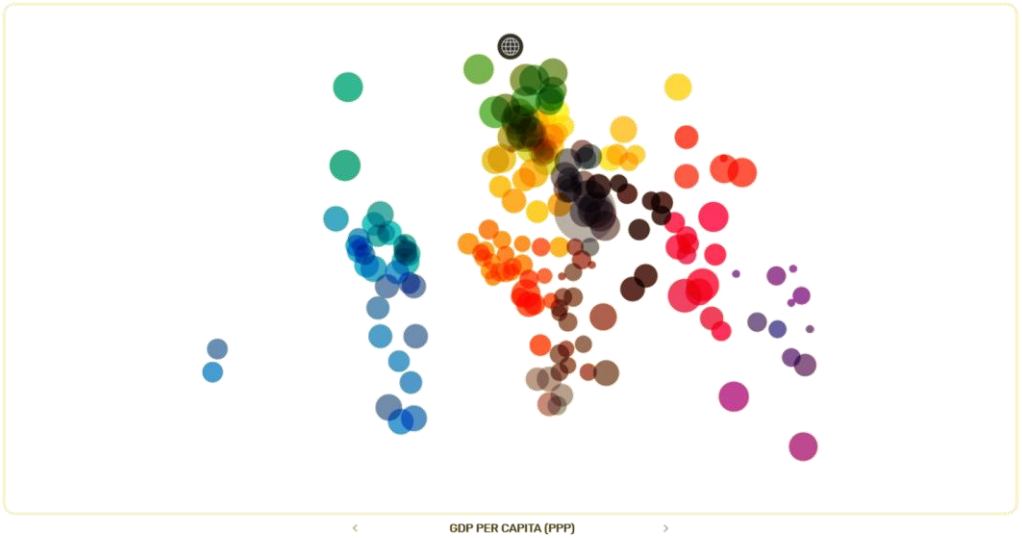
McMaster University ENGINEERING

Chemical Engineering 4H03

Visualizing Data

Jake Nease
McMaster University



https://govdna.frontwise.com/#layout/geo/country/SAU/x/32/y/5/z/8/a/1

Before we Begin…

- The winner of the grading scheme is…
- We need to decide on office hours!
  - Mostly just a reminder for me

Types of Data

What is a data?*

https://i.pinimg.com/736x/1e/b7/d3/1eb7d3bf6a073514960e535622c465ba.jpg

*A miserable little pile of secrets

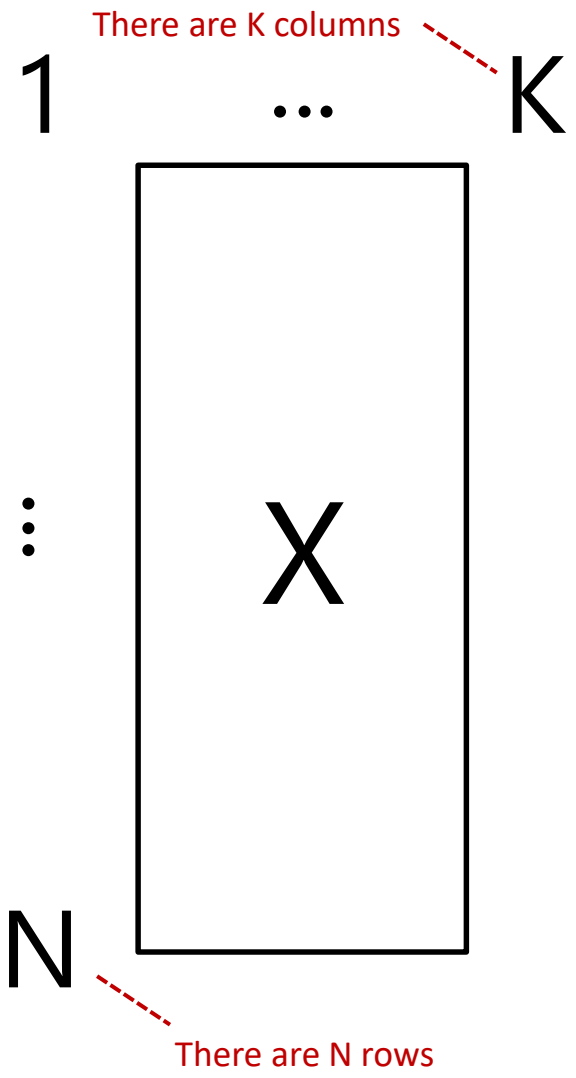
Outline of this Section

- This section is intended to introduce you to the types of data we deal with as engineers
  - A brief history of data
  - Workshop on data analysis in industry/research
- We will also look at visualization tools that will help us visualize and analyze what we are looking at
  - Types of plots
  - Review of plot metrics
- This material is very qualitative
  - Still testable, though! ☺

Types of Data we Deal With

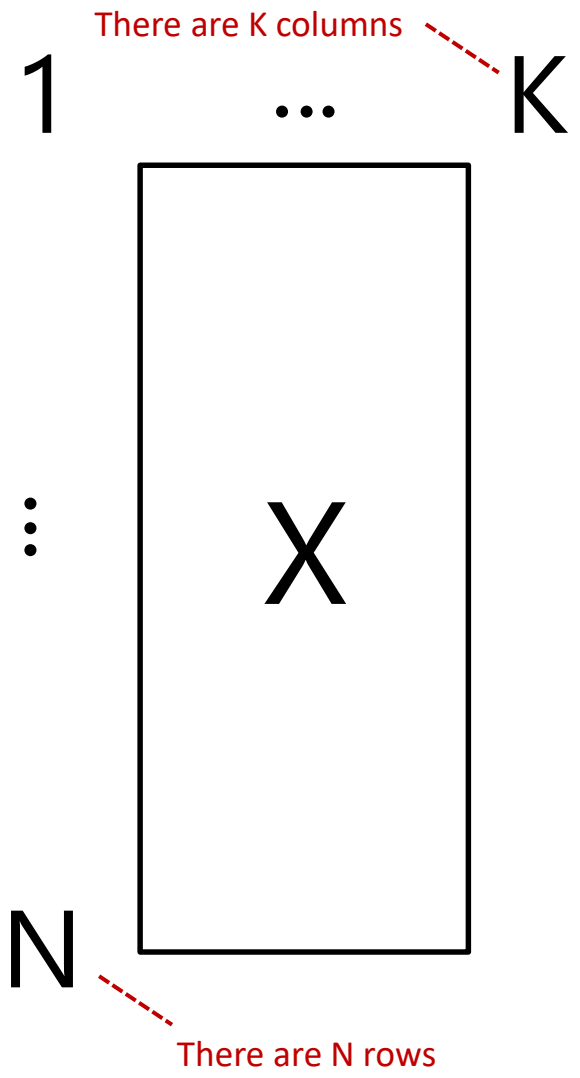
- A data set is typically called **X**

- 1920s – 1950s
  - Small number of columns
  - **K << N**
  - Visualize with scatter plots
  - Can perform Multilinear regression (MLR)
  - Choose which columns to use
    - Independent
    - Low error
    - Low measurement noise
  - Examples?



Types of Data we Deal With

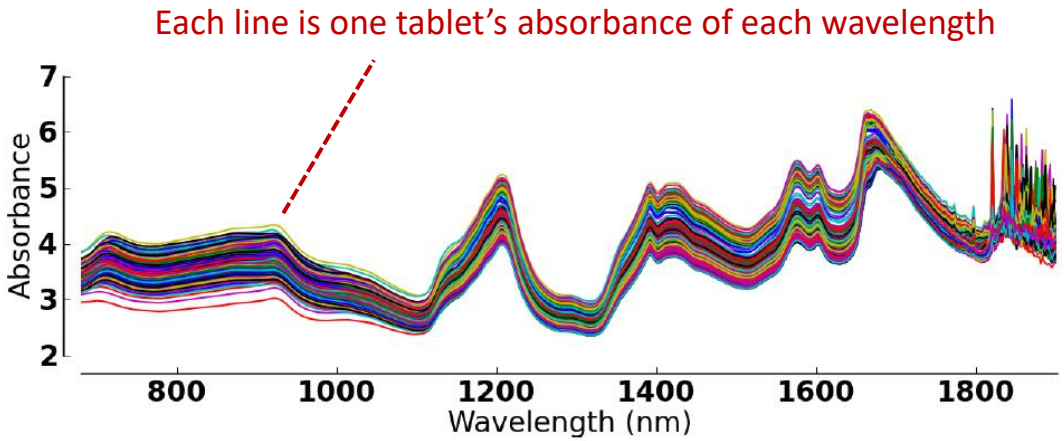
- **Examples** of data sets with small K and (possibly big) N
  - Flow/temp/press measurements of a certain stream
  - Quantified data over time
    - Weight
    - Height
  - Economics or consumer data
  - Others?

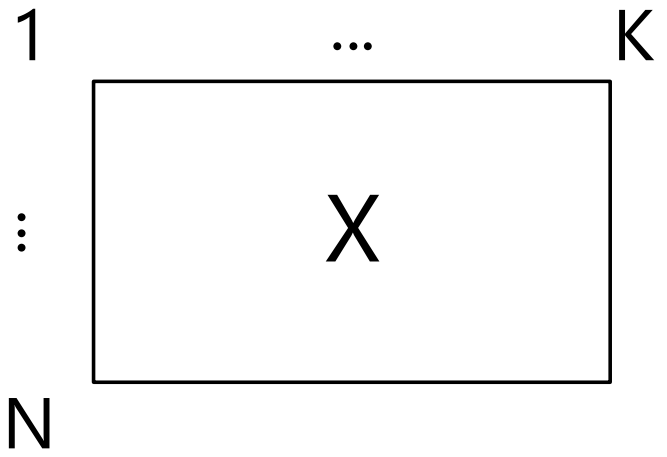


Types of Data we Deal With

- **FAT** data sets (small N and small K, even K > N)
  - Expensive, detailed measurements
  - Low frequency (in the K > N case)
  - Typically a lot of correlated (dependent) data

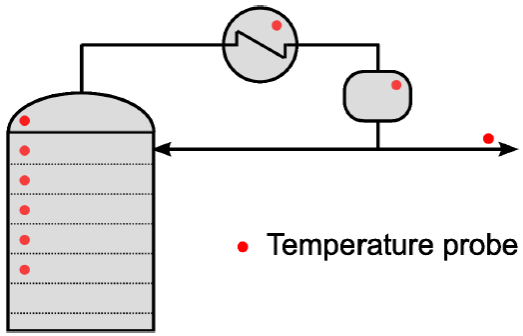
- Example: **spectral data** (how does it work?)



