TECHNISCHE UNIVERSITÄT DRESDEN

Faculty for Computer Science Department of Software and Multimedia technology, Software Technology Group

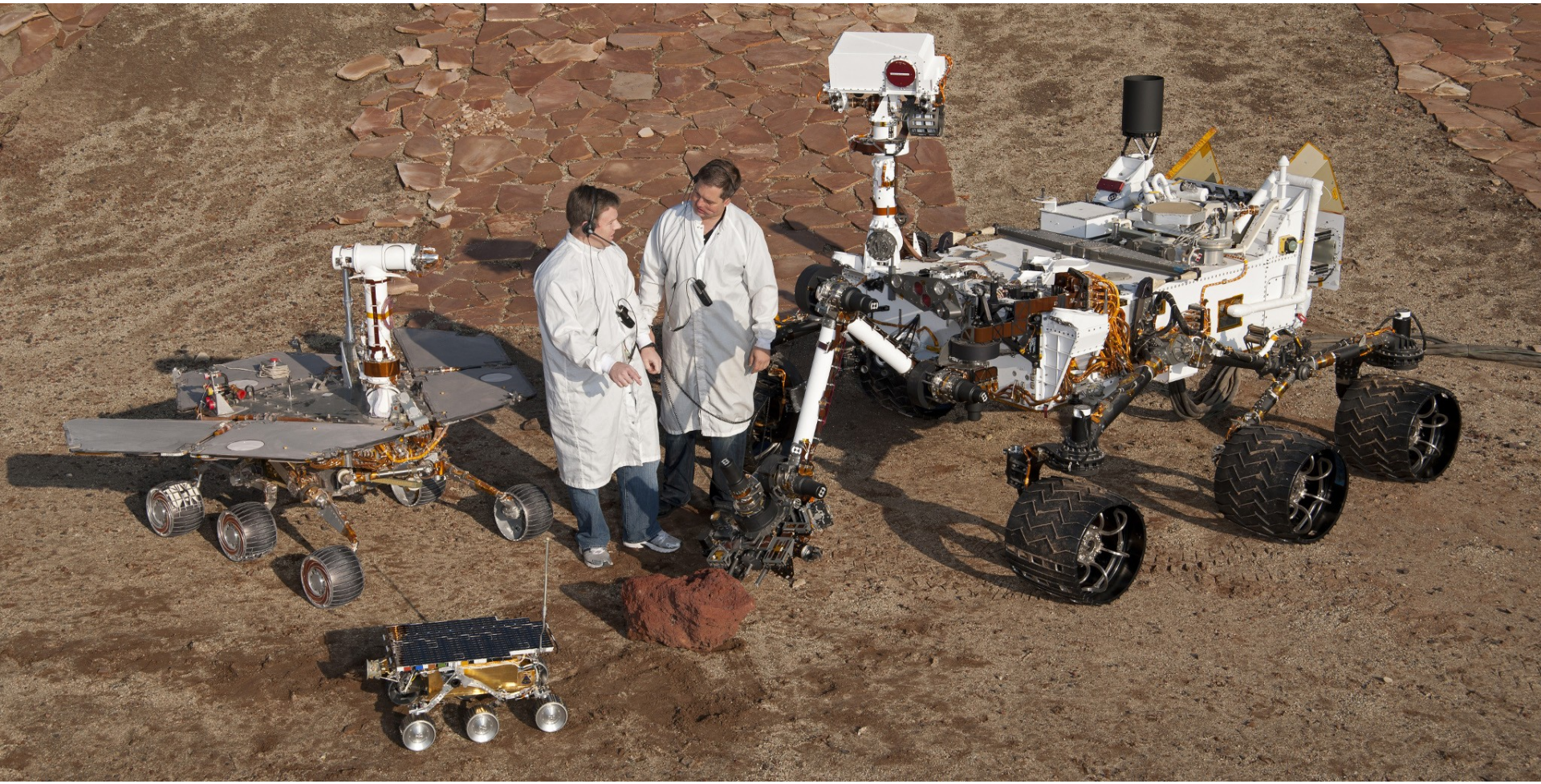
Efficient Scientific Research
with Scripts

Speaker
Thomas Kühn

Output
02.07.2015

DRESDEN concept
Exzellenz aus Wissenschaft und Kultur

Scientific Research



Picture by Nasa (public domain)

Scientific Research
Daily Business

Reading | **Writing** | **Organizing**

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

Scientific Research
Daily Business

Reading **Writing** **Organizing**

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

Scientific Research
Daily Business

Reading

Writing

Panruby[1)]

- Focus on contant rather than ideoms
- Concise ways to structure text
- Direct support for citations, figures, tables
- Transformation to arbitrary formats
- Template engine for (Multi-)markdown
- **One content source many output formats**

1) https://github.com/Eden-06/panruby

Organizing

Organizing

Organizing

Common Tasks

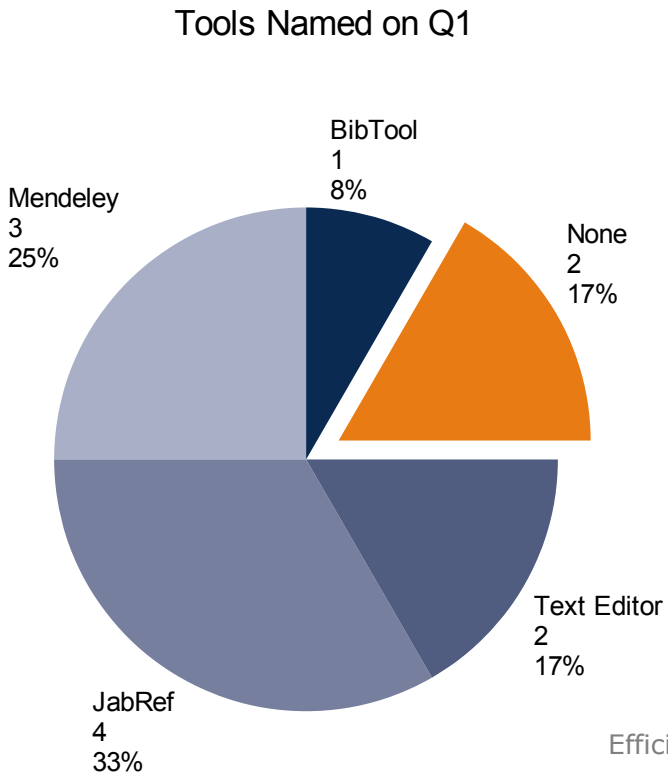
- Managment of stored papers
  - Search text fragments in stored files
  - Look up *BibTex* for stored papers (pdfs)
- Conducting a literature survey
  - Look up *BibTex* for specific Publications from the web
  - Filtering large *BibTex* files
  - Downloading papers you previously referenced

