Community Experience Distilled

Mastering Python Data Visualization

Generate effective results in a variety of visually appealing charts using the plotting packages in Python

Kirthi Raman

[PACKT] PUBLISHING open source* community experience distilled

Mastering Python Data Visualization

Generate effective results in a variety of visually appealing charts using the plotting packages in Python

Kirthi Raman

[PACKT] open source*
PUBLISHING
community experience distilled

BIRMINGHAM - MUMBAI

Mastering Python Data Visualization

First published: October 2015

Production reference: 1211015

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-832-7

www.packtpub.com

Credits

Author
Kirthi Raman

Reviewers
Julian Quick
Hang (Harvey) Yu

Acquisition Editor
Subho Gupta

Content Development Editor
Riddhi Tuljapurkar

Technical Editor
Humera Shaikh

Copy Editors
Relin Hedly
Sonia Mathur

Project Coordinator
Kinjal Bari

Proofreader
Safis Editing

Indexer
Monica Ajmera Mehta

Graphics
Abhinash Sahu
Jason Monteiro

Production Coordinator
Nilesh Mohite

Cover Work
Nilesh Mohite

About the Author

Kirthi Raman is currently working as a lead data engineer with Neustar Inc, based in Mclean, Virginia USA. Kirthi has worked on data visualization, with a focus on JavaScript, Python, R, and Java, and is a distinguished engineer. Previously, he worked as a principle architect, data analyst, and information retrieval specialist at Quotient, Inc. Kirthi has also worked as a technical lead and manager for a start-up.

He has taught discrete mathematics and computer science for several years. Kirthi has a graduate degree in mathematics and computer science from IIT Delhi and an MS in computer science from the University of Maryland. He has written several white papers on data analysis and big data.

I would like to thank my wife, Radhika, my son, Sid, and daughter, Niya, for putting up with my schedule even when I was on vacation. I would also like to thank my dad, Venkatraman, and my sisters, Vijaya and Meena, for their blessings.

About the Reviewers

Julian Quick is pursuing his bachelor's of science degree in environmental resources engineering at Humboldt State University with a specialization in energy resources and energy data analysis. He wrote Python code for the Earth Observing Laboratory, Canary Instruments, home energy monitoring, and the National Wind Technology Center.

I place on record my gratitude towards my family.

Hang (Harvey) Yu graduated from the University of Illinois at Urbana-Champaign with a PhD in computational biophysics and a master's in statistics. He has extensive experience in data mining, machine learning, and statistics. In the past, Harvey has worked on areas such as stochastic simulations and time series in C and Python as part of his academics. He was intrigued by algorithms and mathematical modeling and has been involved in data analytics since then.

Hang (Harvey) Yu is currently working as a data scientist in Silicon Valley. He is passionate about data science and has developed statistical/mathematical models based on techniques such as optimization and predictive modeling in R. Previously, Harvey has also worked as a computational science intern at ExxonMobil.

When Harvey is not coding, he plays soccer, reads fiction books, or listens to classical music. You can reach him at `hangyu1@illinois.edu` or on LinkedIn at `www.linkedin.com/in/hangyu1`.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

https://www2.packtpub.com/books/subscription/packtlib

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

Data visualization is intended to provide information clearly and help the viewer understand them qualitatively. The well-known expression that a picture is worth a thousand words may be rephrased as "a picture tells a story as well as a large collection of words". Visualization is, therefore, a very precious tool that helps the viewer understand a concept quickly. However, data visualization is more of an art than a skill because if you try to overdo it, it could have a reverse effect.

We are currently faced with a plethora of data containing many insights that hold the key to success in the modern day. It is important to find the data, clean it, and use the right tool to visualize it. This book explains several different ways to visualize data using Python packages, along with very useful examples in many different areas such as numerical computing, financial models, statistical and machine learning, and genetics and networks.

This book presents an example code developed on Mac OS X 10.10.5 using Python 2.7, IPython 0.13.2, matplotlib 1.4.3, NumPy 1.9.2, SciPy 0.16.0, and conda build version 1.14.1.

What this book covers

Chapter 1, A *Conceptual Framework for Data Visualization*, expounds that data visualization should actually be referred to as "the visualization of information for knowledge inference". This chapter covers the framework, explaining the transition from data/information to knowledge and how meaningful representations (through logarithms, colormaps, scatterplots, correlations, and others) can make knowledge much easier to grasp.

Chapter 2, *Data Analysis and Visualization*, explains the importance of visualization and shows several steps in the visualization process, including several options of tools to choose from. Visualization methods have existed for a long time, and we are exposed to them very early; for instance, even young children can interpret bar charts. Interactive visualization has many strengths, and this chapter explains them with examples.

Chapter 3, *Getting Started with the Python IDE*, explains how you can use Anaconda from Continuum Analytics without worrying about installing each Python library individually. Anaconda has simplified packaging and deployment methods that make it easier to run the IPython notebook alongside other libraries.

Chapter 4, *Numerical Computing and Interactive Plotting*, covers interactive plotting methods with working examples in computational physics and applied mathematics. Some notable examples are interpolation methods, approximation, clustering, sampling, correlation, and convex optimization using SciPy.

Chapter 5, *Financial and Statistical Models*, explores financial engineering, which has many numerical and graphical methods that make an interesting use case to explore Python. This chapter covers stock quotes, regression analysis, the Monte Carlo algorithm, and simulation methods with examples.

Chapter 6, *Statistical and Machine Learning*, covers statistical methods such as linear and nonlinear regression and clustering and classification methods using numpy, scipy, matplotlib, and scikit-learn.

