Einführung und Administratives

VL Big Data Engineering
(aka Informationssysteme)

Prof. Dr. Jens Dittrich

bigdata.uni-saarland.de
twitter.com/jensdittrich

28. April 2020

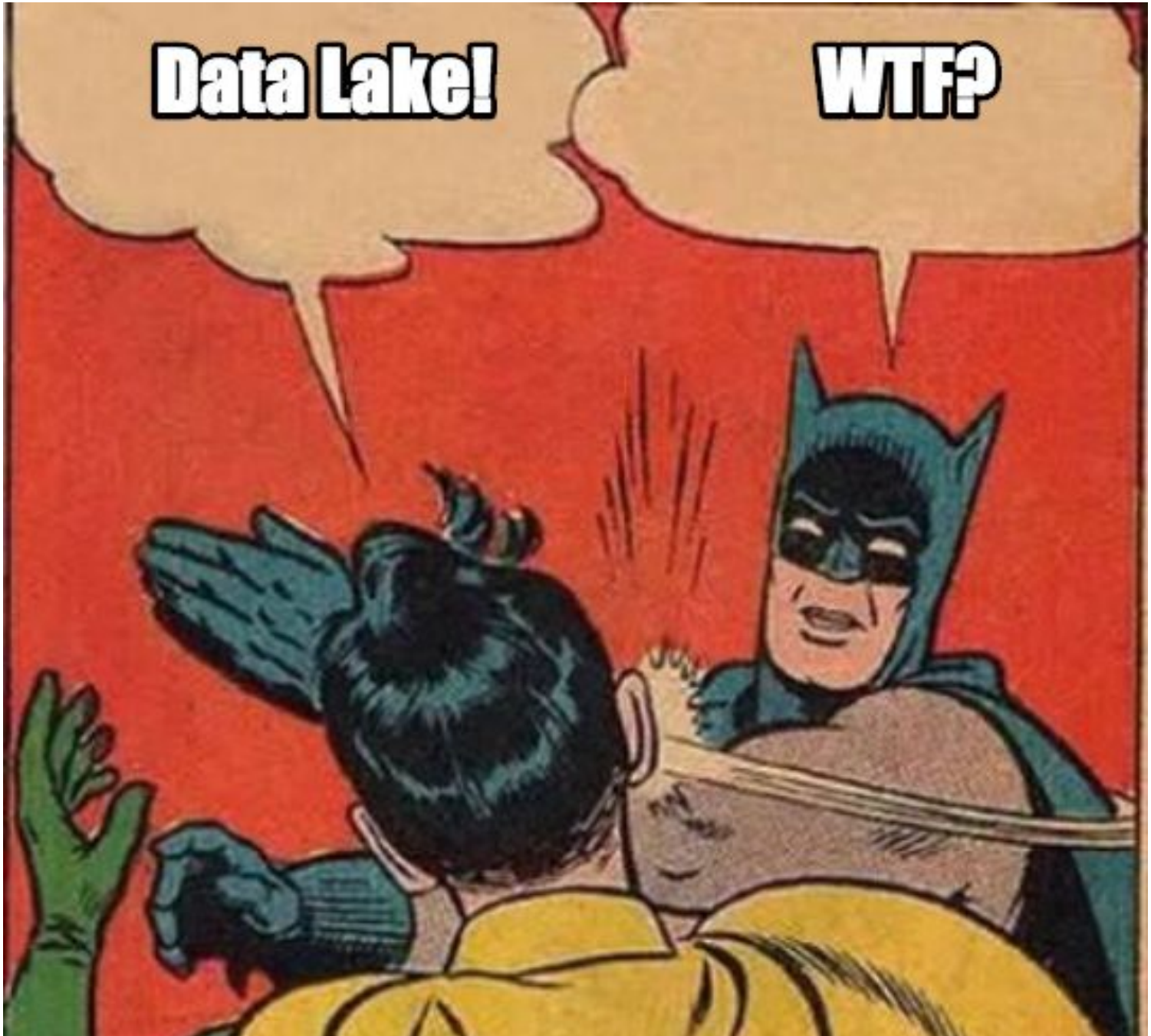
Übersicht über die Vorlesung

- Begriffsbildung
- Inhalt, Konzept
- Lernziele
- Übungen, Zettel
- Klausuren
- Office Hours
- Python

