Chapter 13
Big Data in the Energy and Transport Sectors

Sebnem Rusitschka and Edward Curry

13.1 Introduction

The energy and transport sectors are currently undergoing two main transformations: digitization and liberalization. Both transformations bring to the fore typical characteristics of big data scenarios: sensors, communication, computation, and control capabilities through increased digitization and automation of the infrastructure for operational efficiency leading to high-volume, high-velocity data. In liberalized markets, big data potential is realizable within consumerization scenarios and when the variety of data across organizational boundaries is utilized.

In both sectors, there is a connotation that the term "big data" is not sufficient: the increasing computational resources embedded in the infrastructures can also be utilized to analyse data to deliver "smart data". The stakes are high, since the multimodal optimization opportunities are within critical infrastructures such as power systems and air travel, where human lives could be endangered, not just revenue streams.

In order to identify the industrial needs and requirements for big data technologies, an analysis was performed of the available data sources in energy and transport as well as their use cases in the different categories for big data value: operational efficiency, customer experience, and new business models. The energy and transport sectors are quite similar when it comes to the prime characteristics regarding big data needs and requirements as well as future trends. A special area is

S. Rusitschka (✉)
Corporate Technology, Siemens AG, Munich, Germany
e-mail: sebnem.rusitschka@siemens.com

E. Curry
Insight Centre for Data Analytics, National University of Ireland Galway, Lower Dangan, Galway, Ireland
e-mail: edward.curry@insight-centre.org

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_13

the urban setting where all the complexity and optimization potentials of the energy and transport sectors are focused within a concentrated regional area.

The main need of the sectors is a virtual representation of the underlying physical system by means of sensors, smart devices, or so-called intelligent electronic devices as well as the processing and analytics of the data from these devices. A mere deployment of existing big data technologies as used by the big data natives will not be sufficient. Domain-specific big data technologies are necessary in the cyber-physical systems for energy and transport. Privacy and confidentiality preserving data management and analysis is a primary concern of all energy and transport stakeholders that are dealing with customer data. Without satisfying the need for privacy and confidentiality, there will always be regulatory uncertainty and barriers to customer acceptance of new data-driven offerings.

13.2 Big Data in the Energy and Transport Sectors

The following section examines the dimensions of big data in energy and transport to identify the needs of business and end users with respect to big data technologies and their usage.

Electricity Industry Data is coming from digitalized generator substations, transformer substations, and local distribution substations in an electric grid infrastructure of which the ownership has been unbundled. Information can come in the form of service and maintenance reports from field crews about regular and unexpected repairs, health sensor data from self-monitoring assets, data on end usage and power feed-in from smart meters, and high-resolution real-time data from GPS-synchronized phasor measurement units or intelligent protection and relay devices. An example use case comes from Électricité de France (EDF) (Picard 2013), where they currently "do a standard meter read once a month. With smart meters, utilities have to process data at 15-min intervals. This is about a 3000-fold increase in daily data processing for a utility, and it's just the first wave of the data deluge. Data comes from individual load curves, weather data, contractual information; network data 1 measure every 10 min for a target of 35 million customers. The estimated annual data volume would be 1800 billion records or 120 TB of raw data. The second wave will include granular data from smart appliances, electric vehicles, and other metering points throughout the grid. That will exponentially increase the amount of data being generated."

Oil and Gas Industry Data comes from digitalized storage and distribution stations, but wells, refineries, and filling stations are also becoming data sources in the intelligent infrastructure of an integrated oil and gas company. Down hole sensors from production sites deliver data on a real-time basis including pressure, temperature, and vibration gauges, flow meters, acoustic and electromagnetic, circulation solids. Other data comes from sources such as vendors, tracking service crews, measurements of truck traffic, equipment and hydraulic fracturing, water

usage; Supervisory Control and Data Acquisition (SCADA) data from valve and pump events, asset operating parameters, out of condition alarms; unstructured reserves data, geospatial data, safety incident notes, and surveillance video streams. An example use case comes from Shell (Mearian 2012) where "optical fiber attached to down hole sensors generate massive amounts of data that is stored at a private isolated section of the Amazon Web Services. They have collected 46 petabytes of data and the first test they did in one oil well resulted in 1 petabyte of information. Knowing that they want to deploy those sensors to approximately 10,000 oil wells, we are talking about 10 Exabytes of data, or 10 days of all data being created on the Internet. Because of these huge datasets, Shell started piloting with Hadoop in the Amazon Virtual Private Cloud". Others examples in the industry include (Nicholson 2012): "Chevron proof-of-concept using Hadoop for seismic data processing; Cloudera Seismic Hadoop project combining Seismic Unix with Apache Hadoop; PointCross Seismic Data Server and Drilling Data Server using Hadoop and NoSQL".

Transportation In transportation the number of data sources is increasing rapidly. Air and seaports, train and bus stations, logistics hubs, and warehouses are increasingly employing sensors: Electronic on board recorders (EOBRs) in trucks delivering data on load/unload times, travel times, driver hours, truck driver logs, pallet or trailer tags delivering data on transit and dwell times, information on port strikes, public transport timetables, fare systems and smart cards, rider surveys, GPS updates from vehicle fleet, higher volumes of more traditional data from established sources such as frequent flyer programs, etc. An example use case comes from the City of Dublin (Tabbitt 2014) where the "road and traffic department is now able to combine big data streaming from an array of sources—bus timetables, inductive-loop traffic detectors, closed-circuit television cameras, and GPS updates that each of the city's 1000 buses transmits every 20 s—to build a digital map of the city overlaid with the real-time positions of Dublin's buses using stream computing and geospatial data. Some interventions have led to a 10–15 % reduction in journey times".

