Designing Data Presentations by Using Principles of Data-Driven Storytelling

Francisco José Guerrero Bolaño, Ph.D.

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Sea Grant
University of Wisconsin

WRI
University of Wisconsin

WISCONSIN
DEPT. OF NATURAL RESOURCES

Data Driven Storytelling

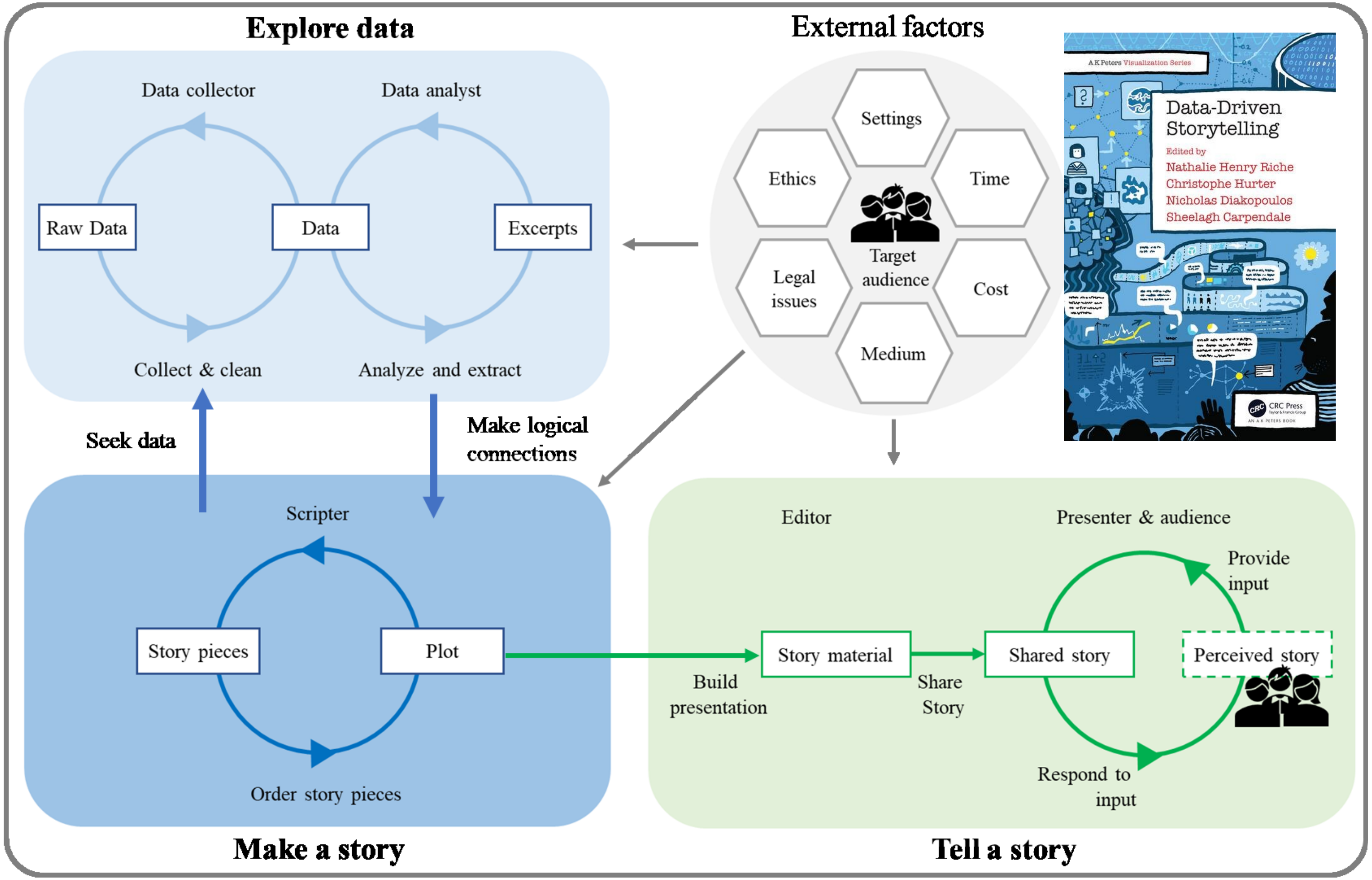
Is a growing field of research known as data-driven storytelling in which computer scientists, journalists and a plethora of other disciplines, are putting together the pieces of the large puzzle that involves effective communication of data-based stories to the public.

The most updated version of the roadmap that could guide our efforts to connect data to the public looks like Figure 1. We are not going to go into the details of this figure and for those interested, you could access the original article at https://hal.inria.fr/hal-01158445/document, written by Lee and co-workers, explaining the core concepts involved in this figure. I want you to pay attention to the boxes and words that may sound more familiar to you. I'm going to make a huge guess and say that probably the top half of this diagram resonates stronger with you.

I am aware of the work that the DNR has been doing in terms of building an architecture to host the development of multiple shiny apps, including those developed by Alex Latzka for PhosMER or by Matt Diebel for the Long-Term Trends of Water Quality. In a recent conversation with Aaron Fisch, I learned about the exciting amount of progress that is being done in that direction with the support of the water quality bureau.

But where I want to drive your attention to is that while for many of us shiny apps might sound like data visualizations, there is a huge difference between a data visualization and a data-driven story. And such difference emerges from the bottom half of this figure. On the one hand, a data visualization can be build by a data analyst to offer supporting evidence. Building data visualizations require data analysis and coding skills nowadays. But telling a story with data requires additional steps that involve developing a story plot and a large amount of time spent on developing visualizations that follow the plot as it thickens and resolve.

Figure 1 | **The storytelling process: from the story idea to visually shared stories***



Explore data

Data collector

Data analyst

Raw Data

Data

Excerpts

Collect & clean

Analyze and extract

Seek data

Make logical connections

External factors

Settings

Ethics

Time

Target audience

Legal issues

Cost

Medium

Scripter

Story pieces

Plot

Order story pieces

Make a story

Editor

Presenter & audience

Build presentation

Story material

Share Story

Shared story

Perceived story

Provide input

Respond to input

Tell a story

*Chevalier, F. et al., 2018. From analysis to communication: supporting the lifecycle of a story. In Henry et al. (Eds.) Data-Driven Storytelling. CRC Press. Taylor & Francis Group.

In the following pages you will find a compilation of key definitions for the guiding principles of data-driven storytelling. A key difference to keep in mind is that between data visualization and data-driven stories. On the one hand, a data visualization is “the use of interactive, dynamic, and responsive visual representation to amplify cognition”. On the other hand, stories that are data-driven start from a narrative that either is based on or contains data and incorporates this data evidence, often portrayed by data graphics, data visualizations, or data dynamics, to confirm or augment a given story. So, data-driven storytelling is the development of strong narrative that contains data visualizations.

You also will find some examples of an editorial analysis on the visualizations incorporated in the Long-Term Trends in River Water Quality across Wisconsin-Shiny App. This Shiny-App was created by Matt Diebel in an extraordinary effort a few years ago. The analysis presented was discussed with Matt during my time at the DNR. The editorial analysis of the data visualizations presented here, besides serving as an example of the myriad of aspects we need to account for as we build visual data stories, it also tries to illustrate that it is perfectly understandable to have blind spots when you are mostly a one-man band. And this is what I want to highlight here. We have people in our organization that have been doing the work of data analyst, storyteller, editor, and web developer because they are exceptionally talented and highly motivated.

