Sinnvolle Datenerhebung für Smart Studies

BETREUER: PROF.DR.MARTIN BECKE

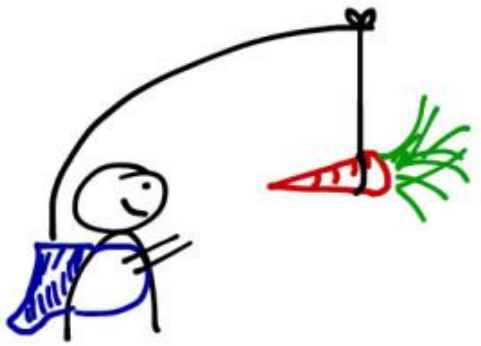
LUKAS KOZLOWSKI - GRUNDSEMINAR WS15/16

Agenda

- Motivation
- Prozesskette
- Forschungsfeld
- Komplexität
- Kontroversität
- Potenziale
- Infrastruktur
- Zielsetzung

Meine Motivation

- Technische Unterstützung der Forschungsgruppe **„Smart Studies“**
- Neues frisches Projekt, keine Restriktionen.
- Die Systemarchitekturbasis **mitentwerfen** und **gestalten**.
- Die **Forschung** und **Implementierung** von Datenerhebung wie z.B. durch Big-Data-Mining und Analytics Konzepte.



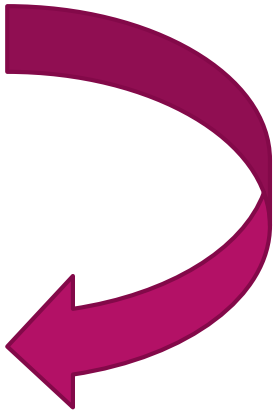
[1]

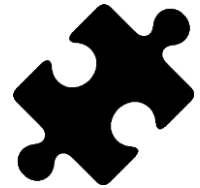
Prozesskette

Es gilt die Frage nach einer sinnvollen Infrastruktur für das individuelle Studium zu beantworten.

- Analyse vorhandener Infrastruktur
- Untersuchung neuer Architekturen

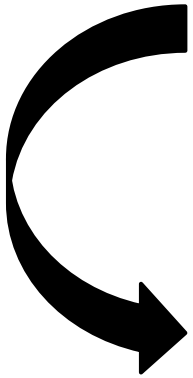
Etablierung
adaptiver Prozessstrukturen





„Wie kann eine **adaptive** **Prozessstrukturkomponente** und **analytische** **Informationsquelle** in die Smart Studies Systemarchitektur implementiert werden, um den Studenten **aktiv** zu helfen und ihn in seinem **individuellen** **Studium** unterstützen?“

Forschungsfeld



Individuelle Indexierte Internetseiten von Google

- 1998 1.000.000 (10^6 **Million**)
- 2000 1.000.000.000 (10^9 **Milliarde**)
- 2008 1.000.000.000.000 (10^{12} **Billion**)
- **2015-2016 ~ 40.000.000.000.000**

Forschungsfeld

Google [3] Suchanfragen pro Tag: 3.5 Milliarden

[4] Tweets pro Tag: 500 Millionen

[5] Nachrichten pro Tag: 64 Milliarden

[6] Updates pro Tag: 800 Millionen

flickr [7] Bilder pro Tag: 3.5 Millionen

YouTube [8] Views pro Tag: 5 Milliarden

Die Quellen erzeugen Informationen einer Größenordnung von Zettabytes täglich!

Der Durchschnittsbürger hat heute **täglich** mehr Zugriff auf Informationsdaten, als ein Individuum aus dem 16. Jahrhundert in seinem **gesamten** Leben.

Jährliche Zuwachsprognose (2015/16) ~50 %

1 Zettabyte = 10^{21} **=** 1.000.000.000.000.000.000.000 Bytes

= 1000 Exabytes **=** 1Million Petabytes **=** 1 Milliarde Terrabytes **=** 1 Billion Gigabytes

90 % der Daten sind in den letzten 2 Jahren entstanden

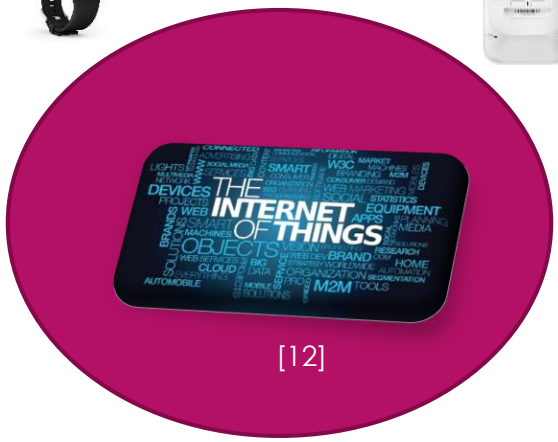
(Diese Annahme basiert auf dem **M**ooreschen **G**esetz welches für die letzten 30 Jahre korrekt war)

