Responsible Data Science: Using Event Data in a "People Friendly" Manner

Wil M.P. van der Aalst(✉)

Department of Mathematics and Computer Science,
Eindhoven University of Technology,
PO Box 513, 5600 MB Eindhoven, The Netherlands
w.m.p.v.d.aalst@tue.nl
http://www.vdaalst.com/

Abstract. The omnipresence of event data and powerful process mining techniques make it possible to quickly learn process models describing what people and organizations really do. Recent breakthroughs in process mining resulted in powerful techniques to discover the real processes, to detect deviations from normative process models, and to analyze bottlenecks and waste. Process mining and other data science techniques can be used to improve processes within any organization. However, there are also great concerns about the use of data for such purposes. Increasingly, customers, patients, and other stakeholders worry about "irresponsible" forms of data science. Automated data decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties. Each step in the "data science pipeline" (from raw data to decisions) may create inaccuracies, e.g., if the data used to learn a model reflects existing social biases, the algorithm is likely to incorporate these biases. These concerns could lead to resistance against the large-scale use of data and make it impossible to reap the benefits of process mining and other data science approaches. This paper discusses *Responsible Process Mining* (RPM) as a new challenge in the broader field of *Responsible Data Science* (RDS). Rather than avoiding the use of (event) data altogether, we strongly believe that techniques, infrastructures and approaches can be made *responsible by design*. Not addressing the challenges related to RPM/RDS may lead to a society where (event) data are misused or analysis results are deeply mistrusted.

Keywords: Data science · Process mining · Big data · Fairness · Accuracy · Confidentiality · Transparency

1 Introduction

Big data is changing the way we do business, socialize, conduct research, and govern society. Data are collected on anything, at any time, and in any place [5]. Organizations are investing heavily in Big data technologies and data science has emerged as a new scientific discipline providing techniques, methods,

S. Hammoudi et al. (Eds.): ICEIS 2016, LNBIP 291, pp. 3–28, 2017.
DOI: 10.1007/978-3-319-62386-3_1

and tools to gain value and insights from new and existing data sets. Data abundance combined with powerful data science techniques has the potential to dramatically improve our lives by enabling new services and products, while improving their efficiency and quality. Big Data is often considered as the "new oil" and data science aims to transform this into new forms of "energy": insights, diagnostics, predictions, and automated decisions. However, the process of transforming "new oil" (data) into "new energy" (analytics) may negatively impact citizens, patients, customers, and employees. Systematic discrimination based on data, invasions of privacy, non-transparent life-changing decisions, and inaccurate conclusions occur regularly and show that the saying "With great power comes great responsibility" also applies to data science.

Data science techniques may lead to new forms of "pollution". Technological solutions that aim to avoid the negative side effects of using data, can be characterized by the term *"Green Data Science"* (GDS) first coined in [4]. The term refers to the collection of techniques and approaches trying to reap the benefits of data science and Big Data while ensuring fairness, accuracy, confidentiality, and transparency. Citizens, patients, customers, and employees need to be *protected against irresponsible uses of data* (big or small). Therefore, we need to separate the "good" and "bad" of data science. Compare this with environmentally friendly forms of green energy (e.g. solar power) that overcome problems related to traditional forms of energy. Data science may result in unfair decision making, undesired disclosures, inaccuracies, and non-transparency. These irresponsible uses of data can be viewed as "pollution". Abandoning the systematic use of data may help to overcome these problems. However, this would be comparable to abandoning the use of energy altogether. Data science is used to make products and services more reliable, convenient, efficient, and cost effective. Moreover, most new products and services depend on the collection and use of data. *Therefore, we argue that the "prohibition of data (science)" is not a viable solution.* Instead we believe that technological solutions can be used to avoid pollution and protect the environment in which data is collected and used.

In this paper we use the term *"Responsible Data Science"* (RDS) rather than "Green Data Science" (GDS). Our notion of *responsible* is inspired by the emerging field of *responsible innovation* [15,21]. From the overall "responsibility" notion, we distill four main challenges specific to data science:

- **Fairness:** *Data science without prejudice* - How to avoid unfair conclusions even if they are true?
- **Accuracy:** *Data science without guesswork* - How to answer questions with a guaranteed level of accuracy?
- **Confidentiality:** *Data science that ensures confidentiality* - How to answer questions without revealing secrets?
- **Transparency:** *Data science that provides transparency* - How to clarify answers such that they become indisputable?

This paper discusses these so-called "FACT" challenges while emphasizing the need for technological solutions that enable individuals, organizations and society

to reap the benefits from the widespread availability of data while ensuring Fairness, Accuracy, Confidentiality, and Transparency (FACT).

The "FACT" challenges are fairly general. Therefore, the second part of this paper focuses on a specific subdiscipline of data science: *process mining* [5]. Process mining can be used to discover what people actually do, check compliance, and uncover bottlenecks. Process mining reveals the behaviors of workers, customers, and other people involved in the processes being analyzed. The unique capabilities of process mining also create a range of "FACT" challenges. For example, analysis may reveal that workers taking care of the most difficult cases are slower than others or cause more deviations. Moreover, the filtering of event data may be used to influence the outcomes in such a way that decision makers are not aware of this. These examples illustrate the negative side-effects that *Responsible Process Mining* (RPM) aims to avoid.

This paper extends the ICEIS 2016/ENASE 2016 keynote paper [4] by introducing the data science discipline and by elaborating on RDS and RPM. The remainder of this paper is organized as follows. Section 2 introduces the field of data science and uses the example of photography to illustrate the impact of digitization in our daily lives. In Sect. 3 we elaborate on the four general "FACT" challenges. Section 4 introduces process mining as a technology to analyze the *behavior* of people and organizations. In this more specific setting, we revisit the four "FACT" challenges and mention possible solution directions (Sect. 5). Finally, Sect. 6 concludes the paper.

2 Data Science

Many definitions have been proposed for data science [11,24]. Here, we use a definition taken from [5]:

> *Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.*

The definition shows that the data science field is quite broad. Data science has it roots in different fields. Like computer science emerged from mathematics, data science is now emerging from a range of disciplines (see Fig. 1).

Within statistics, one of the key areas in mathematics, there is a long tradition in data analysis. Statistics developed over four centuries starting with the work of John Graunt (1620–1674). Although data science can be seen as a continuation of statistics, the recent progress in data science cannot be attributed to traditional statisticians that tend to focus more on theoretical results rather

than real-world analysis problems. The computational aspects, which are critical for larger data sets, are typically ignored by statisticians [5,27]. The focus is on generative modeling rather than prediction and dealing with practical challenges related to data quality and size. It was the data mining community that realized major breakthroughs in the discovery of patterns and relationships (e.g., efficiently learning decision trees and association rules). Data science is also closely related to data processing. Turing award winner Peter Naur (1928–2016) used the term "data science" long before it was in vogue [5]. In 1974, Naur wrote: "A basic principle of *data science*, perhaps the most fundamental that may be formulated, can now be stated: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available" [19].

As Fig. 1 shows, the roots of data science extend beyond mathematics and computer science. Other areas include ethics, law, economics, and operations management.

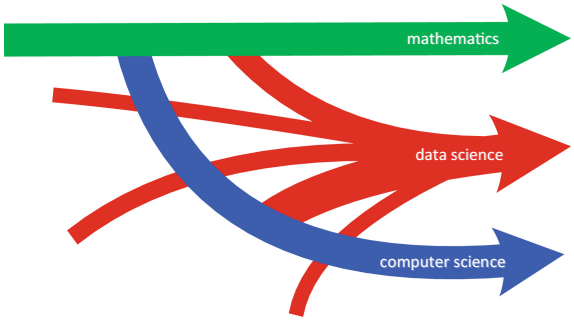

Fig. 1. Just like computer science emerged from mathematics, data science is now emerging from multiple disciplines.

