TECHNISCHE UNIVERSITÄT DRESDEN

Department of Computer Science Institute for Software and Multimedia Technology, Software Technology Group

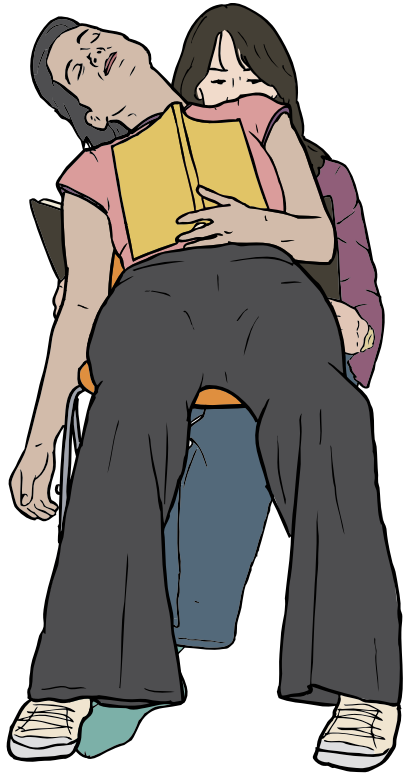
Academic Skills in Computer Science (ASiCS)

Literature Classification, Bulk Collection of Literature

Exercise
Thursday, 6. DS, APB/E001
Thomas Kühn (thomas.kuehn3@tu-dresden.de)

Scientific Research
Daily Business

Reading



Writing

Organizing

Images from OpenClipart.org (Creative Commons by Steve Lambert)

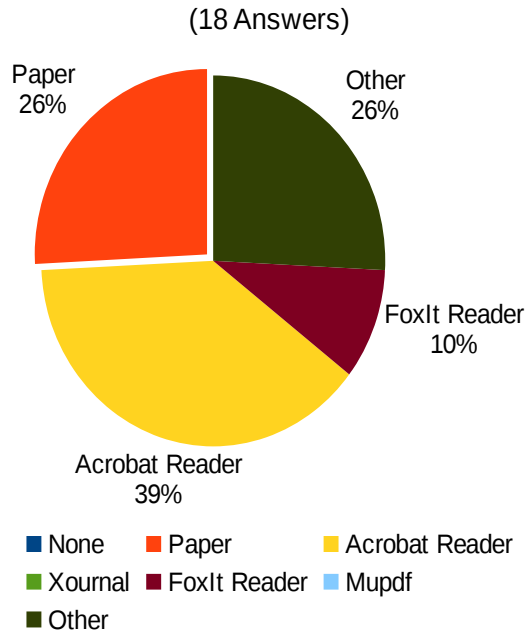
Common Tasks

- Find relevant / related publications
  - Query scientific search engines
  - Look up *BibTex* for specific publications from the web
- Investigate found publications
  - Skim papers
  - Make notes and hints
  - Organize downloaded files
  - Maintain a corresponding **bibliography** of *BibTex* entries

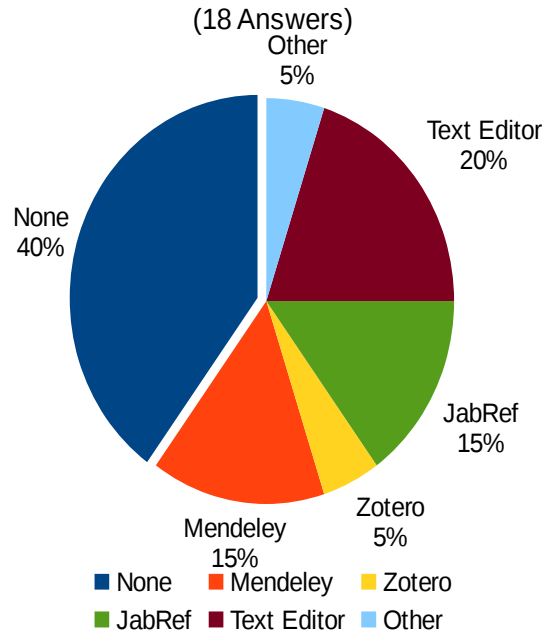
A Small Survey

- Q1: What tools do you use to read and annotate papers?
- Q2: *What tools do you use to organize your bibliography?*
- Q3: *What tools do you use to organize stored papers?*

