Data veracity in intelligent transportation systems: the slippery road warning scenario

Miroslaw Staron, Riccardo Scandariato
Department of Computer Science and Engineering,
Chalmers / University of Gothenburg,
miroslaw.staron, riccardo.scandariato @gu.se

***Abstract*—Intelligent transportation systems rely on the availability of high quality data in order to allow its multiple actors to make correct decisions in diverse traffic situations. Traditionally, high quality is associated with the correctness of the data, its timeliness or integrity. Going beyond data quality, this paper explores the notion of data veracity, which we approach from the perspective of the truthfulness of the data with respect to reality, or, in other words, its ability to be free from 'lies'. Starting from the concrete case of the slippery road warning scenario (which comes from an industrial player), we define an initial taxonomy of data veracity (which is derived from the study of the literature) and use such taxonomy as a means to analyze the threats to data veracity in the above mentioned scenario. Additionally, this paper has the ambition to draw the attention of researchers and practitioners on the emerging challenges in the fiels of data veracity and to define a research roadmap to tackle such challenges.**

I. Introduction

Intelligent transportation systems have the potential to increase the safety of roads users by means of the timely sharing of road-related information. To this aim, a core component is represented by the connected vehicles. A common trend in the automotive industry is to collect the data generated by connected vehicles and store the data in cloud services. Manufacturers like Volvo, Tesla, and Mercedes (among others) are already following this path for their premium cars. Such data is primarily used by manufactures for the purpose of centralized analysis but can also be shared with interested third parties, like for instance the traffic management agencies, road assistance services, first responders, and so on. The data shared by connected vehicles is generated in large quantities and at a fast pace. These two characteristics are often referred to as velocity an volume, respectively, and are a hefty concern in Big Data systems like the intelligent transportation systems [1].

The present state of research in the Big Data community has devoted major attention to the above-mentioned aspects of Big Data systems. However, as the volume of the data and its velocity increases, the quality of the data is more and more essential in order to avoid problems with the reliability, trustworthiness and safety of these systems [1], [2]. As shown later, unfortunately, the concern of *veracity* of the information used in big data systems is disregarded in the state-of-the-art research and underestimated by the state-of-the-practice. For instance, the Gartner definition of Big Data includes high volume, high velocity and high variety, but does not mention veracity. The ambition of this paper is to define a *roadmap* for research in the field of data veracity, with a specific focus on intelligent transportation systems as the decisions made by such systems can become hazardous for road users if they are based on untrustworthy information.

In this paper we explore the notion of veracity of data, which can be commonly perceived as the truthfulness of the data with respect to reality, or, in other words, its ability to be free from *'lies'*. The concept of lies as violations of the veracity property is crucial as it recognizes the ability of agents in transportation systems (car sensors, road devices, humans) to either willingly or unintentionally introduce false information into the system. In the literature, lies are classifies as either lies of commission, i.e., making a misleading statement about something that is not a fact, or lies of omissions i.e., making a misleading omission of a relevant fact [3]. Both types of lies are relevant in the context of data veracity. A faulty car sensor that stops providing information about the road conditions is an example of omission, while a rogue road user producing fake traffic information is an example of commission.

In this paper, we set forward the idea that veracity has several sub-concerns that need to be studied and understood. Commonly, veracity is interpreted from a trustworthiness perspective, meaning that the provenance of data need to be tracked and leveraged in order to assess the trust level of the data used in any analysis or decision process. However, veracity is a multi-faceted concerns that includes timeliness, precision, completeness and several other sub-concerns. Note that a lie could involve any of the above-mentioned aspects. Therefore, we believe it is of uttermost importance that the different facets of veracity are identified in a taxonomy, so that the threats to veracity can be systematically cataloged and explored as well.

