# Grundlagen: Informationstechnologie in Bibliotheken

Sven Koesling

ETH–Bibliothek

Herbst 2017

# Kursus / Übersicht

**19.01.2018 : Datenbanktechnologien II: BigData**

- kurze Wiederholung — wo stehen wir?
- Klärung verschiedener Begriffe und Buzzwords
- Anwendungsszenarien, Anwendung in der ETH
- In Medias Res: BigData am Bsp. Logfiles, DataScience am Bsp. Benutzerdaten

# den Inhalt eines Elements einer Webseite abfragen

```
<script type="text/javascript">
  function einausblenden(id) {
    var meinElement = document.getElementById(id);
    if (meinElement.style.display === "none") {
      meinElement.style.display = "table-row";
    } else {
      meinElement.style.display = "none";
    }
  }
</script>
```

# automatisierte Tests von Primo

- Wir prüfen nach einem Update, ob sich die Suchalgorithmen verändert haben, indem wir die Trefferzahl zwischen den beiden Systemen mit identischem Datenstand vergleichen.

# automatisierte Tests von Primo

- Wir prüfen nach einem Update, ob sich die Suchalgorithmen verändert haben, indem wir die Trefferzahl zwischen den beiden Systemen mit identischem Datenstand vergleichen.
- Der Date–Slider hat in der Vergangenheit Probleme gemacht. Wir checken, ob er plausible Ranges liefert.

# automatisierte Tests von Primo

- Wir prüfen nach einem Update, ob sich die Suchalgorithmen verändert haben, indem wir die Trefferzahl zwischen den beiden Systemen mit identischem Datenstand vergleichen.
- Der Date–Slider hat in der Vergangenheit Probleme gemacht. Wir checken, ob er plausible Ranges liefert.
- Wir testen, ob nicht–lateinische Schriftzeichen gefunden werden.