Designing data presentations by using principles from data-driven storytelling

Key definitions

Data visualization: “the use of interactive, dynamic, and responsive visual representation to amplify cognition”(Henry, et al., 2018, p.7)

Interactive: “interactive computer systems are those where a person acts and the computer system responds to his and her actions and vice versa”(Henry, et al., 2018, p.7)

Dynamic: “on a computer, many sorts of dynamics are possible such as animations, replays, stop motion, etc., that while possible being very useful un telling a story might not be actually interactive”(Henry, et al., 2018, p.7)

Amplify cognition: “visualizations can enhance cognition by assisting memory and by easing comprehension (e.g., by creating representations that appropriately leverage perception)”(Henry, et al., 2018, p.8)

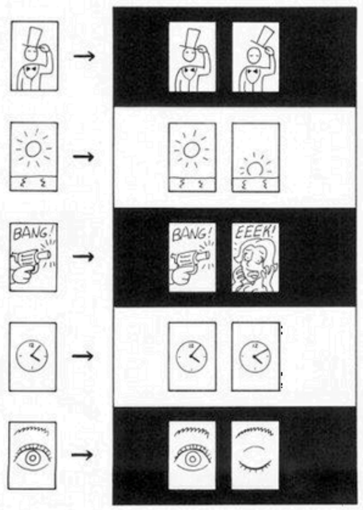
Sequential art (a.k.a. comics): “juxtaposed pictorial and other images in deliberate sequence, intended to convey information and/or to produce an aesthetic response in the viewer”(McCloud, 1994, p.9)

Framing: “Frames are the (typically unconscious) conceptual structures that people have in their brain circuitry to understand a particular issue.” (Lakoff, 2010, p. 71).

“Data-driven storytelling is a movement towards data-driven stories which is apparent in both the data visualization research community and the professional journalism community. It has the potential to form a crucial part of keeping the public informed (i.e., the democratization of data)” (Henry, et al., 2018, p.3).

Data-driven stories

“Stories that are data-driven start from a narrative that either is based on or contains data and incorporates this data evidence, often portrayed by *data graphics*, *data visualizations*, or *data dynamics*, to confirm or augment a given story. A data-driven story often incorporates the visual data representations directly into the presentation of the story. These stories can enhance a narrative with capabilities to walk through visual insights, to clarify and inform, and to provide context to visually salient differences”(Henry, et al., 2018, p.8-9).



“Taken individually, the pictures above are merely pictures. However, when part of a sequence, even a sequence of two, the art of image is transformed into something more: the art of comics” (McCloud, 1994, p.7)

Sequential art (a.k.a. Data Comics)

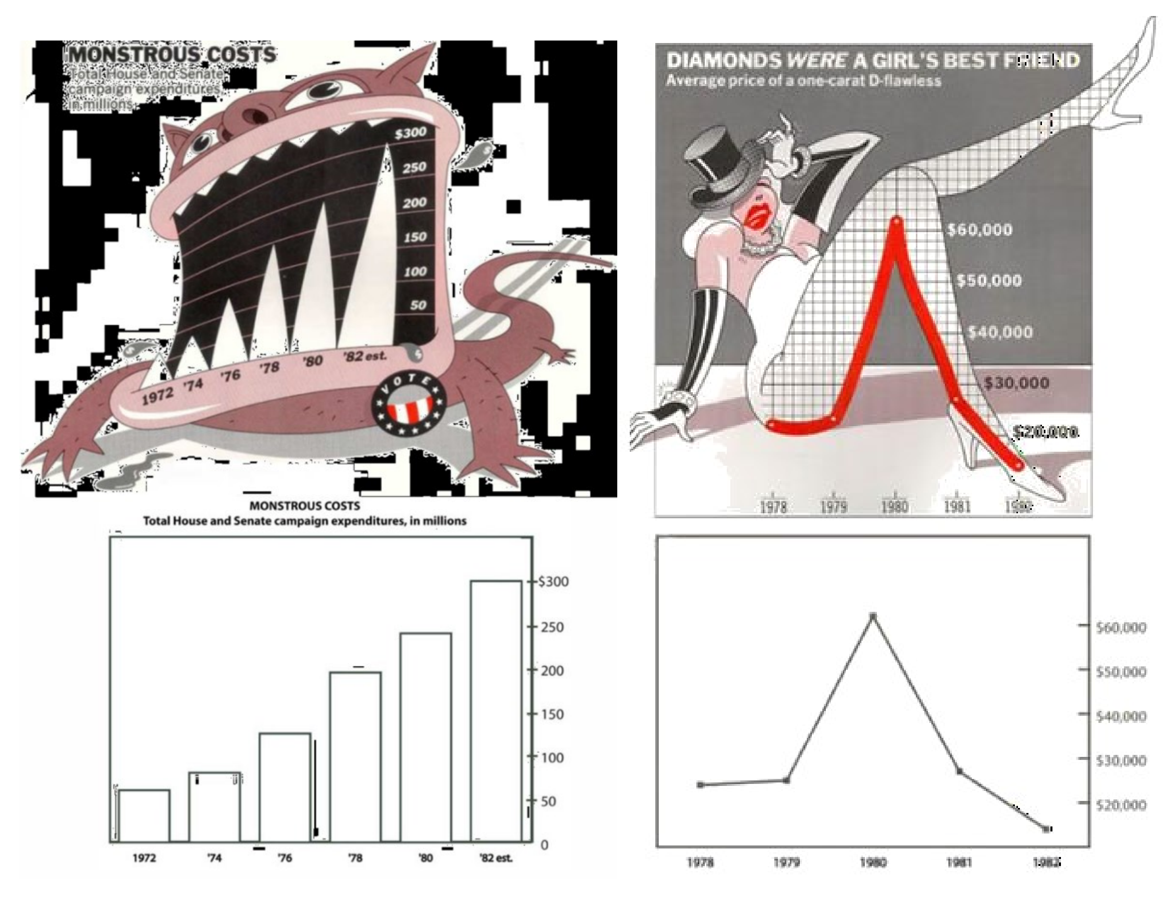
“A new genre inspired on how comics function to convey information in data, tell a story, and communicate through visualization.”(Bach, et al., 2018, p.7). “Data comics can lead to engaging visual and narrative artifacts by combining verbal and visual content, leveraging each one’s strength as well as their synergy: by delivering one message at time and creating an explicit guided tour for the observer, by leveraging the richness of data visualizations to provide visual evidence for facts, and by allowing factual visualizations to blend with other styles and types of pictures and narratives.”(Bach, et al., 2018, p. 12).

In short, “Data Comics is a visual storytelling method based on *sequential images* consisting of data-driven visual representations. Its purpose is to build engaging narratives about data.” (Zhao et al., 2015, p.3).

Framing effects in data-driven storytelling

“How data is framed or presented can significantly affect interpretation.” A data visualization project involves specific techniques connected to different (editorial) layers such as, the data, visual representation, annotation, and interactivity, in such a way that certain interpretations of the data are more probable. (Hullman & Diakopolous, 2011).

