TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Prolog

Programming for Data Science

Methoden der Datenanalyse

Wer für Wen ???

Dresden Database Systems Group

Dozenten

- Dr.-Ing. Claudio Hartmann claudio.hartmann@tu-dresden.de
- Dipl.-Inf. Lucas Woltmann lucas.woltmann@tu-dresden.de

Vorlesung

Course	Informatik				Medieninformatik			Lehrexport (IST, WiWi, Lehramt, VIng)
	Bachelor	Master	Diplom PO2010	Diplom PO2004	Bachelor	Master	Diplom	Bachelor / Master
Programming for Data Science [PDS]	INF-B-510, INF-B-520	BAS-2, VERT-2, BAS-4, VERT-4, INF-PM-FOR		—	INF-B-530, INF-B-540	—	—	D-WW-INF-3421, D-WW-INF-3422, D-WW-INF-3423, INF-LE-WW, WI-MA-08-01, WI-MA-09-01
Komplexpraktikum Methoden der Datenanalyse [KP DA]	INF-B-510 INF-B-520	BAS-2, VERT-2, BAS-4, VERT-4		—	INF-B-530 INF-B-540	—	—	IST-05-KP

Anrechenbare Semesterwochenstunden

- PDS 2/2/0, KP DA 0/0/4

Weitere Informationen zur Vorlesung

Dresden Database Systems Group

Zeit

- Vorlesung: Dienstag, 2.DS (9:20 Uhr bis 10:50Uhr)
  - Zoom: https://tu-dresden.zoom.us/j/89460751682?pwd=MWZBQTlvMzZ5MU1UQlVsTGJNelJvUT09
- Übung/Konsultation: Bitte das Forum im Opal Nutzen, AMCS: https://amcs.website PIN: PDS2020

Skript und aktuelle Informationen

- Folien werden unter http://wwwdb.inf.tu-dresden.de zur Verfügung gestellt (Zugriff von außerhalb der TUD: Login: tud Passwort: dbs - und umgekehrt)
- Ankündigungen sind ebenfalls von http://wwwdb.inf.tu-dresden.de abrufbar

Rückmeldungen und Fragen

- Fragen, Anmerkungen, Kritik, Rückmeldungen sind immer erwünscht
- Kontakt per E-Mail oder während der Zoom-Session

Prüfung

- PDS: Mündliche Prüfung (Schwerpunkt Zusammenhangswissen)
- KP DA: Kolloquium am Ende des Semesters

Hand in your code – A Tale of Instruction

Dresden Database Systems Group

Please send your code to claudio.hartmann@tu-dresden.de

- Regardless of programming language
- Accepted file types: .py, .R, .zip (with one file in it, maybe two if necessary)
- No links! No repos! No data!
- Include your name, first name and task number in the file name.
- Deadline for all tasks: 05.02.2021 (We recommend a weekly hand-in cycle.)

Zeitplan

Dresden Database Systems Group

Oktober						
Mo	Di	Mi	Do	Fr	Sa	So
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

November						
Mo	Di	Mi	Do	Fr	Sa	So
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

Dezember						
Mo	Di	Mi	Do	Fr	Sa	So
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

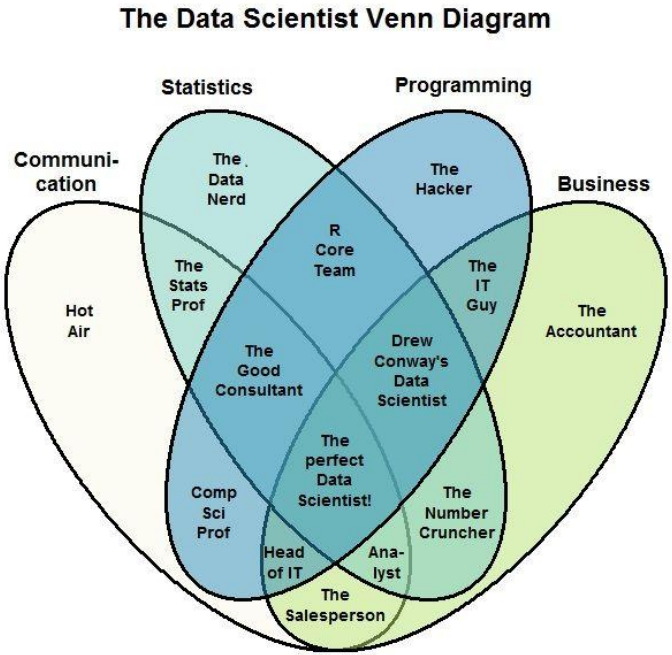
Januar						
Mo	Di	Mi	Do	Fr	Sa	So
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31
1	2	3	4	5		