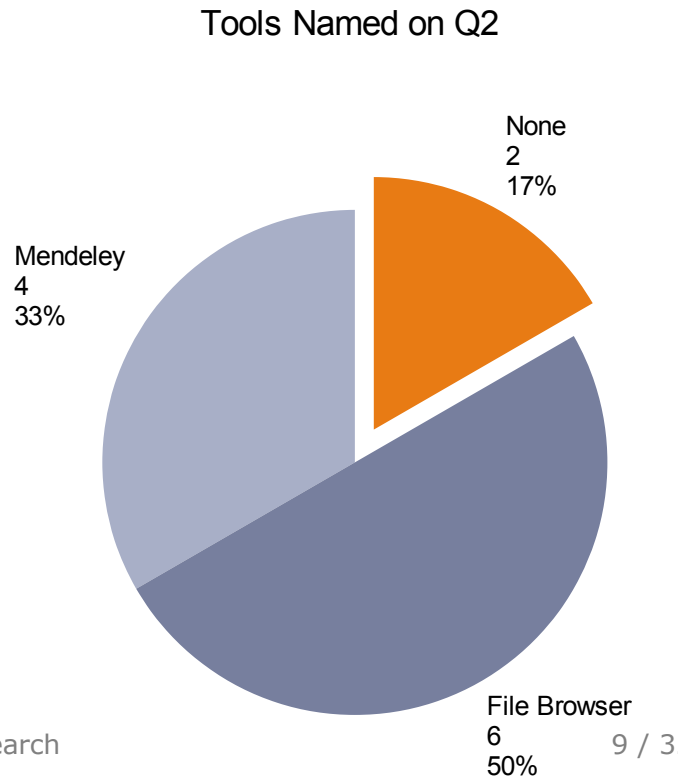
A Small Survey

- Q1: *What tools do you use to organize your bibliography?*
- Q2: *What tools do you use to organize stored papers?*
- 9 Answers named 5 different Tools

A Small Survey

- Q1: *What tools do you use to organize your bibliography?*
- Q2: *What tools do you use to organize stored papers?*
- 9 Answers named 5 different Tools





Survey Results

- Basical only 4 Tools in Use
- Only few participants use special tools *(i.e. Mendeley, JabRef)*
- Most rely on the *File Browser* to manage papers

File Browser

- (Some) Support for search text fragments in stored pdf files

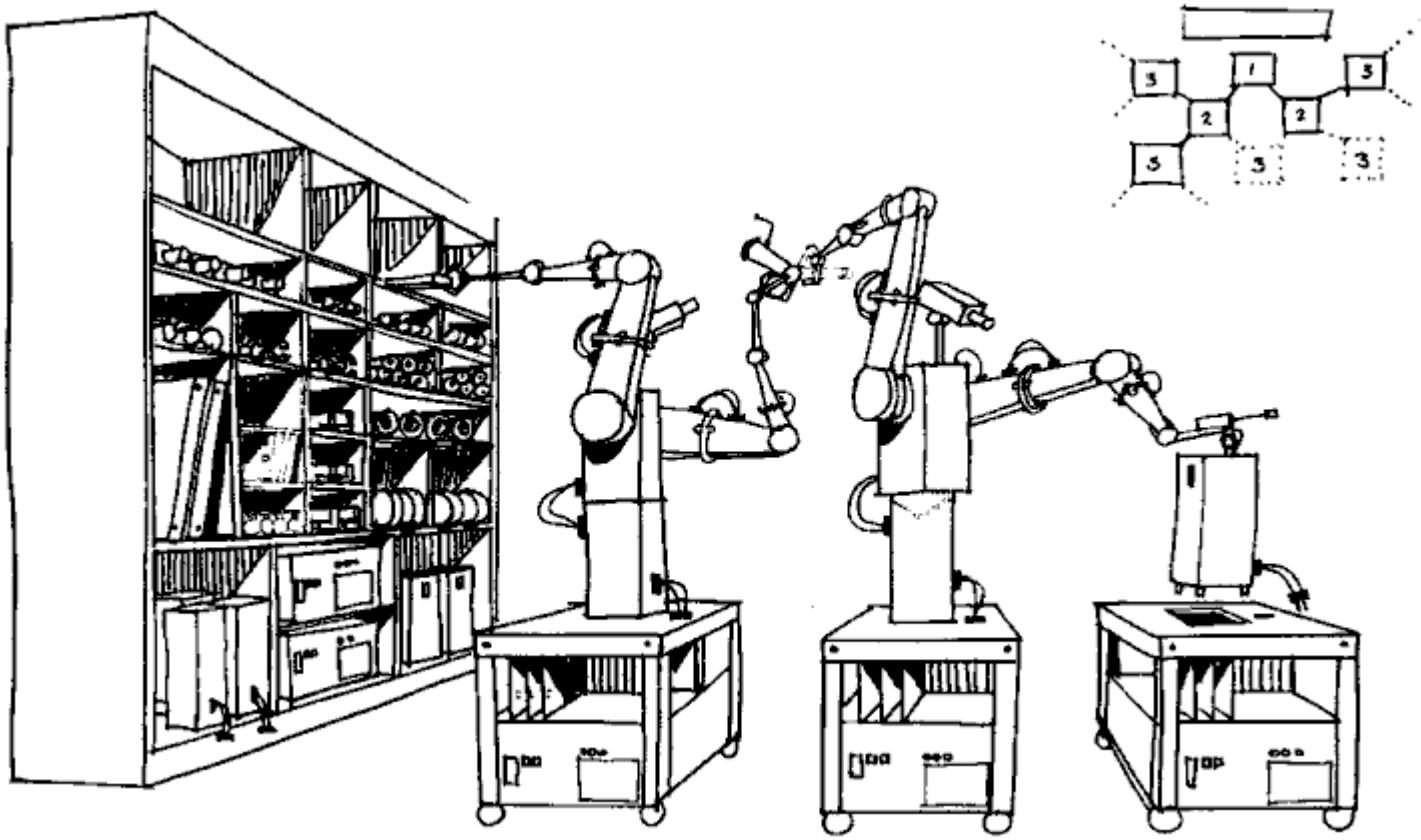
Mendeley[2)]

- Manages Papers and Reference
- Fully-searchable library
- Fetches BibTex entries for stored papers **automatically**

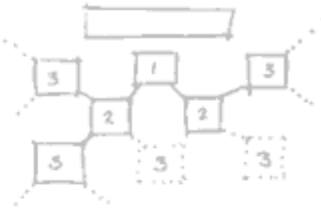
Which other tasks can be automated?

