Exploratory Data Analysis & Visualization

XX-161-A-21 – Big Data Analytics
Dr. Jim Scrofani
jwscrofa@nps.edu

Topics at a Glance

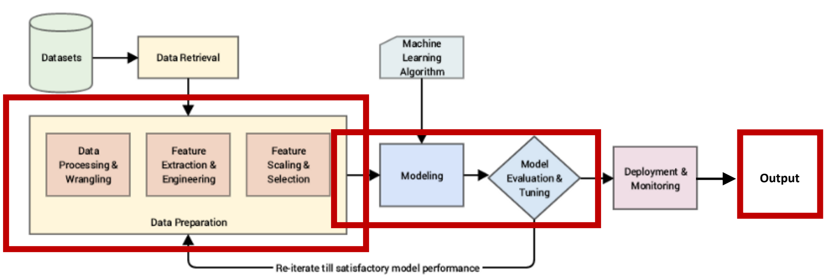
- Data Science Workflow
- Exploratory Data Analysis (EDA)
- Data Summarization
- Visualization
- Domain-specific Representations
- Matplotlib

Data Science Workflow

3

3

Where are we?



A standard machine learning pipeline (source: Practical Machine Learning with Python, Apress/Springer)

4

4

Exploratory Data Analysis

Motivation

- ***Exploratory data analysis*** (EDA) is "detective work - numerical detective work - or counting detective work - or graphical detective work."

  John W. Tukey [1]

- "A philosophy of data analysis where the researcher examines the data without any preconceived ideas in order to discover what the data can tell about the phenomena being studied" [2]

What is Data Exploration?

- **Preliminary exploration of the data to better understand its characteristics**
- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and are also integral to algorithmic solutions
- In our discussion of data exploration, we focus on
  - Data summarization -- Summary statistics
  - Data visualization

Exploratory vs. Confirmatory Analysis

- EDA complements confirmatory data analysis (CDA)
  - Concerned with statistical hypothesis testing, confidence intervals, estimation
  - Uses traditional statistical methods
- "Confirmatory data analysis is judicial or quasi-judicial in character. [1]"
- CDA methods make inferences about or estimates of some population characteristic and then evaluate precision of results
- "Do the data confirm hypothesis XYZ?" Whereas, EDA tends to ask "What can the data tell me about relationship XYZ? [4]

Good Read

For students that are seeking more information:

Exploratory Data Analysis

https://www.itl.nist.gov/div898/handbook/eda/eda.htm

Data Summarization

11

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set
  - Can be obtained from the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Iris
  - From the statistician R.J. Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute. 12

Iris Raw Data

Raw Data Matrix

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
5.0	3.6	1.4	0.2	Setosa
		…		

Data Summarization

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
    - Examples: *location* - mean
      spread - standard deviation
  - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The *frequency* of an attribute value is the percentage of time the value occurs in a dataset
  - For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time
- The *mode* of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

15

Percentiles

- For continuous data, the notion of a *percentile* is more useful
  - Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value x_p of x such that $p\%$ of the observed values of x are less than x_p
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

16

Measures of Location: Mean and Median

- The *mean* is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the *median* or a *trimmed mean* is also commonly used

Let $\{x_1, \ldots, x_m\}$ be the attribute values of x for m objects. Further, let $\{x_{(1)}, \ldots, x_{(m)}\}$ represent the values after they have been sorted in non-decreasing order, thus, $x_{(1)} = \min(x)$ and $x_{(m)} = \max(x)$. Then:

$$\text{mean}(x) = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- *Range* is the difference between the *max* and *min*
- The *variance* or *standard deviation* is the most common measure of the spread of a set of points

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

- Because of outliers, other measures are often used:

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$$

$$\text{MAD}(x) = \text{median}\Big(\{|x_i - \bar{x}|, \ldots, |x_m - \bar{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Multivariate Summary Statistics

- Data comprised of several attributes is multivariate data
- Measure of location:

$$\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n), \text{where } \bar{x}_k \text{ is the } k^{\text{th}} \text{ attribute of } x_k$$

