



DIS08 – Data Modeling

01 - Introduction

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: SS2020

Information Retrieval Research Group



Prof. Dr. Philipp Schaer

- Information retrieval, evaluation of IR systems, digital libraries



Timo Breuer, M.Sc.

- Living Labs Infrastructure for Information Retrieval, project STELLA



Ina Böckmann

- Student assistant for Data Modelling

Some notes on language matters...

Most of the content of this course will be in English:

- The **slides** and **worksheets** of this course will be in English.
- The **text book** we will use, is freely available in English.
- The **additional materials** and **web resources** we point to are mostly in English.
- The **final exam** will be in English.

But:

- The **lecture itself** will be in German!
- The **text book** is also available in German (but you have to pay).
- You can still answer the questions in the **final exam / assignments** in German.



Text book – What text book?

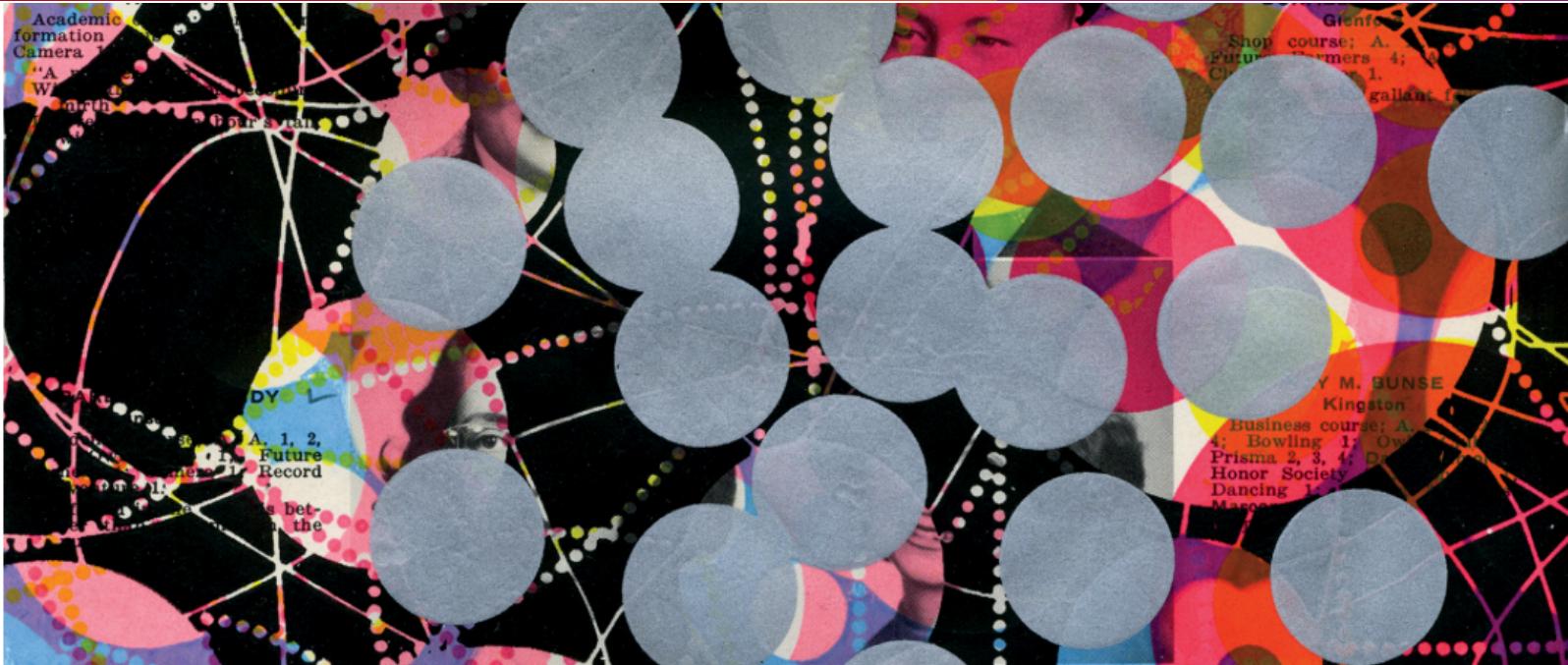
- This textbook covers a lot of technical stuff we will work on in the second half of the semester.
- Prof. Strahringer uses the same text book in her programming course!
- Read it for free under:
<http://automatetheboringstuff.com>
- Buy it for 20 EUR!
- For just \$10 you can get access to the online video lectures:
https://www.udemy.com/automate/?couponCode=FOR_LIKE_10_BUCKS



It's the data, stupid!