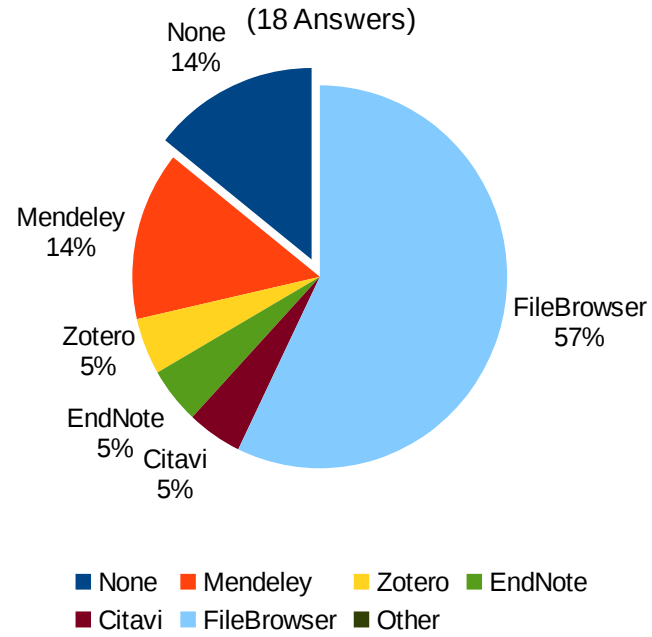
What tools do you use to read and annotate papers?



What tools do you use to organize your bibliography?

