Nadine Pribil

DSI UE 1 Part 1

BigData & DataScience

- Watch the video „Türöffner für die Connected Car Revolution”.
- Consider the topic (in its entirety) according to the Big Data challenges you learned about in the course of this unit (“Challenges 1 through 6” from the “Big Data Challenges” slide set)
- Further consider the topic with regard to the 4 levels of data processing (“challenges level 1 to level 4” from the slide set “Big Data Challenges”)
- Summarize your results / findings in a tabular overview and submit the results (Word, PDF, …) in Moodle

Data Overview Connected Cars	
Data Volumes:	25 GB per car per hour | 130 TB per car per year
Data Types:	Data from ECUs (Speed, RPM, Fuel Efficiency, Temp, Pressure, Braking), Location Data, Safety Data, V2V and V2I, Video
Data Variety:	Streaming (Real-Time), Batch, Multiple Formats
Data Sources:	ECU (Electronic control units), Vehicle Plug-ins, Cameras

Challenges of Big Data	
Challenge 1	How to deal with large amounts of data? - Where to store? ○ E.g.: distributed on many computers → cluster systems ○ Attention: data transfer bottleneck (e.g. limited speed of hard disks, Network cards, …) → velocity becomes a criterion - How to store? ○ Examples: RDBMS, NoSQL, FlatFiles (Text, CSV, …) ○ Large variety of different data types → Variety - Don’t store at all? ○ E.g. Streaming data → requires high processing speed (see Velocity) → Volume can never be considered alone
	Volume (Huge amount of data) As you can see in the Data Overview Table from the Connected Cars above, the volumes of the generated data are very high. There will be 25 GB of data per hour and 130 TB of data per year and per car collected. This huge amount is the first challenge for the Connected Cars concept. Which storing method to use is highly dependent on the data value and variety. For example, data that is used for reducing the risk of the autopilot function, needs to be stored in the car, as it is required to react as fast as the human driver would. It would take too long to retrieve the data from a server or database. On the other hand, data for marketing purposes for example, is not required to be stored in the car itself but can be transferred to a database and further be processed there, so advertisement can be sent later to the car owner via batch. In addition, as 25 GB of data per hour is a huge amount, there should be a filter for data, that needs to have permanent or long-term storing and for data, that can be deleted after a day, as the system only needs the reference data of the last day. To store all accumulated data from every day would be too much and probably not necessary.

Challenge 2	Access to data of various types and sources - A lot of different data formats - Different types to transfer data - Access to data in real time necessary? → Velocity - Detect erroneous data (e.g. in IoT) → Veracity Storing data of different kind - RDBMS can store only strictly structured data - Alternative storage concepts required, e.g. NoSQL → Volume as an additional problem factor Streaming data? - E.g. audio, video, time series (e.g. IoT) → Volume, Velocity, Veracity (e.g. with sensor data) → Variety can never be considered alone
	Variety (Different formats of data from various sources) The Connected Car concept definitely will accumulate many different data formats and transfer types. For example, as can be seen in the Connected Car Data Overview Table, geolocation data, video data, safety data or data from ECUs (like the speed or the fuel efficiency). These data is transferred via streaming, via batch or multiple formats. Again, it depends on the data value, which transfer type should be used. For example the data for the fuel efficiency can be accumulated over a time period and then sent to the servers via batch. On the other hand, the speed data is very important for the autopilot function and needs to be streamed for avoiding dangerous situations. Also, the video data and sensor data is very important to stream – as the car needs to avoid crashes while changing the car track.
Challenge 3	- Extremely strong differences depending on the amount of data → Volume - Analytics part can be extremely time consuming. Performance gain e.g. through: ○ Clustering: parallel processing, program comes to the data and not vice versa (elimination bottleneck network) ○ RAM instead of harddisk (elimination bottleneck harddisk access) ○ Omit storing data (Filter important key data from streaming data, original data is discarded → elimination bottleneck harddisk access) - Special hardware: ○ Performance gain in intensive calculations through the use of graphics cards (GPUs) ○ Hardware manufactures challenged by HW optimized for AI → Velocity can never be seen alone
	Velocity (High speed of accumulation of data) Analyses for the Connected Cars autopilot functions need to be streamed, as it is the most critical part of the concept. The use of Clustering and GPUs can be considered here to improve the performance. Overviewprotocolls on the other hand, don't need to be streamed but rather can be transmitted via batch after a 2 day time period for example. Video material will generate a huge amount of data volume, the autopilot function for example will only need the current video material of the car. Video material from a week ago, will not be used for the decisions and scenarios at hand. Thus, storing these would be not necessary and too expensive.
Challenge 4	- Indirect implications at the infrastructural level