2) http://www.mendeley.com/

Scientific Research
Automating the Organization Process



Picture by Nasa (public domain)



Management of Stored Papers

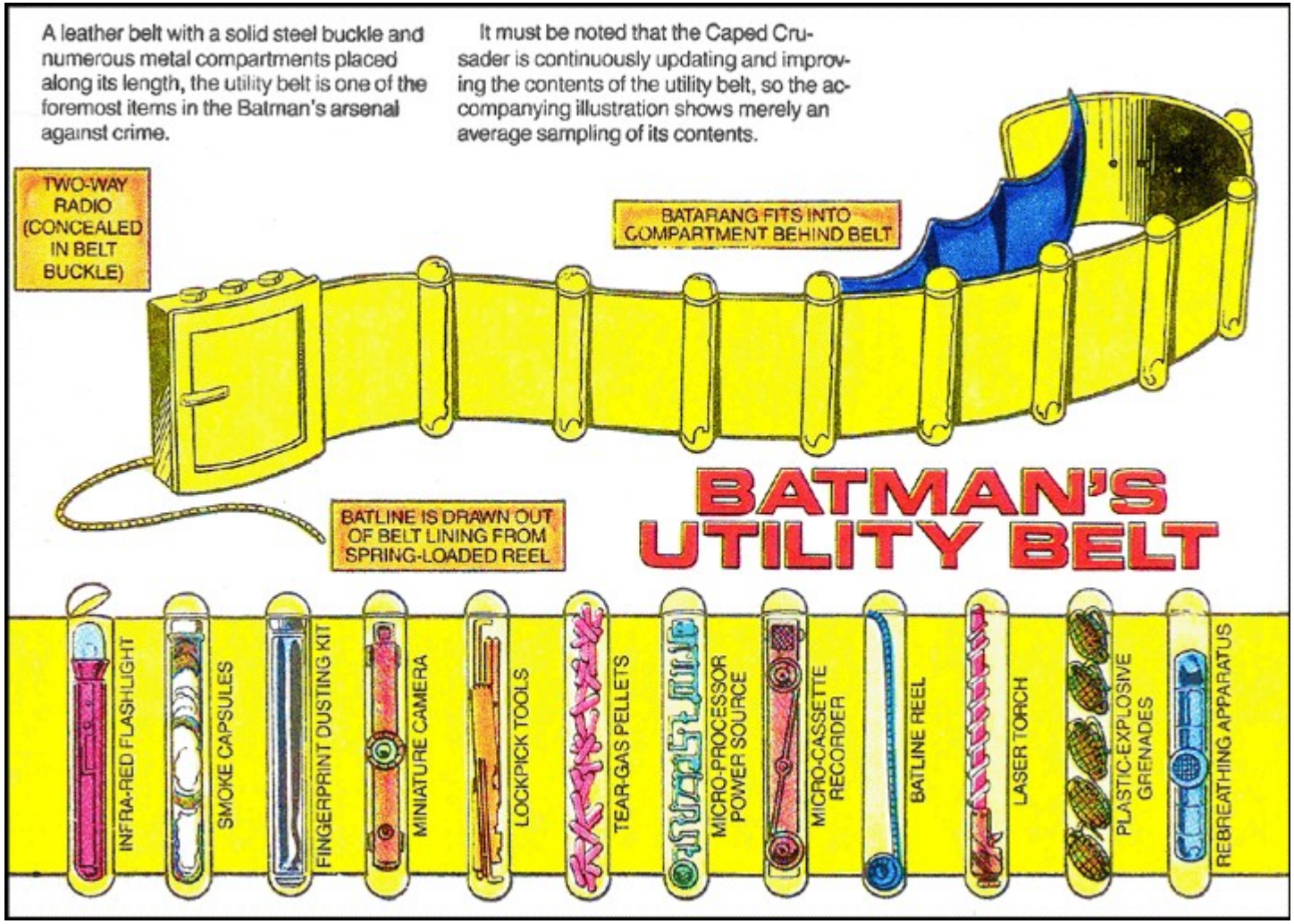
- *Automated* lookup of *BibTex* for stored papers

Conducting a Literature Survey

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

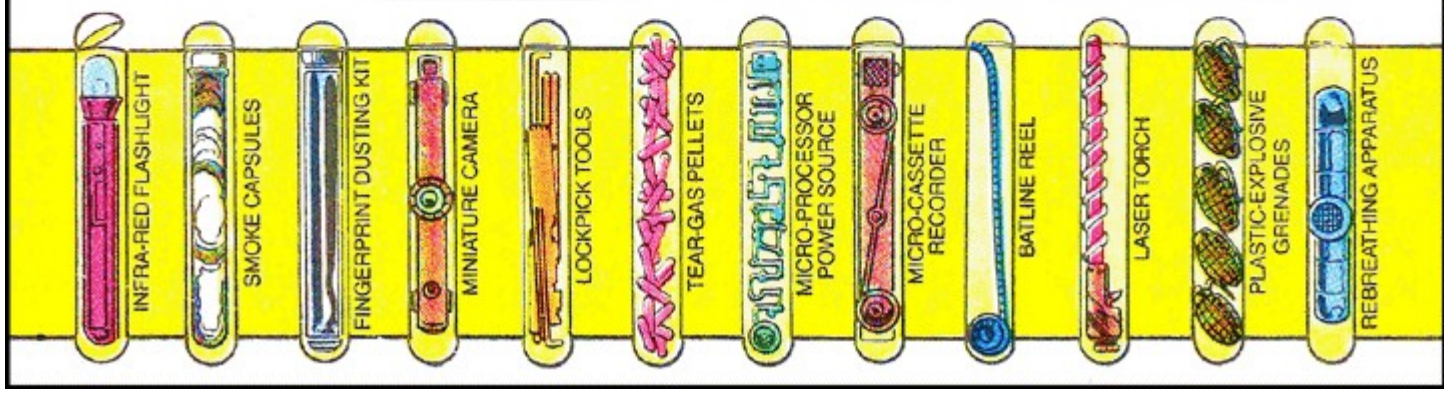
Picture by Nasa (public domain)

Scientific Research
What We Need?

