Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

DATA MINING FÜR BIG DATA

Department Informatik

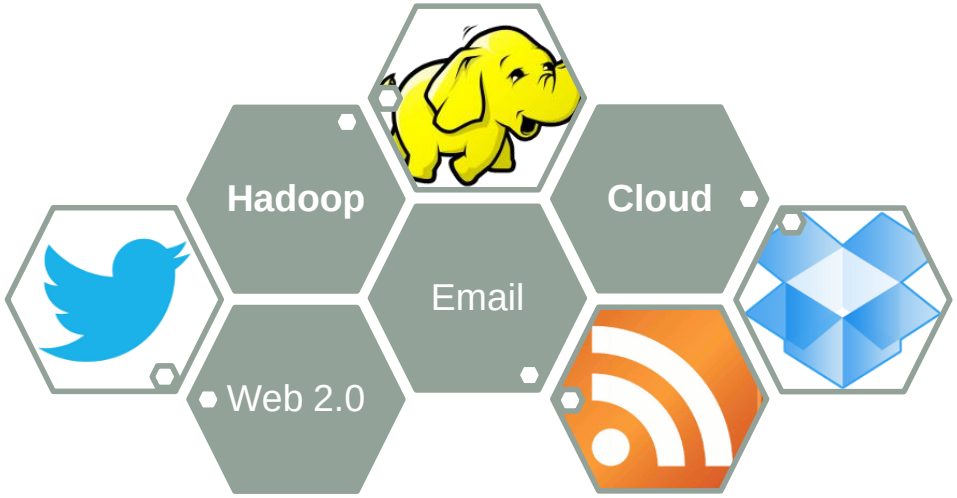
Anwendungen 1 – WiSe 2013/14

Anton Romanov

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Agenda

- Motivation
- Data Mining
  - Assoziationsanalyse
  - Clusteranalyse
- Big Data
  - Map Reduce
  - Apache Hadoop
  - Relevante Projekte
- Vision für meine MA
- Quellen



Motivation „BIG DATA“

- Neugeborenes Baby – 70 x Library of Congress *[Smo 2012]*
- Google – ~25 PB / Tag *[Vat 2012]*

