

**MATH408/MATH524**

**Statistical learning and data mining**

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University

# Lectures

---

- ▶ Lecture: Szu-Chi Chung (鍾思齊)
  - ▶ Office: 理 SC 2002-4
  - ▶ Office hour: Tue. 16:00~18:00 and Wed. 16:00~18:00
- ▶ T.A.: 蔡承翰, 王政德
  - ▶ Office: SC2008, SC2003-1
  - ▶ TA hour: Mon. 10:00~12:00
- ▶ Class hours: Tue. (9:10-12:00)
  - ▶ Classroom: 理 SC-0014
- ▶ Facebook

## Textbook and requirement

---

- ▶ Textbook: *An Introduction to Statistical Learning with Applications in R*
  - ▶ Authors: James, Witten, Hastie, and Tibshirani
  - ▶ <https://www.statlearning.com/>
- ▶ For a more advanced treatment of these topics: Reference book: *The Elements of Statistical Learning*
  - ▶ Authors: Hastie, Tibshirani and Friedman
  - ▶ <https://web.stanford.edu/~hastie/ElemStatLearn/>
- ▶ For the programming patterns: Reference book: *Practical Statistics for Data Scientists 50+ Essential Concepts Using R and Python*
  - ▶ Authors: Peter Bruce, Andrew Bruce and Peter Gedeck
  - ▶ <https://github.com/gedeck/practical-statistics-for-data-scientists>

## Textbook and requirement

---

- ▶ Slides and videos for Statistical Learning MOOC by Hastie and Tibshirani
  - ▶ <https://www.dataschool.io/15-hours-of-expert-machine-learning-videos/>
- ▶ For the exercises of each chapter, there is a GitHub repository of solutions provided by students you can use to check your work.
  - ▶ <http://blog.princehonest.com/stat-learning/>
- ▶ Programming language: Python
  - ▶ You are asked to use python to implement the assignment, midterm and final
  - ▶ Since it is the most popular language in the field of data science
  - ▶ It is free and easy to learn
  - ▶ The lab and related material will be available in the course website
    - ▶ <https://phonchi.github.io/nsysu-math524/>

# Grading policy

---

## ▶ Grading

- ▶ Weekly Homework 35% (Both conceptual and coding part)
- ▶ Midterm exam 35% (Mostly will be coding parts)
- ▶ Final project 30% (You are free to choose any dataset for analysis)

## ▶ Midterm

- ▶ Will be held on 11/16

## ▶ Term project:

- ▶ Organize a team of 2 persons
- ▶ Presentation will be held on 1/11 and 1/18
- ▶ Must hand in a report
- ▶ Score will be the summation of students (10%), TA(10%) and lecturer (10%).

# Dataset and competition

---

## ▶ Dataset

### ▶ 政府資料開放平台

▶ <https://data.gov.tw/>

### ▶ Kaggle

▶ <https://www.kaggle.com/datasets>

### ▶ Google dataset search

▶ <https://datasetsearch.research.google.com/>

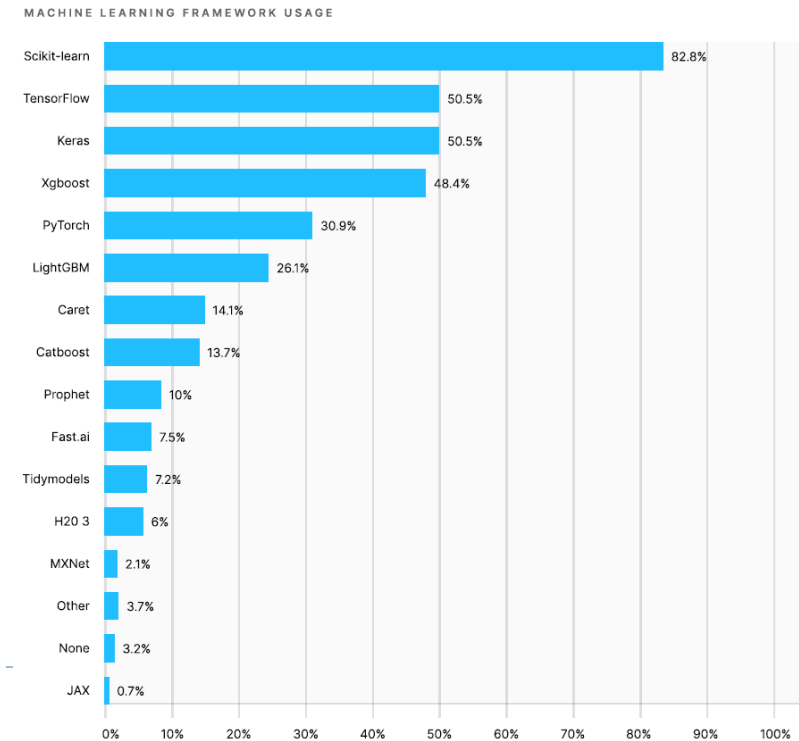
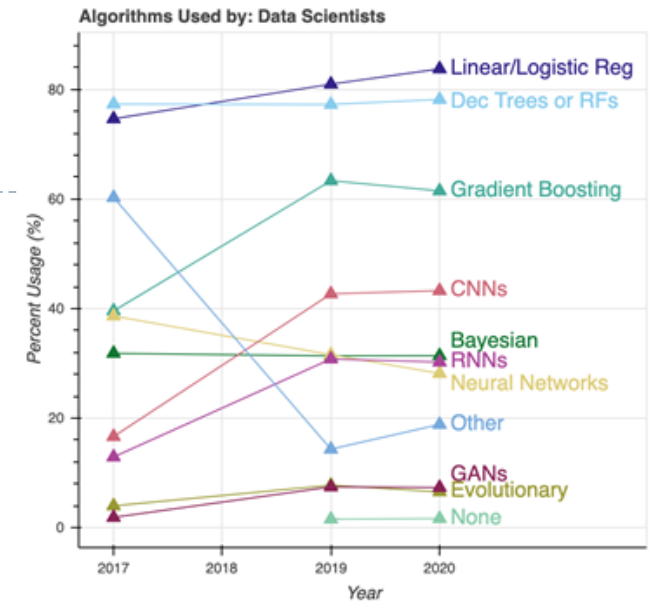
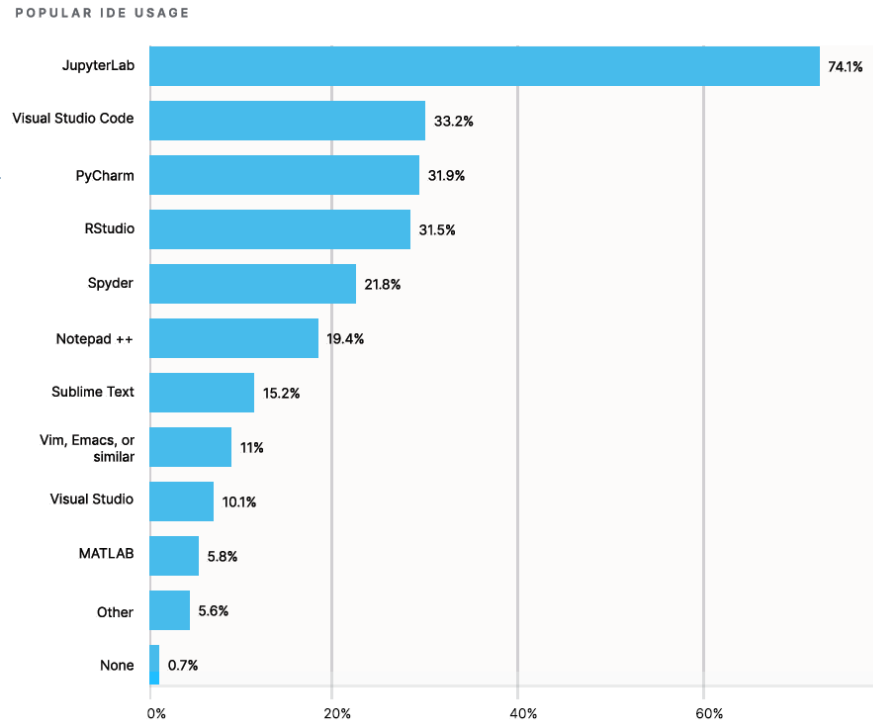
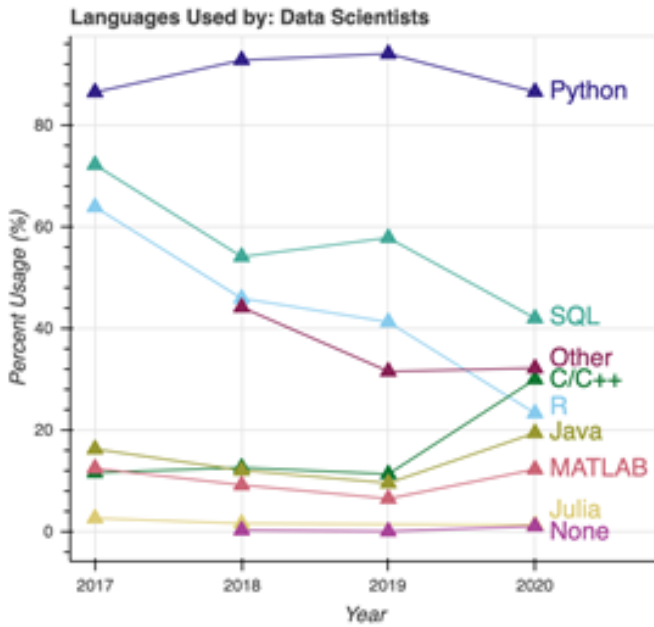
## ▶ Competition

