

Lecture 1

Introduction

Introduction to Python Programming for Data Science MT19

27th January 2022

Welcome to “Python Programming for Data Science – Part 1 (PP4DS-I)”

- Face-to-face online 10-week course.
- Developed for the Department of Continuing Education, starting in Hilary Term 2022.
- Each week:
 - Live online sessions: Thursdays 7-9 pm at Ewert House, starting from Thursday 27 January

Course Overview

- Course aim: offer an introduction to programming for data science, using Python
- At the end of the course, you will be able to write Python programs to:
 - import data from local and/or remote sources and pre-process it
 - extract significant information from the gathered data
 - visualise the relevant features extracted from the data
- Summative Assessments:
 - Students will be asked to submit a portfolio of 3 exercises for their coursework assignment. I will give the first exercise on week 5 for early submission, the second and the third on week 8 and 9.

Components of PP4DS1

- Lectures
- Practical programming exercises, using Python
- Class discussions, off and online
- Formative Assessments: quizzes and small exercises
- Summative Assessment: portfolio of exercises
- Week-to-week class feedback forms

First week

- Introduction to the course
- What is Data Science?
- A brief history of Data Science
- Data science processes
- Programming for Data Science
- Getting set up with the tools for the course

Beyond the first week

Week 2: Python basics: built-in types, functions and methods, if statements

Week 3: Python data structures: list, dictionaries, tuples; for loops (Oct 17)

Week 4: NumPy and the SciPy ecosystem

Week 5: Pandas for data science I

Week 6: Pandas for data science II

Week 7: Data visualisation: matplotlib and seaborn

Week 8: Object-oriented programming: classes, inheritance, and applications

Week 9: Data gathering and cleaning. Text pre-processing

Week 10: Introduction to experimental design and statistical tests. Time Series Analysis

About the tutor

- Originally from Genoa, Italy
- Biomedical Engineering degrees from the University of Genoa
- Worked for 5 years as RSE at the University of Oxford
- Software Engineer at Mind Foundry
- Contacts
 - email: massimiliano.izzo@conted.ox.ac.uk
 - Slack: using the pp4ds-ox.slack.com workspace