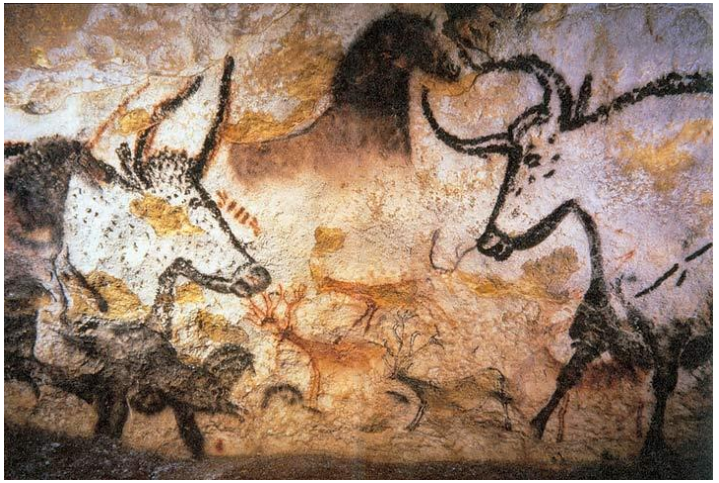


Abb1: commons.wikimedia.org/wiki/File:Lascaux_painting.jpg

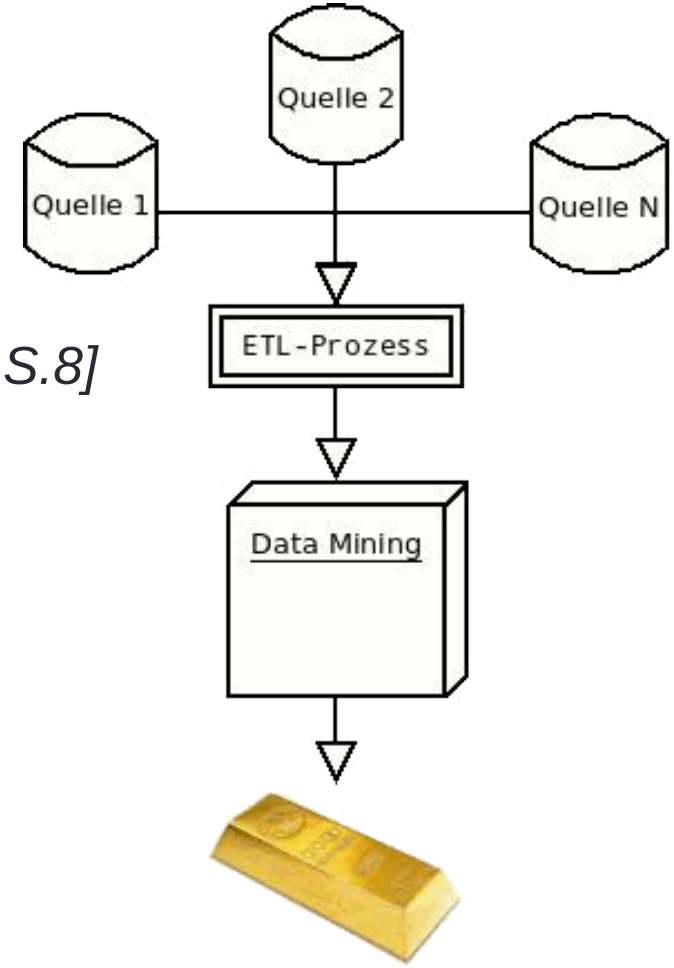
- Klimawanderung, Verkehrslage
- Zahlungsvorgänge, Simulationen

- Konventionelle SW reicht nicht aus
  - Leistungsstarke Rechner
  - Verschiedene Datentypen
  - Unterschiedliche Quellen

Motivation „Data Mining“

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

- SW-gestützte automatische Ermittlung *[Far 2011 S.103]*
  - Zusammenhänge
  - Muster
  - Trends
- Bottom-Up-Vorgehensweise *[Kno 2000 S.8]*
  - „Reverse Engineering“
  - Welches Muster pass zu diesen Daten?
- Ziele sind meistens unklar definiert
  - Im Vergleich zu OLAP



Assoziationsanalyse *[Gab 2009]*

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

- Zusammenhänge zwischen den Produkten bestimmen
  - Aussage in Form von Wenn-Dann
- Support – Häufigkeit der vorkommenden Produkten
  - Support(A → B) = $\frac{Anzahl\ der\ Transaktionen\ mit\ A\ und\ B}{Gesamtanzahl\ der\ Transaktionen}$
- Konfidenz – Stärke der Korrelation
  - Konfidenz(A → B) = $\frac{support\ (A \rightarrow B)}{support(A)}$
- Warenkorbanalyse *[Pet 2005 S.28]*
  - Was wird zusammen gekauft
  - Warenanordnung

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Assoziationsanalyse

Einkauf	Gekaufte Artikel
E1	Käse, Bier, Pizza
E2	Bier, Käse, Popcorn
E3	Bier, Olivenöl
E4	Wasser, Käse, Bier
E5	Käse, Bier, Pizza, Popcorn

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Assoziationsanalyse

Einkauf	Gekaufte Artikel
E1	Käse, Bier, Pizza
E2	Bier, Käse, Popcorn
E3	Bier, Olivenöl
E4	Wasser, Käse, Bier
E5	Käse, Bier, Pizza, Popcorn

Artikel	Enthalten in
Käse	E1, E2, E4, E5
Bier	E1, E2, E3, E4, E5
Pizza	E1, E5
Popcorn	E2, E5
Olivenöl	E3
Wasser	E4

Assoziationsanalyse

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Einkauf	Gekaufte Artikel
E1	Käse, Bier, Pizza
E2	Bier, Käse, Popcorn
E3	Bier, Olivenöl
E4	Wasser, Käse, Bier
E5	Käse, Bier, Pizza, Popcorn

