



# Potential, Methods and Limitations of Text and Data Mining

Prof Dr Albert Weichselbraun



# Agenda

## Definition

- Text mining
- Data mining

## Potential & applications

## Selected methods & their limitations

- Keyword analysis
- Classification

## Conclusions

## Definition

**Text Mining** – derive useful information from textual resources such as Web pages, media articles, document archives, etc.

**Data Mining** – derive useful information from huge data repositories such as databases, sensor data, etc.  
(Chakrabarti et al, 2006)

# Definition

## **Automatically identify patterns in text and data**

- Deduce relevant relations

## **Rely on methods from the fields of**

- Artificial intelligence & machine learning
- Statistics
- Pattern analysis

## **Knowledge poor**

- Compensate missing context with “big data”
- Newer developments: integrate background knowledge from linked open data repositories and ontologies
- Does not require annotations (in contrast to the Semantic Web)

## **High throughput, high performance**

# Potential & applications

## **“Big data” ready**

- Analyze huge amounts of data
- Data-driven decision making (Web and business intelligence)

## **Complements human input**

- Low cost
- Predictable, well described methods
- May require human input for correct interpretation → data scientist

## **Applications**

- Discover “interesting patterns” in large data sets  
→ fraud detection, high-value customers, trends, popular research topics
- Marketing, sales & product development  
→ targeted marketing (product recommendations, clustering), opinion mining
- Semantic integration of heterogeneous databases  
→ genomic and proteomic databases, bibliographies

# Potential & applications

## **SYMPHONY project**

Monitor the publication behavior of Swiss scientists

Qualitative interviews have collected requirements from major stakeholders

- Government
- Funding agencies
- Research institutions
- Libraries

Scenarios

- Text mining to automatically collect bibliographic data
- Data mining to semantically integrate data from heterogeneous sources

Outcome of the requirement analysis and developed scenarios

→ Workshop in Bern

# Keyword analysis

Detects concepts which co-occur statistically significantly more often with the target concept.

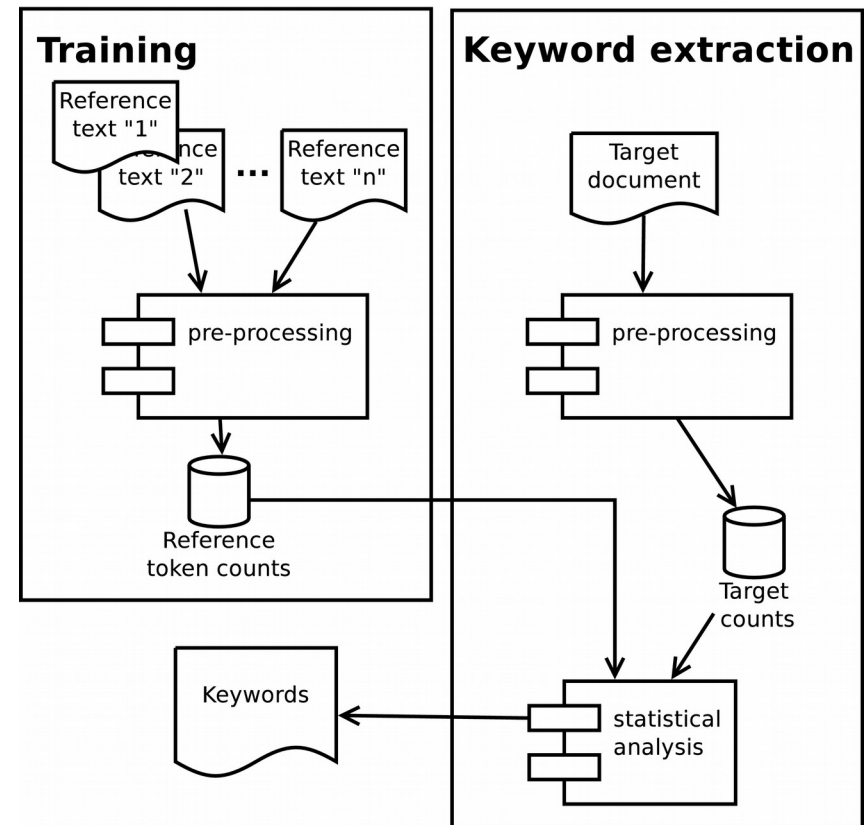
## **Applications:**

- Automatic extraction of document keywords, descriptors
- Topic detection
- Automatic tag clouds
- Search engine optimization

