Green Data Science
Using Big Data in an "Environmentally Friendly" Manner

Wil M.P. van der Aalst
Eindhoven University of Technology, Department of Mathematics and Computer Science,
PO Box 513, NL-5600 MB Eindhoven, The Netherlands.
w.m.p.v.d.aalst@tue.nl

Keywords: Data Science, Big Data, Fairness, Confidentiality, Accuracy, Transparency, Process Mining.

Abstract: The widespread use of "Big Data" is heavily impacting organizations and individuals for which these data are collected. Sophisticated data science techniques aim to extract as much value from data as possible. Powerful mixtures of Big Data and analytics are rapidly changing the way we do business, socialize, conduct research, and govern society. Big Data is 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 illustrate that data science techniques may lead to new forms of "pollution". We use the term "Green Data Science" for technological solutions that enable individuals, organizations and society to reap the benefits from the widespread availability of data while ensuring fairness, confidentiality, accuracy, and transparency. To illustrate the scientific challenges related to "Green Data Science", we focus on process mining as a concrete example. 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. Therefore, this paper poses the question: How to benefit from process mining while avoiding "pollutions" related to unfairness, undesired disclosures, inaccuracies, and non-transparency?

1 INTRODUCTION

In recent years, data science emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. We use the following definition: *"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, 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."* (Aalst, 2016).

Related to data science is the overhyped term "Big Data" that is used to refer to the massive amounts of data collected. Organizations are heavily investing in Big Data technologies, but at the same time citizens, patients, customers, and employees are concerned about the use of their data. We live in an era characterized by unprecedented opportunities to sense, store, and analyze data related to human activities in great detail and resolution. This introduces new risks and intended or unintended abuse enabled by powerful analysis techniques. Data may be sensitive and personal, and should not be revealed or used for proposes different from what was agreed upon. Moreover, analysis techniques may discriminate minorities even when attributes like gender and race are removed. Using data science technology as a "black box" making life-changing decisions (e.g., medical prioritization or mortgage approvals) triggers a variety of ethical dilemmas.

Sustainable data science is only possible when citizens, patients, customers, and employees are *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

W.M.P. van der Aalst. Green Data Science: Using Big Data in an
Environmentally Friendly Manner. Proceedings of the 18th International
Conference on Enterprise Information Systems (ICEIS 2016), INSTICC,
Rome, April 2016.

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.

In this paper, we coin the term *"Green Data Science"* (GDS) to refer to the collection of techniques and approaches trying to reap the benefits of data science and Big Data while ensuring fairness, confidentiality, accuracy, and transparency. *We believe that technological solutions can be used to avoid pollution and protect the environment in which data is collected and used.*

Section 2 elaborates on the following four challenges:

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

Concerns related to privacy and personal data protection triggered legislation like the EU's Data Protection Directive. *Directive 95/46/EC* ("on the protection of individuals with regard to the processing of personal data and on the free movement of such data") of the European Parliament and the Council was adopted on 24 October 1995 (European Commission, 1995). The *General Data Protection Regulation* (GDPR) is currently under development and aims to strengthen and unify data protection for individuals within the EU (European Commission, 2015). GDPR will replace Directive 95/46/EC and is expected to be finalized in Spring 2016 and will be much more restrictive than earlier legislation. Sanctions include fines of up to 4% of the annual worldwide turnover. GDPR and other forms of legislation limiting the use of data, may prevent the use of data science also in situations where data is used in a positive manner. Prohibiting the collection and systematic use of data is like turning back the clock. Next to legislation, positive technological solutions are needed to ensure fairness, confidentiality, accuracy, and transparency. By just imposing restrictions, individuals, organizations and society cannot exploit data (science) in a positive way.

The four challenges discussed in Section 2 are quite general. Therefore, we focus on a concrete subdiscipline in data science in Section 3: *Process Mining* (Aalst, 2011). Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). Event data are related to explicit process models, e.g., Petri nets or BPMN models. For example, process models are discovered from event data or event data are replayed on models to analyze compliance and performance. *Process mining provides a bridge between data-driven approaches (data mining, machine learning and business intelligence) and process-centric approaches (business process modeling, model-based analysis, and business process management/reengineering).* Process mining results may drive redesigns, show the need for new controls, trigger interventions, and enable automated decision support. Individuals *inside* (e.g., end-users and workers) and *outside* (e.g., customers, citizens, or patients) the organization may be impacted by process mining results. Therefore, Section 3 lists process mining challenges related to fairness, confidentiality, accuracy, and transparency.

