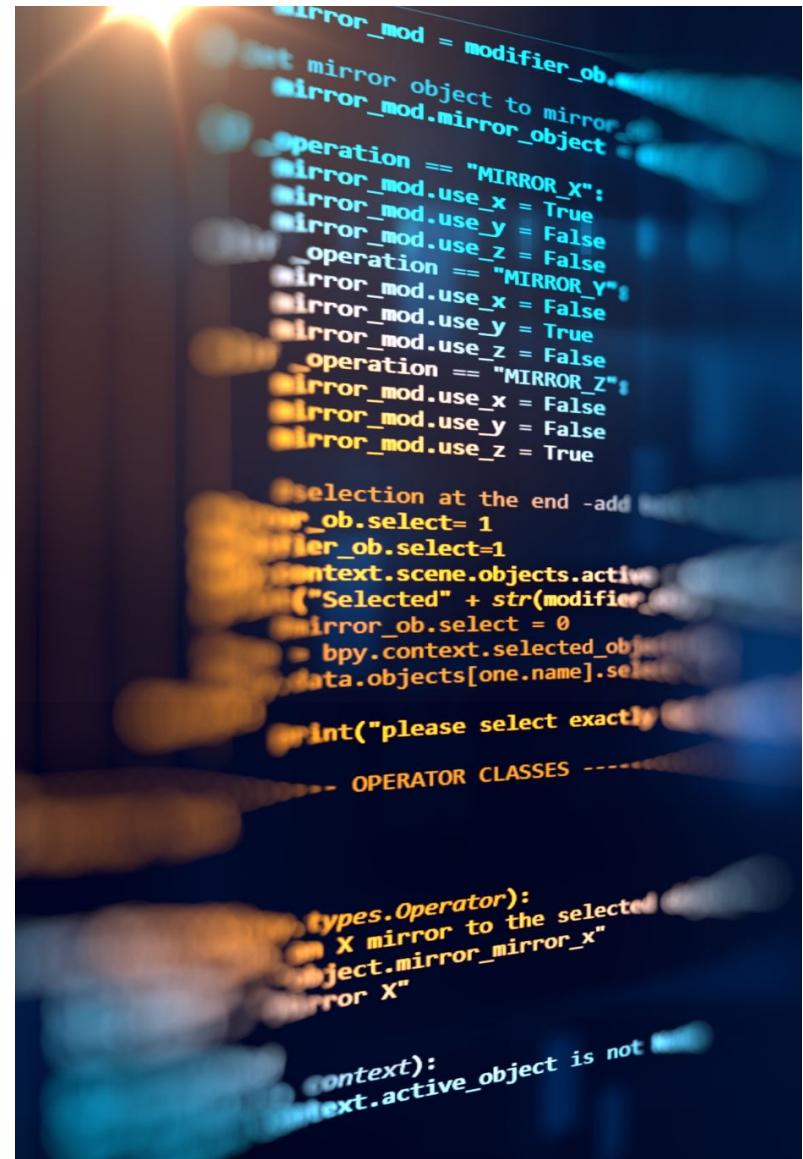


DATA TECHNIQUES FOR ENGINEERS AND SCIENTISTS

J.D. LANDGREBE
DEPT. OF CHEMICAL ENGINEERING
UNIVERSITY OF DELAWARE

SOFTWARE FOR ENGINEERS

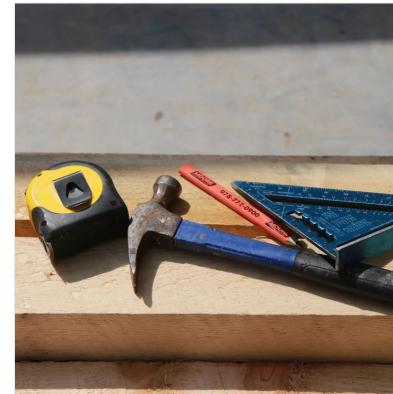
- Pick software by “use case” and use each software in its “lane” where it is best tool for the job
- “Use cases” is a useful way of compartmentalizing tasks or activities in software
- Use cases start with “I need to...” (or “my customer needs to...”)
 - My supervisor needs to be able to scroll through our raw data to understand the results
 - I need to create a correlation matrix for my raw material’s quality data
 - I need to build a validated calculation model for reaction kinetics
- Use case focus
 - Avoids getting (yourself and others) distracted by things that software X is not good at – there is no perfect software tool
 - Puts a premium on making data portable across applications



GENERALIST ENGINEER/SCIENTIST SOFTWARE RECS BY USE CASE

- **JMP® software** - design of experiments (DOE), exploratory analysis and data visualization with accompanying statistical analysis
- **Microsoft Excel®** - make data usable by others, generate “end of the pipeline” reports and create spreadsheet models with calculations for use by yourself or non-coders
- **Python scripts** for data reshaping and for developing coded models and data pipelines (possibly mixing in a little SQL)

Blog



The Right Modeling And Analysis Tools For the Job

Software tools in the data and modeling arena often lead individuals and teams into counterproductive patterns. By being informed and intentional, you can choose the best tool for a particular job. It's good to recognize that software providers, meaning companies and open-source communities, keep...

