DATA SCIENCE: HOW DO YOU SPEAK THE LANGUAGE?

Firas Moosvi

Questions this talk will address

1. What is the language Data Scientists use to present their results to other scientists, to each other, and to the public?

2. How can data analyses be communicated effectively and impactfully ?

3. How should Data Scientists ensure their work is relevant, transferrable, and applicable?

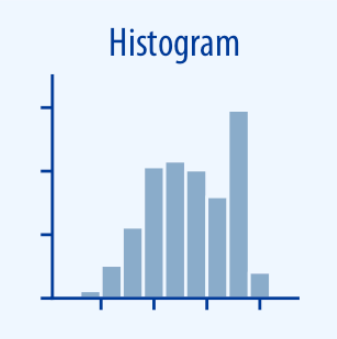
Data Visualizations

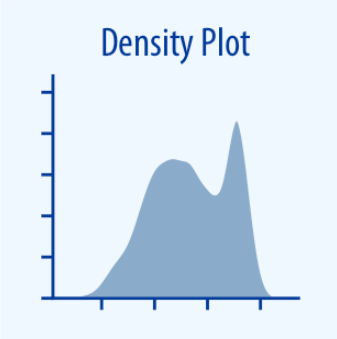
Language of Data Science

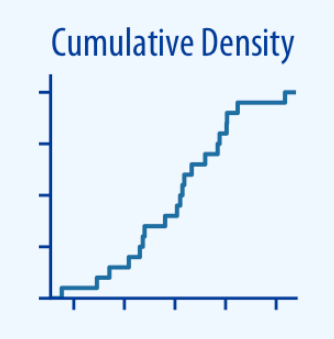
- Similar to most other sciences, data visualizations (or graphs and plots) are **essential** for communicating data analyses.

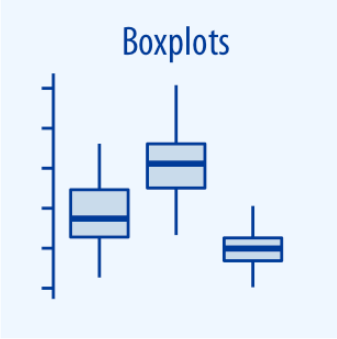
- I have chosen some common visualizations you may encounter when consuming data science analyses.

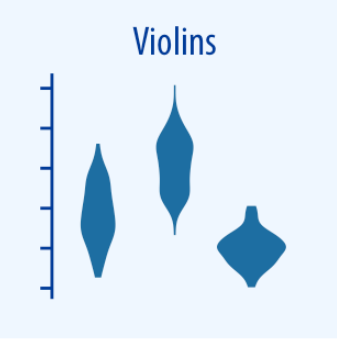
Distributions

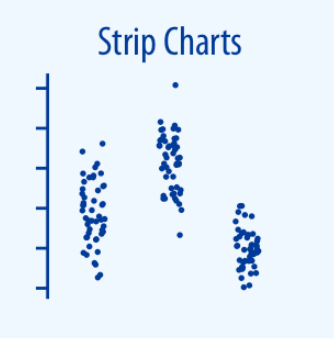


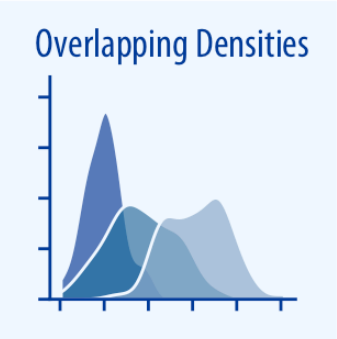






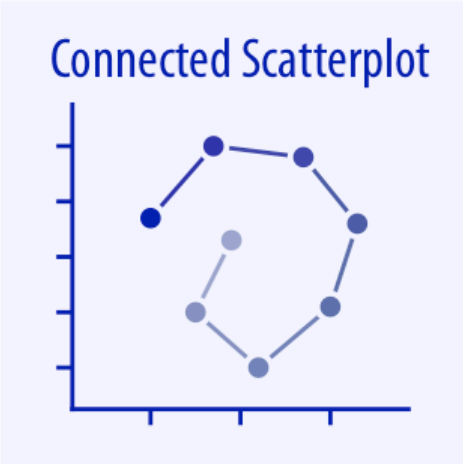


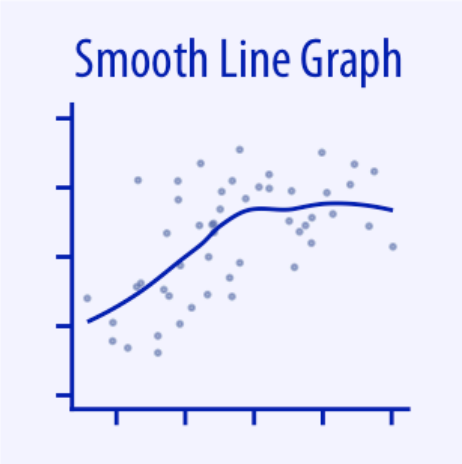




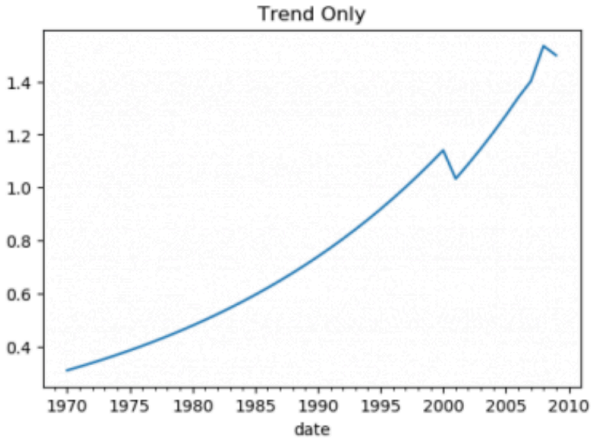
Source: Fundamentals of Data Visualization by Claus Wilke

