

Tools Introductie



Maak werk van je toekomst

© 2023 Olivier Claerbout



Lesmateriaal

- Via GitHub account:
git clone https://github.com/oclaerbout/2024_DS1_Mechelen.git
- © 2024 Olivier Claerbout

Dit materiaal mag niet worden gekopieerd, verspreid, gepubliceerd of anderszins gereproduceerd zonder uitdrukkelijke schriftelijke toestemming van de auteur. Dit geldt eveneens voor de Jupyter Notebooks!

Maak werk van je toekomst

© 2023 Olivier Claerbout



Voorstelling van de tools

Doelen

- Intro tot Git
- Intro tot Docker
- Een Github en een Dockerhub account aanmaken
- Idee krijgen van de omgevingen waarin en waarmee we werken
- Introductie tot Pycharm
 - Github koppelen met Pycharm
- Introductie tot Jupyter Notebooks
- Eerste stukje code runnen in Pycharm en Jupyter

Intro tot Git

- 2005
- Linus Torvalds
- Had eerst BitKeeper (een ander VCS) gebouwd
- Bouwde de eerste versie in 3 dagen
- De source code van Git staat op GitHub:
 - <https://github.com/git/git>



Betekenis van Git?

- random three-letter combination that is pronounceable, and not actually used by any common UNIX command. The fact that it is a mispronunciation of "get" may or may not be relevant.
- stupid. contemptible and despicable. simple. Take your pick from the dictionary of slang.
- "global information tracker": you're in a good mood, and it actually works for you. Angels sing, and a light suddenly fills the room.
- "goddamn idiotic truckload of sh*t": when it breaks

GitHub

- 2007
- Internet hosting service voor code en versiebeheer
- Gratis varianten beschikbaar
- Betaalde varianten eerder voor bedrijven (meerdere mensen die kunnen werken aan een private repo enz)
- Microsoft 2018: USD \$7.5 miljard
- Diepere integratie tussen GitHub en tools zoals VS Code

GitHub

GitHub

- Git is open source software die versiebeheer uitvoert middels de git commandos
- Een voorbeeld is **git push**
- Git wordt veel gebruikt in de command line
- GitHub is de online hosting provider die de code die we op onze machine hebben kan hosten in wat we een repository noemen

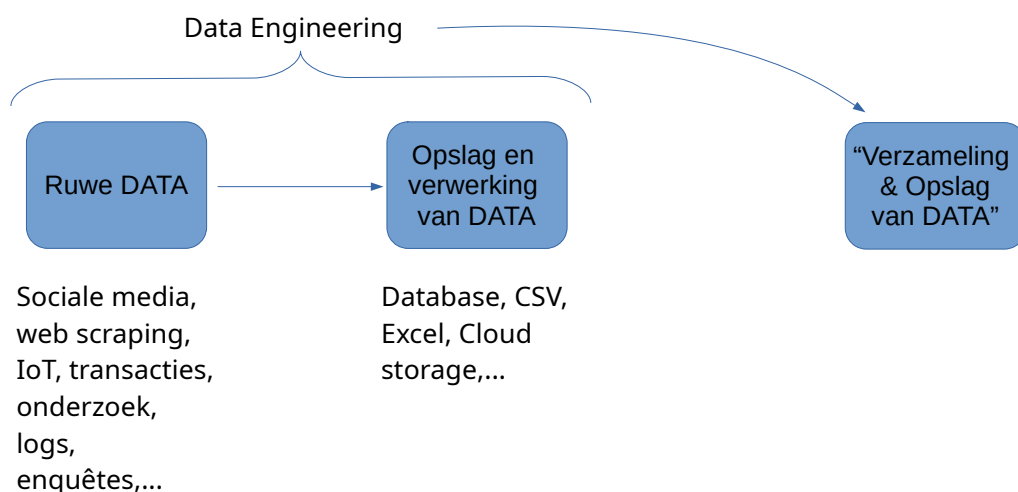
GitHub

- Wie nog geen GitHub profiel heeft:
 - maak er eentje aan
- Stuur jouw GitHub profiel per mail naar olivier@python.exposed, dan voeg ik jou toe aan de repo

