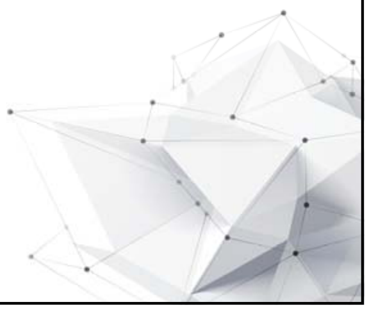
Python for Data Science
Data Handling
Data Visualization

Vetria L. Byrd, Ph.D.
Assistant Professor
Purdue University

Data Science Bootcamp
Georgia Tech
Atlanta, GA

Tuesday, August 6, 2019

PURDUE
POLYTECHNIC



About Me
Vetria L. Byrd, PhD

Academic Preparation

- Computer Science (PhD, MS)
- Biomedical Engineering (MSMBE)

UAB THE UNIVERSITY OF
ALABAMA AT BIRMINGHAM
Knowledge that will change your world

Visualization Initiatives

- Research Experience for Undergraduates in Collaborative Data Visualization Applications (2014/2015)

CLEMSON
UNIVERSITY

Datanauts

Reach higher and explore deeper. Whether you're a software engineer or coding newbie, join our NASA Datanaut community to engage with each other and subject matter experts to solve data challenges. It really IS Rocket [Data] Science.

Water Cooler Chat

MIDWEST
BIG DATA
HUB
ACCELERATING THE BIG DATA
INNOVATION ECOSYSTEM

Steering Committee Member
2016 -1018

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Visualization Webinars

International HPC Summer School
on HPC Challenges in Computational Sciences

Toronto, Canada (2015), Ljubljana, Slovenia (2016), Boulder, CO, US (2017)

HPC wire
Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

Byrd Emphasizes Value of Visualization at XSEDE14
By Trish Barker

July 31, 2014

About Me

Vetria L. Byrd, PhD

Since joining Purdue

New Data Visualization Major for Undergraduates

Courses Taught/Teach

- Undergraduate
  - CGT 270 Data Visualization (for majors)
  - CGT 101 Foundations of Computer Graphics Technology
  - CGT 118 Fundamentals of Imaging Technology
- Graduate Courses
  - CGT 501 Graduate Seminar
  - CGT 575 Data Visualization Tools and Applications
  - CNIT 5700 Certification Course for Rolls Royce

Data Visualization

Learn the art and science of representing data-rich information in a format that enables users to understand, use, communicate, and take action.

Data Visualization

A major in the Computer Graphics Technology Program

https://polytechnic.purdue.edu/degrees/data-visualization

PURDUE POLYTECHNIC

Agent for "Insight"

Data Mine Data Visualization Living Learning Community

- Inaugural cohort this fall
- Goal: 800 students by 2020
- Requirement: Must be an undergraduate

Will incorporate Python Libraries showcased today into the fall 2019 courses.

Faculty Fellow for the Data Visualization Cohort

Fall 2019

- CGT 270 for Non-Majors
- CGT 290 Topics in Data Visualization

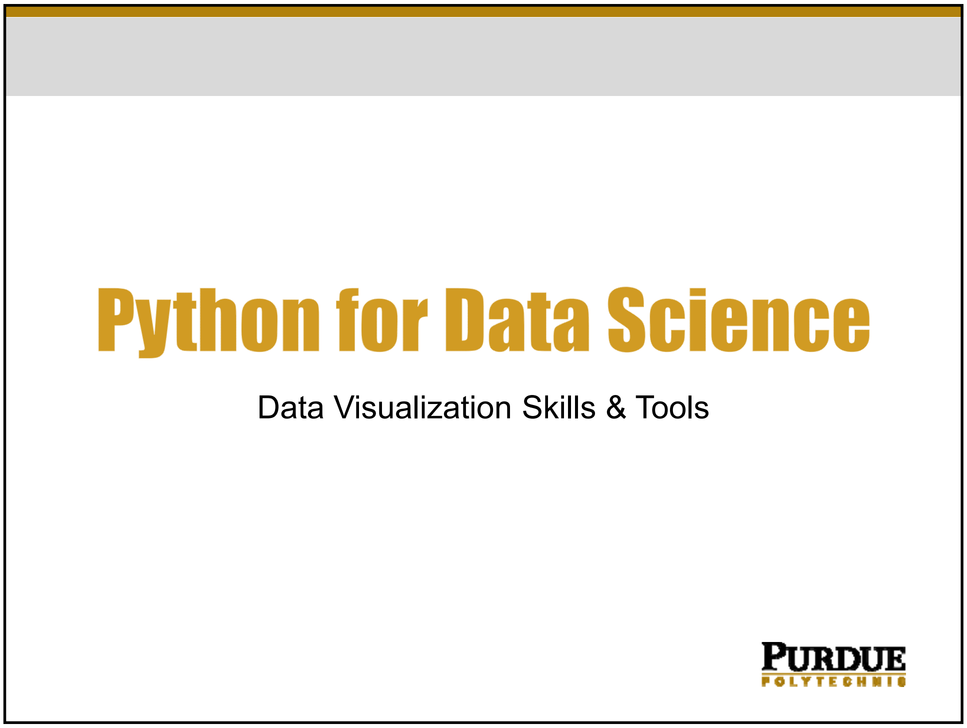
Spring 2020

- Advanced Data Visualization

PURDUE POLYTECHNIC

Will talk about my research on Friday

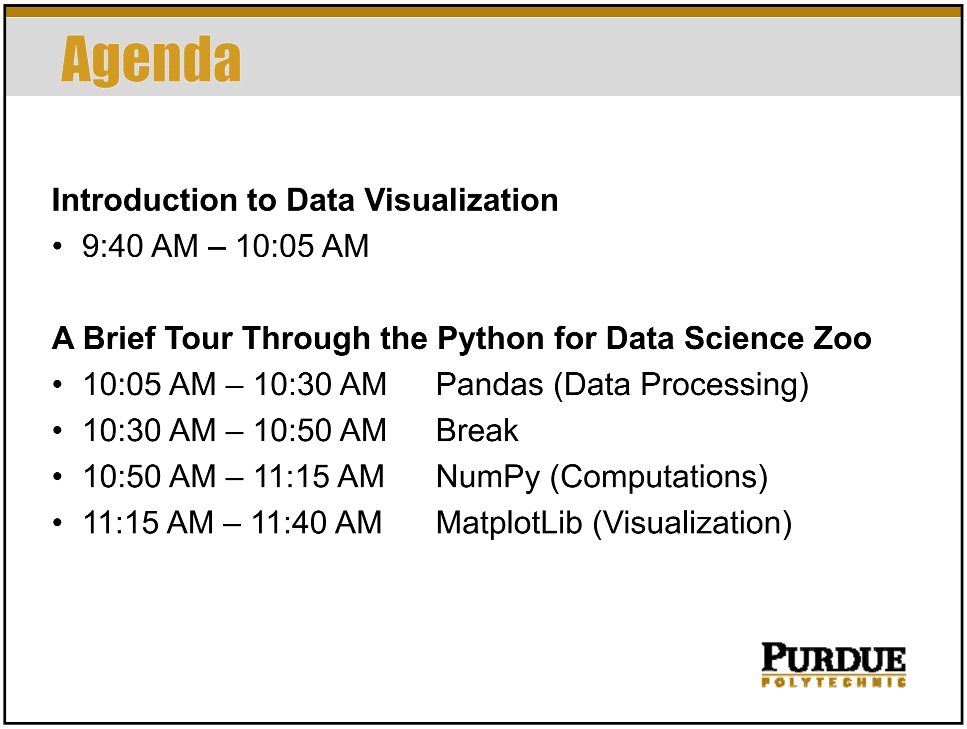
PURDUE POLYTECHNIC



Python for Data Science

Data Visualization Skills & Tools

PURDUE POLYTECHNIC



Agenda

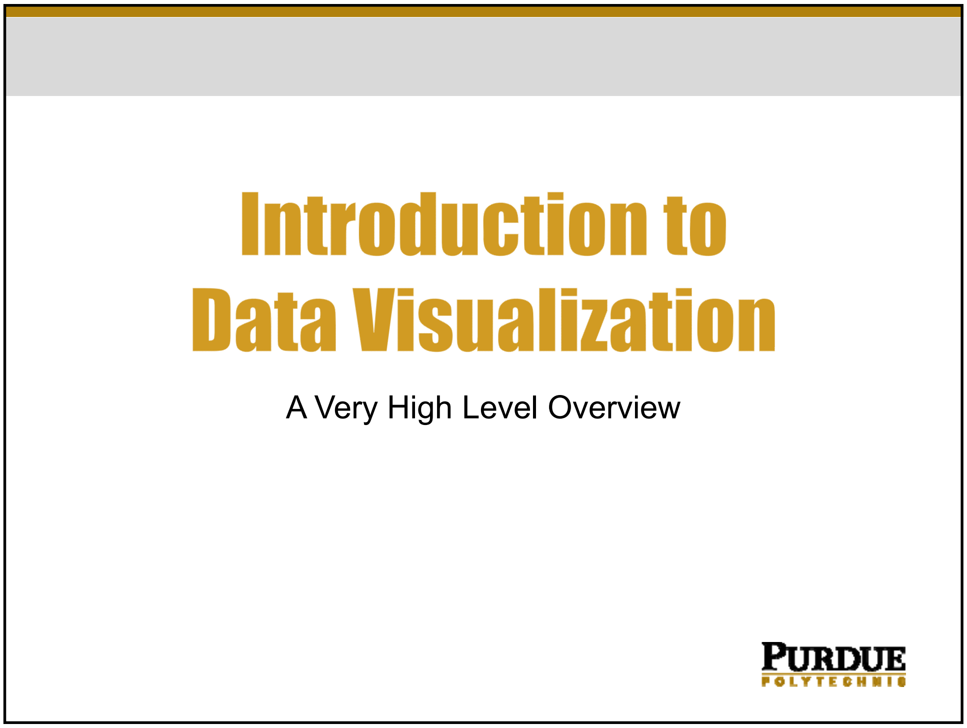
Introduction to Data Visualization

- 9:40 AM – 10:05 AM

A Brief Tour Through the Python for Data Science Zoo

• 10:05 AM – 10:30 AM	Pandas (Data Processing)
• 10:30 AM – 10:50 AM	Break
• 10:50 AM – 11:15 AM	NumPy (Computations)
• 11:15 AM – 11:40 AM	MatplotLib (Visualization)