[READ MORE](#)

<https://datadelveengineer.com/the-right-modeling-and-analysis-tools-for-the-job/>

ADDITIONAL SOFTWARE RECOMMENDATIONS

- Python scripting tools - build model and data cleaning scripts
 - JupyterLab (Anaconda)
 - VS Code script and text editor
 - Github Copilot and Chat extensions VS Code (\$\$)
- VBA Scripting Language – Excel automation
- Advanced visualizations – Python Matplotlib
- Github - Open sharing of projects, trainings etc.

PYTHON LIBRARIES TO EXTEND FUNCTIONALITY

Useful Python libraries for Engineers and Scientists

You can get these with Anaconda installation!!

- [Pandas](#) – Data reshaping and analysis
- [Numpy](#) – Numerical calculations
- [Pytest](#) – Test/Validate your scripts
- [OpenPyXL](#) – Format outputs in Excel
- [Re \(Regular Expressions aka Regex\)](#) – Parse text strings



regex 2023.10.3

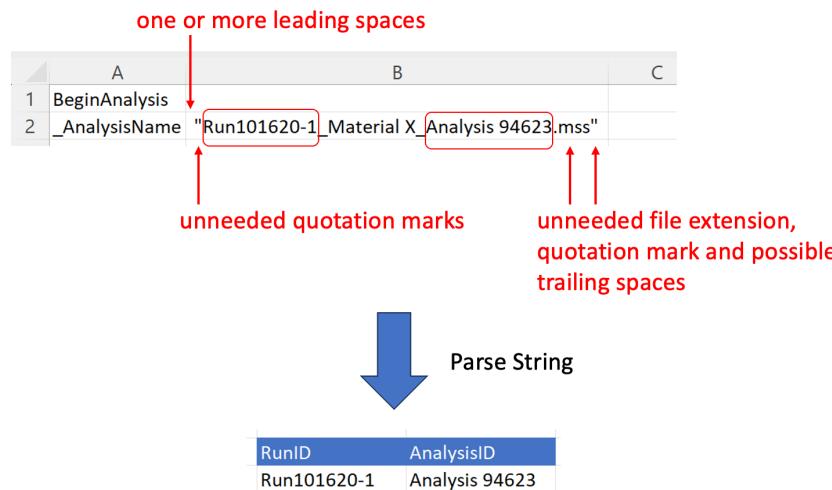
EXAMPLE: PYTHON REGEX PACKAGE TO EXTRACT STRINGS

regex 2023.10.3

Lab instrument software outputs IDs in string needing cleanup

regex Python package can strip out unneeded quotation marks, spaces and (not shown) file extension

(Not shown) Either regex or Python split() command can break the string into pieces using underscore characters as delimiters



```
libs > curve_parse.py > ...
1  #Version 10/5/23
2  import pandas as pd
3  import numpy as np
4  import regex as re
5  import os

174
175
176
177
178
179
180
181
182
183
184
185
186
187
188

def id_string_cleanup(self, id_string):
    """
    Strip leading/trailing spaces and quotes from an ID string
    JDL 10/5/23
    """

    # Define regular expression patterns
    leading_space_pattern = r'^\s+'           ← Regex code for "zero
    trailing_space_pattern = r'\s+$'           ← or more spaces at
                                                beginning of string"
    quote_pattern = r'"|"'                   ("ChatGPT pls help!!")

    # Use regular expression to strip leading/trailing spaces and quotes
    id_string = re.sub(leading_space_pattern, '', id_string)
    id_string = re.sub(trailing_space_pattern, '', id_string)
    id_string = re.sub(quote_pattern, '', id_string)

    return id_string
---
```

INSTALLING PACKAGES (AKA LIBRARIES) IN ANACONDA

The image shows a step-by-step process for installing packages in Anaconda Navigator:

- Step 1:** In the "Installed" environment, a search for "regex" finds no results, indicated by the message: **** regex is not Installed in environment named “Latest” ****.
- Step 2:** In the "Not installed" environment, the "regex" package is selected.
- Step 3:** The "Install Packages" dialog shows the selection of "regex" and its dependencies: `*openssl` and `*ca-certificates`.

Code Snippet (curve_parse.py):

```
curve_parse.py U x
libs > curve_parse.py > TensileParsingRun
1 #Version 10/5/23
2 import pandas as pd
3 import numpy as np
4 import regex as re
5 import os
```

Note: code's import statement will error if package is not installed in Anaconda environment etc.

Anaconda selects related (aka “dependent”) packages needed by the selected package



EXERCISE WITH PANDAS OR PYTHON REGEX

Exercise 1

- Have students ask ChatGPT for a regex to parse a particular string
- Have students try it in ChatGPT or Copilot
- Try their result on some data

Exercise 2

- Ask students/groups to clone Intro_To_Pandas_For_Noncoders
- Run notebook with imported data
- Ask summary question(s) – How many rows in x subset?



OBJECTIVES

WORKING WITH EXPERIMENT DATA

Experiments* have a common data structure (we will call it “architecture”).

Recognizing this lets us master common data techniques to analyze efficiently and have data in good formats for graphing

Objectives

- Learn reusable terms for the data elements from experiments
 - Learn how use software to do needed transformations to go from raw data to analyzed summary
 - Learn how to “curate” experimental data to make it easy to find and share

* Data don't know whether they were generated in a lab or virtual experiment, so this applies to computer modeling data in addition to physical experiments

WHAT ARE EXPERIMENTS?