Artikel	Enthalten in
Käse	E1, E2, E4, E5
Bier	E1, E2, E3, E4, E5
Pizza	E1, E5
Popcorn	E2, E5
Olivenöl	E3
Wasser	E4

$$S(A \rightarrow B) = \frac{A \text{ und } B}{Gesamtanzahl}$$

$$K(A \rightarrow B) = \frac{support\ (A \rightarrow B)}{support(A)}$$

Support	Käse	Bier	Pizza	Popcorn	Olivenöl
Käse		$^4/_5$	$^2/_5$	$^2/_5$	$^0/_5$
Bier	$^4/_5$		$^2/_5$	$^2/_5$	$^1/_5$
Pizza	$^2/_5$	$^2/_5$		$^1/_5$	$^0/_5$
Popcorn	$^2/_5$	$^2/_5$	$^1/_5$		$^0/_5$
Olivenöl	$^0/_5$	$^1/_5$	$^0/_5$	$^0/_5$	

Assoziationsanalyse

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Einkauf	Gekaufte Artikel
E1	Käse, Bier, Pizza
E2	Bier, Käse, Popcorn
E3	Bier, Olivenöl
E4	Wasser, Käse, Bier
E5	Käse, Bier, Pizza, Popcorn

Artikel	Enthalten in
Käse	E1, E2, E4, E5
Bier	E1, E2, E3, E4, E5
Pizza	E1, E5
Popcorn	E2, E5
Olivenöl	E3
Wasser	E4

Regel (S.> 50%)	Konfidenz
Käse → Bier	$\frac{4}{5} / \frac{4}{5} = 1$
Bier → Käse	$\frac{4}{5} / \frac{5}{5} = \frac{4}{5}$

$$S(A \rightarrow B) = \frac{A \text{ und } B}{Gesamtanzahl}$$

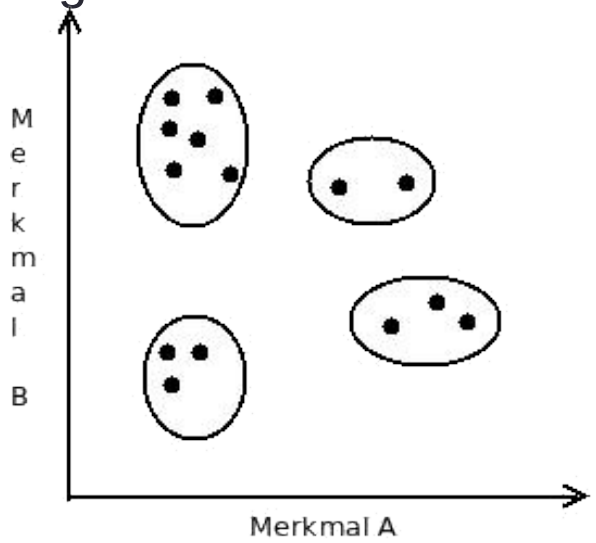
$$K(A \rightarrow B) = \frac{support\ (A \rightarrow B)}{support(A)}$$

Support	Käse	Bier	Pizza	Popcorn	Olivenöl
Käse		$^4/_5$	$^2/_5$	$^2/_5$	$^0/_5$
Bier	$^4/_5$		$^2/_5$	$^2/_5$	$^1/_5$
Pizza	$^2/_5$	$^2/_5$		$^1/_5$	$^0/_5$
Popcorn	$^2/_5$	$^2/_5$	$^1/_5$		$^0/_5$
Olivenöl	$^0/_5$	$^1/_5$	$^0/_5$	$^0/_5$	