### ▶ Kaggle

▶ <https://www.kaggle.com/competitions>

### ▶ Tbrain

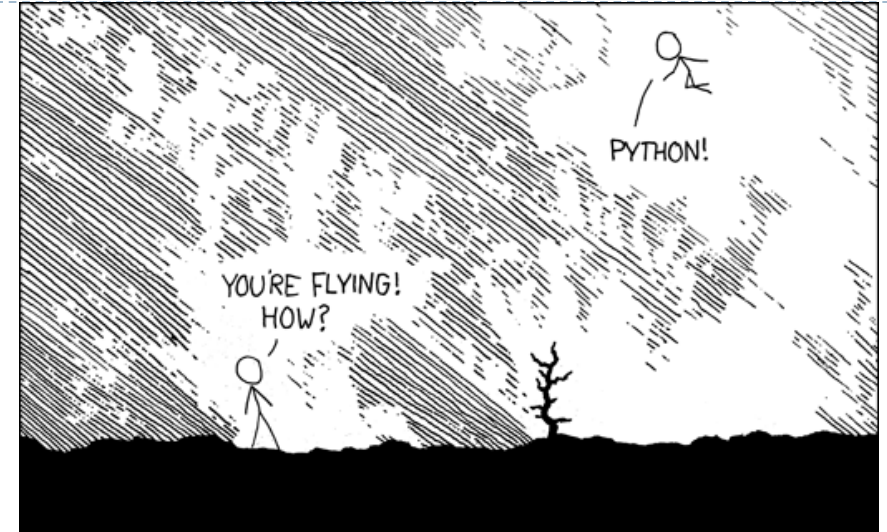
▶ <https://tbrain.trendmicro.com.tw/>



► <https://www.kaggle.com/kaggle-survey-2020>

# Learning Python

- ▶ Python
  - ▶ [Learn X in Y minutes](#)
  - ▶ [Kaggle Python tutorial](#)
  - ▶ [Python for Everybody](#)
  - ▶ [Python 台灣社群](#)
- ▶ Python scientific computing
  - ▶ <https://scipy-lectures.org/>
- ▶ Python for R and Matlab users
  - ▶ <http://mathesaurus.sourceforge.net/r-numpy.html>
  - ▶ <https://numpy.org/doc/stable/user/numpy-for-matlab-users.html>

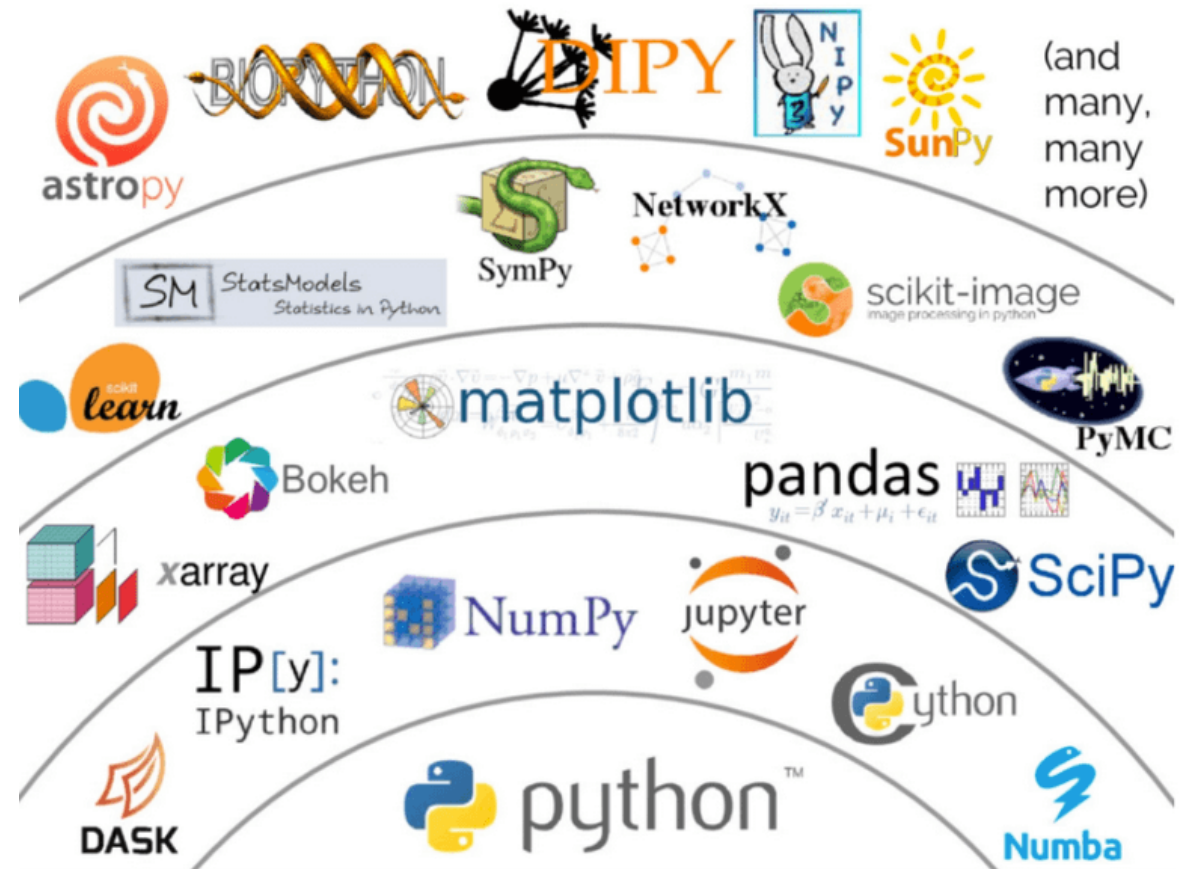


<https://xkcd.com/353/>



# The Pydata Stack

- ▶ In 2017, [a keynote at PyCon](#) presented a schematic of the scientific Python stack
  - ▶ Project [Jupyter](#) and [IPython](#) for interactive computing and IDEs
  - ▶ [NumPy](#) for numerical array computing
    - ▶ [Numba](#) for just-in-time compilation
    - ▶ [Cython](#) for ahead-of-time compilation
  - ▶ [Pandas](#) for dataframe (Labeled array)
  - ▶ [Scikit-learn](#) and [Statsmodel](#) for modeling
  - ▶ [Seaborn](#) for visualization
- ▶ Install Anaconda
  - ▶ <https://www.anaconda.com/products/individual>



Source: <https://coiled.io/pydata-dask/>

# Environment

---

- ▶ Jupyter notebook
  - ▶ Colab - <https://colab.research.google.com/>
  - ▶ Kaggle - <https://www.kaggle.com/docs/notebooks>
  - ▶ Jupyterlab - <https://www.anaconda.com/products/individual>
- ▶ Markdown
  - ▶ Learning
    - ▶ <https://commonmark.org/>
    - ▶ <https://learnxinyminutes.com/docs/markdown/>
  - ▶ Usage
    - ▶ <https://hackmd.io/>
    - ▶ <https://github.com/>
    - ▶ Jupyter notebook

## Related to other course

---

- ▶ More theory
  - ▶ Mathematical statistics
  - ▶ Principles of artificial intelligence
- ▶ More accurate prediction and case study
  - ▶ Machine learning
  - ▶ Deep learning
  - ▶ Data science capstone project
- ▶ Apply to specific domain
  - ▶ Analysis of financial time series
  - ▶ Survival analysis
- ▶ High performance computing, Database systems...