In the long run, data science is only sustainable if we are willing to address the problems discussed in this paper. Rather than abandoning the use of data altogether, we should find positive technological ways to protect individuals.

2 FOUR CHALLENGES

Figure 1 sketches 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.

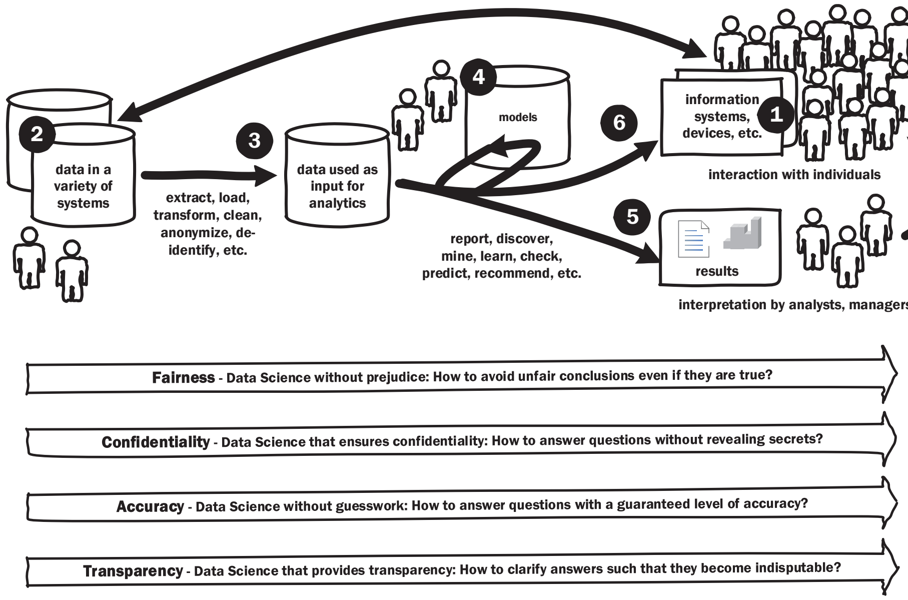

2
data in a variety of systems
extract, load, transform, clean, anonymize, de-identify, etc.
3
data used as input for analytics
4
models
report, discover, mine, learn, check, predict, recommend, etc.
6
information systems, devices, etc.
1
interaction with individuals
5
results
7
interpretation by analysts, managers, etc.
Fairness - Data Science without prejudice: How to avoid unfair conclusions even if they are true?
Confidentiality - Data Science that ensures confidentiality: How to answer questions without revealing secrets?
Accuracy - Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?
Transparency - Data Science that provides transparency: How to clarify answers such that they become indisputable?

Figure 1: The "data science pipeline" facing four challenges.

Figure 1 also lists the four challenges discussed in the remainder of this section. 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.

2.1 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 good 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.

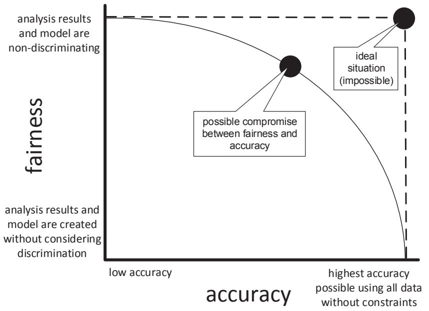


Figure 2: Tradeoff between fairness and accuracy.

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 Figure 2). 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 (Pedreshi et al., 2008). Since then, several papers mostly focusing on fair classification appeared: (Calders and Verwer, 2010; Kamiran et al., 2010; Ruggieri et al., 2010). 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.

2.2 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) currently under development (European Commission, 2015) focuses on personal information: *"The principles of data protection should apply to any information concerning an identified or identifiable natural person. Data including pseudonymized data, which could be attributed to a natural person by the use of additional information, should be considered as information on an identifiable natural person. To determine whether a person is identifiable, account should be taken of all the means reasonably likely to be used either by the controller or by any other person to identify the individual directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development. The principles of data protection should therefore not apply to anonymous information, that is information which does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way 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 Figure 1 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 Figure 3). 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.