DEPARTMENT OF
ENGINEERING
SCIENCE

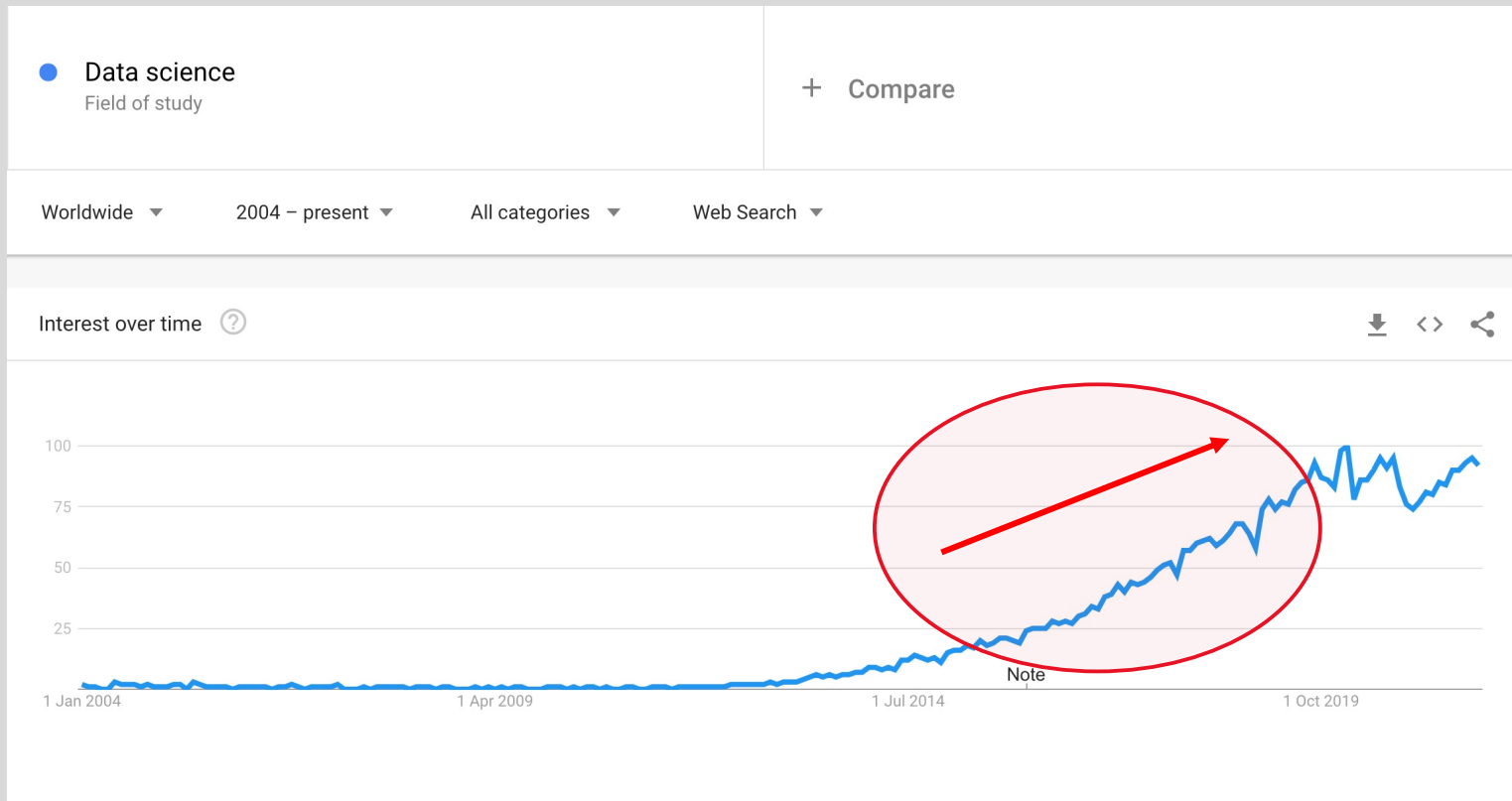


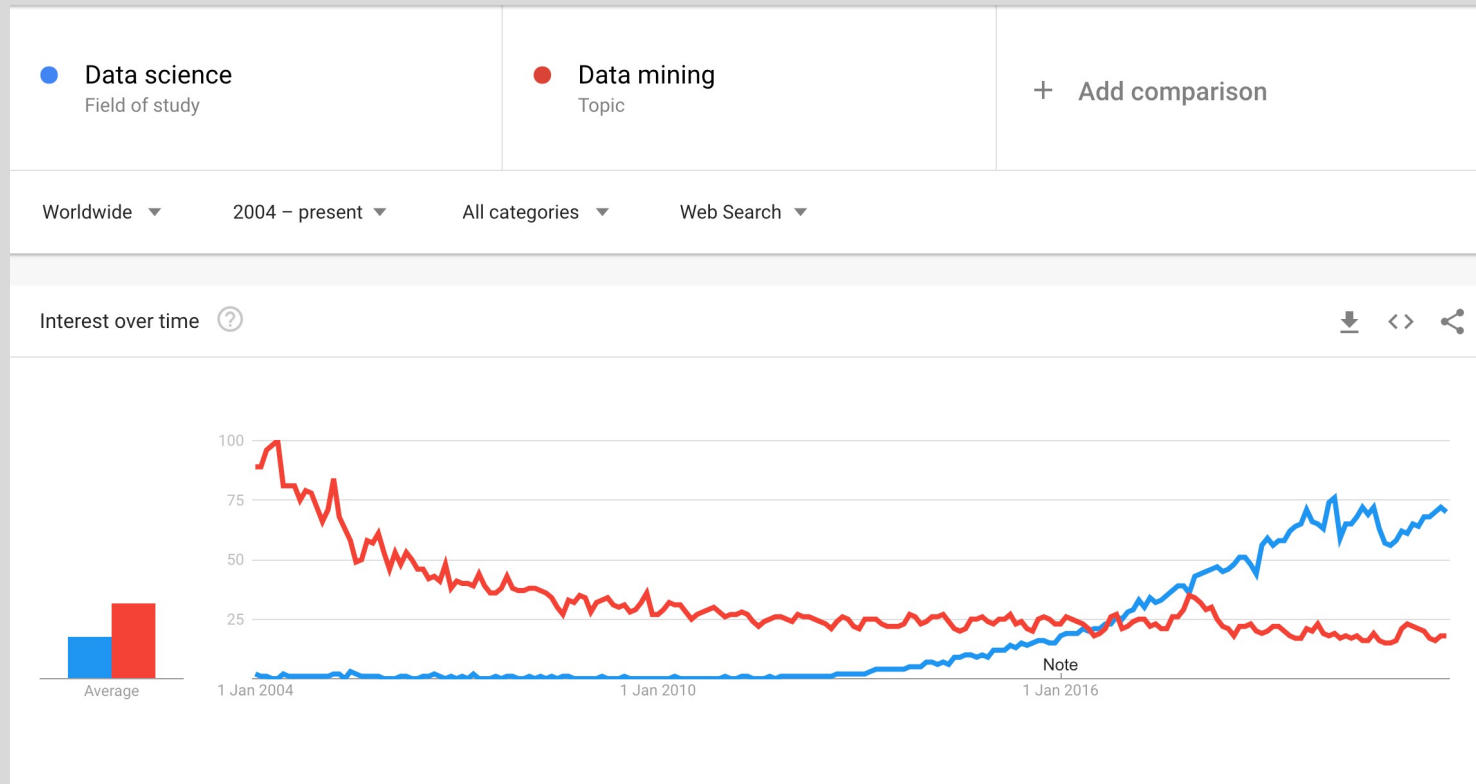
You?

