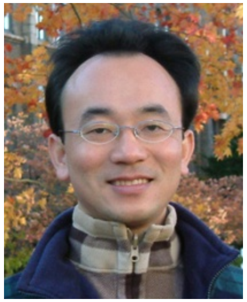
資料探勘
(Data Mining)

資料科學與資料探勘：
發現，分析，可視化和呈現數據

(Data Science and Data Mining: Discovering, Analyzing, Visualizing and Presenting Data)

1092DM04
MBA, IM, NTPU (M5026) (Spring 2021)
Tue 2, 3, 4 (9:10-12:00) (B8F40)



Min-Yuh Day
戴敏育
Associate Professor
副教授
Institute of Information Management, National Taipei University
國立臺北大學 資訊管理研究所
https://web.ntpu.edu.tw/~myday

2021-03-16

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2021/02/23	資料探勘介紹 (Introduction to data mining)
2	2021/03/02	ABC：人工智慧，大數據，雲端運算 (ABC: AI, Big Data, Cloud Computing)
3	2021/03/09	Python資料探勘的基礎 (Foundations of Data Mining in Python)
4	2021/03/16	資料科學與資料探勘：發現，分析，可視化和呈現數據 (Data Science and Data Mining: Discovering, Analyzing, Visualizing and Presenting Data)
5	2021/03/23	非監督學習：關聯分析，購物籃分析 (Unsupervised Learning: Association Analysis, Market Basket Analysis)
6	2021/03/30	資料探勘個案研究 I (Case Study on Data Mining I)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2021/04/06	非監督學習：集群分析，行銷市場區隔 (Unsupervised Learning: Cluster Analysis, Market Segmentation)
8	2021/04/13	監督學習：分類和預測 (Supervised Learning: Classification and Prediction)
9	2021/04/20	期中報告 (Midterm Project Report)
10	2021/04/27	監督學習：分類和預測 (Supervised Learning: Classification and Prediction)
11	2021/05/04	機器學習和深度學習 (Machine Learning and Deep Learning)
12	2021/05/11	卷積神經網絡 (Convolutional Neural Networks)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2021/05/18	資料探勘個案研究 II (Case Study on Data Mining II)
14	2021/05/25	遞歸神經網絡 (Recurrent Neural Networks)
15	2021/06/01	強化學習 (Reinforcement Learning)
16	2021/06/08	社交網絡分析 (Social Network Analysis)
17	2021/06/15	期末報告 I (Final Project Report I)
18	2021/06/22	期末報告 II (Final Project Report II)

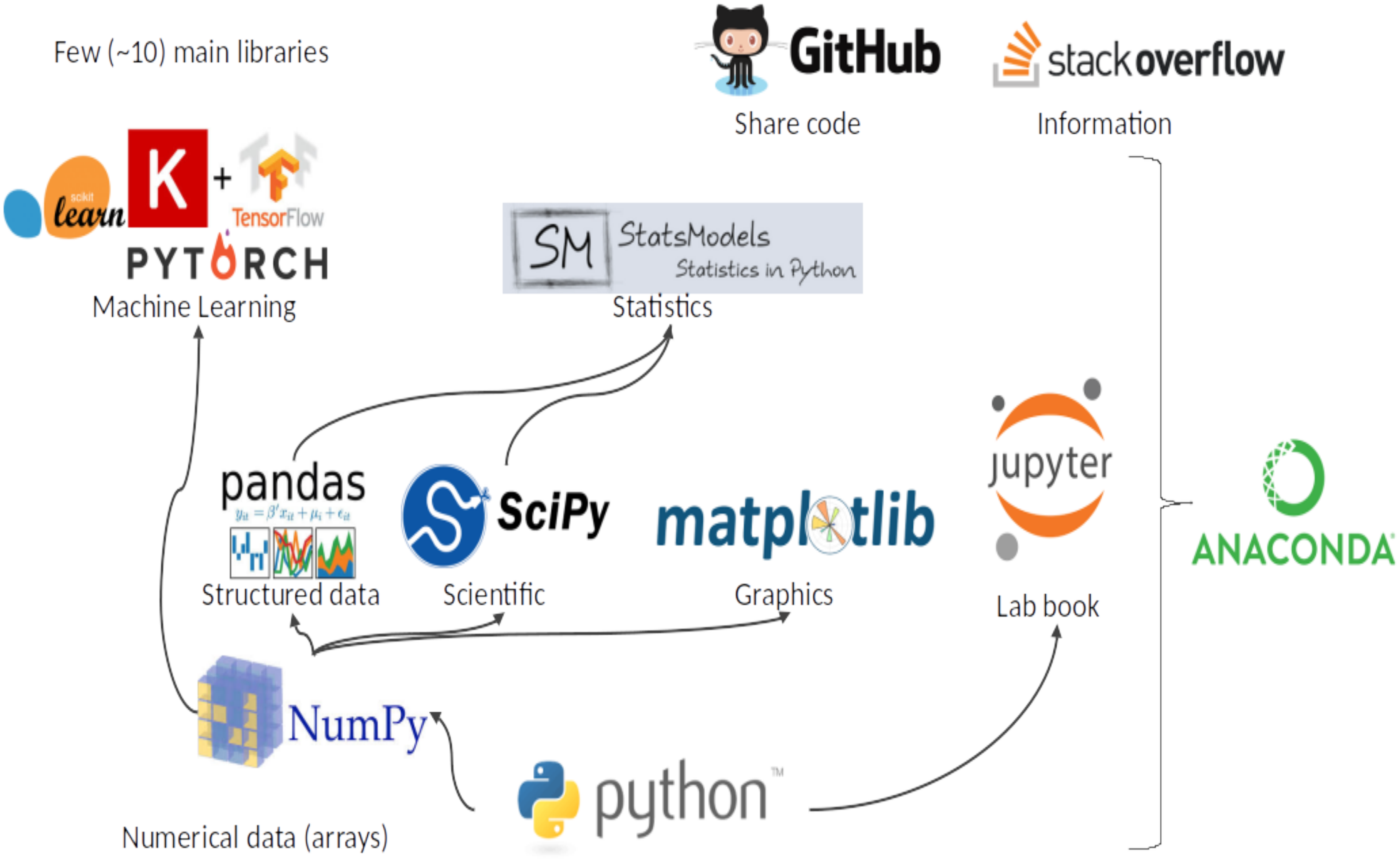
Data Science and Data Mining:

Discovering, Analyzing, Visualizing and Presenting Data

Outline

- Data Science and Data Mining
- Discovering, Analyzing, Visualizing and Presenting Data with Python
    - Pandas
    - Matplotlib
    - Seaborn
    - Plotly
    - Bokeh, Altair

Python Ecosystem for Data Science



Few (~10) main libraries

Share code

Information

Machine Learning

Statistics

Structured data

Scientific

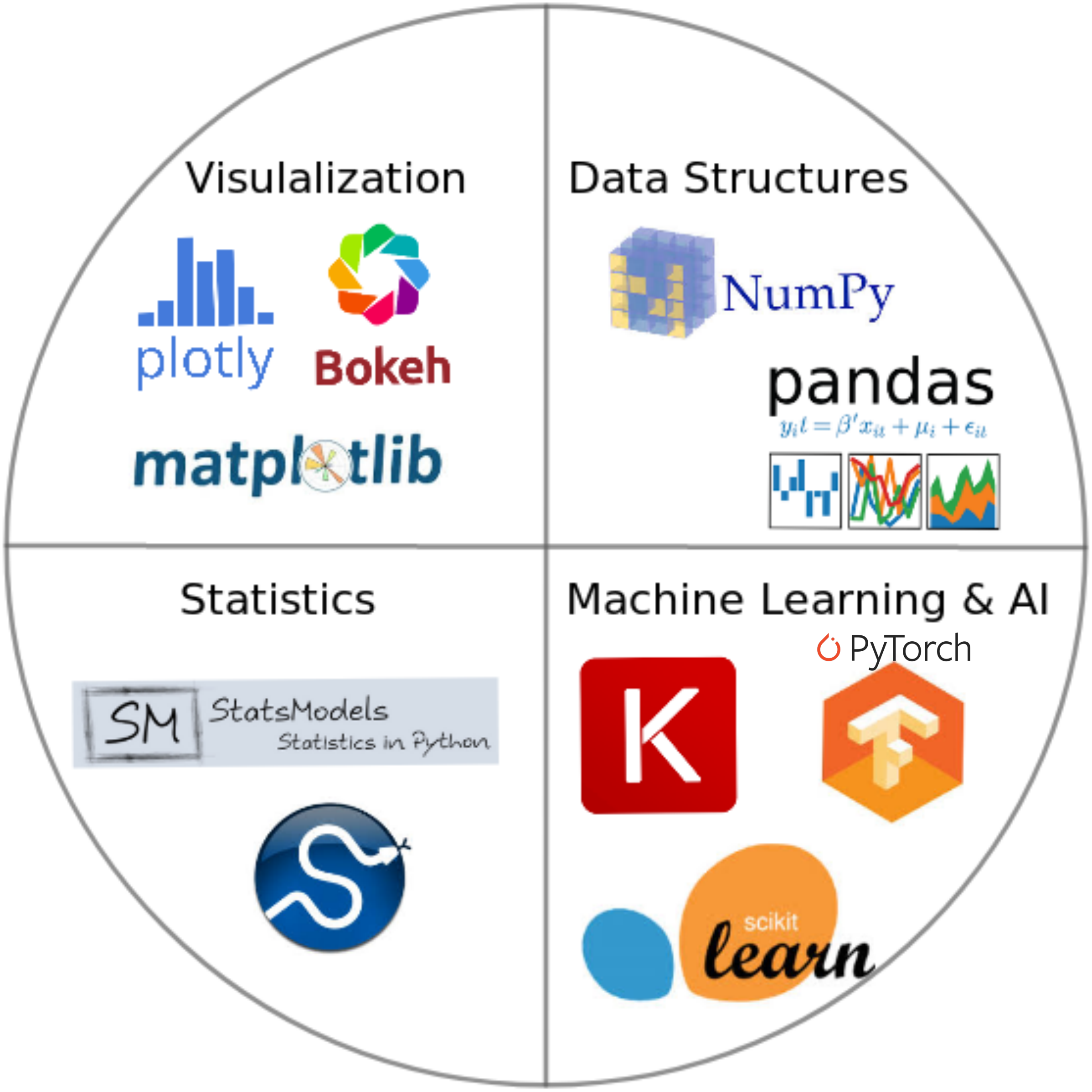
Graphics

Lab book

Numerical data (arrays)

Source: https://medium.com/pyfinance/why-python-is-best-choice-for-financial-data-modeling-in-2019-c0d0d1858c45

Python Ecosystem for Data Science



Visulalization

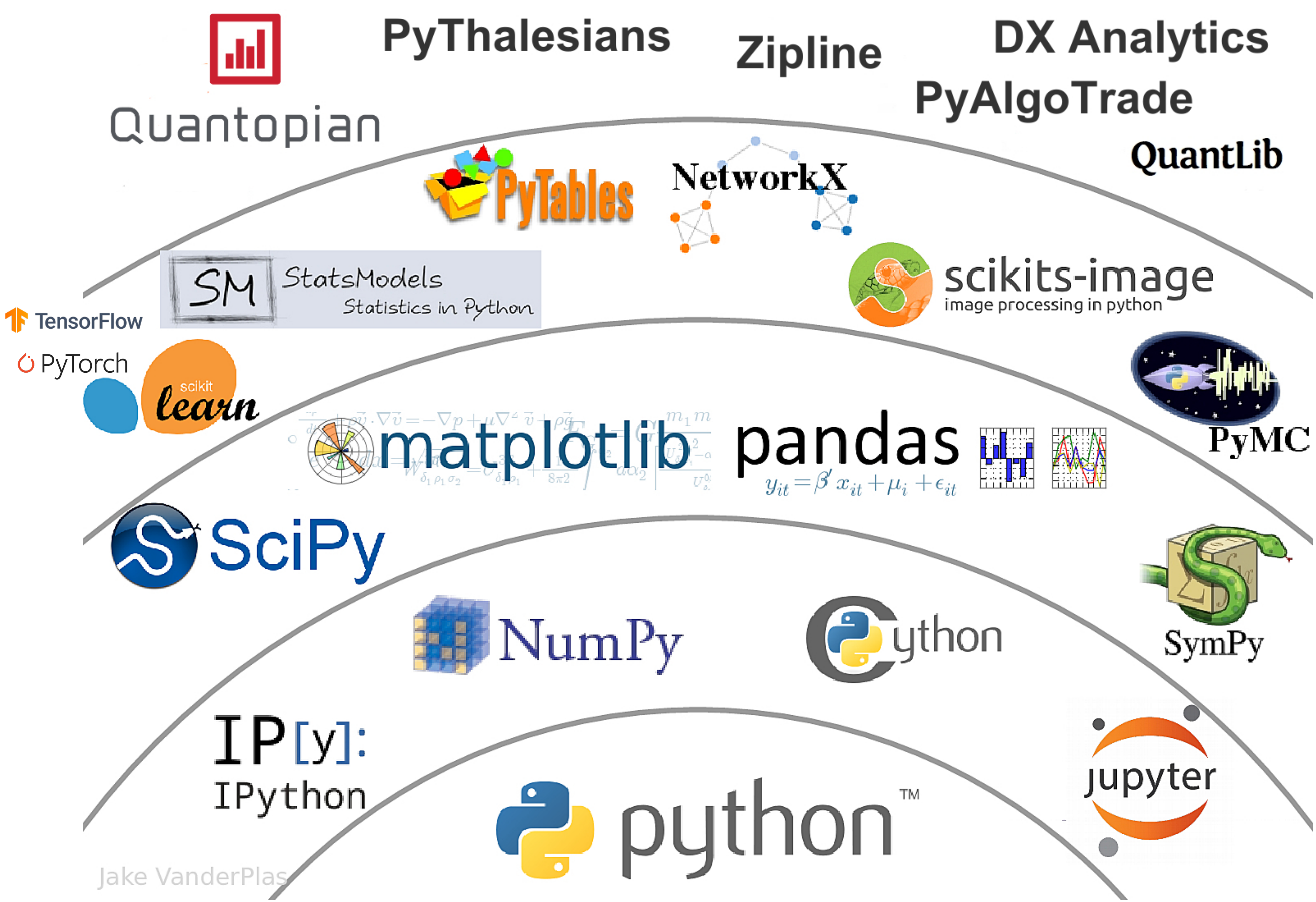
Data Structures

Statistics

Machine Learning & AI

Source:https://duchesnay.github.io/pystatsml/introduction/python_ecosystem.html

The Quant Finance PyData Stack



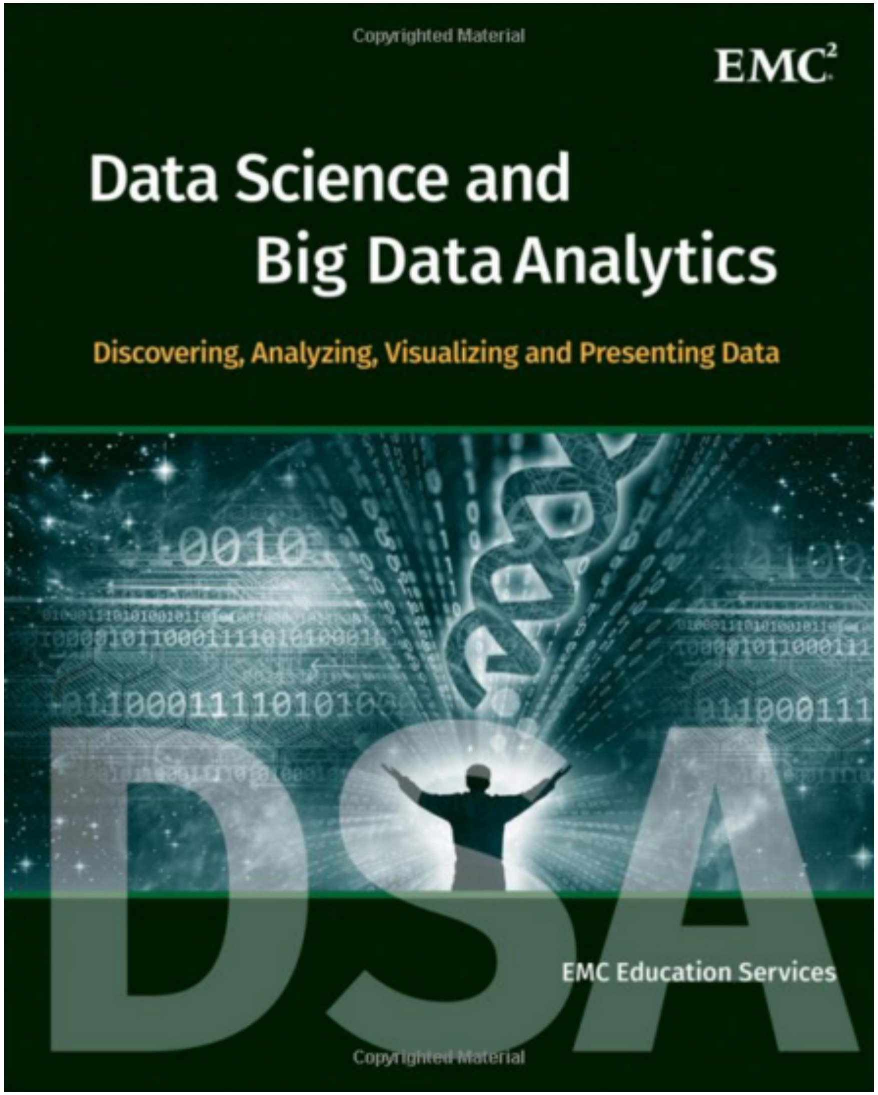
Source: http://nbviewer.jupyter.org/format/slides/github/quantopian/pyfolio/blob/master/pyfolio/examples/overview_slides.ipynb#/5

EMC Education Services,

Data Science and Big Data Analytics:

Discovering, Analyzing, Visualizing and Presenting Data,

Wiley, 2015



Source: http://www.amazon.com/Data-Science-Big-Analytics-Discovering/dp/111887613X

Data Science

Data Analyst

- Data analyst is just another term for professionals who were doing BI in the form of data compilation, cleaning, reporting, and perhaps some visualization.
- Their skill sets included Excel, some SQL knowledge, and reporting.
- You would recognize those capabilities as descriptive or reporting analytics.

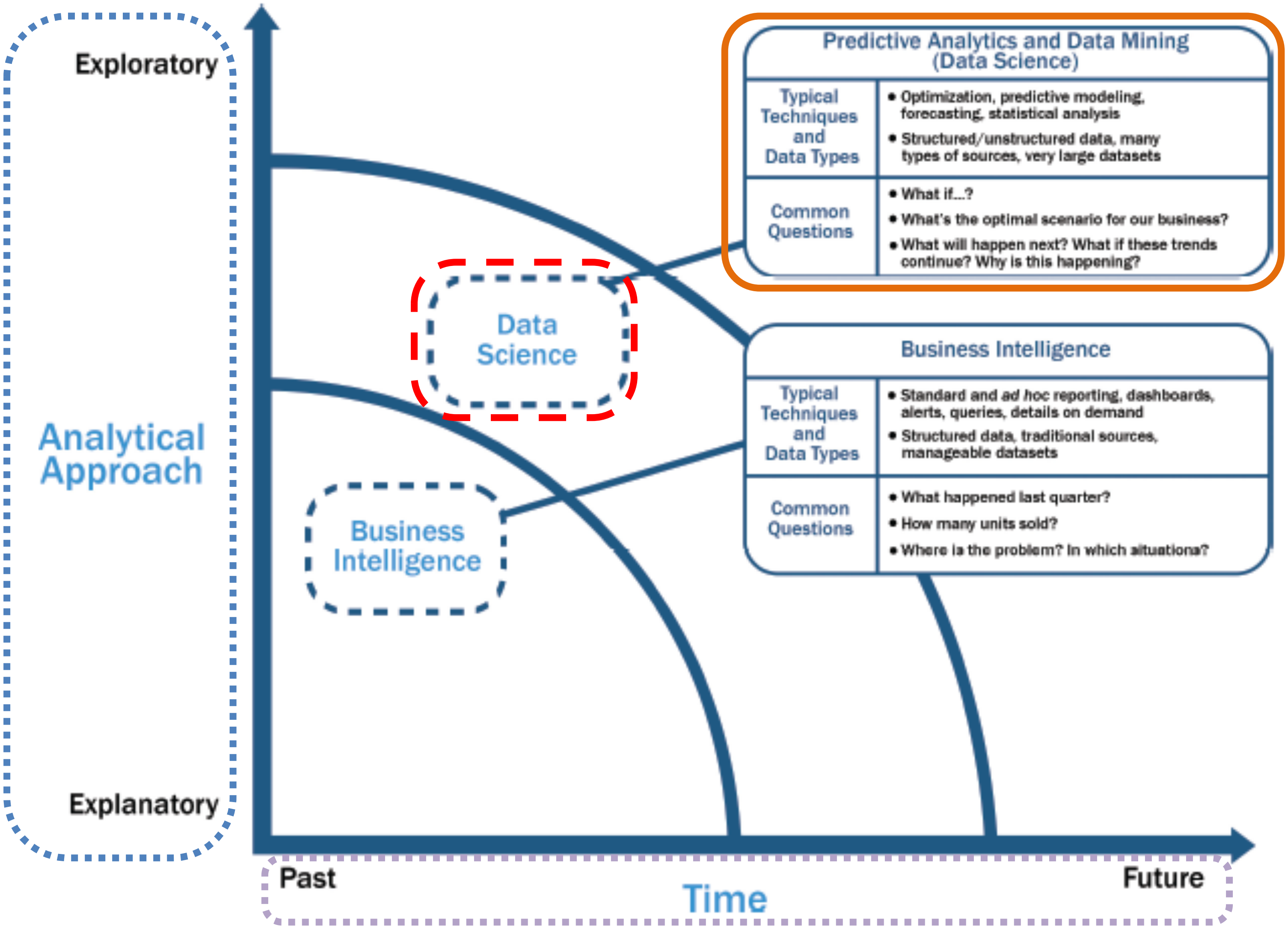
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Data Scientist

- Data scientist is responsible for predictive analysis, statistical analysis, and more advanced analytical tools and algorithms.
- They may have a deeper knowledge of algorithms and may recognize them under various labels—data mining, knowledge discovery, or machine learning.
- Some of these professionals may also need deeper programming knowledge to be able to write code for data cleaning/analysis in current Web-oriented languages such as Java or Python and statistical languages such as R.
- Many analytics professionals also need to build significant expertise in statistical modeling, experimentation, and analysis.

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Data Science and Business Intelligence

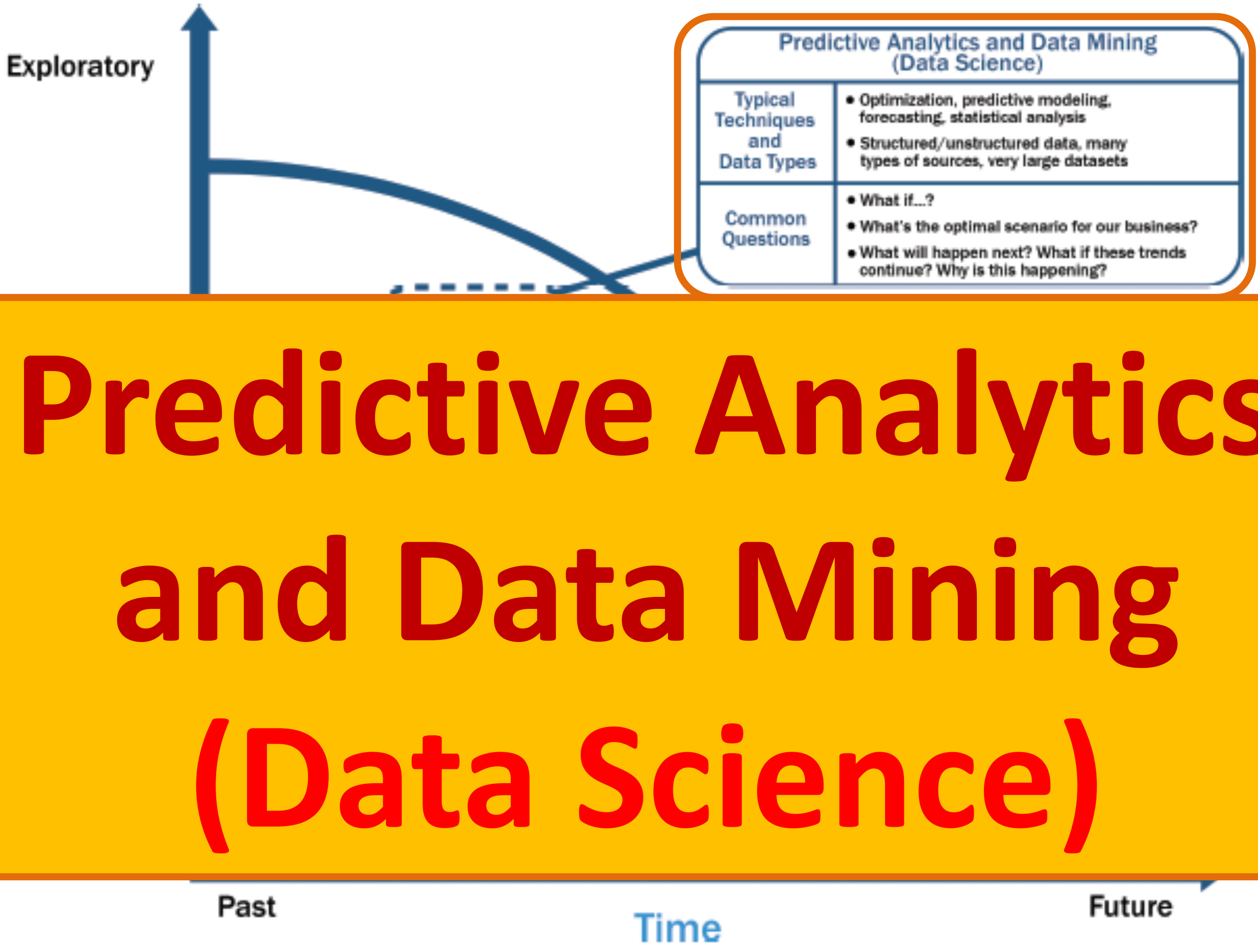


Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which situations?

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Science and Business Intelligence



Predictive Analytics and Data Mining (Data Science)

Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Predictive Analytics and Data Mining (Data Science)

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Predictive Analytics and Data Mining (Data Science)

Structured/unstructured data, many types of sources, very large datasets

Optimization, predictive modeling, forecasting statistical analysis

What if…?

What's the optimal scenario for our business?

What will happen next?

What if these trends countinue?

Why is this happening?

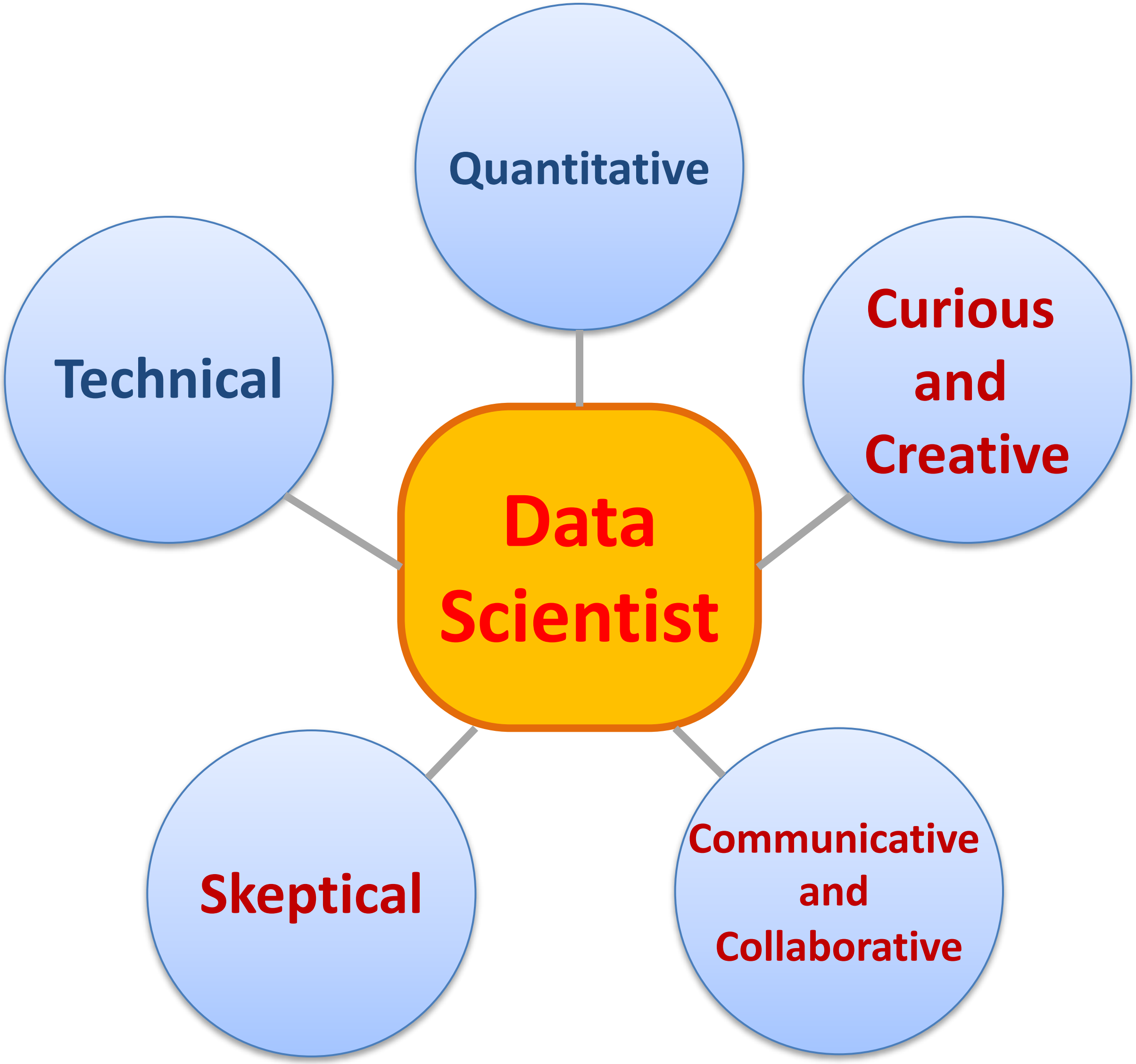
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Profile of a Data Scientist

- **Quantitative**
  - mathematics or statistics
- **Technical**
  - software engineering,
    machine learning,
    and programming skills
- **Skeptical mind-set** and **critical thinking**
- **Curious** and **creative**
- **Communicative** and **collaborative**