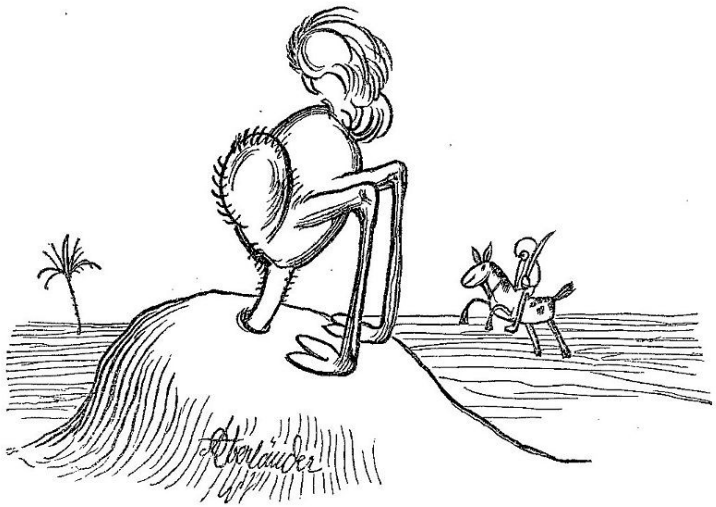


Scientific Research
What We Need?

- ***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



Scientific 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

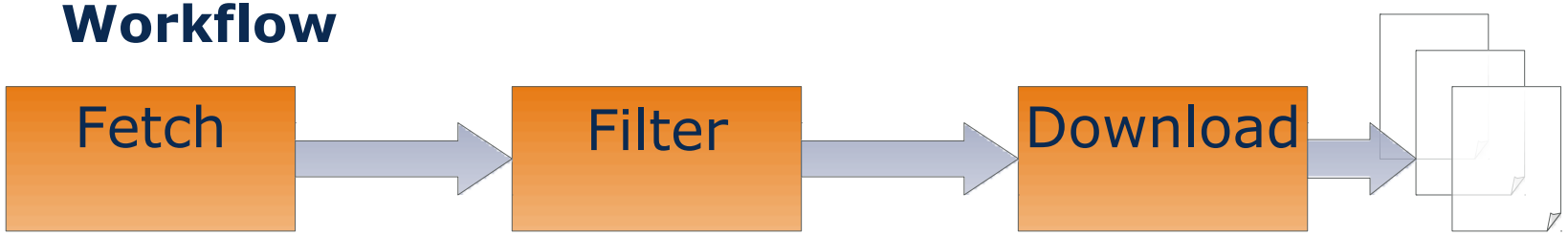
   ```
   $ ruby getbibtex.rb titles.txt my.bib 1>> my.bib
   ```