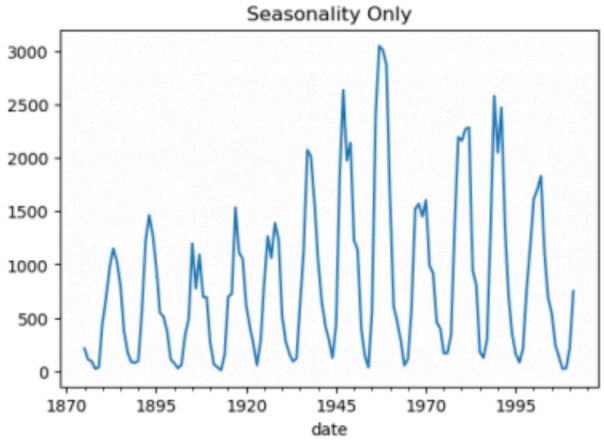
Time series

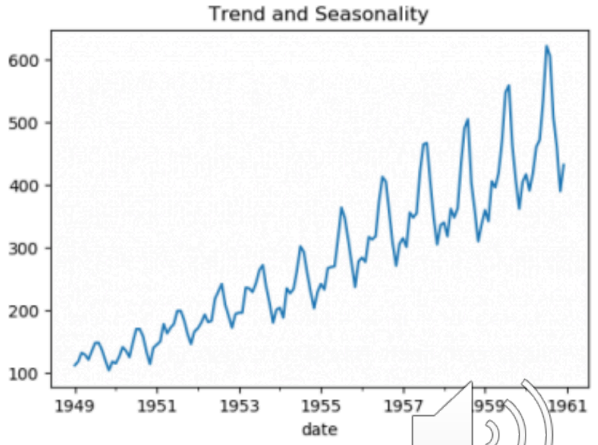




Source: Fundamentals of Data Visualization by Claus Wilke



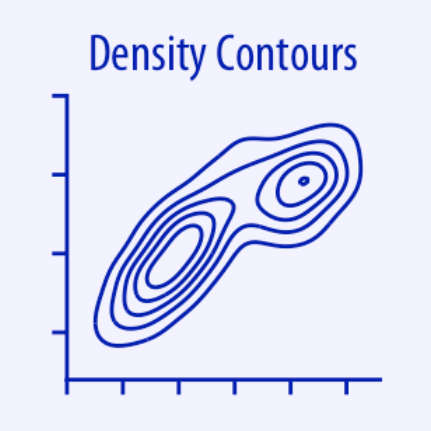


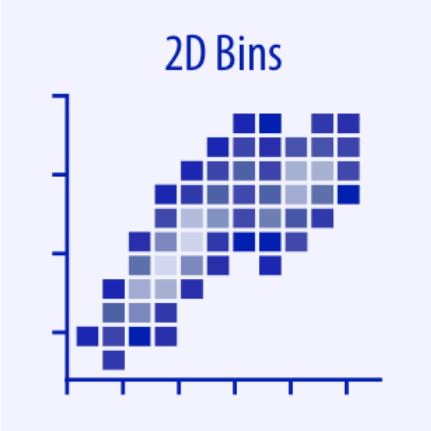


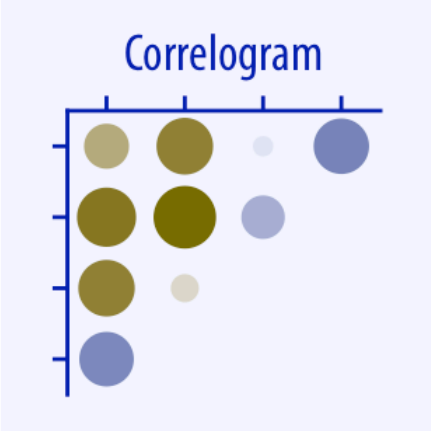
Patterns in Time Series

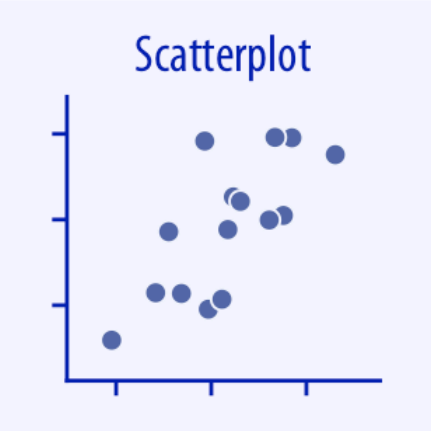
Source: Time Series Analysis in Python

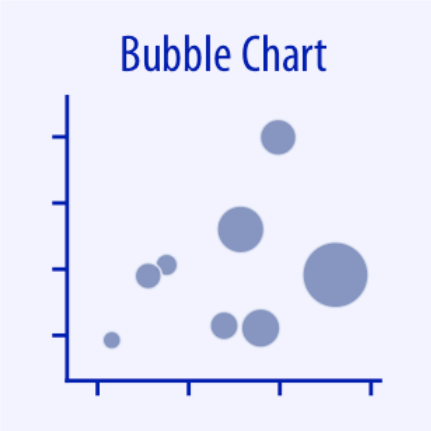
X-Y Relationships

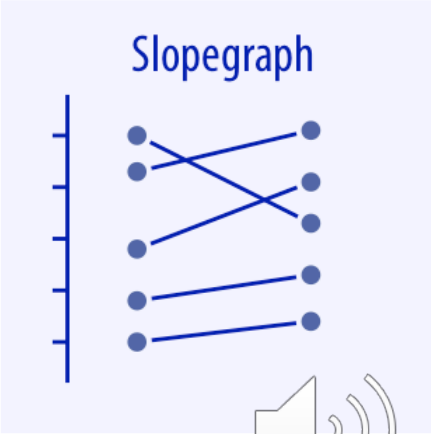






