Applied Statistical Methods

Introduction

Xuemao Zhang
East Stroudsburg University

January 18, 2023

What's covered in this lecture?

- Math 402 Course Outline
  - Course Objectives
  - Tentative Contents
  - Assessments
  - References
- Why programming?
- Python and an IDE Visual Studio Code
- Introduction to Data Science

Math 402 Course Outline

Course Admin Information

- Instructor: Dr. Xuemao Zhang
  - Office: SciTech Rm 128
  - Email: xzhang2@esu.edu
- Lecture Hours:
  - MWF: 10:00–10:50Am
- Department Secretary: Christine Getz
  - Office: SciTech Rm 118
  - Email: cgetz@esu.edu
  - Telephone: 570-422-3447

Course Objectives

- This course (as part of data science) will focus on data visualization, methods in statistical inferences, and statistical(machine) learning
- **Programming:** Python
- **You will learn:**
  - Choose the best chart that fits the data
  - Communicate effectively using statistical graphics
  - Use appropriate statistical methods to conduct statistical inferences
  - Fit appropriate statistical models in statistical learning and assess test errors
- **Prerequisites:** Elementary Statistics (Math 110 or similar courses) and Probability (Math 311 or similar courses).

Tentative Contentes

- Introduction to data science
- Data Wrangling
- Exploratory data analysis and data visualization
- Statistical inferences
  - Review of probability
  - Inference about population means
  - Analysis of variance for design of experiments
  - Analysis of covariance
  - Categorical data analysis
  - Linear regression models
  - Nonparametric statistical methods
- Statistical learning
  - Linear regression models
  - Linear model selection and regularization
  - Regression splines and local regression
  - Classification
  - Principal components analysis and clustering

Assessments

- 24% in-class quizzes (4 sets)
- 30% Homework assignments (6 sets)
- 26% Projects (2 sets)
- 20% Final project, consisting of
  - Oral presentation: 5%
  - Coding and Written report: 15%

References

- Chan, Stanley H. (2021). *Introduction to Probability for Data Science*. https://probability4datascience.com/
- Danielle Navarro and Ethan Weed (2021). *Learning Statistics with Python*. https://ethanweed.github.io/pythonbook/landingpage.html
- Wes McKinney (2016). *Python for Data Analysis* (3rd). https://wesmckinney.com/book/ https://wesmckinney.com/
- Stefanie Molin (2021). *Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization* (2nd). https://github.com/stefmolin/Hands-On-Data-Analysis-with-Pandas-2nd-edition

References

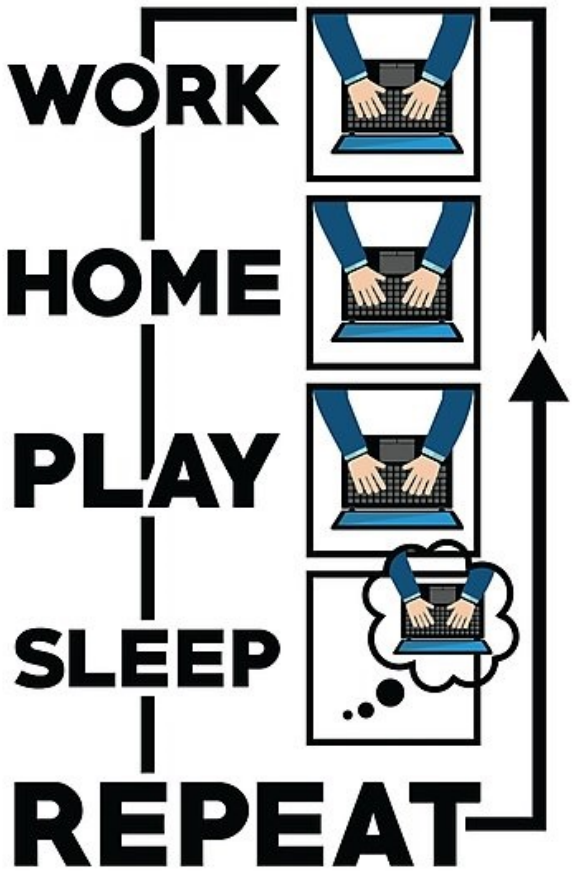
- James, G., Witten, D., Hastie, T. and Tibshirani, R.(2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. https://www.statlearning.com/
- Jordi Warmenhoven (2017). Python code for `An Introduction to Statistical Learning`. https://github.com/JWarmenhoven/ISLR-python
- Shilpa Arora (2020). Statistical Machine Learning in Python - code for `An Introduction to Statistical Learning`. https://github.com/shilpa9a/Introduction_to_statistical_learning_summary_python
- Qiuping Xu (2021). An Introduction to Statistical Learning with Applications in PYTHON. https://github.com/qx0731/Sharing_ISL_python
- Sebastian Raschka and Vahid Mirjalili (2022). *Python Machine Learning*. https://sebastianraschka.com/books/
- Python Cheat Sheets. https://www.utc.fr/~jlaforet/Suppl/python-cheatsheets.pdf