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

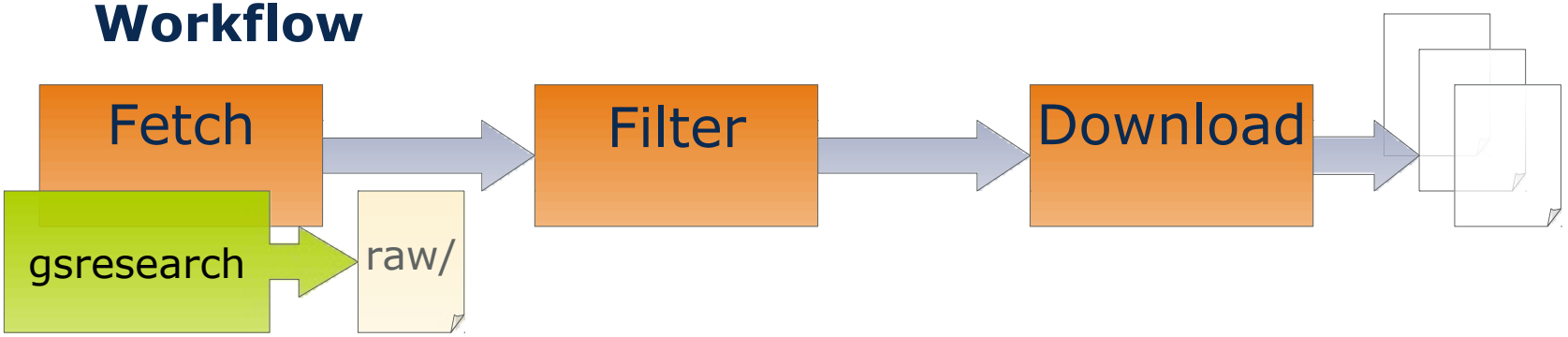


Task

- 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

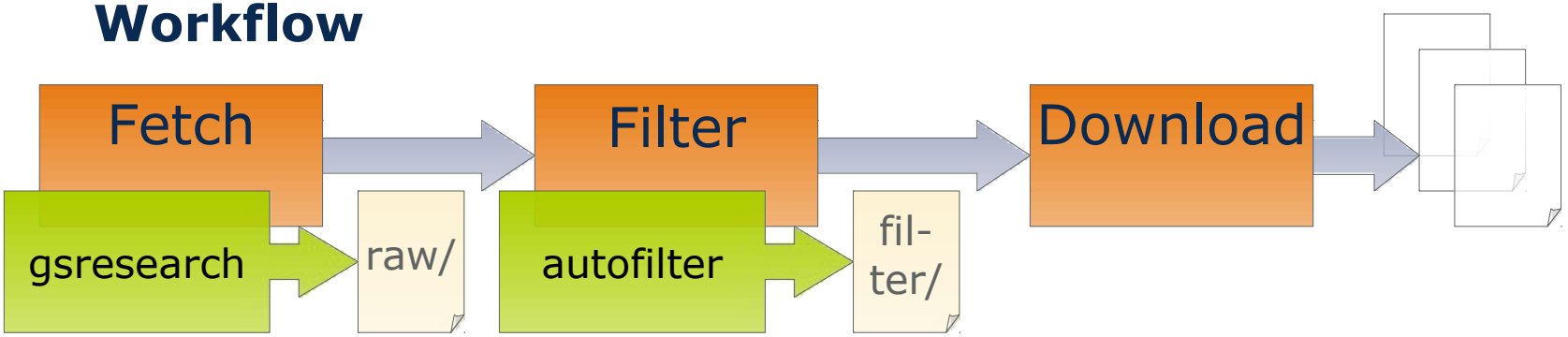


Task

- 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

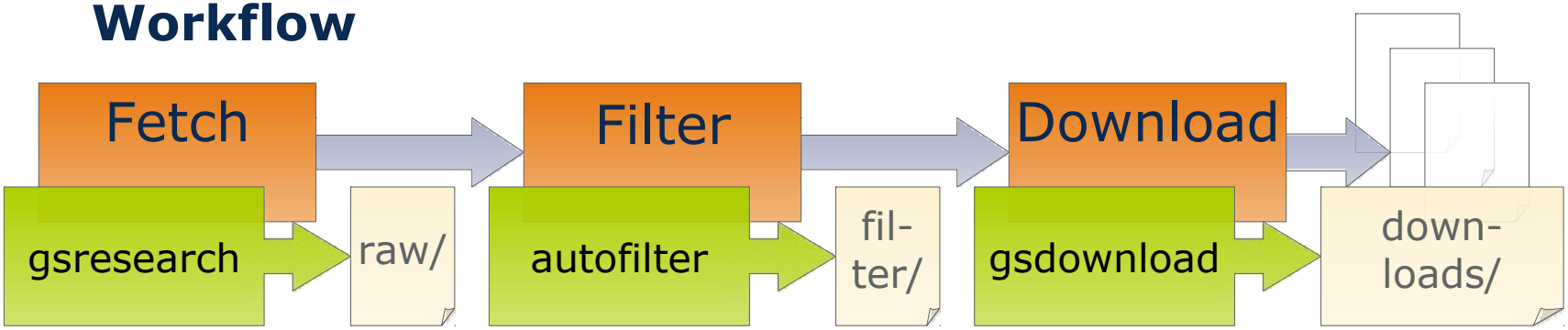


Task

- Fetch all publications matching a query string

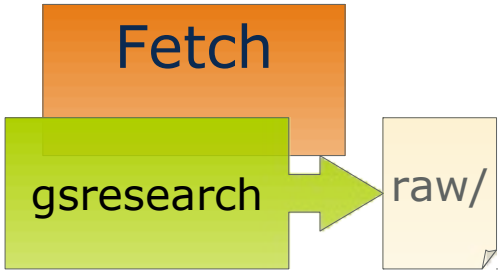
  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*
- Sort out irrelevant publications
- Download PDF files for all relevant publications
- Collect statistics about survey process

Workflow



Automatic Querying

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

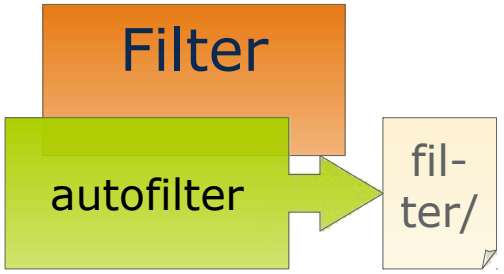


Steps

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

3) https://scholar.google.com

Automatic Filtering



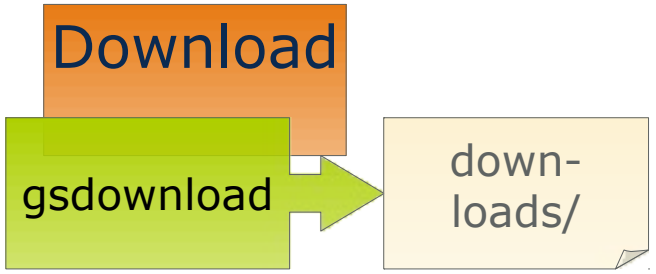
- Further filter the inital dataset
- Using **bibfilter** to select items by
  - document class, publisher, citation count, …
- Two automatic filtering steps in **autofilter**
  1. Select items by publisher
     ACM, IEEE, Springer, ScienceDirect
  2. Filter items with low impact
     Citation Count < 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