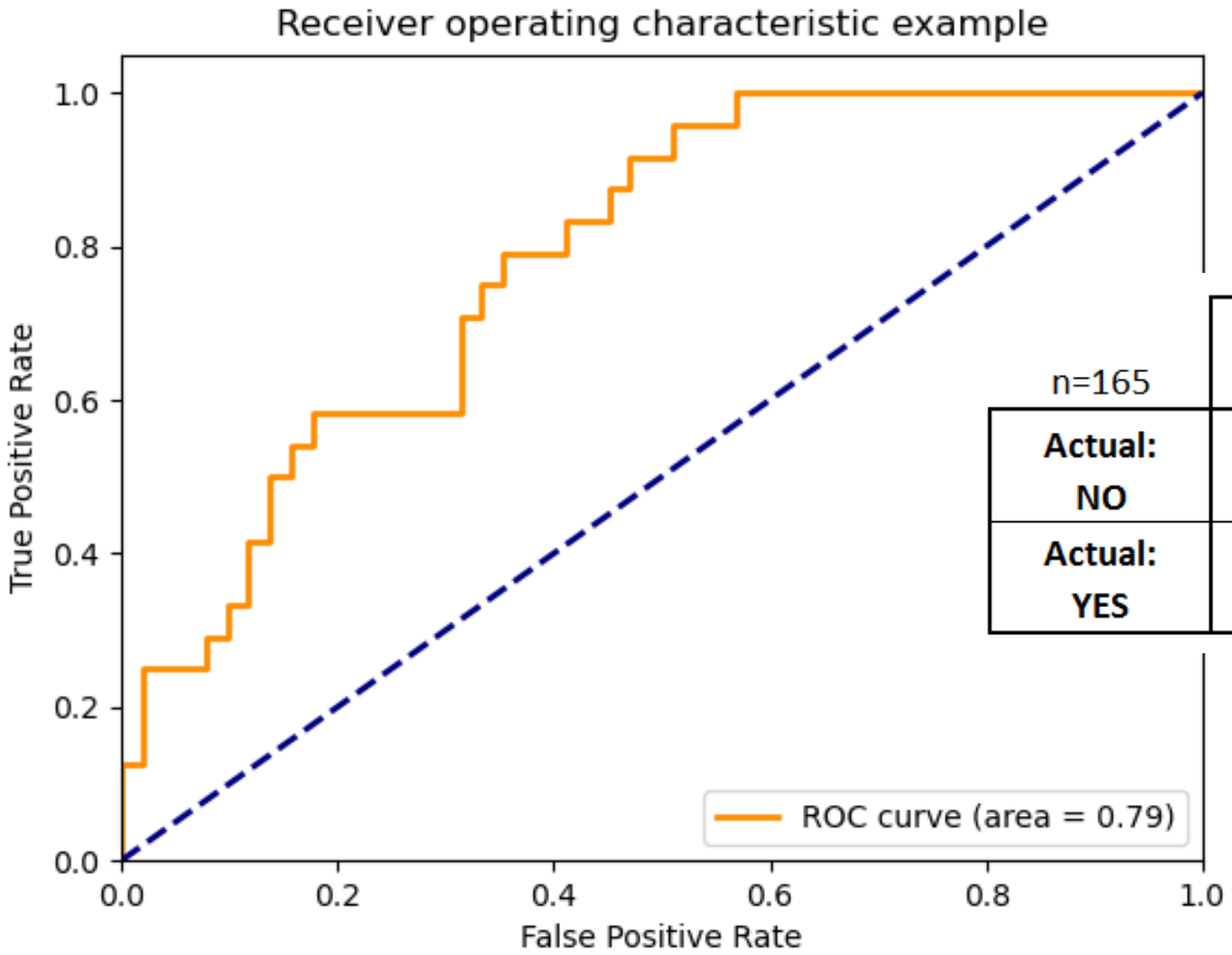






Source: Fundamentals of Data Visualization by Claus Wilke

Confusion Matrix and ROC curves

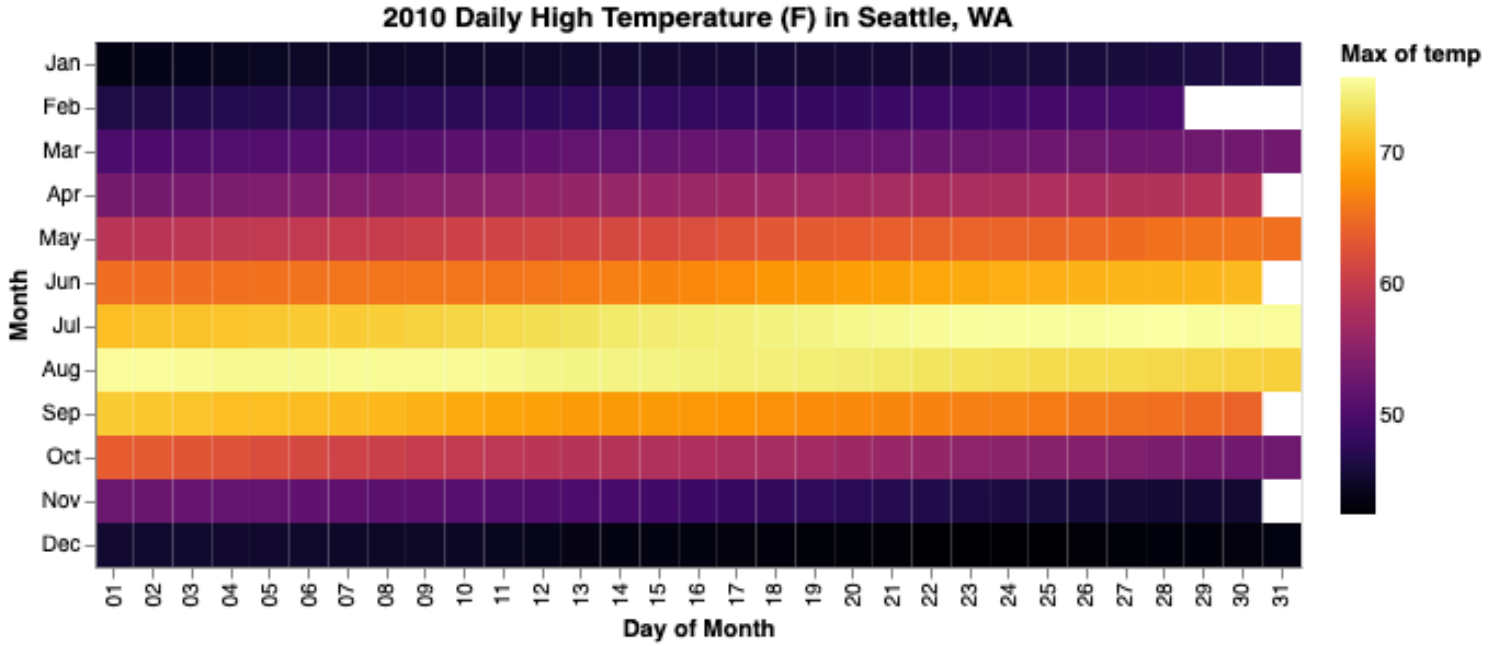


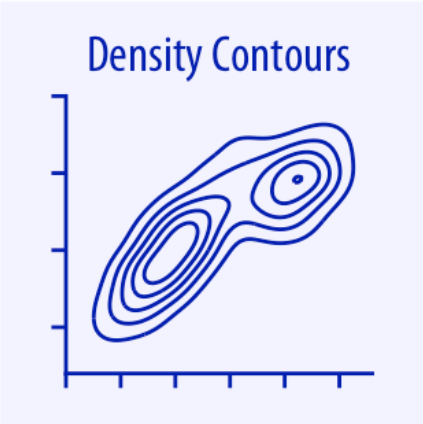
n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

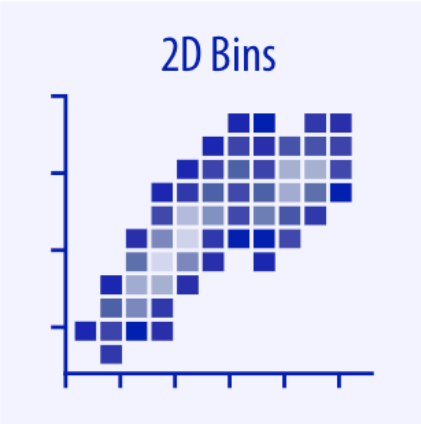
Source: Guide to Confusion Matrix Terminology

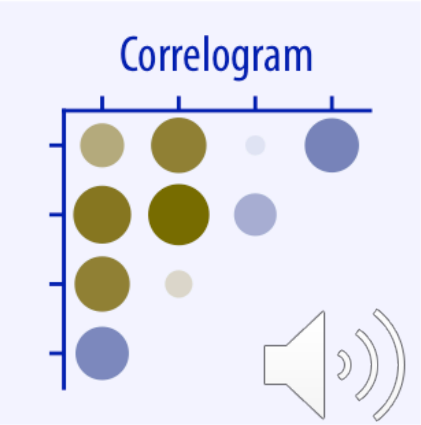
Source: Sci-kit Learn ROC plot

Heat Maps (and similar)



