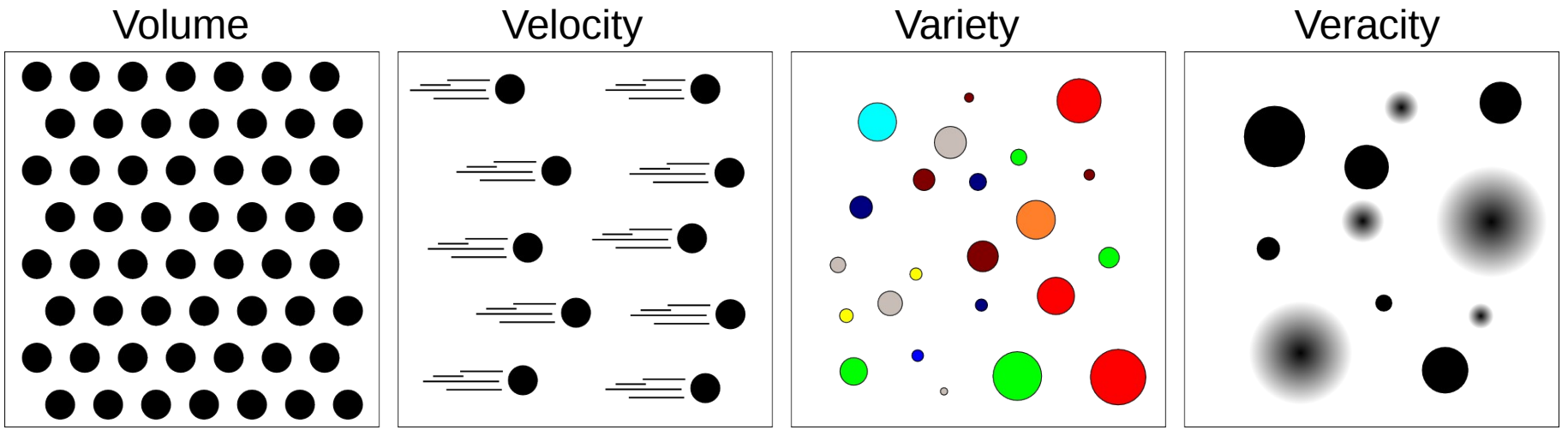
Herausforderungen von Big Data für die Bibliotheks- und Informationswissenschaft

Prof. Dr. Robert Jäschke
Leibniz Universität Hannover

Stand der Vorlesung

Kapitel 10: Big-Data-Technologien

- Einführung
- LIS und Big Data
- Methoden und Technologien



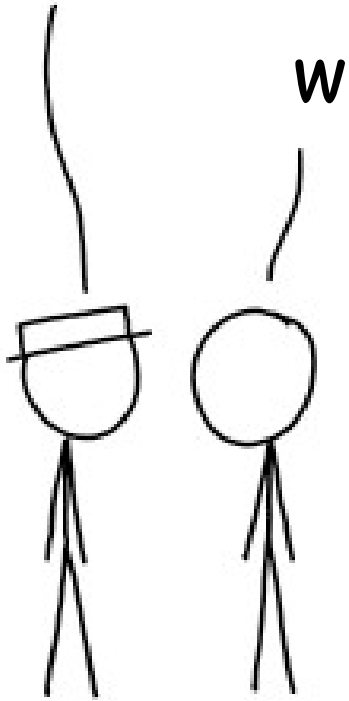
Thema und Lernziele

- LIS und Big Data
  - Fokus auf wissenschaftliche Bibliotheken

- Lernziele
  - verstehen, warum/wie Big Data Bibliotheken betrifft
  - den Unterschied zu traditionellen Medien verstehen
  - Herausforderungen für Bibliotheken (er)kennen und aus Fallbeispielen ableiten können

LIS und Big Data

What is the Library's role with “big data”?

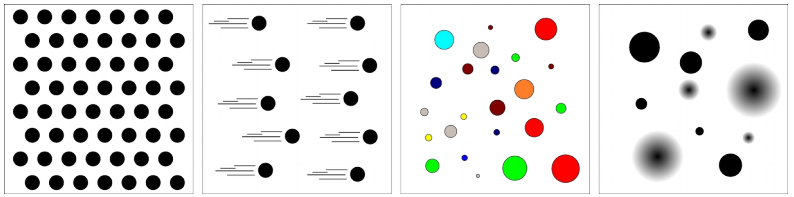


What is the Library's role with any information?

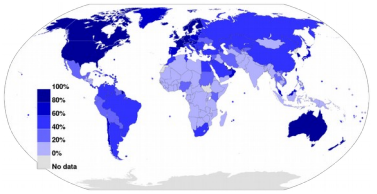
(R. Munroe/M. Furlough)

LIS und Big Data

1. Beispiele



2. Ursachen



3. Herausforderungen

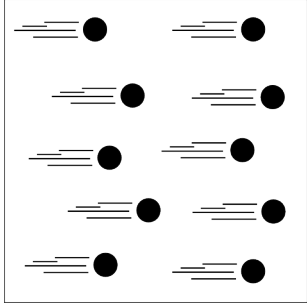
Sammeln

Erschließen

Vermitteln

4. Fallbeispiel: Web-Archivierung

INTERNET ARCHIVE
WayBackMachine



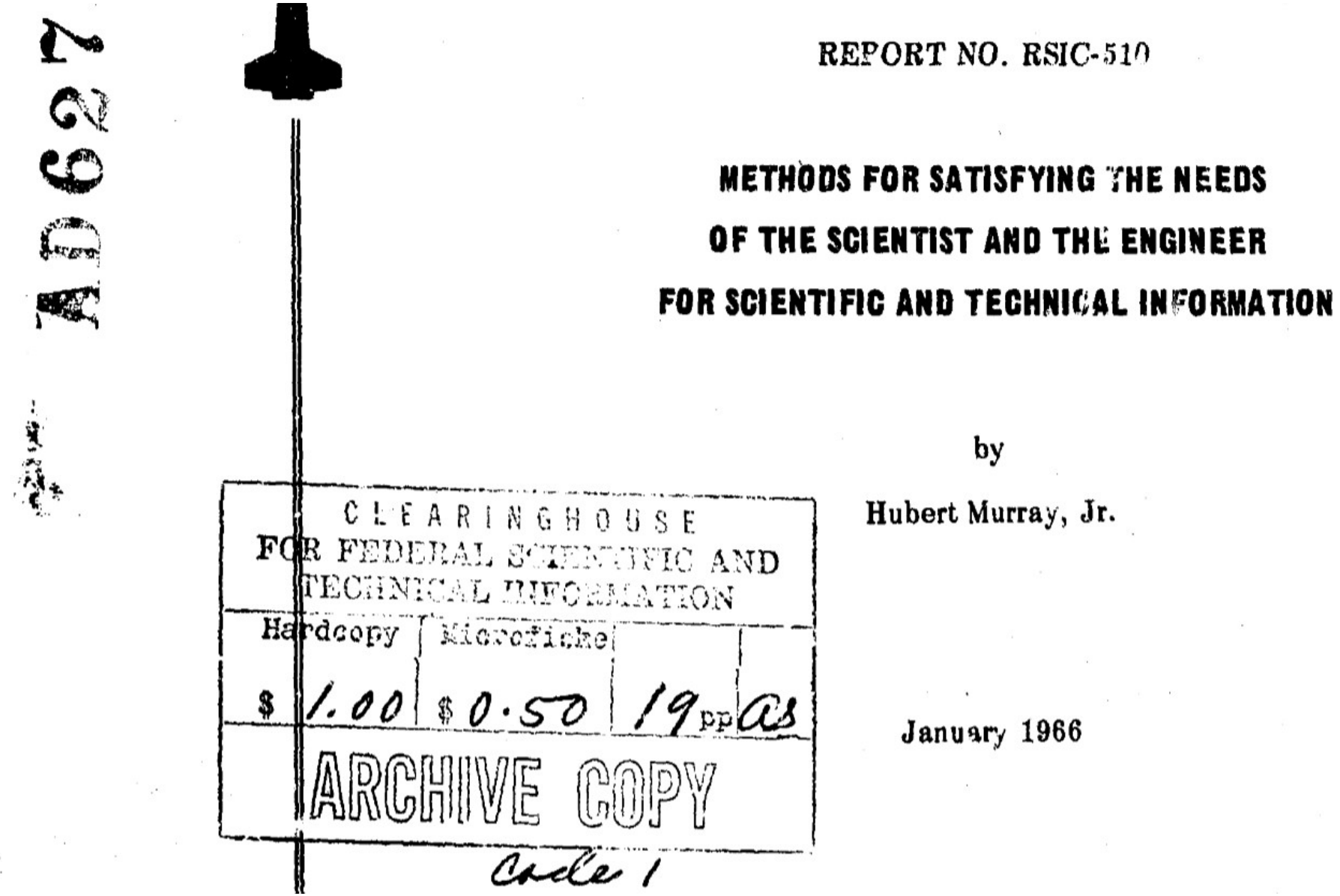
Velocity

In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would ***require 1.5 months****, reading 8 hours every day,* ***to get through 1 day's technical output****, and at the end of that period, he would have fallen 5.5 years behind in his reading!*

Velocity

In order to emphasize the problem, it is worthwhile to cite some figures. The rate at which technical documents are produced at the present time is estimated to be well over 500,000 per year. **In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require $1\frac{1}{2}$ months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen $5\frac{1}{2}$ years behind in his reading!** Even in attempting to read the portion of the literature in a single subject field such as chemistry, he would find himself falling behind an estimated 850,000 pages per year. This production rate of scientific information will undoubtedly increase as countries such as China and India begin to produce technical work commensurate with their size.

Velocity



AD627

REPORT NO. RSIC-510

METHODS FOR SATISFYING THE NEEDS
OF THE SCIENTIST AND THE ENGINEER
FOR SCIENTIFIC AND TECHNICAL INFORMATION

by

Hubert Murray, Jr.