To illustrate the relevance of data science, let us consider the development of photography over time as sketched in Fig. 2. Photography emerged at the beginning of the 19th century. Until 1975 photos were analog and for a long time Kodak was the undisputed market leader. At the peak of its success Kodak developed the first digital camera. It could make 0.01 megapixel black and white pictures and marked both the beginning of the digital photography and the decline of Kodak as a company (see Fig. 2). In 2003, the sales of digital cameras exceeded the sales of traditional cameras for the first time. Today, we make photographs using smartphones and tablets rather than cameras. The remarkable transition from analog to digital photography illustrated by Fig. 2 has had an impact that goes far beyond the photos themselves. The digitization of photography enabled new applications. For example, photos can be shared online (e.g. Flickr, Instagram, Facebook, and Twitter) and changed the way we communicate and socialize (see the uptake of the term "selfie"). Smartphone apps can even be

used to detect eye cancer, melanoma, and other diseases by analyzing photos. Photos capture "events" showing what is really happening. This is enabling new forms of data analysis.

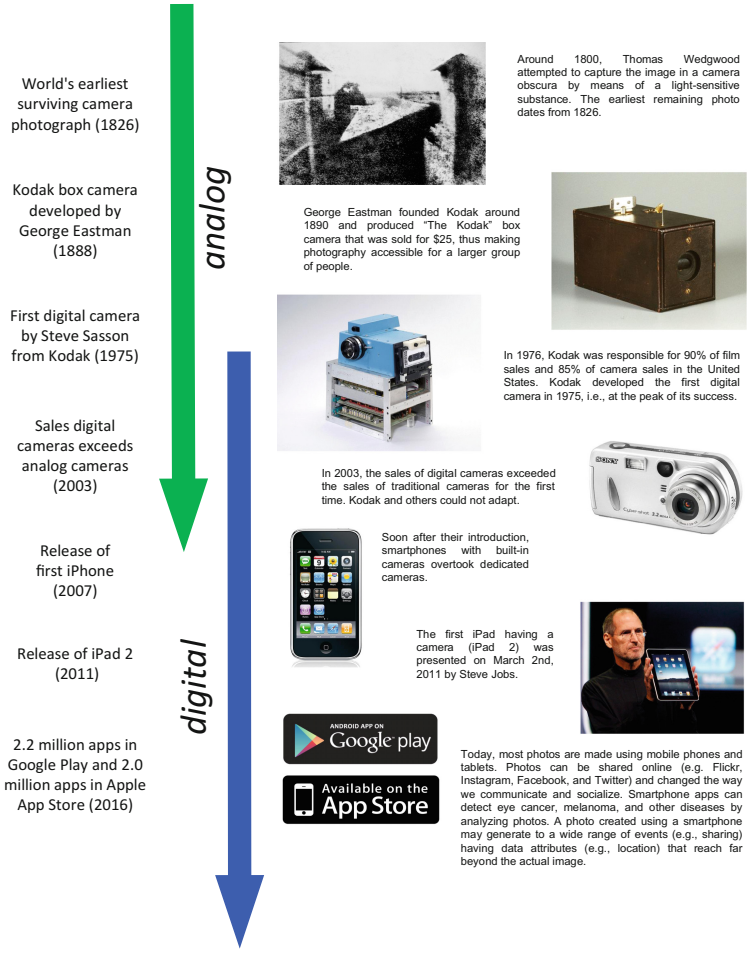

Fig. 2. Example of digitization: digital photography changed the way we make and use photos. Moreover, the digitization of photos enabled new forms of analysis.

Similar developments can be witnessed in all economic sectors. Consider for example the music industry. The transition from analog to digital music has quite some similarities with Fig. 2.

Looking at the timeline in Fig. 2, one can easily see why data science is now emerging as a new discipline. The exponential growth of data over the last decades has now reached a "tipping point" dramatically changing the way we do business and socialize. After explaining why and how data science emerged as a new discipline, we now use Fig. 3 to introduce the three main aspects of data science:

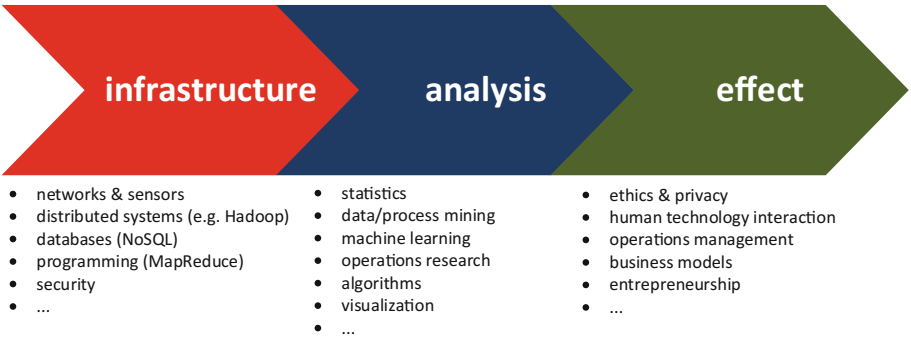

Fig. 3. The data science landscape composed of three main aspects: infrastructure, analysis, and effect.

- **Infrastructure:** *How to collect, store, and process (large amounts of) data?* The infrastructure provides the basis for analysis. Data need to be collected and stored. Systems may need to be distributed to cope with larger amounts of data. Databases may need to be tailored towards the application and special programming models may need to be employed.
- **Analysis:** *How to turn data into insights, answers, ideas, and decisions?* Using the infrastructure different types of approaches can be used to extract value from data. This includes machine learning, data/process mining, statistics, visual analytics, predictive analytics, decision support, etc.
- **Effect:** *How to positively impact reality?* The application of data science may impact individuals, processes, organizations, and society. There may be trade-offs between different goals and stakeholders. For example, privacy concerns may conflict with business targets.

Figure 4 provides yet another view on the data science landscape by sketching the "data science pipeline". Individuals interact with a range of hardware/software systems (information systems, smartphones, websites, wearables, etc.) ❶. Data related to machine and interaction events are collected ❷ and preprocessed for analysis ❸. During preprocessing data may be transformed, cleaned, anonymized, de-identified, etc. Models may be learned from data or made/modified by hand ❹. For compliance checking, models are often normative and made by hand rather than discovered from data. Analysis results based on data (and possibly also models) are presented to analysts, managers, etc. ❺ or used to influence the behavior of information systems and devices ❻. Based on the data, decisions are made or recommendations are provided. Analysis results may also be used to change systems, laws, procedures, guidelines, responsibilities, etc. ❼.

3 Responsible Data Science (RDS)

Figure 4 also lists the four "FACT" challenges mentioned in the introduction. Each of the challenges requires an understanding of the whole data pipeline. Flawed analysis results or bad decisions may be caused by different factors such as a sampling bias, careless preprocessing, inadequate analysis, or an opinionated presentation. We use the term *Responsible Data Science* (RDS) for data science approaches that try to exploit data while avoiding negative side-effects. RDS is synonymous with "Green Data Science" (GDS) [4]. The latter term is based on the metaphor that "data is the new oil" and that we should develop technologies to avoid the "pollution" caused by irresponsible uses of data.

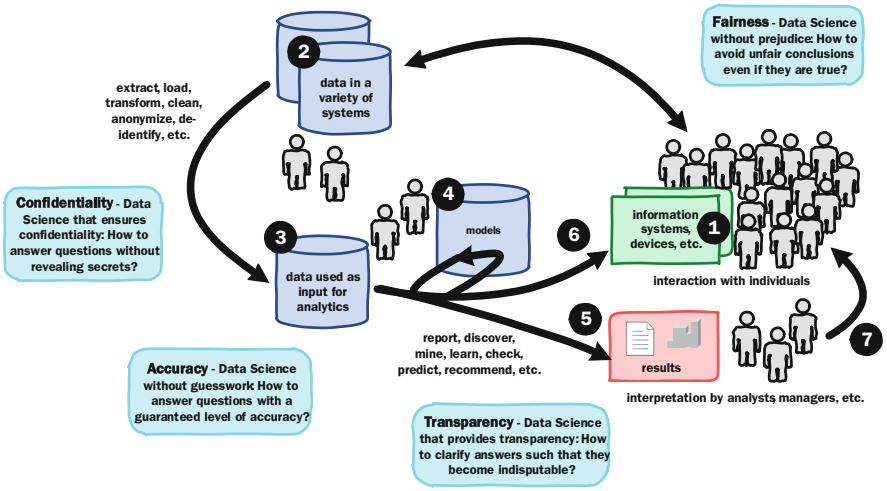

Fig. 4. The "data science pipeline" facing the four "FACT" challenges.