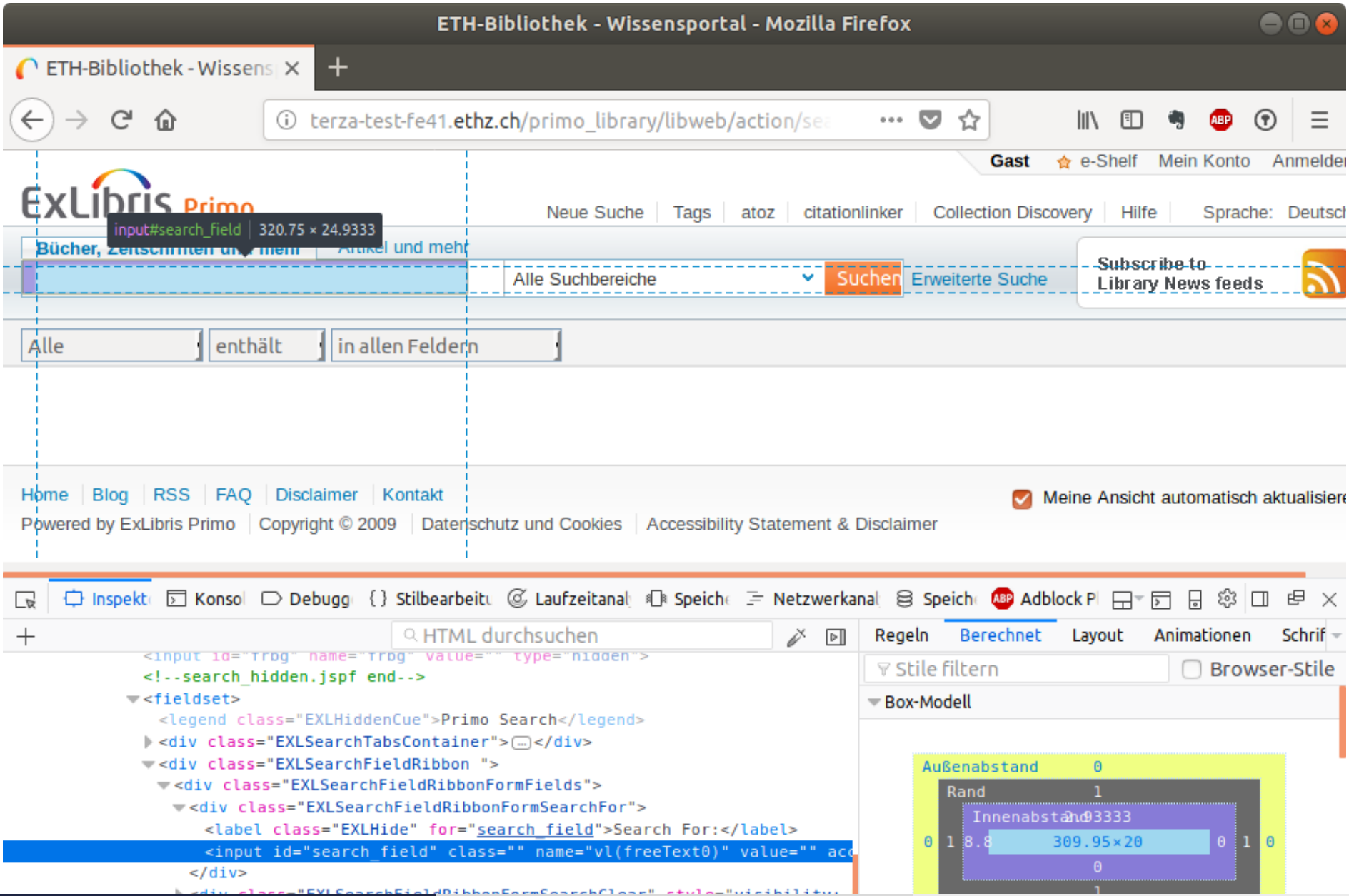
# automatisierte Tests von Primo

- Wir prüfen nach einem Update, ob sich die Suchalgorithmen verändert haben, indem wir die Trefferzahl zwischen den beiden Systemen mit identischem Datenstand vergleichen.
- Der Date–Slider hat in der Vergangenheit Probleme gemacht. Wir checken, ob er plausible Ranges liefert.
- Wir testen, ob nicht–lateinische Schriftzeichen gefunden werden.
- ...

# eine Anforderung in Gherkin formuliert

```
Szenario: Eine Suche ergibt auf den beiden Prod Systemen eine ähnliche
    Anzahl Treffer
    Wenn ich die Seite "http://terza-prod1-fe41.ethz.ch/primo-explore/
        search?vid=DADS&sortby=rank&lang=de_DE" aufrufe,
    Und ich in den Suchschlitz "Wald" eingebe,
    Und die Anzahl der Treffer nehme
    Und dann die Seite "http://terza-prod2-fe41.ethz.ch/primo-explore/
        search?vid=DADS&sortby=rank&lang=de_DE" aufrufe,
    Wenn ich in den Suchschlitz den Suchbegriff "Wald" eingebe,
    Und dort die Anzahl der Treffer nehme
    Dann sollten die Treffermengen ähnlich, d.h. die Abweichung unter
        1\%, sein.
```