13.3 Analysis of Industrial Needs in the Energy and Transport Sectors

Business needs can be derived from the previous dimensioning of big data and examples from within the energy and transport sectors:

Ease of use regarding the typical big data technologies will ultimately ensure wide-scale adoption. Big data technologies employ new paradigms and mostly offer programmatic access. Users require software development skills and a deep understanding of the distributed computing paradigm as well as knowledge of the application of data analytics algorithms within such distributed environments. This is beyond the skillset of most business users.

Semantics of correlations and anomalies that can be discovered and visualized via big data analytics need to be made accessible. Currently only domain and data experts together can interpret the data outliers; business users are often left with guesswork when looking at the results of data analytics.

Veracity of data needs to be guaranteed before it is used in energy and transport applications. Because the increase in data that will be used for these applications will be magnitudes bigger, simple rules or manual plausibility checks are no longer applicable.

Smart data is often used by industrial stakeholders to emphasize that an industrial business user needs refined data—not necessarily all raw data (big data)—but without losing information by concentrating only on small data that is of relevance today. In cyber-physical systems as opposed to online businesses, there is information and communication technology (ICT) embedded in the entire system instead of only in the enterprise IT backend. Infrastructure operators have the opportunity to pre-process data in the field, aggregate data, and distribute the intelligence for data analytics along the entire ICT infrastructure to make the best use of computing and communication resources to deal with volume and velocity of mass sensor data.

Decision support and automation becomes a core need as the pace and structure of business changes. European grid operators today need to intervene almost daily to prevent potentially large-scale blackouts, e.g. due to integration of renewables or liberalized markets. Traffic management systems become more and more elaborate as the amount of digitized and controllable elements increase. Business users need more information than "something is wrong". Visualizations can be extremely useful, but the question of what needs to be done remains to be answered either in real-time or in advance of an event, i.e. in a predictive manner.

Scalable advanced analytics will push the envelope on the state of the art. For example, smart metering data analytics (Picard 2013) include segmentation based on load curves, forecasting on local areas, scoring for non-technical losses, pattern recognition within load curves, predictive modelling, and real-time analytics in a fast and reliable manner in order to control delicate and complex systems such as the electricity grid (Heyde et al. 2010). In the US transportation sector, the business value of scalable real-time analytics is already being reaped by using big data systems for full-scale automation applications, e.g. automated rescheduling that helps trains to dynamically adapt to events and be on time across a wide area.[1]

Big data analytics offer many improvements for the end users. Operational efficiency ultimately means energy and resource efficiency and timeliness, which will improve quality of life—especially in urban mobility settings.

Customer experience and new business models related to big data scenarios are entirely based on better serving the end user of energy and mobility. However, both

[1] https://www.mapr.com/blog/why-transportation-industry-getting-board-big-data-hadoop

scenarios need personalized data in higher resolution. There is significant value in cross-combining a variety of data, which on the downside can make pseudonymization or even anonymization ineffective in protecting the identity and behavioural patterns of individuals, or the business patterns and the strategies of companies. New business models based on monetizing the collected data, with currently unclear regulations, leave end users entirely uninformed, and unprotected against secondary use of their data for purposes they might not agree with, e.g. insurance classification, credit rating, etc.

Reverse transparency is at the top of the wish list of data-literate end users. Data analytics need to empower end users to grasp the usage of their data trails. The access and usage of an end users' data should become efficiently and dynamically configurable by the end users. End users need *practical access to information on what data is used by whom, and for what purpose* in an easy-to-use, manageable way. *Rules and regulations* are needed for granting such transparency for end users.

Data access, exchange, and sharing for both business and end users. In today's complex electricity or intermodal mobility markets, there is almost no scenario where all the required data for answering a business, or engineering, question comes from one department's databases. Nonetheless, most of the currently installed advanced metering infrastructures have a lock-in of the acquired energy usage data to the utilities' billing systems. The lock-in makes it cumbersome to use the energy data for other valuable analytics. These data silos have traditional roots from when most European infrastructure businesses were vertically integrated companies. Also, the amount of data to be exchanged was much less, such that interfaces, protocols, and processes for data exchange were rather rudimentary.

13.4 Potential Big Data Applications for the Energy and Transport Sectors

In the pursuit of collecting the many sectorial requirements towards a European big data economy and its technology roadmap, big data applications in energy and transport have been analysed. A finding that is congruent with Gartner's study on the advancement of analytics (Kart 2013) is that big data applications can be categorized as "operational efficiency", "customer experience", and "new business models".

Operational efficiency is the main driver (Kart 2013) behind the investments for digitization and automation. The need for operational efficiency is manifold, such as increasing revenue margins, regulatory obligations, or coping with the loss of retiring skilled workers. Once pilots of big data technologies are set up to analyse the masses of data for operational efficiency purposes, the businesses realize that they are building a digital map of their businesses, products, and infrastructures—and that these maps combined with a variety of data sources can also deliver

additional insight in other areas of the business such as asset conditions, end usage patterns, etc.

