# Big Data bei unstrukturierten Daten

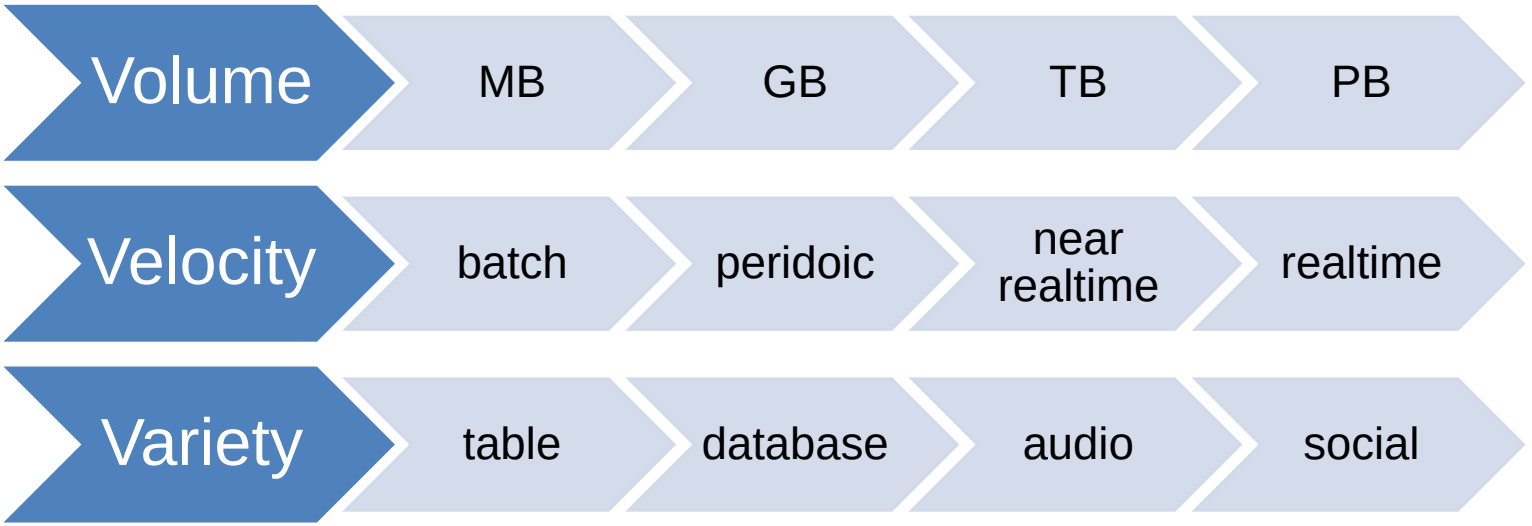
AW1 Vortrag

Sebastian Krome

# Agenda

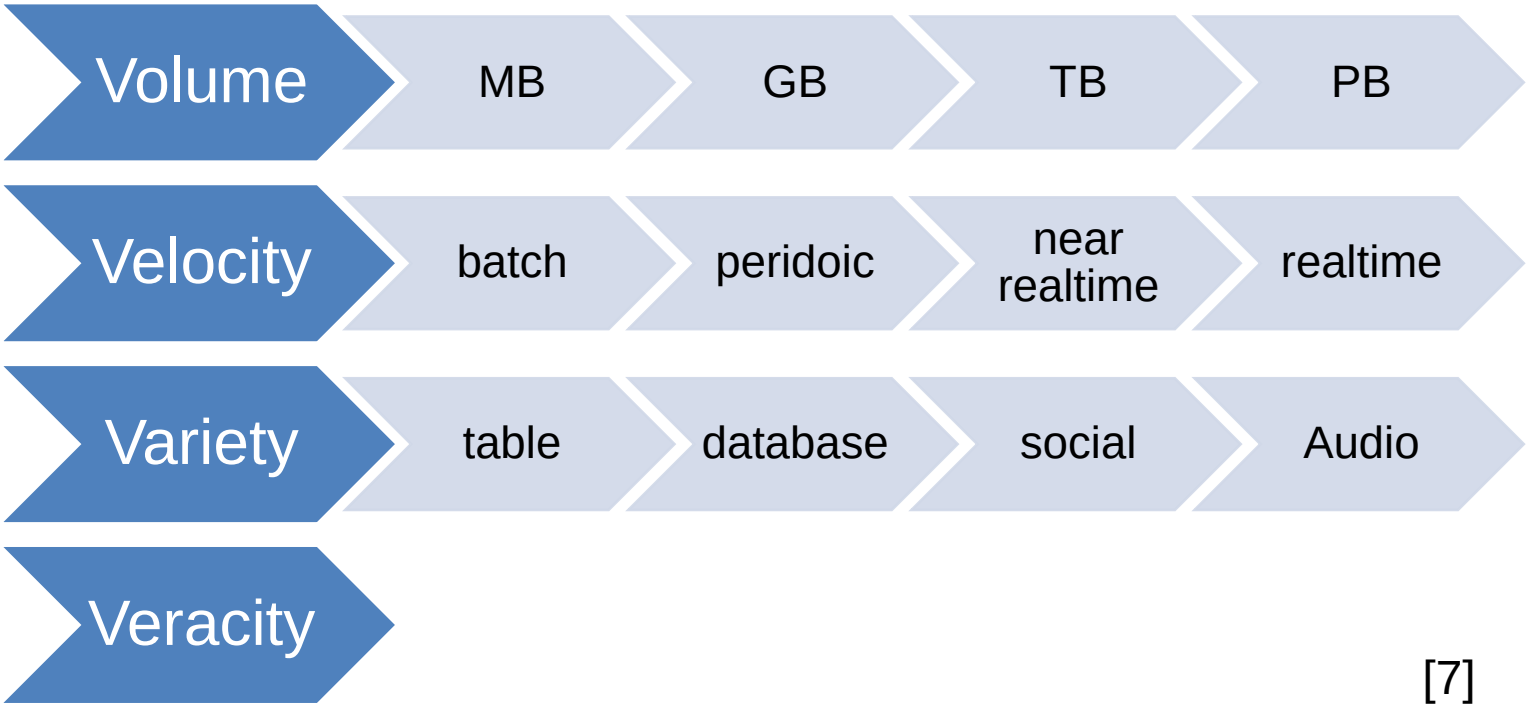
- Wiederholung – Aspekte von Big Data
- Datenverarbeitungsprozess
- TextMining
- Aktuelle Paper
  - Identification of Live News Events Using Twitter
  - Open Domain Event Extraction From Twitter
- Ausblick
- Konferenzen