Forschungsfeld

[10]

[11]

[9]



[12]

[13]

<u>Strukturierte</u> oder **<u>unstrukturierte</u>** Datensätze
die mit der heutigen **Methodologie**
oder durch **Data-Mining-Software-Lösungen** <u>nicht förderbar</u>
sind. Bedingt durch die **Größe** der **Datensätze**
oder die **Größe** der **Datenströme**.

Forschungsfeld

[9]

Unstrukturierte Daten

Daten die numerisch **nicht kategorisiert** und analysiert werden können.

Kundenreviews
Photos und Multimedia
Social Network Kommentare

Daten die **vielfältige** Informationen enthalten.

Strukturierte Daten

Zahlen oder Wörter die **leicht kategorisiert** oder analysiert werden können

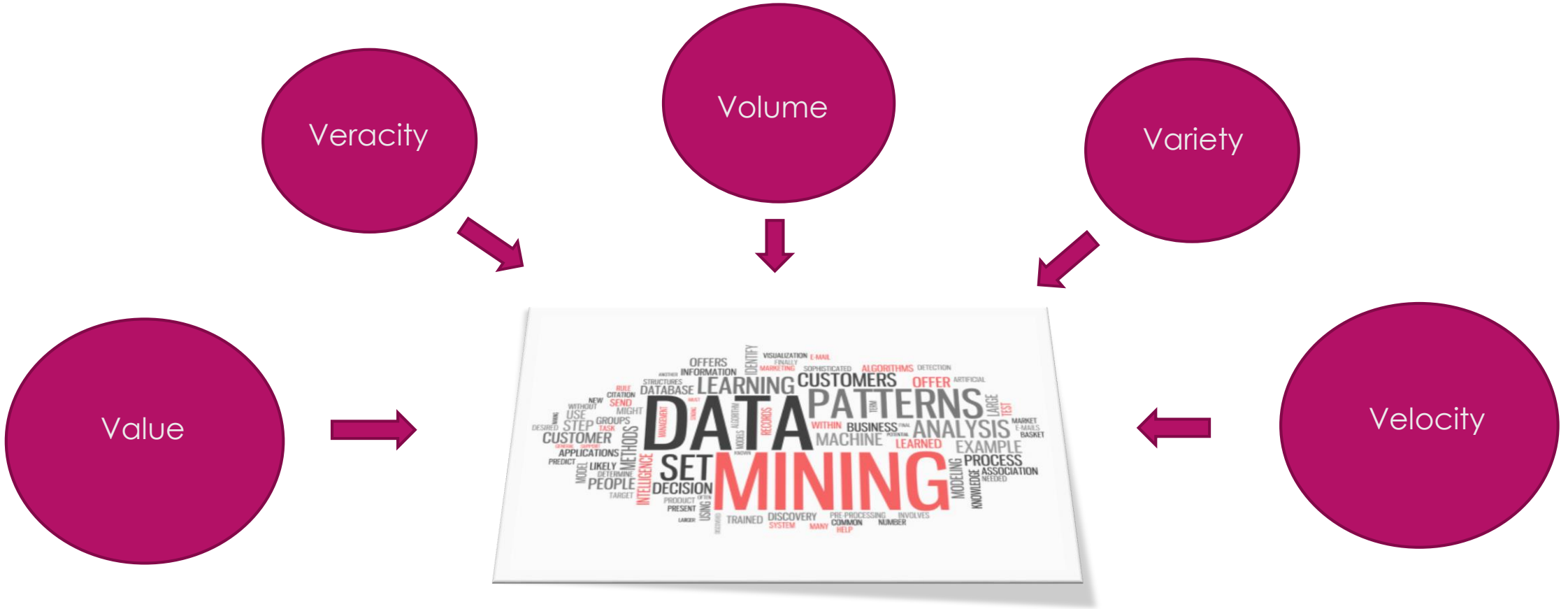
Netzwerk Sensoren
Smartphones
Transaktionsdaten

Hace - Theorem

- **H**eterogeneous, **A**utonomous, **C**omplex, **E**volving
- Big data starts with large volume, **heterogeneous**, **autonomous** sources with distributed and decentralized control, and seeks to explore **complex** and **evolving** relationships among data.

