Data Visualization Best Practices

Guest Lecturer: Julisa (Jules) Gubbins

About Me

- Consultant at Resultant (Formerly KSM Consulting)
- Serves primarily State & Local Governments
- Specialties: Data Viz Development, Self-Service BI Architecture, and Data Engineering
- Traditional Scientific Background
- Really good at Googling & Ctrl+c/Ctrl+v
- Fun Fact: Married to your TA

Learning Objectives

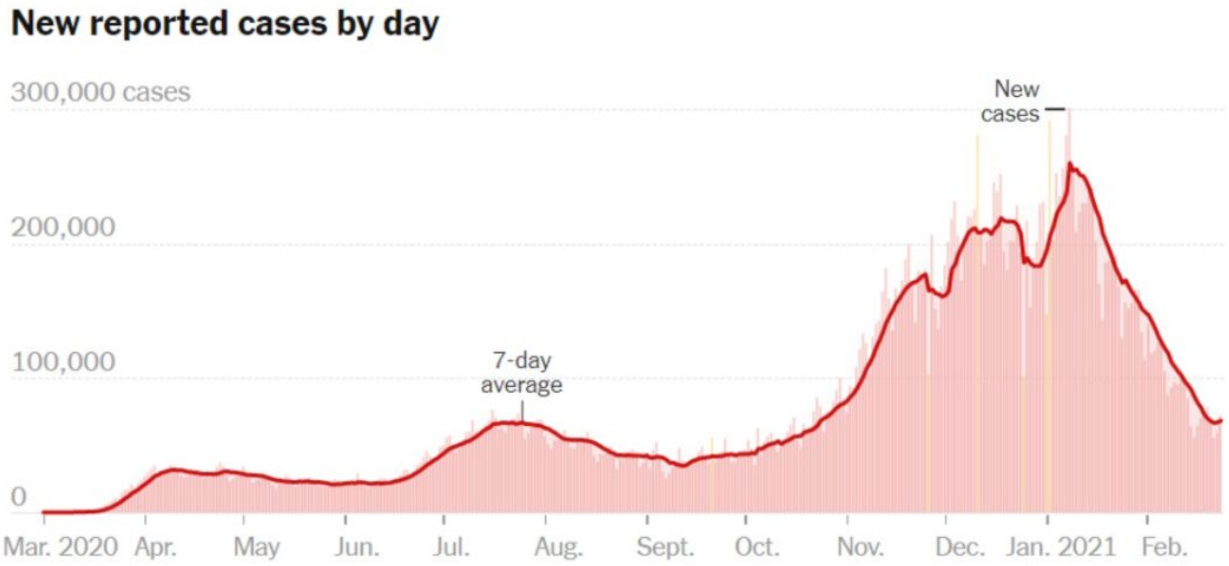
1. Why is Data Visualization Important?
2. How to Develop Quality Data Visualizations
3. How to Avoid Common Data Viz Mistakes
4. What does this look like in practice?

Why is Data Visualization Important?

It allows us to communicate data findings easier.

Why is Data Visualization Important?

It allows us to communicate data findings easier.



Why is Data Visualization Important?

Visualising data makes it easier to identify patterns and trends. Without it, important insights can be lost.

Why is Data Visualization Important?

Visualising data makes it easier to identify patterns and trends. Without it, important insights can be lost.

For example....

Why is Data Visualization Important?

Visualising data makes it easier to identify patterns and trends. Without it, important insights can be lost.

For example....

Why is Data Visualization Important?

(An example you'll hear over and over again....)

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Why is Data Visualization Important?
(An example you'll hear over and over again....)

The summary stats are the same!

Why is Data Visualization Important?