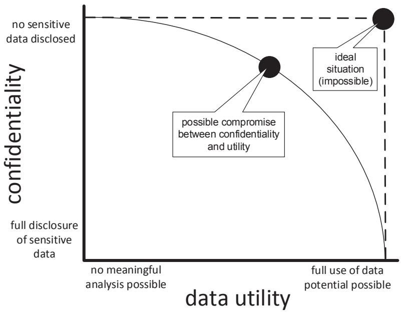

no sensitive data disclosed
confidentiality
full disclosure of sensitive data
possible compromise between confidentiality and utility
ideal situation (impossible)
no meaningful analysis possible
full use of data potential possible
data utility

Figure 3: Tradeoff between confidentiality and utility.

Access rights to the different types of data and analysis results in the "data science pipeline" (Figure 1) 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 Figure 1. Cybersecurity measures should not complicate access, data preparation, and analysis; otherwise people may start using illegal copies and replicate data.

See (Monreale et al., 2014; Nelson, 2015; President's Council, 2014) for approaches to ensure confidentiality.

2.3 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 (Vigen, 2015)):

- 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.

According to *Bonferroni's principle* we need to avoid treating random observations as if they are real and significant (Rajaraman and Ullman, 2011). The following example, inspired by a similar example in (Rajaraman and Ullman, 2011), 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 $\frac{18\times10^6}{100\times100} = 1800$ 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: Bonferroni's principle explained using an example taken from (Aalst, 2016). 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 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 (Casella and Berger, 2002). If we test many hypotheses, we also increase the likelihood of a rare event. Hence, the likelihood of incorrectly rejecting a null hypothesis increases (Miller, 1981). 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.

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 tend to be too "crisp" (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.

2.4 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 Figure 1) need to be as accurate as possible (e.g., to reduce costs and delays). Analysis results (❺ in Figure 1) 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*.

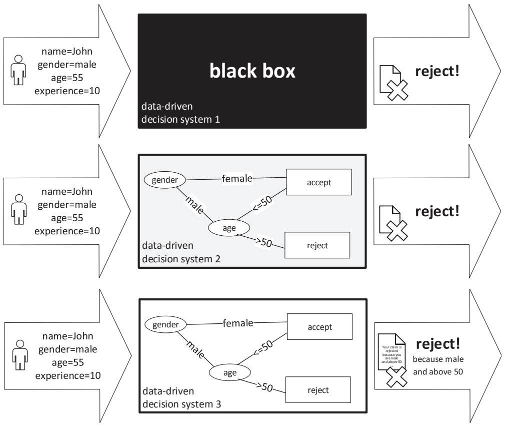


Figure 4: Different levels of transparency.

Figure 4 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 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 challenges depicted in Figure 1 are clearly interrelated. There may be trade-offs between *fairness*, *confidentiality*, *accuracy* and *transparency*. For example, to ensure confidentiality we may add noise and de-identify data thus possibly compromising accuracy and transparency.

3 EXAMPLE: GREEN PROCESS MINING

The goal of *process mining* is to turn event data into insights and actions (Aalst, 2016). 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