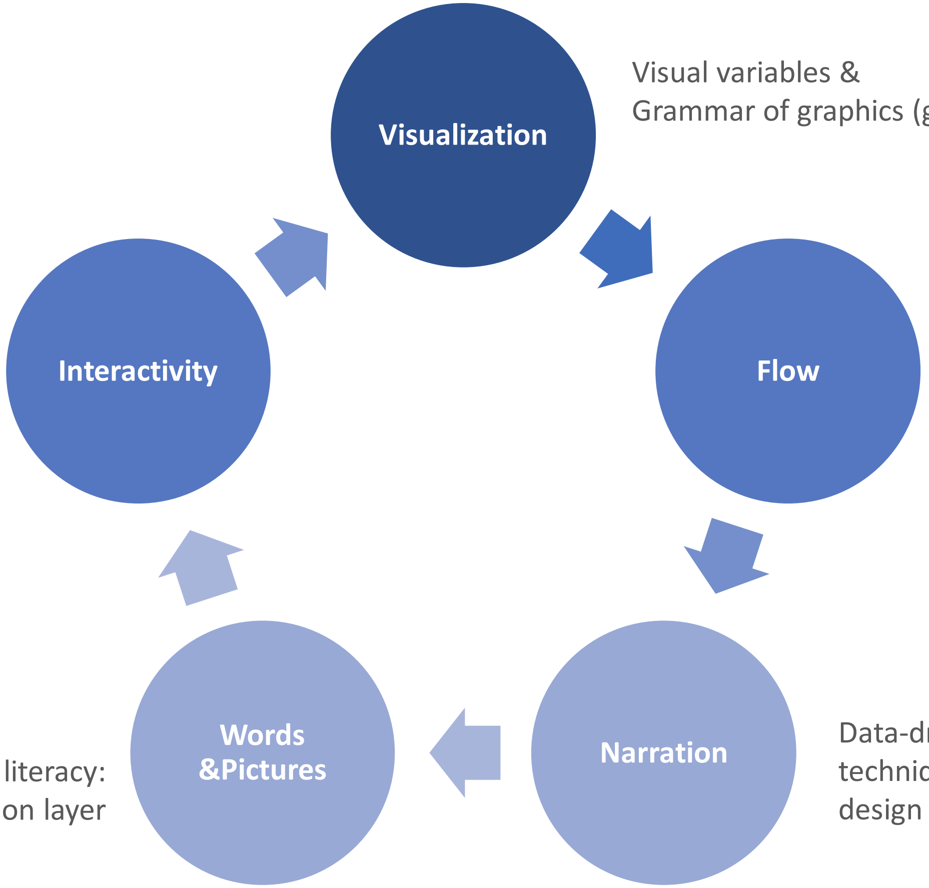
Box 1 | **Data visualization and framing *in extremis***



"Many experts in the area of chart design, such as Edward Tufte, criticize the inclusion of visual embellishment in charts and graphs (and advocate for a minimalist approach); their guidelines for good chart design often suggest that the addition of *chart* junk to a chart can make interpretation more difficult and can distract readers from the data. We found that people's accuracy in describing embellished charts was no worse than for plain charts, and that their recall after a two-to-three-week gap was significantly better. Although we are cautious about recommending that all charts be produced in this style, our results question some of the premises of the minimalist approach to chart design." (Bateman et al., 2010).

Framing can be as subtle as a specific word choice or as explicit as pictorial embellishments, that in some cases are referred as to "chart junk". However researchers have found evidence that such embellishments provide the viewer with a cognitive advantage (e.g., Bateman et al., 2010; Box 1).

Five editorial layers in data-driven storytelling



Visualization

Visual variables &
Grammar of graphics (ggplot2)

Flow

A deeper understanding
of sequence in narrative
visualization

Narration

Data-driven storytelling
techniques & narrative
design patterns.

Words
&Pictures

Data-visualization literacy:
the annotation layer

Interactivity

There are a lot of things that we can do in terms of improving our data-driven storytelling capacities. Data-driven stories require proficiency in both narrative and visualizations and those elements must go together for the story to flow. You would be surprised by the amount of editorial work that these stories need. You need to be aware how visual elements like position, size, transparency, and color interact to communicate your message (See Box 2).

Box 2 | **Visual variables and their syntactics***

		Levels of organization				
Visual variable	**Description**	Associative	Selective	Nominal	Ordinal	Numerical
Position	The placement of some representative graphics within some **display space**, be it one-, two-, or three-dimensional.	Y	Y	G	G	G
Shape	Shapes are **graphic primitives** that represent data by means of points, lines, areas, volumes, and their compositions.	Y	N	G	P	P
Size	Size easily maps to **interval and continuous data** variables, because that property supports gradual increments over some range.	N	Y	G	G	G
Orientation	It describes how **a mark is rotated in connection with a data variable**. The best marks for using orientation are those with a single natural axis. Orientation is typically manipulated in flow maps to represent directionality of flows.	Y	Y	G	M	M
Color hue	Hue provides what most think as of color: the **dominant wavelength** from the visual spectrum.	Y	Y	G	M	M
Color value	Variation in color value results in the **perception of shading** (high/low emission or reflectance of energy) and it is sometimes referred to as "lightness". (it can be conceptualized as the amount of black in a symbol).	N	Y	P	G	M
Texture	Texture can be considered as a **combination of many of the other visual variables** including color, orientation, and size. Due to advances in modern printing and digital devices, visual variables are now manipulated via changes in color.	Y	Y	G	M	M
Color saturation	Saturation is the level of hue relative to gray and drives the **purity of the color** to be displayed. Marks that are more saturated appear brighter than less saturated marks at the same luminance.	▨	▨	P	G	M
Arrangement	It describes the layout of graphic marks going **from regular, grid-like structures, to irregular, cluster-like structures**. Irregular and cluster arrangements tend to rise to figure.	▨	▨	G	P	P
Crispness	It describes the sharpness of the boundary of a symbol, also referred as to "**fuzziness**". Perhaps the most effective visual variable to **represent uncertainty** in point symbolization.	▨	▨	P	G	P
Resolution	It describes the **precision** at witch a mark is displayed and relates to the **removal of detail** in a visualization as the complexity of the real world is abstracted to fit into the visual frame.	▨	▨	P	G	P
Transparency	It describes the amount of **blending between a symbol and its background;** attention is drawn to opaque marks instead. Transparency also has higher potential for **uncertainty depiction** than color saturation**.	▨	▨	G	G	P

Y=yes; N=no; G=good; M=marginal; P=poor; hatched=n/a

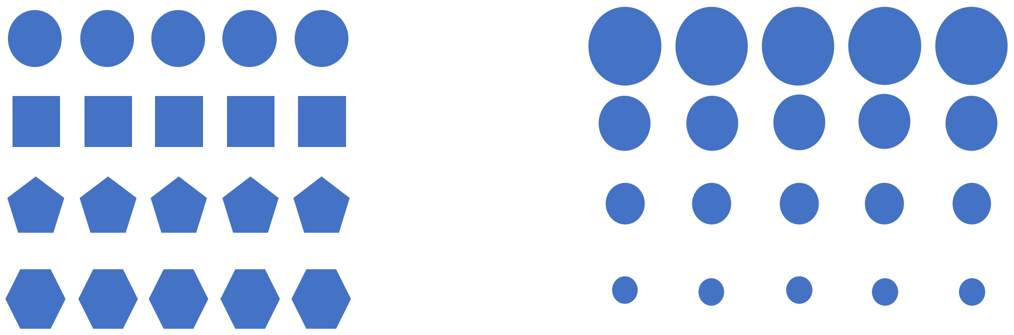
*Modified from Roth, 2017 with notes from Ward et al., (2010).

**Kinkeldey et al., 2014. The cartographic journal, 51(4)

Perceptual basis of the visual variables and their syntactics*

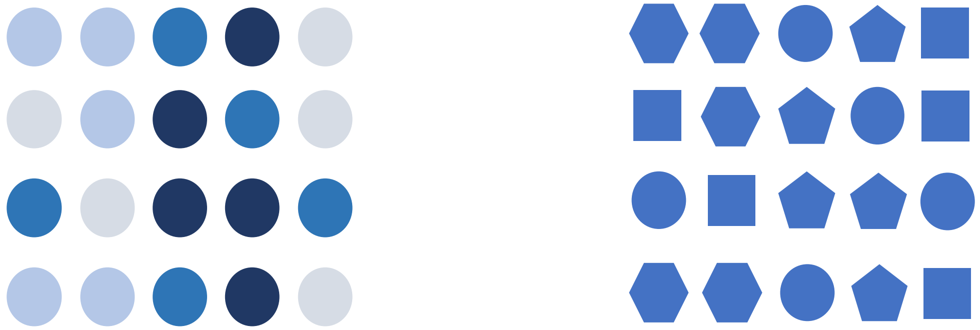
“Bertin (1967/1983) postulated four ways in which visual variables in maps and other visualizations are processed in the eye-brain system, and therefore inform use of one visual variable over others. These four properties were described as level of organization.” (Roth, 2017; p. 5) (Box 2).