# Keyword analysis | Method

## Process:

- Compile a reference corpus (typical documents for that domain)
- Pre-processing
  - tokenization
  - optional word stemming and phrase detection
- Obtain reference counts
- Perform pre-processing for target documents
- compute keywords



## Keyword analysis | Method

**Target corpus**  
News articles on  
George W. Bush  
20'000 tokens

**Reference corpus**  
All news articles  
80'000 tokens

|             | the  | Iraq | war | attributed | Washington | $\Sigma$ |
|-------------|------|------|-----|------------|------------|----------|
| Ref. corpus | 4000 | 24   | 80  | 240        | 60         | 80,000   |

## Keyword analysis | Method

**Target corpus**  
News articles on  
George W. Bush  
20'000 tokens

**Reference corpus**  
All news articles  
80'000 tokens

|                               | the  | Iraq | war | attributed | Washington | $\Sigma$ |
|-------------------------------|------|------|-----|------------|------------|----------|
| Ref. corpus                   | 4000 | 24   | 80  | 240        | 60         | 80,000   |
| Target corpus<br>expected (E) | 1000 | 6    | 20  | 60         | 15         | 20,000   |

# Keyword analysis | Method

**Target corpus**  
News articles on  
George W. Bush  
20'000 tokens

**Reference corpus**  
All news articles  
80'000 tokens

|                               | the  | Iraq | war  | attributed | Washington | $\Sigma$ |
|-------------------------------|------|------|------|------------|------------|----------|
| Ref. corpus                   | 4000 | 24   | 80   | 240        | 60         | 80,000   |
| Target corpus<br>expected (E) | 1000 | 6    | 20   | 60         | 15         | 20,000   |
| Target corpus<br>found (F)    | 1172 | 19   | 48   | 53         | 47         | 20,000   |
|                               |      |      |      |            |            |          |
| $(F - E) / E$                 | 0.17 | 2.17 | 1.40 | -0.12      | 2.13       | 0.00     |

## Keyword analysis | Application

abgabe abstimmungskampf asylpolitik bakom bigler bund  
bundesgerichts bürgerinnen candinas delegierten einreichung  
empfangsgebühren fairness fernsehgebühren fernsehstationen gasche  
gebühr gebühren gegenleistung gesetz gesetzesänderung  
gewerbeverband gewerbeverbandes giacobbo hess initianten  
kanäle lästigen mainstream mediensteuer mehrwertsteuer parolen  
parteiprogramm privaten privatradios präimplantationsdiagnostik  
public radio referendum revision rtvg schawinski schwerpunkte  
service sgv srf Srg steuerhoheit stipendieninitiative substanzielle  
svp systemwechsel tagesschau tritt umstieg unentschlossen  
unterschriften urteils verbände verfassungsänderung  
verhüllungsverbot volksbegehren vorarlberg vorlage