Course Project Page

<https://gitlab.com/data-science-course/pp4ds-pt1-ht2022>

What is Data Science?





Source: Google Trends

A brief history of Data Science (I)

(<https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#29b846a555cf>)

- **1962** John W. Tukey writes in “The Future of Data Analysis”: “Data analysis, and the parts of statistics which adhere to it, must...take on the characteristics of science rather than those of mathematics... data analysis is intrinsically an empirical science... How vital and how important... is the rise of the stored-program electronic computer?”
- https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711



A brief history of Data Science (II)



- **1974** Peter Naur defines data science: “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”
- **1977** [The International Association for Statistical Computing](#) (IASC) is established as a Section of the [ISI](#).

A brief history of Data Science (II)

- **1989** Gregory Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in Databases (KDD) workshop.
- **1996** Members of the [*International Federation of Classification Societies \(IFCS\)*](#) meet in Kobe, Japan, for their biennial conference. For the first time, the term “data science” is included in the title of the conference (“Data science, classification, and related methods”).
- **1996** Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth publish “[From Data Mining to Knowledge Discovery in Databases](#).”
- **1997** In his inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan, Professor C. F. Jeff Wu (currently at the Georgia Institute of Technology), calls for statistics to be renamed data science and statisticians to be renamed data scientists.

A brief history of Data Science (III)

- **2001** William S. Cleveland publishes "[Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics](#)." It is a plan “to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called ‘data science.’” .



Wiley Online Library

Search  [Login / Register](#)

STATISTICAL ANALYSIS AND DATA MINING

Original Article

Data science: An action plan for expanding the technical areas of the field of statistics

William S. Cleveland 

First published: 27 November 2014 | <https://doi.org/10.1002/sam.11239> | Cited by: 7

[Read the full text >](#)  PDF  TOOLS  SHARE

Abstract

An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

 [Volume 7, Issue 6](#)
December 2014
Pages 414-417

Advertisement



 Related  Information

A brief history of Data Science (III)