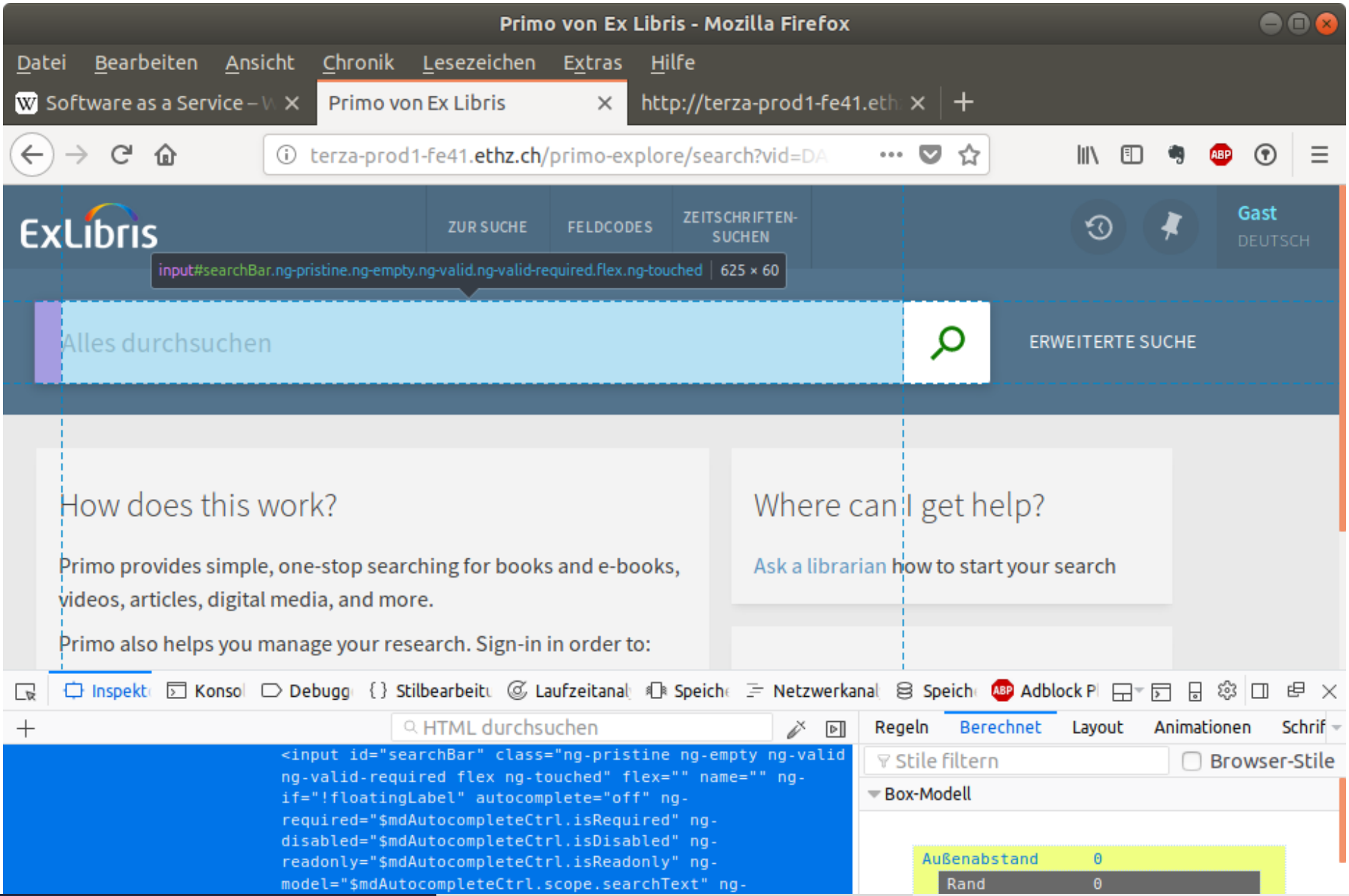
# Der Suchschlitz im alten UI



# der Quellcode des Suchschlitzes (altes UI)

```
<input name="vl(freeText0)" class="" value="" id="search_field"
    accesskey="s" type="text">
```

# Der Suchschlitz im neuen UI



# der Quellcode des Suchschlitzes (neues UI)

```
<input flex="" id="searchBar" name="" ng-if="!floatingLabel"
    autocomplete="off" ng-required="$mdAutocompleteCtrl.isRequired" ng-
    disabled="$mdAutocompleteCtrl.isDisabled" ng-readonly="
    $mdAutocompleteCtrl.isReadonly" ng-model="$mdAutocompleteCtrl.scope
    .searchText" ng-keydown="$mdAutocompleteCtrl.keydown($event)" ng-
    blur="$mdAutocompleteCtrl.blur()" ng-focus="$mdAutocompleteCtrl.
    focus()" placeholder="Alles durchsuchen" aria-owns="ul-0" aria-
    label="Alles durchsuchen" aria-autocomplete="list" role="combobox"
    aria-haspopup="true" aria-activedescendant="" aria-expanded="false"
     class="ng-pristine ng-empty ng-valid ng-valid-required flex ng-
    touched" aria-invalid="false" style="" type="search">
```

# Ruby–Snippet, um „Wald” in den Suchschlitz zu schreiben

```
Wenn(/^ich in den Suchschlitz "([^"]*)" eingebe,$/) do |q|
  fill_in('#searchBar', with: q)
  find(".button-confirm").send_keys(:enter)
end
```