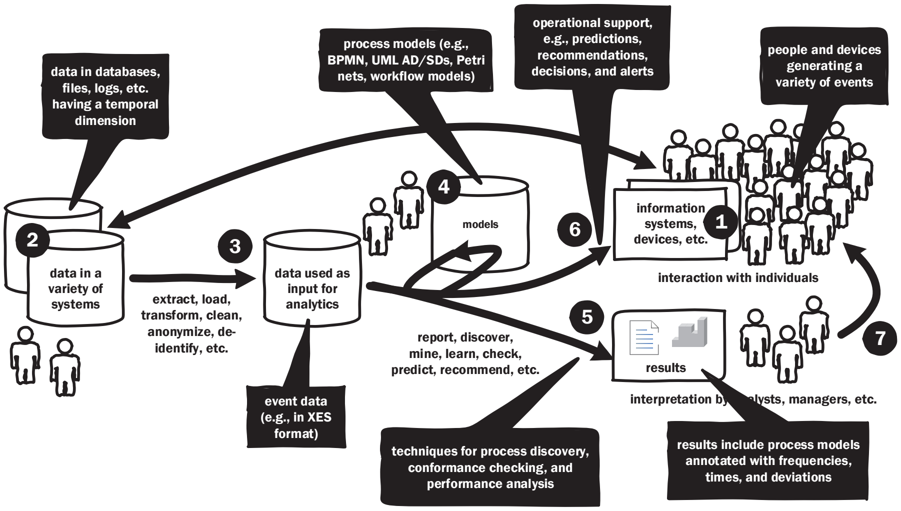

data in databases, files, logs, etc. having a temporal dimension
process models (e.g., BPMN, UML AD/SDs, Petri nets, workflow models)
operational support, e.g., predictions, recommendations, decisions, and alerts
people and devices generating a variety of events
4
models
6
information systems, devices, etc.
1
2
data in a variety of systems
3
data used as input for analytics
extract, load, transform, clean, anonymize, de-identify, etc.
interaction with individuals
5
results
7
report, discover, mine, learn, check, predict, recommend, etc.
interpretation by analysts, managers, etc.
event data (e.g., in XES format)
techniques for process discovery, conformance checking, and performance analysis
results include process models annotated with frequencies, times, and deviations

Figure 5: The "process mining pipeline" relates observed and modeled behavior.

process agonistic whereas process science approaches tend to be model-driven without considering the "evidence" hidden in the data. This section discusses challenges related to fairness, confidentiality, accuracy, 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 when considering processes and event data.*

3.1 What Is Process Mining?

Figure 5 shows the "process mining pipeline" and can be viewed as a specialization of the Figure 1. 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) (Aalst, 2013a) and data science (Aalst, 2014). 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 open-source process mining tool supporting process discovery, conformance checking, social network analysis, organizational mining, clustering, decision mining, prediction, and recommendation (see Figure 6). 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`).

3.1.1 Creating and Managing Event Data

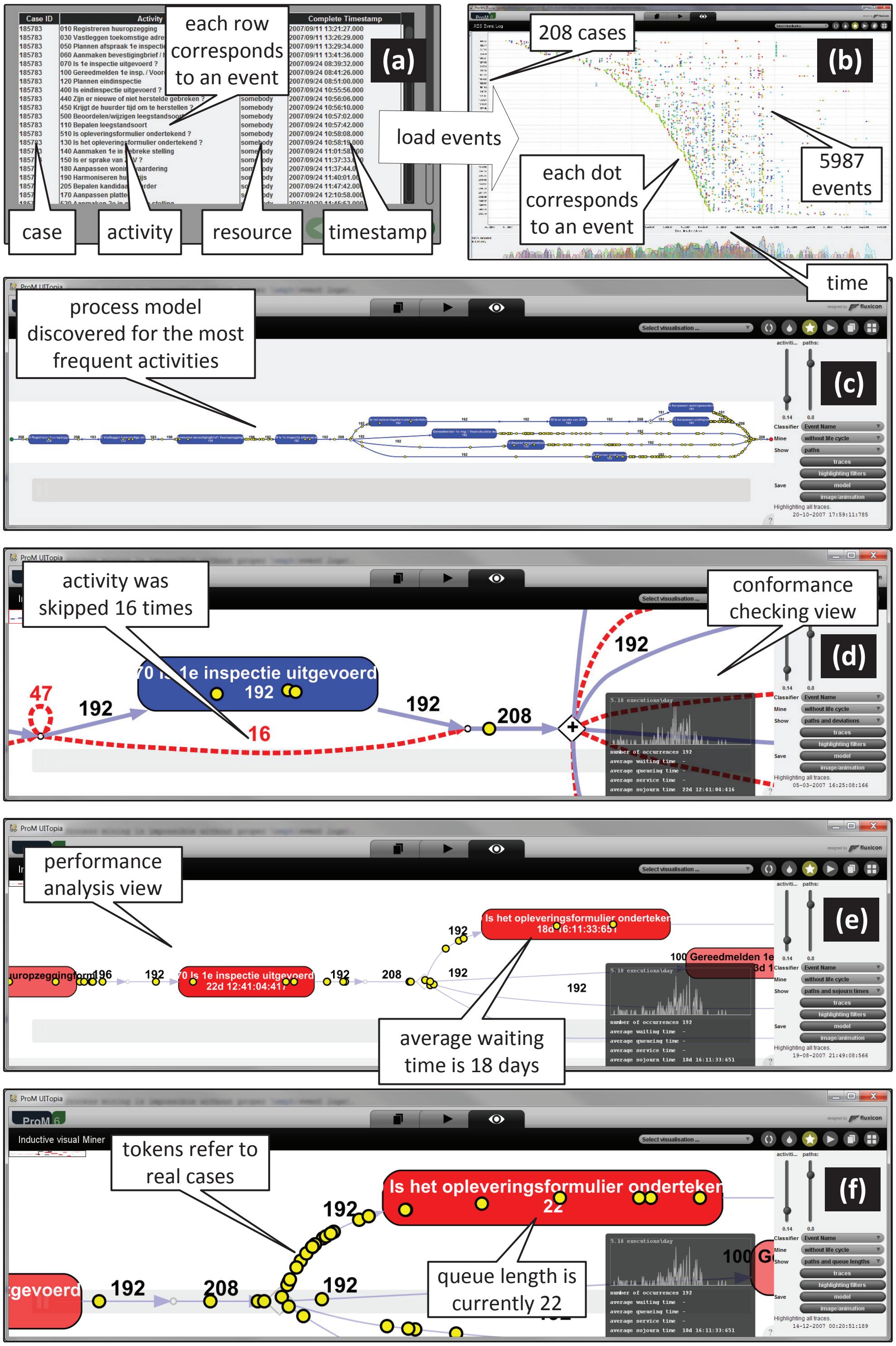
Process mining is impossible without proper *event logs* (Aalst, 2011). An event log contains event data related to a particular process. Each event in an event log refers to one *process instance*, called *case*. Events related to a case are ordered. Events can have attributes. Examples of typical attribute 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 6(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) (IEEE Task Force on Process Mining, 2013). 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 in a variety of systems (see ❸ in Figure 5). Consider for example the hundreds of tables in a typical HIS (Hospital Information System) like ChipSoft, 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.