Chapter 7, *Bioinformatics, Genetics, and Network Models*, covers interesting examples such as social network and instances of directed graphs in real life, data structures that are appropriate for these problems, and network analysis. This chapter uses specific libraries such as graph-tool, NetworkX, matplotlib, scipy, and numpy.

Chapter 8, *Advanced Visualization*, covers simulation methods and examples of signal processing to show several visualization methods. Here, we also have a comparison of other advanced tools out there, such as Julia and D3.js.

Appendix, *Go Forth and Explore Visualization*, gives an overview of conda and lists out various Python libraries.

What you need for this book

For this book, you need Python 2.7.6 or a later version installed on your operating system. For the examples in this book, Mac OS X 10.10.5's Python default version (2.7.6) has been used. Other software packages used in this book are IPython, which is an interactive Python environment. The new version of IPython is called Jupyter, which now has kernels for 50 different languages.

Install the prepackaged scientific Python distributions, such as Anaconda from Continuum or Enthought Python Distribution if possible. Anaconda typically comes with over 300 Python packages. For the Python packages that are not included in the prepackaged list, you may either use pip or conda to install them. Some examples are provided in *Appendix*, *Go Forth and Explore Visualization*.

Who this book is for

There are many books on Python and data visualization. However, there are very few that can be recommended to somebody who wants to build on the existing knowledge about Python, and there are even fewer that discuss niche techniques to make your code easier to work with and reusable. If you know a few things about Python programming but have an insatiable drive to learn more, this book will show you ways to obtain analytical results and produce amazing visual displays.

This book covers methods to produce analytical results using real-world problems. It is not written for beginners, but if you need clarification, you can follow the suggested reading hints in the book. If this book is your first exposure to Python or data visualization, you will do well to study some introductory texts. My favorite is *Introduction to Computer Science and Programming* by Professor John Guttag, which is freely available at MIT OpenCourseWare, and *Visualize This* by Nathan Yau from UCLA.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "First we use `norm()` from SciPy to create normal distribution samples and later, use `hstack()` from NumPy to stack them horizontally and apply `gaussian_kde()` from SciPy."

A block of code is set as follows:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
students = pd.read_csv("/Users/Macbook/python/data/ucdavis.csv")
g = sns.FacetGrid(students, palette="Set1", size=7)
g.map(plt.scatter, "momheight", "height", s=140, linewidth=.7,
edgecolor="#ffad40", color="#ff8000")
g.set_axis_labels("Mothers Height", "Students Height")
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
import blockspring
import json

print blockspring.runParsed("stock-price-comparison",
   { "tickers": "FB, LNKD, TWTR",
   "start_date": "2014-01-01", "end_date": "2015-01-01" }).params
```

Any command-line input or output is written as follows:

```
conda install jsonschema

Fetching package metadata: ....
Solving package specifications: .
Package plan for installation in environment /Users/MacBook/anaconda:

The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    jsonschema-2.4.0           |          py27_0          51 KB

The following NEW packages will be INSTALLED:

    jsonschema: 2.4.0-py27_0

Proceed ([y]/n)?
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "Further, you can select the **Copy code** option to copy the contents of the code block into Canopy's copy-and-paste buffer to be used in an editor."

[Warnings or important notes appear in a box like this.]

[Tips and tricks appear like this.]

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

Downloading the color images of this book

We also provide you with a PDF file that has color images of the screenshots/diagrams used in this book. The color images will help you better understand the changes in the output. You can download this file from: `https://www.packtpub.com/sites/default/files/downloads/8327OS_Graphics.pdf`.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.

1
A Conceptual Framework for Data Visualization

The existence of the Internet and social media in modern times has led to an abundance of data, and data sizes are growing beyond imagination. How and when did this begin?

A decade ago, a new way of doing business evolved: of corporations collecting, combining, and crunching large amount of data from sources throughout the enterprise. Their goal was to use a high volume of data to improve the decision-making process. Around that same time, corporations like Amazon, Yahoo, and Google, which handled large amounts of data, made significant headway. Those milestones led to the creation of several technologies supporting *big data*. We will not get into details about big data, but will try exploring why many organizations have changed their ways to use similar ideas for better decision-making.

How exactly are these large amount of data used for making better decisions? We will get to that eventually, but first let us try to understand the difference between data, information, and knowledge, and how they are all related to data visualization. One may wonder, why are we talking about data, information, and knowledge. There is a storyline that connects how we start, what we start with, how all these things benefit the business, and the role of visualization. We will determine the required conceptual framework for data visualization by briefly reviewing the steps involved.

In this chapter, we will cover the following topics:

- The difference between data, information, knowledge, and insight
- The transformation of information into knowledge, and further, to insight
- Collecting, processing, and organizing data
- The history of data visualization
- How does visualizing data help decision-making?
- Visualization plots

Data, information, knowledge, and insight

The terms **data**, **information**, and **knowledge** are used extensively in the context of computer science. There are many definitions of these terms, often conflicting and inconsistent. Before we dive into these definitions, we will understand how these terms are related to visualization. The primary objective of data visualization is to gain insight (hidden truth) into the data or information. The whole discussion about data, knowledge, and insight in this book is within the context of computer science, and not psychology or cognitive science. For the cognitive context, one may refer to `https://www.ucsf.edu/news/2014/05/114321/converting-data-knowledge-insight-and-action`.

Data

The term **data** implies a premise from which one may draw conclusions. Though data and information appear to be interrelated in a certain context, data actually refers to discrete, objective facts in a digital form. Data are the basic building blocks that, when organized and arranged in different ways, lead to information that is useful in answering some questions about the business.