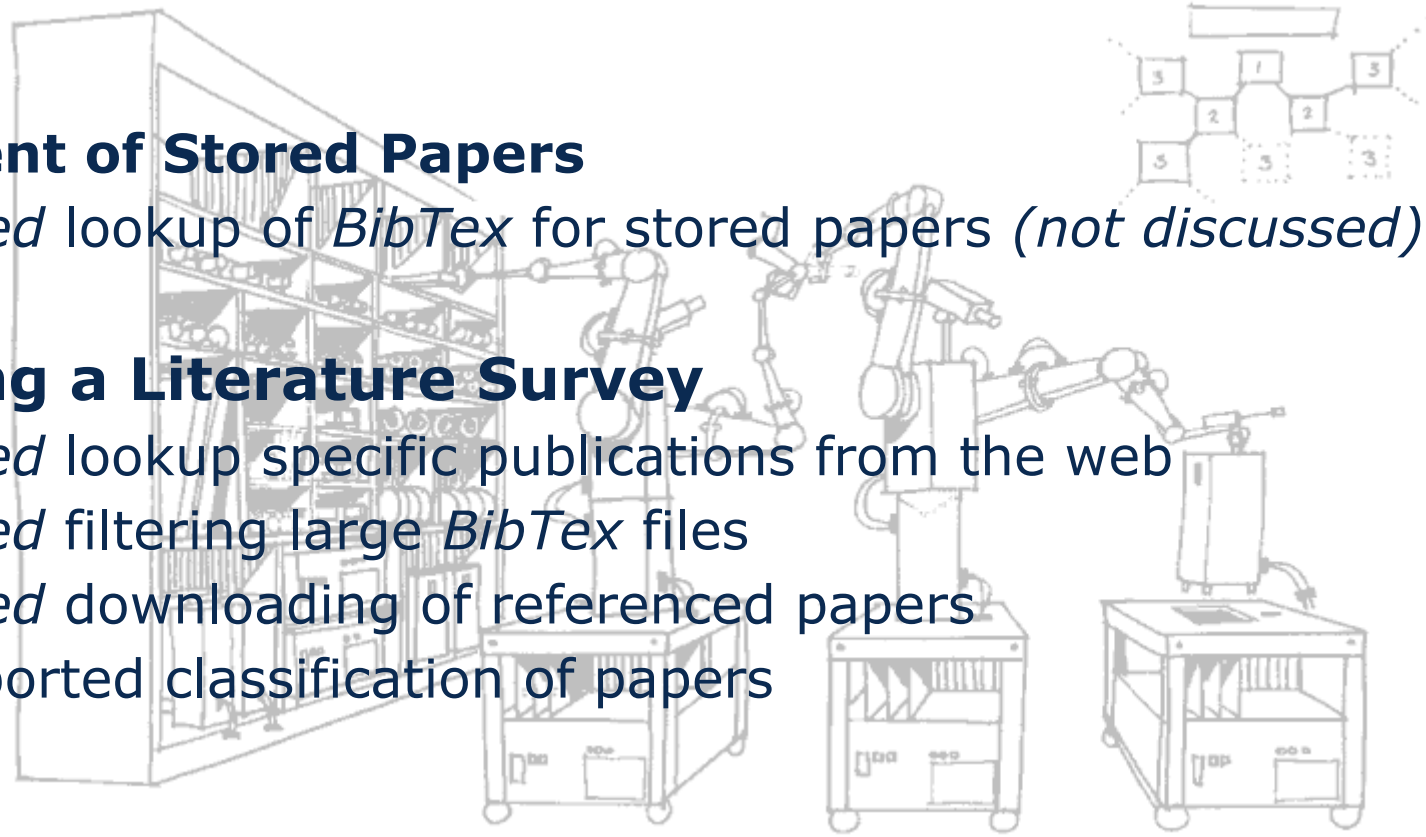


What tools do you use to organize stored papers?



Common Tasks

- Management of stored papers
  - Search text fragments in papers
  - Look up *BibTex* for stored papers
- Conducting a literature survey
  - Look up *BibTex* for specific publications from the web
  - Filtering large *BibTex* files
  - Downloading papers
  - Classifying found papers



Management of Stored Papers

- *Automated* lookup of *BibTex* for stored papers *(not discussed)*

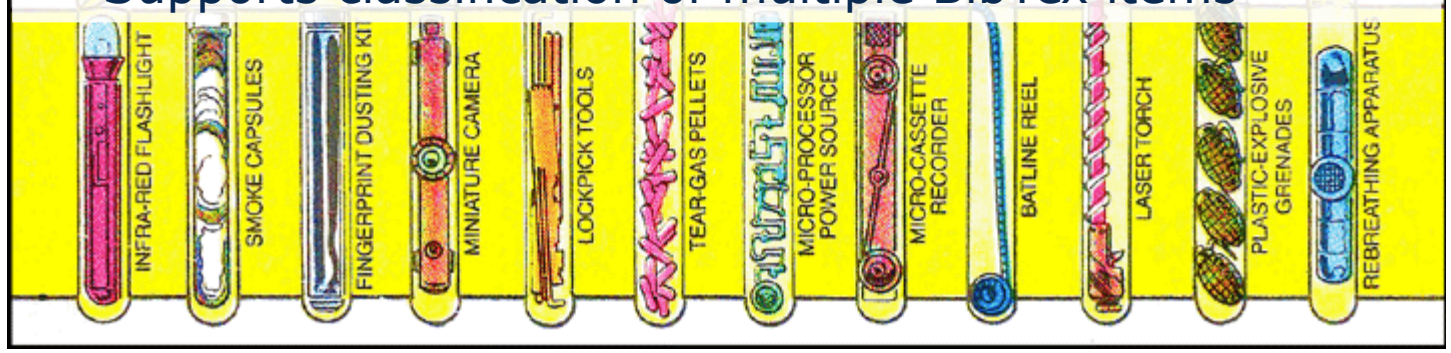
Conducting a Literature Survey

- *Automated* lookup specific publications from the web
- *Automated* filtering large *BibTex* files
- *Automated* downloading of referenced papers
- Tool supported classification of papers

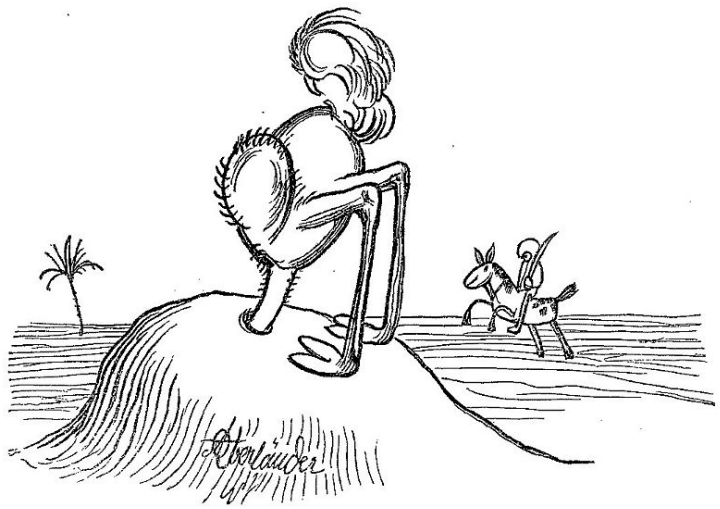
Picture by Nasa (public domain)

Organizing Research
What we Need?