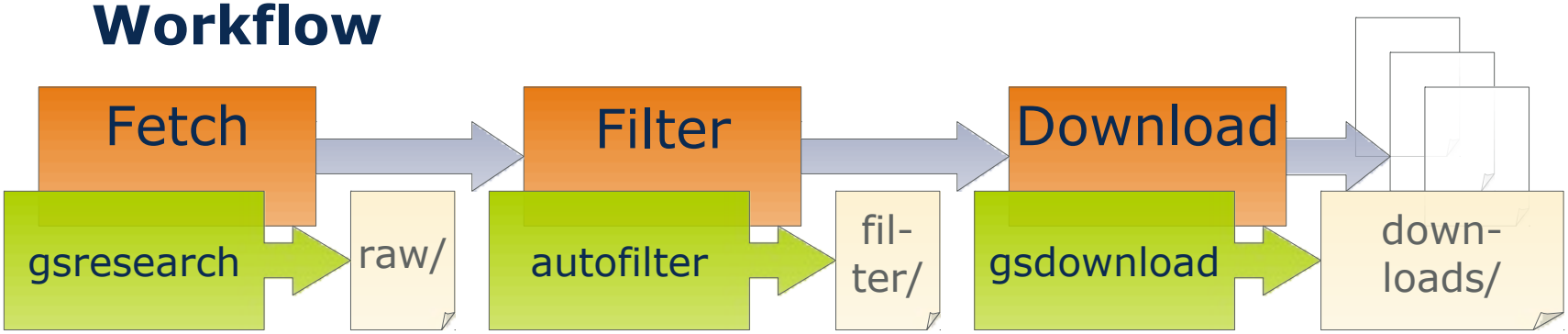
Steps

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

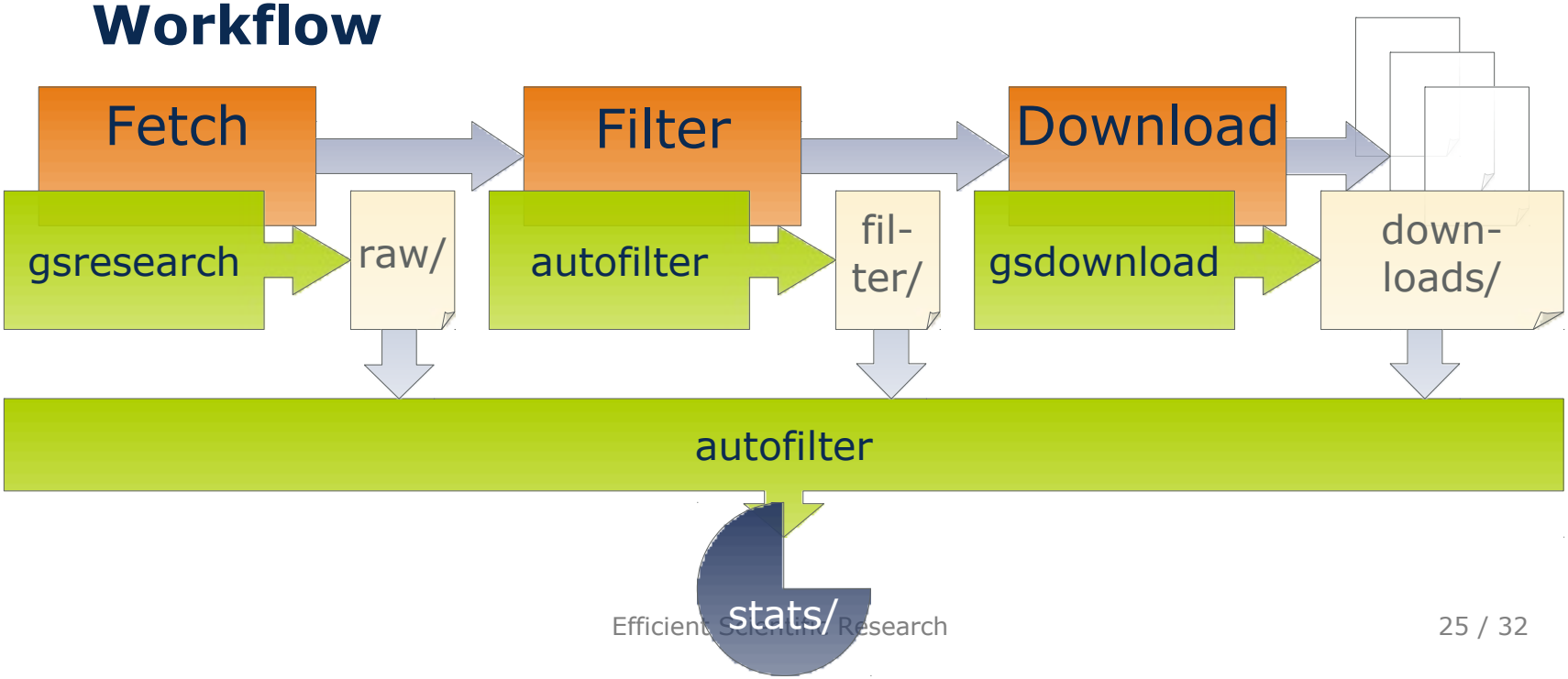
Workflow



Collecting Statistics

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

Workflow

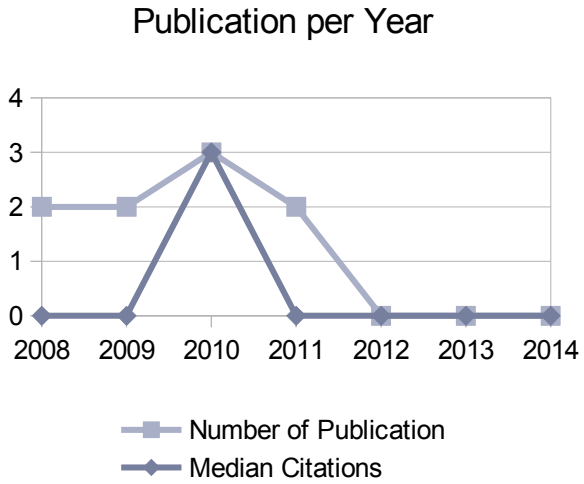


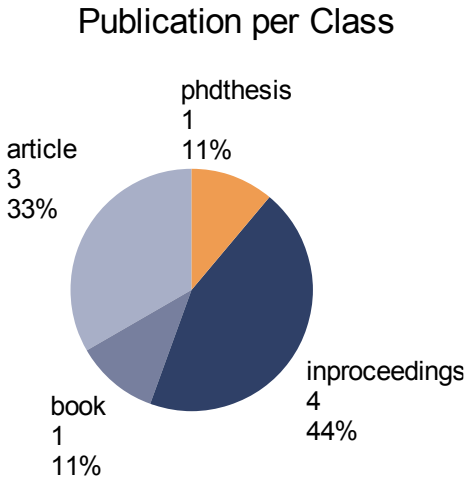
Example

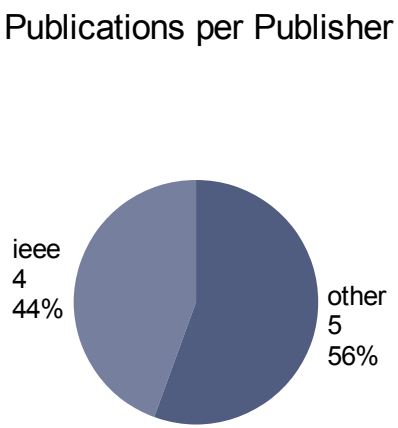
- Query for publications from 2008 to 2014