Schlagworte aus dem Bereich Datenanalyse/Big Data Engineering

Informationssysteme
Big Data
Künstliche Intelligenz
Machine Learning
Deep Learning
Data Mining
Cognitive Computing
NoSQL
SQL
DBMS
RDBMS
ODBMS
Datenbanken
Statistik
Lambda-Architektur
Cloud Computing
Data Warehousing
Data Science
Data Lake
Data Engineering
Data Cleaning
Data Curation
Spark
Hadoop
MapReduce
Data Streaming
IoT
Realtime Analytics
Big Data Analytics
Key/Value-Stores
Column Stores
Blockchain

Industrie vs Universität



The Data Science Cake

Ingredients:
50g statistics
120g linear algebra
200g programming
1kg visualisation
300g software engineering

Additional skills:
creativity
out of the box thinking
grit
team spirit

© istock.com sasilsolutions

Die Sicht eines „Datenbänklers“ (1)

- Unsere Top-Konferenz heißt VLDB (Very Large Databases) seit 1975!
- technisch ist das Verwalten und Anfragen großer Datenmengen längst gelöst (**falls** man weiß, was man tut)
- Performance-Probleme mit großen Datenmengen: es liegt selten an Hardware/Software, das ist in 99,99% der Fälle ein Ausbildungsproblem (des Entwicklers/Informatikers)
- Die Kombination von Datenbanktechnologie mit anderen Teilgebieten von Data Science ist hochspannend.
- Beispiel: angewandtes maschinelles Lernen und Datenbanken, diverse Projekte
- Wichtige Datenbankthemen für Data Science: Datenmodellierung, Relationales Modell, SQL, ETL, ELT, Data Cleaning, Data Curation, Data Warehousing, Scalability, verteilte Datenbanken, permissioned Blockchain, ...

Die Sicht eines „Datenbänklers“ (2)

Auswirkungen des Speicher-/Datenbanksystems wird oft unterschätzt

„Die Daten kommen irgendwie von da unten. Wichtig ist die Komplexität der Algorithmen!”
Nein! Das ist für viele System falsch. Auf moderner Hardware ist nicht mehr die CPU der Flaschenhals sondern die langsame „Anlieferung“ der Daten.

Beispiel:

- Ausgangssituation: Hadoop-Cluster mit Spark und Software in Scala
- Änderung von uns: anderes Storage-Layout + ein paar DB-Tricks
- Vorhersage unseres Kostenmodells: Faktor 10000 schneller
- statt teurem Hadoop-Cluster: Laptop oder Smartphone
- KIWI (kill it with iron) **vs** KIWI (kill it with intelligence)

Die Sicht eines „Datenbänklers“ (3)

Die Performanz der Datenbanktechnologie spielt oft keine Rolle.

Wann? Wenn die Daten klein sind und die Hardware so schnell ist, dass es keinen Unterschied macht.

versus

Die Performanz der Datenbanktechnologie spielt oft eine große Rolle.

Wann? Wenn die Daten größer sind und die Hardware die Mängel der Software nicht löst.

Lernziele dieser Vorlesung

1. Grundlegende Techniken im Bereich „Big Data Engineering“ konzeptuell lernen:
   Folien, Übungsaufgaben
2. Grundlegende Techniken im Bereich „Big Data Engineering“ anwenden lernen:
   Python, SQL, Jupyter
3. Ihnen helfen, später nicht das Rad neu zu erfinden:
   Lernen, neue Probleme auf existierende Probleme abzubilden und mit etablierten Techniken zu lösen.
4. Für Probleme wichtiger Anwendungen sensibilisieren:
   Privatheit, Deanonymisierung, ethische Fragestellungen
5. Für Lösungen wichtiger Anwendungen sensibilisieren:
   Aufwand, Performanz, Robustheit, Erweiterbarkeit, Wartbarkeit

Konzept dieser Veranstaltung: Learning by Application

Geplante Struktur für jeweils zwei Wochen Vorlesung:

1. Konkrete Anwendung: XY
2. Was sind die Datenmanagement und -analyseprobleme dahinter?
3. Grundlagen, um diese Probleme lösen zu können
   (a) Folien
   (b) Jupyter/Python/SQL Hands-on
4. Transfer der Grundlagen auf die konkrete Anwendung