# Classification

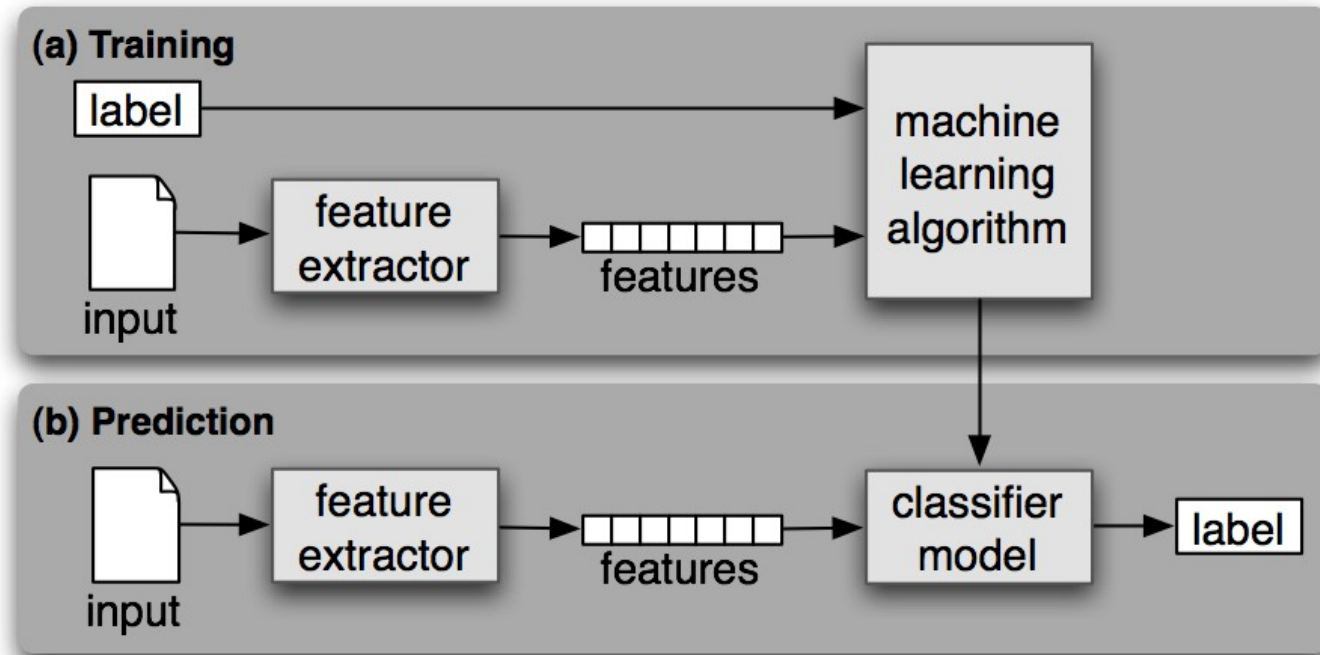
Assign items to target categories or classes based on the features available in text documents or data sets.

→ features: data relevant to the classification problem  
(e.g. words, properties, ...)

## Applications

- Automatically assign subject categories to news articles and images (IPTC taxonomy)
- Assign grant applications to referees and topics
- Classify patient reports in health-care organizations
- Classify customers according to shopping behavior
- Spam filtering

## Classification | Method



|      | dear | best | regards | and  | offer | lottery | win | viagra |
|------|------|------|---------|------|-------|---------|-----|--------|
| Ham  | 530  | 750  | 289     | 3201 | 150   | 1       | 70  | 0      |
| Spam | 15   | 312  | 31      | 3102 | 210   | 397     | 149 | 301    |

# Classification | Swiss Radar Media Criticism

## Gemeindepräsident in U-Haft – Rücktritt vom Amt (20 Minuten)

### **Grenzt an Rufmord**

Auch in der St. Galler Justizszene, wo der Fall Gesprächsthema Nummer eins ist, tappt man im Dunkeln. Es müsse sich aber um «etwas sehr Gravierendes» handeln, schreibt **am Freitag das St. Galler Tagblatt unter Berufung auf die Justizszene**. Dass die Untersuchungshaft so lange dauert, sei ungewöhnlich, sagen mehrere Anwälte gegenüber der Zeitung. Ungewöhnlich vor allem deshalb, weil es sich bei Zellweger um eine Person des öffentlichen Interesses handle. Sollte die Staatsanwaltschaft nicht «etwas sehr Gravierendes» gegen Zellweger in der Hand haben, grenze eine derart lange Untersuchungshaft ohne jede Information für die Öffentlichkeit an Rufmord.