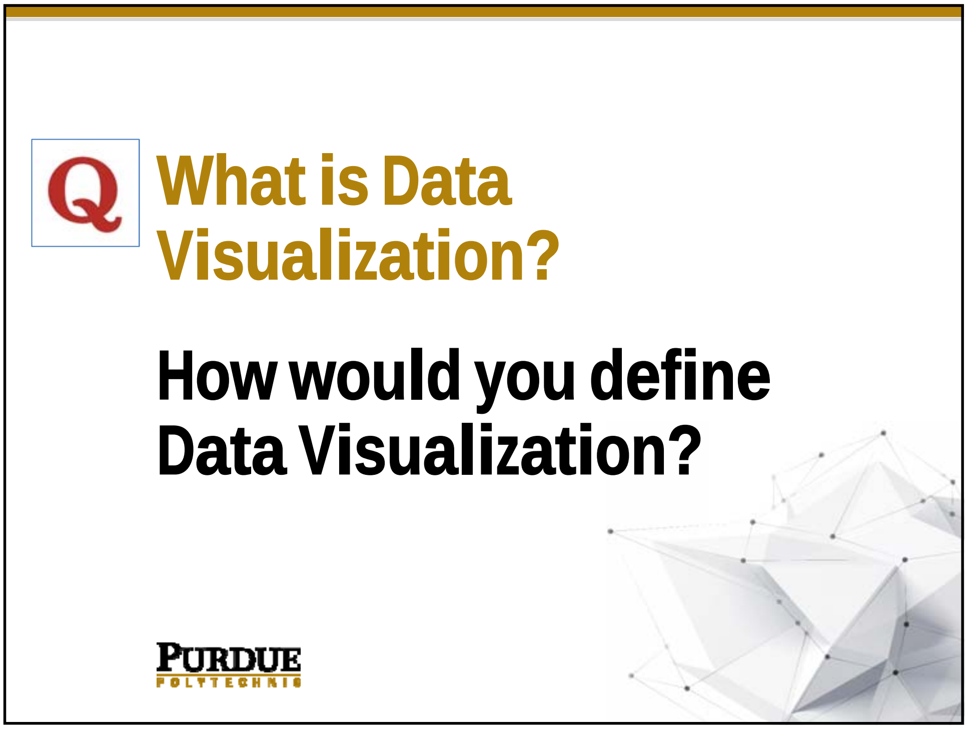
PURDUE POLYTECHNIC



Introduction to Data Visualization

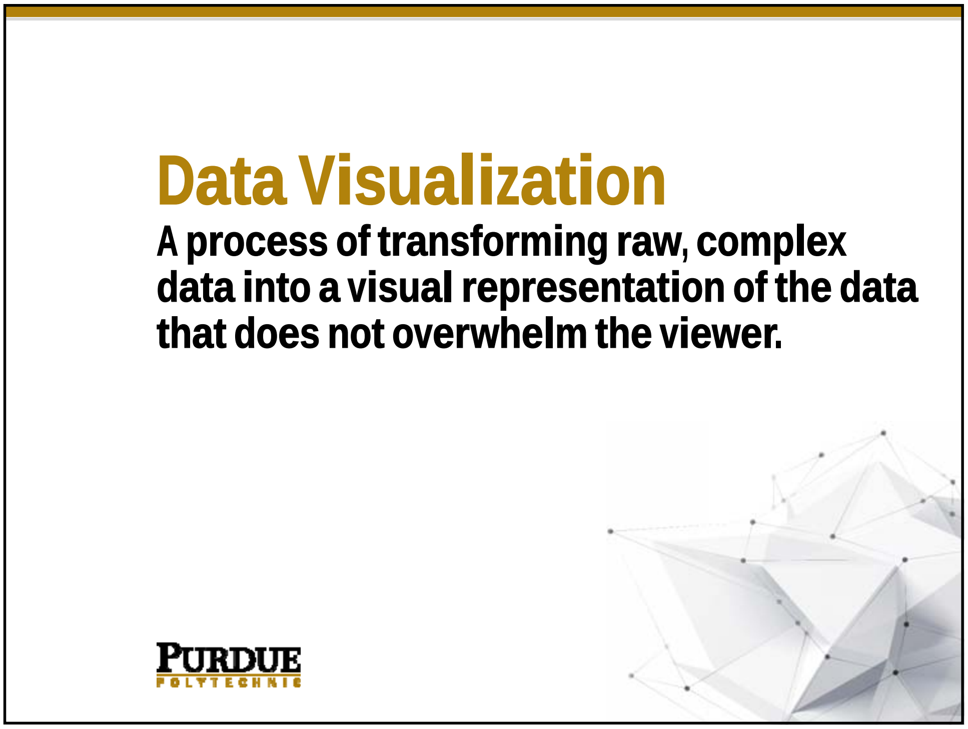
A Very High Level Overview

PURDUE POLYTECHNIC



What is Data Visualization?

How would you define Data Visualization?



Data Visualization

A process of transforming raw, complex data into a visual representation of the data that does not overwhelm the viewer.

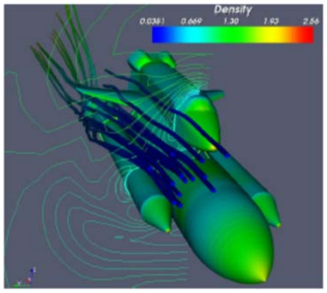
Data Visualization is

A process

Adopted from The ParaView Tutorial, The Basics of Visualization, version 3.98

Data Visualization is

A process of transforming raw, complex data into a visual representation that does not overwhelm the viewer.



Adopted from The ParaView Tutorial, The Basics of Visualization, version 3.98

Principles of Data Visualization

Objective

- Provide foundational understanding of how we process visual information

Outcomes

- Informed opinion on how to communicate more clearly and powerfully using visualizations
- Better analyze visualizations you come across in the newspaper, on the web or in your daily experience

Adopted from FusionCharts White paper, “Principles of Data Visualization - What We See in a Visual”

PURDUE POLYTECHNIC

Why We Visualize Data

PURDUE POLYTECHNIC

Why We Visualize Data

- To meet a very basic need – to tell a story
- One of the most primitive forms of communication known to man
- Cave drawings dated as early as 30,000 B.C.
- Even before written communication (3,000 B.C.)

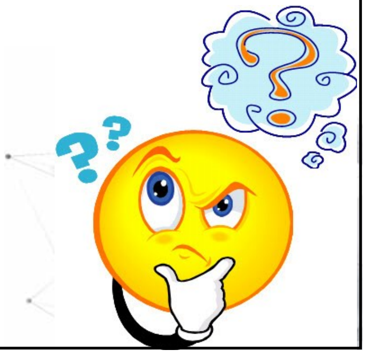
Today

- New ways to visualize information
- Basic chart types
  - Bar chart
  - Line chart
  - Pie chart
- Advanced visualization methods

Adopted from FusionCharts White paper, “Principles of Data Visualization - What We See in a Visual”

Byrd Vis Lab
POLYTECHNIC

What is the purpose of Visualization?



PURDUE
POLYTECHNIC

“The purpose of visualization is “*insight*”, not pictures.”

~Ben Shneiderman

Q What does Insight lead to?

PURDUE POLYTECHNIC

INSIGHT LEADS TO

Discovery

- **Visualizing Patterns Over Time**
- **Spotting Differences**

Decision Making

Analysis of Data

Explanation

Storytelling

```
10101010101010101010101010101001010101
010107010101010101010010700101100110
01100110011001100111001100110010101
0101010701010101011100010111000101
111000101001101010101010101010111000
110010101010101010101000101070100101
000101010101010101010101010101010101
010101010101010101010101010101010101
0101010101010107010101101010701010
101010107010100101010101010101010101
0101010010100101100110011001100110
0110011100110011001010101010101010
10101011100010111000107111000101001
101010101010101011100011001010101010
101010100010101010010100010101010101
01010107010101010101010101010170101
010101010101010101010101010101010101
0101010101110011001100110010101001
```

INSIGHT LEADS TO

Discovery

- Visualizing Patterns over time
- Spotting Differences

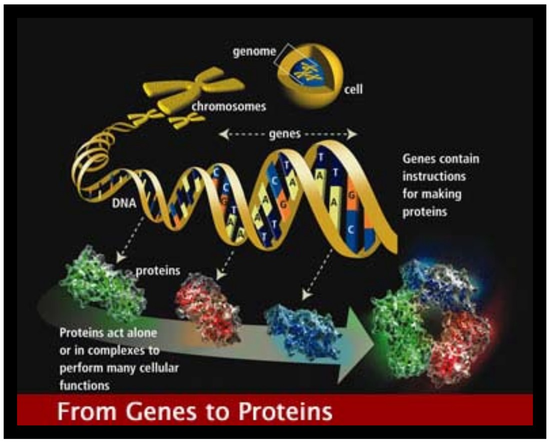
Decision Making

Analysis of Data

Explanation

Storytelling

Allows users to answer questions they didn't know they had


genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
T
A
C
G
proteins
Proteins act alone or in complexes to perform many cellular functions
From Genes to Proteins

Human Genome Project
https://pradipjntu.files.wordpress.com/2011/05/molecularmachine.jpg

INSIGHT LEADS TO

Discovery

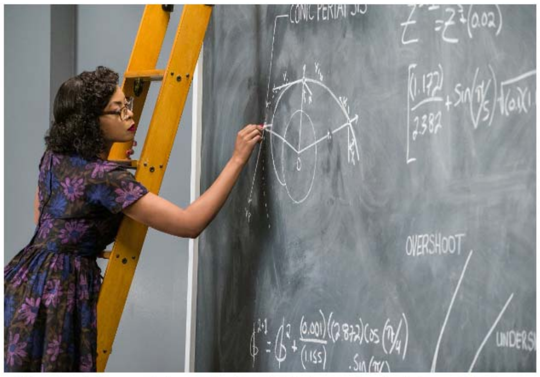
- Visualizing Patterns over time
- Spotting Differences

Decision Making

Analysis of Data

Explanation

Storytelling

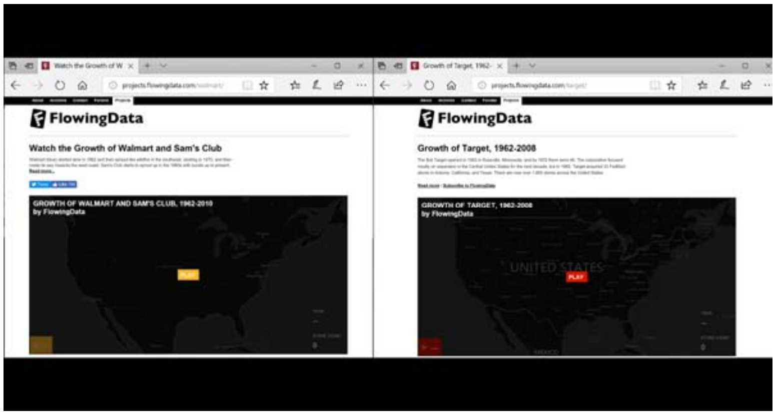


Katherine Johnson (played by Taraji P. Henson) calculates orbital insertion trajectories for the Mercury program using Euler's method in this scene from the movie Hidden Figures. Credit: ™ and © 2017 Twentieth Century Fox Film Corporation. All rights reserved.