- Measures of spread captured by *covariance matrix*

S, whose ij^{th} entry s_{ij} is the covariance of the i^{th} and j^{th} attributes of the data.

$$\text{covariance}(x_i, x_j) = s_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

More on Data Summarization

For students that are seeking more information:

A Gallery of Quantitative Techniques

https://www.itl.nist.gov/div898/handbook/quantgal.htm

Visualization

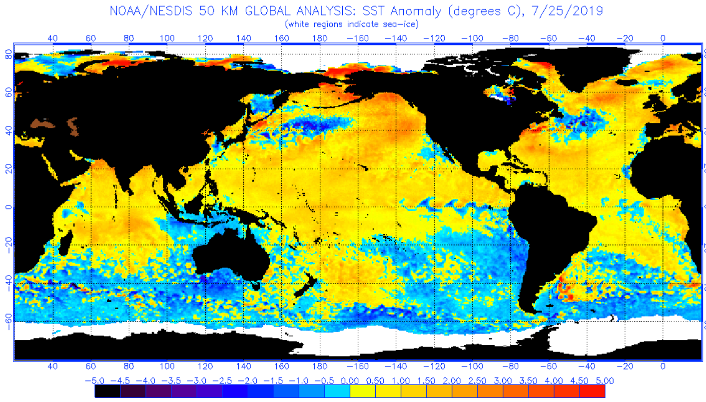
Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported

- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well-developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 2019 [5]
  - Thousands of data points are summarized in a single figure



Why Build Visuals?

Data visualization is a way to show complex data in a form that is graphical and easy to understand

[1] https://cognitiveclass.ai/

Why Build Visuals?

- For exploratory data analysis
- To communicate data clearly
- To share an unbiased representation of data
- Use them to support recommendations to different stakeholders

[1] https://cognitiveclass.ai/

Best Practices

When creating a visual, always remember:

1. Less is more effective
2. Less is more attractive
3. Less is more impactive

[1] https://cognitiveclass.ai/

Best Practices

Review the slides at Darkhorse Analytics

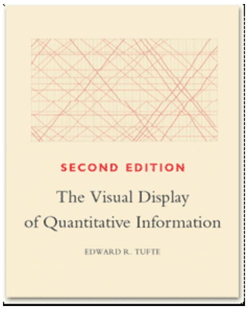
– Data Looks Better Naked (click for the link)

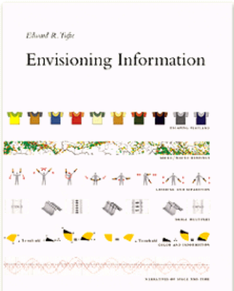
Note that adhering to the *best practices* renders a much more effective visual presentation

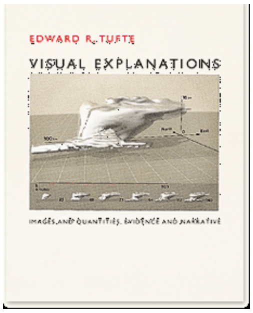
Check out the other blog postings…

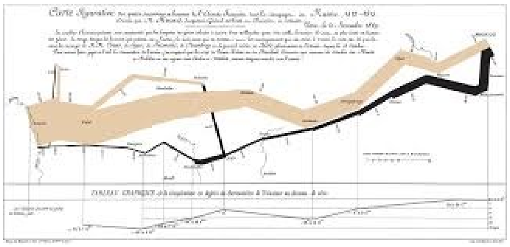
https://www.darkhorseanalytics.com/blog/salvaging-the-pie

Data Visualization Guru, Edward Tufte









Representation

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

29

Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

30

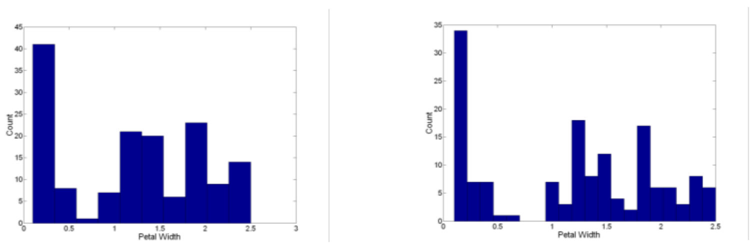
Selection

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve the choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