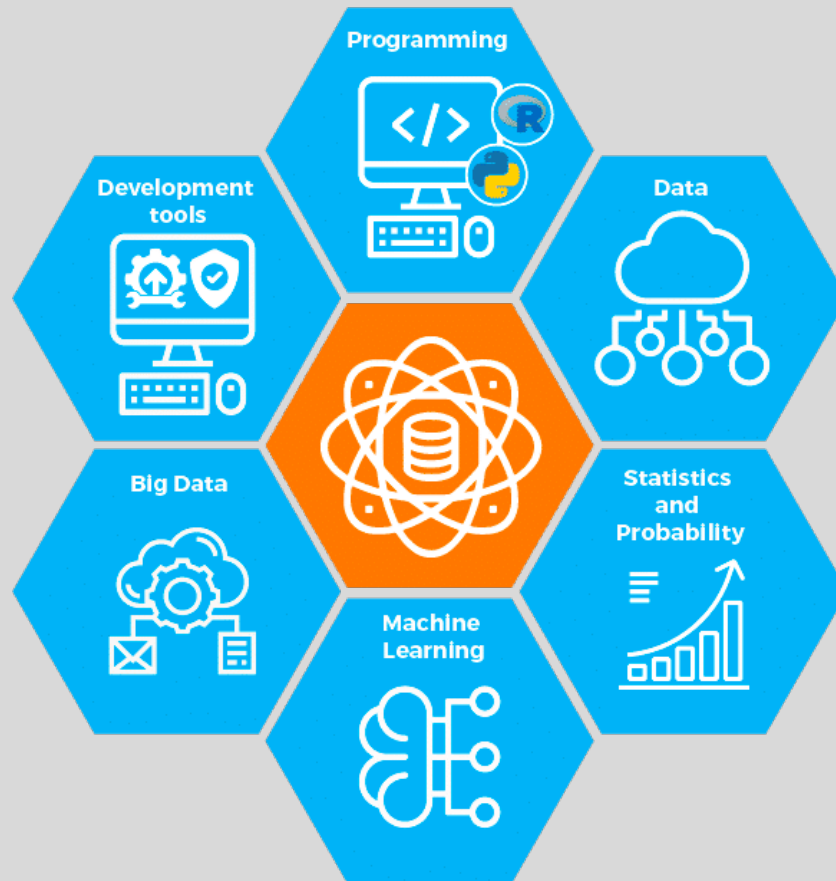
- **April 2002** Launch of [*Data Science Journal*](#)
- **January 2003** Launch of [*Journal of Data Science*](#): “By ‘Data Science’ we mean almost everything that has something to do with data: Collecting, analyzing, modeling..... yet the most important part is its applications--all sorts of applications.”
- **January 2009** [*Harnessing the Power of Digital Data for Science and Society*](#) is published.
- **June 2009** Nathan Yau writes in “[*Rise of the Data Scientist*](#)”

Data Science Life Cycle



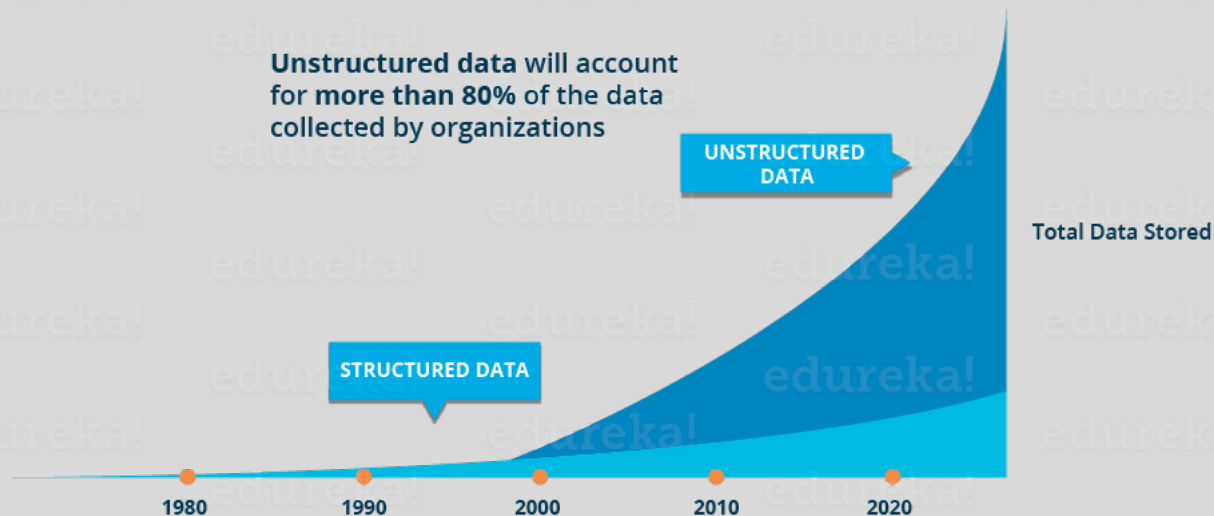
Data Science Components

(<https://intellipaas.com>)