Geplante Struktur jedes Übungszettels:

1. 2 Aufgaben mit Bezug zu Grundlagen: Folien
2. 1 Aufgabe mit Bezug zu Grundlagen: Jupyter/Python/SQL Hands-on
3. 1 Aufgabe mit Bezug zum Transfer der Grundlagen auf die Anwendung

Wochenplan: Geplante Themen&Anwendungen

Thema	**Lernziele**
Python (Teil 1, Videos und/oder 5.5.)	Grundlagen, Funktionen, funktionale Programmierung
IMDb (Teil 1, 7.5.)	Datenmodellierung, relationales Modell
Python (Teil 2, Videos und/oder 12.5.)	Objektorientierung und automatisches Testen
IMDb (Teil 2, 14.5.)	Relationale Algebra
NSA (Teil 1, 28.5.)	SQL Einführung
NSA (Teil 2, 4.6.)	Analytisches SQL, Big Data-Arithmetik, Big Data vs Privatheit, Gegenmaßnahmen
Anfrageoptimierung (Teil 1, 18.6.)	Automatische Anfrageoptimierung, Physische Operatoren, Heuristische Optimierung
Anfrageoptimierung (Teil 2, 25.6.)	kostenbasierte Optimierung, Joinreihenfolge, Planvarianten, Pipelining, Physische Optimierungen
Handel, Banken, Ticketsystem (Teil 1, 2.7.)	Datenbankmanagementsysteme (DBMS), Transaktionen, Serialisierbarkeitstheorie
Handel, Banken, Ticketsystem (Teil 2, 9.7.)	Two-Phase Locking (2PL), Isolationsstufen
~~Datenjournalismus (Teil 1)~~	Pivottabellen, Graphdaten, SQL vs Graphdatenbanken, WITH RECURSIVE, Cypher
~~Datenjournalismus (Teil 2)~~	SQL-Injection, Ablegen von Passwörtern, Grundlegende Sicherheitsmaßnahmen
Zusammenfassung (16.7.)	

Abgrenzung zur Stammvorlesung Database Systems

- Diese VL: Fokus auf Prinzipien, Entwurfsmuster und Anwendung von Big Data-Technologien
- (neukonzipierte) Stamm VL “Database Systems” (WS 20/21): tieferer Einstieg in die zugrundeliegenden Techniken

Dozent: Prof. Dr. Jens Dittrich

Forschung:

- Big Data Analytics, Scalable Data Management, Data Science
- ACM SIGMOD, (P)VLDB, ICDE, ...

Lehre:

- Busy beaver awards 2011, 2013, 2018
- youtube: https://www.youtube.com/user/jensdit
- Programmkoordinator BSc&MSc Data Science and Artificial Intelligence (seit WS 19/20)

Industrie:

- Data Science Startup Daimond GmbH, https://daimond.ai
- Blockchain Startup ChainifyDB, https://chainifydb.com

https://bigdata.uni-saarland.de/people/dittrich.php

Tutorinnen und Tutoren

Marcel Maltry (Cheftutor, Doktorand)

Tanja Bäumel

Daniel Emmel

Gideon Geier

Jakob Görgen

Luca Gretscher

Jonas Lauermann

Laura Plein

Alisa Welter

Moritz Wilhelm

siehe https://cms.sic.saarland/bde20/tutors/

CMS und Vorlesungstermine

CMS:

- https://cms.sic.saarland/bde20/
- bitte registrieren Sie sich im System
- Registrierung im LSF **noch nicht** notwendig

Vorlesungstermine:

- jeden Donnerstag 10:15-12:00
- siehe Kalender: https://cms.sic.saarland/bde20/termine/calendar/index
- Folien nach der VL im CMS
- Code nach der VL auf: https://github.com/BigDataAnalyticsGroup/bigdataengineering

Die Uni in Zeiten von Corona

- Info-Mail der FR
- Aktuelle Infos der FR
- FAQ der Uni
- Robert-Koch Institut

Und was bedeutet das jetzt konkret für diese Vorlesung?

Szenario 1: Back to Normal (unwahrscheinlich)

Die Uni erlaubt ab dem 4. Mai wieder Präsenzbetrieb.

