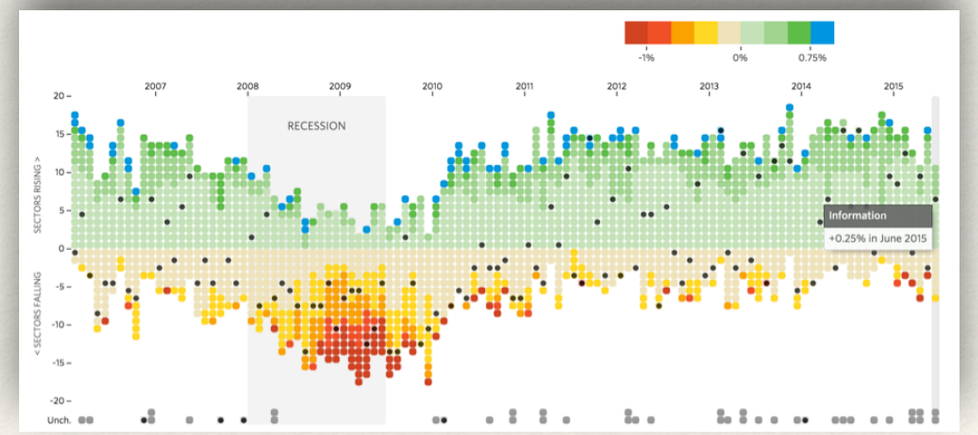
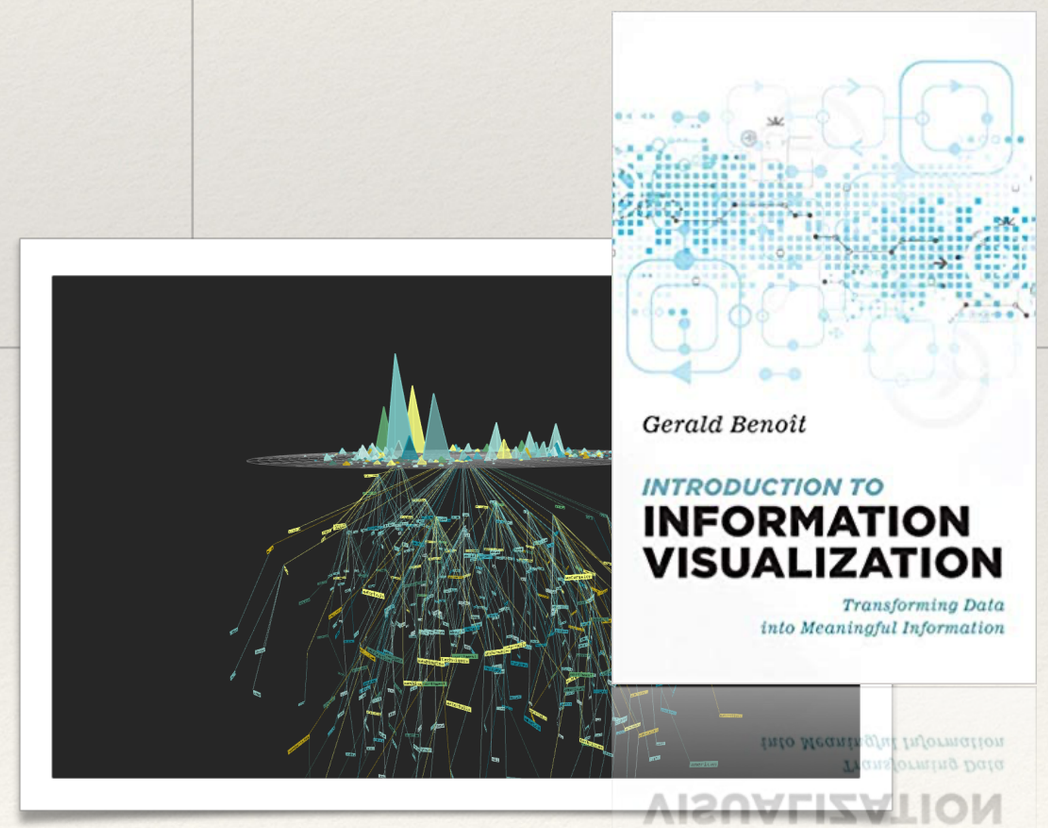