Data can be something very simple, yet voluminous and unorganized. This discrete data cannot be used to make decisions on its own because it has no meaning and, more importantly, because there is no structure or relationship between them. The process by which data is collected, transmitted, and stored varies widely with the types of data and storage methods. Data comes in many forms; some notable forms are listed as follows:

- CSV files
- Database tables
- Document formats (Excel, PDF, Word, and so on)
- HTML files
- JSON files
- Text files
- XML files

Information

Information is processed data presented as an answer to a business question. Data becomes information when we add a relationship or an association. The association is accomplished by providing a context or background to the data. The background is helpful because it allows us to answer questions about the data.

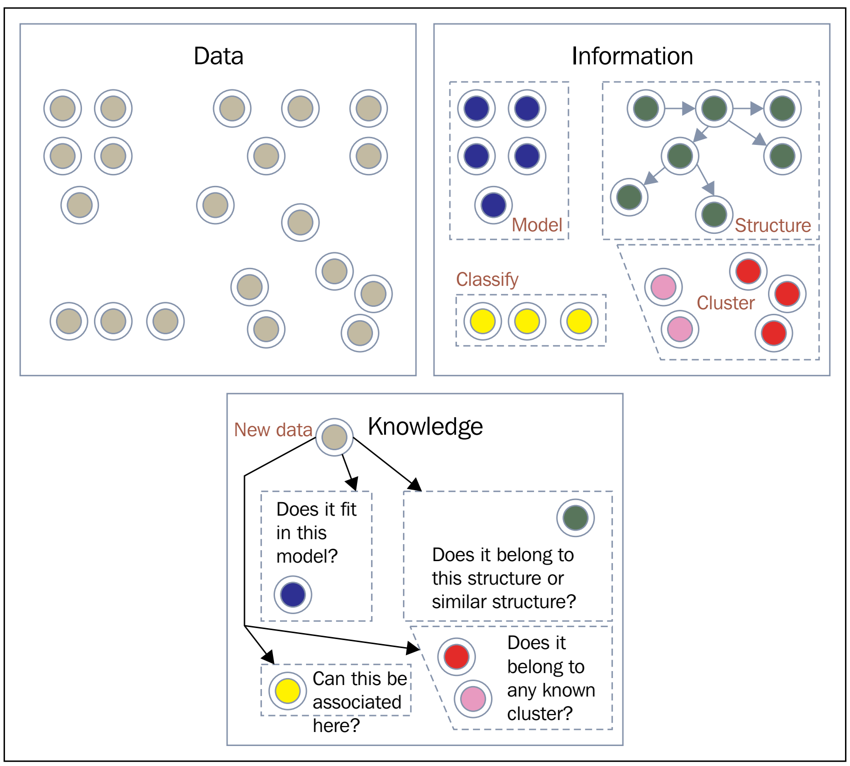
For example, let us assume that the data given for a basketball player includes height, weight, position, college, date of birth, draft pick, draft round, NBA-debut, and recruiting rank. The answer to the question, "Who is the first draft pick with a height of more than six feet and plays on the point guard position?" is also the information.

Similarly, each player's score is one piece of data. The answer to the question "Who has the highest point per game this year and what is his score" is "LeBron James, 27.47", which is also information.

Knowledge

Knowledge emerges when humans interpret and organize information and use that to drive decision-making. Knowledge is the data, information, and the skills acquired through experience. Knowledge comprises the ability to make the appropriate decision as well as the skills to execute it.

The essential ingredient—connecting the data—allows us to understand the relative importance of each piece of information. By comparing results from the past and by recognizing patterns, we don't have to build a solution to a problem from scratch. The following diagram summarizes the concepts of data, information, and knowledge:



Knowledge changes in an incremental way, particularly when information is rearranged or reorganized or when some computing algorithm changes. Knowledge is like an arrow pointing to the results of an algorithm that is dependent on past information that comes from data. In many instances, knowledge is also gained by visually interacting with the results. Insight on the other hand, opens the way to the future.

Data analysis and insight

Before we dive into the definition of insight and how it relates to business, let us see how the idea of capturing insight ever began. For over a decade, organizations have been struggling to make sense of all the data and information they have, particularly with the exploding data size. They all realized the importance of **data analysis** (also known as **data analytics** or **analytics**) in order to arrive at an optimal or realistic business decision based on existing data and information.

Analytics hinges upon mathematical algorithms to determine the relationships between the data that can yield insight. One simple way to understand insight is by considering an analogy: when data does not have a structure and proper alignment with the business, it gives a clearer and deeper understanding by converting the data to a more structured form and aligning it more closely to the business goals. Insight is that "eureka" moment when there is a breakthrough result that comes out. One should not get confused between the terms Analytics and Business Intelligence. Analytics has predictive capabilities while Business Intelligence provides results based on the analysis of historical data.

Analytics is usually applicable to a broader spectrum of data and, for this reason, it is very common that data collaboration happens internally and/or externally. In some business paradigms, the collaboration only happens internally in an extensive collection of a dataset, but in most other cases, an external connection helps in connecting the dots or completing the puzzle. Two of the most common sources of external data connection are social media and consumer base.

Later in this chapter, we refer to real-life business stories that achieved some remarkable results by applying analytics to gain insight and drive business value, improve decision-making, and understand their customers better.