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Scientist Profile



Quantitative

Technical

Curious and Creative

Data Scientist

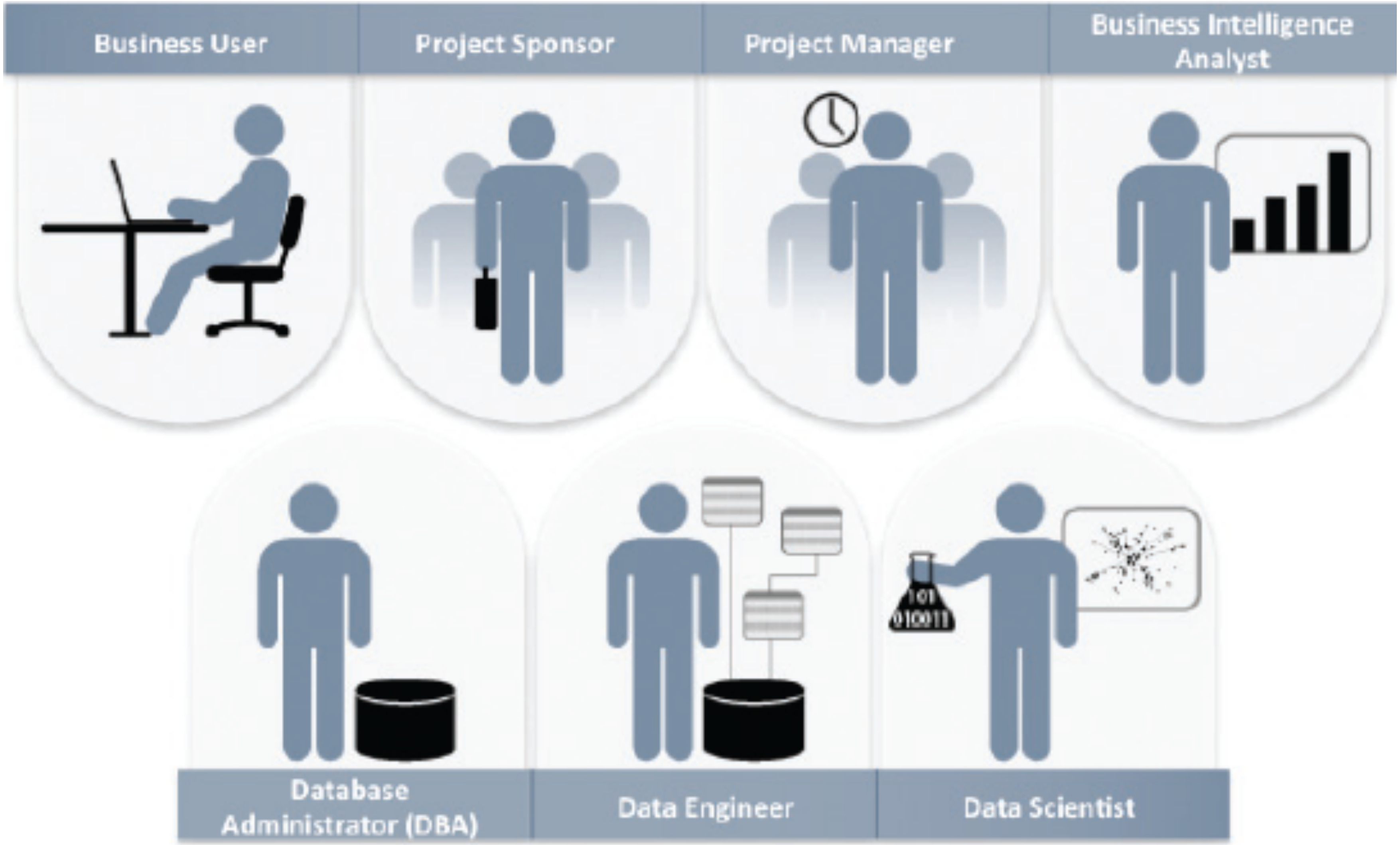
Skeptical

Communicative and Collaborative

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

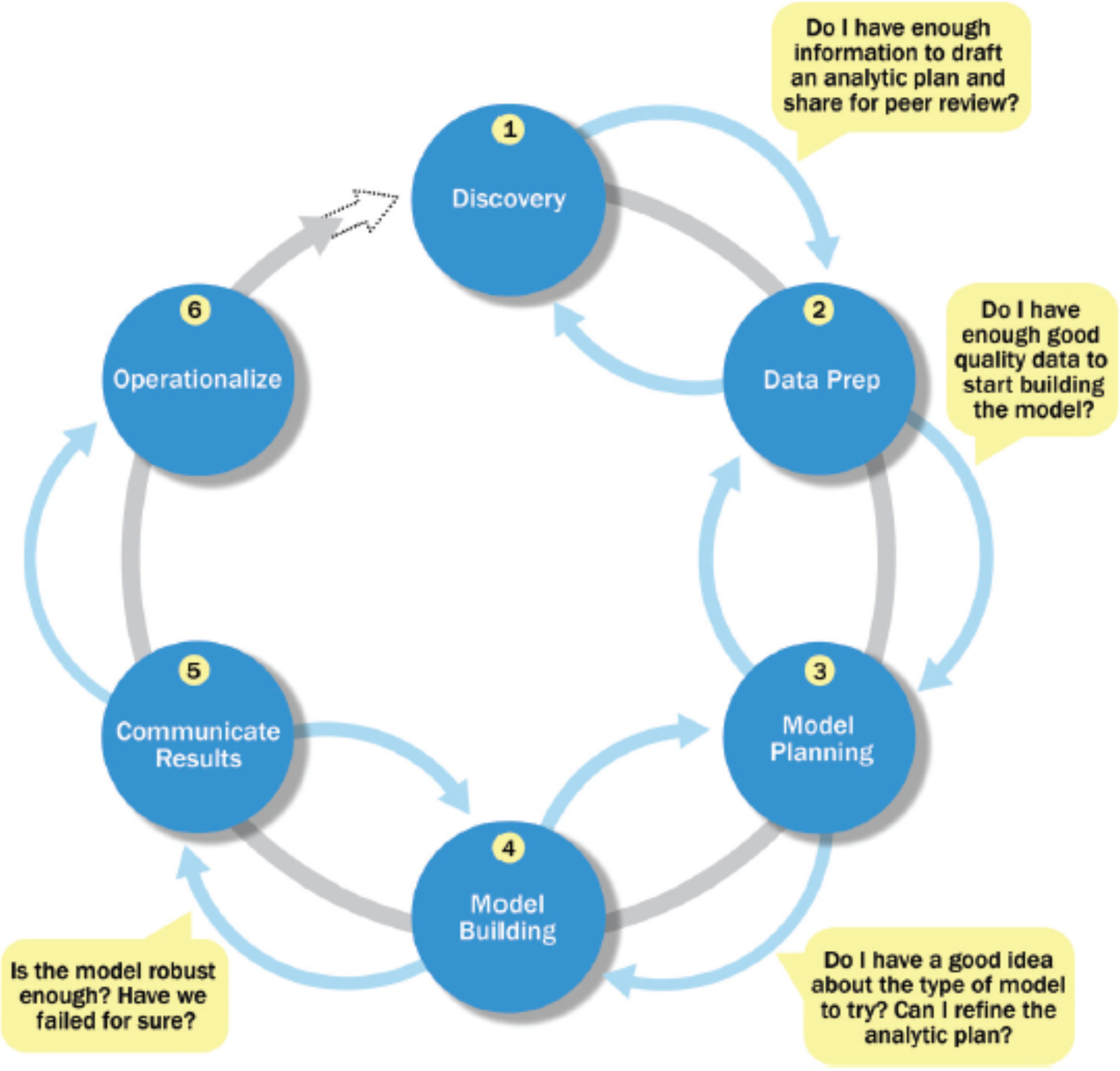
Big Data Analytics Lifecycle

Key Roles for a Successful Analytics Project



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle


1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

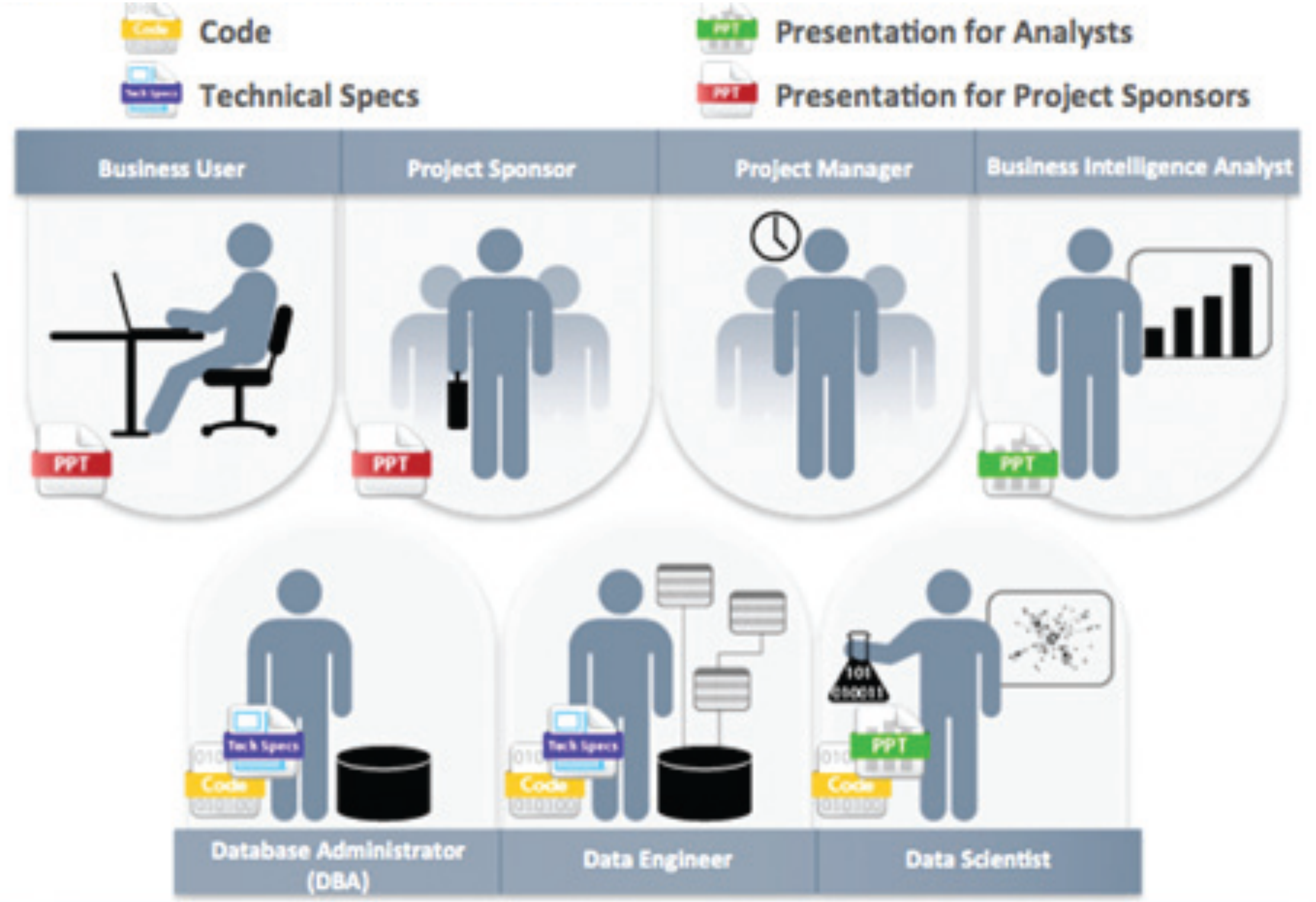
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Key Outputs from a Successful Analytics Project

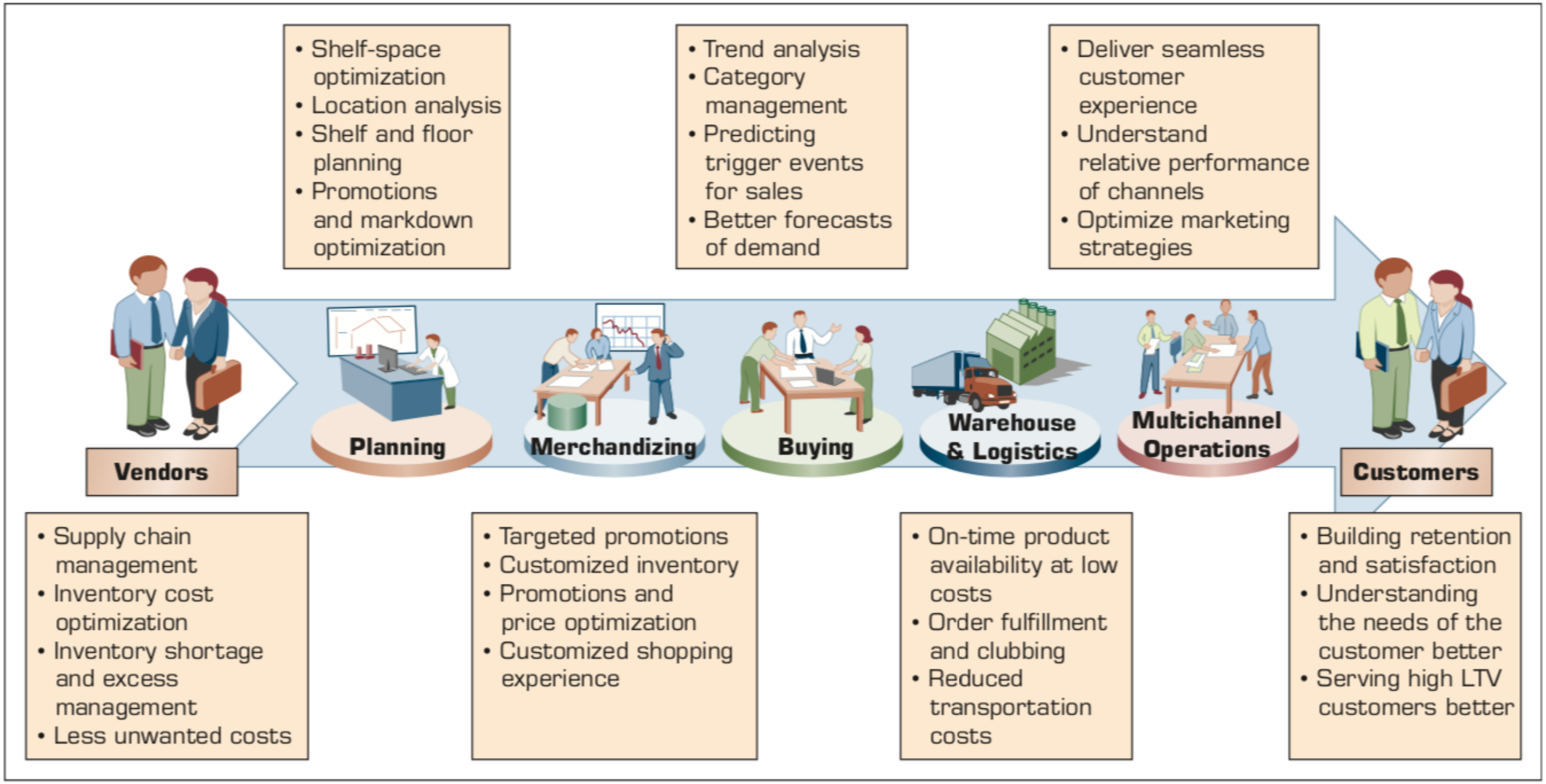


Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Example of Analytics Applications in a Retail Value Chain

Retail Value Chain

Critical needs at every touch point of the Retail Value Chain



Vendors

- Supply chain management
- Inventory cost optimization
- Inventory shortage and excess management
- Less unwanted costs

Planning

- Shelf-space optimization
- Location analysis
- Shelf and floor planning
- Promotions and markdown optimization

Merchandizing

- Targeted promotions
- Customized inventory
- Promotions and price optimization
- Customized shopping experience

Buying

- Trend analysis
- Category management
- Predicting trigger events for sales
- Better forecasts of demand

Warehouse & Logistics

- On-time product availability at low costs
- Order fulfillment and clubbing
- Reduced transportation costs

Multichannel Operations

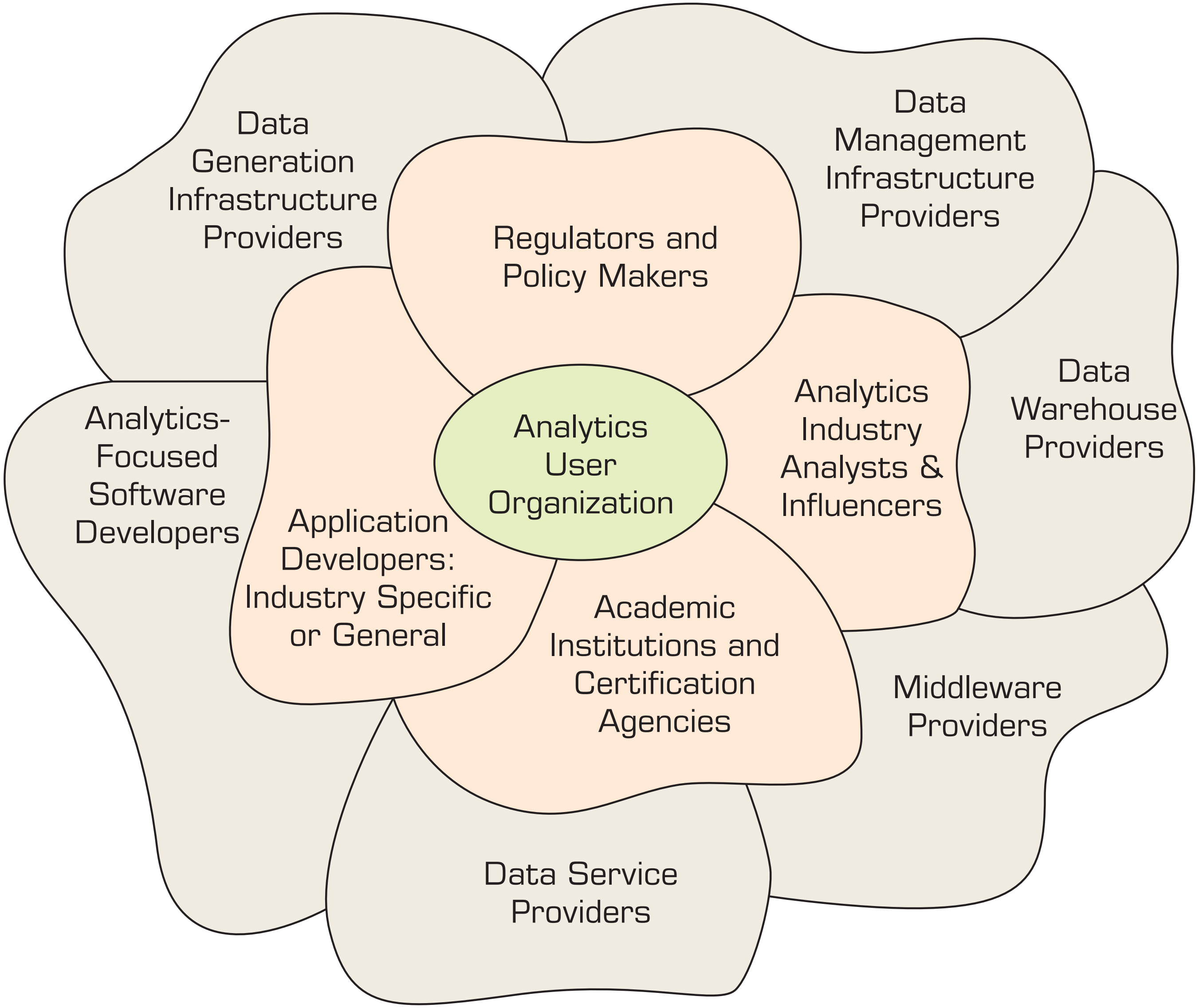
- Deliver seamless customer experience
- Understand relative performance of channels
- Optimize marketing strategies

Customers

- Building retention and satisfaction
- Understanding the needs of the customer better
- Serving high LTV customers better

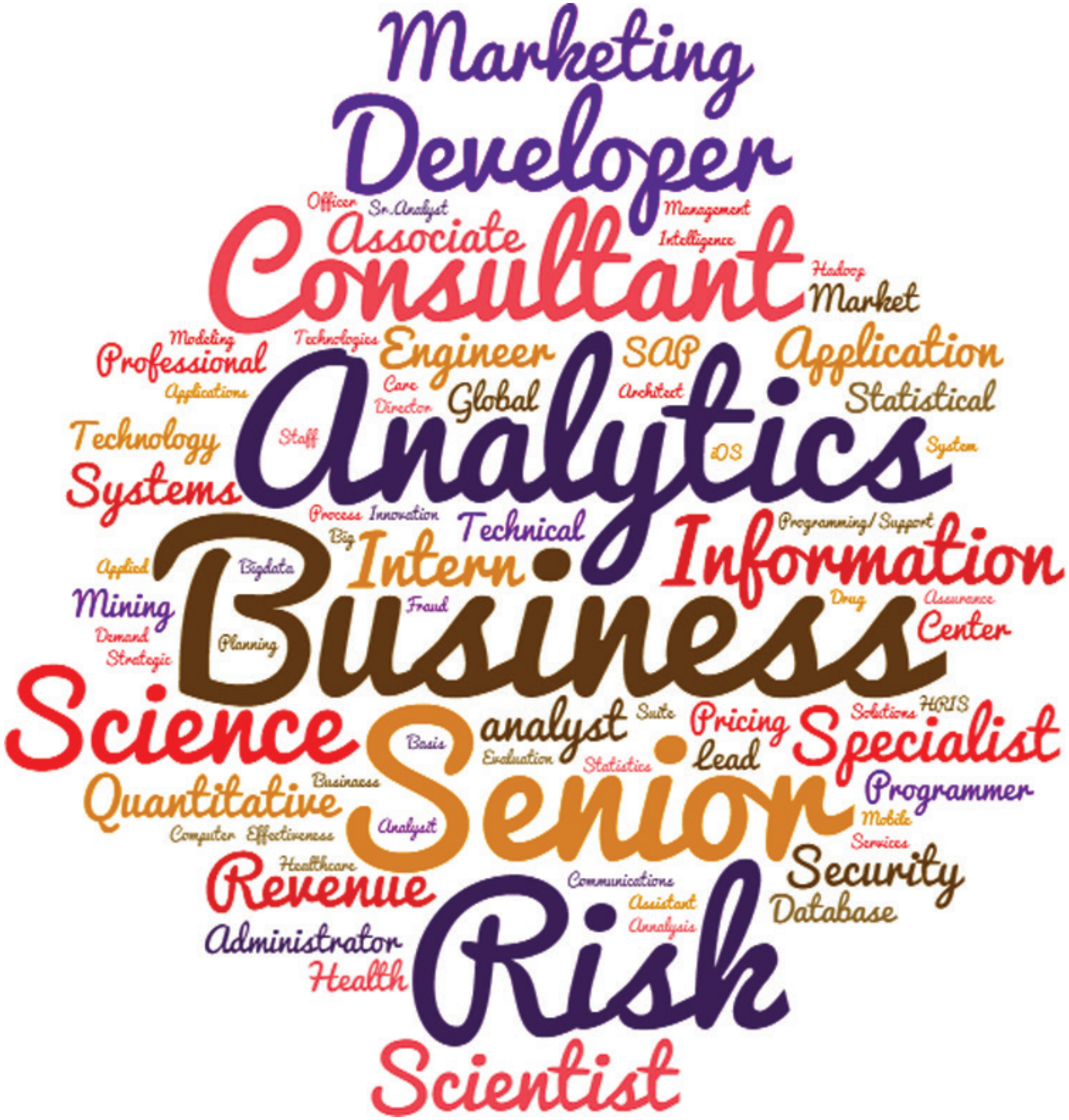
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Analytics Ecosystem


Data Generation Infrastructure Providers
Data Management Infrastructure Providers
Regulators and Policy Makers
Data Warehouse Providers
Analytics Industry Analysts & Influencers
Analytics User Organization
Analytics-Focused Software Developers
Application Developers: Industry Specific or General
Academic Institutions and Certification Agencies
Middleware Providers
Data Service Providers

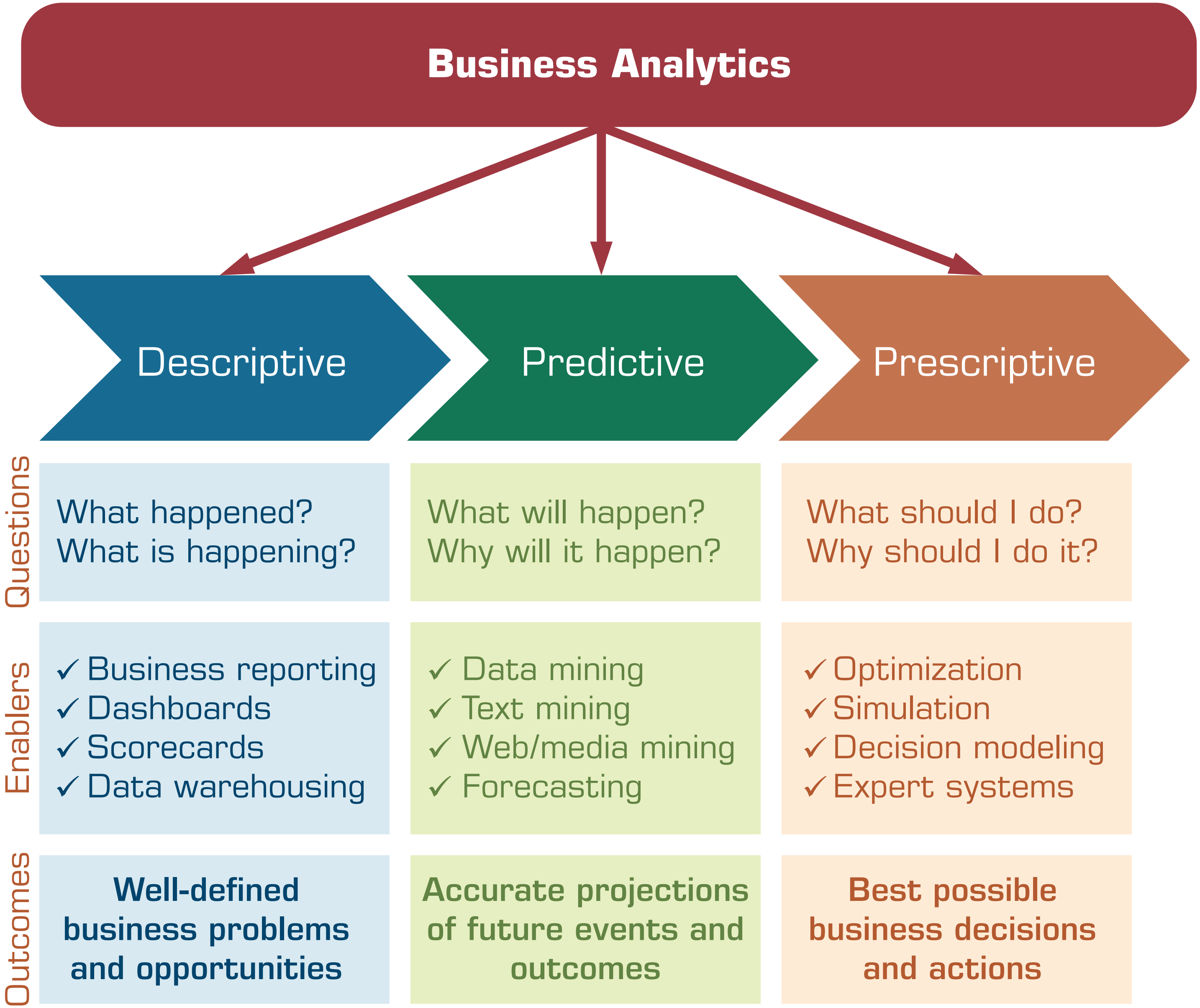
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Job Titles of Analytics



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Three Types of Analytics

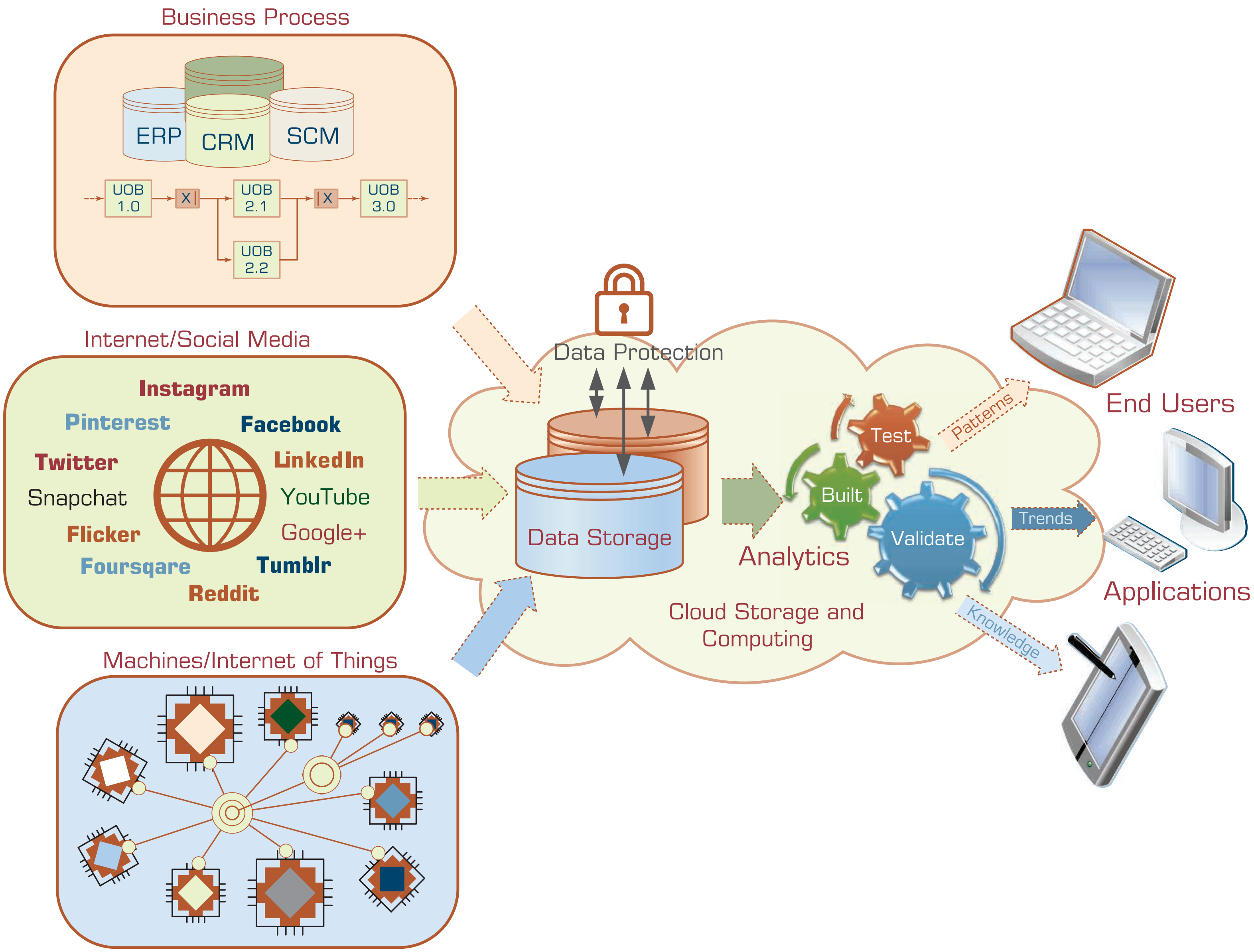


Business Analytics

	Descriptive	Predictive	Prescriptive
Questions	What happened? What is happening?	What will happen? Why will it happen?	What should I do? Why should I do it?
Enablers	✓ Business reporting ✓ Dashboards ✓ Scorecards ✓ Data warehousing	✓ Data mining ✓ Text mining ✓ Web/media mining ✓ Forecasting	✓ Optimization ✓ Simulation ✓ Decision modeling ✓ Expert systems
Outcomes	**Well-defined business problems and opportunities**	**Accurate projections of future events and outcomes**	**Best possible business decisions and actions**

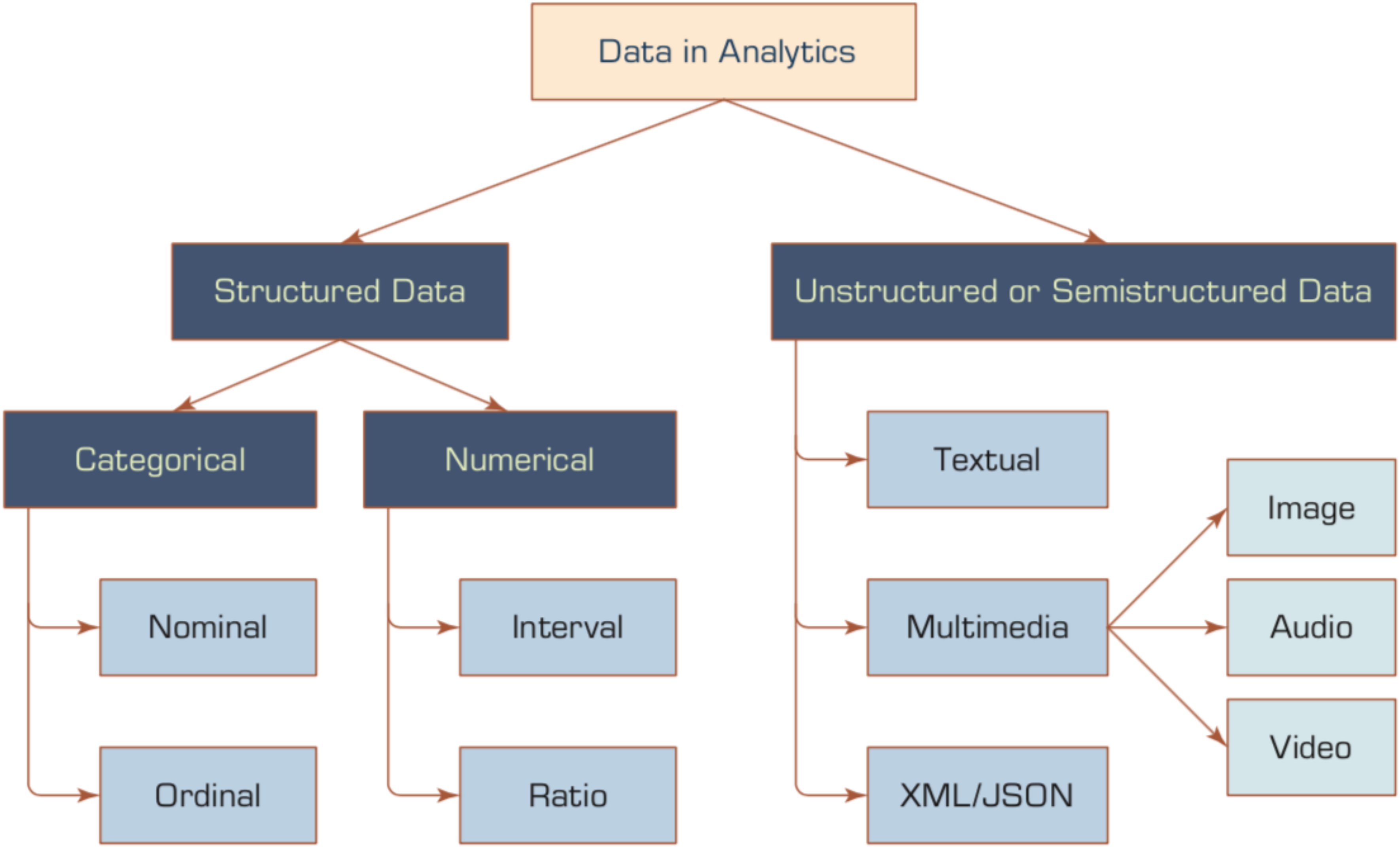
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Data to Knowledge Continuum