UC Berkeley
SCHOOL OF INFORMATION

Module 12

Plotting & Data Visualization

Programming for Data Science



Plotting & Data Visualization

UC Berkeley
SCHOOL OF INFORMATION

- IMHO … "information" is the result of human cognition, engagement with, and establishing significance from, the data, applicable to some need, and which enables a person to explain the rationale behind the significance. Otherwise, it's all just data.
  - Tendency to use as synonyms 'cause our goal is to provide data as quickly and appropriately as possible to fulfill some need (hence "info").
- Visualization is a language for communicating & helping understand data.

9	3-Mar	4-Mar	5-Mar	7-Mar	Unit 9	Text and Binary Data					Exam 1		
10	10-Mar	11-Mar	12-Mar	14-Mar	Unit 10	NumPy - Vectors	Project 1 Presentation	HW unit 9					Project 1 Code
11	17-Mar	18-Mar	19-Mar	21-Mar	Unit 11	Pandas - Dataframes		HW unit 10	HW unit 9			Project 2	
	24-Mar	25-Mar	26-Mar	28-Mar		Spring Break - no classes!							
12	31-Mar	1-Apr	2-Apr	4-Apr	Unit 12	MatPlotLib - Data Visualization		HW on units 11-13	HW unit 10				Project 2 Proposal
13	7-Apr	8-Apr	9-Apr	11-Apr	Unit 13	Advanced Pandas - Aggregation & Groups			HW units 11-13	Exam 2			
14	14-Apr	15-Apr	16-Apr	18-Apr	Unit 14	Testing	Project 2 Presentation				Exam 2		Project 2 Report

Plotting & Data Visualization

UC Berkeley
SCHOOL OF INFORMATION

Data Visualization:

- illustrates raw values;
- delivers data in intelligible graphic forms;
- offers objectivity;
- suggests trends;

DataVis refers primarily to specific techniques to represent *data* in a visual language - graphs, plots, charts ... and more.

Successful visualizations (data + design) convey an overview of the data, help you to explain the raw data, explore the data to discover new events or association, and do so in aesthetic ways that do not distract the viewers and so confuse the message.

If you get lost in creating a visual representation of your data, ask yourself what are the implications (or the "so what?" factor) and focus on that message.

Capture, clean data

→ Uni-, bi-, or multivariate data? Domain / Range of the data

→ Lend itself to expected graphics? (time series -> histograms)

→ Data in Visual Form + Overlaying Layers to provide context (such as labels, legends, details-on-demand, etc.)

Discussion

UC Berkeley
SCHOOL OF INFORMATION

- What's your experience with plotting & data visualization?
- Any specific examples that you find helpful? How and why?
- Why and when is any kind of visualization helpful?
- What's beyond data vis/plotting?
  - Interactive information visualization

- In practice, do you see relationships between Data Science/Big Data/ Data Mining and Statistics? [SAS, SPSS, IBM, Oracle/Java, Tableau, you name it all offers white papers to inform clients of the need and their products!]

Discussion

UC Berkeley
SCHOOL OF INFORMATION

- What message do you want to convey through your visualizations?
- Explain - show the part/whole relationship of groups and contextualize their relationships (e.g., from low to high)
- Explore - visually through graphic techniques (color, placement, difference of visual elements (line, square, circle, icon, etc.)
- Predict - suggest possibilities; provide warrant for trends (in data mining called an "interesting event"; statistically-sound data)
- Help viewers establish meaningfulness - information!