# Wiederholung – Big Data



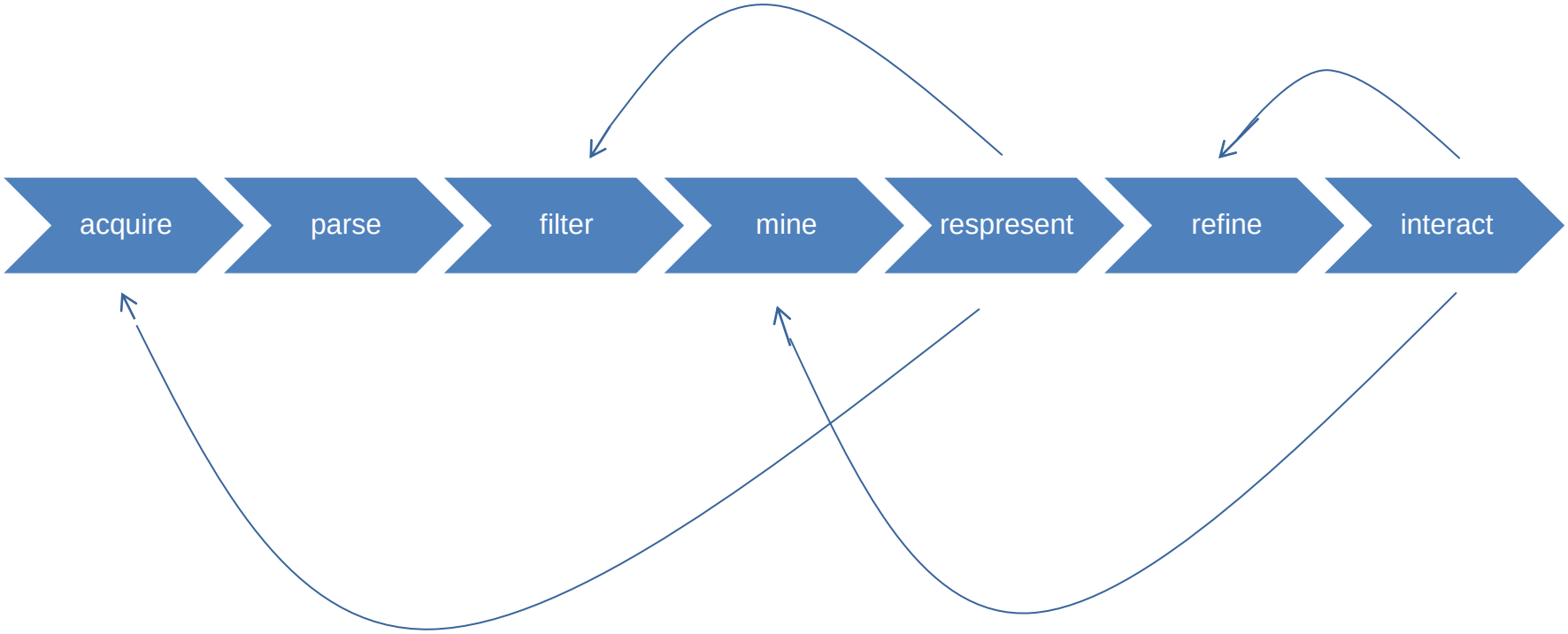

[7]

# Wiederholung – Big Data




[7]

# Datenverarbeitungsprozesse




angeleht an [5]

# Text Mining

- In Text enthaltenes Wissen extrahieren
- Unter Nutzung von
  - Statistischen Verfahren
  - Musterbasierten Verfahren

- Anwendungen
  - Klassifikation
  - Clustering
  - Information Extraction

# Text Mining und Twitter

- Herausforderungen
  - Redundante Informationen
  - Uninteressante Informationen (Alltägliches)
  - Umgangssprache, Rechtschreibung
  - Fehlen von Kontext
  - Zuverlässigkeit/Glaubwürdigkeit
- Chancen
  - Kurze Tweets
  - Keine komplexe Struktur

# Identification of Live News Events Using Twitter [4], Jackoway et al.

- Fragestellungen
  - Welche Tweets sind für die Masse interessant?
  - Welche Tweets betreffen ein reales Ereignis?
  - Wie Zuverlässig sind die Informationen?
- Verbindung von traditionellen Zeitungsartikeln und Twitter
  - Zeitung → Extrahieren von Events
  - Twitter → live-updates zu Events

# Identification of Live News Events Using Twitter [4], Jackoway et al.

- Basiert auf NewsStand
  - Zukünftige Ereignisse in Zeitungsartikeln finden
  - Extrahieren von Keywords zu Ereignissen
    → charakteristische Merkmale
- Live-Updates durch Tweets
  - Input: Tweets, die min. 1 Keyword enthalten
  1. Grobfilterung → Junk/News
  2. Tweet einem Event zuordnen oder verwerfen
     - Winkel zwischen Vektoren (Cosinus-Maß)