Python Resources

- Main website http://www.python.org and SciPy site http://scipy.org.
- Official Python Tutorial http://docs.python.org/2/tutorial/index.html.
- Google's Python Class (2 day class materials including video and exercises) https://developers.google.com/edu/python.
- Think Stats - Exploratory Data Analysis in Python (http://greenteapress.com/thinkstats2/thinkstats2.pdf)
- Python scientific lecture notes (http://www.scipy-lectures.org)
- Learn Python the Hard Way (https://learnpythonthehardway.org/)
- w3schools (https://www.w3schools.com/)

Why programming?

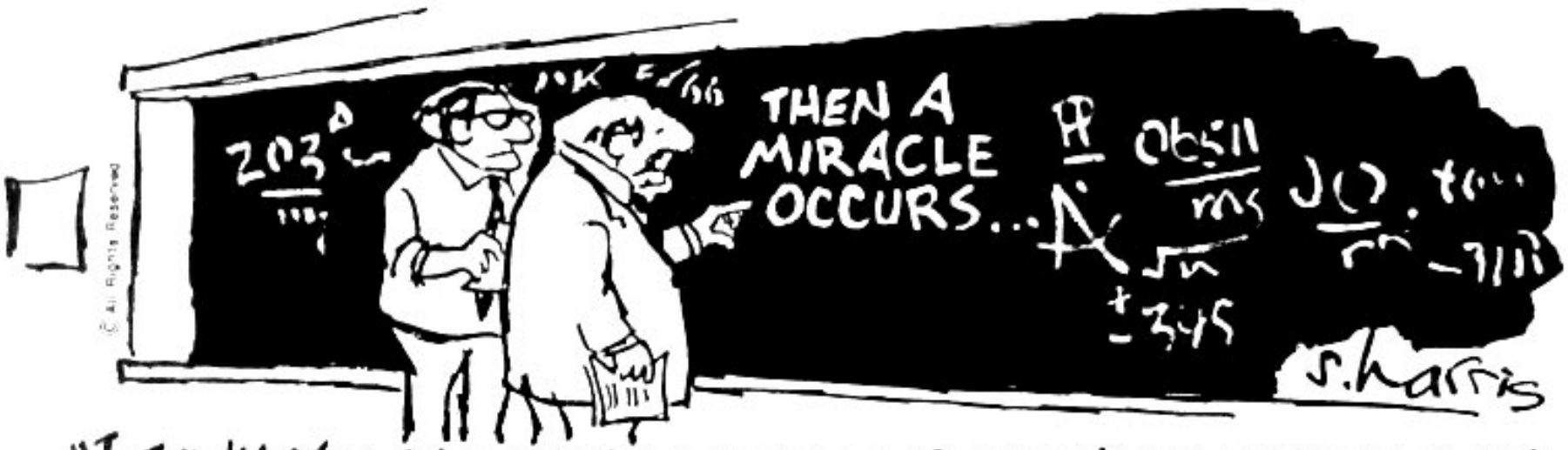
- To be able to easily repeat your own work.