# Der Testablauf auf der Kommandozeile

```
cucumber features/suche.feature:50
# language: de
Funktionalität: Suche

  Grundlage:                                          # features/suche.feature:3
      # Gegeben sei , dass die Seite "http://terza-test-fe43.ethz.ch" aufgerufen ist
    Gegeben sei , dass die Startseite aufgerufen ist # features/step_definitions/suche_steps.rb:1

  Szenario: Slider bietet eine plausible Zeitangabe zur Einschränkung der Trefferliste # features/suche.feature:48
    Wenn ich in den Suchschlitz den Suchbegriff "cucumber" eingebe,                    # features/step_definitions/
vollansicht_steps.rb:1
    Dann sollte der Slider plausible Anfangs- und Endwerte haben.                      # features/step_definitions/
suche_steps.rb:196
      **** alt: 1922--> DS-Startdatum: 1920
      **** jung: 2017--> DS-Enddatum: 2017
      **** Normalisiertes Startdatum: 1920 ***** Normalisiertes Enddatum: 2017

1 scenario (1 passed)
3 steps (3 passed)
0m16.334s
```

Die Ausgabe gibt es auch als Webseite.

# Dienstleistungen in der Cloud

Wir erinnern uns:

- SaaS

# Dienstleistungen in der Cloud

Wir erinnern uns:

- SaaS
- IaaS

# Dienstleistungen in der Cloud

Wir erinnern uns:

- SaaS
- IaaS
- PaaS

*It's not the Cloud, it's just other people's computers.*

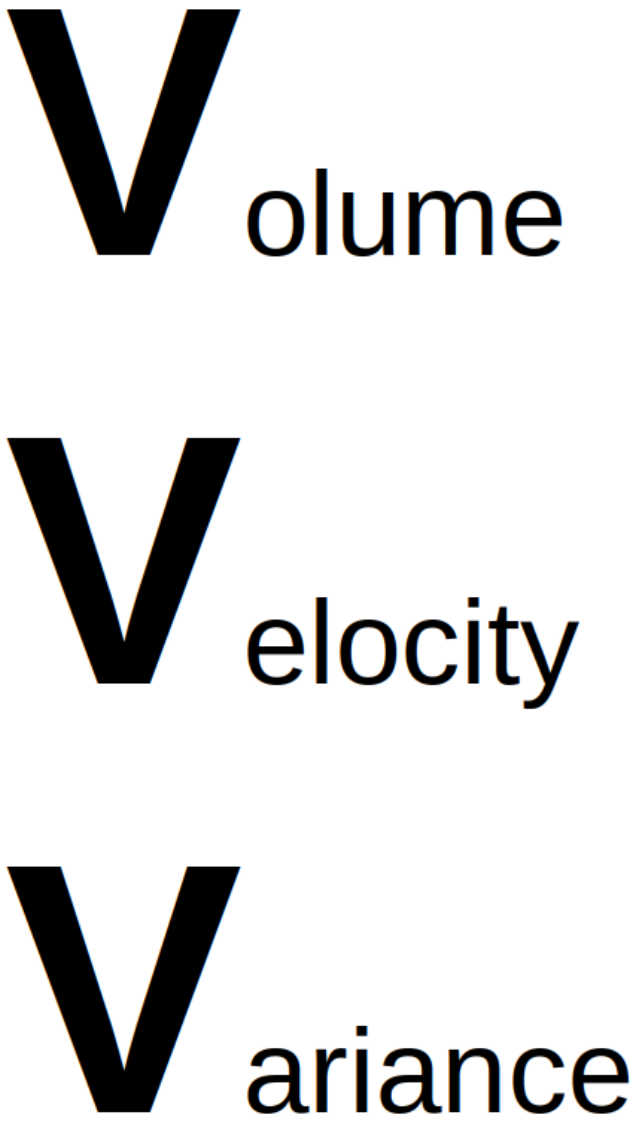
# eine Definition

Der aus dem englischen Sprachraum stammende Begriff Big Data [b detə] (von englisch big ‚groß' und data ‚Daten') bezeichnet Datenmengen, welche

- zu groß,
- zu komplex,
- zu schnelllebig
- zu schwach strukturiert

sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten.

[Seite „Big Data". In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 21. Oktober 2017, 06:34 UTC. URL: https://de.wikipedia.org/w/index.php?title=Big$_D ataoldid$ = 170175574($Abgerufen$ : 26.$Oktober$2017, 14 : 31$UTC$)]


Volume
Velocity
Variance

# Was ist nun BigData?

**DataScience**
Wissensgewinnung

- Sichtung
- Umwandlung
- Analyse
- Visualisierung

**BigData**
Datenhaltung

- Speicherung
- Sicherung
- Abfrage

**MachineLearning**
KI

- Training
- Klassifizierung
- Vorhersage

Und bei uns? 10 Millionen Datensätze sind doch 'ne ganze Menge...