Experiments Cover An amazing range for engineers and scientists!

- Make a production or lab-scale batch of several product formulations
- Run a packing line using pre-set conditions for a set amount of time
- Crash a sensor-equipped car into a wall
- Diaper a baby with several product designs
- Drive a F1 car around a track with several various fuel mixtures
- Ask a man to shave his face with 3 different types of razors on consecutive days
- Make batches of cookies with different types of chocolate
- Wash test, fabric swatches with different detergent formulations
- Use different catalysts to run a chemical conversion with the same starting materials
- Use PCR to replicate and sequence the DNA from multiple virus samples
- Mop floors with mops using different cleaning solutions
- Surgically implant artificial joints manufactured with different polymeric coatings

WHAT DO THESE HAVE IN COMMON?

Experiments (or "studies"...we will use synonymously) are a building block for designing and confirming products.

Experiments have "runs" consisting of one or more conditions having controlled inputs such as the starting materials and test conditions

A run can be described by a list of the "run variable" values which are the pre-set inputs and important ambient conditions recorded during the run.

Experiments generate "raw data" consisting of the original measurements to assess what happened (detailed speed versus time data for F1 car circuit around a track)

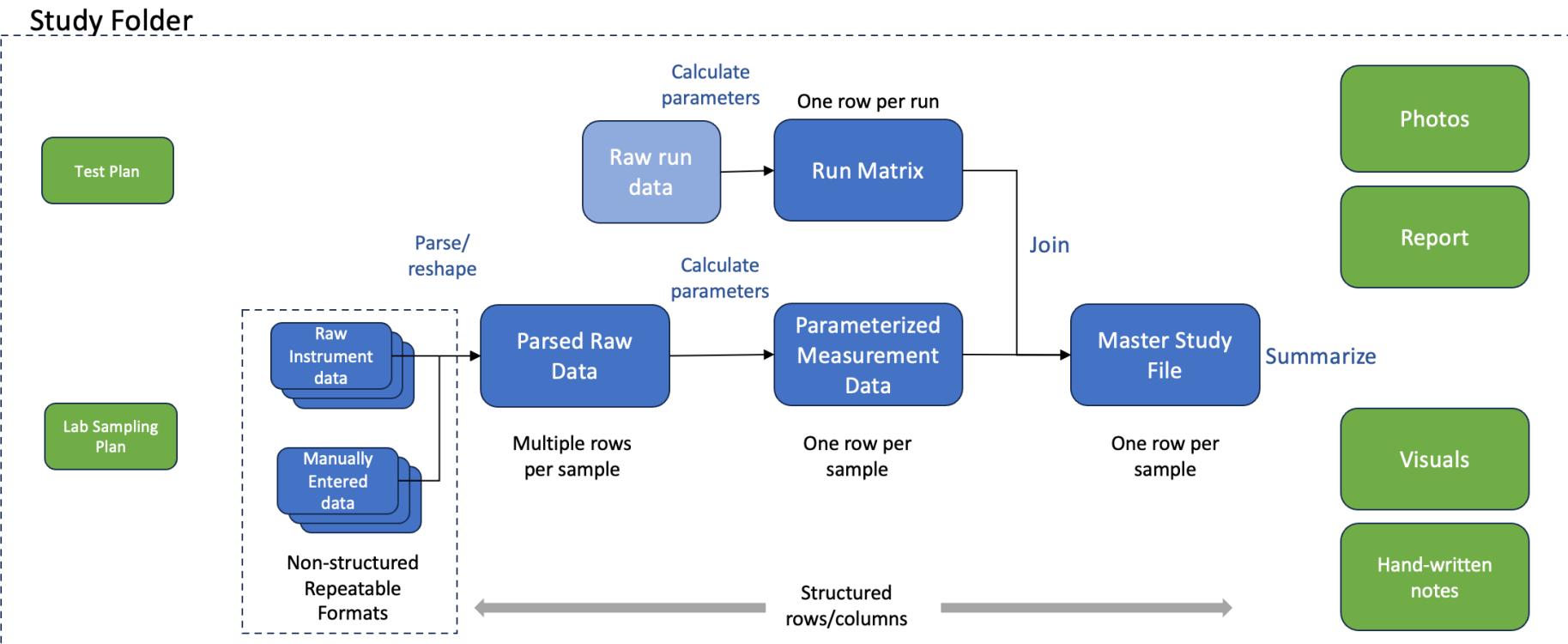
The raw data gets converted to "parameterized data" which is quantities reflecting the outcome (average speed for a complete track circuit)

Run variable values can be assembled into a "run matrix" with a single row of values for each run

The blackboard contains several mathematical notes and diagrams:

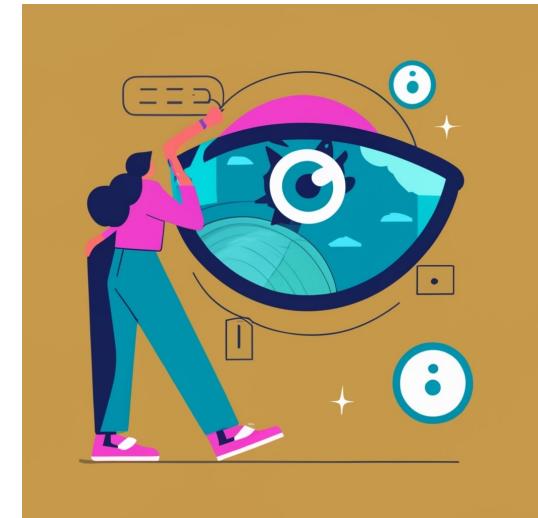
- At the top left, there's a complex fraction involving summations and a square root, with annotations like "h=0" and "sqrt(2434.96) = 49.3".
- To the right of that, a diagram shows a rectangle with a shaded area and a value of 5.4.
- Moving down, there's a formula $\sqrt{a^2 + b^2} = x^2$ with a circled "22" next to it.
- Below that is a circle with a shaded area labeled "c".
- Further down, there's a system of equations involving x, y, c_x, c_y, π , and a note " $\pi = C$ ".
- On the right side, there's a diagram of a circle with a radius and a point labeled "S".
- At the bottom left, there's a formula $u = 584. + n^{av}$ followed by a note " $x^2 + 34,$ ".
- On the right, there's a large bracketed expression involving $u=14!$, $\sum N_{50} \cdot x - \frac{1}{2} [984 + x]$, and $x \leq 5$.
- At the bottom right, there's a diagram of a circle with a radius and a point labeled "B", with the formula $B = 9 + x^2 + y$ written nearby.

EXPERIMENT DATA “ARCHITECTURE”



WHY DOES STANDARD STUDY NOMENCLATURE MATTER?

- Reusable "lens" that you can use in any context
- Knowing nomenclature and architecture gives you a "seeing eye" for how to plan and organize experiments
- Identifies efficient, generic ways of converting the Run Matrix and Raw Data into analyzed data and conclusions
 - Parsers to convert raw data to structured tables
 - "Join" and "Concatenate" to bring data together into a "Master Study File"





EXERCISE WITH JMP

Data Techniques exercise with JMP tables menu for reshaping typical data – Join, Concatenate, Stack, Split

- Break into groups
- Ask them to work the exercises individually + together and compare notes
- Study questions: Share with your group an interesting dialog box option you found
- JDL demo



PARSING RAW DATA

- “Parsing” refers to going from a repeatable non-structured format to structured rows/columns
- Parsing needs to identify samples by its RunID, SampleID (e.g. sampling time/location) and AnalysisID metadata for the result to be considered “structured” (e.g. each sample’s data uniquely identified and traceable back to its source)
- Study owner needs to pre-plan/intervene with lab to ensure metadata are available/parseable post-measurement!!!
- Parsing requirements vary [wildly!] depending on data source
- Manually-entered data are a special case because humans are very creative about how they choose to enter data
- Lab instruments typically have specialized software that outputs raw and/or parameterized data

RAW DATA PARSING EXAMPLE DATA FROM MTS TENSILE TESTER SOFTWARE

- Testing involves pulling apart substrate(s) such as hook and loop fastener or adhesive-coated strip + landing surface
- Measurement of force versus displacement while separating the materials – peak load and average load are key results
- Requires multiple samples to get good data: Testing is noisy due to sample prep of joining substrates and general noise from failure testing



DuraGrip® Brand Adhesive
Backed Hook and Loop Fasteners



VELCRO® Brand Adhesive
Backed Hook and Loop Fasteners

RAW DATA PARSING EXAMPLE DATA FROM MTS TENSILE TESTER SOFTWARE

- Instrument software outputs files containing data for multiple samples from a run
- Raw files contain a combination of calculated parameters and raw data – we want to analyze both!!

A	B	C	D	E	F	G
1	BeginAnalysis					
2	_AnalysisName	"Run101620-1_Material X_Analysis 94623.mss"				
3	_MethodName	"Method 123468_00_Peel Strength.msm"				
4	InitialSpeed	305	"mm/min"			
5	DataRate	50	"Hz"			
6	GageLength	50	"mm"			
7	BeginSample					
8	AverageLoad	0.91	"N"			
9	PeakLoad	2.58	"N"			
10	BeginData					
11	_Load	SlackExt				
12	N	"mm"				
13	0	0				
14	0	0.001				
15	0	0.009				
16	0	0.036				
17	0	0.072				
18	0.02	0.124				
19	0.06	0.189				
457	0.16	44.68				
458	0.13	44.782				
459	0.09	44.884				
460	0.06	44.985				
461	0.03	45.042				
462	0.03	45.042				
463	EndData					
464	EndSample					
465	BeginSample					
466	AverageLoad	0.97	"N"			
467	PeakLoad	2.53	"N"			
468	BeginData					
469	Load	SlackExt				

File header with metadata for Run

Sample 1

Calculated parameter
results for Sample 1

Load versus Extension data
for Sample 1

Sample 2

Calculated parameter

CUSTOM PYTHON PARSER FOR TENSILE DATA

- To try it with example data, clone from:
https://github.com/jlandgre/Python_Curve_Parser
 - Parser consists of a Python Class in curve_parse.py (run with example data from Example_Parse_Files.ipynb)
 - Creates output files: df_params.xlsx and df_raw.xlsx

Example_Parse_Files.ipynb x

Code Python 3

Parse Tensile Raw Data From MTS Testing

This template parses raw data files into rows/columns

- This uses the `read_files` procedure from the `TensileParsingRun` class in `libs/curve_parse.py`
- As inputs, the Class takes a folder path and dictionary with parsing definition