31

Visualization Techniques: Histograms

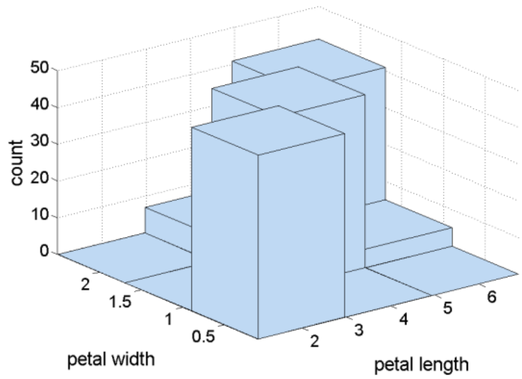
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)



32

Two-Dimensional Histograms

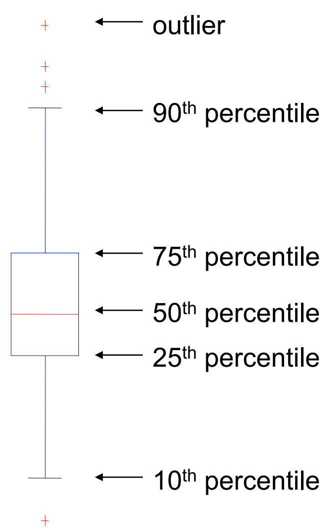
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



Visualization Techniques: Box Plots

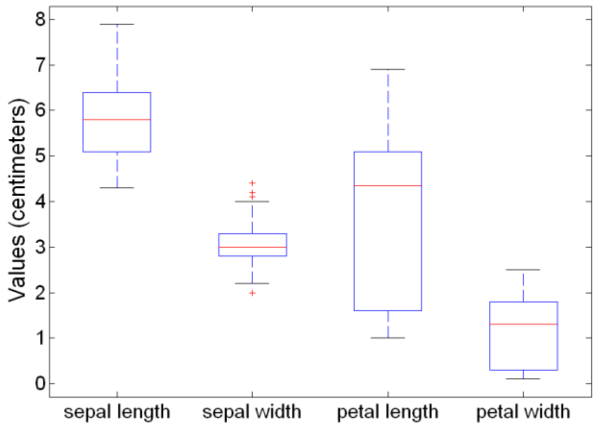
Box Plots

- Invented by J. Tukey
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot



Example of Box Plots

Box plots can be used to compare attributes



35

35

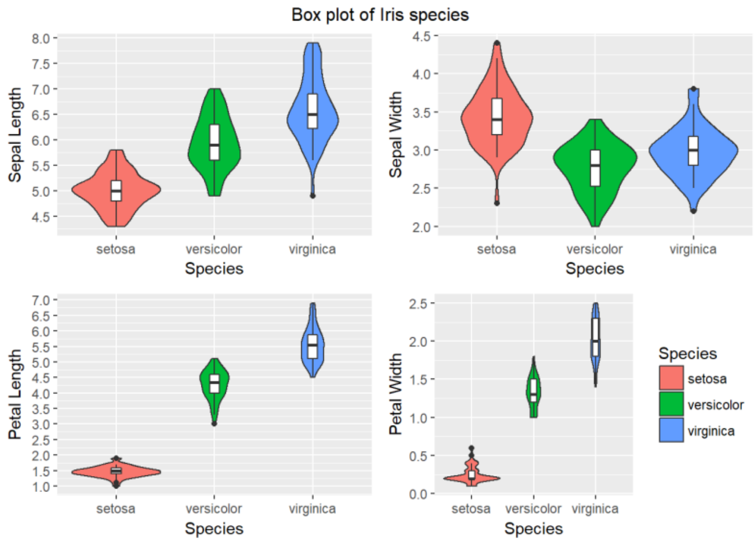
Example of Violin Plots

- Similar to box plot, with the addition of a rotated kernel density plot on each side
- Kernel density plot is based on kernel density estimation, which is a non-parametric way to estimate pdf of a random variable
- NIST reference on visualizations: https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/violplot.htm