January 1966

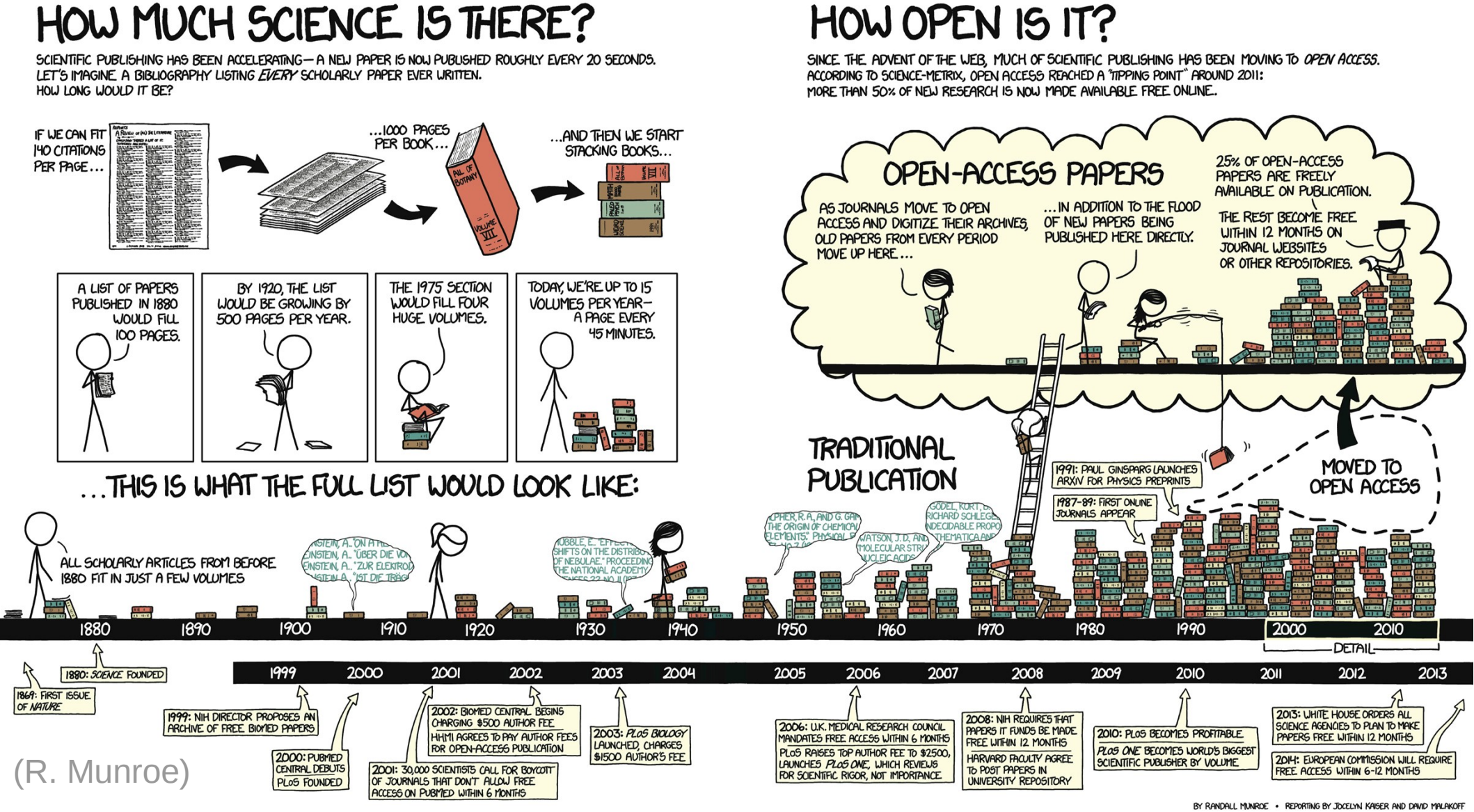
CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION

Hardcopy	Microfiche		
$ 1.00	$ 0.50	19 pp	as

ARCHIVE COPY

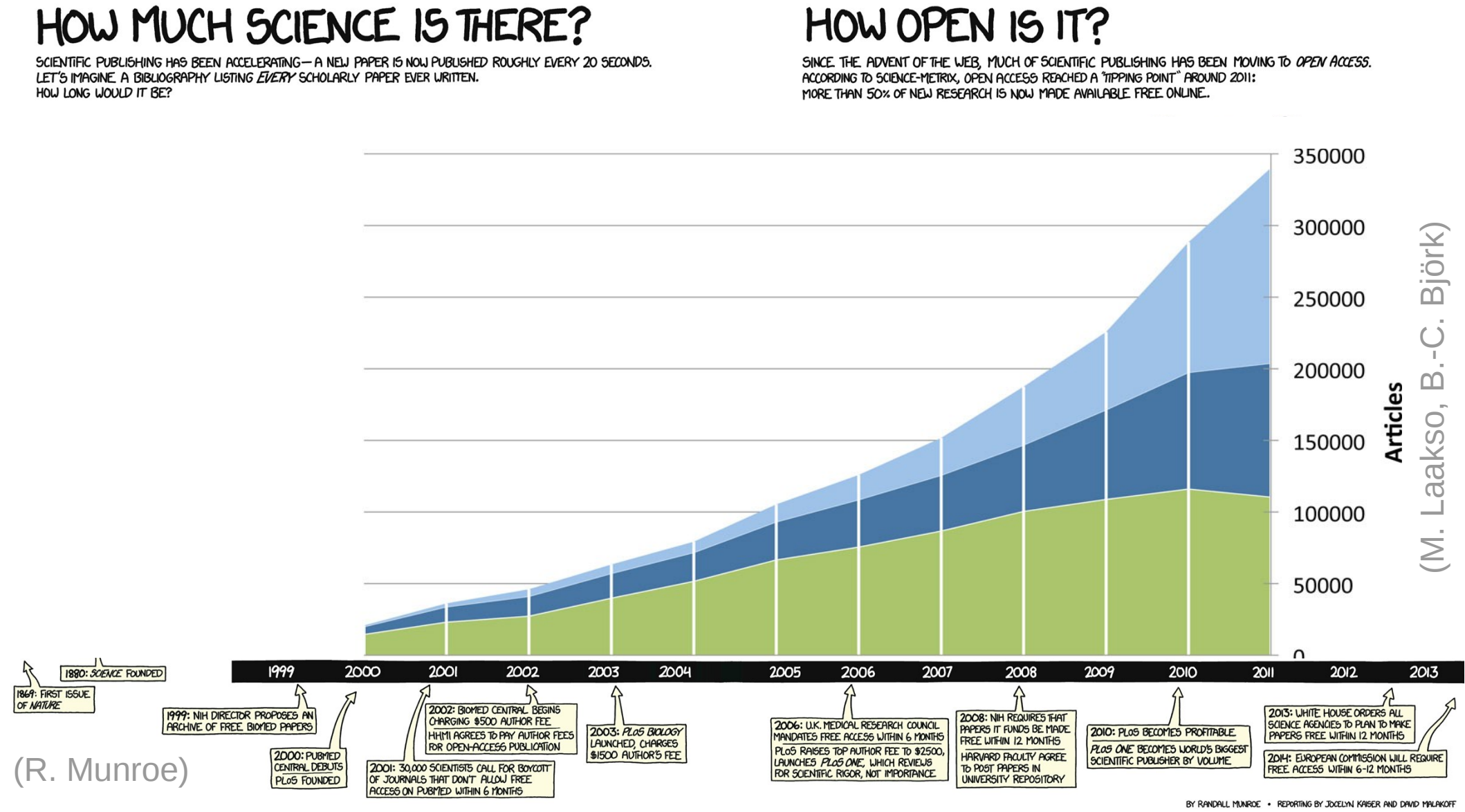
Code 1

Volume



(R. Munroe)

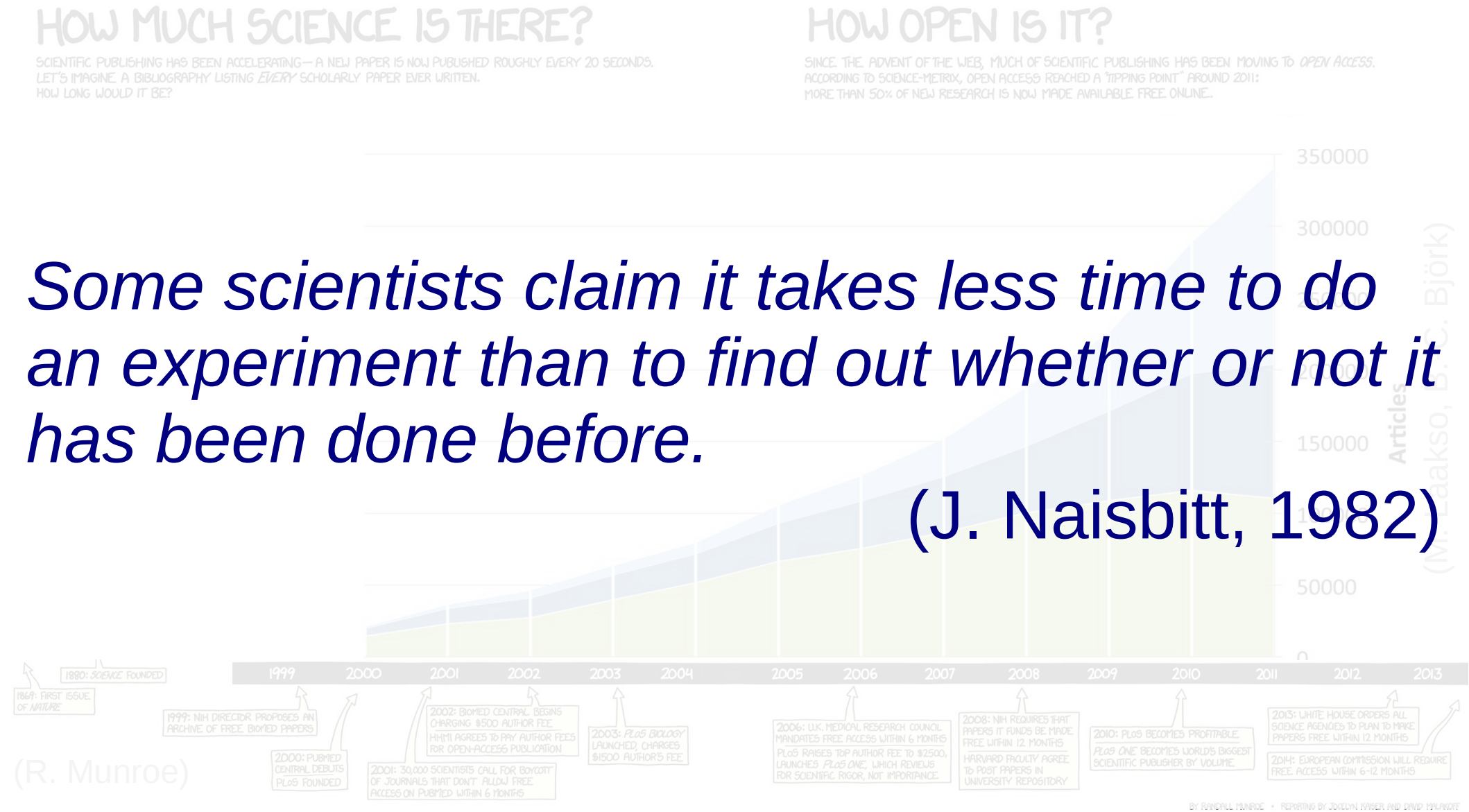
Volume



(M. Laakso, B.-C. Björk)

(R. Munroe)

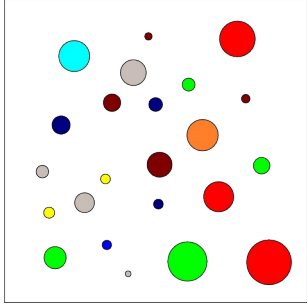
Volume



Some scientists claim it takes less time to do an experiment than to find out whether or not it has been done before.