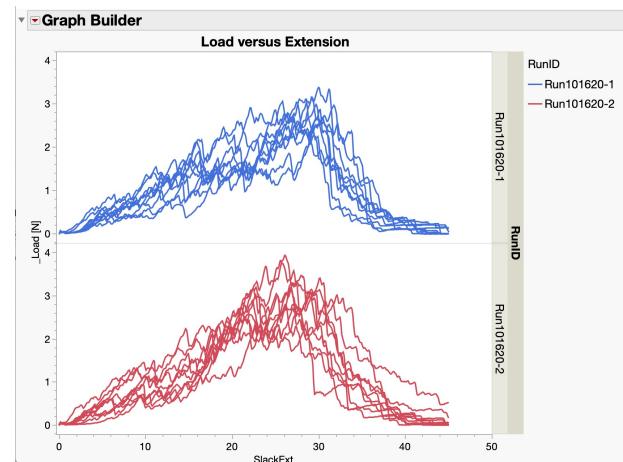
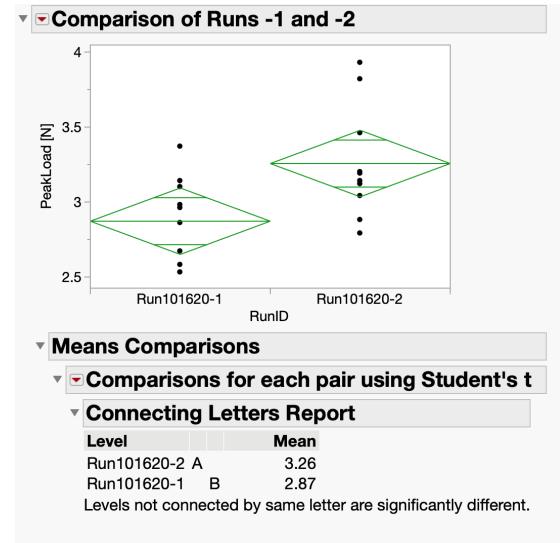
JDL 10/12/23

```
[7]: import pandas as pd
import numpy as np
import sys, os

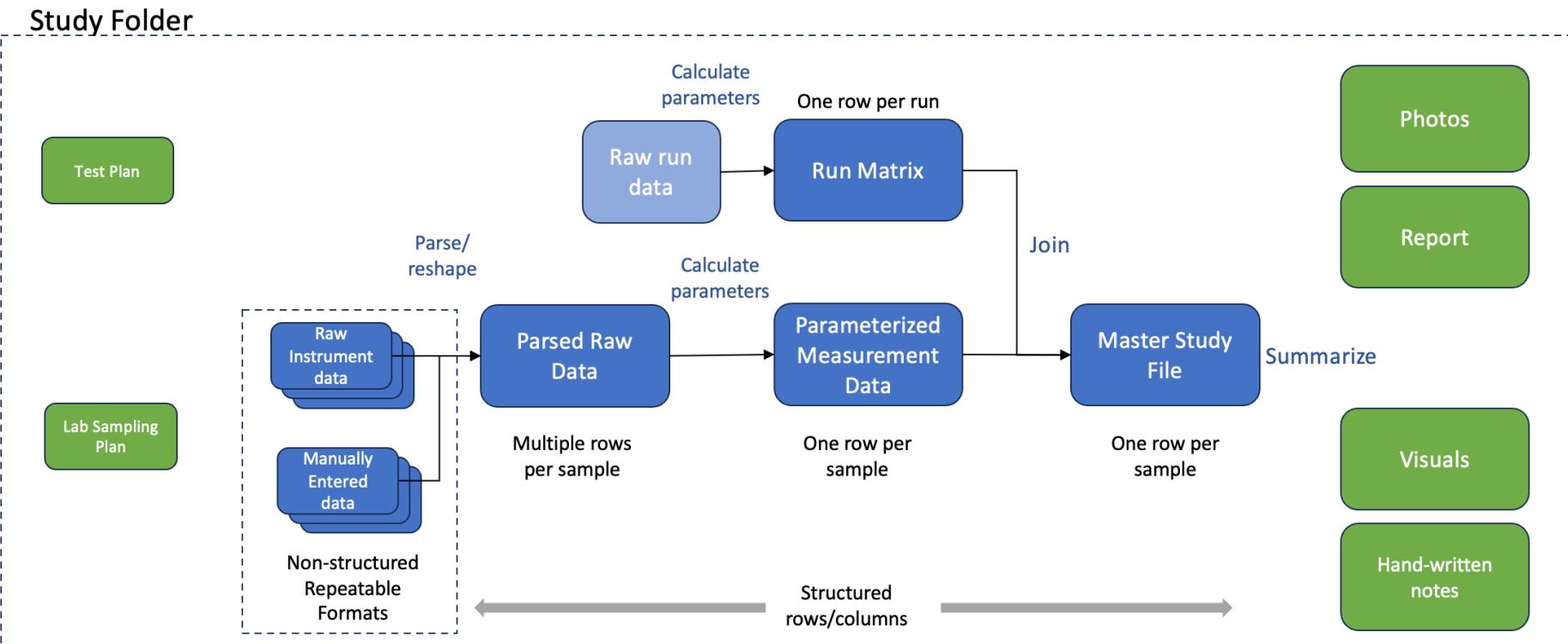
[8]: current_d: 8 =====
libs_dir: 9 TensileParsingRun Class
if not lib: 10 =====
from curve: 11 ===
Create c 12 class TensileParsingRun:
13     def __init__(self, parse_defn, path_folder):
14         """
15             Initializes a ParsingRun object
16             JDL 10/5/23
17
18             Args:
19                 parse_defn [Dictionary] description of how to parse
20                 path_folder [String] directory path of folder containing raw data files
21                 """
22
23             #User inputs
24             self.defn = parse_defn
25             self.path_folder = path_folder
26
27             #Current file while looping
28             self.file = ''
29
30             #Output DataFrames
31             self.df_raw = pd.DataFrame()
32             self.df_params = pd.DataFrame()
33
34             def read_files_procedure(self):
35                 """
```