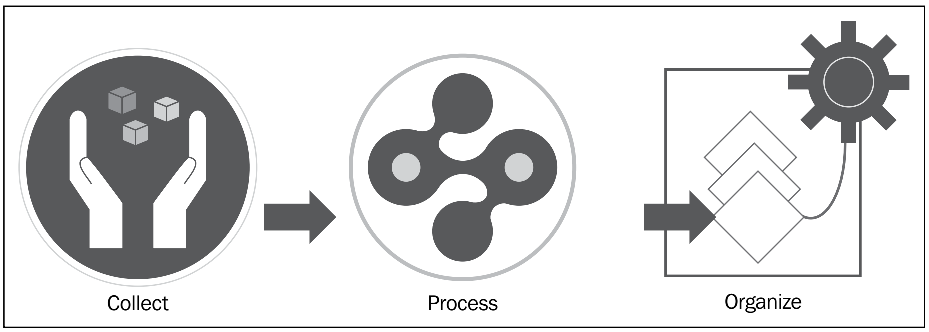
The transformation of data

By now we know what data is, but now the question is: what is the purpose of collecting data? Data is useful for describing a physical or social phenomenon and to further answer questions about that phenomenon. For this reason, it is important to ensure that the data is not faulty, inaccurate, or incomplete; otherwise, the responses based on that data will also not be accurate or complete.

There are different categories of data, some of which are *past performance data*, *experimental data*, and *benchmark data*. Past performance data and experimental data are pretty self-explanatory. Benchmark data, on the other hand, is data that compares the characteristics of two different items or products to a standard measure. Data gets transformed into information, is processed further, and is then used for answering questions. It is apparent, therefore, that our next step is to achieve that transformation.

Transforming data into information

Data is collected and stored in several different forms depending on the content and its significance. For instance, if the data is about playoff basketball games, then it will be in a text and video format. Another example is the temperature recordings from all the cities of a country, collected and made accessible via different formats. The transformation from data to information involves collection, processing, and organization of data as shown in the following diagram:



The collected data needs some processing and organizing, which later may or may not have a structure, model, or a pattern. However, this process at least gives us an organized way of finding answers to questions about the data. The process could be a simple sorting based on the total points scored by basketball players or a sorting based on the names of the city and state.

The transformation from data to information could also be a little more than just sorting such as statistical modeling or a computational algorithm. It is this transformation from data to information that is really important and enables the data to be queried, accessed, and manipulated. In some cases, when there is a vast and divergent amount of data, the transformation may involve processing methods such as filtering, aggregating, applying correlation, scaling and normalizing, and classifying.

Data collection

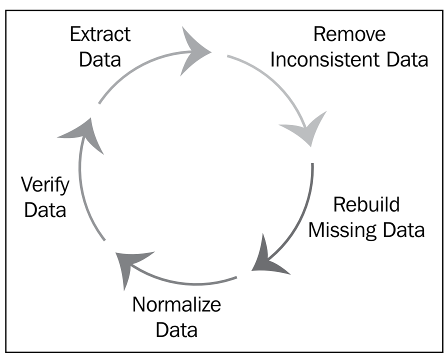
Data collection is a time-consuming process. So, businesses are looking for better ways to automate data capture. However, manual data collection is still prevalent for many processes. Data collection by automatic processes in modern times uses input devices such as sensors. For instance, underwater coral reefs are monitored via sensors; agriculture is another area where sensors are used in monitoring soil properties, controlling irrigation, and fertilization methods.

Another way to collect data automatically is by scanning documents and log files, which is a form of server-side data collection. Manual processes include data collection via web-based methods that get stored in the database, which can then be transformed into information. Nowadays, web-based collaborative environments are benefiting from improved communication and sharing of data.

Traditional visualization and visual analytic tools are typically designed for a single user interacting with a visualization application on a single machine. Extending these tools to include support for collaboration has clearly come a long way towards increasing the scope and applicability of visualizations in the real world.

Data preprocessing

Today, data is highly susceptible to noise and inconsistency due to its size and likely origin from multiple, heterogeneous sources and types. There are several data preprocessing techniques such as *data cleaning*, *data integration*, *data reduction*, and *data transformation*. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges and combines the data from multiple sources into a coherent format, mostly known as data warehouse. Data reduction can reduce data size by, for instance, merging, aggregating, and eliminating the redundant features. Data transformations may be applied where data is scaled to fall within a smaller range, thus improving the accuracy and efficiency in processing and visualizing them. The transformation cycle of data is shown in the following diagram:



Anomaly detection is the identification of unusual data that might not fall into an expected behavior or pattern in the collected data. Anomalies are also known as outliers or noise; for example in signal data, a particular signal that is unusual is considered noise, and in transaction data, an outlier is a fraudulent transaction. Accurate data collection is essential for maintaining the integrity of data. As much as the down side of anomalies, on the flip side, there is also a significant importance of outliers—specifically in cases where one would want to find fraudulent insurance claims, for instance.

Data processing

Data processing is a significant step in the transformation process. It is imperative that the focus be on data quality. Some processing steps that help in preparing data for analyzing and understanding it better are *dependency modeling* and *clustering*. There are other processing techniques, but we will limit our discussion here with the two most popular processing methods.

Dependency modeling is the fundamental principle of modeling data to determine the nature and structure of the representation. This process searches for relationships between the data elements; for example, a department store might gather data on the purchasing habits of its customers. This process helps the department store deduce the information about frequent purchases.

Clustering is the task of discovering groups in the data that have, in some way or another, a "similar pattern", without using known structures in the data.