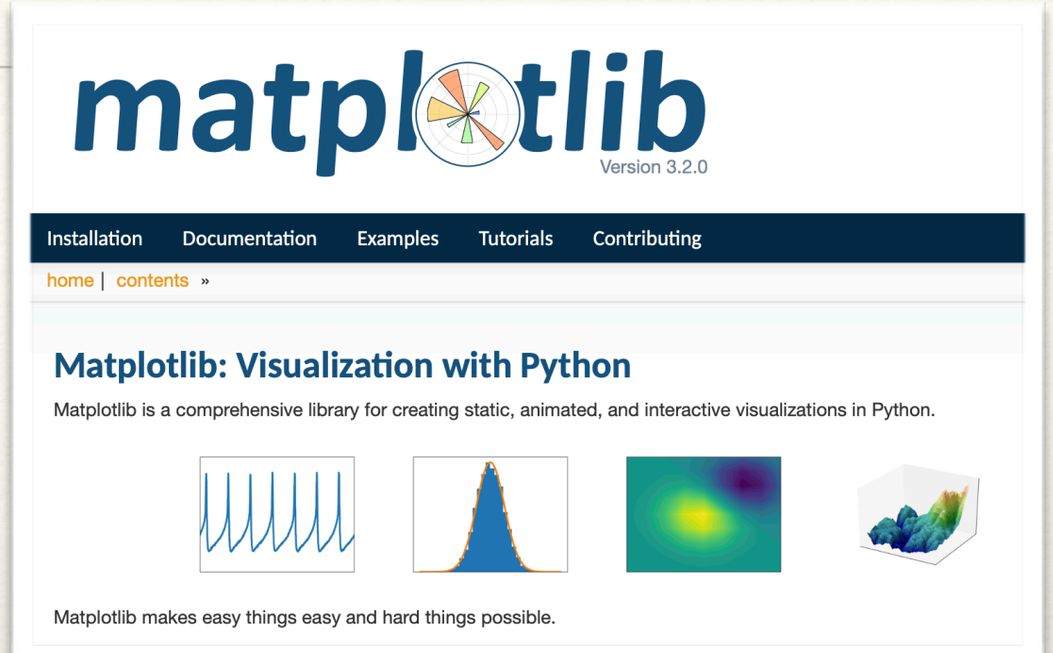
- Your choice: Use third-party products (for "business intelligence", etc.) or use libraries to create your own visualization product (and control the visualization experience). Lots of factors impact our choices.

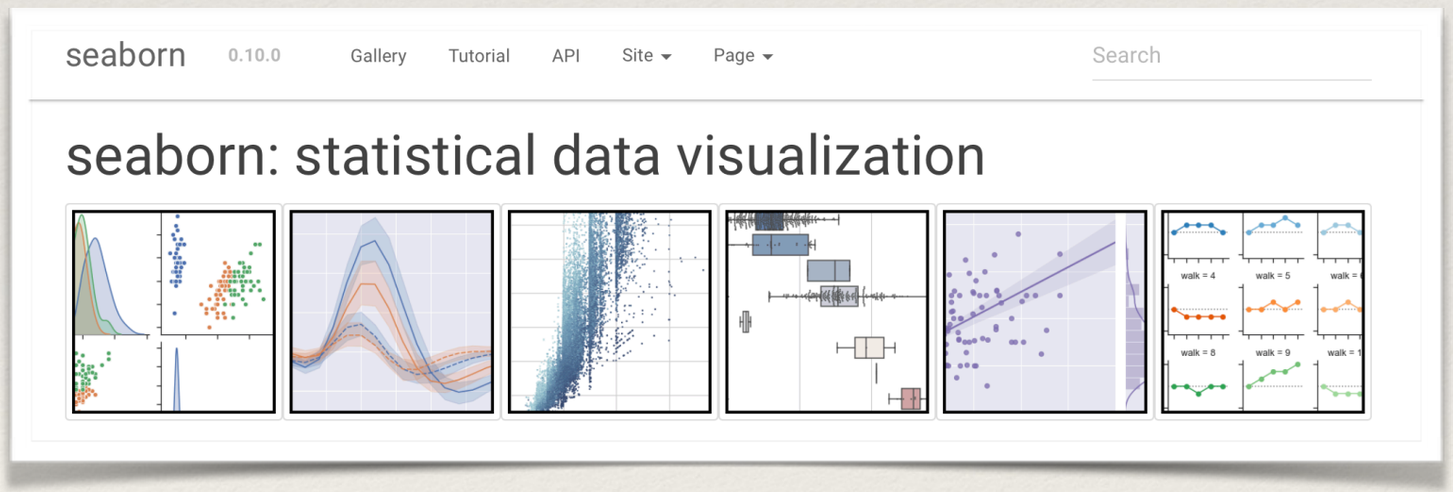
Matplotlib & Seaborn

UC Berkeley
SCHOOL OF INFORMATION

- Two of our most popular libraries for python. Use 'em to explore your data.

https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html.