A thorough understanding of the threats to veracity is a foundation stone in order to identify methods, techniques and algorithms to make intelligent vehicle systems robust vis-a-vis such threats. We remark that engineering robustness to data veracity threats is our ultimate goal. The research problem is motivated by the fact that the intelligent transport systems increasingly rely on using Big Data in decision making [4]. The data can come from multiple actors (e.g. vehicles or infrastructure) and can be heterogeneous in terms of its semantics. As the consequences of the noises, biases and abnormality

[1] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

(low veracity of the data) for decision making in algorithms in advanced driver support systems can be catastrophic for drivers and passengers we need to find automated, scalable and high performance methods for assuring freshness, trust and integrity of the data generated by several sources.

From a practitioner's perspective, the selection and adoption of a data veracity technique, should be driven by the risk associated to the veracity threats. This means that lies can be associated to a likelihood value depending on what poses the threat (e.g., a defective sensor vs. a rogue individual). Similarly, lies have a varying impact depending on the consequences of the decisions taken on the basis of such lies. A risk value associated to a lie could be computed staring from the likelihood and the impact values. In turn, the risk value is crucial in order to determine how to deal with a potential lie, e.g., in terms of countermeasures adopted in the system.

In summary, the research agenda we suggest comprises *three pillars*: (i) understanding what are lies in intelligent transportation systems, (ii) defining ways to assess the risks associated to these lies, and (iii) define techniques to deal with the lies. At present time, we are tackling the first pillar mentioned above and have the ambition to create a taxonomy of the various types of veracity threats in intelligent transportation systems. To this aim, starting from the analysis of a concrete scenario (i.e., the slippery road warning scenario as described in Section II), we have setup a qualitative study in which we interview several key stakeholders, including a major car manufacturer (Volvo) and a traffic management agency (the Swedish Transport Administration). In preparation for the study, we have analyzed several sources and compiled an initial list of the sub-concerns of veracity that could be threatened by lies. The results are reported in Section III. Based on such taxonomy, in Section IV, we report on an initial assessment of the potential data veracity threats in the slippery road warning scenario and elaborate on the consequences of such threats. Finally, the paper sets forward a roadmap in Section V, discusses the related work in Section VI and presents the conclusions in Section VII.

II. Example scenario

To illustrate the problems related to the calculation of the veracity of data we can consider a scenario of slippery road warning, which is considered as a scenario of an intelligent transportation system as defined by Dimitrakopoulos and Demestichas [5]. According to that definition an intelligent transportation system is characterized by implementing functionality for providing knowledge to vehicles, thus jointly managing traffic and safety.

The slippery road warning scenario can be found in advertisements of modern cars such as Volvo [6] and is presented in Figure 1.

The scenario has two goals: (i) to warn the drivers about slippery road (thus contributing to increased safety) and (ii) to notify the road administration of the need to handle the slippery road. In the scenario there are a number of actors such such:

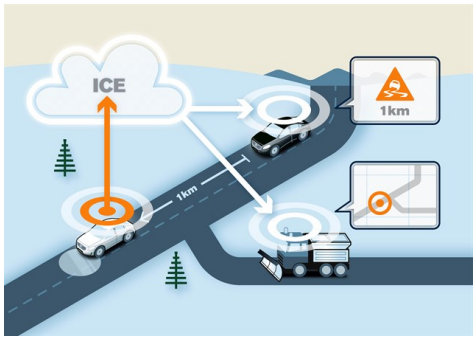

Figure 1. Slippery road warning scenario presented by Volvo Car Group. Picture used with permission from [6].

1) vehicles/drivers which collect the data and get warnings,
2) road administration which gets the notifications and dispatches road maintenance vehicles, and
3) infrastructure providers who store the data, calculate the conditions and notify the road administration.

For the purpose of identifying the actors and the communication channels let us represent the scenario with the focus on actors and the communication channels as presented in Figure 2. In our context of veracity it is both the channels and the actors which are important as the information can become a "lie" in both these kinds of entities.