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

- Inital dataset: 9 entries







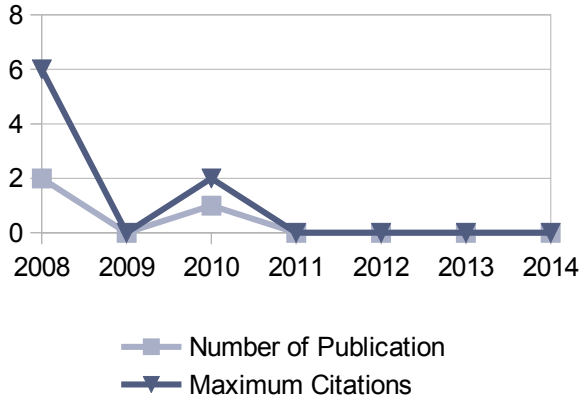
Example

- Query for publications from 2008 to 2014

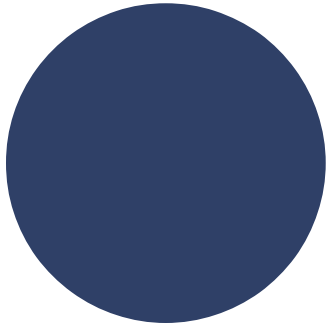
  ***With**: ospp, workflow*
  ***Exact**: sebastian richly*

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

Publications per Year

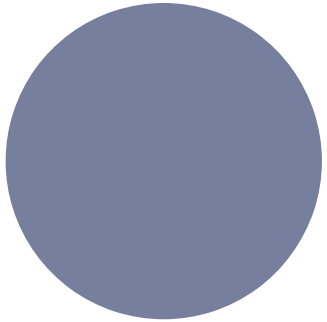


Publications per Class



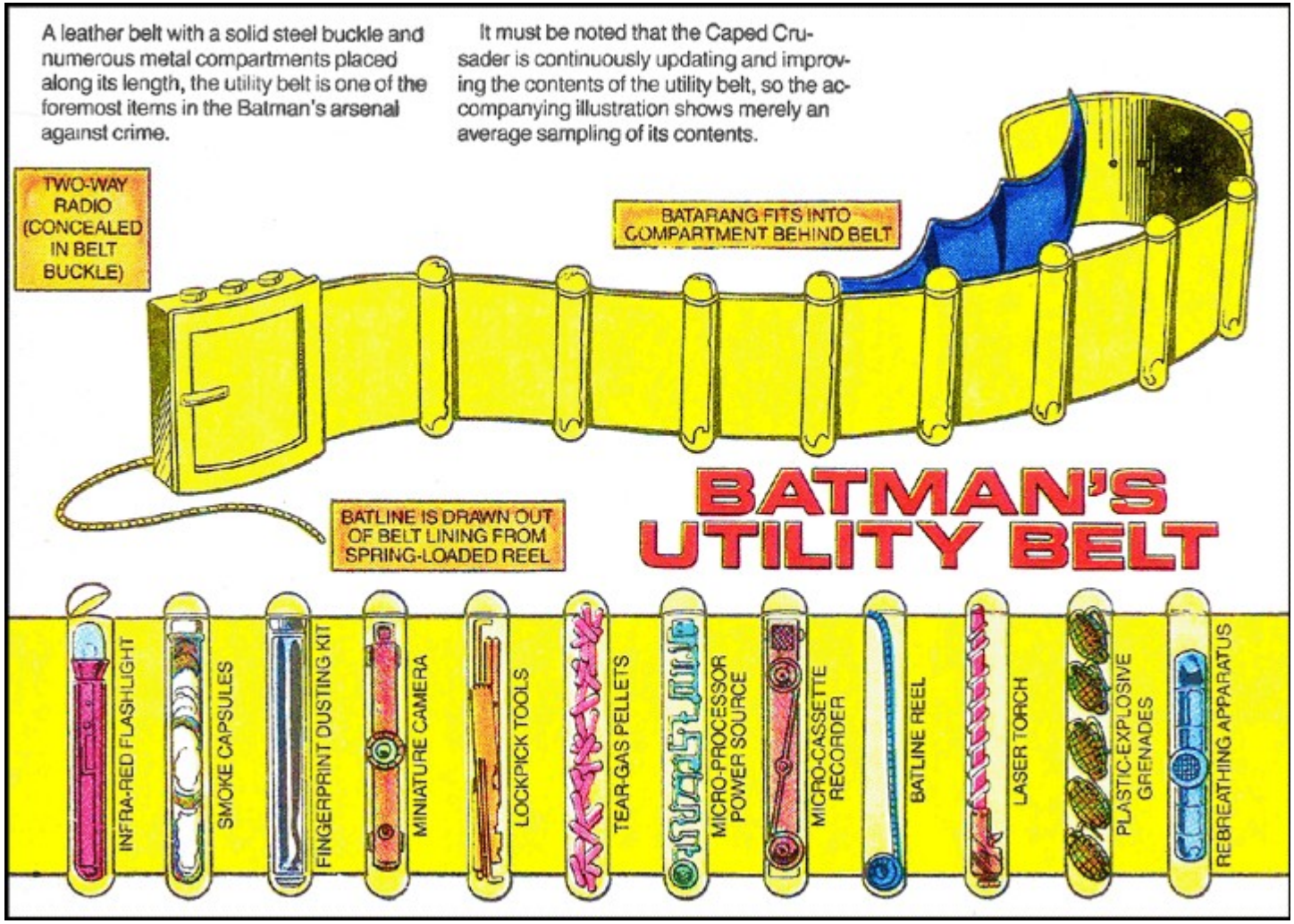
inproceedings
3
100%

Publications per Publisher



ieee
3
100%

How to Get the Utility Belt?

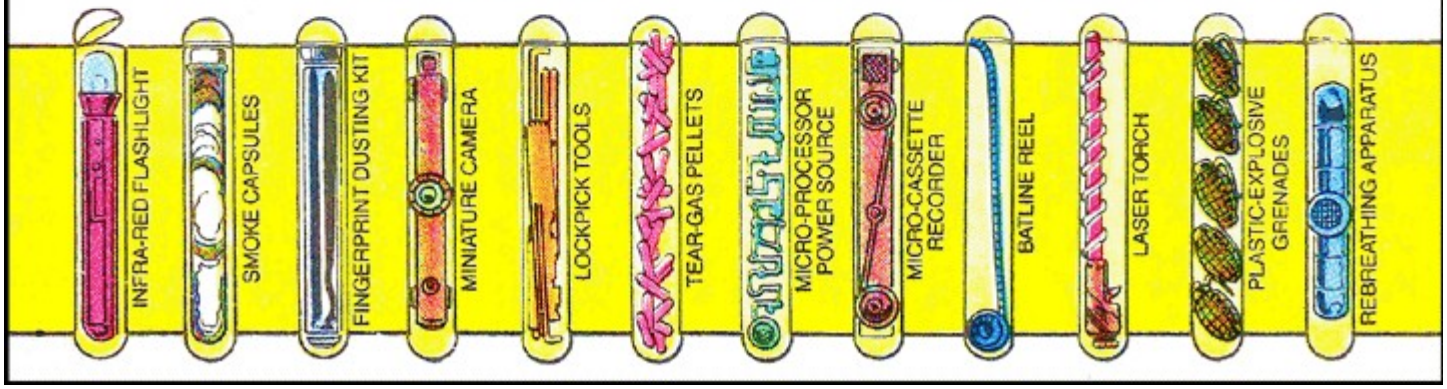


Scientific Research
How to Get the Utility Belt?

