Data Science 101

Janis Keuper

INSTITUTE FOR MACHINE
LEARNING AND ANALYTICS

Data Science 101 - Outline

- What is “Data Science ?”
- Key Skills for Data Scientists
- Data Science Work Flow
- Python for Data Science
- Case Study I: predicting supermarket sales

What is Data Science?

Simple Question: What is Data Science?

What is Data Science?

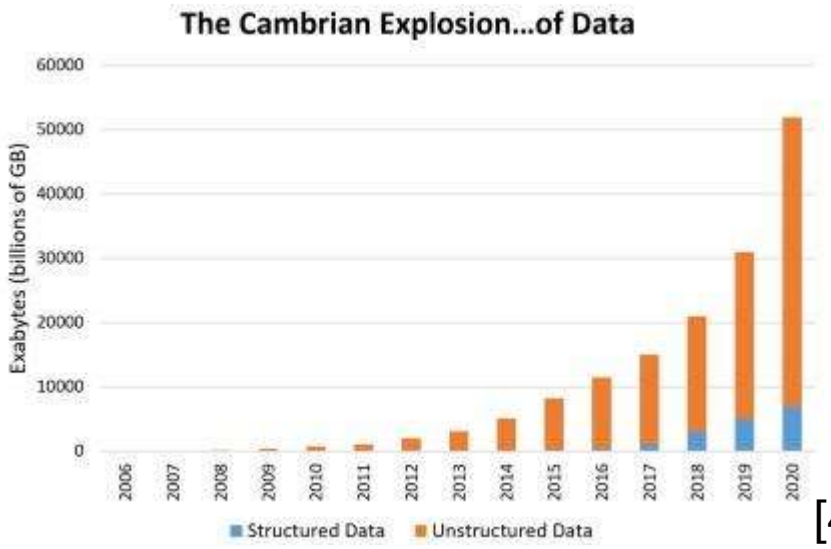
Simple Question: What is Data Science?

→ many different answers

(all right in some way)

What is Data Science?

Historic motivation



[4]

Multiplying Factor	SI Prefix	Scientific Notation	Name
1 000 000 000 000 000 000 000 000	Yotta (Y)	10^{24}	1 septillion
1 000 000 000 000 000 000 000	Zetta (Z)	10^{21}	1 sextillion
1 000 000 000 000 000 000	Exa (E)	10^{18}	1 quintillion
1 000 000 000 000 000	Peta (P)	10^{15}	1 quadrillion
1 000 000 000 000	Tera (T)	10^{12}	1 trillion
1 000 000 000	Giga (G)	10^{9}	1 billion
1 000 000	Mega (M)	10^{6}	1 million
1 000	kilo (k)	10^{3}	1 thousand
0 001	milli (m)	10^{-3}	1 thousandth
0 000 001	micro (u)	10^{-6}	1 millionth
0 000 000 001	nano (n)	10^{-9}	1 billionth
0 000 000 000 001	pico (p)	10^{-12}	1 trillionth
0 000 000 000 000 001	femto (f)	10^{-15}	1 quadrillionth
0 000 000 000 000 000 001	atto (a)	10^{-18}	1 quintillionth
0 000 000 000 000 000 000 001	zepto (z)	10^{-21}	1 sextillionth
0 000 000 000 000 000 000 000 001	yocto (y)	10^{-24}	1 septillionth

[4]

What is Data Science?

Hochschule Offenburg
offenburg.university

Historic motivation

Leads to many new questions:

- ***Where is this data coming from?***
- ***What are we going to do with it?***
- ***Who is going to do it?***

What is Data Science?

Hochschule Offenburg
offenburg.university

Historic motivation

Leads to many new questions:

- ***Where is this data coming from?***
  - ***Internet of Things (IoT: sensors every where)***
  - ***Industry 4.0 (connecting everything)***
  - ***Big Data (storing all the information you can in the hope generate gain)***
  - ***Social Media (and other internet services)***

- ***What are we going to do with it?***

- ***Who is going to do it?***

What is Data Science?