Source: https://www.redbubble.com

Why programming?

- The workflow of using a script makes your research reproducible.



Source: Malanris.ru

Why programming?

- Python, R and SAS are the main data analytics tools which require programming.
- Programming isn't scary. If you've written formulas in Excel, you've already done "programming".

Python

- Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant **indentation**.
- Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming
- Download: https://www.python.org/downloads/

Python

- Python is for data science.
- The library `Pandas` (open source since 2009) uses data structure "data frame" used in R (https://cran.r-project.org/).

```
##   emp_id emp_name salary start_date
## 1      1     Rick 623.30 2012-01-01
## 2      2      Dan 515.20 2013-09-23
## 3      3 Michelle 611.00 2014-11-15
## 4      4     Ryan 729.00 2014-05-11
## 5      5     Gary 843.25 2015-03-27
```

Python libraries and packages

What Is a Python Package?
https://www.udacity.com/blog/2021/01/what-is-a-python-package.html

- A python module is a Python program that you import, either in interactive mode or into your other programs. 'Module' is really an umbrella term for reusable code.
- A python **package** or **library** is a collection of modules. Modules that are related to each other are mainly put in the same package/library.
- Python Package Index (PyPI) is the repository of software for Python at http://pypi.python.org/pypi. As of a day in October 2022, there are over 137,000 python libraries and 198,826 python packages ready to ease developers' regular programming experience. Once a package/library is successfully installed, then you can import the package/library within your script.

Python libraries and packages

- The five most-important Python libraries are `NumPy`, `Pandas`, `Matplotlib`, `statsmodels` and `Scikit-learn`.
- **NumPy** — A library that makes a variety of mathematical and statistical operations easier; it is also the basis for many features of the pandas library.
- **pandas** — A Python library created specifically to facilitate working with data. This is the bread and butter of a lot of Python data science work.
- **Matplotlib** — A visualization library that makes it quick and easy to generate charts from your data.
- **statsmodels** — statsmodels is a Python package that provides a complement to `scipy` for statistical computations including descriptive statistics and estimation and inference for statistical models.
- **Scikit-learn** — The most popular library for machine learning work in Python.

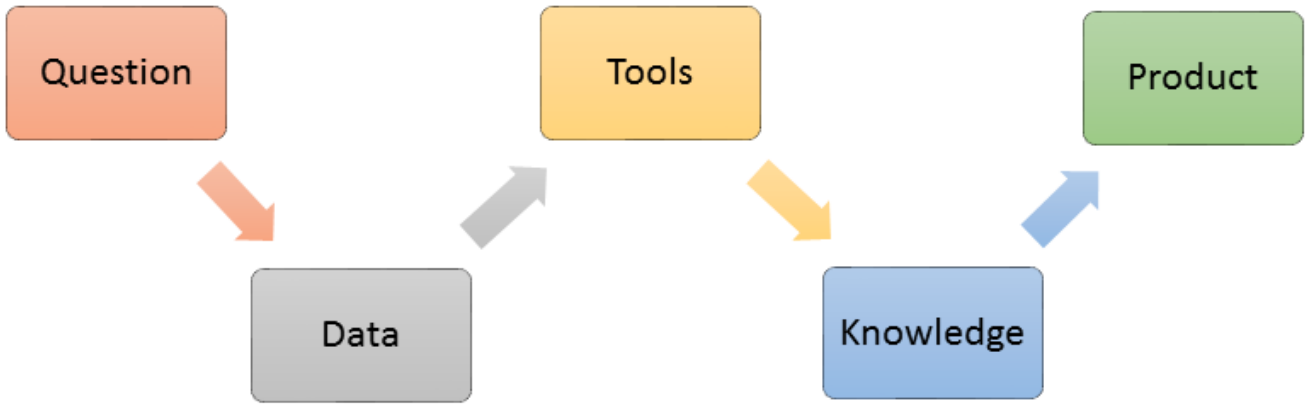
IDE and code editors

- Visual Studio Code (https://code.visualstudio.com/): created by Microsoft, a free and open-source source-code editor that can be used for Python development.
- IDLE (Integrated Development and Learning Environment)is a default editor that accompanies Python.
- Jupyter (https://jupyter.org/): the web-based **interactive** development environment for notebooks, code, and data.
- ipython (https://ipython.org/): a command shell for **interactive** computing.
- Spyder: an open-source IDE usually used for scientific development.
- PyCharm: an IDE for professional developers. It is created by JetBrains, a company known for creating great software development tools.
- Atom: an open-source code editor developed by Github.