36

36

Example of Violin Plots

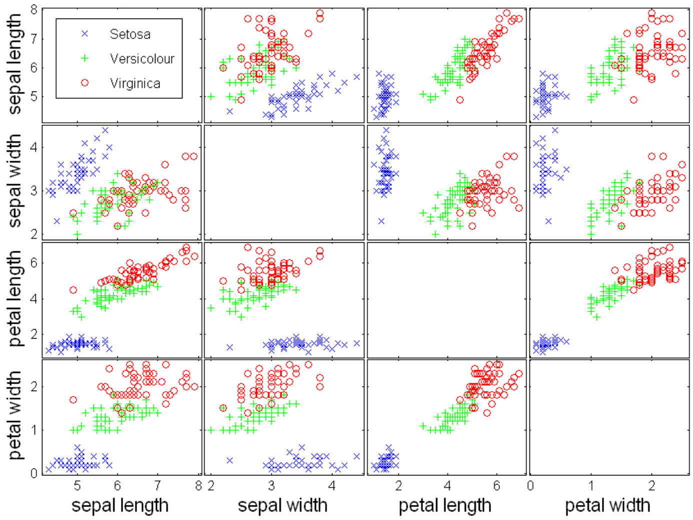


Visualization Techniques: Scatter Plots

Scatter plots

- Attributes values determine the position
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
  - See example on the next slide

Scatter Plot Array of Iris Attributes



39

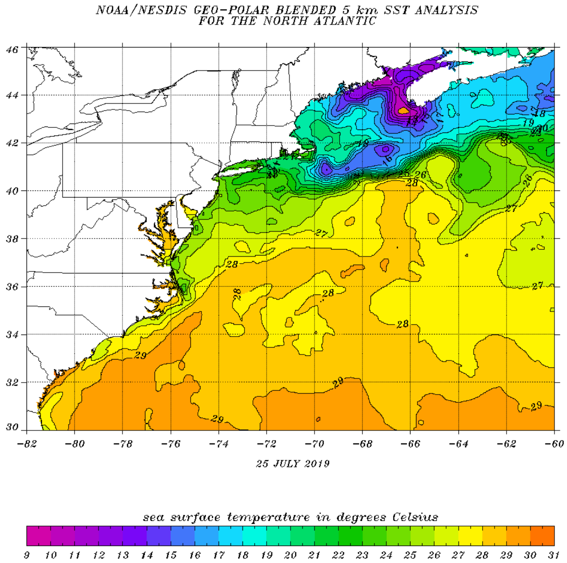
Visualization Techniques: Contour Plots

Contour plots

- Useful when a continuous attribute is measured on a spatial grid
- They partition the plane into regions of similar values
- The contour lines that form the boundaries of these regions connect points with equal values
- The most common example is contour maps of elevation
- Can also display temperature, rainfall, air pressure, etc.
  - An example for Sea Surface Temperature (SST) is provided on the next slide

40

Contour Plot Example: SST July 2019 [5]