Hochschule Offenburg
offenburg.university

Historic motivation

Leads to many new questions:

- ***Where is this data coming from?***
  - ***Internet of Things (IoT: sensors every where)***
  - ***Industry 4.0 (connecting everything)***
  - ***Big Data (storing all the information you can in the hope generate gain)***
  - ***Social Media (and other internet services)***

- ***What are we going to do with it?***
  - ***Analyze the past***
  - ***Predict the future (at least try to)***
  - ***Machine Learning (data driven computing)***
  - ***AI (what ever that is)***

What is Data Science?

Hochschule Offenburg
offenburg.university

Historic motivation

Leads to many new questions:

- ***Where is this data coming from?***
  - *Internet of Things (IoT: sensors every where)*
  - *Industry 4.0 (connecting everything)*
  - *Big Data (storing all the information you can in the hope generate gain)*
  - *Social Media (and other internet services)*

- ***What are we going to do with it?***
  - *Analyze the past*
  - *Predict the future (at least try to)*
  - *Machine Learning (data driven computing)*
  - *AI (what ever that is)*

Biggest Buzz-Words in IT

What is Data Science?

Hochschule Offenburg
offenburg.university

Historic motivation

Leads to many new questions:

- ***Where is this data coming from?***
- ***What are we going to do with it?***
- ***Who is going to do it?*** → ***Data Science !***

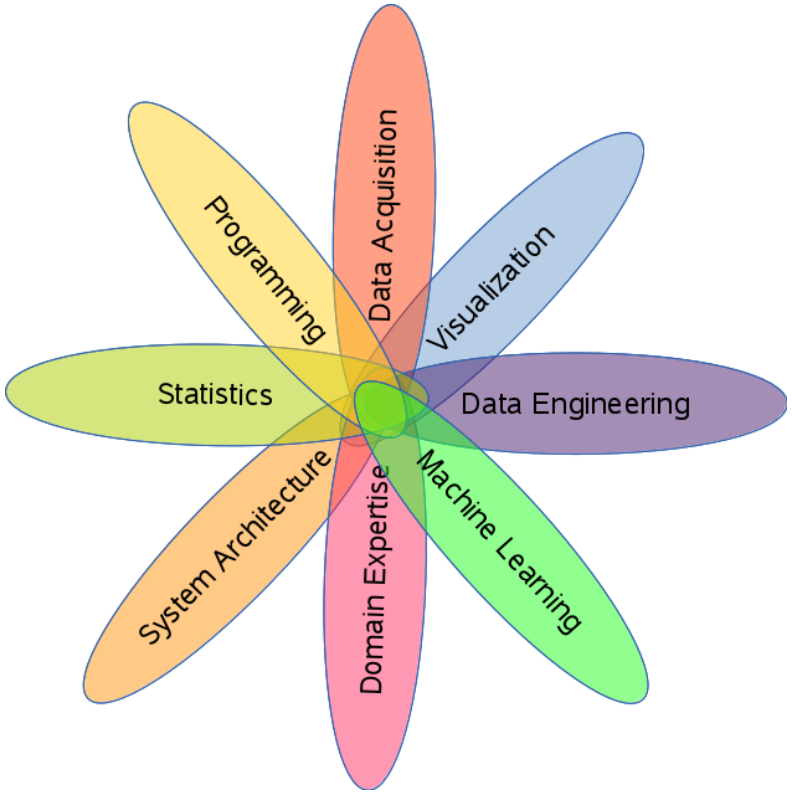
What is Data Science?

Hochschule Offenburg
offenburg.university

Motivation by tasks:

What is expected of Data Science?

What are typical tasks / applications?