PURDUE
POLYTECHNIC

"Insight" Leads to . .

Explanation **Visualizing Spatial Relationships**



Watch the Growth of Walmart and Sam's Club

Read more...

GROWTH OF WALMART AND SAM'S CLUB, 1962-2010
by FlowingData

Growth of Target, 1962-2008

GROWTH OF TARGET, 1962-2008
by FlowingData

Muehlenhaus, I. (2012). **Chapter 8, Visualizing Spatial Relationships,** Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics, pp 271-326.

PURDUE
POLYTECHNIC

"Insight" Leads to . .

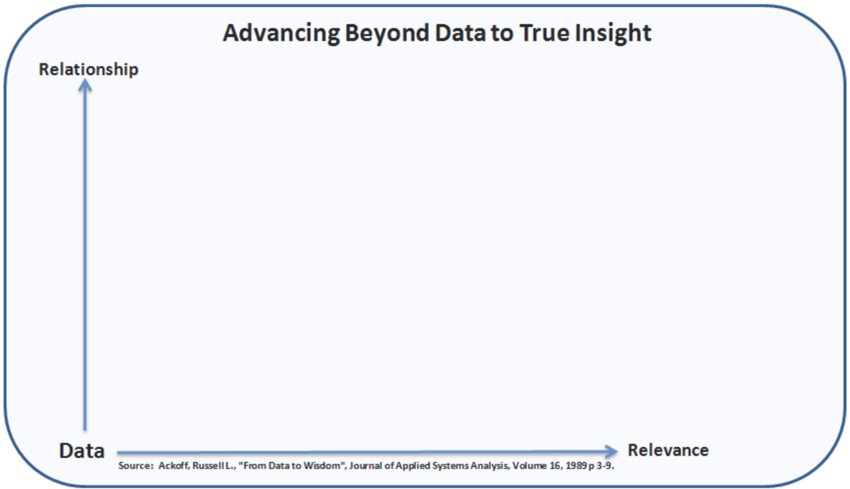
Explanation **Visualizing Spatial Relationships**

http://datafl.ws/197

http://datafl.ws/198

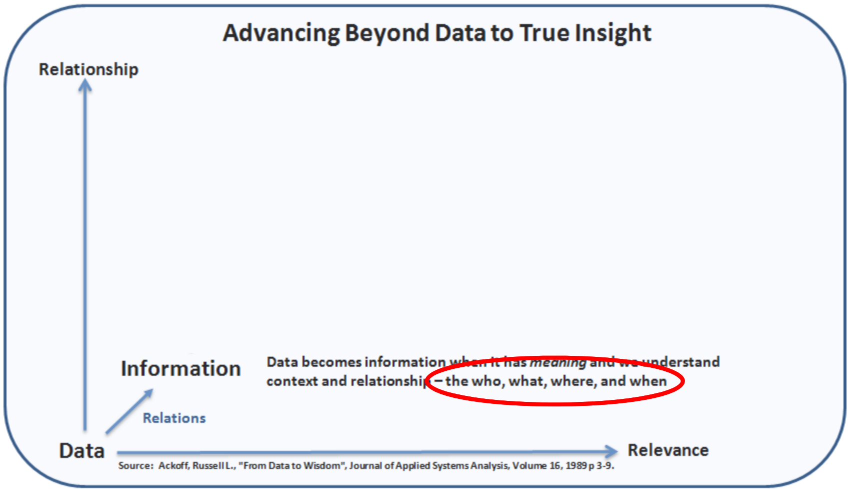
Muehlenhaus, I. (2012). **Chapter 8, Visualizing Spatial Relationships,** Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics, pp 271-326.

FROM DATA TO INSIGHT



Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

FROM DATA TO INSIGHT



Advancing Beyond Data to True Insight

Relationship

Information

Data becomes information when it has *meaning* and we understand context and relationship – the who, what, where, and when

Relations

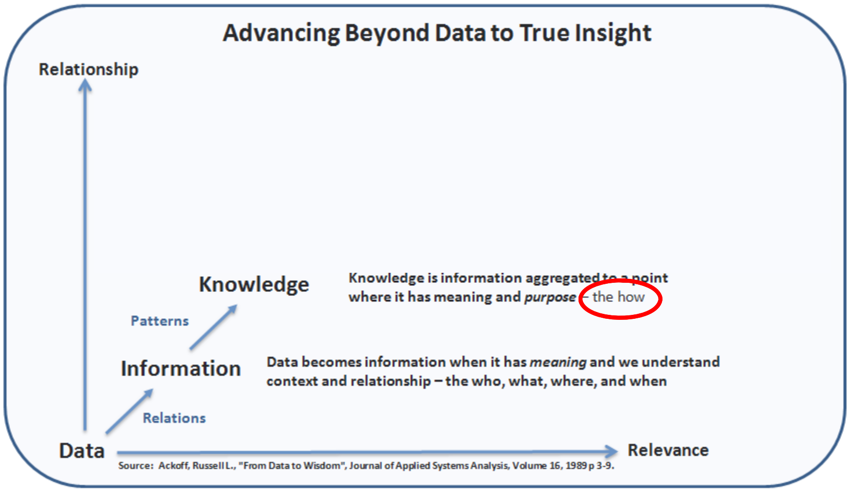
Data

Relevance

Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

Byrd Vis Lab
POLYTECHNIC

FROM DATA TO INSIGHT



Advancing Beyond Data to True Insight

Relationship

Knowledge

Knowledge is information aggregated to a point where it has meaning and *purpose* – the how

Patterns

Information

Data becomes information when it has *meaning* and we understand context and relationship – the who, what, where, and when

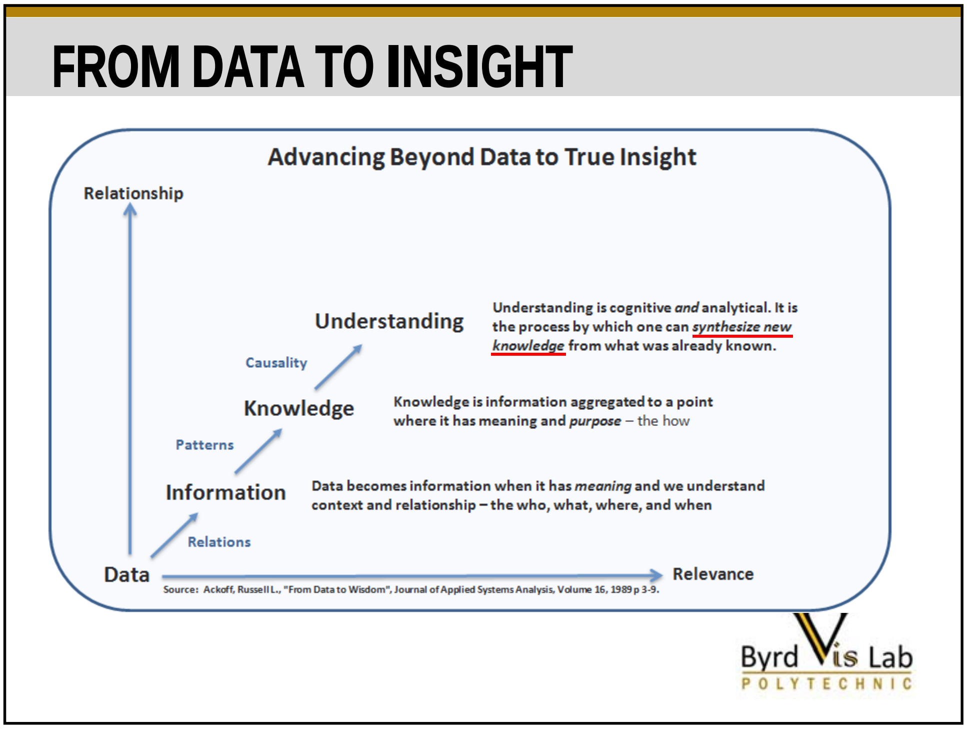
Relations

Data

Relevance

Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

Byrd Vis Lab
POLYTECHNIC



FROM DATA TO INSIGHT

Advancing Beyond Data to True Insight

Relationship

Understanding — Understanding is cognitive *and* analytical. It is the process by which one can *synthesize new knowledge* from what was already known.