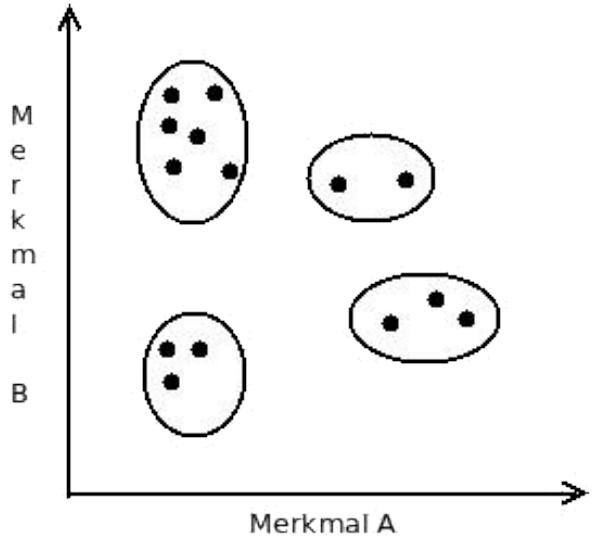
Clusteranalyse

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

- Objekte zu homogenen Klassen gruppieren *[Gab 2009 S. 156]*
- Partitionierende Verfahren
  - Vorgegebene Gruppeneinteilung
- Hierarchische Verfahren
  - Agglomerativ – Aus vielen Klassen werden weniger
  - Divisiv – Aus einer Klasse werden viele
- Zielgruppe direkt ansprechen
  - Werbung anpassen *[Pet 2005 S.25]*
  - Markt- und Kundensegmentierung



K-Means



- Partitionierendes Verfahren *[Pet 2005 S.97]*

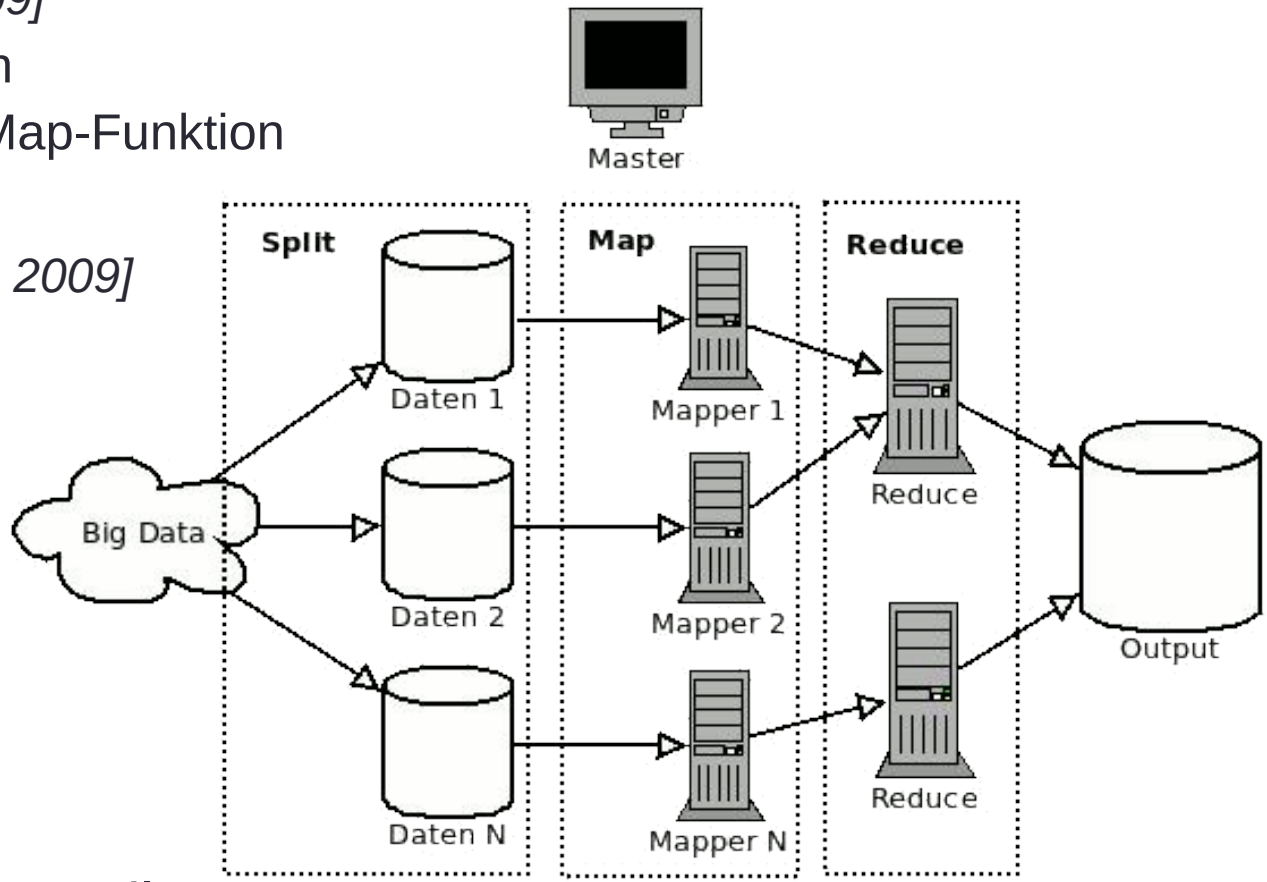
- Algorithmus *[Mac Queen 1967]*
  1. Anfangsposition vorgeben
  2. Klassenzentren berechnen
  3. Objekte verschieben
  4. Abrechnen, wenn X mal hintereinander kein Objekt verschoben wurde. Sonst gehe zu 2.