“Daten sind der Rohstoff
des 21. Jahrhunderts”



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

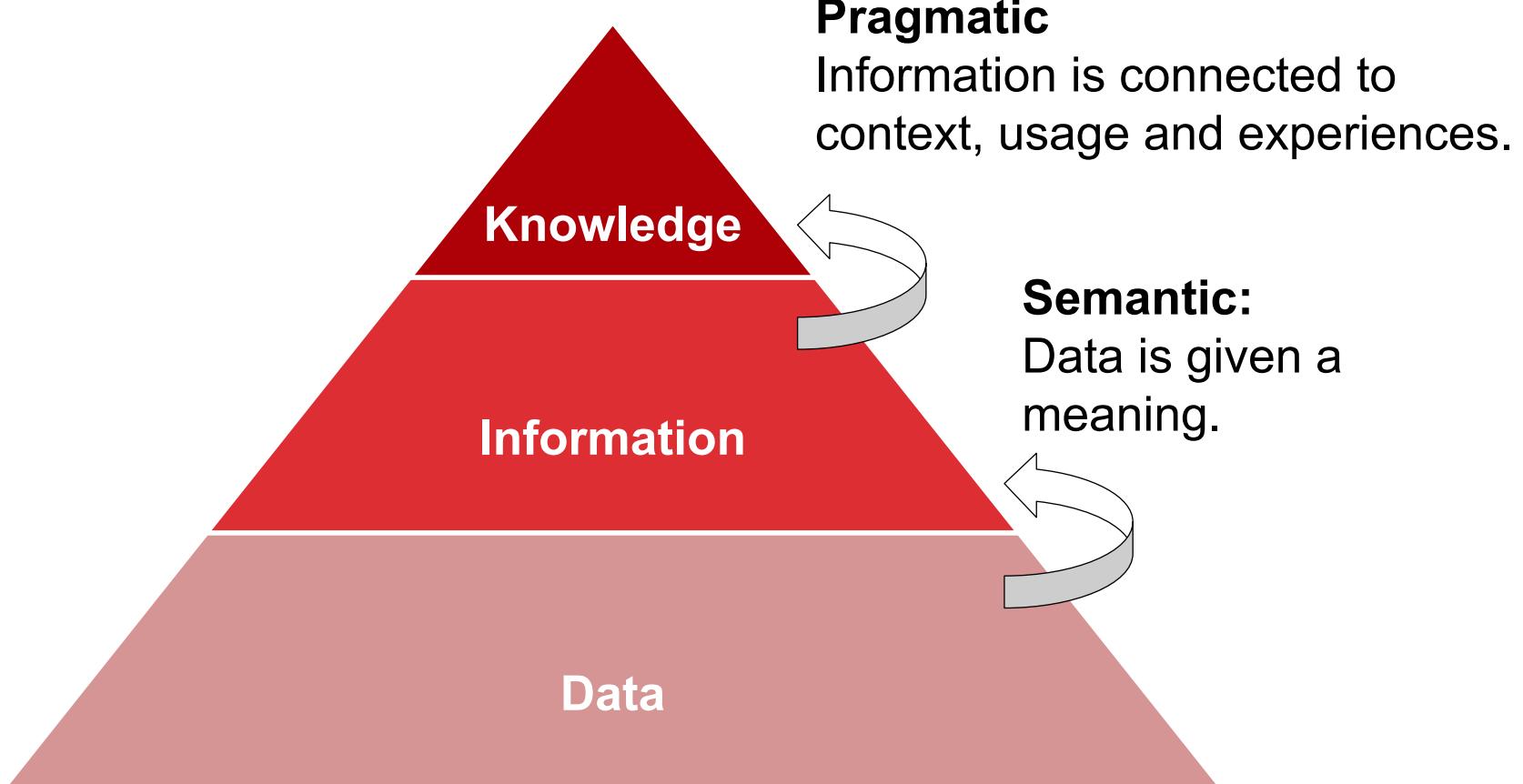
SUMMARY SAVE SHARE COMMENT (13) TEXT SIZE PRINT \$8.95 BUY COPIES