⇒ Dann finden Vorlesungen wie sonst auch im GHH und die Tutorien/Office Hours in Seminarräumen statt.

Szenario 2: Partial Lockdown (wahrscheinlich)

Die Uni erlaubt ab dem 4. Mai weiterhin keinen oder nur einen eingeschränkten Präsenzbetrieb.

⇒ Dann finden Vorlesungen und die Tutorien/Office Hours (ggf. partiell) virtuell statt.

Mehr zu Szenario 2:

Virtuelle Vorlesungen (ff. aktualisiert am 22. April 2020)

- die Vorlesungen werden live stattfinden und gestreamt
- Zeitpunkt: ursprünglicher Vorlesungs-Slot
- die Aufzeichnung der Vorlesung erhalten Sie jeweils nach der Vorlesung
- es wird während der live-Vorlesung zusätzliche Feedback-Kanäle für Rückfragen geben

  z.B. vorbereitete Umfragen und Studierendenfragen, kurze Sprechstunde mit dem Prof

Übersicht: Werkzeuge für die virtuelle Lehre

Konzept	**Werkzeug**	**Link**
Vorlesung	Youtube-Livestream & frag.jetzt	https://www.youtube.com/user/jensdit https://frag.jetzt/participant/room/79834810
Vorlesungspausen	Discord	Discord (Einladung per CMS)
Office Hour		
Office Hour (Prof)		
Tutorium		
Materialien	CMS	CMS
Forum		

Fallback-Systeme:

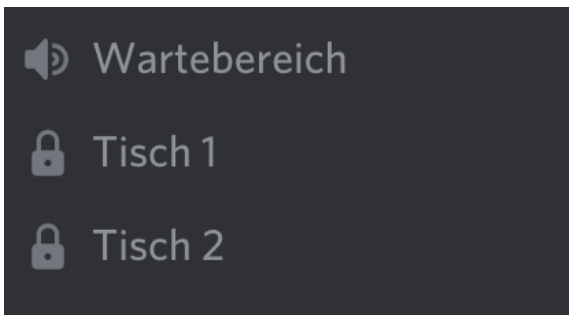
Zoom, MS Teams, BBB, DHL, ...

Youtube Livestreaming

- alle Vorlesungen werden live auf youtube gestreamt (der Stream hat vermutlich in der Regel eine Latenz von 10–15 Sekunden)
- Vorteil für Sie: kein Einloggen notwendig
- danach: archivierter Stream *oder* extra Upload des VL-Videos auf youtube

Discord

- https://discordapp.com
- dies ist ein Tool aus der Gamer-Szene, das Screensharing aber auch Video-Konferenzen unterstützt
- hierfür müssen sie sich einmalig anmelden in Discord
- Sie erhalten von uns eine Einladung zum "Big Data Engineering"-Server in Discord
- klicken Sie dann links einfach einmal auf "Wartebereich": schon "befinden" sie sich im Wartebereich und können alles hören und was andere Anwesende sagen und alle anderen im Wartebereich können Sie hören!



- ein(e) TutorIn, MitarbeiterIn, Prof geht dann mit Ihnen an einen "Tisch" (vom Prinzip her ein Unterraum), um Ihre Fragen zu klären

CMS

Unabhängig von Szenario 1 oder 2:

- CMS inklusive Forum
- alle Folien als pdf **jeweils vor der VL** im CMS/Materialien
- alle Jupyter Notebooks im CMS/Materialien
- umfangreiche Videosammlung zur alten Infosys-Vorlesung auf youtube und datenbankenlernen.de

Office Hours

- zweimal wöchentlich:
  - jeden Dienstag um 12:15
  - jeden Freitag um 12:15
  - in E1 1, R306
- zusätzlich: Office Hour Prof direkt nach jeder Vorlesung
- die Office Hour dienstags wird mehrfach im Semester ersetzt durch unterstützende Veranstaltungen (Hands-On), z.B.: VL: Einführung Python, Einführung in Vagrant, Probeklausuren, ...
- erstmalig am 14.4. um 12:15 Uhr
- siehe Kalender:
  https://cms.sic.saarland/bde20/termine/calendar/index

Tutorien

Prinzip: als LAB gestaltet

1. 15 Minuten Lösungshinweise zum vergangenen (abgegebenen Zettel), dann:
2. 75 Minuten Teamarbeit: einfache Übungsaufgaben lösen

In jedem Tutorium geht es thematisch jeweils um das Material einer 90 Minuten Vorlesungseinheit.