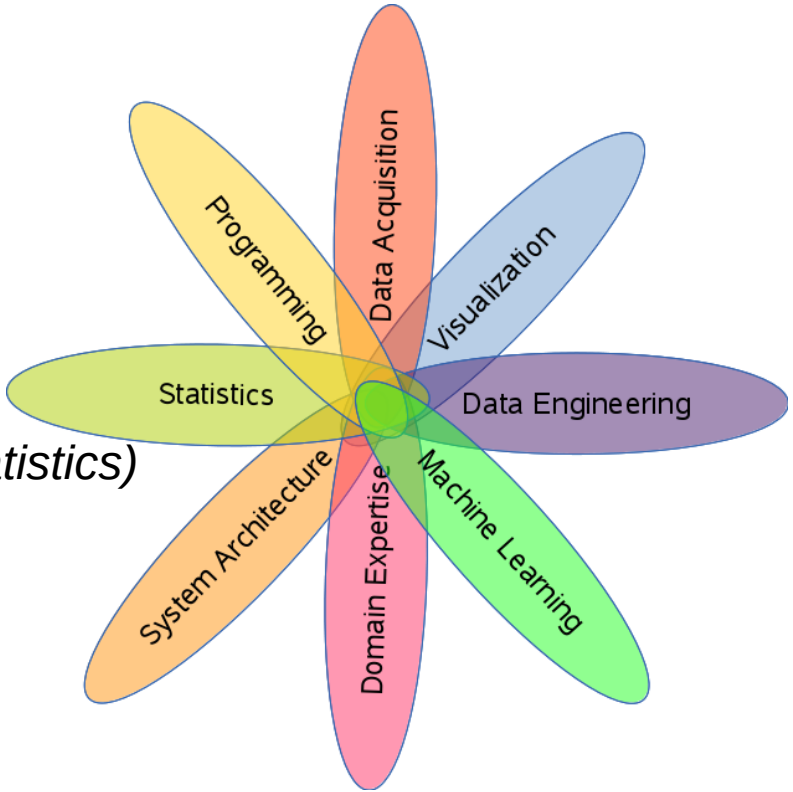


What is Data Science?

Hochschule Offenburg
offenburg.university

Typical Data Science Tasks:

- *Data Acquisition*
- *Data cleaning*
- *Storing data and making it accessible*
- *Selecting and pre-processing data*
- *Structuring and describing data (Descriptive Statistics)*
- *Exploration and Inference (Learning)*
- *Modeling and Abstraction*
- *Verification*
- *Visualization and Reporting*



What is Data Science?

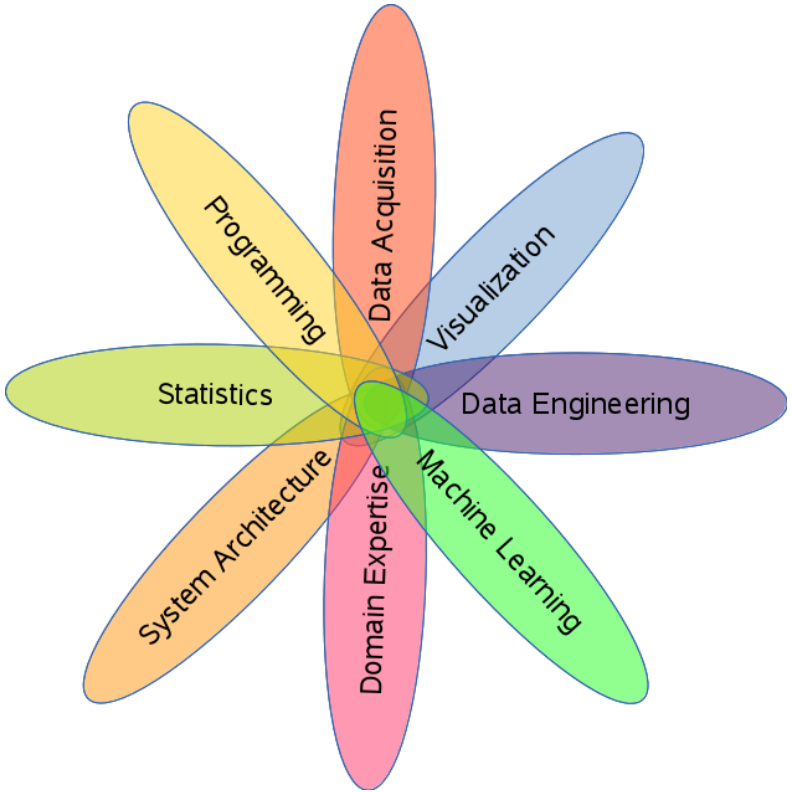
Hochschule Offenburg
offenburg.university

Motivation by tasks:

What is expected of Data Science?

What are typical tasks / applications?

→ **Data Science as mesh-up of many tasks and skills**



What is Data Science?

Data Science – Science of data?

Definition of Data:

Wikipedia:

Data (/ˈdeɪtə/ DAY-tə, /ˈdætə/ DA-tə, or /ˈdɑːtə/ DAH-tə) is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

What is Data Science?