RDS advocates taking the third aspect ("effect") in Fig. 3 as leading when designing or using the first two aspects ("infrastructure" and "analysis"). Whenever possible, infrastructures and analysis techniques should be responsible by design.

The remainder of this section elaborates on the four "FACT" challenges: Fairness, Accuracy, Confidentiality, and Transparency.

3.1 Fairness - Data Science Without Prejudice: How to Avoid Unfair Conclusions Even if They Are True?

Data science techniques need to ensure *fairness*: Automated decisions and insights should not be used to discriminate in ways that are unacceptable from a legal or ethical point of view. Discrimination can be defined as "the harmful treatment of an individual based on their membership of a specific group or

category (race, gender, nationality, disability, marital status, or age)". However, most analysis techniques *aim to discriminate* among groups. Banks handing out loans and credit cards try to discriminate between groups that will pay their debts and groups that will run into financial problems. Insurance companies try to discriminate between groups that are likely to claim and groups that are less likely to claim insurance. Hospitals try to discriminate between groups for which a particular treatment is likely to be effective and groups for which this is less likely. Hiring employees, providing scholarships, screening suspects, etc. can all be seen as classification problems: The goal is to explain a response variable (e.g., person will pay back the loan) in terms of predictor variables (e.g., credit history, employment status, age, etc.). Ideally, the learned model explains the response variable as well as possible without discriminating on the basis of sensitive attributes (race, gender, etc.).

To explain *discrimination discovery* and *discrimination prevention*, let us consider the set of all (potential) customers of some insurance company specializing in car insurance. For each customer we have the following variables:

- name,
- birthdate,
- gender (male or female),
- nationality,
- car brand (Alfa, BMW, etc.),
- years of driving experience,
- number of claims in the last year,
- number of claims in the last five years, and
- status (insured, refused, or left).

The status field is used to distinguish current customers (status = insured) from customers that were refused (status = refused) or that left the insurance company during the last year (status = left). Customers that were refused or that left more than a year ago are removed from the data set.

Techniques for *discrimination discovery* aim to identify groups that are discriminated based on *sensitive* variables, i.e., variables that should not matter. For example, we may find that "males have a higher likelihood to be rejected than females" or that "foreigners driving a BMW have a higher likelihood to be rejected than Dutch BMW drivers". Discrimination may be caused by human judgment or by automated decision algorithms using a predictive model. The decision algorithms may discriminate due to a sampling bias, incomplete data, or incorrect labels. If earlier rejections are used to learn new rejections, then prejudices may be reinforced. Similar "self-fulfilling prophecies" can be caused by sampling or missing values.

Even when there is no intent to discriminate, discrimination may still occur. Even when the automated decision algorithm does not use gender and uses only non-sensitive variables, the actual decisions may still be such that (fe)males or foreigners have a much higher probability to be rejected. The decision algorithm may also favor more frequent values for a variable. As a result, minority groups may be treated unfairly.

Discrimination prevention aims to create automated decision algorithms that do not discriminate using sensitive variables. It is not sufficient to remove these sensitive variables: Due to correlations and the handling of outliers, unintentional discrimination may still take place. One can add constraints to the decision algorithm to ensure fairness using a predefined criterion. For example, the constraint "males and females should have approximately the same probability to be rejected" can be added to a decision-tree learning algorithm. Next to adding algorithm-specific constraints used during analysis one can also use preprocessing (modify the input data by resampling or relabeling) or postprocessing (modify models, e.g., relabel mixed leaf nodes in a decision tree). In general there is often a *trade-off between maximizing accuracy and minimizing discrimination* (see Fig. 5). By rejecting fewer males (better fairness), the insurance company may need to pay more claims.

Discrimination prevention often needs to use sensitive variables (gender, age, nationality, etc.) to ensure fairness. This creates a *paradox*, e.g., information on gender needs to be used to avoid discrimination based on gender.

The first paper on discrimination-aware data mining appeared in 2008 [22]. Since then, several papers mostly focusing on fair classification appeared: [8, 14, 26]. These examples show that unfairness during analysis can be actively prevented. However, unfairness is not limited to classification and more advanced forms of analytics also need to ensure fairness.

3.2 Confidentiality - Data Science That Ensures Confidentiality: How to Answer Questions Without Revealing Secrets?

The application of data science techniques should not reveal certain types of personal or otherwise sensitive information. Often personal data need to be kept *confidential*. The General Data Protection Regulation (GDPR) (see also Sect. 6) focuses on personal information [10]: *"The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed*

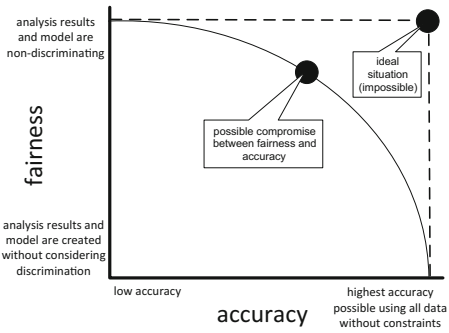

Fig. 5. Tradeoff between fairness and accuracy.

to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

Confidentiality is not limited to personal data. Companies may want to hide sales volumes or production times when presenting results to certain stakeholders. One also needs to bear in mind that few information systems hold information that can be shared or analyzed without limits (e.g., the existence of personal data cannot be avoided). The "data science pipeline" depicted in Fig. 4 shows that there are different types of data having different audiences. Here we focus on: (1) the "raw data" stored in the information system ❷, (2) the data used as input for analysis ❸, and (3) the analysis results interpreted by analysts and managers ❺. Whereas the raw data may refer to individuals, the data used for analysis is often (partly) de-identified, and analysis results may refer to aggregate data only. It is important to note that confidentiality may be endangered along the whole pipeline and includes analysis results.

Consider a data set that contains sensitive information. Records in such a data set may have three types of variables:

- *Direct identifiers*: Variables that uniquely identify a person, house, car, company, or other entity. For example, a social security number identifies a person.
- *Key variables*: Subsets of variables that together can be used to identify some entity. For example, it may be possible to identify a person based on gender, age, and employer. A car may be uniquely identified based on registration date, model, and color. Key variables are also referred to as *implicit identifiers* or *quasi identifiers*.
- *Non-identifying variables*: Variables that cannot be used to identify some entity (direct or indirect).

Confidentiality is impaired by unintended or malicious disclosures. We consider three types of such disclosures:

- *Identity disclosure*: Information about an entity (person, house, etc.) is revealed. This can be done through direct or implicit identifiers. For example, the salaries of employees are disclosed unintentionally or an intruder is able to retrieve patient data.
- *Attribute disclosure*: Information about an entity can be derived indirectly. If there is only one male surgeon in the age group 40–45, then aggregate data for this category reveals information about this person.

– *Partial disclosure*: Information about a group of entities can be inferred. Aggregate information on male surgeons in the age group 40–45 may disclose an unusual number of medical errors. These cannot be linked to a particular surgeon. Nevertheless, one may conclude that surgeons in this group are more likely to make errors.