- jede Woche montags und dienstags
- bitte im CMS registrieren
- bitte im CMS Präferenzen für Termine angeben
- wir werden mit voraussichtlich 7 Tutorien starten

Übungszettel

- jeden Donnerstagabend im CMS
- Abgabe eine Woche später bis 10:00 Uhr im CMS
- Abgabe per pdf bzw. Python-Quellcode: nur die Jupyter-Notebook-Zelle, die Sie ausgefüllt haben! (wir erklären das noch genau)
- **keine** Abgabe per Email, Ausdruck, etc.
- Abgabe in Gruppen von 3 Studierenden
- im Semester maximal zwei Übungszettel mit 0 Punkten oder unbearbeitet
- in der Summe müssen mindestens 50% der Punkte erreicht werden, um zur Abschluss- und Wiederholungsklausur zugelassen zu werden
- wir werden Musterlösungen zur Verfügung stellen

Klausuren und Scheine

Klausuren:

- ~~Midterm am 4. Juni (keine Bestehensgrenze)~~
  dieses Semester nicht wegen des verkürzten Semesters
- Abschlussklausur voraussichtlich am 22. Juli morgens
- Wiederholungsklausur X. September

Scheine:

- ~~Note = 30% Midterm + 70% Abschlussklausur~~
- Note = 100% Abschlussklausur
- keine Scheine, Notenmitteilung über HIS
- Ausnahmen: Erasmus, nicht-Informatik Studiengänge, etc.

Python

- wir benutzen nur elementare Sprachkonzepte
- nur sehr wenig Objektorientierung: Vererbung, Polymorphie (OO ist in Python leider nicht gut umgesetzt)
- Achtung: Python 3.7 **nicht** Python 2.x!
- Jupyter Notebooks

Python-Material/Videos

Python-Video Crashkurs

Es wird bereits vor dem offiziellen Semesterstart am 4. Mai eine Reihe kurzer Videos geben mit dem Fokus auf die Python-Sprachkonzepte, die Sie für die VL benötigen.

- die Videos gehen davon aus, dass Sie Prog 1 und Prog 2 gehört haben
- die Notebooks der Videos stellen wir Ihnen zur Verfügung
- zudem gibt es eine virtuelle Maschine (Vagrant/VirtualBox) mit der kompletten Installation, die Sie für die VL benötigen
- wir bieten umfangreiche Unterstützung hierzu an in Office Hours (bereits ab dem 14. April)!

Ziel

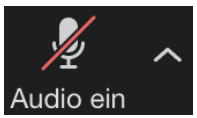
Nutzen Sie die Zeit bis zum 4. Mai, um Python zu installieren und die wesentlichen Python-Sprachkonzepte zu lernen.

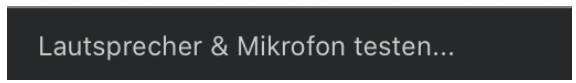
Backup-Systeme

Zoom

- https://zoom.us
- Registrierung in Zoom mit UdS-Email-Account notwendig, Zugang per URL ist ab (spätestens) Anfang Mai im CMS-Kalender eingetragen
- die Sicherheitsdiskussion zu Zoom nehmen wir sehr ernst und haben sie im Blick, wir denken aber, dass Zoom diese Probleme sehr bald behoben haben wird (falls nicht, ist absehbar, dass Zoom sonst ernsthafte wirtschaftliche Probleme bekommen wird)
- siehe Zoom Blog und
- Einschätzung eines IT-Anwalts

Konfiguration von Zoom (WICHTIG)

- laden Sie Zoom herunter
- loggen Sie sich imt Ihrem UdS-Account ein
- testen Sie ihr Audio-Setup, entweder über:
  1. Einstellungen: Audio: Mikrotest und Tontest
     oder:
  2. Neues Meeting: unten links auf den Pfeil neben "Audio ein" klicken:
     

     , dann: 

     , dann den Audio-Test durchführen:
     

  3. ggf. Konfiguration reparieren bis der Test funktioniert