WHAT TO READ NEXT

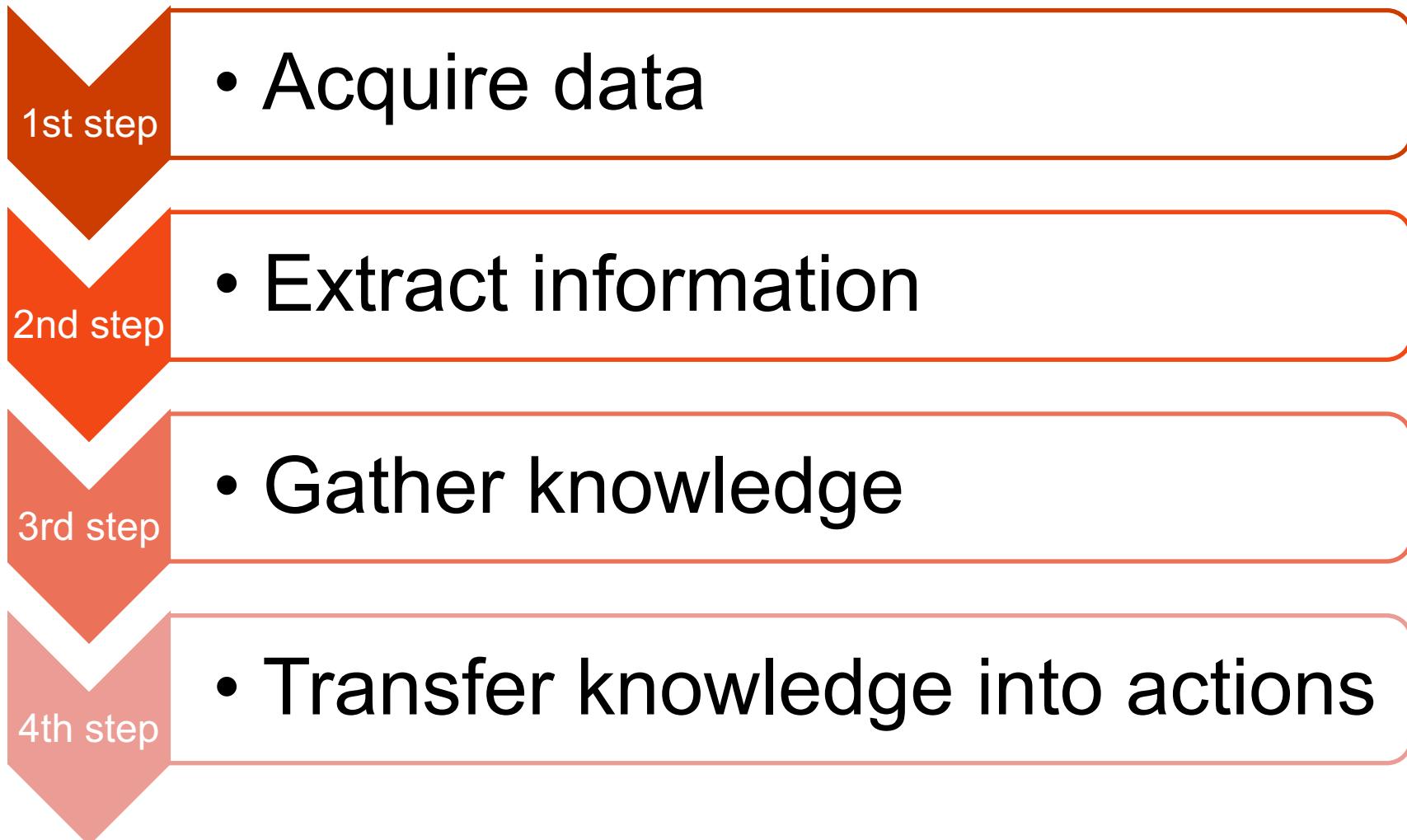


Big Data: The Management Revolution

The Knowledge Pyramid



A transformation pipeline



Learning outcomes

- **(WHAT)** Students learn to **process** and **structure** data and information that is available in electronic form and to convert it into common formats.
- **(HOW)** For this they use **different formats** (e.g. CSV, XML or JSON), **automated transformations** (e.g. with XSLT or on the command line) and editors (e.g. Notepad++).
- **(WHY)** This enables them to process any source data in such a way that it can be used for later applications, e.g. as input for database and retrieval systems or for data mining. They know typical procedures, tools and formats to use the results of their preparation and modelling flexibly. Furthermore, they can adapt them according to the application and requirements.