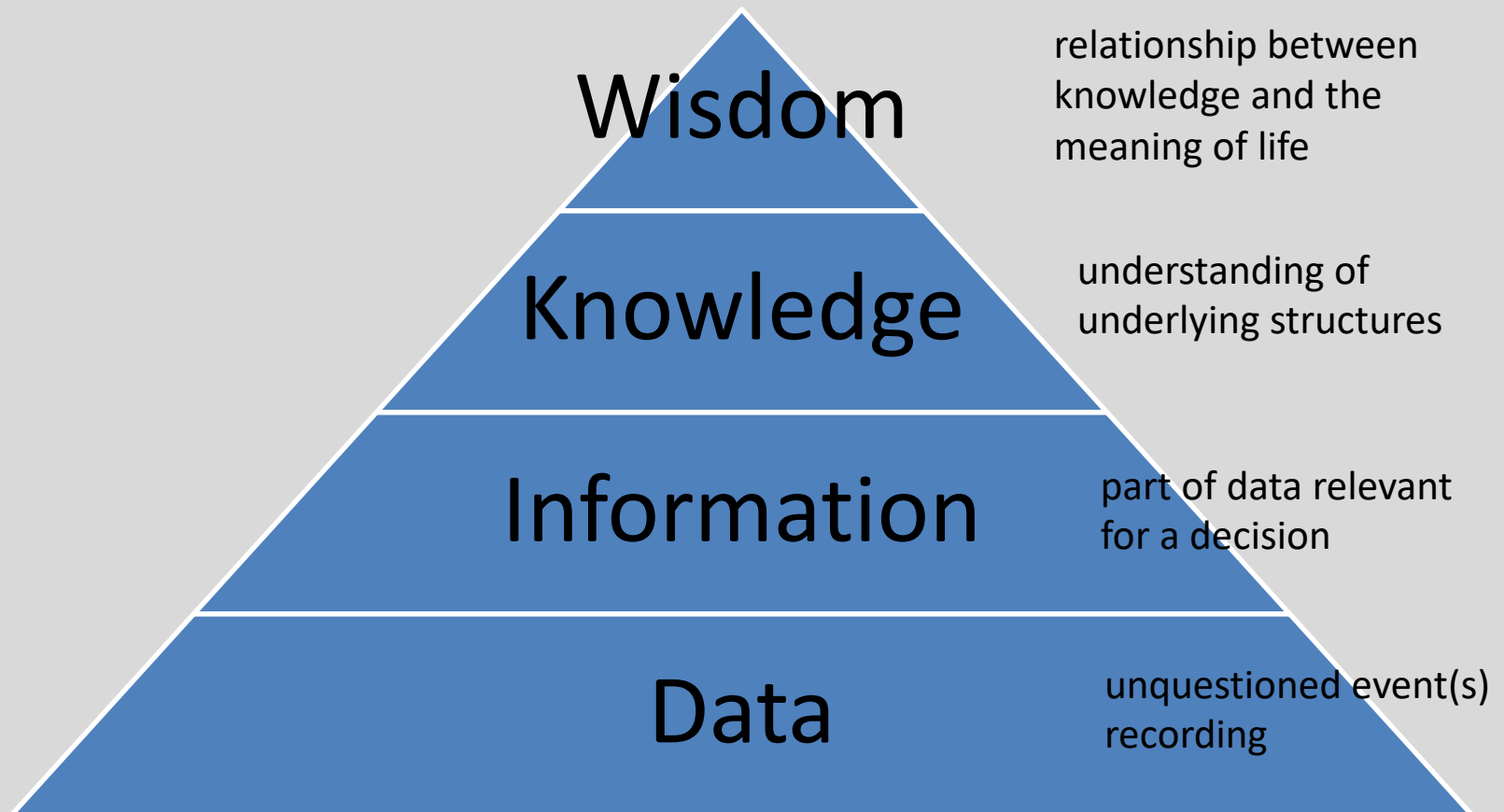
Data

- Structured (tabular)
- Semi-structured (JSON, XML)
- Unstructured (images, text, ...)