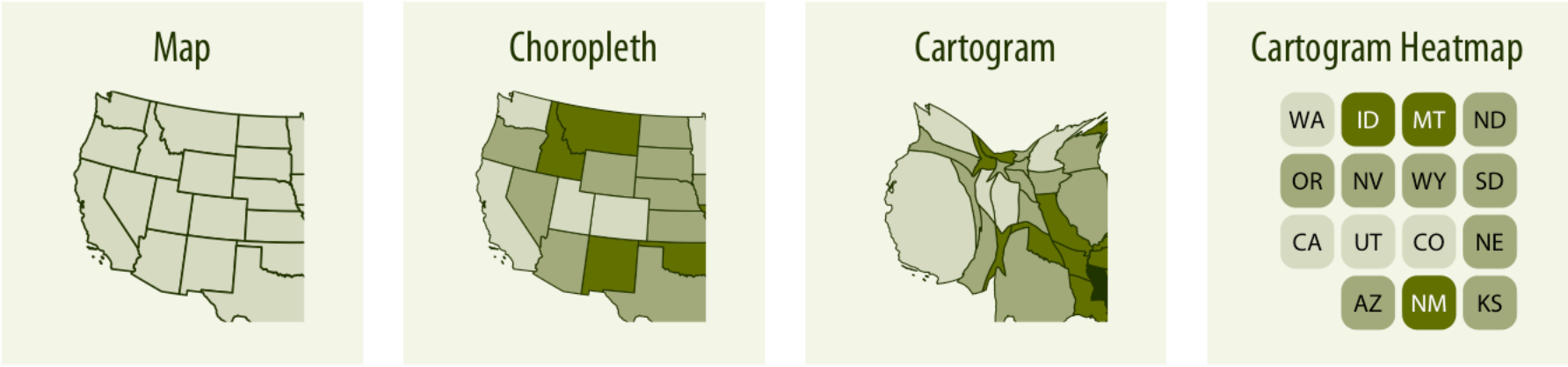




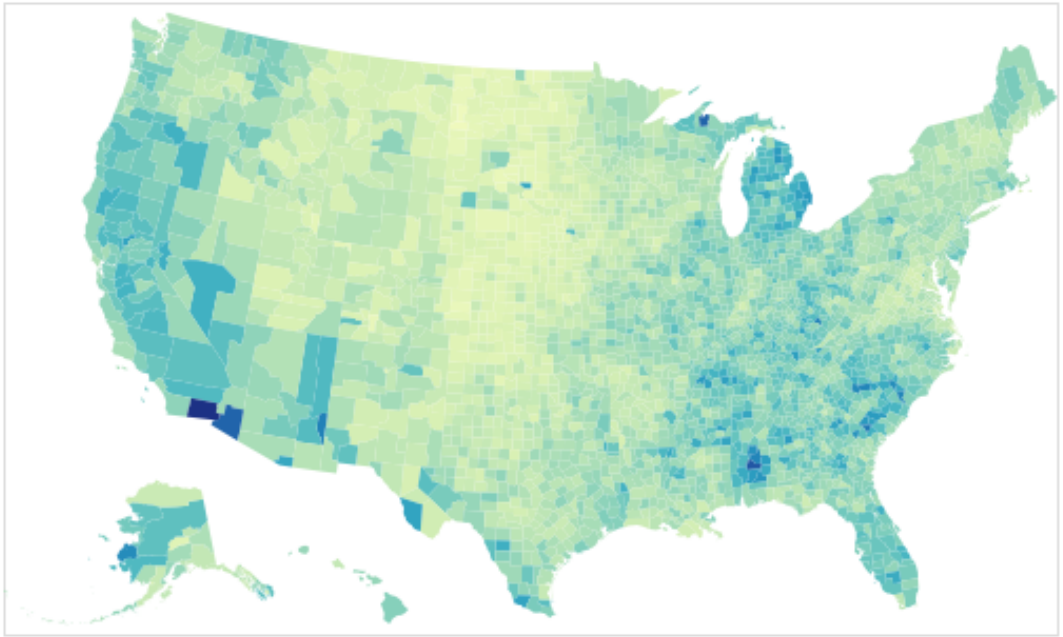


Source: Fundamentals of Data Visualization by Claus Wilke

Choropleths (Geospatial)

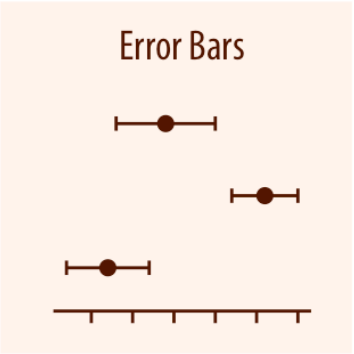


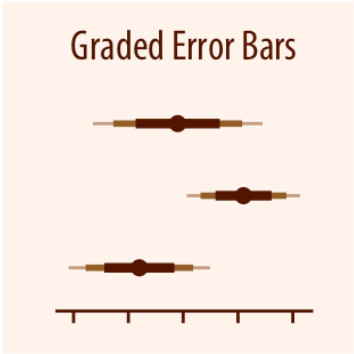
Source: Fundamentals of Data Visualization by Claus Wilke