# Datenmengen an der ETH–Bibliothek

- Bildarchiv Online: ca. 600‘000 Datensätze
- Aleph: ca. 7,9 Millionen Datensätze
- Primo: ca. 10,1 Millionen Datensätze

# ein frustrierender Vergleich:

Ein Beispiel für Datensatzzahlen im BigData–Bereich sind die ca. 15 Milliarden Tweets die über Twitter in einem Monat versendet werden.

Und wie schnell verändern sich die Daten?

# Velocity im WWW

Und die Varianz?

# Varianz

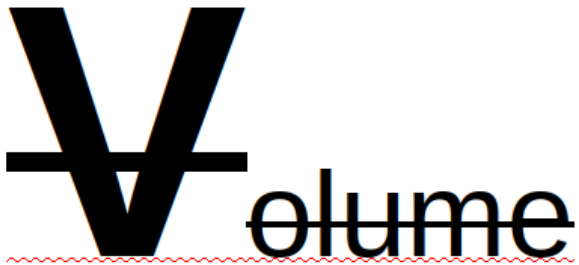
## ein Buch

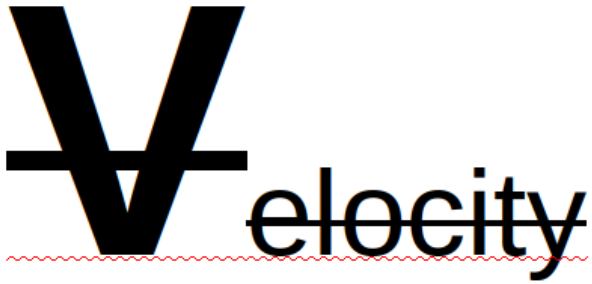
```
<recordid>ebi01_prod010616366</recordid>
<type>book</type>
<title>Sophocles: four tragedies</title>
<creator>Sophocles, v497-v407</creator>
<creator>Oliver Taplin, 1943-</creator>
<edition>First edition</edition>
<publisher>Oxford, United Kingdom : Oxford University Press</publisher>
<creationdate>2015</creationdate>
<subject>Sophocles -- Translations into English</subject>
<subject>Oedipus, Greek mythological figure  Drama</subject>
```

# Varianz

## ein Bild

```
<recordid>ebi01_prod010103021</recordid>
<type>image</type>
<title>[Aias und Kassandra]</title>
<creator>Heinrich Meyer, 1760-1832</creator>
<creator>Johann Heinrich Lips, 1758-1817</creator>
<publisher>[Weimar]</publisher>
<creationdate>[1794]</creationdate>
<format>1 Druckgraphik : Aquatinta und Radierung in Schwarz und
    Braunrot ; Bild 18,5 x 23,1 cm, Blatt 23,4 x 32,2 cm</format>
<identifier><b>DOI: </b>10.3931/e-rara-37966 ; $$CBIBL$$VJoachim Kruse:
     Johann Heinrich Lips 1758-1817, Coburg 1989, S. 220 ;
    $$CBIBL$$VErschienen in: Heinrich Meyer/Carl August Böttiger (Hg.):
     Über den Raub der Cassandra auf einem alten Gefässe von gebrannter
     Erde, Weimar 1794, S. 93</identifier>
<subject>Kassandra, Fiktive Gestalt</subject>
<subject>Athena, Göttin</subject>
```


