What makes for effective data visualization?

Naomi B. Robbins, chair of the Statistical Graphics Section of the American Statistical Association, gives us a useful definition of *effective data visualization*:

> "One graph is more effective than another if its quantitative information can be decoded more quickly or more easily by most observers."
> (*Creating More Effective Graphs* by Naomi B. Robbins, http://www.nbr-graphs.com/resources/books)

Interpretations:

- Effectiveness is measurable. A more effective graph takes less time to decipher and requires less cognitive energy. Effectiveness can be measured with a stopwatch. Effectiveness can also be measured by accuracy, e.g. difference between perceived values/patterns and true values/patterns in the data.
- Effectiveness is defined by the "observers"—the intended audience. A graph's effectiveness can't be ascertained in a vacuum.
- Effectiveness is a continuum. Our goal is to make *more* effective graphs, not THE effective graph.

Where can we turn for effective data visualization?

In the last 50 years, "gurus" in data visualization have given us examples and principles for effective data visualization. Below I list a few of them. In addition, where possible, I provide their website and their most popular book.

John Tukey

Through Exploratory Data Analysis (EDA), Tukey reintroduced graphs as a serious analytical method in scientific discovery and statistical analysis. Tukey created (or made popular) several graphics still widely in use today, like the box plot. Tukey taught at Princeton University and worked at the infamous (AT&T) Bell Labs statistics group.

Book: *Exploratory Data Analysis* (http://www.amazon.com/Exploratory-Data-Analysis-John-Tukey/dp/0201076160)

Edward Tufte

Known as "the Leonardo da Vinci of data" and "the Galileo of graphics", Tufte pioneered the field of data visualization. Tufte is famous for introducing many of the foundational principles of data visualization. Among the most important are *data-ink* and the *data-ink ratio*. "Data-ink is the non-erasable core of the graphic, the non-redundant ink arranged in response to variation in the numbers represented." The data-ink ratio is the proportion of data-ink to the total ink used in the graphic. While at Princeton, Tufte taught a series of joint seminars with John Tukey (which materials became the foundation for his first book).

Website: http://www.edwardtufte.com/tufte/

Book: *The Visual Display of Quantitative Information* (http://www.edwardtufte.com/tufte/books_vdqi)

William (Bill) Cleveland

Cleveland conducted experiments on how we decode the quantitative and qualitative information in graphs. He developed principles of visual perception. As we take advantage of our brain's preattentive processing, we are able to convey information quickly and accurately. Cleveland also coined the term "Data Science" and developed the ever useful trellis (facet) display. Cleveland also worked at Bell Labs.

Website: http://www.stat.purdue.edu/~wsc/

Book: *The Elements of Graphing Data* (http://www.amazon.com/Elements-Graphing-Data-William-Cleveland/dp/0963488414)

Leland Wilkinson

Wilkinson developed a formal description of and system for creating (static) graphics. Wilkinson's system has been implemented in many modern data statistical software, namely, SPSS and R (especially through the package 'ggplot2'). He now works for Tableau, the graphical software company, and, yes, Wilkinson also worked at Bell Labs.

Website: https://www.cs.uic.edu/~wilkinson/

Book: *The Grammar of Graphics* (http://www.amazon.com/The-Grammar-Graphics-Statistics-Computing/dp/0387245448)

Naomi B. Robbins

Robbins is chair of the Statistical Graphics Section of the American Statistical Association and writes data visualization articles for Forbes. She gave us a useful definition for effective data visualization. She also worked at Bell Labs alongside Cleveland.

Website: http://www.nbr-graphs.com/ and http://www.forbes.com/sites/naomirobbins/#6708247e1a23 (blog)

Book: *Creating More Effective Graphs* (http://www.amazon.com/Naomi-B.-Robbins/e/B001IR3KAQ)

Stephen Few

Few makes Tufte practical. Of all the visualization gurus listed here, Few is perhaps the most useful for "corporate" analytics—the basic analyses usually needed/required by many companies. Few is also famous for establishing good practices behind dashboards.