- Heute
- Vorlesung
- Übung
- Ausfall
- Kolloquium

Ziel

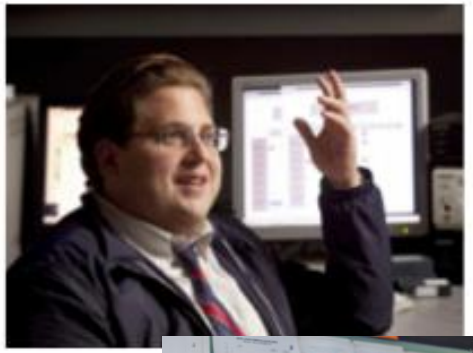
Dresden Database Systems Group

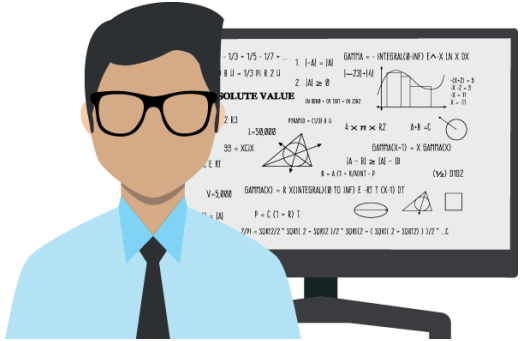
Fokus auf Data Science




Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

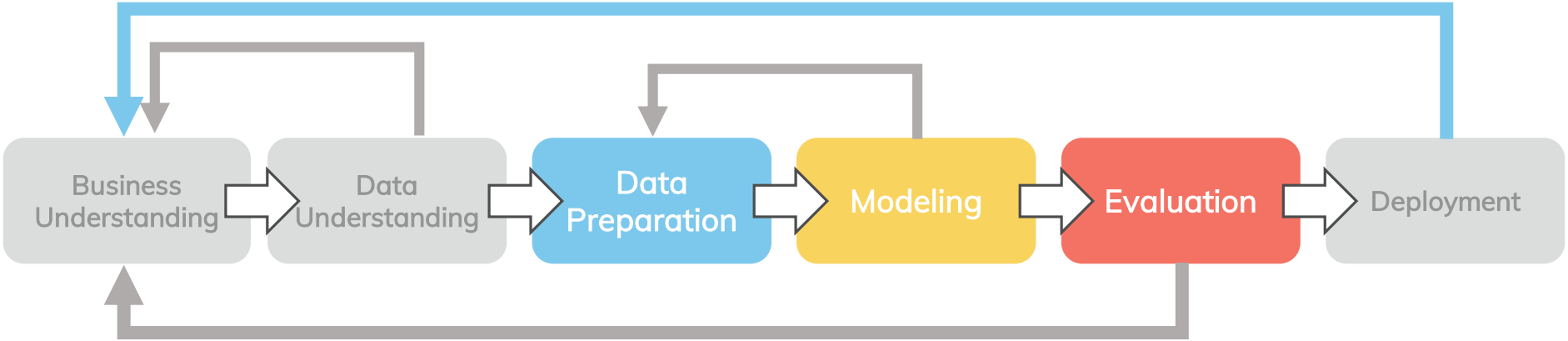




Structure Data Analytics

Dresden Database Systems Group

Cross-Industry Standard Process for Data Mining (Crisp-DM)