- ***getbibtex.rb***†
  Fetches bibtex entries for stored papers
- ***gsresearch.rb***†
  Collects bibtex entries from Google Scholar
- ***bibfilter.rb***
  Filters large BibTex files by various criteria
- ***gsdownload.rb***†
  Downloads all files referenced by a BibTex files
- ***SLR-toolkit***
  Supports classification of multiple BibTex items



Organizing Research
† Disclaimer



Adolf Oberländer (public domain)

- Never use these scripts in jurisdictions, which prohibit automated use of Google Scholar
  - See Google's terms of Use
- Do not use these scripts to attack google services
- These tools are only for research purpose
- *„I would pay for using a Google Scholar API“*

Automated Managment

- Find naming schema for stored publication

  <Full Name of First Author>_<Full Title>.pdf
  (e.g.: *Charles W Bachman_Data Structure Diagrams.pdf*)
- Keep all documents in one folder (e.g.: *library/*)
- Use author's last name for subfolder (e.g.: *library/Bachman/*)

Steps

1. Automated sorting of new files into subfolders

   ```
   $ ./mvtodir.sh
   ```
2. Generating the file list for **getbibtex**

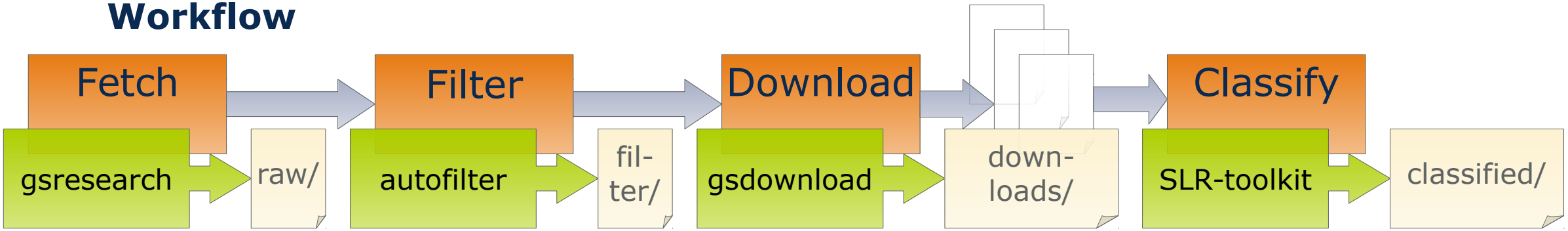
   ```
   $ ./gettitles.sh > titles.txt
   ```
3. Initializing / Updating the bibliography

   ```
   $ ruby getbibtex.rb titles.txt my.bib 1>> my.bib
   ```

Task

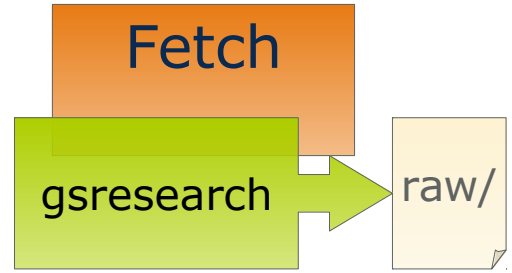
- Fetch all publications matching a query string
  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*
- Sort out irrelevant publications
- Download PDF files for all relevant publications
- Collect statistics about survey process

Workflow



Automatic Querying

- Defining a search query
  - Exact, With, Any, and Without
  - Time span (from year to year)
- Directly supported by **gsresearch**



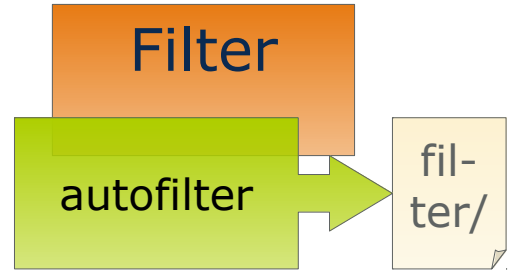
Steps

1. Test your query with Google Scholar[1)] (advanced search)
2. Change the **gsresearch.sh** accordingly
3. Run the script with
   ```
   $ ./gsresearch.sh
   ```
4. Be patient, very patient