Hochschule Offenburg
offenburg.university

Scientific perspectives at data:

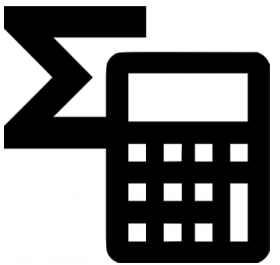
- **Looking at Data like a Statistician, you see the world in terms of**
  - Variables (sets)
  - (joint) Distributions
  - Densities
  - Correlations
  - Likelihoods
  - Hypothesis

What is Data Science?

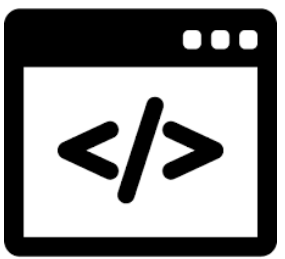
Hochschule Offenburg
offenburg.university

Scientific perspectives at data:

- **Looking at Data like a Mathematician, data forms**
  - Structures (Algebras, Sets, Groups, Graphs ...)
  - Spaces (i.e. high-dimensional Eucledian space)
  - Change (i.e. expressed in differential equations)
  - in practice: a lot of numerical optimization



- **Looking at Data like a Computer Scientist [Programmer]**
  - you think in data structures
  - algorithms and their efficiency
  - levels of abstraction
  - use cases (reusability)
  - parallelization



What is Data Science?

Hochschule Offenburg
offenburg.university

Scientific perspectives at data:

- **Looking at Data like a Data Engineer, you see**
  - Databases
  - Storage
  - System Architectures

- **Looking at Data like a Hacker, you think about**
  - how to get the data (from devices, web databases, streams ...)
  - how to extract the data (from binaries, proprietary formats ...)
  - how to use data in unconventional ways

What is Data Science?

Hochschule Offenburg
offenburg.university

Scientific perspectives at data:

- **Looking at Data like a Visual Artist, you think about**
  - how to visualize the key information hidden in the data
  - how to depict change
  - how to setting data into relation
  - how to present complex problems

- **Looking at Data as a Domain Expert**
  - Domain knowledge is important to set data in the correct meta-relation
    - i.e. helps to ask the "right" questions
    - helps to avoid trivial miss-interpretations (due to missing meta data/information)
  - But, domain knowledge also might be misleading and preventing data-driven discoveries

Key Skills for Data Science

Just be:

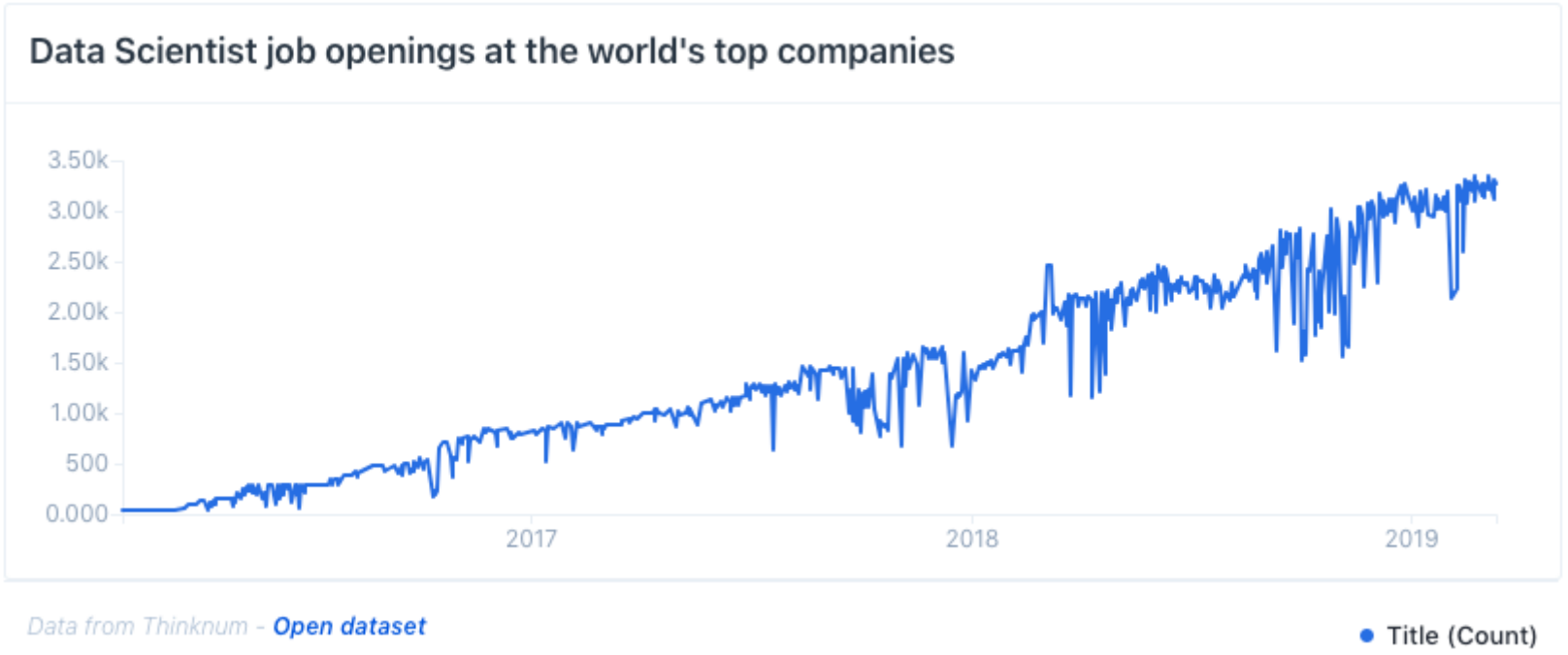
- A Statistician
- Mathematician
- Programmer
- Engineer
- Hacker
- Artist
- And Expert in many domains :-)