⋮

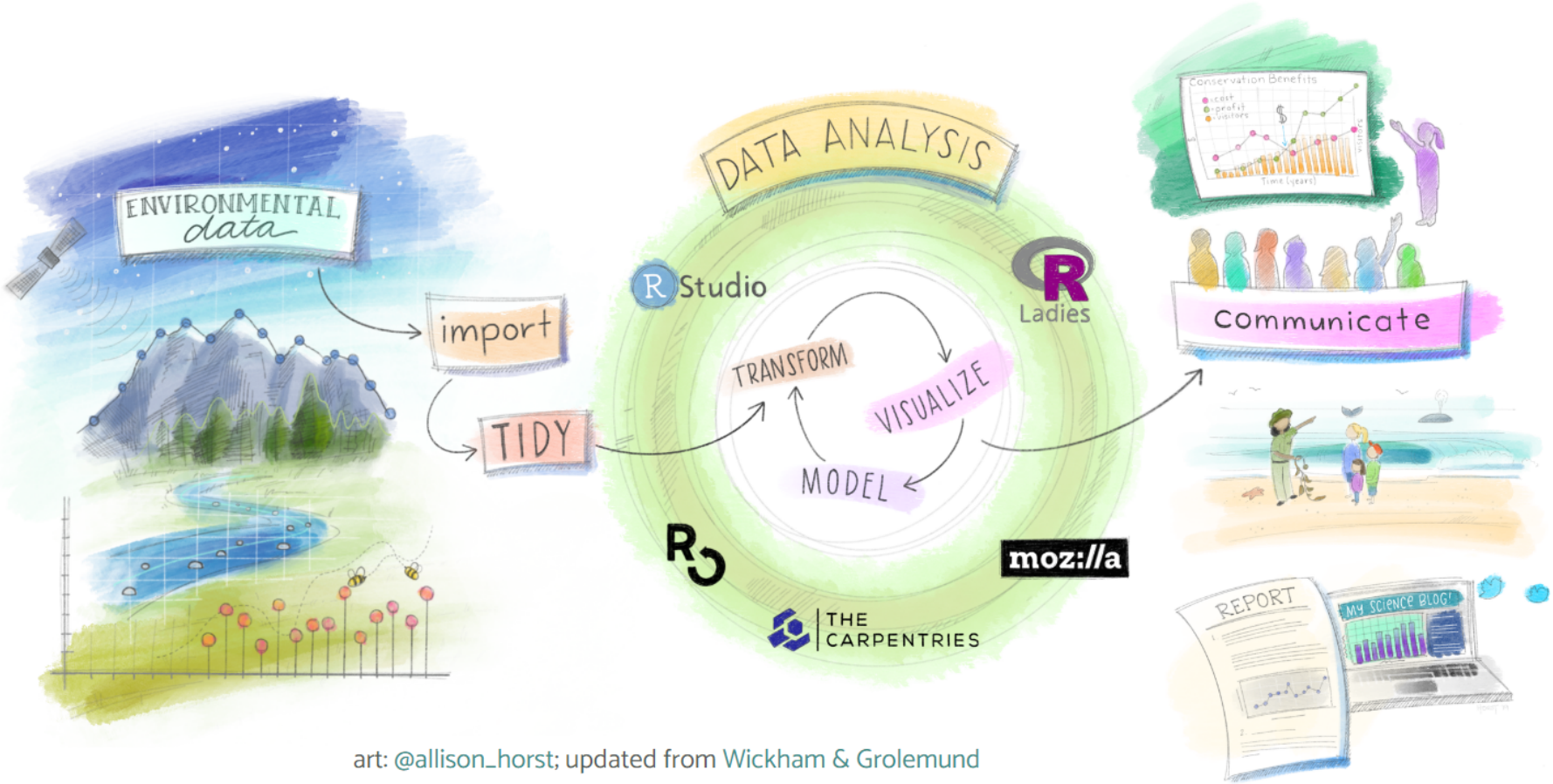
Introduction to data science

- Data science is the study of large sets of data, using computers to look for patterns and trends.