JMP ANALYSIS OF PARSED DATA

- XXX
- Run101620-2 has significantly higher Peak Load than -1 ($p<0.05$)
- yyy



EXPERIMENT DATA “ARCHITECTURE”



CREATING A “MASTER STUDY FILE” TO SUMMARIZE AN EXPERIMENTAL STUDY

- Once raw data are parsed into rows/columns, we can join data and the run matrix to analyze the study
- Example study data has two measurement files and Run Matrix
 - ELN_19_AB1234 Measurements_Activity.csv
 - ELN_19_AB1234 Measurements_pH and Viscosity.csv
 - ELN_19_AB1234 Run Matrix.csv
- Goal is analyze by Run Matrix variables
- Need to also check within-run variability

Measurement Data 1

RunID	SampleID	Position	Replicate	Viscosity	pH
ELN_19_AB1234-1	Begin_1	Begin	1	1031	5.6
ELN_19_AB1234-1	Begin_2	Begin	2	1026	5.9
ELN_19_AB1234-1	Begin_3	Begin	3	1010	5.6
ELN_19_AB1234-1	Middle_1	Middle	1	1001	5.6
ELN_19_AB1234-1	Middle_2	Middle	2	991	5.6
ELN_19_AB1234-1	Middle_3	Middle	3	1022	5.7
ELN_19_AB1234-1	End_1	End	1	1008	5.6
ELN_19_AB1234-1	End_2	End	2	985	5.7
ELN_19_AB1234-1	End_3	End	3	982	5.5
ELN_19_AB1234-2	Begin_1	Begin	1	1214	6.7
ELN_19_AB1234-2	Begin_2	Begin	2	1232	6.5
ELN_19_AB1234-2	Begin_3	Begin	3	1223	6.5
ELN_19_AB1234-2	Middle_1	Middle	1	1189	6.5
ELN_19_AB1234-2	Middle_2	Middle	2	1204	6.3
ELN_19_AB1234-2	Middle_3	Middle	3	1202	6.5
ELN_19_AB1234-2	End_1	End	1	1189	6.6
ELN_19_AB1234-2	End_2	End	2	1182	6.6
ELN_19_AB1234-2	End_3	End	3	1194	6.5

Measurement Data 2

RunID	SampleID	Activity
ELN_19_AB1234-1	1	25
ELN_19_AB1234-1	2	24.4
ELN_19_AB1234-1	3	24.7
ELN_19_AB1234-1	4	24.7
ELN_19_AB1234-1	5	24.1
ELN_19_AB1234-1	6	24.8
ELN_19_AB1234-1	7	24.4
ELN_19_AB1234-1	8	24.3
ELN_19_AB1234-1	9	24.5
ELN_19_AB1234-1	10	25.1
ELN_19_AB1234-2	1	25.6
ELN_19_AB1234-2	2	24.8
ELN_19_AB1234-2	3	25.8
ELN_19_AB1234-2	4	26
ELN_19_AB1234-2	5	26.1
ELN_19_AB1234-2	6	25.9
ELN_19_AB1234-2	7	25.5
ELN_19_AB1234-2	8	25.6
ELN_19_AB1234-2	9	25.2
ELN_19_AB1234-2	10	26



Run Matrix

RunID	Date Made	Activity_Tgt	BatchID	Operator
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG		
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB		