Associative perception
“Variations in the visual dimension are perceived with equal weight, allowing for the eye to perceive all map symbols with the same variation as a group (i.e. associated). Location, shape, orientation, color hue, and texture are associative visual variables. In contrast, a dissociative visual variable inhibits attention to other visual variables that may vary in the visual scene. Size and color value are dissociative visual variables.” (Roth, 2017; p. 6)



“With an associative variable, like shape, the eye is not drawn to one mark over another; as a result the eye is likely to see a series of horizontal rows. With a dissociative variable, like size, the eye is drawn to the larger sizes; as a result the eye perceives a vertical gradient, rather than a set of horizontal rows” (Roth, 2017, Fig. 2C, p.7)

Selective perception
“Variations in the visual dimension are perceived with unequal weight, allowing for the eye to focus individually upon each variation of the visual variable across the visual scene and ignore the other variations. In other words, is relatively easy to isolate visually the distribution of a particular symbol across the visual space. With the exception of shape, most visual variables are selective” (Roth, 2017; p. 6)



“With an selective variable, like color value, the eye is drawn to one mark over another; as a result is easier to see the distribution of the darker symbols on the left than the distribution of the pentagons on the right, despite the pair of figures encoding the same information.” (Roth, 2017, Fig. 2F & 2G, p.6-7)

Ordered, and quantitative perception

“Variations in ordered visual variables are perceived in the sense of “more” or “less”, but it is not possible to quantify those differences based on the visual property. Color value, color saturation, crispness, resolution, and transparency are examples of strongly ordered visual variables. Quantitative perception extends ordered perception, allowing the estimation of numerical values from variations in quantitative visual variables. Bertin believed quantitative perception to be restricted to position and size” (Roth, 2017; p. 6).

“Unordered visual variables –such as color hue, orientation, and shape- are appropriate for encoding nominal information. Visual variables that are ordered but not quantitative –such as color value, color saturation, crispness, resolution, and transparency- are appropriate for encoding ordinal information. Finally, visual variables that are quantitative –such as location and size- are appropriate for encoding numerical information but also can be applied for ordinal and nominal information given their visual dominance” (Roth, 2017; p. 7).

Visual syntactics for shinny app

Resolution G

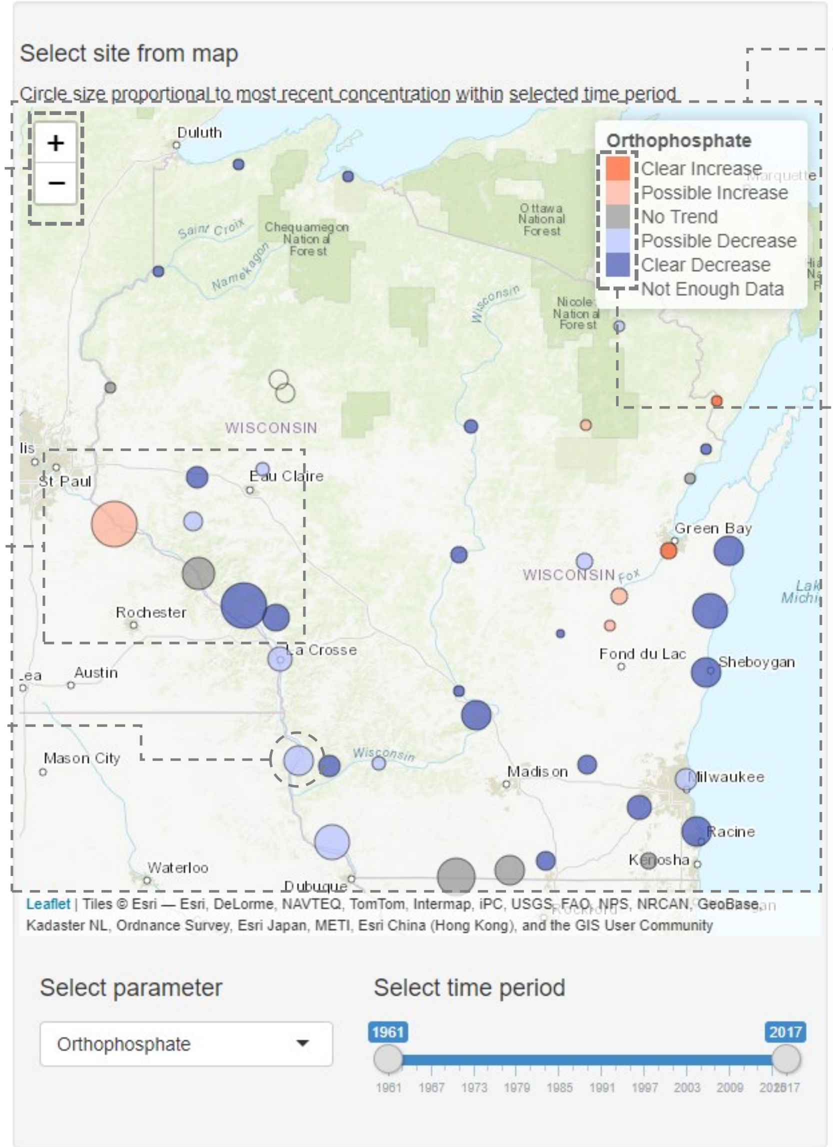
Changes in resolution are possible due to *the interactive interface.* These changes allow for the selective display of varying levels of geographical details (e.g., county names, boundaries) without an overwhelming display of content at first.

Size G

Changes in "bubble" size indicate the length of the record within the selected time period (numerical variable)X**proportional to the concentration**

Transparency G

The amount of transparency draws attention to the superposition of the sampling locations over landscape, while emphasizing location (more opaque). Color values applied on the background, which make it appear "lighter" and more luminous, supplement the effect.



Select site from map

Circle size proportional to most recent concentration within selected time period

Orthophosphate

- 🟥 Clear Increase
- 🟧 Possible Increase
- ⬜ No Trend
- 🟦 Possible Decrease
- 🟦 Clear Decrease
- Not Enough Data

Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

Select parameter: Orthophosphate

Select time period: 1961 – 2017

Position G

A geographical coordinate system informs about the spatial location of monitoring sites and suggest a pattern of coverage (associative).

