



Prolog

Programming for Data Science

Methoden der Datenanalyse

Wer für Wen ???

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

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://wwwwdb.inf.tu-dresden.de> zur Verfügung gestellt (Zugriff von außerhalb der TUD: Login: tud Passwort: dbs - und umgekehrt)
- Ankündigungen sind ebenfalls von <http://wwwwdb.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

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

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	

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			

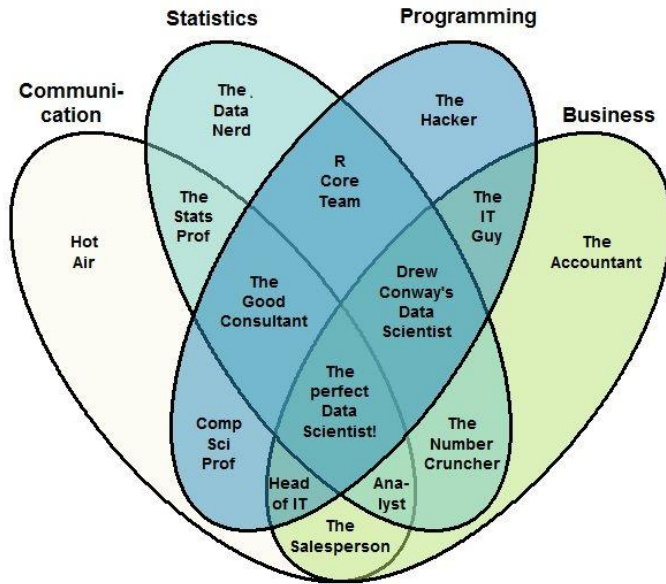
November						
Mo	Di	Mi	Do	Fr	Sa	So
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

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

Fokus auf Data Science

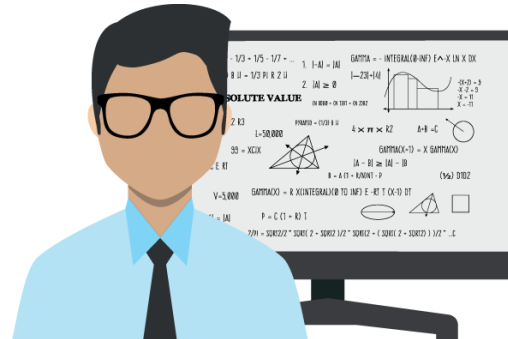
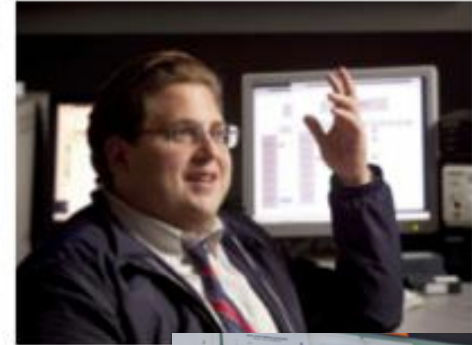
The Data Scientist Venn Diagram



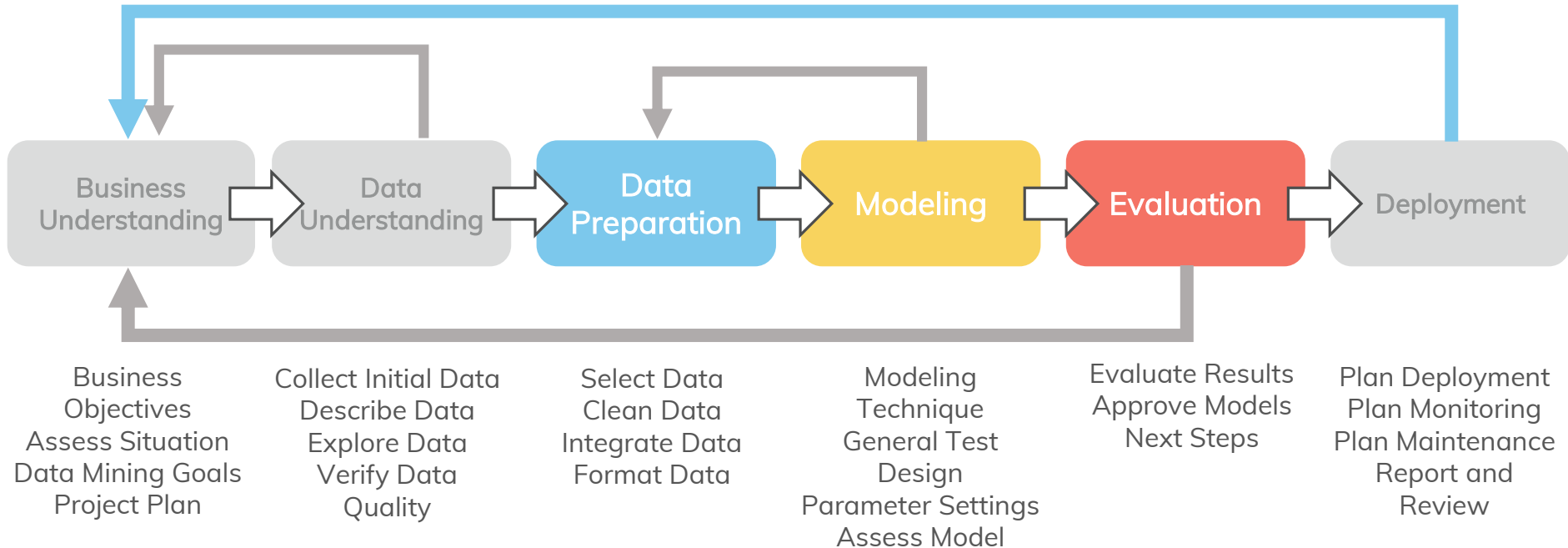
Data Scientist: The Sexiest Job of the 21st Century