In this scenario the first vehicle (A) recognizes the slippery road using its sensors (e.g. the ABS systems) and notifies the infrastructure provider about this using the V-2-I channel. The infrastructure provider makes the calculations of how slippery the road it, how long it was/is slippery using the information from Vehicle A and the information about the road (from its database), other vehicles (if available). The calculations can result in a decision that the slippery conditions are getting worse and notifies the road administration which dispatches the road maintenance trucks to salt the road.

In this description the assumption is that none of the actors (Vehicle A, infrastructure provider and road administration) operate on a veracious data, i.e. all information is true. However, it can be the case that the information is not true, e.g. in the following cases:

- sensors in vehicle A have been tampered and are therefore uncalibrated, providing inaccurate measurement data and not notifying the infrastructure provider (i.e. non-varacious data through omission),
- communication V-2-I have been inaccurate and the information about slippery road is sent to a wrong infrastructure provider or with wrong identification of location, and
- the I-2-I communication is blocked by severe weather conditions resulting in not sending the information about the slippery road to the road administration.

In addition to the slippery road warning scenario, the following scenarios and usage areas are often mentioned in the literature when discussing data-intensive communication. In

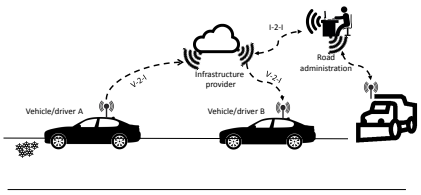

Figure 2. Slippery road warning scenario with the emphasis on actors and their communication channels

the context of intelligent transportation systems these scenarios and areas are:

- Connected vehicles and cooperative driving [7],
- Using robots to enhance the vehicle-to-vehicle communication [8],
- Smoothing of traffic flow [9], and
- Truck fleet optimization [10]

The challenges related to handling of the non-veracious data can differ, but the taxonomy of the assessment methods is similar in all these scenarios.

III. Data veracity

Historically, the notion of veracity is derived from the area of sociology and its major popularity lies in the area of criminology – the ability to detect whether a witness is veracious or not [11], [12]. In that particular context, the term veracity is used both in relation to actors (e.g. witnesses) and their statements [13]. The latter refers to judging the truthfulness of a statement and is in scope for our purposes as well.

In our context, we consider the definition of veracity as quoted by Krotofil [14] who defines the veracity as *the property that an assertion truthfully reects the aspect it makes a statement about*. We can see a direct relation to the field of criminology and also see the challenges related to automated assessment of the veracity in the context of software systems.

For instance, the veracity of the data can be violated by:

- non-adequate measurement of a physical property by a sensor because of the inappropriate design of the sensor
- non-adequate measurement caused by a faulty sensor during the operation
- non-adequate measurement caused by an obstructed sensor
- faulty data caused by a malicious agents tempering with the sensor data

Naturally, there are differences in the countermeasures preventing the malicious manipulation of sensors or other unintentional problem. However, dealing with the non-veracious data does not differ – the system making a decision based on the data needs to assure that its actions do not cause harm to the system and the environment it operates in. Therefore, in our work we consider the ability to function properly in the presence of non-veracious data to be central in order to attain robustness to veracity violations. In general robustness is defined as the ability of a system to operation despite violations in its operational environment [15] and in this context is narrowed to only handle the problems related to detecting and handling of data threats.

As shown in Figure 3, data veracity threats can emerge from the violation of several sub-concerns. The list presented in the figure is not a complete taxonomy, but rather an initial list that we are validating by means of a series of interviews with industrial experts. In particular, veracity entails at least the following aspects:

- Free from error: this is the most fundamental property of veracity. Errors can be generated because of a lack of measurement accuracy or because of bogus data produced maliciously.
- Precision: the data is exact and withing the acceptable tolerance errors. For instance, a location can be reported as a circle and a more precise location involves a smaller radius. Manipulating the way a fact is reported by altering the precision of the account is a common form of lie.
- Objectivity: the data used for analysis, reasoning, and decision making are based on facts rather than opinions or beliefs. This aspect is particularly relevant for human agents.
- Completeness: the data does not contain omissions, i.e., all relevant data is used to make a claim.
- Provenance: the origin of the data can be ascertained with confidence.
- Freshness: the data is not stale and hence is still relevant at the time it is used for analysis, reasoning and decision making.