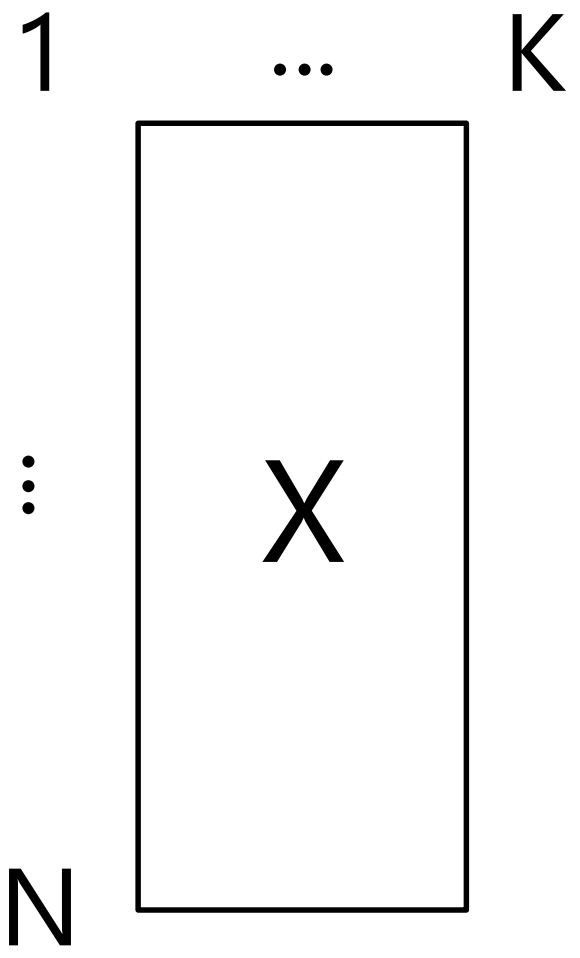


Types of Data we Deal With

- **Really tall** data sets (huge N!)
  - Could have dependent variables
  - Can consume huge amounts of data storage
  - Example: column with redundancies

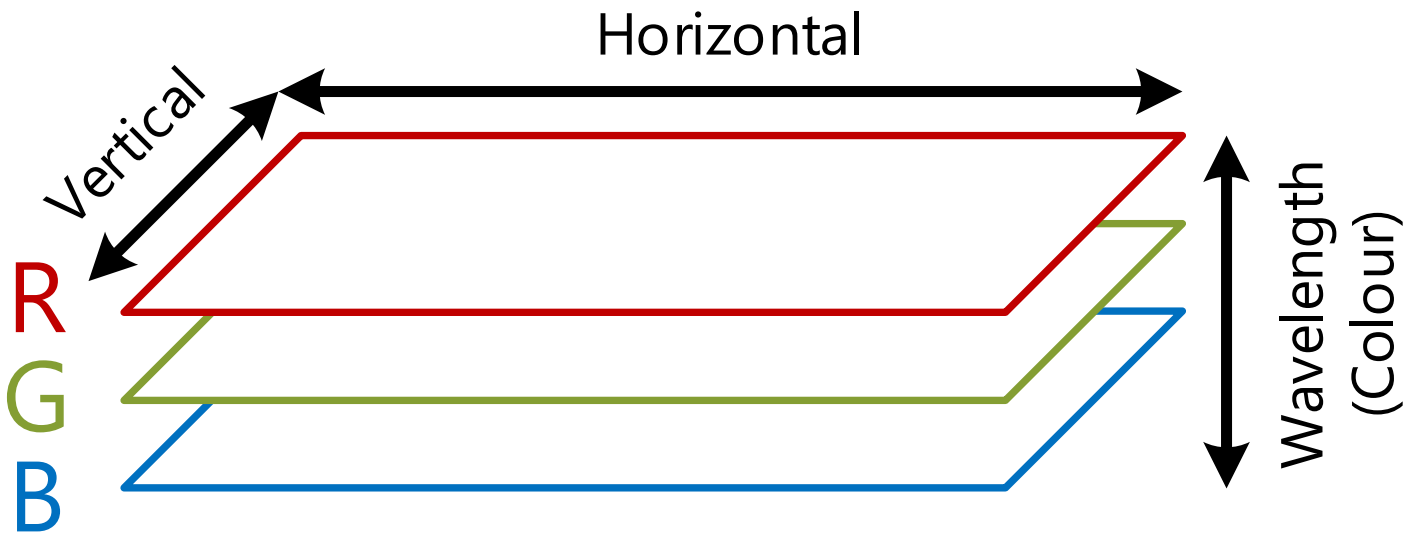


  - 35+ temperature, flow, pressure, and some "calculated" (inferred) values!
  - ISSUE: causes singularities when fitting regression functions!



Types of Data we Deal With

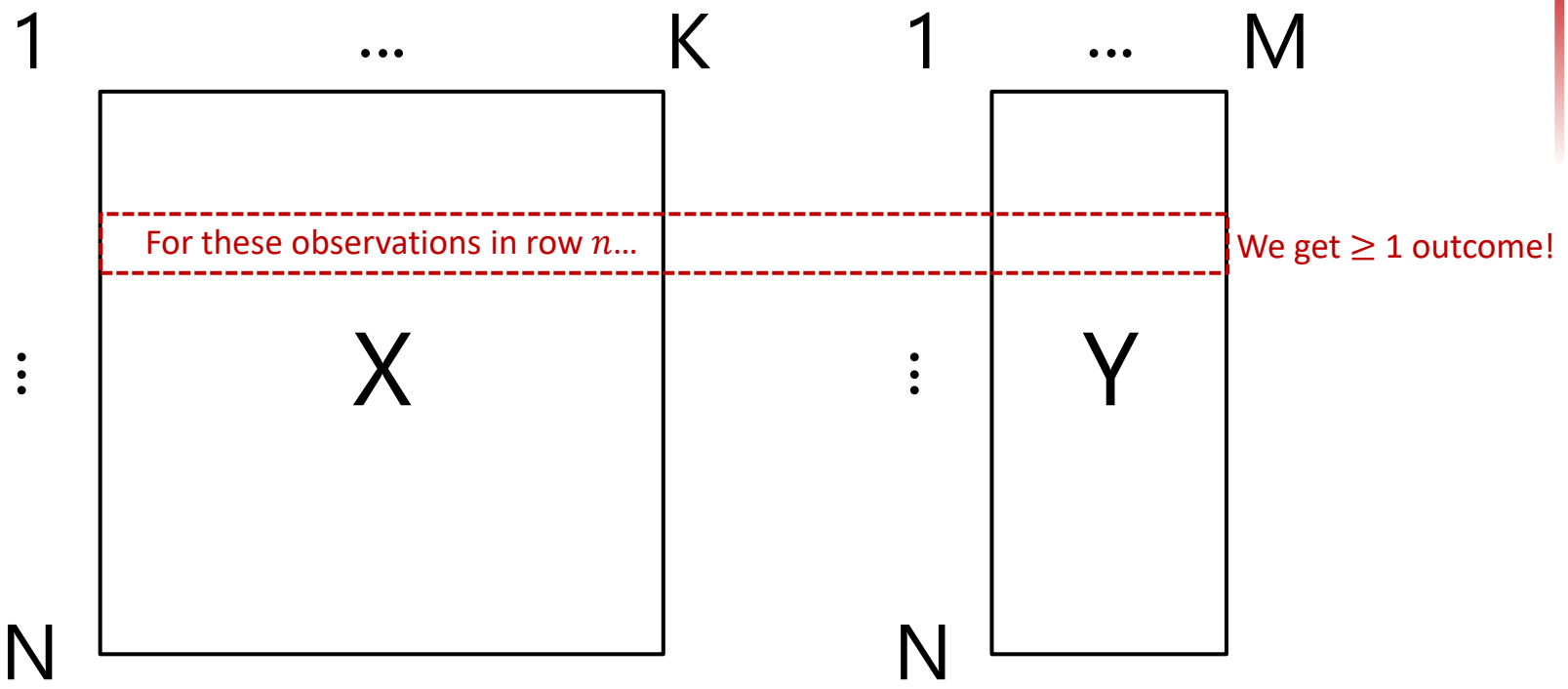
- **3D Data sets** (and even higher dimensions!)
  - Image data
  - Not just for Instagram but also for industry(gram)
  - Example in `MATLAB` of unpacking image data



  - *High measurement redundancy*: neighbouring pixels likely have same (or similar) data! Solution: Convolution