Master Study File

RunID	Date Made	Activity	Tgt	BatchID	Operator	SampleID	Position	Replicate	Viscosity	pH	Activity
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_1	OG		Begin_1	Begin	1	1031	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_2	OG		Begin_2	Begin	2	1026	5.9	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_3	OG		Begin_3	Begin	3	1010	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_1	OG		Middle_1	Middle	1	1001	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_2	OG		Middle_2	Middle	2	991	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_3	OG		Middle_3	Middle	3	1022	5.7	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_1	OG		End_1	End	1	1008	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_2	OG		End_2	End	2	985	5.7	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_3	OG		End_3	End	3	982	5.5	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Begin_1	OG		Begin_1	Begin	1	1214	6.7	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Begin_2	OG		Begin_2	Begin	2	1232	6.5	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Begin_3	OG		Begin_3	Begin	3	1223	6.5	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Middle_1	OG		Middle_1	Middle	1	1189	6.5	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Middle_2	OG		Middle_2	Middle	2	1204	6.3	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	Middle_3	OG		Middle_3	Middle	3	1202	6.5	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	End_1	OG		End_1	End	1	1189	6.6	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	End_2	OG		End_2	End	2	1182	6.6	
ELN_19_AB1234-2	2/5/2020	24.5 XM-2456 OG	End_3	OG		End_3	End	3	1194	6.5	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_1	AB		Begin_1	Begin	1			25
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_2	AB		Begin_2	Begin	2			24.4
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Begin_3	AB		Begin_3	Begin	3			24.7
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_1	AB		Middle_1	Middle	1			24.7
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_2	AB		Middle_2	Middle	2			24.1
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	Middle_3	AB		Middle_3	Middle	3			24.1
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_1	AB		End_1	End	1			24.4
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_2	AB		End_2	End	2			24.4
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456 OG	End_3	AB		End_3	End	3			24.4
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Begin_1	AB		Begin_1	Begin	1			25.6
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Begin_2	AB		Begin_2	Begin	2			24.8
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Begin_3	AB		Begin_3	Begin	3			25.8
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Middle_1	AB		Middle_1	Middle	1			26
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Middle_2	AB		Middle_2	Middle	2			26.1
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	Middle_3	AB		Middle_3	Middle	3			25.9
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	End_1	AB		End_1	End	1			25.5
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	End_2	AB		End_2	End	2			25.6
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633 AB	End_3	AB		End_3	End	3			25.2

Concatenate + Join
(in JMP or Pandas)

EXPERIMENTAL DATA “KEYS”

- Keys [aka data nametags] make data easy to summarize and traceable
- A measurement data point needs up to 6 keys to fully identify the measurement and its source
- Avoid generic Run ID's (e.g. “Run 1”) because makes difficult to combine studies
- Check: Is each row identified by unique keys?

Other Keys Needed

AnalysisID – Uniquely ID's an instance of analyzing Samples from a Run (Lab software often assigns this)

PositionID – Identifies order of data points in curve data

Run ID – Uniquely ID's composition and conditions aka Batch ID, Laundry Load ID, Panelist ID, Modeling Run ID

Study ID – Uniquely ID's collection of Runs; Use ELN Experiment number or created date string as StudyID

RunID	Date Made	Activity_Tgt	BatchID	Operator	SampleID	Position	Replicate	Viscosity	pH	Activity
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Begin_1	Begin	1	1031	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Begin_2	Begin	2	1026	5.9	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Begin_3	Begin	3	1010	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Middle_1	Middle	1	1001	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Middle_2	Middle	2	991	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		Middle_3	Middle	3	1022	5.7	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		End_1	End	1	1008	5.6	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		End_2	End	2	985	5.7	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG		End_3	End	3	982	5.5	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Begin_1	Begin	1	1214	6.7	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Begin_2	Begin	2	1232	6.5	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Begin_3	Begin	3	1223	6.5	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Middle_1	Middle	1	1189	6.5	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Middle_2	Middle	2	1204	6.3	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		Middle_3	Middle	3	1202	6.5	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		End_1	End	1	1189	6.6	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		End_2	End	2	1182	6.6	
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB		End_3	End	3	1194	6.5	
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							25
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.4
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.7
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.7
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.1
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.8
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.8
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							24.5
ELN_19_AB1234-1	2/1/2020	24.5 XM-2456	OG							25.1
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.6
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							24.8
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.8
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							26
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							26.1
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.9
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.5
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.6
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							25.2
ELN_19_AB1234-2	2/5/2020	25.75 XM-2633	AB							26

Sample ID – Uniquely identifies “when/where” tested material came from

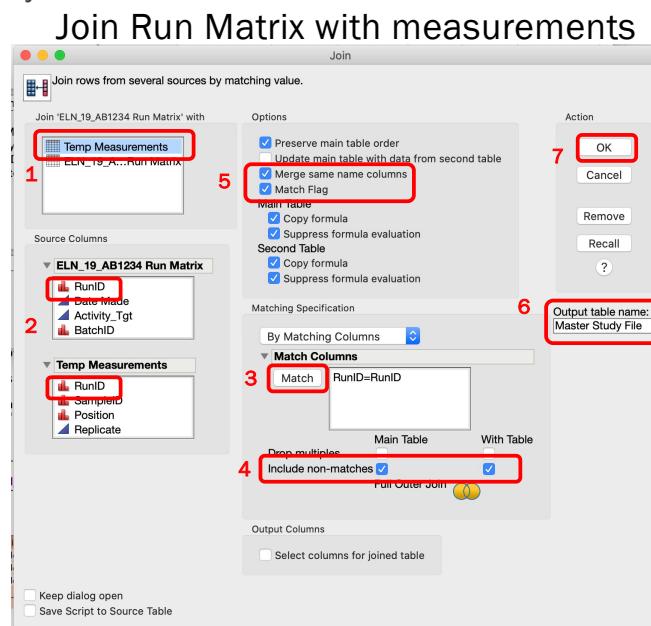
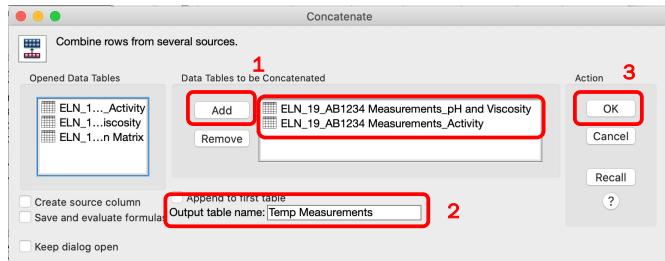
Analyte Name