Organizing data

Database management systems allow users to store data in a structured format. However, the databases are too large to fit into memory. There are two ways of structuring data:

- Storing large data in disks in a structured format like tables, trees, or graphs
- Storing data in memory using data structure formats for faster access

A data structure comprises a set of different formats for structuring data to be able to store and access it. The general data structure types are arrays, files, tables, trees, lists, maps, and so on. Any data structure is designed to organize the data to suit a specific purpose so that it can be stored, accessed, and manipulated at runtime. A data structure may be selected or designed to store data for the purpose of working on it with various algorithms for faster access.

Data that is collected, processed, and organized to be stored efficiently is much easier to understand, which leads to information that can be better understood.

Getting datasets

For readers who do not have access to organizational data, there are plenty of resources on the Internet with rich datasets from several different sources, such as:

- `http://grouplens.org` (from the University of Minnesota)
- `http://ichart.finance.yahoo.com/table.csv?s=YHOO&c=1962`
- `http://datawrangling.com/some-datasets-available-on-the-web`
- `http://weather-warehouse.com` (weather data)
- `http://www.bjs.gov/developer/ncvs/` (Bureau of Justice Statistics)
- `http://census.ire.org/data/bulkdata.html` (census data)
- `http://ww.pro-football-reference.com` (football reference)
- `http://www.basketball-reference.com` (basketball reference)
- `http://www.baseball-reference.com` (baseball reference)
- `http://archive.ics.uci.edu/ml/datasets.html` (machine learning)
- `http://www.pewresearch.org/data/download-datasets/`
- `http://archive.ics.uci.edu/ml/datasets/Heart+Disease` (heart disease)

Transforming information into knowledge

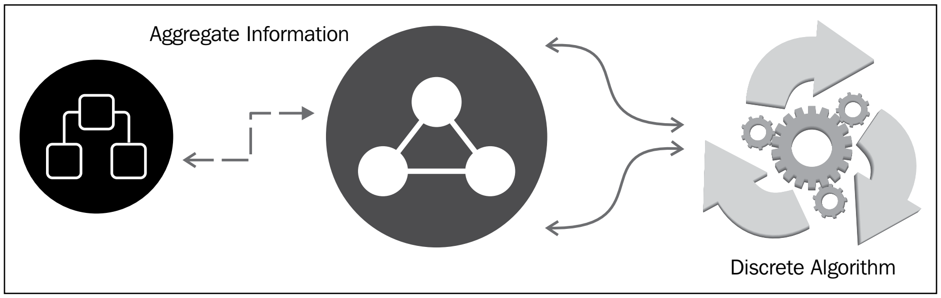
Information is quantifiable and measurable, it has a shape, and can be accessed, generated, stored, distributed, searched for, compressed and duplicated. It is quantifiable by the volume or amount of information.

Information transforms into knowledge by the application of discrete algorithms, and knowledge is expected to be more qualitative than information. In some problem domains, knowledge continues to go through an evolving cycle. This evolution happens particularly when the data changes in real time.

Knowledge is like the recipe that lets you make bread out of the information, in this case, the ingredients of flour and yeast. Another way to look at knowledge is as the combination of data and information, to which experience and expert opinion is added to aid decision making. Knowledge is not merely a result of filtering or algorithms.

What are the steps involved in this transformation, and how does the change happen? Naturally, it cannot happen by itself. Though the word information is subject to different interpretations based on the definition, we will explore it further within the context of computing.

A simple analogy to illustrate the difference between information and knowledge: course materials for a particular course provide you the necessary information about the concepts, and the teacher later helps the students to understand the concepts through discussions. This helps the students in gaining knowledge about the course. By a similar process, something needs to be done to transform information into knowledge. The following diagram shows the transformation from information to knowledge:



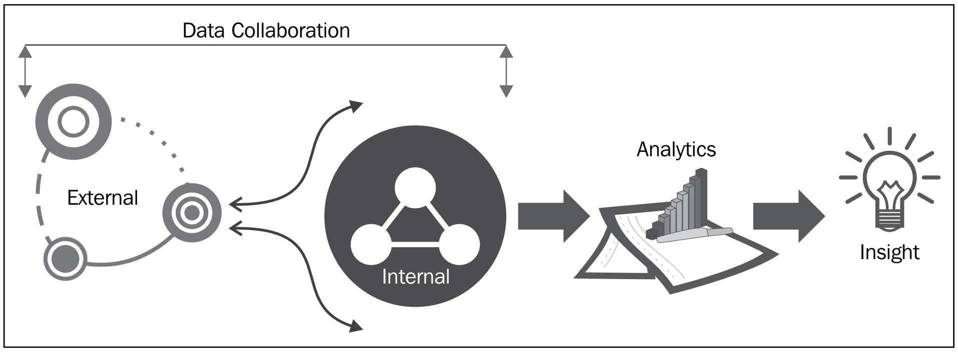
As illustrated in the figure, information when aggregated and run through some discrete algorithms, gets transformed into knowledge. The information needs to be aggregated to get broader knowledge. The knowledge obtained by this transformation helps in answering questions about the data or information such as which quarter did the company have maximum revenue from sales? How much has advertising driven the sales? Or, how many new products have been released this year?

Transforming knowledge into insight

In the traditional system, information is processed, and then analyzed to generate reports. Ever since the Internet came into existence, processed information is already and always available, and social media has emerged as a new way of conducting business.

Organizations have been using external data to gain insights via data analysis. For example, the measure of user sentiments from tweets by consumers via Twitter is used to follow the opinions about product brands. In some cases, there is a higher percentage of users giving a positive message on social media about a new product, say an iPhone or a tablet computer. The analytical tool can provide numerical evidence of that sentiment, and this is where data visualization plays a significant role.