The remainder of this section details big data scenarios and the key challenge that prevents these scenarios from uptake in Europe.

13.4.1 Operational Efficiency

Operational efficiency subsumes all use cases that involve improvements in maintenance and operations in real time, or in a predictive manner, based on the data which comes from infrastructure, stations, assets, and consumers. Technology vendors who developed the sensorization of the infrastructure are the main enablers. The market demand for enhanced technologies is increasing, because it helps the businesses in the energy sector to better manage risk. The complexity of the pan-European interconnected electricity markets, with the integration of renewables and liberalization of electricity trading, requires more visibility of the underlying system and of the energy flows in real time. As a rule of thumb, anything with the adjective "smart" falls into this category: smart grid, smart metering, smart cities, and smart (oil, gas) fields. Some examples of big data use cases in operational efficiency are as follows:

- Predictive and real-time analysis of disturbances in power systems and cost-effective countermeasures.
- Operational capacity planning, monitoring, and control systems for energy supply and networks, dynamic pricing.
- Optimizing multimodal networks in energy as well as transportation especially in urban settings, such as city logistics or eCar-sharing for which the energy consumption and feed-in to the transportation hubs could be cross-optimized with logistics.

All of the scenarios in this category have the *main big challenge of the connecting of data silos*: be it across departments within vertically integrated companies, or across organizations along the electricity value chain. The big data use cases in the operational efficiency scenario require seamless integration of data, communication, and analytics across a variety of data sources, which are owned by different stakeholders.

13.4.2 Customer Experience

Understanding big data opportunities regarding customer needs and wants is especially interesting for companies in liberalized consumerized markets such as electricity, where entry barriers for new players as well as the margins are

decreasing. Customer loyalty and continuous service improvement is what enables energy players to grow in these markets.

Some examples of using big data to improve customer experience are as follows:

- Continuous service improvement and product innovation, e.g. individualized tariff offerings based on detailed customer segmentation using smart meter or device-level consumption data.
- Predictive lifecycle management of assets, i.e. data from machines and devices combined with enterprise resource planning and engineering data to offer services such as intelligent on-demand spare-parts logistics.
- Industrial demand-side management, which allows for energy efficient production and increases competitiveness of manufacturing businesses.

The core challenge is handling confidentiality and privacy of domestic and business customers while getting to know and anticipate their needs. The data originator, data owner, and data user are different stakeholders that need to collaborate and share data to realize these big data application scenarios.

13.4.3 New Business Models

New business models revolve around monetizing the available data sources and existing data services in new ways. There are quite a few cases in which data sources or analysis from one sector represents insights for stakeholders within another sector. An analysis of energy and mobility data start-ups shows that there is a whole new way of generating business value if the end user owns the resources. Then the business is entirely customer- and service-oriented; whereas the infrastructures of energy and transport with their existing stakeholders are utilized as part of the service. These are called intermediary business models.

Energy consumer segment profiles, such as prosumer profiles for power feed-in from photovoltaic, or combined heat and power units; or actively managed demand-side profile, etc., from metering service providers could also be offered for smaller energy retailers, network operators, or utilities who can benefit from improvements on the standard profiles of energy usage but do not yet have access to high resolution energy data of their own customers.

The core challenge is to provide clear regulation around the secondary use of energy and mobility data. The connected end user is the minimal prerequisite for these consumer-focused new business models. The new market segments are diversified through big data energy start-ups like Next: Kraftwerke, who "merge data from various sources such as operational data from our virtual power plant, current weather and grid data as well as live market data. This gives Next Kraftwerke an edge over conventional power traders" (Kraftwerke 2014).

In transportation, cars are parked 95 % of the time (Barter, 2013) and according to a recent study, one car-sharing vehicle displaces 32 new vehicle purchases (AlixPartners 2014). Businesses that previously revolved around the product now

become all about data-driven services. On the contrary to the energy sector, this bold move shows the readiness of the transportation incumbents to seize the big data value potential of a data-driven business.

13.5 Drivers and Constraints for Big Data in Energy and Transport

13.5.1 Drivers

The key drivers in the energy and transport sectors are as follows:

- **Efficiency increase** of the energy and transportation infrastructure and associated operations.
- **Renewable energy sources** have transformed whole national energy policies, e.g. the German "*Energiewende*". Renewable energy integration requires optimization on multiple fronts (e.g., grid, market, and end usage or storage) and increases the dependability of electrification on weather and weather forecasts.
- **Digitization and automation** can substantially increase efficiency in the operation of flow networks such as in electricity, gas, water, or transport networks. These infrastructure networks will become increasingly sensorized, which adds considerably to the volume, velocity, and variety of industrial data.
- **Communication and connectivity** is needed to collect data for optimization and control automation. There needs to be bidirectional and multidirectional connectivity between field devices, e.g. intelligent electronic devices in an electricity grid substation or traffic lights.
- **Open data:** Publication of operational data on transparency platforms[2] by grid network operators, by the energy exchange market, and by the gas transmission system operators is a *regulatory obligation* that fosters grass-roots projects. Open Weather Map[3] and Open Street Map[4] are examples of user-generated free of charge data provisioning which are very important for both sectors.
- **The "skills shift":** As a result of retiring of skilled workers, such as truck drivers or electricity grid operators, a know-how shortage is being created that needs to be filled fast. This directly translates to increasing prices for the customers, because higher salaries need to be paid to attract the few remaining skilled workers in the market.[5] In the mid to long term, efficiency increases and more