3.1.2 Process Discovery

Process discovery is one of the most challenging process mining tasks (Aalst, 2011). Based on an event log, a process



Case ID	Activity		Complete Timestamp
185783	010 Registreren huuropzegging		2007/09/11 13:21:27.000
185783	030 Vastleggen toekomstige adre		2007/09/11 13:26:29.000
185783	050 Plannen afspraak 1e inspectie		2007/09/11 13:29:34.000
185783	060 Aanmaken bevestigingbrief /		2007/09/11 13:41:36.000
185783	070 Is 1e inspectie uitgevoerd ?		2007/09/24 08:39:32.000
185783	100 Gereedmelden 1e insp. / Voor		2007/09/24 08:41:26.000
185783	120 Plannen eindinspectie		2007/09/24 08:51:00.000
185783	400 Is eindinspectie uitgevoerd ?		2007/09/24 10:55:56.000
185783	440 Zijn er nieuwe of niet herstelde gebreken ?	somebody	2007/09/24 10:56:06.000
185783	450 Krijgt de huurder tijd om te herstellen ?	somebody	2007/09/24 10:56:10.000
185783	500 Beoordelen/wijzigen leegstandsoort	somebody	2007/09/24 10:57:02.000
185783	110 Bepalen leegstandsoort	somebody	2007/09/24 10:57:42.000
185783	510 Is opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:08.000
185783	130 Is het opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:19.000
1857 3	140 Aanmaken 1e in gebreke stelling	some ody	2007/09/24 11:01:58. 00
1857	150 Is er sprake van Z V ?	some ody	2007/09/24 11:37:33. 0
1857	180 Aanpassen wonin waardering	som ody	2007/09/24 11:37:44.0
1857	190 Harmoniseren hu ijs	so ody	2007/09/24 11:40:01.00
1857	205 Bepalen kandidaa rder	so ody	2007/09/24 11:47:42.00
1857	170 Aanpassen platte	s ody	2007/09/24 12:10:58.000

```
5.18 executions\day
number of occurrences 192
average waiting time  -
average queueing time -
average service time  -
average sojourn time  22d 12:41:04:416
```

Highlighting all traces.
05-03-2007 16:25:08:166

```
5.18 executions\day
number of occurrences 192
average waiting time  -
average queueing time -
average service time  -
average sojourn time  18d 16:11:33:651
```

Highlighting all traces.
19-08-2007 21:49:08:566

```
5.18 executions\day
number of occurrences 192
average waiting time  -
average queueing time -
average service time  -
average sojourn time  18d 16:11:33:651
```