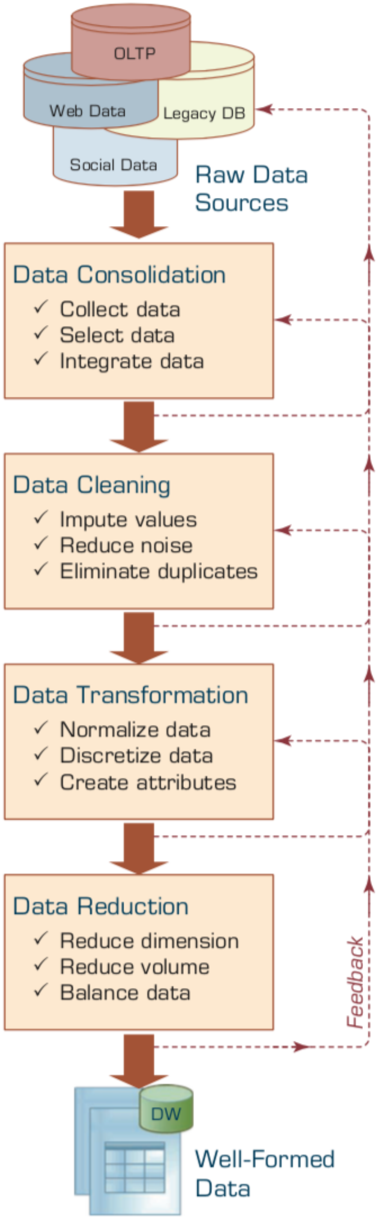
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Simple Taxonomy of Data



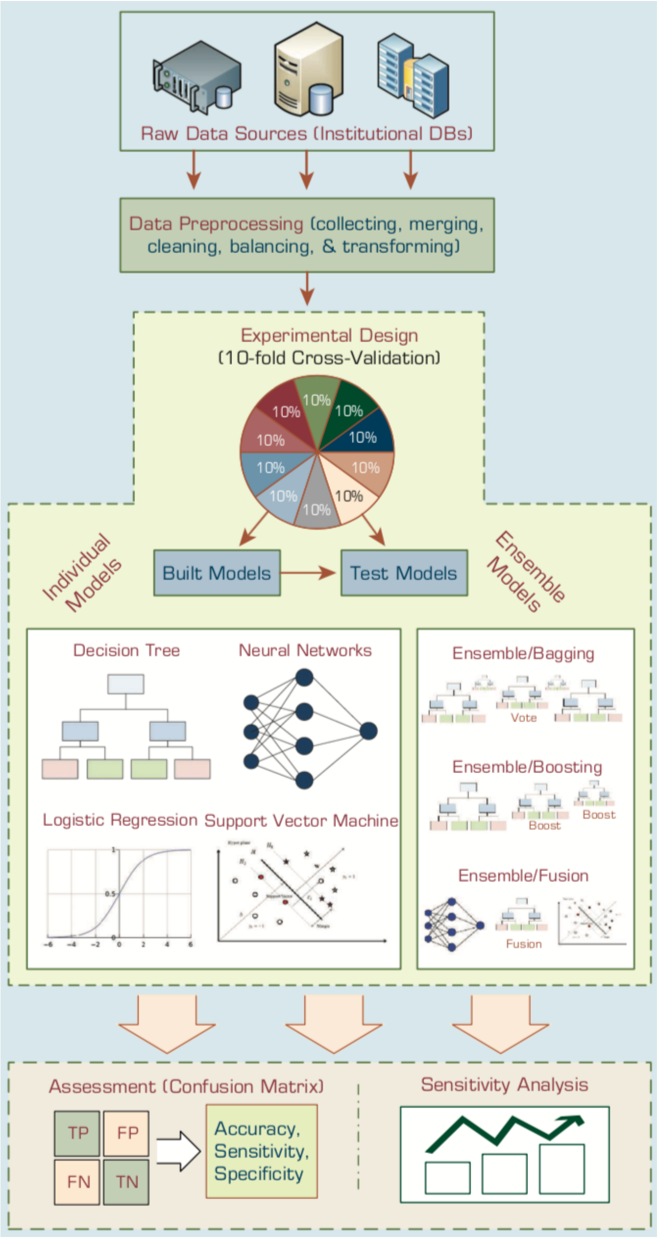
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Data Preprocessing Steps



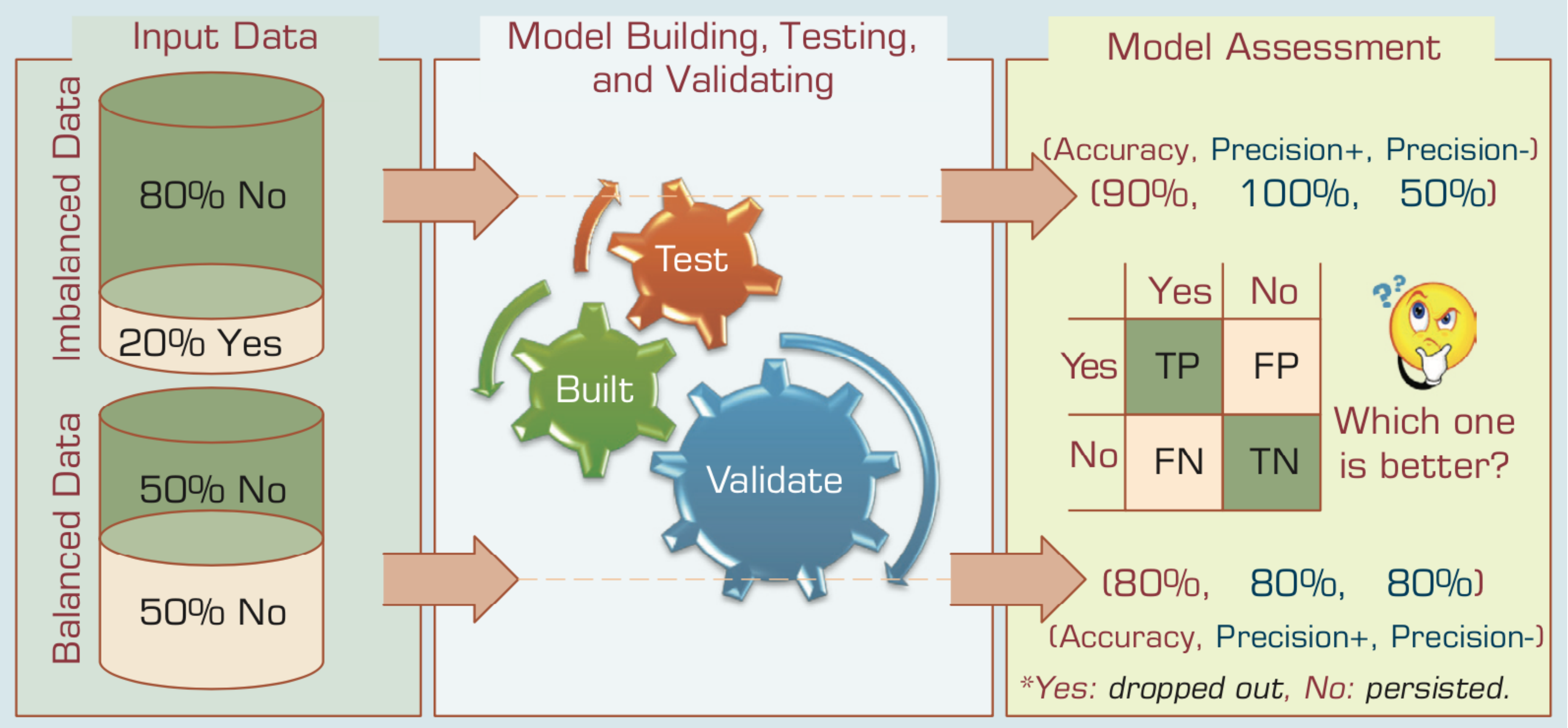
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

An Analytics Approach to Predicting Student Attrition



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

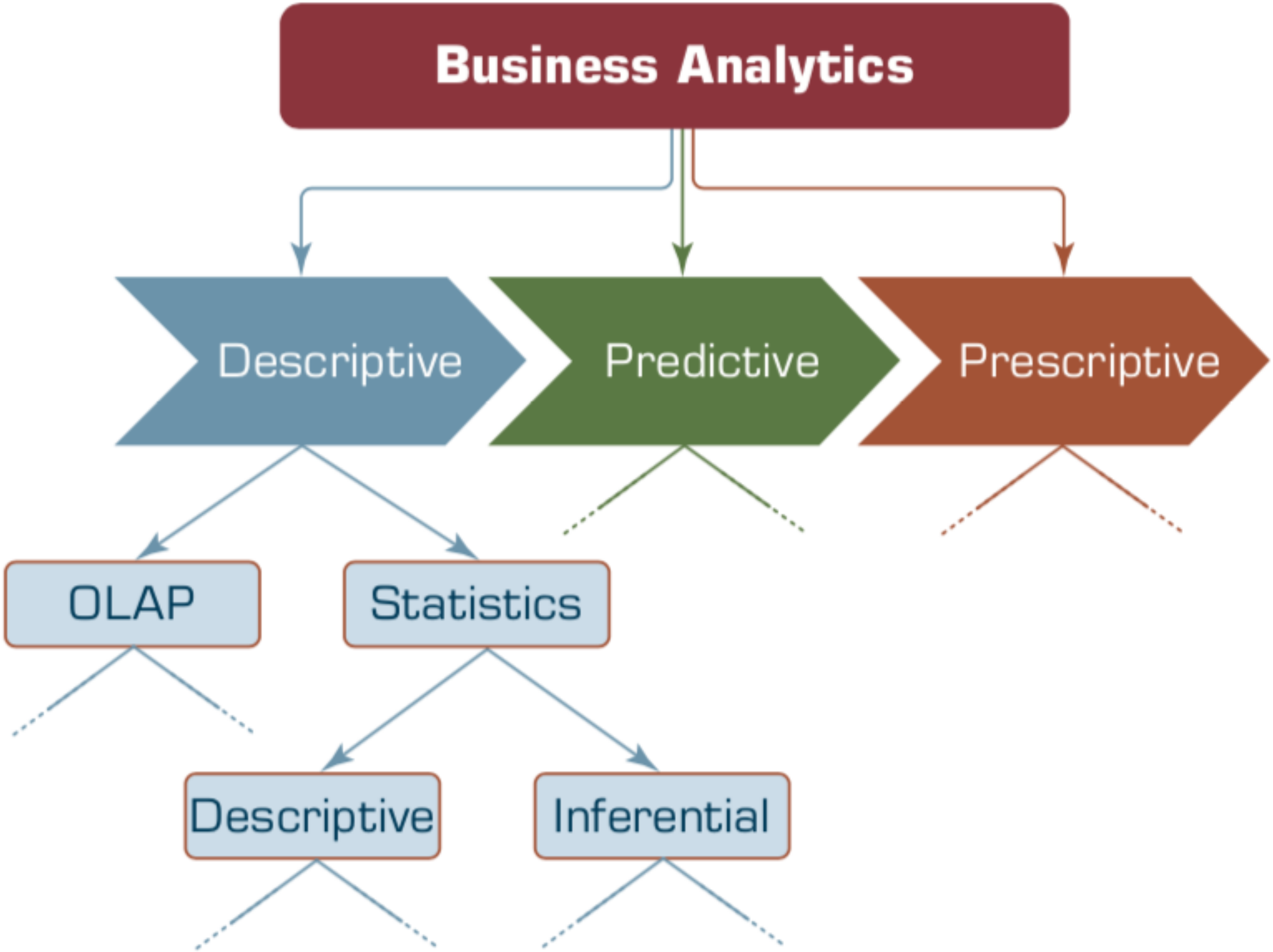
A Graphical Depiction of the Class Imbalance Problem



	Yes	No
Yes	TP	FP
No	FN	TN

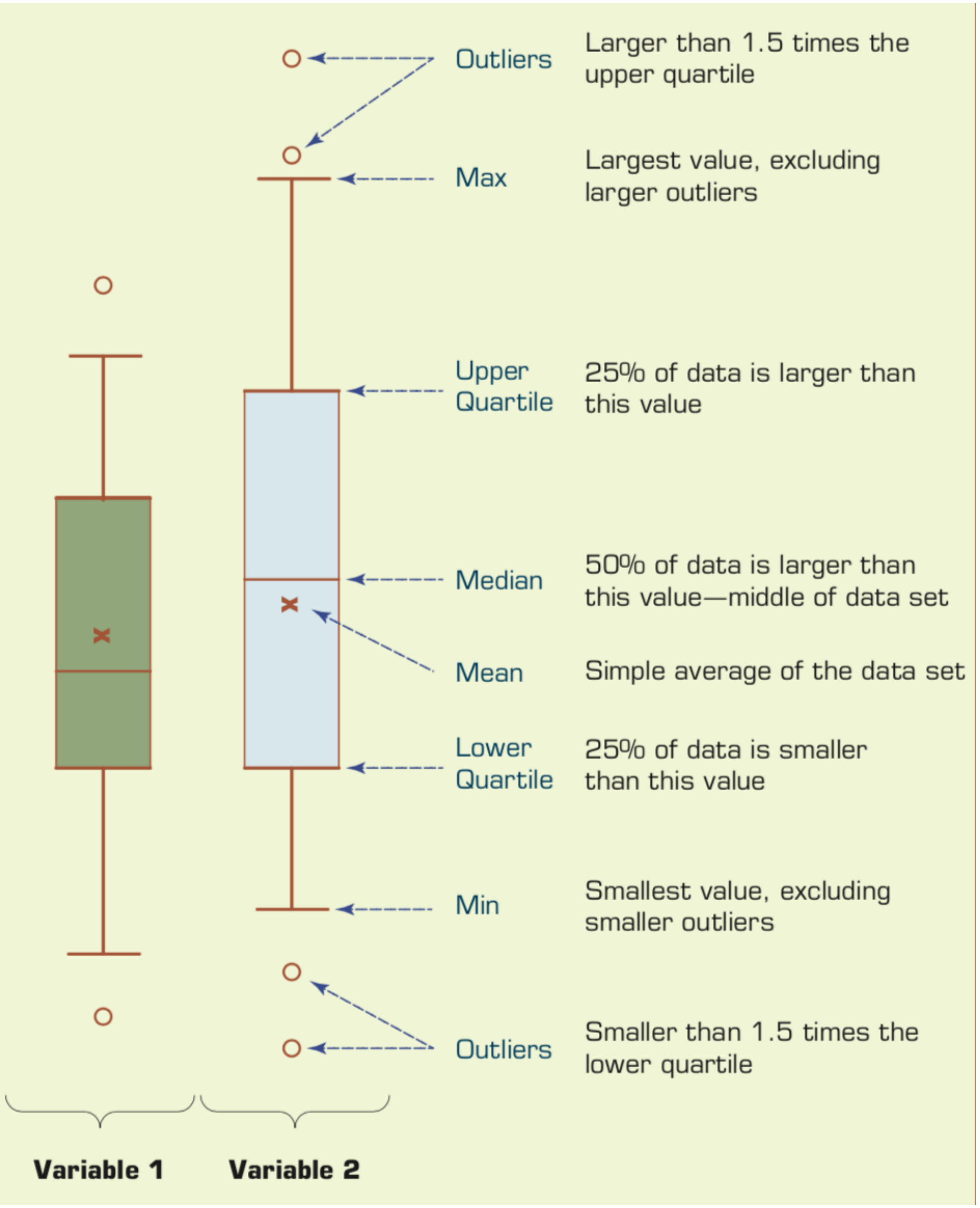
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Relationship between Statistics and Descriptive Analytics



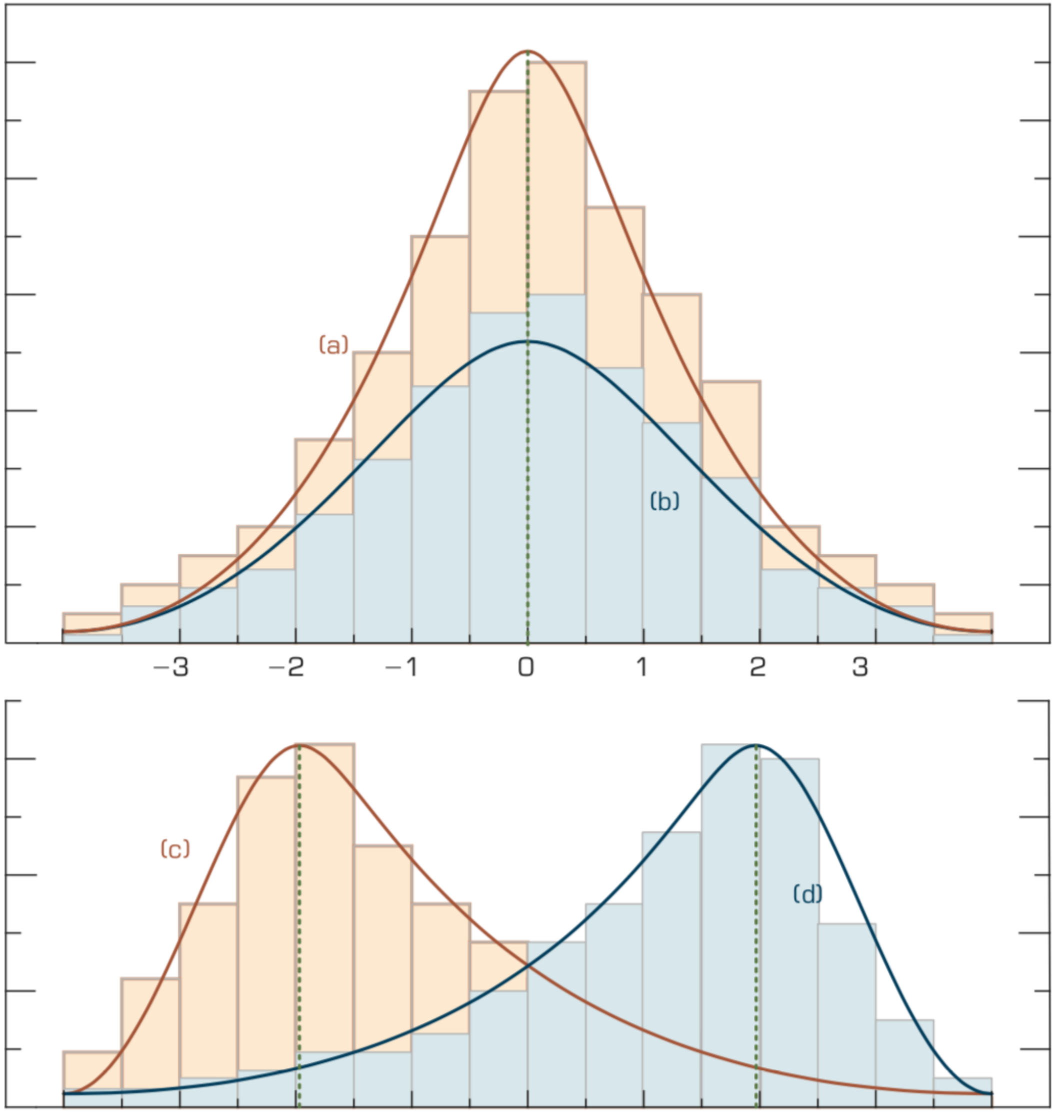
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Understanding the Specifics about Box-and-Whiskers Plots


Outliers
Larger than 1.5 times the upper quartile
Max
Largest value, excluding larger outliers
Upper Quartile
25% of data is larger than this value
Median
50% of data is larger than this value—middle of data set
Mean
Simple average of the data set
Lower Quartile
25% of data is smaller than this value
Min
Smallest value, excluding smaller outliers
Outliers
Smaller than 1.5 times the lower quartile
Variable 1
Variable 2

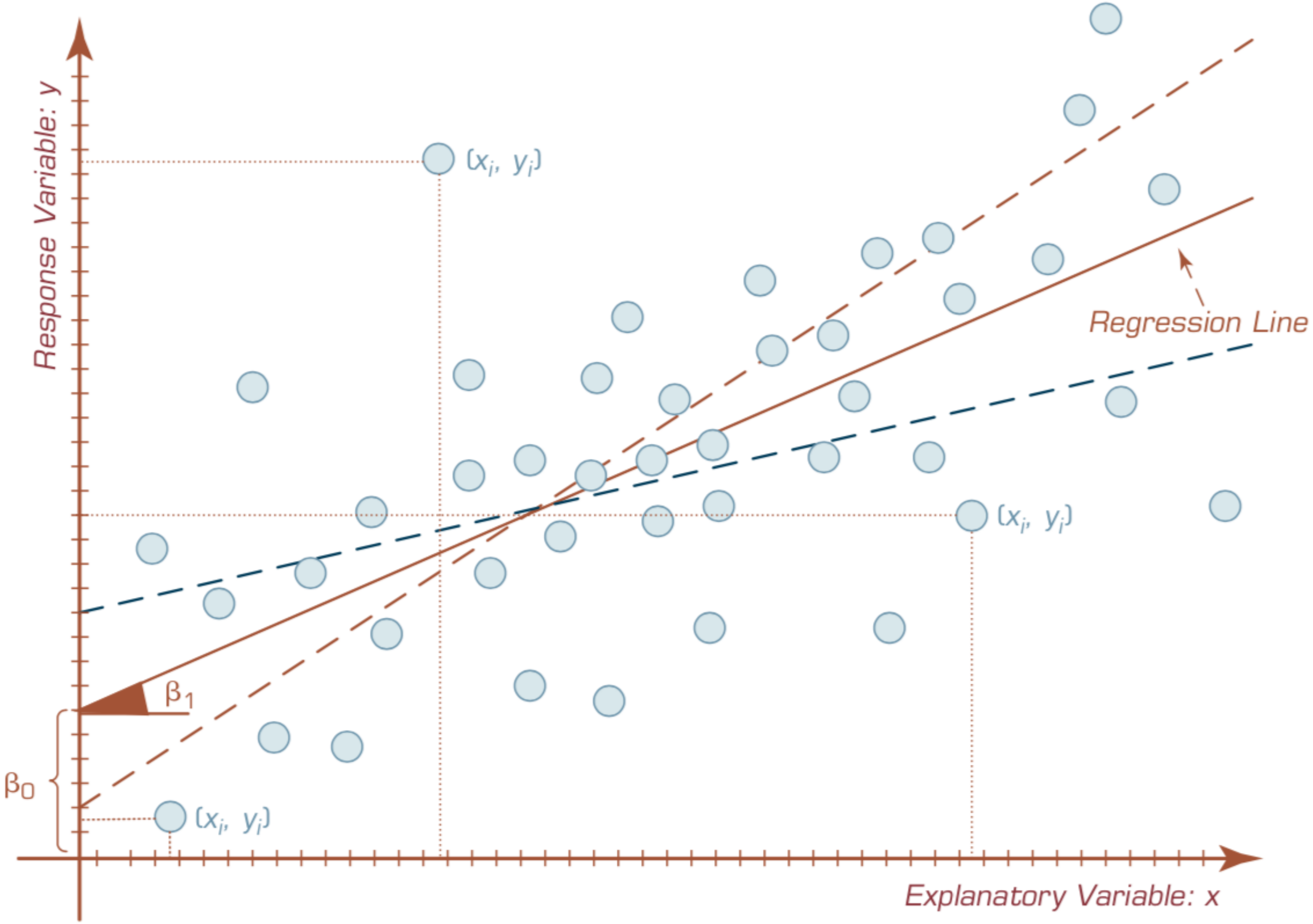
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Relationship between Dispersion and Shape Properties.



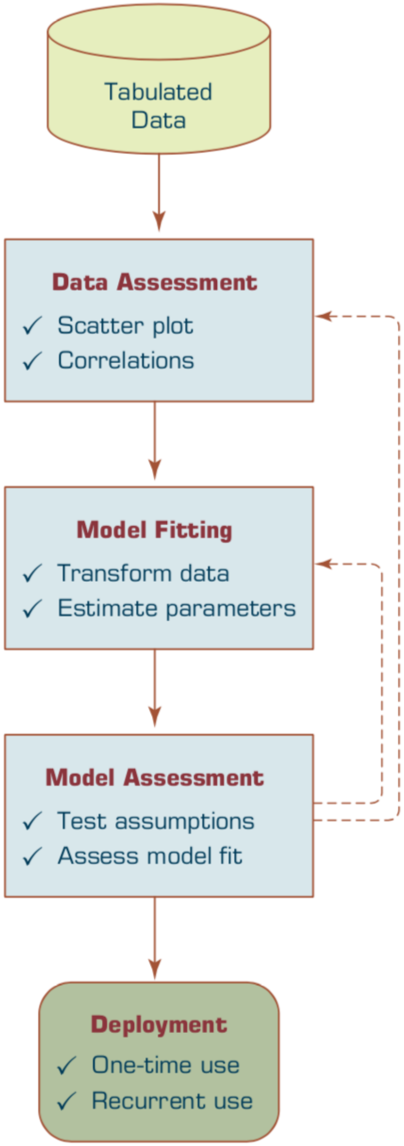
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Scatter Plot and a Linear Regression Line



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

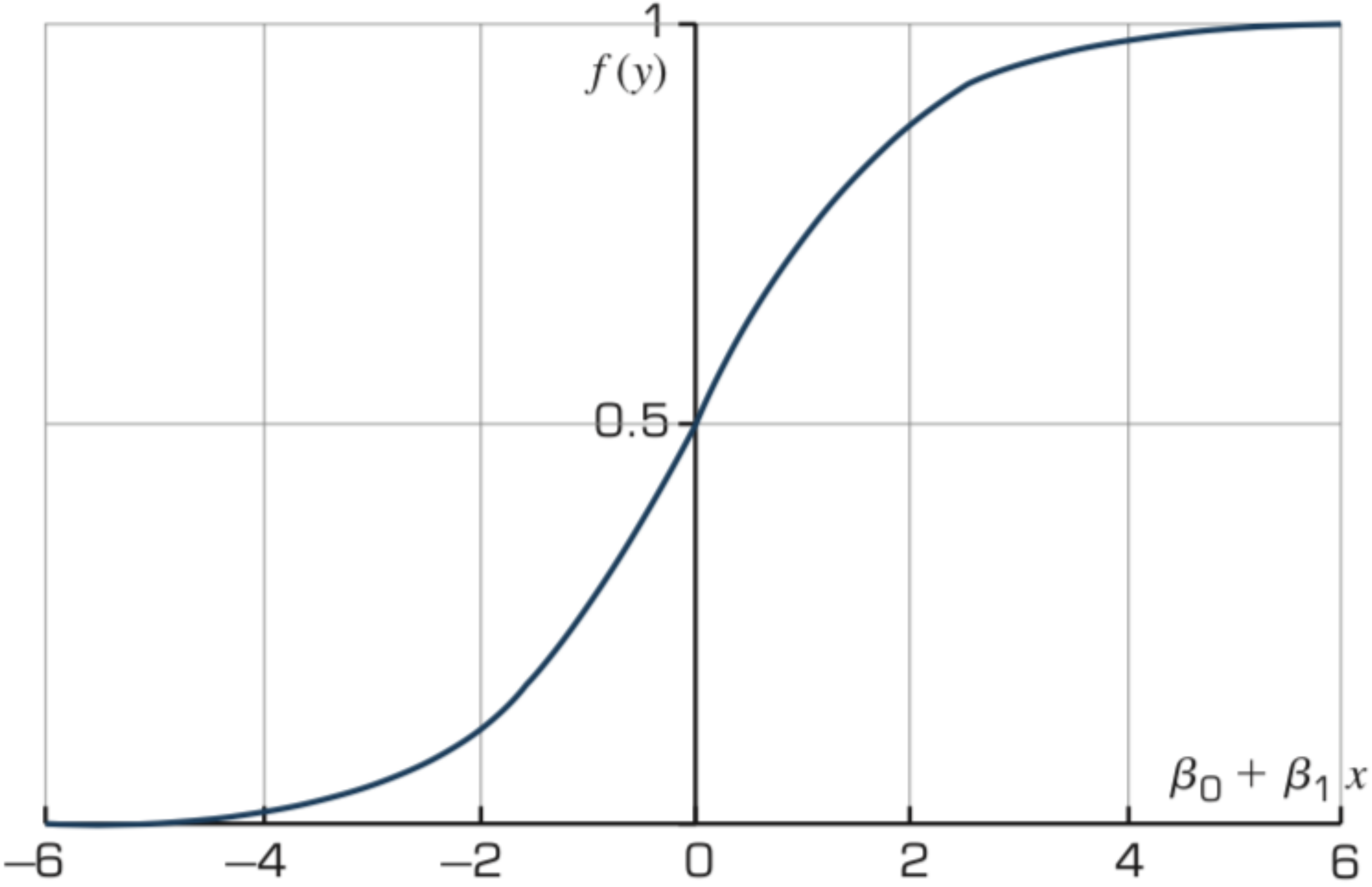
A Process Flow for Developing Regression Models.