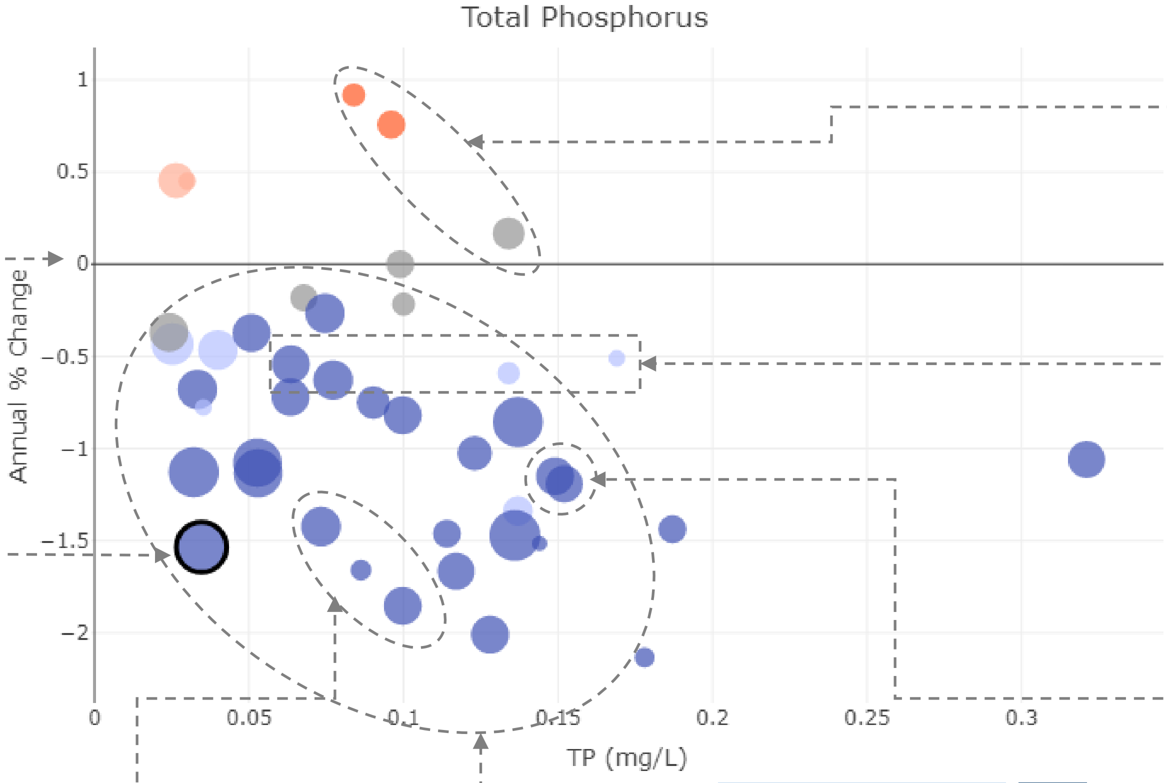
Color hue G M

Color hue allows for the matching of each location with several possibilities of temporal trends (associative).

A relationship between color hue and the direction of the trend is based on an ordinal scale (positive vs. negative). Not strongly recommended.

G=good; M=marginal; P=poor

Concentration summary



Position G

A set of coordinated axes and an emphasized line at “0” on the way axis informs about main directionality in the plot (positive vs. negative).(Associative). Because of the presence of a grid it is possible to make a quantitative assessment of the directionality of the trend.

Texture G

A change in contour width gives a particular data point a different textural appearance, in this case highlighting a selected site (nominal variable)

Size G

Changes in “bubble” size indicate the length of the record within the selected time period (numerical variable)

Arrangement G

Clustering of data points plays an important role in this figure since informs about the number of locations within a certain trend category and also below a specific TP concentration.

Color hue G

Color hue allows for the matching of each location with several possibilities of temporal trends (associative).

Color value M

A relationship between color value and the sign of the trend is apparently suggested. However, points with a similar %change have different color values, implying a “hidden” variable being represented by changes in color.

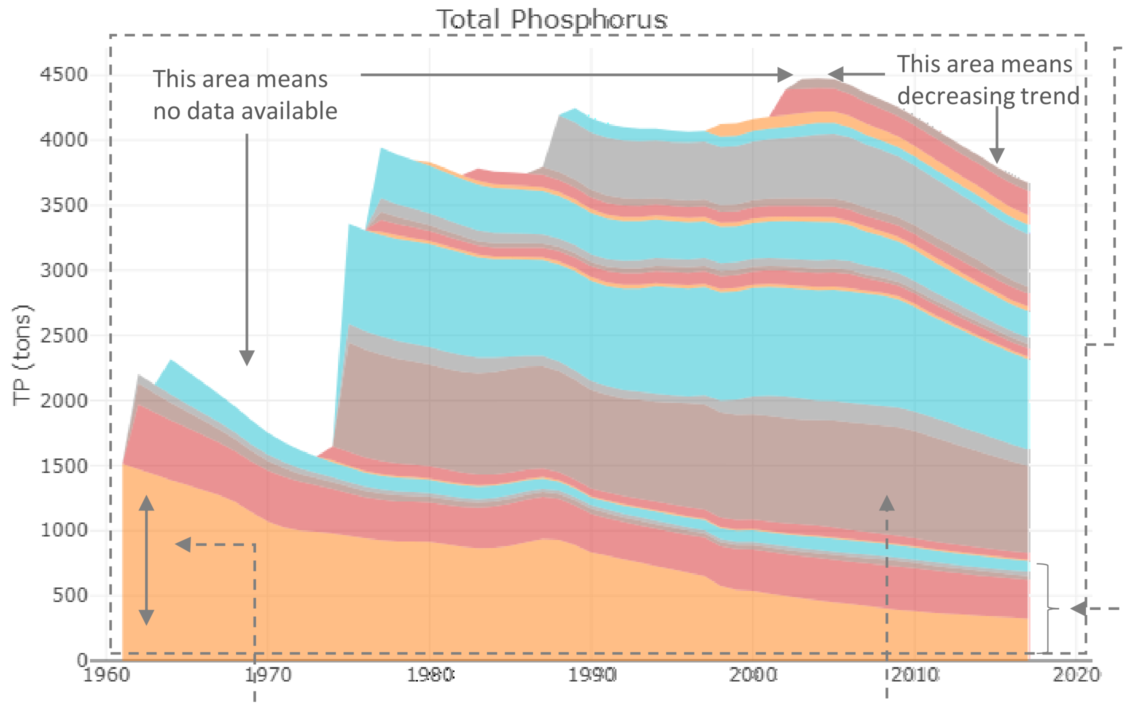
Transparency G

Transparency aids to separate locations with similar trends and values.

G=good; M=marginal; P=poor

Plot explanation: This plot summarizes the status and trends of the selected parameter at all sites over the selected time period, with the selected site highlighted. The horizontal axis is the flow-normalized concentration in the most recent water year of the selected time period. The vertical axis is the annual percent change in the parameter across the selected time period (Note that for non-linear trends, this value may not represent all parts of the time period). The size of each circle is proportional to the length of the record within the selected time period. **(Intensinty of the color is also related to how well the model represents the data).**

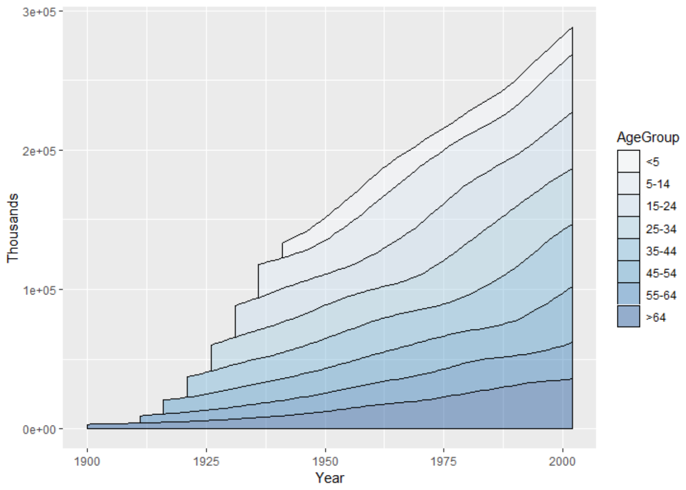
Flux summary



Position G M

Besides the coordinated axes that allow for a quantitative interpretation of the data, the vertical stacking of area features informs about the cumulative nature of the dataset displayed.

White space around the stacked areas plays an ambiguous role: after ~2005, it highlights an overall decreasing trend in TP load, but before that year, it means "no data available". Furthermore, the cuts for the beginning of each series are not "clear" (like in the example on the right) suggesting a false sense of trend instead of the progression in the amount of data available.



An stacked area plot from dataset "uspopage" from the R-package (gcookbook), modified to introduce temporal discontinuities in the record.

Size G

Like size and length, area allows for the representation of quantitative variables, in this case, TP load over time.

Transparency G

Transparency is used to reduce visual cluttering and to allow a better quantitative perception of the data by showing the gridlines in the background.

Color hue M

Each color band represents a sampling location, organized by length of the record. Although color hue is appropriate for the representation of nominal variables, the repeating color palette does not necessarily informs about the identity of each location.

G=good; M=marginal; P=poor

Plot explanation: This plot shows trends in the flux (load) of the selected parameter at all non-nested sites. The width of each colored band in a particular year is the flux at one site and the total height of all bands is the sum of fluxes at all sites with flux estimates for that year. In this plot, sites are sorted by length of record, data gaps are filled by interpolation, and fluxes for sites whose watersheds are partially in other states are reduced by the fraction of watershed outside of Wisconsin.