Dieses Theorem wird zur **Modellierung der Charakteristik** von **Big-Data** verwendet.

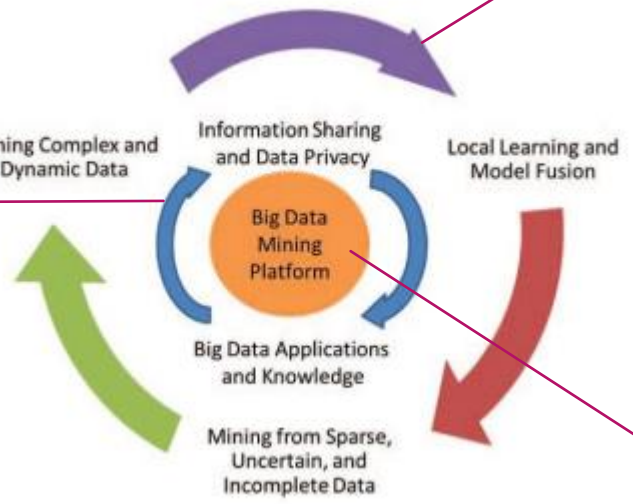
V - Charakteristik



Komplexität

3 Tier

Big Data Mining - Algorithmen



[15]

2 Tier

Data Isolation
&
Domain Knowledge

1 Tier

Access & Computing
parallele Berechnungen
zB. MapReduce auf n Clustern

- Die **Handhabung** eines riesigen Datenvolumens.
- Die **Komplexität** brauchbarer Informationen aus riesigen Datensets oder Datenströmen zu definieren und gezielt zu fördern.
- Die **Forschung** mit neuen Mining-Techniken.
- Der Umgang mit **Datenschutz** und **Kontroversität**

Kontroversität

- Größer bedeutet nicht Besser.

  Es kommt drauf an ob die Daten „**noisy**“ (**Korrupt/Unbrauchbar**) sind und sie das repräsentieren was wir suchen.

- Ethische Besorgnisse

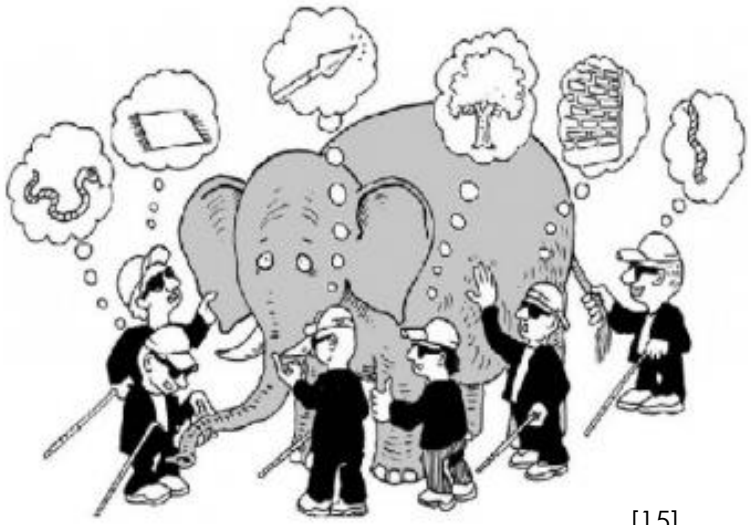
  Ist es ethisch vertretbar wenn Menschen analysiert werden ohne es zu wissen?

- Echtzeit Analysen

  Daten verändern sich daher ist nicht die Größe der Daten wichtig sondern das Alter der Daten.

- Annahmen auf Genauigkeit

  Mit der Variablenanzahl steigt auch die Anzahl falscher **Korrelationen** (**Zusammenhänge**).



[15]

Blind Man - Problematik

Potenziale

- Prozesslaufzeit

Verkürzung der Laufzeit
von Stunden auf Sekunden

- DNA-Mining

Vorhersagen/Trends in
Krankheitsausbreitung/Seuchen/Epidemien

Durch gezielte Big-Data Auswertungen
werden bestimmte
Krankheitsausbreitungen verhindert
andere können früher erkannt werden.

- Economic Development

Statistiken Interpretieren und
Ökonomische Entscheidungen treffen.