- Here's a great Seaborn tutorial https://jovianlin.io/data-visualization-seaborn-part-1/

Other python libraries

UC Berkeley
SCHOOL OF INFORMATION

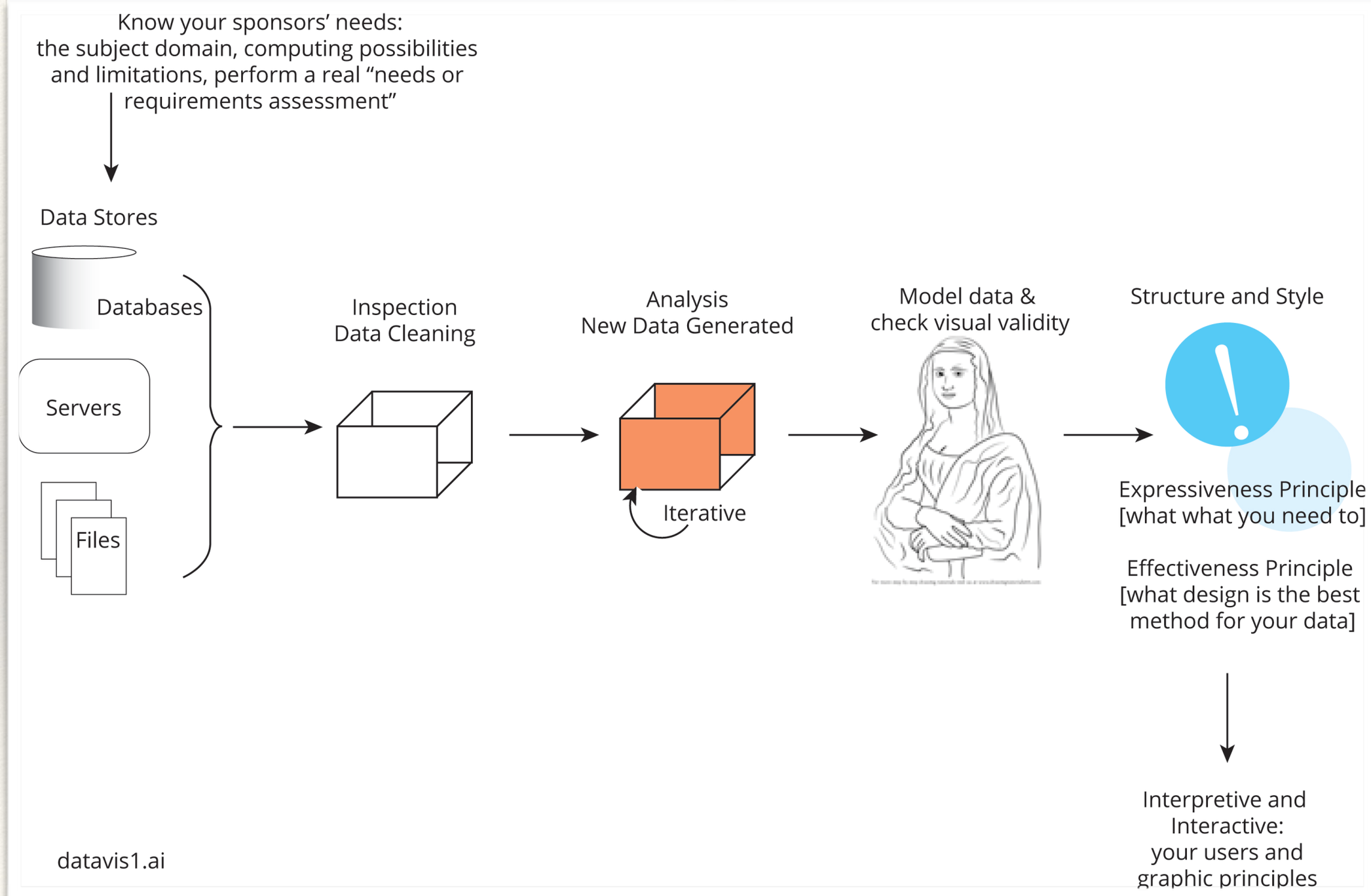
- matplotlib
- seaborn [on top of matplotlib; for statistics and cool visualizations]
- ggplot [library for R and python]
- bokeh [interactive vis]
- pygal
- plotly [interactive vis]
- geoplotlib [maps; chloropleth]
- gleam Other projects
- missingno
- leather
- basemap (for geographic data) [requires installing base map library; pillow (but not PIL) https://pillow.readthedocs.io/en/stable/installation.html