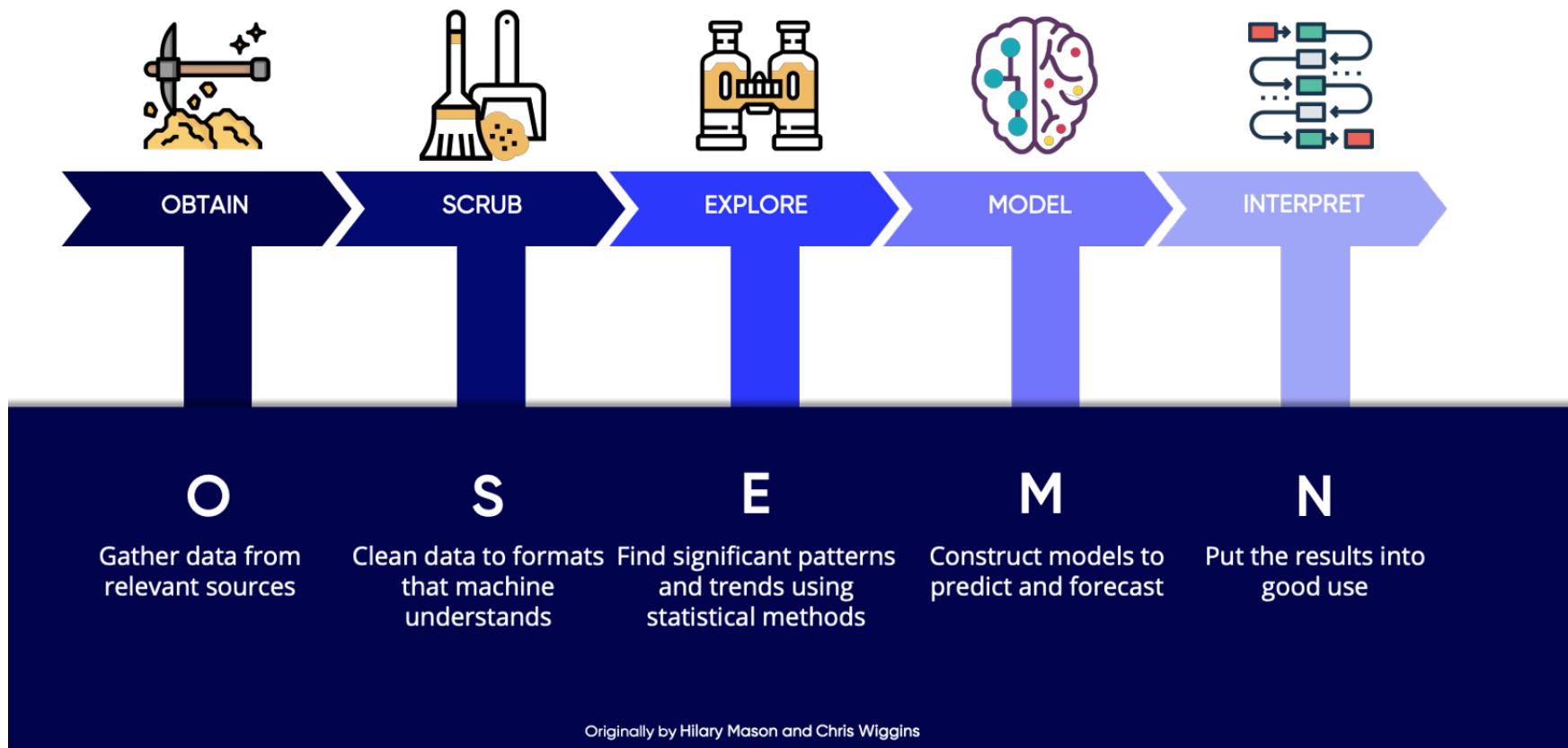
In this course...