1) https://scholar.google.com

Automatic Filtering

- Further filter the inital dataset
- Using **bibfilter** to select items by
  - document class, publisher, citation count, …
- Two automatic filtering steps in **autofilter**
  - Select items by publisher
    ACM, IEEE, Springer, ScienceDirect
  - Filter items with low impact ← **DEEMED EVIL**
    Citation Count < Log(Age)



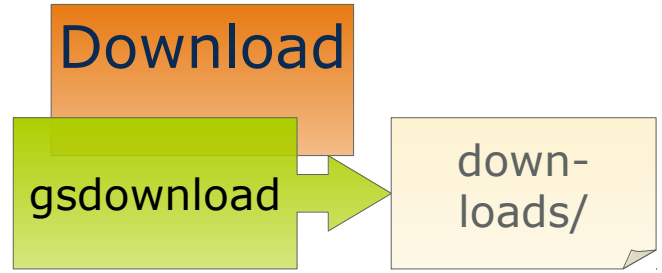
Human Filtering

- Check the title of the paper and (abstract, content)

```
$ mkdir filter_human
$ for f in `ls filter_rel/`; do
      ruby bibfilter.rb 'filter_rel/$f' > 'filter_human/$f' ;
  done
```

Automatic Download

- Download final set of relevant
- Access files via the publisher's site
- Support for the big four:
  ACM, IEEE, Springer, ScienceDirect
- Extensible towards other publishers
- Downloaded files are referenced within bibtex items