Figure 3. Some sub-concerns of data veracity

A. Related concepts

The concept of data veracity is related to a number of information quality attributes, for example a subset of these presented in the AIMQ information quality framework [16]. The relation can be based on the methods for assessing of the veracity (or the related attribute) or the ability to detect the threats to veracity. Figure 4 shows the taxonomy of these concepts. Another related area is the area of information security, which is related to the assessment of veracity of information transmitted over a communication channel (e.g. V-2-I in figure 2).

We present a short description of these related concepts in the list below:

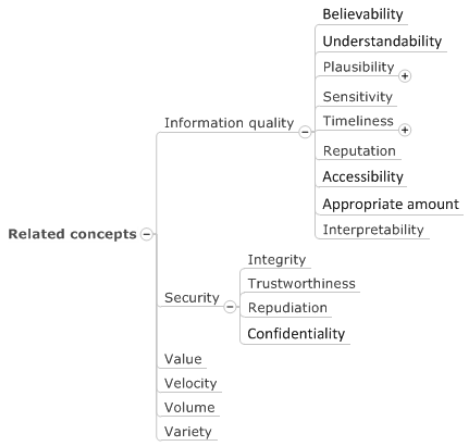

Figure 4. Concepts related to data veracity

- Believability – the degree to which the information can be believed; the non-believeable information should be treated as non-veracious.
- Understandability – the degree to which the data can be understood by the consumer; the data with low understandability should be treated in the same way as low veracity data.
- Plausibility – the degree to which the data is plausible/possible; the concept is very similar to veracity in the sense that plausibility checks can be used to detect non-veracious data.
- Sensitivity – the degree to which the data can be prone to manipulations; the property can indicate that the data should be extra checked for its veracity.
- Timeliness – the degree to which the data is up-to-date (also known as freshness); the data which is out-of-date is no longer useful and can be treated in the same as non-veracious data.
- Reputation – the degree to which the data comes from a reputable source; the low reputation data should be treated in similar way as the uncertain data (sensitive).
- Accessibility – the degree to which the data can be accessed;
- Appropriate amount – the degree to which the data fulfills the requirements for the information richness; if the data does not have the appropriate amount then it should be treated in the same way as low veracity data.
- Interpretability – the degree to which the data can be interpreted; low interpretability data is treated the same as low veracity data.
- Integrity – the degree to which the data has not been tempered with; low integrity data should be treated in the same way as low veracity data.
- Trustworthiness – the degree to which the data can be trusted; treated in the same way as low believability and low veracity.
- Repudiation – the degree to which the data can be traced to its source; non-repudiated data is the same as non-veracious data.
- Confidentiality – the degree to which the data can be kept free from 3rd party reading; low confidentiality data can imply that the data can be non-veracious.

The typical properties of Big Data (volume, value, velocity and variety are only provided for references and more information about them can be found in [17].

In the next section we consider the threats to data veracity in the slippery road warning scenario and the threats to them.

IV. Threats to data veracity in the slippery road warning scenario

In order to present the roadmap for developing robustness algorithms to data veracity threats in section V, we first present a set of example threats to data veracity. As mentioned earlier there are two major sources of threats to veracity: i) actors and ii) communication channels as depicted in figure 2. This first version of the threats taxonomy leads to the understanding of the potential problems with data veracity and helps to explore what needs to be done in order to develop an algorithm for robustness to data veracity threats.