# Introduction

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University

# Statistical Learning In The News

---

- ▶ “Learning from its mistakes”, Watson's software is wired for more than handling natural language processing. David Ferrucci (PI of Watson DeepQA technology for IBM Research), 2011



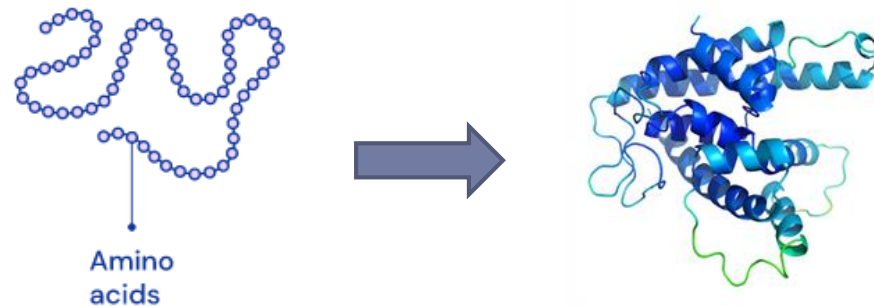
- ▶ “I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative”. Lee Sedol (Winner of 18 World Go Titles). 2016



# Statistical Learning In The News

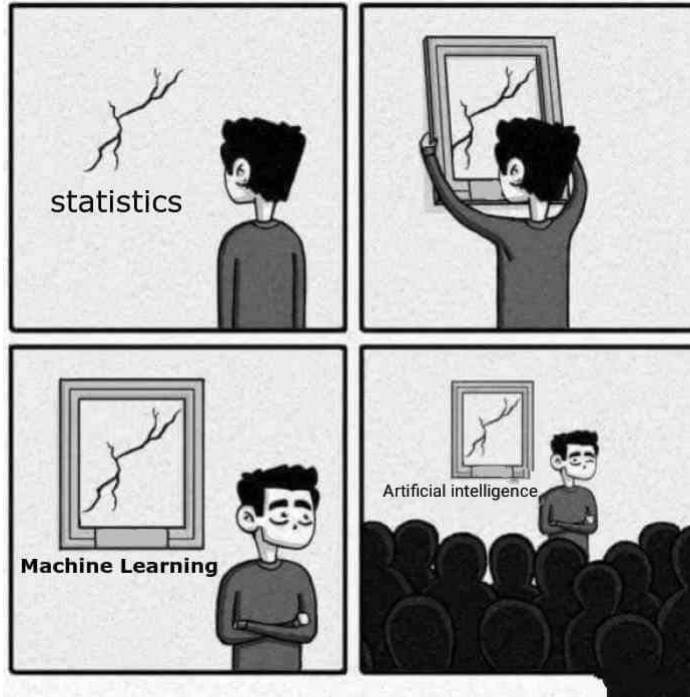
---

- ▶ “We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we’d ever get there, is a very special moment.” John Moult (Co-founder and Chair of CASP, University Of Maryland). 2020



- ▶ It's learning allows the computer to become smarter as it tries to answer questions - and to learn as it gets them right or wrong.

# What is the difference between statistics, machine learning, data mining, statistical learning, AI....?

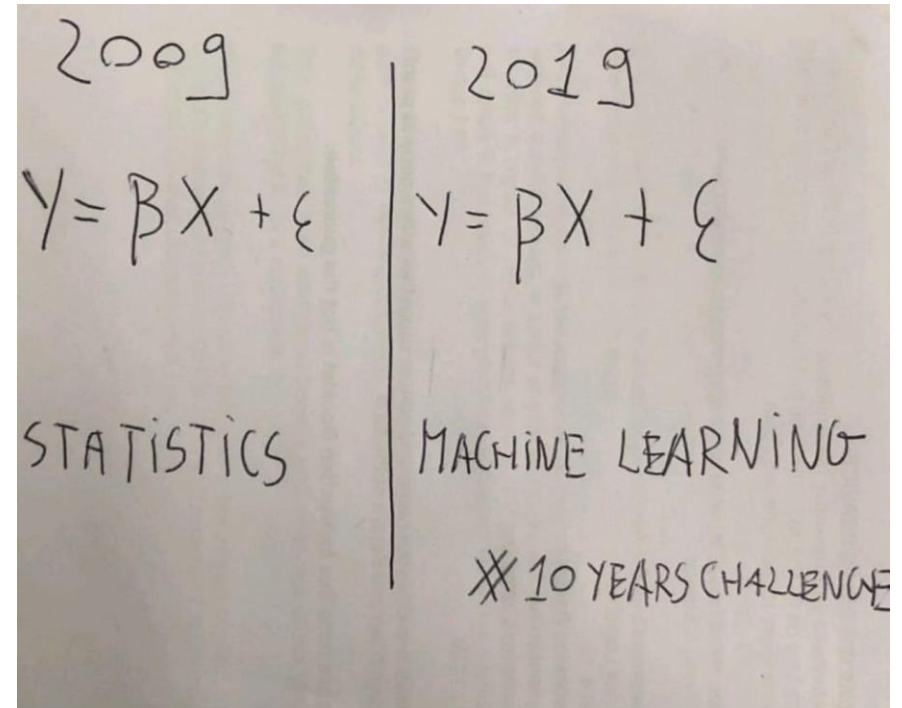


<https://laptrinhx.com/what-actually-is-artificial-intelligence-a-beginners-guide-249021669/>



Let's see who you really are  
machine learning

<https://medium.com/analytics-vidhya/statistics-in-machine-learning-a1eb88b88da2>