<https://www.edureka.co/blog/what-is-data-science/>

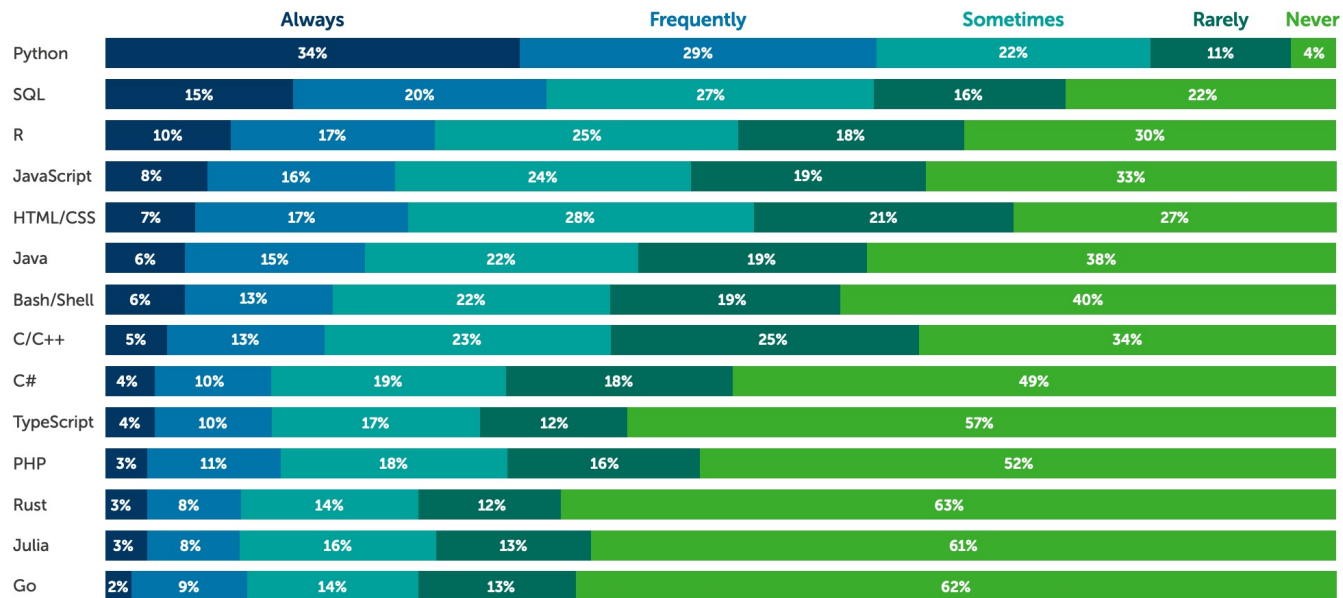
DIKW (epistemological) pyramid



<https://www.adizes.com/articles/cval-where-wisdom.pdf>

State of Data Science 2021 (I)

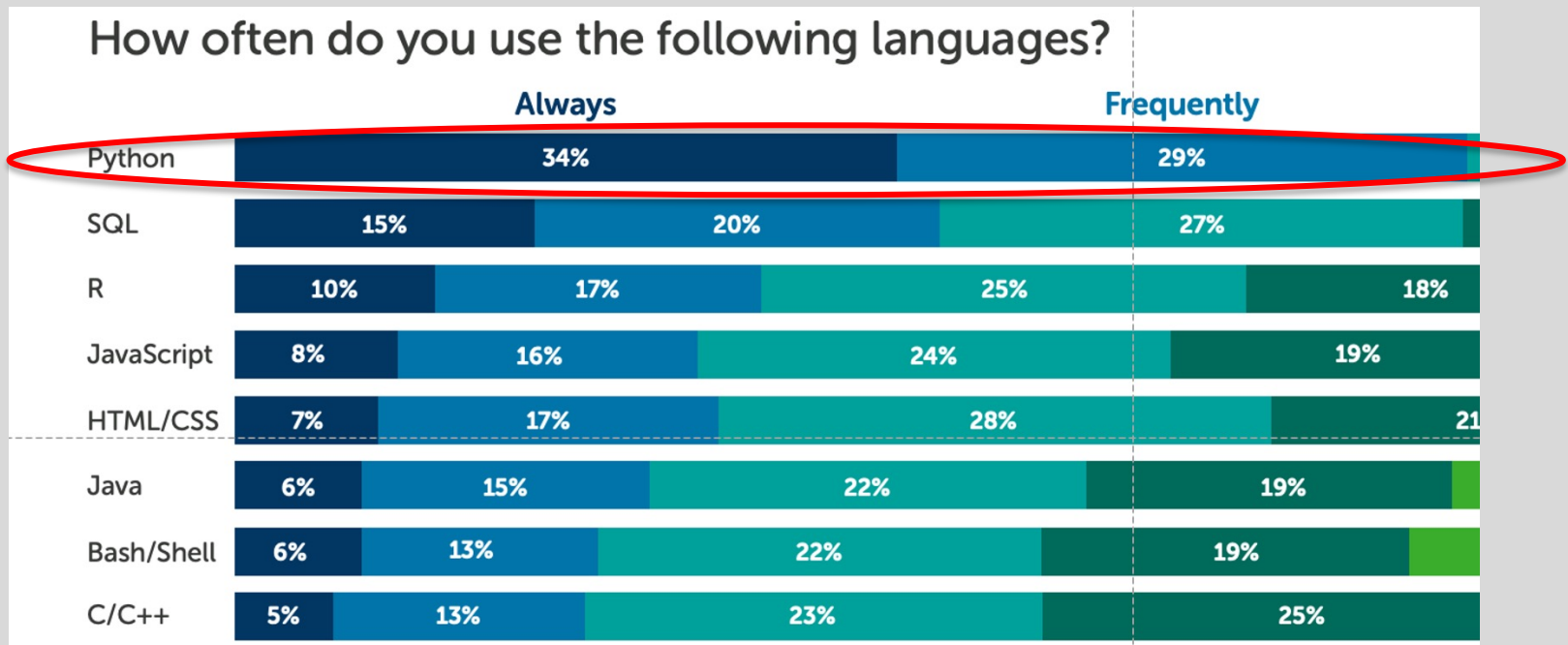
How often do you use the following languages?