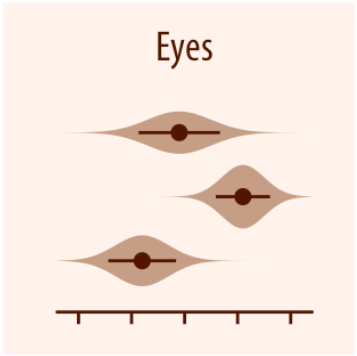


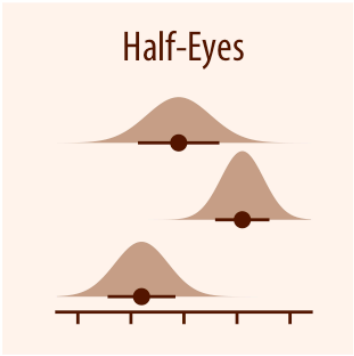
Source: Altair Choropleth Gallery

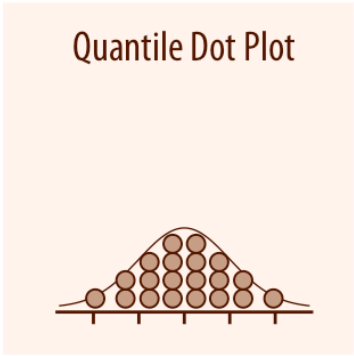
Visualizing Uncertainty

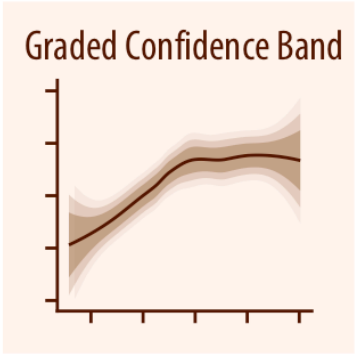


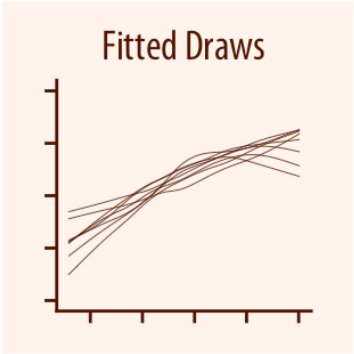












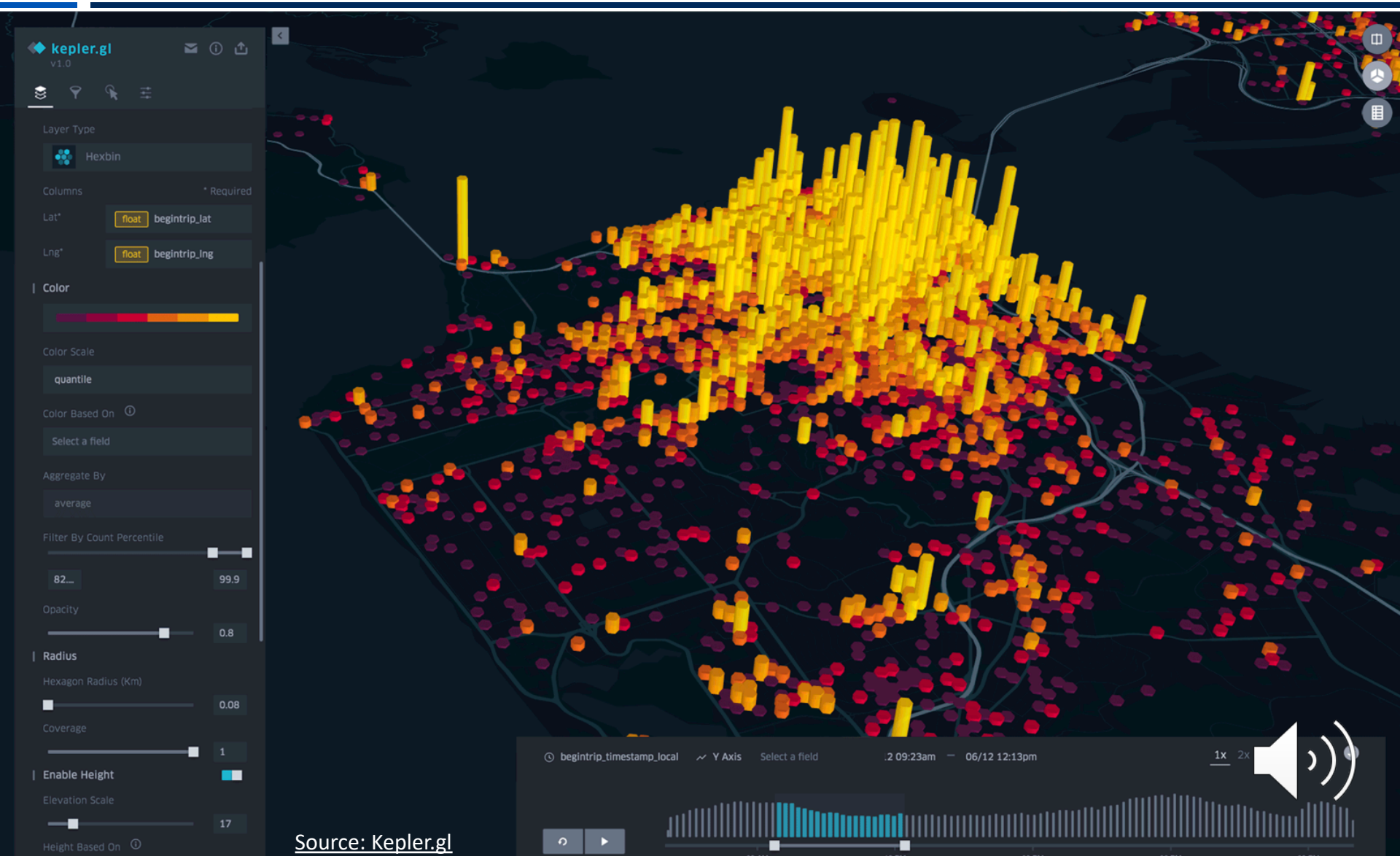
Source: Fundamentals of Data Visualization by Claus Wilke

Word Clouds and Network Diagrams

Source: Data Science Central - Word Clouds

Source: Dash App Gallery - Word Embeddings

Higher Dimensional Plots



Source: Kepler.gl

from Data to Viz

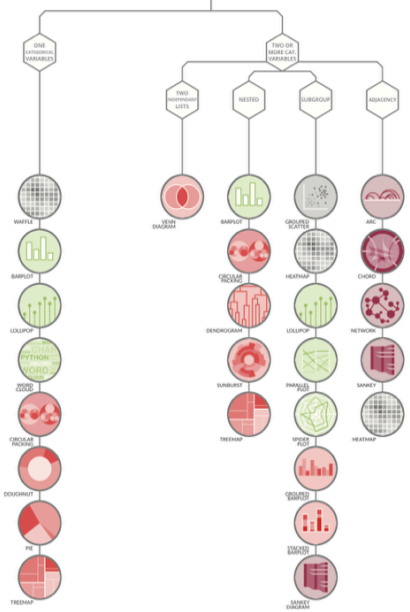
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

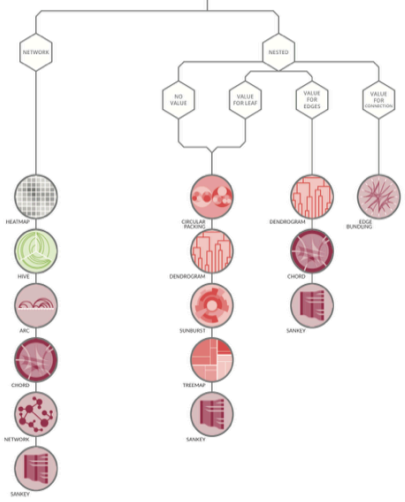
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