Another example to illustrate this transformation, Netflix announced a competition in 2009 for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings. The winner of that competition used the pragmatic theory and achieved a 10.05 percent improvement in predicting user ratings, which increased the business value for Netflix.



Transforming knowledge into insight is achieved using collaboration and analytics as shown in the preceding diagram. Insight implies seeing the solution and realizing what needs to be done. Achieving data and information is easy and organizations have known methods to achieve that, but getting insight is very hard. Achieving insight requires new and creative thinking and the ability to connect the dots. In addition to applying creative thinking, data analysis and data visualization play a big role in achieving insight. Data visualization is considered both an art and a science.

Data visualization history

Visualization has its roots in a long historical tradition of representing information using primitive paintings and maps on walls, tables of numbers, and paintings on clay. However, they were not known as visualization or data visualization. Data visualization is a new term; it expresses the idea that it involves more than just representing data in a graphical form. The information behind the data should be revealed in an intuitive representation using good display; the graphic should inherently aid viewers in seeing the structure of data.

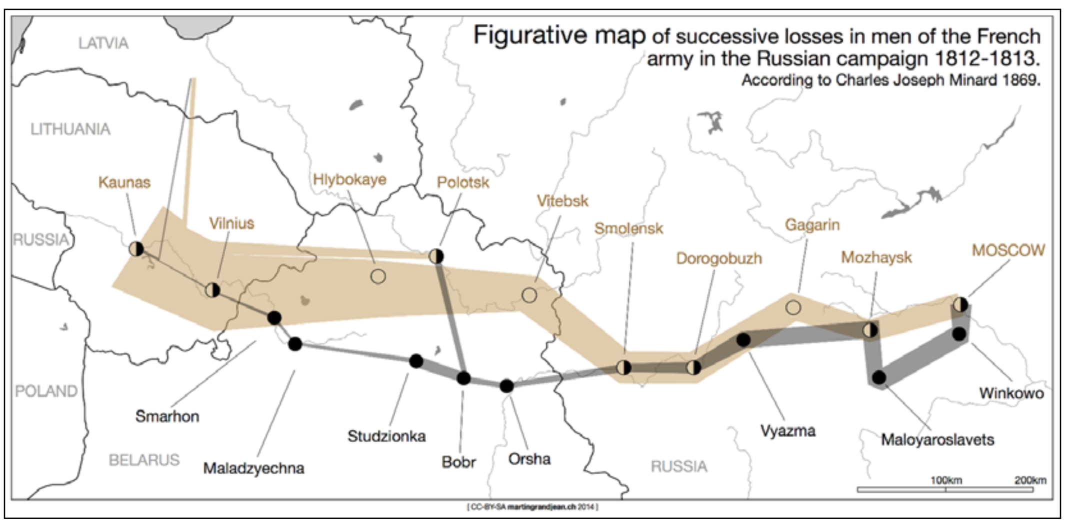
Visualization before computers

In early Babylonian times, pictures were drawn on clay and in the later periods were rendered on papyrus. The goal of those paintings and maps was to provide the viewer with a qualitative understanding of the information. We also know that understanding pictures are our natural instincts as a visual presentation of information is perceived with greater ease. This section includes only partial details about the history of visualization. For elaborate details and examples, we recommend two interesting resources:

- Data visualization (`http://euclid.psych.yorku.ca/datavis/`)
- The work of Edward Tufte and Graphics Press (`www.edwardtufte.com/tufte`)

Minard's Russian campaign (1812)

Charles Minard was a civil engineer working in Paris. He summarized the War of 1812—Napoleon's march on Moscow—in a figurative map. This map is a simple picture, which is both a visual timeline and a geographic map depicting the size and direction of the army, temperature, and the landmarks and locations. Prof. Edward Tufte famously described this picture as possibly being *the best statistical graphic ever drawn*.



The wedge starts with being thick on the left-hand side, and we see the army begin the campaign at the Polish border with 422,000 men. The wedge becomes narrower as it gets deeper into Russia and the temperature gets lower. This visualization manages to condense a number of different numeric and geographic facts into one image: when the army gets reduced, the reason for the reduction, and subsequently, their retreat.

The Cholera epidemics in London (1831-1855)

In October 1831, the first case of Asiatic cholera occurred in Great Britain, and over 52,000 people died in the epidemic. Subsequently, in 1848-1849 and 1853-1854, more cholera epidemics produced large death tolls.

In 1855, Dr. John Snow produced a map showing the deaths due to cholera clustered around the Broad Street pump in London. This map by Dr. John Snow was a landmark graphic discovery, but unfortunately, it was devised at the end of that period. His map showed the location of each of the deceased, and that provided an insight for his conclusion that the source of outbreak could be localized to contaminated water from a pump on Broad Street. Around that time, the use of graphs became important in economic and state planning.

Statistical graphics (1850-1915)

By the mid 18th century, a rapid growth of visualization had been established throughout Europe. In 1863, one page of Galton's multivariate weather chart of Europe showed barometric pressure, wind direction, rain, and temperature for the month of December 1861 (source: The life, letters and labors of Francis Galton, Cambridge University Press).

During this period, statistical graphics became mainstream and there were many textbooks written on the same. These textbooks contained detailed descriptions of the graphic method, discussing frequencies, and the effects of the choice of scales and baselines on the visual estimation of differences and ratios. They also contained historical diagrams in which two or more time series could be shown on a single chart for comparative views of their histories.