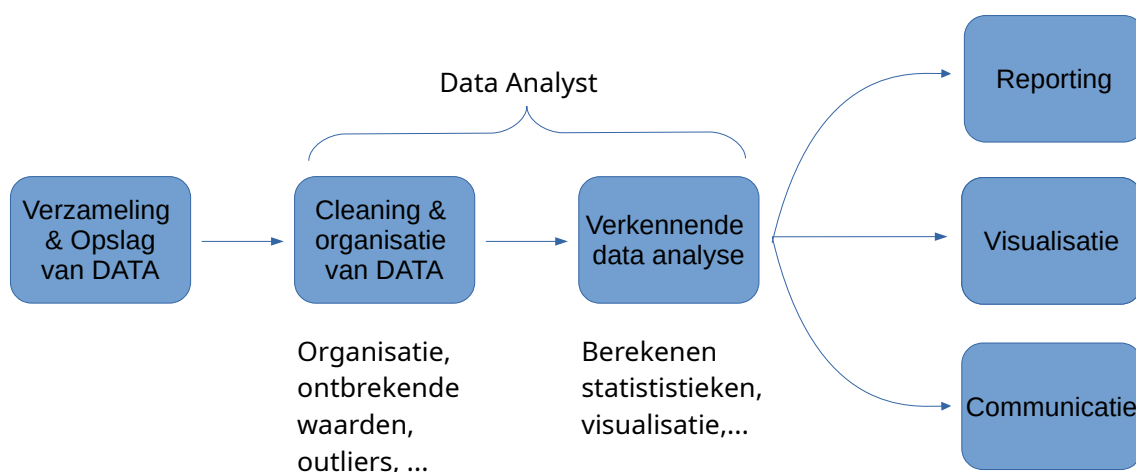
Wat is Data Science?

- Oplossen van problemen: hoe pas ik een parameter aan?
 - Data product: app, service, website,...
- Beantwoorden van vragen: hoe beïnvloedt deze wijziging iets anders?
 - Data analyse: rapport, visualisatie,...
- Overlap!

Wat is Data Science?

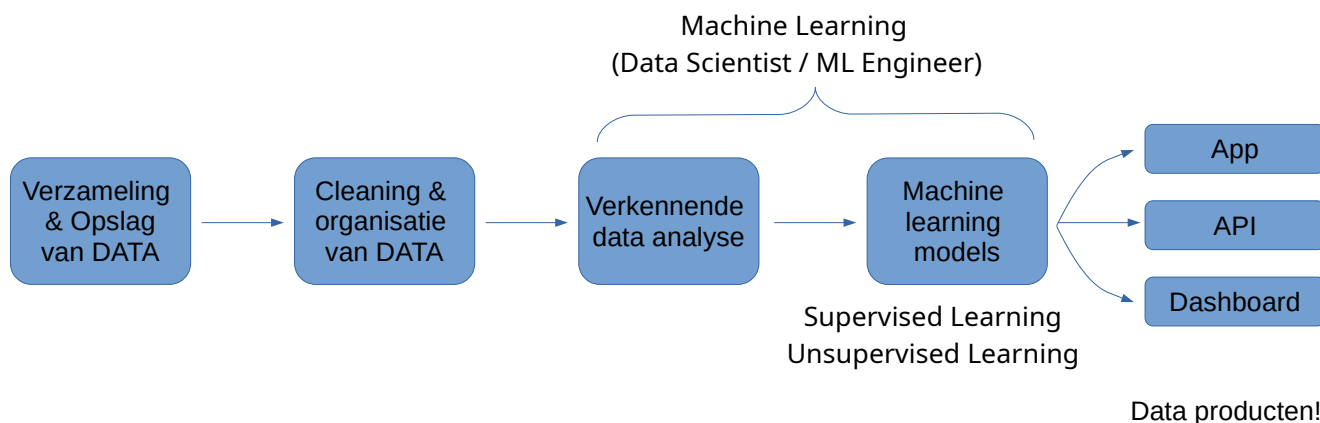


Wat is Data Science?



Misschien is de vraag hier reeds beantwoord...

Wat is Data Science?



Taken van een Data Scientist (in de praktijk)

- **Data Verzamelen en Opschonen:** verzamelen data uit diverse bronnen cleanen om inconsistenties, fouten en ontbrekende waarden te corrigeren. Dit proces staat bekend als data cleaning of data cleansing.
- **Data Verkennen en Analyseren:** exploratieve data-analyse om patronen, trends, en relaties in data te identificeren. Dit omvat het gebruik van statistische methoden en visualisatietools.
- **Machine Learning Modellen:** het ontwerpen, trainen, en implementeren van machine learning modellen om voorspellingen te doen of om patronen in data te identificeren.
- **Data Visualisatie en Rapportage:** visualiseren data en presenteren bevindingen in een begrijpelijke vorm om zakelijke besluitvormers te ondersteunen.

Taken van een Data Scientist (in de praktijk)

- Programmeren en Software: vaardigheden in programmeren (meestal Python) zijn essentieel om datasets te manipuleren, analyses uit te voeren, en algoritmes te ontwikkelen.
- Communicatie en Samenwerking: effectief communiceren met zowel technische als niet-technische stakeholders en vaak samenwerken met andere teams, zoals software-engineers, business analisten, en productmanagers.
- Probleemoplossing en Besluitvorming: Data scientists worden vaak geconfronteerd met complexe problemen en moeten creatieve oplossingen vinden en bijdragen aan data-gedreven besluitvormingsprocessen.