Key Skills for Data Science

On top of that:

- Very good communication skills
  - Presenting your work
  - Getting information from domain experts
- Be a quick learner
  - Adaption to new domains
  - Fast moving field (e.g. machine learning)

Job Opportunities Data Science



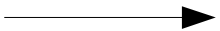
https://media.thinknum.com/articles/massive-increase-in-demand-for-data-science-jobs-in-2019/

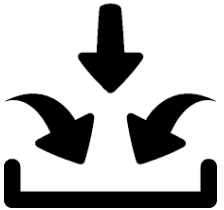
Data Science Workflow

Start: Hypothesis
Idea
Problem setting

Need: Basic understanding of the problem → Domain knowledge

Data Science Workflow



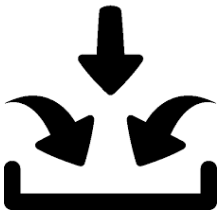


Start: Hypothesis

Collect Data

Data Science Workflow

Start: Hypothesis



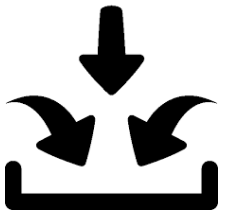
Collect Data

Exploratory Data Analytics (EDA)

- Visualize
- Pre-process
- clean

Data Science Workflow

Start: Hypothesis



Collect Data

Exploratory Data Analytics (EDA)