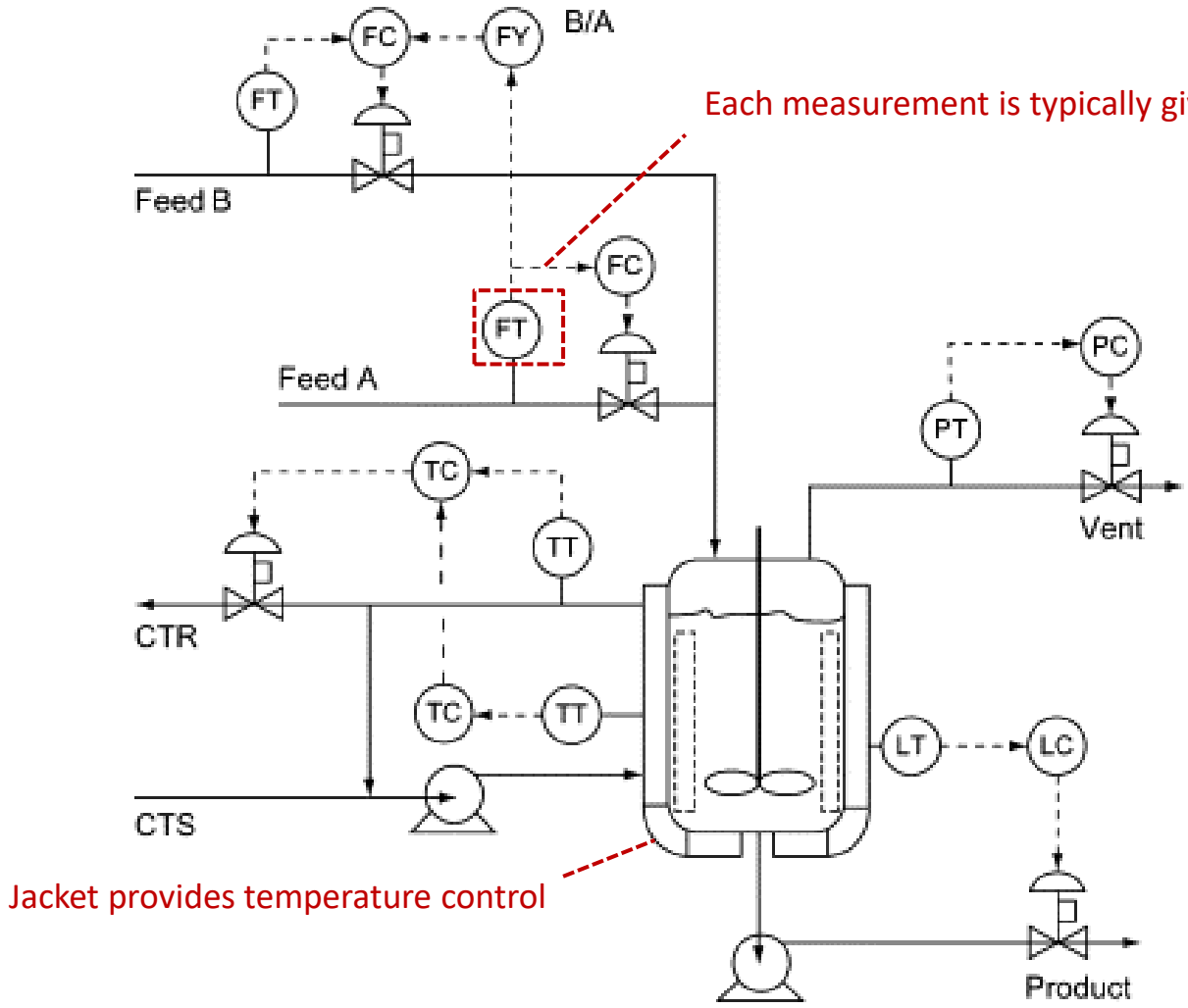
Types of Data we Deal With

- **Introducing outcome variables (Y)**
  - At the end of the day, we may want to *predict* one or more variables!
  - *Ideally*, at each $n \in N$ observation in X we have an $n \in N$ observation in Y



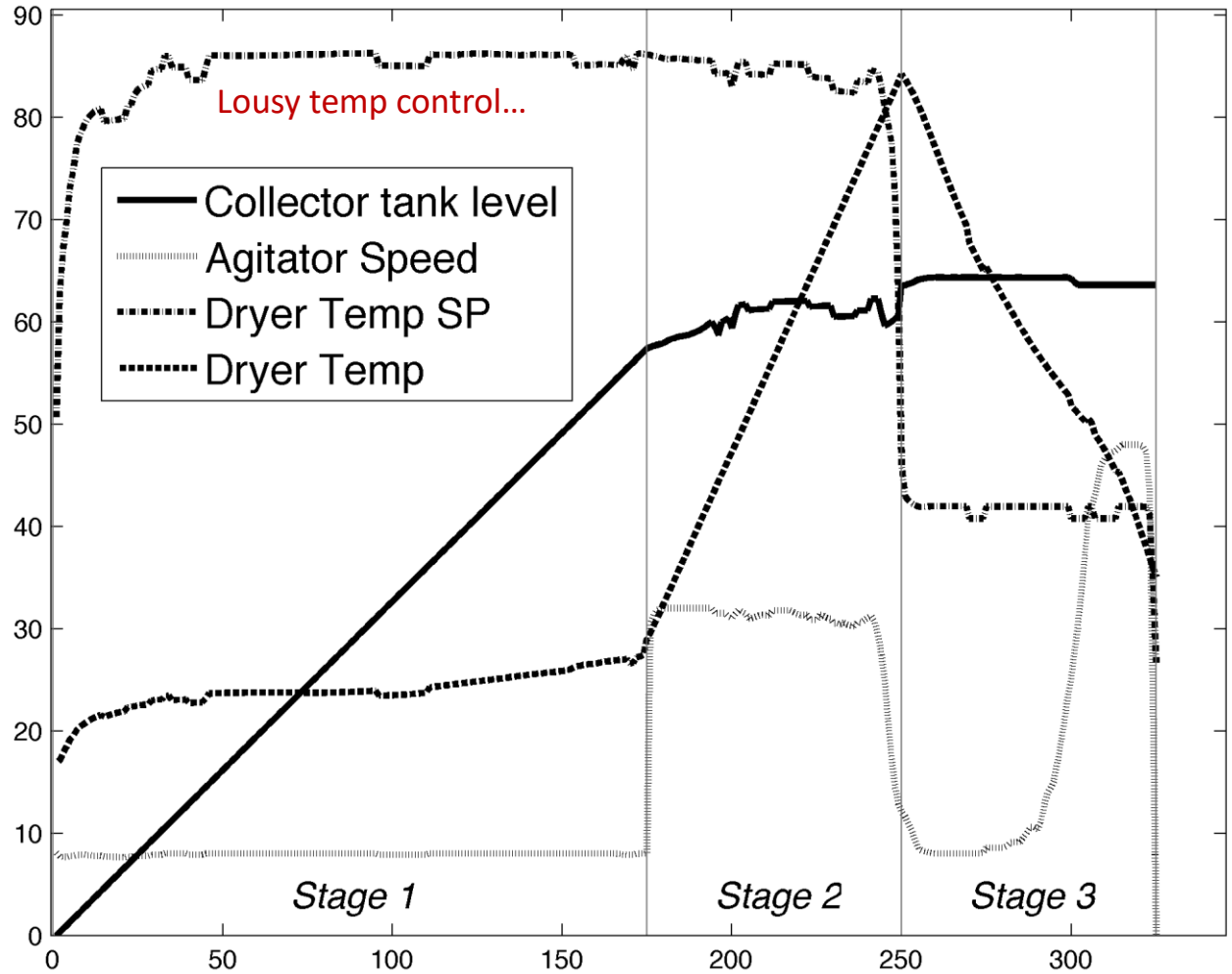
Types of Data we Deal With

- **Batch reactor and measurement reviews**



Types of Data we Deal With

- Batch data sets



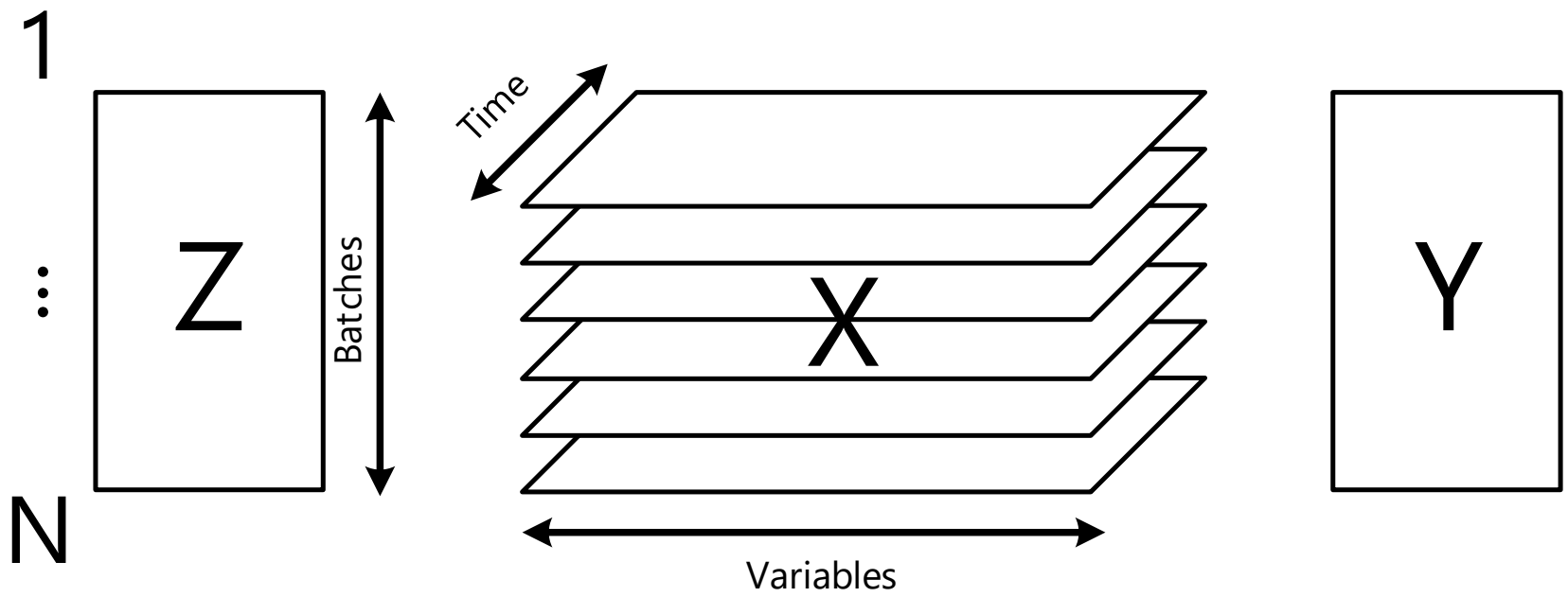
These are variables in X, but we also certainly will have "outcomes" in Y such as MWD, viscosity, purity, blah blah blah

Types of Data we Deal With

- Batch data sets

Lousy temp control…

  - Multidimensional (multiple vars over time)…
  - For multiple batches!



  - Multiblock data sets
  - Stems field of data fusion

Class Workshop

- What are some data sets you have encountered recently?
  - How did you use them?
  - Did you learn anything from them?
  - Talk to your neighbour and lets share!

Visualization

I can see you have an eye for quality

https://www.highsnobiety.com/2014/09/24/10-greatest-minor-simpsons-characters-quotes/

Data Visualization

- **Data visualization** is frequently taken for granted
  - *A picture is worth a thousand words*
  - The human brain is incredible at recognizing patterns and sorting through what is **signal** (bride and groom) and **noise** (that friend photobombing your wedding photo)
    - In fact, it is SO good that they invented this thing called "**machine learning**" that attempts to mimic the brain. Maybe you've heard of it?
  - Humans are so smart that they can deal with bad plots
    - However, good plots are… better.