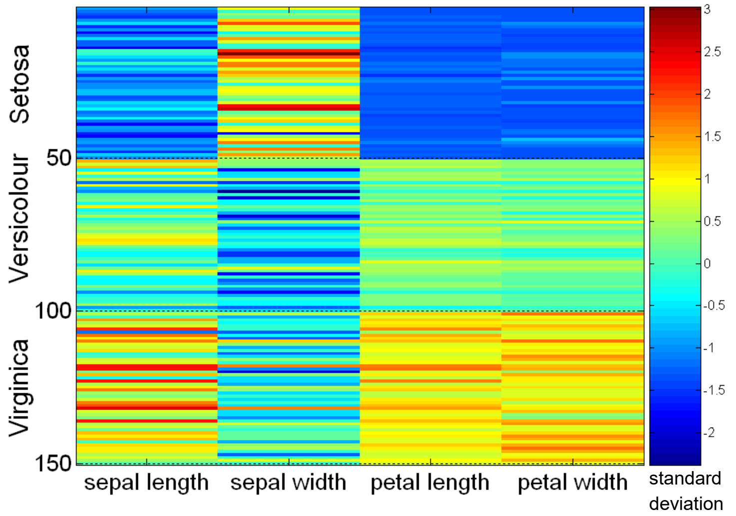


Visualization Techniques: Matrix Plots

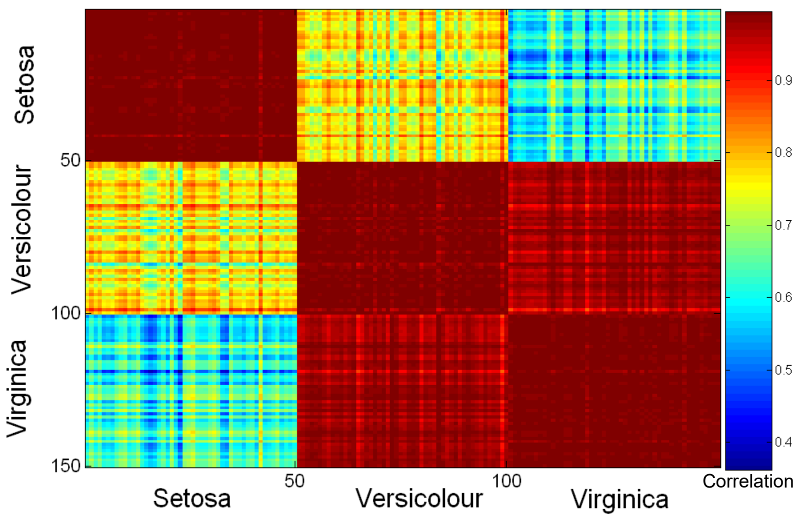
Matrix plots

- Plots of the data matrix
- This can be useful when objects are sorted according to class
- Typically, the attributes are normalized to prevent one attribute from dominating the plot
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
- Examples of matrix plots are presented on the next two slides

Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix



More on Visualization

For students that are seeking more information:

A Gallery of Graphical Techniques

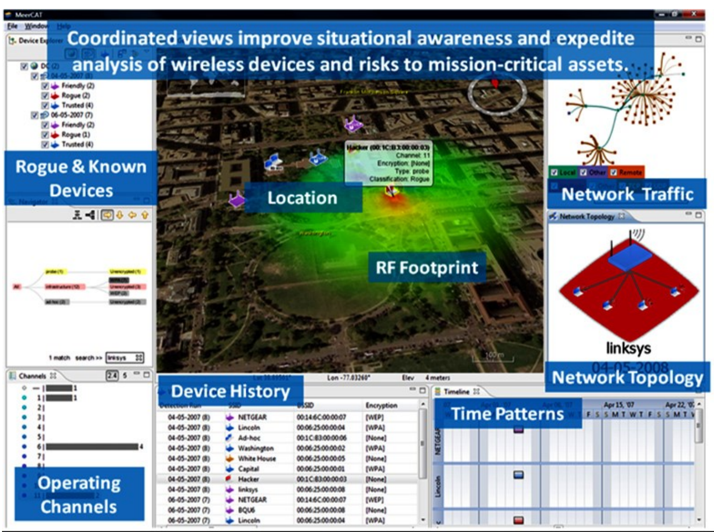
https://www.itl.nist.gov/div898/handbook/graphgal.htm

Domain-specific Representations

Cyber-domain Representations

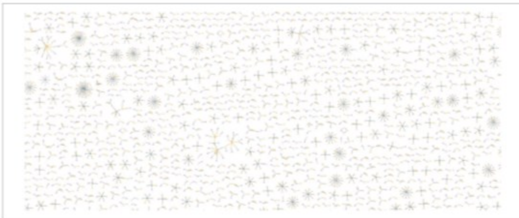
- Network traffic
  - Packet characteristics (protocol fields, IP addresses, Port numbers)
  - Temporal aspects (volume/unit time)
- Network topology
  - End-points
  - L3 devices – routers
  - L2 devices – switches
- Host-based attributes
  - OS/Services/Applications
  - Accounts