- Visualize
- Pre-process
- clean

Feature Engineering

Model building

- ML
- Statistic

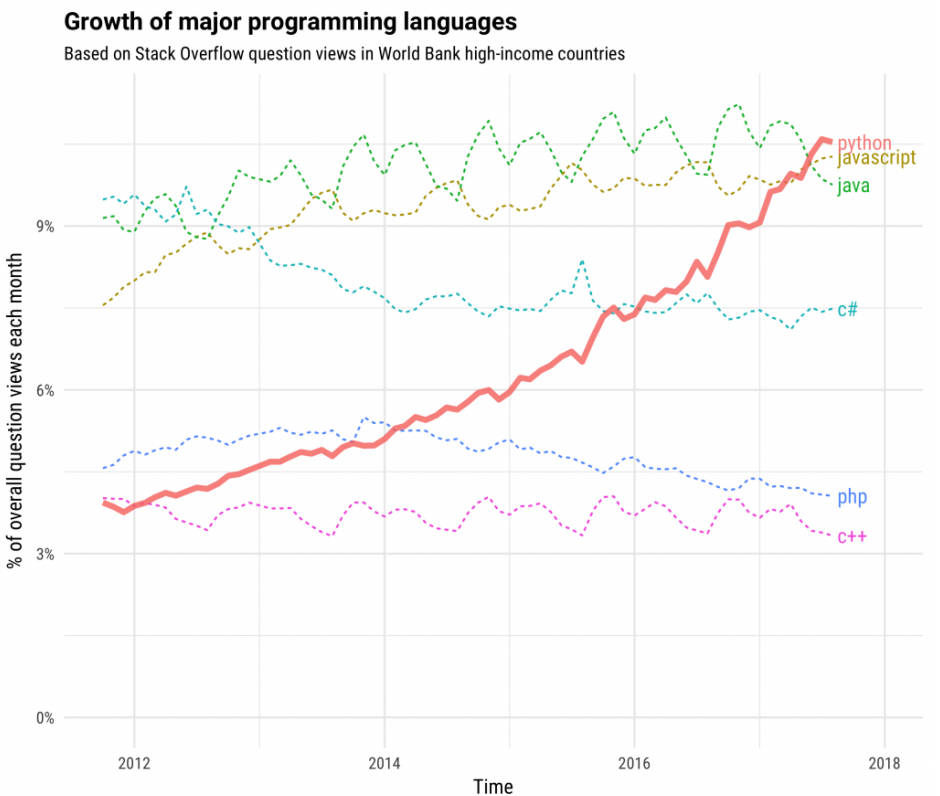
Python for Data Science

- We will use Python (more exactly tools with python interface) for this course

- Python is the most dominant language used in Data Science, Dats Analytics and Machine Learning to day

- Alternatives: R, Julia and Matlab

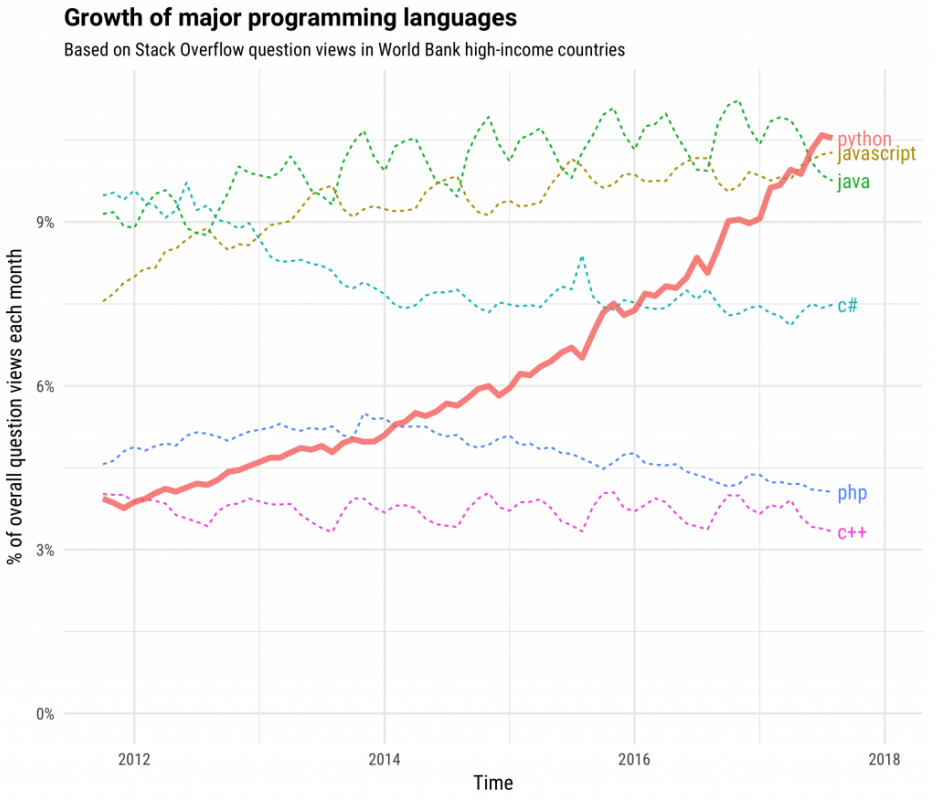
Python for Data Science



[2]

Python for Data Science

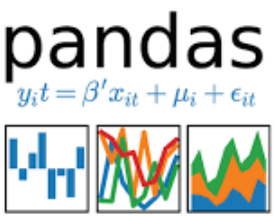
- Triggered by ML + Data Science
- Python pushes convergence of technologies:
  - HPC
  - Cloud
  - Big-Data



[2]

Python for Data Science

- Very low entry barrier (easy to learn)
- Quite universal in programming approaches
- Easy to interface existing high peformance libs
- Huge Community
- **Libraries !!!**
- **Python as „glue code“ for rapid development and markup**



Python for Data Science

https://www.anaconda.com/distribution/

Install on you local computer

- Easy installation on Linux/Mac/Win
  - See Howto in Moodle
  - User space install (no admin needed)
- Conda Python package management
  - Simple install/update/sharing of software
- All open source!