<https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3>

# Statistical Learning versus Machine Learning – textbook author's view

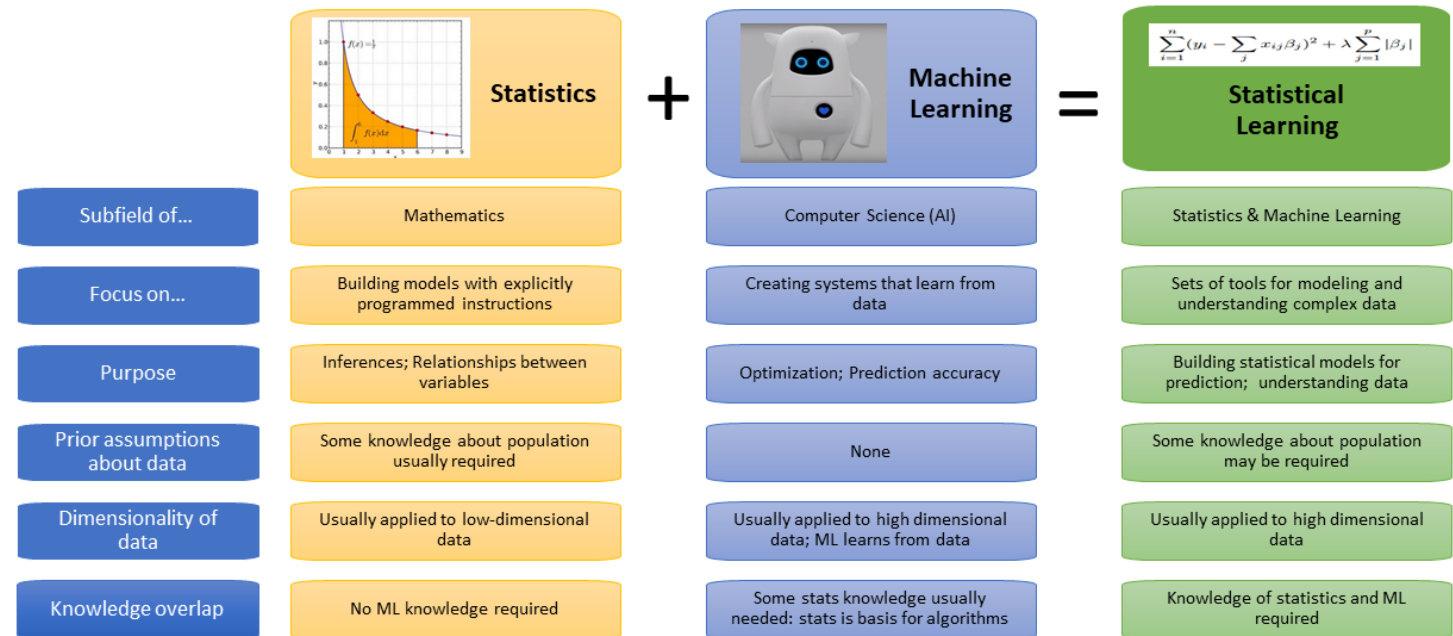
---

- ▶ Machine learning arose as a subfield of Artificial Intelligence.
- ▶ Statistical learning arose as a subfield of Statistics.
- ▶ There is much overlap - both fields focus on supervised and unsupervised problems:
  - ▶ Machine learning has a greater emphasis on large-scale applications and prediction accuracy.
  - ▶ Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
  - ▶ But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- ▶ Machine learning put more focus on the use of computational power to solve a problem



# Statistical Learning versus Machine Learning – a personal view

- ▶ *Statistical learning*, the use of machine learning and statistics techniques with most of the goal is statistical inference: drawing conclusions on the data at hand.

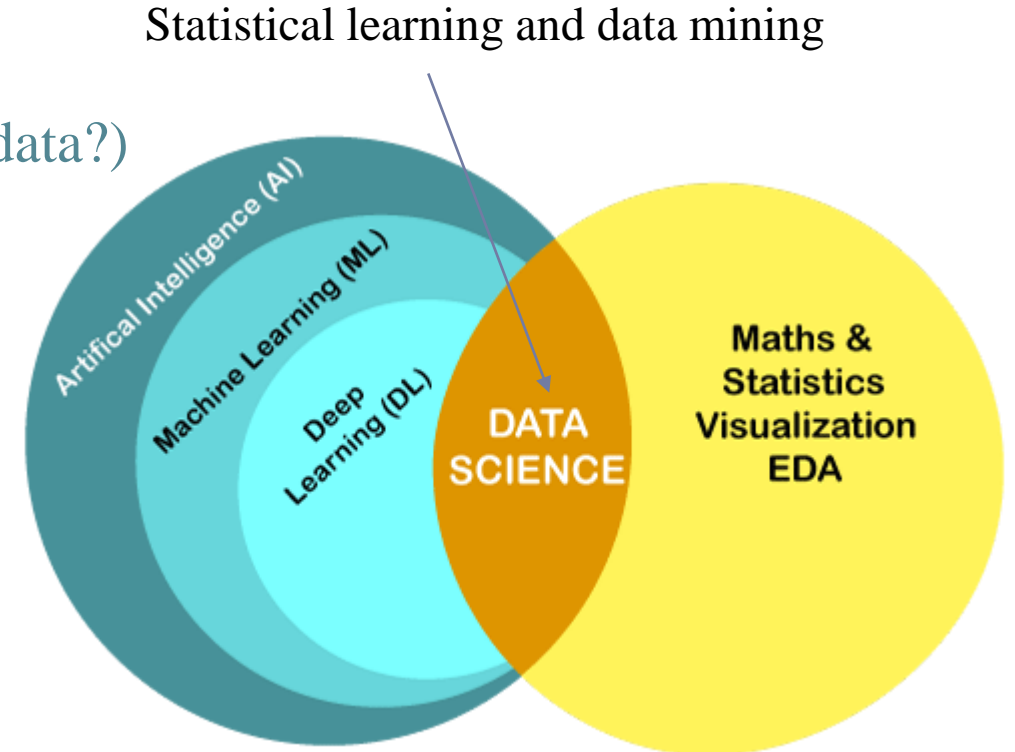


Musio image: Akawikipic [CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>)]

<https://www.datasciencecentral.com/profiles/blogs/machine-learning-vs-statistics-in-one-picture>

# Data Scientist: The Sexiest Job of the 21st Century

- ▶ The shortage of data scientists is becoming a serious constraint in some sectors.
- ▶ Data scientists today are akin to the Wall Street “quants” of the 1980s and 1990s.
- ▶ Data analysis is a process of
  - ▶ Inspecting data (How do we deal with missing data?)
  - ▶ Cleaning data
  - ▶ Transforming data
  - ▶ Modeling data
- ▶ With the goal of
  - ▶ Discovering useful info
  - ▶ Suggesting conclusion
  - ▶ Supporting decision making



<https://www.javatpoint.com/data-science-vs-machine-learning>

# Top 10 Ideas in Statistics That Have Powered the AI Revolution

---

1. Hirotugu Akaike (1973). [Information Theory and an Extension of the Maximum Likelihood Principle](#). *Proceedings of the Second International Symposium on Information Theory*.
2. John Tukey (1977). [Exploratory Data Analysis](#).
3. Grace Wahba (1978). [Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression](#). *Journal of the Royal Statistical Society*.
4. Bradley Efron (1979). [Bootstrap Methods: Another Look at the Jackknife](#). *Annals of Statistics*.
5. Alan Gelfand and Adrian Smith (1990). [Sampling-based Approaches to Calculating Marginal Densities](#). *Journal of the American Statistical Association*.
6. Guido Imbens and Joshua Angrist (1994). [Identification and Estimation of Local Average Treatment Effects](#). *Econometrica*.
7. Robert Tibshirani (1996). [Regression Shrinkage and Selection Via the Lasso](#). *Journal of the Royal Statistical Society*.
8. Leland Wilkinson (1999). [The Grammar of Graphics](#).
9. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). [Generative Adversarial Networks](#). *Proceedings of the International Conference on Neural Information Processing Systems*.
10. Yoshua Bengio, Yann LeCun, and Geoffrey Hinton (2015). [Deep Learning](#). *Nature*.



# The Supervised Learning Problem

---

- ▶ **Starting point:**
  - ▶ Outcome measurement  $Y$  (also called the dependent variable, response, target).
  - ▶ Vector of  $p$  predictor measurements  $X$  (also called inputs, regressors, covariates, features, independent variables).
  - ▶ In the regression problem,  $Y$  is quantitative (e.g., price, blood pressure).
  - ▶ In the classification problem,  $Y$  takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
  - ▶ We have training data  $(x_1, y_1), \dots, (x_n, y_n)$ . These are observations (examples, instances) of these measurements.

# Objectives and Philosophy

---

- ▶ On the basis of the training data, we would like to:
  - ▶ Accurately predict unseen test cases.
  - ▶ Understand which inputs affect the outcome and how.
  - ▶ Assess the quality of our predictions and inferences.
- ▶ It is important to understand the ideas behind the various techniques in order to know how and when to use them.
  - ▶ One has to understand the simpler methods first in order to grasp the more sophisticated ones.
  - ▶ It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
  - ▶ This is an exciting research area, having important applications in science, industry and finance.
  - ▶ Statistical learning is a fundamental ingredient in the training of a modern data scientist.

# The Unsupervised Learning Problem

---

- ▶ No outcome variable, just a set of predictors (features) measured on a set of samples.
- ▶ Objective is more fuzzier - find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- ▶ Difficult to know how well you are doing.
- ▶ Different from supervised learning, but can be useful as a pre-processing step for supervised learning.

## The Netflix prize

---

- ▶ Competition started in October 2006. Training data is ratings for 18,000 movies by 400,000 Netflix customers, each rating between 1 and 5.
- ▶ Training data is very sparse - about 98% missing.
- ▶ Objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- ▶ Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.
- ▶ Is this a supervised or unsupervised problem?

# Recommendation System

---

- ▶ A recommendation system can use supervised or unsupervised learning; it is neither of them because it's a concept at a different level.
- ▶ A recommendation system can:
  - ▶ Use **supervised learning** to classify items into elements to be recommended/not recommended.
  - ▶ “Supervised” because it works with labeled data: user profiles: past items, ratings,...