(J. Naisbitt, 1982)

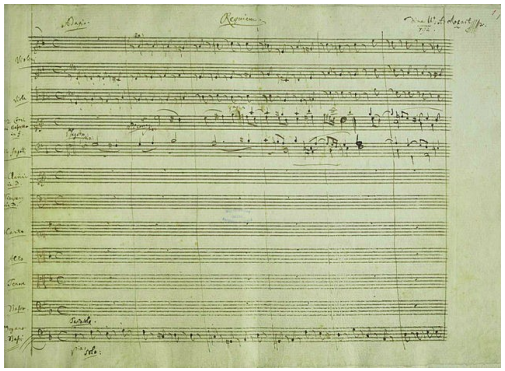
Variety



Variety

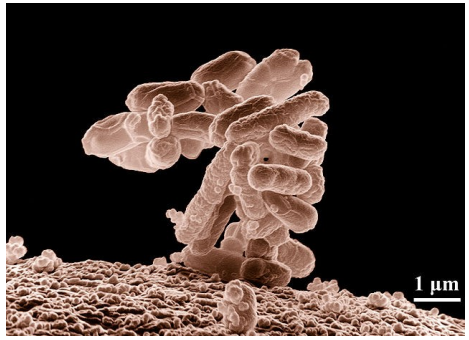

```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38        30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12        30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887209 96        30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60        30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08        30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48        30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12        30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56        30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84        30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77        30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24        30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60        30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72        30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24        30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25        30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38        30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25        30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48        30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44        30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64        30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88        30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68        30198750        http://m.imdb.com/title
:
```

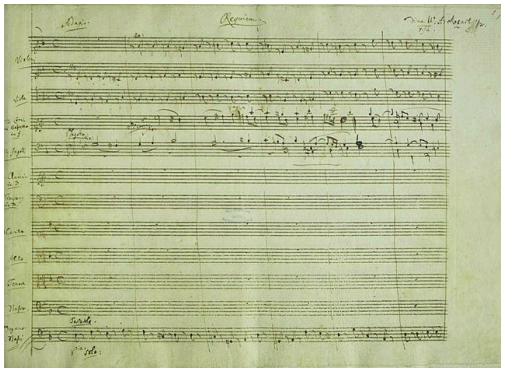
Variety



```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38        30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12        30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96        30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60        30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08        30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48        30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12        30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56        30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84        30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77        30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24        30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60        30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72        30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24        30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25        30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38        30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25        30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48        30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44        30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64        30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88        30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68        30198750        http://m.imdb.com/title
:
```

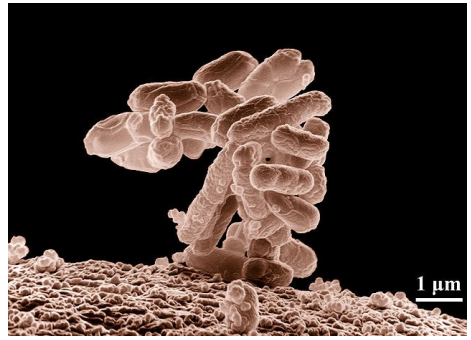
Variety

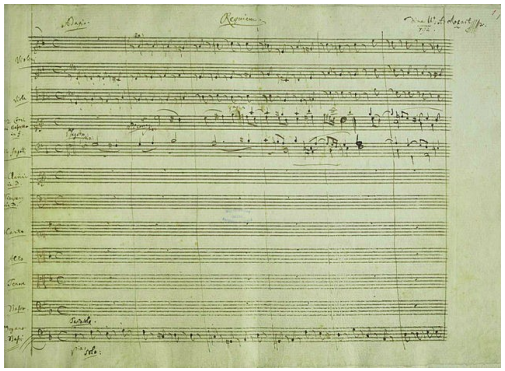




```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38   30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12   30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96   30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60   30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08   30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48   30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12   30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56   30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84   30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77   30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24   30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60   30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72   30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24   30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25   30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38   30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25   30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48   30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44   30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64   30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88   30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68   30198750        http://m.imdb.com/title
:
```

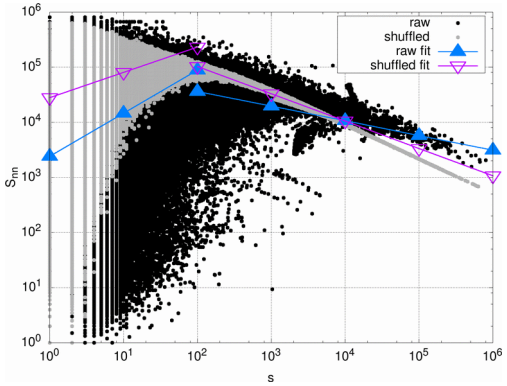
Variety

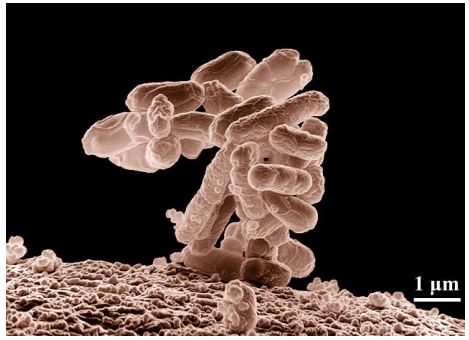


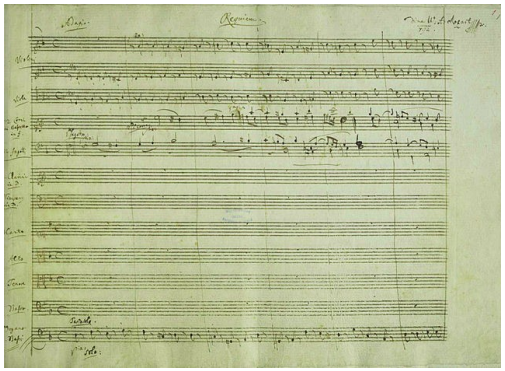


```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38        30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12        30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96        30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60        30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08        30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48        30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12        30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56        30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84        30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48        30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77        30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24        30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60        30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72        30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24        30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25        30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38        30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25        30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48        30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44        30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64        30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88        30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68        30198750        http://m.imdb.com/title
:
```

Variety

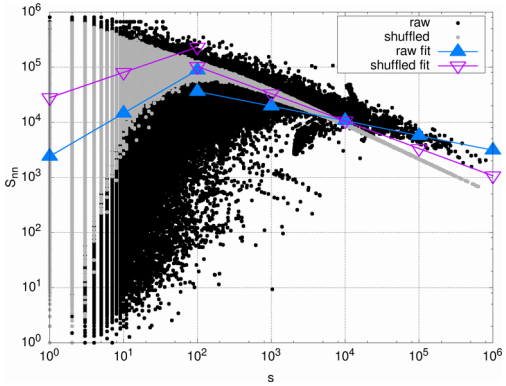


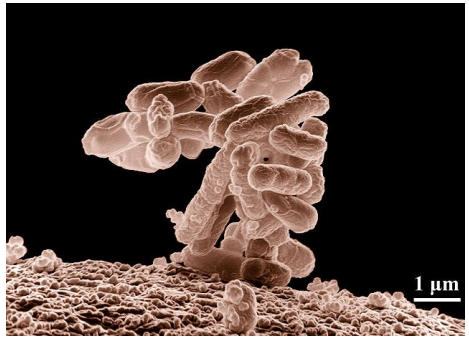


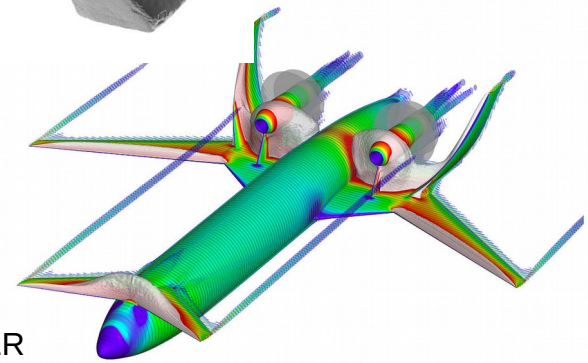


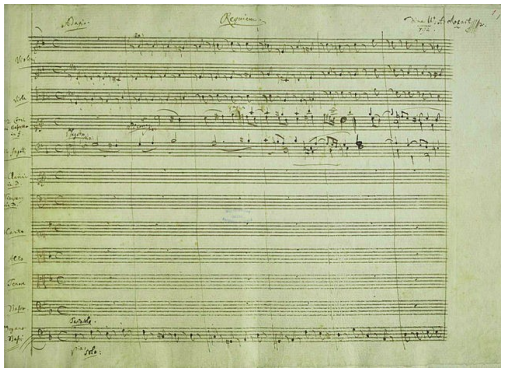
```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38   30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12   30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887209 96   30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60   30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08   30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48   30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12   30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56   30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84   30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77   30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24   30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60   30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72   30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24   30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25   30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38   30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25   30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48   30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44   30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64   30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88   30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68   30198750        http://m.imdb.com/title
:
```

Variety



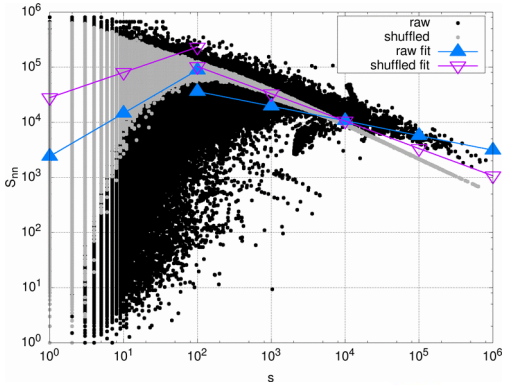


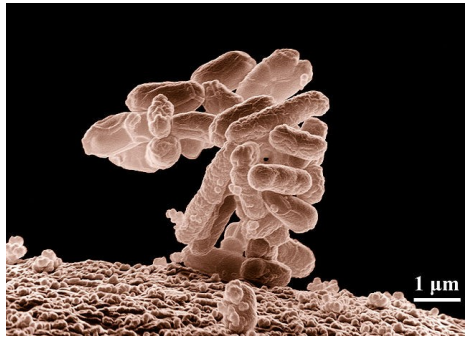


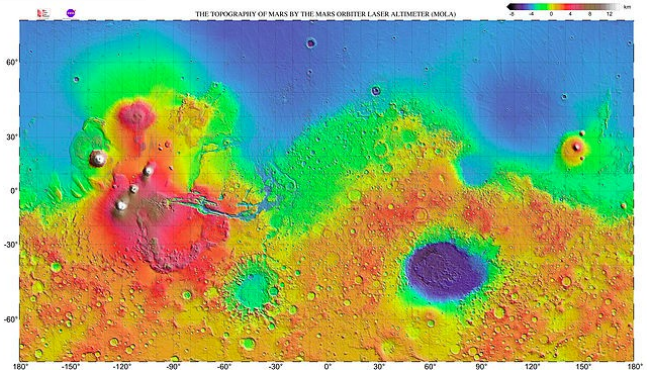


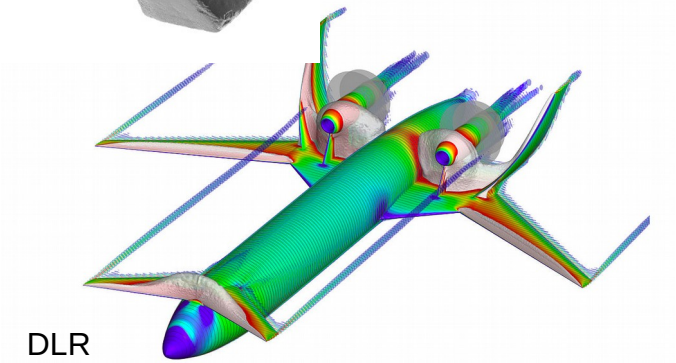
```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38   30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12   30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96   30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60   30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08   30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48   30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12   30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56   30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84   30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48   30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77   30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24   30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60   30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72   30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24   30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25   30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38   30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25   30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48   30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44   30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64   30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88   30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68   30198750        http://m.imdb.com/title
:
```