Website: http://www.perceptualedge.com/

Book: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* (http://www.amazon.com/Stephen-Few/e/B001H6IQ5M)

Hans Rosling

A Swedish medical doctor, academic, statistician, and global health expert, Rosling showed us what happens when you combine really important (health) statistics and animation. Rosling is famous for his TED talks where he uses statistics, data visualization, and storytelling to convey important global health ideas.

Website: https://www.ted.com/speakers/hans_rosling (TED talks) and http://www.roslingsblogger.blogspot.com/ (blog)

Nancy Duarte

Duarte is an expert in presentation design. Duarte is famous for turning Al Gore's message on Climate Change into (probably) the most widely-seen Keynote presentation on the planet (it became the basis for the film *An Inconvenient Truth*).

Website: http://www.duarte.com/

Book: *slide:ology: The Art and Science of Creating Great Presentations* (http://www.amazon.com/Nancy-Duarte/e/B002BMAA0K)

Garr Reynolds

A former manager at Apple, Garr (as he prefers to be called) is a contemporary of Nancy Duarte. Along with Duarte, Garr is a presentation design expert. In addition, Garr does a great job of employing Tufte's principles in slides (like PowerPoint… something Tufte is not found of).

Website: http://www.garrreynolds.com/ and http://www.presentationzen.com/ (blog)

Book: *Presentation Zen: Simple Ideas on Presentation Design and Delivery* (http://www.amazon.com/Garr-Reynolds/e/B001I9TU1W)

Hadley Wickham

Hadley took Wilkinson's grammar of graphics and implemented it in a widely-used, free package in the R statistical programming language called ggplot2 (http://ggplot2.org/). The "gg" in ggplot2 stands for "grammar of graphics". ggplot2 is one of the most popular R packages and is often used by companies and media outlets to produce everything from quick exploratory graphics to production-quality visualizations. ggplot2 has also been implemented in Python, Plotly, and Matlab.

Website: http://hadley.nz/

Book: *ggplot2* (http://www.amazon.com/Hadley-Wickham/e/B002BOA9GI)

Nathan Yau

Yau is the creator of the popular data visualization website: FlowingData. As a statistician with an emphasis in data visualization and one of the youngest experts on the list, Yau brings a unique perspective to the field.

Website: http://flowingdata.com/

Book: *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics* (http://www.amazon.com/Nathan-Yau/e/B004S83IUE)

Alberto Cairo

Cairo is famous for bringing data visualization to journalism. Known, among many things, for his beautiful and insightful infographics (done right!), Cairo has a unique ability to tell a story through data and visualization.

Website: http://www.thefunctionalart.com/

Book: *The Functional Art: An introduction to information graphics and visualization* (http://www.amazon.com/Alberto-Cairo/e/B0050KJ3NK)

How do I create effective data visualizations?

There are many good principles and techniques to creating effective data visualizations. Most of them are found in the resources listed above. Here I share five of them which come from Stephen Few.

In his book, *Show Me the Numbers*, Stephen Few shares two objectives behind all visual communication:

> "We **highlight** important information to give it a voice that comes through loudly and clearly, without distraction. We **organize** information to lead readers through it in a manner that promotes optimal understanding and use."
> (*Show Me the Numbers* by Stephen Few, http://www.perceptualedge.com/library.php, emphasis added)

Within each objective, Few outlines several design principles.