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

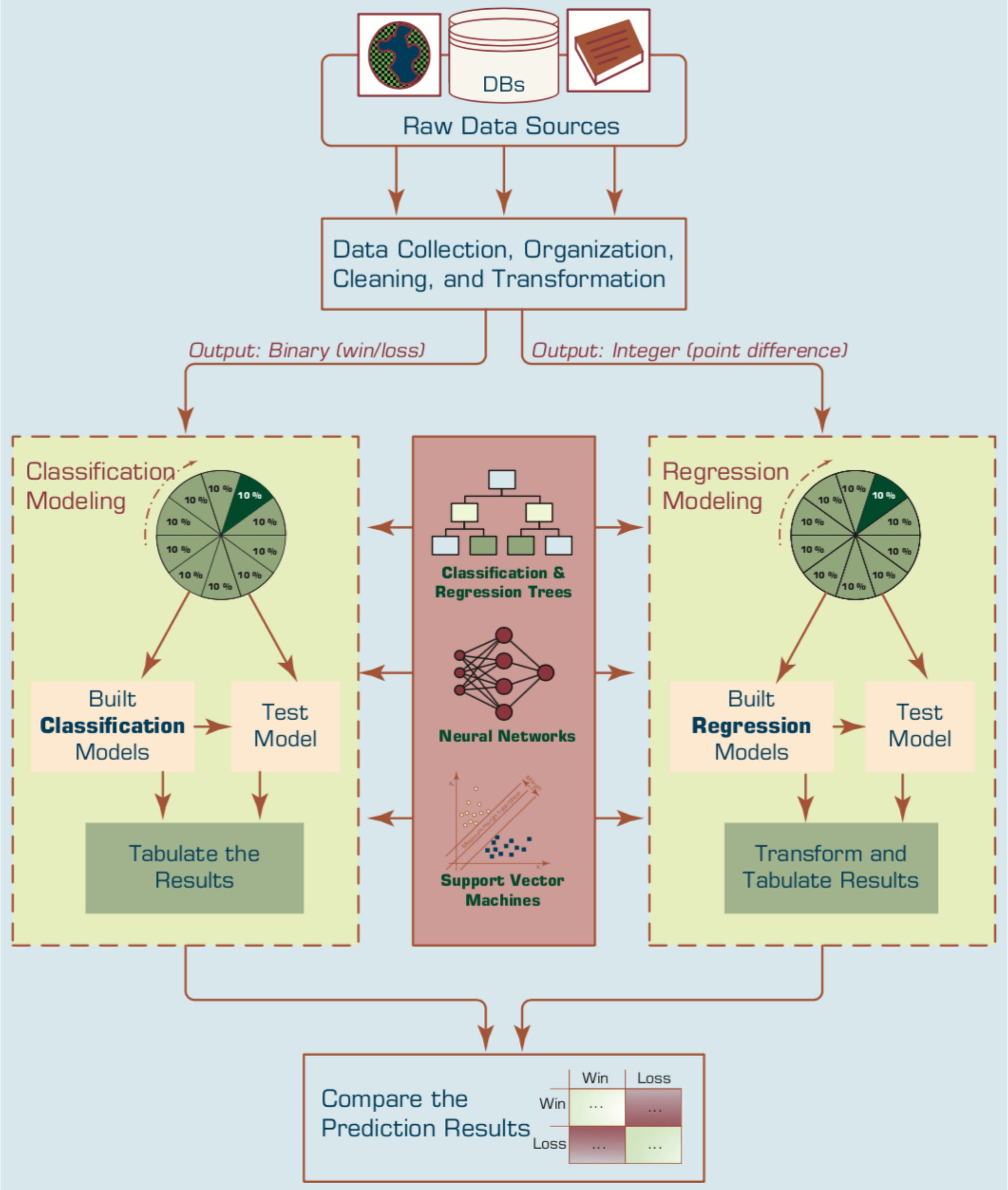
The Logistic Function

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



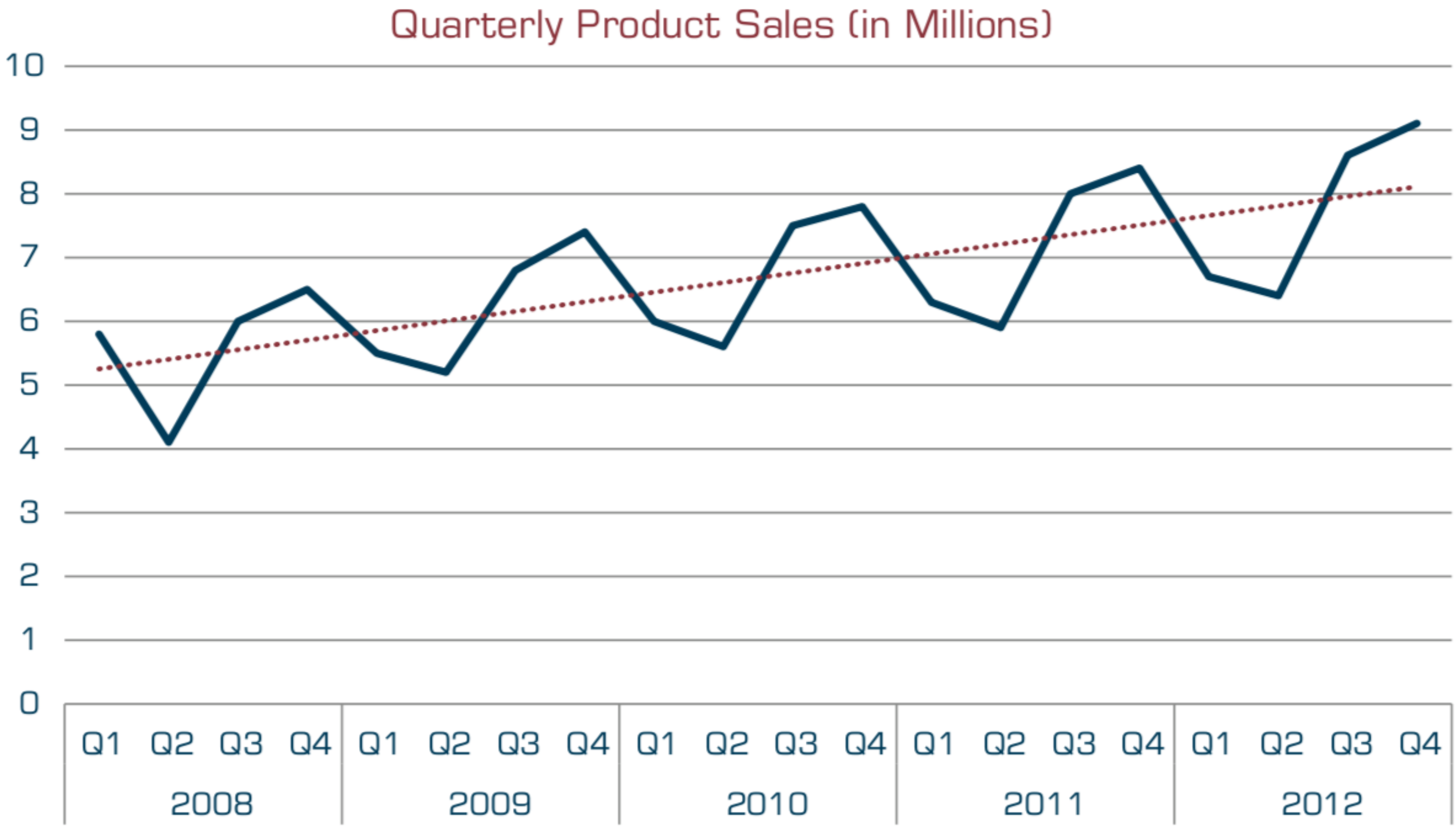
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Predicting NCAA Bowl Game Outcomes


DBs
Raw Data Sources
Data Collection, Organization, Cleaning, and Transformation
Output: Binary (win/loss)
Output: Integer (point difference)
Classification Modeling
10 %
Built Classification Models
Test Model
Tabulate the Results
Classification & Regression Trees
Neural Networks
Support Vector Machines
Regression Modeling
10 %
Built Regression Models
Test Model
Transform and Tabulate Results
Compare the Prediction Results
Win
Loss
...

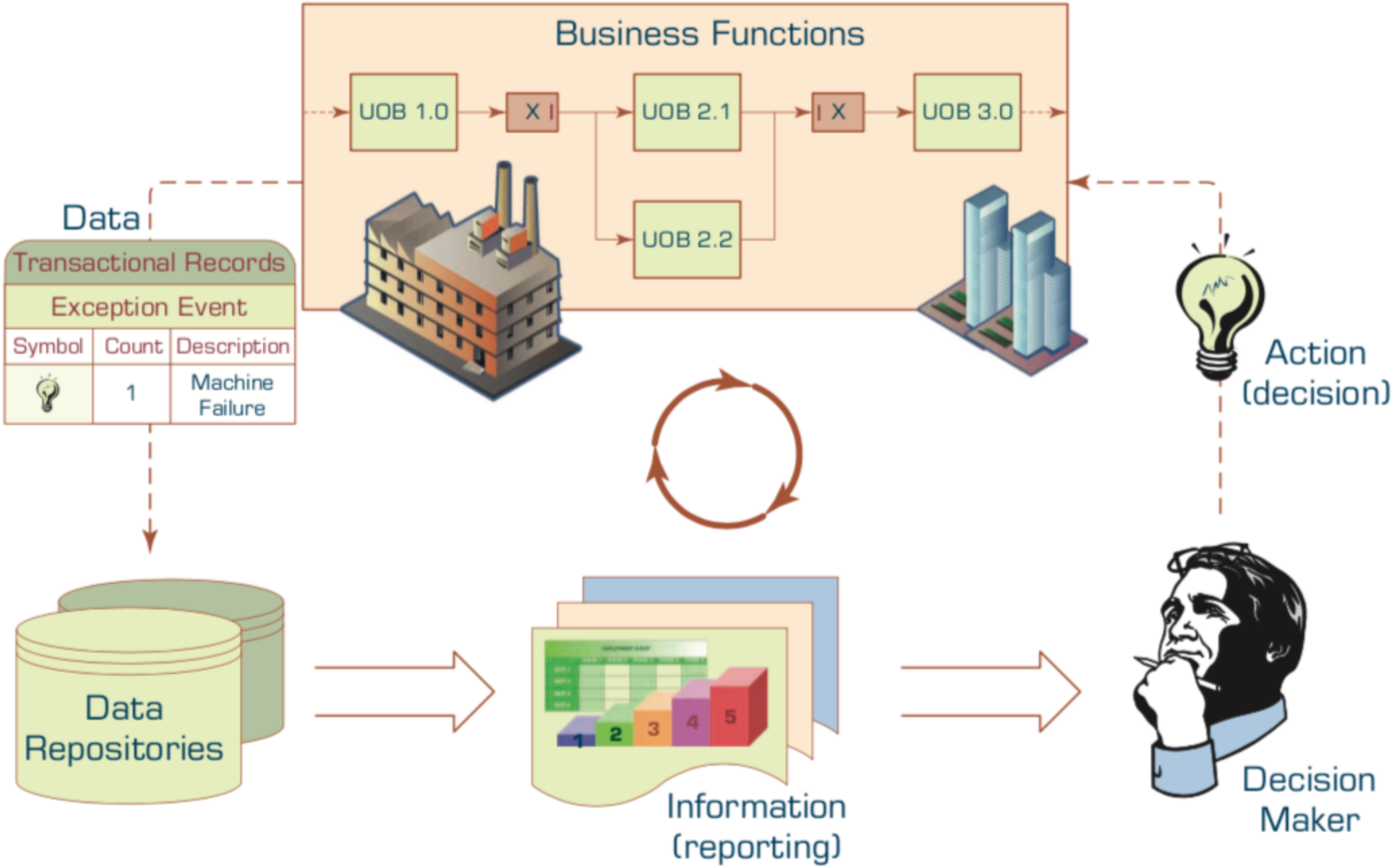
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Sample Time Series of Data on Quarterly Sales Volumes



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

The Role of Information Reporting in Managerial Decision Making



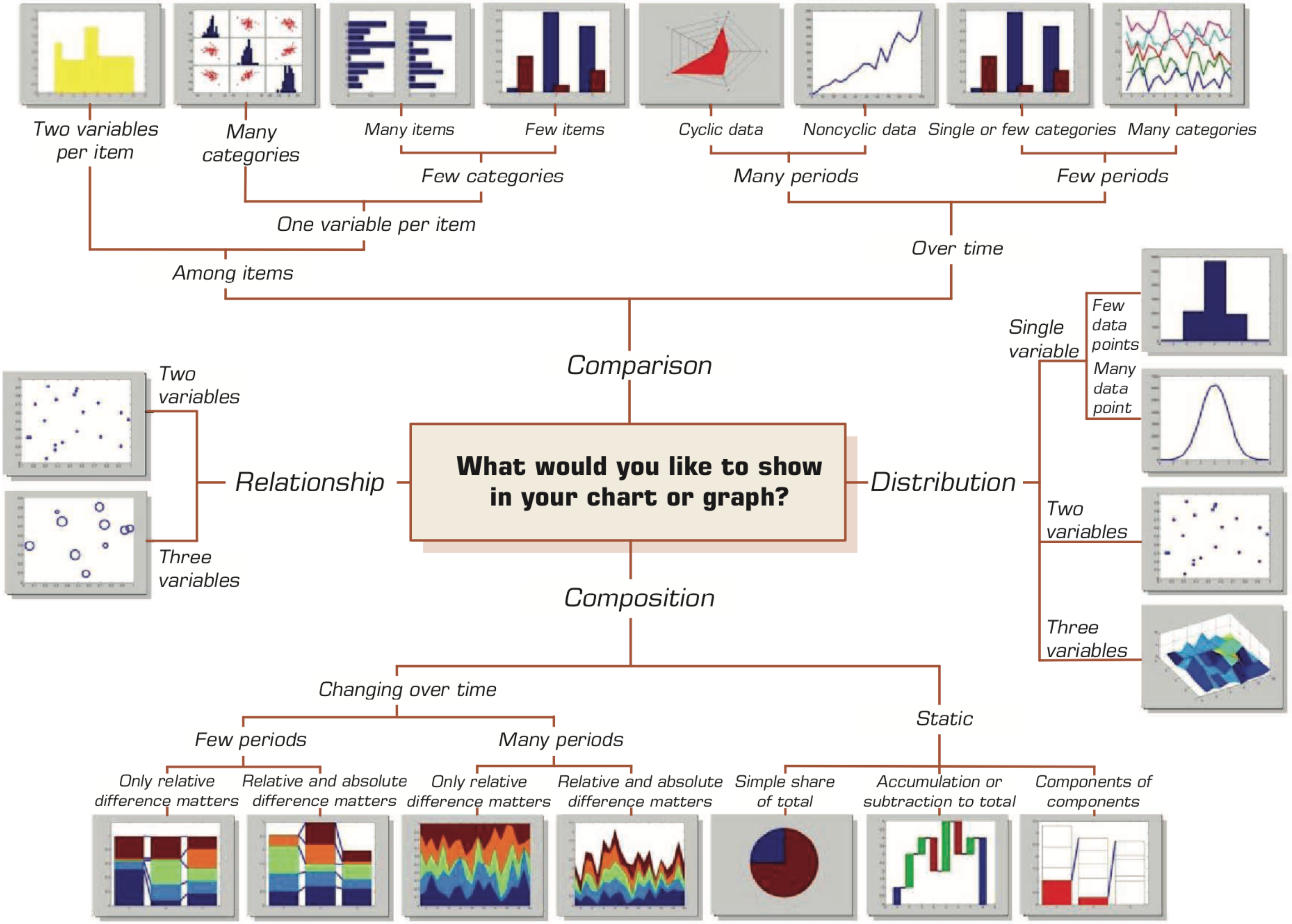
Transactional Records

Exception Event

Symbol	Count	Description
💡	1	Machine Failure

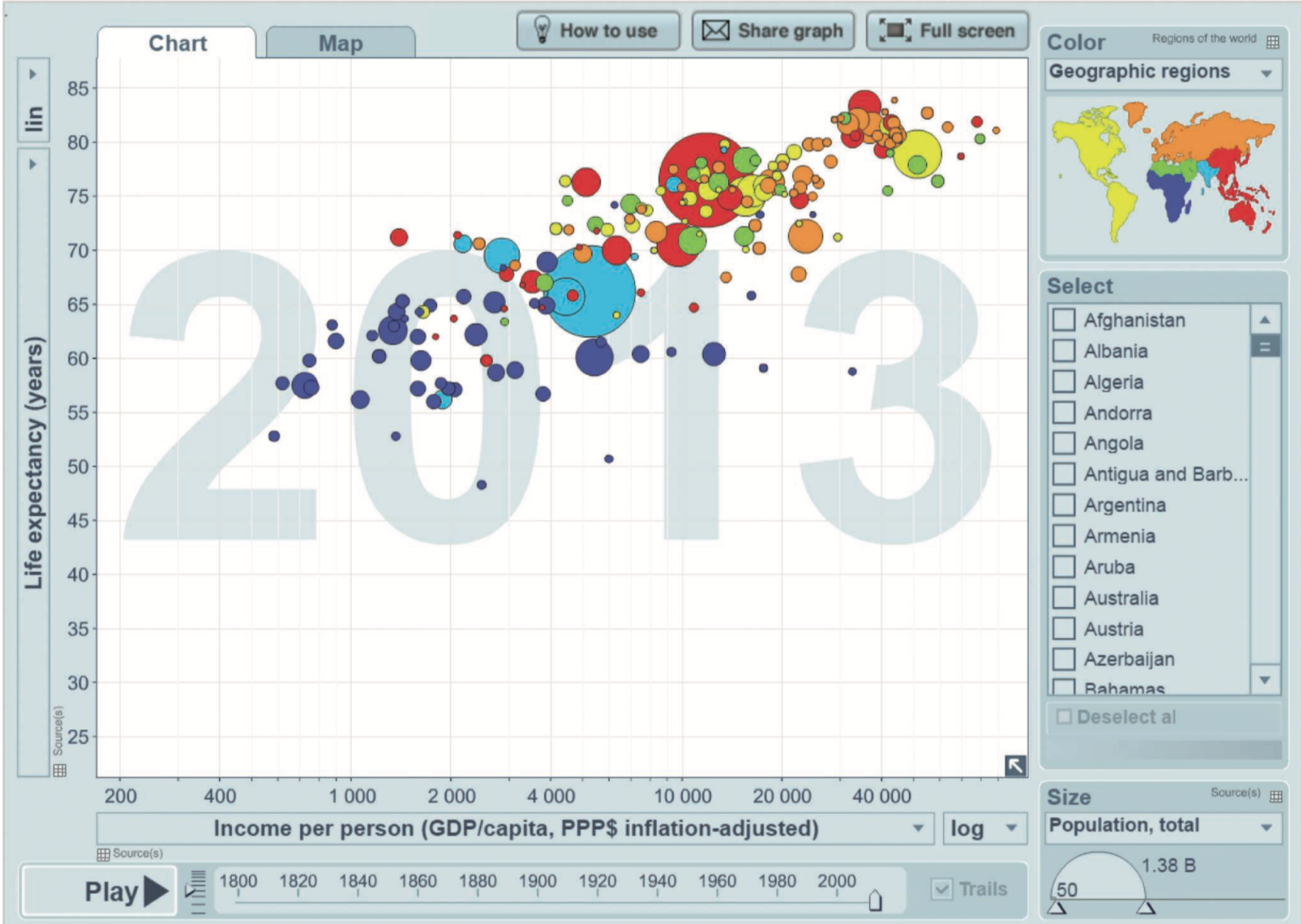
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Taxonomy of Charts and Graphs



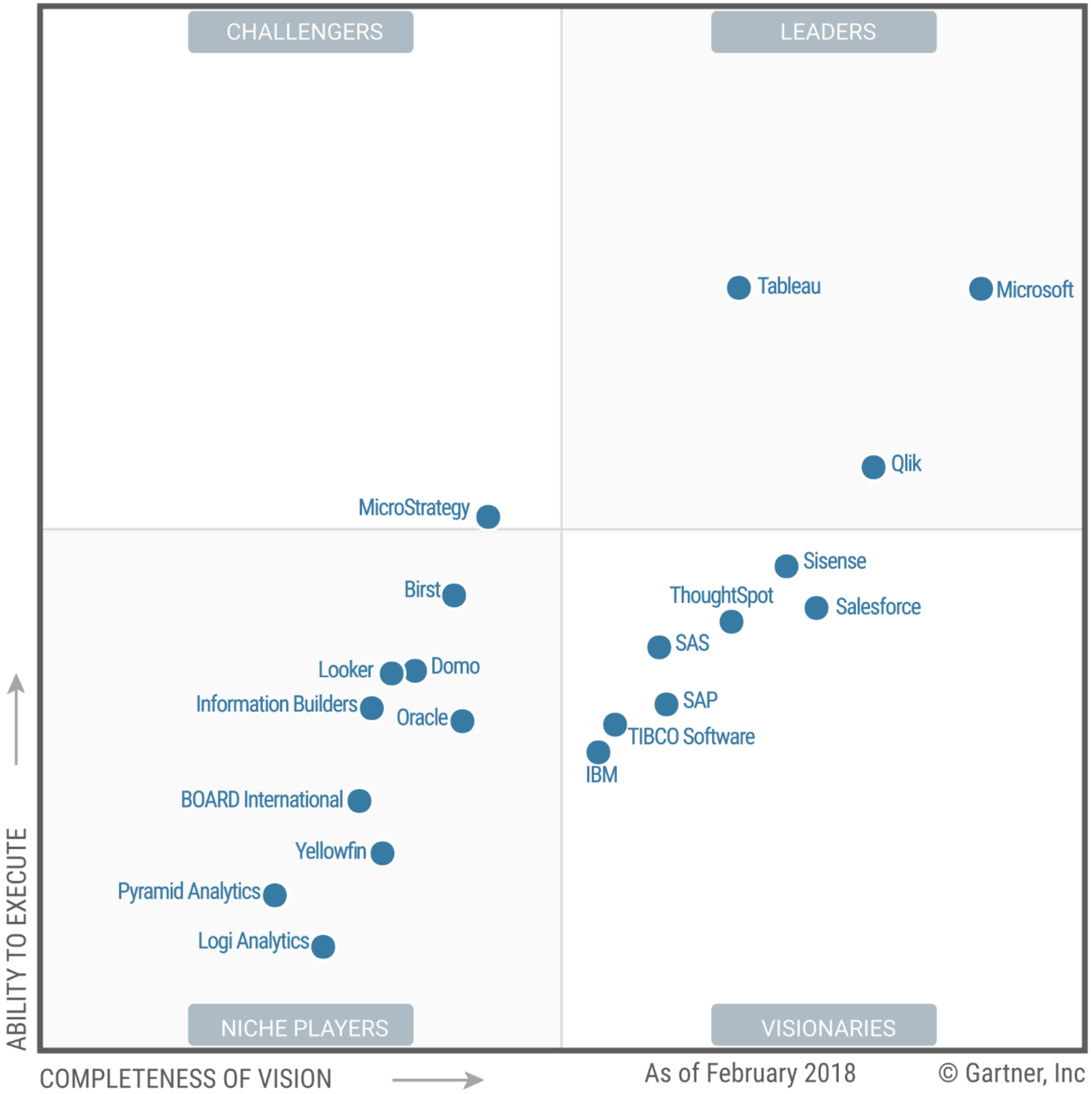
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Gapminder Chart That Shows the Wealth and Health of Nations



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Magic Quadrant for Business Intelligence and Analytics Platforms

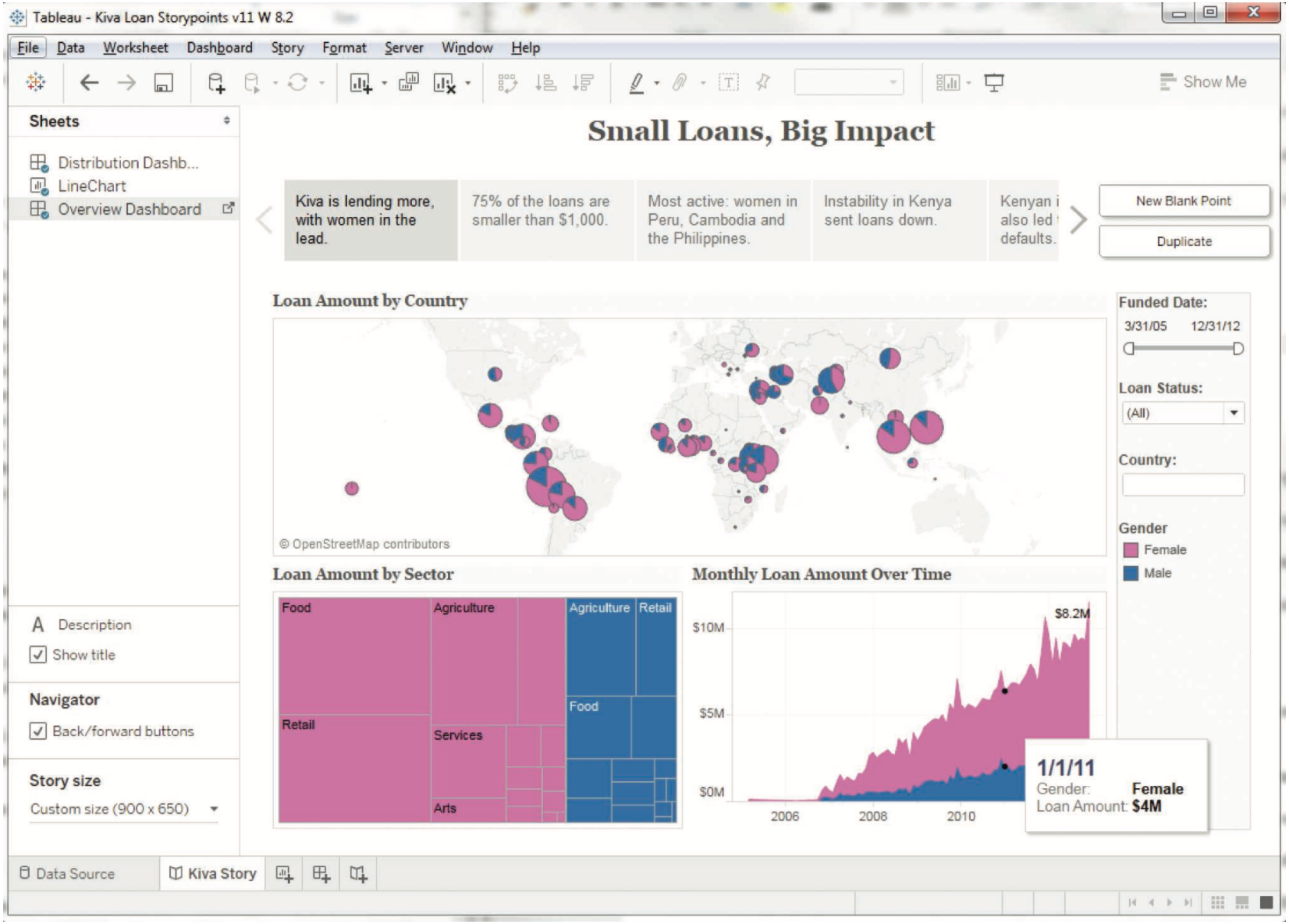


As of February 2018

© Gartner, Inc

Source: https://www.tableau.com/reports/gartner

A Storyline Visualization in Tableau Software



Small Loans, Big Impact

- Kiva is lending more, with women in the lead.
- 75% of the loans are smaller than $1,000.
- Most active: women in Peru, Cambodia and the Philippines.
- Instability in Kenya sent loans down.
- Kenyan i also led defaults.

Funded Date: 3/31/05 12/31/12

Loan Status: (All)

Country:

Gender
- 🟥 Female
- 🟦 Male

1/1/11
Gender: Female
Loan Amount: $4M

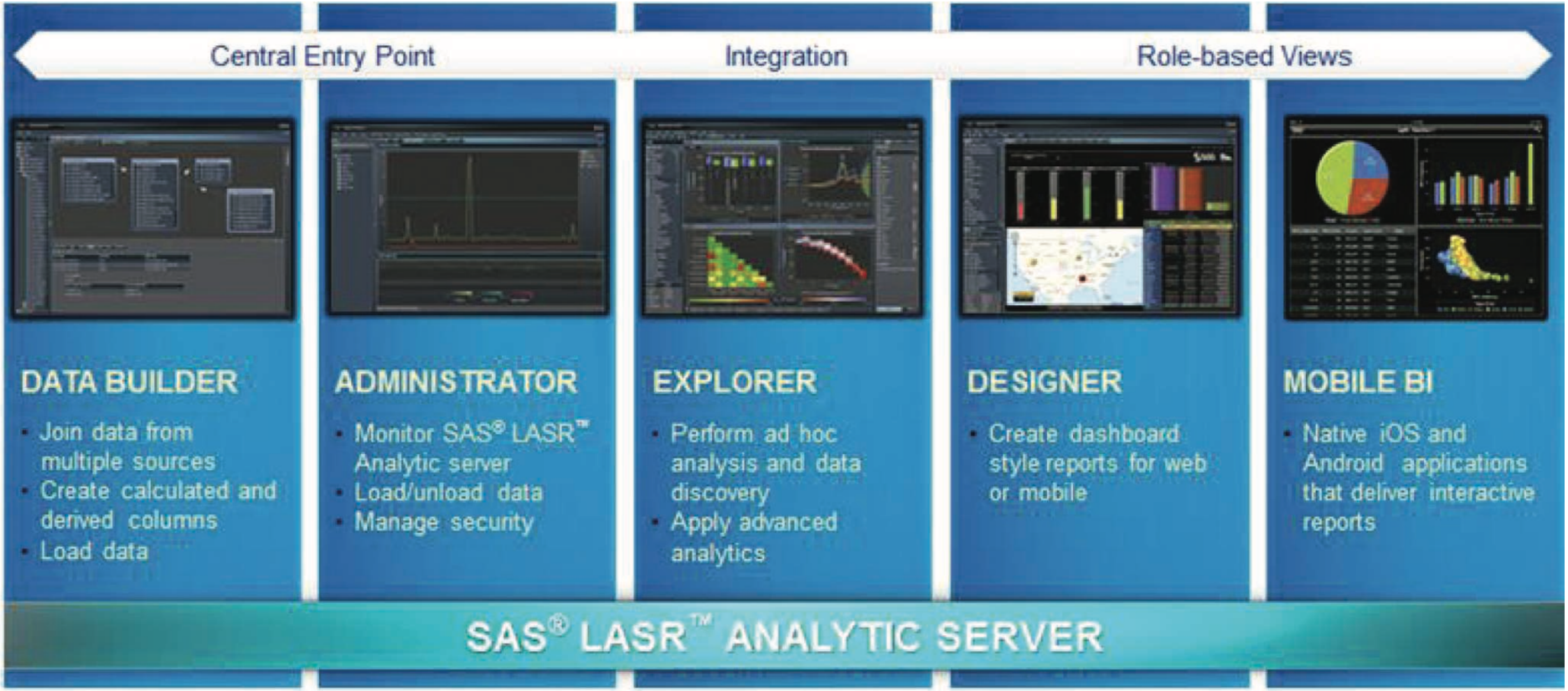
☑ Show title

☑ Back/forward buttons

Story size: Custom size (900 x 650)

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

An Overview of SAS Visual Analytics Architecture



Central Entry Point

Integration

Role-based Views

DATA BUILDER

- Join data from multiple sources
- Create calculated and derived columns
- Load data