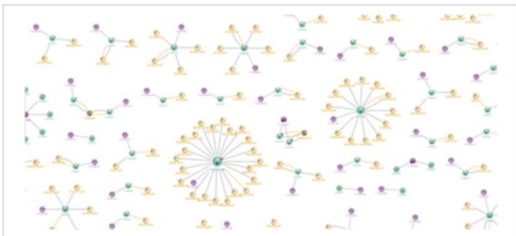
Cyber SA requires complex dashboards



https://securedecisions.com/products/meercat/

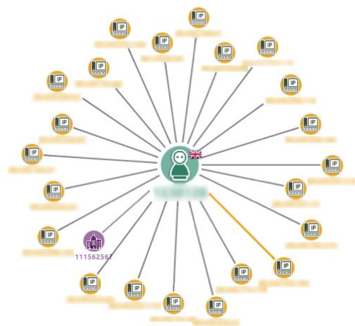
Visualizing User Account Login Data



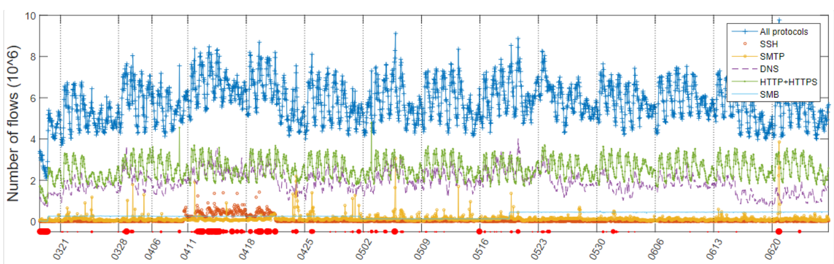


Normal users login from 1-4 IP addresses, so users w/ 20 IP's logging in are out of the ordinary, and merit further exploration.

Source: Cambridge Intelligence,
"Visualizing Cyber Threats with Key Lines"



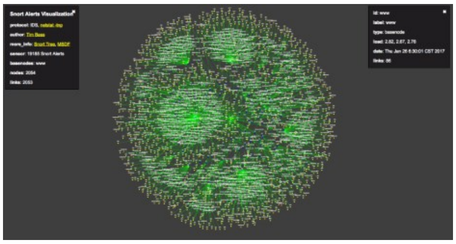
Visualizing Anomalies in Network Flows



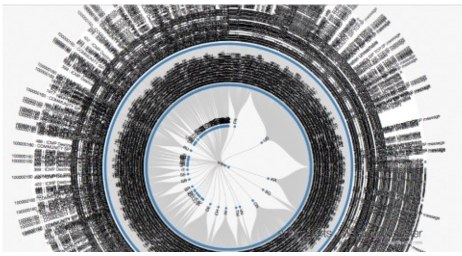
Cyclostationary dataset – normalized for periodic activity
Red dots show anomalies – mostly high SSH traffic and email spam campaigns.
Size of the dot is proportional to volume of anomalous traffic.

Source: UGR'16 Calibration Dataset

Cyber Data is often Big Data: Hard to represent in meaningful figures



2D Javascript D3 force-directed graph visualization over 19000 Snort IDS alerts represented by around 2000 nodes and 2000 edges.



The same set of over 19000 Snort IDS alerts represented by around 2000 nodes and 2000 edges, created as a radial-tree cluster using D3.