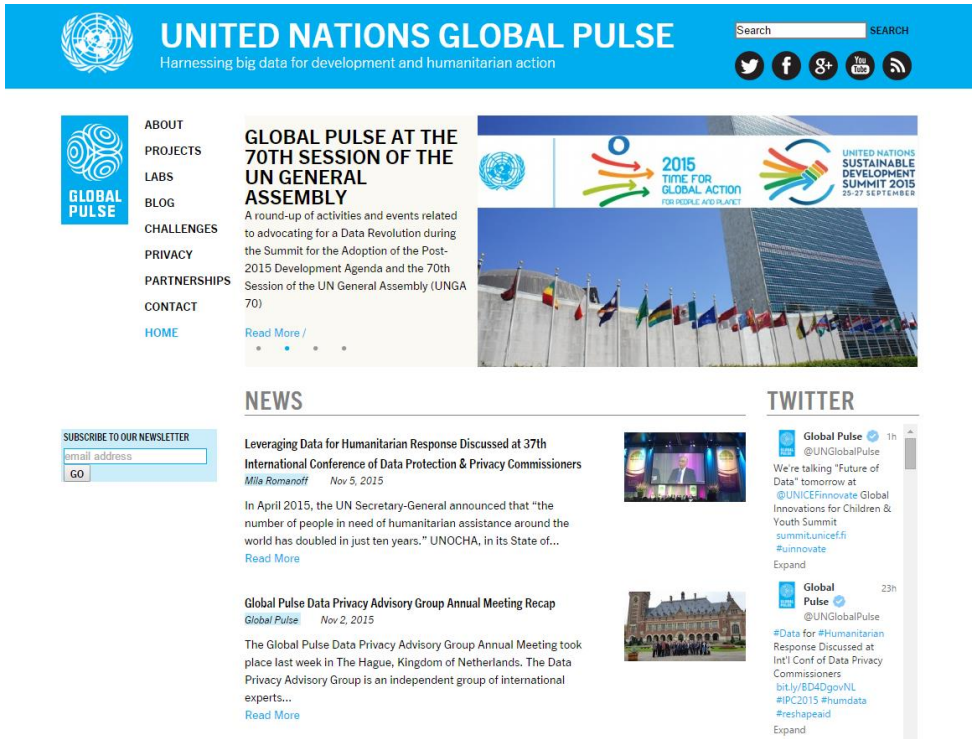
- High Quality of Live

Optimierter Verkehrsfluss (Bahn, Auto usw.)
Smart Emergency Service

- Customer Personalization
- Churn Detection

Potenziale

Projekt der Vereinten Nationen seit 2009



- Early Warning

Schnelle Reaktionszeit in Krisenzeiten

- Real-time Awareness

Realitätsgenauere Entwicklungsprogramme und Richtlinien

- Real-time Feedback

Die Echtzeitüberwachung von laufenden Entwicklungsprogrammen.

Infrastrukturen, Mining und Analyse - Tools

- Apache Hadoop
- Apache S4
- Apache Storm
- Apache Spark
- Apache Mahout
- Apache SAMOA
- Dato (früher GraphLab)
- R
- MOA
- Vowal Wabbit
- Pegasus

[21]

[22]

[23]

[24]

[25]

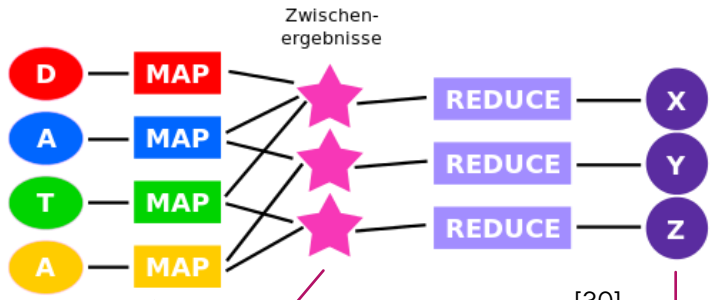
[26]

[27]

Apache Hadoop

- Apache Hadoop – Open Source Plattform für scalable, distributed computing
- Die Softwarebibliothek ist ein Framework was die parallele Berechnung (distributed processing) von großen Datensets über Computercluster ermöglicht.
- Hadoop kann für **strukturierte** und **unstrukturierte** Datensets verwendet werden.
- Hadoop verwendet **MapReduce** als Mining-Technik
- **MapReduce** ist ein Programm das Datensets in einzelne Subsets separiert.
- **Map() Funktion für die Filterung und Sortierung**
- **Reduce() Funktion für die Zusammenfassung**

[29]



[30]

„eine Rose ist eine Rose ist eine Rose“

eine Rose ist
eine Rose ist
eine Rose

eine < *1,1,1* >
Rose < *1,1,1* >
ist < *1,1* >

eine, 3
Rose, 3
ist, 2

Buzzword-Problem

BIG-DATA

BIG-DATA

BIG-DATA … *what else!?*

BIG-DATA

BIG-DATA

BIG-DATA

Zielsetzung.. Aller Anfang ist schwer...