[2] www.entsoe.net, www.transparency.eex.com, http://www.gas-roads.eu/

[3] http://openweathermap.org/

[4] http://www.openstreetmap.org/

[5] http://www.businessweek.com/articles/2013-08-29/germany-wants-more-truck-drivers

automation will be the prevailing trends: such as driverless trucks[6] in transportation or wide area monitoring protection and control systems in energy.

13.5.2 Constraints

Constraints in the energy and transport sectors are as follows:

- **Skills:** There are comparatively few people who can apply big data management and analytics knowledge together with domain know-how within the sectors.
- **Interpretation:** Implicit or tacit models are in the heads of the (retiring) skilled workers. Scalable domain model extraction becomes key, e.g. in traffic management systems rule bases grow over years to unmanageable complexities.
- **Digitization has not yet reached the tipping point:** Digitization and automation of infrastructure requires upfront investments, which are not well considered, if at all, by the incentive regulation by which infrastructure operators are bound. Real-time higher-resolution data is still not widely available.
- **Uncertainty regarding digital rights and data protection laws:** Unclear views on data ownership hold back big data in the end user facing segments of the energy and transport sectors (e.g. smart metering infrastructure).
- **"Digitally divided" European union:** Europe has fragmented jurisdiction when it comes to digital rights.
- **"Business-as-usual" trumps "data-driven business":** In established businesses it is very hard to change running business value chains. Incumbents will need to deal with a lot of changes: change in the existing long innovation cycles, change to walled garden views of closed systems and silos, and a change in the mind-set so that ICT becomes an enabler if not a core competency in their companies.
- **Missing end user acceptance:** In the energy sector it is often argued that people are not interested in energy usage data. However, when missing end user acceptance of a technology is argued, it is more a statement that a useful service using this technology is not yet deployed.
- **Missing trust:** Trust is an issue that could and should be remedied with technology data protection and with regulatory framework (i.e., appropriate privacy protection laws).

[6] http://www.techhive.com/article/2046262/the-first-driverless-cars-will-actually-be-a-bunch-of-trucks.html

13.6 Available Energy and Transport Data Resources

As the potential for big data was explored within the two sectors, the clearer it became that the list of available data sources will grow and still not be exhaustive. A key observation is that the variety of data sources utilized to find an answer to a business or engineering question is the differentiator from business-as-usual.

- **Infrastructure data** includes power transmission and distribution lines, and pipelines for oil, gas, or water. In transportation, infrastructure consists of motorways, railways, air and seaways. The driving question is ***capacity***. Is a road congested? Is a power line overloaded?
- **Stations** are considered part of the infrastructure. In business and engineering questions they play a special role as they include the main assets of an infrastructure in a condensed area, and are of high economic value. The main driving question is current ***status*** and utilization levels, i.e. the effective capacity of the infrastructure. Is a transmission line open or closed? Is it closed due to a fault on the line? Is a subway delayed? Is it due to a technical difficulty?
- **Time-stamped and geo-tagged data** are required and increasingly available, especially GPS-synchronized data in both sectors, but also GSM data for tracing mobility and extracting mobility patterns.
- **Weather data**, besides **geo-location data**, is the most used data source in both sectors. Most energy consumption is caused by heating and cooling, which are highly weather-dependent consumption patterns. With renewable energy resources power feed-in into the electrical grid becomes weather dependent.
- **Usage data and patterns**, indicators, and derived values of *end usage of the respective resource and infrastructure*, in both energy or transport, can be harvested by many means, e.g. within the smart infrastructure, via metering at stations at the edges of the network, or smart devices.
- **Behavioural patterns** both .affect energy usage and mobility patterns and can be predicted. Ethical and social aspects become a major concern and stumbling block. The positive effects such as better consumer experience, energy efficiency, more transparency, and fair pricing must be weighed against the negative side effects.
- Data sources in the **horizontal IT landscape**, including data coming from sources such as CRM tools, accounting software, and historical data coming from ordinary business systems. The value potential from cross-combining historical data with new sources of data which come from the increased digitization and automation in energy and transportation systems is high.
- Finally a myriad of **external third-party data or open data sources** are important for big data scenarios in energy and transport sectors, including macro-economic data, environmental data (meteorological services, global weather models/simulation), market data (trading info, spot and forward, business news), human activity (web, phone, etc.), energy storage information,

geographic data, predictions based on Facebook and Twitter, and information communities such as Open Energy Information.[7]

13.7 Energy and Transport Sector Requirements

The analysed business user and end user needs, as well as the different types of data sharing needs directly translate into technical and non-technical requirements.