Later developments in data visualization

In the year 1962, John W. Tukey issued a call for the recognition of data analysis as a legitimate branch of statistics; shortly afterwards, he began the invention of a wide variety of new, simple, and effective graphic displays under the rubric **Exploratory Data Analysis** (**EDA**), which was followed by **Exploratory Spatial Data Analysis** (**ESDA**). Tukey later wrote a book titled *Exploratory Data Analysis* in 1977. There are a number of tools that are useful for EDA with graphical techniques, which are listed as follows:

- Box-and-whisker plot (box plot)
- Histogram
- Multivari chart (from candlestick charts)
- Run-sequence plot
- Pareto chart (named after Vilfredo Pareto)
- Scatter plot
- Multidimensional scaling
- Targeted projection pursuit

Visualization in scientific computing is emerging as an important computer-based field, with the goal to improve the understanding of data and to make quick real-time decisions. Today, the ability of medical doctors to diagnose ailments is dependent upon vision. For example, in hip-replacement surgeries, custom hips can now be fabricated before surgical procedures. Accurate measurements can be made prior to surgery using non-invasive 3D imaging thereby reducing the number of post-operative body rejections from 30 percent to a mere 5 percent (source: `http://bonesmart.org/hip/hip-implants-specialized-and-custom-fitted-options/`).

Visualization of the human brain structure and function in 3D is a research frontier of far-reaching importance. Few advances have transformed the fields of neuroscience and brain-imaging technology, like the ability to see inside and read the brain of a living human. For continued progress in brain research, it will be necessary to integrate structural and functional information at many levels of abstraction.

The rate at which the hardware performance power has been on the rise tells us that we are already able to analyze DNA sequences and visually represent them. The future advances in computing promises a much brighter progress in the fields of medicine and other scientific areas.

How does visualization help decision-making?

There is a variety of ways to represent data visually. However, there are only a few ways in which one can portray the data in a manner that allows one to see something visually and observe new patterns. Data visualization is not as easy as it seems; it is an art and requires a great deal of practice and experience. (Just like painting a picture—one cannot be a master painter from day one, it takes a lot of practice.)

Human perception plays an important role in the field of data visualization. A pair of healthy human eyes has a total field view of approximately 200 degrees horizontally (about 120 degrees of which are shared by both the eyes). About one quarter of the human brain is involved in visual processing, which is more than any other sense. Among the three senses of hearing, seeing, and smelling, human vision has the maximum sense—measured to be sixty per cent (`http://contemplatingmadness.tumblr.com/post/27478393311/10-limits-to-human-perception-and-how-they-shape`).

Effective visualization helps us in analyzing and understanding data. Author Stephen Few described the following eight types of quantitative messages (via visualization) that may help us with understanding or communicating from a set of data (source: `https://www.perceptualedge.com/articles/ie/the_right_graph.pdf`):

- Time-series
- Ranking
- Part-to-whole
- Deviation
- Frequency distribution
- Correlation
- Nominal comparison
- Geographic or geospatial

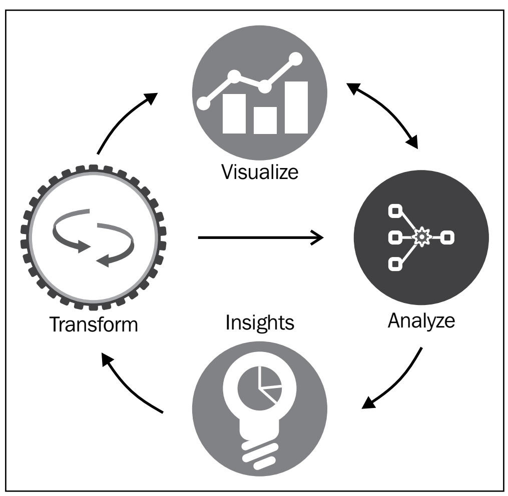
Scientists have mapped the human genome, and this is one of the reasons why we are faced with the challenges of transforming knowledge into a visual representation for better understanding. In other words, we may have to find new ways to visually present the human genome so that it is not difficult for a common person to understand.

Where does visualization fit in?

It is important to note that data visualization is not scientific visualization. Scientific visualization deals with the data that has an inherent physical structure, such as air molecules flowing over an aircraft wing. Information visualization, on the other hand, deals with abstract data, and helps in solving problems involving large datasets. One of the challenges is to ensure that the data is clean and subsequently, to reduce the dimensions so that unnecessary information is discarded.

Visualization can be used wherever we see increased knowledge or value of data. That can be determined by doing more data analysis and running through algorithms. The data analysis might vary from the simplest form to a more complicated one.

Sometimes, there is value in looking at data beyond the mean, median, or total, because these measurements only measure things that may seem obvious. Sometimes, aggregates or values around a region hide the interesting details that need special focus. One classic example is the "Anscombe's quartet" which comprises of four datasets that have nearly identical simple statistical properties yet appear very different when graphed. For more on this, one can refer to the link, `https://en.wikipedia.org/wiki/Anscombe%27s_quartet`.



Mostly, datasets that lend themselves well to visualization can take different forms, but some paint a clearer picture to understand than others. In some cases, it is mandatory to analyze them several times to get a much better understanding of the visualization as shown in the preceding diagram.

A good visualization is not just a static picture that one can look at, like an exhibit in a museum. It is something that allows us to drill down and find more about the change in data. For example, view first, zoom and filter, change the values of some scale of display, and view the results in an incremental way, as described in `http://www.mat.ucsb.edu/~g.legrady/academic/courses/11w259/schneiderman.pdf` by Ben Shneiderman. Sometimes, it is much harder to display everything on a single display and on a single scale, and only by experience, one can better understand these visualization methods. Summarizing further, visualization is useful in both organizing and making sense out of data, particularly when it is in abundance.