Python for Data Science

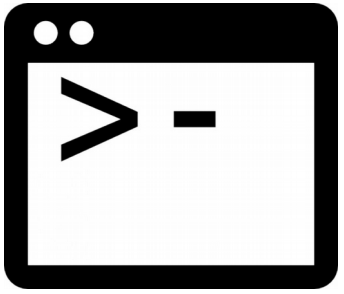
NOTE:

There are two versions of Python:

Python2 and Python3

Support for Python2 ended 31.12.19

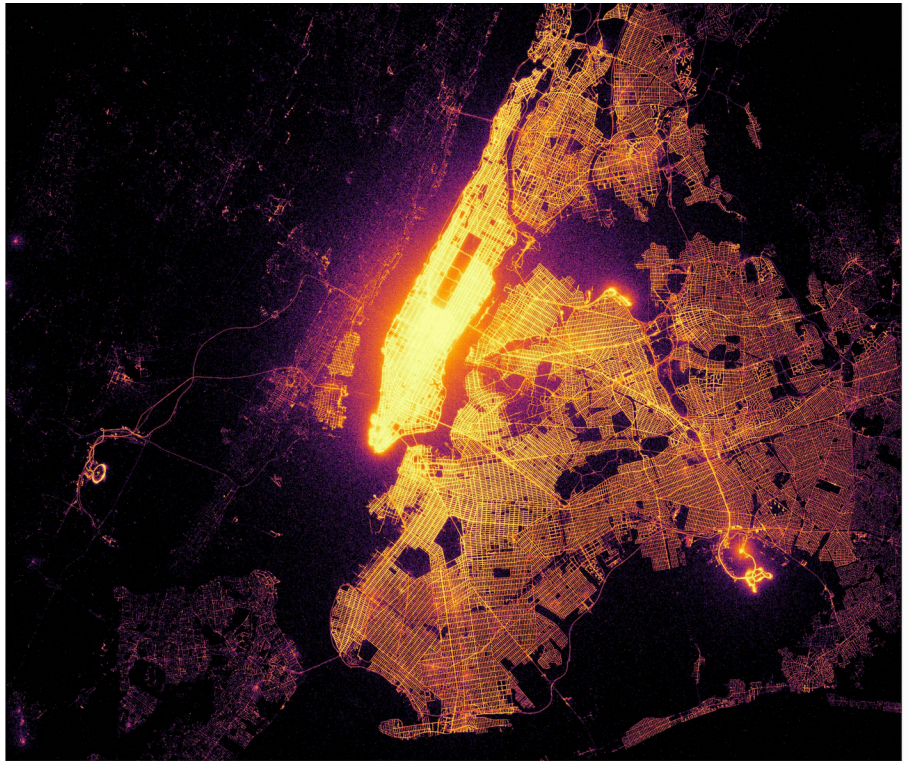
→ get started with Python3!



Interactive Python Introduction

→ Lab session Friday

Case Study I

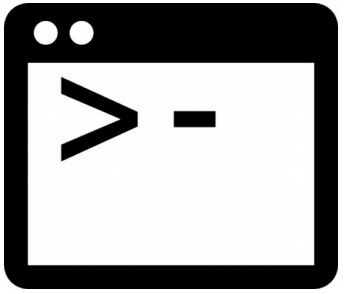


New York Taxi fare prediction

https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/overview

Image: https://towardsdatascience.com/if-taxi-trips-were-fireflies-1-3-billion-nyc-taxi-trips-plotted-b34e89f96cfa

Case Study I



Case Study I:

→ demo on Colab

Literature

[1] free icons taken from https://www.flaticon.com

[2] https://stackoverflow.blog/2017/09/06/incredible-growth-python/

[3] https://datahack.analyticsvidhya.com/contest/practice-problem-bigmart-sales-prediction/

[4] https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462#