## So sieht der Traumpartner der Schweizer aus – (nachrichten.ch)

### **Bilder in den Medien prägen unser Schönheitsideal**

«Durch die starke Medienpräsenz werden wir täglich mit einer Bilderflut schöner Menschen konfrontiert», stellt Diplom-Psychologin Lisa Fischbach von ElitePartner.ch fest. «Das beeinflusst unser Bewusstsein und prägt ein stereotypes Schönheitsideal, das uns intolerant für Individualität macht. Blonde Frauen sind nach wie vor das Traumbild vieler Männer. Doch anstatt das Beuteschema nur auf einen Typ zu beschränken oder selber einem Ideal entsprechen zu wollen, sollte man sich bewusst von diesen Einflüssen befreien und den Blick für den Menschen öffnen. Denn wirkliche Schönheit und positive Ausstrahlung kommen von innen.»

## Barbara Schmid-Federer: Die mustergültige Wahlkämpferin (Limmattaler Zeitung)

Ins Auge stach etwa Schmid-Federers Plakat mit der Aufschrift «Barbara is Bundeshus». Es zeigte unüblicherweise zwei Zeichnungen, eine von Männedorf und eine vom Bundeshaus. Die Gratiszeitung «20 Minuten» fand das Plakat so originell, dass sie es im redaktionellen Teil abdruckte. Schmid-Federer und ihr Mann – das Ehepaar führt eine Apotheke in Zürich – sorgten auch dafür, dass in verschiedenen Apotheken im Kanton Kleinplakate auflagen.

### **Stillzonen lanciert**

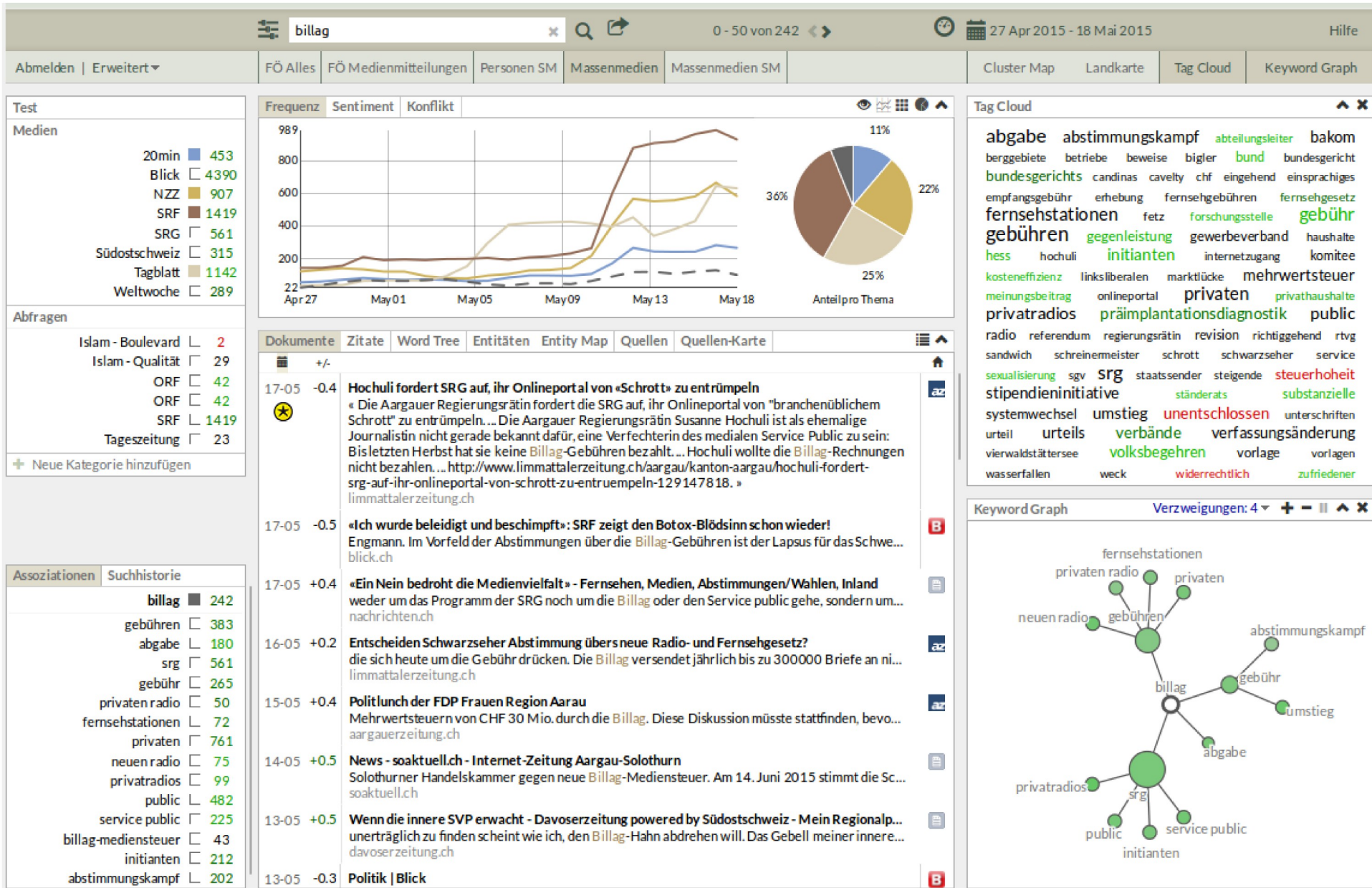
Schmid-Federer erwies sich zudem als reaktionsschnell, als es darum ging, ein emotional aufgeladenes Thema auszuschlachten. Wiederum die erwähnte Gratiszeitung brachte eine Geschichte über eine stillende Frau, die aus einem Restaurant gewiesen wurde, was für Empörung sorgte. Flugs lancierte Schmid-Federer sogenannte «Stillzonen» in Apotheken, eine Idee, die erneut in Gratismedien Echo fand.