Causality

Knowledge — Knowledge is information aggregated to a point where it has meaning and *purpose* – the how

Patterns

Information — Data becomes information when it has *meaning* and we understand context and relationship – the who, what, where, and when

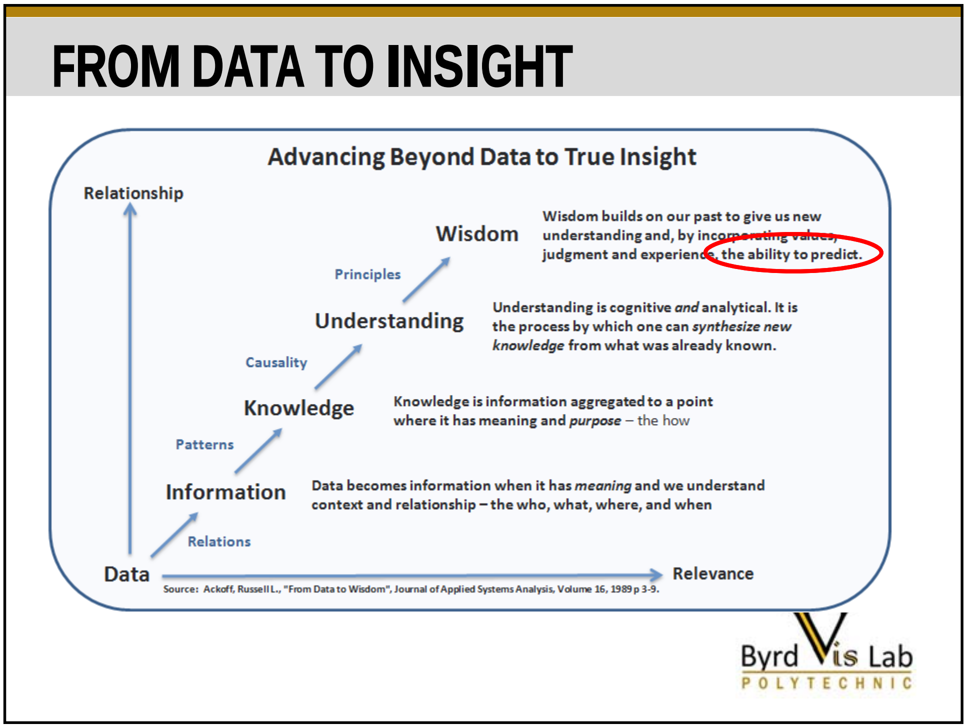
Relations

Data

Relevance

Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

Byrd Vis Lab
POLYTECHNIC



FROM DATA TO INSIGHT

Advancing Beyond Data to True Insight

Relationship

Wisdom — Wisdom builds on our past to give us new understanding and, by incorporating values, judgment and experience, the ability to predict.

Principles

Understanding — Understanding is cognitive *and* analytical. It is the process by which one can *synthesize new knowledge* from what was already known.

Causality

Knowledge — Knowledge is information aggregated to a point where it has meaning and *purpose* – the how

Patterns

Information — Data becomes information when it has *meaning* and we understand context and relationship – the who, what, where, and when

Relations

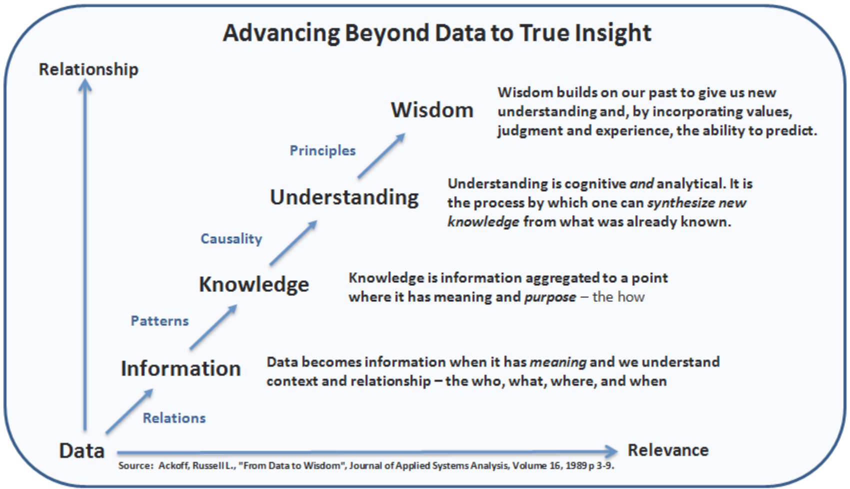
Data

Relevance

Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

Byrd Vis Lab
POLYTECHNIC

FROM DATA TO INSIGHT



Advancing Beyond Data to True Insight

Relationship

Wisdom

Wisdom builds on our past to give us new understanding and, by incorporating values, judgment and experience, the ability to predict.

Principles

Understanding

Understanding is cognitive *and* analytical. It is the process by which one can *synthesize new knowledge* from what was already known.

Causality

Knowledge

Knowledge is information aggregated to a point where it has meaning and *purpose* – the how

Patterns

Information

Data becomes information when it has *meaning* and we understand context and relationship – the who, what, where, and when

Relations

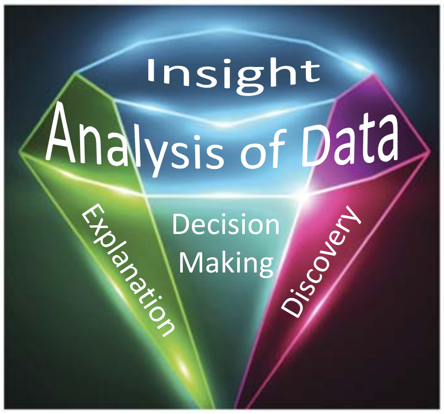
Data

Relevance

Source: Ackoff, Russell L., "From Data to Wisdom", Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9.

Byrd Vis Lab
POLYTECHNIC

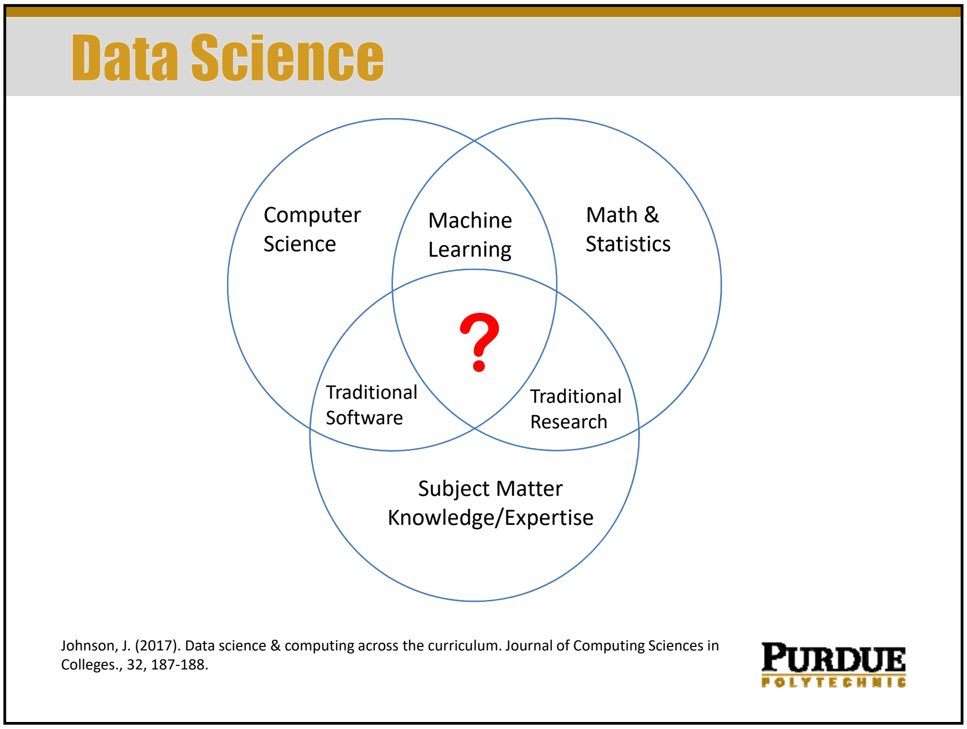
Insight Enables



Storytelling: The Next Step for Visualization

Robert Kosara and Jock Mackinlay, *Tableau Software, Seattle*

Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. Computer, 46(5), 44-50.



Data Science

Computer Science

Machine Learning

Math & Statistics

?

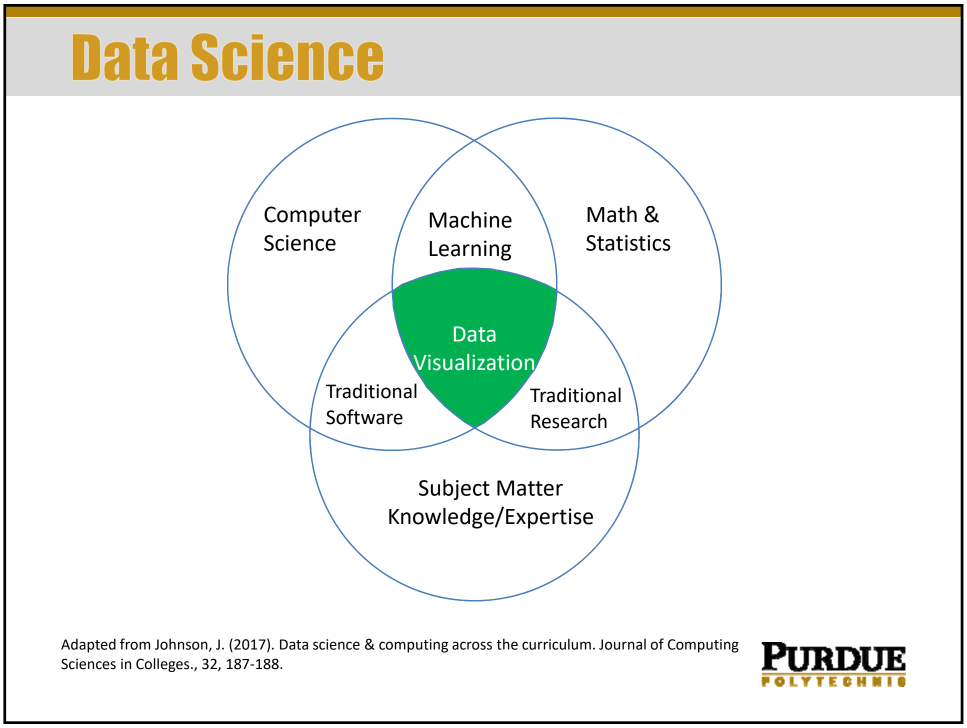
Traditional Software

Traditional Research

Subject Matter Knowledge/Expertise

Johnson, J. (2017). Data science & computing across the curriculum. Journal of Computing Sciences in Colleges., 32, 187-188.