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

When Jonathan Goldman arrived for work in June 2006 in London, the business was in a state of flux. He was like a start-up. The company had just under 10 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But when Goldman arrived, he was not alone. He was the only one who was not a member of the company. He was the only one who was not a member of the company. He was the only one who was not a member of the company.



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



Struktur und Erwartungen

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?

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?

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

The R Console

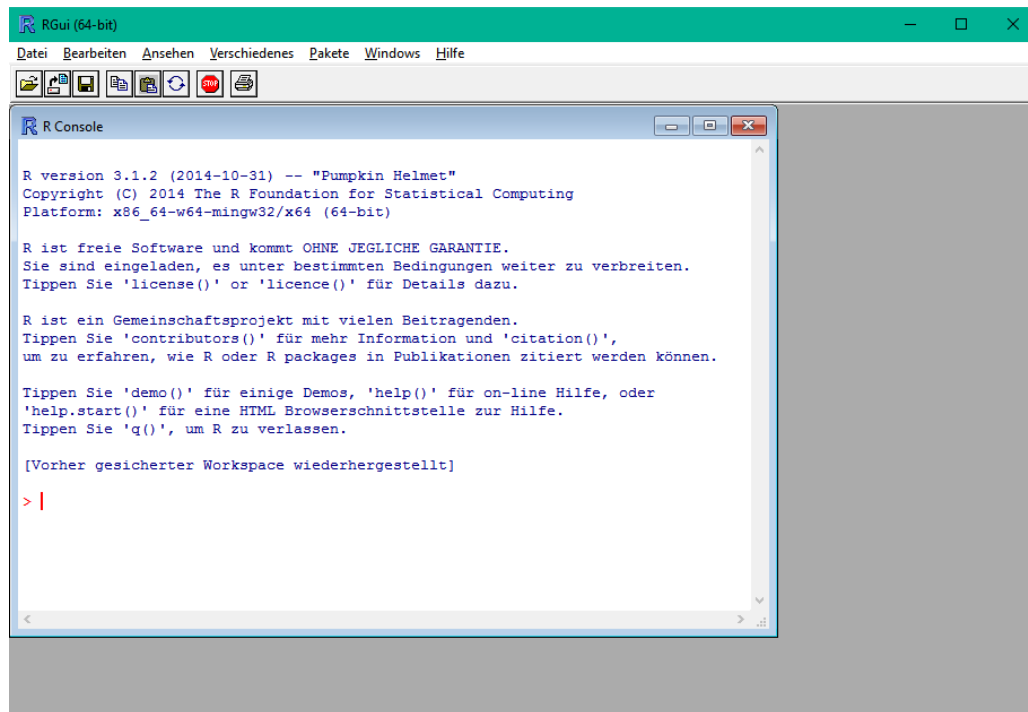
- Any R installation comes with the R console
- It's everything you need to work with R
- But IDEs and editors are very helpful
- Quit R: q()

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>



```
RGui (64-bit)
Datei Bearbeiten Ansehen Verschiedenes Pakete Windows Hilfe

R Console

R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreiten.
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden können.

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.

[Vorher gesicherter Workspace wiederhergestellt]

> |
```

Basic Data Types and Operators

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

```
> 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
```

Sequences

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



Python

What is python?

- 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, jupyter)