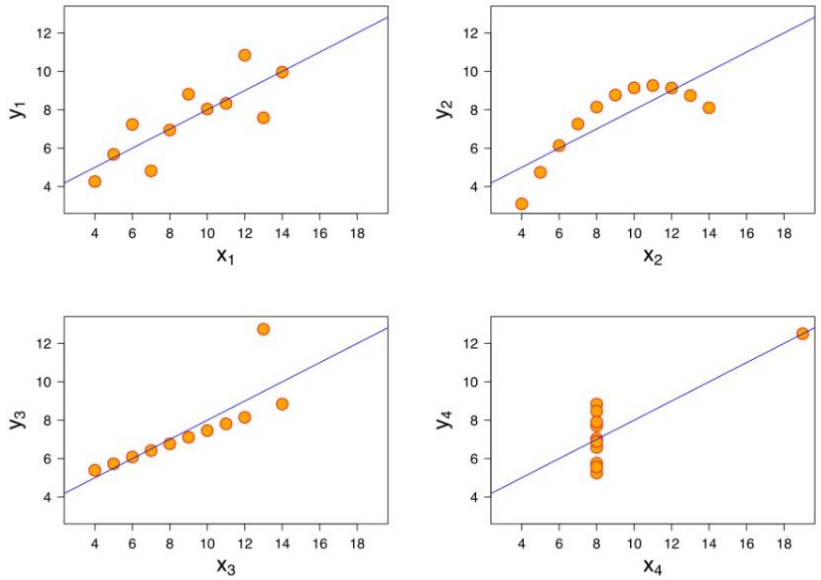
(An example you'll hear over and over again….)

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Why is Data Visualization Important?
(An example you'll hear over and over again….)

The graphs however…..

Why is Data Visualization Important?
(An example you'll hear over and over again….)



How to Develop Quality Data Visualizations

- Answer a question.

How to Develop Quality Data Visualizations

- Answer a question.
- Know your audience.

How to Develop Quality Data Visualizations

- Answer a question.
- Know your audience.
- Use the correct chart/graph type.

How to Develop Quality Data Visualizations

- Answer a question.
- Know your audience.
- Use the correct chart/graph type.
- Stick to conventions.

How to Develop Quality Data Visualizations

- Answer a question.
- Know your audience.
- Use the correct chart/graph type.
- Stick to conventions.
- When in doubt, simplicity.

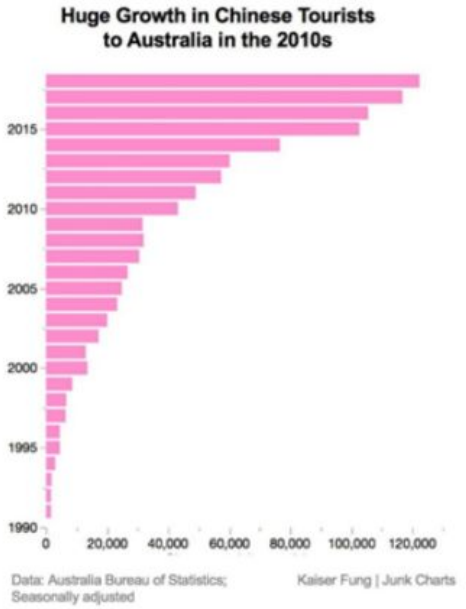
Why does this map seem off?

SOURCE: TOM PATTERSON

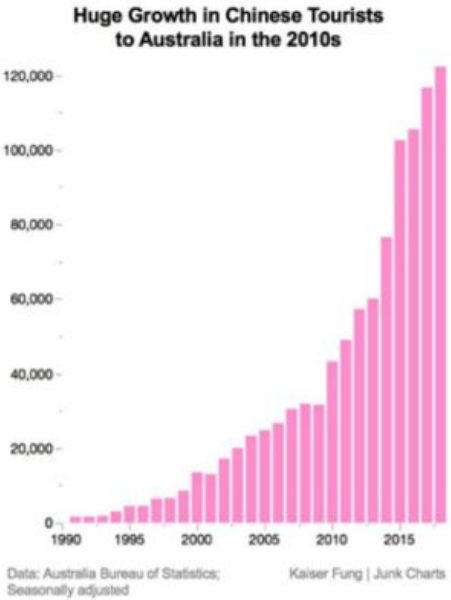
KNOWABLE MAGAZINE

Why does this map seem off?