ADMINISTRATOR

- Monitor SAS® LASR™ Analytic server
- Load/unload data
- Manage security

EXPLORER

- Perform ad hoc analysis and data discovery
- Apply advanced analytics

DESIGNER

- Create dashboard style reports for web or mobile

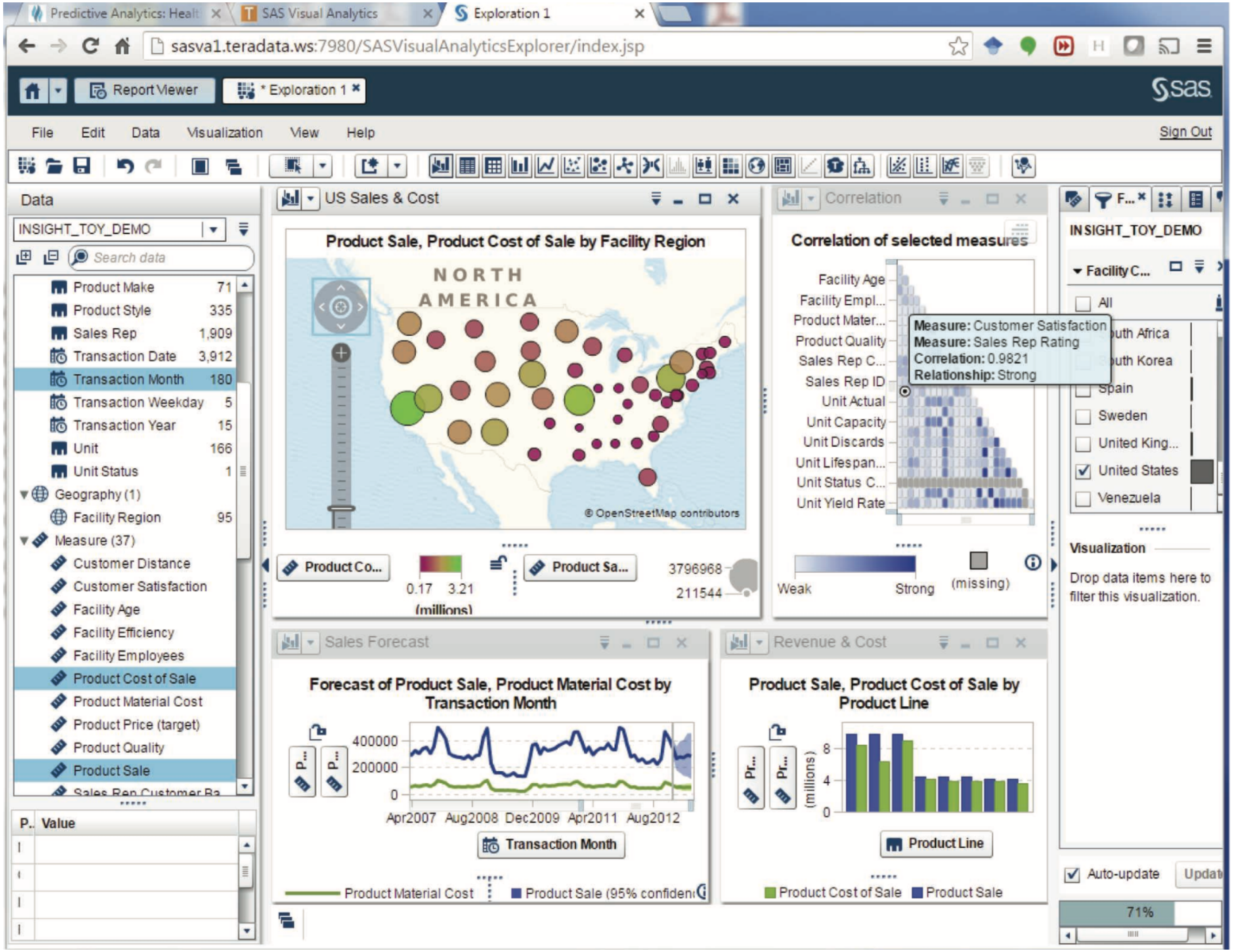
MOBILE BI

- Native iOS and Android applications that deliver interactive reports

SAS® LASR™ ANALYTIC SERVER

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

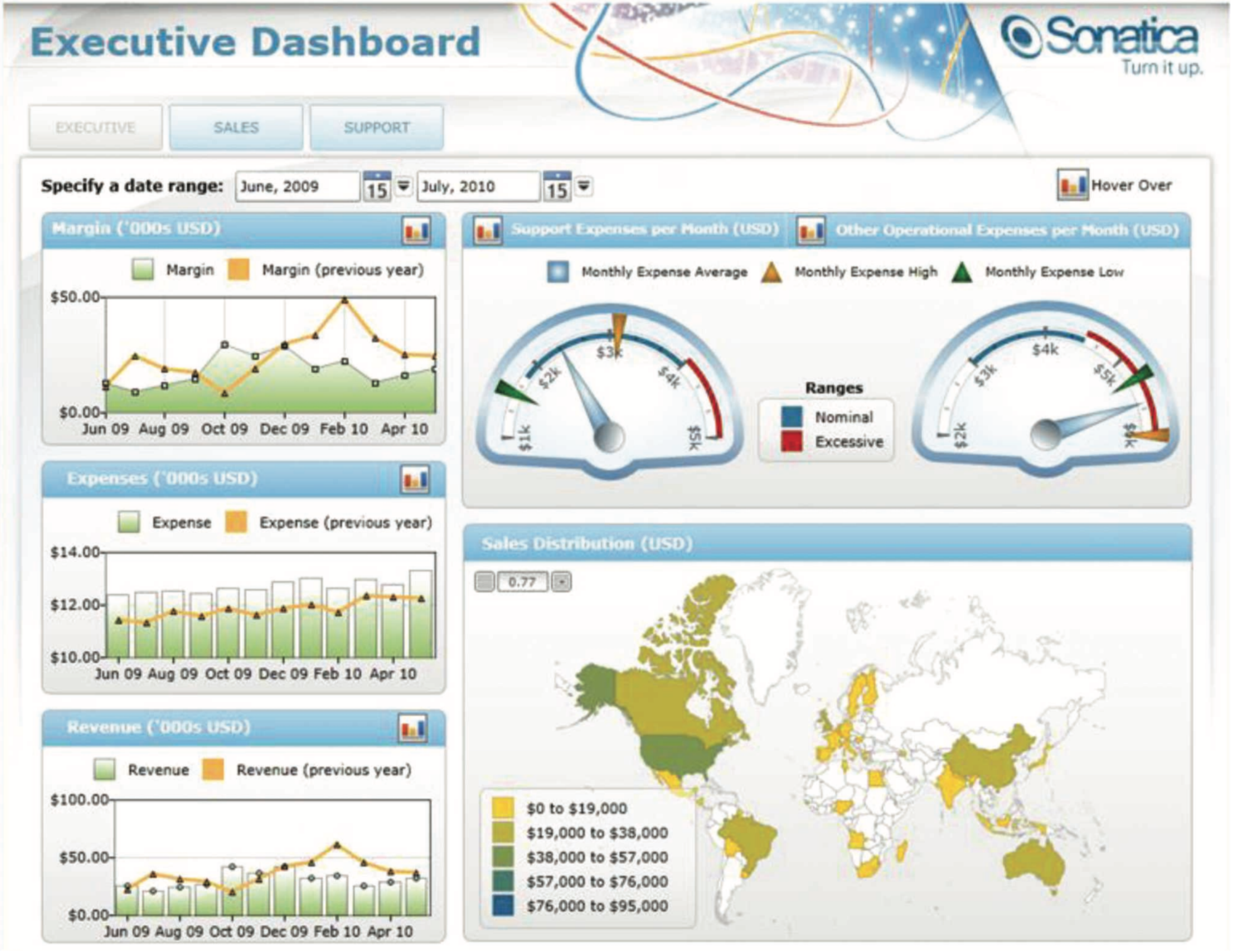
A Screenshot from SAS Visual Analytics



Measure: Customer Satisfaction
Measure: Sales Rep Rating
Correlation: 0.9821
Relationship: Strong

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

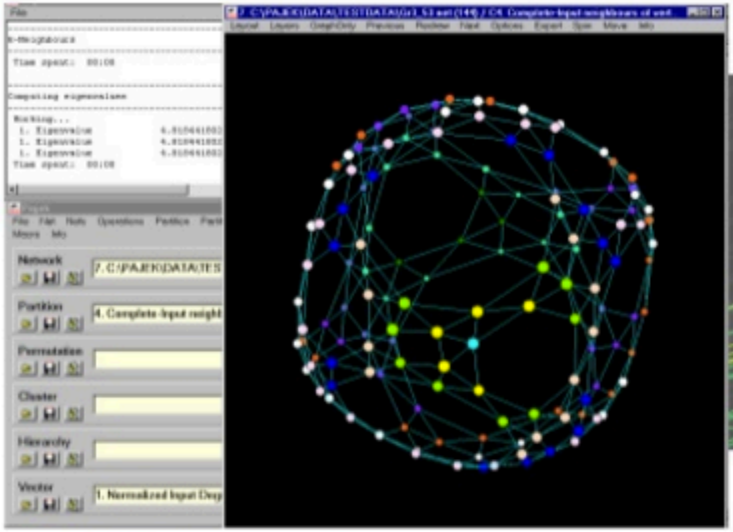
A Sample Executive Dashboard



Specify a date range: June, 2009 – July, 2010

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Exploratory Network Analysis



```
k-Neighbours
Time spent:  00:08
Computing eigenvalues
Working...
 1. Eigenvalue
 1. Eigenvalue
 1. Eigenvalue
Time spent:  00:08
```

Network: 7. C:\PAJEK\DATA\TES

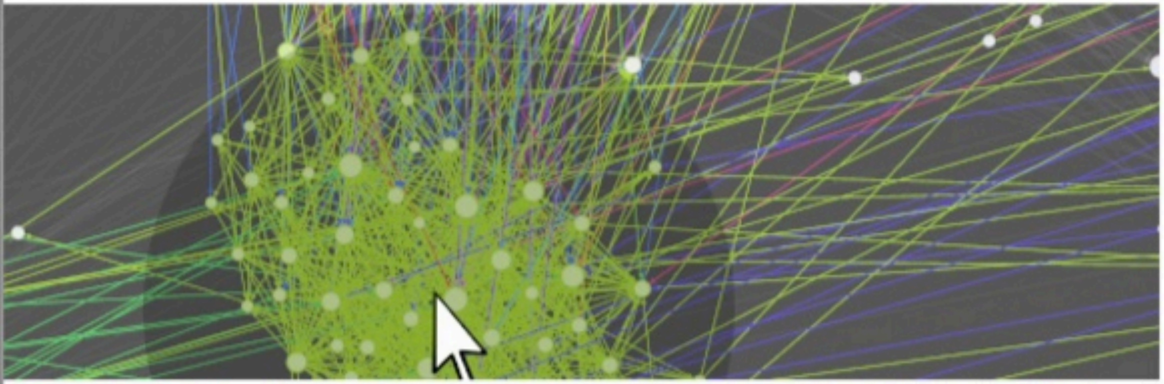
Partition: 4. Complete-Input neight

Permutation:

Cluster:

Hierarchy:

Vector: 1. Normalized Input Dep



1 **see the network**

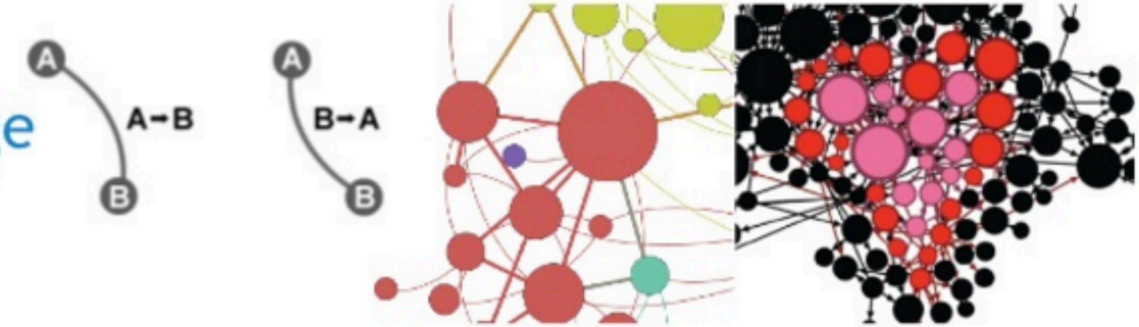
1st graph viz tool: Pajek (1996)
Vladimir Batagelj, Andrej Mrvar

2 **interact in real time**

Gephi prototype (2008)
group, filter, compute metrics…

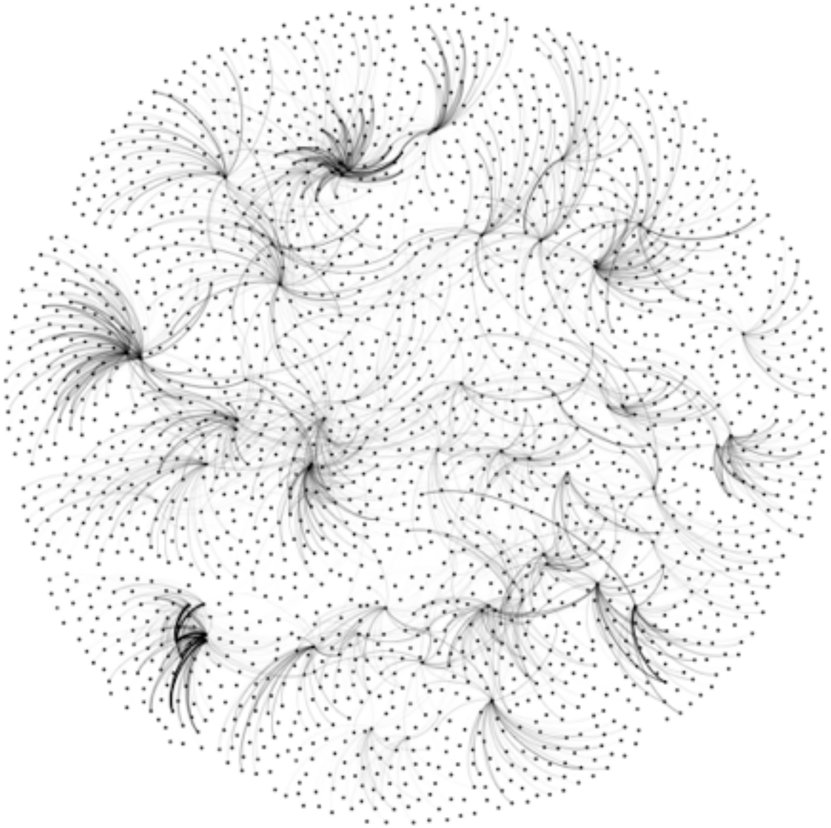
3 **build a visual language**

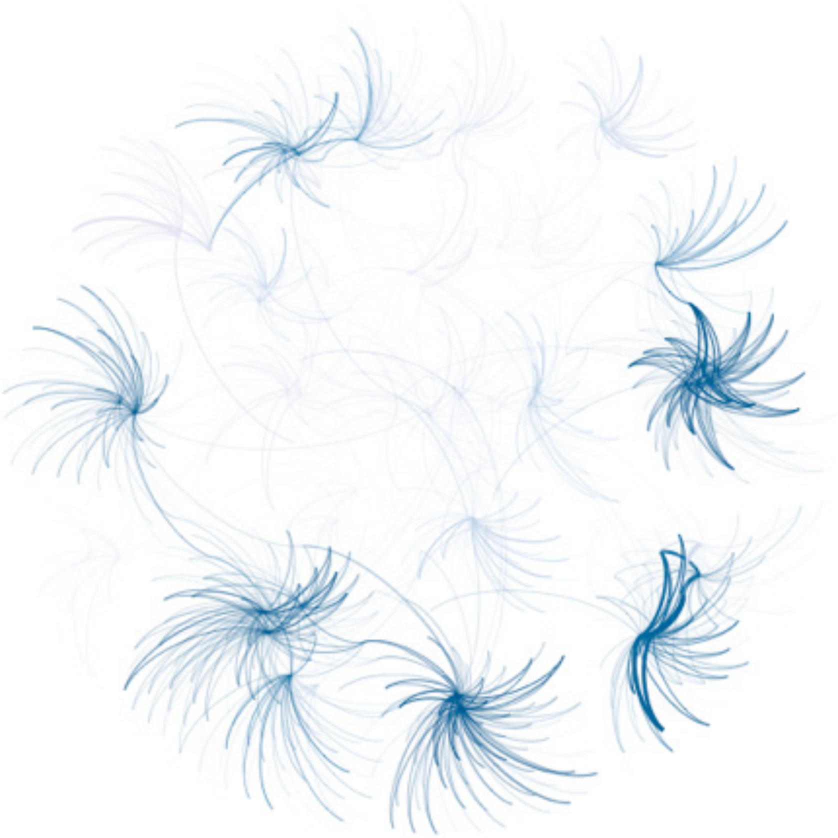
size by rank, color by partition,
label, curved edges, thickness…



Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf

Looking for a "Simple Small Truth"? What Data Visualization Should Do?

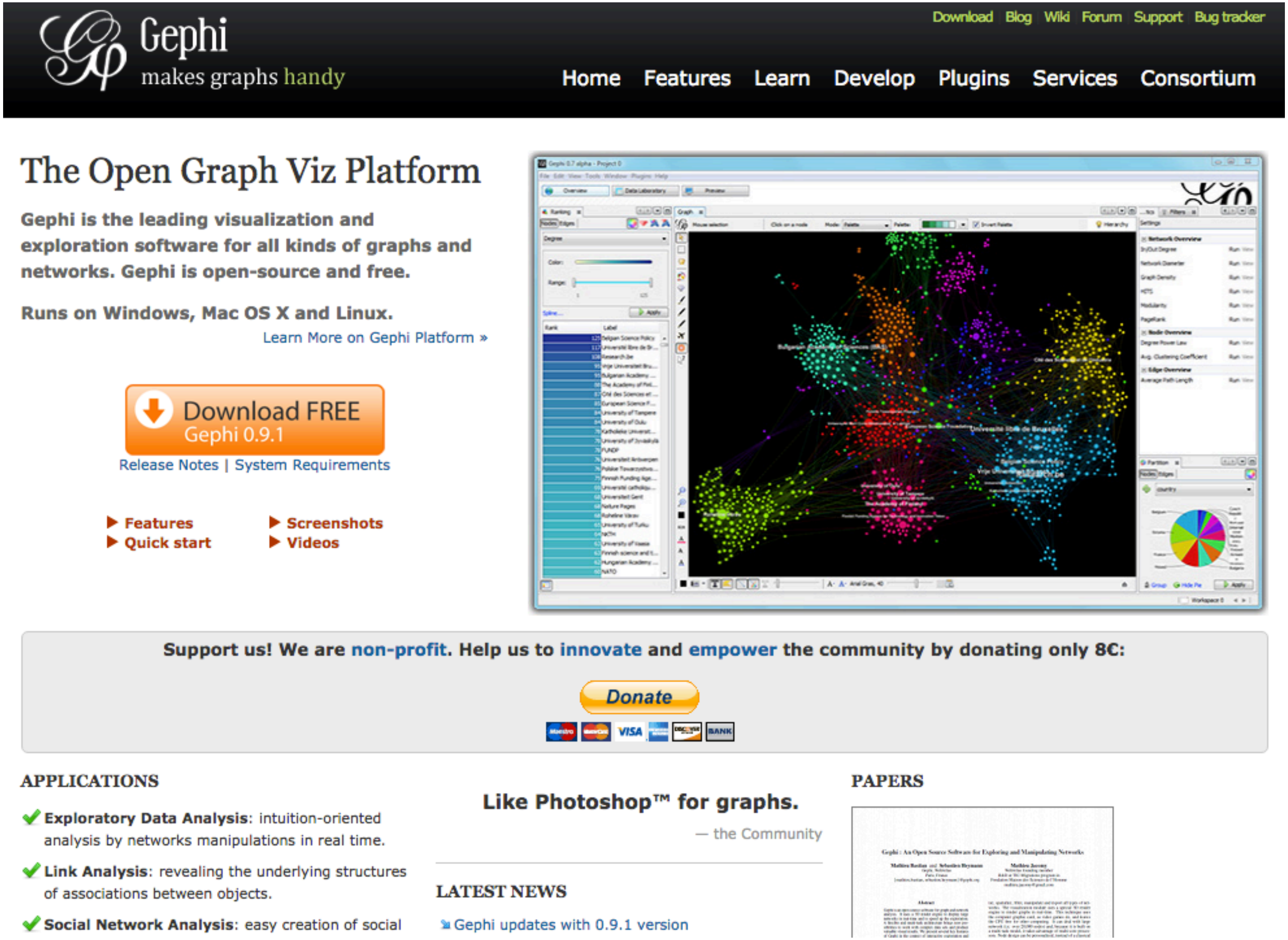




1. Make complex things **simple**
2. Extract **small** information from large data
3. Present **truth**, do not deceive

Source: http://sebastien.pro/gephi-icwsm-tutorial.pdf

Gephi



Gephi
makes graphs handy

Download Blog Wiki Forum Support Bug tracker

Home Features Learn Develop Plugins Services Consortium

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.1

Release Notes | System Requirements

- **Features**
- **Quick start**
- **Screenshots**
- **Videos**

Support us! We are non-profit. Help us to innovate and empower the community by donating only 8€:

Donate

APPLICATIONS

✓ **Exploratory Data Analysis**: intuition-oriented analysis by networks manipulations in real time.

✓ **Link Analysis**: revealing the underlying structures of associations between objects.

✓ **Social Network Analysis**: easy creation of social

Like Photoshop™ for graphs.

— the Community

LATEST NEWS

Gephi updates with 0.9.1 version

PAPERS

Gephi : An Open Source Software for Exploring and Manipulating Networks

https://gephi.org/

Discovering, Analyzing, Visualizing and Presenting Data with Python in Google Colab

Python Data Analysis and Visualization

pandas

plotly

matplotlib

bokeh

seaborn

Altair

Python Pandas

http://pandas.pydata.org/

Python matplotlib

Source: https://matplotlib.org/

Python seaborn

Source: https://seaborn.pydata.org/

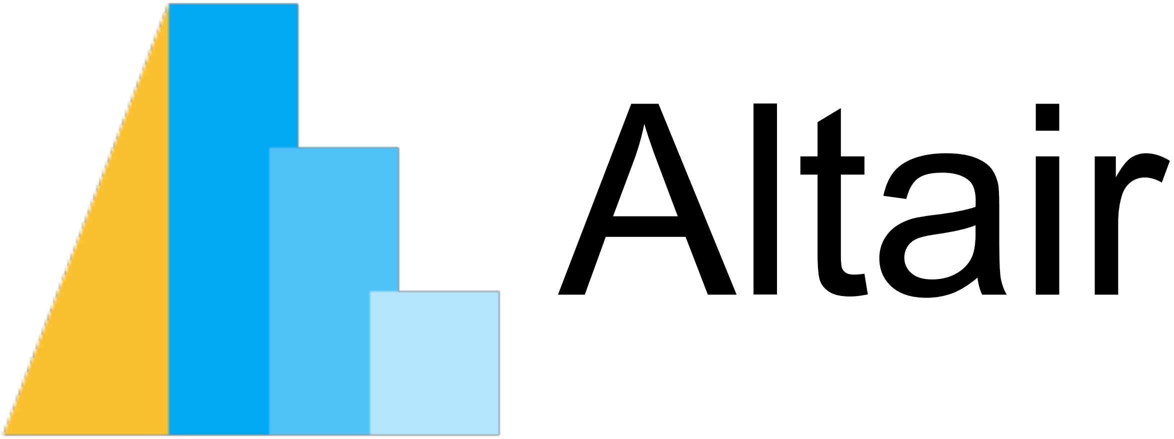
Python plotly

Source: https://plotly.com/python/

Python
bokeh

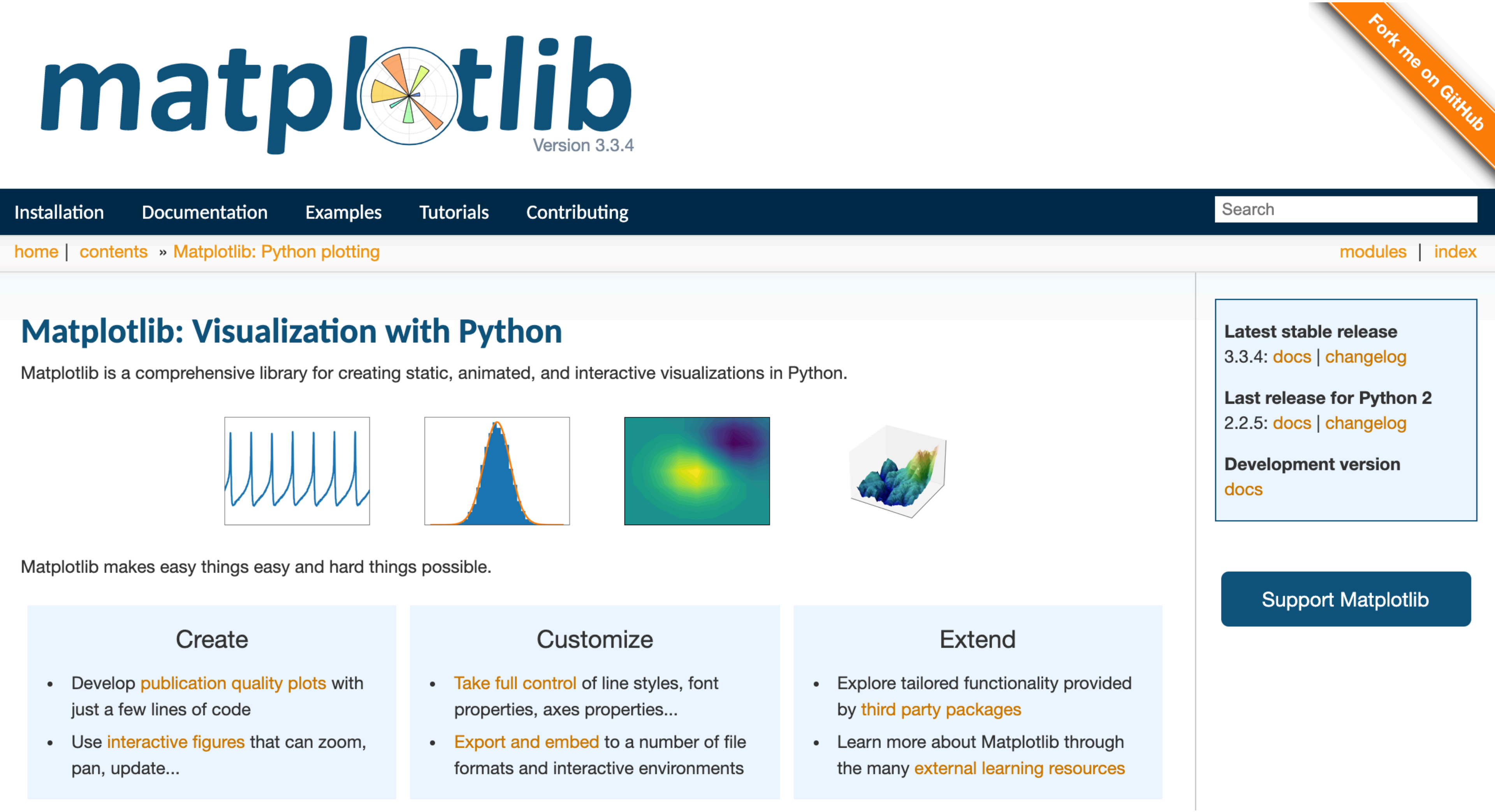
Source: https://bokeh.org/

Python Altair



Source: https://altair-viz.github.io/

Python matplotlib



Matplotlib: Visualization with Python

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Matplotlib makes easy things easy and hard things possible.

Create

- Develop publication quality plots with just a few lines of code
- Use interactive figures that can zoom, pan, update...

Customize

- Take full control of line styles, font properties, axes properties...
- Export and embed to a number of file formats and interactive environments

Extend

- Explore tailored functionality provided by third party packages
- Learn more about Matplotlib through the many external learning resources

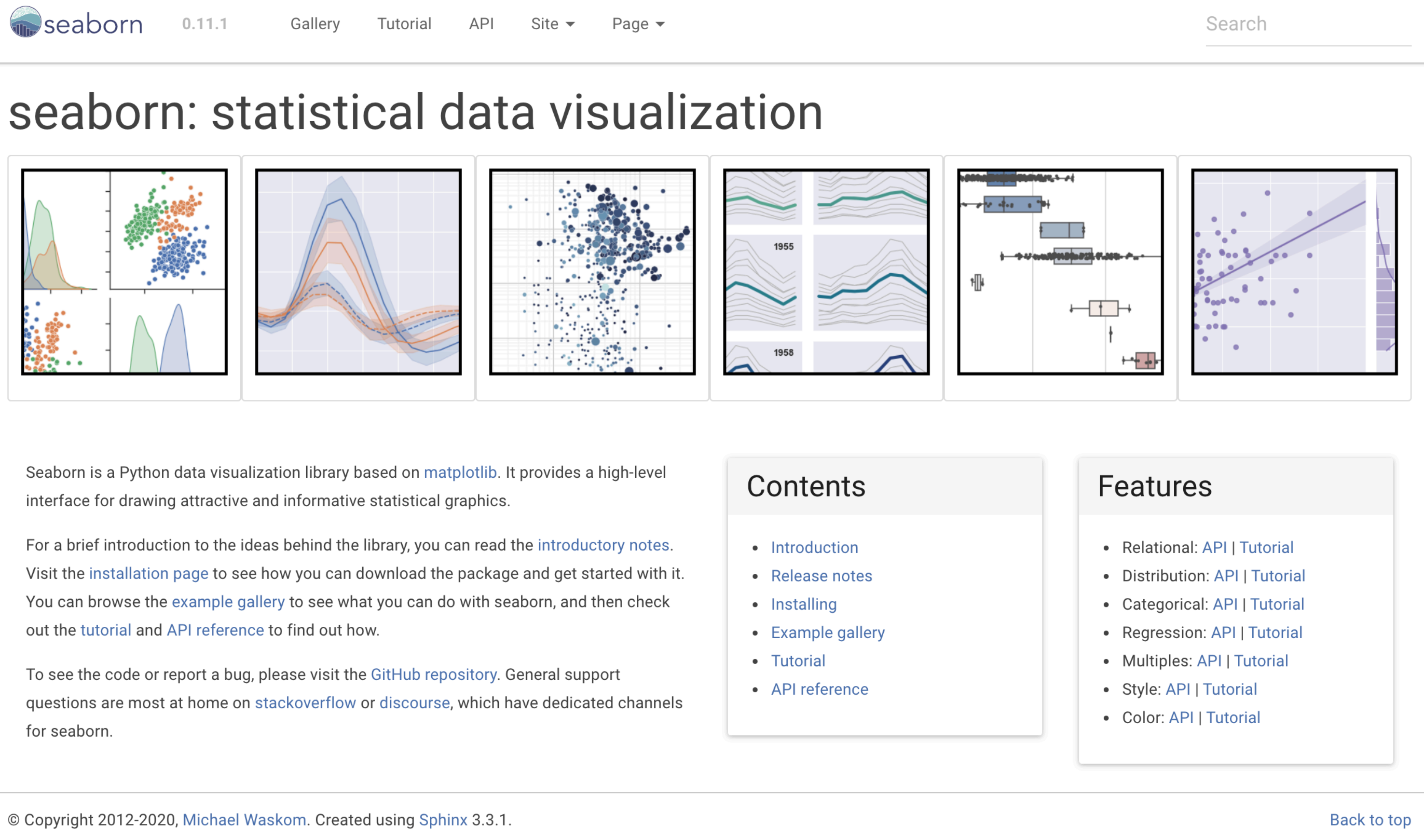
Latest stable release
3.3.4: docs | changelog

Last release for Python 2
2.2.5: docs | changelog

Development version
docs

https://matplotlib.org/

Python Seaborn



seaborn: statistical data visualization

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the introductory notes. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see what you can do with seaborn, and then check out the tutorial and API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow or discourse, which have dedicated channels for seaborn.

Contents

- Introduction
- Release notes
- Installing
- Example gallery
- Tutorial
- API reference

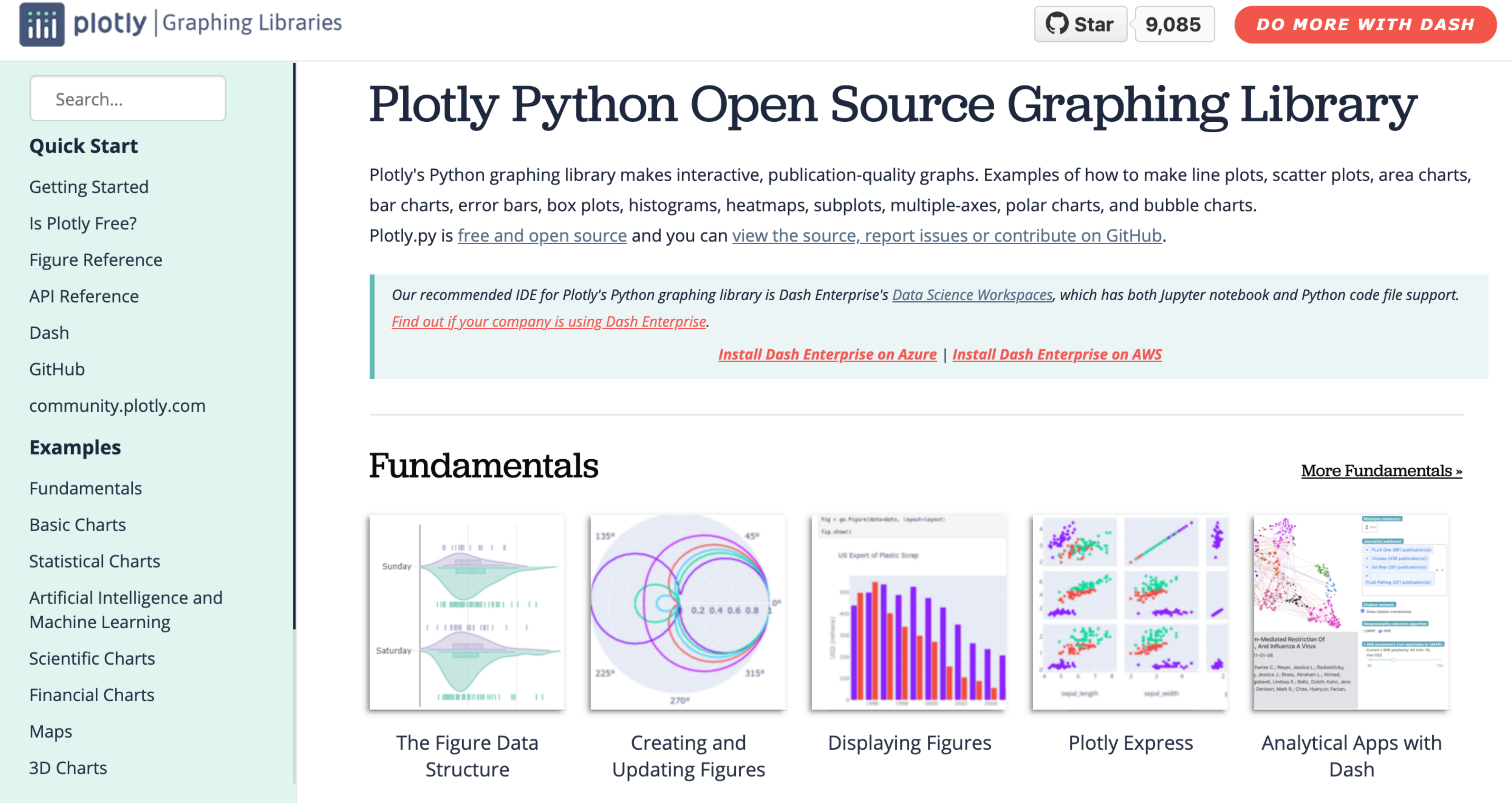
Features

- Relational: API | Tutorial
- Distribution: API | Tutorial
- Categorical: API | Tutorial
- Regression: API | Tutorial
- Multiples: API | Tutorial
- Style: API | Tutorial
- Color: API | Tutorial

https://seaborn.pydata.org/

plotly

Python Plotly Graphing Library



plotly | Graphing Libraries

Star 9,085

DO MORE WITH DASH

Search...