Variety

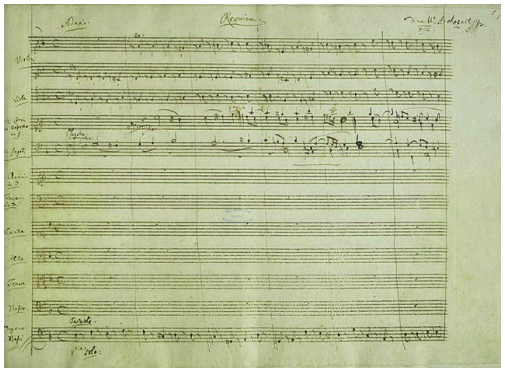






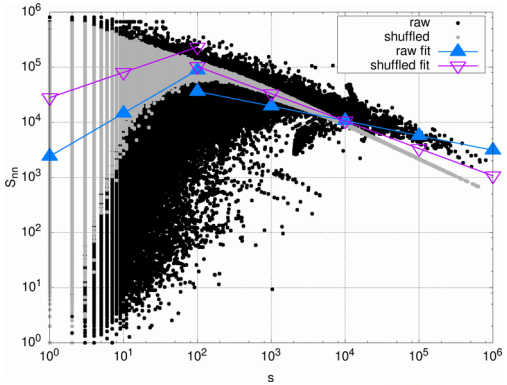


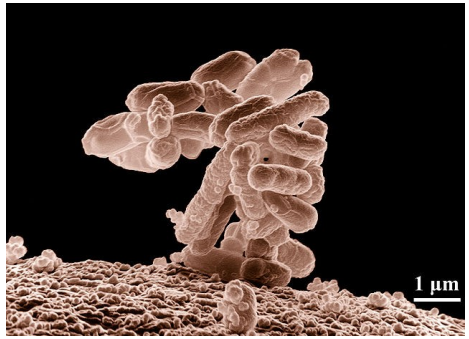
DLR

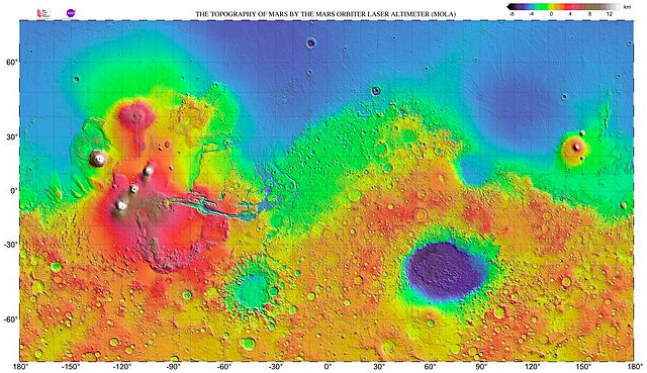


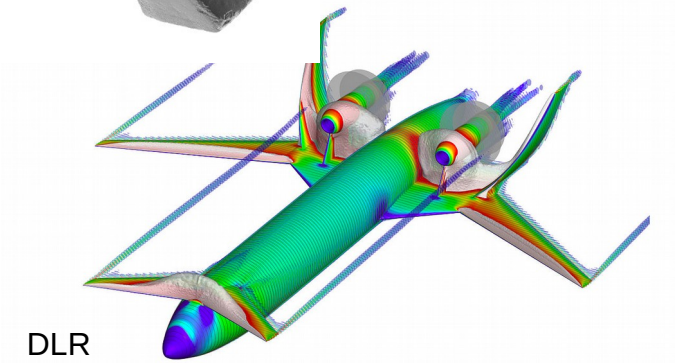
```
Sun Jul 27 17:25:36 +0000 2014  49344712182821273 8   30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  42363841815827251 2   30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  50845946828872089 6   30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  44450533006063616 0   30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  51526731620105420 8   30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  51741356235740364 8   30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  47433915803581644 8   30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  48703045345125171 2   30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  42029890760462745 6   30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  52294595666105958 4   30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  51741356235740364 8   30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  54408951127830937 7   30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  52353512175306342 4   30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  50679164171872256 0   30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  54742155766950707 2   30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  42227704515489382 4   30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  51741452238743142 5   30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  52353628359625113 8   30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  52508956940725862 5   30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  51741413394037964 8   30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  53523003545697894 4   30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  54232213010568806 4   30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  44346107150521548 8   30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  53202385246201036 8   30198750        http://m.imdb.com/title
:
```

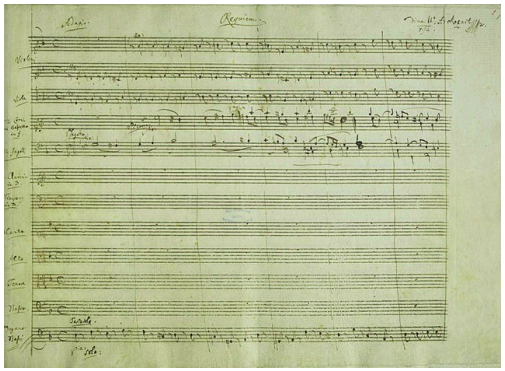
Variety







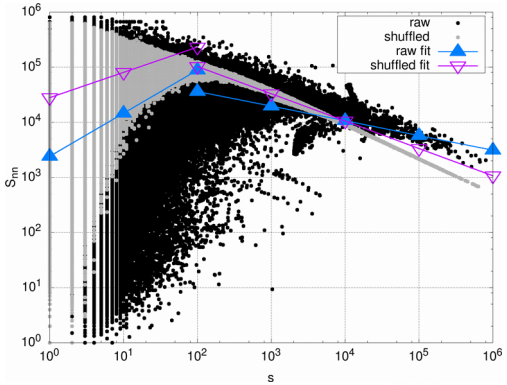


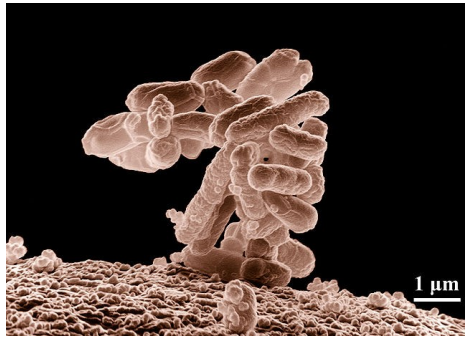


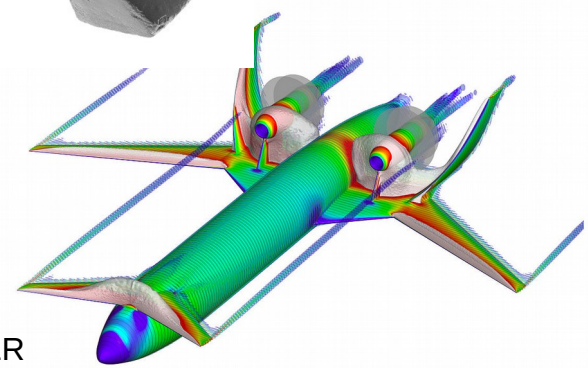
```
Sun Jul 27 17:25:36 +0000 2014  493447121828212738      30198750        http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  423638418158272512      30198750        http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  508459468288720896      30198750        https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  444505330060636160      30198750        http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  515267316201054208      30198750        http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  517413562357403648      30198750        http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  474339158035816448      30198750        http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  487030453451251712      30198750        http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  420298907604627456      30198750        http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  522945956661059584      30198750        https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  517413562357403648      30198750        http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  544089511278309377      30198750        http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  523535121753063424      30198750        https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  506791641718722560      30198750        http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  547421557669507072      30198750        http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  422277045154893824      30198750        http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  517414522387431425      30198750        http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  523536283596251138      30198750        https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  525089569407258625      30198750        http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  517414133940379648      30198750        http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  535230035456978944      30198750        http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  542322130105688064      30198750        http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  443461071505215488      30198750        http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  532023852462010368      30198750        http://m.imdb.com/title
:
```

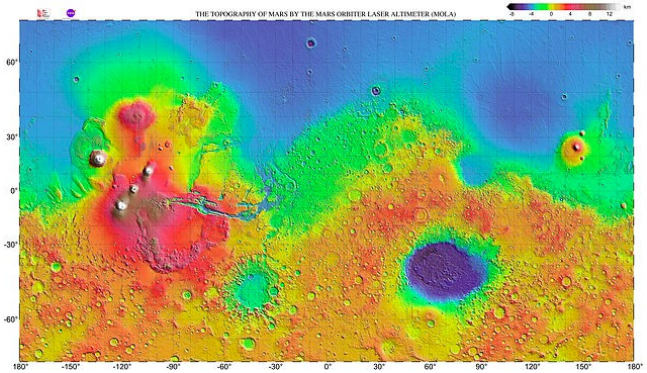
```
for row in reversed(rows):
    url = self.urls[row]
    if self.url_is_publisher(url):
        # a true positive found
        tp_rate.append(tp_rate[-1] + 1)
        fp_rate.append(fp_rate[-1])
        tp_count += 1
    else:
        # a false negative found
        tp_rate.append(tp_rate[-1])
        fp_rate.append(fp_rate[-1] + 1)
        fp_count += 1

# remove first (artificial) element
fp_rate.pop(0)
tp_rate.pop(0)

return np.array(fp_rate)/float(fp_count),
```