PURDUE POLYTECHNIC



Data Science

Computer Science

Machine Learning

Math & Statistics

Data Visualization

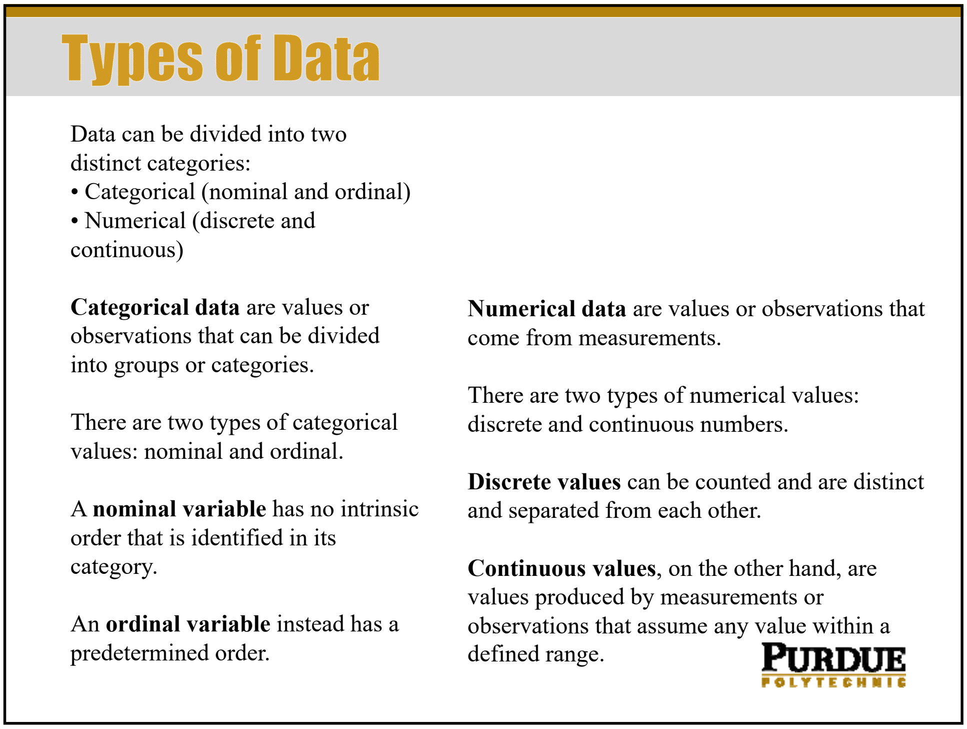
Traditional Software

Traditional Research

Subject Matter Knowledge/Expertise

Adapted from Johnson, J. (2017). Data science & computing across the curriculum. Journal of Computing Sciences in Colleges., 32, 187-188.

PURDUE POLYTECHNIC



Types of Data

Data can be divided into two distinct categories:
- Categorical (nominal and ordinal)
- Numerical (discrete and continuous)

Categorical data are values or observations that can be divided into groups or categories.

There are two types of categorical values: nominal and ordinal.

A **nominal variable** has no intrinsic order that is identified in its category.

An **ordinal variable** instead has a predetermined order.

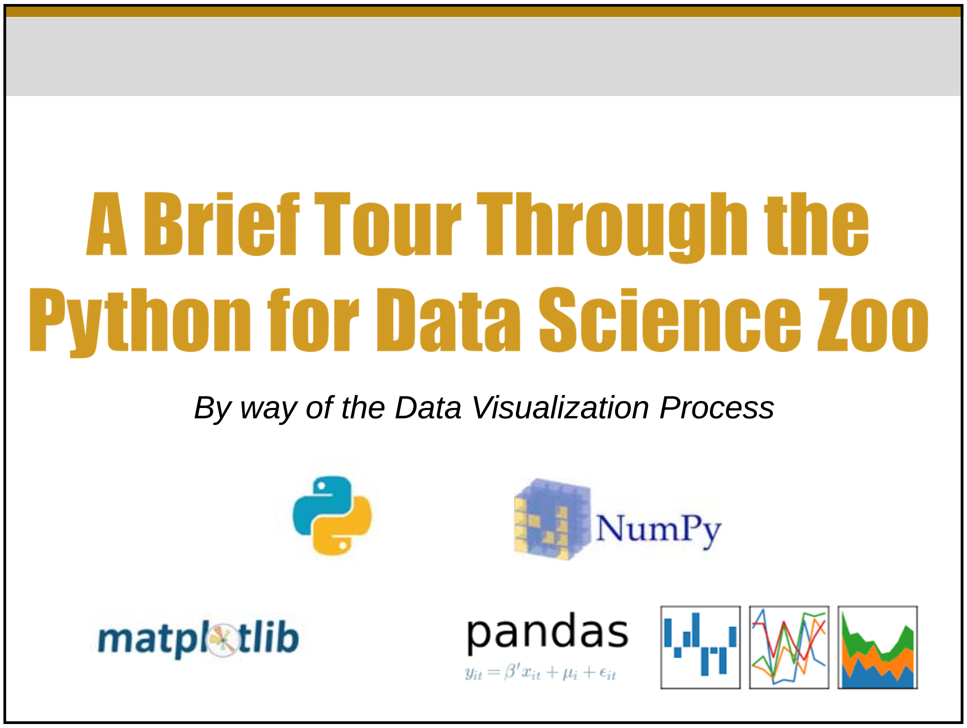
Numerical data are values or observations that come from measurements.

There are two types of numerical values: discrete and continuous numbers.

Discrete values can be counted and are distinct and separated from each other.

Continuous values, on the other hand, are values produced by measurements or observations that assume any value within a defined range.

PURDUE
POLYTECHNIC



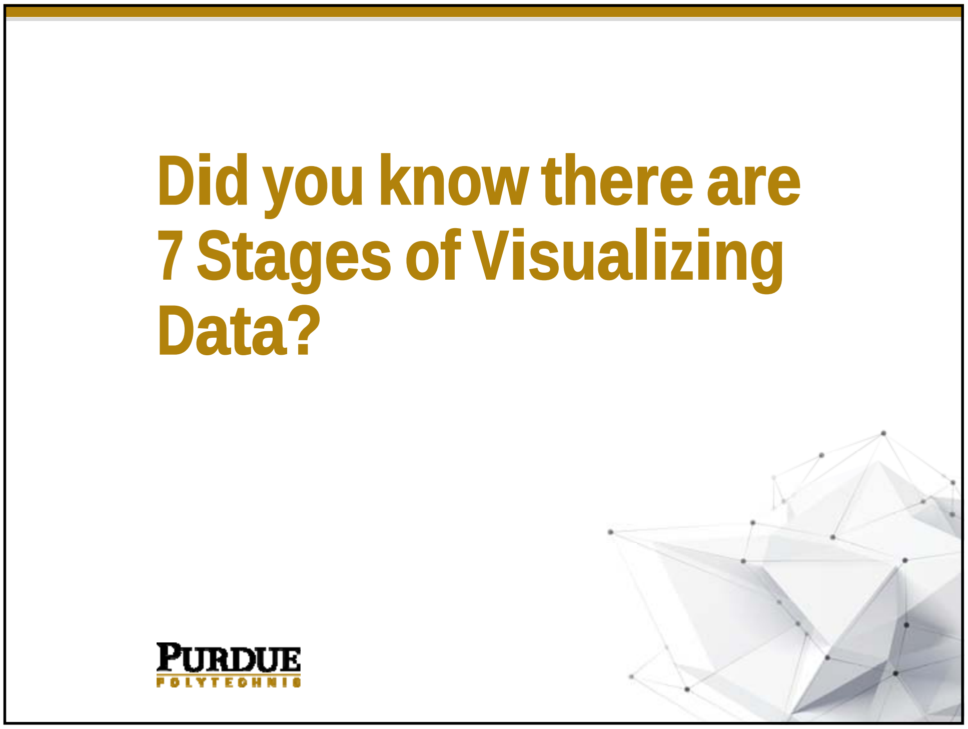
A Brief Tour Through the Python for Data Science Zoo

By way of the Data Visualization Process

NumPy

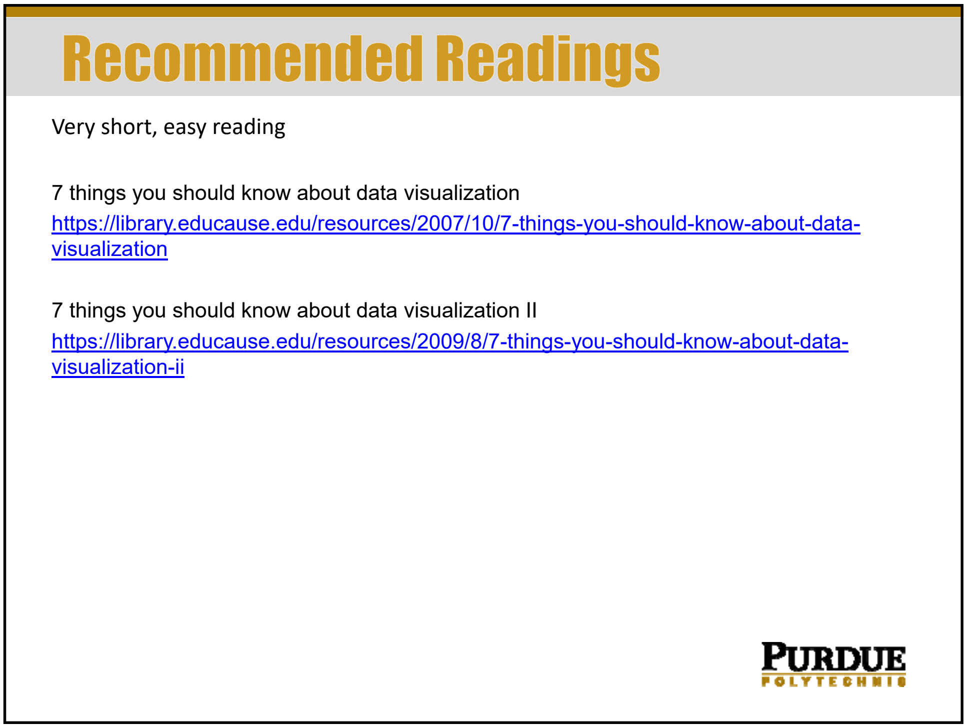
matplotlib

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Did you know there are 7 Stages of Visualizing Data?