Disadvantages

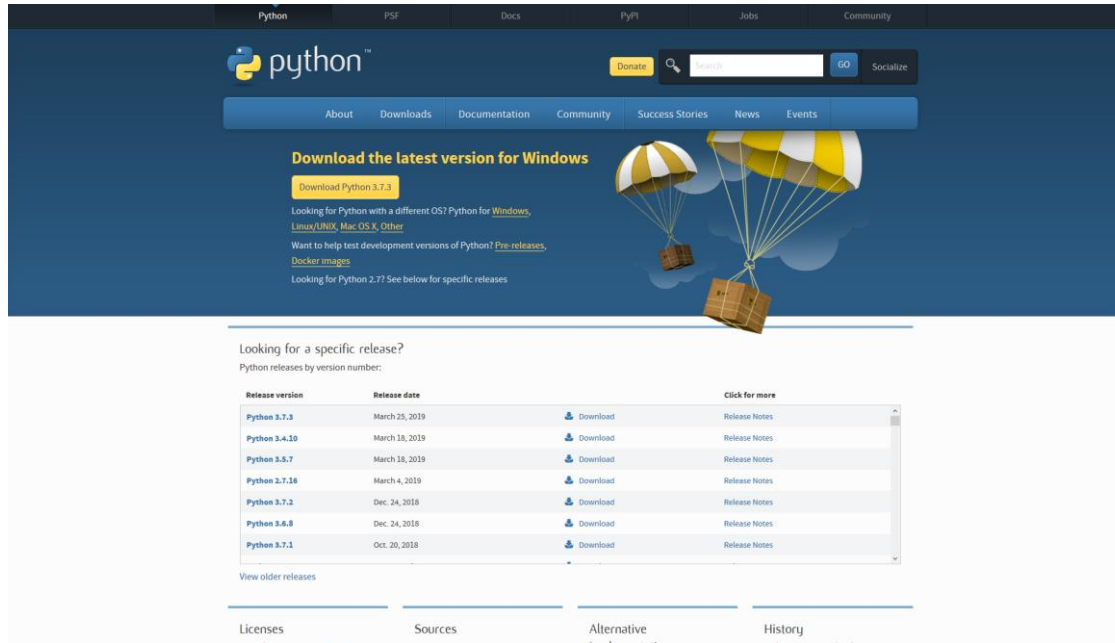
- Needs some time to get used to



How to get python?

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

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



The screenshot shows the Python.org website. The main heading is "Download the latest version for Windows" with a button to "Download Python 3.7.3". Below this, there are links for other operating systems: "Looking for Python with a different OS? Python for Windows, Linux/UNIX, Mac OS X, Other". There is also a link for "Pre-releases" and "Docker images". A section titled "Looking for a specific release?" shows a table of Python releases by version number.

Release version	Release date	Download	Click for more
Python 3.7.3	March 25, 2019	Download	Release Notes
Python 3.4.10	March 18, 2019	Download	Release Notes
Python 3.5.7	March 18, 2019	Download	Release Notes
Python 2.7.16	March 4, 2019	Download	Release Notes
Python 3.7.2	Dec. 24, 2018	Download	Release Notes
Python 3.6.8	Dec. 24, 2018	Download	Release Notes
Python 3.7.1	Oct. 20, 2018	Download	Release Notes

View older releases

At the bottom, there are links for "Licenses", "Sources", "Alternative", and "History".

Package managers are required to extend python's capabilities

pip

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

Anaconda

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

The python Interface

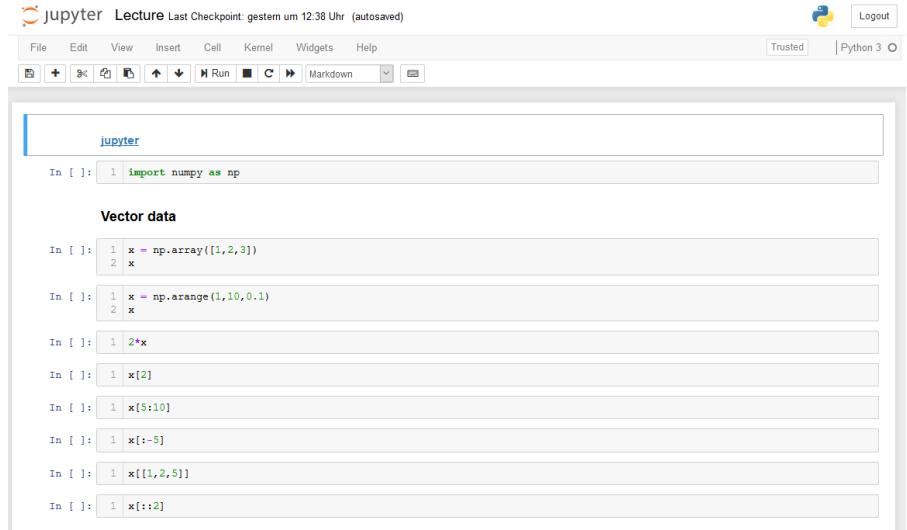
The python Console

- 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 Ctrl + 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



The screenshot shows the Jupyter Notebook interface. At the top, there's a header with the Jupyter logo, the word "jupyter", and a status bar indicating "Lecture Last Checkpoint: gestern um 12:38 Uhr (autosaved)". Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, Help. To the right of the menu bar are buttons for "Trusted" and "Python 3". Below the menu bar is a toolbar with icons for various actions like saving, undo, redo, and running code. The main area contains a series of code cells. The first cell is empty. The second cell contains the code `import numpy as np`. The third cell is titled "Vector data" and contains the code `x = np.array([1,2,3])`. The fourth cell contains the code `x = np.arange(1,10,0.1)`. The fifth cell contains the code `2*x`. The sixth cell contains the code `x[2]`. The seventh cell contains the code `x[5:10]`. The eighth cell contains the code `x[1:-5]`. The ninth cell contains the code `x[[1,2,5]]`. The tenth cell contains the code `x[:,2]`.

Basic Data Types and Operators

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

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

Sequences

- Use the numpy.arange function