Or

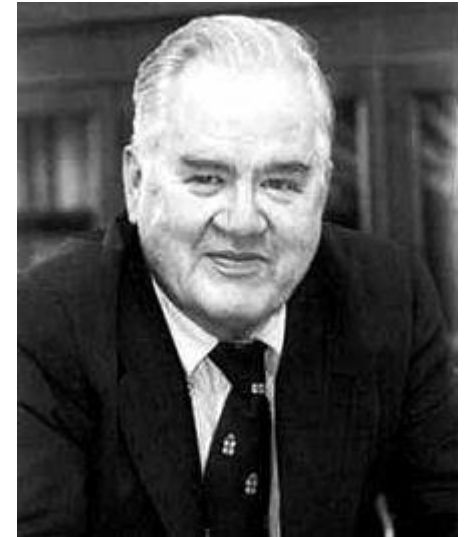
- ▶ Use **unsupervised learning** to make sense of the user-item feature space.
- ▶ For instance, performing clustering analysis or PCA to understand the dataset.



# Exploratory Data Analysis (EDA) and Data Mining

---

- ▶ The field of exploratory data analysis was established with Tukey's 1977 now-classic book *Exploratory Data Analysis* [Tukey-1977]. Tukey presented simple plots (e.g., boxplots, scatterplots) that, along with summary statistics (mean, median, quantiles, etc.), help paint a picture of a data set.
  - ▶ It is important to understand what you can do before you learn to measure how well you seem to have done it
  - ▶ Allow the data to speak for themselves before standard assumptions or formal modeling
  - ▶ The greatest value of a picture is when it forces us to notice what we never expected to see

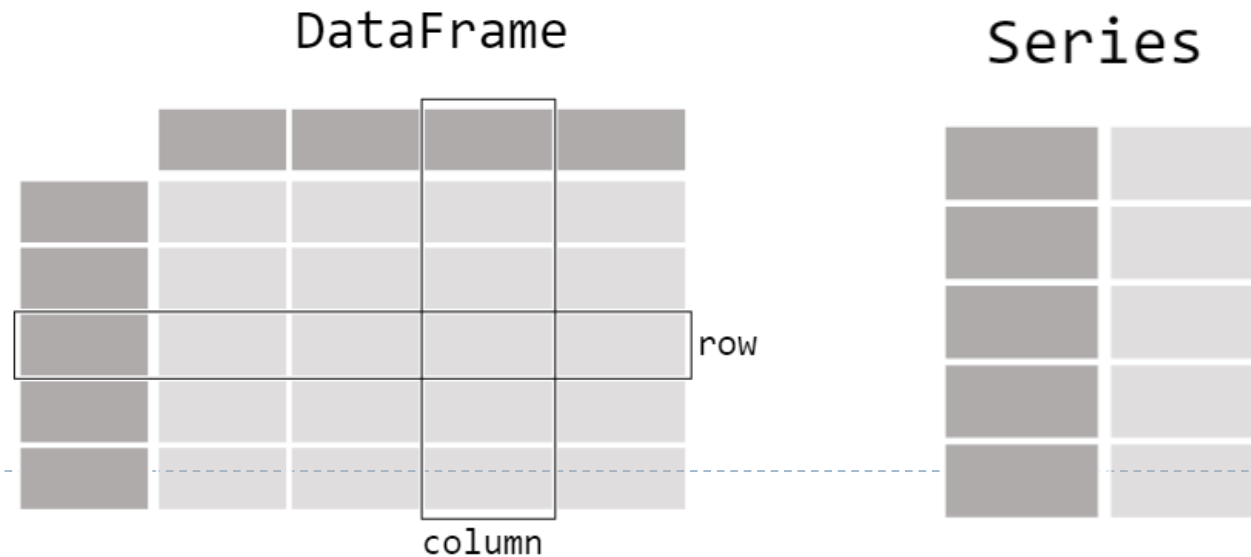


[https://en.wikipedia.org/wiki/John\\_Tukey](https://en.wikipedia.org/wiki/John_Tukey)

# DataFrame

---

- ▶ It is a 2-dimensional data structure that can store data of different types (including characters, integers, floating-point values, categorical data and more) in columns.
- ▶ It is similar to a spreadsheet, a SQL table or the data.frame in R.
  - ▶ [https://pandas.pydata.org/docs/getting\\_started/index.html](https://pandas.pydata.org/docs/getting_started/index.html)
  - ▶ [https://pandas.pydata.org/Pandas\\_Cheat\\_Sheet.pdf](https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)
- ▶ Rows indicating records (cases) and columns indicating features (variables);



# Some common statistics

---

## ▶ Operation on Dataframe

### ▶ Estimate of location

- ▶ Mean, median, percentile, mode

### ▶ Estimate of variability

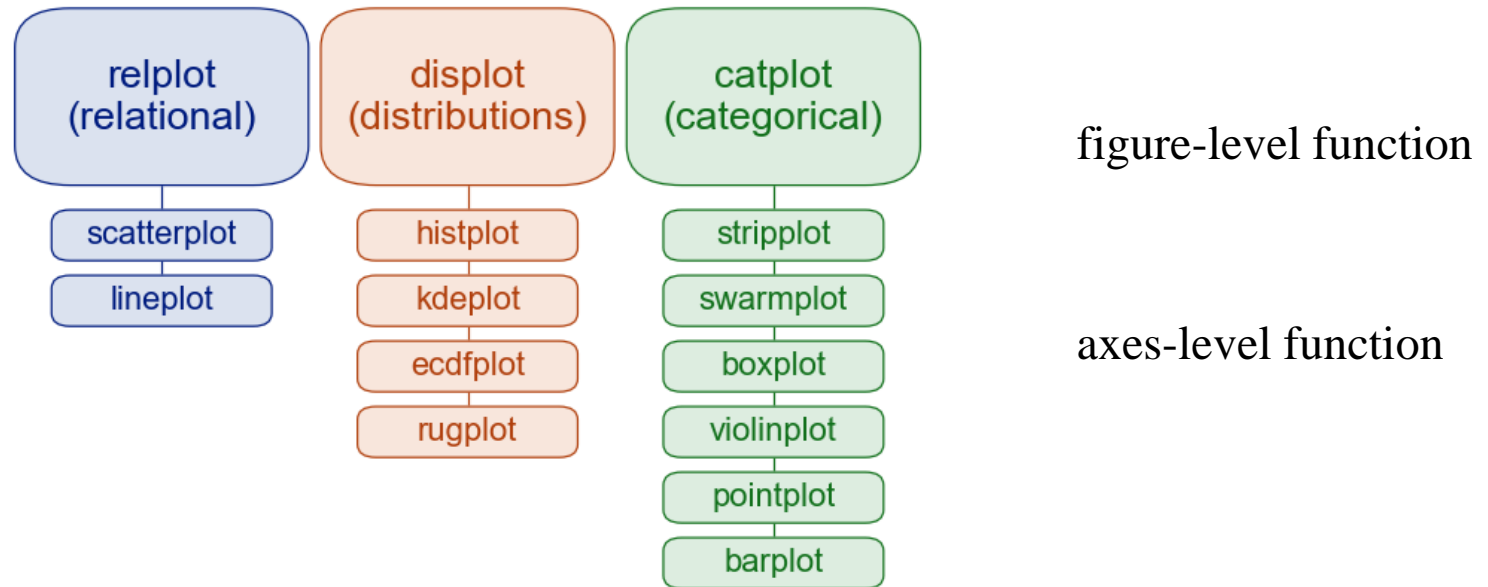
- ▶ Variance, interquartile range (IQR, difference between 25<sup>th</sup> and 75<sup>th</sup> percentile), Mean (median) absolute deviation

### ▶ Filtering, reshaping and combining

# Visualization

---

- ▶ Seaborn combines simple statistical fits with plotting on pandas dataframes.