CATEGORIC

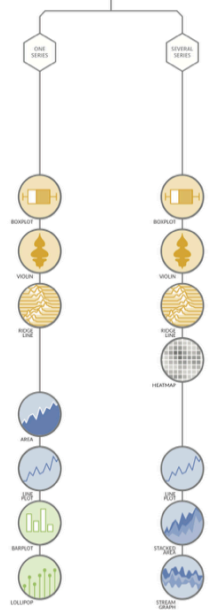


RELATIONAL



MAP

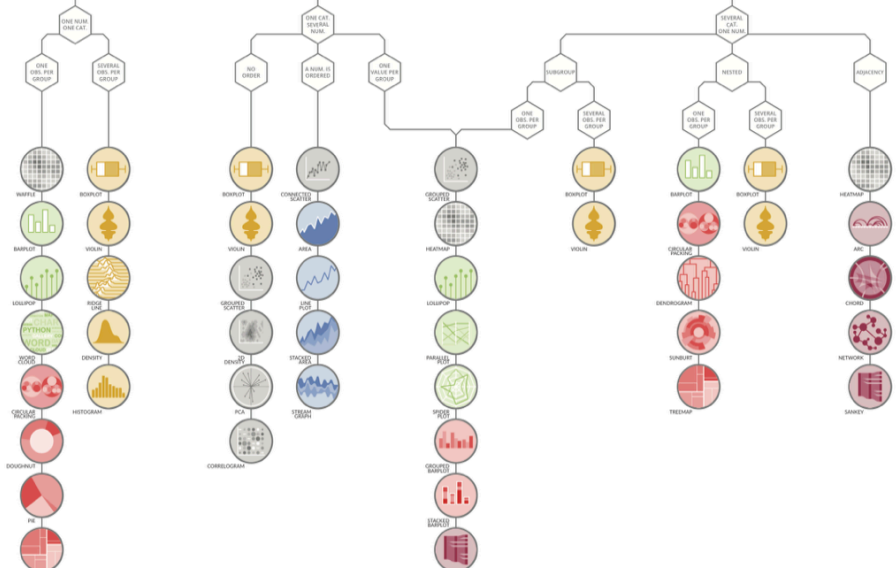
TIME SERIES



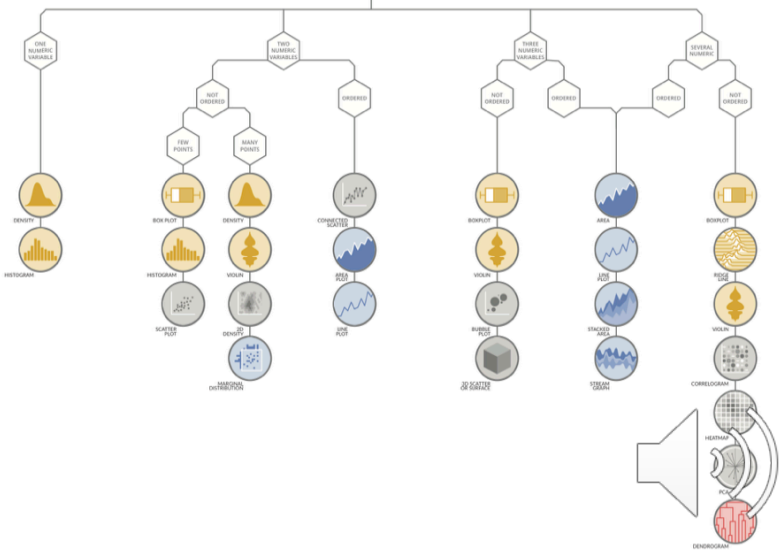
WHAT DO YOU WANT TO SHOW ?

- Distribution
- Correlation
- Ranking
- Part of a whole
- Evolution
- Maps
- Flow

CATEGORIC AND NUMERIC



NUMERIC



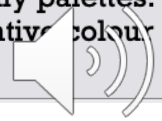
Source: From Data to V

Principles of Effective Visualizations

Source: Darkhorse Analytics

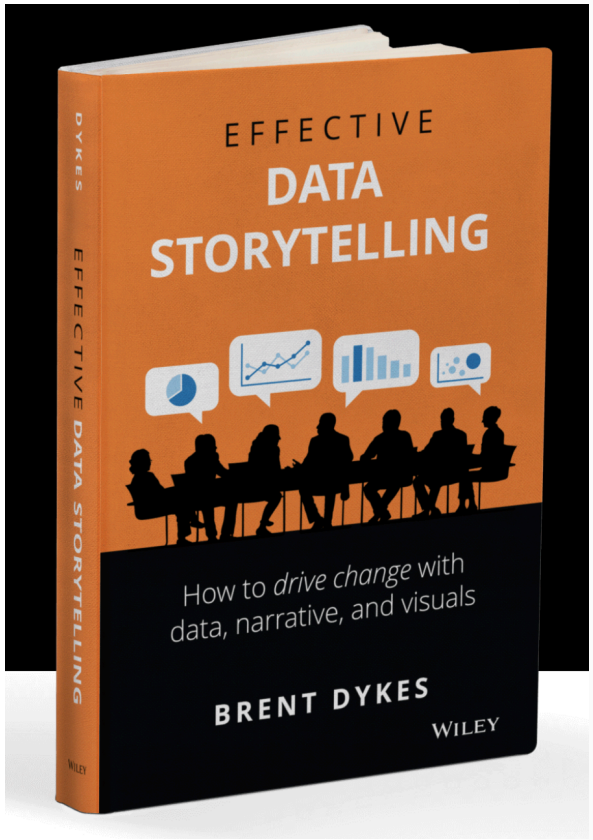
Principles of Effective Visualizations

Principle	Definition	Examples
• Proportional Ink	The amount of ink used to indicate a value should be proportional to the value itself.	Truncating the y-axis on a bar chart to exaggerate the difference between bars violates the principle of proportional ink.
• Data:ink ratio	Remove distracting visual elements to focus attention on the data	Lighten line weights, remove backgrounds, never use 3D or special effects, ~~remove~~ avoid unnecessary/redundant labels.
• Labels & legends	Use axes labels and titles to highlight/ communicate data	Never leave your data column names as axes labels! Generally good to add a title.
• Overplotting	With large datasets, points overlap, resulting in large clouds of data	To fix overplotting, could plot just a sample subset of the data, use alpha, and use smaller points. Or, jitter - but check if appropriate!
• Visualization choice	Must be informed by the **data** you have, the **research question** being asked and the **audience** that cares.	Pick the simplest plot that best shows most/all of the data needed to answer the research question. If you only have summary statistics, cannot show distributions. Tailor the visualization to your audience (within reason) but don't dumb it down.
• Colour & Accessibility	Colour can be used to encode information or for aesthetics/style/ design. However, colour can also be distracting if used inappropriately or poorly.	Choose a perceptually uniform colour palette; can be sequential or diverging for quantitative data. Opt for colour-blind friendly palettes. Categorical data can use qualitative colour schemes.