13.7.1 Non-technical Requirements

Several non-technical requirements in the sectors were identified:

- **Investment in communication and connectedness:** Broadband communication, or ICT in general, needs to be widely available across all of Europe and alongside energy and transportation infrastructure for real-time data access. Connectedness needs to extend to end users to allow them to be continuously connected.
- **A digitally united European union:** Roaming costs have been preventing European end users using data-intensive apps across national borders. European data-driven service providers—especially start-ups looking for scalability of their business models—have mainly focused on the US market, and not the 27 other EU member states due to different data-related regulations. European stakeholders require reliable minimally consistent rules and regulations regarding digital rights and regulations. A digital bill of rights[8] as called for by the inventor of the Web, Tim Berners-Lee, is globally the right move and should be supported by Europe.
- **A better breeding ground for start-ups and start-up culture** is required, especially for techno-economic paradigm shifts like big data and the spreading digitization, where new business widely deviates from business-as-usual. Energy and mobility start-ups require more than just financial investments but also freedom for exploration and experimentation with data. Without this freedom innovation has little chance, unless of course the aforementioned techniques for privacy preserving analytics are not feasible.
- **Open data** in this regard is a great opportunity; however, *standardization* is required. Practical migration paths are required to simplify the adoption of state-of-the-art standards. Data model and representation standards will enable the

[7] http://en.openei.org

[8] http://www.wired.com/2014/03/web25

growth of a *data ecosystem* with collaborative data mining, shareable granularity of data, and accompanying techniques that prevent de-anonymization.
- **Data skilled people:** Programming, statistics, and associated tools need to be a part of engineering education. Traditional data analysts need to grasp the distributed computing paradigm, e.g. how to design algorithms that run on massively parallel systems, how to move algorithms to data, or how to engineer entirely new breeds of algorithms.

13.7.2 Technical Requirements

Several technical requirements were identified in the sectors:

- **Abstraction:** from the actual big data infrastructure is required to enable (a) ease of use, and (b) extensibility and flexibility. The analysed use cases have such diverse requirements that there is no single big data platform or solution that will empower the future utility businesses.
- **Adaptive data and system models** are needed so that new knowledge extracted from domain and system analytics can be redeployed into the data analytics framework without disrupting daily business. The abstraction layer should accommodate plug-in adaptive models.
- **Data interpretability** must be assured without the constant involvement of domain experts. The results must be traceable and explainable. Expert and domain know-how must be blended into data management and analytics.
- **Data analytics** is required as part of every step from data acquisition to data usage. In data acquisition embedded in-field analytics can enhance the veracity of data and can support different privacy and confidentiality settings on the same data source for different data users, e.g. service providers.
- **Real-time analytics** is required to support decisions, which need to be made in ever-shorter time spans. In smart grid settings, near real-time dynamic control requires insights at the source of the data.
- **Data lake** is required in terms of low-cost off-the-shelf storage technology combined with the ability to efficiently deploy data models on demand ("schema-on-read"), instead of the typical data warehouse solution of extract-transform-load (ETL).
- **Data marketplaces, open data, data logistics**, standard protocols capable of handling the variety, volume, and velocity of data, as well as data platforms are required for data sharing and data exchange across organizational boundaries.

13.8 Technology Roadmap for the Energy and Transport Sectors

The big data value chain for infrastructure- and resource-centric systems of energy and transport businesses consists of three main phases: data acquisition, data management, and data usage. *Data analytics*, *as indicated by business user needs, is implicitly required within all steps and is not a separate phase*.

The technology roadmap for fulfilling the key requirements along the data value chain for the energy and transport sectors focuses on technology that is not readily available and needs further research and development in order to fulfil the more strict requirements of energy and transport applications (Fig. 13.1).

13.8.1 Data Access and Sharing

Energy and transport are resource-centric infrastructure businesses. Access to usage data creates the opportunity to analyse the usage of a product or service to improve it, or gain efficiency in sales and operations. Usage data needs to be combined with other available data to deliver reliable predictive models. Currently there is a trade-off between enhancing interpretability of data and preserving privacy and confidentiality. The following example of mobility usage data combined with a variety

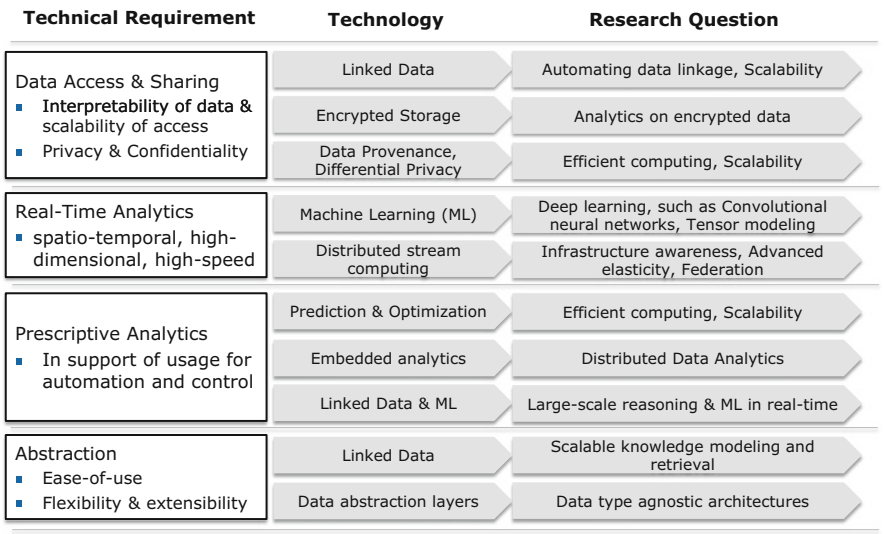

Fig. 13.1 Mapping requirements to research questions in the energy and transport sectors

of other data demonstrates the privacy challenge. de Montjoye et al. (2013) show that "4 spatio-temporal points (approximate places and times) are enough to uniquely identify 95 % of 1.5 M people in a mobility database. The study further states that these constraints hold even when the resolution of the dataset is low". The work shows that mobility datasets combined with metadata can circumvent anonymity.