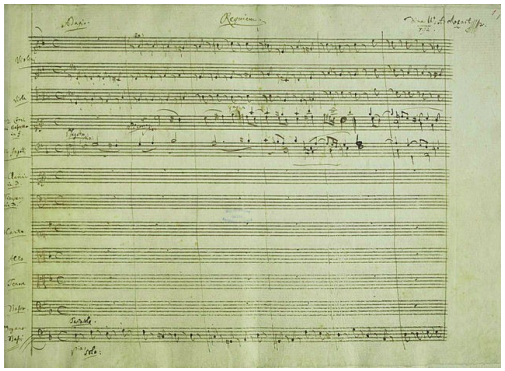
Variety







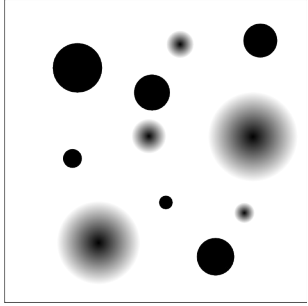




```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38  30198750  http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12  30198750  http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96  30198750  https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60  30198750  http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08  30198750  http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48  30198750  http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48  30198750  http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12  30198750  http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56  30198750  http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84  30198750  https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48  30198750  http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77  30198750  http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24  30198750  https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60  30198750  http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72  30198750  http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24  30198750  http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25  30198750  http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38  30198750  https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25  30198750  http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48  30198750  http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44  30198750  http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64  30198750  http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88  30198750  http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68  30198750  http://m.imdb.com/title
```

```
for row in reversed(rows):
    url = self.urls[row]
    if self.url_is_publisher(url):
        # a true positive found
        tp_rate.append(tp_rate[-1] + 1)
        fp_rate.append(fp_rate[-1])
        tp_count += 1
    else:
        # a false negative found
        tp_rate.append(tp_rate[-1])
        fp_rate.append(fp_rate[-1] + 1)
        fp_count += 1

# remove first (artificial) element
fp_rate.pop(0)
tp_rate.pop(0)

return np.array(fp_rate)/float(fp_count),
```

Veracity



REFERENCES

16. NEED CITATION FROM TELECOM-CITIES
17. PBS. (n.d.). *Forensic anthropology*. Retrieved September 24, 2005, from PBS Web site: http://www.pbs.org/opb/historydetectives/techniques/forensic.html
18. Multidisciplinary Center for Earthquake Engineering Research (2005) Ibid.
19. Washington State Department of Transportation. (1940)*"Galloping Gertie" collapses November 7, 1940.* Retrieved September 30, 2005, from Washington State Department of Transportation Web site: http://www.wsdot.wa.gov/TNBhistory/Connections/connections3.htm
20. Institute for the Future. NEED CITATION FROM TEN YEAR FORECAST?

Ursachen

- Wissensgesellschaft
- Wachstum und Wohlstand
- globaler Wettbewerb
- Digitalisierung
- Open Science
- Anforderungen aus Politik und Gesellschaft
- Data-intensive Science
- Wandel der Rolle von Bibliotheken

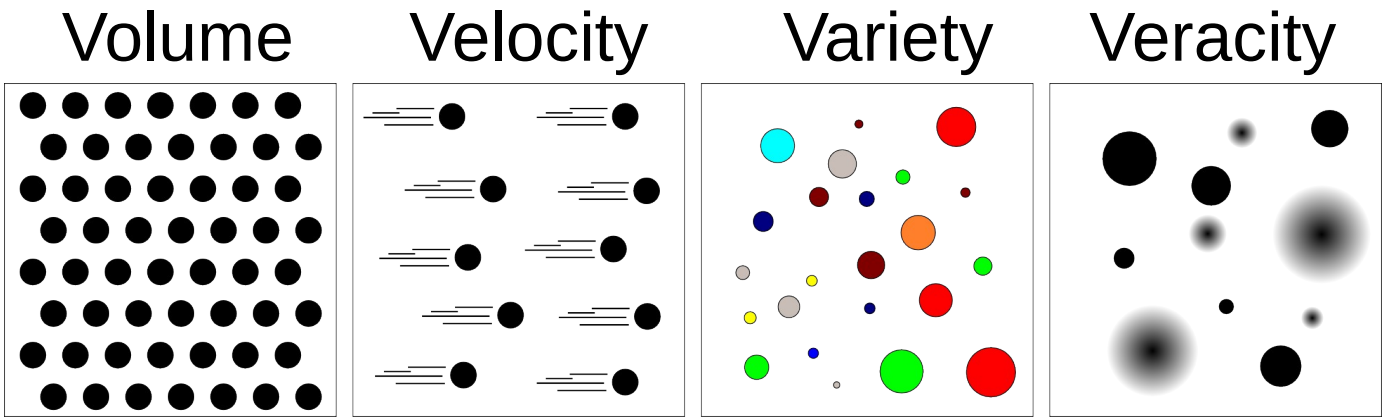
Herausforderungen

Sammeln

Erschließen

Vermitteln

Herausforderungen

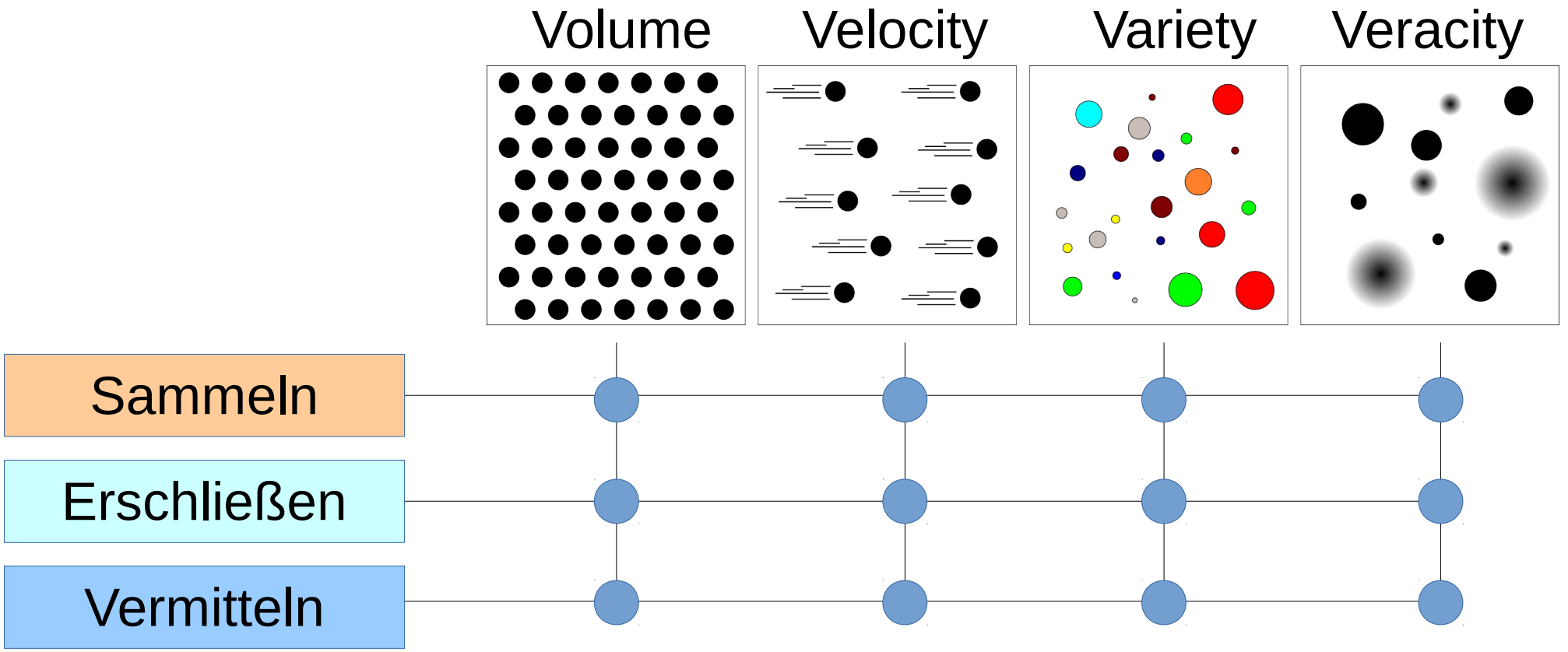


Sammeln

Erschließen

Vermitteln

Herausforderungen

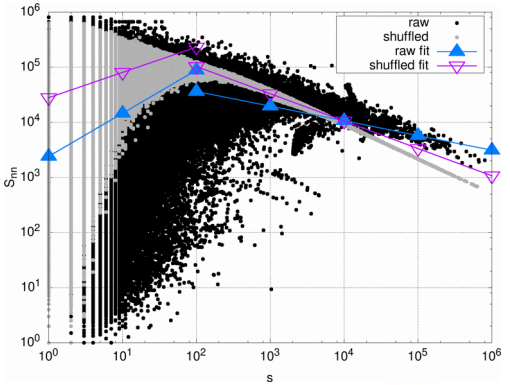


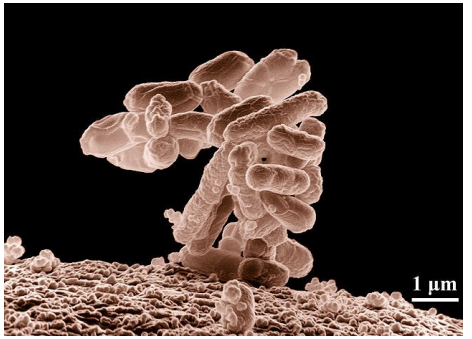
Herausforderungen

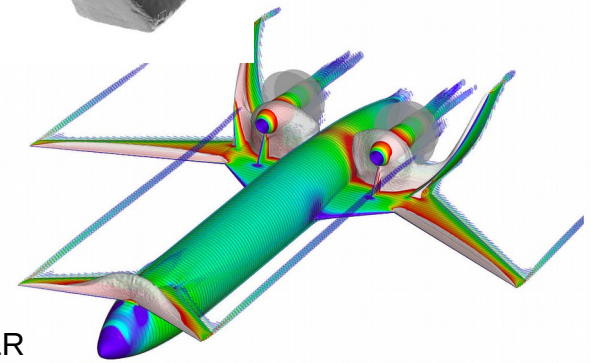
poorly organized collections; poor search tools and lack of accessibility and findability of internal data sets; lack of awareness of available third-party data sets; and copyright and intellectual property issues

(G.-L. Murnane, 2012)