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Business Objectives Assess Situation Data Mining Goals Project Plan	Collect Initial Data Describe Data Explore Data Verify Data Quality	Select Data Clean Data Integrate Data Format Data	Modeling Technique General Test Design Parameter Settings Assess Model	Evaluate Results Approve Models Next Steps	Plan Deployment Plan Monitoring Plan Maintenance Report and Review

Struktur und Erwartungen

Dresden Database Systems Group

1. Data Preparation
2. Data Cleaning
3. Optimization Techniques
4. Regression Models
5. Classification
6. Clustering
7. Dimensionality Reduction
8. Association Rule Mining
9. Time Series Forecasting
10. Performance Optimization
11. & 12. Neural Networks

Warum seid Ihr hier? / Was erwartet Ihr von der VL?

Was erwarten wir von euch?

Voraussetzungen:

- Vorlesung Datenintegration und –analyse ist von Vorteil (Material SoSe2020 online verfügbar)
- Grundkenntnisse in R oder Python, bzw. schnelle Einarbeitung (für mehr Information siehe DIA-Vorlesung Analytic Tools)

R

What is R?

Dresden Database Systems Group

Statistical Programming Environment

- S is an environment for calculating and visualizing answers to statistical questions developed since 1976 at the Bell Labs
- R is an open source implementation of S

Advantages

- R is free (other than e.g. Matlab, Mathematica or SPSS)
- Runs on many (all) platforms
- Many methods for data scientists are already implemented in one of the many user developed packages
- New methods are often first developed in R
- IDE support (RStudio)

Disadvantages

- Needs some time to get used to

How to get R?

Dresden Database Systems Group

Download it - www.r-project.org

- Choose a mirror, and find a the right version for your OS
- Compile it yourself, the sources are available, too

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-03-11, Great Truth) R-3.5.3.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

The R Interface

Dresden Database Systems Group

The R Console

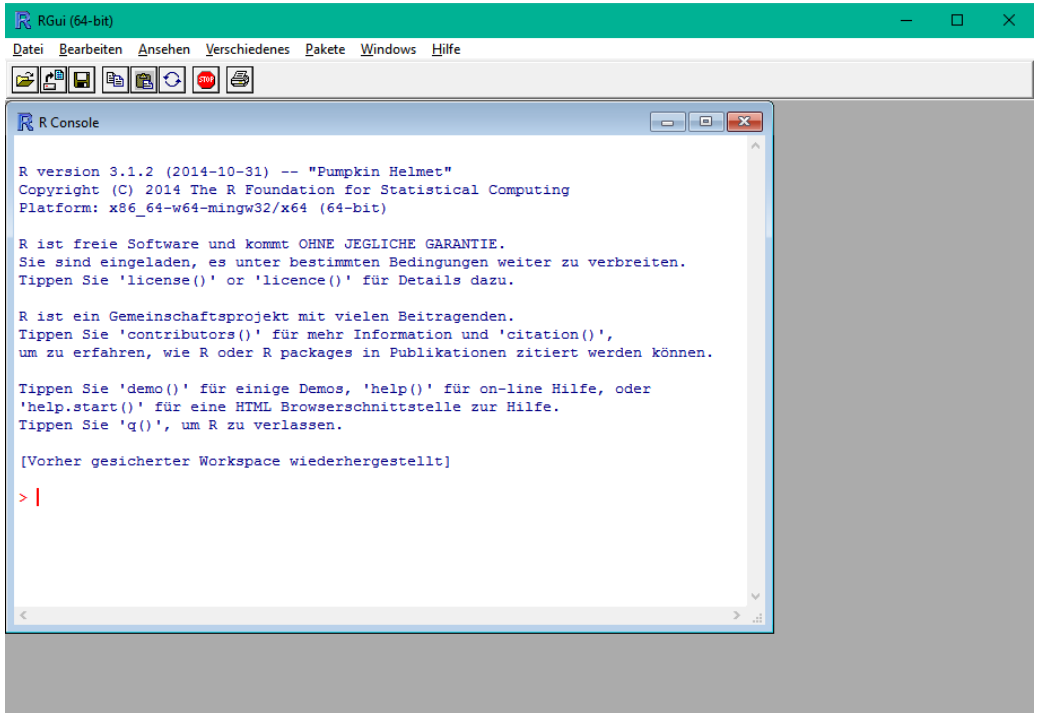
- Any R installation comes with the R console
- It‘s everything you need to work with R
- But IDEs and editors are very helpful
- Quit R: q()