Highlight		
1.	Reduce non-data-ink	Eliminate *unnecessary* non-data-ink. Some non-data-ink is necessary to support the visual components. De-emphasize this remaining non-data ink.
2.	Enhance data-ink	Not all information is equally important. Emphasize the most important data ink using variations in width, orientation, size, enclosure, hue, or color intensity.
Organize		
3.	Group data into meaningful sections	Begin with a clear sense of what belongs together. Use proximity—put related data close together, separate different groups with white space.
4.	Prioritize data by importance	Use contrasts to prioritize—i.e. sort; make bigger/smaller, thicker/thinner; darken/brighten; italicize/bold; change the shape, hue, or position.
5.	Sequence data by how it should be read	Provide clear direction on the best sequence to read the data—e.g. left-to-right and top-to-bottom navigational sequencing. If it's more complex, use sequential labels.

Hospital Charges

Working in a pair or group, how would you **organize** the hospital charges data to "lead readers through it in a manner that promotes optimal understanding and use"?

Pertaining to the hospital charges data, one news report said:

> "In an unprecedented move Wednesday, the Centers for Medicare & Medicaid Services [CMS] made public extensive hospital cost data, jolting healthcare providers, payers, and consumers alike.
>
> The massive file contains *chargemaster data* or what some call the 'sticker price' for the 100 most common Medicare inpatient diagnostic related groups or DRGs. The data does not include physician costs. But it does provide an inside look at how average covered Medicare charges can significantly vary from hospital to hospital within the same city or geographic area.
>
> The data is for [~3,200] hospitals and represent 92% of all hospital inpatient charges in fiscal year 2011."
>
> [HealthLeadersMedia.com, http://www.healthleadersmedia.com/page-1/HEP-292001/CMS-Releases-Hospital-Pricing-Data]

Data source: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html [Note: in the original data, "provider" means hospital.]

Variable description:

Variable Name	Description	Values
DRG Definition	Diagnosis Related Groups (DRGs); classify hospital stays into diagnosis groups for payment purposes	100 DRG categories
Hospital Name	Hospital name	3,201 hospitals
Hospital Location	Hospital street address, city, state, zip code	1,977 cities; 51 states; 3,053 zip codes
Hospital Referral Region (*HRR*)	Delineates regional health care markets; used to make comparisons within geographic regions	306 regions
Total Discharges	Number of discharges billed by the hospital for that particular DRG	Min: 11 Max: 3,383 Median: 27 Mean: 43
Average Covered Charges (*Hospital Charges*)	Average of hospital charges billed to Medicare for that particular DRG	Min: $2,459 Max: $929,100 Median: $25,250 Mean: $36,130
Average Total Payments (*Medicare Payments*)	Average of payments made by Medicare for that particular DRG (difference was not paid)	Min: $2,673 Max: $156,200 Median: $7,214 Mean: $9,707

Example of data—*one row per DRG-hospital combination*:

DRG Definition	Hospital Name	Hospital Location	HRR	Total Discharges	Hospital Charges	Medicare Payments
simple pneumonia & pleurisy w mcc	Mayo Clinic - St Mary's	Rochester, MN	MN - Rochester	117	$26,428	$12,260
simple pneumonia & pleurisy w cc	Mayo Clinic - St Mary's	Rochester, MN	MN - Rochester	193	$16,035	$7,931
heart failure & shock w mcc	Mayo Clinic - St Mary's	Rochester, MN	MN - Rochester	168	$32,429	$13,082
heart failure & shock w cc	Mayo Clinic - St Mary's	Rochester, MN	MN - Rochester	350	$18,874	$9,086
heart failure & shock w/o cc/mcc	Mayo Clinic - St Mary's	Rochester, MN	MN - Rochester	79	$11,643	$5,750
…	…	…	…	…	…	…
simple pneumonia & pleurisy w mcc	Univ of MN - Fairview	Minneapolis, MN	MN - Minneapolis	42	$32,889	$14,797
simple pneumonia & pleurisy w cc	Univ of MN – Fairview	Minneapolis, MN	MN - Minneapolis	46	$25,767	$9,647
…	…	…	…	…	…	…

[w: with, w/o: without, mcc: major complications and comorbidities, cc: complications and comorbidities]

Sketch out your data visualization.