The first group of threats is related to the vehicle as an actor (an excerpt):

- Physically manipulating the sensors in the vehicle (e.g. covering cameras)
- Manipulating with the car without malicious intentions (e.g. putting the snow chains and thus reducing the ability of the sensors to detect wheel slippage and thus slippery conditions)
- Manipulating with the car's software (maliciously or not) and therefore altering the information sent to the cloud
- Reducing the ability of the car to send data to the cloud (e.g. by severe weather conditions)
- Incompatibility between the car's data and the cloud provider's data format (e.g. reporting time stamp in a locale which is different from the locale of the cloud provider)

As it can be noticed, addressing these threats can be done by an algorithm which includes plausibility checks and/or can compare data from two different sources of different believability. However, this cannot be done at the vehicle itself, but needs to be done at the infrastructure provider's side as the vehicle does not have the right reference data (e.g. data from another vehicle).

The second group of threats is related to the infrastructure provider and can be exemplified by:

- Erroneous addressing of data in the databases thus faulty output of decision algorithms
- Using majority voting algorithms to determine which data is "true"
- Marking a non-veracious data points as veracious and spreading this data point to other actors

For these threats we can observe that an algorithm needs to be more complex as the data volume is larger (multiple

vehicles) and therefore new problems arise such as majority voting. Since the infrastructure provider also provides the data to more actors than a single vehicle the consequences can be more severe (e.g. traffic congestions if the faulty warnings are sent or accidents if the warning information is not provided but expected).

The third group of threats is related to communication channels between the actors (an excerpt):

- Non-maliciously altering the data (e.g. data errors caused by severe weather conditions)
- Maliciously altering the transmitted data (e.g. hackers changing the location data)
- Malfunctioning infrastructure altering the data (e.g. broken base station altering the time stamp in the data)

These threats cannot be recognized more easily at the infrastructure provider as the communications can be repeated and the standard algorithms for communication quality assessment can be applied (e.g. checksums). However, the challenge is that from the perspective of the infrastructure provider these threats are similar to the threats for the data veracity at the vehicle (e.g. faulty sensor data).

V. Roadmap

Given the state-of-the-art in veracity assessment of Big Data systems, the complexity of the scenarios in intelligent transportation systems, the development of algorithms to robustly handle veracity issues requires a number of steps. In figure 5 we present a roadmap towards such an algorithm based on the three layered roadmapping method advocated by Phaal et al. [18]. These three layers are (from the top): i) the market pull layer describing the needs for the robustness assessment, e.g. scenarios in intelligent transportation system, ii) features of the algorithm, and iii) technology layer describing the technology enabling the features of the algorithm.

The roadmap presents examples of the features as the work which we present in this paper is still in-progress. To summarize the technology layer of the roadmap we can predict that we are currently on the way of enabling interoperability between systems thanks to open data (i.e. allowing to freely use such data as maps, weather or traffic information). We could also see that the introduction of safety standards (ISO 26262) and common platforms (such as AUTOSAR) provides more possibilities to collect data in similar formats. As the trend continues it enables more advanced features of the robustness algorithm – from the basic information quality assessment (which does not require interoperability), through basic world modelling and threat assessment to advanced world modelling, the algorithm provides the possibility to assess more sub-concerns of veracity and also the related concepts (e.g. plausibility of the data given the semantic information about the location). These kinds of algorithms will contribute to the ability to enable autonomous driving. Please note, however, that veracity assessment is only one small part of the entire algorithm for autonomous driving.

VI. Related work

Veracity of data in the context of a single cyber-physical system has been considered by Krotofil et al. [14]. The authors explored the concept of computational veracity. In our work we expand a similar approach to the transportation systems, in this case consisting of multiple cyber-physical systems.

In general, veracity can be considered as one of the quality attributes of the data together with such attributes (concepts) as believability or timeliness. These attributes are often part of data quality models, such as the AIMQ model by Lee et al. [16].

Automatic assessment of veracity requires a reference point with the annotation which data is true (the reference point) and which is the data that is assessed. A similar approach to the validation of data for another quality attribute of the data has been done in our previous studies at Ericsson [19], [20]. Although the concept of reliability is similar, it does not require the reference point as for veracity and therefore in our current work we study the methods for constructing the reference points for assessing the veracity.