You will learn how to

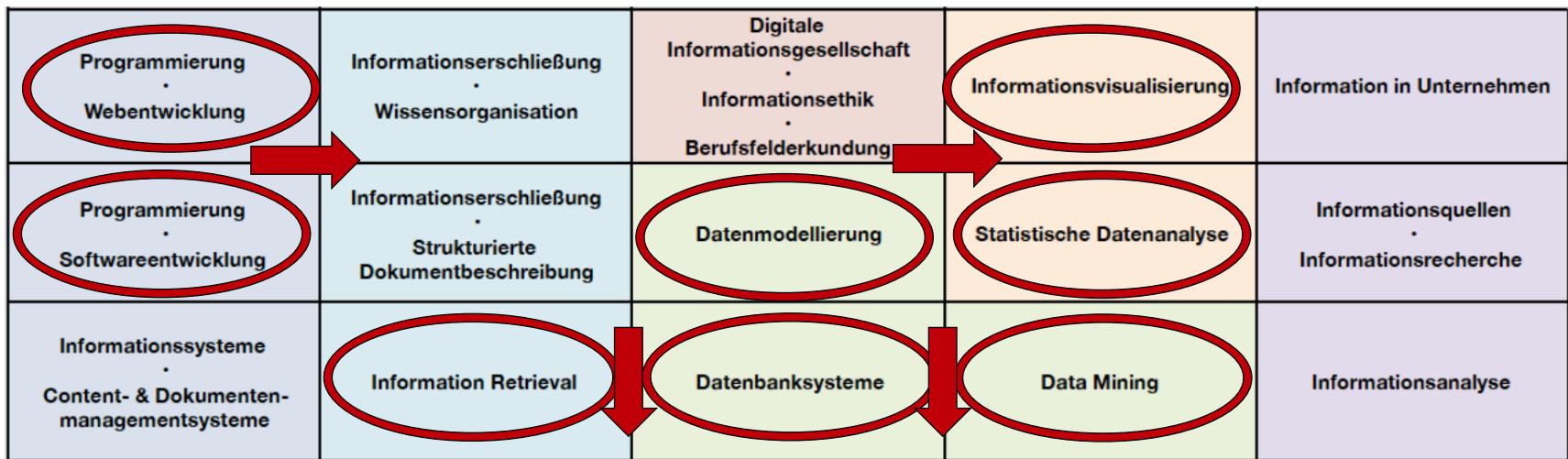
- **extract data** from various sources,
 - **transform data** in different forms and formats,
 - **load data** into applications to make use of it.
-
- This process is known as the Extract-Transform-Load (**ETL**) process often used in **database workflows**.
 - We will extend this process to the **OSEMN principle** which is more related to typical **data science usecases**.



Data Science Process



How does it integrate into DIS?



In this course...

We will work with a **data-science-friendly tech stack** using

- **GitHub** (code repository)
- **UNIX shell / bash** (commandline interactions with the system)
- **Regular expressions** (the Swiss army knife of data processing)
- **Python** (for small scripting / programming tasks)
- Other open-source tools (OpenRefine, etc.)
- ...

On top of that we try to teach you about being a **hacker!**

What is Hacking?

“Hacking is a term that originated in MIT in the late 1950’s in what was then called the Tech Model Railroad Club, where the term **hack** or **hacking** was used in terms of tinkering with machines in ways that surpass the user manual. The term hacking in itself is so exciting that those really interested in solving problems and bringing new ideas to life embrace the term **hacking into** disregarding the fact whether they are judged by the two-timers that break into other peoples computers for fun. Taking things apart and fixing them for purposes of improvement or better efficiency is what Hacking is all about.”

How to get a Hacker Mindset?

- Cultivating a hackers mindset is not just simple, but an **effortless endeavour**. All you have to do is be ready.
- You have to be ready to be wrong. You have to be ready to screw up. You have to be ready to get your hands dirty.
- **Get acquainted with your craft.**
- Hacking doesn't just relate to machines or electronics and computers. You can be a hacker no matter what you do.
- As long as you **obsessively try to solve problems**, and you are not afraid of failing and coming up with **innovative solutions**, you are a **hacker**.
- *The people who break into other people's computers are called **crackers**.*

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE
EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)

		HOW OFTEN YOU DO THE TASK					
		50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
HOW MUCH TIME YOU SHAVE OFF	1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
	5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
	30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
	1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
	5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
	30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
	1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
	6 HOURS				2 MONTHS	2 WEEKS	1 DAY
	1 DAY					8 WEEKS	5 DAYS

Example: Keyboard shortcuts

- Though we will get more computational over the course of the program, we can start our adventure into programming with very simple things like keyboard shortcuts.
- We all have our favorites that are labor saving but also allow us to use this stupid machine in the best possible way.
- You can do all the lessons without keyboard shortcuts, but note that they'll likely come up a lot.

Action	Windows	Mac	+ Keystroke
Save	Ctrl	Command	+ S
Copy	Ctrl	Command	+ C
Cut	Ctrl	Command	+ X
Paste	Ctrl	Command	+ V
Switch Applications	Alt	Command	Tab

Announcements



Open student positions in my team

I am looking for **up to two student assistants** for my DFG-funded project **STELLA**

- approx. 6-10 hrs./week (up to you)
- 10 EUR/hour
- free time scheduling (up to you)

What to do?

- It depends...
- **Supporting our researchers with small programming tasks**
- **Data conversions** (XML, JSON or HTML)
- **Software tests** and bug reporting, ...