	- Special (additional) steps have to be performed, which can have strong effects on Volume, Variety, Velocity: o Checking the data quality (data cleaning) under the aspects of volume, variety, velocity ▪ (Semi) automated techniques are necessary to recognize… • Outliers • Missing values • Inconsistencies o Anonymization, pseudonymization o Development and validation of suitable analysis models ▪ E.g. in Machine Learning: meaningful feature extraction, quality checks (recall, precision, accuracy, …)
	The data quality is very important in the Connected Car concept. For example it is dangerous for analyses of the autopilot, if the received data is transmitted wrong. It is crucial for the safety of the drivers, that the correctness of the data accumulated is always given, as failures can easily end in bad scenarios, crashes and even death. Inconsistencies, missing or incorrect data needs to be checked accordingly. As the amount of data and speed of accumulation is very high, the data check for the data quality will not be easy to handle.
Challenge 5	- Primarily dependent on objectives and economic considerations - Without recognizable benefit, analyses are pointless - The following must be taken into account: o Legal framework o Ethical and moral framework o Validation of results o Data Governance - Implications for volume, variety, velocity, veracity
	Combining the accumulated data to produce analyses are also an important aspect in the Connected Cars concept. Insurance Packages can be individually customised on the driving habits of the drivers. A risk score can be calculated based on the driving habits for each driver. In this score are factors like driving speed, aggressiveness of gaining speed, hitting the brakes, or changing lines are included.
Challenge 6	Was not included in the Slides.

Levels of Data Handling	
Level 1	Data Sources (Data Source Layer) - How to access the data: o Networks (Internet, Wifi, LAN, harddisks, others): speed (Velocity), bandwith (Volume), type o Transfer protocols: standardized, proprietary (e.g. in the SmartHome area, IoT) → Variety o Data provision, access point, interfaces o Data formats: a lot of possibilities → Variety
	Transfer protocols between Connected Cars need to be standardized as it is important to have the same data formats. The Accessibility of the data for each connected car needs to be provided dependently on the importance of the data value and usage. Data used for the autopilot function needs to be streamed at high speed, as the safety of the passengers and the environment (other cars, walkers) is most important. Data used for marketing purposes and analytical reports can be transmitted via Wifi in the parking lots of the owners every day via batch.

Level 2	Data Messaging and Store Layer (Data Storage Layer) - Suitable storage formats have to be found: o In what form and to what extent are the data available (Variety & Volume) - Appropriate criteria are to be identified (what is actually important? – CAP theorem): o Consistency – availability – partition tolerance - Data may need to be converted or quality checked before storage of after readout: o Data Quality, Data Curation, Data Cleaning o ETL / ELT – processes o Variety, Volume and Velocity can be important criteria
	It is important to quality check the data before streaming or storing, as life can depend on it. Afterwards the storing of this data is dependent on the usages. Again, Video Material from the last week, does not to be stored in the car for the autopilot function, as it wont drive in the past, but needs the Video data from the moment. It would not be necessary and to expensive to store all of it. This data should be streamed, data for marketing purposes on the other hand can be stored in a cloud or database for later usage.
Level 3	Analysis Layer (Processing Layer) - Huge and often difficult choice of tools, methods, algorithms - A wide variety of analysis methods depending on data and objectives o Time series analysis o Machine Learning o Text analyses o Predictive Maintenance - Make appropriate feature selection - Machine learning in particular is extremely resource-intensive → Volume, Variety, Velocity
	Algorithms and Machine Learning will be used to predict dangerous situations. Analyses of historical data, current data and video streams will be used to avoid these dangerous situations. It is extremely resource-intensive as the load of data and speed is very high, but as the passengers and the environments safety is crucial for the Product, it has to be used.
Level 4	Consumption Layer (Data Output Layer) - Making results visible - What is the target audience? o Technicians? Managers? Laymen? Experts? o Visualization types o Usability possibly might be an issue - Do results need to be fed back into the system? - Here Veracity and Value are more an issue than the 3 classic Vs Volume, Variety, Velocity
	The output layer varies strong with their intended goal. For the autopilot function for example, the output of the analyses are driving instructions like changing lanes, slowing down or speeding up or signalising that the car needs more fuel/electricity. Outputs for marketing purposes are protocols, reports and the risk score for each driver for example. The risk score will furthermore used for individual insurance packages.