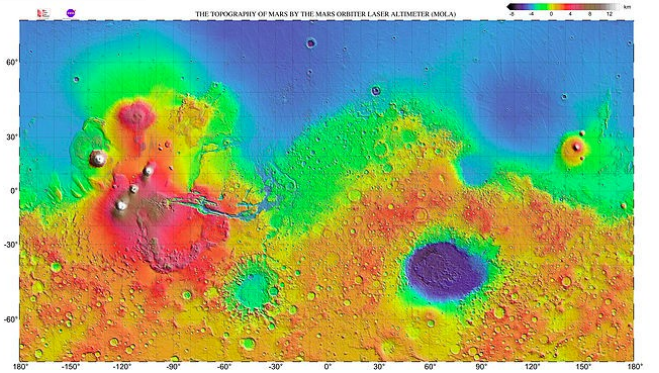
Fallbeispiel: Web-Archivierung

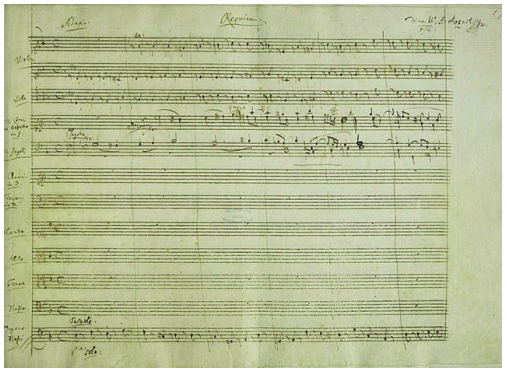






DLR





```
Sun Jul 27 17:25:36 +0000 2014  4934471218282127 38  30198750  http://www.nature.com/n
Thu Jan 16 02:10:44 +0000 2014  4236384181582725 12  30198750  http://oelemento.wordpr
Sun Sep 07 03:39:18 +0000 2014  5084594682887208 96  30198750  https://sites.google.co
Fri Mar 14 16:08:23 +0000 2014  4445053300606361 60  30198750  http://raetschlab.org/s
Thu Sep 25 22:31:16 +0000 2014  5152673162010542 08  30198750  http://people.tuebingen
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48  30198750  http://projectreporter.
Wed Jun 04 23:57:22 +0000 2014  4743391580358164 48  30198750  http://hitseq2014.sched
Thu Jul 10 00:28:03 +0000 2014  4870304534512517 12  30198750  http://hitseq.org/
Mon Jan 06 21:00:43 +0000 2014  4202989076046274 56  30198750  http://www.nytimes.com/
Fri Oct 17 03:03:26 +0000 2014  5229459566610595 84  30198750  https://www.facebook.co
Wed Oct 01 20:39:41 +0000 2014  5174135623574036 48  30198750  http://projectreporter.
Sun Dec 14 11:20:22 +0000 2014  5440895112783093 77  30198750  http://www.biomedcentra
Sat Oct 18 18:04:34 +0000 2014  5235351217530634 24  30198750  https://github.com/ga4g
Tue Sep 02 13:11:57 +0000 2014  5067916417187225 60  30198750  http://www.mskcc.org/mo
Tue Dec 23 16:00:44 +0000 2014  5474215576695070 72  30198750  http://www.fredericksbu
Sun Jan 12 08:01:07 +0000 2014  4222770451548938 24  30198750  http://www.slideshare.n
Wed Oct 01 20:43:29 +0000 2014  5174145223874314 25  30198750  http://projectreporter.
Sat Oct 18 18:09:11 +0000 2014  5235362835962511 38  30198750  https://github.com/ga4g
Thu Oct 23 01:01:23 +0000 2014  5250895694072586 25  30198750  http://ml4chg.org/
Wed Oct 01 20:41:57 +0000 2014  5174141339403796 48  30198750  http://projectreporter.
Thu Nov 20 00:35:59 +0000 2014  5352300354569789 44  30198750  http://www.nytimes.com/
Tue Dec 09 14:17:26 +0000 2014  5423221301056880 64  30198750  http://ml4chg.org/
Tue Mar 11 18:58:53 +0000 2014  4434610715052154 88  30198750  http://nyas.org/ML2014
Tue Nov 11 04:15:45 +0000 2014  5320238524620103 68  30198750  http://m.imdb.com/title
```

```
for row in reversed(rows):
    url = self.urls[row]
    if self.url_is_publisher(url):
        # a true positive found
        tp_rate.append(tp_rate[-1] + 1)
        fp_rate.append(fp_rate[-1])
        tp_count += 1
    else:
        # a false negative found
        tp_rate.append(tp_rate[-1])
        fp_rate.append(fp_rate[-1] + 1)
        fp_count += 1

# remove first (artificial) element
fp_rate.pop(0)
tp_rate.pop(0)

return np.array(fp_rate)/float(fp_count),
```

Fallbeispiel: Web-Archivierung

Fallbeispiel: Web-Archivierung

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Politics › Conservatives

Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern
The Guardian, Wednesday 13 November 2013 15.40 GMT

Jump to comments (1284)

A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media

Fallbeispiel: Web-Archivierung

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Politics › Conservatives

Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern
The Guardian, Wednesday 13 November 2013 15.40 GMT
Jump to comments (1284)

A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media

Losing My Revolution

How Many Resources Shared on Social Media Have Been Lost?

Hany M. SalahEldeen and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA, 23529, USA
{hany,mln}@cs.odu.edu

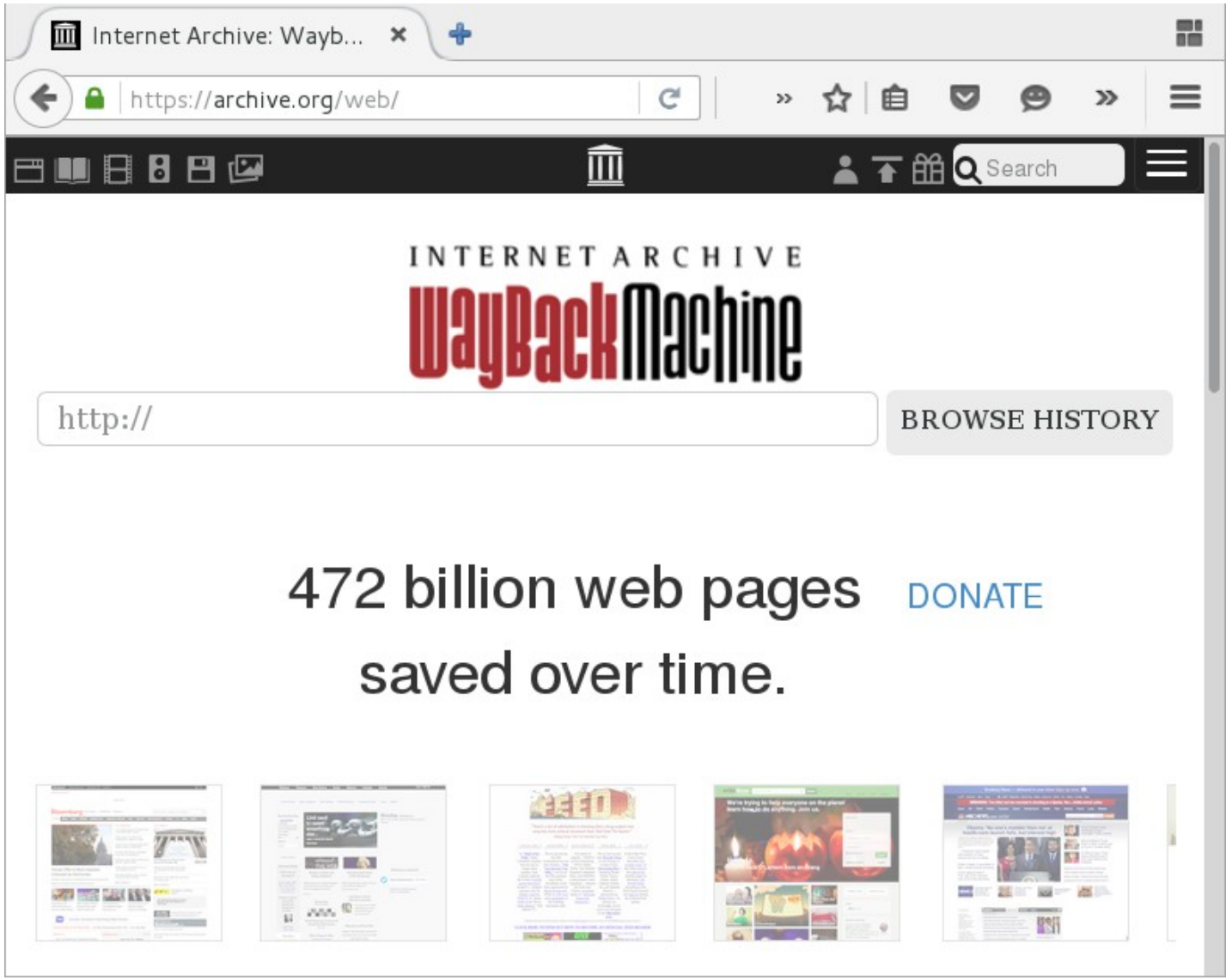
Abstract. Social media content has grown exponentially in the recent years and the role of social media has evolved from just narrating life events to actually shaping them. In this paper we explore how many resources shared in social media are still available on the live web or in public web archives. By analyzing six different event-centric datasets of resources shared in social media in the period from June 2009 to March 2012, we found about 11% lost and 20% archived after just a year and an average of 27% lost and 41% archived after two and a half years. Furthermore, we found a nearly linear relationship between time of sharing of the resource and the percentage lost, with a slightly less linear relationship between time of sharing and archiving coverage of the resource. From this model we conclude that after the first year of publishing, nearly 11% of shared resources will be lost and after that we will continue to lose 0.02% per day.

Fallbeispiel: Web-Archivierung

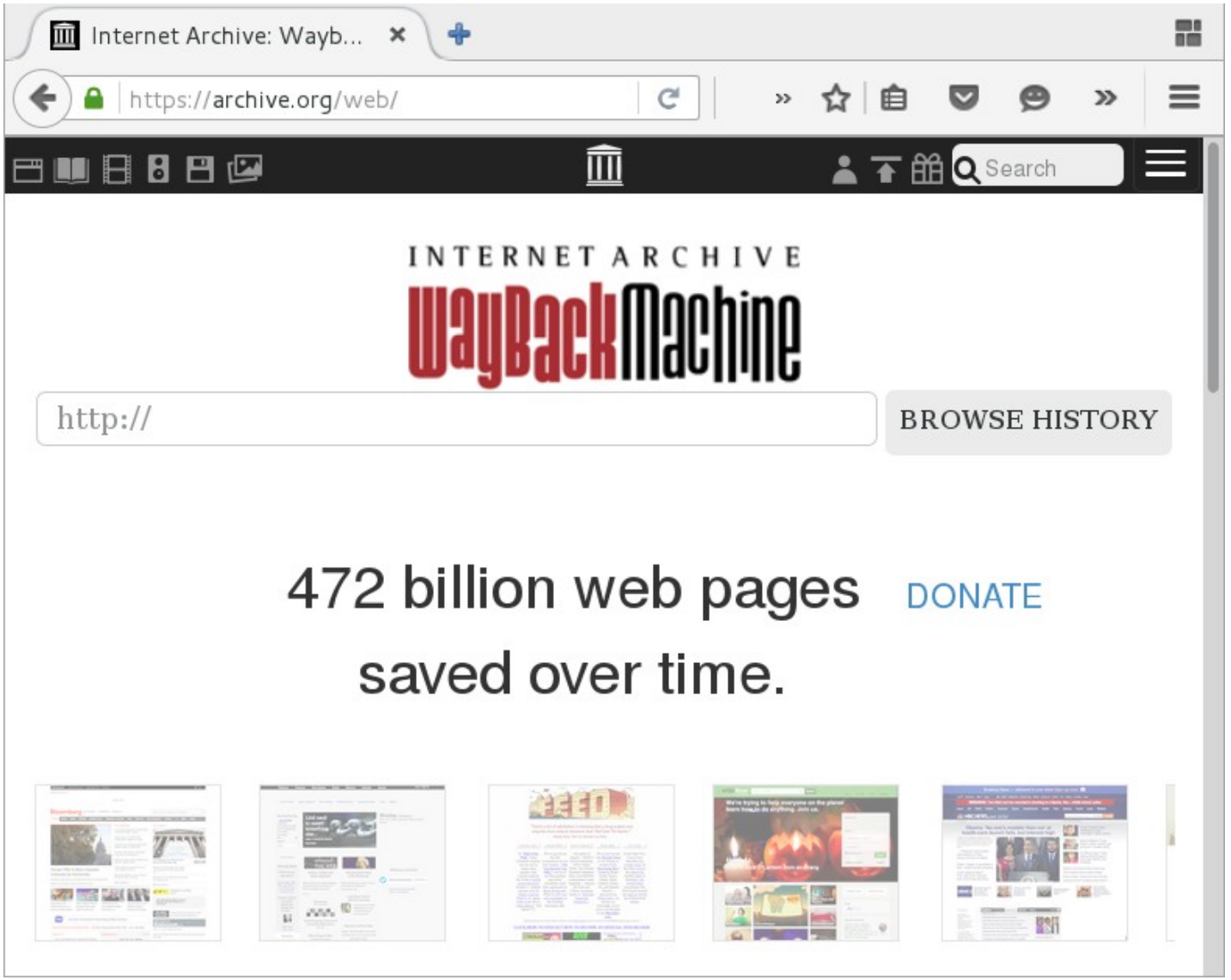
To do a large-scale social, political, or economic history of periods after the mid 1990s will at the very least be considerably enhanced by an understanding of Web data.