Introduction to data science

- Data science is the discipline of turning raw data into understanding



art: @allison_horst; updated from Wickham & Grolemund

Data Scientist The Sexy Job

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

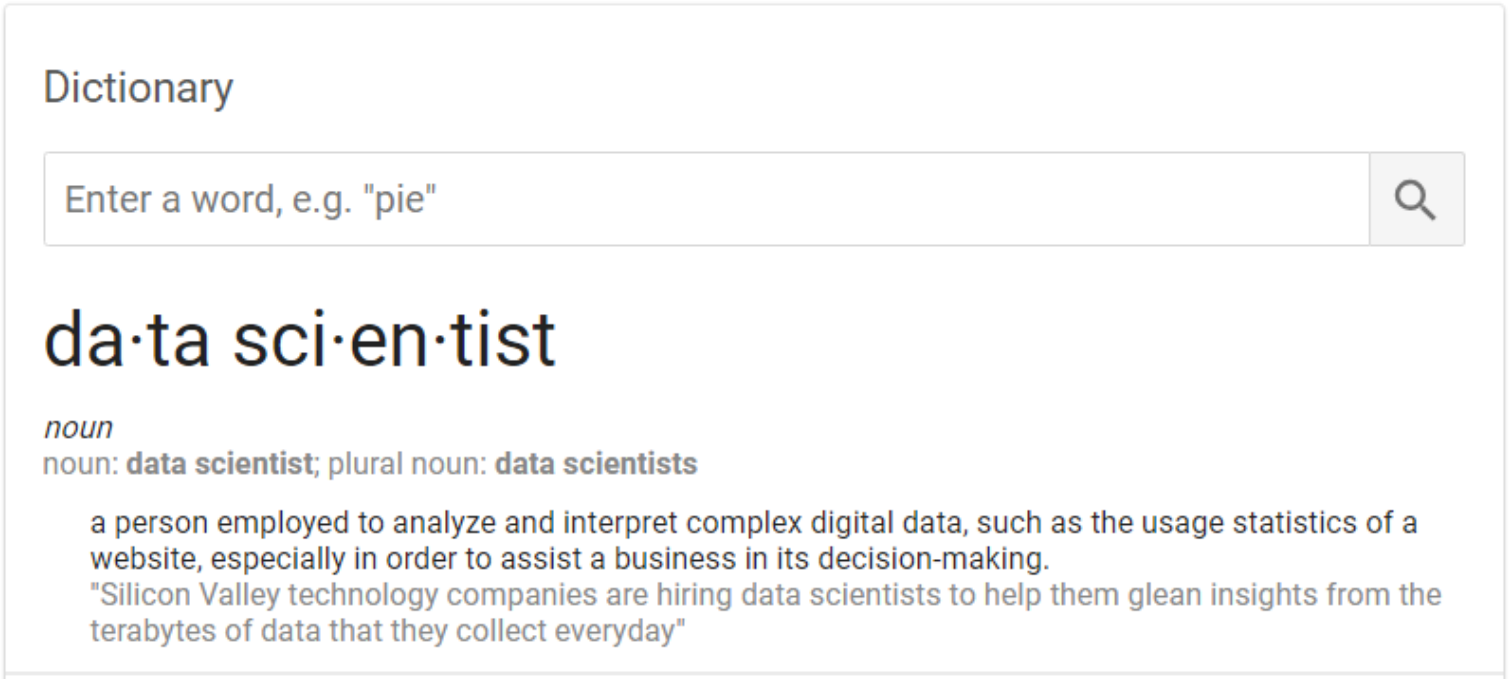
- See also an old article by NYT (2009): For Today's Graduate, Just One Word: Statistics
- And another famous McKinseay 2011 Report: Big data: The next frontier for innovation, competition, and productivity
- Is Data Scientist Still the Sexiest Job of the 21st Century?