Highlighting all traces.
14-12-2007 00:20:51:189

Highlighting all traces.
20-10-2007 17:59:11:785

(a)
each row corresponds to an event
case
activity
resource
timestamp
load events
(b)
208 cases
5987 events
each dot corresponds to an event
time
(c)
process model discovered for the most frequent activities
ProM UITopia
Select visualisation ...
activiti... paths:
0.14 0.8
Classifier Event Name
Mine without life cycle
Show paths
traces
highlighting filters
Save model
image/animation
(d)
activity was skipped 16 times
conformance checking view
47
192
70 Is 1e inspectie uitgevoerd 192
16
192
208
192
Show paths and deviations
(e)
performance analysis view
uropzeggingformul 196
192
70 Is 1e inspectie uitgevoerd 22d 12:41:04:417
192
208
192
Is het opleveringsformulier ondertekend 18d 16:11:33:651
100 Gereedmelden 1e 3d 1
192
average waiting time is 18 days
Show paths and sojourn times
(f)
ProM 6
Inductive visual Miner
tokens refer to real cases
gevoerd
192
208
192
192
Is het opleveringsformulier ondertekend 22
100 G
queue length is currently 22
Show paths and queue lengths
designed by fluxicon

Figure 6: 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)**.

Table 1: Relating the four challenges to process mining specific tasks.

	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.
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.
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.
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 ...”).

model is constructed thus capturing the behavior seen in the log. Dozens of process discovery algorithms are available. Figure 6(c) shows a process model discovered using ProM's *inductive visual miner* (Leemans et al., 2015). Techniques use Petri nets, WF-nets, C-nets, process trees, or transition systems as a representational bias (Aalst, 2016). These results can always be converted to the desired notation, for example BPMN (Business Process Model and Notation), YAWL, or UML activity diagrams.

3.1.3 Conformance Checking

Using conformance checking discrepancies between the log and the model can be detected and quantified by replaying the log (Aalst et al., 2012). For example, Figure 6(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.

3.1.4 Performance Analysis

By replaying event logs on process model, we can compute frequencies and waiting/service times. Using alignments (Aalst et al., 2012) 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 6(d) shows an activity that has an average waiting time of 18 days and 16 hours. Note that such bottlenecks are discovered without any modeling.

3.1.5 Operational Support

Figure 6(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 violated.

3.2 Challenges in Process Mining

Table 1 maps the four generic challenges identified in Section 2 onto the six key ingredients of process mining briefly introduced in Section 3.1. Note that both cases (i.e., process instances) and the resources used to execute activities may refer to individuals (customers, citizens, patients, workers, etc.). Event data are difficult to fully anonymize. In larger processes, most cases follow a unique path. In the event log used in Figure 6, 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 deanonymize or re-identify cases. The same holds for (precise) timestamps. For the event log in Figure 6, 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.

Conformance checking (Aalst et al., 2012) can be viewed as a classification problem. What kind of cases deviate at a particular point? Bottleneck analysis can also be formulated as a classification problem. Which cases get delayed more than 5 days? We may find out that conformance or performance problems are caused by characteristics of the case itself or the people that worked on it. This allows us to discover patterns such as:

- Doctor Jones often performs an operation without making a scan and this results in more incidents later in the process.
- Insurance claims from older customers often get rejected because they are incomplete.
- Citizens that submit their tax declaration too late often get rejected by teams having a higher workload.

Techniques for *discrimination discovery* can be used to find distinctions that are not desirable/acceptable. Subsequently, techniques for *discrimination prevention* can be used to avoid such situations. It is important to note that discrimination is not just related to static variables, but also relates to the way cases are handled.

It is also interesting to use techniques from decomposed process mining or streaming process mining (see Chapter 12 in (Aalst, 2016)) to make process mining "greener".

For *streaming process mining* one cannot keep track of all events and all cases due to memory constraints and the need to provide answers in real-time (Burattin et al., 2014; Aalst, 2016; Zelst et al., 2015). Hence, event data need to be stored in aggregated form. 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.