R-Studio

- Commercial IDE for R
- Free open source edition
- Makes data exploration and organization significantly easier

Short reference

- https://cran.r-project.org/doc/contrib/Short-refcard.pdf



Basic Data Types and Operators

Dresden Database Systems Group

Basics

- Use <- or = for assignments, -> works as well
- Variable names
  - No !, +, -, #, “, ‘ but _ and . are fine
  - Numbers are fine, too, but not as the first sign
  - Case sensitive x and X are two different variables
- Dynamic typing

```
> x=1
> x
[1] 1
> y<-2
> y->x
> x
[1] 2
```

R is optimal to work with vectors/arrays and matrices

- Everything is a vector, until a more complex data type is explicitly used
- Every operation affects ALL vector elements, except only specific elements are selected

Sequences

- Use the seq()-function
- Or the :-notation

```
> x<-c(1,2,3)
> x
  [1] 1 2 3
> x<-seq(1,10,by=0.01)
> x
  [1]  1.00  1.01  1.02 …
> x<-1:10
> x
  [1]  1  2  3  4  5  6 …
> x*2
  [1]  2  4  6  8 10 12 …
> x[2]*2
  [1] 4
```

Python

What is python?

Dresden Database Systems Group

- Python is an open source interpreted programming language

Advantages

- Python is free (other than e.g. Matlab, Mathematica or SPSS)
- Runs on many (all) platforms
- Many methods for data scientists are already implemented in one of the many user developed packages
- New methods are often first developed in Python
- IDE support (PyCharm, juypter)

Disadvantages

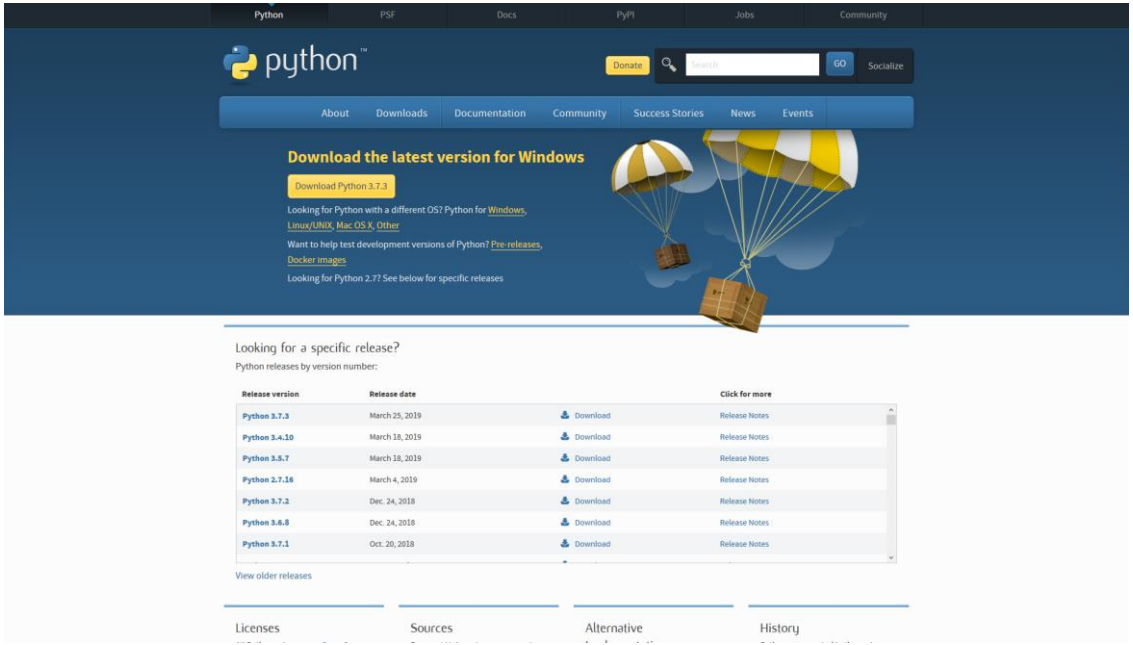
- Needs some time to get used to

How to get python?