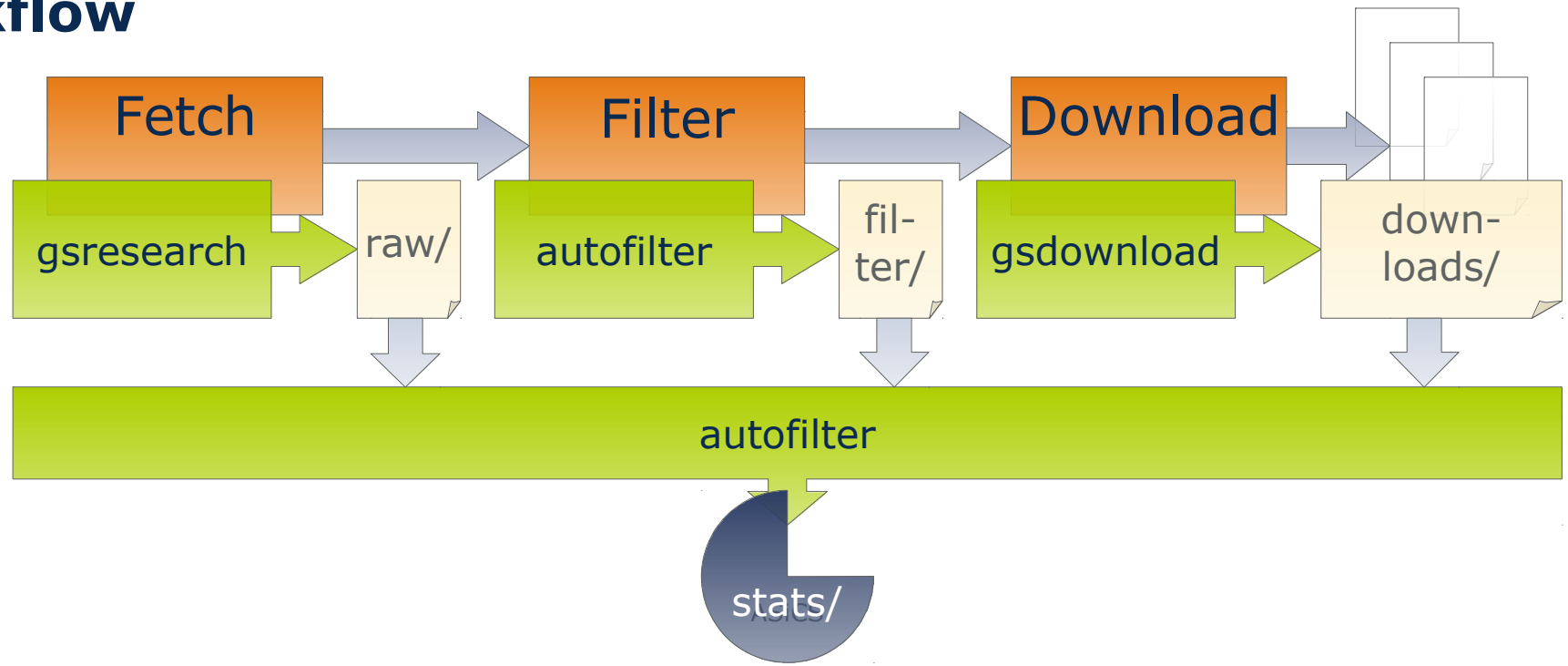
Steps

1. Run the script with
   ```
   $ ./gsdownload.sh
   ```
2. Be patient
3. Rerun
   ```
   $ ./autofilter.sh
   ```

Collecting Statistics

- Crucial to explain selection method of survey
- Generated automatically by **autofilter**
- Stored as csv files in *stats_*/* folder

Workflow

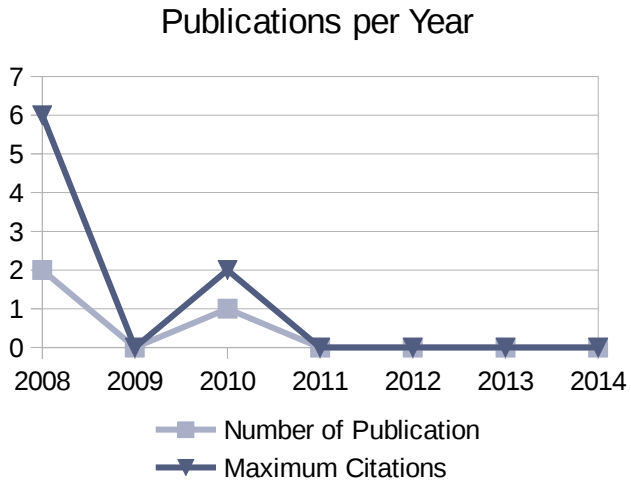


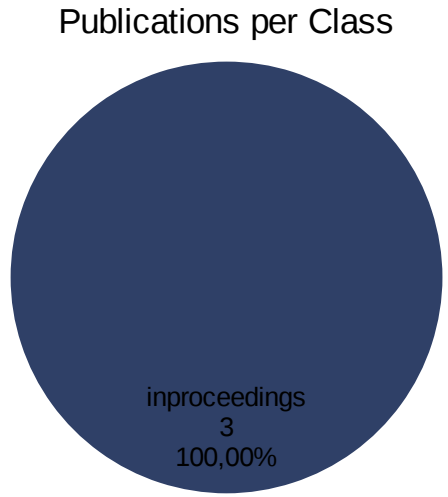
Example

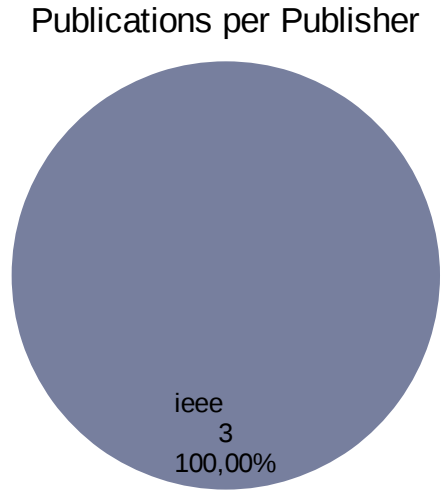
- Query for publications from 2008 to 2014

  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*

- Inital dataset: 9 entries
- Automatic Filter: 4 entries
- Human Filter: 3 entries
- Download: 3 pdf files

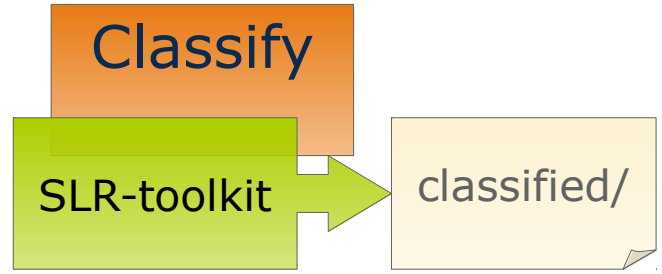






Literature Classification

- Reuse classification scheme of previous *surveys*
- Classification by qualitative and/or quantitative criteria
- Retrieve *goals*, *requirements*, *techniques* from related approach
- Favor orthogonal dimension to group common qualities