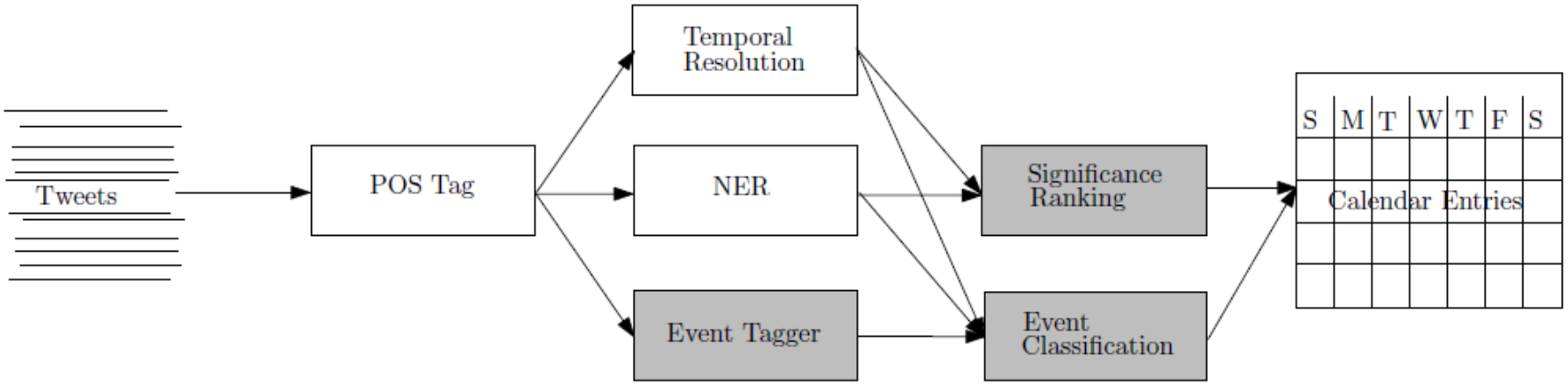
# Open Domain Event Extraction From Twitter [2], Ritter et al.

- 4-Tupel aus Tweet extrahieren
  - Named Entity
  - Event Phrase
  - Date
  - Type

| Entity | Event Phrase | Date | Type |
|---|---|---|---|
| Steve Jobs | died | 10/6/11 | DEATH |
| iPhone | announcement | 10/4/11 | PRODUCTLAUNCH |
| GOP | debate | 9/7/11 | POLITICALEVENT |
| Amanda Knox | verdict | 10/3/11 | TRIAL |

[2]

# Open Domain Event Extraction From Twitter [2 ], Ritter et al.




[2]

# Ausblick

- Zunächst Twitter
  - Intelligente Formel 1 – Tisch
  - Master Next Media
- Später?
  - Evtl. Zusammenarbeit Signal Iduna

# Konferenzen

- KDD
  International Conference on Knowledge Discovery and Data Mining
- WSDM
  International Conference on Web Search and Data Mining
- WWW
  Word Wide Web Conference
- (ISC Big Data)
- HT ACM Conference on Hypertext and Social Media

# Quellen – Twitter Textmining


- [1] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. 2012. Semantics + filtering + search = twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (HT '12). ACM, New York, NY, USA, 285-294. DOI=10.1145/2309996.2310043 http://doi.acm.org/10.1145/2309996.2310043
- [2] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '12). ACM, New York, NY, USA, 1104-1112. DOI=10.1145/2339530.2339704 http://doi.acm.org/10.1145/2339530.2339704
- [3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 851-860. DOI=10.1145/1772690.1772777 http://doi.acm.org/10.1145/1772690.1772777
- [4] Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. 2011. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (LBSN '11). ACM, New York, NY, USA, 25-32. DOI=10.1145/2063212.2063224 http://doi.acm.org/10.1145/2063212.2063224
- [5] Dimitar Robev. 2013. Visualize me. Java Magazin 10.2013
- [6] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 1524-1534.

# Quellen 2


- [7]Big Data http://www.gi.de/nc/service/informatiklexikon/detailansicht/article/big-data.html (1.12.2013)
- [8] AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, and Ba-Quy Vuong. 2009. Information extraction challenges in managing unstructured data. *SIGMOD Rec.* 37, 4 (March 2009), 14-20. DOI=10.1145/1519103.1519106 http://doi.acm.org/10.1145/1519103.1519106
- [9] Gerhard Heyer, Uw Quasthoff, Thomas Witting. Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. W3L-Verlag. 2008