Job Market in Data Science or Data Analytics

Search Data Analytics or Data Scientist on https://www.indeed.com/

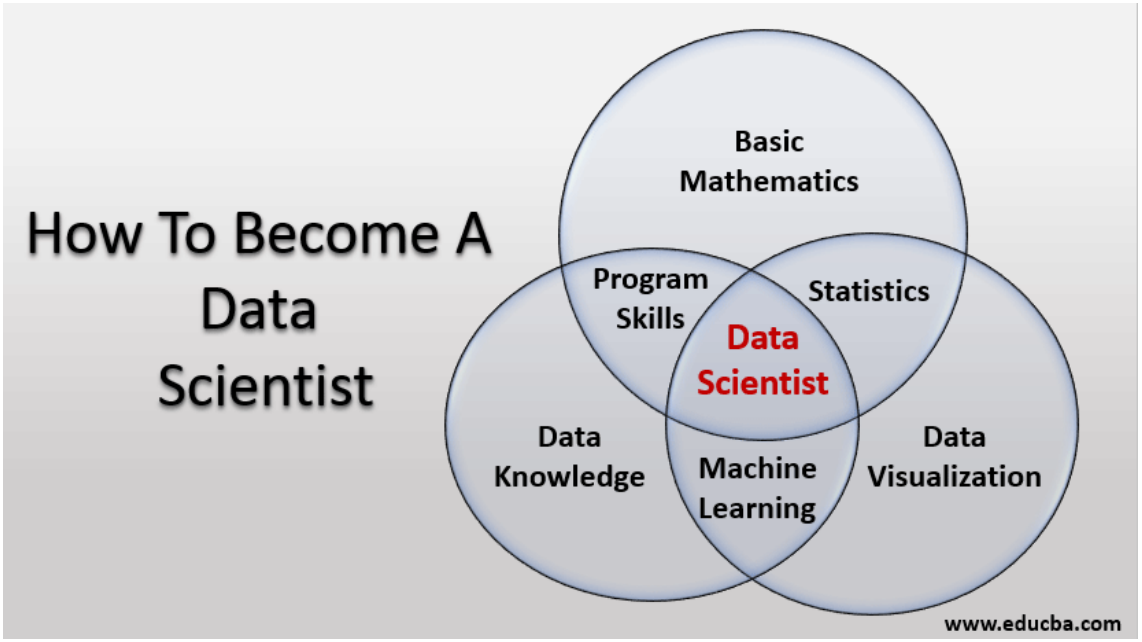
What is a data scientist?

- "A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician." (from Joshua Blumenstock, 2013).