Quick Start

- Getting Started
- Is Plotly Free?
- Figure Reference
- API Reference
- Dash
- GitHub
- community.plotly.com

Examples

- Fundamentals
- Basic Charts
- Statistical Charts
- Artificial Intelligence and Machine Learning
- Scientific Charts
- Financial Charts
- Maps
- 3D Charts

Plotly Python Open Source Graphing Library

Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.
Plotly.py is free and open source and you can view the source, report issues or contribute on GitHub.

> *Our recommended IDE for Plotly's Python graphing library is Dash Enterprise's Data Science Workspaces, which has both Jupyter notebook and Python code file support.*
> *Find out if your company is using Dash Enterprise.*
>
> ***Install Dash Enterprise on Azure* | *Install Dash Enterprise on AWS***

Fundamentals

More Fundamentals »

The Figure Data Structure

Creating and Updating Figures

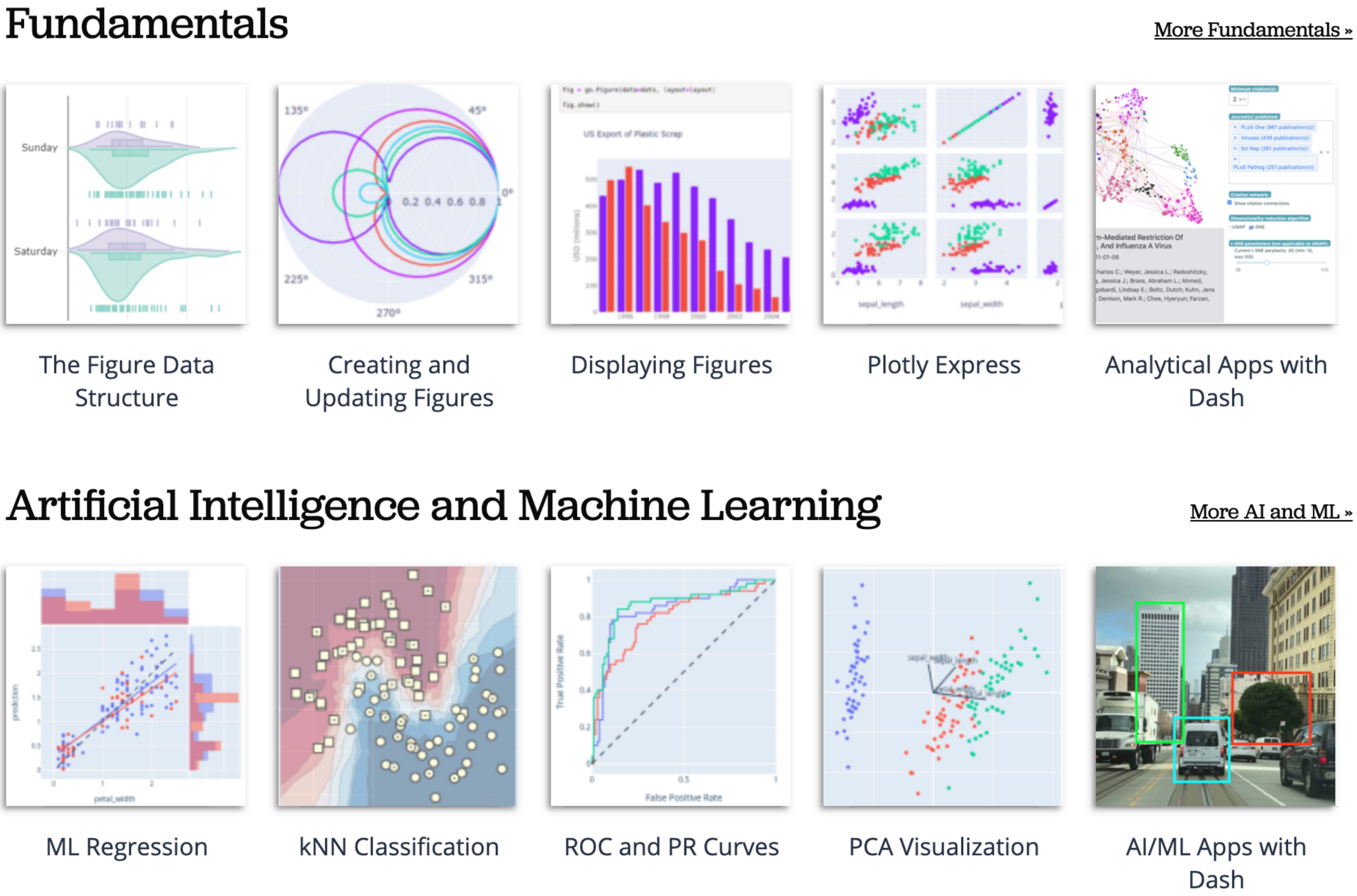
Displaying Figures

Plotly Express

Analytical Apps with Dash

https://plotly.com/python/

Python Plotly Graphing Library



Fundamentals

More Fundamentals »

The Figure Data Structure

Creating and Updating Figures

Displaying Figures

Plotly Express

Analytical Apps with Dash

Artificial Intelligence and Machine Learning

More AI and ML »

ML Regression

kNN Classification

ROC and PR Curves

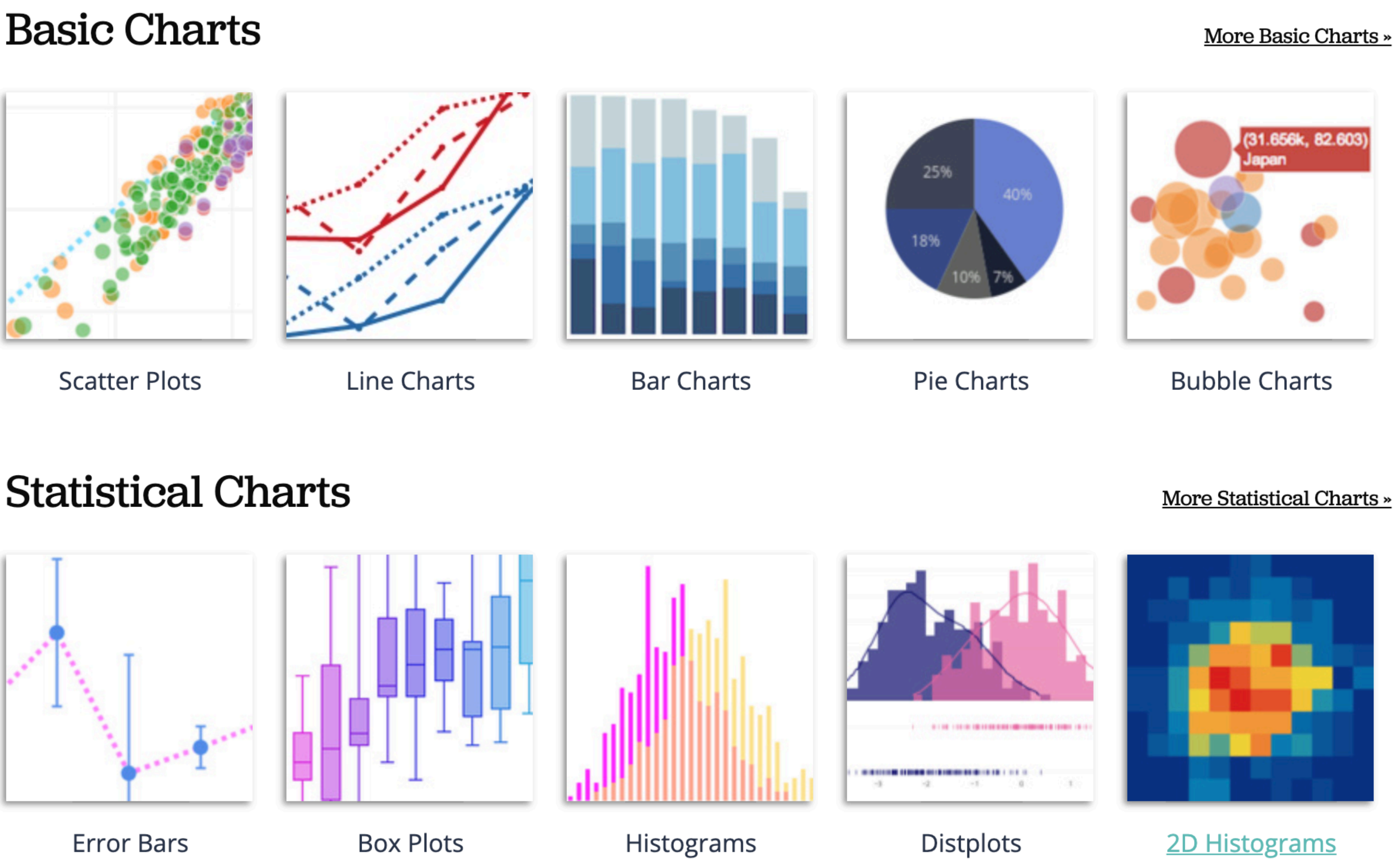
PCA Visualization

AI/ML Apps with Dash

https://plotly.com/python/

plotly

Python Plotly Graphing Library



Basic Charts

More Basic Charts »

Scatter Plots

Line Charts

Bar Charts

Pie Charts

Bubble Charts

Statistical Charts

More Statistical Charts »

Error Bars

Box Plots

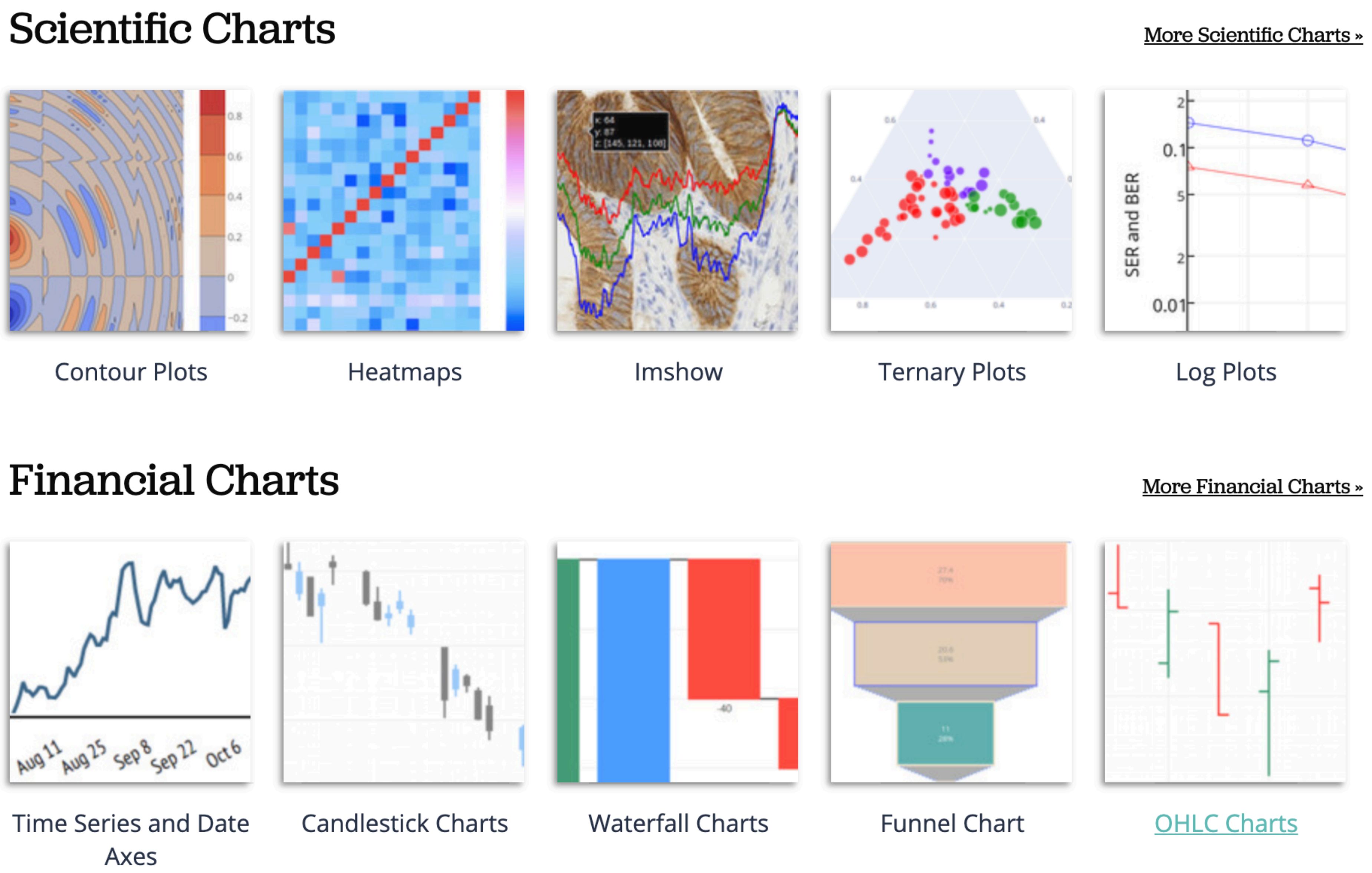
Histograms

Distplots

2D Histograms

https://plotly.com/python/

Python Plotly Graphing Library



Scientific Charts

More Scientific Charts »

Contour Plots

Heatmaps

Imshow

Ternary Plots

Log Plots

Financial Charts

More Financial Charts »

Time Series and Date Axes

Candlestick Charts

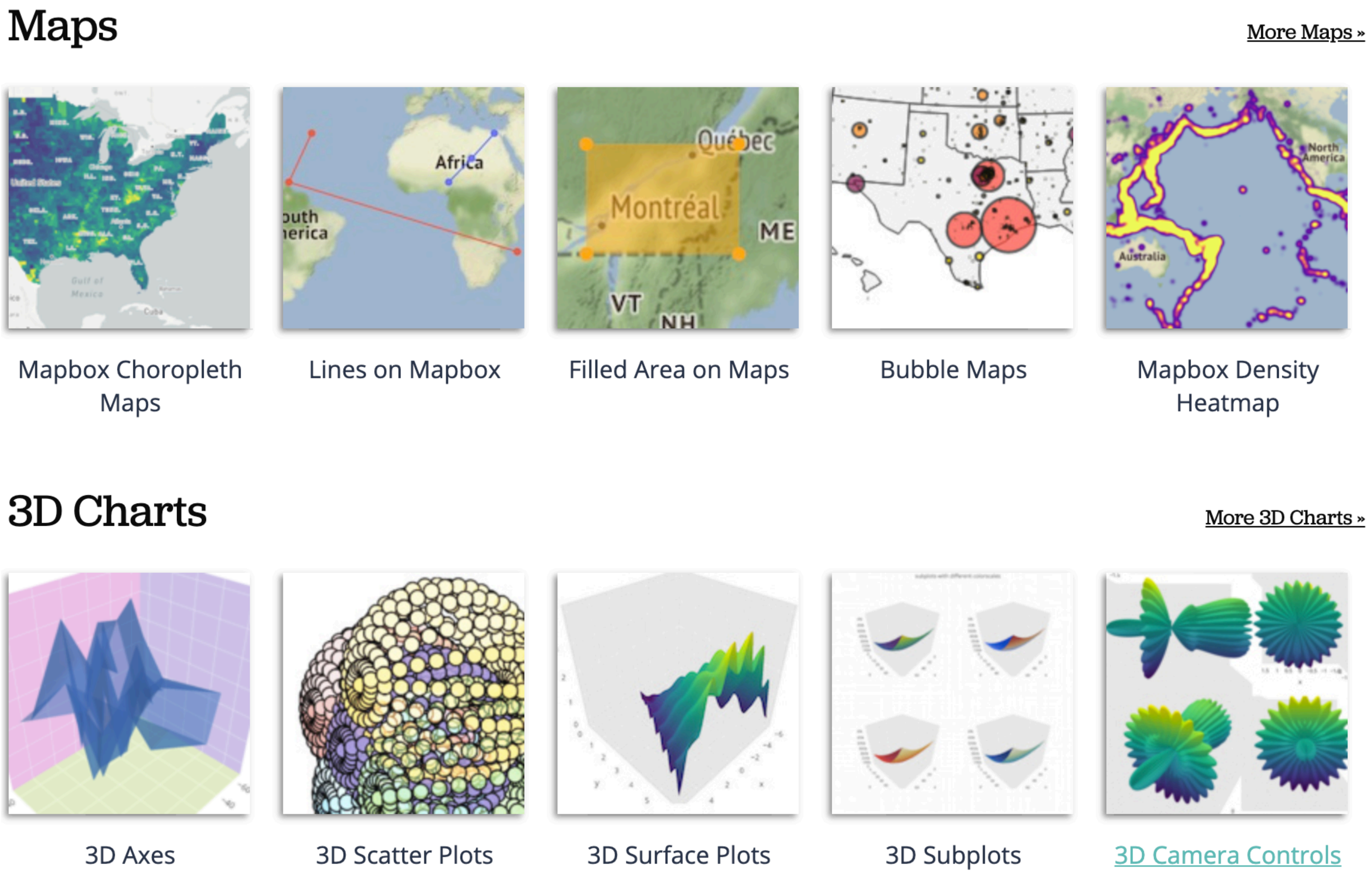
Waterfall Charts

Funnel Chart

OHLC Charts

https://plotly.com/python/

Python Plotly Graphing Library



Maps

More Maps »

Mapbox Choropleth Maps

Lines on Mapbox

Filled Area on Maps

Bubble Maps

Mapbox Density Heatmap

3D Charts

More 3D Charts »

3D Axes

3D Scatter Plots

3D Surface Plots

3D Subplots

3D Camera Controls

https://plotly.com/python/

Python Plotly Graphing Library

Subplots

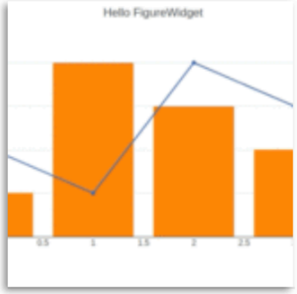
Mixed Subplots

Map Subplots

Table and Chart Subplots

Figure Factory Subplots

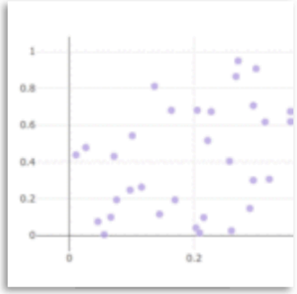
Jupyter Widgets Interaction



Plotly FigureWidget Overview

Jupyter Lab with FigureWidget

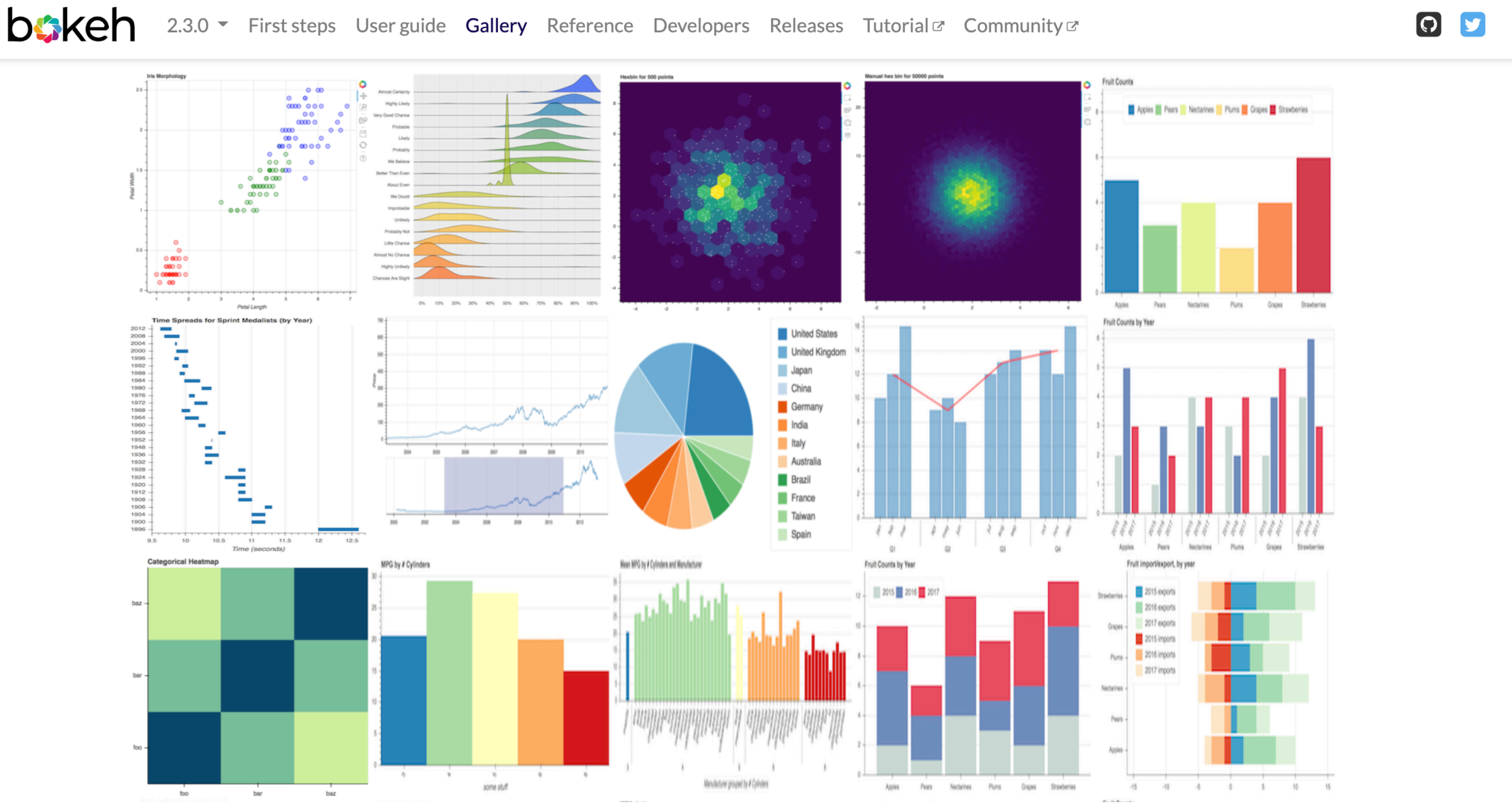
Interactive Data Analysis with FigureWidget ipywidgets



Click Events

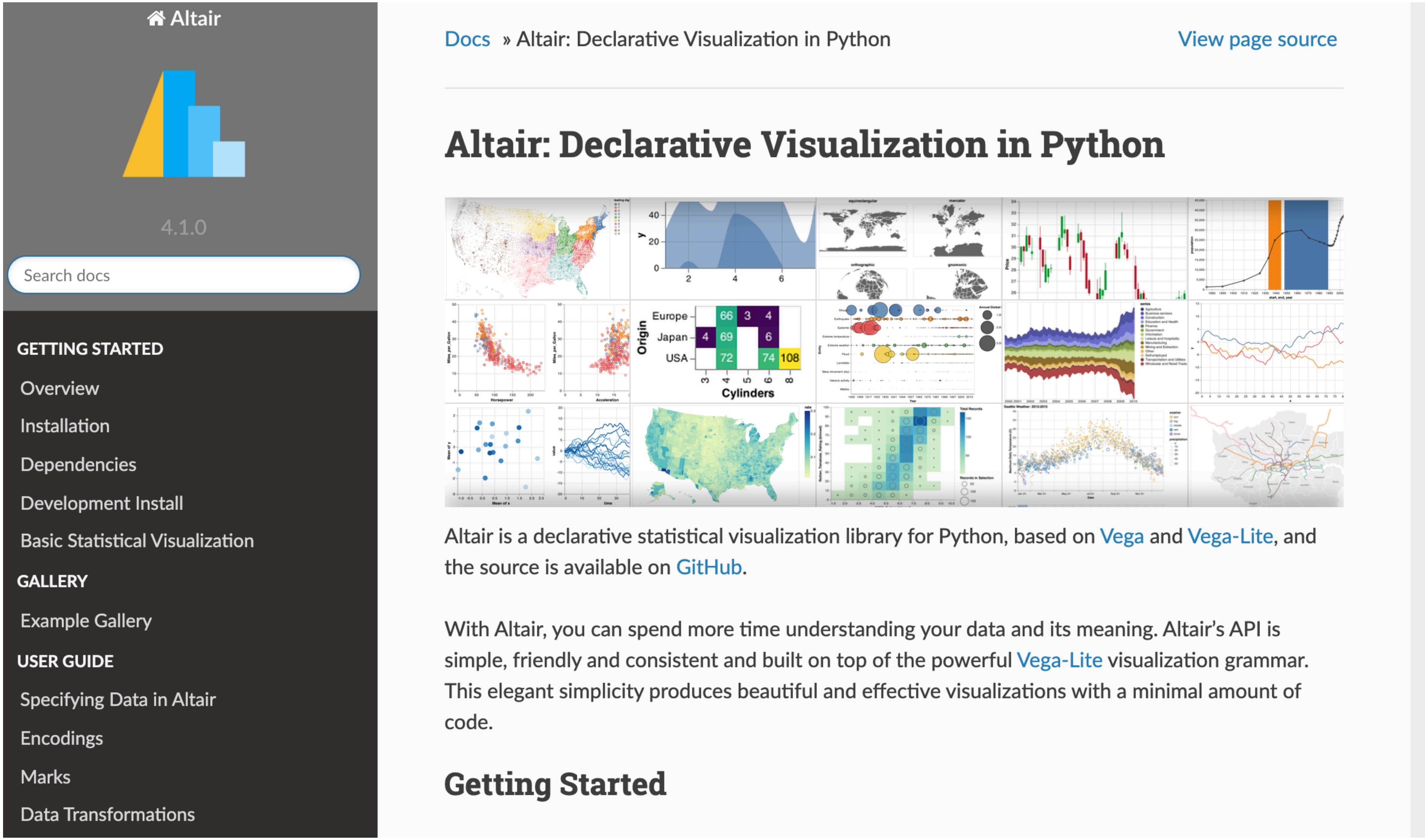
https://plotly.com/python/

Python Bokeh



https://bokeh.org/

Python Altair



Altair

4.1.0

Search docs

GETTING STARTED

- Overview
- Installation
- Dependencies
- Development Install
- Basic Statistical Visualization

GALLERY

- Example Gallery

USER GUIDE

- Specifying Data in Altair
- Encodings
- Marks
- Data Transformations

Docs » Altair: Declarative Visualization in Python

View page source

Altair: Declarative Visualization in Python

Altair is a declarative statistical visualization library for Python, based on Vega and Vega-Lite, and the source is available on GitHub.

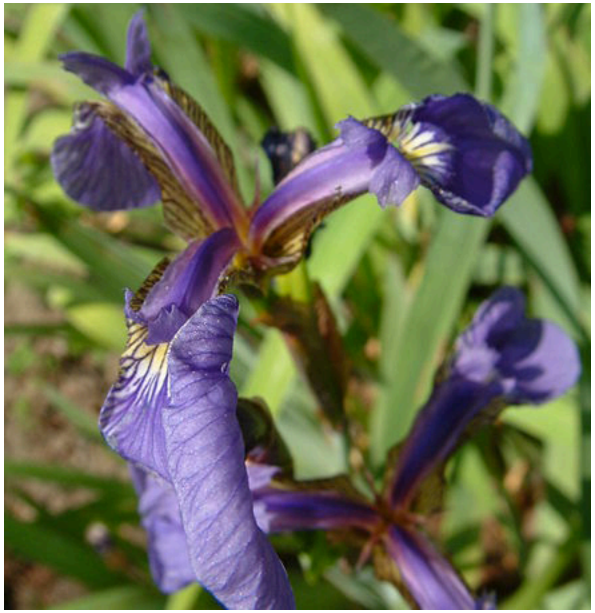
With Altair, you can spend more time understanding your data and its meaning. Altair's API is simple, friendly and consistent and built on top of the powerful Vega-Lite visualization grammar. This elegant simplicity produces beautiful and effective visualizations with a minimal amount of code.