De-identification of data refers to the process of removing or obscuring variables with the goal to minimize unintended disclosures. In many cases *re-identification* is possible by linking different data sources. For example, the combination of wedding date and birth date may allow for the re-identification of a particular person. *Anonymization* of data refers to de-identification that is irreversible: re-identification is impossible. A range of de-identification methods is available: removing variables, randomization, hashing, shuffling, sub-sampling, aggregation, truncation, generalization, adding noise, etc. Adding some noise to a continuous variable or the coarsening of values may have a limited impact on the quality of analysis results while ensuring confidentiality.

There is a trade-off between minimizing the disclosure of sensitive information and the usefulness of analysis results (see Fig. 6). Removing variables, aggregation, and adding noise can make it hard to produce any meaningful analysis results. Emphasis on confidentiality (like security) may also reduce convenience. Note that *personalization often conflicts with fairness and confidentiality*. Disclosing all data, supports analysis, but jeopardizes confidentiality.

Access rights to the different types of data and analysis results in the "data science pipeline" (Fig. 4) vary per group. For example, very few people will have access to the "raw data" stored in the information system ❷. More people will have access to the data used for analysis and the actual analysis results. Poor cybersecurity may endanger confidentiality. Good policies ensuring proper authentication (Are you who you say you are?) and authorization (What are you allowed to do?) are needed to protect access to the pipeline in Fig. 4. Cybersecurity measures should not complicate access, data preparation, and analysis; otherwise people may start using illegal copies and replicate data. See [18,20,23] for approaches to ensure confidentiality.

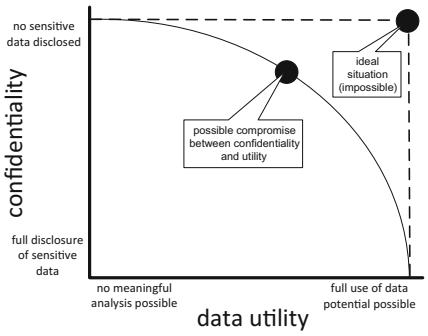

Fig. 6. Tradeoff between confidentiality and utility.

3.3 Accuracy - Data Science Without Guesswork: How to Answer Questions with a Guaranteed Level of Accuracy?

Increasingly decisions are made using a combination of algorithms and data rather than human judgement. Hence, analysis results need to be *accurate* and should not deceive end-users and decision makers. Yet, there are several factors endangering accuracy.

First of all, there is the problem of overfitting the data leading to "bogus conclusions". There are numerous examples of so-called *spurious correlations* illustrating the problem. Some examples (taken from [28]):

- The per capita cheese consumption strongly correlates with the number of people who died by becoming tangled in their bedsheets.
- The number of Japanese passenger cars sold in the US strongly correlates with the number of suicides by crashing of motor vehicle.
- US spending on science, space and technology strongly correlates with suicides by hanging, strangulation and suffocation.
- The total revenue generated by arcades strongly correlates with the number of computer science doctorates awarded in the US.

When using many variables relative to the number of instances, classification may result in complex rules overfitting the data. This is often referred to as the *curse of dimensionality*: As dimensionality increases, the number of combinations grows so fast that the available data become sparse. With a fixed number of instances, the predictive power reduces as the dimensionality increases. Using cross-validation most findings (e.g., classification rules) will get rejected. However, if there are many findings, some may survive cross-validation by sheer luck.

In statistics, Bonferroni's correction is a method (named after the Italian mathematician Carlo Emilio Bonferroni) to compensate for the problem of multiple comparisons. Normally, one rejects the null hypothesis if the likelihood of the observed data under the null hypothesis is low [9]. If we test many hypotheses, we also increase the likelihood of a rare event. Hence, the likelihood of incorrectly rejecting a null hypothesis increases [17]. If the desired significance level for the whole collection of null hypotheses is α, then the Bonferroni correction suggests that one should test each individual hypothesis at a significance level of $\frac{\alpha}{k}$ where k is the number of null hypotheses. For example, if $\alpha = 0.05$ and $k = 20$, then $\frac{\alpha}{k} = 0.0025$ is the required significance level for testing the individual hypotheses.

Next to overfitting the data and testing multiple hypotheses, there is the problem of *uncertainty in the input data* and the problem of *not showing uncertainty in the results*.

Uncertainty in the input data is related to the fourth "V" in the four "V's of Big Data" (Volume, Velocity, Variety, and Veracity). Veracity refers to the trustworthiness of the input data. Sensor data may be uncertain, multiple users may use the same account, tweets may be generated by software rather than people, etc. These uncertainties are often not taken into account during analysis assuming that things "even out" in larger data sets. This does not need to be the case and the reliability of analysis results is affected by unreliable or probabilistic input data.

According to *Bonferroni's principle* we need to avoid treating random observations as if they are real and significant [25]. The following example, inspired by a similar example in [25], illustrates the risk of treating completely random events as patterns.

A *Dutch government agency is searching for terrorists by examining hotel visits* of all of its 18 million citizens (18×10^6). The hypothesis is that terrorists meet multiple times at some hotel to plan an attack. Hence, the agency looks for suspicious "events" $\{p_1, p_2\} \dagger \{d_1, d_2\}$ where persons p_1 and p_2 meet on days d_1 and d_2. How many of such suspicious events will the agency find if the behavior of people is completely random? To estimate this number we need to make some additional assumptions. On average, Dutch people go to a hotel every 100 days and a hotel can accommodate 100 people at the same time. We further assume that there are $\frac{18 \times 10^6}{100 \times 100} = 1800$ Dutch hotels where potential terrorists can meet.

The probability that two persons (p_1 and p_2) visit a hotel on a given day d is $\frac{1}{100} \times \frac{1}{100} = 10^{-4}$. The probability that p_1 and p_2 visit the *same* hotel on day d is $10^{-4} \times \frac{1}{1800} = 5.55 \times 10^{-8}$. The probability that p_1 and p_2 visit the same hotel on two different days d_1 and d_2 is $(5.55 \times 10^{-8})^2 = 3.086 \times 10^{-15}$. Note that different hotels may be used on both days. Hence, the probability of suspicious event $\{p_1, p_2\} \dagger \{d_1, d_2\}$ is 3.086×10^{-15}.

How many candidate events are there? Assume an observation period of 1000 days. Hence, there are $1000 \times (1000 - 1)/2 = 499,500$ combinations of days d_1 and d_2. Note that the order of days does not matter, but the days need to be different. There are $(18 \times 10^6) \times (18 \times 10^6 - 1)/2 = 1.62 \times 10^{14}$ combinations of persons p_1 and p_2. Again the ordering of p_1 and p_2 does not matter, but $p_1 \neq p_2$. Hence, there are $499,500 \times 1.62 \times 10^{14} = 8.09 \times 10^{19}$ candidate events $\{p_1, p_2\} \dagger \{d_1, d_2\}$.

The expected number of suspicious events is equal to the product of the number of candidate events $\{p_1, p_2\} \dagger \{d_1, d_2\}$ and the probability of such events (assuming independence): $8.09 \times 10^{19} \times 3.086 \times 10^{-15} = 249,749$. Hence, there will be around a quarter million observed suspicious events $\{p_1, p_2\} \dagger \{d_1, d_2\}$ in a 1000 day period!

Suppose that there are only a handful of terrorists and related meetings in hotels. *The Dutch government agency will need to investigate around a quarter million suspicious events involving hundreds of thousands innocent citizens.* Using Bonferroni's principle, we know beforehand that this is not wise: there will be too many false positives.

Example 1: Bonferroni's principle explained using an example taken from [5]. To apply the principle, compute the number of observations of some phenomena one is interested in under the assumption that things occur at random. If this number is significantly larger than the real number of instances one expects, then most of the findings will be false positives.