Another great tool of particular value is the JavaScript-oriented d3.js library. Check out Mike Bostock's Data-Driven Documents homepage and on GitHub. By saving your analyzed and processed data; it's easy to stream them into a file to be read later by a website calling the d3.js library.

Commercial products - Tableau is very popular.

Overview of a process

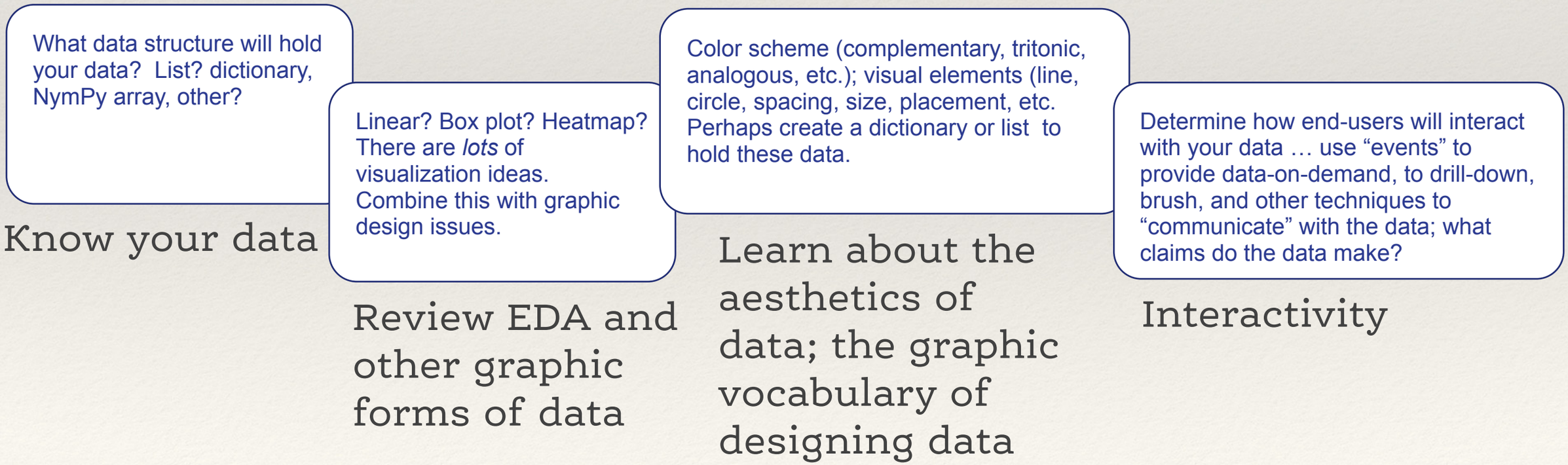
UC Berkeley
SCHOOL OF INFORMATION



Overview example

UC Berkeley
SCHOOL OF INFORMATION

- Using matplotlib as a the first example, let's say your data are stored in a .csv file … and that you've already checked for missing values, NaN, and values that are within range and domain, and that you know the data types of each of the columns & rows (whew!).