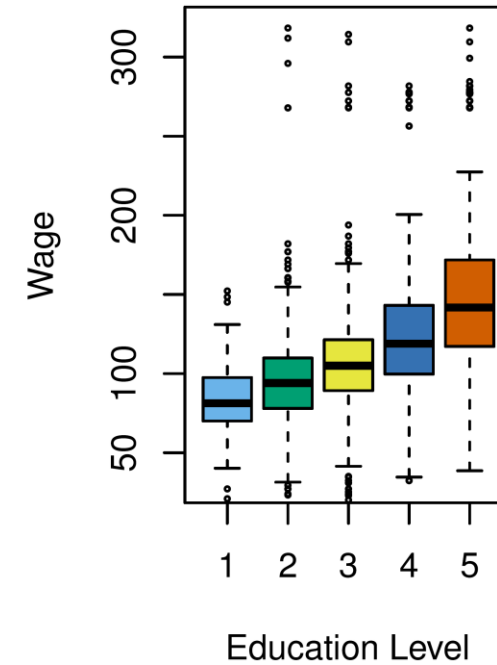
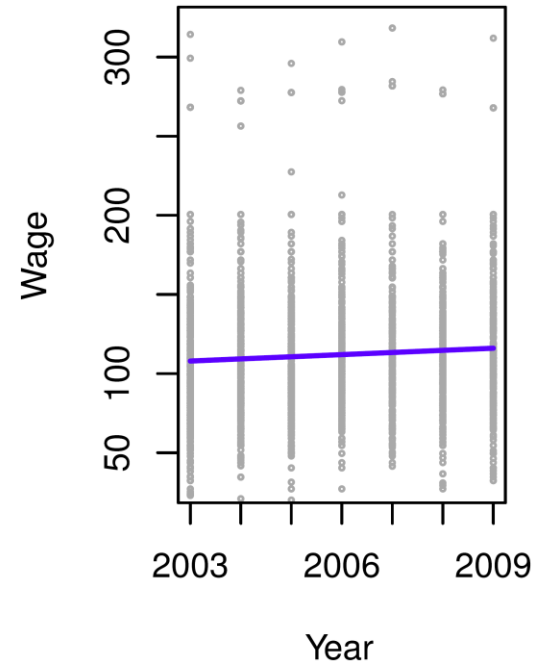
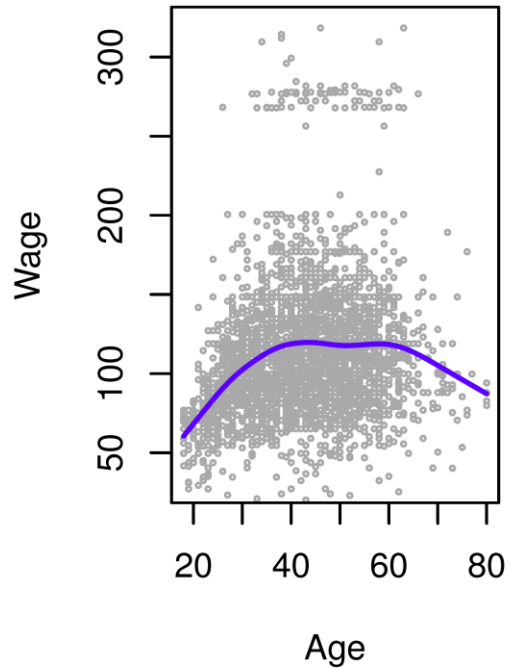


- ▶ Multiple plot - joinplot and pairplot
- ▶ Regression plot – lmpplot, regplot and residplot
- ▶ Matrix plot – heatmap and clusterplot

Name	Description	n	p
<a href="#">Advertising</a>	Sales in different markets, together with advertising budgets in different media channels	200 (with index)	4
Auto	Gas mileage, horsepower, and other information for cars.	392 (with index)	9
Bikeshare	Hourly usage of a bike sharing program in Washington, DC.	8,645 (with index)	15
Boston	Housing values and other information about Boston census tracts.	506 (with index)	13
BrainCancer	Survival times for patients diagnosed with brain cancer.	88 (with index)	8
Caravan	Information about individuals offered caravan insurance.	5,822	86
Carseats	Information about car seat sales in 400 stores.	400	11
College	Demographic characteristics, tuition, and more for USA colleges.	777 (with college name)	18
Credit	Information about credit card debt for 10,000 customers.	400	11
Default	Customer default records for a credit card company.	10,000	4
Fund	Returns of 2,000 hedge fund managers over 50 months.	2,000 (transpose)	50
Hitters	Records and salaries for baseball players.	322	20
Khan	Gene expression measurements for four cancer types.	63 (with index, test in other file)	2,308
NCI60	Gene expression measurements for 64 cancer cell lines.	64 (with index, vector in other file)	6,830
NYSE	Returns, volatility, and volume for the New York Stock Exchange.	6051 (with index)	6
OJ	Sales information for Citrus Hill and Minute Maid orange juice.	1,070	18
Portfolio	Past values of financial assets, for use in portfolio allocation.	100	2
Publication	Time to publication for 244 clinical trials.	244 (with index)	9
Smarket	Daily percentage returns for S&P 500 over a 5-year period.	1,250	9
USArrests	Crime statistics per 100,000 residents in 50 states of USA.	50 (with state name)	4
Wage	Income survey data for men in central Atlantic region of USA.	3,000	11
Weekly	1,089 weekly stock market returns for 21 years.	1,089	9

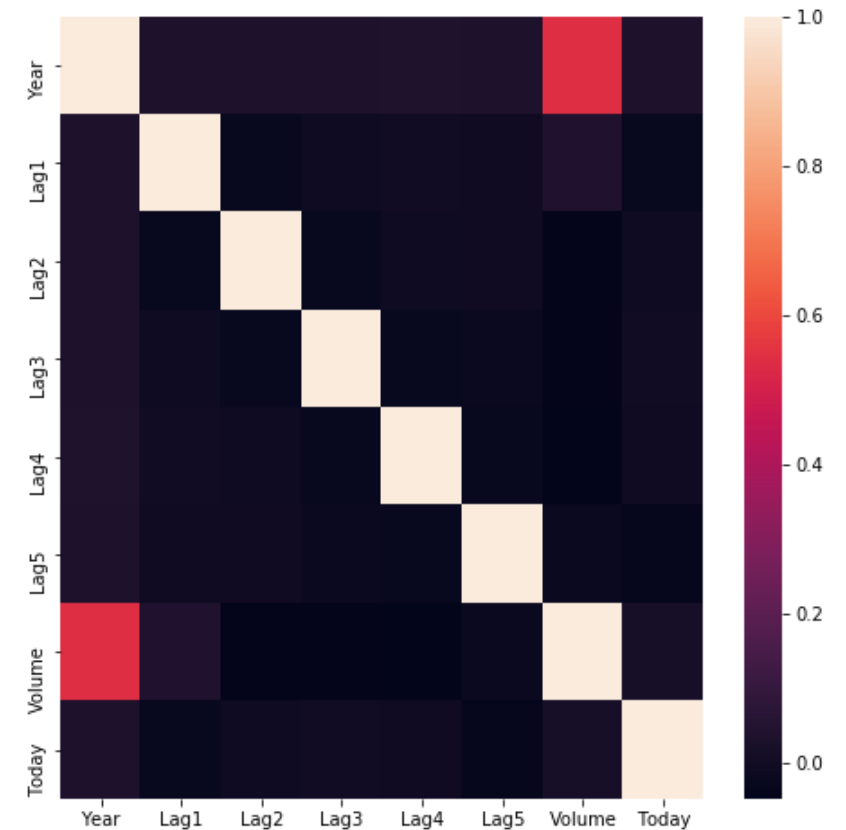
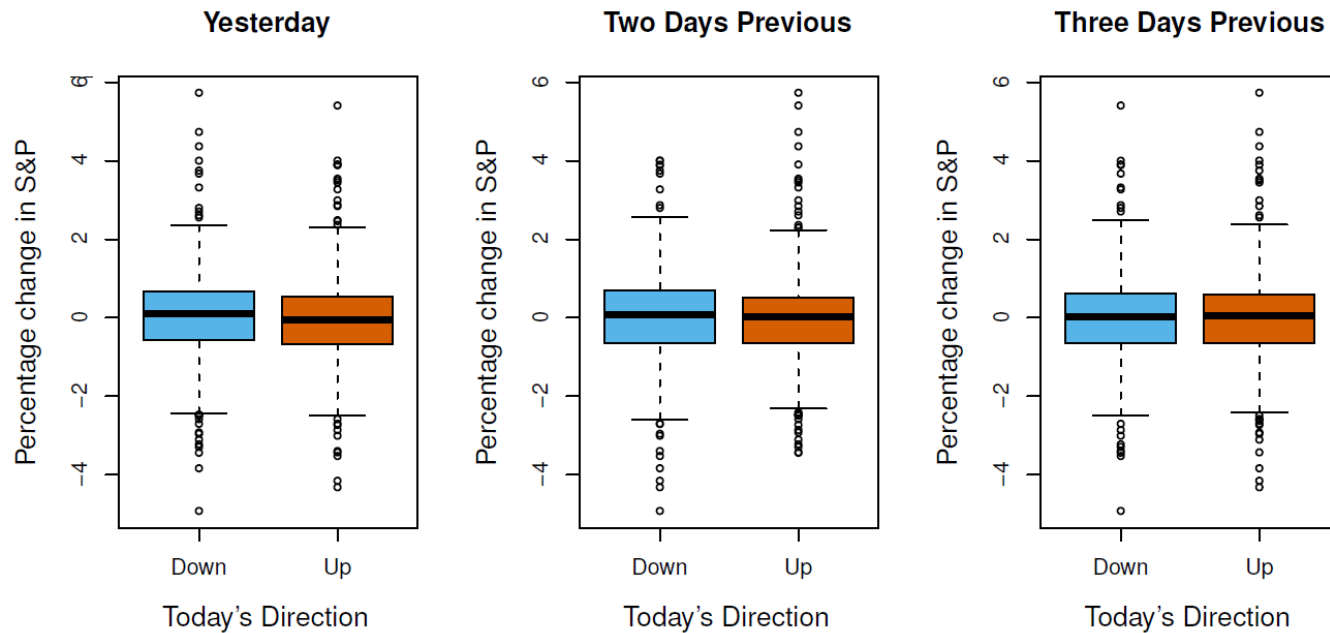
# Wage data