At the same time, insufficient privacy protection options can hinder the sourcing of big data in the first place, as experiences from smart metering rollouts in the energy businesses show. In the EU only 10 % of homes have smart meters (Nunez 2012). Although there is a mandate that the technology reaches 80 % of homes by 2020, European rollouts are stagnant. A survey from 2012 (Department of Energy and Climate Change 2012) finds that "with increasing reading frequency, i.e. from monthly to daily, to half hourly, etc., energy consumption data did start to feel more sensitive as the level of detail started to seem intrusive... Equally, it was not clear to some [participants] why anyone would want the higher level of detail, leaving a gap to be filled by speculation which resulted in some [participants] becoming more uneasy".

Advances are needed for the following technologies for data access and sharing:

- **Linked data** is a lightweight practice for exposing and connecting pieces of data, information, or knowledge using basic web standards. It promises to open up siloed data ownership and is already an enabler of open data and data sharing. However, with the increasing number of data sources already linked, the various types of new data that will come from intelligent infrastructures, and always connected end users in energy and mobility, scalability and cost-efficacy becomes an issue. One of the open research questions is how to (semi-) automatically extract data linkage to increase current scalability.
- **Encrypted data storage** can enable integrated, data-level security. As cloud storage becomes commonplace for domestic and commercial end users, better and more user-friendly data protection becomes a differentiation factor (Tanner 2014). In order to preserve privacy and confidentiality the use of encrypted data storage will be a basic enabler of data sharing and shared analytics. However, analytics on encrypted data is still an ongoing research question. The most widely pursued research is called fully homomorphic encryption. Homomorphic encryption theoretically allows operations to be carried out on the cipher text. The result is a cipher text that when decrypted matches the result of operation on plaintext. Currently only basic operations are feasible.
- **Data provenance** is the art of tracking data through all transformations, analyses, and interpretations. Provenance assures that data that is used to create actionable insights are reliable. The metadata that is generated to realize provenance across the big variety of datasets from differing sources also increases interpretability of data, which in turn could improve automated information extraction. However, scaling data provenance across the dimensions of big data is an open research question.

- **Differential privacy** (Dwork and Roth 2014) is the mathematically rigorous definition of privacy (and its loss) with the accompanying algorithms. The fundamental law of information recovery (Dwork and Roth 2014) states that too many queries with too few errors will expose the real information. The purpose of developing better algorithms is to push this event as far away as possible. This notion is very similar to the now mainstream realization that there is no unbreakable security, but that barriers if broken need to be fixed and improved. The cutting-edge research on differential privacy considers distributed databases and computations on data streams, enabling linear scalability and real-time processing for privacy preserving analytics. Hence, this technique could be an enabler of privacy preserving analytics on big data, allowing big data to gain user acceptance in mobility and energy.

13.8.2 Real-Time and Multi-dimensional Analytics

Real-time and multi-dimensional analytics enable real-time, multi-way analysis of streaming, spatiotemporal energy, and transport data. Examples from dynamic complex cyber-physical systems such as power networks show that there is a clear business mandate. Global spending on power utility data analytics is forecast to top $20 billion over the next 9 years, with an annual spend of $3.8 billion globally by 2020 (GTM Research 2012). However cost-efficacy of the required technologies needs to be proven. Real-time monitoring does not justify the cost if actions cannot be undertaken in real time. Phasor measurement technology, enabling high-resolution views of the current status of power networks in real time, is a technology that was invented 30 years ago. Possible applications have been researched for more than a decade. Initially there was no business need for it, because the power systems of the day were well engineered and well structured, hierarchical, static, and predictable. With increased dynamics through market liberalization and the integration of power generation technology from intermittent renewable sources like wind and solar, real-time views of power networks becomes indispensable.

Advances are needed for the following technologies:

- **Distributed stream computing** is currently gaining traction. There are two different strains of research and development of stream computing: (1) stream computing as in complex event processing (CEP), which has had its main focus on analysing data of high-variety and high-velocity, and (2) distributed stream computing, focusing on high-volume and high-velocity data processing. Complementing the missing third dimension, volume and variety, respectively, in both strains is the current research direction. It is argued that distributed stream computing, which already has linear scalability and real-time processing capabilities, will tackle high-variety data challenges with semantic techniques (Hasan and Curry 2014) and *Linked data*. A further open question is how to ease development and deployment for the algorithms that make use of distributed

stream computing as well as other computing and storage solutions, such as plain old data warehouses and RDBMS. Since cost-effectiveness is the main enabler for big data value, advanced elasticity with computing and storage on demand as the algorithm requires must also be tackled.