Parameter comparison

Position G

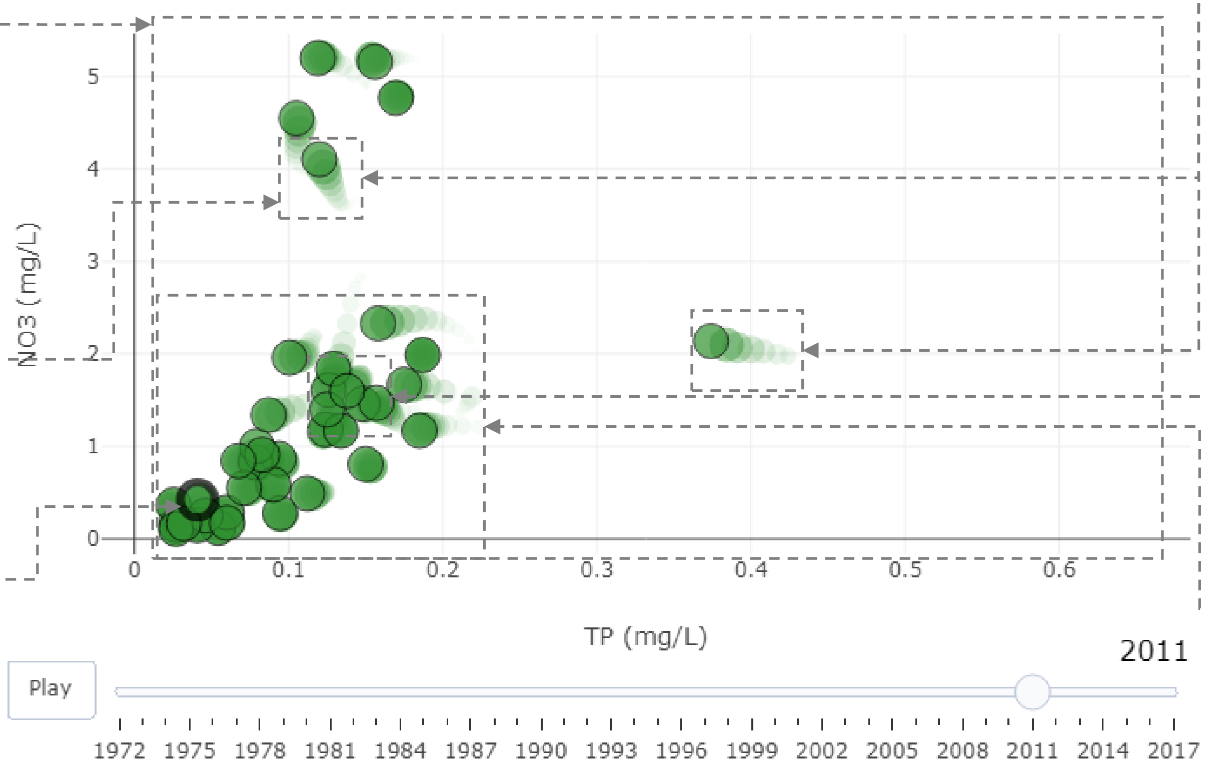
Besides the coordinated axes that allow for a bivariate comparison of sampling locations, a dynamic change in position via an interactive plug-in, tracks the long-term temporal trends in the bivariate space for each sampling location.

Size G

Changes in size of the "trail-bubbles" are made proportional to the temporal separation between marks.

Texture G

A change in contour width gives a data point a different textural appearance, in this case highlighting a selected site (nominal variable).



Orientation G

Orientation of the "trail-bubbles" change according to the temporal trajectory, in doing so, those changes reconstruct short-term temporal trends and provide with a "flow" metaphor for the passage of time.

Transparency G

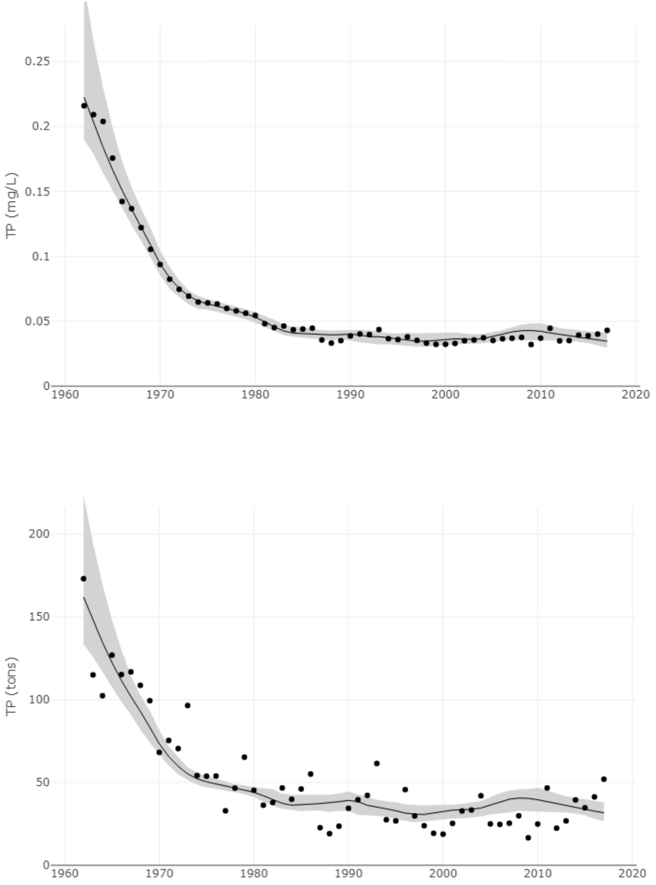
Transparency aids to separate locations with similar trends and values.

Arrangement G

Clustering of data points informs about the overall temporal change in the bivariate relationship presented in the plot.

Plot explanation: This plot shows how the relationship between any two water quality parameters has changed over time at all sites, with the selected site highlighted. Select the parameter for the horizontal axis from the menu below the map. Select the parameter for the vertical axis from the menu above this plot. Then press play to start the animation. The animation may be moved manually and paused by dragging the date slider.

Annual concentration and flux



Plot(s) explanation: These plots shows WRTDS-estimated annual mean concentrations (upper panel) and annual mean fluxes (lower panel) of the selected parameter. Points are annual estimates (not flow normalized), line is a flow-normalized estimate, and gray band (if present) is the 90% confidence interval around the flow-normalized estimate (LCL and UCL are lower and upper confidence limits).

Shape M

Points and lines represent not-flow normalized and flow normalized concentration/fluxes respectively. Since both quantities are (mathematically) related and represent the same variable, same shape with different color saturation could be a better choice.