When we say, "we are 95% confident that the true value of parameter x is in our confidence interval $[a, b]$", we mean that 95% of the hypothetically observed confidence intervals will hold the true value of parameter x. Averages, sums, standard deviations, etc. are often based on sample data. Therefore, it is important to provide a confidence interval. For example, given a mean of 35.4 the 95% confidence interval may be [35.3, 35.6], but the 95% confidence interval may also be [15.3, 55.6]. In the latter case, we will interpret the mean of 35.4 as a "wild guess" rather than a representative value for true average value. Although we are used to confidence intervals for numerical values, decision makers have problems interpreting the expected accuracy of more complex analysis results like decision trees, association rules, process models, etc. Cross-validation techniques like k-fold checking and confusion matrices give some insights. However, models and decisions are often presented unequivocally thus hiding uncertainties. Explicit vagueness or more explicit confidence diagnostics may help to better interpret analysis results. Parts of models should be kept deliberately "vague" if analysis is not conclusive.

3.4 Transparency - Data Science That Provides Transparency: How to Clarify Answers Such That They Become Indisputable?

Data science techniques are used to make a variety of decisions. Some of these decisions are made automatically based on rules learned from historic data. For example, a mortgage application may be rejected automatically based on a decision tree. Other decisions are based on analysis results (e.g., process models or frequent patterns). For example, when analysis reveals previously unknown bottlenecks, then this may have consequences for the organization of work and changes in staffing (or even layoffs). Automated decision rules (❻ in Fig. 4) need to be as accurate as possible (e.g., to reduce costs and delays). Analysis results (❺ in Fig. 4) also need to be accurate. However, accuracy is not sufficient to ensure acceptance and proper use of data science techniques. Both decisions ❻ and analysis results ❺ also need to be *transparent*.

Figure 7 illustrates the notion of transparency. Consider an application submitted by John evaluated using three data-driven decision systems. The first system is a black box: It is unclear why John's application is rejected. The second system reveals it's decision logic in the form of a decision tree. Applications from females and younger males are always accepted. Only applications from older males get rejected. The third system uses the same decision tree, but also explains the rejection ("because male and above 50"). Clearly, the third system is most transparent. When governments make decisions for citizens it is often mandatory to explain the basis for such decisions.

Deep learning techniques (like many-layered neural networks) use multiple processing layers with complex structures or multiple non-linear transformations. These techniques have been successfully applied to automatic speech recognition, image recognition, and various other complex decision tasks. Deep learning methods are often looked at as a "black box", with performance measured empirically and no formal guarantees or explanations. A many-layered neural network is not

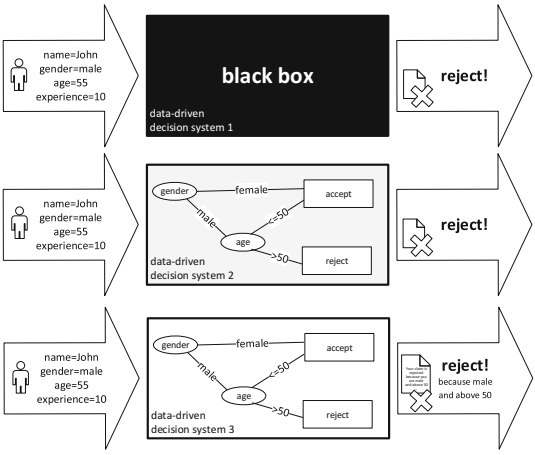

Fig. 7. Different levels of transparency.

as transparent as for example a decision tree. Such a neural network may make good decisions, but it cannot explain a rule or criterion. Therefore, such black box approaches are non-transparent and may be unacceptable in some domains.

Transparency is not restricted to automated decision making and explaining individual decisions, it also involves the intelligibility, clearness, and comprehensibility of analysis results (e.g., a process model, decision tree, regression formula). For example, a model may reveal bottlenecks in a process, possible fraudulent behavior, deviations by a small group of individuals, etc. It needs to be clear for the user of such models (e.g., a manager) how these findings where obtained. The link to the data and the analysis technique used should be clear. For example, filtering the input data (e.g., removing outliers) or adjusting parameters of the algorithm may have a dramatic effect on the model returned.

Storytelling is sometimes referred to as "the last mile in data science". The key question is: How to communicate analysis results with end-users? *Storytelling is about communicating actionable insights to the right person, at the right time, in the right way.* One needs to know the gist of the story one wants to tell to successfully communicate analysis results (rather than presenting the whole model and all data). One can use natural language generation to transform selected analysis results into concise, easy-to-read, individualized reports.

To provide transparency there should be a clear link between data and analysis results/stories. One needs to be able to *drill-down* and inspect the data from the model's perspective. Given a bottleneck one needs to be able to drill down to the instances that are delayed due to the bottleneck. This related to *data provenance*: it should always be possible to reproduce analysis results from the original data.

The four "FACT" challenges depicted in Fig. 4 are clearly interrelated. There may be trade-offs between them. For example, to ensure confidentiality we may add noise and de-identify data thus possibly compromising accuracy and transparency.

4 Process Mining

The goal of *process mining* is to turn event data into insights and actions [5]. Process mining is an integral part of data science, fueled by the availability of data and the desire to improve processes. Process mining can be seen as a means to bridge the gap between data science and process science. Data science approaches tend to be process agonistic whereas process science approaches tend to be model-driven without considering the "evidence" hidden in the data.

4.1 What Is Process Mining?

Figure 8 shows the "process mining pipeline" and can be viewed as a specialization of the Fig. 4. Process mining focuses on the analysis of *event data* and analysis results are often related to *process models*. Process mining is a rapidly growing subdiscipline within both Business Process Management (BPM) [2] and data science [3]. Mainstream Business Intelligence (BI), data mining and machine learning tools are not tailored towards the analysis of event data and the improvement of processes. Fortunately, there are dedicated process mining tools able to transform event data into actionable process-related insights. For example, *ProM* (www.processmining.org) is an

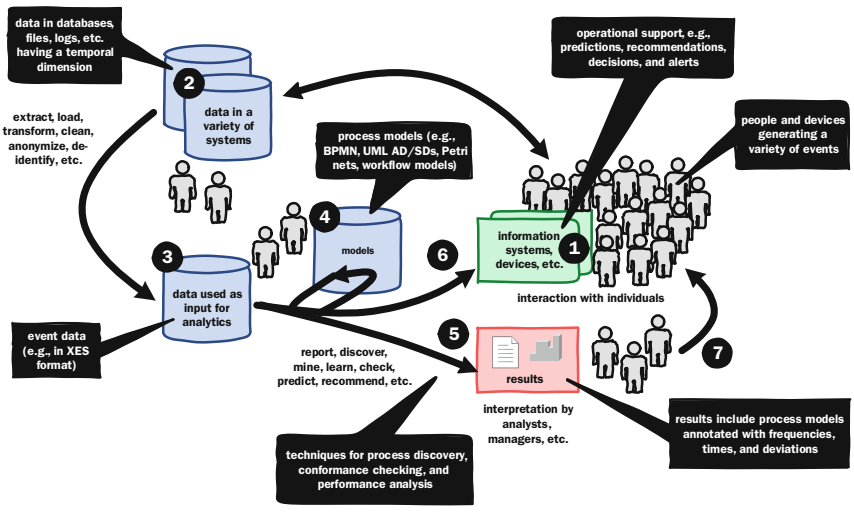

Fig. 8. The "process mining pipeline" relates observed and modeled behavior.

open-source process mining tool supporting process discovery, conformance checking, social network analysis, organizational mining, clustering, decision mining, prediction, and recommendation (see Fig. 9). Moreover, in recent years, several vendors released commercial process mining tools. Examples include: *Celonis Process Mining* by Celonis GmbH (www.celonis.de), *Disco* by Fluxicon (www.fluxicon.com), *Interstage Business Process Manager Analytics* by Fujitsu Ltd. (www.fujitsu.com), *Minit* by Gradient ECM (www.minitlabs.com), *myInvenio* by Cognitive Technology (www.my-invenio.com), *Perceptive Process Mining* by Lexmark (www.lexmark.com), *QPR ProcessAnalyzer* by QPR (www.qpr.com), *Rialto Process* by Exeura (www.exeura.eu), *SNP Business Process Analysis* by SNP Schneider-Neureither & Partner AG (www.snp-bpa.com), and *PPM webMethods Process Performance Manager* by Software AG (www.softwareag.com).