- **Machine learning** is a fundamental capability needed when dealing with big data and dynamic systems, where a human could not possibly review all data, or where humans just lack the experience or ability to be able to define patterns. Systems are becoming increasingly more dynamic with complex network effects. In these systems humans are not capable of extracting reliable cues in real time—but only in hindsight during post-mortem data analysis (which can take significant time when performed by human data scientists). Deep learning, a research field that is gaining momentum, concentrates on more complex non-linear data models and multiple transformations of data. Some representations of data are better for answering a specific question than others, meaning multiple representations of the same data in different dimensions may be necessary to satisfy an entire application. The open questions are: how to represent specific energy and mobility data, possibly in multiple dimensions—and how to design algorithms that learn the answers to specific questions of the energy and mobility domains better than human operators can—and do so in a verifiable manner. The main questions for machine learning are cost-effective storage and computing for massive amounts of high-sampled data, the design of new efficient data structures, and algorithms such as tensor modelling and convolutional neural networks.

13.8.3 Prescriptive Analytics

Prescriptive analytics enable real-time decision automation in energy and mobility systems. The more complex and dynamic the systems are becoming, the faster insights from data will need to be delivered to enhance decision-making. With increasing ICT installed into the intelligent infrastructures of energy and transport, decision automation becomes feasible. However, with the increasing digitization, the normal operating state, when all digitized field devices deliver actionable information on how to operate more efficiently, will overwhelm human operators. The only logical conclusion is to either have dependable automated decision algorithms, or ignore the insights per second that a human operator cannot reasonably handle at the cost of reduced operational efficiency.

Advances are needed for the following technologies:

- **Prescriptive analytics:** Technologies enabling real-time analytics are the basis for prescriptive analytics in cyber-physical systems with resource-centric infrastructures such as energy and transport. With prescriptive analytics the simple predictive model is enhanced with possible actions and their outcomes, as well as an evaluation of these outcomes. In this manner, prescriptive analytics not

only explains what might happen, but also suggests an optimal set of actions. Simulation and optimization are analytical tools that support prescriptive analytics.

- **Machine readable engineering and system models:** Currently many system models are not machine-readable. Engineering models on the other hand are semi-structured because digital tools are increasingly used to engineer a system. Research and innovation in this area of work will assure that machine learning algorithms can leverage system know-how that today is mainly limited to humans. Linked data will facilitate the semantic coupling of know-how at design and implementation time, with discovered knowledge from data at operation time, resulting in self-improving data models and algorithms for machine learning (Curry et al. 2013).
- **Edge computing:** Intelligent infrastructures in the energy and mobility sectors have ICT capability built-in, meaning there is storage and computing power along the entire cyber-physical infrastructure of electricity and transportation systems, not only in the control rooms and data centres at enterprise-level. Embedded analytics, and distributed data analytics, facilitating the in-network and in-field analytics (sometimes referred to as edge-computing) in conjunction with analytics carried out at enterprise-level, will be the innovation trigger in energy and transport.

13.8.4 Abstraction

Abstraction from the underlying big data technologies is needed to enable ease of use for data scientists, and for business users. Many of the techniques required for real-time, prescriptive analytics, such as predictive modelling, optimization, and simulation, are data and compute intensive. Combined with big data these require distributed storage and parallel, or distributed computing. At the same time many of the machine learning and data mining algorithms are not straightforward to parallelize. A recent survey (Paradigm 4 2014) found that "although 49 % of the respondent data scientists could not fit their data into relational databases anymore, only 48 % have used Hadoop or Spark—and of those 76 % said they could not work effectively due to platform issues".

This is an indicator that big data computing is too complex to use without sophisticated computer science know-how. One direction of advancement is for abstractions and high-level procedures to be developed that hide the complexities of distributed computing and machine learning from data scientists. The other direction of course will be more skilled data scientists, who are literate in distributed computing, or distributed computing experts becoming more literate in data science and statistics. Advances are needed for the following technologies:

- **Abstraction** is a common tool in computer science. Each technology at first is cumbersome. Abstraction manages complexity so that the user (e.g.,

programmer, data scientist, or business user) can work closer to the level of human problem solving, leaving out the practical details of realization. In the evolution of big data technologies several abstractions have already simplified the use of distributed file systems by extracting SQL-like querying languages to make them similar to database, or by adapting the style of processing to that of familiar online analytical processing frameworks.

- **Linked data** is one state-of-the-art enabler for realizing an abstraction level over large-scale data sources. The semantic linkage of data without prior knowledge and continuously linking with discovered knowledge is what will allow scalable knowledge modelling and retrieval in a big data setting. A further open question is how to manage a variety of data sources in a scalable way. Future research should establish a thorough understanding of data type agnostic architectures.

13.9 Conclusion and Recommendations for the Energy and Transport Sectors

The energy and transport sectors, from an infrastructure perspective as well as from resource efficiency, global competitiveness, and quality of life perspectives, are very important for Europe.

The analysis of the available data sources in energy as well as their use cases in the different categories of big data value, operational efficiency, customer experience, and new business models helped in identifying the industrial needs and requirements for big data technologies. In the investigation of these requirements, it becomes clear that a mere utilization of existing big data technologies as employed by online data businesses will not be sufficient. Domain- and device-specific adaptations for use in cyber-physical energy and transport systems are necessary. Innovation regarding privacy and confidentiality preserving data management and analysis is a primary concern of the energy and transport sector stakeholders. Without satisfying the need for privacy and confidentiality there will always be regulatory uncertainty, and uncertainty regarding user acceptance of a new data-driven offering.