- **Kleinere Versuche** mit MapReduce() (Wordcount usw.)
- **Tool-Testing** auf Handhabung und Möglichkeiten und Potenziale
- **Performance Analysen** für die verschiedenen Taskarten
- **Tieferes Verständnis** für Big Data und Data-Mining-Techniken entwickeln.

USE-CASE – Beispiel

Der Student durchläuft ein QUIZ, Fragebogen basierend auf dieser Erhebung, durch die Auswertung dieser Daten kann der Student in der Anmeldungsphase ein Feedback erhalten, ob die getroffen Klausurauswahl sinnvoll war oder nicht.

Konferenzen / Workshops

- KDD-2016 2016 : 22nd ACM SIGKDD international conference on knowledge discovery and data mining - San Francisco USA

  [**Konferenz**]

- BIGMINE 2015

  http://bigdata-mining.org/bigmine-15/

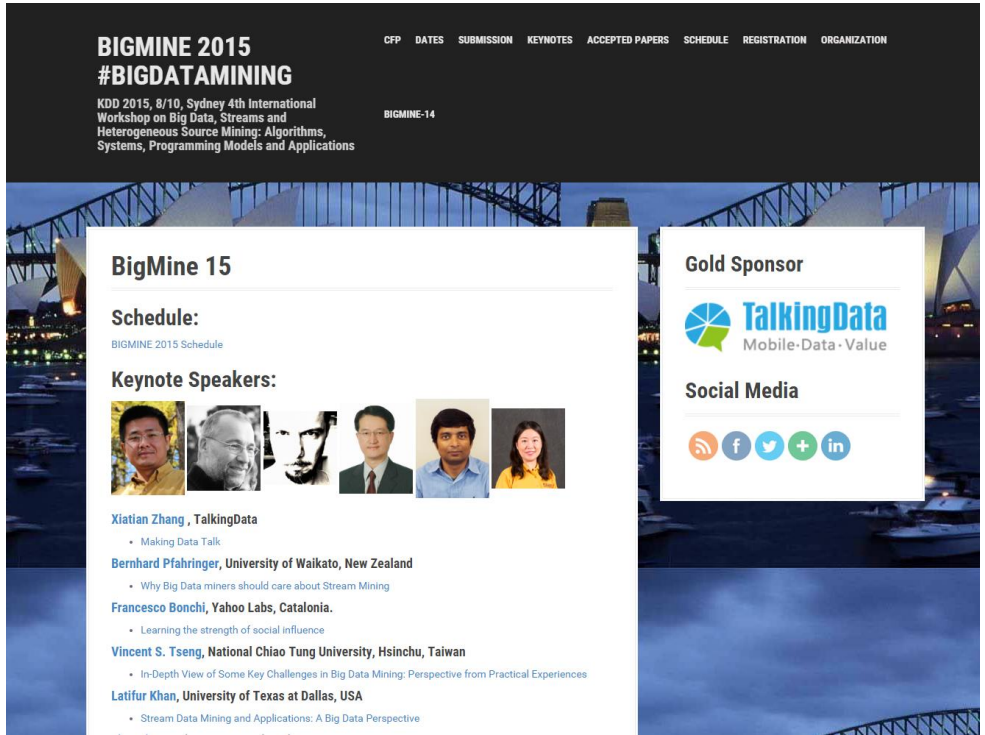
  [**Workshop**]

- Big Data Summit 2016

  http://www.bitkom-bigdata.de/

  25.Februar Congress Park Hanau

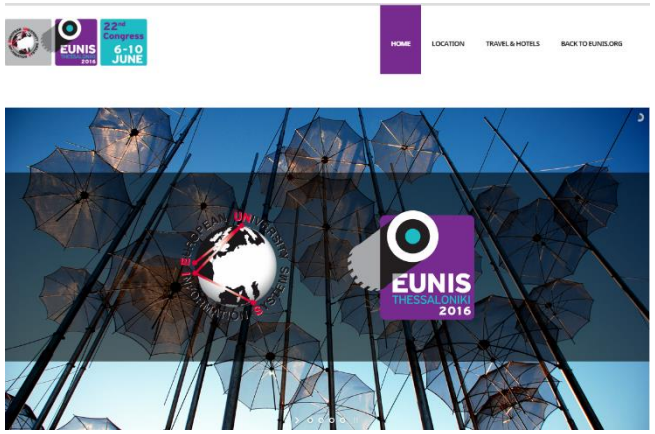
  [**Kongress**]