Interactive visualization is emerging as a new form of communication, which allows users to analyze the information in order to construct their own, new understanding of the data.

Data visualization today

While many areas of computing aim to replace human judgment with automation, visualization systems are unique and are explicitly designed not to replace humans. In fact, they are designed to keep the humans actively involved in the whole process; why is that?

Data Visualization is an art, driven by data and yet created by humans with the help of various computing tools. An artist paints a picture using tools and materials like brushes, and colors. Similarly, another artist tries to create data visualization with the help of computing tools. Visualization can be aesthetically pleasing and helps in making things clear; sometimes, it may lack one or both of those qualities depending on the users who create it.

Today, there are over thirty different visual representations of data, each having a reason to represent data in that specific way. As the visualization methods progress, we have much more than just bar graphs and pie charts. Despite the many benefits of data visualization, they are undermined due to a lack of understanding and, in some cases, due to cluttering together of things on a dashboard that becomes too cumbersome.

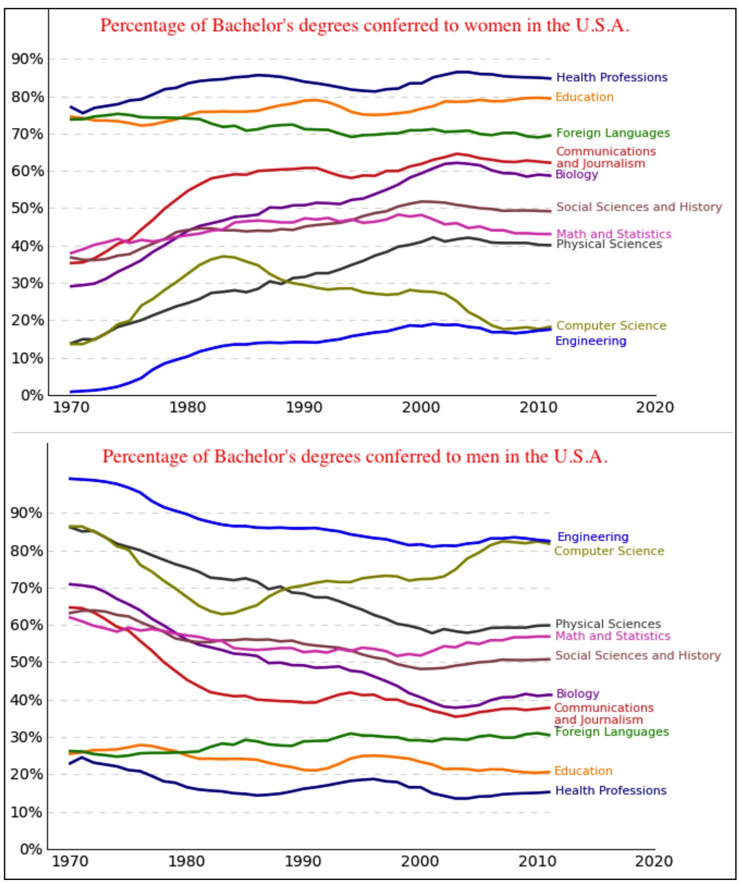
There are many ways to present data, but only a handful of those make sense in most cases; this will be explained in detail in later sections of this chapter. Before that discussion, let us take a look at a list of some important things that make a good visualization.

What is a good visualization?

Good visualization helps the users to explore and understand data, providing value and deep insights. It is effective, visually appealing, scalable, and is easy to understand (good visualization does not have to be too complicated). Visualization is a central tool in finding patterns and trends in the data by carrying out research and analysis, using whichever one can answer questions about the data.

The main principle behind an effective visualization is to identify the main point that you want to make, recognize the level and background of your audience, accurately represent the data, and then create a clear presentation that conveys the message to that audience.

Example: The following representations have been created with a small sample data source that shows the percentage of women and men conferred with degrees in ten different disciplines for the years from 1970-2012 (`womens-undergrad-degrees.csv` and `mens-undergrad-degrees.csv` from `http://www.knapdata.com/python/`):



The full data source available at `http://nces.ed.gov/programs/digest/d11/tables/dt11_290.asp` maintains the complete set of data.

One simple way is to represent them on one scale, although there is no relationship between the numbers between the different disciplines. Let us analyze and see if this representation makes sense, and if it doesn't, then what else do we need? Are there any other representations?

For one thing, all the data about the different disciplines is displayed on one screen, which is an excellent comparison. However, if we need to get the information for the year 2000, there is no straightforward way. Unless there is an interactive mode of display that is similar to a financial stock chart, there is no easy way to determine the information about the degrees conferred in multiple disciplines for the year 2000. Another confusing part of these plots is that the percentage doesn't add up to a sum of 100 percent. On the other hand, the percentage of conferred degrees within one discipline for men and women add up to 100 percent; for instance, the percentage of degrees conferred in the **Health Professions** discipline for men and women are 15.2 percent and 84.8 percent respectively.

Can we represent these through other visualization methods? One can create bubble charts for each year, have an interactive visualization with year selection, and also have a play button that transitions the bubbles for each year.

This visualization better suits the data that we are looking at. We can also use the same slider with the original plot and make it interactive by highlighting the data for the selected year. It is a good habit to visualize the data in several different ways to see if some display makes more sense than the other. We may have to scale the values on a logarithmic scale if there is a very large range of numerical values (for example, from 20 to 200,000).

One can write a program in Python to accomplish this bubble chart. Other alternate languages are JavaScript using D3.js and R using R-Studio. It is left for the reader to explore other visualization options.