Among the energy and transport sector stakeholders, there is a sense that "big data" will not be enough. The increasing intelligence embedded in infrastructures will be able to analyse data to deliver "smart data". This seems to be necessary, since the analytics involved will require much more elaborate algorithms than for other sectors. In addition, the stakes in energy and transport big data scenarios are very high, since the optimization opportunities will affect critical infrastructures.

There are a few examples in the energy and transport sectors, where a technology for data acquisition, i.e. a smart device, has been around for many years, or that the stakeholders have already been measuring and capturing a substantial amount of data. However the business need was unclear, making it difficult to justify investment. With recent advances it is now possible for the data to be communicated, stored, and processed cost-effectively. Hence, some stakeholders run the danger of

not acknowledging the technology push. On the other hand, unclear regulation on what usage is allowed with the data keeps them from experimenting.

Many of the state-of-the-art big data technologies just await adaptation and usage in these traditional sectors. The technology roadmap identifies and elaborates the high-priority requirements and technologies that will take the energy and transport sectors beyond state of the art, such that they can concentrate on generating value by adapting and applying those technologies within their specific application domains and value-adding use cases.

References

AlixPartners. (2014). AlixPartners car sharing outlook study. Retrieved from: http://www.alixpartners.com/en/MediaCenter/PressReleases/tabid/821/articleType/ArticleView/articleId/950/AlixPartners-Study-Indicates-Greater-Negative-Effect-of-Car-Sharing-on-Vehicle-Purchases.aspx

Barter, P. (2013, February 22). 'Cars are parked 95% of the time'. Let's check! [Online article]. Available: http://www.reinventingparking.org/2013/02/cars-are-parked-95-of-time-lets-check.html

Curry, E., O'Donnell, J., Corry, E., et al. (2013). Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics, 27*, 206–219.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Sci Rep, 3*, 1376. doi:10.1038/srep01376.

Department of Energy and Climate Change. (2012, December). Smart metering data access and privacy – Public attitudes research. [Whitepaper]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/43045/7227-sm-data-access-privacy-public-att.pdf

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407. doi:10.1561/0400000042.

GTM Research. (December 2012). The soft grid 2013-2020: Big data and utility analytics for smart grid. [Online]. Available: www.sas.com/news/analysts/Soft_Grid_2013_2020_Big_Data_Utility_Analytics_Smart_Grid.pdf

Hasan, S., & Curry, E. (2014b). Thematic event processing. In *Proceedings of the 15th international middleware conference on - middleware'14*, ACM Press, New York, NY, pp 109–120. doi:10.1145/2663165.2663335.

Heyde, C. O., Krebs, R., Ruhle, O., & Styczynski, Z. A. (2010). Dynamic voltage stability assessment using parallel computing. In *Proceeding of: Power and energy society general meeting*, 2010 IEEE.

Kart, L. (April 2013). Advancing analytics. Online Presentation, p. 6. Available: http://meetings2.informs.org/analytics2013/Advancing%20Analytics_LKart_INFORMS%20Exec%20Forum_April%202013_final.pdf

Kraftwerke. (2014). http://www.next-kraftwerke.com/

Mearian, L. (2012, April 4). Shell oil targets hybrid cloud as fix for energy-saving, agile IT [Online article]. Available: http://www.computerworld.com/article/2502623/cloud-computing/shell-oil-targets-hybrid-cloud-as-fix-for-energy-saving--agile-it.html

Nicholson, R.(2012). Big data in the oil and gas industry. IDC energy insights. Presentation. Retrieved from https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/RICK%20-%20IDC_Calgary_Big_Data_Oil_and-Gas/$file/RICK%20-%20IDC_Calgary_ Big_Data_Oil_and-Gas.pdf

Nunez, C. (2012, December 12). Who's watching? Privacy concerns persist as smart meters roll out [Online article]. Available: http://news.nationalgeographic.com/news/energy/2012/12/121212-smart-meter-privacy/

Paradigm 4. (2014, July 1). Leaving data on the table: New survey shows variety, not volume, is the bigger challenge of analyzing big data. Survey. Available: http://www.paradigm4.com/wp-content/uploads/2014/06/P4PR07012014.pdf

Picard, M.-L. (2013, June 26). A smart elephant for a smart-grid: (Electrical) Time-series storage and analytics within hadoop [Online]. Available: http://www.teratec.eu/library/pdf/forum/2013/Pr%C3%A9sentations/A3_03_Marie_Luce_Picard_EDF_FT2013.pdf

Tabbitt, S. (2014, 17 February). Big data analytics keeps Dublin moving [Online article]. Available: http://www.telegraph.co.uk/sponsored/sport/rugby-trytracker/10630406/ibm-big-data-analytics-dublin.html

Tanner, A. (2014, July 11). The wonder (and woes) of encrypted cloud storage [Online article]. Available: http://www.forbes.com/sites/adamtanner/2014/07/11/the-wonder-and-woes-of-encrypted-cloud-storage/