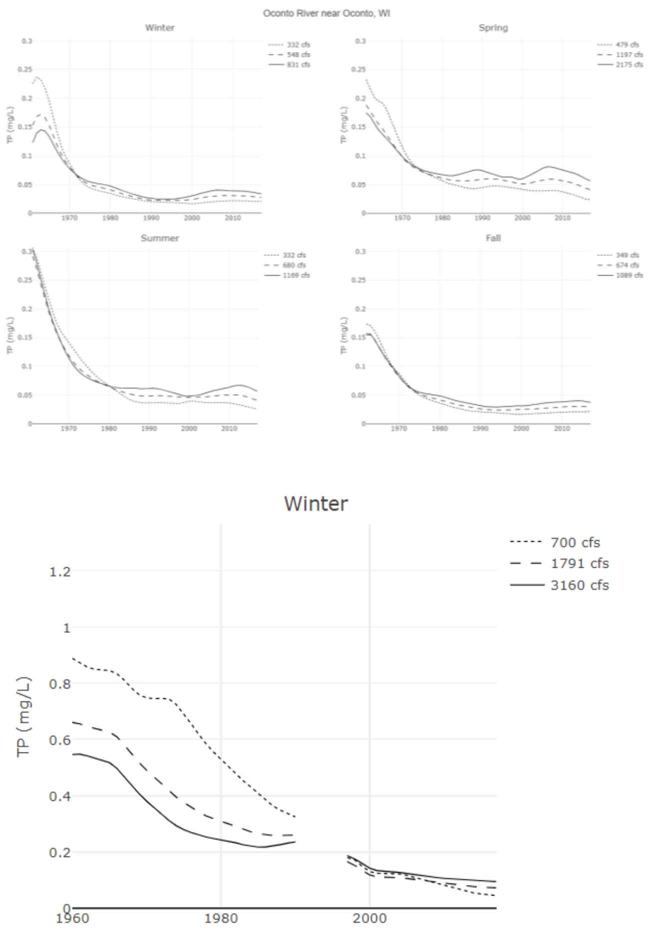
Color saturation M

Color saturation represents uncertainty. Transparency is better suited for that purpose. Also, uncertainty should be displayed for both quantities if possible.

Arrangement M

The superposition between non-flow-normalized estimates and the confidence interval for the flow-normalized trend can be visually confusing by suggesting an average trend with the points being used to estimate such a trend falling outside of the confidence interval.

Season/Flow



Plot explanation: These plots are useful for understanding the influence of season and streamflow on concentration at a point in time and for determining the conditions under which the greatest changes in water quality over time have occurred. For each season, the low and high flows are the 10th and 90th percentiles of flows during that season at that site.

Position M

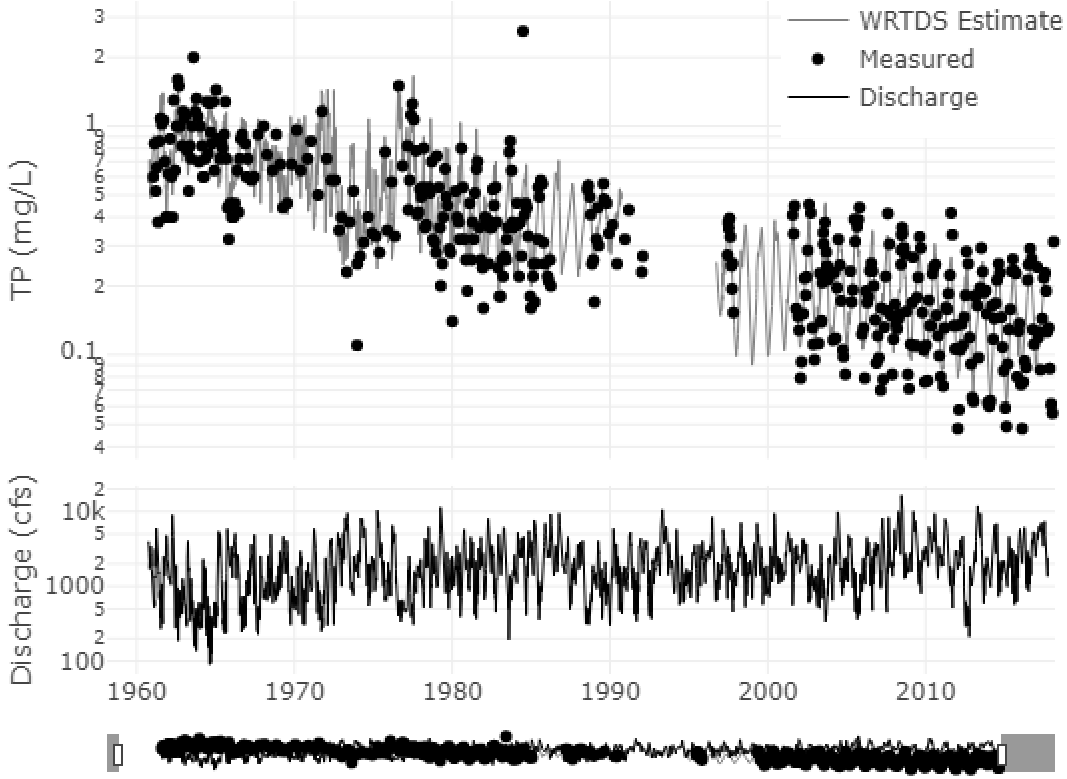
There description of the values displayed in the plot needs improvement. Are those average concentrations within each discharge percentile? Are those flow normalized concentrations?

Additionally, the distribution of the subplots reduces the effectiveness of the visualization to answer the question: Under which conditions have the greatest changes in water quality occurred over time?

Texture M

Different discharge conditions (i.e. low, average, and high flows) are represented by means of lines with different textures. Yet the values for each discharge are provided, indicating that the variable represented is quantitative. In this case, line size could be a better choice to represent discharge values.

Daily concentration



Arrangement

Transparency

Color saturation

Position

M

Vertical layering allows for dynamic comparison between discharge and flow normalized estimates. Yet, the need for flow normalization is not intuitively grasped from the layout. It creates the impression that there is a mismatch between estimated data (WRTDS) and measured data, instead of the idea of flow normalization.

Shape

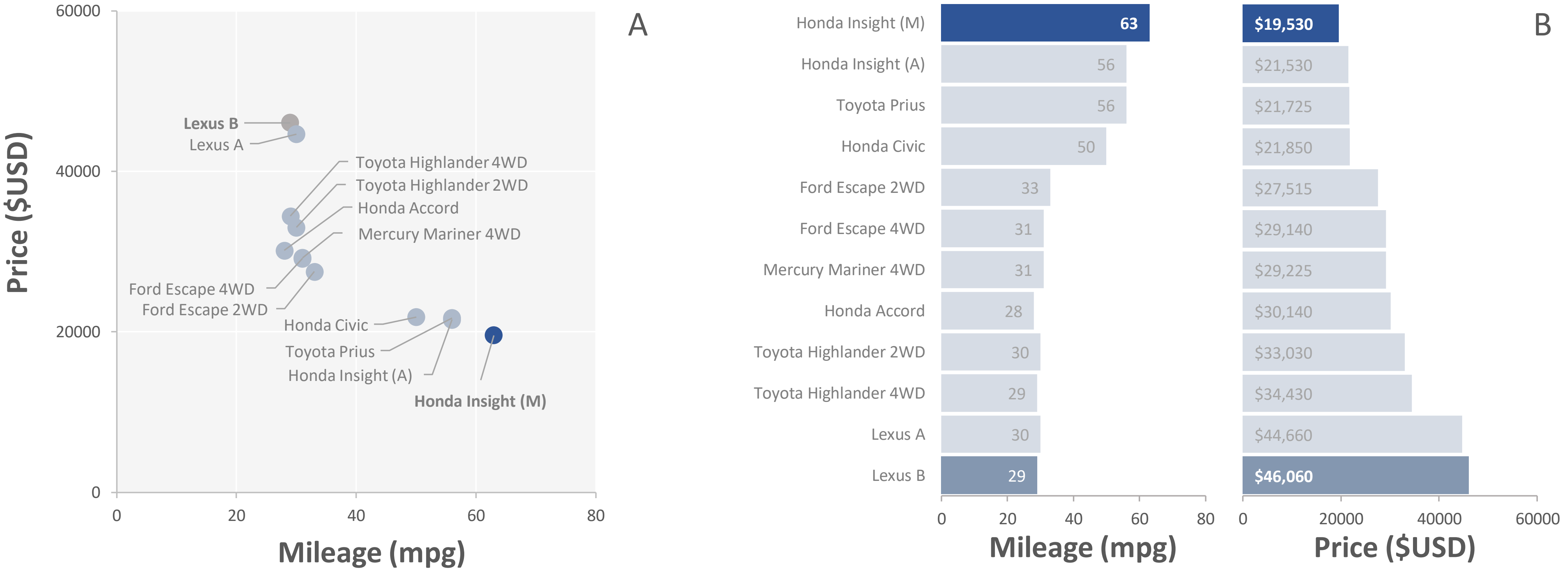
M

Shapes work best for enhancing the associative perception of qualitative variables. In this case, points and lines are used to create a contrast between measured data and WRTDS-estimated data. Being represented by two types of shapes, comparisons about variability is not intuitive. Lines could be used for both (although the discontinuity in the measured time series could be an issue). Points could be used to represent exact location in time, perhaps with less saturation. In this way, comparisons of variability ranges would be done on lines, and location in time would be represented via points.