→ Send a short motivation letter plus a 1-page CV

Scoring and exams

There are two different parts of your exams

1. Practical assignments

- **1st submission** (25 points) – Do it on your own!
Includes all worksheets from the first half of the semester.
- **2nd submission** (25 points) – Teams of up to 3 people!
Includes all Python programming tasks from the second half.

2. Final exam (60 minutes, 50 points)

- You can gather a total of **100 points**
- Your finals grade is determined using this table:

<50	50	55	60	65	70	75	80	85	90	95
n.b.	4,0	3,7	3,3	3,0	2,7	2,3	2,0	1,7	1,3	1,0

Rules for this course

No email support! Ask questions online!

- There will be a Microsoft Teams channel for this course - Use it!!
- I know – Nobody wants to be the one asking “stupid” questions.
- But: Your fellow students have the same issues – Trust me!
- Ask a lot of questions and try to **help your fellows**.

Help each other! Help yourself!

- Tutorial sessions are **interactive** – get the most out of it!
- **Active helpers** will get some **extra points** at the end of the semester!

Be active on GitHub.

- We want to see **weekly commits** in your account.
- A last-minute commit for the assignment is **suspicious**... And we might come up with the idea to ask **embarrassing questions**.

Schedule for lectures SS 2020

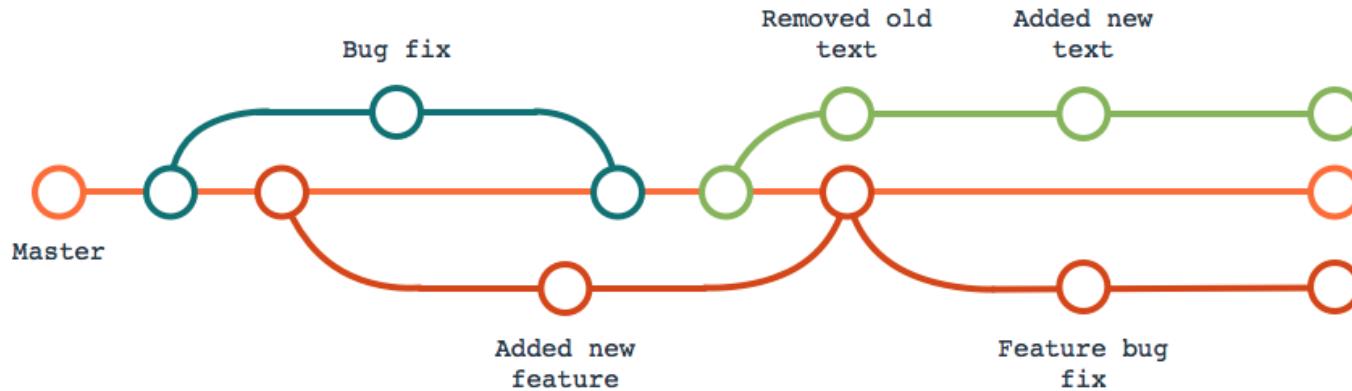
02.04.20	Introduction, Markdown	first week (no tutorial)
09.04.20	Unix Shell	Markdown and style guides
16.04.20	Versioning, Git, GitHub	Shell tutorial
23.04.20	Regular Expressions	Hands-on Git, GitHub
30.04.20	OpenData, CSV, Excel	Regex tutorial
07.05.20	Python: Data structures	open.cologne open data
14.05.20	Projektwoche (no lecture)	Projektwoche (no tutorial)
21.05.20	Christi-Himmelfahrt (no lecture)	Christi-Himmelfahrt (no tutorial)
28.05.20	Python: Files, folders and more	Hands-on Python
04.06.20	Python: Structures file formats	Hands-on Python
11.06.20	Fronleichnam (no lecture)	Fronleichnam (no tutorial)
18.06.20	Python: Web-Scraping	Hands-on Python
25.06.20	Q&A - Summary	Hands-on Python



First steps: GitHub + Markdown

GitHub – In a nutshell (more next week)