- *Let the data speak for themselves*

Univariate Data

- **Univariate data** is data with only **one column**. There may be ultimate instances of that one column for another variable, but the measurement is the same
- Examples
  - Samples with similar measurements or SAME UNITS
    - Temperatures at each probe in a reactor
    - Concentrations at the end of each batch
    - Yield stress results of material samples
  - Data sets that can be compared as groups
    - Course grades
    - Income by demographic
    - Cost of living by geography

Univariate Data: **Box Plots**

- **Box Plots** display a five-number summary of a variable
  - Minimum
  - 25th percentile (1st quartile)
  - 50th percentile (median)
  - 75th percentile (3rd quartile)
  - Maximum

- Notes
  - 25th "percentile" is the value below which 25% of the observations are found
  - Definition: **Interquartile Range (IQR)** is difference between 3rd and 1st quartiles

Univariate Data: **Box Plots**

- Visualization is paramount to success
  - What can you make of these numbers?

Thickness of a wooden board at six positions

	Pos1	Pos2	Pos3	Pos4	Pos5	Pos6
1	1761	1739	1758	1677	1684	1692
2	1801	1688	1753	1741	1692	1675
3	1697	1682	1663	1671	1685	1651
4	1679	1712	1672	1703	1683	1674
5	1699	1688	1699	1678	1688	1705
...						
96	1717	1708	1645	1690	1568	1688
97	1661	1660	1668	1691	1678	1692
98	1706	1665	1696	1671	1631	1640
99	1689	1678	1677	1788	1720	1735
100	1751	1736	1752	1692	1670	1671

2E04 final grades from the last five years

	2015	2016	2017	2018	2019
1	85.231	64.902	105.45	81.065	84.601
2	84.173	49.428	72.425	59.155	74.309
3	84.351	42.788	71.218	81.95	63.684
4	83.209	10.892	76.985	72.165	61.625
5	68.032	74.798	80.017	83.465	93.263
...					
80	77.466	82.492	58.415	88.005	60.97
81	74.566	88.618	80.21	63.235	64.95
82	73.069	64.214	68.25	82.36	88.475
83	72.61	62.697	70.778	79.805	88.005
84	71.119	77.238	95.835	95.415	93.3
85	56.936	56.041	80.895	82.2	97.525

Board data courtesy of ConnectMV

Univariate Data: **Box Plots**

- Visualization is paramount to success
  - Is this any better?

Thickness of a wooden board at six positions

	Pos1	Pos2	Pos3	Pos4	Pos5	Pos6
Min. :	1524	1603	1594	1452	1568	1503
1st Qu.:	1671	1657	1654	1667	1662	1652
Median :	1680	1674	1672	1678	1673	1671
Mean :	1687	1677	1677	1679	1674	1672
3rd Qu.:	1705	1688	1696	1693	1685	1695
Max. :	1822	1762	1763	1788	1741	1765

2E04 final grades from the last five years

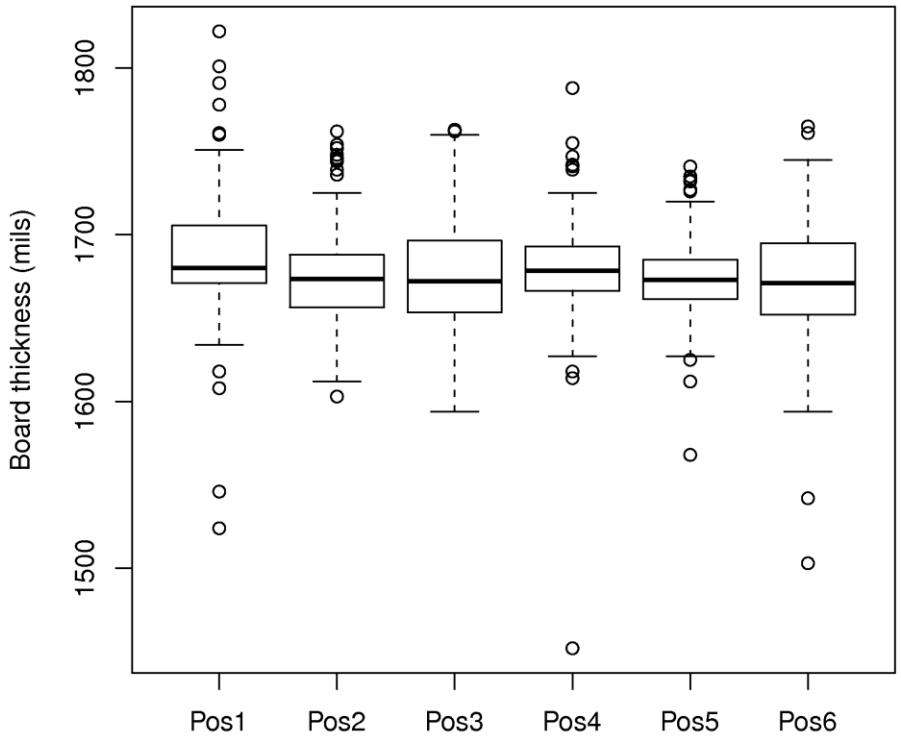
	2015	2016	2017	2018	2019
0	28.92	10.89	4.00	13.20	0.00
1	46.75	60.85	58.19	62.03	61.62
2	55.30	72.78	68.15	72.29	74.31
3	66.13	84.25	78.03	83.45	85.83
4	85.23	101.06	105.45	102.05	103.48

Board data courtesy of ConnectMV

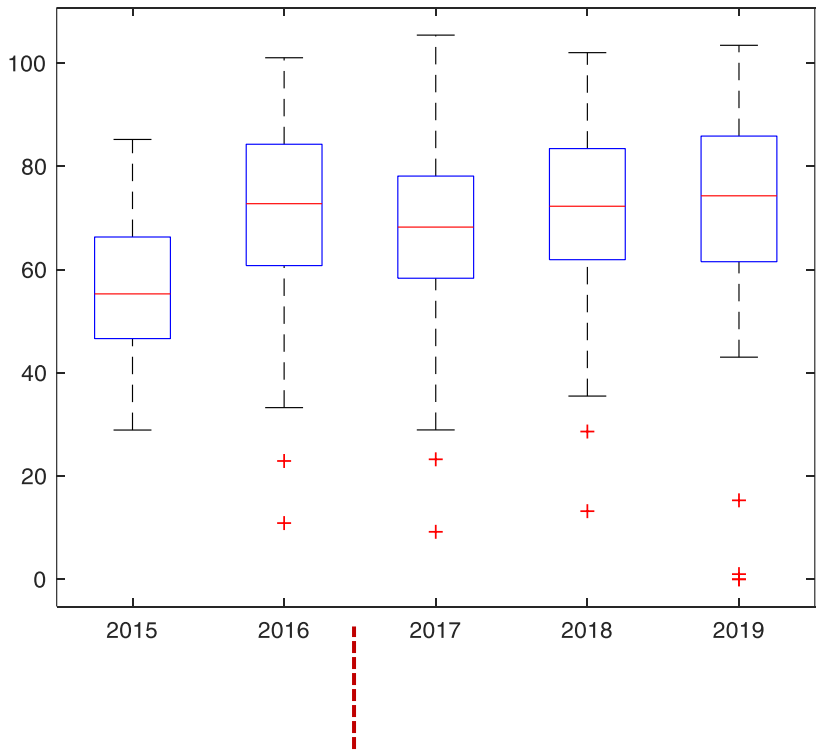
Univariate Data: **Box Plots**

- Visualization is paramount to success
  - How about now?

Thickness of a wooden board at six positions



2E04 final grades from the last five years



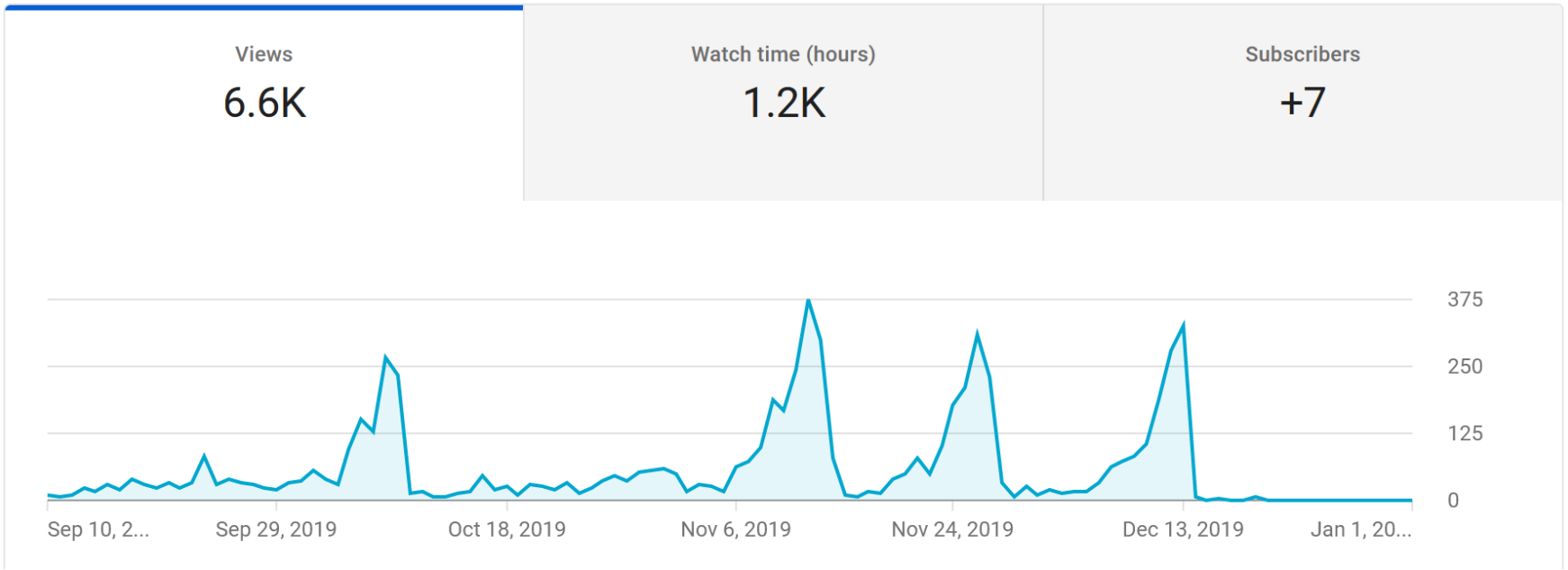
WORKSHOP: Let's talk about the value of visualizing this for a moment…