Plot explanation: The top plot shows measured and daily WRTDS-estimated concentrations of the selected parameter. The bottom plot shows river discharge. Drag or stretch the range slider bar at the bottom of the plot to change the time span. Note that while for some sites and parameters, the WRTDS model does not reproduce measured concentrations very well, for most purposes, we are interested not in the fit to any particular daily value, but rather to summary statistics, such as annual flow-weight mean concentration. The annual plots with confidence intervals illustrate the model fit when aggregated at an annual time step.

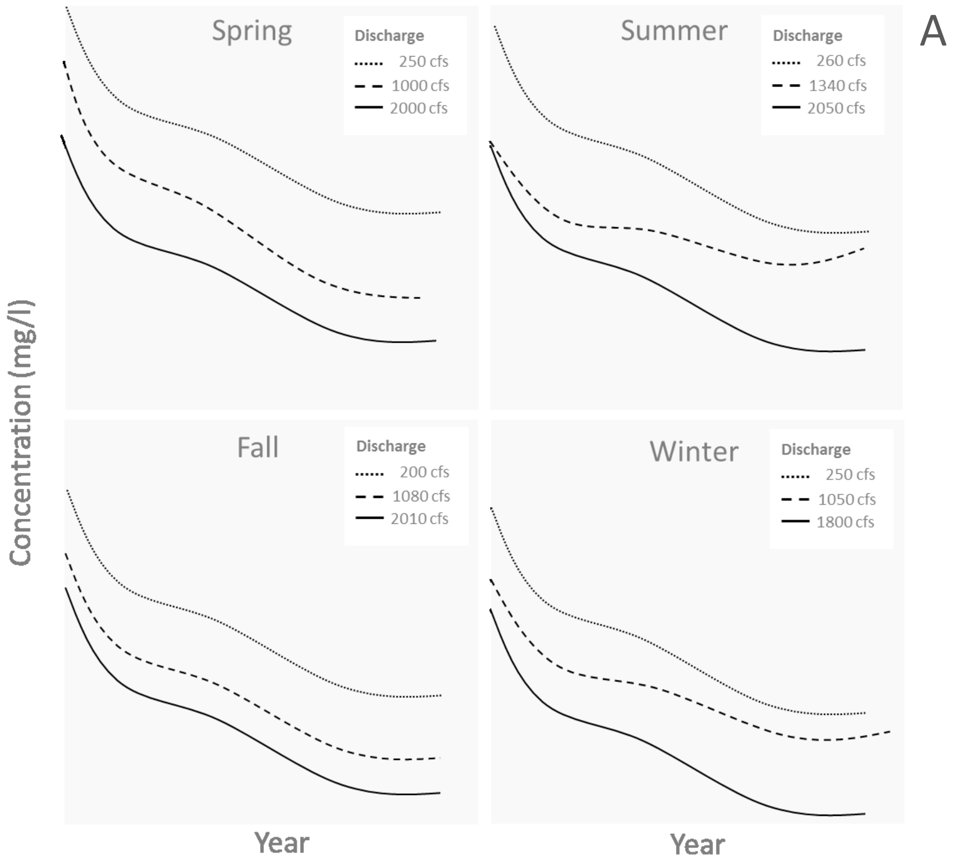
Effectiveness of a visualization

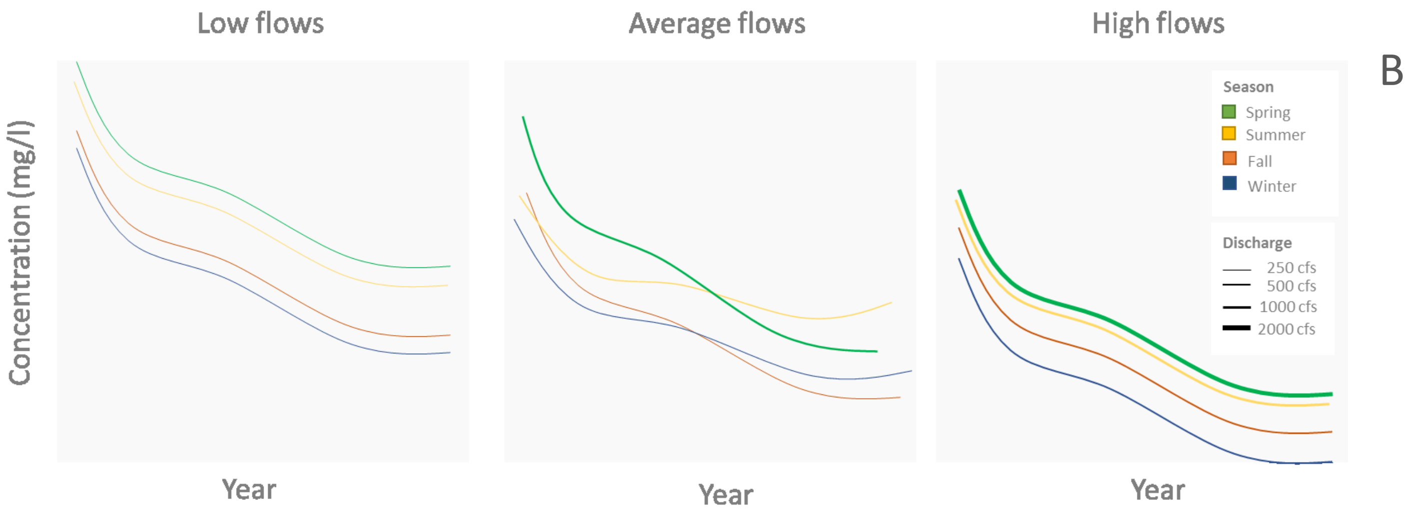
“A visualization is effective when it can be interpreted accurately and quickly and when it can be rendered in a cost-effective manner. Effectiveness is a measure such that for small datasets we measure interpretation time (since rendering is usually fast) and when that time increases, either due to increasing complexity or the size of the dataset, Effectiveness decreases. A visualization with low effectiveness is such that either the interpretation time is very large, or the rendering time is large.” (Ward, et al., 2010; p. 132).



“The information in figures A and B is the same, and both can be rendered quickly (because of the small size of the dataset). However, the effectiveness is different. The information in figure B can be interpreted more accurately and more quickly than that in figure A for some questions. For example, which car has the best mileage? However, if we ask which car has the best mileage under $30,000? Figure B is less efficient. In this case, the scatterplot provides good query-answering capabilities, but is slower for simple one variable queries. Bar charts clearly display mileage and price, but do not provide as much flexibility in answering some other queries.”(Ward, et al., 2010; p. 132-133).

Effectiveness of a visualization: Season/flow plots





Compare the query-answering capabilities between figures A and B:

- *Under which conditions have the greatest changes in water quality occurred over time?*
- Is there a consistent temporal trend in concentrations over the period of record regardless the season or flow condition?
- Is there an interaction between the season and the flow condition influencing the long-term trend in concentrations?
- Which season has the lowest concentrations over the period of record?
- Which season has the highest concentrations at average flow conditions over the period of record?
- What is the approximated range of discharge covered by the flow conditions represented in the plots?
- What is the overall range of fluctuation of the concentrations over the entire period or record?
- What is the range of fluctuation of the concentrations for in year x in winter? (I don’t think this is a valid question though)