Building Narratives of Data Analyses

Brent Dykes on Data story-telling



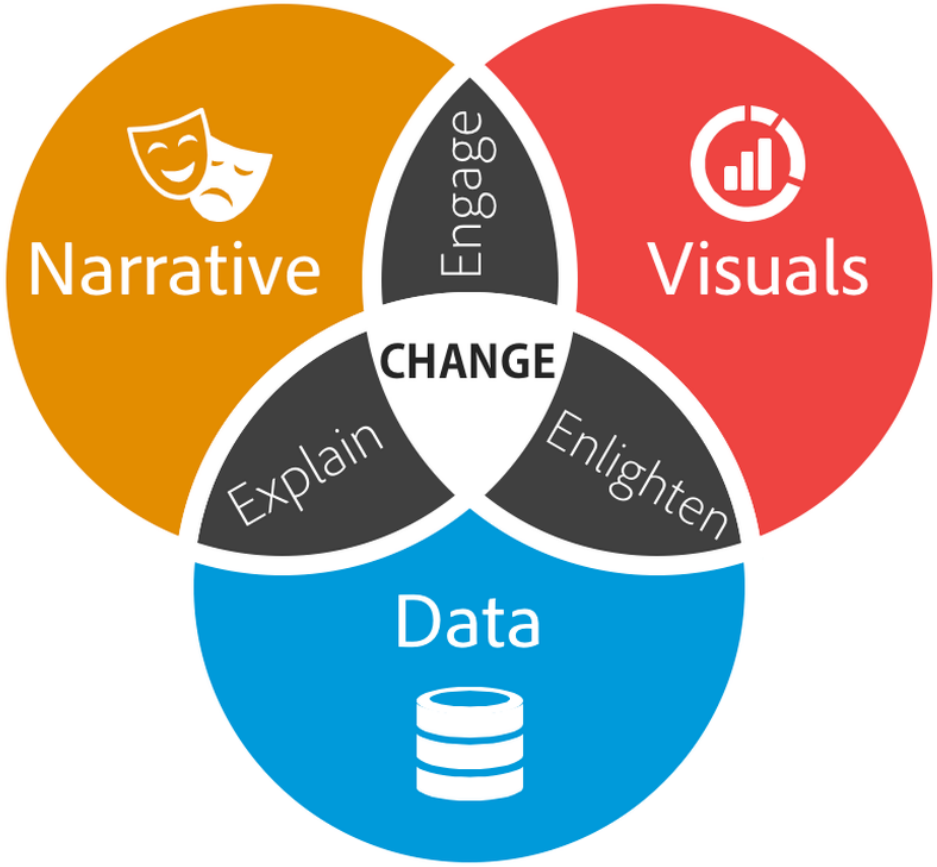
Data storytelling is a structured approach for communicating data insights, and it involves a combination of three key elements: *data*, *visuals*, and *narrative*.

It's important to understand how these different elements combine and work together in data storytelling. When narrative is coupled with data, it helps to ***explain*** to your audience what's happening in the data and why a particular insight is important. Ample context and commentary are often needed to fully appreciate an insight. When visuals are applied to data, they can ***enlighten*** the audience to insights that they wouldn't see without charts or graphs. Many interesting patterns and outliers in the data would remain hidden in the rows and columns of data tables without the help of data visualizations.

Finally, when narrative and visuals are merged together, they can ***engage*** or even entertain an audience. It's no surprise we collectively spend billions of dollars each year at the movies to immerse ourselves in different lives, worlds, and adventures. When you combine the right visuals and narrative with the right data, you have a data story that can influence and drive ***change***.

Source: Effective Data Storytelling (the book) and the Forbes article.

Brent Dykes on Data story-telling



Source: Effective Data Storytelling (the book) and the Forbes article.

Importance of the audience

- The audience of your data analyses matters A LOT!

- You may need to communicate results differently:
  - manager/executive (Image/infographic/sentence)
  - public consumption (Blog post or Jupyter notebook)
  - yourself or others who want to dig deep to explore and understand the data (GitHub repository with code and notebooks)

- Give consumers of your analyses and results choice on how they want to interact with your work

Source: 100 Years of Women in the House of Commons

Consuming data science analyses

- You have choice!

- Do you want to:
  - Get a high-level summary of the results or implications?
  - Dig into the results and extract meaning ?
  - Learn about the methods at an introductory level?
  - Repeat the analyses with different datasets?
  - Understand the methods at the deepest level?
  - Do research in the field to compare or improve the methods?

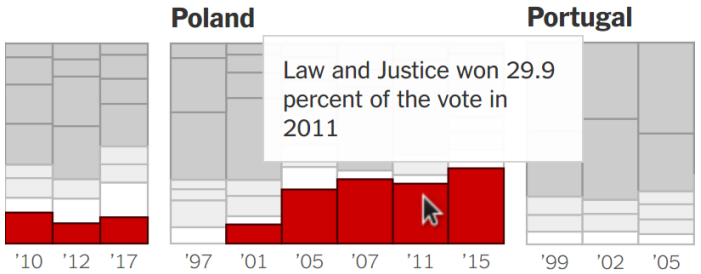
- **Choose the right format for the right level… (interactive notebook, code, dashboard, documentation, blog post, etc…)**

Dashboard and interactivity

Purpose of Interactivity & Dashboards

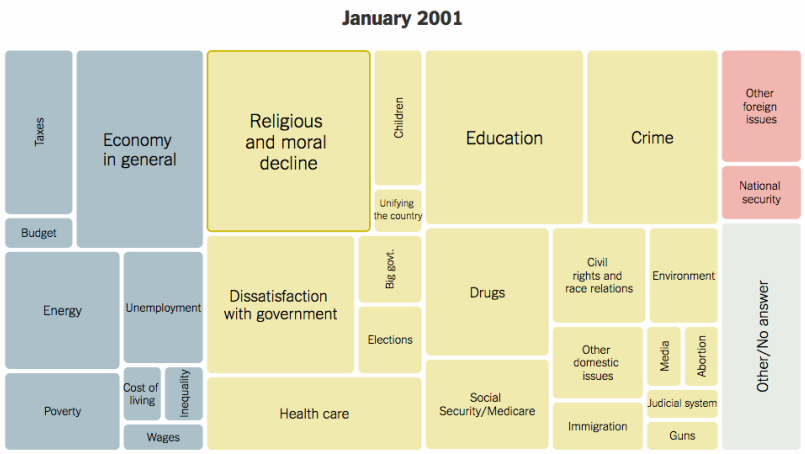
1. Tooltips allow your most interested users to dig deep

Take a look at the following graphic which summarized election results across 20 European countries. Everything you need to see is shown right away. You see the country names, the years and the red bars representing results of right-wing and far-right parties.