Dresden Database Systems Group

Download it - https://www.python.org/

- Choose a version,
- Compile it yourself, the sources are available, too



Package Managers

Dresden Database Systems Group

Package managers are required to extend python's capabilities

pip

- https://pypi.org/project/pip/
- Comes with python (>= 3.4)
- Most common file manager → access to (almost) all packages
- Messy administration

Anaconda

- https://www.anaconda.com/distribution/
- Large bundle of most common packages (incl. juypter)
- Not always up-to-date or not all packages available

The python Interface

Dresden Database Systems Group

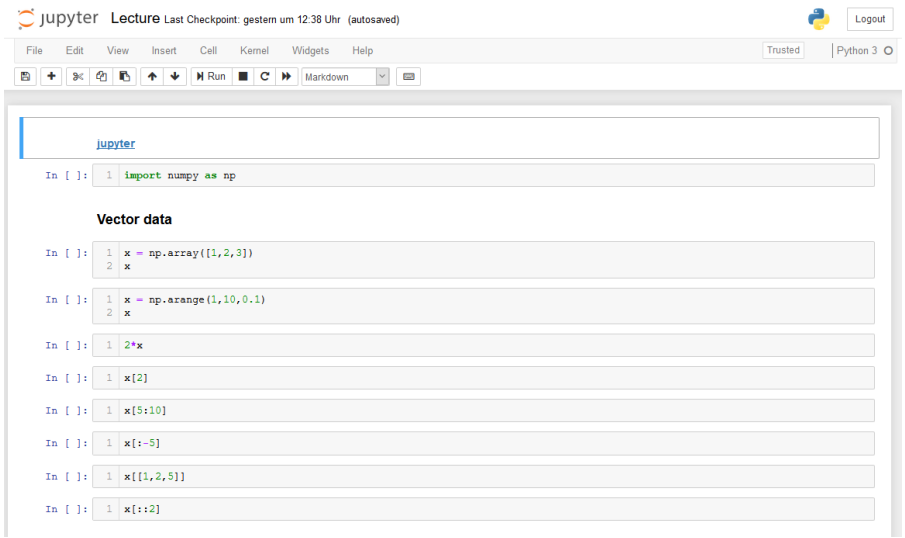
The python Console

- Any python installation comes with the python console
- It‘s everything you need to work with python
- But IDEs and editors are very helpful
- Quit python: quit() or Crtl + D

```
lucas@dbpostgres:~$ python
Python 3.6.8 |Anaconda custom (64-bit)| (default, Dec 30 2018, 01:22:34)
[GCC 7.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

jupyter

- Stateful editor for python
- Free open source edition
- Makes data exploration and organization significantly easier



Basic Data Types and Operators

Dresden Database Systems Group

Basics

- Use = for assignments
- Variable names
  - No !, +, -, #, “, ‘, . but _ is fine
  - Numbers are fine, too, but not as the first letter
  - Case sensitive: x and X are two different variables
- Dynamic typing

```
>>> x=1
>>> x
1
>>> y=2
>>> x=y
>>> x
2
```

Python can work with vectors and matrices through numpy

- The range of a sequence does not include the last element
- Every operation affects ALL vector elements, except only specific elements are selected

Sequences

- Use the numpy.arange function

```
>>> import numpy as np
>>> x = np.arrange(1,3)
>>> x
array([1, 2])
>>> x = np.arange(1,10,0.01)
>>> x
array([0.0, 0.01, …, 9.99])
```