[32]

Konferenzen zum Thema University Information-Systems

- **EUNIS – E-learning Task Force (European University Information Systems)**



[33]

06.06.2016 – 10.06.2016 – Thessalonki - Griechenland (**Kongress**)
http://www.eunis.org/eunis2016/

14.04.2016 - 16.04.2016 Krakow – Polen (**Konferenz**)
http://www.eunis.org/calendar/e-learning-task-force-workshop

09.06.2016 Dundee – Großbritannien (**Workshop**)
http://www.eunis.org/calendar/e-learning-task-force-workshop

- **European MOOCs Stakeholders Summit 2016 (EMOOCs 2016)**

[34]

22.02.2016 – 24.02.2016 – Graz Österreich (**Konferenz**)
http://emoocs2016.eu/

Quellen

- International Journal of Advanced Research in Computer Engineering & Technology – Big Data Analysis using Hace Theorem [January 2015]
- ACM Paper - Mining Big Data: Current Status, and Forecast to the Future [2013]
- ACM Paper - Scaling Big Data Mining Infrastructure: The Twitter Experience [2013]
- ACM Paper - Mining Heterogeneous Information Networks: A Structural Analysis Approach
- ACM Paper - Big Graph Mining: Algorithms and discoveries
- ACM Paper - Mining Large Streams of User Data for Personalized Recommendations
- Buch - Software Architecture in Practice [Len Bass, Paul Clement, Rick Kazman 2013]

Bildquellen

[01] http://www.peterkleinau.com/wp-content/uploads/2014/01/Zeichnung_Motivation_M%C3%B6hre-300x237.jpg

[03] http://static1.squarespace.com/static/533c59ece4b07e844a588995/t/5342ea5ce4b0c23c72915fce/1396894308622/goog

[04] http://cdn.flaticon.com/png/256/8800.png

[05] https://danielrehn.files.wordpress.com/2014/11/iconmonstr-whatsapp-3-icon.png

[06] https://image.freepik.com/free-icon/facebook-logo_318-49940.jpg

[07] http://www.nomarket.org/wp-content/uploads/2014/12/Flickr-Logo.jpeg

[08] http://www.iconsdb.com/icons/preview/black/youtube-3-xxl.png

[09] http://images.offerpop.com/wp-content/uploads/2014/07/bigstock-Big-data-concept-in-word-tag-c-52761202-1-1000x573.jpg?505c96

[10] http://www.fullthoughtcc.com/wp-content/uploads/2014/11/Dollarphotoclub_49104710.jpg

[11] http://iamcr.org/sites/default/files/cloud19-01.JPG

[12] http://ae-lane-report.s3.amazonaws.com/wp-content/uploads/2014/12/InternetOfThings.jpg

[13] http://www.webmarketingpros.com/blog/wp-content/uploads/2014/02/mobile-web-word-cloud.jpg

[14] http://www.mtabsurveyanalysis.com/wp-content/uploads/2014/08/data-mining.jpg

[15] http://ijarcet.org/wp-content/uploads/IJARCET-VOL-4-ISSUE-1-18-23.pdf

Bildquellen

[[20] http://www.unglobalpulse.org

[21] https://hadoop.apache.org/images/hadoop-logo.jpg

[22] http://incubator.apache.org/s4/

[23] http://storm.apache.org/images/logo.png

[24] http://spark.apache.org/

[25] http://www.cs.cmu.edu/~pegasus/index_htm_files/504.png

[26] http://mahout.apache.org/images/mahout-logo-brudman.png

[27] https://dato.com/images/dato_logo.svg

[29] http://wikibon.org/w/images/c/cb/Mapreduce.jpg

[30] https://upload.wikimedia.org/wikipedia/commons/thumb/6/6c/MapReduce2.svg/500px-MapReduce2.svg.png

[32] http://bigdata-mining.org/bigmine-15/

[33] http://www.eunis.org/eunis2016/

[34] http://emoocs2016.eu/

Danke.

Hochschule für Angewandte Wissenschaften Hamburg

Hamburg University of Applied Sciences