PURDUE
POLYTECHNIC



Recommended Readings

Very short, easy reading

7 things you should know about data visualization
https://library.educause.edu/resources/2007/10/7-things-you-should-know-about-data-visualization

7 things you should know about data visualization II
https://library.educause.edu/resources/2009/8/7-things-you-should-know-about-data-visualization-ii

PURDUE
POLYTECHNIC

Stage 1: Acquire

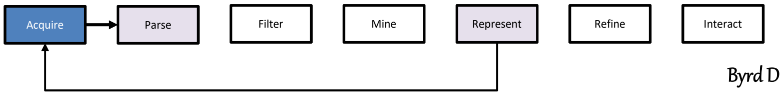
The acquisition step involves obtaining the data.

Like many of the other steps, this can be

- either extremely complicated (i.e., trying to glean useful data from a large system)
- or very simple (reading a readily available text file).

Task: acquire data:

- First name
- Last name
- Major
- Academic status
- Programming Experience
- Visualization Experience



Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Byrd Data Visualization Lab

Stage 2: Parse

- Change the data into a format that tags each part of the data with its intended use.
- Each line of the file must be broken along its individual parts.
- Then, each piece of data needs to be converted to a useful format.

Example data
First name
Last name
Academic status: Fr, So, Jr, Sr
Programming Experience (y/n)
Visualization Experience (y/n)

Parsed Data

First name	Last name	Status	Prog Exp	Vis Exp
String Length: 10	String Length: 12	Char (2) Fr, So, Jr, Sr	Char (1) Y or N	Char (1) Y or N

String

- A set of characters that forms a word or a sentence.

Float

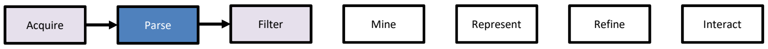
- A number with decimal points (used for the latitudes and longitudes of each location). The name is short for floating point, from programming nomenclature that describes how the numbers are stored in the computer's memory

Character

- A single letter or other symbol.

Integer

- A number without a fractional portion, and hence no decimal points (e.g., –14, 0, or 237).



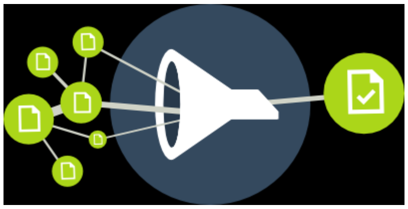
Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

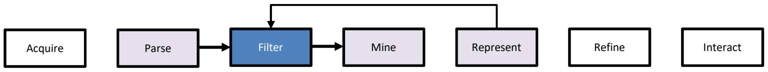
Byrd Data Visualization Lab

Stage 3: Filter

Remove portions not relevant to our use.

Some projects could require significant mathematical work to place the data into a mathematical model or normalize it (convert it to an acceptable range of numbers).





Acquire → Parse → Filter → Mine → Represent → Refine → Interact

Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Byrd Data Visualization Lab

Stage 4: Mine

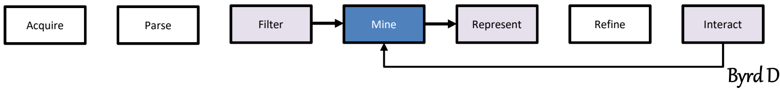
This step involves math, statistics, and data mining.

The data in this case receives only a simple treatment

Most of the time, this step will be far more complicated than a pair of simple math operations.

Tasks:

- Figure out the minimum and maximum values for numeric data
- Figure out the frequency of other values
- What patterns do you see?



Acquire → Parse → Filter → Mine → Represent → Refine → Interact

Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Byrd Data Visualization Lab

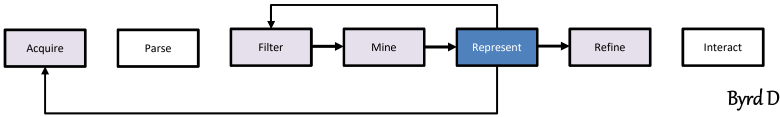
Stage 5: Represent

This step determines the basic form that a set of data will take: List, trees, and so forth.

The Represent stage is a linchpin that informs the single most important decision in a visualization project and can make you rethink earlier stages.

How you choose to represent the data can influence the very first step (what data you acquire) and the third step (what particular pieces you extract).

Task: generate a visualization based on the data received from the Mine stage



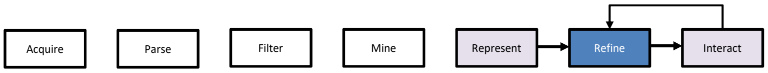
Byrd Data Visualization Lab

Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Stage 6: Refine

Graphic design methods are used to further clarify the representation by calling more attention to particular data (establishing hierarchy) or by changing attributes (such as color) that contribute to readability.

Task: enhance the visualization created in Step 5: Represent



Byrd Data Visualization Lab

Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

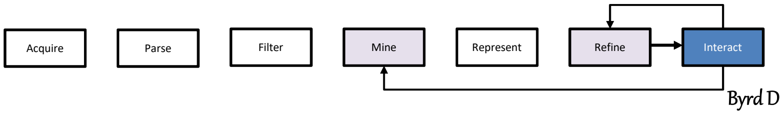
Stage 7: Interact

Letting the user control or explore the data.

Interaction might cover things like selecting a subset of the data or changing the viewpoint.

This stage can also affect the refinement step, as a change in viewpoint might require the data to be designed differently.

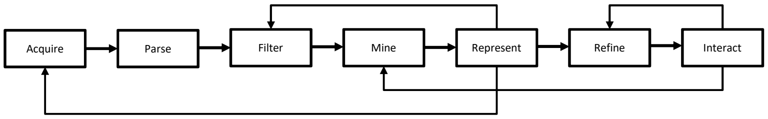
Visually represent the data on the white board.



Byrd Data Visualization Lab

Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

7 stages of Visualizing Data



Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

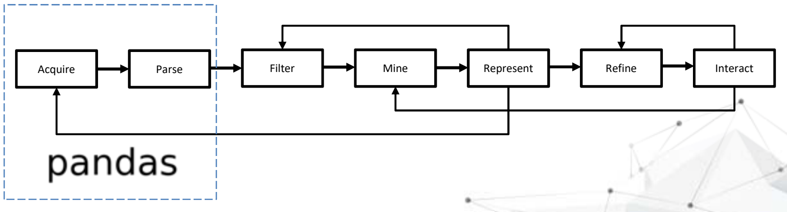
What do we know?

- Output from one stage serves as into the next stage
- Iterative Process
- Your first visualization will **not** be your last visualization

PURDUE POLYTECHNIC

Pandas

Python Data Analysis Library (Some Basics)



Import convention:
Import pandas as pd

PURDUE POLYTECHNIC

Adopted from Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Essential Python Library: Pandas

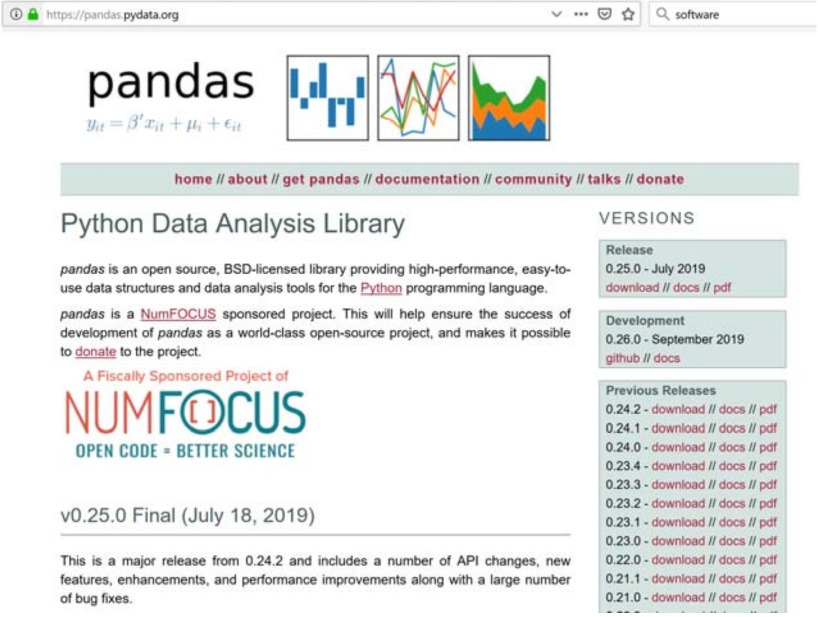
Pandas (http://pandas.pydata.org)

- Provides high-level data structures and functions designed to make working with structured or tabular data fast, easy, and expressive.
- Key objects
  - The *DataFrame*: a tabular, column-oriented data structure with both row and column labels, and
  - The *Series*, a one-dimensional labeled array object.
- Provides sophisticated indexing functionality to make it easy to reshape, slice and dice, perform aggregations and select subsets of data
- Handles:
  - Data structures with labeled axes supporting automatic or explicit data alignment
  - Integrated time series functionality
  - Same data structures handle both time series data and non-time series data

McKinney, Wes. Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython. Second ed. O'Reilly Media, 2017. Web.

PURDUE POLYTECHNIC

Getting Started with Pandas



https://pandas.pydata.org

software

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

home // about // get pandas // documentation // community // talks // donate

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NumFOCUS sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

v0.25.0 Final (July 18, 2019)

This is a major release from 0.24.2 and includes a number of API changes, new features, enhancements, and performance improvements along with a large number of bug fixes.