EDA = Exploratory Data Analysis

As a programmer, working from code to visual product

UC Berkeley
SCHOOL OF INFORMATION

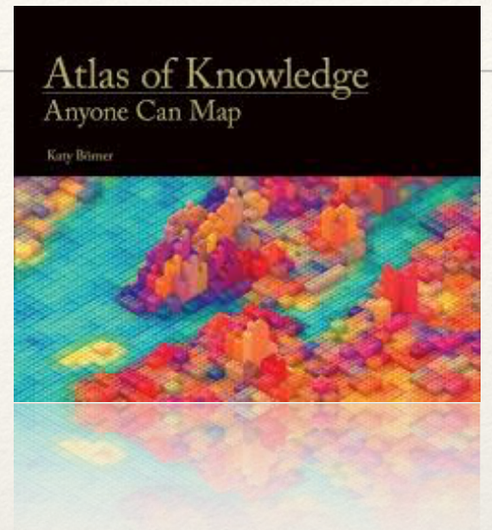
- Sketch your design first; what data where?
- Import your libraries
- Adjust for your IDE (e.g., % to display inline; "retina" option NB)
- Get yer data … prepare the output device to hold the visualizations.
- Create a "container" to hold the viz (e.g., "figure"; compare this to HTML5's canvas or SVG object)
- Contextualize the data in a framework (axes [placement, ticks], labels [titles, subtitles, subplots])
- Design a model and decide on the aesthetic elements (symbols, visual elements, colors, saturation)
- Shortcuts in the code? (e.g., use the code, such as color constants, templates)
- Adjust your design based on the user needs, truthfulness to interpretation of the data; test the plotted data.

Many visual options

UC Berkeley
SCHOOL OF INFORMATION

Dot Graph
Stacked Bar Graph
Proportional Symbol Map
Choropleth
Force-Directed Layout
Radar Graph
Stream Graph
Parallel Coordinate
Stacked Line Graph
Box-and-Whiskers
Radial Tree
Dasymetric
Circular Graph
Tag Cloud
Line Graph
Tree View
Cartogram
Force-Directed Layout
Hive Graph

Pie Chart
Stepped Relief Map
Balloon Tree
Stripe Graph
Arc
Scatter
Bimodal
Crossmap
Steam and Leaf
Bubble Chart
Doughnut
Dot Density
Isoline
Chord
Sankey
Proportional Symbol
Treemap
Arc Graph
History Flow



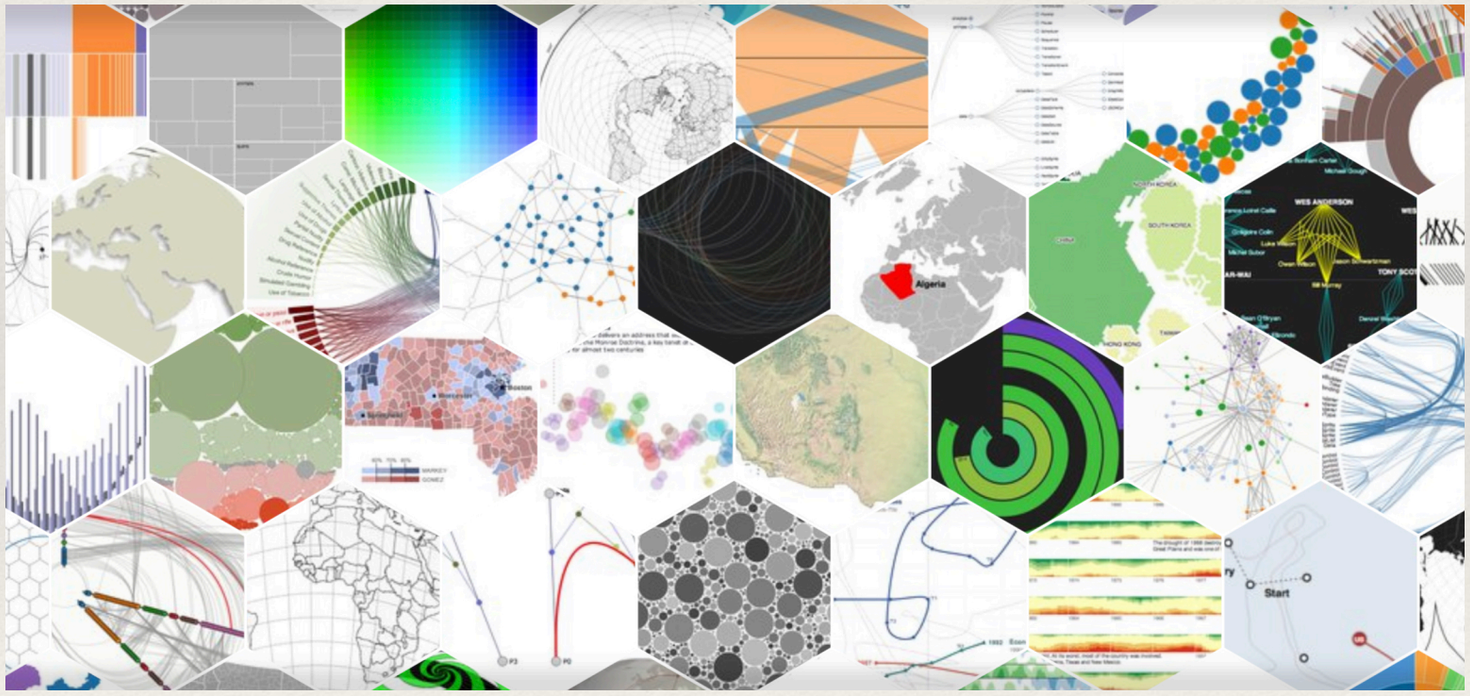
Börner, K. *Atlas of knowledge*. MIT Press