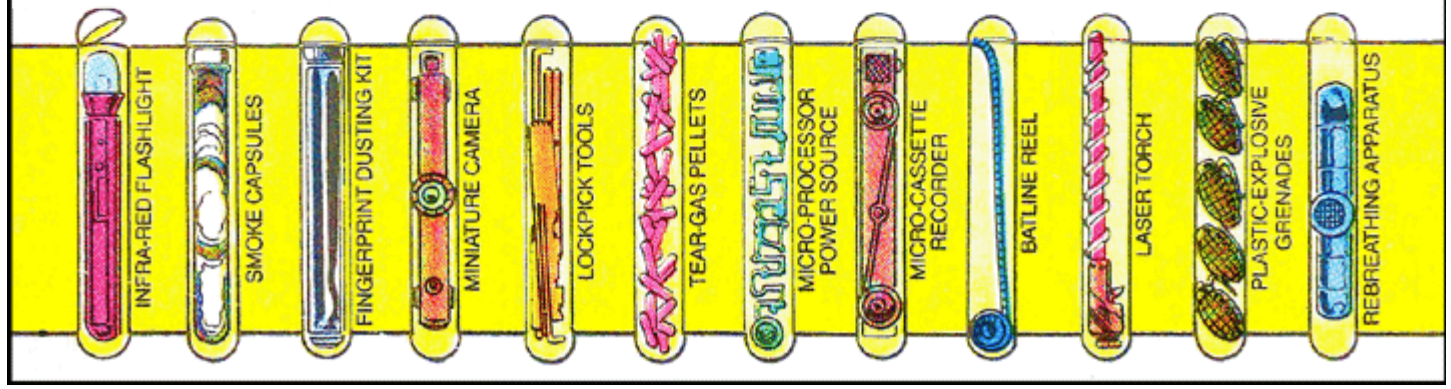


SLR-Toolkit

- Tool Support for Classifying Papers
- Supports arbitrary hierarchical classification schemes
- Classification of papers per *BibTex* annotation
- Synchronization with *Mendeley*

GitHub

- ***bibfilter*** (https://github.com/Eden-06/bibfilter)
  contains the *bibfilter.rb* script as independent tool
- ***gsresearch*** (https://github.com/Eden-06/gsresearch)
  contains the various Ruby scripts
  - *getbibtex.rb,*
  - *gsresearch.rb, and*
  - *gsdownload.rb*
- ***SLR-toolkit*** (https://github.com/sebastiangoetz/slr-toolkit)

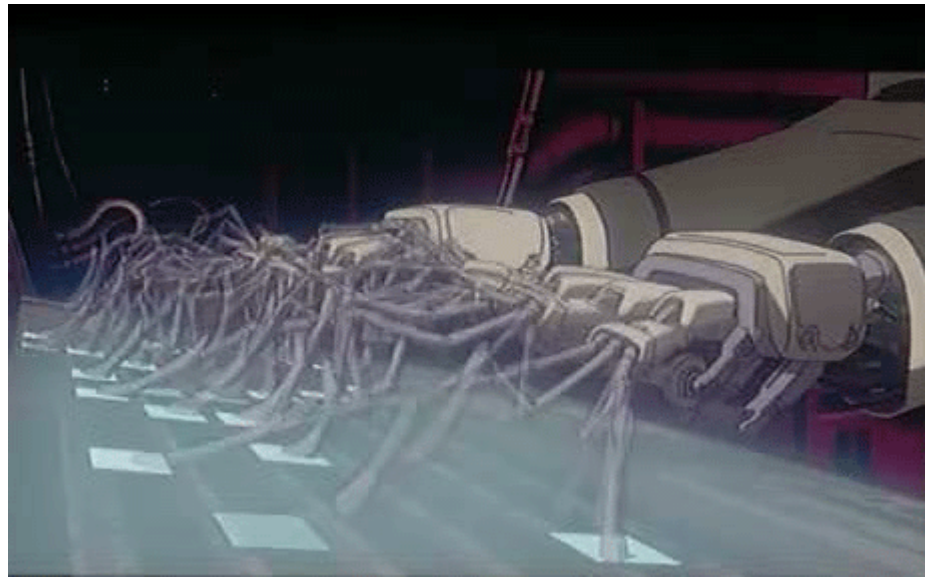


Automated Tasks

- Automated BibTex lookup for stored papers
- Automated BibTex lookup for specific Publications from the web
- Automated filtering of large BibTex files
- Automated download of papers referenced by a BibTex file
- Semi-automatic literature survey

Exercise

1) Create a **Classification Scheme** for your related work.
2) Classify at least 5 papers wrt. this **Classification Scheme**
3) Create and add a comparison table to the *Related Work* section.