VERSIONS

Release
0.25.0 - July 2019
download // docs // pdf

Development
0.26.0 - September 2019
github // docs

Previous Releases
0.24.2 - download // docs // pdf
0.24.1 - download // docs // pdf
0.24.0 - download // docs // pdf
0.23.4 - download // docs // pdf
0.23.3 - download // docs // pdf
0.23.2 - download // docs // pdf
0.23.1 - download // docs // pdf
0.23.0 - download // docs // pdf
0.22.0 - download // docs // pdf
0.21.1 - download // docs // pdf
0.21.0 - download // docs // pdf

Pandas: Basics

Contains data structures and data manipulation tools designed to make data cleaning and analysis fast and easy in Python.

Often used in tandem with numerical computing tools like NumPy and data visualization libraries like Matplotlib

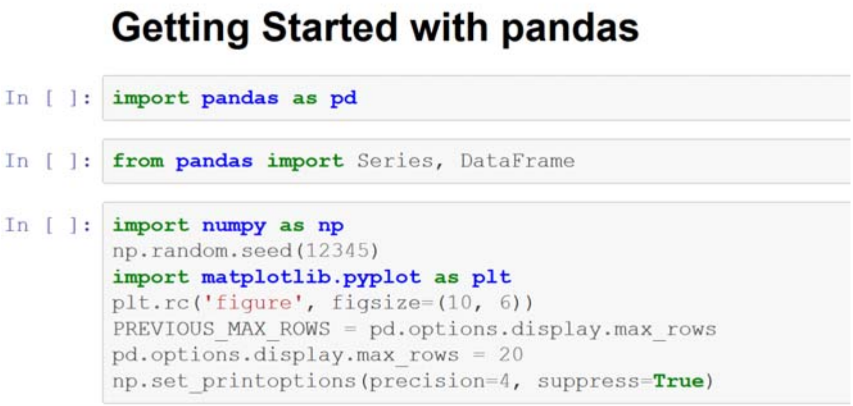
PURDUE
POLYTECHNIC

Pandas: Some Highlights

Used to load data into python from many different file formats

- Time series operations
- Data Frames represents collection off time series
- Can select all data points at a particular time.
- Easy to resample time series data.
- Can specify aggregate data and
- Moving window function

PURDUE POLYTECHNIC



Getting Started with pandas

```
In [ ]: import pandas as pd

In [ ]: from pandas import Series, DataFrame

In [ ]: import numpy as np
        np.random.seed(12345)
        import matplotlib.pyplot as plt
        plt.rc('figure', figsize=(10, 6))
        PREVIOUS_MAX_ROWS = pd.options.display.max_rows
        pd.options.display.max_rows = 20
        np.set_printoptions(precision=4, suppress=True)
```

To run a command in windows interface: press Shift + Enter Key

PURDUE POLYTECHNIC

Pandas Data Frames

Documentation

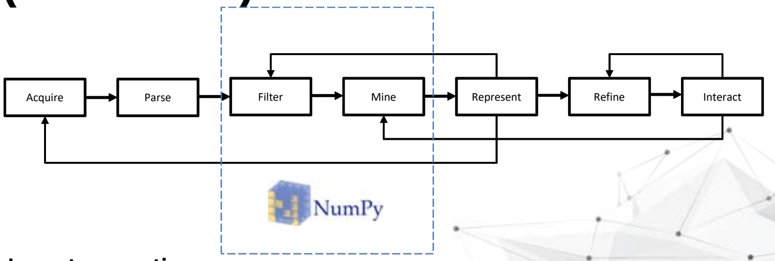
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

Graphical Explanation

https://www.geeksforgeeks.org/python-pandas-dataframe/

NumPy

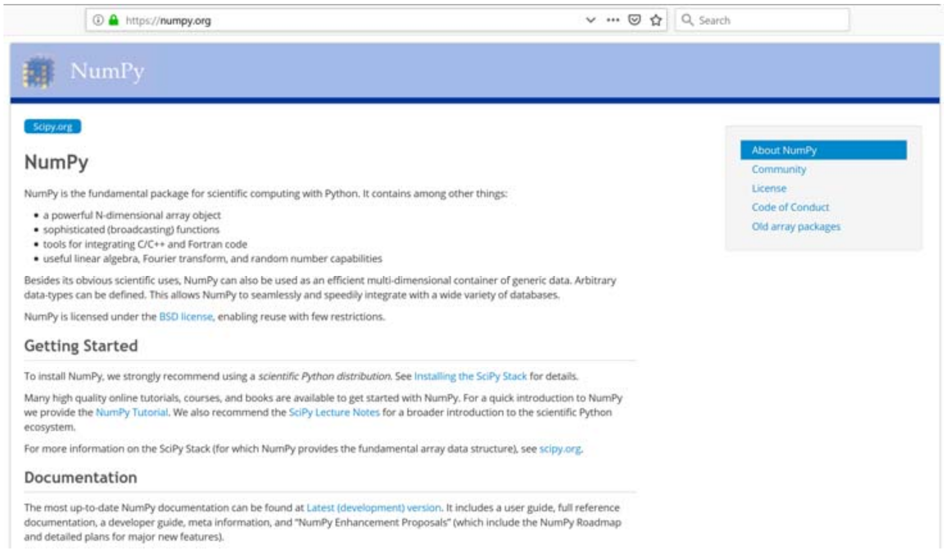
Python Data Analysis Library (Some Basics)



Import convention:
Import numpy as np

Adopted from Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

Getting Started with NumPy



NumPy

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the BSD license, enabling reuse with few restrictions.

Getting Started

To install NumPy, we strongly recommend using a *scientific Python distribution*. See Installing the SciPy Stack for details.

Many high quality online tutorials, courses, and books are available to get started with NumPy. For a quick introduction to NumPy we provide the NumPy Tutorial. We also recommend the SciPy Lecture Notes for a broader introduction to the scientific Python ecosystem.

For more information on the SciPy Stack (for which NumPy provides the fundamental array data structure), see scipy.org.

Documentation

The most up-to-date NumPy documentation can be found at Latest (development) version. It includes a user guide, full reference documentation, a developer guide, meta information, and "NumPy Enhancement Proposals" (which include the NumPy Roadmap and detailed plans for major new features).

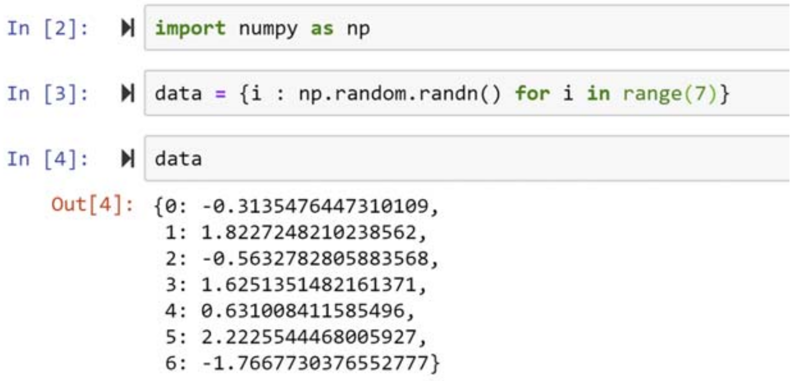
Essential Python Library: NumPy

NumPy (http://numpy.org)

- Aka: Numerical Python
- Provides the data structures, algorithms, and library glue needed for most scientific applications involving numerical data in Python.
- Contains
  - A fast and efficient multidimentional array object ndarray
  - Functions for performing element-wise computations with array or mathematical operations between arrays
  - Tools for reading and writing array-based datasets to disk
  - Linear algebra operations, Fourier transform, and random number generation
  - A mature C API to enable Python extensions and ntive C or C++ code to access NumPy data structures and computational facilities.
- Primary uses in data analysis is as a container for data to be passed between algorithms and libraries.

McKinney, Wes. Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython. Second ed. O'Reilly Media, 2017. Web.

Getting Started with NumPy



```
In [2]:  import numpy as np

In [3]:  data = {i : np.random.randn() for i in range(7)}

In [4]:  data

Out[4]: {0: -0.3135476447310109,
          1: 1.8227248210238562,
          2: -0.5632782805883568,
          3: 1.6251351482161371,
          4: 0.631008411585496,
          5: 2.2225544468005927,
          6: -1.7667730376552777}
```

The NumPy ndarray

:A Multidimensional Array object

- A fast, flexible container for large datasets in Python
- Arrays enable you to perform mathematical operations on whole blocks of data using similar syntax to the equivalent operations between scalar elements.

- Creating arrays in NumPy

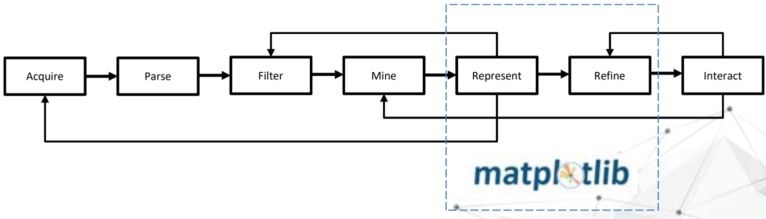
https://towardsdatascience.com/getting-started-with-numpy-59b22df56729

NumPy Resources

http://cs231n.github.io/python-numpy-tutorial/
http://cs231n.github.io/python-numpy-tutorial/#numpy

Matplotlib

Python Visualization Library (Some Basics)



Import convention:
Import matplotlib.pyplot as plt

Figure adopted from Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

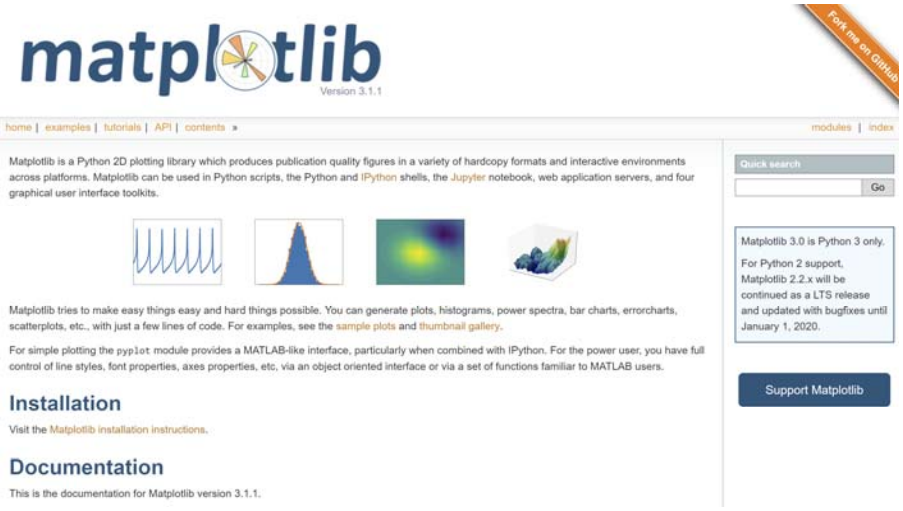
Four types of Visualizations

GEORGES GRINSTEIN (KEYNOTE PRESENTATION, VINCI 2016)

- **Exploratory**
  - Have no hypotheses about the data
  - Explore data interactively as undirected searches
- **Confirmatory**
  - Have specific hypotheses about the data
  - Goal-oriented examination of the hypotheses
- **Presentation**
  - Facts to be presented are fixed a priori
  - Select appropriate presentation techniques
- **Interactive**
  - Interactions with a pre-defined animation

PURDUE POLYTECHNIC

Getting Started with MatplotLib



Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the `pyplot` module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Installation

Visit the Matplotlib installation instructions.