Elevation
Histogram
Science
Isarithmic
Strip Map
Isochrome Map
Population Pyramid
Enclosure Tree
Dendrogram

JavaScript example

UC Berkeley
SCHOOL OF INFORMATION

- Not to be disparaged! Prepare your data; output as .json. Have your webpage (with javascript) just load your data and ta-da! A visualization! [and we have a demo]



Check out d3js.org for a sample of the many types available.

Python example

UC Berkeley
SCHOOL OF INFORMATION

```
import matplotlib
import matplotlib.pyplot as plt
%config InlineBackend.figure_formats = ["retina"]
import numpy as np

x = np.linspace(0, 5, 10)
y = x ** 2

figure()
plot(x, y, 'r')
xlabel('x')
ylabel('y')
title('title')
show()

fig = plt.figure()

axes = fig.add_axes([0.1, 0.1, 0.8, 0.8]) # left, bottom, width, height (range 0 to 1)

axes.plot(x, y, 'r')

axes.set_xlabel('x')
axes.set_ylabel('y')
axes.set_title('title');
```

read each line - know why it is there.

what's up here?

… breakout activity time

Examples & Resources

UC Berkeley
SCHOOL OF INFORMATION

- In the resources folder
  - DataVis_1_Matplotlib.ipynb
  - DataVis-extra-fyi.ipynb
  - Python for Data Analysis (text)
  - "Cheet Sheets"

- InfoVis texts tend to cluster by some main themes:
  - "Empirical" school (Stanko, Ware)
  - Data over People (Mutzner, Illinsky & Steele)
  - Communication (Benoît, Heer)

Resources; White Papers

UC Berkeley
SCHOOL OF INFORMATION

- Benoît (2019) *Introduction to information visualization*
- Ware *Information visualization* and *Design for visual thinking*
- Tufte *Visual display of quantified data*, *Visual explanation*, *Beautiful evidence*, *etc.*
- Steele & Illinsky *Beautiful information* [series of articles]

- Börner, K. *Atlas of knowledge* and *Atlas of science* [various demos of one-off examples]
- Munzner *Visual analysis and design*

- *IEEE Visualization* [professional organization]
- *ACM SIGVIS* [*professional organization*]
- *Information visualization* (journal)

- Gartner "Not all BI platforms are created equal"

- Bobriakov (2018) Comparative analysis of ...

- Tableau (homepage; download and try for free)

About 10 years ago, *ACM*'s lead article was "Should we teach graphic design to computer science students?"

Summary

UC Berkeley
SCHOOL OF INFORMATION

- "data" versus "information"
- Your exposure and your & others' growing expectation of visualizations in data-rich environments
- Explain, Explore, Predict … with statistical and other data to back up and clarify potential interpretations
- Code-your-own: matplotlib, seaborn, many others for python
- JavaScript and others for web-based presentations
- Tableau and other products
- Know your data; know your audience; learn graphic design (!) and integrate the whole in your code.

Enjoy the Beautiful in Information!