4.2 Creating and Managing Event Data

Process mining is impossible without proper *event logs* [1]. An event log contains event data related to a particular process. Each event in an event log refers to one *process instance*, called *case*. Events related to a case are ordered. Events can have attributes. Examples of typical attribute names are activity, time, costs, and resource. Not all events need to have the same set of attributes. However, typically, events referring to the same activity have the same set of attributes. Figure 9(a) shows the conversion of an CSV file with four columns (case, activity, resource, and timestamp) into an event log.

Most process mining tools support XES (eXtensible Event Stream) [13]. In September 2010, the format was adopted by the IEEE Task Force on Process Mining and became the de facto exchange format for process mining. The IEEE Standards Organization is currently evaluating XES with the aim to turn XES into an official IEEE standard.

To create event logs we need to extract, load, transform, anonymize, and de-identify data from a variety of systems (see ❸ in Fig. 8). Consider for example the hundreds of tables in a typical HIS (Hospital Information System) like Chip-Soft, McKesson and EPIC or in an ERP (Enterprise Resource Planning) system like SAP, Oracle, and Microsoft Dynamics. Non-trivial mappings are needed to extract events and to relate events to cases. Event data needs to be scoped to focus on a particular process. Moreover, the data also needs to be scoped with respect to confidentiality issues.

4.3 Process Discovery

Process discovery is one of the most challenging process mining tasks [1]. Based on an event log, a process model is constructed thus capturing the behavior seen in the log. Dozens of process discovery algorithms are available. Figure 9(c) shows a process model discovered using ProM's *inductive visual miner* [16]. Techniques use Petri nets, WF-nets, C-nets, process trees, or transition systems as a representational bias [5]. These results can always be converted to the desired

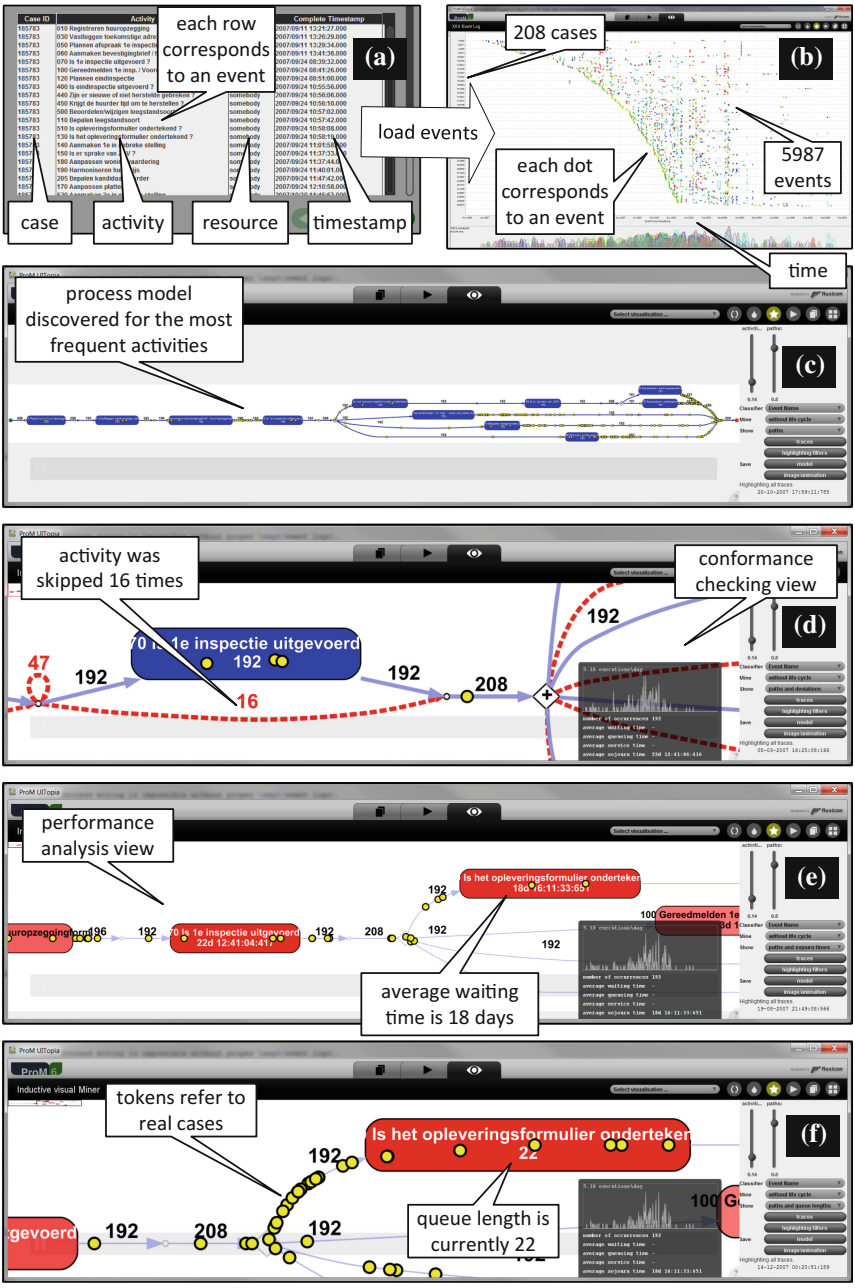

Fig. 9. Six screenshots of ProM while analyzing an event log with 208 cases, 5987 events, and 74 different activities. First, a CSV file is converted into an event log (a). Then, the event data can be explored using a dotted chart (b). A process model is discovered for the 11 most frequent activities (c). The event log can be replayed on the discovered model. This is used to show deviations (d), average waiting times (e), and queue lengths (f).

notation, for example BPMN (Business Process Model and Notation), YAWL, or UML activity diagrams.

4.4 Conformance Checking

Using conformance checking discrepancies between the log and the model can be detected and quantified by replaying the log [6]. For example, Fig. 9(c) shows an activity that was skipped 16 times. Some of the discrepancies found may expose undesirable deviations, i.e., conformance checking signals the need for a better control of the process. Other discrepancies may reveal desirable deviations and can be used for better process support. Input for conformance checking is a process model having executable semantics and an event log.

4.5 Performance Analysis

By replaying event logs on process model, we can compute frequencies and waiting/service times. Using alignments [6] we can relate cases to paths in the model. Since events have timestamps, we can associate the times in-between events along such a path to delays in the process model. If the event log records both start and complete events for activities, we can also monitor activity durations. Figure 9(d) shows an activity that has an average waiting time of 18 days and 16 h. Note that such bottlenecks are discovered without any modeling.

4.6 Operational Support

Figure 9(e) shows the queue length at a particular point in time. This illustrates that process mining can be used in an online setting to provide operational support. Process mining techniques exist to predict the remaining flow time for a case or the outcome of a process. This requires the combination of a discovered process model, historic event data, and information about running cases. There are also techniques to recommend the next step in a process, to check conformance at run-time, and to provide alerts when certain Service Level Agreements (SLAs) are (about to be) violated.

5 Responsible Process Mining (RPM)

This section discusses challenges related to fairness, accuracy, confidentiality, and transparency in the context of process mining. The goal is not to provide solutions, but to illustrate that the more general challenges discussed before trigger concrete research questions in the process mining domain.

5.1 Classification of RPM Challenges

Tables 1 and 2 map the four generic "FACT" challenges introduced in Sect. 3 onto the five key ingredients of process mining briefly introduced in Subsects. 4.2–4.6. Using both dimensions we obtain a classification consisting of $4 \times 5 = 20$ possible problem areas.

It is impossible to discuss all 20 potential problem areas listed in Tables 1 and 2. Therefore, we discuss four selected problem areas in more detail.