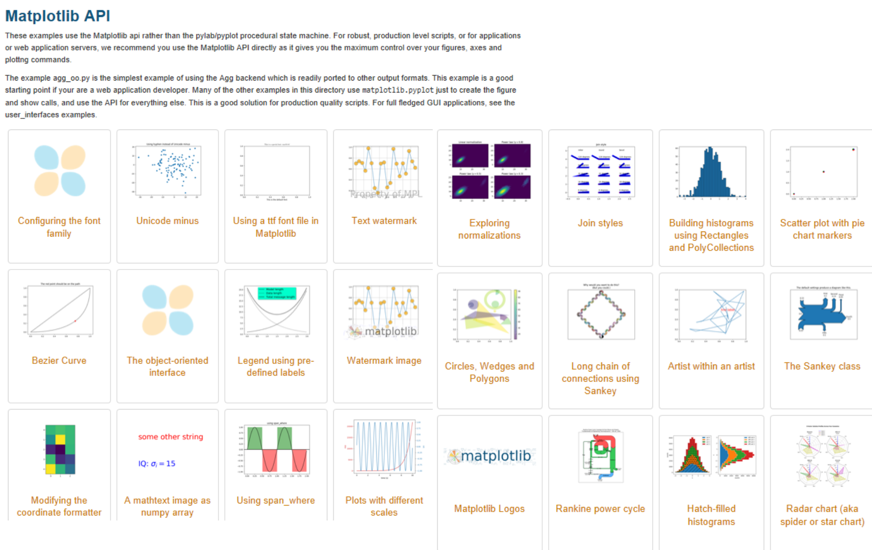
https://www.cyber-situational-awareness.com/2017/01/26/three-views-of-noisy-ids-alert-data-the-scaling-problem-for-big-data/

matplotlib

cheatsheet

https://github.com/matplotlib/cheatsheets#cheatsheets

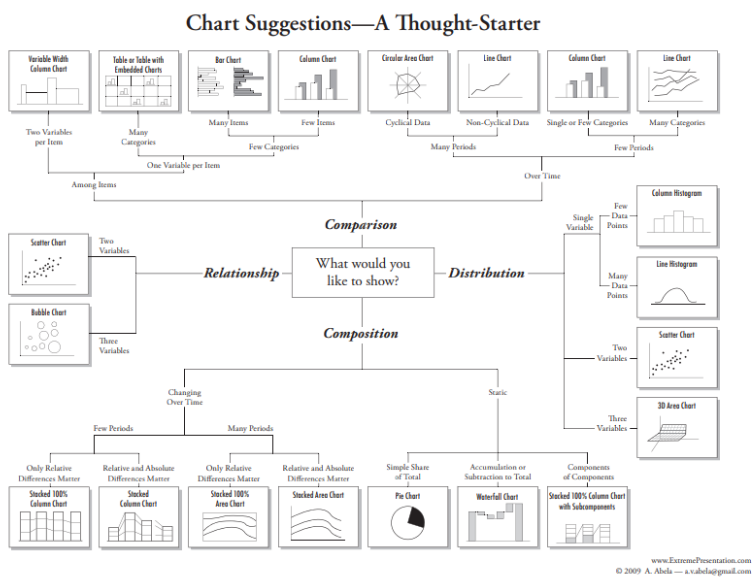
Matplotlib offers TONS of options! [7]



53

53

So how do you choose? [8]



54

54

Recap

- Data Science Workflow
- Exploratory Data Analysis (EDA)
- Data Summarization
- Visualization
- Domain-specific Representations
- Matplotlib

References

[1] Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley

[2] Martinez & Martinez (2005). Exploratory data analysis with MATLAB, CRC Press

[3] Engineering Statistics, https://www.itl.nist.gov/div898/handbook/index.htm, last accessed 5/7/2021

[4] Hartwig and Dearing [1979]. Exploratory Data Analysis, Sage University Press.

[5] Current Operational SST Anomaly Chart https://www.ospo.noaa.gov/Products/ocean/sst/anomaly/index.html, last accessed 5/7/2021

[6] NIST Dataplot Reference https://www.itl.nist.gov/div898/software/dataplot/, last accessed 5/7/2021

[7] https://matplotlib.org/3.1.1/api/index.html, last accessed 7/29/2019

[8] https://extremepresentation.typepad.com/, last accessed 7/29/2019