- ▶ Wage data for a group of 3,000 male workers in the Mid-Atlantic region.
  - ▶ Scatterplot and Boxplot



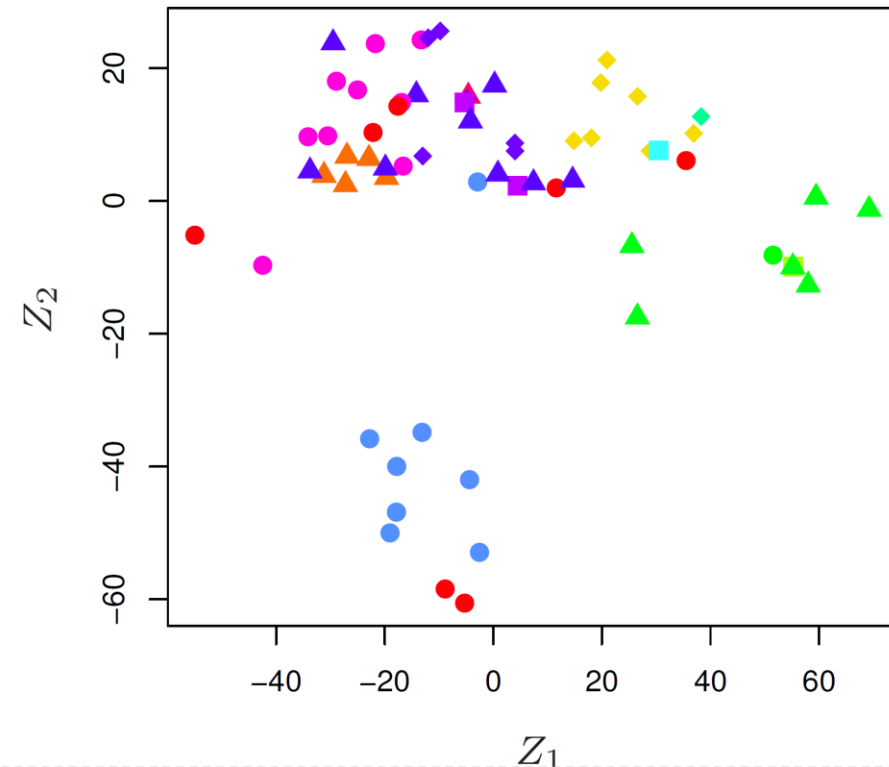
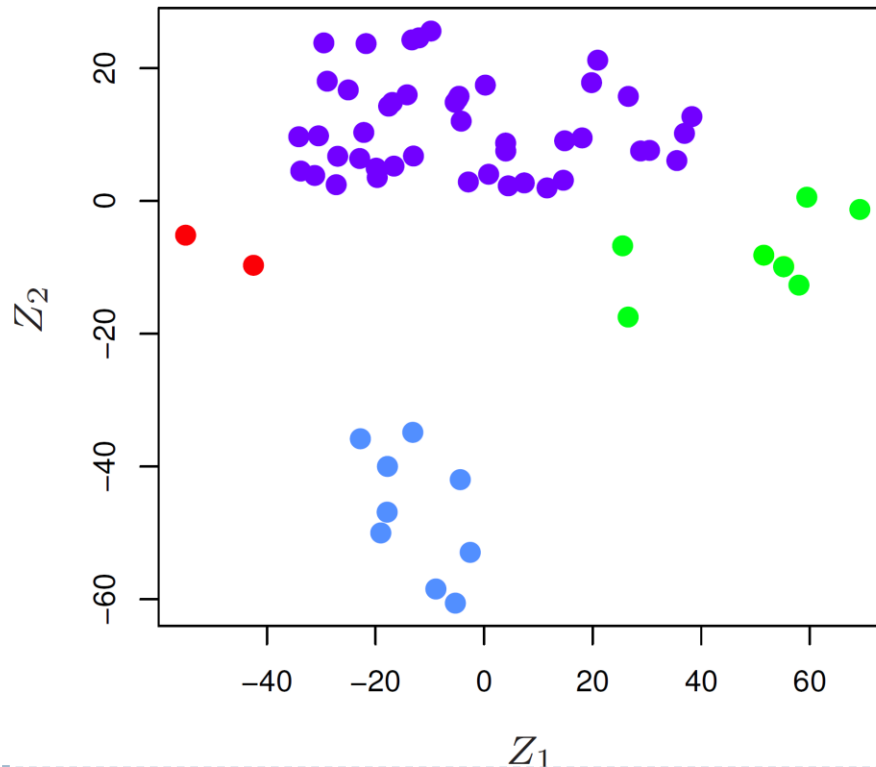
# Stock Market data

- ▶ Daily percentage returns for the S&P 500 stock index between 2001 and 2005.
  - ▶ Boxplot and heatmap