Opbouw van de cursus

Opbouw van de cursus – jaar 1

- Setup van de omgeving (Pycharm, Docker, Git, Jupyter Notebook)
- Python Intro
- Python voor data (databronnen,...)
- Intro databanken
- Numpy
- Pandas
- Data Exploration & Cleaning: Matplotlib, Seaborn
- Statistiek

Python intro

- Intro & expressies
- Variabelen
- Eenvoudige functies
- Condities
- Iteraties
- Functies
- Recursie
- Strings
- Lists
- Dictionaries
- Tuples
- Sets
- Tekstbestanden
- Regular Expressions

Setup van de omgeving

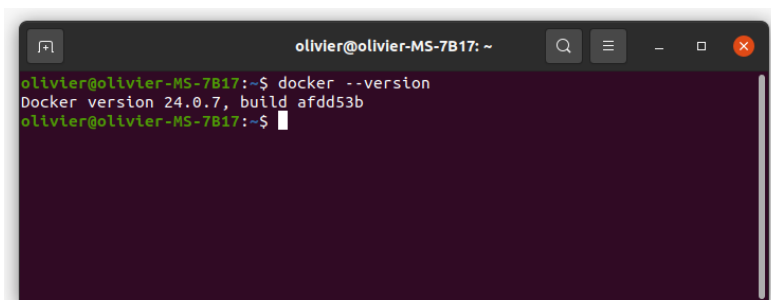


Setup van de omgeving

- Python
- Pycharm
- Docker
- Git
- Jupyter Notebook

Terminal?

- Command line (CLI) of command prompt
- Tekstgebaseerde interface
- Stelt gebruikers in staat om met het OS te communiceren door het invoeren van tekstcommando's
- Directe methode om taken uit te voeren en systeemfuncties aan te sturen



```
olivier@olivier-MS-7B17: ~  
olivier@olivier-MS-7B17:~$ docker --version  
Docker version 24.0.7, build afdd53b  
olivier@olivier-MS-7B17:~$
```

Python

- Eenvoudige syntax en leesbaarheid
- Guido van Rossum, voor het eerst uitgebracht in 1991
- Ondersteunt zowel objectgeoriënteerd als functioneel programmeren
- Open-source
- Grote en actieve gemeenschap
- Breed scala aan bibliotheken en frameworks
- Installatie:
 - <https://www.python.org/downloads/>

Installeer Python

Pycharm

- Populaire IDE
- Door JetBrains
- Specifiek voor Python
- Community en Professional editie
- (-) 'zwaarder' dan Visual Studio Code
- <https://account.jetbrains.com/a/7w50fxc8>
- Installatie:
 - <https://www.jetbrains.com/help/pycharm/installation-guide.html>

Installeer Pycharm Professional

Docker

- 'Container' systeem
- Applicatie en al zijn afhankelijkheden 'verpakken'
 - Consistentie: 'het werkt op mijn machine'
 - Veel lichter en efficiënter dan VM's
 - Snel deployment
 - Schaalbaar
 - Portabiliteit
- Installatie:
 - Check eerst \$ docker --version
 - <https://docs.docker.com/engine/install/>

Installeer Docker en run de 'hello—world' container

GIT

- VCS: Version Control System (cfr DMS in Sharepoint)
- Beheren van broncode en documenten
- Standaard in professionele softwareontwikkeling
- Installatie:
 - Check eerst \$ git --version
 - <https://git-scm.com/book/en/v2/Getting-Started-Installing-Git>

Installeer GIT

Maak werk van je toekomst

© 2023 Olivier Claerbout



Jupyter Notebook

- open-source webapplicatie
- live code, vergelijkingen, visualisaties en tekst
 - code, uitleg, en visuele inhoud combineren
- populair in data science, wetenschappelijk onderzoek, en onderwijs
- Voordelen:
 - Interactieve Data Analyse en Visualisatie
 - Integratie met Data Science en Machine Learning Tools

Via Docker!

Maak werk van je toekomst

© 2023 Olivier Claerbout



Cursusmateriaal

Verdeling cursusmateriaal

- Update op Git voor elke les
- Kies een folder waar je het cursusmateriaal wenst te bewaren
 - Vb: data_science / no_touch /
 - / workdir / lesmateriaal / ...
 - / notebooks / ...
- Initieel in terminal in no_touch folder:


```
git clone https://github.com/oclaerbout/syntra_data_scientist_brugge_2024.git
```
- Wekelijkse update in terminal in no_touch folder: `git pull`
- Breng nooit wijzigingen aan in de 'no_touch' folder (Git = VCS!)
- Copy/Paste de files in /workdir

- ```
olivier@olivier-MS-7B17: ~/workdir
olivier@olivier-MS-7B17:~/workdir$ docker run -it --rm -p 8888:8888 -v "${PWD}"/notebooks:/home/jovyan/work jupyter/scipy-notebook
```

- Figure 1**

<http://0019235178d0:8888/lab?token=c7d50013b862c90d83affa96b54422b2c914e0a2e28319d7>  
<http://127.0.0.1:8888/lab?token=c7d50013b862c90d83affa96b54422b2c914e0a2e28319d7>