Board data courtesy of ConnectMV

Univariate Data: **Time-Series Plot**

- A two-dimensional plot
  - Horizontal axis: time or another variable of **logical order**
  - Vertical axis: data of interest
- Good to see trends (process monitoring, sales…)

ChE 2E04 Youtube Channel Sep 10 – Dec 31 2019

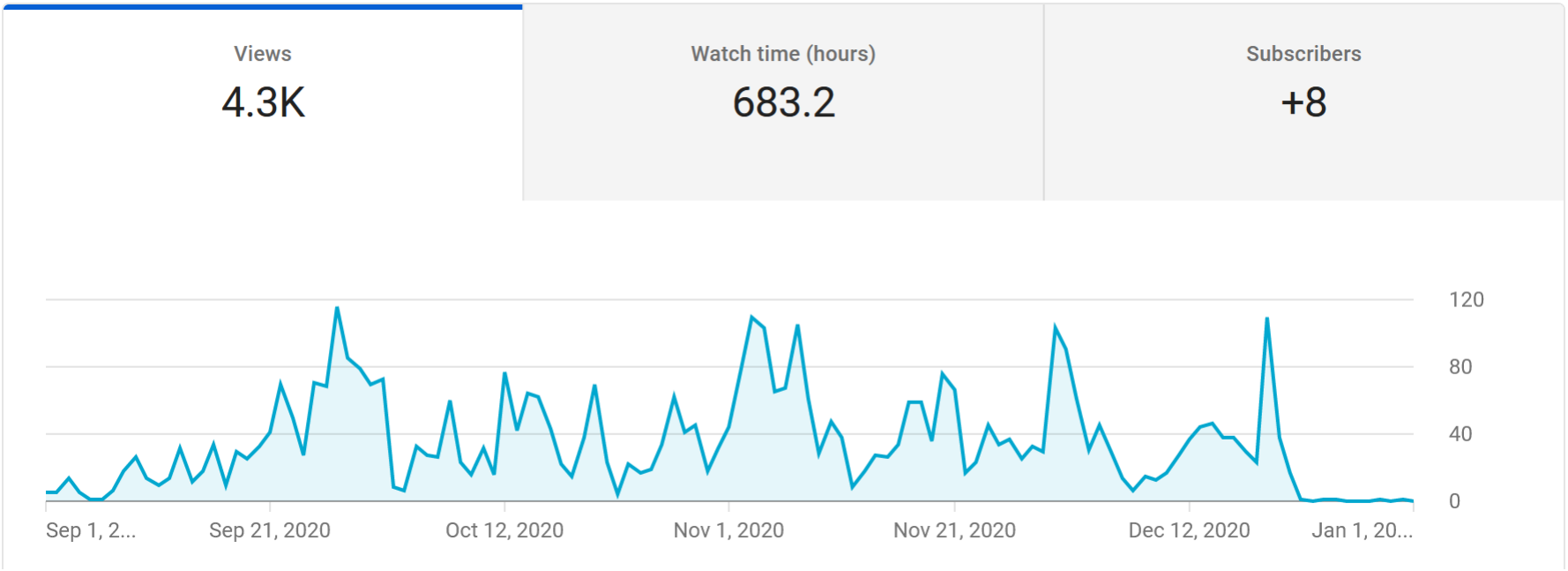


What can we infer from this?

Univariate Data: **Time-Series Plot**

- A two-dimensional plot
  - Horizontal axis: time or another variable of **logical order**
  - Vertical axis: data of interest
- Good to see trends (process monitoring, sales…)

ChE 2E04 Youtube Channel Sep 10 – Dec 31 2019

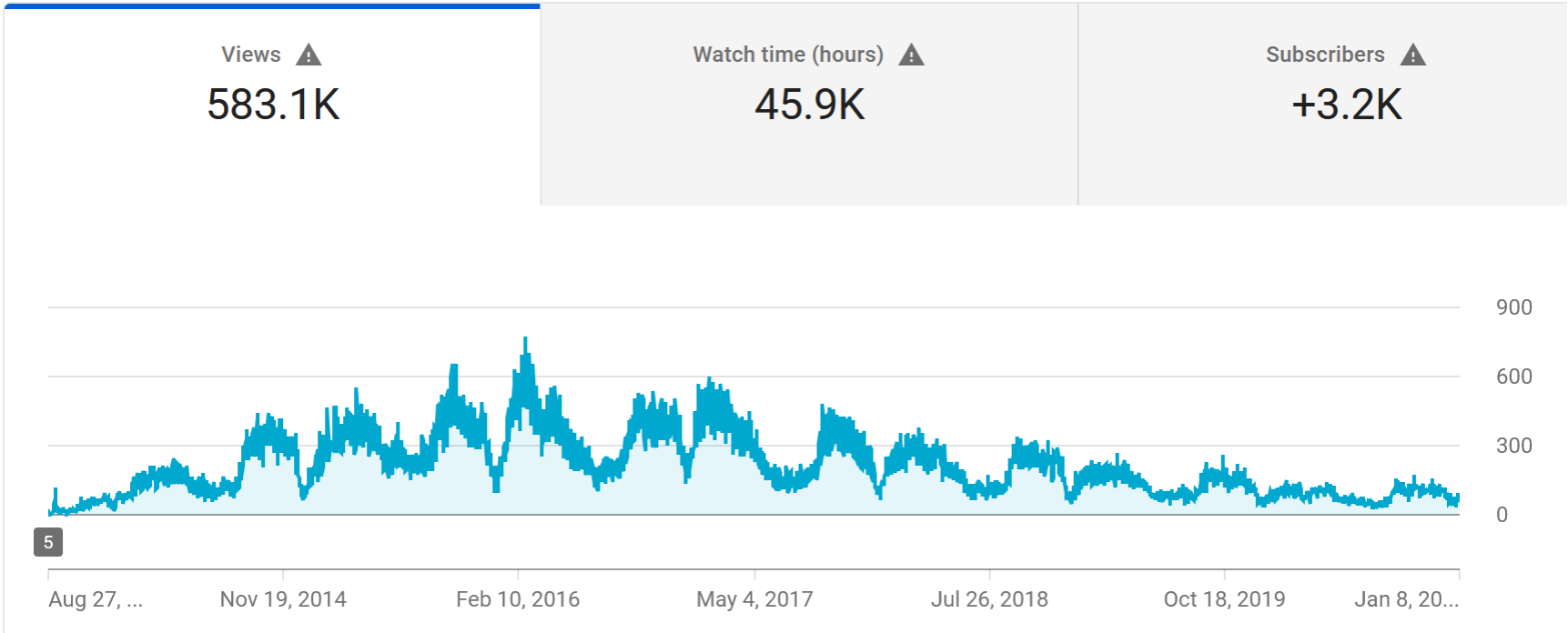


What can we infer from this? How does it relate to 2019?

Univariate Data: **Time-Series Plot**

- More data = more macro-information
  - And even more micro? Anything here?

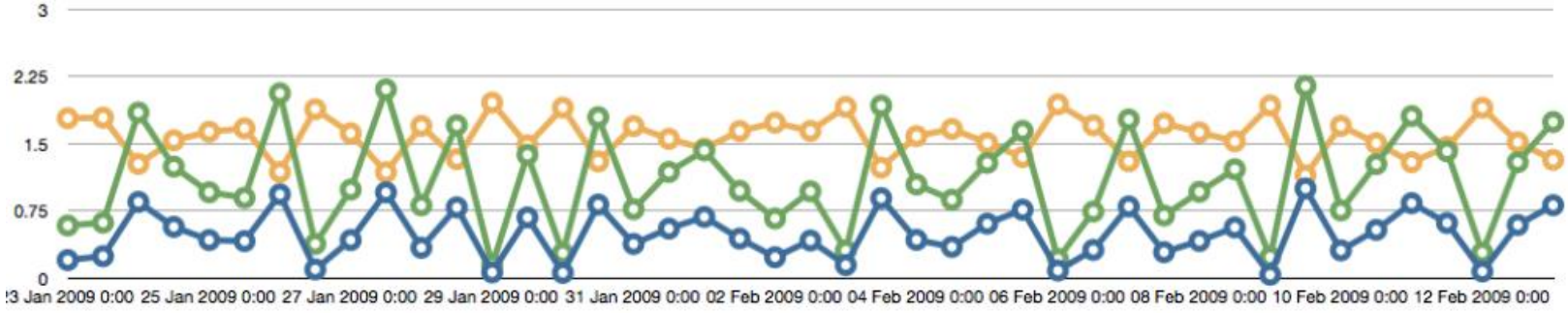
Personal MATLAB Youtube Channel Lifetime



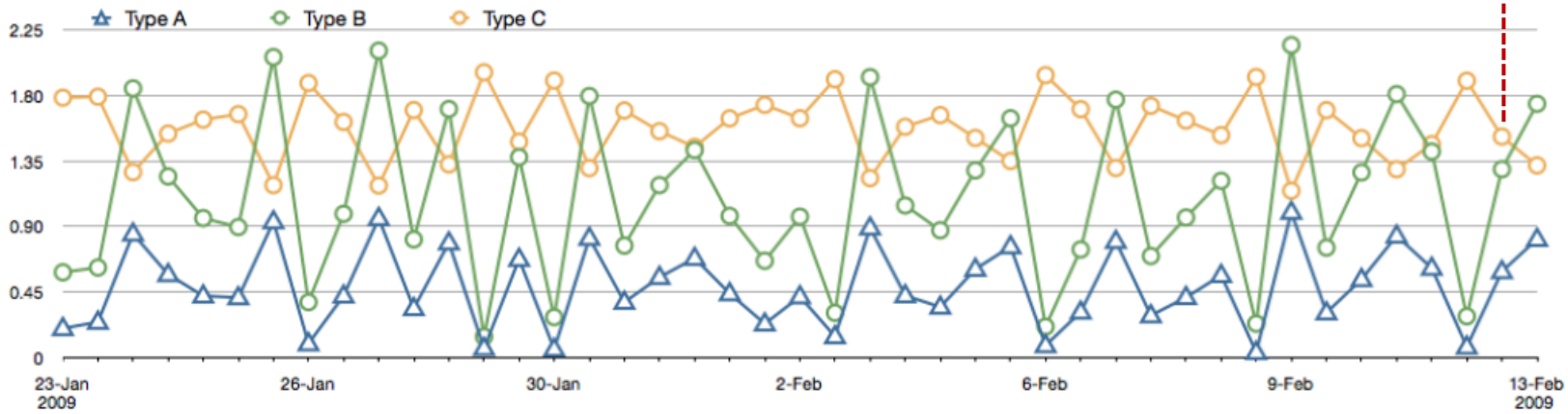
What can we infer from this?