- Anfangsposition und Iterationsreihenfolge beeinflussen die Lösung

MapReduce *[Dea 2008]*

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

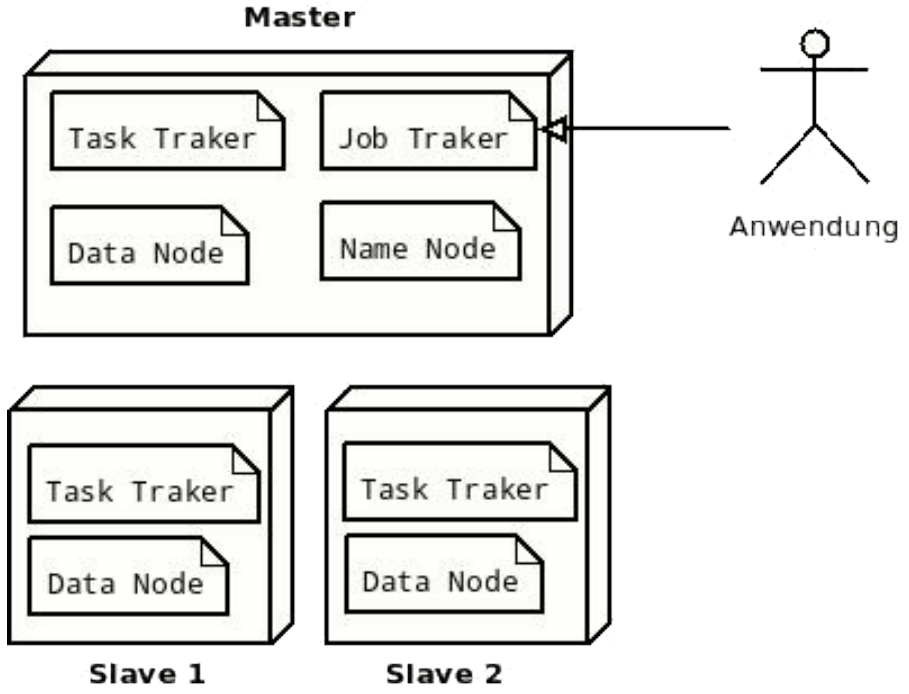
- Map-Phase *[Bro 2009]*
  - Basisberechnungen
  - Anwendung einer Map-Funktion
- Reduce-Phase *[Bro 2009]*
  - Anzahl der Werte wird reduziert
- Weitere Phasen
  - Split
  - Combine
  - …
- Performanz durch Verteilung



Apache Hadoop *[Apa 2013]*

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

- Framework
  - HDFS, MapReduce, Projects
- Slave-Nodes
  - Task Traker – Taskverwaltung
  - Data Node – Datenverwaltung eines Splits
- Master-Node
  - Task Traker, Data Node
  - Job Traker – Job in Tasks zerlegen
  - Name Node – Indiziert die Daten



Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Relevante Projekte

- Apache Hive
  - DWH-Systeme mit Hadoop
  - Arbeitet mit HDFS
- Apache Mahout
  - Verteilte / Skalierbare Algorithmen
  - Maschinelles Lernen
- Apache Pig
  - MapReduce-Programme für Hadoop
  - Pig-Latin als Abstraktion von Java
- Apache Sqoop
  - Bulk-Loader für Hadoop