5.2 Example: Confidentiality and Creating and Managing Event Data

Let us now explore one of the cells in Table 2. Event data may reveal confidential information as highlighted in Fig. 10. The class model shows the information found in event logs using XES [13], MXML, or some other logging format. Process mining tools exploit such information during analysis. In Fig. 10 three levels are identified: *process model level*, *case/instance level*, and *event level*. The case/instance level consists of *cases* and *activity instances* that connect *processes* and *activities* in the model to *events* in the event log. See [5] for a detailed description of the typical ingredients of an event log. For RPM it is important to note that events and cases often refer to individuals. A case may correspond to a customer, patient, student, or citizen. Events often refer to the person executing the corresponding activity instance (e.g., an employee).

Event data are notoriously difficult to fully anonymize. In larger processes, most cases follow a unique path. In the event log used in Fig. 9, 198 of the 208 cases follow a unique path (focusing only on the order of activities). Hence, knowing the order of a few selected activities may be used to de-anonymize or re-identify cases. The same holds for (precise) timestamps. For the event log in Fig. 9, several cases can be uniquely identified based on the day the registration activity (first activity in process) was executed. If one knows the timestamps of these initial activities with the precision of an hour, then almost all cases can be uniquely identified. This shows that the ordering and timestamp data in event logs may reveal confidential information unintentionally. Therefore, it is interesting to investigate what can be done by adding noise (or other transformations) to event data such that the analysis results do not change too much. For example, we can shift all timestamps such that all cases start in "week 0". Most process discovery techniques will still return the same process model. Moreover, the average flow/waiting/service times are not affected by this. However, if one is investigating queueing or resource behavior, then one cannot consider cases in isolation and shift cases in time.

Moreover, event data can also be stored in aggregated form as is done for streaming process mining where one cannot keep track of all events and all cases due to memory constraints and the need to provide answers in real-time [5,7,29]. Aging data structures, queues, time windows, sampling, hashing, etc. can be used to keep only the information necessary to instantly provide answers to selected questions. Such approaches can also be used to ensure confidentiality, often without a significant loss of accuracy.

Table 1. Relating the four challenges to process mining specific tasks (1/2).

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Fairness Data Science without prejudice: How to avoid unfair conclusions even if they are true?	The input data may be biased, incomplete or incorrect such that the analysis reconfirms prejudices. By resampling or relabeling the data, undesirable forms of discrimination can be avoided. Note that both cases and resources (used to execute activities) may refer to individuals having sensitive attributes such as race, gender, age, etc.	The discovered model may abstract from paths followed by certain under-represented groups of cases. Discrimination-aware process-discovery algorithms can be used to avoid this. For example, if cases are handled differently based on gender, we may want to ensure that both are equally represented in the model'	Conformance checking can be used to "blame" individuals, groups, or organizations for deviating from some normative model. Discrimination-aware conformance checking (e.g., alignments) needs to separate (1) likelihood, (2) severity and (3) blame. Deviations may need to be interpreted differently for different groups of cases and resources	Straightforward performance measurements may be unfair for certain classes of cases and resources (e.g., not taking into account the context). Discrimination-aware performance analysis detects unfairness and supports process improvements taking into account trade-offs between internal fairness (worker's perspective) and external fairness (citizen/patient/customer's perspective)	Process-related predictions, recommendations and decisions may discriminate (un)intentionally. This problem can be tackled using techniques from discrimination-aware data mining
Accuracy Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?	Event data (e.g., XES files) may have all kinds of quality problems. Attributes may be incorrect, imprecise, or uncertain. For example, timestamps may be too coarse (just the date) or reflect the time of recording rather than the time of the event's occurrence	Process discovery depends on many parameters and characteristics of the event log. Process models should better show the confidence level of the different parts. Moreover, additional information needs to be used better (domain knowledge, uncertainty in event data, etc.)	Often multiple explanations are possible to interpret non-conformance. Just providing one alignment based on a particular cost function may be misleading. How robust are the findings?	In case of fitness problems (process model and event log disagree), performance analysis is based on assumptions and needs to deal with missing values (making results less accurate)	Inaccurate process models may lead to flawed predictions, recommendations and decisions. Moreover, not communicating the (un)certainty of predictions, recommendations and decisions, may negatively impact processes

Table 2. Relating the four challenges to process mining specific tasks (2/2).

	Creating and managing event data	Process discovery	Conformance checking	Performance analysis	Operational support
Confidentiality Data Science that ensures confidentiality: How to answer questions without revealing secrets?	Event data (e.g., XES files) may reveal sensitive information. Anonymization and de-identification can be used to avoid disclosure. Note that timestamps and paths may be unique and a source for re-identification (e.g., all paths are unique)	The discovered model may reveal sensitive information, especially with respect to infrequent paths or small event logs. Drilling-down from the model may need to be blocked when numbers get too small (cf. k-anonymity)	Conformance checking shows diagnostics for deviating cases and resources. Access-control is important and diagnostics need to be aggregated to avoid revealing compliance problems at the level of individuals	Performance analysis shows bottlenecks and other problems. Linking these problems to cases and resources may disclose sensitive information	Process-related predictions, recommendations and decisions may disclose sensitive information, e.g., based on a rejection other properties can be derived
Transparency Data Science that provides transparency: How to clarify answers such that they become indisputable?	Provenance of event data is key. Ideally, process mining insights can be related to the event data they are based on. However, this may conflict with confidentiality concerns	Discovered process models depend on the event data used as input and the parameter settings and choice of discovery algorithm. How to ensure that the process model is interpreted correctly? End-users need to understand the relation between data and model to trust analysis	When modeled and observed behavior disagree there may be multiple explanations. How to ensure that conformance diagnostics are interpreted correctly?	When detecting performance problems, it should be clear how these were detected and what the possible causes are. Animating event logs on models helps to make problems more transparent	Predictions, recommendations and decisions are based on process models. If possible, these models should be transparent. Moreover, explanations should be added to predictions, recommendations and decisions ("We predict that this case be late, because ...")

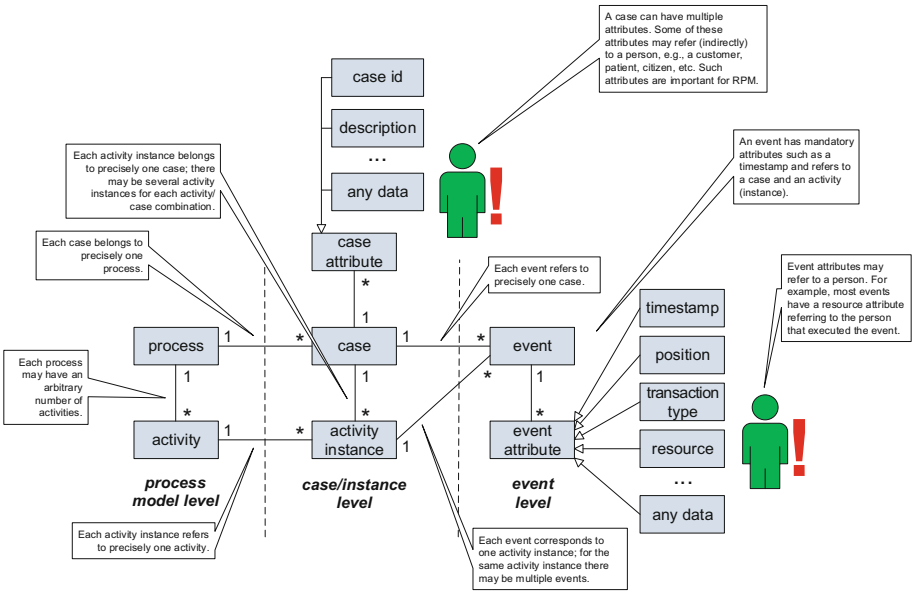

Fig. 10. The typical ingredients of an event log described in terms of a class model highlighting data elements referring to individuals.