(William J. Turkel, Professor of History, 2015)

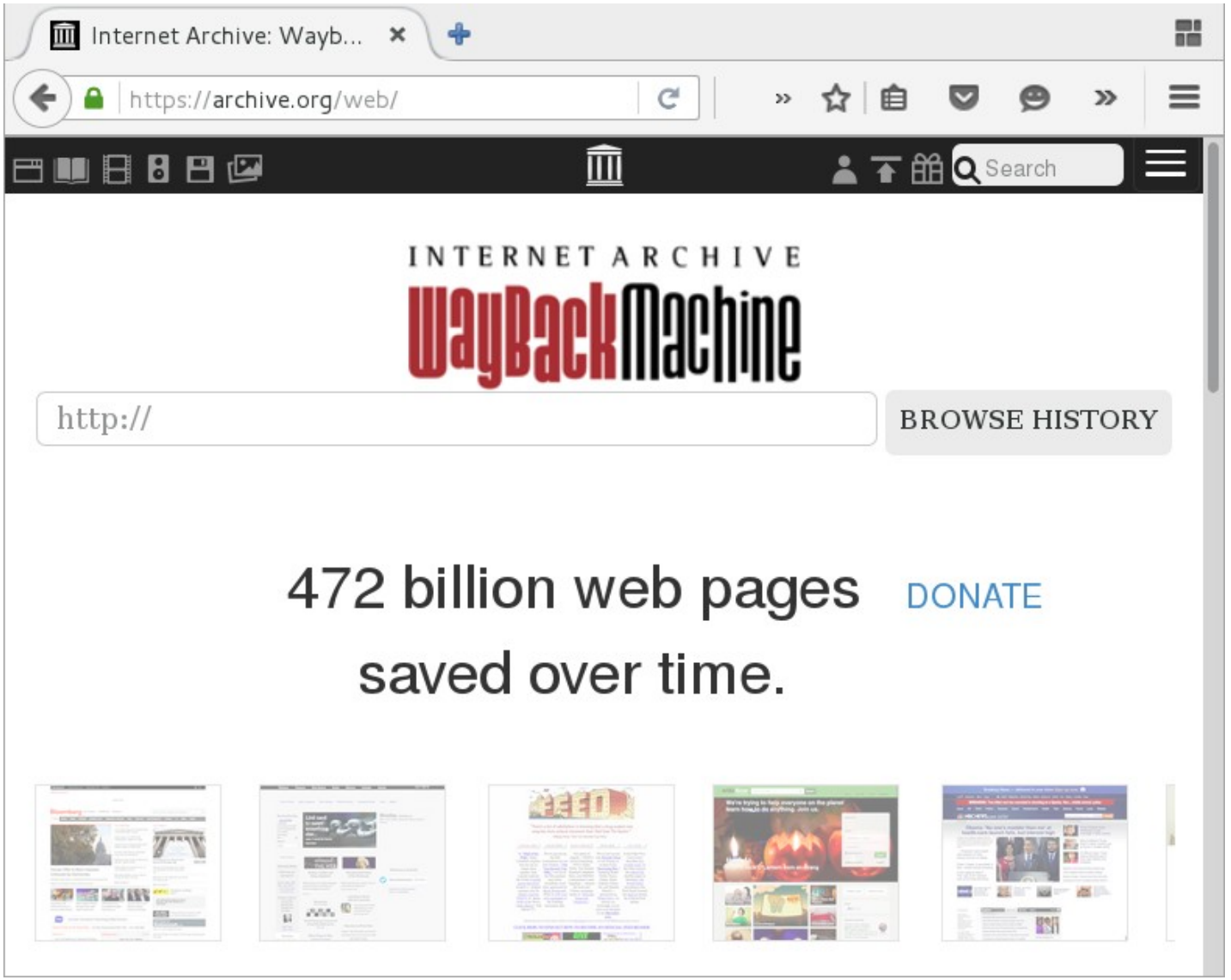
Fallbeispiel: Web-Archivierung



Fallbeispiel: Web-Archivierung

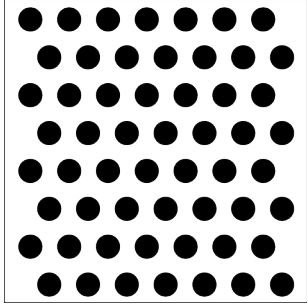


Fallbeispiel: Web-Archivierung



DEUTSCHE NATIONAL BIBLIOTHEK

Volume

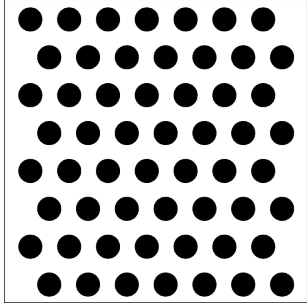


- Google[1]: 60 Bill. Seiten

  100 PB Index
- Internet Archive[2]: 19 PB (2014)
- LOC 2013[3]: 170 Mrd. Tweets

[1] http://www.google.com/insidesearch/howsearchworks/thestory/
[2] https://archive.org/web/petabox.php
[3] https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

Volume



- Google[1]: 60 Bill. Seiten

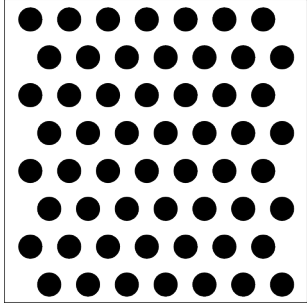
  100 PB Index
- Internet Archive[2]: 19 PB (2014)
- LOC 2013[3]: 170 Mrd. Tweets

Was ist relevant?

Welche physikalischen Ressourcen stehen zur Verfügung?

[1] http://www.google.com/insidesearch/howsearchworks/thestory/
[2] https://archive.org/web/petabox.php
[3] https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

Volume



- Google[1]: 60 Bill. Seiten

  100 PB Index
- Internet Archive[2]: 19 PB (2014)
- LOC 2013[3]: 170 Mrd. Tweets

Was ist relevant?

Welche physikalischen Ressourcen stehen zur Verfügung?

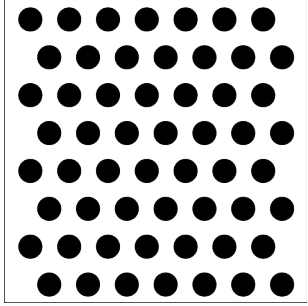
Wie können automatisch Metadaten extrahiert werden?

Wie kann der Crawlprozess dokumentiert werden?

Was sind geeignete Sammlungen?

[1] http://www.google.com/insidesearch/howsearchworks/thestory/
[2] https://archive.org/web/petabox.php
[3] https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

Volume



- Google[1]: 60 Bill. Seiten

  100 PB Index
- Internet Archive[2]: 19 PB (2014)
- LOC 2013[3]: 170 Mrd. Tweets

Was ist relevant?

Welche physikalischen Ressourcen stehen zur Verfügung?

Wie können automatisch Metadaten extrahiert werden?

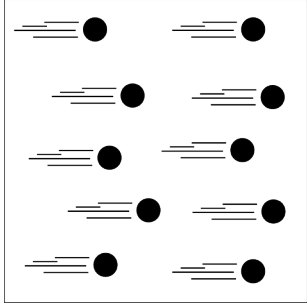
Was sind geeignete Sammlungen?

Wie kann der Crawlprozess dokumentiert werden?

Wie kann die Infrastruktur vorgehalten werden?

Was sind skalierbare Methoden für Indexierung, Suche und Retrieval?

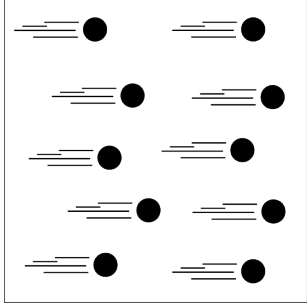
[1] http://www.google.com/insidesearch/howsearchworks/thestory/
[2] https://archive.org/web/petabox.php
[3] https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/



Velocity

- mittlere Lebensdauer einer Webseite[1]: 100 Tage
- max. Tweets pro Minute[2]: 618725 (13.7.2014)

[1] https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/
[2] https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter



Velocity

- mittlere Lebensdauer einer Webseite[1]: 100 Tage
- max. Tweets pro Minute[2]: 618725 (13.7.2014)

Was sind geeignete (Re)Crawl-Strategien?

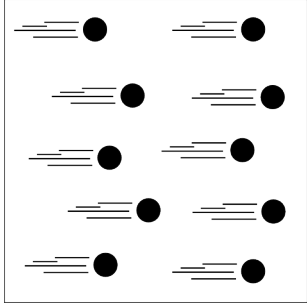
Wann ist ein Crawl komplett?

Wie schnell kann selektiert werden?

[1] https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/

[2] https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter

Velocity



- mittlere Lebensdauer einer Webseite[1]: 100 Tage
- max. Tweets pro Minute[2]: 618725 (13.7.2014)

Was sind geeignete (Re)Crawl-Strategien?

Wann ist ein Crawl komplett?

Wie schnell kann selektiert werden?

Was ist mit verschiedenen Versionen einer Seite?