n = 3,104

<https://www.anaconda.com/state-of-data-science-2021>

State of Data Science 2021 (II)



Why Python? (I)

- Popular in data science
- ...and widely used for general software development
- Quick prototyping

Beautiful is better than ugly. Explicit is better than implicit. Simple is better than complex.

There should be one – and preferably only one – obvious way to do it.

From “The Zen of Python”

Why Python? (II)

- Easy to write code but can handle complex mathematical processing
- Relatively high performance
- Extensive built-in library (i.e. set of modules ready to be used)
- Free and open source ecosystem (i.e. set of modules that can be freely installed and used)

Python libraries for Data Science

Numeric/scientific computation	NumPy, SciPy library
Data Analysis and manipulation	Pandas
Visualization	Matplotlib, Seaborn, Bokeh, Plotly
Machine Learning	Scikit-learn
Deep Learning	Tensorflow, PyTorch, JAX, Keras
Natural Language Processing	SpaCy, NLTK, Gensim
Others	Scrapy, Statsmodels, Spyder, Jupyter

<https://activewizards.com/blog/top-15-libraries-for-data-science-in-python/>

Let's get set up

We won't do any programming today, but we will get you set up with Slack and Jupyter so we can dive into Python next week.

- **Anaconda + Jupyter – our programming environment**
- **Git – our source code management system**
- **Slack – our class discussion forum**

Anaconda Platform

- Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.)
- It aims to simplify package management and deployment.
- Anaconda Individual Edition:
<https://www.anaconda.com/products/individual>

Anaconda (conda) environments

- An environment is an isolated space where you can safely specify the dependencies for a project
- Anaconda comes installed with a default (base) environment
- We will create a dedicated environment for our data science course in Anaconda and we will gradually add dependencies (e.g NumPy, Pandas...) as we need them
- We will then run Jupyter Notebooks from within our environment

Jupyter Notebook (I)

- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more
- Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.

Jupyter Notebooks (II)

- The Jupyter Notebook architecture consists of two parts:
 - frontend (web interface)
 - back-end (kernel)
- The Python backend is IPython kernel (<https://ipython.org/>)
- IPython offers a interactive shell or console as well
- There are other kernels for other languages

Git

- Software that you will use to download and keep track of the changes on the course project page
- The course project page is hosted on the Gitlab.com website:

<https://gitlab.com/data-science-course/pp4ds-pt1-ht2022>

How to install Git

- Windows:
 - install GitBash: <https://gitforwindows.org/>
- macOS:
 - “*brew install git*” (if you have Homebrew) from a terminal
 - download the installer from <https://sourceforge.net/projects/git-osx-installer/>
- Debian/Ubuntu Linux:
 - “*apt-get install git*” from a terminal

Working with Jupyter: Online Option

- Go to our course web page on Gitlab.com (<https://gitlab.com/data-science-course/pp4ds-pt1-ht2022>)
- Start Jupyter notebook clicking on the Binder badge next to the exercise name

Working with Jupyter: Offline Option

- Open GitBash (or a Windows Terminal)
- From your Home Directory run this command (all in one line):

```
git clone https://gitlab.com/data-science-course/pp4ds-pt1-ht2022
```

- Start on your machine Anaconda Navigator
- Start Jupyter notebook from the interface
- Open the notebook (guidance will be provided)

Weekly evaluation form

Please take 5 minutes at the end of the class to fill out the class evaluation form.

Thanks!

<https://forms.gle/9Xx7L7oCnji6gKbDA>