5.3 Example: Accuracy and Process Discovery

As mentioned in Table 1 the accuracy of a discovery process model may depend on a variety parameter settings. A small change in the input data (log or settings) may completely change the result. One of the main problems of existing techniques is that they do not indicate any form of confidence level. Often parts of the model can be discovered with great certainty whereas other parts are unclear and the discovery technique is basically guessing. Nevertheless, this uncertainty is seldom shown in the model and may lead to incorrect conclusions. To support RPM, we need to develop process discovery techniques that indicate confidence information in the models returned.

5.4 Example: Transparency and Conformance Checking

Conformance checking [6] can be viewed as a classification problem. What kinds of cases deviate at a particular point? However, if model and log disagree, then there may be multiple explanations for each deviation. For example, there may be multiple log-model "alignments" having the same costs. Moreover, the costs assigned to deviations may be arbitrary. As mentioned in Table 2 it is vital that conformance diagnostics are interpreted correctly. Moreover, the "process mining pipeline" (Fig. 8) needs to be managed carefully to avoid misleading conclusions caused by, for example, data preparation problems.

5.5 Example: Fairness and Performance Analysis

Process mining provides the ability to show and analyze bottlenecks in processes with minimal effort. Bottleneck analysis can also be formulated as a classification problem. Which cases get delayed more than 5 days? Who worked on these delayed cases? Performance problems can be related to characteristics of the case (e.g., a citizen or customer) or the people that worked on it. The process itself may be "unfair" (discriminate workers or cases) or decision makers can make "unfair" conclusions based on a superficial analysis of the data. Table 1 mentions *internal* fairness (worker's perspective) and *external* fairness (citizen/patient/customer's perspective) as two concerns. Note that the employee that takes all difficult cases may be slower than others. Evaluating employees without taking such context into account will lead to unjustified conclusions.

The above examples illustrate that our classification can be used to identify a range of novel research challenges in process mining.

6 Epilogue

This paper introduced the notion of *"Responsible Data Science"* (RDS) from four angles: *fairness*, *accuracy*, *confidentiality*, and *transparency*. We advocate the development and use of positive technological solutions rather than relying on stricter regulations like the *General Data Protection Regulation* (GDPR) approved by the EU Parliament in April 2016 [10]. GDPR aims to strengthen and unify data protection for individuals and replaces Directive 95/46/EC [12]. GDPR is far more restrictive than earlier legislation. Sanctions include fines of up to 4% of the annual worldwide turnover.

GDPR and other forms of legislation can be seen as environmental laws protecting society against "pollution" caused by irresponsible data use. However, legislation may also prevent the use of data (science) in applications where incredible improvements are possible. Simply prohibiting the collection and systematic use of data would be turning back the clock. Next to legislation, positive technological solutions are needed to ensure RDS. Green data science needs technological breakthroughs, just like the innovations enabling green energy.

The paper also discussed the four "FACT" challenges in the context of process mining. In today's society, *event data* are collected about anything, at any time, and at any place. Today's process mining tools are able to analyze such data and can handle event logs with billions of events. These amazing capabilities also imply a great responsibility. Fairness, accuracy, confidentiality, and transparency should be key concerns for any process miner. There is a need for a new generation of process mining techniques and tools that are responsible by design. However, sometimes painful trade-offs are inevitable. Figure 5 and Table 1 both show the need for trade-offs between fairness and accuracy. Other trade-offs are needed between confidentiality and transparency (see Fig. 6 and Table 2).

We invite researchers and practitioners to contribute to RDS and RPM. These topics are urgent: without proper tools and approaches the use of data may come to a grinding hold. People like Michael Jordan warned for a

"Big data winter", due to the simple-minded and statistically unsound approaches used today. Irresponsible uses of data (science) may trigger restrictive laws and effectuate resistance of customers and citizens.

Acknowledgements. This work is partly based by discussions in the context of the *Responsible Data Science* (RDS) collaboration involving principal scientists from Eindhoven University of Technology, Leiden University, University of Amsterdam, Radboud University Nijmegen, Tilburg University, VU University, Amsterdam Medical Center, VU Medical Center, Leiden University Medical Center, Delft University of Technology, and CWI.

References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Berlin (2011)
2. van der Aalst, W.M.P., Management, B.P.: A comprehensive survey. ISRN Softw. Eng. 1–37 (2013). doi:10.1155/2013/507984
3. Aalst, W.M.P.: Data scientist: the engineer of the future. In: Mertins, K., Bénaben, F., Poler, R., Bourrières, J.-P. (eds.) Enterprise Interoperability VI. PIC, vol. 7, pp. 13–26. Springer, Cham (2014). doi:10.1007/978-3-319-04948-9_2
4. van der Aalst, W.M.P.: Green data science: using big data in an "environmentally friendly" manner. In: Camp, O., Cordeiro, J. (eds.) Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016), pp. 9–21. Science and Technology Publications, Portugal (2016)
5. van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, Berlin (2016)
6. van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis. WIREs Data Mining Knowl. Discov. **2**(2), 182–192 (2012)
7. Burattin, A., Sperduti, A., van der Aalst, W.M.P.: Control-flow discovery from event streams. In: IEEE Congress on Evolutionary Computation (CEC 2014), pp. 2420–2427. IEEE Computer Society (2014)
8. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-aware classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010)
9. Casella, G., Berger, R.L.: Statistical Inference, 2nd edn. Duxbury Press, Delhi (2002)
10. Council of the European Union. General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, April 2016
11. Donoho, D.: 50 years of Data Science. Technical report, Stanford University, September 2015. Based on a Presentation at the Tukey Centennial Workshop, Princeton, NJ, 18 September 2015
12. European Commission: Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Wegard to the Processing of Personal Data and on the Free Movement of Such Data. Official Journal of the European Communities, No L 281/31, October 1995

13. IEEE Task Force on Process Mining: XES Standard Definition (2013). www.xes-standard.org
14. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination-aware decision-tree learning. In: Proceedings of the IEEE International Conference on Data Mining (ICDM 2010), pp. 869–874 (2010)
15. Koops, B.J., Oosterlaken, I., Romijn, H., Swierstra, T., Van den Hoven, J. (eds.): Responsible Innovation 2: Concepts, Approaches, and Applications. Springer, Berlin (2015)
16. Leemans, S.J.J., Fahland, D., Aalst, W.M.P.: Exploring processes and deviations. In: Fournier, F., Mendling, J. (eds.) BPM 2014. LNBIP, vol. 202, pp. 304–316. Springer, Cham (2015). doi:10.1007/978-3-319-15895-2_26
17. Miller, R.G.: Simultaneous Statistical Inference. Springer, Berlin (1981)
18. Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., Pedreschi, D.: Privacy-by-design in big data analytics and social mining. EPJ Data Sci. **1**(10), 1–26 (2014)
19. Naur, P.: Concise Survey of Computer Methods. Studentlitteratur Lund, Akademisk Forlag, Kobenhaven (1974)
20. Nelson, G.S.: Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. Paper 1884–2015, ThotWave Technologies, Chapel Hill (2015)
21. Owen, R., Bessant, J., Heintz, M. (eds.): Responsible Innovation. Wiley, Hoboken (2013)
22. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568. ACM (2008)
23. President's Council of Advisors on Science and Technology: Big Data and Privacy: A Technological Perspective (Report to the President). Executive Office of the President, US-PCAST, May 2014
24. Press, G.: A very short history of data science. Forbes Technology (2013). http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/
25. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2011)
26. Ruggieri, S., Pedreshi, D., Turini, F.: DCUBE: discrimination discovery in databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1127–1130. ACM (2010)
27. Tukey, J.W.: The future of data analysis. Ann. Math. Stat. **33**(1), 1–67 (1962)
28. Vigen, T.: Spurious Correlations. Hachette Books, New York (2015)
29. van Zelst, S.J., van Dongen, B.F., van der Aalst, W.M.P.: Know what you stream: generating event streams from CPN models in ProM 6. In: Proceedings of the BPM2015 Demo Session. CEURWorkshop Proceedings, vol. 1418, pp. 85–89 (2015). http://ceur-ws.org/