Getting Started

https://altair-viz.github.io/

Iris flower data set

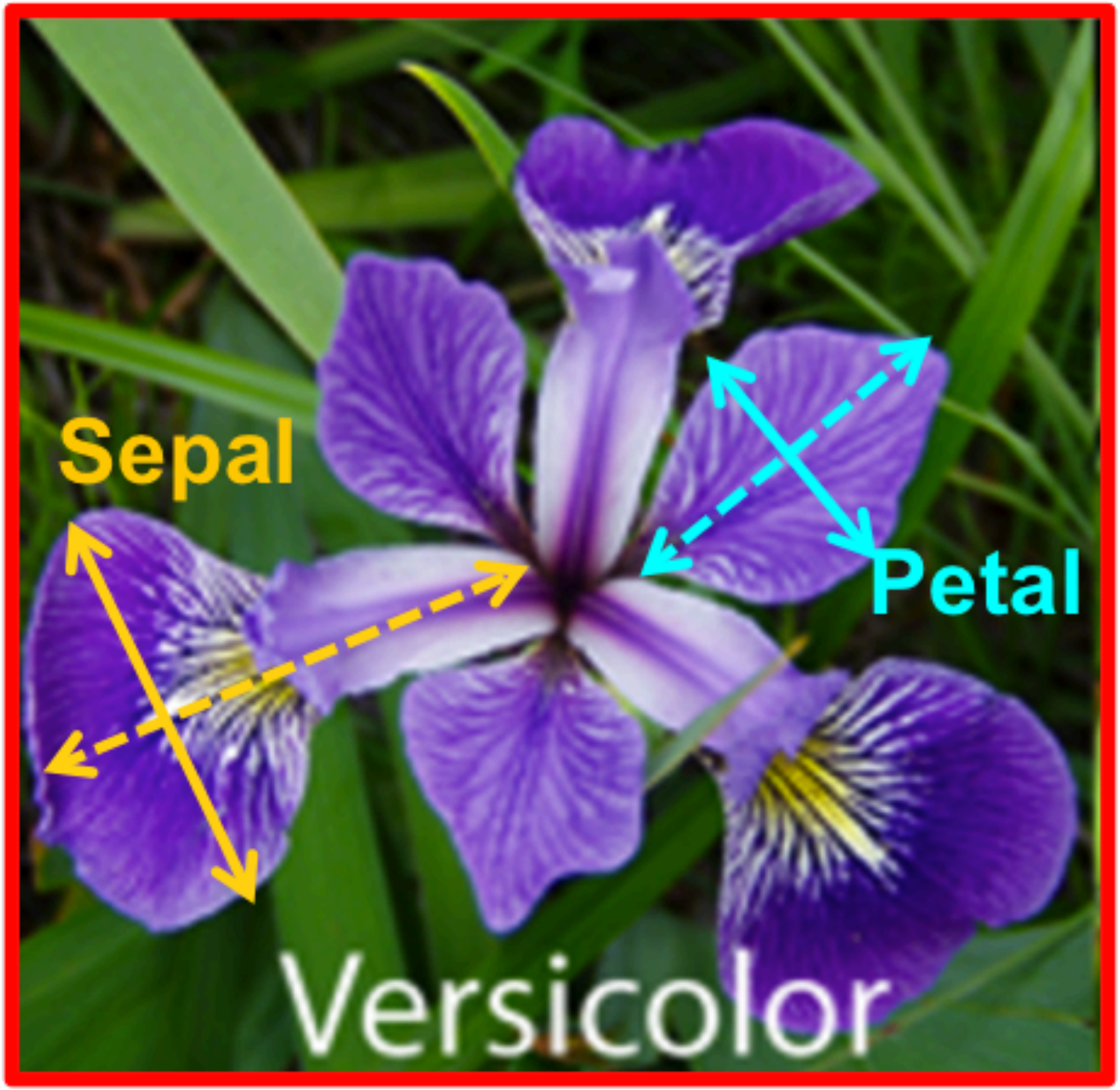
setosa **versicolor** **virginica**



Source: https://en.wikipedia.org/wiki/Iris_flower_data_set

Source: http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/

Iris Classfication



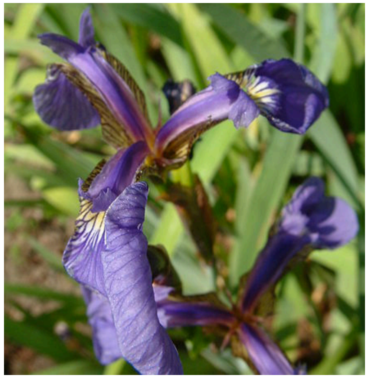
Source: http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/

iris.data

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
```

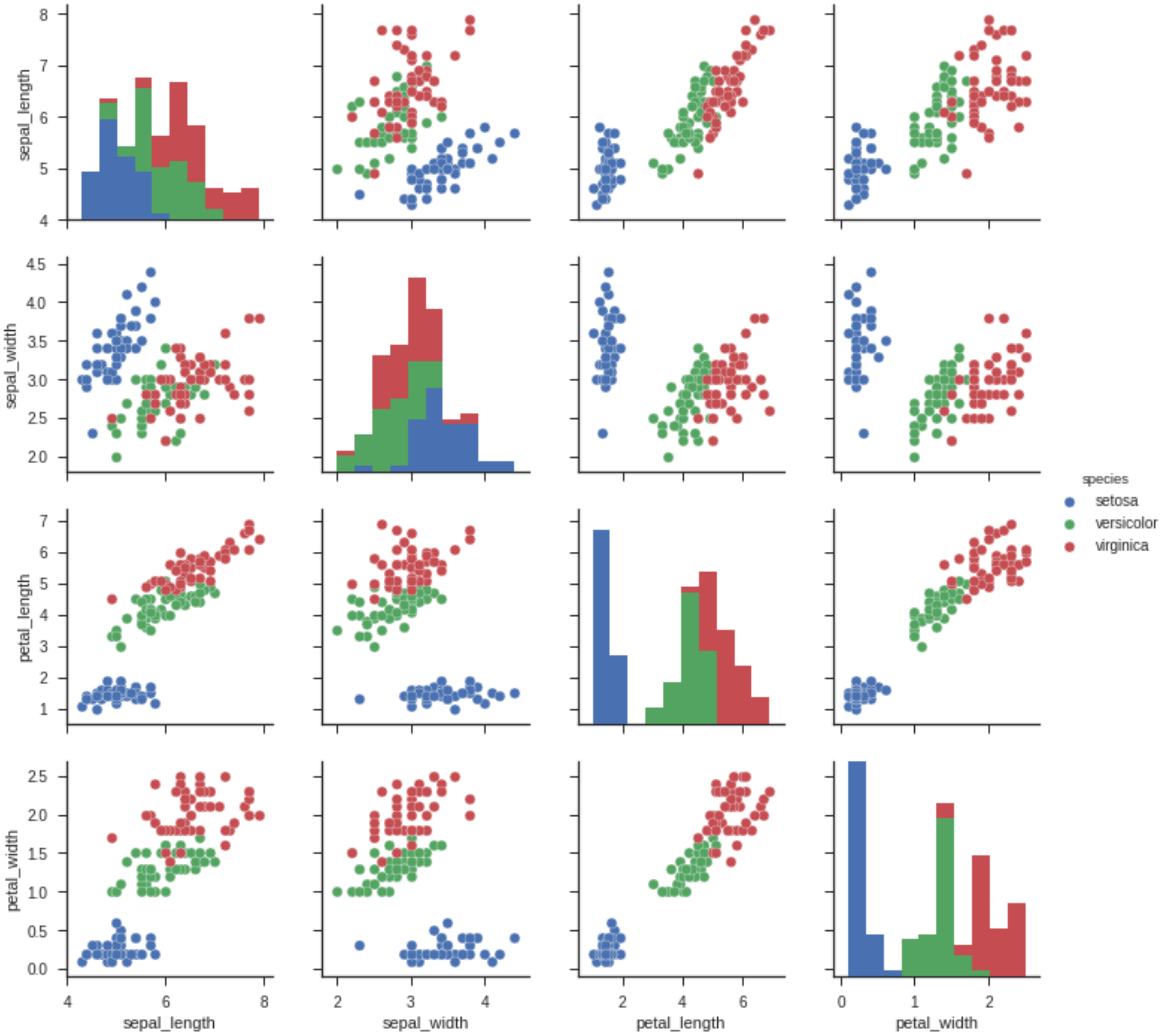
setosa



virginica

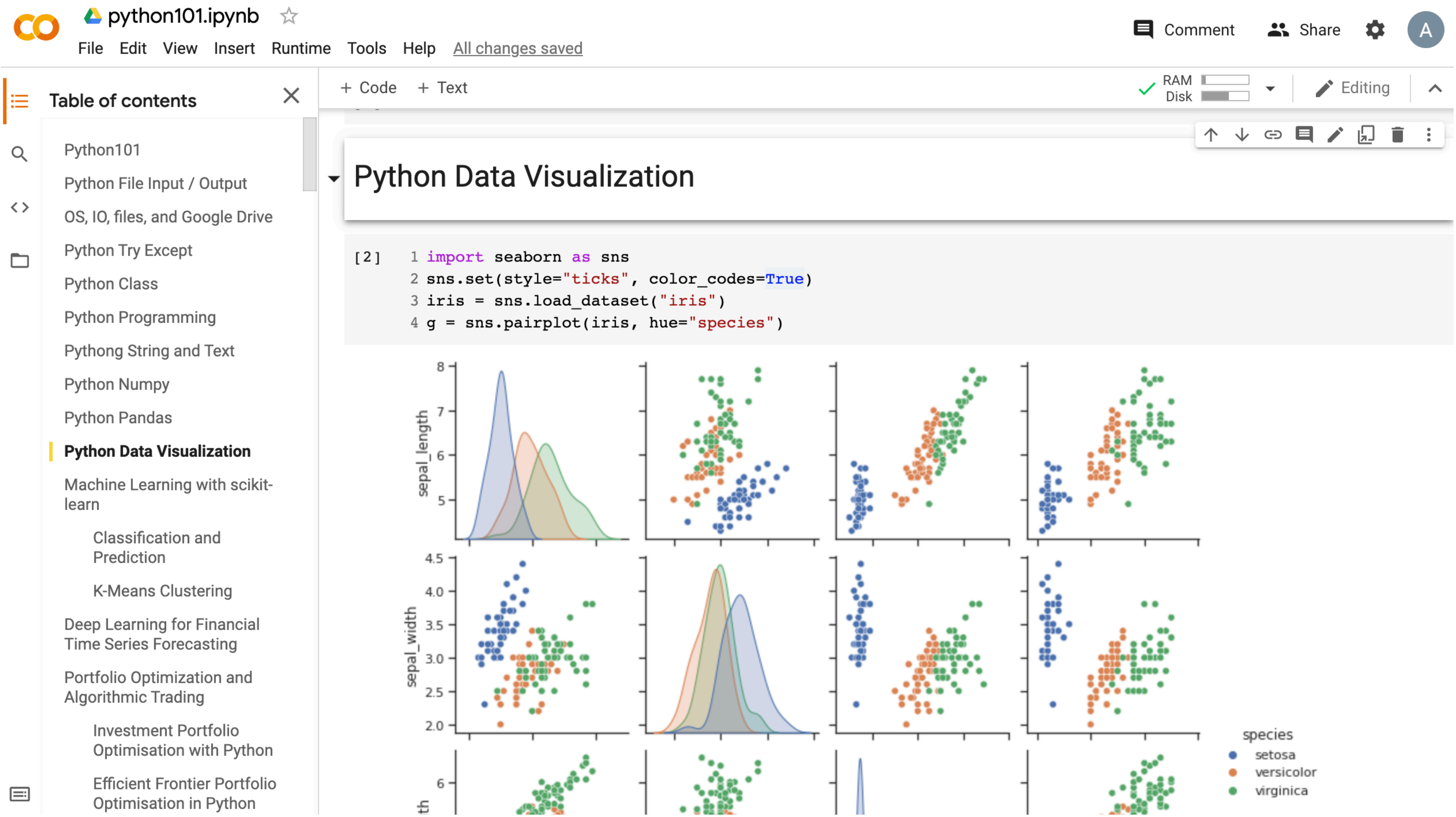
versicolor

Iris Data Visualization



Source: https://seaborn.pydata.org/generated/seaborn.pairplot.html

Data Visualization in Google Colab

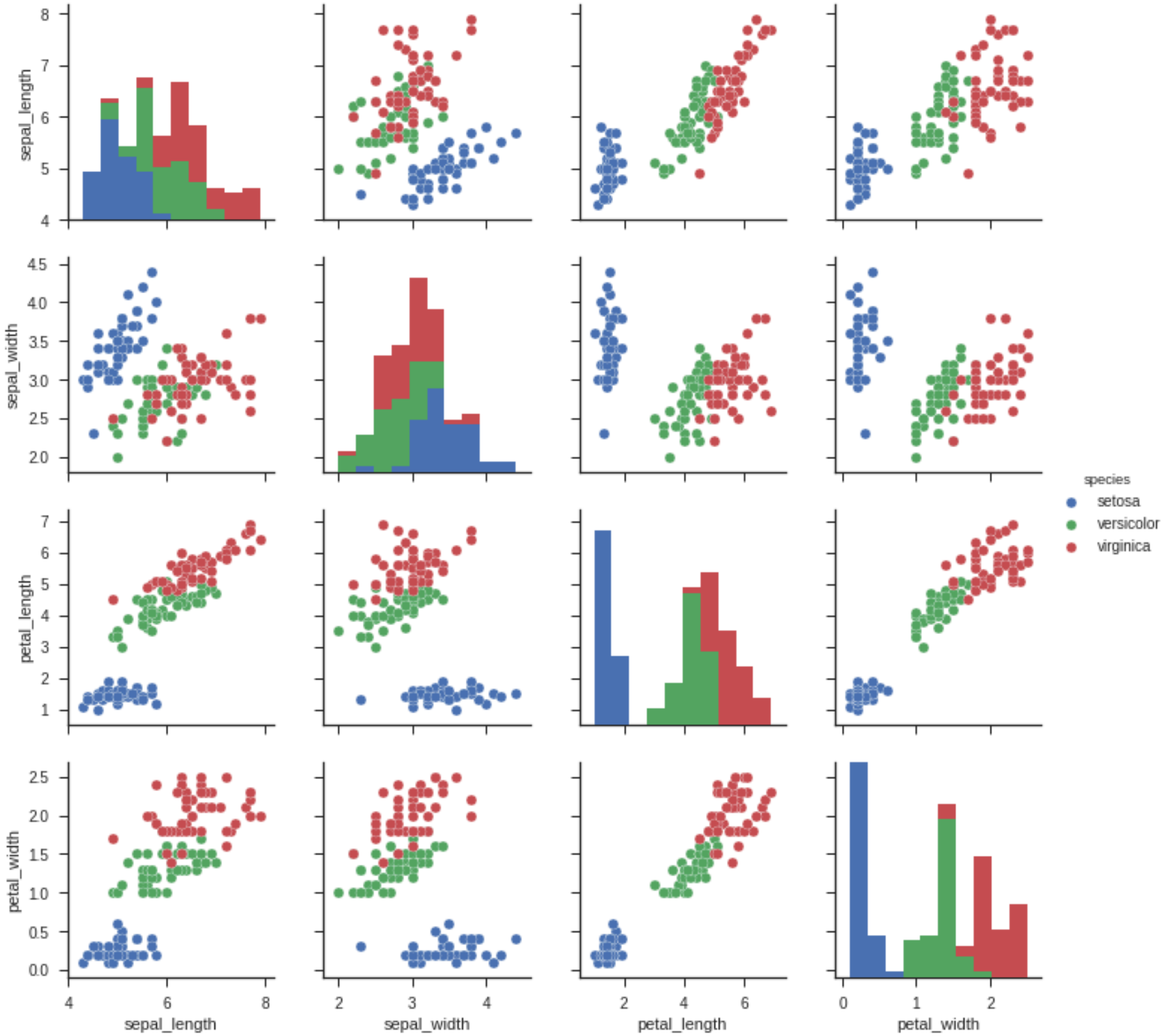


Python Data Visualization

```python
import seaborn as sns
sns.set(style="ticks", color_codes=True)
iris = sns.load_dataset("iris")
g = sns.pairplot(iris, hue="species")
```

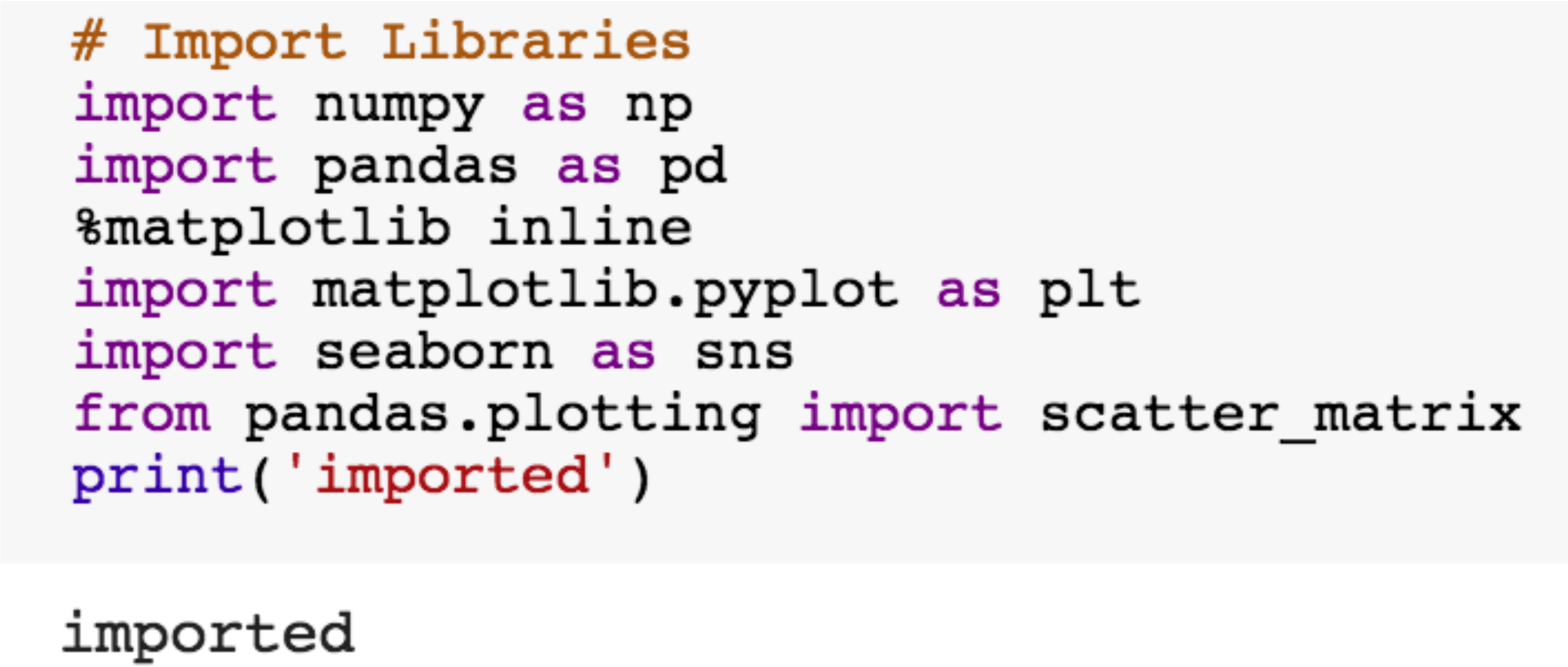
https://tinyurl.com/aintpupython101

```
import seaborn as sns
sns.set(style="ticks", color_codes=True)
iris = sns.load_dataset("iris")
g = sns.pairplot(iris, hue="species")
```



Source: https://seaborn.pydata.org/generated/seaborn.pairplot.html

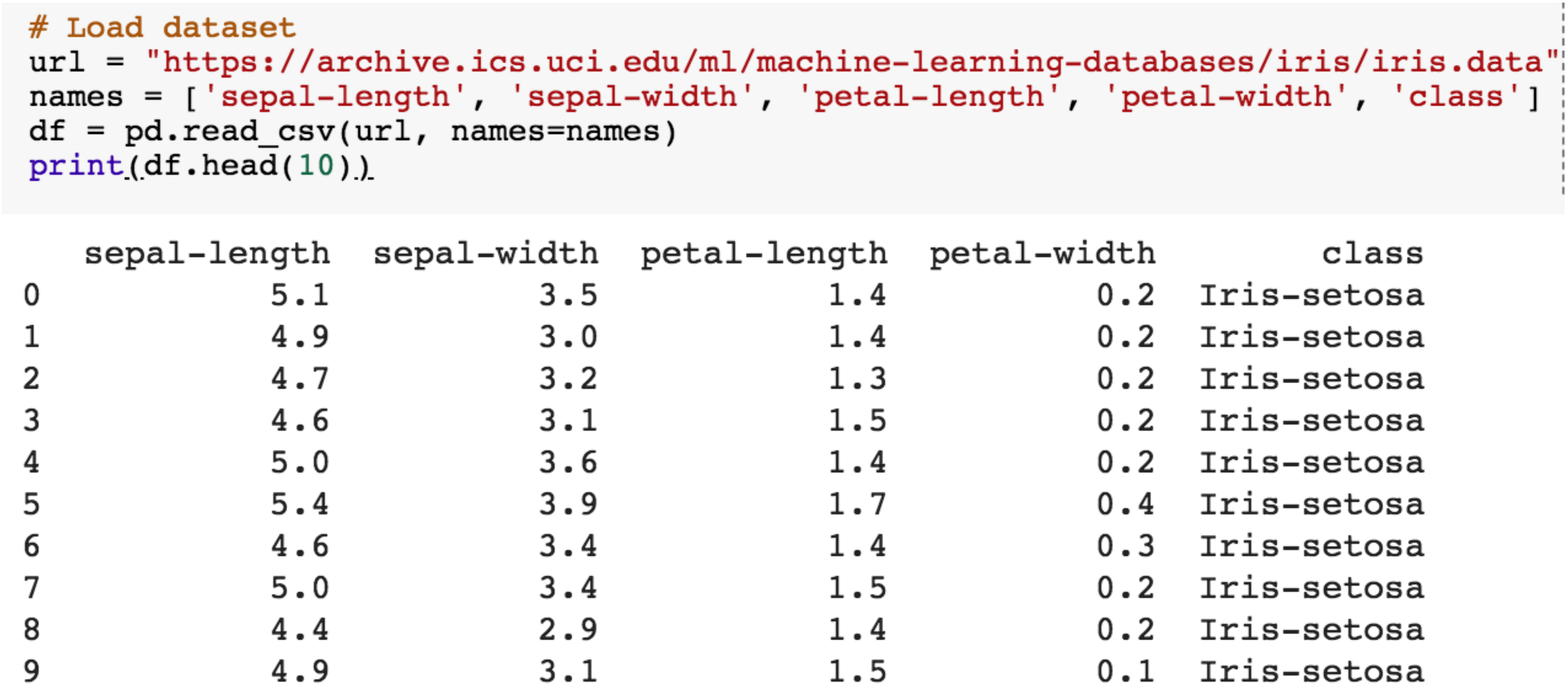
```
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix
```



```
# Import Libraries
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix
print('imported')
```

```
imported
```

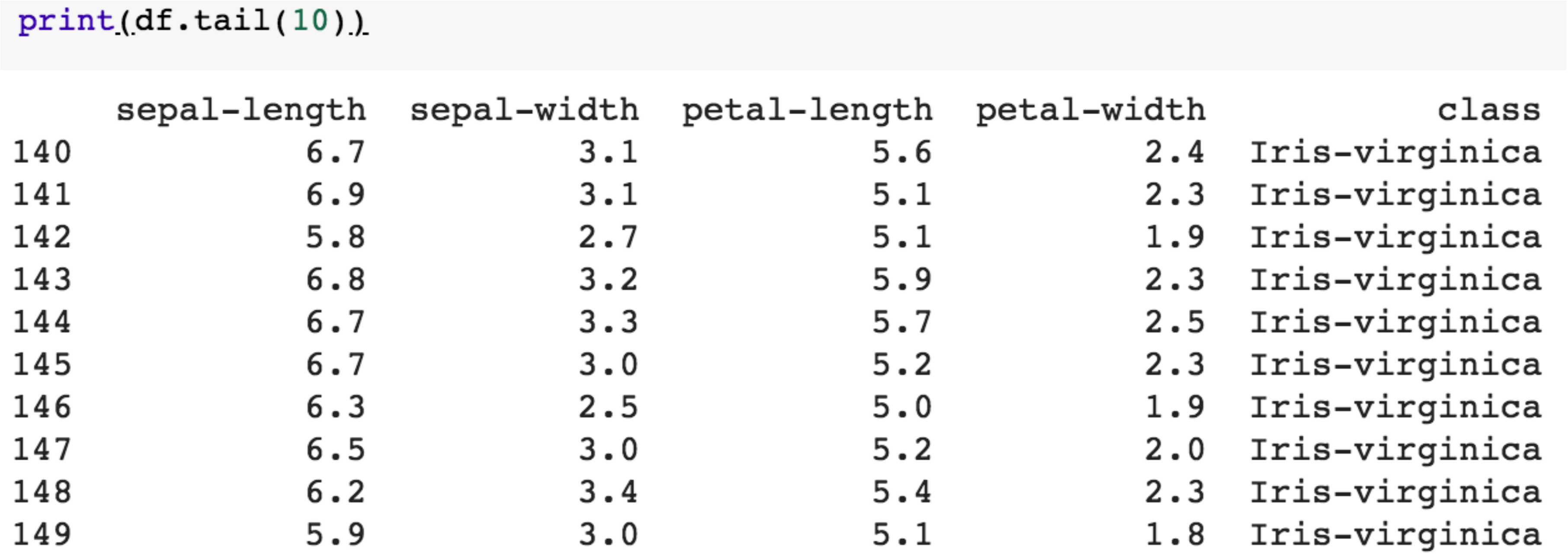
```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
df = pd.read_csv(url, names=names)
print(df.head(10))
```



```
# Load dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
df = pd.read_csv(url, names=names)
print(df.head(10))
```

```
   sepal-length  sepal-width  petal-length  petal-width        class
0           5.1          3.5           1.4          0.2  Iris-setosa
1           4.9          3.0           1.4          0.2  Iris-setosa
2           4.7          3.2           1.3          0.2  Iris-setosa
3           4.6          3.1           1.5          0.2  Iris-setosa
4           5.0          3.6           1.4          0.2  Iris-setosa
5           5.4          3.9           1.7          0.4  Iris-setosa
6           4.6          3.4           1.4          0.3  Iris-setosa
7           5.0          3.4           1.5          0.2  Iris-setosa
8           4.4          2.9           1.4          0.2  Iris-setosa
9           4.9          3.1           1.5          0.1  Iris-setosa
```

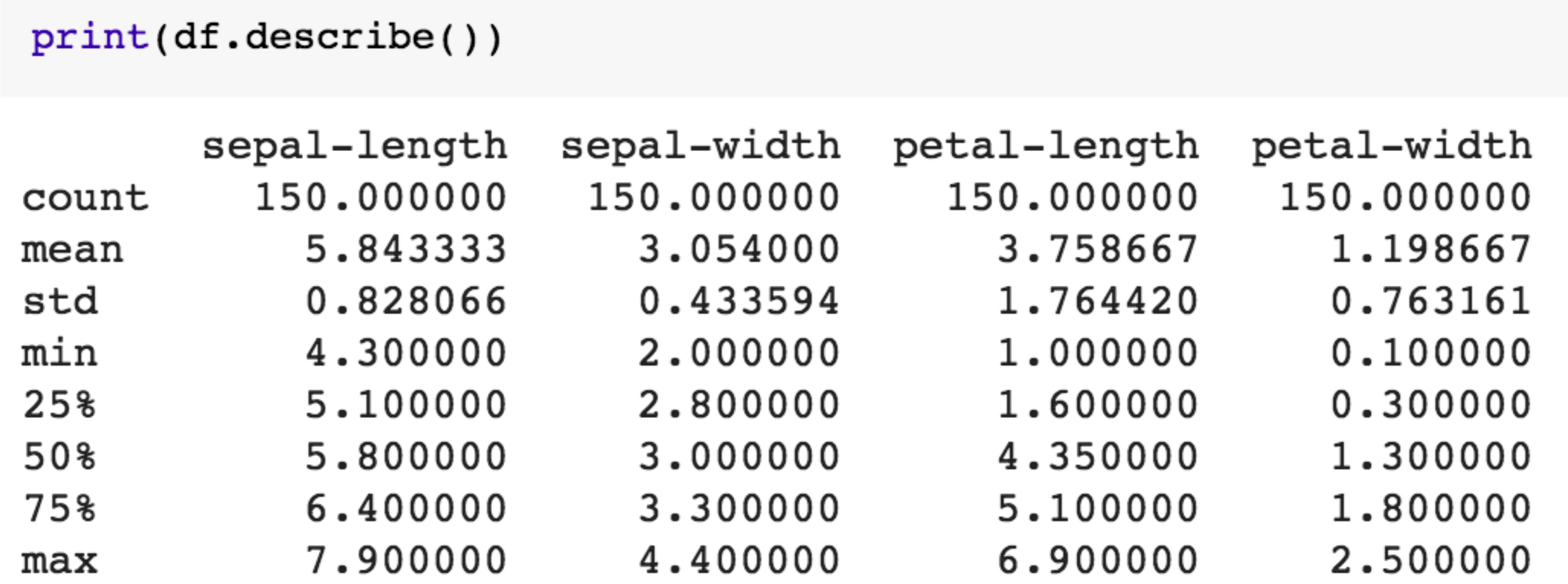
df.tail(10)



```
print(df.tail(10))
```

```
     sepal-length  sepal-width  petal-length  petal-width           class
140           6.7          3.1           5.6          2.4  Iris-virginica
141           6.9          3.1           5.1          2.3  Iris-virginica
142           5.8          2.7           5.1          1.9  Iris-virginica
143           6.8          3.2           5.9          2.3  Iris-virginica
144           6.7          3.3           5.7          2.5  Iris-virginica
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica
```

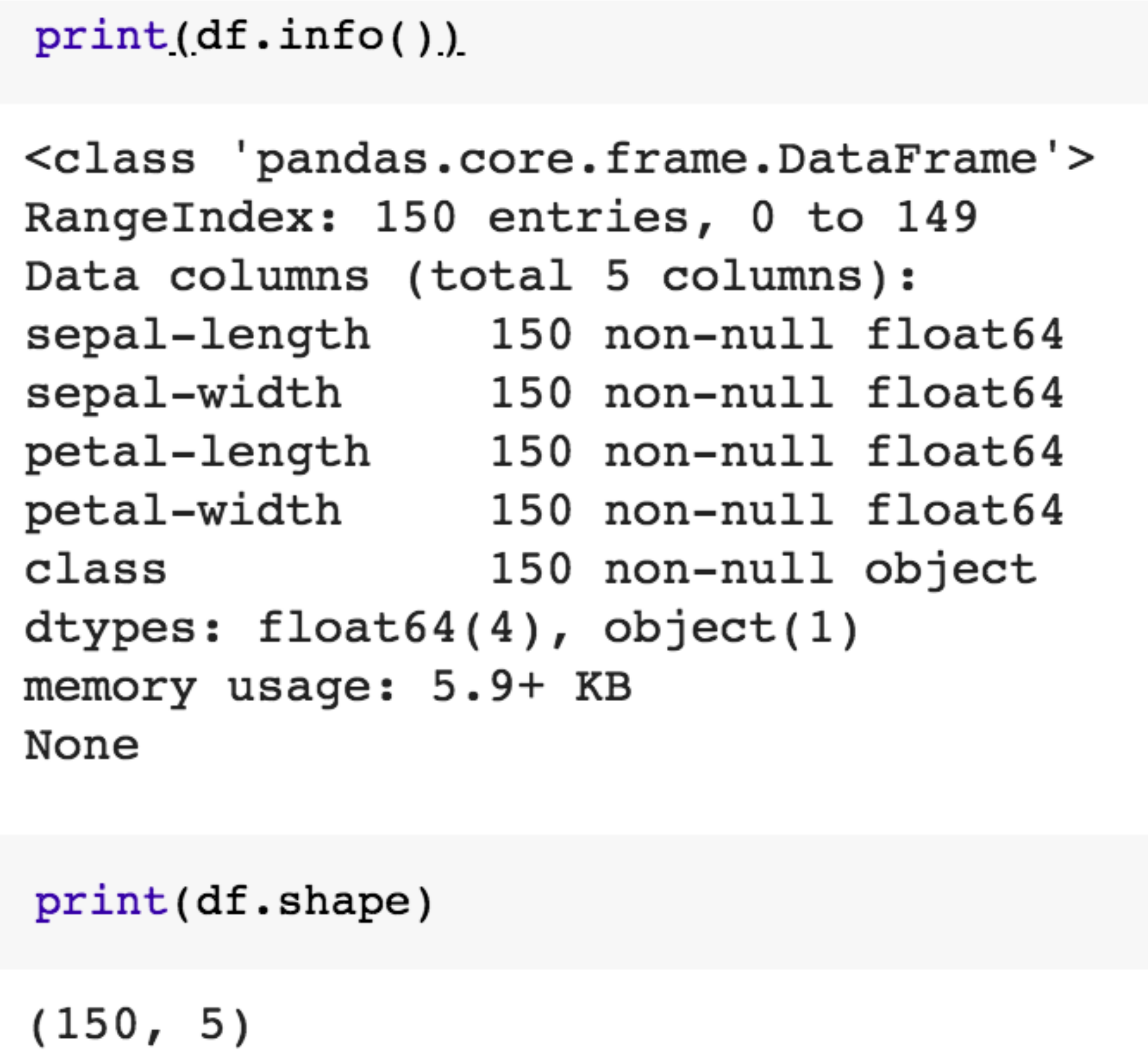
df.describe()



```
print(df.describe())
```

```
       sepal-length  sepal-width  petal-length  petal-width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.054000      3.758667     1.198667
std        0.828066     0.433594      1.764420     0.763161
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```

```
print(df.info())
print(df.shape)
```



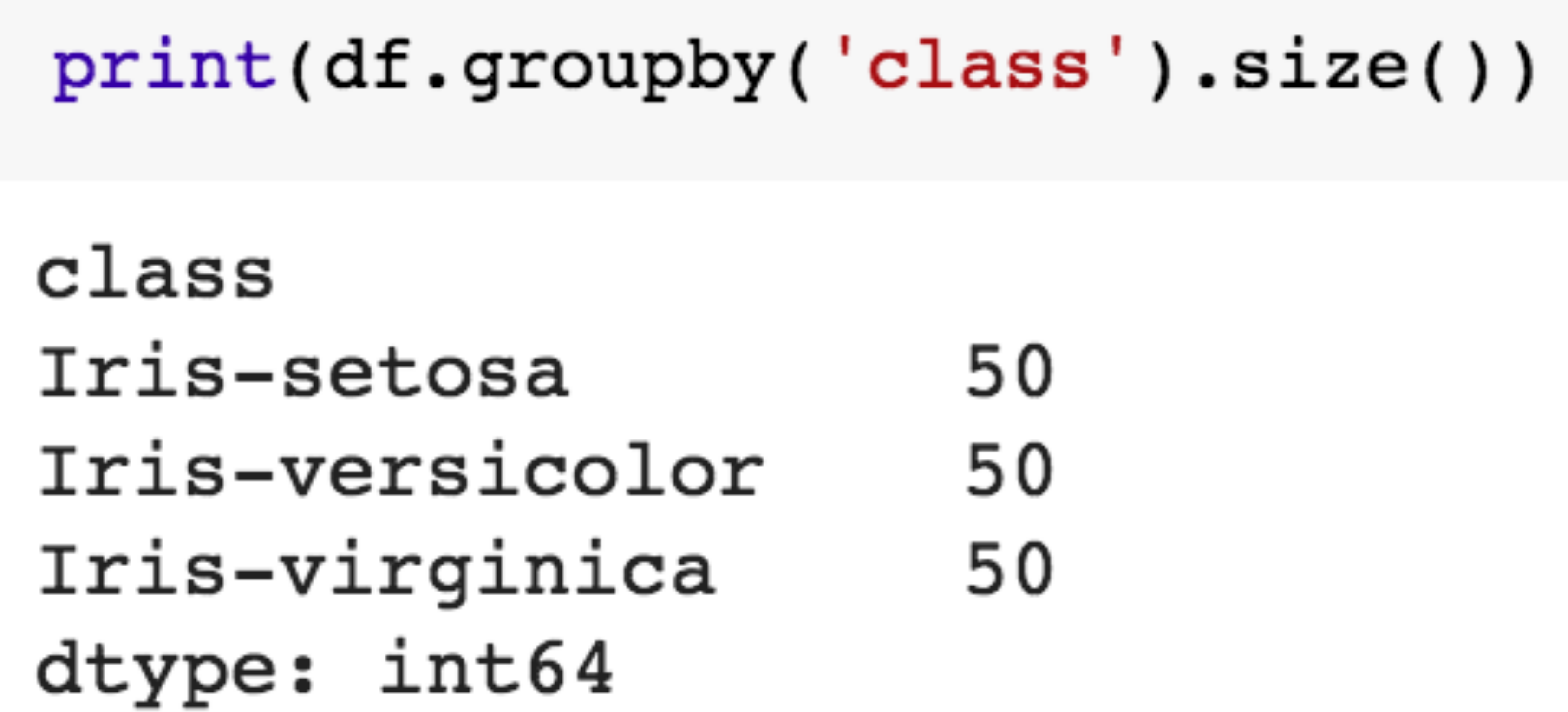
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
sepal-length    150 non-null float64
sepal-width     150 non-null float64
petal-length    150 non-null float64
petal-width     150 non-null float64
class           150 non-null object
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
None
```