The work of Gerlach et al [21] explores the challenges related to the security of data with the focus on plausibility checks. Although, as shown in our work, the concepts are related checking for plausibility does not require a reference point for the data (e.g. the ”true” value) and therefore our work complements the work on the plausibility checking, also visible in such works as Jabbari et al. [22].

As assessing the robustness of a system to data veracity violations is a subset of robustness as a quality of a measuring system, methods from assessing the robustness of a measurement program can be applied [23], which we intend to address in our future work.

Security of big data is an emerging field of research. The Cloud Security Alliance has outlined the Top 10 security and privacy challenges for big data systems [24]. The challenges are organized into four areas: infrastructure security (including secure computation), data privacy (including granular access control), data management (including secure storage) and reactive security (including monitoring and end-point validation).

Big data systems often employ machine learning to recognize patterns and build business intelligence. The adaptability of such systems could be exploited by attackers, e.g., by means of evasion and poisoning attacks. A taxonomy of attacks against machine learning algorithms has been proposed Barreno et al. [25].

VII. Conclusions

Intelligent transportation systems rely on the availability of high quality data in order to allow its multiple actors to make correct decisions in diverse traffic situations. In this paper we have studied the concept of data veracity, broken it down into sub-concerns and presented the related concepts. We have used the slippery road warning scenario as an example of the kind of threats to data veracity that might exist in the context of intelligent transportation systems. Finally we have also presented the roadmap for the development of algorithms

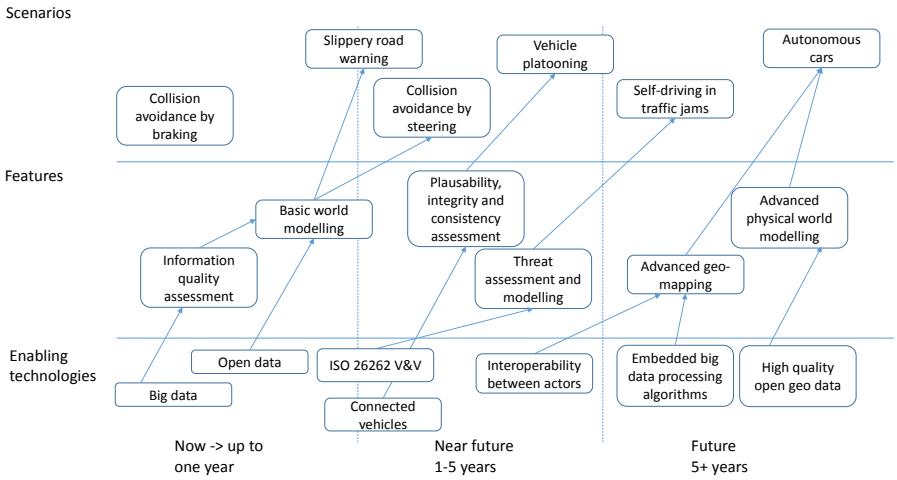

Figure 5. Roadmap for the development of the algorithm for robust handling of veracity threats

which will enable robust handling of big data with respect to veracity threats.

In our current work we focus on interviews with the actors in the Swedish transportation network to validate our models and to provide more details to the veracity threats, namely with the aim at designing a complete taxonomy of veracity threats and their corresponding counter-measures. In the next step we plan to design and implement a demonstrator for the algorithms and to present the theoretical results of how robust such algorithms are.

References

[1] M. Staron, R. Rana, W. Meding, and M. Nilsson, "Consequences of Mispredictions of Software Reliability: A Model and its Industrial Evaluation," in *Mensura*. IEEE, 2014.

[2] M. Staron, "Critical role of measures in decision processes: Managerial and technical measures in the context of large software development organizations," *Information and Software Technology*, vol. 54, no. 8, pp. 887–899, aug 2012. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2012.02.003

[3] S. Bok, *Lying: Moral choice in public and private life*. New York: Pantheon, 1978.

[4] A. Jacobs, "The pathologies of big data," *Communications of the ACM*, vol. 52, no. 8, pp. 36–44, 2009.