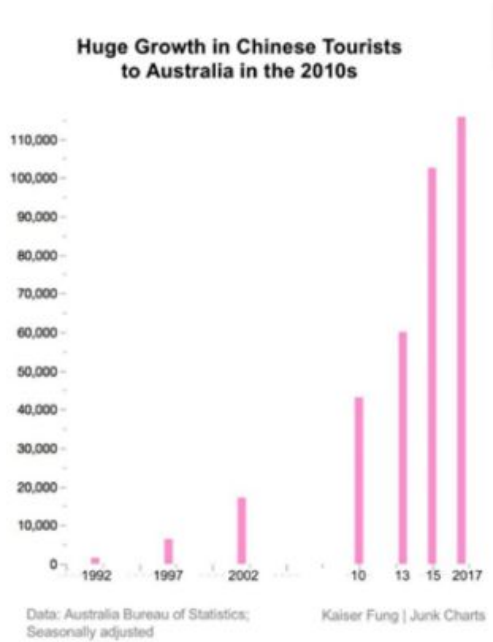
Why does this chart seem off?

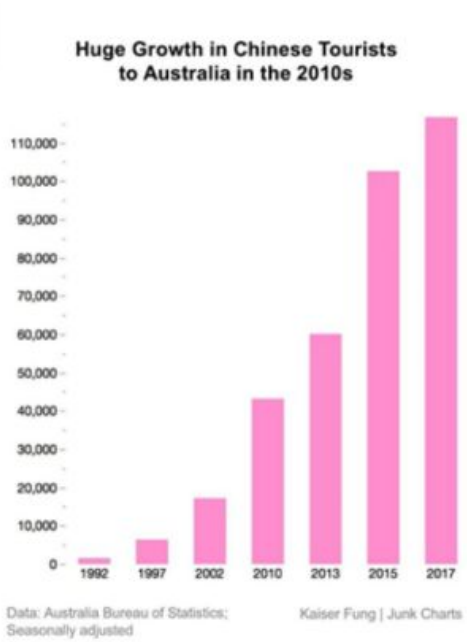


Why does this chart seem off?