```
print(df.shape)
```

```
(150, 5)
```

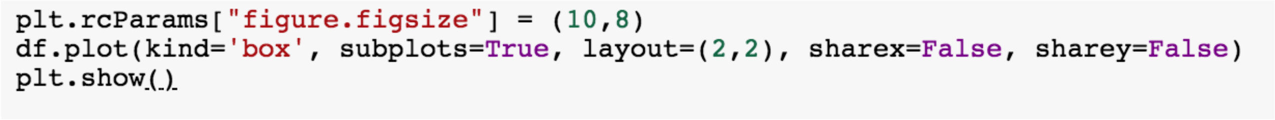
df.groupby('class').size()



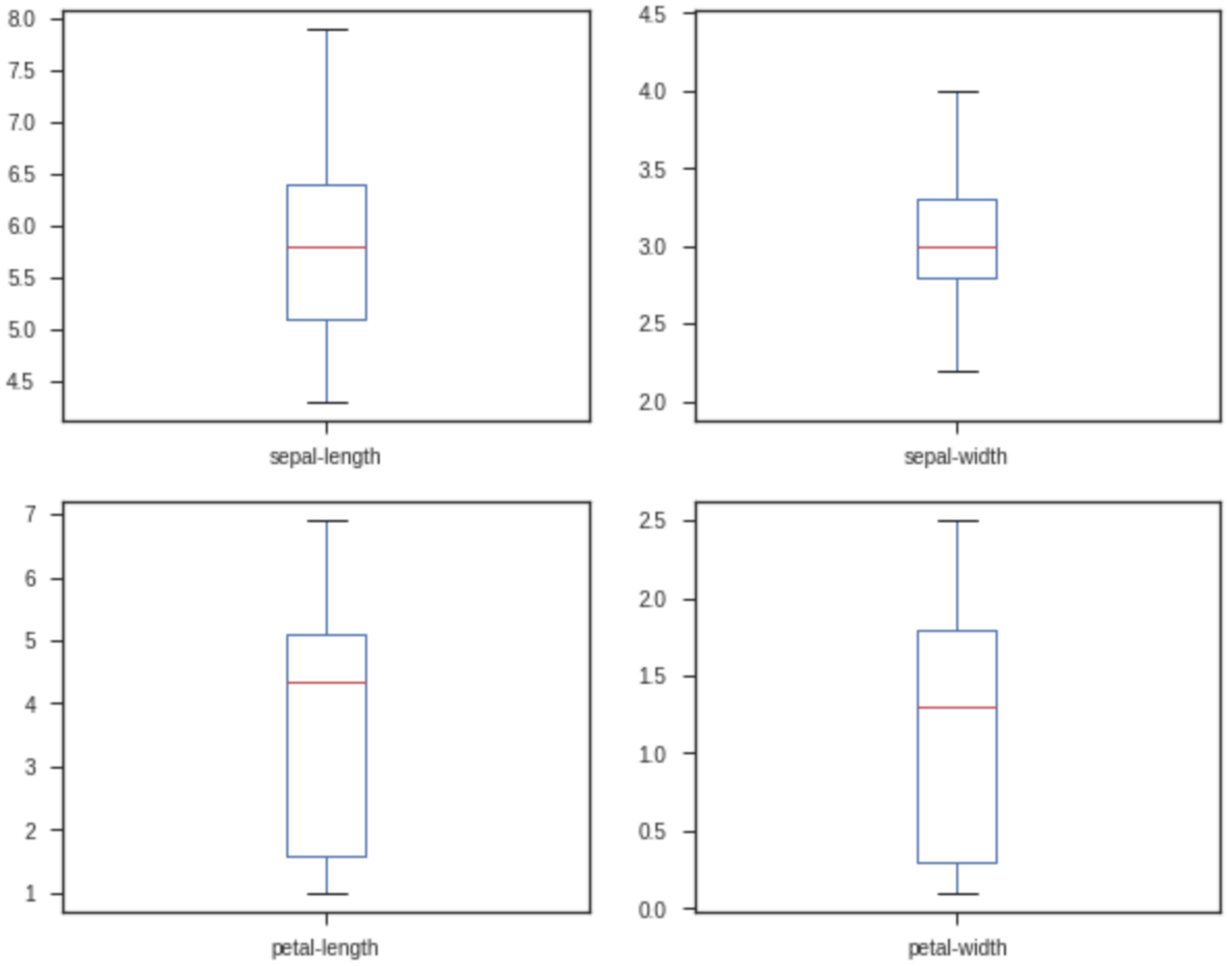
```
print(df.groupby('class').size())
```

```
class
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
dtype: int64
```

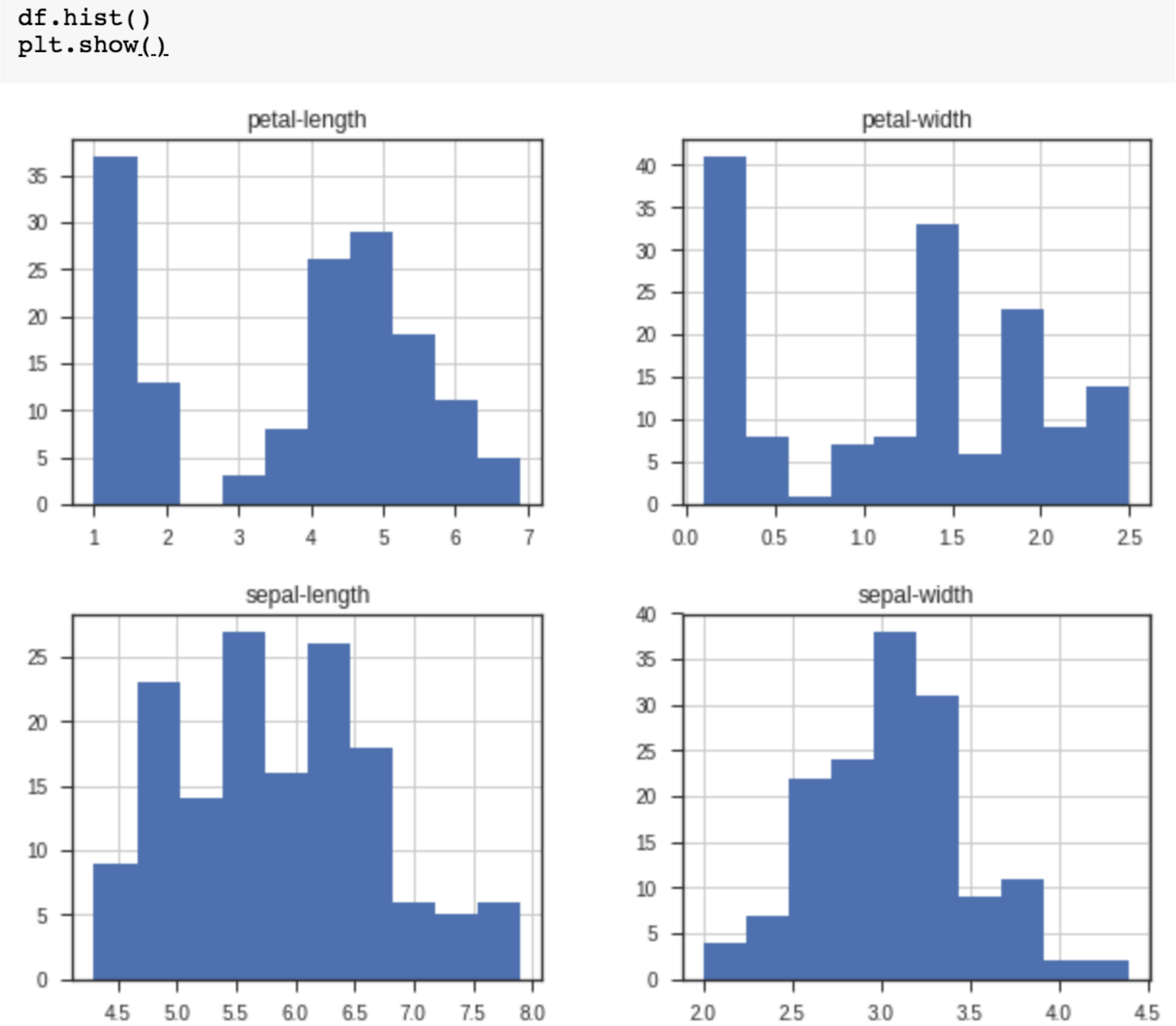
```
plt.rcParams["figure.figsize"] = (10,8)
df.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
```



```
plt.rcParams["figure.figsize"] = (10,8)
df.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)
plt.show()
```

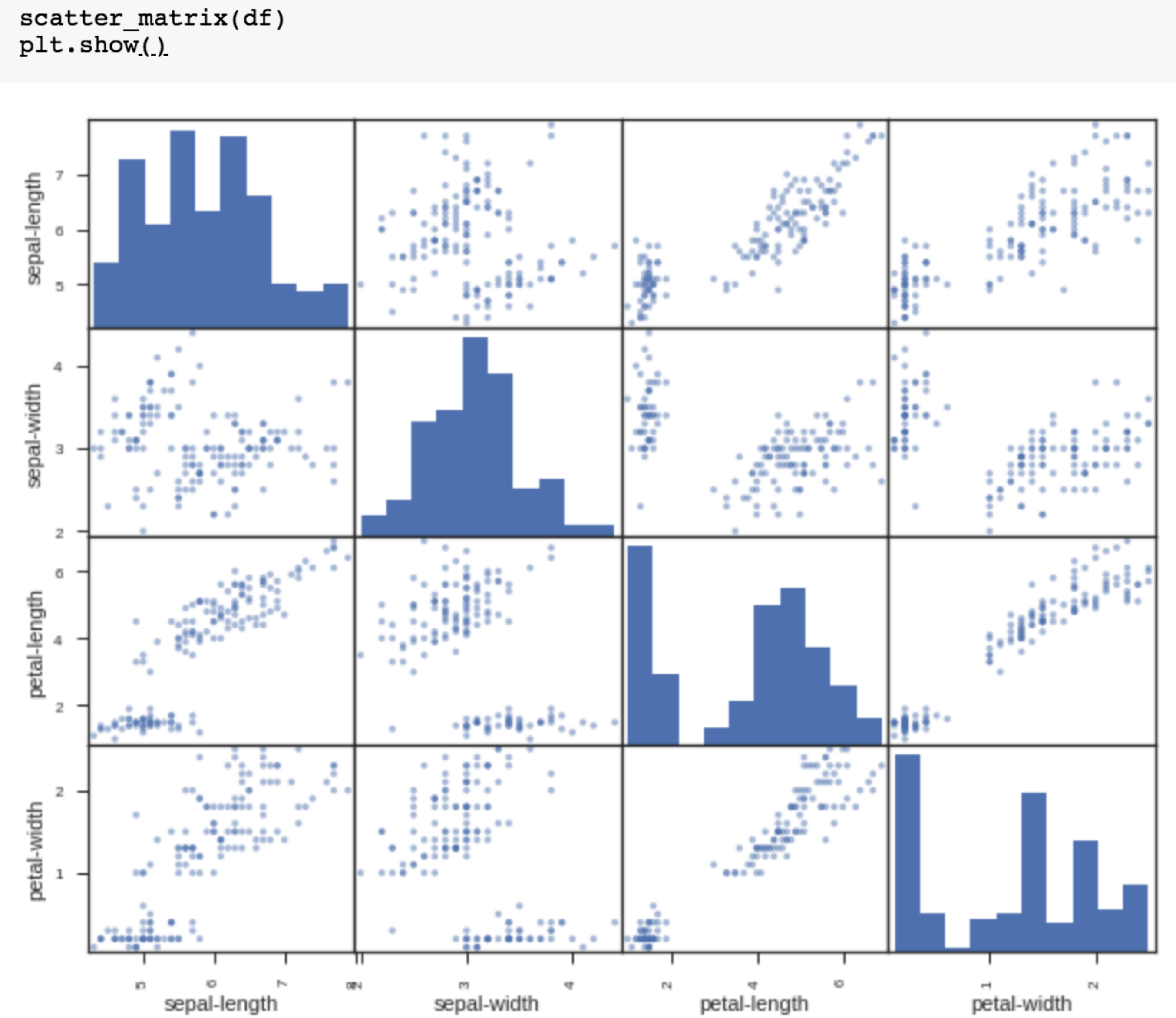


```
df.hist()
plt.show()
```



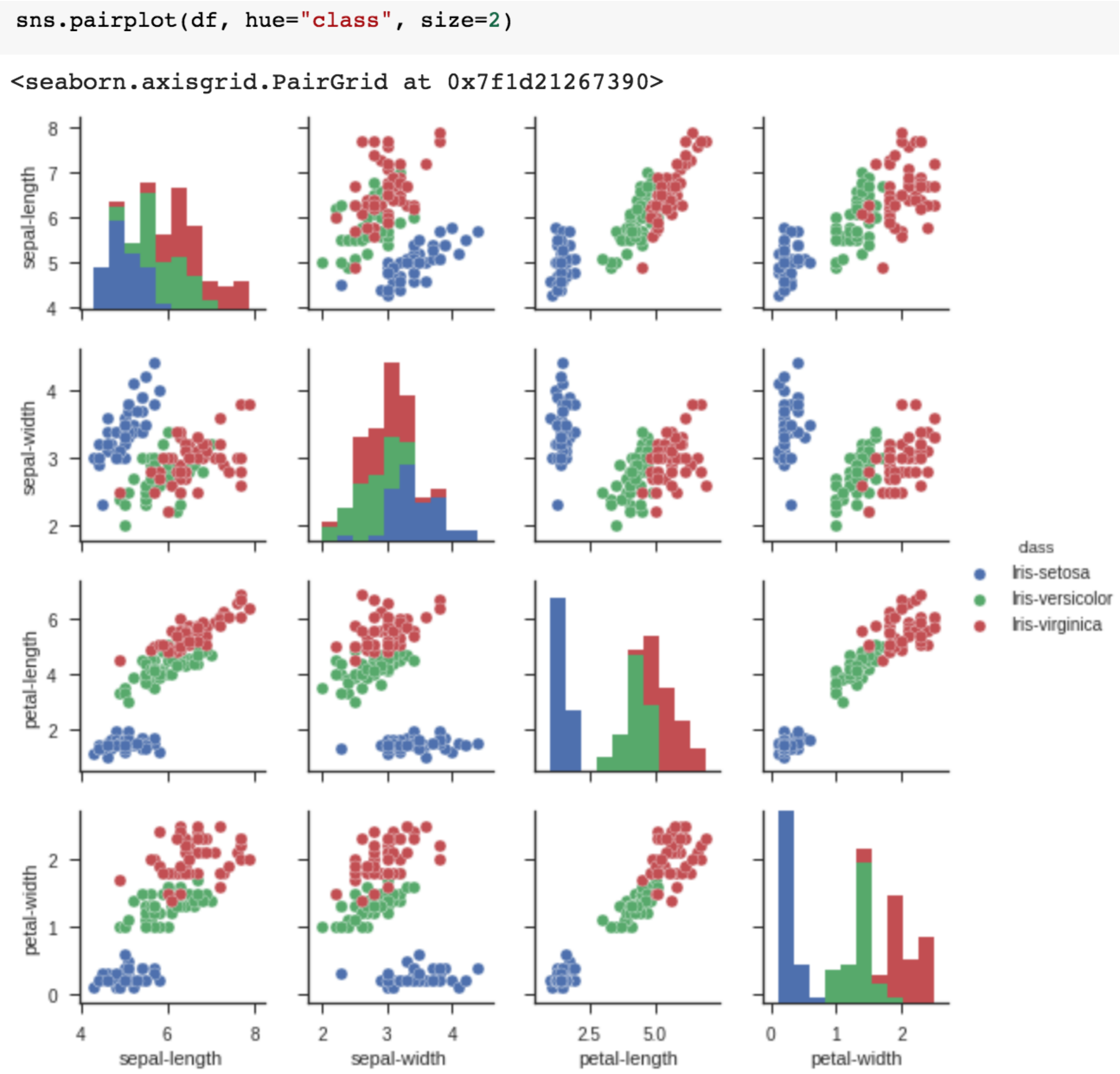
```
df.hist()
plt.show()
```

```
scatter_matrix(df)
plt.show()
```



```
scatter_matrix(df)
plt.show()
```

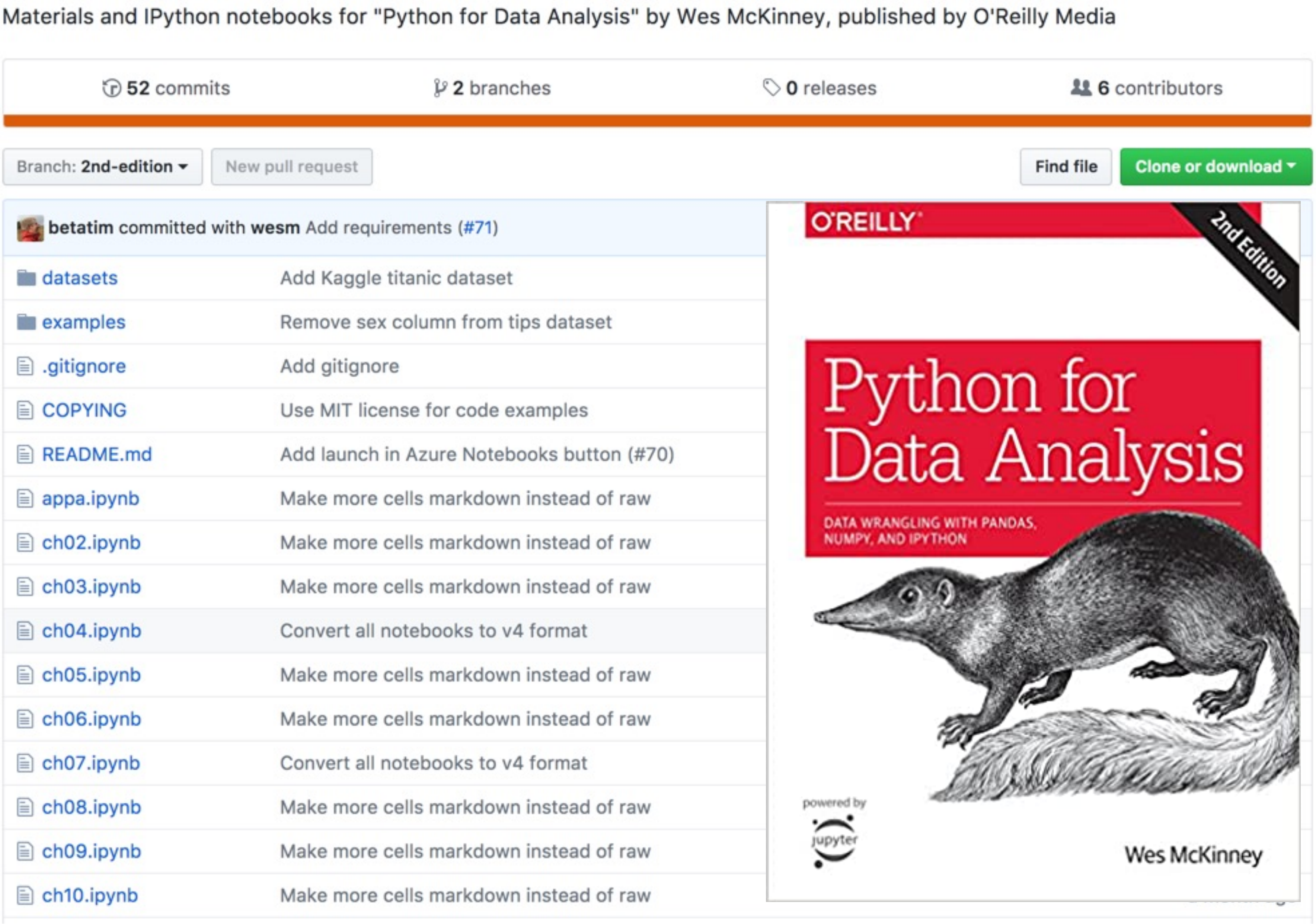
sns.pairplot(df, hue="class", size=2)



```
sns.pairplot(df, hue="class", size=2)
```

```
<seaborn.axisgrid.PairGrid at 0x7f1d21267390>
```

Wes McKinney (2017), "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", 2nd Edition, O'Reilly Media.



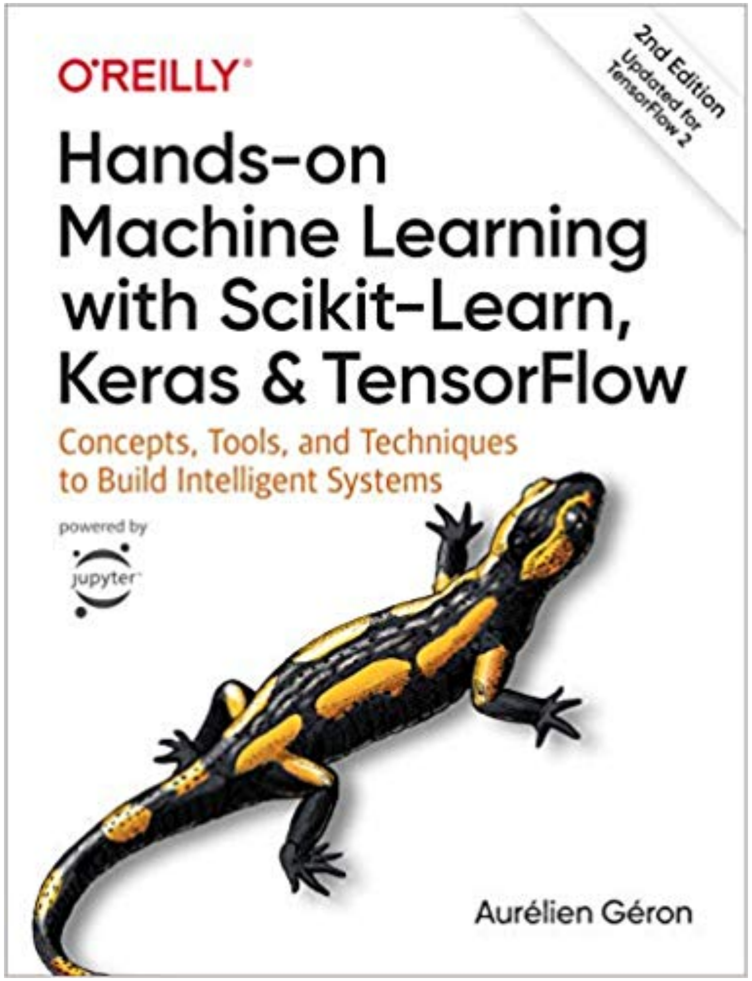
Materials and IPython notebooks for "Python for Data Analysis" by Wes McKinney, published by O'Reilly Media

betatim committed with wesm Add requirements (#71)

datasets	Add Kaggle titanic dataset
examples	Remove sex column from tips dataset
.gitignore	Add gitignore
COPYING	Use MIT license for code examples
README.md	Add launch in Azure Notebooks button (#70)
appa.ipynb	Make more cells markdown instead of raw
ch02.ipynb	Make more cells markdown instead of raw
ch03.ipynb	Make more cells markdown instead of raw
ch04.ipynb	Convert all notebooks to v4 format
ch05.ipynb	Make more cells markdown instead of raw
ch06.ipynb	Make more cells markdown instead of raw
ch07.ipynb	Convert all notebooks to v4 format
ch08.ipynb	Make more cells markdown instead of raw
ch09.ipynb	Make more cells markdown instead of raw
ch10.ipynb	Make more cells markdown instead of raw

https://github.com/wesm/pydata-book

Aurélien Géron (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition O'Reilly Media, 2019


O'REILLY
2nd Edition Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by jupyter
Aurélien Géron

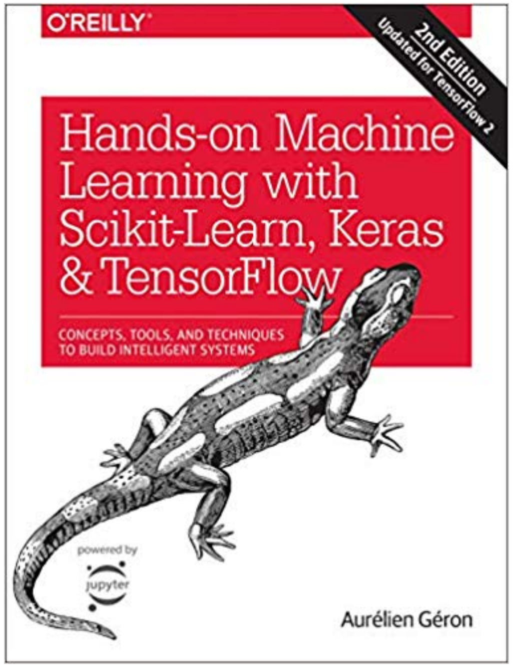
https://github.com/ageron/handson-ml2

Source: https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646/

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Notebooks

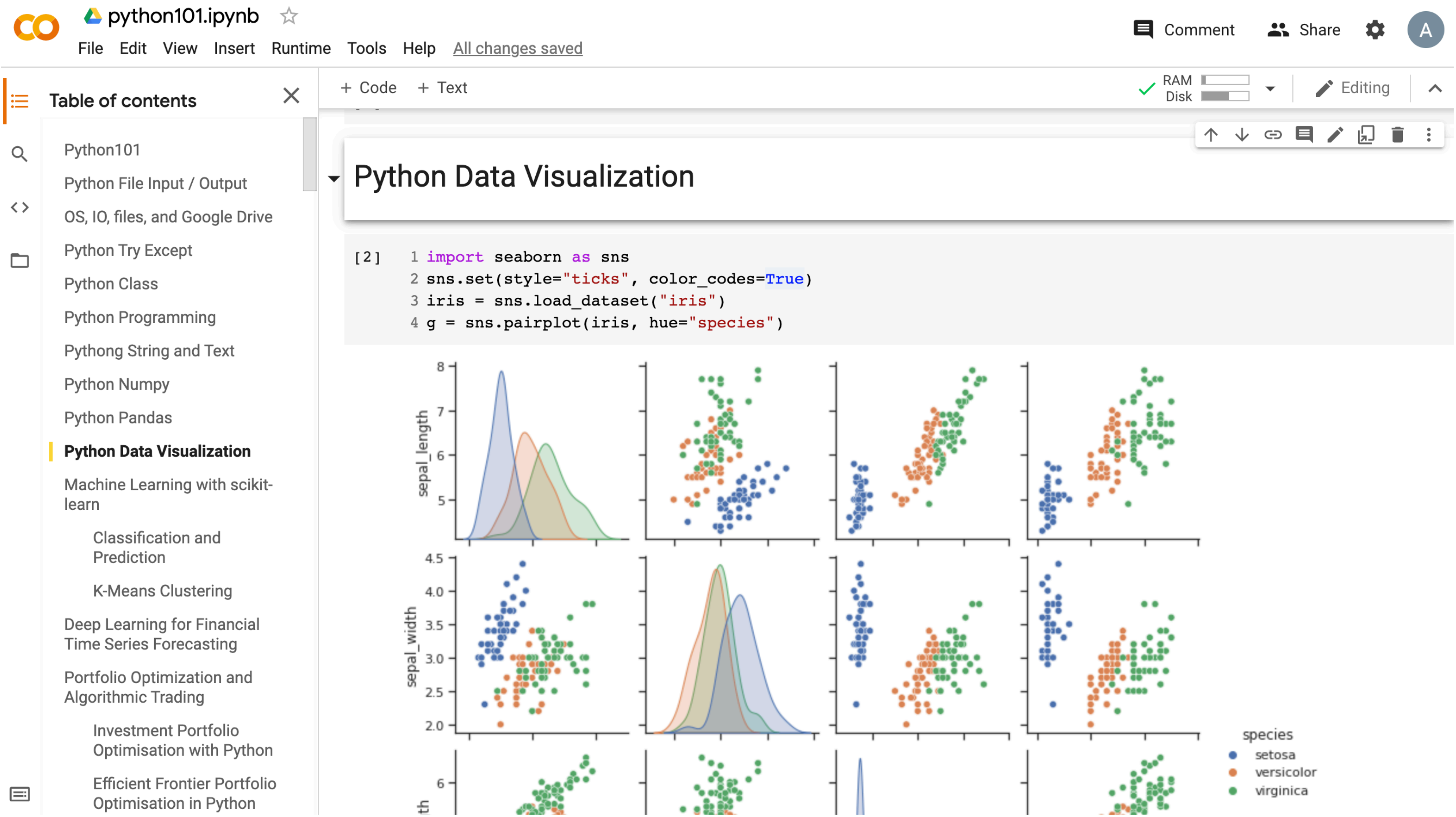
1.The Machine Learning landscape
2.End-to-end Machine Learning project
3.Classification
4.Training Models
5.Support Vector Machines
6.Decision Trees
7.Ensemble Learning and Random Forests
8.Dimensionality Reduction
9.Unsupervised Learning Techniques
10.Artificial Neural Nets with Keras
11.Training Deep Neural Networks
12.Custom Models and Training with TensorFlow
13.Loading and Preprocessing Data
14.Deep Computer Vision Using Convolutional Neural Networks
15.Processing Sequences Using RNNs and CNNs
16.Natural Language Processing with RNNs and Attention
17.Representation Learning Using Autoencoders
18.Reinforcement Learning
19.Training and Deploying TensorFlow Models at Scale


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by
jupyter
Aurélien Géron

https://github.com/ageron/handson-ml2

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT


CO
python101.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
All changes saved
Comment
Share
A
Table of contents
Python101
Python File Input / Output
OS, IO, files, and Google Drive
Python Try Except
Python Class
Python Programming
Pythong String and Text
Python Numpy
Python Pandas
Python Data Visualization
Machine Learning with scikit-learn
Classification and Prediction
K-Means Clustering
Deep Learning for Financial Time Series Forecasting
Portfolio Optimization and Algorithmic Trading
Investment Portfolio Optimisation with Python
Efficient Frontier Portfolio Optimisation in Python
+ Code
+ Text
RAM
Disk
Editing
[2]
sepal_length
sepal_width
8
7
6
5
4.5
4.0
3.5
3.0
2.5
2.0
6
species
setosa
versicolor
virginica

Python Data Visualization

```
import seaborn as sns
sns.set(style="ticks", color_codes=True)
iris = sns.load_dataset("iris")
g = sns.pairplot(iris, hue="species")
```

https://tinyurl.com/aintpupython101

Summary

- Data Science and Data Mining
- Discovering, Analyzing, Visualizing and Presenting Data with Python
  - Pandas
  - Matplotlib
  - Seaborn
  - Plotly
  - Bokeh, Altair

References

- EMC Education Services (2015), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley
- Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
- Robert Layton (2017), Learning Data Mining with Python - Second Edition, Packt Publishing.
- Wes McKinney (2017), "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", 2nd Edition, O'Reilly Media.
- Aurélien Géron (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O'Reilly Media.
  https://github.com/wesm/pydata-book
- Pandas, http://pandas.pydata.org/
- Matplotlib, https://matplotlib.org/
- Seaborn, https://seaborn.pydata.org/
- Plotly, https://plotly.com/python/
- Bokeh, https://bokeh.org/
- Altair, https://altair-viz.github.io/
- Min-Yuh Day (2021), Python 101, https://tinyurl.com/aintpupython101