Univariate Data: **Time-Series Plot**

- Multiple lines should not cross/jumble
  - Use separate axes if possible
  - Using different colours and markers don't help much



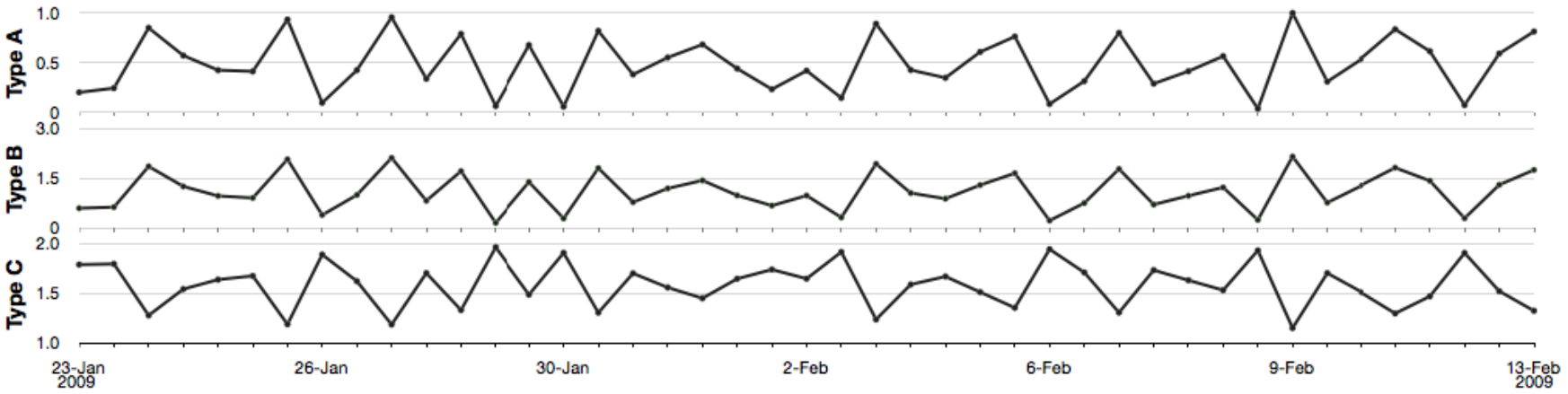
An improvement, but still a jumble



Figures courtesy of ConnectMV

Univariate Data: **Time-Series Plot**

- Multiple lines should not cross/jumble
  - **Use separate axes if necessary**

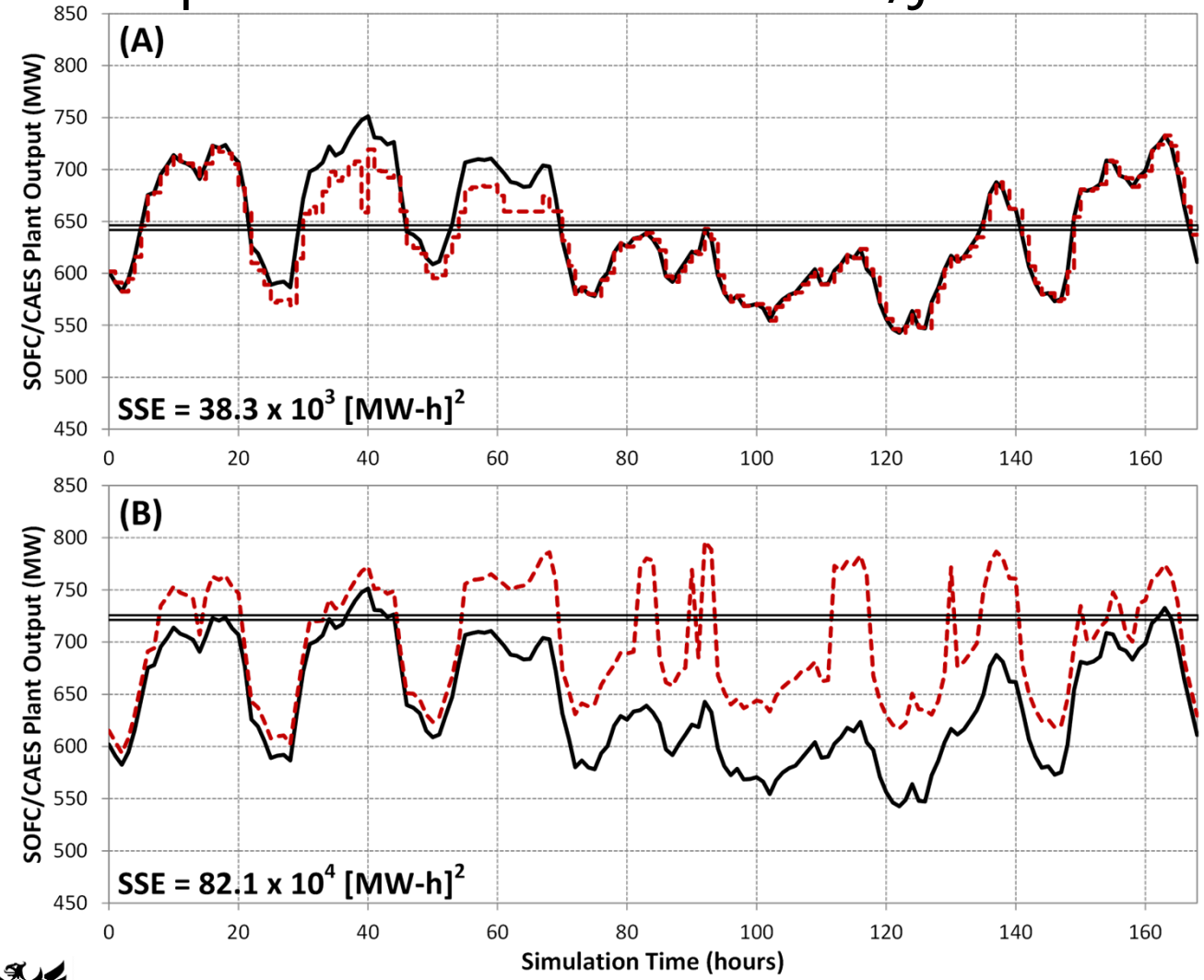


  - In this case we are looking for trends, and do not care as much about relative values (hence different axis scales)

Figures courtesy of ConnectMV

Univariate Data: **Time-Series Plot**

- Multiple lines should not cross/jumble



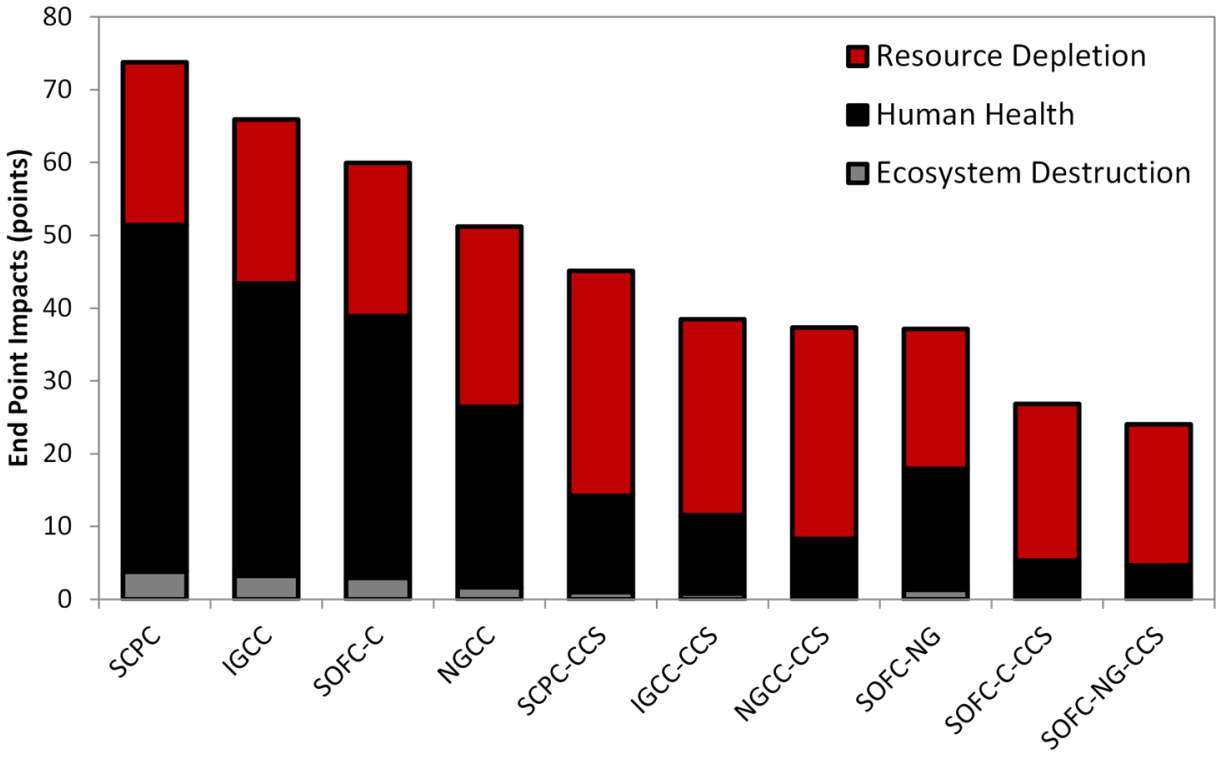
In this case, having the same y-axis scaling helps emphasize the difference in performance between control strategies for the same set points (black line)

Univariate Data: **Bar Plots**

- Bar Plots are used to represent categorical data
- Best to use if:
  - Many categories
  - Axis order does not matter (but a **good order can still help!**)