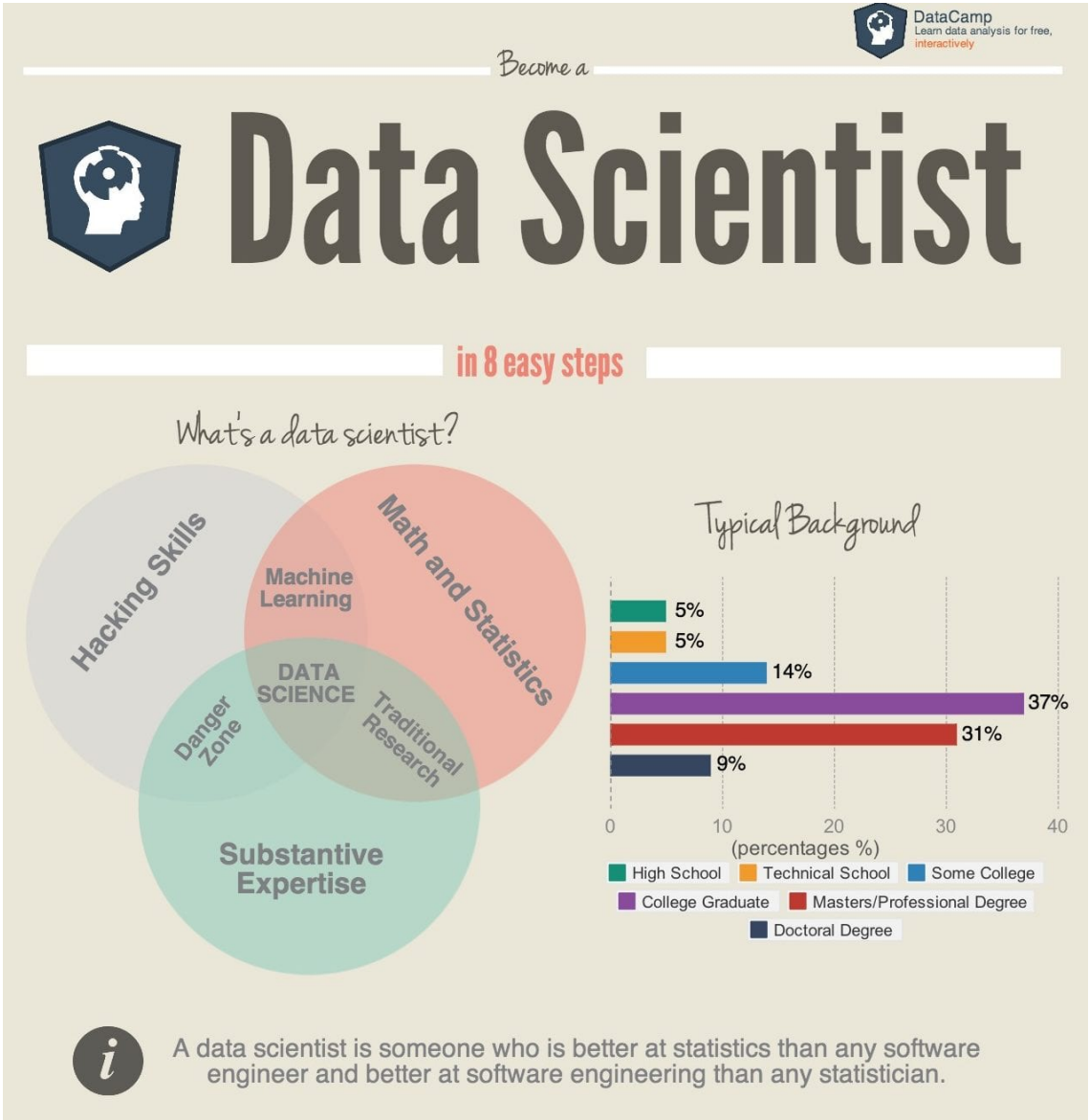
How to become a data scientist

- An article on coursera: How to Become a Data Scientist?
- How to Become a Data Scientist



How to become a data scientist

- Become A Data Scientist in 8 Steps: Infographic



Programming languages

The 10 Best Data Science Programming Languages to Learn in 2021

5 Best Data Analysis Programming Languages in 2022

- Programming languages for data analysis:
  - R
  - Python
  - SAS
  - Julie
- General purpose programming languages
  - Python
  - Java
  - C/C++
  - Scala
  - Julie
- 2020 SAS, R, or Python Survey Results: Which Tool do Data Scientists & Analytics Pros Prefer?
- Databases query: SQL(structured query language)
- Web app development: JavaScript

Data visualization tools

8 Best Data Visualization Tools that Every Data Scientist Should Know

- Microsoft Power BI
- Tableau
- Plotly (free library in R and Python) and Dash (commercial product by https://plotly.com/)
- SAS Visual Analytics
- ⋮
- Excel with VBA

Tools for big data

Top 7 Big Data Analytics Tools | Its Technology And Techniques

- Apache Airflow: "Airflow is a platform to programmatically author, schedule and monitor workflows." — Airflow documentation
- Apache Hadoop
- Apache Spark
- MongoDB
- RapidMiner
- Microsoft Azure
- Zoho Analytics

License