AGGREGATING AND SUMMARIZING DATA (JMP)

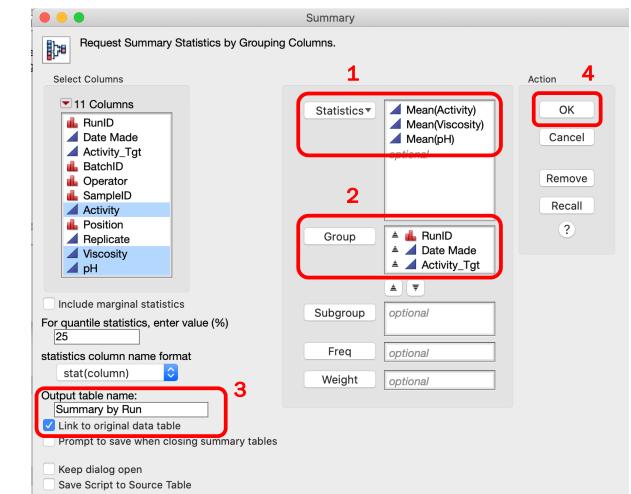
- Open JMP and three *.csv files in Batch Example sub-folder
- To get Master Study File and Summary
 - Tables/Concatenate: Measurements_Activity.csv and Measurements_pH and Viscosity.csv (Makes Temp Measurements.csv)
 - Tables/Join: Run Matrix.csv and Temp Measurements.csv
 - Tables/Summary
- Use Master Study File for statistical analysis

Tables	Rows	Cols	D
Summary			
Subset			
Sort			
Stack			
Split			
Transpose			
Join			
Update			
Concatenate			
JMP Query Builder			

Concatenate measurements



Summarize by Run



AGGREGATING DATA (PANDAS)

See Jupyter Notebook:

Pandas Master Study File Demo.ipynb

This performs the same concatenate and join transformations as the previous JMP example

The screenshot shows a Jupyter Notebook interface with a title bar 'Pandas Master Study File.ipynb'. The main area is titled 'Master Study File Creation' and contains the following text and code snippets:

Imports and concatenates measurement data files. Joins measurement data with a run matrix and outputs a "Master Study File" as a *.csv.

J.D. Landgrebe updated October 25, 2023

Open the three files and set variable lists

```
[ ]: import pandas as pd
lst_Meas_dfs = [pd.read_csv('ELN_19_AB1234 Measurements_pH and Viscosity.csv'),
                pd.read_csv('ELN_19_AB1234 Measurements_Activity.csv')]
df_RunMatrix = pd.read_csv('ELN_19_AB1234 Run Matrix.csv')

#Construct useful lists for later use
lstAllIDs = ['RunID','SampleID','Position','Replicate']
lstMeasVar, lstRunVar = [], []
for df in lst_Meas_dfs:
    for var in df.columns:
        if var not in lstAllIDs: lstMeasVar.append(var)
for var in df_RunMatrix.columns:
    if not var in lstAllIDs: lstRunVar.append(var)

[ ]: df_RunMatrix
```

[]: for df in lst_Meas_dfs:
 print(df.info(),'\n')

Concatenate Measurements and Join with the Run Matrix

Index by Keys + Sample Variables and order the remaining columns by Run Variables and Measurements

```
[ ]: #Concatenate measurement df's
df_Measurements = pd.concat(lst_Meas_dfs,sort=False)

#Merge the Run Matrix with the measurements table
df_MasterStudyFile = df_RunMatrix.merge(df_Measurements ,left_on='RunID',right_on='RunID')

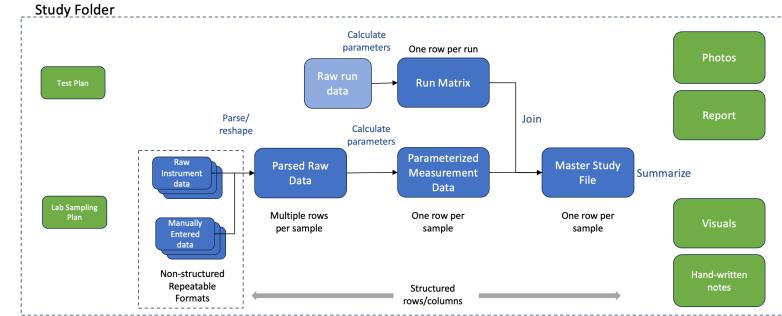
#Set DataFrame index
df_MasterStudyFile.set_index(lstAllIDs,inplace=True)
df_MasterStudyFile
```

Write the file to a CSV for JMP etc.

```
[ ]: df_MasterStudyFile.to_csv('ELN_19_AB1234 MSF.csv')
```

Create Summaries by RunID and RunID + Position

WORKING WITH SOFTWARE AND EXPERIMENT DATA



- Generalist software recos give you a toolbox for working with a broad range of engineering data
- JMP® and Python Pandas are important tools to always have around for working with data
- The Study Folder approach curates experiments to make them efficiently extensible and shareable
- The standard architecture and templates for experiments give you a reusable cookbook for designing experiments and working with the resulting data