2. Interaction allow readers to discover the full dataset

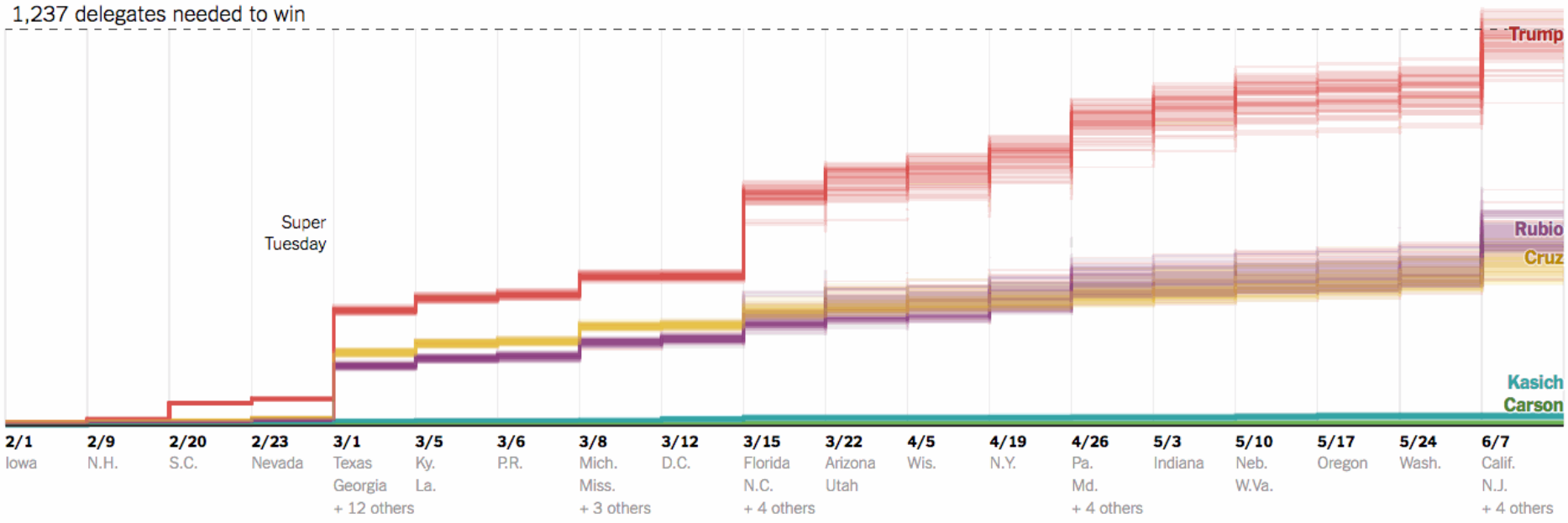
There are cases when you have far more data than fit on a page, which means you have to select which charts to show and which to hide. To avoid cherry-picking we usually try to come up with a selection rule that we apply consistently throughout the piece. For instance, in a recent graphic we decided to show the first poll after the start of the term for each president. Deciding on such a rule is definitely better than just picking charts, but it can still feel arbitrary sometimes. Fortunately we had already set up the graphic in a way that the charts are rendered dynamically. So it didn't cost us much to add in a little bonus feature that allows browsing through the entire dataset.

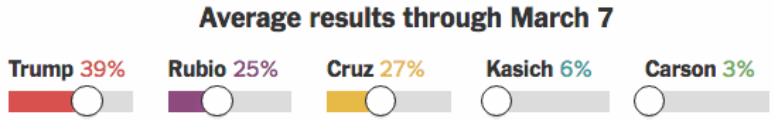


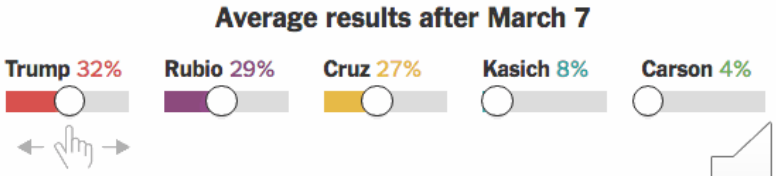
Source: Election results in 20 EU countries and Blog post

Dashboard and interactivity

3. Interaction can help build trust in your data analysis



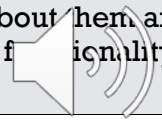




Source: Delegate Calculator and Blog post

Principles of Effective Dashboards I

Principle	Explanation
Audience Matters (a lot!)	You may need to build dashboards with different views: - one for a manager/executive - one for yourself to explore and understand the data - one for the public
Purpose-driven Dashboards	Every dashboard should have a purpose! Resist the idea to bake in the "purpose" as a dropdown or menu option. What are the usage scenarios? List your intent/purpose in your dashboard!
Choose defaults wisely	Interactivity with your dashboard should **NOT** be mandatory! When your audience first arrive at your app, self-sufficient.
Less is more	Resist the urge to "plot everything in every way for every category/option/filter. Go back to the "purpose" of the dashboard, make sure you stay true to that. Put cool charts you want people to look at in an appendix, or build a second app.
Add a narrative and signposts	Have a conversation with your reader, add sign-posts, consider adding a "reset/ home/defaults" button so they can always get back to the main point if they mess around too much.
Aesthetics matter!	Styling, branding, colour schemes (including colour-blind friendly), typography, layout, user interface (UI) and experience (UX) matter! Think hard about them and make good choices. Find the right balance between aesthetics and f[illegible]ionality.



Principles of Effective Dashboards II

Principle	Explanation
Build trust in your analysis	Think about ways you can increase transparency of your data sources and analysis methods. Be upfront about missing data and accuracy of your data. Add tooltips so users can check data.
Think about the "onboarding" experience	What happens when users first visit your site? Related to "set good defaults" but more than that: how do they use it? Where are the controls? What do they do?
Use a consistent layout	Do not burden your users by making them think about the layout of your app and how it's structured ; should be natural!
Use animations sparingly	Animations can be distracting, use them if you think it will help drive your point home (e.g., prison parole example)
Allow users to filter data (if applicable)	If you start with a giant dataset - say, the gapminder dataset - allow users the ability to filter the data and show data for the country they are interested in; have a good default comp
User testing is critical!	Get someone to look at your dashboard during development. Ideally someone who will be using it

Example: Incarceration

FiveThirtyEight

ABC NEWS

Politics Sports Science & Health Economics Culture

Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?

By Anna Maria Barry-Jester, Ben Casselman and Dana Goldstein

Graphics by Matthew Conlen, Reuben Fischer-Baum and Andy Rossback

Filed under Criminal Justice

Published Aug. 4, 2015

Source: FiveThirtyEight

Reproducibility and Transparency

The reproducibility crisis in science

A statistical counterattack

More people have more access to data than ever before. But a comparative lack of analytical skills has resulted in scientific findings that are neither replicable nor reproducible. It is time to invest in statistics education, says **Roger Peng**

Reproducibility in Data Science

Making research reproducible

There are two major components to a reproducible study: that the raw data from the experiment are available; and that the statistical code and documentation to reproduce the analysis are also available. These requirements point to some of the problems at the heart of the reproducibility crisis.

1. Raw Data (anonymized if needed)
2. Code to reproduce the analysis

Source: The Reproducibility crisis in science by Roger Peng

Some tools available to make reproducible Data Science analyses

docker

R Markdown

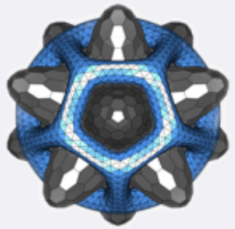
from RStudio

GNU Make

git

Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

binder



The Journal of Open Source Software

Summary

1. Data Visualizations

2. Building Narratives of Data Analyses

3. Reproducibility and Transparency