Documentation

This is the documentation for Matplotlib version 3.1.1.

Matplotlib 3.0 is Python 3 only.

For Python 2 support, Matplotlib 2.2.x will be continued as a LTS release and updated with bugfixes until January 1, 2020.

PURDUE POLYTECHNIC

Essential Python Library: Matplotlib

Matplotlib (http://matplotlib.org)

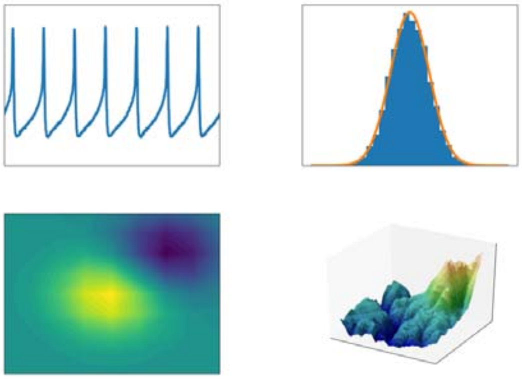
- Most popular Python Library for producing plots and other two-dimensional data visualizations.
- Was designed for creating plots suitable for publication.
- The most widely used visualization library available to Python programmers.

McKinney, Wes. Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython. Second ed. O'Reilly Media, 2017. Web.

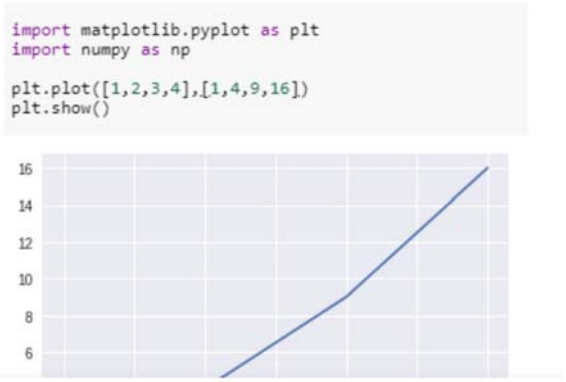
Sample plots in Matplotlib

https://matplotlib.org/tutorials/introductory/sample_plots.html#

http://cs231n.github.io/python-numpy-tutorial/#matplotlib



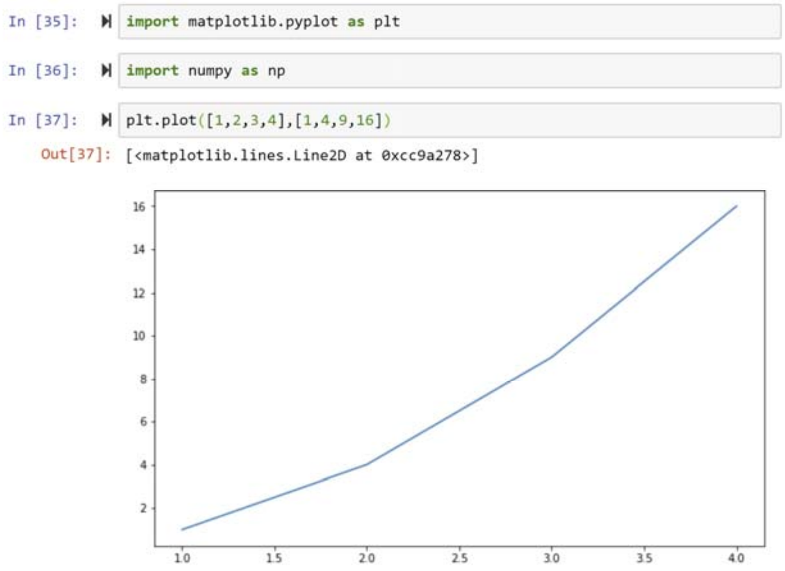
Simple Example



```
import matplotlib.pyplot as plt
import numpy as np

plt.plot([1,2,3,4],[1,4,9,16])
plt.show()
```

From Jupyter Notebook



```
In [35]:  import matplotlib.pyplot as plt

In [36]:  import numpy as np

In [37]:  plt.plot([1,2,3,4],[1,4,9,16])

Out[37]:  [<matplotlib.lines.Line2D at 0xcc9a278>]
```

Matplotlib Resources

https://towardsdatascience.com/matplotlib-tutorial-learn-basics-of-pythons-powerful-plotting-library-b5d1b8f67596

https://realpython.com/python-matplotlib-guide/

PURDUE POLYTECHNIC

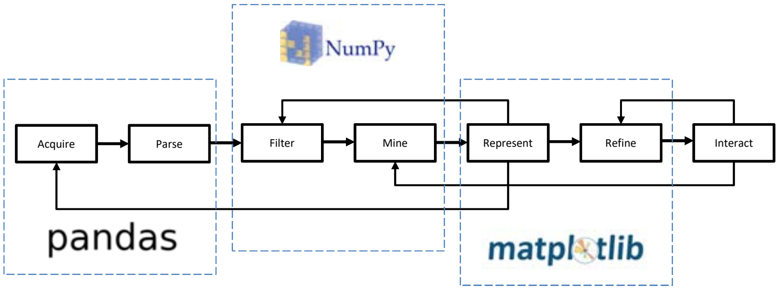
When should you think about visualizing your data?

As early
and
often!

http://howtolaunch.com/howtolaunch/reach-your-audience-early-and-often/

PURDUE POLYTECHNIC

Recap: Data Visualization Tools for Insight



Adopted from Fry, B. (2008). Chapter 1, Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

PURDUE POLYTECHNIC

Additional Resources

7 things you should know about data visualization
https://library.educause.edu/resources/2007/10/7-things-you-should-know-about-data-visualization

7 things you should know about data visualization II
https://library.educause.edu/resources/2009/8/7-things-you-should-know-about-data-visualization-ii

Quispel, and Maes. "Would You Prefer Pie or Cupcakes? Preferences for Data Visualization Designs of Professionals and Laypeople in Graphic Design." Journal of Visual Languages and Computing 25.2 (2014): 107-16.

PURDUE POLYTECHNIC

References Cited

Ackoff, R. (1989). From Data to Wisdom, Journal of Applied Systems Analysis, 16, 3-9.

Fry, B. (2008). Visualizing data (Safari Books Online). Sebastopol, Calif.: O'Reilly Media.

FusionCharts White paper, “Principles of Data Visualization - What We See in a Visual.

Johnson, J. (2017). Data science & computing across the curriculum. Journal of Computing Sciences in Colleges., 32, 187-188.

Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. Computer, 46(5), 44-50.

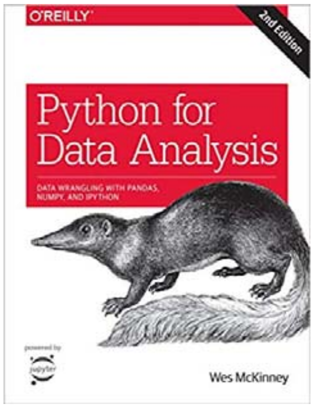
Muehlenhaus, I. (2012). Chapter 8, Visualizing Spatial Relationships, Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics, pp 271-326.

PURDUE
POLYTECHNIC

Main Source for Python Libraries

McKinney, W. (2017). Python for data analysis : Data wrangling with Pandas, NumPy, and IPython (Second ed.). O'Reilly Media.

GitHub: https://github.com/wesm/pydata-book
Sample data and code from book available



2nd Edition

1st Edition

PURDUE
POLYTECHNIC

Web pages referenced

Pandas Links
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
https://www.geeksforgeeks.org/python-pandas-dataframe/
http://pandas.pydata.org)

NumPy Links
http://numpy.org
https://towardsdatascience.com/getting-started-with-numpy-59b22df56729
http://cs231n.github.io/python-numpy-tutorial/
http://cs231n.github.io/python-numpy-tutorial/#numpy

Matplotlib Links
http://matplotlib.org
https://towardsdatascience.com/matplotlib-tutorial-learn-basics-of-pythons-powerful-plotting-library-b5d1b8f67596
https://realpython.com/python-matplotlib-guide/
https://matplotlib.org/tutorials/introductory/sample_plots.html#
http://cs231n.github.io/python-numpy-tutorial/#matplotlib

Other
Degrees in Data Visualization: https://polytechnic.purdue.edu/degrees/data-visualization
Human Genome Project: https://pradipjntu.files.wordpress.com/2011/05/molecularmachine.jpg
http://howtolaunch.com/howtolaunch/reach-your-audience-early-and-often/

PURDUE POLYTECHNIC

Vetria L. Byrd
Assistant Professor
Computer Graphics Technology
vlbyrd@purdue.edu

https://polytechnic.purdue.edu/profile/vbyrd
http://web.ics.purdue.edu/~vbyrd/
@VByrdPhD, @BPViz, @ByrdVisLab
Purdue Polytechnic Institute

PURDUE POLYTECHNIC

Thank You Image Source:
http://careerconfidential.com/category/thank-you-notes/
http://careerconfidential.com/wp-content/uploads/2015/02/ThankYou2.jpg