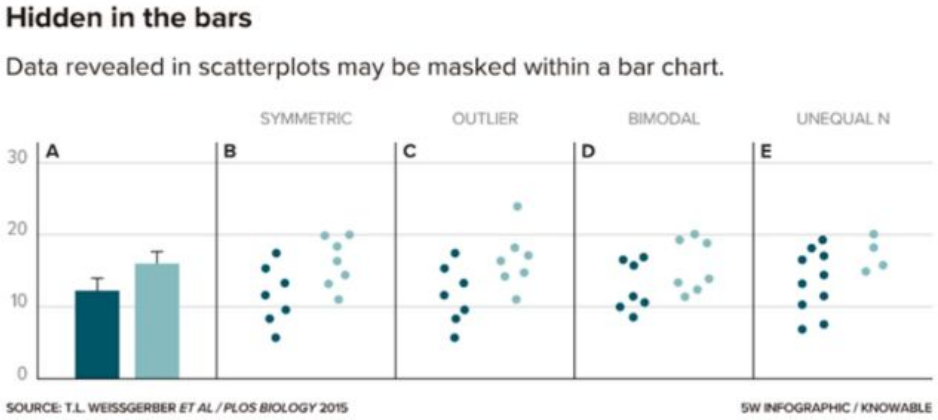


Misleading Visuals



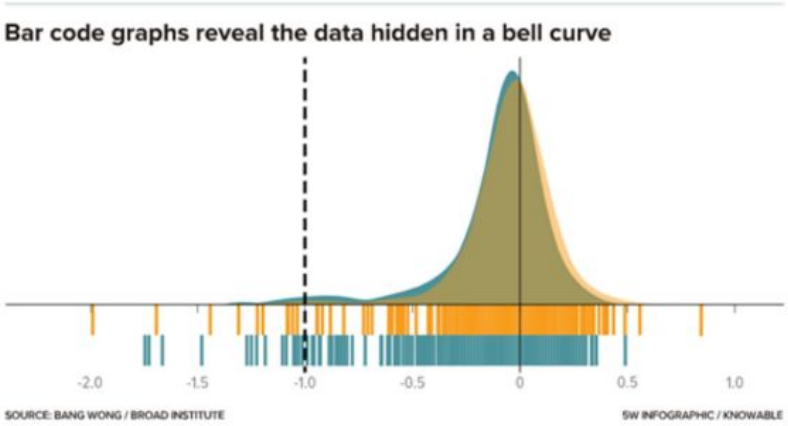


Hidden Data



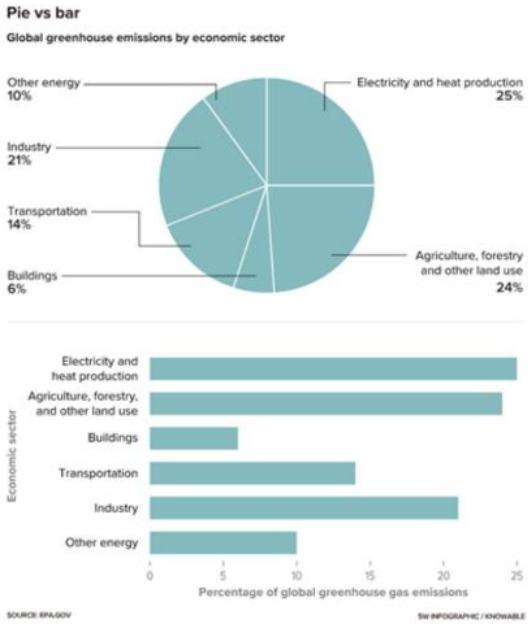
Every one of the four sets of data on the right can be accurately represented by the same bar graph on the left, illustrating how bar graphs can obscure important details about the data, possibly misleading readers.

Hidden Data

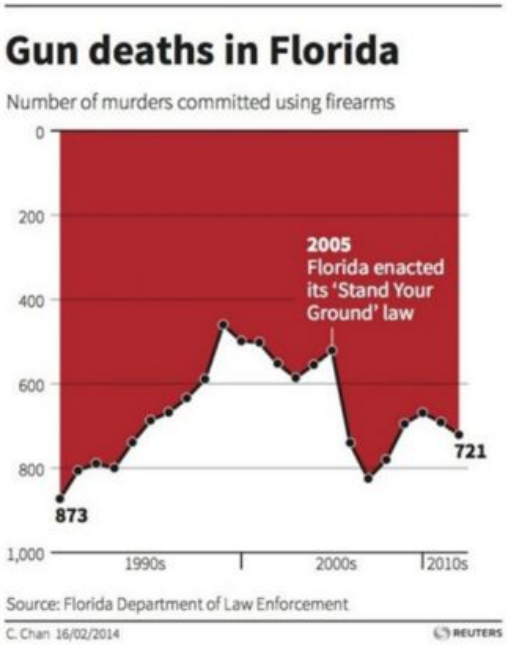


A bell curve draws attention to the distribution of the bulk of the data. But sometimes what's really important is what's on the edges of the data. In that case, a bar code graph may be a better choice. The graphs above show the dependency of cell lines on the gene *FOXA1*; those to the left of the minus-1 reference line need the gene to survive. Those cell lines are difficult to see on the bell curve, but stand out on the bar code.

Unclear Visuals



Misleading Visuals



What could we improve?