# Gene Expression Data

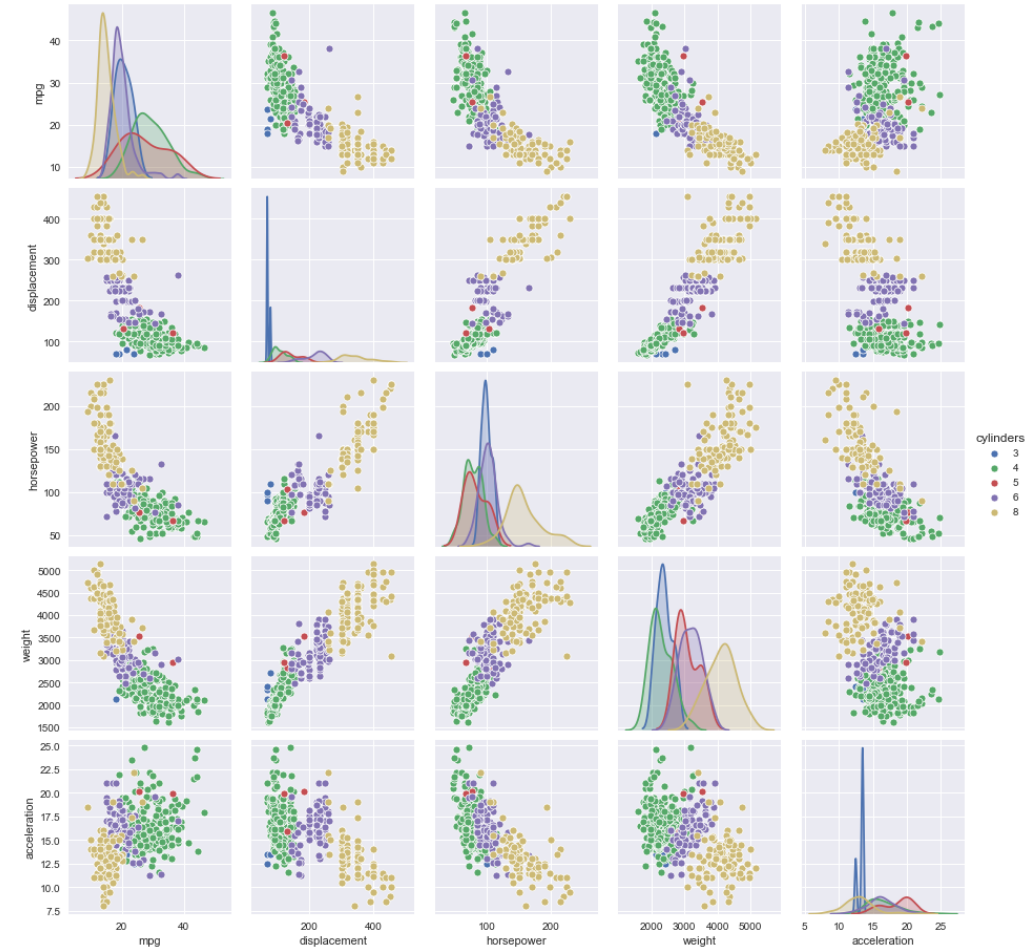
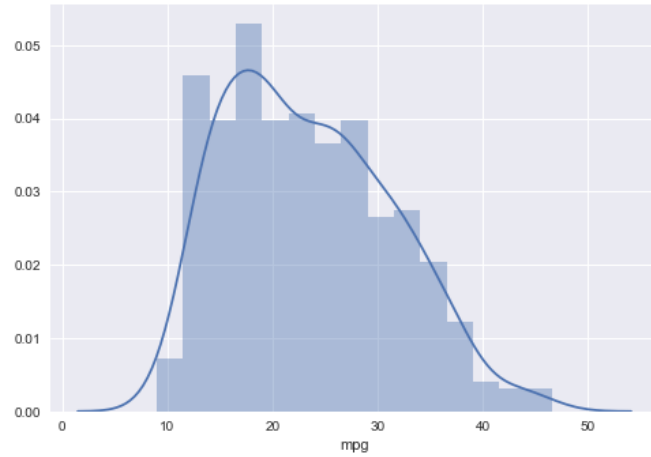
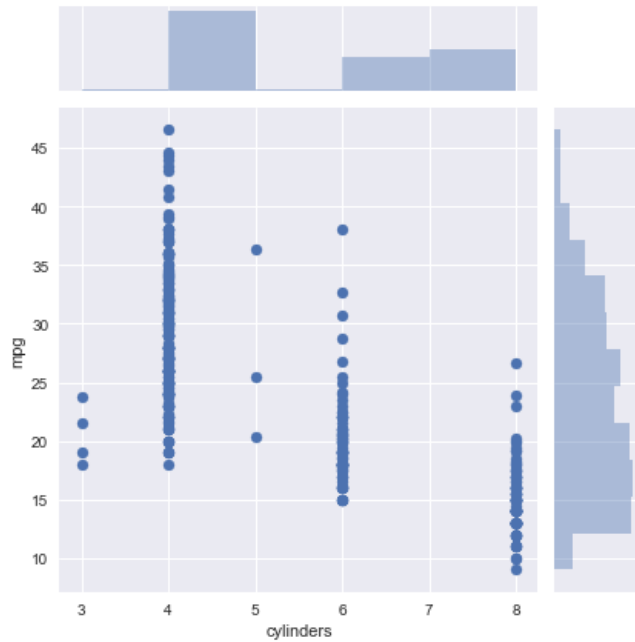
- ▶ NCI microarray data. The data contains expression levels on 6,830 genes from 64 cancer cell lines. Cancer type is also recorded.
  - ▶ Scatterplot





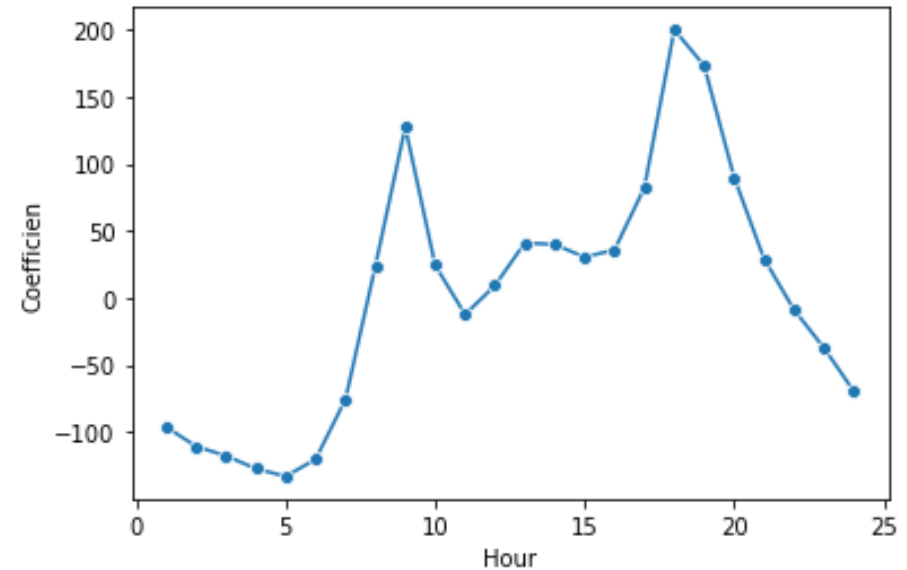
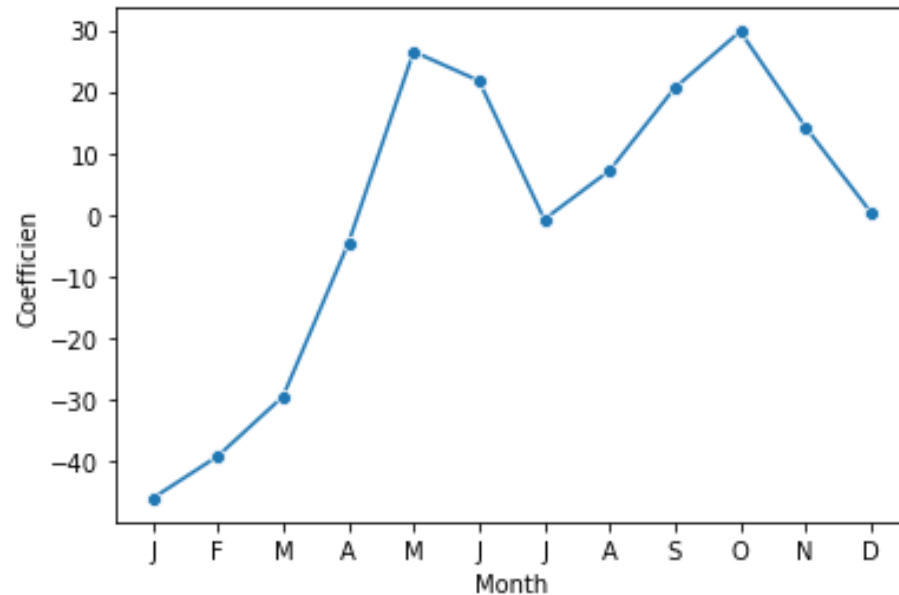
# Auto data

- ▶ Gas mileage, horsepower, and other information for 392 vehicles.
- ▶ Pairplot, displot and joinplot



# Bikeshare Data

- ▶ This data set contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system, along with weather and seasonal information.
- ▶ line plot



# Conclusion

---

- ▶ Exploratory data analysis (EDA), pioneered by John Tukey, set a foundation for the field of data science. The key idea of EDA is that the first and most important step in any project based on data is to *look at the data*. By summarizing and visualizing the data, you can gain valuable intuition and understanding of the project.
  - ▶ Exploratory analysis should be a cornerstone of any data science project
  - ▶ Other tools that use unsupervised learning will be discussed in chapter 12