Volume

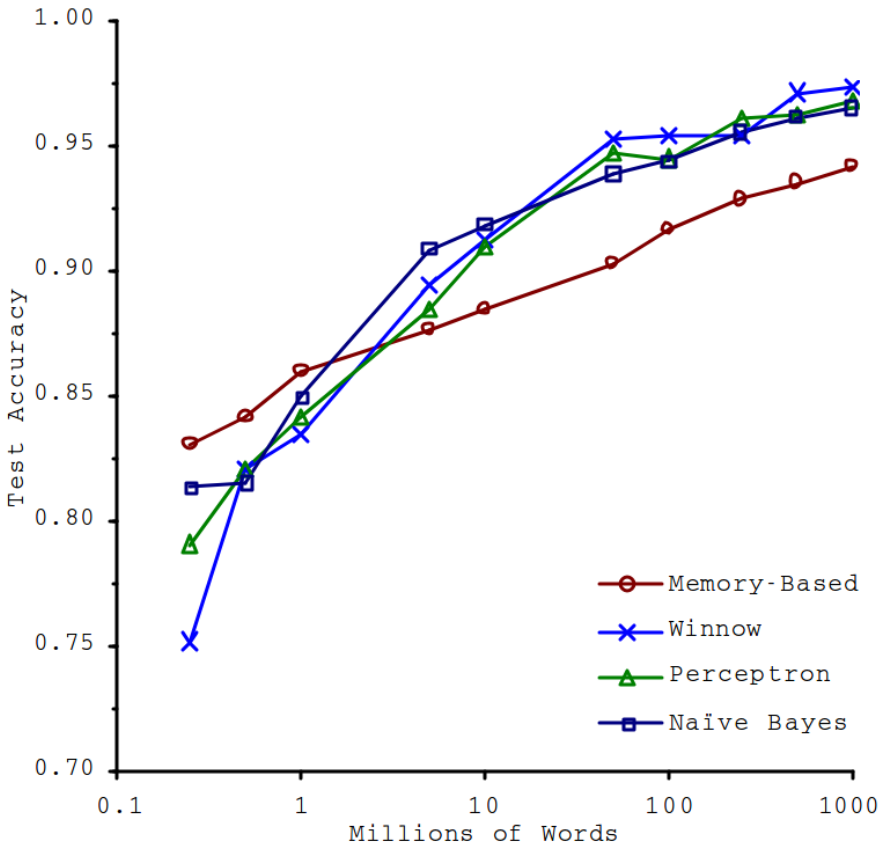

Velocity


Variance

# BigData ohne Big Data?

Wir haben so wenige, so gut strukturierte Daten, dass wir kein Einsatzszenario für BigData entwickeln können, bei dem die Kosten/Nutzen-Rechnung stimmt.

# machine learning




Eine Studie von Microsoft Research kommt bei einem Vergleich verschiedener Klassifizierungsalgorithmen zu dem Ergebnis, das die Tests erst ab ca. 50 Millionen Datensätzen eine befriedigende Genauigkeit aufweisen.

[Michele Banko and Eric Brill, Microsoft Research: Scaling to Very Very Large Corpora for Natural Language Disambiguation]

# anonymisierte Logdaten aus dem Discovery–Portal

## 1. Auszug relevanter Zeilen mit klassischen Methoden

So werden aus insgesamt 158’203’686 Zeilen aus 617 Logfiles 6’405’616 relevante Zeilen extrahiert:

```
awk 'match($7,/(\/.+)+\/search\.do.+freeText/) {sub(/\[/,"NEBIS\t",$4);sub(/(\/.+)+\/search\.do/,"",$7);print
$1"\t"$4"\t"$7}' access_log.2016* > suchanfragen.log
```

# anonymisierte Logdaten aus dem Discovery–Portal

## 2. Visualisierung mit DataScience–Methoden

Mit der Sprache Python werden die Daten in ein Dataframe geladen und können so in beliebigen Ausschnitten gesichtet werden, um anhand von Stichproben die weiteren Bearbeitungsschritte festzulegen...

| | IP | country | Date | FE | url | query_type | query |
|---|---|---|---|---|---|---|---|
| 0 | 3644786527210799 | Germany | 01/Jan /2016:01:07:13 | NEBIS | ?srt=title&srtChange=true&&dscnt=0&vl(21516851... | Basic | ['Verbalphrase'] |
| 1 | 3640797120747109 | Switzerland | 01/Jan /2016:01:10:20 | NEBIS | ?fn=search&ct=search&vl(freeText0)=kadhija&sea... | [] | ['kadhija'] |

# anonymisierte Logdaten aus dem Discovery–Portal

## 2. Auswertung mit DataScience–Methoden

...oder erste Analysen zu machen:

```
len(logframe[logframe['wcount'] == 1 ].index) * 100 / len(logframe.index),
len(logframe[(logframe['wcount'] == 1) & (logframe['query_type'] == "Advanc
len(logframe[logframe['query_type'] == "Advanced"].index) * 100 / len(logfr
```

96.3 Prozent aller einfachen Suchanfragen beinhalten nur ein Suchwort.
62.7 Prozent aller erweiterten Suchanfragen beinhalten ebenfalls nur ein Suchwort.
8.4 Prozent aller Suchanfragen werden über die erweiterte Suche durchgeführt.