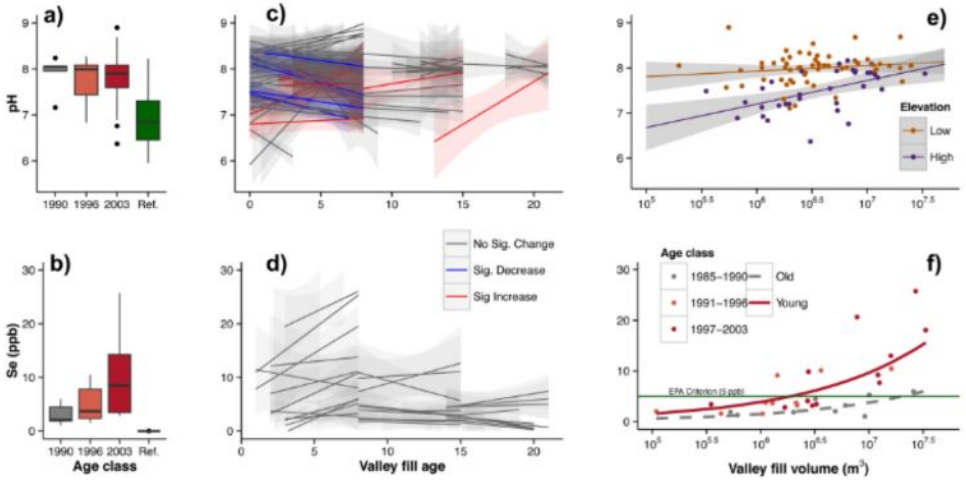
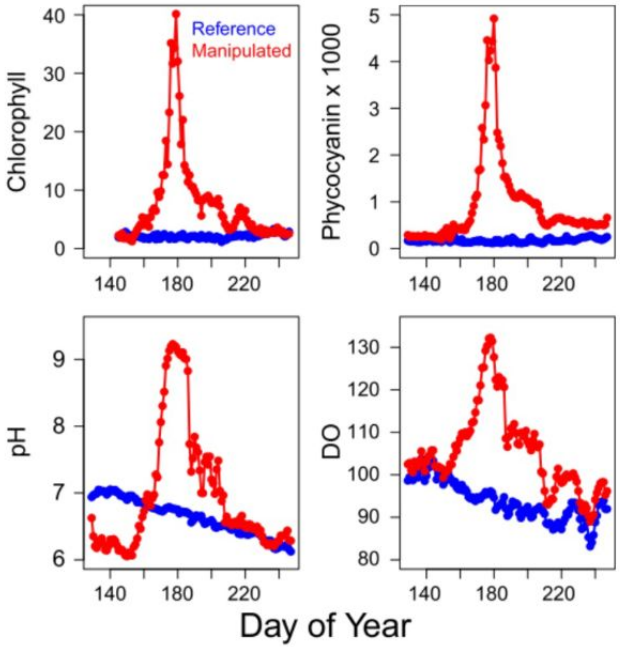


Figure 5. Boxplots for average stream concentration of (a) pH and (b) Se for mines of different ages, shows mean (black bar), 25–75% quanties (box), range (whisker lines), and outliers (black dots). (c) and (d) show regression lines for annual mean stream concentration of (c) pH and (d) Se within individual valley fills. Gray lines indicate no significant trend, whereas red shows significant increases, and blue shows significant declines over time; Se (d) had no trends over time for any valley fills. Relationship between valley fill volume and (e) pH and (f) Se. For pH Valley fill volume is a strong predictor of pH only at high elevation sites ($p < 0.001$ $r^2 = 0.30$) with elevation and volume explaining 37% of the variation in pH ($p < 0.001$). For Se valley fill volume and valley fill age class combine to explain 63% of variance in Se trends between sites, both old (gray), and young (red) sites have the same slope in log–log space, but the *y*-axis shown here is not in log space. There is an order of magnitude increase in valley fill volume showing a 2.6 ppb increase in mean Se concentration.