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*

Additionally, all the presented shell scripts

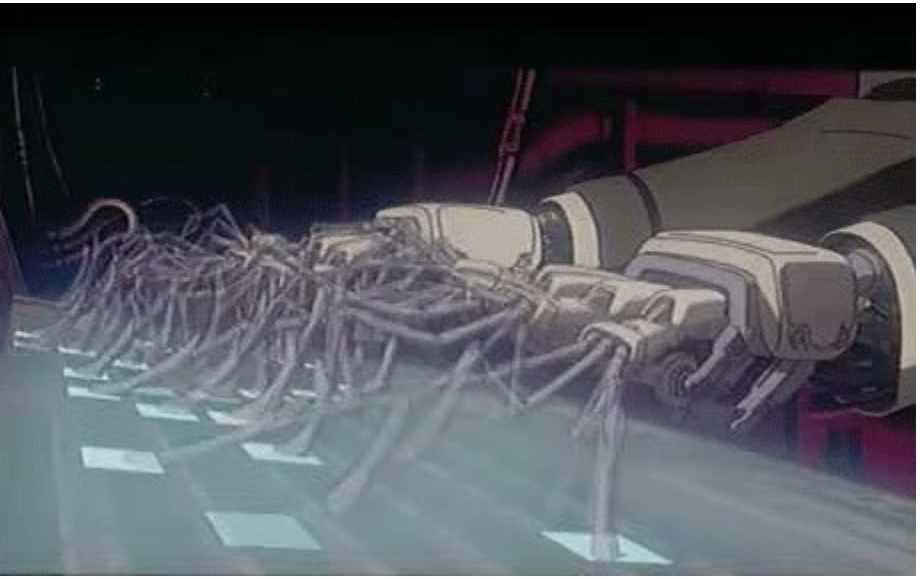


Organizing

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

Writing Papers

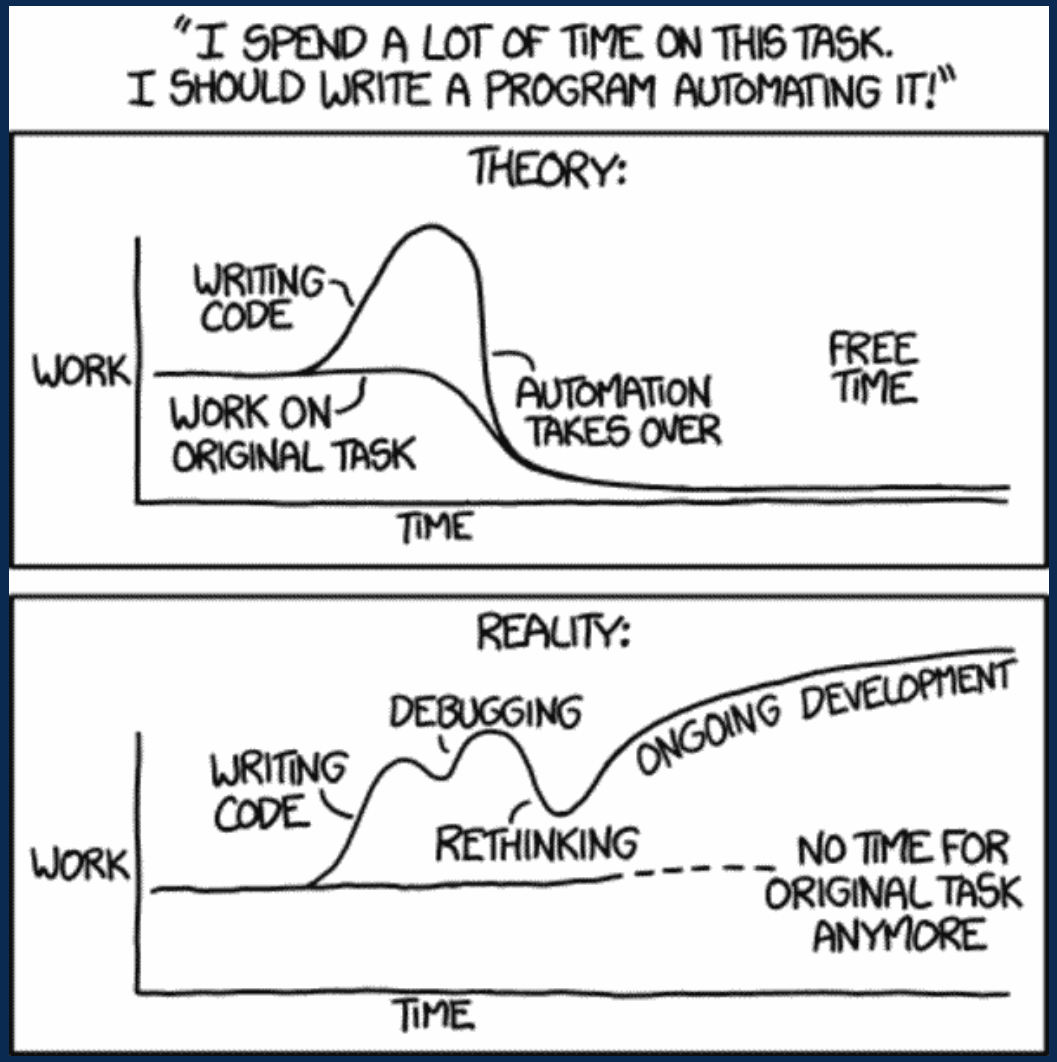


"Ghost in the Shell" by Production I.G ALL RIGHTS RESERVED

Now Automated

- 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)