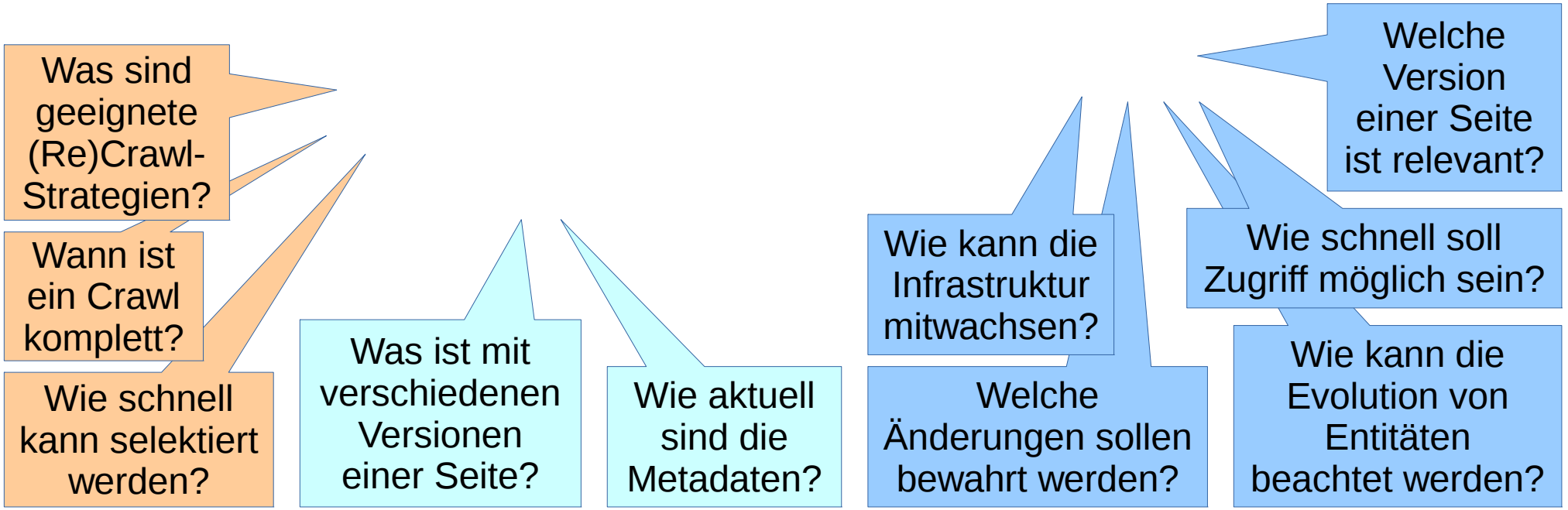
Wie aktuell sind die Metadaten?

[1] https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/

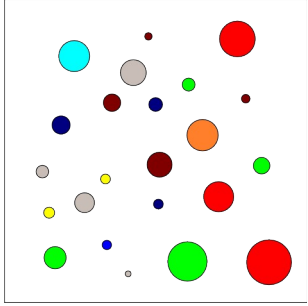
[2] https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter

Velocity

- mittlere Lebensdauer einer Webseite[1]: 100 Tage
- max. Tweets pro Minute[2]: 618725 (13.7.2014)



[1] https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/
[2] https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter

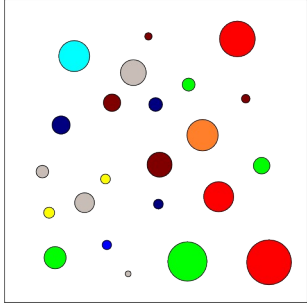


Variety

- 1327 Dateiformate[1]
- Social Media, Datenbanken, APIs, etc.
- Standards: 11x HTML, 5x CSS, …

[1] https://en.wikipedia.org/wiki/List_of_file_formats

Variety

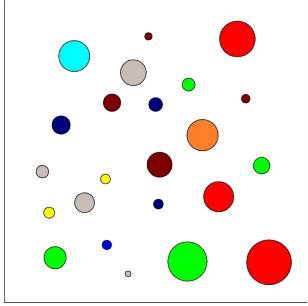


- 1327 Dateiformate[1]
- Social Media, Datenbanken, APIs, etc.
- Standards: 11x HTML, 5x CSS, …

Wie hoch ist der Aufwand?

Was ist handhabbar?

[1] https://en.wikipedia.org/wiki/List_of_file_formats



Variety

- 1327 Dateiformate[1]
- Social Media, Datenbanken, APIs, etc.
- Standards: 11x HTML, 5x CSS, …

Wie hoch ist der Aufwand?

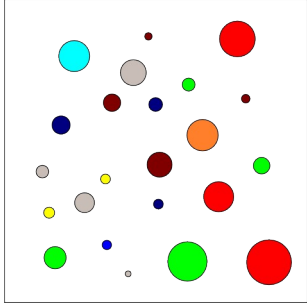
Was ist handhabbar?

Was sind geeignete Beschreibungen?

Wie kann man inhaltlich zusammenfassen?

[1] https://en.wikipedia.org/wiki/List_of_file_formats

Variety



- 1327 Dateiformate[1]
- Social Media, Datenbanken, APIs, etc.
- Standards: 11x HTML, 5x CSS, …

Wie hoch ist der Aufwand?

Was ist handhabbar?

Was sind geeignete Beschreibungen?

Wie kann man inhaltlich zusammenfassen?

Wie sah eine Webseite zu ihrer Erstellungszeit aus?

Wie können wir Zugang bewahren?

[1] https://en.wikipedia.org/wiki/List_of_file_formats

Veracity

error 404 - page not found

Sorry, the requested page does not exist on our server.

Fazit

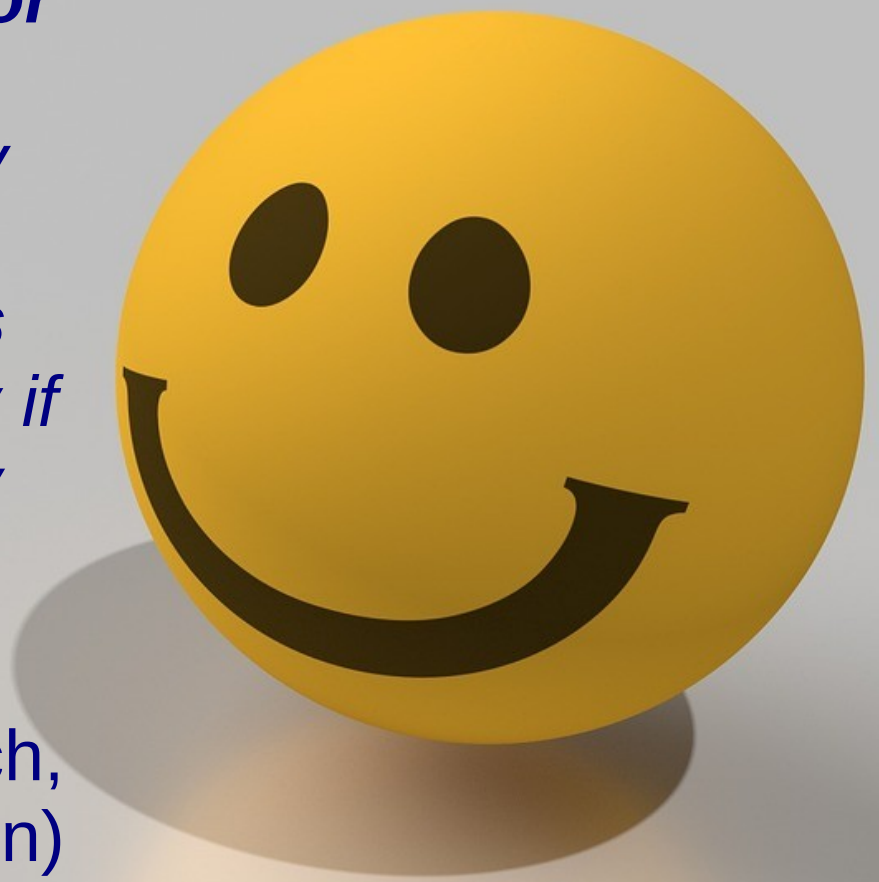
big data affects libraries directly and tangentially: directly because your library can ***use big data tools to analyze your big data sets****; and tangentially, as the faculty at your school will increasingly* ***incorporate big data into their research***

(Mark Bieraugel: Keeping up with... Big Data, 2013)

Fazit

There has probably never been a better time to be a librarian or an information professional.
We live in an information society with access to more information than ever before, and ***librarians have an important role to play*** *if people are going to successfully avoid the much discussed information overload.*

(D. Stuart, Centre for e-Research, King's College London)

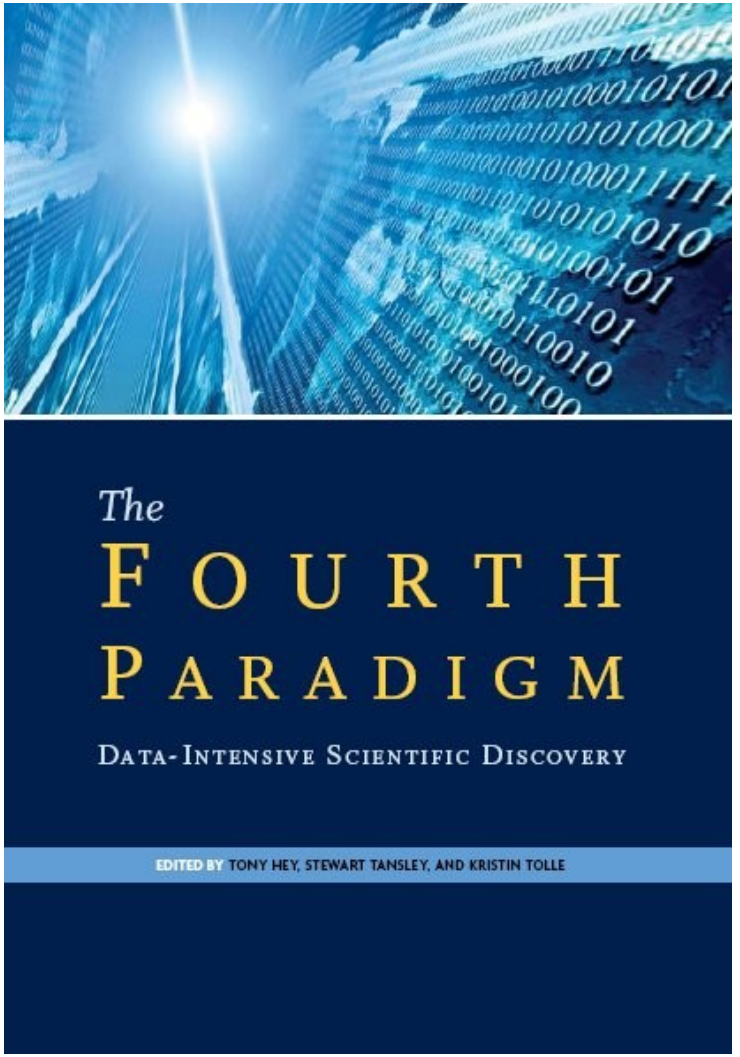


Kontrolle des Lernerfolgs

- Wie sind die LIS-Aufgabenbereiche durch Big Data betroffen?
- Welche Herausforderungen ergeben sich bei der Web-Archivierung?
- Erarbeiten Sie für ein anderes Fallbeispiel die Herausforderungen, die sich für LIS durch die Eigenschaften von Big Data ergeben.

Literatur

- G.-L. Murnane, 2012: Big Data: A Big Opportunity for Librarians
- M. Bieraugel, 2013: Keeping up with... Big Data
- F. Lohmeier, J. Mittelbach, 2014: Offenheit statt Bündniszwang
- D. Salo, 2010: Who owns our work?
- R. Brugbauer, V. Butz, 2015: Die Bibliothek: Raum im digitalen Wandel



Stand der Vorlesung

Kapitel 10: Big-Data-Technologien

- Einführung
- LIS und Big Data
- Methoden und Technologien