What could we improve?



What could we improve?

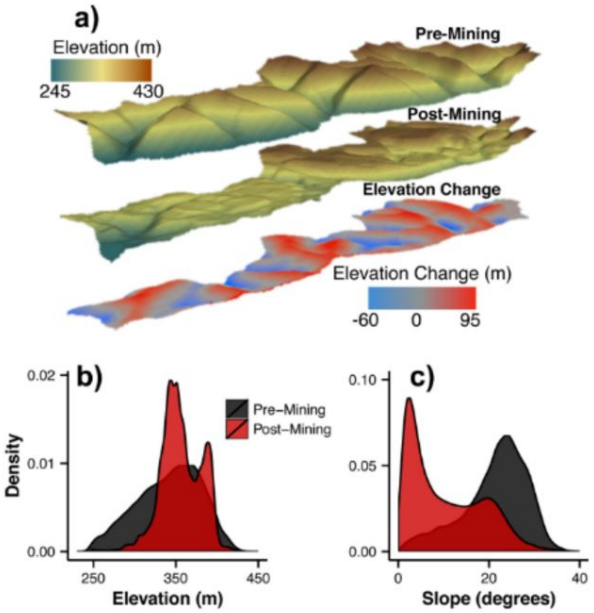


Figure 2. Change in watershed shape from mining at a single valley fill, Connelly Branch in the Hobet Mine Complex, WV. (a) 3-D surfaces highlighting areas with significant elevation change (valley filling in red, ridge cutting in blue). (b) and (c) show probability density functions for

What could we improve?

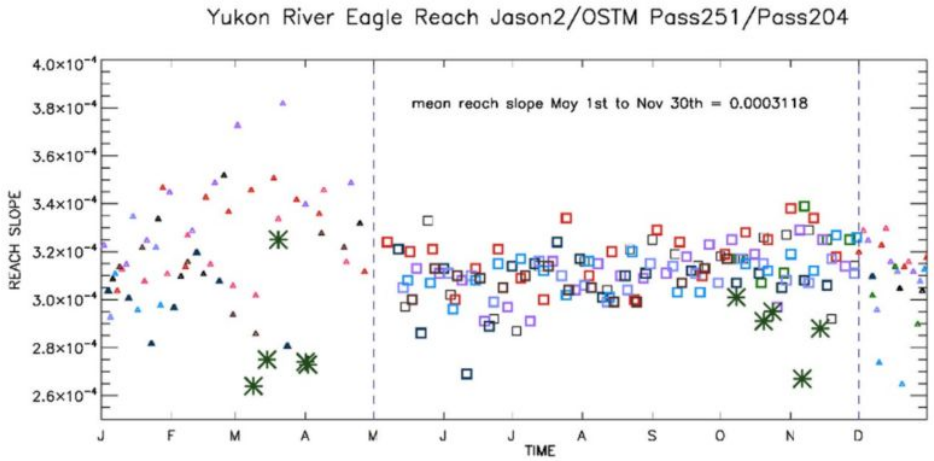


Fig. 7. Seasonal Jason-2 water surface slopes Yukon River at Eagle, Alaska, for the reach between the upstream and downstream river crossings (Pass 251 and Pass 204) over the eight year observation period – each color symbol represents a different year. Because of potential ice/snow penetration, slope estimates during a slightly shorter (December through April) winter season (small triangles) are rejected when estimating a mean reach slope value. Biased ~7% lower than this mean is the average of the multi-year multi-season ICESat water surface slopes (green asterisks) which should not be affected by snow/ice penetration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

What could we improve?