Vision für MA

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

AW2	• Einarbeiten in Hadoop-Framework
PJ1	• Hadoop-Cluster einrichten
PJ1	• Datenbeschaffung
PJ2	• Data-Mining Algorithmus + Mahout
MA	• Mahout-Framework erweitern

Konferenzen

- ISC Big Data'14 in Heidelberg, Germany, October 1–2
  - Efficient & Scalable Big Data Architectures for Search & Mining, Prof. Dr. Felix Wortmann
- IEEE BigData 2013 Main Conference, Santa-Clara 6-9.10
  - Key Usage Patterns for Apache Hadoop in the Enterprise, Amr Awadallah
- Big Data World Conference, München, 3-4 Dezember ’13
  - Understanding Machine Learning and other big data components

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Zusammenfassung

- Data Mining
  - Assozitationsanalyse
    - Support
    - Konfidenz
  - Clusteranalyse
    - Partitionierende Verfahren
    - Hierarchische Verfahren
    - K-Means
- Big Data
  - MapReduce
  - Apache Hadoop
  - Hadoop Projects

Quellen I

- *[Apa 2013]*
  APACHE SOFTWARE FOUNDATION : *Apache Hadoop 2.1.0-beta Documentation*. Stand : 2013 http://hadoop.apache.org/docs/current/ Abruf: 2013-10-09
- *[Bro 2009]*
  Richard A. Brown. 2009. Hadoop at home: large-scale computing at a small college. In Proceedings of the 40th ACM technical symposium on Computer science education (SIGCSE '09). ACM, New York, NY, USA, 106-110. DOI=10.1145/1508865.1508904 http://doi.acm.org/10.1145/1508865.1508904
- *[Dea 2008]*
  DEAN, Jeffry ; GHEMAWAT, Sanjay. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004. *MapReduce: Simplifed Data Processing on Large Clusters*. Stand: 2004 http://research.google.com/archive/mapreduce-osdi04.pdf Abruf 2013-09-30
- *[Far 2011]*
  FARKISCH, Kiumas : *Data-Warehouse-Systeme kompakt : Aufbau, Architektur, Grundfunktionen.* Berlin : Springer-Verlag Berlin Heidelberg, 2011. ISBN 978-3-642-21533-9
- *[Gab 2009]*
  GABRIEL, Roland ; GLUCHOWSKI, Peter ; PASTWA, Alexander : *Data Warehouse & Data Mining*. Witten : W3L-Verlag, 2009. ISBN 3-937137-66-7

Quellen II

- *[Kno 2000]*
  KNOBLOCH, Bernd : Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. In: *Bamberger Beiträge zur Wirtschaftsinformatik №58*, Bamberg, 2000. ISSN 0937-3349. Online verfügbar unter: *www.ceushb.de/forschung/downloads/%5BKnob00%5D.pdf* Abruf: 2013-10-05
- *[MacQueen 1967]*
  MACQUEEN, J. B. ; CAM, L. M. Le (Bearb.) ; NEYMAN, J. (Bearb.): Some Methods for Classification and Analysis of MultiVariate Observations. 1. In: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* : UNIVERSITY OF CALIFORNIA PRESS, 1967, S. 281-297
- *[Pet 2005]*
  PETERSOHN, Helge : *Data Mining : Verfahren, Prozesse, Anwendungsachitektur*. München : Oldenbourg, 2005. ISBN 3-486-57715-8
- *[Smo 2012]*
  SMOLAN, Rick ; ERWITT, Jennifer : *The Human Face of Big Data*. Stand: 2012 http://thehumanfaceofbigdata.com/ Abruf 2013-09-27
- *[Vat 2012]*
  Ranga Raju Vatsavai, Auroop Ganguly, Varun Chandola, Anthony Stefanidis, Scott Klasky, and Shashi Shekhar. 2012. Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (BigSpatial '12). ACM, New York, NY, USA, 1-10. DOI=10.1145/2447481.2447482 http://doi.acm.org/10.1145/2447481.2447482