Can use "stacks" to show individual contributions to a great whole

Ordering highest → lowest (or vice-versa) can add interpretive value

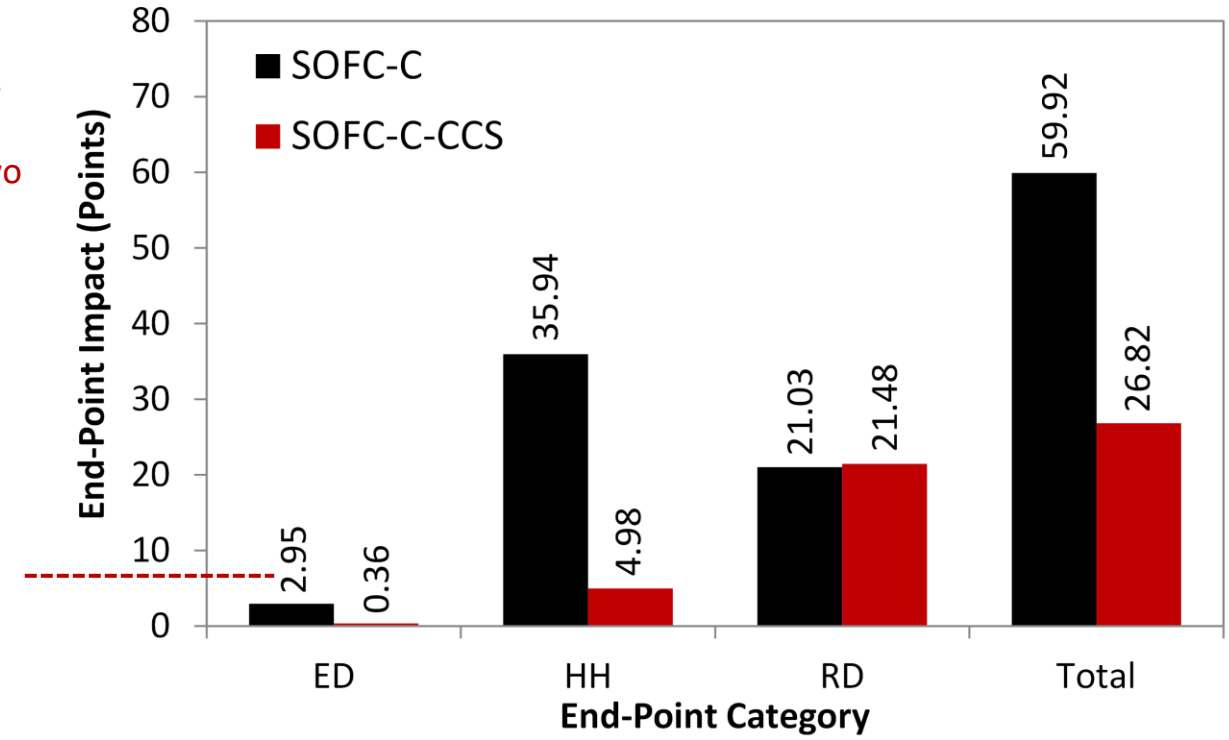


Univariate Data: **Bar Plots**

- Bar Plots are used to represent categorical data
- Best to use if:
  - Many categories
  - Axis order does not matter (but a good order can still help!)

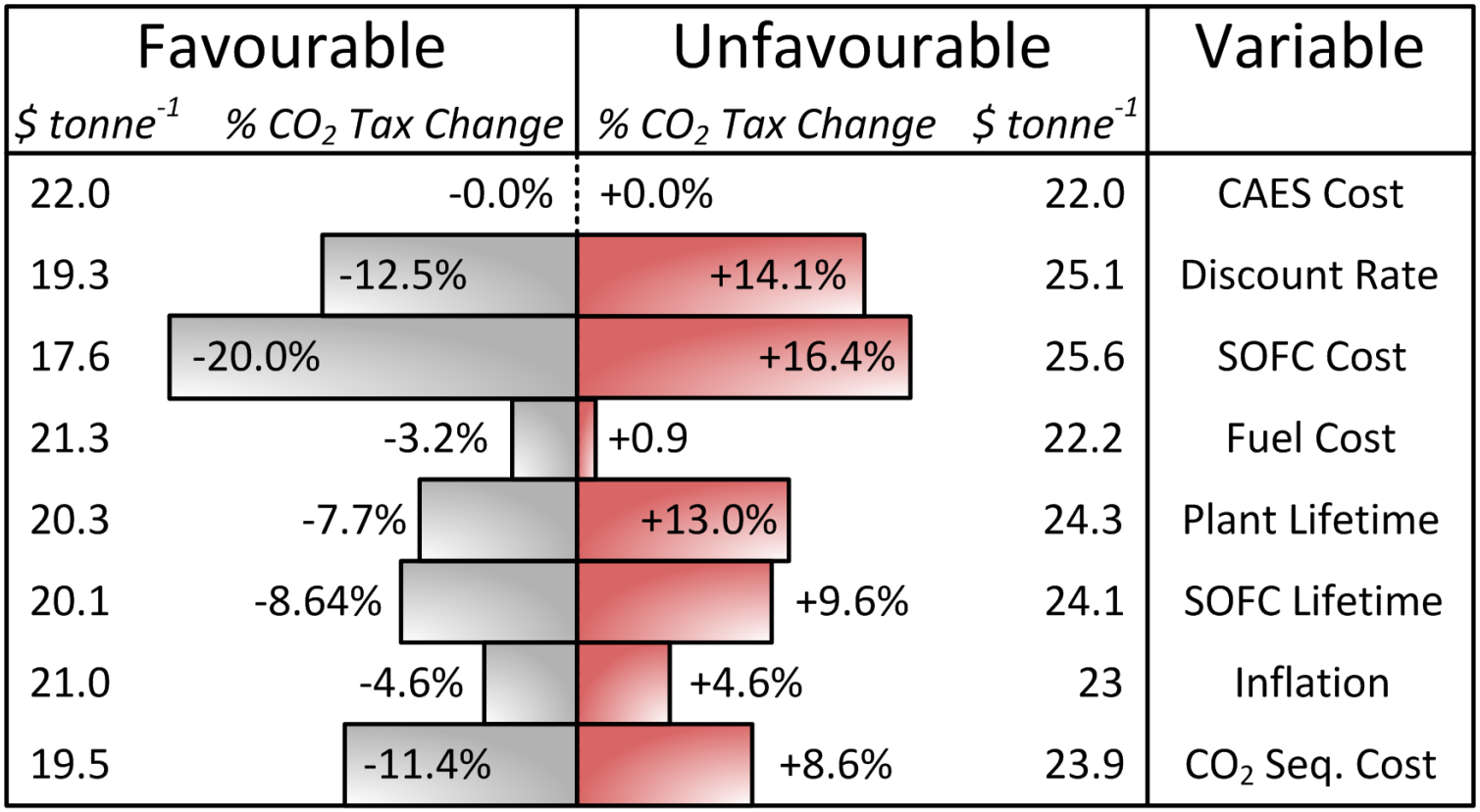
Multi-bars are good at comparing categorical values of data subsets (in this case, life cycle impact results of two plant designs side-by-side)

Do not be afraid of adding data labels or axis labels inside the bar plot for quick reference



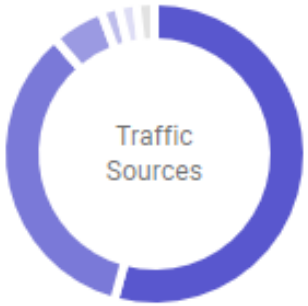
Univariate Data: **Bar Plots**

- Can use horizontal bars if labels are lengthy
  - Prevents then from being squished in the x-axis
  - Be creative! You can display a lot of info on one chart



Univariate Data: **Bar Plots**

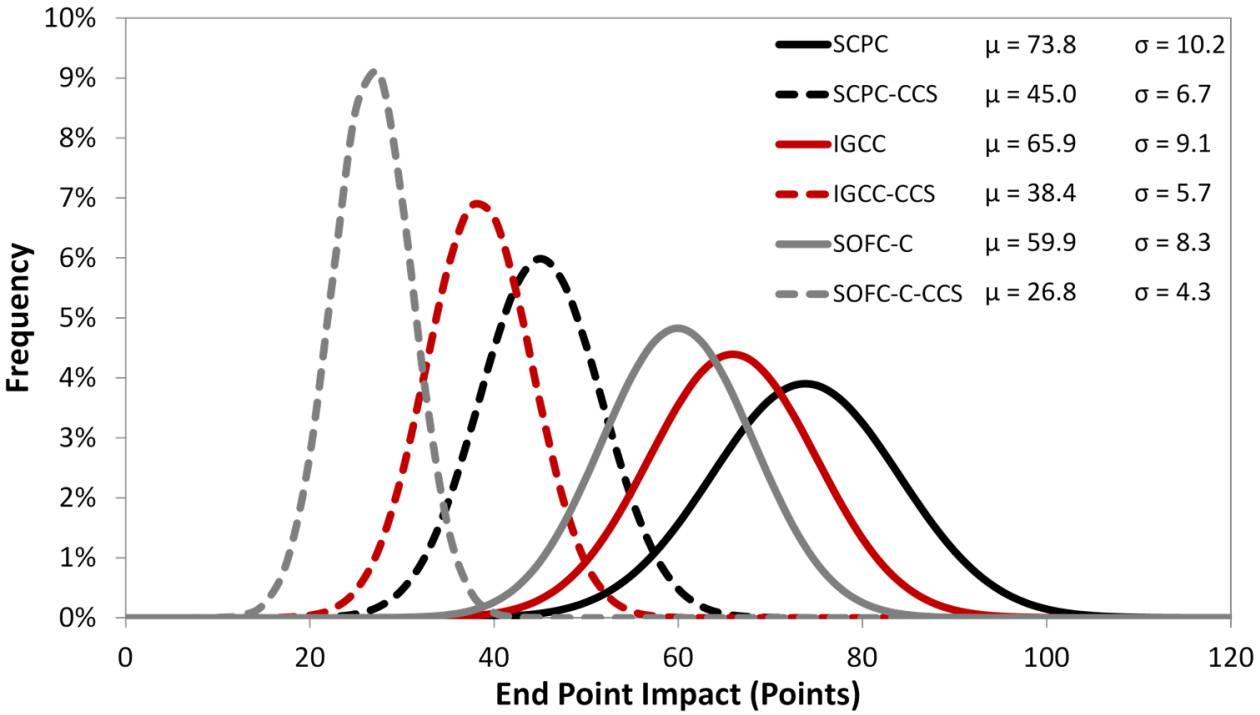
- Don’t use a bar plot when a time series will do better
- Pie charts: the pineapple-on-pizza of plotting tools…



But seriously… Why both?