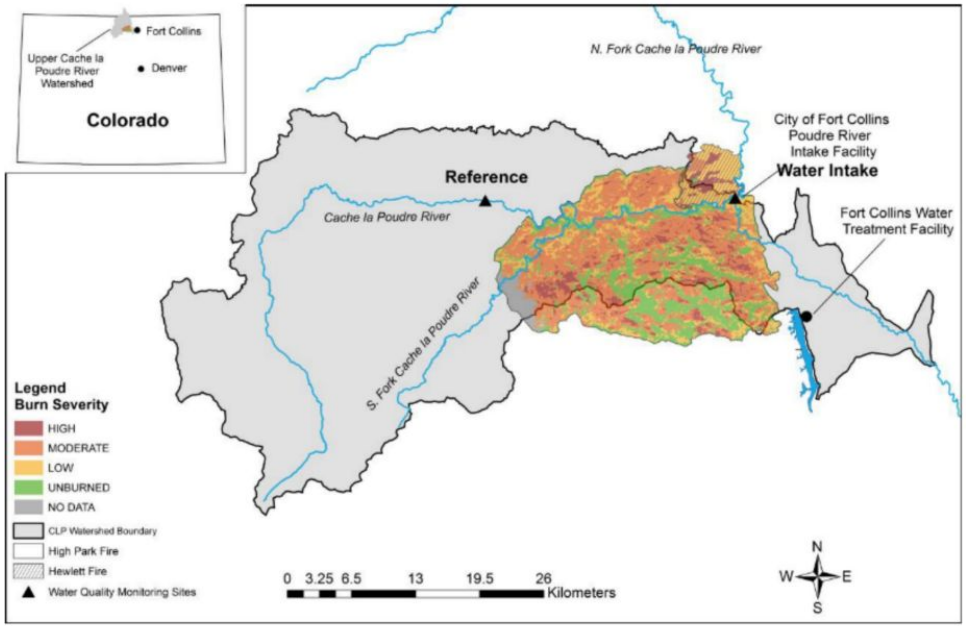


Figure 2.1 CLP River sampling sites include the City of Fort Collins drinking water intake within the High Park fire burn area, and the reference site, upstream and outside of the burned area. The watershed boundary shown includes drainages upstream of the confluence of the North Fork of the CLP River.

What does Data Viz Development look like in practice?

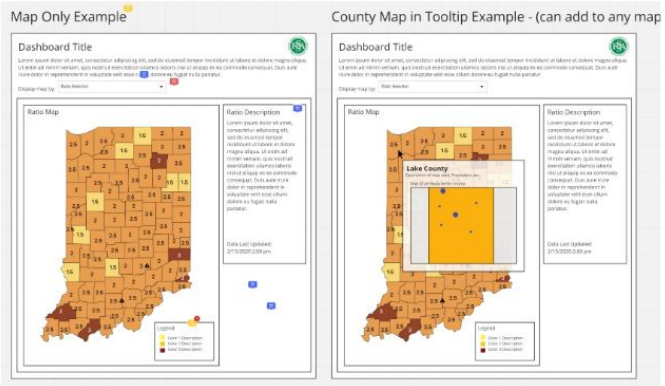
- Requirements Gathering

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development



What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development
- Data Gathering/Cleaning

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development
- Data Gathering/Cleaning
- M.V.P. Development

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development
- Data Gathering/Cleaning
- M.V.P. Development
- Testing

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development
- Data Gathering/Cleaning
- M.V.P. Development
- Testing
- Iterations

What does Data Viz Development look like in practice?

- Requirements Gathering
- Wireframe Development
- Data Gathering/Cleaning
- M.V.P. Development
- Testing
- Iterations

}

REVIEWS, REVIEWS, REVIEWS

DON’T BE LIMITED BY YOUR TOOLBOX

ASSUME SIMPLICITY IN EVERYTHING

Questions?

Guest Lecturer: Julisa (Jules) Gubbins

Send me your visuals! Talk to me about escaping academia! Get reassurance googling syntax is ok! **gubbinsjulisa@gmail.com**