[5] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *Vehicular Technology Magazine, IEEE*, vol. 5, no. 1, pp. 77–84, 2010.

[6] V. C. Group, "Volvo Car Group initiates Scandinavian pilot using cloud-based communication to make driving safer," 2014.

[7] A. Ericsson, "Connected Vehicle - an industry in transformation."

[8] F. M. Company, "FORD STUDIES SPACE ROBOTS FOR CONNECTED VEHICLE COMMUNICATIONS," 2013.

[9] L. Glielmo, "Vehicle-to-Vehicle/Vehicle-to-Infrastructure Control," IEEE Computer Society, Tech. Rep., 2016.

[10] S. Ravi, "Scania Leads a New Era of Trucking Services with Data and the Cloud," 2015.

[11] T. R. Levine, H. S. Park, and S. A. McCornack, "Accuracy in detecting truths and lies: Documenting the "veracity effect"," *Communication Monographs*, vol. 66, no. 2, pp. 125–144, jun 1999. [Online]. Available: http://dx.doi.org/10.1080/03637759909376468

[12] S. Mann and A. Vrij, "Police officers' judgements of veracity tenseness, cognitive load and attempted behavioural control in real-life police interviews," *Psychology, Crime & Law*, vol. 12, no. 3, pp. 307–319, jun 2006. [Online]. Available: http://dx.doi.org/10.1080/10683160600558444

[13] P. W. Carey, J. Mehler, and T. G. Bever, "Judging the veracity of ambiguous sentences," *Journal of Verbal Learning and Verbal Behavior*, vol. 9, no. 2, pp. 243–254, apr 1970. [Online]. Available: http://dx.doi.org/10.1016/s0022-5371(70)80058-5

[14] M. Krotofil, J. Larsen, and D. Gollmann, "The Process Matters," in *Proceedings of the 10th ACM Symposium on Information Computer and Communications Security - ASIA CCS '15*. Association for Computing Machinery (ACM), 2015. [Online]. Available: http://dx.doi.org/10.1145/2714576.2714599

[15] "IEEE Standard Glossary of Software Engineering Terminology." [Online]. Available: http://dx.doi.org/10.1109/ieeestd.1990.101064

[16] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, pp. 133–146, dec 2002. [Online]. Available: http://dx.doi.org/10.1016/s0378-7206(02)00043-5

[17] "Big Data Sources," in *Big Data Analytics*. Wiley-Blackwell, sep 2015, pp. 37–46. [Online]. Available: http://dx.doi.org/10.1002/9781119205005.ch5

[18] R. Phaal, C. Farrukh, and D. Probert, "Technology roadmapping: linking technology resources to business objectives," *Centre for Technology Management, University of Cambridge*, pp. 1–18, 2001.

[19] M. Staron and W. Meding, "Ensuring Reliability of Information Provided by Measurement Systems," in *Software Process and Product Measurement*. Springer Science Business Media, 2009, pp. 1–16. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-05415-0_1

[20] M. Staron, W. Meding, and K. Palm, "Release readiness indicator for mature agile and lean software development projects," in *Agile Processes in Software Engineering and Extreme Programming*. Springer Berlin Heidelberg, 2012, pp. 93–107.

[21] M. Gerlach, A. Festag, T. Leinmüller, G. Goldacker, and C. Harsch, "Security architecture for vehicular communication," in *Workshop on Intelligent Transportation*, 2007.

[22] A. Jabbari, R. Jedermann, and W. Lang, "Application of computational intelligence for sensor fault detection and isolation," *World academy of science, engineering and technology*, vol. 33, pp. 265–270, 2007.

[23] M. Staron and W. Meding, "Mesram – a method for assessing robustness of measurement programs in large software development organizations and its industrial evaluation," *Journal of Systems and Software*, vol. 113, pp. 76–100, mar 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jss.2015.10.051

[24] C. S. Alliance, "Expanded top ten big data security and privacy challenges," https://cloudsecurityalliance.org, 2013.

[25] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, 2010.