## Die meisten Eltern geben sich die Schuld Am Tod ihres Kindes (zsz.ch)

### **Verunglückt ein Kind tödlich, berichten in der Regel die Medien darüber. Ist diese Berichterstattung gerechtfertigt, wenn man bedenkt, wie schwer das Ganze für die Eltern ohnehin schon ist?**

Als Journalistin und Mutter, die selber ein Kind verloren hat, ist diese Frage nicht einfach zu beantworten. Den meisten Betroffenen wäre es am liebsten, wenn nicht über den Tod ihres Kindes berichtet würde. Grundsätzlich aber ist es ein Recht der Öffentlichkeit zu erfahren, was passiert ist, wenn sie vom Unfall oder seinen Folgen betroffen ist – etwa durch abgesperrte Strassen oder Züge, die ausfallen.

# Swiss Radar Media Criticism



# Conclusions

## Text and data mining

- Clearly defined methods, high throughput, low cost
- Replace or complement human efforts
- Knowledge poor
- Future developments: combination with external knowledge sources (linked data source, ontologies, Semantic Web)



**HTW** Chur

Hochschule für Technik und Wirtschaft  
University of Applied Sciences

**web****Lyzard**  
technology

# Thank you for your attention!



**Platform Description**  
[www.weblyzard.com](http://www.weblyzard.com)



**Demo Portals**  
[www.ecoresearch.net/climate](http://www.ecoresearch.net/climate)  
[medienkritik.semanticlab.net](http://medienkritik.semanticlab.net)



**Contact**  
[Albert.Weichselbraun@htwchur.ch](mailto:Albert.Weichselbraun@htwchur.ch)

## Contact

University of Applied Sciences Chur (HTW)  
Pulvermühlestrasse 57, 7000 Chur