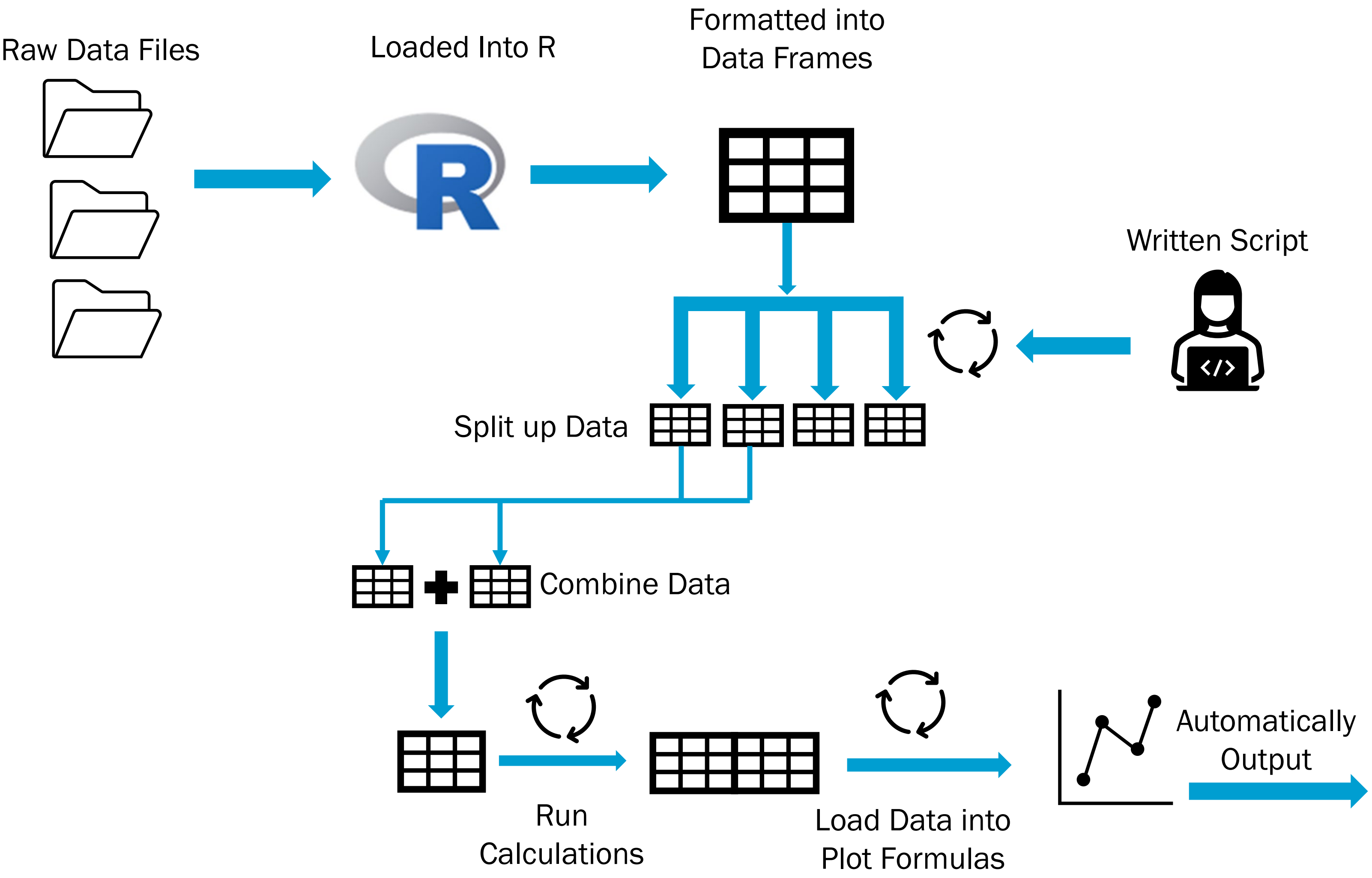
- Big data is here to stay
- More data does not necessarily lead to more conclusions
- Use the right tool for the right task
- Value in engineers with no data analytics background being able to take on data tasks

Call to action!

Learn to script! I promise it’s fun!

Questions?

Scripting Has Fewer Limits



Raw Data Files

Loaded Into R

Formatted into Data Frames

Written Script

Split up Data

Combine Data

Run Calculations

Load Data into Plot Formulas

Automatically Output