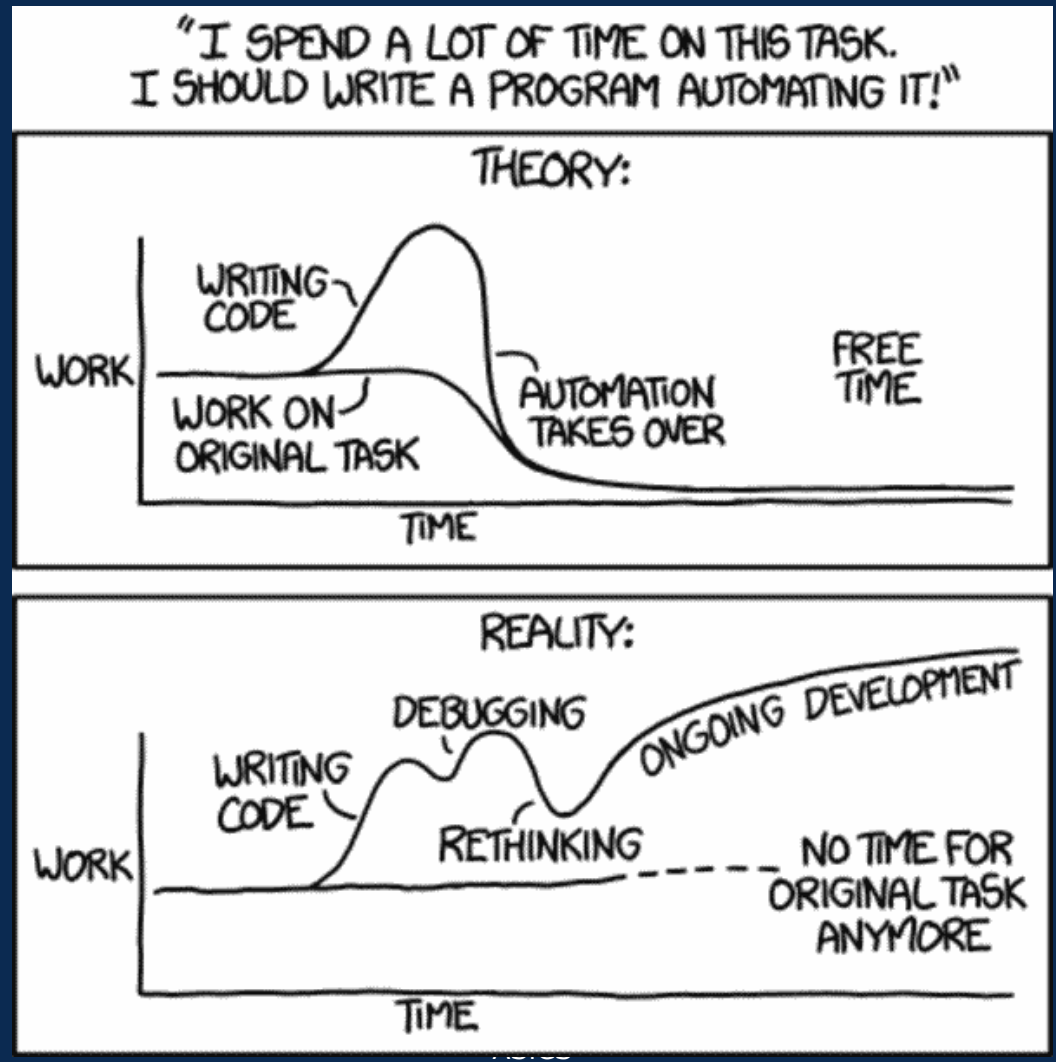


"Ghost in the Shell" by Production I.G ALL RIGHTS RESERVED

Now Automated Writing

- Overview on Paper generators
  - SCIgen[4)]
  - Mathgen[5)]
  - ...
- Automating idea generation
  - Random topic generator
- Predefined Structure

3) http://pdos.csail.mit.edu/scigen/
4) http://thatsmathematics.com/mathgen/



Automation (By Randell Munroe from xkcd.com)