For *decomposed/distributed process mining* event data need to be split based on a grouping activities in the process (Aalst, 2013b; Aalst, 2016). After splitting the event log, it is still possible to discover process models and to check conformance. Interestingly, the sublogs can be analyzed separately. This may be used to break potential harmful correlations. Rather than storing complete cases, one can also store shorter episodes of anonymized case fragments. Sometimes it may even be sufficient to store only *direct successions*, i.e., facts of the form "for some unknown case activity *a* was followed by activity *b* with a delay of 8 hours". Some discovery algorithms only use data on direct successions and do not require additional, possibly sensitive, information. Of course certain questions can no longer be answered in a reliable manner (e.g., flow times of cases).

The above examples illustrate that Table 1 identifies a range of novel research challenges in 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, confidentiality, accuracy and transparency should be key concerns for any process miner.

4 CONCLUSION

This paper introduced the notion of *"Green Data Science"* (GDS) from four angles: *fairness*, *confidentiality*, *accuracy*, and *transparency*. The possible "pollution" caused by data science should not be addressed (only) by legislation. We should aim for positive, technological solutions to protect individuals, organizations and society against the negative side-effects of data. As an example, we discussed "green challenges" in *process mining*. Table 1 can be viewed as a *research agenda* listing interesting open problems.

REFERENCES

Aalst, W. van der (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin.

Aalst, W. van der (2013a). Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, pages 1–37. doi:10.1155/2013/507984.

Aalst, W. van der (2013b). Decomposing Petri Nets for Process Mining: A Generic Approach. *Distributed and Parallel Databases*, 31(4):471–507.

Aalst, W. van der (2014). Data Scientist: The Engineer of the Future. In Mertins, K., Benaben, F., Poler, R., and Bourrieres, J., editors, *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer-Verlag, Berlin.

Aalst, W. van der (2016). *Process Mining: Data Science in Action*. Springer-Verlag, Berlin.

Aalst, W. van der, Adriansyah, A., and Dongen, B. van (2012). Replaying History on Process Models for Conformance Checking and Performance Analysis. *WIREs Data Mining and Knowledge Discovery*, 2(2):182–192.

Burattin, A., Sperduti, A., and Aalst, W. van der (2014). Control-Flow Discovery from Event Streams. In *IEEE Congress on Evolutionary Computation (CEC 2014)*, pages 2420–2427. IEEE Computer Society.

Calders, T. and Verwer, S. (2010). Three Naive Bayes Approaches for Discrimination-Aware Classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Casella, G. and Berger, R. (2002). *Statistical Inference, 2nd Edition*. Duxbury Press.

European Commission (1995). 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.

European Commission (2015). Proposal for a Regulation 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 (General Data Protection Regulation). 9565/15, 2012/0011 (COD).

IEEE Task Force on Process Mining (2013). XES Standard Definition. www.xes-standard.org.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination-Aware Decision-Tree Learning. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2010)*, pages 869–874.

Leemans, S., Fahland, D., and Aalst, W. van der (2015). Exploring Processes and Deviations. In Fournier, F. and Mendling, J., editors, *Business Process Management Workshops, International Workshop on Business Process Intelligence (BPI 2014)*, volume 202 of *Lecture Notes in Business Information Processing*, pages 304–316. Springer-Verlag, Berlin.

Miller, R. (1981). *Simultaneous Statistical Inference*. Springer-Verlag, Berlin.

Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., and Pedreschi, D. (2014). Privacy-By-Design in Big Data Analytics and Social Mining. *EPJ Data Science*, 1(10):1–26.

Nelson, G. (2015). Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. Paper 1884-2015, ThotWave Technologies, Chapel Hill, NC.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568. ACM.

President's Council of Advisors on Science and Technology (2014). Big Data and Privacy: A Technological Perspective (Report to the President). Executive Office of the President, US-PCAST.

Rajaraman, A. and Ullman, J. (2011). *Mining of Massive Datasets*. Cambridge University Press.

Ruggieri, S., Pedreshi, D., and Turini, F. (2010). DCUBE: Discrimination Discovery in Databases. In *Proceedings of the ACM SIGMOD Intetrnational Conference on Management of Data*, pages 1127–1130. ACM.

Zelst, S. van, Dongen, B. van, and Aalst, W. van der (2015). Know What You Stream: Generating Event Streams from CPN Models in ProM 6. In *Proceedings of the BPM2015 Demo Session*, volume 1418 of *CEUR Workshop Proceedings*, pages 85–89. CEUR-WS.org.

Vigen, T. (2015). *Spurious Correlations*. Hachette Books.