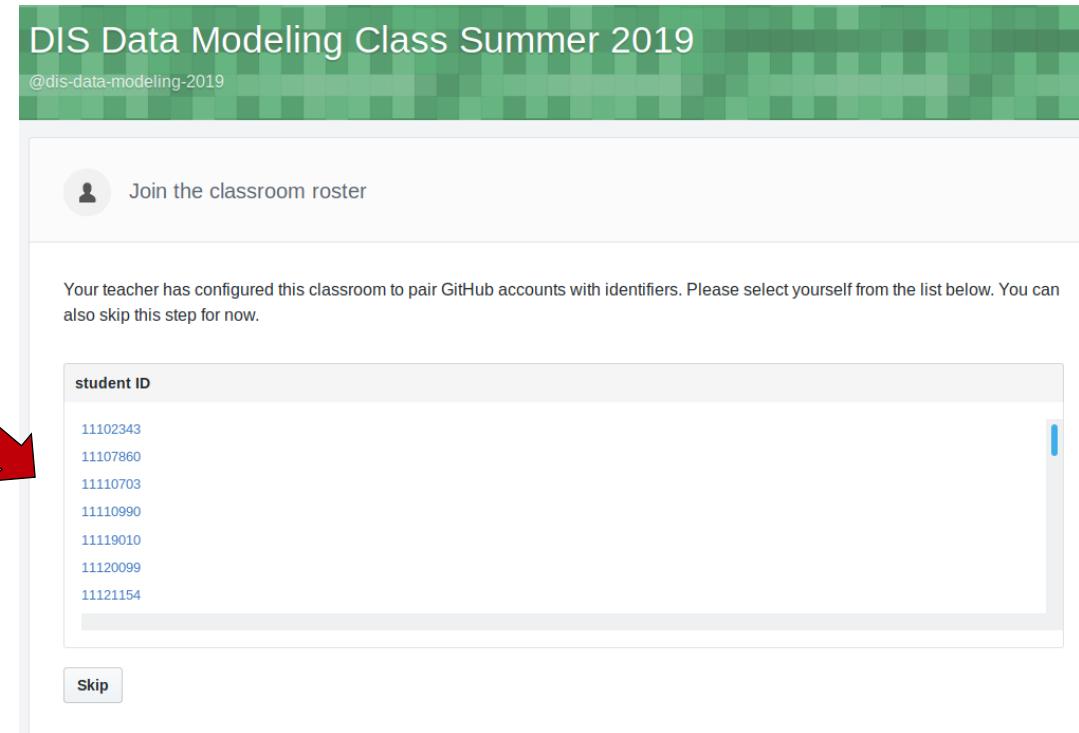
- **Git** is a version control system.



- **GitHub** is an internet platform that offers free Git access, user accounts, disk space, etc. ...
- We use GitHub in this course to see your progress, gather your assignments and to introduce you to one of the state-of-the-art systems for version control.

First task: Create a GitHub account

- Create a GitHub account: <https://github.com>
- Join our GitHub classroom:
<https://classroom.github.com/a/no15MntJ>
- Select your student register number from the list.



DIS Data Modeling Class Summer 2019
@dis-data-modeling-2019

Join the classroom roster

Your teacher has configured this classroom to pair GitHub accounts with identifiers. Please select yourself from the list below. You can also skip this step for now.

student ID

- 11102343
- 11107860
- 11110703
- 11110990
- 11119010
- 11120099
- 11121154

Skip

dis-data-modeling-2019 / assignment-1 Private

Unwatch 2 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

First Assignment, Submission: 16-05-2019 Edit

Manage topics

2 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find File Clone or download

neumannm Update part 1 instructions Latest commit 1adbf1a 11 minutes ago

README.md Update part 1 instructions 11 minutes ago

README.md

Assignment 1

Part 1

Create a new file in this repository via the web interface. Use `introduction.md` as file name.

Inside the file, first enter the following information:

- **Name:** {your name here}
- **E-Mail:** {your smail address here}

Then, answer the following questions:

1. Do you have any prior experience in the field introduced in the introductory session? If so, give a short overview on them.
2. What expectations or wishes do you have on the data modeling course?

Make use of some Markdown formatting like headings and lists when answering these questions when appropriate



GitHub Education

- When you register with your official student mail (...@smail.th-koeln.de), you are able to join the GitHub Education program.
- Premium access to premium features...
- <https://education.github.com/students>

The screenshot shows the GitHub Education homepage. At the top, there's a navigation bar with 'GitHub Education' on the left and links for 'Students', 'Teachers', 'Schools', 'Events', and a button 'Get benefits'. Below the navigation, a breadcrumb trail shows 'Home / Students'. The main headline reads 'With GitHub Education, your work will speak for itself.' followed by the subtext 'Build your portfolio, grow your network, and level up your skills.' A large red arrow points to the 'Join GitHub Education' button. Below this, there are two sections: one featuring a yellow backpack icon with the text 'Get the best developer tools with the GitHub Student Developer Pack', and another featuring a red flag with the GitHub logo and the text 'Grow your leadership skills as a GitHub Campus Expert'.