Die korrekten Zahlen:
23.86981985807454 Prozent aller einfachen Suchanfragen beinhalten nur ein Suchwort.
26.29683387412833 Prozent aller erweiterten Suchanfragen beinhalten ebenfalls nur ein Suchwort.
8.357088529815087 Prozent aller Suchanfragen werden über die erweiterte Suche durchgeführt.
Die Auswertung dauerte 1.91 Sekunden.

# DataScience für alles?

Bsp. Benutzerdatensplitting: Automatisiertes Splitten von 100 Datensätzen, die aus zwei Elementen bestehen, in Vor- und Nachname:

| | | |
|---|---|---|
| mit Python und nltk: | 16.73 | Sekunden |
| mit Python ohne nltk: | 0.69 | Sekunden |
| mit Ruby: | 0.39 | Sekunden |

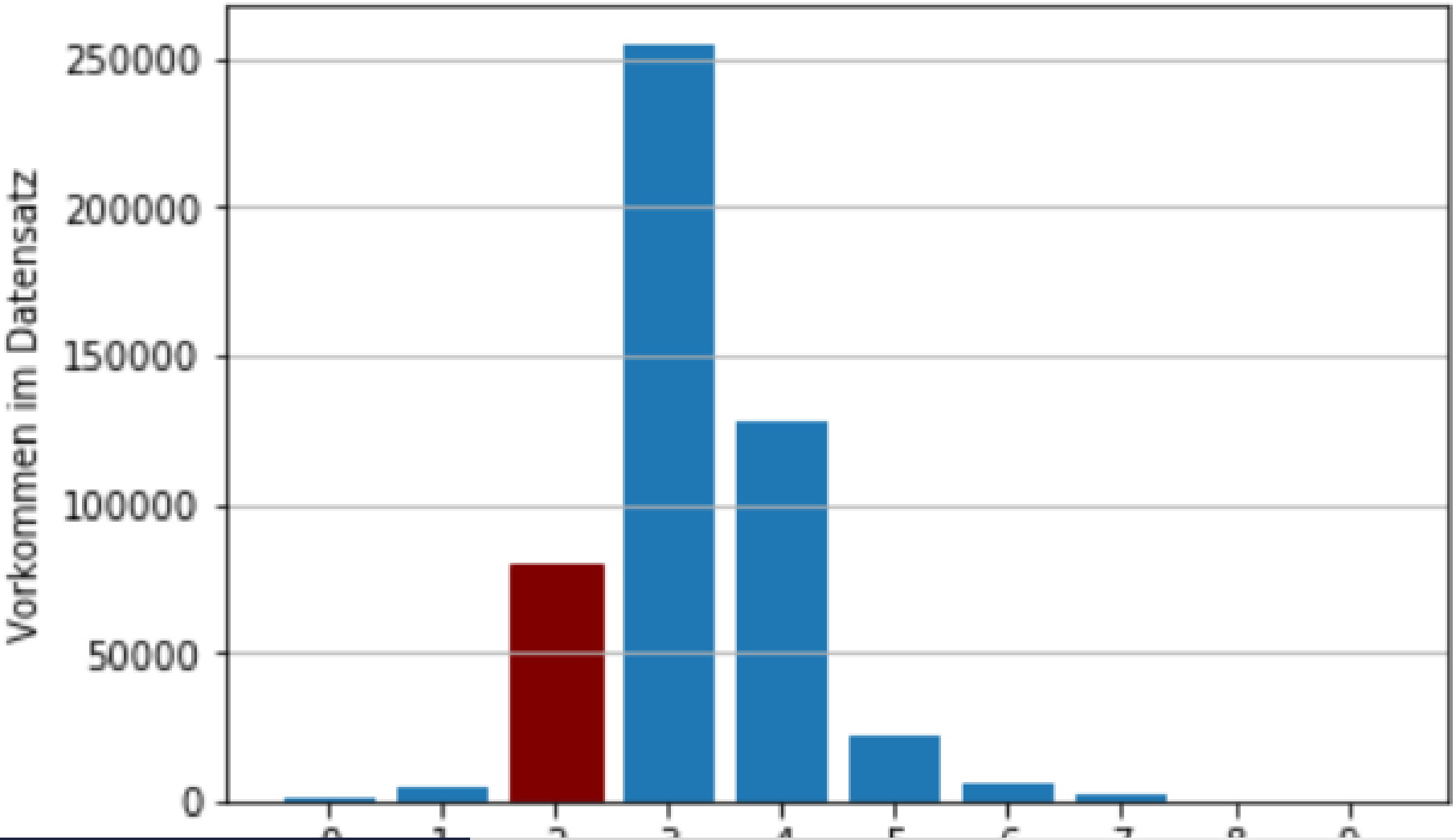
Das wirkt sich bei 650‘000 Datensätzen schon deutlich aus.