Univariate Data: **Histograms**

- Histograms lump discrete or continuous variables into subset (bins)
  - Numerous uses – especially in the realms of statistics
  - Various distributions can be fit
  - Normal distributions are most common

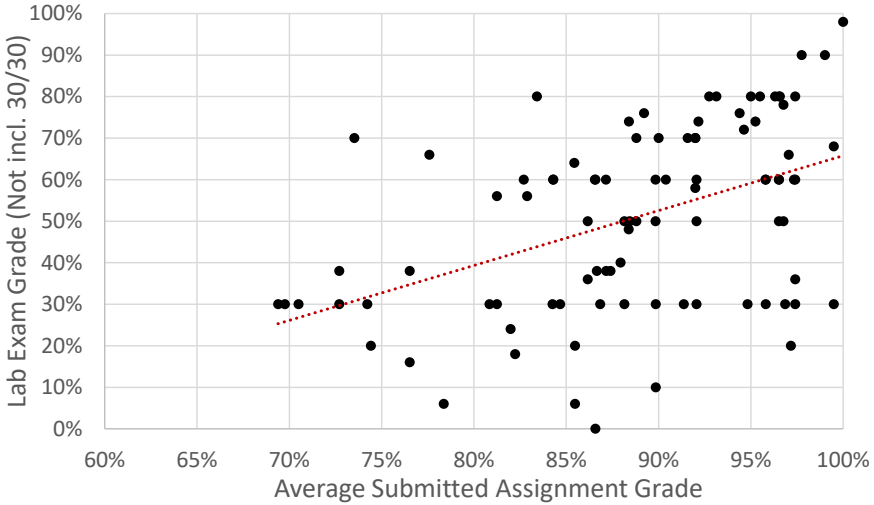


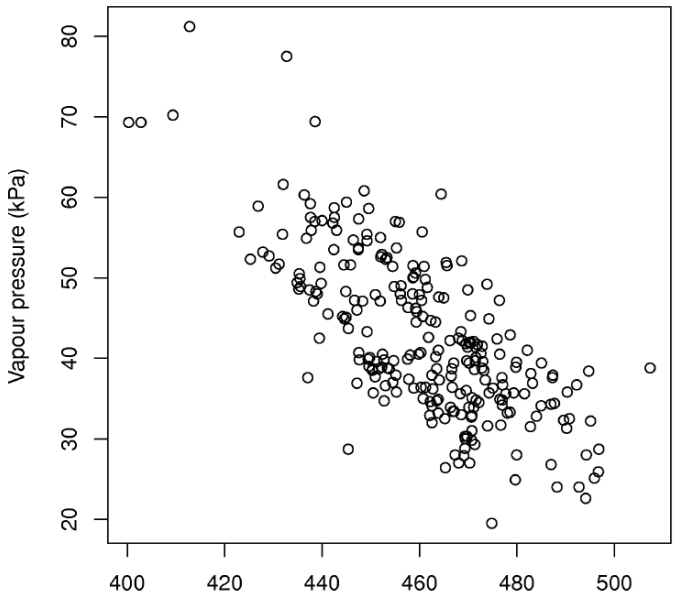
Multivariate Data: **Scatter Plots**

- Used to understand the **relationship** between ≥ 2 vars
  - Collected as **points** on two (or more) axes
  - Each point is the intersection of values on those axes

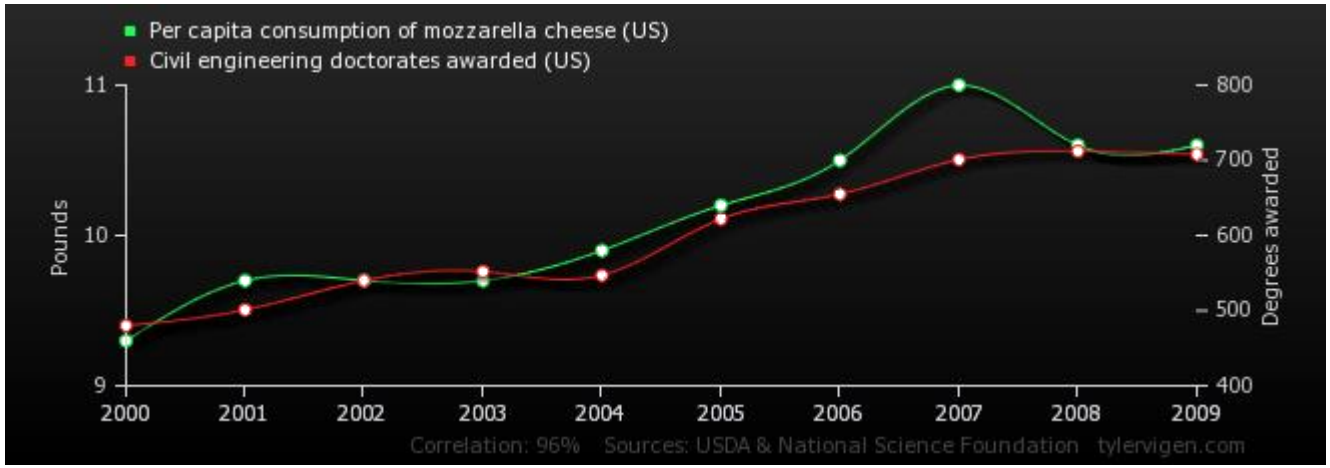
- Intention:
  - Asks the **viewer** to draw a **causal relationship** between the variables
  - Variables should be independent when sampled, but may be related causally (exercise vs. academic results?)
  - CAN be dependent when they are sampled, resulting in a guaranteed relation (final grade vs. midterm grade)
  - Possible to have a relationship that is **not causal**!

Multivariate Data: **Scatter Plots**





Courtesy of ConnectMV

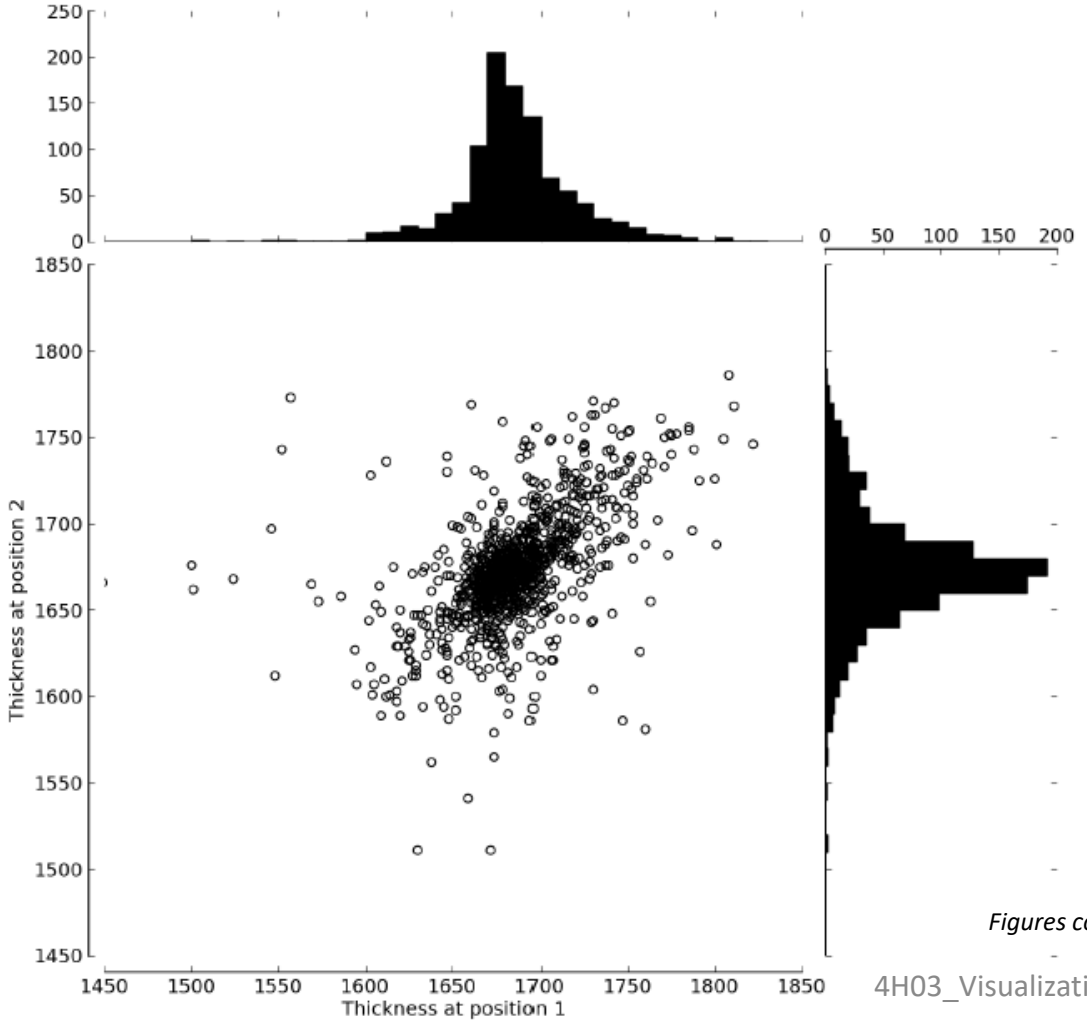


https://tylervigen.com/view_correlation?id=3890

WORKSHOP: Let's talk about all of the things wrong with this figure

Multivariate Data: **Scatter Plots**

- Can add histograms to demonstrate most common outcomes
  - Ex: there seems to be a positive correlation, but deviation from 1675 is low (in both dimensions, really)



Figures courtesy of ConnectMV

Looking Ahead

- Next topic: **regression**
  - But this ain't your grandfather's regression…
    - Linear
    - Derivation of SSE minimization
    - Polynomial and basis functions
    - Multiple dimensions

- Why?
  - Will allow use of eigenspace to derive principal components
  - Basically everything in this course will involve some sort of "regression" of known data to predict known outcomes (supervised learning)

Final Words

- Data visualization is incredibly important
- In engineering, we are often faced with interesting, sometimes HUGE data sets
  - We need to apply the right tool for the job!
  - Never underestimate the power of a **good** plot