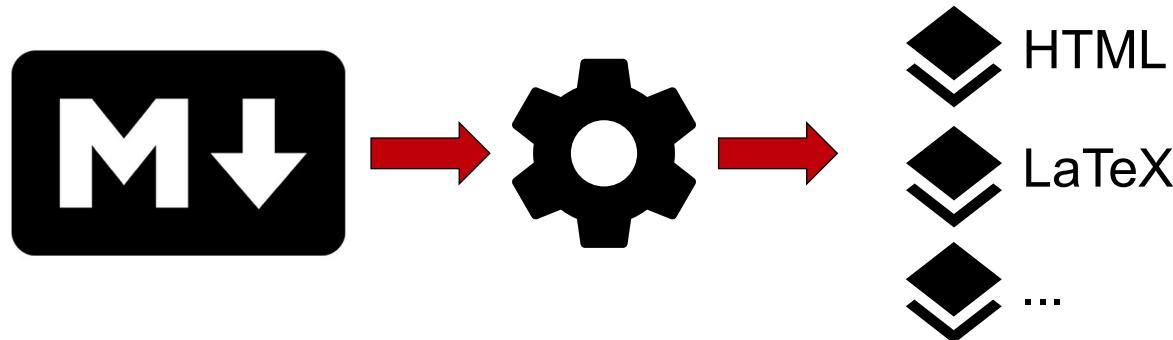
Markdown – In a nutshell

- Created by John Gruber and [Aaron Swartz](#)
- A simplified version of „markup languages“
- Allows to focus on writing opposed to formating
- Simple/intuitive formatting elements
- Easily converted to HTML or other formats
- Full specification: <https://daringfireball.net/projects/markdown/>

Try it out online: <http://pandoc.org/try/>

Other Markdown features

- Markdown files, which use the file extension .md, are machine readable, human readable, and used in many contexts - for example.
- **Markdown is plain text!** Just use a simple text editor like Notepad++ or Atom is sufficient.
- It is a great way to create machine-readable, easily searchable documents that can be repurposed in many ways
- GitHub renders text via Markdown... ☺



Markdown example

```
A First Level Header
```

```
=====
```

```
A Second Level Header
```

```
-----
```

```
Now is the time for all good men to come to  
the aid of their country. This is just a  
regular paragraph.
```

```
The quick brown fox jumped over the lazy  
dog's back.
```

```
### Header 3
```

```
> This is a blockquote.
```

```
>
```

```
> This is the second paragraph in the blockquote.
```

```
>
```

```
> ## This is an H2 in a blockquote
```

```
<h1>A First Level Header</h1>
```

```
<h2>A Second Level Header</h2>
```

```
<p>Now is the time for all good men to come to  
the aid of their country. This is just a  
regular paragraph.</p>
```

```
<p>The quick brown fox jumped over the lazy  
dog's back.</p>
```

```
<h3>Header 3</h3>
```

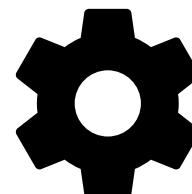
```
<blockquote>
```

```
    <p>This is a blockquote.</p>
```

```
    <p>This is the second paragraph in the blockquote.</p>
```

```
        <h2>This is an H2 in a blockquote</h2>
```

```
</blockquote>
```



Markdown Cheatsheet – Excerpt

- Many examples can be found in the GitHub documentation:
<https://help.github.com/en/articles/basic-writing-and-formatting-syntax>

Styling text

You can indicate emphasis with bold, italic, or strikethrough text.

Style	Syntax	Keyboard shortcut	Example	Output
Bold	<code>** **</code> or <code>__</code>	command/control + b	<code>**This is bold text**</code>	This is bold text
Italic	<code>* *</code> or <code>_ _</code>	command/control + i	<code>*This text is italicized*</code>	<i>This text is italicized</i>
Strikethrough	<code>~~ ~~</code>		<code>~~This was mistaken text~~</code>	This was mistaken text
Bold and italic	<code>** **</code> and <code>__</code>		<code>**This text is _extremely_ important**</code>	<i>This text is extremely important</i>

Summary and next week

For this semester:

- Learn to be OSEMN!
- Learn to be a Hacker!

For this week:

- Login into Moodle and join the course room.
- Get on GitHub.
- Join our Classroom.
- Learn to write Markdown.

Next week:

- Learn more about the hacker tool #1: The Shell