# DataScience–Methoden zum Sichten

Auch bei „wenigen” Daten können Methoden aus der DataScience genutzt werden, um sich schnell einen Überblick zu verschaffen und danach das weitere Vorgehen aufgrund von Fakten besser planen zu können.

# DataScience–Methoden zum Sichten

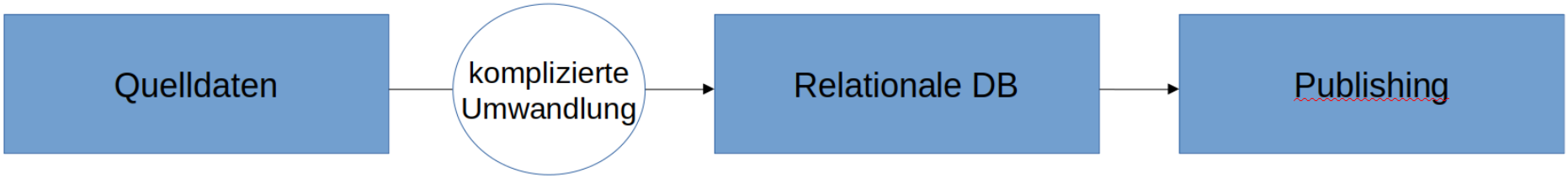
Wieder am Beispiel Benutzerdatensplitting: Visualisierung der Ergebnisse einer maschinellen Bearbeitung:



# BigData Technologien für small data:

## Alter Ansatz mit relationaler DB:

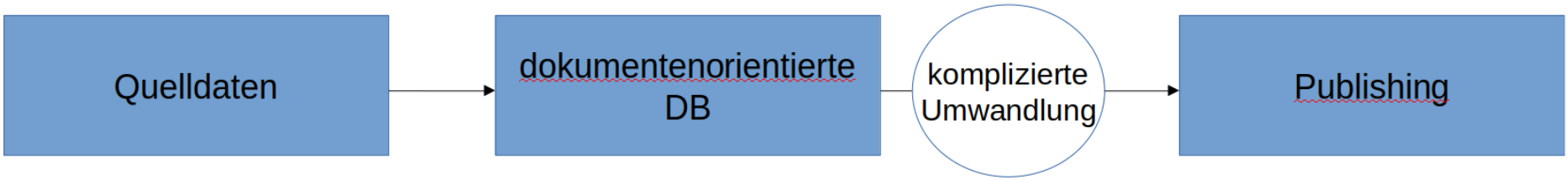
Aufbau eines OAI–Servers für die Daten der graphischen Sammlung: 23‘275 Datensätze, annähernd
$\mathrm{OAI}_D C$, *VernderungindenletztenMonatennichtmaleinDatensatz/Woche*
**gestern:**
Alle Quelldaten werden bereinigt und konvertiert, so dass sie beim Publishen direkt aus der DB gezogen und korrekt ausgegeben werden können.



# BigData Technologien für small data:

## Neuer Ansatz mit dokumentenorientierter DB:

**heute:**
Alle Quelldaten werden unverändert in eine dokumentenorientierte DB, die gut skalierbar ist, geladen. Die Umwandlung passiert erst beim Publishing.



# BigData Technologien für small data:

## Neuer Ansatz mit dokumentenorientierter DB:

**morgen:**
Dadurch dass die Umwandlung erst beim Publishing passiert und somit alle Daten komplett und unverändert in der DB vorhanden sind, lassen sich aus den Quelldaten und Verknüpfungen mit weiteren Daten umfangreiche Auswertungen ziehen.